Oxford Textbook of Medicine 4th edition (March 2003): by David A. Warrell (Editor), Timothy M. Cox (Editor), John D. Firth (Editor), Edward J., J R., M.D. Benz (Editor) By Oxford Press;

By OkDoKeY

Oxford Textbook of Medicine

TABLE OF CONTENTS

Volume 1

Geoffrey L. Smith

7.10.5 Mumps: epidemic parotitis
B. K. Rima

7.10.6 Measles
H.C. Whittle and P. Aaby

7.10.6.1 Nipah and Hendra viruses
James G. Olson

7.10.7 Enterovirus infections
Ulrich Desselberger and Philip Minor

7.10.8 Virus infections causing diarrhoea and vomiting
Ulrich Desselberger

7.10.9 Rhabdoviruses: rabies and rabies-related viruses
M. J. Warrell and D.A. Warrell

7.10.10 Colorado tick fever and other arthropod-borne reoviruses
M.J. Warrell and D.A. Warrell

7.10.11 Alphaviruses
L. R. Petersen and D. J. Gubler

7.10.12 Rubella
P. A. Tookey and S. Logan

7.10.13 Flaviviruses
L. R. Petersen and D. J. Gubler

7.10.14 Bunyaviridae
J. W. LeDuc and J. S. Porterfield

7.10.15 Arenaviruses
Susan Fisher-Hoch and Joseph McCormick

7.10.16 Filoviruses
Susan Fisher-Hoch and Joseph McCormick

7.10.17 Papovaviruses
K. V. Shah

7.10.18 Parvovirus B19
B. J. Cohen

7.10.19 Hepatitis viruses (including TTV)
N. V. Naoumov

7.10.20 Hepatitis C virus
D. L. Thomas

7.10.21 HIV and AIDS
G. A. Luzzi, T. E. A. Peto, R. A. Weiss, and C. P. Conlon

7.10.22 HIV in the developing world
Charles F. Gilks

7.10.23 HTLV-I and II and associated diseases
C. R. M. Bangham, M. Osame, and S. Nightingale

7.10.24 Viruses and cancer
R.A. Weiss

7.10.25 Orf
N. Jones

7.10.26 Molluscum contagiosum
N. Jones

7.11 Bacteria

7.11.1 Diphtheria
Delia Bethell and Tran Tin Hien

7.11.2 Streptococci and enterococci
S. J. Eykyn

7.11.3 Pneumococcal diseases
Keith P. Klugman and Brian M. Greenwood

7.11.4 Staphylococci
S. J. Eykyn

7.11.5 Meningococcal infections
P. Brandtzaeg

7.11.6 Neisseria gonorrhoeae
D. Barlow and C. Ison

7.11.7 Enterobacteria, campylobacter, and miscellaneous food-poisoning bacteria
G. T. Keusch and M. B. Skirrow

7.11.8 Typhoid and paratyphoid fevers

J. Richens and C. Parry

7.11.9 Intracellular Klebsiella infections
J. Richens

7.11.10 Anaerobic bacteria
S. J. Eykyn

7.11.11 Cholera
Michael L. Bennish

7.11.12 Haemophilus influenzae
E. R. Moxon

7.11.13 *Haemophilus ducreyi* and chancroid
Allan R. Ronald

7.11.14 Bordetella
Calvin C. Linnemann, Jr

7.11.15 Melioidosis and glanders
D. A. B. Dance

7.11.16 Plague
T. Butler

7.11.17 Yersinia, Pasteurella, and Francisella
David Lalloo

7.11.18 Anthrax
Thira Sirisanthana

7.11.19 Brucellosis
M. Monir Madkour

7.11.20 Tetanus
F. E. Udwadia

7.11.21 Botulism, gas gangrene, and clostridial gastrointestinal infections
H. E. Larson

7.11.22 Tuberculosis
Richard E. Chaisson and Jean Nachega

7.11.23 Disease caused by environmental mycobacteria
J. M. Grange and P. D. O. Davies

7.11.24 Leprosy (Hansen's disease)
Diana N. J. Lockwood

7.11.25 Buruli ulcer: *Mycobacterium ulcerans* infection
Wayne M. Meyers and Francoise Portaels

7.11.26 Actinomycosis
K. P. Schaal

7.11.27 Nocardiosis
R. J. Hay

7.11.28 Rat bite fevers
D. A. Warrell

7.11.29 Lyme borreliosis
John Nowakowski, Robert B. Nadelman, and Gary P. Wormser

7.11.30 Other borrelia infections
D. A. Warrell

7.11.31 Leptospirosis
George Watt

7.11.32 Non-venereal treponematoses: yaws, endemic syphilis (bejel), and pinta
P.L. Perine and D. A. Warrell

7.11.33 Syphilis
D. J. M. Wright and S. E. Jones

7.11.34 Listeriosis
P. J. Wilkinson

7.11.35 Legionellosis and legionnaires' disease
J. B. Kurtz and J.T. Macfarlane

7.11.36 Rickettsial diseases including ehrlichiosis
D. H. Walker

7.11.37 Scrub typhus
George Watt

7.11.38 *Coxiella burnetii* infections (Q fever)
T. J. Marrie

7.11.39 Bartonelloses, excluding *Bartonella bacilliformis* infections
James G. Olson

7.11.39.1 *Bartonella bacilliformis* infection

A. Llanos Cuentas

7.11.40 Chlamydial infections including lymphogranuloma venerum
D. Taylor-Robinson and D. C. W. Mabey

7.11.41 Mycoplasmas
D. Taylor-Robinson

7.11.42 Newly identified and lesser-known bacteria
J. Paul

7.12 Fungal infections (mycoses)

7.12.1 Fungal infections
R.J. Hay

7.12.2 Cryptococcosis
William G. Powderly

7.12.3 Coccidioidomycosis
John R. Graybill

7.12.4 Paracoccidioidomycosis
M. A. S. Yasuda

7.12.5 Pneumocystis carinii
Robert F. Miller and Ann E. Wakefield

7.12.6 Infection due to Penicillium marneffei
Thira Sirisanthana

7.13 Protozoa

7.13.1 Amoebic infections
R. Knight

7.13.2 Malaria
D. J. Bradley and D. A. Warrell

7.13.3 Babesia
P. Brasseur

7.13.4 Toxoplasmosis
J. Couvreur and Ph. Thulliez

7.13.5 Cryptosporidium and cryptosporidiosis
D.P. Casemore and D.A. Warrell

7.13.6 Cyclospora
D. P. Casemore

7.13.7 Sarcocystosis
V. Zaman

7.13.8 Giardiasis, balantidiasis, isosporiasis, and microsporidiosis
Martin F. Heyworth

7.13.9 *Blastocystis hominis* infection
R. Knight

7.13.10 Human African trypanosomiasis
August Stich

7.13.11 Chagas' disease
Michael Miles

7.13.12 Leishmaniasis
A. D. M. Bryceson

7.13.13 Trichomoniasis
J. P. Ackers

7.14 Nematodes (roundworms)

7.14.1 Cutaneous filariasis
G. M. Burnham

7.14.2 Lymphatic filariasis
R. Knight

7.14.3 Guinea-worm disease: dracunculiasis
R. Knight

7.14.4 Strongyloidiasis, hookworm, and other gut strongyloid nematodes
R. Knight

7.14.5 Nematode infections of lesser importance
D. Grove

7.14.6 Other gut nematodes
V. Zaman

7.14.7 Toxocariasis and visceral larva migrans
V. Zaman

John Dent and Richard H. Holloway

14.7 Peptic ulcer diseases
John Calam

14.8 Hormones and the gastrointestinal tract
P. J. Hammond, S. R. Bloom, A. E. Bishop, and J.M. Polak

14.9 Malabsorption

14.9.1 Differential diagnosis and investigation of malabsorption
Julian R. F. Walters

14.9.2 Small bowel bacterial overgrowth
P. P. Toskes

14.9.3 Coeliac disease
D. P. Jewell

14.9.4 Gastrointestinal lymphoma
P. G. Isaacson

14.9.5 Disaccharidase deficiency
T. M. Cox

14.9.6 Whipple's disease
H. J. F. Hodgson

14.9.7 Effects of massive small bowel resection
R. J. Playford

14.9.8 Malabsorption syndromes in the tropics
V. I. Mathan

14.10 Crohn's disease
D. P. Jewell

14.11 Ulcerative colitis
D. P. Jewell

14.12 Functional bowel disorders and irritable bowel syndrome
D. G. Thompson

14.13 Colonic diverticular disease
N. J. McC. Mortensen and M. G. W. Kettlewell

14.14 Congenital abnormalities of the gastrointestinal tract
V. M. Wright and J. A. Walker-Smith

14.15 Tumours of the gastrointestinal tract
A. F. Markham, I. C. Talbot, and C. B. Williams

14.16 Vascular and collagen disorders
Graham Neale

14.17 Gastrointestinal infections
Davidson H. Hamer and Sherwood L. Gorbach

14.18 Liver, pancreas, and biliary tree

14.18.1 Structure and function of the liver, biliary tract, and pancreas
A. E. S. Gimson

14.18.2 Computed tomography and magnetic resonance imaging of the liver and pancreas
C. S. Ng, D.J. Lomas, and A.K. Dixon

14.18.3 Diseases of the pancreas

14.18.3.1 Acute pancreatitis
C. W. Imrie

14.18.3.2 Chronic pancreatitis
P. P. Toskes

14.18.3.3 Tumours of the pancreas
Julian Britton

14.19 Disease of the gallbladder and biliary tree

14.19.1 Congenital disorders of the liver, biliary tract, and pancreas
J. A. Summerfield

14.19.2 Diseases of the gallbladder and biliary tree
J. A. Summerfield

14.19.3 Jaundice
R. P. H. Thompson

14.20 Hepatitis and autoimmune liver disease

14.20.1 Viral hepatitis - clinical aspects
H. J. F. Hodgson

14.20.2 Autoimmune liver disease

William MacNee

17.7 Chronic respiratory failure
P. M. A. Calverley

17.8 The upper respiratory tract

17.8.1 Upper airways obstruction
J. R. Stradling

17.8.2 Sleep-related disorders of breathing
J. R. Stradling

17.9 Bronchiectasis
D. Bilton

17.10 Cystic fibrosis
Duncan Geddes and Andy Bush

17.11 Diffuse parenchymal lung disease

17.11.1 Diffuse parenchymal lung disease: an introduction
R. M. du Bois

17.11.2 Cryptogenic fibrosing alveolitis
R. M. du Bois

17.11.3 Bronchiolitis obliterans and organizing pneumonia
R. M. du Bois

17.11.4 The lungs and rheumatological diseases
R. M. du Bois and A. K. Wells

17.11.5 The lung in vasculitis
R. M. du Bois

17.11.6 Sarcoidosis
Robert P. Baughman and Elyse E. Lower

17.11.7 Pneumoconioses
A. Seaton

17.11.8 Pulmonary haemorrhagic disorders
D. J. Hendrick G. P. Spickett

17.11.9 Eosinophilic pneumonia
D. J. Hendrick and G. P. Spickett

17.11.10 Lymphocytic infiltrations of the lung
D. J. Hendrick

17.11.11 Extrinsic allergic alveolitis
D. J. Hendrick and G. P. Spickett

17.11.12 Eosinophilic granuloma of the lung and pulmonary lymphangiomyomatosis
D. J. Hendrick

17.11.13 Pulmonary alveolar proteinosis
D. J. Hendrick

17.11.14 Pulmonary amyloidosis
D. J. Hendrick

17.11.15 Lipoid (lipid) pneumonia
D. J. Hendrick

17.11.16 Pulmonary alveolar microlithiasis
D. J. Hendrick

17.11.17 Toxic gases and fumes
D. J. Hendrick

17.11.18 Radiation pneumonitis
D. J. Hendrick

17.11.19 Drug-induced lung disease
D. J. Hendrick and G. P. Spickett

17.12 Pleural disease
M. K. Benson

17.13 Disorders of the thoracic cage and diaphragm
J. M. Shneerson

17.14 Neoplastic disorders

17.14.1 Lung cancer
S. G. Spiro

17.14.2 Pulmonary metastases
S. G. Spiro

17.14.3 Pleural tumours
M. K. Benson

Volume 3

18 Rheumatology

19 Diseases of the skeleton

20 Nephrology

Foreword

by Professor Sir David Weatherall, FRS

It is now 20 years since the first edition of the *Oxford Textbook of Medicine* appeared on the scene, a time when the concept of the all-encompassing textbook of medicine was being questioned. Its predecessor, *Price's Textbook of the Practice of Medicine,* first published in 1922 and by then in its twelfth edition, had come under considerable criticism. One of its most voluble critics, the late J.R.A. Mitchell, had even gone to the trouble of weighing the book, after which he suggested that, because dinosaurs became extinct because of their sheer bulk, medical textbooks would suffer the same fate. In addition, he and many other reviewers suggested that large textbooks are out of date before they are published and hence are of extremely limited value. Notwithstanding Professor Mitchell's outdated views on the extinction of dinosaurs, we thought that he had a point.

After considering these arguments carefully we came to the conclusion that there was still a place for at least one major British work of reference which attempted to cover the whole field of internal medicine. This decision was based largely on the view that, because of the enormous breadth of the subject and the increasing tendency to overspecialization, very few students and practitioners could have immediate access to smaller monographs on every branch of the field; even when they are available they are not always written by those who evaluate their patients in a general medical setting. And if this is true of clinicians in the richer countries, it must apply even more to those in the developing world, where access to libraries and review articles may be limited. Furthermore, although we were well aware that textbooks rapidly become out of date, few advances in medicine lead to major changes in patient care, and those that do often require many years of critical evaluation before they become an integral part of routine clinical practice. For this reason we decided to try to produce a wide-ranging medical textbook which would have a particular emphasis on the global aspects of disease, rather than focus simply on the day-to-day medical problems of the developed world.

Since the *Oxford Textbook of Medicine* first appeared there have been profound changes, both in the practice of medicine and in the problems of the provision of medical care. None of the richer countries has been able to solve the problem of the spiralling costs of health care, which have resulted in part from the introduction of new technology but, even more importantly, from the remarkable increase in the age of their patient populations. If anything, the gap between the quality of the provision of health care between the richer and poorer countries has widened, and although some of the poorer countries have made the epidemiological transition from high death rates due to infection and malnutrition towards a more westernized pattern of illness, particularly in sub-Saharan Africa infectious disease, notably respiratory infection, AIDS, tuberculosis, and malaria, remain the major causes of death; a review of over 11 million childhood deaths in 1998 disclosed, disgracefully, that over 4 million were due to diseases for which adequate vaccines or other forms of prevention already exist. The phenomena of 'globalization', and increasing corporate dominance, are also tending to exacerbate the divide between the rich and poor nations.

Another profound change which has occurred over the last 20 years is the emphasis on the study of disease at the molecular and cellular levels and the increasing role of what is still rather optimistically called 'molecular medicine'. But while this remarkable field promises much for the health of mankind for the future, so far it has had little place in day-to-day clinical practice. Thus, while the fruits of the human genome project offer enormous potential for the better understanding, prevention, and management of the common killers of middle life and old age in richer societies, and the pathogen genome projects offer equal hope for controlling the infectious killers of the developing countries, it is still far from clear when the rich promises of these fields will come to fruition for preventative medicine and clinical care. And there is the danger that when they do, because many of them are likely to be expensive, the gap between the provision of health care in the poorer and richer countries will become even wider. Although many of the solutions to these problems depend on a complete change of attitude of governments and industry in the richer countries, there is no doubt that there will be a rapidly increasing role for their medical schools and doctors to develop collaborative programmes with those of the developing countries and, in general, to take a much more global view of disease, both in medical education and research.

The other major change in the medical field over the last 20 years has been the increasing disquiet about the pattern of medical practice. In many countries doctors have come under increasing criticism for their lack of ability to communicate adequately with patients, for their quality of patient care and, overall, for their lack of humanity. The patient community has become much more sophisticated and demanding, and in most countries there has been a rapid increase in the number of medico-legal actions taken against doctors. This trend has already had wide-ranging repercussions. There has been a major rethink about the pattern of medical education, placing less emphasis on its scientific basis and more on communication skills, ethics, and the social aspects of medicine. The remarkable revolution in the basic biological sciences that underlie medical practice, particularly in the field of genomics, is also raising new ethical issues which would have been undreamed of at the time of the first edition of this book.

In short, medical practice has entered the new millennium in a state of considerable uncertainty. The whole ethos of clinical practice is being questioned, none of the richer countries has got to grips with how to finance the increasing demands of medical care, and many of the poorer countries still have completely dysfunctional health care systems. It is very pleasing therefore to see that the new edition of the *Oxford Textbook of Medicine* reflects so many of these changing issues, as they affect internal medicine. In particular, the textbook has maintained and expanded the aspirations of its original editors towards providing a genuinely global picture of disease, not just as it affects the populations of the richer countries but as it involves the lives of all of those in the poorer countries of the world. As well as continuing to describe the major causes of ill-health and death in the populations of the poorer countries, it includes new sections on screening and the costs of health care, and has greatly increased its coverage of some of the major infectious killers, particularly HIV/AIDS. At the other end of the spectrum it has expanded its sections on the molecular mechanisms of disease and tried to put molecular medicine into perspective by defining its limits. And it has not ignored the remarkable advances in medicine which relate to the richer countries, particularly in its coverage of the problems of the aged. In doing so it has focused on the major killers of Western society, notably cancer, heart disease, and stroke, and has greatly increased the coverage of critical care and emergency medicine. This extensive revision has required the recruitment of many new authors, reflecting a change of over one-third of those from the last edition.

After the publication of the last edition of the *Oxford Textbook of Medicine* my colleague John Ledingham and I decided that it was time to stand aside and pass on our editorial roles to a younger team of editors who are still very active in the fields of medical research and practice. We are delighted to see that our younger colleagues have maintained the tradition of producing a broad-ranging medical textbook which emphasizes the pastoral, scientific, and global aspects of medical care. Despite all its problems medical practice is entering the most exciting and challenging period of its development, and we believe that it still offers the most exciting and enriching of careers for its practitioners. We trust that the 'OTM' will remain their guide and friend for many years to come.

Preface

Textbooks of medicine: *raison d'etre*

Now, in the third millennium, is there any need for a textbook of medicine? Never before has so much information on medical matters been so readily available to so many: physicians are inundated, as are their patients and everyone else. The media seem to carry more and more medical stories in more and more detail every day. The genome has been sequenced. Articulate teenagers speak of stem cells. The internet brings widespread and virtually unlimited access to biomedical information (and misinformation) of a sort: one click of a mouse, and it's all anyone's. A plethora of organizations besieges physicians with guidelines and protocols on every aspect of the practice of medicine. Traditional values are being challenged in all facets of life, including medicine, and there is an unprecedented and entirely appropriate demand for supportive evidence, not just weight of experience, to justify medical interventions.

In these circumstances, some might argue that textbooks of medicine were irrelevant, inappropriate, or even redundant. We strongly refute this. Amidst the maelstrom of 'information' in which physicians now work there is, more than ever, a need for a fixed point of reference, something by which the new, the exciting, and the fashionable can be judged. We make the bold claim that the *Oxford Textbook of Medicine* is just such a fixed point. We argue, unashamedly, that a clinical textbook in the Oslerian tradition is not only required but is essential, to provide expert review, evaluation, and recommendation.

Clinical medicine: changes, challenges, and reconsiderations

This fourth edition of the *Oxford Textbook of Medicine* emerges at a time when discoveries in molecular sciences and advances in technology provide an unprecedented range of diagnostic reagents, drugs, and bioinformatics. Yet, at the same time, there is a widespread recognition that the outcome of treatment for many patients falls short of ideal standards. Microbial resistance to antibiotics, adverse consequences of drugs, and the fallibility of doctors all contribute to failures; and we now realize how dangerous hospitals and clinics can be. Besides this, many contemporary high-tech procedures cannot cure chronic illnesses, and we lack effective weapons to influence the powerful social and behavioural factors that underlie so much illness. The advent of predictive DNA testing also poses complex ethical questions for practitioners, for which few answers are available.

Advances in biomedical science crucially drive innovation and improvement in medical practice. These are not neglected in this book, but the practice of medicine (except in dire emergency) is initiated by a patient talking with a physician and proceeds (as appropriate) through physical examination and investigation to discussion of diagnosis, prognosis, and treatment. These are the core issues of clinical medicine which form the bulk of this textbook.

A culture of public mistrust: the physician-patient relationship

Our political masters in much of the developed world, long tired of being marginalized by old-established networks within the professions, have introduced a new accountability distilled from the concept of audit. This has been exported from the world of finance to embrace the scrutiny of non-financial processes in health care and has created a political climate obsessed with cost effectiveness. The degree of central control often leads to impossible conflicts in the expectations of the public and those entrusted with provision of health care. Baroness O'Neil in her BBC Reith Lectures of 2002 * has pointed out that there is often an inconsistency in the demands raised by such control, providing, as it does, perverse incentives for the specious goals and 'output measures' determined by central bodies. While it is true that much better standards of health care delivery are required and careful surveillance of clinical activities is desirable, the *Oxford Textbook of Medicine* presents an affirmation of the physician-patient relationship in the fight against illness, debility, and suffering: for this relationship should remain sacrosanct, based on professional integrity, knowledge, and human feeling.

Aims and emphases: Sir Archibald Garrod's legacy

Garrod first understood the unique interactions between heredity and environment in the genesis of human disease and asked the question: 'Why did this particular person develop this particular illness in this particular environment?' - a question that we are only just beginning to answer in an era of almost naïve enthusiasm for genetics. While the study of the invariant factors in human genetics is almost intoxicating in its simplicity, we now face the formidable challenge of identifying the contribution of the environment, with all its attendant variables, to the generation of the clinical phenotype we define as illness.

This is the background to this edition of the *Oxford Textbook of Medicine*: its remit stretches from disease as it presents to physicians at the bedside, to the attendant disturbances of cellular, tissue, and organ function, all occurring within an individual, inevitably a part of the turmoil of society. To have a complete description of all these aspects of any medical complaint would not be possible, but we recognize that many readers will not have ready access to the latest sources of scientific information. The book is therefore designed to be a proper reference point for both scientific and clinical aspects of medical practice and bears the fingerprints of Osler, Garrod, Doll, and Weatherall, all Regius Professors of Medicine in Oxford.

Limitations and strengths

The bitter practicalities of writing, editing, and producing any book, especially a work of this size, prevent its referring to the last edition of *The Lancet, Quarterly Journal of Medicine, New England Journal of Medicine,* or any other periodical. But this book can and does provide the medical background against which new information should be assessed and understood. Grounded in the principles that have made the first three editions standard reference textbooks, the new edition has, like medicine itself, evolved to bring all contemporary resources to focus on the teaching and interpretation of medicine. Many new approaches and topics are included and we have incorporated the skill, experience, and perspectives of a truly international complement of highly distinguished authors, including the recently honoured Nobel Laureate in Medicine, Dr Sydney Brenner.

This fourth edition includes, for the first time, an editorial adviser based in the United States (EJB) and a greatly increased and broadened representation of North American authors. By adopting this approach, we hope we have been able to integrate and synthesize in this edition the perspectives on shared medical issues as they confront physicians and medical scientists in different countries.

At a time when there is a tendency for physicians in some parts of the world to be more and more proficient about less and less, this book is a means of their grasping what is happening and what is important in all areas of medical practice. When the movement of people, diseases, and doctors around the world is greater than ever, there is a need for a truly global perspective, which this book provides.

Acknowledgements

This edition contains much that is entirely new, but we wish here to acknowledge that it is built on the firm foundations established by the distinguished co-editors of the previous editions, Professor Sir David Weatherall and Professor John Ledingham. No work of this kind can be produced without the engagement of dedicated professionals who believe in publishing and commit themselves way beyond healthy expectations to see the task through. Mrs Alison Langton has provided guidance and discipline throughout the production and we are enormously grateful to her and her staff at Oxford University Press for their confidence, commitment, and friendship. We are particularly indebted to Dr Irene Butcher who has worked indefatigably to help us realize our aims and at every level has contributed to the organization of the final text and its complex illustrative material. Her experience, knowledge, and uncompromising attention to detail must surely be unique; her forbearance with the editors and, on rare occasions, errant contributors, has been nothing short of miraculous. We thank our contributors for their patience in delivering their sections and review of proofs for which they are responsible. Ultimately, however, the book and any errors it might contain remain the responsibility of the editors.

Finally we thank Mary, Sue, Helen, and Peggy, our constant, supportive, and forgiving wives; Professor Sir David Weatherall, Professor Alastair Compston, Dr Graham Neale, Professor Michael de Swiet, and Dr Michael Sharpe our section advisers; Professor David Lomas, Professor Julian Hopkin, Professor Michael Doherty, Professor David Isenberg, and Dr Christopher Winearls who gave advice and comment for which the editors are very grateful; and our personal secretaries, Eunice Berry (a veteran of four editions), Joan Grantham, Janet Cameron, Naoe Suzuki, and Beverly Comegys for their exceptional dedication.

Oxford, Cambridge, and Boston DAW, TMC, JDF, EJB

January 2003

* *A Question of Trust.* Cambridge University Press 2002.

Contributors

P. Aaby Research Professor (Novo Nordisk Foundation), Bandim Health Project, Bissau, Guinea-Bissau.
7.10.6 Measles

J. P. Ackers Professor of Postgraduate Education in Public Health, London School of Hygiene and Tropical Medicine, UK.
7.13.13 Trichomoniasis

M. W. Adler Professor of Genitourinary Medicine, Department of Sexually Transmitted Diseases, Royal Free and University College Medical School, London, UK.
21.1 Epidemiology

D. Adu Consultant Nephrologist, Queen Elizabeth Hospital, Birmingham, UK.
20.7.4 Minimal-change nephropathy, focal segmental glomerulosclerosis, and membranous nephropathy . 20.10.4 The kidney in rheumatological disorders

Graeme J. M. Alexander University Lecturer in Medicine and Honorary Consultant Physician/Hepatologist, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Cambridge, UK.
14.21.4 Liver transplantation

M. Allison Consultant Hepatologist, Hepatobiliary and Transplant Unit, Addenbrooke's Hospital, Cambridge, UK.
5.5 Innate immune system. 14.21.4 Liver transplantation

Chris Andrews Registrar in Anaesthesia, Mater Misericordiae Hospitals, South Brisbane, Queensland, Australia.
8.5.7 Lightning and electrical injuries

Philip Anslow Consultant Neuroradiologist, Radcliffe Infirmary, Oxford, UK.
24.5 Neuroimaging in neurological diseases

Mark J. Arends Senior Lecturer and Honorary Consultant, Pathology Department, University of Cambridge, UK.
4.6 Apoptosis in health and disease

James O. Armitage Dean, College of Medicine, University of Nebraska Medical Center, Omaha, Nebraska, USA.
22.4.3 Lymphoma

J. K. Aronson Reader in Clinical Pharmacology, Radcliffe Infirmary, Oxford, UK.
15.5.1 Pharmacological management of heart failure

Frances M. Ashcroft Royal Society GlaxoSmithKline Research Professor, University Laboratory of Physiology, Oxford, UK.
4.5 Ion channels and disease

T. C. Aw Professor and Head of Division of Occupational Health, Kent Institute of Medicine and Health Sciences, University of Kent at Canterbury, UK.
8.4.1 Occupational and environmental health and safety . 8.5.10 Noise. 8.5.11 Vibration

M. Bagshaw Head of Occupational and Aviation Medicine, British Airways, Harmondsworth, UK.
8.5.5 Aerospace medicine

E. L. Baker Decatur, Georgia, UK.
8.4.1 Occupational and environmental health and safety

L. R. I. Baker Consultant Physician and Nephrologist, London Clinic, London, UK.
20.14 Urinary tract obstruction

C. R. M. Bangham Professor of Immunology, Imperial College Faculty of Medicine, London, UK.
7.10.23 HTLV-I and II and associated diseases

A. P. Banning Consultant Cardiologist, John Radcliffe Hospital, Oxford, UK.
15.3.3 Echocardiography. 15.14.1 Thoracic aortic dissection

D. J. P. Barker Director, MRC Environmental Epidemiology Unit, University of Southampton, UK.
15.4.1.1 Influences acting *in utero* and early childhood

Roger Barker University Lecturer and Honorary Consultant in Neurology, Department of Neurology, Addenbrooke's Hospital, Cambridge, UK.
24.13.11 Disorders of movement (excluding Parkinson's disease)

D. Barlow Consultant Physician, Department of Genitourinary Medicine, St Thomas's Hospital, London, UK.
7.11.6 Neisseria gonorrhoeae

M. P. Barnes Professor of Neurological Rehabilitation, Hunters Moor Regional Rehabilitation Centre, Newcastle upon Tyne, UK.
24.13.17 Spinal cord injury and its management

John G. Bartlett Chief, Division of Infectious Diseases, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
17.5.2.1 Pneumonia - normal host. 17.5.2.2 Nosocomial pneumonia

Christopher Bass Consultant in Liaison Psychiatry, Department of Psychological Medicine, John Radcliffe Hospital, Oxford, UK.
26.5.3 Medically unexplained symptoms in patients attending medical clinics

M. F. Bassendine Professor of Hepatology, Centre for Liver Research, The Medical School, University of Newcastle upon Tyne, UK.
14.20.2.2 Primary biliary cirrhosis

David Bates Professor of Clinical Neurology, Department of Neurology, University of Newcastle upon Tyne, UK.
24.9 Brainstem syndromes. 24.13.1 The unconscious patient

Robert P. Baughman University of Cincinnati Medical Centre, Ohio, USA.
17.11.6 Sarcoidosis

Peter J. Baxter Consultant Physician, Occupational and Environmental Medicine, University of Cambridge, UK.
8.5.12 Disasters: earthquakes, volcanic eruptions, hurricanes, and floods

Peter H. Baylis Provost and Dean of Faculty of Medical Sciences, University of Newcastle upon Tyne, UK.

20.2.1 Water and sodium homeostasis and their disorders

D. Gareth Beevers Professor of Medicine, City Hospital, Birmingham, UK.
15.16.3 Hypertensive emergencies and urgencies

Michael L. Bennish Director, Africa Centre for Health and Population Studies, Mtubatuba, South Africa.
7.11.11 Cholera

M. K. Benson Consultant Physician, Oxford Centre for Respiratory Medicine, Churchill Hospital, Oxford, UK.
17.12 Pleural disease. 17.14.3 Pleural tumours. 17.14.4 Mediastinal tumours and cysts

V. Beral Head, Cancer Research UK Epidemiology Unit, Radcliffe Infirmary, Oxford, UK.
21.6 Cervical cancer and other cancers caused by sexually transmitted infections

Anthony R. Berendt Consultant Physician-in-Charge, Bone Infection Unit, Nuffield Orthopaedic Centre, Oxford, UK.
18.7.1 Pyogenic arthritis. 19.3 Osteomyelitis

Nancy Berliner Professor of Medicine and Genetics, Yale School of Medicine, New Haven, Connecticut, USA.
22.4.1 Leucocytes in health and disease. 22.4.2 Introduction to the lymphoproliferative disorders

Michael Besser Professor of Medicine Emeritus, Bart's and The London School of Medicine and Dentistry, Queen Mary College, London, UK.
12.2 Disorders of the anterior pituitary. 12.3 Disorders of the posterior pituitary

Delia B. Bethell Specialist Registrar in Paediatrics, Department of Paediatrics, John Radcliffe Hospital, Oxford, UK.
7.11.1 Diphtheria

Ernest Beutler Chairman, Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, California, USA.
22.5.11 Erythrocyte enzymopathies

P. C. L. Beverley Professor and Scientific Head, Edward Jenner Institute for Vaccine Research, Compton, Berkshire, UK.
6.5 Tumour immunology

R. W. Bilous Professor of Clinical Medicine, James Cook University Hospital, Middlesbrough, Cleveland, UK.
20.10.1 Diabetes mellitus and the kidney

D. Bilton Consultant in Respiratory Medicine, Papworth Hospital, Cambridge, UK.
17.9 Bronchiectasis

A. E. Bishop Senior Lecturer, Tissue Engineering and Regenerative Medicine Centre, Chelsea and Westminster Hospital, London, UK.
14.8 Hormones and the gastrointestinal tract

Carol M. Black President of the Royal College of Physicians of London and Professor of Rheumatology, Royal Free and University College Medical School, Royal Free Campus, London, UK.
18.10.3 Systemic sclerosis

S. R. Bloom Professor of Medicine and Head, Division of Investigative Science, Imperial College Faculty of Medicine, Hammersmith Campus, London, UK.
12.10 Non-diabetic pancreatic endocrine disorders and multiple endocrine neoplasia. 14.8 Hormones and the gastrointestinal tract

L. D. Blumhardt Emeritus Professor of Clinical Neurology, University of Nottingham, UK.
24.13.5 Syncope. 24.13.16 Diseases of the spinal cord

N. Boon Consultant Cardiologist, Royal Infirmary of Edinburgh, UK.
15.10.4 Cardiac disease in HIV infection

D. R. Booth Senior Hospital Scientist, Institute for Immunology and Allergy Research, Westmead Millennium Institute, Sydney, New South Wales, Australia.
11.12.3 Familial Mediterranean fever and other inherited periodic fever syndromes

Richard T. Booth Professor, Health and Safety Unit, Aston University, Birmingham, UK.
8.4.2 Occupational safety

Leszek K. Borysiewicz Professor and Principal of the Faculty of Medicine, University of Wales, Cardiff, UK.
7.4 The host response to infection

I. C. J. W. Bowler Consultant Microbiologist, Department of Microbiology, John Radcliffe Hospital, Oxford, UK.
7.9 Nosocomial infections

D. J. Bradley Ross Professor of Tropical Hygiene, London School of Hygiene and Tropical Medicine, UK.
7.13.2 Malaria

Thomas Brandt Klinikum Groshadern, Munich, Germany.
24.12.1 Eye movements and balance

P. Brandtzaeg Professor of Paediatrics, Ullevål University Hospital, University of Oslo, Norway.
7.11.5 Meningococcal infections

P. Brasseur Professor and Head of Department of Parasitology, Faculty of Medicine, Rouen, France.
7.13.3 Babesia

J. Braun Professor and Medical Director, Rheumazentrum Ruhrgebiet, Herne, Germany.
18.6 Spondyloarthritides and related arthritides

Sidney Brenner Research Professor, Salk Institute, La Jolla, California, USA, and Honorary Professor of Genetic Medicine, University of Cambridge, UK.
4.2 The human genome sequence

D. P. Brenton Sub Dean (Curriculum), Royal Free and University College Medical School, London, UK.
11.2 Inborn errors of amino acid and organic acid metabolism

Paul H. Brion Rheumatologist in Private Practice, Vista, California, USA.

18.8 Osteoarthritis

Julian Britton Consultant Surgeon, John Radcliffe Hospital, Oxford, UK.
14.3.1 The acute abdomen. 14.18.3.3 Tumours of the pancreas

Anthony F. T. Brown Associate Professor and Senior Staff Specialist, Department of Emergency Medicine, Royal Brisbane Hospital, Queensland, Australia.
16.4 Anaphylaxis

M. J. Brown Professor of Clinical Pharmacology, University of Cambridge and Honorary Consultant Physician, Addenbrooke's Hospital NHS Trust, Cambridge, UK.
15.16.2.3 Primary hyperaldosteronism (Conn's syndrome). 15.16.2.4 Phaeochromocytoma

A. D. M. Bryceson Emeritus Professor of Tropical Medicine, London School of Hygiene and Tropical Medicine, UK.
7.13.12 Leishmaniasis

Philip J. Burke Johns Hopkins Oncology Center, Baltimore, Maryland, USA.
22.3.3 Acute lymphoblastic leukaemia. 22.3.4 Acute myeloblastic leukaemia

G. M. Burnham Associate Professor of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.
7.14.1 Cutaneous filariasis

Jacky Burrin Professor of Experimental Endocrinology, Bart's and The London School of Medicine and Dentistry, St Bartholomew's Hospital, London, UK.
12.1 Principles of hormone action

Andy Bush Reader in Paediatric Respirology, London, UK.
17.10 Cystic fibrosis

K. Bushby Professor of Neuromuscular Genetics, Institute of Human Genetics, Newcastle upon Tyne, UK.
24.22.2 Muscular dystrophy

Anthony Busuttil Regius Professor of Forensic Medicine, Forensic Medicine Section, Edinburgh University Medical School, UK.
27 Forensic medicine and the practising doctor

T. Butler Professor of Internal Medicine and Chief of Infectious Diseases, Texas Technical University Health Sciences Center, Lubbock, Texas, USA.
7.11.16 Plague

W. F. Bynum Professor of History of Medicine, Wellcome Trust Centre for the History of Medicine at University College London, UK.
2.1 Science in medicine: when, how, and what

I. Byren Consultant in Infectious Diseases and Genito-Urinary Medicine, John Radcliffe Hospital, Oxford, UK.
15.10.3 Cardiovascular syphilis

John Calam* Professor of Medicine, Imperial College London, UK.
14.7 Peptic ulcer diseases

Donald B. Calne Professor Emeritus, University of British Columbia, Vancouver, Canada.
24.13.10 Parkinsonism and other extrapyramidal disorders

P. M. A. Calverley Professor of Medicine (Pulmonary and Rehabilitation), Clinical Science Centre, University Hospital Aintree, Liverpool, UK.
17.7 Chronic respiratory failure

Giovambattista Capasso Professor of Nephrology, Second University of Naples, Italy.
20.13 Urinary stones, nephrocalcinosis, and renal tubular acidosis

Jonathan R. Carapetis Senior Lecturer, Research Fellow, and Consultant in Infectious Diseases, Centre for International Child Health, University of Melbourne Department of Paediatrics, Royal Children's Hospital, Melbourne, Australia.
15.10.1 Acute rheumatic fever

Simon Carette Head, Division of Rheumatology, Toronto Western Hospital, Ontario, Canada.
18.4 Back pain and regional disorders

D. J. S. Carmichael Consultant Renal Physician, Southend Hospital, Westcliffeon-Sea, Essex, UK.
20.16 Drugs and the kidney

D. P. Casemore Senior Research Fellow, CREH, University of Wales, St Asaph, Denbighshire, UK.
7.13.5 Cryptosporidium and cryptosporidiosis. 7.13.6 Cyclospora

D. Catovsky Professor of Haematology, Royal Marsden Hospital and Institute of Cancer Research, London, UK.
22.3.2 The classification of leukaemia. 22.3.5 Chronic lymphocytic leukaemia and other leukaemias of mature B and T cells

Bruce A. Chabner Professor of Medicine, Harvard Medical School and Massachusetts General Hospital, Boston, USA.
6.7 Cancer chemotherapy and radiation therapy

Richard E. Chaisson Professor of Medicine, Epidemiology and International Health, Johns Hopkins University Schools of Medicine and Public Health, Baltimore, Maryland, USA.
7.11.22 Tuberculosis

R. W. Chapman Consultant Gastroenterologist/Hepatologist, John Radcliffe Hospital, Oxford, UK.
14.20.2.3 Primary sclerosing cholangitis

V. Krishna K. Chatterjee Professor of Endocrinology, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK.
12.1 Principles of hormone action

Dominique Chauveau Consultant Nephrologist, Department of Nephrology, Hôpital Necker, Paris, France.
20.9.1 Acute interstitial nephritis

P. F. Chinnery Senior Lecturer in Neurogenetics and Honorary Consultant Neurologist, University of Newcastle upon Tyne and Newcastle upon Tyne Hospitals NHS Trust, UK.
24.22.5 Mitochondrial encephalomyopathies

Seung-Yull Cho Professor, Section of Molecular Parasitology, Sungkyunkwan University School of Medicine, Suwon, Korea.
7.15.4 Pseudophyllidean tapeworms: diphyllobothriasis and sparganosis

Kirpal S. Chugh Professor Emeritus, Department of Nephrology, Postgraduate Institute of Medical Education and Research, Chandigarh, India.
20.7.10 Glomerular disease in the tropics

L. Chwastiak Acting Assistant Professor, Department of Psychiatry, University of Washington, Seattle, USA.
26.5.4 Anxiety and depression

C. M. Clothier Queen's Counsel (retired), London, UK.
1 On being a patient

Andrew J. S. Coats Viscount Royston Professor of Cardiology, Imperial College London and Honorary Consultant Cardiologist, Royal Brompton Hospital, London, UK.
15.2.2 The syndrome of heart failure. 15.5.3 Cardiac rehabilitation

S. M. Cobbe Walton Professor of Medical Cardiology, University of Glasgow, Glasgow Royal Infirmary, UK.
15.2.3 Syncope and palpitation. 15.6 Cardiac arrhythmias

B. J. Cohen Clinical Scientist, Central Public Health Laboratory, London, UK.
7.10.18 Parvovirus B19

J. Cohen Dean and Professor of Infectious Diseases, Brighton and Sussex Medical School, UK.
7.20 Infection in the immunocompromised host

R. D. Cohen Emeritus Professor of Medicine, Bart's and The London School of Medicine and Dentistry, Queen Mary College, University of London, UK.
11.11 Disturbances of acid-base homeostasis

Francis S. Collins Director, National Human Genome Research Institute, Bethesda, Maryland, USA.
4.1 The genomic basis of medicine

R. Collins British Heart Foundation Professor of Medicine and Epidemiology, Clinical Trial Service Unit and Epidemiological Studies Unit, University of Oxford, UK.
2.4.3 Large-scale randomized evidence: trials and overviews

Alastair Compston Professor of Neurology, University of Cambridge, UK.
24.1 Introduction and approach to the patient with neurological disease . 24.16 Demyelinating disorders of the central nervous system

Juliet Compston Reader in Metabolic Bone Diseases and Honorary Consultant Physician, Addenbrooke's Hospital, Cambridge, UK.
19.4 Osteoporosis

C. P. Conlon Consultant Physician in Infectious Diseases, Nuffield Department of Medicine, John Radcliffe Hospital, Oxford, UK.
7.8 Travel and expedition medicine. 7.10.21 HIV and AIDS

Andrew Coop Duke University Medical Center, Durham, North Carolina, USA.
6.2 The nature and development of cancer. 6.3 The genetics of inherited cancers

M. R. Cooper Freelance Science Writer, CAB International, Wallingford, Oxfordshire, UK.
8.3 Poisonous plants and fungi

Susan Copley Consultant Radiologist, Hammersmith Hospital, London, UK.
17.3.1 Thoracic imaging

Fernando F. Costa Professor of Haematology, School of Medical Sciences, Unicamp, Campinas, Brazil.
22.5.10 Disorders of the red cell membrane

J. Couvreur Professeur Associé, Laboratoire de la Toxoplasmose, Institut de Puericulture, Paris, France.
7.13.4 Toxoplasmosis

P. J. Cowen Professor of Psychopharmacology, Warneford Hospital, Oxford, UK.
26.6.1 Psychopharmacology in medical practice

T. M. Cox Professor of Medicine, University of Cambridge, and Honorary Consultant Physician, Addenbrooke's Hospital, Cambridge, UK.
11.3.1 Glycogen storage diseases. 11.3.2 Inborn errors of fructose metabolism. 11.3.3 Disorders of galactose, pentose, and pyruvate metabolism. 11.5 The porphyrias. 11.7.1 Hereditary Haemochromatosis. 11.8 Lysosomal storage diseases. 12.13 The pineal gland and melatonin. 14.9.5 Disaccharidase deficiency. 22.5.4 Iron metabolism and its disorders. 33 Emergency Medicine

Dorothy H. Crawford Professor of Medical Microbiology, Centre for Infectious Diseases, University of Edinburgh, UK.
7.10.3 The Epstein-Barr virus

Robin A. F. Crawford Consultant Gynaecological Oncologist, Addenbrooke's Hospital, Cambridge, UK.
13.17 Malignant disease in pregnancy

A. J. Crisp Consultant Rheumatologist, Addenbrooke's Hospital, Cambridge, UK.
19.5 Avascular necrosis and related topics

D. W. M. Crook Consultant Microbiologist/Infectious Diseases, John Radcliffe Hospital, Oxford, UK.
24.14.1 Bacterial meningitis

J. Cunningham Professor of Renal and Metabolic Medicine, The Royal London Hospital and Queen Mary's School of Medicine and Dentistry, London, UK.
20.8 Renal tubular disorders

Patrick C. D'Haese Associate Professor, Department of Nephrology and Hypertension, University of Antwerp, Belgium.
20.9.2 Chronic tubulointerstitial nephritis

Tim Dalgleish Research Scientist, MRC Cognitions and Brain Sciences Unit, Cambridge, UK.
26.5.1 Grief, stress, and post-traumatic stress disorder

D. A. B. Dance Director/Consultant Microbiologist, Public Health Laboratory, Derriford Hospital, Plymouth, UK.

7.11.15 Melioidosis and glanders

Chi V. Dang Professor of Medicine and Chief, Hematology Division, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
22.3.7 Myelodysplasia

C. J. Danpure Professor of Molecular Cell Biology, Department of Biology, University College London, UK.
11.10 Disorders of oxalate metabolism

John H. Dark Professor of Cardiothoracic Surgery, Freeman Hospital, Newcastle upon Tyne, UK.
15.5.4 Cardiac transplantation and mechanical circulatory support

A. Davenport Consultant Renal Physician/Honorary Senior Lecturer, Centre for Nephrology, Royal Free Hospital, London, UK.
20.3.2 Clinical investigation of renal disease

G. Davey Smith Professor of Clinical Epidemiology, University of Bristol, UK.
15.4.1.2 The epidemiology of ischaemic heart disease

Alun Davies Reader and Honorary Consultant Surgeon, Department of Vascular Surgery, Faculty of Medicine, Imperial College School of Medicine, Charing Cross Hospital, London, UK.
15.14.2 Peripheral arterial disease

P. D. O. Davies Consultant Physician, Fazakerley Hospital, Liverpool, UK.
7.11.23 Disease caused by environmental mycobacteria

Alex M. Davison Professor and Consultant Renal Physician, St James's University Hospital, Leeds, UK.
20.3.1 The clinical presentation of renal disease

Marc E. De Broe Professor in Medicine, Department of Nephrology, University of Antwerp, Belgium.
20.9.2 Chronic tubulointerstitial nephritis

P. de la Motte Hall Professor, Division of Anatomical Pathology, Faculty of Health Sciences, University of Cape Town, South Africa.
14.21.6 Hepatic granulomas

M. de Swiet Professor of Obstetric Medicine, Queeen Charlotte's and Chelsea Hospital, London, UK.
13.7 Thromboembolism in pregnancy. 13.8 Chest diseases in pregnancy

Barbara A. Degar Yale School of Medicine, New Haven, Connecticut, USA.
22.4.2 Introduction to the lymphoproliferative disorders

Eric Demoncheaux Research Associate, Medical School, University of Sheffield, UK.
15.15.1 The pulmonary circulation and its influence on gas exchange

D. M. Denison Emeritus Professor of Clinical Physiology, Royal Brompton Hospital, London, UK.
8.5.5 Aerospace medicine. 8.5.6 Diving medicine

John Dent Director, Department of Gastroenterology, Hepatology and General Medicine and Clinical Professor of Medicine, Royal Adelaide Hospital/Adelaide University, Australia.
14.6 Diseases of the oesophagus

Christopher P. Denton Senior Lecturer/Consultant Rheumatologist, Centre for Rheumatology, Royal Free Hospital, London, UK.
18.10.3 Systemic sclerosis

Ulrich Desselberger Consultant Virologist and Director, Clinical Microbiology and Public Health Laboratory, Addenbrooke's Hospital, Cambridge, UK.
7.10.7 Enterovirus infections. 7.10.8 Virus infections causing diarrhoea and vomiting

Charles A. Dinarello Professor of Medicine, University of Colorado, Denver, Colorado, USA.
4.4 Cytokines: interleukin-1 and tumor necrosis factor in inflammation

A. K. Dixon Professor of Radiology and Honorary Consultant Radiologist, University of Cambridge and Addenbrooke's Hospital, Cambridge, UK.
14.18.2 Computed tomography and magnetic resonance imaging of the liver and pancreas

Michael Doherty Professor of Rheumatology, University of Nottingham Medical School, UK.
18.3 Clinical investigation. 18.9 Crystal-related arthropathies

R. Doll Emeritus Professor of Medicine and Honorary Member, Cancer Studies Unit, Nuffield Department of Medicine, Radcliffe Infirmary, Oxford, UK.
6.1 Epidemiology of cancer

Michael Donaghy Reader in Clinical Neurology, University of Oxford, Honorary Consultant Neurologist, Radcliffe Infirmary, and Honorary Civilian Consultant in Neurology to the Army, Oxford, UK.
24.13.13 The motor neurone diseases

Dominique Droz Unite de Pathologie Renale, Hôpital Necker, Paris, France.
20.9.1 Acute interstitial nephritis

R. M. du Bois Professor of Respiratory Medicine, National Heart and Lung Institute, University College London and Consultant Physician, Royal Brompton Hospital, London, UK.
17.11.1 Diffuse parenchymal lung disease: an introduction. 17.11.2 Cryptogenic fibrosing alveolitis. 17.11.3 Bronchiolitis obliterans and organizing pneumonia. 17.11.4 The lungs and rheumatological diseases. 17.11.5 The lung in vasculitis

C. R. K. Dudley Consultant Renal Physician, The Richard Bright Renal Unit Southmead Hospital, North Bristol NHS Trust, Bristol, UK.
15.14.3 Cholesterol emboli

D. W. Dunne Reader in Immunoparasitology, Department of Pathology, University of Cambridge, UK.
7.16.1 Schistosomiasis

David T. Durack Consulting Professor of Medicine, Duke University, Durham, North Carolina and Vice-President, Corporate Medical Affairs, Becton Dickinson & Co., Franklin Lakes, New Jersey, USA.
7.2 Fever of unknown origin

S. R. Durham Professor of Allergy and Respiratory Medicine, Imperial College Faculty of Medicine, National Heart and Lung Hospital, and Royal Brompton Hospital, London, UK.
17.4.2 Allergic rhinitis ('hay fever')

P. N. Durrington Professor of Medicine, University of Manchester Department of Medicine, Manchester Royal Infirmary, UK.
11.6 Lipid and lipoprotein disorders

M. Eastwood Post-Retirement Honorary Fellow, Department of Medical Sciences, Western General Hospital, Edinburgh, UK.
10.3 Vitamins and trace elements

Jonathan C. W. Edwards Professor in Connective Tissue Medicine, University College London, UK.
18.1 Joints and connective tissue: introduction

Richard Edwards Emeritus Professor of Medicine, University of Liverpool, UK.
24.22.4 Metabolic and endocrine disorders

M. Elia Professor of Clinical Nutrition and Metabolism, Institute of Human Nutrition, University of Southampton, UK.
10.6 Special nutritional problems and the use of enteral and parenteral nutrition

Matthew J. Ellis Associate Professor of Medicine and Director, Breast Cancer Program, Duke University Medical Center, Durham, North Carolina, USA.
6.2 The nature and development of cancer. 6.3 The genetics of inherited cancers

Monique M. Elseviers Department of Nephrology-Hypertension, University Hospital Antwerp, Belgium.
20.9.2 Chronic tubulointerstitial nephritis

M. A. Epstein Emeritus Professor of Pathology, University of Bristol, UK.
7.10.3 The Epstein-Barr virus

E. Ernst Professor and Director, Department of Complementary Medicine, University of Exeter, UK.
2.5 Complementary and alternative medicine

David Eschenbach Professor, Department of Obstetrics and Gynecology, University of Washington, Seattle, USA.
21.4 Pelvic inflammatory disease

S. M. Evans Specialist Registrar in Gastroenterology, Royal Sussex County Hospital, Brighton, UK.
8.5.8 Podoconiosis

S. J. Eykyn Professor (and Honorary Consultant) in Clinical Microbiology, St Thomas' Hospital, London, UK.
7.11.2 Streptococci and enterococci. 7.11.4 Staphylococci. 7.11.10 Anaerobic bacteria. 15.10.2 Infective endocarditis

C. A. Eynon Director of Neurosciences Intensive Care, Southampton University Hospital NHS Trust, UK.
16.3 Cardiac arrest. 33 Emergency Medicine

Christopher G. Fairburn Wellcome Principal Research Fellow and Professor of Psychiatry, Oxford University Department of Psychiatry, Warneford Hospital, Oxford, UK.
26.5.5 Eating disorders

J. J. Farrar Senior Fellow, Wellcome Trust, University of Oxford Clinical Research Unit, The Hospital for Tropical Diseases, Ho Chi Minh, Vietnam.
24.14.1 Bacterial meningitis. 24.14.2 Viral infections of the central nervous system

Ken Farrington Consultant Nephrologist, Lister Hospital, Stevenage, Hertfordshire, UK.
20.6.1 Haemodialysis

D. T. Fearon Wellcome Trust Professor of Medicine, University of Cambridge, UK.
5.5 Innate immune system

John Feehally Professor of Renal Medicine, Leicester General Hospital, UK.
20.7.2 IgA nephropathy and Henoch-Schonlein purpura. 20.7.3 Thin membrane nephropathy

Alvan R. Feinstein* Professor, Yale University School of Medicine, New Haven, Connecticut, USA.
2.4.2 Evidence-based medicine

Eleanor Feldman Consultant Liaison Psychiatrist and Honorary Senior Lecturer, University of Oxford, John Radcliffe Hospital, Oxford, UK.
26.2 Taking a psychiatric history from a medical patient. 26.4 Acute behavioural emergencies

Peter J. Fenner Associate Professor, Schools of Medicine and Health Sciences, James Cook University and National Medical Officer, Surf Life Saving Association of Australia, Mackay, North Queensland, Australia.
8.5.3 Drowning

Robert Ferrari Clinical Assistant Professor, University of Alberta Hospital, Edmonton, Canada.
18.2 Clinical presentation and diagnosis of rheumatic disease

C. ffrench-Constant Professor of Neurological Genetics, University of Cambridge, UK.
24.21 Developmental abnormalities of the nervous system

R. G. Finch Professor of Infectious Diseases, City Hospital and University of Nottingham, UK.
7.6 Antimicrobial chemotherapy

H. Firth Consultant in Medical Genetics, Department of Medical Genetics, Addenbrooke's Hospital, Cambridge, UK.
24.21 Developmental abnormalities of the nervous system

J. Firth Consultant Physician and Nephrologist, Addenbrooke's Hospital, Cambridge, UK.
13.5 Renal disease in pregnancy. 15.15.2.2 Pulmonary oedema. 15.15.3.1 Deep venous thrombosis and pulmonary embolism. 15.18 Idiopathic oedema of women. 16.1 The clinical approach to the patient who is very ill . 20.2.2 Disorders of potassium homeostasis. 20.4 Acute renal failure. 33 Emergency Medicine

Susan Fisher-Hoch Professor, University of Texas School of Public Health at Brownsville, USA.
7.10.15 Arenaviruses. 7.10.16 Filoviruses

Robert A. Fishman Professor of Neurology Emeritus, University of California San Francisco School of Medicine, USA.
24.7 Lumbar puncture

Edward D. Folland Associate Director of Cardiology and Professor of Medicine, UMass Memorial Medical Center/University of Massachusetts Medical School, Worcester, Maryland, USA.
15.3.6 Cardiac catheterization and angiography . 15.4.2.4 Percutaneous interventional cardiac procedures

J. C. Forfar Consultant Cardiologist, John Radcliffe Hospital, Oxford and Honorary Senior Lecturer, University of Oxford, UK.
13.6 Heart disease in pregnancy

I. S. Foulds Consultant Dermatologist, City Hospital, Birmingham, UK.
8.4.1 Occupational and environmental health and safety

Keith A. A. Fox Professor of Cardiology, Royal Infirmary and University of Edinburgh, UK.
15.4.2.3 Management of acute coronary syndromes: unstable angina and myocardial infarction

Richard Frackowiak Vice Provost (Biomedicine), University College London, Institute of Neurology, London, UK.
24.3 Brain and mind: functional neuroimaging

T. J. R. Francis Consultant in Diving Medicine, Tintagel, Cornwall, UK.
8.5.6 Diving medicine

Keith N. Frayn Professor of Human Metabolism, Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, UK.
10.2 Nutrition: biochemical background

Alan Freeman Consultant Radiologist, Addenbrooke's Hospital, Cambridge, UK.
14.2.3 Radiology of the gastrointestinal tract

Peggy Frith Consultant Ophthalmic Physician, The Eye Hospital, Radcliffe Infirmary, Oxford and University College London Hospital, UK.
25 The eye in general medicine

Patrick G. Gallagher Associate Professor, Department of Pediatrics, Yale University School of Medicine, New Haven, Connecticut, USA.
22.5.10 Disorders of the red cell membrane

Clare J. Galton Specialist Registrar in Neurology, Neurology Department, Addenbrooke's Hospital, Cambridge, UK.
24.13.8 Alzheimer's disease and other dementias

Hector H. Garcia Associate Professor, Department of Microbiology, Universidad Peruana Cayetano Heredia and Head, Cysticercosis Unit, Department of Transmissible Diseases, Instituto de Ciencias Neurologicas, Lima, Peru.
7.15.1 Cystic hydatid disease (Echinococcus granulosus) . 7.15.3 Cysticercosis

K. Gardiner Professor and Managing Director, International Occupational Health Ltd., Birmingham, UK.
8.4.1 Occupational and environmental health and safety

Lawrence B. Gardner Assistant Professor of Medicine, Johns Hopkins University school of Medicine, Baltimore, Maryland, USA.
22.3.7 Myelodysplasia

Christopher S. Garrard Consultant Physician in Intensive Care, John Radcliffe Hospital, Oxford, UK.
16.5.2 The management of respiratory failure

J. S. H. Gaston Professor of Rheumatology, University of Cambridge School of Medicine, Addenbrooke's Hospital, Cambridge, UK.
18.7.2 Reactive arthritis

Duncan Geddes Professor of Respiratory Medicine, Royal Brompton Hospital, London, UK.
17.10 Cystic fibrosis

D. G. Gibson Consultant Cardiologist, Royal Brompton Hospital, London, UK.
15.7 Valve disease. 15.9 Pericardial disease

G. J. Gibson Professor of Respiratory Medicine/Consultant Physician, Freeman Hospital, Newcastle upon Tyne, UK.
17.3.2 Respiratory function tests

A. M. Giles Scientific Officer, Health Systems, Oxford, UK.
32 Reference intervals for biochemical data

I. P. Giles ARC Research Fellow, Bloomsbury Rheumatology Unit, London, UK.
18.10.1 Autoimmune rheumatic disorders and vasculitis

Charles F. Gilks Professor of Tropical Medicine and Senior Adviser on Care, HIV/AIDS Department, World Health Organization, Geneva, Switzerland.
7.10.22 HIV in the developing world

Michael D. J. Gillmer Consultant Obstetrician and Gynaecologist, Women's Centre, John Radcliffe Hospital, Oxford, UK.
13.10 Diabetes in pregnancy

Robert H. Gilman Professor, Department of International Health, Johns Hopkins School of Public Health, Baltimore, Maryland, USA and Research Professor, Universidad Peruana Cayetano Heredia, Lima, Peru.
7.15.3 Cysticercosis

A. E. S. Gimson Consultant Physician and Hepatologist, Cambridge Liver Transplantation Unit, Addenbrooke's Hospital, Cambridge, UK.
13.9 Liver and gastrointestinal diseases during pregnancy . 14.18.1 Structure and function of the liver, biliary tract, and pancreas

P. Glasziou Huntington Centre for Risk Analysis, Boston, Massachusetts, USA.
2.4.1 Bringing the best evidence to the point of care

Peter J. Goadsby Professor of Clinical Neurology, Institute of Neurology, University College and The National Hospital for Neurology and Neurosurgery, London, UK.
24.13.2 Headache

D. Goldblatt Reader in Immunology and Consultant Paediatric Immunologist, Institute of Child Health, Great Ormond Hospital for Children NHS Trust, London, UK.

7.7 Immunization

John M. Goldman Professor of Leukaemia Biology and Chairman, Department of Haematology, Imperial College School of Medicine, London, UK.
22.3.6 Chronic myeloid leukaemia

Irwin Goldstein Director, Institute for Sexual Medicine and Professor of Urology and Gynecology, Boston University School of Medicine, Massachusetts, USA.
12.8.4 Sexual dysfunction

Armando E. Gonzalez Department of Public Health, School of Veterinary Medicine, Universidad Nacional Mayor de San Marcos, Lima, Peru.
7.15.1 Cystic hydatid disease (Echinococcus granulosus)

Timothy H. J. Goodship Reader in Nephrology, University of Newcastle upon Tyne and Consultant Nephrologist, Royal Victoria Infirmary, Newcastle upon Tyne, UK.
20.10.6 Haemolytic uraemic syndrome

Sherwood L. Gorbach Department of Community Health and Medicine, TUFTS University School of Medicine, Boston, Massachusetts, USA.
14.17 Gastrointestinal infections

E. C. Gordon-Smith Professor of Haematology, St George's Hospital Medical School, London, UK.
22.3.11 Aplastic anaemia and other causes of bone marrow failure . 22.8.2 Haemopoietic stem cell transplantation

J. M. Grange Visiting Professor, University College London, Centre for Infectious Diseases and International Health, Royal Free and University College Medical School, London, UK.
7.11.23 Disease caused by environmental mycobacteria

R. Gray Professor of Medical Statistics and Director, University of Birmingham Clinical Trials Unit, UK.
2.4.3 Large-scale randomized evidence: trials and overviews

John R. Graybill Professor, University of Texas Health Science Center, San Antonio, Texas, USA.
7.12.3 Coccidioidomycosis

Jackie Green Director, Centre for Health Promotion Research, Leeds Metropolitan University, Leeds, UK.
3.5 Health promotion

Brian M. Greenwood Professor of Clinical Tropical Medicine, London School of Hygiene and Tropical Medicine, London, UK.
7.11.3 Pneumococcal diseases

Roger Greenwood Consultant Nephrologist and Lead Clinician, Lister Hospital, Stevenage, Hertfordshire, UK.
20.6.1 Haemodialysis

B. Gribbin Honorary Consultant Cardiologist, John Radcliffe Hospital, Oxford, UK.
15.10.3 Cardiovascular syphilis . 15.14.1 Thoracic aortic dissection

John Grimley Evans Professor Emeritus of Clinical Geratology, Green College, Oxford, UK.
30.1 Medicine in old age

Michael L. Grossbard Chief, Hematology/Oncology, St Luke's-Roosevelt Hospital and Beth Israel Medical Center, New York, USA.
6.7 Cancer chemotherapy and radiation therapy

David I. Grove Professor and Director, Clinical Microbiology and Infectious Diseases, The Queen Elizabeth Hospital, Adelaide, Australia.
7.14.5 Nematode infections of lesser importance . 7.16.2 Liver fluke infections . 7.16.4 Intestinal trematode infections

J. P. Grünfeld Professor of Nephrology, Université Paris V - René Descartes and Head of Nephrology, Hôpital Necker, Paris, France.
20.11 Renal involvement in genetic disease

D. J. Gubler Director, Division of Vector-Borne Infectious Diseases, Centers for Disease Control and Prevention, Fort Collins, Colorado, USA.
7.10.11 Alphaviruses . 7.10.13 Flaviviruses

Mark Gurnell Specialist Registrar and Research Fellow, Department of Medicine, Division of Endocrinology and Metabolism, Addenbrooke's Hospital, Cambridge, UK.
12.1 Principles of hormone action

David M. Gustin Section of Hematology-Oncology, University of Chicago, Illinois, USA.
22.3.8 The polycythaemias . 22.3.10 Thrombocytosis

M. R. Haeney Consultant Immunologist, Salford Royal Hospitals NHS Trust, Salford, Manchester, UK.
14.4 Immune disorders of the gastrointestinal tract

Davidson H. Hamer Director, Traveler's Health Service, Tufts-New England Medical Center and Assistant Professor of Medicine and Nutrition, Tufts University, Boston, Massachusetts, USA.
14.17 Gastrointestinal infections

P. J. Hammond Consultant Physician and Endocrinologist, Harrogate District Hospital, Yorkshire, UK.
12.10 Non-diabetic pancreatic endocrine disorders and multiple endocrine neoplasia . 14.8 Hormones and the gastrointestinal tract

J. R. Hampton Professor of Cardiology, Queen's Medical Centre, Nottingham, UK.
15.2.1 Chest pain . 15.2.4 Physical examination of the cardiovascular system

M. Hanna Consultant Neurologist and Reader in Clinical Neurology, National Hospital for Neurology and Neurosurgery and Institute of Neurology, University College London, UK.
24.22.1 Introduction: structure and function

David M. Hansell Professor of Thoracic Imaging, Royal Brompton Hospital, London, UK.
17.3.1 Thoracic imaging

P. Harnden Consultant Urological Pathologist, Cancer Research UK Clinical Centre, St James's University Hospital, Leeds, UK.
20.15 Tumours of the urinary tract

J. M. Harrington Emeritus Professor of Occupational Health, University of Birmingham, UK.

8.4.1 Occupational and environmental health and safety

Anthony Harrison Fellow in Health Systems, King's Fund, London, UK.
3.3 The pattern of care: hospital and community

J. R. Harrison Force Medical Adviser, Sussex Police Authority, Lewes, UK.
8.5.9 Radiation

C. Haslett Professor of Respiratory Medicine, Royal Infirmary, Edinburgh, UK.
16.5.1 Pathophysiology and pathogenesis of acute respiratory distress syndrome . 17.1.3 'First line' defence mechanisms of the lung

Adrian R. W. Hatfield Consultant Gastroenterologist, The Middlesex Hospital, London, UK.
14.2.2 Upper gastrointestinal endoscopy

P. N. Hawkins Professor of Medicine, Royal Free and University College Medical School, London, UK.
11.12.3 Familial Mediterranean fever and other inherited periodic fever syndromes . 11.12.4 Amyloidosis

Keith Hawton Professor of Psychiatry, University Department of Psychiatry and Director and Consultant Psychiatrist, Centre for Suicide Research, Warneford Hospital, Oxford, UK.
26.5.2 The patient who has attempted suicide

R. J. Hay Professor and Dean, Faculty of Medicine and Health Sciences, Queens University, Belfast, UK.
7.11.27 Nocardiosis. 7.12.1 Fungal infections

B. Hazleman Consultant Rheumatologist, Rheumatology Department, Addenbrooke's Hospital, Cambridge, UK.
18.11 Miscellaneous conditions presenting to the rheumatologist

Nick Heather Consultant Clinical Psychologist and Director, Centre for Alcohol and Drug Studies, Newcastle, North Tyneside, and Northumberland Mental Health NHS Trust, Newcastle upon Tyne, UK.
26.7.2 Brief interventions against excessive alcohol consumption

David B. Hellmann Professor, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
18.10.7 Polymyositis and dermatomyositis

D. J. Hendrick Consultant Physician and Professor of Occupational Respiratory Medicine, Royal Victoria Infirmary, University of Newcastle upon Tyne, UK.
17.11.8 Pulmonary haemorrhagic disorders. 17.11.9 Eosinophilic pneumonia. 17.11.10 Lymphocytic infiltrations of the lung . 17.11.11 Extrinsic allergic alveolitis . 17.11.12 Eosinophilic granuloma of the lung and pulmonary lymphangiomyomatosis. 17.11.13 Pulmonary alveolar proteinosis. 17.11.14 Pulmonary amyloidosis. 17.11.15 Lipoid (lipid) pneumonia. 17.11.16 Pulmonary alveolar microlithiasis. 17.11.17 Toxic gases and fumes. 17.11.18 Radiation pneumonitis. 17.11.19 Drug-induced lung disease

Mark Herbert Clinical Lecturer in Neonatal Paediatrics, Department of Paediatrics, University of Oxford, UK.
13.15 Infections in pregnancy

Andrew Herxheimer Emeritus Fellow, UK Cochrane Centre, London, UK.
9 Principles of clinical pharmacology and drug therapy

Martin F. Heyworth Chief of Staff and Clinical Professor of Medicine, VA Medical Center and University of Pennsylvania, Philadelphia, USA.
7.13.8 Giardiasis, balantidiasis, isosporiasis, and microsporidiosis

Tim Higenbottam Global Clinical Expert, Astra-Zeneca, Charnwood, Leicestershire and Visiting Professor of Medicine, University of Sheffield, UK.
15.15.1 The pulmonary circulation and its influence on gas exchange . 15.15.2.1 Primary pulmonary hypertension

Katherine A. High William H. Bennett Professor of Pediatrics, University of Pennsylvania School of Medicine and The Children's Hospital of Philadelphia, Pennsylvania, USA.
22.6.4 Genetic disorders of coagulation

S. L. Hillier Research Associate Professor of Obstetrics and Gynecology, University of Washington, Seattle, USA.
21.3 Vaginal discharge

David Hilton-Jones Clinical Director, Oxford MDC Muscle and Nerve Centre, Radcliffe Infirmary, Oxford, UK.
24.17 Diseases of the neuromuscular junction. 24.22.3 Myotonia. 24.22.4 Metabolic and endocrine disorders

John R. Hodges Professor of Behavioural Neurology, MRC Cognition and Brain Sciences Unit and Department of Neurology, Addenbrooke's Hospital, Cambridge, UK.
24.8 Disturbances of higher cerebral function . 24.13.8 Alzheimer's disease and other dementias

H. J. F. Hodgson Sheila Sherlock Professor of Medicine and Director, Centre for Hepatology, Royal Free and University College Medical School, London, UK.
14.9.6 Whipple's disease . 14.20.1 Viral hepatitis - clinical aspects . 14.20.2.1 Autoimmune hepatitis

A. V. Hoffbrand Emeritus Professor of Haematology, Royal Free and University College School of Medicine, London, UK.
22.5.6 Megaloblastic anaemia and miscellaneous deficiency anaemias

Ronald Hoffman Professor, Hematology-Oncology Section University of Illinois College of Medicine, Chicago, USA.
22.3.8 The polycythaemias. 22.3.10 Thrombocytosis

P. A. H. Holloway Consultant Chemical Pathologist in Intensive Care and Honorary Reader in Medicine, John Radcliffe Hospital, Oxford, UK.
32 Reference intervals for biochemical data

Richard H. Holloway Associate Professor of Medicine and Senior Consultant Gastroenterologist, Department of Gastroenterology, Hepatology and General Medicine, Royal Adelaide Hospital, Australia.
14.6 Diseases of the oesophagus

J. M. Hopkin Professor, Experimental Medicine Unit, Swansea Clinical School, University of Wales, Swansea, UK.
17.4.1 Asthma: genetic effects. 17.15 The genetics of lung diseases

Carol Ann Huff Assistant Professor of Oncology, Sidney Kimmel Comprehensive Cancer Care at Johns Hopkins, Baltimore, Maryland, USA.
26.7.3 Problems of alcohol and drug users in the hospital

I. A. Hughes Professor of Paediatrics and Honorary Consultant Paediatric Enterologist, Department of Paediatrics, University of Cambridge, UK.

12.7.2 Congenital adrenal hyperplasia

Lawrence Impey Consultant in Fetal Medicine, The Women's Centre, John Radcliffe Hospital, Oxford, UK.
13.15 Infections in pregnancy

C. W. Imrie Consultant Surgeon and Honorary Professor, Lister Department of Surgery, Royal Infirmary, Glasgow, UK.
14.18.3.1 Acute pancreatitis

H. Irving Consultant Radiologist, St James's University Hospital, Leeds, UK.
20.15 Tumours of the urinary tract

P. G. Isaacson Professor of Histopathology, Royal Free and University College Medical School, London, UK.
14.9.4 Gastrointestinal lymphoma

D. A. Isenberg The Arthritis Research Campaign Professor of Rheumatology at University College London, Centre for Rheumatology, London, UK.
18.10.1 Autoimmune rheumatic disorders and vasculitis. 18.10.2 Systemic lupus erythematosus and related disorders

C. G. Isles Consultant Physician, Medical Unit, Dumfries and Galloway Royal Infirmary, Dumfries, UK.
15.16.1.1 Prevalence, epidemiology, and pathophysiology of hypertension

C. Ison Reader in Medical Microbiology, Department of Infectious Diseases and Microbiology, Faculty of Medicine, Imperial College, St Mary's Campus, London, UK.
7.11.6 Neisseria gonorrhoeae

Alan A. Jackson Professor and Director, Institute of Human Nutrition, University of Southampton, UK.
10.4 Severe malnutrition

H. S. Jacobs Emeritus Professor of Reproductive Endocrinology, University College London Medical School, UK.
12.8.1 Ovarian disorders. 12.8.3 The breast

Robin Jacoby Professor of Old Age Psychiatry, University of Oxford Department of Psychiatry, Warneford Hospital, Oxford, UK.
30.2 Mental disorders of old age

O. F. W. James Head of Clinical Medical Sciences, Medical School, University of Newcastle upon Tyne, UK.
14.21.1 Alcoholic liver disease and non-alcoholic steatosis hepatitis

Paul J. Jenkins Senior Lecturer in Endocrinology, St Bartholomew's Hospital, London, UK.
12.2 Disorders of the anterior pituitary

B. Jennett Emeritus Professor of Neurosurgery, Institute of Neurological Sciences, University of Glasgow, UK.
24.13.6 Brain death and the vegetative state

D. P. Jewell Professor of Gastroenterology, John Radcliffe Hospital, Oxford, UK.
14.9.3 Coeliac disease. 14.10 Crohn's disease. 14.11 Ulcerative colitis. 14.22 Miscellaneous disorders of gastrointestinal tract and liver

Vivekanand Jha Associate Professor of Nephrology, Postgraduate Institute of Medical Education and Research, Chandigarh, India.
20.7.10 Glomerular disease in the tropics

Anne M. Johnson Professor of Infectious Disease Epidemiology and Head, Department of Primary Care and Population Sciences, University College London, UK.
21.2 Sexual behaviour

A. W. Johnson CAB International, Wallingford, Oxfordshire, UK.
8.3 Poisonous plants and fungi

E. Anthony Jones Chief of Hepatology, Academic Medical Centre, Amsterdam, The Netherlands.
14.21.3 Hepatocellular failure

N. Jones Department of Virology, John Radcliffe Hospital, Oxford, UK.
7.10.25 Orf. 7.10.26 Molluscum contagiosum

S. E. Jones Research Associate, Department of Biology, Imperial College of Science, Technology and Medicine, London, UK.
7.11.33 Syphilis

Kenneth C. Kalunian Professor of Medicine, UCLA School of Medicine, Los Angeles, California, USA.
18.8 Osteoarthritis

Eileen Kaner NHS Primary Care Career Scientist, School of Population and Health Sciences, University of Newcastle upon Tyne, UK.
26.7.2 Brief interventions against excessive alcohol consumption

W. Katon Professor and Vice Chair, Director of Division of Health Services and Psychiatric Epidemiology, University of Washington, Seattle, Washington, USA.
26.5.4 Anxiety and depression

Tomisaku Kawasaki Professor and Director, Japan Kawasaki Disease Research Center, Tokyo, Japan.
18.10.8 Kawasaki syndrome

David Keeling Consultant Haematologist and Director, Oxford Haemophilia Centre and Thrombosis Unit, Churchill Hospital, Oxford, UK.
15.5.2 Therapeutic anticoagulation in atrial fibrillation and heart failure . 15.15.3.2 Therapeutic anticoagulation in deep venous thrombosis and pulmonary embolism

David P. Kelsell Non-Clinical Senior Lecturer, Centre for Cutaneous Research, Barts and The London, Queen Mary's School of Medicine and Dentistry, London, UK.
23.2 Molecular basis of inherited skin disease

John G. Kelton Dean and Vice-President, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada.
22.6.3 Disorders of platelet number and function

Christopher Kennard Professor and Head, Division of Neuroscience and Psychological Medicine, Imperial College London, Charing Cross Campus, London, UK.
24.11 Visual pathways

Rose Anne Kenny Professor of Cardiovascular Research, Institute of Ageing and Health, University of Newcastle upon Tyne, UK.
24.13.5.1 Head-up tilt-table testing in the diagnosis of vasovagal syncope and related disorders

M. G. W. Kettlewell Consultant Surgeon, Oxford Radcliffe Trust, UK.
14.13 Colonic diverticular disease

G. T. Keusch Associate Director for International Research, National Institutes of Health, Bethesda, Maryland, and Professor of Medicine, Tufts-New England Medical Center, Boston, Massachusetts, USA.
7.11.7 Enterobacteria, campylobacter, and miscellaneous food-poisoning bacteria

Munther A. Khamashta Senior Lecturer and Consultant Physician, Lupus Research Unit, The Rayne Institute, St Thomas' Hospital, London, UK.
13.14 Autoimmune rheumatic disorders and vasculitis in pregnancy

Maurice King Honorary Research Fellow, University of Leeds, UK.
3.7.2 Health in a fragile future

Keith P. Klugman Professor of International Health, Rollins School of Public Health and Division of Infectious Diseases, School of Medicine, Emory University, Atlanta, Georgia, USA.
7.11.3 Pneumococcal diseases

R. Knight Associate Specialist in General Medicine, Royal Sussex County Hospital, Brighton, UK.
7.13.1 Amoebic infections. 7.13.9 *Blastocystis hominis* infection. 7.14.2 Lymphatic filariasis. 7.14.3 Guinea-worm disease: dracunculiasis. 7.14.4 Strongyloidiasis, hookworm, and other gut strongyloid nematodes. 7.14.8 Angiostrongyliasis. 7.15.2 Gut cestodes

Michael D. Kopelman Professor of Clinical Medicine and Deputy Warden, Bart's and The London, Queen Mary's School of Medicine and Dentistry, University of London, UK.
26.3 Neuropsychiatric disorders

Peter G. Kopelman Professor of Clinical Medicine, Bart's and The London Queen Mary's School of Medicine and Dentistry, London, UK.
10.5 Obesity

Christian Krarup Professor, Department of Clinical Neurophysiology, Rigshospitalet, Copenhagen, Denmark.
24.2 Electrophysiology of the central and peripheral nervous systems

J. B. Kurtz Consultant Virologist (retired), Public Health Laboratory, Birmingham Heartlands Hospital, UK.
7.11.35 Legionellosis and legionnaires' disease

Robert A. Kyle Professor of Medicine and Laboratory Medicine, Mayo Clinic, Rochester, Minnesota, USA.
22.4.5 Myeloma and paraproteinaemias

David Lalloo Senior Lecturer in Tropical Medicine, Liverpool School of Tropical Medicine, UK.
7.11.17 Yersinia, Pasteurella, and Francisella

D. J. Lane Consultant Chest Physician (Retired), Oxford Radcliffe Hospital, UK.
17.2 The clinical presentation of chest diseases

Peter Lanyon Consultant Rheumatologist, University Hospital, Queen's Medical Centre, Nottingham, UK.
18.3 Clinical investigation

H. E. Larson Private Practice in Infectious Diseases, Marlborough, Massachusetts, USA.
7.11.21 Botulism, gas gangrene, and clostridial gastrointestinal infections

S. Lawrie Senior Clinical Research Fellow, University Department of Psychiatry, Royal Edinburgh Hospital, UK.
26.5.6 Schizophrenia, bipolar disorder, obsessive-compulsive disorder, and personality disorder

N. F. Lawton Consultant Neurologist, Wessex Neurological Centre, Southampton General Hospital and Honorary Senior Lecturer, University of Southampton, UK.
24.13.19 Intracranial hypertension

John H, Lazarus Professor of Clinical Endocrinology, University of Wales College of Medicine, Cardiff, UK.
13.11 Endocrine disease in pregnancy

J. W. LeDuc Director, Division of Viral and Rickettsial Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA.
7.10.14 Bunyaviridae

P. J. Lee Consultant in Metabolic Medicine, Metabolic Unit, National Hospital for Neurology and Neurosurgery, London, UK.
11.2 Inborn errors of amino acid and organic acid metabolism

Tak H. Lee Professor of Allergy and Respiratory Medicine, Guy's, King's and St Thomas' School of Medicine, Guy's Hospital, London, UK.
17.4.3 Basic mechanisms and pathophysiology of asthma

William M. F. Lee Department of Medicine, School of Medicine, University of Pennsylvania, Philadelphia, USA.
4.3 Molecular cell biology

T. Lehner Professor of Basic and Applied Immunology, Department of Immunobiology, Guy's, King's and St Thomas' School of Medicine, London, UK.
14.5 The mouth and salivary glands. 18.10.5 Behcet's disease

Irene M. Leigh Professor of Cellular and Molecular Medicine, Bart's and The London Queen Mary's School of Medicine and Dentistry, University of London, UK.
23.2 Molecular basis of inherited skin disease

G. G. Lennox Consultant Neurologist, Addenbrooke's Hospital, Cambridge, UK.
13.12 Neurological disease in pregnancy

E. A. Letsky Consultant Perinatal Haematologist, Queen Charlotte's and Chelsea Hospital, London, UK.
13.16 Blood disorders in pregnancy

Jeremy Levy Consultant Nephrologist, Imperial College, Hammersmith Hospital, London, UK.
20.7.7 Antiglomerular basement membrane disease

L. M. Lichtenstein Professor of Medicine and Director, Asthma and Allergy Center, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
5.2 Allergy

D. C. Linch Professor and Head of Haematology, University College London, UK.
22.2.2 Stem-cell disorders

M. J. Lindop Consultant, Anaesthesia/Intensive Care, Addenbrooke's Hospital, Cambridge, UK.
16.6.3 Brainstem death and organ donation. 16.6.4 The patient without hope

Calvin C. Linnemann, Jr Professor and Director, Infectious Diseases Division, University of Cincinnati Medical Center, Ohio, USA.
7.11.14 Bordetella

Gregory Y. H. Lip Professor of Cardiovascular Medicine, University Department of Medicine, City Hospital, Birmingham, UK.
15.16.3 Hypertensive emergencies and urgencies

P. Little Professor of Primary Care Research, Community Clinical Sciences Division, University of Southampton, UK.
17.5.1 Upper respiratory tract infections

Roderick A. Little Honorary Professor of Surgical Science, University of Manchester, UK.
11.12.2 Metabolic response to accidental and surgical injury

W. Littler Medical Director, University Hospital NHS Trust, Birmingham, UK.
15.10.2 Infective endocarditis

A. Llanos Cuentas Principal Professor, Facultad de Salud Publica y Administracion, Universidad Peruana Cayetano Heredia, Lima, Peru.
7.11.39.1 *Bartonella bacilliformis* infection

Diana N. J. Lockwood Consultant Leprologist and Senior Lecturer, Hospital for Tropical Diseases and London School of Hygiene and Tropical Medicine, UK.
7.11.24 Leprosy (Hansen's disease)

S. Logan Senior Lecturer in Paediatric Epidemiology, Institute of Child Health, London, UK.
7.10.12 Rubella

D. J. Lomas Professor of Clinical MRI, University Department of Radiology, Addenbrooke's Hospital, Cambridge, UK.
14.18.2 Computed tomography and magnetic resonance imaging of the liver and pancreas

David A. Lomas Professor of Respiratory Biology and Honorary Consultant Physician, Department of Medicine, University of Cambridge Institute for Medical Research, UK.
11.13 a_1-Antitrypsin deficiency and the serpinopathies

Thomas Look Professor of Pediatrics, Harvard Medical School and Vice-Chair for Research, Pediatric Oncology Department, Dana-Farber Institute, Boston, Massachusetts, USA.
22.3.1 Cell and molecular biology of human leukaemias

A. D. Lopez Senior Science Adviser, World Health Organization, Geneva, Switzerland.
3.1 The Global Burden of Disease Study

Elyse E. Lower Professor of Medicine, University of Cincinnati, Ohio, USA.
17.11.6 Sarcoidosis

Linda M. Luxon Professor of Audiological Medicine, University of London, Institute of Child Health, London, UK and Director, National Institute for Cancer Research, Genova, Italy.
24.12.2 Disorders of hearing

Lucio Luzzatto Professor, Department of Human Genetics, Memorial Sloan- Kettering Cancer Center, New York, USA.
22.3.12 Paroxysmal nocturnal haemoglobinuria. 22.5.12 Glucose-6-phosphate-dehydrogenase (G6PD) deficiency

G. A. Luzzi Consultant in Genitourinary/HIV Medicine, South Buckinghamshire NHS Trust, Wycombe Hospital, High Wycombe, Buckinghamshire, UK.
7.10.21 HIV and AIDS

D. C. W. Mabey Professor of Communicable Diseases, London School of Hygiene and Tropical Medicine, London, UK.
7.11.40 Chlamydial infections including lymphogranuloma venerum

P. K. MacCallum Senior Lecturer in Haematology, Barts and The London, Queen Mary's School of Medicine and Dentistry, London, UK.
15.1.2.2 The haemostatic system in arterial disease

J. T. Macfarlane Consultant Physician, Nottingham City Hospital, UK.
7.11.35 Legionellosis and legionnaires' disease

K. T. MacLeod Reader in Cardiac Physiology, Cardiac Medicine, NHLI, Faculty of Medicine, Imperial College London, UK.
15.1.3.1 Physical considerations: biochemistry and cellular physiology of heart muscle

William MacNee Professor of Respiratory and Environmental Medicine, University of Edinburgh, and Honorary Consultant Physician, Lothian University NHS Trust, Edinburgh, UK.
17.6 Chronic obstructive pulmonary disease

M. Monir Madkour Consultant Physician, Military Hospital, Riyadh, Saudi Arabia.
7.11.19 Brucellosis

R. N. Maini Professor of Rheumatology in the University of London, Head of the Kennedy Institute of Rheumatology Division, Faculty of Medicine, Imperial College London, and Honorary Consultant Physician, Charing Cross Hospital, London, UK.
18.5 Rheumatoid arthritis

Hadi Manji Consultant Neurologist, National Hospital for Neurology, London and Ipswich Hospital, Suffolk, UK.
24.14.4 Neurosyphilis and neuroAIDS

J. I. Mann Professor in Human Nutrition and Medicine, University of Otago, Dunedin, New Zealand.
10.1 Diseases of overnourished societies and the need for dietary change

D. Mant Professor of General Practice, Department of Primary Health Care, University of Oxford, UK.

3.4 Preventive medicine

Victor J. Marder Orthopedic Hospital/UCLA Vascular Medicine Program, Los Angeles, California, USA.
22.6.2 Evaluation of the patient with a bleeding diathesis

A. F. Markham Professor of Medicine, St James's University Hospital, Leeds, UK.
14.15 Tumours of the gastrointestinal tract

V. Marks Professor of Clinical Biochemistry Emeritus, Post-Graduate Medical School, University of Surrey, Guildford, UK.
12.11.3 Hypoglycaemia

T. J. Marrie Professor and Chair, Department of Medicine, University of Alberta, Edmonton, Canada.
7.11.38 *Coxiella burneti*i infections (Q fever)

Helen Marriott Research Associate, Department of Respiratory Medicine, University of Shefffield, UK.
15.15.2.1 Primary pulmonary hypertension

C. D. Marsden* Professor of Neurology, National Hospital for Neurology and Neurosurgery, London, UK.
24.15 Metabolic disorders and the nervous system

Jay W. Mason Professor and Chair, Department of Medicine, University of Kentucky College of Medicine, Lexington, USA.
15.8.1 Myocarditis

V. I. Mathan Professor, ICDDR, Dhaka, Bangladesh.
14.9.8 Malabsorption syndromes in the tropics

Christopher J. Mathias Professor of Neurovascular Medicine and Consultant Physician, Imperial College of Science, Technology and Medicine at St Mary's and National Hospital for Neurology and Neurosurgery, Institute of Neurology, University College London, UK.
24.13.14 Disorders of the autonomic nervous system

Peter W. Mathiesen Professor of Renal Medicine, Academic Renal Unit, University of Bristol, Southmead Hospital, Bristol, UK.
20.7.5 Proliferative glomerulonephritis . 20.7.6 Mesangiocapillary glomerulonephritis

R. McCaig Head, Human Factors Unit, Health Directorate, Health and Safety Executive, Bootle, UK.
8.5.10 Noise. 8.5.11 Vibration

Mary E. McCaul Professor, Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
26.7.1 Alcohol and drug dependence

Joseph McCormick Regional Dean, University of Texas School of Public Health at Brownsville, USA.
7.10.15 Arenaviruses. 7.10.16 Filoviruses

William J. McKenna BHF Professor of Molecular Cardiovascular Sciences, Department of Cardiological Sciences, St George's Hospital Medical School, London, UK.
15.8.2 The cardiomyopathies: hypertrophic, dilated, restrictive, and right ventricular . 15.8.3 Specific heart muscle disorders

A. J. McMichael Professor and Director, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, UK.
5.1 Principles of immunology.

A. J. McMichael Professor and Director, National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia.
3.2 Human population size, environment, and health

A. McMillan Consultant Physician, Department of Genito-urinary Medicine, Edinburgh Royal Infirmary, UK.
21.5 Infections and other medical problems in homosexual men

Martin McNally Consultant in Limb Reconstruction and Honorary Senior Lecturer in Orthopaedic Surgery, Bone Infection Unit, Nuffield Orthopaedic Centre, Oxford, UK.
19.3 Osteomyelitis

K. McNeil Director of Transplant Services, The Prince Charles Hospital, Brisbane, Australia.
17.16 Lung and heart-lung transplantation

T. W. Meade Emeritus Professor of Epidemiology, London School of Hygiene and Tropical Medicine, UK.
15.1.2.2 The haemostatic system in arterial disease

A. Meheus Professor, University of Antwerp, Belgium.
21.1 Epidemiology

David K. Menon Professor of Anaesthesia, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK.
16.6.2 Management of raised intracranial pressure

Wayne M. Meyers Chief, Mycobacteriology, Armed Forces Institute of Pathology, Washington DC, USA.
7.11.25 Buruli ulcer: *Mycobacterium ulcerans* infection

Anna Rita Migliaccio Dirigente de Ricerca in Transfusion Medicine, Laboratory of Clinical Biochemistry, Istituto Superiore dei Sanità, Rome, Italy.
22.5.1 Erythropoiesis and the normal red cell

M. A. Miles Professor, London School of Hygiene and Tropical Medicine, UK.
7.13.11 Chagas' disease

G. J. Miller Professor of Epidemiology, Barts and The London, Queen Mary's School of Medicine and Dentistry, London, UK.
15.1.2.2 The haemostatic system in arterial disease

Mary Miller Consultant in Palliative Medicine, Sir Michael Sobell House, Churchill Hospital, Oxford, UK.
31 Palliative care

Robert F. Miller Reader in Clinical Infection and Consultant Physician, Royal Free and University College Medical School, London, UK.
7.12.5 Pneumocystis carinii

K. R. Mills Professor of Clinical Neurophysiology, King's College Hospital, London, UK.

24.4 Investigation of central motor pathways: magnetic brain stimulation

Philip Minor Public Health and Clinical Microbiology Laboratory, Addenbrooke's Hospital, Cambridge, UK.
7.10.7 Enterovirus infections

Raad H. Mohiaddin Consultant and Honorary Senior Lecturer, Royal Brompton and Harefield NHS Trust, London, JK.
15.3.5 Cardiovascular magnetic resonance and computed X-ray tomography

Andrew J. Molyneux Consultant Neuroradiologist, Radcliffe Infirmary, Oxford, UK.
24.5 Neuroimaging in neurological diseases

Kevin Moore Senior Lecturer, Centre for Hepatology, Royal Free ospital and University College Medical School, London, UK.
14.21.2 Cirrhosis, portal hypertension and ascites

Pedro L. Moro Fellow, Vaccine Safety Division, National Immunization Program, Centers for Disease Control and Prevention, Baltimore, Maryland, USA.
7.15.1 Cystic hydatid disease (Echinococcus granulosus)

N. J. McC. Mortensen Professor of Colorectal Surgery, Department of Colorectal Surgery, John Radcliffe Hospital, Oxford, UK.
14.13 Colonic diverticular disease

Peter S. Mortimer Professor of Dermatological Medicine and Consultant Skin Physician, St George's Hospital Medical School, Division of Physiological Medicine, London, UK.
15.17 Lymphoedema

Alastair G. Mowat Clinical Lecturer in Rheumatology, Department of Rheumatology, Nuffield Orthopaedic Centre, Oxford, UK.
18.10.4 Polymyalgia rheumatica and giant cell arteritis

E. R. Moxon Head, Oxford University Department of Paediatrics, John Radcliffe Hospital, Oxford, UK.
7.11.12 Haemophilus influenzae

M. F. Muers Consultant Physician, Respiratory Medicine, The General Infirmary at Leeds, UK.
17.3.4 Diagnostic bronchoscopy, thoracoscopy, and tissue biopsy

Tariq I. Mughal Consultant Haematologist and Medical Oncologist and Senior Lecturer in Oncology, Lancashire Teaching Hospitals NHS Trust and Preston and Christie Hospital NHS Trust, Manchester, UK.
22.3.6 Chronic myeloid leukaemia

J. A. Muir Gray Director of the UK National Screening Committee, Institute of Health Sciences, Oxford, UK.
3.6 Screening

P. A. Murphy Professor of Medicine and Microbiology, Johns Hopkins University and Chief, Infectious Diseases Division, Johns Hopkins Bayview Hospital, Baltimore, Maryland, USA.
7.5 Physiological changes in infected patients

C. J. L. Murray Global Programme on Evidence for Health Policy, World Health Organization, Geneva, Switzerland.
3.1 The Global Burden of Disease Study

Iain M. Murray-Lyon Consultant Physician and Gastroenterologist, Charing Cross Hospital and Chelsea and Westminster Hospital, London, UK.
14.21.5 Primary and secondary liver tumours

Jean Nachega Assistant Scientist, Johns Hopkins University, Baltimore, Maryland, USA.
7.11.22 Tuberculosis

Robert B. Nadelman Professor of Medicine, Division of Infectious Diseases, New York Medical College, USA.
7.11.29 Lyme borreliosis

N. V. Naoumov Reader in Hepatology/Honorary Consultant Physician, Institute of Hepatology, University College London, UK.
7.10.19 Hepatitis viruses (including TTV)

R. P. Naoumova MRC Senior Clinical Scientist/Honorary Consultant Physician, MRC Clinical Sciences Centre, Hammersmith Hospital, London, UK.
15.1.2.1 The pathogenesis of atherosclerosis

D. G. Nathan President, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.
22.2.1 Stem cells and haemotopoiesis

Graham Neale Research Fellow, Clinical Risk Unit, University College London, UK.
14.1 Introduction to gastroenterology. 14.1.1.2 Symptomatology of gastrointestinal disease. 14.16 Vascular and collagen disorders

Catherine Nelson-Piercy Consultant Obstetric Physician, Guy's and St Thomas' Hospitals Trust, London, UK.
13.14 Autoimmune rheumatic disorders and vasculitis in pregnancy

A. R. Ness Senior Lecturer in Epidemiology, Department of Social Medicine, University of Bristol, UK.
15.4.1.2 The epidemiology of ischaemic heart disease

Peter Nestor Neurologist, University of Cambridge Neurology Unit, UK.
24.8 Disturbances of higher cerebral function

J. Neuberger Professor of Hepatology and Consultant Physician, Queen Elizabeth Hospital, Birmingham, UK.
14.21.7 Drugs and liver damage. 14.21.8 The liver in systemic disease

John Newell-Price Senior Lecturer in Endocrinology, Division of Clinical Sciences, Sheffield University, Northern General Hospital, Sheffield, UK.
12.3 Disorders of the posterior pituitary

A. J. Newman Taylor Consultant Physician and Head, Department of Occupational and Environmental Medicine, Royal Brompton Harefield NHS Trust, Faculty of Medicine, Imperial College London, UK.
17.4.4 Asthma. 17.4.5 Occupational asthma

C. S. Ng Assistant Professor, Department of Radiology, University of Texas M. D. Anderson Cancer Center, Houston, USA.

14.18.2 Computed tomography and magnetic resonance imaging of the liver and pancreas

S. Nightingale Consultant Neurologist and Honorary Senior Clinical Lecturer, Royal Shrewsbury Hospital and Birmingham University, Shrewsbury, UK.
7.10.23 HTLV-I and II and associated diseases

T. Northfield Professor Emeritus, Department of Biochemical Medicine, St George's Hospital, London, UK.
14.3.2 Gastrointestinal bleeding

John Nowakowski Assistant Professor of Medicine, Department of Medicine, Division of Infectious Diseases, Westchester Medical Center, Valhalla, New York, USA.
7.11.29 Lyme borreliosis

Fujio Numano Director, Tokyo Vascular Disease Institute, Tokyo, Japan.
15.14.4 Takayasu arteritis

D. O'Gradaigh Research Registrar, Department of Medicine, Addenbrooke's Hospital, Cambridge, UK.
18.11 Miscellaneous conditions presenting to the rheumatologist . 19.5 Avascular necrosis and related topics

Stephen O'Rahilly Professor of Clinical Biochemistry, University of Cambridge, and Honorary Consultant Physician, UK.
10.5 Obesity

S. C. O'Reilly Consultant Rheumatologist, Rheumatology Department, Derbyshire Royal Infirmary, Derby, UK.
18.9 Crystal-related arthropathies

P. J. Oldershaw Consultant Cardiologist, Royal Brompton Hospital, London, UK.
15.13 Congenital heart disease in adolescents and adults

James G. Olson Head, Department of Virology, U. S. Navy Medical Research Center Detachment, Lima, Peru.
7.10.6.1 Nipah and Hendra viruses . 7.11.39 Bartonelloses, excluding *Bartonella bacilliformis* infections

M. Osame Professor, Third Department of Internal Medicine, Faculty of Medicine, Kagoshima University, Japan.
7.10.23 HTLV-I and II and associated diseases

Jackie Palace Consultant Neurologist, Radcliffe Infirmary, Oxford, UK.
24.17 Diseases of the neuromuscular junction

Thalia Papayannopoulou Professor of Medicine (Hematology), University of Washington, Division of Hematology, Seattle, USA.
22.5.1 Erythropoiesis and the normal red cell

S. Parish Senior Research Fellow, Clinical Trial Service Unit, Nuffield Department of Clinical Medicine, University of Oxford, UK.
2.4.3 Large-scale randomized evidence: trials and overviews

G. R. Park Director of Intensive Care Research, John Farman Intensive Care Unit, Addenbrooke's Hospital, Cambridge, UK.
16.6.1 Sedation and analgesia in the critically ill

David Parkes Professor of Clinical Neurology, King's College Hospital, London, UK.
24.13.4 Narcolepsy

C. Parry University of Oxford-Wellcome Trust Clinical Research Unit, Centre for Tropical Diseases, Ho Chi Minh City, Vietnam.
7.11.8 Typhoid and paratyphoid fevers

Steve W. Parry Consultant Physician and Honorary Senior Lecturer, Freeman Hospital and University of Newcastle upon Tyne, UK.
24.13.5.1 Head-up tilt-table testing in the diagnosis of vasovagal syncope and related disorders

J. Paul Consultant Microbiologist and Director, Brighton Public Health Laboratory, Royal Sussex County Hospital, Brighton, UK.
7.11.42 Newly identified and lesser-known bacteria . 7.17 Non-venomous arthropods

Malik Peiris Professor, Department of Microbiology, University of Hong Kong.
7.10.1 Respiratory tract viruses

Edmund D. Pellegrino Emeritus Professor of Medicine and Medical Ethics, Georgetown University Medical Center, Washington DC, USA.
2.3 Medical ethics

T. H. Pennington Professor of Bacteriology, University of Aberdeen Medical School, UK.
7.3 Biology of pathogenic micro-organisms

M. B. Pepys Professor and Head of Medicine, Department of Medicine, Royal Free Campus, Royal Free and University College Medical School, London, UK.
11.12.1 The acute phase response and C-reactive protein . 11.12.4 Amyloidosis

P. L. Perine Professor of Epidemiology, School of Public and Community Medicine, University of Washington, Seattle, USA.
7.11.32 Non-venereal treponematoses: yaws, endemic syphilis (bejel), and pinta

G. D. Perkin Consultant Neurologist, Department of Neurology, Charing Cross Hospital, London, UK.
24.13.3 Epilepsy in later childhood and adults

P. L. Perrotta Assistant Professor, Pathology, Stony Brook University Hospital, New York, USA.
22.8.1 Blood transfusion

H. Persson Medical Director and Consultant Physician, Swedish Poisons Information Centre, Stockholm, Sweden.
8.3 Poisonous plants and fungi

M. C. Petch Consultant Cardiologist, Papworth Hospital, Cambridge, UK.
15.4.2.6 The impact of coronary heart disease on life and work

L. R. Petersen Deputy Director for Science, Centers for Disease Control, Division of Vector-borne Infectious Diseases, Fort Collins, Colorado, USA.
7.10.11 Alphaviruses. 7.10.13 Flaviviruses

R. Peto Professor of Epidemiology and Medical Statistics, University of Oxford, UK.
2.4.3 Large-scale randomized evidence: trials and overviews. 6.1 Epidemiology of cancer

T. E. A. Peto Consultant Physician in Infectious Diseases, Nuffield Department of Medicine, John Radcliffe Hospital, Oxford, UK.
7.10.21 HIV and AIDS

A. Phillips Senior Lecturer, Institute of Nephrology, University of Wales College of Medicine, Cardiff, UK.
20.1 Structure and function of the kidney

R. J. Playford Professor, Imperial College School of Medicine, Hammersmith Hospital, London, UK.
14.9.7 Effects of massive small bowel resection

J. M. Polak Professor and Director, Tissue Engineering and Regenerative Medicine Centre, Imperial College School of Medicine, London, UK.
14.8 Hormones and the gastrointestinal tract

Eleanor S. Pollak Associate Director, Clinical Coagulation Laboratory, Hospital of the University of Pennsylvania, University of Pennsylvania Medical Center, Philadelphia, USA.
22.6.4 Genetic disorders of coagulation

P. A. Poole-Wilson Professor of Cardiology and Cardiac Medicine, National Heart and Lung Institute, Faculty of Medicine, Imperial College London, UK.
15.1.3.1 Physical considerations: biochemistry and cellular physiology of heart muscle

F. M. Pope Consultant Dermatologist, West Middlesex University Hospital, London, UK.
19.2 Inherited defects of connective tissue: Ehlers-Danlos syndrome, Marfan's syndrome, and pseudoxanthoma elasticum

Françoise Portaels Professor and Head, Mycobacteriology Unit, Institute of Tropical Medicine, Antwerp, Belgium.
7.11.25 Buruli ulcer: *Mycobacterium ulcerans* infection

J. S. Porterfield Formerly Reader in Bacteriology, Sir William Dunn School of Pathology, University of Oxford, UK.
7.10.14 Bunyaviridae

Jerome B. Posner Attending Neurologist, Memorial Sloan-Kettering Cancer Center, New York, USA.
24.18 Paraneoplastic syndromes

William G. Powderly Professor of Medicine, Washington University School of Medicine, St Louis, Missouri, USA.
7.12.2 Cryptococcosis

J. J. Powell Senior Lecturer - Nutrition and Medicine, GI Laboratory, Rayne Institute, St Thomas' Hospital, London, UK.
8.5.8 Podoconiosis

Janet Powell Medical Director, University Hospitals, Coventry and Warwickshire NHS Trust, Coventry, Warwickshire, UK.
15.14.2 Peripheral arterial disease

J. W. Powles University Lecturer in Public Health Medicine, Institute of Public Health, Cambridge, UK.
3.2 Human population size, environment, and health

M. A. Preece Professor of Child Health and Growth, Institute of Child Health, University College London, UK.
12.9.2 Normal growth and its disorders

J. S. Prichard* Professor of Medicine, St James's Hospital, Dublin, Eire.
15.15.2.2 Pulmonary oedema

A. T. Proudfoot Consulting Clinical Toxicologist, National Poisons Information Service, City Hospital, Birmingham, UK.
8.1 Poisoning by drugs and chemicals

Charles Pusey Professor of Renal Medicine, Faculty of Medicine, Imperial College, Hammersmith Hospital, London, UK.
20.7.7 Antiglomerular basement membrane disease

N. P. Quinn Professor of Clinical Neurology, Institute of Neurology and Honorary Consultant Neurologist, The National Hospital for Neurology and Neurosurgery, London, UK.
24.10 Subcortical structures-the cerebellum, thalamus and basal ganglia

Anisur Rahman Senior Lecturer in Rheumatology, Centre for Rheumatology, Department of Medicine, University College London, UK.
18.10.2 Systemic lupus erythematosus and related disorders

Lawrence E. Ramsay Professor of Clinical Pharmacology and Therapeutics, University of Sheffield and Consultant Physician, Royal Hallamshire Hospital, Sheffield, UK.
15.16.2.1 Hypertension-indications for investigation . 15.16.2.2 Renal and renovascular hypertension . 15.16.2.5 Aortic coarctation . 15.16.2.6 Other rare causes of hypertension. 20.10.2 Hypertension and the kidney

M. Ramsay Consultant Epidemiologist, Immunisation Division, PHLS Communicable Disease Surveillance Centre, London, UK.
7.7 Immunization

A. C. Rankin Reader in Cardiology, Glasgow Royal Infirmary, UK.
15.2.3 Syncope and palpitation . 15.6 Cardiac arrhythmias

C. W. G. Redman Professor of Obstetric Medicine, John Radcliffe Hospital, Oxford, UK.
13.4 Hypertension in pregnancy

Laurence John Reed Academic Unit of Psychiatry, St Thomas' Hospital, London, UK.
26.3 Neuropsychiatric disorders

A. J. Rees Regius Professor of Medicine, Institute of Medical Sciences, University of Aberdeen, UK.
20.10.3 Vasculitis and the kidney

Jeremy Rees Clinical Senior Lecturer in Neuro-oncology, National Hospital for Neurology and Neurosurgery, London, UK.
24.13.18.1 Intracranial tumours

D. Rennie Adjunct Professor of Medicine, Institute for Health Policy Studies, University of California, San Francisco, USA.
8.5.4 Diseases of high terrestrial altitudes

J. Richens Clinical Lecturer, Department of Sexually Transmitted Diseases, Royal Free and University College Medical School, London, UK.
7.11.8 Typhoid and paratyphoid fevers. 7.11.9 Intracellular Klebsiella infections

B. K. Rima Professor of Molecular Biology, Medical Biology Centre, Queen's University of Belfast, UK.
7.10.5 Mumps: epidemic parotitis

A. J. Ritchie Consultant Cardiothoracic Surgeon, Papworth NHS Trust, Cambridge, UK.
15.4.2.5 Coronary artery bypass grafting

Eberhard Ritz Professor and Head, Department of Nephrology, University of Heidelberg, Germany.
20.5.2 Bone disease in chronic renal failure

Harold R. Roberts Sarah Graham Kenan Professor of Medicine and Attending Physician, UNC Hospitals, Chapel Hill, North Carolina, USA.
22.6.1 The biology of haemostasis and thrombosis. 22.6.2 Evaluation of the patient with a bleeding diathesis

William G. Robertson Clinical Scientist, Institute of Urology and Nephrology, University College London, UK.
20.13 Urinary stones, nephrocalcinosis, and renal tubular acidosis

T. A. Rockall Senior Lecturer/Honorary Consultant, St Mary's Hospital, London, UK.
14.3.2 Gastrointestinal bleeding

Allan R. Ronald Professor Emeritus, University of Manitoba, Winnipeg, Canada.
7.11.13 *Haemophilus ducreyi* and chancroid

P. Ronco Professor of Renal Medicine, Université Pierre et Marie Curie (Paris 6) and Director, Renal Division and INSERM Unit 489, Tenon Hospital (Assistance Publique-Hôpitaux de Paris), Paris, France.
20.10.5 Renal involvement in plasma cell dyscrasias, immunoglobulin-based amyloidoses, and fibrillary glomerulopathies, lymphomas, and leukaemias

Antony Rosen Professor and Director, Division of Rheumatology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
5.3 Autoimmunity

Mark J. Rosen Chief, Division of Pulmonary and Critical Care Medicine, Beth Israel Medical Center, New York, USA.
17.5.2.3 Pulmonary complications of HIV infection

Raymond C. Rosen Professor of Psychiatry, UMDNJ-Robert Wood Johnson Medical School, Department of Psychiatry, Piscataway, New Jersey, USA.
12.8.4 Sexual dysfunction

R. J. M. Ross Professor of Endocrinology, Northern General Hospital, University of Sheffield, UK.
12.9.3 Puberty

D. J. Rowlands Honorary Consultant Cardiologist, Manchester Heart Centre, Manchester Royal Infirmary, UK.
15.3.2 Electrocardiography. 15.3.4 Nuclear techniques

M. B. Rubens Director of Imaging and Consultant Radiologist, Royal Brompton and Harefield NHS Trust, London, UK.
15.3.1 Chest radiography in heart disease. 15.3.5 Cardiovascular magnetic resonance and computed X-ray tomography

David Rubenstein Consultant Physician, Addenbrooke's Hospital, Cambridge, UK.
7.1 The clinical approach to the patient with suspected infection

P. C. Rubin Professor and Dean of Medicine, University of Nottingham, UK.
13.18 Prescribing in pregnancy

Anthony S. Russell Professor of Medicine, University of Alberta, Edmonton, Canada.
18.2 Clinical presentation and diagnosis of rheumatic disease

T. J. Ryan Emeritus Professor of Dermatology, University of Oxford, UK.
23.1 Diseases of the skin

Sara S. T. O. Saad Professor and Haematologist, Department of Internal Medicine, Hematology-Hemotherapy Division, Medical Science Faculty, State University of Campinas, Brazil.
22.5.10 Disorders of the red cell membrane

N. J. Samani Professor of Cardiology, Division of Cardiology, Department of Medicine, University of Leicester, UK.
15.16.1.2 Genetics of hypertension

Brian P. Saunders Senior Lecturer in Endoscopy, St Mark's Hospital, Northwick Park, Harrow, Middlesex, UK.
14.2.1 Colonoscopy and flexible sigmoidoscopy

S. J. Saunders Emeritus Professor, Liver Clinic, Groote Schuur Hospital and Medical Research Council/University of Cape Town Liver Research Centre, Cape Town, South Africa.
14.21.6 Hepatic granulomas

M. O. Savage Professor of Paediatric Endocrinology, St Bartholomew's and The Royal London School of Medicine and Dentistry, London, UK.
12.9.1 Normal and abnormal sexual differentiation. 12.9.3 Puberty

John Savill Professor of Medicine, Royal Infirmary, Edinburgh, UK.
20.7.1 The glomerulus and glomerular injury

K. P. Schaal Professor and Director, Institute for Medical Microbiology and Immunology, Faculty of Medicine, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany.
7.11.26 Actinomycosis

Michael Schömig Physician in Charge, Division of Nephrology, Ruperto-Carola-University of Heidelberg, Germany.
20.5.2 Bone disease in chronic renal failure

Ruud B. H. Schutgens Head of Department of Clinical Chemistry, Vrije Universiteit Medical Centre (VUMC), Amsterdam, The Netherlands.
11.9 Peroxisomal diseases

J. Schwebke Associate Professor of Medicine, University of Alabama at Birmingham, USA.
21.3 Vaginal discharge

Neil Scolding Burden Professor of Clinical Neurosciences, University of Bristol Institute of Clinical Neurosciences, Frenchay Hospital, Bristol, UK.
24.15 Metabolic disorders and the nervous system. 24.20 Neurological complications of systemic autoimmune and inflammatory diseases

J. Scott Professor of Medicine, Imperial College Faculty of Medicine, Hammersmith Campus, London, UK.
15.1.2.1 The pathogenesis of atherosclerosis

A. Seaton Professor and Head of Department of Environmental and Occupational Medicine, University of Aberdeen, UK.
17.11.7 Pneumoconioses

G. R. Serjeant Professor Emeritus and Chairman, Sickle Cell Trust, Kingston, Jamaica, West Indies.
20.10.7 Sickle-cell disease and the kidney

N. J. Severs Professor of Cell Biology, National Heart and Lung Institute, Faculty of Medicine, Imperial College London, UK.
15.1.3.1 Physical considerations: biochemistry and cellular physiology of heart muscle

C. A. Seymour Professor of Clinical Biochemistry and Metabolic Medicine and Director for Clinical Advice to The Health Service Ombudsman, St George's Hospital Medical School and Office of Health Service Commissioner, London, UK.
11.7.2 Wilson's disease, Menke's disease: inherited disorders of copper metabolism

K. V. Shah Professor, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.
7.10.17 Papovaviruses

L. M. Shapiro Consultant Cardiologist, Papworth Hospital, Cambridge, UK.
15.4.2.2 Management of stable angina. 15.4.2.5 Coronary artery bypass grafting

Michael Sharpe Reader in Psychological Medicine, University of Edinburgh, Royal Edinburgh Hospital, UK.
7.19 Chronic fatigue syndrome (postviral fatigue syndrome, neurasthenia, and myalgic encephalomyelitis) 26.1 General introduction. 26.5.3 Medically unexplained symptoms in patients attending medical clinics. 26.6.2 Psychological treatment in medical practice

J. M. Shneerson Director, Respiratory Support and Sleep Centre, Papworth Hospital, Cambridge, UK.
17.13 Disorders of the thoracic cage and diaphragm

Tom Siddons Clinical Research Assistant, Pfizer Research and Development (UK), Maidstone, Kent, UK.
15.15.1 The pulmonary circulation and its influence on gas exchange

C. A. Sieff Associate Professor in Pediatrics, Dana Farber Cancer Institute, Boston, Massachusetts, USA.
22.2.1 Stem cells and haemotopoiesis

J. Sieper Head of Rheumatology, Department of Medicine, University Hospital Benjamin Franklin, Berlin, Germany.
18.6 Spondyloarthritides and related arthritides

Leslie Silberstein Professor, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA.
22.5.9 Haemolytic anaemias - congenital and acquired

R. Sinclair Senior Lecturer, Department of Dermatology, University of Melbourne, St Vincent's Hospital, Fitzroy, Victoria, Australia.
23.1 Diseases of the skin

Joseph Sinning Yale School of Medicine, New Haven, Connecticut, USA.
22.4.1 Leucocytes in health and disease

Thira Sirisanthana Professor of Medicine and Director, Research Institute for Health Sciences, Chiang Mai University, Thailand.
7.11.18 Anthrax. 7.12.6 Infection due to Penicillium marneffei

J. G. P. Sissons Professor of Medicine, University of Cambridge and Honorary Consultant Physician, Addenbrooke's Hospital, Cambridge, UK.
7.10.2 Herpesviruses (excluding Epstein-Barr virus)

M. B. Skirrow Honorary Emeritus Consultant Microbiologist, Public Health Laboratory, Gloucester Royal Hospital, UK.
7.11.7 Enterobacteria, campylobacter, and miscellaneous food-poisoning bacteria

Geoffrey L. Smith Professor of Virology and Wellcome Trust Principal Research Fellow, The Wright-Fleming Institute, Faculty of Medicine, Imperial College of Science, Technology and Medicine, St Mary's Campus, London, UK.
7.10.4 Poxviruses

P. H. Smith Department of Urology, St James' University Hospital, Leeds, UK.
20.15 Tumours of the urinary tract

R. Smith Consultant Physician, Nuffield Orthopaedic Centre, Oxford, UK.
19.1 Disorders of the skeleton

E. L. Snyder Professor of Laboratory Medicine, Yale University School of Medicine, New Haven, Connecticut, USA.
22.8.1 Blood transfusion

R. L. Souhami Director of Clinical Research, Cancer Research UK and Emeritus Professor of Medicine, University College London, London, UK.
6.6 Cancer: clinical features and management

C. W. N. Spearman Senior Specialist and Co-Head of Liver Clinic, Groote Schuur Hospital, Cape Town, South Africa.
14.21.6 Hepatic granulomas

C. A. Speed Honorary Consultant Rheumatologist, Addenbrooke's Hospital, Cambridge, UK.
19.5 Avascular necrosis and related topics

G. P. Spickett Consultant Clinical Immunologist, Regional Department of Immunology, Royal Victoria Infirmary, Newcastle upon Tyne, UK.
17.11.8 Pulmonary haemorrhagic disorders. 17.11.9 Eosinophilic pneumonia. 17.11.11 Extrinsic allergic alveolitis. 17.11.19 Drug-induced lung disease

S. G. Spiro Professor of Respiratory Medicine and Medical Director, Medicine, University College London Hospitals NHS Trust, Middlesex Hospital, London, UK.

17.14.1 Lung cancer. 17.14.2 Pulmonary metastases

Jerry L. Spivak Professor of Medicine and Oncology, Johns Hopkins School of Medicine, Baltimore, Maryland, USA.
22.3.9 Idiopathic myelofibrosis

A. Spurgeon Senior Lecturer, Institute of Occupational Health, University of Birmingham, UK.
8.4.1 Occupational and environmental health and safety

Paul D. Stein Director of Research, St Joseph Mercy-Oakland, Pontiac, Michigan, USA.
15.15.3.1 Deep venous thrombosis and pulmonary embolism

Tom Stevens Consultant Psychiatrist, St Thomas' Hospital and Maudsley NHS Trust, London, UK.
26.3 Neuropsychiatric disorders

J. C. Stevenson Reader and Consultant Physician, Endocrinology and Metabolic Medicine, Faculty of Medicine, Imperial College London, UK.
13.20 Benefits and risks of hormone replacement therapy

P. M. Stewart Professor of Medicine, University of Birmingham and Consultant Physician, Queen Elizabeth Hospital, Birmingham, UK.
12.7.1 Disorders of the adrenal cortex

August Stich Consultant in Tropical Medicine, Medical Mission Institute, Unit of Tropical Medicine and Epidemic Control, Wurzburg, Germany.
7.13.10 Human African trypanosomiasis

John H. Stone Associate Professor of Medicine, Johns Hopkins University, Baltimore, Maryland, USA.
18.10.7 Polymyositis and dermatomyositis

J. R. Stradling Consultant Physician and Professor of Respiratory Medicine, Churchill Hospital, Oxford, UK.
17.1.1 The upper respiratory tract. 17.8.1 Upper airways obstruction. 17.8.2 Sleep-related disorders of breathing

Frank J. Strobl Director, Scientific Affairs, Therakos Inc., Exton, Pennsylvania, USA.
22.5.9 Haemolytic anaemias - congenital and acquired

M. A. Stroud Senior Lecturer in Medicine, Southampton University Hospitals Trust, UK.
8.5.1 Environmental extremes - heat. 8.5.2 Environmental extremes - cold

Michael Strupp Associate Professor of Neurology, Department of Neurology, Klinikum Grosshadern, University of Munich, Germany.
24.12.1 Eye movements and balance

P. H. Sugden Professor of Cellular Biochemistry, Imperial College of Science, Technology and Medicine, London, UK.
15.1.3.1 Physical considerations: biochemistry and cellular physiology of heart muscle

Daniel P. Sulmasy Sisters of Charity Chair in Ethics, St Vincent's Manhattan and New York Medical College, New York, USA.
2.3 Medical ethics

J. A. Summerfield Professor of Experimental Medicine, Faculty of Medicine, Imperial College London, UK.
14.19.1 Congenital disorders of the liver, biliary tract, and pancreas . 14.19.2 Diseases of the gallbladder and biliary tree

Pravan Suntharasamai Emeritus Professor of Tropical Medicine, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand.
7.14.9 Gnathostomiasis

J. Swales* Professor of Medicine, University of Leicester, UK.
15.16.1.3 Essential hypertension

P. Sweny Consultant Nephrologist, Royal Free Hospital, London, UK.
20.6.3 Renal transplantation

D. Swirsky Consultant Haematologist, Leeds General Infirmary, UK.
22.4.4 The spleen and its disorders

I. C. Talbot Professor of Histopathology, St Mark's Hospital for Colorectal Disorders, London, UK.
14.15 Tumours of the gastrointestinal tract

D. Tarin Director, UCSD Cancer Center, University of California at San Diego, La Jolla, USA.
6.4 Tumour metastasis

D. Taylor-Robinson Emeritus Professor of Genitourinary Microbiology and Medicine, Division of Medicine, Imperial College of Science, Technology and Medicine, St Mary's Hospital, London, UK.
7.11.40 Chlamydial infections including lymphogranuloma venerum . 7.11.41 Mycoplasmas

P. J. Teddy Consultant Neurosurgeon/Clinical Director, Department of Neurological Surgery, Radcliffe Infirmary, Oxford, UK.
24.14.3 Intracranial abscess

H. J. Testa Professor and Consultant (retired), Royal Infirmary, Manchester, UK.
15.3.4 Nuclear techniques

R. V. Thakker May Professor of Medicine, Nuffield Department of Medicine, University of Oxford, UK.
12.6 Parathyroid disorders and diseases altering calcium metabolism

David G. T. Thomas Professor of Neurological Surgery, National Hospital for Neurology and Neurosurgery, London, UK.
24.13.18.2 Traumatic injuries of the head

D. L. Thomas Associate Professor of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA.
7.10.20 Hepatitis C virus

P. K. Thomas Emeritus Professor of Neurology, Royal Free Hospital School of Medicine and Institute of Neurology, London, UK.
24.6.1 Inherited disorders. 24.13.15.00 Disorders of cranial nerves. 24.19 Diseases of the peripheral nerves

D. G. Thompson Professor of Gastroenterology, University of Manchester, UK.

14.1.1.1 Structure and function of the gut. 14.12 Functional bowel disorders and irritable bowel syndrome

R. P. H. Thompson Consultant Physician, St Thomas' Hospital, London, UK.
8.5.8 Podoconiosis. 14.19.3 Jaundice

S. A. Thorne Royal Brompton and Harefield NHS Trust, London, UK.
15.13 Congenital heart disease in adolescents and adults

Ph. Thulliez Head, Laboratoire de la Toxoplasmose, Institut de Puericulture, Paris, France.
7.13.4 Toxoplasmosis

Tran Tin Hien Vice Director, Centre for Tropical Diseases (Cho Quan Hospital), Ho Chi Minh City, Vietnam.
7.11.1 Diphtheria

J. A. Todd Professor of Medical Genetics, University of Cambridge, UK.
12.11.2 The genetics of diabetes

C. Tomson Consultant Nephrologist, Southmead Hospital, Bristol, UK.
20.12 Urinary tract infection

Keith Tones Professor of Health Education (Emeritus), Leeds Metropolitan University, UK.
3.5 Health promotion

P. A. Tookey Lecturer, Centre for Epidemiology and Biostatistics, Institute of Child Health, London, UK.
7.10.12 Rubella

P. P. Toskes Professor of Medicine, Division of Gastroenterology, Hepatology, and Nutrition, Department of Medicine, University of Florida College of Medicine, Gainsville, USA.
14.9.2 Small bowel bacterial overgrowth. 14.18.3.2 Chronic pancreatitis

Thomas A. Traill Professor of Medicine, Johns Hopkins Hospital, Baltimore, Maryland, USA.
15.11.1 Cardiac myxoma. 15.11.2 Other tumours of the heart. 15.12 Cardiac involvement in genetic disease

David F. Treacher Consultant Physician in Intensive Care, St Thomas' Hospital, Guy's and St Thomas' NHS Trust, London, UK.
16.2 The circulation and circulatory support of the critically ill

A. S. Truswell Emeritus Professor of Human Nutrition, University of Sydney, New South Wales, Australia.
10.1 Diseases of overnourished societies and the need for dietary change

D. M. Turnbull Professor of Neurology, The Medical School, University of Newcastle upon Tyne, UK.
24.22.5 Mitochondrial encephalomyopathies

H. E. Turner Consultant Physician, Radcliffe Infirmary, Oxford, UK.
12.12 Hormonal manifestations of non-endocrine disease

A. Neil Turner Professor of Nephrology, Royal Infirmary, Edinburgh, UK.
20.7.8 Infection-associated nephropathies. 20.7.9 Malignancy-associated renal disease

Robert Twycross Emeritus Clinical Reader in Palliative Medicine, Oxford University, Sir Michael Sobell House, Churchill Hospital, Oxford, UK.
31 Palliative care

F. E. Udwadia Emeritus Professor of Medicine, Grant Medical College and J. J. Hospital, Bombay; Consultant Physician and Director-in-charge of ICU, Breach Candy Hospital; Consultant Physician, Parsee General hospital, Bombay, India.
7.11.20 Tetanus

S. Richard Underwood Professor of Cardiac Imaging, Imperial College of Science, Technology and Medicine, National Heart and Lung Institute, and Royal Brompton Hospital, London, UK.
15.3.5 Cardiovascular magnetic resonance and computed X-ray tomography

Robert J. Unwin Professor of Nephrology and Physiology, Centre for Nephrology, The Middlesex Hospital, London, UK.
20.13 Urinary stones, nephrocalcinosis, and renal tubular acidosis

V. Urquidi Assistant Professor, University of California San Diego Cancer Center and Department of Pathology, La Jolla, California, USA.
6.4 Tumour metastasis

J. A. Vale Director, National Poisons Information Service and West Midlands Poisons Unit, City Hospital, Birmingham, UK.
8.1 Poisoning by drugs and chemicals

P. Vallance Professor of Clinical Pharmacology and Therapeutics, Centre for Clinical Pharmacology, University College London, UK.
15.1.1.2 Vascular endothelium, its physiology and pathophysiology

J. van Gijn Professor and Chairman, Department of Neurology, University Medical Centre, Utrecht, The Netherlands.
24.13.7 Stroke: cerebrovascular disease

Sirivan Vanijanonta Emeritus Professor of Tropical Medicine, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand.
7.16.3 Lung flukes (paragonimiasis)

Patrick J. W. Venables Professor and Honorary Consultant, Kennedy Institute Division, Imperial College London, UK.
18.10.6 Sjogren's syndrome

B. J. Vennervald Senior Research Scientist, Danish Bilharziasis Laboratory, Charlottenlund, Denmark.
7.16.1 Schistosomiasis

C. M. Verity Consultant Paediatric Neurologist and Associate Lecturer, Faculty of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK.
24.21 Developmental abnormalities of the nervous system

M. P. Vessey Emeritus Professor of Public Health, Unit of Health Care Epidemiology, Department of Public Health, Oxford University, UK.
13.19 Benefits and risks of oral contraceptives

R. Viner Consultant in Adolescent Medicine and Endocrinology, University College London Hospitals and Great Ormond Street Hospital, UK.
29 Adolescent medicine

Peter D. Wagner Professor of Medicine and Bioengineering, University of California, San Diego, USA.
17.1.2 Structure and function of the airways and alveoli

Ann E. Wakefield* Professor of Paediatric Infectious Diseases, Department of Paediatrics, Institute of Molecular Medicine, University of Oxford, UK.
7.12.5 Pneumocystis carinii

D. H. Walker The Carmage and Martha Walls Distinguished Chair in Tropical Diseases, Professor and Chairman, Department of Pathology, and Director, WHO Collaborating Center for Tropical Diseases, Galveston, Texas, USA.
7.11.36 Rickettsial diseases including ehrlichiosis

J. A. Walker-Smith Emeritus Professor of Paediatric Gastroenterology, Royal Free and University College Medical School, London, UK.
14.14 Congenital abnormalities of the gastrointestinal tract

Mark J. Walport Professor of Medicine and Head, Division of Medicine, Faculty of Medicine, Imperial College London, Hammersmith Hospital, London, UK.
5.4 Complement

Julian R. F. Walters Reader in Gastroenterology, Imperial College of Science, Technology and Medicine, Hammersmith Campus, London, UK.
14.2.4 Investigation of gastrointestinal function. 14.9.1 Differential diagnosis and investigation of malabsorption

Gary S. Wand Professor of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
26.7.1 Alcohol and drug dependence

Ronald J. A. Wanders Professor of Inborn Errors and Metabolism and Deputy Head of the Laboratory for Metabolic Diseases, Academic Medical Centre, Amsterdam, The Netherlands.
11.9 Peroxisomal diseases

B. Ward Anaesthetic Registrar, Coventry School of Anaesthetics, UK.
16.6.1 Sedation and analgesia in the critically ill

T. E. Warkentin Professor, Department of Pathology and Molcular Medicine and Department of Medicine, McMaster University, Hamilton, Ontario, Canada.
22.6.5 Acquired coagulation disorders

D. A. Warrell Professor of Tropical Medicine and Infectious Diseases and Head, Nuffield Department of Clinical Medicine, University of Oxford, UK.
7.8 Travel and expedition medicine. 7.10.9 Rhabdoviruses: rabies and rabies-related viruses. 7.10.10 Colorado tick fever and other arthropod-borne reoviruses. 7.11.28 Rat bite fevers. 7.11.30 Other borrelia infections. 7.11.32 Non-venereal treponematoses: yaws, endemic syphilis (bejel), and pinta. 7.13.2 Malaria. 7.13.5 Cryptosporidium and cryptosporidiosis. 7.18 Pentostomiasis (porocephalosis). 8.2 Injuries, envenoming, poisoning, and allergic reactions caused by animals. 24.14.1 Bacterial meningitis. 24.14.2 Viral infections of the central nervous system. 24.22.6 Tropical pyomyositis (tropical myositis). 33 Emergency Medicine

M. J. Warrell Clinical Virologist, Centre for Tropical Medicine, John Radcliffe Hospital, Oxford, UK.
7.10.9 Rhabdoviruses: rabies and rabies-related viruses. 7.10.10 Colorado tick fever and other arthropod-borne reoviruses

Paul Warwicker Consultant Nephrologist, Renal Unit, Lister Hospital, Stevenage, Hertfordshire, UK.
20.10.6 Haemolytic uraemic syndrome

J. A. H. Wass Professor of Endocrinology and Consultant Physician, Radcliffe Infirmary, Oxford, UK.
12.12 Hormonal manifestations of non-endocrine disease

Laurence Watkins Consultant Neurosurgeon and Senior Lecturer, Institute of Neurology, London, UK.
24.13.18.2 Traumatic injuries of the head

George Watt Department of Medicine, AFRIMS, Bangkok, Thailand.
7.11.31 Leptospirosis. 7.11.37 Scrub typhus

Richard W. E. Watts Visiting Professor and Honorary Consultant Physician, Imperial College School of Medicine, Hammersmith Hospital, London, UK.
11.1 The inborn errors of metabolism: general. 11.4 Disorders of purine and pyrimidine metabolism. 11.10 Disorders of oxalate metabolism

D. J. Weatherall Regius Professor of Medicine Emeritus, University of Oxford, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, UK.
2.2 Scientific medicine and the art of healing. 22.1 Introduction. 22.5.2 Anaemia: pathophysiology, classification, and clinical features. 22.5.3 Anaemia as a world health problem. 22.5.5 Normochromic, normocytic anaemia. 22.5.7 Disorders of the synthesis or function of haemoglobin. 22.7 The blood in systemic disease

D. K. H. Webb Consultant Paediatric Haematologist, Great Ormond Street Hospital for Children, London, UK.
22.4.7 Histiocytoses

Kathryn E. Webert Clinical Scholar, Hematology and Fellow in Transfusion Medicine, Canadian Blood Services, McMaster University, Hamilton, Ontario, Canada.
22.6.3 Disorders of platelet number and function

A. D. B. Webster Consultant Immunologist, Department of Immunology, Royal Free Hospital, London, UK.
5.6 Immunodeficiency

Anthony P. Weetman Professor of Medicine and Dean, University of Sheffield Medical School, UK.
12.4 The thyroid gland and disorders of thyroid function. 12.5 Thyroid cancer

R. A. Weiss Professor, University College London, UK.
7.10.21 HIV and AIDS. 7.10.24 Viruses and cancer

Peter L. Weissberg BHF Professor of Cardiovascular Medicine, University of Cambridge, UK.
15.1.1.1 Introduction. 15.1.1.3 Vascular smooth muscle cells. 15.4.2.1 The pathophysiology of acute coronary syndromes

Peter F. Weller Professor of Medicine, Harvard Medical School; Chief of Allergy and Inflammation and Co-Chief, Infectious Diseases Division, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA.
22.4.6 Eosinophilia

A. K. Wells Consultant Respiratory Physician, Royal Brompton Hospital, London, UK.
17.11.4 The lungs and rheumatological diseases

Simon Wessely Professor of Epidemiological Psychiatry, Guy's, King's and St Thomas' School of Medicine and Institute of Psychiatry, London, UK.
26.6.2 Psychological treatment in medical practice

Gilbert C. White, II John C. Parker Professor of Medicine and Pharmacology and Director, Center for Thrombosis and Hemostasis, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA.
22.6.1 The biology of haemostasis and thrombosis. 22.6.2 Evaluation of the patient with a bleeding diathesis

Joseph White SPHTM at TUMC, New Orleans, Louisiana, USA.
3.7.1 The cost of health care in Western countries

H. C. Whittle Visiting Professor, London School of Hygiene and Tropical Medicine and Deputy Director, MRC Laboratories, Banjul, The Gambia.
7.10.6 Measles

D. E. L. Wilcken Professor Emeritus of Medicine and Head, Cardiovascular Research Laboratory, University of New South Wales and Prince of Wales Hospital, Sydney, Australia.
15.1.3.2 Clinical physiology of the normal heart

James S. Wiley Professor and Head of Haematology, Nepean Hospital, Penrith, New South Wales, Australia.
22.5.8 Anaemias resulting from defective red cell maturation

P. J. Wilkinson Consultant Medical Microbiologist, University Hospital, Queen's Medical Centre, Nottingham, UK.
7.11.34 Listeriosis

R. G. Will Professor of Clinical Neurology, Western General Hospital, Edinburgh, UK.
24.13.9 Human prion disease

C. B. Williams Consultant Physician in Endoscopy, St Mark's Hospital for Colorectal Disorders, UK.
14.2.1 Colonoscopy and flexible sigmoidoscopy. 14.15 Tumours of the gastrointestinal tract

D. J. Williams Senior Lecturer/Honorary Consultant in Obstetric Medicine, Division of Paediatrics, Obstetrics and Gynaecology, Imperial College of Science, Technology and Medicine, Chelsea and Westminster Hospital, London, UK.
13.1 Physiological changes of normal pregnancy. 13.2 Nutrition in pregnancy. 13.3 Medical management of normal pregnancy

Gareth Williams Professor of Medicine, Department of Medicine, Clinical Sciences Centre, University Hospital Aintree, Liverpool, UK.
12.11.1 Diabetes

J. D. Williams Professor of Nephrology and Consultant Physician, Institute of Nephrology, University of Wales College of Medicine, Cardiff, UK.
20.1 Structure and function of the kidney

Paul F. Williams Consultant Nephrologist, The Ipswich Hospital NHS Trust, UK.
20.6.2 The treatment of endstage renal disease by peritoneal dialysis

Robert Wilson Consultant Physician and Reader, Royal Brompton Hospital and National Heart and Lung Institute, Imperial College of Science, Technology and Medicine, London, UK.
17.3.3 Microbiological methods in the diagnosis of respiratory infections

C. G. Winearls Consultant Nephrologist, Oxford Kidney Unit, Churchill Hospital, Oxford, UK.
20.5.1 Chronic renal failure

F. Wojnarowska Professor of Dermatology and Consultant Dermatologist, Oxford Radcliffe Hospital, Oxford, UK.
13.13 The skin in pregnancy

R. Wolman Consultant in Rheumatology and Sports Medicine, Royal National Orthopaedic Hospital, Stanmore, Middlesex, UK.
28 Sports and exercise

Kathryn J. Wood Professor of Immunology, Nuffield Department of Surgery, University of Oxford, UK.
5.7 Principles of transplantation immunology

Nicholas Wood Professor of Clinical Neurology, Institute of Neurology, London, UK.
24.6.2 Neurogenetics. 24.13.12 Ataxic disorders

Trevor Woodage Clinical Investigator, Celera Genomics, Rockville, Maryland, USA.
4.1 The genomic basis of medicine

H. F. Woods Professor of Medicine, University of Sheffield, UK.
11.11 Disturbances of acid-base homeostasis

Gary P. Wormser Vice Chairman, Department of Medicine, and Chief, Division of Infectious Diseases, New York Medical College, Valhalla, New York, USA.
7.11.29 Lyme borreliosis

D. J. M. Wright Emeritus Reader in Medical Microbiology, Cell and Molecular Biology Section, Imperial College School of Medicine, London, UK.
7.11.33 Syphilis

V. M. Wright Consultant Paediatric Surgeon, Barts and The London NHS Trust, London, UK.
14.14 Congenital abnormalities of the gastrointestinal tract

F. C. W. Wu Senior Lecturer (Endocrinology), Royal Infirmary and University of Manchester, UK.
12.8.2 Disorders of male reproduction

Andrew H. Wyllie Professor and Head of Department of Pathology, University of Cambridge, UK.
4.6 Apoptosis in health and disease

M. A. S. Yasuda Professor, Department of Infectious and Parasitic Diseases, University of São Paulo Medical School, Brazil.
7.12.4 Paracoccidioidomycosis

Newman M. Yeilding Assistant Professor, University of Pennsylvania, Philadelphia, USA.
4.3 Molecular cell biology

Jenny Yiend Postdoctoral Research Assistant, MRC Cognition and Brain Science Unit, Cambridge, UK.
26.5.1 Grief, stress, and post-traumatic stress disorder

V. Zaman Professor, Department of Microbiology, The Aga Khan University, Karachi, Pakistan.
7.13.7 Sarcocystosis. 7.14.6 Other gut nematodes. 7.14.7 Toxocariasis and visceral larva migrans

* It is with regret that we report the death of Professor John Calam during the preparation of this edition of the textbook.
* It is with regret that we report the death of Professor Alvan R. Feinstein during the preparation of this edition of the textbook.
* It is with regret that we report the death of Professor C. D. Marsden.
* It is with regret that we report the death of Professor J. S. Prichard.
* It is with regret that we report the death of Professor J. Swales during the preparation of this edition of the textbook.
* It is with regret that we report the death of Professor Ann E. Wakefield during the preparation of this edition of the textbook.

Color Plate

Plates for Section 5

Chapter 5.4 Complement

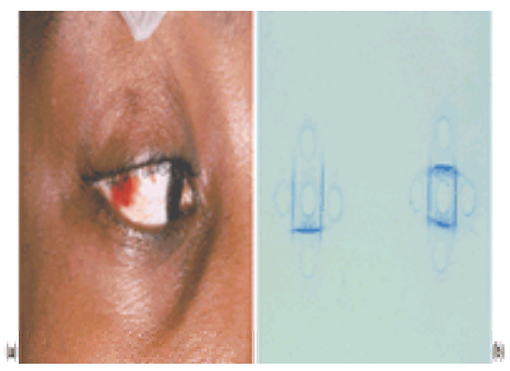

Plate 1 Patient with hereditary deficiency of C6 who presented with meningococcal septicaemia. (a) A subconjunctival haemorrhage. (b) The deficiency of C6. Serum from the patient was placed in the central well of an agarose-coated plate. In each of the outer wells was placed antiserum to, respectively, C5, C6, C7, and C8. The antibody and antigen were allowed to diffuse in the gel and where the antibody encountered its antigen a precipitate formed, which was stained blue. No precipitate formed between the anti-C6 antibody and the patient's serum, indicating the presence of C6 deficiency.

Plates for Section 6

Chapter 6.2 The nature and development of cancer

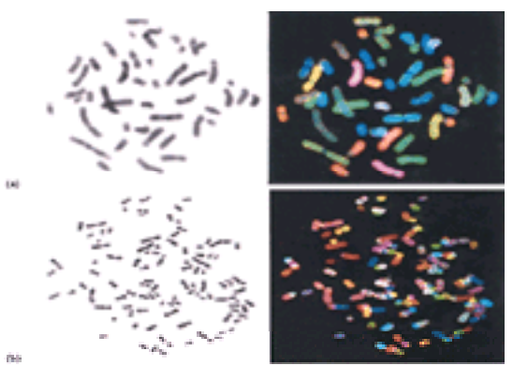

Plate 1 In a spectral karyotype (SKY), each normal chromosome stains homogeneously with a single distinct colour, making translocations evident by the presence of more than one colour in a single chromosome. (a) Forty-five chromosomes are visible in this karyotype (left) from a patient with Turner's syndrome. Despite the loss of the X chromosome, the spectral karyotype (right) clearly shows the homogeneous chromosomal staining pattern typical of normal chromosomes. (b) In contrast, SKY analysis of a metaphase spread prepared from a breast cancer cell line displays both numerical and structural chromosomal aberrations. (By courtesy of Dr Bassem Haddad, Department of Oncology, Georgetown University Medical Center.)

Plates for Section 7

Chapter 7.10.2 Herpesviruses (excluding Epstein-Barr virus)

Plate 1 Primary herpetic gingivostomatitis.

Plate 2 Primary HSV-2 of the buttocks.

Plate 3 (a) Severe chickenpox also involving the lungs. (b) Details of the rash.

Plate 4 Herpes zoster affecting the ophthalmic division of the Vth nerve.

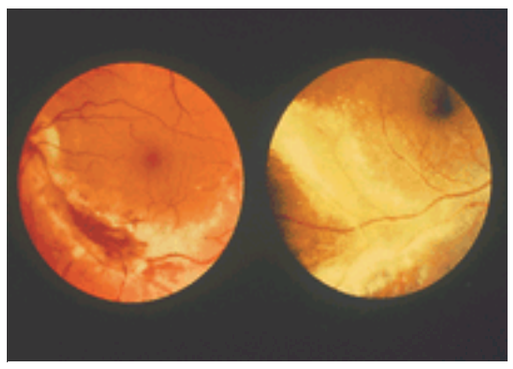

Plate 5 Human cytomegalovirus.

Plate 6 Kaposi's sarcoma affecting the palate and producing symmetrical skin lesions in association with HIV infection.

Chapter 7.10.4 Poxviruses

Plate 1 Ethiopian patient, in 1968, showing classical centrifugal distribution of lesions. (Copyright D.A. Warrell.)

Plate 2 Tanapox lesion on the leg of a Kenyan patient (by courtesy of the late P.E.C. Manson-Bahr).

Chapter 7.10.6 Measles

Plate 1 Measles rash on the legs of an English teenager. (Copyright D.A. Warrell.)

Plate 2 Measles rash (African).

Plate 3 Stomatitis with Herpes simplex ulcers in an African child with severe measles. (Copyright D.A. Warrell.)

Plate 4 Herpes simplex keratoconjunctivitis in an African child with severe measles. (Copyright D.A. Warrell.)

Plate 5 Measles rash (African).

Chapter 7.10.6.1 Nipah and Hendra viruses

Plate 1 Pteropid fruit bat (flying fox), the natural reservoir of Nipah, Hendra, and Menangle paramyxoviruses and Australian bat lyssavirus. (From the painting by John Gould.)

Chapter 7.10.9 Rhabdoviruses

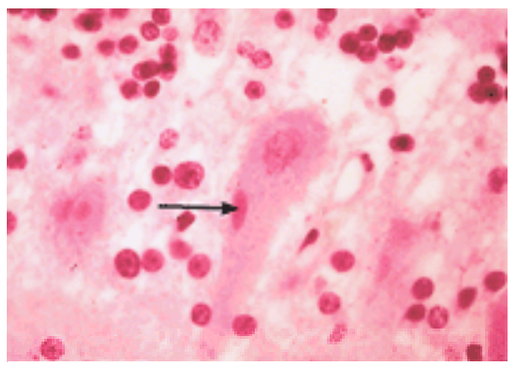

Plate 1 Street rabies virus in human cerebellar Purkinje cells as seen with the light microscope. Several Negri bodies can be seen (one is arrowed). (By courtesy of the Armed Forces Institute of Pathology 73–12330.)

Chapter 7.10.18 Parvovirus B19

Plate 1 'Slapped cheek' rash of erythema infectiosum: note circumoral pallor. (By courtesy of Dr Ken Mutton.)

Chapter 7.11.2 Streptococci and enterococci

Plate 1 Bilateral facial erysipelas. (Copyright S. Eykyn.)

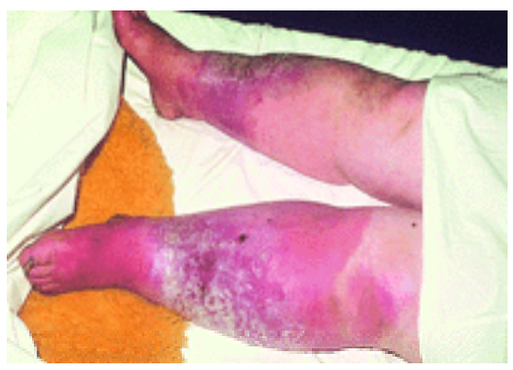

Plate 2 Cellulitis. (Copyright S. Eykyn.)

Plate 3 Streptococcus pyogenes bacteraemia 3 days after a skin graft. (Copyright S. Eykyn.)

Plate 4 Peeling of the skin of the soles of the feet in a patient with *Streptococcus pyogenes* pericarditis. (Copyright S. Eykyn.)

Chapter 7.11.5 Meningococcal infections

Plate 1 Massive skin haemorrhage on the extremities of a 4-year-old girl with fulminant meningococcal septicaemia. The infection was caused by *Neisseria meningitidis* group B. The left leg had to be amputated below the knee. She needed extensive skin transplantation and several fingers had to be amputated.

Plate 2 Macular lesions on the legs, some with a central haemorrhagic spot in a 17-year-old girl with mild meningococcaemia caused by *Neisseria meningitidis* group C. She recovered completely after 5 days treatment with benzylpenicillin.

Plate 3 Macular and haemorrhagic lesions on the legs of a 21-year-old man with mild meningococcaemia caused by *Neisseria meningitidis* group B. He recovered completely after 5 days of penicillin treatment.

Plate 4 The 'glass test' used to differentiate haemorrhagic skin lesions from viral or drug rash in an infant with meningococcal meningitis caused by *Neisseria meningitidis* group B. There was complete recovery after 5 days treatment with benzylpenicillin.

Chapter 7.11.6 *Neisseria gonorrhoeae*

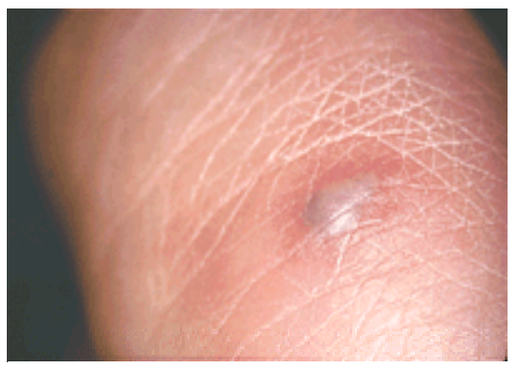

Plate 1 Disseminated gonococcal infection, haemorrhagic vesiculopustule.

Plate 2 Disseminated gonococcal infection: healing lesions with desquamation and deposition of haemosiderin.

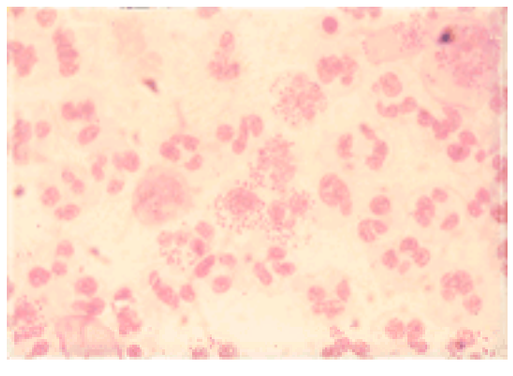

Plate 3 Gram-stained urethral discharge showing Gram-negative intracellular diplococci.

Chapter 7.11.8 Typhoid and paratyphoid fevers

Plate 1 Typhoid rash in a Melanesian child – sparse, purpuric (non-blanching) macules.

Chapter 7.11.17 Yersinia, Pasteurella, and Francisella

Plate 1 Hands in a case of ulcero-(cutano)-glandular tularaemia (by courtesy of A. Berglund, Fallund, Sweden).

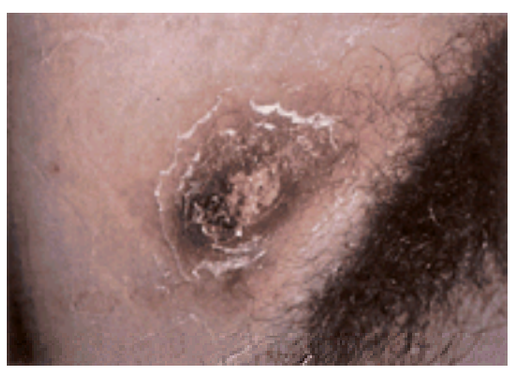

Plate 2 Inguinal lymphadenopathy in ulceroglandular tularaemia (by courtesy of A. Berglund, Fallund, Sweden).

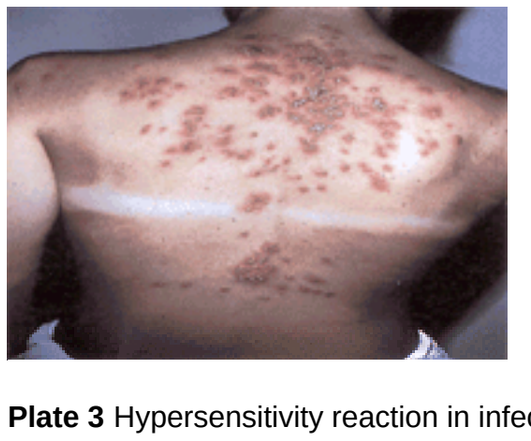

Plate 3 Hypersensitivity reaction in infection with *Francisella tularensis* subsp. *holarctica* (type B) in Scandinavia (by courtesy of A. Berglund, Fallund, Sweden).

Plate 4 Oral tularaemia in a case from northern Sweden (by courtesy of A. Berglund, Fallund, Sweden).

Chapter 7.11.18 Anthrax

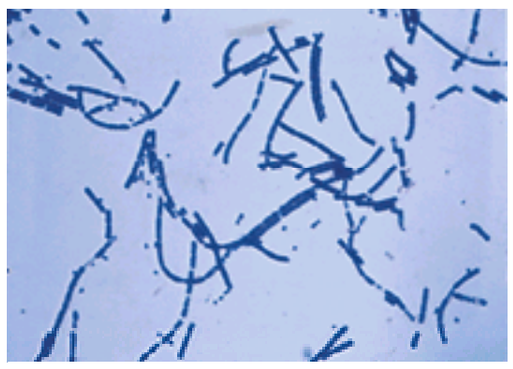

Plate 1 Large Grampositive bacilli in chains are typical of *Bacillus anthracis*. An individual bacillus is 3 to 5 µ m long and 1 to 1.25 µ m wide with a flattened end.

Plate 2 Cutaneous anthrax lesion on the forearm on day 10 showing an ulcer with a depressed black eschar.

Plate 3 Oropharyngeal anthrax on day 9 showing a pseudomembrane covering an ulcer.

Chapter 7.11.20 Tetanus

Plate 1 Facies in tetanus.

Plate 2 Opisthotonos in severe tetanus during seizures.

Plate 3 Brazilian patient with local tetanus confined to muscles innervated by the left VIIth cranial nerve and with trismus, showing the wound causing the infection. (By courtesy of Dr Pedro Pardal, Belém, Brazil.)

Plate 4 Characteristic facies in neonatal tetanus.

Chapter 7.11.24 Leprosy (Hansen's disease)

Plate 1 BT leprosy. This Ethiopian woman has several hypopigmented patches. Testing for anaesthesia will confirm the diagnosis of BT leprosy.

Plate 2 Advanced nodular lepromatous leprosy. This Indian patient presented with ulcerating nodules all over his body.

Plate 3 Reversal (Type 1) reaction. This Ethiopian woman had a postpartum reaction presenting with numerous erythematous raised lesions 8 weeks after delivery.

Plate 4 Severe reversal (Type 1) reaction. This Indian woman has erythematous, oedematous, and desquamating reactional lesions.

Plate 5 Peripheral nerve thickening in leprosy. This young man had marked thickening of his great auricular nerve.

Plate 6 Nerve damage in leprosy. This patient with BT leprosy has damage to the ulnar and median nerves on both sides. This has resulted in hands which are wasted, clawed, and lack finger and thumb opposition.

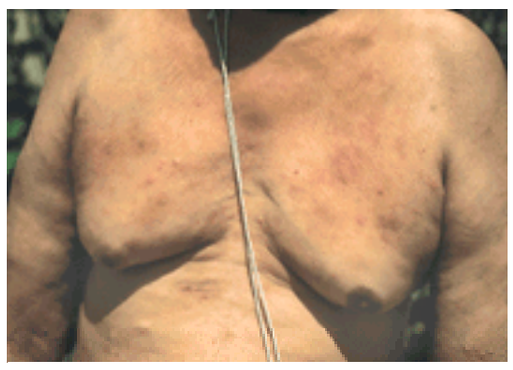

Plate 7 Complications of lepromatous leprosy. Gynaecomastia is visible in this man, secondary to testicular involvement in lepromatous leprosy. Multiple nodules are present, many dark brown, due to clofazimine pigmentation. He also has new erythematous lesions of ENL.

Chapter 7.11.25 Buruli ulcer: *Mycobacterium ulcerans*; infection

Plate 1 Buruli ulcer on the left deltoid area in a 12-year-old Congolese boy who had received a hypodermic injection at this site 3 months previously. Note central necrotic slough in the base of the ulcer, and undermined edges.

Chapter 7.11.29 Lyme borreliosis

Plate 1 Adult female (right) and nymphal (left) – ticks of the *Ixodes scapularis* species.

Plate 2 Erythema migrans rashes from patients who were culture positive for borrelia. (a) A rash with typical central clearing appearance. (b) A rash with more homogenous appearance.

Chapter 7.11.30 Other borrelia infections

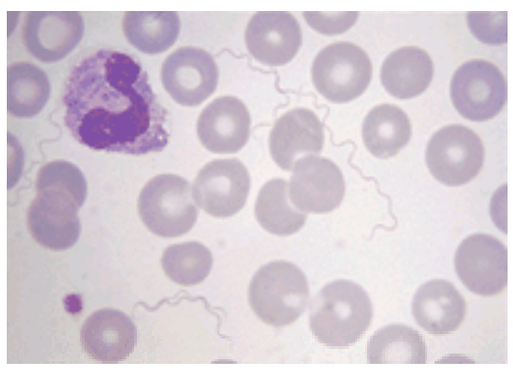

Plate 1 *Borrelia recurrentis* spirochaetes in a Giemsastained thin blood film from a patient with louse-borne relapsing fever. (Copyright D.A. Warrell.)

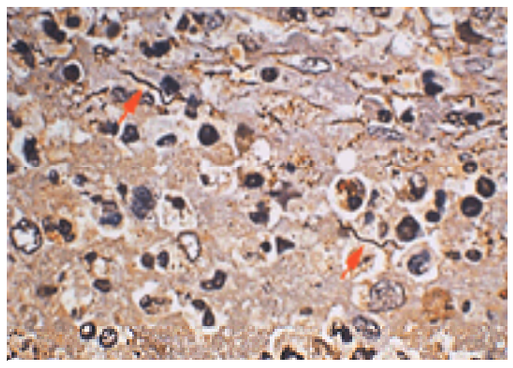

Plate 2 Spleen in louseborne relapsing fever. Warthin Starry stain showing *Borrelia recurrentis* (arrows). (By courtesy of Dr Ken Fleming.)

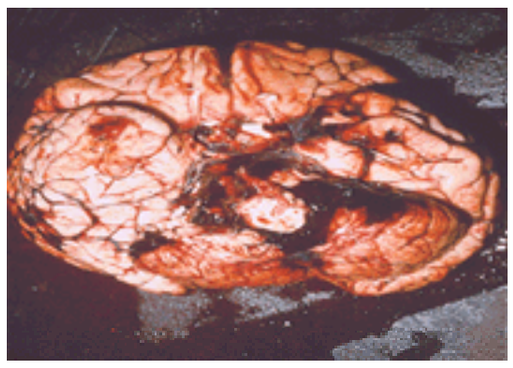

Plate 3 Cerebral haemorrhage in a patient with louse-borne relapsing fever. (Copyright D.A. Warrell.)

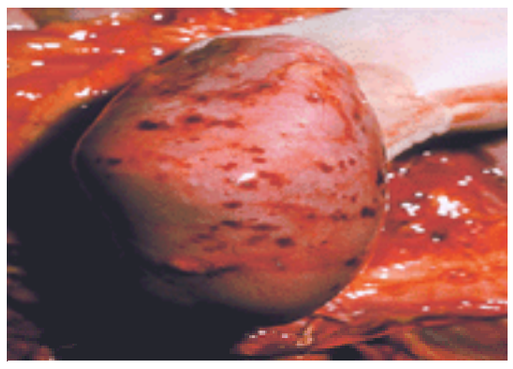

Plate 4 Petechial haemorrhages on the surface of the kidney in a victim of louse-borne relapsing fever. (Copyright D.A. Warrell.)

Plate 5 Ethiopian patient with severe louse-borne relapsing fever. Note emaciation and petechial rash. (Copyright D.A. Warrell.)

Plate 6 Subconjunctival haemorrhages in louse-borne relapsing fever. (Copyright D.A. Warrell.)

Chapter 7.11.31 Leptospirosis

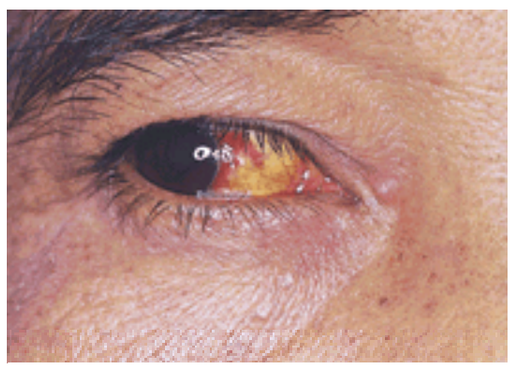

Plate 1 Jaundice, haemorrhage, and conjunctival suffusion in acute leptospirosis.

Chapter 7.11.36 Rickettsial diseases including ehrlichiosis

Plate 1 Boutonneuse fever (South African tick typhus). Eschar with lymphangitic lines spreading towards the femoral lymph nodes in a patient who had visited the Kruger National Park, South Africa, 7 days earlier. (Copyright D.A. Warrell.)

Plates 2 and 3 Boutonneuse fever (South African tick typhus) in a British traveller. (Copyright E. Dunbar.)

Chapter 7.11.37 Scrub typhus

Plate 1 Typical eschars (a, b) and one less typical (c) on the distal foreskin. Lesions in locations such as these can be easily missed during a cursory examination of a febrile patient presenting to a busy outpatient clinic.

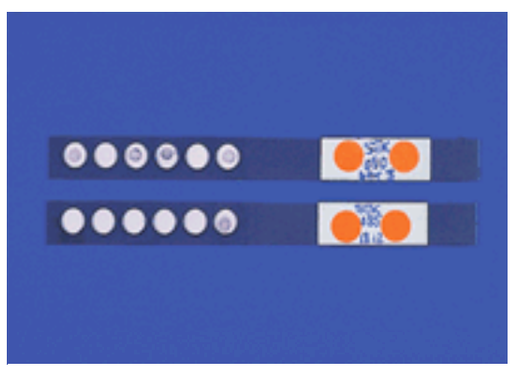

Plate 2 Diagnosis by rapid immunoblot dipstick. The test strip above indicates active scrub typhus, with clearly visible staining within several circles. The test strip below was read as non-reactive, because only the reagent control (last dot on the right) contains staining.

Chapter 7.11.39 Bartonellosis

P><**Plate 1** Miliary haemangioma-like of 'verruga peruana'.

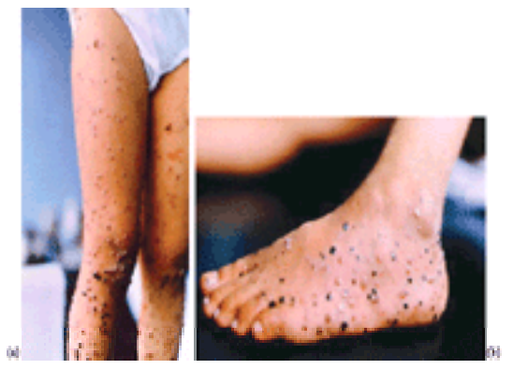

Plate 2 (a, b) Nodular lesions of 'verruga peruana'.

Chapter 7.11.40 Chlamydial infections including lymphogranuloma venerum

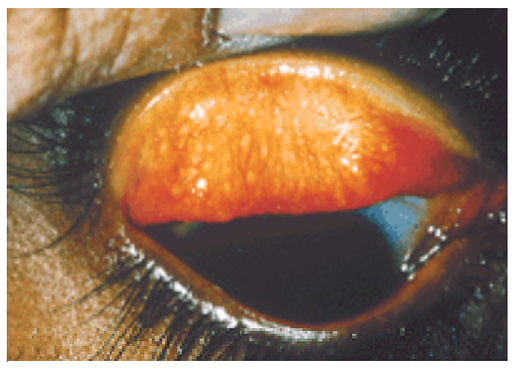

Plate 1 Everted upper eyelid showing follicular trachoma (TF).

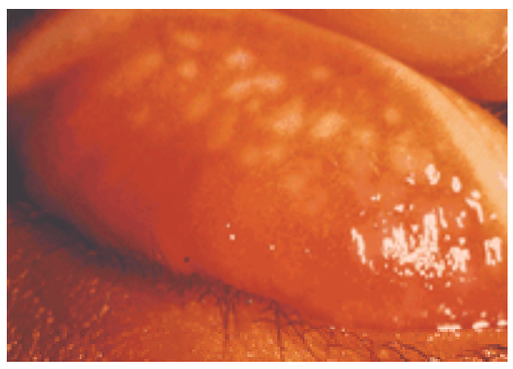

Plate 2 Everted upper eyelid showing intense inflammatory trachoma (TI).

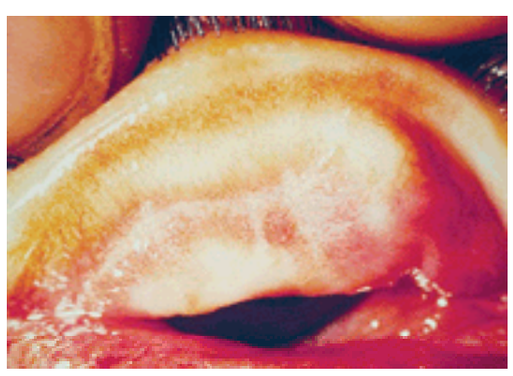

Plate 3 Extensive neovascularization of the cornea (pannus) due to trachoma.

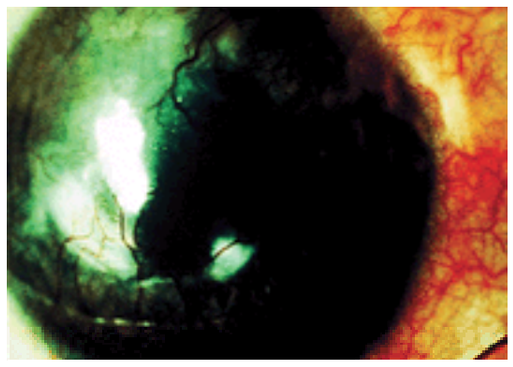

Plate 4 Everted upper eyelid showing trachomatous scarring (TS).

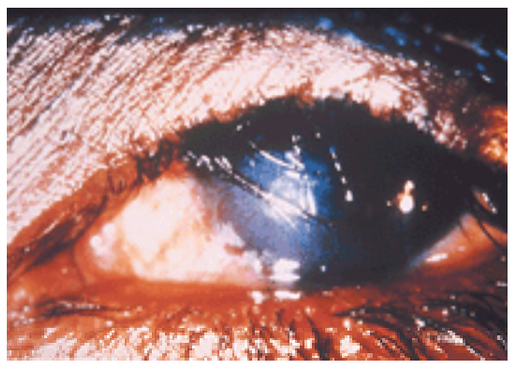

Plate 5 Trachomatous trichiasis (TT).

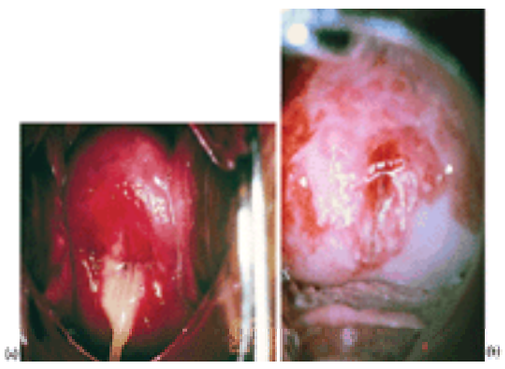

Plate 6 (a) Mucopurulent cervicitis; (b) follicular cervicitis.

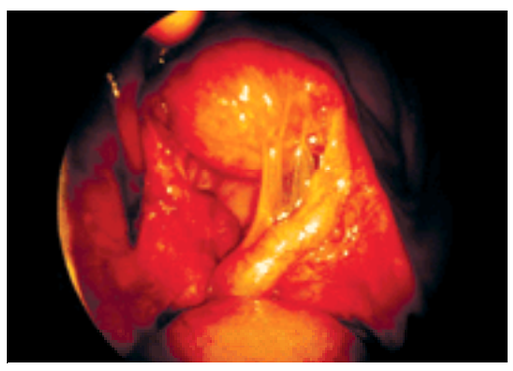

Plate 7 Laparoscopic view of inflamed fallopian tube due to *C. trachomatis*. (By courtesy of P. Greenhouse.)

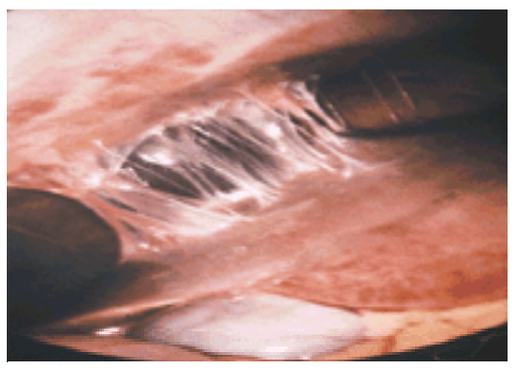

Plate 8 Adhesions in perihepatitis (Curtis Fitz-Hugh syndrome) due to *C. trachomatis*. (By courtesy of P. Greenhouse.)

Plate 9 Mucopurulent neonatal conjunctival discharge due to *C. trachomatis*.

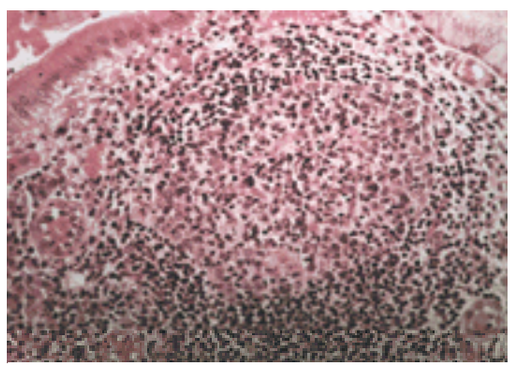

Plate 10 Germinal centre formation in lymphoid follicle of cervicitis due to *C. trachomatis*.

Chapter 7.12.1 Fungal infections

Plate 1 Palmar scaling due to *Trichophyton rubrum*.

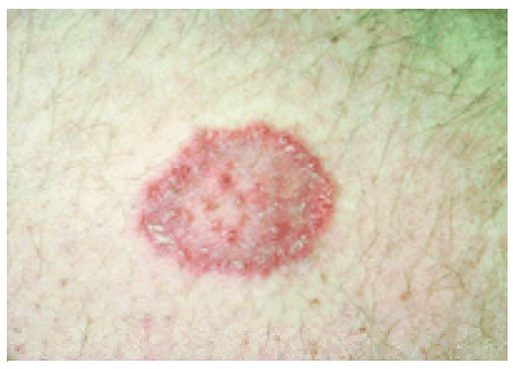

Plate 2 Tinea corporis due to *Microsporum gypseum*.

Plate 3 Oral candidosis in a patient with chronic mucocutaneous candidosis.

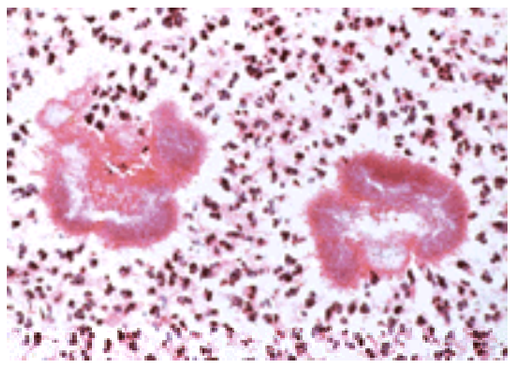

Plate 4 Grains in abscess in actinomycetoma (*Nocardia brasiliensis*) (H & E).

Plate 5 A mycetoma caused by *Madurella grisea*.

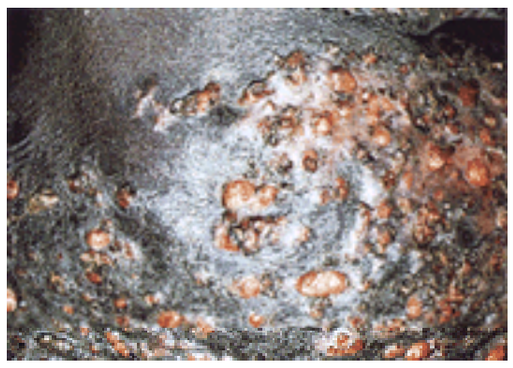

Plate 6 *Nocardia brasiliensis* actinomycetoma draining sinus.

Plate 7 Lobo's disease in a Brazilian man. (Copyright D.A. Warrell.)

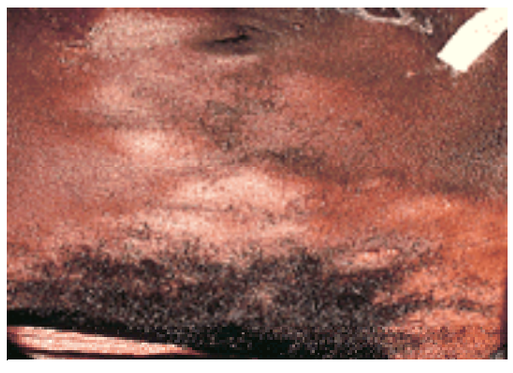

Plate 8 Nodular subcutaneous lesions of African histoplasmosis in a Nigerian man. (Copyright D.A. Warrell.)

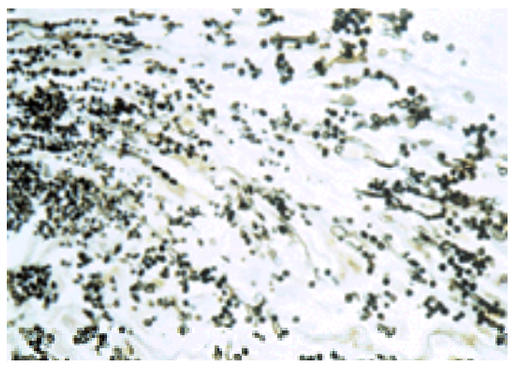

Plate 9 Candidosis disseminated to skin (methenamine silver ×516).

Chapter 7.13.1 Amoebic infections

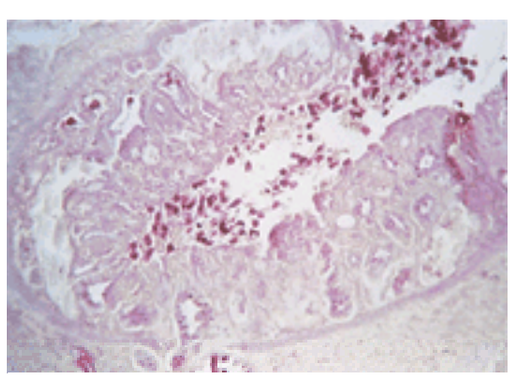

Plate 1 Amoebic colitis. Crypt abscess. PAS stains amoebae red. (Copyright Viqar Zaman.)

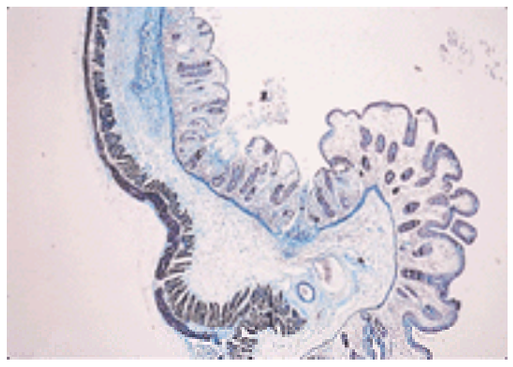

Plate 2 Amoebic colitis. Superficial ulcer breaching the muscularis mucosae. (Copyright Viqar Zaman.)

Plate 3 'Anchovy sauce' pus drained from an amoebic liver abscess. (Copyright Viqar Zaman.)

Plate 4 Sixteen-year-old Peruvian boy with a chronic facial lesion that had been present for 3 years and intracranial space-occupying lesions cuased by *Balamuthia mandrillaris* (a). Perforating lesion of the palate (b). (Copyright D.A. Warrell.)

Chapter 7.13.2 Malaria

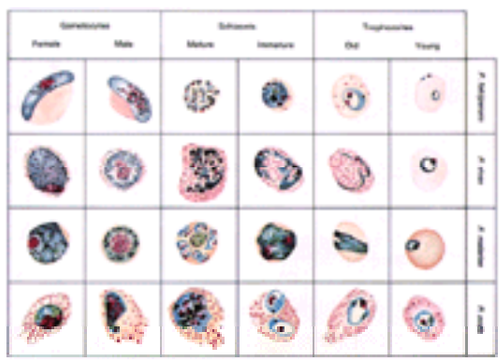

Plate 1 Malaria parasites developing in erythrocytes. (By courtesy of The Wellcome Trust.)

Plate 2 Section of frontal cortex from a Vietnamese patient who died of cerebral malaria, showing sequestration of parasitized red blood corpuscles in blood vessels (N=neurone, V=vessel). (By courtesy of Dr Gareth Turner, Oxford.)

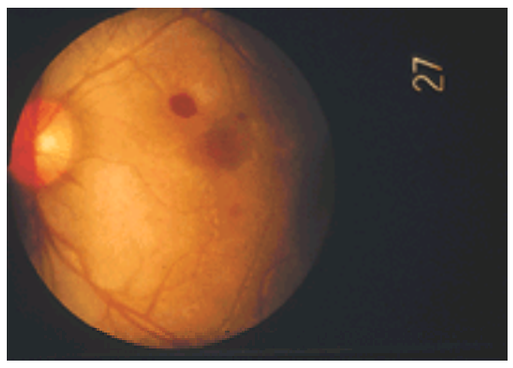

Plate 3 Retinal haemorrhages close to the macula in a Thai patient with cerebral malaria. (Copyright D.A. Warrell.)

Plate 4 Profound anaemia (haemoglobin 1.2 g/dl) in a Kenyan child with *P. falciparum* parasitaemia. (Copyright D.A. Warrell.)

Plate 5 Cerebral malaria. Spontaneous systemic bleeding in a Thai patient with disseminated intravascular coagulation. (Copyright D.A. Warrell.)

Plate 6 Deep jaundice in a Vietnamese man with severe falciparum malaria. (Copyright D.A. Warrell.)

Plate 7 Intravascular haemolysis in a Karen patient with glucose 6-phosphate dehydrogenase deficiency in whom treatment with an oxidant drug resulted in haemoglobinuria and anaemia (normal hand in comparison). (Copyright D.A. Warrell.)

Chapter 7.13.3 Babesia

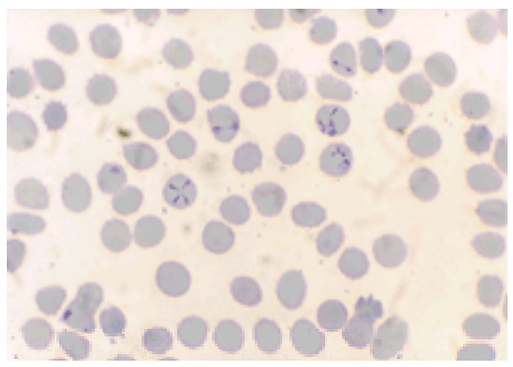

Plate 1 *Babesia divergens* infection in a 29-year-old French man, infected in Normandy. He had been splenectomized 4 months previously for idiopathic thrombocytopenia. Parasitaemia reached 30 per cent. He was successfully treated with exchange transfusion, clindamycin, and quinine. (Copyright P. Brasseur.)

Chapter 7.13.5 Cryptosporidium and cryptosporidiosis

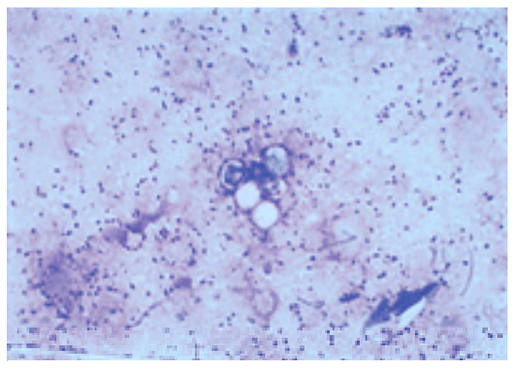

Plate 1 Modified Giemsa-stained faecal smear showing oocysts of *C. parvum*, examined with × 100 oil-immersion objective lens. The uniformity of size (4.5-5 µ m) but variability of staining of oocysts can be seen. The eosinophilic nuclei and basophilic bodies of the sporozoites can be clearly seen within the oocysts that have taken up the stain.

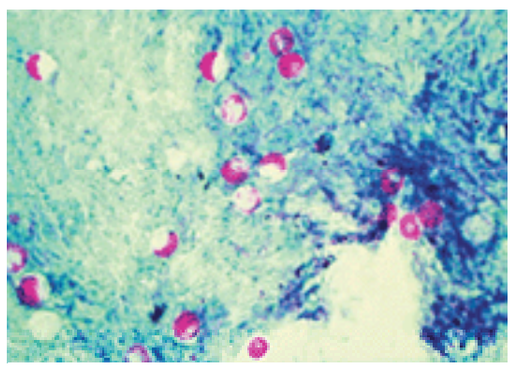

Plate 2 Modified Ziehl-Neelsenstained faecal smear showing oocysts of *C. parvum* examined with × 100 oil-immersion objective lens. The uniformity of size (4.5-5 µ m) but variability of staining of oocysts can be seen.

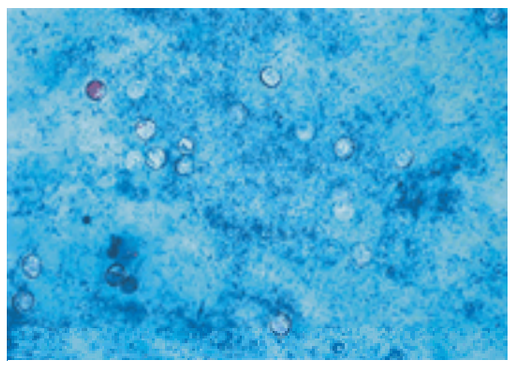

Plate 3 Modified Ziehl-Neelsenstained faecal smear showing oocysts of *C. parvum*. The uniformity of size (4.5-5 µ m) is apparent but the oocysts in this preparation show a definite increase in refractility and marked failure to take up the stain (identity confirmed by immunofluorescence and electron microscopy).

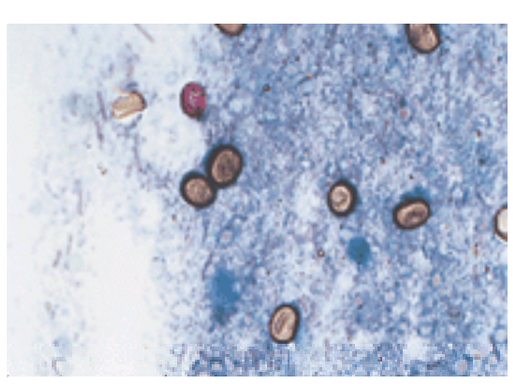

Plate 4 Modified Ziehl-Neelsenstained faecal smear showing oocyst-like bodies (mushroom spores) examined with × 100 oilimmersion objective lens (from specimen submitted to Reference Unit for identification).

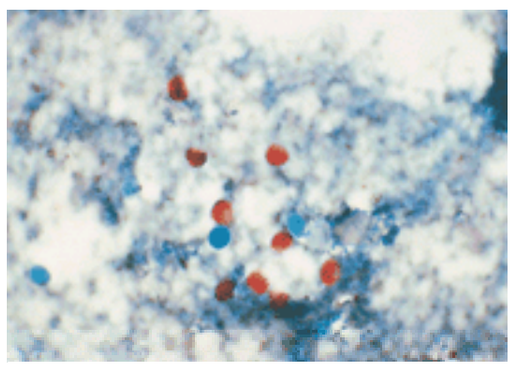

Plate 5 Modified Ziehl-Neelsenstained faecal smear showing oocyst-like bodies (mould spores) examined with × 100 oilimmersion objective lens. The spores are uniform in size but a little smaller (4.0 µ m) than oocysts of *C. parvum*. They are generally more uniform in their acid-fast staining (identity con- firmed by mycological culture and electron microscopy).

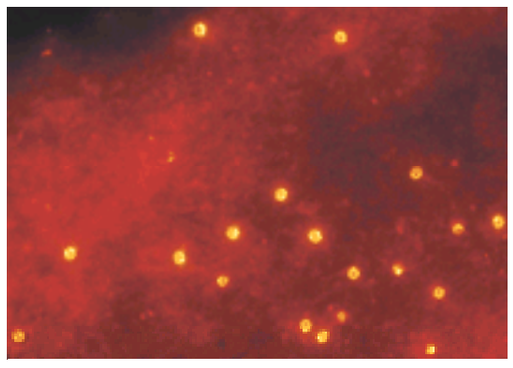

Plate 6 Phenol-auramine/carbol fuchsin-stained faecal smear showing oocysts of *C. parvum*, examined with × 20 dry objective lens (screening magnification) on a fluorescence microscope.

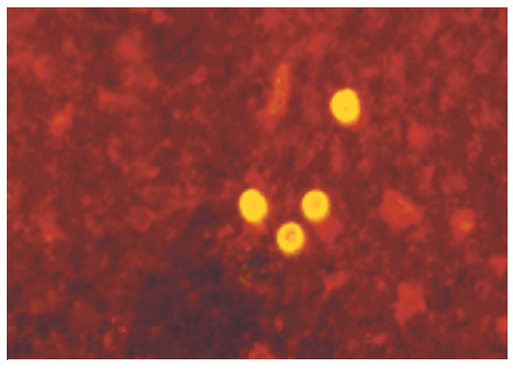

Plate 7 Phenol-auramine/carbol fuchsin-stained faecal smear showing oocysts of *C. parvum*, examined with × 100 oil-immersion objective lens on a fluorescence microscope.

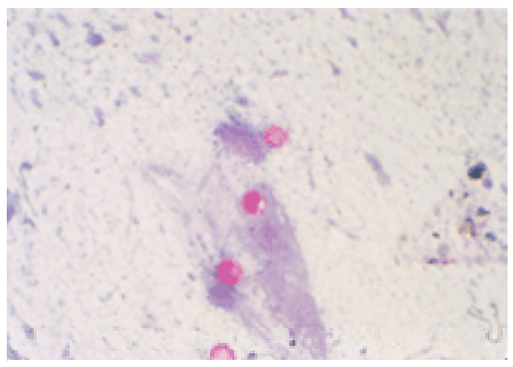

Plate 8 Fluorescent dye-tagged monoclonal antibody-stained faecal smear showing oocysts of *C. parvum*, examined with × 50 oil-immersion objective lens (screening magnification) on a fluorescence microscope. The suture or associated surface cleft or fold, through which the sporozoites are released, can be seen.

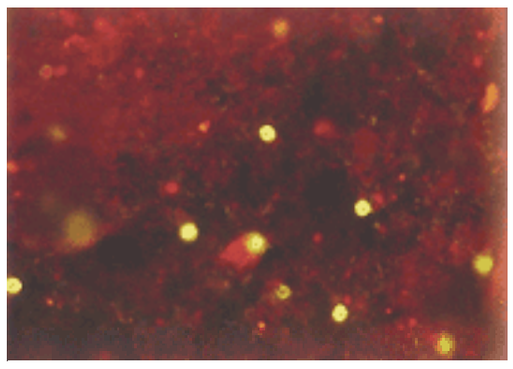

Plate 9 Modified Ziehl-Neelsen-stained sputum smear from an AIDS patient with respiratory involvement (examined with × 100 oil-immersion objective lens). The *C. parvum* bodies present may include endogenous (tissue) stages attached to exfoliated cells. For this reason, oocyst wall-specific indirect immunofluorescence may show a poor reaction. There may also be less uniformity of size and differences in the staining appearance of the internal structures.

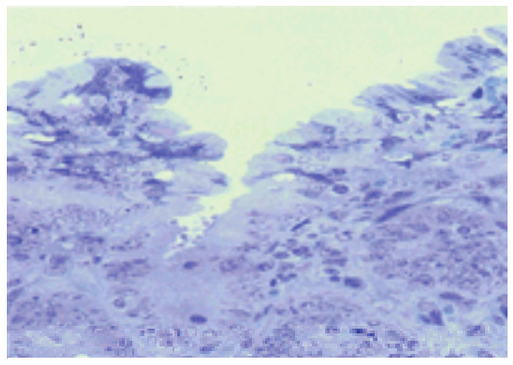

Plate 10 Toluidine blue-stained semithin section of human rectal biopsy tissue of an AIDS patient with cryptosporidiosis. The apparent pseudo-external location of the parasite can be seen, the true location being intracellular but extracytoplasmic. Plates for this Chapter were kindly provided from photographs by A. Curry and D.P. Casemore.

Chapter 7.13.6 Cyclospora

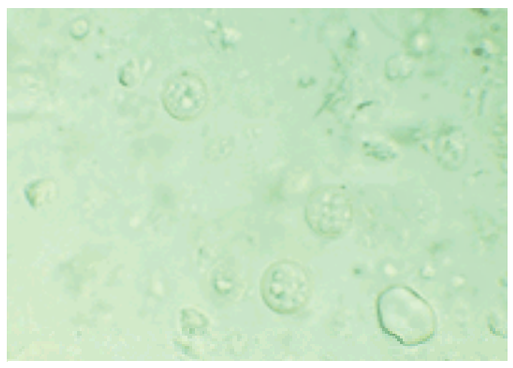

Plate 1 Unstained wet preparation of human faecal material showing oocysts of *Cyclospora* sp., examined with × 100 water-immersion objective lens by phasecontrast microscopy. The uniformity of size (8-10 µ m) and the morular (mulberry) internal structure of the oocysts can be seen.

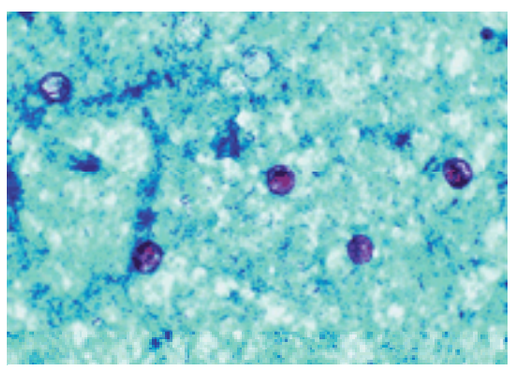

Plate 2 Modified Ziehl-Neelsen-stained faecal smear showing oocysts of *Cyclospora* sp. examined with × 50 oil-immersion objective lens. The uniformity of size (8-10 µm) but variability of staining of the oocysts can be seen. Apart from the greater size, the oocysts can be distinguished from those of *Cryptosporidium parvum* by the different pattern of acid-fast staining. Unstained oocysts within the smear sometimes show the morular structure apparent in wet preparations.

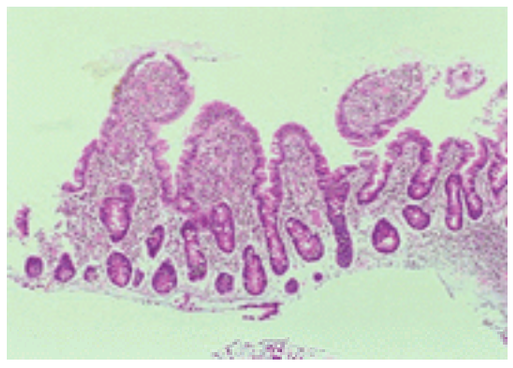

Plate 3 Jejunal biopsy from a patient with cyclosporiasis showing jejunitis with blunting of villi (low power H & E stain). (By courtesy of Dr Sebastian Lucus, London.)

Chapter 7.13.10 Human African trypanosomiasis

Plate 1 Adult tsetse fly (*Glossina morsitans*).

Plate 2 Trypanosomal chancre on the shank of a missionary returning from the Congo.

Plate 3 Patient with late-stage trypanosomiasis.

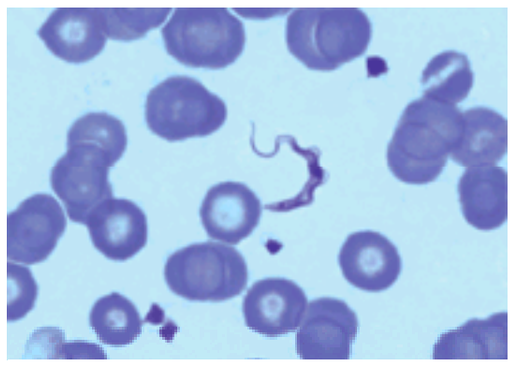

Plate 4 Trypanosomes in thin human blood film (Giemsa stain, × 1000 magnification).

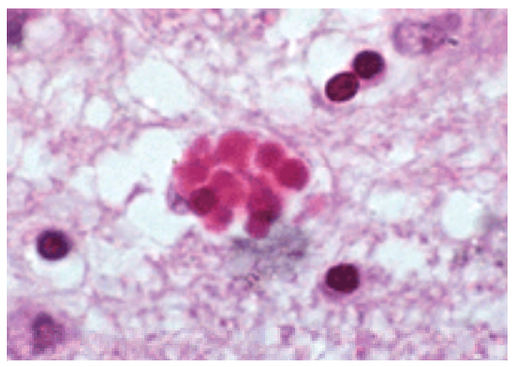

Plate 5 Morular cell of Mott in a histological brain section of a stage II HAT patient (H & E stain, × 1000 magnification).

Chapter 7.13.11 Chagas' disease

Plate 1 Adult female triatomine bug (*Panstrongylus megistus*), with a single egg shown adjacent to the tip of the abdomen. (By courtesy of Dr T.V. Barrett.)

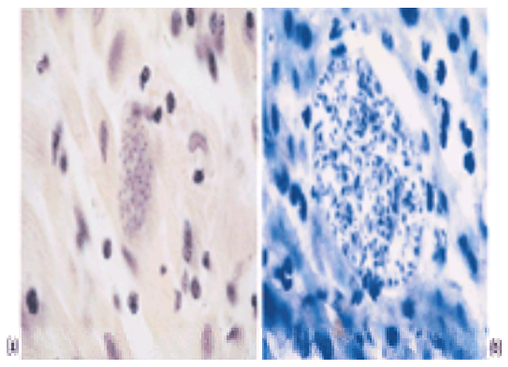

Plate 2 (a) Pseudocyst of *Trypanosoma cruzi* in heart muscle. (By courtesy of J.E. Williams.) (b) Pseudocyst of *Trypanosoma cruzi* in umbilical cord, from a congenital case of Chagas' disease. (By courtesy of Dr Hipolito de Almeida.)

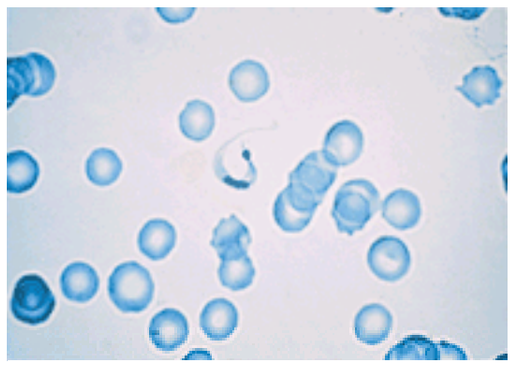

Plate 3 *Trypanosoma cruzi* C-shaped trypomastigote in blood, note large posterior kinetoplast.

Plate 4 Romaña's sign in acute Chagas' disease.

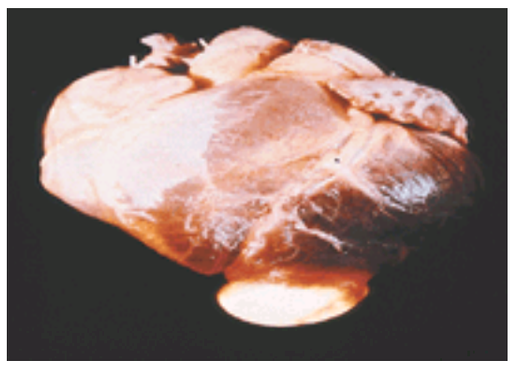

Plate 5 Apical aneurysm of the left ventricle in chronic Chagas' disease. (By courtesy of Dr J.S. de Oliveira.)

Plate 6 Mural thrombus filling the right atrial appendage. (Copyright D.A. Warrell.)

Plate 7 Mega-oesophagus seen by radiography in chronic Chagas' disease. (By courtesy of Dr J.S. de Oliveira.)

Plate 8 Megacolon postmortem in chronic Chagas' disease. (By courtesy of Dr J.S. de Oliveira.)

Chapter 7.13.12 Leishmaniasis

Plate 1 Shallow ulcer with raised edge due to *L. brasiliensis* (copyright A.D.M. Bryceson).

Plate 2 Lupoid or recidivans leishmaniasis in a citizen of Baghdad. (By courtesy of Dr Ahmed.)

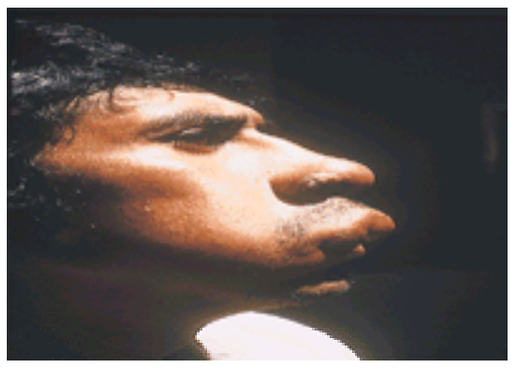

Plate 3 Swollen upper lip and nose due to mucosal leishmaniasis in Peru (copyright A.D.M. Bryceson).

Plate 4 Infiltration of lip and palate due to mucosal leishmaniasis in Peru (copyright A.D.M. Bryceson).

Chapter 7.13.13 Trichomoniasis

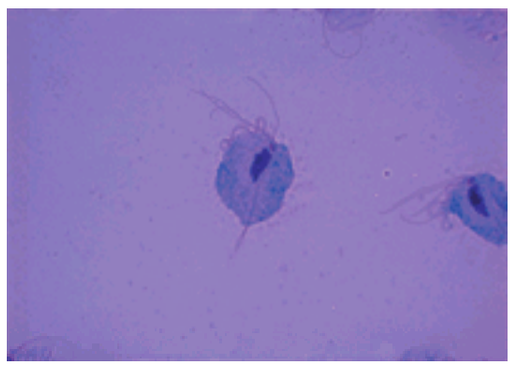

Plate 1 Trichomonads, Giemsa stain, in vaginal secretions. (Copyright J.P. Ackers.)

Chapter 7.14.1 Cutaneous filariasis

Plate 1 A 3-cm subcutaneous nodule.

Plate 2 Excoriated papular lesions of onchocerciasis with hyperpigmentation.

Plate 3 Lichenified skin lesions with atrophy.

Plate 4 Depigmented 'leopard skin'.

Plate 5 Migrating *Loa loa*.

Chapter 7.14.2 Lymphatic filariasis

Plate 1 Microfilaria of *Wuchereria bancrofti* in a blood film from a patient in Samoa. (By courtesy of the Wellcome Museum of Medical Science.)

Chapter 7.14.3 Guinea-worm disease: dracunculiasis

Plate 1 Blister at site of imminent emergence of the female worm. (By courtesy of the late P.E.C. Manson-Bahr.)

Plate 2 Emergent female worm being wound out on a stick. (Copyright D.A. Warrell.)

Plate 3 Guinea worm in the scrotum. (Copyright D.A. Warrell.)

Chapter 7.14.4 Strongyloidiasis, hookworm, and other gut strongyloid nematodes

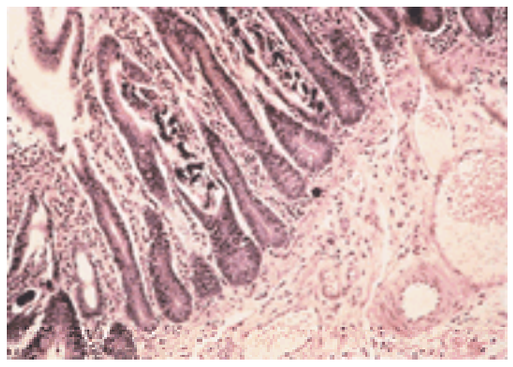

Plate 1 Strongyloides stercoralis in the intestinal mucosa. (Copyright Viqar Zaman.)

Plate 2 Larva currens rash on the back of a Nigerian patient resulting from autoinfection with *Strongyloides stercoralis*. (Copyright D.A. Warrell.)

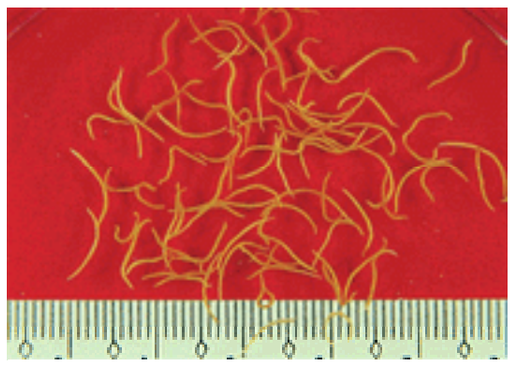

Plate 3 Adult *Ancylostoma duodenale* – scale in millimetres. (Copyright Viqar Zaman.)

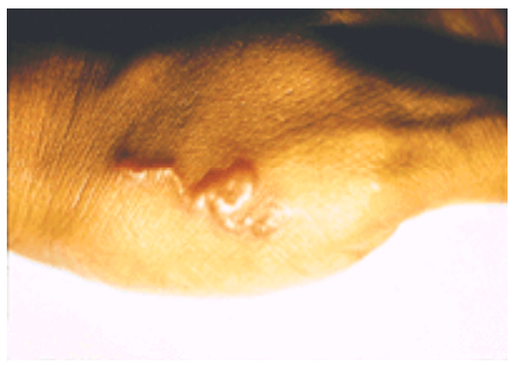

Plate 4 Cutaneous larva migrans of the hand in a Thai patient. (Copyright Sornchai Looareesuwan.)

Chapter 7.14.6 Other gut nematodes

Plate 1 Ascaris – scale in millimetres.

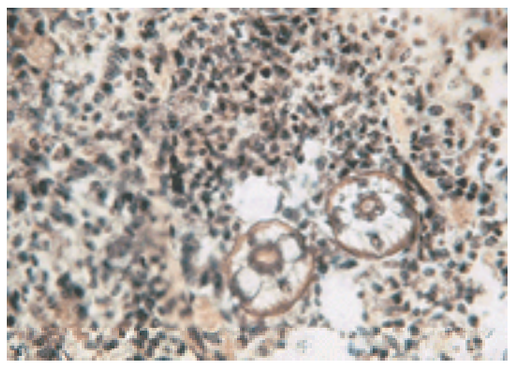

Plate 2 Ascaris in the lungs. (Copyright Viqar Zaman.)

Plate 3 Enterobius – scale in millimetres.

Chapter 7.14.8 Angiostrongyliasis

Plate 1 *Angiostrongylus cantonensis* under the conjunctiva in a Thai girl with a left facial nerve palsy. (Copyright D.A. Warrell.)

Chapter 7.15.1 Cystic hydatid disease (*Echinococcus granulosus*)

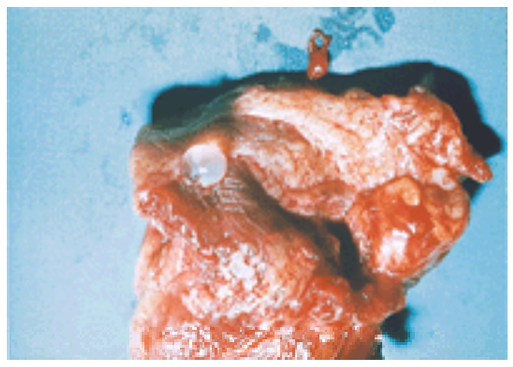

Plate 1 Hydatid cyst in muscles excised from around the femoral head (same case as shown in Fig, 3).

Chapter 7.16.1 Schistosomiasis

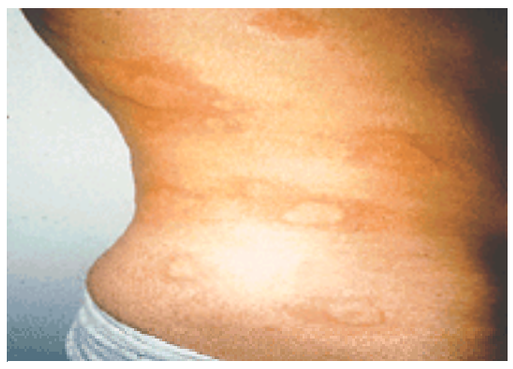

Plate 1 Giant urticarial rash in a patient with Katayama fever (*Schistosoma mansoni* infection). (Copyright R.N. Davidson.)

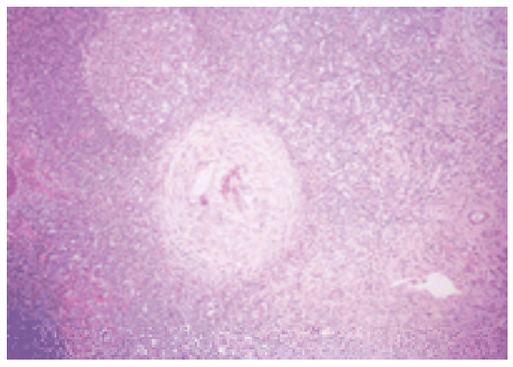

Plate 2 Schistosomal granuloma in the appendix. (Copyright Gareth Turner.)

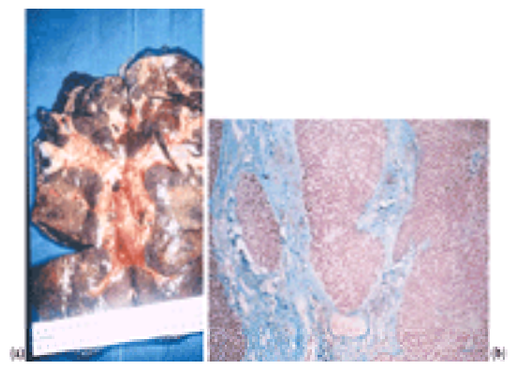

Plate 3 The liver in *Schistosoma mansoni* infection in South Africa. Clay pipestem fibrosis. (Copyright Gareth Turner.) (a) Macrosopic view. (b) Masson trichrome stain.

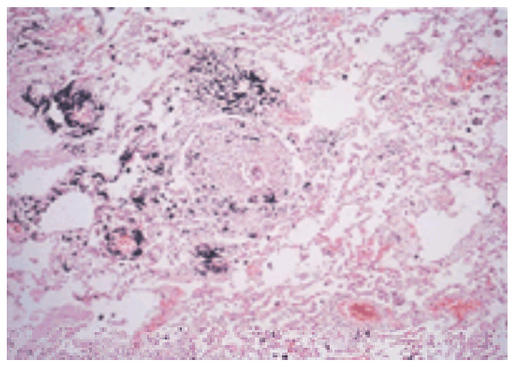

Plate 4 Schistosomal granuloma in the lung. (Copyright Gareth Turner.)

Chapter 7.16.3 Lung flukes (paragonimiasis)

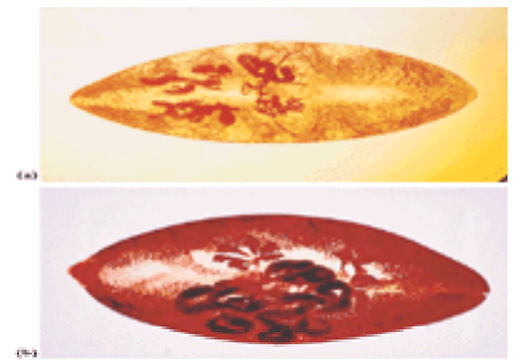

Plate 1 Adult lung fluke. (a) *Paragonimus heterotremus* (1.5 cm). (b) *P. westermani* (1.5 cm). (Copyright Sanan Yaemput.)

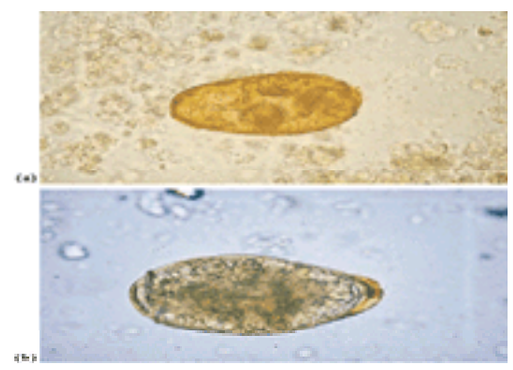

Plate 2 Ova of lung flukes. (a) *Paragonimus heterotremus*. (b) *P. westermani*. (Copyright Sanan Yaemput.)

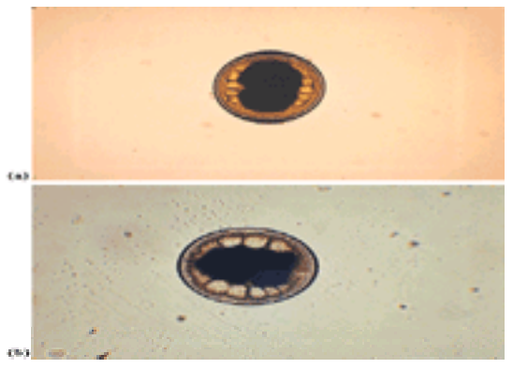

Plate 3 Metacercariae of lung fluke in crabs, the second intermediate host. (a) *Paragonimus heterotremus*. (b) *P. westermani*. (Copyright Sanan Yaemput.)

Plate 4 Freshwater crab *Larnaudia beusekomae* (*Tawaripotamon beusekomae*), the second intermediate host. (Copyright Sanan Yaemput.)

Chapter 7.17 Non-venomous arthropods

Plate 1 Bedbugs, *Cimex lectularius*.

Plate 2 Catflea, *Ctenocephalides felis* : a common cause of flea bites in humans.

Plate 3 Underside of hedgehog tick, *Ixodes hexagonus* to show sucking mouthparts (hypostome).

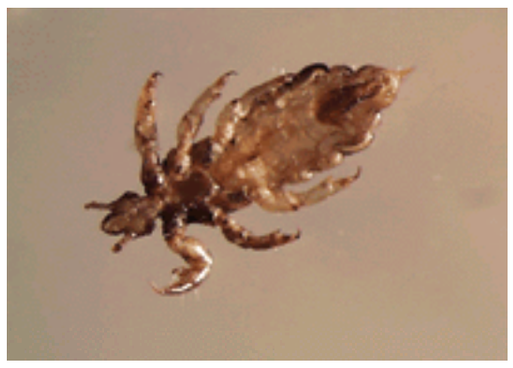

Plate 4 Louse, *Pediculus humanus* : head lice and body lice are morphologically similar.

Plate 5 An Asian carabid beetle, *Sciates sulcatus*, from a patient complaining of vaginal discharge: a rare example of genital cantharíasis.

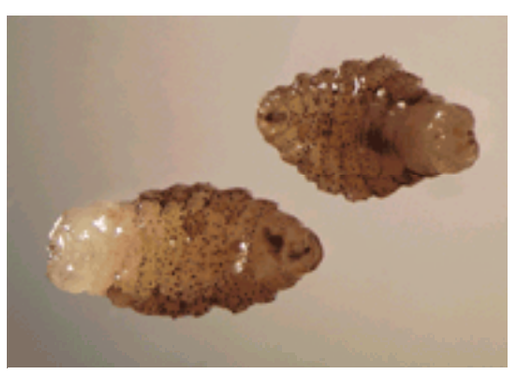

Plate 6 Larvae of African tumbu fly, *Cordylobia anthropophaga* : a common agent of dermal myiasis.

Plates for Section 8

Chapter 8.2 Injuries, envenoming, poisoning, and allergic reactions caused by animals

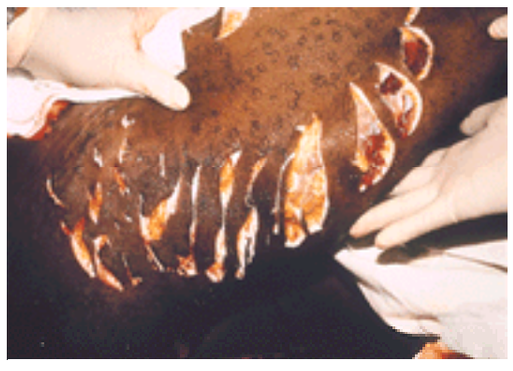

Plate 1 Shark attack: wounds inflicted on the thigh by a tiger shark (*Galeocerdo cuvier*), Madang, Papua New Guinea. (Copyright S. Allen.)

Plate 2 Farmer swallowed by a reticulated python (*Python reticulatus*), Palu, Sulawesi. (Copyright Excel Sawuwu.)

Plate 3 White-lipped pit viper from south-east Asia (*Trimeresurus albolabris*) showing the heat-sensitive pit organ between eye and nostril. (Copyright D.A. Warrell.)

Plate 4 Bleeding from gingival sulci in a victim of the West African saw-scaled or carpet viper (*Echis ocellatus*). (Copyright D.A. Warrell.)

Plate 5 The two species of venomous lizards: left, Mexican beaded lizard (*Heloderma horridum*); right, gila monster (*H. suspectum*). (By courtesy of the Zoological Society of London.)

Plate 6 Poison frog – *Dendrobates histrionicus* (Dendrobatidae) from Bahia Solauo, Colombia. Its skin secretion contains potent nicotinic receptor antagonists, histrionicotoxins. (Copyright D.A. Warrell.)

Plate 7 Poison dart frogs – *Phyllobates terribilis* (Dendrobatidae) from the Chocó region of Colombia, where their skin secretions, containing potent batrachotoxins, are used to coat blow gun darts. (Copyright D.A. Warrell.)

Plate 8 Hooded Pitohui (*Pitohui dichrous*), Vararata National Park, near Port Moresby, Papua New Guinea. (By courtesy of Dr Ian Burrows, Port Moresby.)

Plate 9 Venomous lion fish or butterfly cod (*Brachirus* or *Dendrochirus zebra*), from Madang, Papua New Guinea. (Copyright D.A. Warrell.)

Plate 10 Necrotic and secondarily infected wound at the site of a sting by a freshwater ray (*Potamotrygon hystrix*) in a Brazilian patient. (By courtesy of Dr João Luiz Costa Cardoso, São Paulo, Brazil.)

Plate 11 Extensive weals from contact with the stinging tentacles of the box jellyfish (*Chironex fleckeri*) in an Australian patient stung in Darwin. (By courtesy of Drs B. Currie and P. Nitschke, Darwin.)

Plate 12 Geography coneshell (*Conus geographus*) 10 cm long, responsible for killing a nine-year-old boy at Samarai, Papua New Guinea. (Copyright D.A. Warrell.)

Plate 13 Northern blue-ringed or spotted octopus (*Hapalochlaena lunulatus*) from Madang, Papua New Guinea. (Copyright D.A. Warrell.)

Plate 14 Fourteen-year-old Brazilian boy severely envenomed after more than 1000 stings by Africanized honey bees (*Apis mellifera scutellata*). (Copyright D.A. Warrell.)

Plate 15 Lepidopterism. Lesions caused by urticating abdominal hairs of female moths (*Hylesia* sp.) during an epidemic on the Brazilian coast near São Paulo. (Copyright D.A. Warrell.)

Plate 16 Caterpillar of *Lonomia achelous* whose bristle venom can cause a fatal bleeding diathesis. (By courtesy of Dr Habib Fraiha, Belém, Brazil.)

Plate 17 Beetle (*Paederus crebripunctatus*, Staphylinidae) responsible for causing 'Nairobi eye'. (By courtesy of Dr John Paul, Brighton.)

Plate 18 Scorpion (*Tityus serrulatus*) from Brazil. (Copyright D.A. Warrell.)

Plate 19 Local blistering and necrosis caused by the sting of the scorpion *Hemiscorpius lepturus* (Scorpionidae) found in Iran and Iraq. (By courtesy of Dr M. Radmanesh, Shiraz, Iran.)

Plate 20 Threatening posture of a female Brazilian 'banana spider' (*Phoneutria nigriventer*). Note multiple eyes and large chelicerae. (Copyright D.A. Warrell.)

Plate 21 Female *Loxosceles laeta*. (Copyright D.A. Warrell.)

Plate 22 Australian red back spider (*Latrodectus hasseltii*). (Copyright D.A. Warrell.)

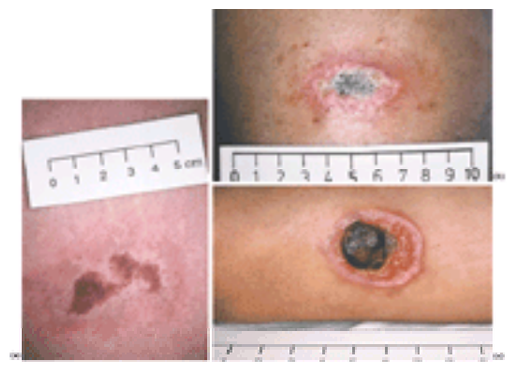

Plate 23 Necrotic araneism. Evolution of the typical lesion following bites by *Loxosceles gaucho* in Brazil. (a) Early ischaemic lesion showing the 'red, white, and blue' sign. (b) 2 weeks later. (c) Necrotic eschar 6 weeks later. (Copyright D.A. Warrell.)

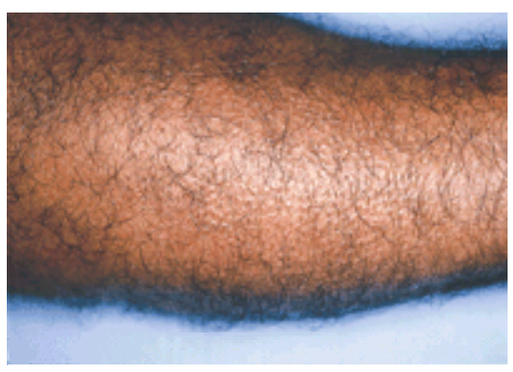

Plate 24 Local sweating and piloerection at the site of a bite by the banana spider *Phoneutria nigriventer*. (Copyright D.A. Warrell.)

Chapter 8.3 Poisonous plants and fungi

Plate 1 Dumb cane, *Dieffenbachia* sp. (GTC).

Plate 2 Laburnum, *Laburnum anagyroides* (GTC).

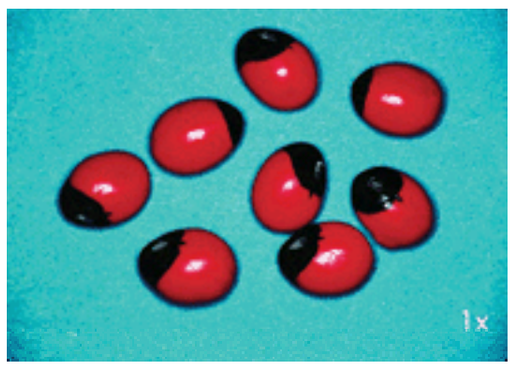

Plate 3 Jequirity beans, *Abrus precatorius* (RBG, Kew).

Plate 4 Castor beans, *Ricinus communis* (GTC).

Plate 5 Woody nightshade, *Solanum dulcamara* (GTC).

Plate 6 Foxglove, *Digitalis purpurea* (GTC).

Plate 7 Oleander, *Nerium oleander* (GTC).

Plate 8 Monkshood, *Aconitum* sp. (GTC).

Plate 9 Yew, *Taxus baccata* (GTC).

Plate 10 Khat, *Catha edulis* (RBG, Kew).

Plate 11 Comfrey, *Symphytum offi- cinale* (GTC).

Plate 12 Poison ivy, *Rhus radicans* (MCC).

Plate 13 Giant hogweed, *Heracleum mantegazzianum* (GTC).

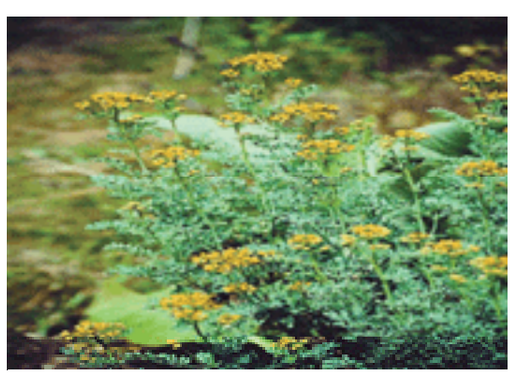

Plate 14 Rue, *Ruta graveolens* (GTC).

Plate 15 Hemlock, *Conium maculatum* (GTC).

Plate 16 Deadly nightshade, *Atropa belladonna* (GTC).

Plate 17 Angel's trumpets, *Brugmansia* sp. (GTC).

Plate 18 Thorn apple, *Datura stramonium* (GTC).

Plate 19 Cycad, *Zamia* sp. (GTC).

Plate 20 Fly agaric, *Amanita muscaria* (GTC).

Plate 21 Death cap, *Amanita phalloides* (GTC).

Plate 22 Roll-rim cap, *Paxillus involutus* (GTC).

Plate 23 Ergot, *Claviceps purpurea* (JW). We thank the following for permission to reproduce their photographs: The Trustees, The Royal Botanic Gardens, Kew (RBG, Kew), G.T. Cooper (GTC), M.C. Cooper (MCC), and Professor J. Webster (JW).

Plates for Section 11

Chapter 11.5 The porphyrias

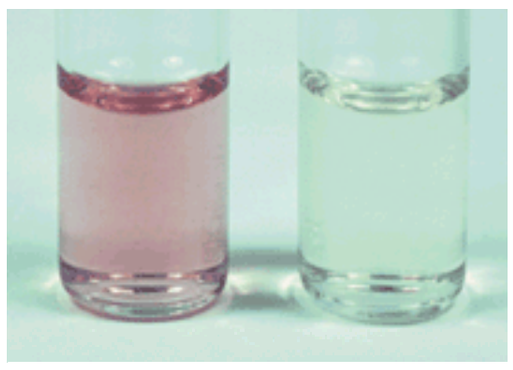

Plate 1 Urine from a patient with acute intermittent porphyria around the time of an acute attack (left); control urine (right). A positive reaction with Ehrlich's diazo reagent is shown in the patient following the addition of 50 µ l of urine to 1 ml of 2 per cent acidic dimethyl benzaldehyde. Subsequent tests showed that the pink diazo adduct was insoluble in chloroform and other organic solvents indicating the presence of excess porphobilinogen. (Urobilinogen in excess may give a positive reaction with the diazo reagent but the product is readily extracted into organic solvents.)

Plate 2 Porphyria cutanea tarda in a 60-year-old heterozygote for the *HFE* C282Y mutation. This man, a taxi driver, had noticed irritation after exposure of his hands to light transmitted through the windscreen. He had noticed fragility and blistering combined with pigmentary changes typical of this disorder. After treatment by controlled phlebotomy his skin complaint has regressed.

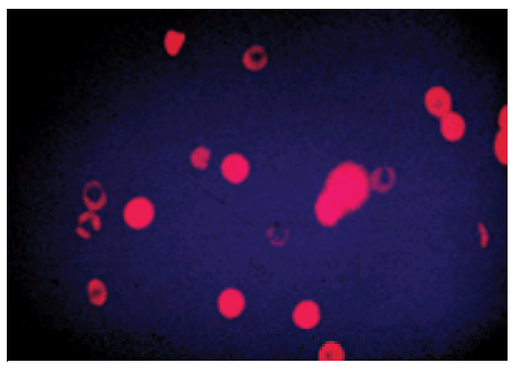

Plate 3 Fluorescent microscopy of an unstained blood film from a patient with erythropoietic protoporphyria. Note the red fluorescence of increased free protoporphyrin within individual young erythrocytes and reticulocytes.

Plate 4 Examination of human plasma under long-wave ultraviolet light. Plasma on the right was obtained from a patient with protoporphyrin hepatopathy and greatly increased photosensitivity and is compared with plasma obtained from a healthy subject on the left. Note the bright red fluorescence due to the presence of high concentrations of free protoporphyrin. Maximum fluorescence was obtained by exposure to visible light in the violet and green–yellow spectral regions corresponding to the absorbance bands of porphyrins.

Chapter 11.6 Lipid and lipoprotein disorders

Plate 1 Achilles tendon xanthoma (heterozygous familial hypercholesterolaemia).

Plate 2 Tendon xanthomata on the dorsum of a hand (heterozygous familial hypercholesterolaemia).

Plate 3 Eruptive and tuberous xanthomata on an arm (type III hyperlipoproteinaemia with marked hypertriglyceridaemia).

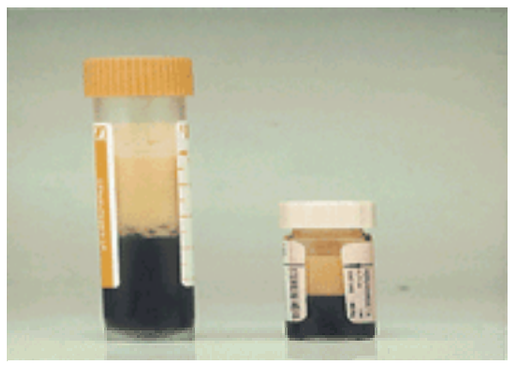

Plate 4 Milky plasma indicating marked hypertriglyceridaemia (blood samples from a patient with acute abdominal pain).

Chapter 11.7.1 Hereditary haemochromatosis

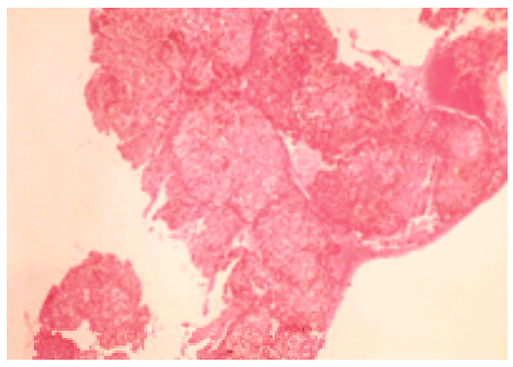

Plate 1 Low-power, needle-biopsy appearance of liver specimen stained with haematoxylin and eosin from a 67-year-old man with adult haemochromatosis due to homozygosity for the *C282Y* mutation. Note the large hyperplastic nodules and fibrosis.

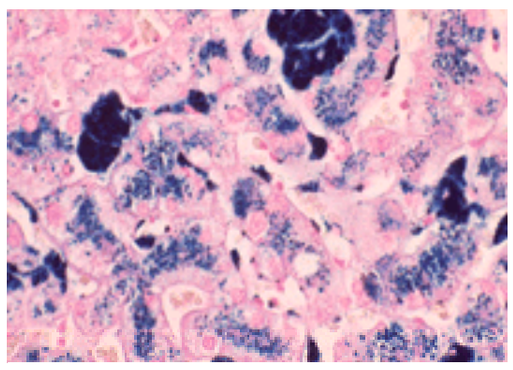

Plate 2 High-power micrograph of the liver biopsy specimen shown in Plate 1 stained with Perls' reagent. Note extensive deposits of ferric iron in all cell types including Kupffer cells, cells lining small biliary radicles, and in a punctate distribution within parenchymal hepatocytes. Liver cells are hyperplastic.

Plate 3 Arthropathy in a man with adult haemochromatosis forced to stop manual work because of painful arthritis especially in the second and third metacarpophalangeal joints; note increased skin pigmentation.

Plate 4 Adult haemochromatosis. Section of liver lobe after surgical resection to remove a primary hepatocellular carcinoma arising in an iron-loaded but, unusually, non-cirrhotic liver in this disorder. The patient, aged 62 years, had been partially treated by venesection but recently noticed increasing lethargy: a raised serum a-fetoprotein concentration led to the diagnosis; moderate histochemical evidence of iron storage was found in the non-malignant tissue excised at surgery.

Chapter 11.7.2 Wilson's disease, Menke's disease: inherited disorders of copper metabolism

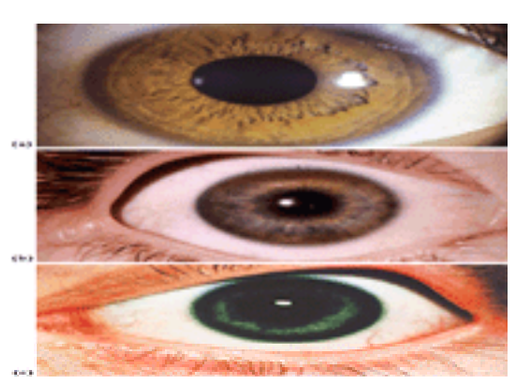

Plate 1 (a–c) Kayser–Fleischer ring in Wilson's disease.

Plate 2 Penicillamine dermatopathy—elastosis perfringens serpiginosa.

Plate 3 Appearance in Menkes' disease.

Chapter 11.8 Lysosomal storage diseases

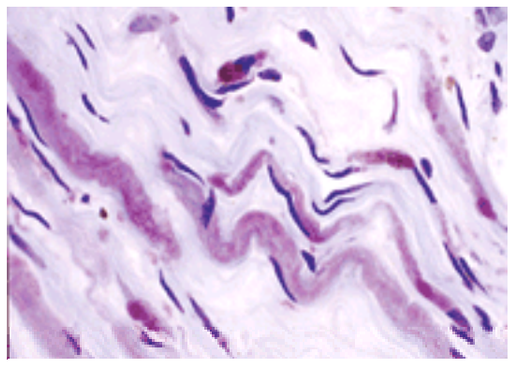

Plate 1 Sural nerve biopsy stained with toluidine blue from the patient shown in Plate 2 with metachromatic leucodystrophy. Note the brown-staining granular material within Schwann and perineurial macrophages typical of this disorder due to the deposition of the glycolipid sulphatide. (By courtesy of Dr J. Xuereb, Addenbrooke's Hospital).

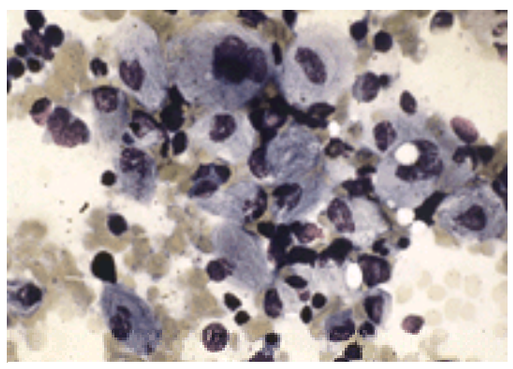

Plate 2 Light micrograph of a Leishmann-stained bone marrow biopsy obtained from a 23-year-old man with type 1 Gaucher's disease. Note that the large, pale-blue staining Gaucher's cells with striated cytoplasm replace the Kupffer cells of the liver, alveolar macrophages of the lung, and of the bone marrow.

Chapter 11.13 a_1-Antitrypsin deficiency and the serpinopathies

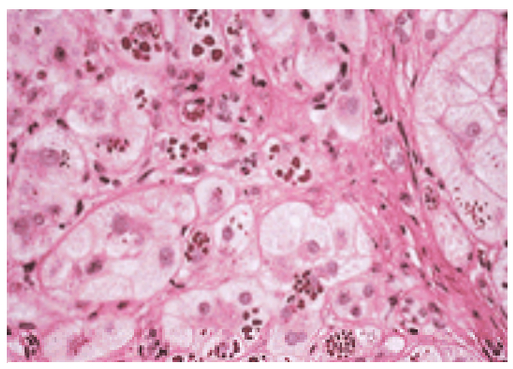

Plate 1 A chain of loop-sheet polymers isolated from a patient with a_1-antitrypsin deficiency. These polymers can form filaments or circlets that tangle within the endoplasmic reticulum of the hepatocyte to form the inclusions which are the hallmark of the disease. These intrahepatic inclusions are characteristically periodic acid–Schiff (**PAS**)-positive and diastase-resistant and stain positive for a_1-antitrypsin on immunohistochemistry.

Plates for Section 12

Chapter 12.4 The thyroid gland and disorders of thyroid function

Plate 1 Thyroid dermopathy (pretibial myxoedema) affecting the lateral aspect of the shin and the dorsum of the foot; the patient also had thyroid acropachy.

Chapter 12.7.1 Diseases of the adrenal cortex

Plate 1 A young woman with Cushing's disease, photographed initially alongside her identical twin sister (a). In this case treatment with bilateral adrenalectomy was undertaken and several years later the patient re-presents with Nelson's syndrome and a right III cranial nerve palsy due to cavernous sinus infiltration from a locally invasive corticotrophinoma.

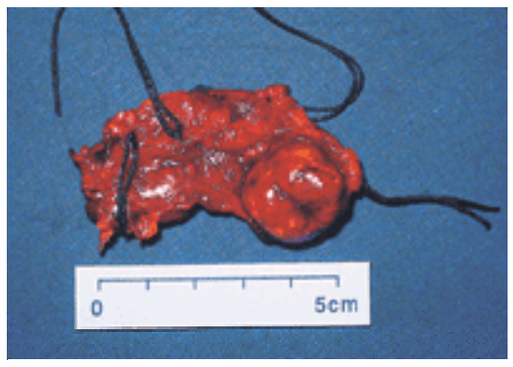

Plate 2 A solitary adrenal adenoma. The characteristic yellow appearance of the cut surface of the excised tumour reflects the high cholesterol content.

Chapter 12.8.1 Approach to the patient with ovarian disorders

Plate 1 Acanthosis nigricans in a young woman with polycystic ovary syndrome.

Chapter 12.10 Non-diabetic pancreatic endocrine disorders and multiple endocrine neoplasia

Plate 1 (a) Necrolytic migratory erythema in a patient with the glucagonoma syndrome. (b) Pigmentation in healed areas.

Chapter 12.11.1 Diabetes mellitus

Plate 1 (a) Diabetic amyotrophy: quadriceps right wasting due to femoral neuropathy (with thanks to Dr Geoff Gill, University Hospital, Aintree, Liverpool). (b) Wasting of small muscles of the hands due to both ulnar and median nerve lesions.

Plate 2 The diabetic foot. (a) Typical punched-out neuropathic ulcer on the lateral aspect of the sole in an ischaemic foot with gangrene of the second, fourth, and fifth toes. (b) Ulceration and digital gangrene, caused by wearing tight shoes on a severely ischaemic foot.

Plate 3 The hands in long-standing diabetes. (a) Limited joint mobility (cheiroarthropathy), showing the 'prayer sign'. (b) Thickening of the skin over the knuckles and proximal interphalangeal joints (Garrod's pads).

Plate 4 Necrobiosis lipoidica diabeticorum (with thanks to Dr Geoff Gill, University Hospital, Aintree, Liverpool.

Plates for Section 13

Chapter 13.13 The skin in pregnancy

Plate 1 Polymorphic eruption of pregnancy: urticated papules and plaques on the thigh.

Plate 2 Pemphigoid gestationis: urticated papules and plaques and blisters (reproduced with permission from Charles-Holmes R, Black MM (1990). Herpes gestationis. In: Wojnarowska F, Briggaman RA, eds. *Management of blistering disease*, pp. 93–104. Chapman and Hall, London).

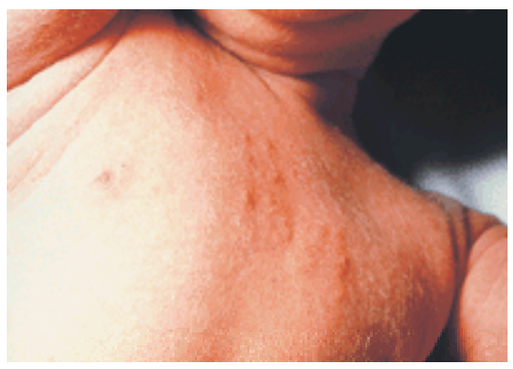

Plate 3 Pemphigoid gestationis: urticated papules in the neonate (reproduced with permission from Charles-Holmes R, Black MM (1990). Herpes gestationis. In: Wojnarowska F, Briggaman RA, eds. *Management of blistering disease*, pp. 93–104. Chapman and Hall, London).

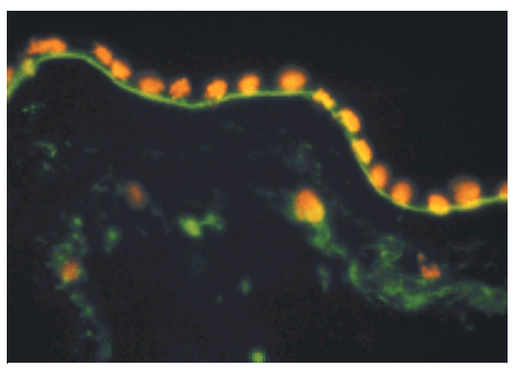

Plate 4 Pemphigoid gestationis: linear deposition of C3 at the amnion basement membrane zone as demonstrated by immunofluorescence. The nuclei are counterstained with propidium iodide. (Provided by B.S. Bhogal and M.M. Black, St John's Institute of Dermatology, St Thomas's Hospital, London.)

Plates for Section 14

Chapter 14.4 Immune disorders of the gastrointestinal tract

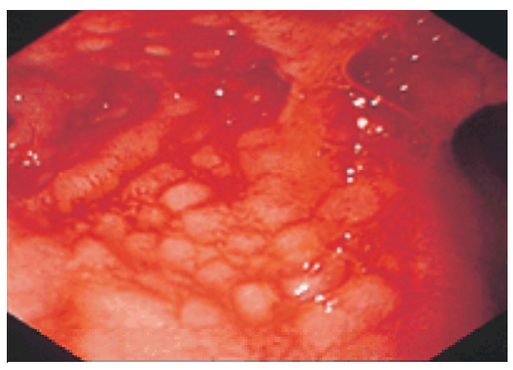

Plate 1 The appearance of nodular lymphoid hyperplasia on upper gastrointestinal endosocpy.

Chapter 14.20.1 Viral hepatitis – clinical aspects

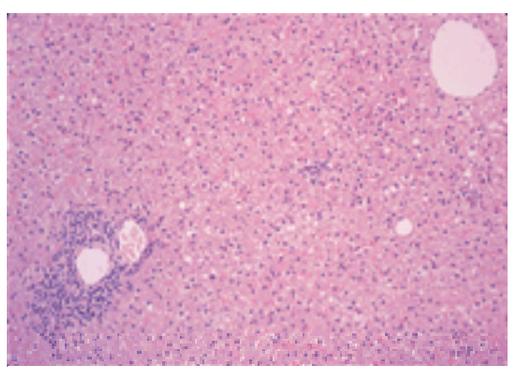

Plate 1 Serological changes during chronic hepatitis C.

Chapter 14.20.2.1 Autoimmune hepatitis

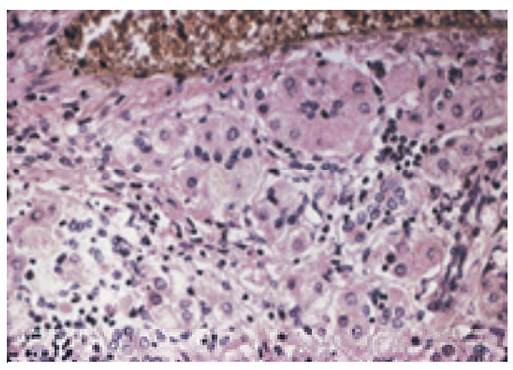

Plate 1 Haematoxylin and eosin stained liver histology showing 'rosettes' of regenerated hepatocytes, surrounded by lymphocytes that have spread into the hepatic parenchyma.

Chapter 14.21.3 Hepatocellular failure

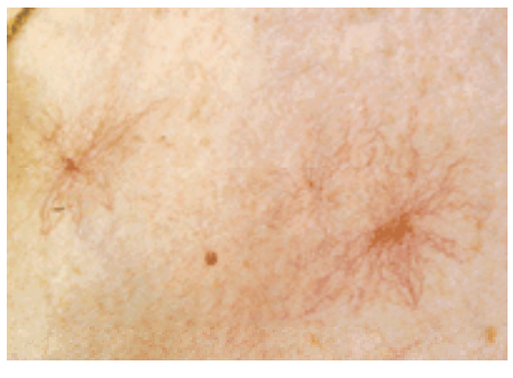

Plate 1 Spider naevi in a patient with cirrhosis.

Plate 2 Palmer erythema in a patient with cirrhosis.

Plate 3 White nails in a patient with cirrhosis.

Plate 4 Finger clubbing in a patient with cirrhosis.

Chapter 14.21.6 Hepatic granulomas

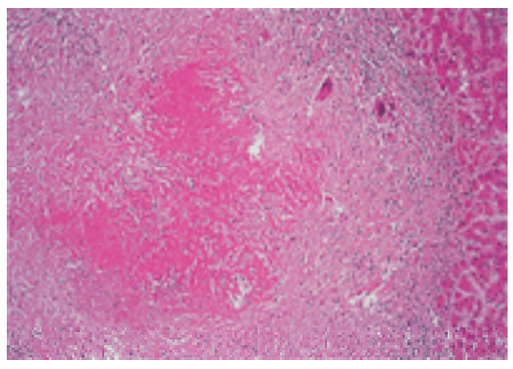

Plate 1 Liver showing a portion of a large caseating granuloma from a patient with miliary mycobacterium tuberculosis. Several Langhans' giant cells are also seen. (Haematoxylin and eosin, total magnification 25.)

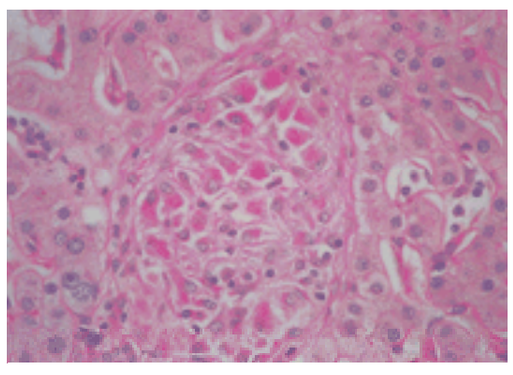

Plate 2 Liver from an HIV/AIDS patient infected with *Mycobacterium avium intracellulare*, showing a granuloma composed of epithelioid cells which contain large numbers of micro-organisms. (Diastase/periodic acid–Schiff stain, total magnification 100.)

Plates for Section 15

Chapter 15.3.3 Echocardiography

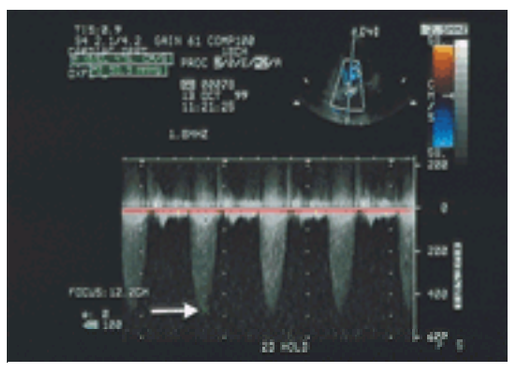

Plate 1 Apical continuous wave Doppler across the aortic valve in a patient with severe aortic stenosis. The peak velocity is greater than 4.5 m/s consistent with a peak instanteneous gradiant across the aortic valve of 90 mmHg.

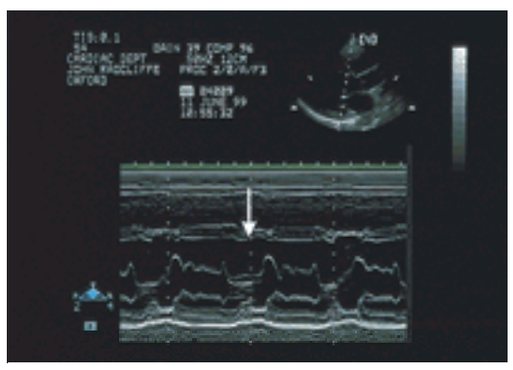

Plate 2 M-mode echocardiogram through the mitral valve in a normal patient. Opening of the leaflets during ventricular diastole and closing during systole (arrow) can be observed.

Chapter 15.8.2 The cardiomyopathies: hypertrophic, dilated, restrictive, and right ventricular

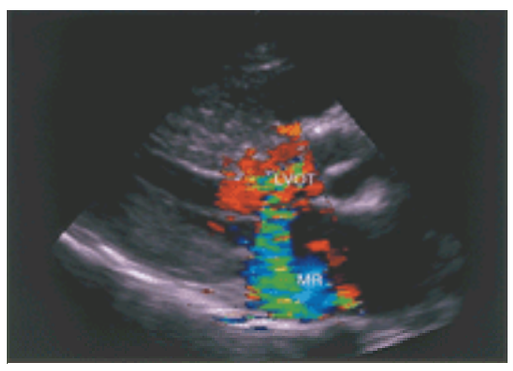

Plate 1 Colour flow Doppler image (parasternal long axis view) of the same patient as shown in Fig. 3 of Chapter 15.8.2, demonstrating left ventricular outflow tract (LVOT) turbulence (shown in red) and mitral regurgitation (MR) with a posteriorly directed jet (shown in blue/green).

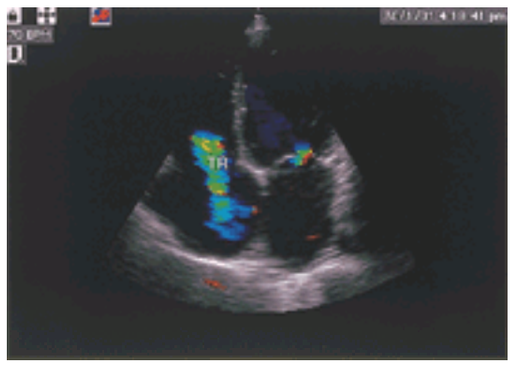

Plate 2 Colour flow Doppler image of the same patient as shown in Fig. 5 of Chapter 15.8.2 showing a regurgitant tricuspid jet (TR, shown in blue).

Chapter 15.10.2 Infective endocarditis

Plate 1 Splinter haemorrhages in a case of infective endocarditis.

Plate 2 Vasculitic rash on lower limb of a patient with infective endocarditis.

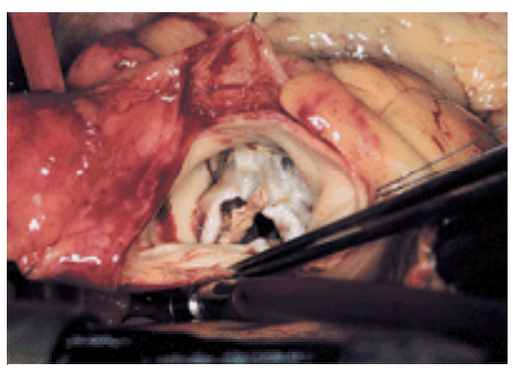

Plate 3 A large vegetation on the aortic valve of a patient with infective endocarditis as seen at the time of surgery.

Chapter 15.14.1 Thoracic aortic dissection

Plate 1 Post-mortem specimen of aortic dissection. The intimal/medial flap is pulled back with a retractor to show the false lumen parallel to the true lumen.

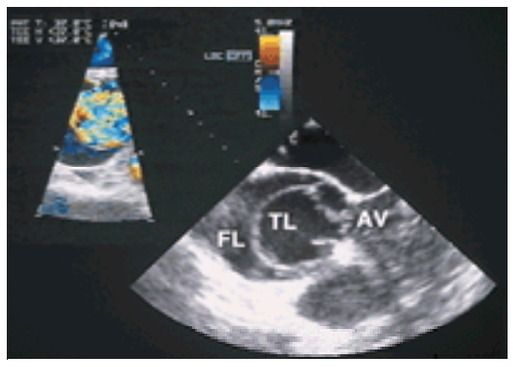

Plate 2 Transoesphageal transverse two-dimensional and colour Doppler echo images of the ascending aorta showing a dissection membrane partitioning the true (TL) and false lumen (FL). Upper left panel shows systolic flow in the true but not the false lumen.

Chapter 15.14.2 Peripheral arterial disease

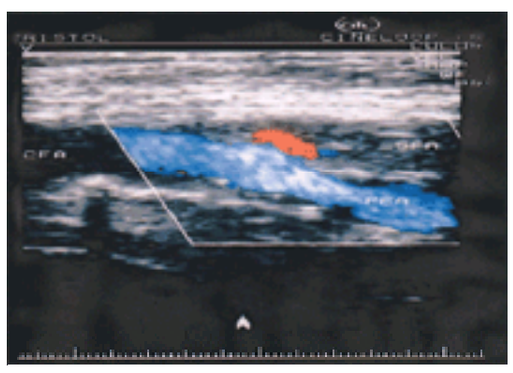

Plate 1 Occlusion of the superficial femoral artery demonstrated by colour-coded duplex ultrasonography. On the left, the common femoral artery (CFA) lies outside the colour box. In the colour box antegrade flow through the profuma femoris artery (PFA) is shown in blue. The red flash represents rebound flow against the occluded origin of the superficial femoral artery (SFA).

Chapter 15.14.3 Cholesterol embolism

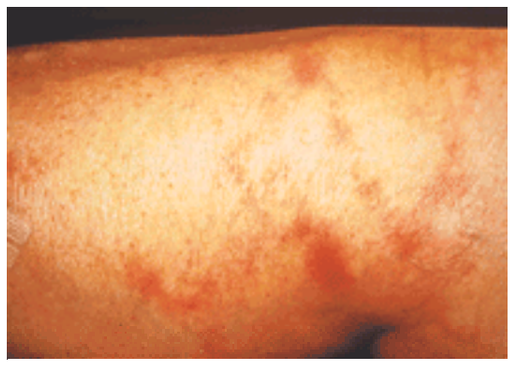

Plate 1 Livedo reticularis and vasculitic-like erythematous nodules on the leg of a patient in whom cholesterolcrystal embolization occurred after coronary angiography.

Plate 2 Purpuric spots and acral cyanosis of the toes from cholesterol embolism after aortic aneurysm repair.

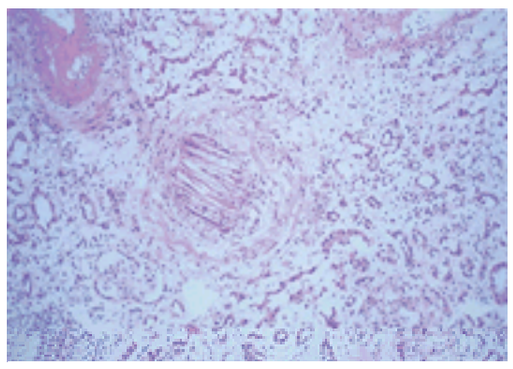

Plate 3 Renal biopsy demonstrating the characteristic needle-shaped cholesterol clefts occluding a medium-sized renal arteriole with surrounding inflammatory cell infiltration, intimal proliferation, thickening, and concentric fibrosis. There is extensive autolysis (postmortem sample).

Chapter 15.14.4 Takayasu arteritis

Plate 1 Typical coronary anastomosis of retinal vessels in Takayasu arteritis.

Chapter 15.15.2.1 Primary pulmonary hypertension

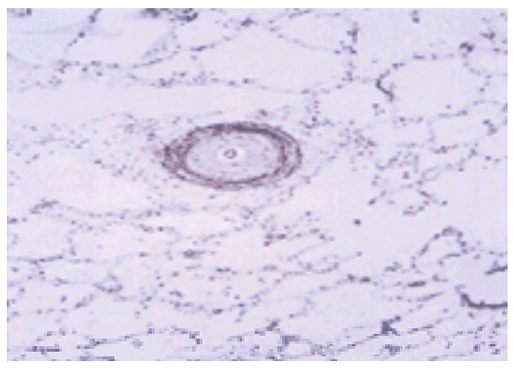

Plate 1 Intimal thickening of a pulmonary artery in pulmonary hypertension (Chazova I *et al.*, 1995. Pulmonary artery adventitial changes and venous involvement in primary pulmonary hypertension. *American Journal of Pathology* **146**, 389–97).

Chapter 15.16.3 Hypertensive emergencies and urgencies

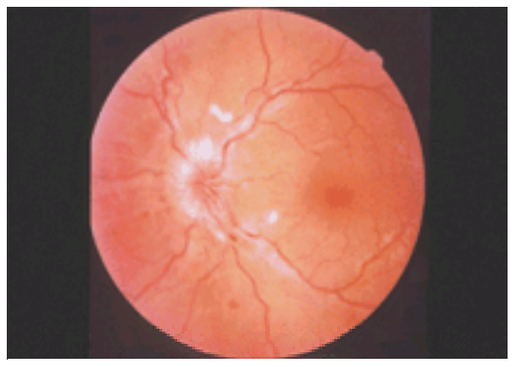

Plate 1 Ocular fundus in hypertension, showing papilloedema, exudates, and a few haemorrhages.

Plates for Section 17

Chapter 17.3.4 Diagnostic bronchoscopy, thoracoscopy, and tissue biopsy

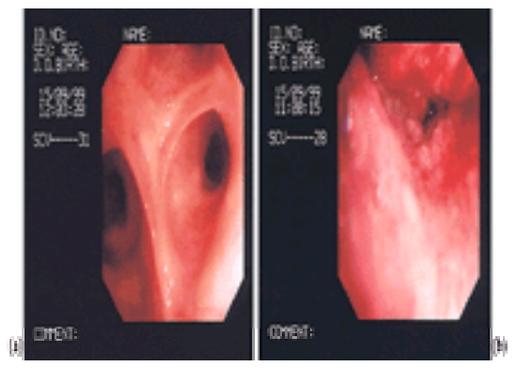

Plate 1 Appearances at bronchoscopy. The normal thin mucosa and sharp interlobar carinae of the normal left side (a) are in contrast to the irregular exophytic appearance of an advanced non-small cell tumour of the right main bronchus (b) in the same 73-year-old patient.

Chapter 17.11.2 Cryptogenic fibrosing alveolitis

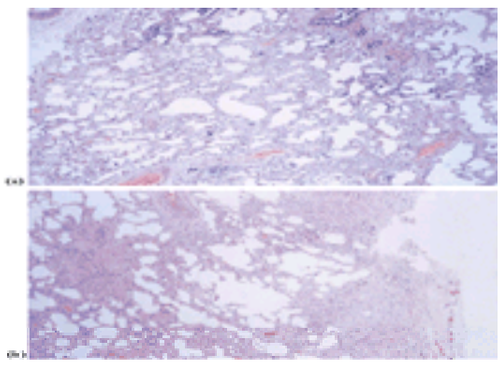

Plate 1 Histopathological appearance of cryptogenic fibrosing alveolitis and the non-specific interstitial pneumonia 'mimic'. (a) Usual interstitial pneumonia, the histopathological pattern seen in cryptogenic fibrosing alveolitis. Note the pale, fibroblast foci that are the hallmark of usual interstitial pneumonia. (b) The nonspecific interstitial pneumonia 'mimic' of cryptogenic fibrosing alveolitis. This is much less common than usual interstitial pneumonia. Note the uniformity of the pathology throughout the section.

Chapter 17.11.3 Bronchiolitis obliterans and organizing pneumonia

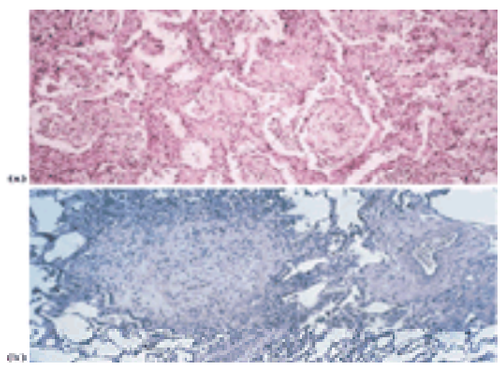

Plate 1 Histopathology. (a) Proliferative bronchiolitis. (b) Constrictive bronchiolitis. Note the loosely packed granulation tissue in (a) in contrast to the more established scarring in (b).

Plates for Section 18

Chapter 18.3 Clinical investigation (rheumatological disorders)

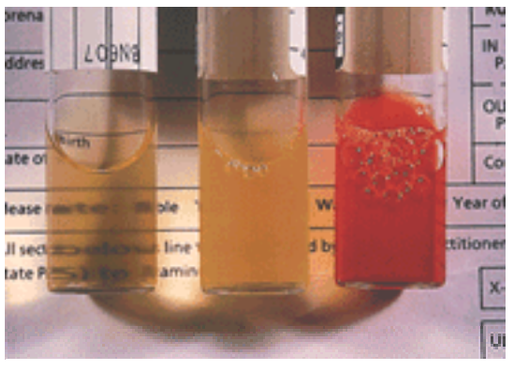

Plate 1 Different macroscopic appearances of synovial fluids: (a) on the left, clear straw-coloured fluid from an osteoarthritic knee (easy to read writing behind it); (b) less viscous, turbid (high cell count) 'inflammatory' fluid from a rheumatoid knee; and (c) uniform bloodstaining (haemarthrosis) due to acute pseudogout.

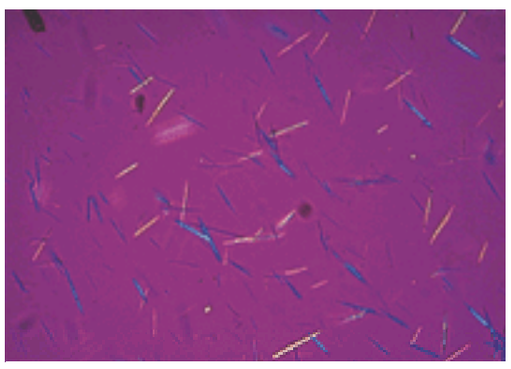

Plate 2 Monosodium urate crystals viewed by compensated polarized light microscopy (×400) showing bright birefringence (negative sign) and needle-shaped morphology.

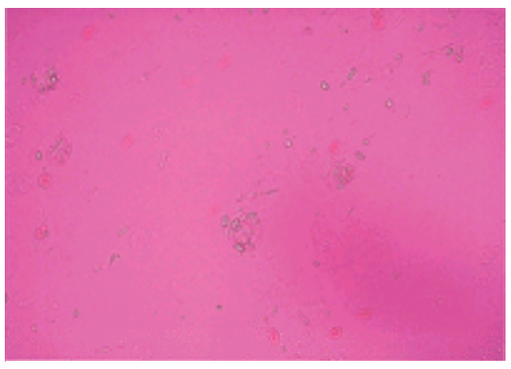

Plate 3 Calcium pyrophosphate crystals viewed by polarized light microscopy (×400) showing weak birefringence (positive sign), scant numbers, and a predominantly

rhomboid morphology. These are clearly more difficult to detect than urate crystals.

Chapter 18.5 Rheumatoid arthritis

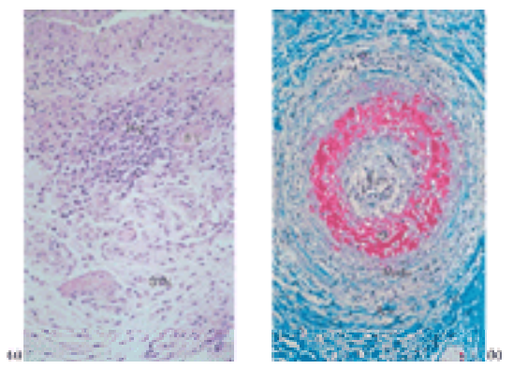

Plate 1 Histology of rheumatoid arthritis. (a) Rheumatoid arthritis synovitis. L.L., lining layer; P.V., perivascular aggregate of lymphocytes and macrophages; B.V., blood vessel; SYN, synoviocytes; (haematoxylin and eosin staining). (b) Small vessel arteritis. Lum., lumen; Int., Intima; P.V., perivascular inflammation; Adv., adventitial tissue. Arterial wall shows a thrombosed vessel with intimal hyperplasia, destruction of internal elastic lamina, and mononuclear cell infiltration of media and perivascular tissue (methylene blue and safranine staining).

Plate 2 The hands of a person suffering from rheumatoid arthritis. Features to note include symmetrical soft tissue swelling of the second and third metacarpophalageal joints, early swan-neck deformity of the left ring finger, ulnar deviation at the metacarpophalageal joints, and wasting of the small muscles of the hand. In addition, several small rheumatoid nodules are present.

Chapter 18.6 Spondyloarthritides and related arthritides

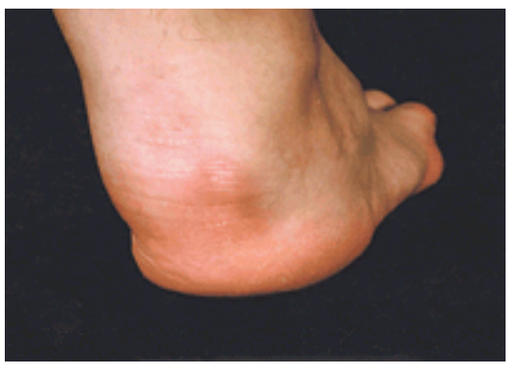

Plate 1 Enthesitis at the insertion of the Achilles tendon in a patient with reactive arthritis.

Plate 2 Dactylitis of the third finger of the right hand in a patient with undifferentiated spondyloarthropathy.

Plate 3 30-year-old man with rapidly progressive ankylosing spondylitis (disease of 5 years duration).

Plate 4 Severe psoriatic arthritis (arthritis mutilans).

Plate 5 Arthritis/hyperostosis of the left sternoclavicular joint in a 52-year-old man with SAPHO syndrome.

Chapter 18.10.2 Systemic lupus erythematosus and related disorders

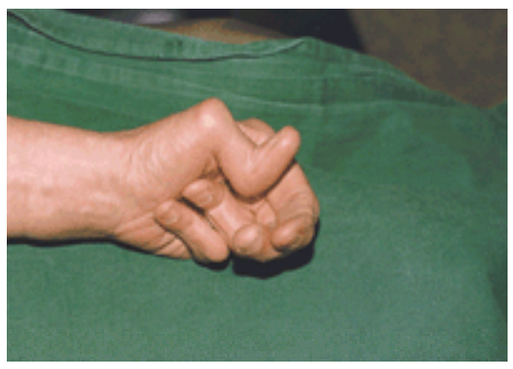

Plate 1 Deforming Jaccoud's arthropathy.

Plate 2 Malar 'butterfly' rash.

Plate 3 Severe scarring alopecia.

Plate 4 Livedo reticularis.

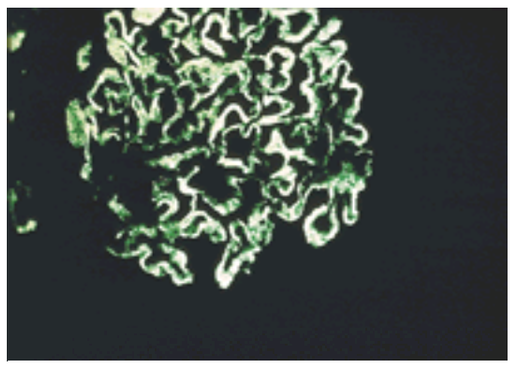

Plate 5 Immuno fluorescence microscopy showing deposition of IgG in the glomerulus of a patient with lupus nephritis.

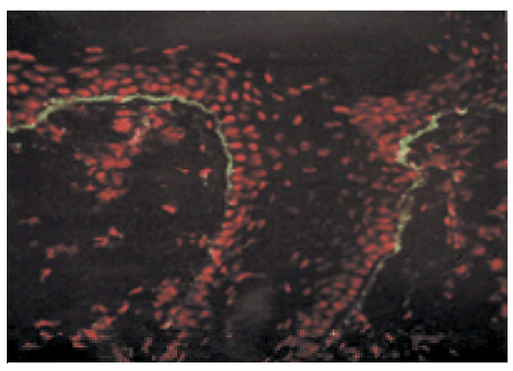

Plate 6 Immunofluorescence microscopy showing deposition of IgG at the dermoepidermal junction in the skin of a patient with systemic lupus erythematosus (sometimes called the lupus band test).

Chapter 18.10.4 Polymyalgia rheumatica and giant cell arteritis

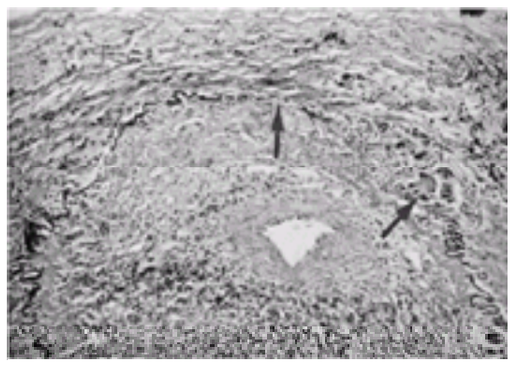

Plate 1 Photomicrograph of a temporal artery biopsy showing giant cells, mononuclear infiltrate, and disruption of the internal elastic lamina.

Chapter 18.10.7 Polymyositis and dermatomyositis

Plate 1 Gottron's sign. Roughened, violaceous papules over the dorsal surfaces of several metacarpophalangeal and proximal interphalangeal joints. Note also the erythema at the bases of the fingernail, caused by capillary loop dilatation.

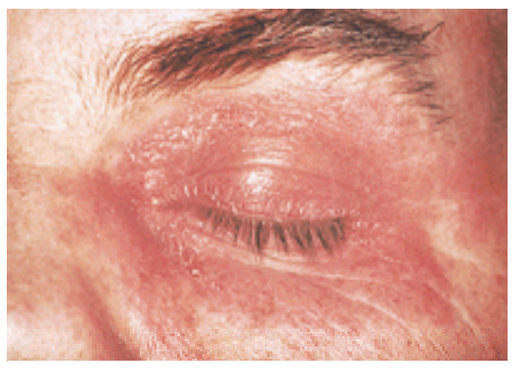

Plate 2 Heliotrope rash. An erythematous (often lilac-coloured) rash over the eyelids in a patient with dermatomyositis (reproduced from Mousari HC, Wigley FM (2000). *Journal of Rheumatology* **27**, 1542-5 with permission).

Chapter 18.10.8 Kawasaki syndrome

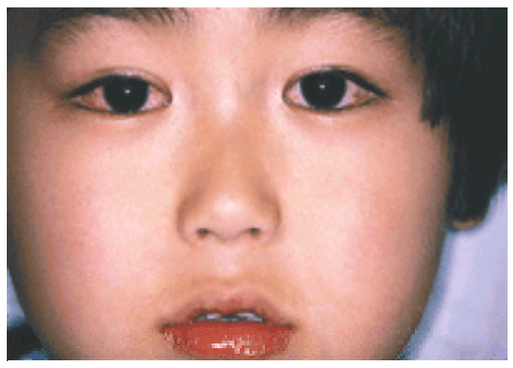

Plate 1 Typical appearance of a patient with Kawasaki disease; note the red eyes and red lips (picture of a 5-year-old boy, taken on the fourth day of illness).

Chapter 18.11 Miscellaneous conditions presenting to the rheumatologist

Plate 1 Pyoderma gangrenosum.

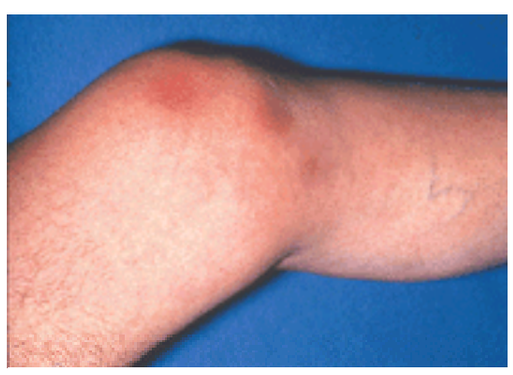

Plate 2 Erythema nodosum.

Plates for Section 19

Chapter 19.2 Inherited defects of connective tissue: Ehlers Danlos syndrome, pseudoxanthoma elasticum, and Marfan syndrome

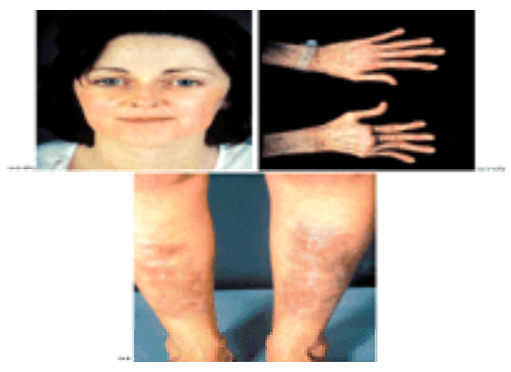

Plate 1 See *caption overleaf*.

EDS type IV (vascular type). (a) Acrogeria–a specific clinical feature of EDS IV. Note the large eyes and thin (b) nose (Madonna facies) with periorial wrinkling. (c) Premature wrinkling of the skin on the dorsum of the hands; note also the joint contractures (d) superficially resembling rheumatoid arthritis. (e) Pretibial bruising and

haemosiderosis.

Plate 2 Skin lesions in pseudoxanthoma elasticum (PXE). (a) Typical flexural skin lesions of PXE of the lateral neck. (b) Widespread cutis laxa in PXE. (c) Mucosal infiltration of the lower lip in PXE. (d) Elastic van Giessen stain of skin section showing mid-dermal elastic fragmentation and degeneration.

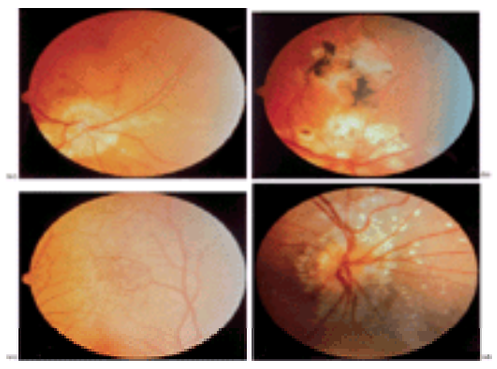

Plate 3 See *caption overrleaf*.

Retinal changes in PXE. (a) Angioid streaks caused by fracture of the retroretinal Bruch's membrane—an early feature. (b) Macular haemorrhage with consequential choroideretinitis. (c) Specked *peau d'orange* mottling. (d) Salmon spotting and drusen.

Plates for Section 20

Chapter 20.3.2 Clinical investigation of renal disease

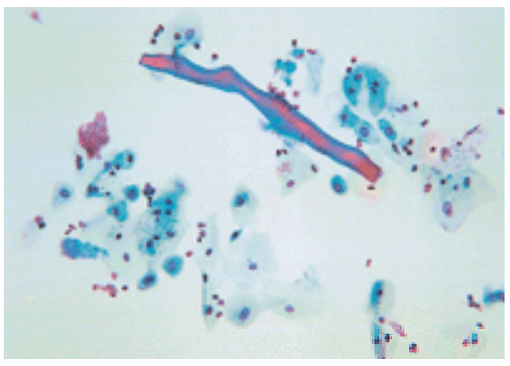

Plate 1 Papanicolaou-stained urine showing a hyaline cast with both normal transitional and squamous cells (blue) and renal tubular cells (pink). (By courtesy of Dr Deery.)

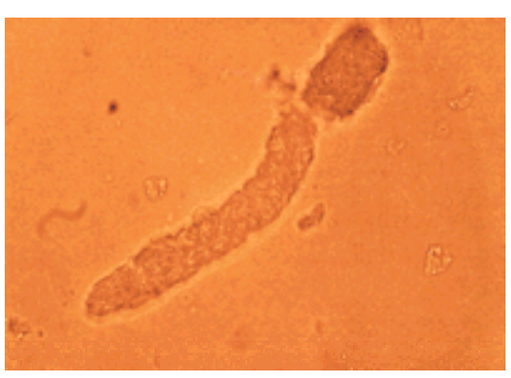

Plate 2 Unstained urine specimen showing a granular cast.

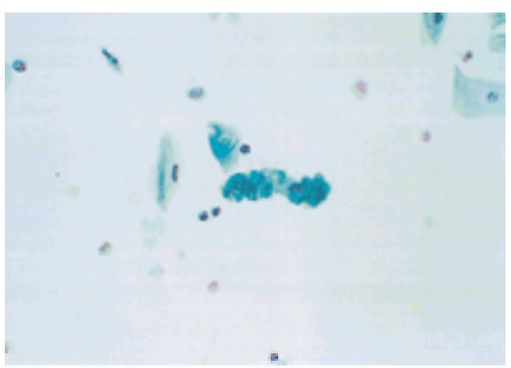

Plate 3 Papanicolaou-stained urine deposit showing a red cell cast.

Chapter 20.7.2 IgA nephropathy and Henoch-Schönlein purpura

Plate 1 Characteristic purpuric rash affecting the lower limbs in Henoch Schönlein purpura.

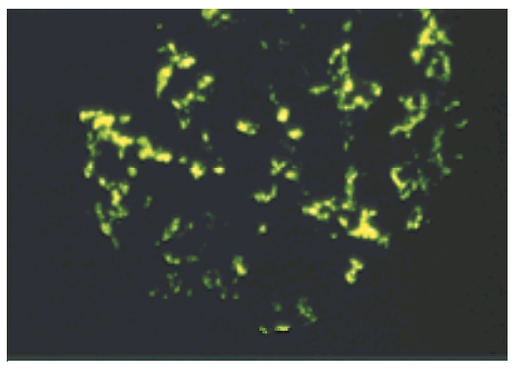

Plate 2 Immunofluorescence of a glomerulus in IgA nephropathy. Bright fluorescent staining is seen within the mesangium with labelled antibodies to IgA. In some cases similar staining is also seen on along capillary walls. A similar distribution of staining for C3 is commonly present (anti-human IgA, × 375).

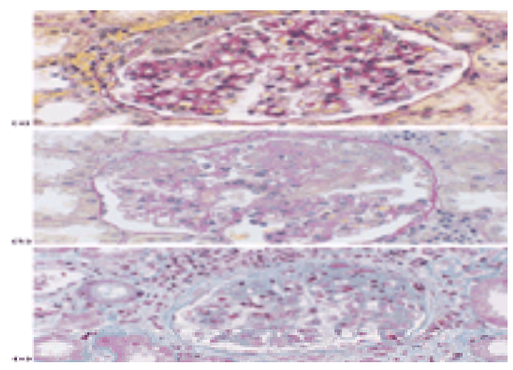

Plate 3 See *caption overleaf*.

Light microscopic appearances of IgA nephropathy. (a) Glomerulus showing global increase in mesangial matrix and cellularity. (Alcian Blue/PAS stain × 375). (b) Glomerulus showing segmental increase in mesangial matrix and hypercellularity with fibrinoid necrosis (solid arrow) and synechia formation (open arrow) between the segmental lesion and parietal epithelium of Bowman's capsule (Alcian Blue/PAS stain, × 375). (c) Glomerulus showing segmental increase in mesangial matrix, segmental sclerosis with synechia formation (open arrows) to overlying Bowman's capsule (Masson Trichrome stain, × 375).

Chapter 20.7.4 Minimal change nephropathy, focal segmental glomerulosclerosis, and membranous nephropathy

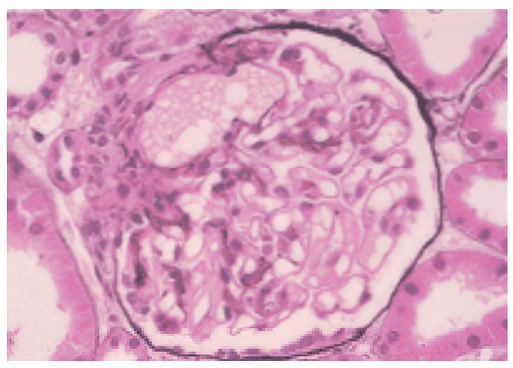

Plate 1 Minimal-change nephropathy. The glomerulus looks normal on light microscopy. Periodic acid-methenamine silver staining (64 ×). (By courtesy of Dr A.J. Howie.)

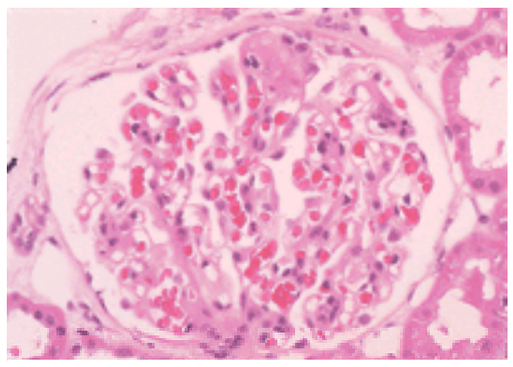

Plate 2 Classical segmental sclerosing glomerulonephritis at an early stage. The glomerulus shows an erratic increase in mesangium with a segmental area of foamy cells and sclerosis opposite the vascular pole, next to the tubular origin. Haematoxylin and eosin staining (50 ×). (By courtesy of Dr A.J. Howie.)

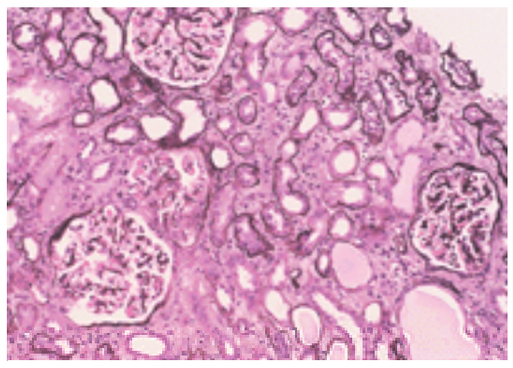

Plate 3 Classical segmental sclerosing glomerulonephritis at a late stage. Four glomeruli show an erratic increase in mesangium and segmental lesions at various sites. Periodic acid-methenamine silver staining (× 64). (By courtesy of Dr A.J. Howie.)

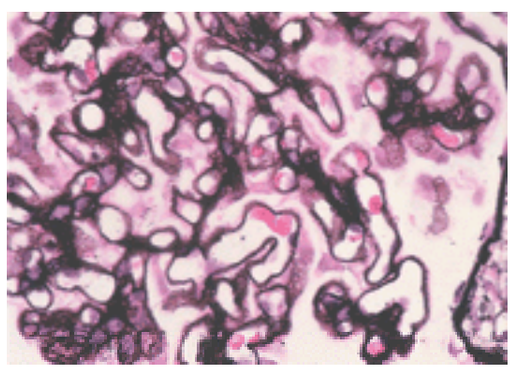

Plate 4 Membranous nephropathy. There are regular short spikes on the outside of glomerular capillary loops. Periodic acid-methenamine silver staining (80 ×).

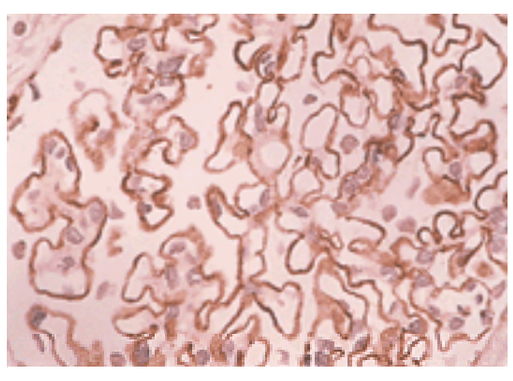

Plate 5 Membranous nephropathy. Immunoperoxidase staining shows uniform granular deposits of IgG on the epithelial side of glomerular basement membranes (80 ×). (By courtesy of Dr A.J. Howie.)

Chapter 20.7.5 Proliferative glomerulonephritis

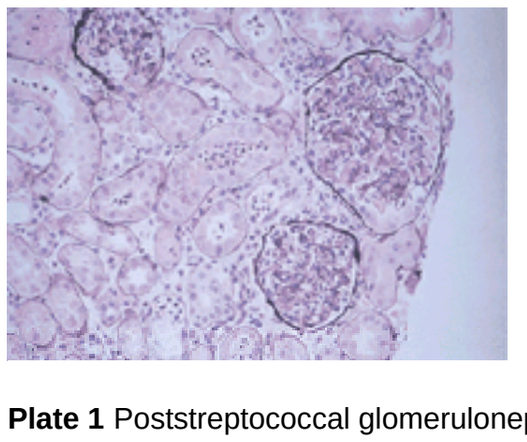

Plate 1 Poststreptococcal glomerulonephritis.

Chapter 20.7.6 Mesangiocapillary glomerulonephritis

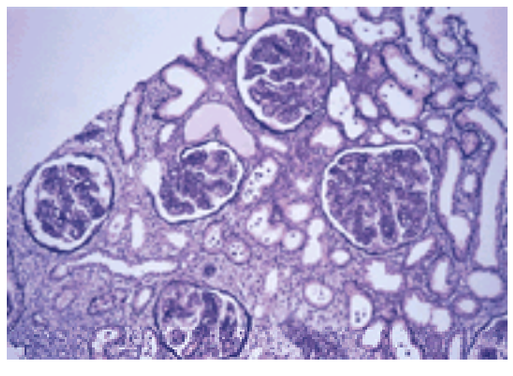

Plate 1 Mesangiocapillary glomerulonephritis. Note characteristic appearance of expanded glomerulus.

Chapter 20.7.7 Antiglomerular basement membrane disease

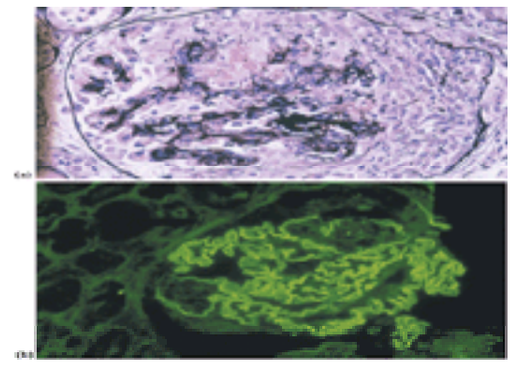

Plate 1 Renal biopsy from a patient with Goodpasture's disease. (a) Light microscopy showing a single glomerulus with cellular crescent and focal necrosis (silver stain). (b) Immunoflourescence of a single glomerulus with linear deposition of IgG along the GBM. (Figure by courtesy of Dr H.T. Cook.)

Chapter 20.7.8 Infection-associated nephropathies

Plate 1 Cutaneous vasculitis in a patient with *Staphylococcus aureus* endocarditis.

Chapter 20.9.1 Acute interstitial nephritis

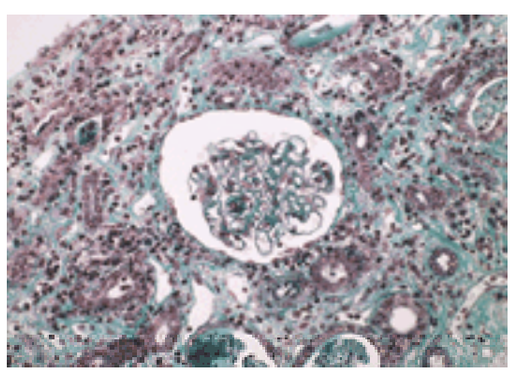

Plate 1 Acute interstitial nephritis. The renal interstitium is invaded by numerous mononuclear cells. The glomerulus is normal. Mason's trichrome 250×.

Chapter 20.10.3 Vasculitis and the kidney

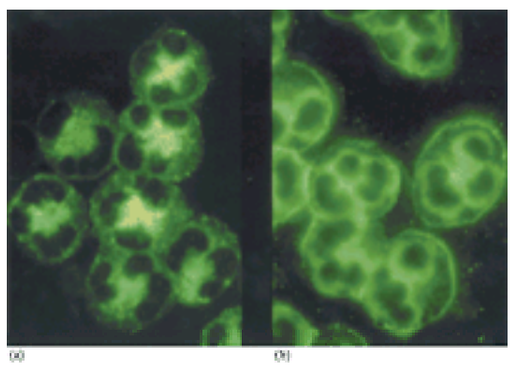

Plate 1 Indirect immunofluorescence assay for ANCA. (a) Typical staining of cytoplasmic ANCA that is usually due to antibodies to proteinase 3. (b) Typical staining pattern of perinuclear ANCA most often due to antimyeloperoxidase antibodies.

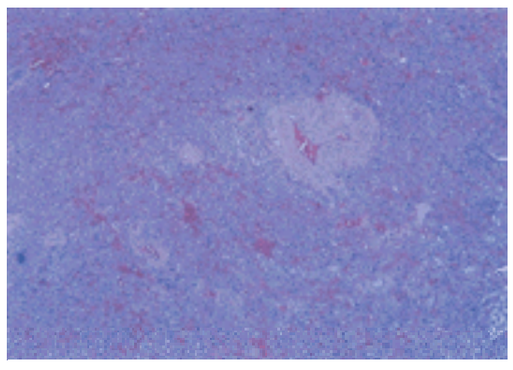

Plate 2 Morphological appearances of pulmonary granulomas in a specimen obtained by video-endoscopic lung biopsy from a patient with Wegener's granulomatosis.

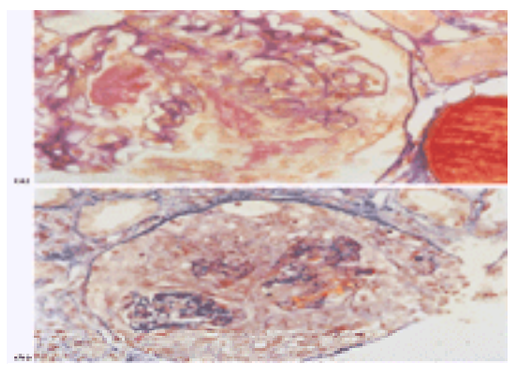

Plate 3 Morphological appearances on a renal biopsy from a patient with pauciimmune focal necrotizing glomerulonephritis. (a) An early lesion with necrosis of one glomerular segment. (b) A much more florid lesion with the whole glomerular tuft surrounded by a crescent.

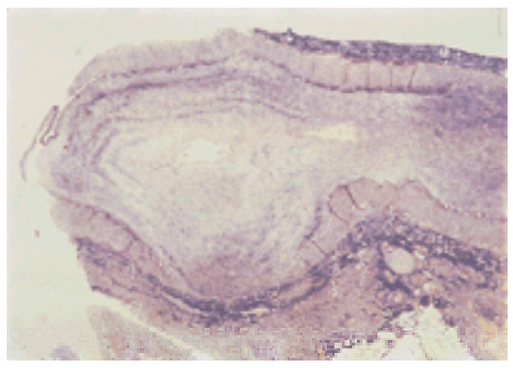

Plate 4 Morphological appearances of a renal artery from a patient with polyarteritis nodosa. The elastic lamina has been destroyed and the artery has become aneurysmal.

Chapter 20.10.4 The kidney in rheumatological disorders

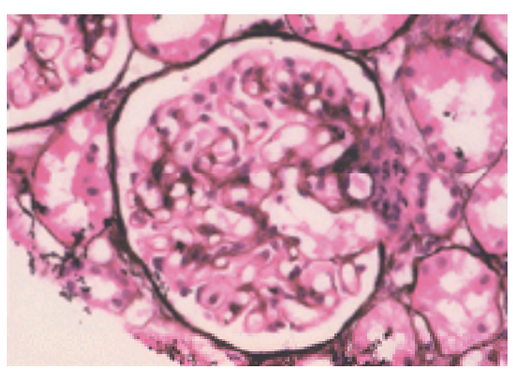

Plate 1 Lupus nephritis. The glomerulus has mild mesangial increase (WHO class II). Periodic acid-methenamine silver staining (×50). (By courtesy of Dr A.J. Howie.)

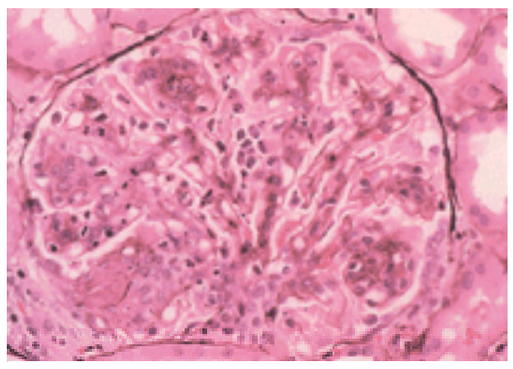

Plate 2 Lupus nephritis. The glomerulus has marked mesangial increase with wire loops, a few doubled basement membranes and segmental lesions (WHO class IV). Periodic acid-methenamine silver staining (×40). (By courtesy of Dr A.J. Howie.)

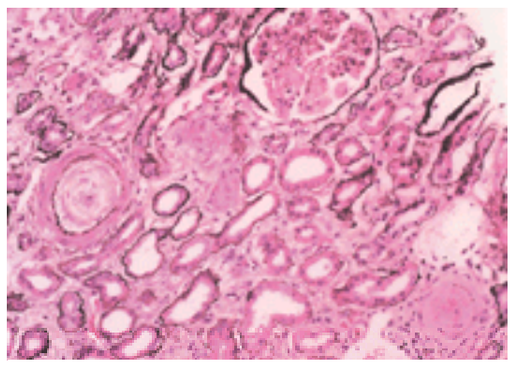

Plate 3 Scleroderma kidney. A small artery has concentric mucoid intimal thickening, an arteriole has thrombosis and fibrinoid necrosis, and tubules and a glomerulus have ischaemic damage. Periodic acid-methenamine silver staining (×25). (By courtesy of Dr A.J. Howie.)

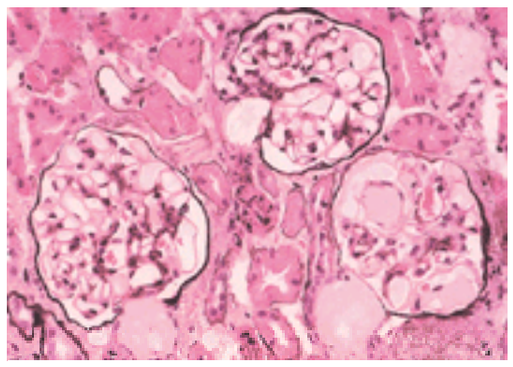

Plate 4 Amyloidosis in rheumatoid arthritis. Arterioles and glomeruli contain acellular masses of amyloid. Periodic acid-methenamine silver staining (×40). (By courtesy of Dr A.J. Howie.)

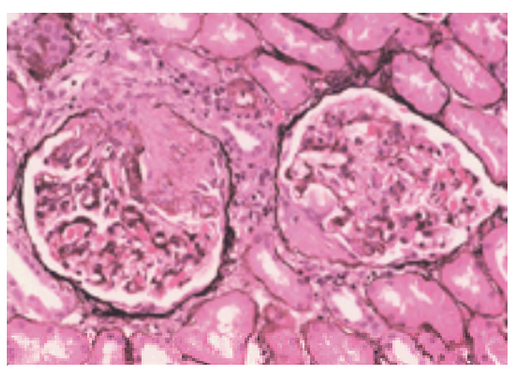

Plate 5 Vasculitic glomerulonephritis in rheumatoid arthritis. Two glomeruli have sharply defined segmental lesions where there has been disruption of the tuft and partial obliteration of Bowman's space. Periodic acid-methenamine silver staining (×32). (By courtesy of Dr A.J. Howie.)

Chapter 20.10.5 Renal involvement in plasma cell dyscrasias, amyloid and fibrillary glomerulopathies, lymphomas, and leukaemias

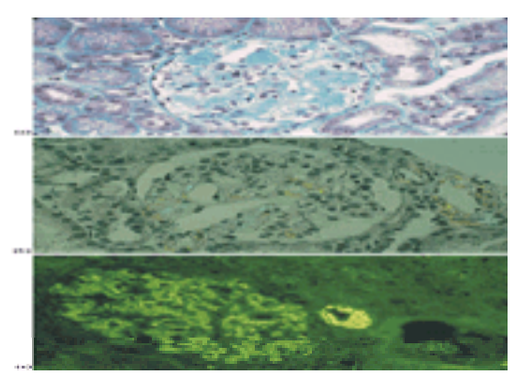

Plate 1 Light-chain amyloidosis. (a) Amyloid deposits in a renal glomerulus (Masson's trichrome stain, ×312). (b) Congo red stain. Apple-green/yellow dichroism under polarized light (×312). (c) Immunofluorescence with anti-lantibody. Note glomerular and arteriolar deposits (×312). (From Béatrice Mougenot's personal collection.)

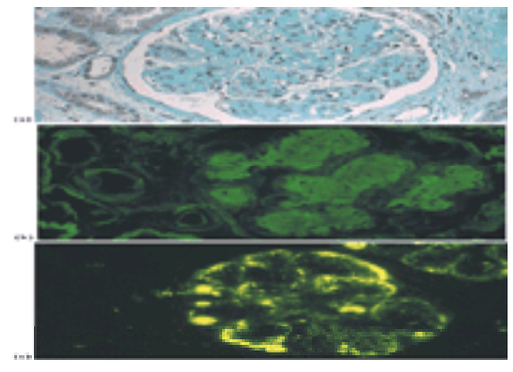

Plate 2 See *caption overleaf*

Cryoglobulinaemic glomerulonephritis. (a) The glomerulus shows a marked endocapillary hypercellularity with massive infiltration of mononuclear leucocytes (Masson's trichrome stain, ×500). (b) Frequent double-contour aspect, and intraluminal thrombi (periodic acid-Schiff stain, ×312). (c) Thrombi and segments of glomerular basement membrane are brightly stained with anti-IgM antibody (immunofluorescence, ×312). (From Béatrice Mougenot's personal collection.)

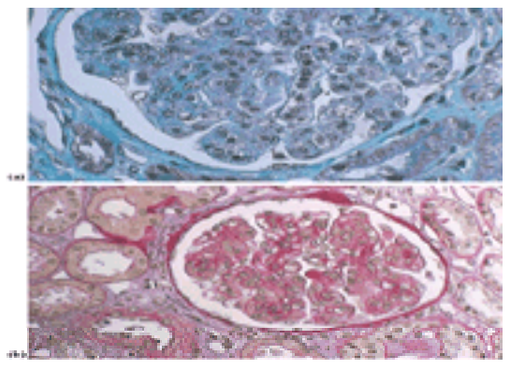

Plate 3 Monoclonal immunoglobulin deposition disease. (a) Typical nodular glomerulosclerosis. Note the membrane-like material in the centre of the nodules and nuclei at the periphery. Some glomerular capillaries show double contours. Note also thickening of the basement membrane of atrophic tubules (Masson's trichrome stain, ×312). (b) Bright staining of tubular basement membranes and mesangial nodules and, to a lesser extent, of glomerular basement membrane with anti-kantibody in a case of klight-chain deposition disease (immunofluorescence, ×312).

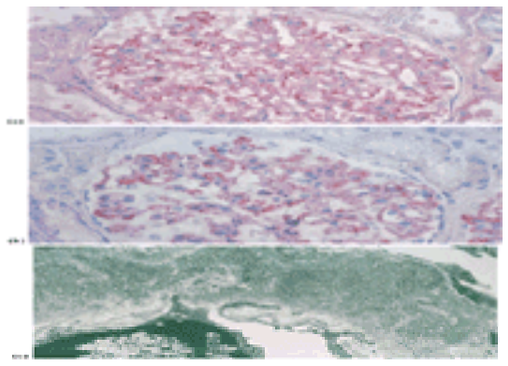

Plate 4 Immunotactoid glomerulopathy in a patient with chronic lymphocytic leukaemia. A typical membranous glomerulonephritis showing exclusive staining of the deposits with anti-g(a) and anti-k(b) antibodies (immunohistochemistry, alkaline phosphatase, × 312). (c) Electron micrograph of glomerular basement membrane, showing the microtubular structure of the subepithelial deposits (uranyl acetate and lead citrate, ×12 000). (From Béatrice Mougenot's personal collection.)

Chapter 20.10.6 Haemolytic uraemic syndrome

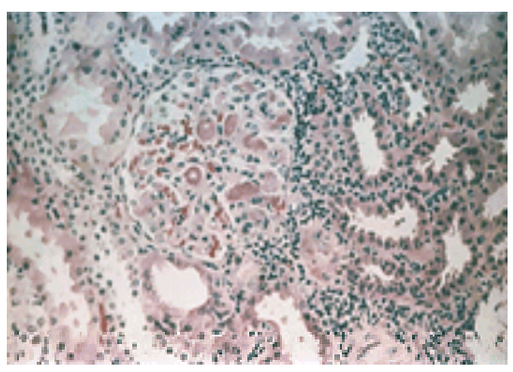

Plate 1 Typical changes of glomerular thrombotic microangiopathy in a patient with HUS (figure kindly provided by Dr Marie O'Donnell).

Plates for Section 21

Chapter 21.5 Infections and other medical problems in homosexual men

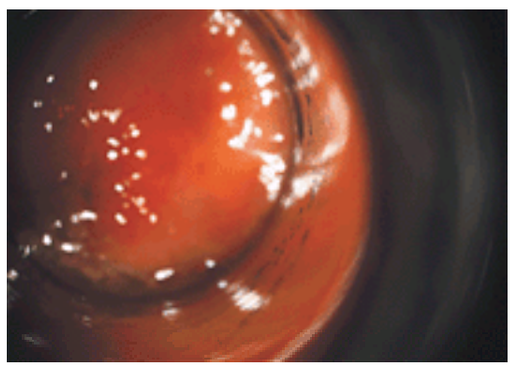

Plate 1 Gonococcal proctitis.

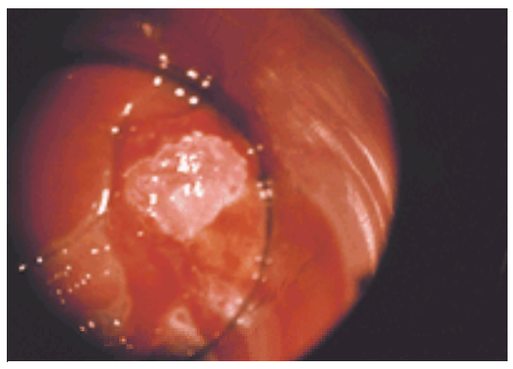

Plate 2 Condylomata acuminata of anal canal.

Chapter 22.3.4 Acute myeloblastic leukaemia

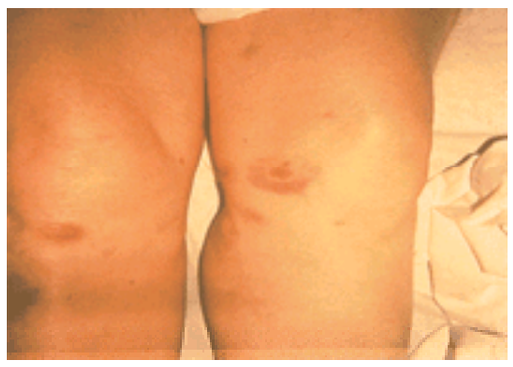

Plate 1 Target purpura. Classic target appearance of purpura formed by infarction of an arteriole by a dividing cluster of leukaemic myeloblasts. Typically, a deep, firm nodule can be felt in the pale centre of the lesion.

Chapter 22.3.5 Chronic lymphocytic leukaemia and other leukaemias of mature B and T cells

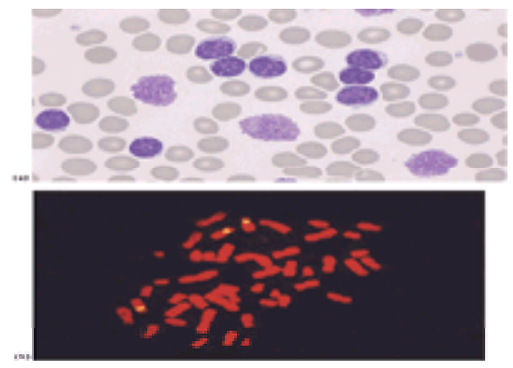

Plate 1 (a) Peripheral blood film from a case of chronic lymphocytic leukaemia showing small lymphocytes and smear cells. (b) Lymphocyte metaphase demonstrating trisomy 12 by *in situ* hybridization with a centromeric probe shown as single fluorescent dots in three chromosomes no. 12.

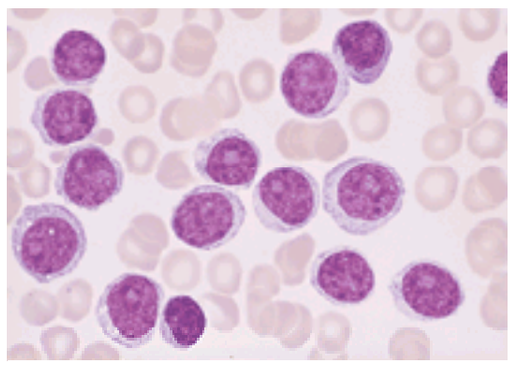

Plate 2 Peripheral blood film from a case of B-prolymphocytic leukaemia with characteristic nucleolated prolymphocytes.

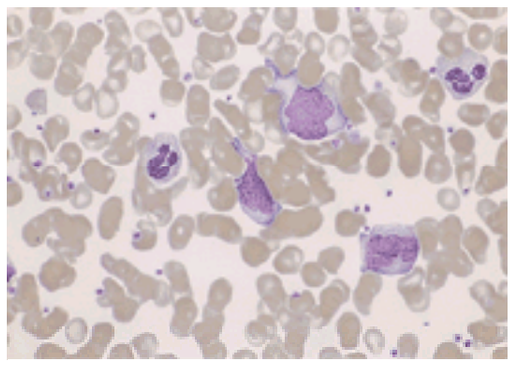

Plate 3 Circulating lymphocytes from a case of mantle-cell lymphoma.

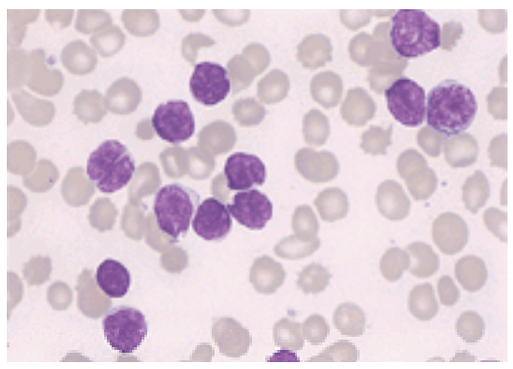

Plate 4 Peripheral blood cells from a case of folicular lymphoma presenting with leukaemia and a high leucocyte count.

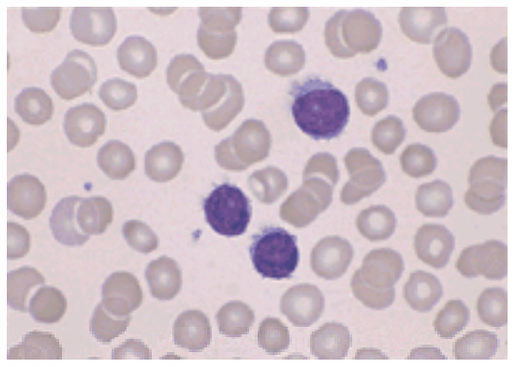

Plate 5 Peripheral blood lymphocytes with short villous projections from a case of splenic lymphoma with villous lymphocytes.

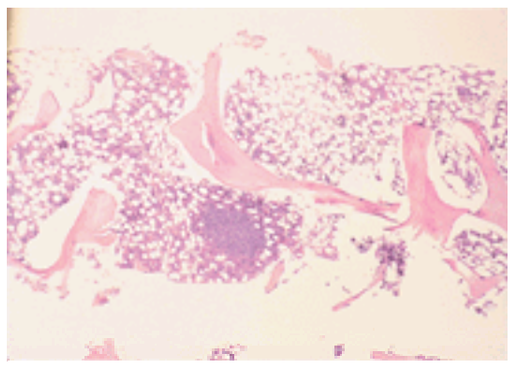

Plate 6 Nodular lymphocytic infiltration pattern in a bone marrow section from a case of mantle-cell lymphoma.

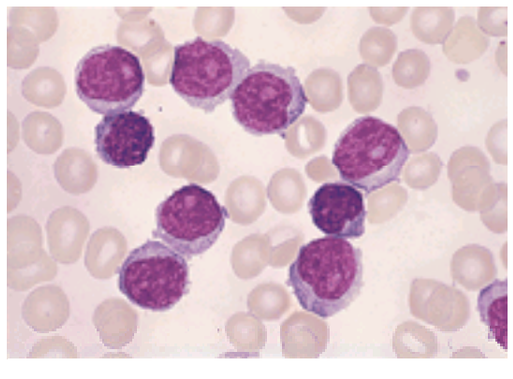

Plate 7 Peripheral blood from a case of T-cell prolymphocytic leukaemia.

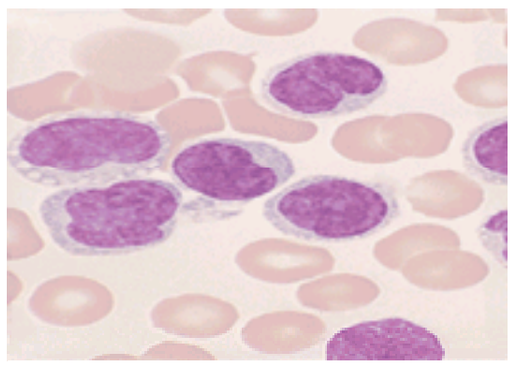

Plate 8 Circulating convoluted T cells from a Caribbean-born patient with adult T-cell leukaemia/lymphoma.

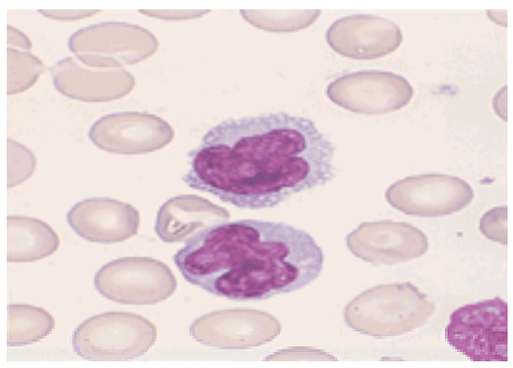

Plate 9 Cerebriform cells from a case of Sezary syndrome evolving with erythroderma and a high lymphocyte count.

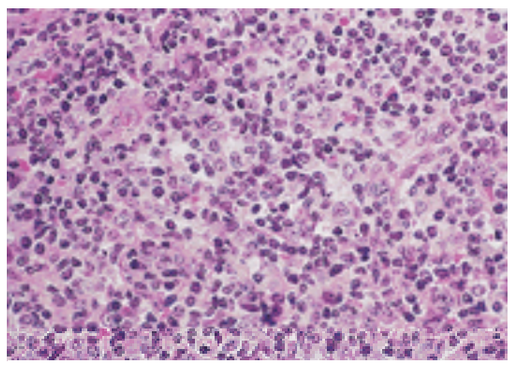

Plate 10 Lymph node section from a case of adult T-cell leukaemia/lymphoma showing diffuse infiltration with pleomorphic small, medium, and large cells.

Chapter 22.3.6 Chronic myeloid leukaemia

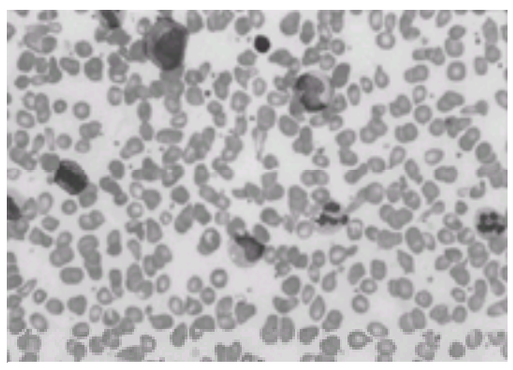

Plate 1 Peripheral blood film from a patient with CML in chronic phase. (Photograph kindly provided by Professor Barbara Bain, Imperial College London.)

Chapter 22.4.1 Leucocytes in health and disease

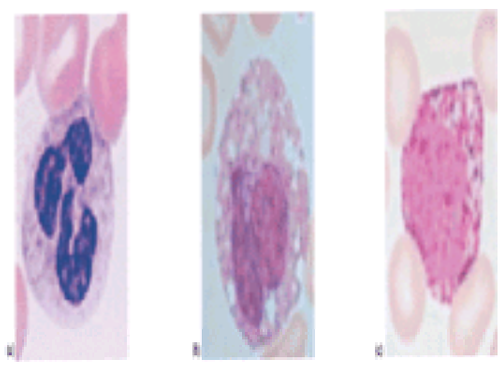

Plate 1 Peripheral blood granulocytes. (a) Polymorphonuclear leucocyte (neutrophil). (b) Eosinophil. (c) Basophil.

Chapter 22.4.7 Histiocytoses

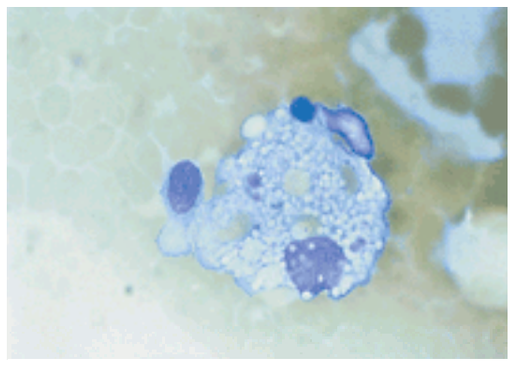

Plate 1 Macrophage exhibiting haemophagocytosis in the bone marrow of a child with haemophagocytic lumphohistiocytosis.

Chapter 22.5.6 Megaloblastic anaemia and miscellaneous deficiency anaemias

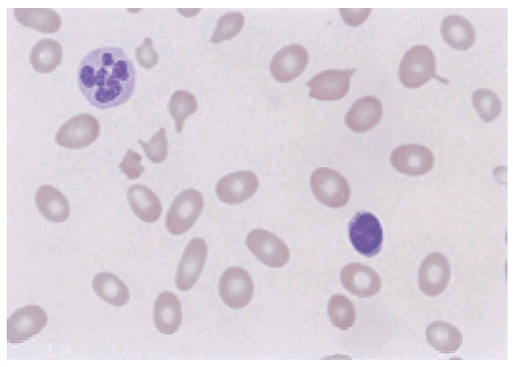

Plate 1 Megaloblastic anaemia. Hb 4.0 g/dl, MCV 120 fl. Hypersegmented neutrophil, oval macrocytes, and a small lymphocyte to show size of macrocytes. The fragmentation of advanced megaloblastosis is present. Thrombocytopenia is marked.

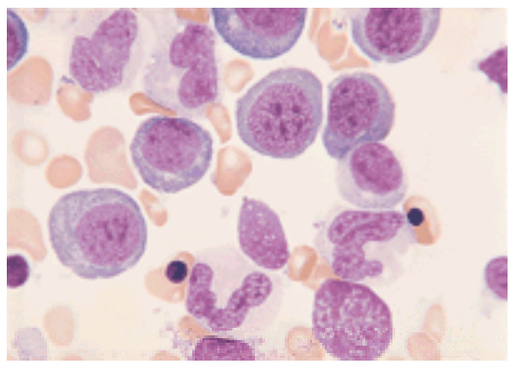

Plate 2 Megaloblastic anaemia. Bone marrow aspirate showing mainly intermediate megaloblasts and four giant metamyelocytes.

Chapter 22.6.2 Evaluation of the patient with a bleeding diathesis

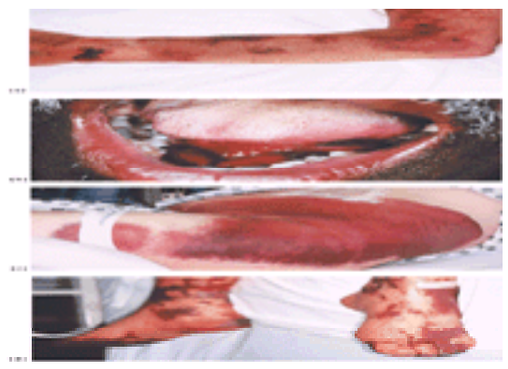

Plate 1 See *caption overleaf*.

Purpura. (a) Confluent ecchymoses of varying size and age on the upper arm in an individual with an acquired factor VIII inhibitor. (b) Spontaneous sublingual haematoma in an individual with severe haemophilia A. (c) Blunt trauma-induced left flank and hip ecchymosis and haemotoma in an individual with severe haemophilia A. (d) Gangrenous ecchymoses in an individual with diffuse intravascular coagulation.

Chapter 22.7 The blood in systemic disease

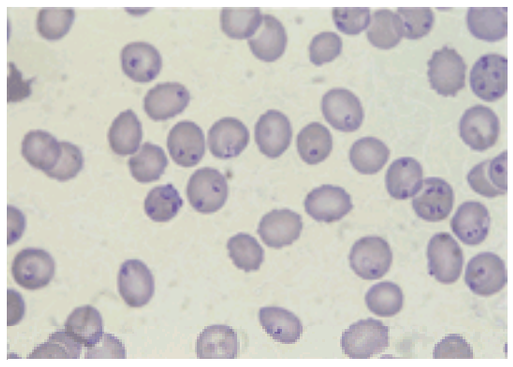

Plate 1 Malaria. Blood film showing fatal *Plasmodium falciparum* infection in a Gambian child.

Plates for Section 23

Chapter 23.1 Diseases of the skin

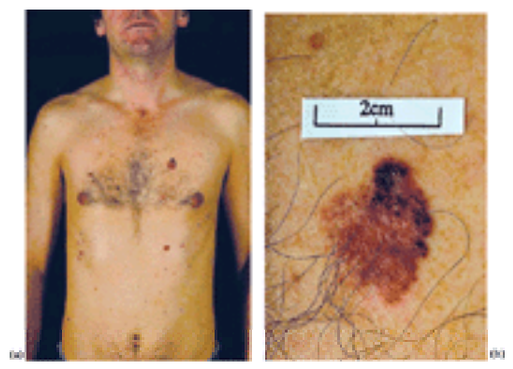

Plate 1 (a) In a patient with multiple atypical naevi one may stand out as different from the others and can be seen to be a melanoma. (b) It has an irregular outline and contains numerous different shades of brown pigmentation.

Plate 2 See *caption next page*.

(a) Most primary melanomas will have some pigmentation, even so-called amelanotic melanoma. (b) Spitz naevi were formerly called juvenile melanoma because of their histological resemblance to melanoma, but their biological behaviour is benign.

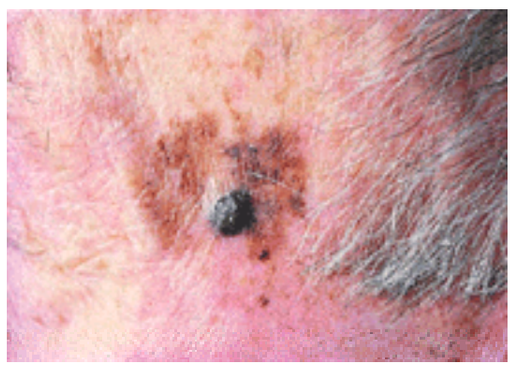

Plate 3 Nodular melanoma arising in a macular lentigo maligna (lentigo maligna melanoma).

Plate 4 Regression in a melanoma making histological assessment of prognosis impossible.

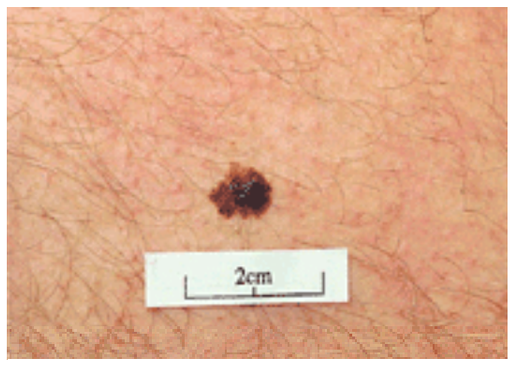

Plate 5 Melanoma most often arises *de novo* on normal skin and grows radially as well as vertically. Early detection requires identification of atypical morphology of smaller lesions.

Plate 6 Acral lentigenous melanoma can be difficult to distinguish from benign junctional naevi on the palms and soles. (a)

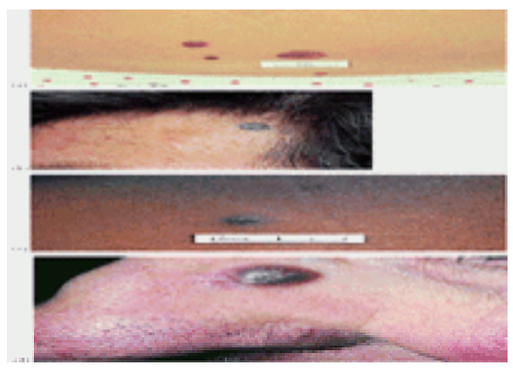

Plate 7 See *caption overleaf*.

Lesions commonly confused with melanoma include (a) naevus *en cocarde*, which are central compound naevi with a surrounding macular junctional component, giving the appearance of a fried egg. Blue naevi (b) are often deeply pigmented, but the pigmentation is uniform and a blue tinge is discernible. Dermatofibromas (c) are sometimes easier to diagnose on palpation as they are hard and tethered to the skin. They may feel like a split pea. Pigmented basal cell epitheliomas (d) often

have a rolled edge; however, sometimes biopsy provides this unexpected diagnosis. (a) (b) (c)

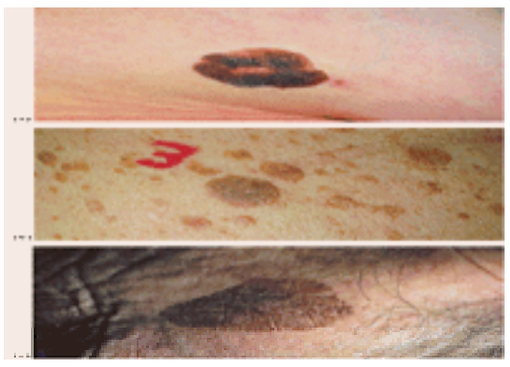

Plate 8 Seborrhoeic warts are often numerous and come in a variety of shapes and sizes. They may be deeply pigmented and elevated (a) or pale (b). They may also be macular (c). Characteristically they have a waxy surface and a 'stuck on' appearance.

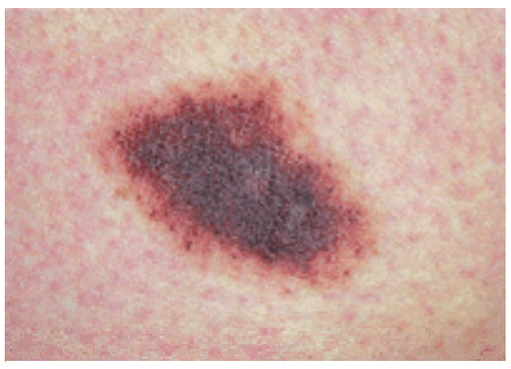

Plate 9 Congenital naevi are often larger than acquired naevi, but are usually evenly pigmented. They have a greater risk of malignant change.

Chapter 23.2 Molecular basis of inherited skin disease

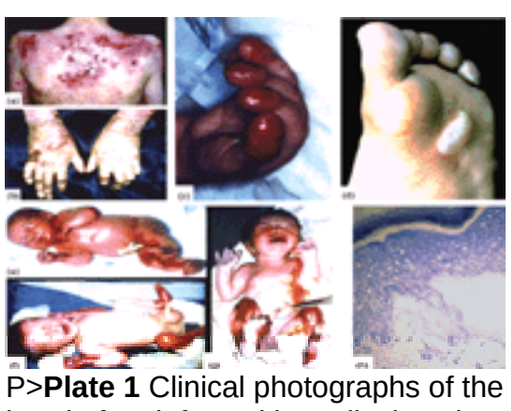

P><**Plate 1** Clinical photographs of the different forms of epidermolysis bullosa. (a) and (b) a patient with Hallopean–Siemans dystrophic epidermolysis bullosa; (c) the hand of an infant with Herlitz junctional epidermolysis bullosa; (d) blister on the foot of a patient with epidermolysis bullosa simplex; (e) baby with epidermolysis bullosa simplex Dowling–Meara; (f) baby with Herlitz junctional epidermolysis bullosa; (g) baby with Hallopean–Siemans dystrophic epidermolysis bullosa; (h) intraepidermal blister from a Weber Cockayne epidermolysis bullosa simplex patient. Skin section stained with Richardson's stain.

Plate 2 Clinical photographs of: (a) bullous ichthyosiform erythroderma (BIE) and three types of keratoderma; (b) focal palmoplantar keratoderma (PPK) associated with a keratin 16 mutation; (c) striate palmoplantar keratoderma associated with a desmoglein 1 mutation; and (d) constriction around the digit from an individual with Vohwinkel's syndrome associated with a C×26 mutation.

Plates for Section 24

Chapter 24.13.3 Epilepsy in later childhood and adults

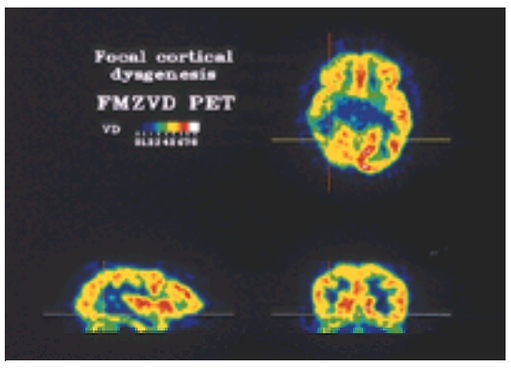

Plate 1 FMZVD PET scan showing a region of probable cortical dysplasia in the right temporal lobe. The 11C-fluamzenil volume of distribution (FMZVD) is an index of GABAa receptor density.

Chapter 24.13.9 Human prion disease

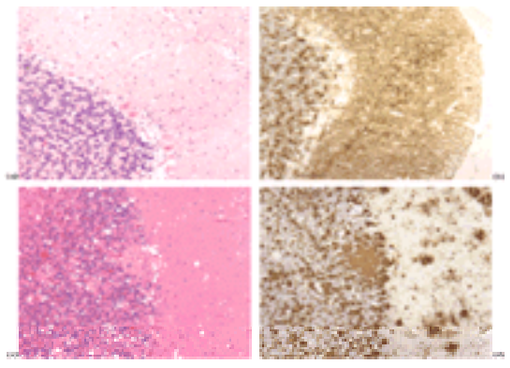

Plate 1 See *caption next page*.

The cerebellum in sporadic CJD shows widespread spongiform change in the molecular layer, with no plaques visible. Haematoxylin and eosin × 250. (b) PrP immunocytochemistry in the cerebellum in sporadic CJD shows a fine granular (synaptic) pattern of deposition in the molecular layer (right) with coarser deposits visible in the granular layer (left). No plaques are visible and the Purkinje cells are unstained. Kg9 monoclonal antibody × 250. (c) The cerebellum in variant CJD shows a group of florid plaques (centre) comprising rounded amyloid deposits surrounded by spongiform change. Spongiform change is also present in the molecular layer (right). Haematoxylin and eosin × 250. (d) Immunocytochemistry for PrP in the cerebellum in variant CJD shows strong staining of the large amyloid plaques (centre) but there is widespread positivity in the form of multiple smaller plaques, with amorphous 'feathery' deposits in the molecular layer (right). Kg9 monoclonal antibody × 250.

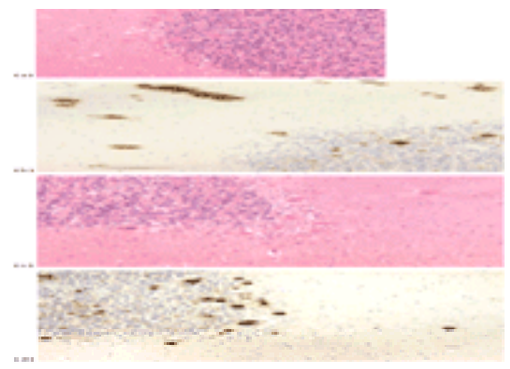

Plate 2 (a) The cerebellum in GSS contains multiple multicentric plaques (centre) which are present both in the molecular layer and in the granular layer. Spongiform change is also present focally in the molecular layer (left). Haematoxylin and eosin ×250. (b) Multicentric plaques in GSS are more easily visualized in the cerebellum using immunocytochemistry for PrP, which shows large deposits of varying size in both the molecular and granular layers. Kg9 monoclonal antibody × 250. (c) The cerebellum in kuru contains typical plaques (the so-called kuru plaques) which are comprised of a rounded structure with a dense centre and a loose fibrillary periphery (centre). Spongiform change is only present to a minimal degree in the molecular layer. Haematoxylin and eosin × 250. (d) Immunocytochemistry for PrP in the cerebellum in kuru shows strong staining of the larger plaques and in addition demonstrates multiple smaller plaques which are not evident on routinely stained preparations. Kg9 monoclonal antibody × 250.

Chapter 24.14.1 Bacterial meningitis

Plate 1 Cutaneous petechiae in a patient with acute meningococcal meningitis. (Copyright D.A. Warrell.)

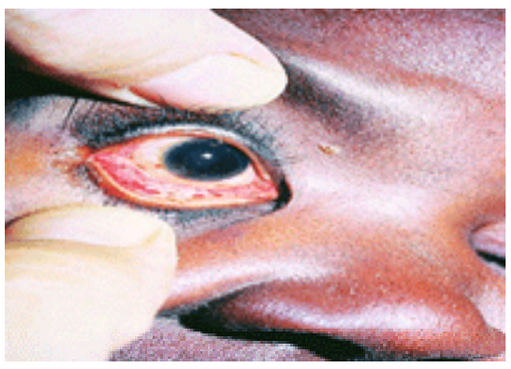

Plate 2 Conjunctival petechiae in a Nigerian boy with meningococcal meningitis. (Copyright D.A. Warrell.)

Plate 3 Haemorrhagic lesions on the face (a) and shin (b) of a 63-year-old Thai man with *Streptococcus suis* meningitis. (Copyright the late Prida Phuapradit.)

Plate 4 The rash of meningococcal septicaemia in an English child.

Plate 5 Healing vasculitic rash in a Brazilian boy with meningococcal meningitis and meningococcaemia. (Copyright D.A. Warrell.)

Plate 6 Septic arthritis of the interphalangeal joints in a 73-year-old Thai man with *Streptococcal suis* meningitis. (Copyright the late Prida Phuapradit.)

Plate 7 Tuberculoma in the brain. (Copyright Gareth Turner.)

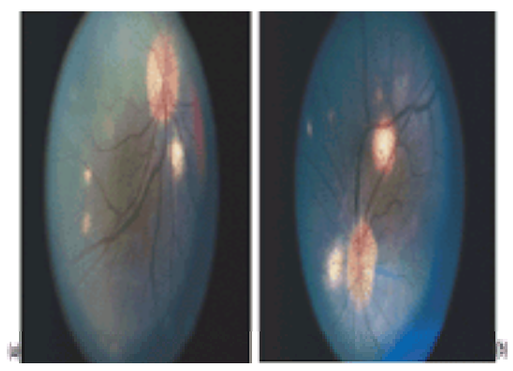

Plate 8 (a and b) Tuberculous choroiditis in a 23-year-old Thai woman. (Copyright the late Prida Phuapradit.)

Plates for Section 25

Chapter 25 The eye in general medicine

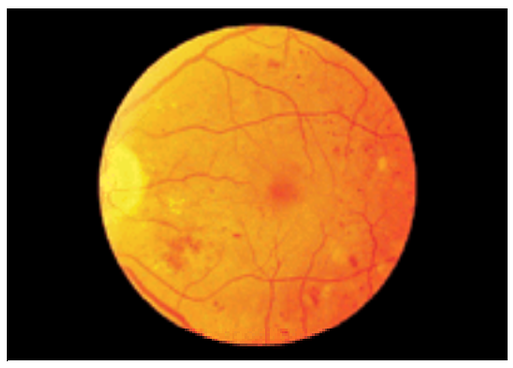

Plate 1 Diabetic, background retinopathy. The hallmarks of background retinal changes are red dots (either microaneurysms or small haemorrhages) and blots (larger haemorrhages) together with glinting hard exudates and. These are no closer than one disc diameter from the central fovea and vision is normal.

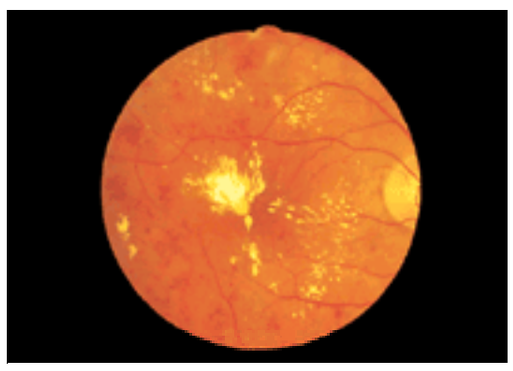

Plate 2 Diabetic, maculopathy. Hard exudate, containing lipid and protein which has leaked from damaged retinal capillaries, has congregated at the fovea. Central vision is irretrievably impaired. Diabetes may present in this way, especially in the elderly.

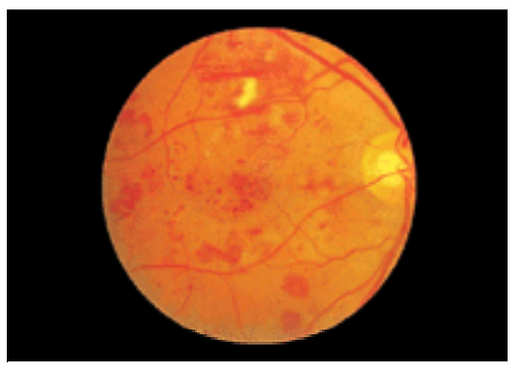

Plate 3 Diabetic, ischaemic retinopathy. Capillary ischaemia creates multiple cotton wool spots—microinfarcts within the nerve fibre layer. Other features are dilatation of retinal veins and multiple blot haemorrhages. Frank proliferation of new vessels is almost inevitable and the retinal changes must be carefully observed.

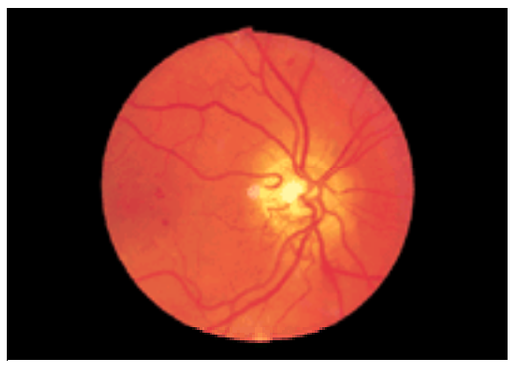

Plate 4 Diabetic, proliferative retinopathy. New vessels have formed on the inferior part of the optic disc. They are fine, looping, and aimless. There may be others in the peripheral retina. If the vessels bleed, vision will become acutely obscured by 'floaters'.

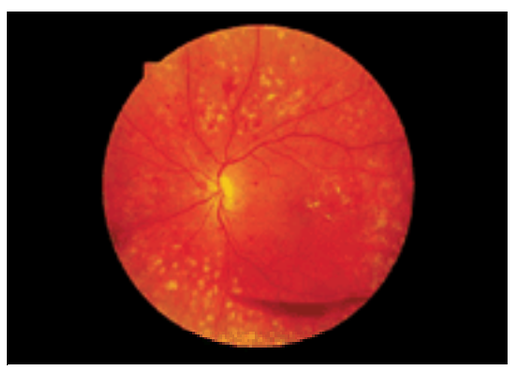

Plate 5 Diabetic, preretinal haemorrhage and laser scars. Neovascular fronds may bleed in front of the retina or into the vitreous, obscuring vision acutely. Here blood

has sedimented into a characteristic 'boat' shape and multiple laser scars have been placed outside the major vascular arcades. There are haemorrhages and hard exudate temporal to the fovea.

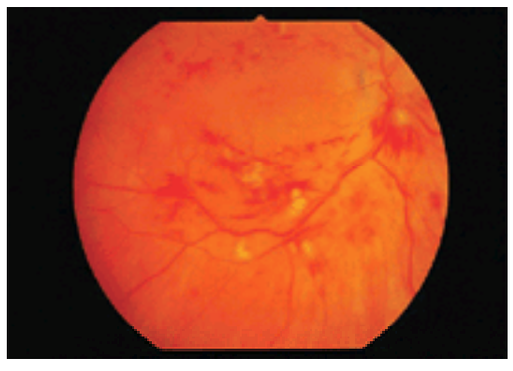

Plate 6 Hypertension, accelerated. Multiple flame shaped haemorrhages, microinfarcts, and swelling of the optic disc margin are characteristic features of accelerated hypertension. Vision may be normal, yet the changes dictate immediate treatment to reduce blood pressure. The diastolic level is usually greater than 110 mmHg and proteinuria is to be expected.

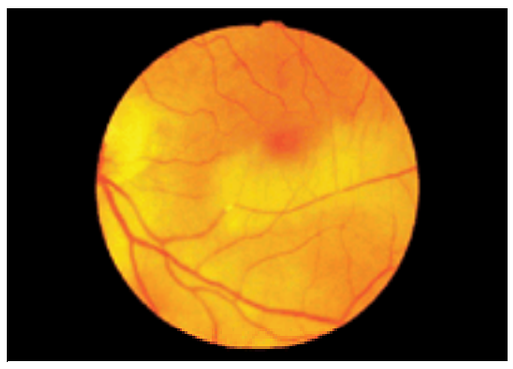

Plate 7 Branch retinal artery occlusion. A small white embolus is lodged at the third bifurcation of the superotemporal branch retinal artery, occluding it. The local retina is oedematous and non-functioning, producing an acute superior scotoma in the left eye. The likely origin is from the ipsilateral carotid or the heart.

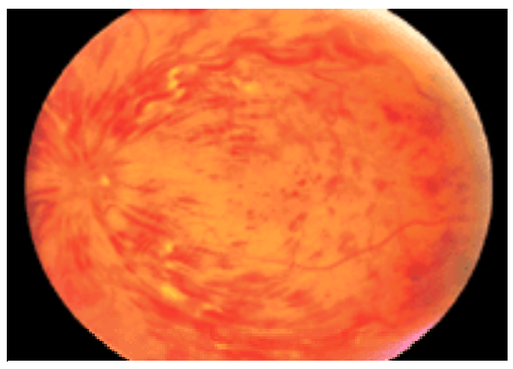

Plate 8 Central retinal vein occlusion. Blockage of the draining central retinal vein results in a 'bloodstorm' appearance with profuse flame haemorrhages forming between the nerve fibres in all quadrants. Cotton wool spots representing microinfarcts are often also present. Vision is acutely blurred as the fovea becomes oedematous.

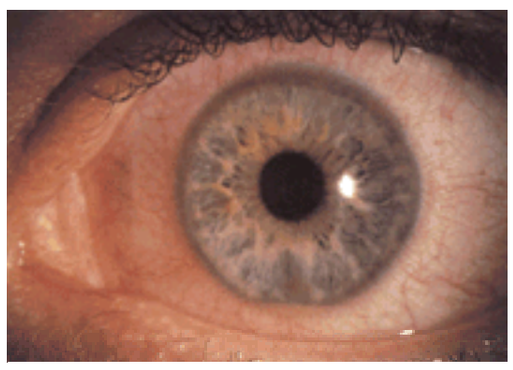

Plate 9 Behçet's hypopyon iritis. The eye is red, painful and photophobic. White cells within the anterior chamber have sedimented into a characteristic hypopyon at the base. If bacterial endophthalmitis is excluded, Behçet's syndrome is a likely cause of this acute, intense, sterile iritis.

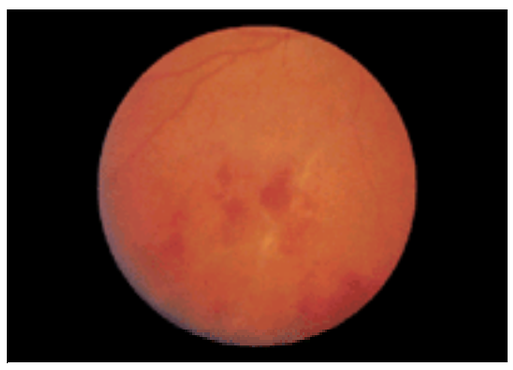

Plate 10 Behçet's retinitis. Occlusion of blood vessels, usually venous, in the peripheral retina produces a wedge of haemorrhage with whitening of the vascular wall. The view is hazy due to inflammatory cells within the vitreous. The retina is ischaemic, function is lost, and neovascularization may occur. Repeated episodes may damage vision irretrievably.

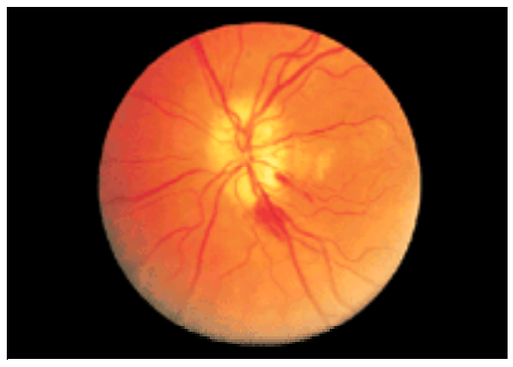

Plate 11 Giant cell arteritis, optic disc infarction. The optic nerve head is infarcted, due to occlusion of multiple ciliary branch arterioles which supply it. The disc is pale and swollen, and juctapapillary haemorrhage has formed. Vision is poor and will not recover. The other eye is at immediate risk unless the systemic inflammatory process is controlled.

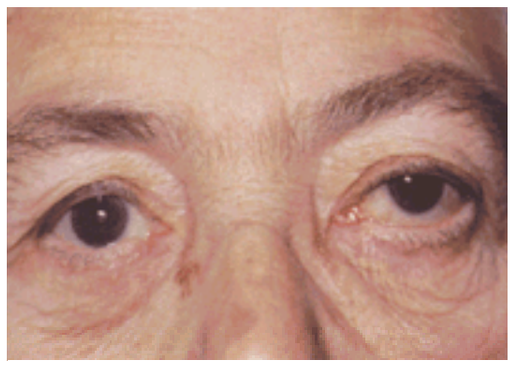

Plate 12 Wegener's granulomatosis of the orbit. An inflammatory mass behind the left eye has displaced it forwards and upwards and the eye moves poorly due to involvement of motor nerves within the orbit. The optic nerve may also be involved. Biopsy confirmed granulomatous vasculitis and ANCA was positive. The adjacent sinuses were involved, with bone loss demonstrated on CT scan.

Plate 13 Scleromalacia in rheumatoid arthritis. Vasculitis results in focal ischaemia, with translucency and thinning of the sclera: the coat of the eye may perforate. The most common associated systemic disorder is rheumatoid arthritis. The eye is usually red, and pain may be intense.

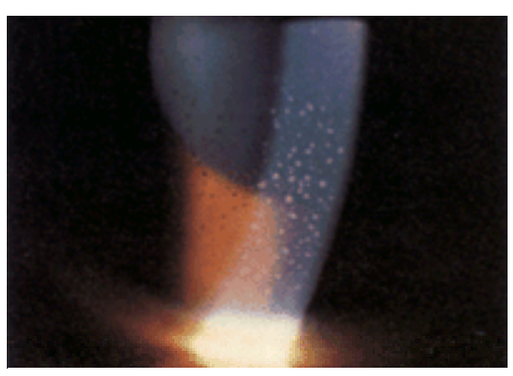

Plate 14 Iritis in ankylosing spondylitis. The slit lamp displays cells within the anterior chamber which have sedimented on to the interior surface of the cornea as white keratic precipitates. These are the hallmarks of iritis (anterior uveitis). The eye is usually red and painful. A frequent association is with ankylosing spondylitis and HLA B27 haplotype.

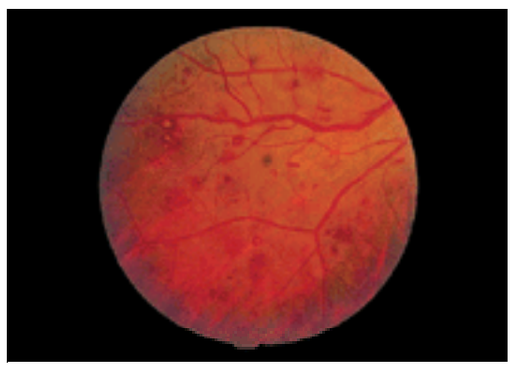

Plate 15 Retinal haemorrhages in leukaemia. Multiple and bilateral retinal haemorrhages suggest a blood dyscrasia, if underlying diabetes and hypertension are excluded. In this case, the peripheral lymphocyte count was considerably raised, consistent with chronic lymphocytic leukaemia. Some haemorrhages have a white centre (Roth spot).

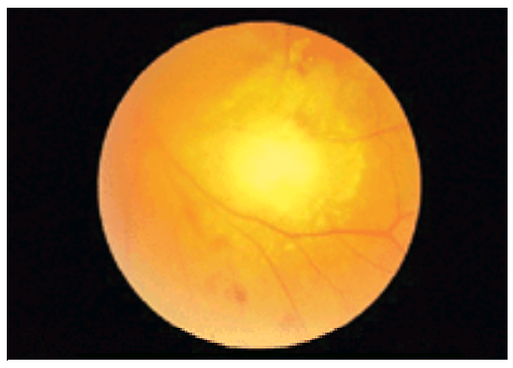

Plate 16 Metastatic staphylococcal endophthalmitis. Blood borne organisms may settle in the eye, forming a focal abscess in the choroid, breaking through the adjacent retina into the vitreous which becomes hazy with inflammatory cells. This patient had poorly-controlled diabetes and a staphylococcal skin infection.

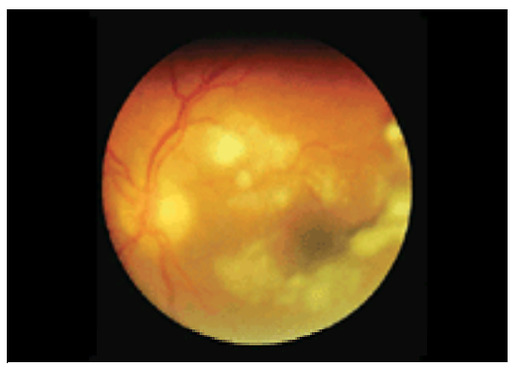

Plate 17 Candida endophthalmitis. Fungal infection of the eye interior forms white 'snowballs' within the vitreous and retina. The organism is usually blood borne and may enter the circulation with intravenously injected agents, including heroin. Infection is indolent, with a relatively white eye and little pain. Vitrectomy and intravitreal antimicrobial treatment may be necessary.

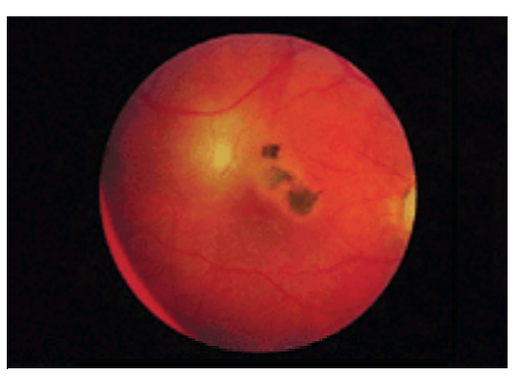

Plate 18 Toxoplasmosis. A fluffy fresh focus of infection within the choroid and retina is found adjacent to an old pigmented scar, typical of toxoplasmosis. The organism encysts within the retina and may reactivate sporadically in this way.

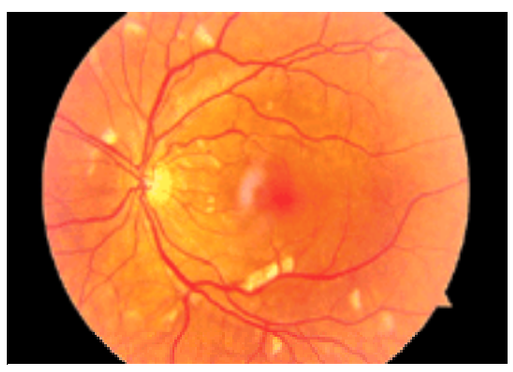

Plate 19 Cotton wool spots. Retinal microinfarcts are due to occlusion of capillaries which supply the nerve fibre layer. These multiple 'cotton wool spots' are found associated with microemboli, as after cardiac surgery employing bypass. In patients with AIDS they may form especially at the time of pulmonary infection, for instance with *Pneumocystis carinii*.

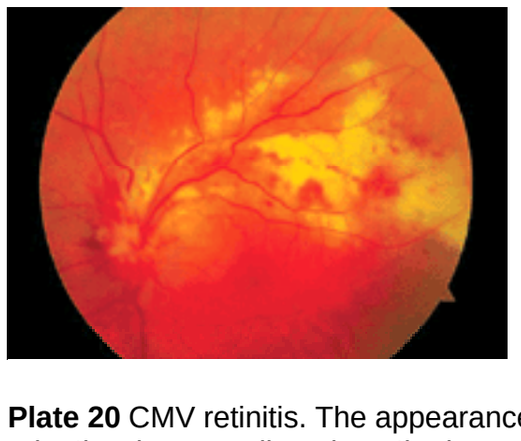

Plate 20 CMV retinitis. The appearance of focal, fluffy, pale retinal necrosis with haemorrhages is characteristic of infection with cytomegalovirus. The area expands relentlessly, spreading along the branch vessels, unless treatment with virustatic agent is instituted or the CD4 lymphocyte count can be improved. The usual underlying disorder is AIDS.

Plate 21 Thyroid eye disease with exophthalmos. Inflammation of orbital tissues— fat and muscles—causes protrustion of the eye—exophthalmos or proptosis. The eyelids also become swollen and the conjunctiva congested. Autoimmune thyroid disease (Graves' disease) is the most common underlying disorder.

Plate 22 Marfan's syndrome. Dislocation of the lens is sometimes easily visible, though lesser degrees may need careful examination using the slit lamp after dilatation of the pupil. The lens may also be unstable, trembling on eye movement. The most common underlying cause is Marfan's syndrome, with deficiency of fibrillin in the suspensory fibres and upward displacement.

Plate 23 Von Hippel Lindau. Angiomas of the retina are an important early feature of this dominantly inherited condition. They begin as small red lesions which expand. Here the angioma is next to the optic disc, a characteristic position which makes management difficult and visual prognosis poor.

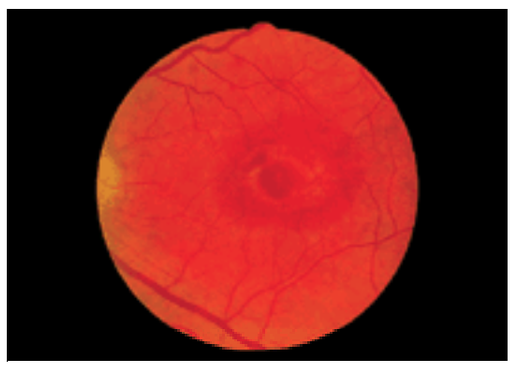

Plate 24 Bull's eye maculopathy. Toxicity at the macula caused by chloroquine results in a concentric target-like pigmentary appearance. The features are reversible in the early, asymptomatic, stages, but once loss of central visual acuity occurs, this may progress despite stopping the drug. Hydroxychloroquine appears to carry a lesser risk.

1 On being a patient

C. M. Clothier

'Patior'—'I am suffering'—one recalls being taught at school. Every student knows that a patient is generally, but not always, one who is suffering, and perhaps not very patiently. Only by derivation has the word become associated with the bearing without complaint, of pain, sorrow, or simple irritation with others. Yet this secondary meaning is important in medicine because it reminds us that the great majority of patients do in fact suffer their illnesses with remarkable fortitude and endurance. Might this be another facet of the urge to survive and to minimize disability in the essentially competitive struggle of life? Is it this instinct, perhaps, which leads the average healthy person who stumbles and falls in the street to declare: 'It's nothing really' or 'I'll be all right in a minute', as they try to rise quickly from the fallen position of the vanquished? For falling to the ground in combat is usually fatal, and voluntary prostration of the body usually signifies submission and defeat.

So when otherwise healthy persons seek help from a doctor, or are admitted to hospital, it may be assumed that they are in some perturbation of mind as well as of body. It is this suffering in the mind that makes every patient different, even when their condition is familiar and well recognized, because the mental element in any illness varies enormously with circumstance and temperament.

Visiting the family doctor is usually less of an ordeal for the patient than admission to hospital. The patient often retains the dignity of an upright position *vis-à-vis* the doctor. However, the element of personal distress may still be there and may mask or distort the objective signs and symptoms for which the doctor is trained to look. The 'dependent well' may only be seeking a listening post for the torments of family dissension, personal tragedy, or just old age. Other patients enjoy being examined, manipulated, or injected because of the personal contact and attention involved, which they otherwise lack. How much time to give to each of the greatly diverse patients of a practice or a hospital is a matter of delicate judgement: most will deserve a sympathetic response, but some will seem to merit a rather positive rejection. Yet one must be careful; for there have been recorded instances of patients who complained of bizarre symptoms, signifying no identifiable illness, who were rejected as malingerers but subsequently were shown to have suffered greatly from insidious disease or poison. Those who are truly ill may be brave, or craven, or something in between, depending upon age, circumstance, or personal quality. However, one thing is certain, each patient will think himself or herself to be reacting normally to the predicament of illness, real or imaginary, and will be expecting the same degree of medical attention. If they do not get it they will be offended and angry: whether that reaction should provoke from the doctor sympathy or dismissal may be as crucial a judgement as writing a prescription.

On the other hand, admission to hospital is a fundamentally different experience both for patients and those caring for them. Here the dominant factors are the concession of defeat and the abandonment of the safe haven of home. To lie down in the presence of others in a strange place, to get undressed in the middle of the day, to give specimens of urine or blood, all on the orders of those who remain upright, are acts of submission. The arrival in hospital is accompanied by feelings of anxious apprehension, fear, isolation, and general mental turmoil which caused Osbert Sitwell to speak of 'First depressions on arrival'. Mingled with these feelings may also be some sense of relief in having finally admitted defeat and agreed to surrender one's body into other hands. When one adds to these varied emotions anxiety about work, about the home, or about the patient's spouse or children, it is a wonder that the admitting doctor gets any sensible or accurate answer out of a patient for the first 24 hours after admission. When histories are taken, these widely variable responses to stress, for the most part concealed, should be in the doctor's mind. Even the lay patient is likely to be aware that blood pressure and heart rate are raised when first they are measured in hospital. Symptoms may be minimized, exaggerated consciously or unconsciously, invented, or simply forgotten. Dates and times of episodes or onsets may be so wide of the mark as to be of little use in diagnosis, if not positively dangerous.

The risk, then, is that the patient in hospital becomes merely an interesting focus of medical attention rather than a person with all those confusions of mind that make him or her an individual. This is a very real danger in modern hospital practice, the more so as we are now equipped with the most sophisticated apparatus for diagnosis, facilities that may be seen by some as reducing the need for any contribution from the patient. The feeling of personal unimportance is a marked cause of unhappiness among patients in hospital and a frequent source of complaint against doctors and nurses. Doctors can and sometimes do speak and look as if the patient was no more than an interesting clinical object and nurses can exacerbate the grievance by discussing their private affairs across the bed of the patient to whom they are attending.

In teaching hospitals, the patient's sense of unimportance may be increased by the ordeal of the professional ward round. It has to be stated that even now, at the start of the twenty-second century, there are consultants who persistently treat patients as the fortunate recipients of their attentions, whose views and feelings are of little relevance in the pursuit of a learned profession. Besides being a technical breach of obligation under the terms of a consultant's National Health Service contract, such conduct is rude. A good doctor, who engages the patients' attention and participation in a discussion of their case, may learn a great deal, besides giving the patient immense satisfaction and inspiring confidence. 'Encourage the patient to talk', said a wise old practitioner, 'and he will eventually tell you what is the matter with him'. It is regrettable that those who practise an excessive clinical detachment are not only the old and authoritarian, but include those who have grown up in a world where medical omniscience is no longer taken for granted. Doctors who cannot naturally feel a surge of sympathy for the body prostrated before them should perhaps consider a career in research.

The patients of today are very different from those of half a century ago. Until the advent of the mass media, the relationship between patient and doctor resembled that between parishioner and pastor, schoolmaster and pupil, or lawyer and client. It was impious to question the wisdom or judgement of the learned professional adviser. The remarkable expression 'sapiential authority' sought to encapsulate this ascendancy of doctor over patient. Such a relationship was often quite a happy one for both parties, and perhaps more conducive to treatment and cure than a less trusting one. However, the mass media, and supremely, television, have changed all that. We have all penetrated behind the camera into the operating theatre and other private places and there seen what doctors and their assistants actually do in the attempt to cure illness or repair damage. However skilful and ingenious, it is obviously not miraculous, and some of it is rather pedestrian. The curtains have been parted and the magic revealed. The magicians themselves have often admitted their humanity and confessed to their failures. When mystery is dispelled, the questions come thick and fast.

It is of no use to resist the tide of doubt and curiosity that now threatens to overwhelm not just medicine but all the learned professions. Family doctors must now expect patients to ask quite penetrating questions about the treatment proposed for them and the drugs it is intended to use. These questions must be answered with some candour if mutual trust is to be maintained. Nothing more disturbs a patient, who may be very intelligent even though not learned in medicine, than hearing their doctor seeking to disguise his own ignorance or doubt by prevarication or deviousness, often easily detected. Besides all of which, it sometimes does professionals in any discipline a great deal of good to be closely questioned about beliefs and practices that they have long held to be unassailably correct; and they should listen to their own explanations and audit them for intelligibility and rationality.

For almost every patient, the general anaesthetic and surgical operation engender particular anxieties which must be recognized and accommodated. It is no small thing to surrender one's consciousness into the hands of others, with all the vulnerability which unconsciousness brings. The intrusion of hands and instruments into the previously intact body equally induces dark fears in many minds. Doctors may not fully appreciate these anxieties, familiar as they are with the procedures used and their general safety and success. Most patients of reasonably resolute temperament face up to these prospects with good enough courage: but having done so, find a postponement of the day especially demoralizing. It is important that surgeons try to arrange lists so that patients do not wait many hours in a state of some tension, only to be told that they must face it all again on another day.

Modern practice and health economics combine nowadays to reduce the patient's stay in hospital to a minimum. Perhaps doctors do not sufficiently realize how much dependence may develop between patients who have been really ill and those who have rescued them from suffering, or even from death, and subsequently cared for them. The cheerful words 'Well, you can go home tomorrow' are not invariably greeted with joy. It is not merely that for some, the attention they receive in hospital is better than anything they get elsewhere: those who have been really ill are often haunted by the prospect of relapse or recurrence and feel a security in hospital, amplified by care and kindness, which they cannot feel at home. Some introductory words of sympathy for the patient's anxiety and reassurance for their ability to survive outside the hospital are often necessary.

An essay entitled 'On being a patient' ought to contain not merely adjurations for doctors but at least some directions for patients. The foremost of these could be to remember the meaning of the newly acquired status and title. Many patients are irritable and demanding, even when those attributes are not produced or justified by some disease process. They are very unattractive qualities in those who have, after all, been obliged to submit themselves to the care and skill of their fellow beings, sometimes through their own fault or neglect. A little humility and gratitude seem called for and no less because the patient is paying for some part of the services rendered, or believes that he has already done so. It is perhaps one of the least likeable of human attitudes to believe that money buys everything and that plenty of it

entitles one to special care and attention. Doctors and nurses do not for the most part do what they do in the expectation of great worldly reward. Patients likewise should recognize and appreciate human kindness when they see it and be grateful for it whether or not they are paying for their treatment.

Finally it is necessary to reflect that, as man is a social animal, the illness of any member of a family affects most of the others. Obviously enough, the spouse of a sick person is liable to be deeply affected by sorrow and by anxiety about the future. Such feelings spring not merely from love and affection but from fear of a future either robbed of economic support or burdened by care for an invalid or disabled person. So a good doctor has more than just the patient to consider and should try to speak to the relatives and to offer them proper and helpful explanations of present treatment and future prospects. It may seem unreasonable to suggest that doctors should treat the relatives of their patients as well as the patients themselves, but human beings are highly interactive and sick people are sensitive to the sorrows and anxieties of their families. If a patient's relatives are much cast down and obviously anxious, this is perceived by the patient and greatly affects morale and the peace of mind that is conducive to recovery.

It is only too tempting to avoid the patient's relatives. Besides being upright and healthy, in contradistinction to the patient, their sorrows and anxieties may make them noisy and demanding. They are apt to ask questions which seem absurd to the doctor and impossible to answer in simple lay terms. However, the effort must be made, not only for humanitarian reasons but because it may rightly be regarded as part of the treatment.

In sum, the patient views those who care for him or her as being in a relationship every bit as confidential and trusting as that which exists within the patient's family, probably more so. The patient has no hesitation in imposing this burdensome connection on one who has hitherto been a total stranger. That is the enormous measure of a doctor's voluntarily assumed responsibilities to the human race.

2.1 Science in medicine: when, how, and what

W. F. Bynum

Introduction

At least since the Hippocratics, medicine has always aspired to be scientific. What has changed is not so much the aspirations but what it has meant to be 'scientific'.

'Science is the father of knowledge, but opinion breeds ignorance', opined the Hippocratic treatise *The Canon*, and Hippocratic practitioners developed an approach to health, disease, and its treatment based on systematic observation and cumulative experience. Even the word physic, the root of physician as well as physicist, derives from the Greek for 'nature'. Further, Hippocratic medicine was experimental, that word stemming from the same classical roots which gave us 'experience'.

Words, however, can be slippery, as philosophers as divergent as Francis Bacon and Ludwig Wittgenstein have stressed. The science and experiment of the Hippocratics can still inspire, but they are not our science and experiment. During the past two or three centuries, an armoury of sciences and technologies has come to underpin medical practice. This essay attempts briefly to describe these, within the context of distinctive and perennial features of medical practice, that is, suffering individuals whose problems and diseases demand attention.

A typology of historical medicine

The late Erwin Ackerknecht always taught that the history of Western medicine revealed five kinds of medicine: bedside, library, hospital, social, and laboratory. Bedside medicine he equated with Hippocratic, with its emphasis on the individual patient, its tendency towards holism, and a concern with the patient within his or her environment. These are some of the reasons why the Hippocratics are still claimed by both orthodox and alternative practitioners. For Ackerknecht, 'library' medicine dominated in the Middle Ages, when learned medicine retreated into the universities and scholars sometimes assumed that everything worth discovering had been discovered by the ancients, and everything worth being revealed could be found in the Bible. The millennium between the sacking of Rome and the discovery of the New World is often dismissed as a sterile period scientifically, but the physicians of the period, linguistically erudite and philosophically inclined, would have been surprised to be described as unscientific. They simply believed that the road to knowledge was through the book.

They also sometimes engaged with nature, although it is undeniable that nature rather than words became an increasing source of truth and knowledge during the Scientific Revolution, a period stretching roughly from just before Andreas Vesalius (1514–1564) to Isaac Newton (1642–1727). Around 1600, it was becoming apparent to many that the Greeks had not left behind a complete and accurate account of the nature of the world, and that scientific knowledge was cumulative. This 'Battle of the Books', the debate over whether the ancients or the moderns knew more, was decided in favour of the moderns. Many of the outstanding scientific achievements of the era were in astronomy and physics, but medicine, both in its theory and its practice, was also affected. Theory has always been easier to change than practice, of course, and it was famously remarked that William Harvey's discovery of the circulation of the blood had no impact on therapeutics. Harvey (1578–1657) also notoriously lamented that his practice actually fell off following the discovery, his patients fearing that he was 'crack-brained'. The fear that too close an identification with science was detrimental to patient confidence recurs in medical history, and is still part of the delicate negotiations between the profession and its public.

Within the discipline of medicine itself there have always been individuals, some of them, like Thomas Sydenham (1624–1689), eminently successful, who believed that experimental science had little to offer to patient care. But these 'artists' of medicine could still invoke the authority of Hippocrates, with its older connotations of knowledge and experience, and during the early modern period, the whole spectrum of the sciences—mathematics, physics, chemistry, the life sciences (not yet called biology)—made their ways into formulations of health and disease. Iatrophysics, iatromathematics, and iatrochemistry all had their advocates in the seventeenth and eighteenth centuries.

That these systems tended to encourage speculation to run ahead of evidence was recognized at the time, and this was part of the reason why 'hospital medicine' had little recourse to those disciplines we now call 'basic medical sciences'. The founders of French hospital medicine, Xavier Bichat (1771–1802), J. N. Corvisart (1755–1821), and R. T. H. Laennec (1781–1826), often referred to chemistry, physiology, and the like as sciences 'accessory' to medicine. The medicine that developed in the Paris hospitals, after the reopening in 1794 of the medical schools closed by the Revolution a couple of years earlier, emphasized above all the study of disease in the sick patient. In a sense, this was Hippocratic medicine writ large, but with some significant differences. First, the hospital offered the curious doctor a vast arena for observing disease. The equivalent of a lifetime's experience of a lone practitioner in the community could be experienced in a few months of hospital work. Hospitals offered the possibility of defining disease based on hundreds of cases. Second, Hippocratic humoralism gradually disappeared as the dominant explanatory framework of health and disease, replaced by the primacy of the lesion, located in the solids—the organs, tissues and, by mid-century, cells. In this new orientation, disease was literally palpable, its lesions to be discovered in life by the systematic use of physical examination—Corvisart rediscovered percussion, Laennec invented the stethoscope—and these findings to be correlated after death by routine autopsy. French high priests of hospital medicine brought diagnosis to a new stage and replaced the older symptom-based nosologies with a more objective, demonstrable one of lesions. The third feature of hospital medicine was what Pierre Louis called the numerical method, the use of numbers to guide both disease classification and therapeutic evaluation.

The philosophy underlying early nineteenth-century French medicine was most systematically expounded by one of the many American students who studied in Paris, Elisha Bartlett, in his *Philosophy of medical science* (1844). The medical science whose philosophy he chronicled was one of facts. All systems of medicine, past and present, were speculative, vague, and useless. Cullen, Brown, Broussais, and Hahnemann were all consigned to the historical dustbin. The new medicine was one of systematic observation and collection of facts, which, properly compared and organized, could provide an objective understanding of disease and a rational basis for its treatment. Bartlett's philosophy was essentially undiluted Baconian inductivism applied to medicine. Unsurprisingly, he counted Hippocrates as well as Pierre Louis among his heroes.

One consequence of the lesion-based medicine was the recognition that not much of what doctors did actually altered the natural history of disease. Therapeutic scepticism, or even nihilism, flourished among doctors whose lives were spent, as Laennec put it, 'among the dead and dying'. It was less likely to be expressed among doctors concerned with earning a living treating private patients, but the concern with medicine's therapeutic impotency also fuelled the movement to prevent disease. Ackerknecht's fourth kind of medicine, social, also flourished in the nineteenth century. Just as hospitals existed long before 'hospital medicine', so epidemics and preventive measures were not invented by the public health movement of the 1830s. Nevertheless, the preventive infrastructures developed partly in response to the cholera pandemics still exist, *mutatis mutandis*. The chief architect of the British public health movement, Edwin Chadwick (1800–1890), was a lawyer who thought that, on the whole, doctors were overrated. (He was neither the first nor the last lawyer to hold that opinion.) He held that epidemic diseases were caused by filth and spread via the foul smells (miasma) of rotting organic matter. His solutions were engineering ones, clean water and efficient waste disposal, which he argued would leave the world an altogether more pleasant and healthier place. His ideas were formed during the 1830s and early 1840s, and they remained more or less fixed for the rest of his long life, which extended well into the bacteriological age. Nevertheless, Chadwick also invoked science in his public health reform programme, above all the science of statistical investigation. His use of statistics can easily be shown to have been naïve, but it was ardent. In his own sphere of enquiry, Chadwick was as much in awe of the unadorned 'fact' as was his contemporary Bartlett. A later generation of Medical Officers of Health and others concerned with disease prevention (or containment) would develop new investigative techniques, more sophisticated statistics, and especially, new theories of disease causation and transmission. But the early public health movement was firmly based on the science of its time.

The final locus of medicine, the laboratory, was also largely a product of the nineteenth century, though of course laboratories (a place where one worked, especially to mutate gold from lead) had existed for much longer. A leading exponent of the laboratory, and one of its most thoughtful philosophers, had experienced Paris hospital medicine as a medical student. Claude Bernard's *Introduction to the study of experimental medicine* (1865) is both an intriguing account of his own brilliant career and a sophisticated analysis of the philosophy of experimentation within the life sciences. Hospitals, he argued, are merely the gateways to medical knowledge, and bedside clinicians can be no more than natural historians of disease. To understand the causes and mechanisms of disease, it is necessary to go into

the sanctuary of the laboratory, where experimental conditions can be better controlled. There are in nature no uncaused causes: determinism is the iron law of the universe, extending equally to living systems and inorganic ones. However, organisms present special experimental problems, and it is only through isolating particular features, and holding other parameters as constant as possible, that reliability and reproducibility can be achieved.

Bernard identified three primary branches of experimental medicine: physiology, pathology, and therapeutics. His own research programme touched all three pillars: his research on the roles of the liver and pancreas in sugar metabolism contributed to understanding normal physiology as well as diseases such as diabetes; his investigations of the sites of action of agents such as curare and carbon monoxide foreshadowed structural pharmacology and drug receptor theory; his work on the functions of the sympathetic nerves buttressed his own more general notion of the constancy of *milieu interieur* as the precondition to vital action (and freedom), a precursor of Walter Cannon's concept of homeostasis. Bernard stands supreme as the quintessential advocate of the laboratory.

Who was the first modern medical experimentalist?

When Bernard wrote, experimental medical science was still a fledgling activity, best developed in the universities of the German States and Principalities. The German university ideal of medical education was to be extolled by the American educational reformer Abraham Flexner in the early twentieth century. It was in the reformed and newly created German universities that the forms of modern scientific research were established. Research careers were created; copublication in specialist journals became common; scientific societies flourished. The microscope became the symbol of the medical scientist even as the stethoscope was becoming the hallmark of the forward-looking clinician. In the hands of scientists like Schwann, Virchow, and Weismann, the modern cell theory was developed and applied to medicine and biology more generally. These researchers established the drive to push units of analysis further and further. Eduard Buchner's identification of cell-free ferments in 1897 firmly established the importance of subcellular functions. Pasteur, Koch, Ehrlich, von Behring, and others advanced new notions of the causes of disease, the body's response to infection, and the possibilities of new drugs to combat disease. Any of these scientists might arguably be the answer to the parlour-game question: who was the first modern medical scientist?

The German-speaking lands perfected the modern forms of scientific research, but a good case can be made for a Frenchman to be crowned the first thoroughly modern experimentalist within medicine. François Magendie (1783–1855) (Fig. 1) was a child of the Enlightenment and product of the French Revolution. One of several eminent individuals (Thomas Malthus was another) raised according to the anarchic principles espoused by Jean Jacques Rousseau, Magendie did not learn to read or write till he was 10. His subsequent precocity was such that he was ready for medical studies by the age of 16, learned anatomy and surgery as an apprentice, and made his way through the Paris hospital system. Although he never lost interest in practical medical issues, his reputation was established primarily within the laboratory. His monographs on physiology and pharmacology marked new beginnings, and his life manifests three emblematic qualities which make him one of us.

Fig. 1 François Magendie. Lithograph by N.E. Maurin. (From Burgess R. Portraits of doctors and scientists in the Wellcome Institute, London, 1973, no.1870.2, by courtesy of the Wellcome Library, London.)

First, he valued facts above theories, evidence above rhetoric. But he went beyond Bartlett and the high priests of hospital medicine in insisting that in experiment, and not simply observation, lay the real future of medical knowledge. Like his pupil Claude Bernard, Magendie was a deft experimentalist. He used animals (and occasionally patients) to probe into a whole range of problems in physiology, pathology, and pharmacology: the functions of the spinal nerves; the physiology of vomiting; important facets of absorption, digestion, circulation, and nutrition; and the actions of drugs and poisons. He described anaphylaxis a century before it was named. He was as philosophically naïve as Bernard was sophisticated: of course he had theories, but his image of himself as a ragpicker with a spiked stick, gathering isolated experimental facts where he found them, is a telling one.

Second, he was modern in sometimes backing the wrong horses. He judged cholera and yellow fever to be non-contagious, was suspicious of anaesthesia, and sometimes claimed more than we might for his newly introduced therapeutic substances, such as strychnine and veratrine.

Finally, Magendie was the scientist who first expunged the double-faced Janus from the medical mentality. William Harvey worshipped Aristotle, Albrecht von Haller was steeped in history, and Isaac Newton popularized the pious conceit of pygmies standing on the shoulders of giants. Magendie looked only in one direction: the future. He had no sense of history and no use for it. He meant what he said when he insisted that most physiological 'facts' had to be verified by new experiments, and he undertook to provide a beginning. He made the laboratory the bedrock of medicine.

What happened next?

Like everyone, Magendie was of his time. Nevertheless, his values were symptomatic of important themes within nineteenth-century medicine and medical science. By the beginning of the First World War, most of the structures and the fundamental concepts of 'our' medicine were in place. Of course, both medical science and medical practice have been utterly transformed since. But the impulse of experimentation and its variable translation into practice were there. We have gone far beyond the cell in our analytical procedures, and our medical, surgical, and therapeutic armamentaria are vastly more sophisticated and powerful.

Our medicine is fundamentally different in one important respect, even if the trend was already evident in the nineteenth century: the fusion of science and technology. Science and technology have become so intertwined that the older distinctions between them are blurred. Technology made a real but minimal impact on nineteenth-century medicine. Some instruments, such as Helmholtz's ophthalmoscope, came into clinical medicine through the laboratory; and German experimental scientists were eager to exploit the latest equipment such as kymographs, sphygmographs, and the profusion of artefacts (Petri dishes, autoclaves, etc.) that Koch and his colleagues devised for the bacteriological laboratory.

Most important of all was probably the X-ray, discovered by Roentgen in late 1895. It made an immediate impact on medical diagnosis, and the associated science of radioactivity soon was felt within therapeutics. Significantly, perhaps, the pioneers of the radioactive phenomena—Roentgen, Becquerel, the Curies—received their Nobel Prizes in physics or chemistry. Hounsfield and Cormack gained theirs for computer-assisted tomography in medicine or physiology. More recently, Kary Mullis's Nobel Prize was for a technological development within molecular biology.

Both medical science and medical practice are now inseparably rooted in technology. So is modern life, another reflection of a perennial historical truth: medical knowledge and medical practice are products of wider social forces with unique historical individualities.

Further reading

Ackerknecht EH (1967). *Medicine at the Paris Hospital, 1784–1848*. Baltimore: Johns Hopkins University Press.

Bynum WF (1994). *Science and the practice of medicine in the nineteenth century*. Cambridge University Press.

Bynum WF, Porter R, eds (1993). *Companion encyclopedia of the history of medicine*. London: Routledge.

Cooter R, Pickstone J, eds (2000). *Medicine in the 20th century*. Amsterdam: Harwood Academic Publishers.

King LS (1982). *Medical thinking: a historical preface*. Princeton University Press, 1982.

Reiser SJ (1978). *Medicine and the reign of technology*. Cambridge University Press.

Weatherall DJ (1995). *Science and the quiet art: medical research and patient care*. Oxford University Press.

2.2 Scientific method and the art of healing

D. J. Weatherall

Further reading

When Henry Dale, the distinguished British physiologist and pharmacologist, arrived at St Bartholomew's Hospital as a medical student in 1900 he was told by his first clinical teacher, Samuel Gee, that, as medicine was not a science but merely an empirical art, he must forget all the physiology that he had learnt at Cambridge. This advice reflects a deep-rooted tension between the art and science of clinical practice, which still permeates the medical profession.

Patient care, from its earliest beginnings to the present day, has always been a mixture of sympathy and kindness backed up with a well-meaning but often empirical effort to alter the natural course of events. In this sense it has been, and still is, an art, practised against a background of incomplete scientific knowledge about the nature of disease processes. Human beings, like all living things, are immensely complex biological systems. Even today, with all our knowledge of their chemistry and physiology, we have a very limited understanding of the mechanisms that underlie most of the diseases that we encounter in day-to-day practice. Caring for sick people involves making considered judgements based on limited evidence and information. At best, we are slowly reaching the stage at which we are aware of how little we know.

In view of the remarkable progress in the biological sciences over the last few hundred years, today's doctors must try to establish the extent to which the balance of medical practice has shifted from 'craft' to 'science'. How far do the contents of a modern textbook of medicine reflect genuine scientific knowledge as compared with received wisdom and experience? And, of particular relevance to current medical practice and its future development, to what extent have advances in patient care in the twentieth century depended on progress in the basic sciences rather than improvements in our environment and lifestyles? In short, how much of our day-to-day clinical practice depends on a scientifically based understanding of the diseases that we encounter? It is important that we address these questions at a time when there is growing public and governmental disillusion with high-technology scientific medicine, and when many believe that the medical profession has lost its way and become more interested in diseases than in those who suffer from them. Before we tackle these difficult questions it may be helpful to define what we mean by 'scientific medicine' and to outline the way in which it has developed over the years.

Philosophers and historians of science and medicine always seem unhappy when it comes to deciding what is meant by 'scientific medicine'; this is dangerous country for the unwary! Here we shall take a pragmatic (if circular) approach, and use the term simply to describe the prevention and management of illness using methods that have been subjected to the same kinds of rigorous experimental, statistical, and observational scrutiny that are applied in other branches of science.

The earliest documentary evidence to survive from the ancient civilizations of Babylonia, Egypt, China, and India suggests, not surprisingly, that longevity, disease, and death are among our oldest preoccupations. From ancient times to the Renaissance, knowledge of the living world changed very little and the distinction between animate and inanimate objects was blurred. The Babylonians and Egyptions believed that water, air, and earth were the primary constituents of the world; a fourth, fire, was added later. This notion of the all-pervading influence of the four elements was extended to form a theory about how the human body is constituted. In short, it was thought to consist of four humours, blood, phlegm, yellow bile, and black bile. The notion that disease results from an imbalance of the humours permeated Graeco-Roman medicine and persisted until the seventeenth century. Health was viewed as a harmonious balance of the humours, while disease was thought to reflect an imbalance, or dyscrasia, leading to an abnormal mixture of the humours. This view of pathology, which provided an explanation for both mental and physical illness, formed the basis for what, at the time, was a rational approach to treatment by bleeding, purging, and dietary modification.

The extraordinary developments in natural philosophy in the seventeenth century created an environment that led to the birth of scientific medicine as we now understand it. Modern physics was founded by Isaac Newton, and the work of Boyle and Hooke finally disposed of the Aristotlean elements of earth, fire, and water. The shape of medical and biological thinking was moulded by the French mathematician, philosopher, and biologist René Descartes, who held that material things, whether animals, plants, or inorganic objects, are ruled by the same mechanical laws. All living things, he held, can be looked on as machines. A sick man is like an ill-made clock; a healthy man a well-made clock. And it was during this time that William Harvey published an almost complete description of the circulation of the blood, work that involved many years of animal experimentation and the application of simple statistical methods to determine the output of blood from the heart, and which, in effect, formed the foundation for modern investigative physiology and, later, medicine.

During the eighteenth and nineteenth centuries, the sciences that underpin medicine were further developed. In particular, the concept of the cell became the centrepiece of biology. As perceived by the French Nobel laureate, Francois Jacob, 'with the cell biology discovered its atom'. In 1858 Rudolph Virchow published his celebrated *Die Cellular Pathologie*, in which cell theory was applied to the study of pathology. All diseases, he held, are diseases of cells. This was the dawning of modern cellular pathology and the study of disease at the microscopic, and later submicroscopic, levels. The nineteenth century also saw the gradual decline of vague theories about life forces and a growing belief, helped by the emergence of organic chemistry, that living processes can be understood in terms of chemistry and physics working through complex interactions between the many different types of cells that constitute all living things, a movement that was to culminate in the extraordinary achievements in biochemistry and molecular biology in the twentieth century.

This was also the time when a start was made at assessing the value of therapeutic practices that had gone on largely unchanged for centuries. For despite these rapid advances in the biological sciences, very little could be done for the majority of the disorders that doctors faced in everyday practice. Blood letting and the administration of a variety of useless and potentially harmful treatments were still rife, and although a few drugs of genuine value had been found, foxglove extract and quinine for example, much of the doctors' armamentarium was of unproven value. In the mid-nineteenth century a French clinician, Pierre Charles Alexandre Louis, pioneered the application of statistical analysis to medical practice. One of his earliest ventures was to compile sufficient data to prove that blood letting, which had been practised for centuries, was not only useless but positively harmful in the management of many diseases. During the latter half of the nineteenth century the focus of medical science moved from France to Germany. It was here that, during the late nineteenth and early twentieth centuries, laboratories were set up where men and women could devote their time to research in the blossoming basic sciences, anatomy, physiology, and, later, biochemistry. In this atmosphere a new generation of clinical scientists evolved who became interested in physiological medicine, that is in understanding the fundamental mechanisms of disease.

These developments led to the establishment of university medical schools in the United States and parts of Europe, based on the German tradition. In 1910 the American educationalist Abraham Flexner, after visiting several German medical schools, wrote a withering critique of medical education and science in North America. This attack stimulated the development of specialist clinical departments in many American and European medical schools. Flexner's revolutionary study advocated that medical education should begin with a strong foundation in the basic sciences followed by the study of clinical medicine in an atmosphere of critical thinking and with adequate time and facilities for research. His philosophy was widely accepted, not only in North America but in many European medical schools.

The development of university clinical academic departments in the period between the two world wars, and particularly after the second, led to the emergence of 'clinical science', experimentation on patients or laboratory animals on problems that stemmed directly from observations made at the bedside. Ultimately, this led to a remarkable improvement in our understanding of disease mechanisms. Together with the expanding pharmaceutical industry, it set the scene for the development of modern, high-technology medical practice. Not surprisingly, it also had a profound effect on medical education. Indeed, those who criticize modern methods of teaching doctors, in particular the Cartesian approach to the study of human biology and disease, believe that the organization of university clinical academic departments along Flexner's lines may have done much to concentrate the minds of doctors on diseases rather than those who suffer from them.

The twentieth century has seen a revolution in the basic sciences, which started in physics, spread to chemistry, and, ultimately, completely changed the face of biology. Remarkable developments in physics at the end of the nineteenth century paved the way to an understanding of how atoms are joined together to form molecules and for the development of a new kind of chemistry, which would start to explain the structure of the molecules that make up living things. The amalgamation of physics and chemistry spawned a new discipline, molecular biology, which was to unravel the way in which genetic information is passed from generation to generation and how individual cells function, both as self-contained units and as part of the complex communication network which is the basis of life itself. In the last 20 years there has been a slow shift of emphasis in medical research from the study of disease at the level of patients or their diseased organs to their cells and molecules. Although major scientific achievements do not always have practical benefits for many years, it is already apparent that molecular and cell biology have enormous potential for the future of medical research and practice.

There is no doubt that a combination of improvements in the environment combined with the fruits of scientific medicine have greatly improved the health of Western

industrialized societies. In England a century ago, four out of ten babies did not survive to adult life, the life expectancy at birth was only 44 years for boys and 47 for girls, and even as recently as the 1930s 2500 women died each year during pregnancy or childbirth. Today, life expectancy at birth is about 73 years for boys and 78 years for girls. The major triumph for scientific medicine and public health has undoubtedly been the control of many infectious diseases. Consequently, the proportion of deaths due to infection and respiratory diseases has declined dramatically and the major causes of mortality in the West are now vascular disease and cancer. Although relatively little progress has been made towards their prevention, their management has been transformed by the ingenuity of the pharmaceutical industry combined with development of high-technology medical practice based on a better understanding of disease mechanisms.

Modern scientific medicine is not without its detractors however. Early this century George Bernard Shaw, in his brilliant Preface to *The doctor's dilemma*, derided medical research of his time. His cry 'stimulate the phagocytes' came straight from the laboratory of Almroth Wright at St Mary's Hospital Medical School. But this work was written with style and humour and was concerned mainly with debunking the pomposity of the medical profession. This was not the case in the book, *Medical nemesis*, written by the philosopher and theologian Ivan Illich, which first appeared in 1975. Using a mass of statistics, Illich set out to show that modern medical practice in general, and scientific medicine in particular, has had no effect whatever on the health of society. His thesis holds that common infections such as tuberculosis and poliomyelitis were disappearing long before the advent of antibiotics and vaccines, and, even worse, that modern medicine is a threat to society as well as to individual patients. It is, he believes, more harmful than good because it generates demands for its services and encourages aspects of behaviour that lead to more ill health and reduce our ability to cope with illness and to face suffering and death. Illich concludes that the medical profession, at least in its present form, should be disbanded.

A series of much more thoughtful critiques of modern medicine were published in the late 1970s by Thomas McKeown and others. McKeown extended Illich's thesis that the advent of vaccination, immunization, and antibiotics has had little effect on the control of infectious disease. He argued that the dominance of the mechanistic approach to the problems of disease, which started in the seventeenth century, had caused doctors to overlook important messages that the patterns of disease origins in the past had left for them, and that it had led them to underestimate their potential value for the organization of health care in the future. McKeown believed that the vast majority of diseases are environmental in origin and that if we had been thinking of disease origins rather than mechanisms it would not have taken us so long to suspect the importance of environmental agents or lifestyles in the genesis of our current killers, smoking or lack of exercise as the cause of heart disease for example. In short, writers like Illich and McKeown believe that, because practically all disease stems from the environment, modern scientific medicine, with its accent on disease mechanisms rather than origins, has had little effect on the health of society. While flawed in many ways, particularly with respect to their lack of appreciation of the relative roles of nature and nurture in the genesis of disease, arguments of this kind have had an important influence on current perceptions of the role of science in medicine.

These criticisms of modern scientific medicine have been mirrored by increasing disenchantment with modern medical practice on the part of the public, media, politicians, and even some doctors themselves. Paradoxically, the origins of this mood of disillusionment can be traced to some of the extraordinary successes of scientific medicine earlier this century. In the period after the Second World War, which saw the emergence of vaccines and antibiotics and the control of many infectious diseases, it appeared that medical science was capable of almost anything. The virtual disappearance overnight of scourges like smallpox, diphtheria, and poliomyelitis led to the expectation that similar successes would soon follow. In effect, society came to expect a state of constant rude health as its right. But this did not happen. The diseases that replaced infection, heart attacks, strokes, cancer, rheumatism, and psychiatric disorders, turned out to be much more intractable. Granted there were some remarkable advances in their symptomatic control, but these new killers could not be prevented or cured. As this became clear there was a move on the part of society to alternative medicine; if medical science could not cope with chronic backache or lung cancer why not turn for help to those who claimed they had the answers? Dietary manipulation, food allergy, herbal remedies, and a variety of other approaches to chronic illness were taken up with enthusiasm.

Yet coincident with this disillusionment with modern medicine and the search for better alternatives, it became apparent that the revolution in the biological sciences, stemming from applications of molecular and cell biology, promised to change completely the face of health care in the future. Today, hardly a week goes by without a new breakthrough being splashed all over our television screens and newspapers; another human gene has been isolated and the cause of a disease of which we have never heard is announced. New cures for heart disease or cancer appear to be just round the corner. Whenever these new remarkable discoveries are announced, excited scientists or journalists tell us that they will have a major impact on health care 'within the near future'. Yet time goes by and this doesn't seem to happen. There is a growing feeling that much which goes on in modern science is motivated more by scientists' wish for self-glorification rather than by any practical goals. Furthermore, many believe that modern science, whether it involves the manipulation of human genes or enquires into the origins of the universe, is a debasing activity that is damaging our environment and moving into areas of knowledge that are best left alone. There is a growing fear about the increasing reductionist approach to medical research. This, combined with concerns about the dehumanizing effect of modern hospitals and the feeling that doctors must return to a more holistic approach to their patients, that is to treat them as individuals rather than diseases, is causing increasing concern to our medical educationalists.

Clearly, younger readers of this book are learning their trade at a time when the whole ethos of scientific medicine is being questioned, and when thoughts are turning more to preventive medicine by modification of our environment and lifestyles, with less emphasis on understanding the basic mechanisms of disease. What are they to make of this confusing scene?

In effect, the doctors of today find themselves in a similar position to their predecessors at the beginning of this century. It was already apparent that many of the infectious diseases that were killing their patients could be partly controlled by better housing, hygiene, and other improvements in the environment; a few could be prevented by vaccination. Yet it was far from clear how far measures of this kind would be successful in controlling these diseases. In the meantime there was little that they could do for their patients with tuberculosis, meningitis, poliomyelitis, or puerperal sepsis, except improve their general well-being and manage their symptoms. They knew that there were some exciting developments in the basic biological sciences, microbiology, and, in particular, immunology, which promised to provide the solution to their problems. Yet these fields had been on the move for over half a century and still appeared to be of limited practical value. Hopes for the development of a cure for tuberculosis, following Koch's discovery of the tubercle bacillus in the 1880s, had still come to nothing. In the event it was to be another 60 years before the discovery of streptomycin provided a definitive cure for tuberculosis.

The situation is more or less the same today. We know that we can reduce the frequency of heart disease and cancer by changes in our environment and lifestyles, stopping cigarette smoking for example. But we have no idea of the extent to which we can control our major killers. For this reason it is essential that we continue to support the basic sciences and to provide the doctors of the future with sufficient understanding of them so that, as practical applications come along, as they certainly will, they are in a position to take advantage of them. If the story of the development of scientific medicine from the seventeenth century onwards has anything to tell us, it is simply this. The bulk of our major advances in health care have stemmed from advances in both public health and scientific research, the latter quite often stemming from fields that were driven by curiosity rather than any practical end in view. Harvey's discovery of the circulation of the blood had no practical value for patients with heart disease at the time; it was to be several hundred years before advances stemming from the disparate sciences of physiology, anatomy, pharmacology, and biochemistry, together with the discovery of anaesthesia and remarkable developments in surgical technology, laid the ground for modern cardiological practice. We must not neglect the role of the basic sciences in medical education and practice simply because they do not appear to have any immediate benefits.

As scientifically trained clinicians we try and analyse our patients' illnesses as far as we can with the tools of modern medical science, but frequently we find ourselves in a situation in which knowledge is incomplete and some form of therapy, even if it is of unproven value, has to be tried. The further scientific knowledge increases, the more difficult it is for caring clinicians to dissociate their scientific training from the practical necessity of doing something to relieve suffering, even though they are aware that they are rarely sure about what they are doing. Medicine has remained an art, but one that has become increasingly difficult to practise as knowledge of the scientific ignorance that underlies it has increased. The central problem for those who educate doctors of today is, on the one hand, how to encourage a lifelong attitude of critical, scientific thinking to the management of illness, yet, at the same time, recognize that moment when the scientific approach, because of ignorance, has reached its limits and must be replaced by sympathetic empiricism. Doctors have to learn to live with uncertainty. For many, this can be one of the most difficult and disturbing aspects of their work.

Textbook descriptions of disease are, of necessity, misleading. Even in the case of the most straightforward of illnesses, for which we know the cause down to the last building-block of DNA, the presentation, course, and management is never the same in any two patients. Not only are they modified by the protean physiological adaptations that occur in response to disease, but also by an individual's reaction to illness, depending on their personality, degree of family support, and many other factors that we do not understand. If, as is frequently done in our better teaching hospitals, we attempt to analyse all the features of a patient's illness and explain them in terms of current scientific knowledge, we always fail. And because we know so little about the mechanisms that underlie most of the illnesses that we encounter, a great deal of what we do must still remain empirical. It is the sheer complexity of the manifestations of illness that is responsible for the notion that medicine is still an art. And if this is the case for the relatively well-defined diseases that we see in hospital, the situation is even more complex in the community. The bulk of a family doctor's work involves non-specific complaints that seem totally foreign to anything that they learnt in the laboratory or lecture theatre as a student, often reflecting an individual patient's reactions to stresses of work, family, and environment rather than clearly defined organic disease.

Thus, apart from pastoral qualities, good doctoring requires an ability to cut through many of the unexplained manifestations of disease, to appreciate what is important and what can be disregarded, and hence to get to the core of the problem, knowing when scientific explanation has failed and empiricism must take over. This is the real art of clinical practice. It comes naturally to some doctors, but for others the difficult transition from theory to practice, from the relative certainty of the preclinical sciences to complexities of sick people, is never quite accomplished. This may be the reason for the notion that good medical practice depends more on the acquisition of experience based on long years of practice, rather than on methods of prevention, diagnosis, and treatment based on sound scientific principles. Unfortunately, this view, which may partly reflect doctors' defence mechanisms against continued ignorance, has been responsible for a great deal of poor practice, often based on fashion and anecdote rather than anything more substantive. An overexaggerated perception of the importance of medicine as a craft may also have been responsible for the dogmatism, unhealthy respect for received wisdom of the past, and extreme pomposity that has characterized many aspects of medical practice over the years. Like most human endeavours, the art of medicine has both its good and bad aspects.

In 1941 Sir Arthur Hall wrote:

> Medicine—however much it develops—must always remain an 'applied science' and one differing from all the rest in its applications to man himself. Were there no sick persons there would be no need for Medicine, either the Science or the Art. So long as there are both, both will be necessary. The application of its Science, to be of value, must be made in such a way that it will produce the maximum relief to the sick man. This calls for certain qualities in the practising physician which differ entirely from anything required in the practice of the other applied sciences. Herein lies the Art of Medicine. The need for it is as great today as it ever was, or ever will be, so long as human sickness continues.

As we have seen, our greatest difficulty is to recognize that moment in caring for a sick patient when the scientific approach, because of ignorance, has reached its limits and has to be replaced by sympathetic empiricism. It is the ability to choose that moment, partly by instinct and partly by experience gained by caring for sick people, that is the main characteristic of a good doctor. Undoubtedly, modern medical science, with its increasingly reductionist approach to the study of disease, has tended to focus our attention more on disease mechanisms than on those who are suffering from the diseases that fascinate us so much. We must redress this balance, and return to a more holistic approach to medical care, without, at the same time, allowing ourselves to develop those uncritical attitudes and reliance on received wisdom which permeated the medical profession for so many centuries. For genuine advances in medicine have stemmed from science, as defined at the beginning of this chapter, regardless of whether it involved cells and molecules, or people and populations.

Further reading

Booth C (1993). History of science in medicine. In: Teeling-Smith G, ed. *Science in medicine: how far has it advanced*?, pp. 11–22. Office of Health Economics, London.

Illich I (1977). *Limits to medicine. Medical nemesis: the expropriation of health*. Penguin Books, Middlesex.

McKeown T (1988). *The origins of human disease*. Blackwell, Oxford.

Weatherall DJ (1995). *Science and the quiet art. The role of research medicine*. Norton, New York.

Weatherall DJ (1999). The conflict between the science and the art of clinical practice in the new Millennium. *Annals of the New York Academy of Sciences* **882**, 240–6.

2.3 Medical ethics

Edmund D. Pellegrino and Daniel P. Sulmasy

Introduction

Clinical ethics, like all ethics, is a practical discipline. Whatever theory it employs, its ultimate aim is a morally defensible decision that is in the patient's best interests. This was the central moral precept of the Hippocratic oath:

> I will follow that system of regimen which, according to my ability and judgement, I consider for the benefit of my patient and I will refrain from whatever is deleterious and mischievous.

Today, this is known as the principle of beneficence, which derives its moral force from the special nature of the relationships between sick persons and health professionals. When patients seek help they are anxious, often in pain, fearful, dependent, and therefore vulnerable and exploitable. In that state, physicians ask then, 'How can I help you?' By that act physicians invite the patient's confidence and trust that they are competent and will use that competence primarily in the patient's interests.

The relationship is therefore not a contract but a covenant of trust, to which physicians must be faithful even if it means some suppression of their own self-interest. The good of the patient is thus a moral compass with four directional guide marks for the physician: medical good (competence), the patient's good expressed in his own preferences (respect for autonomy), the patient's inherent good as a human being (respect for dignity), and the patient's ultimate good (respect for spirituality). To act for the patient's good requires integration of these four elements on behalf of this person who presents as my patient now.

This chapter concentrates on the questions that must be asked, and the conditions that must prevail at the bedside to assure that each of the four levels of the patient's good is attained. Two moral algorithms are provided: one to assess the moral validity of the decision-maker, and the second, to provide a clinical framework or 'work-up' by which to analyse the ethical issues.

Who decides?

Since the good of the patient is far broader than the patient's biomedical good, clinical ethics requires the patient's participation in decision-making. Figure 1 presents an algorithm for determining the morally valid decision-maker. If the patient has sufficient decision-making capacity, the patient is the ultimate decision-maker. If the patient's decision-making capacity is variable, the physician should be guided by the last decision made when the patient was capable. If the patient has never had decision-making capacity (for example an infant or patient with mental retardation), the physician must turn to a morally valid surrogate. If the patient has lost decision-making capacity in a reversible manner (for example through depression) and it can be restored through medical treatment in a timely manner, the decision should be postponed until the patient is treated and capacity restored. If the loss of capacity is irreversible or the decision too urgent, the physician turns to any anticipatory declarations by the patient such as a living will, or the designation of a surrogate decision-maker through a legal document such as a durable power of attorney for health care. Lacking these expressions of the patient's prior wishes, the physician must engage a morally valid surrogate, that is, someone who can responsibly and knowledgeably represent the patient's wishes, and who has intact decision-making capacity and is free of significant conflicts of interest.

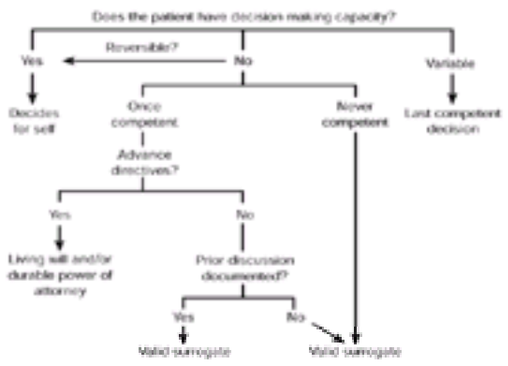

Fig. 1 Who decides?

Decision-making capacity

Accurate determination of a patient's decision-making capacity is an essential clinical skill. In North American jurisprudence, the term 'incompetence' refers to a judge's decision that an individual has lost all capacity to make decisions. The physician, however, is concerned with a narrower question: is this patient capable of making a decision about *this* clinical option in these particular clinical circumstances? The threshold of capacity will vary according to the gravity of the decision. For example, a patient suffering from suicidal depression might be allowed to refuse venepuncture, but not be allowed to sign out of the hospital against medical advice. Psychiatric consultation is often important, but the determination of decision-making capacity is generally the responsibility of the attending physician. The criteria for decision-making capacity are as follows.

1. What is the patient's neurological status? If the patient is profoundly delirious, demented, obtunded, or aphasic, the patient will not have the capacity to participate in medical decision-making.
2. Does the patient have intact judgement? That is, is the patient free of impulsiveness, able correctly to assess the seriousness of situations, to plan, and to appreciate the connections between acts and consequences?
3. Does the patient understand the nature of the procedure, its risks, benefits, and the consequences of deciding either to accept or to forgo the procedure?
4. Can the patient explain the reasons for a decision in a way that is logical and also consistent with his or her life history and previously held values?
5. Does the patient's decision remain relatively stable over time? Patients must certainly be free to change their minds, but a patient whose decision vacillates minute to minute or who refuses to make any decision may not have intact decision-making capacity.

Informed consent

Every physician must be able to obtain a morally adequate informed consent. This is not synonymous with obtaining a signature on a piece of paper. Informed consent is a process. It is one of the fundamental ways in which the physician shows respect for the good of patients as whole persons.

There are four basic elements in informed consent. The first is that the patient must have decision-making capacity. The rudiments of how to assess decision-making capacity were described above.

Second, the decision by the patient must represent an autonomous authorization, that is, it must be free from coercion, or even subtle manipulation by the physician or by others. Information must be presented in a fair and balanced fashion. This does not imply absolute neutrality nor does it imply that the physician cannot make a recommendation or even try to persuade a patient if the physician thinks the patient is making a mistake. But a physician ought not, for instance, purposefully distort the facts or threaten to sever the physician–patient relationship if the patient does not follow the physician's advice.

Third, all relevant information must be disclosed to the patient. The content to be disclosed should generally include the indications and the nature of the procedure, its potential benefits and risks, and the alternatives, including not having any procedure.

The final element of informed consent, but certainly not the least, is comprehension. The patient must not merely have been told; the patient must understand. The common clinical practice of asking, 'Do you have any questions?' is probably inadequate. If one is seriously interested in being sure that the patient has understood, it is better to ask the patient to explain back in his own words the information just disclosed.

Limits to autonomy

While the good of patients as autonomous agents must be respected, patient autonomy is not absolute. Autonomy is limited, for instance, when there is a probable threat of serious injury to an identifiable third party or parties. An example would be a demand for confidentiality by a patient testing positive for HIV who refuses to tell a sexual partner. Autonomy is further limited by the intellectual integrity of medicine as a practice. For instance, a patient cannot demand a treatment that has been proved ineffective, such as laetrile for cancer. Autonomy can also be limited to protect public health, such as in mandatory vaccination in an epidemic of a lethal infection. Finally, autonomy is limited when it violates the moral integrity of health care professionals as individual moral agents whose freedom of conscience must not be violated; for instance, a physician opposed to euthanasia ought not be forced to comply with a patient request even in settings where this is legal.

The ethics work-up

Once the appropriate decision-maker has been identified, and the conditions for autonomy assured, the physician must turn to analysis of the ethical dilemma and its substantive resolution. Ethics committees and consultations sometimes help, but ultimately clinicians are accountable for what they do or agree to. Every clinician is obliged to master the 'work-up' of the ethical problems just as surely as that of a clinical problem like coma, jaundice, or oedema.

The analytical approach we use consists of the following seven steps.

1. Secure the facts

Good ethics begins with reliable clinical and social data, with as accurate an assessment as possible of factors such as diagnosis, prognosis, effectiveness, benefits, burdens of treatments, brain function, patient preferences, and life situations. Each and all may be ethically relevant.

2. Define the ethical issue

The specifically ethical issue must be identified among the communication, interpersonal, and interprofessional problems, which usually intermingle, especially when conflicts arise. The first step in resolving conflict is clarity in the statement of the issues.

3. Frame the issue

By applying generally accepted ethical principles, one can better understand the important moral dimensions of the issue. These principles are: (i) preservation of the good of the patient as a whole person (the principle of beneficence), and (ii) respecting the good and interests of others (the principle of justice). Beneficence for persons, as noted above, demands an examination of all four aspects of the patient's good—the biomedical good, the good of the patient's autonomous choices, the patient's good as a person, and the patient's own beliefs about the ultimate good. So, for example, a severely anaemic Jehovah's Witness might refuse blood transfusion. Transfusion would serve the biomedical good of the patient, but it would violate the patient's autonomy and idea of the higher or ultimate spiritual good. Or a patient's family might demand continued treatment in the intensive care unit when such care was futile or unnecessary. This might impede other patients' access to intensive care and thus violate the principle of justice.

4. Situate the issue

It is helpful to place the case in relation to one's personal experience and that of the profession. This method of moral analysis is called 'casuistry' and asks whether the case at hand is analogous to any paradigmatic case for which a broad moral consensus has been reached. If so, one could reason by analogy to that case. For example, consider a novel case, such as whether one ought to allow a prisoner who has donated one kidney to his daughter to donate his remaining kidney to her after she has rejected the first kidney. It is useful to ask how this case compares with a more familiar case in which there is a broad moral consensus. For example, this case might be likened to the case of a man who jumps in front of a car to save his daughter's life. Such a man would be considered a hero. Casuistic analysis would ask how analogous these cases really are. What is the same about these two cases? What is different? What is the moral relevance of any similarities or differences?

5. Identify the options

In almost every case, a variety of clinical options are possible. Ethics involves selection of the morally correct choice. It is therefore necessary that all the available clinical options be identified and considered from a clinical as well as a moral point of view. This is where the technically correct and the morally good should intersect for the patient's good.

6. Reason

It is important is to weigh all the facts of the case critically and rigorously in light of one's ethical framework and clinical experience. One must interrelate the facts, the relevant principles, and any paradigm cases. Physicians should also play 'devil's advocate' and examine possible objections to their own positions. They should seek colleagues' input if time permits. An ethics committee or an ethics consultation service may be useful at this juncture. Once resolved, a retrospective critique of the reasoning employed in the case is helpful in preparing for the next time such a situation arises.

7. Decide

In clinical ethics, as in all other aspects of medicine, a decision must be made. Taking all of the aforesaid into account, a choice must be made even in the throes of uncertainty. There is no formula that guarantees the right choices. The answer will require a judicious combination of clinical judgement, practical wisdom, and common sense. In the final analysis, the decision rests with the physician's character and commitment to the good of the patient.

The ethics of end-of-life care

The most common ethical issues faced by clinicians arise at the end of life. Here the good of the patient might include decisions to withhold and withdraw life-sustaining treatments, decisions not to resuscitate, and the use of potent opioid analgesics that may hasten death.

The moral propriety of withholding or withdrawing life-sustaining treatment may be analysed systematically by examining the proposed treatment for its effectiveness, benefit, and burdens. Effective treatments are those that alter the natural history of an illness or alleviate an important symptom. Hippocrates counselled that physicians should 'refuse to treat those who are overmastered by their diseases, recognizing that in such cases medicine is powerless'. When, to a reasonable degree of medical certainty, it can be determined that a treatment will not be effective in securing the goals of treatment mutually determined by the medical team and the patient or the patient's surrogates, that treatment can be called clinically ineffective or 'futile'. In general, there is no moral obligation to provide futile treatment, although allowances must often be made for the psychological unpreparedness of the patient or family to accept the idea of futility.

If it is determined that a medical treatment is biomedically effective, the next question is whether it is beneficial. Beneficial treatments are those that bring some good to the patient beyond the biomedical good. Beneficial treatments serve not only the body, but the good as the patient chooses it, the good of the patient as a human person, or the patient's ultimate sense of the spiritual good. For instance, antibiotic treatment of pneumonia in a patient dying of malignancy might be effective, but it might not be beneficial if it merely postpones dying when the patient sees no benefit in it.

Both the effectiveness and benefits of a treatment must be weighed against their burdens—physical, financial, or emotional. When the burdens are disproportionate to effectiveness and benefits, treatment can be withheld or withdrawn. Planned re-examination of the three variables at previously agreed time intervals will avoid much of the confusion that surrounds do-not-resuscitate orders. Cardiopulmonary resuscitation is a treatment which is ineffective, burdensome, and without benefit in many terminally ill patients. When this is the case, a do-not-resuscitate order is morally licit.

It is incontrovertible that there is a moral mandate to treat pain. None the less, some physicians might hesitate to do so adequately because there is a risk that this might unintentionally hasten the death of the patient. The centuries old Rule of Double Effect may be invoked in such cases. According to this rule, a physician completely opposed to euthanasia can act with clear conscience in administering a drug like morphine to a dying patient if several conditions are met. First, the physician must sincerely intend pain relief, not the death of the patient. Second, the dose must be consistent with a plan to relieve pain through the analgesic effects of morphine, not through causing respiratory arrest and death as the means of relieving pain. Finally, the need for pain relief must be great compared with the risk of respiratory arrest and death in that patient. For example, if a patient is dying of metastatic breast cancer and is in severe pain, the potential benefit of intravenous morphine would seem overwhelming compared with the small risk that morphine might contribute to hastening an already imminent death. If these conditions are fulfilled, a physician should be able to control pain with a clear conscience, even knowing that death may unintentionally be hastened as a side-effect.

At present, these is significant controversy about whether physicians should be authorized to hasten the death of the patient intentionally through euthanasia or assisted suicide, actions not permitted by Western medicine since the Hippocratic ethic became dominant many centuries ago. Legal bans on these practices are being challenged through legislative initiatives and civil suits. Almost all professional organizations remain opposed.

Conclusion

This chapter has focused on the heart of clinical ethics, that is, acting for the good of the patient. This is the physician's central moral obligation, from which he cannot be relieved since he is bound in a covenant of trust to respond to the sick person who is in need of his medical knowledge.

We have illustrated this moral theme by analysing the four levels of the good of the patient at the bedside in several ways: first, through defining the appropriate decision-maker, the assessment of capacity, and the elements of informed consent; second, through the explication of an ethical work-up for specific cases; and third, by analysing several important ethical decisions in caring for patients at the end of life.

We acknowledge the great importance of many emerging ethical issues such as those raised by genetics, preventive medicine, and information technology. We also recognize that many ethical issues now arise in the context of team care, cost containment, managed care, and in organizational settings in which the physician is simultaneously an employee, a manager, and perhaps even an investor. These issues are too important for superficial treatment. We would only point out that in these instances too the physician's primary responsibility is the good of the sick person. If physicians default on this commitment, the last moral safeguard of the sick will have been compromised to the peril of us all.

Further reading

Beauchamp TL, Childress JF (2001). *The principles of biomedical ethics*, 5th edn. Oxford University Press, New York.

Faden RR, Beauchamp TL (1986). *A history and theory of informed consent*. Oxford University Press, New York.

Gillon R, ed. (1994). *Principles of health care ethics.* John Wiley & Sons, Chichester, UK.

Hippocrates (1939*). Hippocrates, vols I–IV*. Jones WHS, trans. Harvard University Press, Cambridge, Massachusetts.

Jonsen AR (1991). Casuistry as methodology in clinical ethics. *Theoretical Medicine* **12**, 295–307.

Pellegrino ED (1997). Managed care at the bedside: how do we look in the moral mirror? *Kennedy Institute of Ethics Journal* **7**, 321–30.

Pellegrino ED (1989). Withholding and withdrawing treatments: ethics at the bedside. *Clinical Neurosurgery* **35**, 164–84.

Pellegrino ED, Thomasma DC (1988). *For the patient's good: the restoration of beneficence in health care*. Oxford University Press, New York.

Randall F, Downie RS (1996). *Palliative care ethics: a good companion*. Oxford University Press, Oxford.

Reich WT, ed. (1995). *Encyclopedia of bioethics*, 2nd edn. Macmillan, New York.

Sulmasy DP (1992). Physicians, cost-control, and ethics. *Annals of Internal Medicine* **116**, 920–6.

Sulmasy DP (1997). Futility and the varieties of medical judgment. *Theoretical Medicine* **18**, 63–78.

Sulmasy DP, Pellegrino ED (1999). The rule of double effect: clearing up the double talk. *Archives of Internal Medicine* **159**, 545–50.

2.4.1 Bringing the best evidence to the point of care

P. Glasziou

> 'You must always be students, learning and unlearning till your life's end'.
>
> Joseph Lister

Imagine you, rather than your patient, have just been diagnosed with a serious cancer outside your field of specialty. Wouldn't you prefer that your oncologist colleague had ready access to all the relevant clinical evidence, such as the results of the relevant randomized trials? But we know this is difficult. Our textbooks are often out of date, and the relevant trials scattered across the vast ocean of medical literature. This inaccessibility of the best research data at the point of clinical decision-making has consequences for patient care, and has given rise to the discipline of evidence-based medicine.

Definition

> 'Evidence-based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. This practice means integrating individual clinical experience with the best external clinical evidence from systematic research'.
>
> DL Sackett, *et al*. 1996.

History

With basic research continually developing new diagnostic and treatment modalities, we would like to know: what options have been demonstrated to be effective and which is best? These are not new questions. Ambroise Paré faced them in 1536 as surgeon to French soldiers on campaign in Italy. He followed the advice of the most authoritative texts and treated their battle wounds with cautery using 'the oyle the hottest that was possible into the wounds'. However, he eventually ran short of oil and was 'constrained insteed to apply a digestive'. After a troubled night, he awoke to find those he had cauterized in great pain, whereas those he had not were doing well. This accidental experiment changed Paré's and French treatment. In 1747, James Lind more deliberately set out to examine alternative treatments for scurvy: he 'took 12 cases of scurvy on board the Salisbury at sea. The cases were as similar as I could have them'. Housing and diet were standardized. Of the six pairs of sailors, the two assigned oranges and lemons recovered within 3 weeks. Unlike Paré's results, Lind's took several decades to be implemented.

The methods for conducting, and the criteria for assessing, research have been considerably strengthened in the twentieth century. Bradford Hill introduced the randomized trial to medicine: the Medical Research Council trial of streptomycin for pulmonary tuberculosis in 1948. Since then more than a quarter of a million such trials have been conducted. Almost simultaneously, Yerushalmy introduced greater rigour into the evaluation of diagnostic tests by quantifying the accuracy—the sensitivity and specificity—of chest radiograph screening for pulmonary tuberculosis.

Interest in improving clinical evaluation has grown, giving rise to disciplines such as clinical epidemiology and evidence-based medicine, and a flood of clinical research. This is welcome, but has also hampered the dissemination of research results. Medline started in 1966 and currently adds to its 9 million references over 1000 new articles per day (www.nlm.nih.gov/pubs/factsheets/medline.html). Though these are culled from about 4300 journals in 30 languages and 70 countries, it is only a modest portion of the estimated 13 000 to 14 000 biomedical journals currently being published. No clinician's reading time is sufficient to keep up with this flow directly.

Keeping up to date: two strategies

So how can we cope with our information overload? Fortunately, most of the published information is not sufficient to alter clinical practice: much is 'scientist-to-scientist' communication directed at unravelling mechanisms; and many of the clinically relevant studies are not of adequate quality. Thus filtering for quality and clinical relevance reduces the flow to a manageable trickle as illustrated in Fig. 1.

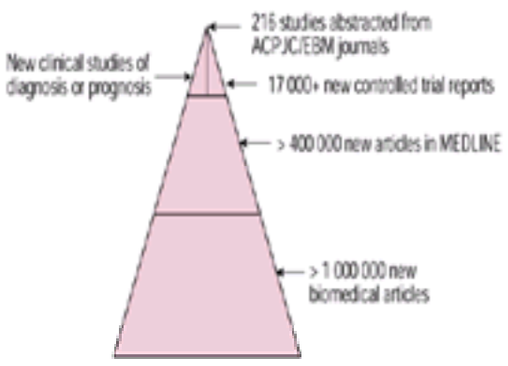

Fig. 1 The yearly flow of new biomedical publications.

There are two complementary ways of obtaining this filtered information. First we need to keep abreast of major new studies that should alter our clinical practice. However, rather than trying to scan hundreds of journals ourselves, it is wiser to enlist a group of our peers to do this. For example, journals such as the *ACP Journal Club*, *Evidence-Based Medicine*, and *Evidence Based Mental Health* review over 100 journals and appraise the articles for the quality of the research methods (less than 1 in 20 pass), relevance, and interest in order to identify new studies that could change the way we practise. The best systematic reviews and studies are reabstracted and an expert commentary helps place the new data in their current context.

The second, and more radical, process is to formulate and answer clinical questions as they arise with our patients. This is a 'just-in-time' approach: instead of trying to keep up to date with all areas of clinical practice, hoping that we have read and remembered the correct articles when we need to apply them, we shift focus to answering questions as they arise. This implies being able to say 'I don't know' and adding 'but I will find out!' When a problem appears, we formulate an answerable question, devise an information-gathering strategy, appraise the information achieved, and take it into account when deciding treatment with our patient. Learning

becomes an active, integral, and daily part of clinical practice.

Asking clinical questions

Answering patient-stimulated questions is unlikely to be done unless we can do it rapidly: finding the information in about 30 s and assimilating it within a couple of minutes. This sounds formidable, but has been shown to be feasible. In many ways it is similar to looking up drug doses: the information must be available in our consulting room, it must be well indexed, and the presentation must be readily usable. Currently none of the continually updated evidence-based resources is as comprehensive and rapid as a pharmacopoeia, and we need some skills to navigate those available.

The steps in answering clinical questions are: (i) formulating an answerable question; (ii) formulating an information-gathering strategy; (iii) assessing the quality and relevance of the information retrieved; and (iv) applying the results to our patient. To illustrate these steps consider the following patient:

Case 1. A 74-year-old man presents with his second episode of trigeminal neuralgia. As with the previous episode, he is is managing to control the pain with carbamazepine, but is requiring such large doses that he is drowsy throughout the day. He presents asking about alternatives.

In answering questions, it is helpful to classify questions into the types presented in Table 1: differential diagnosis, diagnostic accuracy, prediction/prognosis, and therapeutic effectiveness. For case 1, the issue is therapy, and a useful breakdown of such questions is: the patient, the intervention, the comparison, and the outcome. So with our patient this might be: 'In patients with trigeminal neuralgia is there a single or combined therapy which is as effective as carbamazepine at controlling pain but with less drowsiness?'

Finding answers

For treatment, we would generally first seek the results of randomized controlled trials; if there were several then we should seek existing systematic reviews. If we had answered this question previously, then the stored result would provide the fastest answer. However, since we hadn't, the first try might be *Best Evidence*: the electronic accumulation of the abstracted articles in the *ACP Journal Club* (since 1991) and *Evidence-Based Medicine* (since 1995), with over 1300 studies reviewed. Searching on the term 'trigeminal' within Therapeutics and Prevention yields one abstract (within 20 s). This was a systematic review of anticonvulsants, including carbamazepine: three placebo controlled trials showed that at 5 to 14 days follow-up 56 per cent of patients had improved with carbamazepine compared with 18 per cent with placebo ($p < 0.001$). The absolute response difference is 38 per cent, and hence for every three patients we treat there will be one additional responder (the number-needed-to-treat). This confirms carbamazepine's efficacy, but does not give us an alternative.

The next possibility is the Cochrane Library which contains Cochrane systematic reviews (the Cochrane Database Systematic Reviews, CDSR), other systematic reviews (the Database of Abstracts of Reviews of Effectiveness, DARE), and a compendium of randomized trials (the Cochrane Controlled Trials Register, CCTR) identified in Medline, EMBASE, and the handsearching by contributors to the Cochrane Collaboration. Starting the Cochrane Library CD and searching on 'trigeminal neuralgia' identifies (within 50 s) one Cochrane review (an update of the McQuay article we found in Best Evidence), and 54 controlled trials. Among these are several trials studying alternatives to carbamazepine. First, a 1988 double-blind crossover study showed baclofen significantly decreased pain in 7 of 10 patients, and in an open label study was useful in combination with carbamazepine. Second, a tantalizing but single-arm study suggested that topical capsaicin was quite effective (a Zhang's systematic review of randomized trials demonstrated clear efficacy in diabetic neuropathy and postherpetic neuralgia). Having discussed these options the patient chose to add baclofen and decrease his carbamazepine. This controlled his symptoms without drowsiness, but he later switched to the topical capsaicin which, applied to two trigger points, appeared effective.

The application of results in Case 1 was straightforward, but this is not always so. The process varies depending on the type of question (Table 1) but there are some overall similarities across these. First, is the study's illness group sufficiently similar (it need not be identical) to our patient to justify a judgement that the biological behaviour of the test or treatment would not be importantly different? Second, can we implement the test, measure, or treatment in a sufficiently similar manner? If these are fulfilled, then we need to consider how the individual features of our patient might influence the results. The next two sections look at the application of studies of diagnostic tests and of treatments.

Using the results of diagnostic test studies

Most clinical information is imperfect. This includes the history, signs, and laboratory tests. The simplest demonstration of this problem is the extensive data on the lack of agreement among experienced clinicians about the presence or absence of a clinical sign, and even between histopathologists looking at the same image. The sources of this variation and error may be in the patients, in the instruments, or in the observers. For example, true blood pressure varies considerably, but the measured blood pressure varies even more because of ifferent calibration of instruments, cuff sizes, and clinical skill. While it is important to find ways to reduce this variation by standardization and training, some residual error is inevitable.

With experience we learn, implicitly or explicitly, some simple rules to minimize the problems of error. For example, we learn to repeat unexpected abnormal test results: the majority will have disappeared on a second reading, saving us and our patients much anxiety. Experience also teaches us that test results must be interpreted in the light of the clinical picture, or equivalently that we must combine our estimate of the chance a patient has a disease (the pretest probability) with imperfect information from the test. A test's imperfection can be quantified by two measures: (i) the sensitivity, the probability of a positive test result in someone with the target disease, and (ii) the specificity, the probability of a negative test result in someone without the target disease.

Case 2. A 70-year-old man being investigated for fatigue is found to have an iron-deficiency anaemia. As part of the physical examination you do a HemeSelect (a faecal occult blood test) which is negative. Does this obviate the need for a colonoscopy?

Colorectal cancer is clearly high in the differential diagnosis; investigations of consecutive cases of iron-deficiency anaemia suggest a frequency of between 10 and 20 per cent. Let us say our estimate is 16 per cent for our case. To interpret the HemeSelect result we need to know its accuracy, that is, its sensitivity and specificity. A check of the *Best Evidence* CD provides the necessary information (in less than 30 s)—Allison *et al*. followed over 8000 consecutive people screened with three different faecal occult blood tests, with the gold standard being screen-detected cancers or cancers within 2 years of follow-up (which was 96 per cent complete). This study, which is acceptable according to the criteria in Table 1, tells us that the sensitivity is 69 per cent, that is 69 per cent of patients with cancer will have a positive HemeSelect, and the specificity is 94 per cent, that is 94 per cent of patients without cancer will have a negative HemeSelect.

To apply this to our patient with iron deficiency we need to work backwards from his chance of cancer before the test—see Fig. 2.

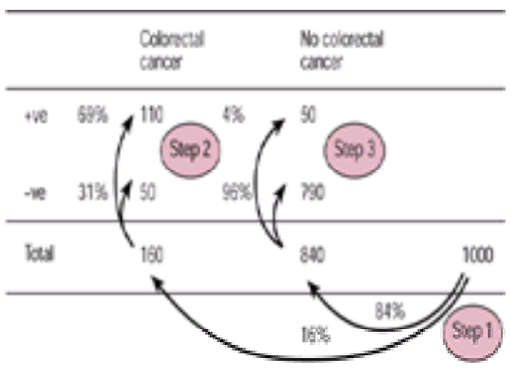

Fig. 2 Breakdown of HemeSelect results for a hypothetical 1000 patients.

Using a hypothetical 1000 patients similar to our Case 2, Fig. 2 works through this probability in three steps.

1. Of the 1000 similar patients, we would expect 160 to have a colorectal cancer and 840 not (bottom row of Table 1.)
2. Of the 160 with cancer 69 per cent (the sensitivity) will have a positive result, that is, 0.69 × 160 = 110 and the remaining 50 will have a negative result (column 1),
3. Of the 840 without cancer 94 per cent (the specificity) will have a negative result, that is, 0.94 × 840 = 790 and the remaining 50 will have a positive result (column 2).

Thus our patient with the negative HemeSelect is among the 50 (false) + 790 (true) negatives, that is, his chance of cancer is 50/(840) = 6 per cent (the post-test probability after a negative test).

We clearly cannot repeat such calculations with every patient, but methods have been developed to simplify the process (see Sackett). However, the important principle illustrated here is the need to use both the clinical picture—quantified as the pretest probability—and the test accuracy. Harold Sox has expressed this succinctly: 'What you believe after the test depends on what you believed before the test'. In particular, it is important not to be misled by false positives when screening; nor to be misled by false negatives when attempting to confirm the most likely diagnosis. Figure 3(a) illustrates this geometrically for our case 2, where even after a negative HemeSelect there is still substantial chance of colorectal cancer, whereas in the screening situation of (b) the positive HemeSelect is more likely to be a false than a true positive.

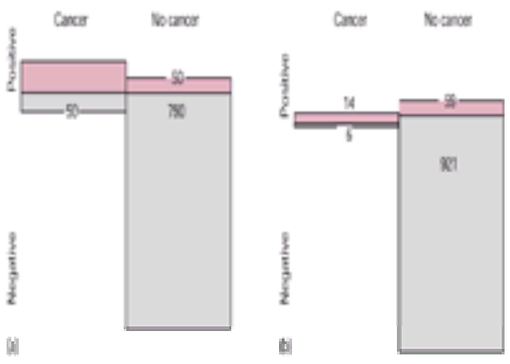

Fig. 3 Interpreting faecal occult blood results in different groups: 2 × 2 tables for (a) patient with iron-deficiency anaemia (16 per cent chance of cancer), and (b) an asymptomatic 70-year-old patient being screened (2 per cent chance of cancer).

Using the results of treatment studies

The overall results of treatment trials apply to the 'average' patient and need to be individualized. If our patient is at a higher or lower risk, then we need to adjust our estimate of the effects of treatment for this. Consider the following case.

Case 3. During a routine check of his blood pressure, a 58-year-old male with stable angina and a history of hypertension was noted to have atrial fibrillation. A check of his chart showed this had been noted several months earlier. Routine investigations revealed no cause, and because of the duration, cardioversion was not warranted. But should he be taking aspirin or warfarin?

The Cochrane Library contains both Cochrane and other systematic reviews of the five relevant randomized trials: warfarin is extremely effective therapy, with a 68 per cent reduction in the risks of ischaemic stroke. However, we must also be concerned about the dangers of antico-agulation— specifically the risks of bleeding, and most crucially the risks of intracranial haemorrhage. Should he be treated? Guidelines seem unhelpful here: a recent review by Thomson showed that the proportion of patients with atrial fibrillation recommended for anticoagulation by the 20 different guidelines ranged from 13 up to 100 per cent!

So how do we apply the systematic review results? The following four questions have been suggested.

1. Is my patient so different from those in the study that the results cannot be applied?

The inclusion and exclusion criteria of clinical trials tell us about the broad category of patients tested in the trials, but are not necessarily a good guide to the applicability of the trials to individuals. A better approach is, first, to think about the potential modifiers of the therapeutic effect, and second, the benefits and harms in the individuals.

The biological effect of an intervention may be modified by several factors: patient characteristics, comorbidities, compliance, or cointerventions. To predict these may require pathophysiological knowledge and empirical data. For example, would a patient with Parkinson's disease having problems with dental hygiene be helped by an electric toothbrush? The randomized trials suggest that certain types of electric brush are clearly better than manual brushing, but did not include patients with Parkinson's disease. However, our knowledge of Parkinson's disease does not suggest there would be any reduction in benefit, and it may be even greater given the effect of bradykinesia on manual brushing.

Treatment decisions must usually balance positive and negative effects of the intervention. For our patient on warfarin the 68 per cent relative reduction in the ischaemic stroke risk must be weighed against the inconvenience of therapeutic monitoring, and more seriously, the risks of major bleeding, particularly the risks of intracranial haemorrhage: an excess of about 1 per cent per year.

2. Is the treatment feasible in my setting?

Barriers to usage include local organization of services, costs, and skills. Patients in remote settings may have difficulty with regular monitoring; service costs and hence access will vary across countries and settings; many new therapies or procedures may require skills or technology that are unavailable, for instance cognitive behavioural therapy is helpful in many conditions but access to a skilled practitioner is often limited. These issues may make the treatment infeasible or threaten the balance of benefits and harms.

3. What are my patient's likely benefits and harms from the treatment?

Low-risk patients usually gain less absolute benefit and high-risk patients more than the 'average' patient in the trials. Hence we need to predict, based on the individual's clinical characteristics, their expected risk. By applying the relative risk reduction seen across the trials, this individualized prognosis can then be used to predict the gains of therapy. Figure 4 summarizes this process. The horizontal axis is the stroke rate per year; the vertical axis is the stroke equivalents prevented by anticoagulation using warfarin. This represents the 68 per cent relative reduction seen across the trials.

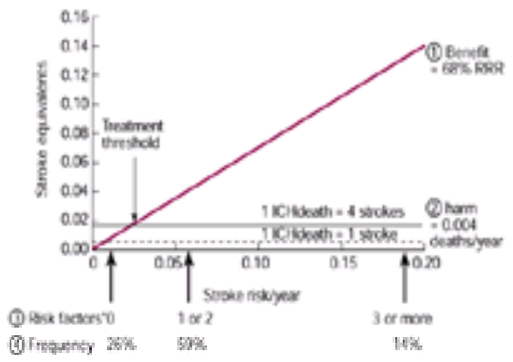

Fig. 4 Warfarin for atrial fibrillation plotting how benefits and harms vary with stroke risk (horizontal axis): (1) Expected benefit from a 68 per cent relative reduction in risk of ischaemic stroke. (2) Expected harms from intracranial haemorrhage: deaths (dashed line) or if one death is considered equal to four strokes (solid line). (3) Predicted risk based on three clinical and two echocardiographic risk factors. (4) The frequency of these risk categories in the Stroke Prevention in Atrial Fibrillation trial.

Where does our patient lie in this spectrum? On the bottom and top axis are marked the clinical risk factors. Our 58-year-old male patient had a normal echocardiogram but ischaemic heart disease and a history of hypertension, and so fitted into the one to two risk-factor category.

4. How will my patient's values influence the decision?

The essence of making wise clinical management is to follow the aphorism of Hippocrates 'Firstly do no (net) harm'. We should now compare the absolute benefits and the absolute harms of therapy, then use the strength of the individual's preferences to weigh these. In large cohort studies of the use of warfarin in the community, the rates of excess intracranial haemorrhage deaths have been about 4 per 1000 per year. This rate is shown as the bottom line in Fig. 4. This line, however, would assume that one death was equivalent to one ischaemic stroke; the line above this values one death equivalent to four ischaemic strokes. The relative value is an individual judgement, but measurements of quality of life in patients after stroke show an average quality of life of roughly 0.75 (on a scale of 0 for death to 1 for normal well health). Where the lines of benefit and harm cross one another, the expected benefit and harms are equal. It is only above this line that we begin to avoid our Hippocratic net harm, and hence the treatment that could be considered worthwhile to the patient, as with case 3.

Conclusions

If we are to advance the use of the best clinical research evidence in patient decision-making then at least two things are required. First, the compilation of the necessary information so that it is quickly accessible by practitioners in the clinic and at the bedside. Answers are needed in minutes not months. The Cochrane Collaboration has gone a long way to achieve this for questions of therapeutic interventions, but similar efforts will be needed for prognosis, diagnosis, and other types of clinical questions. Second, more serious efforts are needed in looking at the applicability and presentation of the results of studies and systematic reviews of studies to allow rapid interpretation and individual application of the results.

Further reading

Allison JE *et al.*(1996). A comparison of fecal occult-blood tests for colorectal-cancer screening. *New England Journal of Medicine* **334**, 155–9.

Anonymous (1996). Anticonvulsant drugs reduce pain in trigeminal neuralgia and diabetic neuropathy and are effective for migraine prophylaxis. *ACP Journal Club* **1124**, 35. *Evidence-Based Medicine* **1**, 89.

Antman EM *et al*. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *Journal of the American Medical Association* **268**, 240–8.

Epstein JB, Marcoe JH (1994). Topical application of capsaicin for treatment of oral neuropathic pain and trigeminal neuralgia. *Oral Surgery, Oral Medicine, and Oral Pathology* **77**, 135–40.

Fromm GH, Terrence CF, Chatta AS (1984). Baclofen in the treatment of trigeminal neuralgia: double-blind study and long-term follow-up. *Annals of Neurology* **15**, 240–4.

Glasziou P *et al*. (1998). Applying the results of trials and systematic reviews to individual patients. *ACP Journal Club* **129**, A-15–16. *Evidence-Based Medicine* **3**, 165–6.

Lind J (1753). *A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease*. Printed by Sands, Murray, & Cochran for A.Kincaid and A Donaldson, Edinburgh. (http://www.rcpe.ac.uk/cochrane/frame.html

McQuay H *et al*. (1995). Anticonvulsant drugs for management of pain: a systematic review. *British Medical Journal* **311**, 1047–52.

Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *British Medical Journal* **ii**, 769–82.

Sackett DL, Haynes RB (1997). Thirteen steps, 100 people, and 1 000 000 thanks. *ACP Journal Club* **127**, A-14. *Evidence-Based Medicine* **2**, 101.

Sackett DL *et al*. (1995). Evidence-based medicine: what it is and what it isn't. *British Medical Journal* **312**, 71–2.

Sackett DL *et al*. (1997). *Evidence-based medicine: how to practice and teach EBM*. Churchill Livingstone, New York.

The Stroke Prevention in Atrial Fibrillation Investigators (1996). Bleeding during antithrombotic therapy in patients with atrial fibrillation. *Archives of Internal Medicine* **156**, 409–16.

The Stroke Prevention in Atrial Fibrillation (SPAF) Investigators (1999). Factors associated with ischemic stroke during aspirin therapy in atrial fibrillation: analysis of 2012 participants in the SPAF I-III clinical trials. *Stroke* **30**, 1223–9.

Thomson R (1998). Guidelines on anticoagulant treatment in atrial fibrillation in Great Britain: variation in content and implications for treatment. *British Medical Journal* **316**, 509–13.

Zhang WY, Li Wan Po A (1994). The effectiveness of topically applied capsaicin: a meta-analysis. *European Journal of Clinical Pharmacology* **46**, 517–22.

2.4.2Evidence-based medicine

*Alvan R. Feinstein**

Evidence-based medicine can be viewed as a novel form of clinical practice, as a special compendium of approved information, or as a revolutionary change in medical education. After all three views are discussed, the rest of this chapter will describe the advantages and disadvantages of the new approach.

Basic views

Each of the three cited views of evidence-based medicine has been often discussed.

'Novel' form of practice

Despite the apparent novelty of the phrase itself, many practising clinicians believe that evidence-based medicine is 'nothing new'. Almost all thoughtful practitioners have regularly assembled evidence when they reviewed their own experience, developed clinical judgement, read medical literature, attended medical meetings, and talked with one another. This activity seems entirely compatible with statements by evidence-based medicine advocates that the information used to practice evidence-based medicine contains 'clinically relevant research, often from the basic sciences of medicine', including studies of diagnostic tests, prognostic markers, and 'the efficacy and safety of therapeutic, rehabilitative, and preventive regimens'. The traditional mode of clinical practice easily seems to fit not only the foregoing description, but also another statement that the practice of evidence-based medicine consists of 'integrating individual clinical expertise with the best available external clinical evidence from systematic research'.

For these reasons, clinicians may not regard the practice of evidence-based medicine as a novelty, and may wonder why it has received so much exhortation and attention. The source of novelty becomes more apparent, however, in the phrase that evidence-based medicine contains 'the best... evidence from systematic research'. Regardless of what might be offered as criteria for what is best or even good evidence, the evidence-based medicine advocates have a clear, unambiguous requirement. The 'gold standard' is 'the randomized trial, and especially the systematic review of several randomized trials'. This constraint on the acceptable 'best evidence' produces the special compendium of information that is the distinctively novel feature of evidence-based medicine.

Special compendium of information

The new journals and books devoted to evidence-based medicine concentrate on data from randomized trials and their meta-analyses (which are sometimes called 'overviews' or 'systematic reviews'). The volume of work and the scope of topics have been prolific. The first pertinent textbook, in 1996, was followed by a series of additional books, all titled as evidence-based topics in clinical practice, general practice, primary care, health care, family medicine, nursing, cardiology, and consultations. The 'evidence-based' prefix has also been applied in titles for individual articles addressing medical education, prescription guidelines, humanitarian relief intervention, organ allocation, budgeting, and health-care reform.

The intellectual centre of evidence-based medicine is the Cochrane Collaboration, based at Oxford and named after the late Archie Cochrane, a pioneering epidemiologist in urging careful evaluation of health-care interventions. The Collaboration, which co-ordinates the activities of acquiring and maintaining the special compendium, comprises an international consortium of research workers who construct an ever-enlarging data base by contributing results of their own randomized trials, by adding discoveries of previously unpublished trials, and by performing the summary aggregations that constitute the meta-analyses. The collected information, which extends through all branches of medicine, becomes the 'best evidence'. It can be published as reports in conventional literary formats or accessed via electronic media, such as the Internet. Coming mainly from activities in clinical epidemiology, the information has sometimes been called 'epi-dense' evidence.

Revolutionary change in learning

A revolutionary change in medical education is produced by the evidence-based medicine demand that clinical decisions be derived from, or sanctioned by, the contents of the new compendium. The new approach drastically alters the traditional pedagogic system in which medicine was learned from presumably knowledgeable personal authorities. They would express their knowledge either in publications or in the direct supervisory instruction given in the distant past to apprentices, and in modern medicine to students, house officers, and fellows. In the education proposed by evidence-based medicine, however, the wisdom and appraisals of a personal expert are no longer encouraged. They are replaced by the meta-analyses and printouts of the evidence-based medicine computer.

A movement that overthrows academic authorities is not a surprising modern development. The urge to purge leaders of the educational 'establishment' began in universities about 30 years ago, provoked by various national and international political discontents, often arising from the war in Viet Nam. Initiated at a time when many current leaders of evidence-based medicine were undergraduates, this drive was later enhanced by four types of new medical events.

One event, at many academic medical institutions, was the exchanging of part-time teaching faculty, who were clinical practitioners in the neighbouring community, for the increasing numbers of full-time faculty, who were clinical investigators. A second event was the reduced clinical expertise of the clinical investigators. Having been chosen mainly for achievements in laboratory research, the new pedagogic leaders were often more knowledgeable about pathophysiology than therapy and patient care. Both of these events tended to remove expert clinical authorities from being readily available in teaching activities.

A third event was the increasing development and use of randomized trials, which were usually applied to demonstrate the efficacy of new therapeutic agents. When older regimens were occasionally tested, however, the trials sometimes produced dramatic contradictions, showing that firmly-held establishment beliefs were either wrong or harmful. Perhaps the most memorable of these refutations occurred in the randomized trial of high concentration oxygen therapy for newborn premature babies. This treatment had been vigorously endorsed by renowned professors of paediatrics, and it was used for more than a decade, particularly at academic medical centres in the United States, for the goal of preventing respiratory distress. About 10 000 infants were permanently blinded with retrolental fibroplasia before the randomized trial ended the 'academic epidemic' by demonstrating that the oxygen, while avoiding respiratory distress, often produced blindness. The results greatly elevated the reputation of randomized trials for resolving controversies about therapy, but sharply diminished respect for academic authorities as sources of wise clinical advice.

A fourth problem was produced by the enormous expansion of technological tests and treatments. A generalist who could formerly keep track of almost all important changes in the field could no longer do so, and became supplanted by an array of specialists in different organ–system domains, such as cardiology. As new information continued to proliferate, the scope of the specialists also became limited. They often could no longer encompass an entire organ system, and became subspecialized in such subdomains as coronary, congenital, rheumatic, or hypertensive heart disease. The expansion and diffusion of domains of expertise would require a large array of individual authorities, not just a few; and all of them would not be readily available for personal consultation at each teaching institution.

In an atmosphere in which personal authorities were often generally viewed with suspicion, sometimes specifically impugned, and seldom always available, the time was ripe for an entirely new system to replace what was derisively called 'eminence-based' medicine. In the new system, the evidence-based medicine compendium

would be the source of 'established wisdom'; and the new authorities would be persons with evidence-based medicine credentials. In medicine, as often in politics, the leaders of the new order would be those who had fomented the overthrow of the old.

Advantages and disadvantages of the new approach

To produce the revolution that would alter long-entrenched patterns of education, the proponents often used extreme vigour and sometimes evangelical fervour in advocating evidence-based medicine and responding to adverse criticism. The zeal itself could evoke either additional admiration or further denunciation. In commenting on complaints, one of the prominent leaders said that 'most of the criticisms have to do with our hubris, style, and conviction'. The value of evidence-based medicine should be judged, however, not by the behaviour of its advocates, but by what it actually does and does not do.

Perhaps the most obvious positive accomplishment of the evidence-based medicine movement is the emphasis on citing explicit data and reasons for clinical decisions. Although this approach had previously been urged for several decades, the renewed demand for citing 'evidence' has helped end the old tradition in which decisions were justified only by non-specific explanations such as 'intuition' or 'judgement'.

A second positive accomplishment has been the demonstration that active clinicians, in an era of extensive and rapid technologic changes, can no longer rely on their previous medical education to provide a permanent basis for clinical practice. Conventional lectures and courses in 'continuing medical education', however, have not offered a satisfactory method to 'keep up' with what is happening. The evidence-based medicine movement has demonstrated a way for clinicians to 'stay alive' by doing computerized searches of accruing literature.

Several other claims of achievement have always been part of good clinical practice, and are not unique to the evidence-based medicine style. Among such claims are the contentions that evidence-based medicine integrates medical education with clinical practice, and that it helps clinicians resist unwarranted pressures.

In the original proposal for evidence-based medicine, a clinical analysis was divided into four main steps. Each step, as discussed in the next four sections, has its own distinctive advantages and disadvantages when conducted with the current evidence-based medicine compendium.

Formulating a question to be answered

The obvious first step in any process of clinical reasoning is to choose a 'prime topic' as the question to be answered. This topic is the doctor's counterpart of the patient's chief complaint. Nevertheless, just as the chief complaint may not always indicate what a patient really wants and expects, the prime topic may not always represent, and may sometimes misrepresent, what is needed for the care of the patient. To be answerable, the chosen question may have to be altered to suit the available data. Thus, the desire to learn about post-therapeutic outcomes, such as relief of symptoms and quality of life, may be diverted to an answer that indicates outcomes such as survival duration and changes in laboratory tests.

A greater, but less apparent, problem is the occasional or frequent mismatch between the available evidence and the nuances of the individual clinical situation. Most published reports of treatment, whether observational studies or randomized trials, will contain results for a stipulated therapy given to patients with a stipulated baseline clinical condition. The stipulations, however, may not include important details—such as concomitant therapy, comorbidity, severity of illness, and functional status—that distinguish the particular patient for whom the question is being asked. The general answer, reflecting results for the larger total group of patients who were treated for the condition, may not be pertinent for the patient's individual distinctions.

This problem is heightened when the evidence comes solely from randomized trials. Designed to answer questions of general efficacy rather than to guide individual treatment, the trials often contain a highly selected group of patients, treated with a relatively rigid therapeutic protocol. Furthermore, with the currently popular intention-to-treat analytic principle, the results of the trials are appraised without regard to whether or how well the patients actually maintained (or even received) the randomly assigned treatment. The results of each trial thus indicate what happens to an 'average' patient assigned to the treatment; and the meta-analyses produce an average of the averages. The average results may be satisfactory for the decisions made by economists, health-plan managers, regulatory agencies, and pharmaceutical companies; but averages are often grossly unsatisfactory for individual decisions about specific patients.

Searching the literature

An important past role of medical textbooks and published 'review articles' was to produce an authoritative summary of pertinent comments and evaluations for each prime topic. The summary may sometimes have been out-of-date, and the authority incorrect, but the search was relatively easy to do and the authority was clearly identified.

With this traditional approach rejected, clinicians are now urged to do their own computerized search of 'the literature'. For only a single topic, among the many others that may be cogent, the computer will regularly produce a large display of multiple reports that can take considerable time to obtain and read. This time can be greatly shortened, however, if clinicians forgo their own full search, and use the approved but highly truncated selection contained in the evidence-based medicine compendium. The clinician thus relies on the evidence-based medicine authorities, who may be relatively anonymous or cited in a multitude of names, instead of the individually identified expert who wrote the section in a chosen textbook or review article.

Relying on the evidence-based medicine compendium, however, can be a frustrating activity in two types of situations. One of them occurs in the 'grey zones of clinical practice' for which gold-standard randomized-trial evidence is available, but inconclusive. In the other situation, the selected prime topic has not been included in the evidence-based medicine collection. Because most randomized trials and their concomitant meta-analyses have been devoted to specific individual therapeutic regimens, very few or no trials have been done for most of the common topics of clinical practice. They include decisions about 'risk factors' and aetiological agents (such as cigarette smoking), pathophysiological challenges (such as restoring electrolyte balance), appraising the relative merits of diagnostic tests, choosing prognostic indicators, or evaluating the 'polypharmacy effects' that occur when several different treatments are used concomitantly. The published literature contains many reports on these topics, but the results come from non-randomized observational studies rather than trials.

Randomized trials (as well as observational studies) are also sparse or non-existent for many interpersonal clinical decisions, such as how to communicate with difficult patients, and how to offer useful reassurance. All of these topics will be omitted from the evidence-based medicine compendium.

Evaluating the literature

The third step in the process is to evaluate what has been found in the literature. The evaluation is relatively quick and easy for topics supplied in the evidence-based medicine compendium, since they have already been assessed and deemed worthwhile. Nevertheless, thoughtful readers may have both qualitative and quantitative difficulties in using the results. Qualitatively, as noted earlier, the evidence-based medicine information may not be suitably pertinent for the individual patient who inspired the search. Quantitatively, the actual magnitude of the cited effects may be difficult to discern and understand when reported in the statistical jargon of proportionate increments, odds ratios, relative risks, and attributable risks.

For example suppose the mortality rates are 24 per cent with treatment A and 18 per cent with the control treatment. This contrast may accurately, but alternatively, be reported as favouring the treatment by a proportional increment of 33 per cent, an odds ratio of 1.44, a relative risk of 1.33, or an attributable risk of 6 per cent.

Aware of the difficulties in interpreting these numbers, the proponents of evidence-based medicine have begun to urge that results be expressed as the inverse of the attributable risk. It is called NNT—the number of patients needed to be treated to produce one extra effect. Its calculation in this example would be 1/0.06 = 16.7, thus indicating that about 17 patients must be treated to save one more life than would occur with treatment in the 'control' group.

For the same set of data, the realization that 17 patients must be treated to get a single extra 'success' is much easier for clinicians and patients to understand than the possibly misleading improvement proportion of 33 per cent, or the often incomprehensible odds ratio of 1.44. The simple, desirable NNT expressions have not yet become ubiquitous in the evidence-based medicine compendium, however, and many results are still summarized, for statistical convenience, as odds ratios or (even worse) as logarithms of odds ratios.

A separate problem is produced when the quantitative magnitude of the difference is obscured by an evidence-based medicine headline such as 'Treatment A is

better than Treatment B for Condition C'. As long as the results have acquired the probabilistic accolade of 'statistical significance', a difference of 0.4 per cent between two treatments may be impressively hailed as 'better', even though 250 patients must receive the 'better' treatment for one to be benefited.

In an era of excessive attention to 'statistical significance', the problem of discerning and interpreting quantitative magnitudes occurs for any type of report, and is not unique to evidence-based medicine. The problem is accentuated, however, when an evidence-based medicine claim of 'better' is accepted uncritically, and particularly when the claim is used to construct guidelines for clinical practice, or criteria for policy recommendations and fiscal reimbursements.

A different type of problem occurs if the desired evidence is not contained in the evidence-based medicine compendium. The clinician must then appraise other sources of information, ranging from published literature to direct discussion with respected colleagues. The evidence-based medicine proponents have offered a hierarchy of rankings for the appraisal of published literature. Randomized trials rank at the top, followed by analyses of non-randomized observational studies, such as the groups appraised in cohort and case–control research. The lowest rank is given to uncontrolled case series, case reports, or the 'anecdotal recommendations' offered by an individual expert.

Although a reasonable generalization, this hierarchy can resemble a ranking of methods for achieving sterile precautions before entering a surgical operating room. Regardless of what is done and how well the precautions are carried out, the most important events occur during the operation, not beforehand. A well conducted observational study that answers the right question can often be more helpful than a randomized trial that is inadequately aimed; and a single case report or small series of cases can sometimes be extraordinarily enlightening. Unfortunately, a concentration on learning the methods of randomized trials and meta-analyses offers no guidance for evaluating non-randomized research, and may lead to underdeveloped critical skills with which 'young physicians who are educated only in evidence-based medicine become completely lost when they have to think about instances in which randomization is impossible'. Evaluating observational research requires special scientific principles for identifying subtle sources of bias that do not occur in randomized trials, but the principles are seldom carefully considered or discussed during an emphasis on randomized trials and meta-analyses.

A separate challenge is to appraise the quality of 'gold' in the 'gold-standard' evidence itself. Diverse 'check-lists' have been proposed for this purpose, but the lists often contain different components; and higher counts of positive components may not always indicate better quality. A major flaw in a crucial single component can invalidate the main results, despite positive counts for all other components.

Implement useful findings

The last step in the recommended process calls for the clinician to implement useful findings. Before beginning the implementation, however, the clinician must first be confident that the findings are indeed useful.

In the old 'eminence-based' system, the authoritative opinions may sometimes have been contradictory, out-of-date, or wrong, but the same phenomena can occur in the evidence-based medicine system. Randomized trials of the same topic have sometimes produced opposing results; different meta-analyses have reached different conclusions for the same set of data; meta-analyses in the evidence-based medicine compendium will often become out-of-date if not promptly revised whenever each pertinent new randomized trial appears; and a later large randomized trial may sometimes contradict results of an existing meta-analysis for smaller previous trials.

A separate problem in usefulness, as discussed earlier, is that the necessary information for an individual patient may be incomplete, inadequate, or wholly absent in the evidence-based medicine compendium. By emphasizing and averaging the 'hard data' of randomized trials, the evidence-based medicine movement can augment the 'statistical reductionism' that tends to dehumanize modern medicine, particularly when evidence-based medicine advocates refer to the care of patients as 'disease management'.

Yet another difficulty in appraising usefulness can arise from the pedagogic revolution that rejects not only the writings of individual authorities, but also their supervisory role in rounds and other teaching activities. When the probing questions of an instructor are replaced by 'evidence carts' or other electronic devices for acquiring evidence-based medicine information, students and house staff are deprived of stimuli that can lead to contemplative thought, and to the learning that comes from justifying decisions and recognizing errors. Without this type of supervisory probing to develop mental agility, young physicians may learn to seek, receive, and excrete the 'best evidence' without simultaneously being challenged to digest, absorb, and evaluate it.

Conclusions

The foregoing comments are not intended to detract from the remarkable accomplishments of the evidence-based medicine movement. Like the Internet itself, evidence-based medicine has had extraordinary growth—developing with unparalleled speed and spreading rapidly throughout the world. The movement has brought a valuable emphasis on the need for using explicit evidence to justify clinical decisions and for maintaining a constant awareness of new and changing evidence. The special evidence-based medicine compendium may contain flaws analogous to those of the old system, but the compendium itself is also a remarkable achievement. It has demonstrated a method to summarize and synthesize a vast plethora of information. The process and the results may be imperfect, but they can serve as a useful basis for future improvements.

The compendium itself, however, can never become fully satisfactory if it continues to rely solely on a 'best evidence' that may sometimes be neither good nor complete, and if the vast bulk of medical evidence, which does not come from randomized trials, continues to be excluded. The methods needed to improve the quality of non-randomized evidence, however, will be delayed or diverted, if talented young clinical investigators, who might construct those methods, are preoccupied with doing meta-analyses for contributions to the evidence-based medicine compendium.

The evidence-based medicine movement can be admired and applauded for its obvious success at inaugurating an exciting new approach to clinical reasoning. The ultimate success of the movement will depend on its ability to escape from self-imposed constraints, and to incorporate all of the contributions, in mind and data, that constitute 'medicine-based evidence'.

---

*It is with regret that we must report the death of Professor A.R. Feinstein during the preparation of this edition of the textbook.

Further reading

Cochrane AL, M Blythe (1989). *One man's medicine. An autobiography of Professor Archie Cochrane*. The Memoir Club (British Medical Journal), London.

Ellrodt G, Cook DJ, Lee J, Cho M, Hunt D, Weingarten S (1997). Evidence-based disease management. *Journal of the American Medical Association* **278**, 1687–92.

Evidence-Based Medicine Working Group (1992). Evidence-based medicine: a new approach to teaching the practice of medicine. *Journal of the American Medical Association* **268**, 2420–5.

Feinstein AR (1967). *Clinical judgment*. Williams and Wilkins, Baltimore.

Feinstein AR (1999). Statistical reductionism and clinicians' delinquencies in humanistic research. *Clinical Pharmacology and Therapeutics* **66**, 211–17.

Haynes RB (quoted in Levin A) (1998). Evidence-based medicine gaining supporters. *Annals of Internal Medicine* **128**, 334–6.

Jacobson RM, Feinstein AR (1992). Oxygen as a cause of blindness in premature infants: 'Autopsy' of a decade of errors in clinical epidemiologic research. *Journal of Clinical Epidemiology* **45**, l265–87.

Knottnerus JA, Dinant GJ (1997). Medicine based evidence, a prerequisite for evidence based medicine. *British Medical Journal* **315**, 1109–10.

Laupacis A, Sackett DL, Roberts RS (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* **318**, 1728–33.

Naylor CD (1995). Grey zones of clinical practice: some limits to evidence-based medicine. *Lancet* **345**, 840–2.

Sackett DL, Richardson WS, Rosenberg W, Haynes RB (1996). *Evidence-based medicine. How to practice and teach EBM*. Churchill Livingstone, London.

Sackett DL, Rosenberg WMC, Muir Gray JS, *et al* (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal* **312**, 71–2.

Vandenbroucke JP (1998). Observational research and evidence-based medicine: What should we teach young physicians? *Journal of Clinical Epidemiology* **51**, 467–72.

Wulff HR, GØtzsche PC (2000). *Rational diagnosis and treatment. Evidence based clinical decision making*, 3rd edition. Blackwell Science Ltd., Oxford.

2.4.3 Large-scale randomized evidence

R. Collins, R. Peto, R. Gray, and S. Parish

Introduction and summary

This chapter is intended principally for practising clinicians who need to use the results of clinical trials in their routine practice, and who want to know why some types of evidence are much more reliable than others. It is concerned with treatments that might improve survival (or some other major aspect of long-term disease outcome), and its chief point is that, as long as doctors start with a healthy scepticism about the many apparently striking claims that appear in the medical literature, trials do make sense. The main enemy of common sense is over-optimism: there are a few striking exceptions where treatments for serious disease really do turn out to work extremely well, but in general most of the claims of vast improvements from new therapies turn out to be evanescent. Hence, clinical trials need to be able to detect or to refute more moderate differences in long-term outcome. Once this common-sense idea is explicitly recognized the rest follows naturally, and it becomes obvious what types of evidence can and cannot be trusted. Although the chapter may also be of some interest or encouragement to doctors who are considering participating in (or even planning) large trials, its main intended readers are practising clinicians. For, even the most definite results from large-scale randomized evidence cannot save lives unless such practitioners accept and apply them. This chapter does not include large amounts of statistical detail: instead, it tries to communicate the spirit that underlies the increasing emphasis on large-scale randomized evidence that has developed since the 1980s.

Unrealistic hopes about the chances of discovering large treatment effects can be a serious obstacle not only to appropriate patient care but also to good clinical research. For, such hopes may misleadingly suggest to some research workers or funding agencies that small or even non-randomized studies may suffice. In contrast, realistically moderate expectations of what a treatment might achieve (or, if one treatment is to be compared with another, realistically moderate expectations of how large any difference between these treatments is likely to be) should tend to foster the design of studies that aim to discriminate reliably between: (1) differences in outcome that are realistically moderate but still worthwhile; and (2) differences in outcome that are too small to be of any material importance. Studies having this particular aim must guarantee strict control of bias (which, in general, requires proper randomization and appropriate statistical analysis, with no unduly 'data-dependent' emphasis on specific parts of the overall evidence) and strict control of the play of chance (which, in general, requires large numbers rather than much detail). The conclusion is obvious: moderate biases and moderate random errors must both be avoided if moderate benefits are to be assessed or refuted reliably. This leads to the need for large numbers of properly randomized patients, which in turn leads to both large but simple randomized trials (or 'mega-trials') and large systematic overviews (or 'meta-analyses') of related randomized trials.

Non-randomized evidence, unduly small randomized trials, or unduly small overviews of trials are all much inferior as sources of evidence about current patient management or as foundations for future research strategies. They cannot discriminate reliably between moderate (but worthwhile) differences and negligible differences in outcome, and the mistaken clinical conclusions that they engender could well result in the undertreatment, overtreatment, or other mismanagement of millions of future patients worldwide. In contrast, hundreds of thousands of premature deaths each year could be avoided by seeking appropriately large-scale randomized evidence about various widely practicable treatments for the common causes of death, and by disseminating such evidence appropriately. Likewise, appropriately large-scale randomized evidence could substantially improve the management of many important, but non-fatal, medical problems.

The value of large-scale randomized evidence is illustrated in this chapter by the trials of fibrinolytic therapy for acute myocardial infarction, antiplatelet therapy for a wide range of vascular conditions, hormonal therapy for early breast cancer, and drug therapy for lowering blood pressure. In these examples proof of benefit, that could not have been achieved by either small-scale randomized evidence or non-randomized evidence, has led to widespread changes in practice that are now preventing tens of thousands of premature deaths each year.

Moderate (but worthwhile) effects on major outcomes are generally more plausible than large effects

Some treatments have large, and hence obvious, effects on survival: for example, it is clear without randomized trials that prompt treatment of diabetic coma or cardiac arrest saves lives (and, indeed, a plaque at the entrance to our own hospital records the first clinical use of penicillin). However, perhaps in part because of these striking successes, for the past few decades the hopes of large treatment effects on mortality and major morbidity in other serious diseases have been unrealistically high. Of course, treatments do quite commonly have large effects on various less fundamental measures: drugs readily reduce blood pressure, blood lipids, or blood glucose; many tumours or leukaemias can be controlled temporarily by radiotherapy or chemotherapy; in acute myocardial infarction, lidocaine (lignocaine) can prevent many arrhythmias and streptokinase can dissolve most coronary thrombi; in early HIV infection, antiretroviral drugs substantially reduce viraemia. However, although all these effects are large, any effects on mortality are much more modest; indeed, there is still dispute as to whether any net improvement in survival is provided by the routine use of radiotherapy for common cancers, lidocaine for acute myocardial infarction, or antiretroviral agents for early HIV infection.

In general, if substantial uncertainty remains about the efficacy of a practicable treatment, its effects on major endpoints are probably either negligibly small, or only moderate, rather than large. Indirect support for this rather pessimistic conclusion comes from many sources, including: the previous few decades of disappointingly slow progress in the curative treatment of the common chronic diseases of middle age; the heterogeneity of each single disease, as evidenced by the unpredictability of survival duration even when apparently similar patients are compared with each other; the variety of different mechanisms in certain diseases that can lead to death, only one of which may be appreciably influenced by any one particular therapy; the modest effects often suggested by systematic overviews (see later) of various therapies; and, in certain special cases, observational epidemiological studies of the strength of the relationship between some disease and the factor that the treatment will modify (for example, blood pressure, blood cholesterol, or blood glucose: see later).

Having accepted that only moderate reductions in mortality are likely with many currently available interventions, how worthwhile might such effects be if they could be detected reliably? To some clinicians, reducing the risk of early death in patients with myocardial infarction from 10 per 100 patients down to 9 or 8 per 100 patients treated may not seem particularly worthwhile, and if such a reduction was only transient, or involved an extremely expensive or toxic treatment, this might well be an appropriate view. Worldwide, however, several million patients a year suffer an acute myocardial infarction, and if just one million were to be given a simple, non-toxic, and widely practicable treatment that reduced the risk of early death from 10 per cent down to 9 or 8 per cent (that is, a proportional reduction of 10 or 20 per cent), this would avoid 10 000 to 20 000 deaths. (For example, about half a million patients a year now receive fibrinolytic therapy for acute myocardial infarction, avoiding about 10 000 early deaths, and large trials have shown that this difference in early mortality persists for several years afterwards.) Such absolute gains are substantial, and might considerably exceed the numbers of lives that could be saved by a much more effective treatment of a much less common disease.

Reliable detection or refutation of moderate differences requires avoidance of both moderate biases and moderate random errors

If realistically moderate differences in outcome are to be reliably detected or reliably refuted, then errors in comparative assessments of the effects of treatment need to be much smaller than the difference between a moderate, but worthwhile, effect and an effect that is too small to be of any material importance. This in turn implies that moderate biases and moderate random errors cannot be tolerated. The only way to guarantee very small random errors is to study really large numbers, and this can be achieved in two main ways: make individual studies large, and combine information from as many relevant studies as possible in systematic overviews (Table 1). However, it is not much use to have very small random errors if there may well be moderate biases, so even the large sizes of some non-randomized analyses of

computerized hospital records cannot guarantee medically reliable comparisons between the effects of different treatments.

Avoiding moderate biases

Proper randomization avoids systematic differences between the types of patient in different treatment groups.

The fundamental reason for randomization is to avoid moderate bias, by ensuring that each type of patient can be expected to have been allocated in similar proportions to the different treatment strategies that are to be compared, so that only random differences should affect the final comparisons of outcome. Non-randomized methods, in contrast, cannot generally guarantee that the types of patient given the study treatment do not differ systematically in any important ways from the types of patient given any other treatment(s) with which the study treatment is to be compared. For example, moderate biases might arise if the study treatment was novel and doctors were afraid to use it for the most seriously ill patients, or, conversely, if they were more ready to use it for those who were desperately ill. There may also be other ways in which the severity of the condition differentially affects the likelihood of being assigned to different treatments by the doctor's choice (or by any other non-random procedure).

It might appear at first sight that by collecting enough information about various prognostic features it would be possible to make some mathematical adjustments to correct for any such differences between the types of patients who, in a non-randomized study, receive the different treatments that are to be compared. The hope is that such methods (which are sometimes called 'outcomes analyses') might achieve comparability between those entering the different treatment groups, but they cannot be guaranteed to do so. For, some important prognostic factors may be unrecorded, while others may be difficult to assess exactly and hence difficult to adjust for. There are two reasons for this difficulty. First, it is often not realized that even if there are no systematic differences between one treatment group and another in the accuracy with which prognostic factors are recorded, purely random errors in assessing prognostic factors can introduce systematic biases into the statistically adjusted comparison between treatments in a non-randomized study. Second, in a non-randomized comparison the care with which prognostic factors are recorded may differ between one treatment group and another. Doctors studying a novel treatment may investigate their patients particularly carefully, and, perhaps surprisingly, this extra accuracy can introduce a moderate bias. For example, an unusually careful search of the axilla among women with early breast cancer will sometimes result in the discovery of tiny deposits of cancer cells that would normally have been overlooked, and hence some women who would have been classified as stage I will be reclassified as stage II. The prognosis of these 'down-staged' women is worse than that of those who remain as stage I, but better than that of those already classified as stage II by less intensive investigation. Paradoxically, therefore, such down-staging improves not only the average prognosis of stage I breast cancer but also the average prognosis of stage II breast cancer, biasing any non-randomized comparison with other average women with stage I or stage II disease for whom the staging was less careful.

The machinery of a properly randomized trial: no foreknowledge of treatment allocation, no bias in patient management, unbiased outcome assessment, and no postrandomization exclusions

No foreknowledge of what the next treatment will be

In a properly randomized trial, the decision to enter a patient is made irreversibly and in ignorance of which of the trial treatments he or she will be allocated. The treatment allocation is made after trial entry has been decided upon. (The purpose of this sequence is to ensure that foreknowledge of what the next treatment is going to be cannot affect the decision to enter the patient; if it did, those allocated one treatment might differ systematically from those allocated another.) Ideally, any major prognostic features should also be irreversibly recorded before the treatment is revealed, particularly if these are to be used in any treatment analyses. For, if the recorded value of some prognostic factor might be affected by knowledge of the trial treatment allocation, then treatment comparisons within subgroups defined by that factor might be moderately biased. In particular, treatment comparisons just among 'responders' or just among 'non-responders' can be extremely misleading unless the response is assessed before treatment allocation.

No bias in patient management or in outcome assessment

An additional difficulty, in both randomized and non-randomized comparisons of various treatments, is that there might be systematic differences in the use of other treatments (including general supportive care) or in the assessment of major outcomes. A non-randomized comparison may well suffer from moderate biases due to such systematic differences in ancillary care or assessment, particularly if it merely involves the retrospective review of medical records. In the context of a randomized comparison, however, it is generally possible to devise ways to keep any such biases small. For example, placebo tablets may be given to control-allocated patients and certain subjective assessments may be 'blinded' (although this is less important in studies assessing mortality).

'Intention-to-treat' analyses with no postrandomization exclusions

Even in a properly randomized trial, unnecessary biases may be introduced by inappropriate statistical analysis. One of the most important sources of bias in the analysis is undue concentration on just one part of the evidence, that is to say on 'data-derived subgroup analyses' (see below). Another easily avoided bias is caused by the postrandomization exclusion of patients, particularly if the type (and prognosis) of those excluded from one treatment group differs from that of those excluded from another. Therefore the fundamental statistical analysis of a trial should compare all those originally allocated one treatment (even though some of them may not have actually received it) with all those allocated the other treatment (in other words it should be an 'intention-to-treat' analysis). Additional analyses can also be reported: for example, in describing the frequency of some very specific side-effect it may be preferable to record its incidence only among those who actually received the treatment. (This is because strictly randomized comparisons may not be needed to assess extreme relative risks.) However, in assessing the overall outcome, such 'on-treatment' analyses can be misleading, and 'intention-to-treat' analyses are generally a more trustworthy guide as to whether there is any real difference between the trial treatments in their effects on long-term outcome.

Problems produced by data-dependent emphasis on particular results

Treatment that is appropriate for one patient may be inappropriate for another. Ideally, therefore, what is wanted is not only an answer to the question 'Is this treatment helpful on average for a wide range of patients?', but also an answer to the question 'For which recognizable categories of patient is this treatment helpful?'. However, this ideal is difficult to attain directly because the direct use of clinical trial results in particular subgroups of patients is surprisingly unreliable. Even if the real sizes of the effects of treatment in specific subgroups are importantly different, standard subgroup analyses are so statistically insensitive that they may well fail to demonstrate these differences. Conversely, even if there is a highly significant 'interaction' (that is to say, an apparent difference between the sizes of the therapeutic effects in different subgroups) and the results seem to suggest that the treatment works in some subgroups but not in others (thereby giving the appearance of a 'qualitative interaction'), this may still not be good evidence for subgroup-specific treatment preferences.

Questions about such interactions between patient characteristics and the effects of treatment are easy to ask, but are surprisingly difficult to answer reliably. Apparent interactions can often be produced by the play of chance and, in particular subgroups, can mimic or obscure some of the moderate treatment effects that might realistically be expected. To demonstrate this, a subgroup analysis was performed based on the astrological birth signs of patients randomized in the very large Second International Study of Infarct Survival (**ISIS-2**) trial of the treatment of acute myocardial infarction. Overall in this trial, the 1-month survival advantage produced by aspirin was particularly clearly demonstrated (804 vascular deaths among 8587 patients allocated aspirin, versus 1016 among 8600 allocated as controls; 23 per cent reduction, two-sided *p* value <0.000001). However, when these analyses were subdivided by the patients' astrological birth signs, to illustrate the unreliability of subgroup analyses, aspirin appeared totally ineffective for those born under Libra or Gemini (Table 2). It would obviously be unwise to conclude from such a result that patients born under the sign of Libra or Gemini should not be given this particular treatment. However, similar conclusions based on 'exploratory' data-derived subgroup analyses, which, from a purely statistical viewpoint, are no more reliable than these, are often reported and believed, with inappropriate effects on practice.

There are three main remedies for this unavoidable conflict between the reliable subgroup-specific conclusions that doctors want and the unreliable findings that direct subgroup analyses can usually offer. However, the extent to which these remedies are helpful in particular instances is one on which informed judgements differ.

First, where there are good *a priori* reasons for anticipating that the effects of treatment might be different in different circumstances then a limited number of subgroup analyses may be **pre**specified in the study protocol, along with a prediction of the direction of such proposed interactions. (For example, it was expected that the benefits of fibrinolytic therapy for acute myocardial infarction would be greater the earlier patients were treated, and so some studies prespecified analyses subdivided by the time from the onset of symptoms to treatment: see later.) These prespecified subgroup-specific analyses are then to be taken much more seriously than other

subgroup analyses.

The second approach is to emphasize chiefly the overall results of a trial (or, better still, of all such trials) for particular outcomes, as a guide to—or at least a context for speculation about—the qualitative results in various specific subgroups of patients, and to give less weight to the actual results in each separate subgroup. This is clearly the right way to interpret the findings in Table 2, but it is also likely in many other circumstances to provide the best assessment of whether one treatment is better than another in particular subgroups. Of course, the extrapolation needs to be performed in a sensible way. For example, if one treatment has substantial side-effects, it may be inappropriate for low-risk patients. (In this case, the side-effects in a particular subgroup and the proportional benefit in that subgroup should be estimated separately, but the estimation for both might be more reliable if based on an appropriate extrapolation from the overall results rather than on the results in that one subgroup alone.)

The third approach is to be influenced, in discussing the likely effects on mortality in specific subgroups, not only by the mortality analyses in these subgroups but also by the analyses of recurrence-free survival or some other major 'surrogate' outcome. For, if the overall results are similar but much more highly significant for recurrence-free survival than for mortality, subgroup analyses with respect to the former may be more stable and may provide a better guide as to whether there are any major differences between subgroups in the effects of treatment (particularly if such subgroup analyses were specified before results were available).

Avoiding moderate random errors

The need for large-scale randomization

To distinguish reliably between the two alternatives that there is no worthwhile difference in survival or that treatment confers a moderate, but worthwhile, benefit (for example, 10 or 20 per cent fewer deaths), not only must systematic errors be guaranteed to be small (see above) compared with such a moderate risk reduction, but so too must any of the purely random errors that are produced just by chance. Random errors can be reliably avoided only by studying large enough numbers of patients. However, it is not sufficiently widely appreciated just how large clinical trials need to be in order to detect moderate differences reliably. This can be illustrated by a hypothetical trial that is actually quite inadequate—even though by previous standards it is moderately large—in which a 20 per cent reduction in mortality (from 10 to 8 per cent) is supposed to be detected among 2000 heart attack patients (1000 treated and 1000 controls). In this case, one might predict about 100 deaths (10 per cent) in the control group and 80 deaths (8 per cent) in the treated group. However, even if this difference were observed it would not be conventionally significant ($p = 0.1$); indicating that even if there is no real difference between the effects of the trial treatments, it would still be relatively easy for a result at least as extreme as this to arise by chance alone. Although the play of chance might well increase the difference enough to make it conventionally significant (for example, 110 deaths versus 70 deaths, $2p < 0.001$), it might equally well dilute, obliterate (for example, 90 deaths versus 90 deaths), or even reverse it. The situation in real life is often even worse, as the average trial size may be only a few hundred patients rather than the several thousand that would ideally be needed.

Mega-trials: how to randomize large numbers

One of the chief techniques for obtaining appropriately large-scale randomized evidence is to make trials extremely simple, and then to invite hundreds of hospitals to collaborate. The first of these large simple trials (or mega-trials) were the ISIS and GISSI studies of heart attack treatment, and a few other mega-trials have now been undertaken. However, in terms of medically significant findings, what has been achieved so far is only a fraction of what could quite readily be achieved by the assiduous pursuit of such research strategies. Any obstacle to simplicity is an obstacle to large size, and so it is worth making enormous efforts at the design stage to simplify and streamline the process of entering, treating, and assessing patients. Many trials would be of much greater scientific value if they collected 10 times less data, both at entry and during follow-up, on 10 times more patients. It is particularly important to simplify the entry of patients, otherwise rapid recruitment may be difficult. The current fashions for unduly complicated eligibility criteria, overly detailed 'informed' consent, excessive 'quality-of-life' assessments, extensive auditing of data, and measurements of the economic costs of treatment are often inappropriate.

Inappropriate inclusion of cost and of 'quality-of-life' indices

Eventually, the cost-effectiveness of various treatments needs to be assessed. However, this does not necessarily imply that costs should be assessed in the same studies in which effectiveness is to be assessed. This is particularly so if attempts to assess costs seriously damage attempts to assess the effects on mortality and major morbidity reliably. Moreover, what really matters is the cost of a treatment in routine practice, not its cost when given in the particular circumstances of a randomized trial.

Likewise, of course, any important ways in which treatments affect the quality of life need to be understood; but again this does not necessarily imply that quality-of-life indices should be assessed in the same trials that assess the main effects of treatment. For, although 20 000 patients may be required for reliable assessment of the effects of treatment on mortality and major morbidity, only a few hundred are likely to be needed for sufficiently reliable assessment of the effects of treatment on various proposed quality-of-life measures (or on costs of treatment). It may be possible to incorporate such assessments within a large mortality study as small sub-studies. But, this may be difficult in practice, and there are many instances where what should be a large simple trial of clinical efficacy should not be jeopardized by the measurement of such factors. Moreover, the effects of a treatment on quality of life in a trial, when both the doctors and the patients are uncertain about any clinical benefits of the treatment, may differ substantially from its effects on quality of life after the treatment has been shown to improve survival. Hence, it may be better to assess these other outcome measures only after having determined whether the treatment has any worthwhile effects on mortality and major morbidity, and if (as is often the case) it does not then any costs and adverse effects on quality of life may be largely irrelevant.

Simplification of entry procedures for trials: the 'uncertainty principle'

For ethical reasons, patients cannot have their treatment chosen at random if either they or their doctor are already reasonably certain what treatment is preferred. Hence, randomization can be offered only if both doctor and patient feel substantially uncertain as to which of the trial treatments is best. The question then arises: 'Which categories of patients about whose treatment there is such uncertainty should be offered randomization?' The obvious answer is all of them, welcoming the heterogeneity that this will produce. (For example, either the treatment of choice will turn out to be the same for men and women, in which case the trial might as well include both, or it will be different, in which case it is particularly important to study both sexes.) In large trials, patient homogeneity is generally a defect, while heterogeneity is generally a strength. Consider, for example, the trials of fibrinolytic therapy for acute myocardial infarction. Some had restrictive entry criteria that allowed inclusion of only those patients who presented between 0 and 6 h after the onset of pain, and so those trials contributed almost nothing to the key question of how late such treatment can still be useful. In contrast, trials with wider and more heterogeneous entry criteria that included some patients with longer delays between pain onset and randomization assessed this question prospectively, and were able to show that fibrinolytic therapy can have definite protective effects when given not only 0 to 6 but also 7 to 12 h after the onset of pain (see later).

This approach of randomizing a wide range of patients in whom there is substantial uncertainty as to which treatment option is best, was used in the Medical Research Council's European Carotid Surgery Trial (**ECST**). This trial compared a policy of immediate carotid endarterectomy with a policy of 'watchful waiting' in patients with partial carotid artery stenosis and a recent minor stroke in that part of the brain supplied by the carotid artery. If a patient was prepared at least to consider surgery, then the neurologist and surgeon responsible for that individual's care considered in their own way whatever medical, personal, or other factors seemed to them to be relevant (Fig. 1), including, of course, the patient's own preferences and values.

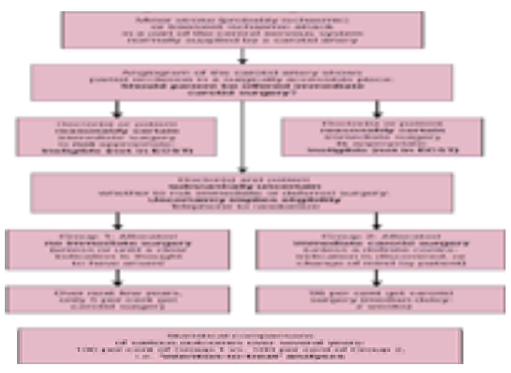

Fig. 1 Example of the 'uncertainty principle' for trial entry: the chief eligibility criterion for the European Carotid Surgery Trial (ECST) was that the doctors and patients

should be substantially uncertain whether to risk immediate or deferred surgery. (Partly because this criterion was appropriately flexible, ECST became the largest ever trial of vascular surgery.)

1. If they were then reasonably certain, for any reason, that they **did wish** to recommend immediate surgery for that particular patient, the patient was ineligible for entry into the ECST.
2. Conversely, if they were reasonably certain, for any reason, that they **did not wish** to recommend immediate surgery, the patient was likewise ineligible.
3. If, but only if, they were **substantially uncertain** what to recommend, the patient was automatically eligible for randomization between immediate versus no immediate surgery (with all patients receiving whatever their doctors judged to be the best available medical care, which generally included advice to stop smoking, treatment of hypertension, and the use of aspirin as an antithrombotic drug).

There were substantial differences between individual doctors in the types of patients about whom they were uncertain (in terms of the severity of carotid stenosis, as well as various other characteristics). This guaranteed that no category—mild, moderate, or severe stenosis—would be wholly excluded, and hence that the trial would yield at least some direct evidence in each case. As a result of the wide and simple entry criteria adopted by the ECST, 3000 patients were randomized, and therefore the study was able to provide some clear answers about who needed carotid endarterectomy. For patients with only mild carotid artery stenosis (0–29 per cent) on their prerandomization angiogram there was little risk of ipsilateral ischaemic stroke, even in the absence of surgery, so that any benefits of surgery over the next few years were small and outweighed by its early risks. Conversely, for patients with severe stenosis (70–99 per cent), the risks of surgery were significantly outweighed by its later benefits over the next few years. The trial stopped early for both categories. However, for the intermediate category of patients with moderate stenosis (30–69 per cent) the balance of surgical risk and eventual benefit remained uncertain, and so recruitment into the study continued with entry still governed by the 'uncertainty principle' as before.

The 'uncertainty principle' simultaneously meets the requirements of ethicality, heterogeneity, simplicity, and maximal trial size. It states that the fundamental eligibility criterion is that both patient and doctor should be substantially uncertain about the appropriateness of each of the trial treatments for the particular patient. With such uncertainty as the fundamental principle of eligibility, informed consent can also be simplified. For, the degree of 'informed consent' that is appropriate in a randomized comparison of different treatments governed by the 'uncertainty principle' should probably not differ greatly from that which is applied in routine practice outside trials when treatment is being chosen haphazardly—or, to put it another way, 'double standards' between trial and non-trial situations are inappropriate. The haphazard nature of many non-randomized treatment choices is reflected in the wide variations in practice between and within countries. Even when a practice is similar it may be similarly wrong: for example, before the ISIS-2 results became available (see later), almost no doctors used fibrinolytic therapy for acute myocardial infarction. Provided that trials are governed by the 'uncertainty principle', there is an approximate parallel between good science and good ethics. Indeed, in such circumstances excessively detailed consent procedures (which can be distressing and inhumane, and so would not be considered appropriate in routine practice) would neither be scientifically nor ethically appropriate.

This 'uncertainty principle' is just one of many ways of simplifying trials and thereby helping them to avoid becoming enmeshed in a mass of wholly unnecessary traditional complexity. If randomized trials can be substantially simplified, as has already been achieved for a few major diseases, and hence made very much larger, then they will play an appropriately central role in the development of rational criteria for the planning of healthcare throughout the world.

Minimizing both bias and random error: systematic overviews (meta-analyses) of randomized trials

Cochrane was one of the first people to emphasize the need to bring together, by specialty, the results from all relevant randomized trials, and the Cochrane Collaboration is now attempting to do this systematically. When several trials have all addressed much the same therapeutic question, the traditional procedure of choosing only a few of them as the basis for practice may be a source of serious bias, since chance fluctuations for or against treatment may affect which trials are chosen. To avoid this, it is appropriate to base inference chiefly on a systematic overview (or meta-analysis) of all the results from all the trials that have addressed a particular type of question (or on an unbiased subset of such trials), and not on some potentially biased subset of the trials. Such overviews will also minimize random errors in the assessment of treatment since, in general, far more patients are involved in an overview than in any contributory individual trial.

The separate trials may well be heterogeneous in their entry criteria, their treatment schedules, their follow-up procedures, their methods of treating relapse, etc. In view of this heterogeneity, at one extreme each trial might be considered in virtual isolation from all others, while at the opposite extreme all might be considered together. Both these extreme views have some merit, and the pursuit of each by different people may prove more illuminating than too definite an insistence on any one particular approach. However, the heterogeneity of the different trials merely argues for careful interpretation of any overviews of different trial results, rather than arguing against any such overviews. For, whatever the difficulties in interpreting overviews may be, without them moderate biases and random errors, which may obscure any moderate treatment effects (or, conversely, may imply effects where none exists), cannot reliably be avoided.

Which overviews are trustworthy?

Since the 1970s, a large (and rapidly increasing) number of meta-analyses of the results of randomized trials have been reported, not all of which are trustworthy. The two fundamental questions are how carefully the overview has been performed and how large it is. The simplest approach is merely to have collected and tabulated the published data from whatever randomized trial reports are easily accessible in the literature, and sometimes this may suffice. At the opposite extreme, extensive efforts may have been made by those organizing the overview to locate every potentially relevant randomized trial, to collaborate closely with the trialists to seek individual data on each patient ever randomized into those trials, and then (after extensive checks and corrections of such data) to produce, in collaboration with those trialists, agreed analyses and publications. The results of some of the largest such collaborations will be described later: the Antiplatelet Trialists' (**APT**) Collaborative Group, the Fibrinolytic Therapy Trialists' (**FTT**) Collaborative Group, and the Early Breast Cancer Trialists' Collaborative Group (**EBCTCG**). Collaboration of the original trialists in the overview process, with collection of individual patient data, can help to avoid or minimize the biases that could be produced by missing trials (for example, owing to the greater likelihood of extremely good, or extremely bad, results being particularly widely known and published), by inappropriate postrandomization withdrawals, or by the failure to allocate treatment properly at random. If randomization was performed properly in the first place, then postrandomization withdrawals can often be followed up and restored to the study for an appropriate 'intention-to-treat' analysis. Knowledge of the exact methods of treatment allocation (backed up by checks on whether the main prognostic factors recorded are non-randomly distributed between the treatment groups in a particular trial) may help to identify trials that were not properly randomized and hence should be excluded from an overview of randomized trials. Overviews based on individual patient data may also provide more information about treatment effects than the more usual overviews of grouped data, for they allow more detailed analyses—indeed, if they are really large then they may actually yield statistically reliable subgroup analyses of the effects of treatment in particular types of patient.

Conversely, even a perfectly conducted overview may not be large enough to be reliable. An overview that brings together complete data from every trial of a certain treatment but which still (because the trials were all small) includes a total of only 100 deaths will have random errors that are no smaller than those for a single trial with 100 deaths among such patients. Small-scale evidence, whether from an overview or from one trial, is often unreliable and will often be found in retrospect to have yielded wrong answers. What is needed is large-scale randomized evidence; it does not matter much whether that evidence comes from a properly conducted overview or a properly conducted trial. The practical medical value of such evidence will now be illustrated by a few recent examples.

Some examples of important results in the treatment of vascular and neoplastic disease that could have been reliably established only by large-scale randomized evidence

Definite result from a single very large trial: benefit from medium-dose aspirin for patients with suspected acute myocardial infarction (and benefits among other groups of patients indicated by overviews of trials)

In the ISIS-2 trial, half of 17 000 patients with suspected acute myocardial infarction were allocated aspirin tablets (162 mg/day for 1 month, which virtually completely inhibits cyclo-oxygenase-dependent platelet activation) and half were allocated placebo tablets. Before 1988, when the ISIS-2 results were published, aspirin was not routinely used in the treatment of acute myocardial infarction, and no other major trial had (or has subsequently) assessed the use of aspirin in cases of suspected acute myocardial infarction. However, the effects of 1 month of aspirin were so definite in ISIS-2 (804/8587 vascular deaths among those who were allocated aspirin versus 1016/8600 among those who were not) that even the lower 99 per cent confidence limit would have represented a worthwhile benefit from such a simple and inexpensive treatment (Fig. 2).

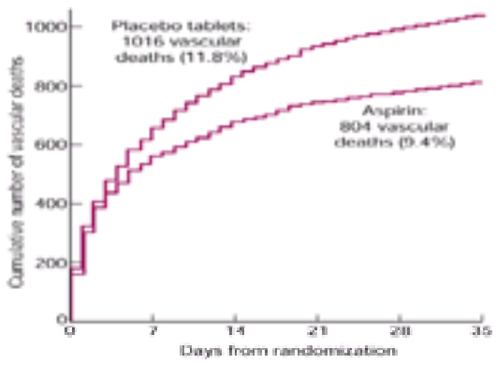

Fig. 2 Effect of administration of aspirin for 1 month on 35-day mortality in the 1988 ISIS-2 trial among over 17 000 patients with acute myocardial infarction. (Absolute survival advantage: 24 SD5 lives saved per 1000 patients allocated aspirin, 2 *p* <0.00001.)

As a result, worldwide treatment patterns changed sharply when the ISIS-2 results emerged, and aspirin is now routinely used in many different countries for the majority of emergency hospital admissions with suspected acute myocardial infarction. In the United Kingdom, for example, two British Heart Foundation surveys found cardiologists reporting that routine aspirin use in acute coronary care had increased from under 10 per cent in 1987 to over 90 per cent in 1989. Worldwide, the annual number of patients with suspected myocardial infarction who would nowadays be given such treatment must be well over a million a year, suggesting that in this clinical context alone aspirin is already preventing tens of thousands of premature deaths each year. However, if the ISIS-2 trial had been a factor of 10 smaller (that is, 1700 instead of 17 000 patients), then exactly the same proportional reduction in mortality as shown in Fig. 2 would not have been conventionally significant and therefore would have been much less likely to influence medical practice—indeed, the result might by chance have appeared exactly flat, greatly damaging future research on aspirin in this context. Likewise, if the ISIS-2 trial had been non-randomized, then it might well have produced the wrong answer (since in a non-randomized study doctors might tend to give active treatment to patients who are particularly ill, or who are rather different in various other ways from those not given active treatment). In addition, even if a non-randomized study did happen to produce an unbiasedly correct answer, it would be impossible to be sure that it had actually done so, and hence again a non-randomized study might have had much less influence on medical practice than did ISIS-2.

In the ISIS-2 trial aspirin significantly reduced the 1-month mortality figure, but it also significantly reduced the number of non-fatal strokes and non-fatal reinfarctions that were recorded in hospital. Combining all these three outcomes into 'vascular events' (stroke, death, or reinfarction), 13 per cent of those who were allocated aspirin versus 17 per cent of the controls were known to have suffered a vascular event in the month after randomization (Table 3)—an absolute difference of 40 events per 1000 treated (or, perhaps more relevantly, of 40 000 per million). The randomized trials of aspirin, or of other antiplatelet regimens, in other types of high-risk patients (for example, a few years of aspirin for those who have survived a myocardial infarction or stroke) have not been as large as ISIS-2, and so, taken separately, most have yielded false-negative results. However, when the results from many such trials are combined, statistically definite reductions in 'vascular events' are seen (Table 3). Since such treatments do not appear to increase non-vascular mortality, all-cause mortality is also significantly reduced.

In principle, these findings could, if appropriately widely exploited, prevent about 100 000 premature vascular deaths a year in developed countries alone, and there are probably at least as many vascular deaths in less-developed as in developed countries. Hence, with realistically achievable levels of the use of 'medium dose' aspirin (75–325 mg/day) for the secondary prevention of vascular disease, it might well be possible in practice to ensure that aspirin is used in enough high-risk patients to prevent, or substantially delay, at least 100 000 vascular deaths per year worldwide, and such use of aspirin would, in addition, prevent a comparable number of non-fatal strokes or heart attacks. (Medium-dose aspirin was the least expensive and most widely tested antiplatelet regimen: it is of proven efficacy, and on review of all the antiplatelet trials no other antiplatelet regimen has been shown to be of greater efficacy in preventing vascular events; see notes to Table 3.) This large-scale randomized evidence regarding medium-dose aspirin is now changing worldwide clinical practice in ways that will, at low cost, prevent much death and disability in high-risk patients. However, small trials, small overviews, or non-randomized studies (however large) could not possibly have provided appropriately reliable evidence about such moderate risk reductions.

Definite result from a very large overview of trials: benefit from 'adjuvant' therapy with tamoxifen for patients with 'early' breast cancer (and possible benefit suggested with ovarian ablation in younger women)

By definition, in 'early' breast cancer all detectable deposits of disease are limited to the breast and the locoregional lymph nodes, and can be removed surgically. However, experience shows that undetectably small deposits may remain elsewhere that eventually cause clinical recurrence at a distant site, perhaps after a delay of several years, which is then usually followed by death from the disease. These micrometastatic deposits may have been stimulated by the body's own hormones during the years before recurrence became detectable. herefore, among women who have had the detectable deposits of breast cancer removed by surgery (or by surgery and radiotherapy), there have been many trials of 'adjuvant' treatments that either reduce the production of endogenous oestrogens (for example, various forms of ovarian ablation) or block the access of those oestrogens to the tumour cells (for example, tamoxifen, which blocks the oestrogen receptor protein in some breast cancer cells).

Taken separately, most of these adjuvant trials have been too small to provide reliable evidence about long-term survival. However, if the results of all of them are combined, some very definite differences in 10-year survival rates emerge (Fig. 3). Among women with stage II disease who are less than 50-years old (and therefore generally pre- or perimenopausal), ovarian ablation appears to produce about a 10 per cent absolute difference in the 10-year survival figure (for example, 50 per cent versus 40 per cent). This finding is based on the analysis of only a few hundred deaths so it is still not as reliable as might ideally be wished and, because substantial uncertainty remains, much larger trials are now in progress. Among older women with stage II disease, ovarian ablation is unlikely to be of much relevance (since most of the endogenous oestrogen at older ages comes from sources other than the ovaries), but, in aggregate, the randomized trials among such women have shown very definitely that a few years of tamoxifen therapy also produces about a 10 per cent absolute difference in the 10-year survival rate. A smaller, but still highly significant, reduction in mortality by tamoxifen is also seen among the 10 000 randomized women with stage I disease. Taken separately, however, 37 of the 42 tamoxifen trials were too small to have yielded statistically reliable evidence on their own (2 *p* >0.01), and the five other trials were significant only because, by chance, they had results that were too good to be true.

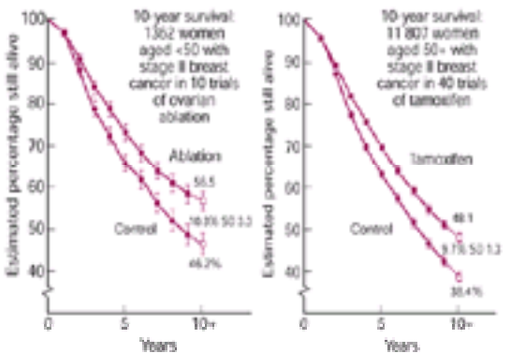

Fig. 3 Effects of hormonal adjuvant treatments for early breast cancer on the 10-year survival rate in a worldwide overview of randomized trials. (Early Breast Cancer Trialists' Collaborative Group, 1992.)

These tamoxifen overview results have already changed clinical practice substantially, and have redirected research towards large randomized trials of the effects of different durations of tamoxifen treatment: should tamoxifen in asymptomatic women be continued for 2 years, for 5 years, or indefinitely? Large randomized studies of tamoxifen in the primary prevention of breast cancer among high-risk women are only just beginning. However, they have been encouraged by the results from the tamoxifen trials' overview in 30 000 patients with established cancer (stage II or stage I) in one breast, among whom there has been a highly significant reduction of

one-third in the likelihood of developing contralateral breast cancer (but a small absolute increase in endometrial cancer). Again, this degree of trustworthy detail would not have been attainable without large-scale randomized evidence.

Promising overview of small trials confirmed by a large trial: benefit from fibrinolytic therapy as emergency treatment for a wide range of patients with acute myocardial infarction

Fibrinolytic drugs that dissolve a thrombus which may be blocking a coronary artery, thereby causing an acute myocardial infarction, were introduced into clinical research in the late 1950s. However, the trials of fibrinolytic drugs in the 1960s and 1970s were too small to be statistically reliable (none involved even 1000 patients). So, by the early 1980s the haemorrhagic side-effects were obvious, the benefits had not been convincingly demonstrated, and these agents were generally considered to be dangerous, ineffective, and hence inappropriate for routine coronary care. Although overviews published in the mid-1980s of the previous small trials (involving a total of only about 6000 patients in 24 trials) indicated a statistically definite benefit, they were not really believed by cardiologists and so such treatments were still not widely used. The situation has been saved by two large randomized trials, ISIS-2 and GISSI-1, both of which involved more than 10 000 patients (and by their aggregation with the seven other randomized trials that involved more than 1000 patients; see below). In ISIS-2, not only were patients randomly allocated to receive aspirin or placebo tablets as described earlier (Fig. 2), but they were also separately allocated to receive intravenous streptokinase (1.5 million units infused over about 60 min) or a placebo infusion. In this 'factorial' design (which allows the separate assessment of more than one treatment without any material loss in the statistical reliability of each comparison), one-quarter of the patients were allocated aspirin alone, one-quarter were allocated streptokinase alone, one-quarter were allocated both streptokinase and aspirin, and one-quarter were allocated neither (that is, they were given placebo tablets and placebo infusion). Streptokinase, like aspirin, produced a highly significant reduction in mortality, and the combination of treptokinase and aspirin was highly significantly better than either aspirin or streptokinase alone (Fig. 4).

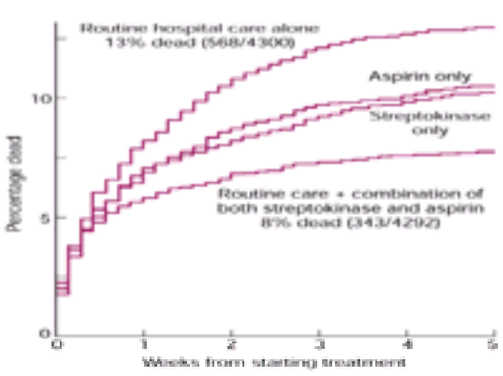

Fig. 4 Effects of a 1-hour streptokinase infusion (and of aspirin for 1 month) on 35-day mortality in ISIS-2 (1988) among 17 187 patients with acute myocardial infarction who would not normally have received streptokinase or aspirin, divided at random into four similar groups to receive aspirin only, streptokinase only, both, or neither. (Any doctor who believed that a particular patient should be given either treatment gave it, but did not include that patient in ISIS-2.)

The results shown in Fig. 4 might suggest that there was no need to collect any more randomized evidence about fibrinolytic therapy, but this ignores the potential hazards of such treatment and the heterogeneity of patients. Taken separately, even ISIS-2, the largest of these trials, was not large enough for statistically reliable subgroup analyses, but when the nine largest trials were all taken together they included a total of about 60 000 patients, half of whom were randomly allocated fibrinolytic therapy. Those entering a coronary care unit with a diagnosis of suspected or definite acute myocardial infarction range from patients who are already in cardiogenic shock with low blood pressure and a fast pulse (half of whom will die rapidly) to those who have merely had a history of chest pain and no very definite changes on their ECG (of whom 'only' a small percentage will die before discharge). Fibrinolytic therapy often causes a frightening blood pressure drop: should it be used in patients who are already dangerously hypotensive? It occasionally causes serious strokes: should it be used in patients who are elderly or hypertensive, and therefore already have an above-average risk of stroke (or who have only slight changes on their ECG, and therefore have only a low risk of cardiac death)? Finally, if the coronary artery has been occluded for long enough, the heart muscle that it supplies will have been irreversibly destroyed: how long after the heart attack starts is fibrinolytic treatment still worth risking—3 h? 6 h? 12 h? 24 h?

These questions needed to be answered reliably before appropriate and generally accepted indications for and against such an immediately hazardous, but potentially effective, therapy could be devised. To address them, all fibrinolytic therapy trialists collaborated in a systematic overview of the randomized evidence. On review of the 60 000 patients randomized between fibrinolytic therapy and control in trials of more than 1000 patients, some of the therapeutic questions were relatively easy to answer satisfactorily. For example, it appears that most of those whose ECG is still normal (or shows a pattern that indicates only a low risk of death) can be left untreated, leaving open the option of starting fibrinolytic treatment urgently if their ECG changes suddenly for the worse in the following few hours. Conversely, among those who already had 'high risk' ECG changes when they were randomized, the absolute benefit of immediate fibrinolytic therapy was, if anything, slightly greater than is indicated by Fig. 4. Age, sex, blood pressure, heart rate, diabetes, and a previous history of myocardial infarction could not identify reliably any group that would not, on average, have their chances of survival appreciably increased by treatment.

The longer that fibrinolytic treatment for such patients was delayed, the less benefit it seemed to produce. Among those whose ECG showed a definite ST-segment elevation or bundle-branch block, the benefit was greatest (about 30 lives saved per 1000) among those randomized between 0 and 6 h after the onset of pain (Fig. 5). However, the mortality reduction was still substantial and significant (about 20 per 1000, 2 p <0.003) when such patients were randomized 7 to 12 h after the onset of pain. Indeed, if they were randomized between 13 and 18 h after the onset of pain there still appeared to be some net reduction in mortality (about 10 per 1000, but not statistically definite). The regression line in Fig. 5 reinforces these separate subgroup analyses in a more reliable way. Yet, before these large trials it was forcefully, but mistakenly, argued that such treatments could not possibly be of any worthwhile benefit if given more than a few hours after the onset of symptoms.

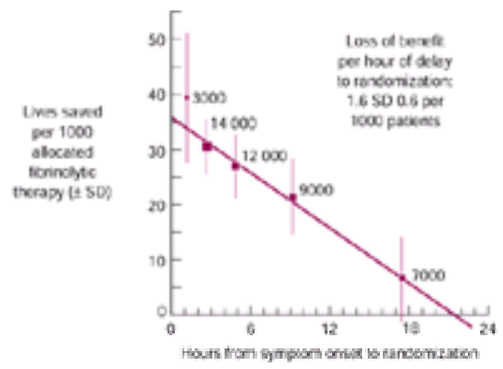

Fig. 5 Benefit versus delay (0–1, 2–3, 4–6, 7–12, or 13–24 h) in the nine largest randomized trials of fibrinolytic therapy versus control in patients with acute myocardial infarction. One-month mortality results for 45 000 patients with ST elevation or bundle-branch block when randomized, showing the definite net benefit even for the 9000 randomized 7–12 h after the onset of pain. (Fibrinolytic Therapy Trialists' Collaboration, 1994.)

Such detailed inferences are difficult enough with large-scale properly randomized evidence, and would be impossible without it; because of their unknowable biases (see above), non-randomized database analyses are simply not a viable alternative to large-scale randomized evidence. Nor would randomization of 'only' several thousand patients have been sufficient. Indeed, in several important respects what is still needed is more, rather than less, randomized evidence about the effects of fibrinolytic therapy in various particular types of patient. First, it is still unclear whether patients who have definite ECG changes such as ST elevation or bundle-branch block, but who present between 12 and 18 h, or even 18 to 24 h, after the onset of pain should be treated; more randomized evidence is still needed (Fig. 5). Second, for one particular poor-prognosis ECG category (ST depression) the 1-month mortality rates still appear unpromising even when all currently available trials are combined (15 per cent dead among those allocated fibrinolytic therapy versus 14 per cent dead among controls, but based on only 4000 patients).

Analogy with the results in other high-risk categories suggests that this result for patients with ST depression may well be a false-negative. Perhaps it has arisen from an unduly data-dependent emphasis on what may, in retrospect, prove to have been a random irregularity in the results in this particular subgroup of only a few thousand individuals. Again, more randomized evidence is needed. Nevertheless, substantial progress has been made in the past decade of mega-trials of fibrinolytic agents. Worldwide, in the mid-1990s, about half a million patients per year were given fibrinolytic therapy, avoiding about 10 000 early deaths each year.

Small trials refuted by a mega-trial: lack of significant benefit from magnesium infusion in suspected acute myocardial infarction

It had been suggested that an infusion of a magnesium salt might reduce early mortality in patients with suspected acute myocardial infarction. Several small trials, involving between them a total of only about 1500 patients, had addressed this question, and their aggregated results indicated a statistically significant, but implausibly large, benefit (42/754 deaths among those allocated magnesium versus 86/740 among the controls, 2 p <0.001). Some argued that such results constituted proof beyond reasonable doubt that magnesium was of sufficient value to justify its widespread usage without seeking further randomized evidence, but others remained sceptical, arguing that the apparent results were far too good to be true.

Therefore two trials, one (LIMIT-2) involving 2000 patients and one (ISIS-4) involving 58 000 patients, were set up to test the possible effects of magnesium more reliably. The former yielded a moderately promising result (Table 4), indicating avoidance of about one-quarter of the early deaths. However, because of its small size this result was statistically compatible with a true benefit that ranged from no effect to about a halving of early mortality. The much larger ISIS-4 trial yielded a completely unpromising result, and so the overall evidence, based on about 60 000 randomized patients, is now non-significantly adverse.

In view of the striking disparity between the apparent effects of magnesium before and after ISIS-4 had provided large-scale randomized evidence, it is of interest to recall some of the expert views that were expressed while ISIS-4 was in progress. Some felt so strongly that magnesium was already of proven benefit (and hence that further randomization was unethical) that the data-monitoring committee of ISIS-4 was lobbied to try to have the study stopped early and all future patients given magnesium. In contrast, the ISIS-4 steering committee was sufficiently sceptical to want large-scale randomized evidence. They believed that the available evidence was consistent with a negligible benefit, or even a small net hazard, although they all thought it more likely that at least some net benefit would be seen. Even after the LIMIT-2 result was available they continued to hold these opinions, and thought that if there was any real benefit then this was likely to be less than LIMIT-2 had suggested (and hence very much less than the other small trials had suggested).

Those who had trusted the implausibly extreme results from the previous small trials may well have been disappointed by the results of the ISIS-4 mega-trial, which now provide strong evidence that the routine use of magnesium has little or no effect on mortality in acute myocardial infarction. However, in a world where moderate benefits are much more plausible than large benefits, striking results in small-scale trials, in small-scale overviews, or in small subgroups will frequently prove evanescent. The medical assumption that both a moderate mortality difference and a zero mortality difference may be plausible, but that an extreme mortality difference is much less so, has surprisingly strong consequences for the interpretation of randomized evidence. In particular, it implies that even quite highly significant (for example, 2p = 0.001) mortality differences that are based on only relatively small numbers of deaths may provide untrustworthy evidence of the existence of any real difference.

Trials in their epidemiological context: effects of lower, and of lowering, blood pressure on the risk of stroke and coronary heart disease

Quantitative epidemiological evidence about the effects of long-term differences in risk factors (such as blood pressure or blood cholesterol level) can help in interpreting the results from trials of the effects of reducing these risk factors for only a few years. This may help not only in interpreting previous trials but also in planning the size and duration of any future risk-factor modification trials. For, epidemiological evidence provides approximate upper limits to the risk reductions that could plausibly be expected in the trials, and may also help to identify populations that are particularly likely to benefit from risk-factor modification.

For example, appropriate analyses of prospective, observational epidemiological studies of diastolic blood pressure and disease indicate that, throughout the range of usual diastolic blood pressure in the populations studied (that is, about 70–110 mmHg), a lower value was associated with a lower risk of suffering a first stroke or episode of coronary heart disease (that is, there seemed to be neither a 'threshold' value nor a 'J-shape' relationship, Fig. 6). The steepness of this continuous relationship suggests that the eventual risk reductions produced by practicable blood pressure lowering measures (for example, with antihypertensive treatment) may well be worthwhile. Not only may this be the case for certain 'hypertensive' individuals, but also for certain individuals who, although considered 'normotensive', are at high risk for some other reason (for example, as a result of a previous stroke or myocardial infarction).

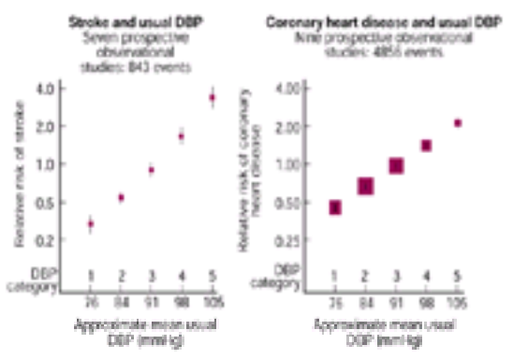

Fig. 6 Relative risks of stroke and coronary heart disease for five categories of diastolic blood pressure from the combined results of prospective observational studies. Solid squares represent the relative risks of disease in each category relative to the risk in the whole study population, and 95 per cent confidence intervals are denoted by vertical lines (MacMahon, 1994).

After making due allowance for the substantial and systematic extent to which the true relationship is diluted by purely random fluctuations in the baseline measurements of blood pressure (that is, the 'regression dilution' bias), the prospective studies suggest that a **prolonged** difference of only 5 mmHg in usual diastolic blood pressure is associated with avoidance of at least one-third of the risk of stroke and at least one-fifth of the risk of coronary heart disease in late middle age. However, although non-randomized, prospective observational studies may be more relevant to the eventual effects of prolonged differences in blood pressure, despite the possibility of confounding by other factors, randomized trials of blood pressure reductions that last for only a few years may be more relevant to assessing the speed with which the epidemiologically expected reductions in stroke or coronary heart disease risk are produced by reducing blood pressure. By comparing the results of a systematic overview of all randomized trials of antihypertensive therapy with the observational epidemiological evidence, it may be possible to estimate the extent to which the eventual effects of a lower blood pressure on disease incidence rates can be achieved within just a few years of treatment in middle or old age. (Ideally, these age ranges should be considered separately, as the fractional avoidance of risk may well be substantially different in middle and in old age.)

Over the past few decades, numerous trials of the treatment of hypertension have been conducted to determine whether blood pressure reduction in middle age reduces the risk of stroke and coronary heart disease. However, although it was fairly rapidly accepted that the treatment of severe hypertension could at least prevent stroke, until recently there has been controversy as to whether the treatment of even severe hypotension could also prevent coronary heart disease. Moreover, questions have also persisted about the effects of the treatment of mild to moderate hypertension on stroke. This continuing uncertainty about the benefits of lowering blood pressure may chiefly have reflected the inability of individual trials (even those with several hundred coronary heart disease events) to detect moderate coronary heart disease reductions reliably, rather than from any important heterogeneity of the real effects of treatment. The mean difference in diastolic blood pressure between treatment and control groups in the trials was only about 5 to 6 mmHg, and the epidemiological evidence suggests that a long-term difference of this magnitude is associated with only about 20 to 25 per cent less coronary heart disease (and about 35–40 per cent less stroke). Even if such trial treatments would eventually produce between 20 and 25 per cent less coronary heart disease after many years, the effects seen within the 2 or 3 years that are available on average between randomization and death in a 5-year trial might well be somewhat smaller (for example, 15 per cent). Considered separately, however, none of the trials recorded enough coronary heart disease events (or enough vascular deaths) for a statistically reliable assessment of 15 per cent risk reductions.

For stroke, the overview of randomized trials provides direct and highly significant evidence that most, or all, of the stroke avoidance associated with a prolonged difference in usual diastolic blood pressure appears soon after the blood pressure is lowered (Fig. 7). In contrast, the significant reduction in coronary heart disease

seen in the trials (16 per cent, SD 4; 95 per cent confidence interval of 8–23 per cent; 2 *p* = 0.0001) falls somewhat short of the difference of about 20 to 25 per cent suggested by the observational epidemiological evidence for a prolonged 5- to 6-mmHg difference in usual diastolic blood pressure. However, this coronary heart disease reduction is substantial and real (2 *p* = 0.0001). Therefore it is reasonable to hope that trials of antihypertensive regimens that can reduce blood pressure to a greater extent than the 5 to 6-mmHg diastolic blood pressure reduction seen in these trials will demonstrate even greater reductions in stroke and coronary heart disease.

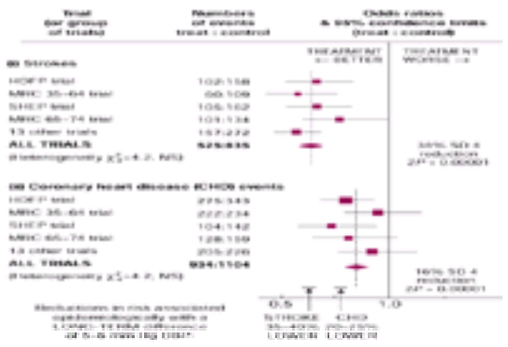

Fig. 7 Reduction in the odds of stroke and coronary heart disease in all unconfounded randomized trials of antihypertensive drug treatment (mean diastolic blood pressure differences of 5–6 mmHg for 5 years). Solid squares represent the odds ratios (treatment:control) for the four larger trials and the properly stratified odds ratio for the combination of the 13 smaller trials. 95 per cent confidence intervals are denoted by horizontal lines (for individual large trials or the combined small trials) and by diamonds (for overviews of all trials) (Collins and Peto, 1994).

The proportional reduction in vascular disease risk observed in the trials appeared to be similar in high- and low-risk individuals, so that the absolute size of the reduction that is produced by treatment may be largely dependent upon the absolute risk. Therefore, for high-risk individuals, the absolute risk reduction produced by antihypertensive treatment might be substantial even among those who are only moderately 'hypertensive'. Indeed, in view of the epidemiological evidence that, for stroke and coronary heart disease risk, there is no 'threshold' level of diastolic blood pressure within the normal range, large randomized trials might even show that blood pressure reduction is of substantial value among many 'normotensive' individuals at high risk of stroke (such as those with a history of cerebral vascular disease) or of coronary heart disease (such as patients with a history of myocardial infarction, angina, peripheral vascular disease, diabetes, or chronic renal failure).

Results from large anonymous trials are relevant to real clinical practice

A clinician is used to dealing with individual patients, and may feel that the results of large trials somehow deny their individuality. This is almost the opposite of the truth, for one of the main reasons why trials have to be large is just because patients are so different from one another. Two apparently similar patients may run entirely different clinical courses, one remaining stable and the other progressing rapidly to severe disability or early death. Consequently, it is only when really large groups of patients are compared that the proportion of patients with a truly good and bad prognosis in each can be relied on to be reasonably similar. One commonly hears statements such as: 'If a treatment effect isn't obvious in a couple of hundred patients then it isn't worth knowing about'. As the previous examples demonstrate, such statements may reveal not clinical wisdom but statistical naïvety.

It is also said that what is really wanted is not a blanket recommendation for everybody, but rather some means of identifying those few individuals who really stand to benefit from therapy. If any criteria (for example, a short-term response to a non-placebo-controlled course of some disease-modifying agent) can be proposed that are likely to discriminate between people who will and will not benefit, then these can be recorded prospectively at entry and the eventual trial results subdivided with respect to them. However, there is a danger in too detailed an analysis of the apparent response of small subgroups chosen for separate emphasis, because of the apparently remarkable effects of treatment in these subgroups. Even if an agent brought no benefit, it would have to be acutely poisonous for it not to appear beneficial in one or two such subgroups! Conversely, if an intervention really avoids an approximately similar proportion of the risk in each category of patient, it will, by chance alone, appear not to do so in some category. The surprising extent to which this happens is evident from the example in Table 2. A large anonymous trial will at least still help to answer the practical question of whether, on average, a policy of widespread treatment (except where clearly contraindicated) is preferable to a general policy of no immediate use of the treatment (except where clearly indicated). Moreover, without a few really large trials it is difficult to see how else many such questions could be resolved over the next few years. For example, digitalis has already been in use for over two centuries, and there is still no reliable consensus as to its net long-term effects on mortality. Trials are at least a practical way of making some solid progress, and it would be unfortunate if desire for the perfect (that is, knowledge of exactly who will benefit from treatment) were to become the enemy of the possible (that is to say, knowledge of the direction and approximate size of the effects of the treatment of many large categories of patient).

Further reading

Antiplatelet Trialists' Collaboration (1994). Collaborative overview of randomised trials of antiplatelet therapy. I: Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *British Medical Journal* **308**, 81–106.

Antithrombotic Trialists' (ATT) Collaboration (writing committee: C Baigent, C Sudlow, R Collins, R Peto) (2002). Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high-risk patients. *British Medical Journal* **324**, 71–86.

Armitage P, Berry G (1994). *Statistical methods in medical research*, 3rd edn. Blackwell Science, Oxford.

Chalmers I (1994). The Cochrane Collaboration: preparing, maintaining and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences* **703**, 156–63.

Chalmers TC, Lau J (1993). Meta-analytic stimulus for changes in clinical trials. *Statistical Methods in Medical Research* **2**, 161–72.

Cochrane AL (1979). 1931–1971: a critical review, with particular reference to the medical profession. *Medicines for the year 2000*, pp 1–11. Office of Health Economics, London.

Collins R, Peto R (1994). Antihypertensive drug therapy: effects on stroke and coronary heart disease. *Textbook of hypertension*, p 1156. Blackwell Science, Oxford.

Collins R, MacMahon S (2001). Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials, *Lancet* **357**, 373–80.

Collins R, *et al.* (1987). Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. *Statistics in Medicine* **6**, 245–50.

Collins R, Doll R, Peto R (1992). Ethics of clinical trials. *Introducing new treatments for cancer: practical, ethical and legal problems*, p 49. Wiley, New York.

Early Breast Cancer Trialists' Collaborative Group (1992). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. *Lancet* **339**, 1–15, 71–85.

Early Breast Cancer Trialists' Collaborative Group writing committee (Clarke M, Collins R, Davies C, Godwin J, Gray R, Peto R) (1998). Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet* **351**, 1451–67.

Early Breast Cancer Trialists' Collaborative Group (writing committee: Clarke M, Collins R, Davies C, Godwin J, Gray R, Peto R) (1998). Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet* **352**, 930–42.

European Carotid Surgery Trialists' Collaborative Group (1991). MRC European Carotid Surgery Trial: interim results for symptomatic patients with severe (70–99 per cent) or with mild (0–29 per cent) carotid stenosis. *Lancet* **337**, 1235–43.

Fibrinolytic Therapy Trialists' Collaborative Group (1994). Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. *Lancet* **343**, 311–22.

Heart Protection Study Collaborative Group (writing committee: Collins R, Armitage J, Parish S, Sleight P, Peto R) (2002). MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in

20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* **360**, 7–22.

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* **ii**, 349–60.

ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group (1995). ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58050 patients with suspected acute myocardial infarction. *Lancet*.

MacMahon S (1994). Blood pressure and the risks of cardiovascular disease. *Textbook of hypertension*, p 46. Blackwell Science, Oxford.

MacMahon S, Collins R (2001). Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. *Lancet* **357**, 455–62.

Peto R, *et al.* (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part I: Introduction and design. *British Journal of Cancer* **34**, 585–612.

Peto R, *et al.* (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part II: Analysis and examples. *British Journal of Cancer* **35**, 1–39.

Yusuf S, Collins R, Peto R (1984). Why do we need some large, simple randomized trials? *Statistics in Medicine* **3**, 409–20.

2.5 Complementary and alternative medicine

*E. Ernst**

Definition

Most doctors feel they know intuitively what is meant by complementary and alternative medicine (**CAM**). Yet an adequate definition is hard to find. Often CAM is described by characteristics that exclude it from mainstream medicine, for example:

- not taught in medical school
- not scientifically proven
- not based on a scientific rationale
- not used in routine health care

CAM can be positively defined as 'diagnosis, treatment, and/or prevention which complements mainstream medicine by contributing to a common whole, by satisfying a demand not met by orthodoxy or by diversifying the conceptual frameworks of medicine'.

CAM encompasses a large variety of techniques which have little in common except that they are excluded from mainstream medicine, claim to offer help for every condition, and pride themselves on a holistic approach to patient care (Table 1). Some relate to therapeutic modalities (e.g. herbalism), some to diagnostic techniques (e.g. iridology), and some include both diagnostic and therapeutic modalities (e.g. acupuncture).

There are considerable local differences in what is regarded as CAM or mainstream medicine. In Germany, massage therapy and herbalism are orthodox whereas in English-speaking countries they are usually regarded as CAM. Acupuncture is CAM in the West, while in China it is a widespread, traditional, and accepted form of treatment.

Since most CAM therapies are used as adjuncts to conventional treatments, 'complementary' is a more appropriate term than 'alternative'. When used as a true alternative to mainstream medicine, CAM almost invariably becomes a hazard to patients.

Prevalence

In the United States the prevalence of CAM increased from 33 to 42 per cent in the general population between 1990 and 1997, involving an annual expenditure exceeding US$20 billion. In the United Kingdom, the figures are 20 per cent and £1.6 billion, respectively.

In industrialized countries, typical users of CAM are:

- middle aged
- female
- well educated
- high socio-economic class

Indications for CAM range from chronic benign conditions where mainstream medicine is unable to offer a cure (e.g. back pain) to life-threatening diseases like cancer and AIDS. Most patients try CAM in parallel with conventional treatment, yet 30 to 50 per cent do not tell their doctor. A comprehensive medical history should therefore include questions about CAM.

Reasons for CAM's popularity

The following motivations may be important:

- to leave no option untried
- to take control over one's own health
- to accord one's health care with one's (slightly alternative) world views
- to be given time, understanding, and empathy by a practitioner
- to avoid adverse effects of conventional treatments

Disenchantment with orthodox medicine is a reason for trying CAM that should be taken seriously.

Examples of CAM methods

Acupuncture

Description

The Chinese believed that the life energy flowing in particular channels (meridians) govern the human body; the energy is a balance of opposite characteristics: yin and yang. Illness was understood as an expression of an imbalance between yin and yang. One way of re-establishing the proper equilibrium would be to insert needles in acupuncture points located along the meridians. Instead of needles one can also use pressure (acupressure), laser light (laser acupuncture), electrical currents (electroacupuncture), or heat (moxibustion). Neither the meridians nor the acupuncture points have a morphological basis and the philosophy of yin and yang is unscientific.

Mode of action

Nevertheless, modern neurophysiological research has created a (hypothetical) rationale for acupuncture: activation of brainstem nuclei and the release of neural transmitters and endorphins in the brain and descending inhibitory control systems.

There are considerable differences between traditional Chinese and Western acupuncture. With the former, no conventional diagnoses are sought, treatment is highly individualized according to each patient's particular yin/yang imbalance and is considered as a 'cure all'. Western acupuncturists tailor the treatment to the conventional diagnosis established beforehand and normally strive to identify those diagnoses for which acupuncture is helpful.

Efficacy

Rigorous trials are possible but fraught with methodological problems, for example:

- What is an adequate sham procedure?
- How can the patient be blinded?
- How can the therapist be blinded?

Several systematic reviews and meta-analyses of clinical trials of acupuncture for defined conditions have been published suggesting that acupuncture is effective for the following conditions:

- back pain
- nausea and vomiting
- dental pain
- migraine

In the following conditions, results are inconclusive:

- addictions (other than nicotine)
- asthma
- headache
- inflammatory rheumatic conditions
- neck pain
- osteoarthritis
- stroke

Acupuncture is no more effective than sham acupuncture or other control interventions for smoking cessation and weight reduction.

Safety

Serious complications include:

- trauma (e.g. cardiac tamponade, pneumothorax)
- infections (e.g. viral hepatitis)

Phytotherapy

Description

Medical herbalism (phytotherapy) is treatment with whole plants, parts of plants, or plant extracts. The term does not cover treatment with single active constituents such as acetylsalicylic acid, originally from willow bark.

Since all plants contain a multitude of chemicals, phytotherapy involves treatment with a mixture of potentially active compounds. In many cases there is uncertainty about the active ingredients and their pharmacological actions. Herbalists claim that the whole plant (extract) will yield more beneficial effects than any single isolated ingredient.

Most medical cultures have their version of traditional herbalism. Traditional Chinese medicine has a long history of employing mixtures of herbs to prevent and treat disease. This tradition was modified by the Japanese and resulted in Kampo medicine. The Indian tradition has generated Ayurvedic medicine, which relies heavily on plant-based remedies. Likewise, European herbalism has a tradition which is as old as European medicine itself. The scientific investigation of medicinal herbs is, however, a relatively recent innovation.

Mode of action

There are few differences in principle between pharmacotherapy and phytotherapy except that herbal remedies are multicomponent systems which render them pharmacologically more complex. There is no reason why the rules of pharmacokinetics and pharmacodynamics do not apply. For every plant-based medicine discernible modes of action exist. In some cases these have been elucidated; in many other cases they are still hypothetical.

Efficacy

Based on authoritative systematic reviews and meta-analysis, good evidence exists for the efficacy of the following herbal remedies:

- garlic for hypercholesterolaemia
- ginger for nausea and vomiting
- *Ginkgo biloba* for intermittent claudication
- *Ginkgo biloba* to delay the clinical deterioration in dementias
- horse chestnut seed extract for primary venous insufficiency
- kava as an anxiolytic drug
- peppermint oil for irritable bowel syndrome
- saw palmetto for benign prostatic hyperplasia
- St John's Wort for mild to moderate depression

For many other popular medicinal herbs, too few clinical trials have been carried out, or the studies are methodologically flawed, or their results are contradictory. The efficacy of such popular herbal remedies as valerian, aloe vera, and ginseng is undetermined.

Safety

Many medicinal herbs have serious adverse effects, for example:

- aconite (cardiotoxic)
- aristolochia (nephrotoxic)
- broom (cardiotoxic)
- chaparal (nephrotoxic)
- comfrey (hepatotoxic)
- liquorice root (hypokalaemia)
- pennyroyal (hepatotoxic)
- skullcap (hepatotoxic)

Herbal remedies can also interact with synthetic drugs (Table 2), and Asian herbal medicines have been shown repeatedly to be adulterated with synthetic drugs or heavy metals. In many countries (e.g. the United Kingdom and the United States) herbal medicines are marketed as food supplements with no stringent quality

control.

Homoeopathy

Description

Samuel Hahnemann, a German physician, believed in two major principles which formed the basis of an entirely new school of medicine, homoeopathy. The first is known as the 'like cures like' principle. Put simply, it postulates that if a given drug induces symptoms (e.g. a headache) in healthy individuals, this very drug can be employed to treat headaches in patients who suffer from this symptom. The second is that 'potentizing' (i.e. shaking and stepwise diluting) drugs makes them more potent for the treatment of illness. Homoeopathic dilutions prepared thus are believed to be clinically more effective than placebo even if not a single molecule of the original medicine is contained in the potentized remedy. Scientists have for 200 years pointed out that homoeopathy cannot possibly work beyond a placebo effect, but homoeopaths insist that homoeopathic remedies work via 'energy' transfer from the original substance to the diluent (the theory of a 'memory of water').

Homoeopaths do not treat diseases but claim to treat the whole individual. A homoeopath would take a detailed history at each patient's first visit. The aim is to match the totality of the symptoms and characteristics of that patient with a 'drug picture' according to the 'like cures like' principle. This homoeopathic remedy given in the correct potency should then be the optimal treatment for that patient. Clinical improvement may, however, take weeks or months, and in about 20 per cent of all cases symptoms may deteriorate before they become better, a phenomenon termed 'homoeopathic aggravation'.

Homoeopathy has to be seen in its historical context. At the time of Hahnemann it was an important discovery—there were very few effective treatments and many that were overtly harmful. At the very least homoeopathic remedies had virtually no adverse effects. And, if nothing else, Hahnemann can be credited with clinically exploiting the placebo effect to the best benefit of his patients. It is therefore hardly surprising that homoeopathy conquered many countries (e.g. France, the United States, India, and South America) by storm. The advent of more effective and less harmful synthetic drugs eventually led to the decline of homoeopathy. The recent boom of CAM has been associated with a strong revival in homoeopathy.

Mode of action

Even though several hypotheses have been developed to explain the transfer of 'energy' from the mother tincture to the diluent, none have so far withstood the scrutiny of independent assessment. Neither has the 'energy' ever been defined in physical terms, nor are there rational explanations as to how this 'energy' (if it exists) might induce a healing process in a diseased body or organ. Homoeopathy is, therefore, among the least plausible forms of CAM.

Efficacy

A meta-analysis of all 89 randomized or placebo-controlled clinical trials published by 1995 calculated an overall odds ratio of 2.45 in favour of homoeopathy. When only the 26 most rigorous studies were meta-analysed the odds ratio fell to 1.66 but remained statistically significant. However, this publication was criticized for pooling data for all medical conditions and all homoeopathic remedies and for including trials that were not randomized nor placebo-controlled and studies of material (low dilution) remedies where efficacy is not disputed.

The results of further systematic reviews are as follows.

1. The most frequently tested homoeopathic remedy (arnica) has not been conclusively shown to be efficacious beyond a placebo effect by two independent research groups.
2. The clinical condition which has been tested more than any other (delayed-onset muscle soreness) does not respond to homoeopathic remedies better than it responds to placebo; neither do clinical conditions that are common in everyday homoeopathic practice (e.g. migraine or headaches).

Safety

Highly diluted homoeopathic remedies are probably safe. Whether 'homoeopathic aggravations' represent a safety issue is unclear at present. One 'indirect' safety problem deserves to be mentioned: homoeopaths who are not medically qualified tend to advise their clients against immunization. If this happens on a large scale, we are in danger of losing herd immunity against important infectious diseases.

Other forms of CAM

CAM is a highly diverse field comprising more than 150 different forms of therapeutic and diagnostic methods (Table 1).

*The constructive comments of Ted Kaptchuk, Harvard School of Medicine, Boston, United States and Adrian White, University of Exeter, United Kingdom are thankfully appreciated.

Further reading

Ernst E, ed. (2000). *Herbal medicine—a concise overview for healthcare professionals*. Butterworth Heinemann, Oxford.

Ernst E, Hahn EG, eds (1998). *Homoeopathy. A critical appraisal*. Butterworth Heinemann, Oxford.

Ernst E, Pittler MH, Stevinson C, White A, Eisenberg D. (2001). *The desktop guide to complementary and alternative medicine*. Mosby, Edinburgh.

Ernst E, White A, ed. (1999). *Acupuncture a scientific appraisal*. Butterworth Heinemann, Oxford.

Fetow CW, Avila JR (1999*). Professional's handbook of complementary and alternative medicine*. Springhouse, Pennsylvania.

Jonas WB, Levin JS (1999). *Essentials of complementary and alternative medicine*. Lippincott, Williams, Wilkins, Philadelphia.

Schulz V, Häusel R, Tyler VE (1998). *Rational phytotherapy*. Springer Verlag, Berlin.

3.1 The global burden of disease study

*C. J. L Murray and A. D. Lopez**

Introduction

Reliable, up-to-date epidemiological information is required so that health authorities can assess priorities and evaluate their health systems. National and subnational mortality and morbidity statistics have been published by many countries for several decades. However, before the Global Burden of Disease (**GBD**) Study began in 1992, there had been no attempt to estimate and project the burden of disease and injury globally and regionally, using the same methods and expressing results in a common unit of measurement.

One goal of the GBD Study was to include measures of morbidity in debates about international health policy, which had largely drawn on the available mortality data, much of it referring to children. There was a need to separate epidemiological assessment from advocacy so that estimates of the mortality or disability from a condition were developed as objectively as possible. There was also a need to quantitate the burden of disease using a measure suitable for cost-effectiveness analysis of intervention packages. The GBD method quantifies the impact of premature death and disability on a population, combining these measures into a single unit of measurement of the overall burden of disease in that population. The Study presented the first global and regional estimates of disease and injury burden attributable to 10 important risk factors such as tobacco, alcohol, poor water and sanitation, and unsafe sex. Quantifiable estimates and projections of disease and injury burden from various exposures, measured in a similar fashion, are a key input into priority setting and for policy debates.

In the original GBD Study, 1990 was chosen as the base year for estimating disease burden. A revised Study is now underway to estimate the Global Burden of Disease in 2000, using more extensive and recent data sources, and improved methods, with a broader range of diseases and risk factors. Results of the GBD 2000 Study will be published in 2003; preliminary findings are given in this chapter.

Measuring disease burden

To combine the burden of premature mortality and disability into one summary measure requires a common unit of measurement. Since the late 1940s, it has been generally agreed that time is an appropriate measure: time (in years) lost through premature death, and time (in years) lived with a disability. A range of these time-based measures has been used in different countries, many of them variants of the so-called Quality-Adjusted Life Year or **QALY**. For the GBD, an internationally standardized form of the QALY was developed, called the Disability-Adjusted Life Year (**DALY**). This expresses years of life lost to premature death and years lived with a disability of specified severity. One DALY is one lost year of healthy life. Premature death is defined as occurring before the age to which a person could have expected to survive if he or she were a member of a model population whose life expectancy at birth was approximately equal to that of the world's longest-surviving population, the Japanese.

To calculate total DALYs for a given condition in a population, years of life lost (**YLLs**) and years lived with disability (**YLDs**) for that condition are estimated, and added together. For example, to calculate DALYs incurred through road traffic accidents in India in 1990, the total years of life lost in fatal road accidents must be added to the total years of life lived with disabilities by survivors of such accidents, weighted by the severity of the disability.

To assess premature mortality, the Study used standard life tables for all populations, with life expectancies at birth fixed at 82.5 years for women and 80 years for men. A standard life expectancy allows deaths at the same age to contribute equally to the burden of disease, irrespective of where the death occurs. Other methods, such as using different life expectancies for different populations that more closely match their actual life expectancies, violate this egalitarian principle. As life expectancy is rarely equal for men and women, the GBD assigned men a lower reference life expectancy than women, the magnitude of the difference (2.5 years) being an estimate of the biological advantage of females.

If people are forced to choose between saving a year of life for a 2-year-old and a 22-year-old, most prefer to save the 22 year-old. A range of studies confirm this broad social preference to value a year lived by a young adult more than one lived by a very young child or an older adult. Adults are widely perceived to play a critical role in the family, community, and society. This is why the GBD Study incorporated age-weighting into the DALY. It was assumed that the relative value of a year of life rises rapidly from birth to a peak in the early twenties, after which it steadily declines.

People commonly discount future benefits in the same way that they may discount future against current wealth. Whether a year of healthy life, like money, is also preferred now rather than later, is a matter of debate among economists, medical ethicists, and public health planners, since discounting future health affects both measurements of disease burden and estimates of the cost-effectiveness of an intervention. There are arguments for and against discounting. In the GBD Study, future life years were discounted by 3 per cent per year. This means that a year of healthy life bought for 10 years in the future is worth around 24 per cent less than one bought for now, as discounting is represented as an exponential decay function. Since the impact of discounting is significant, the findings of the GBD Study were published based on DALYs with and without discounting. Discounting future health reduces the relative impact of a child death compared to an adult death. It also reduces the value of interventions that provide benefits largely in the future, such as vaccinating against hepatitis B, which may prevent liver cancer, but some decades later.

In order to measure time lived with a non-fatal disease and assess disabilities in a way that will help to guide health policy, disability must be defined, measured, and valued in a clear framework that inevitably involves simplifying reality. There is surprisingly wide agreement between cultures about what constitutes a severe or a mild disability. For example, a year lived with blindness appears to most people to be a more severe disability than a year lived with watery diarrhoea, while quadriplegia is regarded as more severe than blindness. These judgements must be made formal and explicit if they are to be incorporated into measurements of disease burden.

To formalize social preferences for different states of health, the GBD Study developed a protocol based on the person trade-off method. In a formal exercise involving health workers from all regions of the world, the severity of a set of 22 indicator disabling conditions—such as blindness, depression, and conditions that cause pain—was weighted between 0 (perfect health) and 1 (equivalent to death). These weights were then grouped into seven classes, where class I has a weight between 0.00 and 0.02 and class VII a weight between 0.7 and 1. Subsequent valuations carried out in various cultures have closely matched the results of the original GBD exercise. For the GBD 2000 Study, disability weights are being determined from an extensive household survey programme to obtain valuations based on a visual analogue scale, with required valuations to be determined on the basis of a more intensive valuation exercise among health professionals to correct for

interval-scale biases among respondents.

Sensitivity analyses

To gauge the impact of changing these social choices on the final measures of disease burden, the GBD assessments were recalculated with alternative age-weighting and discount rates, and with alternative methods for weighting the severity of disabilities. Overall, the rankings of diseases and the distribution of burden by broad cause group are largely unaffected by age-weighting, and only slightly affected by changing the method for weighting disability. Changes in the discount rate, by contrast, may have a more significant effect on the overall results. A higher discount rate results in an increased burden in older age groups, while a lower discount rate results in an increased burden in younger age groups. Changes which affect the age distribution of burden, in turn, affect the distribution by cause, as communicable and perinatal conditions are most common in children while non-communicable diseases are most common in adults. The most significant effect of changing the discount and age weights is a reduction in the importance of several psychiatric conditions.

Ultimately, however, the accuracy of the underlying basic epidemiological data from which disease burden is calculated will influence the final results much more than the discount rate, the age weight, or the disability weighting method. If, for example, estimates of the incidence of blindness are incorrect by a factor of two, the results, whatever the social values used in the unit of measurement, will be substantially incorrect. We conclude that much more effort needs to be invested in improving the basic epidemiological data than in analysing the effects of what are ultimately minor adjustments to the particular summary measure of population health employed.

Estimating mortality and disability

Classification

As most developing countries still have only limited information about the distribution of causes of death in their populations, a primary objective of the GBD has been to develop comprehensive, internally consistent, mortality estimates worldwide for each major cause in 1990. Deaths were classified using a tree structure, in which the first level of disaggregation comprises three broad cause groups:

- *group I*—communicable, maternal, perinatal, and nutritional conditions (ICD-10 codes A00–B99, G00, N70–N73, J00–J06, J10–J18, J20–J22, H65–H66, O00–O99, P00–P96, E00–E02, E40–E46, E50, D50);
- *group II*—non-communicable diseases (ICD-10 codes C00–C97, D00–D48, D51–D89, E03–E07, E10–E16, E20–E34, E51–E89, F00–F99, G03–G99, H00–H61, H68–H95, I00–I99, J30–J99, K00–K92, N00–N64, N75–N99, L00–L99, M00–M99, Q00–Q99); and
- *group III*—injuries (ICD-10 codes V01–Y89).

Each group was then subdivided into categories: for example, cardiovascular diseases and malignant neoplasm are two subcategories of group II. Beyond this level, there are two further disaggregation levels such that 107 individual causes from the International Classification of Diseases (ICD) can be listed separately.

Consistent with the goal of providing disaggregated estimates of disease burden to help set priorities in the health sector, estimates for 1990 were prepared by age and sex and for eight broad geographical regions of the world (Fig. 1)

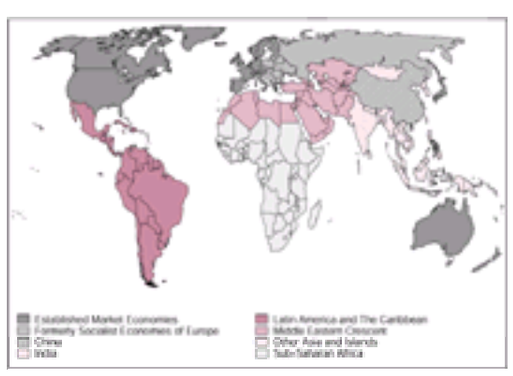

Fig. 1 Regions used for the GBD Study, 1990. EME, Established Market Economies; FSE, Former Socialist Economies of Europe; CHN, China; IND, India; LAC, Latin America and the Caribbean; MEC, Middle Eastern Crescent; OAI, other Asia [countries] and Islands; SSA, Sub-Saharan Africa.

For the GBD 2000 Study, the regional composition has been adapted to the six WHO regions. However, since these groups can be epidemiologically heterogeneous (for instance, the region of the Americas includes the United States, Canada, as well as Peru, Bolivia, and Haiti), each region has been subdivided into five categories of countries (labelled A, B, C, D, E) depending on the relationship between child and adult mortality in each country (Fig. 2).

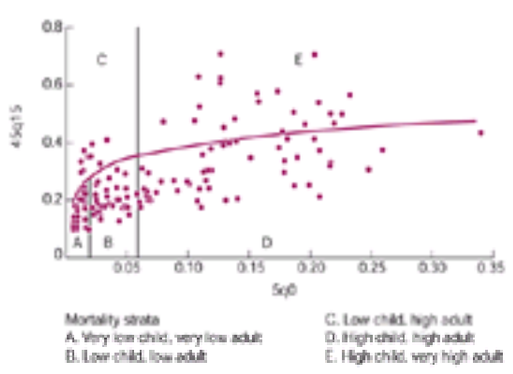

Fig. 2 Child and adult mortality, 1990.

Estimating regional mortality patterns

The Study derived cause-related mortality estimates by using the following four types of data:

1. *Vital registration systems*. Cause-of-death data certified by a physician have been assembled through vital registration systems for over 100 years in some European countries. Data for some 70 countries were available for the 1990s.
2. *Sample death registration systems.* In China, a set of 145 disease surveillance points, representative of both urban and rural areas, and covering about 10 000 000 people, provides useful mortality data. In India, Maharashtra state provides full medical certification for at least 80 per cent of urban deaths, while a rural surveillance system including more than 1300 primary healthcare centres nationwide was used to assess broad rural patterns of mortality. Reliable estimates of age-specific mortality rates in India can be derived, with appropriate adjustment, from the Sample Registration System covering a population of about 6 million but which is representative of all India.
3. *Epidemiological assessments*. Epidemiological estimates exist for specific causes in different regions. These estimates combine information from surveys on the incidence or prevalence of the disease with data on case-fatality rates for both treated and untreated cases.

4. *Cause-of-death models*. These are based on the fact that the broad cause structure of mortality is closely related to the level of mortality in a population. Such models estimate the distribution of deaths by cause in a population from historical studies of mortality patterns in countries with vital registration. The models developed for the initial GBD Study drew on a dataset of 103 observations from 67 countries between 1950 and 1991, and were used primarily to provide plausibility bounds on estimates derived from epidemiological assessments. For the GBD 2000 study, cause-of-death models have been used to ensure that the relative importance of groups I, II, and III as causes of death are consistent with historical observations about the cause structure of mortality, overall mortality, and economic development.

Vital registration data, corrected where necessary for under-registration, were used to construct cause-specific mortality patterns for those regions where registration was complete or virtually complete. For other regions, sex-age-specific mortality rates were estimated from survey and census data on child mortality. Adult mortality levels were inferred from the new WHO model life tables.

Assessing disability

A disease or injury may have multiple disabling effects, or sequelae. To estimate the total burden of disability, the Study measured the amount of time lived with each of the various disabling sequelae of diseases and injuries, in both treated and untreated states, and weighted for their severity, in each population. In all, 483 disabling sequelae of disease and injuries were analysed for the Study, for all regions and age groups, and for both sexes.

Calculating the number of years lived with a disabling condition requires information about its incidence, the average age of onset, the average duration of the disability, and the severity weighting of the condition. Epidemiological experts were asked to estimate each variable for each condition based on a thorough review of published and unpublished studies. For each sequela; prevalence, case fatality, remission, and mortality were estimated. This information allowed correction of the preliminary estimates for internal consistency, ensuring that estimated prevalence and estimated incidence were consistent with one another. Consistency was validated using DisMod software specifically developed for the Study (Fig. 3). When inconsistencies were detected, epidemiological experts were asked to revise their initial estimates. The final disability estimates were the result of several rounds of revision over nearly 5 years.

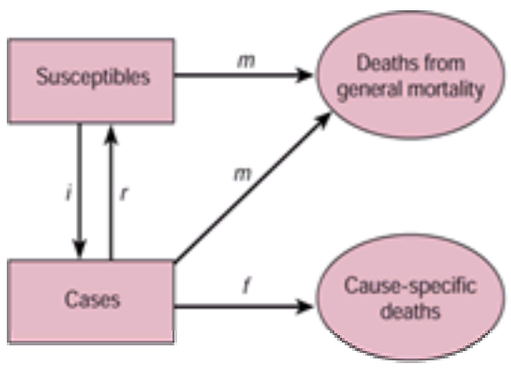

Fig. 3 Basic relationships between susceptibles, cases, and deaths used in developing DisMod.

The number of years that each person had lived with a particular disability was calculated from the incidence of the disability, with the 'stream' of disability arising from it measured from the age of onset, the estimated duration of the disability, multiplied by the condition's severity weight. To calculate the YLDs due to a condition in any given population, the number of YLDs lost per incident case must be multiplied by the number of incident cases. A case of asthma, for example, carries a disability weight of 0.1 if untreated and 0.06 if treated. If the annual incidence of asthma in males aged between 15 and 44 years is 1 million cases, the untreated proportion is 35 per cent, and the average duration is 7 years, then this sequela alone is estimated to cause 664 000 YLDs for that demographic group. Unlike the estimates of years of life lost, not all sequelae of all conditions could be explicitly assessed for YLDs. Estimates for conditions not explicitly considered were made on the basis of information about the ratio of total premature mortality to disability (YLDs) for each broad cause group.

The global burden of disease in 1990 and 2000: main findings

The results demonstrate clearly that disability plays a central role in determining the overall health status of a population. Yet that role has until now been almost invisible to public health. The leading causes of disability are shown to be substantially different from the leading causes of death, which has considerable implications for the practice of judging a population's health from its mortality statistics alone.

A key aim of the GBD was to measure the burden of fatal and non-fatal health outcomes in a single measure, the disability-adjusted life year (DALY). This section presents the main results of the assessments of overall burden for each region. To calculate DALYs due to each disease or injury in a given year and population, the years of life lost through all deaths in that year were added to the years of life expected to be lived with a disability for all new cases of disease or injury occurring in that year, weighted for the severity of the condition.

Regional imbalances in the burden of disease

Sub-Saharan Africa and India together accounted for more than 40 per cent of the total global burden of disease in 1990, although they make up only 26 per cent of the world's population. By contrast, the Established Market Economies and the Formerly Socialist Economies of Europe, with about one-fifth of the world's population between them, together bore less than 12 per cent of the total disease burden. China emerged as substantially the most 'healthy' of the developing regions, with 15 per cent of the global disease burden and one-fifth of the world's population. This means that about 579 years of healthy life were lost for every 1000 people living in Sub-Saharan Africa, compared with just 124 for every 1000 people in the Established Market Economies (Fig. 4).

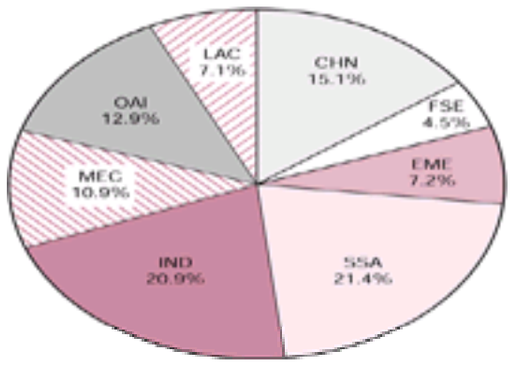

Fig. 4 Distribution of DALYs by region, 1990.

The Study found a sevenfold higher risk of a child dying (that is, a newborn dying before the age of 15) in Sub-Saharan Africa compared to a newborn in the Established Market Economies (Fig. 5). This extraordinary excess mortality in many developing regions must remain a priority for global health programmes. Somewhat surprisingly, the risk of adult death in the FSE region, at least for males, was higher than any other region of the world, except Africa (see Fig. 5). This reflects the rapid increase in adult male death rates in the Russian Federation and neighbouring countries since 1987. In 1990, mortality at these ages (15–59 years) was still rising rapidly in Russia, reaching a peak in 1994. Since then, the probability of death between the ages of 15 and 59 has declined as rapidly as it rose,

although evidence for 1999 suggests that death rates may be rising again; the trends for females are qualitatively similar, though less extreme.

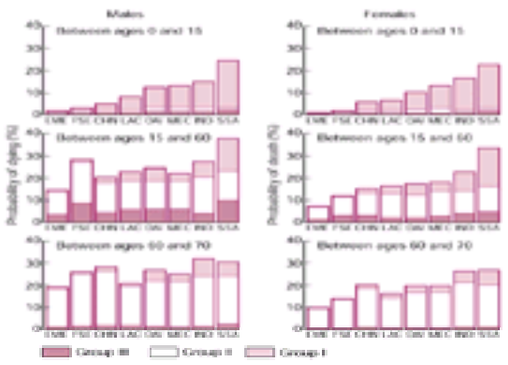

Fig. 5 Regional probabilities of death for males and females by age and group, 1990.

The leading causes of death for 2000 were similar to those estimated for 1990, with two notable exceptions. HIV/AIDS, which killed an estimated 300 000 people in 1990, claimed 10 times that number in 2000 and the annual mortality toll continues to rise. Conversely, interventions to control measles and diarrhoea have had a marked impact on mortality from the two conditions, with measles deaths falling from 2 million in 1990 to less than 800 000 in 2000. Diarrhoeal deaths had fallen to 2.1 million in 2000, 1.3 million of them in children under 5-years-old.

The GBD Study has provided support for the theory that people in high-income, low-mortality populations not only live longer, but also remain healthier for longer. In recent years, opinion has been divided between the view that ill health is compressed into the last few years of life in these populations, and the view that longer life merely exposes people to a longer period of poor health. The results suggest that older people in the developed world are healthier than their counterparts in developing countries. It was also found that babies born in Sub-Saharan Africa could expect to spend about 15 per cent of their lifespan disabled, compared to 8 per cent for babies born in the Established Market Economies. The 60-year-olds in Sub-Saharan Africa can expect to spend about half their remaining years with a disability, whereas 60-year-olds in the Established Market Economies are likely to spend one-fifth of those years disabled. The results suggest that the proportion of the lifespan lived with a disability falls as life expectancy rises.

Major causes of disease burden

The leading causes of disease burden in 1990, were lower respiratory infections, diarrhoeal diseases, perinatal causes, and, perhaps unexpectedly, depression (Table 1). The Study showed that the burden of psychiatric disease had been heavily underestimated. Of the ten leading causes of disability worldwide (in YLDs) in 1990, five were psychiatric conditions: unipolar depression, alcohol abuse, bipolar affective disorder, schizophrenia, and obsessive–compulsive disorder. Unipolar depression alone was responsible for more than 1 in every 10 years of life lived with a disability worldwide. Together, psychiatric and neurological conditions accounted for 28 per cent of all YLDs, compared with 1.4 per cent of all deaths and 1.1 per cent of years of life lost. The predominance of these conditions is not restricted to wealthy countries, although their burden is highest there. They were the most important contributors to YLDs in all regions except Sub-Saharan Africa, where they still accounted for 16 per cent of the total.

Alcohol abuse was the leading cause of male disability, and the tenth largest in women, in developed regions and, perhaps surprisingly, the fourth largest cause in men in developing regions. Other important causes of YLDs were anaemia, falls, road traffic accidents, chronic obstructive pulmonary diseases, and osteoarthritis.

Traditional disease burdens in developing societies—communicable, maternal, perinatal conditions, and nutritional deficiencies—remained of major importance in the 1990s. Even though these group I conditions accounted for only 7 per cent of the burden in the Established Market Economies and less than 9 per cent in the Former Socialist Economies, they accounted for more than 40 per cent of the total global burden of disease in 1990, and for 49 per cent in developing regions. In Sub-Saharan Africa, 2 out of every 3 years of healthy life were lost because of group I conditions. Even in China, where the epidemiological transition is far advanced, a quarter of years of healthy life lost were due to this group. Worldwide, five out of ten leading causes of disease burden (as measured by DALYs) are group I conditions: lower respiratory infections (pneumonia); diarrhoeal disease; perinatal conditions; tuberculosis; and measles.

The burden of injury in 1990 was highest (19%) in the FSE region. China had the second largest injury burden, followed by Latin America and the Caribbean. Even in the Established Market Economies the burden of injuries—dominated by road traffic accidents—was almost 12 per cent of the total. In most regions, unintentional injuries were a greater source of ill health in 1990 than intentional injuries such as interpersonal violence and war. The only exception was the Middle Eastern Crescent, where unintentional and intentional injuries took an approximately equal toll because of a particularly high burden of war in the region at the time.

Preliminary estimates of GBD for 2000 suggest that the leading causes are similar to those in 1990, exception for HIV. In 2000, HIV is estimated to have killed 2.95 million people (2.4 million in Africa). HIV/AIDS is the third leading cause of disease burden, accounting for 6.1 per cent of all DALYs lost in 2000, only marginally behind lower respiratory infections (6.4 per cent) and perinatal conditions (6.2 per cent) (Table 1).

Global burden of disease: risk factors

Exposure to particular hazards, such as tobacco, alcohol, unsafe sex, or poor sanitation, can significantly increase people's risks of developing disease. Health policy makers need accurate information on their impact in order to devise effective prevention strategies. The GBD Study assessed, for the first time, the mortality and loss of healthy life that can be attributed to each of 10 major risk factors in each region: malnutrition; poor water supply, sanitation and personal/domestic hygiene; unsafe sex; tobacco abuse; alcohol abuse; occupation; hypertension; physical inactivity; illicit drug use; and air pollution.

The contributions of risk factors to global burden

Malnutrition, poor water, sanitation, and hygiene, unsafe sex, alcohol, tobacco, and occupation proved the most important, accounting together for more than one-third of the total disease burden worldwide in 1990 (see Table 2). Malnutrition and poor sanitation were the dominant hazards, responsible together for almost a quarter of the global burden. Unsafe sex and alcohol each contributed approximately 3.5 per cent, tobacco and occupation hazards just under 3 per cent each. These are similar to the disease burden due to tuberculosis or measles. Major inequalities exist between regions and between men and women in the burdens of most risk factors. The consequences of unsafe sex, including infections and complications of unwanted pregnancy, are borne disproportionately by women in all regions. In young adult women in Sub-Saharan Africa, unsafe sex accounts for almost one-third of the total disease burden. Tobacco, due to longer exposure, and alcohol caused their heaviest burdens in men in the developed regions, where they accounted together for more than one-fifth of the total burden in 1990. In Asia and other developing regions, the rapid increase in tobacco use over the past few decades is expected to kill many more people in the coming decades than have so far died of this cause in the developed regions.

For the GBD 2000 Study, more than 20 risk factors are being analysed, using a more comparable framework for assessing risk factors. Preliminary results for 2000 suggest little change in the importance of these various risk factors, except for tobacco and unsafe sex, for which disease burdens are rising rapidly, particularly in the developing world. More refined methods have suggested that the blood pressure burden is about twice what was estimated for 1990, although this increase is largely a methodological artefact. Elevated cholesterol is also a major cause of disease burden, about two-thirds that of blood pressure.

Projections of the global burden of disease

To plan health services effectively, policy makers need to know how health needs might change in the future. To meet this need, projections of mortality and disability have been developed for each 5-year period from 1990 to 2020.

Projection methods

Rather than attempt to model the effects of the many different determinants of disease from the limited data available, mortality change has been modelled as a function of a few socioeconomic variables: (1) income per capita; (2) the average number of years of schooling in adults, termed 'human capital'; and (3) time, a proxy measure for the secular improvement in health this century that results in part from accumulating knowledge and technological development. Historically, these variables have been related to mortality rates: for example, income growth was closely related to the improvement in life expectancy that many countries achieved in the twentieth century. Because of their relationships to death rates, these variables may be regarded as indirect, or distal, determinants of health. A fourth variable, tobacco use, was included, because of its overwhelming impact on the occurrence of chronic diseases, using information from more than four decades of research on the time lag between persistent tobacco use (measured as 'smoking intensity') and its effects on health.

Death rates for all major causes based on historical data for 47 countries from 1950 to 1991 were related to these four variables to generate the projections. A separate model was used for HIV, with modifications for the interaction between HIV and tuberculosis.

Mortality projections

In all regions, life expectancy at birth is expected to increase for women. By 2020, infant girls born in the Established Market Economies may expect to survive to almost 88 years. For men, life expectancy will grow much more slowly, mainly because of the impact of the tobacco epidemic. Nevertheless, by 2020, males born in Sub-Saharan Africa, whose life expectancy at birth was below 50 in 1990, may expect to reach 58 years. Males born in Latin America and the Caribbean, who in 1990 could have expected to live to 65, may expect to reach 71 years. However, for men in the Formerly Socialist Economies of Europe, life expectancy is not expected to increase at all between 1990 and 2020. This is partly due to the fact that life expectancy was falling in 1990, so that any positive change is likely to be merely recovering to the 1990 position.

In young children and adolescents under the age of 15, the risk of death is projected to decline dramatically in all regions, falling by about two-thirds in Sub-Saharan Africa and India. In adult women, too, the risk of death is expected to fall in all regions. Men in the Formerly Socialist Economies of Europe and China, because of the tobacco epidemic, may expect a higher risk of dying between the ages of 15 and 60 than they do today. In other regions, the risk of death for men in this age group is expected to fall, but more modestly than in women. Remarkably, by 2020, men of this age group in the Formerly Socialist Economies of Europe could face a higher risk of death even than men in Sub-Saharan Africa.

Deaths from communicable, maternal and perinatal conditions, and nutritional deficiencies (group I) are expected to fall from 17.3 million in 1990 to 10.3 million in 2020. As a percentage of the total burden, group I conditions are expected to drop by more than half, from 34 per cent to 15 per cent. This projected reduction overall, despite increased burdens due to HIV and possibly tuberculosis, runs counter to the now widely accepted belief that infectious diseases are making a comeback worldwide. It reflects, in part, the relative contraction of the world's 'young' population: the under-15 age group is expected to grow by only 22 per cent between 1990 and 2020, whereas the cohort of adults aged between 15 and 60 is expected to grow by more than 55 per cent. In addition, the projection reflects the observed overall decline in group I conditions over the past four decades, due to increased income, education, and technological progress in the development of antimicrobials and vaccines. Even under the pessimistic scenario, in which both income growth and technological progress are expected to be minimal, deaths from these conditions are still expected to fall slightly to 16.9 million.

It should not be taken for granted that the progress against infectious diseases during the past four decades will be maintained. Antibiotic development and other control technologies may not keep pace with the emergence of drug-resistant strains of important microbes such as *Mycobacterium tuberculosis*. If such a scenario were to prove correct, and if, in addition, case-fatality rates were to rise because of such drug-resistant strains, the gains of the twentieth century could be halted or even reversed. None the less, the evidence to date, suggests that, as long as current efforts are maintained, group I causes are likely to continue to decline.

While overall, group I conditions are expected to decline, deaths from non-communicable diseases are expected to climb from 28.1 million deaths in 1990 to 49.7 million in 2020, an increase of 77 per cent in absolute numbers. In proportionate terms, group II deaths are expected to increase their share of the total from 55 per cent in 1990 to 73 per cent in 2020. These global figures do not reveal the extreme nature of the change that is projected in some developing regions because they incorporate the projections for the rich nations, which show little change. In India, deaths from non-communicable diseases are projected almost to double, from about 4 million to about 8 million a year, while group I deaths are expected to fall from almost 5 million to below 3 million a year. In the developing world as a whole, deaths from non-communicable diseases are expected to rise from 47 per cent of the total to almost 70 per cent.

The steep projected increase in the burden of non-communicable diseases worldwide is largely driven by population ageing, augmented by the large numbers of people in developing regions who are now exposed to tobacco. Ageing will result in a rise in the absolute numbers of cases of non-communicable diseases and in their increased share of the total disease burden for the population as a whole, but not in any change in the rates of those diseases in any given age group. As studies in the Established Market Economics show, the age-specific rates of some important non-communicable diseases, such as ischaemic heart disease and stroke, have been falling steadily for at least two decades. Whether these rates are also falling in other regions is much less clear. However, any age-specific decrease in the rates of these diseases that may also emerge in low-income countries is likely to be outweighed by the large and demographically driven increase in the absolute numbers of adults at risk from these diseases, augmented by the tobacco epidemic. As with non-communicable diseases, deaths from injury are also expected to rise for mainly demographic reasons. Young adults are generally exposed to greater risks of injury.

Recent health trends in the 1990s: implications for the GBD projections

The original GBD projections were based on data and information about health conditions worldwide during the late 1980s/early 1990s. At that time, two major epidemics were affecting the health of large population groups: the HIV epidemic, particularly in Africa, which killed an estimated 300 000 people worldwide in 1990 but had, by that time, infected millions more; and the explosive increase in the adult mortality rates in Russia and neighbouring countries, particularly from cardiovascular diseases and injuries, and particularly among men. Making projections at a time of such dramatic epidemiological trends is extremely hazardous.

The 1990 GBD Study's HIV/AIDS projections have severely underestimated the spread of the epidemic in Sub-Saharan Africa, particularly in Southern Africa. By 2000, HIV/AIDS was estimated to have killed 2.4 million Africans, several times more than projected on the basis of what was known in 1990. Whether the disease burden will continue to rise, and how far, is uncertain and new projection methods are being developed to forecast the epidemic better, particularly in Africa.

The other large uncertainty in the projections, namely adult mortality in the FSE region, has confounded epidemiologists with the dramatic change in mortality risks during the 1990s. Death rates had been falling markedly in most large countries in this region, and, if they were to continue to do so, the 1990 forecasts will prove to be unduly pessimistic. It is too early to decide whether the recent declines in mortality are the beginning of a long-term secular trend in mortality. Recent evidence for 1999 and 2000 indicates that death rates, particularly in males, have begun to rise again.

Progress in refining the GBD approach

Measuring and evaluating health

An innovation of the GBD Study was the attempt to measure and evaluate states of ill health in a similar way in different societies. This presupposes a common conceptual framework and measurement strategy. In particular, what are the key domains of health that need to be assessed and what is the minimum number of items and response categories needed to measure them? Self-report instruments currently in use lack cross-cultural comparability, with the result that the measurement of health in various populations cannot be compared. The development and implementation of a conceptual framework to measure and describe health in a way that improves comparability across populations is a key challenge for research on burden of disease.

In the GBD Study, comorbid conditions were evaluated separately, and the time spent with these combined states was measured as the sum of the two. This additive model may not be appropriate. More data are required on the prevalence of major comorbidities to avoid multiple attribution in valuing health states.

Conclusions

The GBD Study has provided a picture of current and projected health needs. It has shown that non-communicable diseases are rapidly becoming the dominant causes of ill health in all developing regions except Sub-Saharan Africa; that mental health problems have been underestimated worldwide; and that injuries are important problems in all regions. The findings pose new and immediate challenges to policy makers.

---

*The authors are extremely grateful to Brodie Ferguson for his assistance in preparing this chapter.

Further reading

Barendregt JJ, Bonneux L, van der Maas, PJ (1998). Health expectancy: from a population health indicator to a tool for policy making. *Journal of Aging and Health* **10**, 242–58.

Murray CJL, Lopez AD, eds (1996). *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Global burden of disease and injury series, Vol. 1. Harvard University Press, Cambridge, MA.

Murray CJL, Lopez AD (1996). *Global health statistics: a compendium of incidence, prevalence and mortality estimates for over 200 conditions*. Global burden of disease and injury series, Vol. 2. Harvard University Press, Cambridge, MA.

Murray CJL, Lopez AD (1999). On the comparable quantification of health risks: lessons from the Global Burden of Disease Study. *Epidemiology* **10**, 594–605.

Murray CJL, Lopez AD (2000). Progress and directions in refining the global burden of disease approach: a response to Williams. *Health Economics* **9**, 69–82.

Murray CJL, Salomon JA, Mathers C (1999). *A critical examination of summary measures of population health.* GPE Working Paper Series, WHO, Geneva.

Peto R, *et al*. (1992). Mortality from tobacco in developed countries: indirect estimation from national vital statistics. *Lancet* **339**, 1268–78.

Sullivan DF (1971). A single index of mortality and morbidity. *Health Reports* **86**, 347–54.

Van de Water HP, Perenboom RJ, Boshuizen HC (1996). Policy relevance of the health expectancy indicator: an inventory of European Union countries. *Health Policy* **36**, 117–29.

Wilkins R, Adams O (1983). Health expectancy in Canada, late 1970's: demographic, regional and social dimensions. *American Journal of Public Health* **73**, 1073–80.

3.2Human population size, environment, and health

A. J. McMichael and J. W. Powles

Introduction

Homo sapiens originated approximately two hundred thousand years ago. Since then it has undergone three population growth surges: (i) an estimated 50-fold increase as hunter–gatherer humans drifted out of north-eastern Africa and dispersed around the world; (ii) a further 100-fold increase following the advent of agriculture, beginning ten thousand years ago; and then (iii) from just before the industrial revolution, another 10-fold increase from half a billion to six billion.

This third, incomplete, increase has occurred much faster than the two previous increases. Absolute additions to human numbers have been biggest in the past quarter-century, capping an almost-fourfold increase from 1.6 to 6 billion during the twentieth century. However, the annual increase expressed in percentage terms has slowed in the last couple of decades. Demographers expect that by the time the demographic transition is completed worldwide, and birth rates equilibrate with death rates (at historically low levels), world population will have reached between 8 and 11 billion. The medium variant projection for 2050 is approximately 9 billion.

Relationship between environment and population

The relationship between the environment and population size is multifaceted. The main components of that relationships are these: (i) the environment sets limits on the size of the supportable population; (ii) human societies find ways of extending that limit; and (iii) that extension process in turn often leads to the depletion and deterioration of the natural environment.

Carrying capacity

In the natural world, the composition and assets of a species' environment determine the maximum number of individuals of the species that can be supported. For that species, this number is the 'carrying capacity' of its local habitat. There are fluctuations of population size around that number as conditions vary over time. Populations of the human species, uniquely, are not fully constrained by given environmental conditions. Through culture and technology, humans can increase the carrying capacity of their local environments—at least temporarily.

The early domestication of plant species increased food yields and hence population carrying capacity. The subsequent domestication of wild animal species further increased carrying capacity. The advent of agriculture led to a substantial increase in fertility—from an average of 4 to 5 births per completed reproductive lifetime (as reported for traditional hunter–gatherers and, coincidentally, for great apes) to 5 to 7 births per reproductive lifetime in agrarian populations. This greater fecundity meant that the approximate equivalence of birth and death rates in slowly enlarging agrarian populations was attained at very high levels of both. This is well illustrated by India a century ago, when fertility was high (7 to 8 births per woman) and life expectancy was 20 to 25 years (reflecting especially the high death rates in infancy and childhood).

Overloading the environment

Human communities exploit local natural resource stocks: soil, water, and plants and animals for food and materials. This exploitation, in time, tends to deplete and ultimately degrade environments. Often, the local consequences have been the restriction or decline in human numbers, and impairment of nutritional status and health.

The extent to which humans are disrupting their environment has increased rapidly over the past two centuries, as numbers have expanded and as the material and energy intensities of productive activity have increased. Over the past century or so, adverse environmental effects have mostly been of a localized kind, such as urban-industrial air pollution, chemical pollution of waterways, and urban squalor. Today, human effects on the environment are being played out on a much larger scale—and the longer-term consequences for health could be commensurately more serious. Recent global assessments point to a significant and increasing 'ecological deficit', with manifest decline in natural environmental and ecological resource stocks.

Further, some of these environmental stresses are likely to cause tensions between human communities, leading to armed conflict—another potential source of damage to health. For example, Ethiopia and the Sudan, upstream of Nile-dependent Egypt, increasingly need the Nile's water for their own crop irrigation.

The central issue in all of this is that the ecological underpinnings of human health are being perturbed or depleted. The sustained good health of any population, over time, requires a stable and productive natural environment that: (i) yields assured supplies of food and fresh water, (ii) has a relatively constant climate in which climate-sensitive physical and biological systems do not change for the worse, and (iii) retains biodiversity (a fundamental source of both present and future value). For the human species, as a 'social animal'in the extreme, the richness, texture, and stability of the social environment (i.e. 'social capital') is also important to population health.

A Malthusian perspective on sustainability

Two hundred years ago, Thomas Malthus, responding to the utopian views of William Godwin, de Condorcet, and others about the perfectibility of human institutions, foresaw a potential crisis arising from excessive human numbers within a food-limited environment. The exponential power of population growth would, he concluded, tend always to outstrip the (arithmetic) power of growth in food production. He grimly predicted that population excess in Europe would lead to starvation and die-off—that is, to nature's 'positive checks' that would bring human numbers back in line with food supplies.

In fact, the crisis did not materialize in Europe. Malthus could not have foreseen the remarkable increase in food-producing capacity that the second agricultural revolution, underwritten by mechanization and fossil-fuel energy, would bring during the nineteenth century—or the bonanza of imported grain and meat that Europe's newly-established colonies would provide. Nor could he have foreseen the marked decline in fertility rates that emerged in European populations during the nineteenth century as social modernization occurred and as contraceptive possibilities became widely understood.

Today, nevertheless, a 'Malthusian' perspective is relevant at another level of analysis. In the past quarter-century we have begun to see the evidence that there are limits to the carrying capacity of the globe as a whole. There is mounting evidence, for example, of a damaged stratosphere and of human-induced climate change. These are signs that we are exceeding the carrying capacity of the world as a whole. What might lie in store?

In the natural world, the tendency of plant and animal species to exponential growth is generally constrained by predation, by limits to food supplies, by infectious disease, and, in many animals, by density-dependent changes in reproductive behaviour. As numbers increase, one of the following patterns operates:

1. logistic (asymptotic) growth, responding to immediate negative feedback, as carrying capacity is approached;

2. domed or capped growth, responding to deferred negative feedback, necessitating compensatory die-off; or
3. irruptive growth, with a chaotic post-crash pattern.

Our recognition, today, of the risk of overshoot and collapse (patterns 2 and 3) underlies the increasing attention being paid to the need for 'environmentally sustainable development'. Simplifying, there are two main adverse outcomes to which an excess of human numbers might contribute: (i) recurrent subsistence crises on a subnational or national level, or (ii) 'planetary overload'.

Local subsistence crises

The focus of concern with 'overpopulation' in much of the latter half of the twentieth century was the likelihood that the growth of many local populations would overload local carrying capacities. Chronic food shortages would ensue and undernutrition would force mortality rates up towards equilibrium with fertility, both at unfavourably high levels. The emiserating effects of population pressure would prevent economic development, prolong population growth, and leave such populations 'demographically entrapped'.

So far, however, there appear to have been few developments in this direction. The famines that have occurred in recent decades, the most serious being the Chinese famine of 1959 to 1961 with around 15 to 20 million deaths, have not been attributable primarily to the progressive reduction of food-producing resources per person. Rather, they appear to have arisen from either economic mismanagement (as in the case of the Chinese 'great leap forward') or warfare (as in parts of Africa)—although the contributory causes of some of that strife may include local population pressures on dwindling natural resources. The recent generally favourable trends in per-person food supplies may not be sustainable as populations double (or more) in size in poor countries such as Bangladesh that already have less than one-tenth of a hectare of cropland per person.

Planetary overload

While local subsistence pressures persist or increase in many parts of the world, there is a newer form of pressure arising at the global level. Human population size and the material intensity of our economies are now so great that, at that global level, we are beginning to disrupt some of the biosphere's life-support systems. The clearest evidence of the occurrence of 'global environmental change' is the documented destruction of stratospheric ozone, particularly in polar and subpolar regions, over the past quarter-century and the apparent incipient changes in world climate due to greenhouse-gas accumulation in the lower atmosphere. There is a net ongoing loss of productive soils on all continents; we have overfished most of the ocean fisheries; we have severely depleted many of the great aquifers upon which irrigated agriculture depends; we are extinguishing at an unprecedented overall rate whole species and many local population;, and, increasingly, persistent human-made organic (especially chlorinated) chemicals are pervading the biosphere.

These various changes are perturbing or weakening systems in the biosphere that provide the stabilization, replenishment, organic production, cleansing, and recycling that our predecessors were able to take for granted in a less populated, less degraded, world. Manifestly we no longer live in such a world. This weakening of Earth's basic life-supporting processes poses a spectrum of long-term risks to human population health. The currently foreseeable risks from these environmental changes include: (i) an anticipated increase in skin cancer rates (and perhaps in ocular disorders and immune system suppression) due to the approximately 10 per cent increase in ultraviolet radiation levels that has now accrued at middle latitudes; (ii) a likely increase in geographical range and seasonality of vector-borne infectious diseases such as malaria and dengue fever under conditions of climate change; (iii) increased exposure, in some parts of the world, to weather extremes and disasters, consequent upon climate change; and (iv) malnutrition and hunger in local populations whose agricultural productivity is adversely affected by changes in climate, soil fertility, freshwater supplies, and the ecology of pests and pathogens.

Contribution of population increase to environmental disruption

The World Wildlife Fund for Nature has analysed in detail the trends over the past three decades in the vitality and function of major categories of ecological systems, including freshwater, marine, and forest ecosystems. Overall, the 'Living Planet Index' has declined by 30 per cent since 1970. Assessments by other international agencies approximately concur.

The three main determinants of human disruption of the environment are population size, the level of material wealth and consumption, and the types of technology. The ongoing climate change debate illustrates well the relativities between the environmental effects of increases in population and consumption. Historically, during the twentieth century, as population increased by just under fourfold the annual fossil fuel emissions of carbon dioxide increased 12-fold. In 1995, the 20 per cent of world population living in high-emission countries accounted for 63 per cent of carbon dioxide emissions, while the lowest-emitting 20 per cent of population contributed just 2 per cent. Over the coming century the projected world population growth will contribute an estimated 35 per cent of growth in carbon dioxide emissions, whereas economic growth would account for the remaining 65 per cent.

If the world were to limit carbon dioxide build-up to a doubling of its preindustrial concentration (i.e. from 275 to 550 p.p.m.)—a level which climatologists think would be tolerable to most ecosystems—then the United Nations medium population projection of 10 to 11 billion by 2100 would allow per-person carbon dioxide emissions similar to those of the 1920 to 1930s. That is approximately two-thirds less than today's level of emissions (see Fig. 1). While that looms as a very demanding task, we actually already have much of the necessary technology to reduce emissions greatly without forfeiture of material standards of living. The real challenge is political—to transform current technologies and economic practice.

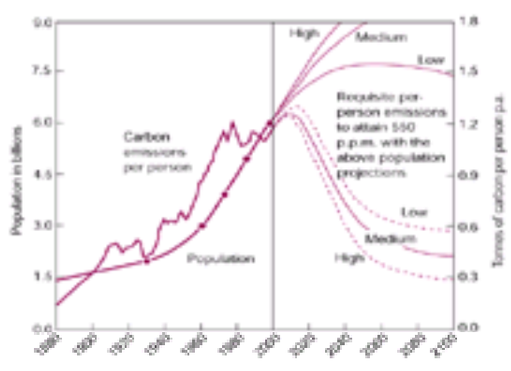

Fig. 1 This shows the configuration of relationships between economic activity, wealth distribution (especially poverty), population size, environmental conditions, and human health. Note that poverty and the material (natural and social) environment are major determinants of health.

Overall, then, the larger potential threat is not from the increase in human numbers *per se* but from mildly environmentally disruptive humans becoming highly disruptive humans—in other words, from a 'development' process that would generalize the patterns of production and consumption typical of today's rich countries. Current practices in rich countries cannot be applied to a human population likely to exceed 10 billion and demanding a higher average standard of living. The Netherlands requires an estimated area 15 times greater than its national size to support its population's way of life. It has been estimated that citizens of high-income countries today each require approximately 4 to 9 ha of the Earth's surface to provide the materials for their lifestyle and to absorb their wastes—while India's population gets by on 1 ha per person. There is not enough Earth to allow more than 1 ha of 'ecological footprint' per average-person when the world population reaches 10 billion during the coming century—and yet that future world population will presumably wish to live like Californians, not Calcuttans.

Serious investment in the development and deployment of less environmentally disruptive technologies, and a much greater commitment to international equity, will be required if a smooth and timely transition to an ecologically sustainable world is to be achieved. Because rich countries remain the main source of new knowledge and new technologies, responsibility for finding paths to sustainability rests mainly with them. Minimizing the probabilities of long-term harm to health will be a major consideration. Indeed, this is now becoming the most important health-related aspect of the 'population debate'.

'Green accounting'

We cannot predict, with certainty, how the adverse health effects of ecological disruption will unfold. It therefore makes sense to concentrate, in the short term, on our society's direction of travel, on whether we are moving closer to, or further away from, sustainable paths of economic development. To this end we need to devise and implement new indicators of material progress.

Sustainability has recently been defined by the Environment Department of the World Bank as leaving to future generations 'as many opportunities as we ourselves have had, if not more'. Sustainability can be more readily expressed in operational terms using measures of economic 'stock' (i.e. capital, including natural capital and human resources) than using measures of 'flow' (income). In this context, conventional national income accounts are both biased (they treat living off natural capital as income) and insensitive (they provide little indication of legacies for the future). A broad measure of wealth would combine the estimated values of natural and human resources with those of produced assets (capital as traditionally considered). Human resources include the 'human capital' embodied in individuals (augmented by health and education levels) and 'social capital' embodied in institutions, customs, and knowledge.

Employing these categories, a pattern of economically sustainable development can be envisaged as one that conserves natural capital while rebuilding (with 'green' technology) the stock of produced assets, and augmenting human and social capital. This shift of emphasis from flows to stocks accords with recent analyses of health trends. For example, in low- and middle-income countries, indicators such as school attendance rates for girls and literacy among adult women are correlated more strongly with the level of child mortality than is income. Countries at similar levels of income may have several-fold differences in child mortality, with the 'better performers' typically showing higher levels of relevant aspects of social capital.

For rich countries, this approach recognizes that health depends less on the consumption opportunities provided by income than on personal and social capacities to protect and enhance health— reflecting, at the individual level, determinants such as schooling and, at the social level, determinants such as food cultures (for example the protection against vascular mortality in Mediterranean populations) and elements of the built environment such as sewers, water supplies, and safe roads. Given that life expectancy differences among high-income countries are, at most, very weakly related to income, it makes little sense to see increasing national income as an important path to sustainable improvements in health.

Conclusion

Over the last two centuries new knowledge of disease and its control, improvements in material conditions, and the enhancement of individual capabilities through education and of social capacities through development of new institutional forms have yielded previously unimaginable improvements in health and longevity. During the historical gap between the fall in the death rates and the fall in birth rates, populations increased rapidly—and are still doing so in many poorer countries. Meanwhile, in high-income countries, attention should be substantially redirected from achieving yet higher local levels of health to developing the conditions necessary to both generalize and sustain good health worldwide.

A first essential is birth control, for without it death control is unsustainable. A second task is to revise our expectations of 'progress' and to reconstruct the milestones by which we measure it, so that the interests of future generations are safeguarded. Physicians are well placed to foster greater public understanding of why large-scale environmental disruption is likely to jeopardize the health of our grandchildren.

Further reading

Intergovernmental Panel on Climate Change (1996). *Second assessment report. Climate change 1995*, Vol. I, II, and III. Cambridge University Press, New York.

Loh J *et al*. (1998). *Living planet report*. World Wildlife Fund International, Gland, Switzerland.

McMichael AJ (1993). *Planetary overload. Global environmental change and the health of the human species*. Cambridge University Press, Cambridge.

UN Department of Economic and Social Affairs (Population Division) (1999). *World population prospects. The 1998 revision*, Document ST/ESA/SER.A/177. United Nations, New York.

World Resources Institute, United Nations Environment Programme, United Nations Development Programme, World Bank (1998). *World resources 1998–99*. Oxford University Press, New York.

3.3 The pattern of care: hospital and community

Anthony Harrison

Introduction

Changes in the treatments available over the last 100 years have transformed the care of individuals. Conditions once untreatable, can now be successfully dealt with, while advances in public health have transformed the pattern of disease confronting the health care system, eliminating or virtually eliminating some diseases. These developments have had a strong influence on the pattern of health care delivery. For much of the twentieth century they led to an expansion of the role of hospitals, but towards the end of the century the scope for moving care to other settings was increasingly recognized.

In 1920, a committee chaired by Lord Dawson set out a blueprint for the way that health care services should be delivered in England. The pattern it proposed consisted of primary and secondary care centres, the first located as near as possible to the local population, and the second in any sizeable urban area. If that area was not large enough to support a teaching hospital, the secondary centre and the teaching hospital should be linked professionally. The primary care centres would provide for a wide range of services including general practice and dentistry and, despite their modest size, inpatient beds. Their location in a single facility would, it was hoped, encourage professional interchange and collaboration. This early attempt to set out a vision for the pattern of health care delivery across all services forming a defined geographical area was not implemented even in the country for which it was devised. However, two central issues were presented that continued to preoccupy policymakers ever since.

1. Exactly which clinical services should be provided locally and which regionally or even nationally?
2. How should the various elements be persuaded to work together?

These questions face any health system but the way they are approached varies according to the way that health care is organized and financed. To outside observers, the British National Health Service (**NHS**), funded almost entirely from a single source and under the control of a single central body, in contrast for example to the pluralist systems obtaining in the United States or Germany, has seemed ideally placed to resolve these issues in a systematic way. In practice, however, solutions have remained elusive in all health care systems.

Since Lord Dawson put forward his proposals, the expansion of clinical knowledge, the specialization that goes with it, and the introduction of expensive diagnostic hardware created pressures for larger concentrations of clinical skills within hospitals. That trend was further strengthened by the growth of clinical research in close association with the patient care and teaching functions of the large acute hospital. Hospital development and clinical progress appeared almost synonymous.

Towards the end of the twentieth century, however, that perception began to change. Although the acute hospital has retained a central role in the health care system, its relative importance has begun to decline. Table 1 sums up some of the key changes within England: similar trends are apparent and indeed are in some cases more advanced in other countries.

The growth in activity as measured by the number of patients treated reflects the massive increase in the scope of hospital work brought about by developments in medical technology such as imaging and endoscopy and the vast new range of surgical procedures such as joint replacement, from which a wide range of the population can benefit.

The large acute hospital has been the physical expression of modern medicine through most of the twentieth century, but it became obvious, towards the end of the century, that technological development was allowing effective care to be provided off the hospital site for conditions once treated within it. Consequently, the scale of some elements of hospital care could be reduced.

A combination of developments in clinical technology, particularly in anaesthesia and surgical techniques, and financial pressures imposed by those paying for care, whether government or private insurers attempting to control health care spending, has resulted in continuing losses of bed capacity even while the numbers of treated patients have gone up. Loss of beds was particularly noticeable in surgical specialties as, from the early 1990s onwards, the proportion of patients treated as day (ambulatory) cases rose rapidly. Medical specialties also lost beds, as lengths of stay fell. Thus the acute hospital sector has continued to expand throughout the twentieth century, but its physical capacity, as defined by its bed-stock, has begun to decline. Turnover has become increasingly rapid. This in turn has meant that its efficient operation is increasingly dependent on effective links with community-based services offering aftercare and other forms of support.

These changes in the acute hospital in themselves implied some degree of shift of care to community settings. Rapid discharge schemes and hospital at home schemes have been developed, often administered by community-based nurses. They complemented the hospital's drive to reduce length of stay and use its remaining bed-stock more intensively. Similarly, other parts of what was once the hospital's role, such as check-ups following an inpatient admission, have tended to become the responsibility of general practice.

These changes were part of a wider and more fundamental development—the dispersal of a wide range of hospital activities to other settings, inspired by developments in technology, changes in clinical practice, and the search for lower costs of provision. The rapid introduction of new drugs from the 1950s onwards allowed general practitioners to deal with a large number of conditions which at one time would have been part of the hospital's workload. In some cases, such as diabetes, technological developments—in this case simple self-monitoring devices—allowed some care to be effectively transferred to the patient's own home. Hospitals were needed only for emergencies.

Lord Dawson's proposal for primary care centres was not widely adopted in its country of origin, but the idea that general practice should be the universal front-line of the health service was, in England and many other countries, seen as the foundation of a properly organized health service. General practice offers ready access and continuity of care, but also some control via the gatekeeping function, to the high-cost services of the hospital. This restriction of patient choice has not been universally accepted. Germany for example continues to allow patients to consult specialists directly. The United States has not had the same tradition of general practice and has, until recently, offered freedom of choice. However, financial pressures have led to the creation of managed care organizations aimed at restricting access to specialist, hospital-based care.

The general practitioner or community-based physician has assumed a co-ordinating role. The range of professional services, closely connected to the practice, has grown and with it the concept of an integrated primary care team emerged, reflecting the range of services, such as nursing, physiotherapy, and counselling, that come within the ambit of general practice or work in close association with it.

While the pattern of acute care was changing, an even more dramatic change was taking place in long-stay hospitals housing the elderly, the mentally ill, and those with learning disabilities. Clinical developments, particularly new drugs for the control of severe mental illness, played their part. However, the main explanation for this switch from hospital to community was the perception, which became widespread in the 1980s and 1990s, that most people housed in these institutions did not need to be there. With appropriate support from housing and social services as well as clinical staff, they could perhaps live reasonably normal lives in the community.

However, problems remain. The switch of some care to the community has not been wholly successful. The transfer of care for the mentally ill away from large isolated institutions to smaller local units supported by a network of health care and other community-based services has led to the emergence of a different pattern. Many patients reach specialist care by self-referral or other routes, and for many, the main point of contact is a community nurse and/or social worker rather than a general practitioner. Their roles are central to the creation of community support networks which provide access to housing, and recreational and other facilities essential to everyday life. Such networks have proved hard to manage and, in practice, the necessary liaison between their various elements has not always been

possible, particularly in times of crisis.

The switch of care for long-stay groups into the community has led to closures of the large hospital institutions that once housed them. These trends have led to the closure of smaller acute hospitals—or a drastic reduction in the scope of their activity through, for example, the transfer of inpatient admission facilities to other sites. This has been bitterly resisted by local communities, but the combination of cost pressures and clinically based arguments in favour of concentration of services has often proved decisive.

Such rationalizations have been based on the presumption that larger hospitals are better, both economically and clinically. The evidence on both is limited. What exists is often of poor quality since confounding factors make it hard to disentangle the effects of size on the outcomes of care. Overall the available evidence does suggest that specialization of clinical roles with concentration of hospital work in fewer sites produces better results. How far this process should go remains unclear and in some areas of care it may already have gone too far.

Working together

Although most episodes of care are 'singletons' consisting of a visit to one health care professional such as a general practitioner or dentist in the community or the accident and emergency department of a hospital, many are complex, involving many different professionals even within a single hospital institution and extending for long periods. Modern medicine has virtually eradicated some diseases and has contributed, with the social and economic trends associated with the growth of personal incomes, to prolonging life. This has meant that a large part of clinical practice is taken up with chronic conditions and with treating frail elderly people who often need continuing support after, for example, an episode of acute hospital care.

An effective health care system requires that these various contributions are made by the most appropriate professionals and are co-ordinated to prevent the patient falling through gaps between different organizations or professionals. These requirements have proved difficult to fulfil. Even within the British National Health Service funded from a single central source and under unified management, there are frequent failures of the various elements to mesh together properly. Failures often occur at the interface between different organizations. A persistent area of difficulty has been hospital discharge, particularly for those requiring postoperative support. Delays have given rise to the notion of the 'blocked bed'. Blocks can sometimes be attributed to poor internal communications within the hospital. They also reflect failure to co-ordinate the hospital discharge process with provision of community health and social care services. This failure reflects differences in source of funding, which create a division—artificial from the viewpoint of the patient—of responsibility for closely related care functions.

Research has revealed important gaps in the care system, particularly in rehabilitation after stroke or heart attack. As a result, too many people may be prematurely admitted to community long-stay institutions. Again the difficulties can in part be attributed to the failure of the financial mechanism to match clinical and patient requirements.

The search for better service integration has stimulated a number of initiatives. In hospitals, these range from large-scale re-engineering of the hospital as a whole to more modest attempts to improve the care pathways for specific groups of patient such as those receiving new joints or other procedures where some degree of care after discharge is required. In some cases, such pathways extend across care providers. The British government has recently introduced the concept of a national service framework. This is intended to define both the elements required from all forms of provision—hospital and community—and the way they should work together to form a clinically integrated system of care. In principle, this should ensure a correct balance between preventive and curative means and should enable patients needing care to be correctly and promptly routed to an appropriately equipped and trained clinical team. In the United States and elsewhere, disease management programmes have been developed with similar objectives.

Cost pressures have meant that the orchestration of the various contributions of complex episodes of care has been combined with a search for the best combination of professional contributions, where necessary across the boundaries of existing disciplines. It is increasingly accepted that different professional roles are to some extent interchangeable. Many health care systems are experimenting with mixtures of different skills, allowing nurses and in some cases technicians to take over roles which were once exclusively medical.

The emphasis on clinical integration has led to the creation of new professional roles such as liaison nurses. They are intended to co-ordinate contributions by different organizations, for instance that hospital and community professionals co-operate at the point of discharge and that the various elements required for post-discharge support are put in place.

The search for the effective integration of care across the boundary of the hospital and the community continues. In the United Kingdom, the Labour Government's first white paper on health policy, *The new NHS*, suggested ways in which health and social care providers might work together harmoniously.

Conclusions

The pattern of health care services will continue to change, as a result of continuing technical developments within medical and information technology and as a result of social and financial pressures for new forms of service delivery. The balance between hospital and community and between hospitals of different sizes and structure will vary according to the relative advantages in access, cost, and clinical effectiveness of treatment in different locations.

For clinicians, the most important development is the move towards the definition of care pathways and whole systems of care for broad disease groups running across the boundaries of individual hospitals and those of community- and hospital-based organizations. While the bulk of medical training and practice will remain focused on the individual intervention and the patient encounter, the vision of health care delivery is of a series of integrated pathways or networks. Clinicians must start to take responsibility for system management as well as patient care.

Critical future developments may lie as much with the user as the professional. Many health systems have been introducing a new front line—the telephone. In England, after a series of experiments in different parts of the country, the first steps were taken in 1998 towards the introduction of a national network of nurse-provided advice lines. Information technology is making it easier to gain access to knowledge that was once the preserve of the professional. Many large health care institutions in the United States now have websites which are accessible worldwide. The NHS has recently opened a site of its own: NHS Online. Patients are becoming better informed. If these developments succeed in changing the balance between professional and user, new patterns of care will emerge that suit the user rather than the professional. In some areas, such as maternity care and mental health, this is already happening.

Further reading

Ferguson B, Sheldon T, Posnett J, eds (1997). *Concentration and choice in healthcare*. FT Healthcare, London.

Grumbach K, Bodenheimer T (1995). The organization of health care. *Journal of the American Medical Association* **273**, 160–67.

Harrison A (2001). *Making the right connections*. King's Fund Publishing, London.

Johnson S, ed. (1997). *Pathways of care*. Blackwell Science, Oxford.

Leutz WN (1999). Five laws for integrating medical and social services: lessons from the United States and the United Kingdom. *The Milbank Quarterly* **77**, 77–110.

Robinson JC (1994). The changing boundaries of the American hospital. *The Milbank Quarterly* **72**, 259–75.

Shortell SM (1995). Reinventing the American hospital. *The Milbank Quarterly* **73**, 131–59.

Shortell SM *et al*. (1996). *Remaking health care in America: building organized delivery systems*. Jossey-Bass Inc., San Francisco.

Starfield B (1994). Is primary care essential? *The Lancet* **344**, 1129–33.

Starfield B (1998). *Primary care: balancing health needs, services, and technology*. Oxford University Press, New York.

Stevens R (1989). *In sickness and in wealth: American hospitals in the twentieth century*. Basic Books Inc., USA.

Stoeckle JD (1995). The citadel cannot hold: technologies go outside the hospital, patients and doctors too. *The Milbank Quarterly* **73**, 3–17.

Wilson J (1997). *Integrated care management: the path to success*. Butterworth-Heinemann, Oxford.

3.4 Preventive medicine

D. Mant

In his millennium address, Nelson Mandela reminded the world that 'we close the century with most people still languishing in poverty, subjected to hunger, preventable disease, illiteracy and insufficient shelter'. The health gap between rich and poor nations is shameful. For example, life expectancy in Malawi is 34 years compared with 79 years in Sweden. But even in the economically developed world many people still die prematurely. In England and Wales almost 1.4 million years of working life are lost each year due to death before age 65 years. Again, there is a marked gap between rich and poor—the death rate from lung cancer in males aged 20 to 64 years is almost three times as great in social classes IV/V (SMR 151) as in I/II (SMR 58). It is naïve to think that medicine will remedy this situation. The fundamental step in achieving good health remains the elimination of poverty, with consequent access to food, sanitation, education, and shelter. But the power of medicine lies in the scientific understanding it provides of the disease process. Preventive medicine uses this understanding both to try to reduce the risk of disease and to detect and treat appropriately emergent disease before it does damage.

What is the scope for prevention? Figure 1 shows the number of women expected to die at different ages if 10 000 were subject to the age-specific death rates in England and Wales of today compared with the1870s (the pattern is similar for men). The dramatic fall in deaths during childhood and early adulthood has meant that the modal age of death is now over 80 years. However, what medicine cannot offer is immortality. The proportion of women surviving to age 65 who live on to age 100 is still very low, about 0.5 per cent. So there seems to be a reasonable expectation that effective preventive medicine might make death before age 70 or 80 years uncommon—but the objective is delay of death, and hopefully better quality of life before death, rather than absolute prevention of death.

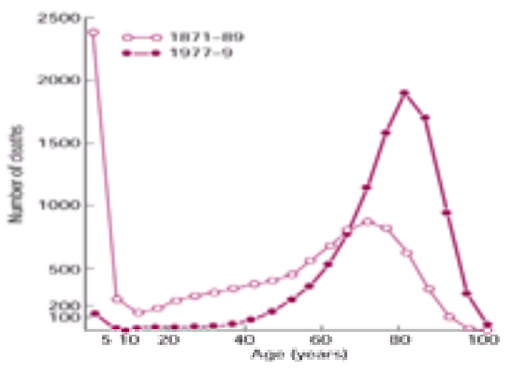

Fig. 1 Numbers of women dying at different ages if 10 000 were subject from birth to the mortality rate current in 1871 to 1880 compared with that in 1977–9. (Figure originally drawn by Doll R. *British Medical Journal*, 1982; **286**: 445–53.)

Preventive strategies

The prevention paradox

The main difference between preventive and curative medicine is the focus on risk. Preventive medicine aims to reduce the risk of disease and the risk of further morbidity and mortality in those who develop disease. It offers hope for the future rather than immediate benefit. The benefit from preventive medicine is the absence of future disease. This is a difficult benefit to champion, particularly to the individual. As Geoffrey Rose pointed out many years ago, not only is the benefit intangible but many people must take precautions in order to prevent illness in only a few. Even in a country where diphtheria is common, several hundred children must be immunized to prevent one death. Rose called this the prevention paradox—a preventive measure which brings large benefits to the community may offer little to each participating individual.

The risk paradox

Epidemiological studies define risk factors for disease—the personal or environmental characteristics which increase the likelihood of developing the disease. One of the risk factors for cardiovascular disease is a high level of cholesterol in the blood. Figure 2 shows the prevalence of high cholesterol in the United Kingdom population, the death rate associated with each cholesterol level, and the proportion of all deaths attributable to cholesterol occurring at each level. The risk paradox is that although those with a blood cholesterol of 7.5 mmol/l or greater are at highest individual risk of disease, the population at highest risk is that associated with a cholesterol level of 5.5 to 6.0 mmol/l. This is simply because of the number of people at risk—there are far fewer people in the high than in the moderate risk group. Targeting just the high-risk group will have relatively little impact on the total number of deaths.

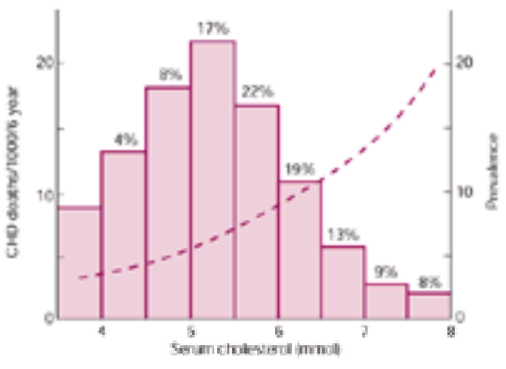

Fig. 2 Proportion of coronary heart disease deaths attributable to raised serum cholesterol occurring at each level (figures above columns). Columns show the

prevalence of different levels of cholesterol in the population. The broken line shows coronary heart disease mortality at each level. (Reproduced from Rose G. *Strategy for prevention.* Oxford University Press, 1992: 23.)

Defining the at-risk population

As preventive medicine seeks to reduce future disease, its patients are usually identified by characteristics which predict risk of developing the disease. It is usual to define this risk in one of three ways—demographic, phenotypic, or familial. Within each category, further subpopulations may be identified as at particularly high risk.

Demographic risk

This is the most common way to define the target group for preventive medicine. Both the United States and Canadian task forces on prevention (see further reading) classify preventive interventions by the age of the target population. Screening programmes, which often target gender-specific diseases (e.g. breast and cervical cancer), may also specify the target population by sex. Geographical specification of risk tends to reflect health-care system boundaries but some preventive programmes target specific ethnic or socially disadvantaged groups.

Phenotypic risk

A phenotype is a set of observable characteristics of an individual or group. Most epidemiological risk factors for disease (e.g. smoking, obesity, hyperlipidaemia) are phenotypic. Other phenotypic categories used to define at-risk populations are behaviours (e.g. smoking, driving), disease states (e.g. diabetes, angina), and physical characteristics (e.g. obesity, cholesterol level). As phenotypic risks are sometimes interactive (i.e. more than one risk factor influences a specific disease risk), multiple risk assessment is an increasingly common practice.

Familial risk

Recent advances in genetic technology have increased our ability to characterize familial risk accurately, and further advance is likely in the next few years. At present, most preventive medicine programmes in this area use genetic assessment to refine assessment of individual risk in phenotypically identified high-risk families (e.g. cystic fibrosis, neurofibromatosis) or populations (e.g. Down's syndrome). However, the characterization of risk based on population-based genetic screening is already technically feasible in the economically developed world.

Identifying the at-risk population

Some public health interventions can be applied without identifying the at-risk population—you can pass seat belt legislation and increase the tax on tobacco without identifying either drivers or smokers. However, most interventions in clinical preventive medicine are delivered to individuals. It is therefore necessary not only to define the at-risk population but also to identify the individuals within it. You don't just need to know that smokers are at risk, you need to know who smokes. Again, this is usually done in one of three ways—registration, screening, and case-finding.

Registration

Most socialized health systems keep registers. These may be simply demographic (e.g. age, sex, and address) or contain phenotypic or genetic details of individuals. Effective preventive medicine is much easier where registration exists and its accuracy is systematically maintained.

Screening

This topic is covered in Chapter 3.6. The objective is to identify early disease or high risk of disease (e.g. neoplastic dysplasia) before significant morbidity occurs. Its most important feature is that it can do harm as well as good (it may generate 'false alarms' and detect disease which would not otherwise present during the patient's lifetime) and benefit needs to be carefully assessed, normally in a randomized trial. Population screening is most efficient when based on an accurate population register.

Case-finding

This involves identifying at-risk individuals during routine clinical work (normally in clinical consultations, but sometimes through contact or family tracing). It is less efficient than systematic population screening, but sometimes provides better access to socially disadvantaged groups who may respond poorly to screening invitations or have no registered address. It also allows some interventions to be given at a particularly appropriate moment (e.g. smoking cessation advice at a consultation for cough or contraceptive advice after termination of pregnancy).

Interventions to modify risk

The marked improvements in health which have been achieved in economically developed countries over the past 200 years are not attributable to medicine. Life expectancy has doubled mainly because of environmental control of infectious pathogens (through sanitation and control of insect vectors) and a lifestyle which reduces individual susceptibility to infectious disease (better food, shelter, and education). So although medical science can play an important role in guiding public health policy by improving understanding of the mechanisms of disease, and specific medical interventions allow us to treat disease when it occurs, the role of preventive medicine should not be overestimated. In particular, the medical profession should not take upon itself responsibilities for public health which are more appropriately assumed by governments and other social and environmental agencies.

However, preventive medicine is an important and integral part of good curative medicine. All doctors have a responsibility to think about why someone is ill. Whatever cause is identified (physiological, social, or psychological) the question of whether the cause can be prevented (and the risk of future disease reduced) should be addressed. Doctors who work in a primary care role (particularly those with a registered population) have the added responsibility to ask themselves whether the risk should be addressed at a population rather than just an individual patient level.

Primary and secondary prevention

Preventive interventions have traditionally been categorized as primary, secondary, and tertiary depending on their objectives. However, the term secondary prevention is now commonly used to cover both secondary and tertiary categories. According to this common usage, primary and secondary prevention can be defined as follows.

1. Primary prevention—interventions to reduce the risk of disease in healthy people (e.g. use of seat belts to prevent injury in car accidents; tobacco control to prevent the occurrence of smoking-related disease; immunization against infectious disease).
2. Secondary prevention—interventions to prevent avoidable morbidity in people with disease (e.g. treatment of vascular disease with aspirin; screening for early cancer).

It is immediately obvious that the distinction between primary and secondary is sometimes difficult. Some interventions can fall into more than one category (e.g. stopping smoking reduces the progression as well as onset of many smoking-related diseases) and the definition of disease is not absolute (e.g. many apparently healthy people will have undetected disease). Nevertheless, the pragmatic categorization of preventive interventions into primary and secondary is often useful in practice.

Immunization

Immunity can be induced against many pathogenic bacteria and viruses. Active immunity is usually achieved by stimulating antibody production by vaccination with an inactivated organism (pertussis, hepatitis A/B), an attenuated live organism (measles, rubella), an antigenic component of the organism (influenza, pneumococcus) or an inactivated toxin (tetanus, diphtheria). Long-lasting immunity can be induced with a single dose of live vaccine. At least two and sometimes three doses are needed for other vaccines, although once an IgG antibody response has been induced, antibody levels are likely to remain high for months or years and can usually be reinforced by a single booster dose. Passive immunity is achieved by injection of human immunoglobulin and therefore protection lasts only a few weeks.

Vaccination can be a very effective preventive strategy. Vaccination against smallpox has led to global eradication of the disease; eradication of polio seems a feasible global objective in the next decade. Vaccination against many other diseases, particularly diseases of childhood such as measles, diphtheria, and polio, has led to rapid and dramatic falls in disease incidence. Figure 3 shows the impact of introduction of the Hib vaccine in 1992 on the incidence of *Haemophilus influenzae* infection in England and Wales.

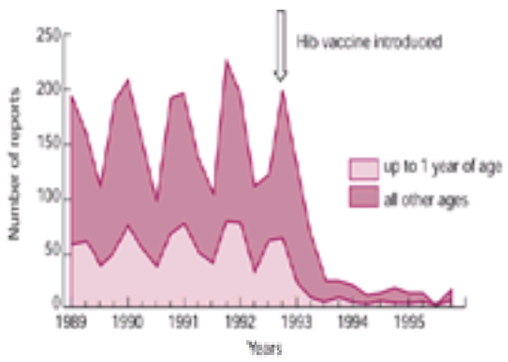

Fig. 3 Laboratory reports of *Haemophilus influenzae* type b before and after introduction of the Hib vaccine (England and Wales 1989 to 1995). (Reproduced by permission of the Controller, Her Majesty's Stationery Office.)

A number of new and important vaccines are on the horizon—for example, a malaria vaccine. But the existence of an effective vaccine does not guarantee the success of an immunization programme. This depends on the effective delivery of the vaccine to the at-risk population. Vaccination programmes are often limited in their effect by affordability (many vaccines are too expensive for developing countries), acceptability (parental anxiety about the adverse effects of pertussis vaccine has limited its uptake in many countries), and deliverability (vaccines may lose potency if stored outside a refrigerator). There are also potential problems with the antigenic variability of organisms (e.g. influenza) and the difficulty of immunizing at an age young enough to prevent morbidity but old enough to stimulate an immune response (e.g. measles). Nevertheless, immunization is probably the most important medical contribution to primary disease prevention.

Prophylactic treatment

Although most people think of medicines as cures for current illness, many medicines are prescribed with a view to preventing future illness. Antibiotics are given before surgery to prevent postoperative infection, antimalarials to prevent malaria in travellers, anticoagulants to prevent stroke in people with atrial fibrillation, and lipid-lowering agents to prevent heart attacks in people at high risk of cardiovascular disease. The duration of treatment may also be extended beyond the initial treatment phase to achieve a preventive effect. Antidepressants are continued after cure to prevent relapse, ACE inhibitors to prevent worsening of ventricular dysfunction, and uricosuric agents to prevent further episodes of gout.

It must be clear from these examples that many, perhaps most, drugs have the potential to be used for prevention as well as cure. In some cases (e.g. treatment of diabetes) the distinction between prevention and cure is unhelpful—treatment aims to prevent morbidity in both the short and long term. However, in all the examples given, prescribing is limited to a defined high-risk group. Prophylactic treatment with drugs is less helpful when a high-risk population cannot be easily defined. It is almost always inappropriate to use prophylactic treatment to reduce population risk for three reasons: the strategy is seldom cost-effective, increasing the reliance of the population on medicine is an adverse social outcome, and uncommon adverse effects can easily outweigh any clinical benefit. The last point has been repeatedly demonstrated in clinical trials, including the use of lipid-lowering agents to prevent heart disease and antioxidants to prevent the development of cancer.

Changing behaviour

Environmental factors contribute substantially to differences in premature morbidity and mortality. Many of these environmental factors reflect individual behaviour. Eating more healthily, taking more exercise, and avoiding riding on a motorcycle are all effective ways of preventing disease. People do listen to doctors, and a number of clinical trials have shown advice on behaviour modification to be cost-effective, even though the effect size may be small (e.g. in most studies only about 1 in 20 to 30 smokers given brief advice to stop smoking actually quit). Brief advice is most effective if practical in nature (giving guidance on how change can be achieved) and if backed up by written advice to take home. More intensive interventions tend to be less cost-effective. Time spent on alternative preventive activities (e.g. effective management of chronic disease) may reap greater rewards.

Environmental change

Many environmental causes of disease are best modified on a public health rather than an individual basis. Such factors include the safety of the workplace, environmental pollution, transport safety, food hygiene, and provision of clean water. However, a number of diseases have environmental causes which need to be recognized and avoided by the individual patient. On a global scale, avoidance of insect and other disease vectors (e.g. by netting) and attention to nutritional hygiene (e.g. by filtering water) are probably the most important. In economically developed countries the most common diseases amenable to individual environmental intervention are those associated with atopy—such as asthma and eczema. Not all patients have an identifiable allergenic cause for their symptoms and, even if one is identified, avoidance (e.g. of house dust mite in asthma) may not be easy. But dramatic improvement can occur, and treating contact dermatitis without giving advice on contact avoidance, or treating louse bites without giving advice on how to rid clothes of lice, is bad medical practice.

What interventions work?

It is impossible in a single chapter to cover every possible preventive intervention. It is also undesirable, because many preventive interventions are better seen as part of good routine clinical care (and are included in the relevant chapters on specific diseases). Nevertheless, it is worth listing the preventive interventions which may not be included elsewhere, and for which there is very good evidence of effectiveness from clinical trials. The best sources of evidence are the task force reports from Canada and the United States (see further reading). The evidence cited below is based on the last (1997) update of the Canadian report on clinical preventive health care. It focuses on issues of importance in economically developed countries in which most of the research has been done.

Table 1 lists the preventive interventions shown to be effective for mothers and babies. The target for all but one intervention (screening for haemoglobinopathies) is the whole population. Many of the interventions are usually delivered by midwifery or nursing staff (health visitors in the United Kingdom), but medical staff may be involved in child development and antenatal examinations and it is very important that they reinforce the advice and guidance given by other staff.

Table 2 deals with children and adolescents. Immunization, dental care, and protection of hearing are interventions which should be offered to all children. All other interventions are targeted at specific high-risk groups, particularly children living in conditions of social disadvantage or with families identified as being at 'high risk' of providing unacceptable levels of child care. Chemoprophylaxis is effective for child contacts of open tuberculosis and for immunocompromised children exposed to influenza. Screening for sexually transmitted disease has been shown to be effective in at-risk adolescents and children.

Table 3 deals with interventions for adults, including the elderly. Only two (hearing impairment and dental care) are targeted at the whole population, although post-fall assessment is targeted at all elderly patients and mammography at all women in the relevant age group. The effectiveness of mammography has again been questioned recently, but the balance of evidence remains in its favour. All other interventions are aimed at high-risk groups. It must be stressed that the omission of

many secondary preventive interventions (e.g. management of high blood pressure; rehabilitation after myocardial infarction) reflect their inclusion in other chapters rather than lack of evidence of effectiveness.

What is achievable?

Programme effectiveness

The interventions cited above are known to work because they have been tested in clinical trials. However, clinical trials are often done in settings far removed from everyday life—participants are compliant, those delivering the intervention are highly trained, the technology is of high specification, and quality control is rigorous. These conditions will not hold on a wet Tuesday morning in the boondocks. When preventive interventions fail, the most common reason is lack of effective implementation of the implementation programme, rather than lack of effectiveness of the intervention itself.

The importance of considering programme effectiveness is seen most vividly in relation to immunization and screening programmes. The three most important issues which determine programme effectiveness are the following.

1. Coverage—What proportion of the population at risk receives the intervention?
2. Delivery—Are factors which effect the delivery of the intervention (like the maintenance of equipment, the training of staff, and the storage of biological materials) up to scratch?
3. Quality control—Are standards set and monitored for key indicators of the intervention process (e.g. immune response or predictive value of screening)?

The effect on programme effectiveness of failure in just one of these areas is well documented in the United Kingdom in relation both to immunization (e.g. the resurgence of pertussis after media publicity about potential adverse effects of the vaccine led to a fall in uptake) and to cervical screening (e.g. lack of quality control in cervical sampling and cytological assessment led to false-negative results and avoidable mortality).

Cultural constraints

Many preventive interventions aim to change behaviour. Most behaviour has a strong sociocultural component and reflects prevalent attitudes and norms in society. Preventive interventions are severely constrained by this social context—convincing individuals to stop smoking, eat less salt, drink less beer, or drive more slowly is difficult if everyone else is doing the opposite. For example, the mean blood cholesterol level in Finland is almost twice that in Japan (Fig. 4). Migrant studies suggest that this difference is dietary rather than genetic in origin and so medical advice to reduce fat consumption should have the potential to reduce substantially blood cholesterol level. However, even in the context of a clinical trial, dietary advice from health professionals in a community setting seldom achieves a reduction in blood cholesterol of more than 3 to 5 per cent. Studies of salt restriction (to lower blood pressure) show a similar result—intensive intervention and support is needed for an individual patient to achieve a physiologically significant reduction in intake, with many finding such a diet unpalatable. Countercultural change is difficult to achieve.

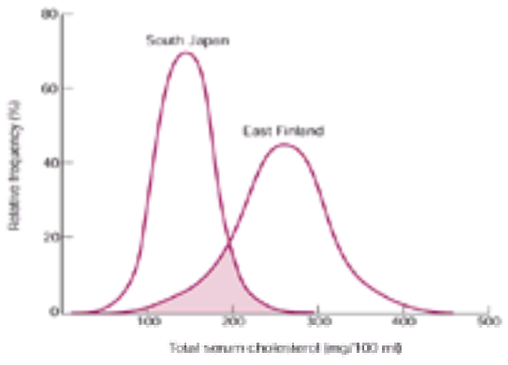

Fig. 4 Distribution of serum cholesterol in southern Japan and eastern Finland. (Reproduced from Rose C. *Strategy of prevention*. Oxford University Press, 1992: 57.)

Time effects

Things change over time. The North Karelia project was a large-scale long-term programme to reduce mortality from cardiovascular disease in northern Finland started in 1972 which involved both public health and individual intervention. Figure 5 compares mortality from cardiovascular disease in North Karelia with that in 10 other provinces in Finland before and during the intervention by plotting two regression lines. The difference in slope of these two lines shows that the intervention was to some extent effective. However, far more impressive in magnitude is the absolute fall in mortality over time in both North Karelia and the other provinces. The lessons for preventive medicine are twofold: the effect of medical intervention may be small in relation to the effect of other economic and social influences; and the change in baseline risk and social context over time may be so rapid that it will substantially influence the absolute benefit of any preventive intervention.

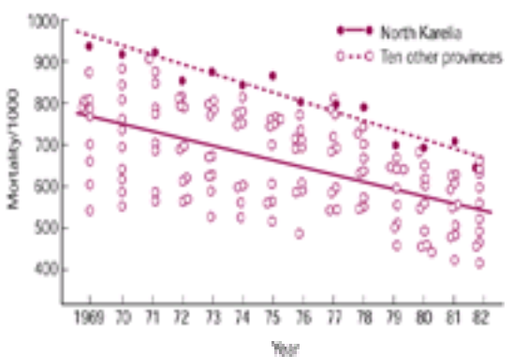

Fig. 5 The North Karelia project. Age-standardized annual mortality from cardiovascular disease in men aged 35 to 64 in Finland, 1969 to 1982. (Redrawn from original data published by Tuomilehto J *et al*. *British Medical Journal,*1986: **293**: 1068–71.)

Conclusion

Preventive medicine is an integral part of clinical practice for all doctors. It is our responsibility as clinicians not only to cure the presenting illness but also to take action where possible to prevent future morbidity. However, we must display both humility and assertiveness in our approach. We need to be humble in our approach to patients and to recognize that medicine is not the main determinant of health. At the same time we must display assertiveness in our advocacy of prevention. In the United Kingdom, the Royal College of Physicians' reports, the campaigning of medical charities, and the decision of virtually all doctors to stop smoking have played a major role in influencing both public and political opinion against tobacco use. As a profession, we can make a unique and powerful contribution to prevention by identifying the existence and causes of ill health. We also have a unique and powerful responsibility to act as advocates for our patients in seeking to ensure that these causes are addressed and the risk to their health is minimized. Good clinical practice entails preventive medicine, but good preventive medicine is more than just good clinical practice.

Further reading

Canadian Government (1994, reprinted 1998). *The Canadian guide to preventive health care.* Canadian Government Publishing, Ottawa. [Encyclopaedic (more than 1000 page) summary of current evidence on the effectiveness of preventive health care—updated electronic version at www.fedpubs.com/subject/health/clinpre.htm.]

Goldbloom R, Lawrence R (1990). *Preventing disease—beyond the rhetoric.* Springer Verlag, New York. [Published 10 years ago but a landmark text (based on early work done for the Canadian and US Task Forces on prevention) which made the case for evidence-based prevention.]

Rose G (1992). *The strategy of preventive medicine*. Oxford University Press. [The definitive text on the theory of preventive medicine—short, readable, brilliant.]

UK Joint Committee on Vaccination and Immunisation (1999). *Immunisation against infectious disease. Report of UK Joint Committee on Vaccination and Immunisation*. HMSO, London. [Annually updated short publication which provides a practical, but evidence based, summary of immunization recommendations in the United Kingdom.]

US Preventive Services Task Force (1995). *Guide to clinical preventive services—Report of the US Preventive Services Task Force*, 2nd edn. US Department of Health and Human Services, Washington, DC. [Similarly encyclopaedic (more than 900 page) summary of current evidence on the effectiveness of preventive health care—updated electronic version at www.odphp.osophs.dhhs.gov/pubs/guidecps.]

World Health Organization (1999). *Removing obstacles to healthy development*. World Health Organization, Geneva. [Report focusing on the potential for global prevention of infectious disease—electronic version at www.who.org/infectious-disease-report.]

World Health Organization (2000). *World Health Report 1999. Making a difference*. World Health Organization, Geneva. [Short (121 page) report focusing on what are seen as the two main preventable threats to global health at the millennium—tobacco and malaria.]

3.5 Health promotion

Keith Tones and Jackie Green

Health promotion

Health: meaning and aetiology

Health: the positive dimension

Health promotion is a controversial concept—it means different things to different people. This is not surprising since the notion of health itself is open to many interpretations. The definition of health embodied in the Constitution of the World Health Organization (1946) confirmed that health is not just the absence of disease and infirmity. It is also concerned with positive well being and has mental and social as well as physical dimensions. Although there is a temptation for medical practitioners to dismiss such preoccupations with well being as vague philosophizing, medicine cannot, and should not, discard these broader concerns. For instance, the notion of tertiary prevention has traditionally acknowledged that medicine should aim to maximize the quality of life of those it cannot cure. There is also increasing evidence that life expectancy, in addition to well being, can be influenced by a number of 'positive' individual and social attributes. For instance, the notion of 'social capital' is currently very popular with health policy-makers—it is argued that a high level of social capital supports health through giving access to a range of social networks which provide support and foster a sense of social connectedness. People will be healthier if they feel they are in control of their lives and that their lives are meaningful and make sense emotionally—in other words, they have a 'sense of coherence'. The central concern of health promotion should be not only to add years to life but life to years!

Determinants of health

Whether health is defined as the absence of disease or as a broader more holistic state, it is widely agreed that four major factors determine the extent to which people are or are not healthy. These are shown in Fig. 1.

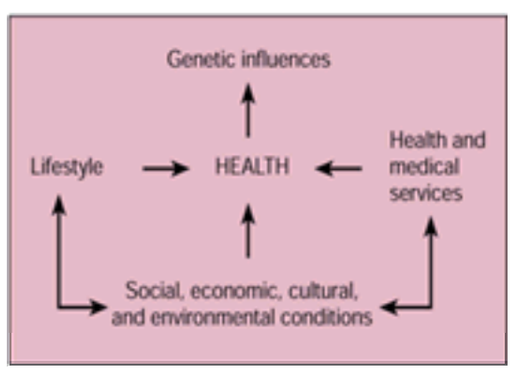

Fig. 1 Determinants of health

Traditionally, the individual's lifestyle has been the main target of health education—together with exhortations to make proper use of medical services. The current view is that the environment in which we live, work, and play has the most powerful influence, either directly or through the mediation of individuals' behaviour. Accordingly, the most effective health promotion strategies are those that modify the environment to make it safer and more health-enhancing and, most important of all, those which tackle the social and economic factors that cause health inequalities. Environmental circumstances may make it more or less easy for individuals to adopt healthy behaviours. For this reason a central concern for health promotion is to go beyond merely providing education about healthy choices. It should provide a supportive environment such that the 'healthy choice becomes the easy choice'. Of course, changing the environment often requires substantial changes in public policy and associated political action. Health promotion is therefore considered to include both health education and 'healthy public policy'.

Human behaviour in health and illness

Although some medical practitioners have been involved in a wide range of different health promotion activities—including community development and creative arts projects—the face-to-face encounter between doctor and patient is the most common focus of interaction. Every such encounter offers an opportunity for one-to-one health promotion. However, quite often this face-to-face interaction fails to capitalize on the potential offered.

Lessons from failure

Promoting healthy lifestyles is not an easy task and requires a thorough understanding of human behaviour at the social and individual level. The interaction between health professionals and their clients has been subjected to extensive research. This research confirms that, on average, as many as 50 per cent of people fail to co-operate with health advice of any kind, and changes in lifestyle are particularly challenging. A variety of reasons accounts for this failure, for example:

- recipients of, or participants in, health promotion often misinterpret the information provided and a smaller proportion forget key points;
- many people do not believe the information provided because it conflicts with their own ideas and beliefs;
- people may want to change lifestyle but do not believe they can change—because of real or perceived obstacles to action;
- environmental circumstances may be such that there is an absolute barrier to change.

We can and must draw lessons from examples of failure. Typically they are due to inadequate and naïve attempts at health promotion and lack a sound theoretical foundation. There is, however, considerable evidence that allows us to state categorically that health promotion can be effective—provided that appropriate and sufficiently sophisticated interventions are employed. Such interventions should be derived from analysis of the determinants of health and health behaviour, and must be based on a thorough understanding of the psychological, social, and environmental factors that influence people's behaviour in health and illness.

The psychosocial and environmental determinants of health actions

Figure 2 summarizes the important factors that determine whether or not individuals will adopt a new and healthier course of action.

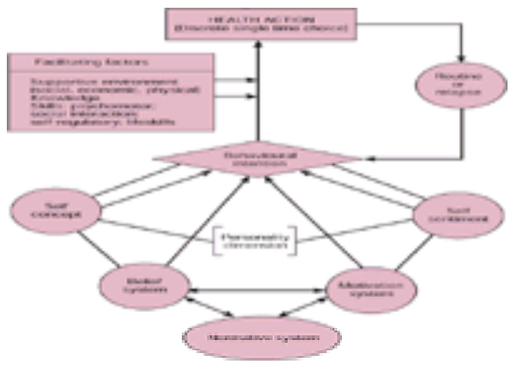

Fig. 2 Determinants of health actions.

Three major systems influence individuals' intention to act, that is the strength of their determination to adopt new types of behaviour or change old habits. These comprise the sum total of their beliefs about carrying out the action, together with their level of motivation to take action. This, in turn, will be the product of all their values and attitudes relating to the action and other emotional states. Finally, their intentions will also be determined by the cumulative influence of social pressures. These pressures will range from the relatively unimportant effect of general norms, conveyed by mass media, through the more powerful community norms, to the much greater influence of close friends, peers, and family.

Even quite firm intentions to adopt a healthy course of action will only be translated into practice if the environment and social circumstances are supportive and there is adequate provision to overcome the various barriers that often stifle action. In other words, it is essential to identify those circumstances that 'make the healthy choice the easy choice' and help avoid relapse into earlier unhealthy habits. The psychosocial and environmental prerequisites for adopting and routinely using condoms as part of a package of safer sex practices provide an illustration of what is required.

Beliefs and motivation

Certain key beliefs will contribute to the intention to use condoms with a sexual partner. First of all, individuals must believe they are susceptible to infection. Furthermore, they must accept that using a condom really will prevent HIV infection (and other sexually transmitted diseases). This belief will, in turn, depend on their having sufficient understanding about the nature of the virus and its mode of transmission—and by accepting that a condom's thin membrane is an effective barrier to the virus. More importantly, they must be persuaded that using a condom will neither be expensive nor inconvenient. Unfortunately, there is a common belief that condoms reduce sensitivity and gratification, interrupt lovemaking, and generate embarrassment!

A number of key motivational factors may militate against safer sex. Most people value health and are worried about the risk of disease—especially fatal disease. However, many women have been conditioned to believe that casual sex is immoral, a powerful disincentive to taking anticipatory action. This may well create a situation in which a powerful moral value may rule out precautionary action, but be insufficiently powerful to resist sexual passion and their partner's preference.

Social pressure

It is widely recognized that intentions are influenced not only by personal beliefs and values but also by various pressures, real or imagined, from other people. These pressures will include the general influence of social norms in defining 'normal' sexual behaviour, peer expectations and pressures, and also the particularly powerful influence of the partner or potential partner.

Skills and the environment

Good intentions will rarely be implemented unless barriers to action have been minimized and appropriate support has been provided. Three kinds of supportive strategy may be necessary. The first and simplest of these involves access to the information needed to adopt the health action. In this specific example—where to obtain and how to use condoms.

The second prerequisite for making healthy choices is to acquire essential skills:

- assertiveness to negotiate condom use with a partner;
- interpersonal skills and associated confidence needed to obtain condoms;
- psychomotor skills needed for safe and efficient use of the condom.

Third, the provision of a supportive environment is essential. A range of environmental measures will be relevant, including physical measures such as access to condoms and social measures that will create a climate in which condom use is acceptable.

However, the most difficult environmental barriers of all are the socioeconomic circumstances associated with poverty and its accompanying feelings of hopelessness and helplessness. It is no coincidence that the prevalence of preventable disease worldwide—including AIDS—is highest in the lowest socioeconomic groups.

The principle of reciprocal determinism

The principle of reciprocal determinism—drawn from social psychology—has proved to have great explanatory value for understanding efficient health promotion interventions. It asserts that there is typically a continuing interplay between the environment and individual action. On the one hand, social, material, and economic circumstances can facilitate or inhibit individual choice and action; on the other hand, individuals are capable of taking action to change their environment thus effecting change in those circumstances. Health promotion, therefore, is concerned both to reduce environmental barriers to action and, at the same time, strengthen individuals' capacity to challenge and modify their environmental circumstances. For these reasons, empowerment occupies a central place in the philosophy and practice of contemporary health promotion.

Health promotion as 'empowerment'

The World Health Organization has defined the main purpose of health promotion as helping people to take control of their lives and their health by:

- providing individuals with those 'empowerment' skills that generate competence and confidence; and
- removing at least some of the environmental barriers to action.

We noted earlier that health promotion involves a close relationship between health education and 'healthy public policy'. Health education should not be concerned primarily with persuading and cajoling to gain compliance, but rather with reassuring and supporting people to gain their co-operation. A supportive environment will be dependent on healthy public policy, including fiscal and economic measures, environmental engineering, and legislation. The pursuit of healthy public policy itself requires education and lobbying of key figures and activists in the policy arena.

However, it is clear that although lobbying and advocacy are essential to successful health promotion, they will have little impact on entrenched power structures—especially if there are major financial implications—unless they are supported by public pressure. Accordingly, one of the main functions of health education is to create 'critical consciousness'—a process that involves:

- building a sense of community;
- raising people's awareness of, and indignation about, major social and health problems, and developing the skills they need to exert political pressure on

government or organizations.

In some instances, this dual process may be sufficient in itself to bring about change but frequently the public's efforts must be supplemented by advocacy on their behalf—and the powerful voice of the medical profession can be especially important in this context.

Furthermore, medical services can contribute in a more general way to both individual and community empowerment and provide a model of good practice for other agencies. The service offered should be 'user-friendly' and responsive to the felt needs of communities.

Settings for health promotion

The 'health career'

The 'health career' is one of the most useful devices for identifying the cumulative influences on health behaviour—and for devising appropriate intervention strategies. In short, it plots key influences that operate over a lifetime. Figure 3 describes its main features. It identifies key agencies, such as the family, and more general influences, such as the community or mass media, that shape beliefs and motivations about health, exert social pressure, and erect barriers to action or, alternatively, provide support for healthy choices.

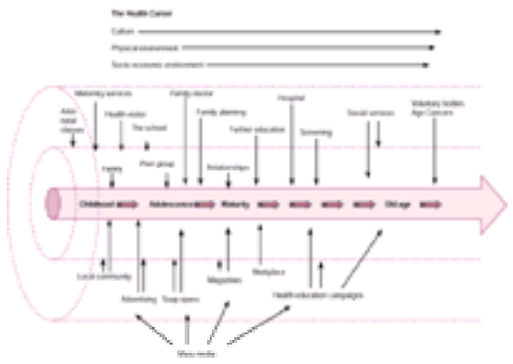

Fig. 3 The Health Career.

Settings, alliances, and mass media

The 'health career' analysis makes it possible not only to identify negative influences on health, but also, more positively, to identify agencies, organizations, and settings that can be used to 'deliver' health promotion. For instance, particular interest has been shown recently in the Health Promoting School, the Health Promoting Workplace, and the Health Promoting Hospital; three points merit emphasis:

- the whole ethos of an organization should be health promoting and empowering;
- the impact of any one organization will be maximized by intersectoral working and the creation of healthy alliances;
- mass media can support community health programmes but cannot act as a substitute for interpersonal working.

The doctor's role

It should by now be self-evident that health promotion is a multidisciplinary endeavour. What is the doctor's role? It is clear that the medical profession enjoys a high level of credibility with public and politicians. They are also uniquely placed to have awareness of the broad range of psychosocial and environmental factors influencing the health and well being of patients. Their major contribution may be summarized as follows:

- providing advice on lifestyle;
- delivering medical services in ways that are empowering and contribute to a sense of control;
- acting as advocates for social change—by lobbying and by using creative epidemiology (presenting health data in a dramatic, readily understandable, and compelling way) to create public concern and pressure for political action.

Health promotion methods: the consultation

Ultimately, the effectiveness of health promotion will be directly related to the quality of the specific methods used by health workers. It is clear that the achievement of certain important goals—such as remedying health inequalities—depends on the strength in unity that results from the kinds of alliance and political action mentioned above. However, individual practitioners and organizations such as hospitals can make a substantial impact—if the right techniques and strategies are employed. Box 1 provides guidelines for a health promoting consultation that helps clients take charge of their lives and their health. This approach should avoid the kinds of failure mentioned earlier in this chapter and maximize the chances of achieving real change.

Box 1 The health promoting consultation

Needs assessment

- Establish and maintain rapport using appropriate counselling skills.
- Check patients' state of readiness. Have they given any thought to changing their behaviour? Do they think they have a problem? Do they feel it is important for them to change? Are they committed to change but just need some support?

Communication phase

- Provide information at a level appropriate to the patient.
- Ensure that patients pay attention to the message.
- Check that patients have correctly interpreted the message—and fully understood key points. Take account of non-verbal communication.
- Try to ensure that patients remember important points—and provide a written aide-memoire.
- Ensure that any written material has been checked for 'readability'.

Motivation phase: helping patients to change

- Explore patients' existing beliefs and attitudes and those skills necessary for the proposed behaviour change.
- Analyse patients' environmental and social circumstances.
- Seek to modify beliefs and attitudes where appropriate and provide skills training.
- Negotiate and agree a 'contract'.
- Check that the relevant changes in beliefs, attitude, and skills have taken place. Are patients now committed to action?

Support phase

- Provide opportunities for patients to acquire any extra knowledge and skills necessary for translating good intentions into actual practice.
- Help mobilize social and environmental support.
- Consider particularly any skills or support needed to minimize the chance of relapse.
- Keep a check on patients' progress.

Where consultation time is limited, the task outlined above may appear daunting. However, the following points should be noted:

- The full procedure would be necessary in only a few cases.
- Many interactions with health professionals are continuing and cumulative. A series of meetings may be needed to cover the full 'recipe' for a successful outcome.
- Careful assessment of the stages of readiness to change will allow selection of only those strategies which are pertinent.
- It is frequently possible and desirable to delegate some or all of the task to other staff who may have more time.
- Different methods may be combined to achieve the various subtasks described above, e.g. a support group may use lay 'peer leaders' as health educators.
- Although the consultation offers only limited opportunities for addressing major socioeconomic problems, some action is possible. For instance, a number of health centres have offered 'welfare benefits clinics' in collaboration with a social worker to ensure that patients claim their full entitlement to financial help. Furthermore, by acting as advocates for patients, health workers as a professional group can and should seek to influence local and national policy.

Further reading

Ewles L, Simnett I (1999). *Promoting health: a practical guide*. Baillière Tindall/RCN, London. [A practitioner's guide.]

Jones L, Sidell M, eds. (1997). *The challenge of promoting health: exploration and action.* Macmillan/Open University, London. [Useful on features of health policy, intersectoral working, and the community.]

Tones BK, Tilford S (2001). *Health promotion: effectiveness, efficiency and equity*, 3rd edn. Thornes/Nelson, London. [Standard reference on contribution of health education to health promotion with a slant on evaluation and effectiveness.]

Tones BK (2001). Health promotion, health education and the public health. In: Detels R, McEwen J, Beaglehole R, Tanaka H, eds. *Oxford textbook of public health*, 4th edn, Vol. 2, Ch. 24. Oxford University Press, Oxford. [An overview of the contribution of health promotion to public health, incorporating discussion of philosophy, ethics, programme planning and evaluation.]

Websites

http://www.who.ch/ [World Health Organization website with search facility. Useful access to health documents, e.g. *Health For All*, *Ottawa Charter*, *Jakarta Declaration*.]

http://www.hda-online.org.uk/ [Health Development Agency. Access to research and information on good practice.]

http://www.healthpromis.had-online.org.uk/ [Access to 'Healthpromis' the health promotion database for England.]

3.6 Screening

J. A. Muir Gray

All screening programmes do harm; some also have the potential for doing good. In this respect screening is no different from the rest of clinical practice, but there are important differences between screening and clinical practice.

First, the contract, implicit or explicit, between the person being screened and the screener is qualitatively and ethically different from the contract that exists between clinician and patient. A patient who presents with a problem is seeking help and is usually aware of the facts about the limitations of medicine, and does not expect a guaranteed cure or risk-free treatment. This is part of the deal that they make with the clinician when entering the process of care. When, however, a professional or health service, or indeed the government through one of its agencies, writes to a healthy member of the population and asks them to come along for screening, the responsibility on the person issuing the invitation is much heavier. The traditional principle and the first priority to do no harm is even more strongly reinforced when dealing with people who perceive themselves as being healthy. In addition, the limitations and risks of screening must be explicitly and clearly spelt out.

The second way in which screening differs from clinical practice is that screening focuses on populations or on subgroups of the population, although it is often delivered by an individual clinician to an individual member of the public.

Screening is defined by the National Screening Committee in the United Kingdom as 'The systematic application of a test or inquiry, to identify individuals at sufficient risk of a specific disorder to warrant further investigation or direct preventive action, among persons who have not sought medical attention on account of symptoms of that disorder'.

This definition emphasizes that a subgroup of the whole population is selected on the basis of gender, age, or other characteristics, and then offered a test or asked a question. The results of the test may be either positive or negative and both positive and negative results occur in people with the disease or risk factor, and in those without. The relationship between positive and negative test results and those who are positive and negative for the disease is most easily set out in a figure (Fig. 1).

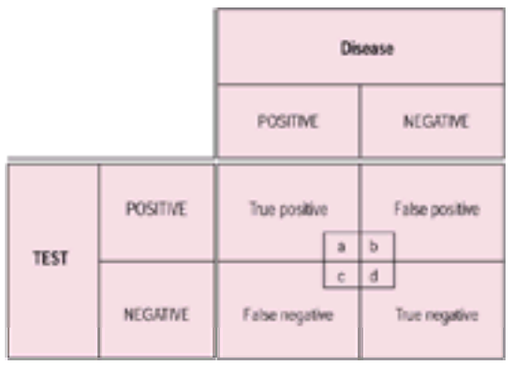

Fig. 1 The relationship between test results and the presence of disease.

On the basis of these results, it is possible to define certain characteristics of a screening test of which the most important are its sensitivity and specificity.

The sensitivity of a screening test is measured by the proportion of people who actually have the disease or the risk factors sought and who are detected by a positive test.

The specificity of a screening test is measured by the proportion of the people who do not have the disease who are classified as negative by the test result (Fig. 2)

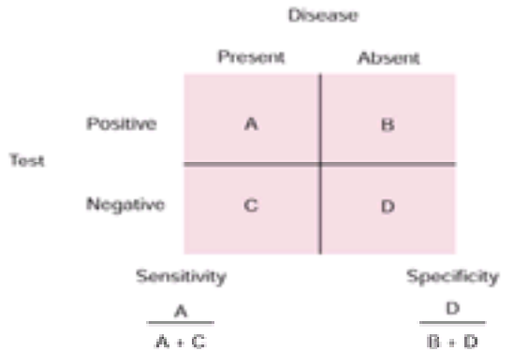

Fig. 2 The calculation of sensitivity and specificity.

These are the true traditional epidemiological parameters for a screening test, but one of the principles of screening is that screening consists not simply of tests but of a whole set of activities ranging from the identification of the population at risk right through to the diagnosis and treatment of affected individuals.

Appraising the balance of good and harm

The first step in deciding whether or not to introduce a screening programme is to appraise the balance of good and harm.

Screening is delivered as a programme and not as a single test. The sensitivity and specificity of the test, or tests, used in the screening programme are important but criteria have to be used to assess the programme as a whole.

The criteria proposed by the World Health Organization in 1968 have been widely used since then. However, the National Screening Committee in the United Kingdom considered that these criteria, sometimes called the Wilson–Jungner criteria, were not sufficiently robust for the twenty-first century because:

1. they did not emphasize the need to take into account the adverse effects of screening; and
2. they paid insufficient attention to the strength of the evidence on which decisions about screening were to be made.

Accordingly a new set of criteria, set out in Box 1 have been developed.

Box 1 Criteria for appraising the viability, effectiveness, and appropriateness of a screening programme

Ideally all the following criteria should be met before screening for a condition is initiated.

The condition

1. The condition should be an important health problem.
2. The epidemiology and natural history of the condition, including development from latent to declared disease, should be adequately understood and there should be a detectable risk factor, disease marker, latent period, or early symptomatic stage.
3. All the cost-effective primary prevention interventions should have been implemented as far as practicable.

The test

4. .There should be a simple, safe, precise, and validated screening test.
5. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.
6. The test should be acceptable to the population.
7. There should be an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals.

The treatment

8. There should be an effective treatment or intervention for patients identified through early detection, with evidence of early treatment leading to better outcomes than late treatment.
9. There should be agreed evidence-based policies covering which individuals should be offered treatment and the appropriate treatment to be offered.
10. Clinical management of the condition and patient outcomes should be optimized in all health care providers prior to participation in a screening programme.

The screening programme

1. There should be evidence from high-quality randomized controlled trials that the screening programme is effective in reducing mortality or morbidity.
2. There should be evidence that the complete screening programme (test, diagnostic procedures, and treatment/ intervention) is clinically, socially, and ethically acceptable to health professionals and the public.
3. The benefit from the screening programme should outweigh the physical and psychological harm (caused by the test, diagnostic procedures, and treatment).
4. The opportunity cost of the screening programme (including testing, diagnosis, and treatment) should be economically balanced in relation to expenditure on medical care as a whole.
5. There should be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards.
6. Adequate staffing and facilities for testing, diagnosis, treatment, and programme management should be available prior to the commencement of the screening programme.
7. All other options for managing the condition should have been considered (e.g. improving treatment, providing other services).

Measuring the benefits of screening

The first step in appraising a proposed screening programme is to try to measure or estimate the benefits that will result from that programme. But there are two traps that lie in the way of anyone wishing to assess the effectiveness of screening—lead-time bias and length time bias.

Lead-time bias

Proponents of the introduction of any screening programme sometimes base their argument on cohort studies, which are designed to follow a series of people who have had a screening test and compare their survival with that of the general population. However, this is a poor method of evaluating screening, principally because of what is called lead-time bias.

Imagine a disease that has a natural history of 10 years from its beginning to its fatal end, and that causes symptoms after 5 years, which usually prompt the sufferer to visit a doctor; the survival time from the point of symptomatic diagnosis is 5 years (Fig. 3(a)). A test that enables a diagnosis of the disease to be made at an earlier, presymptomatic, stage—for example, at 3 years—will apparently increase survival time (Fig. 3(b)). This apparent increase in survival time does not necessarily mean that screening is effective; it may simply mean that the person with the presymptomatic disease found by screening is aware of the condition for 7 years as opposed to 5—this is referred to as lead-time bias. It is essential that any screening programme is evaluated within a randomized controlled trial which has been designed with death as the outcome in order to control for lead-time bias.

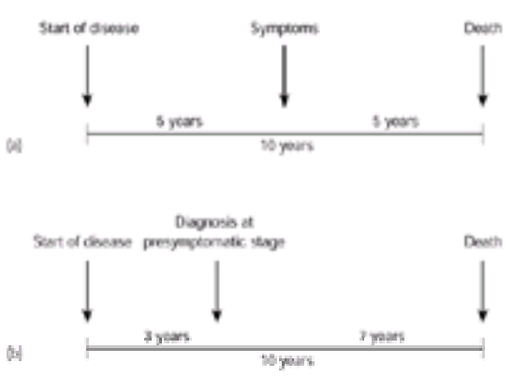

Fig. 3 Lead-time bias.

Length-time bias

Imagine a disease that consists of a number of subtypes; almost every disease has been shown to consist of subtypes when more sophisticated methods of diagnosis and classification have been developed. Imagine those subtypes are of two sorts, those that kill very quickly and those that kill very slowly. A screening programme based on regularly repeated tests is inevitably going to identify more of the slow-growing diseases. Thus it might be possible to describe an improvement in survival following the introduction of a screening programme, but if the only people whose survival has been estimated are those with slowly progressing disease, this does not warrant the conclusion that screening is effective.

This problem, sometimes called length-time bias, is classically demonstrated in the case of prostatic cancer. At the age of 80, about one-third of men have evidence of prostate cancer, but only about 8 per cent of men develop symptoms of prostate cancer. It is now thought that there are at least two types of prostate cancer, sometimes called the tigers and the pussycats, with the pussycats being very slow growing and the tigers very rapidly growing. One interpretation of the results of prostate cancer screening is that it is very good at diagnosing the pussycats but has little or no impact on the tigers.

For these reasons it is important to conduct randomized controlled trials of proposed screening programmes and to prepare a systematic review of the evidence from several trials if more than one has been done.

The harm from screening

No screening test is 100 per cent sensitive and specific; furthermore, initiatives to increase sensitivity almost always decrease specificity and, vice versa, initiatives to increase specificity decrease sensitivity.

People who are true positives, who have both the test and the disease, suffer least harm from a screening programme. However, not all the people identified with the disease will benefit from screening; for example, for a proportion of women whose breast cancer is detected by screening the outcome remains unchanged—they still die of breast cancer. The effect of screening on the whole population has been to increase the probability of survival but has also meant that when they die this subgroup will have known that they have had cancer for a longer period of time.

People who are false negatives can be said to be harmed by screening if they are falsely reassured, but that harm is minimal. The main harm from screening affects those who have false positive tests. People with false positive tests suffer psychological harm, usually short term, but some may also suffer physical harm from the additional tests or treatments that they undergo. In colorectal cancer screening, for example, it is inevitable that someone who does not have colonic cancer will die from colorectal cancer screening as a result of a complication of colonoscopy.

The special ethical issues of screening

All health care involves risk and many people suffer side-effects from treatment. The ethical difference in screening is, however, that screening tests and programmes are offered to people invited to come for screening by the health service, and the ethical contract between the screening programme and those who are screened is different from that which exists between a clinician doing the best that they can for a patient who has sought help.

The problems that might result from prostate cancer screening illustrate the ethical dilemma. Prostate cancer screening leads to the diagnosis of cancers that would not appear in the lifetime of the individual. In a country the size of the United Kingdom, more than 10 000 people would be identified each year whose cancer would never have become clinically evident in their lifetime. If those people were offered radiotherapy or radical prostatectomy, there would be a mortality of about 1.5 per cent from the effects of treatment alone, and about 30 per cent of those treated would suffer significant side-effects such as incontinence and impotence. Thus the screening programme will harm a large number of people with significant numbers dying as a direct result of screening. These harms must be clearly assessed before screening is started.

Special issues in antenatal screening

For antenatal screening, in which the ethical issues are heightened because the 'outcome' is often abortion, the randomized controlled trial is not necessarily the most appropriate design and for some screening this type of trial is impossible. Of crucial importance in appraising antenatal screening is the sensitivity and specificity of the diagnostic test. With this information it is possible to model the effects of a screening programme. The model for Down's syndrome screening is shown in Fig. 4.

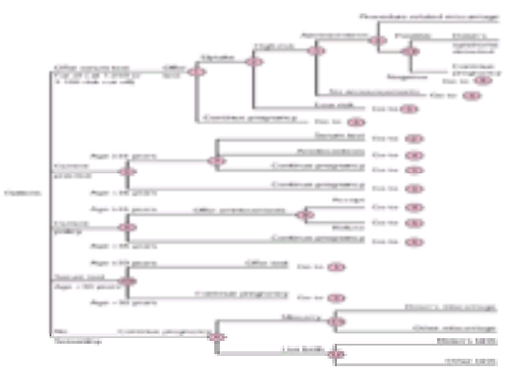

Fig. 4 Decision analysis to show implications of offering screening for Down's syndrome.

However, even the most elegant research study does not actually make a policy decision; policy-makers in public health have to take the best evidence available and then decide whether or not to introduce the screening programme.

Policy-making: assessing the costs

Once the benefits and harms that might result from screening have been assessed, it is necessary to consider whether or not the introduction of a screening programme that has been shown to do more good than harm is a reasonable use of resources for the population concerned, and policy-makers need to take into account a number of issues, as shown in Fig. 5.

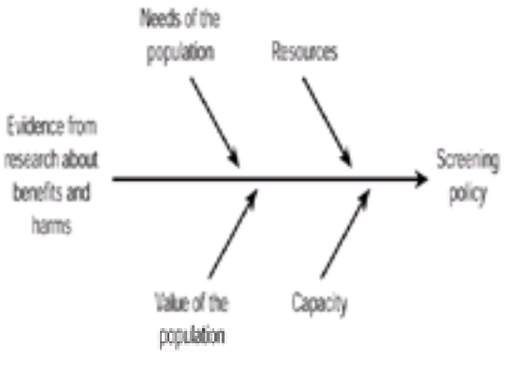

Fig. 5 The ingredients of a screening policy decision.

Research produces the evidence of benefit and harm but other factors are equally or more important.

The needs of the population

The needs of the population and the many other demands made on resources need to be taken into account. The decision about whether or not to screen antenatally for HIV infection should be influenced by the prevalence of HIV infection in the population. Similarly, the decision about whether or not to screen for haemoglobinopathies will be influenced by the prevalence of those disorders in the population.

Resources

The cost of the potential screening programme has to be taken into account. One way to think of screening is to think of opportunity costs, namely what else could be done for people with that disease with the money that is being invested in screening. For this reason it is sometimes useful not to think of the proposed screening programme in isolation—for example, not to ask 'shall we screen or not screen?', but to pose the question 'would the resources invested in screening obtain better results if they were put into primary prevention or treatment services for this health problem?'.

The values of the population

Like any other public health service, the values associated with screening are important. Society places a high value on prevention but, for example, if the values of the population are against abortion then many antenatal screening programmes will not even be considered.

Capacity

The results of screening done in research settings are helpful but have their limitation. For example, in considering the results from the Swedish Two Counties study of breast cancer screening, the Department of Health in the United Kingdom had to take into account not only the results of the research but also the facts that:

1. the research was being done by people who were among the best in the world at mammography and the management of early breast cancer; and
2. 100 screening teams of the same size as the research team would have to be recruited, trained, and supported for the results obtained in research to be reproduced in practice.

In considering the relevance of research to everyday practice, it is sometimes useful to pilot the proposed screening in an ordinary service setting, principally to see if a sufficient level of quality can be obtained, for without good quality management screening should not be introduced.

Quality management

The quality of a service, as defined by the guru of quality management, Avedis Donabedian, is that quality is measured by the degree to which the service conforms to preset standards of goodness.

The importance of quality assurance

The balance between the good and the harm of a clinician or a service is a function of the quality of the service. If the quality improves, the benefit increases and the harm decreases, and there comes a certain point, as shown in Fig. 6, where a service can do more harm than good.

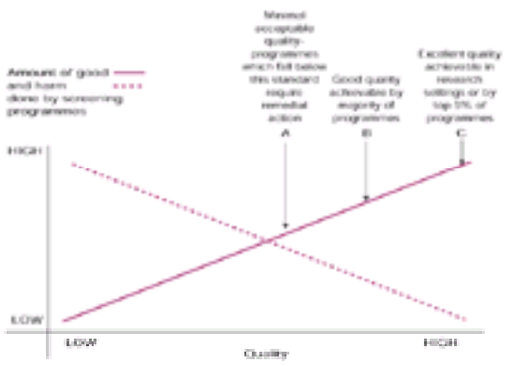

Fig. 6 The relationship between programme extremes and quality, showing the balance between good and bad effects at different levels of quality.

The challenge for the person responsible for a public health service such as screening is that they are presented with evidence from a research setting which usually has a very high level of quality and they have to ensure that they can achieve the quality standards at point B on Fig. 6, even if they cannot, immediately, achieve the quality standards at point A. Standards are arbitrarily chosen and different levels of standards can be set. The standard set by the research team working to a very tight protocol with highly committed and expert staff is a standard of excellence. The pursuit of excellence is obviously an objective of quality assurance but it is not realistic to expect every service to achieve the level achieved by the best in the world, and a second level, indicated at point B in Fig. 6, is usually defined as the achievable standard. It is also important to identify the minimal acceptable standard below which no service should fall, point C in Fig. 6, for it is important to ensure that of the hundred or so programmes covering an entire population not one of the programmes should fall below the minimal acceptable standard and do more harm than good.

Running screening programmes as systems

A system is a set of activities with a common set of objectives and once the objectives of the programme have been chosen it is possible to identify criteria that can be used to measure progress towards those objectives. Standards can also be set on the basis of the criteria with targets for annual quality improvement being selected.

This approach to systems thinking in screening programme management is shown in Table 1.

The objectives of quality management

Quality management of screening programmes has four objectives.

1. The prevention of quality failures, e.g. by good training and the appropriate choice of equipment.
2. To identify and resolve speedily quality failures that occur; failures in screening programmes are often the focus of intense press and media interest with public alarm being generated by the discovery that one screening programme has a higher number of false negatives than is acceptable.
3. To help all those involved in screening continually to improve their performance; screening is essentially a boring repetitive activity and a high level of

commitment is needed to ensure that staff remain motivated; this is particularly important in those screening programmes that depend on the perception of an observer, for example visual perception of cervical smears or mammograms, or the auditory perception of Koratkow sounds.

4. To reset standards regularly; it is not possible to stand still in screening programmes—unless one is constantly striving for improvement there is a risk that the programme will fail because boredom and loss of motivation may reduce quality. The resetting of standards provides new targets at which clinicians and managers can aim.

Achieving continuous quality improvement

Continuous improvement in quality is the objective of screening quality assurance and it requires:

1. an evidence-based screening system with objectives, criteria, and standards;
2. an information system that allows performance to be measured and compared with the standards; and
3. authority to drive quality improvement in all aspects of the programme and to take action if necessary if any part of the programme falls below minimal acceptable standards.

A screening programme requires careful management. The main responsibility for assuring quality rests with the clinicians and managers who are actually responsible for delivering the screening programme, but it is also important for management to invest in quality assurance, activities that are not directly under control of clinicians and managers but which play a part in helping the two groups assess their performance, compare it with the performance of others, and take action either to solve problems or improve performance.

The future of screening

Much has been written about 'genetic screening' but it is in fact hard to generalize about genetic screening for the term is not particularly useful. Certainly new knowledge about genetic risk is raising new opportunities for screening, either in hunting for a particular gene that increases risk or in considering the use of family history as a means of identifying subsets of the population that might be offered screening earlier than the general population, for example breast cancer screening before the age of 50; or offered more intensive screening, for example referral to a cancer genetic clinic. However, the same criteria as outlined above in the section on the good and harm of screening are as relevant to screening based on genetic risk factors as for any other risk factor.

Perhaps more important implications for the future are the changing social attitudes towards technology and risk. Certainly it appears to be much more important to inform people being offered a screening test about the possibility of risk and harm associated with screening. In the past, for example, people have simply been offered blood pressure screening or sometimes been given an antenatal screening test without the implications being fully explained to them. In the twenty-first century this will be unacceptable and all choice will need to be informed. The choice has hitherto been based on statistics defined for populations so people can be told, for example, that the benefits of screening for hypertension outweigh the risks. However, in the twenty-first century, with members of the public much better informed and the World Wide Web providing unprecedented access to knowledge, of both good and poor quality, people will increasingly want to know about how the risks relate to them as individuals. Screening is a public health service that focuses on individuals and which brings clinicians in contact with individuals in a manner analogous to traditional clinical practice. In the twenty-first century screening will be based on even sounder epidemiological evidence, derived from studies of whole populations, than has been the case in the twentieth century. Individuals offered screening will need help and advice in deciding whether or not to accept the clinician's proposal and will have an increasingly important part to play in helping them decide whether or not the screening programme offered is appropriate for them as individuals.

Further reading

Donabedian A (1980). The definition of quality: a conceptual exploration. In: *Explorations in quality assessment and monitoring: Vol. I: The definition of quality and approaches to its assessment.* Health Administration Press, Ann Arbor.

Gray JAM (2001). *Evidence-based healthcare*, 2nd edn Churchill Livingstone, Edinburgh.

Wilson JMG, Jungner G. *Principles and practice of screening for disease*. Public Health Paper No. 34. World Health Organization, Geneva.

3.7.1 The cost of health care in Western countries

Joseph White

Reasons for interest

The countries conventionally called 'Western' (a term that refers to level of economic development, not geography) have the physical and financial resources needed to provide a high standard of health care to all citizens. Yet doing so at a socially acceptable cost is difficult. Efforts to limit cost may affect the quality of care, access to it, and the work and incomes of medical providers.

Defining the issues

One view of costs emphasizes health care's share of a national economy, and thus the economic 'burden' of care. Another view emphasizes costs relative to service or outcome, and therefore value for money. Those who object most to spending levels, such as employers who contribute to insurance, or government budget-makers, often claim they are concerned with value for money, to show that they are not trying to reduce needed care.

Cost control should be distinguished from 'cost-shifting'. If tight budgets for the British National Health Service cause many patients to 'go private', that is cost control from the government's perspective. From the patients' and providers' vantage point, some costs have been shifted, not reduced.

Why is cost a policy issue?

The cost of health care is more of a public issue than that of most other commodities or services because all Western countries remove much health care from the market's allocation processes. Markets allocate resources by price. A person who cannot afford an item then may save for it, borrow for it, do without, or buy a lower-quality alternative. All of these options can be impracticable or socially unacceptable for a large number of medical needs, so some other method of allocation is needed.

All advanced industrial countries free sick people from worrying (too much) about cost by combining individual contributions into pooled funds. Whether those pools are private or public, and cover all or part of the population of an area, their effect and intent is to protect members from the price constraints that otherwise would limit spending. Some other method of controlling spending on individuals' care is therefore needed, in order to limit contributions to the pool.

Spending trends

Table 1 shows health care spending as a percentage of national economies for 11 countries with constant borders from 1975 to 1995 (Germany is therefore excluded). Generally the 'burden' of health care spending has risen, but much more quickly in some countries than others, and at different rates over time within the same country.

One common explanation of spending increases emphasizes the ageing populations in Western countries. Yet careful analyses suggest that longer life expectancies explain only a small part of the trend, because greater life expectancy implies that mortality, and its attendant costs, are lower at each particular age.

Some analysts believe that individuals simply prefer health care to other goods so that, once other basic needs are met, higher income will lead to disproportionately higher consumption of health care services. The methods of payment, however, make it hard to determine to what extent spending represents individual preferences.

Rapid innovation in treatments and diffusion of knowledge about them usually interacts with pooled financing to encourage increases in cost. Particularly when policies promise a high standard of care, knowledge of a new, supposedly useful treatment creates pressures to make it available to everyone, as opposed to the norm with market goods, which can be rationed legitimately by price.

In spite of this generally upward trend, Table 1 shows that both spending levels and rates of increase vary greatly. A nation's cost level reflects its history: for instance, the development of national medical capacity, provider income expectations, and practice norms. Spending can also vary due to differences in underlying causes of illness, such as poverty, diet, smoking, alcoholism, violence, and genetic endowments. Yet these factors are unlikely to explain the different rates of growth.

Cost control policies

The main reason why trends in medical costs differ is that nations have made different efforts to restrain them. We can distinguish cost control efforts in terms of the targets of control and the systems of health care finance.

Targets

National policies vary in both the emphasis on and success of targeting different types of services, such as inpatient care or pharmaceuticals. Until recently, national medical norms and the economic and political power of different providers favoured, for example, surgery in the United States and pharmaceutical treatment in Japan. Sometimes advocates argue that higher spending on one service would reduce expenses for others: for instance, that expanded insurance for ambulatory care or pharmaceuticals would reduce hospital costs. Whether such expansions of insurance ever reduce total costs (as opposed to providing better value) is doubtful.

Policies may target different factors that influence spending for a given type of service. For example, they may focus on restraining prices per service or numbers of services or, somehow, both. Another common approach is to limit physical capacity to provide care, and so both volume and prices. For instance, fewer MRI machines per person will probably mean fewer scans and, since each machine is used more often, lower costs per scan.

Research showing that many procedures lack scientific support, and demonstrating wide, medically unexplained variations in practice among communities, encourages another type of policy: measures to make care more 'appropriate', such as treatment guidelines. Sometimes termed 'evidence-based medicine', these measures are promoted as offering savings without harm to patients.

Systems

Cost control is affected by the systems used to pool funds and pay for services.

The extent to which individuals pool their resources may range, in theory, from not at all (individuals pay all costs out of pocket) to total (all care is paid from a single pooled fund). In Western countries, the range is from most people being insured for most services, with some cost-sharing (as in the United States); to virtually

everyone insured, for most services, in a single pool (as in Canadian provinces, and the British National Health Service, **NHS**). Both the low and high ends of this continuum are likely to control costs better than some position in the middle. If people pay substantial amounts from personal funds they will face considerable price constraints. But extensive pooling can limit costs, because payment through a 'single pipe' provides a single point of control: limit the total flow, and you can limit spending.

A similar pattern of effectiveness can be seen in choices about how individual caregivers are paid. At one end of this payment continuum, virtually all providers of a given service are paid on the same terms, even if there are multiple payers. Such co-ordination gives the payer(s) dominant market power, which can be used to enforce (relatively) low prices. At the other extreme, payers contract selectively with providers. Then hospitals or doctors may agree to reduce their fees, or accept some outside management of their practices, so as to preserve access to patients. Co-ordinated payment is usual in most Western countries. Selective contracting is the basis of much of American 'managed care', and of 'internal market' reforms such as fundholding by British general practitioners (**GPs**). Unselective contracting by uncoordinated payers, as in the United States around 1990, seems to be the most costly possible approach.

Highly pooled funds tend to be associated with co-ordinated payment. Yet selective contracting may be implemented within a large pool, as in NHS GP fundholding, while high cost-sharing could be combined with standard fees. Thus distinguishing the two dimensions allows more careful analysis than a simple division into more 'market-like' or 'regulatory' approaches.

Choices about targets and systems affect only potential for cost control; performance depends on the specific measures chosen and the commitment of payers to control costs. Extreme positions on these two dimensions none the less have very different risks, both for other health care variables and interests. For example, higher cost-sharing is probably a more direct threat to equity of care, but a lesser threat to the incomes of providers, than some other cost control measures. The much-vaunted 'marketization' of Western systems has generally involved moves towards selective contracting rather than lesser socialization. The evidence does not yet show that highly selective contracting controls costs more effectively than highly co-ordinated payment.

Paying doctors

Cost control choices have many implications for doctors in particular. They tend to prefer higher cost-sharing to alternative cost control measures. But cost-sharing is particularly unpopular with voters, so policies tend to focus instead on the structure of doctors' compensation.

Economists worry that doctors may respond to fee restraints by encouraging patients to consume more services. Capitation or salary would not be vulnerable to that response, but may create incentives to provide fewer services than are necessary. Both private and public policy-makers therefore seek mixed compensation schemes that might limit either failure. In most countries doctors have much preferred fee-for-service payment. Yet methods that reduce fees automatically in response to increased services, as prevailed in Germany from 1985 to 1997, seem to counter the effect of any volume increases, leaving doctors as a group working more for the same money. In many countries, therefore, doctors' organizations are rethinking their preferences about payment methods.

Policy-makers virtually everywhere would like to get the medical profession to 'manage care', making it more appropriate, and to manage the balance among price, volume, and quality within a budget. Unfortunately, even German physicians, with a supportive tradition and institutions, find corporate self-management difficult.

As an alternative, some schemes make individual doctors or groups of doctors responsible for the cost of pharmaceutical, hospital, or other services. Unless these responsibilities are limited and spread across a large payment base, they can create dangerous incentives to deny necessary care.

Fairy tales and the future

Policy-makers and critics dream of more 'rational' measures than those discussed above. One is that policies to keep people healthy would save lots of money on acute care. There is little or no confirming evidence. Another is that societies could prioritize medical services within budgets, excluding those that science shows to be least cost-effective. But such general rules would not be appropriate for many individual cases. The state of Oregon's well-publicized prioritization of treatments in its insurance for the poor has contributed little to cost control, both because the covered list was quite extensive, and capitated providers found that distinguishing covered from uncovered treatments was not worth the trouble. Instead, each Western country must choose its own imperfect policy, with attendant consequences. The trend towards relatively increased spending on health is likely to continue, because longer, less painful lives seem desirable to most people, and modern medicine is believed capable of providing that. But systems will vary in both total costs and cost-effectiveness.

Further reading

Barros PP (1998). The black box of health care expenditure growth determinants. *Health Economics* **7**, 533–44.

Dudley RA *et al*. (1998). The impact of financial incentives on quality of health care. *The Milbank Quarterly* **76**, 649–86.

Jacobs L, Marmor T, Oberlander J (1999). The Oregon health plan and the political paradox of rationing: what advocates and critics have claimed and what Oregon did. *Journal of Health Politics, Policy and Law* **24**, 161–80.

Reinhardt U (1996). Our obsessive campaign to 'gut' the hospital. *Health Affairs* **15**, 145–54.

Rice T, Morrison KR (1994). Patient cost-sharing for medical services: a review of the literature and implications for health care reform. *Medical Care Review* **51**, 235–87.

White J (1999). Targets and systems of health care cost control. *Journal of Health Politics, Policy and Law* **24**, 653–96.

3.7.2Health in a fragile future

Maurice King

Politics is health. Health is also politics—the operation of power in society—to the grave disadvantage of today's poor and tomorrow's everybody. The traditional enemies of public health—infection, etc.—have been reinforced by those three serpents: the world, the flesh, and the devil, newly armed and globalized for the coming millennium. For 'the world' read 'Mammon'—the overarching quest for wealth, and the power that it brings. For the devil and the flesh read on.

This is written at the start of the new millennium. The human population of the earth has just exceeded 6 billion (12 October 1999), having tripled during the writer's lifetime. Mankind has already passed the highpoint in per capita oil consumption (1979) and in grain consumption (1985). The average human is unlikely ever to have so much energy at his disposal again, or to be so well fed.

Wealth

One superpower with 4.6 per cent of the world's population now dominates the earth. Half the world's health expenditure is spent there, and a quarter of its fossil fuel is burnt there. Most serious, that superpower dominates what the world thinks, and it dominates the United Nations system. Within it, money dominates. With a public 'dumbed down' by the media, power is increasingly concentrated, not in Congress and Senate, but in the Department of State, in lobbyists, in think tanks, and in the boardrooms of transnational corporations, especially those which control the media. Domestically, the proper functions of government are being increasingly abrogated to lucrative struggles between lawyers.

The market economy reigns supreme, with no credible alternative on the horizon. The market is now advertisement-driven to produce ever more luxurious resource consumption, and ever less sustainable lifestyles, to the point that tourism has now become the largest industry on earth. The rich become ever richer, and the gap between rich and poor ever wider. More and more wealth is in fewer and fewer hands. For the fortunate, wealth effortlessly creates more wealth as never before. Meanwhile, during the last 20 years, 60 countries have been getting steadily poorer.

With the coming of the new millennium, and the stock markets buoyant, the mood (for the fortunate) is bullish. Unfortunately, the prosperity bubble is fragile. Besides the technical problems of the stability of the global economic system, the bubble can only continue to swell if certain major problems are overlooked—population, global warming, global food, the rising energy costs of fossil fuel exhaustion, and the effects of the media on the stability of the family. One or more of these could prick the bubble and all have the gravest implications for health. In the United States, the market economy requires a per capita carbon dioxide production of 20 tonnes annually, and in the United Kingdom 10 tonnes. Globally, the sustainable fair share is probably about 1.4 tonnes. I argue for radical adaptations in lifestyle, North and South, for equity, frugality, ecology, economy, and for the intense dialogue that must accompany the solution of our major problems. Susan George reminds us that the fossil fuel based economic activity of the present 2 billion 'haves' who burn that fossil fuel, is destroying the world, and that the 9 billion, who by 2050 will want that economic activity, will destroy the world even more disastrously — with no adequate means of either controlling or sharing such economic activity as the world might support sustainably. Meanwhile, the 'have nots' who will not enjoy such benefits are likely to become increasingly violent, North and South. The market economy is therefore destroyed either way. She concludes that we cannot both sustain the liberal free-market economy and continue to tolerate 'the superfluous billions'. This was published in 1999. September 11 2001 has proved how right she is. Before discussing the difficulties of opening this dialogue on a sufficient scale, there is a more immediate problem for physicians.

Health as a free market commodity

Now that the service sector of the economy is growing faster than manufacturing, it has become a tempting source of profit, with the result that the unfettered free market is becoming increasingly perilous for those who cannot afford private health care. Traditionally, most governments have sought to provide some form of insurance-based and tax-supported 'national health service' which was available to everyone, and which strove to 'generalize the adequate' in health care. Of these, the British National Health Service was much admired.

High on the agenda of the World Trade Organization (WTO) is the privatization of education, health, welfare, and social housing. The race is on to capture that share of gross domestic product (GDP) which governments spend on public services, and to open the European public market to transnational competition to the disadvantage of universal coverage, of solidarity through risk pooling, of equity, of comprehensive care, and to the European tradition of democratic accountability—largely in the pursuit of profit.

Opposition to all this is disorganized and inchoate. Over a thousand non-governmental organizations (NGOs) opposed the WTO's 1999 meeting in Seattle. Out of this soup of opposition a coherent, high profile, alternative has yet to crystallize. The problem is that, whereas the WTO's objective is simple—profit—that of the 'the global new green left', if such it can be called, is complex, and lacks a tangible manifesto. It is part green, part animal rights, part local autonomy, part antitransnational, and part much else. It is also divided North/South. Since it is commonly at variance with itself, there is no single voice to articulate the opposition to the free market status quo, and to debate the many problems connected with it. The greatest of these is population.

Demographic disentrapment

Demographic disentrapment is the crux—communities exceeding the carrying capacities of their ecosystems with nowhere to go, and with economies that do not enable them to compete for food on the world market. The Belgian administration had long considered Rwanda trapped. The classic case of as yet peaceable entrapment is Malawi, which has outgrown the carrying capacity of its fragile ecosystem, and where chronic malnutrition is endemic. Half its children are stunted. Expert opinion considers (privately) that much of Africa is trapped. There are also grave doubts about the ability of India to feed a population which is expected to increase by 50 per cent and exceed that of China. Its per capita grain is already falling gently, and water is an even greater constraint than land.

Starvation permitting, Malawi's population, without AIDS, was set to have tripled by 2050, and eventually to have quadrupled. With AIDS it is expected to increase by only 50 per cent. Without AIDS, immediate one-child families would have been essential. Now that AIDS has removed Malawi's demographic momentum, two-child families may be sufficient. The tragedy of AIDS has made the disentrapment of Malawi possible. This opportunity has to be grasped courageously on a continental scale.

Demographic entrapment is tightly taboo to demographers, to development economists, and particularly to the United Nations system. It also taboo to most NGOs. There amounts to what the United States demographer Jason Finkle has termed a 'population policy lockstep', in which demography, the great foundations, and the United Nations agencies support the same policies, and in particular deny (publicly) that demographic entrapment exists. The Cairo population conference in 1994 avoided all population targets, and by 1999 at 'Cairo plus five', population control was even further out of fashion.

There are many reasons for this lockstep. The most politically sensitive one is that that the United States Department of State appears to be actively interested in maintaining it—presumably, since if the South is to restrict its fertility, the North will be expected to restrain its resource consumption, with the serious implications that this has for the market economy, and for United States resource consumption in particular (see www.leeds.ac.uk/demographic.disentrapment).

In effect, there appear to be many more 'Chinas' and 'Malawis' in urgent need of one-or-two-child families, if they are to avoid starvation and slaughter, but no demographer dare say so publicly. China's triumph is that one-child families have become the socially-responsible norm. Equity requires that if any community is to be counselled to reduce its fertility drastically, we all should. If therefore the United Nations is to suggest that any country should do so, it will have to be in the context of

a United Nations programme for 'a one-or-two-child world'—we are all in it together. This cannot be a tight directive, but has rather to be a 'general political direction' in which the world needs to go. In the uproar over population that would result, Italy for example, with a total fertility of 1.2, and which is afraid that it might disappear, might try to raise its fertility.

Demographic entrapment is only one of many problems in which population plays a part, often a large part. Others include poverty, hunger, malnutrition, increasing global inequality, deforestation, street children, and, indirectly, global warming. To make demographic entrapment taboo is also to hinder gravely the resolution of these other problems —in effect to obstruct the solution of all the major human problems. The consequences of doing this, particularly if the desire to preserve Northern levels of resource consumption, is ultimately the major factor, are diabolical—'overpowering evil' which, incidentally, the *British Medical Journal* saw fit to list as a technical term.

A common fear underlies all this—that of an anarchic, overcrowded world. '*Quand viennent les Africains?'*, a Swiss acquaintance asked, not 'if they come' but 'when are they actually coming?' Many have already come. Equity demands that more should come. Reason argues that all cannot come, say a billion from Africa which is having increasing difficulty feeding itself, and where the fertility of the land is falling widely, and perhaps another half a billion from Asia. The future of the world is brown. How brown? What does seem clear is that for the majority of the trapped, disentrapment will have to take place *in situ*, if they are not to starve or slaughter themselves on the spot. But there is another threat to the family, besides that to its size.

The flesh and the family

The most important institution for human welfare is the family—that cradle of the virtues, 'in sickness and in health', 'for richer for poorer'. The fact that the family is falling apart so widely is thus of the gravest consequences for health and human well being. Why is the family in such decline? Taboos, it seems, are important for maintaining the structure of society. The taboo on entrapment helps to preserve the current North/South power structure. It may well be that the many kinds of taboo with which all societies have surrounded sex—'who may say, or show what, when, where, and to whom', play an important part in maintaining the structure of the family, and that the weakening of the taboos, under the persistent attrition of the media, to which we are all increasingly exposed, from the age of three onwards—in effect 'anything goes'—bears a large part of the blame. We are each the recipients of many years of nurturing in the bosoms of our families. When we die we take these blessings with us. A society which fails to replace the 'social capital' generated by the family, does so at its peril. The consequences of this loss are all around us.

So where is the devil?

Any serious writer or *mediaste* hopes to influence what other people think. The devil lies in what he wants them to think—and not to think—and why, and on the scale and the means by which he does it. Even modest deviance from 'political correctness' becomes ever more difficult in an increasingly 'one-think', one-language, one-superpower, CNN world, with its ever more narrowly focused centres of wealth and power. Frances Stonor Saunders has documented the multifarious ways in which the United States Department of State manipulated the intelligentsia, both left and right, during the Cold War. There is no reason to think that this manipulation has ceased. The policing of the population policy lockstep is a particularly clear example. A really free unlockstepped press capable of debating anything, however alarming, and whatever implications it has for northern lifestyles, is critical.

The idea that, in a world of supposed free speech, the media need to be controlled in order to protect the family, and to prevent our behaviour being too gravely manipulated by scientifically crafted advertisement, and if so how, and by whom, is a thought so outrageous as to defy credulity. But so was the *contagium vivum,* and a spinning earth. We are just starting to think about tempering the free market, the proposed Tobin tax on financial transfers being the outstanding example. But this must be only the beginning.

So as we pore in triumph over the atoms in our chromosomes, let us not forget that we have no clue as to how those atoms bring us conscious minds, and that the devil is as busy as ever in influencing what those minds think, and in narcotizing our consciences so us to prevent our getting to grips with the terrifying problems that beset us—and those who come after us. The first millennium started with a command to care for all men. The important difference now is that we have 30 times as many to care for, and that for 'our fellow men' we should now read 'all terrestrial creation'. Such, then, is the fragile future of public health in a new millennium.

Further reading

Campbell CJ and Lacherrère JH (1998). The end of cheap oil. *Scientific American*, **278**, 78–83.

George S (1999). *The Lugano Report on preserving capitalism in the 21st Century*. Pluto Press, London.

King MH (1999). Commentary: bread for the world—another view. *British Medical Journal* **319**, 991.

King MH (1999). The US Department of State is policing the population policy lockstep. *British Medical Journal* **319**, 998–1001.

King MH and Elliott CM (1997). To the point of farce: A Martian view of the Hardinian taboo—the silence that surrounds population control. *British Medical Journal* **315**, 1441–3.

Saunders FS (1999). *Who paid the piper?* Granta Books, London.

Willey D (1997). Population control: a necessity for the preservation of individual liberty. *Politics and the Life Sciences*, **16**, 228–30.

4.1 The genomic basis of medicine

Trevor Woodage and Francis S. Collins

Introduction

Much of the progress made in biomedical science in the past century has been due to an increasingly sophisticated understanding of the cellular and molecular mechanisms underlying disease processes. Deciphering the nucleotide sequence of the DNA molecules that make up the 46 chromosomes found within each human diploid cell will represent a major milestone in our ability to understand these processes. In particular, knowledge of the sequence of the human genome will help us understand the structure and function of the proteins and RNA molecules that control the developmental programmes and functions of human cells in health and disease. The Human Genome Project is a large, multinational effort that has among its goals determining the complete human genetic sequence by the year 2003, although most of the sequence was already available in 2000, and two groups published draft sequences in 2001. Along with its technical goals, the Human Genome Project has brought dramatic changes in the ways in which biological and medical processes are viewed. This chapter will discuss some of these influences on modern medical theory and practice.

The terms genetics and genomics are often incorrectly used interchangeably. Genetics refers to the study of inherited traits or characteristics, while genomics describes the study of the large-scale structure and composition of the material encoding these genetic instructions. Applying a genomic approach to medicine involves an appreciation of the complex set of interactions that govern the repertoire of genes that are expressed in different tissues under conditions of health and disease.

Scope of the Human Genome Project

Although initially met with widespread scepticism, proposals to determine the complete DNA sequence of the human genome began to circulate in the mid 1980s. This objective was ambitious both in its scientific scope and in its technological requirements. The suggestion that efforts to discover the sequence of individual genes, answering specific biological questions, were to be supplanted by industrial-scale programmes to analyse large numbers of genes or chromosome regions using novel methods and equipment was not well received in all quarters. Debates about the relative merits of gene-based (cDNA) versus whole genome sequencing projects eventually moved to a consensus (but not unanimous) position that both approaches were worth pursuing.

The development of several generations of semiautomated DNA sequencing apparatus that helped to drive down costs gave impetus to arguments that large-scale sequencing efforts were feasible. Even so, the international Human Genome Project was expected to require funding of 3 billion dollars over 15 years to sequence the estimated 3×10^9 base pairs (bp) making up the haploid genome. A substantial proportion of that expense was not directly attributable to DNA sequencing, but rather to providing infrastructure in the form of genetic and physical maps of the human genome, and work on the genomes of model organisms. Construction of these maps, in turn, depended upon the development of new DNA cloning and analytical methods, especially large-insert cloning vehicles such as yeast artificial chromosomes (YACs) and bacterial artificial chromosomes (BACs). YACs and BACs allow the propagation of extended regions of genomic DNA (100 000 to more than 1 million bp) in multiple copies in cultured yeast and bacterial cells respectively. Availability of these cloned DNA molecules allows analysis of large, but manageable, chromosomal fragments that encompass complete genes.

Exceeding expectations, and with the entry of private sector efforts to spur progress, it appears that the complete sequence of the human genome will be available substantially before the initial target date, 2005. In fact, in June 2000 announcements of the completion of the first draft of the human genome sequence were made, and in February 2001 publications appeared describing the draft sequence and its initial analysis. Both public and private sector sequencing efforts arrived at similar estimates of human genome size, at approximately 3.1×10^9 bp, remarkably close to earlier estimates based upon measurements of the amount of DNA contained in an average cell's nucleus. Continuing attempts to provide an essentially complete, ordered representation of the sequence of the human genome are now planned to be finished by 2003.

Generation of these huge amounts of DNA sequence is of little benefit if there is no effective way of interpreting what it means. The creation of large databases containing different kinds of genomic information together with sophisticated software algorithms that can detect protein coding regions, compare related sequences to each other and identify biologically relevant patterns or sequence motifs have given rise to the rapidly expanding discipline of bioinformatics. Although most public attention paid to the Human Genome Project has, naturally enough, been given to sequencing the human genome, early realizations that one of the most effective ways of interpreting sequence data was by cross-species comparison led to the establishment of parallel efforts to sequence a series of model organisms. The complete, or near complete, genomic sequence has been determined for over 20 bacterial species including *E. coli* (5 million bp); *Saccharomyces cerevisae* (12 Mb), a simple, unicellular eukaryote, the common baker's yeast; *Caenorhabditis elegans* (97 Mb), a simple multicellular nematode used in many types of basic research; and *Drosophila melanogaster* (120 Mb), the fruit fly used by biologists for almost 100 years as a model for studying genetics and development. Major efforts to sequence the genomes of the mouse, rat, and zebrafish are underway. These efforts are providing a solid foundation for understanding the genomic basis of many aspects of molecular and cellular physiology.

Implications of genomics for medical practice

Disease gene identification

Initial successes (prior to 1985) in cloning genes that were mutated in human genetic diseases required a detailed understanding of the pathophysiology of the disorder being investigated and the nature of the proteins whose sequences were altered (such as the roles of the a-globin and b-globin genes in the thalassaemias). Unfortunately, such approaches could not be applied in most cases. For example in cystic fibrosis, the most common serious autosomal recessive disease to affect Caucasians, little was known about the molecular causes of the respiratory and gastrointestinal problems that occurred in patients with this uniformly fatal disease. The lack of any detailed knowledge about the protein whose function was disrupted in cystic fibrosis meant that it was impossible to use standard methods to proceed from phenotype to protein to gene. In order to discover the gene that was mutated in cystic fibrosis, it was necessary to follow a novel strategy, which has become known as positional cloning.

Because of its pattern of inheritance, the gene that was mutated in cystic fibrosis was known to lie on one of the autosomes (chromosomes 1–22) rather than on a sex chromosome (the X or Y chromosomes). Cystic fibrosis carriers would have one abnormal copy of the gene and affected individuals would have mutations in both

copies of the gene. By examining the pattern of inheritance of a panel of variable DNA markers, it was possible to identify genetic markers lying on the long arm of chromosome 7 whose inheritance patterns correlated with carrier or disease status more often than expected by chance alone. These markers were thus said to show genetic linkage to the cystic fibrosis gene. The cystic fibrosis gene candidate interval was narrowed by identifying additional polymorphic DNA markers that were more finely distributed throughout this part of chromosome 7. Even when this process, linkage analysis, is successful, the chromosomal region containing the disease gene usually consists of several hundred thousand to several million base pairs, typically containing 10 to 100 genes. Clearly, this is a considerable improvement over individually investigating the 30 000 to 40 000 genes that the human genome is estimated to contain, but it is still a daunting target.

The DNA sequences that make up the protein-coding regions of genes occupy only 1.5 per cent of chromosomal DNA (the rest is largely repetitive DNA and spacer elements of unknown function). Thus, it is not a trivial matter to identify all of the genes in a candidate interval of a million or more base pairs. A number of techniques have been developed to identify the genes within a defined interval of chromosomal DNA, most requiring large cloned DNA molecules covering the chromosomal interval under investigation. Recent improvements in the ability to identify and obtain these cloned molecules using BACs and YACs have greatly facilitated gene identification, but all will soon be superceded by the availability of the complete, ordered, human DNA sequence, together with ever-improving annotations of gene location and structure.

Once the genes within the chromosomal segment containing the disease gene have been discovered, the laborious task of finding which of these genes contains the pathogenic mutation is undertaken. In some instances, easily recognized abnormalities such as large deletions or chromosomal translocations disrupt the genes, but in most cases subtle changes in DNA sequence produce disease. The most definitive subtle alterations are mutations that result in the premature termination of a growing peptide chain during the translation of an mRNA coding sequence into protein. Thus, the simple alteration C to A, changing the codon TAC (encoding tyrosine) to TAA (a stop signal) will prematurely terminate protein manufacture. Similarly, the deletion or insertion of a small number of bases (not divisible by three) will cause a frame-shift in which interpretation of codons downstream of the mutation site will be scrambled, and a premature stop codon is usually formed by chance. Simple sequence changes that do not produce premature termination signals may be harder to recognize as being responsible for disease. For example changing a TCG codon to TCA would be unlikely to result in a disease state because both of these nucleotide triplets code for the amino acid serine. Substituting ACG for TCG produces threonine instead of serine. These two amino acids are chemically closely related and substituting one for the other may or may not result in a significant functional alteration. A TCG to CCG change, however, produces proline instead of serine. This amino acid change is non-conservative and more likely to produce a protein that functions differently from the wild-type form. Other subtle mutations that may be difficult to recognize can involve alterations in the DNA sequences that direct the occurrence of normal messenger RNA splicing, or regulatory elements such as promoters, or enhancers which control gene expression.

There are several ways to distinguish between mutations that cause inherited disease and harmless genetic polymorphisms that do not significantly change protein function or patterns of gene expression. A mutation may segregate in families with the presence of disease but not be found in unaffected controls. Cases of highly penetrant autosomal dominant disease that arise spontaneously in pedigrees where the illness has not previously been noted should manifest mutations in the affected proband that are not present in germline DNA from either biological parent. Comparing amino acid sequences of homologous proteins from different species can also help to distinguish between disease-causing mutations and harmless polymorphisms. If the amino acid in question is conserved across a range of species then it is more likely that this specific amino acid residue performs an important role and that altering it would have a significant effect on protein function. Finding pedigree-specific mutations involving different amino acids in the same gene can also add confidence to the search. A more direct test of the significance of a particular sequence change is to examine its effect directly in a functional assay. It may be possible to introduce a copy of a gene containing the sequence change into an appropriate cell line and measure whether protein function has been altered. To perform this type of experiment meaningfully, the protein being studied must be sufficiently understood to develop an assay.

In the case of cystic fibrosis, one of the genes in the candidate interval of chromosome 7 being investigated was found to have a three base pair deletion that segregated with disease status in many affected pedigrees. This results in the absence of a phenylalanine residue at position 508 of the protein, named the cystic fibrosis transmembrane conductance regulator because of its potential effects on ion flow across cell membranes. Loss of the phenylalanine from the protein disturbs intracellular processing, preventing the mature protein from reaching the cell membrane.

In 1989, the feat of identifying the specific 3-bp sequence alteration represented by the mutation known as ‡F508, amidst a genome 3×10^9 bp in size, was a testament to the potential power of positional cloning. However, when practised without a pre-existing framework of polymorphic DNA markers and large-insert DNA clones spanning the genome, this process proved very laborious, time-consuming, and expensive. Fortunately, as a consequence of the Human Genome Project, high-resolution genetic maps now exist for all of the chromosomes. Thus, it is usually no longer necessary to identify novel polymorphic DNA markers when trying to narrow the chromosomal intervals containing disease genes. Further, once a candidate chromosomal interval has been identified, it is now a straightforward matter to obtain DNA clones and most of the sequence covering the region containing the disease gene. The identification of genes lying within a specific genomic interval was previously greatly facilitated by the mapping of a large number of genes to well-defined chromosomal intervals, and is now even more enhanced by the availability of richly annotated genomic sequence databases. Significant technical advances have also been made in the field of identifying sequence variants that represent disease-causing mutations although, in many ways, this area remains the rate-limiting step in positional cloning projects.

Common diseases as complex traits

While great advances have been made in the realm of identification of the causes of single-gene disorders, genetic factors contributing to the development of more common diseases are, on the whole, much less well understood. It is generally accepted that a hereditary component contributes to the aetiology of almost all types of disease. Evidence in support of this contention is provided by observations that family members of affected individuals have the same disease more often than do members of the general population. Of course, a portion of disease familiality may also be due to shared environmental exposures, and care must be taken to consider this when examining such statistics. Nevertheless, careful study designs have allowed the identification of substantial genetic contributions to such common diseases as diabetes mellitus (both types I and II), asthma, essential hypertension, atherosclerosis, degenerative neurological disorders such as senile dementia and Parkinson's disease, the major mental illnesses, and many forms of cancer. Unlike single-gene disorders, common diseases tend not to segregate in families in easily recognized mendelian patterns. For example first-degree relatives of someone with an autosomal dominant disease such as familial hypercholesterolaemia have a 50 per cent chance of carrying the same mutation in the LDL receptor and second-degree relatives have a 25 per cent chance of carrying the mutation. Risk of developing the disease falls with this predictable pattern, depending directly upon the degree of relatedness to the index case. In common polygenic diseases, however, disease risk declines much more rapidly with the genetic distance from the index case. For this reason, familial clusters tend to be fairly small, making them difficult to recognize and study. Such clusters can also be the result of chance, when the disease is common. It is at least partly for these reasons that intensive genetic analyses of many common diseases remain in relatively early stages.

Common diseases are often known as complex traits because they are thought to be due to complex interactions between multiple genetic and environmental factors. Thus, genetic factors that might contribute to type I diabetes mellitus include a locus within the HLA complex, a polymorphic region just upstream of the coding region of the insulin gene and over a dozen other as yet, uncloned genetic loci in addition to presumed environmental exposures such as Coxsackie virus infection. While the presence of HLA-DQw8 has been associated with type I diabetes, possessing this allele does not guarantee that an individual will develop diabetes. In fact, most people with HLA-DQw8 will not become diabetic unless they also carry other predisposing genetic variants and are exposed to certain, largely undefined, environmental factors. In the absence of a complete understanding of these interactions, it is not yet possible, in most instances, to develop models capable of describing precise risk patterns for complex trait inheritance. Nevertheless, a variety of parameters have been used to quantify the relative contribution of various genetic and environmental factors that play a role in the development of disease. One of these quantitative measures is denoted l, and represents the ratio between the risk of a relative of a patient developing disease compared with the risk of an individual selected at random from the general population. l_s refers to the l value for siblings of index cases. Thus, l_s for type I diabetes is approximately 20, representing the relative risk of a sibling of a diabetic patient developing the disease (6 per cent) compared to the population risk of developing type I diabetes (0.3 per cent). This compares with much higher values for genetic disorders displaying simple mendelian patterns of inheritance. For example the l_s for cystic fibrosis is approximately 750 in Caucasians (l_s will tend to be large for rare disorders with low population prevalence). It should be remembered that l measures the effect of combined genetic and shared environmental factors. Each individual component necessarily must confer a smaller relative risk than the total l. The manner in which distinct risk factors combine to yield the overall value of l for a condition can be additive or multiplicative depending upon whether they are independent of each other or synergistic. Values of l_s for a number of disorders are shown in Table 1. In any case, individual genes that play a role in the development of common diseases may only cause comparatively small increases in the relative risk of developing disease. Assessing the clinical importance of each of these effects may pose substantial challenges in individual patients and they will need to be interpreted in the context of the presence of other genetic variants, exposure to environmental risk factors, and the overall clinical setting. When considered in large populations, however, even sequence variants that are responsible for modest increases in disease risk can be associated with a major public health burden.

While the distinction between 'simple' and 'complex' diseases is a useful one in helping to understand their genetic underpinnings, it is an artificial division. For example approximately 1 in 10 000 individuals possess frame-shift mutations that produce truncated and non-functional versions of one copy of their APC genes, causing familial adenomatous polyposis. This condition is characterized by the development of hundreds to thousands of intestinal polyps. Without prophylactic

colectomy or other preventive measures, there is a high likelihood that one or more of these polyps will undergo malignant transformation. Untreated, most patients will develop cancer of the colon by 40 years of age. Approximately 7 per cent of Ashkenazi Jews carry a non-truncating T®A sequence variant in APC, rendering the gene unstable and prone to further mutation in dividing cells of the colonic epithelium. This APC mutation appears to be associated with a 1.5 to 2-fold increased risk of developing colorectal carcinoma. This contrast between levels of cancer risk for low-frequency, high-penetrance and high-frequency, low-penetrance alleles of the same gene is an important one, and helps to demonstrate the spectrum of risk associated with different disease-associated alleles. Moreover, it helps to remind us that, strictly speaking, genes do not cause genetic disease, specific alleles of genes confer risk.

The challenge of identifying sequence variants contributing to complex disease

Several of the factors that make it difficult to find DNA sequence changes associated with complex disease have already been alluded to above. They include the absence of large numbers of affected family members sharing identical mutations, the fact that many of the sequence variants found in chronic disease are likely to be subtle changes that are difficult to recognize, and the expectation that the predictive power of particular sequence changes will generally be quite low. Patients with the sequence variant may not have the disease (low penetrance) and some patients with the disease will not carry the sequence variant (phenocopies). In contrast to the situation for simple mendelian disorders in which it may be possible to identify pathogenic mutations by analysing DNA samples from only a small number of cases, for complex disease samples will be needed from hundreds or perhaps thousands of cases. Rather than being able to make simple decisions about whether or not a sequence variant is the causative allele, sophisticated statistical analyses will generally be necessary.

Several approaches are being used to search for disease-associated sequence variants. Instead of relying on collecting large pedigrees, one type of study analyses pairs of affected siblings looking for evidence of excess allele sharing at loci that might be linked to disease. Two siblings can have none, one, or two alleles in common at a particular polymorphic genetic marker, with an expected frequency distribution of 25 per cent, 50 per cent, and 25 per cent respectively. If a particular allele were associated with disease risk, then both siblings with the disease would be expected to carry this allele more often than by chance alone. In large sets of sibling pairs, this would be manifested by statistically significant deviation from the expected distribution of shared alleles.

It is believed that many DNA variants that predispose to common disease traits will be ancient, derived from a common ancestor at some time in the past. More recent examples of shared ancestral alleles causing disease have come from studies of monogenic disorders. For instance in the case of porphyria variegata, two apparently unrelated Afrikaaners with the disease are highly likely to have the same underlying mutation and be descended from a pair of Dutch immigrants who arrived in South Africa in the early seventeenth century. In more outbred populations, the sources of such founder mutations will generally lie in the more distant past. Individuals who carry disease gene variants with common origins will also carry the same alleles at nearby polymorphic markers that were present on the ancestral chromosome on which the mutation first occurred. The distance over which these shared polymorphisms will be observed depends, in part, on the age of the variant. The greater the number of generations that have passed, the greater the chance that meiotic chromosomal recombination will have occurred, creating new patterns of associated markers.

The concept of association between disease status and specific alleles at nearby polymorphic markers in apparently unrelated subjects is known as linkage disequilibrium. Although there remains some uncertainty relating to these estimates, it is likely that, in groups that have not gone through recent population bottlenecks, linkage disequilibrium will extend for regions estimated to be 25 to 30 kb, on average. This means that in order to perform genome-wide scans for disease-associated alleles, roughly 200 000 to 300 000 polymorphic markers will need to be examined (in contrast to the 300–400 markers that are often used to search for monogenic disease genes by linkage analysis in affected families).

Even though characterizing common variants in human DNA was not part of the original goals of the Human Genome Project, that goal has been added. The need for very large numbers of polymorphic markers has meant that a great deal of energy is now being devoted to identifying and cataloguing human genetic variation, and within the next 1 to 2 years it is expected that more than 1 000 000 variants will be found. On average, any two randomly selected chromosomes differ from each other in about one base pair in every thousand. Several efforts have been undertaken to establish databases containing very large numbers of single nucleotide polymorphisms (SNPs). Some methods identify SNPs that are randomly scattered throughout the genome while other approaches target SNP discovery to particular parts of the genome, such as protein coding regions. All of these SNPs may be useful for genetic mapping and linkage disequilibrium studies, though some would argue that focusing on coding region SNPs (cSNPs) will give a higher yield of sequence changes that are more likely to be associated with alterations in gene function. A number of issues relating to the technical and statistical challenges associated with SNP analysis, high through-put genotyping, and the mathematics underlying linkage disequilibrium analysis remain to be resolved. These issues are the subjects of active research.

Changes to medical practice in the age of genomics

Given the great effort needed to define allelic variants contributing to complex disease, it is reasonable to ask whether such a large investment of resources is warranted. To be able to answer in the affirmative, it is necessary to demonstrate that benefits will accrue to everyday medical practice and patient health. Understanding genetic factors that contribute to disease could help establish a more rational basis for many aspects of patient care by providing deep insights into molecular pathogenesis and through improved molecular diagnostic tools that allow individually tailored preventive and/or therapeutic regimens.

Better understanding of molecular mechanisms of disease

Despite the extraordinary advances in our understanding of the functions of cells and organ systems in states of health and disease, it is somewhat humbling that fewer than 5000 human genes have been functionally characterized—many in only a cursory fashion. Clearly, it is difficult to provide full descriptions of the ways in which disease processes perturb cellular function in the absence of a comprehensive catalogue of genes that are either affected by these disease processes or are involved in the response to disease. The Human Genome Project will provide such a catalogue, giving a complete description of the DNA and protein sequences of all of these genes.

The genome project is providing important tools that will help researchers discover functions of novel proteins. About half of the new genes identified by large-scale DNA sequencing bear some resemblance to other genes that have previously been studied, either in humans or in model organisms. Sequence similarity to previously characterized genes can provide important clues to protein function. In addition to comparisons of primary DNA and protein sequences, computerized approaches to predicting three-dimensional protein structures are becoming increasingly feasible, and may also allow generation of testable predictions about gene function.

Of course, many novel genes do not have any (or enough) similarity to known genes for useful predictions to be made, and direct experimental investigation may be required to determine their function. To deal with the large number of new genes that are being identified by the Human Genome Project, an innovative set of methods has emerged, known collectively as functional genomics, which explore the roles of many genes in parallel. Functional genomics experiments will obtain a great deal of information about patterns of gene expression, protein interactions, and metabolic pathways.

Diagnostic tests

One advance in genomics that is already finding its way into clinical practice is the use of diagnostic tests based on DNA sequence changes. Such tests can be used for several purposes. As with more conventional tests, genetic tests can confirm a specific diagnosis or contribute to the evaluation of problematic differential diagnoses. Presymptomatic diagnostic testing can also be performed in subjects without disease. However, even in the case of high penetrance mutations, such as those found in Huntington's disease, it may be difficult to predict the time at which clinical signs or symptoms will develop. In some cases, measures may be available that will prevent or ameliorate the onset or course of illness. Examples of such conditions include haemochromatosis, in which regular venesection can prevent the sequelae of iron overload, or hereditary non-polyposis colon cancer, in which case colonoscopic removal of premalignant lesions can help to prevent the development of cancer. In the absence of such effective interventions, especially careful consideration must be given to the circumstances in which predictive genetic testing is performed. Some patients find presymptomatic genetic diagnosis helpful because it gives them the opportunity to make long-term plans which include the likelihood (or not) of developing illness. Others prefer 'not to know' and would rather forego testing unless an intervention is available. In many centres, teams that include qualified genetic counsellors are in place to ensure that patients receive appropriate education and non-directive counselling before and after making decisions about whether to undergo testing for serious illnesses.

An important way that DNA testing can differ from conventional diagnostic investigations is in its implications for relatives of the tested proband. A positive (or negative) result may allow one to infer the genotype of individuals other than the proband. For example if a subject had a paternal grandparent with Huntington's disease and was found to have an expanded glutamine-encoding triplet repeat in the Huntington gene herself, then it is virtually certain that her father is also at risk of developing Huntington's disease. Even without his consent to be tested, the father's genotype could be inferred with a high degree of certainty. Such situations can

complicate issues relating to informed consent. It is not unusual for several members of a family with a history of a genetic disorder to present for presymptomatic evaluation. Genetic counsellors can help to explore complex issues related to the needs and expectations of different family members while respecting individual autonomy.

To date, most conditions being investigated with DNA-based tests are relatively uncommon, single-gene disorders, and the tests are usually carried out in specialized centres with considerable experience in their execution and interpretation. As the technologies needed to perform these tests become more widely available and the sequence changes being evaluated more frequent, attention will be needed to ensure that procedures relating to DNA-based diagnostics continue to be carefully executed. Such concerns include issues related to quality control, skills in the interpretation of complex test results, provision of adequate genetic counselling, and other matters such as control of record-keeping systems to ensure genetic privacy. As the use of predictive genetic testing becomes more widespread, the associated obligations will fall to an increasingly diverse range of health-care professionals and it is important that they receive adequate training in this area.

Pharmacogenomics

Although not generally thought of in the same way as DNA variations that predispose to complex disease, germ-line sequence changes can also be important determinants of response to pharmacological treatment. The study of DNA sequence polymorphisms affecting metabolism of, and response to, drugs has become known as pharmacogenomics. It has long been recognized that individuals metabolize pharmacological agents at different rates. A substantial proportion of this variation can be due to genetic effects. A classic example of this phenomenon was described well before the modern era of molecular genetics and genomics, with the division of the population into slow and rapid acetylator groups. Approximately 60 per cent of Caucasians, but only 5 to 10 per cent of Asians, show reduced rates of *N*-acetylation and elimination of a number of drugs, including isoniazid, hydralazine, and caffeine, compared with the remainder of the population who exhibit comparatively rapid *N*-acetylation. Although initially characterized by standard biochemical tests, the genotypic bases for these metabolic differences in drug metabolism are now known to be several polymorphic variants in the gene *N*-acetyl-transferase-2 or *NAT2*.

In addition to showing different rates of metabolism and clearance of drugs from the system, subjects are often found to show variable therapeutic responses to these agents. In some cases, this may be due to subtle changes in the structure of drug targets between subjects. For instance several studies have found that schizophrenics with the amino acid histidine at residue 452 of the 5-HT $_{2A}$ serotonin receptor have a poorer antipsychotic response to clozapine than do patients with a tyrosine at this position. In diseases such as essential hypertension that are treated symptomatically, rather than with specific measures, some patients respond better to certain interventions than to others. If the genetic factors that are responsible for different forms of hypertension could be identified, then it might prove possible to tailor treatment regimens specifically directed towards generating responses in the physiological pathways that are perturbed in particular patients. For instance we might be able to predict that a particular patient would be more likely to respond to the antihypertensive effects of b-blockers than ACE inhibitors. We may then be able to use relatively inexpensive, one-time DNA tests, to determine which drugs can be used safely and efficaciously in patients with chronic illness. Such approaches promise to be more effective than the trial and error processes that are now often needed to discover optimal drug treatment regimens.

Development of therapeutic agents

In parallel with the government-funded human genome project, much effort has been expended in the private sector to identify and sequence novel genes. A large part of the motivation behind these efforts is the expectation that new therapeutic agents and targets will be discovered. The distinction between therapeutic agents and targets is one that has important implications for the development strategies that must be used. Therapeutic agents developed from genomic sequence information are usually secreted proteins or hormones that can be made for direct *in vivo* administration. A prime example is the use of recombinant erythropoietin for the treatment of anaemia in chronic renal failure. On the other hand, therapeutic targets can potentially be any proteins that are involved in a disease process or a compensatory response to disease. Once such a target is identified, further effort is needed to develop agents (usually small molecules) that can affect function of the target. Although it remains difficult to predict the three dimensional structure of novel proteins based on amino acid sequence alone, improvements in computational analysis, X-ray crystallography, and nuclear magnetic resonance studies are making it possible to produce increasingly sophisticated models of the active sites of target proteins. Using principles of rational drug design, it will become a more straightforward and less expensive matter to develop new drugs for particular diseases than using the trial-and-error methods that were relied upon in the past. The use of genomic sequence information in identifying targets for drug development is also advancing rapidly in the field of new antibiotic discovery. The genomes of several important microbial pathogens including *H. influenzae*, *S. aureus*, *M. tuberculosis*, and *T. pallidum* have been completely sequenced. Newly identified proteins that are important for bacterial replication or virulence but that do not have close relatives in mammalian cells may provide excellent drug or vaccine targets.

Another exciting possibility that arises from an understanding of the genomic sequence of disease genes is the possibility that the expression of particular proteins might be able to be modulated by developing small molecules that interact specifically with promoter and enhancer elements of the genes coding for these proteins. Considerable advances have been made in both the understanding of factors that influence sequence-specific DNA-binding and the ways in which gene expression are controlled, and it is likely that this knowledge will be used to help develop drugs that can up- or down-regulate gene expression in useful ways for disease treatment or prevention.

Gene therapy

Though they have been slow to come to fruition, the concepts underlying somatic cell gene therapy predate the inception of the Human Genome Project. Initial plans to induce expression of normal proteins in patients with rare genetic defects, such as adenosine deaminase deficiency, have been broadened to include a range of more common diseases such as cancer, atherosclerosis, and AIDS. Achieving clinical benefits from gene therapy has proven to be problematic because of technical difficulties preventing sufficiently high-level expression of recombinant proteins in appropriate tissues. Substantial effort is now being directed to the development of viral and other vectors that will allow more effective delivery of introduced genes to the desired sites. The recent death of a patient involved in a clinical trial using adenovirus as the delivery system has underlined the risks of triggering an overwhelming immune response with such vectors. A full discussion of gene therapies is beyond the scope of this chapter; but clearly, as our understanding of the human and other genomes increases, the range of conditions that are amenable to treatment by various forms of somatic genetic manipulation will also increase.

Public policy implications

With the advances in medical and scientific knowledge that come from the recent, rapid developments in the field of genomics come a host of issues that require consideration by the broader community. Some of these issues have been faced before in other contexts, and some are peculiar to the fields of genetics and genomics.

Genetic privacy

In addition to issues relating to control of access to medical records that are shared with other forms of medical data, there are several privacy concerns that are particular to genetic and DNA sequence-based information. Most people would agree that individuals should have a high degree of control over who has access to information concerning the makeup of their genome. Because relatives share genetic markers, it may be possible to determine a person's genetic constitution without the knowledge or consent of that individual. This can lead to a conflict between one person's desire to be aware of their genetic makeup with another's desire that other people not know his or hers. Deciding 'ownership' of genetic information is not a simple matter and may be difficult to resolve if disputes occur within families. Such disputes also have the potential inadvertently to reveal previously known or unknown instances of non-paternity, with unsettling consequences for those involved.

The use of combinations of polymorphic markers for the purpose of uniquely identifying subjects has found increasing application in non-medical settings, especially forensic science. Privacy concerns have been raised about databases established by government agencies that record DNA profiles of large numbers of individuals. It is important that appropriate safeguards be instituted to ensure that databases that might be used for socially justifiable purposes, such as criminal forensic analysis or identification of deceased military personnel, are not abused.

Health insurance and employment discrimination

Especially in countries such as the United States that do not have universal health insurance programmes, there is concern that people will be denied access to affordable health care because of their genetic makeup. There have been occasional, but well documented, examples of denial of health insurance because patients are at risk of developing genetic illnesses. As the factors underlying genetic contributions to complex disease are elucidated, eventually the majority of people could

be found to be at an increased risk of developing one serious medical condition or another. It would be most unfortunate if the possession of an allele that predisposes to cancer of the colon put someone in the position of losing health insurance coverage, rather than the institution of a regular colonoscopy programme that could prevent them from dying of cancer. Legislative solutions are needed to ensure that the availability of health care does not depend upon genetic constitution.

Fears have also been expressed that potential employers could screen applicants for the presence of 'undesirable' genetic traits and use the results of these tests to decide who should be offered positions. Should someone be denied the chance to become an air traffic controller because they were more likely than average to abuse alcohol? Should a person possessing a neurotransmitter receptor variant that makes them prone to depression be prevented from carrying a firearm as a police officer? In general, however, predictions based on such correlations will be rather weak—and thus the strong consensus is that employment decisions should be made solely on the basis of ability to do the job. Here again, legislative protections are needed to prevent discriminatory practices from becoming widespread.

Regulation of novel genetic tests and therapies

Novel diagnostic tests and therapies arising from rapid progress in the field of genomics can pose substantial challenges to statutory regulatory agencies such as the Food and Drug Administration in the United States or the Committee on the Safety of Medicines in the United Kingdom. These authorities may need to acquire quickly the expertise to assess innovative modalities and ensure that they are used appropriately and safely. Many of these issues will best be considered within frameworks similar to those that have been used in other areas of medical practice. The establishment of the Secretary's Advisory Committee on Genetic Testing (SACGT) signals an era of increased scrutiny of predictive genetic tests in the United States.

Many aspects of the traditional drug approval process are being adapted to assess the safety of somatic cell gene therapy and recombinant biotherapeutic agents. Nevertheless, certain of these processes will need to be refined to cope with changes brought about by advances in genomics. Authorities responsible for the governance of genetic testing procedures must address issues relating to informed consent and genetic privacy. Techniques with potential to alter the makeup of the human genome, such as germ-line gene therapy and cloning, have already raised thorny moral and philosophical questions.

Education of medical and lay communities

Several factors appear to be responsible for the relatively little attention given by many medical schools to clinical aspects of genetics and genomics. In part because of the relatively large cost of establishing comprehensive research programmes in this area there, has been a tendency for cutting-edge genomics research to be carried out in only a few centres. This, together with the fact that most of the physicians and scientists with experience in genomics tend to be relatively young, has meant that few medical school curricula have received input from senior academics who appreciate the importance of the subject. Clinical genetics has often been taught under the auspices of paediatrics, and has focused on rare diseases. Given the cross-disciplinary nature of the topics that need to be covered to understand genomic aspects of human disease (including clinical medicine, pathology, biochemistry, epidemiology, and biostatistics) and its application to many common diseases of adult onset, it may be that the study of genomics should be considered within a broader context.

With the extensive amount of coverage that has been given to recent advances in genetics and genomics by the print and electronic media, it is not unusual for patients to ask their doctors questions on topics about which the physicians have little information. To meet their patients' needs, and to allow them to be more proactive in addressing preventive medicine and other health concerns, it is important that established medical practitioners and physicians in training have access to continuing medical education resources informing them about genomic aspects of human disease. In the United States, the recent formation of the National Coalition for Health Care Professional Education in Genetics (by the National Human Genome Research Institute, American Medical Association, and American Nursing Association) promises to fuel this effort. Several databases and other resources are already widely accessible on the world wide web. As a useful starting point, the National Center for Biotechnology Information site (*http://www.ncbi.nlm.nih.gov/*) provides access to On-line Mendelian Inheritance in Man (OMIM), a comprehensive collection of information related to inherited disease phenotypes. Other highly useful sites include GeneClinics (*http://www.geneclinics.org/*), which provides authoritative reviews on the diagnosis and management of specific genetic disorders, and GeneTests (*http://www.genetests.org/*), which provides a guide to availability of genetic testing services.

At the same time, members of the lay public will benefit from an increased understanding of how genetic factors might be important in managing their health. Challenges that have to be met in this arena include the sometimes technically complex types of information that need to be understood, as well as the probabilistic nature of many of the clinically relevant associations between genotype and phenotype under consideration. While many patients will take the initiative to try to learn about medical aspects of genomics for themselves, most will benefit from education from a number of sources, especially their physicians and other health-care providers.

Conclusion

It is not an exaggeration to assert that the field of genomics has revolutionized biomedical research in recent years. The process of transition from the laboratory to the clinic is still in its early phases but holds great promise. Insights into the pathological processes underlying many illnesses will expand the range of diagnostic and therapeutic options available to the clinician. Perhaps the most exciting near-term advance that the fruits of the Human Genome Project offers is the opportunity to develop personalized risk profiles based on the particular patterns of DNA sequence variation that individual patients exhibit. These profiles will give us many chances to ameliorate the effects of disease by instituting appropriate preventive health measures and tailoring treatment regimens to specific patients based on their genotypes.

We must be cautious, however, not to overestimate the power of genomics to explain clinical outcomes and other biological phenomena. A comprehensive understanding of the importance of genetic factors in human disease should only serve to highlight the significance of environmental factors as modifiers of genetically coded molecular processes. The effects of genes or DNA sequence variation should not be treated in too deterministic a manner but only regarded as one facet contributing to our understanding of the processes involved in human health and disease.

Further reading

Baxevanis AD, Ouellette BFF, eds (1998). *Bioinformatics: A practical guide to the analysis of genes and proteins*. John Wiley and Sons, New York.

Collins FS (1997). Sequencing the human genome. *Hospital Practice* **15**, 35–43.

Collins FS (1999). Shattuck lecture—medical and societal consequences of the Human Genome Project. *New England Journal of Medicine* **34**, 28–37.

Collins FS, Guyer MS, Chakravarti A (1997). Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1.

Evans WE, Relling MV (1999). Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–91.

Holtzman NA, Murphy PD, Watson MS, Barr PA (1997). Predictive genetic testing: from basic research to clinical practice. *Science* **278**, 602–5.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

McCarthy JJ, Hilfiker R (2000). The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nature Biotechnology* **18**, 505–8.

Miklos GL, Rubin GM (1996). The role of the genome project in determining gene function: insights from model organisms. *Cell* **86**, 521–9.

Risch N, Merikangas K (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–7.

Venter JC *et al*. (2001). The sequence of the human genome. *Science* **291**, 1304–51.

4.2 The human genome sequence

Sydney Brenner

The modern period in biological research began in 1953 when JD Watson and FHC Crick discovered the double helical structure of DNA. Within the very short period of time of about a decade, the new science of molecular biology uncovered the basic mechanisms of information transfer from genes to proteins in living cells. The nucleotide sequence of a gene was shown to be colinear with the amino acid sequence of the protein that it specified, the genetic code was found to be a triplet code and the correspondences of the three-base codons to each of the 20 amino acids was established. Although there were chemical methods for determining the amino acid sequences of proteins, the structure of the gene was accessible only through mutational changes as recorded by phenotypes. Such mutations defined genes, and these could be mapped by recombination. Research in molecular genetics proceeded most rapidly with bacteria and their viruses which could be handled easily in the laboratory. A beginning was also made to study more complex systems, such as *Drosophila* which had been long established as a laboratory organism for experimental genetics, and *Caenohabditis elegans*, a small, free-living nematode worm.

In the mid 1970s there were two revolutionary, technical innovations which changed the entire course of genetics. The first was a method of cloning and propagating fragments of DNA in bacteria and yeasts and the second was the invention of methods for sequencing DNA. Thus the genome of any organism could be obtained as a library of fragments and the sequences of these fragments could be determined. In principle, therefore, the complete sequence of the genome could be obtained, and since it was possible to clone and sequence cDNA copies of messenger RNA, the expressed genes could be identified and the amino acid sequences of their proteins inferred from the nucleotide sequences. cDNA characterization became important when it was quickly discovered, by the new techniques, that the genes of higher organism were interrupted by intervening sequences called introns which were removed by splicing leaving the coding exons to form a coherent sequence.

For some time the only complete genomes that were sequenced were small, of the order of 100 kb. In 1985, when it was suggested that the complete sequences of the human genome might be obtained, it was realized that not only would there have to be considerable technical improvements, but also the project would have to be on a large scale and require international co-operation. The technical improvements were the automation of many of the laborious steps of the sequencing process, and the availability of sequencing machines and their progressive enhancement in throughput. Larger genomes were tackled and an early accomplishment was the 14-Mb sequence of yeast. This was followed by the sequences of *C. elegans* and *Drosophila*, each of around 100 Mb. In 2001, two groups announced that they had more or less completed the first draft of the human genome sequence, and no doubt this will be subjected to much improvement in future years.

The human genome sequence was seen by many to provide new approaches to human biology and, in particular, to medicine. For example it was claimed that once all the genes were found, all the proteins specified by those genes would be known, and this would allow the uncovering of an enormous number of new targets for the development of new therapeutic agents. In the long run this may prove correct, but before we can use a protein as a drug target we have to know how it functions in the body, and whether any alteration of its activity will have the desired effects. Although the molecular function of a protein can often be specified by comparison of its sequence with proteins of known function, this is insufficient to decide how this activity is translated into a particular cellular process and how this is, in turn, integrated into the physiology of the whole human organism. Thus, for example, while we can readily identify domains with resemblances to proteases, we require additional information to decide what the proteolysis is doing; it may be involved in digestion, in gene regulation, in cell death, in blood coagulation, or the complement pathway. For any protein to be a target we must have something much more than the protein sequence, we must have a therapeutic hypothesis and this will be made possible only by the continuation of conventional biological and clinical research.

There can be no doubt, however, that knowing all or nearly all of the proteins made by cells will spur research on the biochemistry of cells. In particular, this has already had a profound effect on the development of knowledge about cell signalling pathways, DNA repair, protein traffic within cells, and ion channels, to name only a few. In addition, because many of these processes are common to other organisms such as *Drosophila*, the nematode, and even yeast and bacteria, we can draw on research in these organisms to illuminate function in mammalian cells. A whole area of research, called functional genomics by some, relies extensively on this comparative approach. Since mutations in genes are easily obtained in these model organisms, such mutant homologues can inform us about the cellular function of the gene, which can be carried over to the mammalian systems. The main experimental organism for functional analysis is the mouse, a mammalian model organism amenable to a special form of gene manipulation. A line of embryonic cells, ES cells, can be propagated in tissue culture, and when these are injected into a mouse blastula they become incorporated into the embryo and populate both the germline and somatic ells. In this way, transgenic mice can be constructed in which genes have been added or removed from the mouse genome. Mice in which genes have been deleted are called knockout transgenics and can reveal the contribution of the gene to the total phenotype. These methods were used to prove that prions cause the endogenous prion protein to adopt an incorrect form and so cause the neurological disease. Complete removal of the prion gene has no effect of its own but the animals become resistant to infection.

The most significant contribution made by the new genetics has been the identification of the genes involved in single gene mutations in humans. More than 1000 such monogenic, inherited disorders have been identified; some rare, others, such as cystic fibrosis, quite common. These provide the direct test of function and they throw light on the pathogenesis of disease and the underlying molecular causes. In certain areas, such as cholesterol metabolism, the analysis of the changes in certain monogenic diseases has revealed the connections between cholesterol and the ensuing cardiovascular pathology. In addition, even very rare monogenic diseases which resemble the much more common diseases, such as the cases of breast cancer or Alzheimer's disease, can lead us to understand the pathogenic pathways involved.

There are, however, several very common diseases which can be shown to have a genetic component but are not due to single gene mutations. Schizophrenia, diabetes, and Crohn's disease are common and are about 50 per cent correlated in identical twins. There is, therefore, a large environmental component and the fact that they are polygenic makes it difficult to discover the genes involved. However, the possibility that one might identify the genes with polymorphisms correlating with the disease state has led to the development of what is called predictive medicine or probabilistic medicine. We only have a few of these disease susceptibility marks, but the fact that genetic analysis may be predictive has raised many questions about the ethical, social, and legal consequences of genetic testing. There is very little established work in this field but already there are problems, mostly created by health insurance.

It is clear that in the rush to obtain the sequence, the understanding of the connections between genotype and phenotype has remained superficial. There is a tendency to talk about genes for homosexuality, alcoholism, criminality, and so on. The unravelling of the complex skein of connections between the genes and the final phenotype has only just begun and it will occupy biomedical scientists for the next few decades, at least. DNA sequencing is a unique technology; one can feed a machine with DNA derived from anything—plants, bacteria, humans—and the linear sequence of bases which is the essential information in the DNA can be extracted. There has been an explosion in the amount of data available and there will be much more to come. However, data are not knowledge and only knowledge can lead to understanding the meaning of the sequence and allow us to diagnose and treat human disease.

4.3Molecular cell biology

William M. F. Lee and Newman M. Yeilding

The molecular mechanisms underlying many cellular and biological functions have been unravelled in recent years and, with this, has come an understanding of the molecular basis of many pathological conditions. This progress promises to provide improved insights into disease pathogenesis, suggest novel opportunities for therapeutic intervention, and launch efforts to make these interventions a part of medical practice. In this chapter, we provide a synopsis of salient features of cell function at the molecular level. Important decisions and events that all cells undergo, such as proliferation, differentiation, and death, are then discussed in molecular terms. Finally, to provide a medical context for this information and illustrate the potential clinical relevance of the knowledge gained, we discuss emerging concepts of the molecular pathogenesis of two important human diseases, cancer and cardiomyopathy. Clearly, none of these sections can hope to be comprehensive given the mass of accumulated knowledge and relatively scant number of pages allotted to this chapter. Accordingly, emphasis has been placed on concepts and principles, with details and examples presented only to better illustrate these paradigms. A further reading list is provided for those seeking more in-depth discussions of the various topics. These are in the nature of review articles and books in order to limit the size of the bibliography. For those with greater interest in the science and who wish to know the details and read the primary articles, the bibliographies of the reviews and book chapters will provide the appropriate references.

Genetic information and information retrieval

No biological process is more universal or fundamental than the way cells store, access, and use the information that is needed for the all the processes involved in creating and sustaining life. Deoxyribonucleic acids (DNA) in the genes of the cell are the repository of this information. DNA is an unbranched, linear polymer constructed of deoxyribonucleotide monomers that are directionally, covalently linked by phosphate bonds between the 5' hydroxyl group of the deoxyribose sugar moiety of one deoxyribonucleotide and the 3' hydroxyl group of the deoxyribose moiety of the next deoxyribonucleotide (Fig. 1). Four different deoxyribonucleotides (adenine, A; guanine, G; cytosine, C; and thymine, T) comprise the genetic 'alphabet' with which genetic information is written. Cellular DNA usually exists in a double-stranded (duplex) helical configuration in which one strand is 'zippered' by hydrogen bonds to a second, fully complementary DNA strand. This bonding between complementary strands is established by the nucleotide on one strand interacting with the paired nucleotide on the opposite strand with which it forms the most stable hydrogen bonds: G bonds best with C, C with G, A with T, and T with A. The two strands in duplex DNA maintain full complementarity because, during DNA replication, the DNA polymerase enzymes responsible for synthesis use one strand as template while synthesizing the other through covalent addition of complementary nucleotides to the nascent strand. This type of replication is termed 'semiconservative' because each of the two new daughter DNA strands contains one newly synthesized strand and one previously synthesized strand derived from the original duplex.

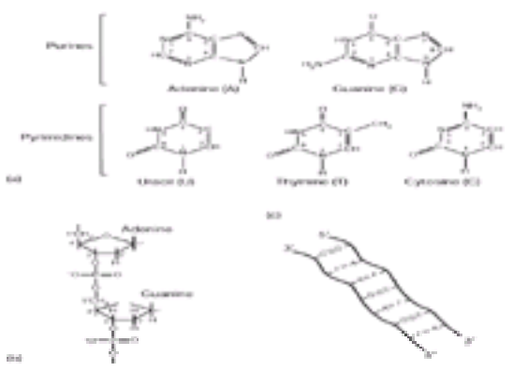

Fig. 1 Chemical structure of DNA and RNA. The chemical structures of the bases that comprise DNA and RNA are depicted in (a). In the nucleotides comprising DNA and RNA, the nitrogen atom at position 9 of purines or position 1 of pyrimidines is covalently bonded to the 1' carbon atom of a ribose or deoxyribose sugar. DNA and RNA polymers are directionally linked via a phophodiester bond of the 5' carbon of the sugar moiety of one ribonucleotide or deoxyribonucleotide and the 3' carbon of another (b). While RNA is most commonly single stranded, DNA primarily exists as a hydrogen-bonded antiparallel duplex between complementary DNA strands (c).

In eukaryotic cells, the library of genetic information is stored in chromosomes which consist of exceedingly long strands of DNA with associated proteins that help organize and compact the DNA. The most prominent proteins in chromatin, as chromosomal DNA with associated proteins is called, are the various types of histones. Certain histones form a core around which the DNA is more or less tightly wound, forming a structure called a 'nucleosome'. Nucleosomes are stacked, and stacks of nucleosomes are further organized to allow the very long DNA in chromatin both to fit in chromosomes and, importantly, to be accessible for information retrieval and replication. Arrayed on each chromosome are a multitude of genes, each of which contains the information necessary to synthesize a functional protein or RNA molecule. Normal somatic cells are diploid, meaning that they possess a pair of every non-sex chromosome (autosomes) and two sex chromosomes, which are either a pair of X chromosomes in cells from females or one X and one Y chromosome in cells from males. This means that two copies of every autosomal gene exist in each cell (one derived from each parent), two copies of X chromosome genes exist in female cells but only one copy in male cells, and one copy of Y chromosome genes exists in male cells and none in female cells. Human cells have 22 pairs of autosomes and two sex chromosomes, so they have 46XX chromosomes if they are female, and 46XY chromosomes if they are male.

An organism's genome contains a complete catalogue of the information required for its development, growth, and function. Proteins are the primary effector molecules responsible for carrying out the instructions written in the genome, and the properties, structure, and function of each protein is largely determined by its amino acid sequence. The principal information carried in the genome is the amino acid sequence of all the proteins produced by that organism. To create these proteins, genomic information is first accessed by transcribing DNA into ribonucleic acid (RNA) format, after which the information in RNA is translated into proteins. The process of gene transcription is carried out by complexes of proteins collectively termed transcription factors and involves the synthesis of RNA copies from DNA templates. RNA differs from DNA structurally in that the sugar moiety is ribose instead of deoxyribose, uracil (U) is incorporated where thymidine (T) is present in DNA, and RNA is usually single-stranded. The absence of a complementary strand means that nucleotides in RNA are free to form hydrogen bonds with other nucleotides in the same RNA (intrachain hydrogen bonding), a property that allows RNA to become highly folded and adopt secondary structures that are functionally important. The three major types of RNA present in cells are:

1. messenger or mRNA which encode the amino acid sequence of proteins;
2. ribosomal or rRNA which are components of ribosomes that translate mRNA into proteins; and

3. transfer or tRNA which actually decode the codons in mRNA transcripts into their respective amino acids during protein synthesis.

Information contained in mRNA for constructing proteins is deciphered through the process of translation. Proteins are polymers composed of 20 different amino acids covalently linked together by peptide bonds. Instructions for the sequence of amino acids in a protein are contained in sequential groups of three consecutive nucleotides, or codons, in mRNA (and the DNA from which it was transcribed), with each codon designating a specific amino acid. Thus, the sequence of specific codons defines the sequence of amino acids that are to be polymerized to create the encoded protein. Since each position in a given codon contains one of four possible nucleotides from which RNA and DNA are made, there are 4^3 or 64 possible nucleotide sequence combinations in a codon triplet. Of these 64 codons, 61 code for 20 different amino acids and 3 (TAA, TAG, and TGA) signal termination of translation. These numbers indicate redundancy in the genetic code, and, in fact, most amino acids are encoded by more than one codon.

The flow of information from DNA to RNA to protein just outlined is the fundamental basis upon which all life on earth is built. However, while this principle is relatively straightforward, the actual process of retrieving information from the genome is much more complicated. Consider that some proteins are needed in large quantities, others are needed in small quantities, and many are needed only some of the time. Yet, two copies of most genes exist in each diploid cell, regardless of the demand or lack of demand for the information they bear. To provide for this requirement that gene expression varies independently of gene copy number, cells have devised elaborate mechanisms for regulating the expression of genes and the activity of their products. That expression of genes is a highly regulated process is hinted at by the fact that the sequences that actually encode protein occupy only part of the DNA of most genes and only a small proportion of the human genome. Many non-coding regions of the genome have been shown to have important regulatory functions, such as regulating the transcription of genes.

The process of making mRNA transcripts suitable for translation into protein involves the initial synthesis of a primary mRNA copy of the gene and subsequent maturation of this primary transcript to ready it for translation (Fig. 2). Non-coding regions of genes near the start site of transcription or further upstream, termed promoters, initiators, and enhancers, regulate transcription. These elements act only on genes to which they are covalently linked (i.e. they act 'in *cis*') and function by interacting with proteins that regulate transcription, so-called transcription factors. These proteins non-covalently bind to specific DNA sequence motifs in gene regulatory elements and act 'in *trans*' to modulate expression of genes by assembling multiprotein complexes that start and control the process of transcription. The abundance and activity of many of these transcription factors are regulated by specific stimuli and environmental factors. Together, these *cis*- and *trans*-acting factors regulate mRNA production and determine which genes are expressed, when, and to what extent in different types of cells.

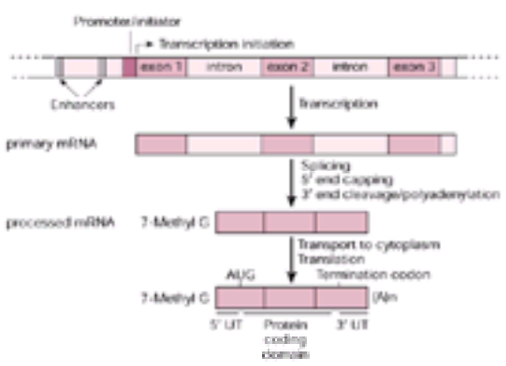

Fig. 2 Transcription and post-transcriptional processing of mRNA. Features of a gene transcriptional unit are shown. Following transcription and generation of a primary mRNA transcript, post-transcriptional processing converts this to a mature mRNA that is ready for export to the cytoplasm and translation. Landmarks in the mature mRNA relevant to translation are shown (UT, untranslated region).

After transcription, mRNAs must undergo additional processing to ready them for translation. The protein coding sequences of many eukaryotic genes are discontinuous and separated by intervening non-coding sequences (introns) which are transcribed and are part of the primary mRNA transcript. Introns must be removed by the process of splicing before the mRNA can be translated, resulting in a spliced mRNA that contains only exon sequences. While their origins are unclear, introns effectively divide protein coding regions of the gene into distinct segments. This gives rise to a form of modularity that can be envisaged as facilitating the modification of existing genes and proteins and the evolution of new ones if exons are moved, mixed, shuffled, duplicated, or deleted. Besides splicing out introns, mRNA maturation involves other events, such as the addition of a special nucleotide 'cap' (7-methyl G) at the proximal or 5' end, cleavage of the distal or 3' end and the addition there of an untemplated string of A residues (polyadenylation). These post-transcriptional modifications of mRNA protect it from degradation by nucleases and make it more translatable once it reaches the cytoplasm.

Mature mRNAs are transported from the nucleus to the cytoplasm where they are translated into the specified amino acid sequence. Ribosomes, which are macromolecular assemblages of multiple RNA and protein components, are the engines of translation and protein synthesis, while multiple species of tRNAs, each charged with its specific amino acid, 'read' the appropriate codons using complementary sequences in their structure. Mature mRNAs contain non-coding sequences preceding and following their coding sequences, and these have been shown to regulate both the efficiency of translation and longevity of the mRNA. During translation, ribosomes must exclude these 5' and 3' non-coding sequences from translation or, put another way, they must know where translation should begin and end. Current models of translation suggest that ribosomes attach to the mRNA via its 5' cap structure and 'scan' down the mRNA until they find the first AUG trinucleotide that is located in a nucleotide sequence context appropriate for translation initiation. This prevents false translation initiation by distinguishing the authentic translation initiation AUG from other AUG trinucleotides that are frequently present upstream. Translation is initiated at the authentic initiation AUG, which encodes methionine, then proceeds down the mRNA in the nucleotide triplet 'reading frame' established by the initiation AUG. Amino acids encoded by each subsequent trinucleotide are covalently added to the elongating polypeptide strand until an 'in-frame' termination codon (UAA, UGA, or UGA) is encountered, at which point translation stops. The reading frame is all important for correctly interpreting the message in mRNA nucleotide sequence and synthesizing the correct protein, since translation of the same sequence in either of the two alternate reading frames will produce totally different, incorrect proteins.

Mutations

Many systems have been built into cells to insure the fidelity of the genetic information transmitted from cell to progeny cell. If DNA errors arise, mechanisms exist to correct them, and if this is not possible, mechanisms exist to cull out the damaged cells. Despite these protective mechanisms, genetic changes and mutations do arise. This is not necessarily bad, for heritable changes drive diversity and the evolution of species. The prejudice against genetic mutations derives, justifiably, from a focus on the individual and the medical problems they may cause. In truth, however, most genetic changes probably have no detectable phenotypic consequences and give rise to so-called DNA sequence or genetic 'polymorphisms' that distinguish the genome of different individuals and have been used for a variety of purposes (e.g. forensics analysis, determining parentage, tracing inheritance of genes, etc.). Because they cause no problems, their occurrence goes unnoticed. On the other hand, mutations that cause disease soon come to medical attention and tend not to remain undetected, especially with the rapid pace of modern genetic analysis. These mutations can cause organ malfunction and developmental abnormalities, particularly those that arise in the germline and are inherited, because all cells in the organism are affected. More frequently, however, genetic mutations arise in somatic cells during a person's life. These may be spontaneous or induced by DNA damaging agents (e.g. gamma or UV radiation or chemical mutagens). Most somatic mutations probably have few if any consequences, because they affect only individual cells, and the vast majority are unaffected. However, when somatic mutations involve important regulatory genes, this may lead to aberrant cell behaviour which, if it involves loss of normal control over cell proliferation, can eventually cause disease, such as cancer.

The nature of genetic errors and their potential consequences are diverse. The complement of chromosomes in somatic cells may deviate from the normal diploid. Aneuploidy may involve either fewer (hypodiploid) or greater (hyperdiploid) numbers of chromosomes, and an aberrant complement of even a single chromosome can be pathogenic (e.g. trisomy 21 causes Down's syndrome). Segments of chromosomes rather than entire chromosomes may be amplified, which can give rise to visible cytogenetic abnormalities, such as chromosomes with 'homogeneously staining regions' (HSR). 'Double minutes' are amplified chromosomal segments that exist free in the nucleus, separated from their chromosome of origin. These cytogenetic abnormalities indicate the presence of genes that are present in many copies in the cell and whose expression may be deregulated. More common than gene amplification are gene deletions, which may or may not be detectable cytogenetically. Chromosomal translocations are yet another aberration and are frequently associated with lymphoid malignancies, probably because gene breakage and rejoining naturally occurs during certain stages of T and B cell development. Translocations juxtapose genetic elements that are normally separate and may cause aberrant regulation of gene expression or fusion of two different genes resulting in the production of a chimeric protein.

The genetic anomalies described so far are detectable cytogenetically, but most mutations are on a much smaller scale. Deletion of whole or portions of genes, which are usually not detectable cytogenetically, cause total loss of the protein product or creation of a truncated protein. Loss (or addition) on a much smaller scale (e.g. of one or two nucleotides), particularly if they occur in protein coding regions of genes, may produce equally devastating effects by shifting the reading frame during translation. Sequences downstream of the nucleotide loss or addition are normal but will be read by ribosomes in an alternate reading frame, resulting in incorporation of totally different amino acids from those intended. Needless to say, unless the frameshift mutation is near the end of the protein coding domain, functionality of the whole protein will be lost. Nucleotide substitutions or 'point' mutation can cause more or less profound disruptions of gene expression and protein function. At one extreme, substitutions may convert a codon for an amino acid into a termination codon ('nonsense' mutation) leading to premature termination of translation. Point mutations outside protein coding regions may also have profound effects if they occur in sequences important for exon–intron splicing or for binding of critical transcription factors. At the other extreme, nucleotide substitutions may be 'silent' when they occur in a nucleotide position that causes no change in the encoded amino acid (due to redundancy in the genetic code). Between these extremes, nucleotide substitutions may alter a codon so that it encodes a different amino acid. These 'missense' mutations may have little effect on the properties of the encoded protein (essentially becoming an amino acid sequence polymorphism) or may critically alter its function. The spectrum of genetic mutations briefly described here will assume greater relevance when the genetic aetiology of cancer is discussed.

Post-translational protein regulation

Cells are highly ordered structures and properly function only when their molecular assemblages and organelles function and interact appropriately. This organization requires that newly synthesized proteins are targeted to the areas of the cell where they are designed to function. For example haemoglobin must be retained within the cytosol of erythrocytes for optimal function, while insulin is secreted and carried via the vasculature to sites distant from the cell that produced it. The molecular beacons that designate the cellular location of proteins are contained in their amino acid sequence. Proteins that integrate in the cell membrane or that are destined for secretion contain a signal peptide, typically an amino-terminal peptide sequence containing one or more positively charged amino acids followed by six to 12 hydrophobic residues. The signal peptide targets proteins for translocation across the endoplasmic reticulum membrane into the lumen as they are being synthesized, after which the peptide is cleaved off, giving the mature protein a different amino terminus from the one initially produced by translation. Endoplasmic reticulum channels are lined by membrane that is contiguous with the plasma membrane, and their lumen is contiguous with the extracellular milieu, so that proteins translocated across the endoplasmic reticulum membrane have access to the cell surface and beyond. Transmembrane proteins are distinguished from secreted proteins by the presence of a transmembrane domain which is comprised of a linear stretch of about 22 hydrophobic amino acids that fixes the protein in the lipid bilayer of cell membranes as it transits into the endoplasmic reticulum during translation. Secreted proteins, in contrast, have no such domain and fail to arrest in the lipid bilayer during membrane transit. The orientation and topography of transmembrane proteins may vary and can be complex, such as in the case of proteins with multiple membrane spanning domains that weave in and out of the cell. Cytosolic and nuclear proteins do not have a signal peptide and are deposited during translation in the cytosol. Nuclear proteins are characterized by the presence of nuclear localization signals which consist of peptides with a preponderance of basic amino acids (lysine or arginines) that target the protein for transport to the nucleus.

Appropriate protein localization within cells and cellular subcompartments is critically important for function. During signal transduction, for example, signals received at the cell surface frequently need to reach the nucleus in order to transcriptionally reprogramme the cell to generate a response. The proper localization of proteins in these signal transduction pathways is crucial so that they are correctly positioned to receive and pass on these signal. As examples Ras and Src proteins function proximally in signal transduction pathways and are kept at the plasma membrane by covalently linked lipid moieties added after translation. Here, they are positioned to interact with upstream signalling molecules, such as transmembrane receptors. The functional importance of this localization is shown by the fact that loss of membrane attachment resulting in cytosolic localization renders Ras and Src inactive. Protein location may be used to regulate their biological activity. Transcription factors obviously have to be in the nucleus to modulate gene expression, and one such factor, NF_B, is kept inactive in the cytoplasm through binding to its inhibitor, I_B. When a signal for NF_B activation is received, I_B is phosphorylated, the complex dissociates, and free NF_B enters the nucleus where it can function. Keeping proteins out of the nuclear compartment until they are needed is a mechanism of regulation for several transcription factors.

Even after appropriate localization, many proteins must have their activity carefully regulated. This is frequently achieved by modifying proteins in one of a number of ways to alter their behaviour and function. Such post-translational modifications enable the cell to regulate protein activity without having to alter gene expression, resulting in faster, more responsive, and often reversible control. A variety of mechanisms have evolved to regulate protein activity. They may be proteolytically cleaved to generate fragments that have more or different activity from the uncleaved, parent protein. Many secreted enzymes are initially secreted in inactive, proenzyme form and require proteolytic modification for activation. Prime examples include blood coagulation factors which are widely distributed in inactive form and capable of being rapidly activated by proteolysis where and when clot formation is needed. Hormones provide additional examples, some being initially secreted in 'pre' or 'pro' form that require proteolytic processing to become active hormones. Protein function can be altered by covalently modifying the parent protein. For example phosphorylation (attachment of phosphate groups) of the hydroxyl groups of specific serine, threonine, and tyrosine amino acids modifies the activity of many proteins. The enzymes that catalyse such reactions are termed kinases; serine/threonine kinases catalyse phosphorylation of serine and threonine residues, and tyrosine kinases perform the same function for tyrosine residues. Phosphorylation confers a negative charge to these otherwise uncharged amino acids and can change the functional properties of the proteins in which they reside. For example many kinases are themselves substrates for phosphorylation which, in turn, activates their kinase activity. Phosphorylation can also confer upon proteins the ability to interact specifically with other proteins; phosphorylation of certain tyrosine residues allows proteins to bind proteins that have so-called SH2 domains, which are phosphotyrosine binding elements. Modification of proteins by phosphorylation can be reversed through dephosphorylation catalysed by protein phosphatases. The opposing activities of protein kinases and phosphatases and their antagonistic effects on substrate function clearly sets up a highly regulable system for controlling the activity of target proteins. Other examples of post-translational protein modifications with functional consequences include glycosylation, acetylation/deacetylation, ADP ribosylation, sulfation, and attachment of lipid groups (myristoylation, farnesylation, geranylation, etc.).

Signal transduction

Described above are the basic materials and processes important for the normal function of cells, but normal function of the body as a whole requires more highly co-ordinated, integrated, and orderly function of all of its constituent cells. Significant progress has been achieved in recent decades towards defining the molecular mechanisms involved in the complex and multifaceted inter- and intracellular signalling pathways through which cells and tissues communicate with each other. When functioning normally, these signalling networks allow us to perceive, integrate, and respond to local, environmental, and behavioural stimuli. Deregulated or dysfunctional signalling pathways, on the other hand, are pathogenically associated with many disease states. Cell-to-cell communication occurs through a variety of mechanisms, and the proximity of interacting cells dictates how such communications occur. Cells that are in direct contact with each other can establish direct lines of communication through plasma membrane junctions or pores that allow exchange of small molecules or the propagation of electrical signals to help co-ordinate metabolic, mechanical, or behavioural response. In the absence of direct contact, however, cell-to-cell communication occurs primarily via signalling molecules that are synthesized and released by a signalling cell and elicit a specific response in a target cell. Signalling may be paracrine, with the signalling and responding cells adjacent or nearby, as in the case of signalling at a neuromuscular junction where release of acetylcholine from a neurone elicits a contractile response from its target myocyte. Endocrine signals, on the other hand, use the circulatory system to target more distant cells and may elicit a tissue-specific or tissue non-specific response. For example release of follicle stimulating hormone from the anterior pituitary gland stimulates maturation of an ovarian follicle in a tissue-specific manner, while release of insulin from pancreatic islet cells elicits a more general physiological response from cells throughout the body resulting in their increased uptake of glucose. Cells can also respond to stimuli that they themselves elaborate through autocrine signalling pathways. For example interleukin-2 (IL-2) produced by activated T lymphocytes can stimulate their own IL-2 receptors, leading to autocrine stimulation of proliferation and clonal expansion. This section will address how cells perceive and respond to these signals.

Identification of many of the molecules mediating cell-to-cell communication has facilitated the use of molecular approaches to defining the mechanisms by which target cells respond to signals and modify their behaviour accordingly. Signals are perceived when signalling molecules or ligands bind their cognate receptors. The chemical characteristics of the signalling molecule dictate how and where this interaction occurs. Hydrophilic signalling molecules, for example peptide hormones and small charged molecules, such as epinephrine and histamine, that cannot freely diffuse across the plasma membrane bind receptors located on the cell surface. Some lipophilic signalling molecules, such as prostaglandins, also interact with cell-surface receptors. Others, for example steroid hormones, thyroxine, and retinoids, diffuse across the plasma membrane and interact with receptors located within the cell. Interaction of signalling molecules with their receptors initiates a cascade of chemical reactions that culminates in the modification of the cell's behaviour. These modifications may be rapid and transient, such as the contraction of a muscle cell, or they may be more prolonged, such as the metabolic response initiated by insulin. Some signals initiate a complete and irreversible reprogramming of the cell, as is seen when haematopoietic progenitor cells are induced to terminally differentiate into myeloid or erythroid cells.

The stimulus initiated when a ligand binds to its receptor is usually transduced by a variety of intermediary molecular mechanisms before it affects cell behaviour or produces its response. Conceptually, the most straightforward of these mechanisms is exemplified by the activities of receptors for most steroid hormones, thyroxine, and retinoids. Upon binding, the hormone–receptor complexes are induced to bind specific DNA sequences, called hormone response elements, and regulate the

transcription of associated genes (Fig. 3). Up- or down-modulation of the transcription of specific genes reprogrammes the cells toward a defined biological end. For example oestrogens stimulate endometrial growth in preparation for embryonic implantation and growth of the mammary ductal system in preparation for future lactation.

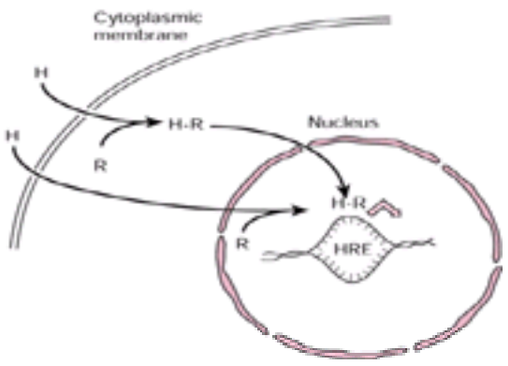

Fig. 3 Activation of gene transcription by hormones. After diffusing across the cell membrane, the lipophilic hormone (H) binds to its cytoplasmic or nuclear hormone receptor (R). This induces binding of the receptor to specific DNA sequences, called hormone response elements (HRE), which modulates transcription of associated genes and results in altered gene expression.

Signalling molecules that cannot diffuse across the plasma membrane are limited to interactions with cell-surface molecules. Therefore, the signals transmitted by hydrophilic ligands or by lipophilic ligands that interact with cell-surface receptors must be transduced by intracellular messengers. Cell-surface receptors transduce signals through three primary mechanisms: activation of ion transport channels; generation of small molecular intermediates ('second messengers') that modulate the activity of specific cellular proteins; or induction of covalent modifications of proteins, thereby modulating their enzymatic activity. The first mechanism, activation of ion transport channels, can alter the electrical potential across the cell membrane as seen when the nicotinic acetylcholine receptor is stimulated by acetylcholine resulting in activation of its Na^+/K^+ ligand-gated ion channel, membrane depolarization, and muscle contraction. This will not be discussed further, and attention will focus on the other two mechanisms by which cell-surface receptors transduce their signals.

Second messengers

Cell surface receptors modulate cell behaviour by activating intracellular signalling networks. Many of these signalling networks involve second messengers, intracellular signalling molecules whose concentration increases or decreases in response to activation of a cell-surface receptor. Second messengers transmit the signal by binding to and altering the activity of specific proteins that modulate cell behaviour. Ionized calcium, Ca^{2+}, serves as a second messenger, as do many small molecular weight intermediates, for example cyclic AMP (cAMP), cyclic GMP (cGMP), 1,2-diacylglycerol, and phosphatidylinositides, produced by enzymes activated directly or indirectly by cell surface receptors. An overview of the mechanisms by which second messengers transmit signals provides insight into the complexities of intracellular signalling networks.

Cyclic AMP

Cytosolic cAMP levels regulate many cellular metabolic responses. Increases in cAMP levels cause an increase in contraction rate in cardiac myocytes, increased gluconeogenesis in hepatocytes, increased thyroxine synthesis in thyroid cell, and many other cell type-specific metabolic responses. Levels of cytosolic cAMP are regulated by the enzyme adenylate cyclase, which is located on the inner surface of the plasma membrane and which, when activated, converts ATP to cAMP. Adenylate cyclase activity is regulated by members of a family of cell surface receptors, called G-protein-coupled receptors. These receptors are characterized by a primary amino acid sequence containing seven transmembrane domains that thread the receptor back and forth seven times through the plasma membrane (hence, they are sometimes called seven transmembrane-spanning proteins). These receptors regulate the activity of a family of GTP-dependent regulatory proteins (G-proteins) located at the plasma membrane, which, in turn, regulate adenylate cyclase activity (Fig. 4). Both stimulatory (G_s) and inhibitory (G_i) G-proteins exist which induce or inhibit the activity of adenylate cyclase, respectively, thus providing a mechanism for either increasing or decreasing cAMP levels.

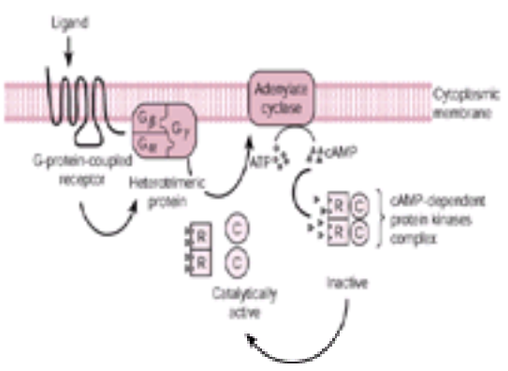

Fig. 4 Activation of cyclic AMP-dependent protein kinases by G-protein-coupled receptors. Ligand-induced activation of the G-protein-coupled receptor activates the heterotrimeric G-protein leading to dissociation of the Ga subunit from the Gb–Gg complex. Ga activates adenylate cyclase leading to the synthesis of cAMP from ATP. Two cAMP molecules bind to each of the two regulatory subunits of the heterotetrameric cyclic AMP-dependent protein kinase, causing dissociation and activation of the catalytic subunits.

Cytosolic levels of cAMP regulate the activity of a family of cAMP-dependent protein kinases, referred to as protein kinase A. These enzymes are composed of two regulatory (R) and two catalytic (C) subunits that form R_2C_2 tetramers. When cAMP binds to the R subunits, the C subunits dissociate and become catalytically active serine/threonine protein kinases. Phosphorylation of target proteins by protein kinase A modulates their enzymatic activity and leads to the downstream physiological consequences of receptor activation (Fig. 4). These effects of increased levels of cAMP persist until cAMP is hydrolysed to AMP by cAMP phosphodiesterase. The C and R subunits of protein kinase A then reassociate into a catalytically inactive R_2C_2 tetramer, and signalling is terminated.

The cascade of reactions initiated by the signalling network, from hormone to receptor to G-protein to adenylate cyclase to cAMP to protein kinase A activity to biological effect, permits amplification of the hormone signal, comparable to the amplification of chemical reactions initiated by activation of the blood coagulation cascade. One G-protein-coupled hormone receptor can activate as many as 100 G_s-proteins, and activated adenylate cyclase catalyses the production of many cAMP molecules. Such amplification can lead to dramatic augmentation of an initially small signal, and explains why epinephrine levels as low as 10^{-10} M can lead to the generation of as much as 10^{-6} M cAMP, an amplification of 10^4.

Adenylate cyclase is activated by a variety of hormone receptors, yet it leads to different metabolic responses depending on the cell type. For example activation of adenylate cyclase in adipocytes by epinephrine, ACTH, or glucagon results in decreased amino acid uptake and increases lipolysis. In contrast, its activation in hepatocytes increases amino acid uptake, as well as activating pathways that lead to increased gluconeogenesis, ketogenesis, and glycogenolysis. This observation leads to the question: How do an apparent limited number of signalling proteins generate such diverse cellular responses? While the mechanism of the diversity of biological responses is only partially understood, it is clear that cell context is a major determinant of the biological effect of second messengers. Different cell types express different repertoires of proteins, enzymes, and transcription factors. Second messengers can have diverse biological readouts depending on which factors they interact with, and furthermore, the same signal can have different effects depending on the strength with which it is delivered. Thus, subtle differences in gene

expression, signal strength, and amplification can have important biological and physiological consequences.

The clinical relevance of cAMP-mediated signalling pathways is highlighted by the mechanism by which *Vibrio cholerae* induces massive diarrhoea. These bacteria produce cholera toxin, a peptide that irreversibly activates G_s by covalently adding an ADP-ribose moiety to a specific arginine residue. This leads to continuous activation of adenylate cyclase and dramatic increases in cAMP levels in intestinal epithelial cells. Increased cAMP levels alter the activity of ion transport proteins and potentiate the flow of water through intestinal epithelial cells into the intestinal lumen leading to massive diarrhoea.

Other second messengers

The principles outlined above in which receptor activation leads to changes in second messenger concentrations which, in turn, alters the activity of specific effector proteins hold true for other second messengers, including cGMP, Ca^{2+}, and phosphatidylinositides. cGMP levels are regulated by soluble and membrane-bound forms of the enzyme guanylate cyclase. Guanylate cyclase is regulated by a broad spectrum of factors, including atriopeptins (e.g. atial natriuretic peptide, brain natriuretic peptide, and some enterotoxins) which regulate the membrane-bound form, and nitric oxide, nitroglycerine, nitroprusside, and sodium nitrite which diffuse across the plasma membrane to regulate the soluble form. cGMP levels and cGMP-regulated protein kinases play important roles in the regulation of vascular smooth muscle tone, endothelial cell permeability, cardiac contractility, platelet aggregation, intestinal motility and ion transport channel function, bone growth, and neuronal function.

The inositol-lipid signalling pathways are regulated by second messengers derived from two phospholipids located mainly in the inner layer of the plasma membrane lipid bilayer, PIP (phosphatidylinositol 4-phosphate) and PIP_2 (phosphatidylinositol 4,5-bisphosphate). This signalling pathway is activated when G-proteins, G_0 and G_q, activate the enzyme phospholipase C (PLC), which catalyses hydrolysis of PIP_2 to two second messengers, IP_3 (inositol 1,4,5-trisphosphate) and diacylglycerol (Fig. 5). IP_3 diffuses into the cytosol and stimulates Ca^{2+} release from the endoplasmic reticulum by activating IP_3-gated Ca^{2+}-release channels. Ca^{2+}, itself a second messenger, mediates many of its cellular effects through a calcium-dependent regulatory protein, calmodulin. The Ca^{2+}/calmodulin complex is an important regulator of ion transport channels, and structural elements of the cell, for example actin–myosin complexes in smooth muscle cells and microfilaments that mediate many processes such as cell motility, conformation changes, and mitosis. It also regulates a group of enzymes known as Ca^{2+}/calmodulin-dependent protein kinases (CaM-kinases) which affect glycogen breakdown and synthesis of catecholamine neurotransmitters.

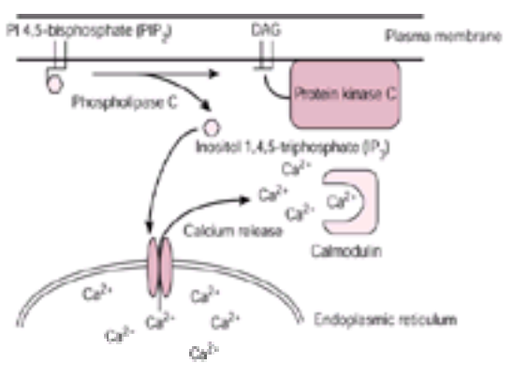

Fig. 5 Inositol–phospholipid signalling pathways. Activation of phospholipase C catalyses the hydrolysis of phosphatidylinositol 4,5-bisphosphate to inositol 1,4,5-trisphosphate (IP_3) and diacylglycerol (DAG). IP_3 induces calcium release from the endoplasmic reticulum which acts as a second messenger binding to calmodulin and modulating the activity of CaM kinases. DAG activates protein kinase C which transduces the signal by modulating the activity of a number of downstream enzymes.

In addition to its effects on CaM kinases, increases in intracellular Ca^{2+} mediated by IP_3 stimulates the migration of a cytosolic protein, protein kinase C, to the plasma membrane where it is activated by diacylglycerol, the other product of PIP_2 hydrolysis. Protein kinase C regulates the activity of a number of enzymes complementary to those regulated by CaM-kinases in mediating glycogen breakdown. It also regulates various transcription factors, such as NF_B, to alter the transcriptional programme of the cell. Thus second messengers transduce signals that are important in regulating numerous cell functions, including metabolism, structure, function, proliferation, and differentiation.

Signalling via covalent modifications of proteins

Many recently identified signalling pathways are initiated and conducted by proteins with latent enzyme or functional activity. Stimulation of these pathways by ligand binding to cell surface receptors induces the activity of these signalling proteins. These signalling pathways are initiated by one of four classes of receptors: (1) receptor tyrosine kinases; (2) tyrosine kinase-associated receptors; (3) receptor tyrosine phosphatases; and (4) receptor serine/threonine kinases. Once activated, these receptors induce either phosphorylation or dephosphorylation of target proteins on specific tyrosine, serine, or threonine residues, thereby altering target protein structure and/or function, and initiating a cascade of downstream events.

Receptor tyrosine kinases

Receptor tyrosine kinases (RTKs) make up a family of cell-surface proteins with several common structural features: an extracellular domain that interacts with a specific ligand; a transmembrane domain; and a cytoplasmic domain with regulated tyrosine kinase enzymatic activity. When ligand binds to the extracellular domain, the kinase activity of its cytoplasmic domain is induced through a process involving dimerization or multimerization of the receptor. The kinase domains of the dimerized receptors phosphorylates specific tyrosine residues in the cytoplasmic domain of its dimerization partner, a process termed 'autophosphorylation'. These phosphorylated tyrosine residues serve as high-affinity binding sites for intracellular proteins that transduce the receptor signal.

Once activated, the RTK signal is transduced through two classes of proteins that bind to the receptor: adapter proteins that have no intrinsic enzymatic or signalling properties and enzymes involved in activating downstream events. These two classes of proteins share SH2 domains which are known to mediate binding to phosphotyrosine residues. Through their SH2 domains, adapter proteins and other signalling proteins bind the newly phosphorylated tyrosine residues on the RTK cytoplasmic domain (Fig. 6). The specific residue that each protein binds is determined by the distinct sequence of amino acids surrounding the phosphotyrosine residue. For example Src binds the amino acid sequences phosphotyrosine–glutamate–glutamate–isoleucine through its SH2 domain. By binding adapter proteins and other enzymatically active proteins, the RTK transduces its signal into the cell.

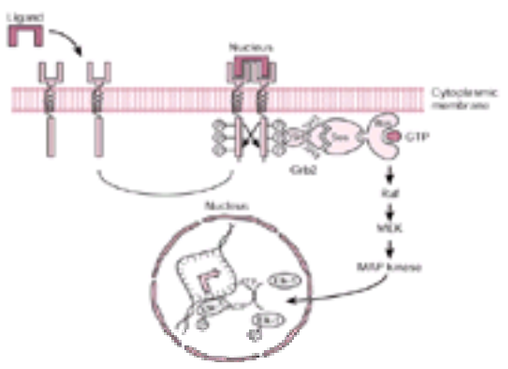

Fig. 6 Activation of intracellular signalling pathways by receptor tyrosine kinases. Ligand-induced dimerization of receptor tyrosine kinases leads to activation of latent kinase activity and *trans*-phosphorylation of the receptor's cytoplasmic domain. The adapter protein, Grb2, binds specific phosphotyrosine residues in the receptor via its SH2 domain, and it binds Sos via its SH3 domains. Sos functions as a guanine exchange factor, activating Ras by facilitating GDP–GTP exchange. The activated Ras signal is transduced in a linear fashion through Raf to MEK, to MAP kinase. MAP kinases translocate to the nucleus and modulate cell behaviour by regulating

the activity of transcription factors, for example Elk-1, and hence gene expression.

Intracellular signalling pathways activated by receptor tyrosine kinases

Many of the signals initiated by RTKs are transduced by Ras proteins, which are members of a family of low molecular weight proteins that bind guanine nucleotides and possess GTPase activity. They, in turn, regulate a cascade of serine/threonine kinases that control cell proliferation and differentiation. Ras proteins were initially identified through the role of mutant Ras proteins in carcinogenesis. Ras mutations are estimated to be present in about 30 per cent of human cancers and are among the most common molecular abnormalities found. Ras proteins cycle between an active 'on' state when GTP is bound and an inactive 'off' state when GDP is bound. Cycling between the on and off states is regulated by two classes of signalling molecules: GTPase-activating proteins (GAPs) which increase Ras hydrolysis of bound GTP to GDP, thereby inactivating Ras, and guanine-nucleotide exchange factors which facilitate dissociation of GDP from Ras (Fig. 7). Since the cytosolic concentration of GTP is approximately 10-fold higher than GDP, Ras will tend to bind GTP after dissociation from GDP, resulting in its activation. Many RTKs activate Ras through the actions of two cytosolic proteins, Grb2 and Sos. Grb2 is an adapter protein that contains an SH2 domain through which it bind the tyrosine-phosphorylated RTK, and two Src homology 3 (SH3) domains through which it binds Sos. Sos functions as a guanine-nucleotide exchange factor and activates Ras by facilitating GDP–GTP exchange. In many human cancers, Ras has been mutated so that it no longer hydrolyses GTP to GDP normally, resulting in a Ras protein that is continuously bound to GTP and a signalling pathway that remains turned on.

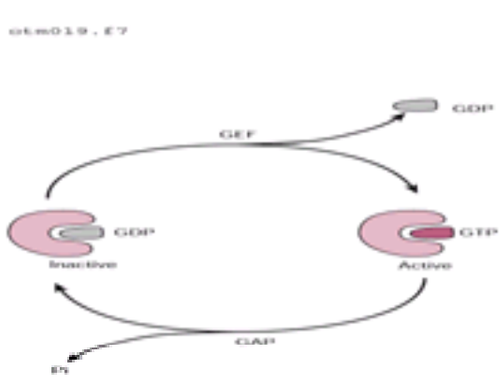

Fig. 7 Activation of Ras proteins. Ras proteins cycle between active and inactive forms. Ras becomes activated when guanine nucleotide exchange factor (GEF) facilitates its dissociation from GDP, enabling it to bind GTP, thus resulting in Ras activation. Ras binding to GTPase-activating proteins (GAP) induces hydrolysis of GTP to GDP resulting in Ras inactivation.

Many of the cellular changes induced by Ras are mediated by a family of proteins called mitogen-activated protein (MAP) kinases. These kinases are unusual in that they require phosphorylation of both threonine and tyrosine residues to stimulate their full activity, while most kinases catalyse phosphorylation of either tyrosine residues or serine and threonine residues, but not both. The dual function kinase that activates MAP kinase is called MAP-kinase-kinase (MKK, also known as MAPKK or MEK). MKK is activated by a serine/threonine kinase called Raf (also known as MAP-kinase-kinase-kinase or MKKK) which, in turn, is activated by Ras. Once activated by this linear cascade of reactions (Fig. 6), MAP kinases modulate cell behavior by phosphorylating other cellular proteins, including cytosolic proteins and other kinases. They also translocate to the nucleus where they phosphorylate and activate transcription factors, e.g. Elk-1 and serum response factor (SRF).

Transcription factors are final participants in afferent signal transduction pathways and initiators of cellular responses to these signals. In general, they are proteins that bind specific DNA sequences and modulate the expression of genes to which they bind. Most bind DNA as dimers, and different transcription factors use specific peptide motifs to dimerize with their partner(s), e.g. 'leucine zippers' and helix-loop-helix domains. Upon binding to DNA, they interact with the basal transcription machinery either directly or via intermediary proteins ('coactivators' and 'corepressors') to initiate, enhance, or inhibit transcription. Transcriptional gene regulation is highly complex, not only due to the multitude of transcription factors present in cells but also due to the ability of many factors to heterodimerize and form combinatorial pairs that have different DNA-binding, transactivation and/or regulatory properties. By transcriptionally reprogramming the cell, these factors regulate cell proliferation or differentiation, survival or apoptosis, and cell structure and function.

Organization of MAP kinase signalling pathways

The RTK–MKKK–MKK–MAP kinase cascade transduces many different signals from the cell surface to the nucleus. To date, 14 MKKKs, five MKKs, and 12 MAP kinases have been identified. Different signal transduction pathways are created by combining different components of each of these kinase families, thus generating diversity in cellular responses. However, some limitations exist in the different components that can be combined. For example MKKs show a fair amount of specificity for the MAP kinases that they activate and are generally coupled to a specific MAP kinase. The MKKs, MEK1 and MEK2, activate MAP kinases involved in cell growth and differentiation, for example ERK1 and ERK2, but they do not activate the MAP kinases involved in response to stress, for example JNK/SAPK kinases or p38 kinases. In contrast, MKKKs are quite promiscuous and can activate multiple MKKs. This promiscuity raises the question of how specificity in signal transduction pathways is generated. This mystery was recently solved by the discovery that scaffold proteins bring together specific components of the signalling cascade so that they are sequestered into a signalling unit. Scaffold proteins assure that when a MKKK is activated by a specific signal, only those MKKs and MAP kinases bound to the same scaffold protein are activated (Fig. 8). Thus scaffold proteins bring specificity to signalling pathways so that only the intended cellular responses are generated.

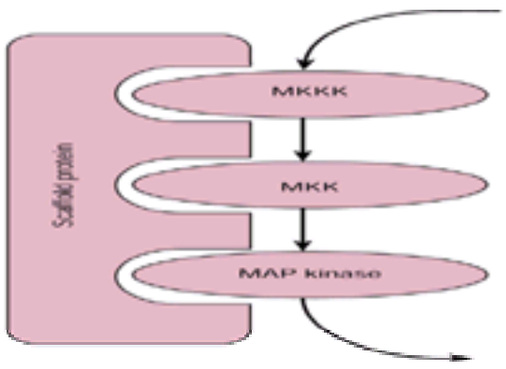

Fig. 8 Organization of MAP kinase signalling pathways by scaffold proteins. Scaffold proteins can bind multiple members of a signalling pathway, restricting their access to substrates and providing specificity in the signalling response.

Tyrosine kinase-associated receptors

While RTKs have latent enzymatic activity that is activated by ligand binding, receptors of the tyrosine kinase-associated receptor family lack intrinsic kinase activity, but associate via their cytoplasmic domains with tyrosine kinases located in the cytosol and/or the plasma membrane. Ligand binding induces activity of the associated kinase to transduce the receptor signal. This family includes antigen-specific receptors on T and B lymphocytes and receptors for many of the cytokines that regulate the proliferation and differentiation of haematopoietic cells. These non-catalytic receptors primarily associate with members of the Src family or the Janus family of cytosolic tyrosine kinases (JAK) to transduce their signals. Members of the Src family of tyrosine kinases play important roles in regulating the cell cycle,

regulating activation of immune effector cells induced by antigen and Fc receptors, and modulating osteoclast behaviour in bone remodelling.

The more recently described JAKs are involved in signalling pathways initiated by many cytokines, and mediate their effects by activating transcription factors from the STAT (signal transducers and activators of transcription) family. Binding of cytokines such as interferon a, b, or g or many interleukins to their cognate receptors activates the latent kinase activity of their associated JAKs which, in turn, phosphorylates specific members of the STAT family. Phosphorylation induces the STATs to dimerize and translocate to the nucleus where they activate transcription of specific cytokine-regulated genes.

Receptor tyrosine phosphatases

The currency of signal transduction initiated by RTKs and tyrosine kinase-associated receptors is phosphorylation of specific tyrosine residues. These signalling pathways illustrate the dramatic effect that an apparently simple modification, tyrosine phosphorylation, has on protein function. Receptor tyrosine phosphatases, in contrast, transduce signals by removing phosphate residues from, or dephosphorylating, specific proteins. The CD45 protein is a transmembrane protein expressed on the surface of leucocytes and plays an important role in lymphocyte activation. Upon activation, CD45 acts as a phosphatase to dephosphorylate target proteins. One such target is Lck, a member of the Src family of tyrosine kinases, whose kinase activity is induced when it is dephosphorylated.

Receptor serine/threonine kinases

The final class of receptors that transduce signals by inducing covalent modifications of target proteins is the serine/threonine kinase receptor family, which, upon activation, catalyses the phosphorylation of target proteins on specific serine and/or threonine residues. Receptors for the transforming growth factor-b (TGF-b) superfamily of ligands, including TGF-b proteins, activins, and bone morphogen proteins (BMP), contain serine/threonine kinase domains through which they modulate signalling pathways. These receptors activate members of the Smad family of proteins by inducing phosphorylation of target serine residues. Activated Smads form heteromeric complexes with other Smad family members, and translocate to the nucleus where they act as transcription factors.

Signalling networks

The events involved in transducing signals from cell surface receptors have been presented as simplified linear pathways leading to predictable cellular responses. However, biological systems are in actuality much more complex with numerous interconnections between the different signalling pathways. These interconnections enable a single signalling event to activate a network of signalling cascades and, thus, to orchestrate complex metabolic, structural, or functional cellular changes. For example activated RTKs typically phosphorylate several tyrosine residues in the RTK cytoplasmic domain which enables the docking of multiple proteins containing SH2 domains. Such proteins, in addition to Grb2, include the g isoform of phospholipase C (PLC), phosphatidylinositol-3 (PI-3) kinase, tyrosine phosphatases (e.g. Shp2 and Syp), RasGAP (a negative regulator of the Ras GTPase), Src family tyrosine kinases, and multiple adapters (including Shc, Grb7, Nck, etc.). Thus, in addition to activating the MKKK–MKK–MAP kinase pathway, RTKs can also activate numerous accessory signalling pathways (Fig. 9). Moreover, interconnections between signalling pathways exist distal to cell surface receptors. For example, protein kinase C can activate the Ras signalling pathway, and Ras can activate PI-3 kinase to activate the inositol–lipid signalling pathways. Thus complex interconnections exist between different signalling cascades that enable signalling proteins to build a multifaceted signalling network to marshal a cellular response.

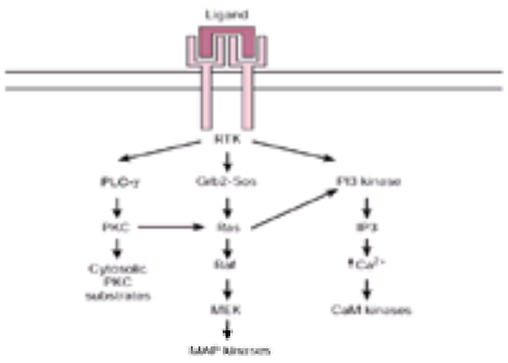

Fig. 9 Signalling networks. Signals from activated RTKs can be transduced via multiple signalling pathways, and numerous interconnections exist between different signalling pathways. Thus a single stimulus can orchestrate a complex biological response.

Fates of a cell: proliferation, differentiation, death

Somatic cells undergo one of three general fates: they (a) proliferate by mitotic cell division; (b) differentiate and acquire specialized functions; or (c) die and are eliminated from the body. Cell proliferation is necessary for growth of the organism and insures repletion of cells lost to terminal differentiation, death, or cell loss. In the immune system, lymphocyte proliferation serves the important function of amplifying responses to antigens. Differentiation provides the organism with cells that execute specific and specialized functions. Differentiation tends to be an incremental process, going through stages, but at the end may be 'terminal' so that further cell proliferation is precluded. Cell death as an active process, initiated by the cell itself, is apoptosis. Perhaps not obvious is the fact that this cell fate is as important physiologically as cell proliferation and differentiation. It allows tissue renewal and changes in cellular composition without undesirable or harmful cell accumulation. In the event of exposure to toxic agents, apoptosis eliminates the damaged cells, preventing them from being a burden or harmful to the organism. In complex multicellular organisms, when any of these three cellular processes becomes deregulated and unbalanced, the consequences are usually dire and result in either functional insufficiency or neoplasia. In recent years, some of their mechanisms and regulation have been defined at a molecular level.

Cell proliferation

The cell cycle

Somatic cells proliferate by mitosis, a process that produces two identical progeny from one parental cell. Mitotic cells pass through an ordered series of states collectively termed the 'cell cycle' (Fig. 10). This cycle has four sequential phases, labelled G_1, S, G_2, and M, which are defined biochemically, morphologically, and on the basis of cellular DNA content. S phase is the period of wholesale DNA synthesis during which the parental diploid cell with a '2N' complement of DNA replicates its entire genetic content and becomes a cell with 4N DNA content. M phase or mitosis is the period of nuclear and cell division during which the duplicated DNA complement of the 4N parental cell is divided equally between the two progeny cells which are consequently 2N. M phase is morphologically obvious as the period during which chromosomes condense into their familiar, microscopically visible forms, the nuclear envelope breaks down, the chromosomes segregate into two identical sets, the nuclear envelopes reforms (which completes nuclear division or 'karyokinesis'), and the two progeny cells separate (which completes cell division or 'cytokinesis'). G_1 and G_2 phases were originally conceived as 'gaps' between the distinctive M and S phases of the cell cycle. G_2 is the period between S and M, when cells have finished replicating their DNA, have 4N DNA, and are preparing to divide. G_1 is the period between M and S when cells are 2N, have finished one round of cell division, and have not yet initiated the next.

The durations of S, G_2, and M tend to be relatively constant, in contrast to that of G_1 which can be highly variable depending on the cell type and is subject to regulation by environmental factors, such as the availability of mitogens and nutrients. It is the period of cell growth, and a certain increase in mass may be required before the cell can enter the next S phase. When conditions are unsuitable for cell proliferation, they arrest in G_1, and those that are already in S, G_2, or M usually complete the round they have entered and arrest only when they reach G_1 again. A point in late G_1 called the 'restriction point' or 'R' has special significance and is the point past which cells become committed to enter S, even if mitogens are withdrawn. Cells may withdraw from the cell cycle and remain for prolonged periods in a metabolically active but non-proliferative state. These cells have 2N DNA content and are described as being in G_0. Terminally differentiated cells are examples of cells in G_0. However, other cells reversibly enter G_0 and may be induced to return to G_1 and begin cycling again under certain conditions (distinction between cells in G_0 and prolonged G_1, admittedly, may be difficult). Hepatocytes are in G_0 unless partial hepatectomy or hepatotoxic insults induce them to proliferate to reconstitute

functional liver mass. Resting, antigen-specific lymphocytes remain in G_0 until antigen and cytokine stimulation induces them to proliferate.

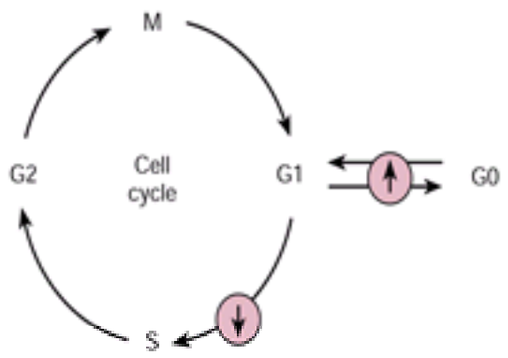

Fig. 10 Cell cycle. The sequential phases of the cell cycle, G_1, S, G_2, and M, are depicted, as well as the resting G_0 phase. Common regulatory points near the end of G_1 and between G_1 and G_0 are shown by circles with arrows within.

Adherence to the G_1–S–G_2–M sequence during normal progression through the cell cycle means that a cell must duplicate its DNA before dividing and that it must divide before duplicating its DNA again. This insures a normal genetic complement in the progeny cells and maintains genetic constancy. The dependence of later events in the cell cycle upon normal completion of earlier events is insured by 'checkpoint' control mechanisms that prevent a cell that has not successfully completed one phase of the cycle from entering the next. Checkpoint activity is seen after cell exposure to DNA-damaging agents, such as ionizing radiation, and is manifest as delayed cell entry into S and M by inducing temporary arrest in G_1 or G_2. This delay allows cells time either to repair its damaged DNA or, if the damage is irreparable, to execute a programme of self-destruction or apoptosis.

Cell cycle regulation: cyclins and cyclin-dependent kinases

Notable progress has been made in understanding the regulation of cell entry and progression through the cell cycle. In brief, cell cycling is regulated by serine/threonine kinases of the Cdk (cyclin-dependent kinase) family. As implied by their name, the catalytic activities of these kinases are dependent on associated regulatory proteins called cyclins. Cyclins were so named because levels of the first to be described were seen to fluctuate periodically with the cell cycle. Numerous Cdks and cyclins exist in the cell and form various combinatorial pairs with distinct activities. Control of cyclin/Cdk activity exists at many levels and occurs by the appearance and disappearance of the cyclins at specific phases of the cell cycle, by post-translational modification of Cdks, and by association with Cdk inhibitors. Cyclin/Cdks, in turn, regulate cell cycling by modulating the activity and behaviour of other proteins, such as transcription factors and structural proteins.

The function of cyclins and Cdks is best exemplified by their regulation of cell entry into M phase. Studies of mutant fission yeast called cdc2 (cell division cycle 2) that tended to arrest in G_1 or G_2 led to the cloning of a 34 kDa serine/threonine kinase ($p34^{cdc2}$) that is required for yeast to enter S or M phase. This protein is evolutionarily conserved, being present in a structurally and functionally similar form in humans. Another line of study showed that cytoplasmic extracts from mature frog eggs, when injected into immature oocytes, induced the oocytes to mature and undergo typical M phase changes. The 'maturation promoting factor' (MPF) activity in these extracts was found to reside in two proteins, frog $p34^{cdc2}$ and a B-type cyclin (Fig. 11). Cyclin B has no intrinsic enzymatic function and plays a regulatory role by associating with $p34^{cdc2}$ which then exhibits kinase/MPF activity. Cyclin B levels increase during S and G_2, and levels of the cyclin B/$p34^{cdc2}$ complex sufficient for the G_2/M transition are reached well before the onset of M. Mitosis is not prematurely triggered, however, because the complex accumulates in an inactive form. During S and G_2, the $p34^{cdc2}$ complexed with cyclin B accumulates as a multiply phosphorylated protein. In human $p34^{cdc2}$, phosphorylation of a specific threonine (thr161) stabilizes its association with cyclin B and is essential for activity. On the other hand, phosphorylation of another threonine (thr14) and a tyrosine (tyr15) in $p34^{cdc2}$ suppresses its kinase activity and keeps the cyclin B/$p34^{cdc2}$ complex inactive. Activation of this complex just prior to entry into M requires dephosphorylation of both thr14 and tyr15 which is accomplished by a dual-specificity phosphatase, Cdc25. The kinase and phosphatase that regulate $p34^{cdc2}$ activity and time cell entry into M are themselves regulated by phosphorylation (which inhibits the kinase responsible for tyrl5 phosphorylation and enhances the phosphatase function of Cdc25). Once activated, cyclin B/$p34^{cdc2}$ can phosphorylate Cdc25 and create a self-amplifying feedback loop that generates more oocyte MPF activity from a small initial amount of active MPF and the large pre-existing stock of inactive MPF. What starts this sequence of events by initially phosphorylating Cdc25 is unclear, although cyclin A/Cdks are candidates because they are active prior to cyclin B/$p34^{cdc2}$ activation and have MPF activity; also inhibition of cyclin A during S prevents entry into M.

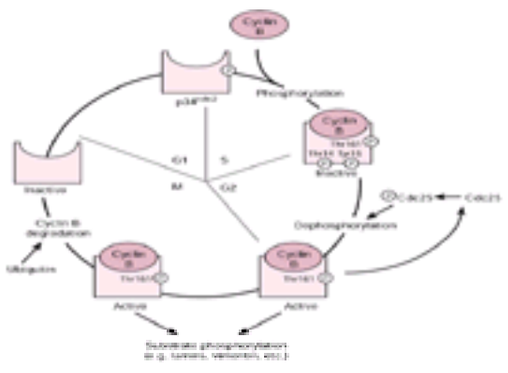

Fig. 11 Activation of mitosis. The G2–M transition activated by cyclin B/$p34^{cdc2}$ is depicted, as well as the phosphorylation and dephosphorylation events leading up to cyclin B/$p34^{cdc2}$ activation and cyclin B degradation during M phase.

As its name indicates, activated cyclin B/$p34^{cdc2}$ phosphorylates serine and threonine residues in cellular proteins. Discerning its physiological substrates is difficult, however, because there are many potential substrates and other cyclin/Cdk complexes and kinases are active at the same time. Candidates include the lamins and vimentin which are proteins important for the structural organization of cells. They are substrates for cyclin B/$p34^{cdc2}$ kinase activity *in vitro* and undergo M phase phosphorylation *in vivo*. Phosphorylation of lamins is important for nuclear lamina disassembly and envelope breakdown, and phosphorylation of vimentin may cause depolymerization of intermediate filaments in the cytoplasm. If these are physiological substrates, cyclin B/$p34^{cdc2}$ activity may initiate the structural reorganization that is part of mitosis. As M phase progresses, cyclin B/$p34^{cdc2}$ is inactivated by degradation of the cyclin B component. Mutant cyclin B that is resistant to proteolysis induces cell arrest in M, demonstrating that cyclin B degradation is important for cells to exit M.

Cyclins other than cyclin B and Cdks other than $p34^{cdc2}$ regulate other parts of the cell cycle. The behaviour and activity of these others resemble those of cyclin B and $p34^{cdc2}$, so that the activity of Cdks are regulated by the cyclins with which they pair, and the permitted partnerships determine where and how in the cell cycle the individual cyclin/Cdk complexes function. The portion of the cell cycle that has received particular attention is G_1 and the G_1/S boundary, because events here determine commitment to and rates of cell proliferation and have a bearing on neoplastic transformation of cells. Among cyclins, cyclin A and B are unlikely to be important in G_1 and the G_1/S boundary because they disappear during M and reappear only in S; cyclins D and E are better candidates from a timing standpoint, and D cyclins (there are three types, Dl, D2, and D3) may be especially significant because abnormalities in these have been implicated in the pathogenesis of parathyroid adenomas and certain B cell lymphomas. Cdks that associate with these cyclins are probably important in G_1/S regulation: Cdk4 and 6 associate only with cyclin D, and Cdk2 associates with cyclins A, D, and E. These cyclins and Cdks are thought to regulate the G_1/S transition through phosphorylation of Rb, the protein product of the retinoblastoma susceptibility gene (Fig. 12). In its hypophosphorylated state, Rb inhibits cell entry into S phase probably by binding to certain members of the

E2F family of transcription factors. However, during passage through G_1, Rb is phosphorylated on many serine and threonine residues, and hyperphosphorylated Rb releases E2F. E2F, in turn, activates transcription of genes needed for S phase activity and other aspects of cell proliferation (dihydrofolate reductase, thymidine kinase, myc, myb, etc.). Thus, Rb's ability to bind E2F is determined by its phosphorylation state and, in turn, regulates E2F transcriptional activity and cell cycling. Cyclin D/Cdk4 and 6 are among the kinases that can phosphorylate Rb with substrate specificity being conferred, at least in part, by the ability of cyclin D to bind Rb. Others cyclin/Cdk complexes may play important roles in cell cycle regulation as well. For example cyclin E associates primarily with Cdk2, and in G_1 cells is found in a quaternary complex with Rb-related p107 protein, E2F, and Cdk2. This complex disappears as cells enter S, just as a similar complex containing cyclin A instead of E makes its appearance. Cyclin E can also phosphorylate Rb and reverse the G_1 growth arrest induced by hypophosphorylated Rb. The appearance of cyclin A at the beginning of S, its decline in G_2 and M, and its presence in S phase in a quarternary complex with Cdk2, E2F, and p107 suggest that cyclin A plays a role in driving S phase events. Cyclin A in complex with $p32^{cdc2}$ may trigger the G_2/M transition by phosphorylating Cdc25 and initiating cyclin B/$p32^{cdc2}$ activation.

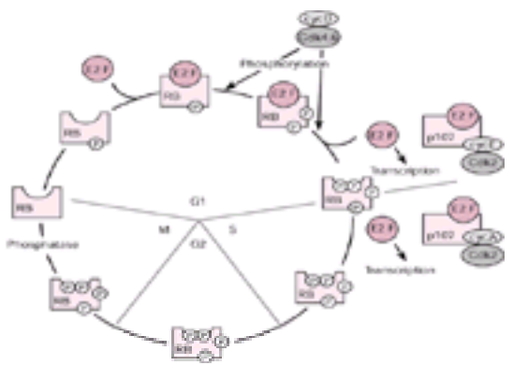

Fig. 12 Rb regulation of cell cycling. Cell cycle regulation by Rb and its association with E2F is depicted, as well as the regulation of the Rb/E2F association by cyclin D/Cdk4,6 phosphorylation of Rb on serine and threonine residues. Other potentially important kinase and transcription factor complexes involving E2F and Rb-like p107 are shown.

Cell cycle regulation: cyclin-dependent kinase inhibitors

Another layer of cell cycle regulation complexity has been revealed with the discovery of inhibitors of Cdk and cyclin/Cdk complexes. The first inhibitor identified and cloned was p21 (Waf1, Cip1, Sdi1) which binds several different cyclin/Cdk complexes and has been classified as a 'universal' inhibitor. While this and other properties of p21 (such as its ability to inhibit the process of DNA replication) account for its ability to induce cell cycle arrest, the regulation of p21 expression sheds the most light on its cellular function. Its expression is transcriptionally induced by p53 which is a transcription factor activated following DNA damage. This suggests that p21 is responsible for halting cell proliferation after DNA damage and allowing the cell time for damage assessment and repair. Other Cdk inhibitors similar to p21 exist. These include p27 (Kipl) which, like p21, bind and inhibit multiple cyclin/Cdk complexes. Upregulation of p27 is thought to mediate the growth arrest of cells under crowded conditions (contact-inhibition of cell growth) or after treatment with TGF-b, a growth arrest cytokine. Other Cdk inhibitors are more specific and inhibit specific Cdks. For example, p16 (INK4a, MTS1, Cdk4I) and p15 (INK4b, MTS2) are Cdk inhibitors that bind Cdk4 and Cdk6 exclusively, and binding inhibits their association with cyclin D and kinase activity. Since Rb is an important target of cyclin D/Cdk4,6 kinase activity, p15 and p16 may inhibit cell proliferation by preventing Rb phosphorylation by cyclin D/Cdk4,6. The observation that p16 overexpression inhibits proliferation of cells expressing Rb but not of cells devoid of Rb supports this idea.

Pathophysiological relevance

Evidence from a variety of spontaneously arising and experimental cancers indicate that Rb/E2F and their regulation by cyclin/Cdk play a central role in cell cycle regulation. In many types of human cancer cells, both copies of the *Rb* gene are disrupted, normal Rb protein is not made, and Rb regulation of E2F activity does not occur. The prototypic tumour in which Rb is pathogenically involved in this manner is retinoblastoma, but Rb abnormalities occur in other, more common cancers as well. Inactivation of Rb regulatory activity can also be achieved by means that do not alter the gene. In tumours induced by certain DNA tumour viruses, oncogenic proteins produced by the virus bind to and functionally inactivate Rb. Large T antigen (the transforming protein of SV40 virus), E1A (one of two transforming proteins of adenovirus), and E7 (one of two transforming proteins of human papilloma virus) proteins all bind Rb in its hypophosphorylated form and prevent it from inhibiting E2F activity. These examples of disruption of Rb cell cycle regulation by different viral oncoproteins provide compelling evidence that transforming viruses 'recognize' the importance of this pathway for maintaining normal cell behaviour and the need to disrupt it if they are to induce neoplasia.

Cell differentiation

Proliferation may provide organisms with the cells needed for growth, replacement, and repair, but differentiation endows these cells with the specialized characteristics and functions that make them useful. Cell differentiation to ultimate form and functionality does not occur in a single step but, rather, is an incremental process. In many tissues, cells, such as haematopoietic or intestinal epithelial cells, go through discernible stages of progressive lineage commitment leading up to full differentiation. In some differentiation lineages, the earliest cells are so-called 'stem cells' which display few if any signs of lineage commitment but have the potential for proliferation and self-renewal. Following cell division, progeny of stem cells undergo one of two fates—they either remain as stem cells or begin to commit to differentiate. The former path maintains the size of this critical pool of cells, while the latter supplies the organism's need for differentiated cells. Once stem cell progeny commit to differentiate, a process that may be progressive, incremental, and span several cell generations, they begin to express proteins that make them recognizable as 'blast cells' of their particular lineage. These remain largely undifferentiated but are fully committed and restricted to particular paths of differentiation. Though blast cells proliferate, they and their progeny inevitably acquire more differentiated characteristics and become fully differentiated cells of their type, that is they lack self-renewal potential and have limited or singular differentiation potential.

This model of progressive cell commitment and differentiation is well illustrated by haematopoietic differentiation. The elusive pluripotent haematopoietic stem cell is believed to divide infrequently, but when it does, it self-renews and gives rise to lymphoid progenitor cells and/or myeloid progenitor cells. These progenitors have the potential to develop into one of several lineages but, as their names imply, do not retain the full potential of the haematopoietic stem cell. Further along and perhaps some cell generations later, lymphoid and myeloid progenitor cells give rise to more committed progeny which are unipotential, that is they can differentiate only along one specific lineage, and show discernible features of differentiation. For example lymphoid progenitors may become T or B lymphocyte precursor cells (which are identifiable by evidence of T-cell receptor or imunoglobulin gene rearrangement), while myeloid progenitors give rise to erythroid, megakaryocytic, granulocytic, or monocytic precursors (which are identifiable as erythroblasts, megakaryoblasts, myeloblasts, monoblasts, etc. by special stains of bone marrow or *in vitro* colony forming assays). Even after reaching this stage, several days and a few cell divisions elapse before their progeny appear in the bloodstream as terminally differentiated blood cells. This multistep process of haematopoiesis clearly involves many fateful decisions about self-renewal, lineage commitment, restriction of potential, and differentiation. Through it, however, the stem cell pool is preserved as a resource for the life of the organism, and mitotic amplification of committed and differentiating cells supplies the body's huge requirement for the end products—an estimated 10^{10} to 10^{11} erythrocytes and granulocytes are normally produced every day.

The molecular switches that determine commitment, lineage-specification, and differentiation are incompletely understood. What is not in doubt is the involvement of transcriptional reprogramming of cells during this process, evidenced by the fact that cells committed to differentiate express a different complement of genes from their uncommitted counterparts. A stochastic model proposes that the stem cell decision to self-renew versus commit is based on probability. What determines this probability is unknown but, based on the likelihood that it involves transcriptional reprogramming, processes involved may include 'opening up' of critical regions of the genome so that they are permissive for gene transcription and binding of a requisite combination of transcription factors to their cognate sites in regulatory genes. When all these conditions are met, the cell may move to the next phase of differentiation (e.g. commit) but, otherwise, will remain as it was (e.g. self-renew). An alternative model is an inductive model of differentiation which suggests that environmental signals, perhaps originating from neighbouring cells and/or the cell stroma, induce stem cells to commit. However the decision is made, embarking on commitment and differentiation involves expression of lineage-specific genes. Among these genes are those encoding receptors for growth and survival factors pertinent to the development of that lineage. These may be crucial for the further progress of differentiation, because they provide the ability to receive instructions that promote the further proliferation, survival, and differentiation of cells of that lineage. For example erythropoietin signalling through its receptor promotes the survival, proliferation, and differentiation of cells of erythrocytic lineage, so that

production of the erythropoietin receptor is an essential component of erythrocyte lineage-specific gene expression. The signals generated by these lineage-specific factors often provide an autocrine feedback loop by inducing specific transcription factors which activate *cis*-regulatory sequences and expression of additional genes that contribute to the survival and differentiation of cells committed to these, but not other, lineages.

Many of our current insights into the molecular basis and transcriptional regulation of cell differentiation have been gained from the study of cell culture systems and gene knockout mice to examine myogenic differentiation. *In vitro* studies of myogenesis indicated that certain mesenchymal cells without detectable features of myocyte differentiation could become myoblasts and differentiate into myocytes following treatment with agents (e.g. 5-azacytidine) that can derepress expression of certain silent cellular genes. Transfer of genomic DNA from these 5-azacytidine-induced myoblasts conferred myogenic differentiation on the parental cells, indicating that the myogenic phenotype was heritable. Identification of mRNAs expressed in myoblasts but not in the parental cells led to the cloning of the *MyoD* gene which was subsequently shown to be sufficient for myogenic conversion of the parental cells. MyoD is a basic–helix–loop–helix transcription factor that activates expression of muscle-specific genes by binding to consensus DNA sites present in enhancer elements of muscle-specific genes. It is a member of a family of myogenic transcriptional regulators, including Myogenin and Myf-5, that can induce development of the myogenic phenotype in certain undifferentiated mesenchymal cells. MyoD binds its DNA sites preferentially as a heterodimer with another basic–helix–loop–helix transcription factor, E2A. The ubiquitous presence of E2A suggested that myogenic differentiation and transcriptional regulation of muscle specific genes are regulated by the abundance of MyoD, but this was contradicted by the presence of MyoD in undifferentiated myoblasts. his led to the discovery of the *Id* (inhibitor of differentiation) gene whose protein resembles MyoD in having a helix–loop–helix domain that allows it to dimerize with MyoD or E2A, but differs from MyoD in that it lacks a basic region for binding DNA. Heterodimers containing Id are unable to bind DNA, and Id antagonizes myogenic differentiation by inhibiting the ability of MyoD to activate expression of muscle-specific genes. Expression of Id decreases in cells undergoing myogenic differentiation and cell cycle withdrawal, allowing MyoD to function unhindered in these cells. Neat though this model may be, myogenic differentiation is more complex, because other transcriptional regulators, such as Myogenin and Myf5, can also induce muscle differentiation programs *in vitro.* Furthermore, the finding that mice lacking either MyoD or Myf5 (generated by knocking out both germline copies of these genes) had normal skeletal muscle development, whereas those lacking both MyoD and Myf5 had no signs of skeletal muscle development indicated that MyoD and Myf5 either function redundantly or act in separate myogenic cell populations that can compensate for the other's absence. In mice lacking Myogenin, skeletal muscle differentiation is blocked, even though MyoD and Myf5 are expressed and myoblasts are present. Thus, Myogenin is needed to activate myocyte differentiation following myogenic commitment that seems to depend on MyoD and Myf5 activity.

Apoptosis (see also Chapter 4.6)

Apoptosis is the process of cell death, also called 'programmed cell death', in which the mechanism of cell killing is instituted from within the cell itself. Apoptotic cells undergo fairly stereotypic changes characterized morphologically by early compaction of nuclear chromatin and condensation into clumps at the nuclear periphery. Nuclear and cellular outlines become mildly convoluted, the nucleus fragments, and the dying cell sheds or breaks up into membrane-enveloped 'apoptotic bodies'. These bodies are taken up by adjacent cells or nearby phagocytic cells, and apoptosis typically engenders little or no inflammatory response *in vivo.* Biochemically, the genomic DNA of apoptotic cells undergo strand breaks due to cleavage between nucleosome loops. Since each nucleosome loop is about 180 base pairs in length, this fragmentation leads to a characteristic 'laddering' of DNA fragments in size increments of 180 base pairs on gel electrophoresis. The presence of so many free genomic DNA ends in apoptotic cells allows them to be detected sensitively by DNA end-labelling techniques.

Cell death by necrosis is different and generally occurs when cells are exposed to severe physical, thermal, or other injury imposed from without. In necrosis, nuclear clumping may also be evident, but the most obvious changes involve marked swelling of the cell and organelles, such as mitochondria, and eventual internal disintegration of the nucleus and other structures. Genomic DNA of cells undergoing necrosis is degraded without regard to nucleosomal organization and, upon electrophoresis, appears smeared (indicating random cleavage) rather than laddered. Necrosis tends to be accompanied by inflammatory responses *in vivo*. It is important to point out that when pathologists use the term 'necrosis' to describe areas of cell death in tissues and organs, these terms are histopathology descriptors and are not generally meant to convey death mechanism, which may be apoptosis, necrosis, or a combination of the two. Indeed, apoptosis, rather than necrosis, is the most common mechanism by which cells die in the body, for example during tissue remodelling, after immunological attack, following cytotoxic chemotherapy or radiation therapy, and in pathological states such as congestive heart failure.

Cells can be induced to undergo apoptosis by a variety of factors and stimuli. Death programmes may be initiated following exposure of cells to toxic insults. The insults may be metabolic, such as severe hypoxia, acidosis, or a combination of the two stemming from ischaemia. They may be genotoxic, such as exposure to radiation that damages DNA beyond the point of repair. Chemotherapy-induced cell death is also attributed to apoptosis. Apoptosis can also be induced by factors unrelated to toxins and insults. Many cells thrive only in the presence of specific survival signals, and absence of these signals leads to apoptosis. These signals may come in the form of soluble factors, such as cytokines and interleukins, and 'permission' for cell survival may work hand in hand with cell differentiation mechanisms to establish specific populations of differentiated cells. For example erythropoietin induces development of late erythroid precursor cells but also promotes the survival of these precursors. Survival signals may also come from interaction of cell surface receptors, such as integrins, with specific components of the extracellular matrix in the cell's environment. Presumably, this insures that cells survive when they are in their 'correct' location or environment and not otherwise. In the last few years, specific receptor–ligand systems have been identified that function specifically to receive or stimulate cell death signals, and these provide another way that apoptosis can be induced. Among the best studied of these is the Fas/Fas ligand system. Fas is a member of the tumour necrosis factor (TNF) receptor 1 family of cell surface receptors which exist as trimers. Present only on certain cells, its interaction with crosslinking anti-Fas antibodies or Fas ligand (FasL; also a cell surface protein) induces apoptosis. Fas is conditionally expressed on lymphocytes and Fas/FasL interaction is known to be important for deletion of autoreactive lymphocytes (to prevent autoimmunity), following activation by antigen (to control immune responses and limit the number of antigen-reactive lymphocytes following antigen clearance), and in immunological sanctuary sites. The physiological importance of this system for maintaining lymphocyte and immune homeostasis is shown by the development of lymphoproliferative disease and autoimmunity in mice deficient in either Fas *(lpr* mice) or FasL *(gld* mice). Other members of this family of ligands and receptors, such as the prototypic TNF and TNF receptor 1, also generate apoptosis signals.

Cell killing during apoptosis is accomplished by members of the caspase family of cysteine proteases with specificity for aspartic acid residues in their substrates (Fig. 13). Upon activation of apoptosis, caspases cleave critical cellular proteins causing irreversible cell damage and death. Caspases exist in cells in inactive single-chain proenzyme form, and activation involves limited proteolytic cleavage of the proenzyme into a large and a small subunit, both of which are needed for catalytic activity. In some caspases, this proteolysis also removes an N-terminal regulatory domain, called the prodomain. Two large and two small caspase subunits assemble into a tetramer which form an active complex that possesses full proteolytic activity and executes cell death programmes. Caspases are activated in a hierarchical order, with caspases 8 and 9 activated first and cleaving procaspases 3, 6, and 7 to activate them so that they, in turn, can cleave and inactivate target cellular proteins (e.g. poly(ADP-ribose) polymerase or PARP, DNA-dependent protein kinase, actin, etc.). Thus the mechanism of caspase 8 and 9 activation provides the link between apoptosis stimulation and apoptosis execution. In the case of apoptosis signalling initiated by Fas ligation, a cytoplasmic domain of the receptor, termed the 'death domain', binds a similar domain in an adapter protein known as FADD. FADD also contains a 'death effector domain' which binds the prodomain of caspase 8. Assembly of this multiprotein complex brings several procaspase 8 molecules into close proximity, which may allow its very low level of proteolytic activity to cleave and activate nearby procaspase 8 molecules (i.e. 'autoactivation'). This would generate fully active, soluble caspase 8 and the subsequent cascade of downstream events. Procaspase 9, on the other hand, interacts via a 'caspase recruitment domain' in its prodomain with a similar domain in a protein called Apaf1 (apoptotic protease-activating factor 1). Apaf1 also self-associates, but this and procaspase 9-Apaf1 binding occur only in the presence of cytochrome c, which is made available by release from damaged mitochondria. Thus, mitochondrial release of cytochrome c triggers clustering and autoactivation of procaspase 9.

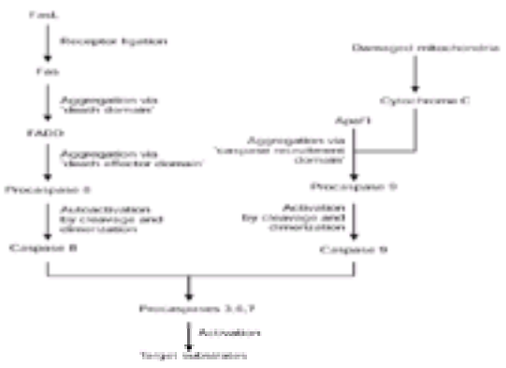

Fig. 13 Apoptosis mechanisms involving caspases. Apoptosis pathways leading from Fas death receptor activation and from cytochrome C release from damaged mitochondria to the hierarchical activation of different caspases are shown.

A system for activating efficient and irreversible self-destruction mechanisms cannot exist safely in cells without safeguards and regulators. A family of regulators exists in the Bcl-2 family of proteins. *Bcl-2* was identified as the gene juxtaposed to the immunoglobulin heavy chain gene in the t(14;18) translocation present in human follicular lymphomas. This type of lymphoma is characterized by a low growth fraction and indolent clinical behaviour early in its course. The accumulation of lymphoma cells in patients is attributable to their decreased rate of cell death due to decreased rates of apoptosis resulting from deregulated *bcl-2* expression. Bcl-2 is the prototypic member of a complex family of proteins, some of which are antiapoptotic (e.g. Bcl-2 and Bcl-X_L) and some of which are proapoptotic (e.g. Bax). Bcl-2 family proteins influence cellular responses to apoptosis signals by as yet poorly defined mechanisms but which may involve formation of solute channels in mitochondrial membrane and/or modulation of release of mitochondrial contents, including cytochrome c. Alternatively, certain antiapoptotic Bcl-2 proteins appear to bind Apaf 1, while certain proapoptotic Bcl-2 proteins dissociate this complex and allow Apaf 1 to activate caspase 9. Other proteins are also important in regulating apoptosis, including the tumour suppressor protein, p53. It is a regulator of gene transcription that becomes activated by DNA strand breaks and induces G_1 cell cycle arrest. Overexpression of p53 in cells has been associated with enhanced apoptosis, while inactivation of p53 seems to confer resistance to apoptosis. However, p53 only appears to affect apoptosis induced by certain stimuli and not others, for example p53 null thymocytes undergo less apoptosis following ionizing radiation but have unchanged apoptosis following exposure to glucocorticoid. These observations suggest that some pathways to apoptosis are p53-dependent, while others are p53-independent. This has great clinical relevance to cancer therapy, because many chemotherapy agents kill tumour cells by inducing apoptosis. p53 is believed to be the most frequently inactivated gene in human cancers (an estimated 50 per cent), and the p53 status of tumours may be an important determinant of chemotherapeutic response. Other regulators of apoptosis may also exist, but the Bcl-2 family of proteins and p53 are clearly clinically important regulators of apoptosis.

Molecular basis of cancer (see also Chapter 6.3)

Molecular studies have been applied to the study of cancer pathogenesis with notable success. The knowledge gained has led to a much better understanding of the mechanistic basis of cell transformation and promises to improve our ability to categorize, diagnose, and treat cancer. The problem of cancer pathogenesis may be viewed from the perspective of cell fate determination. As described above, all cells in the body undergo one of three general fates: proliferate to produce more cells, differentiate to carry out specialized functions, or die by the process of apoptosis and be eliminated. Organisms require an appropriate balance of cells undergoing each of these fates for normal function and homeostasis, and neoplasia arises when proliferation consistently and aberrantly exceeds apoptosis in a clonal population of cells, leading to their inappropriate and pathological accumulation. To the extent that cell differentiation and proliferation are opposing cell fates, the deregulation of cell proliferation associated with neoplasia is generally accompanied by a loss of differentiation. While this reductionist view greatly oversimplifies the pathogenesis of cancer which proceeds through multiple stages (e.g. dysplasia, adenoma, carcinoma *in situ,* invasive carcinoma, etc.) and involves complex interactions with the host environment (e.g. induction of angiogenesis, invasion, metastasis, evasion of immune response, etc.), it has directed research towards mechanisms that deregulate cell proliferation and apoptosis. This effort has led to the discovery of a rich cache of cellular genes that normally regulate cell behaviour and have the potential to contribute to neoplastic transformation when they go awry (Table 1). The underlying lesson of these discoveries is that cancer has a genetic basis. It results when cells lose or deregulate the function of genes responsible for their normal behaviour and it is heritable. While the transformed phenotype is transmitted from cell to progeny cell, it is not done without modification. Genetic mutability is a characteristic of many, perhaps all, cancer cells, and evidence suggests that selective forces operate on these cells *in vivo* to allow those with selective growth advantage to thrive and predominate, much as 'natural selection' acts at the organism and species levels. Cancer involves a cell evolutionary process in which the end product is the result of multiple genetic anomalies accumulated as once-normal cells progress to malignancy.

Oncogenes

Cellular genes potentially involved in neoplastic cell transformation have been identified by a number of approaches. Some were identified by their close relationship to transforming genes or oncogenes present in rapidly oncogenic retroviruses (e.g. *src* from Rous sarcoma virus, *myc* from avian myelocytomatosis virus, *ras* from Harvey and Kirsten sarcoma viruses, etc.). These retroviral oncogenes were originally derived from normal host genes present in infected cells by a process of gene capture ('retroviral transduction') that can occur during the retrovirus life cycle because they integrate into the host genome. This aetiology explains the close homology between retroviral oncogenes and their cellular precursors or 'proto-oncogenes' and makes the case that the latter have latent transforming potential. Cellular proto-oncogenes were also identified through study of genome integration sites of slowly transforming retroviruses. These retroviruses typically cause specific types of tumours in certain species (e.g. mouse mammary tumour virus or MMTV causes mammary tumours in mice, avian leukosis virus or ALV causes bursal or B cell lymphomas in chickens, etc.) but do not possess oncogenes. Instead, they transform by integrating into the host genome near specific cellular proto-oncogenes, deregulating their expression and activating their transforming potential (e.g. *int* in the case of MMTV mammary tumours, *myc* in the case of ALV bursal lymphomas, etc.). A different approach that identified cellular oncogenes employed experimental transfer of genomic DNA from cancer cells into phenotypically normal cells. Rare recipient cells acquired the transformed properties of the donor cells, and cloning of the donor genes in these recipients led to identification of the cellular oncogenes responsible. Finally, cancer cells commonly have chromosomal anomalies that are pathogenically significant because they are frequently or consistently associated with certain cancers (e.g. chromosomal translocations) or with aggressive clinical behaviour and advanced stage of disease (e.g. chromosomes with homogeneously staining regions, double minute, etc.). Cloning of the genes involved in these anomalies led to identification of additional oncogenes and proto-oncogenes.

The genes identified by the approaches outlined are transforming or have the potential to transform cells. One only has to consider that many proto-oncogenes produce proteins that participate in mitogenic signalling and cell cycling or are highly related to proteins that do to get a sense of their nature, that is proto-oncogenes are cellular genes whose products promote cell proliferation. More specifically, the types of genes that have been found to have oncogenic potential include genes that encode growth factors (e.g. platelet derived growth factor), growth factor receptors with tyrosine kinase activity (e.g. epidermal growth factor receptor or its homologue, Her2/neu), non-receptor tyrosine kinases (e.g. Src), Ras proteins (e.g. H-, K- and N-Ras), serine/threonine kinases (e.g. raf-1), transcription factors (e.g. Myc, Fos, Rel, etc.) and cyclins (e.g. cyclin D). While scores of candidate transforming genes have been revealed through the study of avian and rodent tumours and cells, only a few of these have actually been shown to contribute to human cancer pathogenesis *(ras, myc, Her2/neu* are examples of these). As indicated previously, genes regulating cell death may also play a role in oncogenesis, and an inhibitor of cell apoptosis (*bcl-2*) is involved in the pathogenesis of certain human lymphomas.

The importance of proto-oncogenes for maintaining normal control of cell proliferation and apoptosis indicates that cell transformation is not an expected result of their normal function. Indeed, their oncogenic potential is only unveiled through 'gain-of-function' or activating gene mutations that release the proto-oncogene or its protein from normal regulation of their activity (Table 2). These genetic alterations tend to upregulate or deregulate gene expression or to enhance or alter the function of their protein products. For example deregulation can result from chromosome translocations which involve breakage and aberrant rejoining of chromosomes to link genetic segments that are not normally juxtaposed. If translocation breakpoints occur in the coding region of genes, it may result in the synthesis of truncated or chimeric proteins with altered regulatory or functional properties. Examples include the chimeric proteins produced by the *bcr–abl* [t(9;22) or 'Philadelphia' chromosome] translocation characteristic of chronic myelogenous leukaemia and the PML–RARa [t(15;17)] translocation characteristic of acute promyelocytic leukaemia. If the translocation breakpoint is outside the coding region of genes, it may bring genes under the regulation of alien transcriptional control elements and alter gene expression. Examples are common in B-cell and T-cell lymphomas where immunoglobulin gene or T-cell receptor gene enhancers and promoters, respectively, become juxtaposed to proto-oncogenes, such as *myc* and *bcl-2.* Other genetic alterations, such as gene amplification, produce multiple copies of genes which may augment or deregulate their expression and activate oncogenic potential. Clinically relevant examples include amplification and overexpression of N *-myc* in neuroblastoma and of *Her2/neu* in breast and ovarian cancer, both of which are associated with a worse prognosis. Point mutations in coding regions of genes can result in amino acid substitutions that alter protein functional properties. The oncogenic potential of *ras* proto-oncogenes typically become activated by this mechanism. Finally, viruses whose genomes integrate into the cell's genome as part of their life cycle (e.g. retroviruses) can activate nearby genes by insertion of their foreign transcription regulatory elements. While retroviral insertional activation of proto-oncogenes is important in the pathogenesis of several types of animal tumours, it has yet to be shown to be a significant mechanism of human oncogene activation.

Tumour suppressor genes

The transforming genes or oncogenes just described generally promote cancer by favouring cell proliferation. Tumour suppressors are entirely different kinds of genes involved in cancer causation. They are cellular genes that normally function to restrain cell proliferation and contribute to cell transformation only following 'loss-of-function' or inactivating mutations. Since the activity of only one copy of a tumour suppressor gene is generally sufficient for function in a diploid cell, the function of both copies must be lost before they promote transformation. Thus, tumour suppressor genes act recessively at the cellular level and stand in marked contrast to oncogenes, which act dominantly (only one activated copy of an oncogene has to be present in a cell to promote transformation). Discovery and identification of tumour suppressor genes came much later than discovery of oncogenes because of the difficulties inherent in tracking down genes that produce effects only when they are absent or functionless. Almost all known tumour suppressor genes have been identified through laborious genetic mapping studies of cancer-prone kindreds. The telltale sign of tumour suppressor genes is loss of heterozygosity (LOH; loss of one of the two alleles normally present in diploid cells) at specific genetic loci in cancer cells that are not present in the normal cells of an individual with cancer. When the two copies or alleles of a gene in an individual's

cells can be distinguished molecularly (e.g. by restriction fragment length polymorphisms (RFLP) or microsatellite repeat length polymorphisms), LOH can be detected by molecular analysis. A given LOH found in tumour cells can be an incidental occurrence of little or no pathogenic significance, because these cells tend to be genetically unstable and may have lost alleles of many cellular genes as a consequence. If, on the other hand, LOH at a particular genetic locus is found in tumour cells from many different patients, the probability is high that the loss is pathogenically significant. With LOH, the remaining alleles are physically present but, when these have been studied, they have been found to contain inactivating mutations (e.g. small deletions or premature termination and frameshift mutations). Thus, non-random LOH at specific loci in the cancer cell genome suggests the functional loss of tumour suppressor genes at these sites and provides the starting point for more precise localization studies and eventual cloning.

Inherited cancers and cancer-prone kindreds have been instrumental in providing insights into the presence and identity of tumour suppressor loci and genes. *Rb* (retinoblastoma susceptibility gene), *APC* (adenomatous polyposis coli gene), *WT1* (Wilms' tumour gene), *p16*$^{INK4\varepsilon}$(melanoma susceptibility gene), and *BRCA1* and *BRCA2* (breast cancer susceptibility genes 1 and 2) are examples of tumour suppressor genes that were identified through study of families whose members had strong predispositions to developing the respective tumours. In each case, the inherited predisposition to cancer results from functional loss of one allele of the corresponding tumour suppressor gene in the germline (either through complete or partial deletion of the gene or the presence of inactivating mutations). Examination of the actual tumours in affected family members reveal that these have lost or inactivated the remaining, functional tumour suppressor allele and, thus, have undergone the obligate functional loss of both alleles for cancer pathogenesis. Thus, the high incidence of specific cancers in affected individuals in these kindreds is attributable to the pre-existing loss of one tumour suppressor allele, so that, as these individuals go through life, their cells have only to lose the remaining allele to embark on cancerous changes. Cells of normal individual, in contrast, need to lose both alleles before they embark on similar changes. While this obviously happens, attested by the fact that sporadic (non-familial) retinoblastomas consistently lose both copies of *Rb* and sporadic colon and breast cancers frequently lose both copies of *APC* and *BRCA2*, respectively, the likelihood must be substantially less given the much lower incidence of these tumours in the general population than in affected families. Finally, the vastly greater chance of cells losing one versus both tumour suppressor alleles accounts for the dominant inheritance pattern of cancer predisposition in families, despite the fact that the defect inherited resides in a gene that acts recessively at the cellular level. Thus, in familial retinoblastoma, children who inherit a defective *Rb* gene are virtually assured of developing these tumours, indicating that at least one retinoblast in their eyes will almost certainly lose its remaining good copy of *Rb*.

The number of known tumour suppressors is quite small compared to the number of putative oncogenes, but most if not all of the former have a role in human tumour pathogenesis. They are functionally diverse, act in seemingly distinct cellular pathways, and defy ready categorization. The function of Rb and pl6 ^{INK4a} as a negative regulator of the cell cycle was described previously. WT1 is a transcriptional repressor that inhibits expression of genes that encode certain growth factors and growth factor receptors. The APC protein binds b-catenin (among other proteins) and targets it for degradation. b-Catenin is a protein with many functions, one of which is to interact with transcription factors of the Tcf (T cell factor)/Lef (lymphoid enhancer factor) family and activate transcription of genes that probably stimulate proliferation or inhibit apoptosis. In colon carcinoma cells deficient in APC, b-catenin levels are high and activity of b-catenin-transcription factor complexes is constitutive, which leads to activation of target genes and the beginnings of cell transformation and polyposis. Interestingly, the tumour suppressors responsible for hereditary non-polyposis colon cancer (HNPCC; a familial syndrome characterized by right-sided colon carcinomas without antecedent polyposis) are entirely different. Defects in several related genes and proteins are responsible for HNPCC, with defects in *hMSH2* and *hMLH1* being the most common. These proteins repair DNA that contain mismatches arising from errors during DNA synthesis, and defects in these proteins impede repair of these replication errors and allow them to accumulate. This 'mutator' phenotype presumably begins the process of cell transformation and colon carcinogenesis when errors affect genes that influence cell proliferation or apoptosis. These mismatch repair tumour suppressors differ from oncogenes and tumour suppressors, such as *Rb* and *APC*, in that the function of their products does not directly affect cell proliferation, apoptosis, or differentiation. Rather, they are responsible for maintaining the fidelity of genomic information transmitted from cell to progeny cell, and their absence contributes to tumourigenesis by favouring mutagenesis of genes that affect cell fate directly.

The relevance to cancer causation of genes that allow cells to repair defective DNA is reinforced by the role of the *p53* tumour suppressor in human cancers. Originally described as the gene producing a 53 kDa protein that is overexpressed in many transformed cells, *p53* was believed to be an oncogene, because overexpression of *p53* cloned from cell lines had transforming properties. Subsequently, comparison with normal *p53* cloned from fetal tissues revealed mutations in the *p53* genes previously used, and normal *p53* was found to suppress rather than promote cell transformation. These studies and findings of frequent *p53* LOH in tumours have led to the current view that *p53* is a tumour suppressor gene with a negative effect on cell proliferation. It is probably the most frequently deleted or mutated cancer-associated gene (involved in about 50 per cent of human tumours), and Li–Fraumeni syndrome kindreds, who are prone to developing a variety of cancers, are heterozygous for mutant *p53* in their germline. The observation that mutant *p53* transforms is explained by its 'dominant negative' effect: p53 modulates transcription of genes to which it binds as a tetramer, and mutant p53 not only loses this ability but prevents normal p53 from functioning when they are together in a mixed tetramer. Thus, a mutant *p53* allele may be transforming despite the presence of a normal *p53* allele in cells, especially since mutant p53 proteins are frequently longer lived and more abundant than normal p53 protein. Given the importance of p53 in preventing oncogenesis, it was somewhat surprising to find that mice with both their *p53* alleles deleted developed normally and were abnormal only in being prone to developing tumours. How p53 prevents tumourigenesis but is not essential for normal growth and development is explained by the conditional requirement for its activity. When DNA is damaged, for example by exposure to radiation or chemotherapy agents, p53 becomes functionally activated and modulates transcription of genes that induce cell cycle arrest, apoptosis, and repair of damaged DNA. Cell cycle arrest presumably allows the cell time to repair damaged DNA before it is replicated, passed on to progeny cells and becomes a ongoing source of genetic misinformation. Alternatively, induction of apoptosis presumably eliminates cells that are genetically damaged beyond repair. Cells without functional p53 fail to cell cycle arrest following DNA damage, are predisposed to genomic instability, and are less likely to be eliminated by apoptosis. These p53 functions obviously are not needed for normal growth and development but are needed for protection against DNA damage leading to cell transformation.

Regulatory pathways disrupted during cell transformation

As more and more oncogenes and tumour suppressor genes are identified, the proteins of many have been found to interact with each other and function in the same cell regulatory pathways. This suggests that these regulatory pathways are important for maintaining normal cell behaviour and must be disrupted for neoplasia to occur and that cells undergoing transformation may achieve this by altering one pathway participant or another (Fig. 14). For example p53, activated by DNA damage, transactivates expression of a gene, *mdm2*, whose protein binds p53, promotes its degradation, and inhibits its transcription regulatory function—that is Mdm2 is a negative regulator of p53 activity. Interestingly, Mdm2 is an oncogene that induces cell transformation when overexpressed in cells, and in tumours that overexpress Mdm2, p53 is normal. Thus, inactivating p53 is important for cell transformation, but this can be achieved either by inactivating p53 directly or activating Mdm2 expression. More recently another protein, p19 ^{ARF}, has been found to interact with Mdm2 and inhibit its ability to counter p53. The genetic locus of p19 ^{ARF} (which coincides with the gene for the Cdk4,6 inhibitor, p16 ^{INK4a}, but is in an alternative reading frame) is disrupted in familial melanomas and other types of cancers, suggesting that it is a tumour suppressor gene affecting the p53 functional pathway.

The APC-b–catenin interaction provides another example of the importance of disrupting certain regulatory pathways during cell transformation, no matter the means. Many sporadic colon carcinomas resemble carcinomas arising in those with familial adenomatosis polyposis in that the tumour cells have no functional APC. In these carcinomas, both alleles have been inactivated by somatic mutations. Other cases of sporadic colon carcinoma, however, have normal APC. A substantial proportion of these cases have mutant b-catenin genes that produce proteins that do not interact with APC but retain the ability to interact and function with Tcf/Lef transcription factors, that is b-catenin is a mutationally activated oncogene in these cells. One can infer that deregulated expression of genes regulated by b-catenin–Tcf/Lef is important for colon carcinogenesis, and less important is whether this is achieved by eliminating APC, an inhibitor of b-catenin activity, or mutating b-catenin so that it is no longer susceptible to APC regulation.

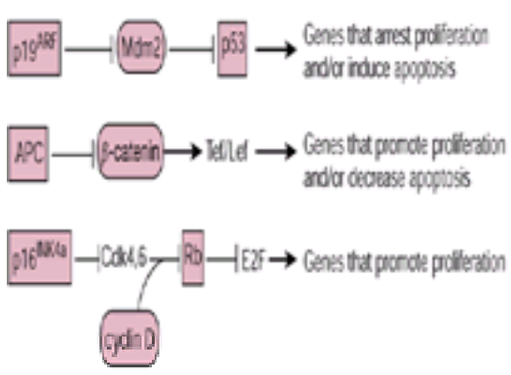

Fig. 14 Regulatory pathways targeted for deregulation in cancer. Three pathways important for the regulation of cell proliferation and apoptosis which are frequently deregulated during neoplastic cell transformation are shown. Components of these pathways that are frequently activated or inactivated by mutations in cancer cells are surrounded by ovals and rectangles. Rectangles designate tumour suppressor genes that are subject to loss-of-function mutations, and ovals designate

proto-oncogenes that are subject to gain-of-function mutations.

The pathway of cell cycle regulation involving Rb perhaps best illustrates the importance of certain regulatory pathways for controlling cell behaviour and how these pathways may be subverted in a number of different ways during cell transformation. In many types of human cancer cells, both copies of the *Rb* gene are inactivated, and there is no functional Rb. While this may be the most effective way to eliminate Rb regulation, other events pushing cells towards neoplasia may achieve a similar effect without altering Rb itself. Enhanced cyclin D expression due to chromosome translocation is seen in certain neoplasms. As cyclin D/Cdk4,6 phosphorylates Rb and prevents it from regulating E2F, deregulating cyclin D expression produces an oncogenic effect resembling Rb inactivation. p16 ^{INK4a} inhibits Cdk4,6 function, and loss of this Cdk inhibitor has been shown to result from gene deletion or epigenetic mechanisms in different tumour types. Its loss eliminates a regulator of cyclin D/Cdk4,6 activity, allowing the latter to phosphorylate Rb when it should not and producing an effect similar to Rb inactivation (note that disruption of p16 ^{INK4a} is likely to disrupt the p19^{ARF} gene as well, in which case two important pathways, Rb and p53, are disrupted by one genetic event). The complexity of the Rb regulatory pathway—p16^{INK4a} inhibiting cyclin D/Cdk4,6 activity which inhibits Rb control of E2F function—allows inactivation of p16 ^{INK4a}, deregulation of cyclin D/Cdk4,6, or inactivation of Rb to effectively deregulate E2F function and cell cycling. Recall that DNA tumour viruses (SV40, certain adenoviruses, and certain human papilloma viruses) also know the importance of this pathway and produce viral proteins that neutralize Rb function. Together, these argue compellingly for the central importance of the Rb pathway for maintaining normal cell regulation and preventing neoplastic behaviour, and cells on the path to transformation need to find a way to subvert it.

Limitations and uses of the information available

No one questions that investigations into the molecular basis of cell transformation over the past two decades have radically altered and refined our view of cancer pathogenesis. Major concepts and important subtleties now known were not even suspected two to three decades ago. However satisfying this may be, clearly much more needs to be learned. Cancer as a clinical entity is the product of many changes undergone by the cancer cell, not only intrinsically but also in terms of how it relates to the host. Our understanding of the molecular mechanisms involved in cell transformation, the initial steps in pathogenesis, have outpaced our understanding of how the neoplastic cell masters its environment, the steps responsible for most of the symptoms, signs, and clinical character of cancer: how transformed cells evade the host immune system and invade the surrounding extracellular matrix, induce host blood vessel growth necessary to supply the demands of enlarging tumours, enter and exit lymphatic and blood vessels to initiate metastasis, and thrive in distant, ectopic sites to form metastases.

Despite these gaps in our understanding, the knowledge gained so far has proved clinically useful. Knowing the oncoproteins involved in cell transformation, specific inhibitors of their function are being developed. Fortunately, even though neoplastic transformation involves multiple steps and has many participants, interrupting only one of these steps may be sufficient to inhibit transformed behaviour. Perhaps the closest to clinical application are putative inhibitors of Ras function, the farnesyltransferase inhibitors. Ras and related G-proteins must attach to cell membranes to function and do so by adding lipid groups (e.g. farnesyl and geranyl moieties) post-translationally. Compounds that inhibit this modification prevent proper localization and function of Ras and relatives of Ras. Tumour suppressor proteins, in contrast, are generally missing in tumour cells, and consideration is being given to functional reinstatement of these proteins in tumour cells to reverse their neoplastic behaviour. Currently, most attempts have focused on gene transfer technologies ('gene therapy') to accomplish this goal. Tumour-associated proteins that are mutant or overexpressed in transformed cells are potential neoantigens and have been suggested as targets for tumour-specific immunological attack. Mutant Ras proteins, novel Bcr–Abl fusion proteins, overexpressed HER2/neu receptors, and abundant mutant p53 proteins have each been suggested as the basis for tumour vaccine formulations. Whether these ideas and approaches turn out to be practical and therapeutically useful will probably be determined with another decade of research, but, in that interval, many additional new ideas and approaches will undoubtedly surface.

The genetic changes and molecular participants in neoplastic transformation also may be used for diagnostic and classification purposes. With the extreme detection sensitivity of polymerase chain reaction (PCR) technology, genetic alterations characteristic of certain cancers may be detected even when they are present at extremely low frequency. For example this has been applied to the *bcr–abl* translocation that characterizes chronic myelogenous leukaemia and used to detect minimum residual disease in patients following therapy. The association between this translocation and this leukaemia is so strong that the presence of the *bcr–abl* translocation is considered diagnostic of chronic myelogenous leukaemia, whether or not a t(9;22) Philadelphia chromosome is detected on cytogenetic examination. For many cancers, the molecular 'genotype' may affect and even define clinical behaviour and, perhaps, therapeutic response. For example the t(15;17) *PML–RARa* translocation is responsible for the distinctive phenotype of acute promyelocytic leukaemia and its initial dramatic response to all *-trans* retinoic acid therapy. The influence of genotype on therapeutic response may be especially important with regard to the p53 status of tumours. Mutations in p53 are so common in human cancers and its influence on apoptosis seems so likely to affect the killing of cancer cells by certain chemotherapy agents that tumour p53 status may become an important factor in deciding between different types of cancer therapy. These examples just begin to illustrate the potential applications and utility of molecular genotyping of tumours in the future.

Molecular basis of dilated cardiomyopathy

Defining the molecular mechanisms by which cells respond to environmental stimuli and communicate with each other has led to greater understanding of the role played by intracellular signalling pathways in the pathogenesis of many common diseases. While it is intuitive that deregulated signalling might lead to uncontrolled cell growth, and hence malignancy, it is increasingly recognized that signalling cascades also play a significant role in the pathogenesis of many non-malignant conditions. For example the molecular events that culminate in dilated cardiomyopathy and congestive heart failure are mediated by signalling cascades initiated by increases in biomechanical stress.

Cardiomyocytes respond to increased biomechanical stress by increasing myocardial mass commensurate with the increase in work load. This increase in mass, or hypertrophy, is characterized by an increase in the number of contractile protein units contained in individual cardiomyocytes, but not an increase in the number of cells. Depending on the inciting event, the level of work required, and the physiological state of the functioning myocardium, this adaptive response can result in physiological hypertrophy (a proportional increase in the length and width of cardiomyocytes), concentric hypertrophy (increases in width of cardiomyocytes out of proportion to increased length), or eccentric hypertrophy (increased length relative to width). For example, in normal individuals, exercise can lead to physiological hypertrophy in response to increased cardiac workload. Loss of myocardium, as occurs following myocardial infarction, similarly increases the workload on residual cardiomyocytes and initiates compensatory hypertrophy. The magnitude of myocardial loss, and hence the level of biomechanical stress on residual cells, is a major determinant of whether the heart responds with physiological or eccentric hypertrophy. Increased biomechanical stress can also result from defects intrinsic to cardiac myocytes (e.g. genetic abnormalities of cytoskeletal proteins as seen in some forms of muscular dystrophy) or from extrinsic defects (e.g. haemodynamic overload resulting from chronic hypertension).

The molecular events that induce cardiac hypertrophy are only beginning to be defined. Biomechanical stress is known to induce release of a number of growth factors and cytokines, including proteins that activate G-protein-coupled receptors (e.g. endothelin-1 and angiotensin II), receptor tyrosine kinases (e.g. insulin-like growth factor I or IGF-1), and tyrosine kinase-associated receptors (e.g. the interleukin-6-related cytokine, cardiotrophin 1). Furthermore, each of these receptors transduce signals that induce cardiac hypertrophy. Receptors for endothelin-1 and angiotensin II induce hypertrophy by activating the heteromeric G $_q$ protein, which, in turn, activates protein kinase A and specific isoforms of phospholipase (PLC-b and phospholipase D) and protein kinase C. Activation of these pathways leads to transcriptional reprogramming of cardiomyocytes and produces compensatory hypertrophy through, as yet, poorly characterized mechanisms (Fig. 15). Factors activating receptor tyrosine kinases induce cardiac hypertrophy through multiple, different MAP kinases, including the ERKs and the stress-activated MAP kinases, JNK/SAPK and p38, each of which has been shown to be capable of activating hypertrophy programmes. The tyrosine kinase-associated receptor for cardiotrophin, gp130/LIF, activates specific JAK and STAT isoforms (JAK1, JAK2, and Tyk2; and STAT1 and STAT3) that regulate cardiac hypertrophy. These same growth factors and cytokines have pleiotropic effects in various other tissues, but they all activate signalling cascades that lead to hypertrophy in cardiomyocytes. Whether these different pathways represent functional redundancy and how they interact is unclear.

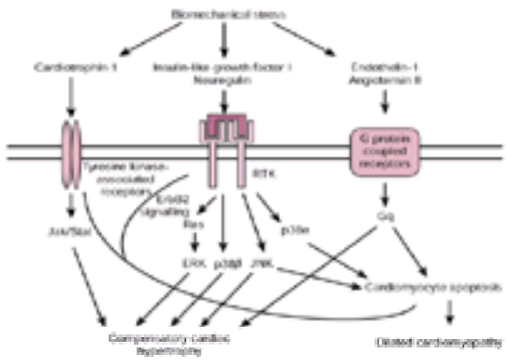

Fig. 15 Signalling pathways activated by biomechanical stress that mediate cardiac hypertrophy and heart failure. Biomechanical stress activates multiple signalling pathways that converge to induce compensatory cardiac hypertrophy, including signals mediated by receptor tyrosine kinases, tyrosine kinase-associated receptors, and G-protein-coupled receptors. However, many of these same pathways can also stimulate apoptosis pathways. Loss of cardiomyocytes through apoptosis results in increased biomechanical stress on residual cell and leads ultimately to a dilated cardiomyopathy and congestive heart failure. The end physiological response to biomechanical stress depends on the balance of these two opposing signal transduction pathways.

Prolonged, uninterrupted biomechanical overload ultimately leads to a dilated cardiomyopathy which is characterized by eccentric hypertrophy, loss of cardiac contractile function, and loss of cardiomyocytes due to apoptosis. The molecular events that tip the balance from adaptive hypertrophy to eccentric hypertrophy and heart failure are poorly understood, but increasing evidence implicates activation of apoptosis pathways in the development of dilated cardiomyopathy. Like most other cells, cardiomyocytes survival depends on activation of cell survival signals, and withdrawal of survival signals can induce apoptosis. This important property is highlighted by the unanticipated clinical observation that patients with metastatic breast cancer who received therapy with herceptin (a monoclonal antibody that blocks signalling via the receptor tyrosine kinase, ErbB2) had an approximately 15 per cent incidence of dilated cardiomyopathy, with patients previously exposed to anthracyclines being at greatest risk. The molecular basis of the heart failure seen in these trials is unclear, but was thought to result from apoptosis of cardiomyocytes. If true, this observation suggests that cardiomyocyte survival is dependent on ErbB2 signalling cascades and that apoptosis results when cardiomyocytes are deprived of these signals.

Interestingly, many of the signalling pathways that induce cardiac hypertrophy also regulate cell survival and cell death signals. The receptor for cardiotrophin-1, gp130/LIF, and the beta isoform of p38 transduce cell survival (antiapoptotic) signals, while the alpha isoform of p38, JNK/SAPK, and G_q all transduce proapoptotic signals. The balance of these pro- and antiapoptotic signals is likely to play a critical role in the adaptive response of the heart. When cell survival signals predominate, the heart responds with an adaptive hypertrophy, but a predominance of apoptotic signals induces cardiac decompensation with eccentric hypertrophy and heart failure. The magnitude and duration of biomechanical stress may play an important role in determining whether pro- or antiapoptotic signals predominate. For example moderate overexpression of G_q induces cardiac hypertrophy, but high expression results in cardiomyocyte apoptosis.

As heart failure progresses, signalling pathways are activated that appear to further exacerbate biomechanical overload and accelerate cardiac decompensation. To compensate for depressed cardiac function, b-adrenergic signalling pathways are activated in an attempt to augment cardiac contractility. While b-adrenergic signalling transiently improves contractility, prolonged stimulation depresses calcium fluxes resulting in impeded cardiac contractility. In this setting, b-adrenergic blockade improves calcium transport from the cytosol to the sarcoplasmic reticulum, and, in so doing, improves both cardiac relaxation and contractility. This observation provides the rationale for using b-adrenergic blockade in the treatment of patients with congestive heart failure.

Defining the signalling pathways that lead to cardiac hypertrophy has begun to provide a clearer understanding of the molecular basis for heart failure and the molecular events that link mechanical factors (e.g. workload) to biological effect. This understanding of what has traditionally been considered in mechanical and physical terms promises not only to provide important new mechanistic insights into these processes but also to suggest novel approaches for prophylactic and therapeutic intervention.

Further reading

Alberts B, *et al*.(2002). *Molecular biology of the cell*. New York, Garland Publishing, Inc.

Chien KR (1999). Stress pathways and heart failure. *Cell* **98**, 555–8. [Review.]

Lewin B, Lewin B (2000). *Genes VII*. New York, Oxford University Press, Inc.

Lodish H, *et al*. (1999). *Molecular cell biology*. New York, Scientific American Books, Inc.

4.4 Cytokines: interleukin-1 and tumour necrosis factor in inflammation

Charles A. Dinarello

The biology of cytokines

Cytokines are small, non-structural proteins with molecular weights ranging from 8000 to 40 000 Daltons. Originally called lymphokines and monokines to indicate their cellular sources, it became clear that the term 'cytokine' is the best description since nearly all nucleated cells are capable of synthesizing these proteins and, in turn, responding to them. There is no amino acid sequence motif or three dimensional structure that links cytokines; rather, their biological activities allow us to group them into different classes. For the most part, cytokines are primarily involved in host responses to disease such as infection and inflammation; involvement with homeostatic mechanisms is primarily at the level of host defence mechanisms in order to combat the constant challenge of micro-organisms from the environment. For example mice deficient in specific cytokines will spontaneously develop inflammatory bowel disease but when maintained in a germ-free environment, the disease does not occur.

It is not accurate to think of cytokines as hormones. First, hormones tend to be constitutively expressed by highly specialized tissues whereas cytokines are synthesized by nearly every cell. Hormones are the primary synthetic product of a cell (e.g. insulin, thyroid, ACTH), but cytokines account for a rather small proportion of the synthetic output of a cell. In addition, hormones are expressed in response to homeostatic control signals, many of which are part of a daily cycle. In contrast, most cytokine genes are not expressed unless specifically stimulated by noxious events. For example ultraviolet light, heat shock, hyperosmolarity, or adherence to a foreign surface activate cytokine gene expression. One concludes, then, that cytokines themselves are produced in response to 'stress' whereas most hormones are produced according to a daily, intrinsic clock.

Cytokine responses to infection and inflammation

There are at present 23 cytokines termed 'interleukin' (IL). Other cytokines have retained their original biological description such as 'tumour necrosis factor' (TNF). Most cytokines possess more than one biological activity and hence are often called 'pleiotropic'. Some cytokines clearly promote inflammation and are called proinflammatory cytokines whereas others suppress the activity of proinflammatory cytokines and are called anti-inflammatory cytokines. For example IL-4, IL-10, and IL-13 are potent anti-inflammatory agents. They are anti-inflammatory cytokines by virtue of their ability to suppress the expression of genes for proinflammatory cytokines such as IL-1 and TNF. Interferon-g (IFNg) is an example of the pleiotropic nature of cytokines. IFNg is an activator of the pathway which leads to cytotoxic T cells. However, IFNg is considered a proinflammatory cytokine because it augments TNF activity, upregulates vascular endothelial adhesion molecules, and induces nitric oxide (NO). Therefore, listing cytokines in various categories should be done with an open mind in that, depending upon the biological process, any cytokine may function differentially.

The concept that some cytokines function primarily to induce inflammation whereas others suppress inflammation is fundamental to cytokine biology and also to clinical medicine. The concept is based on the observation that during inflammation, the expression of genes coding for enzymes that synthesize small mediator molecules is upregulated. Examples of these proinflammatory enzymes are phospholipase A2 type-II, cyclo-oxygenase-2 (COX-2), and inducible nitric oxide synthase (iNOS), which synthesize platelet activating factor and leukotrienes, prostanoids, and nitric oxide. Another class of genes code for chemokines, small, proinflammatory peptides (8000 Daltons) that facilitate the passage of leucocytes from the circulation into the tissues. The prototypic chemokine is the neutrophil chemoattractant IL-8. IL-8 also activates neutrophils to degranulate and cause tissue damage. IL-1 and TNF are inducers of endothelial adhesion molecules, which are essential for the adhesion of leucocytes to the endothelial surface prior to emigration into the tissues.

Cytokine-mediated inflammation involves a cascade of proinflammatory gene products that are usually not produced in health. What triggers the expression of these genes? Although inflammatory products such as endotoxins can act directly, the cytokines IL-1 and TNF (and in some cases IFNg) are particularly effective in stimulating the expression of these genes. Moreover, IL-1 and TNF act synergistically in this process. Whether induced by a infection, trauma, ischaemia, immune-activated T cells, or toxins, IL-1 and TNF initiate the cascade of inflammatory mediators by targeting the endothelium. Figure 1 illustrates the inflammatory cascade triggered by IL-1 and TNF.

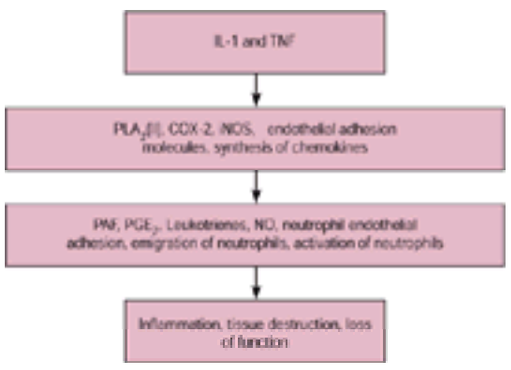

Fig. 1 The scheme of inflammatory cascade triggered by IL-1 and TNF.

IL-1 and TNF

The two cytokines discussed in detail here, IL-1 and TNF, have been selected because of their association with the pathogenesis and progression of disease, particularly autoimmune/inflammatory diseases such as rheumatoid arthritis and inflammatory bowel disease. IL-1 and TNF are the primary members of two 'families' of cytokines. Thus, seven cytokines have been identified as members of the IL-1 family; each has a different but related primary amino acid sequences and different but related receptors. The TNF 'family' and TNF receptor family is considerably larger (over 16 members). For the purposes of this chapter, IL-1 is used to denote IL-1b and IL-1a, the two major IL-1 agonists and TNF is used to denote TNFa, the major member of the TNF family. The case for a causative role of any cytokine in a particular disease can not be deduced from an association with production levels but can be proven only by specific inhibition of that cytokine in controlled, human trials. Fortunately, such trials are increasingly taking place since neutralizing antibodies to TNF and soluble receptors for TNF, which bind TNF in the extracellular

space, have been approved for clinical use. The IL-1 receptor antagonist IL-1Ra is likely to be approved for use in patients with rheumatoid arthritis.

The biology of IL-1 relevant to disease

Although animal experiments revealed that IL-1 was a highly proinflammatory cytokine, a great deal of information has been learned from studies in which humans were injected with either recombinant IL-1a or IL-1b. Results from trials for treating cancer or bone marrow suppression in which humans received IL-1, suggested a role for IL-1 in those conditions. Although there was no reduction in tumour growth, IL-1 treatment resulted in a more rapid recovery of platelets in subjects given high-dose chemotherapy. However, patients receiving 30 to 50 ng/kg of IL-1 developed fever, hypotension, and profound systemic, flu-like symptoms. Acute toxicities of either IL-1a or IL-1b were greater following intravenous compared to subcutaneous injection; subcutaneous injection was associated with significant local pain, erythema, and swelling. Chills and fever were observed in nearly all patients, even in the 1 ng/kg dose group. The febrile response increases in magnitude with increasing doses and chills and fever were abated with indomethacin treatment. In patients receiving IL-1a or IL-1b nearly all subjects experienced significant hypotension at doses of 100 ng/kg or greater. Systolic blood pressure fell steadily and reached a nadir of 90 mmHg, or less, 3 to 5 h after the infusion of IL-1. At doses of 300 ng/kg, most patients required intravenous pressors. By comparison, in a trial of 16 patients given IL-1b from 4 to 32 ng/kg subcutaneously, there was only one episode of hypotension at the highest dose level. These results suggest that the hypotension is probably due to induction of NO and elevated levels of serum nitrate have been measured in patients with IL-1-induced hypotension.

At 30 to 100 ng/kg of IL-1 patients exhibited a sharp increase in cortisol levels 2 to 3 h after the injection. In addition, there were increases in ACTH and thyroid stimulating hormone but a decrease in testosterone. No changes were observed in coagulation parameters such as prothrombin time, partial thromboplastin, or fibrinogen degradation products. This latter finding is in contrasted to TNFa infusion into healthy humans which resulted in a distinct coagulopathy syndrome. Not unexpectedly, IL-1 infusion into humans significantly increased circulating IL-6 levels in a dose-dependent fashion. These elevations in IL-6 are associated with a rise in C-reactive protein and a decrease in albumin.

Patients receiving only 3 ng/kg IL-1, given by intravenous infusion over 30 min, exhibited elevated IL-6 and IL-8 levels after 1 to 2 h; IL-6 reached a peak of 25 pg/ml after 1 h, IL-8 reached a peak of 311 pg/ml at 2 h, and nitrite/nitrate peaked after 10 h at 89 µmol/l (all statistically significant). If one calculates the maximum possible plasma concentration of IL-1b in these patients, the range would be 50 to 65 pg/ml. However, this calculated concentration is falsely high since the cytokine was given as a 30-min infusion rather than a bolus injection. Nevertheless, at 3 ng/kg, maximal plasma levels of IL-1b is 3 to 4 pM which is the same concentration of IL-1b needed *in vitro* for a biological response. Of interest is that the concentration of soluble IL-1R type II in healthy subjects is about 175 to 200 pM. When IL-1b is injected intravenously, it immediately encounters a 50-fold molar excess of soluble receptor to ligand. Why is there any biological effect to IL-1b given the high affinity of IL-1b for the soluble receptor? And the biological effect is dramatic with fever, hypotension, increased IL-6, IL-1Ra, IL-8, etc. There are two possibilities: the soluble receptor is already bound or the biological response is seen even at low concentrations.

Effects of TNFa injected into humans

Similar to IL-1, TNF is a highly proinflammatory cytokine when injected into humans. Unlike IL-1, TNF induces cell death, particularly in tumour cells. TNF has been injected as part of anticancer protocols, either systemically or regionally, in order to bring about death of tumour cells. Similar to the effects of IL-1 in humans, TNFa induces fever, headaches, myalgias, nausea, vomiting, and profound hypotension. Similar to IL-1 administration, elevated levels of IL-6, other cytokines, soluble cytokine receptors, and acute phase proteins are observed. In fact, the systemic responses to TNFa and IL-1 are mostly indistinguishable. However, activation of the coagulation pathways is observed with TNFa but not IL-1. The injection of TNF, 50 µg/m ^{2} body surface area, induced an early and short-lived rise in circulating levels of the activation peptide factor X and an increase in prothrombin fragment F1+2. These findings demonstrate that a single injection of TNF elicits a rapid and sustained activation of the coagulation pathway, probably induced through the extrinsic route. In these subjects, TNF induced a short-term increase in circulating plasminogen activator activity with rises in the antigenic levels of urokinase-type plasminogen activator and tissue-type plasminogen activator. This was followed by an eight-fold increase in plasminogen activator inhibitor type I and a sustained coagulation activation for 6 to 12 h. These findings may explain the microvascular thrombosis seen in septicaemia. Sequential measurement of the plasma concentrations of von Willebrand factor antigen was determined after a bolus intravenous injection of TNF (50 µg/m^2) in six healthy men. TNF induced a marked increase in von Willebrand factor antigen plasma levels, becoming significant after 45 min and peaking after 4 h (351 per cent increase). This increased von Willebrand factor secretion may explain the release observed in acute and chronic inflammatory disease and in systemic infection.

TNF also induced a transient stress hormone response, associated with an early and sustained rise in plasma glucose, free fatty acid, and glycerol concentrations. Resting energy expenditure showed a transient rise after TNF injection. In six healthy males, an intravenous bolus injections of TNF (50 µg/m ^{2}) produced the characteristic changes in circulating thyroid hormones and TSH observed in the sick euthyroid syndrome. In these subjects, TNF elicited a neutropenia after 15 min, followed by a neutrophilia. Decreased numbers of lymphocytes and monocytes were observed for several hours. There were increases in the concentrations of elastase–a1-antitrypsin complexes, lactoferrin levels, and neopterin, suggesting neutrophil and monocyte activation.

Diagnostic and prognostic value of measuring cytokines

Studies in human subjects have consistently revealed that circulating levels of pro- and anti-inflammatory cytokines as well as soluble cytokine receptors are elevated; moreover, levels may correlate with disease severity and outcome. In addition, spontaneous gene expression for certain cytokines in peripheral blood leucocytes or bone marrow cells are present in patients with various inflammatory and malignant disease whereas cells from healthy subjects rarely exhibited spontaneous cytokine production. A variety of animal models of infection, inflammation, or autoimmunity reveal that nearly all pro and anti-inflammatory cytokines are produced during disease and, as in humans, correlate with disease intensity. The best correlations of cytokine levels with disease activity are reported for IL-6 and IL-1Raa rather than with IL-1 or TNF; this includes patients with local infection, sepsis, and autoimmune diseases such as rheumatoid arthritis. Both IL-6 and IL-1Ra are considered 'acute phase reactants' and other correlate with C-reactive protein measurements. In addition, soluble receptors for IL-1 or TNF are better markers of disease intensity than the IL-1 or TNF itself. In patients with sepsis or septic shock, IL-6 and IL-1Ra levels correlated with mortality. In a study of patients with acute coronary artery syndromes, the admission and 48 h levels of IL-6 and IL-1Ra indicated which will undergo a corrective procedure such as angioplasty or by-pass graft and which patients will respond to conservative therapy and be discharged. In patients with rheumatoid arthritis who receive anti-TNF or anti-IL-1 therapy, the levels of IL-6 fall and are associated with a decrease in disease activity. Since IL-1 and TNF are the primary inducers of IL-6 and IL-1Ra, it is not surprising that therapeutic interventions to reduce IL-1 or TNF activity are associated with a decrease in IL-6 and IL-1Ra.

Reducing IL-1 and TNF activities in human disease

The reduction of the biological activity of any cytokine is based on a class of heterogeneous agents termed 'anticytokine-based therapies'. Treating patients with agents designed to reduce the production of cytokines or their biological effects has entered clinical medicine. Therapeutic strategies for reducing the effects of proinflammatory cytokines such as IL-1 or TNF are either specific or non-specific. This is an important distinction. For example corticosteroids inhibit the synthesis of proinflammatory cytokines but also suppress T-cell function as well as several metabolic pathways. Thalidomide also suppresses the synthesis of TNF, IL-1, and other cytokines and has been used to treat refractory multiple myeloma and the oral aphthous lesions of HIV-1 disease. IL-10, an anti-inflammatory cytokine, suppresses the synthesis of IL-1 and TNF but also of IL-6, IL-12, IFNg, and the entire family of chemokines. In contrast, a specific anticytokine therapy targets only one cytokine or closely related members of a single cytokine family. Neutralizing antibodies, soluble cytokine receptors, or receptor antagonists have the advantage of specificity, preventing the activity of a single cytokine in a particular disease. The first anticytokine-based therapy for rheumatoid arthritis employed a monoclonal anti-TNFa antibody which neutralizes only TNFa. IL-1 receptor antagonist blocks the IL-1 receptor; in doing so, it prevents the binding of either IL-1a or IL-1b. Soluble IL-1R type I binds IL-1Ra as well as IL-1a. Nevertheless, soluble receptors for TNF, the IL-1 receptor antagonist, and soluble IL-1 receptors (either type I or type II) are examples of specific anticytokine therapies.

Specific anticytokine-based therapies usually require parenteral administration. Non-specific agents, even when their primary mode of action is anticytokine based, are usually administered orally. The precursors for IL-1b, TNFa, and IL-18 each require cleavage by specific proteases before they are active molecules. Therefore, inhibiting these proteases is a logical anticytokine-based, anti-inflammatory strategy. Cleavage of the IL-1b or IL-18 precursor molecules is carried out by the cysteine protease IL-1b converting enzyme (ICE, also known as caspase-1); inhibition of this enzyme is effective in reducing the secretion of biologically active IL-1b and IL-18. For cleavage of the TNFa precursor, a specific metalloprotease is needed. Small, orally-active molecules, which inhibit these proteases specifically, reduce disease activity in animals and are at present undergoing clinical trials.

In animals, blocking either IL-1 or TNF reduced disease severity despite the fact that several cytokines were overexpressed in the disease model. This finding led to the concept that IL-1 and TNF are 'upstream' from many of the other cytokines produced in disease. Thus, IL-1 or TNF, or the combination of IL-1 plus TNF acting

synergistically, induce chemokines, IL-6, and several other 'downstream' cytokines. These observations explaines why blocking just one cytokine (either IL-1 or TNF) reduces disease. Moreover, of these two cytokines, TNF induces IL-1 production. Thus, in patients with rheumatoid arthritis who receive neutralizing monoclonal antibodies to TNFa, there is a rapid fall in circulating IL-1b as well as IL-6 within 24 h of the initiation of the antibody treatment.

Treating rheumatoid arthritis with IL-1Ra

IL-1Ra was initially tested in a trial in 25 patients with rheumatoid arthritis. In patients receiving 4 mg/kg per day for 7 days, there was a reduction in the number of tender joints from 24 to 10, the erythrocyte sedimentation rate fell from 48 to 31 mm/h and C-reactive protein decreased from 2.9 to 1.9 µg/ml. In an expanded double-blind trial, IL-1Ra was given to 175 patients. After 3 weeks, a significant reduction in disease parameters such as number of swollen joints, the investigator and patient assessments of disease activity, pain score, and C-reactive protein levels. Optimal improvement was in patients receiving 70 mg/day.

A double-blind, placebo-controlled, multicentre trial of IL-1Ra in 472 patients has been reported. There were three doses of IL-1Ra, administered subcutaneously: 30, 75 and 150 mg/day for 24 weeks. After 24 weeks, 43 per cent of the patients receiving 150 mg/day of IL-1Ra met the criteria for response (the primary efficacy measure), 44 per cent met the Paulus criteria, and statistically significant (p=0.048) improvements were seen in the number of swollen joints, number of tender joints, investigator's assessment of disease activity, patient's assessment of disease activity, pain score on a visual analogue scale, and duration of morning stiffness. In addition, there was a dose-dependent reduction in C-reactive protein level, and erythrocyte sedimentation rate.

Importantly, the rate of radiological progression in the patients receiving IL-1Ra was significantly less than in the placebo group at 24 weeks, as evidenced by the Larsen score and the erosive joint count. The reduction in new bone erosions was assessed by two radiologists who were blinded to the patient treatment as well as blinded to the chronology of the radiographs. This finding suggests that IL-1Ra is blocking the osteoclast activating factor property of IL-1, as has been reported in myeloma cell cultures. This study confirmed both the efficacy and the safety of IL-1Ra in a large cohort of patients with active and severe rheumatoid arthritis.

A trial of IL-1Ra in combination with methotrexate in patients with rheumatoid arthritis has also been reported; 419 patients were randomized to receive either placebo or increasing doses of IL-1Ra. Patients were also being treated with methotrexate (mean 17 mg/week). After 24 weeks, patients taking IL-1Ra (1.0 mg/kg) had significantly decreased parameters of disease compared to the placebo. For examples the proportion of patients with a 50 per cent reduction in disease activity was significantly greater in the IL-1Ra treated group (24 per cent) compared to the placebo (4 per cent). These studies suggest that the addition of IL-1Ra to optimal methotrexate treatment results in a further decrease in disease.

Treating rheumatoid arthritis with soluble IL-1R type I

Soluble IL-1R type was administered subcutaneously to 23 patients with active rheumatoid arthritis in a randomized, double-blind study. Patients received subcutaneous doses of the receptor at 25, 250, 500, or 1000 µg/m^2 per day or placebo for 28 consecutive days. In patients receiving 1000 µg/m^2 per day, only one showed improvement in measures of disease activity. One possible explanation for the lack of clinical response despite efficacy in suppressing immune responses could be the inhibition of endogenous IL-1Ra. This was observed in volunteers receiving soluble IL-1R type I before challenge by endotoxin.

Neutralizing TNF in rheumatoid arthritis

Strategies for reducing the biological activity of TNF have focused on neutralization by either monoclonal antibodies or soluble TNF receptors. The initial studies on the efficacy of blocking TNF in rheumatoid arthritis were carried out with the monoclonal antibody cA2 which is specific for neutralization of TNFa. Using a double-blind, placebo-controlled trial, increasing amounts of the antibody clearly showed that specific neutralization of TNFa in rheumatoid arthritis reduced the severity of both clinical and laboratory parameters of the disease. There were dramatic reductions in the circulating levels of IL-1b, IL-6, soluble vascular adhesion molecules, and other markers of systemic inflammation such as C-reactive protein and erythrocyte sedimentation. When combined with methotrexate, the efficacy of monoclonal anti-TNFa in this disease is improved over either agent alone. The biological basis for the effectiveness of blocking TNFa in rheumatoid arthritis is experiments showing a unidirectional cascade of cytokines in explants of synovial tissues in which the presence of anti-TNFa antibody inhibited the spontaneous production of IL-1, IL-6, and IL-8. IL-1Ra reduced IL-6 and IL-8 production but not TNFa production.

Treating rheumatoid arthritis with soluble TNFR p75-Fc

There are two distinct receptors for TNF termed p55 and p75, based on their molecular size. The extracellular forms of these receptors (also termed 'soluble' receptor), which circulate in healthy humans, can be administered in pharmacological concentrations in order to neutralize TNF. To increase the affinity of TNFa and to prolong its plasma concentration, the p75 receptor was synthesized as a fusion protein linked to human Fc domain of the IgG1 (TNFR p75-Fc). A randomized, double-blind, placebo-controlled trial involving 234 patients with active rheumatoid arthritis was carried out using soluble TNFR p75-Fc administered twice-weekly by subcutaneous injection. After 3 months of treatment, there was a improvement in 62 per cent of the patients compared to 23 per cent of the placebo using a 20 per cent ACR response. After 6 months of treatment, 59 per cent of the treated group but only 11 per cent of the placebo group achieved a 20 per cent ACR response. Clearly, blocking TNF using this construct resulted in a significant and sustained benefit in patients with active rheumatoid arthritis. However, similar to monoclonal anti-TNFa, the inflammatory component of the disease returned upon cessation of therapy.

When combined with methotrexate, the use of TNFR p75-Fc resulted in a ignificant reduction in disease activity using a lower-dose of methotrexate. Eighty-nine rheumatoid arthritis patients were treated for 24 weeks with a stable dose of methotrexate and randomized to receive placebo or TNFR p75-Fc; 71 per cent of the patients receiving TNFR p75-Fc plus methotrexate but only 27 per cent of the placebo group plus methotrexate showed a reduction in the ACR 20 criteria. TNFR p75-Fc has also been used to treat patients with juvenile rheumatoid arthritis and Still's disease.

Treating patients with Crohn's disease with anti-TNFa

Animal and clinical studies suggest a role for TNF in the pathogenesis of inflammatory bowel disease, particularly Crohn's disease. During a12-week multicentre, double-blind, placebo-controlled trial of a single infusion of anti-TNFa monoclonal antibody treatment in patients with moderate-to-severe Crohn's disease, 64 per cent of those given the antibody had a clinical response, as compared to 17 per cent of patients in the placebo group. Thirty-three per cent of the patients given the antibody went into remission. After 12 weeks, 41 per cent of the antibody-treated patients exhibited a clinical response as compared with 12 per cent of the patients in the placebo group. On the basis of this and other trials, the use of anti-TNFa monoclonal antibody to treat Crohn's disease was approved. In another study, 94 patients with enterocutaneous fistulas of draining abdominal or perianal fistulas of at least 3 months' duration were randomly assigned to receive placebo or anti-TNFa antibody; 38 per cent of the patients who received the antibody had closure of their fistulas compared to 13 per cent of the patients in the placebo group.

Neutralizing TNF in the treatment of congestive heart failure

Initial studies measuring biologically active TNF in the circulation of patients with severe (New York Heart Association class III and IV) congestive heart failure showed that the levels of TNF correlated with the severity of cachexia and exercise tolerance. Therefore, treating patients with severe heart failure with strategies to block TNF activity were tested. To date, 18 (class III) heart failure patients have been treated with p75-Fc, given as a single intravenous infusion, in a randomized, double-blind trial. A significant overall increase in quality-of-life scores, 6-min walk distance, and ejection fraction was observed in the cohort that receiving p75-Fc. There was no significant change in these parameters in the placebo group. It should be noted that in these studies, the patients were already receiving optimal treatment for heart failure with ACE inhibitors and b-blockers.

The biological basis for TNF and IL-1 in heart failure is well-established. Using strips of human atrial heart tissue *ex vivo*, the presence of concentrations of TNFa or IL-1b as low a few pg/ml results in depression of contractility. Moreover, the effect of these two cytokines to depress myocardial function is highly synergistic.

Blocking IL-1 and TNF in patients in septic shock

The great promise of anticytokine therapy in human disease was to reduce the mortality of sepsis and particularly septic shock. Despite the impact of intensive care units and optimal physiological support, the mortality in septic shock remain high at 35 to 40 per cent. A great number of animal models of septic shock provided overwhelming data that blocking IL-1 or TNF activity would reduce mortality in these patients. Over 10 000 patients have been studied in carefully designed trials to examine IL-1Ra, neutralizing monoclonal antibodies to TNFa, soluble TNF receptor p55, and soluble p75 TNF receptor, the same agents described above as showing efficacy in the treatment of rheumatoid arthritis, Crohn's disease, and congestive heart failure. There has been no statistically significant improvement in all cause

mortality 28 days after the onset of anticytokine treatment. However, in a meta-analysis of these trials, it was concluded that anticytokine intervention resulted in a small but consistent reduction (2–6 per cent) in all-cause 28-day mortality. In only one trial, mortality increased as a result of the anticytokine therapy. Some conclusions can be made from these trials since they offer insights into the effect of the same anticytokine agents as in patients with rheumatoid arthritis. First, despite being infected, the vast number of patients did not worsen or become vulnerable to more infection. There was one exception, the p75-Fc which is effective in treating rheumatoid arthritis was associated with a dose-dependent increase in 28 day all cause mortality. In each study, there was a subgroup with clear benefit but the overall group did not benefit significantly. The heterogeneity of the patients probably prevented the detection of a statistically significant improvement. Fortunately, these anticytokines are now receiving considerable attention for use in local inflammatory diseases.

Further reading

Bresnihan B, Alvaro-Gracia JM, Cobby M, *et al.* (1998).Treatment of rheumatoid arthritis with recombinant human interleukin-1 receptor antagonist. *Arthritis and Rheumatism* **41**, 2196–204.

Campion GV, Lebsack ME, Lookabaugh J, *et al.* (1996). Dose-range and dose-frequency study of recombinant human interleukin-1 receptor antagonist in patients with rheumatoid arthritis. *Arthritis and Rheumatism* **39**, 1092–101.

Elliott MJ, Maini RN, Feldmann M, *et al.* (1994). Randomised double-blind comparison of chimeric monoclonal antibody to tumour necrosis factor alpha (cA2) versus placebo in rheumatoid arthritis. *Lancet* **344**, 1105–10.

Moreland LW, Baumgartner SW, Schiff MH, *et al.* (1997). Treatment of rheumatoid arthritis with a recombinant human tumor necrosis factor receptor (p75)-Fc fusion protein. *New England Journal of Medicine* **337**, 141–7.

Torcia M, Lucibello M, Vannier E, *et al.* (1996). Modulation of osteoclast-activating factor activity of multiple myeloma bone marrow cells by different interleukin-1 inhibitors. *Experimental Hematology* **24**, 868–74.

4.5 Ion channels and disease

Frances M. Ashcroft

Ion channels are membrane proteins that act as gated pathways for the movement of ions across cell membranes. They are found in both surface and intracellular membranes, and play essential roles in the physiology of all cell types. An ever-increasing number of human diseases are found to be caused by defects in ion channel function. Ion channel diseases may arise in a number of different ways:

- From mutations in the coding region of the gene, or its control elements, leading to the gain, or loss, of channel function (Table 1). Diseases that result from ion-channel mutations are often known as 'channelopathies'. As with all single-gene disorders, their frequency in the general population is usually very low. Many channelopathies are genetically heterogeneous and the same clinical phenotype may be caused by mutations in different genes, as is the case for Long-QT syndrome. Conversely, mutations in the same gene may produce different phenotypes. For example, gain-of-function mutations in the epithelial Na^+ channel produce Liddle's syndrome, whereas loss-of-function mutations cause psuedohypoaldosteronism type-1. Disease severity may also vary with different mutations in the same gene.
- From defective regulation of channel activity by intracellular or extracellular ligands, or by channel modulators, due to mutations in the genes encoding the regulatory molecules themselves, or defects in the pathways leading to their production. For instance, glucokinase mutations cause one type of maturity-onset diabetes of the young (**MODY2**), by impairing the metabolic regulation of ATP-sensitive K^+-channels in pancreatic b-cells.
- From autoantibodies to ion channel proteins, which may either downregulate or enhance channel function (see Table 2). These diseases are discussed elsewhere.
- From ion channels that act as lethal agents. These are secreted by cells and insert into the membrane of the target cell to form large non-selective pores that cause cell lysis and death. Examples include bacterial toxins such as staphylococcal a-toxin and the amoebopore of *Entamoeba histolytica.* The membrane-attack complex of complement, perforin, and the defensins also act in this way.

To understand how ion channel defects give rise to disease, it is helpful to understand how these proteins work. The next section therefore considers what is known of ion channel structure, explains the properties of the single ion channel and shows how single-channel currents give rise to action potentials and synaptic potentials.

Properties of ion channels

Ion channel structure

Some ion channels consist of a single subunit, as in the case of the Ca^{2+}-release channel of the sarcoplasmic reticulum. In other cases, the channel pore is formed from a single (a) subunit but associated regulatory subunits may modify the ion channel properties, as in the case of voltage-gated Na^+ and Ca^{2+} channels. Yet other ion channels are multimeric and several subunits are involved in pore formation—the nicotinic acetylcholine receptor comprises five subunits (2a, b, g, †), while the voltage-gated K^+ channels are composed of four subunits (which are sometimes, but not invariably, identical). Mutations in both pore-forming and regulatory subunits can cause disease.

The multimeric nature of an ion channel may influence whether or not a channelopathy is inherited in a dominant or recessive fashion. Individuals who are heterozygous for voltage-gated K^+ channel mutations will express both mutant and wild-type subunits in the same cell. If the mutant subunits co-assemble with wild-type subunits to form heteroligomeric channels that are non-functional, the resulting K^+ current will be much smaller than if heteromultimerization does not occur. This is known as the 'dominant negative' effect, and may give rise to a disease that is dominantly inherited.

Single-channel properties

An ion channel can either be open or closed. When it is open, permeant ions are able to move through the channel pore. The current flowing through the open pore is known as the single-channel current. Its magnitude is determined by the ion concentrations on either side of the membrane (the chemical gradient), by the membrane potential (the electrical gradient), and by the ease with which the ion can move thorough the channel pore (its permeability). At the equilibrium potential of an ion, the electrical and chemical gradients are equal in magnitude but opposite in direction, and thus there is no net ion flux. The single-channel conductance g is a measure of the permeability of the ion and is given by the single-channel current (i) divided by the membrane potential ($g = i/V$).

Ion channels are often highly selective in the ions they conduct. K^+ channels, for example, are about 100 times more permeable to K^+ than Na^+ ions, while Na^+ channels conduct Na^+ ions but discriminate against K^+ ions. Ion selectivity takes place within a narrow region of the pore known as the selectivity filter. The basis of ion selectivity is only just beginning to be understood, but it is clear that while some ions are excluded on the basis of their size or their charge, hydrophobic interactions and the energy required to remove the waters of hydration are also important.

The fraction of time the channel spends in the open state is known as the open probability. Some channels open and close at random, but gating is regulated in other channels. In voltage-gated channels, the open probability is determined by the membrane potential, whereas in ligand-gated channels it is regulated by the binding of extracellular or intracellular ligands. Gating may also be subject to modulation, a process in which channel opening or closing is modified, usually by one of a number cytosolic substances (for example, by Ca^{2+} binding, phosphorylation, etc.). Gating is believed to involve conformational changes in the channel structure that result in the opening or closing of the pore.

At the resting potential of the cell, most voltage-gated channels are closed. In response to a membrane depolarization, the open probability of the channel is increased. This voltage-dependent activation may be followed by a further conformational transition (inactivation) to an inactivated state in which the channel no longer conducts ions. Recovery from inactivation occurs after a variable period following repolarization to the resting potential. Although most voltage-gated ion channels are opened by depolarization, a few types of voltage-gated channel are activated by hyperpolarization. Ligand-gated channels are opened (or more rarely closed) by binding of an appropriate ligand to a specific site on the channel protein, which induces a conformational change that allosterically opens the ion pore. The channel may open and close several times while the ligand remains bound to its receptor, but this intrinsic gating ceases on ligand dissociation.

There are numerous different types of channel. For example, even among the inwardly rectifying K^+ channels, there are seven subfamilies, most of which have several members. In general, ion channels are named after their gating and/or selectivity properties.

Single-channel currents summate to produce macroscopic currents

The cell membrane contains many hundreds of ion channels. The macroscopic current (I) flowing through all ion channels of the same type is determined by the

product of the number of channels in the membrane (*N*), the channel open probability (*P*), and the single-channel current (*i*); in other words *I = NPi.* Disease-causing mutations may affect any or all of these parameters and thereby influence the macroscopic current.

Cell membranes also contain several different types of channel. The total current that flows across the cell membrane (the membrane current) represents the sum of the ion fluxes through all the different kinds of ion channel open in the membrane. If it is sufficiently large, the membrane current may cause a change in membrane potential. The size of this voltage change is given by Ohm's law (*V = IR*) and is therefore influenced by both the current amplitude (*I*) and by the membrane resistance (*R*) (which in turn reflects the number of open channels). A change in the membrane potential to a more positive value is known as 'depolarization'; 'hyperpolarization' is a change to more negative potentials. The resting potential of most cells lies between –60 to –100 mV.

Action potentials

In excitable cells, a depolarizing stimulus may elicit an action potential. In nerve axons and skeletal muscle fibres, the action potential results from the initial activation of voltage-gated Na^+ channels followed shortly afterwards by activation of voltage-gated K^+ channels. Because Na^+ channels open rapidly on depolarization, there is an initial inward Na^+ current. If this is greater than the outward current flowing through (voltage-independent) K^+ channels which are open at the resting potential, it will produce a further depolarization. This activates more Na^+ channels and depolarizes the membrane even more. In this way, a regenerative increase in membrane potential (an action potential) is produced. The membrane is returned to its resting level by inactivation of the Na^+ channels (which reduces the inward current) and the opening of K^+ channels (which produces an outward, hyperpolarizing current).

The potential at which the inward Na^+ current exactly balances the outward resting K^+ current through resting K^+ channels is known as the threshold potential. It is a critical potential: any increase in the Na^+ current will elicit an action potential, while any reduction in the inward current (or increase in the outward current) will prevent action-potential generation. Ion channel mutations may increase nerve or muscle excitability either by enhancing the inward current (as in hyperkalaemic periodic paralysis), or by reducing the outward current (as in benign familial neonatal convulsions). This will produce a larger depolarization, so that the threshold potential is reached more easily and a subsequent action potential is initiated. Other mutations produce a depolarizing block of action-potential activity. This results from a maintained membrane depolarization of sufficient amplitude to inactivate the voltage-dependent Na^+ channels.

In some cells, additional types of ion channel contribute to the action potential—the ventricular action potential is mediated by voltage-dependent Na^+, Ca^{2+}, and at least four kinds of K^+ channel; several different kinds of K^+ channel contribute to the repolarization of action potentials in mammalian neurones; and chloride channels play an important role in the electrical activity of skeletal muscle. The functional importance of these different ion channels is exemplified by the fact that mutations in the genes which encode them produce a range of nerve and muscle diseases.

Synaptic potentials

When a nerve impulse arrives in the presynaptic terminal it opens voltage-gated Ca^{2+} channels, producing a rise in the intracellular Ca^{2+} concentration ($[Ca^{2+}]_i$) that triggers the exocytosis of synaptic vesicles. The amount of transmitter released varies with $[Ca^{2+}]_i$ and thus with the magnitude of the presynaptic Ca^{2+} current. In turn, this is influenced by the duration of the membrane depolarization and thus by the amplitude of the voltage-gated K^+ current that underlies membrane repolarization. A reduction in the presynaptic K^+ current therefore leads to excess transmitter release and postsynaptic hyperexcitability, as in episodic ataxia type 1 and acquired neuromyotonia. Conversely, a reduction in the presynaptic Ca^{2+} current is associated with reduced transmitter release, as occurs in the Lambert–Eaton myasthenic syndrome when the density of presynaptic Ca^{2+} channels is decreased by receptor internalization induced by the binding of autoantibodies.

Once released, the transmitter diffuses across the synaptic cleft and binds to receptors in the postsynaptic membrane. At the neuromuscular junction, for example, acetylcholine (**ACh**) binds to the nicotinic acetylcholine receptor (**AChR**), and opens an intrinsic ion channel. The resulting synaptic current produces a depolarization of the postsynaptic membrane (the endplate potential) which, if it is sufficiently large, triggers an action potential in the muscle fibre. A reduction in AChR density, as in myasthenia gravis, decreases effective transmission and leads to muscle weakness. Gain-of-function mutations in AChR may also induce myasthenia, by causing prolonged depolarization of the postsynaptic membrane and thereby Na^+ channel inactivation. This depolarizing block is the basis of the slow-channel syndromes. Mutations in the voltage-gated Na^+ channel of skeletal muscle may cause paralysis, or myotonia.

In skeletal muscle, the action potential is conducted into the interior of the fibre via invaginations of the surface membrane known as the transverse tubules (**T-tubules**). Depolarization of the T-tubule membrane stimulates the opening of Ca^{2+}-release channels (**RyR**) in the membrane of the sarcoplasmic reticulum (**SR**), the intracellular Ca^{2+} store. The T-tubule and SR membranes are not directly connected and the precise mechanism by which they interact is not fully understood. However, there is evidence that the a_1-subunit of the voltage-gated Ca^{2+} channel in the T-tubule membrane acts as the voltage sensor for the Ca^{2+}-release channels in the SR membrane. Mutations in the Ca^{2+}-release channel of skeletal muscles cause malignant hyperthermia and central core disease.

The channelopathies

This section provides brief descriptions of a range of channelopathies. Additional details may be found elsewhere in the *Oxford textbook of medicine* or in the books and Websites referenced.

Neuronal channelopathies

Generalized epilepsy with febrile seizures

Generalized epilepsy with febrile seizures (**GEFS**) has been linked to a mutation in the gene (*SCN1B*) that encodes the b_1-subunit of the voltage-gated Na^+ channel. Affected individuals exhibit febrile seizures in childhood and afebrile generalized epilepsy in later life. The presence of the b-subunit accelerates both the rate of inactivation, and the rate of recovery from inactivation, of the voltage-gated Na^+ channel. This modulatory effect is abolished if the b_1-subunit carries the GEFS mutation. It is predicted to cause a persistent inward Na^+ current that leads to neuronal hyperexcitability and seizures.

Benign familial neonatal convulsions

Benign familial neonatal convulsions (**BFNC**) is characterized by neonatal convulsions within the first 3 days after birth that show spontaneous remission by the third month of life. There is an increased risk of epilepsy in later life in 10 to 15 per cent of individuals. Mutations in the voltage-gated K^+ channel genes *KCNQ2* and *KCNQ3* are associated with BFNC.

KCNQ2 and KCNQ3 associate in a heteromeric complex to form the M-channel. This channel plays a critical role in determining the electrical excitability of many neurones. It is slowly activated when the membrane is depolarized to around the threshold level for action potential firing, thereby hyperpolarizing the membrane back towards its resting level. This reduces neuronal excitability by limiting the spiking frequency and decreasing the responsiveness of the neurone to synaptic inputs. All BNFC mutations studied to date result in reduced expression of the mutant protein. This may be expected to lead to neuronal hyperexcitability, accounting for the epileptic seizures. Because the M-channel is a heteromer of KCNQ2 or KCNQ3, mutations in either gene will disrupt channel function and cause BNFC.

Episodic ataxia type-1

Episodic ataxia type 1 (familial periodic cerebellar ataxia with myokymia) is an autosomal dominant disorder that causes ataxia accompanied by myokymia, nausea, vertigo, and headache. It results from mutations in the voltage-gated K^+ channel $K_V1.1$, which is expressed in the synaptic terminals and dendrites of many brain neurones. These mutations either prevent the formation of functional channels or result in a reduced K^+ current. This is expected to prolong the neuronal action potential, inducing repetitive firing and excessive and unregulated transmitter release, and thereby produce the clinical symptoms of ataxia and myokymia.

Familial hemiplegic migraine, episodic ataxia type-2, and spinocerebellar ataxia type-6

There are three human diseases with different phenotypes that are associated with mutations in the same Ca^{2+}-channel gene, *CACNL1A4*. These are familial hemiplegic migraine (**FHM**), episodic ataxia type-2 (**EA-2**), and spinocerebellar ataxia type-6 (**SCA-6**). FHM is associated with missense mutations, EA-2 is caused by truncation of the protein within the third repeat, and SCA-6 is produced by expansion of a polyglutamine repeat in the C-terminal coding region of the protein. All three diseases result in progressive cerebellar atrophy, but they differ in the extent and rate of progression of neuronal degeneration, with SCA-6 showing the greatest, and FHM the least, atrophy. Migraine-like symptoms also occur in all three diseases, and are most severe in patients with FHM. EA-2 and SCA-6 are also characterized by ataxia and nystagmus. It remains unclear how the different mutations in *CACNL1A4* give rise to the different phenotypes.

Startle disease (hyperekplexia)

Glycine is the major inhibitory transmitter in the brainstem and spinal cord. It binds to a ligand-gated Cl^- channel, producing an increase in Cl^- permeability that reduces the membrane depolarization and neuronal firing induced by excitatory neurotransmitters. The glycine receptor is a pentamer of three a-subunits, which contain the glycine-binding site, and two b-subunits. In humans, two types of the a-subunit have been identified. Mutations in the gene encoding the a_1-subunit of the glycine receptor give rise to startle disease (hyperekplexia). This is an autosomal dominant, neurological disorder characterized by muscle spasm in response to an unexpected stimulus. It manifests as facial grimacing, hunching of the shoulders, clenching of the fists, and exaggerated jerks of the limbs. Startle disease mutations produce a dramatic decrease in glycine-activated currents. Because glycinergic interneurones are important for normal spinal cord reflexes, muscle tone, and the pattern of motor neurone firing during movement, this leads to excessive and uncontrolled movements.

Charcot-Marie-Tooth disease

Charcot-Marie-Tooth disease type 1 (CMT1) causes progressive degeneration and demyelination of the peripheral nerves. It is genetically heterogeneous, but the X-linked form of the disease results from mutations in the gap junction channel connexin 32 (Cx32). It shows incomplete dominant inheritance, with heterozygous females being affected less severely than hemizygous males. The phenotype may vary from mild, in which the patient has a normal gait, to a severe form which may necessitate the use of a walking stick or wheelchair.

Over a hundred mutations in *CX32* have been identified. These fall into two main groups–those in which the protein never reaches the plasma membrane, and those where the protein reaches the membrane but forms channels with altered functional properties. The former give rise to a severe phenotype, whereas the latter may be associated with either mild or severe phenotypes, according to whether they partially or completely disrupt channel function.

The Cx32 protein is primarily expressed in the Schwann cells of peripheral myelinated nerves, at the nodes of Ranvier and at Schmidt–Lanterman incisures. In these regions, the myelin is not complete and there is a thin layer of cytoplasm between each of the enveloping turns of the Schwann cell. This suggests that Cx32 may serve as short-cut pathway for nutrients and other substances moving to the innermost layers of the Schwann cell, and perhaps also to the axon itself. This might explain why loss of Cx32 function causes axonal degeneration and demyelination.

Cardiac muscle channelopathies

Long-QT syndrome is a congenital cardiac disorder associated with an abrupt loss of consciousness and sudden death from ventricular arrhythmia in children and young adults. It is characterized by an abnormally long QT interval in the electrocardiogram, which reflects the delayed repolarization of the ventricular action potential. The duration of the cardiac action potential is determined by the balance between the inward and outward currents flowing during the plateau phase. Prolongation of the action potential can therefore be caused by a persistent inward current or by a reduction in outward K^+ currents.

Some six different genetic loci for Long-QT syndrome have been mapped, five of which have been identified and shown to encode cardiac ion channels. The I_{Ks} channel is a complex of two different proteins, KCNQ1 and minK, and mutations in these genes cause LQT1. Likewise, I_{Kr} is a complex of HERG and Mirp1, and is associated with LQT2. Mutations in these four genes either abolish, or markedly decrease, the repolarizing K^+ currents I_{Ks} and I_{Kr}, and are therefore expected to prolong the cardiac action potential and increase the QT interval. Mutations in the cardiac muscle sodium-channel gene (*SCN5A*) cause LQT3. These mutations affect Na^+-channel inactivation, producing a sustained inward current that results in an increased action potential duration. The larger the component of non-inactivating current, the more severe the phenotype.

In many cases, LQT syndrome is not inherited but acquired. For example, drugs that block I_{Kr} or I_{Ks} currents prolong the cardiac action potential and induce the Long-QT syndrome. Among these are class III antiarrhythmic agents such as sotalol, dofetilide, and quinidine, which selectively block I_{Kr}, and the antihistamine H_1-receptor antagonists terfenidine and astemizole, which block HERG. In most people, terfenidine does not produce cardiac problems as it is rapidly broken down in the liver and its metabolite, terfenidine carboxylate, does not block I_{Kr}. However, if the activity of the P-450 enzymes that break down terfenidine is impaired (due to liver disease or drugs such as ketoconazole and the macrolide antibiotics), there is a risk of *torsade de pointes.*

Skeletal muscle channelopathies

Myasthenia and slow-channel syndromes

Myasthenia gravis is produced by autoantibodies directed against the nicotinic acetylcholine receptor (**nAChR**), as discussed elsewhere. These antibodies lead to receptor internalization and thus to a smaller endplate potential that fails to reach the threshold for action-potential initiation.

Slow-channel syndrome (**SCS**) is a congenital myasthenia that results from mutations in the muscle nAChR channel. These mutations have been found in all four types of adult nAChR subunits (a, b, †, e) and result in protracted channel activation by acetylcholine. The increase in channel open probability produces a prolonged synaptic current and endplate potential. Consequently, temporal summation of endplate potentials can occur at physiological rates of stimulation, leading to prolonged depolarization of the muscle membrane, inactivation of voltage-gated sodium channels, and failure of muscle excitability. This explains why patients with SCS experience muscle weakness and rapid fatigue. A similar 'depolarization block' is observed with acetylcholinesterase inhibitors or with AChR agonists like suxamethonium.

The prolonged endplate potential also causes enhanced Ca^{2+} entry, which may account for the progressive destruction of the postsynaptic neuromuscular junction observed in SCS—loss of junctional nAChRs and destruction of the junctional folds has been reported. Abnormal channel openings in the absence of acetylcholine may also contribute to the 'endplate myopathy'. This may explain why the SCS mutations that cause spontaneous openings are often associated with a more severe phenotype.

In contrast to myasthenic gravis, the symptoms of SCS are exacerbated by acetylcholinesterase inhibitors and patients do not respond to immunotherapies.

The periodic paralyses

Hyperkalaemic periodic paralysis, paramyotonia congenita, and the potassium-aggravated myotonias result from mutations in the a-subunit of the human skeletal muscle Na^+ channel. All are inherited as dominant traits and usually present within the first or second decade of life.

Hyperkalaemic periodic paralysis (**HyperPP**) may occur spontaneously, but attacks are commonly precipitated by exercise, stress, fasting, or eating K^+-rich foods. Paralysis is often preceded by signs of muscle hyperexcitability such as myotonia or fasciculations. The duration is variable (minutes to hours) and may be so severe that the patient is unable to remain standing. It is associated with a raised blood K^+ concentration (5–7 mM). Paramyotonia congenita is precipitated by cold and (in contrast to most classical myotonias) aggravated by exercise. In some patients, the myotonia may be followed by prolonged paralysis. Potassium-aggravated myotonia is characterized by myotonia without muscle weakness or paralysis. It can be distinguished from classical myotonias by the fact that the myotonia is exacerbated by a mild elevation of the plasma K^+ concentration.

All three types of disorder result from mutations in the a-subunit of the skeletal muscle Na^+ channel (*SCN4A*), which disrupt Na+-channel inactivation. As a consequence, they produce a persistent inward current that causes a tonic depolarization of the muscle membrane (the larger the current, the greater the

depolarization). The magnitude of the depolarization determines whether myotonia or paralysis occurs. A small depolarization causes membrane hyperexcitability by lowering the action-potential threshold, whereas a large depolarization can lead to Na^+-channel inactivation and thereby paralysis. It is still not understood how cold or an elevated plasma potassium level precipitate attacks.

Myotonia

Loss-of-function mutations in the gene (*CLCN1*) encoding the skeletal muscle Cl^- channel produce two forms of myotonia—the autosomal dominant myotonia congenita (Thomsen's disease) and the autosomal recessive generalized myotonia (Becker's disease). Clinical descriptions of the disease can be found elsewhere.

In normal skeletal muscle, the Cl– conductance accounts for between 70 and 80 per cent of the resting membrane conductance. Mutations in *CLCN1* that result in a loss of functional Cl^- channels will therefore produce a marked increase in the input resistance of the muscle fibre. Consequently, muscle excitability will be enhanced (because a smaller Na^+ current will be sufficient to trigger an action potential). The elevated input resistance also produces a reduced rate of action potential repolarization, which enhances muscle excitability. An important role of the muscle Cl^- conductance is to counteract the depolarizing effect of K^+ accumulation in the transverse tubular (T-) system that accompanies muscle activity. During an action potential, K^+ ions leave the muscle fibre. In normal muscle, the amount of K^+ entering the T-system during a single action potential is not sufficient to alter the membrane potential, because the tubular Cl^- conductance is very high. But in myotonic muscle, the Cl^- conductance is very low and a small rise in tubular K^+ produces a significant depolarization following an action potential. If several action potentials occur in rapid succession, summation of the after-depolarizations may be sufficient to trigger spontaneous action potentials and thereby myotonia.

Mutations in *CLCN1* give rise to both recessive and dominant forms of myotonia. This may be because the muscle Cl^- channel comprises more than one subunit. In heterozygotes, mutant subunits might combine with wild-type subunits to form heteromeric channels. The extent to which the mutant subunit reduced the function of the heteromeric channel would thus dictate the severity of myotonia. Total inactivation of the channel by a single mutant subunit (the dominant-negative effect) would produce dominant myotonia, whereas recessive myotonia might occur if the heteromeric channel was unaffected by the mutant subunit.

Malignant hypothermia and central core disease

Mutations in the ligand-gated Ca^{2+} channel of skeletal muscle cause malignant hyperthermia and central core disease. This channel mediates Ca^{2+} release from the sarcoplasmic reticulum (SR), allowing Ca^{2+} to enter the cytoplasm and activate the contractile proteins. It is also known as the ryanodine receptor (or RYR1) because it binds the alkaloid ryanodine with high affinity.

Malignant hyperthermia (**MH**) is one of the main causes of death due to anaesthesia. In susceptible individuals, common inhalation anaesthetics or depolarizing muscle relaxants trigger accelerated skeletal muscle metabolism, muscle contractures, hyperkalaemia, arrhythmias, respiratory and metabolic acidosis, and a rapid rise in body temperature (as much as 1°C every 5 min). It is thought that this is due to stimulation of Ca^{2+} release from the SR, which produces a sustained increase in intracellular Ca^{2+}. This activates both metabolic and contractile activity; the former results in respiratory and metabolic acidosis and the latter produces the elevation in body temperature. The syndrome can be treated with dantrolene sodium, which blocks Ca^{2+} release from the SR. Malignant hyperthermia is genetically heterogeneous and is not linked to *RYR1* in all families.

Central core disease (**CCD**) is an autosomal dominant, non-progressive myopathy that presents in infancy as proximal muscle weakness and hypertonia. Diagnosis is by muscle biopsy, which reveals that regions of type 1 skeletal muscle fibres (known as 'central cores') are depleted of mitochondria and oxidative enzymes. The disease is often associated with a predisposition to malignant hyperthermia and results from mutations in *RYR1*. Thus CCD and MH are allelic disorders of the same gene. It is not clear how the different phenotypes arise, especially because the same mutation can give rise to MH in some individuals and CCD in others. Because all CCD patients are MH-susceptible, it is possible that additional factors are necessary for the development of central core disease.

Kidney channelopathies

Liddle's syndrome

Liddle's syndrome is a congenital form of salt-sensitive hypertension characterized by a very high rate of renal Na^+ uptake despite low levels of aldosterone, secondary hypokalaemia, and metabolic acidosis. It is caused by gain-of-function mutations in the epithelial sodium channel (**ENaC**). This channel consists of three subunits (a, b, g), and disease-causing mutations have been identified in both the b- and g-subunits. All are located in the C-terminus of the protein and result in constitutive channel hyperactivity.

The increase in ENaC current causes enhanced Na^+ uptake. This is accompanied by increased water uptake, thereby producing a chronic increase in blood volume and ultimately hypertension. An increased Na^+ uptake also has secondary consequences: in particular, K^+ secretion into the tubule lumen is stimulated because the apical membrane depolar*ize*s and so increases the driving force for K^+ efflux. In addition, more K^+ enters the cell due to the enhanced activity of the Na+/K+-ATPase. This explains why excess ENaC activity in Liddle's syndrome is associated with hypokalaemia and, conversely, why reduced ENaC activity (as in pseudohypoaldosteronism type 1 (**PHA-1**) disease) is accompanied by hyperkalaemia.

Pseudohypoaldosteronism type 1

While gain-of-function mutations in ENaC cause enhanced Na^+ uptake and hypertension, loss-of-function mutations produce salt-wasting, hypotension, and dehydration in newborns and infants. Pseudohypoaldosteronism type 1 (PHA-1) results from loss-of-function mutations in the a, b, or g ENaC subunits. The marked reduction in ENaC activity leads to decreased Na^+ absorption by the kidney. This stimulates renin and aldosterone secretion, but salt reabsorption cannot be augmented as ENaC is not functional. The high Na^+ concentration in the tubular fluid causes water to be osmotically retained in the tubule lumen, leading to diuresis and dehydration.

Bartter's syndrome

Bartter's syndrome is characterized by severe salt-wasting, with elevated plasma renin and aldosterone levels. The syndrome is both phenotypically and genetically heterogeneous, and several subtypes have been distinguished. Antenatal Bartter's syndrome or hyperprostaglandin-E syndrome presents *in utero* with a marked fetal polyuria. Newborns fail to thrive and show severe salt-wasting, moderate hypokalaemia, and metabolic acidosis, and elevated urinary excretion of prostaglandins. There is also marked calcinuria, osteopenia, and nephrocalcinosis.

Antenatal Bartter's syndrome results from loss-of-function mutations in the genes encoding proteins involved in salt transport in the cells of the distal kidney tubules. These include the inwardly rectifying K^+ channel Kir1.1 (*KCNJ1*; Bartter's syndrome type II), the NaK2Cl cotransporter (*SCL12A1*, Bartter's syndrome type I), and the voltage-gated Cl^- channel CLC-Kb (*CLCNKB*, Bartter's syndrome type III). These variants may be distinguished clinically, because hypokalaemia is less pronounced (3.0–3.5 mM) in patients with mutations in *KCNJ1*, and the course of the disease is less severe. And in contrast to patients with Bartter's syndrome types I and II, patients with mutations in *CLCNKB* do not suffer from nephrocalcinosis, despite elevation of the urinary calcium concentration.

Disease-causing mutations in Kir1.1 or CLC-Kb impair NaCl uptake in the distal tubules by impairing channel function or decreasing protein expression. This leads to a high salt concentration in the urine and thus to an osmotic diuresis, which accounts for the salt-wasting, polyuria, and low plasma volume characteristic of Bartter's syndrome. A similar phenotype is observed with loop diuretics, such as frusemide (furosemide), which inhibit the NaK2Cl cotransporter.

Nephrolithiasis

Mutations in the voltage-gated, renal chloride channel gene, *CLCN5*, cause Dent's disease, a congenital form of congenital nephrolithiasis. It is usually associated with proteinuria. Different mutations may produce phenotypically distinct syndromes (Table 1), but there is as yet no clear explanation for how this occurs. The mechanism by which mutations in a Cl^- channel impair calcium handling by the kidney is also not fully resolved.

Nephrogenic diabetes insipidus

Familial nephrogenic diabetes insipidus (**NDI**) results from impaired water uptake by the kidney tubules. The diseases manifests within the first few weeks of life and is characterized by the excretion of large amounts of hypotonic urine and excessive thirst. In early infancy these may not be noticed and the disease is often recognized by signs of dehydration, such as poor feeding, poor weight gain, irritability, and fever. In most cases, familial NDI is caused by a mutation in the vasopressin receptor, but in some families it results from loss-of-function mutations in the aquaporin 2 (*AQP2*) gene. AQP2 is expressed exclusively in the collecting duct of the kidney and plays a fundamental role in the production of a concentrated urine because it acts as a water channel. Vasopressin stimulates water uptake by causing the insertion of AQP2 channels into the apical membranes of the principal cells of the collecting duct, thereby enhancing water uptake. Loss-of-function mutations in *AQP2* result in a dramatic reduction in water channels, thereby accounting for the polyuria.

Other channelopathies

Cystic fibrosis

Of all the channelopathies, the best known is probably cystic fibrosis (**CF**). Its clinical features are described in Chapter 17.10. Cystic fibrosis results from mutations in an epithelial chloride channel known as the cystic fibrosis transmembrane conductance regulator (**CFTR**). Although its primary sequence is highly homologous to that of the ATP-binding cassette transporters, it is now well established that CFTR functions as a chloride channel. It also regulates the activity of the outwardly rectifying Cl^- channel and the epithelial Na^+ channel.

All disease-causing CF mutations result in the complete absence or a marked reduction in CFTR function. Those which result in the total loss of channel activity, either because the protein does not reach the plasma membrane or because it is present but completely inactive, give rise to a severe form of the disease. Mutations that result in a reduced Cl^- current are associated with a milder form of the disease. Compound heterozygotes carrying one allele with a severe mutation and another with a mild mutation will have significant residual channel activity and therefore a mild form of the disease.

While a large number of mutations (more than 450) have been identified in CFTR, it is still far from certain how the loss of Cl–-channel function gives rise to the clinical features of the disease, especially, in the lungs.

Congenital hyperinsulinaemia

Familial congenital hyperinsulinaemia (**CHI**) is characterized by an unregulated insulin secretion and profound hypoglycaemia that presents at birth or within the first year of life. Some patients respond to treatment with diazoxide, but in others the most effective treatment is resection of the pancreas (more than 90 per cent is usual). Many patients develop diabetes in later life.

CHI results from mutations in the genes encoding the pancreatic b-cell ATP-sensitive K^+ (K_{ATP}) channel. This channel plays a key role in glucose-stimulated insulin secretion. When the plasma glucose level is low (less than 3 mM), the channel is open and keeps the b-cell membrane potential at a hyperpolarized level. When plasma glucose levels rise, increasing glucose uptake and metabolism by the b-cell, the K_{ATP} channels close. This produces a membrane depolarization that activates voltage-gated Ca^{2+} channels, increases Ca^{2+} influx, and so stimulates insulin release. Two classes of therapeutic drugs modulate insulin secretion by interacting with K_{ATP} channels. Sulphonylureas inhibit channel activity and are used to enhance insulin secretion in patients with type 2 diabetes mellitus, whereas K-channel openers (for example, diazoxide) activate K_{ATP} channels, hyperpolarizing the b-cell and preventing insulin release.

The K_{ATP} channel consists of two types of subunit: a pore-forming subunit Kir6.2, and a regulatory subunit SUR1. Mutations in either subunit can cause CHI, but those in SUR1 are more common. CHI mutations result in the loss of K_{ATP} channel activity, even at low blood glucose levels. This results in a continuous depolarization of the b-cell, that leads to persistent Ca^{2+} influx and thus constitutive insulin secretion.

Non-syndromic deafness

About 70 per cent of all cases of prelingual deafness are non-syndromic. The disorder shows marked genetic heterogeneity, but in some families it results from loss-of-function mutations in the gene (*GJB2*) encoding the gap junction channel connexin 26. Both recessive and dominant mutations have been described. Connexin 26 (Cx26) is expressed in the cochlea, but the mechanism by which the lack of functional Cx26 leads to hearing loss remains obscure. In some individuals, mutations in Cx26 are associated with Vohwinkel's syndrome, a disorder characterized by keratoderma. Many patients also suffer from deafness.

Further reading

Ashcroft FM (2000). *Ion channels and disease* p.481. Academic Press, San Diego.

Lehmann-Horn F, Jurkatt-Rott K (1999). Voltage-gated ion channel and hereditary disease. *Physiological Reviews***79**, 1317

Websites

http://www.ncbi.nlm.nih.gov/omim/ [Online Mendelian inheritance in man (**OMIM**), a database of human genes and genetic disorders.]

http://www.neuro.wustl.edu/neuromuscular/mother/chan.html [A website concerned with neurological and CNS disorders, including those associated with ion channel defects, maintained by the Neuromuscular Disease Center at Washington University School of Medicine, St Louis, USA.]

http://www.ncbi.nlm.nih.gov/ [The National Center for Biotechnology Information, providing access to genetic sequence databases (e.g. GenBank)]

4.6 Apoptosis in health and disease

Andrew H. Wyllie and Mark J. Arends

Introduction

Apoptosis is the process by which single cells die in the midst of living tissues. It is responsible for most—perhaps all—of the cell-death events that occur during the formation of the early embryo and the sculpting and moulding of organs. It continues to play a critical role in the maintenance of cell numbers in those tissues in which cell turnover persists into adult life, such as the epithelium of the gastrointestinal tract, the bone marrow, and lymphoid system including both B- and T-cell lineages. It is the usual mode of death in the targets of natural killer (**NK**)- and cytotoxic T-cell killing, and in involution and atrophy induced by hormonal and other stimuli. It also appears in the reaction of many tissues to injury, including mild degrees of ischaemia, exposure to ionizing and ultraviolet radiation or treatment with cancer chemotherapeutic drugs. Excessive or too little apoptosis plays a significant part in the pathogenesis of autoimmunity, infectious disease, acquired immunodeficiency syndrome (**AIDS**), stroke, myocardial disease, and cancer. When cancers regress, apoptosis is usually part or all of the mechanism involved.

Structural changes in apoptosis

Apoptosis can be recognized because of its characteristic, stereotyped sequence of structural changes (Fig. 1). The dying cells lose contact with their neighbours. They undergo a rapid loss of volume, often of the order of 50 per cent. There is explosive blebbing from the cell surface, in which multiple cytoplasmic protrusions extend and are immediately withdrawn, and this is followed by fragmentation into a cluster of subcellular bodies (apoptotic bodies) each membrane-bounded and containing a variety of compacted cytoplasmic organelles. The nucleus undergoes similar distortion and fragmentation. Chromatin condenses under the nuclear membrane in granular aggregates with a knob-like, hemilunar or toroidal distribution. Nuclear membranes overlying residual uncondensed chromatin are rich in pores but these are absent adjacent to condensed chromatin, suggesting that redistribution takes place. The nucleolus segregates so that its argyophilic fibrillar centre lies close to the peripheral aggregates of chromatin, whilst the osmophilic particles that are associated with transcription complexes disperse in the central nucleoplasm. Eventually the nuclear membrane disappears and the entire nuclear remnant becomes a mass of condensed granular chromatin.

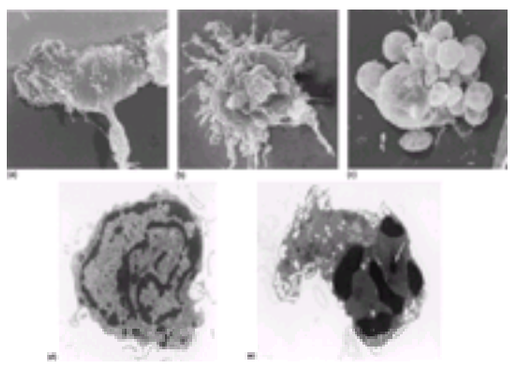

Fig. 1 The structure of apoptosis. (a) Scanning electron micrograph of a normal macrophage shows its surface sprouting many pseudopodia. In (b), the cell has been injured (in this case by oxidized lipid of the type often present in high concentration in atheromatous plaques) and is throwing out and retracting multiple surface blebs. In (c) the whole cell has fragmented into roughly spherical apoptotic bodies. Some of these are cratered by the orifices of the dilated endoplasmic reticulum. (d) Transmission electron micrograph of a thin section. The condensed chromatin (arrowheads), nucleolar remnant (arrow), and highly convoluted surface are clearly visible. The scale bar in μm. (Micrographs by courtesy of Dr Jeremy Skepper and Dr Jing Xia, Cambridge School of Biology Multi-imaging Centre.)

Within the cytoplasm, bundles of microfilaments often appear in a side-to-side configuration. Sometimes free ribosomes pack into semicrystalline arrays. There is dilatation of the endoplasmic reticulum. The cell surface loses any pre-existing microvilli or other indices of polarity. The shrunken cell and the apoptotic bodies into which it fragments tend to become spherical.

Isolated apoptotic cells lose the ability to maintain ionic homeostasis within an hour or so, lose density, swell in volume, and permit the entry of various dyes classically used to mark dead cells (such as Trypan blue and propidium iodide) to which they had been previously impermeable. Within tissues, however, this phase is seldom seen, because the apoptotic cell and its fragments undergo phagocytosis. Often this is undertaken by 'professional' phagocytes—the resident tissue macrophages—but where unusually large numbers of apoptotic cells are generated, other cell types share in ingesting them, including their viable neighbours. Once within the phagosome of the ingesting cell, the apoptotic cell and its fragments rapidly become indistinguishable from the contents of any other large secondary lysosome.

For reasons to be expanded later, the process of apoptotic-cell phagocytosis inhibits the neutrophil-dominated inflammatory reaction that is often seen when macrophages are activated in other circumstances. Cell loss by apoptosis can therefore be effected with little disruption of the tissue concerned. Moreover, apoptosis, once initiated, is completed swiftly. Although the interval from the initial application of a lethal stimulus to the first manifestations of shrinkage and blebbing can vary greatly, phagocytosis may be complete within an hour thereafter. Hence, the evidence for cell loss by apoptosis is transient and often surprisingly scanty relative to the reduction in cell number it effects.

Apoptosis is not the only mode of cell death. Dying cells sometimes show a very different pattern of change, dominated by volume overload and, eventually, plasma membrane breakdown and leakage of intracellular contents into the extracellular space. At first, the nucleus retains its general structure, although the chromatin patterns coarsen. Usually there is an associated acute inflammatory reaction. This pattern of death is frequently found when tissues are overwhelmed by high

concentrations of toxic substances or in severe ischaemic damage, where vascular perfusion has been arrested. Classically, it is called necrosis.

Later, following equilibration of the cytosol with extracellular calcium, and the resultant widespread activation of degradative enzymes such as cathepsins, vestiges of nuclear structure fade away (karyolysis) and only ghost-like cellular outlines remain.

Caspases: effectors of apoptosis

Many of the morphological features of apoptosis are attributable to activation of a family of proteases called 'caspases', because of the presence of the amino acid **c**ysteine in their catalytic site, and their preferential cleavage of peptides immediately C-terminal to **asp**artate residues. There are at least 12 mammalian caspases. All are initially synthesized as relatively inactive proenzymes and undergo proteolysis to generate two fragments of around 10- and 20-kDa molecular weight, together with a fragment of variable length from the original N-terminus (Fig. 2). These 10- and 20-kDa fragments oligomerize in pairs to form a tetramer, which is the active enzyme. Long N-terminal sequences provide the opportunity for regulation through interaction with various binding proteins.

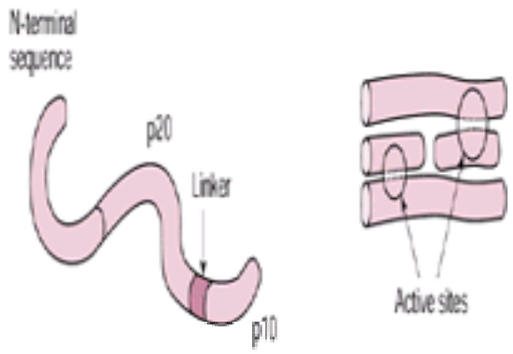

Fig. 2 Schematic diagram of caspase activation. The proenzyme is on the left. Following processing, as shown on the right, the N-terminal sequence and the linker are lost. The active sites of the enzyme each contain elements from both p10 and p20 subunits.

Caspases recognize 4-amino acid motifs that are present in many proteins. Significantly, such caspase target sites are often highly conserved between species, and frequently occur in strategic intramolecular locations, such that caspase cleavage would radically alter the function of the substrate protein. In particular, the cleavage of caspase substrates accounts for many of the structural changes of apoptosis already described. Particularly interesting substrates include proteases, structural proteins of the cytoskeleton, members of various signalling pathways, proteins involved in DNA damage and repair, and molecules of importance in the regulation of the cell cycle. Some caspase substrates appear to be intermediates in self-amplifying positive feedback systems required to complete the death process.

Caspases and proteases

For most caspases, the cleavage sites are themselves typical caspase target sequences, indicating that caspase activation might proceed by cascade-like autocatalysis. There is very good evidence for this, the initiator caspases with long N-terminal sequences (caspases 8–12 and probably 2, 4, and 5) being activated prior to the short effector caspases (3, 6, and 7). Caspases can also activate other proteases. There is, for example, a caspase site in the calpain-inhibitor protein, calpastatin. The inhibitor is rendered inactive by cleavage, so turning on calpain digestion within the dying cell.

Caspases and cytoskeletal proteins

Actin (the major protein of the cytoskeleton), fodrin (which provides the deformable shell underlying the plasma membrane), vimentin (an intermediate filament protein of the cytoskeleton), and the lamins (which form a major component of the nuclear envelope) are all caspase substrates. Caspase cleavage of these large polymeric proteins provides a means whereby they can be rapidly disassembled to the monomers of which they are composed. Gelsolin, a further caspase substrate, is an actin-binding protein that cleaves actin filaments in a calcium-dependent manner. Caspase cleavage of gelsolin separates the calcium-sensitive negative regulatory domain from the protease domain, and hence actin-filament cleavage is effected under normal intracellular calcium concentrations. These cytoskeletal proteolytic events probably contribute to the rounded shape of apoptotic bodies and to the eventual dissolution of the nuclear membrane.

Caspases and signalling proteins

The small G-protein, rho, regulates the mobility of the cell surface. Two rho-dependent kinases, PAK2 and ROCK-1, are rendered constitutively active by caspase cleavage, through excision of their negative regulatory domains. PAK2 activity is a factor in the early retraction of the apoptotic cell from its neighbours or from substrate attachment, whilst ROCK-1 activity is responsible for the enhanced action of a myosin light-chain kinase that drives the phase of cell-membrane blebbing that immediately precedes fragmentation of the apoptotic cell.

FAK^{p125} is the kinase associated with focal adhesion plaques. It is a critical element in the signalling pathway that links cellular awareness of substrate attachment (through integrins) to other cellular functions, including movement, attachment, and new transcription. FAK^{p125} is cleaved and inactivated by caspases, hence isolating the cell from such signals, many of which would normally promote survival. Somewhat similarly, the adenomatous polyposis coli protein (**APC**) and b-catenin are both elements in the wnt-1 signalling pathway, connecting cell-to-cell signals with regulation of cell function. Both are cleaved by caspases, although the precise significance of this is still speculative.

Certain members of the MAP kinase pathway are also targets of caspase cleavage. In particular, MEKK1 (MAPK/ERK kinase kinase; MAPK is mitogen activated protein kinase; ERK is extracellular signal-related kinase; all these enzymes form part of an autocatalytic cascade) a kinase upstream of the p38 stress-associated protein kinase, is activated through the removal, by caspase proteolysis, of an N-terminal inhibitory domain. As p38 itself initiates apoptosis, which would lead in turn to further activation of MEKK1, this could generate a positive feedback loop ensuring that the process of death is executed swiftly and effectively.

DNA damage and repair

ICAD (inhibitor of caspase-activated DNase) is a cytoplasmic chaperone that binds a double-stranded nuclease, **CAD** (caspase-activated DNase). The ICAD–CAD complex is normally cytoplasmic. ICAD, however, is a caspase substrate and once cleaved can no longer retain CAD in the cytoplasm. On translocation to the nucleus, CAD initiates the digestion of DNA, first to large fragments of around 50 kilobase pairs (**kbp**), and eventually—through cleavage of chromatin at internucleosomal sites—to a series of fragments that are multiples of the 180- to 200-bp unit wrapped around each nucleosome. These DNA fragments can be extracted from apoptotic cells and, on electrophoresis, produce the characteristic ladder pattern that historically was one of the distinctive signatures of apoptosis. Intranuclear DNA cleavage is still the basis of a variety of diagnostic methods that depend on the presence of large numbers of free double-stranded DNA ends. A further protein, acinus, also apparently activated by caspases, contributes to the extreme condensation of chromatin seen in the nuclei of apoptotic cells.

DNA-PK, ATM, PARP, and Rad51 are all DNA repair proteins concerned with the recognition and response to double-strand breaks. Significantly, all are cleaved in apoptosis at sites that separate their DNA-binding and catalytic domains, thus ablating their ability to effect non-homologous end-joining. This may be important in preventing re-ligation of the heavily digested DNA of the apoptotic nucleus, so avoiding the generation of large numbers of undesirable recombinant DNA molecules. Cleavage of PARP (poly-ADP-ribose polymerase) may have an additional function. PARP normally responds to DNA breaks by adding poly-ADP ribosyl tails. It is an abundant nuclear protein and has the capacity to exhaust cellular adenine nucleotide stores if presented with the large number of free DNA ends in the apoptotic nucleus. Apoptosis is an energy-requiring process, and it may therefore be advantageous to conserve adenine nucleotides, even during the process of death.

Caspases and cell-cycle proteins

Unexpectedly, several proteins that normally inhibit movement around the cell cycle are targets for caspase cleavage. These include p21 ^{WAF1} and p27^{KIP1}, inhibitors of cyclin-dependent kinases that catalyse movement through the G_1 and S phases of the cell cycle. Wee-1 is a further caspase substrate: normally it blocks movement from G_2 to mitosis. CDC27, a component of the anaphase-promoting complex which destroys cyclin B and hence inhibits entry into mitosis, is also subject to caspase cleavage. The purpose of this reactivation of elements of the cell cycle during the process of death is still quite obscure. It occurs during the apoptosis of cells such as neurones which have long since ceased movement around the cycle.

Pathways that activate caspases

Two well-documented pathways converge on and activate the effector caspases, one leading from cytokine-activated, death-signalling receptors on the cell surface, the other from mitochondria.

Death-signalling receptors coupled to caspase activation

The death-signalling receptors are all members of the tumour necrosis factor-a (**TNF-a**) receptor family. They are type 1 membrane receptors (that is, with the N-terminus on the external surface), containing a series of cysteine-rich incomplete repeats in the ligand-binding domain, a single transmembrane domain, and a cytoplasmic moiety with one or more signalling domains (Fig. 3). Their ligands are homologues of the cytokine TNF-a. The prototype death-signalling receptor is fas (also called CD95 or Apo-1). On binding its ligand, FasL, this receptor trimerizes and immediately recruits to its cytoplasmic moiety a cluster of proteins collectively called the death-initiating signalling complex (**DISC**). The aggregation of DISC proteins is the result of protein–protein interaction at an a-helical region called the death domain (**DD**), because without it fas signalling is ineffective. Through the DD, fas interacts with an adapter protein called **FADD** (fas-associated protein with death domain) that contains a further interactive region called **DED** (for death-effector domain). Through DED, FADD recruits procaspase 8 to the DISC, an initiator caspase with two DEDs in its N-terminal sequence. Because they are at high local concentration in the DISC, the procaspase 8 molecules can catalyse their own activation, and so initiate the proteolytic cascade that ultimately turns on the effector caspases. Whilst fas is widely expressed in many tissues, fasL expression is largely restricted to cytotoxic lymphocytes and to cells in immunologically privileged sites. In this way, the fasL/fas system plays a major role in cell killing by cytotoxic T-cells (**CTLs**) but can repulse CTLs at immunologically privileged sites. As will be described, upregulation of fas may play a much wider role in the sensitization of damaged cells to other types of apoptotic stimuli. Caspase 8 is not the only output from fas activation. A DISC component called daxx acts as an adaptor, connecting fas activation to the apoptosis signal-regulating kinase-1, **Ask-1**. This links fas signalling to the stress-activated kinases JNK and p38 (Fig. 3 and Fig. 4).

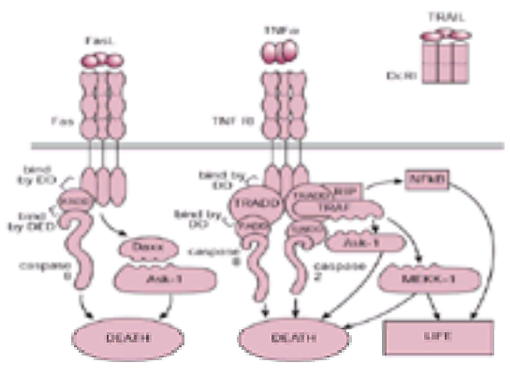

Fig. 3 Death-signalling receptors, shown diagrammatically. The fas receptor, with its ligand and DISC, signalling exclusively to death, is shown on the left. The more complex TNF-a receptor 1 is shown in the centre. Both prosurvival and proapoptosis signals can emanate from this receptor. On the right is shown one of the decoy receptors for **TRAIL** (TNF-related apoptosis-inducing ligand), DcR1. This receptor has no membrane anchor and so competes for TRAIL with the death-signalling membrane receptors, DR4 and DR5.

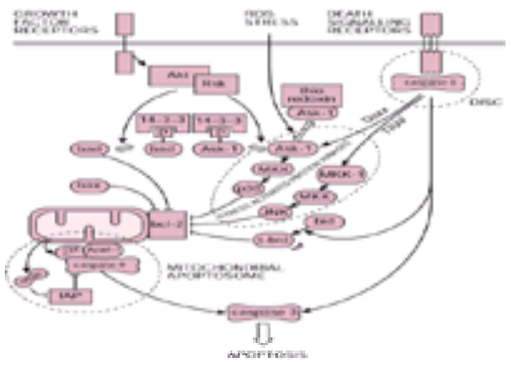

Fig. 4 Interaction between membrane receptors and mitochondrial signals coupled to apoptosis. In this summary diagram, proapoptotic elements are shown with rounded profiles, whilst prosurvival elements have rectangular borders. Cyt c, cytochrome *c*; t-bid, truncated bid.

TNFR1, the high-affinity TNF-a receptor, also trimerizes and signals to caspase 8 through FADD, in a DISC, but its action is substantially more complex. TNFR1 first recruits an additional DD-containing adaptor protein called TRADD. TRADD then recruits FADD, but can also bind other proteins into the DISC, including TRAF2 and a serine–threonine kinase called RIP. TRAF2 activates MEKK1 and Ask-1, elements of the MAP kinase pathway that transmit activation signals to JNK and hence the jun-containing transcription factor AP-1. RIP activates the transcription factor NF_B. NF_B provides a consistent survival signal, whilst JNK activation can be either prosurvival or proapoptosis. TNFR1 can also signal through yet another DD-containing adaptor protein, RAIDD, that in turn activates procaspase 2. Hence TNFR1-mediated signals can be interpreted as proapoptotic (through activation of caspase 8, caspase 2, or the p38 kinase), prosurvival (through NF_B), or ambivalent (through JNK and AP-1). This design means that incoming cytokine signals can be interpreted in opposite ways—for life or death—depending on precise conditions in the cell at the time.

DR3 is a receptor closely similar in structure to TNFR1 but with a narrower tissue distribution. Whereas TNFR1 is ubiquitous, DR3 is expressed predominantly in the lymphocytes of spleen, thymus, and peripheral blood. Interestingly, the expression of the ligands appears to adopt the opposite pattern, with TNF-a being a product predominantly of activated macrophages and lymphocytes, whereas the DR3 ligand (variously also called Apo3L and TWEAK) is expressed in many tissue types.

DR4 and -5 are similar receptors that bind a ligand called TRAIL (TNF-related apoptosis-inducing ligand). The downstream signalling appears to involve both FADD and caspase 8, although the possibility that certain cell types may couple DR4 and -5 to other signalling molecules is not excluded. Both TRAIL and its receptors are expressed in many tissue types. TRAIL has excited particular attention because it is frequently cytotoxic to tumour cells under conditions in which normal cells are unharmed. Variant receptors that lack the cytoplasmic signalling moieties are expressed in many normal tissues and appear to act as inhibitory decoys for TRAIL.

The death-signalling receptor pathway has the capacity to integrate several different types of environmental information in formulating its ultimate message to the caspases. Although its primary function is response, through the TNF-a family of receptors, to death-inducing ligands, its connection to the stress-activated kinases through Ask-1 also allows it to be sensitive to the redox status of the cell. Ask-1 binds to thioredoxin in a reversible, redox-sensitive manner. Hence the conditions under which Ask-1 is available for signalling are also redox-sensitive. Further, Ask-1 is inactivated, through binding to the chaperone 14–3–3, when serine phosphorylated. In this way the Ask-1 signalling pathway may be taken out of circuit by serine phosphorylation, as would pertain under conditions of growth-factor stimulation through raf kinase, PI3 kinase, and protein kinase B (Akt).

Mitochondrial signals coupled to apoptosis activation

The mitochondrial pathway depends upon the release of cytochrome *c*, together with dATP, from the intermembranous space of mitochondria. Cytochrome *c* and deoxyATP (**dATP**) bind to and effect a conformational change in a protein of the outer mitochondrial membrane, Apaf-1, so that it exposes a protein-binding domain (generically called a **CARD**, for caspase-activating recruitment domain) capable of recruiting and activating procaspase 9. This molecular assembly has been called the apoptosome, although other procaspase-containing complexes exist, including the DISC and probably others, referred to below. This then activates the effector caspases (Fig. 4). Triggers for the release of cytochrome *c* include reactive oxygen species, cellular redox stress, pH changes, and proteins of the BCL-2 family.

BCL-2 is a protein with a C-terminal hydrophobic domain that allows it to anchor to intracellular membranes, of which the outer mitochondrial membrane is one. It was first identified because of its consistent activation (through a chromosome translocation) in follicular B-cell lymphoma. In this sense, BCL-2 can function as an oncogene, but it differs from most other oncogenes in failing to stimulate cell proliferation—indeed, when expressed in isolation it inhibits cell-cycle progression. Its major role, however, is that of a survival factor, and thus it can cooperate with other oncogenes to sustain the life of clones of cells that otherwise might be deleted by apoptosis. This survival factor function appears to be a basic property of metazoan cells. The normal development of the nematode *Caenorhabditis elegans*, for example, depends on a BCL-2 homologue (ced 9), without which the embryo undergoes widespread apoptosis at an early stage. Transgenic mammalian BCL-2 can rescue the development of ced 9-deficient nematode larvae.

The mammalian BCL-2 family contains at least 15 members in three major branches, distinguished on the basis of their function, which can be either prosurvival or prodeath, and the presence or absence of certain conserved domains (called BH1–4). Amongst the prosurvival molecules are BCL-2 itself, BCL-xL, and BCL-w, all of which share all four BH domains, MCL-1 which lacks the BH4 domain, Boo which lacks BH3, and A1 which lacks both BH4 and BH3. In contrast, bax, bak, and bok form a branch of the BCL-2 family that possesses BH3, BH2, and BH1 domains but exerts prodeath functions. The third family branch consists of proteins whose sole region of homology with the others is a BH3 domain (sometimes amounting to no more than an 8-amino acid motif): bid, bad, bim, blk, hrk, and BNIP3. The BH1, BH2, and BH3 domains form part of a hydrophobic pocket in the molecule, into which another BH3 domain can fit in much the same way as a ligand binds to its receptor. Through this interdomain binding, BCL-2 family members can homo- and heterodimerize with each other. Whilst the prodeath action of the BH3-only proteins is known to depend upon such heterodimerization, the mechanism of action of the prolife and bax-related proteins is still uncertain.

Proteins in the prosurvival and bax-like prodeath branches of the BCL-2 family have a tertiary structure similar to bacterial haemolysins such as diphtheria toxin, suggesting that they could form transmembrane pore-like structures. Moreover, the majority of the members of the extended BCL-2 family possess C-terminal membrane anchors, and some (notably BCL-2 itself, BCL-xL, and bax under conditions of apoptosis activation) clearly localize to the mitochondrial outer membrane. Superficially at least, this could provide an explanation for the regulated escape of cytochrome *c* from mitochondria, through opening or closing of transmembrane channels formed from the BCL-2 proteins. This proposition is supported by the observation that BCL-2 family proteins can insert into artificial lipid membranes, thereby changing their electrical conductance. Further, the voltage gradient across the mitochondrial membrane (‡jm) collapses around the time of apoptosis activation. There remains some doubt, however, both over the capacity of such channels to permit the exit of molecules as large as cytochrome *c* and whether the relationship between the collapse of ‡jm and the initiation of apoptosis is cause or effect. Another model proposes that BCL-2 family members modify the function of the mitochondrial permeability transition pore (a multimolecular transmembrane structure containing a voltage-dependent anion channel and an adenine nucleotide translocator) leading eventually to rupture of the outer membrane.

BCL-2 possesses an important regulatory serine site that, on phosphorylation, neutralizes the prosurvival function. This critical phosphorylation can be effected by the stress-activated kinase pathways mentioned above, that lead from Ask-1 to activation of the JNK and p38 kinases. In this way, cellular stresses of many kinds, including exposure to reactive oxygen species and ultraviolet light, can negate the survival functions of BCL-2 family members and so swiftly lower the cellular threshold for apoptosis.

The BH3-only, prodeath members of the family appear to play important roles in coupling the powerful mitochondrial pathway to lethal stimuli in the cellular environment (Table 1). Notably, bid is activated through cleavage by caspase 8 of a small peptide from its N terminus. The truncated, activated bid translocates from the cytosol to mitochondria and effects the mitochondrial release of cytochrome *c*. In this way, stimuli too small to activate the effector caspases directly can be amplified by recruitment of the mitochondrial pathway through bid. Put another way, bid can lower the threshold at which cytokines trigger apoptosis. Somewhat similarly, bad is involved in a mechanism to raise the threshold at which apoptosis is engaged, depending on the availability of cytokine growth factors. The bad protein is phosphorylated by the kinases Akt (protein kinase B) and Rsk, both in turn dependent on PI3 kinase and the growth factors responsible for its activation. Normally, phosphorylated bad is sequestered in the cytoplasm by the chaperone 14–3–3. In conditions of growth-factor deprivation, however, unphosphorylated bad becomes available, translocates to the mitochondria, and activates cytochrome *c* release. BNIP3 is a mitochondrial protein that accumulates under conditions of hypoxia. It may thus provide a trigger linking hypoxia to apoptosis. Normally, bim binds to the light chain of dynein, a cytoskeletal protein. It may act as a sensor, triggering apoptosis under conditions of cytoskeletal disruption.

Perhaps surprisingly, BCL-2 itself is a caspase substrate: caspase 9 can cleave it to produce a C-terminal peptide with proapoptotic function. This is a further example, along with activation of the mitochondrial pathway through caspase truncation of bid and activation of MEKK1 as a proapoptotic stress-responsive pathway, that emphasizes the subtle strategy embodied in the caspase cascade. Mechanisms of this type would have no meaning if activation of initiator caspases inevitably led to commitment to death through the effector caspases. Rather, the initiator caspases can adjust a threshold, influenced by and integrated with many other incoming stimuli, that determines whether the death sentence will be enacted or repealed. Cells under stress of various types may start nearer that threshold, but the system is designed to process and interpret the continuous arrival of potentially lethal stimuli from around and within the cell.

Additional pathways for caspase activation

Despite the importance of the mitochondrial and death-receptor pathways in activating the effector caspases, it is clear that other means of caspase activation exist. Procaspase 12 localizes to the endoplasmic reticulum, whilst procaspase 2 can be found in the Golgi apparatus and the nucleus. The nucleus also contains CARD-proteins, that are recruited to foci of DNA damage, and BCL-2 is anchored to nuclear and ER membranes as well as mitochondria. Thus the death and survival of the cell, as modulated by caspase activation, may depend upon the synthesis of signals arising at many intracellular sites: some resulting from cell injury, others reflecting physiological stimuli. Damage to nuclear DNA is a particularly important source of injury-related stimuli for caspase activation.

A remarkable set of DNA-binding nucleoproteins is responsible for the recognition of DNA damage of different types, and for the initiation of repair. Thus, separate molecular mechanisms exist for responding to the presence of inappropriately inserted bases (base excision–repair, **BER**), nucleotides that have become modified through crosslinking or the formation of covalently bonded adducts (nucleotide excision–repair, **NER**), nucleotide mismatch or abnormal methylation (mismatch repair, **MR**), and double-strand breaks (homologous recombination or non-homologous end-joining, **HR** or **NHEJ**). In MR, MSH-2 and MLH-1 are recruited sequentially into a molecular complex at the injury site, which activates p53, effects cycle arrest, and, in the meantime, initiates repair at the site of damage. Similarly, amongst the first molecules to bind to DNA double-strand breaks in NHEJ are the DNA kinases ATM, ATR, and DNA-PK. In turn, these recruit and activate p53 and other molecules (for example, chk1 and chk2). In surviving cells, these effect arrest at a variety of points around the cell cycle, so ensuring that there is an opportunity to load the repair machinery on to the damaged DNA template before this is further altered by passing through DNA replication or chromatid separation.

A profoundly different means of limiting the effect of genome damage, however, is to commit the damaged cell to apoptosis; activation of the repair complex in both MR and NHEJ can also do this. The molecular basis for the decision between apoptosis or survival with repair is still unknown: indeed, there are circumstances in which repair is activated despite clear evidence that the decision to engage the machinery of apoptosis has already been taken. Activation of p53 is common to both outcomes, and it is therefore reasonable to search in this molecule for clues to the nature of the life or death decision. Activated p53 alters the transcription of a large number of genes. Some are well-known inhibitors of cell-cycle progression, such as p21waf1/cip1, but others (for example bax, fas, and a membrane protein called PERP) are associated exclusively with apoptosis. The situation is further complicated by the fact that p53 also signals to the apoptosis effector process by non-transcriptional means, via an N-terminal sequence that does not appear to be instrumental in effecting cell-cycle arrest. Phosphorylation provides one of the critical signals for p53 activation, and there are several different phosphorylation sites that respond preferentially to the various kinases. Thus the precise phosphorylation status of p53 could provide a molecular signature indicative of the nature, and perhaps the outcome, of the DNA damage.

Another potential factor in controlling DNA injury is a kinase (called **DAP** kinase because it was originally discovered as a death-associated protein) that influences the selection of $p14^{ARF}$ rather than $p16^{INK4A}$—alternative splice forms from the same gene. Whereas $p16^{INK4A}$ is a cell-cycle regulator, inhibiting the cyclin-dependent kinases, $p14^{ARF}$ displaces p53 from its inhibitor, MDM2, so generating a sustained p53 signal that may favour apoptosis. Curiously, DAP kinase normally docks on to the actin cytoskeleton. It is possible therefore that it affects the threshold at which apoptosis is triggered by DNA injury, depending on cytoskeletal-related factors such

as cell-to-cell and cell-to-substrate contacts.

Further clues to the way in which a variety of factors may be integrated in the response to DNA injury have emerged from detailed study of the injured nucleus. Within an hour of DNA damage, very large molecular complexes form around the damaged site. The first arrival is the phosphorylated histone gH2AX, which may recruit later members. These complexes contain molecular species that appear to generate platforms on which many injury-response proteins can associate, including p53, ATM, MSH-2, and Rad50 and -51. One type of platform is constructed from a protein called **PML** (so-called because of its abnormal synthesis in association with promyelocytic leukaemia). PML polymerizes to form intranuclear bodies (called PML bodies) into which p53 and many other proteins in the DNA-injury response are recruited. Indeed, nuclei without PML cannot mount the expected p53-dependent responses to DNA damage, even though p53 itself is available. Another such injury-related subnuclear body (the BASC body) incorporates the large DNA helicases BRCA1 and -2, proteins often defective in inherited susceptibility to breast cancer. These molecular platforms may provide the means whereby complex decisions, including the appropriate response to injury, are informed and initiated.

The replicative status of the cell is a further important determinant of its sensitivity to apoptosis following DNA injury. The proto-oncogene c-*myc* is normally amongst the earliest gene products to be synthesized when cells are stimulated by growth factors to leave quiescence and enter their replicative cycle. Thereafter, continuous expression of c-*myc* is required to sustain repetitive re-entry to the cycle following each cell division, rather than return to quiescence. Paradoxically, however, c-*myc* expression is also a powerful factor lowering the threshold for apoptosis. In particular, c-*myc* expression without concurrent molecular evidence of external growth-factor stimulation (such as PI3 (phosphatidylinositol-3) kinase and Akt activation) is interpreted as a death signal. Similarly, other early regulators of cell-cycle entry, including inhibition of function of the retinoblastoma protein and the release of the transcription factor E2F-1 from its binding pocket, also trigger apoptosis in the absence of concurrent evidence of external mitogenic stimulation. Perhaps this represents a means whereby tissues are protected from autonomous cell replication: survival of replicating cells is made conditional on the presence of appropriate stimuli in the cellular environment. The benefits of removing cells that show a tendency for such replicative autonomy are obvious, but the precise mechanism that couples replication to death except in acceptable circumstances is far less clear. It seems probable that the cell cycle itself includes checkpoints at which the decision to engage the apoptosis machinery can be taken should any of the appropriate conditions for replication be absent. Indeed, it is possible that injured cells may force the activation of such checkpoints as one way to access their apoptosis programme. This might explain the paradoxical activation of cyclin-dependent kinases by caspases in cells such as neurones that normally do not engage in replicative cycles at all, as mentioned earlier.

Inhibitors of caspase activation

The role of the BCL-2 family proteins in the activation and inhibition of apoptosis has been described, but there are other powerful endogenous inhibitors of caspase-associated cell death. One is FLIP, a DED-containing version of procaspase 8 that lacks caspase activity. High local concentrations of FLIP compete with procaspase 8 for recruitment into the DISC and so inhibit further propagation of death signals originating from the TNF family of receptors.

The heat-shock proteins hsp70 and hsp90 inhibit caspase processing and block mitochondrial cytochrome *c* release. The hsp90 forms a complex with Apaf-1, which is required for the inhibition of procaspase 9 processing, and is reversed by exposure to lethal doses of DNA-damaging agents.

IAPs (inhibitors of apoptosis proteins) inhibit caspase activity after autocatalytic processing of the procaspase has begun. All contain an element called a BIR domain, that binds to the N-termini of the short fragment of partially processed caspases, in such a way that adjacent elements of the IAP molecule drape across the caspase active site and sterically hinder substrate attachment. There are several such proteins—IAP1 and -2, ILP, the neuronal NAIP, and an X-linked family member X-IAP, all of which possess several BIR domains, and livin and survivin which contain a single BIR domain. One manifestation of the importance of IAPs is the presence of an IAP inhibitor, variously called smac or DIABLO, which is released from mitochondria along with cytochrome *c* during caspase activation by the mitochondrial pathway. The inhibitor smac has an N–terminal sequence that competes with partially processed caspase for the binding site in the BIR domain, and so allows the caspase to escape from the inhibitory embrace of the IAP (Fig. 5).

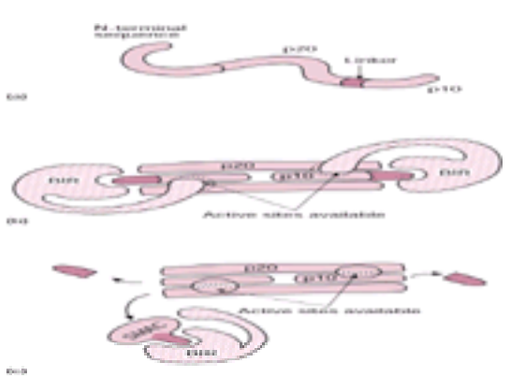

Fig. 5 Representation of the mode of action of IAPs. Following partial processing of the caspase, shown in (a), the IAP binds through its BIR domain, which identifies the newly exposed N-terminus of the linker region between the p20 and p10 caspase fragments. Thus bound, adjacent parts of the IAP molecule drape across the caspase active site (b). The inhibitor smac displaces the BIR domain, by presenting as a bait its own N-terminus, which is closely homologous to that of the partially processed caspase (c). With its active site now empty, the caspase can complete its processing and becomes fully active.

The IAPs provide a further example of the extraordinary interconnections between the cell cycle and cell death. Survivin, apparently associated with caspase 9, forms a complex with and is phosphorylated by active **cdk1** (cyclin-dependent kinase-1) during mitosis. Loss of phosphorylation leads to dissociation of the survivin–caspase-9 heterodimer, activation of caspase 9, and apoptosis. As normal mitosis proceeds, survivin associates with kinetochore proteins, the spindle microtubules, and finally, at cytokinesis, with the mid-body. Complexes of survivin with cyclin-dependent kinases active earlier in the cycle (for example, cdk4) have also been identified and promote transit through G_1. Thus, survivin may form part of the cell-cycle checkpoint system postulated earlier, providing a means whereby apoptosis can be activated by abnormalities in progression through the cell cycle. Finally, IAPs are themselves potential substrates of caspase attack, transactivators of the survival factor NF_B, and downstream products of NF_B-directed transcription. They thus form part of positive-feedback systems for both survival and death.

Recognition of apoptotic cells

Macrophages recognize and bind to the surface of apoptotic cells by virtue of multiple molecular 'eat me' signals (Fig. 6). The disposition of phosphatidyl serine (**PS**) residues on the apoptotic cell surface is one of the most characteristic of these. Normally PS appears only on the inner leaflet of the cell membrane, but this strict polarity is lost very early in apoptosis: around the time of rounding up, substantially earlier than chromatin condensation and DNA cleavage, and probably prior to evidence of caspase activation. Macrophages possess a PS receptor that binds to the exposed PS residues. The exposed PS residues may also bind to molecules in the extracellular environment that then form linkers to receptors on macrophage surfaces. Thus thrombospondin helps bind PS on the apoptotic cell surface to b_1, b_3, and b_5 integrins on the macrophage surface. Similarly, the complement fragment iC3b links to macrophage b_2 integrins, whilst the near-ubiquitous extracellular molecule b_2 glycoprotein-1 links to a macrophage receptor specific for it. In the same way, extracellular complement component C1q links specific binding sites on the apoptotic cell surface to receptors on the macrophages. A group of scavenger receptors (SRA, CD36, CD68, LOX-1) may tether directly to poorly defined oxidized lipid groups (similar to those in oxidized low-density lipoproteins) exposed on the surfaces of apoptotic cells. CD14 on the macrophage binds to ICAM3, exposed on apoptotic cells. Endogenous macrophage surface lectins also bind to sugars (such as *N*-acetyl glucosamine) selectively exposed on apoptotic cell membranes.

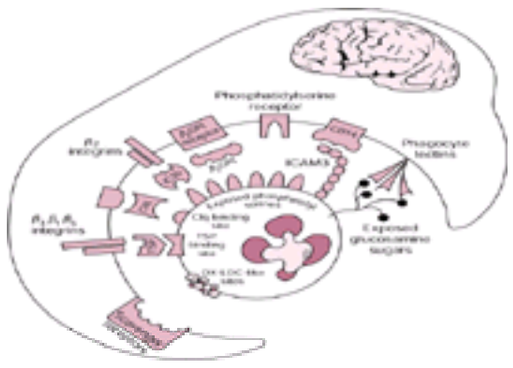

Fig. 6 Receptor–ligand interaction in the recognition of apoptotic cells by macrophages.

A distinctive feature of macrophage binding to apoptotic cells is the concurrent effect on macrophage function. Macrophages that engage in the phagocytosis of particles opsonized by immunoglobulin or complement component C3b engage in a sharp increase in oxygen usage (the respiratory burst), generation of reactive oxygen species and nitric oxide, and the release of inflammatory cytokines such as TNF-a. these recruit other acute inflammatory cells to the site. In contrast, macrophages that ingest apoptotic bodies show suppression of proinflammatory responses, mediated through the release of different cytokines, such as TGF-b. The basis of these contrasting effector responses appears to be the different signalling pathways that are activated by the macrophage receptors engaged by apoptotic bodies as opposed to opsonized particles.

Are caspases necessary and sufficient for cell death?

Although caspase activation appears to play a dominant role in the effector phase of apoptosis, it is clear that it is not responsible for all the phenomena of apoptosis. One striking example is the surface exposure of 'eat me' signals described above: this is not affected by pharmacological blockade of caspases using inhibitors that are unquestionably effective in blocking other aspects of apoptosis. These inhibitors are usually molecules that jam the caspase catalytic site through presentation to it, in uncleavable form, of a motif similar to the caspase target tetrapeptide. Inhibitors of this type, applied to lethally injured cells *in vitro*, often prolong cellular lifespan, but seldom totally reverse the cellular commitment to death. Moreover, developmentally programmed cell death can sometimes occur on schedule in embryonic tissues in which caspases have been inhibited, or key members of the caspase activation system (such as Apaf-1) are rendered deficient through germline gene knockout. In all these circumstances, the morphology of the caspase-free death is not that of apoptosis. The nuclei swell rather than undergoing the widespread condensation of chromatin. The cytoplasm shows signs of fluid overload, sometimes with the formation of dramatic fluid-filled vacuoles. Some of these changes are reminiscent of necrosis rather than apoptosis. Rather similar changes take place during the developmental death of phylogenetically ancient multicellular organisms that do not possess recognizable close homologues to the caspases, such as the slime mould *Dictyostelium discoides*.

These observations suggest that caspase activation, although intrinsic to the subtle and highly co-ordinated death process recognized as apoptosis, may not be the only event that commits cells to die. The existence of at least one caspase-independent death pathway is highlighted by a flavoprotein released from the mitochondria of injured cells called 'apoptosis-inducing factor' (**AIF**). AIF translocates to the nucleus, where it can effect chromatin cleavage to large fragments, but not the extreme condensation observed in apoptosis. It also appears to reproduce the cellular volume overload described above, even in the presence of caspase inhibition. Phylogenetically it is found in bacteria and plants as well as invertebrate and vertebrate animals.

Apoptosis and disease

There are few disease processes in which apoptosis does not feature, but the examples below are chosen because they exemplify how various steps in the apoptosis pathways may be critical for, or are subverted in, the course of disease pathogenesis.

Immunity and its disorders

Apoptosis is used extensively in the normal function of the immune system to facilitate the process of clonal selection. Antigen stimulation of T-cell proliferation is usually followed by expression of both fas and fasL, a recipe for apoptosis on a grand scale (called activation-induced cell death, **AICD**) unless there is rescue by a survival stimulus. This can be provided by co-stimulation from the immediate environment—adhesion molecules or cytokine receptors. A particularly important route for co-stimulation is through CD28, a receptor on T cells for signals transmitted from antigen-presenting cells, which increases the expression of several cytokines and their receptors. Similarly, clonally expanded populations of stimulated B cells in the bone marrow or those undergoing affinity maturation in lymph-node follicle centres are deleted by fas signalling, but can be selectively rescued by co-stimulation through CD40.

Cytotoxic T cells (CTLs) kill their targets by delivering to them the contents of their granules. Amongst these are perforin, which creates regions in the target-cell membrane of enhanced permeability at the points of contact with the CTL, and granzyme B, a protease that directly activates the caspases of the target cell. In this way, CTLs induce target-cell apoptosis.

The importance of apoptosis for the normal function of the immune system is underscored by the effects of genetic defects. Strains of mice with deficiency in the fas or fas ligand (called *lpr* and *gld*, respectively) show similar immunological phenotypes, characterized by massive lymphoproliferation and autoimmune disorders. The human homologue is the rare condition of the Canale–Smith syndrome (childhood autoimmune lymphoproliferative syndrome or **ALPS**) in which there is a mutation in the DD of fas. Inherited deficiency in C1q also leads to an autoimmunity syndrome: affected individuals almost always develop systemic lupus erythematosus. The pathogenesis here appears to be ineffective recognition and phagocytosis of endogenous apoptotic cells, so that their intracellular antigens are inappropriately processed.

Infective disorders

Shigella dysentery is due to pathogenic strains of *Shigella flexneri*. Pathogenicity is conferred by plasmid-borne genes that neutralize the primary host defence: phagocytosis and destruction of the bacteria by macrophages in the intestinal lamina propria. The plasmid-encoded protein Ipa B activates macrophage caspase 1, so annihilating the defence by inducing macrophage apoptosis. This strategy appears to be successful, because the bacterium that would normally be destroyed if it persisted within the phagosome of the ingesting macrophage, can escape from the cytoplasm of macrophages that undergo apoptosis.

The initial response to *Trypanosoma cruzi*, the parasite responsible for Chagas' disease, is dominated by T-lymphocyte activation. The resultant AICD generates a population of apoptotic lymphocytes. These impinge upon the macrophages that, suitably armed by proinflammatory cytokine stimulation, would be one of the most effective elements in the host defence against the parasite. As described earlier, sustained macrophage phagocytosis of these large numbers of apoptotic cells leads to suppression of proinflammatory cytokine release. The parasite subverts this aspect of the physiology of apoptosis into a source of protection from the host-defence reaction.

Viruses engage with the machinery of apoptosis in many ways. Even lytic viruses have strategies designed to conserve the life of their host cells for some time. DNA viruses, in particular, require means to abort apoptosis, as they must activate the cellular DNA synthesis machinery in order to replicate their own genomes, yet must then avoid the apoptosis that would otherwise follow DNA synthesis unaccompanied by commensurate external stimuli. The *E6* gene of human papillomavirus 16 (**HPV16**) encodes a protein that targets p53 for ubiquitination and subsequent degradation, and so permits cellular survival as the viral *E7* gene inactivates Rb and initiates entry into S-phase. The transforming genes of adenoviruses pair up to effect rather similar outcomes: whilst E1A binds Rb and initiates DNA synthesis, the 55-kDa subunit of E1B binds and inhibits p53, and the 19-kDa subunit neutralizes proapoptotic members of the BCL-2 family. Human herpesviruses such as HHV8 encode their own version of FLIP (v-FLIP). They also have their own prosurvival BCL-2 family members, such as BHRF1 in the Epstein–Barr virus (**EBV**) and KS-BCL2 in HHV8. The HHV8 strategy is particularly subtle, because the virus also destroys the endogenous BCL-2. Unlike endogenous BCL-2, this viral surrogate lacks an internal caspase site, and cannot be converted into a killer peptide by caspase cleavage. Baculovirus encodes a 35-kDa protein with BIR domains that is a prototypical IAP.

Apoptosis plays a key role in the pathogenesis of AIDS. The progressive loss of circulating CD4+ T cells, by which the course of HIV-1 infection to clinical AIDS can be charted, involves numbers of cells that are several orders of magnitude greater than the numbers that ever carry the virus. It is therefore clear that the overwhelming majority of the dying cells must be bystanders, sensitized to apoptosis by the presence of infection but not infected themselves. Viral proteins, released from infected cells, effect this sensitization by several parallel routes. The HIV proteins Tat and Nef induce fas, fasL, and TRAIL. Tat alters the cellular redox equilibrium in a manner that may activate Ask-1. Vpr binds to the mitochondrial permeability transition pore. A type of AICD may be induced by stimulation of CD4 and the cytokine receptor CXCR4 (both of which bind HIV epitopes). In infected cells, however, Nef inhibits ASK-1, and so may selectively protect these from apoptosis. Rather similar mechanisms underlie the deletion of neurones in HIV-associated dementia.

Cardiovascular disease

Pathogenetic mechanisms that interface with apoptosis are relatively poorly understood in cardiovascular disease, but there are several observations of potential relevance. Laminar flow inhibits Ask-1 in endothelium, whilst the generation of reactive oxygen species (**ROS**) induces the p38 and JNK stress kinase pathways. Thus, turbulence and the presence of ROS generators such as oxidized low-density lipoproteins—both known risk factors in the genesis of atheroma—are liable to promote apoptosis in endothelium. Other elements of the vascular wall are also abnormal in atheroma. Vascular smooth muscle cells from atheromatous vessels express p53, induce fas, and undergo apoptosis in increased numbers, particularly in the shoulders of the plaque, thus weakening attachment of the fibrous cap and rendering plaque rupture more probable. Macrophages also undergo apoptosis in response to the oxidized lipids that are present in atheromatous plaques. Death of the lipid-filled macrophages (foam cells) produces extracellular depots of oxidized lipid in the plaque core, a key step in plaque progression.

Whilst necrosis is the pattern of the cell death that immediately follows episodes of infarction, there is now substantial evidence that apoptosis occurs in the surrounding tissue over several hours thereafter, probably in response to relative ischaemia and the local generation of ROS. In animal models of stroke, this apoptosis can be downregulated by a variety of manoeuvres, including caspase inhibition, with objective evidence of improved cerebral function. These observations have generated enthusiasm for the development of antiapoptotic drugs for use following stroke and myocardial infarction. Another approach, potentially applicable to ischaemic myocardium, is to promote angiogenesis, perhaps by the use of angiogenic stem cells. Experimental models suggest that this improves the remodelling of the peri-infarct tissue, including decreased apoptosis of myocytes and improved cardiac function.

CNS degeneration

Despite the importance of the subject, there is still much doubt over the role of apoptosis in the chronic degenerative disorders such as Alzheimer's and Parkinson's diseases. Much of the problem stems from the relative inaccessibility of the brain for sequential studies following injury. In both conditions there is clear evidence of a loss of neurones, and those that remain accumulate abnormal cytoplasmic material, such as presenilins 1 and 2, and amyloid protein Ab in Alzheimer's disease. Cell culture and animal models suggest that the presence of these proteins may induce oxidative stress, which can lower the threshold for apoptosis. The protective effect of BCL-2 and caspase inhibition has also been recorded. The difficulties are compounded by the fact that neurones that undergo severe overstimulation (for example, by local high concentrations of the neurotransmitter glutamate) can also be induced to die (a phenomenon called excitotoxicity), but it is not clear whether the pathways involved overlap with or are identical to those of apoptosis.

Genetic studies of the inherited disorder, spinal muscular atrophy, provided what might have been the most definitive evidence linking the apoptotic pathways to CNS degeneration. A nearly consistent germline mutation has been identified that involves the IAP neuronal apoptosis inhibitory protein, **NAIP**. Unfortunately, another gene of quite different properties is also mutated in a high proportion of cases and it is still not clear which is responsible for the progressive loss of spinal motor neurones in this condition.

Tumour biology

Malignant tumours almost invariably show evidence of genomic instability. This commonly manifests itself as a tendency to undergo repeated episodes of chromosome breakage and recombination, or of mutation at microsatellite sites. The latter is due to defective function in the mismatch repair genes such as *MSH2* or *MLH1*, but there is less certainty over the basic causes of the former. Both can be associated with the loss of a p53-controlled checkpoint, chromosome instability more consistently so than microsatellite instability. Since MSH-2, MLH-1, and p53 are all connected to the activation of apoptosis, it follows that cancer cells inappropriately survive checkpoints normally controlled by these molecules. This gives cancer cells the opportunity to explore the consequences of genomic rearrangements that are denied normal cells. Some of these prove incompatible with continuing life but others lead to selective growth advantage (Fig. 7).

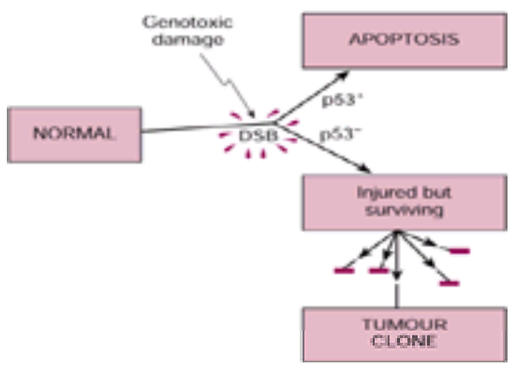

Fig. 7 Failure to activate apoptosis following damage by a genotoxic carcinogen, because of the absence of functional p53, leads to the inappropriate survival of clones of cells bearing double-strand breaks (DSB) and illegitimate recombination events. Whilst some of these clones may fail to proliferate further, others survive to become the founder clones of tumours. Constitutionally, these survivors have unstable genomes, as on further exposure to similar genotoxic stimuli they will again fail to enact apoptosis. Whilst the example given is for cells lacking p53, and hence unable to respond appropriately to DNA DSBs, exactly the same argument applies to cells that fail to identify nucleotide mismatch through defective MSH-2 or MLH-1. Such cells sustain extremely high mutations rates, as mismatches occur (and are normally recognized and repaired) in the course of normal DNA replication, even in the absence of genotoxic carcinogens.

An apoptotic view of carcinogenesis has implications for tumour management. The earlier sections of this chapter have indicated how proapoptotic pathways can cooperate. The activation of p53, for example, increases fas expression and this in turn, perhaps through daxx, Ask-1, and p38 kinase, sensitizes cells to the action of many xenobiotics. By contrast, however, cells lacking a critical signal that couples DNA damage to apoptosis (in this case p53) may often survive further damage by these agents. Whatever other changes occur in the course of tumour progression, the entry to malignant behaviour is likely to involve the loss of one critical link to the apoptosis effector pathway rather than loss of the pathway itself. Thus there are few cancer cell lines, and no reported primary tumours, in which, for example, caspase activity is absent, and levels of caspases appear to bear no relationship to tumour chemoresistance. Restoration of the links that couple DNA damage and other cellular stresses to the effector pathways of apoptosis is thus a real objective for cancer drug discovery.

Further reading

Adams JM, Cory S (1998). The Bcl-2 protein family: arbiters of cell survival. *Science* **281**, 1322–5. [These three articles are part of a special issue in *Science* devoted to the biology of apoptosis, and include reference to many of the historic breakthroughs in this rapidly expanding subject.]

Bennett MR, Boyle JJ (1998). Apoptosis of vascular smooth muscle cells in atherosclerosis. *Atherosclerosis* **138**, 3–9. [A review, with new findings and speculation on therapeutic implications, on the role of apoptosis in atherogenesis.]

Cohen O, Kimchi A (2001). DAP-kinase: from functional gene cloning to establishment of its role in apoptosis and cancer. *Cell Death and Differentiation* **8**, 6–15. [A review of the discovery of a new family of death-regulating proteins, by the group leader concerned.]

Evan G, Littlewood T (1998). A matter of life and death. *Science* **281**, 1317–21.

Freire-de-Lima CG, *et al.* (2000). Uptake of apoptotic cells drives the growth of a pathogenic trypanosome in macrophages. *Nature* **404**, 904–6. [This paper shows how *T. cruzi* exploits macrophage handling of apoptotic cells to its own advantage, and how this might be approached therapeutically.]

Green DR, Reed JC (1998). Mitochondria and apoptosis. *Science* **281**, 1309–11.

Hengartner MO (2000). The biochemistry of apoptosis. *Nature* **407**, 770–6. [A racy introduction to some detailed cell biology. This article, together with those marked with an asterisk, form part of a special issue devoted to somewhat visionary reviews of apoptosis.]

Hilbi H, *et al.* (1998). Shigella-induced apoptosis is dependent on caspase-1 which binds to IpaB. *Journal of Biological Chemistry* **273**, 32895–900. [A good source paper on bacterial pathogenicity factors that interact with apoptosis effectors.]

Kang PM, Izumo S (2000). Apoptosis and heart failure. *Circulation Research* **86**, 1107–13. [Helpful literature-rich review of developing area.]

Kocher AA, *et al.* (2001). Neovascularization of ischemic myocardium by human bone-marrow-derived angioblasts prevents cardiomyocyte apoptosis, reduces remodelling and improves cardiac function. *Nature Medicine* **7**, 430–6. [A glimpse of the future in the use of circulating endothelial stem cells to repopulate tissues damaged by ischaemia.]

*Krammer PH (2000). CD95's deadly mission in the immune system. *Nature* **407**, 789–95.

Lin Y, *et al*. (2001). Laminar flow inhibits TNF-induced ASK1 activation by preventing dissociation of ASK1 from its inhibitor 14–3–3. *Journal of Clinical Investigation* **107**, 917–23. [A good introduction to ASK-1, with relevance to endothelial apoptosis.]

Mills JC, Stone NL, Pittman RN (1999). Extranuclear apoptosis: the role of the cytoplasm in the execution phase. *Journal of Cell Biology* **146**, 703–8. [A detailed account of the mechanisms underlying the classical structural changes of apoptosis.]

*Nicholson DW (2000). From bench to clinic with apoptosis-based therapeutic agents. *Nature* **407**, 810–16.

Ojala PM, *et al.* (2000). The apoptotic v-cyclin—CDK6 complex phosphorylates and inactivates Bcl-2. *Nature Cell Biology* **2**, 819–25. [Critical new information on phosphorylation of Bcl-2, in the context of the subtle strategy employed by HHV8 to ensure its own survival.]

*Savill J, Fadok V (2000). Corpse clearance defines the meaning of cell death. *Nature* **407**, 784–8.

Wellington CL, Hayden MR (2000). Caspases and neurodegeneration: on the cutting edge of new therapeutic approaches. *Clinical Genetics* **57**, 1–10. [Good review of potential role of apoptosis in CNS degenerative disease.]

Wyllie AH, Golstein P (2001). More than one way to go. *Proceedings of the National Academy of Sciences, USA* **98**, 11–13. [A brief review of non-caspase death.]

*Yuan J, Yanker BA (2000). Apoptosis in the nervous system. *Nature* **407**, 802–9.

Zhou B-B, *et al.* (1998). Caspase-dependent activation of cyclin-dependent kinases during fas-induced apoptosis in Jurkat cells. *Proceedings of the National Academy of Sciences, USA* **95**, 6785–90. [A classic paper on the controversial but intriguing subject of cell-cycle activation in apoptosis.]

5.1 Principles of immunology

A. J. McMichael

Introduction

There are two features which distinguished immune responses from the non-specific defence mechanisms such as inflammation. The first is the specificity of the reaction, which is easiest to appreciate in terms of antibody responses but it is also true of the cellular immune responses; an essential part of this specificity is the remarkable ability to distinguish between self and non-self. The second is memory by which a second challenge with a stimulus provokes a more rapid and more vigorous immune response. These concepts began to be formulated by Jenner and were developed by Pasteur, Erlich, Landsteiner, Medawar, Burnet, and many others. In the last 30 years, the cellular and molecular basis of these two characteristics has become much clearer.

Immune reactions play important roles in most disease processes. Much of what is observed at the bedside involves immune responses, although the visible, palpable, or audible end-result can be quite distant from the primary event. End-immune reactions can be divided into two: those dependent on antibody (humoral responses) and those on T lymphocytes (cell-mediated immune responses, CMI). Antibody reactions themselves are normally quite silent *in vivo* (e.g. neutralizing virus infectivity), but sometimes antibodies may trigger various secondary events that become literally visible (e.g. anaphylaxis) or revealed on investigation (e.g. haemolysis). Cell-mediated immune responses may also be silent (e.g. clearing of some virus infections by T cells lysing infected cells or releasing interferon) or visible (e.g. delayed hypersensitivity reactions in the skin) or revealed by investigation (e.g. kidney graft rejection).

These two different types of immune response are indicative of the basic division of lymphocytes into two types, B and T cells. They interact with each other and with a third important cell, the antigen presenting cell. The following sections explain how these cells work, as far as possible at a molecular level.

Antigens

Both B and T lymphocytes make the fundamental distinction between self and non-self. This is quite remarkable; for example each species will make antibody to cytochrome c of other species but not self, even though they are closely related proteins differing in very few amino acids. Similarly, T cells respond to all HLA (transplantation) antigens of other members of the species but not self. The mechanisms that underlie this self tolerance are complex and will be explained later. First, the nature of antigens recognized by B and T cells need to be examined, because there are important differences.

Antigens recognized by B cells

B lymphocytes recognize antigen through their surface antibody receptors. In chemical terms, the recognition is identical to that by secreted antibody. The antibody reacts with the native proteins. The availability of large amounts of purified antibodies, particularly monoclonal antibodies secreted by hybridomas, has facilitated detailed analysis of how protein or carbohydrate macromolecules are recognized. By inhibiting antibody reactivity with polysaccharides with small sugars, Kabat was able to show that an antibody binds to a small part (epitope) of a macromolecule, about the size of seven sugars. For protein antigens, knowledge of their sequence has helped to define their epitopes. A good example is the haemagglutinin (HA) of influenza A virus. The amino acid sequence and three-dimensional structure of this molecule are known. In addition, the amino acid sequence of variant HA molecules that react with different antibodies are known. It is possible, thereby, to locate the parts of the molecule that bind to antibody. Four have been identified in this way, situated on the outside of the globular head of the molecule, each involving less than ten amino acids; because the antibodies bind to a folded molecule these amino acids are not in continuous sequence. Thus for both protein and polysaccharide the epitopes that react with antibodies and B cells are discrete and small.

A protein antigen must be greater than about eight amino acids in size to stimulate an immune response. It is possible, however, to make antibodies to smaller molecules if these are coupled to larger carrier proteins. In this way antibodies can be made to virtually all drugs and other small chemical (hapten) groups. This implies a further, very special feature of the humoral immune response; it can recognize molecules that do not occur in nature or that have not been previously synthesized. This is because of the spatial configuration of many small molecules may be similar in the sense that they fit antigen binding sites of antibodies. Such molecular mimicry has been invoked to explain some autoimmune reactions, for instance the cross-reaction between streptococcal and cardiac antigens in rheumatic carditis.

B cells can react with antigen in its natural form, and special processing is not required. However, the signal delivered to B cells by antigen is on its own insufficient to activate them to proliferate and secrete antibody. Growth and differentiation factors must be provided by helper T lymphocytes. As discussed below, T lymphocytes respond to processed antigen. Certain carbohydrate antigens, however, which have multiple repeating units that can function as epitopes, can stimulate B cells directly to divide and secrete IgM antibody. Some polysaccharide carbohydrate antigens facilitate this by non-specifically activating cells to which they bind (lipopolysaccharide from Gram-negative bacteria for example).

Antigens recognized by T cells

In principle, T cells can recognize as antigen the same range of molecules that are seen by antibodies. However, they are more fastidious about the nature of antigen and its presentation; they do not respond to soluble or free antigen, but to antigen at cell surfaces. On the presenting cells the antigen has to be associated with histocompatibility antigens, HLA (inappropriately standing for human leucocyte antigens) in humans. As foreign antigens are invariably presented *in vivo* with self HLA antigen, the T cells show specificity for self HLA plus foreign antigen. T cells so activated will not recognize the same foreign antigens if presented with foreign HLA *in vitro*. However, foreign HLA antigens provoke very strong immune reactions and these alloreactive T cells do not have to see foreign HLA in association with self HLA. There is therefore something very special about T cell recognition of HLA antigens.

Foreign protein antigens are presented to Th cells (helper T cells that will stimulate B cells to make antibody) by specialized antigen presenting cells. When the parts

of the protein molecule that T cells respond to are analysed it is clear that these are often different from those seen by antibody. Using the same example as above, the T cells can recognize parts of the influenza A virus HA that are buried in the molecule. They respond well to short (15–20 amino acid residues) synthetic peptides that represent fragments of the known amino acid sequence. Experiments have indicated that if antigen presenting cells (macrophages) are lightly fixed with paraformaldehyde they will successfully present peptide fragments but not whole protein antigens to helper T cells. All this implies that helper T cells see digested fragments of protein antigens, rather than native proteins, in association with self major histocompatibility complex (MHC) antigens. The role of the antigen presenting macrophage or monocytes is to capture protein, or other antigen, ingest it, and present it in processed form on its surface in association with HLA antigen (Fig. 1). B lymphocytes, by binding antigen to their antibody receptors, can also process antigen and present to Th cells.

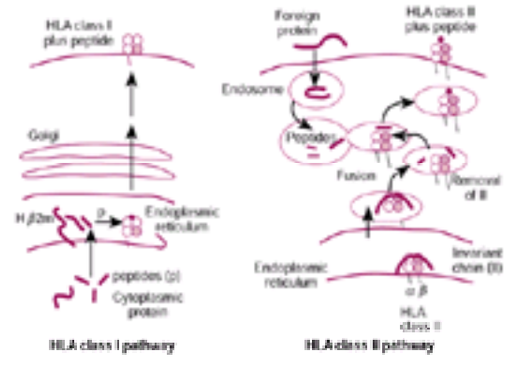

Fig. 1 Antigen processing: the two pathways are shown, through HLA class I on the left and through HLA class II on the right. For the class I route, cytoplasmic (e.g. virus) proteins are degraded to peptides in the cytosol. The peptides are transported to the lumen of the endoplasmic reticulum. Here, nonamer peptides bind to the newly synthesized class I molecules, stabilizing folding with b_2-microglobulin. The stable complex is translocated to the cell surface where foreign peptides, as well as self peptide, are displayed. For the class II route, foreign protein antigen is ingested by the cell and degraded to peptides in the endosome. The newly synthesized class II molecules fold in the endoplasmic reticulum and the groove is protected by binding to a polypeptide, 'invariant chain'. As the complex formed is transported towards the cell surface the invariant chain is digested away by proteases exposing the groove. HLA class II-containing endosomes fuse with peptide-containing vesicles and peptide binds to the class II HLA groove, stabilizing the structure. The HLA class II molecules reach the cell surface and display foreign peptides. (For further details see Townsend A and Trowsdale J (1993). The transporters associated with antigen presentation. *Seminars in Cell Biology* **4**, 53–61.)

The other chief subset of T cells, the cytotoxic T lymphocytes, see foreign antigen at the surface of non-specialized presenting cells, particularly virus antigen at the surface of infected cells. Cytotoxic T cells recognize peptide fragments of virus antigens at the surface of target cells bound to HLA class I molecules. In this form, antigen would not be recognized by antibody specific for the intact protein. This suggests a reason for this apparently complicated process of T-cell recognition. If T cells could react with native antigen free from cell surfaces they would be inactivated because the receptor would be engaged but the secondary signals needed for T cell activation, provided by the antigen presenting cell, would be lacking.

Histocompatibility antigens

These antigens, HLA in humans, play a central role in T lymphocyte function because, as described above, they associate with foreign antigen to stimulate T cells. HLA antigens are controlled by the major histocompatibility complex (MHC) which is a cluster of genes on the short arm of chromosome 6 (Fig. 2). There are two types of MHC antigen: class I and class II.

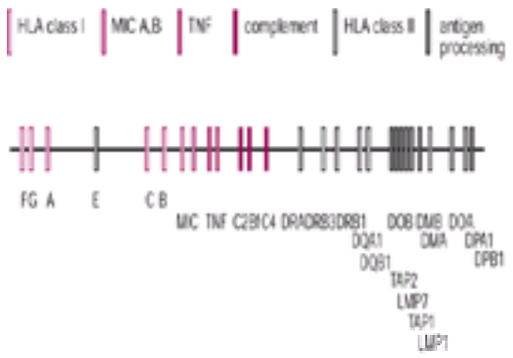

Fig. 2 Gene map of the HLA complex on the short arm of chromosome 6, showing the immunologically important genes. Shown are the HLA class I loci, the related MIC A and B loci, genes relevant to inflammation including the complement components C2, C4, and Bf, the HLA class II loci, and genes relevant to antigen processing. The map is based on that published by Aguado *et al*. 1999.

Class I antigens are dimers with a 45 000 molecular weight heavy chain, encoded in the major histocompatibility complex (MHC) of genes on the short arm of chromosome six, and an invariant light chain, b_2-microglobulin (b_2m), that is coded on chromosome 15. In 1987, the three-dimensional structure of HLA A2 was determined by Bjorkman, Strominger, Wiley, and colleagues, a finding that has had a profound impact on our understanding of T cell recognition. The heavy chain is divisible into three extracellular domains of about 90 amino acids, of which the membrane-proximal a3 domain resembles an immunoglobulin domain. Together with b_2-microglobulin, which is also immunoglobulin like, these form a stalk on which sits a structure made up by the a-1 and -2 domains. The molecule is folded to form a groove, bounded on its sides by two alpha helices and on its floor by an eight-stranded beta sheet; nearly all of the amino acids that differ between different class I molecules contribute to the fine structure of this groove. The ends of the groove are closed and contain tyrosines and threonines that are conserved in nearly all class I molecules. The crystal structure showed that the groove contains bound peptides, now known to be derived from degraded cytoplasmic proteins. Because different class I molecules (HLA types) have different shaped grooves they bind different peptides. The peptides are usually nine amino acids in length and are bound as extended chains with the amino- and carboxy-termini bound in pockets at either end of the groove. Between two and four of the amino acids that make up the peptides are involved in binding to the HLA molecule with their side chains fitting into other pockets in the groove. These anchoring residues are different in different class I molecules; for instance all peptides that bind to HLA B27 have arginine at position two, those binding to HLA A2 usually have leucine at this position, compared to proline for peptides binding to HLA B35. The class I molecules on a given cell probably bind several hundred peptides, sharing the common anchor residues; these give the cell a 'signature' that is monitored by T cells which are tolerant to normal self peptides. If a foreign peptide, derived from an intracellular parasite or mutated self protein enters the system, T cells can react and destroy the cell. The class I molecules thus serve the function of displaying abnormalities within the cell at its surface.

Most nucleated cells express the classical class I HLA antigens (HLA-A, B, and C) although trophoblast is negative and very low amounts seem to be expressed on hepatocytes, muscle cells, and nerve cells. Some tumours are negative, which may be one way in which they evade T-cell immunity. However, expression on many cell types is increased by the action of g-interferon released by activated T cells.

HLA class I heavy chains are encoded in the MHC. Products of two loci, HLA A and B, are expressed on cell surfaces in large amounts (10–100 000 molecules per cell); a third series, HLA C, is expressed at much lower levels. There are, in addition, three loci that express non-classical HLA class I molecules. HLA G is expressed almost exclusively on extra villous trophoblast and probably serves to inhibit natural killer cells and macrophages that might attack the foreign trophoblast. HLA E also inhibits natural killer cells; its surface expression is controlled by the availability of a particular peptide derived from the leader sequence of the classical HLA A, B, and C class I proteins, as well as HLA G. HLA E binds to a receptor CD94-NKG2 on natural killer (NK) cells that delivers inhibitory signals. Other receptors on NK cells (killer inhibitory receptors, KIR) detect the presence on classical HLA class I molecules, particularly HLA C, on the potential target cell but deliver an inhibitory signal that protects the target from lysis. Cells that fail to express HLA, such as certain tumour cells selected to evade T cell attack, are lysed by NK cells. Thus the

HLA class I molecules are intimately involved in the maintenance of cell survival.

The HLA A, B, and C antigens are highly polymorphic with multiple alleles at each locus (Table 1). As their structure implies, class I antigens present peptide fragments of foreign antigens, usually virus, on the surface of normal cells to cytotoxic T lymphocytes (CTL). The polymorphism of HLA A, B, and C means that, when tested *in vitro*, cytotoxic T cells (primed to virus *in vivo*) will only recognize virus antigen on self cells or HLA matched cells (HLA restriction). It is likely that the extreme degree of HLA polymorphism results from evolutionary selection of multiple HLA alleles by pathogens. Individual HLA molecules show varying efficiency in presenting particular antigens to CTL, thus affecting the quality of the immune response to various intracellular pathogens.

HLA class II proteins have two chains, a and b. Each is around 30 000 daltons molecular weight and is composed of two extracellular domains (Fig. 2). At least four different members of the family are expressed on cells and most are polymorphic. The best studied series are the DR antigens, which have a highly polymorphic b-chain (encoded by the DRB1 locus) giving rise to over 20 DR types (Table 1). There is a second, less polymorphic DR-b chain that gives the specificities DR 51, 52, and 53, encoded by the DRB3, DRB4, and DRB5 loci respectively. The DQ molecules are polymorphic in both a and b chains, which means that hybrid molecules can occur in heterozygotes. The DP molecules are polymorphic in the b-chain and are expressed in smaller amounts. The HLA class II molecules are expressed on a limited set of cells that includes B lymphocytes, monocytes, dendritic cells, activated T cells, and some endothelial and epithelial cells. They can be induced on many more cell types by interferon-g.

The crystal structure of HLA DR1 is remarkably similar to HLA class I. The main difference is that the groove is open-ended allowing the bound peptides to hang out; thus the peptides are longer, 12 to 16 amino acids. Like the peptides that bind class I, there are anchor residues, determined by the fine structure of the groove. There is a particularly prominent pocket in the floor at the left hand end of the groove which binds large aromatic side chains in some instances.

The function of the class II antigens is to present foreign processed antigen to helper T cells. As discussed above, this process is mediated by specialized antigen presenting cells, although the wider expression of class II antigens induced by interferon suggests that other cells may be able to process and present under certain circumstances. A particularly important antigen-presenting cell type is the dendritic cell; this expresses large amounts of HLA class II as well as important costimulatory molecules. Dendritic cells appear to be particularly good at processing antigen for both the class I and class II pathways.

Human individuals differ in the epitopes of an antigen that they recognize, according to their HLA class II type. The effects of this could be manifest, for instance, in the degree of cross-reactivity between related viruses that stimulate antibody responses. Similarly, HLA class II type could determine whether antigens that differ from self in only a few epitopes are recognized. Self antigens that are slightly altered could fall into this category and it is striking that several organ-specific, autoimmune diseases are associated with HLA DR3. However, the exact reasons for this and other HLA and disease associations (Table 2) remain unresolved. It should be noted that the disease associations imply either a direct role for the HLA molecule or that there is a nearby disease-susceptibility gene in linkage disequilibrium with the HLA marker. In the case of HLA-DR3, there is strong linkage disequilibrium with HLA B8 and everything in between, including a complement C4 null allele and at least 50 other genes, most of which are non-polymorphic and many of unknown function.

The action of HLA class II antigens in presenting antigens to helper T cells gives them a key role in initiating immune responses. In contrast, class I antigens are involved at a later and more specialized stage. This is important in transplantation, where matching class II antigens is more important than matching class I, although the final damaging activity of cytotoxic T cells and antibody is directed at class I antigens.

Antigen processing

There are two pathways of antigen processing, generating the peptides that bind to HLA class I and II molecules, recognized by cytotoxic T lymphocytes (CTL) (which carry the CD8 surface marker) and Th (CD4 positive) respectively.

As indicated above, class I HLA molecules present short peptides derived from cytoplasmic proteins. The most likely pathway by which this happens is given in Fig. 1. Proteins in the cell are naturally turning over, broken down by cytoplasmic proteases. The multicatalytic proteasome complex is a major contributor to this process. It degrades cytoplasmic proteins that have been coupled to ubiquitin and digests them to small fragments. It is made up of 28 components; two of these are encoded in the MHC in the class II region, are interferon-g inducible and they may affect the protease specificity. The peptides generated are transported by an ATP-dependent transporter (TAP, transporters associated with antigen processing), made up of two chains encoded in the class II region of the MHC, into the endoplasmic reticulum. Here the peptides are degraded further unless they bind to newly synthesized class I HLA molecules. The latter take the peptide to the cell surface bound in the groove. Expression of most class I molecules at the cell surface is dependent on the integrity of this pathway. Patients described recently who have a deficiency in the TAP, display very low levels of HLA class I molecules on their cell surfaces; interestingly, they suffer from recurrent bacterial infections and a severe granulomatous vasculitis with highly activated NK cells.

Processing of antigens to be presented by class II HLA takes place in specialized antigen-presenting cells, particularly monocytes, dendritic cells, and B lymphocytes. Protein antigens are taken into the cell from outside. Monocytes may take up antigen passively by endocytosis or by receptor-mediated uptake, acquiring proteins bound through complement or Fc receptors. The latter bind to the constant regions of antibodies and so mediate uptake of proteins bound to cells as immune complexes—antigen may bind to small amounts of 'natural antibody' or the early specific antibody (giving an early positive feedback). Also B lymphocytes with specific antibody receptors can bind foreign proteins directly, endocytose, and process antigen to present peptides bound to class II molecules. Digestion of proteins takes place in an endosome compartment; these fuse with class II HLA-carrying vesicles bringing class II molecules from the endoplasmic reticulum. The newly synthesized class II molecules do not bind peptides in the ER; instead they bind to a third chain the 'invariant chain'. This protects the groove from peptide binding and the complex is exported through the Golgi complex where glycosylation is completed and addition of sialic acid residues to both the class II molecules and the invariant chain takes place. In a late endosomal compartment, at low pH, the invariant chain is degraded and removed from the class II molecules. This process involves another HLA class II molecule, HLA DM that somehow facilitates the exchange of invariant peptide for foreign peptide. Binding of peptide stabilizes the class II structure and the class II molecule containing the peptide reaches the surface. As for class I HLA, several peptides derived from normal proteins are displayed and elicit no T cell response because of tolerance. Foreign peptides that stimulate T cells with appropriate receptors initiate Th responses.

Priming the immune response

It is becoming clear that antigen presentation by class I or class II HLA on its own is insufficient to initiate a T-cell immune response. Specialized cells play a role, particularly B lymphocytes, dendritic cells, and others of the monocyte series. In addition to presenting peptides on their MHC molecules, they display other accessory molecules on their surface important for cell–cell interactions. In addition to the adhesion molecules, such as LFA-1 and CD2 that are also necessary for T cell recognition of target cells, the B7 molecule (CD80 and not to be confused with HLA B7) seems especially important. This binds to the T-cell marker CD28 and delivers a signal to naive T cells initiating the T-cell response. Once activated in this way these T cells can interact with antigen-presenting cells, such as virus-infected epithelial cells, and react; from here on, CD28 may be less crucial though it can inhibit activation-induced cell death and so ensure the responding T cell remains alive. The importance of the CD28–CD80 pathway is that there can be cells in the body that display antigenic peptides but because they lack CD80 do not initiate an immune response and suffer immune attack. However, a virus infection could sometime initiate a cross-reacting T-cell response and trigger an autoimmune response.

Antibodies

Structure

The basic structure of an antibody molecule is illustrated in Fig. 3. An IgG molecule consists of four chains, two identical heavy chains (50 000 daltons) and two identical light chains (25 000 daltons). The immunoglobulin molecule can be broken into segments by enzymes, giving the peptides illustrated in Fig. 3. Sequence analysis of light and heavy chains and crystallographic studies have revealed that they are composed of domains of about 100 amino acids, held by a disulphide loop between two cysteines (a domain structure that has been found to be present in many other cell surface molecules). There are two for each light chain (L) and four for an IgGg heavy chain (H). The N terminal domains of both L and H chains are highly variable when different antibody molecules are compared; they contain the antigen binding site. The constant domains of the heavy chain define the isotype of the antibody, IgG, A, D, M, or E, each of which has particular functions. The main differences between the various heavy chain constant regions are described in Table 3. The isotypes determine essential properties of the antibodies, particularly binding of Clq of the complement system, binding of Fc receptors that enable antibody to cross the placenta, and binding of Fc receptors on macrophages, mast cells, and basophils. IgA binds to a specific Fc receptor, known as a secretory component, which is on epithelial membranes. The antibody can then be endocytosed and transported across these cells and released on the outside, that is the gut lumen, biliary tract, respiratory tract, or milk duct. Light chain constant regions are one of

two classes, _ or l, but there are no known differences in function between the two.

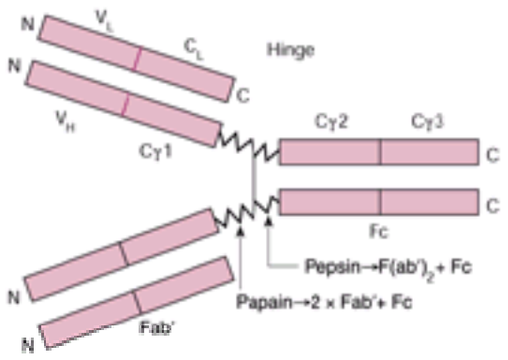

Fig. 3 Basic structure of an immunoglobulin, IgG, molecule. Two identical light (L) and two identical heavy (H) chains are shown. At the amino terminal (N) there is a variable (V) domain of 110 amino acids. In the light chain (L) this is connected to one constant domain of 100 amino acids which is one of the two types: l or _. The heavy-chain variable domain links to three constant-region domains, Cl1, Cl2, and Cl3. Between Cl1 and Cl2 is the hinge region, which gives the binding site (VL+VH) flexibility and contains interchain disulphide (—S—S—) bridges (see Table 3). The site of cleavage by the proteolytic enzymes papain and pepsin are shown, which give the antigen-binding fragments Fab and F(ab)$_2$ and the constant fragment Fc.

The N-terminal domains of L and H chains are variable regions. They vary between different antibodies to a remarkable degree. Sequence comparison of several variable regions has shown that there are three short, hypervariable patches between amino acids 28–34, 45–56, and 91–97 in both light and heavy chains. The crystalline structure of antibody indicates that these hypervariable regions form a surface to which antigen binds. In agreement with the studies on antigen epitopes, this is about the right area to bind to six amino acids on the surface of the antigen. Sequence analysis has also revealed that there are hundreds of genes for variable regions, explaining some of the enormous variability. Estimates, by various methods, agree that an individual must be capable of generating over several million different antibody molecules, probably nearer 10^8. As the variability is controlled by both the L and the H chains, which appear to associate independently, this means that there should be several hundred VL and VH sequences. How these are generated is discussed below.

Polyclonal or monoclonal antibodies

A normal antibody response includes multiple antibody molecules that bind to each epitope on a given antigen. B-cell activation requires binding of antigen, so any B cell that binds with affinity above a certain threshold stimulates B cells (with appropriate signals from helper T cells) and ultimately antibody secretion. A single antigen epitope can evoke a response comprising several hundred antibody molecules. This means that each antibody must be capable of binding more than one antigen, which is not surprising considering that antibodies recognize surfaces rather than sequences. An antiserum made up of hundreds or thousands of antibodies maintains its specificity because the immunogen is the common denominator. As each antibody molecule is the product of a single clone of B lymphocytes, and many are involved, this type of response is known as polyclonal (Fig. 4).

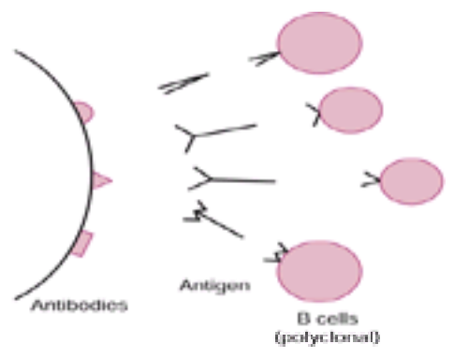

Fig. 4 Polyclonal antibodies. An antigen such as a protein has many antigenic sites (epitopes) on the surface of the molecule. To each can bind a set of antibodies with a range of affinity depending on the amino acid sequence of their variable regions. These are indicated diagrammatically by the shape of the binding site that binds to epitope. Each species of antibody is the product of one clone of B lymphocytes that can only make antibody with that shape, that is V-region sequence.

Under certain circumstances a monoclonal antibody response is generated. This may happen when the stimulating antigen is limited in its presenting epitopes or when it is present in extremely low amounts, thus selecting out only cells with the receptor with the highest binding avidity. Oligoclonal responses are also found in the cerebrospinal fluid in multiple sclerosis, possibly as a consequence of limiting the immune response to the few clones of cells that manage to cross the blood–brain barrier. More commonly, a monoclonal antibody *in vivo* is the product of abnormal proliferation of a single clone of B cells or plasma cells.

Kohler and Milstein, in 1975, devised a way of fusing normal (mouse) B lymphocytes to cultured plasmacytoma cells to generate immortal hybrid cell lines (hybridomas) that secrete the antibody of the B-cell parent. Because the donor of the spleen cells can be immunized at will, these monoclonal antibodies can be generated to any antigen. They have been used to explore the structure of human cells, particularly their surfaces, because unlike polyclonal antibodies made in another species, each monoclonal antibody reacts with a single component of the immunogen. They also have a multitude of practical applications in the study of growth factors, cell surface receptors, micro-organisms, and many other antigens. They have become routine tools for many diagnostic assays and are increasingly therapeutic applications are being developed, for instance the use of antitumour necrosis factor-a (anti-TNF-a) in the treatment of rheumatoid arthritis or the use of antibodies to respiratory syncitial virus to treat severe bronchiolitis in infants. For these clinical applications, it is possible to transplant the antigen binding sites of mouse antibodies into human immunoglobulins, retaining the antigen specificity. Antibodies that are 'humanized' in this way are less likely to provoke an immune response in recipients of the treatment.

Genetics of antibody production

The development of methods of DNA cloning, the use of the polymerase chain reaction, and sequencing have had a big impact on our understanding of immunoglobulin genes. The genes are arranged as shown in Fig. 5 on three chromosomes: 14 for heavy chain, 22 for _, and two for l. The antibody genes are arranged in exons, each coding for a single immunoglobulin domain, variable and constant region, with further exons for the hinge and transmembrane regions. The constant region exons are grouped according to antibody class. There are additional exons for the leader sequences which take the newly synthesized antibody chains across the membrane (and is then cleaved off). The transmembrane exon is only expressed on antibody that is attached to cell surfaces.

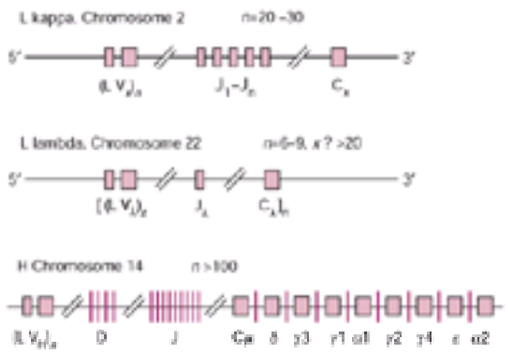

Fig. 5 Arrangement of the immunoglobulin genes. These are arranged on three chromosomes, chromosome 2 for *L*_, 22 for *L* l, and 14 for the heavy chain. The basic arrangement of the human genes are shown. *L*_ is arranged with a large number (*n*) of *V* genes, each preceded for a gene for the leader sequence *L*. Downstream are five *J* genes and further on the one *C* gene. *L*l is arranged in an unknown number (*n*) of sets, with a group of X (?*x* > 20) *L*–*V* pairs followed by one *J* and *C*. There may be six to nine such sets. *H* is arranged with a large number (>100) of *L*–*V* pairs followed by a set of *D* minigenes and a set of *J* genes. Downstream are the constant-region *C* genes, each split into separate exons for each domain, and an optional membrane segment, which are placed in the order shown.

The variable (V) genes are on the 5' side of the constant genes. (i.e. upstream, as DNA is read from 5' to 3'). There are many of these and it appears that any variable gene can combine with a constant gene on the same region of the chromosome. The antibody genes are therefore unusual and complex in their arrangement. Besides the multitude of variable genes there are also D (diversity) (for heavy chains but not light chains) and J (joining) genes which are short coding sequences (exons) found between V and C. The arrangement of the genes in B lymphocytes and plasma cells is different from that in all other cells. Early in B-cell development, a single heavy chain variable gene joins to a D and J gene to make a VDJ rearrangement. Similarly, in one light chain gene region a VJ rearrangement is made. Coupling of VDJ or VJ to the C gene occurs after transcription to nuclear RNA which is then processed to make messenger RNA, where, for both heavy and light chain transcipts, J is connected to C, and this is translated. The D segment, which is one to three amino acids long, is in the third hypervariable and the very many combinations of VD and J that are possible contribute substantially to antibody diversity. In addition to this, breaks in this region may occur in early B cells and be repaired in a random fashion by an enzyme called terminal deoxynuceotide transferase, which again adds further diversity.

Further variation in both heavy and light chain variable regions occurs by somatic mutation as B lymphocytes proliferate in large numbers in the germinal centres of lymphoid organs. This involves selection by antigen and a mechanism that locates the mutations in particular sites in the antigen binding site. As the concentration of antigen declines, only high avidity B cell clones will be stimulated to divide and this exerts selective pressure on mutations in the hypervariable regions. It is now known that developing B lymphocytes divide rapidly in the germinal centres and that a large proportion die by apoptosis (programmed cell death), a process not unlike that occurring in the thymus as T cells develop.

The process of gene rearrangement described above occurs on only one of the chromosomes encoding the heavy chain, and a similar process occurs on one of the four chromosomes encoding a light chain (allelic exclusion). A B cell is therefore committed to making one antibody, one heavy chain VDJC and one VJC by the time it is immunocompetent. It is not clear why this only happens on one chromosome for each chain. It is possible that it is such a complex process that the chances of a single cell making two rearrangements successfully is extremely remote. Alternatively, there may be some suppressive signal generated once a successful VDJ arrangement has occurred. This property provides a useful way of determining whether proliferating or infiltrating lymphocytes (for example in a tissue section) are polyclonal or monoclonal. The latter will all express either a _ or l chain; the former will include both light-chain types in roughly equal numbers.

The heavy chain VDJ sequence is attached to the C genes in an orderly progression during B cell development. The gene order is Cμ, C†, Cg3, Ca1, Cg 1, Cg 2, Cg 4, Ce, and Ca 2. All the B cells appear to go through a Cμ, stage, most go through C† and Cg, but may end up at one of the g subtypes, or a or e, probably jumping segments in the process. This progression involves deletion of DNA coding the no-longer-used C genes as the B cell develops. It is therefore a one-way process. Switching of B cells from production of IgM to IgG is a striking and normal feature of a simple antibody response. The process is, in part, regulated by external factors, particularly the cytokines IL4 and IL5. IL4 drive switching to IgE and IL4-secreting T cells (Th2 cells) probably play a key role in the development of allergic responses. It is also of practical importance that fetal and cord B lymphocytes do not switch from IgM to IgG production.

The T-cell receptor

Although the T-cell receptor was elusive for many years it is now well understood and some crystal structures have been determined. Early studies with monoclonal antibodies revealed a two-chain glycoprotein of 85 000 molecular weight, reducing to two components of 40 000 and 45 000 daltons. When its DNA was cloned it was found to rearrange in ways very similar to that described above for B-cell immunoglobulin receptors. Unexpectedly, two families of T cell receptors were found, the a–b receptor present of all conventional T cells and the g–† receptor present on a subset of T cells whose function is still less clear.

Both chains of ab T-cell receptors have been sequenced from multiple T-cell clones of known antigen specificity. Both chains are similar to immunoglobulin light chains with two external domains, each of about 100 amino acids. The N terminal domains of both a and b chains are variable. They include V and J segments, with a D segment for the b chain. Several families of Vb, Jb, Va, and Ja have been identified. The genes for these rearrange as T cells develop, whilst the cells are in the thymus. There are two constant region classes for the b chain, Cb1 and Cb2, which are very similar in structure. The number of possible T-cell receptors has been estimated to be a staggering 10^{14}, generated by the multiple combinations of the gene segments on the two chains. A substantial part of this diversity comes from the activity of the terminal deoxynucleotide transferase creating much variability at the V–(D)–J junction.

The crystal structure of the T-cell receptor shows that the three hypervariable regions (complementarity determining regions, CDR) form the MHC–peptide antigen binding site. This structure places the CDR-3 region, made up by the most hypervariable part of the T-cell receptor where the V–(D)–J join occurs, in a critical site to interact with the peptide part of the complex. The rules for this engagement are now being worked out. The T-cell receptor lies diagonally across the peptide binding groove of a class I MHC molecule and more at a right angle to the peptide in a class II MHC molecule. These orientations put the most variable parts of the receptor over the peptide and more conserved, but still variable, parts of the receptor over the MHC molecule.

Why do the T cells not simply use the antibodies as their receptor? The advantage of using MHC molecules as the antigen presenter is that it means that T cells do not see native antigen and thus cannot be inactivated by, for instance, free virus. In addition, by responding only to cell surface antigens the response can be controlled by signals derived from the presenting cell. Thus T cells that react with self-antigens in the thymus may be eliminated while those that react with foreign antigens in the periphery are stimulated.

The function of the T cells that carry the gd T-cell receptor remains enigmatic. In mice they are abundant in the intestinal mucosa. In humans they have been found at sites of chronic inflammation, particularly mycobacterial infections. Some bind to non-classical MHC molecules including CD1 and MIC-1, but why is less clear. Some are probably important in antibacterial immunity. One particular subset of T cells that carry some natural killer cell markers express a particular receptor and interact with CD1d; these T cells are thought to play a role in regulating Th1 and Th2 responses (see below).

Lymphocytes

Lymphocytes mediate immune reactions and can be divided into two groups: B and T cells.

B lymphocytes

B lymphocytes are the precursors of antibody-secreting plasma cells. They express immunoglobulin on their surface, which acts as antigen receptor. The B cell expresses only one pair of immunoglobulin VH and VL gene products and thus antigen receptors of only one molecular type and sequence. The progeny, or clone, of this cell retains the same commitment and the antibody secreted uses the same variable genes. Thus, antigen on first immune challenge selects B cells that already express appropriate receptors. These divide and some mature to antibody-producing cells while others develop into memory cells. The latter greatly exceed the original population in number and as they can in turn be activated by antigen to generate antibody-producing cells this provides a basis for the memory phenomenon.

Immature B cells express IgM or IgM plus IgD antibody as their receptor. As the B cells differentiate they switch their heavy chain VDJ gene product to associate with a g, a, or e chain and thus switch secretion from IgM to IgA or IgE. Memory cells have g, a, or e receptors and secrete this class without an intermediate phase.

A number of B-cell differentiation antigens have been found using murine monoclonal antibodies raised against human lymphocytes. The best characterized of them are shown in Table 4. As has been the case with the T cell antigens described below, it is gradually emerging that these have important functions. Thus, antibodies to CD20 trigger B cells to divide, CD22 is involved in B-cell signalling, CD19 is part of the B cell antibody receptor complex, CD21 is the complement C3b receptor and is also the receptor for Epstein–Barr virus (EBV) which readily transforms B lymphocytes *in vitro* and probably *in vivo*. Another molecule that is crucial in B cell signalling is CD40; its ligand, CD40L, is expressed on T lymphocytes. In the hyper-IgM syndrome, where there is a failure of B cells to switch from IgM to IgG production, DC4L is mutated, implying a role for CD40 in immunoglobulin isotype switching. CD40 is also expressed on dendritic cells and is important for receiving signals from CD40-ligand bearing T-helper cells to make them efficient at stimulating primary CD8+ T-cell responses by releasing the Th1-inducing cytokine IL-12.

Activation of B lymphocytes requires antigen and a signal from helper T (Th) lymphocytes which are themselves responding to the same antigen. The T cells, of the Th2 subset (see below) release cytokines including IL-4 and IL-5 as well as initiating signals through CD40; these activate B lymphocytes in the presence of antigen to divide and differentiate. Note that two kinds of signal are needed, an antigen specific trigger through the antigen receptor and a second type mediated in a non-antigen specific fashion. Inappropriate signalling, such as an antigen signal in the absence of the second type, can lead to inactivation of the B cells.

Besides secreting antibody, B cells have a role in antigen presentation that is increasingly recognized as important. B cells can bind foreign antigen directly through their immunoglobulin receptor or as an immune complex through the Fc or complement (immune complexes bind complement—see Chapter 5.4) receptors. Such antigen can be internalized and digested in endosomes to generate peptides that bind to class II MHC. In this way, small amounts of circulating antibody can enhance primary T-cell responses by facilitating processing of antigen and, in the early immune response, can act as a positive feedback.

B cells have a complex life-cycle that includes a selection process in the germinal centres of lymphoid organs. Here, many cells die by apoptosis and there is selection by antigen in an environment that favours somatic mutation. The ontogeny of B lymphocytes is of some clinical relevance because various leukaemia and lymphomas express surface antigens characteristic of B cells at various stages of development. Figure 6 gives the scheme for B-cell differentiation indicating the corresponding leukaemias and lymphomas. Studies on the immunoglobulin light chains expressed indicate that these malignancies are monoclonal diseases.

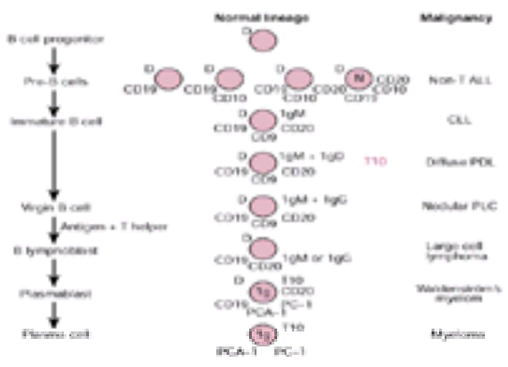

Fig. 6 Expression of B-lymphocyte differentiation antigens on the surface of differentiating B cells. The circles represent developing B cells. The CD antigens are listed in Table 4. In addition, D is HLA class II, μ is an intracellular immunoglobulin μ-chain. IgM, IgD, and IgG are surface immunoglobulins. Plasmablasts and plasma cells are shown with intracellular immunoglobulin (Ig). On the right are the B-cell malignancies that are thought to arise at the levels shown. They express the surface antigens indicated for their level. The B cell lymphomas and their abbreviations are described in Chapter 22.4.3.

Burkitt's lymphoma is a B-cell malignancy which is caused by Epstein–Barr virus (EBV). In addition to the presence of EBV DNA and protein in the malignant cells, there are some chromosomal rearrangements. The c-*myc* oncogene on chromosome 8 is translocated to chromosome 2, 14, or 22. There, it comes into close proximity with one of the three sets of immunoglobulin genes. It is likely that the tissue-specific enhancer that is present between the J and C exons activates the oncogene.

T lymphocytes

T lymphocytes require the thymus for their development and show a set of characteristic surface glycoproteins and their own form of receptor, as described above. They can be divided into types: cytotoxic T lymphocytes (CTL) that carry the CD8 marker and helper T cells that carry CD4. The latter can be divided according to the cytokines they release on antigen contact into Th1 (IL-2, IL-15, and interferon-g) and Th2 (IL-4, IL-5, IL-10, and IL-13). There are also intermediate T cells, Th0 cells, and some claim for more immunosuppressive T cells in the gut which secrete transforming growth factor (TGF-b). T cells may also be divided into those in an inactive state (virgin T cells) and preactive state (memory T cells) by the CD45 isotype on the cell surface. The short version of the molecule CD45RO marks cells that have recently seen antigen; however, fully differentiated CD8+ T cells can revert to CD45RA expression. Another marker that characterizes a subset of memory T cells is the chemokine receptor CCR7. This is involved in the trafficking of T cells to lymph nodes and is present on the surface of long-term memory T cells but is lost as these are stimulated to become effector T cells, functional outside the lymphoid organs.

T-cell differentiation antigens

Hybridoma-generated monoclonal antibodies have been used to explore the surface of T lymphocytes and these have revealed a series of molecules that play essential, accessory roles in antigen recognition by T cells (Table 5). Monoclonal antibodies to these structures are now often used to define T-cell subpopulations or leukaemias and are being used therapeutically. Their expression on T cells as they develop is described in Fig. 7.

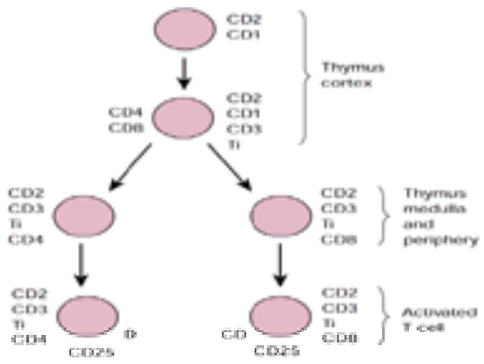

Fig. 7 Expression of T-lymphocyte differentiation antigens on differentiating T cells. The arrows indicate the probable pathway of T-cell development in the thymus and periphery. The CD antigens are described in Table 4 and in the text. Ti is the T-cell receptor. D is HLA class II. Negative selection (deletion of self-reactive T cells) occurs at the double positive (CD4+CD8+) stage and is followed by positive selection as either CD4 or CD8 is selected.

The T-cell receptor and CD3

The CD3 antigen is present on T-cell surfaces in association with the T-cell receptor. Antibodies to one will precipitate the other from detergent-solubilized cell membranes. The antigen is made up of three chains, which are closely associated with the T-cell receptor and form a complex on the inside of the T cell. When T-cell

receptor binds antigen the receptors cluster on the cell surface in lipid rafts, bringing the coreceptors CD4 or CD8 (see below) with them. This activates a cascade of kinases within the cell, involving kinases such as Lck and Zap 70. Through a complex series of interactions between enzymes and a parallel flux of calcium into the cytosol, several genes are activated including IL-2 and interferon-g. This causes a cytokine response within 6 h in memory T cells. Virgin T cells undergoing a primary T-cell response require additional activation through the cell surface glycoprotein CD28. This is activated by binding to its ligand CD80, which is expressed on dendritic cells, specialist antigen presenters. CD28 also sets up a kinase cascade important for T-cell activation and survival.

CD4

Recognition of processed antigen on antigen-presenting cells by helper T cells involves a specific binding interaction between CD4 and HLA class II molecules. Helper T cells carry the CD4 antigen. CD4 is also expressed at low levels on macrophages and some dendritic cells. The surface glycoprotein gp120 of human immunodeficiency virus (HIV) binds to CD4, giving the virus specificity for the cells that carry CD4, helper T cells, dendritic cells, and macrophages. The virus needs a second receptor to enter cells, either the chemokine receptor CCR5 or the chemokine receptor CXCR4. After high affinity binding to CD4, the gp120 undergoes a conformational change to expose the chemokine receptor binding site and then virus entry occurs. The cells targeted, with devastating consequences, by HIV are defined by the presence of these receptors, particularly CD4 which is restricted in its distribution to vital cells of the immune response.

CD8

This antigen is a glycoprotein of molecular weight 33 000. It is heavily glycosylated on a long stalk that holds an immunoglobulin domain away from the T-cell membrane. CD8 binds to the a-3 domain of HLA class I molecules and through its cytoplasmic domain recruits the tyrosine kinase Lck. It is the counterpart of CD4. In T-cell development in the thymus, immature T cells express both CD4 and CD8; if their receptor interacts strongly with self HLA antigens those T cells are deleted, if they react less strongly, but receive enhancing signals though CD4 or CD8, they are positively selected. In this process they lose expression of the inappropriate accessory molecule, CD4 or CD8, and become either CD4+ or CD8+ mature T cells.

Other important molecules on the surface of T lymphocytes are the chemokine receptors CCR5, CXCR4, and CCR7 which help to determine the trafficking of the T cells. The adhesion molecule LFA-1 (CD11a–CD18) is a two chain glycoprotein important for T and B cells to stick to endothelium and target cells. CD2 and CD44 have similar functions for different ligands. CD45 is an abundant glycoprotein that has several different isoforms and is expressed differently on lymphocytes at particular stages of differentiation; its cytoplasmic tail is a phosphatase that appears to control the state of activation of kinases such as Lck involved in T-cell activation.

T-cell subpopulations

Functionally, T lymphocytes can be divided into two major subtypes helper T cells (Th) which carry the CD4 glycoprotein and recognize antigens presented by class II HLA, and CTL which carry CD8 and respond to peptides presented by HLA class I. Th cells are divisible into two further major subtypes with different functions, Th1 and Th2. They differ in the cytokines they release on antigen activation. Th1 T cells release interferon-g, IL-2, and IL-15. These have direct and indirect antiviral effects and some inflammatory activity. They are potent in successful T-cell responses to infectious agents including mycobacteria, most viruses, and a range of parasitic infections. In contrast, Th2 cells release IL-4, IL-10- and IL-13. IL-10 has some immunosuppressive properties, IL-4 and IL-13 favour B-cell responses, facilitating the IgG immunnoglobulin switching. The extreme of a Th2 response is the IL-4 mediated switching to IgE production, resulting in allergic reactions. Th2 cells also produce IL-5 which stimulates eosinophils. Th2 immune responses are in combating parasitic worm infestations.

The balance between Th1 and Th2 is of considerable pathological interest. The two extreme forms of leprosy exemplify this beautifully. In tuberculoid leprosy, there is a Th1 response, with a strong inflammatory response of the delayed hypersensitivity type. In lepromatous leprosy, there is a Th2 response, the infection is poorly controlled with abundant organisms despite a good antibody response. Similar polarization can be found in tuberculosis (minimal disease (Th1) versus miliary (Th2)) and leishmania infection (minimal disease where there is a Th1 response and kala azar with a Th2 response). Whilst Th2 responses appear bad in these contexts, there are other situations where Th1 responses can be harmful, particularly in autoimmune diseases such as juvenile onset diabetes and rheumatoid arthritis. The polarization is not always complete and a variety of intermediate CD4+ T cells (Th0) have been described. Also, a Th3 type associated with the intestinal tract has been claimed, predominantly secreting TGF-b and protecting against immunization against food proteins.

For many years there were claims that there were a distinct population of specialist suppresser T cells that carried the CD8 marker. These do not exist as such but there are clearly suppressive phenomena mediated by subsets of CD4 or CD8 T cells through their cytokines or cytolytic effects. The term 'suppressor T cell' should be rested for the time being.

Cytotoxic T lymphocytes (CTL)

The effector T cells that mediate cellular immune responses are the CTL and Th1 cells that secreted cytokines such as interferon-g and tumour necrosis factor (TNF). Cytotoxic T cells recognize antigen plus class I HLA antigen on presenting cells. The antigen is peptide-derived, from intracellular proteins such as virus proteins. In fact, any cellular protein can enter the pathway that puts peptides into HLA molecules. The proteins are degraded by intracellular proteases, most important of which is the proteasome complex. The short peptides generated are transported by the transporter associated with antigen processing (TAP) into the endoplasmic reticulum where most are further degraded, but some that have the right sequence characteristics to bind to HLA class I molecules, and are then presented on the cell surface. CTL monitor the surface of cells for abnormalities and react when their receptors bind. In the case of an acute virus infection, the magnitude of the CTL response is extraordinary. Before infection, T cells with specificity for any single epitope (antigenic region) on a virus are extremely rare, less than 1 in a million lymphocytes. In several acute virus infections they have now been recorded at between 2 and 20 per cent of all lymphocytes, a massive expansion. These T cells are functional, able to kill virus-infected cells, and release cytokines such as interferon-g and TNF-a. The massive expansion is controlled by programmed cell death of the T cells and, as the virus antigen is removed, the number of antigen-specific T cells declines rapidly, leaving a memory T-cell response for rapid re-expansion should the virus attempt to reinfect.

There is considerable evidence that these T cells are crucial in clearing acute virus infections (such as measles, influenza) and controlling persistent infections, such as Epstein–Barr virus (EBV) or cytomegalovirus (CMV). Where CTL responses are impaired (e.g. by deliberate immunosuppression for transplantation) these viruses may escape control and cause disease, for example EBV-associated lymphomas. The CTL response is normally accompanied by a Th1 response that is important for initiating the CTL response and maintaining its functional activity and memory.

The CTL and Th1 responses, together often known as the 'cellular immune response' is central to the control of infections with intracellular pathogens that include not only viruses but some bacteria (e.g. *Listeria*, *Mycobacteria*) and protozoal infections (e.g. malaria liver stage). For some microbes, cytotoxicity is the important mediator of protection, for others interferon-g release or chemokine release is crucial. The CTL response could well be important in the control of some cancers. For those caused by viruses (EBV lymphomas, HBV-associated liver cancers, HPV-associated cervical cancer) this is already clear. There is increasing evidence that CTL responses are generated to some solid tumours, melanomas being the best example. For tumours and also many viruses, particularly HIV, CMV, EBV and HCV, selection of escape pathways arises. Many tumours and several viruses can decrease expression of HLA class I molecules on cell surfaces, making them invisible to CTL. Some viruses and many tumours mutate the epitopes recognized by CTL. Viruses and tumours can also increase expression of Fas-ligand on the cell surface, triggering apoptosis in the attacking T cells that express the ligand Fas. The presence of such mechanisms implies that the T cells are worth escaping from.

Graft rejection is a classical cellular immune response that has a long history in immunology. Both CTL and Th1 cells are involved and infiltrate the graft. HLA class II is essential to induce T helper cells and thus to activate T cells, the CTL recognize the foreign HLA class I molecules.

The role of thymus in T-cell function

T lymphocytes by definition are thymus derived. Thymectomized or congenitally athymic animals lack mature T cells and T-cell-mediated functions. Athymic (nude) mice and human infants with thymic hypoplasia (Di George syndrome) show no cellular immune responses and impaired antibody responses. B cells, however, can respond by secreting IgM to some polysaccharide antigens in the absence of T cells. Normal thymocytes carry the T-cell differentiation antigens described above and also CD1, which has some similarities to an HLA class I molecule but is not MHC encoded and has an unknown function in the thymus; the molecules are, however, also expressed on dendritic cells where they are involved in presentation of glycolipid antigens to T cells.

Immature lymphocytes enter the thymus from the bone marrow and first rearrange their T-cell receptor b genes, expressing first a primitive receptor associated with a

temporary second chain. Then the a chain genes rearrange and a functional receptor is expressed. At this point these thymocytes express both CD4 and CD8. Then two selection processes occur. The first is first deletion of any T cell that is self reactive (negative selection); dendritic cells enter the thymus from the periphery bringing in peptides from self tissues bound to self HLA class I and class II molecules. Self-reacting T cells that carry both CD4 and CD8 die by apoptosis. The second selection step is positive selection; a repertoire of T cells that express either CD4 or CD8 and have receptors that can recognize non-self peptide bound to self HLA molecules is generated. It is thought that the receptors on these T cells react weakly with the self peptides on the self HLA molecules, enough to trigger a survival signal rather than the stronger apoptotic signal of negative selection. During this process either CD4 or CD8 is lost. The T cells that then populate the periphery have an affinity for self that is too low to trigger a signal but can respond to foreign peptides bound to the self HLA molecules. In addition to the active selection processes in the thymus, many T cells whose receptors confer neither negative nor positive selection, also die. Thus more than 90 per cent of thymocytes die.

Some self-reactive T cells may escape to the periphery. Many of these are probably deleted but some may instead be put into an inactive (anergic) state. Under some circumstances these cells may be activated, for instance by a cross-reactive microbial peptide antigen, and autoimmunity can result.

A third way of preventing autoimmunity is for the self cell to be immunologically inaccessible. This may occur at certain sites such as the eye. It can also occur on cells that fail to express the costimulatory molecule B7 (CD80) which is necessary to initiate immune responses.

Cytokines

A number of cytokines have already been referred to, particularly in the context of understanding the function of Th cells and immunoglobulin class switching. They are small polypeptides, released by immunocytes and other cells, with normally short-range functions on target cells that carry the appropriate receptor. Specificity is therefore conferred by the nature and state of activation of the cell that makes the cytokine and by the cell that bears the receptor. The actual effects of cytokines tend to be pleiomorphic—activating cells, triggering general differentiation, and activation of specific genes (e.g. co-ordinated expression of HLA class I and II genes plus antigen processing genes, TAP and components of the proteasome by interferon-g). Of particular interest is the control of immunoglobulin isotype switching by cytokines (see Table 6) as well as by the CD40–CD40L interaction. Some cytokines, such as IL-12, seem more effective at orchestrating the others (the Th1 set for IL-12). Cytokine activity in terms of T-cell function can be described by the Th1–Th2 division. Also, Th3 cells that secrete the inhibitory cytokine TGF-b have been described. Although this is clear in mice it is more complex in human T cells, possibly because the original activation of the T cells *in vivo* is not under experimental control and also because the source of T cells, peripheral blood lymphocytes in man and spleen cells in mice, is different. In humans there are intermediate phenotypes known as Th0 cells. Nevertheless, it is a useful paradigm, implying that the response of Th cells may be set at the time of original activation and with implications for understanding disease processes. The cytokines relevant to the immune response are listed in Table 6. Some of these are being tried for therapy, for example interferon and IL-2. Also antibodies to some are useful, particularly anti-TNF for treatment of rheumatoid arthritis.

Accessory cells in the immune response

Antigen-presenting cells

Besides T and B lymphocytes, certain accessory cells play crucial roles in immune responses. Antigen-presenting cells are clearly important. B cells and CTL can react with antigen directly in native form or on altered cells, respectively, but both require signals from Th cells which are dependent on specialized presenting cells. The latter are either B lymphocytes or monocyte/ dendritic cells which can internalize antigen, degrade it, and display derived peptides in HLA class II molecules. They also carry specialized accessory molecules such as CD80 (also known as B7, not to be confused with HLA-B7) which binds to CD28 on T cells, delivering a cosignal. Dendritic cells are found not only in the T-cell areas of lymph nodes (see below) but also widely distributed in many organs. Here they may be important in activating local immune responses, for example to localized virus infection. Dendritic cells have to be activated to initiate CTL responses; this is achieved through the CD40 molecule recognized by CD40L on Th cells, although some viruses such as influenza may be able to activate directly. Dendritic cells seem to be able to internalize particulate antigens and can put these into the class I antigen presenting pathway. They are also able to take up apoptozing cells which may be an important pathway for initiating immune responses to viruses; macrophages are important for taking up necrotic cells and are not very effective in stimulating CTL responses. The requirements for initiating a primary CTL response are much more stringent than a secondary response, where the T cells may be able to respond and divide on contact with non-specialist cells that present antigen.

In the B-cell area of lymph nodes, the secondary follicles, there is a network of follicular dendritic cells. Unlike the dendritic cells referred to above, they are HLA class II negative but display receptors for C3b (of complement) and immunoglobulin Fc. They can therefore capture immune complexes which are particularly good at initiating primary immune responses. They are probably able to capture small amounts of antigen percolating through the sinuses of lymphoid organs. They can hold antigen at their surface for long periods, possibly months or years.

Adjuvants are chemicals given with antigen that are able to localize antigens at the site of injection, giving a local inflammatory response, activating macrophages, and antigen-processing cells and thus initiating immune responses more effectively. An example that was used clinically for several decades was potassium alum added to diphtheria and tetanus toxoids. By triggering a non-specific inflammatory reaction at the site of injection, an adjuvant may direct the type of immune response, Th1 or Th2, as well as enhancing the level. Recently, a hypothesis has been proposed by Matzinger, that non-specific danger signals are crucial in initiating immune responses to invading pathogens because self antigens do not provoke non-specific inflammatory responses and do not therefore provoke autoimmune reactions. Whilst the concept of the danger signal is probably correct, it is unlikely that the whole of self–non-self discrimination can be explained this way.

B cells are particularly important in priming Th cells. Because they carry antibody receptors, they can bind the foreign protein and internalize it for degradation. This can also be achieved if the antigen is bound to serum antibody; internalization occurs through the Fc receptor. In addition, complement receptors can facilitate presentation of antigen that is in immune complex form.

Natural killer cells

Interferon is a potent activator of natural killer cells. These are large, granular lymphocytes, neither classical T nor B cells, which lyse cultured tumour cells and virus-infected cells *in vitro* very efficiently. Their role *in vivo* is uncertain but they have been implicated in rejection of histocompatible bone marrow grafts, tumour immunity, antivirus immunity, and autoimmunity. They may thus form a general surveillance system, eliminating tumour cells and virus-infected cells as they arise. Although these effects are non-specific, antigen-specific T cells, by releasing g-interferon, could activate them to give a vital enhancement of T-cell killing.

Recently, much has been discovered about the specificity of their function. It is clear that they are particularly effective at recognizing and killing cells that lack expression of classical HLA class I molecules. Two series of receptors have been identified that specifically bind HLA class I molecules and deliver inhibitory signals to NK cells. One series KIR (killer inhibitory receptors) interacts with particular HLA-C or B molecules—there are at least three series of these receptors and different isoforms with differing specificities in each series. Expression of these receptors is complex and seems to vary on different NK cells within one individual. In addition, there are receptors in a related series that deliver stimulatory signals to NK cells. The second type of receptor is the CD94/NKG2 series which recognize a non-classical HLA molecule, HLA-E, which is expressed on a cell surface only if it has bound a peptide derived from the signal peptide of classical HLA class I molecules. This means that HLA-E can signal to NK cells that the classical class I molecules are present and thereby inhibit NK cell attack. Although one of the CD94/NKG2 family is activating, the predominant effect is inhibition of NK cell activity. Thus NK cells express a variety of receptors with differing specificities for HLA class I molecules, most with inhibitory function but some activators. It would appear that these cells are controlled by a set of modulating receptors that signal the type of cell that is in contact with the NK cell.

Mast cells

In addition to antigen-presenting cells, there are other accessory cells at the other end of the immune response which might be termed enhancers. The best characterized of these is the mast cell which has a receptor for the Fc portion of the e chain of IgE. When this antibody, bound by its Fc region to mast cells, binds antigen the cross linking of neighbouring Fc receptors triggers the cell to degranulate. This results in the release of histamine, kinins, and leucotrienes, which give the anaphylactic of type 1 allergic reaction. Mast cells also have receptors for some of the peptides released during complement activation.

Macrophages

Macrophages are derived from monocytes and are long-lived, potent phagocytic cells. Differentiation to macrophages and their activation is a response to local events such as contact with foreign material, lectins, and complement fragment C5a, but is also under the control of immune cells, with immune complexes and interferon-g, which is released by antigen-specific T cells, being potent activators. Macrophages are larger than monocytes and differ in their surface glycoproteins with less HLA class II antigen and increased amounts of receptors for immunoglobulin Fc fragments and complement. Within the cells, lysosomes are increased in number.

The main function of macrophages, and also granulocytes, is phagocytosis. This is greatly enhanced (several thousand-fold) if the foreign material is coated with antibody and/or the complement fragment C3b. Ingested particles are taken into the cell in a phagosome which fuses to a lysosome. Similar processes in granulocytes and macrophages are associated with a respiratory burst; there is a sudden uptake of oxygen and generation of hydrogen peroxide and hydroxyl radicals which are toxic to micro-organisms. Nitric oxide production through activation of NO synthase is also of considerable importance. Activated macrophages also produce a variety of enzymes that are important in inflammatory processes, including proteases, elastase, collagenase, plasminogen activator, and procoagulants. The last may account for the deposition of fibrin, which is responsible for the characteristic induration of delayed-type hypersensitivity reactions.

Macrophages also release monokines and synthesize complement components. The former include interleukin-1 and tumour necrosis factor (Table 6). The latter may help to amplify the local inflammatory responses. Bioactive lipids, including prostaglandins and leucotrienes, are also made.

Macrophages are thus highly active cells which are crucially important in converting immune responses into inflammatory reactions. They feature prominently in granuloma formation where antigen persists or forms immune complexes. Mycobacteria are thought to inhibit fusion between phagosomes and lysosomes and thus evade the toxic mechanisms. Under these circumstances, the macrophages form epithelioid and giant cells in forming the granuloma.

Like natural killer cells, macrophages may be dangerous to the body and a number of inhibitory receptors have been identified. These include the 'immunoglubulin-like transcripts', ILT-2 and ILT-4, which bind HLA class I molecules, both classical and non-classical. Again normal cells deliver a 'hands off' signal whereas HLA-negative cells would be targets. A similar situation applies to the ligand for CD46, the latter is expressed on red cells and protects them from destruction by splenic macrophages in the white pulp.

Complement

The complement pathway is described in Chapter 5.4. This is an important, intrinsic pathway of immunity, activated directly by micro-organisms or by immune complexes. Complement can activate cells such as monocytes, macrophages, and neutraphils, contributing to inflammatory responses.

Organization of the immune system

The immunocytes are divided into those which are circulating in blood and lymph, and those localized in lymphoid organs: the thymus, bone marrow, fetal liver, spleen, lymph nodes, and the gut-associated immune system of tonsils, Peyers patches, and intraepithelial lymphocytes.

The circulating lymphocytes follow precise routes. From the efferent lymph nodes they travel to the thoracic duct and thence to the venous blood. From the blood they return to lymph nodes or spleen through high-walled capillary venules. Gut-associated lymphocytes also circulate and tend to home back to gut lymphoid tissue, or bronchial or mammary tissue, thus distributing antigen-sensitive cells widely to all possible sites of entry of organisms.

The structure of a lymph node is shown in diagrammatic form in Fig. 8 (see also Chapter 23.1). The B cells are congregated in follicles. Primary follicles contain mostly early B cells with IgD on their surface. Activation of B cells by an immune response in a germinal centre results in division and accumulation of B cells to generate secondary follicles which contain mostly IgG-bearing B cells. Scattered through the follicle are helper T cells and follicular dendritic cells. From the follicles, stimulated B cells mature to plasma cells in cords in the medulla where they secrete immunoglobulin.

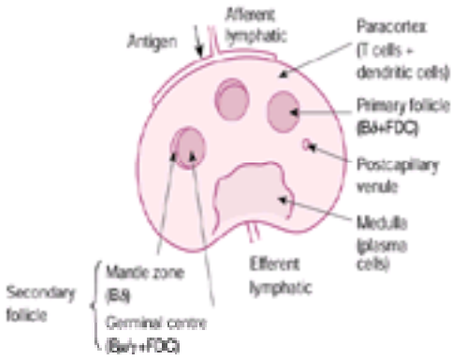

Fig. 8 Schematic representation of the structure of a lymph node. The drawing is not to scale but shows the main components. The cell types found in each region are indicated: B†, B cells expressing surface IgD; Bμ/g, B cells expressing surface IgM or surface IgG; FDC, follicular dendritic cells; T, T cells (a few T cells are also found in the germinal centres where B cells may present antigen).

The majority of T cells in the lymph nodes are in the paracortex surrounding the follicles. The T helper cells and Tc/Ts are found there. They are clearly associated with the HLA class II-positive dendritic cells. Antigen enters by the afferent lymphatics and percolates through the lymph nodes, activating immune cells via the antigen presenting cells.

The spleen is similarly organized in the white pulp, which surrounds the end arterioles, whence the cells and antigen enter. Circulating lymphocytes leave by venous sinuses (Fig. 9).

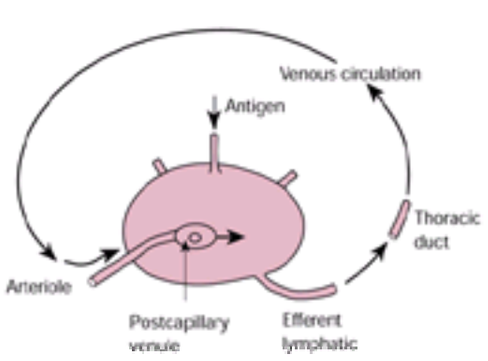

Fig. 9 Lymphocyte recirculation. Antigen enters the lymph node by the afferent lymphatics. Lymphocytes enter the lymph node through the walls of the postcapillary (high endothelial) venules. They leave by the efferent lymphatics, whence they go to the thoracic duct and the superior vena cava.

Conclusions

Immune responses are occurring in healthy individuals continuously. The immune system should be thought of as an essential part of the homeostatic mechanism, continually keeping out and destroying invaders and abnormal cells as they appear. At the same time it is regulating its own cells as they develop and circulate. Although the immune system has been described in terms of its individual components they do not react in isolation. Thus, an infecting virus, for instance, will evoke an antibody response, activate complement through the classical and alternative pathways, stimulate T-cell immunity involving both regulatory and effector T cells, stimulate lymphokine release, and activate natural killer cells. In a later chapter the consequences of abnormal immune responses are described, both where there is a deficiency in immune reactivity and where, for various reasons, the immune reaction itself is harmful.

Further reading

Aguado B *et al*. (1999). Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921–23.

Alberts B *et al*. (1994). *Molecular biology of the cell*, 3rd edn. Garland, New York.

Barclay AN *et al*. (1993). *The leucocyte antigen facts book*. Academic Press, London.

Bjorkman PJ *et al*. (1987). Structure of the human major histocompatibility antigen, HLA-A2. *Nature* **329**, 506–12.

Burke F *et al*. (1993). The cytokine wall chart. *Immunology Today* **14**, 147.

Davis MM, Bjorkman PJ (1988). The T-cell receptor genes and T cell recognition. *Nature* **344**, 395–402.

Elliott T, Smith M, Driscoll P, McMichael AJ (1993). Peptide selection by class I molecules of the major histocompatibility complexs. *Current Biology* **3**, 854–66.

5.2

Allergy

L. M. Lichtenstein

Introduction

Approximately 20 per cent of the population suffers from allergy. In allergy, exposure to common environmental substances induces the production of specific antibodies of the IgE class that arm mast cells and basophils to initiate a complex response which leads to the tissue inflammation. It is believed that the IgE response originally developed to combat parasitic diseases but this has never been proved.

Allergens are antigens which induce IgE antibody responses. They may be large molecules, usually proteins, or they may be small molecules, 'haptens', such as penicillin, which link to protein molecules to induce the immune response. Not all individuals who develop an IgE antibody response develop allergic symptoms, but after exposure to the allergen most will suffer rhinitis, asthma, or anaphylaxis. The synthesis of allergen-specific IgE by B lymphocytes is thought to be mediated by activated CD4+ T lymphocytes. The IgE occupies high-affinity receptors on mast cells and basophils and the response is triggered when these cells bind allergen. Murine CD4 T-cell populations have been defined as T_{H1} cells, which produce interleukin 2 (**IL-2**) and g-interferon, and T_{H2} cells, which produce IL-4 and IL-5. Both populations of cells secrete other cytokines. In humans, the distinctions between T_{H1} and T_{H2} cells are not so clear, but none the less help us to characterize allergic disease. Allergic individuals have much higher IgE levels than non-allergic individuals, which contribute to diagnosis. Allergen skin testing allows specific diagnosis.

Cytokines influence all phases of the allergic inflammatory response. IL-4 is required to initiate IgE synthesis, while ongoing synthesis is enhanced by IL-5. Other cytokines such as g-interferon downregulate IgE synthesis. Granulocyte–macrophage colony-stimulating factor (**GM-CSF**) and IL-3, -4, and -9 promote the differentiation and expansion of mast cells, whereas g-interferon interferes with their growth. Another group of recently discovered cytokines are the histamine-releasing factors, first identified in monocytes. Mast cells and, particularly, basophils, also synthesize cytokines which include IL-4 and IL-13. The stages of an allergic reaction are seen in Fig. 1, Fig. 2 and Fig. 3.

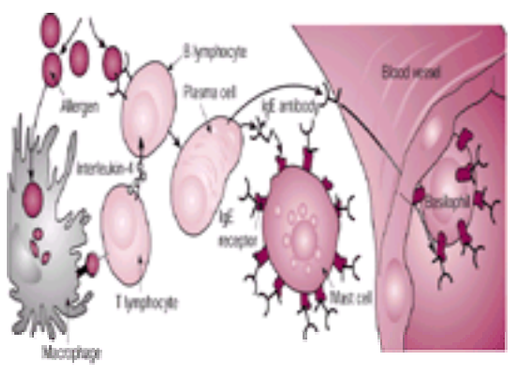

Fig. 1 Stages of an allergic reaction: sensitization. The initial meeting of an allergen and the immune system yields no symptoms; rather it may prepare the body to react promptly to future encounters with the substance. The sensitization process begins when macrophages degrade the allergen and display the resulting fragments to T lymphocytes (bottom left). The steps that follow are somewhat obscure, but in a process involving secretion of interleukin 4 by T cells, B lymphocytes mature into plasma cells able to secrete allergen-specific molecules known as immunoglobulin E (IgE) antibodies. These antibodies attach to receptors on mast cells in tissue and on basophils circulating in blood.

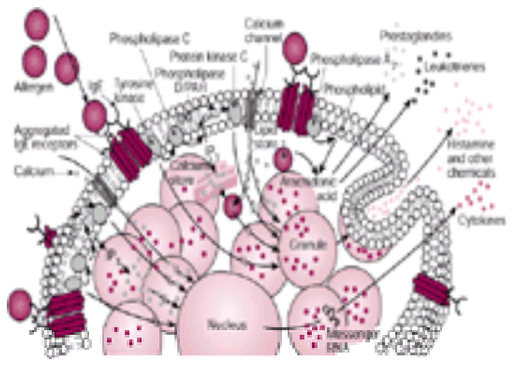

Fig. 2 Stages of an allergic reaction: activation of mast cells. In later encounters between the allergen and the body, allergen molecules promptly bind to IgE antibodies on mast cells (top left). When one such molecule connects with two IgE molecules on the cell surface, it draws together the attached IgE receptors, thereby directly or indirectly activating various enzymes (green spheres) in the cell membrane. Cascades involving tyrosine kinase enzymes, phospholipase C, protein kinase C, and an influx of calcium ions (black arrows) induce chemical-laden granules to release their contents. These cascades also appear to promote the synthesis and extrusion of chemicals known as cytokines (brown arrows). Other sequences of molecular interactions (green arrows) end in the secretion of such lipids as prostaglandins and leukotrienes. The various chemicals released by mast cells are responsible for many allergic symptoms. The reaction pathways shown are simplified and are only a few of those thought to occur; many are also only partly understood (broken arrows).

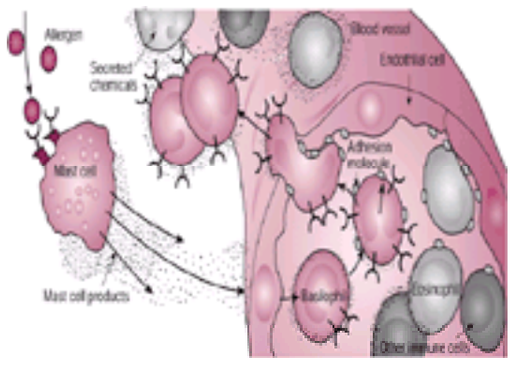

Fig. 3 Stages of an allergic reaction: prolonged immune activity. Chemicals emitted by activated mast cells (left) and their neighbours in tissue may induce basophils, eosinophils, and other cells flowing through blood vessels (right) to migrate into that tissue. The chemicals facilitate migration by promoting the expression and activity of adhesion molecules on the circulating cells and on vascular endothelial cells. The circulating cells then attach to the endothelial cells, roll along them, and eventually, cross between them into the surrounding matrix. These recruited cells secrete chemicals of their own (orange speckles), which can sustain immune activity and damage tissue.

Allergens

Modern techniques have facilitated characterization of important allergens, many of which have enzymatic activities that may favour penetration of mucosal surfaces. In the past, the allergen preparations used for diagnosis and therapy have not been standardized, resulting in unacceptable inconsistencies in diagnosis and treatment. The current trend is to insist upon standardization of such preparations, based on their major defined allergens. The mapping of T-cell epitopes has allowed development of peptide vaccines for immunotherapy; some of these have clinical efficacy and merit further therapeutic exploration.

Genetic basis of atopic disease

There is a genetic basis to the allergic diathesis. Allergic susceptibility is clearly polygenic. For example, certain ragweed allergens such as Amb a V and Amb a V1 are recognized in association with HLA DR2, the W2 and HLA DR11 haplotypes, respectively. Lineage studies of large families have suggested autosomal dominant or recessive patterns of inheritance. Rather than focusing on atopic individuals, early studies sought associations with one or more criterion: high serum IgE, positive skin tests, or clinical allergic disease. Linkage data suggested the presence of a gene for atopy on the long arm of chromosome 11. This finding implied that non-MHC linked genes may determine the overall predisposition to allergic disease, while MHC gene products control the specificity of the allergic response.

The cells involved in allergic disease

Mast cells and basophils

Tissue mast cells account for most acute allergic phenomena while the basophil, which infiltrates into the regions of an acute response, is responsible for chronic allergic manifestations. Both cell types originate in the bone marrow and have cytoplasmic granules. Tissue mast cells are found mostly in the lungs, skin, and gastrointestinal track, the location of most immediate hypersensitivity reactions. IL-3 and stem cell factor (**SCF**) are required for the growth, differentiation, and survival of basophils, while the added effects of SCF are critical for mast cell maturation and survival. The signal transduction events which are induced by cross-linking of cell-bound IgE by allergen have been studied with great interest, since interference with these signals would seem to be the best target for blocking the entire allergic response. Intracellular signals mediated by increased levels of intracellular calcium and activation of protein kinase C lead to phosphorylation of granule membrane proteins. These signals lead to the release of preformed mediators such as histamine and to the production of arachidonic acid metabolites including prostaglandin D and leukotriene C. This mode of activation is associated with cytokine generation.

Eosinophils

Circulating and tissue eosinophilia are cardinal manifestations of allergic diseases. Eosinophils are derived from the bone marrow aided principally by IL-5 but also by IL-3 and GM-CSF, and normally constitute less that 3 per cent of the circulating granulocytes. Tissue retention of eosinophils is determined by chemotactic factors which include chemokines (e.g. eotaxin), leukotrienes (e.g. LTC_4), and platelet-activating factor. Eosinophils have very few if any IgE receptors on their surface. Eosinophils contain highly cationic proteins including the so-called major basic protein. These are highly toxic products which have been implicated in the denudation of bronchial epithelium in asthma and in other types of cellular injury.

Monocytes, macrophages, and lymphocytes

Monocytes and macrophages may be activated by allergen binding to a low-affinity IgE receptor and contribute thereby to the non-specific inflammation of allergic diseases. Stimulated monocytes and macrophages produce many proinflammatory mediators and cytokines, including lysosomal enzymes, superoxide anions, lipid mediators, and diverse cytokines. Lymphocytes contribute critically to allergic disease.

The mediators of allergic disease

Histamine

This is the principal mediator of immediate hypersensitivity reactions; it is produced by basophils and mast cells and reacts with three specific histamine receptors. The first recognized, H_1, mediates hypersecretion of mucus, pruritus, contraction of non-vascular smooth muscle, and relaxation of vascular smooth muscle. There are now several non-sedating H_1 antihistamine drugs which are useful in controlling the manifestations of allergic disease. The H_2 receptor mediates enhanced gastric acid, mucus secretion, and bronchodilatation. There are also specific H_2 antihistamines. An H_3 receptor mediates the synthesis and release of histamine and neurotransmitters; anti-H_3 drugs are not available.

Eicosanoids

Eicosanoids are synthesized by mast cells, basophils, and eosinophils by the oxygenation of arachidonic acid metabolites via two pathways: the cyclo-oxygenase pathway forms prostaglandins and thromboxanes and the 5-lipoxygenase pathway generates leukotrienes. Prostaglandin D_2 and thomboxane A_2 are bronchoconstrictors. Leukotriene A is an unstable intermediate in the formation of leukotrienes B, C, D, and E. Leukotriene B is a chemotactic for leucocytes and leukotrienes C, D, and E increase systemic vascular permeability, smooth muscle contraction, and mucus secretions. The recent introduction of leukotriene antagonists has been a useful adjunct to the control of allergic diseases.

Platelet-activating factor

Platelet-activating factor is a proinflammatory molecule whose true function in allergic phenomena is not clearly understood. It is derived from the membrane phospholipids of mast cells and other cell types. Inhalational challenge with platelet-activating factor promotes bronchoconstriction and airways hyperreactivity in both asthmatic and non-asthmatic individuals and increases vascular permeability and mucus release. Platelet-activating factor is also a chemoattractant for eosinophils and neutrophils.

Kinins

Kinins are generated by kallikrein action on serum kininogens. Their biological activities are quite similar to histamine.

Allergic diseases

Anaphylaxis

Anaphylaxis results from the rapid degranulation of mast cells and basophils, usually, but not always, caused by the parental administration of drugs such as penicillin, or insect stings. Anaphylaxis is a medical emergency that can result in death either by cardiovascular collapse, laryngeal oedema, and/or bronchial smooth muscle constriction and anoxia. Anaphylaxis can also be induced by several food allergens such as cow's milk, shellfish, or peanuts. Anaphylactoid reactions, which do not act through IgE-related mechanisms, may follow ingestion of aspirin or other non-steroidal anti-inflammatory drugs or by the injection of radiocontrast media, metabisulphites, or opiates.

The clinical manifestations of anaphylaxis or anaphylactoid reactions are similar and may involve the skin (erythema, urticaria, angio-oedema), laryngeal oedema, bronchoconstriction, vascular shock, and gastrointestinal symptoms (abdominal pain, nausea, vomiting, or diarrhoea). These reactions cause tachycardia, arrhythmias, hypotension, and anoxia, leading to myocardial or cerebral damage. Treatment involves the removal of the inciting agent when possible with the immediate therapeutic use of adrenaline intravenously, if necessary. Antihistamines may be given but probably have little effect, and corticosteroids, while often recommended, are not useful in the acute phase. Attention to airway and cardiovascular support is crucial. After anaphylaxis, patients should be maintained under observation for a minimum of 24 h as the anaphylactic reaction may recur. Disseminated intravascular coagulation with widespread haemorrhage due to thrombocytopenia is a a severe late complication of anaphylaxis. Individuals allergic to insect venoms whose treatment is not complete, and patients who have food allergy should be advised to carry prepackaged adrenaline syringes.

Asthma (see also Section 17.4)

Extrinsic asthma is sometimes differentiated from intrinsic asthma. There is real doubt, however, as to whether intrinsic asthma exists or whether the causal allergen is simply not identified. Asthma affects 10 per cent of children and 3 to 5 per cent of adults and may range from a modest shortness of breath to a severe, life-threatening, obstructive ventilatory disease. By definition, asthma is characterized by widespread narrowing of the airways, which is reversible either spontaneously or as a result of treatment. However, chronic asthma may ultimately cause fixed airway obstruction. Asthma is usually associated with a personal family history of allergic disease and offending allergens may be known to precipitate attacks. However, a definitive clinical history and appropriate skin testing may be needed for diagnosis. Asthma may often be perennial, as a result of exposure to house dust mites or pet allergens.

Asthma is increasing all over the world, and this is especially true in inner-city African-American people. When wealth and environmental status are taken into account, the increase in African-American patients is still greater than that in Caucasian populations. The reasons for this are not known. Environmental pollution and cockroach allergy have been suggested to account for this, but evidence is lacking. The increase in asthma, for example, was shown to be greater in western Germany than in heavily polluted eastern Germany. The pathogenesis of asthma is exceedingly complex. When an atopic or asthmatic individual is challenged by bronochial allergens there is an early bronchocontraction, which is due to the release of mediators such as histamine and leukotrienes, followed by a recovery and a so-called late-phase response, which has been shown to be due to the infiltration of basophils, eosinophils, and other proinflammatory cells as well as the mediators of the acute phase. Asthmatic individuals have a heightened non-specific response to bronchoconstrictors such as histamine, methacholine, and especially, bradykinin.

The diagnosis and management of asthma are discussed in Section 17.4. However, a comment on a new therapy is appropriate here. A humanized anti-IgE antibody has been developed at Genentech and has been used in a variety of clinical trials. The use of this antibody in appropriate dosage intravenously decreases the serum IgE by more than 90 per cent. Moreover, as we predicted 20 years ago, based on the close correlation between serum IgE and mast cell and basophil IgE receptor number, it also decreases the IgE receptors on these cells. When this reduction is sufficient, there is a partial or complete inability to elicit basophil histamine release and the results of skin tests can be depressed or negative. The appropriate individuals for treatment with this regimen are just those who most need it: asthmatic patients with sensitivities to multiple allergens. In these individuals, the number of allergen-specific IgE antibodies can be reduced below the necessary level for initiating mediator release. In most individuals with asthma caused by a single or a few allergens, anti-IgE therapy will not be helpful. Clinical trials in a mixed population, including those who should and should not be receiving anti-IgE, showed a modest amelioration of asthma. It seems likely that if therapy were limited to appropriate individuals, the relief may be greater.

Allergic rhinitis

Allergic rhinitis affects about 10 per cent of the population and may be seasonal, in response to the pollens of weeds, trees, and grasses, or perennial, usually linked to house dust mite sensitivity or animal dander. The symptoms are sneezing, nasal congestion, rhinorrhoea, and often pharyngeal and conjunctival pruritus. Inspection of the nasal passages usually reveals a pale mucosa with swollen turbinates. The disease may be diagnosed by the timing of symptoms since grass, tree, or weed pollination occurs at predictable times, or symptoms may appear in association with pet exposure. In order to induce allergy, the particles of pollens or dust mite must be in the range of 10 to 100 μm in diameter. Nasal polyps may accompany the mucosal oedema, particularly in perennial rhinitis. Vasomotor or non-allergic rhinitis has many of the same symptoms but it is not due to IgE-mediated events. The diagnosis of allergic rhinitis is made by clinical history and evidence of specific IgE antibodies by skin testing or by the presence of serum specific IgE detected in the laboratory. The former diagnostic test is preferable.

Insect venom allergy (see also Chapter 8.2)

Insect venom hypersensitivity can be demonstrated in approximately 20 per cent of individuals in the United States, although only 3 to 5 per cent have a history of an anaphylactic reaction. While the number of deaths caused by this sensitivity is few (perhaps 50 per year in the United States), the social morbidity induced by insect sensitivity is very significant and this condition is one of the few where allergen immunotherapy is clearly indicated and effective. The history of the treatment of insect sensitivity shows how misleading uncontrolled studies can be. The disorder was formerly treated with 'extracts' derived by grinding-up the entire insect and injecting it. When controlled studies were initiated, it was found that treatment with placebo and whole-body extract were identical, while, as noted, venom immunotherapy was almost completely effective.

Insect reactions are characteristic of the order Hymenoptera including honey bees and vespids, such as wasps, white and yellow hornets, and yellow jackets, as well as polistes and fire ants. Another form of response to an insect sting is a large local reaction which occurs immediately contiguous to the sting and does not involve any life-threatening response of blood vessels or airways. These individuals rarely develop systemic reactions and usually do not require specific treatment, but in some these reactions are very large, crossing a joint space and causing marked discomfort. Immunotherapy should be considered in these individuals because it is effective. Diagnosis, as with other allergies, can be established by history, skin testing, or by the measurement of specific IgE antibodies. Rarely, an individual will have an anaphylactic episode despite being unaware of the sting, so this possibility must be considered in all cases of idiopathic anaphylaxis. While patients are unprotected, as in early stages of immunotherapy, they should carry emergency kits containing adrenaline.

Since immunotherapy is extremely effective, it should be used in all adult patients. Immunotherapy is usually continued for 4 or 5 years and may then be stopped, although there is some risk even after the several years of therapy. There is a difference between adults and children with respect to the manifestations of insect hypersensitivity. In adults, most reactions involve the airways or vascular system and are life-threatening. This occurs only rarely in children, with most manifestations being cutaneous. Children who have had such a cutaneous reaction usually do not have a further reaction on re-sting and need not be treated.

Urticaria (see also Section 23)

Urticaria is characterized by well-defined areas of transient pruritic dermal oedema, demarcated by a red border, which usually resolve spontaneously within a few hours, although episodes may continue for days. If the oedema spreads through the underlying epidermis, then it is called angio-oedema. The latter occurs mostly in the periorbital regions, the lips, the tongue, and the oropharynx and does not itch. In these instances, the possibility of pharyngeal obstruction may develop rapidly. Urticaria is probably due to mast cell degranulation, whether this is immunologically or non-immunologically mediated. Basophils may play a role in the longer episodes. Biopsies of acute urticarial lesions simply show oedema, but infiltration with neutrophils and eosinophils with perivascular monocytes may occur. Most

commonly, urticaria is 'idiopathic' as no offending allergen can be identified. However, certain foods such as eggs, shellfish, and peanuts, or drugs are implicated occasionally. There is also a syndrome caused by aspirin and other non-steroid anti-inflammatory agents. Underlying atopy is not usually associated with urticaria.

Acute urticaria usually resolves within hours or days. In urticaria which is protracted or recurrent, the causal agents are usually undiscovered. However, increasingly, antihypertensive drugs, antirheumatoid arthritis drugs, and hormone-based medications such oral contraceptives are identified as triggering agents. There are forms of urticaria that have physical triggers: dermatographism is the condition where individuals suffer from urticaria on scratching or pressure. Cholinergic urticaria is associated with a tendency to show symptoms due to perspiration. Urticaria may also be evoked by cold, heat, pressure, or sunlight. There is a rare condition called hereditary angio-oedema, which is an autosomal dominant disorder characterized by severe episodic attacks of intractable bowel, laryngeal, and cutaneous angio-oedema. A quantitative or qualitative defect of the C1 esterase inhibitor is associated with an uncontrolled activation of complement components, C4 and C2.

If attacks of urticaria are recurrent and without obvious precipitating factors, patients should keep a diary to document food, beverage, and drug intake. In this way, triggers may be identified. It should also be remembered that urticaria sometimes is associated with systemic diseases such as lymphoma or systemic lupus erythematosis.

Treatment at first is with antihistamines, and since they may have to be used quite vigorously, the non-sedating antihistamines are probably preferable. These may be ineffective and, if so, some of the other sedating H_1-antihistamines such as cyproheptadine or cetirizine should be used. Systemic corticosteroids may be required in very severe cases that do not respond to other treatments.

Food allergy

True IgE-mediated food allergy is far more rare in adults than is generally believed. As noted, this usually occurs with the ingestion of specific allergens such as eggs, nuts, or shellfish. Food allergy is far more common in children. Skin tests are rather ineffective and double-blind, controlled, oral challenge is usually necessary for diagnosis.

The diagnosis of allergic disease

Skin testing

Skin testing was first described in the 1860s by Charles Blackley. It may be performed at any site, but is usually on the lower aspect of the forearm. Allergen testing must be accompanied by saline as a negative control and histamine as a positive control. Methods differ with intradermal skin testing being most common in the United States while prick testing, a lancet through a drop of allergen extract, is in frequent use in the United Kingdom. Results are read at 15 to 20 min and are interpreted by several techniques which measure the magnitude of the weal and flare response. Large numbers of skin tests are not usually indicated; skin testing with 5 to 10 different allergens, based on history, is the proper clinical procedure. The identification of food allergies is less useful, since there are many false positives requiring appropriate double-blind ingestion of suspected foods.

In vitro tests for specific IgE antibodies

Such testing is necessary in certain conditions such as in dermatographism and in young children. Allergen extracts are immobilized on an insoluble particle and then are allowed to react with a patient's serum. After appropriate washing, antihuman IgE is added, either radiolabelled or conjugated to an enzyme.

The treatment of allergic disease

The most effective measure is allergen avoidance, and this should be rigorously attempted by scrupulous house-cleaning, impenetrable mattress and pillow covers, and the use of air conditioning and filters. The first line of medical treatment is with long-acting antihistamines. The standard antihistamines should be tried first as they are much cheaper than the non-sedating antihistamines and only a small percentage of the population are sedated by them. The next line in therapy is a topical corticosteroid spray. Used properly, these have no systemic side-effects and are very effective. Topical vasoconstrictors are not recommended as they cause a rebound chemical rhinitis.

For rhinitis unresponsive to these measures, immunotherapy involving weekly injections of gradually increasing doses of the allergens to which the patients are sensitive is effective, as shown by double-blind trials in the United Kingdom and the United States. The new standardized allergen extracts should be administered by an experienced physician. Immunotherapy is not effective for asthma in children and has marginal utility in adults.

The underlying immunological mechanisms of successful immunotherapy have not been fully determined. There is an increase in specific IgGantibody to the allergen that is associated with successful immunotherapy, but this association is not felt to be causal. Other types of immunotherapy are being evaluated: in one, the relevant allergen peptides are determined by reaction with T lymphocytes. This has been shown to lead to significant clinical improvement unassociated with any changes in allergen-specific IgG or IgE. The method has been largely abandoned because it was shown to be less clinically effective than standard immunotherapy; however, only a single regimen was examined. The therapy has considerable potential as the peptides are completely unreactive with mast cells or basophils and is thus extremely safe. Another new therapy uses allergens linked to bacterial DNA, which leads to a complete switch of the established T_{H2} response to a T_{H1} response in mice. Recent experiments in primates shows a similar type of immune response to these materials and the results of clinical trials are eagerly awaited.

Further reading

Lichtenstein LM, Fauci AS, eds (1996). *Current therapy in allergy, immunology, and rheumatology*, 5th edn. Mosby-Year Book, Inc., St. Louis.

Marone G *et al.*, eds (1998). *Asthma and allergic diseases.* Academic Press, London.

Middleton E Jr *et al.*, eds (1998). *Allergy principles and practice*, 5th edn. Mosby-Year Book, Inc., St. Louis.

Naclerio RM, Durham SR, Mygind N, eds (1999). *Rhinitis mechanisms and management*, Marcel Dekker, Inc., New York. Vol 123 in the series, *Lung biology in health and disease*, Claude Lenfant, ed.

5.3 Autoimmunity

Antony Rosen

Introduction

The effector mechanisms that the immune system utilizes to destroy extracellular pathogens, or host cells that harbour intracellular foreign bodies (such as mycobacteria or viruses) must be appropriately targeted if indiscriminate damage to normal host tissue is to be avoided. Under most inflammatory circumstances, some bystander tissue damage is unavoidable. In most circumstances, this damage is self-limited, due to efficient clearance of the exogenous antigen source and appropriate down-modulation of the immune response. Tissue damage in autoimmune diseases differs fundamentally from bystander damage in that the host immune system is specifically activated and driven by self-components, focusing damaging immune effector pathways on host tissues expressing those components, in an autoamplifying and self-sustaining way. The danger inherent in initiating a self-sustaining, specific immune response directed against components of self-tissues is intuitively apparent, since antigen clearance under these circumstances is necessarily associated with complete tissue destruction.

It is now clear that an autoimmune component is a feature of many human diseases. Indeed, there are some estimates that autoimmune diseases afflict more than 3 per cent of inhabitants of Western countries, and impose a significant personal and economic burden on individuals and nations. This chapter will illustrate many of the principles unifying various autoimmune states, and will present a conceptual framework within which to understand their aetiology, pathogenesis, and pathology. The rapid advances in knowledge being made in this group of disorders predict that disease mechanisms will soon be more clearly understood, and will greatly impact therapeutics.

Definitions

Autoimmune disease occurs when a sustained, specific, adaptive immune response is generated against self-components, and results in tissue damage or dysfunction. Although a single immune effector pathway may predominate in generating tissue dysfunction and damage in some autoimmune diseases, it is much more frequent for multiple effector pathways to participate in generating the final phenotype. Those pathways which generate tissue damage or dysfunction include autoantibody binding to target cells, immune complex-mediated activation of complement and Fc receptor pathways, cytokine pathways, and lymphocyte-mediated cytotoxicity of target cells. The nature and sites of the tissue damage are what determine the pathological and clinical features of the specific diseases.

Tissue-specific autoimmune diseases (Table 1)

These occur where immune-mediated damage is restricted to a particular tissue or organ that specifically expresses the targeted antigen. Pertinent examples include: (i) Graves' disease (where autoantibodies bind to and stimulate the thyroid-stimulating hormone receptor, resulting in thyrotoxicosis); (ii) myasthenia gravis (where autoantibodies target the acetylcholine receptor at the neuromuscular junction, resulting in muscular weakness and fatigue due to the inefficient transmission of the acetylcholine signal); and (iii) insulin-dependent diabetes mellitus (where a cytotoxic T-cell response to the b-cells of the pancreatic islets results in destruction of the insulin-producing cells).

Systemic autoimmune diseases (Table 2)

These are frequently characterized by simultaneous damage in multiple tissues (such as kidney, lung, skeletal muscle, nervous system, and skin). Unlike tissue-specific autoimmune diseases which target tissue-specific antigens, autoantibodies in systemic autoimmune diseases are frequently directed against molecules expressed ubiquitously in multiple tissues. Examples include the tRNA synthetases targeted in autoimmune myositis, the small nuclear ribonucleoproteins (snRNPs) targeted in systemic lupus erythematosus, and topoisomerase-1 targeted in scleroderma. Each of these molecules is expressed in all cells, where they play critical roles in essential cellular processes (such as protein translation, mRNA splicing, and DNA replication and remodelling, respectively). Recent studies have suggested that novel forms of these ubiquitously expressed antigens are generated when cells undergo some forms of apoptotic death, and that apoptotic cells may represent an important source of immunogens in this group of disorders. While tissue damage is frequently mediated by numerous mechanisms in systemic autoimmune diseases, deposition of immune complexes at sensitive sites (such as skin, joints, and kidney) represents a prominent mode of tissue damage (see below).

Non-sustained autoimmune diseases

These are characterized by organ or tissue damage and dysfunction, which tends to be self-limited and resolve after the first attack, and are very unlikely to recur (e.g. epidemic Guillain–Barré syndrome). These processes typically occur in the setting of infection, and are associated with cross-reactive antibody responses that recognize both components of the infecting organism as well as the target tissue.

Epidemiology

Autoimmune diseases may affect individuals at all stages of life. In general, diseases have a predilection for beginning after the second decade, with peak incidence in the third to sixth decades. In many instances, females predominate, with the magnitude of this sex difference varying among the different diseases. Thus, for the systemic autoimmune diseases (such as systemic lupus erythematosus, rheumatoid arthritis, Sjögren's syndrome, scleroderma, and autoimmune myositis) and autoimmune thyroid disease, the female:male (F:M) ratio is approximately 4:1 to 9:1, whilst for insulin-dependent diabetes mellitus, multiple sclerosis, and myasthenia gravis, the female predominance is much less prominent (F:M ratio less than 2:1). The exact mechanisms underlying this female predominance remain unknown, but this striking biological difference provides a major clue to pathways underlying susceptibility to autoimmunity.

Aetiology

A general theme in the autoimmune diseases is that the diagnostic phenotype appears subacutely over weeks to months, even though non-specific symptoms and signs frequently predate this. Examples include the fatigue and constitutional symptoms that predate diagnosis of systemic lupus erythematosus and rheumatoid arthritis. The well-developed phenotype represents a highly driven immune response directed against self-antigens, which is amplified between the moment of initial immunization and development of diagnostic disease features. It is therefore operationally useful to divide autoimmune diseases into separate kinetic phases: (i) susceptibility (pre-disease, in which inherited or acquired defects in pathways required to maintain tolerance to self-antigens render the individual susceptible to

disease initiation); (ii) initiation (the interface of susceptibility genes and unique environmental events, which initiate an immune response directed at and driven by self-antigens); and (iii) propagation (a self-amplifying phase in which the specific immune response to self-antigens causes damage of tissues, with the release of more antigens, which further drive the immune response). Although these phases are conceptually distinct, they probably overlap considerably *in vivo*. Diagnostic symptoms and signs represent the highly amplified form of the phenotype generated by the immune system, and generally are kinetically widely separated from the initiating event. This has greatly complicated the study of events underlying disease initiation, since this phase is frequently non-specific and difficult to classify clinically.

Both genetic and environmental factors play important roles in initiation and propagation of autoimmune diseases. They probably play their central roles by regulating the activation, function, and targets of the host immune system. There is also evidence that stochastic processes play an important role in disease initiation, greatly complicating studies to define the causes and mechanisms of autoimmune disease (see below).

Genetic factors

There is accumulating evidence that numerous genes interact to determine the susceptibility threshold for initiating and propagating a self-sustaining autoimmune process in a given individual. Relevant genes include genes that: (i) regulate the immune response; (ii) facilitate efficient, non-inflammatory clearance of apoptotic cells (such as C1q and C-reactive protein); or (iii) influence the target tissue. Recent genetic studies in mice by Wakeland and colleagues in systemic lupus erythematosus (see below), and by Todd and colleagues in human and mouse insulin-dependent diabetes mellitus, have underscored several important observations.

1. Multiple genes interact to generate the final phenotype. Some of these genes render an individual susceptible to initiating an autoimmune response; others affect the target tissue and contribute to the fine disease phenotype.
2. Background genes can have a profound effect on the ability to generate a self-sustaining phenotype. The presence or absence of suppressor genes appears to be particularly important.
3. Genes in the MHC region as well as non-MHC genes appear to play critical roles. While there is a particularly striking contribution from MHC alleles (particularly MHC class II) to disease susceptibility and protection, the mechanisms underlying this phenomenon are still incompletely defined.

MHC class II genes

The associations of MHC class II alleles with disease susceptibility and phenotype can be grouped into the two broad categories that follow.

Association with an increased frequency of the disease itself

Some MHC class II alleles are found at increased frequency in patients with different autoimmune diseases. For example, patients with rheumatoid arthritis have an increased frequency of HLA DR4. HLA DR4 (initially defined serologically) encompasses numerous different alleles that have been defined by sequencing. Interestingly, not all subtypes of HLA DR4 are associated with an increased frequency of rheumatoid arthritis, but those alleles that are associated with this disease share a short amino acid sequence (QKRAA) at positions 70 to 74 of the b-chain of the HLA DR molecule. This sequence, termed the 'shared epitope' is located along the peptide-binding groove of HLA DR4, which presents peptides to the antigen receptor of T cells. Interestingly, this same 'shared epitope' is present in many individuals with rheumatoid arthritis who are positive for HLA DR1.

A similar principle appears to hold for patients with insulin-dependent diabetes mellitus, where there is a strong association of disease with a specific DQb genotype. Whereas the DQb sequence from most normal individuals has an aspartic acid at position 57, most patients with diabetes from the same population group had valine, serine, or alanine at that position.

Since MHC class II molecules function as a scaffold for presentation of specific peptides to T cells (see below), it is possible that this MHC-encoded susceptibility to disease reflects the ability of these alleles to present unique self-peptides to autoreactive T cells. Presentation of these specific peptides may play a critical role in disease initiation and propagation. It is important to note that an added level of complexity appears to be present that has not yet been accounted for. The presence of significant linkage disequilibirum within the MHC region (i.e. large stretches of DNA do not undergo recombination, generating functional cassettes of associated genes) also creates the potential for the disease association of particular MHC alleles to be influenced by additional genes on the extended haplotype in affected individuals.

Determination of which autoantibodies are produced in patients with a particular disease

In some autoimmune diseases, certain MHC class II alleles are strongly associated with the ability to mount a particular autoantibody response in that disease. For example, glutamate at position 34 and leucine at position 26 of the DQa1 and DQb1 chains, respectively, have the strongest association with the ability to make antibodies to Ro and La (ribonucleoprotein antigens in systemic lupus erythematosus/Sjögren's syndrome). Similar observations have been made for numerous other autoantibody specificities, for example anti-DNA, antiphospholipid, and antiribonucleoprotein antibodies. This specificity may again reflect the ability of a particular MHC class II molecule to capture and present self-peptides to T cells. Where specific autoantibodies are associated with unique elements of disease phenotype (such as anti-DNA with renal disease, anti-Ro with photosensitive skin disease), MHC alleles may also be associated with specific disease phenotypes.

Non-MHC genes

Genes outside the MHC affect both susceptibility to, and the phenotypic expression of, autoimmune disease. For example, in studies performed by Wakeland and colleagues in lupus-prone mice, susceptibility and severity of lupus nephritis have been mapped to several different genetic intervals, including regions on chromosomes 1 (*sle1*), 4 (*sle2*), 7 (*sle3*), and 17 (*sle4*). While *sle1* mice exhibit loss of tolerance to chromatin and make autoantibodies to nucleosomes, they do not develop lupus nephritis. Similarly, mice having *sle3* exhibit low-grade polyclonal B- and T-cell activation, and only mild glomerulonephritis. Mice having both *sle1* and *sle3* and female gender show robust autoantibody response (targeting numerous antigens including nucleosomes), splenomegaly, and severe, fatal glomerulonephritis. The definition of the complex genetics underlying autoimmune diseases may delineate those critical pathways required for development of self-sustaining disease, which might be amenable to therapy (see Fig. 1).

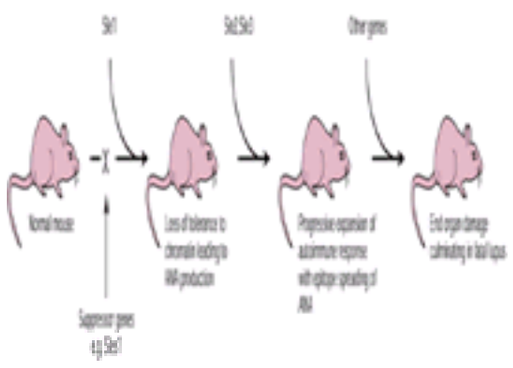

Fig. 1 Genetic susceptibility to autoimmune disease involves multiple interacting loci. Wakeland and colleagues have shown that numerous genes (*sle1*, *sle2*, *sle3*) interact to generate lupus-like autoimmunity in mice. By taking a non-autoimmune mouse (B6), and making a series of mouse strains carrying discrete disease susceptibility intervals, these investigators reproduce fully the phenotype in a lupus-susceptible host (NZM2410). In addition to genes that have a positive effect on generating the autoimmune phenotype, other 'suppressor' genes may counteract the effects of susceptibility genes. (Redrawn from Wakeland EK *et al*., 1999, *Immunity* **11**, 131–9).

Similar types of observations have been made in insulin-dependent diabetes mellitus, where multiple non-MHC genes are associated with development of disease. There has also been the recent description of a single non-MHC gene (AIRE—for autoimmune response—that appears to be a transcription factor) which is strongly

associated with the development of autoimmune polyendocrinopathy with candidiasis and ectodermal dysplasia (APECED) syndrome. The mechanisms whereby single or combinations of genes render individuals susceptible to development of autoimmune diseases are not yet clear, but major strides in this area are likely to be forthcoming within the next 5 to 10 years.

Environmental factors

That environmental insults and stochastic events influence the development of autoimmunity is clear from twin studies and in animal models, as individuals with an identical genotype may be variably affected by disease. For example, the concordance of systemic lupus erythematosus in identical twins is approximately 30 to 50 per cent, and for rheumatoid arthritis it is only about 15 per cent. Many potential environmental insults have been suggested to play a role in autoimmune diseases. These include infections, irradiation, and exposure to drugs and toxins. For example, exacerbations of systemic lupus erythematosus can follow sunlight exposure, and there are numerous reports that disease initiation may have a similar association with ultraviolet irradiation in rare patients. Numerous infections have been postulated to play a role in disease initiation across the spectrum of human autoimmune diseases (see below). In rare cases, the association between antecedent infection and subsequent development of disease is evident (for instance autoimmune myocarditis induced by Coxsackie virus infection, acute rheumatic fever following streptococcal infection, and Epstein–Barr virus infection with childhood systemic lupus erythematosus).

In the majority of autoimmune diseases an environmental connection has not been possible to confirm with any certainty. This does not imply that a causal connection does not exist in these instances, but rather reflects several features of the diseases that greatly complicate firm establishment of such a connection. These are (i) the kinetic complexity of the autoimmune diseases—since establishment of a recognizable disease phenotype often takes months, evidence of the initiating insult may have disappeared by the time the environmental component is sought for the first time; (ii) several different environmental insults may induce a similar response; and (iii) the environmental force may be extremely frequent in the population, and may only induce autoimmune disease in a unique subset of individuals with appropriate susceptibility genes.

How environmental forces influence initiation of autoimmune diseases is not yet known for most autoimmune diseases, but several plausible mechanisms have been advanced. These include: (i) the disruption of cell and tissue barriers, allowing previously sequestered antigens access to a previously ignorant immune system (see below); (ii) inducing novel pathways of antigen presentation; (iii) alteration of the structure of self-antigens; and (iv) molecular mimicry. Some of these mechanisms are dealt with in more detail below.

Pathogenesis

Although extraordinarily complex in detail, the adaptive immune response operates by a set of relatively simple principles: (1) the immune system has the capacity to discern molecular structure in extremely fine detail; (2) it has a uniquely adapted set of signalling systems that computes the amount of antigen; and (3) it responds in a binary way to contextual information, that is, seeing an antigen in the setting of a dangerous context (such as infection) initiates an immune response, while seeing the antigen in the absence of such costimulatory signals leads to tolerance. Numerous studies over the past two decades have underscored that the sustained autoimmune response is extremely similar to adaptive immune responses directed against foreign pathogens, except that the driving antigens in autoimmune disease are self-molecules. For example, autoantibodies in most autoimmune diseases display evidence of isotype switching (for example from IgM to IgG or IgA), and show features of having undergone affinity maturation through somatic hypermutation. These properties of autoantibodies require the activity of antigen-specific CD4+ T cells, and have therefore focused much attention on defining the mechanisms whereby self-reactive T cells are activated in autoimmunity. Since this is such a central issue in the understanding of autoimmunity, and since there are numerous mechanisms employed by the normal individual to prevent activation of autoreactive T cells, it is important to review briefly the mechanisms that the normal immune system uses to maintain tolerance against self-proteins.

Thymic and peripheral T-cell tolerance purges the T-cell repertoire of receptor specificities that recognize self-peptide/MHC complexes

In order to prevent the survival of lymphocytes that will probably encounter their cognate antigens in healthy self-tissues, with potential autoimmune destruction of tissues, the immune system spends significant energy on testing the specificity of all receptors generated during antigen-independent development of lymphocytes, initially in the thymus and subsequently in the periphery. When the T-cell receptor generated through somatic recombination recognizes a peptide/MHC complex in the thymus with high affinity/avidity, cells expressing this receptor are negatively selected (since they are probably self-reactive, and will recognize their cognate antigens at additional peripheral sites). These self-reactive cells undergo apoptosis in the thymus, and never make it into the periphery. In contrast, those T-cell receptors that have some affinity for the selecting MHC molecule, but not for the peptide ontained in the groove, are likely to recognize foreign peptides, and are positively selected. This process of establishing tolerance to self-proteins in the thymus is called 'central tolerance'.

T cells exiting the thymus therefore encompass cells which can recognize peptides within the scaffold of the MHC molecule used to select that T cell, but which have not encountered their specific peptide in the thymus. Since not all self-antigens are expressed in the thymus, there is still a chance that these T cells will encounter a self-peptide/MHC complex in the periphery for which they have high affinity. Since cells that have left the thymus no longer have the developmental context that is likely to denote a self-peptide (that is, recognition with high affinity of a peptide/MHC complex during development in the thymus), peripheral T cells utilize another binary system to define whether a high affinity interaction should lead to activation or inactivation. This binary system uses additional cell surface molecules (called costimulatory molecules) to denote context. Thus, when peripheral T cells recognize a peptide/MHC complex with high affinity in the absence of costimulation (through ligation of CD28 by surface B7.1 or B7.2 on the antigen-presenting cell), T cells are inactivated or tolerized. This is known as peripheral tolerance. In contrast, when peripheral T cells recognize a peptide/MHC complex with high affinity in the presence of costimulation, these T cells are activated.

In addition to T-cell tolerance, B-cell tolerance to self-components is also actively maintained. Thus, if B cells encounter either soluble or membrane-bound antigen during development in the bone marrow, these cells are either deleted (tolerance) or inactivated such that they become refractory to specific stimulation by their antigen (anergy).

Mechanisms which allow an immune response to be directed against self-antigens

Although tolerance to self-molecules is stringently maintained at the T- and B-cell levels, reactivity against self-molecules may still be possible for several reasons. These include the following.

Abnormal immunoregulation

There are numerous mechanisms used to establish and maintain T- and B-cell tolerance. There is accumulating evidence that defects in regulation of these pathways may result in the failure to eliminate autoreactive lymphocytes, or an altered activation threshold for lymphocytes. Examples include defects in the Fas/Fas–ligand system, a receptor–ligand pair which is required for removal of activated, self-reactive lymphocytes. Mice or humans with defects in this pathway manifest profound lymphadenopathy and a spectrum of autoimmunity. Similarly, defects in regulatory molecules which normally function to dampen the immune response (such as CTLA-4, the inhibitory T-cell receptor for the costimulatory molecules B7.1 and B7.2) may result in profound autoimmune responses. Mice lacking CTLA-4 develop fatal autoimmunity, with widespread T-cell infiltrates. Whether similar defects occur in human autoimmune diseases remains to be determined. It should be remembered that the immune system is a highly complex system, with interdependent regulation present at numerous levels. It is likely that many of the susceptibility genes in human autoimmunity impinge on these immunoregulatory pathways.

Existence of sites of immune privilege

Strict sequestration of tissue-specific antigens behind anatomical and immunological barriers prevents the development of tolerance to molecules expressed preferentially at these sites. Events (such as penetrating trauma) which breach this tight boundary may allow initiation of an immune response to these previously hidden self-molecules. Relevant examples include antigens within the eye, testis, and central nervous system. In the eye for instance, penetrating injury to one eye may be followed by development of inflammation in the contralateral eye (sympathetic ophthalmia) approximately 1 to 2 weeks after injury. Several mechanisms have been proposed to be responsible for maintaining the immune-privileged status of these tissues. One powerful mechanism appears to involve the constitutive expression of Fas–ligand in the relevant tissue (such as the eye). When this molecule binds to and activates its receptor on lymphocytes, these cells undergo apoptotic death, and are prevented from entering the tissue.

Cryptic determinants within self-molecules that are not normally revealed during antigen processing by default pathways, and allow the persistence of

potentially autoreactive lymphocytes

Not all regions of a molecule are equally immunogenic. Regions of the molecule that are captured well by class II MHC molecules during natural processing of self-antigens are able to tolerize T cells (these determinants have been termed 'immunodominant' by Sercarz and colleagues). In contrast, regions of self-molecules that are not generated in significant amount during natural antigen processing (so-called 'cryptic determinants') cannot effectively tolerize T cells, since they are never seen by these cells either in the thymus or peripherally. This immunodominance appears to be influenced by the intrinsic affinity of the peptide for MHC class II, as well as by neighbouring structural determinants on the antigen that may influence its binding to the peptide-binding groove. On self-molecules, two sets of determinants can therefore be defined (Fig. 2):

1. those that are easily processed and presented (comprising the dominant self), which readily tolerize developing T cells.
2. those that are not presented in appreciable amounts after natural processing (comprising the cryptic self), which do not tolerize.

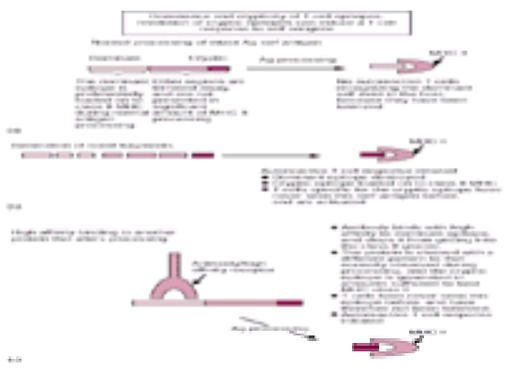

Fig. 2 Dominant and cryptic T-cell epitopes in autoimmune disease. (a) The default processing pathway for intact antigen results in the preferential and reproducible loading of the 'dominant' peptide determinant into the antigen-binding groove of MHC class II. During establishment of thymic and peripheral tolerance, T cells recognizing this dominant epitope are purged from the repertoire, but T cells recognizing cryptic epitopes do not encounter their antigens, and are not deleted or anergized. (b) and (c) When the processing of self-antigens is altered (for example by novel proteolysis or through high-affinity binding to another molecule), a different hierarchy of epitopes is loaded on to class II MHC. If cryptic epitopes are loaded in sufficient amounts, these peptides can stimulate autoreactive T-cell responses directed against the cryptic self, and drive the autoimmune process.

There are unusual circumstances in which natural processing of self-antigens is altered from the default pathway. Examples include novel proteolysis of autoantigen (which destroys the dominant epitope or generates a new dominant epitope) prior to entry into the processing pathway, as well as high-affinity binding to specific receptors or antibodies, which can hinder access of the dominant epitope to the antigen-binding groove of MHC class II molecules, or optimize the loading of a previously cryptic epitope. Since T cells recognizing these cryptic peptides have not previously been tolerized, such 'autoreactive' T cells can now be activated (Fig. 3).

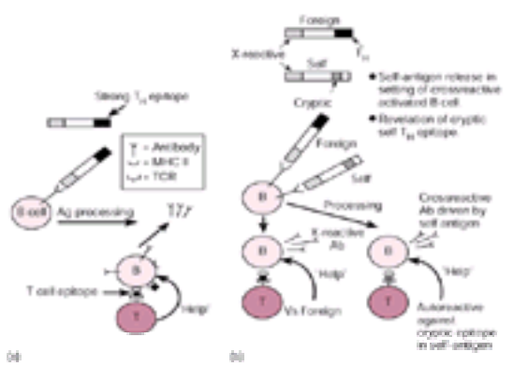

Fig. 3 Molecular mimicry. (a) Foreign antigens, which clearly differ from their homologous self-antigens in some areas, may nevertheless bear significant structural similarity to self-antigens in other regions. Initiation of an immune response to the foreign antigen may generate a cross-reactive antibody response that also recognizes the self-protein. When the self-antigen is a cell surface molecule, antibody-mediated effector pathways can lead to host tissue damage. Although the antibody response is cross-reactive with self-molecules, the T cells that drive this response are directed exclusively at the foreign antigen. (b) Under highly novel conditions, the simultaneous liberation of significant amounts of self-antigen in the setting of a cross-reactive antibody response may allow effective presentation of cryptic epitopes in the self-antigen to autoreactive T cells by activated cross-reactive B cells. These autoreactive T cells can now continue to drive an autoantibody response to the self-antigen. If continued release of self-antigen occurs as part of this process, a specific, adaptive immune response to self will be sustained.

There are several clear demonstrations that autoreactive T cells recognizing cryptic epitopes can be activated *in vivo* through altered processing of self-molecules to reveal these previously immunocryptic epitopes. For instance, high-affinity binding of the HIV surface protein gp120 to CD4 alters the processing of CD4, and activates T cells which recognize epitopes of CD4 not generated during normal antigen processing. This mechanism may account for the autoimmune response to CD4 seen during HIV infection. Similarly, although intact mouse cytochrome c is not immunogenic in mice, cleavage of the molecule into smaller peptides induces a robust T-cell response to cryptic areas of cytochrome c, which were never previously presented by the natural processing pathway, and therefore did not induce tolerance.

The revelation of cryptic epitopes in self-antigens is likely to be a highly relevant mechanism in many human autoimmune diseases, but the studies to demonstrate the importance of this mechanism have only recently begun in earnest. Since the structure of autoantigens influences the hierarchy of dominant and cryptic and determinants generated when the molecule is processed, unique processes which alter the structure of molecules may play critical roles in initiation of autoimmune diseases. These unique events are not likely to occur during normal homeostasis, but may occur preferentially during infectious or other pro-immune events at the host–environment interface. Relevant examples include the following.

1. Unique proteolytic pathways are activated that specifically alter the structure of autoantigens during immune effector pathways. It has recently been observed that the majority of autoantigens targeted across the spectrum of human autoimmune diseases are specifically cleaved by granzyme B during killing of infected target cells by cytotoxic lymphocytes. This cleavage generates unique molecular fragments never generated in the organism during development or homeostasis. Interestingly, this cleavage is a unique feature of autoantigens, and does not affect non-autoantigens. Although it has been proposed that these cleavage events allow the efficient presentation of previously cryptic epitopes, this remains to be formally demonstrated.
2. Additional post-translational modifications alter conformation of antigens and modify their subsequent processing. It is noteworthy that numerous post-translational modifications of autoantigens occur, and that in some cases the autoimmune response is strictly dependent on the occurrence of these modifications. Examples include phosphorylation, acetylation, deimination, and isoaspartyl formation, amongst others.
3. High-affinity complexes are formed between autoantigens and other viral or self-proteins.

In all these examples it should be remembered that the initiating event in autoimmunity requires that, on the background of appropriate susceptibility genes, several stringent criteria needed to initiate a primary immune response must be simultaneously satisfied. These include the generation of suprathreshold concentrations of self-molecules that have a structure not previously tolerized by the immune system, and the presentation of these unique molecular forms to T lymphocytes in the presence of costimulation (that is, in a pro-immune context).

Molecular mimicry

Foreign antigens, which clearly differ from their homologous self-antigens in some areas, may nevertheless bear significant structural similarity to self-antigens in other regions. Initiation of an immune response to the foreign antigen may generate a cross-reactive antibody response that also recognizes the self-protein (molecular mimicry). When the antigen is a cell surface molecule, antibody-mediated effector pathways can lead to host tissue damage. Although the antibody response is cross-reactive with self-molecules, the T cells that drive this response are directed at the foreign antigen (see below). Diseases involving this sort of 'antigen mimicry' therefore tend to be self-limited. It is important to realize that molecular mimicry alone cannot explain self-sustaining autoimmune diseases, which are driven by self-antigens and autoreactive T cells. In these cases, there is a requirement for overcoming T-cell tolerance to the self-protein. The simultaneous liberation of self-antigen in the presence of the cross-reactive antibody response is likely to play a critical role in this regard (see below).

Mechanistic insights into molecular mimicry: source of cross-reactive antibodies and potential role in overcoming T-cell tolerance to self-proteins

Although a number of microbial and viral antigens have regions of high homology with various human autoantigens, a causal link between exposure to these foreign antigens and the onset or exacerbation of autoimmune disease has been extremely difficult to establish. However, there are clear examples which suggest the existence of 'one-shot' autoimmune processes, in which cross-reactive antibodies directed against surface self-antigens are generated following infection, and result in tissue damage. This persists until infection is cleared, and the immune response wanes. Although the mechanistic details of this scheme are difficult to prove *in vivo*, several pertinent examples exist. One of these is a seasonal epidemic form of Guillan–Barré syndrome seen in northern China, which follows *Campylobacter jejuni* infection. Affected patients make antibodies recognizing gangliosides, and the disease has a self-limited course, which rarely recurs. The anti-ganglioside antibodies generated are probably responsible for the pathological findings of acute motor axonal neuropathy. Another plausible example of this mechanism (although with meager *in vivo* evidence) is immune thrombocytopenia in children. This process characteristically: (i) follows an infectious process; (ii) demonstrates antiplatelet antibodies, and (iii) frequently shows durable remissions. The mechanistic details of this process have been difficult to prove *in vivo*, and cross-reactive epitopes on potentially initiating pathogens have not yet been defined.

The single episodes of tissue damage in the setting of a cross-reactive immune response following infection must be contrasted to the sustained, autoamplifying disease frequently seen in other autoimmune syndromes. The central issues in this regard are: (i) how T-cell tolerance to self-antigens might initially be broken, and (ii) once this has occurred, why these antigens continue to drive the immune response to self. Examination of tolerance to cytochrome c, a ubiquitous protein that has regions of homology and divergence across different species, has been very useful in understanding molecular mimicry of cross-reactive epitopes. Mouse cytochrome c shares significant homology with human cytochrome c, although they are entirely different in other areas. When Mamula and colleagues used mouse cytochrome c to immunize mice, no T-cell or antibody response to the murine protein was observed. When human cytochrome c was similarly used to immunize mice, strong T-cell epitopes on the foreign cytochrome c were able to induce a strong antibody response to the foreign protein. The antibodies induced recognized both the murine and the human forms of cytochrome c, that is, cross-reactive antibodies that recognize the self-protein were produced. However, the T-cell response to cytochrome c was directed entirely against the foreign (human) form of the protein, and no T cells against the murine protein could be found. These cross-reactive antibodies disappear as the immune response to the foreign protein wanes.

Interestingly, when mouse cytochrome c was included with human cytochrome c during the immunization, a T-cell response to human cytochrome c, and a humoral response to the human protein that cross-reacts with the murine protein were induced. Within a few days, a strong helper T-cell response specific for murine cytochrome c was detected. This breaking of T-cell tolerance to murine cytochrome c was dependent on activated B cells specific for cytochrome c, which probably exert their effect through altering the processing of mouse cytochrome c, potentially uncovering previously cryptic epitopes in the self-protein (see Fig. 3). In the presence of continued release of self-antigen, this response may become self-sustaining—self-antigen driving autoreactive T cells, providing help to autoantibody-producing B cells (Fig. 3).

Molecular mimicry may therefore induce the production of cross-reactive antibodies, which in the absence of liberation of significant amounts of self-antigen, should disappear when the foreign pathogen is cleared. The form of epidemic motor axonopathy described above is probably representative of this scenario. Under highly novel conditions, the simultaneous liberation of significant amounts of self-antigen in the setting of a cross-reactive antibody response may allow effective presentation of cryptic epitopes in the self-antigen to autoreactive T cells by activated cross-reactive B cells. If continued release of self-antigen occurs, a specific, adaptive immune response to self will be sustained. Antigen release from tissues probably plays a critical role in driving this autoimmune process. Understanding the mechanisms of ongoing antigen release at sites of tissue damage in autoimmune disease (such as unique pathways of cell injury and death) is a high priority for future work, as it provides a novel target for therapy (see below).

It is clear from the above discussion that extraordinary complexity is operative in initiation of the human autoimmune diseases. The patient population is genetically heterogeneous, the human immune system is complex and extremely plastic, and it interacts with a plethora of environmental stimuli and stochastic events. The simultaneous confluence of susceptibility factors and initiation forces to set off the self-sustained and autoamplifying process is therefore an extremely rare occurrence. In contrast, once activation of autoreactive T cells has occurred, the ability of the immune system to respond vigorously to vanishingly low concentrations of antigen, to amplify the specific effector response to those antigens, and to spread the response to additional antigens in that tissue, greatly reduces the stringency that must be met to keep the process going (Fig. 4).

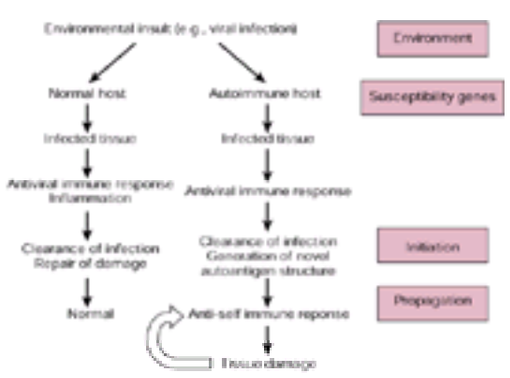

Fig. 4 Model of initiation and propagation of autoimmune disease. Autoimmune diseases are highly complex disorders which require the simultaneous co-operation of multiple factors for their development. Numerous susceptibility genes (some of which regulate the immune response) appear to determine the threshold for disease initiation. In many diseases, a discrete, pro-immune environmental trigger probably plays a role in disease initiation, but is infrequently recognized. A critical requirement for disease initiation is the generation of suprathreshold concentrations of self-antigen with novel structure. Development of a recognizable disease phenotype generally requires marked antigen-driven amplification of the autoimmune response, in which immune effector pathways play a role in generating the ongoing supply of antigen to sustain the process.

Effector mechanisms in autoimmune diseases

The initiation phase of autoimmunity requires co-operation between many different cell types, including antigen-presenting cells, T cells, and B cells, as well as numerous soluble mediators including antibodies, chemokines, and cytokines. The effector phase of autoimmunity uses the same immune and inflammatory effector mechanisms that the immune system has evolved for removing and destroying pathogens. These include activation of the complement cascade, which generates signals that effect inflammatory cell recruitment and activation. Similarly, ligation of activating Fc receptors on inflammatory cells by immune complexes activates macrophage and neutrophil effector function. Autoantibodies directed against cell surface antigens initiate antigen-dependent cellular cytotoxicity, probably mediated by macrophages and natural killer cells. Cytokines and chemokines play a central role in inflammatory cell recruitment and activation in the target tissue. Tissue damage can also be effected by cytolytic lymphocytes. The pathology characteristic of each autoimmune disease reflects the particular antigens targeted as well as the predominant effector mechanisms activated.

One principle of central importance in the effector phase of autoimmunity is autoamplification (Fig. 4), which appears to play a central role in the self-sustaining nature

of the autoimmune process. Thus, immune effector pathways cause damage to cells in the target tissue, liberating antigen which further stimulates the immune response and effector pathways, thus liberating more antigen. Although this is probably an oversimplification, the view that the immune system plays a role in generating an ongoing supply of autoantigen is useful therapeutically, since it focuses attention on controlling both the supply of antigen as well as immune effector pathways (see below).

Clinical features

The clinical features of the different autoimmune diseases are extremely diverse, and reflect the specific tissue dysfunction which results from activity of immune effector pathways. Almost all tissues may be affected, including prominent involvement of endocrine organs, nervous system, eye, bone marrow elements, kidney, muscle, skin, liver and gastrointestinal tract, blood vessels, lung, and joints. For tissue-specific autoimmune processes (such as insulin-dependent diabetes mellitus, immune thrombocytopenia, autoimmune haemolytic anaemia (AIHA)), symptoms may relate to tissue hypofunction resulting from: (i) target cell destruction (for insulin-dependent diabetes mellitus, destruction of the b-cells of the pancreatic islets; for immune thrombocytopenia and AIHA, destruction and phagocytosis of platelets and erythrocytes); or (ii) antibody-mediated interference with function or down-regulation of autoantigen expression (for example in myasthenia gravis, bullous pemphigus). In other cases, symptoms may arise from tissue hyperfunction (for example in Graves' disease) due to activating effects of antibody binding (where antibodies to the thyroid-stimulating hormone receptor induce non-physiological thyroid hormone secretion).

In the case of systemic autoimmune processes, symptoms frequently result from localized target tissue destruction (for example skeletal muscle in polymyositis, skin disease in systemic lupus erythematosus) as well as from the more general activities of inflammatory effector pathways. The latter result from: (i) immune complex deposition at multiple sensitive sites (such as joints, kidney, skin, and blood vessel walls) with activation of the complement cascade and recruitment and activation of myelomonocytic cells; and (ii) ongoing secretion of pro-inflammatory cytokines. In this regard, the profoundly positive effects of tumour necrosis factor inhibition recently observed on the inflammatory symptoms and joint destruction in rheumatoid arthritis underscore the central role of these general inflammatory mediators in generation and maintenance of the disease phenotype in systemic autoimmune diseases.

Prognosis

While the barriers that need to be overcome in terms of initiating an autoimmune disease are highly stringent, and are very difficult to satisfy even in the setting of appropriate susceptibility genes, the immune system is equipped with a powerful memory. The mechanisms of memory are still incompletely defined, but include the generation of a population of memory cells specific for the antigen that initiated the response, which respond vigorously (in terms of clonal expansion as well as effector function) to very low concentrations of antigen if they encounter it again. Since the autoimmune diseases are disorders driven by the ongoing release of self-antigen, this immunological memory constitutes a major barrier to complete cure. Autoimmune diseases therefore tend to be self-sustaining over long periods, and are often punctuated by clinical exacerbations (flares), which are probably due to re-exposure of the primed immune system to antigen (for example in systemic lupus erythematosus, autoimmune myositis, and rheumatoid arthritis). The possibility of disease recurring, even after long clinical remission, remains present in most of the autoimmune diseases. Tissue-specific autoimmune diseases may result in the complete destruction of the target tissue over time, with loss of function of that tissue accompanied by a waning immune response (for example in insulin-dependent diabetes mellitus). Interestingly, in cases where immune-mediated tissue pathology results from effector pathways being driven by a cross-reactive T-cell response to a foreign antigen (for example in epidemic Guillain–Barré syndrome), disease has a finite duration, and generally does not recur.

Therapy

It is not possible to discuss the therapy of this broad group of disorders in any detail in this chapter, but a few principles that underlie current approaches to therapy are discussed. Autoimmune diseases cause significant tissue dysfunction through: (i) inflammation, (ii) tissue destruction with loss of functional units, (iii) the consequences of healing, and (iv) functional disturbances (such as interference with acetylcholine signalling by autoantibody to the acetylcholine receptor and inducing receptor down-modulation in myasthenia gravis). Therapeutic interventions in autoimmune diseases are generally focused on controlling immune and inflammatory pathways, and at replacing or accommodating lost function.

Controlling the immune and inflammatory pathways responsible for ongoing damage

Since in the majority of instances, the critical autoantigens and effector pathways responsible for unique diseases have not been defined, this goal is frequently extremely challenging to accomplish. Thus, frequent use is made of anti-inflammatory and immunosuppressive therapies which broadly target many aspects of the immune response (such as steroids, azathioprine, cyclophosphamide, methotrexate, and mycophenolate). Since a robust immune response is required to protect the host from a myriad of infectious threats, this non-targeted suppression of the immune system can have deleterious consequences in terms of increased susceptibility to infection, with its attendant high morbidity and mortality. In this regard, therapeutic targeting of specific inflammatory pathways is extremely attractive, and there are recent examples in which this approach has been highly successful. In rheumatoid arthritis, the maintenance of chronic inflammatory joint pathology appears to be dependent on the activity of tumour necrosis factor. Specific inhibition of tumour necrosis factor through the use of either soluble tumor necrosis factor receptors or humanized monoclonal antibodies has led to an astonishing effect on disease activity in rheumatoid arthritis, with abolition of systemic symptoms, and a striking decrease in the rate of joint destruction. These positive effects were associated with only a minimal increase in susceptibility to infection, although this risk is certainly present. Inhibition or stimulation of specific inflammatory pathways as therapy for autoimmune diseases may well be more broadly applicable, and will certainly be tested in other autoimmune processes once critical roles of additional specific inflammatory pathways are defined.

Another example of specific targeting of pro-inflammatory pathways is that of intravenous immunoglobulin. This is prepared from pooled serum and its major component is immunoglobulin G. Intravenous immunoglobulin therapy has been used as a treatment for several autoimmune diseases, including immune thrombocytopenia, autoimmune myositis, and acute demyelinating polyneuropathy, but is only available at prohibitive cost. Recent data from mice have demonstrated that intravenous immunoglobulin induces surface expression of the inhibitory Fcg receptor (FcgRII $_{B}$) on macrophages, and shifts the balance of signalling through Fc receptors towards inhibition, down-regulating the pro-inflammatory response to immune complexes. It is likely that continued identification of additional agents that precisely modulate specific inflammatory pathways will have a major therapeutic impact on this group of diseases.

Interventions aimed at replacing or accommodating lost function

The majority of autoimmune diseases are associated with loss of function of organs and tissues, many of which perform essential physiological functions. Indeed, recognition of the autoimmune phenotype in many instances requires that tissue damage is sufficiently severe to have led to characteristic loss of function. For example, loss of insulin-secreting b-cells of the pancreatic islets results in insulin-dependent diabetes mellitus, and blockade and down-regulation of the nicotinic acetylcholine receptor causes striated muscle weakness and fatigue in myasthenia gravis. Similarly, chronic immune complex deposition in glomeruli causes renal inflammation and scarring in systemic lupus erythematosus. Where significant functional reserve is still present in a particular disease, a strong argument can be made for preventing further damage through specific or general immunosuppressive strategies described above. This is particularly relevant where the 'supply' of tissue that could be damaged is essentially inexhaustible (for example most instances of systemic autoimmune disease). Where functional impairment is already established, interventions aimed at replacing or accommodating lost function are indicated. For example, insulin replacement is required for insulin-dependent diabetes mellitus, and treatment for hyperthyroidism is indicated in Graves' disease.

Summary

Autoimmune disease results when the immune system becomes activated to recognize self-antigens. The response is antigen-driven and T cell-dependent, and is directed against highly specific autoantigens that are in many instances disease specific. The genetic contribution to autoimmunity is important, with MHC and non-MHC genes playing significant roles. MHC genes may confer susceptibility to disease in some cases, and determine the autoantibodies produced in others. The essence of sustained autoimmune disease is the breaking of T-cell tolerance to self-molecules, resulting in a sustained immune response to self, and consequent tissue damage. Although the mechanisms by which tolerance is broken remain unclear, it is likely that, in the genetically susceptible host, environmental influences play an important role in the initiation of an autoimmune response. Possible mechanisms include alteration of antigen structure, location, concentration, processing, presentation, and context. The pathology characteristic of each autoimmune disease reflects the particular antigens targeted as well as the immune effector mechanisms activated. Ongoing immune-mediated damage probably plays a central role in providing autoantigen to drive the continuing autoimmune response.

Further reading

Diamond B *et al*. (1992). The role of somatic mutation in the pathogenic anti-DNA response. *Annual Review of Immunology* **10**, 731–57. [Comprehensive review of the evidence that autoantibodies are antigen driven and T-cell dependent.]

Feldmann M, Brennan FM, Maini RN (1996). Rheumatoid arthritis. *Cell* **85**, 307–10. [Short review of the complex pathogenesis of rheumatoid arthritis.]

Gammon G, Sercarz EE, Benichou G (1991). The dominant self and the cryptic self: shaping the autoreactive T-cell repertoire. *Immunology Today* **12**, 193–5. [Concise introduction to the concepts and consequences of immunodominance.]

Gianani R, Sarvetnick N (1996). Viruses, cytokines, antigens, and autoimmunity. *Proceedings of the National Academy of Science USA* **93**, 2257–9. [Review of the pathogen–host interface and autoimmunity.]

Kotzin BL (1996). Systemic lupus erythematosus. *Cell* **85**, 303–6.

Lanzavecchia A (1995). How can cryptic epitopes trigger autoimmunity? *Journal of Experimental Medicine* **181**, 1945–8. [Important review of potential mechanisms of autoimmunity.]

Lin R-H *et al*. (1991). Induction of autoreactive B cells allows priming of autoreactive T cells. *Journal of Experimental Medicine* **173**, 1433–9. [Important demonstration that autoreactive B cells may alter the processing of self-antigens to allow activation of autoreactive T cells.]

Morel L *et al*. (1999). Epistatic modifiers of autoimmunity in a murine model of lupus nephritis. *Immunity* **11**, 131–9. [Clear definition of the complex genetics of systemic lupus erythematosus.]

Naparstek Y, Plotz PH (1993). The role of autoantibodies in autoimmune disease. *Annual Review of Immunology* **11**, 79–104.

Radic MZ, Weigert M (1994). Genetic and structural evidence for antigen selection of anti-DNA antibodies. *Annual Review of Immunology* **12**, 487–520. [Comprehensive review of evidence that autoimmunity is driven by self-antigen.]

Rosen A, Casciola-Rosen L (1999). Autoantigens as substrates for apoptotic proteases: implications for the pathogenesis of systemic autoimmune disease. *Cell Death and Differentiation* **6**, 6–12. [Revew of role of altered autoantigen structure in autoimmune diseases.]

von Muhlen CA, Tan EM (1995). Autoantibodies in the diagnosis of systemic rheumatic diseases. *Seminars in Arthitis and Rheumatism* **24**, 328–58. [General review of autoantibodies and their specificities.]

Wicker LS, Todd JA, Peterson LB (1995). Genetic control of autoimmune diabetes in the NOD mouse. *Annuual Review of Immunology* **13**, 179–200. [Review of the complex genetics of mouse autoimmune diabetes.]

5.4 Complement

Mark J. Walport

Introduction

Jules Bordet first discovered complement as an activity in serum that complemented the activity of antibody in the killing of bacteria—hence its name. Complement comprises a group of more than 20 plasma and cell-bound proteins and is part of the innate immune system. The direct binding of particular complement proteins to potential pathogens activates a cascade of sequentially acting complement proteins.

Complement is a triggered enzyme cascade, initiated by the binding of any of three complement proteins, C1q, mannose-binding lectin, or C3, to acceptor molecules, especially on the surface of potential pathogens. The binding of these molecules may be direct or, in the case of C1q, to antibody bound to antigens (immune complexes). The binding of C1q to immune complexes represents an important bridge between the adaptive immune system and the innate immune system.

After initiation, the activation of complement is amplified by the sequential activation of a series of enzymes, which lead ultimately to the cleavage of C3 and C5. Thus a small initiating signal, for example from binding of a few mannose-binding lectin molecules to the surface of a bacterium, leads to cleavage of a large number of C3 and C5 molecules. These amplification steps in complement activation increase the effectiveness of complement as a host defence mechanism but also carry the risk to the host that inappropriate complement activation may cause bystander inflammatory injury to host tissues. To prevent this, there is a large array of regulatory mechanisms that prevent inappropriate complement activation and reduce the chance of complement injury to self tissues.

A detailed description of the biochemistry of complement is beyond the scope of this chapter; recent reviews that describe this are provided at the end of the chapter. This chapter will focus on the diseases associated with abnormalities of the complement system. We will first consider the different physiological activities of complement, which is necessary in order to understand the role of complement in disease. We will then review the diseases associated with hereditary disorders of complement, followed by diseases in which there are acquired complement abnormalities. The chapter will end with a consideration of when and how assays of the complement system should be performed in the assessment and management of disease.

Physiology of complement

The physiological activities of complement are summarized in Table 1. There are three overarching activities. The first is the role of complement in host defence against infectious disease. Complement provides mechanisms for the killing and clearance of micro-organisms; it does this by the covalent binding to their surface of C3 and C4 fragments that are ligands for receptors on phagocytic cells that ingest and kill the organism. The activation of complement also causes the generation of the anaphylatoxins, C5a and C3a, which have chemotactic activity and recruit leucocytes to sites of infection and inflammation. A further role of complement in host defence against infection is generation of the membrane attack complex. This may disrupt the cell membrane and kill the micro-organism.

The second activity of complement is as a bridge between the humoral adaptive immune system (antibody) and innate immunity. Activation of complement by immune complexes facilitates the clearance of antigen and thereby helps to prevent immune complexes from causing inflammatory damage to tissues, although, as we shall see, complement may contribute to inflammatory injury to tissues in circumstances when immune complexes persist. Activation of complement also augments antibody responses and thereby enhances host defence against pathogens. The binding of complement to antigens reduces the threshold of B lymphocytes for activation. It enhances antigen presentation and B cell memory by helping to localize antigen on antigen-presenting cells and on the follicular dendritic cells that are key to the maintenance of B cell memory for foreign antigens.

The third activity of complement is in the resolution of inflammatory responses. It is in this role that complement may prevent the development of systemic lupus erythematosus (SLE) by promoting the clearance of tissue debris.

Complement in disease

Hereditary disorders

Studies of the inherited abnormalities of the complement system have illuminated our understanding of the major roles of the complement system *in vivo*. There are three types of disease associated with hereditary complement deficiency. The first is immunodeficiency, which illustrates the role of complement in host defence against infection. The second is the association of systemic lupus erythematosus with deficiency of certain classical pathway proteins. This association has led to a greater understanding of the role of complement in the resolution of inflammation and in the waste disposal mechanisms of the body. The third category of disease is caused by deficiencies of proteins of the regulatory mechanisms of the complement system. This small group of diseases illustrate the effects of unrestrained activation of the complement system. We will consider each of these associations.

Complement deficiency and infection

Patients lacking C3 and the pathways leading to C3 activation show increased susceptibility to pyogenic infections with bacteria such as *Streptococci* and *Staphylococci*. There is a similar susceptibility to infections in patients lacking antibodies or normal phagocytic function (see Chapter 5.5 and Chapter 5.6). This shows that the normal pathway for host defence against such bacteria is binding of antibody, followed by complement, providing opsonins for uptake and bacterial killing by phagocytes. Disruption of any of the links in this chain causes increased susceptibility to infection by these pyogenic bacteria.

Neisserial infection and complement deficiency

Humans who lack one of the proteins of the membrane attack complex (C5, C6, C7, C8, or C9) display a unique susceptibility to neisserial infection, especially by *Neisseria meningitidis*, which is frequently recurrent (Fig. 1 and Plate 1). This pattern of infection shows that host defence against these bacteria, which are capable of intracellular survival, is mediated by their lysis by the membrane attack complex. Individuals lacking the earlier components of the complement system, which are the necessary precursors for the formation of the membrane attack complex, are also at increased risk of neisserial infection. Deficiency of properdin is also especially associated with neisserial infections. This protein stabilizes the alternative pathway C3 convertase enzyme and augments the cleavage of C3. It is encoded on the X chromosome and therefore properdin deficiency is found almost exclusively in males. Increased susceptibility to neisserial infections is also a feature of acquired complement deficiency, such as may be seen in patients with SLE or with C3 nephritic factor.

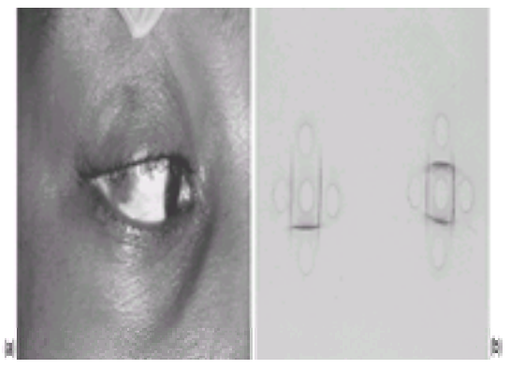

Fig. 1 Patient with hereditary deficiency of C6 who presented with meningococcal septicaemia. (a) a subconjunctival haemorrhage. (b) The deficiency of C6. Serum from the patient was placed in the central well of an agarose-coated plate. In each of the outer wells was placed antiserum to, respectively, C5, C6, C7, and C8. The antibody and antigen were allowed to diffuse in the gel and where the antibody encountered its antigen a precipitate formed, which was stained blue. No precipitate formed between the anti-C6 antibody and the patient's serum, indicating C6 deficiency. (See also Plate 1.)

When should complement deficiency be suspected in a patient with infectious disease? Immunodeficiency should be suspected in any individual who has recurrent or unexplained major infections. The type of infection provides a clue to the relevant investigations of the immune system. Recurrent pyogenic infections imply a need to assay the activity of antibodies, the complement system, and phagocytic function.

In the specific case of meningococcal sepsis, factors that point to complement deficiency are recurrent attacks, a family history of meningococcal infection (especially if disseminated in time), or infection by unusual strains of *N. meningitidis*. Individuals with complement deficiency remain susceptible to neisserial infection throughout life and may present at any age.

Mannose-binding lectin deficiency

Mannose-binding lectin is a protein homologous in structure to C1q that initiates complement activation by a pathway similar to the classical pathway. Mannose-binding lectin binds to terminal mannose groups in a spatial orientation that is present on many micro-organisms, including certain Gram-positive and Gram-negative bacteria, mycobacteria, yeasts, and parasites, but absent on mammalian cells. It is one of the 'pattern-recognition' molecules of the innate immune system that binds molecules present on potential pathogens but not on the cells of the host. The importance of its role in host defence was identified when it was discovered that a group of children with recurrent bacterial infections in early childhood were deficient in the ability of their serum to opsonize yeast with C3 *in vitro*. It turned out that this *in vitro* opsonic deficiency was caused by a subtotal deficiency in the expression of mannose-binding lectin. The most important causes of this common deficiency are mutations of residues in the collagen domain of mannose-binding lectin, which cause misassembly of the multimer and thereby have a dominant effect suppressing mannose-binding lectin levels.

The clinical effects of mannose-binding lectin deficiency are most apparent in young children from the ages of 2 to 5, when maternal passive immunity has waned and the antibody response has not yet matured. At this stage of life the innate immune system is of particular importance in host defence against infection.

Complement deficiency and autoimmune disease

A dramatically increased prevalence of SLE is found amongst patients with deficiencies of proteins of the classical pathway of complement. There is a hierarchy of susceptibility and severity of SLE according to the position of the missing protein in the pathway of classical pathway activation (Table 2).

These cases of SLE, associated with inherited complement deficiency are, *in toto*, extremely rare and only account for a tiny minority of the population of patients with SLE. However, they provide an important clue to the aetiology of the disease. They show that there is an important activity of the early classical pathway of complement that protects against the development of SLE. The source of the autoantigens that drive the autoimmune response in SLE is thought to be apoptotic cells. Complement has been found to play a role in the disposal of apoptotic cells and in the processing of immune complexes. Loss of these activities might lead to abnormal processing of effete cells, that, in the context of an inflammatory response, could initiate and drive an autoimmune response leading to the development of SLE.

Hypotheses such as this may be tested in animal models of disease. A series of mice have been developed lacking molecules that have been implicated in the 'waste-disposal' mechanisms of the body. These include mice lacking C1q, serum amyloid P component (which coats and may mask extracellular chromatin from the immune system), DNase 1 (which digests extracellular chromatin), and IgM (which may augment the clearance of effete cells and cellular debris). Each of these spontaneously develops SLE and supports the hypothesis that effective mechanisms of cellular waste disposal are essential to prevent the development of SLE.

Abnormalities of complement regulation

C1 inhibitor deficiency

The disease hereditary angioedema is caused by deficiency of C1 inhibitor. This is inherited as an autosomal dominant disorder with partial penetrance. The disease is dominantly inherited because the production of C1 inhibitor from a single, normal allele is insufficient to maintain normal homeostasis of the complement and kinin pathways. The mutations may have two effects on protein production. In type 1 hereditary angioedema, which accounts for approximately 85 per cent of cases of the disease, the mutant prevents any expression of protein from the mutant allele. This variety of disease is therefore associated with reduced levels of C1 inhibitor.

Type 2 hereditary angioedema is caused by a series of point mutations in the C1 inhibitor gene that alter one of the amino acids at the active centre of the protein and abolish its activity as a serine proteinase inhibitor. These mutations allow expression of normal amounts of protein, which is non-functional, or even abnormally high C1 inhibitor levels, because the mutant protein is not consumed by normal interaction with activated serine proteinases. It is easy to miss the diagnosis of this variant of hereditary angioedema, if it is not appreciated that levels of C1 inhibitor can be normal or high in patients with the disease.

The clinical manifestations of hereditary angioedema are caused by vascular leakage of fluid that cause angioedematous swellings. The swellings are caused by the action of small peptides, called kinins, in particular bradykinin, that cause increased vascular permeability by their actions on vascular endothelium and smooth muscle. These kinins are produced by the action of serine proteinases that are ineffectively regulated in the presence of reduced activity of C1 inhibitor.

Allergy is much commoner than hereditary angioedema as a cause of angioedema. In hereditary angioedema, the swelling is not itchy and is not accompanied by other features of allergy such as asthma and urticaria. The disease is potentially life-threatening if there is major pharyngeal or laryngeal swelling, causing airways obstruction. Swelling of the submucosa of the intestines may cause severe abdominal pain and temporary obstruction of the bowel.

Diagnosis of hereditary angioedema is made on the basis of the clinical findings described above, the presence of family history, and blood tests. A family history of angioedema makes diagnosis much easier but is not always present. This is because some cases of the disease are due to new mutations in the C1 inhibitor gene. In other families, other members with C1 inhibitor deficiency may have no clinical symptoms.

The abnormal blood tests associated with hereditary angioedema are reduced C1 inhibitor protein levels (usually below 30 per cent of normal levels) in patients with type 1 hereditary angioedema. However, in the 15 per cent of patients with type 2 disease, protein levels may be normal or high. In these patients functional assays of C1 inhibitor are necessary. These are based on the ability of C1 inhibitor to block cleavage of a chromogenic substrate by activated C1s. In addition to these abnormalities, levels of C2 and C4 are typically low. This is because the reduced C1 inhibitor activity allows C1s to cleave C4 and C2 in an unregulated fashion.

Treatment of the disease involves, firstly, the treatment of acute attacks and, secondly, prophylaxis to attempt to prevent their recurrence. Acute attacks of hereditary angioedema do not respond to epinephrine, though if there is any cause to suspect allergic rather than hereditary angioedema, then administration of epinephrine is

unlikely to cause any harm. If attacks involve the airways, then respiratory support is the first priority. Acute attacks of angioedema may be arrested by infusion of purified C1 inhibitor concentrate. If this is not available, fresh frozen plasma may be infused. This is less satisfactory, as plasma not only contains C1 inhibitor but also kallikrein, C1r, and C1s, which may generate further kinin production. In patients with repeated attacks of angioedema or infrequent but life-threatening attacks of disease, prophylactic treatment should be given. C1 inhibitor levels originating from the single normal allele increase in response to treatment with impeded androgens, such as danazol and stanozolol. This is a moderately effective treatment, although these compounds still retain some virilizing activity. An alternative, though probably less effective therapy (there are no randomized trials), is the proteinase inhibitor tranexamic acid, which may reduce the consumption of C1 inhibitor by blocking the activity of the serine proteinases that interact with C1 inhibitor.

Diseases associated with unregulated C3 activation

Factor I and Factor H deficiency

A key step in the regulation of complement activation is control of the fate of the C3 fragment, C3b. This acts as the nucleus for formation of further C3 convertase enzyme, unless it is catabolized by the serine esterase enzyme, Factor I, in conjunction with the cofactor protein, Factor H. Deficiency of either of these proteins allows the unregulated formation of C3 convertase enzyme and continuing cleavage of C3 (Fig. 2). Patients with deficiency of Factor I or H have a severe secondary deficiency of C3 and are susceptible to the pyogenic infections associated with this C3 deficiency. In addition, the enormous turnover of C3 associated with these deficiencies allows some C3 to deposit in glomerular basement membrane, which leads to the development of glomerulonephritis, which may proceed to renal failure. A number of families have been identified with inherited mesangiocapillary glomerulonephritis and some of these have mutations in just one Factor H allele, suggesting a dominant form of glomerulonephritis caused by partial Factor H deficiency. The mechanism of this form of nephritis is not understood.

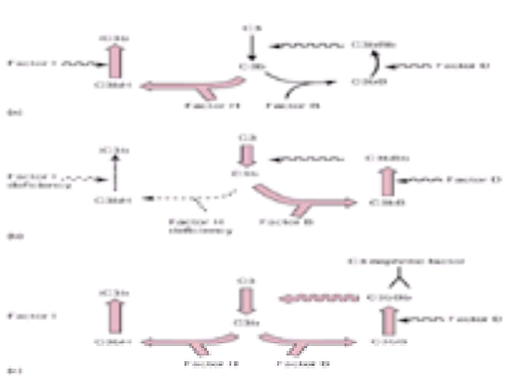

Fig. 2 Unregulated activation of C3 caused by Factor H or I deficiency, or by C3 nephritic factor. Panel A illustrates the normal control of C3 cleavage. In plasma and tissues any C3b that is formed by the normal low-grade turnover of C3 is bound by Factor H and catabolized by Factor I to inactive products. Panel B shows the effects of Factor H or Factor I deficiency. C3b can not be catabolized to inactive products. Instead there is increased formation of the alternative pathway C3bBb C3 convertase gene causing accelerated cleavage and depletion of C3. Panel C shows the effects of C3 nephritic factor. This antibody stabilizes the C3bBb C3 convertase enzyme. This results in accelerated cleavage and depletion of C3.

Acquired disorders of complement

Complement is activated *in vivo* by many stimuli, which include invading organisms, the formation of immune complexes, and tissue necrosis. When complement activation occurs on a substantial scale, this causes depletion of complement proteins, which may be measured as a reduction in their antigenic levels or as a reduction in the activity of the classical and/or alternative pathway. The measurement of complement activation may be useful in both the diagnosis and monitoring of some diseases.

In the case of sepsis associated with endotoxic shock, the large-scale systemic activation of the complement system may play an important part in the pathogenesis of this lethal condition. Activation of the classical and alternative pathways by bacterial endotoxin causes the generation of large amounts of the anaphylatoxins, C3a and C5a, and of membrane attack complex which activate neutrophils and endothelial cells causing vascular and pulmonary injury, leading to death. Diagnosis of this condition is sadly all too easy and the measurement of complement in such patients does not play an important role in assessment or management.

Tissue necrosis is also an important cause of complement activation. Therapeutic studies of experimental models of myocardial infarction, and of ischaemia–reperfusion injury in other organs, including the brain, have shown that inhibition of complement causes a significant reduction in tissue injury and final infarct size.

The diseases associated with acquired complement activation may be divided into two categories. The first category is the diseases associated with abnormal regulation of complement, which is most commonly caused by certain autoantibodies to complement. Paroxysmal nocturnal haemoglobinuria is a further example of an acquired disorder of regulation of the complement system. The second category is the diseases in which infection or autoimmunity cause clinically important activation of the complement system.

Diseases associated with abnormal complement regulation

We shall consider four diseases caused by acquired abnormalities of the regulation of complement. The first three of these are associated with the development of high-affinity autoantibodies to complement proteins, known as C3 nephritic factor, anti-C1q antibodies, and anti-C1 inhibitor antibodies. The fourth disease is paroxysmal nocturnal haemoglobinuria, in which a clone of haematopoietic cells loses expression of a family of cell surface molecules including regulatory proteins of the complement system.

Autoantibodies to complement proteins

C3 nephritic factor

C3 nephritic factor is an autoantibody that binds to and stabilizes the C3bBb C3 convertase enzyme. This increased stability of the C3 convertase enzyme disrupts the normal regulation of C3 activation and leads to chronic consumption of C3 (Fig. 2). Patients with C3 nephritic factor have very low C3 levels in serum, accompanied by normal C4 levels. When serum from a patient with C3 nephritic factor is mixed with normal serum, the C3 in the normal serum is activated and converted to C3b, which forms the basis of an assay for C3 nephritic factor.

The presence of C3 nephritic factor is associated with three clinical manifestations. The first of these is partial lipodystrophy, in which there is disfiguring loss of fat from the face and upper part of the body. Adipocytes, or fat cells, produce several complement proteins including C3 and factor D, which was independently discovered in fat cells and named adipsin. It is thought that C3 nephritic factor stabilizes the assembly of a C3 convertase enzyme on adipocytes causing the activation of complement on these cells leading to their destruction. The second clinical feature is of mesangiocapillary glomerulonephritis type II, in which electron-dense deposits of unknown composition, associated with C3 deposited at the margins of the electron-dense deposits, are found in the glomerular basement membrane. This form of nephritis may be severe, leading to renal failure. The third clinical feature is of recurrent infections, caused by the severe acquired deficiency of C3 associated with this condition.

The conventional approach to treatment of patients with type II mesangiocapillary glomerulonephritis is the use of corticosteroids, often in combination with immunosuppressive agents such as azathioprine or cyclophosphamide.

Anti-C1q antibodies

These autoantibodies react with a neoepitope, exposed in the collagenous region of C1q, which has dissociated from the other proteins of the C1 complex, C1r and C1s. Up to a third of patients with SLE develop anti-C1q autoantibodies. These are associated with activation of the classical pathway, causing very low C4 levels and, to a lesser extent, reduced C3 levels. The presence of anti-C1q autoantibodies is a marker for severe SLE, especially for the presence of lupus nephritis.

Anti-C1q antibodies are also found as the sole autoantibody in the uncommon disease hypocomplementaemic urticarial vasculitis (HUVS). In this condition, very high titres of anti-C1q antibodies are typically found, which may sometimes cause precipitation of C1q *in vitro*, hence they are known as C1q precipitins. The main clinical feature of HUVS is chronic urticaria, which is found on biopsy to be associated with a cutaneous vasculitis. Other clinical features of HUVS include glomerulonephritis, neuropathy, and chronic obstructive bronchitis. There is a considerable degree of overlap between the clinical manifestations of HUVS and SLE. This is analogous to the relationship between SLE and the primary antiphospholipid syndrome, discussed in Chapter 18.11.2.

Anti-C1 inhibitor antibodies

The third disease associated with an autoantibody to a complement protein is angioedema associated with autoantibodies to C1 inhibitor. The symptoms and signs of this are very similar to the disease of hereditary angioedema, though typically occur with a late onset. Measurements of complement proteins in the serum from patients with this disease show a similar abnormal profile to that seen in the blood of patients with hereditary angioedema, with low C1 inhibitor and low C4 levels. Additional abnormalities are low C1q levels and the presence of autoantibodies to C1 inhibitor. This serious condition is frequently associated with the presence of a B cell lymphoma.

Paroxysmal nocturnal haemoglobinuria

Paroxysmal nocturnal haemoglobinuria (PNH) illustrates the role of membrane-bound complement regulatory proteins in protection against the activation of complement on normal cells. Haemolysis in this disease is caused by the loss of expression of a membrane protein named CD59. This prevents the formation and assembly of the membrane attack complex of complement in cell membranes and thereby inhibits the lysis by complement of autologous cells (see Chapter 22.3.12).

Complement in autoimmune disease

The measurement of complement is a useful diagnostic tool as part of the assessment of patients with vasculitis and glomerulonephritis (Table 3). Some of the causes of these conditions are associated with systemic activation of the complement system on a sufficient scale that plasma levels of classical pathway proteins and C3 are significantly reduced below normal. In these diseases, it is the formation of immune complexes, either in the circulation or *in situ* in tissues, that is responsible for the activation of the complement system.

SLE

The relationships between the complement system and SLE are complex. As we have discussed, inherited deficiency of classical pathway complement proteins causes SLE. However, the vast majority of patients with SLE do not have homozygous deficiencies of complement proteins. Indeed, in these patients, SLE is associated with large-scale activation of the classical pathway of complement. The deposition of complement proteins in tissues, associated with the presence of immune complexes, has been thought to play a role in causing inflammatory lesions in tissues in SLE. Deficiency of C1q protein is most strongly associated with the development of SLE, yet as we learnt above, approximately one-third of patients with SLE develop autoantibodies to C1q.

The explanation for these complex relationships between complement and SLE is partially understood. Studies in animal models of SLE show that the predominant manner in which immune complexes cause inflammation is by ligation of Fc receptors. Mice lacking Fc receptors were protected from glomerulonephritis caused by immune complexes, whereas mice lacking complement developed full-blown lupus glomerulonephritis. A key role of complement may be to protect against the development of tissue injury by immune complexes by promoting their clearance from tissues, rather than playing a major role in the causation of injury.

We have already discussed how C1q deficiency might cause the development of SLE. How might SLE cause the development of anti-C1q antibodies? The essential feature of SLE is the formation of autoantibodies to complexes of autoantigens, such as the spliceosome complex and chromatin. C1q binds to the cellular debris that is thought to be the source of the autoantigens that drive the autoimmune response in SLE. As part of the debris, C1q may become antigenic and evoke an autoimmune response. This is a situation analogous to the association in SLE, and the primary antiphospholipid syndrome, of the presence of anticardiolipin autoantibodies with anti-b2 glycoprotein I antibodies. b2-glycoprotein I is a plasma protein that binds to negatively charged phospholipids that are exposed on the cell membranes of apoptotic cells and may thereby become part of the cellular debris that drives the autoimmune response in SLE.

The measurement of complement and of anti-C1q antibodies in SLE is of clinical value in both the diagnosis and management of patients. Serum from patients with active disease typically shows evidence of classical pathway activation with reduced C4 and, to a lesser extent, reduced C3 levels. In patients with persistently very low C4 levels, there is a high likelihood that anti-C1q antibodies will be present and such patients are more likely to have, or to develop, glomerulonephritis. Patients with persistent, severe hypocomplementaemia are at increased risk of pyogenic infections and there are strong arguments for the use of prophylactic penicillin in such patients.

Haemolytic anaemia

There is sometimes sufficient systemic complement activation associated with the haemolytic anaemias caused by autoantibodies to erythrocyte surface antigens to cause reduction in the levels of complement proteins measured in serum. This is most prominent in cold agglutinin disease (see Section 22) in which IgM cold agglutinins, which bind to I antigen, cause the deposition of many thousands of C4 and C3 molecules per erythrocyte.

The accelerated clearance of erythrocytes in autoimmune haemolytic anaemias is mainly caused by ligation of cells bearing Fc receptors in the spleen, in the case of IgG autoantibodies. In the case of cold agglutinin disease, mediated by an IgM autoantibody which cannot bind to Fc receptors, there is typically low-grade intravascular haemolysis by complement.

Human red cells are well protected from complement-mediated lysis by complement regulatory proteins expressed on their cell membranes. As we have learnt, the activity of these proteins is illustrated dramatically by the disease PNH. Rarely, if there is extensive complement fixation, as in the case of a transfusion reaction caused by an ABO mismatch, then intravascular complement-mediated lysis of red cells may cause severe injury.

Infectious disease

We have already discussed the role of complement in the innate immune system and as an effector arm of humoral adaptive immunity by illustration of the infections that accompany hereditary or acquired complement deficiency. Complement is also involved in the pathogenesis of infections in other ways. For example several viral pathogens use the complement system in a subversive manner as part of their pathogenesis (Table 4). Several infections cause hypocomplementaemia through systemic activation of complement in a similar fashion to autoimmune disease and we shall consider some examples.

Complement activation is a feature of chronic bacterial sepsis, for example in subacute bacterial endocarditis or ventriculoatrial shunt infection. In both of these conditions there is chronic release of bacterial antigens in the presence of an antibacterial antibody response that cannot eliminate the infection because of its relative inaccessibility to the immune system. This causes the chronic production of immune complexes with complement activation by the classical pathway, associated with low C4 and C2 levels and glomerulonephritis and small vessel vasculitis.

Chronic viral infection by hepatitis C is a further important cause of acquired hypocomplementaemia. This infection stimulates the production of large amounts of rheumatoid factor, which in some patients may lead to cryoglobulin production, causing complement consumption and vasculitis. In one survey in Japan of hypocomplementaemia in blood donations, infection with hepatitis C was the major cause.

Another example of hypocomplementaemia associated with infection is the complement activation associated with poststreptococcal glomerulonephritis. In this disease, which is thought to be due to an immune response to a pathogen cross-reacting with host tissues, there is marked complement activation, which includes

activation of the alternative pathway, associated with low C3 levels.

Measurement of complement in clinical practice

Throughout this chapter examples have been given of diseases which are associated with abnormal levels of complement proteins in the blood. Complement levels and activity can be assayed in clinical practice. It is useful to consider the value of measuring complement proteins in two categories, firstly in diagnosis of disease and, secondly, measurement repeatedly to monitor the activity of particular diseases.

When to measure complement

Complement in the diagnosis of disease

There are four groups of diseases in which it is important to be able to measure complement activity in serum. The first is the immunodeficiencies—it is essential to measure complement in patients with recurrent pyogenic infections, particularly in the context of recurrent or familial meningococcal disease. In this group of diseases, simple antigenic measurement of C4 and C3 levels is insufficient—it is necessary to use tests that assay the activity of the whole complement system, preferably haemolytic assays of the classical and alternative pathways. If absent or severely reduced activity is detected, then the sample should be referred to a specialist laboratory to try to identify the precise nature of the deficient component. Treatment should comprise counselling, prophylactic penicillin, and vaccination against meningococci.

The second group of diseases is vasculitis and glomerulonephritis. We saw in Table 3 how a very useful diagnostic subdivision of these diseases can be made on the basis of whether or not there is evidence of systemic complement activation. It is in these diseases that it can also be helpful to use assays of complement to monitor disease activity.

The third group are the chronic infections, which may masquerade as primary systemic vasculitis and, in this context, there should be a high index of suspicion for the presence of bacterial endocarditis or hepatitis C.

The fourth group of diseases are those specifically associated with abnormalities of the complement system, including hereditary and acquired angioedema, the familial glomerulonephritis associated with factor H deficiency, and the syndrome of partial lipodystrophy with or without mesangiocapillary glomerulonephritis.

Complement in the monitoring of disease

There are very few diseases in which the repeated monitoring of complement levels is useful. In SLE, no single test acts as a reliable surrogate for the measurement of disease activity. However, there are some patients in whom fluctuation in complement levels correlates with the waxing and waning of disease activity and, in these individuals, it is useful to monitor regularly C4 and C3 levels. It can also be useful to measure complement levels regularly in patients with autoantibodies to complement proteins; in these individuals the complement levels are a useful surrogate marker for the continuing presence of the autoantibody.

How to measure complement

Complement can be measured in a several ways. The simplest is antigenic measurement of the concentration of individual proteins, and measurement of the levels of C3 and C4 are the most widespread assays in clinical use. The results of such assays need to be interpreted cautiously. The ranges of normality are wide, because there is substantial genetic variation in the levels of these proteins. Furthermore, protein levels are a product of both synthetic and catabolic rates and both of these may vary in health and disease. Both C3 and C4 are acute phase reactants and concentrations of these proteins may rise, in the case of C3 by as much as 0.5 g/l in response to acute phase stimuli.

Measurement of C4 and C3 levels act as very crude surrogate markers of classical and alternative pathway activation respectively. However, further measurements are needed if there is any suspicion of the possibility of inherited complement deficiency or of an abnormality elsewhere in the complement system. Functional assays of complement are fairly straightforward and have the advantage that they measure the activity of all of the proteins in the complement system between activation and the end point, which is the lysis of target erythrocytes. The classical pathway is usually measured by assessment of the lysis by serum of sheep erythrocytes coated with a rabbit polyclonal antisheep erythrocyte antibody. The alternative pathway is measured by assay of the lysis of guinea pig erythrocytes, which directly activate the alternative pathway of complement in the absence of antibody, in the presence of a buffer that prevents classical pathway complement activity. Results of these assays are normally expressed as CH50 or AP50 units, which are measurements of the haemolysis of 50 per cent of respective erythrocyte preparations.

Other approaches have been devised to assess the presence of complement activation *in vivo*. Many assays have been developed which identify the products of activation of the complement system. Although these assays are attractive in principle, the products of complement activation are only present in plasma very transiently and, in practice, assays of total C4 and C3 levels, together with measurement of CH50, have not been supplanted as the best 'rough and ready' estimates of complement activation in routine clinical practice.

Further reading

Janeway C Jr, Travers P, Walport MJ, Shlomchik MJ (2001). *Immunobiology: the immune system in health and disease*, 5th edn. Garland Publishing, New York.

Liszewski MK, Farries TC, Lublin DM, Rooney IA, Atkinson JP (1996). Control of the complement system. *Advances in Immunology* **61**, 201–83.

Moffitt MC, Frank MM (1994). Complement resistance in microbes. *Springer Seminars in Immunopathology* **15**, 327–44.

Morgan BP, Walport MJ (1991). Complement deficiency and disease. *Immunology Today* **12**, 301–6.

Pickering M, Botto M, Taylor P, Lachmann PJ, Walport MJ (2000). Systemic lupus erythematosus, complement deficiency and apoptosis. *Advances in Immunology* **76**, 227–324.

Ross GD, ed (1986). *Immunobiology of the Complement System*. Academic Press, Orlando.

Turner MW (1996). The lectin pathway of complement activation. *Research in Immunology* **147**, 110–15.

Volanakis JE, Frank MM, eds (1998). *The human complement system in health and disease.* Marcel Dekker, New York.

Walport MJ (2001). Complement. *New England Journal of Medicine*, **344**,1058–66, 1140–4.

Williams DG (1997). C3 nephritic factor and mesangiocapillary glomerulonephritis. *Pediatric Nephrology* **11**, 96–8.

5.5 Innate immune system

D. T. Fearon and M. Allison

Innate immunity is an ancient system of host defence found in invertebrates as well as vertebrates. There are similar mechanisms and parallels between aspects of the innate immune systems of the fruit fly, *Drosophila*, and man. The innate immune system can therefore be considered to have two general functions; first, as a mechanism of primary host defence and, secondly, as a means of priming the adaptive immune system. Innate immunity performs the first of these functions in both invertebrates and vertebrates, having evolved many systems that facilitate the recognition and elimination of pathogens. The second role of the innate immune system in vertebrates is the instruction of the adaptive immune system. The enhanced specificity of the adaptive response is guided by the uptake recognition mechanisms of the innate immune system with the consequent generation of immunological memory for components of infectious organisms.

A key concept in understanding why the innate immune system has evolved the way it has is that an essential requirement is fulfilled—recognition of the presence of foreign organisms and, in effect, the differentiation of infectious from non-infectious. During the evolution of the innate immune system certain properties of micro-organisms that differentiate them from the host have been identified and diverse ways of detecting pathogens have gradually developed. These recognition systems have evolved in association with a range of effector mechanisms, some innate themselves and others induced through the instigation of an antigen-specific, adaptive immune response. The innate recognition systems are described in this context.

Recognition systems, soluble and membrane-bound

The innate immune system has evolved several mechanisms for detecting the presence of 'infectious' substances as such, through the development of molecular species that recognize the lowest common structural denominators of invading micro-organisms. These receptors are both soluble proteins in blood and extracellular fluid and also membrane-bound species.

Complement

The complement system is an ancient system of host defence, found in invertebrates (e.g. sea-urchin) as well as mammals. The initial methods of activation and the specific members of the early component pathways differ between the classical pathway, the lectin pathway, and the alternative pathway. All, however, use as one of their effector arms the formation of the membrane attack complex, which forms a pore in the cell membrane, as the method of killing cells. The recognition reaction central to the tagging of surfaces of micro-organisms as infectious is the covalent attachment of activated C3 component to the microbial substance. This facilitates the elimination of the pathogen through recognition of the C3-coated surface by receptors which then phagocytose and kill the organism directly or indirectly, by formation *in situ* of the membrane attack complex and subsequent cytolysis (Fig. 1).

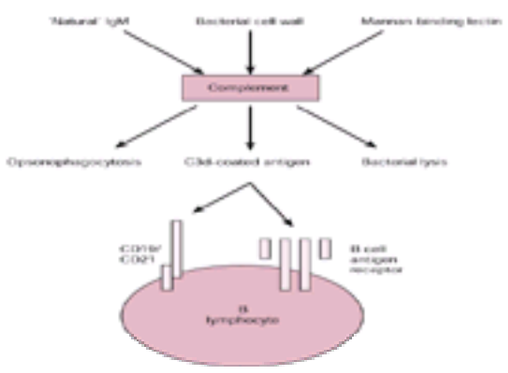

Fig. 1 The role of complement in the immune response.

The complement system is essential for protective immunity against certain bacterial infections with polysaccharide antigens to which natural IgM reacts. In models of acute peritonitis, activation of the classical pathway of complement is required for the generation of an inflammatory response and resolution of the infection. The ability of complement-coated antigen to augment the antigen-specific response has also been proven through co-ordinated ligation of complement receptor 2 on B cells with the surface immunoglobulin within the B-cell receptor complex.

The complement system is thus essential for protective immunity to polysaccharide antigens against which natural IgM is reactive, in the generation of an effective inflammatory response in models of acute septic peritonitis, and in facilitating the production of an antigen-specific antibody response.

Soluble recognition molecules

In addition to the complement proteins, there are other secreted proteins that have a significant role in the systemic or local immune response (Table 1). In respect of local defence systems, there is increasing recognition that antimicrobial peptides may have a important role, particularly at mucosal surfaces. These include the defensins, which are polypeptides of 29 to 40 amino acids that exert microbicidal activity, most commonly by disruption of bacterial membranes. These proteins are able to augment the immune response by the activity of the human b-defensins 1 and 2, which are chemotactic for immature dendritic cells and memory T cells.

Mannose-binding protein recognizes carbohydrate moieties and has selectivity for foreign, as opposed to self, cell surfaces or glycoproteins as a result of the spacing of its terminal carbohydrate-recognition domains. Mannose-binding protein exists as an oligomer in the circulation and therefore achieves high-affinity interaction with multivalent carbohydrate molecular species, such as those found on microbial surfaces. Several different mutant alleles of mannose-binding protein have been described and individuals who are homozygous or heterozygous for the mutant alleles seem to be more susceptible to a certain infections, and exhibit a defect in opsonophagocytosis. There is evidence that other proteins, such as C1q, may act similarly since mannose-binding protein is recognized by a C1q receptor recently

cloned from monocytes.

Mannose-binding protein is a member of the collection family of soluble proteins that also includes the surfactant proteins. These proteins possess collagenase domains and carbohydrate recognition domains. The latter mediate recognition of structures on the outer surface of micro-organisms. *In vitro*, surfactant protein-A and D bind carbohydrates on the surface of viruses, bacteria, and fungi and thereby may lead to the aggregation of the micro-organisms, thus encouraging phagocytosis by neutrophils and by alveolar macrophages. There is evidence that the surfactant proteins have a role in pulmonary defence; they bind to components of *Pneumocystis carinii* as well as *Cryptococcus neoformans*, which are important pathogens in immunocompromised individuals. Furthermore, patients with cystic fibrosis, who are susceptible to repeated pulmonary infections, have diminished levels of surfactant protein-A and D; the deficiency correlates with reduced effective killing of pathogen.

A further family of soluble proteins that have an emerging role in the innate immune response are the pentraxins. The members of this family, which includes C-reactive protein and serum amyloid protein, exist as a radial pentameric structure in solution and recognize a variety of ligands such as phosphate esters and polysaccharides. Their production is induced by inflammatory cytokines, such as IL-1 and IL-6, and, as a consequence of being multimeric, they are able to activate the classical pathway of complement by directly binding C1q. In animal models, C-reactive protein has been found to have a role in protection against endotoxin-induced mortality. Recently, mice deficient in C-reactive protein have been found to develop autoantibodies, suggesting that C-reactive protein may also participate in the maintenance of immunological tolerance to self proteins, possibly by mediating the clearance of apoptotic cells.

Membrane-bound recognition receptors

In addition to soluble recognition systems, there are several cell-associated receptors whose cellular distribution reflects the effector arms of the immune system to which they are linked (Table 2). There are two broad types of receptor; those that have primarily an endocytic capacity and those that have signalling capacity.

The mannose receptor is an example of an endocytic recognition system and is expressed on tissue macrophages, immature dendritic cells, and some endothelial cells. This receptor recognizes saccharide residues commonly expressed on microbial surfaces but not those frequently found exposed on self glycoproteins. Hence this receptor is able to differentiate infectious from non-infectious and binds a large number of different organisms including *Mycobacteria*, *Trypanosoma*, yeast, and both Gram-positive and Gram-negative bacteria. Significantly, this receptor, which constitutively recycles to the cell surface through endosomal compartments, is able to target bound antigen for presentation on MHC class II molecules, and, in the case of a component of mycobacteria, lipoarabinomannan, on the non-classical MHC class I molecule, CD1b**.**

Within the same lectin family, but containing carbohydrate-recognition domains with a different structure, is DEC-205. Initially used as marker of dendritic cells but now known to be expressed on a subset of dendritic cells, DEC-205 may also be able to enhance presentation of antigen in the context of MHC class II molecules.

A further family of cell-surface receptors that is increasingly being recognized as having an important role in recognition and clearance of invading pathogens are the scavenger receptors. This group of recognition molecules is characterized by their broad ligand specificity; it was first identified by Brown and Goldstein in 1979 as responsible for binding modified, but not native, low-density lipoprotein. As well as binding acetylated and oxidized low-density lipoprotein, members of the scavenger receptor family also can recognize Gram-positive and Gram-negative bacteria, lipotechoic acid and lipopolysaccharide, and aldehyde-modified proteins. There are considered to be five classes of scavenger receptors, differentiated by structure and binding characteristics as well as the cells on which they are expressed. The class A scavenger receptors, expressed on macrophages, have a role in clearance of micro-organisms from the circulation. Mice deficient in two allelic forms of the class A scavenger receptors have increased susceptibility to herpes simplex virus and *Listeria*, as well as to the lethal effects of endotoxin.

A further exciting aspect of the biology of scavenger receptors is that they recognize apoptotic cells. This is achieved, at least in part, through their ability to bind phosphatidylserine, which becomes exposed on the surface of apoptotic cells during the process of cell death—a mechanism which may represent a means of clearing apoptotic cells in a way that avoids an inflammatory response.

Recognition of cell-surface components of pathogens is an important warning mechanism, indicating to the host the presence of foreign micro-organisms. It has long been recognized that a major constituent of the walls of Gram-negative bacteria, polysaccharide, is a potent inducer of an inflammatory response; lipopolysaccharide can be lethal through the induction of shock. The recognition of lipopolysaccharide, and hence the generation of an inflammatory cytokines in response to this stimulus, occurs through the CD14 surface receptor which binds lipopolysaccharide once lipopolysaccharide is itself bound by lipopolysaccharide-binding protein, a constituent of serum. The means of signalling upon ligation of lipopolysaccharide–lipopolysaccharide-binding protein by CD14 was, until recently, unclear, because CD14 is a glycophosphatidylinositol-linked membrane protein, and therefore would have no intrinsic signalling capacity. However, in 1997, Medzhitov and Janeway demonstrated the presence in humans of a protein homologous to the *Drosophila* protein Toll; they further showed that this facilitated the production of inflammatory cytokines by monocytes. Subsequently, four further members of the same family have been discovered and mutations in the gene encoding one of these toll-like receptors (Tlr4) have been found in two different strains of mouse that are lipopolysaccharide-non-responsive. Different members of the Tlr family can differentiate between various classes of micro-organisms—Tlr4 responds to Gram-negative organisms while Tlr2 is activated by Gram-positive pathogens and yeasts. Downstream consequences of ligation of CD14 and Tlr4 by lipopolysaccharide–lipopolysaccharide-binding protein, which include generation of inflammatory cytokines, are mediated through activation of the transcription factor, NF_B.

Cellular components of the innate immune system

Just as B and T lymphocytes are clearly constituents of the acquired response, macrophages, dendritic cells, and polymorphonuclear cells must be considered as key members of the innate response. Recently, greater attention has been given to the roles of these cells and the way they orchestrate an immune response, not just in their own right but also in co-ordinating the involvement of different arms of the adaptive immune system.

Macrophages

The monocyte/macrophage is beautifully adapted for its role of ingestion and intracellular killing of pathogens. It has several receptors that are able to recognize organisms or components of organisms and it is extremely active phagocytically—thus it can internalize bound micro-organisms and kill them after formation of phagosomes. In addition to the well-established production and activity of reactive oxygen species macrophages also produce reactive nitrogen intermediates (RNIs), including nitric oxide, by inducible nitric oxide synthase (iNOS or NOS2). Studies in mice deficient in NOS2 have demonstrated that the production of reactive nitrogen intermediates plays a non-redundant role in protection against viruses and intracellular pathogens. There is also synergy between nitric oxide formation and reactive oxygen intermediates, through the generation of peroxynitrite ($ONOO^-$) which possesses additional microbicidal activity.

Dendritic cells

These cells are also bone-marrow derived (with the exception of a small subtype of dendritic cells). During maturation they pass through two functionally very different stages (Fig. 2). As immature dendritic cells, they populate various non-lymphoid tissues (e.g. Langerhans cells in the skin are immature dendritic cells) and actively sample their microenvironment through receptors, including the mannose receptor, DEC-205, and Fc receptors. Immature dendritic cells are thus well equipped to take up a large variety of antigens, and once activated by bacterial constituents, such as lipopolysaccharide and the CpG motifs within bacterial DNA, their surface phenotype changes and they are induced to migrate out of the tissues to draining lymphoid tissue. By the time they reach the draining lymph nodes they acquire a fully mature phenotype; such antigen-capturing receptors as the mannose receptor and Fc receptor are down-regulated, and surface expression of MHC class II, as well as costimulatory molecules, CD80 and CD86, and, vitally, CD40, are up-regulated. At the same time, there is increased activity of the antigen-processing machinery within these cells, rendering them powerful 'professional' antigen-presenting cells.

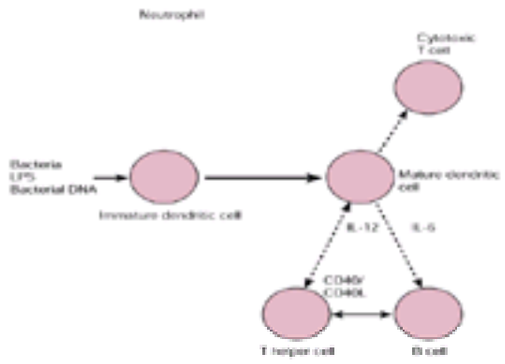

Fig. 2 The central role of the dendritic cell in the innate and acquired immune responses.

The efficiency of dendritic cells is utilized by the immune system in several settings, and it is increasingly clear that they participate critically in the development of immune responses to certain tumours and viruses, both of which are mediated through CD8+ T cells. For some viral infections, the production of virus-specific CD8+ T cells requires the presence of antigen-specific CD4+ T cells. In this setting, activated dendritic cells provide the link, as these cells are able to stimulate CD4+ T cells, that then reciprocate by upregulating CD40 ligand on their surface. CD40 ligand then binds to CD40 expressed on the dendritic cells, giving a signal to these cells that enable them to activate antigen-specific CD+ T cells. Dendritic cells are absolutely required for initiation of antigen-specific CD8+ T cell responses to certain viral infections and tumours through a process called 'cross priming'. Antigens derived from one cell are captured by a second cell and this cell processes and presents antigenic determinants from the first cell in the context of surface MHC class I to CD8+ T cells. These CD8+ T cells are activated and expand clonally; they are then able to exhibit cytolyic activity for the first cell type, be it a tumour cell expressing the antigen or a virally-infected cell.

Neutrophils

Polymorphonuclear phagocytes constitute one of the principal arms of the innate immune system. As one of the first immune cells to arrive at the site of an infectious challenge, their response is important not only in the initial control of infection, but also in guiding the migration of other components to the local environment. Neutrophils, like macrophages, are phagocytically active and express receptors for bacterial peptides, complement components, and the constant regions of specific immunoglobulin isotypes. Hence, specific phagocytosis of either immunoglobulin- or complement-opsonized bacteria can occur. Signalling though some of these receptors induces activation of the neutrophil, with immediate release of reactive oxygen species and phospholipid-derived inflammatory mediators which exhibit microbicidal activity. The importance of this function in neutrophils and macrophages is reflected in the phenotype of individuals with chronic granulomatous disease who have a defect in the generation of the reactive oxygen species and who suffer from recurrent, non-resolving infections that are usually caused by organisms that express catalase.

Natural killer cells

Natural killer cells, although clearly related to T and B lymphocytes, are distinct in not expressing rearranged receptors for antigen or peptides fragments of antigen. First discovered as responsible for tumour cytolytic activity in mice devoid of lymphocytes, these cells can develop both from the bone marrow and from fetal thymocytes.

Natural killer cells act as a means of non-adaptive cellular cytotoxicity against tumour cells and some virally infected cells. Although these cells do not have antigen-specific receptors, a surface molecule, MICA, up-regulated on tumour cells and stressed cells (e.g. when virally infected) is recognized by a natural killer cell receptor, with consequent activation of the cell-killing machinery. Protection against lysis of healthy host cells is provided through the recognition by receptors on natural killer cells of self MHC class I molecules. Ligation of these receptors negatively regulates the natural-killer-cell-mediated lysis. Natural killer cell activation can be induced through ligation of CD16, the low affinity Fc receptor for IgG, on the surface of the cell leading to antibody-dependent cellular cytotoxicity. Inflammatory cytokines, including IL-15, TNF-a, IL-12, and IL-18, also trigger natural killer cell activation, with the concomitant production of interferon-gamma (IFN-g).

Natural killer cells play an important and individual role in antiviral responses. For example adenovirus- and human cytomegalovirus-infected cells secrete proteins that inhibit surface MHC class I expression, which protect against antiviral cytotoxic T lymphocyte (CTL) activity, but these proteins also render the cells more susceptible to natural-killer-cell-mediated killing. A further protein, also produced by cytomegalovirus, however, is homologous to class I heavy chain and is able to bind to b_2-microglobulin. This protein facilitates transport of the complex to the cell surface and confers on the cell resistance to natural-killer-cell-mediated lysis.

Mast cells

Mast cells are most widely recognized for their role in allergic responses. They are bone marrow-derived cells whose proliferation and maturation are stimulated by a number of cytokines including, in particular, IL-3 and stem cell factor. The expression of FceRI, the high affinity Fc receptor for IgE, is characteristic of mast cells. Once multiple molecules of IgE have bound to an allergen, ligation and cross-linking of this receptor can occur, which causes mast cell degranulation. The degranulating mast cell releases histamine, neutral proteases (for example tryptase), proteoglycans, and lipid mediators with diverse effects, including vasodilatation and increased vascular permeability, bronchoconstriction, and intestinal smooth muscle constriction.

Mast cells also aid the immune response to micro-organisms. They can phagocytose and, *in vitro* at least, present antigen in the context of MHC class I and class II molecules—although the functional importance of this in the host has not been demonstrated. These cells do, however, have a prominent role in the development of a protective response in animal models of acute septic peritonitis, in which they promote the clearance of organisms, in part by the production of TNF-a and the subsequent recruitment of neutrophils. In this setting, it is thought that mast cells are activated by complement products, and the generation of C5a anaphylatoxin induces mast cell degranulation as a result.

g† T cells

These cells show features of both the innate and the adaptive immune systems. They have a limited ability to rearrange their T cell receptor genes, and *in vivo* are composed of prominent clones of g† receptor combinations that respond to alkylamine components of micro-organisms.

The cytokines of the innate immune system

Interferons

The interferons (IFNs) were originally considered to have mainly antiviral activity. There are two groups—the type I interferons, IFN-a and IFN-b, and the type II, IFN-g. While the type I IFNs are produced by virally-infected cells, IFN-g is secreted by T lymphocytes and natural killer cells. The IFNs exert their antiviral activity through several mechanisms, that may be specific to the infected cell or involve the induction of broader host responses. The two groups of IFN inhibit cellular protein synthesis and induce activation of cellular RNAses, thereby destroying viral double-stranded RNA.

The broader actions of the IFNs affect innate as well as adaptive responses in host immunity. IFN-g is produced by natural killer cells, once these are activated by IL-12. The IFN-g itself activates the natural killer cell cytolytic machinery and also simulates macrophage microbicidal activity by inducing transcription of the *NOS2* gene and activating NADPH oxidase. Both classes of IFN, but particularly IFN-g, can encourage the generation of the adaptive immune response, by inducing transcription of several genes that encode proteins involved in antigen processing and presentation. In these ways, interferons enhance humoral immune responses as well as antiviral T cell responses.

The inflammatory cytokines—TNF, IL-1, and IL-6

Tumour necrosis factor-alpha (TNF-a) plays a central role in the host response to bacterial infection, and while its production is essential for protection, excess TNF-a can also be lethal to the host. TNF-a is produced by macrophages when stimulated by bacterial products (including lipopolysaccharide) from Gram-negative bacteria, but can also derive from many other cell types from diverse tissues in response to inflammatory stimuli. TNF-a propagates the inflammatory response; it induces the microbicidal activity of macrophages, stimulates production of macrophage IL-1, IL-6, IL-8, as well as TNF-a, and augments the cytotoxic activity of natural killer cells. An additional feature of TNF-a is that it can induce the maturation of immature dendritic cells, that then leave the local environment where they have encountered micro-organisms and migrate to local lymphoid tissue to initiate an adaptive immune response. TNF-a can be considered as a double-edged sword because it also plays a role in the syndrome of septic shock, characterized by hypotension, capillary leak, and multiorgan failure.

IL-1 has many proinflammatory activities, most of which overlap with TNF-a. As well as influencing almost all cells of the immune system, it exerts activity in neuronal tissue, the liver, adipose tissue, and the endothelium. IL-1 results in the increased expression of adhesion molecules, accumulation of neutrophils at a site of inflammation, the hepatic acute phase response—and is involved, with IL-6 and $PGE_{2,}$ in the generation of fever. Clearly, regulation of the activity of IL-1 is important; of the two types of IL-1 receptor, one (the type II IL-1 receptor) acts merely as an inactive ligand or decoy for IL-1, thereby preventing it binding to the type I IL-1 receptor. Furthermore, there is also an inactive analogue of IL-1 called IL-1 receptor antagonist (IL-1Ra) that binds to the type I IL-1 receptor but does not induce an activating signal.

IL-6 is produced by a large number of cell types, and its expression is increased in almost all tissues in response to infection. Inflammatory stimuli such as lipopolysaccharide, TNF-a, and IL-1 are responsible for the induction of IL-6 in infections. IL-6 is partly responsible for the hepatic acute phase response exemplified by enhanced transcription of the pentraxin C-reactive protein discussed earlier. IL-6 also induces B-lymphocyte differentiation and can induce activation of T cells.

IL-12 and IL-18

Both macrophages and dendritic cells possess further means by which they can stimulate their antimicrobial activity and that of other cells. IL-12 is a cytokine which is produced by macrophages and dendritic cells, either upon direct stimulation by certain microbial products, for example lipopolysaccharide or CpG motif-bearing DNA, or more commonly by stimulation of these antigen-presenting cells by CD4+ T cells, themselves activated by the antigen-presenting cells. The cognate interaction-mediating IL-12 generation by these cells is mediated by CD40 on the antigen-presenting cells, the stimulation of IL-12 production having a number of effects. IL-12 stimulates IFN-g secretion by natural killer cells, which can then both stimulate CD8+ T cells and Th1-type CD4+ T cells. IL-12 also directly activates the microbicidal activity of macrophages by inducing the transcription of NOS2, with the consequent generation of reactive nitrogen intermediates as discussed earlier.

Another cytokine whose full role in the immune response to micro-organisms is still being elucidated, is IL-18. Initially described as a product of activated Kupffer cells and called IFN-g-inducing factor (IGIF), IL-18 is now known to be produced by macrophages. IL-18 is synergistic with IL-12 in the induction of IFN-g and, independently of IL-12, it also enhances natural killer cell activity.

Conclusions

The innate immune system, through various receptor species, is able to detect motifs common to pathogen subtypes, thus keeping the requirement for such receptors to a minimum. The receptors are linked to diverse effector mechanisms that facilitate inactivation or death of the micro-organism. An emerging pattern is that while the adaptive immune system recognizes protein structures, it is carbohydrate moieties that are the determinants for innate immune ligands.

Within the innate immune system, new components continue to be uncovered, and, at the same time, additional functions are being ascribed to previously characterized molecules. Thus, the field of innate immunity and especially its role in directing the subsequent adaptive response is one of the most active and exciting areas in contemporary immunology.

Further reading

Aderem A, Ulevitch RJ (2000). Toll-like receptors in the induction of the innate immune response. *Nature* **406**, 782–7.

Bogdan C, Rollinghoff M, Diefenbach A (2000). Reactive oxygen and nitrogen intermediates in innate and specific immunity. *Current Opinion in Immunology* **12**, 64–76.

Carroll MC (1998). The role of complement receptors in induction and regulation of immunity. *Annual Reviews of Immunology* **16**, 545–68.

Feizi T (2000). Carbohydrate-recognition systems in innate immunity. *Immunological Reviews* **173**, 79–88.

Jack DL, Klein NJ, Turner MW (2001). Mannose-binding lectin: targeting the microbial world for complement attack and opsonophagocytosis. *Immunological Reviews* **180**, 86–99.

Travis SM, Singh PK, Welsh MJ (2001). *Current Opinion in Immunology* **13**, 89–95.

5.6 Immunodeficiency

A. D. B. Webster

Introduction

The primary immunodeficiencies have provided a valuable insight into the critical components of the immune system for protection against infection. Although 10 years ago only two of these conditions were understood at a molecular level, over 80 defective genes causing a variety of clinical phenotypes have now been identified. This section focuses on lymphocyte disorders causing susceptibility to infection; defects in phagocytes and the complement pathways, which are important components of the innate immune system, are described in Chapter 5.4.

In the United Kingdom, there is still an unacceptable delay of 5 years on average for the diagnosis of some types of immunodeficiency, emphasizing a need for clinicians to be more aware of these disorders.

Classification

The primary immunodeficiencies (**PIDs**) are mostly inherited single-gene disorders presenting in infancy and early childhood. However, the one important exception is common variable immunodeficiency (**CVID**) that is still not precisely defined, most patients having a complex polygenic disorder of immune regulation which frequently presents in adults. PID includes a wide variety of cellular disorders of both the adaptive and innate immune systems, some causing autoimmunity rather than susceptibility to infection. The International Union of Immunological Societies (**IUIS**) supports a committee to meet every 5 years to review these classifications.

Secondary immunodeficiency occurs in a wide range of diseases, the immunodeficiency often being caused or exacerbated by therapy with immunosuppressive drugs.

History and examination

The family history is important, particularly since it may suggest X-linked or autosomal inheritance. There are few characteristic physical features, but the total absence of tonsils in infants and children is a feature of X-linked agammaglobulinaemia and severe combined immunodeficiency, the latter also being associated with failure to thrive and an absent thymic shadow on a chest radiograph. Signs of chronic otitis media, sinusitis, conjunctivitis, bronchitis, and bronchiectasis are typical, and splenomegaly is common in some types. Growth retardation, dysmorphic features, and severe skin disease (e.g. eczema, erythroderma) occur individually or in combinations in some of the PID syndromes. Chronic infection with atypical mycobacteria in young children suggests a defect in the g-interferon and interleukin-12 signalling circuit. Massive lymphadenopathy and splenomegaly with autoimmune disease is a feature of lymphocyte apoptotic defects.

Primary immunodeficiency

Antibody-deficiency syndromes

Prevalence

The lifetime prevalence of the severe antibody-deficiency syndromes is about 16 per million of the Caucasian population in Western countries, but there is no reliable information on prevalence in developing countries; there are currently about 1000 diagnosed patients in the United Kingdom. However, selective IgA deficiency is common, occurring in about 1 in 700 of Caucasians, most of whom are healthy.

Aetiology

Most of the disorders are caused by single-gene defects leading to blocks at various stages of the maturation and differentiation of B lymphocytes (Fig. 1). However, common variable immunodeficiency, the most common of all the PIDs, appears to be a complex polygenic disorder of immune dysregulation.

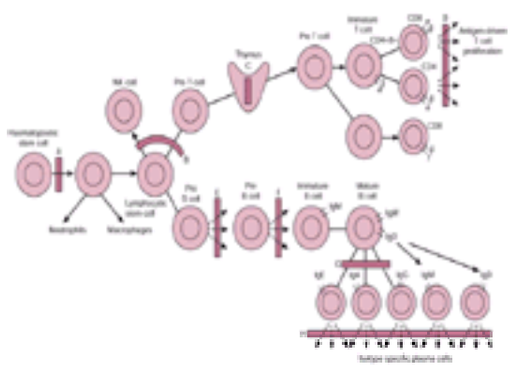

Fig. 1 Various blocks (hatched bars) in the maturation and differentiation of lymphocytes in the primary immunodeficiencies. (a) Reticular dysgenesis; B: X-linked severe combined immunodeficiency (**SCID**) (gc defect), JAK3 deficiency, adenosine deaminase (ADA) defects; C: purine nucleoside phosphorylase, *RAG1* and *-2*, *Artemis*, ZAP-70, CD45, IL-7Ra, and lymphocyte HLA class II defects; D: common variable immunodeficiency (CVID), $p56^{lck}$ defects. (b) E: defects in surface µ

heavy-chain expression, Blnk, Iga, and l5 surrogate light chain, RAG1 and -2, Artemis, ADA; F: XLA (Btk deficiency); G: X-HIM (CD40-ligand deficiency, CD40) and activation-induced cytidine deaminase deficiencies; H: CVID. Note that all defects causing SCID, other than those affecting B-cell development, will compromise antibody production through the lack of T-cell interactions in the lymphoid apparatus at stage G.

Nomenclature (Table 1)

These follow IUIS guidelines, the individual types usually being referred to as a clinical description (e.g. X-linked agammaglobulinaemia), or alternatively by the molecular defect (e.g. CD40-ligand deficiency). The common use of acronyms by immunologists can be confusing for those outside the field.

Major types of antibody deficiency

Common variable immunodeficiency (CVID)

This is the most common of all the primary immunodeficiencies, affecting about 1 in 30 000 Caucasians. It is likely that the majority of patients labelled as CVID have a consistent combination of molecular abnormalities, the remaining few having as yet unidentified single-gene defects. Patients become symptomatic at any age, but usually in early childhood or late adolescence. Serum immunoglobulin levels are variable, but typically the serum IgA is below 0.1 g/l, the IgG below 2 g/l, and the IgM below 0.2 g/l. However, some patients have near-normal IgM levels and can occasionally make IgM antibodies. At least 20 per cent of patients with CVID can be shown to have a polygenic inherited disorder which is genetically linked to selective IgA deficiency. The pedigrees of affected families show a variable phenotype, even in affected siblings, ranging from IgG subclass defects, IgA deficiency, to CVID. Mothers with IgA deficiency and circulating antibodies to IgA are more likely to have affected offspring. The major predisposing genetic locus is located in the MHC region on chromosome 6, covering part of the class III and adjacent class II region. None of the genes has yet been identified, and there are at least three minor susceptibility genetic loci on other chromosomes.

A third of patients are lymphopenic, with circulating CD4+ T-cell counts between 0.15 and 0.4 × 10^{9}/l, often with a relative increase in CD8 T cells. Splenomegaly occurs in about 30 per cent of patients, and splenectomy may be necessary for those who develop hypersplenism. The spleen usually contains numerous non-caseating granulomas, with excessive numbers of activated macrophages in the surrounding tissue. A smaller number of patients develop granulomatous disease, requiring steroid therapy, in the lungs and liver; other organs such as the skin, brain, and kidneys are less commonly involved. Scandinavian patients are much less prone to granulomatous complications, suggesting an environmental factor is involved.

Chronic or recurrent diarrhoea, not related to known pathogens, occurs in at least 20 per cent of patients. This is often associated with a mild colitis and an excess of intraepithelial T lymphocytes. A minority have a Crohn's-like condition with a florid ileitis, and occasionally strictures. A few patients have upper intestinal villous atrophy with a florid inflammatory infiltrate; a minority of these will respond to a gluten-free diet, while the others need steroids to induce remission. About 10 per cent of patients have a pan-gastritis, sometimes with anaemia due to a lack of intrinsic factor and poor vitamin B12 absorption. Submucosal lymphoid nodules are common in the small bowel, but can occur elsewhere in the gut; this nodular lymphoid hyperplasia probably represents an aborted attempt at a local immune reaction to antigens in the gut.

The mechanism of CVID is complex and not well understood, but the evidence suggests that the fundamental abnormality is a failure to generate the ppropriate microenvironment in the lymphoid apparatus for B-cell differentiation and antibody production. There is evidence of an excessive production of g-interferon and interleukin-12 by lymphocytes and monocytes, respectively; this cytokine dysregulation is likely to cause a marked skewing towards a TH1-type response and increased susceptibility to chronic inflammatory disease. The antibody deficiency has recovered after HIV (human immunodeficiency virus) infection in five reported cases, probably by altering these abnormal cytokine patterns.

The differential diagnosis of CVID depends on excluding the other rarer single-gene PIDs and secondary immunodeficiency (see below).

Thymoma and hypogammaglobulinaemia

This has some distinctive features but many clinical and laboratory similarities with CVID. The thymoma, usually benign and well encapsulated, occurs in patients over 40 years of age, the hypogammaglobulinaemia being of varying severity. There may be autoimmune phenomena such as neutropenia, haemolytic anaemia, and red-cell aplasia. The disease has a poor prognosis, with most patients dying within 15 years from opportunistic viral or fungal infections due to deteriorating cellular immunity. Surgical removal of the thymoma has no effect on the immunodeficiency or the prognosis, but is usually necessary to exclude malignancy and/or involvement of neighbouring structures. The mechanism of the association with hypogammaglobulinaemia is not understood.

X-linked (Bruton's) agammaglobulinaemia (XLA)

Affected males usually develop recurrent infections in the first 2 years of life, often at about 6 months when maternal IgG is exhausted. Most patients have some residual IgG production (less than 50 mg/100 ml), but make no IgA and IgM. T-lymphocyte function is normal, but there are very few circulating B cells due to a block in the differentiation at the pre-B-cell stage in the bone marrow. The relevant gene on the X-chromosome codes for **Btk** (Bruton's tyrosine kinase), an intracellular signalling molecule involved in pre-B-cell development. A similar phenotype occurs in males and females with rare, autosomal recessive defects in critical molecules for B-cell differentiation upstream of Btk (see Table 1). Provided serious infection can be prevented, the patients have an excellent prognosis, and suffer from none of the chronic inflammatory/granulomatous complications seen in CVID. This may be partly explained by Btk having a role in the signalling cascade for macrophage activation, causing XLA patients to have a downregulated inflammatory response.

Hyper-IgM syndrome (HIM)

There are three known rare molecular defects causing a failure of immunoglobulin class switching from IgM to IgG, and then to IgA and IgE. Two involve either the CD40 ligand (CD154), an activation-induced surface protein on CD4+ T lymphocytes, or its ligand CD40 on B cells. The CD154 gene is on the X chromosome, affecting males who have a poor prognosis due to an unexplained susceptibility to sclerosing cholangitis, cirrhosis, and liver cancer. X-HIM patients are also prone to opportunistic infections with, for example, *Pneumocystis carinii* and *Cryptosporidium parvum*, suggesting that the failure to express CD40 ligand has wider implications for T-cell immunity. Sometimes female carriers of the genetic mutation can present with mild antibody deficiency due to incomplete Lyonization.

A rarer cause of HIM is caused by deficiency of a lymphocyte-specific cytidine deaminase, an enzyme involved in RNA editing in activated B cells. This autosomal recessive condition has a milder phenotype than X-HIM, resembling CVID and not being associated with opportunistic infections and liver disease.

X-linked lymphoproliferative syndrome (XLPS)

Affected males have a defect in the control of T-lymphocyte immunity to the Epstein–Barr virus (**EBV**), either dying during acute infectious mononucleosis or developing Burkitt's-like B-cell lymphomas and/or hypogammaglobulinaemia. XLPS is caused by mutations in the gene coding for **SAP** (surface lymphocyte activation molecule (SLAM) associated protein), a cytoplasmic protein that regulates the activation of cytotoxic CD8+ lymphocytes; those dying in the acute phase have a massive multiorgan infiltration by these cells. There appears to be a defect in the control of EBV reactivation leading to lymphoma in some survivors of acute infection. The mechanism of the immunodeficiency is not known, with some patients being misdiagnosed as having CVID.

Transient hypogammaglobulinaemia in infancy and childhood

Maternal IgG crosses the placenta in the last trimester of pregnancy and helps to protect the infant against infection for the first few months of life. Between 4 and 6 months of age the normal infant will develop an increasing repertoire of IgG antibodies, mainly of the IgG1 subclass. The capacity to make IgG2 and IgA is not fully developed until adolescence. This sequence may be retarded in some infants, who present in early childhood with infections and hypogammaglobulinaemia. There is no consensus on the precise definition of this disorder except that recovery should have occurred by 5 years of age. There is evidence that the mechanism has

similarities with CVID.

Infections associated with hypogammaglobulinaemia

Patients are prone to bacterial septicaemia and respiratory infections, with a minority being also susceptible to mycoplasma and enteroviral infection (Fig. 2). There is usually uneventful recovery from most common childhood viral infections (e.g. measles, varicella, and mumps).

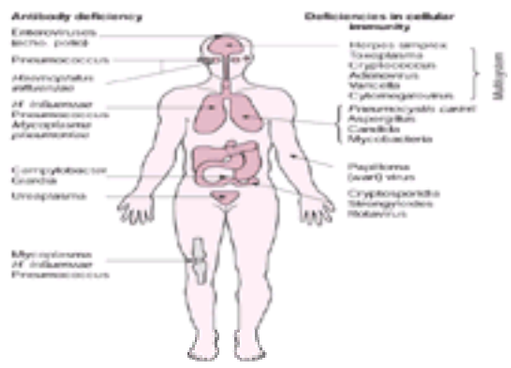

Fig. 2 Pattern of infections associated with severe defects in antibody production or cellular immunity. Patients with severe combined immunodeficiency (**SCID**) disease suffer from all the infections listed, while those with the acquired immunodeficiency syndrome (**AIDS**) are prone mainly to infections shown in the right-hand column.

Bacteria

Patients may present with pneumococcal, *Haemophilus influenzae* (capsulated), or meningococcal septicaemia, but more often there is a history of recurrent respiratory infection, the main organism involved being non-encapsulated *Haemophilus influenzae.* This semi-commensal organism colonizes the upper respiratory tract of many normal individuals, spreading to involve the bronchi following common viral infections. Patients with antibody deficiency suffer from chronic infection in the ears, sinuses, and bronchi, often leading to bronchiectasis and deafness. *Pneumococcus* spp. and *Moraxella catarrhalis* are other common respiratory pathogens in these patients. Staphylococcal skin infection is common in children.

Mycoplasmas

Antibodies inhibit the growth of mycoplasmas on mucosal surfaces. About 10 per cent of XLA or CVID patients develop chronic mycoplasma arthritis with destruction of joints. Overgrowth of these organisms on mucosal surfaces (usually in the respiratory or genitourinary tracts) apparently leads to the uptake of viable organisms by phagocytes, which then transport them to joints where the microenvironment supports growth. A variety of mycoplasma species have been implicated (e.g. *M. hominis* and *Ureaplasma urealyticum* from the urogenital tract, *M. pneumoniae* from the lungs). Rarely, infection can be acquired from animals as a zoonotic infection. Local infection can lead to chronic cystitis or urethritis, and there is the possibility that some species cause chronic bronchitis. Although methods based on the polymerase chain reaction (**PCR**) are being developed for rapid molecular diagnosis, very few laboratories can culture these organisms and provide antibiotic sensitivities. A working diagnosis of mycoplasma infection should be made in hypogammaglobulinaemic patients with arthritis, cystitis, or urethritis when samples taken for routine microbiological testing are negative; doxycycline should be given in the first instance, and specialist advice sought to confirm the diagnosis and provide antibiotic sensitivities in case the organism is resistant.

Viruses

Enteroviruses

These include polio-, coxsackie-, and echoviruses. Coxsackie- and echoviruses, of which there are many different strains, are a common cause of self-limiting mild enteritis and/or meningitis in normal individuals, but cause chronic meningoencephalitis and myositis in patients with severe hypogammaglobulinaemia. Echoviruses are usually involved, the classical features being convulsions, VIIIth nerve deafness, headache, and myositis, the last leading to fibrosis of the limb muscles with contractures. The diagnosis is usually made by culturing enteroviruses from the cerebrospinal fluid, or by a positive PCR for viral RNA. There is usually a gradual deterioration in the central nervous system features and death within 5 years; however, in a few people the disease can be modified or even cured by giving pooled immunoglobulin containing specific antibodies to the virus intravenously and into the cerebrospinal fluid on a regular basis. Standard prophylactic immunoglobulin therapy (see below) probably reduces the risk of enteroviral infection, but is not fully protective and often obscures the diagnosis by preventing culture of the virus from cerebrospinal fluid; patients in this situation may present with insidious mild symptoms of cerebral involvement such as altered personality and decreased mental ability. It is important to make the diagnosis because there is a new antienteroviral drug, pleconaril, which has cured most of treated patients.

There is a risk of paralytic poliomyelitis after live oral polio vaccination which is contraindicated in these patients. Fortunately, regular immunoglobulin therapy appears to prevent poliovirus infection from recently immunized family members, probably because enough neutralizing IgG leaks into the saliva and prevents faecal/oral transmission. However, rarely, a patient may become a chronic excreter of a virulent polio (vaccine related) strain. The World Health Organization, who are planning to discontinue routine polio immunization, are concerned about such patients who could start a new polio epidemic if immunity waned in the general population.

Other viruses

Patients with CVID are prone to recurrent *Varicella zoster* skin infection (shingles) but this rarely recurs after treatment with immunoglobulin. Reactivation of *Herpes simplex* (cold sores) or vaginal herpes is uncommon. The role of persistent picornavirus (e.g. rhinovirus, enterovirus) infection in the respiratory tract is unclear, but these viruses may have a role in the susceptibility to recurrent sinusitis and bronchitis.

Gastrointestinal infections/complications

Infections

Giardia lamblia

This is the only protozoal parasite to often cause symptoms in these patients. Mild to severe malabsorption may follow, with some patients complaining of abdominal distension, colicky pain, and flatulence. A secondary lactose intolerance may occur. The parasite may be difficult to eradicate with a standard course of metronidazole (2 g daily for 3 days), and tinidazole may be needed. It may be useful to give high-dose intravenous immunoglobulin therapy (2 g/kg body weight every 2 weeks) in resistant cases. Giardiasis is an uncommon complication nowadays in the Western world, probably because of higher dose immunoglobulin prophylaxis and improved cleanliness in the preparation of food.

Campylobacter

Campylobacter jejuni is the most frequent cause of bacteria-associated diarrhoea, usually responding to a course of a erythromycin (in adults, 500 mg four times a day for 10 days). Infection is currently uncommon in the United Kingdom, presumably because of improved hygiene. Stool culture will differentiate between shigella and salmonella infection, which occur no more frequently in these patients than in the general population.

Liver disease

The dependence on blood products led to a number of outbreaks of hepatitis C virus (**HCV**) infection before the early 1990s, with most infected patients dying from cirrhosis within 15 years. Hepatitis B virus (**HBV**) contamination of immunoglobulin products has not been a problem since the routine screening of blood donors started in the 1970s. No specific infection has yet been linked to the granulomatous hepatitis that occurs in at least 10 per cent of patients with CVID, the presinusoidal inflammatory reaction causing portal hypertension and oesophageal varices. The fact that about 30 per cent of patients with CVID in the United Kingdom have raised liver alkaline phosphatase serum levels is of concern as this is a good marker of liver involvement. This will lead to cirrhosis in some patients who should be considered for liver transplantation as their post-transplant survival is no worse than that for immunocompetent patients.

Sclerosing cholangitis is an important complication of X-HIM, and patients should be screened for abnormalities in liver function tests every 4 months, followed by cholangiography in those with persistently elevated liver enzymes. The cause is unknown, although cryptosporidial infection in the bile ducts may be the trigger in some cases. Affected patients should be considered for bone marrow transplantation, as well as liver transplantation for those with cirrhosis.

Malignancy

Apart from liver cancer and EBV-associated lymphoma in X-HIM and XLPS, respectively, malignancy is not a major complication of patients with hypogammaglobulinaemia. There is a small increase in lymphoma in XLA patients, with this being more impressive in patients with CVID (threefold increase over general population). However, a very high incidence of gastric carcinoma in patients with CVID was reported in the 1980s (50-fold increase over general population); this is now rare in the United Kingdom, possibly due to the wider use of prophylactic antibiotics for respiratory infection, which may have incidentally reduced the incidence of gastric infection with the cancer-promoting bacterium, *Helicobacter pylori.*

Prognosis

Recent surveys show that about 80 per cent of patients with CVID survive for 30 years, but morbidity and mortality depend on early diagnosis and management in an expert centre. The prognosis for X-linked agammaglobulinaemia is even better and is improving as patients are diagnosed earlier. Pneumonia and bronchiectasis remain the most common causes of death, liver and lung fibrosis being an additional complication in CVID. The overall prognosis for X-HIM and XLPS is poor, with most patients dying within 20 years of diagnosis.

Diagnosis

Diagnosing severe antibody deficiency is simple when the levels of serum IgG, IgA, and IgM are all very low or unrecordable. Death in childhood resulting from infection in a family member suggests one of the rare single-gene causes of immunodeficiency. For males presenting in childhood, blood should be sent to an expert laboratory to screen for Btk (XLA) and SAP protein (XLPS) in lymphocytes by Western blotting, and for expression of CD40 ligand on activated lymphocytes (X-HIM); XLA is likely if there is an absence of circulating B lymphocytes. Other rare syndromes should be considered in those presenting in childhood (see Fig. 3(a)). The interpretation of low immunoglobulin levels in children under 1 year of age is difficult, and further follow up is always needed to confirm the presence of a significant antibody deficiency. In affected families with single-gene immunodeficiencies, a molecular diagnosis can be made at birth and the infant started on treatment. Similarly, fetal diagnosis can be offered at about 14 weeks' gestation by screening DNA from fetal blood, or amniotic or chorionic villous cells.

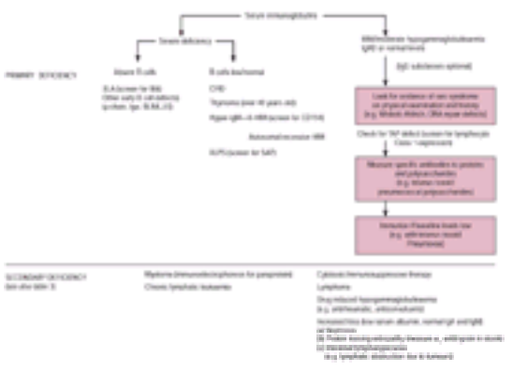

Fig. 3 A scheme for the diagnosis of primary and secondary immunoglobulin (antibody) deficiency.

Diagnosis of mild/partial antibody deficiencies

Selective IgA deficiency (IgAD)

The class switch to IgA requires a co-ordinated sequence of events within the germinal centre, involving continuing B-lymphocyte proliferation and T/B-cell interactions. It is therefore not surprising that IgA deficiency is associated with a variety of defects in lymphocyte function. IgA deficiency, defined as a serum IgA level below 0.1 g/l, is the most common of the primary immunodeficiencies, and is often genetically linked to CVID with which it shares a major susceptibility genetic locus in the MHC region. It occurs mainly in Caucasians, with about 1 in 700 of the population affected in northern Europe; it is rare in Africans (~1:6000) and very rare in Japanese (1:18 000). IgA deficiency is also associated with inherited single-gene defects in DNA repair (for example, ataxia telangiectasia) and major cytogenetic defects in chromosome 18. Various antirheumatic and anticonvulsant drugs can induce IgA deficiency. Most patients with IgA deficiency are healthy, and the defect is discovered either by chance or during surveys of families with CVID. A small percentage are discovered during investigation for recurrent infections, but these patients usually have additional defects in IgG antibody production, and therefore have a mixed partial deficiency.

Some IgA-deficient individuals have high levels of serum anti-IgA antibodies, which may cause anaphylactic reactions during blood or blood product infusion. There is a slightly raised incidence of IgA deficiency in patients with coeliac disease, probably because of shared susceptibility genes in the MHC region.

Other selective class deficiencies

Complete selective IgM deficiency, the mechanism not being understood, is rare and usually discovered by chance in patients not susceptible to infection. Low IgM levels occur in Bloom's and the Wiskott–Aldrich syndrome (see below). Selective IgE deficiency has been described but is not clinically important.

IgG subclass deficiencies

The clinical significance of IgG subclass deficiency is controversial. As in IgA deficiency, the complete absence of a major IgG subclass is compatible with normal health in the Western world. There are four IgG subclasses: IgG1 having the highest serum level; and IgG2, IgG3, and IgG4 having sequentially lower levels. Many healthy individuals have IgG4 levels close to the limit of detection, and most immunologists in the United Kingdom no longer measure this subclass. Apart from rare individuals with inherited genetic defects in the constant-region genes for IgG1, 2, and 4, the mechanism of subclass deficiency is unknown, although some susceptibility genes are probably shared with CVID and IgA deficiency because all three types of defect can occur in the same family. IgG2 deficiency and IgA deficiency can occur together, particularly in patients with ataxia telangiectasia.

The four subclasses have different functional capacities in relation to activating the first component of complement (IgG2 being weak) and Fc-g receptors on phagocytes. Specific IgG antibodies to bacterial components are also skewed towards certain subclasses, with those to polysaccharides being predominantly IgG2 in adults, and those to viral proteins being mainly IgG3. However, attempts to link subclass deficiencies with a predisposition to particular infections has been

unsuccessful, probably because of the flexibility and redundancy in the immune system.

IgG subclasses can be measured in most routine immunopathology laboratories, but it is difficult to show that the results influence clinical management. Moreover, there is no official International Standard serum for the subclasses, making it difficult to compare results from different laboratories. Experience has shown that it is not worth measuring subclass levels in patients with total IgG levels above 8 g/l; furthermore, measuring them in children under 5 years is of little use because of the wide range of levels in healthy children in this age group. The current consensus is moving towards measuring the levels of functional IgG antibodies as a better indicator of immune status in those with recurrent infections.

Functional immunoglobulin deficiencies

Functional deficiency is defined as a complete or partial failure to produce antibodies to specific proteins or polysaccharides, in the presence of normal total serum immunoglobulin levels. The mechanism is not understood and its prevalence in the general population is not known. In practice, only functional IgG antibody deficiency is considered clinically important for protection against infection. The standard practice is to measure baseline levels of antibodies to proteins such as tetanus toxoid, and polysaccharides such as those purified from the capsules of pneumococci; if these are low then the response after immunization is measured. Antibodies to other antigens such as diphtheria toxin, measles and polio viruses can also be measured to provide a broader range of responses. There are workable normal values for baseline levels of antibodies to tetanus and diphtheria toxins, and for pneumococcal and *Haemophilus influenzae* B polysaccharides, but the interpretation of responses after vaccination is difficult because of the paucity of data from healthy age-matched individuals. Nevertheless, complete failure to respond following a second vaccination is evidence of an abnormality that may influence the clinical management.

Treatment

Immunoglobulin replacement therapy

Immunoglobulins for therapeutic use are manufactured from large pools of donor blood (about 20 000 donations). Those used for intramuscular (**IMIG**) or subcutaneous (**SCIG**) injection are approximately 16 per cent solutions, while those for intravenous use (**IVIG**) are less concentrated (6–12 per cent solutions). The manufacturing process involves alcohol precipitation of plasma to produce an IgG concentrate with very little IgA or IgM remaining. All preparations should be subjected to rigorous safety measures, which include screening donors for HIV, HBV, and HCV infection, and manufacturing procedures which inactivate a wide range of viruses. Fortunately, HIV is inactivated by alcohol, but there were outbreaks of HCV hepatitis caused by contaminated batches prior to the introduction of improved safety measures in the early 1990s.

Immunoglobulin prophylaxis protects against pneumococcal and *H. influenzae* septicaemia, parvovirus, and probably reduces the susceptibility to infection from *Giardia* and *Campylobacter* spp. However, it is much less effective in preventing infection with mycoplasmas and enteroviruses. It reduces the frequency of acute bronchitis in antibody-deficient patients, probably by preventing common respiratory viral infections. However, there is poor penetration into the respiratory mucosa with little effect on the growth of *H. influenzae* in the respiratory tract.

Indications and dose

Most patients with severe hypogammaglobulinaemia and a history of recurrent infections should be offered immunoglobulin replacement therapy. IVIG at a dose of 400 mg/kg every 4 weeks is usually given in the United Kingdom and the United States, but an alternative regime is to give SCIG at an equivalent total dose every week; this route is popular in Scandinavia. Intramuscular therapy is now rarely used because it is painful. With adequate training, many patients infuse at home, with a nurse or partner inserting the intravenous lines for IVIG, or using infusion pumps for SCIG. The aim is to maintain the preinfusion (trough) IgG level towards the lower limit of the normal range (~8 g/l). Failure to maintain this level on standard doses suggests an increased loss or hypercatabolism of IgG, the latter being a useful marker of chronic infection and/or inflammation. The majority of patients tolerate IVIG and SCIG well. About 10 per cent of patients experience mild 'reactions' during IVIG infusions (for example, headaches, mild fever, backache), but these can usually be controlled by reducing the infusion rate and/or giving an antihistamine; reactions needing adrenaline (epinephrine) and steroid therapy are rare. Changing the immunoglobulin product may be helpful in those with recurrent reactions. Some reactions are caused by high plasma levels of anti-IgA antibodies, although there is poor correlation between the level of these antibodies and susceptibility to reactions. Nevertheless, anti-IgA antibodies are usually measured routinely in patients with a serum IgA level below 0.1 g/l, and an immunoglobulin preparation chosen with minimal contaminating IgA for those with very high levels of anti-IgA. In practice, patients benefit from being referred to specialist centres for immunoglobulin therapy where the response to infusions can be assessed by experienced staff.

Reaching a decision on the treatment of patients with mild or moderate antibody deficiency is more difficult. It may be best to use prophylactic or intermittent courses of antibiotics, particularly in children. Objectively assessing the efficacy of immunoglobulin prophylaxis is not straightforward, particularly since there is likely to be a significant placebo effect.

General management

Patients should be encouraged to take antibiotics early to treat bronchitis, and those with structural lung damage may require long-term prophylaxis. The quinolones (for example, ciprofloxacin) are particularly effective because they are concentrated in the mucous layer lining of the respiratory tract and have a very low minimal inhibitory concentration for non-typable *H. influenzae*. Amoxicillin, alone or in combination with clavulanic acid, or cotrimoxazole, is a useful alternative. Postural drainage and regular exercise are useful in promoting the removal of secretions from the lungs. Patients should be encouraged to join support groups which provide educational literature in lay language and assistance for social problems.

Major defects in cellular (T cell) immunity

Thymic aplasia (Di George syndrome)

This rare condition (about 1 in 3500 live births) is caused by fetal malformation of the third and fourth branchial arches at about 7 weeks of gestation, apparently due to abnormal cephalic migration of neural crest cells into these regions. These cells contribute to the development of the skull, palate, thymus, and parathyroid glands, explaining why the syndrome is associated with dysmorphic facies, palatal abnormalities, and hypoparathyroidism—affected infants sometimes presenting with tetany and convulsions due to hypocalcaemia. Although originally thought to be caused by teratogens or maternal disease, the majority of cases are now known to be associated with a chromosomal deletion at 22q11. The condition overlaps with other genetic disorders such as the velocardiofacial syndrome and conotruncal anomaly face syndrome, the phenotype broadening into endocrine, cognitive, and neurological defects. The cardiac defects may require major cardiac surgery in infancy.

Most affected infants retain nests of thymic tissue in the neck and have only a moderate T-cell lymphopenia, mainly affecting CD8 T cells, which improves over the first few years of life. Antibody production is usually adequate, and there is an increased incidence of autoimmune disease such as thyroiditis and haemolytic anaemia. A minority of infants have a more severe T-cell defect and are prone to severe infections.

Treatment

Apart from treating the associated abnormalities (for example, hypocalcaemia), infants with circulating T-lymphocyte counts above 500/µl usually need no specific immunological intervention, but should be followed up regularly to confirm recovery of T-cell immunity. Those with a more profound T-cell lymphopenia may need either a bone marrow or thymic graft. Transplantation of partially HLA-matched postnatal thymic tissue has been successful, with the appearance of mature functional 'educated' host T cells in the blood a few months later. As for other patients with severe T-cell defects, live vaccines should be avoided and blood for transfusion irradiated to avoid graft-versus-host disease (**GvHD**). Infants with major cardiac defects should be screened for the condition before cardiac surgery.

Severe combined immunodeficiency disease (SCID)

These rare immunodeficiencies (about 1:30 000 live births) are caused by inherited mutations of genes that influence the maturation of lymphocytes, particularly T cells (Fig. 1). Those affected, who are usually infants or children, are susceptible to life-threatening infection with a wide range of pathogens and opportunistic microbes. The emphasis is on early diagnosis and transfer to a specialist centre for bone marrow transplantation. Most of the rare adult patients who present with

severe infection associated with 'idiopathic CD4+ lymphopenia' have a 'leaky phenotype' of one of the known types of SCID.

Clinical features

Symptoms usually start before 7 months of age with failure to thrive, diarrhoea due to parasitic or viral infection (rotavirus, enteroviruses), *Pneumocystis carinii* pneumonia, and oral candidiasis. Other common viruses causing serious disease are adenoviruses, cytomegalovirus, human herpesvirus-6, and respiratory syncytial virus; systemic aspergillus or candida infections are less common but usually fatal. Routine immunization with live vaccines, particularly **BCG** (bacille Calmette–Guérin), may cause severe infection in patients with SCID. Most patients die within 2 years without a successful bone marrow graft. There are no specific features for any of the different types of SCID. The tonsils are usually absent and the thymic 'shadow' absent on a chest radiograph. Skin rashes due to a GvHD reaction, and more rarely lymphadenopathy and hepatosplenomegaly (Omenn syndrome), may confuse the diagnosis (see below).

Immunology

Most affected infants and children are lymphopenic, a characteristic finding which is often overlooked on the routine blood count. Table 2 outlines the differential diagnosis of lymphopenia in this age group. SCID can be classified into subgroups depending on the presence of T, B, and natural killer (**NK**) cells (Fig. 4), those with no NK cells being more likely to engraft following bone marrow transplantation. The levels of serum immunoglobulins are of little diagnostic use during the first year of life. The proliferation of T lymphocytes in culture in response to mitogenic and antigenic stimulation is nearly always very depressed. The diagnosis may be obscured when there are normal numbers of circulating T lymphocytes, for example in the Omenn syndrome, or following engraftment by maternal cells *in utero* or cells from an unirradiated blood transfusion.

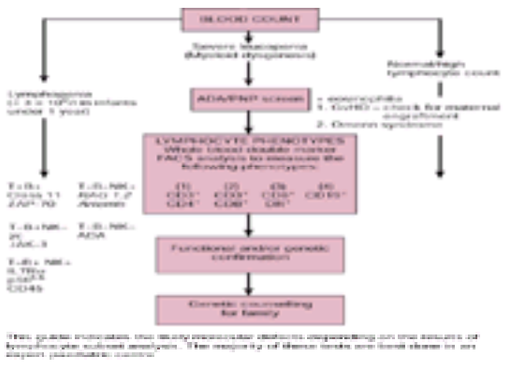

Fig. 4 Guide to the diagnosis of severe combined immunodeficiency (SCID) in infants and young children.

Molecular causes of SCID

There follows a brief description of the most frequent types of SCID, highlighting the contribution the discovery of the molecular defect has made to our understanding of the human immune system.

Defects of lymphocyte signalling

It was surprising when mutations in the g chain of the interleukin (**IL-2**) receptor were associated with X-linked SCID because the IL-2 'knock out' mouse had only a mild immunodeficiency. This led to a search for a wider role for the g chain, now known to be the signalling chain for the IL-2, IL-15, IL-4, IL-9, and IL-7 receptors, the latter being critical in the early development of T lymphocytes. This common g chain (gc) contains the appropriate tyrosine motifs on its cytoplasmic tail to facilitate the phosphorylation of Janus kinase-3 (JAK-3) which activates the signal cascade. Inherited defects in the IL-2-receptor a chain cause a multi-organ autoimmune disease, whereas defects in the IL-7-receptor a chain cause severe SCID. Failure to express HLA class II molecules on lymphocytes leads to severe SCID; research into the various molecular defects involved has improved our understanding of the factors that regulate class II gene transcription. Defects in the g and e chains of the CD3 complex are not always associated with severe immunodeficiency, suggesting some redundancy in the complex; inherited defects in ZAP (zeta chain-associated protein)-70 kinase, which interacts with the z chain, cause a predominantly CD8 lymphopenia. A defect in the p56 ^{lck} cytoplasmic kinase, which is important for signalling through the T-cell receptor, and complete absence of the surface protein tyrosine phosphatase receptor, CD45, are very rare causes of SCID.

*Defects in **VDJ** (variable, diversity, joining) recombination*

Molecular defects in the recombinase activating genes (*RAG1* or *RAG2*) lead to a variable phenotype depending on the amount of residual RAG protein produced. In cases of complete RAG1 or -2 deficiency there is failure to express a functional T- and B-cell receptor, the result being severe SCID. Partial RAG expression is one cause of Omenn syndrome, characterized by severe skin rash, lymphadenopathy, hepatosplenomegaly, eosinophilia, and autoimmune disease. Recently, mutations in another gene (Artemis), which appears to be involved in VDJ recombination and DNA repair, has been found in SCID infants who have a generalized sensitivity to ionizing radiation.

Defects in purine metabolism

Lymphocytes need efficient salvage and interconversion pathways for purines and pyrimidines during rapid bursts of proliferation, particularly in the lymphoid germinal centres and fetal thymus. Adenosine deaminase (**ADA**) deficiency was the first established cause of SCID: the condition also being the first example of enzyme replacement therapy (initially using red cell transfusions which contain ADA) in clinical medicine, and later being the first disease to be treated by gene therapy, although with only partial success. ADA has an important role in the intermediate pathways of purine metabolism. Purine-nucleoside phosphorylase (**PNP**) is also active in this pathway, but is a much rarer cause of SCID. It is interesting that deficiencies in both these enzymes predominantly affect lymphocytes, despite their presence in most other cells of the body.

Diagnosis and early management

In most cases the clinical diagnosis is obvious, the infant failing to thrive with a severe infection and profound lymphopenia. Congenital HIV has to be considered in the differential diagnosis, but is easily excluded by PCR tests for the virus. A detailed family history is important, asking whether the parents are related and whether there have been early deaths from infection in family members. Lymphoreticular malignancy may need to be considered in patients with lymphadenopathy and hepatosplenomegaly; biopsies from bone marrow, skin, liver, and lymph nodes should be reviewed by a histopathologist with experience of SCID and GvHD; it may be necessary to confirm T-cell engraftment (maternal or otherwise) using chromosome markers.

Affected infants often present in respiratory distress due to a viral or fungal pneumonia; co-trimoxazole should be given immediately for possible *Pneumocystis carinii* infection and the patient transferred urgently to an expert paediatric centre where the diagnosis can be confirmed and specific treatment arranged; time should not be lost trying to diagnose and treat infection before transfer.

Subsequent management

Having identified and treated any current infection, the next priority is to confirm the diagnosis of SCID and then counsel the parents on the practicalities and risks of bone marrow transplantation (**BMT**), regardless of whether the molecular defect is identified. Screening tests for some SCIDs can be completed within a few days (ADA, PNP, HLA class-II defects) but others may take some weeks/months unless the molecular defect is already known in the affected family. BMT from an HLA-identical sibling is the ideal treatment and has close to 100 per cent success, but haploidentical sibling or HLA-matched unrelated marrow has about an 80 per cent success. Transplantation of stem cells from HLA-matched cord blood is an alternative. Unlike BMT in non-immunodeficient patients (e.g. for leukaemia), no

preconditioning of the recipient is required to provide space within the marrow, but T-cell depletion of the donor marrow is necessary to minimize GvHD. The long-term outlook is good; there are now about 15 patients in the United Kingdom who have survived for between 12 and 20 years following BMT; the majority are healthy, although 20 per cent require regular immunoglobulin infusions because of failure to engraft donor B cells. For those rare adult patients who are diagnosed with SCID, bone marrow transplantation has previously been considered too risky; however, this view is changing since they rarely survive beyond a few years after presenting with severe infection. Some patients with ADA deficiency can be maintained on regular injections of bovine ADA, but this is very expensive and often does not completely correct the immune defect.

SCID is an ideal condition for gene therapy, but there have been problems in transfecting enough copies of the relevant gene into host bone marrow stem cells. However, this has been successfully achieved in at least four infants with X-linked (gc-deficient) SCID, possibly due to a selective advantage for the transfected stem cells; these patients now have normal immunity at up to 1 year of follow up.

Defects in the interferon-g/IL-12 pathway and susceptibility to mycobacteria

The study of rare children with a familial susceptibility to fatal mycobacterial infection has confirmed animal studies showing the critical importance of interferon-g (IFN-g) in stimulating macrophages to kill mycobacteria. Affected children have mutations in components of the circuit involved in delivering the signal to macrophages: which are the IL-12 p40 subunit, the IL-12 receptor b1 subunit, both chains of the IFN-g receptor (**IFN-gR**), and STAT1 (STAT, signal transducer and activator of transcription; a signalling component downstream of the IFN-gR). These are all autosomal inherited conditions in which the heterozygote carriers might be expected to be healthy; however, some families with heterozygotes for IFN-gR defects show a dominant inheritance pattern due to disruption of the receptor complex by non-functional chains. A similar dominant-negative effect has been described in a patient with a STAT1 defect because two functional molecules must be recruited to the cytoplasmic domain of the IFN-gR to provide a signal. Regular IFN-g therapy is useful for those with normal IFN-gR function and downstream signalling, while the others require bone marrow transplantation. It is extraordinary that affected patients are so selectively susceptible to mycobacterial disease, particularly BCG and atypical strains such as *M. avium* and *M. fortuitum*. This suggests that this lymphocyte/macrophage interactive circuit has been selected by humans to specifically cope with mycobacterial infection (Fig. 5).

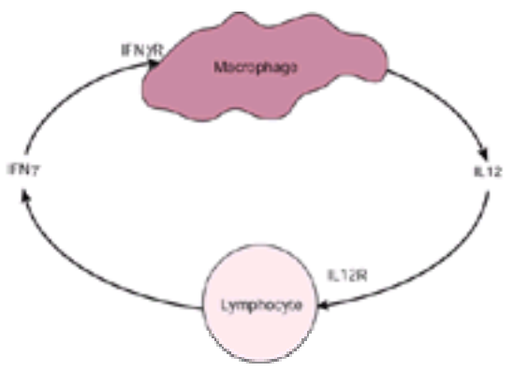

Fig. 5 This circuit is crucial for the effective killing of mycobacteria. Macrophages secrete IL-12 following uptake of bacteria, which amplifies the production of interferon-g by sensitized T lymphocytes, which in turn stimulates the macrophage to kill the organism.

Inherited syndromes associated with immunodeficiency

Defects in DNA repair

Efficient repair of DNA damage is fundamental to cell survival. Our knowledge of the cascade involved in the excision of damaged nucleotides, insertion of new nucleotides, and rejoining (ligation) of the DNA strands is rapidly expanding, helped by the study of rare syndromes caused by genetic defects in this pathway. Ataxia telangiectasia (**A-T**) is an autosomal recessive disease characterized by progressive cerebellar ataxia, chromosomal instability, telangiectasia on exposed areas of skin, early death from cancer, and immunodeficiency of variable severity. About 80 per cent of patients have IgA deficiency, with a third having complete absence of IgA; a minority have additional defects in IgG production, often IgG2 deficiency, while a few have severe panhypoimmunoglobulinaemia. T-lymphocyte function is often depressed. The relevant gene codes for a protein involved in the regulation of the cell cycle, probably having a role in the suspension of DNA replication after damage from ionizing radiation to allow time for repair. The defective gene in A-T leads to chromosomal instability and susceptibility to cancer, particularly lymphoma associated with translocations between chromosomes 4 and 7 involving the genes that code for immunoglobulin heavy chains and T-cell receptor a, b, g chains. Most patients with A-T die before their third decade from either tumours or respiratory infection, the latter usually caused by a combination of immunodeficiency and progressive neurological deterioration. About 1 in 200 of the general population is heterozygous for the genetic defect, and there is some evidence that they are at an increased risk of malignancy. Furthermore, the gene is mutated in some types of leukaemia cells (e.g. T-prolymphocytic leukaemia), suggesting its product acts as a tumour suppressor.

The Nijmegen breakage syndrome (**NBS**) has a similar phenotype with additional craniofacial abnormalities, including progressive microcephaly. The normal physiological function of the *NBS* gene is not known, but like *ATM* leads to chromosomal instability following exposure to DNA-damaging agents. Other recessive chromosomal instability syndromes predisposing to cancer are caused by mutations in the DNA ligase-1 gene (with severe immunodeficiency and dwarfism), and the helicase mutated in Bloom's syndrome which is associated with moderately low immunoglobulin levels.

Wiskott–Aldrich syndrome

This X-linked disease is characterized by thrombocytopenia, moderate immunodeficiency, eczema, autoimmune disease (including vasculitis), and susceptibility to EBV-induced B-cell lymphomas. Patients have a dysregulated humoral response with depression of IgM antibody production to polysaccharides, and often a raised serum IgE. There is a milder variant resulting in only thrombocytopenia. The defective gene codes for a cytoplasmic protein (**WASP**, Wiskott–Aldrich syndrome protein) which is involved in cytoskeletal reorganization following the activation of platelets and T lymphocytes. The diagnosis is based on the presence of small platelets and on demonstrating the absence of WASP in white cells by Western blotting. Splenectomy may be needed to reduce the thrombocytopenia, and bone marrow transplantation is recommended for most patients because of the poor prognosis.

TAP deficiency

The transporter associated with antigen processing (**TAP**) is composed of two subunits (TAP-1 and -2) and facilitates the transport of HLA class I molecules from the endoplasmic reticulum to the *cis*-Golgi compartment. Inherited defects in TAP (so far only confirmed for TAP-2) lead to the failure to express class I molecules on the lymphocyte surface, preventing cytotoxic T and NK cells from recognizing antigen in the context of 'self' class I molecules. However, adequate cytotoxic function against virus-infected cells is retained using mechanisms that are not completely understood. Affected patients are prone to progressive bronchiectasis that is not entirely explained by infection. Some patients have developed nose and mid-face destruction, similar to midline granuloma, probably caused by a failure to inhibit NK-cell self-destruction via class I mediated inhibitory signals.

Other rare syndromes associated with severe infection

Chronic mucocutaneous candidiasis is a very rare sporadic disease of unknown cause, which in some patients is associated with multiple endocrine abnormalities. Patients have subtle defects in humoral and cellular immunity that do not explain the severity of the candida infection. Most patients can be managed satisfactorily with long-term antifungal therapy (fluconazole or itraconazole). The hyper-IgE (Job's) syndrome is another poorly defined disorder characterized by eczema, deep staphylococcal abscesses, and serum IgE levels usually in excess of 10 000 kU/l. Many patients have consistent facial features, delayed shedding of primary teeth, and hyperextensible joints suggesting they share the same underlying genetic defect. The Chediak–Higashi and Griscelli syndromes are autosomal-recessive diseases characterized by the presence of giant lysosomes in all granulated cells that compromises the function of neutrophils and NK cells. The relevant genes have been identified but their precise function is still unknown. Patients often die from infection or bleeding due to thrombocytopenia during an 'accelerated phase', which is

similar to the virus-associated haemophagocytic syndrome. Bone marrow transplantation will correct the haematological abnormalities but not the other features, which include albinism and various neurological abnormalities.

Immunodeficiency associated with other congenital or inherited disorders

There are many rare disorders causing major multisystem disease in infants and young children that are associated with variable immunodeficiency states. Examples are inherited metabolic defects such as transcobalamin-2 deficiency (causing immunoglobulin deficiency secondary to severe vitamin B12 deficiency) and biotin-dependent carboxylase deficiency (causing a severe T-cell defect). A variety of skeletal (e.g. cartilage–hair hypoplasia), growth disorders (e.g. Schimke immuno-osseous dysplasia), and major dermatological abnormalities (e.g. ectrodactyly ectodermal dysplasia) are associated with T-cell defects and early death from infection. (For a comprehensive list of these disorders see reference to the IUIS Report, 1999.)

Secondary immunodeficiencies (Table 3)

Lymphoid malignancies, immunosuppressive agents, and AIDS are common causes of severe immunodeficiency, while nutritional deficiencies, metabolic disturbances (for example, uraemia), and trauma have a less severe effect on the immune system. In many of these situations the primary disease usually overshadows the immunodeficiency, although attention to the latter can improve the patient's quality of life.

Recurrent pneumonia and bronchitis suggest antibody deficiency, whereas varicella-zoster and herpes-simplex reactivation, oral candida, and rapid growth of skin warts are often early indications of a defect in cellular immunity. The presence of lymphopenia, often overlooked, indicates that the immune system is compromised but is a poor guide to the clinical significance of the defect. In practice, measuring the numbers of circulating CD4+ T cells and serum immunoglobulins are useful simple tests for monitoring the severity of the immunodeficiency.

Lymphoid malignancy

Various types of lymphoreticular malignancy are associated with both humoral and cellular immunodeficiency, exacerbated by the use of cytotoxic drugs. There is no consistent pattern of immunodeficiency for any particular lymphoid malignancy, presumably the severity depending on the genetic background of the patient and immunomodulating factors released from the malignant cells. However, an important exception is chronic lymphatic leukaemia (**CLL**) in which the majority of patients develop hypogammaglobulinaemia during the course of their disease. Although the immunoglobulin deficiency in most patients is mild, a few have severe hypogammaglobulinaemia and suffer from recurrent infections, particularly of the upper and lower respiratory tract; these patients will benefit from regular immunoglobulin replacement therapy, while others can be managed with prophylactic or intermittent courses of antibiotics. The cause of the antibody deficiency is complex and seems to be due to a combination of inhibitory factors released by the malignant clone and interference with the normal traffic of T and B lymphocytes through the lymphoid apparatus by proliferating CLL cells.

Patients with myeloma often have antibody deficiency, which explains their predisposition to pneumococcal pneumonia and septicaemia. In the past, few haematologists paid attention to the immunodeficiency because of the very poor prognosis of the underlying condition. However, modern cytotoxic therapy can now induce prolonged remissions, during which the immunodeficiency recovers, so it may be worth treating the more severely immunocompromised patients with immunoglobulin during the induction period. There is evidence that the malignant plasma cells produce factors that inhibit normal antibody production.

The increasing use of bone marrow transplantation to treat leukaemia carries a legacy of persistent antibody deficiency in a minority of patients due to inadequate B-cell engraftment and/or the drugs used to prevent rejection. Follow-up protocols should include appropriate screening to identify those patients who may require immunoglobulin replacement.

Drugs

The extensive literature on the immunological effects of cytotoxic agents and steroids will not be reviewed here. Many of these drugs have a profound effect on cellular immunity, as shown by the severity of varicella infection in patients treated with corticosteroids, and the risk of cytomegalovirus and EBV reactivation in those on immunosuppression therapy to prevent graft rejection. Some of these drugs, particularly cyclophosphamide and azathioprine, may compromise antibody production after prolonged use. A variety of antirheumatic and anticonvulsant drugs induce a partial (often IgA) deficiency, and occasionally a severe antibody deficiency in a small minority of treated patients, probably due to their genetic susceptibility to the metabolic effects of the drug on B-cell differentiation and/or antigen presentation. The effects are reversible, but it may take up to 2 years for antibody production to recover after stopping the drug.

Viruses

HIV is the most common and important immunosuppressive virus, and is described in Chapter 7.10.21. Many other viruses cause moderate immunosuppression during active infection, particularly measles which depresses cellular immunity. Fetal infection with the rubella virus may, rarely, lead to prolonged depression of IgG and IgA antibody production after birth, sometimes with a high serum IgM level. Fetal cytomegalovirus infection can have a similar effect. There is evidence of prolonged alteration in the type of immune response after common childhood virus infections, some researchers suggesting that these events 'programme' the system towards a TH1 response and reduce the risk of allergy; the marked reduction in measles and other severe childhood infections due to vaccination has been suggested as one reason for the increase in childhood allergy, including asthma.

Immunodeficiency secondary to metabolic and nutritional defects

This is probably the most common cause of immunodeficiency worldwide and contributes to the high infant death rate in the Third World. Protein-calorie malnutrition and deficiency of vitamins and trace elements, particularly vitamin A, zinc, and probably selenium, can lead to significant depression of T-lymphocyte function and reduced antibody production. Poor nutrition in the very elderly in Western countries probably contributes to their poor antibody responses and an increased risk of pneumococcal pneumonia. Vitamin A supplementation has been shown to reduce childhood mortality from infection in New Guinea.

Prolonged metabolic disturbances associated with liver and renal failure will compromise immunity; this persists in about 10 per cent of patients on ambulatory peritoneal dialysis who have low IgG levels and are susceptible to infection, and may be due to a combination of persistent uraemia and hypercatabolism of IgG by activated peritoneal macrophages.

Severe trauma and major surgery often compromises both T- and B-lymphocyte function, but is usually clinically masked by the routine use of broad-spectrum prophylactic antibiotics and immunoglobulin provided in blood transfusions. Even full-thickness burns involving less than 10 per cent of surface area in young children appear to suppress IgG2 and IgG3 subclass production for at least a week. This observation provided an explanation for the high incidence of deaths from the toxic-shock syndrome in one centre and prompted the routine use of prophylactic antibiotics on admission. In major surgery, particularly when hypothermic cardiopulmonary bypass is used in elderly patients, attempts are being made to reduce the risk of postoperative infection by 'boosting' the nutritional requirements of the immune system with supplements such as L-arginine and nucleotides.

Increased catabolism/loss of immunoglobulin

Loss of immunoglobulin from the kidney or bowel is an important cause of mild/moderate hypogammaglobulinaemia, but is rarely of clinical significance. Serum IgM, being a larger molecule, is usually normal, with low IgA and IgG levels. The nephrotic syndrome and protein-losing enteropathy are the most common causes, the latter being difficult to diagnose when the serum albumin level is normal. Leakage of protein and lymphocytes occurs in primary or secondary intestinal lymphangiectasia—the combination of hypogammaglobulinaemia, low serum albumin level, and lymphopenia being a useful clue to this diagnosis. An increase in the catabolism of many proteins occurs in chronic infection/inflammation, but this is never severe enough to cause severe hypogammaglobulinaemia unless there is an associated primary defect in immunoglobulin synthesis. A selective increase in the catabolism of IgG occurs in dystrophia myotonica, but the mechanism is unknown.

Further reading

IUIS Scientific Committee (1999). Primary immunodeficiency diseases. Report of an IUIS Scientific Committee. *Clinical and Experimental Immunology* **118**(Suppl 1), 1–34.

Ochs HD, Smith CIE, Puck JM, eds (1999). *Primary immunodeficiency diseases. A molecular and genetic approach*. Oxford University Press, Oxford.

Webster ADB (2001). Common variable immunodeficiency. In: Roifman C, ed. *Immunology and Allergy Clinics of North America*, Vol 21, pp 1–22. WB Saunders, Philadelphia.

5.7 Principles of transplantation immunology

Kathryn J. Wood

Introduction

Transplantation of an organ, tissue, or cells between genetically disparate individuals within the same species, allografts, or between species, xenografts (Table 1), almost inevitably results in rejection of the graft if active steps are not taken to control the destructive immune response that is triggered immediately after transplantation.

Studies on the behaviour of tumour grafts in the early part of the twentieth century led Peter Gorer to formulate the concept of graft rejection in 1938. Gorer's description of what triggers rejection still holds today, even if the language he used does not fit with current immunological jargon: 'isoantigenic factors present in the graft tissue and absent in the host are capable of eliciting a response which results in the destruction of the graft'. The recognition that the immune system was involved came nearly 10 years later when Gibson and Medawar clearly identified specificity and memory as hallmark features of the rejection response.

The rejection process is complex. Many factors, including the nature of the tissue transplanted and the genetic disparity between the donor and recipient, the site of transplantation, as well as the immune status of the recipient, all contribute to determining the character of the rejection response (Table 2).

The events that lead to allograft rejection are summarized in Fig. 1. In brief, inflammation as a result of the removal of the graft from the donor and implantation into the recipient is always triggered as a result of the transplantation procedure itself, irrespective of whether the tissue is allogeneic or xenogeneic in origin. These 'danger' signals are responsible for activating both the innate and adaptive immune systems that act in concert to destroy the graft. For acute allograft rejection, activation of the adaptive immune system requires recognition of molecules that are mismatched or polymorphic between the donor and the recipient. Antigen recognition in combination with additional signals, termed costimulation, leads to the activation of donor reactive lymphocytes, both T cells and B cells. Clonal expansion, meaning proliferation of donor reactive lymphocytes, is triggered such that many more daughter cells with donor antigen specificity are produced rapidly. The environment created by such lymphocyte activation results in the differentiation of the activated donor reactive lymphocytes into effector cells, including cytotoxic T cells and mature B cells or plasma cells that secrete anti-donor antibodies. These antigen-specific effector cells in combination with activated components of the innate response, such as activated macrophages and natural killer cells, orchestrate the destruction of the graft (Fig. 1).

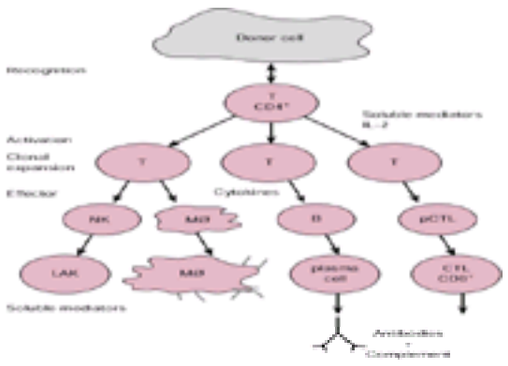

Fig. 1 Overview of allograft rejection. There are three phases to the responses—recognition (direct pathway allorecognition is illustrated), activation, and the generation of effector mechanisms. Each step involves the orchestrated interaction of cells and molecules to ensure that the response is driven towards an aggressive phenotype that will result in the destruction of the transplant. Immunosuppressive drug therapy is designed to interfere at different stages in the response to ensure effective inhibition of rejection.

If immunosuppressive drugs such as cyclosporin, tacrolimus, mycophenolate mofetil, or azathioprine are administered at the time of transplantation, many of the events that lead to acute allograft rejection can be inhibited. As a result of the effective use of these drugs in clinical transplantation the short-term, 1-year, graft survival rates for all solid organ grafts have increased dramatically in the last 20 years (up-to-date summaries of graft survival data can be obtained from the websites listed at the end of this chapter). Unfortunately, this short-term success has not translated into significantly improved long-term, more than 10-year, graft survival outcome. Following the first year after transplantation there is still a steady attrition of grafts; this delayed or late graft loss occurs due to a variety of different processes and factors, only some of which are immunological. Late graft loss is often referred to as chronic allograft rejection (Table 2). Unfortunately, the drugs in use in clinical transplantation at present are relatively ineffective at preventing chronic stimulation of the immune system by the graft in the longer term after transplantation.

When tissues are transplanted between species (xenotransplantation) where the recipient species has preformed natural antibodies against the donor (so-called discordant species that include pig to human), additional immunological events contribute to the destruction of the graft, resulting in the very rapid elimination of the graft through a process known as hyperacute rejection (Table 2). In the pig to human species combination, preformed natural antipig antibodies bind to carbohydrate determinants present on pig cells. As a result the endothelial cell surface develops procoagulant activity causing leucocytes to accumulate in the vessels, complement is activated, and the tissue is rejected very rapidly. If hyperacute rejection can be inhibited, for example by removal of the preformed antibody before transplantation or by controlling complement-mediated damage to the graft, the downstream events involving the adaptive immune system will be triggered resulting in acute vascular or delayed xenograft rejection.

Hyperacute rejection can also occur when an allograft is transplanted into a recipient who has already been sensitized to the histocompatibility antigens of the organ donor (Table 2). In allotransplantation, anti-donor antibody formation can occur as a result of the rejection of a first graft, pregnancy, or blood transfusion. Rigorous screening processes, whereby sera from the recipient are cross-matched against tissue from the donor, ensure that the recipient does not have preformed antidonor antibodies and that hyperacute rejection of allografts hardly ever occurs in current clinical practice.

This chapter will outline the key cellular and molecular events that lead to the destruction of a graft by the immune system of a naïve recipient. The events that lead to allograft rejection will be dealt with in most detail alongside a summary of the sequelae that also need to be considered when xenogeneic tissue is transplanted.

Transplantation sends danger signals to the host

The removal of tissue for transplantation from the donor and its implantation into the recipient will result in a series of changes in gene expression within the donor tissue that will markedly influence the way the recipient's immune system responds. When the organ or tissue to be transplanted is harvested from a cadaver donor some of these changes are a direct consequence of brain death. In addition, the trauma associated with the surgical procedures required to remove and transplant the tissue contributes to the very early events that initiate rejection. These factors are often referred to collectively as the events associated with ischaemia and reperfusion injury. Indeed, it has been suggested that there is a link between the ischaemia time and increasing immunogenicity of the graft.

The consequences of these events include the release of preformed P-selectin (CD62P) from the Weibel–Palade bodies contained within endothelial cells. This is an adhesion molecule responsible for the earliest step in leucocyte migration into the tissue. There is also *de novo* expression of a variety of genes, including those encoding chemokines (chemoattractant cytokines) and other adhesion molecules by the transplanted tissue. Expression of these molecules by the graft creates a proinflammatory environment and results in changes in endothelial cell function and the recruitment of inflammatory leucocytes into the graft, as well as the exodus of donor-derived passenger leucocytes from the graft and their migration to recipient lymphoid tissue. Thus the graft itself initiates a vicious circle of events that contribute to its own destruction.

It is important to note that some of these initial changes will occur even when there are no antigenic differences between the donor and recipient, as is the case when a graft is transplanted between genetically identical individuals—a syngeneic graft. These events are associated exclusively with organ retrieval and the transplantation procedure itself. Of themselves, they are not sufficient to lead to the destruction of the graft, as evidenced by the lack of rejection of autografts and isografts (Table 1). However, they can have a marked influence on early graft function and they will have a significant effect on the way in which the innate and adaptive immune responses to the graft are both triggered and evolve when an antigenic disparity does exist. Moreover, it has been suggested that these early events can predispose the graft to late dysfunction or chronic rejection, the distinctive feature of which is transplant vasculopathy (Fig. 2).

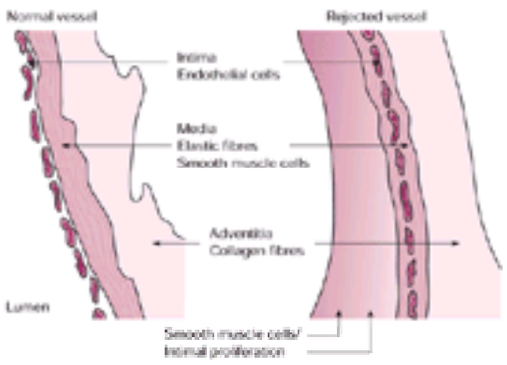

Fig. 2 Histological features of vascular rejection—the hallmark of chronic graft rejection. A normal and rejected vessel are shown in the cartoon. The rejected vessel (right) exhibits severe intimal proliferation compared with the normal vessel (left) as a result of the proliferation of smooth muscle cells.

One way of thinking about the changes that arise as a direct result of the removal and transplantation of tissue is in terms of the trauma of these events initiating a series of 'danger signals'. Receipt and integration of these signals by the host immune system, along with information about the genetic disparity of the tissue transplanted with the recipient, will determine whether and how the recipient immune system is triggered.

Role of the innate immune system

The innate immune system is used by the host as the first line of defence against any adverse event, including transplantation. It comprises a series of cells and molecules that are poised for action as soon as the normal resting situation in the body is perturbed. Elements of the innate immune system will be triggered by the danger signals arising from the trauma associated with the transplantation procedure. The nature of the components of the innate immune system involved in this phase of the rejection response are relatively poorly characterized but are likely to include the components of the complement system, particularly C3, and phagocytic cells such as macrophages.

The complement system is a cascade of proteolytic enzymes whose activation leads to opsonization of targeted cells as well as the generation of a membrane attack complex that can initiate cell lysis. Complement can be activated in a variety of ways, including by some of the proteolytic enzymes produced by the clotting cascade, as well as by contact with damaged or altered endothelial cells. Once activated the enzymes of the complement cascade release soluble mediators, such as C3a and C5a, that will attract leucocytes to the site of the graft, and also produce molecules that can bind covalently to the cells within the graft forming a focus for the damaging events that follow.

Macrophages express a series of pattern recognition receptors, including those that recognize carbohydrate structures, reactive oxygen species, and activated complement components. When these receptors engage their ligands the macrophage is triggered to release a battery of inflammatory cytokines—including tumour necrosis factor (**TNF**), interleukin-1 (**IL-1**), and IL-6 amongst others—that further augment the proinflammatory environment and promote the activation of the adaptive response.

Role of the adaptive immune system

Antigens that stimulate allograft rejection

The degree of histocompatibility (tissue compatibility) between the donor and recipient determines whether a graft is rejected or accepted when transplanted between two members of the same species. In molecular terms this arises from a series of molecules, both cell surface and intracellular, that are polymorphic or variant between different members of the species—so-called histocompatibility antigens. These were originally classified as either major or minor depending on the location of the gene encoding the polymorphic molecule in the genome.

A series of cell surface molecules encoded by genes present within one region of the genome, the major histocompatibility complex (**MHC**), are known as the major histocompatibility antigens or MHC antigens. Many of these molecules are well characterized. Any other polymorphic molecules that trigger rejection are called minor histocompatiblity (**miH**) antigens. The genes for miH antigens are scattered throughout the genome.

Incompatibility or mismatching for either MHC or miH antigens can trigger graft rejection. In general, in naïve recipients the greater the number of incompatibilities for MHC and miH antigens, the more vigorous the rejection response. However, the type of tissue transplanted as well as the site of transplantation will have a marked influence on graft outcome, even when the matching for MHC and miH antigens between the donor and the recipient is identical. For solid organ grafts such as the kidney, matching for MHC antigens between the donor and the recipient improves graft outcome in immunosuppressed recipients. However, in bone marrow transplantation even grafts transplanted between individuals who are identical for MHC antigens can still trigger an immune response, either rejection or graft-versus-host disease, as a result of mismatching for miH antigens.

MHC class I and class II molecules

The MHC encodes a series of polymorphic genes in every species of vertebrate (Fig. 3). Within any one species a large number of variant forms of each of these genes exists within the population as a whole. Of the genes present in the MHC there are two families that code for cell surface molecules known as the MHC class I and MHC class II molecules (Fig. 3). Some of the loci that form part of the class I and class II families have been well characterized and in humans these are called HLA A, HLA B, and HLA C, and HLA DR, HLA DQ, and HLA DP, respectively. Additional class I and class II genes are present in the MHC, but they are less well

characterized than those mentioned above and polymorphisms in these molecules are not considered routinely before either organ or bone marrow transplantation at present and they will not be discussed further here.

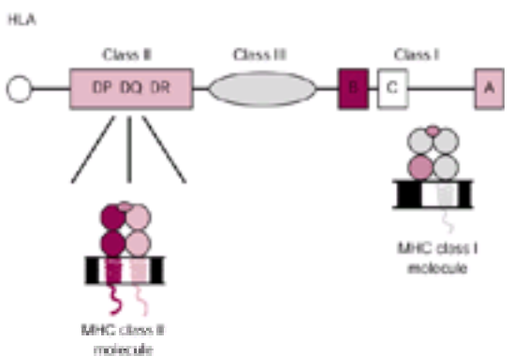

Fig. 3 Outline map of the major histocompatiblity complex (MHC) in man. The HLA gene complex maps to the short arm of chromosome 6. It is divided into regions and subregions that in simple terms each contain a family of genes. Only the well characterized loci are shown in this representation: HLA A, HLA B, and HLA C class I a-chain genes, and HLA DR, HLA DQ, and HLA DP class II A and B genes. Additional class I and class II genes have been described. A full map for the HLA region is available at the website attached to *Nature* (1999) **401**, 921–3.

MHC class I moleucles are cell surface glycoproteins comprising two polypeptide chains; the polymorphic a chain (molecular mass, **MM**, 45 kDa), which is anchored in the plasma membrane and encoded by a gene in the MHC, and b_2-microglobulin (MM: 12 kDa), which is not anchored in the membrane and is encoded by a gene on another chromosome (Fig. 3). MHC class I molecules are expressed on virtually all somatic nucleated cells, albeit at different levels in the resting state. Their expression is rapidly upregulated in response to cytokines such as interferon-g (IFN- g) and tumour necrosis factor- a (TNF-a) that are produced during an immune response. After transplantation, mismatched intact donor MHC class I molecules expressed by donor cells can be recognized and trigger the activation of recipient CD8+ T cells.

Class II molecules are also cell surface glycoproteins built up of two polypeptide chains. However, in contrast to class I, both chains—a and b (MM: 35 and 28 kDa, respectively)—are anchored in the plasma membrane. The two chains are encoded by genes found in the MHC; class II A and B genes for the a and b chains, respectively. Both genes can be polymorphic. MHC class II molecules are not expressed by all cells in the body, their tissue distribution is therefore much more restricted than for class I molecules and expression is only found constitutively on some cells, including dendritic cells, B lymphocytes, macrophages, and some endothelial cells. Importantly, expression of MHC class II molecules can not only be increased on the cells that already express class II molecules but can be induced on other cell types during an immune response. After transplantation, mismatched MHC class II molecules expressed by donor cells can be recognized and trigger the activation of recipient CD4+ T cells.

During the biosynthesis and transport of MHC molecules to the cell surface they become associated with short peptides derived from both intracellular and extracellular proteins. This process is known as antigen processing and presentation. As a result of these antigen processing pathways, MHC class I and class II molecules expressed at the cell surface report the status of the internal and external environment of a cell to the immune system. When the cell is functioning normally the peptides associated with MHC molecules are derived from self proteins, that is, the proteins belonging to the tissue itself, and including peptides derived from the MHC molecules themselves. However, when there is an adverse event such as a pathogen invading either the cell itself or its local environment, the MHC molecules will become loaded with peptides derived from the invader. It is this peptide–MHC complex that is recognized by T cells.

In the context of transplantation the situation is slightly more complex. Before transplantation, donor MHC molecules expressed by the transplanted tissue will contain peptides of donor origin. After transplantation, these donor-derived MHC–peptide complexes can be recognized by recipient T cells via the so-called direct pathway of allorecognition (Fig. 4). However, recipient antigen-presenting cells also come into contact with donor cells and molecules (see below) and through the normal pathways of antigen processing and presentation peptides of donor origin become associated with recipient MHC molecules in just the same way as any other foreign antigen. Recipient MHC–donor peptide complexes can then be recognized by recipient T cells via the so-called indirect pathway of allorecognition (Fig. 4), and this pathway is also used for recognition of mismatched miH antigens. Thus, after transplantation there are two routes of presentation of donor MHC molecules to the recipient immune system, the direct and the indirect pathways of allorecognition.

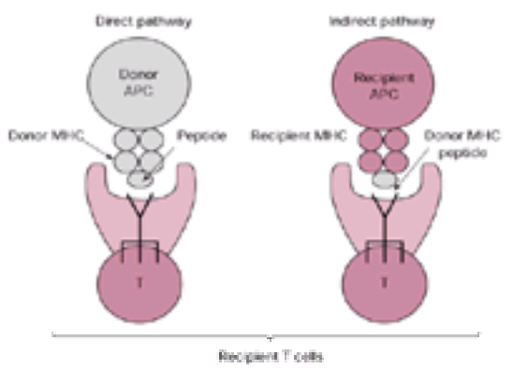

Fig. 4 Direct and indirect pathways of allorecognition. (1) Direct pathway: Donor MHC–peptide complexes are presented to recipient T cells by donor-derived antigen-presenting cells. Two hypotheses have been proposed to explain the high frequency of T cells, between 1 and 10 per cent of the repertoire, that can respond to alloantigens presented in this way. (i) High determinant density: the similarity in structure between MHC molecules results in T-cell receptors exhibiting cross-reactivity for donor MHC molecules irrespective of the peptide that is bound to each molecule. When donor molecules are expressed at high levels, as is the case on donor-derived passenger leucocytes, a sufficient number of T-cell receptors will engage the molecule to trigger a response. (ii) Multiple binary complexes: each donor MHC–peptide complex can be recognized by a different clone of T cell in the recipient giving rise to a high overall frequency of responding cells. (2) Indirect pathway: Donor MHC and miH antigens are processed by recipient antigen-presenting cells and presented as peptides by recipient MHC molecules. Each recipient MHC–donor peptide complex can be recognized by T cells in the recipient. The frequency of responding cells is of the same order of magnitude to T cells responding to other nominal antigens, such as viral antigens.

Two pathways for presentation of donor antigens to recipient T cells

Bone marrow-derived passenger leucocytes are present in non-lymphoid tissues throughout the body and have the characteristics of immature dendritic cells. After transplantation, in response to inflammatory cytokines and other danger signals, the donor-derived passenger leucocytes migrate out of the graft very rapidly and end up in the recipient lymphoid tissue. The migration process results in the passenger cells acquiring the phenotype and function of mature dendritic cells. Mature dendritic cells are often referred to as professional or immunostimulatory antigen-presenting cells as they express high levels of MHC class I and class II molecules as well as other cell surface and soluble molecules that enable them to stimulate naïve CD4+ and CD8+ T cells to respond (Fig. 5). The additional molecules required for an antigen-presenting cell to stimulate the activation of naïve T cells include costimulatory molecules such as members of the B7 family, in particular CD86, CD40, and adhesion molecules. Thus the donor passenger leucocytes that end up in the recipient lymphoid tissue have all of the attributes required for them to present any donor MHC molecules that were mismatched between the donor and the recipient to recipient T cells via the direct pathway of allorecognition.

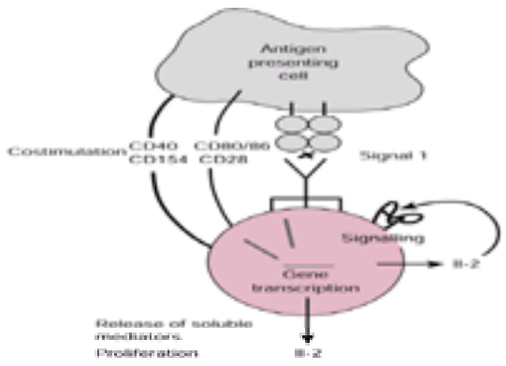

Fig. 5 Antigen presentation to naïve T cells. In conjunction with antigen recognition, additional signals or costimulation are required to trigger T-cell activation. Some of the molecules involved are illustrated, including on the T-cell side CD28 and CD154 (CD40L) and on the antigen-presenting cell side CD86 and CD40.

Evidence that the donor-derived passenger leucocytes play an important role in initiating rejection comes from studies showing that in certain situations removal of the passenger cells from grafts before transplantation can result in prolonged graft survival. However, this is not the case in every situation and the second route of antigen presentation—the indirect pathway—has also been shown to contribute to acute rejection as well as playing a significant role in the evolution of chronic rejection.

At the same time that donor antigen-presenting cells are migrating from the graft, recipient leucocytes are being attracted to the graft in response to chemokines (along with other mediators) released by the transplanted tissue. Amongst the cells recruited into the graft are circulating antigen-presenting cells. These take up debris arising from the tissue damage caused by the transplantation procedure itself, and then migrate to the draining lymphoid tissue. In addition, soluble antigens released as a result of damage to the tissue at the time of transplantation are also transported to the draining lymphoid tissue where they can be picked up by resident antigen-presenting cells. The captured antigens are then processed and presented as peptides with recipient MHC molecules to T cells in the T-cell areas of the recipient lymphoid tissue. In the context of transplantation this route of presentation is known as the indirect pathway of allorecognition (Fig. 4). It is clearly the more physiological of the two pathways that are used to trigger the activation of the response after allotransplantation. Moreover, indirect presentation of donor antigens is likely to continue in the long term after transplantation. Once all of the donor-derived passenger leucocytes have migrated from the graft they are obviously not replaced and therefore only so-called 'non-' or less-professional antigen-presenting cells, such as endothelial cells of donor origin, are available for the continued stimulation of direct pathway T cells.

It has been shown recently that migrating antigen-presenting cells are drawn to the correct area within the lymphoid tissue by chemokines, thereby ensuring that they come into contact with naïve T cells maximizing the chances of antigen presentation. Similarly, once a T cell has been triggered it migrates to other areas of the lymphoid organ, notably the B-cell area, in order to propogate the response and initiate the development of effector cells (Fig. 1).

Activation of recipient T cells

Recipients deprived of T cells either through manipulation of the immune system or through genetic mutations are unable to reject allografts. T cells are therefore a key element of the rejection response. The relative roles of CD4+ and CD8+ T cells in the initiation of the response will depend on the donor–recipient combination and the context in which the activation takes place.

As has become clear in the preceding sections, for T cells to become activated they need to recognize antigen. Every T cells bears a recognition structure, the T-cell receptor (**TCR**). The majority of T cells in the peripheral lymphoid organs and peripheral blood express a TCR comprising an a and b chain—the recognition structure—that is associated with a complex of polypeptides which form the signalling moiety known as CD3 (Fig. 5). abTCRs can recognize MHC–peptide complexes with exquisite specificity, each being specific for one MHC–peptide complex. Once recognition has taken place, signals are delivered to the intracellular machinery by CD3.

At this stage in the process the cell membrane in the vicinity of the TCR–CD3 complex becomes very active and reorganization of the molecules in the membrane occurs to form an immunological synapse. This results in all of the elements required for productive T-cell activation being brought into close proximity with the TCR, including the accessory molecules, CD4 or CD8, and molecules required for the delivery of costimulation or second signals to the T cell, CD28 and CD154. Other structures that are important for adhesion of the antigen-presenting cell and the T cell localize to the edges of the synapse, thus ensuring that the two remain in close contact with one another for long enough for information to be transmitted in both directions.

The localization of CD4 or CD8 in the immunological synapse brings them into close proximity with the TCR–MHC–peptide complex. CD4 is expressed by T cells that recognize MHC class II–peptide complexes (class II-restricted T cells) and CD8 by T cells that recognize MHC class I–peptide complexes (class I-restricted T cells). Each of these molecules can interact with conserved elements of the class II or class I structure, respectively, and they fulfil both an adhesion and signalling function when antigen recognition occurs.

In addition to signals coming through the TCR–CD3 complex and accessory molecules—also known as signal 1—additional signals arising from other cell surface receptors are required to ensure that the responding T cell is activated. These additional signals are often referred to as signal 2 or costimulation. In the presence of signal 1 but the absence of signal 2, T cells become unresponsive or anergic and fail to proliferate in response to further signals from antigen-presenting cells. Thus, during the initial phase of activation it is important that the antigen is presented by a professional antigen-presenting cell that can provide costimulation in addition to presenting donor antigen, either as the intact molecule or as recipient MHC–donor peptide complexes.

Costimulation is a complex process involving many cell surface structures. In strict terms costimulatory molecules can be defined as those that are essential for the initiation of a response from naïve T cells. The best characterized of these on the T-cell side is a molecule known as CD28 (Fig. 5). This is expressed by naïve T cells at rest, interacts with two cell surface ligands on antigen-presenting cells (CD80 and CD86), and is reported to have a preferential interaction with CD86 at the initiation of the response. CD86 is expressed at low levels by immature antigen-presenting cells, but upregulated rapidly during maturation of the antigen-presenting cell and following contact with T cells. By contrast, CD80 is expressed at lower levels than CD86 at the beginning of the response, but once expressed can also interact with CD28. The current interpretation of these data suggests that CD86 is more important for interaction with CD28 during the initiation of the response and that CD80 participates more actively in the downregulation of the response by preferentially interacting with another T-cell molecule, CD152 or CTLA4 (see below).

Signals delivered through CD28 result in increased cytokine synthesis by the responding T cell resulting from the stabilization of cytokine mRNA species. Signals through CD28 are independent of those delivered through the TCDR–CD3 complex and can be blocked independently by different immunosuppressive drugs. However, when the two signalling pathways occur in the same context the signals are integrated by the responding T cell, resulting in an augmented response. The complex series of phosphorylation and dephosphorylation events that take place results in the production of transcription factors, including nuclear factor of activated T cells (NF-AT), that translocate the nucleus of the T cells and switch on transcription of genes such as that for IL-2.

Regulation of immune responses is always critically important: dysregulated immune responses can have very dramatic and harmful consequences for the host. A pathway that counterbalances the positive signals coming through CD28 therefore exists to ensure that the process of T-cell activation does not continue indefinitely in an uncontrolled manner. Later during the course of T-cell activation a new molecule, CD152, is expressed by the activated T cells and acts as a negative regulator of the response. Evidence for this has been obtained by analysing mice that have a targeted disruption in the CD152 gene, so-called CD152 knockout mice. These mice have uncontrolled T-cell expansion when they are housed under normal environmental conditions where they are exposed continuously to a wide variety of antigenic stimuli. CD152 has been shown to have a higher binding affinity for CD86 and CD80 than CD28. Once it is expressed by the activated T cell it can therefore compete for binding with these molecules on the antigen-presenting cell. In addition, the interaction of CD152 with CD80 has been shown to deliver a negative signal to the T cell, shutting down further clonal expansion.

The construction of a fusion protein from the extracellular domains of CD152, CTLA4Ig or CTLA4Fc, and its use as a therapeutic agent has provided evidence that blocking costimulation through CD28 is sufficient to inhibit graft rejection, and confirming that T-cell costimulation through this pathway is a critical step in the

activation steps of the rejection response.

Following the initial stages of T-cell activation, CD4+ T cells also express another cell surface molecule, CD154 or CD40L, that can provide additional costimulatory signals for the responding cell. CD154 interacts with its ligand CD40, which is expressed by antigen-presenting cells, including dendritic cells, B cells, and monocytes (Fig. 5). Non-haematopoietic cells can also express CD40, including endothelial cells, fibroblasts, and epithelial cells. Interestingly, signalling through this pathway is a two-way event, not only leading to modification of the functional capabilities of the T cell but also those of the antigen-presenting cell. Thus, signalling through CD40 results in the augmented expression of CD86 and CD80 by antigen-presenting cells, potentially setting up an amplification loop for augmenting the response. For example, in the kidney, tubular epithelial cells have been shown to express CD40 and engagement by CD154 results in the increased production of chemokines, including IL-8 and Rantes.

CD40 and CD154 are members of the TNF and TNF receptor families, respectively. They utilize different signalling molecules to both the TCR–CD3 and CD28 pathways. Blockade of CD154 by a monoclonal antibody has been shown, either in combination with CTLA4Ig or at high doses alone, to prevent acute allograft rejection and lead to long-term rejection-free survival of vascularized as well as non-vascularized grafts. This pathway is therefore also critical for the early events in T-cell activation. Although CD4+ T cells are highly dependent on it for activation, evidence is emerging that activation of CD8+ T cells is much less dependent upon or independent of the CD154–CD40 pathway. Thus in some donor–recipient combinations rejection can still be initiated by CD8+ T cells even in the presence of high doses of anti-CD154.

Determining the character of the rejection response

The character of the downstream response is critically dependent on the context in which the initial activation and restimulation of donor antigen-specific T cells takes place. Once activated, T cells recruit other cells into the response and play a role in determining how these differentiate into effector cells. The antigen-presenting cells involved, the cytokine environment, and the immune status of the host will all have a marked influence on the downstream response. Following chronic antigen stimulation, such as will occur with time after transplantation, a marked divergence in cytokine production by the responding cells can take place. This was first noted following chronic stimulation of antigen-specific mouse T cells *in vitro*, but has subsequently been demonstrated to occur in humans as well.

T-cell activation in the presence of IL-12 has been shown to result in the differentiation of T cells that secrete IFN-g and IL-2. By contrast, if the initial contact between the T cell and antigen takes place in the presence of IL-4, then the cell will differentiate along a different pathway and secrete IL-4, IL-5, and IL-10, so-called signature cytokines (Fig. 6). The two types of T cell have been referred to as T_{H1} and T_{H2} or T_{C1} and T_{C2} depending on whether they express CD4 or CD8. T cells secreting IFN-g and IL-2 orchestrate cell-mediated immunity, resulting in the activation of cytotoxic T cells (T_C) and macrophages predominantly, whereas T cells secreting IL-4 and IL-10 trigger the differentiation of B cells into plasma cells producing certain isotypes of immunoglobulin (humoral immunity) and the activation of eosinophils. Both types of T-cell response have been shown to lead to rejection. Therefore the hypothesis that a T_{H1} response is aggressive and results in rejection whereas a T_{H2} response promotes tolerance—the T_{H1}–T_{H2} paradigm for transplantation—is not clear cut and the context in which the response evolves will have a marked influence on whether rejection occurs.

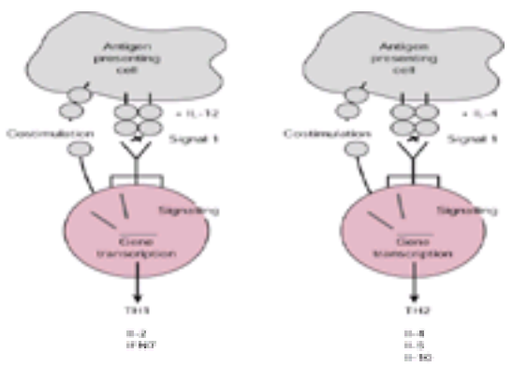

Fig. 6 T-cell differentiation. The microenvironment present when a T cell is activated will have a marked influence on the way in which it differentiates upon restimulation with antigen. T cells that are activated in the presence of IL-12 will differentiate into cells that secrete the signature cytokines IFN-g and IL-2, whereas T cells that encounter antigen in the presence of IL-4 will differentiate into cells producing IL-4, IL-5, and IL-10.

Migration of activated leucocytes into the graft

Once activated, the donor-specific leucocytes must migrate from the recipient lymphoid tissue back to the graft if they are going to be effective in destroying the transplanted tissue. To enter the graft the leucocytes have to cross the donor vascular endothelium. Leucocyte extravasation is a multistep process, controlled by the production of chemokines by the transplanted tissue and multiple interactions between cell surface molecules expressed by the endothelium and the migrating leucocytes. Chemokine receptors involved in the recruitment of leucocytes into tissues are only expressed at low levels on resting leucocytes, therefore activation of the different leucocyte populations that participate in either the innate or adaptive immune response to the graft is a key step in the recruitment process.

Characterization of the chemokines and chemokine receptors involved in recruiting leucocytes to the graft is currently in progress. Chemokines produced within the tissue can be tethered to endothelial cells by interaction with carbohydrate structures on the endothelial cell surface as well as secreted from the tissue. Thus leucocytes flowing in the blood through the vessel can become attracted to the endothelial cells as a result of chemotatic gradients being established from the vessel wall into the tissue. Cell–cell interactions between the leucocytes and the endothelial cells are initiated such that the leucocytes flowing past the tissue in the blood are taken out of the flow and begin to roll along the vessel endothelium. As the leucocytes roll they sample the environment of the endothelial cells. If there is nothing wrong with the endothelial cell surface, then the leucocytes detach and return to the blood flow. However, in the presence of 'danger' signals the leucocytes express new cell surface structures, including P-selectin. The rolling leucocytes then becomes tethered to the endothelial cell. As a result of interaction between additional families of cell surface molecules, both on the endothelial cell and the leucocytes, including integrins and immunoglobulin superfamily members, firm adhesion of the leucocytes to the endothelial cell surface occurs, allowing the cells to transmigrate between endothelial cell junctions into the tissue along the chemokine gradient.

Graft destruction

Unlike some immune responses, for example to certain viruses, where a single effector mechanism dominates the final stages of the process, for allograft rejection the immune system uses many strategies to destroy the graft (Fig. 1). Once the vascular endothelium of the graft has been damaged by one or more of the mechanisms outlined below, the blood supply to the graft will be lost and rapid necrosis of the transplanted tissue will occur. Later in the rejection process the parenchymal cells of the graft will also become targets for these destructive mechanisms.

Antibody

Alloantibodies have been shown to play a role in hyperacute, acute, and chronic rejection. As mentioned above, hyperacute rejection of allografts is very rarely seen in clinical transplantation as rigorous screening of recipients for antidonor reactive antibodies is carried out before transplantation to eliminate any patients who have preformed antibodies against the donor. Hyperacute rejection of xenografts is the first immunological barrier that needs to be overcome if xenotransplantation is to be successful in the future. Different approaches are being investigated with varying degrees of success. Understanding how antibodies can trigger the destruction of a graft is clearly one of the important pieces of information required to facilitate the design of effective strategies to prevent antibody-mediated damage.

Antibodies that react with the graft can trigger its destruction in two ways: by activating complement or through antibody-dependent cellular cytotoxicity via killer cells.

Complement is a cascade of proteases that are triggered sequentially following the initial activating event. The system can be activated when antidonor antibodies formed as a result of T-dependent B-cell activation (Fig. 1) bind to donor antigens. As a result, inflammatory mediators will be released, increasing the vascular

permeability of vessels in the graft and thereby facilitating the migration of leucocytes into the graft. The graft will become coated with antibody and activated complement components, targeting donor cells for opsonization by phagocytic cells which express receptors both for complement components and antibody. The membrane attack complex of the complement system is then formed, resulting in the lysis of donor cells. Many of the pathological changes that are associated with acute rejection, such as arteriosclerosis, interstitial haemorrhage, and fibrinoid necrosis of arteriolar cell walls, may result from the binding of antibodies and complement activation.

Antibody-dependent cellular cytotoxicity is cell dependent and occurs when the antibodies act as a bridge between the graft and killer cells, activating their lytic machinery. Killer cells are heterogeneous and many different types of leucocyte can participate in antibody-dependent cellular cytotoxicity when they are present in the correct microenvironment.

Although the appearance of antidonor antibodies can trigger rejection, their appearance does not necessarily mark the rejection of the graft. Indeed, the presence of antidonor antibodies may be perfectly compatible with continued graft survival. The specificity and the effector properties of the antibodies produced hold the key to whether particular antidonor antibodies are destructive.

Donor-specific cytotoxic T cells

Donor-specific cytotoxic T cells (**TC**) mature from precursor T_C (**pTC**) following activation of donor-specific T-helper cells (**TH**) (Fig. 1). T_C are activated either as a result of the formation of a three-cell cluster with the helper cell and the antigen-presenting cell, or as a result of the activated T_H cell 'licensing' the antigen-presenting cell to activate the pT_C. Once mature, effector T_C exhibit potent, antigen-specific cytotoxic activity. Their cytotoxic activity arises through a variety of mechanisms: these include the release of proteases called granzymes; the deposition of perforins, proteins that punch holes in the membrane of the target cell; the triggering of Fas-dependent cytotoxicity; and the release of soluble molecules such as tumour necrosis factor (TNF). The exact mechanism that is used *in vivo* may vary depending on the conditions that prevail within the graft.

There is considerable evidence to suggest that T_C can be involved in graft rejection. Most convincingly, when CD8+ cells are eliminated from the recipient before transplantation (most T_C recognize donor class I molecules and express CD8), graft rejection is often delayed or prevented. However, the presence of T_C is not mandatory for rejection as in some circumstances this has been shown to occur in the absence of demonstrable T_C activity. Moreover, the demonstration that donor-specific T_C activity can be detected *ex vivo* is not a guarantee that rejection is taking place. Again, the precise microenvironment in the graft markedly influences the ability of the effector cells to elicit graft destruction.

Natural killer cells

Natural killer (**NK**) cells form part of the innate immune response and are a potent source of cellular cytotoxicity. They are only triggered to kill when certain non-polymorphic MHC class I molecules, HLA E molecules, are missing from the target cells. In other words, NK cells are not used to destroy normal cells of the host unless they have been modified such that they no longer express HLA E molecules.

The receptors involved in NK-cell activation have been characterized and under normal circumstances these receive both positive and negative signals from both activating and inhibitory receptors when they engage their ligands on the target cell. Only when the inhibitory signals are missing do the cells exhibit cytotoxic activity. The role of NK cells in the rejection of solid organ allografts is still uncertain. NK cells with the ability to kill target cells *ex vivo* can be found in rejecting allografts, but to date there has been no direct demonstration that they play a role in rejection. By contrast, NK cells have been shown to be capable of rejecting bone marrow cells that express very low levels of MHC class I molecules and are thought to be very important in the rejection of xenografts where the graft will express no human class I.

Macrophages, eosinophils, and cytokine release

When T cells are activated they can elicit a non-specific effector mechanism referred to as a delayed-type hypersensitivity reaction (**DTH**). DTH reactions are characterized by the infiltration of leucocytes, including lymphocytes, macrophages, and eosinophils into the target site, in this case the graft. Damage to the graft occurs as a result of the production of non-specific mediators, such as nitric oxide, reactive oxygen species, IL-1, and TNF-a by the infiltrating cells. This activity is triggered in an antigen-specific manner by the T_H cell, but the effector mechanisms that lead to the destruction of the graft are non-specific. DTH reactions have been shown to be capable of playing a role in acute and chronic allograft rejection.

Conclusion

The immune response to a transplant is complex. The precise nature of the response will depend on many factors: the donor–recipient incompatibility, the type of graft, the site of transplantation, and not least the cocktail of immunosuppressive drugs that are used to try and prevent or control the response.

Further reading

Bach F *et al*. (1995). Barriers to xenotransplantation. *Nature Medicine* **1**, 869–73.

Banchereau J, Steinman R (1998). Dendritic cells and the control of immunity. *Nature* **392**, 245–52.

Brent L (1997). *A history of transplantation*. Academic Press, San Diego.

Cyster J (1999). Chemokines and cell migration in secondary lymphoid organs. *Science* **286**, 2098–102.

Ginns L, Cosimi A, Morris P, eds. (1999). *Transplantation*. Blackwell Science, Oxford.

Gould D, Auchincloss H (1999). Direct and indirect recognition: the role of MHC antigens in graft rejection. *Immunology Today* **20**, 77–82.

Matzinger P (1994). Tolerance, danger and the extended family. *Annual Reviews of Immunology* **12**, 991–1045.

Medzhitov R, Janeway C (2000). Innate immune recognition: mechanisms and pathways. *Immunological Reviews* **173**, 89–97.

Transplantation Websites

The Eurotransplant Foundation. _ HYPERLINK http://www.eurotransplant.org/ __http://www.eurotransplant.org/_

Anthony Nolan Bone Marrow Trust. _ HYPERLINK http://www.anthonynolan.com/ __http://www.anthonynolan.com/_

United Network of Organ Sharing. _ HYPERLINK http://www.unos.org/ __http://www.unos.org/_

6.1 Epidemiology of cancer

R. Doll and R. Peto

Introduction

All cancers have certain pathological and clinical characteristics in common, but those arising in different organs often have very different causes. The epidemiology of cancer, by which is meant the study of the incidence of the disease in people under different conditions of life, is, therefore the epidemiology of specific types of cancer, usually, but not always, defined as cancers of specific organs. In this sense, the subject has a history dating back nearly 300 years to Ramazzini's observation that cancer of the breast occurred more often in nuns than in other women of similar age and to Pott's observation, 200 years ago, that scrotal cancer in young men occurred characteristically in chimney sweeps. The high risk in nuns (which largely reflected the protective effect of multiple pregnancies in the general population) helped the realization that hormonal factors can substantially affect the incidence of several types of cancer, while the latter led to the recognition that the combustion products of coal to which sweeps had been exposed could cause cancer on any part of the skin with which they came into repeated contact and to the isolation of the first specific chemical carcinogen. Many other similar observations were made over the next 150 years, mostly as a result of the acumen of individual doctors who were struck by the observation that clusters of cases of a particular type of cancer occurred in patients with a similar occupational or cultural background. Lip and tongue cancers were found in pipe smokers, bladder cancer in certain aniline dye workers, buccal cancer in those who habitually chewed mixtures of tobacco and betel in India, lung cancer in miners of particular ores (who, it was subsequently realized, were heavily exposed to radon and its daughter products), and skin cancer in the early radiologists and radiographers who were heavily exposed to X-rays and in farmers and seamen heavily exposed to sunlight. Gradually, however, clinical anecdotes were replaced by statistics as the epidemiological methods that are described below began to be applied to the study of cancer and other non-infectious disease. As a result, many other causes were identified with sufficient certainty to justify preventive action and data were obtained to suggest hypotheses that could be tested in the laboratory.

Preventability of cancer

Perhaps the most important result of such observations has been the realization that any type of cancer that is common in one population is rare in some other, and that the differences between populations are mostly not genetic. Hence, where they are common these cancers occur, in large part, as a result of the way people behave and the circumstances in which they live and they are, therefore, at least in principle, preventable. This does not mean that we can at present envisage a society in which any of the common cancers are completely eliminated (although this may prove to be possible when we understand more clearly the mechanisms by which they are produced). What it does mean is that we can envisage a society in which the age-specific risk of developing any type of cancer is low.

Differences in incidence between communities

Reliable evidence of variation in the incidence of particular types of cancer between different communities was slow to emerge because of differences in the standards of medical care, and hence in the extent to which any cancers are diagnosed, the absence of population-wide systems for the registration of any cases that were diagnosed, and differences in the reliability with which cases were reported when registration systems were established. Nowadays, however, the large differences that are reported between good cancer registries throughout the world are, for the most part, real, particularly if comparisons are restricted to the limited range of ages between 35 and 64 years. This excludes the youngest ages, at which cancer is rare, and the oldest ages, at which the records of the incidence of the disease are least reliable.

Table 1 shows, for selected types of cancer, the range of variation recorded by cancer registries that have produced data sufficiently reliable for the purpose of international comparison (International Agency for Research on Cancer, 1992) or, in a few instances, the range determined by special surveys. Types of cancer have been included if they are common enough somewhere to have a cumulative incidence among men or women of at least 1 per cent by 75 years of age. The ranges of variation shown are for incidence rates between 35 and 64 years (see above). The range of variation is never less than seven-fold and is sometimes more than a hundred-fold. Despite the selection of reasonably reliable registries, some of this tabulated variation may still be an artefact, due to different standards of medical service, case registration, and population enumeration; but in many cases the true ranges will be greater. First, there are still gaps in the cancer map of the world, so that some extreme figures may have been omitted because no accurate surveys have been practicable in the least developed areas and it is just these areas that are

likely to provide the biggest contrasts (both high and low) with Western society, as Chen *et al.* (1990) have shown in rural China. Secondly, the figures cited refer, with one exception, to cancers of whole organs and do not distinguish between different histological types or different locations within an organ and the more one learns about each type of cancer the more disadvantageous this is found to be. It is obvious in the case of skin cancer, which includes melanomas that have increased in incidence dramatically in the last 50 years, basal-cell carcinomas of the face, which affect more than half the fair-skinned population of Queensland by 75 years of age, scar epitheliomas of the leg, which develop on the site of old ulcers in some African populations and account for 10 to 20 per cent of all cancers seen in some hospitals in Malawi and Rwanda Burundi, '*dhoti*' cancers of the groin in India, and occupational cancers on the forearm due to exposure to tar and oil in industrialized countries. But it also applies, to a greater or lesser extent, to most of the cancers listed in Table 1.

The variation in incidence is not limited to the common cancers, but is also shown by many others. Burkitt's lymphoma, for example, never affects more than 1 in 1000 of the population, but it is at least 100 times as common among children in parts of Uganda as it is in Europe and North America; while Kaposi's sarcoma, which was extremely rare in most of the world until the advent of the acquired immunodeficiency syndrome (AIDS), was so common in children and young adults in parts of Central Africa, even before 1970, that it accounted for 10 per cent of all tumours seen in one of the African hospitals surveyed by Cook and Burkitt. Some few cancers occur with approximately the same frequency in all communities; but all of these are uncommon. Acute myeloid leukaemia at 15 to 25 years of age is an example; nephroblastoma is another, except that it appears to be only half as common in Japan as elsewhere.

The figures that have been cited so far all refer to the incidence of cancer in different communities defined by the area in which they live. Communities, can, however, be defined in other ways and no matter what method is used, including categorization by ethnic origin, religion, or socioeconomic status, substantial differences may be found. Jewesses, for example, have a low incidence of cervical cancer irrespective of the country in which they live, and the Mormons of Utah and the Seventh Day Adventists of California suffer fewer cancers of the respiratory, gastrointestinal, and genital systems than members of other religious groups living in the same American states.

Few of the large differences observed between communities can be explained by genetic factors, apart from some of the differences observed in the incidence of cancer of the skin, the risk of which is much greater for whites than blacks, and possibly also for some of those in the incidence of testis cancer, which rarely affects black populations, and in the incidence of chronic lymphocytic leukaemia, which rarely affects people of Chinese or Japanese descent. Genetic factors cannot explain the differences observed on migration or with the passage of time, which are discussed below, nor can they explain the correlations observed between the national rates for particular types of cancer and some measures of the lifestyle of the different countries.

Changes in incidence in migrant groups

That changes in the incidence of cancer occur on migration is certain. Many groups have been studied, including Indians who went to Fiji and South Africa, Britons who went to Fiji and Australia, and Central Europeans who went to North America. Among the most reliable data are those for the black Africans whose ancestors were taken to America and the Japanese who went to Hawaii. The former now experience incidence rates for internal cancers that are generally much more like those of white Americans than those of the black populations in West Africa from which most of their ancestors came, while the latter have experienced rates that are much more like those of the Caucasian residents in Hawaii than those of the Japanese still living in Japan (Table 2). The ancestors of black Americans and Hawaiian Japanese will have come from many different parts of West Africa and Japan, some of which are likely to have cancer rates somewhat different from those that have been cited in Table 2. Nevertheless, the contrasts are so great that there can be no serious doubt that new factors were introduced with migration.

Changes in incidence over time

Within one population there may be substantial changes in the incidence of a particular type of cancer over a period of a few decades that provide conclusive evidence of the existence of preventable factors. Changes in incidence over time may, however, be difficult to assess reliably, chiefly because it is difficult to compare the thoroughness of the selection and registration of particular types of cancer at different periods and partly because few incidence data have been collected for long enough, so we often have to fall back on changes in mortality rates even though these may be influenced by changes in treatment as well as by changes in incidence.

There are no simple rules for deciding which of the many changes in recorded cancer incidence and mortality rates are reliable indicators of real changes in incidence. Each set of data has to be assessed individually. It is relatively easy to be sure about changes in the incidence of cancer of the oesophagus, as the disease can be diagnosed without complex investigations and its occurrence is nearly always recorded, at least in middle age, because it is nearly always fatal. It is much more difficult to be sure about changes in the incidence of many other types. The common basal-cell carcinomas of the skin, for example, are also easy to diagnose, but they seldom cause death and can be treated effectively outside hospital, so they often escape registration. What appears to be a change in incidence may, therefore, be a change only in the completeness of registration. Cancers of the pancreas, liver, and brain, and myelomatosis, in contrast, usually cause death, but even when they do they may be misdiagnosed as another disease (for example brain tumours in old people could frequently in the past be misdiagnosed as other neurological conditions), so that an increased incidence or mortality rate may be wholly or partly due to improvements in diagnosis, in the availability of the medical services, or in the readiness of physicians to inform cancer registries of the cancers they find. Such changes are particularly likely to affect the rates recorded for people over 65 years of age, as many old people who were terminally ill used not to be intensively investigated.

Despite these difficulties, some of the decreases and increases in the recorded rates of particular types of cancer have been so gross that there must have been real changes in their incidence. Examples include the increase in oesophageal cancer in the black population of South Africa, the increase in lung cancer throughout most of the world (and its recent large decrease in men in the United Kingdom), the increase in mesothelioma of the pleura in men in industrialized countries, the decrease in cancer of the tongue in the United Kingdom, and the decrease in cancers of the cervix uteri and stomach throughout western Europe, North America, and Australasia. For a fuller account see *Trends in the incidence of cancer*, Doll *et al.*, 1994.

Identification of causes

Finally, it has been possible to obtain evidence of the preventability of cancer by defining agents or circumstances that are a cause of the disease and are capable of control. In general, reliable evidence of causality (and particularly of the magnitude of any risks) has to come from epidemiology and not from laboratory experiments, although the latter can often provide reinforcement of epidemiological findings and essential guidance or completely novel hypotheses for epidemiological study. Reliable epidemiological evidence does not require randomized trials within particular populations, but it does require the study of different individuals within populations and not just the comparison of incidence rates in different populations. Non-randomized epidemiological studies of individuals have often yielded proof of causation beyond reasonable doubt (like that required to convict in a court of law). Action based on such evidence has, moreover, often been followed by the desired result—for example a reduction in the incidence of bladder cancer in the chemical industry on stopping the manufacture and use of 2-naphthylamine and, on a national scale, the reduction in the incidence of lung cancer in men in the United Kingdom following the decrease in smoking over the previous half century. Cancer research workers have, therefore, accepted that the type of human evidence that has been obtained (often, but not invariably, combined with laboratory evidence that the suspected agents are carcinogenic in animals) is strong enough to conclude that a cause of human cancer has been identified and that, as a corollary, the disease can be prevented if this cause is controlled.

Biological factors

Speculations about the causes of cancer and the mechanisms that lead to its occurrence have been constrained by some of its biological characteristics. These include the relationships between incidence and genetic susceptibility, age, sex, and the delay (which is sometimes misleadingly called the 'latent period') that occurs between exposure to a causative agent and the appearance of clinical disease.

Genetic susceptibility

Genetic differences in susceptibility are discussed in Chapter 6.3. We note here only the role of epidemiology in (i) detecting familial clusters that are so marked that no statistical analysis is needed to show the reality of their existence, or (ii) demonstrating by large studies that if one member of a family develops a specific type of cancer, other members are somewhat more likely to develop that same type than would be expected in the population as a whole.

The first has shown that several rare genes have such a great effect on susceptibility that bearers of one such gene (if it is dominant) or two (if they are recessive) almost invariably develop a particular type of cancer. Examples include the dominant genes for polyposis coli and Gardner's syndrome that lead to cancer of the large bowel, and the recessive genes for retinoblastoma and xeroderma pigmentosum that lead (in the latter case) to squamous carcinoma and (less commonly) melanoma of the skin. Similar evidence has shown that other genetic syndromes frequently, but not invariably, lead to cancer, such as von Recklinghausen's neurofibromatosis

leading to fibrosarcoma, the Peutz–Jeghers syndrome leading to carcinoma of the small bowel, the Wiskott–Aldrich syndrome leading to non-Hodgkin's lymphoma, and ataxia telangiectasia, Bloom's syndrome, and Fanconi's anaemia leading to leukaemia. The recognition of these syndromes is important to the individual, as it may provide an opportunity for prophylactic surgery, or enable the diagnosis of malignancy to be made at an early stage when treatment is more likely to be effective, or (rarely) enable precautions to be taken to prevent exposure to the relevant carcinogens, as in the case of sufferers from xeroderma pigmentosum or albinism, who can be protected against sunlight. The proportion of all cancers that occur in people who are highly susceptible to cancer in this way is, however, very small.

The second sort of epidemiological evidence has shown that there is no material tendency for cancer as a whole to cluster in families and that there are no common genetic polymorphisms that substantially increase the risk of developing cancer in all organs (although mutations in the p53 gene may increase the risk in many). It has also shown, however, that several of the common types of cancer do tend to cluster in families to some extent. Differences of this sort do not necessarily imply that the familial clusters are genetic in origin; they could be due to familial similarities of behaviour. Nor, however, do they necessarily imply that any genetic difference in susceptibility is particularly small. Calculations show that they are compatible with 50- to a 100-fold differences in genetic susceptibility if the genes for high susceptibility have an appropriate prevalence in the population. That socially important genetic variants exist is demonstrated by the greatly increased risk of developing basal-cell and squamous carcinomas of the sun-exposed skin in fair-skinned populations compared with dark-skinned, and there may be other genes associated with localized populations, which, for example, diminish the risk of chronic lymphatic leukaemia and myelomatosis in Chinese, Japanese, and Indians. Other genes may have only a minor effect, such as the gene for blood group A, the possession of which increases the risk of gastric cancer by about 20 per cent over that of people belonging to blood groups O or B.

Discovery of genetic factors that affect particular types of cancer is unlikely to explain much of the social and geographical differences in the distribution of cancer other than skin cancer, but it should help to elucidate mechanisms and in extreme cases may help to focus health education and costly methods of early diagnosis on the sections of the populations that are most at risk.

Age

Some risk of cancer occurs at every age, but the risk of developing any particular type varies with age. The most common relationship with age is a progressive increase in incidence from near zero in childhood and adolescence to a high rate in old age. This type of relationship is shown by carcinomas of the skin, lung, and gastrointestinal and urinary tracts, and by myelomatosis and chronic lymphatic leukaemia. The rate of increase is rapid, being typically proportional to the fourth, fifth, or sixth power of age in years, so that the annual incidence may be 100 or 1000 times greater above age 75 than before age 25. With most of these cancers, the recorded incidence may stabilize, or even decrease, in the oldest age groups; but this is partly or wholly an artefact due to incomplete investigation of the terminal illnesses of old people. This pattern is observed for skin carcinoma due to exposure to ultraviolet light and for bronchial carcinoma, both in non-smokers and in men who regularly smoke a constant number of cigarettes a day, and can, under certain circumstances, be observed in the laboratory in skin-painting experiments on mice. It is probable that it reflects the cumulative effect of processes that operate steadily throughout life, starting at around the time of birth (or, for lung cancer among habitual smokers, in adolescence).

A less common pattern is a peak incidence early in life, which may be followed either by a decline virtually to zero or by a slow rise in middle and old age. Retinoblastomas and nephroblastomas occur only in childhood, with peak incidences (respectively) in the first and second years of life. Teratomas and seminomas of the testis have peak incidence rates at about 20 and 30 years of age, respectively, and later almost cease to occur, while osteogenic sarcomas have a peak incidence in adolescence and then show a slow increase with age from a lower rate in young adult life.

The remaining cancers show a variety of patterns. Carcinomas of the breast and cervix uteri of women, for example, begin to appear in adolescence and become rapidly more common up to the menopause. After the menopause the incidence of carcinoma of the breast may remain approximately constant, or may even become slightly reduced for a few years, before increasing again with age, though at a slower rate. Carcinoma of the cervix continues to increase fairly steeply for a few years after the menopause, before showing a stable or declining rate. Hodgkin's disease, on the other hand, appears in childhood and then continues to occur more or less evenly throughout life with only minor peaks in young adult life and in old age, while connective tissue sarcomas become progressively more common from childhood on, but with a much slower rate of increase than is shown by the common carcinomas.

Some of these relationships with age, like that for retinoblastoma in early childhood, seem to be invariant everywhere and, as far as is known, at all times. Others vary from community to community, or from time to time. In postmenopausal women, for example, cancer of the breast becomes progressively less common with increasing age in parts of Asia, but more common in Europe, while carcinoma of the lung used to show a peak incidence at about 60 years of age in the United Kingdom, which gradually moved to older ages, as a generation that had not smoked substantial numbers of cigarettes throughout adult life was replaced by one that had, and the same process is now being repeated in many developing countries.

These various patterns provide information, either about the period of activity of the stem cells from which the cancers derive, or about the period when the main exposure to causative agents occurs and the duration of that exposure. Some of this variation has already helped to explain some of the causes of cancer, as was the case with the shift in the peak incidence of bronchial carcinoma; but much of it still awaits elucidation.

Sex

Cancer used to be more common in women than in men in many countries due to the great frequency of carcinoma of the breast and of the cervix uteri and to the rarity of bronchial carcinoma, and this is still the case in populations for which similar conditions persist, as in parts of Latin America. Elsewhere, cancer is now more common in men, among whom lung cancer often predominates. This overall male preponderance hides, however, a wide range of sex ratios for cancer of different organs. If the sites of cancer that are peculiar (or almost peculiar) to one sex are ignored, the sex ratio varies (in Britain) from a male excess of about 6 to 1 for pleural mesothelioma and carcinoma of the larynx, through many types of cancer with only a small male preponderance, to carcinomas of the right side of the colon, thyroid, and gallbladder, which may be up to twice as common in women.

For many types of cancer the sex ratio is much the same in different countries and at different times. For some, however, and particularly for cancers of the mouth, oesophagus, larynx, and bronchus, the sex ratio is extremely variable—not only between countries and at different times, but sometimes also between different ages at the same time and in the same country. The most marked variation is shown by cancer of the oesophagus, which may affect both sexes equally or be 20 times more common in men than in women. As with the various patterns of incidence with age, these different sex ratios and their variation can provide useful clues to the causation of the particular type of cancer, not all of which have yet been successfully followed up.

Delay between cause and effect

One reason why it has been difficult to recognize causes of cancer in humans is the long delay that characteristically occurs between the start of exposure to a carcinogen and the appearance of the clinical disease. This 'latent period', as it is commonly, but rather misleadingly, called is often several decades, although it may be as short as 1 year or as long as 60. The exact relation between the date of exposure and the date of the appearance of different cancers is still uncertain, partly because the interval is subject to random factors, partly because few cancers are induced by a single, brief exposure, and partly because there are still very few sets of quantitative data with detailed information about the dates when exposure began and ended.

When cancer is induced by short but intensive exposure to ionizing radiation, as following the explosions of the atomic bombs in Hiroshima and Nagasaki or in patients treated by radiotherapy, the excess incidence of solid tumours rises for 15 to 20 years and then may continue to rise, level off, or decline. In the case of acute leukaemia, however, a peak incidence occurs much earlier (about 5 years after irradiation) and relatively few cases appear after more than 30 years.

Short, intensive exposure to a carcinogen is, however, exceptional. The more usual situation is for sporadic or continuous exposure to a carcinogen to be prolonged for years—a decade or two in the case of occupational exposure, several decades in the case of tobacco smoking, and a lifetime in the case of ultraviolet sunlight. In this situation the incidence of cancer increases progressively with the length of exposure. In the last two cases cited, the incidence appears to increase approximately in proportion to the fourth power of the duration of exposure so that the effect after (say) 40 years is more than 10 times as great as that after 20 years, and more than 100 times as great as that after 10 years. Whether the same holds for occupational exposure is not known; but it has been shown to hold in some experiments in which chemicals were repeatedly applied to the skin of genetically similar mice and it may prove to be a general biological rule for many types of carcinoma and many carcinogens.

There is still less quantitative information about what usually happens when exposure ceases; but in the case of cigarette smoking the rapidly rising annual risk

among those who continue to smoke stabilizes for one or two decades after smoking ceases before increasing again slowly. The exsmoker consequently avoids the enormous progressive increase in risk suffered by the continuing smoker.

These delayed effects accord with the idea that the appearance of clinical cancer is the end-result of a multistage process in which several mutations have to be produced in a single stem cell to turn it into the seed of a growing cancer. From the practical point of view, the important conclusions are that cancer may be very much more likely to occur after prolonged exposure to a carcinogen than after short exposure, that it is seldom likely to appear within a decade after first exposure (except in the case of leukaemia and the specific cancers of childhood), that it commonly occurs several decades after first exposure, and that some excess risk may continue to occur for decades after exposure has ceased. The exact relationship may, however, differ for different carcinogens and different types of tumour. Bladder tumours, for example, began to appear within 5 years of intensive exposure to 2-naphthylamine in the dye industry, while mesotheliomas of the pleura have seldom, if ever, appeared within 10 years of exposure to asbestos, but they continue to increase in incidence for up to 50 years after first exposure, even if the exposure was relatively brief.

Luck

There remains the influence of luck, which is commonly ignored; yet it is important for the individual as it is the reason why two animals of identical genetic constitution that have been treated in the same way do not, in general, develop cancer in the same place at precisely the same age. It reflects the element of chance that determines whether a particular series of events all occur in one particular stem cell out of the many thousands of stem cells that exist that don't give rise to a malignant clone. For any one individual the role of good or bad luck in determining the occurrence of cancer may be large (just as luck plays a substantial part in whether or not an individual driver has a traffic incident); but in a large population luck has little net effect on the incidence of cancer and only nature and nurture are important.

Avoidable factors

Tobacco

Tobacco is by far the most important single cause of cancer in developed countries. Chewed it can cause cancers of the mouth and oesophagus; smoked it is a major cause of cancers of the mouth, pharynx (other than nasopharynx), oesophagus, larynx, lung, pancreas, renal pelvis, and bladder. For these eight cancers, epidemiological evidence indicates that prolonged smoking of average numbers of cigarettes per day increases the risk 3 to 20 times. It is, however, now clear that cigarette smoking also causes a proportion of several other types of cancer, increasing the incidence up to twice that in non-smokers: namely, cancers of the lip, nose, nasopharynx, stomach, liver, and renal body and also myeloid leukaemia. Although the proportional increases are not large, the consistency of the findings in different countries, the evidence of dose–response relationships, the lower mortality in exsmokers than in continuing smokers, the lack of evidence for important confounding, and the presence in the smoke of many different carcinogens provide strong grounds for believing that most or all of these observed associations are causal.

In sum, smoking is estimated to have caused 30 per cent of all fatal cancers in the United Kingdom in 1995, down from 34 per cent 20 years earlier. The reduction was substantial in men (down from 52 per cent to 40 per cent) but it was largely counteracted by the increase in women (from 12 per cent to 20 per cent). Comparable figures from the United States and from some other developed countries are shown in Table 3. In men, there have been decreases in some developed countries, but increases in others, particularly in Central and Eastern Europe. In women, the proportion of cancer deaths attributed to smoking was generally low in 1975, but has subsequently increased in all developed countries and must be expected to increase further. It was, however, still small in countries such as France, where few middle-aged or elderly women had been smoking for long enough for any material effect to be produced.

In developing countries, the effects of smoking have only recently begun to be studied systematically and much remains unclear. In general, women in developing countries do not smoke (or if they do they smoke very little). In men, however, there has been a very large increase in cigarette consumption, the full effects of which have yet to materialize. China, with 20 per cent of the world's population, smokes 30 per cent of the world's cigarettes and by 1990 smoking was already responsible for about 20 per cent of male cancer deaths. In India, where many men have smoked 'bidis' (small home-manufactured cigarettes) for decades, the proportion may be even greater (chiefly because smoking can act as a cofactor for the production of cancers of the mouth, oesophagus, or stomach in those who habitually chew quids containing betel and tobacco). In some parts of South America, the male lung cancer rates from smoking are already as high as in developed countries. Overall, tobacco may be causing about as many cancer deaths in developing as in developed countries, in which case it would be responsible for about 20 per cent of cancer deaths throughout the world.

Alcohol

At least six types of cancer are caused in part by the consumption of alcohol. One, liver cancer, is produced only secondarily to the production of liver cirrhosis and is, consequently, caused only by heavy and prolonged consumption. Four are causally related to smoking as well as to alcohol: namely, cancers of the mouth, pharynx (other than nasopharynx), oesophagus, and larynx. The two agents act synergistically, increasing each other's effect, so that the risk from alcohol in non-smokers or long-term exsmokers is very small, while that in heavy smokers is disproportionately large. The remaining type, cancer of the breast, has been shown to be related to alcohol only within the last decade or so. Cohort studies show that the risk increases progressively with the amount drunk (at least up to moderately high levels) and laboratory studies that show that alcohol increases the level of oestrogen in the blood suggest a plausible mechanism.

Cancers of the large bowel have also been associated with alcohol in many studies, but the relationship is weak and its nature uncertain: it could be due to confounding with smoking and diet.

Ionizing radiations

Ionizing radiations, of whatever sort, share the characteristic of being able to penetrate animal tissues and damage DNA. It is not surprising, therefore, that they have been found to increase the incidence of cancer in practically every organ. It has not been possible to detect by direct observation the effect of the small amounts that adults receive as a result of exposure to (for example) radiological examination, atmospheric pollution, and normal levels of natural background; but it has been possible to make an estimate of their effect by extrapolating from the observed effects of the much larger doses received by the survivors of the atomic explosions at Hiroshima and Nagasaki, patients given radiotherapy or repeatedly screened radiologically, and people exposed occupationally to radium or to high concentrations of radon in mines or exceptionally in houses. Theoretical considerations and the dose–response relationship observed with these relatively large doses both indicate that there is unlikely to be any threshold below which no effect is produced. This conclusion is reinforced by the discovery that children who received doses of 10 to 20 mGy *in utero* (because their mothers were irradiated for diagnostic purposes whilst they were pregnant) were subject to an added risk of developing cancer in childhood of approximately 1 in 2000. At low doses (less than about 20 mGy) it seems probable that the carcinogenic effect is linearly proportional to the dose; at higher doses the same is true for most cancers other than leukaemia, for which the risk is approximately proportional to the square of the dose. It is unlikely, however, that we should be far out in our estimate if we accepted the conclusions of the International Committee on Radiological Protection (1991) and assumed that the lifetime risk of developing a fatal cancer is approximately 10 per cent per Gray (or per Sievert) to the whole body if the radiation dose is moderate and given acutely and about half that if the dose is low and spread out over time (that is 5 per 100 000 per mGy (or mSv)) with corresponding reductions if only part of the body is exposed.

People are exposed to different amounts of radiation in different countries, depending principally on the build up of radon in the air in domestic houses and the medical use of radiation for diagnosis and therapy. In the United Kingdom, the average annual dose is about 2.6 mSv, which, in a population of about 55 million, is estimated to cause about 7000 deaths a year from cancer, about 5 per cent of the total. In the United States, the average annual dose is about 50 per cent greater. The estimated hazard depends critically on the effects of chronic exposure to radon in houses, which, in the United Kingdom, contributes about half the total dose from all forms of radiation. In some parts of the country, however, most notably Devon and Cornwall, the average domestic radon dose is three or four times greater and in a few houses may be 10 or even 100 times greater. Its effect is discussed later under lung cancer. Nearly all the rest comes from other sources of natural radiation (35 per cent) and medical uses (14 per cent). The last, however, causes less cancer than would be deduced from the dose, as a large proportion is received by old or ill people who will not survive for long enough for a radiation-induced cancer to appear or because the doses given radiotherapeutically are so large that most of the cells that might have been made cancerous are destroyed. Less than 0.5 per cent of the average annual dose from all sources can be attributed to occupational exposure, fall-out from past bomb tests, and man-made products or radioactive waste.

Ultraviolet light

Photon energies in the ultraviolet (UV) range are sufficient to excite electrons in atoms to chemically active higher energy states permitting the formation of pyrimidine dimers between adjacent pyrimidine bases in DNA and these may, as a result of misrepair, be the origin of mutations. UV does not penetrate much below the skin, so that it is chiefly within the skin that it is directly carcinogenic. Within the skin, however, it is the principal cause of all types of cancer, other than Kaposi's carcinoma. Whether it has any indirect carcinogenic effect on other tissues (notably the lymphopoietic tissue) by (for example) destroying Langerhans cells and so modifying immune reactions, has yet to be proved.

Infection

Infection, principally viral, but also in some cases bacterial and parasitic, is a major cause of avoidable cancer.

Viral

Viruses that are known to cause human cancers, or suspected of doing so, are listed in Table 4, along with the types of cancer with which they are associated. Not all infected people develop the disease. In some cases the proportion doing so is quite small, unless other factors are also present. These include heavy malarial infection for Burkitt's lymphoma, the consumption of a type of salted fish for nasopharyngeal cancer, and the consumption of aflatoxin, a metabolic product of fungal infection with *Aspergillus flavus,* for liver cancer. What they are for the cancers produced by the human papilloma virus is not known.

Quantitatively, chronic infection with hepatitis B virus is one of the most important causes of cancer in many parts of the world. In China, for example, liver cancer accounts for about 20 per cent of all cancer deaths, the large majority of which are due to chronic lifelong infection with the virus. Infant vaccination against the virus is now being introduced and will protect the new generation, but will not provide retrospective protection for those born previously.

Bacterial

Only one bacterial infection has been closely linked with the development of cancer: namely, *Helicobacter pylori.* Persistent infection acquired early in life leads to chronic gastritis in the antrum of the stomach and increases the risk of gastric cancer two to three-fold. Non-specific chronic infection in the bladder may also increase the risk of bladder cancer by the local formation of carcinogenic nitrosamines.

Parasitic

In parts of Africa and Asia, parasitic infection is a major cause of cancer. Infestation with *Schistosoma haematobium*, which excretes its eggs through the bladder wall, causes a high incidence of bladder cancer in Egypt and East Africa while infestation with *Schistosoma japonicum*, which excretes its eggs through the wall of the large bowel, is responsible for a high incidence of intestinal cancer in parts of China. Liver flukes (*Clonorchis sinensis* and *Opisthorcis viverrini*) are similarly responsible for the high incidence of cholangiosarcoma of the bile ducts in parts of South East Asia. The parasites may not cause cancer directly but chronic infection may start a chain of events that leads to cancer in other ways, such as chronic bacterial infection and the local formation of nitrites and nitrosamines.

Medical drugs

Apart from ionizing radiations, some 20 agents have been used therapeutically that are known to cause cancer in humans. These are listed in Table 5. That so many carcinogenic agents should have been prescribed medically is not surprising when it is borne in mind that treatment often requires modification of cellular metabolism and is sometimes intended to interfere with DNA. The hazard of cancer, however, need not necessarily be a bar to the use of a drug if the risk to life due to iatrogenic cancer is materially less than the chance of saving life that is achieved by its use—as is commonly the case with antineoplastic agents, immunosuppressive drugs, and radiotherapy.

Some of the chemotherapeutic agents listed in Table 5 were soon abandoned, while others have continued to be used for the treatment of uncommon conditions, and the sum of the cancers that these now produce cannot amount to more than a 100 or so a year in the United Kingdom.

Three of the listed drugs are, however, used extensively: hormonal replacement therapy (HRT) for postmenopausal women, selected steroids for contraception, and tamoxifen for the treatment of hormone-sensitive breast cancer. The first two increase the risk of breast cancer and all three can increase the risk of endometrial cancer, but HRT does so substantially only when given in the form of oestrogen alone and steroid contraceptives do so only in the form (now abandoned) in which oestrogen and progestogen are given sequentially. The combined steroid contraceptives currently in use can also rarely cause liver cancer and they may possibly increase the risk of cervix cancer. In contrast to these effects, tamoxifen reduces the incidence of breast cancer, and combined steroid contraceptives reduce the incidence of endometrial cancer and halve the risk of ovarian cancer for many years after they have been used. HRT and the combined steroid contraceptive are, moreover, associated with a reduction of some 20 per cent in the risk of colorectal cancer, but whether this is causally related to their use remains to be proved.

Other drugs that may inhibit cancer rather than cause it are the non-steroidal analgesics, most notably aspirin, the prolonged use of which may somewhat reduce the risk of colorectal cancer.

Taken altogether it seems unlikely that medically prescribed drugs can be responsible for more than 1 per cent of all today's fatal cancers and may, in total, reduce the risk by somewhat more.

Occupation

In the years that followed Pott's observation that chimney sweeps tended to develop cancer of the scrotum, many other groups of workers were found to suffer from specific hazards of cancer and more substances that are known to be carcinogenic to humans have been unearthed by the search for occupational hazards than by any other means. These hazards, many of which are described in relation to individual types of cancer, are listed in Table 6. Many of the hazards that have been recognized caused large, or at least relatively large, risks, albeit for limited populations, and it may well be that other occupational hazards exist that have not yet been detected, either because the added risk is small in comparison with that due to other causes, or because only a few workers have been persistently exposed, or simply because the hazards have not been suspected and so not looked for. It must also be borne in mind that cancer in humans seldom develops until one or more decades after exposure to the carcinogen first occurs and it is, therefore, too soon to be sure whether agents that have been introduced into industry only during the last 20 years are carcinogenic or not.

Many groups of workers not listed in Table 6 have been suspected of having a special risk, but it has not been possible to decide whether the risk is real and attributable to their work. Many types of cancer have been examined in these groups and, in these circumstances, some differences that are conventionally 'statistically significant' are bound to have arisen by chance alone. Such differences can provide substantial evidence of a hazard only if highly significant ($p<0.001$), or if excess rates are confirmed in other studies, or if a risk of the specific type of cancer could be predicted from the nature of the agent to which they were exposed.

Other excess rates may be due to confounding; that is, they may be produced by social factors that are associated with the occupation in question rather than by the occupation itself. The potential importance of such confounding was illustrated by Fox and Adelstein's analysis of the occupational mortality statistics for England and Wales over the period 1970 to 72. They aggregated the occupations of men aged 15 to 64 years into 25 major categories and found that the lung cancer rates in these categories differed up to two-fold, a spread that was far too wide to be attributed to chance alone. However, data obtained from random samples of the general population showed that the proportions of men who smoked in each of these 25 large occupational categories varied from about 65 per cent of the national average to about 130 per cent, and that the differences in these proportions could account for most of the variation in the mortality from lung cancer.

Given sufficient details and the ability to repeat the observations, it is usually possible to obtain a fairly clear idea of whether or not an excess incidence in an occupational group does or does not reflect an occupational hazard by (for example) seeing whether the effect is related to the length of employment, the time after first exposure, and a specific type of work within the industry. Unfortunately these details are not always available and the reasons for many of the moderate excesses

of cancer that have been reported in certain industries are still uncertain.

At present it seems unlikely that occupational hazards account for more than 2 or 3 per cent of all fatal cancers in developed countries such as the United Kingdom; but the quantitative evidence is uncertain and this estimate could be out by a factor of two. The three principal causes are probably asbestos dust (lung and pleural cancer), the combustion products of fossil fuels (skin and lung cancer), and ionizing radiations (any cancer).

Pollution

The idea that pollution might be an important cause of cancer has been in the forefront of the minds of cancer research workers since it was realized that the incidence of lung cancer tended to be higher in towns than in the countryside and that the combustion products of coal, which used to produce a pall of smoke over all large cities in Britain, contained carcinogenic hydrocarbons. Subsequently, with the rapid expansion of the chemical industry and the discovery that some of its products are mutagenic *in vitro* and carcinogenic in laboratory animals, anxiety increased about the possible effects of distributing such products ubiquitously in the air we breathe, the water we drink, and the food we eat.

The effects of pollution of this sort are, however, peculiarly difficult to assess directly by epidemiological methods, as pollutants are likely to be present in most areas, the absolute risk from each is likely to be small, and there may be little difference in the extent to which individuals are exposed over a wide area. Reliance is, therefore, often placed on two indirect methods: extrapolation from the effects of chronic exposure to much larger amounts in an occupational setting, and prediction of the effects on humans from laboratory tests. Both, however, (but particularly the latter) involve substantial uncertainties.

So far as atmospheric pollution is concerned, the epidemiological picture is complicated by the personal pollution produced by tobacco smoke and the social distribution of smoking habits; but, despite this complication, the various methods that have been discussed under lung cancer all lead to the conclusion that the pollution of the past may have contributed to the production of a few per cent of all lung cancers; but that the levels over the last three decades (principally from the combustion of fossil fuels, but also from asbestos, dioxins, and various other materials) are unlikely to be responsible for more than a fraction of 1 per cent of future cancers—although there may be exceptions awaiting discovery in the neighbourhood of particular factories. The greater effect of the modern type of pollution with ultra fine particles and of the intense indoor pollution with smoke that occurs in parts of China is examined later under lung cancer.

The effect of polluted drinking water and food is more obscure. Until recently no serious consideration had been given to the possibility that either might be important, except for the possible effect of the contamination of food with smoke from urban air. Now, however, analytical techniques permit the detection of chemicals at concentrations of less than 1 p.p.b. in both food and water and, in consequence, many have been detected that might arguably be carcinogenic, including pesticide residues and a variety of halogenated organic materials produced by the chlorination of water supplies. Relationships have been reported between the concentrations of some of these compounds in water and the mortality from cancers of the bladder and, possibly, the large intestine, in different localities; but it is extremely difficult to know what these relationships mean as there are many potentially confounding factors.

Mortality rates from cancers of the gastrointestinal and urinary tracts are, for the most part, stable or decreasing in early middle age, when the effects of new agents might be expected to show themselves first, and, in the absence of more specific evidence, it seems unlikely that chemical pollution of water and food could have a greater effect than the small effect already estimated for pollution of the air. It is important, however, to monitor the situation and, in particular, to seek an explanation for any increase in incidence (as has occurred with testis cancer and with non-Hodgkin's lymphoma not attributable to AIDS) and for any irregularities in the geographical distribution of any type of the disease.

Diet

For many years there has been suggestive evidence that most of the cancers that are currently common could be made less so by modification of the diet; but, with few exceptions, there is still little reliable evidence as to the modifications that would be of major importance. If we define diet to include all materials that occur in natural foods, are produced during the processes of storage, cooking, and digestion, or are added as preservatives or to give food colour, flavour, and consistency, the ways in which diet could influence the development of cancer are legion.

Ingestion of preformed carcinogens

The most obvious is the ingestion of small amounts of powerful carcinogens or precarcinogens. Several have been identified in foodstuffs but only two have been related at all clearly to the production of cancer in humans: that is aflatoxin, a metabolic product of *Aspergillus flavus*, which contaminates stored or oily foods in many countries, and is a major cause of liver cancer in the tropics among those individuals who are also chronic carriers of the hepatitis B (or less commonly hepatitis C) virus. Likewise, the salted fish eaten extensively in South China, probably acts synergistically with EBV to cause nasopharyngeal cancer. A third possible source is bracken fern, an extract of which is carcinogenic in animals. It is eaten extensively in Japan and has been tentatively linked with the development of oesophageal cancer. The polycyclic hydrocarbons and other mutagens that are produced in food by grilling or smoking have often been suspected of playing a role, but intensive investigation has failed to detect one.

It seems, therefore, that if diet does affect the incidence of cancer in the Western world in any material way, it is likely to do so by more indirect means, such as those described in Table 7.

Overnutrition

That overnutrition could affect the incidence of cancer was first suggested by Tannenbaum's experiments on mice during the Second World War. These showed that the incidence of spontaneous tumours of the lung and breast and of a variety of tumours produced experimentally could be halved by moderately restricting the intake of food without modifying the proportions of the individual constituents. This protective effect has subsequently been demonstrated repeatedly, but has attracted little attention (perhaps because reports of such results emphasized the benefits of restriction rather than the harm of overeating). It is now clear, however, that what is considered normal nutrition in developed countries increases the risk of breast cancer (by bringing forward menarche and increasing body size) and possibly also that of testis cancer. With greater consumption obesity (that is a BMI greater than 25 kg/m ^{2}) has been estimated to be responsible for 5 per cent of all incident cases in Europe and 10 per cent of all cancer deaths in non-smokers in the United States (Peto, 2001): most notably cancer of the breast in women after the menopause and cancers of the endometrium, large bowel, pancreas, gallbladder, prostate, and kidney and myelomatosis. For some of these increases, the explanation is obvious: namely, those of the two female cancers, which in postmenopausal women are attributable to the formation of oestrogen from androstenedione in adipose tissue while the increased risk of gallbladder cancer may be due to a greater secretion of bile salts. For others, the explanation is obscure.

Meat and fat

Figures for food consumption and cancer incidence and mortality rates in different countries show fairly close correlations between the consumption of fat and, to a less extent, the consumption of meat and the incidence of several types of cancer. The correlations are closest for breast cancer and cancer of the large bowel and are less strong for cancers of the endometrium, pancreas, and prostate. When, however, attempts are made to associate the consumption of either type of food with the disease in individuals within a country the evidence is commonly conflicting. This could be because the international correlations are misleading, indicating only that the risks are correlated with something that is correlated with fat and meat consumption (for example some other aspect of a high gross national product), but it could be partly because of the inaccuracy of dietary histories and partly because people in developed countries, and particularly in North America, eat such similar diets. Overviews of the published data, however, do suggest that a high consumption of fat is associated with a high risk of colorectal cancer, but the claim that a high consumption of fat (or of particular types of fat) is associated with high risks of breast and endometrial cancer after the menopause, other than by providing a high calorie diet leading to obesity, is controversial.

Whether meat increases the risk of any type of cancer, apart from the contribution it makes through its calorie content, is also uncertain. The low incidence of several types of cancer that is commonly observed in vegetarian communities is not necessarily due to the absence of meat from the diet, as it can generally be explained by the increased consumption of protective foods (vegetables and fruits) and commonly by associated behavioural characteristics (below average use of tobacco and alcohol). Some studies that make allowance for these confounding factors have claimed that meat specifically increases the risk of large-bowel cancer, but the evidence is weak and the increase in risk, if any, is small.

Fibre

That fibre may play a part was suggested by Burkitt's observation that several intestinal diseases, including cancer of the colon, were common in countries in which cereals were processed to remove the fibre and rare in rural Africa and Asia where they were not. The idea was attractive, as 'fibre' passes through the small bowel unchanged and serves as pabulum for the colonic bacteria, thus increasing faecal bulk and possibly protecting against the development of cancer mechanically by diluting any carcinogens present and hastening their transit through the bowel. The idea was, however, too simple and has not been confirmed (using the original definition of fibre) by either epidemiological studies on individuals in developed countries or by experiments aimed at reducing the recurrence of colorectal adenomas. In fact, fibre is difficult to define and the term is better replaced by 'non-starch polysaccharides' as there are many that share the characteristics of passing through the small bowel unchanged and being, for the most part, partially or wholly degraded by bacteria in the large bowel. Some starch, moreover, known as 'resistant starch' and found in green bananas and cold potatoes, has similar physiological characteristics. Further studies that take these complexities into account are, therefore, needed before 'fibre' in any of its manifestations can be considered as having any place in protecting against the development of cancer.

Retinoids and carotenoids

Experiments on animals and on cell cultures *in vitro* have suggested that vitamin A (retinol) and its esters and analogues (retinoids) may, in appropriate circumstances, reduce the risk of cancer by reducing the probability that partially transformed cells become fully transformed and proliferate into clinically detectable tumours, although in other circumstances they appear to have opposite effects. Human studies, however, fail to support the idea that serum levels of retinol are related to the risk of any type of cancer, at least in countries in which clinical symptoms of vitamin A deficiency seldom or never occur. Such studies suggested that the risks were inversely related to the serum level of b-carotene, which acts as an antioxidant and is broken down to produce retinol. When, however, b-carotene was put to the test of clinical trials it provided no benefit and the inverse relationship commonly observed in epidemiological studies is presumably due to confounding with some other protective factors in vegetables.

Other components

Many other components of the diet, including lycopene in tomatoes, indoles in cabbages and sprouts, vitamins C, D, and E, and calcium and selenium have also been proposed as protective agents; while nitrates, nitrites, secondary amines, and the preservation of food by salting, have been thought to increase the risk of cancer. For some the evidence is strongly suggestive: notably for vitamin C as protective against gastric cancer and for salt-preserved foods predisposing to it. In general, however, the evidence of benefit or harm is too weak to justify any firm conclusion.

Conclusion

Some of the uncertainties about the effect of diet could be resolved only by means of controlled trials in which volunteers are allocated at random micronutrient supplements (such as vitamin C, lycopene, calcium, and selenium) or a dietary schedule that requires a substantial reduction in the consumption of fat. Several such studies are under way and if any such factors affect the later stages of cancer induction, clear answers may be obtained to some of the outstanding questions within a few years. Practicable modifications of the diet may well provide the means for reducing cancer deaths in developed countries by a third; but the range of uncertainty about this figure is large. Meanwhile the only dietary changes that can be recommended with confidence in developed countries are a general increase in the use of fresh fruit and vegetables and a sufficient limitation of calories to avoid obesity.

Reproduction and other factors affecting secretion of reproductive hormones

Epidemiological observations have shown clear relationships between a woman's reproductive history and the risk of cancers of the sex organs, which are generally thought to reflect changes in hormonal secretions; but which hormones are concerned and the mechanisms by which they act are, for the most part, still uncertain. An exception is endometrial cancer, the risk of which is directly related to the degree of exposure to oestrogen not followed after an appropriate interval by progestogen. Proof that oestrogenic stimulation of the mammary tissues is a cause of most cases of breast cancer in developed countries has been provided by randomized trials of tamoxifen, an antioestrogenic drug that blocks the oestrogen receptors in the cells of the normal breast. The effect is large and rapid: 5 years of tamoxifen approximately halves the incidence of breast cancer not only while the drug is being taken but also for some years afterwards. Exogenous oestrogen also increases the risk of breast cancer when given as hormonal replacement therapy and endogenous oestrogen accounts for the increased risk associated with adiposity after the menopause, as androstenedione, which continues to be produced by the adrenals, is converted to oestrogen in adipose tissue. It is presumably oestrogens, too, that cause a small increase in risk of breast cancer during and immediately after pregnancy and the oestrogen component of the steroid contraceptives that causes a similar small increase in risk during their use and for a few years after their use is stopped. It is, however, unclear which hormone-related processes are involved in reducing the long-term risk for the rest of a woman's life that occurs some years after the occurrence of each pregnancy and it is equally unclear why the use of oral contraception and the consequent suppression of ovulation reduces the long-term risk of ovarian cancer.

Hormones, it is thought, must also be involved in producing cancers of the testis and prostate in men, but the epidemiological evidence is, as yet, unhelpful. Randomized trials of the effects of physical or medical castration in men who already have prostate cancer have, however, shown that progression of the disease can be slowed substantially, presumably by the reduction of androgenic stimulation.

Physical inactivity

Physical inactivity contributes to the risk of cancer indirectly by increasing the risk of obesity but it may also contribute directly to the risks of cancers of the colon and breast. An association with colon cancer has been consistently reported, which may be due to a reduction in colonic mobility with corresponding prolonged exposure of the colonic mucosa to faecal carcinogens. An association with breast cancer may simply be due to confounding, but it may also reflect an effect on hormone secretion.

Interaction of agents

Attribution of the risk of cancer to different causes is complicated by the fact that some agents interact with others to produce effects that are much greater than the sum of the separate effects of each on its own. An example is provided by smoking and asbestos, which multiply each other's effects so that, compared with non-smokers in general, the incidence of cancer of the lung was increased six-fold among a group of asbestos insulation workers in the United States who did not smoke, but were heavily exposed to asbestos dust in the 1940s, 10- to 20-fold among cigarette smokers in general who did not work with asbestos, and nearly 90-fold among such asbestos insulation workers who also smoked cigarettes regularly. Other examples are provided by smoking and radon (which interact similarly, though somewhat less than multiplicatively, to produce cancer of the lung), by smoking and alcohol (which interact to produce cancers of the mouth, pharynx, larynx, and oesophagus), and by infection with the hepatitis B virus and aflatoxin (which interact to produce cancer of the liver).

Such interactions may come about through some analogue of the dual mechanism of initiation and promotion that was suggested 60 years ago by the experiments of Rous and Kidd and Berenblum and Shubik. They could, however, also be produced in many other ways, some of which have been discussed in the section on diet (see above). From a statistical point of view, they complicate the attribution of risk, as we may find ourselves appearing to claim that more cancer can be prevented than actually occurs by attributing, say, 80 per cent of lung cancers in men heavily exposed to asbestos to their occupational exposure and 90 per cent of the same cancers to cigarette smoking.

Conclusion

Estimates of the proportions of fatal cancers that can be attributed to environmental and behavioural factors are given in Table 8, along with the proportions that are now known to be avoidable in practicable ways. For this purpose, individual causes (such as the various carcinogenic drugs and occupational hazards) have been grouped into 12 main categories. The evidence on which these estimates are based is summarized in this chapter and for earlier periods in greater detail by Doll and Peto (1981) and the International Agency for Research on Cancer (1990).

The sum of the best estimates in Table 8 amounts to less than 100 per cent, despite the fact that some of the listed agents interact with one another to augment each other's effect and that some fatal cancers are consequently counted twice. The total would be somewhat more than 100 per cent, however, if the true proportions attributable to some of the categories turn out to be nearer the upper end of the acceptable estimates and it would be a great deal more than 100 per cent if it had

been possible to characterize the factors that have been classified as 'other and unknown'.

The estimates in columns two and three of Table 8 do not distinguish between factors (such as tobacco) that are sufficiently understood to enable specific action to be taken with a guarantee of success and those (such as diet) that are not. They should not, therefore, be taken as guides to the proportion of cancer deaths that can now be prevented by practicable means, without reference to the paragraphs in which they are individually discussed. This is illustrated by the fourth column in Table 8, which shows the proportions of United Kingdom cancer deaths in 1995 that are reliably known to be avoidable by practicable means. The only percentage that is both large and reliably known is that attributed to tobacco and that is more than twice the sum of the percentages reliably attributable to other specific factors for which practicable preventive measures are available: and tobacco causes about twice as many deaths from other diseases as it does from cancer. The position is different in countries such as China, where hepatitis B virus causes about as many cancer deaths as tobacco and the hazard for future generations can be avoided in a cost-effective way by infant vaccination.

Epidemiology of cancer by site of origin

In the following account of the epidemiology of cancers arising in specific organs, the description of each type is preceded by notes showing its importance in England and Wales. One figure gives the proportion of all cancers that arise in the site, as indicated by the national cancer register for England and Wales for 1994 (Office of National Statistics, 2000) and another gives the proportion of all cancer deaths allocated to the site in the national mortality statistics for 1998 (Office of National Statistics, 1999). A third gives the ratio of the age-standardized incidence rates for each sex. The way in which the incidence of the disease varies with age is shown for males and females in a series of graphs, using the data for England and Wales over a 5-year period (International Agency for Research on Cancer, 1992). Under 25 years the numbers of most types of cancer that occur in a 5-year age group, even in a 5-year calendar period, are small and the rates are consequently unreliable due to chance variation of small numbers. Trends in incidence and mortality for each type, along with the trends in possible causative factors, are given by Swerdlow *et al.* (2001). Major differences between Britain and other countries are commented on in the text and are described more fully in the report on *Cancer causes, occurrence and control* by the International Agency for Research on Cancer (1990).

Lip

- 0.1 per cent of all cancers and 0.02 per cent of cancer deaths.
- Sex ratio of rates 5.0 to 1. Age distribution like skin (non-melanoma).

Carcinoma of the lip was one of the first types of cancer to be related to an extrinsic cause when, more than 200 years ago, it was noted to occur characteristically in pipe smokers. Many years later it was realized that the disease could also be produced by smoking cigarettes, although much less readily, so that it must be produced by the chemicals in smoke rather than by the non-specific effect of local heat. It is also much more common in outdoor than in indoor workers and is induced by ultraviolet light in the same way as other cancers of the exposed skin. Ultraviolet light and tobacco account, between them, for the great majority of all cases in Britain, probably multiplying each other's effects. The disease is much less common than it used to be, because of the decrease in both pipe smoking and outdoor work.

Oral cavity and pharynx (excluding salivary glands and nasopharynx)

- 1.0 per cent of all cancers and 1.2 per cent of cancer deaths.
- Sex ratio of rates 2.3 to 1. Age distribution, see Fig. 1.

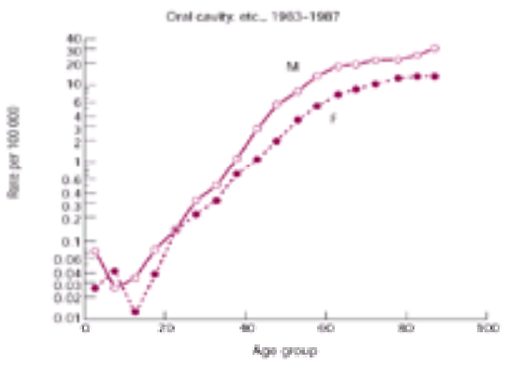

Fig. 1 Annual incidence of cancers of the oral cavity and pharynx by age and sex (excluding cancers of the salivary glands and nasopharynx).

Cancers of the tongue, mouth, and pharynx (other than nasopharynx) are all related to smoking (of pipes, cigars, and cigarettes) and to the consumption of alcohol. The two factors act synergistically and cancers in these sites are extremely rare in non-smokers who do not drink alcohol.

Cancer of the tongue is much less common in Britain than it was early this century, but the reason for the sharp decline in incidence is unknown. One explanation could be the decrease in syphilis, which was commonly believed to be a predisposing factor because of the clinical association with syphilitic leucoplakia. Recent increases in oral and pharangeal cancer in men are partly due to increased consumption of alcohol and possibly, in the case of pharyngeal cancer, to human papilloma virus infection.

Cancers that occur low in the hypopharynx are distinguished by a tendency to affect women who have suffered from iron-deficiency anaemia and dysphagia.

Cancers of the mouth and pharynx (excluding nasopharynx) are particularly common in South-East and Central Asia where tobacco smoking is largely replaced by chewing tobacco, betel nut or leaf, and lime (calcium hydroxide). A close association with such chewing habits has been established by studies that have shown that the cancers tend to originate in the part of the mouth in which the quid is usually held—a characteristic that varies both between individuals and between areas. The materials chewed differ in different places and, although the disease is commonly described as 'betel chewer's cancer', betel is not invariably a component of the quid and the most characteristic constituent seems to be a small amount of lime and, in most cases, some form of tobacco. In parts of Asia, the disease is so common that it accounts for 20 per cent of all cancers and in those populations the abandonment of chewing would be the single most effective means of reducing the total incidence of cancer—so long as the habit was not replaced by an increase in tobacco smoking. Among habitual quid chewers, the risks are particularly elevated in those who both chew and smoke—indeed, in parts of India the majority of deaths from 'betel chewer's cancer' could have been avoided if those affected had not also smoked. The incidence might also be reduced by improved nutrition, as the disease in Southern Asia tends to be associated with vitamin A deficiency.

In parts of India where women tend to smoke local cigars and cigarettes with the burning end inside the mouth to prevent them going out, the habit is associated with cancer of the palate.

Salivary glands

- 0.2 per cent of all cancers and 0.1 per cent of cancer deaths.
- Sex ratio of rates 1.3 to 1. Age distribution like non-Hodgkin's lymphoma.

The salivary glands are not common sites for cancer anywhere. They are, however, relatively more common in the Asiatic populations of Hawaii and in Canadian Indians than elsewhere. A small proportion of cases occurs specifically in families that also have a high incidence of breast cancer. No causative factors are known and no notable changes in incidence over time have been reported.

Nasopharynx

- 0.1 per cent of all cancers and of cancer deaths.

- Sex ratio of rates 2.0 to 1. Age distribution, see Fig. 2.

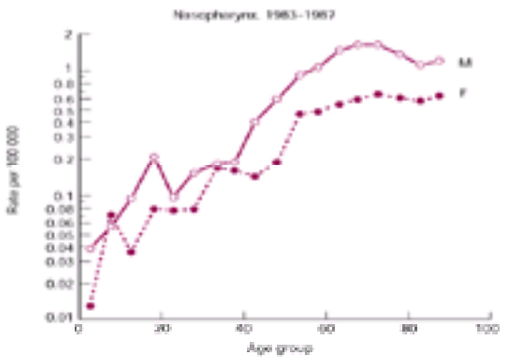

Fig. 2 Annual incidence of cancer of the nasopharynx by age and sex.

Cancers of the nasopharynx, unlike those in other parts of the pharynx, are not related to alcohol and are only weakly related to tobacco. They are rare in most populations but are common in those that originated from parts of Guangdong, in southern China, where the disease is the most common type of cancer. Moderately high rates have been observed in Alaskan Eskimos and American Indians, with intermediate rates in Malaysia, Kenya, and North Africa. A weakly significant relationship with HLA type has been reported from Singapore, but the existence of a specific genetic predisposition remains to be proved. Incidence rates appear to have been decreasing among Chinese Americans.

DNA characteristic of the Epstein–Barr virus (EBV) has been detected in the nuclei of nasopharyngeal cancer cells and patients with the disease tend to have unusually high antibodies against EBV-related antigens. Among adults, sudden increases in certain EBV antigens in the blood often precede by a few years the appearance of a pathological cancer. Infection with the EBV is, however, almost universal and can be only one of several agents that act in combination to produce the disease. One such agent in Southern China occurs in the 'salted fish' on which children are commonly weaned. This strongly flavoured delicacy bears little relation to the salted fish eaten elsewhere, and might better be described as decomposing fish: it contains various mutagens, and exposure to it in childhood when infection with EBV first occurs may alter the usual lifelong balance between host and virus in some hazardous way.

Oesophagus

- 2.3 per cent of all cancers and 4.4 per cent of cancer deaths.
- Sex ratio of rates 2.0 to 1. Age distribution like gastric cancer.

Cancer of the oesophagus, like other cancers of the upper respiratory and digestive tracts, is closely related to prolonged smoking and the consumption of alcohol. All types of smoking have comparable effects and, so it appears, do all alcoholic drinks, although spirits may be slightly more effective per gram of ethyl alcohol than other alcoholic drinks. Alcohol and tobacco act synergistically and, in the absence of either, the incidence of the disease in Britain would be greatly reduced. In France, where the consumption of alcohol is greater than in Britain, it would be reduced even more. A few cases originate from the scars produced by poisoning with corrosive substances and a very few in conjunction with a particular hereditary form of tylosis (presenting with keratoses of the palms and soles). The relatively small excess in men probably reflects the existence of other unknown causes in women, possibly nutritional in origin and similar to those responsible for cancers of the hypopharynx. Mortality (which, because of the high fatality rate, approaches incidence) fell progressively in the first half of the twentieth century in line with the fall in the consumption of alcohol, and rose again after 1950 when the trend in the consumption of alcohol reversed. Since pipe smoking affects oesophageal cancer risks at least as strongly as cigarette smoking, no large effects on male oesophageal cancer trends could be predicted from the male switch from pipes to cigarettes, although the switch by females from non-smoking to cigarettes should, other things being equal, produce a large upward trend. It appears, however, that other things were not equal and some other, possibly nutritional, cause of oesophageal cancer seems to have decreased, for any upward trends in oesophageal cancer are moderate. In men, in contrast, the rates have increased when they might have been expected to decrease. To some extent this can be accounted for by the increased consumption of alcohol and possibly by an increase in the nitrosamine content of tobacco smoke, which has resulted from changes in the method of curing tobacco and which could have a specific effect on the oesophagus. A small part of the increase is due to an increased risk of adenocarcinoma at the lower end of the oesophagus, which may be associated with a decreased prevalence of *Helicobacter pylori* and gastritis, and an increase in oesophageal reflux.

In Africa and Asia, the epidemiological features are quite different and present some of the most striking unsolved problems in the field of cancer epidemiology. In parts of China (particularly in North Henan but also elsewhere) and on the east coast of the Caspian Sea in Turkmenistan and Iran, oesophageal cancer is the most common type of cancer, with incidence rates in both sexes that are equal to the highest rates observed for lung cancer in men in European cities. Within China, the disease varies more than 10-fold from one county to another; alcohol and tobacco cannot account for these geographic differences, but when people within one particular Chinese county or city are compared with each other the disease is more common among those who smoke. In parts of Africa, particularly in the Transkei region of South Africa and on the east coast of Lake Victoria in Kenya, extremely high rates are also observed, sometimes equally in both sexes and sometimes only in men. In these and several other areas, as in Asia, the high incidence zones are strictly localized and the incidence falls off rapidly over distances of 200 or 300 miles.

When tobacco and alcohol are used, they increase the hazard, but they are not the principal agents in these high-incidence areas. Many causes have been proposed, including molybdenum deficiency in the soil (resulting in a deficiency of the plant enzyme nitrate reductase and a build-up of nitrosamines), contamination of food and pickled vegetables by fungi (particularly by species of *Fusaria*) with the production of carcinogenic metabolites, an agent associated with the production of beer from maize, and the residues left behind in pipes from smoking opium (which are commonly swallowed). None, however, is supported by any impressive epidemiological data. The high incidence area in Iran, which has been intensively investigated, is characterized by extreme poverty and a restricted diet consisting chiefly of home-made bread and tea, with some sheep's milk and milk products, and very little meat, vegetables, or fruit. In this area the disease has been common for centuries. In Southern Africa, however, it seems to have become common only since the First World War. In China, where cancer of the oesophagus was the second most important neoplastic cause of death in the 1970s, the high incidence has persisted.

Stomach

- 3.7 per cent of all cancers and 4.7 per cent of cancer deaths.
- Sex ratio of rates 2.5 to 1. Age distribution, see Fig. 3.

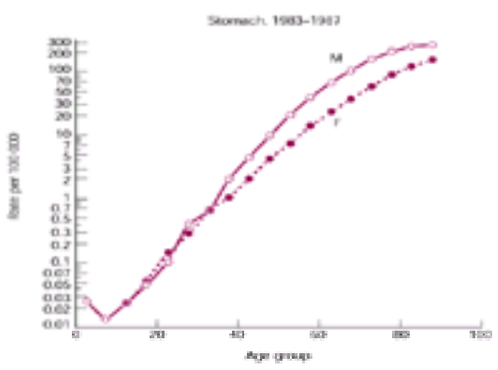

Fig. 3 Annual incidence of cancer of the stomach by age and sex.

Until about 1980, gastric cancer was responsible for more deaths from malignant disease world-wide than any other. Over the last 50 years, however, the incidence has declined in Western Europe, North America, and Australasia and recently it has begun to do so in South America and Japan. High rates are now confined to China, Japan, Russia and other countries of the old Soviet Union, and Central and South America, while the lowest rates are found equally in North America, Australasia, and some of the least developed parts of Africa. Irrespective of whether the incidence is high or low, the sex ratio is between 1.5 and 3 to 1.

Within Britain, cancer of the stomach is most common in North Wales and becomes progressively less common from north to south and from west to east. Over the last 70 years it has consistently been some five times more common in unskilled labourers than in members of the major professions, a gradient with socioeconomic status that has been one of the most marked for any disease. Relatively high rates have been observed in coal miners and in some chemical workers; but no specific occupational hazards have been identified and the excess in coal miners was paralleled by a similar excess in their wives. A hazard has been suspected from exposure to asbestos, but the apparent excess may be due to misdiagnosis of lung cancer and mesothelioma.

Four factors are known to predispose to the disease: blood group A constitution, gastritis associated with infection by *Helicobacter pylori* (sometimes leading to atrophic gastritis), a diet deficient in fruit and green and yellow vegetables, and a poor diet with large amounts of salt and salt-preserved food. Chronic infection with *H. pylori* is a major cause of peptic ulcer, a finding that is of considerable practical value in patients with ulcers, because the infection can generally be eliminated from the stomach by a short course of appropriate antibiotic therapy and this provides long-term protection against recurrence. Whether such treatments will have any material effect on the incidence of stomach cancer remains, however, to be shown. How these various factors influence the production of the disease is unclear. One possibility is that they encourage or discourage the formation of carcinogens *in vivo*, particularly perhaps the production of nitrosamines; but if they do, the intake of nitrates (which can be converted into nitrites by bacterial enzymes) is not a rate-limiting factor. Changes in the prevalence of the three environmental factors could have contributed to the decline in the incidence of the disease, but they could not have brought about such a large and widespread reduction in risk, and it seems probable that the better preservation of food, resulting from the extensive use of refrigeration, has played the major part.

No risk has been detected from the consumption of mutagens produced by the different methods of cooking meat and fish, nor from food additives or pesticide residues. Some food additives may, on the contrary, have served to reduce risk (by avoiding food spoilage and hence improving nutrition, by avoiding contamination by carcinogen-producing micro-organisms, or by some antioxidant or other protective effect on the gastric epithelium).

Large bowel

- 11.1 per cent of all cancers and 11.0 per cent of cancer deaths.
- Sex ratio of rates 1.4 to 1. Age distribution, see Fig. 4.

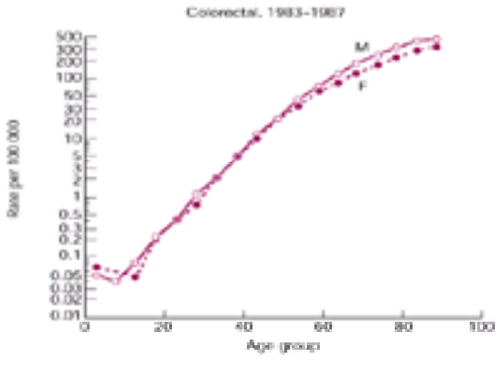

Fig. 4 Annual incidence of cancer of the large bowel by age and sex.

Cancers of the colon and rectum ought to be considered separately, as their causes are not identical. Cancer of the colon, for example, tends to occur more often in women than in men, particularly when it occurs on the right side, while cancer of the rectum is nearly twice as common in men. The geographical distribution also differs slightly, colonic cancer varying in incidence more than rectal cancer. Separate consideration may, however, sometimes be misleading as cancers commonly occur at the rectosigmoid junction and the site of origin of these cases is not recorded consistently. Moreover, there is a growing tendency to describe both diseases merely as 'cancers of the large bowel', which, according to the internationally agreed rules, are classed as cancers of the colon. The two diseases will, therefore, be considered together.

Over the past few decades the male incidence rates in both the United Kingdom and the United States have been approximately constant, while the female rates have decreased slightly. More recently, decreases in early middle age have begun to be seen in both sexes. In contrast, the incidence in Japan, which used to be very low, has begun to increase and the disease in Japanese migrants in Hawaii has become as common as in Caucasians. In most other parts of Asia, and in Africa and Eastern Europe, large-bowel cancer continues to be relatively uncommon (except in areas where chronic schistosomal infestation of the large intestine is prevalent; for example, high rectal cancer rates are found in Chinese counties in which *S. japonicum* was, until recently, a major cause of death). Incidence rates in different countries correlate closely with the per capita consumption of fat and meat and crudely with the consumption of processed foods from which the natural fibre has been removed. Ways in which these and other dietary constituents might influence the development of the disease have been discussed under diet. Other factors increasing risk are obesity and physical inactivity. A weak association with smoking has been observed in several cohort studies, which may be the result of confounding with the consumption of alcohol and a high fat diet. It is possible, however, that smoking may cause a few cases indirectly by causing the diet to be modified in the direction of a higher fat content.

Within Britain, there is no clear relation to socioeconomic status and no occupational hazard has been established. The association that has been reported with exposure to asbestos may be due to misdiagnosis as in the case of cancer of the stomach. Cases in childhood or early adult life occur as a complication of olyposis coli or (more rarely) Gardner's syndrome. These conditions are determined by dominant genes, which increase the susceptibility to the disease so much that it almost invariably develops before middle age. Many other cases develop from adenomatous polyps and a few occur as a complication of long-standing ulcerative colitis.

Anal intercourse causing infection with types 16, 18 or some other specific types of the human papilloma virus is a probable cause of some anal carcinomas, but patients who have sexually transmitted anal warts that are due to other types of human papilloma virus are not for this reason at special risk of anal cancer.

Liver

- 0.6 per cent of all cancers and 1.4 per cent of cancer deaths.
- Sex ratio of rates 2.1 to 1. Age distribution, see Fig. 5.

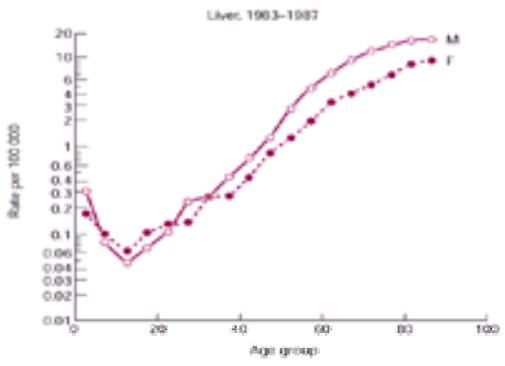

Fig. 5 Annual incidence of cancer of the liver by age and sex.

Incidence rates have tended to be overestimated in developed countries because the primary condition is often confused with metastases to the liver from cancer in various other organs, particularly over 65 years of age when carcinomas of the gastrointestinal and respiratory tracts are common. Recently, however, there has been a small increase in the United Kingdom and the United States from the very low level that had come to be recorded, which is probably due to an increased prevalence of infection with hepatitis C.

The disease is much more common in South-East Asia and tropical Africa; in China it accounts for about 20 per cent of all cancer deaths and in parts of Africa it is the most common cancer in men. Most cases derive from the main cells of the organ (hepatocellular carcinomas) and are attributable primarily to chronic active infection, established early in life, with the hepatitis B virus, exacerbated by consumption of some specific metabolite (e.g. aflatoxin) of particular types of fungi that contaminate stored foods. Neonatal vaccination against the virus produces a marked decrease in the proportion of children who, at 5 years of age, are chronically infected. This has begun in Japan, Taiwan, and parts of China and tropical Africa and has already produced a decreased risk of hepatocarcinoma at young ages. Some cases, however, are caused by the hepatitis C virus, which is an RNA virus spread by blood transfusion, and these cannot be avoided by immunization.

In developed countries, some cases are also due to infection with hepatitis B and C viruses, but more arise as complications of cirrhosis of the liver attributable to heavy and prolonged consumption of alcohol or, rarely, to haemochromatosis. Occasionally, liver cancer is produced by drugs. A few cases have occurred in young men who have taken androgenic-anabolic steroids to increase their muscular strength and a few from the use of steroid contraceptives, either arising *de novo* or from benign adenomas, which are themselves rare complications of the use of steroid contraceptives. Some can be attributed to smoking for an association has been observed in parts of China where little alcohol is drunk and case–control studies in Europe have shown an association after alcohol consumption has been taken into account.

A second histological type (cholangiosarcoma) arises from the intrahepatic bile ducts, tends to occur at a somewhat later age than hepatocellular carcinoma, and, although generally less common than hepatocellular carcinoma, nevertheless accounts for an appreciable proportion of cases. In China, Thailand, and other parts of Asia it can be produced by chronic infection with liver flukes (*Clonorchis sinensis* or *Opisthorchis viverrini*).

A third histological type that is extremely uncommon everywhere has been variously described as reticuloendothelioma or angiosarcoma. It was first recognized as a complication of the use of 'Thorotrast' as a contrast agent in neuroradiology, a long-abandoned practice that led to chronic retention of insoluble thorium radionuclides in the marrow, spleen, and liver. In 1973, the disease was found to be an occupational hazard for men exposed to vinyl chloride monomer. A few hundred cases have occurred throughout the world in men who were heavily exposed in the manufacture of vinyl chloride polymer, and it seems improbable that the minute amounts that have leached out of the plastic consumer products can have caused more than a dozen or so cases altogether in the general public, if indeed they have produced any. A third, and even rarer, cause is prolonged exposure to inorganic arsenic, such as used to result from the medical prescription of Fowler's solution. Despite these multiple causes only one case of hepatic angiosarcoma normally occurs annually in some 10 million people, which is why the recognition of new causes has been easy.

The relative rarity of cancer of the liver in most developed countries is intriguing, since most of the carcinogens thus far discovered in experimental animals induce, perhaps with other cancers, tumours of the liver. The lack of any high or increasing liver cancer rate in Britain and America consequently suggests that, on average, people have not been substantially exposed to the sort of chemical carcinogens that are currently recognized by such studies.

Gallbladder and extrahepatic bile ducts

- 0.5 per cent of all cancers and 0.4 per cent of cancer deaths.
- Sex ratio of rates 1.0 to 1. Age distribution like colorectal cancer.

Cancers of the gallbladder and extrahepatic bile ducts are nearly always classed together, which is unfortunate as the causes differ. The former is more than twice as common in women as in men, is strongly associated with obesity, and is usually preceded by (and probably caused by) cholelithiasis. The latter is slightly more common in men and is increased in incidence by clonorchiasis and (to a less extent) by long-standing ulcerative colitis. Both types are uncommon, and their aggregate varies only slightly from one population to another. Relatively high rates are recorded among Jewesses in Israel, especially among those born in Europe and America.

The incidence of cancer of the gallbladder has fallen sharply in the United States in the last 25 years, which may be partly due to the decreased consumption of animal fat and, perhaps more importantly, to an increase in the rate of cholecystectomy in people who, having gallstones, are at greatest risk of cancer of the gallbladder.

Pancreas

- 2.3 per cent of all cancers and 4.3 per cent of cancer deaths.
- Sex ratio of rates 1.5 to 1. Age distribution like stomach cancer.

Cancer of the pancreas is two to three times more common in regular cigarette smokers than in lifelong non-smokers. The chemicals in cigarette smoke that specifically cause pancreatic cancer have not been identified, but the volatile nitrosamines in smoke that are absorbed from the alveoli and carried to the pancreas in the bloodstream are likely candidates. The disease is twice as common in diabetics as in the population as a whole. It is, therefore, not surprising that the highest rate is recorded among New Zealand Maoris, who smoke heavily and are prone to obesity, hypertension, myocardial infarction, and diabetes.

Cancer of the pancreas is generally regarded as a disease of the developed world, but the diagnosis is difficult in the absence of a well-developed medical service and some of the relatively small geographical and temporal variations may be due to variation in diagnostic standards. Mortality rates in Britain and the United States have begun to decrease under 65 years of age, and this is more likely to reflect a reduction in incidence from reduction in smoking than to any improvement in treatment, as the 5-year survival rate remains well under 10 per cent.

Nose and nasal sinuses

- 0.1 per cent of all cancers and of cancer deaths.
- Sex ratio of rates 2.0 to 1. Age distribution like non-Hodgkin's lymphoma.

Surprisingly, in view of the widespread exposure of the human nose to tobacco smoke and other airborne toxins, cancers of the nasal cavity itself are extremely rare. Most arise from the paranasal sinuses. Several occupational hazards have been recognized, including the refining of nickel, processes giving rise to exposure to strong sulphuric acid mists, and the manufacture of hardwood furniture and leather goods. It would be wrong, however, to conclude that all contact with nickel, hardwood dust, and leather creates a hazard. The hazards have been observed in special occupational situations in which exposure has been intensive and prolonged. The nickel-refining hazard was first observed in South Wales where the nickel carbonyl process was used, but similar hazards were subsequently observed with other refining processes in Canada, Norway, and the Soviet Union. In the Welsh refinery the workplace exposures were much heavier before the Second World War, and (despite the continued use of the nickel carbonyl process in Wales) no hazard of nasal sinus cancer has been observed among men first employed there since 1950. The hazard in furniture workers was first observed in High Wycombe and appears to have followed the introduction of high-speed wood-working machinery early in the twentieth century. A hazard certainly affects some other groups of wood-workers, but should not be assumed to affect furniture workers in general.

Most nasal and nasal sinus cancers are squamous carcinomas, but the hazard from hardwood dust characteristically produced adenocarcinomas. In some of the groups exposed to this hazard, as many as 5 per cent of the men developed the disease. This meant that the risk of adenocarcinoma was increased 1500 times (as this histological type of the disease is normally very rare) and the hazard was, in consequence, easy to confirm once suspicion had been aroused.

Chromate workers are sometimes said to experience a hazard of nasal cancer, but this may be an error due to confusion with the characteristic 'chrome ulcer' of the

nasal septum. Such ulcers have not generally been found to become malignant.

Larynx

- 0.8 per cent of all cancers and 0.5 per cent of cancer deaths.
- Sex ratio of rates 5.7 to 1. Age distribution, see Fig. 6.

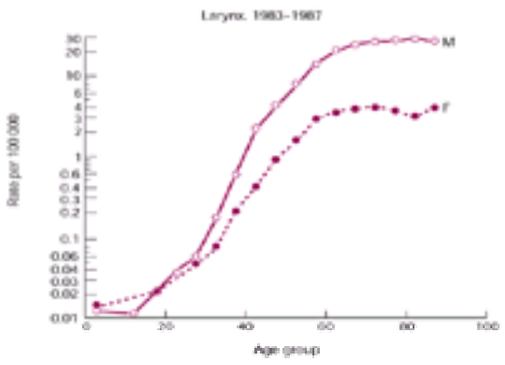

Fig. 6 Annual incidence of cancer of the larynx by age and sex.

Cancers of the larynx, like cancers of the oesophagus and buccal cavity, are closely associated with tobacco smoking and with the consumption of alcohol. The two agents act synergistically and in the absence of either the disease is rare. The different parts of this small organ are, however, related to the two agents differently. Cancers of the glottis are strongly related to smoking, particularly to cigarette smoking, and only weakly to alcohol; while cancers of the epilarynx resemble cancers of the neighbouring hypopharynx and are strongly related to both agents and to pipe and cigar smoking equally with cigarette smoking.

In Scandinavia, the incidence has increased in line with the increase in cancer of the lung. A similar increase has not, however, been seen in Britain and it seems probable that some other aetiological factor, perhaps nutritional in character, has become less prevalent. That there are other causal factors is evident from the relatively high incidence rates in parts of India, Turkey, North Africa and Brazil, which cannot be accounted for by tobacco and alcohol.

The disease has also occurred as an occupational risk in the manufacture of mustard gas and in processes that cause exposure to strong sulphuric acid mists.

Lung

- 13.6 per cent of all cancers and 22.2 per cent of cancer deaths.
- Sex ratio of rates 3.2 to 1. Age distribution, see Fig. 7.

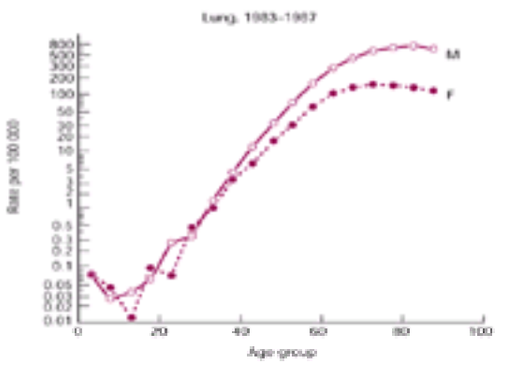

Fig. 7 Annual incidence of cancer of the lung by age and sex.

Nearly all lung cancers are bronchial carcinomas and should properly be so described. The term 'lung cancer' is, however, in such common use that it is used here as synonymous with bronchial carcinoma, although it actually includes a very small proportion of alveolar-cell carcinomas and other rare types of cancer with different characteristics.

Until the 1920s, lung cancer was uniformly rare (except in the Hartz mountains, see below). In the next two decades, German and then British pathologists began to comment on an apparent increase, but this tended to be dismissed as an artefact of the greatly improving methods of diagnosis and the establishment of special centres for thoracic disease. Gradually, however, the increase became so pronounced and the change in the sex ratio so marked that the increase could no longer be dismissed as wholly artefactual and, by the late 1940s when the age-standardized mortality rate in men in the United Kingdom had increased 20 times, it was clear that the developed world had begun to see an epidemic of lung cancer comparable in severity to, though with a longer time scale than, the epidemics of infectious disease of the past. Until the 1940s, the increase among British women was largely a diagnostic artefact, and so provides a useful indication of the quantitative extent to which such artefacts may have affected the male rates. Since 1950, however, diagnostic standards in middle age have changed very little, the increase in British men has been replaced by a decrease, while the increase among middle-aged women has continued for longer, before also reversing. As a result, the sex ratio (male rate divided by female rate) at, for example, 50 to 54 years of age, which rose from 1.8 after the First World War to 8.9 after the Second World War, was reduced to 1.6 in 1998. In the first quarter of the twentieth century, the male excess may have been largely due to the effects of pipe smoking which was an almost exclusively male habit in the nineteenth century.

Smoking

Changes in treatment have had little effect on the fatality rate, which remains extremely high, and real changes in mortality closely reflect real changes in incidence. These can be explained almost entirely by the effect of smoking tobacco, particularly in the form of cigarettes, which caused more than 90 per cent of all lung cancers in Britain in the early 1990s. Evidence of this effect was first obtained in the middle of the last century by comparing the smoking histories of patients with different diseases. It was found that the proportion of patients who had never smoked was much smaller if they had lung cancer than if they had some other disease, and the proportion who had smoked heavily was correspondingly greater.

Further evidence was obtained by asking large numbers of apparently healthy men and women what they smoked and then following them up to determine the causes of death of those that had died. Cohort studies of this type—in over a million American men and women studied by the American Cancer Society, in 34 000 male British doctors, in a regional population of nearly 300 000 Japanese, and in a random sample of the Swedish population—have all shown similar results, the risk increasing with the amount smoked, but varying quantitatively depending on the length of time cigarettes had been smoked. If attention is restricted to populations in which most cigarette smokers had been smoking cigarettes regularly since early adult life, lung cancer is about 20 times more common in regular cigarette smokers than in lifelong non-smokers and up to 40 times more common in very heavy smokers. At first the relationship was less marked in women than in men, but this was because female smokers who were old enough to have a high risk of cancer either had not begun smoking cigarettes so early in adult life or had smoked them less intensively when they began, and the sex differences in behaviour and risk have both been progressively eliminated with the passage of time.

No other exposure has been identified that can account for the extreme difference in lung cancer risk between regular cigarette smokers and lifelong non-smokers and most or all of the excess must have been caused by smoking. Further quantitative studies have found that the relative risk of lung cancer has increased with decreasing age of starting to smoke and decreased with the number of years that smoking has been stopped; that the national increases in incidence have appeared

at appropriate times after the increase in cigarette sales (after due allowance is made for a spurious increase due to improved diagnosis and appropriate differences in consumption by men and women), and that there is a general parallelism between the incidence of the disease in different countries and social and religious groups and the prolonged consumption of cigarettes. Furthermore, it has been found that when extracts of cigarette smoke are applied repeatedly to the skins of laboratory mice many tumours develop. Finally, and most encouragingly, the trend in mortality has reversed following reduction in smoking. By 1998, the mortality from lung cancer among men in their 30s in Britain was only about a quarter of that of men of the same ages some 40 years earlier, corresponding to the earlier changes in the prevalence of smoking. The reduction in tar delivery between 1939 and 1965 contributed to the reduction in lung cancer in young men after the war, but the later reduction had little effect because of changes in the way cigarettes were manufactured and in the way they were smoked to ensure an adequate intake of nicotine. At older ages the decreases are less striking, but they are now seen at all ages in British men, and up to 60 in British women.

Occupation

Several other causes of lung cancer have been discovered as a result of observations in industry. Many thousands of men and women have experienced significant hazards from exposure to asbestos or to polycyclic hydrocarbons (from the combustion of fossil fuel). The former has given rise to hazards in asbestos mines, asbestos textile works, and insulation work in the shipbuilding and construction industries and the latter to specific hazards in the manufacture of coal gas in coking ovens, in steel works, in aluminium foundries, and wherever substantial amounts of incompletely combusted fumes were released into the working environment. Much smaller numbers of men have experienced substantial hazards from radon in the air of mines (not only when mining radioactive materials, but also when mining haematite and fluorspar under conditions in which radon seeped into the mine air from streams and the surrounding rock), from the manufacture of chromates and chrome pigments, from the refining of nickel, from arsenic (in the manufacture of arsenical pesticides and in the refining of copper, which is always contaminated with arsenic), from the manufacture of mustard gas, and, to a small extent, from exposure to silica if sufficient to cause silicosis. In one extreme situation, the absolute risk of contracting lung cancer due to the occupational hazard of radon was so large that more than half the workers contracted the disease (in the cobalt mines of the Hartz mountains in Central Europe, that were subsequently mined for radium and uranium). In several other situations with heavy exposure to asbestos or the early stages of nickel refining, the occupational hazard has affected as many as 20 to 30 per cent of the exposed men.

Atmospheric pollution

Some of the materials responsible for these occupational hazards—particularly the combustion products of fossil fuels—are or have been widely distributed in the air of towns and it is still uncertain how far they have, in this way, contributed to the production of the disease in the general population. That lung cancer was more common in big towns than in small towns and rural areas is certain, but this held as strongly for Oslo and Helsinki, two relatively unpolluted cities, as it did for London, Birmingham, Manchester, Chicago, Los Angeles, and Pittsburgh. Differences between the largest towns and the least populated areas have seldom been more than three-fold and much of the difference can be accounted for by past differences in cigarette smoking, a habit which has tended to spread outwards from the major cities. Attempts to 'allow for' cigarette smoking have usually been inadequate, as it is impossible to take full account of such factors as the age of starting to smoke cigarettes, the amount smoked daily at different periods, and the method of smoking (number of puffs, depth of inhaling, etc.). It is clear, however, that in the absence of cigarette smoking any effect of urban pollution in developed countries is relatively small. Estimates, based on extrapolation from the heavy pollution with coal-smoke that used to occur in large towns, suggest that in such towns it may have contributed, in synergism with smoking, to as much as 10 per cent of the risk of lung cancer, but would have caused very little risk in non-smokers. On this basis, the present levels of pollution with benzo(a)pyrene and the other known lung carcinogens in town air can be only very small. Modern pollution with ultra-fine particles (<10 μm diameter) may, however, be more hazardous. Study of residents in six contrasting cities in the United States in which information about personal smoking habits had been obtained suggests that the risk in the most polluted city compared to that in the least polluted could be increased by about a quarter in both smokers and non-smokers. The position in some developing countries is different: notably in parts of China, where intense indoor pollution with smoke and fumes from heating and cooking more than doubles the risk of lung cancer in non-smokers.

The effect of another form of pollution—that of house air with radon leaked from underground rocks and building materials—can be estimated by extrapolation from the effects of the much larger doses to which some groups of underground miners have been exposed and by direct observation in case–control studies of people with and without lung cancer. Both methods suggest that indoor radon may contribute to about 6 per cent of lung cancers in the United Kingdom and 12 per cent in the United States. Most cases are caused in synergism with smoking, so that in the absence of smoking only few cases would be produced.

Geographic differences

The development of the male lung cancer epidemic and the early signs of its departure have been most prominent in Britain and Finland, since the switch of young men to cigarettes was largely complete in these countries by the 1920s. In the United States, where cigarette consumption doubled during the Second World War, the benefits of recent reductions in tobacco exposure are superimposed on the increasing lung cancer rates due to the delayed effects of past increases in smoking in early adult life by those who are now reaching middle and old age. Hence, it is thus far only among younger men in the United States that the benefits of reduced smoking and a switch to low-tar cigarettes are causing net decreases in lung cancer mortality. In some other developed countries, the development of the epidemic is still further behind and it is only just beginning to appear in many developing countries. Chinese males, for example, who now consume about 30 per cent of the world's cigarettes, experienced a three-fold increase in cigarette consumption during the 1980s that may well eventually cause almost a million cancer deaths a year when the young men of today reach middle age.

In women, the development of the epidemic has been later than in men. Only in the Maori population of New Zealand did it occur at the same time. In the United Kingdom, United States, and a few other developed countries, the female lung cancer rates are already substantial, but in others, such as France and Spain, the epidemic in women has scarcely begun. A relatively high risk has long been noted in Chinese women who are non-smokers, irrespective of their country of residence, which is probably due to their exposure to mutagens in the fumes from oils used in cooking with a wok and from the coal smoke with which many Chinese homes are heavily polluted.

Pleura and peritoneum

- 0.5 per cent of all cancers and 0.4 per cent of cancer deaths.
- Sex ratio of rates 6.0 to 1. Age distribution like laryngeal cancer.

The existence of a specific type of tumour arising from the pleura, or less commonly the peritoneum, was debated by pathologists until 1960 when Wagner and his colleagues reported that six African patients with a similar type of 'peripheral lung cancer' had all lived in villages that were heavily polluted with dust produced by the mining of blue asbestos (i.e. crocidolite). Since then, many cases have been reported throughout the world, the great majority of which have been specifically associated with exposure to asbestos at work. They are much less likely to be produced by white asbestos (chrysotile) than by brown asbestos (amosite) or blue, as the two last persist for longer in the lungs. A few cases arise from neighbourhood pollution with asbestos or secondary contamination (e.g. from household contact with asbestos workers) and some in Turkish villages are due to the weathering into the general atmosphere of mineral fibres in local rock that are physically similar to, but chemically different from, asbestos. A few cases have been caused by radiotherapy and natural ionizing radiations may be responsible for most of those that are not associated with asbestos. An SV-40 like virus has been found in some tumours; but it is uncertain whether it plays a part in causing the disease.

Mesotheliomas seldom occur less than 15 years after first exposure to asbestos, commonly occur 25 to 30 years afterwards, and may be delayed for 50 years or more. Almost all cases are fatal, so that the mortality would reflect incidence, if all cases were correctly diagnosed. Due to confusion with lung or other types of cancer, it is still uncertain how many cases have occurred each year and some of the large increase since 1960 may be artefactual. In the last few years, the recorded mortality under 70 years of age has begun to decrease.

Pleural mesothelioma is not related to cigarette smoking and the occupational hazard affects smokers and non-smokers alike.

Bone

- 0.1 per cent of all cancers and of cancer deaths.
- Sex ratio of rates 1.3 to 1. Age distribution, see Fig. 8.

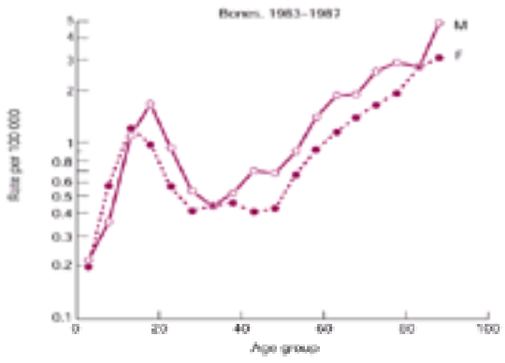

Fig. 8 Annual incidence of cancer of bones by age and sex.

Sarcomas can affect any bone, but characteristically affect the long bones in adolescence. After 45 years of age they occur most commonly in bones affected by Paget's disease (osteitis deformans), which predisposes to sarcoma so strongly that as many as 1 per cent of all people affected by the disease eventually develop a bone tumour.

Many different histological varieties occur, some of which appear to have different causes. Osteogenic sarcomas and chondrosarcomas are the most common, the former accounting for nearly all the adolescent peak. One rare type (Ewing's tumour) occurs only in childhood and adolescence and is almost unknown in black people, irrespective of the society in which they live.

Ionizing radiations are the only known extrinsic cause. Cases have been produced after intensive radiotherapy or the medicinal use of thorium (a bone-seeking radionuclide). In industry they have occurred in 'luminizers' who, in previous decades, used delicate paint brushes to apply radium compounds, and ingested radium, possibly as a result of 'pointing' the paint brushes in their mouths.

National statistics record a reduction in mortality over the last 50 years, but are unreliable indicators of incidence as many deaths attributed to tumours of bone are due to cancers that have metastasized from other sites. The recorded decrease in mortality is, therefore, largely an artefact due to improved diagnosis (though it has been contributed to in recent years by higher survival rates in childhood) and the true incidence may have remained roughly constant.

Connective tissues

- 0.4 per cent of all cancers and 0.5 per cent of cancer deaths.
- Sex ratio of rates 1.3 to 1. Age distribution, see Fig. 9.

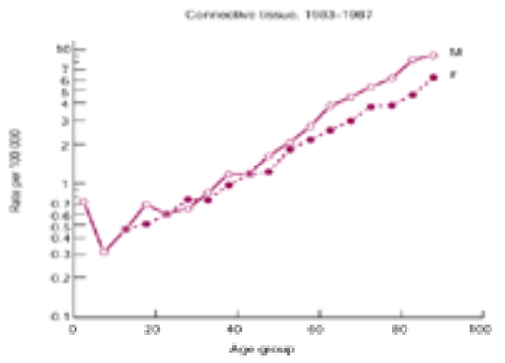

Fig. 9 Annual incidence of connective tissue sarcoma by age and sex.

Sarcomas of the soft tissues include a variety of different diseases, all of which are rare everywhere. Some are genetic in origin and others are caused by ionizing radiations. A few may be caused by intensive immunosuppression or exposure to chlorophenols, but the evidence is inconclusive.

Skin (melanoma)

- 1.8 per cent of all cancers and 1.0 per cent of cancer deaths.
- Sex ratio of rates 0.6 to 1. Age distribution, see Fig. 10.

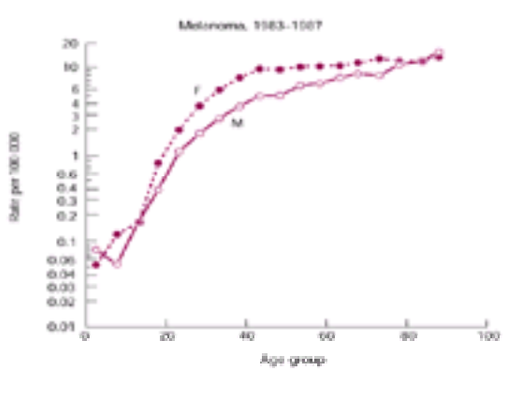

Fig. 10 Annual incidence of melanoma of the skin by age and sex.

The incidence of the disease varies inversely with the amount of skin pigmentation. In white people the tumour occurs most commonly on the legs (in women) and the trunk, head, and neck (in men). It is extremely rare in blacks in the United States, but is more common in Africa, where it occurs at the junction of the pigmented and unpigmented skin on the sole of the foot. Like basal-cell and squamous carcinoma of the skin, it is particularly common in sufferers from xeroderma pigmentosum.

Incidence rates in white people vary roughly in proportion to the flux of ultraviolet light in the countries in which they live. For all sites combined, the incidence is not, however, greater in outdoor than indoor workers (rather the reverse, in fact, perhaps due to the protective effects of a semipermanent suntan) and it seems to be associated with periodic bouts of sunbathing and sunburn. Incidence and mortality rates have increased in Britain, the United States, and in many other countries with mainly white populations. The increase began in cohorts born early this century, who exposed their skin to the sun more than their predecessors had done, and it still continues at ages over 60 years. The totality of the evidence suggests that ultraviolet light is the principal cause, but the relationship is not simple and other factors may also be important.

Skin (non-melanoma)

- 14.1 per cent of all cancers and 0.3 per cent of cancer deaths.

- Sex ratio of rates 1.5 to 1. Age distribution, see Fig. 11.

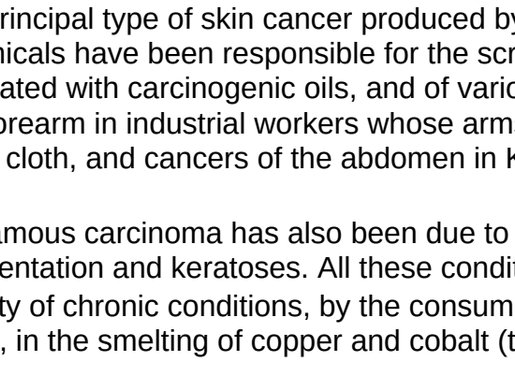

Fig. 11 Annual incidence of non-melanomatous skin cancer by age and sex.

Non-melanomatous skin cancers are of two main types, basal-cell and squamous carcinomas. The former, also known as rodent ulcers, are produced by ultraviolet light. They occur mainly on parts of the body that are regularly exposed to the sun and, in particular, on the face, head, and neck. They are more common in outdoor workers, such as seamen and farmers, than in indoor workers, more common in fair-skinned than in dark-skinned people, and are almost unknown in blacks (except those who suffer from albinism). Some few cases have been produced by exposure to X-rays, but the risk is very small unless the dose is very large and they seldom occur after normal courses of radiotherapy. People who suffer from xeroderma pigmentosum, a hereditary condition in which there is a defect in the enzyme responsible for the repair of the damage done to DNA by ultraviolet radiation, develop large numbers of skin tumours at an early age in response to even quite mild exposure to diffuse sunlight (see Chapter 23.1).

Squamous carcinoma is also produced by ultraviolet light, but less easily, so that it accounts for only about 20 per cent of cancers on the exposed skin. It is, however, the principal type of skin cancer produced by various carcinogenic chemicals, and particularly by polycyclic hydrocarbons in the combustion products of coal. These chemicals have been responsible for the scrotal cancers of chimney sweeps, who accumulated soot in the folds of the scrotum, of mule spinners, whose clothes were saturated with carcinogenic oils, and of various other groups of workers whose clothes were contaminated with tar. They have caused (and still do cause) cancers of the forearm in industrial workers whose arms are regularly splashed with tar or carcinogenic oils, cancers of the groin in India, localized by the continued friction of the *dhoti* cloth, and cancers of the abdomen in Kashmir associated with the habit of carrying a *kangri*, or small stove, inside the clothes in winter to keep warm.

Squamous carcinoma has also been due to prolonged exposure to arsenic, which is excreted by the skin and in the hair, when it may be accompanied by arsenical pigmentation and keratoses. All these conditions have been produced by prolonged medical treatment with inorganic arsenic, which used to be prescribed for a variety of chronic conditions, by the consumption of well water from arsenic-rich soils, and by occupational exposure, sometimes to as much as 1000 µg of arsenic/m^3 of air, in the smelting of copper and cobalt (the ores of which often contain arsenic), and in the manufacture of arsenical pesticides.

How large a part human papilloma viruses play in the development of squamous carcinoma of the skin is unclear. The type 5 virus is responsible for the warty lesions of epidermodysplasia verruciformis, some of which progress to cancer, and other types of the virus may contribute to the increased risk that follows the intensive immunosuppression given to permit the survival of organ transplants.

A third type is Kaposi's sarcoma, which is now classed as a skin cancer. It is associated with AIDS when AIDS results from homosexual intercourse, but probably only when this is accompanied by orofaecal contact. Frequent at first, particularly in the United States, the association has become progressively less common. Before the advent of AIDS, Kaposi's sarcoma was common in some parts of Central Africa, where it occasionally affected children, progressed rapidly, and could account for as many as 10 per cent of all hospital patients with cancer. Elsewhere it was rare, but indolent cases occurred occasionally in developed countries, principally on the legs of middle-aged and elderly men. The disease is initiated by infection with the human herpes virus type 8, but cofactors are required for tumour development.

Breast

- 12.2 per cent of all cancers and 8.6 per cent of cancer deaths.
- Sex ratio of rates 0.01 to 1. Age distribution, see Fig. 12.

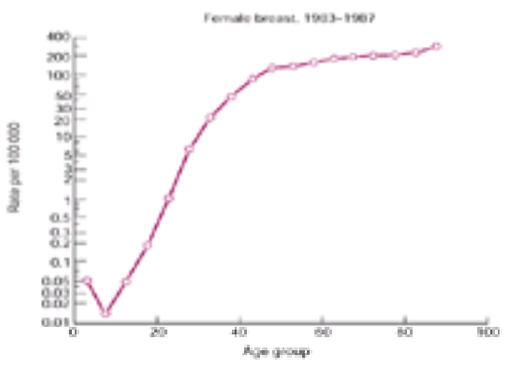

Fig. 12 Annual incidence of breast cancer in women by age.

Cancer of the breast was the most common fatal cancer in women throughout most of the developed world, but is now being displaced by lung cancer in the United Kingdom and the United States. It is less common in Eastern Europe and much less common in Asia and in black African populations south of the Sahara. Incidence rates have tended to rise slowly in many countries, but the changes have been relatively small and decreases have recently been recorded in young age groups. The geographical differences are unlikely to be chiefly due to genetic factors, as black women in the United States and Japanese women in Hawaii have rates that are similar to those in their white American compatriots and much greater than those in their countries of origin.

Hormonal factors are important in the production of the disease, particularly oestrogens, but others may also be important. The duration of ovarian activity is relevant, as the disease is particularly common in women who have an early menarche and a late menopause (the former being more important than the latter). Pregnancy produces a short-term increase in risk, followed after a few years by a lifelong decrease, particularly after teenage or early adult pregnancies. The incidence in later life increases progressively with a woman's age at the time of her first full-term pregnancy, being about three times greater when the first birth occurs after 35 years of age than when it occurs before 18 years. Full-term pregnancies after the first have an additional protective effect. Pregnancies that end in abortion have little or no effect, however, suggesting that the effects of pregnancy depend on the induction of lactation. The duration of lactation has an additional protective effect but is not marked unless it continues for a year or more.

Parity and menstrual differences are insufficient to account for the large variations in the incidence of the disease in different countries, which seem to be correlated with a 'high' standard of living: that is, with life in a developed country. Diet may play an important part, but the evidence is complex and inconclusive. Obesity is associated with a reduced risk before the menopause, as it tends to be associated with ovarian dysfunction. After the menopause, obesity increases both the incidence and probably the fatality of the disease. Oestrogens prescribed medically, as hormone replacement therapy (HRT) after the menopause, increase the risk by about 2 per cent for each year of use; combined with progestogens in the contraceptive pill they increase it by about 25 per cent during use, but the increased risk gradually disappears over 10 years, when use is stopped, as it does after HRT is stopped. Tamoxifen, an antioestrogen prescribed for the treatment of breast cancer, reduces the subsequent incidence of the disease in the unaffected breast.

Much of the recent increase in incidence is the result of intensive case finding in association with mammography and there is no reason to attribute it to any form of environmental pollution.

Cervix uteri

- 1.2 per cent of all cancers and 1.1 per cent of cancer deaths.
- Confined to women. Age distribution, see Fig. 13.

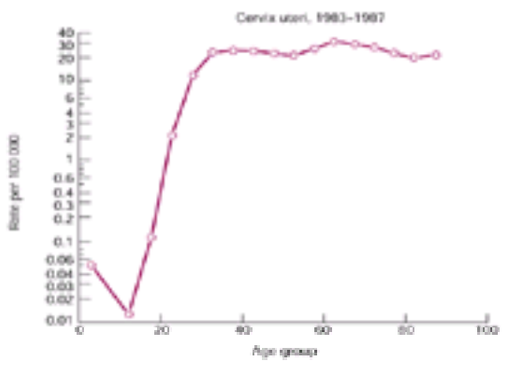

Fig. 13 Annual incidence of cancer of the cervix uteri by age.

Carcinoma of the cervix is the most common type of cancer throughout much of Africa, Asia, and Latin America, and used also to be common in Europe and North America. It has always been rare in Jewesses and has tended to be less common in Muslim women than in women of other faiths living in the same country (as, for example, Hindus in India).

Changes in incidence have been difficult to assess, partly because mortality data have not always distinguished between deaths due to cancer of the cervix and those due to cancer of the corpus (or endometrium), partly because the introduction of screening programmes has made it possible to diagnose and treat premalignant lesions (see below), and partly because hysterectomy for benign conditions has become progressively more common, with a corresponding reduction in the number of uteri in which the disease could occur. Despite these complications there can be no doubt that the disease has become substantially less common in Europe and North America than it was before the Second World War.

The rarity of the disease in Jewesses and its relative rarity in Muslims suggest that male circumcision may reduce the risk of its development, but this is unlikely as the state of circumcision of her husband has no substantial effect on a woman's risk of developing the disease in communities in which only some men are circumcised. Cleanliness is likely to be protective, as the disease is relatively uncommon in communities that practise ritual ablution before and after intercourse and, within each community, it becomes less common with rising socioeconomic status.

Squamous carcinoma, which constitutes the vast majority of all cases, is intimately connected with sexual activity. It almost never occurs in virgins and increases in frequency with the number of sexual partners that a woman or her partner has had. The great majority of cases are attributable in part to infection with some types of the human papilloma virus, most notably types 16 and 18.

The development of squamous carcinoma is preceded by pathological changes limited to the epithelium, known as cervical intraepithelial neoplasia (CIN) types I, II, and III. CIN III is associated with the same types of virus as squamous carcinoma, but CIN I and CIN II generally are not. The changes may progress from one to another, finally leading to carcinoma, but the early lesions (CIN I and II) commonly regress and even CIN III (previously known as carcinoma *in situ*) may do so occasionally. The lesions can be recognized in cervical smears and destroyed by lasers or extensive biopsy and the occurrence of clinical disease can be greatly reduced by the examination of all sexually active women every 2 or 3 years and the treatment of advanced CIN lesions.

Other factors associated with the production of the disease are the use of oral contraceptives and cigarette smoking. Both tend to be associated with behaviour conducive to venereal infection, but it is uncertain whether this tendency can wholly account for their association with the disease. That smoking may be responsible for some cases is suggested by the presence of mutagens in the cervical mucus of smokers that are not present in the secretions of non-smokers.

Adenocarcinoma of the cervix uteri is generally rare, but may have become somewhat more common recently. Its causes are unknown.

Endometrium (corpus uteri)

- 1.5 per cent of all cancers and 0.7 per cent of cancer deaths.
- Confined to women. Age distribution like cancer of ovary.

The epidemiological features of endometrial cancer are in many respects the opposite of those of cervical cancer. Histologically, it is nearly always an adenocarcinoma. It is common in developed countries, rare in poor populations, and is, if anything, becoming more common with the passage of time. It is inversely related to parity, but not otherwise related to coitus, and is unaffected by the number of sexual partners. Like cancer of the breast, it is positively associated with early menarche and late menopause.

The one factor known to produce the disease is regular exposure to oestrogens, unopposed by progestogens. This leads to endometrial hyperplasia and eventually, in some cases, to cancer. Known causes include oestrogen-secreting tumours of the ovary, the use of oral contraceptives in which oestrogens and progestogens are prescribed sequentially (types that have now been abandoned), the use of 'natural' conjugated oestrogens to relieve menopausal and postmenopausal symptoms, and adiposity. The last causes the disease because oestrogens are produced in the body after the menopause in adipose tissue from the adrenal hormone, androstenedione. Tamoxifen, an analogue of the natural oestrogens, which blocks oestrogen receptors in the breast and hence acts as an antioestrogen, can, due to differences between the hormone receptors in different tissues, have a pro-oestrogenic effect in some other organs, and slightly increases the incidence of endometrial cancer.

It is improbable that oestrogens are initiating agents. They are not mutagens *in vitro* and the changes that took place in the incidence of the disease in the United States following the increase and the subsequent reduction in the use of premarin (a conjugated oestrogen) for the treatment of menopausal symptoms occurred so quickly that they make sense only if oestrogens act on some late stage(s) of the carcinogenic process.

Ovary

- 2.0 per cent of all cancers and 2.9 per cent of cancer deaths.
- Confined to women. Age distribution, see Fig. 14.

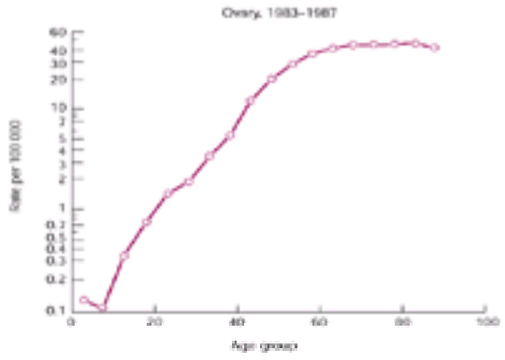

Fig. 14 Annual incidence of cancer of the ovary by age.

Prostate

- 7.4 per cent of all cancers and 6.3 per cent of cancer deaths.
- Confined to men. Age distribution, see Fig. 15.

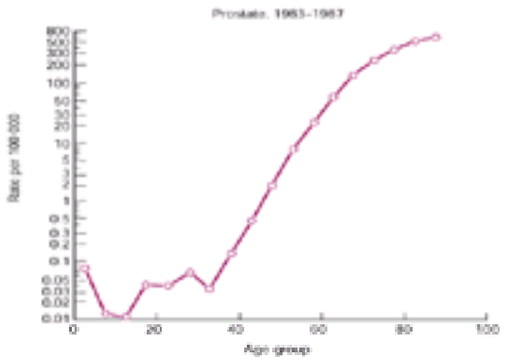

Fig. 15 Annual incidence of cancer of the prostate by age.

Cancer of the prostrate is more characteristically a disease of old age than any other cancer, so that it comes to play a much larger part in clinical experience as the proportion of old people in the population increases. It is unusual in that foci of cells resembling cancer can be found in a high proportion of clinically normal prostates, so that the recorded incidence is drastically increased by increasing the number of prostatic biopsies. Some increase in mortality has been recorded in Britain and North America, but the weight of evidence suggests that the disease is principally due to factors that have affected society for many years. What these factors are remains obscure. Associations have been reported with both increased and decreased sexual activity. On general grounds it seems likely that the disease is dependent on hormonal imbalance (particularly as castration slows the progression of clinical disease) but the nature of the imbalance is unknown. Vasectomy was thought to increase the incidence of the disease, but probably does not.

Two epidemiological observations stand out: the high incidence in black populations throughout the world, and the low incidence in Japanese. Both may be partly due to genetic factors, but they are not wholly so, as both Japanese and blacks have higher rates in the United States than they have in Japan and Africa.

Testis

- 0.5 per cent of all cancers and 0.1 per cent of cancer deaths.
- Confined to men. Age distribution, see Fig. 16.

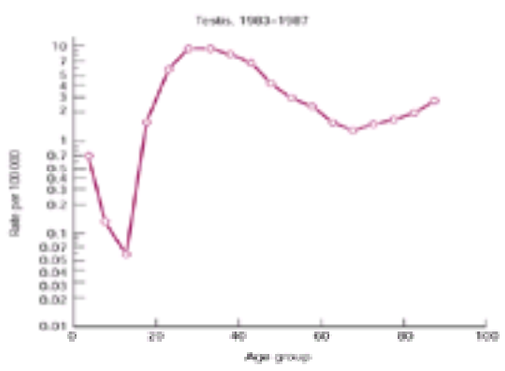

Fig. 16 Annual incidence of cancer of the testis by age.

Testicular cancers are of two main types. Seminomas, which are the more common, have a peak incidence at about 30 years of age and teratomas, commonly called embryomal carcinomas in the United States, which have a peak incidence about 10 years earlier. Tumours after 50 years of age are mostly lymphomas and are now classed as such. Both genetic and environmental factors are important. On the one hand, the disease is uniformly rare in black populations, whether in Africa or in the United States. On the other, it has increased in incidence in many countries, notably in Denmark and Britain. In Britain, the increase began in the 1920s and affected first the higher socioeconomic groups. The increase trebled the mortality at 15 to 34 years of age and produced a sharp peak in young adult life that had not previously been present. In the United States, the increase started later and has been less marked. The disease is much more likely to occur in an undescended than in a normal testis, but otherwise its causes are unknown.

Penis

- 0.1 per cent of all cancers and of cancer deaths.
- Confined to men. Age distribution like skin (non-melanoma).

Carcinoma of the penis is at all common only in some parts of tropical Africa and Brazil, where it has accounted for 10 per cent of all cancers in men. It is avoided almost entirely by circumcision at birth and is very rare if circumcision is carried out in boyhood. In developed countries it is rare even in the absence of circumcision if the glans, coronary sulcus, and foreskin are kept clean.

The oncogenic types of the human papilloma virus (principally types 16 and 18) can usually be identified in the malignant cells and are important causes of the disease.

Bladder

- 4.5 per cent of all cancers and 3.3 per cent of cancer deaths.
- Sex ratio of rates 3.6 to 1. Age distribution like gastric cancer.

Cancer of the bladder can be produced by cigarette smoking, occupational exposure to a group of chemicals classed together as aromatic amines, infestation of the bladder with *Schistosoma haematobium*, and the medical prescription of chlornaphthazine (*N*,*N*'-bis(2-chloroethyl)-2-naphthylamine) and cyclophosphamide. Most bladder cancers are transitional cell carcinomas, but those associated with schistosomiasis are characteristically squamous carcinomas. It is not surprising that the bladder should be affected by many chemicals, as any noxious small molecules in the blood will tend to be found at greatly increased concentration in the urinary tract. Cigarette smoke contains several mutagenic chemicals that enter the bloodstream and thence the bladder, so that when tested *in vitro* on bacterial DNA the urine of cigarette smokers is found to be mutagenic, while that of non-smokers is barely active.

Occupation

An occupational cause was first suspected in 1898 in Germany, when Rehn commented on a cluster of cases in men using aniline for the manufacture of dyes. Aniline, however, is not carcinogenic in experimental animals, more recent studies have failed to incriminate it epidemiologically, and it seems likely that other carcinogenic chemicals were present as impurities. Four aromatic amines that are carcinogenic in experimental animals have been shown to cause bladder cancer in humans: 2-naphthylamine, benzidine, 3,3'-dichlorobenzidine, and 4-aminobiphenyl. The first is one of the most powerful human carcinogens yet known and was responsible for the development of bladder cancer in all the 19 men who were employed in distilling it in a British factory. Its manufacture in Britain was stopped in 1949; but small amounts continued to be imported until the 1960s. Other aromatic amines that may cause bladder cancer include auramine, magenta, and, perhaps, 1-naphthylamine. The last is dubiously carcinogenic in experimental animals and it seems probable that the cases associated with its use have been due to a few per cent of 2-naphthylamine present as an impurity in the commercial material. These chemicals were used in the manufacture of dyes, in the rubber industry as antioxidants (1-naphthylamine and 4-aminobiphenyl) and hardeners (benzidine), and in laboratories as a reagent (benzidine). 2-Naphthylamine is also found in the combustion products of coal and may have been responsible for the hazard of bladder cancer in men who made coal gas. As many as 10 per cent of cases were, at one time, attributable to occupational causes in Britain and North America; but the proportion should now be much less.

Smoking

The most important cause numerically is cigarette smoking, which probably accounts for about half the total number of cases in Britain and North America. 2-Naphthylamine and 4-aminobiphenyl are present in cigarette smoke, but whether the amounts are sufficient to account for the carcinogenic effect is uncertain.

Medicines

The two medicinal causes have, by contrast, been responsible for only a handful of cases. Chlornaphthazine was used briefly for the treatment of myelomatosis, until it was found to be metabolized into 2-naphthylamine. Cyclophosphamide is used primarily for the treatment of malignant disease, but it is also used as an immunosuppressant. In large doses it may cause sloughing of the bladder mucosa and, occasionally, cancer.

Parasitic infection

Heavy infection of the bladder with *Schistosoma haematobium* has been found to be a cause of the disease, most notably in Egypt and Tanzania.

Diet

The evidence linking bladder cancer to diet is weak. Several case–control studies suggested a positive relation with the consumption of coffee, but the results were inconsistent and it is difficult to exclude the effect of confounding by the stronger relation with cigarette smoking. Artificial sweeteners came under suspicion because of the results of animal experiments in which, first, mixtures of cyclamates and saccharin and then saccharin alone were shown to cause bladder cancer in rats. The human use of cyclamates was banned before saccharin came under suspicion and it now appears that the 'positive' results of animal experiments with cyclamates alone were due to impurities. Saccharin has been shown to cause bladder cancer in rats in feeding experiments, especially when given over two generations and when given after a single instillation into the bladder of a powerful carcinogen. In both instances the quantities that had to be given were large, constituting a few per cent of the feed. The human evidence is extensive and could hardly be more negative, except that it does not cover lifelong use.

Kidney

- 1.8 per cent of all cancers and 2.0 per cent of cancer deaths.
- Sex ratio of rates 2.1 to 1. Age distribution, see Fig. 17.

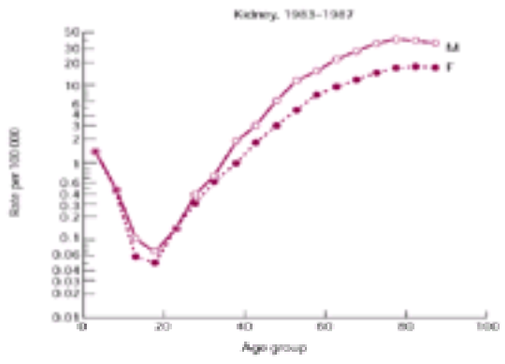

Fig. 17 Annual incidence of cancer of the kidney by age and sex.

Cancers of the kidney are of three main types: nephroblastomas (or Wilms' tumours), adenocarcinomas (or hypernephromas), and transitional- and squamous-cell carcinomas of the renal pelvis. The first are limited to childhood, occur with almost equal frequency everywhere, and apart from a few of genetic origin, are of unknown aetiology. The second constitute by far the majority of all cases, are more common in Western Europe and North America than in Africa and Asia, and have been slowly increasing in incidence. Cigarette smoking is one cause, but the association is weak and it does not account for more than about a quarter of the cases.

The third type of renal cancer (carcinoma of the pelvis) constitutes some 10 per cent of all cases. Three established causes are occupational exposure to the chemicals that cause cancer of the bladder, smoking, and the consumption of phenacetin in large enough amounts to produce analgesic nephropathy. In all three cases the hazards are relatively small (two to three-fold). A fourth cause, Balkan nephropathy (see Section 20.7) increases the risk several hundred-fold.

Brain

- 1.3 per cent of all cancers and 2.1 per cent of cancer deaths.
- Sex ratio of rates 1.4 to 1. Age distribution, see Fig. 18.

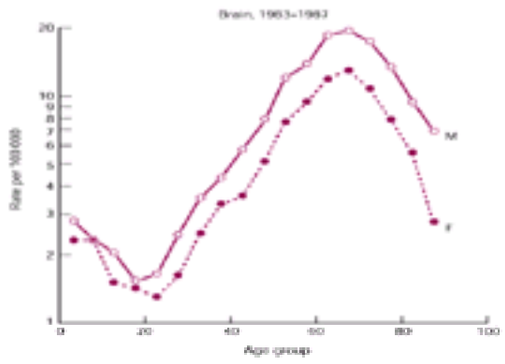

Fig. 18 Annual incidence of cancer of the brain by age and sex.

Tumours of the brain and nervous system are of several different histological types, some of which may not be clearly either benign or malignant. One type occurs characteristically in childhood (medulloblastoma), another in adult life (glioblastoma), and a third (astrocytoma) at all ages. Despite the overall male excess, one type (meningioma) is more common in women.

A moderately large increase in incidence in old age has been recorded in many countries, which can be attributed to improved diagnosis with computerized tomographic scans and nuclear magnetic imaging. Little or no increase in mortality has been reported in or before middle age and the recorded increases in incidence are certainly largely, and possibly wholly, artefactual. No new environmental cause has been established, but many have been suspected, including electromagnetic fields associated with the use of electricity (50–60 Hz) and mobile phones.

Thyroid

- 0.4 per cent of all cancers and 0.2 per cent of cancer deaths.
- Sex ratio of rates 0.5 to 1. Age distribution, see Fig. 19.

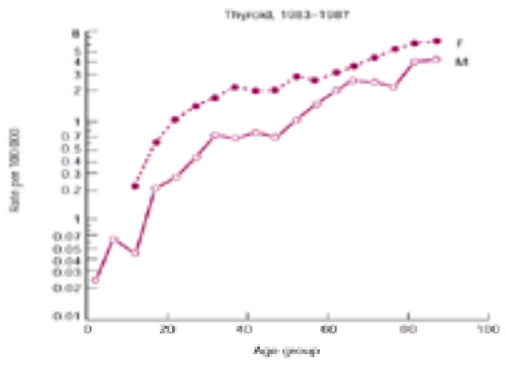

Fig. 19 Annual incidence of cancer of the thyroid by age and sex.

The thyroid is particularly sensitive to ionizing radiation in childhood. Substantial numbers of cases have occurred among the survivors of the atomic explosions in Hiroshima and Nagasaki, children who were exposed to large amounts of radioactive iodine following the Chernobyl accident, and young people whose necks were irradiated in infancy for the treatment of an enlarged thymus (a condition now considered to be perfectly normal, but at one time thought to be a cause of sudden death). Fortunately, the thyroid tumours produced by ionizing radiations are nearly all of the papillary and follicular types, which respond well to treatment. No causes are known of the medullary and anaplastic types, which have a high fatality and occur only in adult life.

The disease is several times more common in Iceland, northern Norway, Hawaii, Fiji, and Israel than elsewhere.

Hodgkin's disease (Hodgkin's lymphoma)

- 0.5 per cent of all cancers and 0.2 per cent of cancer deaths.
- Sex ratio of rates 1.6 to 1. Age distribution, see Fig. 20.

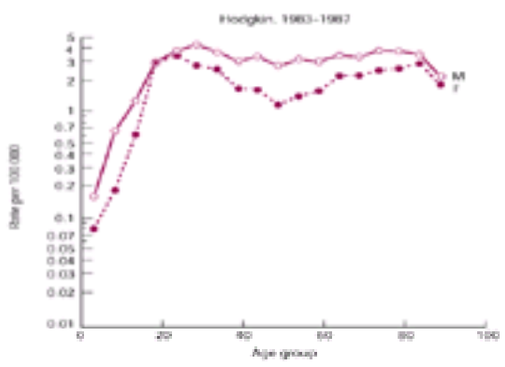

Fig. 20 Annual incidence of Hodgkin's lymphoma by age and sex.

Hodgkin's disease is best thought of as at least two diseases, one affecting primarily youths and young adults, the other primarily the middle aged and elderly. This division is suggested partly by the existence of two peaks in the age-specific incidence rates, partly by the histological appearances (younger patients tending to have the nodular sclerotic form of the disease and older patients the mixed cellular form), and partly by the clinical distinction that young patients show mediastinal involvement in more than 50 per cent of cases and infradiaphragmatic involvement in less than 5 per cent, while the reverse tends to be true in the elderly.

There are several reasons for thinking that the type characteristic of young people is infective in origin. In developing countries, Hodgkin's disease occurs in childhood, but as the standard of living rises, the childhood cases disappear and are replaced by a larger number in young adults. This is reminiscent of what happened to poliomyelitis in the first half of the century and suggests that the disease may be due to a ubiquitous infective agent that becomes less widespread as hygiene improves. That the agent was likely to be the Epstein–Barr virus (EBV or human herpes virus type 4) was suggested by the finding that the incidence was increased 5 to 20 years after a clinical attack of infectious mononucleosis and the virus has now been found in the DNA of the malignant cells characteristic of the tumour (the Reed–Sternberg and tumour reticulum cells). As with other virus-induced cancers, there are likely to be cofactors (at present unknown) that determine whether cellular infection leads to the production of a malignant clone.

Non-Hodgkin's lymphoma

- 2.8 per cent of all cancers and 2.9 per cent of cancer deaths.

- Sex ratio of rates 1.5 to 1. Age distribution, see Fig. 21.

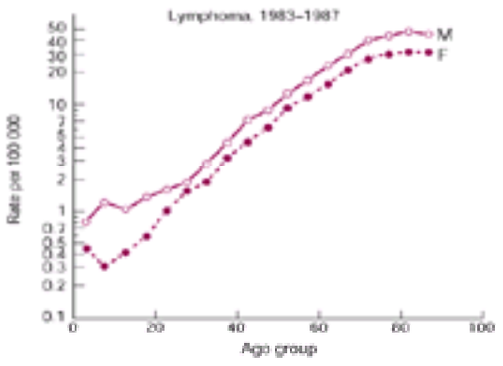

Fig. 21 Annual incidence of non-Hodgkin's lymphoma by age and sex.

Non-Hodgkin's lymphoma is a non-specific term that embraces several diseases with different histological appearances. The histological classification has, however, varied from place to place and from time to time, and it has been difficult to collect epidemiological information about the individual types.

One type that has been clearly distinguished is Burkitt's lymphoma, derived from B lymphocytes. This affects children everywhere, but is common only in a few areas in which malarial infection is both heavy and widespread. In parts of Uganda, Tanzania, and Nigeria the disease is 100 times more common than in Europe and North America. In high incidence areas, EBV can nearly always be recovered from the lymphomatous cells and part of its genome is identifiable in the cells' DNA. Infection with the virus is, however, not necessary for the development of the disease, as some cases occur in its absence; nor is it sufficient, as infection is almost universal and occurs at a very young age in high incidence areas. It seems, therefore, that EBV is a potential cause and that its carcinogenic effect is precipitated by the intense stimulation of the reticuloendothelial system that is characteristic of heavy and chronic malarial infection.

Another type occurs as part of the adult T-cell leukaemia–lymphoma syndrome that follows infection with the human T-cell leukaemia–lymphoma virus (HTLV-1). The disease is common in South Japan and the Caribbean, but may occur occasionally anywhere.

A third type, primary upper small-intestinal lymphoma (PUSIL), affects young people in many populations with a low standard of living, not only in North Africa and the Middle East (where its frequency gave it the earlier name of Mediterranean lymphoma) but also in South Africa and Central and South America. Malnutrition is not, however, a sufficient cause as it is uncommon in Bangladesh and several other malnourished populations.

A fourth type, the mucosa-associated lymphoid tissue tumour known as a maltoma, occurs in the stomach as a result of *H. pylori* infection and can be cured by aggressive treatment of the infection.

The remaining lymphomas, which constitute the majority in developed countries, should probably be divided further. Some in childhood might be better classed with acute lymphatic leukaemia from which they are distinguished arbitrarily only by the number of lymphocytes in the blood. At present, however, they have to be considered as a group. As such they constitute one of the few types of cancer that have been increasing in incidence at all ages.

Two factors that have contributed to the increase, but which cannot account for it all, are the use of immunosuppressive drugs and the spread of AIDS. Intense immunosuppression is followed within 1 or 2 years by an increase in the incidence of the disease of the order of 50- to 100-fold, and smaller increases follow the less intensive use of immunosuppressive drugs for the medical treatment of patients with arthritis, Crohn's disease, and other similar conditions. Many, but not all, of the lymphomas that occur in these circumstances are associated with EBV and some of these may, unusually, arise in the brain. Greatly increased incidence rates are also seen in a variety of rare hereditary disorders characterized by major immunological impairment, such as the Wiskott–Aldrich syndrome.

The rare hairy-cell leukaemia, of unknown aetiology, is now regarded as another type of lymphoma rather than as a leukaemia.

Myelomatosis

- 1.1 per cent of all cancers and 1.6 per cent of cancer deaths.
- Sex ratio of rates 1.5 to 1. Age distribution like large bowel cancer.

Myelomatosis has been much easier to diagnose since marrow puncture and then serum electrophoresis became standard diagnostic tools and since the improvement in the management of renal failure, which is often the presenting symptom. As a result it is difficult to be sure whether the increase that was recorded until recently, in both incidence and mortality rates, was due solely to improved diagnosis, or whether it also reflects the introduction of major new causes into Europe and North America between the two World Wars. In southern Sweden, where there has been a long-standing interest in, and search for, cases of myelomatosis, no large increase was seen over the same period; the rates were higher than in other developed populations, but in recent decades those in other populations have caught up.

The disease is uncommon in undeveloped areas, where it is almost certainly underdiagnosed. Genetic factors could be important, as it is twice as common in blacks in the United States as in whites and is rare in Japanese irrespective of where they live.

Leukaemia

- 2.1 per cent of all cancers and 2.6 per cent of cancer deaths.
- Sex ratio of rates 1.5 to 1. Age distribution, see Fig. 22.

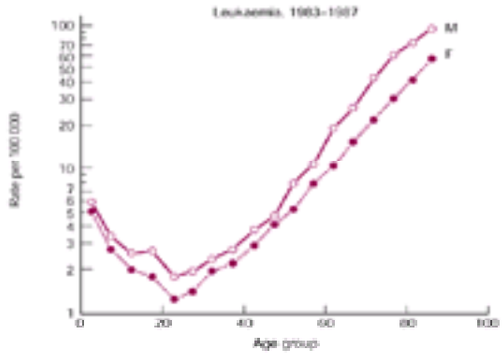

Fig. 22 Annual incidence of leukaemia by age and sex.

Leukaemia may be divided primarily into chronic lymphatic leukaemia (CLL), chronic myeloid leukaemia (CML), acute myeloid leukaemia (AML), and acute lymphatic leukaemia (ALL). CML, AML, and ALL are, in turn, amalgams of two or more different types, with different causes, different age distributions, and different prognoses, but the distinctions between them are still undergoing evolution and, with the exception referred to later, the epidemiological descriptions of each subtype are unclear.

CLL increases progressively with age in the same way as myelomatosis and most of the common epithelial cancers. It is extremely rare in Chinese, Japanese, and Indians, which is presumably due to genetic differences in susceptibility as it continues to be rare in these racial groups even when they migrate to other countries.

AML occurs at all ages. It becomes slowly, but progressively, more common from childhood on and is the most common type in young adult life. In this age group, its incidence is probably less variable throughout the world than that of any other reasonably common type of cancer. CML, by contrast, is very rare in youth, but becomes more common than AML in later middle age. The few cases that occur in childhood should perhaps be regarded as constituting a separate disease, as they lack the Philadelphia chromosome that normally characterizes CML in adult life.

ALL is the most common type of childhood cancer. Three main types can be distinguished. Common (c) ALL arises from B-lymphocyte precursors and is responsible for a peak incidence of the disease at 2 to 3 years of age. Null ALL also arises from B-cell precursors, but lacks the common antigen and accounts for most cases in the first year of life. T-cell ALL occurs more or less equally at all ages in childhood. ALL in adult life can be either B cell or T cell.

Many causes of leukaemia are known. The most important is ionizing radiation, which causes all types except CLL. The sparing of CLL may be because the relevant stem cells are so radiosensitive that they are killed by small doses that would otherwise be carcinogenic. The other main types are induced by ionizing radiation more easily than most other types of cancer and constitute about 10 per cent of all fatal cancers from exposure of the whole body to moderate doses.

Whether extremely low frequency non-ionizing radiation can cause leukaemia, particularly ALL in childhood, is uncertain. There is evidence to suggest that the risk of ALL is approximately doubled by exposure to power frequency magnetic fields of an intensity greater than 0.4 μT that occurs rarely in the United States but only very rarely in the United Kingdom. The evidence for a causal relationship is, however, inconclusive.

One type of the disease (adult T-cell lymphoma/leukaemia) is caused by a virus (HTLV/1) and has been described under non-Hodgkin's lymphoma.

Other causes include smoking, which causes a small increase in myeloid leukaemia, several chemicals, and genetically determined diseases. The most important chemical is benzene, which is used widely in industry. Prolonged occupational exposure to large amounts has caused a substantial risk of AML (particularly one of its subtypes, erythroleukaemia) and, less commonly, acute lymphatic leukaemia. Many cases are preceded by periods of aplastic anaemia and there is still some doubt whether leukaemia can be caused by small doses.

Two other chemicals, melphalan and busulphan, are used in the treatment of cancer and the small risk of AML that follows their use is unimportant in comparison to the benefit obtained if they are used appropriately. Melphalan is an alkylating agent and so presumably mutagenic. Busulfan, however, has been observed to produce a substantial risk of leukaemia only after being given in such high doses that it produces aplastic anaemia.

Of the hereditary causes, Down's syndrome is the most common and is probably responsible for the greatest number of cases, although the relative risk in some of the other rarer syndromes may be greater than the 20-fold increase in childhood leukaemia that occurs with Down's disease. Ataxia telangiectasia and Bloom's syndrome predispose to ALL, while Fanconi's anaemia predisposes to AML.

Further reading

Chen J, Campbell TC, Li J, Peto R (1990). *Diet, life-style, and mortality in China: a study of the characteristics of 65 Chinese counties*. Oxford University Press.

Dockery DW, Pope III CA, Du X, Spengler JD, Ware JH, Martha EF, Ferris BG, Speizer FE (1993). An association between air pollution and mortality in six US cities. *New England Journal of Medicine* **329**, 1753–9.

Doll R, Peto R (1981). The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *Journal of the National Cancer Institute* **66**, 1191–308. (Reprinted Oxford University Press, 1981.)

Doll R, Fraumeni J, Muir C, eds (1994). *Trends in cancer incidence and mortality. Cancer surveys*, Vol. 19 and 20. Cold Spring Harbor Laboratory Press, New York.

Hammond EC, Selikoff IJ, Seidman H (1979). Asbestos exposure, cigarette smoking, and death rates. *Annals of the New York Academy of Sciences* **330**, 473–90.

International Commission on Radiological Protection (1991). Recommendations of the International Commission on Radiological Protection. Publication 60. *Annals of the ICRP* **21**, Nos. 1–3.

Office of National Statistics (1999). *Review of the Registrar General on deaths by cause, sex and age, in England and Wales, 1998.* Stationery Office, London.

Office of National Statistics (2000). *Registrations of cancer diagnosed in 1994, England and Wales*. Stationery Office, London.

Parkin DM, Muir CS, Whelan SL, Gao Y-T, Ferlay J, Powell J, eds (1992). *Cancer incidence in five continents*, Vol. 6. International Agency for Research on Cancer, Lyon.

Peto R (2001). Cancer epidemiology in the last century and the next decade. *Nature* **411**, 390–5.

Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R (2000). Smoking, smoking cessation, and lung cancer in the UK since 1950, combination of national statistics and two case-control studies. *British Medical Journal* **321**, 323–9.

Swerdlow A, Dos Santos Silva I, Doll R (2001). *Cancer incidence and mortality in England and Wales: trends and risk factors*. Oxford University Press, Oxford

Tomatis L, ed. (1990). *Cancer causes, occurrence and control*. IARC Scientific Publications No. 100. International Agency for Research on Cancer, Lyon.

6.2 The nature and development of cancer

Andrew Coop and Matthew J. Ellis

Introduction

The disruption of proteins with pivotal roles in cell growth, death, and the regulation of gene expression is the underlying cause of cancer. The most common causes of these disturbances are somatic mutations that accumulate in cellular DNA over time, induced by chemicals in the environment (carcinogens), radiation, or simply the background rate of error in DNA replication. On occasion, carcinogenic mutations are inherited ('germline' mutations). The study of individuals with a genetic predisposition to cancer ('inherited cancer') has contributed greatly to our understanding of malignancy by pinpointing individual genes involved in this process (see Chapter 6.3). In a several instances, including Burkitt's lymphoma, Kaposi's sarcoma, cervical cancer, and hepatocellular carcinoma, cancer is initiated by viral infection, not somatic mutation ('endemic cancer'). 'Tumour viruses' encode proteins that either mimic or disrupt the functions of essentially the same set of cellular proteins whose genes are targeted by mutation in sporadic and inherited cancers.

Extensive functional and epidemiological studies of the genes responsible for cancer show that they can be broadly divided into two operational classes: oncogenes and tumour suppressor genes. Oncogenes are associated with mutant proteins demonstrating a gain in function, which overstimulate cell division or support the survival of genetically aberrant cells. In contrast, tumour suppressor genes are characterized by mutations that cause a loss of function. Typically, tumour suppressor genes encode proteins that suppress cellular proliferation, activate cell death pathways, or protect the integrity of the genome in the presence of DNA damage. In general, inactivation of both alleles of a tumour suppressor gene is required before aberrant cellular behaviour is evident. In contrast, gain-of-function mutations in oncogenes act in a dominant manner, so that typically only one allele is mutated. To qualify as an oncogene or tumour suppressor gene, there must be evidence for cancer-specific mutations in the gene concerned. Furthermore, introduction of the gene in question into appropriate recipient cells should either generate cellular responses typical of cancer cells (in the case of an oncogene) or suppress malignancy in cells that harbour defects in both alleles of the gene (in the case of a tumour suppressor gene).

In addition to cell growth and death, mutation in a oncogene or tumour suppressor gene causes disturbances in cellular physiology that ultimately lead to all the characteristics of a lethal tumour. These include an ability to acquire a blood supply (angiogenesis), tumour extension across anatomical boundaries (invasion), and growth in tissues beyond the organ of origin (metastasis). For several decades now, cancer researchers have focused on identifying the key genes responsible for these cellular processes. Their key assumption has been that the identification of so-called 'cancer genes' will lead to rational and more effective therapies. Today this is reality, with gene-targeted treatments a routine aspect of cancer management and clinical cancer research.

Cancer as a defect in cellular society

Unlike unicellular species, cells within a multicellular organism must co-operate in a way that favours the survival of the whole organism rather than that of the individual cell. If cellular co-ordination is disrupted, the unrestrained growth of even a single cell, by competing with its neighbours for space and nutrients, will eventually cause the death of the organism. Cancer is the term used to describe this breakdown in cellular society. Multicellular organisms must depend on complex gene networks that ensure full cellular co-operation. Unfortunately somatic mutation causes these networks to degrade over time, ultimately leading to cancer. Current research on the pathogenesis of cancer therefore focuses on the identification of genes that serve to maintain cellular order. Research has been greatly assisted by the recognition that cancer genes are organized into gene families that are conserved in evolution, from the simplest worm through to *Homo sapiens*. This observation has profoundly influenced the investigation of human cancer because the results of gene manipulation in experimentally tractable lower organisms, such as the nematode worm *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, mice, yeast, and even bacteria, have direct implications for their human homologues.

Cellular transformation

In many instances cancer cells can be grown in tissue culture, which has greatly facilitated the study of this disease. Unlike their healthy counterparts, cancer cells have been found to have some or all of the following *in vitro* characteristics:

1. Reduced requirement for growth stimulatory molecules, termed 'growth factors'.
2. Loss of contact inhibition, so that the cells tend to pile up and form foci.
3. Anchorage-independent growth, usually manifest as an ability to grow in soft agar.
4. The ability to divide indefinitely, a characteristic termed 'immortalization'.

When a cell has acquired these *in vitro* characteristics, it is referred to as 'transformed'. Empirically, cellular transformation can be achieved in human cells in tissue culture by disabling or disrupting the activities of four cellular components (Fig. 1). However, the process of becoming a cancer cell begins, not ends, with transformation. Not all experimentally transformed cells are capable of establishing a tumour when transplanted into a suitable host, and fewer still are capable of metastasis. The process of becoming a lethal, angiogenic, invasive, and motile cancer involves the activities of additional families of genes that serve these cellular programmes. For research purposes, the pathways involved in cancer tend to be categorized along functional lines, i.e. transformation, genetic instability, aberrant growth and survival, angiogenesis, tissue invasion, and cellular motility. While this provides an excellent framework for discussion, these distinctions are somewhat artificial because many of the genes implicated in cancer serve not one but several of these processes.

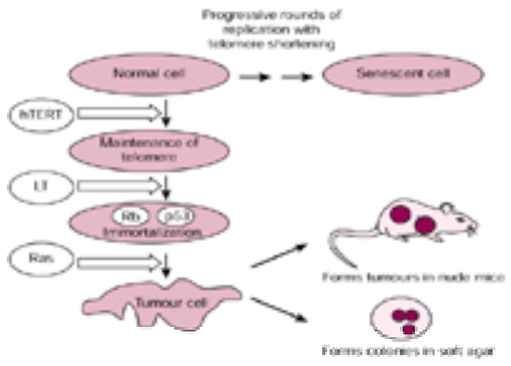

Fig. 1 *In vitro* experiments demonstrate that at least four events are required to convert a normal human cell into a tumour cell. Addition of the catalytic hTERT subunit of telomerase (event 1) prevents telomere shortening (normal telomerase expression results in progessive telomere shortening, until further replication cannot be sustained and the cells become senescent). Introduction of SV40 large T antigen (LT), an oncoprotein from simian virus 40, disrupts the Rb and p53 protein pathways, which are essential for normal cell proliferation control (events 2 and 3). Finally, introduction of the oncogene *ras* is sufficient to cause malignant transformation (event 4). That the cells are indeed malignant can be demonstrated by their ability to form tumours in nude mice and colonies in soft agar.

Somatic mutation and clonal evolution

The theory that cancer is a multistep or multigene process is supported by epidemiological evidence. The incidence of common cancers increases as individuals age, with kinetics dependent on the fourth or fifth power of elapsed time. This observation implies that a minimum of four to five events must take place before a tumour is evident clinically. If each event represents a somatic mutation, how does tumour evolution occur at the cellular level? Estimates of the rate of somatic mutation in normal tissues indicate that for any gene there is a one in a million chance of a somatic mutation each time a cell divides. The baseline somatic mutation rate is therefore very low and a somatic mutation that significantly affects the function of a 'cancer gene' must be a very rare event. In fact, the background rate of somatic mutation appears to be too low to be the driving force behind cancer. However, carcinogenic mutations favour the growth or survival of a cell at the expense of neighbouring cells. As a result, cancer-promoting mutations are subject to a powerful positive selection process that magnifies the impact of these low-frequency events. In the model depicted in Fig. 2, a mutation has occurred that stimulates growth. As a result, a clone of cells has arisen that has replicated the initial mutation thousands of times. As the size of the clone approaches 10^6 cells, the chance of a second growth-stimulating mutation in one of the cells in the clone increases significantly. When a second mutation does occur, the cell with two growth-promoting mutations begins to outgrow the original clone. In this way a tumour evolves continuously, with each successive mutation adding a new facet to the repertoire of cellular properties required for a cell to be malignant. This process continues after diagnosis, with treatment resistance also driven by somatic mutation. Evidence for the 'clonal selection ' theory of cancer abounds in the literature. For example, in Barrett's oesophagitis, where normal tissue, premalignant lesions, and cancers often coexist in the distal segment of the oesophagus, genetic analysis of oesophageal biopsies over time reveals multiple competing cell clones evolving simultaneously, each with a different complement of somatic mutations.

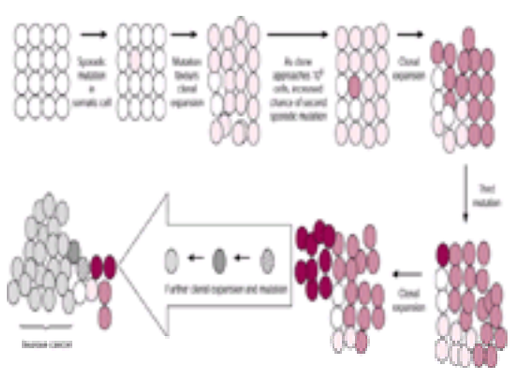

Fig. 2 Sporadic mutations which give a somatic cell a selective growth advantage encourage the outgrowth of mutated clones. Successive clones continue to accumulate growth-favouring mutations and outgrow preceding clones. The process of clonal expansion and mutation eventually results in a clone of cells with a malignant phenotype (invasive cancer). Even after the emergence of a malignant clone, the process of clonal expansion and mutation continues.

Gatekeepers and caretakers

As already pointed out, cancer-inducing mutations are rare events and despite the amplifying effect of clonal selection, the rate of somatic mutation in normal cells may still be insufficient to generate cancer at a high frequency. However, the evolution of cancer is also accelerated by 'genetic instability'. This term implies an increase in the rate of mutation in cancer cells when compared with normal cells, not simply that cancer cells possess a large number of mutations. A detailed mutational analysis of human tumours indicates that there are at least two general categories of genetic instability. In a minority of cancers there is instability at the nucleotide level, with a higher rate of nucleotide substitution, deletion, or insertion than in normal cells. More commonly, instability is observed at the level of the chromosome, with large-scale deletions, duplications, translocations, and amplification of entire chromosomal regions (Fig. 3 and Plate 1).

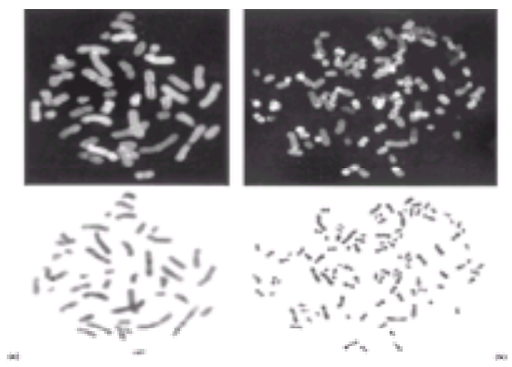

Fig. 3 In a spectral karyotype (SKY), each normal chromosome stains homogeneously with a single distinct colour, making translocations evident by the presence of more than one colour in a single chromosome. (a) Forty five chromosomes are visible in this karyotype (left) from a patient with Turner's syndrome. Despite the loss of the X chromosome, the spectral karyotype (right) clearly shows the homogeneous chromosomal staining pattern typical of normal chromosomes. (b) In contrast, SKY analysis of a metaphase spread prepared from a breast cancer cell line displays both numerical and structural chromosomal aberrations. (By courtesy of Dr Bassem Haddad, Department of Oncology, Georgetown University Medical Center.) (See also Plate 1.)

How does genetic instability arise? The integrity of DNA is protected by two classes of genes referred to as 'caretaker' and 'gatekeeper' genes. Caretaker genes are involved in recognizing and/or participating in the repair of damaged nucleotides or DNA strand breaks. Gatekeeper genes have a cell cycle arrest function that is triggered by DNA damage. Cell cycle arrest is coupled to DNA damage so that caretaker proteins can repair DNA before replication is reinitiated. If gatekeeper functions fail, replication continues through tracts of damaged DNA, generating mutations that are passed on to daughter cells, courting disaster in the form of

cancer-promoting mutations.

Gatekeepers function through 'checkpoint' signal transduction pathways that operate at transition states in the cell cycle, when the cell normally pauses for an analysis of DNA integrity and an integration of internal and external signals that either promote or inhibit progression of the cell cycle. Activation of gatekeeper pathways can prevent the onset of DNA synthesis (G_1 checkpoint) or entry into mitosis (G_2 checkpoint) or prevent completion of chromosomal segregation after chromosomes are aligned on the mitotic spindle (spindle checkpoint). Importantly, gatekeepers can also trigger cell death, so that if genetic damage is so extensive that it cannot be repaired, or the repair process fails in some way, genetically altered cells with the potential to become malignant are deleted. Cancer cells that possess both gatekeeper and caretaker defects can have astonishing rates of genetic instability and often display rapid progression towards a lethal tumour phenotype.

Genetic instability at the nucleotide level

Subtle nucleotide changes are frequently observed in cancer cell genomes and cause 'gain-of-function' or 'loss-of-function' mutations in oncogenes and tumour suppressor genes. The majority of these mutations are probably not due to defects in gatekeeper and caretaker functions, but reflect the impact of environmental carcinogens or the background rate of somatic mutation. Under two circumstances, however, defective DNA repair leads to errors in nucleotide replication at an abnormally high rate. Nucleotide excision repair is responsible for restoring the correct DNA sequence after damage by exogenous mutagens, particularly ultraviolet light. Individuals with inherited nucleotide excision repair defects (xeroderma pigmentosum, see Chapter 6.3) have a marked increase in the incidence of skin cancer. Interestingly, the incidence of internal cancer is not increased to the same degree as in the skin. Furthermore, xeroderma pigmentosum heterozygotes are not at increased risk for cancer. Thus, nucleotide excision repair appears to be relevant to genetic instability only in the rare circumstance of an inherited cancer predisposition syndrome.

In contrast to nucleotide excision repair, mismatch repair defects accelerate the mutation rate in both hereditary and sporadic cancers of the colon, stomach, and endometrium. Conclusions concerning mismatch repair defects originated from the observation that a group of colorectal tumours have frequent alterations in short polynucleotide tracts in their genomes ('microsatellites') and the term 'microsatellite instability' arose to refer to this type of error. The similarity of these changes to mismatch repair defects in bacteria and yeast led to the identifications of six human homologues of bacterial mismatch repair genes (*mutS* and *mutL*) which, when inactivated by mutation, cause human tumours with microsatellite instability. A mismatch repair defect can be detected in about 13 per cent of all colorectal, stomach, and endometrial cancers, and in virtually all tumours arising in individuals with inherited *mutS* and *mutL* defects (hereditary non-polyposis colon cancer). Microsatellite sequences can be found throughout the genome, frequently in non-coding regions. It is noteworthy that some genes with roles in growth suppression have been found to be mutated at microsatellite sequences within their coding sequences (Table 1).

Instability in chromosome number

In contrast to the relative rarity of nucleotide excision repair and mismatch repair defects, another form of genetic instability, characterized by gains and losses of large segments of chromosomes, occurs in most human tumours. True chromosomal instability appears to underlie this observation, since the rate of losses and gains of chromosomal material in cancer cells with aneuploid genomes (more or less DNA than the normal diploid DNA complement) are at least tenfold higher than in cells with a normal chromosomal complement. The 'average' aneuploid cancer cell of the colon, breast, or prostate loses up to 25 per cent of its alleles, with the figure exceeding 50 per cent in some cases. Interestingly, in colon cancer and endometrial cancer, there is an inverse relationship between chromosomal instability and mismatch repair defects, suggesting that the two routes to genetic instability are somewhat mutually exclusive. In both cases, the hallmarks of genetic instability can be detected early in tumorigenesis and increase markedly over time, as each wave of cell clones arises bearing a new set of mutations, with a relentless progression in tumour size, aggression, and the degree of genomic disorganization. The molecular basis of chromosomal instability is heterogeneous, and unlike mismatch repair defects, no unifying mechanism has been proposed. However, defects in several cell cycle checkpoints have been implicated. In one example, an aberrant mitotic spindle checkpoint is proposed in which an abnormal mitosis is tolerated despite the presence of a lagging chromosome, causing an imbalance in chromosomal segregation. Several 'spindle checkpoint' genes have been isolated, and in the case of *hBUB1* or *hBUBR1*, somatic mutations in tumours have been detected. Defects in a second checkpoint, referred to as the 'DNA damage checkpoint', is probably a more frequent cause of chromosomal instability. As discussed earlier, DNA damage checkpoints exist because gross structural alterations in chromosomes will occur if DNA replication occurs in the presence of single or double DNA strand breaks. As with mismatch repair defects and nucleotide excision repair, many of the genes within the 'DNA strand break pathway' have been identified through an association with an inherited predisposition to cancer and are further discussed below. A third, and by no means final, potential mechanism concerns disruption of centrosome function. Centrosomes nucleate the ends of the mitotic microtubule spindle along which the sister chromosomes segregate into two daughter cells. An abnormal number of centrosomes (i.e. more than two) have been observed in a variety of common cancers. When more than two poles form during mitosis, the potential for abnormal chromosomal segregation is high. Recently, overexpression of two centrosome-associated kinase genes, *STK15* and *PLK1*, both homologues of *Drosophila* genes known to regulate centrosome function, have been associated with an increase in centrosome number and defective chromosome segregation in human cancer cells. Despite considerable insights into these complexities, it is fair to state that for the majority of cancers the cause of aneuploidy is unclear. There are a number of pathways that lead to chromosomal instability, and perhaps this is the reason why chromosomal instability is so common in human tumours.

Chromosome translocation

There are two types of chromosomal translocations in human cancers. The first type is common and is without any clear pattern of repetition within tumours of the same histopathological type. The driving force behind this so-called 'complex' type of translocation is probably the same set of gatekeeper and caretaker defects that are a frequent cause of chromosomal instability. A favoured postulate is that these complex translocations reflect cells that enter mitosis before double-strand breaks are repaired, leading to random joining of free DNA ends through non-homologous recombination. The second pattern of translocation is referred to as the 'simple' type. These are common in leukaemia, lymphomas, and in a subset of sarcomas. The breakpoints in these translocations have been analysed and several distinct types of molecular effect have been identified. For example, when an oncogene is repositioned next to transcription regulatory sequences, overexpression occurs. In other cases, translocation generates a fusion gene that combines a coding sequence from the two genes at each end of the break point (Table 2).

Genetic instability is not thought to be the cause of simple translocations. In lymphoid cells, DNA strand breaks are generated as part of the normal recombination process that generates diversity in immunoglobulin genes and the T-cell receptor. The translocations characteristic of lymphoid malignancies are therefore thought to reflect low-frequency aberrations in physiological recombination events. The genesis of simple translocations in sarcoma is more enigmatic, however, as physiological gene rearrangement does not occur in the soft tissues from which sarcomas arise.

Gene amplification

Gene amplification occurs in late stage cancers and almost certainly reflects an aspect of genetic instability, because these aberrations occur at a higher rate in cancer cells than normal cells. Gene amplification is also thought to represent a relatively late step in the evolution of cancer and can lead to massive overexpression of oncogenes. It has been argued that the key defect underlying amplification concerns an inability to trigger a cell suicide or 'apoptosis' signal that usually deletes cells with amplified chromosomal segments. Table 3 lists a series of genes that are commonly amplified in human cancers.

Cell cycle deregulation

The cell cycle is governed by the activities of cyclin/cyclin-dependent protein kinases that oscillate through the cell cycle, orchestrating the complex process of cell division (Fig. 4). The activities of cyclin and cyclin-dependent protein kinase are closely regulated through checkpoint pathways that exist to monitor the integrity and replication status of DNA. Genes encoding proteins that promote cell cycle progression often operate as oncogenes, and are subject to activation in human cancers through gain-of-function mutations or gene amplification. In contrast, genes encoding the components of checkpoint pathways are tumour suppressors that may be inactivated during tumorigenesis. The importance of the restraining influence of cell cycle checkpoints is underscored by the finding that the tumour suppressor genes involved are amongst the most frequently mutated in human cancer.

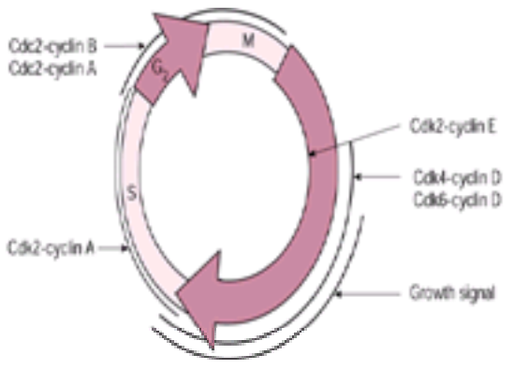

Fig. 4 Following growth stimulation, resting cells (G_0) move into and through the cell cycle. Cycle progression is tightly regulated by the temporal expression of cyclins and cyclin-dependent kinases, which are not expressed in the resting state.

The G_1/S checkpoint

Transition through the G_1/S checkpoint is promoted by the cyclin D family and their partner kinases, cyclin-dependent protein kinase 4 and 6 (Fig. 5). The cyclin D/cyclin-dependent protein kinase complex ('Rb kinase'), after activation by a cyclin-dependent protein kinase-activating kinase (**CAK**), phosphorylates the retinoblastoma tumour suppressor gene product, Rb. An increase in Rb phosphorylation releases a set of transcription factors termed E2F. E2F proteins stimulate the expression of genes required for the S phase, for example the nucleotide biosynthetic genes dihydrofolate reductase and thymidine kinase. E2F also activates the functions of cyclins A and B to promote further cell cycle progression. Finally, E2F activates a negative feedback loop through a gene called *CDKN2A*. The p16*CDKN2A* protein inhibits Rb kinase by interfering with CAK activity and disrupting the cyclin-dependent protein kinase 4/Rb complex. Mutations in these regulatory genes are frequently present in cancer cells, underscoring the importance of this pathway in the regulation of normal cell growth. Gain-of-function mutations may occur in cyclin-dependent protein kinase 4, or cyclin D may be subject to gene amplification, deregulating Rb kinase activity and increasing E2F-dependent transcription. Loss-of-function mutation in *Rb* or *CDKN2A*, both common events in tumorigenesis, also result in deregulated E2F.

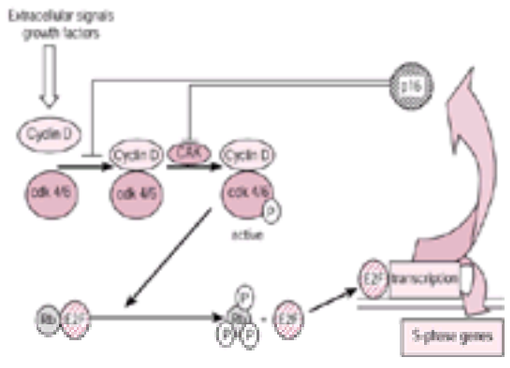

Fig. 5 Growth factor stimulation of cells promotes accumulation of cyclin D and the assembly of cyclin D and cyclin-dependent protein kinase 4 or 6 into complexes. Phosphorylation of the cyclin-dependent protein kinase component by CAK activates the complex, which phosphorylates the Rb protein. Phosphorylated Rb releases a transcription factor of the E2F family. E2F drives transcription of genes required for cell cycle progression as well as the gene *CDK2NA,* from which the inhibitory protein p16*CDKN2A* is translated. In a negative feedback mechanism, p16 disrupts activity of CAK and cyclin D/cyclin-dependent protein kinase complexes, marking the completion of the initial phase of the cell cycle.

The G_1/S checkpoint is also under strict control from a gatekeeper pathway activated by DNA strand breaks. The components of this pathway have recently been put together as a result of research on a series of inherited cancer predisposition syndromes (Fig. 6). Perhaps the most critical component of the DNA strand break pathway is the tumour suppressor gene *p53*. Mutations in *p53* are possibly the most frequent genetic lesion in cancer cells. This, and the critical role of p53 in the DNA damage pathway, has inspired the name 'the guardian of the genome'. In fact, *p53* is a member of a regiment of 'genome guards' that act in concert to protect the integrity of DNA. At the apex of the DNA strand break pathway is the ataxia telangiectasia gene product ATM. ATM is a phosphatidylinositol-3 kinase that is activated by double-strand DNA breaks and orchestrates the function of proteins that repair DNA and arrest the cell cycle. To achieve this, ATM directs the phosphorylation of a series of targets, including p53, CHK2, NBS1, and BRCA1. When p53 is activated through the ATM-dependent kinase CHK2, the expression of p53-regulated genes is greatly increased. It is the spectrum of p53-regulated genes that actually conduct the caretaker and gatekeeper functions associated with p53. For example, p53 activates the expression of ribonuclease reductase, the enzyme responsible for a rate-limiting step in the production of deoxyribonucleotide triphosphates, 'sending' nucleotides to repair DNA. The G_1/S checkpoint function of p53 is mediated by *CDKN1* gene activation. The p21*CDKN1* protein acts in a similar fashion to p16*CDKN2A*, inhibiting Rb kinase and preventing E2F activation. p53 is also instrumental in activating cell death after DNA damage. This is achieved in part by activating the expression of the cell death protein BAX (see below). *p53* mutations are therefore a major contributor to DNA instability since the cell continues to replicates in the presence of DNA strand breaks and fails to undergo cell death after DNA damage.

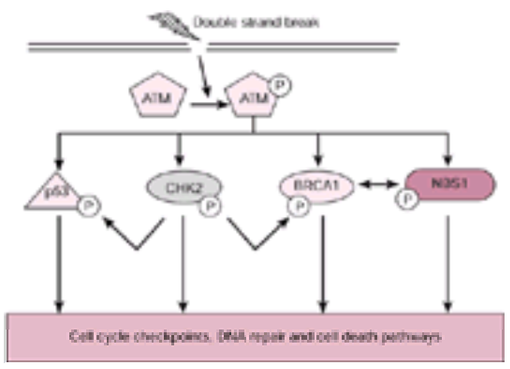

Fig. 6 Double-strand DNA breaks result in the activation of ATM. ATM activates, by phosphorylation, a variety of proteins (including p53, CHK2, BRCA1, and NBS1) whose checkpoint and repair functions prevent propagation of DNA damage. Each of the depicted proteins was identified as a product of a gene mutated in familial cancer syndromes. p53 (Li–Fraumeni syndrome), also activated by the ATM-dependent kinase CHK2 (Li–Fraumeni), increases the expression of genes including p21*CDKN1* and *BAX*. Interaction between BRCA1 (familial breast and ovarian cancer) and NBS1 (Nijmegen breakage syndrome) may co-ordinate DNA repair.

The G_2/M checkpoint

Normal cells arrest in G_2 in the presence of DNA damage. Cells mutant for p53 do not maintain arrest at G_2 and enter mitosis in the presence of DNA strand breaks or other types of chromosomal damage. A major component in sustaining p53-dependent G_2 arrest is the p53-regulated signalling protein 14-3-3s, an inhibitor of the cyclin B/cyclin-dependent protein kinase 2 complex required for the initiation of mitosis. The integrity of these biochemical events is thought to have a major impact on the action of cytotoxic drugs, because agents that disrupt the G_2 checkpoint selectively potentiate the cytotoxic effects of DNA-damaging agents on p53 mutant cancer

cells.

The discovery of oncogenes

A major feature of cancer cells is a loss of dependence on growth factor stimulation as a result of mutations in growth factor signal transduction pathways. Signal transduction genes were the first oncogenes identified, initiating the genetic revolution in cancer research in the late 1970s and early 1980s. In experiments initiated by Peyton Rous, retroviruses were isolated from chicken and rodent tumours that were capable of rapidly inducing cancer in infected animals. DNA sequence analysis of these 'acutely transforming' retroviruses showed that the transforming activity was due to a growth-stimulating 'oncogene' of non-viral origin that had been picked up in a rare recombination event from a host cell. The study of acutely transforming retroviruses led to the identification of a family of oncogenes that when subjected to gain-of-function mutations or overexpression induce aberrant signal transduction, cellular transformation, and tumour formation (Table 4).

Plasma membrane receptors as oncogenes

Ligand-activated, tyrosine kinase-linked plasma membrane receptors (receptor tyrosine kinases, **RTKs**) that operate as oncogenes include erbB2 and epidermal growth factor receptor. Amplification of the genes for these receptors occurs in a wide spectrum of common malignancies, including breast, lung, pancreatic, and head and neck cancer. Receptor overexpression reduces or bypasses the requirement for the presence of a ligand, removing dependence on external growth signals. For example, erbB2 normally signals as a heterodimer in partnership with a second member of the erbB2 family, because erbB2 cannot bind to a ligand directly. When overexpressed, erbB2 is forced into homodimers that are active in the absence of ligand. The RTK oncogenes c-*met* and c-*ret* are involved in a more limited spectrum of tumours, including papillary renal cancer (c-*met*) and medullary thyroid cancer (c-*ret*). In these cases, missense mutations introduce cysteine residues in the extracellular binding domain, resulting in inappropriate disulphide bond formation, dimerization, tyrosine kinase activation, and ligand-independent growth.

Plasma membrane receptors as tumour suppressor genes

The transforming growth factor b (**TGF-b**) pathway provides an example of a plasma membrane receptor that suppresses cell growth (Fig. 7). Transforming growth factor b is a peptide growth factor with complex effects on both tumour cells and on host stromal cells. Responses to TGF-b include inhibition of tumour growth, induction of cell death, extracellular matrix synthesis, and angiogenesis. Transforming growth factor b signals through a heterodimer of two receptors, TGF-b RI and TGF-b RII, both plasma membrane serine–threonine kinases. TGF-b RII binds TGF-b and then dimerizes with and phosphorylates TGF-b RI. TGF-b RI in turn phosphorylates the signal transduction proteins Smad2 and Smad3. Cytoplasmic Smad proteins then migrate to the nucleus to activate gene expression in concert with a third Smad protein, Smad4. One of the genes activated is p27*kip1*, a cyclin-dependent kinase inhibitor in the same family as p21 *CDKN1* and p16*CDKN2A* with roles in both cell cycle arrest and the induction of programmed cell death. Loss-of-function mutations, characteristic of tumour suppressor genes, occur in the TGF-b/Smad pathway. TGF-b RII is inactivated in 20 per cent of colorectal cancers with mismatch repair defects because of a vulnerable poly(A) microsatellite in the TGF-b RII coding sequence (Table 1). In addition, Smad4 is inactivated in up to 50 per cent of pancreatic cancers.

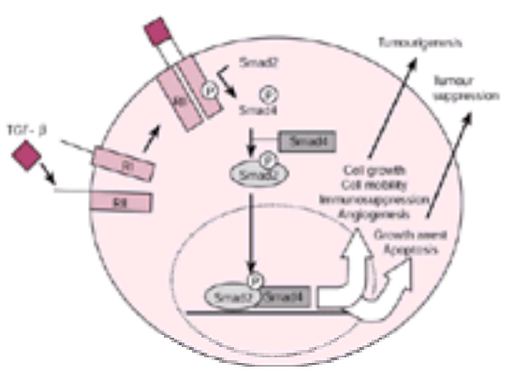

Fig. 7 On TGF-b binding to TGF-b RII, this receptor dimerizes with and phosphorylates TGF-b RI. Activated TGF-b RI phosphorylates the cytoplasmic proteins Smad2 and Smad 3 (not shown), which in turn interact with Smad4. The Smad complex migrates to the cell nucleus where, in concert with various transcription factors, it drives transcription of genes involved in both tumorigenesis and tumour suppression.

Cytoplasmic signal transduction components as oncogenes (*ras*)

When an RTK is activated by dimerization, the cytoplasmic portion of the receptor becomes autophosporylated and a variety of intracellular docking proteins are recruited to the cell membrane. These docking proteins create a scaffold in the inner surface of the plasma membrane around which signalling components congregate further downstream. Critical amongst these components is the 'Ras' family of proteins (K-, N-, and H-Ras). *ras* genes are frequently mutated in human cancer and at least partially bypass the need for RTK signalling in cell growth. Ras proteins operate a GTP-dependent switching mechanism that alternates between a GTP-bound active form and a GDP-bound inactive form. When an RTK is activated, docking proteins, together with a protein called SOS, stimulate the release of GDP from Ras, replacing it with GTP. Activated, GTP-bound Ras then signals to a host of downstream targets to stimulate multiple cellular changes, including activation of DNA synthesis, and changes in lipid metabolism, cellular morphology, cell adhesion, and gene expression (Fig. 8). The system is switched off by GTPase-activating proteins that stimulate the hydrolysis of Ras-bound GTP to GDP. Mutations in *ras* genes cause the protein to be held in the GTP-activated form, generating a continuous unregulated signal. K-*ras* and N-*ras* mutations are found in common solid tumours including lung, brain, colon cancer, and nearly 95 per cent of pancreatic cancers and 30 per cent of acute leukaemias. H-*ras* mutations are found only in a small subset of bladder and head and neck tumours.

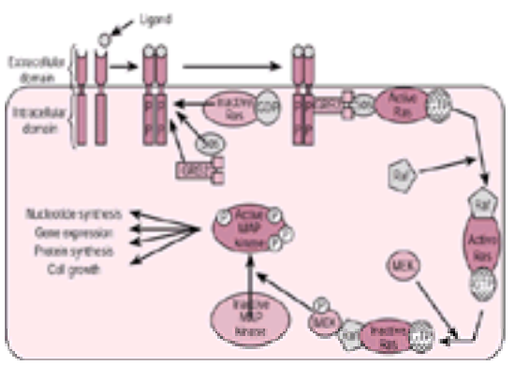

Fig. 8 Dimerization and activation of most cell surface receptor tyrosine kinases (RTKs) is dependent on ligand binding. The ligand-activated receptor autophosphorylates specific tyrosine residues in the cytosolic domain. These phosphotyrosine residues act as binding sites for a variety of proteins. GRB2 binds to both activated RTK and a protein called SOS. SOS facilitates the exchange of GDP for GTP, converting inactive Ras/GDP to active Ras/GTP. Active Ras binds to Raf (a serine/threonine kinase), and this complex binds to and phosphorylates the kinase MEK. MEK phosphorylates MAP kinase; MAP kinase subsequently activates proteins important for cell growth.

Cytoplasmic signal transduction components as tumour suppressor genes

Neurofibromatosis is caused by germline mutation of the tumour suppressor neurofibromin (**NF1**), a GTPase-activating protein. Loss of NF1 increases Ras activity; however, NF1 is not commonly targeted for mutation in common solid malignancies. The oncogenic consequences of NF1 inactivation are tissue specific since patients with type 1 neurofibromatosis do not have a marked increase in the incidence of common solid malignancies. The adenomatous polyposis coli (**APC**) gene

product is a stronger example of a cytoplasmic tumour suppressor gene that operates to downregulate a growth regulatory pathway (Fig. 9). The *APC* gene is mutated in most human colon cancers and germline mutations are associated with familial colon polyps and cancer. APC normally regulates the signal transduction protein b-catenin by targeted degradation, so that in an APC mutant cell, b-catenin accumulates abnormally. b-catenin is an essential component of a signal transduction pathway that connects cell adhesion and the extracellular matrix to the cell nucleus. b-catenin operates by activating the TCF/LEF family of transcription factors, leading to gene transcription and cell cycle progression. For example, in APC mutant cells, overexpression of cyclin D1 occurs because of deregulated b-catenin /LEF activity. b-catenin is also subject to somatic mutation, particularly in colon cancers without germline APC mutations, as well as melanoma, prostate cancers, gastric cancer, hepatocellular carcinoma, and medulloblastoma. Mutations in the b-catenin gene tend to stabilize the b-catenin protein, increasing TCF/LEF activity and cyclin D expression.

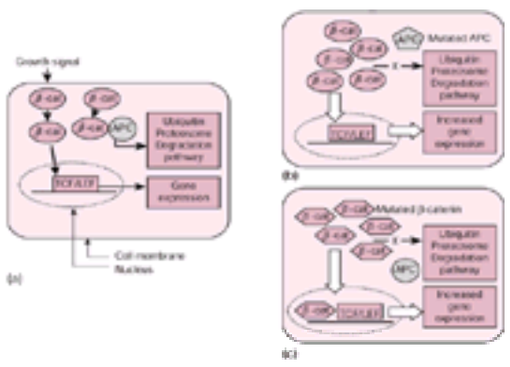

Fig. 9 (a) In the presence of a growth signal, cytoskeleton-associated b-catenin migrates from the inner plasma membrane to the nucleus, where it increases the activity of TCF/LEF transcription factors. APC binds b-catenin and targets it for degradation by the ubiquitin pathway. (b) Certain APC mutations disrupt interaction between APC and b-catenin. This prevents ubiquitin targeting, so that b-catenin accumulates and migrates to the nucleus, where the increased levels drive excessive transcription of TCF/LEF-regulated genes. (c) In a similar manner, certain b-catenin mutations disrupt interaction with functional APC, with the same consequences.

Transcription factors as oncogenes and tumour suppressor genes

Transcription factors regulate the rate of transcription by binding to specific DNA motifs in non-coding sequences, thereby recruiting a multicomponent protein complex that promotes (or represses) the activity of RNA polymerase. A broad spectrum of mutations activates the oncogenic potential, or disrupts the tumour suppressor functions, of transcription factor genes. In examples discussed already, point mutations inactivate or alter the functions of the transcription factors p53, b-catenin, and Smad4. In other instances, transcription factor genes are subject to activation through gene amplification, including c- *myc*, a gene required for cell cycle progression and cyclin expression, and amplified in breast cancer-1 (*AIB1*), a gene encoding a protein that boosts the transcriptional activity of steroid receptors.

Perhaps the most remarkable examples of aberrant transcription factors in cancer result from translocations that create transcription factor gene fusions (Table 2). For example, in promyelocytic leukaemia, the retinoic acid receptor a gene (*RARa*) on chromosome 15 is involved in a translocation with a gene on chromosome 17 termed *PML*. The resulting fusion transcription factor, RARa/PML, demonstrates aberrant localization within the cell nucleus where it acts as a 'decoy', binding normal retinoic acid receptor family members and inhibiting their role in myeloid differentiation. High doses of retinoic acid induce remission of PML by inducing the degradation of the RARa/PML fusion protein, thereby releasing normal retinoic acid receptors to activate the terminal differentiation of leukaemic cells. Ewing's sarcoma family of tumours provides further examples of transcription factor gene fusion, in this instance between the *EWS* gene and the ETS family of transcription factors, either FLI1 or ERG (95 per cent and 5 per cent of Ewing's tumours respectively). In these cases, the fusion ETS transcription factor becomes hyperactive because of the anomalous presence of *EWS* gene sequences. In a further example of transcription factor overactivity, the c- *myc* gene becomes overexpressed in Burkitt's lymphoma as a result of an 8®14 translocation that places the c- *myc* gene under the influence of immunoglobulin gene sequences.

Mutations in cell death pathways (Fig. 10)

Programmed cell death or apoptosis is an essential aspect of the ability of a multicellular organism to develop and survive (see Chapter 4.6). For example, during embryogenesis the targeted removal of cells in tissue remodelling is achieved through cell death programmes, generating digits and body cavities. In the immune system, apoptosis deletes lymphocytes that recognize 'self-antigens', thereby preventing the growth of B- and T-cell clones capable of autoimmune damage. Apoptosis is also essential for the deletion of cells with DNA damage, chromosomal aberrations, and abnormalities of mitotic spindle formation. These events stimulate the release of cytochrome C from mitochondria. Cytoplasmic cytochrome C activates a cascade of 'caspase' proteases that effect DNA fragmentation, plasma membrane destruction, and the characteristic morphological features of apoptosis (see *section 4.6*). Receptor tyrosine kinases play a critical role in cell survival by activating the signal transduction enzyme phosphatidylinositol-3-kinase. Phosphatidylinositol-3-kinase in turn activates a serine threonine kinase, AKT (an oncogene, discovered through the study of acutely transforming retroviruses). AKT promotes survival by phosphorylating and inactivating BAD, a protein that promotes mitochondrial release of cytochrome C. Excess phosphatidylinositol-3-kinase activity and aberrant survival are also consequences of loss-of-function mutations in a tumour suppressor termed PTEN (phosphatase tensin homologue deleted in chromosome 10). The *PTEN* gene encodes a protein phosphatase mutated in a wide spectrum of cancers, including breast, thyroid, lung, and a small proportion of lymphomas. When PTEN is lost through somatic mutation, AKT becomes constitutively active, holding the cell death-promoting protein BAD in an inactive conformation. The activity of BAD is also tightly controlled through the BCL2 family of BAD dimerization partners. BAD and BCL2 are members of a family of at least 15 proteins that promote or suppress cell death. BCL2 itself is an oncogene activated by a translocation that increases expression in indolent lymphoma. BCL2 is a very potent cell survival factor that binds to BAD (and other proapoptotic partners, BAX, BID, and $BCLX_L$) to prevent cytochrome C release, and when overexpressed BCL2 potently inhibits cell death. The tumour suppressor gene *p53* is also a key regulator of apoptosis. When *p53* is activated, the transcription of the proapoptotic gene *BAX* is increased, also promoting mitochondrial cytochrome C release after DNA damage.

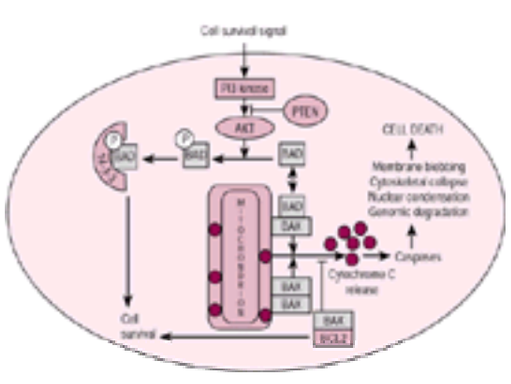

Fig. 10 Apoptosis is triggered by cytochrome C release from the mitchondrion which activates caspase proteases that destroy the cell. Cytochrome C release is therefore a tightly regulated process that is under the control of several signal transduction pathways that operate through the BCL2 family of cell death regulators.

Genetic programmes serving tissue invasion

Tissue invasion is a complex process that, like all other aspects of the malignant phenotype, is driven by a complex array of genetic lesions. Like many of the pathways described thus far, the molecular details of the cancer invasion pathway are incompletely resolved. Since non-invasive tumours are essentially benign, how tumour cells acquire an invasive phenotype is at the core of understanding what makes a tumour benign or malignant. Invasion requires tissue degradation and

several classes of enzyme have been identified that are expressed by invasive tumours, including metalloproteinases and collagenases. Inhibitors of these enzymes are currently in clinical trial as antimetastasis or anti-invasion agents.

Another important aspect of invasion is increased cell motility, required for a malignant cell to migrate away from its site of origin. Excess growth factor signalling, typical of cancer cells, promotes motility, and specific motility genes, for example the *C-MET* 'scatter factor' receptor, are subject to activation through somatic mutation. Building on the concept of cell cycle checkpoints, a checkpoint for invasion has recently been proposed based on a receptor–ligand pair termed amphoterin/RAGE (receptor for advanced glycation end products). Amphoterin assists in the generation of the protein-degrading protein plasmin, which then activates metalloproteinases. RAGE is the receptor for amphoterin, which promotes, through MAP kinase, motility, 'adhesion' receptors, and growth pathways. While genetic lesions in this pathway have yet to be described, the outline of a motility and invasion pathway is emerging. As part of this developing concept, 'invasion suppressor genes' are being identified. Mutations in the E-cadherin gene have been described in breast, colon, and gastric cancers and are associated with changes in cell morphology, increased motility, and activation of the b-catenin/LEF transcriptional pathway discussed in the context of the *APC* tumour suppressor gene.

Genetic programmes serving angiogenesis

Angiogenesis is also a critical aspect of the cancer phenotype. Due to the limitations of tissue diffusion, a tumour cannot increase beyond 1 mm in size without a blood supply. Exactly how a tumour manipulates the process of angiogenesis is beginning to be understood, and once again mutations in regulatory genes are at the root of the issue. Normal cells respond to hypoxia by increasing the expression of a set of hypoxia-inducible genes, of which that for vascular endothelial growth factor (**VEGF**) is a prime example. VEGF profoundly stimulates the growth of the endothelial cells that line blood vessel walls. The transcription factors concerned, hypoxia-inducible factors 1a and 2a (**HIF1a** and **HIF2a**), are tightly regulated by oxygen tension. Unlike normal cells, tumour cells may show HIF activity even when oxygen tension is adequate, i.e. the connection between hypoxia and HIF activity is severed. In tumours that arise in the cancer predisposition syndrome von Hippel–Lindau syndrome, as well as in sporadic renal cell carcinoma, uncoupling of hypoxia-inducible gene expression from hypoxia is due to loss-of-function mutations in the von Hippel–Lindau gene, *VHL*. The VHL protein normally targets HIF1a and HIF2a for degradation. When VHL activity is lost, excess HIF1a and HIF2a activity drives increased VEGF expression and excessive blood vessel formation. Of course, aberrant angiogenesis also occurs in tumours in which VHL is not mutated. VEGF is only one of many angiogenic factors. For example, fibroblast growth factors are also angiogenic, and the fibroblast growth factor-3 gene is subject to gene amplification in breast cancer, squamous cell carcinoma of the head and neck, and nasopharyngeal cancer.

Human cancers caused by infection

Up to 20 per cent of cancers worldwide may be due to viral infection. The discovery of infectious agents that induce cancer has proved invaluable for the identification of genes implicated in tumorigenesis. The insights gained from the study of acutely transforming retroviruses (RNA tumour viruses) were discussed earlier. As the name implies, these viruses are capable of causing tumours in the affected species within a few weeks to months of infection. As far as we know, such viruses are not a cause of human cancer. However, endemic retroviruses are strongly implicated in several malignant diseases of humans. Unlike the acutely transforming retroviruses, endemic retroviruses do not contain host (human) gene sequences and are associated with cancer only after a long latency period of years to decades. The most prominent example is the HTLV-I virus, associated with endemic T-cell leukaemia in Caribbean and Japanese populations. An HTLV-I viral protein essential for T-cell transformation, Tax, activates the transcription factor NF_B.

The immune system normally provides a considerable measure of protection from a class of viruses termed 'DNA tumour viruses'. By weakening this immune surveillance, the human immunodeficiency virus (**HIV**) is responsible for a significant cancer burden. Three DNA tumour viruses are prominently responsible for HIV-associated malignancies: Epstein–Barr virus (**EBV**), human herpesvirus 8 (**HHV-8**), and human papilloma virus (**HPV**). These viruses can also cause disease in individuals with apparently normal immune systems.

It has long been appreciated that EBV is capable of immortalizing B cells, thereby stimulating polyclonal populations of B cells in which secondary transforming events occur. In HIV-infected patients, lack of a T-cell-dependent antiviral response allows proliferation of EBV-infected polyclonal B-cell populations, from which highly aggressive lymphomas evolve. Through a similar mechanism, EBV is a cofactor in endemic childhood Burkitt's lymphoma in Africa. In this case, poor nutrition and malarial infection deplete the immune system, allowing an EBV-expanded population of lymphocytes to proliferate. Eventually a c- *myc*/immunoglobulin gene translocation generates an aggressive, poorly differentiated Burkitt's-type lymphoma. Finally, EBV is implicated in the pathogenesis of nasopharyngeal cancer, an endemic malignancy of the nasal sinus epithelium in southern China. The genome of EBV encodes up to 100 potential genes and several have oncogene-like properties, including BHRF1, a BCL2-like protein. While the exact role of all EBV-encoded proteins in transformation has yet to be elucidated, viral gene products are likely to initiate the malignant process, even if EBV genes may not be required for the continued growth of the malignancies that ultimately arise.

Like EBV, HHV-8 is a herpesvirus that is associated with malignancy, including Kaposi's sarcoma, primary effusion lymphomas, multiple myeloma, angioimmunoblastic lymphadenopathy, and Castleman's disease. Kaposi's sarcoma is an indolent sarcoma affecting the skin (and more rarely internal organs). At least 95 per cent of Kaposi's sarcoma lesions can be shown to contain the HHV-8 virus. While first recognized in eastern European males at the beginning of the twentieth century, Kaposi's sarcoma became a common problem only after the onset of the HIV epidemic. Sequencing of the HHV-8 genome reveals several genes that may be associated with malignant transformation, including a cyclin-like gene that inhibits the function of Rb, and a BCL2-like protein that inhibits apoptosis. Therefore, like EBV, HHV-8 may initiate Kaposi's sarcoma and other malignancies by encoding proteins that mimic several steps in the multistep carcinogenesis process.

HIV-infected individuals are also prone to develop squamous carcinoma of the anus and, in women, cervical cancer. Both diseases are caused by infection with HPV. Cervical cancer is an endemic disease that does not require an immunodeficient state for the virus to be pathogenic. The viral genome of HPV has been extensively examined, and two viral genes in particular have been implicated in HPV-induced cellular transformation, *E6* and *E7*. The E6 protein interferes with the function of p53 by targeting the protein for degradation. The E7 protein binds to and interferes with Rb. Thus, HPV mimics two of the most common loss-of-function mutations in human malignancy.

Another example of a strong link between malignancy and viral infection is the hepatitis B virus. In several parts of the world, including parts of China, hepatocellular carcinoma is the most common malignancy and reflects the very high incidence of hepatitis B infection in these areas. Portions of the hepatitis B virus can be found integrated into the genome of hepatocellular carcinomas. These hepatitis B virus 'X gene' fragments may be critical to carcinogenesis by providing viral promoter sequences that activate neighbouring oncogenes.

Epigenetic gene silencing in cancer

Epigenetic gene silencing is another non-mutational pathway for inactivation of tumour suppressor genes that has recently come into focus. In a number of instances, expression of tumour suppressor genes is lost in cancer cells, even though genetic analysis had shown that the sequence of the gene is intact. Tumour suppressor genes subject to gene silencing in cancer cells include *VHL*, *hMLH1* (a mismatch repair gene) *CDKN2A*, and *E-cadherin*. The mechanism for gene silencing involves methylation of cytosine residues in CpG dinucleotide sequences found within gene regulatory sequences. Methylation is believed to hold the gene in a closed conformation that cannot be accessed by RNA polymerase. How epigenetic gene inactivation is faithfully passed on to daughter cells during cellular replication is not clear. The mechanism may represent an aberration of the process that suppresses the expression of genes on the inactive X chromosome in females.

In addition to epigenetic gene silencing, there is also evidence for epigenetic oncogene activation due to loss of methylation. A good example is the gene for insulin-like growth factor 2 (**IGF2**). IGF2 is a potent growth factor that signals both cell growth and survival. Normally in adult tissues only the paternal IGF2 allele is expressed. The maternal allele is silenced through methylation, in an embryonic process termed 'imprinting'. In paediatric tumours (Wilm's tumour, rhabdomyosarcoma, and hepatoblastoma) the imprinting pattern is lost, with the occurrence of biallelic IGF2 expression and excess IGF2 activity in these malignancies. In these instances, loss of imprinting has been found to be due to mutation in the *H19* locus that is usually targeted for imprinting on chromosome 11p15, the vicinity of the *IGF2* gene.

Cancer therapy in the twenty-first century

The completion of the human genome project represents an enormous opportunity in medicine. Cancer therapy is likely to be an early beneficiary from this triumph because detailed information on the genetic nature of cancer will be translated into new, less toxic, and more effective cancer treatments. The recent development of a gene-targeted treatment for chronic myeloid leukaemia illustrates this potential well. Chronic myeloid leukaemia is characterized by excess proliferation of cells from the myeloid lineage. The hallmark of this disease is a reciprocal translocation between chromosomes 9 and 22 creating the so-called 'Philadelphia chromosome'. The

molecular consequence of this translocation is to fuse the c-*abl* gene, a nuclear tyrosine kinase, with the *BCR* gene. The fusion BCR–ABL gene product has enhanced tyrosine kinase activity. Having identified the enzyme target that drives chronic myeloid leukaemia, selective ABL tyrosine kinase inhibitors were developed that were able to cure up to 90 per cent of mice with human chronic myeloid leukaemia cells in their bone marrow and circulation. In early clinical trials with chronic myeloid leukaemia patients, the ABL tyrosine kinase inhibitor was highly effective, and certainly less toxic than the alternatives that include chronic treatment with interferon or allogeneic bone marrow transplant.

Like gene translocation, gene amplification also provides an opportunity for gene-targeted therapy. For example, a humanized antibody against the *erbB2* oncogene produces a 40 per cent response rate in breast tumours with *erbB2* gene amplification. Exposure to the erbB2 antibody downregulates erbB2 tyrosine kinase activity by internalization of the receptor/antibody complex. These early successes have spurred the development of new inhibitors of essentially the entire spectrum of signal transduction enzymes discussed in this chapter. Promising gene-specific therapeutic approaches are summarized in Table 5.

Further reading

Kaelin WG (1999). Choosing anticancer drug targets in the postgenomic era. *Journal of Clinical Investigation* **104**, 1503–6.

Lengauer C, Kinzler KW, Vogelstein B (1998). Genetic instabilities in human cancers. *Nature* **396**, 643–9.

Tycko B (2000). Epigenetic gene silencing in cancer. *Journal of Clinical Investigation* **105**, 401–7.

6.3 The genetics of inherited cancers

Andrew Coop and Matthew J. Ellis

Introduction

The description of families with multiple members afflicted by cancer provided an early and persuasive argument that the aetiology of cancer has a genetic component. This conclusion is now beyond doubt, since germline mutations responsible for inherited cancer syndromes have been identified, and a role for the encoded proteins in malignant transformation established. While the development of essentially every cancer is influenced by the genetic complement of the patient, only in these inherited cancer syndromes is the effect of a single gene mutation powerful enough to generate a mendelian pattern of cancer predisposition. The study of these diseases, despite their rarity, has provided critical insights into the genesis of more common forms of cancer because the same genes are frequently affected. In the case of sporadic cancer, however, mutations are not present in the germline but arise entirely through the process of somatic mutation and selection, whereby mutations that enhance cellular survival or proliferation accumulate in tissues over prolonged periods. DNA-based diagnostic tests are now available for gene abnormalities underlying inherited cancer predisposition syndromes. A positive test may mandate intensive cancer screening and, in carefully selected cases, surgery to remove organs at risk. Researchers are also examining the use of medications designed to prevent cancer (chemoprevention), allowing cancer-prone individuals to 'escape' their genotype. This chapter is intended to provide insights that will facilitate the recognition of cancer predisposition syndromes, help in understanding the clinical implications of the diagnosis, and render further study comprehensible. For a completely exhaustive list of familial cancer syndromes the reader is referred to the section on further reading at the end of the chapter.

The identification of cancer predisposition genes

Hereditary retinoblastoma: a classical example of a cancer predisposition syndrome

The first cancer predisposition genes were identified through the careful analysis of rare but remarkable syndromes. Hereditary retinoblastoma provides a classical example. A cancer of retinal cells, retinoblastoma usually occurs before the age of 3 years. One in 13 500 to one in 25 000 children are affected, with an equal sex distribution. About 40 per cent of patients have a family history of the disease, with an autosomal dominant pattern of inheritance. Eighty-five per cent of retinoblastomas presenting at less than 6 months of age affect both eyes and are likely to be the inherited form. The proportion of bilateral cases declines to 6 per cent by 24 months, when most cases are of the sporadic type, with no risk of genetic transmission. Overall only 10 per cent of patients with single tumours transmit susceptibility to the next generation. Individuals with hereditary retinoblastoma are at an incresed risk of developing a variety of cancers (especially osteosarcoma). Retinoblastoma illustrates the cardinal clinical features of inherited cancer predisposition syndromes: early onset, bilateral or multifocal cancer, and an association with similarly affected close relatives (see Table 1).

In a series of now classical studies, Knudson carefully examined family histories and tumour characteristics from patients with retinoblastoma and generated a 'two-hit' model to explain his statistical observations (Fig. 1). He hypothesized that the gene for retinoblastoma, *Rb*, is a tumour suppressor gene, whose function must be lost for a retinoblastoma to develop. In familial cases, he postulated that affected individuals had inherited an inactive mutant *Rb* allele. In these individuals every cell in the retina, indeed every cell in the body, carries an inactive copy of the *Rb* gene. As a result, a single second somatic mutation, or 'hit', in the remaining *Rb* allele is sufficient to inactivate *Rb* function and initiate a tumour. The barrier to the development of retinoblastoma in the absence of an inherited *Rb* mutant allele is much higher, as the coincidence of two somatic mutations (two hits) in a single cell to inactivate both *Rb* alleles is considerably less likely. This explains the low incidence, later onset, and absence of multiple tumours in non-familial retinoblastoma. The experimental verification of Knudson's model arose from studies that identified a region of chromosome 13 that was consistently lost in both inherited and non-inherited retinoblastoma. By examining multiple cases, investigators were able to identify the common region of loss on chromosome 13q14, and ultimately the gene concerned. True to Knudson's model, germline loss-of-function *Rb* mutations were present in familial cases, and the loss of chromosome 13 in familial tumours always eliminated the remaining wild type allele. Furthermore, sporadic adult onset cancers, for example of the lung, breast, and bladder, also have *Rb* mutant alleles, acquired as somatic mutations. This testifies to the importance of *Rb* in the formation of not just a rare childhood tumour, but common tumours as well. The final proof that *Rb* is a tumour suppressor gene came from studies that showed that malignant tumour formation can be suppressed when *Rb* function is returned to *Rb*-null cancer cells in gene transfer (transfection) experiments. These seminal studies on inherited retinoblastoma defined key experimental approaches and concepts that were subsequently used to identify other cancer predisposition genes.

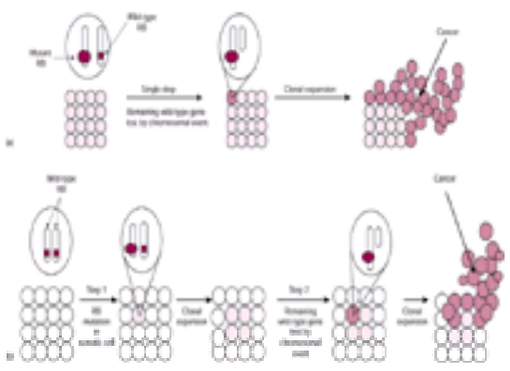

Fig. 1 Knudson's hypothesis. (a) In individuals with an inherited predisposition to retinoblastoma every somatic cell contains one intact *Rb* allele and one mutant *Rb* allele. A single somatic mutation is therefore sufficient for loss of *Rb* activity, with subsequent clonal expansion of the double mutant cell and tumour formation. (b) In normal individuals both copies of *Rb* must be targeted by somatic mutation for *Rb* function to be disrupted. Since the somatic mutation rate is low the risk of two *Rb* mutations occurring in the same cell is low. This explains the later onset and unifocal nature of retinoblastoma cases that occur in the absence of a family history.

Familial clustering of common cancers

Retinoblastoma is striking because cancer is a rare event in childhood, but up to 5 to 10 per cent of adults with cancers at commonly affected sites, such as breast and colon, have a family history of similar cancers. How many of these instances of 'familial clustering' are due to germline mutations in tumour suppressor genes?

Careful analysis of many of these family trees suggests that cancer predisposition is inherited in an autosomal dominant fashion. However, locating the genes associated with familial clustering is a daunting prospect. The presence of a family history of cancer alone is not particularly reliable evidence for an inherited cancer predisposition gene because many cases of familial clustering may be due to common exposure to environmental risk factors (smoking or poor diet for example). Other issues that frustrate the geneticist include small family size, incomplete knowledge of a relative's medical problems, or, in the case of breast and ovarian cancer, lack of female relatives in which the phenotype can be expressed. Geneticists also define the twin problems of 'incomplete penetrance' and 'phenocopy' as particularly awkward because they mask a mendelian inheritance pattern (Fig. 2). Incomplete penetrance refers to the situation where an individual has inherited a mutated tumour suppressor gene but does not develop cancer. By masking the presence of a mutant allele, incomplete penetrance generates confusing phenomena such as 'generation skipping'. The phenocopy problem arises because 'common cancers occur commonly'. A sporadic tumour can occur in an individual in a family with cancer predisposition who did not inherit the predisposing mutation. These false negative and false positive situations led to the development of statistical tools to predict the likelihood of mutations in tumour suppressor genes in any particular family. These models are now frequently used to help decide who should undergo expensive genetic testing. An example of one approach for breast cancer is provided in Table 2. A history of early onset or bilateral disease is used to weight these analyses, as these features substantially increase the likelihood that a cancer predisposition gene is present. Ethnic background is also taken into account. For example, certain mutations in the breast cancer genes *BRCA1* and *BRCA2* are more common in individuals with Jewish ancestry. These statistical approaches are critical because, unlike retinoblastoma, there were initially no cytogenetic clues from tumour analysis to help identify the chromosomes that contained the genes responsible for familial clusters of common tumours. Tumour chromosomes from common solid malignancies often show multiple gains and losses of chromosomal material, and no obvious single change associated with cases that arose in families with multiple affected members can be identified.

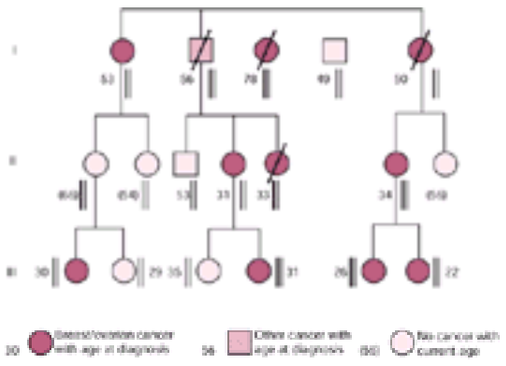

Fig. 2 Incomplete penetrance and phenocopies obscure mendelian inheritance patterns for cancer predisposition genes. In this hypothetical family, cancer predisposition is due to a mutant *BRCA1* allele on chromosome 17 (represented by the thick black line). Molecular analysis identified a 66-year-old woman in generation II who carries a mutant *BRCA1* allele but has not yet developed cancer. None the less she did transmit a mutant allele to one of her daughters who developed breast cancer at the age of 30. This is an example of incomplete penetrance that caused hereditary breast cancer to 'skip' a generation. In generation I, a 78-year-old woman developed breast cancer, but did not carry the mutant *BRCA1* allele. In this instance breast cancer was of the sporadic type, typically occurring in an older patient population. This case is referred to as a 'phenocopy' that mimicked the consequences of inheriting a disease allele.

Inherited cancer predisposition syndromes

These classical syndromes can be distinguished from the problem of familial clustering of common cancers because the frequent occurrence of cancer is not the only feature of these unusual but instructive diseases. A diagnosis of one of these syndromes should be considered when cancer arises in organs not commonly affected by sporadic cancer or when tumours have unusual 'premalignant' manifestations or other disease characteristics that suggest the presence of a disorder of tissue regulation or formation. An example that illustrates these concepts is neurofibromatosis type 1. In excess of 90 per cent of the individuals who inherit a mutated neurofibromatosis type 1 gene (*NF1*) develop *café au lait* spots and skin neurofibromas. Other clinical manifestations may also be present, but with much lower frequency, for example learning difficulties, skeletal abnormalities, and more deforming 'plexiform' neurofibromas. Cancer is actually fairly uncommon and occurs in less than 5 per cent of cases. Tumours arise at unusual sites, including the adrenal medulla (phaeochromocytoma), the central nervpus system (optic nerve gliomas, ependymomas, and meningiomas), and also within neurofibromas (neurofibrosarcomas). The reason for the extreme variability in the clinical features of neurofibromatosis is unclear. The effects of other genes (modifier genes) segregating within an affected family probably enhance or suppress the consequences of an *NF1* mutation. In addition, *NF1* mutations are diverse and lead to highly variable losses of coding sequence. Future studies may therefore define a relationship between the site of the mutation in the *NF1* gene and the clinical severity of the syndrome.

Linkage analysis

An effective, although labour intensive, approach to identifying genes that cause cancer predisposition is called DNA linkage analysis. Scattered throughout the human genome are sequence polymorphisms that are frequently heterozygous. Examples of these markers are single-nucleotide polymorphisms ('snips') or repetitive sequences (microsatellites) that can be identified through DNA sequencing or restriction fragment length polymorphisms. To identify the chromosomal location of a cancer predisposition gene these polymorphic markers are used in a genome-wide search to find a marker that cosegregates, or is 'linked' to the diagnosis of cancer because it is on the same chromosome. Once the chromosome carrying the cancer predisposition gene is identified, the segregation of further markers along the chromosome in question is determined. Because DNA is exchanged between chromosomes (recombination), the statistical association between a marker and the presence of cancer increases the closer the marker in question is to the disease locus. Eventually a marker is identified that is considered (statistically) close enough for the investigator to clone the segment of DNA that encompasses the marker and look for 'candidate genes'. Ultimately the gene is identified by sequencing each gene candidate to identify mutations that are present in individuals that have developed cancer but are not present in unaffected family members. Tumour DNA is examined for evidence for loss-of-function somatic mutations in the remaining wild type allele to confirm Knudson's postulate. The number of families necessary for linkage analysis depends of the nature of the disorder. For the breast and ovarian cancer genes *BRCA1* and *BRCA2*, a large number of families were required to generate sufficient linkage information because of the family history problems discussed above. For inherited cancer syndromes with characteristic clinical features and high penetrance, fewer families are required because affected and unaffected individuals within a family can be more reliably identified (which is just as well considering how rare these diseases are).

A classification of inherited cancer syndromes based on gene function

Textbook descriptions have classified cancer predisposition syndromes by mode of inheritance (dominant or recessive), by site of organ involvement (for example breast and ovarian cancer syndromes or colon cancer syndromes), or by the presence of associated clinical features (for example gastrointestinal polyps or neurofibromas). However, there is now sufficient insight into the function of genes associated with cancer predisposition to use a mechanistically based, rather than clinical, classification. This functional approach has the advantage of illustrating the relationship between gene function and the clinical manifestations of these disorders. The conversion of a normal cell to a cancer cell requires multiple mutations in genes that control the fundamental cellular processes of proliferation, cell death, and differentiation. An inherited mutation can therefore contribute to the process of malignant transformation by disrupting signal transduction, regulation of the cell cycle, chromosome stability, or DNA repair. All known cancer predisposing mutations, in which the function of the encoded protein has been established, involve one, or more than one, of these four cellular processes.

Cancer predisposition syndromes associated with mutations in signal transduction proteins (Table 3)

Signal transduction is the key to normal tissue homeostasis, and mutations in genes that encode components of these pathways are extremely common in cancer cells. Many cancer predisposition syndromes could really be considered inherited 'signal transduction disorders' because germline mutations in signal transduction proteins generate not only malignant tumours but also benign growths and unusual syndrome stigmata that represent the effects of abnormal signal transduction on the formation of normal tissue. Furthermore, tumours induced by mutant signal transduction proteins tend to arise in an organ-restricted manner, presumably reflecting the tissue-specific roles of the signalling pathways concerned. Examples of signalling proteins disrupted in these diseases include plasma membrane tyrosine kinases, cytoplasmic protein kinases, protein phosphatases, GTPase activating proteins, and transcription factors. All the conditions discussed in this section are autosomal dominant, but with variable penetrance, except ataxia telangiectasia, which is autosomal recessive.

Mutations in peptide growth factor receptors

Cell growth is regulated by peptide growth factors that bind and activate plasma membrane receptors. These receptors are enzymes that activate downstream cytoplasmic signal transduction proteins to alter gene expression and regulate the cell cycle. Four syndromes have been identified that are associated with activating mutations in either the RET or MET growth factor receptors, i.e. the mutation activates the receptor in the absence of the cognate peptide ligand (constitutive activation). These syndromes represent a major exception to Knudson's model, because the mutations concerned are 'gain-of-function'. As a result, there are no second mutational events in the remaining wild type allele. A fifth syndrome involves the *PTCH* gene that encodes a transmembrane receptor involved in body patterning. In this case the gene does conform to Knudson's postulate with respect to a 'second hit' and the germline and somatic mutations are all 'loss-of-function' in nature.

Multiple endocrine neoplasia types 2A and 2B and familial medullary thyroid cancer

Three disorders are due to activating mutations in the RET tyrosine kinase-linked cell surface receptor encoded by the *RET* gene at 10q11.2. RET is a neurotrophic pepide receptor and is predominantly expressed in cells of neural crest origin. Medullary thyroid cancer is common to all three conditions and about 20 per cent of all cases of this rare cancer are associated with a germline *RET* mutation. When no other stigmata are present, a diagnosis of familial medullary thyroid cancer is made. In the case of multiple endocrine neoplasia types 2A and 2B, additional unusual tumours arise including phaeochromocytoma and parathyroid adenomas (particularly in multiple endocrine neoplasia type 2A). Multiple endocrine neoplasia type 2B is associated with ganglioneuromas of the gastrointestinal tract and mucosal neuromas. In 95 per cent of cases of multiple endocrine neoplasia type 2A, mutations affect cysteine residues in the extracellular binding domain of RET, resulting in inappropriate disulphide bond formation, dimerization, and constitutive (ligand independent) activation of the RET tyrosine kinase. Familial medullary thyroid cancer results from mutations which similarly involve cysteine residues in the majority of cases but at different sites. The mutation found in multiple endocrine neoplasia type 2B is distinct and involves a methionine to threonine substitution in the ATP binding site of the receptor tyrosine kinase, leading to excessive receptor activity.

Familial papillary renal cell carcinoma

Some cases of this recently recognized syndrome, characterized by multiple bilateral papillary renal cell carcinomas, are due to germline mutations in the gene for MET tyrosine kinase at 7q31.1–q34. MET is a transmembrane tyrosine kinase receptor for hepatocyte growth factor or scatter factor, a peptide with essential roles in embryogenesis, cell motility, and tumour invasion. Germline *MET* missense mutations in cysteine residues, homologous to those involved in aberrant dimerization and activation of the RET receptor, are associated with familial papillary renal cell carcinoma. Monoallelic activating mutations in *MET* are also found the sporadic form of the disease. The spectrum of mutations found in sporadic papillary renal cell carcinoma is wider and includes activating mutations in the MET tyrosine kinase domain. However, recent studies have indicated that *MET* mutations occur in only about 15 per cent of sporadic papillary renal cell carcinomas.

Basal cell naevus syndrome (Gorlin's syndrome)

This condition should be considered in any patient presenting with a basal cell carcinoma before the age of 30, or with a personal or family history of multiple basal cell naevi/carcinomas. Associated abnormalities include abnormalities of skin, bone, and tooth formation (including polyostotic bone cysts, odontogenic keratocysts (jaw cysts), ectopic calcification, and palmar or plantar pits). An increased incidence of other cancers, including medulloblastoma, ovarian carcinoma, and sarcomas, may also occur. The incidence has been estimated at one in 55 600 in the United Kingdom. The gene believed to be responsible in the majority of cases, *PTCH* on 9q22.3, is a homologue of the *Drosophila* patched gene that encodes a transmembrane receptor for an extracellular ligand (hedgehog). This pathway is unrelated to classical peptide growth factor signalling and controls the fate of cells, body patterning, and growth by forming gradients in embryonic tissues. Initial analysis indicates that the *PTCH* gene behaves as a classical tumour suppressor gene.

Mutations in cytoplasmic signal transduction proteins

Plasma membrane receptors are linked to a highly complex array of cytoplasmic signal transduction proteins that transmit signals generated by an activated receptor to the nucleus to activate the cell cycle and alter gene expression. A number of highly characteristic syndromes are the result of loss-of-function mutations in proteins that have a negative role in signal transduction. The molecular genetics of these disorders generally conform to Knudson's model. However, disorders of tissue formation in these syndromes may occur in the absence of a second mutation, presumably because two normal functioning copies of the gene are required to maintain tissue homeostasis. In some cases, syndrome stigmata may arise because the mutated signal transduction protein interferes with the function of the wild type protein (a dominant negative effect).

Peutz–Jeghers syndrome

This autosomal dominant disorder has an incidence of one in 120 000. The syndrome stigmata are obvious with multiple pigmented spots on the lips and buccal mucosa, and multiple benign hamartomatous polyps throughout the gastrointestinal tract, most frequently affecting the jejunum. Malignant transformation of the gastrointestinal polyps is not common but may account for the increased incidence of colon cancer in these patients. In addition there is an increased risk of other cancers, including ovarian, cervical, testicular, and pancreatic cancer. The gene for Peutz–Jeghers syndrome, *STK11*, a serine threonine kinase, is located on 19p13.3. *STK11* is a tumour suppressor gene because germline mutations are predicted to disrupt the kinase domain of the protein. Molecular studies of hamartomas and gastrointestinal cancers associated with Peutz–Jeghers syndrome suggest that additional somatic mutation events are required for the transition of a hamartoma to an adenocarcinoma.

Neurofibromatosis type 1

The clinical features of neurofibromatosis were described in the introduction. The *NF1* gene (located on chromosome 17q11.2) encodes a guanosine triphosphatase activating protein known as the NF1-GAP-related protein, or neurofibromin. GAP proteins downregulate the activity of RAS, a critical mediator of the mitogenic response. An association between neurofibromin and the tropomyosin fibres of the cytoskeleton indicates a role in RAS-related signal transduction from the cell membrane to the cytoskeleton. A variety of mutations have been identified, and most result in a truncated protein. The rate of new germline mutations in *NF1* is high, with one-third to one-half of cases arising without affected parents. The *NF1* gene is prone to new mutational events because of its extremely large size (59 exons).

Neurofibromatosis type 2

This disease has an autosomal dominant pattern of inheritance. The neurological effects of neurofibromas predominate in neurofibromatosis type 2. There is a predisposition to development of tumours of the central nervous system, including schwannoma of the eighth cranial nerve ('acoustic neuroma'), meningioma, spinal cord schwannoma, and malignant gliomas. Deafness and tinnitus due to acoustic neuromas as well as muscle weakness and wasting due to spinal cord compression are not unusual. The *NF2* gene, located at 22q12.2, encodes a protein named schwannomin (or merlin). The majority of mutations within *NF2* result in the synthesis of a truncated protein. Schwannomin shows a close relationship to the family of ezrin–radixin–moesin proteins that link membrane proteins to the cytoskeleton and thus may function to maintain cytoskeletal organization. The incidence of neurofibromatosis type 2 is one in 35 000, of which half represent new germline mutations.

Tuberous sclerosis

Tuberous sclerosis is a disease of variable severity characterized by the development of multiple hamartomas involving many organs. Characteristic skin lesions often suggest the diagnosis. Often there is no family history since as many as 60 per cent of cases are due to a new germline mutation. There is a 5 to 15 per cent incidence of childhood brain tumours in affected individuals, mostly subependymal giant cell astrocytomas. In addition a weak association with renal cell cancer has been reported. A wide variety of benign tumours, including hamartomas, angiofibromas, and renal lesions, occur. Linkage studies have identified two genes, *TSC1* at 9q34 and *TSC2* at 16p13.3. *TSC1* encodes a protein called hamartin. Most mutations described within this gene result in a truncated protein. *TSC2* encodes tuberin, a protein showing some homology to GTPase activating proteins.

Von Hippel–Lindau disease

The diagnosis of von Hippel–Lindau disease depends on the presence of either coexisting central nervous system and retinal haemangioblastoma or one or other of these features, plus multiple renal, pancreatic or hepatic cysts, phaeochromocytoma, or clear cell renal cancer. Only one of these clinical features need be present if

there is a family history of von Hippel–Lindau disease. The incidence of the disease in the United Kingdom has been estimated as one in 36 000, with near complete penetrance by the seventh decade of life. Life expectancy is markedly reduced with death usually due to haemangioblastoma (benign vascular tumours) or renal cell cancer. The *VHL* gene at 3p25–p26 contains three exons that encode a 213 amino acid protein. The von Hippel–Lindau protein plays a role in the transduction of growth signals generated by changes in oxygen tension, promoting the translation of target genes that include vascular endothelial growth factor. *VHL* is a classical tumour suppressor gene, with a second, somatic mutation required for the development of cancer. Mutations in *VHL* are common in sporadic renal clear cell carcinoma.

Familial adenomatous polyposis or adenomatosis polyposis coli including Turcot's syndrome and Gardener's syndrome

Familial adenomatous polyposis (alternatively referred to as adenomatosis polyposis coli) has an incidence of one in 6000 to one in 13 000. Adenomatous polyps of the colon appear at an early age, with multiple polyps present in more than 90 per cent of cases by the age of 20 years. The polyps are premalignant, with the risk of adenocarcinoma of the colon approaching 100 per cent by the fifth decade of life. There is an increased risk of gastrointestinal carcinomas at other sites, thyroid cancer, childhood hepatoblastoma, and central nervous system tumours. Sometimes the combination of these tumours generates a recognizable syndrome. For example, multiple adenomatous colonic polyps in combination with a medulloblastoma is referred to as Turcot's syndrome. There is also an increased frequency of benign neoplasms including duodenal polyps, gastric polyps, and lipomas. In addition, there is an elevated risk for desmoid tumours. The gene for these disorders, *APC*, is located on chromosome 5q21–q22. Up to a third of cases are due to new germline mutations that usually cause protein truncation. Mutations mostly occur in the middle of *APC*, causing multiple colonic polyps appearing at puberty. *APC* mutations outside of this region are associated with fewer polyps and later onset. Some mutations correlate with the presence of unusual extracolonic syndrome stigmata, including congenital hypertrophy of the retinal pigment epithelium and benign bone tumours (osteomas). When these features are present the condition is referred to as Gardener's syndrome. These abnormalities occur in the absence of second mutations and may be due to the ability of mutant APC protein to interfere with the function of the remaining wild type protein (dominant negative effect). The APC protein is a negative regulator of b-catenin, a critical component of a signal transduction pathway that regulates cell–cell adhesion, cellular polarity, and tissue architecture.

Cowden's disease (gingival multiple hamartoma syndrome)

This autosomal dominant condition is most often recognized on the basis of characteristic skin lesions and intestinal hamartomas. Craniomegaly and mental subnormality occur in about 50 per cent of affected individuals. The pathognomic mucocutaneous lesions include trichilemmomas, acral keratoses, papillomatous papules, hyperkeratoses, and oral fibromas. Breast cancer occurs in 30 per cent of female gene carriers. Thyroid cancer is also prevalent (3 to 10 per cent), as well as glial masses that may present as cerebellar ataxia and seizures (Lhermitte–Duclos disease). The gene concerned, 'phosphatase tensin homologue deleted in chromosome 10' or *PTEN*, is located on 10q23 and is frequently deleted in sporadic breast cancer and glioblastoma. The PTEN phosphatase, by operating in opposition to the phosphoinositol-3-kinase pathway, inhibits cell survival and growth.

Ataxia telangiectasia

This rare recessive condition (one in 30 000 to one in 100 000) is not easily classified. Traditionally ataxia telangiectasia has been grouped with Fanconi's anaemia, Bloom syndrome, and other disorders associated with an increased rate of somatic mutation. Ataxia telangiectasia patients, however, have a distinct signal transduction defect, hence the inclusion of ataxia telangiectasia in this section. The *AT* gene, at 11q22.3, encodes a 350 kDa protein which contains a domain sharing homology to members of the phosphatidylinositol-3-kinase family. The AT gene product is believed to be a signal transduction protein that regulates cell cycle checkpoints. In the presence of DNA damage, cells with an AT mutation fail to activate p53-dependent cell cycle arrest (see next section). Although the precise connection to p53 and other cell cycle regulators is unclear, it is understood that AT-deficient cells exhibit extreme sensitivity to agents which damage DNA, genetic instability, and spontaneous chromosome aberrations. There is a 30 to 40 per cent lifetime risk of malignancy including epithelial tumours, solid tumours, chronic T-cell leukaemia, and lymphoma. Ataxia telangiectasia patients exhibit multiorgan defects including progressive cerebellar degeneration, general neuromotor dysfunction, and humoral and cellular immune defects. In fact, neurological problems and infection dominate the clinical picture. *AT* heterozygotes do not exhibit any of these defects, but may suffer a two- to threefold increase in the risk of cancer.

Mutations in transcription factors that control tissue-specific gene expression

The signal transduction pathways discussed above ultimately have an impact on the activity of DNA binding transcription factors that either activate or suppress transcription from target genes. At the time of writing there is only one example of a transcription factor involved in tissue-specific gene expression that functions as a tumour suppressor, the Wilm's tumour gene *WT1*. It is unclear why mutations in transcription factors that regulate tissue formation are rarely the cause of familial cancer, while mutant transcription factors occur frequently in sporadic cancer. One can extrapolate from transgenic experiments in mice that this disparity is explained by fact that the regulatory functions encoded by transcriptional suppressors are essential for normal embryogenesis.

Wilm's tumour (nephroblastoma)

Wilm's tumour is a poorly differentiated tumour of the kidney associated with developmental abnormalities of the genitourinary tract. Males and females are equally affected and usually present early in childhood, most often with an abdominal mass. Two sites of loss of heterozygosity have been identified in Wilm's tumours, *WT1* at 11p13 and *WT2* at 11p15.5. There are also rare familial cases in which linkage to neither 11p locus has been established (referred to as the ' *WT3'* group). In 10 to 30 per cent of patients, the disease is bilateral or multifocal, but less than 1 per cent of all cases are truly familial. Most cases of bilateral nephroblastoma are due to new germline mutations in *WT1*. The protein encoded by the *WT1* gene is a 'zinc finger' DNA-binding transcription factor. WT1 interacts with another tumour suppressor, p53, to bind and suppress transcription from the epidermal growth factor receptor and insulin-like growth factor 2 gene promoters. When WT1 function is compromised, transcription from these growth- and survival-promoting proteins is increased, initiating tumour development. *WT1* is not, however, a strictly Knudson-type tumour suppressor. Statistical analysis of age at diagnosis and proportion of bilateral and unilateral tumours does not follow the pattern described for retinoblastoma. Furthermore, the children of patients who survive Wilm's tumour are at lower risk of the disease than would be expected from a dominant-acting tumour suppressor gene. There is evidence that 'genomic imprinting' may explain some of these anomalies. Imprinting is a process of gene inactivation through DNA methylation that preferentially favours expression from genes inherited from one or other parental lineage. The reader is referred to the list of further reading for a comprehensive discussion of this complex issue.

Cell cycle checkpoint defects (Table 4)

Cell cycle checkpoint defects and cancer

The entry of a cell into the cell cycle is regulated by a 'G $_1$/S' checkpoint that controls the transition of cells from a resting (G $_0$/G$_1$) state into a DNA synthetic or S phase. When S phase is complete, cells enter premitosis (G $_2$). A second major checkpoint, G$_2$/M, regulates entry into mitosis (M phase). A major component of these two checkpoints is a family of regulatory proteins called cyclins that regulate the activity of enzymes called cyclin-dependent kinases. As the cell cycle progresses, cyclin and cyclin-dependent kinase activities oscillate, with peak activities corresponding to transition through each cycle checkpoint. Cyclin-dependent kinase activity is subject to negative regulation by Rb, p53, and a family of cyclin-dependent kinase inhibitors. These proteins operate to arrest the cell cycle in response to a wide variety of signals, including inhibitory factors, radiation, and other 'genotoxic' stresses. By blocking the cell cycle in cells that have undergone DNA damage, cell cycle inhibitory genes provide important protection against the development of cancer. Without this protection, cells exhibit genetic instability, with the accumulation of chromosomal deletions, translocations, and duplications. Cancer predisposition syndromes that are due to mutant cell cycle regulators are not generally associated with obvious clinical stigmata because cell cycle proteins do not regulate tissue-specific gene expression. Premalignant tumours may occur, however. Despite the fundamental role these genes have in cell cycle regulation, the pattern of tumour development is surprisingly restricted. This may be due to a major role for environmental stresses as a cofactor in tumour development (ultraviolet irradiation of the skin for example) or poorly understood tissue-specific roles for the genes concerned.

The retinoblastoma gene product Rb

The clinical features of retinoblastoma were described in the introduction. The *Rb-1* gene encodes a 928 amino acid protein. Rb functions by binding to the 'E2F' family of transcription factors. When Rb becomes phosphorylated by a cyclin-dependent kinase, E2F transcription factors are released to drive transcription from genes required for DNA synthesis. In addition, transcription of cyclins that stimulate entry into S phase is activated. Mutations in Rb therefore disrupt the ability of Rb

to interact with and inhibit E2F. As a result, the G_1/S checkpoint is lost and cells initiate unscheduled DNA replication.

Li–Fraumeni syndrome

This uncommon autosomal dominant disease exhibits a high incidence of a wide variety of early onset tumours, including rhabdomyosarcoma, osteogenic sarcoma, breast cancer, brain cancer, leukaemia, and adrenal corticoid carcinoma. Penetrance is high, with almost half of genetically affected individuals developing cancer by the age of 30 (compared with 1 per cent in the general population), rising to almost 90 per cent by the age of 70. Cancer at multiple sites is common. The clinical diagnosis is made by a patient with sarcoma under the age of 45, a first-degree relative under the age of 45 with cancer (not specified), and a third affected family member (first- or second-degree relative) with either sarcoma or any cancer under the age of 45. Approximately half of Li–Fraumeni syndrome families have mutations within the *p53* gene located at 17p13.1 resulting in a truncated p53 protein. *p53* is a classical tumour suppressor that stimulates transcription of cyclin-dependent kinase inhibitors in response to cellular stress. *p53* has been referred to as the 'guardian of the genome' because of a critical role in arresting the cell cycle in the presence of DNA damage.

Familial melanoma with or without dysplastic naevi

An autosomal dominant susceptibility to melanoma occurs in some families. Affected individuals in these families tend to have a large number of moles and are at risk for development of multiple melanomas and dysplastic naevi at a young age. Linkage to three loci, *CMM1*, *CMM2*, and *CMM3*, has been established. *CMM2* (9p21) encodes the cyclin-dependent kinase inhibitor 2A (known as p16). *p16* mutations are associated with a markedly elevated risk of both melanoma and pancreatic cancer and mutations occur frequently in a wide variety of sporadic tumours. Two transcripts are encoded by the same gene. One transcript, p16 (INK4a), induces a G_1 cell cycle arrest by inhibiting the phosphorylation of the Rb protein by cyclin-dependent kinases 4 and 6. The second (b) transcript encodes p14 (ARF), and is believed to function as a tumour suppressor by stabilizing p53 and leading to p53 accumulation. Mutation in the gene for cyclin-dependent kinase, identified as the gene at the *CMM3* locus 12q14, generates a dominant acting oncogene that promotes cell cycle progression. The gene responsible for the locus at chromosome 1p36, termed *CMM1*, has not been identified.

Hereditary breast/ovarian cancer (*BRCA1*)

In this condition there is an autosomal dominant predisposition to breast and/or ovarian cancer. Cancer occurs at a young age and is more frequently bilateral. Several hundred deletions, insertions, and point mutations have been identified within the *BRCA1* gene, with most predicted to result in truncated protein. The risk of breast cancer for individuals inheriting mutant forms of *BRCA1* is 3 per cent by the age of 30 years, 19 per cent by 40 years, and 85 per cent by 70 years. The risk for ovarian cancer is not as dramatic, but is increased greatly over that of the general population. An increased risk for colon and prostatic cancer is also present. The *BRCA1* gene on 17q21 has a role in the maintenance of genetic stability and the cell cycle response to DNA damage. This genetic surveillance role is similar to that of p53; BRCA1 and p53 are known to directly interact with each other.

Hereditary breast/ovarian cancer (*BRCA2*)

As in *BRCA1*, *BRCA2* mutations cause an autosomal dominant pattern of early onset breast and/or ovarian cancer predisposition. The onset of breast cancer in *BRCA2* families tends to occur earlier than in *BRCA1* families. Unlike *BRCA1*, there is an increased frequency of male breast cancer and pancreatic cancer in *BRCA2* patients. *BRCA2* has been localized to 13q12–q13. BRCA1 and BRCA2 share some homology and presumably serve similar functions in cell cycle checkpoint control and response to DNA damage.

Increased somatic mutation due to DNA helicase defects (Table 5)

The following recessive chromosome instability syndromes are rare and distinct from the classical cancer predisposition syndromes. There is no evidence for somatic mutation in these genes in tumours and so Knudson's model is not pertinent. The clinical features of these disorders overlap, with all three showing features of premature ageing. On a molecular level these syndromes are due to recessive mutations in a family of helicases that operate to maintain DNA topography during replication, recombination, and DNA repair.

Bloom syndrome

This an autosomal recessive disease of unknown incidence, more common in Ashkenazi Jews. Manifestations include growth retardation, sensitivity to the sun, skeletal abnormalities, and susceptibility to infection. An increased frequency of malignant neoplasms occurs throughout life, with dramatically reduced life expectancy. Lymphoma and leukaemia predominate before the age of 25; those that survive into their fourth and fifth decades are prone to a variety of common solid tumours. The age at diagnosis for these carcinomas is usually 20 or more years earlier than usually expected in the general population. Multiple mutations have been documented in the gene responsible, *BLM*, located on chromosome15q26.1. Loss of BLM, a RecQ DNA helicase, generates genetic instability with spontaneous chromosomal abnormalities and increased sensitivity to radio- and/or chemotherapeutic agents. Males are infertile due to a defect in meiosis. Heterozygotes do not seem to have an increased cancer risk, reflecting apparently normal genetic stability.

Werner syndrome

Werner syndrome is characterized by a multisystem premature aging phenotype also due to a RecQ helicase defect. The incidence is one in 50 000 to one in 100 000. Affected individuals have an excess of neplasms (especially osteosarcoma, meningioma, and thyroid cancer). A variety of loss-of-function mutations in the *WRN* gene located at chromosome 8p12–p11.2 have been identified. Loss of WRN helicase function leads to inappropriate expression of inhibitors of DNA synthesis, early cellular senescence, and genetic instability.

Rothmund–Thompson syndrome

This is the third recessive cancer predisposition syndrome due a defect in a RecQ-type DNA helicase (in this case, *RECQ4* at 8q24.3). The clinical features are similar to Werner and Bloom syndromes, comprising poikiloderma (marbleized pigmentation), telangiectasia, growth deficiency, cataracts, some aspects of premature ageing, and a predisposition to malignancy, especially osteogenic sarcomas and skin tumours.

Increased somatic mutation due to DNA repair defects (Table 6)

The fourth major class of cancer predisposition syndromes involves defects in DNA repair. Some of these defects cause severe problems in childhood and others are associated with a delayed onset, with cancer in adult tissues the major phenotype. These conditions can be recessive or dominant, depending on the consequences of the heterozygous state on the rate of somatic mutation. All promote the development of cancer by increasing the rate of somatic mutation. The genetics of these conditions is very complex—since DNA repair involves multiple protein components, a defect in any one will generate essentially the same phenotype.

Xeroderma pigmentosum

This is a spectrum of autosomal recessive disorders, with an incidence of one in 1000 000 in the United States. Childhood onset of photosensitivity and freckling leads to progressive degenerative skin changes and early development of skin and eye cancers. Approximately one-quarter of affected patients have concurrent neurological abnormalities. Basal cell and squamous cell carcinomas of the skin are increased 2000-fold and so the differential diagnosis includes Gorlin's syndrome. An increased risk of melanoma has also been observed. An increased risk of other tumours has been reported, including brain, lung, stomach, and haematological tumours. Benign neoplasms include conjunctival papillomas, actinic keratoses, lid epitheliomas, keratoacanthomas, angiomas, and fibromas. Defects of several enzymes involved in excision repair of ultraviolet light-induced pyrimidine dimers are responsible for this syndrome including *XPA* on chromosome 9q34.1, *ERCC3* on 2q21, *XPC* on 3p25.1, *ERCC2* on 19q13.2, *XPE* on 11p12–p11, *ERCC4* on 16p13.2–p13.1, and *ERCC5* on 13q32–q33.

Fanconi's anaemia

Fanconi's anaemia is another spectrum of recessive diseases characterized by a complex variety of developmental abnormalities, progressive marrow failure, and a

predisposition to acute myeloid leukaemia (15 000 times that of the general population). Fanconi's anaemia commonly presents in early to middle childhood with anaemia and bruising. Progressive pancytopenia and chromosome breakage, worsened by exposure to alkylating agents, is characteristic. Fanconi's anaemia homozygotes may develop a wide range of common cancers occurring at an early age, and are vulnerable to the hepatocarcinogenic effects of androgens used to treat Fanconi's anaemia. Squamous cell carcinomas, especially of the head and neck, oesophagus, cervix, vulva, and anus, occur with increased frequency, as do liver adenomas. Life expectancy is poor, around 12 years, with most deaths resulting from marrow failure and cancer. Approximately one-fifth of childhood aplastic anaemia is associated with Fanconi's anaemia. The heterozygote frequency is estimated to be one in 300 to one in 600, and is even commoner in Ashkenazi Jews. At the cellular level, Fanconi's anaemia homozygotes display chromosomal instability in response to DNA damage and reactive oxygen species and increased sensitivity to DNA crosslinking agents such as mitomycin C. Spontaneous chromosome aberrations are seen in a variety of cell types and an increase in chromosome deletion is seen. Five genes have been defined by complementation studies: *FA-A* at 16q24.3, *FA-C* at 9q22.3, and *FA-D* at 3p26–p22; *FA-B* and *FA-E* are currently unmapped. Several mutations have been found for *FA-C* and *FA-A*. The *FA-C* gene has been cloned but the function of the gene product is still not understood as it is a novel protein with no recognized functional motifs. However, the cellular phenotypes strongly suggest a role in DNA regulation or repair. Most studies have shown that Fanconi's anaemia heterozygotes do not display spontaneous genetic instability and do not have an increased risk of malignancy.

Hereditary non-polyposis colon cancer

This autosomal dominant condition may account for 6 to 10 per cent of all colorectal cancers. Clinically a diagnosis of hereditary non-polyposis colon cancer requires the following: three cases of colon cancer in a family in which two of the affected individuals are first degree relatives of the third; colorectal cancers occurring in two generations; and one colon cancer diagnosed before the age of 50 years. Five genes have so far been identified (located at 3p21.3, 2p22–p21, 2q31–q33, 7p22, and 2p16), each encoding proteins that participate in multimeric DNA mismatch repair complexes. Two of the genes, *hMLH1* and *hMSH2*, account for more than 90 per cent of cases of hereditary non-polyposis colon cancer. Most characterized mutations yield truncated gene products. The dominant inheritance pattern of these repair defects clearly distinguishes hereditary non-polyposis colon cancer from xeroderma pigmentosum and Fanconi's syndrome. Heterozygotes have 50 per cent expression of the mismatch repair protein in question, which presumably raises the somatic mutation rate sufficiently to increase the risk of cancer. Tumours that arise in patients with hereditary non-polyposis colon cancer exhibit very dramatic genetic instability in nucleotide repeat sequences (microsatellites). In fact, microsatellite instability can be used as a clinical test to distinguish adenomatosis polyposis coli from hereditary non-polyposis colon cancer. The lifetime risk of colon cancer for patients with hereditary non-polyposis colon cancer is 80 per cent, the average age at diagnosis is 45 years, and most cancers occur in the right side of the colon. Hereditary non-polyposis colon cancer is associated with an increased risk of other cancers, in particular endometrial adenocarcinoma, but also ovarian, gastric, small intestine, hepatobiliary tract and pancreatic cancer, skin cancer (sebaceous carcinomas), and transition cell cancer of the renal collection system. Glioblastoma multiforme is associated with hereditary non-polyposis colon cancer. 'Turcot's syndrome' of hereditary brain and colon cancer can therefore occur with both adenomatosis polyposis coli and hereditary non-polyposis colon cancer.

Syndromes in which the underlying gene defect has not been identified or the function of the gene has not been fully established (Table 7)

This review has focused on syndromes with established genetic aetiologies. Of course there are many other instances of familial clustering of common cancers that are under investigation, including those for prostate cancer, gastric cancer, carcinoid tumour, Hodgkin's disease, pancreatic cancer, and testicular cancer. In addition there are cancer predisposition syndromes with unusual phenotypes that should yield to genetic investigation soon, including oesophageal carcinoma with tylosis, Carney syndrome (hereditary myxoma), familial chordoma, osteochondromatosis, and familial paraganglioma. Again, the reader is referred to the further reading list.

Interventions for patients with inherited cancer predisposition

The role of clinical genetic testing in the management of cancer predisposition syndromes is in a constant state of flux. When the genetic basis for an inherited cancer syndrome is established there is an immediate opportunity to develop a reliable genetic test. However, genetic testing is fraught with methodological, psychological, and clinical difficulties and should not be applied without careful genetic counselling of the entire family concerned. In general, genetic testing is most rationally applied when the patient can be offered an intervention to prevent the cancer. This may simply be increased surveillance, a strategy most successfully applied when the target organ is easily accessible. For example, melanoma screening with regular skin examinations and sun avoidance for patients with hereditary melanoma syndromes is almost certainly successful in reducing the incidence of lethal melanoma. In contrast, for patients with Li–Fraumini syndrome, a condition in which cancers arise in a variety of internal organs with a high frequency, surveillance strategies are cumbersome and patients are generally reluctant to comply. A genetic test can be used to establish an indication for prophylactic surgery. For example colectomy should be offered to patients with adenomatosis polyposis coli or hereditary non-polyposis colon cancer and mastectomy and ovariectomy for patients with *BRCA1* or *BRCA2*. These are not easy decisions, however, since prophylactic surgery is disfiguring and/or is associated with significant functional and/or psychological impairment. The most hopeful approaches to these conditions involve the application of chemoprevention. Here a positive genetic test is an indication for a medication that attenuates the effect of a cancer predisposing mutation. The use of tamoxifen, an anti-oestrogen, to prevent breast cancer for patients with *BRCA* mutations is under investigation and inhibitors of prostaglandin synthesis have been shown to inhibit the development of polyps for patients with adenomatosis polyposis coli.

In the future it may be possible to develop preventative drugs specifically designed to interfere with the deregulated enzymatic activity generated by a mutant protein. For example, it is conceivable that RET or MET specific tyrosine kinase inhibitors could be employed to prevent the onset of cancer in multiple endocrine neoplasia type 2 and familial papillary renal cancer. Finally, using gene therapy strategies it may be possible to restore tumour suppressor function. Recent reports on p53 gene therapy certainly offer hope in this regard.

Further reading

Foulkes WD, Hodgson SV, eds (1998). *Inherited susceptibility to cancer: clinical, predictive and ethical perspectives*. Cambridge University Press, Cambridge.

Lindor NM, Greene MH (1998). The concise handbook of family cancer syndromes. *Journal of the National Cancer Institute* **90**, 1039–71.

Ponder BAJ (1997) Inherited cancers. In: Cox TM, Sinclair J, eds. *Molecular biology in medicine*, pp 172–90. Blackwell Science, Oxford.

6.4Tumour metastasis

V. Urquidi and D. Tarin

Introduction

Tumour dissemination with colonization of distant sites in the body and the formation of secondary tumours, termed metastasis by Recamier in 1829, is an enigmatic phenomenon. It is also by far the most common cause of death in cancer patients, because of the failure of vital organs replaced by deposits of disorganized, malfunctioning tumour tissue.

In the human embryo, progenitor cells undergo complex migrations and interactions during the formation of body structure, but after arriving at their final positions, they remain fixed throughout adult life. The only normal cell lineages exempt from this fate are leucocytes which continue to patrol the blood vessels, tissues, and lymphatics. Neoplastic non-haematological cells can acquire the capability to do likewise, and the characteristic patterns of colonization displayed by specific types of tumours and cell lines derived from them demonstrate that tumour metastasis is not a random process. Such programmed disorderly behaviour by a subpopulation of host cells, compromising the regimented anatomy of the other cell lineages, is a unique phenomenon in animals. Its mysterious quality and its grave clinical implications both hinge upon the failure thus far to identify any properties unique to metastatic cells that can illuminate the mechanisms involved or help treatment to control or eliminate disseminated malignant disease.

The metastatic process

Formation of a secondary tumour colony in a new site is the culmination of a complicated series of sequential and highly selective events. It begins with the emergence within the expanding tumour cell population of one or more clones whose descendants can cross tissue boundaries and infiltrate adjacent cellular populations, progressively disrupting the structure and function of the organ. As the tumour grows it induces blood vessels to penetrate and arborize within it to supply its metabolic needs. This in turn provides opportunities for tumour cells with appropriate properties to break into the circulation and be carried away to seed distant sites. Survivors that attach to the vascular endothelium at a distant site gain access to the resources of the tissue or organ where they have alighted by diapedesing through the endothelium and focally destroying the surrounding sleeve of basement membrane. To complete the process and produce a secondary tumour or metastasis in this site they then proliferate and attract a new fibrovascular scaffolding from the host organ, to sustain growth and prepare for the next metastatic event.

Clinicopathological correlations of metastasis

Not only are molecular events within the spreading tumour cells responsible for the phenomenon of metastasis, but microenvironmental conditions within the mixed tumour cell populations, composing the organ in which the secondary deposits are established, also play an important role. This interaction between tumour and host cells causes characteristic patterns of organ distribution of secondary tumours, linked to the site of origin of the primary tumour, that are clinically well recognized. Thus, for instance, the distribution of metastatic deposits in patients with breast cancer most commonly involves the regional lymph nodes, the long bones, the lungs, and the liver. Similar, largely predictable, patterns of spread are seen in patients with colon cancer, prostate cancer, and many other common malignancies. These patterns are exceptionally useful clinically, in guiding physicians where to search in order to stage the degree of spread of the disease and seek early evidence of recurrence.

Many studies in animals and humans have demonstrated that these distribution patterns result primarily from preferential colonization of some sites rather than from simple vascular drainage patterns, although such mechanical factors can be superimposed on the biological processes involved. This 'seed and soil' synergy between the properties of the spreading tumour cells and the cells comprising the colonized organ dictating the distribution of secondary deposits, first recognized by Paget in 1889, has been corroborated by many converging lines of evidence in recent years and finally proven by studies in ambulatory humans. The cumulative evidence shows that the disseminating tumour cells are not omnipotent, and are still dependent upon co-operative interactions between them and the cells of the stroma and vasculature of the host organ in order to establish a secondary colony.

The routes of dissemination are also clinically and experimentally important in that the tumour cells sometimes preferentially go by lymphatic or transcoelomic pathways rather than by the blood vascular system. Additionally it has been shown that certain types of mouse tumours suppress the growth of dispersed micrometastases by the release of blood-borne antiangiogenic molecules that inhibit the growth of the secondary tumours. Removal of the primary tumour can result in the simultaneous and sudden growth of many secondaries in these model systems. However, although some anecdotal literature suggests that similar phenomena have been observed in humans, it is certainly not accorded sufficient clinical credence to alter the current best standard of clinical practice, which is to excise as much tumour tissue as one can without compromising vital functions of the patient.

Clinical consequences of metastasis

The clinical effects of metastasis relate to the exponentially increasing tumour burden in scattered locations, which ultimately leads to the failure of organs vital for survival of the host. These effects are compounded by the proven capability of the progeny of cells in metastatic colonies to metastasize again. This makes it impossible to effectively focus treatment on the expanding and leap-frogging tumour population and therefore increases the risk of collateral damage to surrounding normal tissue. The idiosyncratic metabolic and physical effects of this increasing tumour load, such as paraneoplastic effects, pain, anxiety, haemorrhage, pathological fractures, and compressed or eroded vital structures, also need to be alleviated. Furthermore, the significant possibility of recurrence of disease in a distant site long after successful eradication of a locoregional tumour necessitates careful and prolonged monitoring, sometimes assisted by appropriate tumour markers.

Therapeutic considerations relating to metastasis

The current practice of tailoring combined therapeutic approaches, incorporating various radiotherapeutic and/or chemotherapeutic regimens, after surgical excision to staging of disease in the patient is increasingly effective for some cancers. Spectacular successes can result if multimodality therapy is used fairly early in the course of the disease. However, the main difficulty of treating widely disseminated metastatic cancer remains intractable and its solution is the 'Holy Grail' of oncology. New research on genetic mechanisms offers the best currently available hope of finding target molecules for rational drug design, gene therapy, and novel approaches to biological therapy, all aiming specifically at the eradication of late stage disease. Meanwhile, however, some significant and useful advances in treatment of such patients have been made by:

1. combining therapeutic regimens (for example surgery and hormone therapy and/or chemotherapy for patients with breast cancer)
2. delivering chemotherapy by appropriate routes (for example intrathecally for cerebral deposits, intraperitoneally for ovarian carcinomatosis peritonei), and
3. increasing dose intensity within a short period.

Although very high-intensity chemotherapy with subsequent (autologous or heterologous) bone marrow transplantation has been effective for some haematological malignancies, recent trials have failed to show clear evidence of benefit in patients with breast cancer or other solid cancers. Radiotherapy, like surgery, remains more suitable for the management of locoregional disease or isolated secondaries.

The problem of deciding how much treatment is needed to eradicate dormant micrometastases (small collections of clinically undetectable tumour cells sprinkled in

many sites), otherwise known as minimal residual disease, without exposing non-neoplastic stem cells in many of the patient's tissues to irreversible toxicity, is a continually present dilemma. The detection and treatment of metastatic tumour deposits in the brain, leptomeninges, bone marrow, and bones pose particularly difficult challenges. In the neural axis, treatment is fraught with the problems of delivery of chemotherapeutic agents across the blood–brain barrier and the difficulty of adjusting the dose of radiation therapy to be effective without damaging sensitive, normal nervous tissue. Sampling difficulties complicate evaluation of whether the marrow is involved in a patient and recent advances in purging it of malignant cells using monoclonal antibodies vary in their efficacy, depending on the type of tumour. Bony involvement is effectively controlled by radiotherapy if it is localized, but widely scattered deposits are refractory to reduction or elimination by most therapeutic regimens.

Current metastasis research and its importance for clinical oncology

There is, therefore, a real need for novel research to identify special characteristics of metastatic cells that could render them susceptible to therapies controlling further spread. The prevention of escalating dissemination whilst the existing tumour deposits are attacked would be a valuable contribution to the therapeutic armoury. Current research on the mechanisms of tumour cell metastasis is opening promising avenues for such advances. These are still in their infancy, but new observations on cellular, genomic and postgenomic (gene expression) aspects of the metastatic phenotype permitted by recent technological advances (i.e. cDNA microarray and proteomic methods) are providing clues to the mechanisms involved. At the cellular and biochemical level, the data demonstrate that tumour-induced angiogenesis, proteolytic activity by tumour-derived matrix metalloproteases, and epithelial–mesenchymal interactions, dependent on specific growth factor receptor–ligand binding events, facilitate the metastatic process in many experimental tumour systems. Other studies have indicated the presence of putative metastasis suppressor genes on a variety of chromosomes, depending on the tumour system under investigation. Interestingly, to date, comparable information on metastasis promoting genes is less abundant, although some studies have provided data supporting this concept. Additionally, a substantial body of data has been published indicating that expression of individual effector molecules, such as integrins, selectins, CD44, and cadherins, may accomplish various parts of the metastatic process under the direction of co-ordinating genes, as yet unknown. Despite decades of research, no single consistent marker or effector of metastatic behaviour has yet been identified and modern emphasis has shifted to using high-throughput gene expression analysis, such as gene chips and spotted microarrays, to identify whether co-ordinated patterns of expression of clusters of genes are involved in this complex process.

As the basic events in the metastatic process are comparable in many different species and in individuals with tumours of differing histogenetic origin, it seems likely that the underlying causal mechanisms will be initiated and driven by the same group of regulatory genes. If this interpretation is correct, the recognition of patterns of gene expression meaningfully correlated with metastatic phenotypes should guide investigators to the identity of the co-ordinating genes governing the phenotype and more effective targets for therapy. Results from our own laboratory using gene chips and other methods of genome-wide expression analysis support this view and are revealing previously invisible reliable differences in mRNA species between metastatic and non-metastatic tumour cells cloned from a patient with breast cancer. Computational analysis of several such datasets, utilizing clustering algorithms, to screen for commonalities in gene expression within the metastatic versus the non-metastatic phenotype should facilitate identification of the functional implications of such associations. Using such methods, genes at ascending levels in the hierarchy of regulation of the phenotype will be catalogued. The ultimate objective of modern clinical cancer research is to identify better markers for reliable evaluation of prognosis and detection of disease recurrence and to develop improved targets and strategies for novel therapies.

Summary

Although considerable progress has been made in the clinical management of patients with cancer, the best immediate hope of surviving the grim sequelae of metastatic spread of cancer still lies in early detection of the primary tumour, before it has started to shed cells into the circulation. The hunt for molecules or combinations of molecules which indicate the imminent onset of malignancy and which are released into easily accessible body fluids (blood, saliva, sputum) and waste products (urine and faeces) therefore has practical and scientific appeal, although success may not necessarily provide much information on the underlying mechanisms of the metastatic process. For new methods to contain or eradicate cancer that has already begun to disseminate, we still need to understand the mechanisms involved. Suffice it to say for the present, however, that understanding the fundamental basis of the metastatic process will also profoundly illuminate how cellular societies, which compose all metazoan and especially vertebrate organisms, are assembled and kept in strict topographical and developmental order. Such an intellectual 'spin-off' adds a further level of interest to the challenge of finding a practical solution to the vexing and challenging biological problem of human malignant disease.

Further reading

Fidler IJ (1999). Critical determinants of cancer metastasis: rationale for therapy. *Cancer Chemotherapy and Pharmacology* **43** (Suppl.), S3–S10.

Fidler IJ, Kripke ML (1977). Metastasis results from preexisting variant cells within a malignant tumour. *Science* **197**, 893–5.

Tarin D (1992). Tumour metastasis. In: McGee, JO'D, Isaacson PG, Wright NA, eds. *Oxford textbook of pathology*, Vol. 1, pp 607–33. Oxford University Press, Oxford.

Tarin D (1997). Prognostic markers and mechanisms of metastasis. In: Anthony PP, MacSween RNM, eds. *Recent advances in histopathology*, Vol. 17, pp 15–45. Churchill Livingstone, Edinburgh.

Tarin D, Matsumura Y (1994). Recent advances in the study of tumour invasion and metastasis. *Journal of Clinical Pathology* **47**, 385–90.

Tarin D *et al.* (1984). Mechanisms of human tumour metastasis studied in patients with peritoneovenous shunts. *Cancer Research* **44**, 3584–92.

Welch DR, Wei LL (1998). Genetic and epigenetic regulation of human breast cancer progression and metastasis. *Endocrine-related Cancer* **5**, 155–97.

6.5Tumour immunology

P. C. L. Beverley

Historical perspective

The strategies that have been tried in tumour immunology are based on the idea that tumour cells harbour differences that can be detected by the immune system. The first is to use antibodies to distinguish between tumour and normal cells; this difference can then be exploited for diagnostic, prognostic, or therapeutic purposes. The second is to stimulate the host by specific or non-specific immunization to react more vigorously to autologous tumour. All these avenues have been pursued for close to a hundred years so that it is pertinent to ask why it is that immunology has had such a minor impact in oncology, and particularly in cancer therapy.

The early attempts involved experiments in tumour transplantation and the description of tumour rejection. Unfortunately, the tumour rejection experiments were carried out in outbred animals and it was later realized that allograft, rather than tumour rejection, was being demonstrated. Nevertheless, the experiments did show the power of the immune response to destroy a large, growing tumour mass. Non-specific stimulation with bacterial products (Coley's toxin) was also tried as therapy but later more rigorous examination of the effect of immunostimulation showed that the success rate was extremely low. After the second world war, the development of inbred animals allowed properly controlled experiments on tumour-specific immune responses to be performed and the theory of immune surveillance provided a stimulus and a basis for a new wave of experiments in tumour immunology. Nevertheless, although tumour-specific immune responses were soon detected, it remained difficult to define their target antigens until the development of monoclonal antibodies and methods for gene cloning. One other advance was also essential for understanding of antitumour immune responses, the realization that thymus-derived (T) lymphocytes can only recognize antigens processed inside cells and displayed at the cell surface in association with major histocompatibility complex (MHC) antigens (Fig. 1).

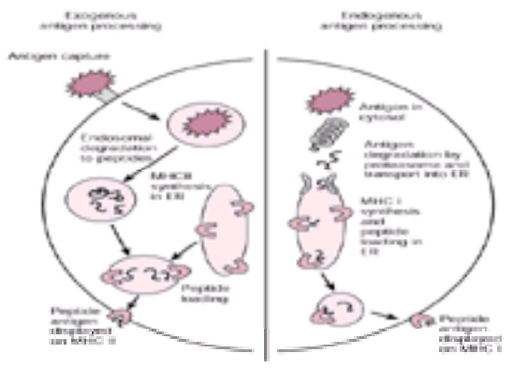

Fig. 1 Antigens recognized by T lymphocytes may be extracellular, in which case they are captured and internalized by antigen presenting cells to enter the exogenous antigen processing pathway. The antigens are broken down to peptides for loading on to newly synthesized MHC II molecules and translocated to the cell surface. Antigens synthesized within cells, which may originate from viral or self proteins, enter the endogenous pathway. They are broken down to peptides by proteasomal enzymes and the peptides are transported into the endoplasmic reticulum for loading and translocation to the cell surface.

Here are discussed immune surveillance, the evidence for immune responses to tumours, and aspects of immune function relevant to immunotherapy. Current understanding of the nature of tumour antigens recognized by the host immune system and how differences between tumour and normal cells may be exploited for diagnostic and therapeutic purposes are also reviewed.

Immune surveillance

The theory of immune surveillance postulated that tumours arise frequently but that most are eliminated by the immune system before becoming clinically apparent. Diverse evidence was adduced in support of the theory (Table 1) and many experiments performed to test it. Most of these were studies of the effects of immunosuppression, since the strongest prediction of the theory was that tumours should arise with overwhelming frequency in the absence of an immune response. Examination of congenitally immunodeficient or deliberately immunosuppressed mice supported the theory, since a higher frequency of tumours was observed than in controls. However, many tumours were characteristic of those caused by murine oncogenic viruses. Similarly, in man, deliberately immunosuppressed individuals or AIDS sufferers do have a greatly increased risk of tumours (Table 2) but not the most common types. Strikingly, viruses are implicated in the aetiology of most of the tumours types arising in immunocompromised individuals (Table 3). These data are clearly not in accord with a straightforward immune surveillance mechanism, since this would postulate an increased frequency of all tumours but in practice the strongest effects attributed to surveillance against tumour cells are most probably due to immune responses against potentially oncogenic viruses.

In spite of strong immune responses against oncogenic viruses, these agents do cause tumours in normal individuals as well as the immunocompromised. This may be because the virus-transformed cells adopt several strategies for escaping from immune attack (see below). In turn, this leads to the suggestion that the most effective means of preventing these tumours is by prophylactic immunization against the oncogenic virus, preventing infection and therefore transformation. Immunization against hepatitis B virus has already been demonstrated to decrease the incidence of liver cancer in endemic areas and vaccines against Epstein–Barr virus (EBV) and papillomaviruses are areas of active research.

In the case of many of the most common tumours, there is no evidence for a viral aetiology so that for these tumours the existence and importance of immune responses to tumour-associated antigen remains questionable. This is discussed below.

Tumour-specific immunity

The first convincing evidence that there were tumour-specific antigens came from experiments with inbred mice using chemically-induced tumours. Animals immunized with irradiated tumour cells or by grafting and then excising tumour, were protected against rechallenge with the same, but not a different, tumour line. Immunity could be transferred to naïve animals with cells but not serum. Notably, spontaneous tumours were usually much less immunogenic, raising the possibility that non-viral as well as virus-induced tumours may exhibit escape mechanisms.

In man, the evidence was rather slower to accrue but the development of the colony inhibition test, in which tumour cells were cocultured with patients' lymphocytes and the number of growing colonies counted, allowed an analysis of antitumour responses. Early studies showed that many patients were resistant to tumour and it was also suggested that 'blocking factors' present in serum might be responsible for the progressive growth of tumours in the face of a cellular colony-inhibiting response. Later results threw doubt on the specificity of the early research and it became clear that both normal and cancer patients' lymphocytes could often inhibit

the growth of tumour targets. This led to the definition of natural killer cells. These cells play a role in the early innate (non-specific) response to infectious agents but there is little to suggest that they play an important role in protection against non-viral tumours.

More persuasive evidence of specific T-cell responses to tumours became available following the realization that T cells can only recognize antigen in the context of self MHC (Fig. 1) and that specific immune responses can often only be revealed following an *in vitro* boost. Since that time, a number of authors have shown that both MHC class II restricted tumour-specific T-helper and class I restricted T-cytotoxic responses can be generated by coculture of lymphocytes and autologous tumour cells. Although it is clear that these responses occur in many tumour-bearing patients, the fact that they have been difficult to detect reproducibly indicates that they are often weak, in other words that there is a low frequency of responding T cells. That these responses may nevertheless play a role in host protection is suggested by the observation that loss of MHC molecules is very common in tumours. At least 50 per cent of human tumours show either down regulation of all class I or allele-specific loss. Clearly loss of MHC molecules which could present a tumour antigen is a very effective mechanism of tumour escape.

What are the tumour antigens recognized by T cells?

The nature of the tumour-specific antigens of non-viral experimental tumours and of most human tumours remained inaccessible until the development of gene expression cloning methods. In the 1980s Boon and his colleagues set out to identify tumour antigens of a mouse tumour, P815. The tumour was initially non-immunogenic in syngeneic DBA/2 mice, so that it grew readily even in preimmunized mice. A set of tumour variants were produced by treating the tumour with a mutagen. Some of these were highly immunogenic and unable to grow in preimmunized mice (tum– variants). From mice immunized with tum– cells, it was possible to isolate cytotoxic T-cell clones, which could kill only the immunizing tumour variant and not parental P815 or other tum– variants. The cytotoxic T-cell clones were used to screen parental P815 cells, which had been transfected with a cosmid library constructed from tum– variant cells. Eventually tum– antigen-positive transfectants were identified, the cosmid recovered, and, after subcloning, transfecting the subclones, and rescreening, the gene coding for the cytotoxic T-cell target antigen could be identified. Several different tum– antigens have now been cloned in this way and it has become clear that they fall into two categories (Table 4). The first is due to a genetic alteration in the coding sequence of a gene and the second to altered expression of a normal gene product.

Boon and his colleagues also investigated human tumour antigens in a similar fashion. Melanoma has long been recognized as capable of stimulating a relatively strong host T-cell response. Melanoma-specific cytotoxic T-cell clones were therefore produced and used to screen a melanoma cDNA library. They first identified MAGE-1 (melanoma antigen 1), belonging to the category of over-expressed, unaltered tumour antigens (Table 5). More remarkably, the antigen belongs to a large, previously undiscovered gene family which is widely expressed in tumours and seldom in normal tissues. Subsequently, Boon's group and others have cloned several other melanoma antigens as well as antigens of other tumours (Table 6). The majority of antigens identified by this methodology are unaltered. However, mutations and genetic rearrangements creating new sequences are common in tumours (Table 7) and computer analysis has shown that many of the neoepitopes formed could bind to common HLA alleles such as HLA A2. It has been shown also that some tumour patients and mice can respond to peptides of mutant *ras* oncogene and some cancer patient can generate cytotoxic T cells to the HER2/*neu* oncogene product. This work also established that most tumour antigens recognized by T cells are similar to other antigens recognized by T cells (Fig. 1, Table 4), that is they are short peptides presented by MHC molecules.

One other form of T-cell antitumour response is noteworthy. Mucins are heavily glycosylated molecules which are expressed at the surface of many epithelial cells and corresponding tumours. However, the glycosylation of the molecule is frequently altered in tumour cells leading to the detection of new epitopes by monoclonal antibodies on the tandem repeat structure of the protein core of the molecule. Cytotoxic T cells generated from breast cancer patients lyse tumour cells expressing the mucin but in a MHC-unrestricted fashion. The cytotoxic T cells may be CD4 or CD8 and have an ab T-cell receptor but unusually killing can be blocked a mAbs to the mucin tandem repeat. Exactly how these unusual cytotoxic T cells recognize their antigen remains to be determined.

The demonstration that many tumour patients make a cellular immune response to their tumour raises the question as to why tumours continue to grow and suggests that the response to tumours may be weak. This is discussed below.

Why are the responses to tumour antigens weak?

Just as for other T-cell-recognized antigens, several factors will determine whether a tumour antigen stimulates an immune response in the host. The first is that processed peptides from the molecule must reach the cell surface in association with MHC molecules. Whether a particular peptide does so depends on the amount produced by processing enzymes, the ability of the peptide to reach intracellular compartments where it may bind to MHC molecules, and its affinity for the MHC alleles expressed by the cell.

The second factor determining whether there is a response is that there must be T cells capable of responding. During normal development of T lymphocytes in the thymus, those cells that react with high affinity to self antigens are deleted to prevent the development of autoimmunity. It is, therefore, somewhat surprising that so many human tumour antigens detected by T cells appear to be unaltered self molecules. The most likely explanation is that thymic deletion and T-cell responsiveness in the periphery are governed by the affinity of the interaction of the T-cell receptor with peptide–MHC complexes. The thymic deletion mechanism selects against the highest affinity cells, while cells with lower affinity for self escape to the periphery. If the antigen is up-regulated, there may now be enough peptide–MHC complexes to stimulate lower affinity cells that have escaped deletion in the thymus. For antigens containing new sequences introduced through genetic alteration, there is in principle no reason why there should not be responsive T cells in the periphery. However, the frequency of T cells able to respond to a single epitope will be low (contrast this situation with that of a pathogen which has many proteins and even more peptide epitopes).

There is a second reason why responses to tumours may be poor. This is the need for 'danger signals' in initiation of immune responses. Recently, it has been recognized that activation of the innate immune system is essential for development of a subsequent, specific immune response. Pathogens carry 'danger signals' which perform this function. These are conserved structures such as bacterial carbohydrates, lipopolysaccharides, and signature DNA and RNA sequences that are recognized by conserved host receptors including complement components, receptors for carbohydrates (mannose and scavenger receptors) and lipopolysaccharides (CD14), and as yet ill defined recognition systems for nucleic acid sequence motifs. Interaction of danger signals with these conserved receptors leads to the production of cytokines by dendritic cells, macrophages, natural killer, and tissue cells that are essential for initiation of immune responses (Fig. 2).

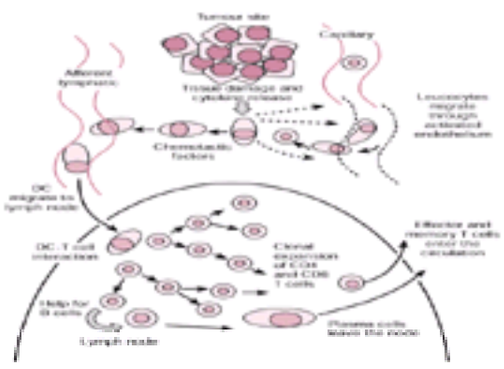

Fig. 2 Initiation of immune responses to pathogens is dependent on 'danger' signals carried by micro-organism, which stimulate production of cytokines by tissue cells and tissue-resident dendritic cells and macrophages, leading to influx of more leucocytes and migration of antigen-loaded dendritic cells to the draining node. Tumour sites are unlikely to stimulate these events until tissue damage and cell death occur as a relatively late event, so that the immune response to tumours may be delayed.

In the case of a tumour it is not clear what would provide such danger signals during the early stages of growth; later, tissue damage and repair processes would be expected to lead to cytokine production. Tumours may therefore be immunologically silent until a relatively late stage, by which time the tumour burden may be too great for the immune system to deal with, especially in the face of tumour escape mechanisms.

An important second stage in initiation of immune responses, induced by T-cell receptor–peptide–MHC contact, is the up-regulation of costimulatory molecules on both T and antigen presenting cells (usually dendritic cells) (Fig. 3). The interaction between CD28 and B7 is particularly important in promoting T-cell proliferation

and that between CD40 and CD40 ligand has recently been shown to play a key role in differentiation of cytotoxic T cells. It has also been shown that T-cell receptor–peptide–MHC interaction in the absence of costimuli may lead to inactivation rather than activation of T cells. Since most tumour cells lack costimulatory molecules, they are usually poor antigen-presenting cells. Even when activated effector cells return to the tumour site, the lack of costimuli may provide a mechanism for tumour escape, as may the production by tumour cells of down-regulatory cytokines, such as transforming growth factor-b (TGF-b). T lymphocytes isolated from tumour-bearing patients have been reported to show defects in signalling through the T-cell receptor, supporting the idea that tumours are indeed immunosuppressive.

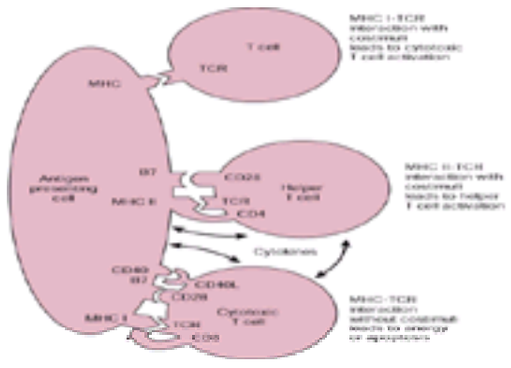

Fig. 3 Most T cells responses are initiated by contact between T cells and dendritic cells, which are specialized for antigen presentation and express key costimulatory molecules such as B7 and CD4O. Many tumour cells lack these and are therefore likely to be poor antigen-presenting cells.

Tumour antigens defined by antibodies

The exquisite specificity of antibodies has long persuaded investigators that it should be possible to use them to distinguish between tumour and normal cells. However, until the development of monoclonal antibodies (mAbs) there had been few successful examples. Many mAbs have now been raised against tumours and the vast majority recognize not tumour antigens but normal differentiation antigens. There are a few exceptions—antibodies to the idiotype of T and B cell tumours are tumour specific and antibodies to mutant forms of the p53 tumour suppressor gene product identify this as a tumour-associated molecule. In spite of this apparent failure, mAbs have proved extremely useful in studying the biology of tumours and lack of absolute specificity for tumour cells does not preclude their use as diagnostic or therapeutic agents.

More recently, serological analysis of tumour antigens by recombinant cDNA expression cloning (SEREX) has been developed. Sera from patients are used to screen cDNA expression libraries from fresh tumour material. Isolation of antigens detected by high titre IgG or IgA antibodies implies that there is a T helper cell response to the same antigen, as immunoglobulin class switching does not occur without this. That there is often a concurrent T-cell response when antibody to autologous tumour is present, is confirmed by the finding that some SEREX-defined antigens are identical to those recognized by T cells (see Table 6 and Table 8). The SEREX method has the disadvantage that it may not detect all conformational epitopes of a protein and does not identify carbohydrate antigens. Nevertheless, over 900 sequences of genes cloned using the SEREX method have already been deposited in a data base set up for this purpose. These include previously identified antigens such as MAGE-1 and tyrosinase, sequences identical (or nearly identical) to known genes not previously identified as eliciting an autoantibody response (e.g. kinectin, a transporter of golgi vesicles), and a large group of previously unknown genes (Table 8). A full description of the expression patterns of these genes in normal and tumour tissue, let alone a analysis of their functions, will be a major undertaking, but will eventually identify many new targets for immunotherapy.

The potential for immunodiagnosis and immunotherapy

Monoclonal antibodies have already found a place in the diagnosis of malignant disease (summarized in Table 9) although they seldom play a role in deciding whether a cell is malignant (although antibodies to some of the gene products identified by SEREX may eventually change this). So far, they have been most useful in cell identification. MAbs readily allow identification of single gene products in sections or smears and this is often a simple way of assigning a cytologically undifferentiated cell to a cell lineage. A relatively small panel of mAbs (Table 9) allows the identification of most tumours, which might otherwise cause diagnostic confusion. An extension of this is the detection of rare malignant cells. Thus, mAbs have been used to identify micrometastases in lymph nodes, blood, bone marrow, or cerebrospinal fluid, which are below the limit of detection by conventional histology or cytology. This data indicates that the presence of micrometastases at the time of diagnosis is much more common for many carcinomas than had been appreciated before. The clinical importance of this observation remains to be determined.

MAbs may also be used to provide prognostic information. Estimation of the proportion of cells expressing molecules associated with progression through cell cycle, such as the transferrin receptor or nuclear cyclins, correlates to some degree with the malignancy of tumours. However, it is not to be expected that expression of any single marker will correlate absolutely with tumour behaviour since one result of studies with mAbs is to reinforce the view that tumours are very heterogeneous. Nevertheless, as understanding of the function of gene products expressed in tumour cells improves, studies of tumours with carefully designed panels of mAbs will provide increasingly useful prognostic information.

Serological tests for a number of cancer-associated molecules have been developed; examples are carcinoembryonic antigen, a-fetoprotein, and prostate specific antigen. These are useful confirmatory diagnostic tests and can be used for follow up after treatment. A rise in serum antigen may indicate a recurrence of tumour.

Antibodies *in vivo*

As discussed above, there are few antibodies against molecules which are truly tumour specific. For *in vivo* use it is also necessary that the target antigen should be at the cell surface. In spite of these limitations, many mAbs have been used *in vivo* for diagnostic or therapeutic purposes. For both uses, the problems are similar. The first is to obtain sufficient penetration of antibody into tumour to give a useful signal-to-noise ratio. Penetration depends on the size of the molecule, its half-life in the serum, the vascularity of the tumour, and the abundance and availability of the antigen. The second problem is that heterologous antibody is immunogenic as are most of the larger molecules which are commonly attached to it, such as plant toxins. Both of these problems have been addressed by immunochemical or molecular engineering methods. It is possible to produce a variety of sizes of antibody fragments and to replace most of a rodent immunoglobulin by human sequences. This reduces greatly the immunogenicity of the antibody (although the idiotypic epitopes associated with the antibody binding site may remain immunogenic) but will not affect the ability of attached toxins to generate an antibody response. Strategies designed to circumvent this problem include further engineering of toxin moieties to render them less immunogenic, induction of tolerance to immunogenic molecules, and the use of a succession of different antibody–toxin or drug conjugates. Some of these strategies have now been explored in human clinical trials (see below). It should also be noted that the anti-idiotype response to an antitumour mAb can generate a further anti-idiotypic response (Ab3) which may react with the original tumour antigen. The presence of Ab3 has been associated with a favourable outcome in some studies.

Experience in man with antibodies for diagnosis suggests that their sensitivity is of the same order as that of computer assisted tomography. However, optimization of the antibody and radioisotope selection and coupling procedures will undoubtedly yield improvements.

For therapy, the problem of penetration into large tumour masses may dictate that antibodies are most useful in an adjuvant setting. In a randomized trial of a murine mAb to an epithelial differentiation antigen given at the time of surgery to colon cancer patients, treated patients showed significantly increased survival. Here the target is micrometastases, which would be expected to be relatively easily reached by serum antibody, and indeed it could readily be demonstrated that antibody had bound *in vivo* to tumour cells in the bone marrow. A number of other trials of antibodies have been carried out. In perhaps the most promising, a humanized antibody to the *neu* antigen (a member of the epidermal growth factor (EGF) receptor family) has been used alone or combined with chemotherapy to treat metastatic breast cancer; increased objective response rates have been reported. The antibody has received a licence from the Federal Drug Administration. Trials of anti-CD20 (a B lymphocyte differentiation antigen) mAb in lymphoma have also shown encouraging results.

CD20 antibody has also been used as a carrier for a radioisotope, as has an antibody to the mucin, MUC1, to deliver high doses of yttrium to the peritoneal cavity in

ovarian cancer patients with minimal residual disease after chemotherapy. In the latter trial, a significant benefit was seen compared to historical controls. A phase three trial is underway.

One possible avenue of progress in achieving antibody penetration is to target not the tumour itself but tumour vasculature. Endothelial cells respond to their microenvironment by changes in surface phenotype and the vessels of tumours may therefore express molecules which could be targeted. Initial animal experiments are encouraging. So also is the successful treatment of melanoma by limb perfusion with very-high-dose tumour necrosis factor (TNF-a), in which the tumour vasculature appears to be susceptible while normal endothelium remains unscathed.

T-cell immunotherapy

For those tumours in which a virus is implicated (Table 3), prophylactic vaccination would be the most obvious long-term solution. For other common cancers where there is no obvious viral target, T-cell-based therapy, if it does have a role, will have to be directed at established tumours. Viral antigens may also be targeted for active immunotherapy as has been tried in the case of carcinoma of the cervix, where the human papillomavirus transforming proteins E6 and E7 have been used (Table 10).

As discussed above, there is now good evidence for a specific T-cell response to at least some human tumours. In principle, it seems that any altered and expressed gene product, or even a normal but over-expressed product, may be recognized as a tumour antigen. It seems likely that most tumours will accumulate several possible tumour antigens during their evolution, though not all will be immunogenic in a particular individual because of the necessity for peptides of the tumour antigen to bind to self HLA molecules (genetic restriction). Although cataloguing the target epitopes present in tumours is a major task, it is certainly now possible and methodology will continue to improve. Given this increasing list of target epitopes, how can it be exploited?

Active immunization

One possibility will be to attempt active immunization against the antigens identified. This can be done using whole cells expressing the antigen, with recombinant viruses (most commonly recombinant vaccinia virus; see Table 10), with proteins or peptides, and most recently with DNA. Animal experiments have indicated that transduction of tumour cells with costimulatory molecules such as CD80 (B7) or cytokines (particularly GM-CSF and IL-2) reduces their tumorigenicity and renders them better able to immunize against subsequent tumour challenge. Recombinant viruses and DNA offer the possibility of introducing similar costimuli along with antigen and also ensure that the antigen can enter the class I pathway of antigen processing so that cytotoxic T cells should be induced. A related strategy is to isolate dendritic cells from patients and load them *in vitro* with tumour antigen before reintroducing them into the patient. Tumour antigen can range from tumour cell lysate, through recombinant antigens, to peptides, DNA, or RNA. Table 10 summarizes some of the approaches and antigens that have been tried in at least phase 1 trials. Inevitably, most of these trials have been performed in patients with advanced disease who have often received chemo- and radiotherapy. These patients are unlikely to make optimal immune responses, often have a large disease burden, and their tumours often show loss of HLA antigens. Nevertheless, objective clinical responses have been demonstrated in several trials.

Non-specific immunotherapy

Once cytokines became more readily available through the introduction of recombinant technology, several clinical studies were carried out. So far, the overall results have not been dramatic although high-dose IL-2, with or without lymphokine activated killer cells, does appear to cause some remissions in renal carcinoma and melanoma, but not without considerable toxicity. In retrospect, it is perhaps not surprising that the effects of high-dose cytokine therapy have been less than dramatic. Few cytokines are directly cytotoxic to tumour cells, although they may be cytostatic, and their effects are often transient because of receptor down-regulation. In addition, the indirect *in vivo* effects of high doses of cytokines are little understood. Much more work on the biology of cytokines is needed if they are to be used for tumour therapy in a rational fashion.

Cytokines have found application in cancer therapy, since colony stimulating factors can significantly shorten the period of aplasia following bone marrow transplantation. Another possibility is to use cytokines as vehicles for targeting. A recombinant IL-2–diphtheria toxin molecule has been engineered and shows some promise as an immunosuppressive agent. In principle, such a recombinant molecule can be used to target IL-2-receptor-bearing tumours such as those caused by HTLV-1.

Non-specific immunotherapy has a long history, dating back to the experiments of Coley. More recently Matthé used BCG to treat leukaemia, though with uncertain effect. However, there is now good evidence that intravesical BCG is an effective treatment for bladder cancer. Large-scale clinical trials have show that BCG can give very high response rates in recurrent superficial transitional cell carcinoma and the duration of remission is significantly prolonged. The mechanism of action of BCG remains incompletely understood but is presumed to be a consequence of altered local cytokine production in the tumour environment.

Passive immunotherapy

An alternative may be to develop passive immunotherapy. T-cell lines can be grown to very large numbers *in vitro* and in animal models tumour-specific cells can be effective especially when given with recombinant interleukin-2. Definition of tumour antigens will make it easier to isolate human tumour-specific clones and grow the large numbers of cells required for this strategy. Initial experiments using marker genes have shown some localization of *in vitro* expanded tumour infiltrating lymphocytes. More encouragingly, cytotoxic T cells against EBV antigens have been demonstrated to be effective in eliminating post-transplant lymphoma. To circumvent the difficulty of growing tumour specific T cell clones, methods are being developed for introducing genes for tumour-antigen-specific receptors into normal lymphocytes using retroviral vectors. A potential advantage of *in vitro* grown cells is that they might also be used to carry damaging molecules to the tumour. A disadvantage is that this method will always be labour intensive and will have to be carried out individually for each patient.

Conclusions

In non-malignant disease, immunological manoeuvres have been most effective when applied prophylactically. It is not expected that immunotherapy of established tumours will be easy. Antibody-based therapy shows promise and some antibodies may become standard adjuvant therapy in the near future. The future of T-cell-based therapy is less clear. Although active immunization against tumour antigens is now possible, its effectiveness will depend on the frequency of responding T cells and how quickly tumour escape mutants develop. Much work remains to be done to optimize immunization schedules and methods for monitoring the success of immunological approaches in cancer patients. New methods, such as the use of MHC tetramers and sensitive elispot assays, will help in enumerating tumour-antigen-specific cells.

Further reading

Boon T (1995). Tumor antigens and perspectives for cancer immunotherapy. *Immunologist* **3**, 262–3.

Lee MS *et al*. (1998). Hepatitis B vaccination and reduced risk of primary liver cancer among male adults: a cohort study in Korea. *International Journal of Epidemiology* **27**, 316–19.

Matzinger P (1998). An innate sense of danger. *Seminars in Immunology* **10**, 399–415.

Riethmuller G *et al*. (1998). Monoclonal antibody therapy for resected Dukes' C colorectal cancer: seven-year outcome of a multicenter randomized trial. *Journal of Clinical Oncology* **16**, 1788–94.

Schoenberger SP *et al*. (1998) T-cell help for cytotoxic T lymphocytes is mediated by CD40–CD40L interactions. *Nature* **393**, 480–3.

Sheil AG (1998). Cancer in immune-suppressed organ transplant recipients: aetiology and evolution. *Transplantation Proceedings* **30**, 2055–7.

6.6 Cancer: clinical features and management

R. L. Souhami

Introduction: cancer in general medical practice

Cancer is a common disease—approximately 20 per cent of the population of the United Kingdom will die of cancer. It is a source of concern and perplexity to oncologists that so many of their patients are referred to them late in the disease. Symptoms may have been present for a long time, during which their significance has been overlooked, or multiple (and sometimes futile) investigations have been performed with a failure to appreciate the need for speed. To this delay can be added a frequent lack of understanding of the possibilities of treatment and a failure to inform patients either of the nature of the diagnosis or of its implications and possibilities for therapy.

Almost every specialist sees patients with cancer affecting their particular field; unfortunately, these specialists may not have been taught the principles of cancer medicine. Oncologists therefore frequently see patients with advanced disease, who have not had a proper explanation of their illness, and who have little idea of what treatment might involve. Many patients dying with lung cancer in the United Kingdom have never seen a specialist in cancer medicine at any stage in their illness. While some will have had disease so far advanced that nothing but palliative treatment was needed, others may have been denied the opportunity for treatment and prolongation of life. The principles of cancer management are important for every physician.

Management of suspected cancer would be improved greatly if the following simple rules were adhered to:

1. Cancer should be suspected with any unexplained illness, especially in the elderly.
2. Imaging with isotopic and computed tomography (**CT**) and magnetic resonance imaging (**MRI**) (see Chapter 6.5), will often accelerate diagnosis. However, a tissue diagnosis cannot be made by these means and every attempt should be made to make a histological or cytological diagnosis expeditiously.
3. Patients with many tumours should start a planned programme of treatment within days and not weeks of diagnosis. The need for speed in diagnosis and treatment is tacitly recognized in specialist centres for breast cancer where patients can, in well-regulated clinics, reasonably expect to have a diagnosis made within a few days of first consultation and to begin definitive treatment within 2 weeks. This admirable efficiency should be attainable for nearly all cancers.

Common symptoms and signs of cancer

Many of the symptoms and signs of cancer are due to the local effects of the tumour infiltrating surrounding tissues and causing pressure and distortion of neighbouring structures. Tumours also produce symptoms that are, to some extent, common to all cancers. These are general symptoms due to the metabolic disturbances caused by the tumours and specific symptoms related to hormonal effects and immunological effects of the particular tumour—so-called paraneoplastic syndromes.

Pain

Most patients with cancer experience pain at some stage in their illness, either as a direct result of the tumour or of its treatment. Pain is a feature at presentation in about 30 per cent of patients with cancer, but the incidence varies greatly with the site of the tumour. For example, over 90 per cent of patients with primary bone tumours or with metastases to bone have pain, and this has the characteristic feature of being worse at night-time. In contrast, only 5 per cent of patients with leukaemia develop pain. Pain also varies according to the rate of progression of the disease and is more likely to occur with rapidly growing tumours. No symptom of cancer causes greater demoralization than unremitting pain. Any patient with unexplained, persistent pain should be suspected of having malignant disease and appropriate investigations performed. In pain clinics, 80 per cent of patients seen with cancer have pain due to direct tumour infiltration. If the pain is due to neurological infiltration it may be felt at the distribution of the nerve root. Certain pain syndromes are sufficiently common and misleading to warrant separate consideration.

Direct tumour infiltration of bone

The origin of the pain that occurs with tumour infiltration of bone is not fully understood. The periosteum is a pain-sensitive structure and may be the source in many patients. It is probable that osteolytic processes involving prostaglandins are also involved. Pain is a common feature of metastasis at the base of the skull. If the tumour is situated around the jugular foramen, pain is often referred to the vertex of the head and the ipsilateral shoulder and arm. Movement of the head may exacerbate the pain and, later, cranial nerve involvement may cause hoarseness, dysarthria, and dysphasia. Involvement of the 9th to the 12th cranial nerves, and the development of ptosis and Horner's syndrome indicates involvement of the sympathetic nervous system extracranially adjacent to the jugular foramen. When metastases occur in the sphenoid sinus, severe headache, usually felt in both temples or retro-orbitally, is a common feature. There may be a full sensation in the head, nasal stuffiness, and a 6th nerve palsy.

When metastases occur in vertebral bodies the pain frequently precedes neurological signs and symptoms. Persistent thoracic vertebral pain and a positive bone scan is an indication for urgent investigation and treatment. In small-cell lung cancer, for example, a patient with thoracic vertebral pain and a positive bone scan has a 30 per cent chance of developing a paraplegia. Ninety per cent of patients who have epidural spinal cord compression have vertebral body metastasis as the source of the epidural tumour (the management of spinal compression is described later). With metastasis to the odontoid process, patients complain of severe neck pain and stiffness radiating over the skull, up to the vertex. This is then followed by progressive neurological signs, often associated with autonomic dysfunction. In the

lower cervical vertebras pain is felt as an aching sensation, often radiating over both shoulders. If nerve root compression occurs at this site, there will be pain in the root distribution felt in the back of the arm, the elbow, and the ulnar aspect of the hand. The association with Horner's syndrome suggests involvement of the paravertebral sympathetic system. Lumbar metastases are associated with local pain, worse on lying or sitting and relieved by standing. In lesions in L1 the pain is often felt over the superior iliac crests. In the sacrum, pain may be accompanied by neurological signs with symptoms of bowel and bladder dysfunction and perianal sensory loss and impotence.

When tumours infiltrate peripheral nerves they are often accompanied by an alteration in sensation, with hyperaesthesia, dysaesthesia, and sensory loss. This is a particularly common presentation when tumours invade the paravertebral or retroperitoneal region. Here the pain is often in a root distribution and is unilateral. Another common site is when a metastasis in a rib entraps the intercostal nerves. When tumour infiltrates the brachial plexus the pain is felt in the C7 or T1 distribution. Pain in this site is frequent with the Pancoast syndrome, where an apical lung cancer infiltrates the lower brachial plexus roots. Pain in the C5 distribution occurs with upper root infiltration.

Visceral pain is a frequent symptom of cancer; it can cause diagnostic confusion and be difficult to control when a tumour has already been diagnosed. Poorly localized abdominal pain is a frequent feature of ovarian and pancreatic cancer and of peritoneal carcinomatosis. Retroperitoneal pain may be particularly difficult to diagnose. It may vary greatly in position (being relieved on leaning forwards) and be felt variably in the back. Left upper quadrant pain may be a presenting feature of carcinoma of the tail of the pancreas involving the mesentery of the splenic flexure of the colon.

Weight loss

Weight loss is an invariable accompaniment of advanced cancer and also a frequent presenting symptom. Often it results from the physical presence of the tumour interfering with gastrointestinal function, such as in carcinoma of the stomach, pancreas, or colon, or with peritoneal carcinomatosis. Mechanical obstruction of the bowel and loss of appetite commonly accompany these tumours. Loss of appetite is a frequent symptom of any cancer that has metastasized to the liver and usually appears at a point when metastasis is replacing much of the normal liver tissue. The mechanism is not known. Pancreatic cancers, and cancers metastatic to the porta hepatis, cause weight loss from a malabsorption syndrome due to obstructive jaundice or blockage of the pancreatic ducts.

Nevertheless, many tumours cause weight loss without direct involvement of digestive organs. It is well recognized that a weight loss of more than 5 per cent is a very adverse prognostic feature in almost all cancers. Usually it indicates that the disease is more widespread than is apparent on clinical investigation, but the mechanisms of this symptom, which is often accompanied by alteration of taste, anorexia, and a general feeling of ill health, are obscure. Sometimes quite profound weight loss can accompany non-metastatic tumours, which are relatively small. As with advanced cancer, the cachexia syndrome is then accompanied by anorexia and altered taste. These tumours may produce circulating factors responsible for the weight loss and loss of appetite. Tumour necrosis factor-a and interleukin 1b have both been shown to produce cachexic syndromes experimentally. Tumours may themselves contribute to weight loss by alteration in protein and energy metabolism. Negative nitrogen balance has been frequently documented in patients with cancer, particularly when advanced. An increase in whole-body glucose recycling via pyruvate and lactate has also been described in patients with cancer.

The loss of body weight is therefore due to an accumulation of events involving direct interference with digestive function, production of factors leading to weight loss and anorexia by the tumour, and possible alteration in protein and energy metabolism. Later in the course of the illness, antineoplastic treatment with chemotherapy, radiation, and surgery may exacerbate weight loss.

Tumour mass

It is astonishing that patients sometimes report the appearance of a swelling only to have the significance of the finding overlooked by their doctors. The appearance of any mass should lead to prompt investigation. Although imaging techniques can sometimes distinguish benign from malignant swellings, a biopsy will usually be necessary and should be taken without delay. Nowadays it is often unnecessary to undertake surgical excision biopsy. Indeed, doing so may sometimes make subsequent management almost impossible. Where there is doubt about the nature of a swelling the correct procedure will usually be needle biopsy. Biopsy is in general preferable to aspiration cytological diagnosis because the precise diagnosis of many cancers depends on architecture as well as on cytology. The great advantage of early biopsy diagnosis is that a planned approach to treatment can then be undertaken by oncologists, radiotherapists, and surgeons together. Injudicious, and often marginal, surgical excision may lead to a greatly increased risk of local recurrence of the tumour. This is a particularly frequent occurrence in sarcomas where an avoidable amputation may then be necessary or the local recurrence provoked by inadequate excision prove uncontrollable and fatal. Furthermore, for some tumours chemotherapy may be the first line of treatment and will allow assessment of response to drug treatment before surgery is undertaken; a good response may modify the need for surgery.

Fever

In cancer, fever is usually caused by infection. However, about 30 per cent of patients with cancer develop fever at some stage in their illness and it may be the presenting feature of some tumours, particularly lymphomas, renal carcinoma, and any cancer metastatic to the liver. The fever may be accompanied by sweating, particularly at night. The characteristic feature of the sweats that accompany malignant lymphomas and other cancers is that the patient falls asleep and wakes in the middle of the night to find themselves drenched with sweat. Rigors are very uncommon with febrile episodes in cancer, and should always lead to a suspicion that an infective complication is present. Characteristic patterns of fever are seldom observed; usually it is of a low-grade remittent type. The Pel–Ebstein fever of Hodgkin's disease, in which febrile periods are interspersed with several days of normal temperature, is well known but very uncommon.

The cause of the fever of malignant disease is unknown. Endogenous pyrogens may be liberated from mononuclear phagocytes in the liver or bone marrow. Tumour cells have also been shown to produce 'pyrogens'. The nature of the cytokines responsible is not clear. Exogenously administered tumour necrosis factor and interleukin 2 both produce fever and may be secreted in patients with cancer.

The fever of malignant disease may respond to simple antipyretics such as aspirin or paracetamol. In malignant lymphoma it will disappear with successful treatment of the tumour. In advanced cancer non-steroidal anti-inflammatory agents may also help, but corticosteroids are more effective, at least for a short period.

Anaemia

The anaemia of malignant disease is multifactorial (see Section 22). Chronic blood loss may occur in cancer of the gastrointestinal tract, as a result of vaginal bleeding, or because of malabsorption of iron. Usually the anaemia is normochromic, or slightly hypochromic in nature, and the plasma transferrin and serum iron are low. The iron stores are not reduced as judged by stainable iron in the bone marrow. The mechanism is discussed in Section 22.

Hypercalcaemia

Malignant disease is responsible for most of the very severe cases of hypercalcaemia seen in clinical practice. The patient will usually have widespread skeletal metastases, but occasionally the syndrome is paraneoplastic (see below). Parathyroid hormone-related protein (**PTH-rP**) may contribute to the pathogenesis of both paraneoplastic hypercalcaemia and that produced by bone metastases. For some cancers it appears that metastases in bone release PTH-rP locally and stimulate osteclastic resorption of bone. Resorption releases cytokines such as transforming growth factor-b and insulin-like growth factor 1 which, in turn, provokes more release of PTH-rP from the metastatic tumour. These cytokines may also cause proliferation of the tumour. Bisphosphonates may arrest the process by decreasing the activity of the osteoclasts. In the same way, bisphosphonates interrupt the activity of PTH-rP when this is liberated from a non-metastatic tumour as a paraneoplastic phenomenon. Bisphosphonates are important both for the treatment of hypercalcaemia and containing growth of bone metastases.

Hypercalcaemic symptoms include anorexia, weight loss, and mental confusion, all of which may simulate metastatic disease. The symptoms and signs of hypercalcaemia and its management are discussed in Section 12 and Section 22.

Paraneoplastic syndromes

Many patients with cancer have complications that are not due to direct invasion of adjacent tissues by the cancer or its metastases. The tumour produces hormones or cytokines, which are responsible for symptoms at a remote site. Alternatively, the tumour provokes an immune response to altered cellular constituents and the paraneoplastic syndrome arises from the resulting immunological reaction. Paraneoplastic syndromes are not rare but each syndrome only occurs in a minority of

patients with cancer. Furthermore, although some syndromes, such as the production of parathyroid hormone-related peptide, are found in many cancers, others, such as Cushing's syndrome, are found in a few neuroendocrine tumours.

It is important to be aware of paraneoplastic syndromes because their appearance may be the first sign of malignant disease. Furthermore, they may lead the physician into believing that the patient has metastases and thus alter management inappropriately. The syndromes themselves may cause considerable disability, which is amenable to treatment. The diversity of paraneoplastic syndromes is such that only a brief description can be given in this chapter. A summary is shown in Table 1.

Cancer can cause almost any clinical syndrome, however bizarre, and should, therefore, enter the list of differential diagnosis in any unusual clinical disorder. There are, however, dangers in making a diagnosis of a paraneoplastic syndrome as a cause of symptoms. For example, most neurological problems in cancer are not due to paraneoplastic manifestations but to the local presence of the tumour. This means that spinal cord signs in a patient with cancer are much more likely to be due to direct compression of the cord than due to transverse myelitis as a paraneoplastic syndrome. Prompt treatment of the space-occupying lesion is essential and a mistaken diagnosis of paraneoplasia is potentially disastrous. Similarly, endocrine syndromes from cancer are often caused by resectable endocrine cancers themselves. Anaemia or thrombosis may be paraneoplastic in origin but more frequently a deep venous thrombosis is due to a direct compressive effect of cancer in the pelvis and iron-deficiency anaemia should always raise the possibility of occult bleeding. Unless obviously paraneoplastic in nature, symptoms from cancer should, in the first instance, be regarded as likely to be produced by a direct effect of the tumour since this distinction has important therapeutic consequences. Some of these syndromes are described in detail in later sections: endocrine in Section 12, renal in Section 20, and neurological in Section 24.

Investigation and staging

Histopathological diagnosis

The foremost investigation of a cancer is to verify that the diagnosis is correct. Oncologists are completely dependent on the quality of the histopathological examination. Errors are not common but may be very serious, as they may lead to inappropriate investigation or the denial of curative treatment. The latter merits fuller consideration.

Misdiagnosis of a lymphoma

Lymphomas may present with histological appearances that resemble anaplastic or undifferentiated carcinoma. The diagnosis should therefore always be considered when this is the pathology report. Nowadays diagnosis of lymphoma has been made much easier as a result of immunohistochemical techniques. An example is shown in Fig. 1. The use of antibodies to leucocyte common antigen, or a combination of B-cell and T-cell markers are invaluable in diagnosis. If tumour cells do not stain, it makes lymphoma unlikely, but does not rule out the possibility. If positive, the diagnosis of lymphoma is virtually certain. Nevertheless, histologists may have difficulty either because the immunohistochemical technique is not sufficiently standardized in the laboratory, or because they mistake infiltrating lymphocytes for the tumour cells. Some undifferentiated pleomorphic lymphomas may be negative for leucocyte common antigen. These present formidable diagnostic difficulties, which may be resolved by examination of the tissue by molecular genetic techniques looking for rearrangement of the T-cell receptor genes or for immunoglobulin gene rearrangement. Other situations in which lymphoma may be overlooked, or a mistaken diagnosis made, are in the pulmonary lesions (or metastases) from small-cell carcinoma, which may be mistaken for lymphoma, or in biopsies from gastric ulcers where malignant lymphoma cells may wrongly be regarded as a chronic inflammatory cellular infiltrate. As non-Hodgkin's lymphomas can present in many different sites, where the diagnosis is not clear a prudent physician will always ask the pathologist whether the diagnosis of lymphoma has been firmly excluded when a diagnosis of chronic inflammation is made in an atypical clinical setting.

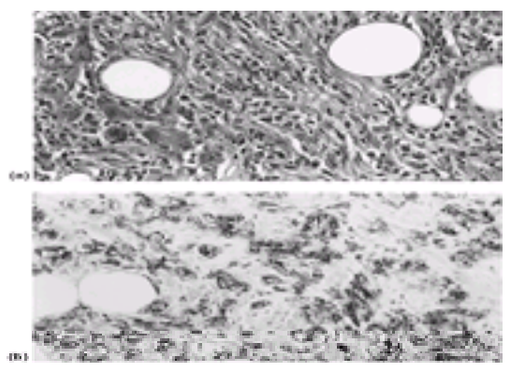

Fig. 1 (a) Section stained with haematoxylin and eosin of excision specimen of a retroperitoneal mass. The tumour is poorly differentiated, with fibrosis and infiltration of retroperitoneal fat. (b) Immunostaining with an antibody of CD20 (a B-lymphocyte marker) shows intense staining of tumour cells confirming that this is a B-cell non-Hodgkin's lymphoma.

Mediastinal or metastatic germ cell tumours

These tumours may be mistaken for anaplastic carcinoma, but the recognition of a germ cell tumour is exceedingly important because many of them are curable. Mediastinal germ cell tumours typically present in young adults and with cervical node metastases. Special stains or serum tests for a-fetoprotein or b-human chorionic gonadotrophin may be very helpful, but if negative, do not exclude the diagnosis. Several reports have indicated that the use of intensive combination chemotherapy, as for germ cell tumours at other sites, may result in lasting remissions of mediastinal poorly differentiated tumours in young adults, even when there were no other features of the germ cell nature of the neoplasm. In contrast, poorly differentiated adenocarcinoma in the mediastinum of young adults can seldom be ascribed to germ cell tumour, although occasional cases may respond dramatically to chemotherapy.

Investigating the local extent of the tumour

Following diagnosis, clinical staging is the most important first procedure. Clinical examination will often establish the likely extent of the tumour. This may require specialized techniques such as ear, nose, and throat examination and bronchoscopy. The extent of infiltration and fixation to surrounding structures is assessed. CT scanning and MRI have greatly improved the preoperative determination of tumour extent. They have largely replaced more invasive techniques such as angiography and lymphangiography. MRI is particularly valuable in the staging of sarcomas and central nervous system tumours. Both techniques show the extent of the tumour and infiltration of surrounding structures (Fig. 2(a) and Fig. 2(b)). CT scanning is a particularly valuable aid to needle biopsy diagnosis of deep-seated tumours.

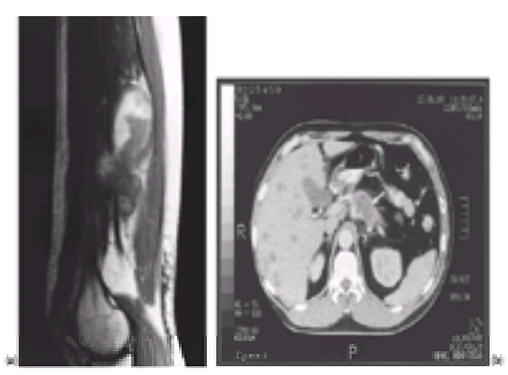

Fig. 2 (a) Longitudinal MRI of lower thigh. A large, soft tissue mass is seen displacing the muscle groups posteriorly. It lies behind the femur and the femoral artery is in close proximity to the mass, which at one point surrounds it. (b) CT scan of abdomen. A carcinoma of the body of the pancreas is shown (arrowed). The liver

contains numerous small metastases.

Staging of lymph node spread

Spread to adjacent lymph nodes may be noted clinically or on straightforward investigations, such as chest radiography. Lymphangiography was used formerly to examine pelvic and lower para-aortic nodes but has largely been supplanted by CT. In fact the two techniques give slightly different information, as a CT scan will show enlarged lymph nodes (the assumption being that these are replaced by tumour when the lymph nodes become more than 2 cm in size), while lymphangiography may show abnormal appearances even when the nodes are not enlarged but contain foci of tumour cells within them.

Staging for distant metastases

Bone metastases are usually demonstrated by $^{99}Tc^{m}$-polyphosphate isotopic scanning. The sensitivity of the examination is high and abnormalities frequently precede detectable changes on plain radiography. However, the specificity is rather lower because any traumatic or inflammatory disorder in bone can give areas of increased uptake. When areas of increased uptake are seen on technetium scanning it is important to follow up with plain skeletal radiography, particularly in the long bones of the limbs. This is because isotope scanning gives no indication of the structural integrity of the bone and the risk of pathological fracture in a limb cannot be assessed on an isotope scan.

Liver metastases are detected by an increase in circulating enzyme levels, particularly alkaline phosphatase and serum glutamic oxaloacetic transaminase. Lactic dehydrogenase is also elevated in a somewhat greater frequency. Nevertheless, liver metastases can be present without alteration in serum enzyme levels and ultrasound scanning is an invaluable non-invasive method of detecting liver metastases. CT and modern ultrasound scanning are approximately equal in sensitivity. Metastases down to 1 cm in size can be detected reliably.

Pulmonary metastases may be detected on chest radiograph but may be present even when the chest radiograph is normal if they are below 1.5 cm in size. Metastases larger than this may also be overlooked if they are situated behind the heart or behind the diaphragm. CT scanning is the best method for demonstrating pulmonary metastases and lesions as small as 0.4 cm in diameter may be seen. CT scanning is therefore an essential investigation in patients who are to undergo extensive or mutilating surgery, such as for sarcomas where metastases to the lungs are particularly frequent and the presence of metastases may influence the surgical decision.

Brain metastases are detected by CT scanning or, more reliably, by MRI. In a patient who is neurologically normal there is only a low chance of detecting asymptomatic cerebral metastasis by these methods (about 5 per cent). For this reason the technique is not worthwhile as a routine investigation of most cancers.

Surgical staging of cancer

Surgery specifically for staging rather than treatment is reserved for a few specific tumour sites. In lung cancer, investigation of the mediastinum is extremely important in deciding whether a tumour is operable. CT scanning may demonstrate inoperability either because the tumour is infiltrating the mediastinum or because there is lymph node spread to both ipsilateral and contralateral hilar nodes. However, in other patients, the mediastinum may appear normal and a mediastinoscopy may reveal tumour in mediastinal nodes implying the inoperability of the condition. Staging laparotomy used to be performed in localized Hodgkin's disease, but is now reserved for specific indications. In ovarian cancer thorough surgical staging is performed at the time of the initial resection, but surgical staging is in this case (as in many other tumour resections) part of the treatment.

An important recent development has been the introduction of so-called 'sentinel node' biopsy. In this technique a radioactive tracer, or a blue dye, is injected in the vicinity of the tumour, or into the tumour itself, and the lymph node at the first adjacent site of uptake is sampled, either by biopsy or surgical removal. In the case of breast cancer, the disease in which the technique is exciting most interest, the sampling may be at the time of operation. The presence or absence of tumour cells in the lymph node is taken as an indication of whether lymphatic spread has occurred and surgery, and subsequent treatment, can be modified accordingly. There are unresolved issues about accuracy of the technique, and the confidence with which the results can be used to plan therapy, but it is likely that the procedure will become part of the lymph node staging of some cancers.

The use of a staging notation

This has been valuable in the reporting of results of cancer treatment and is also helpful, in an individual patient, in focusing attention on the extent of the disease and the subsequent planning of treatment. A widely used system is the **T** (tumour) **N** (nodes) **M** (metastasis) system. This is particularly valuable for tumours that follow an orderly progression of spread from the primary site to adjacent lymph nodes and then to metastatic sites. Thus, tumours of the head and neck, breast, non-small-cell lung cancer, renal carcinoma, bladder carcinoma, and rectal carcinoma are all well-defined by this means. In addition to the TNM system, many classifications contain a stage grouping, by which tumours with varying TNM assignments are grouped together because of equivalence of prognosis or similar approaches to management. An example of the TNM staging system and stage grouping for lung cancer is given in Table 2.

Not all tumours can be summarized by the TNM system. For example, small-cell lung cancer is usually widely metastatic at the time of presentation and a simpler classification into limited (confined to one side of the thorax with ipsilateral supraclavicular node) or extensive (disease that is bilateral within the chest or metastatic) is used. This simple classification serves to separate patients in whom radiation treatment may be worthwhile and those in whom it is unlikely to have any benefit. In leukaemia and myeloma, other staging criteria, which are based on prognostic factors and are not related to anatomical stage, have been developed. In Hodgkin's disease and non-Hodgkin's lymphoma, the presence (B) or absence (A) of constitutional symptoms is added to the anatomical staging system, which is used to define the degree of lymph node spread. These additions were made because the presence of constitutional symptoms confers an adverse prognostic significance in addition to the prognosis related to the anatomical stage (see also Section 22).

Principles of cancer management

The principles and details of cancer chemotherapy are discussed elsewhere. This section summarizes an integrated approach towards cancer management.

Nowadays the management of cancer will nearly always involve more than one specialist and more than one type of treatment. Increasingly, patients with cancer are being seen in joint clinics where surgeons, medical oncologists, and radiotherapists plan treatment. Often there will be several possible approaches towards treatment, and these require discussion and assessment by the appropriate experts. It is of inestimable value if a patient is referred for expert opinion before any definitive procedure is undertaken. For example, more information about gynaecological malignancy can often be obtained if a patient with abdominal swelling and ascites and an ultrasound-demonstrable mass in the pelvis is assessed preoperatively by a gynaecological oncologist. The subsequent laparotomy is likely to reveal much more information than if it is carried out as an emergency by an inexperienced surgeon. Similarly, a mass on a limb should be investigated thoroughly, including a biopsy diagnosis, before surgery is undertaken, because the nature of the histological diagnosis may profoundly alter management in the case of a sarcoma. Table 3 lists tumours in which radiotherapy and chemotherapy have an important part to play in management and where these modalities of treatment may sometimes be curative.

Surgery

Surgeons see over 80 per cent of patients presenting with cancer for the first time. Following diagnosis and staging to exclude metastases, curative surgery may be undertaken, for example in breast or colorectal cancer. The aim of the operation is complete excision of the tumour with a margin of normal, uninvolved tissue around the main tumour mass. The risk of local recurrence is very high with a marginal excision in which a tumour has been 'shelled out' because the pseudocapsule around the tumour is likely to be infiltrated with tumour cells. Removal or sampling of the draining lymph nodes will often be undertaken, for example in breast cancer and other tumours where involvement of regional lymph nodes is likely (see the discussion of sentinel node biopsy above). In some cancers, such as breast cancer, it has become clear that very extensive primary tumours are usually accompanied by distant metastasis. In this situation the role of surgery is to prevent local recurrence rather than to be curative. With other tumours, for example of the head and neck, extensive surgery may be the only means of gaining effective control and in these cases a considerable degree of surgical expertise is necessary. In some situations the tumour may be approached either by surgery or by radical radiotherapy and

there may be little to choose between the results. An example is in early prostate cancer where the results of radical radiation and surgery are probably equivalent, and in operable oesophageal cancer, particularly of squamous histology, where long-term results of radiation may be the same as those of surgery. In these situations the benefits of local control, survival, and long-term side-effects have to be judged together in making a decision.

Nowadays local treatment frequently involves surgery and radiation to maximize the chances of local control. Wide local excision is increasingly practised in carcinoma of the breast and radiation to the breast and to axillary nodes is used as an adjunct. Radiation reduces the risk of local relapse, both in the breast and in the axilla. 'Lumpectomy' and radical radiotherapy have now replaced mastectomy for many patients with small primary breast cancers. Preoperative radiation of soft tissue sarcoma may increase the chance of successful compartmental excision of the tumour, and postoperative radiation decreases the risk of local recurrence in patients in whom the excision has been marginal. These are two examples of the way in which the definitive local management of the primary tumour is a matter of discussion between surgical and radiation oncologists.

Optimum local management has become further complicated by the successful use of chemotherapy in many tumours. An example is the treatment of Ewing's tumour. In this highly malignant round-cell tumour of childhood, initial chemotherapy usually produces a prompt regression of the main tumour mass, both in the bone and in the surrounding soft tissues. However, the tumour permeates widely through the bone, and local irradiation, given after initial chemotherapy, is a standard means of maintaining local control. There is an increasing tendency to continue both chemotherapy and radiation synchronously. However, in large tumours, even with full-dose radiation, the risk of local relapse is still present. For this reason surgery is being used increasingly, provided that the cosmetic and functional results are reasonable. Surgical excision alone may be successful after chemotherapy, but frequently, because of the permeating nature of the tumour, viable tumour is present right up to the resection margins of the bone. In this situation radiation will be needed in addition to the chemotherapy and surgery. In these tumours, very detailed planning of the approach to treatment by experienced specialists is essential for optimum results.

Specific management problems

Spinal cord and cauda equina compression

Cord and cauda equina compression are common and devastating complications of metastatic cancer. For successful management it is essential to remember one rule—every hour counts. Even if early treatment is not always successful, delay ensures that the patient will end his or her days bed- or chair-bound, paralysed, and incontinent.

The metastasis often develops in a vertebra, from which it spreads directly or via the intervertebral foramina to compress the cord (or cauda equina below L1) from the extradural space (Fig. 3). Alternatively, the malignant mass may originate in a mass of retroperitoneal nodes, or the primary tumour (for example a bronchial carcinoma) may be in the posterior mediastinum or retroperitoneum (Fig. 3). Damage to the cord is by direct compression and by interruption of the arterial supply leading to infarction. It is uncommon for the tumour to be metastatic to the cord itself, although meningeal spread occurs and may cause compression (see carcinomatous meningitis, below). Cord compression may be the first manifestation of cancer but more commonly arises with metastases from a known primary.

Fig. 3 MRI of thoracic vertebras showing destruction of the body of T10 by a mass of Hodgkin's disease. The tumour extends posteriorly and compresses the spinal cord. The tumour mass has passed to the side of the vertebra (not shown), and is also compressing the cord posteriorly after infiltrating into the intervertebral foramen.

Pain often precedes the onset of neurological symptoms. In the case of cord compression it is felt in the thoracic and cervical vertebras. It is worse on coughing. An exceedingly sinister symptom is vertebral pain with a root distribution. A patient with this symptom needs urgent investigation, as cord compression may be imminent. The next symptom is usually weakness of the legs combined with sensory loss, of which loss of proprioception is especially characteristic. Loss of bladder and bowel sensation is late and once weakness and bladder disturbance begins, progression to irreversible paraplegia occurs in hours or a few days.

The patient often has a sensory level, motor weakness, brisk leg reflexes, and extensor plantar responses. The bladder may be palpable. Radiography of the spine often shows vertebral destruction—loss of a pedicle or compression of the body being typical. Myelography will demonstrate the block, but MRI (and, less reliably, CT scanning) is very valuable and has now largely replaced myelography. Treatment usually consists of surgical decompression, although for radiation-sensitive tumours such as lymphoma or Ewing's sarcoma, high-dose corticosteroids and radiation will produce quick relief of compression. If there are multiple sites of block, radiation and steroids may be the only feasible option. The surgical approach to decompression varies according to the nature of the lesion—whether anterior or posterior, cervical or thoracic. Anterior decompression may involve removal of part of the vertebral body, but the risk of destabilization of the spine means that immediate stabilization may be necessary. It is not clear whether radiation is inferior to surgical decompression in patients with tumours that show sensitivity to radiation. Radiation is, in any event, usually given after laminectomy.

Outcome is crucially dependent on the functional state of the patient before treatment. Less than 10 per cent of those who are paraplegic before treatment will be able to walk later, 25 per cent will do so if they have some motor function preserved, while almost all patients who can still walk will continue to be able to do so.

Cerebral metastasis

Cerebral metastasis is clearly a serious complication of cancer occurring in about 30 per cent of all patients. Metastases are over 10 times as common as primary brain tumours. About 15 per cent of patients with cancer will develop symptomatic brain metastases during life. Thus, there will be approximately 15 000 deaths each year in the United Kingdom of patients with symptomatic cerebral metastasis. Metastasis at this site is life-threatening and disabling, causing severe deterioration in quality of life and great difficulty for the patient and his or her carers.

Most cerebral metastases are intradural, usually in the substance of the brain extending to the meningeal surface. About 80 per cent of these are situated in the cerebrum and the rest in the cerebellum and other regions. Lung cancer and breast cancer are the most common primary sites, and certain tumours are particularly associated with single metastases (cancer of the breast, ovary, and kidney). Usually cerebral metastases occur following diagnosis and treatment of a primary tumour.

In the brain substance, the metastases are vascularized from the cerebral circulation, but there is no evidence that a vascularized metastasis maintains a 'blood–brain' barrier—the vascularization is, after all, of non-nervous tissue without the tight endothelial junctions which characterize cerebral capillaries. Indeed, capillary leakiness appears to be a feature of cerebral metastasis and is responsible for the substantial amount of oedema of the brain, which typically accompanies cerebral metastasis. The blood–brain barrier may, however, be an important impediment to cytotoxic treatment when the metastasis is being established before it is vascularized and at the periphery of an established metastasis. It will be a very significant factor in failure of treatment of leptomeningeal cancer.

Symptoms and signs

The typical signs are headache, disturbance of cognitive function and effect, focal fits or grand mal convulsions, and limb paresis. Headache usually reflects a rise in intracranial pressure. It is typically present in the morning and increases in duration and frequency until other signs of raised intracranial pressure become apparent.

Focal weakness is present in about half of all patients, and disturbance in higher cerebral function in about 60 per cent.

Investigation

CT scanning or MRI is the essential diagnostic investigation. On a CT scan most metastases appear hypodense but enhance with contrast material. Typically there will be oedema around the metastases. Occasionally CT scans may be normal even in patients whose symptoms strongly suggest cerebral metastasis and where cerebral metastasis is sometimes proved by further scanning some weeks later or at autopsy. In these patients there may be multiple small metastases without oedema or leptomeningeal spread. MRI has a greater degree of sensitivity and is particularly valuable in detecting leptomeningeal spread of tumour. In the presence of a known primary it is not usually necessary to subject patients to histological confirmation of the tumour. However, after a very long disease-free interval, or where the primary is unknown, histological diagnosis will be essential.

Treatment

Dexamethasone is started as soon as the diagnosis is made. The usual dose is approximately 16 mg/day, although higher doses can be used if the patient does not respond. The clinical effects are rapid and usually noticeable within 24 h. The maximum effect is achieved in about 4 days. Approximately 80 per cent of patients will respond. Phenytoin or carbamazepine are used to control focal fits.

The most useful non-surgical treatment is radiation therapy. The therapeutic doses depend on the likely primary site, but usually consist of 30 Gy in 10 fractions in 2 weeks, or 40 Gy in 15 fractions in 3 weeks. The former is the most widely used schedule in the United Kingdom and no schedule has been proved to be superior over another. Solitary cerebral metastases may be removed if they are in an accessible site. The criteria for operation are usually that a solitary metastasis is present, that the diagnosis is uncertain, or that the response to radiation is unpredictable because of doubt about the nature of the primary tumour. The patient must be clinically fit in other respects to undergo surgery, and without life-threatening metastatic disease elsewhere.

There has been recent interest in the use of chemotherapy in the treatment of cerebral metastasis, as it is now clear, for example in small-cell lung cancer, that the response to chemotherapy in cerebral metastases is equal to that in metastases at other sites. Responses to chemotherapy in tumours such as small-cell lung cancer may be rapid and dramatic but cranial radiation will usually be necessary as an adjunct to chemotherapy.

Prognosis

The prognosis depends on the clinical setting. If there is a solitary metastasis with no disease elsewhere then a long disease-free interval may result, particularly if the metastasis has occurred after a considerable interval following the primary treatment. In other tumours, where multiple metastases occur either synchronously with the primary tumour or after a short disease-free interval, and where the tumour is a particularly difficult type to treat (such as melanoma and non-small-cell lung cancer), the prognosis is very poor indeed. Overall, only 30 per cent of patients will be alive at 1 year and the median survival is about 7 months. A small randomized trial has suggested that surgical resection of a solitary metastasis adds to survival when compared with radiation and steroids alone.

Carcinomatous 'meningitis'

Leptomeningeal spread of cancer seems to be increasing in frequency. In autopsy series about 4 per cent of patients dying of advanced cancer have leptomeningeal spread. The frequency is higher in breast cancer (5 to 10 per cent). This complication is increasing in lymphoma, small-cell lung cancer, ovarian cancer, and some sarcomas. Curiously, adenocarcinomas seem to have a greater propensity for this form of metastasis than other epithelial tumours. There may or may not be intracerebral metastasis at the same time. Malignant cells may enter the cerebrospinal fluid from intracerebral tumour via the arachnoid, or from vertebral deposits growing along nerve roots into the subarachnoid space. However, the most likely source of seeding appears to be directly from the bloodstream. Tumour is present as a thin covering of malignant cells, but in some cases the tumour cells penetrate deeper into the substance of the brain along blood vessels. The tumour may also penetrate cranial and spinal cord nerves as they pass through the subarachnoid space.

Clinical features

The onset is usually over a few weeks and may be subtle at first. Headache is often severe and is due to raised intracranial pressure. Cranial nerve dysfunction is frequent, with diplopia, hearing loss, and facial numbness. There is often back pain and sometimes bladder and bowel dysfunction. A change in mental state may occur. Focal fits are uncommon. On examination there may be an abnormal mental state, signs of raised intracranial pressure, and extensor plantar responses. Focal neurological signs in the limbs are uncommon. Cranial nerve weaknesses are frequent, the most common being ocular muscle palsy, facial weakness, and hearing loss.

Diagnosis and treatment

The diagnosis is made by examining the cerebrospinal fluid. Typically, the opening pressure is high, the white count is raised, the cerebrospinal fluid sugar low, and the protein increased. Cytological confirmation on the first lumbar puncture is obtained in about 60 per cent of patients, but a negative examination does not exclude the diagnosis. Myelography may show typical appearance of multiple small tumour seeds in the subarachnoid space, but MRI is proving invaluable and is now the preferred initial investigation if cerebrospinal fluid cytology is negative and the diagnosis strongly suspected.

Treatment is difficult and often unsuccessful. Temporary improvement can be obtained by the insertion of an intraventricular reservoir to deliver chemotherapy. Chemotherapy administered by lumbar puncture is uncomfortable and may not be effective if there is meningeal invasion supratentorially, since the drugs do not penetrate in high concentration beyond the foramen magnum. In breast cancer and lymphoma, intrathecal methotrexate is effective and may be administered in combination with thiotepa or, in the case of lymphoma, cytosine arabinoside. In addition, whole-brain irradiation is often given if the patient is improving and the clinical situation indicates that this treatment would produce further benefit. In general, however, the prognosis is poor when the meningeal infiltration is from an epithelial tumour, with a median survival of only 4 months.

Pleural effusion

Malignant pleural effusions occur either as a sign of metastasis or due to direct invasion of the pleural space from an underlying primary bronchial carcinoma, or pulmonary metastasis. The effusions are typically exudates with a protein content of more than 3 g/dl. There is increased capillary permeability through inflammation and abnormal capillary endothelium in the tumour lining the pleural space. Typical primary sites are: breast and ovarian cancer, as common epithelial tumours metastasize into the pleural space; lung cancer, as a cause of pleural effusion with underlying lung disease; and sarcomas, as a cause of pleural effusion due to invasion of the pleura by pulmonary metastasis.

Clinical features

The typical features are dyspnoea, which is directly related to the size of the effusion, dry cough, and chest wall discomfort. Even a small effusion may cause dyspnoea in a patient who has underlying lung disease such as chronic bronchitis and emphysema. Many patients have asymptomatic pleural effusions detectable on chest radiograph. The sequence of radiological appearances includes blunting of the costaphrenic angle (with volumes of 2 to 3 ml), increasing effusion, and, finally, mediastinal shift, which usually occurs when amounts in excess of 2 litres have accumulated. Ultrasound examination may assist in localizing the effusion and any loculi, which may influence the procedure for aspiration.

Diagnosis

The diagnosis, if the primary tumour is not known, is made by demonstrating malignant cells in the pleural fluid. The rate of positivity, in patients known to have an underlying cancer, is about 60 per cent with a low false-positive rate. If pleural cytology is negative on the first aspiration it should be repeated using fresh aspirates. Occasionally, pleural biopsy will be necessary to make a diagnosis, and the combination of the two increase the diagnostic yield to about 90 per cent. If both techniques fail, thoracoscopy is more successful, but is, of course, more invasive.

Treatment and prognosis

The primary tumour should be treated if possible. When a pleural effusion persists after treatment of the primary tumour, or if such treatment has been unsuccessful, treatment may need to be directed to the effusion itself. Frequently the effusion will need to be aspirated in order to make the patient comfortable, and pleural sclerotherapy considered. For best results of sclerotherapy it is important to drain the pleural cavity as completely as possible. A small flexible chest drain is ideal and is left in place for some time (12 to 24 h if possible) to allow the fluid to drain as far as possible. If there has been loculated effusion, the insertion of the drain is best done under ultrasound control. Sclerosis of the two pleural surfaces can be achieved by a variety of means; all give approximately equivalent results. The most favoured techniques are the instillation of talc, tetracycline, bleomycin, or *Corynebacterium parvum*. They all cause an inflammatory reaction in the pleural space and have an approximately 60 per cent success rate in preventing immediate recurrence of the effusion. When pleural effusion complicates an underlying bronchial carcinoma it is more difficult to control than when it is a metastatic manifestation of a distant neoplasm, such as ovarian cancer. If the effusion is recurrent and is the major cause of morbidity, pleuroperitoneal shunting can be carried out, whereby the pleural fluid drains into the peritoneal cavity.

Pericardial effusion

The most common malignancies to cause pericardial effusion are breast, lung, ovary, and gastrointestinal cancers and non-Hodgkin's lymphomas. Pathologically, the pericardium may be infiltrated with tumour or diffusely nodular. The accumulation of fluid is due to obstruction of lymphatic and venous drainage of the pericardium.

Symptoms and signs

The symptoms are usually vague in onset, including orthopnoea, dyspnoea, and cough. Fatigue and dizziness also develop. If cardiac tamponade occurs it is associated with severe dyspnoea, vague central chest pain, and anxiety. The physical signs are usually minimal, although when tamponade occurs there will be jugular venous distension, pulses paradoxus, hypotension, and tachycardia.

Investigation

Investigations include a chest radiograph, which shows enlargement of the cardiac silhouette, and echocardiography, which is a rapid non-invasive technique for demonstrating pericardial effusion.

Diagnosis and management

The diagnosis is made by finding malignant cells in the pericardial fluid. False negative results occur and the test may need to be repeated. Once the diagnosis has been established, the pericardial fluid may need drainage using a small rubber catheter. Installation of sclerosants can be carried out as for pleural effusions, but troublesome pericardial effusions can be controlled by the formation of a pericardial window through a small left anterior thoracotomy. Some patients, particularly those who have lymphoma, will respond to external-beam radiation with a dose of approximately 30 Gy given in 15 fractions over a 2- to 3-week period. Radiation is also considered for control of chronic pericardial effusion in breast cancer. The management of cardiac tamponade is discussed in Section 15.

Metastatic cancer from an unknown primary site

Approximately 3 per cent of patients present with a metastasis from a cancer where the primary site is not known after full history, physical examination, blood count, and chest radiograph. This clinical situation requires considerable clinical expertise, as the diagnosis creates especial anxiety for the patient. The clinician has to decide on the most effective therapy and to sustain the patient without indulging in futile, invasive, and expensive investigations which will not alter management. The problem with extensive investigations is that they seldom alter management and the overall prognosis in this position is poor (4 to 6 months median survival). As one investigation after another fails to reveal the primary site, the patient and the doctor may come to consider this a failure and confidence can be badly shaken. Nevertheless, some tumours are potentially curable and, for these, investigation is justified. The common primary sites, when one is discovered, are cancers of the lung, pancreas, liver, gut, and stomach. The tumours for which therapy is possible, and which therefore must not be overlooked, are listed in Table 4.

Presentation

If the presentation is exclusively in cervical nodes, a full ear, nose, and throat examination is mandatory as local treatment with surgery and/or radiation may produce prolonged survival or even cure. The higher the cervical node, the more likely it is that and ear, nose, and throat tumour is the primary source. Supraclavicular lymph nodes carry a worse prognosis because the likely primary site on the right-hand side is the lung or breast, and on the left-hand side intra-abdominal malignancy via the thoracic duct. Patients presenting with lymph node enlargement in the axilla are likely to have breast cancer as the primary site and this may not be excluded even with normal mammograms. Malignant melanoma is another possibility at this site and a careful examination for skin lesions should be made. Inguinal lymph nodes usually point to a primary site in the pelvis, vulva or rectum, or prostate. Malignant melanoma may present with an inguinal mass. Cutaneous metastasis typically occurs from carcinomas of the lung, breast, and melanomas. A pulmonary metastasis may arise from a variety of different sites, including breast, kidney, gut, melanoma, and sarcoma. In the liver, the likely source for the primary will be the gastrointestinal tract, although breast and lung primaries are other possibilities. A metastasis presenting in bone is particularly likely to occur from a cancer of the lung, breast, or prostate, the last being particularly likely if there is a mixed lytic and osteoblastic radiological appearance.

Investigation

The most important single investigation is a review of the histology. The clinician should discuss the diagnosis with the pathologist so that appropriate tests can be carried out. It is absolutely essential to distinguish between an epithelial tumour, a sarcoma, and a lymphoma. Immunohistochemistry may be invaluable in this respect. If there is any question of a germ cell tumour, the section should be stained for a-fetoprotein, b-human chorionic gonadotrophin, and placental alkaline phosphatase. If the histology is that of adenocarcinoma, the diagnosis will be more difficult and special stains may not serve to elucidate the diagnosis further. Where possible, the tissue should be examined for the presence of oestrogen or progesterone receptor, as this would make carcinoma of the breast or ovary more likely. The protein S100 is typically present in melanoma and may be invaluable in distinguishing this diagnosis from anaplastic carcinoma.

Further investigation and management

Investigation must be selective. Since there is specific treatment available for breast and prostate cancer, these diagnoses must always be considered when the histology is adenocarcinoma. Mammography is therefore justifiable, and measurement of serum acid phosphatase and prostatic specific antigen are simple and non-invasive. A pelvic ultrasound may show an ovarian mass, which may influence management as platinum-based combination chemotherapy might then be used, whereas it would not be contemplated in many patients with metastasis from an unknown primary site in view of its toxicity. The possibility of a germ cell tumour must always be considered in a young person, and in these circumstances full investigation is necessary if this diagnosis is possible.

Treatment follows pragmatic lines. Locally troublesome or painful metastases are treated with irradiation. If breast cancer seems a possible diagnosis a trial of hormone therapy is fully justified and, similarly, hormone treatment of prostatic cancer should be introduced if this seems a likely diagnosis. As mentioned above, radiation is frequently given to patients with enlarged cervical nodes when the diagnosis is poorly differentiated carcinoma, even if a head and neck primary has not been found.

The use of combination chemotherapy when the primary site is not known is much more controversial. In general, responses are infrequent and are not long lasting. This drug treatment should be reserved for patients with more than one lesion and particularly when symptoms occur. It is important not to be dogmatic about this issue because many patients find it quite unacceptable to be told that no treatment of any kind is available to them, and are willing to accept the possible toxicities of chemotherapy in exchange for the chance of response. Most chemotherapy programmes will include an alkylating agent and some include doxorubicin or a taxane.

Supportive care of the patient with cancer (see alsoSection 31)

Psychological support

Nearly everyone will have had friends or relatives who have had cancer and who may have died of it, and they will have read articles and seen television programmes about cancer and its management. Many patients will have been worried about the possibility of cancer before they ever consult their general practitioner, or are subjected to a series of diagnostic tests, the effect of which may be to increase their anxiety. At each stage in the diagnostic process physicians should be aware of patients' feelings and be prepared to talk openly to them about why investigations are being performed. When the diagnosis is established it is essential for the physician to sit quietly with the patient, explaining the nature of the diagnosis and the broad principles that treatment will follow. Sometimes patients will like to have a member of the family with them during this conversation, in case they forget aspects of what is said. The conversation should take place quietly, not on a ward round, with both the patient and the physician seated and the physician calm and unhurried in approach. Avoidance of the word 'cancer', body language that indicates discomfiture or embarrassment, and evasion and vagueness are very likely to be interpreted by the patient as signs of a serious or hopeless outlook.

Many patients will be unable to take in all that is said in the first conversation, and the physician needs to make it quite plain that he or she will be very pleased to talk again the next day, to go over points that need further clarification. There is much useful literature for the patient to take home, there are professional and expert support groups that the patient can contact and, in many hospitals, skilled counsellors who can provide follow-up support after the physician has outlined the basis of treatment. It is essential that all members of the medical team understand what was said and what words were used. The members of the family also need to understand exactly what information has been imparted. It may be necessary to hold back on a precise prognosis; first, because one may not be known until treatment starts, and second, because patients naturally tend to become fixated on the numerical prognosis, which is likely to be extremely inaccurate. If referral to an oncologist is to be made it is critical to indicate exactly what has been said to the patient. Oncologists are put in an extremely difficult position when patients arrive with a diagnosis of cancer, without any indication at all of whether they know the diagnosis, or what words have been used.

A new difficulty in communication is now displacing that which formerly arose from concealment of the diagnosis. Modern cancer management is often complex, with equivalent results sometimes being obtained from approaches that have different early and late effects. A well-intentioned wish to 'share' the treatment decisions with the patient, and an increasing resort to litigation when events don't turn out well, has led doctors sometimes to present treatment options as a series of uncertainties in which the outcome will be strongly influenced by chance and fate. It is bad enough to be told you have cancer. Worse still if your treatment seems mired in uncertainty. For some patients, treatment options will have been made even less clear by access to unfiltered advice on the internet, from reputable sources, charlatans, and cranks. There is nothing paternalistic in sensible advice from a well-informed, kind, sensitive, and experienced specialist. Questions which arise from complexity of choice and outcome in management increase the need for competent advice; they are not answered by passing the problem to the patient. Much distress, and a feeling of being abandoned, can come from lack of clear guidance.

When treatment is to be palliative, after relapse or with widespread metastatic disease, it should none the less be made clear to the patient that it is 'treatment'. Patients dislike feeling that they are being abandoned. Indeed, many wise oncologists see their patients more frequently when they are having palliative treatment than they do during routine treatments or follow-up. They do this because palliative treatment requires great attention to detail, especially with respect to control of pain and other symptoms, and also to provide psychological support for the patient and the family. One of the most common reasons for patients seeking second opinions is that they have been given no feeling that there are possibilities for treatment in their case. Continuity in management is one of the most rewarding aspects of cancer medicine for the physician and for the patient. There is no place for impersonal clinics where patients see different doctors each time they attend, and where the emotional component of their illness cannot be properly explored.

Management of cancer pain

Pain is a common and distressing feature of cancer. A careful history is essential to determine the exact site and nature of the pain and to establish a close and trusting relationship with a patient who feels that the symptom is being taken seriously. Exacerbating factors should be noted and an anatomical diagnosis made as far as possible. If the pain is arising in a bone it may be quickly and effectively helped by radiation treatment. The primary tumour or metastasis may be responsive to treatment with irradiation or chemotherapy. If specific antitumour treatments of this kind are not appropriate, then the only approach is to control the pain with analgesics.

Non-narcotic analgesics are used for mild or moderate pain. Useful agents include aspirin, paracetamol, and non-steroidal anti-inflammatory drugs such as ketoprofen or naproxen. A combination may be useful. Combination drugs such as co-proxamol or co-dydramol are also helpful. Although prescribing each drug separately allows greater control over the constituents, in practice this may not be helpful, particularly for elderly patients who often find it difficult to take multiple medication. The aim of treatment should be to prevent pain as far as possible by taking regular analgesics, and to have additional analgesics on hand for an acute exacerbation. Side-effects of non-opiate analgesics include gastric irritation (and they should therefore be used cautiously if steroids are being used at the same time), nausea, and constipation, particularly with codeine, oxycodone, or propoxyphene.

If these analgesics do not control the pain, opiate analgesics are essential. Two preparations have made an enormous contribution to pain relief. The first is long-acting morphine sulphate, which can be given twice daily, and the second is short-acting morphine sulphate (Sevredol). The former has a duration of action of 8 to 12 h and the latter of about 4 h. One curious feature of the use of morphine-like drugs is that the dose required to control pain varies greatly from person to person. It must therefore be found by trial and error and the patient must be prepared to increase the dose under medical supervision. The aim is to produce background pain relief for most of the day and night. Sevredol is particularly useful for dealing with acute exacerbations of pain.

If oral opiates are unable to control pain fully, continuous subcutaneous infusion is a useful alternative. This approach is particularly valuable in patients who cannot tolerate oral analgesics because of gastrointestinal symptoms, or where the tumour causes nausea or intestinal obstruction. Many pumps are now available, which are designed for continuous infusion through a small-gauge butterfly needle implanted subcutaneously. Patients can manage at home with these infusion pumps, with a nurse calling daily to change the infusion mixture.

Specialized forms of analgesia

A detailed discussion is beyond the scope of this chapter. Amongst the specialized techniques available are continuous epidural and intrathecal opiate infusion, nerve block procedures (including coeliac plexus block, peripheral nerve block, and epidural blocks), neurosurgical procedures, such as ablation of the peripheral nerve by neurectomy or, more radically, interruption of pain pathways by cordotomy. Each of these procedures has its value and limitations and the advice of specialists in the field of pain relief will be necessary.

Further reading

deVita VT, Hellman S, Rosenberg SA (1998). *Cancer: Principles and practice of oncology*, 5th edn. Lippincott, Philadelphia.

Souhami RL *et al.* (2001). *Oxford textbook of oncology*, 2nd edn. Oxford University Press.

6.7 Cancer chemotherapy and radiation therapy

Michael L. Grossbard and Bruce A. Chabner

The last two decades have brought significant improvements in cancer therapy. Patients with previously fatal diseases, including acute leukaemia, non-Hodgkin's lymphoma, Hodgkin's disease, and germ cell tumours, can have a reasonable expectation of cure. For patients with commonly occurring solid tumours, including lung cancer and breast cancer, several new chemotherapeutic agents have been developed which offer improved treatment options and enhanced survival. An increased understanding of tumour biology and the explosion in knowledge in molecular biology have paved the way for new developments in cancer treatment.

Nevertheless, cancer remains the second leading cause of death in the United States. Although exciting advances continue to be made in cancer therapeutics, nearly 40 per cent of patients diagnosed with cancer will die of their disease. An estimated 1.2 million new cancer cases will occur in the United States in 2000 and 552 000 patients will die cancer-related deaths. Virtually all patients diagnosed with advanced stage solid tumours will succumb to their tumour or complications of its therapy. Because a medical oncologist and radiation oncologist will manage many patients diagnosed with cancer only after initial referral from an internist or surgeon, it is critical for the primary care physician to understand the general principles of cancer therapy. The dramatic increase in the number of effective chemotherapy agents since nitrogen mustard was introduced more than 50 years ago has made this challenge greater.

Chemotherapy and radiation therapy, together with surgery, are the major modalities of cancer therapy. Chemotherapy has the advantage of targeting tumour cell throughout the patient, while external beam radiation therapy acts to provide local control. Often, the two modalities are combined to take advantage of synergistic cytotoxicity. This chapter will describe major principles of chemotherapy and radiation therapy.

Chemotherapy

A knowledge of cancer chemotherapy requires an appreciation of some general principles of tumour biology. Cancer results from the uninhibited growth of a single clone of cells. As cancer cells grow, they move through the cell cycle, characterized by several phases: resting (G0), pre-DNA synthesis (G1), DNA synthesis (S), post-DNA synthesis (G2), mitosis (M). Most chemotherapy drugs are active in the S phase of the cell cycle, although some directly block cells entering mitosis and most directly promote apoptosis (programmed cell death).

Chemotherapy can be used in several different settings (Table 1). Foremost, chemotherapy is applied as primary therapy for the treatment of advanced-stage cancer. A few diseases, including leukaemias, lymphomas, and advanced-stage germ cell tumours are sensitive to multiple chemotherapy agents and can be cured with combination chemotherapy. More often, combinations of therapeutic agents are used to diminish tumour-related symptoms, improve the quality of life, and extend survival in patients with advanced-stage tumours. For example randomized clinical trials of chemotherapy versus best supportive care have demonstrated a survival advantage and quality of life improvement when patients with advanced-stage lung cancer receive chemotherapy.

Second, chemotherapy can be used as neoadjuvant therapy, given prior to radiation or surgery for locally advanced disease. In this setting, the drugs are used to decrease the tumour mass, reduce the extent of the subsequent surgery or radiation, and to determine disease sensitivity to drugs. Clinical trials have identified a potential role for neoadjuvant therapy in the treatment of lung cancer, oesophageal cancer, and locally advanced breast cancer, among other diseases. In the case of osteosarcomas, neoadjuvant therapy can provide important information about tumour sensitivity, thereby permitting a more tailored approach to further management. Finally, the drugs can be used as adjuvant therapy, administered after the completion of local definitive surgery and/or radiation therapy in order to decrease the risk of recurrence. For instance adjuvant chemotherapy reduces the risk of tumour recurrence and improves survival in node-positive colon cancer and in breast cancer following surgical resection. In all of these settings, chemotherapy can be administered in conjunction with radiation therapy to optimize local effects of treatment.

Only in rare circumstances, such as the use of methotrexate as therapy for choriocarcinoma, can therapy with single agents cure advanced-stage cancer. Single agents tend to select for drug resistant cells. Most often, therapy with multiple drugs has been required to effect cure. In combining drugs, it is imperative to employ agents that have independent activity, have toxicities that do not overlap significantly, and that can be used at an optimal dose and schedule. Chemotherapy schedules are designed to permit marrow recovery prior to the next dose administration. Typically, peripheral blood counts will reach a nadir at 5 to 10 days post-therapy, with recovery seen by day 21.

Combinations of drugs may circumvent tumour cell resistance, so treatments have been designed in which non-cross-resistant drugs are administered either together or in sequence. One of the earliest combination therapy programmes to cure a cancer was the use of MOPP (mechlorethamine, vincristine, prednisone, and procarbazine) for the treatment of Hodgkin disease. Similarly, acute lymphoblastic leukaemia can now be cured in 80 per cent of children using multidrug chemotherapy administered at frequent dosing intervals to avoid the development of resistant cells. Alternatively, extremely high doses of therapy can be used to overcome tumour resistance, but this may obligate the use of stem cell support to overcome the resulting profound bone marrow toxicity.

Because of the potentially severe toxicities of chemotherapy, physicians should administer regimens that have been reported in the peer-reviewed medical literature. Alternatively, the development of new regimens in the context of well-designed institutional review board approved clinical trials can permit the development of novel investigational treatment programmes. Clinicians should not routinely administer drug combinations based on anecdotal evidence.

Classes of chemotherapy agents

There are several distinct classes of chemotherapy agents (Table 2). Because these drugs can have major side-effects, only physicians knowledgeable in their dosing and side-effects should administer them. In order to reduce variability in exposure to drugs, doses of most chemotherapy agents are administered based on the patient's body surface area, a calculation determined by the patient's height and weight. In addition, doses of chemotherapy need to be adjusted for renal (methotrexate, bleomycin, fludarabine) and hepatic function (anthracyclines, vinca alkaloids, taxanes).

Adequate intravenous access must be secured since many of the drugs are vesicants and extravasation can lead to tissue necrosis. Similarly, patients must be adequately hydrated prior to the administration of cisplatin and cyclophosphamide, to prevent renal toxicity and bladder toxicity, respectively. Careful attention must be given to fluid and electrolyte balance with the administration of many agents. Cisplatin renal toxicity can cause profound hypomagnesaemia.

Antimetabolites exert their cytotoxicity by serving as substrates in pathways vital to cellular function and replication. Many of these agents are incorporated into DNA or RNA or act on enzymes involved in the synthesis of nucleic acids. Methotrexate acts by inhibiting the enzyme dihydrofolate reductase, which maintains intracellular pools of reduced tetrahydrofolates required for the synthesis of purine nucleotides and thymidylate. 5-Fluorouricil is another commonly used antimetabolite. A metabolite of this drug, fluorodeoxyuridine monophospate inhibits thymidylate synthase, an enzyme required for the synthesis of deoxythymidine triphosphate and DNA. In addition, fluorodeoxyuridine triphosphate is incorporated into RNA, interfering with its function, and fluorodeoxyuridine triphosphate is incorporated into DNA, leading to strand breakage.

A third important antimetabolite is cytarabine (ara-C) which is converted to cytarabine triphosphate (ara-CTP) in the cell. Cytarabine triphosphate is incorporated into DNA and serves as a chain terminator. A related deoxycytidine analogue, gemcitabine, has the additional actions of inhibiting the conversion of ribonucleotides to

deoxyribonucleotides, which are DNA precursors. Prolonged exposure of tumour cells to some of the antimetabolites, such as 5-fluorouracil and cytosine arabinoside, through continuous intravenous infusion may be more effective than bolus injections alone.

Purine analogues also have important roles as antimetabolites; 6-mercaptopurinre (6-MP) and 6-thioguanine (6-TG) are converted in the cell to monophosphates which inhibit the first step of purine synthesis. Moreover, the triphosphate nucleotides of 6-mercaptopurinre and 6-thioguanine are incorporated into DNA resulting in an increase in strand breaks. Another purine analogue is fludarabine phosphate, which serves as an adenosine analogue. Fludarabine is converted to 2-fluoro-ara-A in plasma and subsequently is phosphorylated intracellularly. The resulting triphosphate inhibits DNA polymerase and ribonucleotide reductase, interfering with DNA and RNA synthesis.

Alkylating agents exert their cytotoxicity by binding to DNA and forming DNA adducts which alter DNA structure and function enough to disrupt DNA replication and transcription. They act throughout the cell cycle, but have their greatest activity on rapidly proliferating cells. These agents, including cyclophosphamide, nitrogen mustard, melphalan, busulfan, and chlorambucil were among the first chemotherapy drugs and remain important agents in cancer therapy, with particular activity in haematological malignancies and breast cancer. In a similar manner, the platinum derivatives bind to and cross-link DNA, leading to DNA breaks and apoptosis.

The anthracyclines intercalate into DNA and disrupt DNA synthesis. The antitumour activity of doxorubicin and daunorubicin, the two most commonly used agents in this drug class, results in part from triggering of topoisomerase II dependent DNA breaks. Etoposide also inhibits topoisomerase II. In a similar way, other topoisomerase inhibitors interfere with topoisomerase I, which is critical in the repair of normal DNA; these agents include irinotecan and topotecan. Vinca alkaloids interfere with microtubule formation and disrupt cell division. In contrast, the taxanes stabilize microtubule assembly, also inhibiting mitosis.

Along with the traditional cytotoxic agents, hormone-directed therapy can be critical in the regulation of tumours. The growth of many normal tissues and tumours is influenced by hormone exposure. Many breast cancers express receptors for oestrogen and progesterone and most prostate cancers have androgen receptors. Depriving these tumours of the hormonal stimulus can exert both cytocidal and cytostatic effects on the cell. Thus, more than 50 per cent of breast cancers expressing the oestrogen receptor will respond to treatment with tamoxifen, an antioestrogen. Similarly, the use of luteinizing hormone releasing hormone (LHRH) agonists (which reduce testosterone synthesis) or antiandrogens can have dramatic effects on prostate cancer growth.

Chemotherapy resistance

Unfortunately, there are several mechanisms by which tumours may become resistant to the effects of cytotoxic chemotherapy (Table 3). Decreased accumulation of drug in the cell through alteration in transport mechanisms permits resistance to methotrexate. Alternatively, the intracellular target for methotrexate may be altered, as in the case of amplification of dihydrofolate reductase. Similarly, with 5-flurouracil, the target enzyme thymidylate synthase may be amplified. Altered drug metabolism, as occurs with ring reduction of 5-fluorouracil, can contribute to drug resistance. In tumours resistant to alkylating agent, the alkylating agents may be inactivated through reactions with thiol-containing compounds or through enhanced DNA repair.

A particularly important mechanism of resistance, conferred by a P-glycoprotein, leads to resistance to multiple drugs. Multidrug resistance results from enhanced drug efflux from the cell secondary to P-glycoprotein over-expression. In this setting, the cancer cells becomes resistant to anthracyclines, vinca alkaloids, taxanes, etoposide, and other drugs simultaneously. Several agents currently in clinical trials are designed to inhibit the multidrug resistance pump. A different type of multidrug resistance can occur with the topoisomerase inhibitors where quantitative and qualitative changes in topoisomerase II activity have been associated with decreased tumour sensitivity.

Side-effects of chemotherapy

The commonly used chemotherapy agents have several side-effects in common (Table 4). Myelosuppression occurs with virtually all agents, although the timing of its onset and its duration differs with different groups of drugs. Cyclophosphamide causes an acute, short-onset depression in counts, affecting the white blood count more than platelets. By contrast, the nitrosoureas lead to delayed-onset reductions in both neutrophils and platelet with nadir counts typically reached 4 to 6 weeks after therapy.

Nausea and vomiting remain a troubling side-effect of chemotherapy, though the present use of serotonin uptake inhibitors has diminished the incidence of vomiting with the most emetogenic agents, including cisplatin. Alopecia is a major cause of concern to patients and occurs uniformly with some agents such as doxorubicin, but rarely with agents such as methotrexate and fludarabine. Profound neuropathy frequently accompanies the use of vinca alkaloids, and occurs less frequently with cisplatin and paclitaxel.

Moreover, many of the agents have unique side-effects that are of concern to the practising internist. Doxorubicin can cause cardiac toxicity, including an acute syndrome characterized by arrhythmias and congestive heart failure. In addition, doxorubicin can cause a cumulative, dose-dependent decline in left ventricular ejection fraction, with a higher incidence of myocardial dysfunction seen in patients receiving a cumulative dose of greater than 500 mg/m ^{2}. Bleomycin causes lung toxicity, including pneumonitis, which can progress to interstitial fibrosis. The carbon monoxide diffusing capacity of the lung diminishes with increasing cumulative bleomycin doses. Methotrexate in high doses can cause acute renal failure due to drug precipitation in the renal tubules. The administration of paclitaxel can cause anaphylaxis in response to cremaphor, the vehicle in which it is delivered. Hence, premedication with dexamethasone and antihistamines is required to reduce the risk of adverse reactions. Cytarabine administered in high single doses (3 gm/m ^{2} or greater) can cause irreversible cerebellar dysfunction, so a neurological exam should be performed daily on patients receiving therapy so that it can be discontinued at the earliest sign of such toxicity.

A major, delayed side-effect of cancer chemotherapy is the development of secondary leukaemias due to therapy. These are most commonly seen in patients receiving therapy with multiple alkylating agents, as was the case for the treatment of Hodgkin disease with MOPP chemotherapy. Five to six per cent of patients receiving this therapy developed leukaemia as a consequence of therapy. Newer regimens for Hodgkin disease treatment avoid this devastating complication of therapy. More recently, topoisomerase II therapy (etoposide, anthracyclines) in high total doses has been associated with a risk of secondary leukaemias. High-dose therapy followed by autologous stem cell infusion has a risk of secondary leukaemias that exceeds 10 per cent. As survival rates are improved with combination chemotherapy regimens used in the treatment of diseases such as leukaemia and lymphoma, the long-term complications of cancer chemotherapy become more evident.

Finally, infertility occurs with many chemotherapy regimens, especially when patients are treated with alkylating agents or high-dose therapy. While reduced sperm counts may occur only transiently in males, most alkylating-induced azoospermia is irreversible, and a discussion should be initiated regarding the possibility of banking sperm for any patient at risk. For premenopausal woman, the risk of infertility and early menopause increases with age. These fertility issues must be addressed with patients prior to the administration of chemotherapy.

Biological therapy

Cancer chemotherapy not only refers to the traditional, cytotoxic agents but also encompasses novel biological therapies including monoclonal antibody-based treatments and cytokine therapies. Currently, two unconjugated monoclonal antibodies are available for commercial use in the United States. Rituximab binds to the CD20 antigen that is expressed on the surface of both normal and malignant B lymphocytes. Nearly 50 per cent of patients with low-grade B-cell lymphoma respond to this targeted therapy, which lacks the typical myelosuppressive side-effects of chemotherapy as well as the typical alopecia and emesis. The most common side-effects with antibodies are infusion-related fevers, chills, and hypotension. Another biologically active antibody is herceptin, which binds to the Her-2 receptor that is over-expressed in 15 to 20 per cent of breast cancer cases. When given in conjunction with paclitaxel, the combination appears to prolong survival for patients with metastatic breast cancer.

The two most extensively studied cytokine therapies are interferon and interleukin-2 (IL-2). The interferons are a class of proteins produced by the body in response to viral infections. These agents have relatively disappointing antitumour activity, although they can induce major responses in patients with low-grade lymphoma and hairy cell leukaemia. Toxicities include fevers, chills, liver function test abnormalities, and cytopenias. IL-2 is a cytokine produced by activated T cells that plays a role in triggering the immune system. IL-2 has activity in renal cell carcinoma and melanoma, with occasional patients achieving long-duration remissions. However, its toxicities include fevers, renal dysfunction, and capillary leak syndrome.

Recent advances in molecular biology have led to the development of additional therapies which are currently in clinical trials. Antisense oligonucleotides of 15 to 20 bases in length can be used to target specific messenger RNAs. A bcl-2 antisense compound has shown biological activity against low grade non-Hodgkin's

lymphoma (which over-express the bcl-2 protein). Another exciting class of compounds target new blood vessel formation, which may be critical for the implantation and growth of tumour cells. These antiangiogenic compounds also are in clinical trials.

Radiation therapy

Understanding the cytotoxicity of radiation therapy requires knowledge of radiation physics and tumour biology. Over the last two decades, the ability of CT scans and MRI scans to carefully localize the tumour within the patient, along with technical improvements in treatment machines, have radically improved the accuracy of radiation therapy.

Electromagnetic radiation, used most commonly in patient care, consists of roentgen (X-rays, photons) and gamma radiation. In general, gamma rays are produced by the degradation of nuclear isotopes while electrical machines produce photons. Alternatively, electron beams can be used for the treatment of superficial tumours (such as cutaneous lymphoma) because of the sharp reduction in dose that occurs beyond a certain tissue depth. Proton beam therapy can be used to delivery high treatment doses to a highly localized lesion because of its very sharp margins of dose deposition, and has particular relevance in the treatment of spinal tumours and some central nervous system tumours.

The dose of irradiation is defined as the unit of energy absorbed by each kilogram of tissue. This is conventionally expressed in 'Grays' (Gy). As the dose of radiation is increased the percentage of cells that is killed increases.

Radiation used for the treatment of patients generally consists of either external beam irradiation delivered from outside the body or brachytherapy, in which the radiation device is placed within or near the target tumour. Brachytherapy has been used effectively in the treatment of head and neck cancer, cervical cancer, and endometrial cancer. At some centres, intraoperative radiation therapy can be used to deliver a single, large fraction of radiation directly to the tumour bed. In some circumstances, radioisotopes themselves can be used for systemic treatment. For example iodine-131 is taken up by thyroid tissue both locally and at sites of metastatic disease.

The target for radiation-induced cytotoxicity is DNA. Radiation therapy generates free radicals that damage DNA. Because the presence of oxygen is important in the generation of free radicals, hypoxic tissues are less sensitive to the toxic effects of radiation. Thus, well vascularized tissues are most sensitive to radiation therapy. A theoretical advantage of preoperative radiation is the chance to treat a tumour while its vasculature remains intact. However, resection of a larger, poorly vascularized tumour may improve the chances of curing a locally advanced head and neck cancer with irradiation.

The delivery of radiation therapy requires careful planning. CT scans or MRI scans are used to identify the target tissue and surrounding normal tissues accurately. Next, treatment technique and volume are tested on a radiation simulator. The simulator duplicates the treatment plan, but uses only superficial radiation for imaging and accurately assessing the location of the treatment beam. To make certain that treatments are delivered to the same tumour volume each day, tattoos are placed on the patient. Blocks are made to exclude treatment from normal tissue such as the heart and lungs. Often the maximum tolerated dose of radiation is administered to the total treatment volume with a boost administered to the treatment bed.

Radiation doses are usually delivered in a number of daily fractions with the total fractionated dose dependent on tumour sensitivity and normal tissue tolerance. Seminoma is an exquisitely radiation sensitive tumour and requires a lower therapy dose (30 Gy) than solid tumours such as lung cancer (60 Gy). Some tumours, such as melanoma, are relatively radioresistant.

Within 6 h after radiation, cells begin to recover from the effects of therapy. Thus, fractions placed too close together can offer increased toxicity to normal tissues, but those too far apart can permit repair of sublethal damage. While conventional therapy usually provides a daily radiation fraction of 1.8 to 3 Gy over 15 to 35 treatments, to total doses of 40 to 65 Gy, alternative schemes have been investigated. In hyperfractionated therapy, a smaller fraction size is used, and more fractions are administered. This permits a higher total radiation dose to be administered. For palliation of painful bone metastases, radiation therapy can be given in either a single large dose or four to five moderate doses to minimize the patient's travel to the treatment centre. More recently, radioisotopes such as iodine-131 have been conjugated to antibodies, with radiation delivered directly to the surface of the targeted cell. In this scenario, low-dose continuous radiation therapy is effectively delivered at the tumour site.

Complications of radiation therapy

Adverse effects of radiation therapy can be considered with respect to both acute and delayed toxicities. Tissues that normally proliferate rapidly, such as skin, mucosal linings, and bone marrow, are most susceptible to radiation cytotoxicity. Thus, erythema and desquamation are important acute, local effects of therapy. For patients receiving radiation therapy for gastrointestinal tumours, diarrhoea, nausea, and vomiting are common. If a significant radiation dose is delivered to the bone marrow, patients may develop cytopenias. In the case of whole body irradiation, the lymphocyte count falls and significant immune suppression occurs. On occasion, these acute side-effects are severe enough to require delays in treatment to allow the normal tissues to repair themselves. When patients receive irradiation to the lung, a resultant radiation pneumonitis may occur characterized by fevers, cough, dyspnoea, and pulmonary infiltrates. Occasionally, this may require treatment with corticosteroids.

Long-term sequelae are tissue specific and occur most commonly if normal tissue tolerance is exceeded. Thus, careful dosimetry and radiation planning must be done to verify that tissues do not receive treatment beyond their maximum tolerated dose. For example radiation doses to the spinal cord in excess of 45 Gy can cause myelitis, doses to the small bowel in excess of 45 Gy can cause strictures, and doses to the kidney above 20 Gy can cause renal dysfunction. The liver tolerates radiation therapy poorly. Accelerated coronary artery disease has been seen in patients with Hodgkin disease who received radiation to the heart in an effort to encompass a mediastinal mass.

Perhaps the most severe side-effect of radiation therapy is the development of secondary tumours. Ordinarily, this is not an issue for patients with metastatic cancer receiving radiation therapy for palliation of disease related symptoms since their survival will be short. However, in patients with Hodgkin disease, who can be cured with radiation therapy, the development of second solid tumours in the radiation field, including sarcomas, lung cancers, and oesophageal cancer, can limit survival.

Role of radiation therapy in cancer treatment

In the clinical management of patients, radiation therapy is used as primary therapy for tumour treatment, as an adjuvant or neoadjuvant therapy either alone or in conjunction with chemotherapy (which often acts as a radiation sensitizer), and as palliative therapy for advanced stage treatment (Table 5).

Radiation therapy has a role in the management of several acute complications of cancer. Radiation can be valuable in the treatment of bone metastases, both to decrease painful lesions and to diminish the risk of pathological fractures. Radiation therapy can be delivered as an emergency procedure in patients with spinal cord compression to reduce the risk of permanent neurological toxicity. Likewise, radiation has an important role in the management of brain metastases, either as primary therapy for patients with multiple lesions or as a prophylactic therapy for patients undergoing excision of a solitary brain metastasis. In lung cancer, radiation can be used to palliate obstructive symptoms. In bleeding tumours, radiation therapy can often assist in local control of haemorrhage.

In the management of several tumours, radiation therapy can serve as the definitive treatment. In early-stage Hodgkin's disease, patients can be cured with either mantle radiation therapy alone or with mantle and para-aortic radiation. Similarly, 35 to 50-Gy doses of radiation therapy can induce long-term remissions in 50 to 60 per cent of patients with stage I/II low-grade non-Hodgkin's lymphoma. Seminoma is an exquisitely radiation sensitive tumour and early stage disease can be cured in a high percentage of patients with radiation therapy alone. Radiation therapy can provide equivalent long-term survival to surgery in patients with prostate cancer and laryngeal cancer, with potentially reduced morbidity as compared with surgery. Finally, in early-stage breast cancer, lumpectomy and radiation therapy provides an equivalent survival outcome to a modified radical mastectomy.

In other diseases, such a squamous cell carcinoma of the anus, combined modality therapy using radiation therapy in conjunction with 5-FU and mitomycin C chemotherapy yields a high cure rate without surgery. Similarly, in patients with limited stage small cell lung cancer, combined modality therapy using cisplatin-based chemotherapy and radiation therapy reduces local tumour recurrence and improves survival. Likewise, in cervical cancer, a combination of cisplatin and radiation post-resection reduces tumour recurrence.

Radiation therapy also has an important role in adjuvant therapy. In the adjuvant treatment of rectal cancer, randomized clinical trials have demonstrated that radiation

therapy administered in conjunction with 5-FU chemotherapy can reduce local recurrence and systemic recurrence and can improve both disease free and overall survival. In node-positive gastric cancer, a combination of 5-FU based chemotherapy administered as adjuvant therapy can reduce the risk of recurrence and improve survival. Recent studies have demonstrated that the administration of prophylactic cranial irradiation to patients with small cell lung cancer who achieve a complete remission can reduce the risk of central nervous system spread of disease and improve survival. In the neoadjuvant setting, radiation in combination with cisplatin-based chemotherapy has been shown to improve survival in patients with stage IIIA lung cancer and oesophageal cancer.

As noted earlier, the gamma radiation from radioactive isotopes also can be used to treat tumours. Well-differentiated thyroid cancer metastases take up iodine-131 which then can offer tumour-specific toxicity. Antibodies have been conjugated to iodine-131 and yttrium-90 to deliver targeted radiation therapy to the surface of lymphoma cells.

Conclusions

Advances in radiation therapy and chemotherapy have revolutionized the care of cancer patients. Significant improvements in supportive care and the development of new, active anticancer agents have improved the prospects for long-term survival even for patients with metastatic disease. The internist has a pivotal role in co-ordinating care for such patients and should be aware of these advances and new option for patients.

Further reading

DeVita VT Jr (1997). Principles of cancer management: chemotherapy. In: DeVita VT Jr, Hellman S, Rosenberg SA, eds. *Cancer: principles and practice of oncology*, pp. 333–73. Lipincott-Raven, Philadelphia. [Describes general aspects of chemotherapy and concepts of drug resistance, cell cycle biology, and dose intensity.]

Goldie JH (1987). Scientific basis for adjuvant and primary (neoadjuvant) chemotherapy. *Seminars in Oncology* **14**, 1–7. [Discusses principles of adjuvant and neoadjuvant therapy as well as theoretical benefits and concerns.]

Greenlee RT, *et al*. (2000). Cancer statistics, 2000. *CA: a Cancer Journal for Clinicians* **50**, 7–33. [Summarizes data on incidence and survival for all types of cancer in the United States.]

Hellman S (1997). Principles of cancer management: radiation therapy. In: DeVita VT Jr, Hellman S, Rosenberg SA, eds. *Cancer: principles and practice of oncology*, pp. 307–32. Lipincott-Raven, Philadelphia. [Overview of biological and physical properties of radiation therapy.]

Leibel SA, Phillips TL, eds (1998). *Textbook of radiation oncology.* WB Saunders, Philadelphia. [Comprehensive textbook describing current techniques in radiation oncology and their clinical application.]

Marino P, *et al*. (1994). Chemotherapy vs supportive care in non-small-cell lung cancer. *Chest* **106**, 861–5. [A meta-analysis demonstrating improved survival and quality of life for patients with metastatic lung cancer receiving chemotherapy.]

Mauch PM, *et al*. (1996). Second malignancies after treatment for laparotomy staged IA-IIIB Hodgkin's disease: long-term analysis of risk factors and outcome. *Blood* **87**, 3625–32. [Describes risk of secondary haematological malignancies and solid tumours in 794 patients receiving either radiation therapy or combined modality therapy.]

Multani PS, Grossbard ML (1998). Monoclonal antibody-based therapies of hematologic malignancies. *Journal of Clinical Oncology* **16**, 3691–710. [Provides an overview of this developing field.]

Perez CA and Brady LW, eds (1997). *Principles and practice of radiation oncology*, 3rd edn. Lippincott-Raven, [New York. Comprehensive textbook of radiation oncology.]

Pinedo HM, Longo DL, Chabner BA, eds (1999). *Cancer chemotherapy and biological response modifiers: annual 18*. Elsevier, New York. [Up-to-date text on mechanisms of action and resistance of chemotherapy agents and their major indications for use.]

Walsh TN *et al*. (1996). A comparison of multimodal therapy and surgery for esophageal adenocarcinoma. *New England Journal of Medicine* **335**, 462–7. [Improved survival for patients receiving preoperative chemotherapy and radiation versus surgery alone in a phase III trial.]

7.1 The clinical approach to the patient with suspected infection

David Rubenstein

Presentation of illnesses and the history

Most systemic infections produce fever but not all fevers are, at cause, infectious. Patients with systemic infections range symptomatically from apparent full fitness to near moribund and initial appearance will indicate the speed of clinical response required. Appearances can be deceptive and many serious, potentially fatal infections do not present acutely. This gives time for more careful assessment although not for too much delay.

The standard undergraduate history produces a reliable range of questions covering all systems, which usually indicates the system or systems at fault (except perhaps for over- and underactivity of the thyroid) and should always be completed thoroughly. Where infection is common or possible, the addition of a detailed travel history can be invaluable, particularly as some infections occur in local geographical pockets, for example schistosomal infection from swimming in Lake Malawi, histoplasmosis along the Ohio river valley, and coccidioidomycosis in the San Joaquin Valley, California. Also air passengers may contract illness either during travel, from air droplet transmission, and even after a very brief stop-over. A detailed drug history is also most important. Not only do drugs produce fever but prophylactic agents may themselves be the cause of illness. As general physicians, in parallel with the rest of humanity, we have blind areas and ours are ear, nose, and throat, gynaecology, and sexually transmitted diseases, all of special importance when considering infection sites, if not obvious on initial history taking. The past and social history rarely give a pointer to the cause of infection, but previous surgery has commonly created a site of infection even weeks or months after operation, and social and sexual contacts and their illness may offer an important clue—HIV infection now occurs worldwide.

Examination

A complete and well learned and rehearsed examination system is an essential component of a clinician's skill. With time this becomes almost second nature and will prevent serious omission. ('The more I practise the better I get'—Arnold Palmer, but also variously attributed.)

The initial history almost invariably focuses the examiner on the likely site of infection, but occasionally a clinical finding will add to the clinical picture. If so, this will guide further management, but if not, it is essential to assess lymph nodes carefully at all common sites, and listen to the heart for murmurs. Careful examination of the skin may give important clues. The ear, nose, and throat region and the pelvis may hide infection as can the abdomen and re-examination should be performed daily to include these and the 'physician's blind areas' of the midline—the epigastrium, umbilical, and suprapubic regions. I have on first examination missed enlarged lymph nodes, heart murmurs, suprapubic masses, splenic enlargement, scrotal ulcers, and circinate balanitis—and probably many other signs, but the chance of missing a key sign is greatly reduced by repeated examination and by different observers.

Key investigations

> 'Ned, why do you keep robbing the banks?'
>
> 'Well, Judge, that's where all the money is'
>
> (Ned Kelly, Australian bush ranger, nineteenth century and variously attributed.)

The history and examination findings usually point to the system involved and often the diagnosis and investigation is aimed at confirming clinical suspicion. This is rarely a problem as most patients present with both a fever and some other marker, such as rash, vomiting, diarrhoea, cough, and sputum. It becomes more difficult when there are no key features other than fever. In all cases it is worth rechecking the drug and travel history, particularly if malaria is suspected, and all patients require a full blood count, blood cultures, a routine urine check, and a chest radiograph. At the same time and for future reference, studies may reveal early evidence of dehydration, a common feature of acute infectious diseases.

The white blood count is of particular value in septicaemia/bacteraemia and also in returning travellers, as the neutrophil count is not raised in malaria. The presence of a neutrophilia does not exclude malaria but suggests bacterial or amoebic infection. Some returning travellers have both. Blood cultures will allow confirmation of bacterial blood infection and guide antibiotic therapy, particularly alteration in antibiotic cover if the initial choice is ineffective. Routine urine testing and culture is often revealing even in the absence of urinary tract symptoms. As it is now simple, it is often omitted—even on renal units. In fairness, blood may not be sent for culture prior to treatment even on specialist infectious disease wards. A chest radiograph in the absence of respiratory symptoms may reveal tuberculosis, an abscess, or hilar lymph node enlargement, this being the first indication of tuberculosis, lymphoma, or sarcoid. If the diagnosis remains uncertain, more detailed studies including serology and ultrasound CT or MR scanning should be guided by clinical suspicion and probability, followed if indicated, by guided biopsy. Persistent fever has many causes, some of which are non-infectious and investigation of these should continue in parallel if clinical suspicion is sufficient (see Chapter 7.2).

Management

This is guided by the clinical picture and investigation findings. It becomes more difficult if the patient is too ill to wait for results or if these initially fail to reveal the diagnosis. This is commonly encountered in 'ill' patients with suspected septicaemia. This is a medical emergency and treatment should not be delayed until results are through. Antibiotics chosen to cover the likely infective organisms should be given intravenously as soon as key investigations, including blood and urine for culture, have been taken. Results and the patient's progress may change the plan, but delay may have serious consequences.

It is essential not to forget the important general supportive measures of rehydration and short-term anticoagulation for bed-bound and dehydrated patients. Good communication and reassurance are also very important, but nothing like as important as getting the right answer and starting the right treatment.

The future

New infectious diseases

These continue to appear and Lassa fever, legionnaire's disease, Lyme disease, and HIV infection were little recognized or unknown 30 years ago. Remaining alert to new diseases is critical but almost impossible for any single clinician, who must remain up to date with the literature and particularly the regular and superb publications from the Centers for Disease Control, Atlanta, the Public Health Laboratory Service in London, the World Health Organization, and the Pro Med web site.

Prevention

This is the key to future success and currently at a very exciting phase. Smallpox is we hope eradicated, although some of us remain slightly nervous in the era of

biological warfare. Current research, if fruitful, may produce successful vaccines against a wide range of infections—malaria, leishmaniasis, and even HIV are all good candidates.

Advanced diagnostic technique

It is impossible for our generation to appreciate the 'miracle' of chest radiography. Likewise, those of us who practised prior to scanning, particularly CT and now MR scanning, still remain amazed at the detail and accuracy of current radiology and the great improvement in diagnostic accuracy which can be quickly, safely, and almost atraumatically achieved. No doubt future advances, and greater availability with lowered cost, will make these techniques much more widely available. The polymerase chain reaction is now well established, but not universally available. Of perhaps even greater diagnostic help may be results of research into the use of 'antigen strips'. If the example of urine dipsticks is a guide, these may become a universally available, accurate, inexpensive, and possible bedside aid.

Final thoughts

Given sufficient knowledge and experience, your initial diagnosis is probably right and investigation will be aimed at confirmation. If the answer remains obscure after careful reassessment of history, examination, and clinical notes, seek a second opinion. In the hour of need, microbiologists and dermatologists make excellent friends.

Always consult with ease but selectively, remembering that colleagues who refer everyone or no one might have something to hide. Working alone makes clinical medicine a scary occupation and the best clinicians hunt in packs. It makes for more certain diagnosis, finer tuned investigation, better overall therapy, and peace of mind.

Further reading

Web sites

Pro Med

Mobile texts

Chin J, ed. (2000). *Control of communicable diseases manual*, 17th edn. American Public Health Association, Washington DC.

Wilkes D, Farrington M (1975). *The infectious diseases manual.* Blackwell Scientific Publications, Oxford. [2nd edn—2002.]

Essential reviews

Morbidity and Mortality Weekly Reports. United States Department of Health and Human Diseases, Washington DC.

Communicable Disease Report Weekly. PHLS (Communicable Disease Centre), London NW9 5EQ.

Ethics

Cronin AJ (1996). *The citadel*. Victor Gollancz, London.

7.2Fever of unknown origin

David T. Durack

Febrile episodes are common, often transient, and often due to an obvious cause. In many cases the cause is obvious, such as an upper respiratory infection in a child. In a few cases, fever is persistent and the cause is not easily diagnosed. Such episodes are termed 'fever of unknown origin' (**FUO**) or 'pyrexia of unknown origin'.

Definitions and terminology

Normal body temperature is 37.0 °C or 98.6 °F. The normal range is quite wide, being affected by site of measurement, diurnal variation, heavy exercise, hormonal and menstrual status, individual variation, and environmental factors. A patient's body temperature is often estimated by measurements taken in the mouth for reasons of convenience, but oral temperatures can be affected by mouth-breathing, by the respiratory rate, and by recent drinking of hot or cold liquids. Many modern thermometry instruments take readings from the ear canal, so that the temperature is not affected by these factors. The core body temperature is more closely reflected by rectal measurements, which are usually 0.3 to 0.6 °C higher than oral measurements. Heavy exercise can temporarily raise the core temperature of healthy people by 2 °C or more. Another factor is normal circadian variation, which cycles through a range of approximately 0.5 °C (0.9 °F) daily, with the lowest temperatures occurring between 0400 and 0600 h and the highest between 1600 and 2000 h. The normal circadian rhythm varies between individuals and is likely to be affected by jet travel between time zones, by work and sleep patterns, and by illnesses. The menstrual cycle alters the baseline temperature of normal women by 0.3 to 0.5 °C, with a small spike at ovulation and higher temperatures from about the 15th to the 25th days of a 28-day cycle. In addition to these factors, there is considerable variation in normal temperature patterns between individuals. Some normal young people, especially women, persistently exhibit slightly 'high' temperatures that are of no pathological significance. This common normal variant, which does not require investigation, may be termed 'habitual hyperthermia'.

Fever and hypothermia may be defined, respectively, as core body temperatures above or below the normal range, allowing for all the factors listed above. For practical clinical purposes, oral or rectal temperatures falling outside the range of 35.5 to 38.0 °C (95.9–100.4 °F) can be regarded as abnormal. Specific circumstances should be considered; for example, an oral temperature of as low as 37.5 °C taken at 0600 h in an elderly patient could represent a clinically significant fever.

FUO has many possible causes. To help classify these, four distinct types of prolonged fever have been defined: classical FUO, nosocomial FUO, neutropenic FUO, and human immunodeficiency virus (**HIV**)-associated FUO.

Symptoms and signs of FUO

The symptoms and signs of FUO are highly variable. Some patients have mild feverish symptoms, while others may be incapacitated by debilitating chills, rigors, and sweats. The clinical findings may be limited to manifestations of the fever itself, or may also reflect the underlying disease. The physician should evaluate every symptom or sign, especially new ones, as potential clues to the primary diagnosis.

Certain diseases can produce characteristic patterns of fever, notably malaria, brucellosis, typhoid fever, and some lymphomas, but in practice the shape of the fever curve is seldom of major value in the diagnosis of FUO. Individual host reactions to disease and the common use of antipyretic analgesic drugs confuse the picture. There is a common misconception that drug-induced fevers are usually low-grade ones, with relatively little variation from peak to trough and a relatively low pulse rate, but in fact the clinical characteristics of drug-induced fevers are highly variable.

Classical FUO

Most of the many causes of classical FUO can be classified into five categories: infections, malignancies, connective tissue diseases, miscellaneous conditions including factitious fever and habitual hyperthermia, and undiagnosed cases. Within the first three categories, certain diagnoses predominate (Table 1 and Table 2). The leading infectious aetiologies for classical FUO are intra-abdominal infections, complicated urinary tract infections, tuberculosis, and infective endocarditis. The leading malignancies are lymphomas, leukaemias, and some solid tumours, including adenocarcinomas and hypernephromas. Vasculitides, including the temporal arteritis–polymyalgia syndromes, Still's disease, systemic lupus erythematosus, and rheumatic fever, are important among the connective tissue diseases. Among the miscellaneous conditions that can cause FUO, alcoholic hepatitis and granulomatous conditions such as sarcoidosis or granulomatous hepatitis are important. Self-induced or factitious fever is surprisingly common. Some of the many other miscellaneous, uncommon, or rare diseases that can cause FUO are listed in Table 2. In all published series, a sizeable subgroup of patients with FUO remains undiagnosed.

FUO in children

The proportion of cases of FUO due to infections is higher in children, and the proportion due to malignancy is correspondingly lower. Viral syndromes and urinary tract infections are particularly common infections in children. Still's disease and rheumatic fever are more likely to cause FUO in children than in adults, and children are less likely to have factitious fever. The overall mortality of FUO in children is lower than in adults.

FUO in the elderly

In patients over 65 years of age, intra-abdominal abscesses including hepatic abscesses, malignancies, and vasculitides cause a higher proportion of cases of FUO. The proportion of FUOs that remain undiagnosed in the elderly is lower, being only about half that in children and younger adults. The higher rate of underlying malignancies in any series of elderly patients with FUO means that the long-term prognosis is less favourable than in a younger group. The temporal arteritis–polymyalgia rheumatica syndromes are particularly important because they are common in the elderly, and their many non-specific symptoms may be missed or misdiagnosed. This diagnosis is easily suspected if the erythrocyte sedimentation rate (**ESR**) is over 100 mm/h, but is easily overlooked if an ESR is not obtained. Another hint is a high platelet count. Other connective tissue diseases are less common than in younger patients. Bacterial prostatitis and related urinary tract infections are more common in elderly men due to prostatic hypertrophy. In developed countries, infective endocarditis has become more common in older patients. Occult pulmonary emboli always should be considered in the differential diagnosis. Factitious fever is rare in the elderly.

Nosocomial FUO

Fever that develops after a patient has been admitted to hospital, and which remains undiagnosed, is termed 'nosocomial FUO'. These patients are usually being treated for one or more major pre-existing conditions, and have multiple possible reasons for developing fever. Several of these factors may be contributing simultaneously to the development of fever. After common bacterial infections such as pneumonia, urinary tract infection, and bacteraemia have been excluded, many other conditions remain in the differential diagnosis: for example, local or disseminated candidiasis, *Clostridium difficile* diarrhoea or colitis, cytomegalovirus infection,

hepatitis, sinusitis (especially if the patient is intubated), intravascular catheter-related local or bloodstream infections, and infective endocarditis. The possibility that a non-infectious inflammatory condition such as acalculous cholecystitis, gout, or pseudogout has flared during hospital admission for another condition should be considered. Occult pulmonary emboli are an important cause of nosocomial FUO. Drug fever is especially common in this patient group.

Neutropenic FUO

The number of patients with neutropenia caused by cytotoxic chemotherapy for various diseases is increasing, although the duration of neutropenia is now often curtailed by the timely administration of colony-stimulating factors. Fevers in neutropenic patients are very different from the classical FUO defined above. The leading causes of neutropenic FUO are bacteraemias, pneumonias, and skin/soft tissue infections. Urinary tract infections are less common than in nosocomial FUOs (above). Focal bacterial infections of intravascular lines and puncture wounds, skin folds, and the perianal area all are common, and often associated with bacteraemia. In the early stages of neutropenia, fevers are usually caused by bacteria, but if neutropenia persists, fungal, viral, and other conditions become relatively more common. However, this well-known sequence loses diagnostic value when the patient has received multiple cycles of chemotherapy and antimicrobial drugs.

The duration of neutropenic FUOs tends to be much shorter than that of classical FUOs. The onset of fever often occurs within days of the onset of neutropenia; immediate empirical antimicrobial treatment is usually given, and improvement is frequently rapid. The aetiology of these FUOs often remains unconfirmed. In the majority of neutropenic FUOs the fever is probably due to infection, but the aetiological organism(s) will be identified in only 40 to 60 per cent of cases. Recurrent episodes are likely to occur for as long as the patient remains neutropenic.

HIV-associated FUO

A self-limited episode of fever often occurs during primary HIV infection. After a long asymptomatic interval, fevers and FUOs become extremely common during the later stages of HIV infection. This justifies the introduction of the term 'HIV-associated FUO' in the definitions listed above. The single most common cause of FUO in this setting is mycobacterial infection (tuberculosis in developing countries, *Mycobacterium avium* complex (**MAC**) in the developed world). MAC infection eventually affects up to 40 per cent of patients with acquired immunodeficiency syndrome (**AIDS**) in developed countries. Many other diagnoses must be considered, especially *Pneumocystis carinii* infection, cytomegalovirus infection, disseminated cryptococcosis, toxoplasmosis of the central nervous system, lymphomas, and nocardiosis. In the appropriate geographical regions, disseminated leishmaniasis, histoplasmosis, coccidiodomycosis, and *Penicillium marneffei* infection must be considered. Recently, *Bartonella* species, which cause bacillary angiomatosis and peliosis hepatitis, have also been found to cause febrile bacteraemic syndromes and endocarditis in patients with AIDS.

Investigation of FUO

At the first encounter with the patient, a meticulous history should be taken and a complete physical examination performed. The theme should be attention to detail: for example, careful ophthalmoscopy after dilating of the pupils could reveal Roth spots or retinal tubercles in patients with classical FUO, retinal candidiasis in those with nosocomial FUO, or cytomegalovirus retinitis in cases of HIV-associated FUO. Routine test results (chest radiograph, routine blood count, differential cell count, erythrocyte sedimentation rate, and serum biochemistry) should be scanned for clues. A raised serum uric acid level could signal rapid cell turnover in lymphoma, and a raised alkaline phosphatase level can indicate liver involvement. The peripheral blood smear should be carefully examined for abnormalities such as thrombocytosis, leukaemoid reactions, the presence of nucleated red blood cells, and other clues that the marrow is reacting to a pathological stimulus. The initial findings should be reviewed in relation to the tempo of disease progression before deciding upon the next round of investigations. What major tests have already been performed elsewhere? Repetition of costly radiographs and scans may be unnecessary. Can further testing be safely postponed? Sometimes more will be learned by waiting, or the FUO may resolve spontaneously.

The next level of investigation will usually involve blood cultures, skin testing for delayed hypersensitivity to tuberculosis, and selected serological tests for infections and connective tissue diseases. In older patients, tests for prostate-specific antigen and carcinoembryonic antigen should be obtained. Echocardiography should be performed if any clues are found that increase the pretest probability of infective endocarditis (for example, unexplained heart murmurs or emboli).

Selection of further investigations requires careful consideration of the likely yield, risks, and costs of each. Because many FUOs are associated with intra-abdominal conditions, computed tomography (**CT**) of the abdomen often is valuable. Sinus radiographs and pulmonary CT can reveal the lesions of Wegener's granulomatosis. Radiographs of the bowel with contrast can reveal abnormalities needing further investigation. Gastrointestinal endoscopy with biopsy is often appropriate if symptoms or imaging studies suggest enteric conditions such as inflammatory bowel disease or cancer. Adjunctive imaging with magnetic resonance imaging scan and/or ^{67}gallium- or ^{111}indium-labelled leucocytes can be helpful, but these tests should be used selectively because they are costly and of limited sensitivity; the chance that one of these will reveal a diagnosis is quite low if radiographs and CT scans are negative. Positron-emission tomography using isotopic fluorodeoxyglucose (**FDG-PET** scan) is a new imaging technique which may prove better than older scanning methods. Transoesophageal echocardiography should be performed if the transthoracic echocardiogram is normal or indeterminate but endocarditis still seems likely.

Biopsies of bone marrow, lymph nodes, lung tissue, liver, skin, and temporal arteries or other vessels are essential for the diagnosis of many FUOs. Exploratory laparotomy, previously often performed for the diagnosis of FUO, is now rarely necessary because of improved imaging techniques and directed biopsies. Abnormalities of cytokine levels in blood occur in over two-thirds of patients with FUO, but this finding has not yet proven useful in diagnosis.

For HIV-associated FUO, if the chest radiograph is abnormal or the patient is hypoxic, bronchial washings or biopsy may reveal *Pneumocystis* spp. *Mycobacteria* spp. *Cryptococcus* spp. or cytomegalovirus infection. Direct staining of stool may reveal the presence of *Mycobacteria* spp. If the patient is stable, the results of blood cultures for *Mycobacteria* spp. and unusual bacteria such as *Rhodococcus* or *Bartonella* spp. should be awaited before further invasive tests are done. If the fever remains undiagnosed at this stage, bone marrow and liver biopsies are most likely to be informative.

Approach to treatment

Treatment of the fever itself is indicated if fever distresses the patient, exacerbates heart failure, or is severe enough to cause catabolism and wasting. Otherwise, the temperature curve can be observed in the absence of treatment, often yielding useful new information while investigations continue. If the fever must be treated, aspirin, paracetamol, or a non-steroidal anti-inflammatory drug in standard doses will usually suffice. A regular dosage schedule rather than occasional or 'as required' dosing is recommended.

Treatment of classical FUO

If an aetiological diagnosis cannot be made at first, it is usually best to withhold treatment while observing the patient's progress at frequent intervals. Ideally, a diagnosis will eventually be made, so allowing specific therapy. If an undiagnosed patient is too ill to permit prolonged observation, empirical treatment for FUO may be considered. The most common choice for an empirical therapeutic trial is a corticosteroid. The recommended dose for an adult is prednisone 30 mg orally twice daily initially, or the equivalent dose of another corticosteroid. The possibility that the fever may be eliminated by empirical corticosteroid therapy while the primary disease is unaffected (or even exacerbated) should be kept in mind. The next most common choice for empirical therapy is a broad-spectrum antibiotic such as oral amoxicillin or a fluoroquinolone, or a parenteral regimen such as ampicillin plus gentamicin. In patients who are desperately ill, a combination of parenteral corticosteroid plus antibiotics may be administered. Less commonly, empirical therapy for possible tuberculosis may be tried.

Treatment of neutropenic FUO

After performing a focused physical examination, obtaining a chest roentgenogram, and sending two blood samples and a urine sample for cultures, empirical broad-spectrum antibacterial therapy should be started immediately, before the results of laboratory tests are available. If initial cultures are positive, an antibiotic regimen specific for the aetiological micro-organism can be chosen. If necessary, antifungal or antiviral therapy may be added or substituted, according to the patient's progress and the results of investigations.

Treatment of HIV-associated FUO

HIV-infected patients with an acute onset of fever and hypoxia will usually be treated immediately for possible *Pneumocystis carinii* infection, even if the chest

radiograph is normal. Ideally, a diagnosis should be made as soon as possible through standard investigations. Once the cause of HIV-associated FUO has been diagnosed, specific treatment regimens as described in Chapter 7.10.21 can be prescribed. For patients who remain undiagnosed, various empirical treatments may be tried. Antimycobacterial drugs are commonly included because of the high prevalence of both tuberculous and non-tuberculous mycobacterial infections in HIV-infected subjects.

Prognosis

Classical FUO is a serious condition. Although most of the causes of this type of FUO can be treated, the 1-year mortality is still between 20 and 30 per cent. Obviously, the prognosis varies depending upon the underlying disease and the age of the patient. If FUO persists undiagnosed for more than 6 to 12 months, the likelihood that a specific diagnosis will ever be made decreases, and the prognosis improves greatly, to less than 5 per cent mortality.

The prognosis for nosocomial FUO depends largely on the underlying diagnoses. The short-term prognosis for neutropenic FUO is excellent, with over a 90 per cent response to initial empirical antimicrobial therapy (with appropriate modification as laboratory results return). Again, the long-term prognosis is determined largely by the underlying disease.

Most of the causes of HIV-associated FUO can be treated, but these patients have a relatively poor prognosis, with death likely within 2 years because HIV disease is usually advanced by the time the patient has FUO. The prognosis has improved somewhat with the introduction of highly active antiretroviral therapy. Atypical *Mycobacteria* spp. (which are the commonest cause of HIV-associated FUO) can be suppressed but seldom eliminated, and are likely to develop resistance during therapy.

Further reading

Armstrong WS, Katz JT, Kazanjian PH (1999). Human immunodeficiency virus-associated fever of unknown origin: a study of 70 patients in the United States and review. *Clinical Infectious Diseases* **28**, 341–5.

Blockmans D, *et al* (2001). Clinical value of (18F) fluoro-deoxyglucose positron emission tomography for patients with fever of unknown origin. *Clinical Infectious Diseases* **32**, 191–6.

Durack DT, Street AC (1991). Fever of unknown origin—reexamined and redefined. In: Remington JS and Swartz MN, eds. *Current clinical topics in infectious diseases*, Vol 11, pp 35–51. Blackwell Scientific, Boston.

Cunha BA (1998). Fever of unknown origin. In: Gorbach SL, Bartlett JG, and Blacklow NR, eds. *Infectious diseases*, 2nd edn, pp 1678–89. WB Saunders, Philadelphia.

Hughes WT, *et al* (1997). 1997 guidelines for the use of antimicrobial agents in neutropenic patients with unexplained fever. *Clinical Infectious Diseases* **25**, 551–73.

Kazanjian PH (1992). Fever of unknown origin: review of 86 patients treated in community hospitals. *Clinical Infectious Diseases* **15**, 968–73.

Kjaer A, Lebech AM (2002). Diagnostic value of (111)In-granulocyte scintigraphy in patients with fever of unknown origin. *Journal of Nuclear Medicine* **43**, 140–4.

Knockaert DC, Bobbaers HJ (1996). Long-term follow-up of patients with undiagnosed fever of unknown origin. *Archives of Internal Medicine* **156**, 618–20.

Knockaert DC, Vanneste LJ, Bobbaers HJ (1993). Recurrent or episodic fever of unknown origin. Review of 45 cases and survey of the literature. *Medicine* **72**, 184–96.

Mackowiak P, ed. (1997). *Fever: basic mechanisms, and management*, 2nd edn. Raven Press, New York.

Maschmeyer G (1999). Interventional antimicrobial therapy in febrile neutropenic patients. *Diagnostic Microbiology, and Infectious Diseases* **34**, 205–12.

Norman DC (2000). Fever in the elderly. *Clinical Infectious Diseases* **31**, 148–51.

Petersdorf RG, Beeson PB (1961). Fever of unexplained origin: report on 100 cases. *Medicine* **40**, 1–30.

Sepkowitz KA (1999). FUO and AIDS. *Current Clinical Topics in Infectious Diseases* **19**, 1–15.

7.3Biology of pathogenic micro-organisms

T. H. Pennington

Why clinicians need to know about the biology of pathogens

Joseph Lister in Glasgow in the late 1860s was the first successfully to apply a hypothesis about the biology of pathogenic micro-organisms to the management of individual patients. It was that 'putrefaction.... as it occurs in surgical practice'. is caused by the 'germs of various low forms of life' that could be 'deprived of energy' by various chemicals without seriously injuring patients. The subsequent development of microbiology paved the way for the replacement of his antiseptic method by aseptic procedures—one of the most successful innovations in direct patient care that has ever happened—as well as the rational development of vaccines. Antibiotics followed. All of these have been enormously successful. But, even in the richest countries, the current importance of nosocomial infections, the long list of organisms for which vaccines are still desired, and increasing antibiotic resistance show that, despite a successful track record, much unfinished business remains. In poor countries, infections such as tuberculosis are still common causes of premature death. Capping all these things is the regular— but unpredictable—appearance and intercontinental spread of totally new pathogens such as HIV.

Progress towards resolving these problems will very often depend on achieving a better understanding of the relevant pathogens. Genomics promises much. Equally important for practitioners and for medicine, however, is the more effective application of the science that is already known. Both of these considerations explain why a general understanding of the biology of pathogens is important in clinical practice.

The ecology of pathogens

For a parasite to be successful, it must be capable of invading and maintaining itself in a host population. For it to spread, its basic reproductive rate— the average number of progeny that it produces— must be greater than one. Our aim for pathogens is to reduce this below unity, both in patients and in human communities. World-wide, polio may be eradicated soon, but smallpox is our sole success so far. Only infecting humans, it had a low efficiency of transmission which could effectively be interrupted by isolating patients and surrounding them with a ring of individuals made resistant to infection by immunization— which had already reduced the number of susceptibles to a very low figure. These factors— host range, patterns of transmission, and those affecting susceptibility—are key biological parameters for understanding why infections occur and for devising countermeasures. Knowledge about them enables us to explain why, for example, the pattern of *Salmonella* infections has undergone such a dramatic change in Britain since the end of the nineteenth century. It is a reasonable estimate that the overall incidence of human infections with species of this genus was much the same in 1880 as 1990. But infections with *S. typhi*—which accounted for nearly all those contracted in 1880—have almost completely disappeared to be replaced by those caused by *S. enteritica* serovars. This is because the human-to-human faecal–oral spread of *S. typhi* has been interrupted by the provision of clean water and safe sewers, while the recent cultivation of poultry on a massive scale in crowded hen houses with ample opportunities for faecal–oral spread, and subsequent processing that contaminates carcasses with gut contents, has provided bacterially contaminated food on a grand scale.

The lists of important transmission routes and sources of pathogens are short. Person-to-person skin contact, sex, mother to fetus transmission, inoculation of pathogens by arthropod or animal bites and other wounds and by needles, ingestion in food and water, inhalation in droplet nuclei, and transmission by fomites are the important routes. Exogenous sources of infection are other individuals, animals, the environment, and arthropods (as an intermediate but replicative step in transmission from other humans or animals). Many bacteria and viruses are facultative pathogens. Organisms such as *Staphylococcus aureus* and *Streptococcus pneumoniae* spend most of their time living harmlessly on the skin or in the throat, and a significant proportion of the clinically important infections they cause are endogenous, in that they occur in the individuals who carry them. Consideration of these factors provides explanations as to why some public health measures have been much more successful than others, and why current problems, such as nosocomial infections, still occur. Thus, it is easier to filter and chlorinate water than to change human sexual behaviour. The difficulty of changing human behaviour has also provided a major obstacle to the implementation of the critically important safety measure in preventing person-to-person spread and in food safety—hand washing.

The classification of pathogens

Many different bacteria and viruses cause disease in humans. Biodiversity is their hallmark. The modern classification of bacteria is based on DNA sequence data. So far, only a few genomes have been completely sequenced and relationships have been worked out using data from limited regions of the genome, particularly those coding for 16S ribosomal RNA—a relatively stable molecule from an evolutionary standpoint (Fig. 1). Viruses are classified according to schemes which take account of the nature of the nucleic acid in the virus particle—DNA or RNA—and its size and dispositions, particle structure—particularly crystallographic symmetry, size, and the presence or absence of a lipid envelope—and the pathways of virus messenger RNA synthesis (Fig. 2). Only a few general rules emerge from these groupings. Thus, the ability to form spores (which drives the need for autoclaves) only occurs in two medically important Gram-positive genera, *Clostridium* and *Bacillus.* Gram-negative and Gram-positive bacteria have different cell wall structures which play an important role in determining the different antibiotic sensitivities of organisms in these categories. Endotoxin, a major virulence factor, is the lipopolysaccharide component of the Gram-negative cell wall.

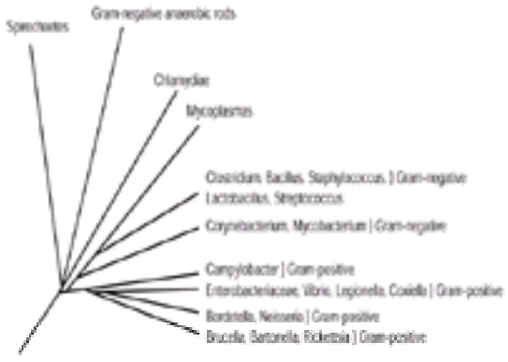

Fig. 1 Phylogenetic tree of medically important bacteria based on 16S rRNA sequence comparisons.

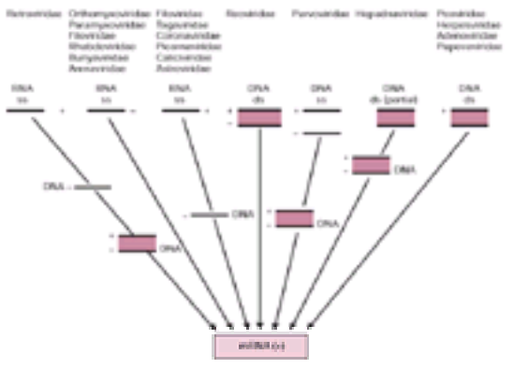

Fig. 2 Grouping of viruses by genome composition and pathway of mRNA synthesis.

Only viruses with DNA genomes (such as herpesviruses) or with RNA genomes that are copied into DNA as part of the infection cycle (such as retroviruses) set up long-term or permanent residence in the host; sometimes this is associated with the formation of malignant tumours. Genome size is inversely proportional to the rate of evolution of natural populations. A small genome, coupled with high mutation and replication rates and high yields, explains why the evolution of some viruses proceeds millions of times faster than that of the human host. In one or two cases, notably HIV, the evolution rate is so great that the virus population in a patient comprises a 'quasispecies' made up of a 'cloud' of variants a few mutational steps away from a master sequence—the sequence of highest fitness—and the master sequence itself. For the influenza viruses a high evolution rate is facilitated by a high recombination rate (strictly speaking a reassortment of segments of its naturally fragmented genome). Other RNA viruses, such as measles virus and poliovirus, show much less genetic variation during natural infections—one reason why their empirically developed vaccines continue to be successful many years after their development.

The fact that they have not been designed, but have evolved, explains why pathogens are so diverse and why so few general principles can be drawn from the groupings that emerge from their classification. Nevertheless, important guides to practice do emerge from evolutionary considerations. Thus the use of antibiotics, antiviral drugs, and vaccines exert strong selection pressures on pathogens and so should only be used in the context of evolutionary thinking. Evolutionary theory also generates explanations as to why virulence has evolved to certain levels; these may in turn generate hypotheses about how to reduce it.

Virulence determinants

Just as pathogens show enormous biodiversity, so many different mechanisms—virulence determinants—are involved in the causation of disease by them. The coincidental operation of multiple factors is nearly always needed to cause disease. A classical and notable exception is botulism, where the ingestion of preformed exotoxin in food is sufficient.

For most pathogens our knowledge of which virulence factors they possess and how they work together is significantly incomplete. The practical impact of this on clinical work is that it often limits the help that laboratory tests can provide in assessing the virulence of organisms isolated from patients. Indeed, tests for the production, or capability of producing, exotoxins, as in the cases of *Corynebacterium diphtheriae* and enterohaemorrhagic *E. coli* (such as *E. coli* 0157 H7), provide about the only examples where routine laboratory testing for virulence factors is used to distinguish between virulent and avirulent strains of the same species. Lack of knowledge about virulence factors and how they work in concert is a particularly important deficiency in the area of emerging and rapidly evolving pathogens such as influenza virus, and antibiotic resistant bacteria such as MRSA (methicillin-resistant *Staphylococcus aureus*). Thus it is still not possible, confidently, to predict from the details of its genetic structure whether an influenza virus will be more or less virulent—despite the very great depth of knowledge accumulated over many years about the sequence, function, and variation of the genes of the virus, and its structure, replication, and natural history in general. Likewise, laboratory tests have not helped to resolve the debate about whether—drug resistance apart—MRSA have added clinical importance because they are more virulent than other *Staphylococcus aureus* strains.

Adhesion and cell entry

To cause disease after being transmitted to its host, a pathogen must adhere to cells or other structures such as connective tissue or foreign bodies. Some bacteria produce clinically apparent effects when growing extracellularly; for others, entry into cells is an essential part of the disease process. Obligate intracellular parasites such as viruses and some bacteria must go through this step in order to replicate. The rates and extent of subsequent local or distant growth are usually important determinants of disease outcomes. Specialized structures, molecules, and processes are used to carry out these steps. For bacteria, surface structures and molecules important for adhesion include fimbriae (pili)—hair-like structures synthesized by Gram-negative bacteria—, flagella (in *Vibrio cholerae* and *Campylobacter jejuni*), outer membrane proteins (particularly in *Neisseria gonorrhoeae*), exopolysaccharides (in cariogenic streptococci and some staphylococci), and collagen, fibronectin, and fibrinogen binding proteins (as in *Staphylococcus aureus*). Many important pathogens employ molecular mimicry. In some, their adhesion molecules may mimic natural ligands, such as the *Bordetella pertussis* haemagglutin which mimics integrins. In others, the organism binds to host ligands which then bind to their natural receptors, as with *Legionella pneumophila* which becomes coated with complement components which then bind to complement receptors.

One of the main functions of a virus particle is to optimize the likelihood of infection by providing structures that interact with molecules on the surface of host cells and prepare the way for virus entry into them. All animal viruses carry multiple attachment proteins or sites on their surface. In some, these are discrete structures, such as the glycoprotein spikes on the surface of influenza virus. In others, such as picornaviruses, receptor binding sites are created at the interfaces between different viral polypeptides. In some large viruses, such as *Herpes simplex*, an initial low-affinity interaction with one receptor is followed by a slower, tighter, binding to another. Viruses enter cells either by fusing their envelopes with the cell membrane and internalizing the nucleocapsid, or by receptor-mediated endocytosis into endosomes, subsequent fusion of the virus with the endosome membrane, and entry of its genome into the cytoplasm.

The bacterial species that enter cells when infecting humans are all important pathogens. They include *Mycobacterium* species, *Salmonella*, *Shigella*, pathogenic *E. coli, Yersinia, Legionella*, and *Listeria.* For *Yersinia* and *Listeria* internalization correlates with a strong adhesion of bacteria to the cell surface (mediated in *Yersinia* by interactions between bacterial invasion and multiple integrins, and in *Listeria* by its internalin interacting with cellular E-cadherin). Cellular pseudopods extend around the bacterium and engulf it in a tight phagosome. *Shigella* and some pathogenic *E. coli* inject proteins into cells which cause rearrangements which promote and lead to entry. The *Shigella* Ipa protein enters the cell and associates with vinculin. A major rearrangement of the host cytoskeleton follows in which actin linked to the cell membrane causes the formation of projections which engulf the bacterium, cell entry occurring by micropinocytosis. Other pathogens, as taxonomically diverse as *Listeria* and vaccinia virus, also associate with the cytoskeleton. They use actin cables to move about the cell during the intracytoplasmic part of their life cycles.

Spread of pathogens within the body

Local, restricted spread is sufficient for some pathogens to cause common and self-limiting or, with a few bacterial species, lethal conditions. The commonest clinically apparent staphylococcal infections—boils—fall into the first category, as do rhinovirus infections of the upper respiratory tract. Localized infections may be life threatening because of the productions of exotoxins that act at a distance, as with *Clostridium tetani*, or exotoxins that act locally as well, as with *Corynebacterium diphtheriae*, *Clostridium perfringens* when causing gas gangrene, and *Streptococcus pyogenes* when causing necrotizing fasciitis. A few bacterial pathogens, such as *Salmonella typhi* or *Treponema pallidum*, invade and grow in organs such as the brain, bones, and skin as a matter of course. For the majority of common bacterial pathogens such outcomes occur in only a small minority of infected individuals. Thus the number of individuals that develop meningitis or septic arthritis is much smaller than the number who carry *Neisseria meningitidis* in their throats or suffer from gonorrhoea. Nevertheless, all these pathogens must possess mechanisms—many of which are poorly understood at present— that allow them to reach blood vessels, traverse the endothelial lining, and use the circulation to travel to distant organs. Similar considerations apply to viruses. While some remain localized, others spread widely. The circulation is the most important route. Viruses enter either via lymphatics or are shed from infected endothelial cells or circulating mononuclear leucocytes. Then they circulate free in the plasma or associated with formed elements. During this process some replicate in monocytes, T or B lymphocytes, or, occasionally, erythrocytes. A particularly important example is HIV, which infects CD4+ T lymphocytes and monocytes in the blood, particularly the former. Some viruses are disseminated in the body by spreading along peripheral nerves, an essential route for rabies and certain herpesviruses, and a supplementary one (to viraemia) for polio- and reoviruses.

Factors that neutralize host defences

Many pathogens employ mechanisms which optimize their chances of establishing themselves and growing in a host. Some bacteria that live in the respiratory tract, such as *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Neisseria meningitidis*, produce a protease that selectively destroys host secretory IgA. Bacterial polysaccharide capsules protect against phagocytosis— unless opsonized by the attachment of specific antibody—and against the activity of the alternative complement pathway. Some bacteria evade the host immune response by taking up an intracellular location; it is considered that *Mycobacterium tuberculosis* can survive inside macrophages for long periods, setting up a latent infection which can reactivate and cause disease many years after the initial infection. Large DNA viruses code for proteins that interfere with host mechanisms that respond to infection. A major herpes simplex glycoprotein binds the C3b and iC3b components of the complement cascade protecting against complement-mediated virus neutralization and lysis of infected cells; likewise the vaccinia virus complement control protein binds to C4b—mutations in this gene result in reduced virulence. Poxviruses also induce the synthesis of many other proteins that interfere with host cytokine functions, including soluble receptors for interleukin-1b (IL-1b), interferon g, and tumour necrosis factor (TNF), serpins that inhibit cytokine activation, and proteins that inhibit the expression of MHC-1 antigens at the cell surface.

Factors that damage the host

Bacterial infections damage hosts in three different ways. Bacterial cell components are directly toxic; some bacterial species produce exotoxins which damage cells, cell function, or physiological processes; and the host response to infection may itself contribute to pathology.

The lipopolysaccharide component of the cell walls of Gram-negative bacteria, particularly its core region lipid A, is an important virulence factor. Lipid A stimulates the production of cytokines such as IL-1 and TNF by macrophages. Gram-positive cell wall and other surface structures components may also be toxic. A diversity of bacterial species produce exotoxins. They are important because their toxicity can be very great and because medical measures can be targeted against them by developing toxoids for prevention, and antitoxins for treatment. The former have been much more successful in practice than the latter. Exotoxins are nearly all medium sized to large proteins. Multicomponent ones, such as cholera toxin with one A subunit and five B subunits, are the exception. Some are inactive when produced and need to be cleaved by proteolysis, such as cholera toxin, the *Clostridium botulinum* toxins, and diphtheria toxin. The B region of the latter molecule, for example, binds to cells and is proteolytically nicked to release another part, the A region, which enters the cell and exerts its toxic effect. Some exotoxins cause very different diseases but have similar actions at the molecular level. Thus both the cholera toxin and a pertussis toxin activate adenylate cyclase by enzymatically ribosylating its GTP-binding regulator, and diphtheria toxin causes the ADP ribosylation of elongation factor 2. The end result of these actions on individuals is, of course, quite different. Activation of adenylate cyclase by pertussis toxin damage inhibits the functions of immune effector cells, while in cholera it stimulates chloride secretion and the inhibition of sodium uptake in villus and crypt gut cells causing profuse, watery diarrhoea. In diphtheria, cell protein synthesis is turned off. Some toxins are highly organ specific, such as the neurotoxins of *Clostridium botulinum* which cause flaccid paralysis by acting presynaptically at neuromuscular junctions where they inhibit the exocytosis of vesicles containing acetylcholine. Others have a more widespread effect, such as the staphylococcal toxins that act directly on cell membranes, the a toxin causing cytotoxic damage by creating transmembrane channels, and the b toxin, a phospholipase active on sphingomyelin, being cytolytic for red cells, platelets, and macrophages.

Viruses do not produce exotoxins, neither do their virions contain toxic components such as endotoxins. They produce disease by killing cells (directly or indirectly), by altering patterns of cell growth, and by immunopathological mechanisms. The irreversible inhibition of host cell protein synthesis is a common consequence of infection. A variety of mechanisms selectively turn off cell protein synthesis. They include host mRNA degradation and the inactivation and subversion of components involved in the translation of capped mRNAs. The induction of apoptotic cell death by virus infection has been clearly shown in experimental models. Its role in the causation of human disease is not clear. The induction of benign or malignant tumours by viruses is not uncommon. Representatives of all the major subdivisions of viruses with double-stranded DNA have been implicated in neoplasia, and it has been estimated that up to 20 per cent of human tumours have a viral risk factor. Oncogenes occur in retroviruses and some DNA viruses; the association of other virus infections with tumour formation is indirect and usually associated with cofactors such as malaria in the case of Epstein–Barr virus and Burkitt's lymphoma. The mechanisms responsible for virus-induce immunopathology are exemplified by hepatitis B, where the destruction of virus-infected hepatocytes can follow attack by CD8 T cells, and rashes, glomerulonephritis, and polyarteritis nodosa can be caused by immune complexes.

Why do pathogens cause disease in particular patients?

The possession of virulence factors by a pathogen are necessary, but not usually sufficient, for it to cause disease. In other words, infections with most pathogens are asymptomatic in at least a minority of patients. Factors which may increase the probability of host damage include pathogen dose and route of transmission, particularly if the latter by-passes host defences. In individual patients, it is hardly ever possible to measure pathogen dose and prognostic attention usually focuses on host factors. Many are important, including the efficiency of non-specific and immunological host defences in neutralizing virulence factors and inhibiting pathogen growth. Previous infections can play roles through immunosuppressive effects, as in HIV and AIDS, or by causing tissue damage, as in the bacterial pneumonia that sometimes follows influenza. Genetically determined host susceptibilities to infection include various immunodeficiencies, and conditions which enhance the ability of pathogens to cause tissue damage. Human B19 parvovirus infections provide a good example. The much shortened red cell survival times in sickle-cell disease and hereditary spherocytosis combine with the preferential replication of this virus in erythroid precursor cells to lead to complete red cell aplasia in individuals with these conditions, in contrast to the transient reduction in reticulocyte count seen in normal individuals.

Age is also an important factor in determining the outcome of infection. Although as a general rule disease severity is more severe in the very young and the very old, pathogen-specific features such as the severe organ damage produced by rubella in the first trimester of pregnancy are also important. The biological basis of age-determined susceptibility is not known for most pathogens, even when the effect is consistent and marked. In untreated typhus, for example, mortality rates increase steadily with age, being negligible in children, about 5 per cent in young people under 20, 50 per cent in 50-year olds, and 100 per cent in those over 60.

All the factors discussed in this chapter are the result of evolution. Evolutionary studies on pathogenicity have only just started but they are beginning to provide explanations about how organisms evolve towards virulence. Bacteria reproduce asexually and so have clonal population structures. Horizontal gene exchange is not infrequent, however, and genomes have mosaic structures built up of a basic clonal framework peppered with DNA regions derived from other organisms. These include 'pathogenicity islands'—stretches of DNA containing many virulence genes. Toxin genes also move around on bacteriophages and plasmids. Viruses have incorporated virulence genes, such as oncogenes, from their hosts. Why organisms become pathogens is much less certain.

Further reading

General

Anderson RM, May RM (1991). *Infectious diseases of humans. Dynamics and control*. Oxford University Press.

Collier L, Balows A, Sussman M, eds (1998). *Topley and Wilson's microbiology and microbial infections*. Vol. 1 Chapters 1–13, Vol. 2 Chapters 1–12, Vol. 3 Chapters 1–11. Arnold, London.

Stearns SC, ed (1999). *Evolution in health and disease*. Oxford University Press.

Specific reviews illustrating general principles

Abu Kwaik Y (1998). Fatal attraction of mammalian cells to *Legionella pneumophila. Molecular Microbiology* **30**, 689–95.

Frankel G *et al.* (1998). Enteropathogenic and enterohaemorrhagic *Escherchia coli:* more subversive elements. *Molecular Microbiology* **30**, 911–21.

Hinton JCD (1997). The *Escherichia coli* genome sequence: the end of an era or the start of FUN? *Molecular Microbiology* **26**, 417–22.

O'Brien V (1998). Viruses and apoptosis. *Journal of General Virology* **79**, 1833–45.

Petit L, Gilbert M, Popoff MR (1999). *Clostridium perfringens:* toxinotype and genotype. *Trends in Microbiology* **7**, 104–10.

Smith DB *et al*. (1997). Virus 'quasispecies': making a mountain out of a molehill? *Journal of General Virology* **78**, 1511–19.

Sokurenko EV, Hasty DL, Dykhuisen DE (1999). Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends in Microbiology* **7**, 191–5.

Wallinga J, Edmunds WJ, Kretzschmar M (1999). Perspective: human contact patterns and the spread of airborne infectious diseases. *Trends in Microbiology* **7**, 372–7.

7.4 The host response to infection

Leszek K. Borysiewicz

Introduction

The expansion in our knowledge of the pathogenesis of infectious diseases over the last two decades is principally a consequence of the development of analytical techniques for investigating biological molecules. Our greater understanding of the biology of infectious agents, including complete genomic sequences of pathogens, as well as the nature of the host response, provides new opportunities for therapeutic research. The interdependence of host and parasite shows the importance of selective pressure from microbes on the evolution of human immune responses. None the less, the central and perplexing question confronting clinicians remains unanswered—what are the principal determinants of disease in an individual patient? This chapter cannot answer this question, but it does describe developments in the field and how they may impact on our understanding of infectious diseases.

The nature of the host–pathogen interaction

Although the parasite–host interaction is often portrayed as a balance of host and parasite factors (Fig. 1), the full range of interactions is large and complex. Symbiosis occurs in the gastrointestinal tract even though factors involved in pathogenesis, such as bacterial adherence and colonization, also operate—but usually with benefit to the host. True pathogenic effects are seen when the host–parasite interaction is modified, such as antibiotic-associated colitis. The survival advantage to a pathogen of host damage is variable and often the best-adapted parasites find an ecological niche and disseminate with minimal disturbance to the host.

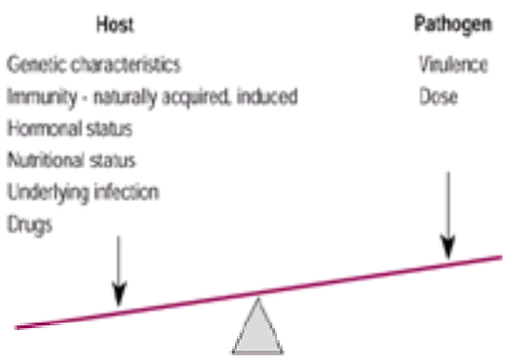

Fig. 1 Factors affecting the host/parasite balance.

Pathogen factors

The molecular basis of microbial pathogenicity has advanced significantly through a genomic approach. Since the sequence of *Haemophilus influenzae* was reported in 1995, more than 50 complete bacterial genomes have now been sequenced. They range from 0.58 Mbp for *Mycoplasma genitalium* to 4.4 and 4.6 Mbp for *M. tuberculosis* and *Escherichia coli*, respectively. Analysis of these genome sequences supports earlier work suggesting that bacterial populations are derived from clonal lineages. This concept implies that genetic exchange (through recombination or horizontal exchange using mobile genetic elements) occurs at low frequency, otherwise a greater heterogeneity of bacterial populations would be observed. However, horizontal transfer is an important source of diversity in bacterial genomes (~18 per cent in *E. coli*). Some pathogenic bacteria are remarkably homogeneous; there are few clonal populations of *Shigella sonnei* and *Bordetella pertussis* in which a single clonal type represents the whole species. Other species such as *Neisseria gonorrhoeae* and *N. meningitidis* show random chromosomal rearrangement, although, during outbreaks of meningitis, a few clonal types dominate the bacterial isolates. Multiple clones of *E. coli* are present during colonization but pathogenic clones persist, suggesting a process of selection within the host environment.

Virulence factors

Observations based on genome analysis combined with information on pathogenic factors, suggest that three mechanisms, each operating on a different time-scale, permit the acquisition of pathogenicity within a clonal subpopulation of bacteria:

- Transfer of virulence genes on mobile genetic elements, such as bacteriophages and plasmids, is rapid, and in the clinical context that allows the acquisition of antibiotic resistance in a population of bacteria during a single episode of infection. Other virulence factors are similarly transferred: for example, cholera toxin is encoded by a vibriophage, which integrates into the *Vibrio cholerae* chromosome.
- Random chromosomal transfer and bacterial transformation allows the acquisition of factors by *N. gonorrhoeae* to replicate in a host.
- The above mechanisms may be coupled with chromosomal relocation of genes in pathogenic bacteria to create 'pathogenicity islands'. These are segments of the bacterial genome (up to 200 kbp) that are flanked by repeat or insertion elements. The segments encode clusters of virulence genes including adhesion molecules, toxins, and secretory and regulatory proteins.

Bacterial virulence factors are regulated such that they are expressed in particular environmental circumstances. Common factors triggering expression include: changes in ionic concentration, pH, iron concentration, and temperature: for example, induction of K-88 and K-99 and pyelonephritis-associated fimbriae in uropathogenic *E. coli*. The factors regulating transcription of these genes are often aggregated into 'regulons' as exemplified by *Bordetella pertussis*. The bacterium expresses a number of virulence factors, such as pertussis toxin, haemagglutinin, fimbrial protein, and pertactin, regulated in a co-ordinated manner by the *bvgAS* chromosomal locus. The locus encodes a BvgS protein that spans the periplasmic membrane. At 37 °C, BvgS autophosphorylates with later transfer of the phosphorylation signal to BvgA, a bacterial regulator that can bind to specific regions of *Bordetella* DNA, thereby promoting the specific transcription of virulence factors as well as amplifying its own signal and repressing other genes. Thus a single signalling cascade can induce a variety of changes responding to several environmental stimuli in a co-ordinated manner.

Although many virulence factors are secreted, others are surface molecules that promote adhesion and bacterial entry by intracellular pathogens as well as interfering with specific host immunity. Intermittent expression of such factors on bacterial pili in *N. gonorrhoeae* reduces the ability of the host to mount a neutralizing antibody response. The mechanisms regulating antigenic variation at this locus and the correct orientation of such factors in pili are the consequence of gene rearrangements in chromosomal DNA and are essential for immune evasion by *N. gonorrhoeae*. Similar mechanisms operate in larger parasites such as *Trypanosoma brucei*.

Route of entry

Invasion of the host to access its environmental niche requires a pathogen to overcome physical barriers, as well as specific defence mechanisms, to establish itself. Many organisms have more than one phase in their lifecycle, either in another host (for example, malaria) or persisting outside a host by an alteration of its lifecycle (for example, encystment in *Amoeba* spp. or spore formation in *Clostridia* spp.). The skin is a relatively impermeable barrier which the pathogen must bypass. This can be achieved for some pathogens by insect vectors or direct trauma with penetrating wounds: for example, a needlestick injury for transmission of the human immunodeficiency virus (**HIV**) and hepatitis C. Even those organisms that produce local skin infections have to either penetrate the skin at sites of injury to gain access to susceptible cells in the basal layers of the epithelium (for example, human papillomaviruses) or infect adnexal structures (for exam-ple, staphylococcal folliculitis).

Alternatively, pathogens enter at mucosal surfaces; but here again physical barriers limit the ability of the pathogen to be retained, let alone replicate at such sites. Many pathogenic factors encoded by bacteria are adhesion molecules that interact with specific host receptors, for example to enable the pathogen to avoid the mucociliary carpet of the respiratory tract or the urothelial/gastrointestinal tract. However, such interactions are not inert because receptor–ligand binding may trigger cellular activation, which itself can be utilized to promote pathogenicity.

For intracellular parasites such as viruses, host interactions also provide the mechanism of entry to the intracellular environment. One of the best-studied mechanisms for the initiation of infection is the interaction of the haemagglutinin of orthomyxoviruses, such as influenza A, with sialic acid. The influenza haemagglutinin (**HA**) is present in the viral envelope as homotrimers of 550 amino-acid chains attached to a membrane-spanning and a short cytoplasmic domain. The molecule has a globular head that includes a conserved receptor pocket in the HA_1 globular region and a long stalk that incorporates an HA_2 domain. Depending on minor variations in HA_1, there are preferences between influenza strains for the optimum configuration of the presented sialyl residues that are bound. After the virus has bound to the cell surface, it is endocytosed and transferred to the endosomal compartment where the pH is reduced to between 5 and 6. Acidification results in an irreversible conformational change, which exposes a hydrophobic domain of amino acids in HA_2, normally concealed in the stalk of the intact molecule. The hydrophobic domain interacts with the lipid bilayers to result in membrane fusion and release of the viral genome into the intracellular environment to initiate virus replication. This mechanism of entry is blocked by amantadine, which can inhibit virus replication by raising the pH within cellular endosomes thus preventing haemagglutinin cleavage and exposure of the hydrophobic domain. Influenza mutants resistant to amantadine have been identified, these are characterized by changes in the hydrophobic region to initiate membrane fusion at higher pH, thereby bypassing the action of this drug.

The molecular structure of the binding domain for the host cell receptor is central to the process of infectivity and the establishment of virus infection. It is a common structural feature of viruses that this domain is often shielded by numerous loops—giving rise to the concept of a receptor 'canyon', first promulgated in the context of rhinovirus infection and its interaction with the intercellular adhesion molecule, **ICAM-1**. Such loops can enhance the specificity of a receptor–ligand interaction, but they also deflect the generation of antibodies from the receptor-binding site itself. This is an important virus evasion mechanism, since greater variation in these loops can be accommodated than in the sterically sensitive receptor site.

Adhesion is essential for bacterial pathogenesis. This is usually achieved through fimbriae or pili, present as 100 to 1000 2- to 7-nm rod-like structures often identified as virulence factors. Uropathogenic *E. coli*, express so-called P-fimbriae in up to 90 per cent of patients with pyelonephritis. These *E. coli* bind to uroepithelial cells through glycolipids expressing a galactose–galactose (**Gal–Gal**) carbohydrate moiety. P-fimbriae genes are expressed from the *pap* gene operon in *E. coli*—*papA* gene products form the core of the fimbrial rod, attached to the membrane through *papH*, and this assembly is protected by two-pap operon-encoded chaperones. The ends of the P-fimbriae consist of fibril-associated proteins, with the *papG* product at the tip acting as the Gal–Gal binding unit. There are at least three 'alleles' of *papG* with slightly different affinities for Gal–Gal, which in turn affect disease association in the urinary tract. However, there is added complexity, in that the whole P-fimbriae complex is regulated, possibly by the *papI* and *papB* units of the operon, enabling the detachment and migration of bacteria to new sites. Therefore local pathogenicity and transmission in a single host may be regulated in response to environmental change.

The ability to interfere with bacterial–host cell binding is attractive for prophylaxis or therapy. The presence of excess soluble Gal–Gal subunits blocks bacterial adhesion, and immunization with *E. coli* bearing class-II PapG protects primates against pyelonephritis by the development of P-fimbrial antibodies. However, protection is limited to a specific allele on the P-fimbriae and little cross-protection against other strains of *E. coli* is observed.

The adhesion of *E. coli*, in particular enteropathogenic types, alters epithelial-cell morphology to facilitate bacterial colonization. Binding of the bacterial adhesin intimin to a membrane receptor (**Tir**, translocated intimal receptor), results in the host membrane developing pseudopodia or 'pedestals', via a cytoskeletal modification process, to aid bacterial attachment but not cell invasion. This is regulated through a bacterial pathogenicity island—'locus for enterocyte effacement'. However, this effacement is unusual in that the 'cell' receptor is also of bacterial origin. Tir is secreted from the bacterial cell through a specialized secretory pathway along with other molecules that cause the epithelial cell to phosphorylate Tir, as well as producing other changes in the activation state of the epithelial cell. Other gastrointestinal pathogens induce similar epithelial-cell adaptations such as the epithelial ruffling associated with *Salmonella* spp., in this instance facilitating bacterial uptake into the optional intracellular environment.

Bacteria also colonize surfaces as 'biofilms'—arrangements of populations of bacteria, which coexist on an inert surface surrounded by a matrix that contains microchannels which facilitate the circulation of micronutrients required for colony survival. Microfilms have evolved to promote bacterial survival in adverse environmental conditions; they occur naturally in dental plaque. Microfilms have assumed great importance because of the bacterial biofilm contamination of indwelling devices including intravenous catheters, peritoneal dialysis tubing, urinary catheters, mechanical valves, and orthopaedic prostheses. The formation of biofilms is associated with a 500-fold enhanced antibiotic resistance related to the presence of the matrix encapsulating the micro-organisms. *Staphylococcus epidermidis*, a common skin commensal, is often associated with the development of such films on indwelling devices. The initial adherence to plastic is through fimbrial adhesion proteins SSP-1 and SSP-2, which is followed by bacterial replication. Secretion of a polymer of b1–6-linked, 2-deoxy-2-amino-glucopyranosyl residues, in a structure that resembles cellulose, is regulated by a bacterial operon, and this allows the formation of bacterial colonies in the biofilm.

Pathogen dose

One important pathogenicity factor, frequently forgotten, is the dose of pathogen that is delivered at the portal of entry. The numbers of organisms required for the development of clinical syndromes are dependent on the site of entry as well as on the nature of the pathogen. Conclusions are often drawn from experimental infections that use inappropriate routes of infection, and excessive pathogen doses. It is difficult to establish the infectious dose needed to produce disease through natural routes in humans. Studies rely entirely on volunteers, which in some instances have been held as unethical. The infectious doses for respiratory virus based on direct inoculation studies suggest that 1 $TCID_{50}$ is sufficient to produce infection via the nasal cavity but that a higher dose is required at the posterior pharyngeal wall.

Large doses (>10^8 bacteria) of water-borne pathogens such as *Vibrio cholerae* are needed to produce a gastrointestinal infection, this is because many bacteria are inactivated by gastric acidity—the infectious dose administered with bicarbonate is reduced to 10^4 bacteria. About 10^5 salmonellae are required for infection, while only 10 to 100 shigellae are needed because of their resistance to gastric acid.

Encysted forms of organisms are particularly resistant to gastric acid—mathematical modelling of a water-borne outbreak of cryptosporidiosis in Milwaukee suggested that ingesting a single oocyst may have been sufficient to produce disease.

Host factors

The susceptibility of the host to infection is dependent on the ability of a pathogen to break down or bypass host barriers, physical and non-specific defence mechanisms which developed early in phylogeny.

Mechanical and local barriers to infection

Skin and epithelial surfaces

Provided the morphological integrity of local epithelial surfaces is maintained, epithelia represent a harsh environment for micro-organisms. Skin is dry and slightly acidic (pH 5–6) due to fatty acids, which are often produced by the local bacterial microflora from the hydrolysis of triglycerides. In addition, secretions such as sebum may be protective, while physical desquamation itself prevents effective colonization by adherent pathogenic species.

Mucosal barriers are more inviting to micro-organisms because of humidity, but this is offset by an established microflora and secretions with antimicrobial properties: for example, lysozyme and *N*-acetyl muramyl-L-alanine amides which cleave the amino-acid backbone of the peptidoglycans found in Gram-positive organisms. Each anatomical site has specific antimicrobial barriers, as discussed below.

Respiratory epithelium

- Air filtration provided by the upper respiratory tract and large bronchi ensures that larger particles are deposited on bronchial mucus and that penetration to the bronchioles and alveoli only occurs with particles smaller than 5 µm in diameter. The large numbers of suspended bacteria in air (approximately 400–900 organisms/m^3 in buildings, resulting in 10^4–10^5 organisms inhaled and cleared each day) emphasizes the efficiency of this barrier.
- A mucociliary blanket clears 90 per cent of surface-deposited material in less than 1 h
- Secretion of defensins and collectins in alveoli help in the opsonization of organisms for clearance by alveolar macrophages.

Intestinal epithelium

- Gastric acidity provides a direct barrier—bacteria penetrate more readily to intestinal sites in the presence of achlorhydria.
- Pancreatic, biliary, and intestinal secretions contain antimicrobial agents, including enzymes and defensins.
- Peristalsis—use of atropine-like agents can prolong the duration of infection with *Shigella* spp.
- Normal bowel flora

Genitourinary epithelium

- Urinary pH and osmolarity is inhibitory to bacterial growth.
- Tamm–Horsfall protein may compete for bacterial binding to the epithelium.
- Urinary voiding—bacterial urinary tract infection is more common where there is a genitourinary obstruction.
- Lactobacilli, present at 10^8/ml in the vaginal microflora, degrade glycogen to lactic acid thereby producing an acidic environment.

The role of commensal microflora at epithelial surfaces

This important barrier to pathogenic micro-organisms is acquired from the environment after birth and is modified throughout life. Although the bacterial interactions in this environment are poorly understood, perturbations caused by the therapeutic use of antibiotics that bring about the establishment of an inappropriate flora readily exemplify its importance. The precise mechanisms of protection are unknown, but commensal bacteria may:

- compete for adhesion sites with pathogenic organisms;
- produce antibacterial compounds to limit bacterial expansion from outside the ecosystem;
- compete with pathogens for nutrients;
- produce metabolic products that limit the growth of other species: for example, lactobacilli in vaginal secretions;
- induce generation of 'natural' antibodies.

Non-specific protective immunity

This involves several defensive mechanisms mediated through cellular and humoral mechanisms.

Complement

Complement is an evolutionary conserved cascade of proteins that not only provides direct protection against invading micro-organisms but also augments the effector functions of specific immune responses (Chapter 5.4). These include:

- *Complement-mediated lysis* is largely dependent on the presence of antibody; polysaccharide-coated bacteria can directly activate this alternative pathway, as can certain virus-infected cells (for example, measles), but this is seldom enough to mediate direct damage. There is little direct evidence that complement lysis in isolation is protective, although deficiency of the terminal pore-forming C8/9 components predisposes to neisserial infection.
- *Opsonization and neutralization*.
- *Induction of inflammation.*

Febrile response

Although body temperature is tightly regulated, no single 'temperature centre' has been identified in the central nervous system. A series of structures in the reticular formation, limbic system, and lower brainstem, including the hypothalamic preoptic region with thermosensitive neurones, can initiate behavioural thermoregulatory responses. These regions become responsive by increasing the rate of neuronal firing to a variety of interactive cytokines, either by direct receptor interactions or through intermediary mediators, such as prostaglandins. Micro-organisms induce the release of pyrogenic cytokines from macrophages, including interleukin (**IL**)-1a and -1b, IL-6, tumour necrosis factor-a (**TNF-a**), and interferon-g (**IFN-g**). Unfortunately, because of their complex interaction and involvement of intermediaries, it is difficult to dissect individual roles. Although cytokines may not cross the blood–brain barrier, they can alter the thermo-regulatory setpoint either by inducing phospholipase A_2 to release prostaglandin E_2 or using a common pathway involving IL-6.

Acute-phase response

Similar stimuli generating a febrile response also induce acute-phase proteins, which may modulate inflammation and tissue repair: for example, the C-reactive protein (**CRP**) binding of phospholipids on damaged cells or micro-organisms. It may also activate the complement cascade and promote anti-inflammatory products. Serum amyloid A may potentiate the inflammatory response by improving adhesion between effector cells and lipid uptake. Complement components and proteins, such as haptoglobin, with antioxidant activity are increased alongside a number of protease inhibitors (for example, a $_1$-antichymotrypsin).

Acute inflammatory responses and the role of the polymorphonuclear leucocyte

Polymorphonuclear neutrophils are produced at the rate of 10^{11} cells daily, with a marrow reserve allowing a 10-fold increase in the presence of infection. Neutrophils circulate in rapid flow as a marginated population of cells through transient endothelial interactions. Following establishment of an infection, polymorphs are recruited to the site by their interaction with specific endothelial and chemokine receptors. There are four distinct phases of migration recognized—rolling adhesion, integrin activation, firm adhesion, and transmigration—mediated, respectively, by selectins, integrins, immunoglobulin-like proteins, and mucin-like selectin ligands. The process is active both in the endothelial cell and neutrophil, necessitating the differential expression of receptors as well as an altered affinity of receptors. This is regulated by multiple interacting signals requiring accessory molecules and signalling through chemokines such as IL-8.

Phagocytosis of microbes is enhanced by receptor-mediated entry, especially if the microbe is opsonized by antibody or complement components. Once the microbe is ingested, the changes in the cytoskeletal architecture adjacent to the vacuole allow the orderly fusion of cellular proteins to ensure a sequential transfer to the phagosome and, ultimately, fusion with neutrophil granules. The pH in the granules falls to 3.5, and enzymes aiding the breakdown of bacteria are released. A distinct microbicidal mechanism is the generation of a respiratory burst enabling the production of reactive oxygen species that are toxic to bacteria. In addition, non-oxygen-dependent killing also occurs through a variety of mediators, including pH, enzymes, defensins, bactericidal/permeability-increasing protein (BPI), lactoferrin, and a range of cationic proteins. This complex interaction has resulted in the identification of numerous clinical syndromes associated with a failure of neutrophil function, all of which enhance susceptibility to infection.

A neutropenia of less than 1000 cells/mm^3, regardless of its clinical cause, carries a significant risk of Gram-positive and Gram-negative bacterial sepsis. More specifically, defects in intracellular killing as found in chronic granulomatous disease caused by different mutations in the NADPH-dependent oxidase complex, can have multiple effects, often depending on the severity of the impairment of the oxidative burst. In its most severe form, there is increased susceptibility to staphylococcal with Gram-negative bacterial infections which normally would be low-grade pathogens for example, *E. coli*, *Klebsiella*; failure to clear the pathogen often results in granuloma formation with viable pathogens in the presence of accumulated polymorphs and macrophages. The selective nature of the defect is illustrated by the observation that most neutropenic patients have minor problems with streptococcal infection. A mild counterpart in the family of neutrophil deficits is myeloperoxidase deficiency. Previously thought to be rare, this disorder is now identified as one of the most common gene defects (1 in 2–4000 of the population), but as yet has not been associated with susceptibility to infections, except perhaps an increased frequency of candidiasis.

While identification of these defects underlines the importance of the polymorph in host protection, their cellular nature does not render them easily accessible to passive replacement therapy—thus antibiotic prophylaxis is the mainstay of treatment. However, in the future, gene therapy-based approaches may offer the best option for cure.

Natural killer (NK) cells

These cells have long been recognized by their ability to kill both virus-infected and tumour cells *in vitro* without the requirement of MHC restriction. This activity is markedly enhanced by the presence of cytokines such as IFN-g during infection *in vivo*. The capacity of NK cells to recognize transformed or infected target cells does not depend on conventional antigen presentation. NK cells carry receptors which are C- type lectins that bind diverse carbohydrates or inhibitory receptors interacting with MHC molecules, that is to say killing occurs when MHC is downregulated. The specificity of the three immunoglobulin-like receptors is being defined, but their role in protective immunity may be broader. Human NK-cell defects are rare, but in one that was identified, the deficit was only identified as a result of an enhanced susceptibility to herpesvirus infections. Therefore, although the direct protective role may be limited, recent studies suggest that the initial NK response at sites of infection may be important for initiating local-early specific responses to the pathogen. Such initiation may represent a fundamental link between innate and specific immunity.

Specific immune responses

Vertebrates, and especially mammals, have evolved specific immunity, probably in response to the selective pressure of exposure to intra- and extracellular pathogens. Although originally protective, enhanced immunity may render some responses detrimental (for example, atopy and asthma) or occasionally augment the effect of an infection by immune-mediated injury. Each effector mechanism is tightly regulated to minimize this possibility and immune responses show unique properties of: specificity for antigens; the ability to turn off the response once an exposure is cleared; and memory with augmentation, such that repeated exposure results in a more rapid and augmented response. These detailed aspects and associated pathology are discussed elsewhere (Section 5), and only mechanisms associated with protective antimicrobial immunity are considered here.

Antibody-mediated protection

Antibodies consist of glycoprotein immunoglobulin molecules secreted in response to infection by T-cell regulated B cells. Antigenic specificity is provided by the primary protein structure at one end of the heavy and light chains that make up the molecule. These specificities are produced by a recombination to bring one of a large number of variable gene segments alongside constant-region genes before transcription and translation of the resultant complex. The biological properties of antibodies are a property of the heavy chains, and immunoglobulins are divided into classes based on the genes that encode the heavy-chain polypeptides. The molecules are invested with unique protective properties against microbial pathogens:

- *Neutralization of toxins and viruses*. One of the earliest observations of the clinical role of antibodies was the recognition that a soluble factor could block both bacterial toxins as well as pathogen entry, especially with viruses. This has led to the use of passive immunotherapy where antibodies can be administered postexposure in toxin-mediated disease (for example, tetanus) or in susceptible patients to block infection (for example, varicella-zoster virus infection in immunocompromised children). However, a wider utility of this approach with the development of monoclonal-antibody technology has not proved successful, for example the use of anti-J5 antibody in *E. coli* bacteraemia. Neutralization is often viewed as steric hindrance consequent on an antibody binding close to a receptor/ligand domain. But, particularly in the context of virus neutralization, this view is simplistic. Titration analysis of picornaviruses has shown that one or two antibody molecules may neutralize poliovirus infectivity despite the presence of 60 potential receptor-binding sites. The exact mechanism of inactivation is unknown, but it could be due to antibody-induced changes in the virion capsid or an inability of the virus to uncoat on entry into the cell.
- *Agglutination*. This may be considered another form of neutralization but is a result of polyvalent antibodies of IgM, IgA, and IgG classes crosslinking particles, thereby blocking adherence and infectivity—as well as enhancing susceptibility to phagocytosis.
- *Complement activation*. IgM and IgG are able to bind complement through the classical pathway via C1q, particularly when there is aggregation of IgG on the surface of the microbe. Binding of complement augments local inflammation and enhances phagocytosis consequent on opsonization.
- *Opsonization and phagocytosis*. Antibodies together with any bound complement components enhance phagocytosis by promoting receptor-mediated entry into phagocytes. Such interactions may also trigger cytokine release, as well as promoting enhanced bactericidal activity in the phagocyte by direct activation of the oxidative burst in these cells.
- *Antibody-dependent cellular cytotoxicity*. Binding through the Fc receptor on leucocytes can kill the target cell, which has been well described with NK cells and the low-affinity FcgRIII. However, the relative importance of this mechanism in host protection is unknown.

Perhaps the best measure of the importance of antibody-mediated protection against microbes is from naturally occurring deficiency diseases. These include X-linked agammaglobulinaemia, common variable immunodeficiency, and the X-linked hyper-IgM syndrome (Chapter 5.6). These diseases present with common clinical features of early-onset recurrent pulmonary infections with *S. pneumoniae* and *H. influenzae* often complicated by chronic sinusitis and middle-ear infections, with gastrointestinal infections caused by common pathogens, such as *Salmonella* spp., *Shigella* spp., and *Campylobacter* spp., as well as more chronic infection with *Giardia* spp. Increased susceptibility to viral infection is limited to rotavirus and enteroviral infection. The latter are unusual in that they can induce a chronic viral meningitis in the face of immunoglobulin-replacement therapy. Enteroviruses are also associated with an unusual polymyositis syndrome.

Cell-mediated immunity

Specific cell-mediated immunity to infection is mediated by T cells of thymic origin, which recirculate to specific sites because of specific cell-surface receptors. Naïve T cells first encounter antigen that has been processed by professional antigen-presenting cells such as dendritic cells. As has been shown by cannulation of lymphatic vessels draining to and from local lymph nodes following local visna virus infection, these cells pick up antigen locally, are activated, and then migrate to the lymph nodes. Here the cells mature and express the processed antigen in the context of MHC class I and II molecules, as well as numerous accessory molecules (for example, CD40 and B7) required for efficient maturation of specific T cells. Specific T cells leave the lymph node after 4 to 5 days and aggregate at sites of infection and/or inflammation. Here the T cells encounter processed antigen either in infected cells or in local antigen-presenting cells. CD8 T cells may act by directly killing infected cells through perforin or the FAS–FASL (FAS ligand) interaction, with the consequential activation of apoptotic pathways. CD4 and, to a lesser extent, CD8 T

cells release a range of cytokines such as the interleukins, IFN-g, TNF-a and -b, granulocyte–macrophage colony-stimulating factor (**GM-CSF**), etc., which may have a direct anti-infective action but that also allow the development of delayed-type hypersensitivity, particularly important for the control of intracellular bacteria. Cytokines and their specific functions are discussed in Chapter 4.4.

One feature of the specific cell-mediated response is the development of memory T cells. These cells are characterized by the expression of specific cell-surface proteins and their ability to recognize antigen without the need for priming through professional antigen presentation. It has been shown for a number of specific cell-mediated immune responses, notably CD8 T-cell responses to the lymphocytic choriomeningitis virus (**LCMV**), that a circulating pool of memory T cells is established following the first *in vivo* encounter with antigen. Subsequent encounter with the virus results in the more rapid activation from this pool of memory T cells, which are often retained for the lifetime of the host. Numerous studies are currently focusing on the nature of the memory T-cell pool: whether it can be augmented; whether antigen is required to sustain this T-cell population; and whether continuous exposure can result in specific depletion. These issues are of particular importance to the development of effective subunit vaccines, as well as to the development of optimum methods of vaccination for maximum protection especially against weaker antigens.

Mucosal immunity

Since the primary encounter with antigen occurs at mucosal surfaces, the nature of the local response can often determine whether infection is initially established. Various elements of the innate immune response are active at surfaces, including cytokines, complement, lysozyme, lactoferrin, NK cells, and phagocytes, as well as specific immune responses such as secretory IgA (**sIgA**). IgA is secreted at most surfaces (including in colostrum) by the addition of a 'secretory piece' to dimeric IgA. Its secretion follows initial encounter of antigen by a mucosal B cell, which circulates to localize at distant mucosal sites, such as salivary glands. Their mucosal localization is determined by specific receptors to endothelial cells at these locations. This has given rise to the concept of a mucosal, as distinct from a systemic, immune system. This system permits the use of active vaccines to block the entry of pathogens at mucosal sites.

Evasion of the host immune response

To establish effective parasitism, it is essential to bypass the immune response in the host. Several strategies are used, such as the speed of replication—the organism replicates rapidly and spreads to infect the next host before an effective specific immune response can be generated (for example, picornaviruses). If a specific locally active antibody response is generated, then some organisms can bypass it by destroying the antibody through IgA proteases. Alternatively, the antibody may be deflected to components of a pathogen that are irrelevant: for example, failure to develop neutralizing antibodies or even antibodies that may facilitate the entry of the micro-organism to optimum sites of replication as in dengue infection. Some viruses may coat themselves with antibody by expressing Fc receptors to block the development of neutralizing activity.

Similarly, the infected cell can be 'masked' by pathogens so that it becomes invisible to the immune response; for example, DNA viruses, especially herpesviruses, downregulate the expression of MHC class I molecules and block antigen processing/presentation so that specific CD8 cytotoxic T cells cannot recognize the infected cell. Inhibition of cytokine action is similarly employed by viruses, notably vaccines that interfere with cytokine action by virally encoded inhibitors and decoy cytokine receptors (for example, for IL-1, IL-8, TNF).

Alternatively, many pathogens such as *Trypanosoma brucei* vary their antigenicity. The parasite is surrounded by a surface coat of carbohydrate and a 67-kDa glycoprotein. The trypanosome genome consists of a large number of chromosomes, including around 150 mini-chromosomes. The coat glycoprotein is encoded by basic copy genes lying near telomeric sites and also internally within chromosomes, so that more than 200 copies of these genes exist. In humans, trypanosomes express one copy of the protein which is transcribed from a telomeric site. Variation occurs by gene conversion, replacing the gene at the active transcription site by a silent copy from elsewhere in the genome, although additional novel sequences may be created. During infection, changes in the basic copy gene expressed occur spontaneously once in 10 000 organisms, resulting in an antigenic change that can always bypass the host immune response generated. Although the organism commits approximately 10 per cent of its genome to this evasion mechanism, it ensures its long-term survival in the host in the presence of an otherwise effective immune response.

Antigenic variation also occurs outside the host to ensure pathogen survival within the population. This is exemplified by influenza whose mode of entry via haemagglutinin and sialic acid interaction has been described. During virus replication in the host, point mutations are introduced into haemagglutinin which result in the virus bypassing the prevalent immune response to the original infecting virus type. Hence, a small number of hosts will always be susceptible to such changes. As the immune response in the population naturally wanes, the population again becomes susceptible to the parental strain every 4 years, resulting in local outbreaks of influenza. This phenomenon is termed 'antigenic drift'. However, influenza A also demonstrates 'antigenic shift', which is the consequence of the genetic recombination of large segments of the genome in an intermediary host such as birds. If such recombination involves the haemagglutinin gene, against which most neutralizing antibodies are directed, then the host population is severely exposed and global pandemics of influenza A as occurred during 1918, 1957 (Asian 'flu), and 1968 (Hong Kong 'flu) can result.

The ability to maintain a reservoir of pathogen in the face of an active immune response, either in the host or the population, is a prerequisite for the effective survival of many pathogens. Each of the mechanisms briefly considered requires elaborate genomic mechanisms or the utilization of a considerable part of the genome to be committed to this end—for example, a trypanosome utilizes approximately 10 per cent of its genome by encoding multiple copies of the surface glycoprotein, while human cytomegalovirus, in the more restricted genome of viruses, utilizes at least six open-reading frames for genes that interfere with MHC class I antigen presentation in the infected cell. Such genomic commitment, of itself, recognizes the importance of evading immunity for parasite survival.

Pathogen persistence and latency

Pathogens, especially DNA herpesviruses, have evolved a virus cycle that enables them to remain dormant at anatomical sites with little virus gene expression, thereby minimizing the potential for immune activation. Following primary infection with herpes simplex, multiple cutaneous lesions develop and virus is shed. An effective immune response develops that clears most of the infectious virus, and the skin lesions heal. However, virus passes to nerve ganglia where it enters a latent state. Periodically it can be triggered to reactivate; infectious particles track down axons with epithelial reinfection and temporary shedding until the immune response again clears the cutaneous site. Therefore a virus–host equilibrium is established, although the relationship is never symbiotic. Not only can activation of the virus by non-specific triggers alter the equilibrium, but immunosuppression can result in the more widespread peripheral dissemination of reactivation. The virus gains two major advantages from this virus cycle: evasion of the immune response and a host reservoir that enables it to bypass a host generation for horizontal transmission, thereby sustaining itself in small isolated populations in the face of an active immune response.

Immune-mediated injury

In most instances the development of an active immune response is beneficial to the host and results in pathogen clearance or the establishment of persistent or latent infection, in which a host–parasite equilibrium is established that minimizes damage to the host. Occasionally, the infectious agent may not itself be as damaging to the host as the immune response that is generated against it.

Immunopathology

At the simplest level, many clinical features of infection are the result of the host's response to the infectious agent; fever is caused by the release of cytokines to the infectious agent. However, in some circumstances the phenotype of clinical disease is altered depending on the nature of the immune response. *Mycobacterium leprae* infection results in a spectrum of disease phenotype with lepromatous and tuberculoid leprosy at either extreme. In the former, the host is relatively 'anergic' to the infectious agent or products of *M. leprae*, with the result that the clinical lesions are less physically destructive but contain high levels of viable organisms, allowing easier spread of infection. In the tuberculoid form, the cell-mediated response to *M. leprae* is strong, with local mononuclear infiltrates at sites of infection and few organisms but more intense tissue destruction and local injury.

Similarly, respiratory syncytial virus infection can result in bronchiolitis in young infants. An immunopathological explanation for this condition was suggested following a failed vaccination campaign using an alum-precipitated killed virus, which resulted in vaccinated children suffering worse disease. The mechanism was thought to be a failure to induce neutralizing antibody, but a T_H2 response capable of releasing IL-3, IL-4, and IL-5 was generated. On infection, release of these cytokines resulted in bronchospasm and local inflammation with eosinophil infiltration. Similar changes can be observed in experimental murine infection.

Autoimmunity

If immunopathology can be induced by an infecting agent through the generation of an aberrant or misdirected immune response, the question arises of whether such a response could be sustained and result in end-organ damage as in autoimmune disease. While attractive as a hypothesis, the 'hit and run' nature of infectious triggers for common autoimmune disorders is difficult to establish. However, in the experimental setting autoimmunity is induced using bacterially derived adjuvants in genetically susceptible hosts. In such circumstances the infectious agent may break tolerance, although direct evidence for this mechanism in humans is lacking.

Alternative mechanisms have been proposed that include molecular mimicry, in which an immune response is generated against an antigen from an infectious agent to trigger a crossreactive response to a host-cell protein, due to its structural similarity. Experimentally transgenic mice expressing the LCMV nuclear protein under an insulin promoter (such that it is expressed only in b cells in the Islets of Langerhans) do not generate a response against the protein and do not develop diabetes. However, if the mouse is infected with the virus then the CD8 T cells generated will destroy the islets and render the mouse diabetic. Evidence for this phenomenon in humans is lacking; infection with *Streptococcus* spp. elicits crossreactive antibodies resulting in rheumatic fever, but even in this case there is little evidence that the response can be sustained. Similarly, the demonstration of prior chlamydial infection in the identification of a specific MHC type and late Reiter's syndrome with reactive arthritis is suggestive but not conclusive.

Host susceptibility to infection

In addition to specific and non-specific immune mechanisms, there are several other host factors that mediate susceptibility to infection.

Genetic factors

The overall impact of genetic factors in host susceptibility to infection is difficult to estimate. Perhaps the best evidence comes from twin studies—susceptibility to infections ranging from malaria to *Helicobacter pylori* are inconclusive, with associations being established for disease severity rather than a global susceptibility to an infectious agent. However, a Swedish study of adoptees reported a sixfold increased risk of an infectious cause of death where a biological, as opposed to an adoptive, parent had also died from an infectious illness. Alternatively, the role of candidate genes in specific infections has been suggested.

MHC and infection

The central position of the MHC in regulating specific immunity make this locus an obvious candidate. However, studies have required large numbers of subjects because this locus is highly variable. Perhaps the best known association is the linkage of HLA-B*5301 with resistance to severe malaria in children in The Gambia. This study was given additional impetus with the identification of a malaria-derived peptide that shows variability and can be presented by HLA-B*5301. An MHC class II locus *HLA-DRB1*1302* was also associated with disease resistance. Taken together with the observation that the HLA-B53 association did not pertain in East African populations, this suggests that the association is complex and that there could be geographical variation in the pathogen which could influence susceptibility to infection by affecting the presentation of microbial peptides by different MHC molecules.

Other associations have been described including: HLA-DQ with cervical cancer associated with human papillomavirus infection (**HPV**); and HLA-DRB1 with the clearance of hepatitis B infection. Numerous studies have been performed to try to establish an association of resistance or disease progression with HIV infection but none are conclusive to date. The subtleties of such associations are exemplified by the association of HPV infection and cervical cancer. While an MHC class II association has been recognized, differences in the survival of patients with invasive cervical cancer have been linked to HLA-B7. This is not evident when all patients with HLA-B7 who have the disease are studied. However, in cervical cancer, there is selective loss of expression of individual MHC class I alleles on the surface of tumour cells. In this instance, patients who lost only HLA-B7 expression on tumour cells had a reduced survival rate. This could be the result of tumour-specific MHC class I-restricted immune responses, but whether the restriction affects the initial infection with HPV or the maintenance of the oncogenic process, is unknown.

The recognition that infections are associated with specific antigens in the host's MHC is important. With improved knowledge of the regulation of cellular immunity and an ability to generate specific immunity with peptides that bind to appropriate MHC molecules, it is possible to develop a new range of vaccines for the treatment and prevention of a range of infectious diseases. However, the vaccine-based approach must take into account possible protective and adverse associations to ensure that the real risk of immunopathology is minimized.

Other genetic factors

A range of other candidate genes have been identified that predispose to specific susceptibility. Blood groups are well described; peptic ulceration is associated with blood-group secretor status, which may confer susceptibility to *H. pylori* infection. Mutations in haemoglobin, particularly in heterozygotes, results in resistance to malaria, which may allow the persistence in the population of what would otherwise be a detrimental defect. Several authors have reported alterations in receptor and chemokine and cytokine genes and susceptibility to infection: elevated TNF concentrations consequent on a mutation in the promoter of TNF-a have been associated with cerebral malaria in Africa, mucocutaneous leishmaniasis in South America, and lepromatous leprosy in India. Mutations in chemokine CCR5 receptors are associated with a reduced susceptibility to HIV infection. Other molecules that may be linked to immune or inflammatory responses are also associated with susceptibility to infection, for example mannose-binding lectin may be associated with invasive disease by encapsulated bacteria, although enhanced susceptibility to individual pathogens has not been demonstrated. The natural resistance-associated macrophage protein type 1 (**NRAMP1**) protein is found on macrophages. The gene encoding NRAMP1 is homologous to a murine gene that has been associated with a susceptibility to leishmania and salmonella infection in defined strains, but mutations in the human gene product are associated with severe pulmonary tuberculosis. Mutations in the human vitamin D receptor also appear to be associated with severe *Mycobacterium tuberculosis* infections.

Conclusions

Despite a number of interesting associations with widely disparate molecules, it is difficult to draw conclusions with respect to pathogenesis at this stage. As the most intriguing associations are with the MHC complex, it is tempting to infer that microbial infections are a powerful evolutionary driver to polymorphism at these loci. Although the specific examples described here are intriguing, it is likely that the factors that determine susceptibility and resistance to individual pathogens in the general population are highly polygenic.

Environmental and intercurrent susceptibility to infection

Several additional factors that operate in a patient have to be considered as influencing susceptibility, severity, and the likelihood of recovery from infection.

Age

Common infectious diseases occur more frequently in older patients: for example, pneumonia (twofold), bacteraemia (threefold), urinary tract infections (fivefold), reactivation of varicella-zoster virus (**VZV**), and tuberculosis (**TB**). The explanation is frequently given that cell-mediated immunity declines with age and that this is reflected in the increased incidence. While this may be the case with VZV and TB, it is difficult to reconcile this alongside other common physical problems of ageing, such as urinary obstruction due to prostatic hypertrophy, and confounding variables such as relative malnutrition, social isolation, and socioeconomic deprivation.

Hormonal influences

Hormones can affect immunity and impact on infections. 'Stress' is often cited as an important factor in the severity of infection. Chronic stress and 'life events' are strongly associated with the increased duration of virus shedding and the intensity of symptoms in upper respiratory tract infections. Pregnancy is associated with an increased severity of poliomyelitis and influenza (among other viral infections) during the third trimester.

Malnutrition

Worldwide, calorie malnutrition is the most common cause of increased susceptibility to infections such as bacterial septicaemia, middle-ear infection, dental caries,

and perioral infection. Enhanced severity of measles has been widely reported, although the effect of overcrowding, which frequently accompanies malnutrition in developing countries, may also contribute. However, improvement in diet has historically been associated with reduced mortality from several infectious diseases. Specific nutrients are difficult to identify, although the best evidence supports a role for vitamin A; deficiency of this vitamin in pregnancy has been associated with enhanced vaginal HIV shedding and increased perinatal transmission.

Intercurrent illness and infections

Many severe illnesses, including other infections, can impair host immunity. Although the obvious example is HIV infection, impairment can also be associated with many conditions such as alcoholism, liver failure, renal failure, and late-stage cancer. Some specific examples are also well documented: measles infection is associated with the loss of delayed-type hypersensitivity in the Mantoux response, with the possibility that tuberculosis may be exacerbated.

Therapy

There are obvious examples of immunosuppressive drugs that render the host susceptible to numerous opportunistic infections.

Altering the host response to prevent and treat infection

Vaccination

The decline in common infectious diseases throughout the last 150 years can be attributed to improved public health measures (sanitation and hygiene) and the development of cheap and effective vaccines. The success of mass vaccination was highlighted by the eradication of smallpox in 1980, and may soon be followed by the eradication of poliomyelitis. However, the vaccination programme has also controlled important childhood infections such as measles, mumps, rubella, diphtheria, haemophilus meningitis, etc. The development of these vaccines has relied primarily on empiricism based on the principle of producing a mild form of the illness, either by using a low infecting dose or an attenuated live agent. Killed vaccines are effective where the appropriate immunity can be generated to a pathogenic toxin, for example antitetanus toxoid.

With a greater understanding of the nature of immunity induced by vaccination, a rational basis exists to pursue the more difficult pathogens that remain major targets for vaccine development, including: malaria, schistosomiasis, hookworm, tuberculosis, infantile diarrhoea, pneumonia, and HIV infection.

A successful vaccine not only has to induce effective immunity, but it must also be acceptable to the public. Such acceptability is difficult to gauge and varies from one community to another. Public acceptance may be partly based on a relative-risk perception between the hazards of vaccination and the burden of morbidity and mortality. There are several examples of difficulties in implementing vaccination programmes such as the *Bordetella pertussis* vaccine and the recent difficulties of sustaining high levels of immunity against measles due to a suggested link between autism and the triple vaccine. Failure to sustain high levels of herd immunity result in resurgent epidemics of diseases previously held under control, for example the diphtheria outbreaks in the former Soviet Union. In addition, for worldwide use a vaccine has to be stable, cheap, and easy to administer. Thus the development of a theoretically effective immunogen may not necessarily translate directly to a successful vaccine for clinical use.

None the less, improvements in technology utilizing the nature of T- and B-cell interaction have seen the development of cellular conjugate vaccines such those undergoing trial against *S. pneumoniae*. Adjuvant technology is also improving to enhance immune responsiveness, and cytokines have been shown to enable vaccine non-responders to generate effective immunity to hepatitis B. Furthermore, DNA vaccines are being produced and undergoing efficacy testing alongside appropriate delivery systems; such as prime–boost immunization—where two different vaccines are administered in sequence: for instance DNA vaccination followed by a booster of the same agent in another recombinant form or the pure cognate protein. Similarly, specific antigens are being delivered to different anatomical sites to enhance mucosal immunity utilizing isolated properties of pathogens, for example the fimbrial proteins of *S. typhimurium*.

This expansion in the utilization of new methods for vaccine development has raised our expectations of tackling difficult pathogens—especially those where a persistent or latent state can be established in the host. There is also the possibility that vaccines may be able to alter the immune response against established infections to control clinical disease. In infections such as leprosy, leishmaniasis, and schistosomiasis, cytokine therapy may modify the effects of the infection, especially where the disease itself is mediated by immune events. Some viruses that persist or even transform cells may continue to express a range of proteins in the transformed state; these products too may serve as appropriate targets for eliminating the infected or transformed cell.

Passive immunotherapy

Perhaps one area that has still not seen the full exploitation of its potential is passive immunotherapy. Passive antibody is used to prevent disseminated varicella in immunocompromised subjects and in the treatment of tetanus; it is also used as an adjunct to the postexposure prophylaxis of rabies. The ability to modify heterologous antibodies to render them less immunogenic, alongside the ability to select the antibody repertoire *in vitro*, makes passive immunotherapy an attractive option for the control of established infection or for modifying immunopathology. In addition, the adoptive transfer of specific cell-mediated immunity in specific circumstances, such as following bone marrow transplantation to prevent human cytomegalovirus pneumonia, may be coupled with the modification of cellular responses including the repertoire of T-cell receptors. Currently, these approaches are still experimental but they may soon be explored in clinical trials.

Conclusions

Although there have striking advances have been made in the control of infectious disease by modifying host responses, many challenges remain. Common childhood infections pose major difficulties for healthcare in many parts of the world, and here the availability of effective prophylaxis must engender greater international effort for its effective implementation. Finally, a greater understanding of the nature of the host–parasite relationship will, coupled with increased knowledge of immune mechanisms, be required to develop new methods to prevent and treat the more intractable common infections.

Further reading

Costerton JW, *et al*. (1995). Microbial biofilms. *Annual Review of Microbiology* **49**, 711–45. [Review of formation and factors involved in biofilm formation.]

Cotter PA, Miller JF (1998). *In vivo* and *ex vivo* regulation of bacterial virulence gene expression. *Current Opinion in Microbiology* **1**, 17–26. [Description of regulation of the control of virulence island genes.]

Gander S (1996). Bacterial biofilms: resistance to antimicrobial agents. *Journal of Antimicrobial Chemotherapy* **37**, 1047–50. [Antibiotic resistance in biofilms.]

Hill AVS (1998). The immunogenetics of human infectious diseases. *Annual Review of Immunology* **16**, 593–617.

Mackowiak PA (1998). Concepts of fever. *Archives of Internal Medicine* **158**, 1870–81. [Discussion of major factors in the regulation of temperature.]

Mims C, Nash A, Stephen J (2001). *Mims' pathogenesis of infectious disease*, 5th edn. Academic Press, San Diego. [A full and readable account of major factors involved in microbial pathogenesis.]

Nicholson KG, Webster RG, Hay AJ (1998). *Textbook of influenza*. Blackwell Science, Oxford. [A detailed account of the virology and pathogenesis of influenza infection in man and associated species, including a good section on the development and problems of influenza vaccines.]

Plotkin SA, Orenstein WA (1999). *Vaccines*, 3rd edn. WB Saunders, Philadelphia. [The standard reference work to currently used vaccines, future developments, as well as regulatory and delivery issues.]

Rosen FS, Cooper MD, Chapel HM (1995). The primary immunodeficiencies. *New England Journal of Medicine* **333**, 431–40.

Rossman MG (1989). Neutralisation of small RNA viruses by antibodies and antiviral agents. *FASEB Journal* **3**, 2335–43. [A good discussion of the mechanisms of virus neutralization and drugs that block entry of picornaviruses.]

Sorenson TI, *et al.* (1988). Genetic and environmental influences on premature death in adult adoptees. *New England Journal of Medicine* **318**, 727–32. [Suggestion of global genetic susceptibility to

infection.]

'Vaccines and immunology' (2001). *Science* **293**, 233–56. [A series of articles defining the current issues in immunology impacting on future vaccine development.]

Van der Woude M, Braaten B, Low D (1996). Epigenetic phase variation of the *pap* operon in *Escherichia coli*. *Trends in Microbiology* **4**, 5–9. [Discussion of the regulation of expression of pathogenicity-associated fimbriae in *E. coli*.]

Ziegler E, *et al.* (1991). Treatment of Gram-negative bacteraemia and septic shock with HA-1A human monoclonal antibody against endotoxin: a randomised, double blind, placebo-controlled trial. *New England Journal of Medicine* **324**, 429–36. [Failure of monoclonal antibody immunotherapy to protect against bacteraemia.]

7.5 Physiological changes in infected patients

P. A. Murphy

Further reading

All of us are exposed to potentially lethal infectious agents from shortly after birth until we die. And while there are viral, bacterial, fungal, and parasitic infections, there is no doubt that the most dangerous organisms are bacteria. Bacteria in logarithmic phase divide every 15 to 20 min: no eukaryotic cell can match that rate, nor can any virus, fungus, or parasite. The immune system evolved primarily to deal efficiently with bacterial infections. It is true that the immune system has minor effects on some tumours, but if mice with severe inherited immune defects are maintained in a sterile environment they have a normal lifespan and do not have an unusual incidence of cancer. Out of the plastic bag, they last a week or two before dying of overwhelming infection.

The immune system is divided into the natural immune system and the acquired immune system. The natural immune system is a series of defences against bacteria and other infections which is present from birth. None of its mechanisms depend on previous experience with the organism, and they do not improve with experience. It is so efficient that the overwhelming majority of bacteria are destroyed shortly after entry into tissue without eliciting any noticeable host response. There are responses, but they function at a microscopic, sometimes a molecular, level. The only infectious diseases of which we become aware are those in which the natural immune system has failed, the population of invading organisms has become large, and inflammatory reactions become sufficiently pronounced to cause symptoms that we can feel and signs that we can see. A healthy adult may go months or years between clinically apparent infections.

Acquired immunity is called on when the initial battle has gone badly and bacterial populations have become large. Initiating an immune response of any kind usually requires antigen concentrations in the microgram range. Since a bacterium weighs about a picogram, specific acquired immune responses are only likely to occur when bacterial populations are measured in millions or billions. Organisms which can multiply to these levels in healthy people generally have some feature, such as a capsule, which enables them partially to evade the natural immune response. Small numbers of B and T lymphocytes which can respond to essentially any immunodeterminant exist in everyone. Uncontrolled infection leads to rapid clonal expansion of specific B and T cells which mediate immunity to the pathogen, most usually by antibody formation. Antigen is maintained in tissue depots so that antibody secretion is prolonged long after the initiating infection has resolved. Both B and T cells generate long-lived memory cells which can initiate secondary immune responses if the organism is encountered again. Secondary immune responses require ten to a hundred times less antigen to elicit them, and occur faster than do primary immune responses. The increased efficiency means that many clinical infections occur only once. Lifelong immunity is often, but not always, maintained by periodic asymptomatic rechallenge with virulent organisms.

Bacteria and fungi are recognized as foreign by the natural immune system. Both classes of organism have thick cell walls which have no counterpart in mammalian cells. The recognition molecule is C3; one important foreign structure is a hydroxyl group on each of two adjacent carbon atoms, which is found in most sugar residues. C3 has a very low but finite rate of spontaneous activation to C3a and nascent C3b. Nascent C3b has an unstable thioether bond which has about 60 μs to find a carbohydrate before it reacts with water and is inactivated. Binding to carbohydrate results in covalent attachment of a molecule of C3b to the bacterial or fungal surface. This is the primary activator of the alternative pathway of complement fixation, which causes fixation of C3b to the foreign surface in amounts which are opsonic for phagocytic cells. Mammalian glycoproteins have carbohydrate groups but many of the hydroxyl residues on terminal sugars are oxidized to carbonyl groups which do not react with C3b.

There are whole families of proteins, both soluble and cell bound, which recognize specific sugar residues such as mannose. Since many fungi have cell walls which are largely polymannose, some of these are likely to be important for antifungal defence.

Bacteria are prokaryotic and present many more foreign labels than do fungi. Probably most important are the major cell wall constituents, endotoxin in Gram-negative organisms and peptidoglycan in Gram-positive organisms. These are recognized as foreign by many systems, both soluble proteins such as lysozyme and cell bound proteins. The key cell in initial defence against organisms is clearly the macrophage. Every tissue contains macrophages which are often organized anatomically in ways which facilitate defence against infection. Alveolar macrophages patrol the alveoli, Kupffer cells have processes which stretch across hepatic sinusoids, and splenic macrophages inspect the blood which filters past them in the red pulp. Most normal tissues contain few or no polymorphonuclear leucocytes. CD14 protein is exposed on macrophage membranes. It contains two high-affinity binding sites, one for endotoxin and one for peptidoglycan. These sites are not identical, but they are so close together that monoclonal antibodies exist which can block binding of both ligands. This single protein is a bacteria detector, and binding of its ligand activates the macrophage to secrete interleukins IL-8 and IL-1 and tumour necrosis factor. IL-1 and tumour necrosis factor act on capillary endothelium to initiate the cascades which result in increased permeability and display of receptors for polymorphonuclear leucocytes. IL-8 is powerfully chemotactic for polymorphonuclear leucocytes. Thus recognition of bacteria by the CD14 protein initiates the natural immune response.

Additional properties of bacteria which initiate inflammation are their ability to fix complement and the fact that bacterial proteins start with an *N*-formyl methionine residue rather than the methionine employed by eukaryotic cells. Almost all bacteria fix complement in the cell wall, though pathogenic species may fix complement in deep layers which are not accessible to polymorphonuclear leucocytes. The soluble complement component C5a is strongly chemotactic for polymorphonuclear leucocytes and even though the bacteria are not opsonized, if polymorphonuclear leucocytes are attracted to their vicinity the bacteria may be destroyed by surface phagocytosis. Similarly, there is a high-affinity receptor for *N*-formyl methionyl peptides in the membranes of polymorphonuclear leucocytes; binding to this receptor initiates a chemotactic response which may be useful in the same way.

Some bacteria, such as mycobacteria, and some fungi, such as *Histoplasma capsulatum*, can evade the natural immune response and the antibody mediated arm of the acquired immune response. These infections tend to be chronic, asymptomatic at first, but progress to extensive destructive reactions such as caseous necrotic cavitation and scarring. The destructive reactions are mediated by macrophages highly activated by parasite-specific T cells.

In general, viruses do not contain components which are directly inflammatory, but typically produce inflammation by killing cells. Some viral infections are not cytolytic; symptoms occur only when the acquired immune response is activated and large populations of virus-specific T cells have been generated. When these T cells kill cells infected by virus, inflammation is induced. In measles, cells can be replaced and the rash signals the end of the disease. If the cells are not replaceable (for example neurones) the immune response may be lethal.

Parasites which live in tissue are almost never directly inflammatory, and they invariably have some mechanism for suppressing or evading the acquired immune response.

A few organisms have specific properties which determine most of the features of the illness. Sore throat due to diphtheria bacilli would be inconsequential if it were not for the toxin which binds to cardiac muscle cells, stops protein synthesis, and leads to heart failure. Many diarrhoeal syndromes are mediated by bacterial toxins, and the absence of toxin means absence of disease. However, even where there are distinctive features some symptoms are shared with other infections—there is almost always fever and inflamed areas hurt.

The general characteristics of infections are explained by the features discussed above. The vast majority of infections are dealt with at a microscopic level by the natural immune system and never give rise to any clinical symptoms. The relative importance of natural and acquired immunity is given by observations on untreated individuals born with congenital defects of one or other system. A child born with no B cells will generally die of infection at about the age of 6, a child born with no T cells will generally die at 2 or 3. A child born with no neutrophils will last a few weeks, and no child with congenital absence of macrophages has ever been described. Admittedly, that may have more to do with the role of macrophages in remodelling tissues during fetal development than with resistance to infection.

Bacterial infections are characterized by acute onset, intense local general inflammation, rapid progression to severe local and general symptoms, and resolution towards death or recovery in a week or two. Infections by slowly dividing organisms like mycobacteria are characterized by less initial inflammation but deterioration over months or years. Fungal infections are far less inflammatory than bacterial ones, though there may be intense inflammation in the later course due to acquired immunity. Viral infections are not usually serious unless, as in poliomyelitis, the virus grows in a population of cells which cannot be replaced. Patients with fungal and viral infections may be highly febrile, but are-notably less generally ill than those with bacterial infections, because the infecting organisms are less inflammatory.

Whether or not an infection becomes established depends on the balance between the host's ability to mobilize an adequate inflammatory response and the parasite's ability to evade that response. Numbers are crucial: it is highly improbable that one organism by itself can initiate a serious or lethal infection. There is good evidence that many infections begin with the survival of a single organism from the initial inoculum, and that all the myriad of organisms which eventually overwhelm the patient are descendants of that one bacterium. However, the probability of that happening is so low that it is almost never observed. As a practical matter, most experimental infections in small groups of animals are initiated by inocula of many millions of organisms so that the individual tiny probabilities are multiplied to a level which can be observed. Spontaneous infections in people are ordinarily initiated by small inocula, but there are hundreds of millions of people. Even though serious infection is improbable, the number of people at risk is so huge that some serious infections are observed.

If there is a defect in host defences, the probability of a serious infection initiating from a small inoculum is greatly increased. Worldwide, the most common cause of poor immune performance is malnutrition, especially lack of adequate protein. Every component of the natural and acquired immune system depends on proteins, and if the building blocks are not available these components cannot be made. Premature infants and elderly patients have serious defects of T-cell function; in one case the system has not completely developed and in the other it has atrophied. There are also numerous acquired immune defects in people who have had transplants, been given chemotherapy for malignancy or immunosuppressive treatment for diseases such as lupus erythematosus and rheumatoid arthritis, and infected with HIV. All these people frequently develop infections with organisms which rarely or never trouble normal people because they cannot control bacterial populations at a low level.

Large numbers of bacteria are inherently inflammatory. If bacterial populations reach hundreds of millions, inflammation will occur whether the bacteria are alive or dead. It makes no difference whether or not the species is 'pathogenic'; the response is to the cell wall components which are found in all bacteria. Pathogenic species possess some attribute such as a capsule or a leucocidin which enables them to evade the initial natural immune response and to multiply in tissue. Multiplication to large numbers does several things. Local inflammation is induced wherever the organisms are, and one sees a clinical sore throat, a boil, pneumonia, or a urinary tract infection. Initially mediators of inflammation leak out of the inflamed area into the circulation and induce a systemic inflammatory response. The systemic inflammatory response includes new aspects of protection by the natural immune response. Generally speaking, bacteria in numbers which induce a systemic inflammatory response also initiate acquired immune responses. Bacteria may invade the bloodstream from the local lesion; they may be able to grow in the bloodstream (sepsis), or they may set up distant foci of infection such as endocarditis, a septic joint, or a splenic abscess. Finally, certain aspects of the systemic inflammatory response are deleterious and may progress to circulatory failure (septic shock) or multiple organ failure.

The concept that large numbers of essentially any bacterium in tissue will cause serious or fatal inflammatory reactions explains a number of clinical situations: aspiration of saliva containing 10^8 bacteria/ml causes pneumonia; a burst appendix causes peritonitis. Infusion of infected fluids intravenously has caused fatal shock in otherwise healthy people. Most of the fatal cases have been due to Gram-negative species; if the organism is Gram positive high fever is common but death is rare. This corresponds to the fact that endotoxin is more inflammatory than peptidoglycan by a factor of 10 to 1000, depending on the test used. In all these cases, most of the bacteria are of species which are ordinarily not pathogenic, This is particularly clear with infected blood transfusions, where the selective factor is the ability to grow in a refrigerator at 4 °C, not the ability to evade the immune response.

It is possible to explain, at least in general terms, almost all the changes which we see in infected patients. The most common reaction to infection in normal people is to develop no clinical symptoms. However, a great deal is going on unobserved. The natural immune system controls and eliminates most invading pathogens. However, the natural immune system has as its key cell the macrophage, which is also a key cell for the acquired immune response. Macrophages display pieces of dismembered bacteria on their cell membranes, in the context of major histocompatibility determinant and IL-1. This is recognized by T cells, which help B cells to make antibodies. Only about 1 per cent of healthy people can remember an attack of pneumococcal pneumonia, but every healthy adult has serum antibodies to 30 or 40 pneumococci, as well as antibodies to common bacteria of most other species. Most infections result in symptomless seroconversion.

If the infection is not controlled at a low population, local inflammation will develop. The classical signs of local inflammation are redness, warmth, swelling, pain, and loss of function. As mentioned above, inflammation is initiated by macrophages which secrete tumour necrosis factor-a, IL-1, IL-6, and IL-8 in response to sensing bacteria through the CD14 protein. Macrophages secrete a powerful thromboplastin which initiates thrombosis and fibrinolysis. Local vasodilatation (redness) is caused by many molecules: in no particular order some of them are histamine from local mast cells, several arachidonic acid derivatives from the membrane phospholipids of activated cells of many types, kinins from serum kininogens activated by Hageman factor exposed to endotoxin, and nitric oxide generated by capillary endothelial cells. There are two distinct waves of nitric oxide synthesis, both induced by IL-1 and tumour necrosis factor acting synergistically. In the first few hours preformed enzyme is activated. After about 12 h newly synthesized enzyme becomes most important.

Local warmth (heat) occurs because of increased blood flow, which raises the temperature of the skin closer to aortic blood temperature. In the fingers this is very obvious because the temperature of the skin of the hand is normally about 30 °C. However, even shoulder skin normally has a temperature of about 35 °C and covered skin such as that on the stomach is still a degree or so below central temperature. The increased rate of local metabolism must contribute something to local warmth, but the effect is thought to be trivial.

Swelling occurs because of local accumulation of extravascular fluid. Normally the amount of albumin in blood is just sufficient to pull back into the low-pressure venous end of the capillaries almost all the fluid which passes into the extracellular space under the higher pressure of the arterial end of the capillary. The rate of lymph flow from a resting non-inflamed tissue is almost zero, and the albumin content is less than 0.5 per cent. Inflamed capillary endothelium is leaky; initially IL-1 and tumour necrosis factor synergize to cause retraction of the cell edges so that gaps develop between adjacent cells. Later, polymorphonuclear leucocytes bound to integrins and palisaded along the endothelial cells may degranulate directly onto the endothelial cell surface and may cause further endothelial cell damage due to lipid peroxidation by generation of reactive oxygen derivatives such as superoxide and hydroxyl radicals. Eventually, there is widespread death of endothelial cells, some of which seems to be due to apoptosis. The increased capillary permeability is of extraordinarily rapid onset—1 h after injecting Gram-negative bacteria into the veins of a dog, pulmonary lymph flow had quadrupled, and the albumin content of pulmonary lymph was over 2 per cent.

Pain appears to be relatively simple. In experiments in which various inflammatory mediators were dropped onto denuded blisters in people, bradykinin was by far the most effective at eliciting pain.

Pus forms when so many polymorphonuclear leucocytes accumulate locally that their secreted proteases are able to overwhelm the local concentrations of antiprotease control proteins such as a_1-antitrypsin. Pus has a thick consistency because it contains large amounts of DNA from the nuclei of broken down polymorphonuclear leucocytes, and it is sometimes greenish because of the presence of large quantities of myeloperoxidase. Pus formation causes local tissue destruction, and, although bacteria do not grow well in it, spontaneous or surgical evacuation of pus often brings an infection to an end.

In serious infections, cytokines such as IL-1, IL-6, and tumour necrosis factor leak out of the local lesion and cause a generalized inflammatory response. This is characterized by fever, polymorphonuclear leucocyte lecocytosis, a striking lowering of serum iron and zinc, and major changes in hepatic protein synthesis. All these changes are adaptive, systemic aspects of the natural immune response which often enable survival in serious infections for long enough to allow production of specific antibody so that the host can recover. A most striking example of this was sometimes found in untreated pneumococcal pneumonia, in which the patient was desperately ill for 6 or 7 days and then, if lucky, had an almost miraculous recovery in a 24-h period. This crisis corresponded to the appearance of free antipneumococcal antibody in the serum, indicating that enough antibody had been made to opsonize all the pneumococci and leave some unbound antibody in the serum.

Fever is due to resetting of the anterior hypothalamic thermostat by IL-1, IL-6, and tumour necrosis factor. The body temperature is controlled just as precisely as it is in health, but the set point is higher. These three cytokines probably work by inducing hypothalamic synthesis of prostaglandin E_2 or a similar compound; aspirin and other non-steroidal anti-inflammatory drugs are antipyretic because they inhibit prostaglandin synthesis. The adaptive value of fever in infectious illness appears to be that it potentiates the immune response.

Polymorphonuclear leucocyte leucocytosis is largely induced by IL-1; initially they are mobilized from the reserve pool of mature polymorphonuclear leucocytes in bone marrow. In a healthy person, this reserve contains about 100 times more polymorphonuclear leucocytes than there are in the blood. Blood smears show more polymorphonuclear leucocytes than average with an increased proportion of band forms. In serious infections, the mature pool of polymorphonuclear leucocytes in the bone marrow is exhausted, and production of polymorphonuclear leucocytes is stimulated from earlier levels. Development is hurried and polymorphonuclear leucocytes are allowed into blood with markers of rushed development. 'Toxic granules' are large, blue perfectly normal polymorphonuclear leucocyte tysosomal granules; the cell did not have time to make the specific granules which would ordinarily have covered them up. 'Dohle bodies' are bluish patches in the cytoplasm of a polymorphonuclear leucocyte; they are pieces of endoplasmic reticulum which the cell did not have time to discard. Vacuoles are not markers of immaturity:

polymorphonuclear leucocytes with vacuoles are cells which have had a 'near miss' with bacteria and have partially degranulated into their own cytoplasm. Vacuoles are highly correlated with the presence of bacteraemia. The adaptive value of mobilization of polymorphonuclear leucocytes is obvious.

The specific granules of polymorphonuclear leucocytes contain lactoferrin an iron binding protein which has a dissociation constant two orders of magnitude less than that of transferrin, the normal serum iron transport protein. Polymorphonuclear leucocytes exposed to bacteria are 'messy feeders' and much of their lactoferrin is released into the surrounding fluid. Lactoferrin finds its way into the circulation, strips iron off transferrin, and the lactoferrin–iron complex is taken up and stored in the reticuloendothelial system. Iron is not available for haemoglobin synthesis, and so the 'anaemia of chronic disease' develops in prolonged infections. More important, iron is not available for bacterial growth. There are many organisms which become highly virulent in iron-overloaded individuals. An example is *Vibrio vulnificus* septicaemia, which is very rare. More than half of the published cases have occurred in people with haemochromatosis.

Normally, the liver makes mostly albumin. It also makes smaller amounts of plasma proteins such as complement components and coagulation cascade proteins, both participating proteins such as fibrinogen and control proteins such as protein C. During serious infections, there is a switch. Little or no albumin is made, and the production of complement and coagulation proteins is increased. Serious infections in individuals with extensive atherosclerosis are often punctuated or terminated by thrombotic events such as strokes and myocardial infarctions. The increased erythrocyte sedimentation rate found in acute infections is mostly due to increased plasma fibrinogen. The adaptive significance of increased levels of complement is obvious; the adaptive significance of faster blood clotting is not so obvious. In addition, the liver makes huge quantities of two proteins, C reactive protein and serum amyloid associated protein, which are not found at all in normal serum. C reactive protein is a phosphoryl choline binding protein which binds to the cell walls of many bacteria. Once bound, it fixes complement, and the bacterium becomes opsonized. Serum amyloid associated protein behaves in a similar way. These proteins act in a non-specific and low-affinity manner, and the protection they provide against any one organism is far inferior to that provided by specific antibody. But their non-specificity is their virtue: they provide low-grade protection against a wide variety of bacteria, and they are available in quantity early in infection. In 24 h, the liver can synthesize 15 g of C reactive protein. The main inducer of changes in liver protein synthesis is IL-6, with assistance from IL-1.

If infection becomes overwhelming, especially if bacteraemia is present, a generalized inflammation of capillaries occurs which is clearly maladaptive. The inducers of the sepsis syndrome are the same as those which induce local inflammatory changes; sepsis is simply inflammation writ large and affecting every tissue in the body. The shock of sepsis is due to peripheral circulatory failure—there is generalized vasodilation in all the vascular beds of the body. Patients with normal hearts attempt to compensate by tachycardia and increased left ventricular stroke volume, The cardiac output in a young man may be 15 to 20 litre/min, with a blood pressure of 80/50 and a systemic vascular resistance of about 4×10^{7} N s/m^{-5}. This pattern is completely different from either shock due to left ventricular failure or shock due to major blood or fluid loss. In both of these, the blood pressure may be identical, but the cardiac output is very low, and the peripheral resistance is high. The circulatory pattern is so reliable that the diagnosis of sepsis can be made with virtual certainty if it is present. The only real differential diagnosis is anaphylaxis, which can generally be ruled out on the history.

The capillaries become leaky, and there is a steady loss of fluid and albumin from the blood into the extravascular space ('third spacing'). Unless fluid is replaced, the contracting intravascular volume exaggerates the shock state because the heart cannot maintain a high cardiac output and cardiac output eventually falls below normal. By this time the patient is well into multiple organ failure—the lungs are heavy and waterlogged and PO_2 plunges to levels which require supplemental oxygen. Eventually, the lungs become so stiff that the patient cannot sustain the work of breathing and needs intubation. Even then very high inflation pressures, high positive end expiratory pressure, and high concentrations of inspired oxygen may be necessary. This is full-blown adult respiratory distress syndrome and carries a 50 per cent mortality because the measures required to keep the patient alive themselves induce progressive pulmonary damage.

Other organs react to the anoxia in their own ways. Delirium is an early sign of sepsis: neurones are adapted to a lower pH and slightly different electrolyte concentrations from those found in serum. When the blood–brain barrier becomes leaky, the concentrations of constituents of the cerebrospinal fluid approach those in serum and there is clouding of consciousness. Initially, septic patients are lethargic and inattentive, but can be roused. Late in sepsis patients are usually stuporous. The kidney stops making urine, the liver may stop conjugating bilirubin, and the intestine may become permeable to the endotoxins and bacteria in its lumen. There is lactic acidosis, probably because of glycolysis in ischaemic tissues. Eventually the heart stops beating.

Sepsis has for 40 years carried a mortality of about 30 per cent. If there is adult respiratory distress syndrome the mortality is 50 per cent, and if three or more organs fail the mortality is 85 per cent. For most of this time the treatment has been with antibiotics to eliminate bacteria, fluid replacement to restore intravascular volume, vasoconstrictors if necessary to get the mean arterial pressure above 70, provision for oxygenation by whatever means necessary, correction of obvious electrolyte disorders, hypoglycaemia, and perhaps extreme acidosis, and support for organ failure. In 28 clinical trials, attempts have been made to lower the mortality of sepsis using antagonists of known septic mediators such as tumour necrosis factor, IL-1, platelet activating factor, etc. And it has not worked: the mortality of sepsis has not changed over that 40-year period.

Recently some progress may have occurred. First, experiments in animals have made it clear that the high flow rates in early sepsis are deceptive: much of the bloodflow is bypassing the tissues, and from the outset there is tissue ischaemia. The most convincing demonstration of this was in septic rats: their livers showed strong fluorescence when illuminated with ultraviolet light, indicating that most of the cellular NAD was in the reduced state, and therefore that the hepatocytes were anoxic. Livers of normal rats had low fluorescence. Second it was realized that at least some of the progressive capillary obliteration, which underlies multiple organ failure, is due to exhaustion of clotting control proteins and widespread thrombosis.

Septic patients were shown to have very low or zero levels of activated protein C, and the lower the level the worse the outlook. Replacement with human recombinant activated protein C was shown to reverse some of the circulatory disorder. A placebo controlled double blind multicentre trial of the efficacy of human recombinant activated protein C in 1000 septic patients with at least one organ failing has just been published. For the first time in 40 years, a significant change in mortality was demonstrated. Mortality fell from 30 per cent to 25 per cent—not much, but significant in this number of patients. Where this will lead is difficult to say, but a way has been found to start attacking the capillary dysfunction which is the main problem in septic people.

Further reading

Bernard OR *et al.* (2001). Efficacy and safety of recombinant human activated protein C for severe sepsis *New England Journal of Medicine* **344**, 699–709.

Djiarski R, Tapping RI, Tobias PS (1998). Binding of bacterial peptidoglycan to CD14. *Journal of Biological Chemistry* **273**, 8680–90.

Moxon ER, Murphy PA (1978). *Haemophilus influenzae* bacteremia and meningitis resulting from survival of a single organism *Proceedings of the National Academy of Sciences of the USA* **75**, 1534–6.

7.6 Antimicrobial chemotherapy

R. G. Finch

Introduction

The discovery and clinical application of antibiotics and antimicrobial chemotherapeutic agents is one of the major achievements in medicine. Life-threatening infections such as meningitis, endocarditis and typhoid fever are now treatable, whereas before they were generally fatal. Likewise, the morbidity associated with many infectious diseases of a less life-threatening nature, such as urinary tract infections, skin and soft tissue infections, and bone and joint sepsis, has been substantially reduced. Perioperative prophylactic use of antibiotics has reduced the risk of infections complicating surgical procedures, such as large bowel and gall bladder surgery, vaginal hysterectomy, and implant surgery, such as the insertion of prosthetic heart valves, joints, and neurosurgical shunting devices.

Antimicrobial chemotherapy is the use of antibiotics and chemotherapeutic substances to control infectious disease. The term 'antibiotic' was coined by Waksman to describe a substance derived from naturally occurring micro-organisms and possessing antimicrobial activity in high dilution. The latter characteristic is essential in defining its selective toxicity to other micro-organisms. True antibiotics include penicillin, derived from the mould *Penicillium notatum*; streptomycin from *Streptomyces griseus,* and the cephalosporins from *Cephalosporium* spp. Many chemotherapeutic substances with antimicrobial activity have been artificially synthesized, such as the sulfonamides, quinolones, and isoniazid. However, the term 'antibiotic' is loosely applied to both the true antibiotics and other antimicrobial agents.

Antibiotics are among the most widely prescribed drugs, accounting for an international expenditure of $25b. In the United Kingdom, some 80 per cent of all prescribing is in the community where the emphasis is largely on oral agents; the remainder are used in hospitals where there is a greater emphasis on injectable drugs. More than 125 different antibiotics are available, but a relatively small number is necessary to deal with most prescribing needs. It is important that clinicians who prescribe are familiar with the principles of antimicrobial chemotherapy and that they adopt a continuous learning approach throughout their professional lives to ensure safe and effective prescribing. Table 1 summarizes the agents available for the treatment of bacterial, mycobacterial, fungal, viral, protozoal, and helminthic infections. More agents have been developed for the treatment of bacterial infections, but globally viral, fungal, and parasitic infections predominate. In recent years, there have been major advances in the availability of antiviral drugs particularly for the treatment of the herpesviruses and HIV. Likewise, safe and effective systemic antifungal agents have resulted from the discovery of azoles and triazoles.

The very success of antimicrobial chemotherapy has led to widespread and often excessive use, particularly in community practice where prescribing is largely empirical and clinical distinction between viral and bacterial infections is difficult. Antibiotics are used extensively in animal husbandry both for the treatment and prevention of infectious disease and, more controversially, as growth-enhancing agents among commercially raised poultry and swine. This has raised concerns about the emergence and spread of antibiotic resistance which affects many classes of antibiotic, may be intrinsic to a particular pathogen, or may result from genetic mutation. Resistance may be caused by enzymatic inactivation (b-lactamase), failure of drug penetration into the bacterial cell (porin mutation), alteration of the target binding site (e.g. penicillin-binding protein alteration in penicillin-resistant *Streptococcus pneumoniae*), or from efflux resistance whereby the drug is extruded from the bacterial cell (e.g. chloroquine-resistant *Plasmodium falciparum*). Organisms can also develop alternative metabolic pathways which by-pass drug inactivation. Resistance may be transferable between the same species or genera but may also spread between genera. Coding for multiple antibiotic resistance has been increasingly observed and results from a number of mechanisms, in particular plasmid transfer.

Despite the advances in antimicrobial chemotherapy, fresh challenges remain. These include the treatment and prevention of viral causes of enteric infection, meningitis, and hepatitis which are still without effective chemotherapy. Tuberculosis and malaria are among the world's major infectious disease killers and here problems of antibiotic resistance and, in the case of tuberculosis, the continuing need for long and complex regimens continue to frustrate disease management.

Among the more worrying trends in antibiotic resistance is the emergence within hospitals of methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant enterococci (VRE). Among community pathogens, *Strep. pneumoniae* has rapidly become less sensitive to penicillin causing clinical failures in the treatment of meningitis and otitis media. Internationally, multidrug resistant tuberculosis and multidrug resistant salmonellae, including *Salmonella typhi,* are of major concern.

Resistance is not confined to bacteria. Fungal resistance is increasing (e.g. of *Candida albicans* and *C. krusei*. to fluconazole). Resistance by the human immunodeficiency virus (HIV) to the nucleoside, non-nucleoside, and protease inhibitors is rapidly emerging. Within a decade of the introduction of antiviral agents, failure of chemotherapy may become a major factor responsible for progression of HIV disease.

Pharmacology

Mode of action

Knowledge of the pharmacological mode of action of an antimicrobial agents permits an understanding of the diverse mechanisms of microbial inhibition and the

opportunities for drug resistance. This is best established for antibacterial and antiviral agents. In the case of antifungal, and especially antiparasitic agents, the modes of action are less well defined. This reflects the process of drug discovery whereby an understanding of the biochemical and molecular action of agents derived from natural or chemical sources has not always been a priority in establishing efficacy and safety.

Antibacterial drugs

Antibacterial agents may affect cell wall or protein synthesis, nucleic acid formation, or may act on critical metabolic pathways (Table 2).

The b-lactams (penicillins, cephalosporins, and monobactams (aztreonam)) and the glycopeptides (vancomycin and teicoplanin), inhibit cell wall synthesis. The b-lactams, which share the common b-lactam ring, act on cell wall transpeptidases to inhibit cross-linking of peptidoglycan. The glycopeptide antibiotics act at an earlier stage of cell wall synthesis by binding to acyl- D-alanyl-D-alanine. Despite their similar mode of action, they are less efficient bactericides than the b-lactams.

Inhibitors of protein synthesis

Antibacterial agents that inhibit protein synthesis act on the 30S ribosomal subunit responsible for binding mRNA, or the 50S subunit which binds aminoacyl tRNA. The aminoglycosides, tetracyclines, and macrolide antibiotics are the most widely used inhibitors of protein synthesis. Chloramphenicol, clindamycin, and the recently introduced agent quinupristin/dalfopristin (Synercid) also act at this site.

Inhibitors of nucleic acid

Nucleic acid synthesis is targeted by quinolones, metronidazole, and rifampicin. The bacterial DNA gyrase is essential for the supercoiling of bacterial DNA. This, together with the enzyme topoisomerase IV, are the major targets for the quinolones. These enzymes are absent in humans, explaining the selective activity of these drugs. Rifampicin and other rifamycins interfere with DNA-dependent RNA polymerase, preventing chain initiation.

Metabolic inhibitors

The best known metabolic inhibitors are the sulfonamides and trimethoprim which interfere with folic acid synthesis by sequentially inhibiting the enzymes pteroic acid synthetase and dihydrofolate reductase. By acting sequentially, a combined bactericidal effect results. The selective activity of these compounds is dependent upon the fact that humans are unable to synthesise folic acid and require preformed folic acid in their diet.

Antiviral agents

Viruses live and replicate within the host cell. Antiviral chemotherapy therefore presents a particular challenge if it is to be selectively toxic. The cycle of viral replication provides a number of opportunities for therapeutic intervention. Most available antiviral agents are nucleoside analogues, largely used in the treatment of HIV or herpesvirus infections (Table 3). The recent growth in numbers of antiviral agents has benefited greatly from HIV related research through the identification of new drug targets (Fig. 1). Interference with cell surface attachment through ligand blockade of surface receptors provides a theoretical, as yet unfulfilled, target. Penetration into the host cell may be through a process of translocation or direct fusion between the outer lipid membrane of the virus and the cell membrane, before uncoating and release of viral nucleic acid. Replication differs among viruses, thereby providing a number of therapeutic options. Viral mRNA becomes translated into multiple copies of viral proteins encoded by the viral genome either as a result of virus-specific enzymes or by co-opting host-derived protein. HIV, for example, employs its own reverse transcriptase to convert RNA to DNA before integration into the host cell chromosome. Transcription and translation follow. Before the virus can be released, new viral particles must be assembled for which host cell proteins and mechanisms of phosphorylation and glycosylation may be recruited. The protease inhibitors act at this stage and have been particularly successful. Virus release is either the result of transportation and budding, or host cell lysis.

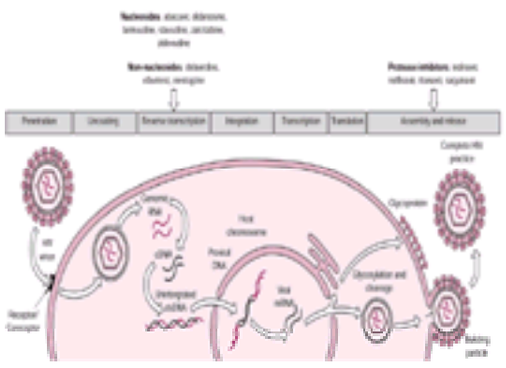

Fig. 1 Sites of inhibition of HIV replication by current antiretroviral drugs.

Antifungal agents

The polyene antifungals (amphotericin B and nystatin) act on ergosterol within the fungal cell membrane. Ergosterol is largely absent from bacteria and humans, explaining the selective toxicity of these agents. The azole antifungals include the imidazoles (e.g. clotrimazole, miconazole, ketoconazole) and the triazoles (fluconazole and itraconazole) which bind preferentially to fungal cytochrome P450 to inhibit 14-a-methylsterol demethy-lation to ergosterol.

Antiparasitic agents

The mechanism of action of many antiparasitic drugs is only partially known. Among the antimalarials, chloroquine interferes with the digestion of haemoglobin taken up by *Plasmodia*. Quinine is thought to act in a similar manner. Metronidazole is active against a number of protozoa such as *Entamoeba histolytica* as well as anaerobic bacteria. It acts as an electron sink, by reducing of its 5-nitro group and activated by nitroreductase within the target pathogen, thus interrupting DNA synthesis.

Among the antihelminthic drugs, piperazine and paraziquantel act by selectively inducing muscle paralysis in the target helminth. Others, such as thiabendazole, inhibit parasitic ATP synthesis and energy production.

Antimicrobial spectrum of activity

The antimicrobial spectrum of an agent is dependent upon target site susceptibility among pathogenic organisms at clinically achievable drug concentrations. Some micro-organisms are intrinsically resistant to certain antibiotics. For example the aminoglycosides are inactive against anaerobic bacteria because cell entry is an energy dependent process relying on respiratory quinones, which are absent in anaerobic bacteria. Certain strains of *Pseudomonas aeruginosa* are resistant to the aminoglycosides as a result of altered protein porin channels which inhibit antibiotic penetration.

The antimicrobial spectrum of a drug in part dictates its clinical indications. While information on this spectrum is more easily determined *in vitro*, *in vivo* efficacy can only be confirmed through clinical use, which can be supported by animal model data during drug development. For example *in vitro Salmonella typhi* is susceptible to gentamicin, but the drug is not effective clinically.

Narrow spectrum and broad spectrum agents

There are few truly narrow spectrum agents. Fusidic acid, mupirocin, and the glycopeptides (vancomycin and teicoplanin) target specific pathogens and are mainly

used to treat microbiologically confirmed infections.

Broad spectrum agents, such as the quinolone antibiotics and the parenteral cephalosporins, such as cefotaxime and ceftriaxone, are active against many Gram-positive and Gram-negative pathogens. Metronidazole has activity against a large number of anaerobic bacteria and, because of this restricted activity, is considered to have a narrow spectrum. The aminoglycosides, although active against staphylococci and aerobic Gram-negative bacilli are inactive against streptococci and anaerobes and are therefore frequently prescribed in combination. The carbapenems (imipenem and meropenem) possess the broadest spectrum of activity which includes most aerobic and anaerobic bacterial pathogens. Broad spectrum agents are often used empirically in the initial management of severe infection. However, they frequently affect the normal flora so that super-infection with *Clostridium difficile* and yeasts are more likely to arise.

Susceptibility testing

Antibiotic susceptibility testing of clinical isolates is important for appropriate prescribing and for gathering epidemiological data. It is determined *in vitro* by using either broth- or agar-based methods. Pathogens are exposed to known concentrations of an antibiotic and their degree of inhibition compared to a standard control. Disk susceptibility testing is the most widely used method. Zones of inhibition around the antibiotic-containing disk are measured, compared to a standard, and the pathogen designated sensitive, resistant, or of intermediate susceptibility to the drug. Currently, such methods require the isolate to be tested in pure culture. It is therefore difficult to obtain information on the susceptibility of a pathogen in less than 36 to 48 h from sample collection.

The minimal inhibitory concentration (MIC) in mg/l provides more precise *in vitro* information on the activity of a drug against bacterial pathogen. It is more time consuming and costly to determine although automated systems and commercial strip tests are available (Fig. 2). Defining susceptibility by MIC determination permits greater predictive benefit in the treatment of certain infections such as gonorrhoea, bacterial endocarditis, and pneumococcal meningitis. Knowledge of the *in vitro* susceptibility of common pathogens to antimicrobial agents (Fig. 3) is helpful in selecting drug therapy but is only relevant to the achievable drug concentrations, which is important in predicting performance as discussed below.

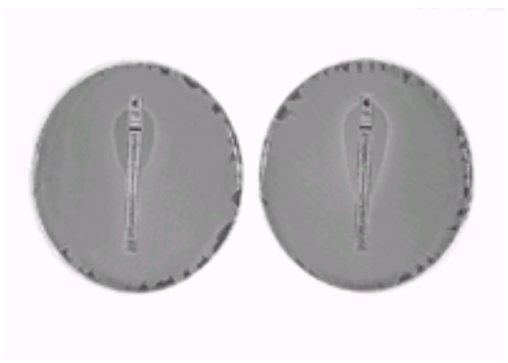

Fig. 2 *Staphylococcul aureus* resistant to penicillin (minimum inhibitory concentration 8 mg/l) on the left and sensitive to vancomycin (minimum inhibitory concentration 1.0 mg/l) on the right as demonstrated by a commercial strip test.

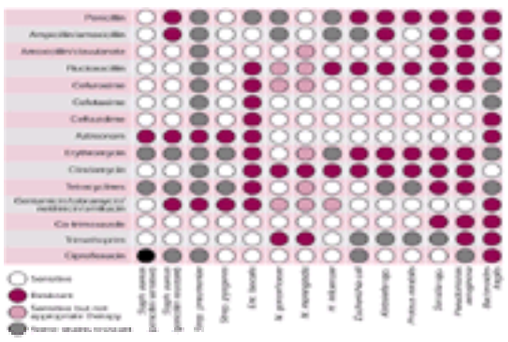

Fig. 3 Sensitivity of selected pathogenic bacteria to some common antibacterial agents.

Combined drug therapy

In hospital practice, it is common to combine agents when dealing with mixed infections or where initial broad spectrum empirical therapy is required. Another important reason for combining drugs is to prevent the emergence of antibiotic resistance, such as in the treatment of tuberculosis and HIV infections. Antituberculosis regimens have been developed to ensure that naturally occurring minority populations of *Mycobacterium tuberculosis* resistant to isoniazid or rifampicin do not emerge during therapy. By combining isoniazid and rifampicin with pyrazinamide and ethambutol for the initial phase of therapy (2 months), resistance is usually avoided. Therapy can be restricted to isoniazid and rifampicin for the continuation phase (4 months). The regimen is extended in those patients unable to tolerate pyrazinamide and in the treatment of tuberculous meningitis (Table 4).

HIV infection is treated with multidrug regimens. The success of highly active antiretroviral therapy, in which nucleoside analogues and protease inhibitors are combined in a three-drug regimen, is not only based on greater potency of the combined regimen but also on its ability to slow the emergence of drug-resistant mutants. The more recently introduced non-nucleoside reverse transcriptase inhibitors, such as efavirenz, appear to be equally potent in combination with nucleoside analogues and can delay the need for using protease inhibitors. This may increase the period of time in which antiretroviral therapy remains effective in an individual. The options for treating HIV infection are summarized in Table 5 (see also Chapter 7.10.21).

Occasionally, drugs are combined for the purpose of achieving a synergistic effect based on evidence that the *in vitro* activity of the combination is shown to be greater than the sum of the activity of the individual agents. Most drugs in combination will simply be additive in effect. One of the more frequently prescribed synergistic combinations is that of penicillin (or ampicillin) and streptomycin (or gentamicin) in the treatment of endocarditis caused by *Enterococcus* spp. The aminoglycoside alone is generally inactive against enterococci but in combination with ampicillin achieves synergistic killing (Fig. 4). A similar effect is employed in the treatment of viridans streptococcal endocarditis with this combination.

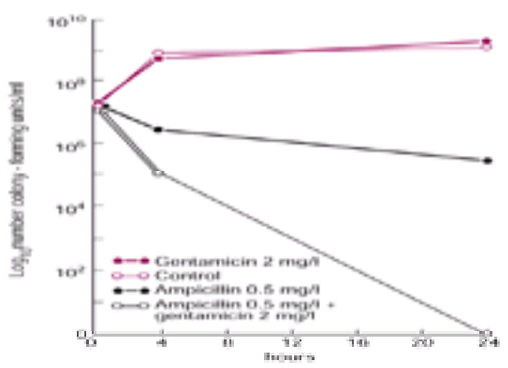

Fig. 4 Effects of ampicillin (0.5 mg/l) and gentamicin (2 mg/l) alone and in combination on a strain of *Enterococus faecalis* from a patient with infective endocarditis. A

synergestic effect is observed with the combined agents.

Another widely used example of synergistic inhibition is the combined effects of an antipseudomonal b-lactam, such as ceftazidime or piperacillin, and an aminoglycoside, such as gentamicin, tobramycin, or amikacin. This combination is used to treat documented or suspected *Pseudomonas aeruginosa* infections occurring in neutropenic states complicating bone marrow transplantation, cytotoxic chemotherapy, and burn wound infections.

Antibiotic resistance

General considerations

Antibiotic resistance has been recognized since the introduction of effective antibiotics. For example penicillin-resistant strains of *Staphylococcus aureus* became widespread shortly after the introduction of this agent; penicillin sensitive strains are now uncommon. Resistant strains of Gram-negative bacteria, such as *Klebsiella, Enterobacter, Acinetobacter,* and *Pseudomonas* spp. are commonly found in high dependency units where they may cause epidemics. The international emergence of epidemic MRSA infections, primarily within hospitals and nursing homes, is very worrying. Conventional approaches to controlling these infections have been largely unsuccessful. The emergence of MRSA together with multiple-antibiotic-resistant coagulase-negative staphylococci has rapidly increased the use of vancomycin. Vancomycin-resistant enterococci has emerged in specialist hospital facilities such as dialysis and haematology units; therapeutic options are limited. Other problems include the emergence of penicillin-resistant pneumococci and b-lactamase producing *Haemophilus influenzae*.

At present, there is great international concern among professionals, politicians, and, increasingly, the public about antibiotic resistance. In the United Kingdom, the House of Lords published an influential document in 1998 reviewing the issues surrounding this problem. This has led to a number of initiatives: reducing the use of antibiotics, particularly in the treatment of minor upper respiratory tract infections in the community; education strategies for prescribers and the public; and better enforcement of infection control policies. Within the European Union, similar measures have been proposed together with a ban on the use of antibiotics as growth promoters in livestock animals. However, antibiotic resistance is a global problem. An increasing number of multidrug resistant infections caused by *Salmonella* spp. and *Mycobacterium tuberculosis* are being imported from developing countries where the availability or prescribing of antibiotics is less controlled.

Antibiotic resistance drives changes in patterns of prescribing and is a major impetus to the pharmaceutical industry in its search for new therapies. Micro-organisms differ in their ability to develop resistance, which may affect a particular drug, a class, or multiple classes of antibiotics. Genetic mutations select for antibiotic resistance which frequently occur under the influence of antibiotic pressure. The major mechanisms of resistance are summarized in Table 6. Resistance to single or multiple antibiotics may be either chromosomally, or plasmid mediated, or both. In turn, genes may code for resistance to a single or multiple antibiotics. In addition to plasmid-mediated resistance, other transposable genetic elements (transposons), and insertion sequences incapable of self-replication, may exist within a chromosome, plasmid, or bacteriophage.

Resistance genes are most frequently transferred between organisms by conjugation. This occurs between the same or different species of bacteria and also between different genera. Other mechanisms of transferring resistance include transduction via a bacteriophage, and less commonly transformation in which naked DNA released during cell lysis is taken up by other bacteria.

Transposon-mediated resistance reflects transfer of discreet sequences of DNA between chromosomes or plasmids whereby individual or groups of genes can be inserted into the host bacterial cell. More recently, molecular structures known as integrons have been identified which facilitate new combinations of resistance genes within the bacterial chromosome, plasmid, or transposons. The antibiotic resistance genes are bound on each side by conserved segments of DNA. These individual resistance genes can be inserted or removed between the conserved structures and act as an expression vector for antibiotic resistance genes.

While the molecular mechanisms of antibiotic resistance are legion, the ability of drug-resistant micro-organisms to survive, disseminate, and cause disease varies widely. In many instances, antibiotic resistance may give a survival advantage only in the presence of continued antibiotic exposure to such agents. This is reflected in the occurrence of epidemic disease in high-dependency units such as intensive care facilities where antibiotic usage is often high. However, it is also clear that once the genetic mechanism for evading antimicrobial activity has been acquired, it is rarely lost and adds to the continuously expanding genetic memory that has steadily eroded the efficacy of many antimicrobial drugs.

Enzymatic inactivation

Aminoglycoside-modifying enzymes include adenylating, acetylating, and phosphorylating enzymes. Gentamicin is the most susceptible and amikacin the least susceptible to such inactivation. However, the largest group of inactivating enzymes are the b-lactamases which hydrolyse the b-lactam ring common to all penicillins and cephalosporins. Penicillinase was the first b-lactamase to be identified and is the reason why most strains of *Staphylococcus aureus* are resistant to this drug. Another important b-lactamase is that known as TEM-1 which is responsible for resistance to ampicillin by *Haemophilus influenzae*. The major impetus to the development of the penicillins and cephalosporins was to extend their spectrum of activity by resisting inactivation by b-lactamases present in many aerobic Gram-negative bacilli. However, new inactivating enzymes continue to emerge, including the extended spectrum b-lactamases, which are now limiting the clinical utility of third-generation cephalosporins, and the carbapenemases which hydrolyse imipenem.

Impermeability resistance

Drug uptake of antibiotics such as the penicillins, tetracyclines, and quinolone antibiotics by bacteria is through protein channels (porins) which cross the outer membrane. Alterations in the permeability of the outer membrane of Gram-negative bacteria is an increasingly important mechanism of drug resistance. Mutations in porin structure are responsible for resistance among pathogens such as *P. aeruginosa* and *Serratia marcescens.*

Alterations in target site

Another important mechanism of resistance is mutational modification of drug binding sites. This affects susceptibility to b-lactams, erythromycin, chloramphenicol, and rifampicin. Erythromycin and chloramphenicol bind to the bacterial 50S ribosomal subunit which is subject to genetic mutation. In contrast, the quinolones target DNA gyrase which is subject to subunit structure alteration resulting in one variety of resistance to drugs such as ciprofloxacin. The increasing resistance to penicillin among *Streptococcus pneumoniae* is the result of reduced binding of penicillin to several binding proteins (PBP-2a and PBP-2x). *Staph. aureus* resistance to methicillin is due to the presence of penicillin binding protein (PBP-2a) which has reduced affinity for methicillin and other b-lactams and is encoded by the *mecA* gene.

The recently recognized problem of vancomycin resistant enterococci, which largely affects *Enterococcus faecium,* is the result of the production of enzymes (ligases) which permit continued cell wall synthesis despite the presence of vancomycin. To date, five different genes have been found responsible for this phenomenon (Van A-E) which result in different phenotypic patterns of resistance to the glycopeptides vancomycin and teicoplanin.

Metabolic bypass resistance

Bacteria must synthesize folic acid from the precursor *p*-aminobenzoic acid. The sulfonamide antibiotics competitively inhibit the enzyme dihydropteroate synthetase. Trimethoprim acts on the same metabolic pathway by inhibiting dihydrofolate reductase. The sequential inhibitory effects of trimethoprim and sulfamethoxazole (co-trimoxazole) result in synergistic bactericidal activity against many pathogens. Resistant organisms are able to synthesize their own enzymes thereby evading such competitive inhibition.

Surveillance of antibiotic resistance

Information on the susceptibility of pathogenic micro-organisms is important. Such data can provide information on the relative frequency of pathogens and the pattern of susceptibility to prescribed agents. Surveillance, therefore, has a role in guiding prescribing, in developing prescribing policies and in identifying and monitoring

organisms which are subject to infection control measures. On a broader front, surveillance is also of value in alerting industry and health-care planners to the need for new drug and vaccine strategies for disease control.

To be of maximum benefit, surveillance needs to be sensitive to a defined geographical base which may simply reflect the catchment area of specimens submitted to a particular laboratory, providing information on the trends in community and hospital isolates. Within hospitals, more specific information can be provided about susceptibility patterns in high dependency units, where antibiotic consumption is often greater, and more resistant pathogens, such as *Klebsiella, Serratia, Enterobacter, Acinetobacter spp.,* and *Pseudomonas aeruginosa* are found. Among Gram-positive pathogens, *Staphylococcus aureus* and enterococci present an increasing challenge to prescribing and infection control practice.

National networks of surveillance often vary in their focus and include data on enteric pathogens, *Staphylococcus aureus*, penicillin resistance among pneumococci, and, more recently, vancomycin resistant enterococci. There are important international networks which collect information on such pathogens as *Legionella pneumophila* and *Mycobacterium tuberculosis*. Drug resistant tuberculosis is increasingly prevalent in the United Kingdom and overseas.

Surveillance of resistance to antiviral agents is rudimentary. Patient-specific data is increasingly sought in those with HIV infection, to assess drug failure, guide change in management, and direct primary therapy in selected cases of person-to-person and mother-to-infant transmission. Determination of phenotypic resistance is still costly and time consuming and most data relate to genotypic patterns of resistance to antiretroviral drugs among HIV isolates.

Pharmacokinetics

To be effective, antimicrobial agents must achieve therapeutic concentrations at the site of the target infection. This may be localized to a single anatomical site, such as the bladder or the cerebrospinal fluid, or involve a major organs such as the lung. Infections may also be generalized and affect many body sites. Drug selection must take into consideration the fact that pathogens such as *M. tuberculosis, Legionella pneumophila,* and *S. typhi* replicate intracellularly. Antimicrobial drugs may be administered arenterally, orally, or topically to the skin, external auditory meatus, conjunctiva, and by intraocular application. In the case of systemically active agents, the effective drug concentrations are determined by the standard pharmacokinetic parameters of absorption, distribution, metabolism, and elimination. Since selective toxicity is crucial to safe prescribing, the dose regimen for each agent aims to avoid concentrations toxic to the host but inhibitory to the micro-organism. This 'therapeutic window' varies by drug.

Bioavailability

The rate and degree of absorption from the gastrointestinal tract is not only important for plasma concentrations reflected in the pharmacokinetic parameters of C_{max} and T_{max} of a drug, but also for potential adverse effects on the bowel (Table 7). For example ampicillin, the first of the aminopenicillins, commonly caused gastrointestinal side-effects, most notably diarrhoea. These effects have been reduced by increasing the bioavailability of the active drug through the introduction of hydroxy-ampicillin (amoxicillin) and various esters and prodrugs of ampicillin.

Some agents such as cefalexin, doxycycline, and a number of the quinolone antibiotics are extremely well absorbed, achieving 80 to 100 per cent bioavailability. In the case of some recent quinolones, the excellent bioavailability has raised the possibility of treating with oral antibiotics some severely ill patients who might normally require parenteral therapy. In contrast, drugs which are poorly bioavailable, such as cefixime and cefuroxime axetil, not only have a higher incidence of gastrointestinal side-effects but are also more likely to select for *Clostridium difficile* associated large bowel disease.

Distribution

Most drugs are distributed in the blood via the plasma before gaining access to the extracellular fluid. Tissue concentrations of a particular agent are affected by pH, drug ionizability, lipid solubility, and the presence of an inflammatory reaction whereby the capillary fenestrations are increased in size. In the case of agents administered intravenously by infusion or by bolus injection, the distribution phase is rapid in comparison with orally, rectally, or intramuscularly administered drugs. Drugs which are poorly lipophilic, such as the b-lactams and aminoglycosides, achieve low concentrations in tissues such as the brain. However, the b-lactams achieve therapeutic concentrations in the cerebrospinal fluid as a result of the inflammatory reaction which accompanies meningitis.

Drugs may also be taken up intracellularly, as in the case of macrolides and quinolones, resulting in a large volume of distribution compared to drugs confined to the extracellular space, such as the b-lactams and aminoglycosides. This is important in relation to the treatment of intracellular pathogens such as *Mycoplasma pneumoniae*, *Legionella pneumophila,* and *Mycobacterium tuberculosis* which can only be effectively treated by drugs which are concentrated and remain biologically active intracellularly.

The plasma half-life (T½), which is the time required for the concentration of a drug in the plasma to fall by half, is affected by drug distribution and, in particular, its rate of elimination as a result of metabolism and excretion. This in turn affects the time taken to reach steady state. In the treatment of life-threatening infections, it is important that steady state kinetics are achieved rapidly. This may require the administration of a loading dose. This applies to the use of agents such as gentamicin for the treatment of serious Gram-negative infections and intravenous quinine in the case of life-threatening malaria where the pharmacokinetic behaviour can be altered by the severity of the disease in comparison with healthy subjects (Fig. 5).

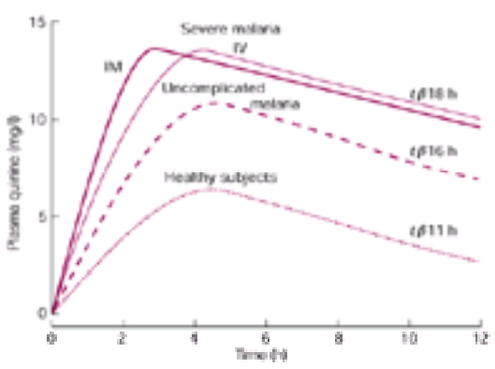

Fig. 5 Average plasma quinine concentrations following administration of a loading dose of 20 mg (salt)/kg to patients with severe and uncomplicated malaria, compared with those predicted to occur in normal subjects. (Reproduced from White (1992), with permission.)

Drugs are commonly distributed in the blood and tissues bound to plasma proteins, largely albumin. Drugs vary in their degree of protein binding. With agents such as flucloxacillin and ceftazidime it exceeds 95 per cent. The importance of protein binding lies in the fact that the active moiety is the unbound drug. Dissociation from the bound to the unbound state is usually rapid but this equilibrium may affect drug performance at certain sites such as the joints. The relationship between protein binding and drug performance has been emphasized in recent studies of the pharmacodynamics of drug activity (see below).

Metabolism

Antibiotics, like other drugs, are degraded at various sites in the body but, predominantly within the liver. Degradation involves conjugation, hydrolysis, oxidation, glucuronidation, or dealkylation, according to the particular drug. Members of the hepatic cytochrome P450 group of enzymes play a dominant role in this process. Drug metabolites are usually but not always biologically inactive. For example cefotaxime is degraded to desacetyl-cefotaxime and clarithromycin to hydroxy-clarithromycin, both of which are biologically active and contribute to the overall antibacterial activity of these agents.

Excretion

Most drugs are excreted in the urine by glomerular filtration, tubular secretion, or a combination of these mechanisms. Thus high concentrations of drug will often be present in the urine; this has therapeutic importance in the treatment of urinary tract infections. Urinary pH affects the biological activity of many drugs; for example the activity of ciprofloxacin is markedly reduced at pH 5.5. Tubular excretion can be blocked by probenecid. This was formerly used to ensure higher plasma concentrations of penicillin and is still recommended in the treatment of gonorrhoea with single doses of amoxicillin, ampicillin, or intramuscular procaine penicillin. It is also important to note that any reduction in glomerular filtration rate will not only affect urinary concentrations of drug but also the plasma half-life and, in turn, serum concentrations of drugs which are primarily excreted by this route. In the case of antibiotics such as the aminoglycosides and vancomycin, the dose must be reduced in renal failure.

Biliary excretion is another important route for drug elimination either as the active compound or as a microbiologically active or inactive metabolite. Reabsorption from the gastrointestinal tract can result in enterohepatic recirculation, which in turn may affect plasma half-life. Drugs which achieve high concentrations in the bile are effective in the treatment of infections at this site such as cholecystitis. However, biliary obstruction or hepatic impairment may reduce therapeutic efficacy and require dose reduction to avoid toxic effects. Examples include clindamycin, efavirenz, mefloquine, and tetracyclines.

Therapeutic drug monitoring of some antibiotics is essential in order to ensure therapeutic yet non-toxic concentrations. This applies particularly to aminoglycosides which have a relatively narrow therapeutic index. Trough concentrations of gentamicin in excess of 2 mg/l, if sustained, can result in nephrotoxicity and ototoxicity. The target cells for such toxicities are the renal tubular lining cells and the cochlear hair cells of the inner ear respectively. Vancomycin is also frequently monitored, particularly in patients with impaired renal function.

Pharmacodynamics

The inter-relationship between drug, micro-organism, and the infected host creates an important pharmacological dynamic. Antibiotics are unique in therapeutics in that they are targeted at an invading micro-organism which may be present at a particular site or be more widely distributed in the body. The host's response to infection may modify the pharmacokinetic handling of a drug. Many antibiotics have a measurable effect on a variety of bacterial and host cell functions, even at subinhibitory concentrations. It is difficult to establish the exact role that these factors play clinically, but they are likely to contribute to the overall effect of an antibiotic. Macrolides, such as erythromycin, illustrate this point since they affect a variety of virulence characteristics (Table 8) as well as affecting the host's response to infection.

Exposure of micro-organisms to sublethal concentrations of an antibiotic may temporarily inhibit growth which recommences following removal of the drug. The time to recovery is known as the postantibiotic effect. This varies with the drug and the micro-organism; for example the quinolones have a longer postantibiotic effect than b-lactams (Table 9). The relevance of this observation to the *in vivo* situation, where plasma drug concentrations are often well above the inhibitory concentration and are sustained through repeat dosaging, remains uncertain. It may have greater relevance to tissue concentrations which tend to be lower than plasma concentrations. Postantibiotic effect certainly contributes to the effects of agents that are administered once daily, such as gentamicin.

The relationship between the pharmacokinetic characteristics of a drug and bacterial inhibition is critical to therapeutic outcome (Table 10). In the case of agents such as penicillins and cephalosporins, the time that drug concentrations are maintained above the minimum inhibitory concentration (MIC) predicts the response. This contrasts with agents such as the quinolones and aminoglycosides, where it is more important to achieve high C_{max} to MIC ratios. Modelling the MIC of a particular organism against the dose response curve for a drug (Fig. 6) has established a number of important pharmacodynamic parameters which have been supported by studies in animal models and man. For example dosage regimens of quinolones, such as ciprofloxacin and levofloxacin, have been based on pharmacodynamic data. The ratio of C_{max}:MIC has been refined in the parameter, area under the inhibitory concentration, which is the ratio of the area under the time curve (AUC):MIC. This is more predictive of outcome. The importance of protein binding for drug performance has also emerged as an important modifying factor in this modelling. The AUC:MIC of the free drug is the most sensitive predictor of response. The manner in which these ratios differ for selected quinolones is shown in Table 11.

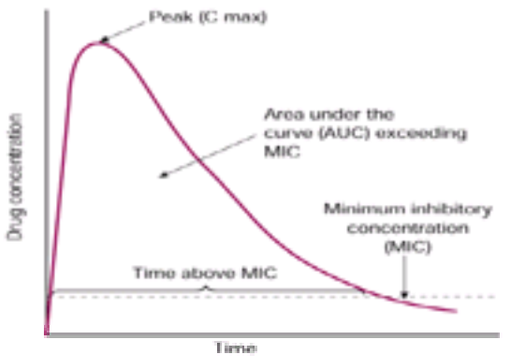

Fig. 6 Relationship between the MIC of a drug and its pharmacokinetic profile.

Principles of use

In comparison with many other classes of drugs, antimicrobial agents are usually prescribed in short courses ranging from a single dose to a few days. Prolonged therapy is required for certain infections such as tuberculosis and bone and joint sepsis, and for HIV infection, treatment is usually life-long.

Most antibiotic prescribing, especially within community practice, is empirical. Even among patients in hospital, where there are greater opportunities for diagnostic precision based on laboratory investigations, the exact nature of the infection is established in only a minority of cases. Most therapeutic prescribing requires a presumptive clinical diagnosis which, in turn, is linked to a presumptive microbiological diagnosis based on knowledge of the usual microbial causes of such infections. Among the most widely treated infections are those affecting the upper and lower respiratory tracts, the urinary tract and skin and soft tissues for which the likely microbial aetiology is restricted. For example urinary tract infections arising in the community are usually caused by *Escherichia coli* and other Gram-negative enteric pathogens and, less commonly, by enterococci or *Staphylococcus saprophyticus*. Local knowledge of the susceptibility of these pathogens to commonly used agents such as trimethoprim, ampicillin, and a quinolone such as norfloxacin is helpful in recommending initial empirical antibiotic management.

In more severe infections, such as community-acquired pneumonia, prompt empirical therapy is essential. Although the range of possible pathogens is more extensive (Table 12) *Streptococcus pneumoniae* predominates and must always be targeted. Assessment of severity, based on validated criteria, assists in defining the initial empirical antibiotic regimen. This is illustrated by the British Thoracic Society's recommendations for the initial empirical antibiotic management of community acquired pneumonia (Table 13).

The use of empirical therapy depends on the ease with which a clinical diagnosis can be made, disease severity, and toxicity. In the case of herpesvirus infections, the empirical use of aciclovir for the treatment of mucocutaneous herpes simplex infections or of shingles in the elderly is now common. However, it would be inappropriate to start treatment for HIV or CMV infections without laboratory support for these diagnoses in view of the toxicity and cost of the antiviral agents used to treat these infections.

Antibiotic prophylaxis

Antibiotics are used widely in the prevention of infection, in association with surgery and in a range of medical conditions (see above). Antibiotic prophylaxis is used for selected surgical procedures where the risk of infection, although relatively low, is of serious import should it occur. Examples include prosthetic joint implantation and cardiac surgery in which prosthetic valves and intracardiac patches are inserted.

The principles of antibiotic prophylaxis are based on the selection of an agent active against the known potential target pathogen(s). The drug should be present in

high concentrations at the site and time of surgery and be relatively free from adverse reactions. One or two doses are generally effective depending on the length of the procedure. No regimen can be effective against all potential pathogens hence the importance of postoperative follow-up.

An important medical indication for the use of prophylactic antibiotics is the prevention of bacterial endocarditis in those with established valvular heart disease undergoing selected dental, urinary tract, or gastrointestinal procedures during which transient bacteraemia carries the risk of endocardial infection. Here again, the principles governing the selection of the regimen are based on the recognition of the likely target pathogens, their pattern of susceptibility, and the necessity to ensure high bactericidal concentrations of drug at the time of the procedure. In community dental practice, single-dose oral therapy with high-dose amoxicillin/clavulanate (co-amoxiclav) or in those allergic to this agent, clindamycin, are currently recommended. Another example of effective prophylaxis is the use of low-dose suppressive therapy to prevent *Pneumocystic carinii* pneumonia in those with advanced HIV infection. Co-trimoxazole is the preferred agent. Dapsone and inhaled pentamidine are also used.

Anatomical or functional asplenia is associated with a 12.6-fold increased incidence of severe sepsis compared with the general population. This risk is related to the patient's age and, in those splenectomized, the reason for surgery and the period of time that has elapsed. Young children are particularly at risk but this declines substantially after the age of 16 years. Hence the recommendation for prophylactic oral penicillin (erythromycin for the intolerant) to prevent fulminant pneumococcal sepsis which predominates. It will not offer protection against other pathogens such as *Escherichia coli* and *Pseudomonas aeruginosa.* Apart from good evidence for the benefit of prophylaxis in children with sickle cell disease, there is poor support for efficacy in splenectomized patients.

There remain, therefore, differences of opinion about the recommendation for the continued use of chemoprophylaxis in adults. Issues of cost, compliance, and drug resistant pathogens add further fuel to the debate. What is clear is that the patient or legal guardian(s) should be educated concerning this risk.

Dose selection

Few antibacterial drugs are specific to a single pathogen, hence the dosage regimen must capture a range of susceptibilities of the various target micro-organisms to ensure a successful response. The dosage regimen is determined initially by pharmacokinetic studies in healthy volunteers. This is supplemented by information from standardized animal models which simulate infections such as peritonitis, endocarditis, meningitis, thigh abscess, otitis media, pneumonia, and sepsis complicating neutropenia. In man, information on drug penetration into CSF, bile, joint fluid, and cutaneous blisters can be supplemented by data from biopsy specimens from sites such as tonsils, bronchus, and prostate. The role of pharmacodynamic assessment is of increasing importance in defining dose and predicting outcome as discussed earlier. Despite all this information, the definitive dosage regimen still requires support from large clinical trials in which the endpoints of response are precisely determined.

Bactericidal versus bacteristatic agents

In the treatment of many common community infections which are usually of mild or moderate severity, the choice of either a bacteristatic or bactericidal antibiotic is of limited importance. However, in patients with severe infection, particularly when complicating an immunocompromised state, a bactericidal agent must be used. This applies particularly to those with severe granulocytopenia which is a common accompaniment of cytotoxic chemotherapy, especially in the treatment of haematological malignancies and following bone marrow transplantation. Another important indication for selecting a bactericidal regimen is in the treatment of infective endocarditis; although the infected vegetations are in the bloodstream, they are relatively protected from host phagocytic control. Effective penetration into the fibrin–platelet mass requires high concentrations of a bactericidal drug to sterilize the infected vegetations.

Duration of treatment

The duration of therapy for many common infections has not been rigorously determined. The treatment of many common conditions is based on custom and practice and often varies internationally. The duration of treatment has been more thoroughly determined in the following cases:

- Gonococcal urethritis responds promptly to single dose treatment with agents such as ceftriaxone, spectinomycin, or a quinolone antibiotic such as ciprofloxacin or ofloxacin.
- Uncomplicated urinary tract infection, particularly when affecting women of child bearing years, responds promptly to single dose and 3-day treatment regimens with selected agents such as trimethoprim and norfloxacin, although bacteriuria can be eliminated with a single dose, the symptoms of dysuria and frequency take longer to subside, hence a 3-day course is preferred.
- Pharyngitis caused by *Streptococcus pyogenes* improves symptomatically within a few days of antibiotics such as penicillin but eradication of the infecting organism from the throat often takes up to 10 days. It is acknowledged that this presents major difficulties with regard to drug compliance.
- Pulmonary tuberculosis—the current recommendation of 6-months treatment with rifampicin, isoniazid, pyrazinamide, and ethambutol, reducing to isoniazid and rifampicin for a further 4 months provided the isolate is confirmed to be susceptible, is based on extensive clinical trials (Table 4).
- Bacterial endocarditis—knowledge of the *in vitro* susceptibility of the infecting organism is crucial in determining dose, duration, and outcome of therapy. Highly penicillin-sensitive strains (MIC < 0.1 mg/l) of viridans streptococci are treated effectively with a 2-week regimen of parenteral penicillin which may be supplemented with sequential high-dose oral amoxicillin for a further 2 weeks. Less sensitive strains should be treated with parenteral penicillin and gentamicin for a total of 4 weeks, which is essential if the infecting organism is an enterococcus.

Infections caused by *Staph. aureus* are a particular challenge since the severity is highly variable and yet the potential for metastatic infection and chronicity as in the case of osteomyelitis must be kept in mind. The isoxazolyl penicillins, such as flucloxacillin are preferred with, or without, the addition of fusidic acid. Clindamycin is a useful alternative agent. Many *S. aureus* of the skin and soft tissues respond promptly to 7 to 14 days oral therapy. Where there is a severe systemic response to infection, parenteral therapy is appropriate initially. Where there is evidence of dissemination, treatment should be extended for periods up to 4 weeks.

In the case of septic arthritis, antibiotics should be given promptly and joint aspiration carried out, sometimes repeatedly, to avoid damage to the articular cartilage. The duration of therapy has not been rigorously determined. Most infections will resolve in 2 to 3 weeks. One of the most challenging infections is staphylococcal osteomyelitis. To avoid chronicity, it is customary to treat for 4 to 6 weeks. Treatment is generally administered parenterally. In centres where skill, experience, and administrative support exist, patients are increasingly being managed in the community by parenteral administration through peripherally inserted venous catheters. Under these circumstances, a glycopeptide such as teicoplanin is convenient since it is administered once daily.

For most infections, the duration of therapy remains uncertain. However, many mild to moderate uncomplicated infections will defervesce within a 3 to 5-day period suggesting that 5 to 7-days treatment is usually adequate. There is little evidence to suggest that treatment periods of 7 to 14 days, or longer, are any more effective and are likely to be associated with an increased risk of side-effects, superinfection, and the selection of antibiotic resistant organisms, apart from being more costly.

The parenteral administration of antibiotics is appropriate in the management of severe life-threatening infections and when oral therapy is contraindicated, such as in the postoperative period, if the patient is vomiting, or where gastrointestinal absorption cannot be relied upon. However, the need for continued parenteral therapy should be reviewed regularly. In the treatment of many common infections, the acute features of infection such as temperature, tachycardia, and an elevated circulating neutrophil count usually improve within a period of 48 to 72 h. Provided there is no contraindication to oral therapy, this should be considered early in the course of patient management. The advantages are not just in the reduced cost of medication. The risk of intravenous line associated complications, such as infection, is also eliminated and discharge from hospital may be hastened.

Adverse drug reactions

Overall, antimicrobial agents have an outstanding record of safety. Nonetheless, no drug is without the potential for side-effects. The risk varies by agent, sometimes dose, while host genetic factors and pathophysiological status can also be important.

Oral antibiotics are largely used in the community where they are generally well tolerated and used in the treatment of minor infections in large populations. Injectable agents selected for short course perioperative prophylaxis have a well established safety record. However, agents such as the antiretroviral drugs and amphotericin B carry a higher risk of more serious adverse drug reactions which must be balanced against the life-threatening nature of their target infections.

While drug safety is assessed during drug development, the full repertoire of adverse reactions becomes apparent only during widespread clinical use, hence, the importance of adverse drug reaction reporting systems. In the United Kingdom, the 'yellow card' system has been very successful and relies on voluntary reporting of

possible adverse drug events to the Medicines Control Agency by physicians and, more recently, pharmacists. It is important to distinguish between adverse event reporting and adverse drug reaction reporting. The latter is more difficult to establish with certainty and may require rechallenge which raises medical and ethical concerns.

It is essential to enquire about previous drug reactions as well as other forms of drug toxicities before prescribing. The relationship to a previously prescribed drug requires careful assessment. Hypersensitivity is among the more common of drug reactions and, in the case of b-lactam drugs, appears to be more a function of the five-membered thiazolidine ring (Fig. 7) of the penicillin molecule since hypersensitivity reactions are less common with the cephalosporins which have a six-membered dihydrothiazine ring. The monobactam, aztreonam, has neither ring structure and hypersensitivity reactions appear to be rare. However, it is important to note that accelerated systemic hypersensitivity reactions (anaphylaxis) can be life-threatening such that any previous association with a b-lactam drug is an absolute contraindication to the use of all b-lactams.

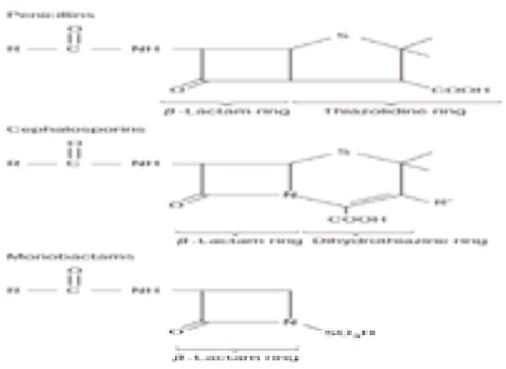

Fig. 7 Chemical structure of the b-lactam antibiotics (penicillins, cephalosporins, and monobactams) identifying the common b-lactam ring component which is subject to hydrolysis by b-lactamases.

Some drug toxicities are genetically determined. For example people who are genetically slow acetylators of isoniazid are more at risk of side-effects such as peripheral neuropathy. Those genetically deficient in the enzyme glucose-6-phosphate dehydrogenase are at risk of drug-induced haemolysis. This risk is more common in those of African, Mediterranean, or Far Eastern descent. Hence, it is important to screen for this red cell enzyme deficiency before the administration of oxidant drugs such as primaquine.

Adverse drug reactions may not always be acute in their presentation but reveal themselves after prolonged drug exposure. Oral flucloxacillin and co-amoxiclav when administered for several weeks, particularly in the elderly, are more likely to induce drug-associated hepatotoxicity. Likewise, parenteral formulations of selected drugs may be more toxic than their oral formulation, as is the case with a fusidic acid where prolonged parenteral administration frequently gives rise to hepatotoxicity.

Concentration-dependent adverse reactions (Table 14) are more likely to occur in the presence of organ system failure. Aminoglycoside toxicity is more common in the elderly, in those with pre-existing renal failure, and after repeated aminoglycoside doses or other nephrotoxic drugs. Concentration-dependent bone marrow suppression characterizes the use of chloramphenicol whereby pancytopenia arises when plasma concentrations are in excess of 25 mg/l. This is to be distinguished from the idiopathic aplastic anaemia that is a rare accompaniment of chloramphenicol use, but unfortunately is rarely reversible.

Much has been learned about the structure activity determinants of drug toxicity. For example the quinolone antibiotics as a class have the potential to induce phototoxicity, arthrotoxicity, CNS toxicity, cardiotoxicity, and interact with agents such as caffeine, theophylline, and non-steroidal anti-inflammatory drugs (Fig. 8). Knowledge of such predictors has lead to the selection of agents with safer structural profiles. Despite this, adverse drug reactions have led to the withdrawal or modification of the licensed indications for several quinolones, notably temafloxacin, trovafloxacin, sparfloxacin, and grepafloxacin, emphasizing the importance of clinical recognition and reporting of adverse events.

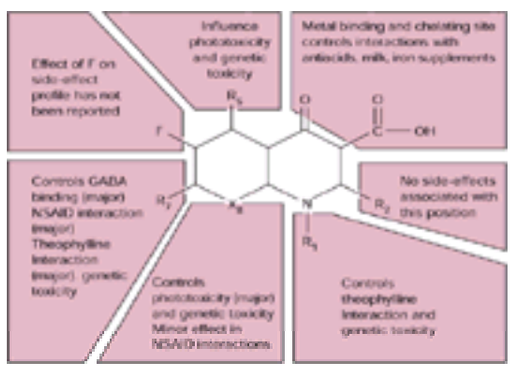

Fig. 8 Structure:activity side-effect relationships of the fluoroquinolone antibacterial drugs (redrawn after Domagala 1994).

Few infectious conditions require life-long therapy. The management of HIV infection has challenged this tenet. To date, drugs directed at the causative viruses or complicating opportunistic infections are suppressive rather than achieving eradication. It is also important to note that the drugs used in the treatment of HIV and AIDS are often licensed with limited information concerning their long-term safety. The potential for adverse reactions and especially interactions is considerable and requires careful attention to their detection and management. This has become an increasingly important challenge as life expectancy for those with HIV infection improves. It is important to balance drug safety while encouraging compliance and the maintenance of a reasonable state of health.

Failure of antibiotic therapy

Antimicrobial therapy may fail for a number of reasons. The agent selected may be inappropriate for the particular infection and fail to inhibit the target organism, or fail to reach the site of infection in sufficient concentration. For example drugs such as nitrofurantoin and norfloxacin, while achieving high urinary concentrations, fail to deal adequately with parenchymatous infection of the kidney or bacteraemia which may complicate acute pyelonephritis.

The prostate also presents a chemotherapeutic challenge owing to the relatively low pH, of about 6.4, in chronic bacterial prostatis. Drugs which are weak bases, such as trimethoprim either alone or in combination with sulfamethoxazole (co-trimoxazole), are preferred, especially since they are also lipid soluble. Ciprofloxacin has similar characteristics and has also produced favourable results. However, treatment of acute bacterial prostatitis sometimes needs to be prolonged (4–6 weeks and occasionally for longer duration), especially if there is a history of chronic relapsing infection.

The drug may be appropriate, but the dose selected may be inadequate. This may apply to such conditions as unsuspected bacterial endocarditis where high-dose parenteral antibiotic is required. Likewise, the concentration of penicillin required to deal with pneumococcal meningitis greatly exceeds that effective in the treatment of pneumococcal pneumonia; occasionally the two diseases may coexist. Infections caused by *Legionella pneumophila* and *Chlamydia* spp. require drugs that achieve high intracellular concentrations such as the macrolides, tetracyclines, or quinolones.

Resistance emerging during treatment is an uncommon cause of clinical failure but should be considered. Drug resistant *Mycobacterium tuberculosis* can develop on therapy as a result of the emergence of minority populations of organisms resistant to such first-line drugs as rifampicin and isoniazid. The current multidrug regimens are, in part, designed to avoid this occurrence. Likewise, in those with HIV infection, drug resistant virus is an increasingly important cause of treatment failure and

requires good compliance with multidrug regimens to slow its rate of emergence.

Mixed infections are commonly associated with intrabdominal sepsis and occasionally with infections of the lung. They may fail to respond to treatment unless the regimen covers the full range of bacterial pathogens. In the case of intrabdominal sepsis, the regimen should be active against anaerobic as well as aerobic bacterial pathogens.

Another important cause of antibiotic failure is the continued presence of a focus of infection. This may be an abscess which requires surgical drainage or the removal of an implanted medical device such as an intravascular catheter. Much more serious is infection of a prosthetic heart valve, hip joint, or CNS shunt where revision surgery carries significant risks. Many antibiotics fail to achieve therapeutic concentrations within abscess cavities, or are pH sensitive. Implant-associated infections present a similar challenge since bacteria often replicate slowly within a biofilm that is protective against normal host defences.

Finally, it should be remembered that a persistently elevated temperature in the presence of what appears to be adequate antibiotic treatment can reflect drug fever or indeed fever complicating a non-microbial diagnosis. This emphasizes the importance of monitoring the response to treatment and repeated patient assessment.

Practice guidelines and formularies

The plethora of therapeutic agents currently available presents a considerable challenge to the prescriber. Guidance on the choice of agent and the management of disease is becoming increasingly important. This is not only to ensure that the selection of treatment is appropriate for the target infection and consistent with current patterns of antimicrobial susceptibility but that it reflects an acceptable safety profile as well as being sensitive to the appropriate use of health-care resources. Such guidance is increasingly provided within formularies designed for local use, either within a hospital or community practice. These frequently offer information on preferred and alternative regimens for particular infections. Formularies should include drugs currently tested by the diagnostic laboratory since changing patterns of susceptibility may require modification of recommended drugs.

Within hospital practice, it is common for such formularies to identify drugs which may be prescribed freely according to specific indications, and those for which expert advice from a clinical microbiologist or infectious disease specialist should be sought. The latter applies particularly to drugs that require specific skill and experience in their use, need drug levels to be monitored, or are expensive. For example the treatment of deep-seated fungal infections with amphotericin B requires careful clinical assessment and guidance on dosage and monitoring. Likewise, the treatment of HIV infection is increasingly a specialist area. Antibiotics which are expensive to prescribe such as parenteral quinolones, third generation cephalosporins, and the carbapenems may be restricted. The policy may also have recommendations for the timing of transfer from parenteral to oral therapy in order to minimize the use of injectable agents.

Formularies are educational and allow the prescriber to become familiar with indications and safety of the most commonly used agents. Their use should be supported by educational activities both at undergraduate and postgraduate level.

Ideally, the selection of agents for inclusion in the formulary should be based on sound evidence of efficacy, safety, and economic benefit. However, such evidence-based medicine is often lacking or incomplete for commonly treated infections, since clinical trials of antibiotics, although increasingly robust in their design, are largely conducted to support licensing requirements rather than to address clinical use. They generally demonstrate the equivalence of a new agent in comparison with existing therapies. As a result, the recommendations of formularies and practice guidelines are based on a matrix of information derived from knowledge of the *in vitro* profile of an agent, its pharmacokinetic parameters, its clinical and microbiological efficacy, and safety profile. This in turn is modified by custom and practice which explains why there is local and, sometimes, national and international variation in recommendations for some common indications such as community-acquired pneumonia and bacterial meningitis.

In developing countries where medical resources are much more limited, greater reliance is placed on low-cost agents. The World Health Organization regularly updates its list of recommended essential drugs which includes anti-infective agents (Table 15). Despite the emphasis on low cost agents, the drugs offered cover the majority of infections and prescribing needs of developing countries. The agents available in individual countries often vary according to local interpretation of the needs for these 'essential' drugs.

Recent developments in economically advanced countries have included an assessment of health-care technologies for current management, national need, and the resources available. In the United Kingdom, the National Institute of Clinical Excellence (NICE) was established in 1999 to assess a variety of health-care technologies including procedures as well as new therapies. At the ime of its initial assessment by NICE, zanamavir, a recently licensed drug effective in the treatment of influenza virus infection, lacked sufficient evidence of efficacy in people known to be at higher risk of complicated influenza virus infection. Such assessments place greater emphasis on ensuring that new technologies are evaluated in a manner that more closely resembles clinical practice as well as demonstrating economic benefit, in contrast to drug licensing which addresses the quality, safety, and efficacy of new therapies. This new emphasis is likely to require a greater partnership between health-care systems and pharmaceutical companies to ensure that the place of new technologies is not only rapidly assessed but that they are consistent with health-care strategies.

Further reading

Bennett WM, Aronoff GR, Golper TA, Morrison G, Brater DG, Singer I (1994). *Drug prescribing in renal failure: Dosing guidelines for adults*, 3rd edn. American College of Physicians, Philadelphia.

Combined Working Party (1998). Revised guidelines for the control of methicillin-resistant *Staphylococcus aureus* infection in hospitals. *Journal of Hospital Infection* **39**, 253–90.

Davey PG, Parker SE, Malek MM (1993). Pharmacoeconomics of antimicrobial prophylaxis. *Journal of Antimicrobial Chemotherapy* **31** (Suppl. B), 107–18.

Domagala JM (1994). Structure–activity and structure–side-effect relationships for the quinolone antibacterials. *Journal of Antimicrobial Chemotherapy* **33**, 685–706.

Finch RG and Williams RJ (1999). *Bailliére's clinical infectious diseases: antibiotic resistance*. Bailliére Tindall, London.

Joint Tuberculosis Committee of the British Thoracic Society (1998). Chemotherapy and management of tuberculosis: recommendations. *Thorax* **53**, 536–48.

Kerr KG (1999). The prophylaxis of bacterial infections in neutropenic patients. *Journal of Antimicrobial Chemotherapy* **44**, 587–91.

Kucers A, Crowe S, Grayson ML, Hoy J (1997). *The use of antibiotics*, 5th edn. Butterworth Heinemann, Oxford.

Macfarlane J, *et al*. (2001). The British Thoracic Society Guidelines for the Management of Community Acquired Pneumonia in Adults. *Thorax* (in press).

O'Grady FW, Lambert HP, Finch RG, Greenwood D (1997). *Antibiotic and chemotherapy*, 7th edn. Churchill Livingstone, Edinburgh.

Raviglione MR, Snider DE, Kochi A (1995). Global epidemiology of tuberculosis—morbidity and mortality of a world-wide epidemic. *Journal of the American Medical Association* **273**, 220–6.

Read RC and Finch RG (1994). Prophylaxis after splenectomy. *Journal of Antimicrobial Chemotherapy* **33**, 4–6.

Russell AD and Chopra I (1996). *Understanding antibacterial action and resistance*, 2nd edn. Ellis Horwood, London.

Shyrock TR, Mortensen JE, Baumholtz M (1998). The effects of macrolides on the expression of bacterial virulence mechanisms. *Journal of Antimicrobial Chemotherapy* **41** 505–12.

Simmons NA (1993). Recommendations for endocarditis prophylaxis. *Journal of Antimicrobial Chemotherapy* **31**, 437–8.

Standing Medical Advisory Committee Subgroup on Antimicrobial Resistance (1998). *The path of least resistance*. Department of Health, London.

Wenzel RP and Edmond MB (1998). Vancomycin-resistant *Staphylococcus aureus*: infection control considerations. *Clinical Infectious Diseases* **27**, 245–51.

White NJ (1992). Antimalarial pharmacokinetics and treatment regimens. *British Journal of Clinical Pharmacology* **34**,1–10.

Wise R and Honeybourne D (1999). Pharmacokinetics and pharmacodynamics of fluoroquinolones in the respiratory tract. *European Respiraatory Journal* **14**, 221–9.

7.7 Immunization

D. Goldblatt and M. Ramsay

Introduction

Infectious diseases remain a major cause of mortality and morbidity worldwide. The prevention of certain infectious diseases by effective immunization programmes represented one of the major triumphs of twentieth-century medicine. Most of this was achieved in the final third of that century during which rapid strides in the understanding of the biology and causality of infectious agents and improved techniques for the purification of infectious agents or their components led to the development of safe and effective vaccines. The greatest triumph in the field of immunization was the eradication of smallpox. In 1959 the World Health Organization (**WHO**) declared its intention to eradicate smallpox, and in 1966 began to allocate sufficient resources to accomplish this ambitious goal. Thirteen years later, in 1979, the global eradication of smallpox was officially declared. Effective vaccines can eliminate infectious diseases, but to do this they must be implemented and used appropriately. Over 12 million children under the age of 5 years die annually. Two million of these deaths are from diseases that could be prevented by vaccines already available through the WHO's Expanded Programme of Immunization (**EPI**). While rapid advances in vaccine science have introduced new techniques such as DNA vaccines, delivering vaccines to those most at risk must remain a priority.

Immunology of active immunization

Both non-specific (innate) and specific adaptive immune systems are responsible for protecting humans against infectious diseases. The ability of the adaptive immune system to refine its antigen recognition domains and establish immunological memory is the basis of successful active immunization. The specific immune system contains both cellular and humoral elements whose relative importance differs depending on the nature of the infecting organism. Cell-mediated immune responses depend on T lymphocytes and their secreted factors derived from the thymus, while humoral responses involve B lymphocytes derived from the bone marrow which produce antibodies (immunoglobulins IgG, IgM, IgA, IgD, or IgE).

Cellular responses are induced when antigen-presenting cells, such as dendritic cells, present antigens to T cells. T cells do not respond to soluble, unmodified antigens but only recognize peptide antigens in association with self major histocompatability complex (**MHC**) molecules. Two major forms of MHC molecules exist. The majority of nucleated cells express MHC class I molecules, which stimulate a subset of T cells expressing the CD8 differentiation antigen. These T cells recognize and lyse infected target cells, hence their designation as cytotoxic T lymphocytes. In contrast, MHC class II molecules are expressed on cells that participate in the immune response, and are recognized via a subset of T cells expressing the CD4 differentiation antigen. A major role of such T cells is to augment the immune response and so they are known as T helper cells. At least two subsets of T helper cells have been described: T helper 1 cells are involved in cytotoxic and delayed hypersensitivity type responses, while T helper 2 cells support antibody production.

Immunoglobulin receptors on the surface of B cells are able to recognize soluble antigens and so initiate the process of B-cell activation and differentiation. During differentiation, naïve B cells become antibody secreting plasma cells. In addition, there is endocytosis of antigen bound to surface immunoglobulin, and processed antigen in the form of small peptides is re-expressed on the surface of the B cell in the context of MHC class II molecules. Thus B cells act as antigen-presenting cells and recruit T-cell help. The signals and soluble factors that result from such T-cell help drive the B-cell process of affinity maturation and memory formation. This takes place in the germinal centres of lymph nodes where there is intimate contact between B cells, T cells, and dendritic cells. It is here that memory B cells are formed and then migrate to the bone marrow, spleen, and the submucosa of the respiratory tract and gut. On re-encounter with antigen, memory B cells undergo rapid activation and differentiation into plasma cells and secrete large amounts of switched, high-affinity antibody.

Thus, the ideal vaccine antigen will lead to the activation, replication, and differentiation of T and B lymphocytes. Ideally, the antigen will persist, conformationally intact, in lymphoid tissue to allow the continuing production of cells that secrete antibody of high affinity and the generation of memory cells.

Vaccine antigens

The ideal vaccine antigen is safe with minimal side-effects, promotes effective resistance to the disease (although it does not necessarily prevent infection), and promotes immunity that is lifelong. It needs to be stable and remain potent during storage and shipping and also has to be affordable to allow widespread use. Most currently licensed vaccines contain live or killed bacterial or viral constituents, bacterial polysaccharides, or bacterial toxoids (Table 1).

Live vaccines are ideal for certain diseases as replication in the body mimics natural infection thereby inducing appropriate and site-specific immunity. Live vaccines must be attenuated to produce the beneficial effects of inducing immunity without the danger of clinical disease. Some live vaccines may be spread from person to person and thus enhance herd immunity although such spread may endanger immunocompromised individuals in whom live vaccines should be avoided. Live vaccines are inherently less stable than killed vaccines and the possibility of reversion of vaccine virus to wild type exists (as in polio). Killed vaccines do not carry the risk associated with person-to-person spread and are inherently more stable, but often require two or three doses to induce optimal immunity, especially when used in the first year of life.

New developments in vaccine antigens

Developments in molecular biology have begun to revolutionize the field of vaccine science and provide a glimpse of the future when traditional reliance on live attenuated viral vaccines or purified bacterial or viral products as vaccine antigens may be reduced. The first licensed vaccine to contain recombinant genetic material was the hepatitis B vaccine. Despite the licensing of highly effective plasma-derived hepatitis B vaccines in the early 1980s, fears about safety and their high cost led to the search for other hepatitis B vaccines. Several vaccine manufacturers used recombinant DNA technology to express hepatitis B surface antigen in other organisms, which has led to the development of new vaccines.

Recent developments have focused on the use of DNA as a vaccine antigen. The utility of naked DNA as a vaccine antigen was discovered by chance in 1989 during a gene therapy experiment when it was shown that a gene inserted directly into a mammalian cell could induce the cell to manufacture (express) the protein encoded by that gene. In early experiments, DNA was injected directly into muscle and the resulting immune response was measured (see Fig. 1).

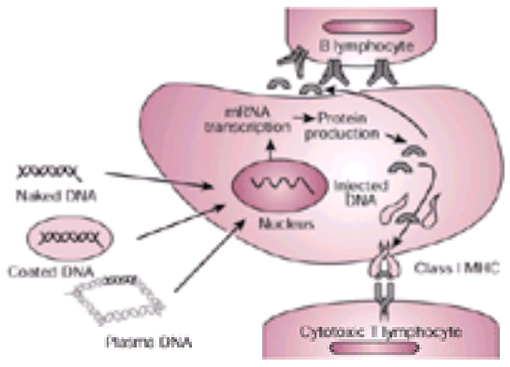

Fig. 1 Injection of DNA encoding a foreign protein can elicit antibodies and a cytotoxic T-lymphocyte response.

DNA vaccines can induce protective immunity in animals to a variety of pathogens, but data in humans are limited. As DNA has the theoretical potential to be incorporated into the host genetic makeup and subvert the genetic working of cells, safety concerns have delayed studies in humans. Phase I studies, however, have assessed DNA vaccines designed to protect against hepatitis B, herpes simplex type 1 and 2, HIV, influenza, and malaria. So far clinical trials have proved disappointing, either because the level of the response was inadequate or because excessive doses of DNA were required to achieve an adequate response. To improve the response to DNA vaccines, a number of newer techniques have been developed. These include:

1. Incorporation of DNA into microprojectiles that are then shot into the target cell via the skin (the so-called 'gene gun' technique).
2. The coating of DNA with cationic lipids or other material that neutralizes its charge; the lipids facilitate cellular uptake and membrane transfer.
3. Delivery of DNA by incorporating it into a viral delivery system using disabled viruses.
4. Delivery of DNA by incorporation into a bacterial delivery system such as attenuated *Salmonella typhimurium*.
5. Delivery of DNA together with traditional adjuvants such as alum.
6. Improving immunogenicity by including a cytokine gene in the plasmid, adjacent to the gene encoding the protective antigen. Local expression of the appropriate cytokine (for example granulocyte–macrophage colony stimulating factor) may augment the immune response in a fashion similar to that seen with adjuvants.
7. Combination of 'priming' immunization with DNA vaccine with subsequent 'boost' with recombinant vaccine.

The huge potential of DNA vaccines, which offer the promise of cheap and stable vaccines that do not require a cold chain for distribution, will stimulate further development of these exciting products.

New developments in vaccine delivery

Research into different routes of vaccine delivery has been driven by the limitations of the parenteral route. These include the difficulty associated with the use of live viral vaccines in the first 6 to 9 months of life (due to the neutralizing effect of passively transferred maternal antibody) and the difficulty and expense of delivering mass immunization by injection. Mucosal delivery of vaccine via the intranasal route has been studied for a number of antigens including measles, influenza, rubella, varicella, and *Streptococcus pneumoniae*. The induction of local immunity for pathogens that either enter the body via the nasopharynx (measles, influenza) or are commonly carried in the nasopharynx (*S. pneumoniae*) is attractive.

Edible vaccines are attracting increasing attention, providing as they do both a means of antigen production and delivery. Studies in animals and Phase I studies in humans have demonstrated their potential. Mice fed with potatoes expressing a non-toxic fragment of the cholera toxin developed mucosal antibodies to the toxin which reduced diarrhoea on challenge with whole cholera toxin. Humans fed raw potatoes expressing the B subunit of enterotoxigenic *Escherichia coli* also showed mucosal immune responses and an increase in neutralizing antibody levels. There are some problems with stability, but edible vaccines are a potentially simple and convenient method of vaccine delivery on a wide scale.

The aim of immunization programmes

Once a vaccine has been developed and shown to be effective it can be used in different ways. Many vaccines are used selectively in groups of the population who are at increased risk of infection (for example because of occupation or travel) or of the severe consequences of the disease that results from infection (because of an underlying medical condition for example). Other vaccines are employed for mass immunization targetting the whole population. Mass immunization can aim to eradicate, eliminate, or to control an infectious disease. Eradication, the state where a disease and its causal agent have been removed from the natural environment, has been achieved only for smallpox. Once eradication has been certified, mass immunization programmes can cease and resources can be transferred to other programmes.

The next target for the WHO is the global eradication of poliomyelitis. Characteristics that favour eradication are the absence of an animal host, the absence of a carrier state, and lifelong protection from vaccination. Poliovirus infection has now been eliminated from the Pan American and Western Pacific WHO Regions although there was a recent resurgence in Haiti. In recently endemic countries, wild poliovirus transmission has been interrupted by a series of National Immunization Days—where live attenuated polio vaccine is delivered to a high proportion of the childhood population on a single day. In 1997, almost 450 million children under 5 years of age were immunized during National Immunization Days. In addition to the resources within each country, this effort has required massive financial support from international donors. Between 1988 and 1998, the number of reported cases of polio worldwide had fallen from 35 251 to 3228 and only three major foci of transmission remain—South Asia and West and Central Africa.

For some infections, eradication by immunization is not possible. A good example is tetanus where the agent is distributed widely in the environment. For these programmes, the aim is to control infection to the point where it no longer constitutes a public health burden. To maintain control, immunization will need to be continued indefinitely.

For diseases that are transmitted from person to person, a good immunization programme provides protection by conferring both individual and herd immunity. For many vaccines, herd immunity can be achieved by vaccinating a high proportion of the childhood population—older individuals are generally immune from previous natural infection. If such a situation can be sustained, transmission of the infection may be interrupted and elimination or eradication becomes possible. If vaccine coverage or efficacy is suboptimal, however, then, in the absence of natural transmission, the number of susceptible people will gradually increase. Eventually the proportion of susceptible people (those who did not receive vaccine or who failed to respond to it) may reach a level sufficient to support an epidemic. Although the size of these epidemics may be small by prevaccine standards, the average age of those infected will be higher than in the prevaccine era. For infections that have more severe consequences in older individuals the morbidity associated with such outbreaks can be substantial. A tragic example of this has been recently observed in Greece where mass vaccination with rubella in childhood has been recommended since 1975. Implementation was poor, however, and during the 1980s coverage was below 50 per cent. The low level of coverage, however, was sufficient to interrupt transmission for several years. By the time rubella infection recurred in 1993, a high proportion of pregnant women were susceptible to rubella and an epidemic of congenital rubella syndrome occurred.

The Expanded Programme of Immunization

In 1974, the WHO, in recognition of the major contribution of vaccines to public health, launched the EPI. At the start of the programme fewer than 5 per cent of the world's infants were immunized against the six target diseases—diphtheria, tetanus, whooping cough, polio, measles, and tuberculosis. Between 1990 and 1997, around 80 per cent of the 130 million children born each year were immunized by their first birthday—preventing around 3 million deaths each year. Each year, over 500 million immunization contacts occur with children and these have provided an opportunity for the delivery of other primary health care interventions.

During the 1990s, EPI has added immunization against yellow fever and hepatitis B to its target (see Table 2) The introduction of these vaccines, however, has been less impressive, particularly in those poorest countries in greatest need. Of 33 African countries at risk of yellow fever, only 17 have included the vaccine in the childhood schedule. By 1998, hepatitis B vaccine had been incorporated into the national programmes of 90 countries, but it is estimated that 70 per cent of the

world's hepatitis B carriers live in countries without programmes. The major barrier to using new vaccines in the developing world is likely to be sustainable funding.

Delivery of immunization programmes

For mass immunization to achieve its aims, high and uniform coverage of immunization must be reached and sustained. Coverage of immunization is associated with a variety of factors including sociodemographic characteristics of the population, organization of health services, knowledge among health professionals, and parental attitudes.

Sociodemographic factors that may influence vaccine coverage include deprivation, maternal education, and family size. Centrally co-ordinated health services with few barriers to access and standard record systems with facilities for call and recall are likely to achieve higher vaccine coverage. Health professionals with accurate knowledge of the indications and true contraindications to immunization are important. Excessive lists of contraindications for DTP immunization in the newly independent states of the former USSR contributed to a massive resurgence of diphtheria in the early 1990s. The number of cases rose from 2000 in 1990 to over 47 000 in 1994; 2500 deaths from diphtheria occurred between 1990 and 1995.

Whether or not parents decide to have their children vaccinated depends on their perceptions of the severity of the disease and of the safety and effectiveness of the vaccine. Knowledge of parental perceptions can be used successfully to target health promotion campaigns. When coverage is high, the incidence of vaccine preventable disease declines and parental perception of the severity of that disease may decrease. In this situation, concerns about the safety of vaccine become paramount and can lead to a decline in vaccine coverage. Such a situation arose in the United Kingdom in the early 1970s when concern about the safety of pertussis vaccine led to a fall in vaccine coverage. This resulted in resurgence of disease with consequent mortality and morbidity (see Fig. 2). Over the next decade, vaccine coverage improved again and the incidence of disease fell to the lowest levels ever.

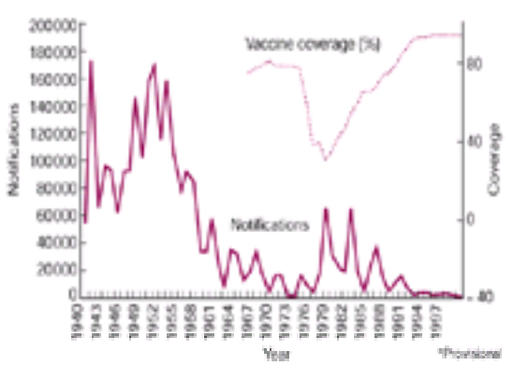

Fig. 2 Whooping cough cases and vaccine coverage in England and Wales between 1940 and 1998.

A more recent example of where largely unsubstantiated concerns about vaccine safety have damaged the success of a vaccination programme have occurred in France. Anecdotal reports of multiple sclerosis following hepatitis B vaccine produced a fall in vaccine coverage in France. Although the number of reports were in line with expected numbers, in October 1998, the French Ministry of Health decided to suspend the schools programme but to continue immunization in general practice. Three case–control studies failed to confirm the link and it is clear that the risk of hepatitis B infection always outweighs the potential risk of adverse events. The WHO and the national committees of several neighbouring countries have reviewed the data and concluded that there is no evidence of a causal association. Despite this, coverage of hepatitis B vaccine has continued to fall in France and in French-speaking Belgium.

In the United Kingdom, speculation that the combined measles–mumps–rubella (**MMR**) vaccine produced bowel problems which in turn could result in autism was published in *The Lancet* in February 1998. This anxiety was based upon a case-series of only 12 patients and the validity of the study was questioned in an accompanying editorial. Subsequent population-based studies of children with autism in the United Kingdom have failed to demonstrate a link with measles-containing vaccines - a finding which is supported by data from Finland. Despite the lack of scientific evidence, a majority of mothers in the United Kingdom now believe the MMR vaccine to be unsafe. This, together with a low proportion of mothers who believe that measles is a severe disease, has resu1ted in a small but sustained fall in MMR vaccine coverage in the United Kingdom.

Evaluation of immunization programmes

Evaluation of an immunization programme may include the measurement of vaccine coverage, surveillance of disease incidence, assessment of prevalence of immunity, and the monitoring of adverse events.

Vaccine coverage

Timely measurement is important for monitoring trends in vaccine coverage and to identify pockets of low coverage. Low coverage may be apparent before any increase in disease incidence is observed. Since the late 1970s, three outbreaks of poliomyelitis have been observed amongst groups with religious objection to immunization in The Netherlands. Despite national coverage of 96 per cent for MMR vaccine, the same group has recently been the focus of a large epidemic of measles. Between April and December 1999, 1750 cases of measles occurred in The Netherlands compared with only nine in the whole of 1998.

Disease surveillance

Once an immunization programme has been implemented, disease incidence data can be used to monitor the effectiveness of that strategy. For example, the dramatic decline in the incidence of invasive *Haemophilus influenzae* infection described in both The Netherlands and the United Kingdom can be used to demonstrate the impact of conjugate vaccination. The age distribution of infection may change as children above or below the target age will form an increasing proportion of those infected. Various epidemiological methods, including case–control studies, cohort studies, and the screening method, can be used to estimate the efficacy of the vaccine in the field.

Seroprevalence studies

Seroprevalence studies are used to assess population immunity to infection. Such immunity results either from immunization or from natural infection. This can detect groups including a high proportion of susceptible individuals who may be the focus of future outbreaks. In 1991, seroprevalence studies in the United Kingdom identified that a large proportion of school-age children was susceptible to measles and therefore that an epidemic of measles was likely. A large campaign was mounted to immunize children from 5 to 16 years of age in November 1994. The number of cases of measles fell rapidly and remained at low levels over the next 5 years.

Adverse events

Monitoring of adverse events is important in maintaining public confidence in an immunization programme and for detecting rare events that could not be identified before licensing the vaccine. The detection of such events may lead to the withdrawal of certain vaccines. In August 1998, rotavirus vaccine was licensed for use in the United States and recommended for mass immunization of infants. During prelicensing studies, five cases of intussusception had been reported in around 10 000 recipients, compared with only 1 in almost 5000 controls; this difference was not statistically significant. During postlicensing surveillance, however, 15 cases were reported to the Vaccine Adverse Event Reporting System. On 22 October 1999, a review of scientific data concluded that there was an increased frequency of intussusception in the 1 to 2 weeks after vaccination which led to withdrawal of the vaccine in the United States.

Further reading

Chen RT (1999). Vaccine risks: real, perceived and unknown. *Vaccine* **17**, S41–S46.

Czerkinsky C *et al.* (1999). Mucosal immunity and tolerance: relevance to vaccine development. *Immunological Reviews* **170**, 197–222.

Leitner WW, Ying H, Restifo NP (1999). DNA and RNA-based vaccines: principles, progress and prospects. *Vaccine* **18**, 765–77.

Orenstein WA, Bernier RH, Hinman AR (1988). Assessing vaccine efficacy in the field. Further observations. *Epidemiologic Reviews* **10**, 212–41.

Tacket CO *et al.* (1998). Immunogenicity in humans of a recombinant bacterial antigen delivered in a transgenic potato. *Nature Medicine* **4**, 607–9.

World Health Organization and the United Nations Children's Fund (1996). *State of the world's vaccines and immunization*. WHO, Geneva.

World Health Organization (1997). *Polio. The beginning of the end*. WHO, Geneva.

7.8 Travel and expedition medicine

C. P. Conlon and D. A. Warrell

United Kingdom citizens make 56 million visits abroad each year, 8 per cent of these to developing countries which carry a higher risk of illness (600-fold risk in Mexico, 1835-fold in the Indian subcontinent) than travel to continental Europe (for example, to France).

Pretravel advice

This can be obtained from medical practitioners interested in travel medicine, embassies of the countries to be visited, travel agencies, organizations, specialist travel clinics, and the internet (see below). Members of immigrant communities in Western countries, especially from the Indian subcontinent and West Africa, are vulnerable to endemic diseases, including malaria and typhoid, when they return on holiday to their country of origin, perhaps to visit their families. This group of travellers is less likely to receive good pretravel advice and perhaps less willing to seek or accept it. A certificate of vaccination against yellow fever may be needed for entry to some countries. Details of other immunizations, allergies, blood group, and regular medications should also be carried by the traveller. Adequate insurance is essential. The geographical area to be visited, the age and health of the traveller, and any special risks of the journey (for example, mountain climbing) are taken into account. In remote areas or those with inadequate health facilities the travel insurance policy must cover repatriation.

General advice about health

The basic first-aid kit should include: a topical antiseptic solution; bandages; plasters; proprietary drugs for pain relief, diarrhoea, dyspepsia, allergy, and itch; sunscreen preparations; water purification tablets; and insect repellents.

For motion sickness, antiemetic drugs such as cyclizine are effective, but they may cause sedation and a dry mouth. Long-acting transdermal skin patches containing scopolamine are preferable. Long-haul air flights lead to jet lag: sleep disturbance, fatigue, a feeling of light-headedness and unreality, and poor concentration. These symptoms may be attributable to a 'hangover' if excessive alcohol has been drunk on the flight! A short-acting benzodiazepine, such as temazepam, taken for the first couple of nights after flying, helps to re-establish a regular sleeping pattern. Some travellers have found that melatonin is helpful (Chapter 12.13), but a recent trial demonstrated no efficacy for this product. People with diabetes may need advice on adjusting their insulin regimen or diet for changes in time zones.

Climatic and environmental extremes

At high altitudes, snow blindness and severe sunburn can occur under clear skies even at very low ambient temperatures. Those going to high altitudes should acclimatize slowly and build up their level of physical activity gradually (see Chapter 8.5.4). They should be aware of the symptoms and signs of altitude sickness. Acetazolamide ('Diamox'), in an adult dose of 250 mg twice a day, starting 12 h before starting the ascent, is effective prophylaxis for mild mountain sickness, especially if the traveller has to ascend rapidly (e.g. flying from sea level to more than 3000 m). But gradual ascent, allowing acclimatization is preferable and, if severe symptoms develop, there is no substitute for rapid descent. In the tropics, heat, dehydration, and salt depletion may cause problems. Several days of relative inactivity are needed to acclimatize safely to hot climates.

Strict food and water hygiene are important in countries with relatively poor sanitation. 'Boil it, peel it, or forget it'; is a useful adage for the traveller. Water purification tablets and many types of portable water filters are available. Beverages made with boiled water are generally safe, whereas bottled water and, particularly, ice cubes are unreliable. Treated water should be used even for tooth cleaning.

In many developing countries, blood-borne pathogens, such as hepatitis B and C viruses, human immunodeficiency virus (**HIV**), human T-cell leukaemia/lymphoma virus type 1 (HTLV-1), and, in some areas, malaria, trypanosomiasis, and other infections are prevalent. Screening of donated blood may not be rigorous and needles are commonly reused, sometimes without adequate sterilization. As a result, travellers have been advised to take 'AIDS kits', usually containing needles, cannulas, intravenous giving sets, syringes, and artificial plasma expanders. These are too bulky and expensive for most travellers, but it is worth taking a few 21-gauge needles and 10-ml syringes in case blood must be taken for a laboratory test or an injectable drug is needed. A covering letter from a doctor may allay the suspicion of customs officials that they are to be used for drug abuse.

Travellers seem to become unusually disinhibited and foolish and are particularly likely to engage in promiscuous unprotected sexual activity, especially if they are taking alcohol or other recreational drugs. Since sexually transmitted diseases, including HIV, are highly prevalent in many holiday resorts (not only in prostitutes), good-quality condoms, often not available when travelling, should be carried and used.

Patients with chronic illnesses, such as diabetes or asthma, should take plenty of their current medications as these may not be available abroad. It is a good idea to carry separate supplies in case of luggage loss or theft.

Immunizations

The record of routine childhood immunizations should be reviewed. Many adults will require booster doses for tetanus, polio, and diphtheria.

Yellow fever is only endemic in tropical Africa and South America, not in Asia. Recently, a Belgian tourist acquired fatal yellow fever in The Gambia, emphasizing the continuing importance of this immunization. Cholera vaccine is no longer recommended by the World Health Organization as its adverse effects outweigh its usefulness, but a new oral vaccine is promising. Other immunizations may be recommended after considering the travel itinerary and risk of exposure (Table 1).

The risk of hepatitis A in developing countries ranges from 300 to 2000/100 000 unprotected travellers per month of stay. Active immunization is safe, effective, and durable (see Chapter 7.10.19).

In the 'meningitis belt' of sub-Sahelian Africa, from Senegal east to the Sudan, and in some other areas, dry season meningococcal meningitis outbreaks are so common that immunization is recommended.

Pre-exposure rabies vaccination is being used increasingly. Although the risk of transmission is fairly low, the lack of effective treatment for rabies encephalitis and the fear engendered by a dog bite justifies considering immunization.

Plague vaccine is effective but may give rise to serious side-effects. An alternative in endemic areas is prophylactic or postexposure doxycycline treatment (see Chapter 7.11.16). Anthrax is endemic in many tropical countries, but, despite anxieties raised by its use for bioterrorism in the United States, vaccination or chemoprophylaxis are unnecessary. Japanese (B) encephalitis vaccine is safe and there is a risk of infection in many parts of Asia (see Chapter 24.14.2). Hepatitis B is a risk for medical staff, whose work involves contact with human blood and to those receiving unscreened blood transfusions in some developing countries (see

Chapter 7.10.19). It is also a risk of unprotected sexual activity.

Prevention of malaria

Both travellers and non-specialist physicians must be educated about the prevention and recognition of malaria (see Chapter 7.14.2).

Travellers' diarrhoea (Table 2)

This is the most common health problem of travellers. Symptoms are usually mild, lasting only about 3 to 5 days, but holiday and business plans may be disrupted. The most common cause is enterotoxigenic *Escherichia coli* (**ETEC**). *Salmonella* spp., *Campylobacter* spp., *Shigella* spp., and other pathogenic *E. coli* are also common. Protozoan pathogens, such as *Giardia lamblia, Entamoeba histolytica*, *Cryptosporidium parvum*, and viruses are less common causes. Fish and shellfish poisoning cause similar symptoms starting within minutes or hours of exposure.

Strict food and water hygiene reduces the risk of gastroenteritis. Heating water to 100 °C will kill most pathogens, as will chemical treatment with chlorine or iodine (iodine is contraindicated in pregnant women and some patients with thyroid disease). Water filters are useful additions. Antimicrobials such as co-trimoxazole, doxycycline, and the 4-fluoroquinolones are protective to some extent but are not cheap, may cause side-effects, cannot be taken for prolonged periods, and may encourage antimicrobial resistance. Colloidal bismuth salts are cheaper, safer, and reasonably effective, but the large volumes are inconvenient. An effective vaccine against ETEC may soon be available.

Treatment is by maintaining an adequate fluid intake and using sachets of oral rehydration salts that can be made up with boiled water. Eating solid food may stimulate bowel action by the gastrocolic reflex. Antidiarrhoeal agents, such as codeine phosphate, imodium, or loperamide, often relieve symptoms sufficiently for normal activities to be continued. Short courses of empirical antimicrobials, for example ciprofloxacin (500 mg for 3 days, adults only), can be useful, particularly for patients with underlying diseases. Localized abdominal pain or bloody diarrhoea are indications for seeking medical help immediately.

Immunocompromised travellers

Except for asplenic patients, immunocompromised travellers—including those who have received radiotherapy for lymphomas—should not be given live vaccines such as yellow fever, oral polio, and oral typhoid. Killed or synthetic vaccines are safe. Those patients with mild to moderate immune suppression, including those with early HIV infection, will probably make a reasonable response to immunization; those with more severe immunosuppression may still make a useful, though less durable, response. Influenza, pneumococcal, and *Haemophilus influenzae* b (**Hib**) conjugate vaccines are recommended, as the risk of respiratory infection and bacteraemia is increased. Gammaglobulin is the preferred prophylaxis against hepatitis A in these patients, as the response to hepatitis A vaccine may be unreliable. Asplenic individuals should be on prophylactic antibiotics, such as amoxicillin, particularly if travelling, and should be dissuaded from travelling to areas with high rates of malaria transmission.

Immunocompromised patients should carry antimicrobials with them for treating respiratory or gastrointestinal infections, should seek medical help when abroad, and should carry a letter from their physician outlining their condition and medication.

Pregnant travellers

Commercial airlines will not normally convey a woman who is 36 weeks or more pregnant without a covering letter from her physician. Insurance to cover the costs of delivery abroad should be considered.

The risk–benefit assessment of immunizations and chemoprophylaxis is of particular importance for the pregnant woman and the fetus. Live vaccines should be avoided, but if there is a genuine risk of yellow fever the vaccine should be given as there is no recognized associated teratogenicity. Inactivated polio vaccine may be given parenterally and tetanus immunization is safe. Heat-killed typhoid vaccine is best avoided as it might cause a febrile reaction, stimulating premature labour. However, the modern polysaccharide capsular Vi vaccine should be safe. Pneumococcal, meningococcal, and hepatitis B vaccines are safe in pregnancy, as is gammaglobulin.

Malaria is specially dangerous in pregnant women (see Chapter 7.14.2). Chloroquine and proguanil are safe chemoprophylactic drugs, and quinine, in normal therapeutic doses, is safe for treatment. Mefloquine is best avoided in pregnancy. Pregnant women should take special care with food and drink when abroad, as dehydration may threaten the fetus. There are concerns about congenital goitre when pregnant women use iodine to purify water—the maximum recommended daily intake is 175 µg. Loperamide as an antidiarrhoeal agent is safe, but antimicrobials such as tetracyclines and quinolones should be avoided.

Extremes of age

Young children should have completed their routine immunizations before travelling. Malaria chemoprophylaxis is recommended for all ages. Yellow fever vaccine should only be given to children older than 9 months as a few cases of vaccine-associated encephalitis have occurred in younger children. Most other vaccines, including rabies, are safe. Hepatitis A is rarely symptomatic in children under 5-years old. Families planning to live in developing countries should be offered BCG vaccination for their children to reduce the risk of tuberculous meningitis.

The elderly should have the same immunizations as younger adults and should take antimalarial drugs. They are more prone to respiratory infection and should, therefore, be given influenza, pneumococcal, and *Haemophilus influenzae* vaccines. Jet lag and changes in time zones may be very disturbing. Older people are more likely to have an underlying medical condition requiring medication. It is important that sufficient supplies of medicines are taken abroad and that the patient has a detailed list of these medicines and their dosages, in case the tablets are lost or stolen, and the name and contact address of their physician at home in case of emergency.

Explorers and expeditions

Because of their adventurous aims, expeditions are likely to involve exposure to greater environmental extremes and hazards than ordinary travel. Expeditions usually take place in areas remote even from rural health centres, and so a greater responsibility for dealing with medical problems will devolve on the expedition members. The explorer's greatest fear may be to fall victim to a lethal tropical disease or an attack by a wild animal, but the reality is more mundane: road traffic accidents, mountaineering disasters, drowning, and attacks by humans. Prevention and treatment of medical problems must be planned well in advance. Detailed advice and information can be obtained from a number of organizations, such as the Expedition Advisory Centre of the Royal Geographical Society in London (Tel: 0207–581–2057; Fax: 0208–584–4447), from clubs specializing in mountaineering, cave exploring, diving, and other activities, and from books, journals, and websites. All expeditions should have a designated medical officer and all their members should receive first-aid training, aimed ideally at the particular needs of the expedition. The basics are clearing the airway, controlling bleeding, treating shock, relieving pain, and moving the injured person without causing further damage. Expedition medical kits should be more comprehensive than those carried by ordinary tourists and travellers. Lists of essential drugs are given in Anderson and Warrell (2002). Scissors, and a generous supply of large triangular and crêpe bandages, adhesive plasters, and an 'AIDS kit', to reduce the risk of infection from dirty needles and intravenous fluids are important. Lightweight emergency insulation must be taken if there is any risk of exposure in severe weather conditions, a lightweight collapsible stretcher for mountaineering, and an adequate water supply must be assured or taken if the expedition is into desert areas. A covering letter on official notepaper, signed by a doctor, may be helpful in getting drugs, even apparently innocuous ones such as codeine, through Customs (for example, the Russian Federation) and explaining the need for needles and syringes. The medical facilities nearest to the site of the expedition must be identified and contacted in advance. An emergency plan must be drawn up for the first-aid treatment and evacuation of severely ill or injured expedition members. In some areas, 'Flying doctor' and air evacuation services (such as AMREF in East Africa) are available. Medical insurance must be generous and comprehensive, to include repatriation of the injured. Before leaving their home country, expedition members should have a thorough dental check and treatment for outstanding medical or surgical problems. Control of chronic medical problems such as diabetes mellitus, hypertension, and asthma should be stabilized. In selecting members for an expedition, the most important attributes are experience, possession of the necessary skills (for example, diving and mountaineering), physical fitness, and proven psychological stability under stress. It is advisable always to appoint a reliable local agent in the country where the expedition will take place.

Illness in returning travellers

Details are needed about the countries visited, activities while travelling, immunizations, and antimalarials taken. Common problems are fever, rash, diarrhoea, and eosinophilia (Table 3, Table 4, and Table 5).

In travellers with acute diarrhoea, a dietary history, assessment of hydration state, stool microscopy and culture, abdominal films, and sigmoidoscopy may be needed. There are many possible causes (see Table 2). Patients with chronic diarrhoea may be infected with *Giardia* spp., *Cryptosporidium* spp., *Entamoeba histolytica*, shigellae, or salmonellae. Investigations should include a search for *Clostridium difficile* and its toxin, especially if the patient took antimicrobials while abroad. A minority of patients may develop a postinfective enteropathy, the most common problem being a secondary lactose intolerance. Rarely, bacterial overgrowth or tropical sprue develop.

The commonest causes of eosinophilia are allergy and helminths (see Table 5).

Further reading

Anderson S, Warrell DA (eds) (2002). *Expedition medicine*, 2nd edn. Profile Books, London.

Auerbach PS (1995). *Wilderness medicine. Management of wilderness and environmental emergencies*, 3rd edn. Mosby, St Louis.

Backer HD, *et al*. (eds) (1998). *Wilderness first aid. Emergency care for remote locations*. Jones and Barlett, Boston.

Bradley DJ, Bannister B (2001). Guidelines for malaria prevention in travellers from the United Kingdom for 2001. *Communicable Diseases and Public Health (PHLS)* **4** (2), 82–101.

Dawood R (2002). *Travellers' health*. Oxford University Press, Oxford. [New edition in press]

Department of Health (2001). *Health information for overseas travel*, 2nd edn. London, Stationary Office.

Forgey WW (2000). *Wilderness medicine. Beyond first aid*. Globe Pequot Press, Guilford, Connecticut.

Freedman DO, eds (1998). Travel medicine. *Infectious Disease Clinics of North America* **12**, 249–554.

Journal of Wilderness Medicine (1990–). Published for the Wilderness Medical Society by Chapman and Hall Medical, London.

Milne AH, Siderfin CD, eds (1995). *Kurafid. British Antarctic Survey medical handbook*. British Antarctic Survey, Natural Environment Research Council, Cambridge, UK.

Monath TP, Modlin JF (2002). Prevention of yellow fever in persons traveling to the tropics. *Clinical Infectious Diseases* **34**, 1369–78.

Potter SA, ed. (1992). *Anare Antarctic field manual*, 4th edn. Australian Antarctic Division, Kingston, Tasmania.

Salisbury DM, Begg NT, eds (1996). *Immunisation against infectious disease*. HMSO, London. [New edition in press]

Steedman DJ (1994). *Environmental medical emergencies*. Oxford University Press, Oxford.

Voluntary Aid Societies (1997). *First aid manual*, 7th edn. Dorling Kindersley, London.

Ward MP, Milledge JS, West JB (1989). *High altitude medicine and physiology*. Chapman and Hall Medical, London.

Internet sites

www.cdc.gov/travel/

www.who.int/ith/

www.the-stationery-office.co.uk/doh/hinfo/index.htm

www.isid.org/

http://www.premedmail.org/

www.doh.gov.uk/traveladvice/index.htm

www.istm.org/

7.9 Nosocomial infections

I. C. J. W. Bowler

Definitions

Hospital-acquired or nosocomial (Greek voso_oμeiov, hospital) infections are distinct from community-acquired infections. They may affect patients and hospital staff. A useful epidemiological tool for the study of these infections is to define them as any infection manifesting more than 48 h after admission. However, some nosocomial infections may not be so easily identified as hospital acquired; for example hospital-acquired hepatitis B infection may not become clinically apparent until months after the patient has been discharged because of the prolonged incubation period. **Iatrogenic infections** are acquired as the direct consequence of a therapeutic intervention (e.g. insertion of a urinary catheter). **Opportunistic infections** are caused by organisms that do not ordinarily harm healthy people; they occur in people with impaired defences. **Endogenous (autogenous) infections** are produced by the patient's normal flora, while **exogenous infections** result from transmission of organisms to the patient from elsewhere. Although in practice it may not always be possible to distinguish endogenous from exogenous infections, this differentiation must be attempted because of important implications for control. Rapid changes in health-care provision in hospitals mean the frequency and nature of nosocomial infection are also changing. The increasing trend to early discharge, particularly for surgical patients, can lead to an under assessment of the burden of nosocomial infection. New interventions provide new opportunities for infection. For instance flexible endoscopes, which have revolutionized the investigation and management of a wide variety of diseases, can transmit Hepatitis B between patients if they are not decontaminated.

Nosocomial infections are preventable. Systematic surveillance to assess the size of the problem and an organized programme aimed at preventing or minimizing the impact of nosocomial infection should be an important part of the hospital's quality assurance system. Hospital managers must ensure appropriate staffing and resources. The programme involves surveillance, feedback of data on infections to staff, plans for outbreak management, and agreed policies for antibiotic prophylaxis, the management of patients with infections, and for carrying out procedures likely to increase the risk of infection. Staff should be educated, through an organized teaching programme, and results should be systematically audited.

Scale and costs of nosocomial infections

The World Health Organization (WHO) recognizes the serious global problem of nosocomial infections. Epidemic nosocomial infections frequently receive greater attention because of the alarm caused by the obvious spread of an infectious disease. A common point source or person-to-person spread are usually involved. Immediate application of control measures to prevent transmission often curtails such outbreaks. However, only about 3 per cent of nosocomial infections are accounted for by epidemic infection. Prevention of the other 97 per cent of (endemic) nosocomial infections requires a systematic approach based on careful surveillance. Host susceptibility, the infectious risk of medical procedures, and the type of hospital environment (e.g. intensive care units) are responsible. The risk of endemic nosocomial infections is reduced, for example, by using prophylactic antibiotics for contaminated surgical procedures.

Rates of nosocomial infections between 5.7 and 8 infections per 100 admissions have been reported. The urinary tract, surgical wounds, and the lower respiratory tract are the most common sites, in that order (Table 1). In the United States, it is estimated that, of 200 000 deaths in patients with nosocomial infections, 20 000 were directly attributable to the infection. In a further 60 000, it contributed to death. Costs were estimated at $4.5 billion in the USA in 1992 and approximately £120 million in England and Wales in 1987, based on an average additional stay of 4 days for each hospital-acquired infection. These are likely to be underestimates. The WHO published comparative world-wide costs in 1984 based on an extra 5 days of admission per infection and a minimum cost of a hospital stay, per day, of $45.00 (Table 2). For developed countries, this hospital stay cost is too low.

Host factors

The principal risk factor is the severity of the underlying disease (e.g. neutropenia, organ system failure). In multivariate analysis, the number of medical diagnoses on admission, especially diabetes, renal failure, or alcohol abuse, are most strongly associated with risk. Treatment itself may lower host defences, for example surgical incisions, bladder catheterization, mechanical ventilation, and neutropenia following cancer chemotherapy. Pathogens are able to form biofilms on the increasingly used prosthetic devices (totally implantable, e.g. hip replacement, or transcutaneous, e.g. intravascular devices) subverting normal host clearance mechanisms.

Patients with similar clinical problems, who are likely to share similar risk factors for infection, tend to be nursed together for convenience but the introduction of a micro-organism into such a group can infect a number of patients. A good example is the rapid spread of small round structured virus gastroenteritis in geriatric wards. A poorly maintained hospital environment is a threat to vulnerable patients; an example is outbreaks of legionellosis in units caring for patients with solid organ transplants, resulting from defective ventilation and hot water systems.

Micro-organisms

Bacteria (*Escherichia coli, Staphylococcus aureus, Enterococcus* spp., *Pseudomonas* spp., and coagulase-negative staphylococci, in decreasing order of frequency) are the most important. Viruses, fungi, and protozoa play a minor part.

Whether endogenous or exogenous, the organisms causing nosocomial infection are usually part of a patient's colonizing flora. It may be difficult to distinguish infecting from colonizing organisms using bacteriological tests alone. They are frequently multidrug resistant. Empirical antibiotic therapy must accommodate the shift towards more resistant colonizing flora occurring in hospitals, particularly in burns and intensive care units. For example *Pseudomonas aeruginosa*, methicillin-resistant *Staph. aureus* (MRSA), and enterococci exhibit multiresistance to antimicrobials, making them difficult and expensive to treat.

Principles of hospital infection control

The main goal of hospital infection control is to prevent nosocomial infection. First, hospital-acquired infections must be identified as endemic or epidemic by clinical and epidemiological investigations. The identification and typing of isolates causing nosocomial infection allows recognition of organisms that are epidemiologically linked. Invasive multiresistant organisms, such as MRSA, often require infection control measures to prevent their spread, and so minimize the use of expensive, sometimes toxic, antibiotics required for their prophylaxis and treatment.

Epidemic outbreaks are usually amenable to measures that interrupt the spread of infection, such as use of gowns and gloves and careful hand washing by those attending patients. Transfer of colonized or infected patients to a single room or an isolation ward is a physical means of preventing spread. Patients infected with the same organism can be grouped together and attended to by a cohort of nurses not involved with uninfected patients. Identification of additional carriers and elimination of colonization may be necessary for some epidemic outbreaks. Controlled trials demonstrating the efficacy of such measures have not been made, but

many observational studies support their use.

Endemic nosocomial infections are less straightforward to control. The size of the problem may not be apparent because attack rates in individual units may be low or because some infection is seen as a normal consequence of certain interventions. It is important that information about endemic infections is collected systematically, analysed, disseminated, and discussed so that preventive strategies can be improved. Control measures are applied to selected patients according to risk; for example correctly timed antimicrobial prophylaxis and meticulous sterile technique in prosthetic joint replacement surgery.

Site of nosocomial infections

Urinary tract

A bacterial count of 10^5 organisms or more per ml in cultured urine indicates infection. However, counts as low as 10^2 organisms/ml are included by some. Urinary-tract infection accounts for 30 to 40 per cent of all nosocomial infections. Most patients remain asymptomatic, but 20 to 30 per cent develop the symptoms of urinary-tract infection, about 1 in 100 of whom develop bacteraemia. Causes of nosocomial urinary-tract infections are listed in Table 3.

Indwelling urinary catheters account for 80 per cent of nosocomial urinary-tract infections; 50 per cent of patients catheterized for longer than 7 to 10 days develop bacteriuria. Most of the others result from instrumentation of the urinary tract. The main source of organisms is the periurethral flora. Bacteria gain access to the bladder, usually by spreading up the outside of the lumen of the catheter. Occasionally, infection is acquired exogenously during an epidemic of nosocomial infection. Most symptomatic or bacteraemic infections occur within 24 h of the organisms gaining access to the bladder. Early recognition, by daily urine culture, of a urinary-tract infection before it becomes symptomatic is not helpful.

Treatment is with broad-spectrum antimicrobials administered empirically after obtaining appropriate cultures and later adjusted after receiving results of bacteriological studies. Asymptomatic patients need not be treated.

Since the important risk factor is the duration of catheterization, prevention is by avoiding catheterization or reducing the period of catheterization. Catheters should be inserted aseptically, and closed sterile drainage systems, uninterrupted gravity drainage, or intermittent or suprapubic catheterization employed. Prophylactic antimicrobial treatment is not useful.

Surgical wound infection

One acceptable definition requires the presence of a purulent discharge in, or exuding from, a wound. *Staph. aureus* (15–33 per cent of all wound infections) and *E. coli* (12–19 per cent) are leading causes. Many other aerobic and anaerobic bacterial may be implicated.

The main risk factor is the degree of wound contamination at operation. Operations may be 'clean' (e.g. herniorrhaphy), 'clean contamination' (e.g. appendicectomy which requires incision of bowel), or 'contaminated' (e.g. gross spillage from the gastrointestinal tract during surgery). *Staph. aureus* causes most infections complicating clean surgery. 'Contaminated' surgery is associated with polymicrobial infections, especially with *E. coli* and mixed anaerobes. Other risk factors include the length of the operation, obesity, a remote infection, and underlying disease. Most wound infections follow direct inoculation of organisms into the wound at surgery or spread of bacteria to open wounds such as burns.

Wound infections present with local symptoms and signs (pain, erythema, pus, dehiscence) with general features of infection such as fever. Appropriate cultures, including blood cultures, are taken, pus is drained, and broad-spectrum antimicrobials are given empirically, directed at the likely flora but later adjusted according to bacteriological results. Prevention is by meticulous aseptic surgical technique. Prophylactic antimicrobials, given no more than 2 h before the surgical incision, are indicated for clean-contaminated and contaminated procedures, and in clean surgery when a prosthesis is inserted (e.g. vascular grafting).

Nosocomial pneumonia

Pneumonia is defined clinically by production of purulent sputum, chest signs, a fall in arterial PO_2 and the appearance of new infiltrates on the chest radiograph not ascribable to pulmonary emboli, collapse, or pulmonary oedema. Between 0.55 and 1.5 per cent of patients admitted to hospital develop lower respiratory-tract infections. Crude case fatalities of between 20 and 30 per cent are quoted but death may be due to underlying disease. Intubated and ventilated patients have the highest risk of acquiring pneumonia. Bacteria colonizing the gastrointestinal and upper respiratory tracts are probably aspirated. This flora is often acquired after admission to hospital. Organisms cultured from bronchoscopic samples are listed in Table 4.

Culture of expectorated sputum or tracheal aspirate is poorly predictive of the bacterial cause of nosocomial pneumonia, which is best determined by quantitative culture of specimens obtained by sampling the terminal airways (e.g. by bronchoalveolar lavage). Initially, broad-spectrum antimicrobials appropriate for likely infecting flora should be given empirically. Once the susceptibility of the causative pathogen has been determined, specific antimicrobial treatment can be instituted. Selective decontamination of the digestive tract has reduced occurrence of nosocomial pneumonia but there has been no reduction in the mortality of ventilated patients. Epidemic nosocomial pneumonia usually results from bacterial contamination of respiratory equipment such as nebulizers, ventilators, or bronchoscopes and can be prevented by cleaning and disinfection of the equipment and hand washing after patient contact.

Intravascular device-associated infections

The most important result of intravascular device-associated infection is bacteraemia, varying in incidence from about 0.04 per cent for subcutaneous central venous ports, to about 0.2 per cent for peripheral intravenous cannulae, and approximately 10 per cent for central venous haemodialysis catheters.

Duration of intravascular cannulation is the greatest risk factor. Bacteria usually gain access to the blood by direct spread from the skin surface along the subcutaneous catheter tunnel to its tip in the blood vessel. Bacteraemia from intraluminal bacteria results from contamination of connecting devices. This is particularly important in catheters with subcutaneous cuffs, such as Hickman catheters, where the periluminal route of infection is less likely. The leading organisms causing intravenous device-related sepsis are *Staph. aureus, Pseudomonas* spp., and *Candida* spp. In patients with haematological malignancies, coagulase-negative staphylococci and enterococci are also frequently implicated.

Line-related sepsis presents with local inflammation or signs of thrombophlebitis but usually with features of bacteraemia. Blood cultures are obtained, the affected catheter is removed and cultured, and empirical antimicrobials are given. Sometimes, long-term intravenous catheters, such as Hickmen lines, can be 'sterilized' by giving parenteral antibiotics down the line. Exit site infections involving these devices can usually be treated with antibiotics with the line *in situ*. Tunnel infections usually require line removal for resolution. Prevention is by using aseptic technique when inserting catheters, maintaining a high standard of line care, and removing catheters as soon as possible. The insertion site should be disinfected with a reliable disinfectant such as an iodine-containing agent, 70 per cent alcohol, or 2 to 4 per cent chlorhexidine. At the time of insertion the operators should wash their hands and for long-line insertion, wear sterile gloves, gown, face mask, and hat. Removal of peripheral intravascular devices should be considered after 3 days. Central venous catheters are usually only removed if blocked or suspected as a source of sepsis. The skin at the exit site should be checked daily and the device removed if sepsis is suspected. Subcutaneous tunnelling, insertion of a subcutaneous cuff (Hickman line), burying them subcutaneously (e.g. portacaths), and incorporating antimicrobials on to the surface of the device can all reduce the infection rate significantly. Replacing the entire intravenous delivery set every 72 h is sufficient to reduce sepsis secondary to intraluminal contamination of 'giving' sets.

Prosthetic device-related infection

Infections of prosthetic devices such as heart valves, vascular grafts, cerebrospinal fluid shunts, artificial lenses, and joints are usually caused by the normal skin flora, for example coagulase-negative staphylococci. The devices become coated with a layer of host-derived macromolecules such as fibronectin and fibrin which have specific adhesion receptors for bacteria, particularly staphylococci. Once attached, these organisms multiply on the surface of the coated prosthesis forming a biofilm in a state physiologically different from rapidly dividing, 'free' micro-organisms. They are inherently more resistant to antimicrobials, which explains the frequent failure of antimicrobial treatment. Bacteria gain access to prosthetic devices by direct inoculation, usually at surgery, or by settling on a prosthesis after bacteraemic spread. Direct inoculation at surgery is responsible for prosthetic-device infections occurring more than 1 year after insertion since the organisms involved are usually

skin commensals of low virulence. Except for organisms that are exquisitely susceptible to antimicrobials, these infections are seldom cured with antimicrobial agents. Surgical removal of the device is frequently necessary. However, infection of artificial lenses in the eye are frequently cured by antimicrobial treatment.

Prevention is by avoiding contamination of the wound at surgery, by using strict aseptic surgical technique. Sometimes, as when inserting prosthetic joints, there is an advantage in providing operating theatres with ultra-clean air. Prophylactic antimicrobials reduce the risk of some prosthetic devices becoming infected during insertion.

Antibiotic-associated diarrhoea

Up to 30 per cent of patients treated with antibiotics will develop diarrhoea as a result of the disturbance of the complex gut flora. In a few, loss of 'colonization resistance' predisposes to acquisition of *Clostridium difficile*. Faecal/oral colonization by this organism is usually harmless, but in about a third, particularly the elderly, the organism may overgrow, produce a cytotoxin and causing colitis. The clinical picture varies from mild diarrhoea with fever to fulminating toxic megacolon requiring colectomy. *Clostridium difficile* related diarrhoea entails a delay in discharge of about 3 weeks. Since attack rates in the elderly are around 5 per cent, the disease can have a major impact on hospital resources. Diagnosis is by detection of the cytotoxin in stool, but the test has poor disease specificity since toxin may be found for many weeks after full recovery. Patient management includes adequate rehydration, avoiding drugs which inhibit gut motility, and stopping the provoking antibiotics. More severe cases will require metronidazole or vancomycin given by mouth and surgical review. Prevention is by restricting the use of antibiotics according to agreed and audited protocols. Hand washing after patient contact, isolation of patients with diarrhoea, and cleaning the ward environment are employed on microbiological grounds, despite a lack of prospective studies showing their efficacy.

Nosocomial bacteraemia

Bacteraemia may occur secondarily to the infections mentioned above. The incidence is approximately 3/1000 hospital discharges. The case fatality is about 40 per cent, but varies with the severity of the underlying disease, being as low as about 2 per cent in obstetric patients. The focus must be identified and, if possible, removed surgically. Appropriate antimicrobials are given after obtaining blood and other relevant cultures.

Other nosocomial infections

These include viral infections such as varicella-zoster, hepatitis C, hepatitis B, Norwalk, and rotaviruses, bacterial infections such as tuberculosis and legionellosis, and fungal infections such as aspergillosis.

Further reading

Ayliffe GAJ, Lowbury EJL, Geddes AM, Williams JD, eds (1992). *Control of hospital infection: a practical handbook*, 3rd edn. Chapman and Hall, London.

Bennett JV and Brachman PS, eds (1992). *Hospital infections*, 3rd edn. Little, Brown, New York.

Emmerson AM, Enstone JE, Griffin M, Kelsey MC, Smyth ETM (1996). The second national prevalence survey of infection in hospitals—overview of the results. *Journal of Hospital Infection* **32**, 175–90.

Haley RN, Culver DH, White JW, *et al*. (1985). The efficacy of infection surveiellance and control programs in preventing nosocomial infections in US hospitals. *American Journal of Epidemiology* **121**, 182–205.

Infection Control Standards Working Party (1993). *Standards in infection control.* HMSO, Southampton, UK.

Meers P, Jacobsen W, McPherson M (1992). *Hospital infection control for nurses*. Chapman and Hall, London.

Wenzel RP, ed. (1997). *Prevention and control of nosocomial infections*, 3rd edn. Williams and Wilkins, Baltimore.

7.10.1 Respiratory tract viruses

Malik Peiris

Introduction

Infections of the respiratory tract are one of the most common afflictions of mankind. In economically developed countries, acute respiratory infections account for 20 per cent of all medical consultations, 30 per cent of work absences, and 75 per cent of antibiotic prescriptions. Viruses account for the majority of acute respiratory infections, although when they occur, bacterial infections tend to be more severe. Longitudinal family studies suggest that an individual has on average 2.4 respiratory viral infections per year, a quarter of them leading to a medical consultation. With the exception of influenza, these infections are not a major cause of mortality in the developed world, but it is estimated that they contribute to 20–30 per cent of the 4.5 million deaths annually associated with acute respiratory infection in the developing world.

The term 'respiratory virus' is imprecise, but for the purpose of this discussion it will include those that have the respiratory tract as their primary target. Taxonomically, they belong to diverse virus families (Table 1) and are global in distribution. Other viruses cause systemic disease with respiratory tract involvement as part of an overall disseminated disease process in immunocompetent (e.g. measles, Hantavirus pulmonary syndrome) or immunocompromised (e.g. cytomegalovirus) patients. These are dealt with elsewhere.

A respiratory virus may cause a range of clinical syndromes. Conversely, a respiratory syndrome may be caused by more than one virus. The major viral respiratory syndromes and their common aetiological agents are shown in Table 2. The pattern seen in tropical countries is similar, a notable difference being the role of measles as a major cause of lower respiratory tract infections and fatality.

The anatomical demarcation between upper and lower respiratory tract infections is the larynx. Influenza and adenoviruses are well-recognized lower respiratory tract pathogens in adults as well as in children. Respiratory syncytial virus and parainfluenza viruses, hitherto diseases associated mainly with children, are now being increasingly recognized as important lower respiratory tract pathogens in adults and the elderly.

Transmission

The routes of respiratory virus transmission are through direct contact, contaminated fomites, and large airborne droplets (mean diameter >5 μm, range of transmission <1 m). Influenza may be spread over longer distances by small particle aerosol (mean diameter <5 μm), but even here, direct contact, fomites, and large droplets are more important. Adenoviruses are transmitted by the faeco-oral route as well as by direct contact and large droplets.

Factors increasing transmission of respiratory viruses include the time of exposure, close contact (e.g. spouse, mother), crowding, family size, and lack of pre-existing immunity (including lack of breast-feeding). School-age children often introduce an infection into the family and the commencement of school term may affect transmission patterns in the community. Infected children shed higher titres of viruses than adults. The duration of virus excretion is shown in Table 1. Infectivity usually precedes the onset of clinical symptoms. Immunocompromised patients shed virus for a longer time.

Seasonality

Some respiratory viruses have a predictable seasonality, which varies regionally. For example in temperate regions, influenza A is a typically winter disease while in tropical regions it is a spring/summer disease (e.g. Hong Kong) or occurs all year round (e.g. Singapore, India). Similarly, respiratory syncytial virus (RSV), a primarily winter disease in temperate countries, is a summer disease in Hong Kong. Rhinoviruses occur year round (with increases in the spring and fall) in temperate climes while adenoviruses have no predictable seasonality. The basis for seasonality is unclear but climatic factors such as high humidity may help virus survival and transmission. Factors affecting population congregation such as commencement of school-term and seasonal effects on social behaviour may also play a role.

Laboratory diagnosis

A well-collected specimen is the first (and often most important) determinant in successful laboratory diagnosis. Nasopharyngeal aspirates (secretions aspirated from the back of the nose into a mucus trap) or nasopharyngeal washes are superior to nasopharyngeal or throat swabs for the isolation of many respiratory viruses. They offer the advantage that rapid ('same day') diagnosis for a number of viruses is possible provided the appropriate methods are available. Swabs for viral culture are

placed in viral transport medium immediately upon collection and kept cool (around 4°C) until processed. More invasive specimens such as endotracheal aspirates, brochoalveolar lavage or lung biopsy, when available, usually provide better information. However, the likely site of pathology must be kept in mind—the more invasive specimen is not always better.

The laboratory methods used for detecting a virus in the clinical specimen/s are viral culture, antigen detection, and, more recently, nucleic acid detection. Serology is an option for diagnosing some respiratory virus diseases, but is impractical for others such as rhinoviruses where the large number of antigenically distinct serotypes have no common immunodominant antigen/s. On the other hand, adenoviruses (or influenza viruses), though having many antigenic types or variants, have common antigen/s and a single antigen can detect serological responses to many of them. IgM assays are not routinely available for diagnosis of respiratory viral diseases and paired sera are required so that significant increases in antibody titres can be documented. Complement fixation tests are widely used for this purpose though their sensitivity is not ideal. Haemagglutination inhibition tests are more sensitive for diagnosis of influenza and ELISA tests (though still only available in research settings) provide better sensitivity for diagnosis of RSV and parainfluenza infections.

'Near patient testing' is becoming a reality for some viruses (e.g. influenza) with availability of tests that can be performed in a general practice setting. These become more relevant with the greater availability of antiviral drugs.

Rhinoviruses

Rhinoviruses are adapted to replicate at temperatures of 33–35°C, as found in the external airways. There are over 115 distinct serotypes, but only a few will circulate in a region at any given time. Most rhinoviruses use ICAM-1 on the cell surface as the receptor for attachment but a minority of rhinoviruses use other receptors.

Epidemiology

Rhinoviruses remain one of the most common infections of humans, with 0.5 infections per person per year being a conservative estimate. Secondary attack rates in a family setting may be around 50 per cent overall and 70 per cent in those who are antibody negative.

Immunity

In experimental challenges, immunity is serotype specific, and homologous type specific protection lasts for at least 1 year and correlates with serum IgA, IgG, and secretory IgA antibody levels.

Pathogenesis

Viral replication occurs predominantly in the ciliated epithelial cells of the nasopharynx. The structure of the epithelium is preserved. Mucosal secretions associated with coryza appear to be due to the release of inflammatory mediators and neurogenic reflexes.

It was thought that the preference of the virus for a lower temperature for replication restricted it to the upper respiratory tract, however, this is not strictly true. The temperature of the mucosa of the trachea and bronchi is also lower than core body temperature and does not preclude rhinovirus replication. The virus has been isolated from the lower respiratory tract (including bronchial brushings) and viral RNA has been demonstrated by in situ hybridization in bronchial epithelial cells. Rarely, the virus has been isolated from post mortem lungs of immunocompromised patients.

Clinical manifestations

Rhinorrhoea, nasal obstruction, pharyngitis, and a cough are common features of rhinovirus infections. Fever and systemic symptoms are rare, but more common in the elderly in whom disease can be more severe. Rhinoviruses are a major cause of exacerbations of asthma and chronic obstructive respiratory disease. Lower respiratory tract symptoms are uncommon in the healthy young adult, but may occur in children (bronchiolitis), the immunocompromised, and the elderly.

Treatment and prevention

There are no established antiviral drugs for treatment. Topical interferon-a prevents symptoms if given before onset of disease, but cannot be used for prophylaxis over prolonged periods because of side effects. A viral capsid-binding agent (pleconaril) blocks viral attachment and uncoating and is undergoing clinical trials at present. Antibiotics are ineffective in preventing bacterial complications of the common cold. Mucopurulent discharges are part of the natural course of the common cold and are not an indication for antimicrobial treatment, unless it persists (e.g.>10 days). Given the large number of rhinovirus serotypes, vaccination is not an option.

Coronaviruses

There are two distinct serotypes of respiratory coronavirus—0C43 and 229E. They cannot be cultured from primary specimens and laboratory investigations rely on serology or molecular methods which are only available for research.

Epidemiology

Infection occurs in early childhood and 85 to 100 per cent of adults have antibody to both virus types.

Immunity

Volunteer reinfection studies show that 1 year after initial infection protection from reinfection and illness following a challenge from the homologous virus is incomplete.

Pathogenesis

In common with rhinoviruses, coronaviruses induce little or no damage to the respiratory mucosa. The mucosal discharge is caused by the release of mediators from affected host cells.

Clinical findings

Coronaviruses typically cause upper respiratory tract infections and the common cold. Involvement of the lower respiratory tract is probably more frequent than with rhinoviruses. The virus contributes to exacerbation of asthma in children and adults, but is less important in this role than rhinoviruses. Coronaviruses are also significant pathogens of the elderly.

Treatment and prevention

There are presently no options for antiviral treatment or prevention.

Adenoviruses

Adenovirus subgroups A to D cause respiratory, ocular, hepatic, genitourinary, or gastrointestinal system disease in immunocompetent or immunocompromised individuals. Only respiratory diseases are considered here.

Productive replication and excretion of infectious virus can occur for a prolonged period (see below). In addition, adenoviruses can establish chronic persistence or

'latency', the virological basis and clinical significance of which is poorly understood.

Epidemiology

Adenovirus infections are common during childhood (usually serotypes 1, 2, 5 in early childhood, 3 and 7 during school years or later), but continue to occur throughout life. Reinfection with the same serotype occurs but is usually asymptomatic. Serotypes 1,2,5, and 6 are typically endemic, types 4 and 7 more typically associated with outbreaks, and type 3 can occur in either situation.

Clinical features

Adenovirus respiratory illness often leads to upper respiratory tract disease with coryza and sore throat. Fever may last up to 2 weeks. The sore throat may be exudative and clinically difficult to differentiate from streptococcal infection. Adenoviral infection may present as pharyngoconjunctival fever. Otitis media is a complication in children. Unlike other respiratory viral infections, adenoviruses may be associated with elevated white blood cell counts (>15 × 10^{9}/l), C-reactive protein, or ESR and thus more easily confused with bacterial diseases.

Though uncommon, pneumonia may occur sporadically or in epidemics (caused by serotypes 4 and 7 for example), particularly in closed communities such as the military where stress and physical exertion may predispose to lower respiratory tract involvement. Community outbreaks of adenoviral pneumonia have been reported. Radiological appearance varies from diffuse to patchy interstitial infiltrates and pleural effusion may be present. Adenovirus type 7 pneumonia can lead to permanent lung damage, including bronchiectasis, bronchiolitis obliterans, and unilateral hyperlucent lung syndrome.

Adenoviral infection may disseminate and present as 'septic shock' in the newborn baby. Manifestations in the immunocompromised patient includes hepatitis (especially in liver transplant recipients), colitis, and haemorrhagic cystitis (in renal and bone marrow transplant recipients) in addition to pneumonia. The serotypes associated with disease in these patients may differ from those typically found in the immunocompetent patient, and include the subgroup B2 serotypes 11, 34, and 35. With improving control of other common viral diseases of the immunocompromised (e.g. cytomegalovirus), the role of adenoviruses infections is being increasingly appreciated.

Isolation of an adenovirus from a clinical specimen presents a challenge in interpretation. Adenoviruses are excreted for a prolonged period after initial infection, especially, but not exclusively, from faeces. In children, one-third of patients shed viruses longer than 1 month and 14 per cent longer than 1 year. The clinical significance of a positive result depends on the specimen, the method, and the serotype. Isolation of viruses from the respiratory tract carries greater significance than that from faeces. Patients who have symptomatic adenoviral diseases have higher viral loads than those with asymptomatic carriage. Thus, a rapidly growing virus, a positive antigen detection test (both reflecting higher virus load), or a detectable serological response all point to greater clinical significance. Antigen detection applies only to nasopharyngeal aspirates or bronchial washings.

The above guidelines may not apply to immunocompromised patients who may be infected with unusual serotypes. The presence of the virus in multiple body sites or in peripheral blood possibly points to clinical significance, although further data are needed.

Treatment and prevention

Most adenoviral infections in immunocompetent patients are self-limited and require no specific therapy, however, some infections, especially but not exclusively in the immunocompromised, are severe and life threatening. Ribavirin, vidarabine, cidofovir, and ganciclovir are active against adenoviruses *in vitro.* There are anecdotal reports of their therapeutic use with variable success. However, there are no clinical trials on which to base firm recommendations.

Live attenuated oral vaccines containing serotypes 4 and 7 (associated with outbreaks in military conscripts) are safe and effective, but not licensed for general use.

Respiratory syncytial virus (RSV)

Respiratory syncytial virus (RSV) infects human and non-human primates and was first isolated from a chimpanzee with a 'cold'. Related viruses affect cattle and sheep but do not directly affect humans. The virus has two surface glycoproteins on its envelope (G and F) and the immune responses to them correlate with protection. Two subgroups (A and B) are recognized on the basis of antigenic differences of the G glycoprotein.

Epidemiology

Over two-thirds of infants acquire RSV infection during the first year of life. Of patients hospitalized with RSV disease, 75 per cent are younger than 5 months. The peak of morbidity occurs around 2 months of age, a time when passive maternal antibodies protect against most other viral infections. Primary infection does not lead to solid immunity and reinfection is common. The first reinfection can still be associated with lower respiratory tract involvement. Subsequent reinfection occurs throughout life leading to asymptomatic or upper respiratory tract infections. However, significant diseases may result in the immunocompromised or elderly.

Immunity

Both antibody and cell mediated immunity are important in protection. Antibody to the G proteins prevents attachment of viruses to the cellular receptor, but immunity to the F protein is required to prevent cell to cell spread via fusion of virally infected cells. Cell mediated immunity is important in eliminating established viral infection.

Pathogenesis

The virus leads to a ballooning degeneration of the ciliated epithelial cells, lymphocytic infiltration, and necrosis of the epithelium. There is oedema and increased secretion from the mucous cells and the formation of plugs of mucous and cellular debris in the bronchioles. This results in obstruction and air trapping leading to collapse or over-distension of the distal alveoli. Cells throughout the respiratory tract are affected but the alveoli are spared unless there is RSV pneumonia. Degranulation products of mucosal eosinophils and mast cells and cytokines released by infected macrophages contribute to disease pathogenesis. The cell-mediated immune response contributes to immunopathology in some circumstances. For example when patients with severe combined immunodeficiency and chronic RSV infection receive a bone marrow transplant (for correction of the immunodeficiency), engraftment may be associated with exacerbation of the lung pathology, sometimes with fatal consequences.

Severe RSV bronchiolitis is strongly associated with subsequent childhood asthma. RSV appears to promote type-1 hypersensitivity responses following subsequent exposure to unrelated antigens.

Clinical features

RSV infections of infants may lead to bronchiolitis and pneumonia. Bronchiolitis in infants is associated with expiratory wheeze, subcostal recession, hyperinflation of the chest, nasal flaring, and hypoxia with or without cyanosis. Fever is not prominent in half of the patients. Complete obstruction of a small airway leads to subsegmental atelectasis. Apnoea may occur (particularly in premature infants or in those <3 months of age) and may precede the development of bronchiolitis. Interstitial pneumonitis is uncommon but carries a bad prognosis. Otitis media is a common complication of RSV infection in children. Infants at highest risk from severe RSV disease are those under 6 months, those with pre-existing congenital heart disease, chronic lung diseases (e.g. bronchopulmonary dysplasia), and those born premature.

Infection in adults is often asymptomatic or leads to upper respiratory tract infection. During the RSV season, it is an important cause of lower respiratory tract infection in adults and the elderly—estimated to cause 2 to 9 per cent of the hospitalizations and deaths associated with pneumonia in the elderly. Much of this morbidity is clinically indistinguishable from influenza.

RSV (as well as parainfluenza and influenza) infections in the immunocompromised patient can be life threatening. They usually occur during community outbreaks, but a significant proportion are nosocomially acquired. Once infected, immunocompromised patients have a prolonged period of viral shedding and pose a risk of

transmission to other high-risk patients. The disease typically commences as an upper respiratory tract infection but may progress to involve the lower respiratory tract with more serious consequences. Factors that increase risk of disease progression appear to include bone marrow transplant recipients who acquire the infection in the period prior to engraftment and oncology patients with neutrophil counts less than 0.5×10^{6}/l. Those immunocompromised by HIV appear to tolerate community acquired respiratory viruses better than oncology patients and transplant recipients. This may, however, reflect inadequacy of data rather than reality.

Treatment and prevention

Ribavirin has activity against RSV *in vitro.* Aerosol administration is recommended because it results in much higher concentrations in the respiratory tract than can be achieved by intravenous administration. A number of controlled clinical trials in patients with severe RSV disease have reported clinical benefits associated with its administration by small particle aerosol via a mist tent, mask, oxygen hood, or ventilator, but these findings remain controversial.

In adult bone marrow transplant recipients, an uncontrolled study of ribavirin together with intravenous immune globulin (selected batches with high neutralizing antibody titre) appeared to be beneficial when compared to historical controls. More information is required for deciding the best management strategy.

Monthly intravenous administration of human hyperimmune RSV immunoglobulin during the RSV season, protects against disease of the lower respiratory tract and otitis media in patients with pre-existing risk factors, but is not yet widely available. It did not confer benefits to children with cyanotic heart disease and is not recommended for this group. Side-effects included reversible fluid overload, decrease in oxygen saturation, and transient fever. High titre RSV intravenous immunoglobulin by itself is ineffective in treatment of established RSV disease.

Candidate vaccines for RSV are undergoing clinical trials at present but none is yet available for routine use. Experience of early trials with inactivated RSV vaccines that led to enhanced RSV disease, rather than protection, continues to haunt the field.

Parainfluenza virus

Parainfluenza viruses, despite their name, are not related to influenza viruses, and are more akin to respiratory syncytial virus with which they are classified. They carry two envelope glycoproteins; HN containing both haemagglutinin and neuraminidase activity and F carrying fusion activity.

Epidemiology

The total impact on hospitalization of children by all four types of parainfluenza viruses taken together is comparable to that of RSV but, in contrast to RSV, their impact is in later infancy and childhood. In temperate countries, parainfluenza virus type 3 occurs annually and infects two-thirds of all infants in their first year of life. Parainfluenza type 1 and 2 tend to occur in alternate years and infection is acquired more slowly over childhood. Reinfection with parainfluenza viruses occurs, but rarely leads to lower respiratory tract infection.

Pathogenesis

The virus is confined to the respiratory epithelial cells, macrophages, and dendritic cells within the respiratory tract. Dissemination, even in immunocompromised patients, is rarely documented.

Immunity

Reinfection with parainfluenza viruses continues throughout life. Presence of virus specific IgE in nasopharyngeal secretions has been implicated in the development of parainfluenza croup or bronchiolitis.

Clinical features

Parainfluenza type 1 predominantly causes croup, while type 2 and 3 also cause bronchiolitis and pneumonia. Croup (or laryngotracheobronchitis) is associated with fever, hoarseness, and a barking cough and may progress to inspiratory stridor due to narrowing of the subglottic area of the trachea. The differential diagnosis is epiglottitis due to *Haemophilus influenzae* type b. Parainfluenza type 4 infection is rare, but causes bronchiolitis and pneumonia in children, often in those with underlying disease.

Reinfection in adults, when symptomatic, is a coryzal illness with hoarseness being prominent. Parainfluenza viruses (type 3 in particular) are significant causes of lower respiratory tract disease in adults when the virus is active in the community.

Parainfluenza viruses cause problems in immunocompromised patients (see section on RSV). Lower respiratory tract involvement is associated with wheezing, rales, dyspnoea, and diffuse interstitial infiltrates, and a fatal outcome in one-third of patients. When pneumonia occurs, the histological appearance of the lung is that of a giant cell or an interstitial pneumonia.

Treatment and prevention

The need for specific antiviral therapy arises, particularly in the immunocompromised. Ribavirin is effective *in vitro* but there are no controlled trials documenting its clinical efficacy. There are anecdotal reports of clinical efficacy as well the lack of it.

There are no options for prevention at present, either using vaccines or passive immunization. A live attenuated bovine-derived vaccine strain is currently undergoing clinical trials.

Influenza viruses

Influenza viruses contain a segmented RNA genome. Types A, B, and C are antigenically distinct; of these, types A and B are important in human disease. The viral envelope contains two glycoproteins, the haemagglutinin (H) and neuraminidase (N), which are critical in host immunity. Influenza viruses are designated by the virus type, place of isolation, strain designation, year of isolation, and the H and N antigen subtype, for example A/Sydney/5/95 (H3N2).

Epidemiology

The H and N genes of influenza types A, B, and C undergo mutational change resulting in the emergence of antigenic variants ('antigenic drift'). Every few years, a variant successful in evading the prior immunity of the human population emerges, to cause a global epidemic.

Fifteen H and 9 N subtypes of influenza A are found in aquatic birds, the natural reservoir of the virus. Human influenza A viruses in the first half of this century carried H1N1 surface antigens. In 1957, this virus acquired the genes for different H and N antigens (H2N2) by reassortment of its segmented genome with an avian virus ('antigenic shift'). The human population had no immunity to these new antigens and the virus caused the 'Asian flu' pandemic. A similar reassortment event gave rise to the H3N2 virus and the 'Hong Kong influenza' pandemic of 1968. While all three influenza pandemics this century resulted in significant morbidity and mortality, the toll exacted by the 'Spanish flu' of 1918 was horrendous—over 20 million deaths, greater than that of both World Wars combined. Since influenza B and C have no significant zoonotic reservoirs, antigenic shift and pandemics do not occur.

Avian viruses (e.g. subtype H5N1, H9N2) can occasionally infect humans without undergoing prior reassortment with existing human strains. The H5N1 virus that recently emerged in Hong Kong clearly had the potential to cause disease of unusual severity. However, such non-reassorted avian viruses do not appear to be efficiently transmitted between humans, a prerequisite for a pandemic virus. What might have happened had the H5N1 virus had the opportunity to undergo reassortment and adapt to transmission in humans is too horrifying to contemplate.

Pathogenesis

Viral replication occurs in the columnar epithelial cells leading to its desquamation down to the basal cell layer. The pathology involves the entire respiratory tract. Infection results in decreased ciliary clearance, impaired phagocyte function, and increased adherence of bacteria to viral infected cells, all of which promote the occurrence of secondary bacterial infection.

While there are differences in viral virulence (e.g. current H1N1 strains cause milder disease than H3N2), pre-existing cross-reactive immunity is a major determinant in reducing disease severity. Virus dissemination outside the respiratory tract is uncommon in humans, though it has been occasionally detected in the brain, heart, and fetus.

Immunity

Infection by an influenza virus results in long-lived immunity to homologous reinfection. However, the continued antigenic change in the virus allows it to keep ahead of the host immune response. Cross-immunity to 'drifted' strains within the same H or N subtype may provide partial protection, but there is little cross protection between different subtypes. Local and systemic antibody responses and cytotoxic T cells contribute to host protection.

Clinical features

Influenza ranges from asymptomatic infection, through the typical influenza syndrome, to the complications of influenza. While it cannot always be distinguished from other viral infections on clinical grounds, the typical influenza syndrome is relatively characteristic. It is associated with fever, chills, headache, sore throat, coryza, non-productive cough, myalgia, and sometimes prostration. The onset of illness is abrupt and the fever lasts 1 to 5 days. The pharynx is hyperaemic but does not have an exudate. Cervical lymphadenopathy is often present and crackles or wheezing are heard in around 10 per cent of patients. While the acute illness usually resolves in 4 to 5 days, the cough and fatigue may persist for weeks thereafter.

Common (>10 per cent of symptomatic patients) complications of influenza include otitis media (in children) and exacerbation of asthma, chronic obstructive airways disease, and cystic fibrosis. Less common complications are acute bronchitis, primary (viral) and secondary (bacterial) pneumonia, myocarditis, febrile convulsions, encephalopathy, encephalitis, and myositis (especially in patients with influenza B infection). Age, prior immunity, virus strain, the presence of underlying diseases, pregnancy, and smoking all influence morbidity and severity.

Treatment and prevention

Antiviral therapy

Antiviral drugs with proven clinical efficacy for treatment of influenza A are amantadine, rimatadine and a new class of antivirals, the neuraminidase inhibitors (e.g. zanamivir, oseltamivir). The neuraminidase inhibitors are also active against influenza B, while amantadine and rimatadine are not. All these drugs have maximal efficacy if administered early (within the first 48 h) in the illness.

Rimatadine has fewer neurological side-effects than amantadine. The former is mainly eliminated by the liver while amantadine is excreted by the kidney, a point relevant for patients with compromised renal or liver function. These drugs may be used for containing institutional outbreaks. However, the prophylactic efficacy may be lost if the index case is also treated, probably due to the emergence and transmission of resistant strains.

Preliminary data suggests that antiviral resistance may be less of a clinical problem with the neuraminidase inhibitors. Zanamivir is administered by inhalation, oseltamivir orally. There is limited clinical data suggesting efficacy of aerosolized ribavirin in therapy of influenza A and B.

Aspirin should be avoided in children with influenza because of the increased risk of Reye's syndrome.

Vaccines

Influenza vaccines contain antigens from the two subtypes of human influenza A (H3N2 and N1N1) and B viruses. To keep abreast of change in the surface antigens of the virus, its composition must be modified on an annual basis and annual reimmunization is required. To make global recommendations on vaccine composition and to maintain surveillance for emergence of influenza viruses with pandemic potential, the World Health Organization maintains a global network of collaborating laboratories.

Vaccines in use hitherto have been made from egg-grown viruses and contain: (a) inactivated whole virus, (b) detergent-treated virus (split virus vaccines), or (c) purified surface antigens (subunit of surface antigen vaccines). Split virus and subunit vaccines are associated with fewer side-effects in children (<12 years) and are therefore preferable. Previously unvaccinated children require two doses at least 1 month apart, whereas a single dose appears adequate for adults. These vaccines are generally safe, the most common side-effect being soreness at the injection site lasting a few days. Efficacy is best when there is a good antigenic match between the vaccine and outbreak virus. Immunogenicity and clinical protection are better in healthy young adults compared to patients with chronic renal failure, the immunocompromised, and the elderly (all groups most at need of the vaccine). However, the vaccine is still effective in reducing influenza and pneumonia-related hospitalization and mortality in the elderly and is cost-saving. In young adults, vaccination is associated with decreased absenteeism from work. The duration of protection is limited and therefore vaccine administration should be timed to precede the expected peak of influenza activity.

Influenza vaccine is recommended to those groups at highest risk of morbidity. Recommendations vary from country to country, but they usually include patients in chronic care facilities (especially the elderly), those with chronic cardiopulmonary, lung, or renal diseases, diabetes mellitus, haemoglobinopathies, and the immunocompromised. Some countries, such as the United States, extend the recommendation to all persons over 65 years, pregnant women who will be in the second or third trimester during the influenza season, children receiving long-term aspirin therapy (potentially at risk from Reye's syndrome if they acquire influenza), health-care workers (particularly those in contact with the high-risk patient groups above), and household members of persons in high-risk groups. Currently, there is no consensus on the use of influenza vaccine in HIV infected patients.

An intranasally administered, cold-adapted, live attenuated vaccine has undergone clinical trials with promising results and may in future offer advantages of easier administration and greater patient acceptability.

Nosocomial infection

Respiratory viruses are efficient nosocomial pathogens. Although influenza and RSV are the most notorious, even rhinoviruses cause problems when transmitted to immunocompromised patients. Though paediatric units face the brunt of the problem, adult wards are not exempt. Transmission may occur from patient to patient, patient to staff, and staff to patient, with visitors making their own contribution.

While transmission occurs by large respiratory droplets gaining access to the mucosa of a susceptible individual, their dispersal range is short (<1 m). Much of the transmission occurs by direct hand contact. Adherence to strict hand-washing is the most critical preventive measure. Gloves are useful in reinforcing the 'hand-washing message', but will only be effective if they are changed between patients. Cohorting infected patients, either by symptoms (during the outbreak season) or by rapid viral diagnostic results, is useful. Influenza A vaccination of health-care workers, especially those caring for high-risk children, is to be recommended. Staff education is vital, including awareness of the fact that some of these viruses manifest themselves as a mild 'cold' in adults, and that infected staff members can transmit to patients under their care.

Further reading

Centers for Disease Control and Prevention (2001). *Prevention and control of influenza: recommendations of the Advisory Committee on Immunisation Practices* (ACIP). MMWR 50 (No.RR-4), pp.

1–65. Atlanta, GA. [Reviews the use of vaccines and antiviral therapy for influenza prophylaxis.]

Dolin R, Wright PF, eds (1999). *Viral infections of the respiratory tract.* Marcel Dekker, Basel, pp. 1–432. [Comprehensive monograph with chapters on each of the respiratory viruses, antiviral therapy, and on infections in immunocompromised patients.]

Dowell SF, ed (1998). Principles of judicious use of antimicrobial agents for pediatric upper respiratory tract infections. *Pediatrics* **101** (Suppl.), 163–84. [Journal supplement reviewing the use and abuse of antibiotics in upper respiratory tract infections.]

Gem JE, Busse WW (1999). Association of rhinovirus infections with asthma. *Clinical Microbiology reviews* **12**, 9–18.

Han LL, Alexander JP, Anderson U (1999). Respiratory syncytial virus pneumonia among the elderly: An assessment of disease burden. *Journal of Infectious Diseases* **179**, 25–30. [Key paper reviewing data on the role of RSV in respiratory disease of the elderly.]

Jacob John I, *et al*. (1991). Etiology of acute respiratory tract infections in children in tropical Southern India. *Reviews in Infectious Diseases* **13** (Suppl. 6), S463–9. [Describes the epidemiology of respiratory viruses in a tropical setting.]

Madeley CR, Peiris JSM, McQuillin J (1996). Adenoviruses. In: Myint S, Taylor-Robinson D, eds. *Viral and other infections of the human respiratory tract,* pp.169–90. Chapman and Hall, London. [Reviews the adenoviral respiratory disease and laboratory diagnosis.]

Nicholson KG, Webster RG, Hay AJ, eds. (1998). *Textbook of influenza.* Blackwell Scientific, Oxford. [Comprehensive review of the ecology, clinical features, and control of influenza.]

Shortridge KF (1995). The next pandemic influenza virus. *Lancet* **346**, 1210–12. [Reviews the genesis of pandemic influenza viruses in the context of its zoonotic origin.]

Siddell S, Myint S (1996). Coronaviruses. In: Myint S, Taylor-Robinson D, eds. *Viral and other infections of the human respiratory tract.* Chapman and Hall, London, pp. 141–67.

Treanor J (1997). Respiratory infections. In: Richman DD, Whitley RI, Hayden FG, eds. *Clinical virology,* pp. 5–33. Churchill Livingstone, New York. [Reviews viral respiratory infections.]

Yuen KY, *et al*. (1998). Clinical features and rapid viral diagnosis of human disease associated with avian influenza A H5N1. *Lancet* **351**, 467–71. [Describes the clinical features of an avian influenza outbreak with high morbidity and mortality.]

7.10.2 Herpesviruses (excluding Epstein–Barr virus)

J. G. P. Sissons

Human herpesviruses

General introduction

The family of Herpesviridae is widely distributed in the animal kingdom. More than 100 herpesviruses have been isolated from humans, primates, and other mammals, and from reptiles and fish. Comparative sequence analysis suggests they have been coevolving with their individual hosts for millions of years. To date, eight human herpesviruses have been identified, as summarized in Table 1. Herpesviruses have been assigned to three subfamilies, the alpha-, beta-, and gammaherpesvirinae, on the basis of shared genomic and biological properties. All the herpesviruses are characterized by a linear double-stranded DNA genome, contained inside an icosahedral capsid, which is surrounded by a protein tegument and an outer lipid envelope containing virus glycoprotein spikes. They are large viruses whose genomes consist of unique segments of DNA flanked by repeat sequences and encode most of the proteins needed for replication. Although differing in many of their biological properties, all herpesviruses share an important biological feature: their capacity to produce latent infection in their natural host, during which the viral genome persists in cells as a closed circular episome, only a limited subset of virus genes being expressed. Although individual viruses establish latency in different types of cell, this property is key to their ability to produce persistent lifetime infection of the host, and thus to persist in the population. The individual human herpesviruses, and their associated diseases are described in the succeeding chapters: these diseases may result from primary infection, or reactivation of the virus from latency, and tend to be more severe in immunosuppressed patients. The gammaherpesviruses can induce cell transformation and are associated with specific tumours.

Herpes simplex virus infections

Historical introduction

Herpes is a word derived from the Greek meaning to creep or crawl, apparently used since antiquity to describe the evolution of the skin lesions caused by herpes simplex virus (**HSV**) and varicella zoster virus (**VZV**). HSV was the first of the herpesviruses to be isolated, during the 1930s, after the infectious nature of the mucous membrane lesions it causes had been demonstrated by transmission to animals in 1919. The serological distinction of the two types of HSV, HSV-1 and HSV-2, and the association of HSV-2 with genital herpes was made in the 1960s. HSVs are now some of the most intensively studied human viruses.

Aetiology

HSV has a genome size of 150 kbp and codes for some 80 proteins. The genomes of HSV-1 and HSV-2 are largely colinear, but have different restriction

endonuclease sites. Gene expression occurs in three temporally regulated phases of immediate-early, early, and late genes. Immediate-early proteins are largely regulatory proteins which prepare the cell to produce further virus, the early genes code particularly for enzymes involved in the replication of virus DNA, and the late genes for the structural proteins of the virion. Antigenic differences in the surface glycoprotein G are used to distinguish between HSV-1 and HSV-2. Release of progeny virus is normally accompanied by host-cell death—that is, the infection is lytic. The virus infects a relatively wide range of cells *in vitro*, and can also infect experimental animals thereby allowing studies of its pathogenesis.

Epidemiology

HSV is a ubiquitous virus, widely distributed in all populations of the world. Although the virus can produce experimental infections of animals, there are no natural animal hosts and humans are the only reservoir for the virus. Transmission occurs by direct contact of a susceptible individual with infected secretions from an HSV carrier. This is usually via transmission from oral, genital, or skin lesions to mucous membranes or abraded skin of the recipient. HSV carriers can excrete virus asymptomatically, and 1 to 15 per cent of adults excrete HSV at any one time. The prevalence of infection is conventionally assessed by the demonstration of antibody to HSV-1 or HSV-2. The prevalence of HSV-1 increases with age, although the time of acquisition of HSV-1 antibody varies depending on socioeconomic factors. There is a higher seroprevalence in lower socioeconomic groups in early life: 70 to 90 per cent of individuals have antibodies by the age of 10 years, whereas only about 30 per cent of higher socioeconomic groups have antibody by this time. By mid-life 80 to 90 per cent of all individuals have antibody to HSV-1.

HSV-2 infection is usually acquired through sexual contact: consequently seroconversion correlates with the onset of sexual activity, and there is a progressive increase in seroprevalence to HSV-2 beginning in adolescence. The number of sexual contacts is a major risk factor for acquisition of HSV-2. Cumulative seroprevalence rates in adults vary from 10 to 80 per cent depending on the population and risk factors.

HSV can be transmitted to the neonate by infection (usually HSV-2) from maternal genital secretions at the time of delivery. Neonatal HSV infections usually occur in children born to mothers who are asymptomatic excretors of the virus and have no history of genital herpes.

Pathogenesis

HSV infects and replicates in epithelial cells at the site of inoculation on to mucous membranes or abraded skin, with an incubation period of between 4 and 6 days before clinical lesions appear. There is a marked local inflammatory response, but viraemia and dissemination may occur in the immunocompromised host. Following local epithelial replication, HSV enters the peripheral sensory nerves innervating the site of epithelial replication, and ascends the axons by retrograde transport to reach the dorsal root ganglia, or the trigeminal ganglion in the case of oral or conjunctival inoculation. The virus then becomes latent in sensory ganglia, but, despite extensive study, the mechanism of virus latency remains uncertain. Latent HSV-DNA is in an inactive state with minimal gene expression. RNA species called 'latency-associated transcripts' (**LATs**) are the only detectable transcripts. LATs have no detectable protein product and their deletion from the genome does not prevent the establishment of latency, although reactivation is impaired. Latent HSV is carried for the lifetime of the host, but may reactivate in response to certain stimuli including stress, menstruation, ultraviolet light, and immunosuppression. Upon reactivation, infectious virus is produced, travels down peripheral nerves by anterograde transport, and replicates in epithelial cells at the nerve ending. The neuronal latency of HSV and VZV is an extremely effective method of virus persistence. Latent virus in neuronal cells appears to be inaccessible to the immune response, and as it does not replicate, is not susceptible to the action of antiviral drugs. In addition to specific antibody, normal HSV carriers mount a cytotoxic T-lymphocyte response to the virus, which is presumed to control reactivation at local sites. HSV encodes genes which interfere with antigen processing by the class I MHC pathway, and are presumed to help the virus to evade the T-cell immune response. There is no good evidence that the immune response to HSV of people who have symptomatic reactivation episodes differs from that of asymptomatic carriers.

Clinical features

Primary infection with HSV is often asymptomatic: in a recent study of sexually active subjects, only 60 per cent of primary infections with HSV-1, and 40 per cent with HSV-2, were symptomatic. HSV-1 is the predominant cause of orofacial infections and HSV-2 the predominant cause of genital HSV infection, but the clinical manifestations overlap.

Gingivostomatitis

This is the commonest clinical form of primary infection with HSV-1. It is most often seen in children and has an incubation period of 2 to 12 days. Primary infection may be associated with a considerable systemic reaction, with fever, sore throat, pharyngeal oedema, and redness. Painful vesicles appear a few days later on the pharynx and the oral mucosa, lips, and the skin around the mouth: there may be cervical lymphadenopathy. Affected patients may have difficulty in eating, and the lesions last from 3 days to 2 weeks. The differential diagnosis includes other causes of pharyngitis including bacterial pharyngitis and herpangina (due to coxsackie A virus infection): anterior vesicles and ulceration affecting the lips and skin around the mouth are more suggestive of HSV infection. Stevens–Johnson syndrome and severe aphthous ulceration may appear similar, and staphylococcal impetigo affecting the skin around the mouth can give a similar external appearance, but is not associated with oral ulceration.

Reactivation of HSV may give rise to recurrent orolabial lesions: appearing as intraoral mucosal ulcers, but more frequently as the classical 'cold sore' on the lips or skin around the mouth (Plate 1). Patients may first experience a tingling sensation in the area of impending ulceration 1 to 2 days prior to the appearance of vesicles. The lesions usually recur in the same site in individual patients. Around 25 per cent of HSV-1 seropositive people develop recurrent orolabial lesions: the majority have only one or two reactivation episodes per year, although a minority (less than 10 per cent) have more than one attack a month. These episodes are not associated with systemic symptoms and diagnosis is usually straightforward.

Infection at other cutaneous sites

Herpetic whitlow

HSV infection of the finger (herpetic whitlow) may complicate primary oral or genital herpes by autoinoculation of virus, or may occur through occupational exposure (for instance, in nursing, medical, and dental staff). There is oedema, erythema, and local tenderness of the infected finger. Lesions at the fingertip may be confused with pyogenic bacterial paronychias and incised (which is contraindicated for herpetic whitlow, and may even spread infection).

Herpes gladiatorum

This is a term which refers to mucocutaneous HSV infection occurring by transmission of infection through skin trauma, resulting from wrestling or other contact sports.

Eczema herpeticum

HSV infections of the skin are more severe in patients with pre-existing skin disease. In patients with eczema, burns, or other blistering skin diseases, HSV infection may become disseminated.

Cutaneous HSV infection can be confused with herpes zoster, although the latter is usually easy to diagnose by its unilateral dermatomal distribution.

Herpes simplex and erythema multiforme

About 15 per cent of all cases of erythema multiforme are preceded by a symptomatic attack of recurrent herpes simplex, and in susceptible individuals the characteristic rash can be induced by the intradermal inoculation of inactivated herpes simplex virus antigen. The rash of erythema multiforme starts several days after the onset of the herpetic vesicles, and in severe cases can involve the mucous membranes (Stevens–Johnson syndrome). The frequency of these attacks can be reduced by aciclovir prophylaxis.

Keratitis

HSV keratitis is characterized by the acute onset of pain, blurred vision, conjunctival injection, and dendritic ulceration of the cornea. HSV keratitis can cause corneal blindness and its treatment is urgent. Topical aciclovir is the drug of choice, for topical steroids may make the infection worse. HSV can also cause an acute necrotizing retinitis, usually seen in immunosuppressed subjects such as those with HIV infection, but rarely in immunocompetent people.

Genital herpes

Primary genital HSV infection is sexually transmitted and may be associated with systemic symptoms such as fever, headache, and myalgias. Symptoms tend to be more severe in women than in men. There is local pain and itching, dysuria, and vaginal discharge with inguinal lymphadenopathy, with vesicles and ulcers on the vulva, perineum, vagina, and cervix, and sometimes on the skin of the buttocks (Plate 2). In males, primary HSV lesions are seen as vesicles on the shaft or glans of the penis and there may be an associated urethritis. Most genital HSV infections are due to HSV-2, with a variable smaller proportion due to HSV-1. Only 40 per cent of primary HSV-2 genital infection is symptomatic: in patients who have had prior HSV-1 infection, the symptoms of primary genital herpes tend to be less severe. HSV has been isolated from the urethra in 5 per cent of women with the 'urethral syndrome', in the absence of obvious genital lesions. Other manifestations of genital tract disease due to primary HSV infection are, rarely, endometritis and salpingitis in women, and prostatitis in men.

HSV proctitis may follow rectal intercourse. There is anorectal pain and discharge with ulcerative lesions visible on sigmoidoscopy. Perianal lesions are seen in immunosuppressed patients, and spreading perianal HSV infection and HSV proctitis occur in patients with human immunodeficiency virus (**HIV**) infection.

Recurrent genital herpes is frequent in the first year following primary genital infection (90 per cent for HSV-2 and 55 per cent for HSV-1). Thereafter the recurrence rate tends to decrease with time and is around 3 or 4 attacks per year for HSV-2, but less for HSV-1. Severe recurrent genital herpes is particularly troublesome to women.

Complications of primary genital HSV infection include a sacral radiculomyelitis with urinary retention and hyperaesthesia of the perineal area, which usually resolves over several weeks. Aseptic meningitis requiring admission to hospital occurs in up to 7 per cent of women and 2 per cent of men, although suggestive symptoms are more common. Occasionally, and more seriously, transverse myelitis may occur.

HSV encephalitis (see also Viral infections of the CNS)

Encephalitis is the most serious type of disease produced by HSV in the normal immunocompetent host, and has an estimated annual incidence of 2 or 3 cases per million of population. It is the commonest identified cause of acute sporadic encephalitis in Western countries, with the great majority of cases due to HSV-1. There is a reported biphasic age incidence, with higher rates between the ages of 5 and 30 and in those over 50 years. The clinical presentation is that of a focal encephalitis, with an acute onset of fever, confusion and unusual behaviour, impaired consciousness, and possibly focal neurological abnormalities. However, as there are no clinical features specific to HSV encephalitis, the diagnosis should be considered in any patient who presents with clinical features that could indicate an encephalitis.

The cerebrospinal fluid (**CSF**) shows lymphocytic pleocytosis, although neutrophils and red cells may also be present, with a raised protein level. Computed tomographic (**CT**) scans of the brain may show changes in the temporal lobe: magnetic resonance imaging (**MRI**) is a more sensitive method of detection. The electroencephalogram (**EEG**) classically shows spike- and slow-wave activity localized in the temporal lobes. The most definitive way of establishing the diagnosis is by brain biopsy: in the original trial of aciclovir for the treatment of HSV encephalitis, brain biopsy was an entry criterion and confirmed the diagnosis in only 50 per cent of clinically suspected cases. Brain biopsy is very rarely used now, since the advent of non-toxic effective chemotherapy for HSV. There is good correlation between a positive polymerase chain reaction (**PCR**) test for HSV-DNA in the CSF and the diagnosis of HSV encephalitis by brain biopsy and virus isolation. Evidence for intrathecal production of specific HSV antibody is also diagnostic, but as it usually becomes positive a week after onset, PCR-based diagnosis is more useful. Serum or CSF titres of antibodies to HSV do not usually increase in the first week of the illness. In practice, the diagnosis is established by a compatible clinical picture, evidence of characteristic temporal lobe involvement on CT or MRI imaging and EEG, and by PCR-based detection of HSV-DNA in the CSF.

The pathogenesis of HSV encephalitis remains uncertain. Up to half of patients have primary infection and in the rest the disease is presumed to result from reactivation. However, where HSV has been isolated from the brain and the mouth simultaneously in the same patient, the two isolates differ by restriction endonuclease analysis in about 30 per cent of the patients, suggesting a new exogenous virus infection in an already seropositive patient. HSV-DNA can be detected in the brain at autopsy of normal virus carriers, but the factors precipitating HSV encephalitis are not known. Immunosuppression is not usually associated with HSV encephalitis, which predominantly affects normal immunocompetent adults, and very rarely patients with advanced HIV infection. The pathology is that of a focal haemorrhagic necrotizing encephalitis, affecting the temporal lobes.

Treatment

Treatment should be started immediately with intravenous aciclovir (in doses as given below) in any patient in whom HSV encephalitis is clinically suspected, without waiting for confirmation of the diagnosis. Prior to effective antiviral therapy, the mortality from HSV encephalitis was over 70 per cent with very few patients making a full neurological recovery. Intravenous aciclovir was established to be more effective than the previous best therapy of vidarabine in a randomized trial reported in 1986. Mortality in the aciclovir-treated group was 28 per cent, although a lower Glasgow Coma Score on entry carried a higher risk of mortality. However, only 38 per cent of those who received aciclovir had fully recovered at 6 months: there is thus still a high incidence of permanent neurological events, particularly seizures, defects of memory, and personality changes and the prognosis of HSV encephalitis remains serious.

Meningitis

HSV can cause an aseptic meningitis which is quite independent of, and not associated with progression to, HSV encephalitis. The commonest association is with primary genital HSV-2 infection: the incidence of proven HSV meningitis is 7 per cent in women, and 2 per cent in men, with primary genital HSV. There is a pleocytosis, usually lymphocytic, but neutrophils may predominate in early meningitis. HSV may be isolated from CSF by culture, but is now more reliably detected by PCR for HSV-DNA.

In a high proportion of patients with Mollaret's meningitis (a recurrent aseptic meningitis of unknown aetiology), HSV-DNA is reported to be detectable in cerebrospinal fluid by PCR.. The role of HSV in this syndrome remains uncertain.

An association of HSV with Bell's palsy has been reported, but a recent Cochrane review considered the evidence was inconclusive.

Neonatal HSV infection and pregnancy

The incidence of neonatal HSV infection is approximately 1 in 3500 deliveries per annum in the United States, but appears to be rarer in the United Kingdom at 1.65 in 100 000 live births. About 70 per cent of cases are caused by HSV-2 and result from fetal acquisition of HSV-2 from maternal genital secretions during delivery. Most infants with neonatal HSV are born to mothers without clinically evident HSV infection. The risk of transmission from a woman with symptomatic primary HSV infection is about 50 per cent, and 20 per cent from a woman with clinically evident recurrent HSV-2 infection. A small proportion (about 10 per cent) of infections are acquired postnatally by contact with other family members with active lesions.

Neonatal HSV infection appears as lesions on the skin, eye, and mouth or may present as encephalitis or disseminated visceral infection. Although initial superficial infection may progress to visceral infection, visceral infection can present with no evidence of cutaneous lesions, and the diagnosis should be considered in severely ill neonates. Untreated, visceral infection has a high mortality (around 60 per cent).

Primary infection in early pregnancy can lead to congenital HSV infection. This is rare, but can produce serious congenital abnormalities.

HSV in the immunosuppressed patient

HSV infections in immunosuppressed subjects are usually due to reactivation rather than primary infection. They tend to be more severe, are more likely to progress, and take longer to heal than in the normal immunocompetent host. Clinical manifestations in patients with HIV infection include severe perineal, orofacial, and oesophageal infection. HSV pneumonitis, hepatitis, and colitis are also described in immunosuppressed patients.

Pathology

The histological appearance of HSV infection is identical whether infection is primary or recurrent. There is ballooning of infected cells with condensed chromatin in the cell nuclei. Intranuclear inclusion bodies (Cowdrie type A bodies) may be seen and multinucleated giant cells form. VZV produces similar appearances.

Laboratory diagnosis

Definitive diagnosis is made by isolation of virus: swabs from vesicular fluid or other body fluids in virus transport medium can be inoculated into tissue culture and typical cytopathic effects seen. Virus from vesicle fluid can also be identified rapidly as a herpesvirus by electron microscopy after negative staining. The use of PCR-based techniques to detect viral DNA is becoming more widespread. It is particularly applicable to the detection of HSV-DNA in cerebrospinal fluid.

Serological tests for antibody to HSV are only useful in making a retrospective diagnosis. Seroconversion provides proof of primary infection and absence of antibody to HSV-1 or HSV-2 rules out a diagnosis of recurrent HSV infection. However, making a diagnosis of reactivation by demonstrating rising antibody titres is of limited value.

Treatment

The introduction of aciclovir heralded a new era of targeted antiviral drugs and superseded other drugs previously used for the treatment of HSV infections such as vidarabine and idoxuridine. Aciclovir is an acyclic nucleoside which is preferentially phosphorylated in HSV-infected cells by the virus-encoded thymidine kinase to aciclovir monophosphate. Cellular kinases then phosphorylate the aciclovir monophosphate to the triphosphate, which is incorporated into nascent HSV-DNA where it acts as a chain terminator; aciclovir also directly inactivates the HSV-DNA polymerase. Two newer related drugs with the same mechanism of action are famciclovir, a prodrug of penciclovir, and valaciclovir, the valyl ester of aciclovir, which has greater bioavailability and requires less frequent dosage. All these drugs are relatively free of side-effects, although intravenous aciclovir can crystallize in the renal parenchyma and produce renal impairment: it should be given by infusion over 1 hour, and patients should be adequately hydrated. The doses should be reduced in patients with renal impairment.

Primary mucocutaneous infection

In primary oral and genital infection aciclovir 200 mg, 5 times daily, given orally for 10 to 14 days from the onset, reduces the severity of infection, the duration of symptoms, and the duration of viral shedding. There is little evidence that treatment of primary infections reduces the incidence of subsequent symptomatic reactivation episodes. If swallowing is difficult, intravenous aciclovir (5 mg/kg, every 8 h) may need to be given. Famciclovir, 250 mg thrice daily, or valaciclovir, 500 mg twice daily, are alternatives.

Symptomatic reactivation of mucocutaneous infection

Treatment of recurrent infections in the immunocompetent host is often unnecessary as the symptoms are usually very mild. However, aciclovir can shorten the duration of symptoms if it is given very early in the course of the recurrence, preferably during the prodrome before vesicles appear. Oral aciclovir is effective and anecdotal reports suggest that topical aciclovir is effective symptomatically. The same dosage as above for treating a primary infection can be given for 5 days: famciclovir and valaciclovir can be used as alternatives.

Long-term suppressive therapy

This can be considered in immunocompetent patients with genital herpes who have frequent reactivation episodes. Trials of aciclovir in recurrent genital herpes have shown that a dose of 400 mg twice a day significantly reduces the frequency of attacks. However, patients may be able to find a lower effective dose, and in some 200 mg daily prevents attacks. Although it is advised that treatment is discontinued for a month every 6 to 12 months, there is little evidence that resistant virus is a problem in this population. Valaciclovir, 500 mg daily, or famciclovir, 250 mg twice daily, are alternatives.

CNS infection

For HSV encephalitis intravenous aciclovir (10 mg/kg, every 8 h for 10–14 days) should be given to any patient in whom the diagnosis is clinically suspected (see above).

For HSV meningitis intravenous aciclovir (5 mg/kg, every 8 h) can be used with conversion to oral valaciclovir (1 g, twice daily) when improvement occurs, for a total of 10 days.

Systemic infection in the immunosuppressed

Oral treatment as for primary HSV can be used for mild mucocutaneous infection, but for more severe infection and for visceral involvement intravenous aciclovir 5 mg/kg every 8 h should be used. After resolution, continued prophylaxis is usually necessary until immunocompetence is restored, particularly in patients with HIV.

Aciclovir resistance

Resistance of HSV to aciclovir develops readily *in vitro* but is clinically rare: it is due to mutations in the HSV thymidine kinase or DNA polymerase gene. It is seen almost exclusively in immunocompromised patients who have received prolonged aciclovir prophylaxis, especially those with HIV infection, and is manifest as unresponsive or worsening HSV disease despite treatment with aciclovir. There is usually cross-resistance to famciclovir and valaciclovir, and intravenous foscarnet (more usually used to treat human cytomegalovirus infection, see the HMCV section) is the most useful alternative drug for use in severe infection due to resistant HSV.

Prevention and control

There is no vaccine yet available for HSV, although several candidates are approaching phase III trials. There is particular interest in the use of vaccines for postinfective immunization to reduce the frequency of recurrent genital HSV attacks (which has been shown to be possible experimentally in guinea-pigs).

Special problems in pregnant women

Prevention of neonatal HSV infection is best achieved by preventing genital HSV infection late in pregnancy. There is no reason to give aciclovir prophylactically to women with a history of recurrent genital herpes who are asymptomatic, as the incidence of neonatal HSV infection is low in their children. However, women with clinically apparent genital herpes during the last trimester (and probably at any other time in pregnancy) can be treated with aciclovir (although the drug is not licensed for treatment in pregnancy). Women with no clinical lesions may have a vaginal delivery, but the presence of active lesions at the time of labour is regarded as an indication for Caesarean section. Babies born to mothers with clinically apparent genital HSV infection or with a history of recurrent genital HSV infection should be screened for HSV by cultures from the nasopharynx and eyes after birth.

Proven neonatal HSV infection should be treated with high-dose intravenous aciclovir (60 mg/kg per day in three divided doses for 21 days).

Varicella zoster virus infection

Historical introduction

There are clinical descriptions of varicella (chickenpox) and herpes zoster (shingles) in very early medical literature, although the skin lesions of herpes simplex and herpes zoster were grouped together under the term 'herpes'. The similarities between the exanthematous rashes associated with smallpox and with varicella meant they were not distinguished until the latter part of the nineteenth century. Because of its characteristic clinical appearance in a dermatomal distribution, shingles was recognized as a discrete entity in the early Greek literature: the term 'zoster' is said to derive from the Greek term for a girdle, and shingles from the Latin *cingere* meaning to encircle. Von Bocquet in 1892 observed that children developed varicella after contact with adults who had herpes zoster, and in 1925 it was shown that vesicular fluid from patients with zoster produced chickenpox in susceptible individuals when directly inoculated. The idea that zoster resulted from reactivation of latent virus remaining in the individual following childhood varicella was put forward by Garland in 1943, and was strengthened by the work of Hope-Simpson, a British general practitioner. Varicella zoster virus (VZV) was isolated in 1958, and Weller and colleagues showed the similarity of the viral isolates from varicella and zoster. Identity by restriction endonuclease analysis has been shown between the isolates from chickenpox and from later zoster in the same individual, although this was an immunocompromised patient: the long interval between the two illnesses in normal subjects means such studies have never been conducted in this population.

Aetiology

VZV is structurally similar to other members of the Herpesviridae family. The genome is a linear double-stranded DNA of 125 kilobase pairs. The virus is closely cell-associated and spreads from cell to cell in tissue culture. VZV is an alphaherpesvirus, and encodes sets of genes which are largely colinear to those of HSV, and are also expressed in immediate-early, early, and late phases.

Epidemiology

VZV only infects humans, who are thus the only reservoir. The virus is presumed to spread by the respiratory route. Varicella is predominantly a disease of childhood affecting both sexes: 90 per cent of cases occur in children under the age of 13. The incubation period of varicella from exposure to VZV to development of the initial rash is about 2 weeks (with a range of 10–20 days). Patients with varicella are infectious for about 48 h before the vesicles appear, and remain so for 4 or 5 days afterwards until all the vesicles have crusted over. Secondary attack rates in susceptible contacts where there is an index case in the household are between 70 and 90 per cent. The prevalence of VZV varies in different ethnic groups. After 15 years of age only about 10 per cent of subjects in Europe are seronegative for, and consequently susceptible to, infection, although in tropical countries only 50 per cent of young adults may be seropositive. Varicella in adulthood is uncommon in Europe, with less than 2 per cent of all cases occurring in patients older than 20 years.

After primary infection, VZV becomes latent in dorsal root ganglia. Reactivation appears clinically as herpes zoster (shingles), which is a common disease that can affect all age groups but particularly the elderly: about 20 per cent of the population will experience an attack. There is no evidence that exposure to people with active VZV infection predisposes to herpes zoster in their contacts, but a seronegative subject may catch varicella from contact with the vesicles of a patient with shingles. As nosocomial varicella infection is well recognized, isolation of patients with varicella and immunocompromised patients with herpes zoster should be ensured in hospitals. Local unidermatomal zoster is less likely to cause infection and consequently to need isolation. Subclinical infection is unusual and accounts for less than 5 per cent of all infections, but the disease may be mild and in some surveys only 10 per cent of people with a negative history were in fact seronegative for VZV. One attack of chickenpox usually confers life-long immunity.

Pathogenesis

Upon primary infection, initial virus replication probably occurs in the epithelial cells of the upper respiratory tract mucosa, followed by a phase of viraemia during which VZV can be isolated from leucocytes. This blood-borne spread is associated with the production of the disseminated rash. In the skin, the virus infects capillary endothelial cells and adjacent fibroblasts and epithelial cells. During the viraemic phase virus may spread to visceral organs, the lung including alveolar epithelial cells, and transient subclinical hepatitis is probably a normal feature of varicella. Infection of the brain can occur, and VZV encephalitis may be a feature of primary infection, particularly affecting the cerebellum. The encephalitis usually recovers completely (unlike that associated with HSV), and it has been suggested it may have an immune-mediated pathogenesis. Following recovery from primary infection, the virus persists for life in a latent state in dorsal root ganglia. VZV reaches the ganglia by retrograde axonal transport from the lesions in the skin during primary infection, and all dorsal root ganglia and the trigeminal ganglion can potentially carry latent VZV in neurones and possibly in satellite cells.

As for other herpesviruses, the host response is critical in containing the initial infection. The cellular immune response is of particular importance since varicella may be progressive in patients with severely impaired T-cell immunity. Both CD4 and CD8 cytotoxic T lymphocytes specific for VZV are present in normal people carrying latent VZV. The cellular immune response presumably plays a part in controlling reactivation, since impaired T-cell immunity is associated with an increased risk of developing zoster, with zoster in multiple dermatomes, and with cutaneous dissemination of reactivated virus. The increasing risk of herpes zoster with age may reflect waning cellular immunity to VZV.

Clinical features

Primary infection and varicella (chickenpox)

The most striking feature of varicella is the rash, which is centripetal (mainly on the trunk) in distribution (Plate 3). Initially, lesions are present on the face and scalp, before progressing to the trunk and later to the limbs. A macular erythematous rash, papules, and vesicles may all be present together. Individual lesions progress from being papular to vesicles to pustules and then crust over. The scabs normally separate after 10 days without scarring. The systemic symptoms associated with varicella vary considerably. In the majority of children there is a mild illness with fever. Adults characteristically have a more severe illness with myalgias, headache, arthralgias, malaise, and higher fever, with the complications listed below. Symptoms may precede the rash by 1 to 2 days.

The principal complications of varicella in the immunocompetent person are pneumonitis and encephalitis.

Pneumonitis

A prospective study showed that 6 per cent of young adults with chickenpox had respiratory symptoms, but 16 per cent had changes on chest radiography, although the rate of admission to hospital with pneumonia in adults with varicella is only about 0.3 per cent. Patients present with dyspnoea, cough, hypoxia, and bilateral infiltrates on the chest radiograph occurring 1 to 6 days after the appearance of the rash. Hypoxia may be more severe than expected from the physical signs or the chest radiograph. The interstitial pneumonitis can progress to respiratory failure requiring artificial ventilation and intensive care, but it is more commonly transient and resolves completely within 2 to 3 days. Varicella pneumonia is said to be commoner in smokers. Fatalities do occur but the great majority of patients survive, and VZV pneumonia is not associated with long-term respiratory problems. Benign nodular calcification throughout the lung occasionally follows.

Encephalitis

Central nervous system involvement during varicella infection most commonly presents as an acute cerebellar ataxia within a week of the onset of the rash, although it may present as late as 21 days after the rash: it resolves completely over 2 to 4 weeks. A frequency of 1 in 4000 cases of children under the age of 15 years has been quoted. The cerebrospinal fluid of these patients shows a lymphocytosis and elevated protein concentration.

A more serious encephalitis can occur in between 0.1 and 0.2 per cent of cases of varicella. This begins earlier in the course of infection than the cerebellar ataxia, with headache, vomiting, confusion, and impaired consciousness. There is evidence of diffuse cerebral oedema but no defined pattern of CT or MRI abnormality. The encephalitis may be progressive and the mortality is between 5 and 20 per cent with neurological sequelae in up to 15 per cent of survivors.

A meningitis can occur with varicella. Other rare neurological complications reported include optic neuritis, transverse myelitis, and Reye's syndrome.

Other complications

Other complications of primary VZV infection include acute thrombocytopenia with petechiae and purpura and haemorrhage into vesicles and other haemorrhagic manifestations. The platelet count can remain low for weeks after the illness has resolved. Secondary infection of the skin lesions with *Staphylococcus aureus* or *Streptococcus pyogenes* can occur. Purpura fulminans is a rare complication associated with arterial thrombosis and haemorrhagic gangrene. Nephritis and arthritis have been reported as rare complications and myocarditis, pericarditis, pancreatitis, and orchitis are even more rare.

Special problems in pregnant women

Varicella in pregnant women can be severe, with a quoted maternal mortality of 1 per cent. Varicella in the first trimester can cause 'varicella embryopathy': in affected infants there may be a scarred atrophic limb, microcephaly, and cortical atrophy, as well as eye defects, including chorioretinitis, micro-ophthalmia, and cataracts. The autonomic nervous system may be damaged. Varicella embryopathy is rare in recent reported series, giving a risk of about 1 to 2 per cent in mothers with varicella in the first 20 weeks of pregnancy. Varicella zoster immune globulin (**VZIG**) should be considered for pregnant women in contact with varicella (see below), and varicella during pregnancy should be treated with aciclovir on a named patient basis. Neonatal varicella occurs in babies whose mothers contract varicella just prior to or after delivery, and is more severe when the maternal onset is from 2 days before to a week after delivery.

Herpes zoster

The clinical syndrome associated with reactivation of VZV from sensory ganglia is herpes zoster (shingles). This typically presents with pain followed by erythema and vesicular lesions occurring in a dermatomal distribution. The thoracic dermatomes, especially T5 to T12, are involved in about 50 per cent, lumbosacral dermatomes in about 16 per cent, and cranial nerves, mainly the Vth, in 14 to 20 per cent of patients. The first symptoms are usually paraesthesias and shooting pains in the affected dermatome, which precede the eruption of vesicles by several days and occasionally a week or more. Erythematous maculopapular lesions then appear and quickly evolve into a vesicular rash, nearly always in a unilateral dermatome with no vesicles beyond the midline. The vesicles usually form scabs after 3 to 7 days and these separate after 2 weeks or so. There is a risk of secondary infection, particularly with *Staphylococcus aureus*. There may be malaise and low-grade fever, but there are usually no abnormalities of laboratory investigations, although up to 40 per cent of patients with uncomplicated zoster may have lymphocytes and elevated protein concentrations in the CSF. Involvement of the mandibular branch of the Vth cranial nerve can give intraoral lesions on the palate, floor of the mouth, and tongue. Involvement of the geniculate ganglion results in the Ramsay Hunt syndrome, with pain and vesicles in the external auditory meatus, a loss of taste in the anterior two-thirds of the tongue, and a lower motor neurone VIIth cranial nerve palsy.

Complications of zoster

Ophthalmic zoster

VZV reactivation from the trigeminal ganglion can affect the ophthalmic division of the trigeminal nerve resulting in ophthalmic zoster (Plate 4). The features include conjunctivitis, anterior uveitis, a keratitis, and sometimes iridocyclitis with secondary glaucoma and panophthalmitis. However, these latter sight-threatening complications of ophthalmic zoster are unusual. A rare association of ophthalmic zoster is granulomatous cerebral angiitis, which can be associated with arterial thrombosis: cerebral angiography shows segmental narrowing in cerebral arteries on the side of the ophthalmic zoster. CT scanning may show cerebral infarcts, particularly in the distribution of the middle cerebral artery.

Motor zoster

Weakness or paralysis can sometimes be associated with zoster, and is due to involvement of the anterior horn cells in the same segment of the spinal cord as the involved dorsal root ganglion. Depending on the segment involved, this can lead to a monoparesis affecting the upper or lower limb or to diaphragmatic palsy (with involvement of C5/6). Paralysis usually recovers completely, although the outlook for recovery of facial nerve palsy is more variable. It is suggested VZV may be responsible for some cases of 'idiopathic' VIIth nerve (Bell's) palsy.

Autonomic zoster

Lumbosacral herpes zoster can be associated with a neurogenic bladder and acute retention of urine, which may be accompanied by haemorrhagic cystitis due to vesicles on the bladder wall. Intestinal ileus and obstruction may occur.

Zoster meningoencephalitis

A meningoencephalitis may accompany zoster at any site and is characterized by impaired consciousness, headache, photophobia, and meningism. The interval from the onset of skin lesions to symptoms of encephalitis is around 9 days, but may be as long as 6 weeks. Symptomatic encephalitis usually lasts around 2 weeks and is nearly always followed by full recovery without neurological sequelae.

A transverse myelitis, although rare, can occur at any level of the spinal cord.

Postherpetic neuralgia

The incidence of postherpetic neuralgia rises with increasing age of the patient. It is uncommon in young people, but can occur in 50 per cent of patients over the age of 50 years. It is characterized by pain in the affected dermatome persisting for 1 month or more after the acute attack of zoster has resolved. The pain may be steady and burning or paroxysmal and stabbing in nature: it may occur spontaneously or be triggered by stimuli such as temperature or touch.

Zoster sine herpete

This term refers to a syndrome characterized by radicular pain, similar to that experienced in zoster, but without any antecedent skin lesions of zoster. It was originally applied to patients who did have obvious zoster, but had dermatomal pain in areas distinct from those areas where there was rash: however, subsequently it has usually been applied to patients with radicular pain and no rash at all. There are more recent reports describing the use of PCR testing for the detection of VZV-DNA in the CSF of patients with presumed zoster sine herpete. The literature is anecdotal, and it is difficult to regard zoster sine herpete as a diagnostic entity unless there is good evidence for VZV involvement, for instance as shown by the detection of VZV-DNA in CSF and/or blood mononuclear cells. It should be kept in mind, however, as a possible explanation for radicular pain of unknown cause: any possible mechanism is speculative.

Varicella zoster virus infection in the immunosuppressed patient

In patients with immunosuppression, particularly of cellular immunity, varicella can be much more severe (Plate 3). The skin lesions are more diffuse and can take up to three times as long to heal. There may be visceral dissemination to the lungs, liver, and central nervous system. Patients with lymphomas being treated with chemotherapy are particularly susceptible.

Herpes zoster in immunosuppressed patients is also more severe than in normal subjects. Prior to the availability of effective antiviral therapy, skin lesions were more extensive and could take several weeks longer to heal. Dissemination, presumably due to viraemic spread, with widespread skin lesions as in varicella, occurs in 10 to 40 per cent of patients. Cutaneous dissemination is more likely to be associated with visceral dissemination to the same sites as those associated with varicella.

Patients with HIV infection and the acquired immunodeficiency syndrome (**AIDS**) are prone to multidermatomal zoster, which can be one of the defining features of AIDS.

VZV retinitis

This presents with pain and blurred vision in one eye, with progressive necrotizing retinitis seen on ophthalmoscopy. Adjacent cutaneous zoster indicates the diagnosis, but occasionally VZV retinitis can present in immunocompetent patients as the sole manifestation of VZV reactivation. VZV retinitis may be difficult to distinguish from cytomegalovirus (**CMV**) retinitis. A severe form of the disease, seen particularly in patients with HIV infection, is known as progressive outer retinal necrosis: it is associated with a high incidence of retinal detachment and may require treatment with ganciclovir as aciclovir is often ineffective.

Differential diagnosis

Varicella is usually recognized relatively easily. Other causes of a vesicular rash are generalized herpes simplex in the immunosuppressed patient and enterovirus disease, particularly hand, foot, and mouth disease due to coxsackievirus infection, but the rash on the hands and feet is unlike that of varicella which has a centripetal distribution. Human cases of infection with animal pox viruses (monkey pox and camel pox) have rarely been described.

Pathology

Histological appearances of VZV infection are similar or indistinguishable from those of HSV infection.

Laboratory diagnosis

The diagnosis of varicella and herpes zoster are usually made on clinical criteria alone. Virus can be seen in vesicular fluid by electron microscopy or isolated in culture. Serological diagnosis of varicella can be made by demonstrating seroconversion or VZV-IgM antibody. Urgent serology is needed to confirm the seronegative status of contacts at risk of severe VZV infection, to determine the need for VZV immunoglobulin (see below). PCR-based tests for the detection of VZV-DNA are available, and are most useful in testing CSF in cases of suspected central nervous system disease.

Treatment

Pruritus may be alleviated in patients with chickenpox by calamine lotion and antihistamines. Fingernails should be closely cut to minimize scratching. Skin care is important to prevent secondary bacterial infection in patients with varicella and zoster. Aspirin should be avoided in children with chickenpox because of the risk of Reye's syndrome. Strong analgesia may be needed in patients with zoster.

VZV is sensitive to the nucleoside analogues, aciclovir (aciclovir), famciclovir, and valaciclovir: as for HSV, VZV encodes a thymidine kinase which preferentially phosphorylates these drugs in infected cells. The median 50 per cent inhibitory concentration of aciclovir against HSV is 0.1 µmol/l, but it is 2.6 µmol/l against VZV and consequently 800 mg orally is necessary to achieve inhibitory concentrations against VZV.

Treatment recommendations for varicella and herpes zoster are summarized in Table 2.

Varicella

Whether to treat normal children with varicella (who are the great majority of patients) has been much debated. The argument can be made that it is not possible to predict which child may have a severe case and the disease is not always mild. Therapy with aciclovir is safe and, although it has been suggested that widespread treatment with antivirals might result in viral resistance or failure to develop normal immune responses, there is no evidence of this in controlled trials. Treatment with aciclovir begun within 24 h of the onset of rash leads to a 25 per cent decrease in the duration and severity of chickenpox. The argument for treating all adolescents and adults is easier, as chickenpox is more severe for them than it is for young children. Chickenpox in neonates, children with leukaemia, and transplant recipients, should always be treated with aciclovir: intravenously administered aciclovir limits the visceral spread of the virus if given immediately on diagnosis. Treatment in these immunosuppressed patients can be changed from intravenous to oral aciclovir once the fever has settled, if there is no evidence of visceral varicella.

Herpes zoster

The major justification for the antiviral treatment of herpes zoster in immunocompetent patients has been to limit postherpetic neuralgia. Although there are difficulties in accurately and objectively quantifying the pain of postherpetic neuralgia, trial data indicates that acyclovir, valaciclovir, and famciclovir can limit the duration of zoster-associated pain and that valaciclovir is slightly more effective. Acyclovir, valaciclovir, and famciclovir accelerate the healing of cutaneous lesions by 2 days over placebo. Valaciclovir and famciclovir have the advantage of more convenient dosage, as well as being probably slightly more effective.

Patients with zoster over the age of 50 are at the highest risk of postherpetic neuralgia and should consequently be offered antiviral treatment. Patients younger than this may warrant treatment if they have marked pain. All patients with ophthalmic zoster should be treated with antiviral agents, even if they present relatively late, as acyclovir reduces the incidence of keratitis. Immunosuppressed patients with herpes zoster should receive intravenous acyclovir to prevent cutaneous and visceral dissemination of VZV. Valaciclovir and famciclovir may be used if zoster presents in a localized form in less severely immunosuppressed patients.

Corticosteroids have been advocated in patients with herpes zoster, in order to reduce the severity of postherpetic neuralgia. However, in recent trials the addition of oral prednisone to aciclovir resulted in a slightly increased rate of healing of skin lesions but did not affect the incidence of postherpetic neuralgia. The place of corticosteroids thus remains unproven.

Prevention and control

Varicella zoster immune globulin (VZIG), prepared from high-titre immune human serum, has been shown to prevent or ameliorate varicella in seronegative individuals at high risk, such as immunocompromised patients and pregnant women. Seronegative immunodeficient patients (including those on high-dose corticosteroid treatment) and pregnant women with definite contact with varicella are candidates for VZIG administration. It should be administered within 10 days (preferably 2–4) of exposure. Neonates whose mothers have had varicella less than a week before delivery or within 28 days after delivery are also recommended for VZIG administration.

A vaccine is available for VZV. This is the Oka strain of VZV, which is a live attenuated vaccine developed in Japan. It induces 90 per cent protection from natural varicella when administered to non-immune immunosuppressed individuals (such as patients with leukaemias and lymphomas treated with chemotherapy), but it can produce a vaccine-induced rash in up to 40 per cent of such recipients. In immunized healthy children the risk of subsequent varicella after community exposure is reduced to less than 5 per cent, and vaccine-induced rash is much less common (about 5 per cent of recipients). This vaccine is licensed in Japan, some European countries, and the United States, where it is recommended for routine immunization in children aged between 12 and 18 months, but not in the United Kingdom where it is available on a named patient basis only for use in non-immune immunosuppressed subjects. Trials are in progress to assess whether postinfective immunization with the vaccine can diminish the incidence of zoster in those over the age of 50 years: it has already been shown to boost pre-existing cell-mediated immunity to VZV in this age group.

Nosocomial transmission of VZV from those patients with varicella who require admission to hospital is a significant risk as 10 per cent of adults are seronegative. The nursing and management of inpatients with varicella should be restricted to those staff known to be seropositive for VZV. Patients with varicella in hospital should ideally be isolated in negative-pressure rooms to prevent airborne transmission.

Human cytomegalovirus infection

Historical introduction

The syndrome of congenital cytomegalovirus infection 'cytomegalic inclusion disease' was described in children with fatal infection in 1904, but the intranuclear inclusions were attributed to a protozoan parasite. In 1921, the pathologist Goodpasture suggested the inclusions in the parotid glands of infants were caused by a virus, because a filterable agent produced similar histology in guinea-pig salivary glands, and the lesions were attributed in 1926 to 'salivary gland virus'. Human cytomegalovirus (**HCMV**) was finally isolated in 1956, and so named by Weller for the characteristic 'owl's eye', or cytomegalic, inclusions it produces in the nucleus of

infected cells.

HCMV produces little morbidity in the immunocompetent, but can produce severe disease in the fetus if infection is acquired *in utero*, and in the immunosuppressed patient.

Aetiology

HCMV is the largest human herpesvirus with a linear double-stranded DNA genome of 250 kb encoding over 200 proteins. It can be grown in tissue culture in human fibroblasts. Mammalian cytomegaloviruses are species-specific, and so HCMV cannot be studied in animal models. The most widely studied laboratory strain, AD169, shows significant genomic variation from recent clinical isolates which possess an additional 15 kb of DNA. They can infect and replicate in macrophages, which are probably an important site of latency for this wild-type virus. HCMV replicates slowly compared to other herpesviruses: gene expression occurs in sequential immediate-early, early, and late phases.

Epidemiology

Following primary infection, HCMV persists for life as a latent infection with periodic asymptomatic excretion of virus in saliva, breast milk, urine, semen, and cervical secretions. Infection is spread by close contact with these body fluids. In less developed countries HCMV is usually acquired in childhood and seropositivity is nearly 100 per cent in young adults. In more developed countries seropositivity increases with age, but seroprevalence is higher in lower socioeconomic groups. Overall, about 50 per cent of adults are seropositive. During childhood, HCMV is acquired from breast milk or contact with other infected children excreting virus in their saliva or urine: studies in day nurseries have shown transmission between children as well as to susceptible adult carers. Later, sexual transmission becomes a major route of infection: seroprevalence approaches 100 per cent in homosexual men and sex workers.

Blood and blood products from seropositive donors can transmit HCMV. Transfusion recipients at risk of HCMV disease now usually receive screened seronegative blood, otherwise the risk of transfusion-related HCMV infection is 2.5 per cent per unit of blood. Transmission results from virus in leucocytes, but leucodepletion of blood (now being widely adopted as a preventive measure against transmissible spongiform encephalopathies) greatly reduces the risk of HCMV transmission. Finally, solid organ and bone marrow transplants from seropositive donors can transmit HCMV, and produce particularly severe disease in seronegative recipients.

Pathogenesis

Current evidence suggests myeloid lineage cells are a principal site of HCMV latency, and that virus may be reactivated when monocytes acquire a permissive phenotype as they differentiate into macrophages. Other cells, including endothelial cells and possibly epithelial cells, may also be sites of latency.

The immune response is critical in controlling infection in the normal host. Normal immunocompetent infected individuals mount a strong T-cell response, with very high frequencies of cytotoxic (CD8+) T lymphocytes in the peripheral blood particularly targeted at the HCMV major tegument protein (pp65) and the major immediate-early protein (IE1). Impairment of this response is associated with the risk of disseminated infection. HCMV possesses at least four genes whose products interfere with the class I MHC antigen-processing pathway, and a number of other potential 'immune evasion' genes, which may help the virus reactivate by delaying immune recognition of infected cells. Natural killer cells may be active early in the infection. Antibody probably limits the blood-borne dissemination of HCMV, as maternal IgG appears to be especially important in preventing viral transmission to the fetus. Subclinical reactivation occurs requently in the normal host but is controlled by the immune response. Immune deficiency, particularly of the T-cell response, as occurs with iatrogenic or disease-induced immunosuppression, may allow uncontrolled replication and result in HCMV disease. Pathology is presumably produced by direct cytopathic effects of the virus, although indirect effects produced by soluble virus-encoded proteins or the host response are also possible. The presence of HCMV in a diseased organ does not necessarily implicate the virus as a cause of disease, because reactivation of virus may sometimes be non-pathogenic and a 'bystander' effect to some other pathogenic process.

This difficulty in unequivocally attributing disease to HCMV is illustrated by its postulated role in arterial disease. HCMV has been detected in atherosclerotic lesions by immunohistology. A more recent study associated HCMV with the smooth muscle cell proliferation responsible for coronary artery restenosis following angioplasty, although subsequent reports failed to confirm this. Other microbial agents have now been described in atherosclerotic lesions in humans, and Marek's disease virus (also a herpesvirus) is associated with atherosclerosis in chickens. The association thus remains plausible but speculative.

Clinical features of HCMV disease

Primary infection in immunocompetent subjects

Primary infection in children and adults is asymptomatic in most cases, but HCMV can produce an illness clinically indistinguishable from infectious mononucleosis caused by primary Epstein–Barr virus (**EBV**) infection, characterized by fever, myalgia, cervical lymphadenopathy, and mild hepatitis. Tonsillopharyngitis is much less common than in primary EBV infection, and lymphadenopathy and splenic enlargement are less prominent features. The fever lasts 2 to 3 weeks, but it can persist for up to 5 weeks. In more developed countries an increasing proportion of HCMV seroconversion illness is seen in older adults, and the diagnosis should still be considered in patients over 50 or 60 years of age. Myocarditis, pneumonitis, and aseptic meningitis are rare complications. A proportion (5–10 per cent) of patients with Guillain–Barré syndrome (**GBS**) show serological evidence of primary HCMV infection: they are more likely to have antibodies to the GM2 ganglioside than other patients with GBS, and a causal relationship is postulated.

Primary HCMV infection acquired from blood transfusion is characterized by a similar clinical picture occurring 3 to 6 weeks after transfusion, and is usually self-limiting in the normal host. The distinction of primary HCMV infection from other causes of mononucleosis syndromes (such as EBV and toxoplasmosis) depends on serological testing (the Paul-Bunnell and Monospot tests are negative in HCMV mononucleosis).

HCMV disease in the immunosuppressed patient

HCMV produces its most severe disease in immunosuppressed patients, particularly solid organ and bone marrow transplant (**BMT**) recipients, and those with AIDS, all characterized by impaired T-lymphocyte function: this strongly supports the importance of T cells in controlling infection.

Disease in solid-organ transplant recipients

The risk of HCMV disease is three- to fivefold greater in a seronegative recipient receiving a graft from a seropositive donor than in a seropositive recipient, and disease is much more severe. Many centres 'match' seronegative donors to seronegative recipients, although this is often thwarted by organ shortage. Disease presents with specific organ involvement not seen in the normal subject. Interstitial pneumonitis due to HCMV carries a poor prognosis; gastrointestinal disease includes oesophagitis, gastritis and peptic ulceration, and colitis; HCMV retinitis may occur in severely immunosuppressed patients. HCMV has been reported to be associated with increased graft rejection and renal artery stenosis in renal transplant recipients, with accelerated coronary artery stenosis in heart transplant recipients, and with 'vanishing bile duct' syndrome in liver transplant recipients. However, none of these associations is definitively established as causal.

Disease in bone marrow transplant (BMT) recipients

HCMV disease is a major problem in allogeneic BMT recipients, with a 30 to 50 per cent incidence of clinically significant infection. It is a lesser problem in autologous BMT. Seropositivity in donor or recipient, or both, carries a risk of HCMV disease, but the risk can be eliminated when both donor and recipient are seronegative if HCMV seronegative blood products are used to support the patient. Pneumonitis is the most serious manifestation of HCMV infection after BMT, occurring in 10 to 15 per cent of allogeneic BMT recipients, with a mortality of 80 per cent without antiviral therapy. There is interstitial pneumonitis in the absence of any other identifiable pathogen, with increasing arterial hypoxaemia, and progression to respiratory failure. It is suggested that graft-versus-host disease (**GvHD**) may contribute to the lung injury in HCMV pneumonitis in BMT recipients. The relationship of HCMV to GvHD is controversial, with propositions that HCMV may predispose to GvHD, and vice versa.

Disease in patients with AIDS

HCMV disease is one of the most frequent opportunistic infections in patients with advanced HIV infection, of whom 40 per cent develop sight- or life-threatening HCMV disease. A CD4 count below 50/µl carries a high risk of disease, although the widespread use of antiretroviral therapy in developed countries means that relatively few patients now have such low CD4 counts, and the incidence of HCMV disease in patients with AIDS has consequently declined significantly.

HCMV retinitis was seen in up to 25 per cent of patients with AIDS prior to effective antiretroviral therapy (Plate 5). Characteristically, haemorrhagic retinal necrosis spreads along retinal vessels and threatens sight when disease encroaches on the macula. Patients present with visual impairment and have an increased risk of retinal detachment and haemorrhage: hence those with low CD4 counts should undergo regular examination of the optic fundi to detect retinitis before it becomes symptomatic. Diagnosis is made by the ophthalmological detection of typical retinal changes, preferably with accompanying evidence of HCMV viraemia. Without treatment, HCMV retinitis almost invariably progresses to affect both eyes and destroy vision.

HCMV is reported to produce a diffuse encephalitis in AIDS patients but, although the virus is sometimes seen in neuronal cells at autopsy, encephalitis attributable to HCMV is relatively rare in clinical practice in comparison to the other causes of encephalitis in those with AIDS. HCMV can also produce a progressive radiculopathy causing low back pain, which radiates to the area supplied by the affected spinal nerve root, and the development of flaccid paraparesis.

In the gastrointestinal tract, HCMV is associated with oesophagitis, gastritis, and enterocolitis. Virus can be seen in biopsies from these sites, usually in shallow ulcers.

HCMV pneumonitis is rare in patients with AIDS, suggesting there must be additional factors to account for its frequency in BMT recipients.

Congenital and neonatal HCMV infection

HCMV infection of the neonate may be congenital from intrauterine infection, perinatal transmission during birth, or postnatal from breast milk. The frequency of congenital HCMV infection in developed countries is around 0.5 to 1 per cent of live births: it results from either primary maternal infection in pregnancy, or from reactivation of HCMV during pregnancy in a previously infected mother. The risk of primary maternal infection in pregnancy is about 1 per cent, and it carries a 40 per cent risk of congenital infection. Fetal infection is more likely to occur, and to be severe, when a seronegative mother acquires primary infection in early pregnancy. The risk of symptomatic congenital infection from reactivation of maternal HCMV in pregnancy is much lower, although not absent. Pre-existing maternal immunity limits spread to the fetus.

Approximately 5 to 20 per cent of congenitally infected babies are symptomatic at birth: the higher figure applies to babies of mothers with primary infection, who are also more likely to have serious disease. In its most severe form congenital HCMV infection is associated with microcephaly, chorioretinitis, nerve deafness, hepatitis with jaundice and hepatosplenomegaly, and thrombocytopenia with petechiae: such classical 'cytomegalic inclusion disease' has a high mortality, and 80 per cent of all infants symptomatic at birth who survive have serious sequelae such as mental, visual, and hearing impairment. However, the majority of congenitally infected babies are asymptomatic at birth: only 5 to 15 per cent subsequently develop sequelae on long-term follow-up, the commonest being sensorineural deafness, which occurs in isolation in otherwise normal babies.

Perinatal or postnatally acquired HCMV infection is rarely symptomatic or associated with long-term sequelae, if the mother is seropositive.

Malignancy

Although associations between HCMV and malignancy have been postulated in the past, there is currently no good evidence to associate the virus with any human malignancy.

Pathology

On light microscopy, typical HCMV-infected cells appear large with a relative reduction in cytoplasm, and nuclei that contain prominent intranuclear inclusions, surrounded by a clear halo (described as 'owl's eye' inclusions). These cells contain replicating virus and are associated with active infection and disease. They are diagnostic when seen in biopsies of affected organs. In patients dying of severe disease, histological evidence of HCMV involvement can be found in most organs, whereas it infects a restricted range of cells *in vitro*.

Laboratory diagnosis

Primary infection is usually diagnosed by the detection of IgM antibody to HCMV in the absence of IgG antibody: there is a marked atypical lymphocytosis (mainly due to increased CD8+ T cells) but heterophile antibody (as detected in primary EBV infection by the Monospot or Paul-Bunnell tests) is absent. Serology is of limited use in confirming HCMV disease in the immunosuppressed patient: IgG antibody is a useful marker of HCMV carriage, but titres do not rise reliably in disease. IgM antibody is found during primary infection, and also sometimes with reactivation in patients who are immunosuppressed. Culture of virus from urine may only indicate asymptomatic reactivation; culture from the blood buffy coat is more definitively associated with HCMV disease as virus can never be cultured from the blood of normal HCMV carriers, and culture from an organ site (such as bronchoalveolar lavage fluid) may indicate locally active disease. Rapid culture methods, such as the **DEAFF** (detection of early antigen fluorescent foci, by a monoclonal antibody against an immediate-early viral protein) or shell vial tests facilitate virus isolation. The HCMV antigenaemia assay detects the presence of the HCMV pp65 protein in peripheral blood neutrophils (where it may be taken up passively, rather than expressed by natural infection): the number of positive cells correlates with the level of viraemia. PCR-based techniques are increasingly used to detect and quantitate the HCMV load in blood or plasma, and in many laboratories are now the standard assay for detecting HCMV. As virus can never be detected in plasma (as opposed to leucocytes) in normal carriers, the presence of HCMV-DNA in plasma indicates active viral replication. Detection of virus in biopsy specimens by histological and immunohistological techniques implies organ disease due to HCMV.

In practice, HCMV disease is usually diagnosed by the combination of an appropriate clinical syndrome, and HCMV detection in blood or plasma, or in biopsies from involved organs, in the absence of any other likely causal microbial pathogen.

Treatment

Several drugs are now available for the treatment of disease due to HCMV. Aciclovir has little *in vitro* activity against HCMV, which, unlike HSV, does not possess a thymidine kinase (see above), and has no place in its therapy (although it is used in prophylaxis—see below). Ganciclovir, another nucleoside analogue, is monophosphorylated in infected cells by the *UL97* gene product of HCMV, and is active against HCMV: its most limiting side-effect is myelotoxicity with leucopenia and thrombocytopenia, but it has many other potential side-effects including azoospermia in males. Intravenous administration is necessary since an oral form of ganciclovir has bioavailability of only about 10 per cent, restricting its use to prophylaxis (oral valganciclovir, a new valyl ester of ganciclovir, has much higher bioavailability: it also produces equivalent plasma concentrations to intravenously administered ganciclovir, which initial trials suggest it may replace). Resistance to ganciclovir results from mutations in the drug target, the HCMV-DNA polymerase, or in the *UL97* gene, and is seen mainly in AIDS patients in whom prolonged use is necessary. An alternative drug to ganciclovir is foscarnet (trisodium phosphonoformate), which is a competitive inhibitor of the viral DNA polymerase and shows no cross-resistance with ganciclovir. This also has to be given intravenously and its side-effects include renal impairment and hypocalcaemia. Cidofovir, a nucleotide analogue acting on the viral DNA polymerase is also licensed for use in the United Kingdom, but is highly nephrotoxic (probenecid has to be given concurrently to prevent irreversible renal damage) and therefore relatively little used.

Primary infection

In the immunocompetent host, this requires no specific antiviral treatment.

HCMV disease in the immunosuppressed (due to primary or secondary infection, or reactivation)

This is usually treated with ganciclovir or foscarnet for 2 to 3 weeks of full-dose induction intravenous therapy: for ganciclovir this is 5 mg/kg body weight every 12 h;

and for foscarnet 60 mg/kg every 8 h. Secondary prophylaxis may well be needed if immunosuppression persists (see below).

HCMV pneumonitis in BMT recipients

HCMV pneumonitis responds poorly to ganciclovir or foscarnet alone, but the combination of full-dose ganciclovir with intravenous immunoglobulin has been shown to reduce mortality. Although specific anti-CMV immunoglobulin was initially used, recent trials suggest normal pooled intravenous immunoglobulin is equally effective. Many centres monitor BMT recipients, especially of allogeneic grafts, for CMV viraemia and if detected commence 'pre-emptive therapy' with ganciclovir prior to the development of symptomatic or obvious organ disease.

HCMV retinitis in patients with AIDS

This is treated with an induction course of ganciclovir or foscarnet (both drugs have also been used in combination). Continued prophylaxis is needed to prevent relapse unless significant recovery of the CD4 count can be induced with antiretroviral therapy: for this, intravenous ganciclovir—5 mg/kg per day for 5 days per week—is effective, but high-dose oral ganciclovir (1000 mg three times per day) may be adequate if retinal disease is peripheral to the macula. Given the difficulty of these regimes, implantable intraocular devices giving a sustained release of ganciclovir into the vitreous humor have also been used. The use of combination antiretroviral therapy in HIV-infected patients is associated with a much improved long-term control of HCMV infection.

However, the syndrome of 'immune recovery vitritis', characterized by posterior segment inflammation, can occur in patients with inactive treated CMV retinitis as their CD4 count reconstitutes on antiretroviral therapy.

Congenital HCMV infection

In a phase II evaluation of ganciclovir (8 or 12 mg/kg body weight daily for 6 weeks) for the treatment of symptomatic congenital HCMV infection, excretion of CMV in the urine decreased: however, after cessation of therapy viruria returned to near-pretreatment levels. Hearing improvement occurred in 5 out of 30 babies at 6 months or later, suggesting some efficacy, but the role of antiviral therapy in congenital HCMV infection remains to be established.

Prevention and control

The problem posed by HCMV in immunosuppressed patients has led to several approaches to prophylaxis.

Antiviral prophylaxis

Ganciclovir has been used for primary prophylaxis in solid-organ and BMT recipients, particularly those at high risk of disease (seronegative recipients of a seropositive graft, or seropositive recipients), and in AIDS patients with less than 100 CD4 cells/μl. Oral ganciclovir has been shown to be effective in many of these settings. Despite their limited *in vitro* activity against HCMV, and lack of efficacy as therapy, oral aciclovir and valaciclovir have also been shown to provide significant prophylaxis against HCMV disease in renal transplant recipients. Moreover, valaciclovir prophylaxis was also associated with a lower rate of graft rejection. This evidence, combined with their lack of toxicity compared to ganciclovir, has led to their widespread use.

Passive immunization

CMV hyperimmune globulin has been reported to reduce the risk of HCMV disease in renal transplant recipients but is expensive and little used in practice.

There are initial reports that HCMV-specific T-cell immunity can be reconstituted in BMT recipients by the adoptive transfer of virus-specific T lymphocytes from the immune donor, but this is still a research therapy.

Active immunization

A live laboratory (Towne) strain of HCMV has been tested as an experimental candidate vaccine in renal transplant recipients and found to have some evidence of protective immunity, perhaps equivalent to having previous natural HCMV infection. However, there is currently no available licensed vaccine, although some candidates are in early development.

Special problems in pregnant women

Pregnant women who are seronegative should avoid contact with possibly infected children in day nursery settings, although this may be impractical. This population would be the target for a vaccine were one available. Ganciclovir must not be used in pregnancy.

Human herpesvirus-6 and -7

Human herpesvirus-6

Introduction

Human herpesvirus-6 (**HHV-6**) was originally isolated in 1986 from cultured human lymphocyte lines and initially named 'human B lymphotropic virus'. However, it was subsequently shown to be tropic principally for T cells, although it can also replicate in macrophages, glial cells, and EBV-transformed B cells. HHV-6 is widely distributed in humans and primary infection is associated with roseola infantum (also known as exanthem subitum or sixth disease), an aetiological association first described in Japanese children by Yamanishi and colleagues in 1988.

Aetiology

HHV-6 has typical herpesvirus morphology and is genetically classified in the betaherpesvirus subfamily. Two groups of HHV-6 isolates, HHV-6A and HHV-6B, are now clearly distinguished by their genetic sequence and some variation in their biological properties. HHV-6B is associated with roseola, whilst HHV-6A has not been clearly associated with human disease.

Epidemiology

There is high seroprevalence of HHV-6 in all populations. More than 90 per cent of children are seropositive at 2 years of age. The virus (usually the HHV-6B variant) can be detected in peripheral blood mononuclear cells by PCR in nearly all healthy people. It is most probably transmitted via maternal saliva, although intrauterine and perinatal transmission could occur. The virus is not detectable in breast milk.

Pathogenesis

Upon primary infection with HHV-6, the virus probably replicates in regional lymphoid tissue in the oropharynx and can be found in circulating lymphocytes. HHV-6 replicates *in vitro* in CD4+ T-cell lines. However, during persistent infection in a normal adult, the virus can be detected by PCR in both CD4+ T cells and in monocytes/macrophages in peripheral blood. Monocytes/macrophages are probably the principal site of carriage during persistent infection. The mechanism of viral latency is uncertain. HHV-6 induces CD4 expression on $CD4^-$ lymphocytes, and *in vitro* may thus facilitate HIV entry into previously $CD4^-$ cells.

Although HHV-6 cannot normally be isolated in culture from the peripheral blood of normal individuals, it is easy to detect HHV-6 DNA in the peripheral blood of immunosuppressed subjects: this and other evidence indicates immunosuppression is associated with reactivation of HHV-6. Mechanisms by which HHV-6 may

produce clinical manifestations remain unclear.

Clinical features

Primary infection with HHV-6 in young children is associated with roseola and also with a febrile illness without rash.

Roseola infantum (exanthem subitum, sixth disease)

Roseola is an acute illness of infants and young children, characterized by 3 to 5 days of high fever with upper respiratory tract symptoms and sometimes cervical lymphadenopathy. As the fever defervesces, a rash appears and lasts for 1 to 3 days. The rash is diffuse, macular or maculopapular, and appears similar to that of rubella. The illness is accompanied by mild atypical lymphocytosis and there may be neutropenia. Rarely, infections are complicated by febrile convulsions, meningitis, encephalitis, and hepatitis, which is usually mild but occasionally severe.

Roseola has been estimated to occur in only 10 to 20 per cent of children, as primary HHV-6 infection is commonly subclinical.

Febrile illness

Fever without rash is a more usual manifestation of primary HHV-6 infection than roseola. In one major study of 1600 children under the age of 3 years presenting to a North American hospital emergency department with acute febrile illness, 10 per cent of cases were ascribed to primary HHV-6 infection. In children between the ages of 6 and 12 months within this study, 20 per cent of acute febrile illness was due to HHV-6, but only 17 per cent of all these children with documented primary HHV-6 infection had clinical roseola.

Febrile convulsions

There is accumulating evidence that HHV-6 has a major association with febrile convulsions in young children. In the study quoted above, 13 per cent of all the children under 3 years of age had febrile convulsions associated with acute HHV-6 infection, and the infection was reported to account for one-third of all febrile seizures in children up to the age of 2 years. It is believed that this association is not solely because HHV-6 induces high fever, but because it also specifically infects the nervous system. HHV-6 DNA can be detected in the cerebrospinal fluid of children with primary infection.

HHV-6 infection in immunosuppressed patients

A number of studies have shown increases in antibody titres to HHV-6 and increased HHV-6 DNA levels in peripheral blood by PCR in immunosuppressed solid-organ and bone marrow transplant recipients. In bone marrow transplant recipients, HHV-6 has been associated with fever, skin rash, graft-versus-host disease, encephalitis, delayed engraftment, marrow suppression, and pneumonitis: however, it is not clear whether HHV-6 plays a specific aetiological role in these syndromes. In the case of pneumonitis, BMT recipients have higher levels of HHV-6 DNA in the lung, but other opportunist infections such as human cytomegalovirus were not always excluded as the cause.

There is good evidence that HHV-6 reactivates in patients with advanced HIV infection and AIDS. However, there is no firm evidence for any HHV-6 associated disease in patients with AIDS.

In summary, current evidence suggests HHV-6 is infrequently associated with disease in immunosuppressed patients, but that it could be responsible for occasional cases of pneumonitis in BMT recipients.

Other disease associations

Some studies have associated HHV-6 with the chronic fatigue syndrome and with multiple sclerosis. The present consensus is that there is no convincing evidence for any significant aetiological association between HHV-6 and these illnesses.

Malignancy

HHV-6 DNA has been detected in the blood of patients with a number of lymphoproliferative disorders, but this probably reflects reactivation rather than any causal association with the tumour. There are reports of HHV-6 DNA being detected in some tumours, such as the nodular sclerosis variant of Hodgkin's disease, but there is no convincing aetiological association between HHV-6 and any tumour.

Differential diagnosis

Primary HHV-6 infection may be confused with many febrile childhood illnesses associated with a rash. The rash associated with roseola may mistakenly be attributed to sensitivity to recent antibiotic treatment. Other virus infections (EBV, HCMV) may also be associated with atypical lymphocytes and a mononucleosis syndrome.

Pathology

HHV-6 replicates in cells of central nervous system origin, particularly glial cell lines, *in vitro*. HHV-6 DNA can be detected in the brain of apparently normal individuals, suggesting viral persistence in the central nervous system. No distinctive histopathology has yet been attributed to HHV-6.

Laboratory diagnosis

Commercial assays for HHV-6 antibody do not distinguish between antibody to HHV-6A and HHV-6B and may cross-react with antibodies to HHV-7. Seroconversion is evidence of primary infection. IgM assays for HHV-6 antibody are not reliable indicators of primary infection as some HHV-6 carriers may have IgM antibody periodically.

Although HHV-6 can be cultured from peripheral blood mononuclear cells during acute primary infection, few laboratories will undertake this. PCR-based techniques for the detection of HHV-6 DNA in plasma and cerebrospinal fluid are the method of choice for clinical diagnosis, and are becoming increasingly available.

Treatment

HHV-6 sensitivity to antiviral drugs corresponds to the sensitivity of cytomegalovirus. Thus, HHV-6 replication is inhibited *in vitro* by ganciclovir and foscarnet but not aciclovir, but there are no controlled clinical trials of these agents. Their use may be considered for an immunosuppressed patient in whom HHV-6 associated pneumonitis is suspected.

Prevention and control

There is currently no place for measures to prevent HHV-6 transmission. It seems unlikely there will be a case for the development of a vaccine because infants may be infected so early in life, while they still have maternal antibody.

Special problems in pregnant women

Nearly all pregnant women will be carriers of HHV-6. There is no evidence to associate HHV-6 with a specific risk to the fetus or neonate.

Human herpesvirus-7

Human herpesvirus-7 (HHV-7) was isolated in 1990. It is also a betaherpesvirus, similar to but distinct from HHV-6. HHV-7 predominantly infects CD4+ T cells and can be reactivated from latency upon T-cell activation.

Although there is serological cross-reactivity between HHV-6 and HHV-7, the evidence is that HHV-7 infects nearly all humans during childhood (but later than HHV-6), with greater than 90 per cent of children being infected by the age of 5 years. The virus is excreted in saliva.

HHV-7 has been associated with some cases of roseola, and in a Japanese study was reported to cause roseola in infants who had already had a previous episode of roseola proven to be due to HHV-6. The similarity between HHV-7 and HHV-6 suggests that they may be associated with a similar range of diseases, but the association with roseola is the only one so far identified.

The best method of diagnosis is by PCR on serum or CSF. Laboratory tests for HHV-6 often also test for HHV-7 in a multiplex PCR. In the absence of any disease associations, apart from that with roseola, there is no reason to consider any treatment for HHV-7.

Human herpesvirus-8

Introduction

Human herpesvirus-8 (HHV-8) is the most recently isolated of the human herpesviruses. In 1994 Chang and colleagues, using the technique of representational difference analysis (which selectively amplifies DNA present in diseased tissue, but not the corresponding normal tissue), reported the detection of novel DNA sequences with homology to herpesviruses in Kaposi's sarcoma tissue. This herpesvirus is most closely related genetically to a well-characterized simian herpesvirus (herpesvirus saimiri) and less so to EBV, and it was consequently assigned to the g2-herpesvirus subfamily. It was initially named Kaposi's sarcoma-associated herpesvirus (**KSHV**), but was subsequently designated HHV-8. Current culture techniques are unreliable, but the virus can be detected by PCR. Serological assays depend on the use of infected cell lines or synthetic antigens from predicted open-reading frames. The seroepidemiology, biology, and disease associations of the virus are still being analysed, but HHV-8 is clearly closely associated with Kaposi's sarcoma, a tumour which has long been suspected of having a viral aetiology, with primary effusion lymphoma, and with Castleman's disease. Reported associations with multiple myeloma and other cancers are unconfirmed.

Aetiology

HHV-8 has the characteristic morphology of a herpesvirus. The viral genome is composed of a 141-kbp long unique segment flanked by multiple 801-bp direct repeats. Sequence analysis suggests that HHV-8, like other herpesviruses, is an ancient human virus, with several major subtypes reflecting the migrationary divergence of human populations. HHV-8 contains eight genes homologous to mammalian genes encoding cell-cycle regulatory proteins (the cyclins), chemokines, and inhibitors of apoptosis. Comparative genetic analyses show a high rate of amino acid variation in *ORF-K1* and another gene, *K15*. Analysis of the variable genes indicates there are at least four virus subtypes, A–D. On the evidence to date the normal cellular site of latency of HHV-8 almost certainly includes the B cell.

Epidemiology

The emerging epidemiology of HHV-8 suggests it is less ubiquitous than other human herpesviruses. Initial serological assays using indirect immunofluorescence on infected cell lines to detect antibodies to a latent nuclear antigen (**LANA**), give a seroprevalence of approximately 80 per cent in patients with Kaposi's sarcoma, and 25 to 30 per cent in HIV-positive homosexual men without Kaposi's sarcoma. Seroprevalence in normal adults is reported as being more than 50 per cent in African adults in West Africa, 20 per cent in Black South African blood donors, and less than 5 per cent in blood donors in the United Kingdom and United States, with intermediate rates in Italy and other Mediterranean countries. Seroprevalence in HIV-positive and -negative adults in Uganda was equal at 53 per cent. HHV-8 can be detected by PCR in nearly all cases of Kaposi's sarcoma, so the failure of this assay to detect antibody in all cases of Kaposi's sarcoma implies it has limited sensitivity. An assay using lytic-cycle antigens gave higher rates of seroprevalence, but these may result from cross-reaction with EBV antibodies. Newer assays using multiple HHV-8 antigens are currently being applied.

The normal route of transmission of the virus is unknown, but sexual transmission occurs between homosexual men. A LANA-based assay detected seroconversion to HHV-8 in HIV-infected homosexual men at a median of 33 months before they subsequently developed Kaposi's sarcoma. HHV-8 infection in children correlates with seropositivity in their mothers, but whether this reflects vertical or horizontal transmission is unknown.

Pathogenesis

There has been much uncertainty over the cell of origin of Kaposi's sarcoma, but the spindle cells of which the tumour is largely composed are thought to be of lymphatic endothelial origin. In Kaposi's sarcoma tumour tissue, HHV-8 DNA and LANA are present in every spindle cell, suggesting an aetiological role for the virus.

In HIV-associated Castleman's disease, the HHV-8 LANA antigen is present in immunoblasts in the mantle zone of the tumour. HHV-8 is present in the tumour cells of all cases of primary effusion lymphoma so far studied, although so also is EBV. HHV-8 latently infected cell lines derived from these tumours can be induced *in vitro* to release infectious virus, and are used to detect antibodies to the virus.

These clear associations of virus DNA with tumour cells suggest a definite oncogenic role for HHV-8. The possession of the genes encoding cyclin and antiapoptotic protein homologues suggests they may be involved in cellular transformation and oncogenesis by HHV-8.

It has been suggested that HHV-8 may be involved in the pathogenesis of multiple myeloma. The virus has been reported in bone marrow dendritic stromal cells of myeloma patients, but not in myeloma cells; however, this putative association remains speculative. The individual HHV-8 subtypes are not associated with any distinct pathology.

Clinical features

Apart from these malignancies, the only reported clinical syndrome accompanying primary or reactivated HHV-8 infection is fever and bone marrow failure in immunosuppressed transplant recipients.

Kaposi's sarcoma

Kaposi's sarcoma manifests clinically as purplish brown macules, papules, or plaques. It is described in four characteristic clinical settings: the classic form in elderly Mediterranean or Jewish males, the endemic African form (accounting for 10 per cent of cancer in equatorial Africa), in patients with immunodeficiency states such as transplant recipients, and the AIDS-associated form. The classic and African forms are characterized by lesions on the extremities, systemic and mucosal involvement is rare, and the disease is indolent. In immunosuppressed patients (other than those with AIDS) lesions are more widespread and more rapidly progressive, although visceral involvement is still unusual, and lesions may regress if immunosuppressive drugs are stopped. AIDS-associated Kaposi's sarcoma is seen predominantly in homosexual men in Western countries, but is commonly associated with heterosexually acquired HIV infection in African countries: it is characterized by widespread cutaneous lesions with involvement of the oral mucosa (Plate 6), and visceral lesions may occur in the lungs or gastrointestinal tract. Progression can be much more rapid than the other forms. HHV-8 has been isolated from all these forms of Kaposi's sarcoma.

Primary effusion lymphomas

Previously known as body cavity-based lymphomas, these are a rare and aggressive type of B-cell lymphoma presenting in patients with AIDS as lymphomatous effusions of the peritoneal, pleural, or pericardial spaces, usually with no identifiable tumour mass. HHV-8 is present in the tumour cells of all cases so far studied, although so also is EBV.

Castleman's disease

Also known as angiofollicular lymph node hyperplasia, Castleman's disease can be localized and is amenable to curative excision. However, a multicentric form is seen particularly in HIV-infected patients and is more aggressive: HHV-8 is found in a high proportion of these multicentric cases, especially those associated with HIV.

Pathology

No distinctive histopathology, independent of the pathology of the tumours with which it is associated, has been identified as attributable to HHV-8 infection in cells.

Laboratory diagnosis

HHV-8 can best be detected by PCR-based tests. Antibody assays are described above, and may become commercially available in the near future.

Treatment

Assays based on the HHV-8 infected lymphoma cell lines suggest that HHV-8 replication is moderately sensitive to foscarnet, ganciclovir, and cidofovir. It has been suggested that patients with AIDS treated with foscarnet and ganciclovir may have been less likely to develop Kaposi's sarcoma. There is no established use of antiviral drugs in treating HHV-8 tumours.

The treatment of Kaposi's sarcoma is discussed in Chapter 7.10.24. Kaposi's sarcoma confined to the skin can be treated with radiotherapy or with intralesional interferon-alpha. More widespread cutaneous or visceral disease can be treated with single-agent or combination chemotherapy.

Kaposi's sarcoma lesions may regress with antiretroviral treatment, possibly due to the improved cellular immunity that occurs with a reduction of the HIV load.

Prevention and control

Given limited knowledge of the epidemiology and disease associations of HHV-8, no prevention or control measures are used. No special problems of infection have been identified in pregnant women.

Cercopithecine herpesvirus-1 (herpes B virus)

Introduction

Cercopithecine herpesvirus-1 is the formal name now given to herpes B virus (also previously known as herpesvirus simiae, a term no longer used), whose natural hosts are members of the *Macaca* genus of Old World monkeys. It produces minimal disease in its natural host, but its transmission to humans results in a high incidence of severe disease. Although more than 30 other herpesviruses have been isolated from non-human primates, none of these have been unequivocally associated with a disease in humans. The virus was first isolated in 1932 from the brain of Dr WB who died of encephalitis after a bite from a macaque monkey (hence the name herpes B virus). There have since been about 45 cases of human infection resulting from accidental transmission from captive monkeys.

Aetiology

Herpes B virus is an alphaherpesvirus closely related to HSV and appears to behave in an analogous manner to HSV in its natural primate host. Herpes B virus can also infect and produce disease in other non-human primates and small mammals.

Epidemiology

Herpes B virus is enzootic in Old World macaque monkeys, principally rhesus (*Macaca mulatta*) and cynomolgus (*Macaca fascicularis*) macaques. The epidemiology in its primate host is similar to that of HSV in humans, with 80 per cent or more of natural and captive adults monkeys being infected. Infected monkeys may develop vesicular oral lesions and can shed virus intermittently from oral, conjunctival, and genital secretions.

Rhesus and cynomolgus macaques have been quite widely used in medical research, particularly for polio vaccine development during the mid-1950s and for studies of retroviruses from the late 1980s following the AIDS epidemic. Nearly all the reported human cases resulted from occupational exposure through bites and scratches in workers handling monkeys, but transmission from needlestick injuries and a splash in the eye have also been reported. One case of human-to-human transmission apparently occurred by inoculation on to inflamed skin.

Two clusters of infections have been described in the United States (in 1987 and 1989), the earlier one involving the case of human-to-human transmission. A seroprevalence study of over 300 monkey handlers revealed no seropositive subjects, and asymptomatic infection documented by seroconversion appears to be extremely uncommon.

Clinical features

The incubation period from occupational exposure to the development of symptoms has been 3 to 5 days in most cases, but it can range from 3 to 30 days. Cutaneous vesicles may occur at or near the site of inoculation accompanied by regional lymphadenitis. A prodrome of fever, malaise, headache, and abdominal pain occurring in the first 2 weeks is common, but the dominant and characteristic feature of reported cases is a progressive myelitis and encephalitis. Herpes B virus produces a multifocal haemorrhagic myelitis and encephalitis. Visceral spread is recorded in fatal cases. Before the advent of aciclovir and later antiviral drugs, the mortality rate was 70 per cent.

It is not clear whether herpes B virus can establish latency and then reactivate in humans. Viral shedding has recurred when antiviral treatment was stopped relatively early, so most patients have been maintained on antivirals for long periods.

Laboratory diagnosis

Herpes B virus is a category IV pathogen and only a few designated laboratories can undertake culture and isolation of the agent. In the United Kingdom the designated laboratory is at the Central Public Health Laboratory, Colindale, London, and in the United States at the South West Foundation for Biomedical Research, San Antonio, Texas. Suspected infected monkeys should be bled to determine seropositivity. Serodiagnosis in humans is made difficult because of antigenic cross-reactivity between herpes B virus and herpes simplex virus. The inoculation site should preferably be biopsied for possible culture and analysis. PCR-based methods are available in specialized centres and are the standard for definitive diagnosis.

Treatment

Suspected injuries from macaques carry the risk of herpes B virus infection, although most captive macaque colonies are now maintained free of the virus. A suspected contaminated wound should be debrided and cleaned with chlorhexidine or iodine soap. There may be a case for initiating immediate antiviral treatment if infection in the monkey is suspected or for a deep wound. Otherwise presumptive therapy may be initiated if the monkey is subsequently shown to be positive for herpes B virus, although the report of transmission by an eye splash favours early presumptive treatment.

Aciclovir and ganciclovir both inhibit herpes B virus replication *in vitro*. For presumptive therapy, oral aciclovir 800 mg, 5 times daily, or preferably valaciclovir in

equivalent dose can be given for at least 2 weeks.

If symptomatic disease is suspected, intravenous aciclovir should be used (10 mg/kg body weight every 8 h for peripheral disease, and 15 mg/kg every 8 h for central nervous system involvement). If progression occurs, ganciclovir (5 mg/kg every 12 h) should be considered as an alternative. Treatment has been associated with limitation of disease and recovery in some patients.

Prevention and control

Those working with macaques should follow standard procedures to avoid infection. Screening of newly imported monkeys, and the creation of colonies of macaques free of herpes B virus, are now becoming standard practice.

Further reading

Herpes simplex virus infection

Balfour HH, Jr (1999). Review article: drug therapy: antiviral drugs. *New England Journal of Medicine* **340**, 1255–68. [A good review of antiviral therapy including coverage of drugs for HSV.]

Lakeman FD, Whitley RJ (1995). Diagnosis of herpes simplex encephalitis: application of polymerase chain reaction to cerebrospinal fluid from brain-biopsied patients and correlation with disease. NIAID collaborative antiviral study group. *Journal of Infectious Disease* **171**, 857–63. [Study showing detection of HSV-DNA by PCR in 98 per cent of 54 patients with brain-biopsy proven HSV encephalitis.]

Langenberg AGM, *et al*. (1999). A prospective study of new infections with herpes simplex virus type 1 and type 2. *New England Journal of Medicine* **341**, 1432–8. [A recent study of incident HSV-1/2 infections (undertaken for an unsuccessful vaccine trial), which reports the proportion of symptomatic infections.]

Roizman B, Pellett PE (2001). Herpesviridae. In: Knipe DM, Howley PM, eds. *Fields virology*, pp 2381–48. Lippincott, Williams and Wilkins, Philadelphia. [Authoritative chapter on the herpesviruses in major virology text, accompanied by another on the basic virology of HSV.]

Tookey P, Peckham CS (1996). Neonatal herpes simplex virus infection in the British Isles. *Paediatric and Perinatal Epidemiology* **10**, 432–42. [Comprehensive survey of neonatal HSV in the UK over a 5-year period.]

Whitley RJ (2001). Herpes simplex viruses. In: Knipe DM, Howley PM, eds. *Fields virology*, pp 2461–510. Lippincott, Williams and Wilkins, Philadelphia. [Chapter on clinical aspects of HSV in major authoritative virology text.]

Whitley RJ *et al*. (1986). Vidarabine versus acyclovir therapy in herpes simplex encephalitis. *New England Journal of Medicine* **314**, 144–9. [The classic original trial showing efficacy of aciclovir in, and probably the largest case series of, HSV encephalitis.]

Varicella zoster virus infection

Arvin AM (2001). Varicella-zoster virus. In: Knipe DM, Howley PM, eds. *Fields virology*, pp. 2731–68. Lippincott, Williams and Wilkins, Philadelphia. [Chapter in major authoritative virology text, accompanied by another by JI Cohen and SE Straus on the basic virology of VZV.]

Enders G, *et al.* (1994). Consequences of varicella and herpes zoster in pregnancy: prospective study of 1739 cases. *Lancet* **343**, 1547–50. [Large study from Germany and the UK assessing risk of varicella embryopathy.]

Gilden DH, *et al*. (2000). Medical progress: neurologic complications of the reactivation of varicella-zoster virus. *New England Journal of Medicine* **342,** 635–46. [A good recent review of the subject including postherpetic neuralgia.]

Pastuszak AL, *et al.* (1994). Outcome after varicella infection in the first 20 weeks of pregnancy. *New England Journal of Medicine* **330,** 901–5. [A North American study of 106 women with varicella in pregnancy.]

Wallace MR, *et al.* (1992). Treatment of adult varicella with oral acyclovir—a randomized placebo-controlled trial. *Annals of Internal Medicine* **117,** 358–83. [Describes the influence of aciclovir on the course of varicella in young adults.]

Wood MJ, *et al.* (1994). A randomised trial of acyclovir for 7 days or 21 days with and without prednisolone for treatment of acute herpes zoster. *New England Journal of Medicine* **330**, 901–5. [UK study showing that longer courses of aciclovir and prednisone do not reduce the frequency of postherpetic neuralgia.]

Human cytomegalovirus infection

Crumpacker CS (1996). Review article: drug therapy: ganciclovir. *New England Journal of Medicine* **335**, 721–9. [A comprehensive review of the use of ganciclovir in the treatment of HCMV.]

Fowler KB, *et al*. (1992). The outcome of congenital cytomegalovirus infection in relation to maternal antibody status. *New England Journal of Medicine* **326**, 663–7. [United States study of sequelae of congenital HCMV infection in relation to primary infection.]

Lowance D, *et al*. (1999). Valacyclovir for the prevention of cytomegalovirus disease after renal transplantation. *New England Journal of Medicine* **340**, 1462–70. [Trial showing efficacy of valaciclovir for HCMV prophylaxis in this group.]

Minton EJ, Sinclair JH, Sissons JGP (1995). Biological aspects of cytomegalovirus infection in marrow transplantation. In: Sullivan KM, Koppa SD, eds. *Marrow Transplantation Reviews 1991–1994*, pp 171–5. Kluge, Virginia. [Review of HCMV biology in relation to marrow transplantation.]

Ramsay ME, Miller E, Peckham CS (1991). Outcome of confirmed symptomatic congenital cytomegalovirus infection. *Archives of Diseases in Childhood* **66**, 1068–9. [United Kingdom study of congenital HCMV outcome.]

Pass RF (2001). Cytomegalovirus. In: Knipe DM, Howley PM, eds. *Fields virology*, pp. 2675–706. Lippincott, Williams and Wilkins, Philadelphia. [Chapter in major authoritative virology text, accompanied by another by ES Mocarski on the basic virology of HCMV.]

Ross R (1999). Atherosclerosis—an inflammatory disease. *New England Journal of Medicine* **340**, 115–26. [A review which includes discussion of the possible role of HCMV and other microbial pathogens in the aetiology of arterial disease.]

Whitley RJ, *et al*. (1998). Guidelines for the treatment of CMV diseases in patients with AIDS in the era of potent antiretroviral therapy. *Archives of Internal Medicine* **158**, 957–69. [Recommendations of an international panel on treatment of CMV disease in AIDS.]

Human herpesvirus-6 and -7

Hall CB, *et al*. (1994). Human herpesvirus-6 infection in children: a prospective study of complications and reactivation. *New England Journal of Medicine* **331**, 432–8. [A comprehensive study of primary HHV-6 infection in children presenting with febrile illness to a hospital emergency department.]

Knox KK, Carrigan DR (1994). Disseminated active HHV-6 infections in patients with AIDS. *Lancet* **343**, 577–8. [Provides evidence for HHV-6 reactivation in AIDS.]

Pellett PE, Dominguez G (2001). Human herpesvirus 6. In: Knipe DM, Howley PM, eds. *Fields virology*, pp. 2769–84. Lippincott, Williams and Wilkins, Philadelphia. [Chapter in major authoritative virology text, including the basic virology of HHV-6.]

Yamanishi K (2001). Human herpesvirus-6 and 7. In: Knipe DM, Howley PM, eds. *Fields virology*, pp. 2785–802. Lippincott, Williams and Wilkins, Philadelphia. [Description of the clinical disease associations of HHV-6 and 7.]

Human herpesvirus-8

Antman K, Chang Y (2000). Medical progress: Kaposis sarcoma. *New England Journal of Medicine* **342**, 1027–39. [Review of Kaposi's sarcoma and the association with HHV-8, including review of therapy of Kaposi's sarcoma.]

Cesarman E, *et al*. (1995). Kaposi's sarcoma-associated herpesvirus-like DNA sequences in AIDS-related body-cavity-based lymphomas. *New England Journal of Medicine* **332**, 1186–91. [Describes the association between HHV-8 and primary effusion lymphomas.]

Chatlynne LG, Ablashi DV (1999). Seroepidemiology of Kaposi's sarcoma-associated herpesvirus (Kaposi's sarcomaHV). *Seminars in Cancer Biology* **9**, 175. [Summarizes current knowledge of seroepidemiology of HHV-8.]

Hayward GS (1999). Kaposi's sarcomaHV strains: the origins and global spread of the virus. *Seminars in Cancer Biology* **9**, 187. [Summarizes current molecular evidence for the evolution of the

virus.]

Martin JN, *et al*. (1998). Sexual transmission and the natural history of human herpesvirus 8 infection. *New England Journal of Medicine* **338**, 948–54. [Provides evidence for sexual transmission of HHV-8 and association with Kaposi's sarcoma in homosexual men.]

Moore PS, Chang Y (1995). Detection of herpesvirus-like DNA sequences in Kaposi's sarcoma in patients with and those without HIV infection. *New England Journal of Medicine* **332**, 1181–5. [The original detection of HHV-8 in Kaposi's sarcoma.]

Cercopithecine herpesvirus-1 (herpes B virus)

Davenport DS, *et al.* (1994). Diagnosis and management of human B virus (*Herpesvirus simiae*) infections in Michigan. *Clinical Infectious Diseases* **19**, 3. [Review of clinical aspects of herpes B virus infection.]

Holmes GP, *et al.* and the B Virus Working Group (1995). Guidelines for the prevention and treatment of B virus infections in exposed persons. *Clinical Infectious Diseases* **20**, 421–39. [Current US recommendations for management of human herpes B virus infection.]

Sabin AB, Wright AM (1934). Acute ascending myelitis following a monkey bite, with the isolation of a virus capable of reproducing the disease. *Journal of Experimental Medicine* **59**, 115–36. [The original description of herpes B virus and the case of Dr WB.]

Straus SE (2000). Herpes B virus. In: Mandell GL, Bennett JE, Dolin R, eds. *Principles and Practice of Infectious Diseases*, pp. 1621–4. Churchill Livingstone, Edinburgh. [Recent chapter in major textbook of infectious disease.]

Whitley RJ, Hilliard JK (2001). Cercopithicine herpesvirus (B virus). In: Knipe DM, Howley PM, eds. *Fields virology*, pp. 2835–48. Lippincott, Williams and Wilkins, Philadelphia. [Review of herpes B virus and disease.]

7.10.3 The Epstein–Barr virus

M. A. Epstein and Dorothy H. Crawford

Background

The virus

Epstein–Barr virus (**EBV**), discovered in 1964, is one of the eight herpesviruses of man. It consists of an outer envelope, a protein capsid, and an inner double-stranded linear DNA genome.

Viral infectious cycle

Natural infection is limited to man and susceptible target cells are circulating B lymphocytes and, in certain circumstances, squamous epithelial cells of the oropharynx. Lytic infection of these cell types leads to production of viral progeny and cell death. The virus also causes a latent infection of B cells *in vivo* and can transform normal B lymphocytes *in vitro* into continuously growing, latently infected, immortalized lymphoblastoid lines. Specific sets of virus-coded proteins are expressed in each type of infection.

Virus-coded proteins

EBV-coded proteins are categorized according to the time of their appearance during the infectious cycle as latent, early, or late antigens. Most elicit cytotoxic T-cell responses and serum antibodies; both are important for controlling the infection and the latter are used in diagnosis.

General epidemiology

The virus is widespread in all human populations. Primary infection usually occurs in early childhood, at which age it is clinically silent, but leads to the generation of antibodies to the virus-determined antigens and of specific cytotoxic T lymphocytes. A lifelong carrier state ensues, in which both humoral and cellular immune responses are maintained continuously. The virus persists as a latent infection in a few circulating B lymphocytes and as a productive, lytic infection in intraepithelial B cells of the mouth and pharynx, and perhaps also the urogenital tract and salivary glands. EBV is shed into the buccal fluid in considerable amounts in about 20 per cent of those who have been infected and in small amounts in the remainder; the virus has also been detected in genital secretions. Virus in the buccal fluid provides the main source for transmission of the infection in the population; in children this occurs via droplets or when objects are casually contaminated with saliva and sucked, whereas amongst the sexually active, transmission is by salivary transfer during kissing. In developing countries, 99.9 per cent of children are infected by the second to the fourth year of life but in industrialized countries with high standards of hygiene many do not meet the virus as young children. The percentage of teenagers or young adults remaining free of infection in Western societies depends on socio-economic group—the higher the standard of living, the greater the percentage; 50 per cent of very affluent young adults may escape childhood infection (Fig. 1).

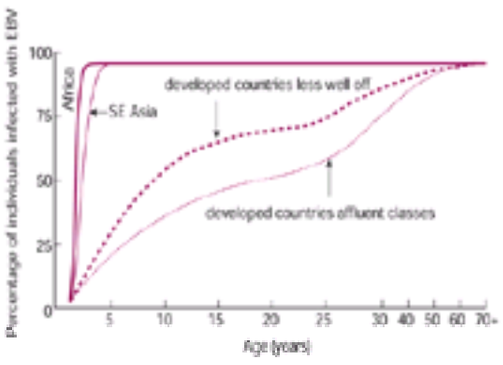

Fig. 1 Comparison of the ages at which individuals in different populations become infected with EBV. In developing countries, almost all children have acquired the virus by 2 to 4 years of age, depending on geographical region. In developed countries with high standards of living and hygiene, the time of infection is delayed for many, more markedly among the affluent than the less well off. Amongst the very rich, as many as 50 per cent may reach adolescence or young adulthood without having encountered the virus and will undergo delayed primary infection with a high risk that this will be accompanied by the symptoms of infectious mononucleosis. (Reprinted with permission from Epstein MA (2002). Infectious mononucleosis. In: *Encyclopedia of life sciences*, vol. 10, pp.211–16. http://www.els.net/, London: Nature Publishing Group.)

Infectious mononucleosis

Infectious mononucleosis occurs in about 50 per cent of those who miss EBV infection in childhood when, sooner or later, they undergo delayed primary infection; the other 50 per cent of delayed infections are symptom free. Because teenagers and young adults in the affluent classes of Western countries escape infection as children, infectious mononucleosis is a disease of upper socio-economic groups; conversely it is exceptionally rare in developing countries (Fig. 1). Although most cases occur in adolescents and young adults, children and the middle aged may sometimes develop the disease, and rarely also the elderly. Infectious mononucleosis is associated with kissing and is acquired when a healthy carrier, who is shedding virus in his/her saliva, passes this during close buccal contact directly into the oropharynx of a partner who has not been primarily infected in the usual way as a child. This explains why case-to-case infection and epidemics are not seen and why the incubation period, perhaps 30 to 50 days, is difficult to calculate. Primary EBV infection giving infectious mononucleosis-like symptoms may also be transmitted by blood transfusion or organ grafting from an infected donor to a previously uninfected recipient.

Symptoms

Classic infectious mononucleosis may follow some days of vague indisposition or may start abruptly. It presents with sore throat, fever with sweating, anorexia, headache, and fatigue, together with malaise quite out of proportion to the other complaints. Dysphagia may be noticed and also brief orbital oedema. Erythematous and maculopapular rashes occur in a small number of untreated patients, but in many more who have been taking ampicillin for sore throat before infectious mononucleosis has been diagnosed. Tonsillar and pharyngeal oedema can rarely cause pharyngeal obstruction (Fig. 2).

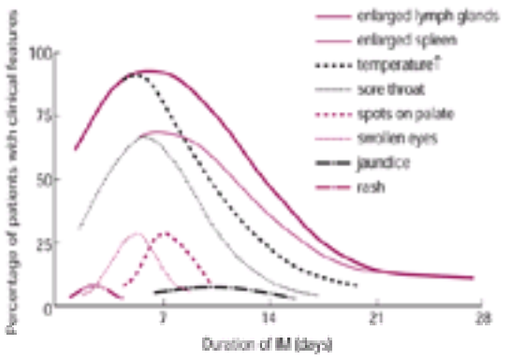

Fig. 2 Percentage of patients with infectious mononucleosis showing various clinical features during the course of the disease, and the timing and average duration of each. (Reprinted with permission from Epstein MA (2002). Infectious mononucleosis. In: *Encyclopedia of life sciences*, vol. 10, pp.211–16. http://www.els.net/, London: Nature Publishing Group.)

Signs

The fever may rise to 40°C but high levels and swings are not seen. There is redness and oedema of the pharynx, fauces, soft palate, and uvula, and about half the patients develop greyish exudates. Generalized lymphadenopathy is almost always present, most marked in the cervical region; the glands are symmetrical, discrete, and slightly tender, and are accompanied by splenomegaly in about 60 per cent of cases and an enlarged liver in 10 per cent. There is usually a moderate bradycardia. Besides the rash, characteristic palatal enanthematous crops of reddish petechiae are found in about one-third of patients, and jaundice occurs in about 8 per cent (Fig. 2).

Clinical course

Mild cases may resolve in days, but 1 to 2 weeks is more usual, followed by a period of lethargy. The duration of this convalescence is influenced by psychological factors, particularly the speed with which patients are encouraged to resume full activity. About 1 case in 2000 may continue in a truly chronic or recurrent form for several months or years, and here exhaustive investigations have shown immunological defects. Most other cases of so-called chronic infectious mononucleosis are in reality manifestations of 'chronic fatigue syndrome', but it is highly controversial as to whether this is a true entity rather than a form of depression or a belief disorder; credible connections with EBV have not been established. In contrast, there is an extremely rare, genetically determined, X-linked, lymphoproliferative condition (XLP disease, or Duncan syndrome after one of the first families to be recognized) in which the affected young males of certain kindred die from infectious mononucleosis owing to a specific inability to respond normally to EBV during primary infection; the disease progresses inexorably, with necrotic destruction of vital organs and multisystem failure. There is evidence that an aberrant immune response to EBV in XLP results in unregulated cytotoxic T-cell activity being directed against the normal cells of vital organs instead of targeting solely on EBV-infected cells displaying EBV antigens. The gene responsible for this defect has recently been cloned.

Complications

Minor non-specific complications may occur; rare more serious complications include secondary bacterial throat infections, traumatic rupture of the enlarged spleen, asphyxia from pharyngeal oedema, massive hepatic necrosis, Guillain–Barré syndrome, and autoimmune manifestations such as thrombocytopenia and haemolytic anaemia.

Differential diagnosis

Classic infectious mononucleosis is diagnosed on the basis of the clinical picture considered in conjunction with serological and haematological laboratory investigations (see below). An infectious mononucleosis-like disease can occur in primary cytomegalovirus infection and in toxoplasmosis, but in both conditions the sore throat is much less severe and with cytomegalovirus the lymphadenopathy may be minimal or absent.

Laboratory diagnosis

A rapid screening test (Monospot test) can be used to detect the presence of heterophil antibodies in the patient's serum. Although these heterophil antibodies are not directed against viral-coded proteins they are present in up to 85 per cent of acute infectious mononucleosis sera. Cases of Monospot-negative infectious mononucleosis tend to be those outside the classic 15- to 25-year age range, and false-positive tests may occur in pregnancy and autoimmune disease. However, the presence of serum IgM antibodies to EB virus capsid antigen (**VCA**) are diagnostic of infectious mononucleosis. An additional important feature of infectious mononucleosis is the presence of lymphocytosis of up to $15 \times 10^{9}/1$, with the majority of cells having an 'atypical' morphology.

Treatment

Bed rest and aspirin for headache and pharyngeal discomfort are the only treatments required. When the fever resolves the patient should be encouraged to get up and resume some activities as fast as is practicable, but violent exercise should be avoided for 3 weeks after an enlarged spleen ceases to be palpable. Only complications need active therapy: splenic rupture requires surgery, bacterial infections call for appropriate antibiotics, airway obstruction must be relieved by tracheostomy, and corticosteroids should be given for life-threatening pharyngeal oedema and for neurological and haematological complications.

Pathogenesis

Why children do not have symptoms during primary EBV infection whereas adolescents and young adults frequently react by developing infectious mononucleosis is not fully understood. The immunological reactions of young adults on first encountering EBV are more exuberant than those of children reflecting physiological differences in responsiveness. The mode of infection and consequent size of infecting dose also play an important part; children come into contact with small amounts of virus in saliva from a shedder in droplets or casually contaminating some sucked object, whereas a young person may take in large amounts of virus-containing saliva from a carrier during kissing so that virions reach susceptible cells in large numbers. Intraepithelial B cells become productively infected in infectious mononucleosis and at this stage EBV can easily be found in patients' saliva. The newly replicated virus infects masses of B lymphocytes which are released from the throat into the blood and lymphatics from where they accumulate in lymphoid tissue throughout the body. In response, there is a generation of even greater numbers of cytotoxic T cells specifically directed against EBV-determined antigens displayed by the infected B lymphocytes. An exaggerated immunological response thus underlies the changes seen in infectious mononucleosis since all these lymphoid cells present in the circulation, in lymph nodes, in tonsils, in lymphoid centres in the mouth and pharynx, and in spleen and liver are responsible for causing the sore throat, fever, malaise, lymphadenopathy, and hepatosplenomegaly by immunopathological mechanisms.

Endemic (or 'African') Burkitt lymphoma

The classic form of this B-cell tumour, first described by Burkitt in 1958, is found in certain parts of Africa and Papua New Guinea where the temperature does not fall below 16°C or the annual rainfall below 55 cm. Endemic Burkitt lymphoma is distinct from the 'Burkitt-like' tumours that occur sporadically everywhere in the world (sometimes called 'American' Burkitt lymphoma) and that have a different age incidence, anatomical distribution, and response to therapy, and arise from B cells with different phenotypic characteristics.

The association between EBV and endemic Burkitt lymphoma is so close that it is generally accepted that the virus is an essential link along with cofactors in a complicated chain of events which leads to the malignancy. Hyperendemic malaria has been identified as the important cofactor, and its spread by anopheline mosquitoes requiring warmth and moisture explains the climate dependence of Burkitt lymphoma.

Burkitt lymphoma is a disease of childhood, is extremely rare over the age of 14 years, and in the endemic areas it is more common than all other childhood tumours added together.

Symptoms

The tumour is usually multifocal and the symptoms depend entirely on the anatomical location. Jaw tumours are present in 70 per cent of patients, are the usual presenting feature, may be multiple in up to all four quadrants, and are almost always accompanied by tumours elsewhere. They give a rapidly growing mass with loosening of teeth and exophthalmos from orbital spread. Abdominal tumours involve retroperitoneal nodes, liver, ovaries, intestines, and kidneys. Burkitt lymphoma sometimes presents in thyroid, the adolescent female breast, testicles, and salivary glands; extradural tumours in the spine cause rapid paraplegia, and skeletal tumours also occur. Characteristically Burkitt lymphoma does not involve the spleen or peripheral lymph nodes.

Signs

The tumours are firm, very rapidly growing, painless, and cause minimal constitutional disturbance. Their sites determine the clinical signs.

Clinical course

Tumour growth is relentless and death ensues within a few months in the absence of treatment.

Differential diagnosis

In endemic areas Burkitt lymphoma can be diagnosed from the clinical picture. Unlike Burkitt lymphoma, retinoblastoma is intraocular, rhabdomyosarcoma is extraorbital and does not involve teeth, nephroblastoma is not multifocal, and neuroblastoma and ovarian tumours can be distinguished histologically. Paraplegia of tuberculous origin causes vertebral collapse and acute transverse myelitis is preceded by pain and fever. Other lymphomas have a markedly dissimilar anatomical distribution.

Laboratory diagnosis

Histological examination of a biopsy sample gives ready confirmation. Antibodies to EBV antigens show a unique pattern and titres rise or fall with disease progression or response to therapy. IgG anti-VCA titres are around 10 times higher than in controls and antibodies to EBV-restricted early antigens (EA-R) and membrane antigens (MA) are also detectable.

Treatment

Surgery and radiotherapy are ineffective, but moderate courses of chemotherapy give excellent results with cyclophosphamide being the drug of choice.

Pathogenesis

The molecular pathways whereby EBV-coded antigens transform normal B lymphocytes *in vitro* into continuously growing immortalized cell lines are partially understood. There are now credible explanations as to how EBV combines with cofactors such as hyperendemic malaria to cause the tumour. The virus is a necessary, but not on its own sufficient, element in the aetiology of the disease.

Lymphoproliferations in immunosuppressed states

In primary and secondary suppression of cellular immunity there is diminished immune control of persisting EBV infection which leads to increased virus replication in the oral cavity, increased numbers of circulating, virus-carrying B lymphocytes, and increased levels of serum antibodies to EBV antigens. This is sometimes described as a 'reactivated infection' although the condition is clinically silent. However, on occasions the loss of control leads to the development of EBV-associated lymphoproliferations.

In transplant recipients

Transplant recipients who receive lifelong immunosuppressive drugs to prevent graft rejection have a 28 to 100 times increased risk of developing lymphoproliferative disease and lymphoma compared with normal controls; most of these conditions are of B-cell origin, contain the EBV genome, and express viral antigens in their cells. Lymphoproliferative disease has two forms. About 50 per cent of cases are associated with primary EBV infection in patients who were seronegative at the time of grafting, occur within the first year after transplantation in a young age group, and have infectious mononucleosis-like symptoms. The remainder of the cases occur in older patients late after transplantation as a localized mass, commonly in the gut, central nervous system, or transplanted organ; biopsy shows large-cell lymphoma, which is usually monoclonal. Reduction of immunosuppressive therapy, with or without acyclovir therapy, is the first line of treatment, with cytotoxic drugs and/or radiotherapy being used only where there is no response or after recurrence. Recently, experimental treatments with EBV-specific cytotoxic T-cell infusions have shown encouraging results.

In acquired immunodeficiency syndrome (AIDS)

Two types of lymphoma are seen in patients with AIDS; large-cell lymphoma and Burkitt lymphoma, and both may be associated with EBV.

Large-cell lymphomas similar to those found in transplant recipients (see above) occur in severely immunocompromised patients with AIDS; their distribution is

extranodal, involving many unusual sites, most commonly the central nervous system. These lymphomas show a strong association with EBV which reaches 100 per cent in cerebral tumours; the progress is rapid, with a mean survival time from diagnosis of 3 to 4 months. Treatment (radiotherapy) is disappointing because patients with terminal AIDS are in such poor general health.

Burkitt lymphoma occurs earlier in the course of human immunodeficiency virus (HIV) disease while the immune system is still relatively intact and is therefore more amenable to treatment. About 50 per cent of these lymphomas contain EBV DNA.

Hodgkin's disease and T-cell lymphomas

There has long been a suspicion that EBV is involved in the induction of Hodgkin's disease because of the similar socio-economic epidemiology of Hodgkin's disease and infectious mononucleosis, and because within 5 years of infectious mononucleosis there is a four- to sixfold increase in the likelihood of developing Hodgkin's disease. There is now evidence that in Hodgkin's lymphomas EBV DNA is carried and expressed in both the Reed–Sternberg and the mononuclear Hodgkin's cells. These findings are as yet insufficient to implicate EBV in the aetiology of Hodgkin's disease, but point to the need for further investigation. A similar situation exists with oral T-cell lymphoma in patients with AIDS, and nasal T-cell lymphomas.

Nasopharyngeal carcinoma

This tumour is restricted to the postnasal space where it arises from squamous epithelial cells. The tumour is seen rarely throughout the world but has a remarkably high incidence in southern Chinese, and in the Inuit and related circum-Arctic races. In high incidence areas, nasopharyngeal carcinoma is the most common cancer of men and the second most common of women. A rather high incidence of nasopharyngeal carcinoma is seen amongst Malays, Dyaks, Indonesians, Filipinos, and Vietnamese people, and a medium-high incidence belt stretches across North Africa, through the Sudan, to the Kenya highlands. The tumour usually occurs in middle or old age, but in North Africa it has bimodal age peaks, one involving young people up to 20 years of age and a second much later in life. Irrespective of geographical region, nasopharyngeal carcinoma cells always carry the EBV genome.

Symptoms

Nasopharyngeal carcinoma causes nasal obstruction, discharge, or bleeding; deafness, tinnitus, or earache; headache; and ocular paresis from tumour spread to involve cranial nerves. Patients may present with a single symptom caused locally by the tumour or with several symptoms, and about one-third complain only of cervical lymph-node enlargement due to metastatic spread from an occult primary tumour.

Signs

Direct spread from the primary tumour may involve the soft tissues, bone, parotid gland, buccal cavity, and oropharynx. The neoplasm may extend into the nasal fossas, the paranasal sinuses, or the orbit, and can invade the eustachian canal or the parapharyngeal space where cranial nerves IX, X, XI, and XII can be involved. Invasion of the skull or cranial foramina may damage cranial nerves II, IV, V, and VI. Lymphatic spread causes enlarged cervical lymph nodes and subsequently extends to the supraclavicular glands. If bloodborne metastases occur, they are most frequent in the bones, liver, and lungs, but may be in any organ.

Clinical course

Untreated nasopharyngeal carcinoma progresses inexorably to death.

Differential diagnosis

Nasopharyngeal carcinoma must be distinguished from other tumours of the nasal cavities, namely adenocarcinomas, sarcomas, malignant lymphomas, and rare malignancies such as chordoma, teratoma, and melanoma.

Laboratory diagnosis

The diagnosis of nasopharyngeal carcinoma is made histologically on a biopsy sample either from the primary tumour or from an enlarged cervical lymph node. In addition, serum antibody titres to EBV antigens show a characteristic reaction pattern—IgG and IgA antibodies to VCA and diffuse early antigen (EA-D) are raised, with the titre correlating with the tumour burden. Uniquely, IgA antibodies to VCA and EA are also found in the saliva from patients.

Treatment

Nasopharyngeal carcinoma responds well to radiotherapy, which is the treatment of choice. In the earliest stages of the disease, radiotherapy gives 5-year survival rates of 50 per cent or more, and of those surviving 5 years, 70 per cent remain permanently free of relapse. More advanced stages of nasopharyngeal carcinoma have correspondingly worse prognoses.

Pathogenesis

EBV is now widely accepted as necessary for the causation of nasopharyngeal carcinoma, but is not sufficient on its own. Besides the racial predisposition there is also a genetic predisposition since southern Chinese people with an A2BW36 haplotype are four to six times more likely to have nasopharyngeal carcinoma than those without. Epidemiological studies suggest that important environmental cofactors associated with the Chinese way of life play a role and two likely candidates are: traditional herbal medicines, taken as snuff, and containing tumour-promoting substances of phorbol ester type; and traditional salt fish which has been shown to contain carcinogenic nitrosamines.

Hairy leukoplakia in AIDS

This lesion occurs in people with HIV and in other immunosuppressed individuals; it usually presents as painless white patches on the tongue or on the lateral buccal mucosa. The lesions are slightly raised, poorly demarcated, and have a 'hairy' or corrugated surface; the patches are usually multiple and measure up to 3 cm in diameter.

The squamous epithelial cells of this condition contain large amounts of actively replicating EBV, providing an unusual example of the production of the virus by such cells. Treatment with acyclovir arrests the EBV replication and the lesions regress, but only for as long as the drug is continued.

Smooth muscle tumours and gastric carcinoma

Recently EBV has been implicated in various types of gastric carcinoma and in leiomyomas and leiomyosarcomas in children immunosuppressed by AIDS or after organ transplantation. Much further study of these relationships is required.

Further reading

Bar RS *et al.* (1974). Fatal infectious mononucleosis in a family. *New England Journal of Medicine* **290**, 363–7. [The first account of an XLP (Duncan) syndrome family.]

Burkitt D (1958). A sarcoma involving the jaws of African children. *British Journal of Surgery* **46**, 218–3. [The first description of Burkitt lymphoma.]

Burkitt D (1963). A lymphoma syndrome in tropical Africa. *International Review of Experimental Pathology* **2**, 67–138. [An early comprehensive review of Burkitt lymphoma.]

de Thé *et al.* (1978). Epidemiological evidence for a causal relationship between Epstein-Barr virus and Burkitt's lymphoma: results of the prospective Ugandan study. *Nature* **274**, 756–61. [A massive

investigation linking EBV to the causation of Burkitt lymphoma.]

Epstein A (1999). On the discovery of Epstein-Barr virus: a memoir. *Epstein-Barr Virus Report* **6**, 58–63. [Details of how EBV was discovered.]

Epstein MA, Achong BG, eds (1979). *The Epstein–Barr virus*. Springer Verlag, Berlin. [A complete survey of the first 15 years of EBV research.]

Greenspan JS *et al.* (1985). Replication of Epstein-Barr virus within the epithelial cells of oral hairy leukoplakia, an AIDS-associated lesion. *New England Journal of Medicine* **313**, 1564–71. [The first description of the condition.]

Henle G, Henle W, Diehl V (1968). The relation of Burkitt's lymphoma tumor-associated herpesvirus to infectious mononucleosis. *Proceedings of the National Academy of Sciences (USA)* **59**, 94–101. [The account of the original findings identifying EBV as the cause of infectious mononucleosis.]

Herbst H, Niedobitek G (1994). Epstein–Barr virus in Hodgkin's Disease. *Epstein–Barr Virus Report* **1**, 31–5. [A very useful review.]

Hoagland RK (1955). Transmission of infectious mononucleosis. *American Journal of Medical Science* **229**, 262–72. [The first recognition of infectious mononucleosis as the 'kissing disease'.]

Rickinson AB, Kieff E (2001). Epstein–Barr virus. In: Fields BN *et al*. eds. *Fields virology*, 4th edn, Vol 2, pp 2575–627. Lippincott, Williams and Wilkins, Philadelphia. [A comprehensive review of recent work on EBV.]

Rickinson AB *et al*. (2001). T-cell recognition of Epstein–Barr virus associated lymphomas. *Cancer Surveys* **13**, 53–80. [An excellent survey.]

Schlossberg D, ed. (1989). *Infectious mononucleosis*, 2nd edn. Springer Verlag, Berlin. [A multiauthor work covering many aspects of the disease.]

Shanmugaratnam K (1971). Studies on the etiology of nasopharyngeal carcinoma. *International Review of Experimental Pathology* **10**, 361–413. [An excellent review.]

Sprunt TP, Evans FA (1920). Mononuclear leucocytosis in reaction to acute infections ('infectious mononucleosis'). *Bulletin of the Johns Hopkins Hospital* **31**, 410–17. [The first description of infectious mononucleosis.]

Thomas JA, Allday MJ, Crawford DH (1991). Epstein–Barr virus-associated lymphoproliferative disorders in immunocompromised individuals. *Advances in Cancer Research* **57**, 329–80. [A good review.]

7.10.4 Poxviruses

Geoffrey L. Smith

Poxviruses are large DNA viruses that replicate in the cell cytoplasm. The most infamous was variola virus, which caused smallpox, a disease responsible for devastating epidemics with up to 40 per cent mortality. Smallpox was eradicated (1977) by immunoprophylaxis with vaccinia virus, a related orthopoxvirus. Poxvirus infections in humans have since been restricted to molluscum contagiosum and rare zoonoses such as monkeypox, cowpox, orf, pseudocowpox, yaba tumour virus, and tanapox.

Classification

The chordopoxviruse subfamily is subdivided into eight genera, of which the orthopoxviruses have been the most important (Table 1). Viruses within different genera are antigenically distinct, while those within a genus are cross-reactive and cross-protective. Four of the nine poxviruses that infect humans are orthopoxviruses: cowpox, variola, monkeypox, and vaccinia virus. Different orthopoxviruses are distinguishable by their biological properties such as pock type, ceiling temperature on the chorioallantoic membrane, or by the restriction pattern of genomic DNA. Following the sequencing of several virus genomes, species-specific DNA probes are becoming available. Vaccinia virus has no known natural animal reservoir and its origin remains a mystery. It caused human disease only as a rare complication after vaccination against smallpox. Cowpox and monkeypox viruses were named after the species from which they were isolated but the natural reservoir of each virus may be rodents. Infections in cows or monkeys, like the occasional transmission to man, are rare. Cowpox, monkeypox, and vaccinia virus have a broad host range, while variola virus infected only humans and the lack of an animal reservoir aided the smallpox eradication campaign.

Poxvirus biology

Poxviruses replicate in the cytoplasm, encode enzymes for transcription and DNA replication, and have large, complex virions (Fig. 1) and double stranded DNA genomes of 150 to 300 kb. Vaccinia virus is the most intensively studied poxvirus. It encodes about 200 genes of three classes (early, intermediate, and late) that are expressed in a strictly regulated manner. Transcription of each class is dependent upon the prior expression of the previous class.

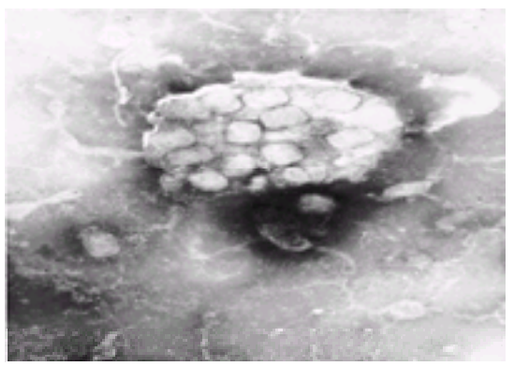

Fig. 1 Electron micrograph of material from smallpox lesion, viewed by negative contrast, showing a clump of poxvirus particles. (By courtesy of the late Henry Bedson.)

Virus morphogenesis is complex (Fig. 2(a)) and produces two forms of infectious virion: intracellular mature virus (IMV) and extracellular enveloped virus (EEV). IMV remains within the cell until it is lysed and forms most of the progeny, whereas EEV is released from the cell before death and represents less than 1 per cent of infectivity (Fig. 2(b)). EEV possesses an additional lipid envelope with which several virus and cellular proteins are associated, giving it distinct immunological and biological properties. EEV is necessary for virus dissemination *in vitro* and within the infected host. Immunity to EEV-specific antigens, which are highly conserved among orthopoxviruses, is required for protection against disease.

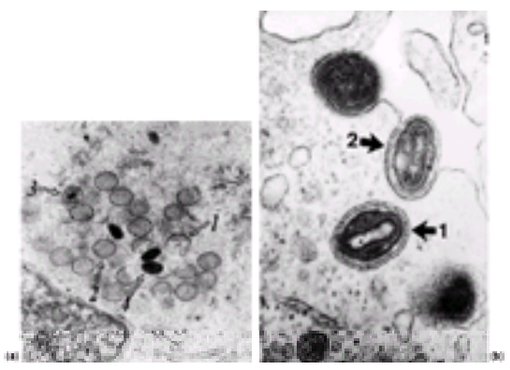

Fig. 2 Electron micrographs showing (a) a cytoplasmic vaccinia virus factory containing maturing virus particles with stages of morphogenesis numbered 1 to 4 and (b) fully enveloped virus particles, one of which is leaving the cell.

Pathogenesis

Poxvirus infections cause a local skin lesion or generalized pustular rash. Detailed experimental analysis of human smallpox was impossible, but generalized poxvirus infections have been studied in experimental models, namely monkeypox in monkeys, rabbitpox (a neurovirulent vaccinia virus) in rabbits, and ectromelia virus in mice. The spread of variola virus in man was probably similar to that of ectromelia virus in mice and is characterized by sequential phases of virus infection, replication, and release accompanied by cell necrosis.

Virus enters through skin abrasions (ectromelia and cowpox) or inhalation of airborne virus and establishes a respiratory infection (ectromelia, rabbitpox, and variola). In smallpox, the respiratory route was most important and sometimes the only possible route of transmission from index cases to contacts; also patients became infectious only after enanthem developed. A respiratory infection was established in the epithelial cells of the alveoli and small bronchioles. Here, alveolar

macrophages became infected and transmitted the virus via lymphatics to the local lymph node, where further virus replication occurred. Virus released into the blood (primary viraemia) was mostly cell-associated and spread to other organs of the reticuloendothelial system, notably the liver, spleen, and lymph nodes.

Extensive replication here released larger amounts of virus into the blood (secondary viraemia) enabling the virus to infect other organs such as the kidneys, lungs, and intestines and to reach the skin and produce the skin lesions with the characteristic centrifugal distribution (Fig. 3) (Plate 1). Lesions started with a papule that became pustular and then crusted. After 2 to 3 weeks the scab was shed leaving a scar. The incubation period of smallpox was approximately 12 days. Symptoms included headache, fever, malaise, vomiting, and, in severe cases, prostration, toxaemia, and hypotension. Delayed onset of the skin eruptions usually correlated with a grave prognosis. Haemorrhagic or flat confluent-type smallpox had very high mortality rates.

Fig. 3 Smallpox in a 9-month-old boy in Pakistan, photographed on the eighth day of the rash. (By courtesy of the World Health Organization.)

The outcome of infection depended upon the age and physiological and immunological status of the patient and the strain of virus. Variola major was more virulent and produced fatality rates in unvaccinated patients of between 5 and 40 per cent, while the milder variola minor, called alastrim in the Americas, caused only 0.1 to 2 per cent mortality. Morphologically, the viruses were indistinguishable, and vaccination with vaccinia virus was equally effective against both. However, alastrim virus was consistently more thermolabile and had a lower ceiling temperature of 37.5°C compared to 38.5°C for variola major, 39°C for monkeypox, 40°C for cowpox, and 41°C for vaccinia virus.

Very young and elderly patients were most susceptible to smallpox and those aged 5 to 20 years most resistant. Pregnancy and immunological deficiency, particularly in cell-mediated immunity, increased the severity of infection. Pregnant women were more likely than any other group to develop haemorrhagic-type smallpox, which was usually fatal. The greater importance of cell-mediated immunity rather than antibody in recovery from poxvirus infections was illustrated in several ways. Firstly, in children with severe defects in cell-mediated immunity there was a progressive and uncontrolled virus replication from the vaccination site that was usually fatal. In contrast, defects in antibody production were usually tolerated if the cell-mediated immune response was normal. Secondly, passive administration of antivaccinia virus serum had little effect on mice infected with ectromelia virus, whereas prior infection with vaccinia virus was protective. Thirdly, in mice infected with ectromelia virus, the effective mechanisms that combated infection in the liver and spleen were operative by 4 to 6 days postinfection and coincided with the maximum levels of cytolytic T cells, but preceded the development of systemic antibody.

The eradication of smallpox

Early attempts to control smallpox relied upon variolation or inoculation, in which material isolated from a mild case of smallpox was administered by sniffing or scratching. This was replaced by vaccination in 1798 after Jenner noticed that milkmaids, who often acquired cowpox infections on their hands from the teats of cows, were protected from smallpox. Jenner infected a boy (James Phipps) with poxvirus material (probably cowpox), derived from a cow via a milkmaid (Sarah Nelmes), and subsequently challenged him with smallpox. Protection was achieved and due to the efficacy and greater safety of this procedure it rapidly replaced variolation. Sometime between 1798 and the twentieth century vaccinia virus replaced cowpox as the smallpox vaccine. In 1959, the World Health Organization (WHO) adopted a recommendation to achieve the global eradication of smallpox. With fresh funding and a plentiful supply of potent freeze-dried vaccine this goal was achieved in 1977. Two years later, the WHO certified that eradication was complete. This triumph of preventive medicine justifies the saying 'prevention is better than cure' but also demonstrates that prevention is best achieved by eradication.

Sequencing of poxvirus genomes

The genomes of vaccinia virus strains Copenhagen, Western Reserve and modified virus Ankara (MVA), variola virus strains India-1967 and Bangladesh-1975, and camelpox virus have been determined. In addition, regions of variola virus Harvey-1947, Garcia-1966, Congo-1970, and Somalia-1977, rabbitpox, and cowpox virus GRI-90 have been sequenced. These analyses showed that the central region of these orthopoxvirus genomes are very closely related with greater than 96 per cent nucleotide identity between vaccinia and variola viruses, but that there is significant divergence in the terminal regions. A notable difference between the vaccinia and variola genomes is the fragmentation of several genes in variola that are intact in vaccinia virus. It is possible that the disruption of genes of an ancestral poxvirus may have contributed to the evolution of variola major as a highly pathogenic virus for man. The retention of these non-functional genes in the variola virus genome suggests that they became non-functional in the relatively recent evolutionary past, and perhaps that variola virus is a 'recent' human pathogen that never became fully adapted to man.

Other poxvirus genomes that have been sequenced are molluscum contagiosum virus, Shope fibroma virus and myxoma virus, yaba-like disease virus, and the *Melanoplus sanguinipes* entomopoxvirus.

Poxvirus expression vectors

Vaccinia virus recombinants expressing foreign genes were developed in 1982 and have become a widely used laboratory expression system and have potential as live vaccines for infectious disease and cancer. Infection with a recombinant vaccinia virus allows expression and simultaneous delivery of the foreign antigen to the immune system. Moreover, the large capacity of vaccinia virus allows expression of multiple foreign genes from a single virus so creating polyvalent vaccines. Safer vaccinia virus strains that do not cause vaccination complications (eczema vaccinatum, generalized vaccinia, progressive vaccinia, encephalopathy (<2 years), or encephalitis (>2 years)) are being developed by deletion of virulence genes from conventional vaccinia virus strains. An alternative strategy is to use poxviruses that establish only abortive infections in human cells, such as MVA or the avipoxviruses fowlpox virus and canarypox.

Human monkeypox

Monkeypox was discovered in captive primates in 1958, but in 1970 was isolated in tropical rain forests of West and Central Africa from humans who had suffered generalized poxvirus rashes visibly very similar to smallpox. The virus is distinct from variola in pock morphology, ceiling temperature, genomic restriction endonuclease pattern, lesion morphology on rabbit skin, and its ability to be passaged indefinitely in mouse brain. However, although monkeypox produced a very similar disease to smallpox in man, person-to-person transmission was too inefficient for establishing epidemics. Thus human monkeypox infections are single or multiple sporadic cases restricted to dense tropical rain forests in Central and West Africa. Clinically, human monkeypox closely resembles ordinary, discrete-type smallpox except that there is a pronounced lymph node enlargement (Fig. 4). Mortality rates in unvaccinated patients between 1970 and 1986 were 11.2 per cent but these were all in children less than 8 years old and the highest rate (18.7 per cent) was in infants less than 2 years. The virus is probably acquired from infected monkeys or rodents such as squirrels.

Fig. 4 Moderately severe monkeypox in a girl of 7 years from Equateur Province, Zaire. (By courtesy of the World Health Organization.)

Cowpox and pseudocowpox

Cowpox virus has a broad host range including cattle, humans, large felines, and even elephants, but it is not enzootic in cattle and its natural hosts are rodents. It is distinguishable from vaccinia virus by the pock type, ceiling temperature, rate of replication in tissue culture, genome size and restriction pattern, and the production of cytoplasmic A-type inclusion bodies. Pseudocowpox is enzootic in cattle, unlike cowpox. Historically, it was important since it was sometimes used mistakenly for vaccination and, being a parapoxvirus, was ineffective in preventing smallpox. Its misuse compromised Jenner's correct assertion that true cowpox was an effective smallpox vaccine.

In man, cowpox produces an acutely inflamed, local lesion, similar to a primary smallpox vaccination. There is usually fever, enlargement of the local lymph nodes, and pain. Unlike vaccinia virus, which occasionally produced a generalized infection, cowpox lesions are always local. Human lesions caused by pseudocowpox virus (milker's nodules) are extremely rare and are less painful than those caused by cowpox.

Tanapox and yaba tumour virus

Tanapox virus was first isolated from the Tana valley in Kenya from humans suffering from localized skin lesions typical of poxviruses (Plate 2). Subsequently, a similar virus was found in humans in Zaire during surveillance for monkeypox. It is a rare zoonosis of monkeys, that in Kenya may have been transmitted by mosquitoes. Serologically, it is related to yaba-like disease virus and yaba tumour virus. In man it usually produces a solitary lesion that is preceded for a few days by a mild fever. The lesion takes 5 to 6 weeks to clear and is distinguished from other poxvirus lesions by its failure to become pustular. This virus cannot be cultured on the chorioallantoic membrane.

Yaba tumour virus is a monkey virus that can cause histiocytomas if injected subcutaneously or intradermally into man. The lesions are not neoplastic and are cleared by the immune response.

Cutaneous poxviruses (orf and molluscum contagiosum)

See Chapter 7.10.25 and Chapter 7.10.26.

Further reading

Binns MM (1992). *Recombinant poxviruses*. CRC Press, Boca Raton, Florida.

Fenner F, Wittek R, Dumbell KR (1989). *The orthopoxviruses*. Academic Press, London.

Moss B (1996). Poxviridae: the viruses and their replication. In: Fields BN, Knipe DM, Howley PM, Chanock RM, Melnick J, Monath TP, Roizman B, Straus SE, eds. *Virology*, 3rd edn, pp. 2637–71. Lippincott-Raven, Philadelphia.

7.10.5 Mumps: epidemic parotitis

B. K. Rima

Aetiology

Mumps is an acute, generalized, communicable infection of children and young adults, caused by a paramyxovirus. Almost any organ can be infected—salivary glands, pancreas, testis, ovary, brain, mammary gland, liver, kidney, joints, and heart. Swelling of the face is only one of the symptoms of the disease, albeit the most common and important one for diagnosis.

Epidemiology

The incubation period lies between 14 and 18 days. In any outbreak, 30 to 40 per cent of those exposed infected have subclinical illness. Mumps is highly infectious. Transmission depends on close personal contact with a patient who is excreting virus in the saliva and spreading it in droplets. In the prevaccine era, the peak incidence was in the late winter or early spring, in 3- to 7-year cycles. Most morbidity is associated with meningitis and orchitis. Case fatality is about 2 per 1000.

Virology

Mumps virus (**MuV**) can be grown in tissue cultures of chick embryo, monkey kidney, human amnion, or HeLa cells. Cytopathic changes may be seen as early as 24 h postinfection. With immunofluorescence the virus can be detected in a matter of hours. The virus can also be cultured in the yolk sac or embryonic cavity of chick embryos.

MuV is thermolabile. It can be stored for years at –70°C but infectivity is lost in a few days at room temperature. Treatment with ether or paraformaldehyde inactivates the virus rapidly, but neither of these processes destroys the antigens responsible for the complement fixation, haemagglutination, or reactivity in the skin test.

A patient excretes culturable MuV in the saliva for between 2 and 6 days before parotitis develops and for up to 4 days afterwards. Virus can be cultured from the urine for up to 14 days around the onset of disease. It is almost as easily cultured from ultracentrifugates of urine as from saliva. There is, however, no evidence of viral spread of the virus by urine. During the acute disease MuV can also be cultured from throat washings or a swab of the orifice of Stensen's duct and be detected by reverse-transcriptase polymerase chain reaction (**RT-PCR**), in saliva, throat swabs, and urine. In the blood it can be cultured only for a day or two around the start of disease. Virus can be isolated from cerebrospinal fluid for the first 3 or 4 days of the meningeal illness.

MuV is an enveloped RNA virus with a genome of 15 384 nucleotides. Its inner core is a ribonucleoprotein complex (the nucleocapsid) containing the non-segmented, negative-strand, RNA molecule encapsidated by the major nucleocapsid protein (N). The nucleocapsid has the herring bone structure characteristic of paramyxoviruses (Fig. 1(a)). Attached to this are two further proteins involved in transcription and replication of the RNA genome: the phosphoprotein (P) and the large replicase protein (L). The nucleocapsid is surrounded by a lipid bilayer membrane derived from the host cell (Fig. 1(a) Fig. 1(b)). On the inner leaflet there is a membrane or matrix protein (M) that plays an essential role in virus budding. On the outer surface are two glycoproteins, one carrying the haemagglutinin-neuraminidase activity (HN), the other responsible for fusion activity (F). A non-structural, small, hydrophobic protein (SH) has been described in the membrane of MuV-infected cells. The sequence of the SH protein is hypervariable. This has been exploited in molecular studies of the epidemiology of MuV. The functions of V, and other non-structural protein, and SH are unknown. The gene order (Fig. 1(c)) leads to an expression gradient in which the abundance of mRNAs decreases with increasing distance to the promoter at the 3_ end of the genome so that the N mRNA is more abundant than the L mRNA.

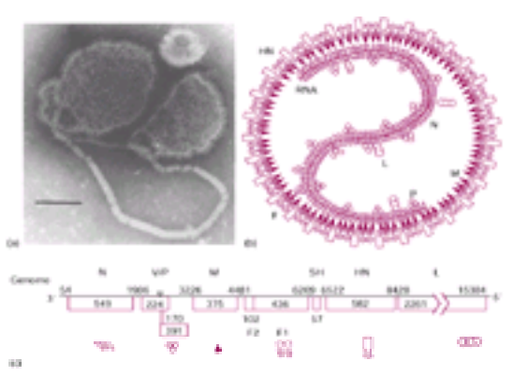

Fig. 1 Structure and genome organization of mumps virus. (a) A disrupted, negatively stained, mumps virion. The viral nucleocapsid protrudes from the particle and the fringe of viral spikes is visible (bar = 100 nm). (b) Diagram of the localization of the nucleocapsid (N), phospho- (P), large (L), matrix (M), haemagglutinin-neuraminidase (HN), and fusion (F) proteins in the mumps virion. (c) Structure of the genome of mumps virus indicating the localization of the genes, the nucleotide number of their starting and stopping position, and (in boxes) the number of amino acid residues in each of the viral proteins.

Pathogenesis and pathology

MuV causes an infection of the upper respiratory tract that spreads to draining lymph nodes. The subsequent viraemia and infection of the lymphocytes and macrophages causes spread to many organs, but because mumps is so rarely lethal, details are scant. Lymphocytic infiltration and destruction of periductal cells lead to blockage of the ducts both in salivary glands and in the seminiferous tubules of the testes. The lymphatics in the tissues surrounding and overlying the parotid glands become obstructed, producing a gel-like oedema that may spread down over the chest wall, especially when the swelling of the salivary glands is severe.

MuV frequently invades the nervous system: changes can be detected by electroencephalography or by examination of the cerebrospinal fluid in at least half the patients. However, in most of these cases there are no neurological symptoms or signs. Mumps virus is one of the most common known causes of lymphocytic meningitis. Neuronal damage probably does occur, explaining the occurrence of quadriplegia or single-nerve paralysis in some patients. Apart from transient weakness of the facial nerve, which may be due to pressure of a swollen gland or damage by mumps virus, these complications are very rare.

Mumps encephalitis is a different entity; cerebrospinal fluid is normal and contains no virus. At autopsy there is perivascular demyelination as in other forms of

postinfectious encephalitis (see Chapter xxxx).

Clinical features and diagnosis

Parotitis

A patient with mumps parotitis may have a fever without rigors (40 to 40.5°C) as well as pain near the angle of the jaw. The face and neck become distorted with swelling. The skin over the gland is hot and flushed but there is no rash, unlike in the swelling of erysipelas. If the swelling is severe, the mouth cannot be opened for pain and tightness, and is dry because the flow of saliva is blocked. This discomfort lasts for 3 or 4 days. Sometimes as one side clears, the parotid on the other side swells. When there is bilateral parotitis, clinical diagnosis is usually obvious. One condition that must be excluded is bull-neck diphtheria (see Chapter 7.11.1), which can look very like mumps.

Rarely, the submaxillary and sublingual salivary glands may also be affected. The symptoms are similar to those in parotitic mumps, but it is difficult or impossible to distinguish the swelling from other forms of submaxillary swellings, especially inflammation of various groups of lymph nodes and Ludwig's angina. In mumps, the neck swelling is ill defined and the angle of the jaw is impalpable. To determine if cervical lymph nodes are swollen from some other cause, the pharynx must be examined carefully. The fauces must be examined for signs of tonsilitis that might cause cervical adenitis. The lymph nodes in contact with the submaxillary and sublingual salivary glands drain the corner of the eye, the side of the nose, the cheeks, the lips, and the floor of the mouth, all of which must be explored, before a diagnosis of submaxillary or sublingual mumps can be made. Laboratory tests are needed to confirm the diagnosis.

In infectious mononucleosis, the glands stand out distinctly and the parotid is not affected. In septic parotitis there is more parotid tenderness; there may be fluctuation, and pus exudes from the orifice of Stensen's duct. Calculus causes spasmodic pain and swelling and may be detected radiographically. Recurrent parotitis and Mikulicz's syndrome are unlikely to be confused with mumps except in the earliest stages, nor are uveoparotid fever and tumours of the gland, as they are chronic conditions.

Orchitis

Orchitis may occur 4 or 5 days after the onset of parotitis. Quite often it occurs without preceding parotitis. It is an acute condition, with chills, sweats, headache, and backache, and a swinging temperature as well as severe local testicular pain and tenderness. The scrotum is swollen and oedematous, and the testicles are impalpable. Usually only one testicle is affected but sometimes both: the second testicle may become affected just as the swelling of the first is subsiding. The illness lasts 3 or 4 days before the swelling begins to subside. Orchitis is unusual before the age of puberty, though it has occurred in young boys and even in infants. In adolescent and young males it develops in 1:5 cases. Some degree of atrophy of the testicle occurs in at least one-third of patients with orchitis. Azoospermia after mumps is rare and only temporary. The fear of sterility after mumps orchitis has been exaggerated, so the doctor can reassure the patient. Orchitis when it occurs without parotitis is difficult to distinguish from gonococcal epididymo-orchitis unless there has been contact with mumps. The rare case of orchitis in infancy may resemble torsion of the testis and perhaps it is safer to operate than risk a serious misdiagnosis.

Meningitis and encephalitis

Lymphocytic or viral meningitis may develop a few days after the start of parotitis, but almost as often it occurs in the absence of parotitis. In the cerebrospinal fluid, protein and lymphocytes are increased. Occasionally the patient develops transient paralysis of limbs. Polyneuritis, neuritis of the trigeminal or facial nerve, and retrobulbar optic neuritis have been described in mumps but all are rare. The meningitis is usually mild and self-limiting. In encephalitis the outlook is different. The patient is confused and may lapse into coma and remain comatose for days, weeks, or months. Almost 2 per cent of the encephalitis cases are fatal.

Other complications

Deafness is sometimes reported after mumps, but it is rarely permanent. Women sometimes complain of ovarian pain during an attack of mumps, but it is rarely as severe as in men with orchitis. There is no evidence that it affects fertility. Mastitis occurs in 15 per cent of the cases, both in females and males, but it is usually mild and fleeting. Mild upper abdominal pain in about 50 per cent of the cases may be related to viral changes in the pancreas. The amount of amylase in duodenal fluid may be less than normal. This is probably caused by a blockage of the ducts in the pancreas. Although there are anecdotal reports of diabetes occurring after an attack of mumps, there is no virological or immunological evidence for a direct link.

Mumps in the fetus and infant

Abortion may occur in women with mumps in the first trimester of pregnancy. It is not common and probably not caused by viral damage to the fetus. The connection between primary endocardial fibroelastosis, which is declining in incidence, and mumps is rather vague, but recent evidence indicates that by RT-PCR viral RNA could be amplified from myocardial samples in more than 70 per cent of the cases. Mumps virus has not been isolated from heart tissue at autopsy and these infants have no mumps antibody in their blood. They may show a delayed hypersensitivity response to the skin test. This has not been explained, but may reflect some immune defect in the fetus which could cause myocarditis and fibroelastosis.

In the normal infant, maternal IgG passes to the fetus and seems to protect the infant against mumps during the first year of life. The typical disease of mumps in infants is a rare clinical finding even in populations with no previous experience of the disease. Orchitis has been reported in infants, and mumps virus may be isolated in vague respiratory infections in infants.

Laboratory diagnosis

In patients without parotitis, especially meningitis, and in the absence of contact history, serological tests, RT-PCR, and virus isolation are the only means of reaching a firm diagnosis. MuV isolation is now rarely used. MuV contains several different antigenic components, which provoke distinct antibodies that are useful for laboratory confirmation. The most important are the HN protein (V antigen) and the N protein (S antigen). S antibody rises in the first 2 weeks of infection but then declines rapidly. V antibody appears at the end of the first week, usually in high titre: it may persist for years and indicates past infection. Neutralizing antibodies also develop. Nowadays, sensitive enzyme immunoassay (EIA) allows early diagnosis by detection of mumps-specific IgM and IgA. IgA can be detected in saliva or mouth washings on about the fourth day after infection, and in the serum early in the disease. Measurement of antibodies in acute and convalescent sera is a reliable method for diagnosis, especially in patients who have no parotitis. Viral antigen produces a tuberculin-like reaction when injected into the skin of people who have been infected with mumps virus before, with or without clinical symptoms. The test is of value in assessing immunity or the need for vaccination.

Treatment

There is no specific treatment. Symptomatic treatment includes simple analgesics, but for the severe pain of orchitis, morphine (15 to 30 mg) may be required for a day or two. Corticosteroids are worth trying in severe cases of parotitis, more especially in orchitis. An adult dose of 60 mg prednisolone daily for 2 or 3 days sometimes gives dramatic relief from pain though it may not reduce the swelling.

Prevention and control

The mainstay of prevention is vaccination of susceptible individuals. Isolation is not effective as the patient has been infectious for days before parotitis occurs and inapparent cases are frequent. Attenuated live vaccine, licensed since 1967, gives 95 per cent seroconversion, and protection lasts for at least 15 years. In developed countries, mumps vaccine is currently given between 14 and 16 months of age as one component of a trivalent mumps/measles/rubella (**MMR**) vaccine, using live attenuated strains of all three viruses. This has succeeded in suppressing the incidence of mumps by more than 98 per cent in the United States and in the United Kingdom. Mumps vaccination is contraindicated in pregnant women and patients with immunodeficiency due to immunosuppressive therapy or disease. However, HIV seropositive children should be vaccinated with the MMR vaccine.

Further reading

Christie AB (1980). *Infectious diseases: epidemiology and clinical practice*, 3rd edn. Churchill Livingstone, Edinburgh.

Feldman HA (1989). Mumps. In: Evans AJ, ed. *Viral infections of humans*, 3rd edn, pp 471–91. Plenum Medical, New York.

Rima BK (1999). Mumps virus. In: Webster RG, Granoff A, eds. *Encyclopedia of virology*, 2nd edn, pp 988–94. Academic Press, London.

Wolinsky JS (1996). Mumps virus. In: Fields BN, Knipe DM, Howley PM, eds. *Virology*, 3rd edn, pp 1243–65. Lipincott–Raven Publishers, Philadelphia.

7.10.6

Measles

H. C. Whittle and P. Aaby

Measles is an acute, highly transmissible viral infection of man spread by aerosolized droplets, which causes much death and suffering, especially among children of the so-called Third World. Its severity varies according to host and socioeconomic factors, not to antigenic variation or alteration in virulence of the virus. There is no reservoir of infection other than in man and no evidence of a carrier state. The virus causes a generalized infection coupled with severe immunosuppression. The chief clinical features result from infection of the skin, mucous membranes, and respiratory tract. Attack rates in home contacts are very high (of the order of 90 per cent), subclinical infection is infrequent, and children are the main victims. Long-life immunity follows the disease. Although global coverage by measles immunization in 1998 was 72 per cent, at least 36 million children are infected annually and 1 million die mainly in sub-Saharan Africa where immunization coverage is low.

Epidemiology

The epidemiology of this global infection varies markedly between developed and developing countries.

In the West, most children are infected between 3 and 6 years of age, when they attend nursery and primary schools. Mortality is low (under 0.05 per cent) and morbidity, although considerable when compared to many other common viral infections, is limited. Most cases occur in the winter and spring, with a biannual epidemic pattern. Recently the epidemicity has been influenced by widespread immunization (Fig. 1), which has dramatically reduced both the number of cases and complications. However, even in the United States, which has the longest experience of systematic immunization, there is now evidence of a resurgence of measles with a higher case fatality in non-immunized subgroups of the population, such as religious minorities who do not believe in vaccines, refugees, illegal immigrants, and the poor in the inner cities.

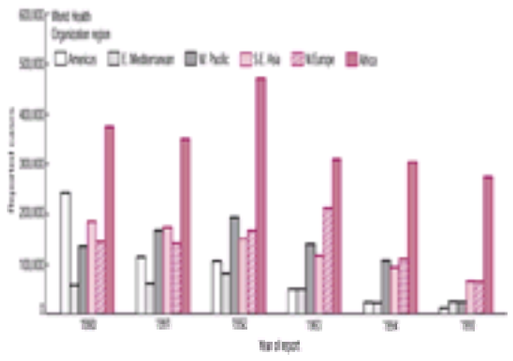

Fig. 1 Reported incidence of measles—World Health Organization regions, 1990–1995.

In the Third World, measles is severe and different: it kills between 3 and 15 per cent of children in the community and some 10 to 20 per cent of those admitted to hospital. Mortality from measles is considerably higher in Africa (5–15 per cent) than in Asia or South America (1–3 per cent); within Africa, West Africa has the highest case-fatality rates. Contrary to the early European experience, when the highest mortality was among the overcrowded urban poor, studies from the developing world indicate a higher mortality in rural rather than urban populations. In communities where females tend to stay at home and are more constrained in their social contacts, mortality is higher in girls than boys. There is a high fatality rate in children with chronic disease, including kwashiorkor, tuberculosis, and human immunodeficiency virus (**HIV**) infection.

There are many reasons for this increase in severity: children are infected at a young age (median age, 12–24 months); severe malnutrition leads to prolonged, severe measles that kills up to 40 per cent of those infected. Overcrowding is another strong determinant of outcome, for secondary and tertiary cases in large families are at great risk of death. Exposure to a large dose of the virus when in close contact with the index case may be an important factor. Furthermore, the severity of measles and the chances of the secondary case dying are dependent on the severity of disease in the index case. Transmission of measles from one sex to the other has been found to increase mortality two- to threefold compared to transmission from the same sex. The high mortality found in West Africa is probably due to the very large, polygamous, and extended families, which increase the risk of intense exposure.

The epidemiology of measles used to vary according to the degree of urbanization, but now the coverage of immunization may be a more important determinant. In remote villages, outbreaks were less frequent and with more susceptible children in each family the risk of intensive exposure and severe disease was increased. These outbreaks resembled the severe epidemics that in the past have devastated the remote island populations of the Faeroes, Fiji, Greenland, and Tristan da Cunha and killed more than 80 per cent of the Inca people in the wake of the Conquistadors in fifteenth-century Middle America. In urban areas, migration, and overcrowding have led to a hyperendemic pattern of infection with a low age of infection but fewer cases per family, the case fatality therefore being lower than in rural outbreaks. Hospital wards and clinics in the developing world are usually important centres of infection; in a hospital in the north of Nigeria, 35 per cent of children with measles had acquired the infection by attending the outpatient clinic 2 to 3 weeks before.

Acute measles is often less severe among children under 6 months of age, among previously immunized children, and among those who have received immunoglobulin when exposed. In these cases the course of infection is characterized by a prolonged incubation period, a short prodrome, mild symptoms, and a favourable outcome. Recent research has provided limited support for the previous belief in a general increase in long-term morbidity and mortality after the first 6 weeks of measles infection. These long-term consequences may have been due to the initial severity of infection as secondary cases have a higher long-term mortality than index cases. However, nowadays in areas with high vaccination coverage the disease is milder. This is particularly true for index cases who tend to be older than the secondary cases and have a lower mortality than uninfected children (suggesting that there is a beneficial effect of mild measles). Long-term morbidity is most likely to be experienced by young children who have severe measles following intensive exposure.

The severity and mortality of measles in developing countries has decreased dramatically as a consequence of measles immunization, which reduces the severity of infection, increases the number of subclinical cases, and lessens the likelihood of transmission. However, even if coverage is maintained, the epidemiology of measles alters as a result of changes in herd immunity from exposure to natural measles, and subclinical infection is important in maintaining protective antibody

levels among vaccinated individuals. Thus, as antibody levels fall and the number of unvaccinated and unexposed subjects rise, there will be an increasing potential for epidemics among young adults. This is particularly true in developing countries where fertility rates are high; for measles is severe in young pregnant women and may attack both mother and child at the same time.

Popular beliefs

In most cultures, measles has a specific local name and is a much feared disease. Popular understanding is centred around the rash, which if it stays within the body will lead to severe disease. This belief has some basis in truth for the prodrome is prolonged in severe cases, and a proportion of deaths reportedly occur before the appearance of the rash during very severe epidemics. Therapeutic practices, such as rubbing the skin with palm oil or kerosene, are aimed at eliciting the rash quickly. In West Africa it is believed that cooling keeps the rash within the body, so the child may be bedded in warm sand or covered with blankets, and is not washed or given cold water to drink. Such customs may aggravate dehydration. In West Africa, as a result of popular awareness of measles, good correspondence exists between parental diagnosis and that based on clinical and immunological assessments. The mother's diagnosis, which can be used for epidemiological surveillance, is nearly always correct.

The virus and its antigens

Measles mainly infects humans, but like the other closely related morbilliviruses (such as rinderpest or canine distemper virus) it is able to cross species to infect other primates and, on occasions, dogs. It contains a single strand of RNA, is highly pleomorphic, and ranges from 100 to 300 nm in diameter. The virus propagates by budding from the cell membrane, from which it acquires an envelope. The membrane of infected cells and the virion envelope contain two surface glycoproteins, the haemagglutinin (**H**) and fusion (**F**) proteins, and a non-glycosylated matrix (**M**) protein, which forms the inner layer. The H protein, which allows attachment of the virus to cells, via the CD46 or CDw150 receptors, is the main target for neutralizing antibodies; the F protein is responsible for fusion and syncytium formation of infected cells. CD46 is a ubiquitous membrane cofactor protein, which together with five other proteins, protects cells from complement activation and lysis. Vaccine strains of measles virus bind to CD46 to downregulate the protein, resulting in complement-dependent cell lysis which limits viral replication. Some wild-type viruses, but not all, bind to the receptor but do not downregulate it, thus preventing lysis and allowing efficient viral replication. The CDw150 receptor (also known as signalling lymphocyte-activator molecule or **SLAM**) is expressed on immature lymphocytes and on effector memory T cells, and is rapidly induced on T and B cells after activation. The internal components or nucleocapsid consist of RNA, the nucleoprotein (**N**), which is the major protein, the phosphoprotein (**P**) and the large protein (**L**). The F protein is remarkably stable, the H protein shows minor antigenic variation, but the N protein, which contains a variable region in the C terminus, is highly divergent among different strains of virus. Genetic analysis of *H* and *N* genes allowed molecular surveillance of the measles virus in the United States, which suggested that the majority of cases were the result of international spread of the virus. There is also variation in the M protein, which some claim is related to persistent infection. The replication and assembly of measles virus is shown in Fig. 2.

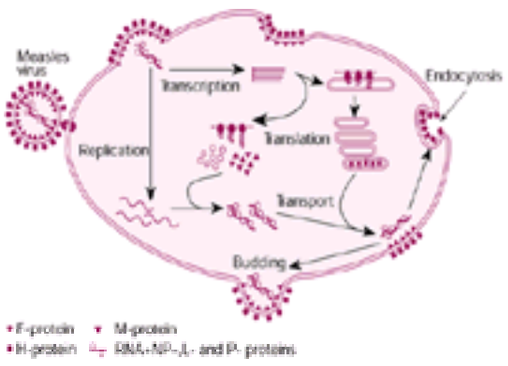

Fig. 2 Replication and assembly of measles virus. (Reproduced by courtesy of van Binnedijk RS (1992). T-cell function in measles. PhD thesis. University of Utrecht, Holland.)

Pathogenesis and the immune response

The course of infection and the immune response to this invasion are shown in Fig. 3. The measles virus, which is thermolabile and survives best at low humidities, is spread to susceptible contacts in droplets during sneezing and coughing. First it infects and multiplies in the epithelium of the upper respiratory tract or the conjunctivas. Some 4 to 6 days later the virus is found in the reticuloendothelial tissue of the liver and the spleen after passage through lymph nodes and spread via the blood. Here it multiplies, causing fusion of cells to form giant cells with many nuclei. Viral antigens, which can be found by immunofluorescent techniques in and on the surface of both these cells and lymphocytes, now induce the immune response. First, natural killer cells and cytotoxic T cells mount a cell-mediated reaction that contains the virus and limits its spread within cells. Later, B cells are primed to produce antibody. Defects in the cellular immune system, as in severe malnutrition, cancer or primary immunodeficiencies, allow widespread multiplication of the virus to cause fatal giant-cell pneumonia.

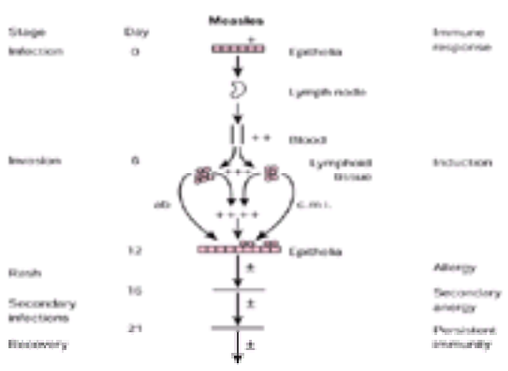

Fig. 3 Pathogenesis of measles. + Denotes amount of virus; ab, antibody. (Reproduced with permission from Parry EHOP (1984). *Principles of medicine in Africa*, 2nd edn. Oxford University Press, Oxford.)

Around day 8, the measles virus is carried by the blood, either free or in mononuclear cells, to the target tissues, which are the epithelia of the eye, lung, and gut. Again, the agent multiplies to cause a bright redness of the mucosa and Koplik's spots (see below), which are foci of viral multiplication. At this stage, measles virus may be cultured from nasopharyngeal secretions, and antigen can be detected by immunofluorescent techniques in the characteristic giant cells of the buccal mucosa, in epithelial cells, and in both B and T lymphocytes in the blood.

The rash, appearing around days 14 to 16, is the sign of a strong and complicated allergic reaction to the virus in the epithelia. The extent and severity of the rash, which reflects the clinical severity of the disease, is determined by the number of target cells infected. Histological examination shows virus in the disrupted epidermis, in the corium, and in capillary endothelium. These tissues are infiltrated by mononuclear cells together with antibody, immune complexes, and complement. An intact cell-mediated immune response is essential to generate the rash and clear the virus, for if impaired, as in the case of children with leukaemia, or occasionally in severe kwashiorkor, the virus multiplies unchecked and no rash appears. Some 2 or 3 days after the start of the rash, around day 17 or 18, the virus can no longer be cultured in the epithelia, for infected cells have been disrupted and the free virus neutralized by antibody. The first antibody to appear is to the nucleoprotein antigens. The second, which is largely responsible for neutralization of the virus, is to the haemagglutinin. Finally, the antibody to the fusion glycoprotein appears in a low titre. This antibody stops cell-to-cell spread of the virus. At this stage the child is markedly immunosuppressed and thus susceptible to secondary infections of the eyes, mouth, gut, and lungs. Latent viruses, such as herpes simplex or cytomegalovirus, may be reactivated and in turn cause further immunosuppression. The delayed

hypersensitivity reaction, as measured by skin tests to old tuberculin or candida antigen, is absent or severely impaired.

By the third week, day 21, as the patient recovers, antibody is in full production. Levels remain elevated for the rest of the patient's life, either because of repeated subclinical infections or because the virus persists in latent form in the spleen and other organs, so stimulating antibody. Occasionally the virus persists in the brain in a damaging form to cause subacute sclerosing panencephalitis (see below). In this rare condition, virus can be isolated from the brain up to 8 years after measles, and antibody levels to all but the M protein antigen are raised in the cerebrospinal fluid and blood. The immune system, for unknown reasons, has failed to clear the virus, which is probably aberrant, for such strains are unable to produce normal amounts of protein.

The mechanisms of immunosuppression are complex. The cytotoxic T-cell response, which is exuberant, may result in the destruction of infected T cells and dendritic cells thus leading to their depletion, deficient antigen processing, and generalized immunosuppression. Crossbinding of the CD46 cellular receptor downregulates IL-12, a crucial cytokine in the development of TH-1 and delayed hypersensitivity responses (Fig. 4). Infection of CDw150+ lymphocytes, which are predominantly of the T_{H0}/T_{H1} type, results in suppression of lymphoproliferation and cell death. Thus measles ultimately dampens the TH1 response, resulting in a skewing towards a TH2 cytokine response and susceptibility to intracellular and other pathogens. However, this immunosuppression may be in the interest of the host by limiting further autoallergic damage of infected tissues.

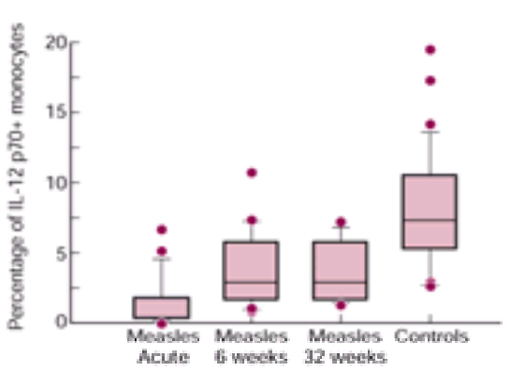

Fig. 4 Production of interleukin-12 by monocytes is depressed long after measles. (Reproduced with permission from Atabani SF *et al.* (2001). *Journal of Infectious Diseases* **184**, 1–9.)

Pathogenesis in the underprivileged, in the malnourished, and in the HIV-infected

Measles in the children of the Third World, as was formerly the case in the underprivileged in Europe, is severe, prolonged, and carries a high case-fatality rate due to secondary infections. Two explanations are offered. Crowding leads to a high dose of measles virus and also increases the chances of secondary infection. The period of incubation has been found to be short in severe and fatal cases, which is consistent with the emphasis on infecting dose as a mechanism of severe disease. Alternatively, or in tandem, malnutrition diminishes the immune response to the virus, allowing great proliferation of virus and subsequent damage to the host. There is experimental evidence, although only in severely malnourished children with marasmus or kwashiorkor, that the lymphocytes of these patients may be more readily infected during the induction phase. A normal immune response follows, which generates a severe and widespread rash followed by prolonged immunosuppression. Secondary bacterial infections with, for example, *Streptococcus pneumoniae*, or latent infections such as herpes simplex or *Mycobacterium tuberculosis* follow in the wake of this intense immunosuppression, often killing or maiming the child. Virus persists in lymphocytes and epithelial cells for up to 30 days after the start of the rash. Antibody production occasionally fails and secretory IgA is deficient, which may explain why the virus persists for so long in the gut. Anorexia, increased catabolism, protein loss from the gut, and further malnutrition exaggerate the problem, which is worst in the weanling child (Fig. 5).

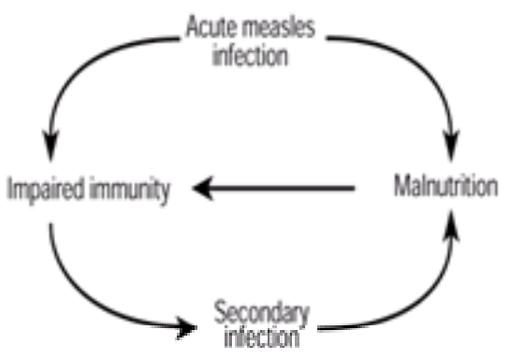

Fig. 5 The complex interaction between infection, nutrition, and impaired immunity seen in measles. (Reproduced with permission from Greenwood BM (1996). The host's response to infection. In: Weatherall DJ, Ledingham JGGL, Warrell DA (1996). *Oxford Textbook Medicine*, 3rd edn, p. 282. Oxford University Press.)

Over 1 million children under the age of 5 years are living with HIV infection in sub-Saharan Africa. The impact of measles and measles vaccination on disease and death in this population will probably depend, as in malnutrition, on the degree of immune damage at the time of infection. The death rate after measles in hospitalized infants is higher in HIV-infected children, and prolonged viral shedding, as detected by the polymerase chain reaction (**PCR**), occurs in the majority of these children. Thus, in regions of high prevalence, HIV-infected children may be important unrecognized transmitters of the virus. Asymptomatic HIV-infected children respond normally to vaccination, but those with **AIDS** (acquired immunodeficiency syndrome) are less likely to respond and may be threatened by persistent infection. Further research is needed on the interaction of the two infections and their impact on the epidemiology and eradication of measles.

Clinical features

There is a spectrum of severity that ranges from mild in the privileged and well nourished to severe in the blatantly malnourished or immunosuppressed. However, the rule is not inviolate and other factors such as the age and dose of infection are probably as important in determining the severity of disease. Measles, often severe, occasionally infects unvaccinated young adults or individuals who have lived in isolated communities. The clinical features of measles and some complications are shown in Fig. 6 and discussed below.

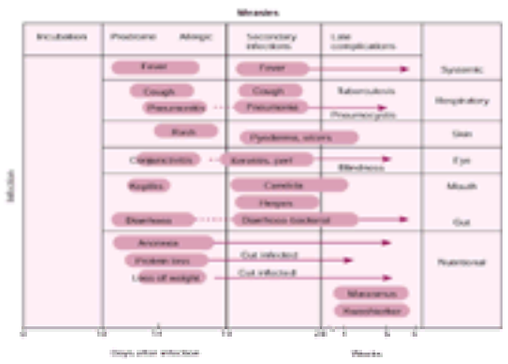

Fig. 6 Clinical features of measles and some of its complications. (Reproduced with permission from Parry EHOP (1984). *Principles of medicine in Africa*, 2nd edn,

Oxford University Press, Oxford.)

Prodrome (days 10–14)

A diagnosis of measles is often missed at this stage, when fever coupled with a runny nose, and sometimes complicated by convulsions, is the main feature. Other signs are mild conjunctivitis, red mucosa, Koplik's spots, and diarrhoea. Koplik's spots are found in the buccal mucosa. They are 'small irregular spots of bright-red colour; in the centre of each spot is seen a minute bluish-white speck'. A useful diagnostic test is to scrape the buccal mucosa with a spatula, place the scraping on a microscope slide, stain with Leishman's stain, and examine for giant cells under a microscope. The prodrome is prolonged in severe cases, and reduced in individuals with modified measles due to maternal antibodies or the prophylactic use of immunoglobulin.

Rash (days 14–18)

The morbilliform rash first appears on the forehead and neck and then spreads, over a period of 3 to 4 days, to involve the trunk and finally the limbs (Plate 1).

In children in Africa and other parts of the developing world the rash is often red, confluent, raised, very extensive, and sometimes accompanied by bleeding into the skin and gut (Plate 2). Later the rash blackens, then the skin peels causing extensive desquamation. Other epithelial surfaces are inflamed, the severity matching that of the rash. Cough, a cardinal sign, may be hoarse and coupled with inspiration difficulty if the larynx and trachea are inflamed. Signs of pneumonitis are apparent, which in severe cases may cause cyanosis or be complicated by mediastinal and subcutaneous emphysema. Conjunctivitis, especially in those who are vitamin-A deficient, can be severe; enteritis may cause profuse diarrhoea with a resulting loss of protein, and malabsorption of food and water. The mouth is painful and red, which adds to the misery of the child, who becomes anorexic and may even refuse to suck the breast. In the uncomplicated case, as is usual in the West, the convalescent period is short, usually lasting less than a week. Complications should be suspected if fever persists while the rash is fading or desquamating.

Complications

Early complications (days 18–30)

As a result of the widespread, severe allergic reaction to the measles virus signified by the rash, the patient is left severely immunosuppressed and is susceptible to infection.

Pneumonia

This causes the most deaths (Table 1) and is heralded by a rise in fever, leucocytosis, and respiratory difficulties. Lobar pneumonia is usually caused by *S. pneumoniae*, but bronchopneumonia, which is more common, results from other bacteria, such as *S. aureus*, or secondary viral infections with, for example, herpes simplex or adenovirus. A variety of other organisms such as Gram-negative bacteria, cytomegalovirus, fungi, *M. tuberculosis*, and *Pneumocystis carinii* should be considered as potential lung pathogens in the malnourished or immunocompromised child.

Stomatitis and enteritis

Chronic diarrhoea and a sore mouth caused by candidal infection are common complications of measles in children in the Third World. The gut is often superinfected with *Bacteroides* spp., *Escherichia coli*, *Pseudomonas* spp., and *S. aureus*, which results in malabsorption and protein loss. Deep ulcers caused by herpes simplex virus erode the corners of the mouth, gums, and inner surface of the lips causing much misery, illness, and pain (Plate 3).

Eye infections

Corneal ulceration leading to impaired vision or blindness is common after measles, especially in malnourished and vitamin A-deficient children (Plate 4). Several studies from Africa have shown that more than half of childhood blindness is related to measles. The mechanisms are still under discussion. In northern Nigeria, herpes simplex was found in 47 per cent of active corneal ulcers after measles, and measles virus in 12 per cent: the children often had evidence of oral herpes. In a study in Tanzania, blindness precipitated by measles was associated with vitamin A deficiency (50 per cent), *herpes simplex* infection (21 per cent), and the use of traditional eye medicine (17 per cent).

Skin and other infections

Pyoderma is common after measles. In the malnourished patient, deep eroding ulcers may bore through the skin even into bone. When originating in the mouth they are know as cancrum oris or noma (Plate 5). Otitis media is also common.

Encephalitis

This is a rare, but much feared, complication found in approximately 1 to 2 per 1000 cases. The onset is usually between 4 and 7 days after the start of the rash, but, rarely, it may occur within 48 h or up to 2 weeks from the onset. In addition to seizures, there is often fever, irritability, headache, and a disturbance in consciousness that may progress to profound coma. The disorder is probably attributable to a neuroallergic process: lymphocytes from the cerebrospinal fluid have been shown to respond to myelin basic protein, as in experimental allergic encephalomyelitis. The virus cannot be isolated from cerebrospinal fluid, which contains lymphocytes and raised levels of IgG but normal levels of measles antibody. Mortality and morbidity is high: 10 to 15 per cent of victims die and 25 per cent of children are left with permanent brain damage. Treatment is supportive; dexamethasone has no convincing beneficial effect.

Late complications

Malnutrition

This is the most frequent complication, for children of the developing world often lose a lot of weight during measles and may take many weeks to regain it. Those originally underweight, who have had severe measles, are at greatest risk, for anorexia in these children is prolonged, much protein is lost from the gut, and secondary infections, which lead to marasmus or marasmic kwashiorkor, are frequent. Measles has been shown to persist in the epithelia and lymphocytes of the severely malnourished for 30 or more days after the rash.

Persistent infection

Pneumonitis

A giant-cell pneumonia is found in patients with defects in cell-mediated immunity; children with leukaemia or kwashiorkor are particularly vulnerable as are those with symptomatic HIV infection. The lung disease may develop weeks after measles, and in most cases the rash of measles has been absent and thus the diagnosis may not be suspected. The diagnosis is made by virological and/or histological examination of lung tissue. Most of these children die.

Subacute sclerosing panencephalitis (SSPE)

Persistent measles virus infection is responsible for this rare, progressive disease of the brain, which is found in 0.1 to 1.4 per million children after measles. The virus is detected in the brain by electron microscopy and by immunofluorescent methods, and has only been isolated using cocultivation techniques. The child with SSPE has usually experienced normal measles, albeit at a young age, 5 to 10 years earlier. The first indication is a disturbance in intellect and personality; behaviour

disorders and deterioration in school work are frequently mentioned. There then follows, over a period of weeks and months, myoclonus-like seizures, signs of extrapyramidal and pyramidal disease, and finally a state of decerebrate rigidity followed by death. The electroencephalogram shows a characteristic regular series of high-amplitude, spike-like waves. Very high titres of measles complement-fixing and haemagglutinin-inhibiting antibody are present both in serum and cerebrospinal fluid; the haemagglutin-inhibiting antibody is probably produced within the nervous system. Treatments for SSPE have included the use of transfer factor, plasmapheresis, and antiviral drugs, but to no avail.

Multiple sclerosis, autism, Crohn's disease

There is no convincing evidence that measles virus or immune responses to it have a causative role in these diseases. The alleged association between the measles, mumps, and rubella (**MMR**) vaccine, autism, and Crohn's disease was based on weak science and has now been convincingly refuted by larger and stronger epidemiological studies. Subsequent molecular studies have failed to confirm the original finding of measles virus and genomic RNA in diseased bowel.

Diagnosis

This is primarily clinical, although signs may be less clear-cut in vaccinated subjects. Thus in areas of high vaccine coverage the detection of measles-specific IgM antibody by enzyme-linked immunoassay (**ELISA**) or, better still, the detection of measles antigen in saliva or urine may clinch the diagnosis if the rash is mild or atypical. Subclinical measles is common in vaccinated children after exposure to measles: the diagnosis is made by detecting a fourfold or greater rise in measles antibody within 2 to 6 weeks of exposure. It is not clear if such cases are infectious.

Treatment of measles and its complications

No effective antimeasles drug exists, yet some children do benefit from treatment in hospital. The following criteria indicate severe measles and a need for hospital admission: a widespread, confluent rash darkening to deep red or purple; signs of laryngeal obstruction; subcutaneous emphysema; marked dehydration; blood in the stool or more than five stools a day; convulsion or loss of consciousness; severe secondary pneumonia; corneal ulceration; or severe ulceration of the mouth and skin. These signs should be taken particularly seriously when the child is underweight or frankly malnourished.

Hydrate the child orally or intravenously. Treat lobar pneumonia with benzylpenicillin, and bronchopneumonia with combined antibiotics such as gentamicin and cloxacillin, or with co-trimoxazole. Antibiotic eye ointments relieve discomfort and possibly prevent secondary infections of measles conjunctivitis. Antibiotics (topical and systemic) and vitamin A should be given routinely for the treatment of eye ulcers. If *herpes simplex* virus is the cause, use aciclovir topically or, when severe, systemically. Candidal infections of the mouth or gut often respond dramatically to nystatin. Feeding, by tube if necessary, needs careful planning and presentation, for the anorexic infected child will be in severe negative energy balance due to a greatly increased catabolic rate. Case fatality rates are 30 to 50 per cent lower in those children in hospital treated with vitamin A. This should be given orally at the time of diagnosis in a dose of 100 000 IU for children below 12 months of age and in a dose of 200 000 IU for older children. If eye signs of vitamin A deficiency are present, the initial dose should be repeated the next day and again 1 to 4 weeks later.

No specific effective treatment for encephalitis or SSPE exists.

Prevention

Passive immunization with human immunoglobulin is highly effective if given within 2 or 3 days of exposure, in a dose for children of 0.2 ml/kg. Immunoglobulin should be given to those in whom vaccination is contraindicated: children immunosuppressed by kwashiorkor, by malignancies such as leukaemia or lymphoma, or by steroids or cytotoxic drugs. Passive immunization may also be used to protect pregnant women, those with active tuberculosis, and those with AIDS.

The currently used vaccines are live strains, attenuated by culture in chick fibroblasts. The complications of vaccination are few and generally mild. Fever of moderate severity is infrequent, and a mild rash with some signs of upper respiratory tract infection occurs rarely. Encephalitis or SSPE is exceedingly rare after vaccination. Underweight children respond normally to the vaccine, as do ill children attending the outpatient department and those on the ward. As clinics and hospitals are major sites of transmission of the virus in the developing world, all susceptible children in these places—unless seriously immunocompromised because of, for example, leukaemia, kwashiorkor, or AIDS—should be vaccinated.

The first measles vaccines introduced in the United States and Europe in the 1960s were inactivated with formalin. Although they produced high levels of H antibody, they failed to raise antibody to the F protein. This gave poor protection, and on exposure a severe local reaction at the site of injection or a bizarre form of measles resulted. The rash was unusual, having urticarial and vesicular features, fever was high, oedema of the limbs frequent, and severe pneumonia present. This syndrome is still occasionally seen in adults who were vaccinated with killed vaccine during childhood. Rhesus macaques immunized with formalin-inactivated vaccine followed by challenge with measles virus developed an atypical rash and pneumonitis accompanied by immune-complex formation and eosinophilia. This experiment shows that atypical measles results from a non-protective type-2 CD4 T-cell response.

The optimal age for vaccination in the developed world is between 14 and 16 months, when maternal antibody, which neutralizes the virus to cause vaccine failure, has disappeared. This recommendation does not apply to children in the Third World, because there measles infects at an early age. The World Health Organization (**WHO**) recommends vaccination at 9 months of age but, by then, 5 to 15 per cent of children may have had measles. Edmonston–Zagreb, a different strain of vaccine, passaged in human diploid cells, and given in high dose (in excess of 10^5 infectious particles) subcutaneously, has proved to be immunogenic at the age of 4 to 6 months. Subsequently, the WHO recommended its use at 6 months of age in areas of the world with a high incidence of measles occurring below 9 months of age. However, long-term follow-up demonstrated lower survival rates among female recipients of high-titre vaccine than among female recipients of standard-dose measles vaccine. Although the biological explanation for this unexpected finding is unknown, the use of high-titre vaccines is now no longer recommended.

Some scientists have argued that measles vaccines will have a limited impact on childhood survival—for disadvantaged children, saved from measles by vaccination, will only die at a later date from other infections or malnutrition. However, a variety of epidemiological studies has documented remarkable reductions in mortality after standard measles vaccine. In Bangladesh, measles vaccination was associated with a 36 per cent reduction in all-cause mortality from the age of 9 months, despite the fact that acute measles accounted for only 4 per cent of deaths in the community. The reason for this unexpected benefit is unknown but it is not related to the prevention of measles. The benefit is particularly marked for girls and is most likely due to non-specific immune stimulation.

Primary vaccine failure may occur because maternal antibody neutralized the vaccine, or it may be due to heat-inactivation of improperly stored vaccine. Secondary vaccine failure may occur because the children are intensively exposed or because protective antibody levels wane. This has not been a problem in developed countries where most children have received their vaccine after 15 months of age, but in developing countries early vaccination at 6 to 9 months of age may increase the likelihood of antibody concentrations declining to non-protective levels.

Waning immunity may become an increasing problem as vaccine coverage increases: because more mothers will have been vaccinated and since they have not been exposed or had natural measles, they will transmit lower levels of maternal antibody. Thus their babies become susceptible to measles by 3 to 5 months of age. This increases the problem of measles control: the two groups of children, born to vaccinated or naturally infected mothers, will have low or high levels of maternal derived antibody, respectively, and thus will need to be vaccinated at different ages.

...or eradication?

Measles eradication has yet to be made an official global policy. However, several regions are pursuing a policy of measles elimination. The United States has been successful with a policy of obligatory vaccination before school entry. The Pan-American Health Organization (**PAHO**) has pursued a successful but expensive policy of national immunization days for all children aged between 1 and 14 years, high routine vaccination coverage, and periodic follow-up campaigns during which all children aged 1 to 4 years are vaccinated. However, the real test of global measles eradication will be in Europe and Africa. Europe, which at best has a coverage of 85 to 90 per cent, has no tradition of an obligatory use of vaccines and an increasing proportion of the population are averse to vaccination. Africa will be a stern test of the PAHO strategy, for due to political instability it is difficult to maintain a sufficiently high coverage for a long period. In addition, because of high fertility rates there will be a constant renewal of susceptible children, combined with an accumulation of susceptible young adults due to waning immunity.

New vaccines, which can be given in early infancy, or two-dose strategies using the standard vaccine at 6 and 9 months of age, will be necessary to contain measles in the developing world. Coverage of at least 95 per cent of all susceptible children, including those between 3 and 9 months of age, with a vaccine that is at least 95 per cent effective will be necessary if the virus is to be eradicated. Current vaccines do not meet these standards. New vaccines such as the Modified Vaccinia Ankara recombinant virus, a non-replicating mutant of horsepox made to express the F and H proteins, or a DNA vaccine expressing these proteins may possibly fulfil such exacting requirements, for they have been shown to protect macaques from measles. However, in those countries with a high child mortality rate where measles vaccine has been shown to have non-specific beneficial effects, vaccination programmes may need to be continued even if measles is eradicated.

Further reading

Aaby P, *et al.* (1983). Measles mortality, state of nutrition, and family structure. A community study from Guinea-Bissau. *Journal of Infectious Disease* **147**, 693–701.

Aaby P, *et al.* (1995). Non-specific beneficial effects of measles immunization: analysis of mortality studies from developing countries. *British Medical Journal* **311,** 481–5.

Cutts FJ, Steinglass R (1998). Should measles be eradicated? *British Medical Journal* **316**, 765–7.

Fenner F (1948). The pathogenesis of the acute exanthems. An interpretation based on experimental investigations with mouse-pox (infectious ectromelia of mice). *Lancet* **ii,** 915–20.

Karp CL (1999). Measles immunosuppression, interleukin 12, and complement receptors. *Immunological Reviews* **168**, 91–101.

Morley D (1969). Severe measles in the tropics. *British Medical Journal* **1,** 297–300; 363–5.

Polack FP, *et al*. (1999). Production of atypical measles in rhesus macaques: evidence for disease mediated by immune complex formation and eosinophils in the presence of fusion—inhibiting antibody. *Nature Medicine* **5**, 629–34.

Whittle HC, *et al.* (1979). Severe ulcerative herpes of mouth and eye following measles. *Transactions of the Royal Society for Tropical Medicine and Hygiene* **73,** 66–9.

Whittle HC, *et al*. (1999). Effect of sub-clinical infection on maintaining immunity against measles in vaccinated children in West Africa. *Lancet* **353**, 98–101.

7.10.6.1Nipah and Hendra viruses

James G. Olson

Nipah virus

Introduction

An outbreak of severe, febrile encephalitis associated with human deaths occurred in Malaysia, beginning in September 1998. Initially recognized near Ipoh in the northern state of Perak, it was attributed to an endemic mosquito-borne viral encephalitis caused by Japanese encephalitis virus which is amplified by swine. Veterinary and public health measures for control of Japanese encephalitis failed to have an impact on the encephalitis in humans. In January 1999, the outbreak spread to the state of Negri Sembilan and cases increased dramatically.

Aetiology

A new virus of the family Paramyxoviridae, genus *Megamyxovirus*, shows characteristic syncytia and giant cell formation in Vero cell culture. Nipah virus, named for the location where the first isolate was obtained, is approximately 1.1 µm in length with an average diameter of 21 nm. Extracellular, pleomorphic virus particles average 500 nm but vary greatly in size. Surface projections along the envelope are seen only sporadically and measure 10 nm in length. Nipah virus differs genetically from its closest relative, Hendra virus, at the nucleotide level in the phosphoprotein (11 per cent), nucleoprotein (26 per cent), and matrix protein (31 per cent) genes, respectively.

Epidemiology

Nearly all patients had a history of direct contact with pigs. Exposures were primarily occupational and frequently included close contact with pigs with respiratory disease. Since late 1996, a new disease of swine has spread among pig farms in Malaysia. It was not identified as a new syndrome because morbidity and mortality were not considered excessive, and the clinical signs were not markedly different from those of a range of other swine diseases. Transmission of virus among pig farms was by trade and movement of asymptomatic pigs. Other species have become infected with Nipah virus, but only in circumstances where they were exposed to infected pig farms. Serological studies have implicated two species of pterapid fruit bats (flying foxes) as possible wildlife reservoirs of infection (Plate 1).

Clinical features (see also Chapter 24.14.2)

Nipah virus encephalitis is a severe, acute, febrile disease with prodrome of fever, headache, and drowsiness. On admission, patients have fever (70 per cent), drowsiness (50 per cent), disorientation (18 per cent), neck stiffness (12 per cent), myoclonus (10 per cent), unresponsiveness or coma (9 per cent), and seizures (6 per cent). Most (58 per cent) cerebrospinal fluid white blood cell counts were normal, while cerebrospinal fluid protein levels were increased (68 per cent). Platelet counts were mildly decreased (median, 137 × 10^9/l). Of patients who had CT scans, 16 per cent showed cerebral oedema and 14 per cent had focal hyper densities. MRI scans of brain revealed multiple small hyper intensity signals on T2-weighted images. New findings following admission include hypotension (40 per cent), signs of autonomic dysfunction (26 per cent), seizures (22 per cent), and myoclonus (20 per cent). Almost half (49 per cent) of patients became unresponsive or comatose and 50 per cent required intubation. The case fatality ratio was 43 per cent, with the median time from admission to death of 4 days. Intubation, seizures, myoclonus, hypotension, and autonomic dysfunction were grave prognostic indicators.

Pathology

Histological and immunohistochemical findings at autopsy included a systemic vasculitis with fibrin thrombi and fibrinoid necrosis. Endothelial cells of affected vessels showed occasional multinucleated syncytial giant cell formation. Endothelial cells also showed lytic necrosis and sloughing into the lumen of the blood vessels. In the central nervous system, the vasculitis was diffuse involving cerebral cortex and brain stem and was associated with extensive areas of rarefaction necrosis. In these areas, neuronal degeneration and death, neuronophagia, microglial nodules, and mild perivascular inflammatory infiltrates were present. Eosinophilic, mainly intracytoplasmic, viral inclusions were seen in affected neurons and parenchymal cells.

Laboratory diagnosis

Enzyme immunoassay and immunohistochemical tests for Nipah virus infections are sensitive and specific methods for laboratory confirmation of clinical illness. Viral isolation in Vero cells can be successful using throat washings, serum, and urine from acutely ill patients and from brain of patients who die. Isolation of Nipah virus is not recommended for laboratory confirmation because of the extreme hazard (biosafety level 4) of the virus.

Treatment

Clinical management of patients is primarily supportive. Most patients required intubation. Patients with hypotension were treated with intravenous fluids and pressor agents, while those with seizures and myoclonus received anticonvulsive drugs. Oral tribavirin (ribavirin) was administered to several patients but data are not available on its efficacy.

Prevention and control

Control areas with active disease were identified, all pig farms within the designated areas were culled, and the pigs buried. This broke the cycle of transmission to humans. Serological testing of swine showed that on infected farms most of the adult pig population had been exposed, an observation which formed the basis of a national testing and eradication programme. Serum samples from every pig farm in Malaysia outside the control were sampled twice at a minimum interval of 3 weeks. Nipah virus should be eradicated from the Malaysian pig herd.

Hendra virus

Since its first description in Australia in 1994, there have been two outbreaks of fatal disease in horses and humans in Queensland, caused by Hendra virus (formerly known as equine morbillivirus). This is closely related to Nipah virus (see above) and Menangle virus, which was responsible for disease in pigs and humans in New South Wales in 1997. Clinical features of Hendra virus infection in humans were pneumonitis and meningoencephalitis. The natural hosts of Hendra, Menangle, and Nipah viruses and Australian bat Lyssavirus (see Chapter 7.10.10) are species of flying foxes (genus *Pteropus*). In eastern Australia, 20 per cent of flying foxes (*P. poliocephalus* and *P. alecto*) were seropositive for Hendra virus. In these bats, the disease is usually subclinical. Virus is found in uterine fluid and is transmissible

oronasally to horses, cats, and other mammals.

Further reading

Chua KB, Goh KJ, Wong KT, *et al*. (1999). Fatal encephalitis due to Nipah virus among pig-farmers in Malaysia. *Lancet* **354**, 1256–9.

Department of Health (UK) (2000). *Hendra virus and Nipah virus. Management and control*. www.doh.gov.uk/jointunit/jip.htm.

Goh KJ, Tan CT, Cheu NK, *et al*. (2000). Clinical features of Nipah virus encephalitis among pig farmers in Malaysia. *New England Journal of Medicine* **342**, 1229–35.

Halpin K, Young PL, Field HE, Mackenzie JS (1999). Newly-discovered viruses of flying foxes. *Vetinary Microbiology* **68**, 83–7.

Halpin K, Young PL, Field HE, Mackenzie JS (2000). Isolation of Hendra virus from Pteropid bats: a natural reservoir of Hendra virus. *Journal of General Virology* **91**, 1927–32.

Lee KE, Umapathi T, Tan CB, *et al*. (1999). The neurological manifestations of Nipah virus encephalitis, a novel paramyxovirus. *Annals of Neurology* **46**, 428–32.

Paaton NI, Leo YS, Zaki SR, *et al*. (1999). Outbreak of Nipah-virus infection among abattoir workers in Singapore. *Lancet* **354**,1253–6.

7.10.7 Enterovirus infections

Ulrich Desselberger and Philip Minor

Enteroviruses are a major group of viruses causing systemic infection in man. They form two genera of the Picornaviridae family (the Enterovirus and Parechovirus genera) and occur in at least 66 serotypes in humans. They infect via the gastrointestinal tract and are mostly clinically inapparent. However, viraemia can be followed by infection of organs distant from the site of entry with often devastating affects in the form of meningitis, encephalitis, myopericarditis, and also rashes and conjunctivitis.

The viruses

Viruses of the Picornaviridae are unenveloped icosahedral particles of 27 to 30 nm in diameter, containing single-stranded RNA of positive polarity and approximately 7.5 kb size as their genome. The nucleic acid is poladenylated at the 3' end and carries a small protein, VPg, covalently linked at its 5' end. The enteroviruses and parechoviruses form two of the six current genera of the Picornaviridae family, the other four being rhinoviruses, cardioviruses, aphthoviruses, and hepatoviruses. Three serotypes of poliomyelitis virus (polio virus), 23 types of Coxsackie A virus, six types of Coxsackie B virus, and various types of enterocytopathic human orphan (echo) viruses are recognized within the enterovirus genus. The parechoviruses comprise echoviruses 22 and 23 and were established as a separate genus on the basis of the highly distinctive nature of the sequence of their genomes. Other classical features of the enteroviruses, such as their stability at acid pH (in contrast to rhinoviruses or aphthoviruses), their buoyant density in caesium chloride gradients, and the nature of their broad clinical effects and persistence in the environment are also shared by the parechoviruses.

The three-dimensional structure of the poliovirus particle has been elucidated by crystallographic analysis. The viral capsid consists of 60 protein subunits, each containing the four unglycosylated viral proteins VP1 to VP4. The capsid proteins are arranged in such a way that VP1 molecules form the apices at the five-fold symmetry axis of the icosahedron whereas two other proteins, VP2 and VP3, are arranged in the centre of the triangular face near the three-fold axis of symmetry. VP4 is an internal protein. All proteins interact with each other. The N terminus of VP4 is myristylated.

Viruses initiate replication by attaching to their cellular receptors, and some of these have been characterized. The poliovirus receptor is a member of the immunoglobulin superfamily. Transgenic mice expressing the human poliovirus receptor become susceptible for poliovirus infection with a pathology similar to that of infected primates. They may eventually replace primates for vaccine testing (see below). Other enterovirus receptors are the decay accelerating factors (DAF; receptor for echovirus 7), implicated in the complement pathway, and the integrin VLA-2 (echovirus 1). Other cell surface molecules may be involved in the virus–cell receptor interactions of many enteroviruses, as the expression of an identified receptor is not always sufficient to make a previously resistant cell line susceptible to productive infection. It is also of interest that some strains of poliovirus, mainly of serotype 2, are able to paralyse mice if injected. The receptor involved in mice has not been identified.

The positive sense RNA genome acts as a messenger molecule. All enterovirus RNAs have a long 5' end untranslated region (UTR) of approximately 750 nucleotides length, containing an internal ribosomal entry site (IRES) which is important for binding of ribosomes and translation of RNA into protein. Downstream of the 5' UTR is a large single open reading frame containing three parts: P1, coding for structural proteins VP1 to VP4; P2, coding for proteins 2A, 2B, and 2C; and P3, coding for proteins 3A to 3D. Protein 3C is a viral protease and protein 3D the RNA-dependent RNA polymerase. P2 and P3 proteins (with the exception of VPg = 3B) are only found in infected cells. The P1 to P3 proteins are synthesized as one large precursor, from which the individual proteins are produced by a complex autocleavage and cleavage cascade. RNAs replicates via double-stranded replicative intermediates. The ratio of positive to negative stranded RNA molecules in infected cells is approximately 100:1. Naked enterovirus RNA is infectious upon transfection, and can be transcribed from full length cDNA clones permitting biochemical manipulation and studies of structure and function at the molecular level.

The extensive antigenic variation of enterovirus capsid proteins allows typing into polio-, Coxsackie-, and echoviruses using type-specific neutralizing antisera, but there is some cross-reactivity. The main antigenic sites are located on all three major virion proteins (VP1–VP3), and some involve sequences from more than one protein.

Comparison of complete RNA genome sequences of many enteroviruses show a very close relationship between some enterovirus and rhinovirus sequences. Within the echoviruses, however, there is great diversity, for example echovirus 22 shows a very low degree of homology with any other enteroviruses.

Four subdivisions of human enteroviruses have been proposed according to genomic relatedness:

- Group 1: polioviruses, Coxsackie viruses A1, A11, A13, A17, A18, A21, and A24;
- Group 2: enterovirus types 68–70;
- Group 3: Coxsackie viruses B1–B6, A9, all echoviruses except types 22 and 23, and enterovirus 69;
- Group 4: Coxsackie viruses A2, A3, A5, A7, A8, A14, and A16, and enterovirus 71.

Hepatitis A virus has previously been designated enteroviruses 72, but is now in its own genus Hepatovirus. Echoviruses type 22 and 23 form the Parechovirus genus.

Pathogenesis

The most widely accepted model of the pathogenesis of enterovirus infection is based on that developed by Bodian for poliovirus, in which the virus infects via the gastrointestinal tract and undergoes primary replication in lymphoid cells lining the alimentary tract (oropharyngeal, intestinal). A viraemic phase follows, allowing infection of distant target organs: spinal cord and brain, meninges, myocardium, skeletal muscles, skin, mucous membranes. Other tissues, for example lymph nodes

and brown fat tissue, can also become infected. Intensive multiplication in the CNS leads to the destruction of motor neurones and results in paralysis.

A slightly different and more subtle model of poliovirus pathogenesis was proposed by Sabin, in which the virus infects the mucosal surface, so accounting for the fact that virus can be shed in faeces long after it has become undetectable in lymphoid tissues and when neutralizing antibody is detectable in the blood. The primary replication creates a viraemia which seeds distant, still unknown, sites and virus replication there results in a second viraemia which may be detected about 1 week postinfection and can lead to systemic infection including CNS involvement.

Shedding of virus occurs from the throat and faeces for many weeks and even months after infection and thus ensures transmission (see below). Virus replication in sites distant from the port of entry normally terminates with the appearance of neutralizing antibody, first IgM at 8 to 12 weeks after infection, and then IgG. Children with B cell immunodeficiencies may develop persistent infections.

Most enterovirus infections are silent or produce a 'minor illness' with the symptoms of a mild upper respiratory tract infection with or without fever. In a minority of infections (1 per cent or less) a systemic 'major disease' may develop:

- paralytic poliomyelitis and aseptic meningitis (polioviruses);
- aseptic meningitis, herpangina, conjunctivitis, hand, foot and mouth disease (Coxsackie A viruses);
- aseptic meningitis, myopericarditis, encephalitis, pleurodynia (Coxsackie B viruses; enterovirus 71);
- aseptic meningitis, rashes, conjunctivitis (echoviruses);
- polio-like illness, aseptic meningitis, hand, foot and mouth disease, epidemic conjunctivitis (enterovirus types 68–71).

Symptoms of clinical illness caused by enteroviruses are summarized in Table 1 and are discussed in more detail below.

Clinical symptoms

Central nervous system infections

Poliomyelitis

While there is evidence of poliomyelitis as an ancient human disease as shown on a funerary stele from Middle Kingdom Egypt, about 1300 BC, there is little documentation of its occurrence until near the end of the nineteenth century, when it appeared in epidemics in children (hence the alternative name 'infantile paralysis'). The appearance of poliomyelitis coincided with improvement in standards of public hygiene and is explained by the consequent exposure of infants to infection at a later age. Maternal antibody is capable of confining infection to the gut, where the virus can persist until the immune response develops to eliminate it. In contrast, when maternal antibody has declined in older infants, the virus can spread to sites outside the intestine, causing paralysis.

Even under modern conditions of hygiene, infection with all three poliovirus types is normally inapparent, but illness with neurological symptoms results in about 1 per cent of infections. This can present as aseptic meningitis with neck stiffness, usually recovering after 10 days (abortive or non-paralytic poliomyelitis). Meningitis is also caused by several other enteroviruses (see below). The more serious presentation is paralytic poliomyelitis, appearing 5 to 10 days after a mild upper respiratory tract infection ('minor illness') and progressing to flaccid paralysis resulting from motor neurone destruction ('major illness'). This may be accompanied by spasms and inco-ordination of non-paralysed muscles. Various forms of the 'major illness' reflect infection of different parts of the CNS. Paralysis of limbs results from destruction of motor neurones in the lower part of the spinal cord ('spinal form'), while the more life threatening bulbar poliomyelitis ('bulbar form') involves infections of the medulla oblongata or bulb. Respiratory functions can be affected in both the spinal and bulbar forms of the disease. Encephalitis is rare. In children under 5 years old, paralysis of one leg is most common; in children 5 to 15 years of age, weakness of one limb or paraplegia are frequent; quadriplegia is most common in adults, often accompanied by urinary bladder and respiratory muscle dysfunction. Muscular function in limbs may return slowly but there is residual paralysis in 90 per cent of survivors. Ten to 25 per cent of paralytic cases have bulbar symptoms with hypertension, shock, and dysphonia. Complications are nosocomial pneumonias (by staphylococci or Gram-negative bacteria), urinary tract infections, and emotional problems. The mortality from paralytic polio is 2 to 5 per cent among children and 15 to 30 per cent among adults. Muscle weakness may develop many years after the initial polio disease (postpolio syndrome or postpolio neuromuscular atrophy). A persistent poliovirus infection as cause of this has been assumed, based on the presence of viral RNA in cerebrospinal fluid and neural tissue. However, such RNA has also been found in patients with other neurological and non-neurological diseases and is therefore less likely to be related to the postpolio syndrome. The alternative view is that the postpolio syndrome is anatomical in origin, such that the initial attack of polio destroys motor neurones and reduces the backup available as the patient ages.

Aseptic meningitis

Aseptic meningitis is the most frequent clinical presentation of enterovirus infection and can be caused by Coxsackie viruses of both groups A and B, and echoviruses, mainly types 4, 6, 11, 14, 16, 25, 30, and 31 (see Table 1). The disease starts with fever, headache, neck stiffness, and photophobia. Sensory or motoric deficits are unusual but confusion is common. The symptoms may persist for 4 to 7 days. The CSF usually shows pleocytosis consisting of 10 to 500 leucocytes per µl, mainly lymphocytes. Polymorphonuclear cells may predominate at the onset, but, should they persist, bacterial infection and possibly abscesses should be considered. The protein concentration in CSF may be normal or slightly increased; the glucose level is normal. Complete recovery is the usual outcome of aseptic meningitis.

Encephalitis

Enterovirus encephalitis is rare but may follow aseptic meningitis. Enterovirus infection in patients with hypo- or agammaglobulinaemia may persist for years with chronic meningitis or encephalitis as sequelae, and a high mortality.

Enterovirus 71 infection which is normally associated with hand, foot, and mouth disease has been found to cause severe meningoencephalitis (with brain stem involvement), polio-like acute flaccid paralysis, and a high case fatality rate in children during several recent outbreaks in Bulgaria, Taiwan, and Malaysia. In some of the fatal cases there may have been coinfections with a species B adenovirus. Enterovirus 71 occurs in three genotypes and is rapidly evolving; it is most closely related to Coxsackie virus A16.

Neonatal infections

Neonatal infection followed by severe, generalized disease may be caused by Coxsackie B viruses and echovirus, mainly of types 6, 7, and 11. These viruses seem to be transmitted late in pregnancy, perinatally, or postnatally by the mother or other virus-infected infants in neonatal wards or special care baby units. The infants develop heart failure due to a severe myocarditis, or a meningoencephalitis. Hepatitis and adrenalitis may also occur. The mortality is high. Viruses may be recovered from brain, spinal cord, myocardium, and liver at autopsy.

Bornholm disease (epidemic pleurodynia)

This is usually caused by Coxsackie B viruses but also by echoviruses of types 1, 6, 9, 16, and 19, and by Coxsackie A viruses of types 4, 6, 9, and 10. The disease can strike families in small outbreaks. It typically starts abruptly with fever and chest pain due to the involvement of the intercostal muscles, or abdominal pain resulting from involvement of muscles of the abdomen. There may be severe frontal headache. The symptoms last 3 to 14 days, followed by complete recovery.

Myopericarditis

Enterovirus-induced myocarditis is mostly due to infection with Coxsackie B viruses in the young. The onset of disease is usually acute, very severe, and may be fatal in neonates, however in adolescents and adults it is normally mild. The virus may persist after the initial infection and cause dilated cardiomyopathy. In fatal cases (usually neonates 2–11 days after onset of disease) there is cardiac dilatation, myocyte necrosis, and an inflammatory reaction. The diagnosis is often difficult, particularly in older patients, as pericarditis, coronary artery occlusion, or heart failure may have been diagnosed initially. Typical clinical findings are often

tachycardia, arrhythmias, murmurs, rubs, and cardiomegaly.

Besides causing acute myocarditis, chronic enterovirus infection can lead to chronic myocarditis and dilated cardiomyopathy, possibly due to immunopathologial mechanisms. In chronic disease, neither infectious virus nor viral antigens are normally detected in heart biopsies; however, viral RNA is regularly found in cardiac muscle, suggesting that the viral genome persists. The true significance of the presence of the viral genome in such cases is still under discussion.

The disease can be produced with Coxsackie B viruses in mice. In this animal model there is also initial viraemia and replication in myocytes, but this is followed by disappearance of infectious virus and destruction of myocytes, possibly by autoimmune mechanisms.

Herpangina

This is caused by Coxsackie viruses of types A1 to 6, 8, 10, and 22. Children and young adults between 2 and 20 years of age are mainly affected. The disease presents with acute onset of fever, sore throat and pain on swallowing, also vomiting and abdominal symptoms. Small vesicular lesions or white papules surrounded by a red halo can be seen on the fauces, pharynx, palate, uvula, and tonsils. The disease is mild and self-limiting.

Exanthemata

Rubella-like rashes can be produced by echoviruses of types 4, 9, and 16, but also Coxsackie viruses A9, A16, and B5. Those usually occur in the summer and may be accompanied by fever, malaise, cervical lymphadenopathy, and aseptic meningitis.

Hand, foot, and mouth disease

A typical distribution of vesicular lesions in hands, feet, and mouth (but also buttocks and genitalia) is produced by infection with Coxsackie virus type A16 and enterovirus 71, less frequently with Coxsackie viruses A4, A5, A9 and A10, B2, and B5. Enterovirus 71 may produce more severe clinical symptoms (see above).

Foot and mouth disease

The aphthovirus causing foot and mouth disease in cloven hoofed animals, is endemic in Africa, Asia, and South America. Virus is secreted before blisters of mouth and feet appear in animals. The zoonosis in humans is very rare, with about 37 recorded cases. Human infection occurs from virus entering through broken skin, drinking unpasteurised milk, or by inhalation of droplets. A 2 to 6-day incubation period is followed by blisters of hands, feet, and mouth, fever and sore throat. Complete recovery ensues. No person-to-person spread is recorded.

Conjunctivitis

Several enterovirus types cause conjunctivitis, often affecting large numbers of people epidemically. Most notable causes are echovirus types 7 and 11, Coxsackie virus A24 and B2, and enterovirus 70 that often produces a haemorrhagic conjunctivitis.

Diabetes and pancreatitis

Insulin-dependent diabetes mellitus (IDDM, or type 1 diabetes) is likely to be an autoimmune disorder in which the insulin-secreting pancreatic islet cells (beta cells) are destroyed. The human disease has long been thought to be caused by infectious agents, particularly since association between enterovirus infection and the development of IDDM has been shown in animal model studies (infection of mice with Coxsackie B3–B5 viruses). However, there is also a strong genetic component in the development of IDDM.

Chronic fatigue syndrome (CFS)

Chronic fatigue syndrome (CFS), also known under the names of myalgic encephalomyelitis (ME), Royal Free disease, Iceland disease, postviral fatigue syndrome, and neuromyasthenia, can occur both sporadically and epidemically. The main clinical feature is excess fatiguability of skeletal muscle, accompanied by pain. Other symptoms include headaches, inability to concentrate, paraesthesia, and impairment of short-term memory. A major problem in diagnosis is a clear definition of the clinical entity. Several virus infections have seemed to precede the development of CFS. Those are mainly enterovirus infections, chronic Epstein–Barr virus (EBV) infection and also infections with *Toxoplasma* and *Leptospira* species. The stringency of the association of chronic enterovirus infection with the appearance of CFS is controversial. A recent report of a joint working group of the Royal Colleges of Physicians, Psychiatrists and General Practitioners has concluded that persistence of enteroviruses is unlikely to play a role in the development of CFS. Similar conclusions have been drawn for the possibility of a causal link between chronic EBV infection and CFS (see Chapter 7.10.3).

Gastroenteritis

Although enteroviruses infect via the gastrointestinal tract and readily replicate there, they very rarely cause diarrhoea. Outbreaks of diarrhoea with echovirus type 11 have been reported. In Japan, an entervirus termed Aichivirus which is proposed as the type species of a new genus of the Picornaviridae family has been identified as the cause of multiple outbreaks of gastroenteritis in humans, mostly associated with the consumption of raw oysters. This virus seems to circulate widely in populations of Japan and other South-east Asian countries, with subclinical infections likely to be common (see Chapter 7.10.8).

Laboratory diagnosis of enterovirus infections

Virus isolation

Virus isolation is an excellent procedure to diagnose enterovirus infections. Virus is shed for weeks, and sometimes months, from the primary infection sites (cells lining the gut, see above). Starting from a few days after infection, virus can be found in concentrations of 10^5 to 10^6 tissue culture infectious doses 50 per cent per g faeces ($TCID_{50}$/g faeces). Throat swabs are also a good source for virus, particularly early in infection and when there are respiratory symptoms. In cases of meningitis, enteroviruses can be propagated in cell culture from the CSF, but the method is much less sensitive than genome detection (see below). Viruses are readily isolated in secondary cultures of monkey kidney cells, or in cultures of permanent cell lines derived from human embryonic kidney, human amnion, or human fetal lung. The cytopathic effect (CPE) produced by enteroviruses is non-specific. Typing of a cytopathic agent is carried out using antiserum pools or in multistep procedures. Most Coxsackie A viruses (with the exception of Coxsackie virus A9) do not grow well in cell culture but can be readily isolated by intracerebral, intraperitoneal, or subcutaneous infection of mice, causing flaccid paralysis and death. In contrast, Coxsackie B viruses cause spastic paralysis. Polio- or echoviruses do not usually grow in mice although polioviruses will replicate in transgenic animals that have appropriate receptors (see above).

Serology

Neutralization assays are the method of choice for typing enteroviruses. These tests are labour intensive and not apt for rapid diagnosis. They are mainly used for seroepidemiological studies. Recurrent enteroviruses infections during a lifetime often result in elevated serum antibody titres which obscure diagnostic changes. Significant antibody rises are therefore rarely observed in paired sera (taken at the onset of and during convalescence from disease).

A Coxsackie B virus-specific IgM test (using an IgM antibody capture technique) has been developed for rapid diagnosis. However, there is cross-reactivity between the IgM responses to different enteroviruses, including different genera of the picornaviruses, and so this test is not very specific. Prolonged presence of enterovirus-specific IgM has also been observed. Thus, in summary, the usefulness of serology for the diagnosis of enterovirus infection is limited.

Genome detection

Hybridization techniques and, more recently, reverse transcription polymerase chain reaction (RT-PCR) techniques have been applied to test for the presence of

enterovirus genomes. This approach has been very productive, particularly in diagnosing CNS infections from CSF specimens, and has become the 'gold standard' of diagnosis, surpassing viral culture. Enterovirus RNAs have also been detected in myocardial biopsies from patients with myocarditis and dilated cardiomyopathy, in muscle of people with inflammatory muscle disease and chronic fatigue syndrome, and in brain biopsies. The significance of these findings is not clear as infectious virus can rarely be isolated, and viral antigen cannot be detected. Highly conserved sequences in the 5' UTR of enterovirus genomes have allowed the design of PCR primers detecting most enterovirus RNAs. As the echovirus 22 genome is very different from that of the other enteroviruses, tailor-made primers have to be added in a multiplex RT-PCR to include these viruses, which occur particularly in neonates and infants. A modified RT-PCR procedure can differentiate between wild-type and vaccine-derived poliovirus infections.

Epidemiology of enterovirus infections

Enteroviruses are mainly transmitted by the faecal–oral route, due to the fact that viruses are shed in faeces for weeks or months after infection. Spread is particularly intense within families, usually starting from young children's primary infection. In temperate climates, there are seasonal peaks (July–September in the northern, and December–February in the southern hemisphere), whereas in subtropical/tropical climates enterovirus infections occur all the year round. The vast majority of primary human enterovirus infections occur during the first decade of life. Type-specific surveillance in several geographical regions has shown that Coxsackie viruses A9, A16, B4, and echovirus types 6, 9, 11, 19, 22, and 30 are most frequently found.

Prevention of enterovirus infections

As there are only three poliovirus types and no significant animal reservoir, it has been possible to develop very successful poliovirus vaccines. In 1954, a formalin-inactivated poliovirus vaccine (IPV) was introduced by Dr Jonas Salk in the United States, and in 1962 Dr Albert Sabin introduced a vaccine consisting of attenuated strains of the three poliovirus types which could be given orally (OPV). Protection by the live-attenuated vaccine is effected mainly at the site of entry by eliciting locally virus-specific IgAs and IgGs. Inactivated vaccine mainly elicits serum IgGs which prevent infection of the CNS and other sites distant of the port of entry by neutralization of viraemic virus. The main characteristics of IPV and OPV are summarized in Table 2.

Inactivated poliovirus vaccine

This vaccine was and is used with high acceptance rates in Scandinavia and Holland and has virtually eliminated poliomyelitis in these countries. There was a small outbreak of poliomyelitis due to poliovirus type 3 in Finland in the early 1980s, which seemed to be possible due to the fact that type 3 antibody levels in the community were comparatively low. However, the poliovirus strain isolated during the outbreak was antigenically unusual and less well neutralized by antisera to the reference strains of type 3 poliovirus. IPV is the vaccine of choice in cases of immunodeficiency.

Live attenuated poliovirus vaccine

This vaccine has a number of advantages compared to the inactivated vaccine as it:

- parallels the natural infection;
- stimulates both local secretory IgA in the pharynx and alimentary tract, and systemic circulating virus-specific IgG antibody;
- is easy to administer as an oral vaccine;
- is more cost effective.

The disadvantage is that in a few cases the attenuated vaccine strains have reverted to virulence. Since the early 1980s, all cases of polio in the United States and Europe were found to be caused by vaccine-related, that is reverted poliovirus, or were imported from endemic countries and were not indigenous, original wild-type strains. The risk of vaccine-associated poliomyelitis is between 0.5 and 3.4 cases per million of susceptible children immunized. Vaccine-related polio is mostly caused by type 2 or type 3 viruses, probably due to the fact that the number of point mutations in type 1 vaccine virus compared to wild type is much higher than in type 2 and type 3 vaccine viruses. However, this finding raises the question of whether oral vaccination should be continued. In the United States, guidelines have been developed recently which replace the oral vaccination programme by a mixed procedure, initially using inactivated vaccine, followed by booster doses of oral vaccine. For a variety of reasons many countries (United States, France, Germany) have either subscribed to the exclusive use of IPV or are likely to do so in the near future. The decision was influenced by the good progress made towards the eradication of polio due to wild-type virus.

OPV is the vaccine of choice for people travelling into poliovirus endemic areas if their immune status is unknown or in doubt. The vaccine should be given at least 2 weeks before departure.

Polio eradication and surveillance

For many years it was thought that the Sabin oral poliovaccines were ineffective in tropical countries. While many reasons were put forward, the lack of impact of polio vaccination programmes was probably due to loss of vaccine potency through failure to maintain storage at cool temperatures, and also the epidemiology of poliovirus infection. In temperate countries, poliomyelitis is seasonal with infections peaking in the summer months. A strategy of vaccination based on immunization of young children at a set age (usually a few months) is therefore able to build up a highly immune population in the winter so that transmission of the wild-type virus becomes more difficult. In tropical countries, where exposure is year round, it is a matter of chance whether a child is first naturally infected, or immunized. This was recognized by Sabin in 1960, but not acted upon until some 20 years later, when the strategy of National Immunization Days was developed in South America. This approach involves immunizing all children below a certain age in a country within a very short period, so that all susceptible children's intestinal tracts are occupied by vaccine virus and are therefore resistant to infection by the wild type. Transmission of wild-type virus is therefore broken, and the virus dies out.

WHO have pronounced the intention of eliminating poliomyelitis due to wild-type virus. The Americas have been free of polio since 1992, but vaccine-derived poliomyelitis occurred in Haiti and the Dominican Republic in 2000/1. The last case of polio in South East Asia occurred in March 1997. In 2000, the Western Pacific Region was declared polio-free by WHO, and polio is now endemic in only 30 countries (Fig. 1). The scale of the undertaking is colossal, and the progress towards eradication is extraordinary. For example in 1992 in China, all children aged 5 or less were immunized over a 1-week period. This amounts to one-quarter of the world's children. At the time of writing, virus was still known to be endemic in India and surrounding countries, countries of West and Central Africa, and of the Eastern Mediterranean region, but eradication before long is a real possibility.

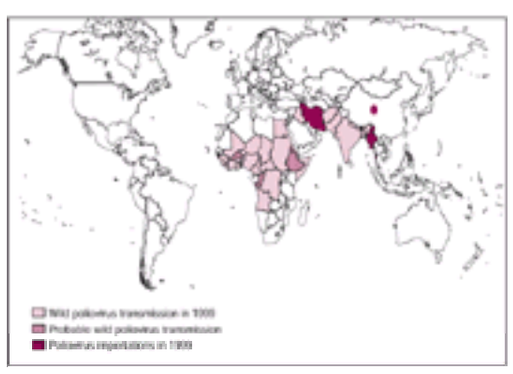

Fig. 1 Countries with known or probable wild poliovirus transmission —World Health Organization, 1999, as of March 2000. (From Centers for Disease Control (2000), with permission.)

Part of the challenge is to demonstrate that the virus has in fact been eliminated, and this depends on rigorous, effective surveillance. One approach is to obtain data on cases of acute flaccid paralysis of whatever cause, including the Guillain–Barré syndrome. All cases should be investigated to see whether they are due to poliovirus infection or not, and it is considered that the background rate in the absence of poliomyelitis should be one case per 100 000 members of the population,

providing a control for the adequacy of the surveillance scheme. Alternative approaches include the investigation of poliovirus isolates to establish whether they derived from vaccine or wild-type strains. There are possible concerns over the adequacy of either approach.

Once wild-type poliovirus has been eradicated, the only sources of the virus will be manufacturers of vaccines, laboratories holding stocks, and recipients of live attenuated vaccine. While manufacturers and laboratory workers can be required to work under high containment level conditions to avoid escape of virulent virus, vaccinees pose a particular problem. The vaccine works by establishing an infection in the recipient but the virus may adapt to the gut and eventually undergo major molecular changes to improve its fitness. In principle, such viruses could spread to others forming a focus for a return of poliovirus infections and poliomyelitis. In practice, the vaccine virus seems to be poorly transmissible compared to the wild type. In countries such as Cuba where it has been given only in the early part of the year as a matter of policy, virus is not detectable after 6 months. Thus, it might be possible to stop vaccinating with no further precautions, as the vaccine strain of poliovirus will die out more rapidly than susceptible individuals will accumulate to provide a population to maintain it. However, people with B cell immunodeficiency can remain infected but apparently healthy for up to 15 years. During this time the virus may adapt to an extent that neurotropism is regained, and an unvaccinated population will again be highly susceptible. The numbers or geographical distribution of such long-term excretors are unknown. One possible approach to the problem would be to use inactivated vaccine for some unspecified period. The fact that serious consideration has to be given to how to deal with the cessation of vaccination is a tribute to the extraordinary progress which has been made towards polio eradication.

Further reading

Centers for Disease Control (2000). Progress toward global poliomyelitis eradication, 1999. *Morbidity and Mortality Weekly Report* **49**, 349–54.

Joint Working Group of the Royal Colleges of Physicians, Psychiatrists and General Practitioners (1997). *Chronic fatigue syndrome*, pp. 58. Royal College of Physicians, Publication Unit, London.

King AMQ, Brown F, Christian P *et al*. (2000). *Picornaviridae.* In: van Regenmortel MHV, Fauquet CM, Bishop DHL, *et al*. eds. *Virus taxonomy. classification and nomenclature of viruses. Seventh report of the International Committee on Taxonomy of Viruses*, pp, 657–78. Virology Division, International Union of Microbiological Societies. Academic Press, San Diego.

Melnick JL (1996). Enteroviruses: polioviruses, coxsackieviruses, echoviruses, and newer enteroviruses. In: Fields BN, Knipe DM, Howley PM *et al*., eds. *Fields virology,* 3rd edn, pp. 655–712. Lippincott-Raven, Philadelphia.

Mendelsohn C, Wimmer R, Racaniello VR (1989). Cellular receptor for poliovirus: molecular cloning, nucleotide sequence and expression of a new member of the immunoglobulin superfamily. *Cell* **56**, 855–65.

Minor PD (1990). Antigenic structure of picornaviruses. *Current Topics in Microbiology and Immunology* **161**, 122–54.

Minor PD (1996). Poliovirus. In: Nathanson N, Ahmed R, Gonzalez-Scarano F, *et al*., eds. *Viral pathogenesis,* pp. 555–74. Lippincott-Raven, Philadelphia.

Minor P (1999). Picornaviruses. In: Mahy BW, Collier IL, eds. *Topley and Wilson's microbiology and microbial infections*, Vol. 1: Virology, 9th edn, pp. 485–509. Arnold, London.

Racaniello VR and Baltimore D (1981). Cloned poliovirus complementary DNA is infectious in mammalian cells. *Science* **214**, 916–19.

Yamashita T *et al*. (2000). Application of a reverse transcription-PCR for identification and differentiation of Aichi virus, a new member of the picornavirus family associated with gastroenteritis in humans. *Journal of Clinical Microbiology* **38**, 2955–61.

7.10.8 Virus infections causing diarrhoea and vomiting

Ulrich Desselberger

Introduction

Acute gastroenteritis and vomiting in humans is a well-characterized clinical entity caused by several different agents (viruses, bacteria, parasites, etc.). Viral gastroenteritis is a global problem, particularly in infants and young children.

Many viruses are found in the gut but not all of them produce acute gastroenteritis (Table 1). Viral infections normally associated with gastroenteritis are caused by: rotaviruses; enteric adenoviruses; small, round structured viruses (**SRSVs**) and classic human caliciviruses; and astroviruses. Other viruses found in the gastrointestinal tract (enteroviruses, reoviruses, non-group F adenoviruses, toroviruses, coronaviruses, parvo-viruses) are not regularly associated with diarrhoeal disease in humans. Finally, there are viruses found in the gut of immunosuppressed patients (most often those with human immunodeficiency virus (**HIV**) infection), including herpes simplex virus (**HSV**), cytomegalovirus (**CMV**), and picobirnaviruses. HIV itself can also infect the gut directly.

Only the major virus groups regularly causing gastroenteritis in man are described here separately. Clinical symptoms, diagnosis, treatment, epidemiology, and vaccine development are reviewed under common headings.

Rotaviruses

Structure

Rotaviruses are the major cause of infantile gastroenteritis worldwide and also of acute diarrhoea in the young of many mammalian species. They are members of the Reoviridae family, with a genome of 11 segments of double-stranded RNA encoding six structural viral proteins (VP1–4, VP6, and VP7) and six non-structural proteins (NSP1–NSP6). The structural proteins VP1–3 are located in the inner or core layer, VP6 forms an inner shell or intermediate layer, and VP7, and VP4 are components of the outer shell or outer layer (VP4 protrudes as spikes). Thus, a double-shelled/triple-layered particle, 75 nm in diameter, appears in a characteristic form on electron micrographs (Fig. 1), the name of the virus being derived from *rota* (Latin = wheel).

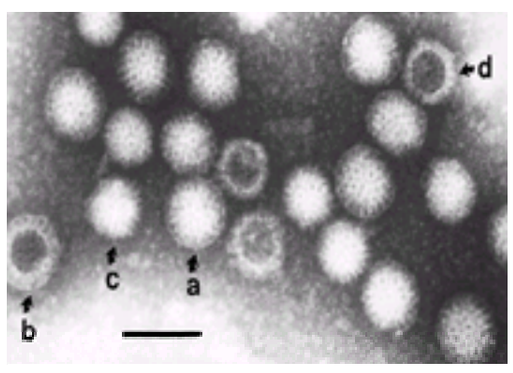

Fig. 1 Rotavirus particles in the faeces of a child admitted to hospital with acute gastroenteritis. Negative staining with aqueous 2 per cent potassium phosphotungstate, pH 7.0. The scale bar represents 100 nm. Four different morphologies of particles are shown: (a) double-shelled (ds) particles containing RNA; (b) ds particle without RNA (empty; core penetrated with stain); (c) single-shelled (ss) particle containing RNA; (d) single-shelled empty particle. (Figure by courtesy of M. Jenkins, Regional Virus Laboratory, East Birmingham Hospital). (Taken from U. Desselberger, Reoviruses. In: *Medical Microbiology,* 14th edition, (eds. D. Greenwood, R. Slack, J. Peutherer), page 620. Churchill Livingstone, Edinburgh, 1992, with permission of the publisher.)

Classification

Rotaviruses are classified according to the immunological reactivities and genomic sequences of three of their structural components.

1. Specific epitopes on the inner-shell protein VP6 allow five groups (A–E) to be distinguished, and two more groups (F, G) probably exist. Within group A rotaviruses, there are at least four subgroups. Group A rotaviruses constitute the vast majority of human infections.
2. Both surface proteins, VP4 and VP7, elicit neutralizing antibodies and thus confer type-specificity. A dual-type classification system has been devised for group A rotaviruses, which differentiates G types (VP7-specific, G for glycoprotein) and P types (VP4-specific, P for protease-sensitive protein)—for example, G1P1A[8] is G serotype and genotype 1, P serotype 1A, P genotype 8. At least 11 G types and 9 P types have been found in humans (see reviews by Estes 1996 and Desselberger 2000). As group A rotaviruses reassort readily in doubly infected cells, various combinations of VP4 and VP7 types occur in natural isolates.

Replication

Rotaviruses replicate in the mature epithelial cells at the tips of the villi of the small intestine. Viruses enter cells either by receptor-mediated endocytosis or directly, and single-shelled subviral particles (devoid of VP4 and VP7) produce and protrude large numbers of mRNAs from all 11 segments into the cytoplasm. Viral proteins are synthesized from mRNAs, and morphogenesis proceeds in a complex fashion. Mature rotavirus particles are released from cells by lysis, resulting in very high concentrations of up to 10^{11} virus particles per ml of faeces at the peak of the acute diarrhoea.

Pathogenesis

The diarrhoea arises by cellular necrosis and atrophy of the epithelium, leading to a reduction in the breakdown and absorption of carbohydrates and to increased osmotic pressure in the gut lumen. The villous damage is repaired by cells emerging and differentiating from the crypts of the gut epithelium, which show a reactive hyperplasia. This is accompanied by increased secretion, which also contributes to the diarrhoea. Several viral proteins have been shown to be determinants of the pathogenicity of rotaviruses and VP4 is of major significance. Most recently, the non-structural protein NSP4 (= VP10) has been characterized as the first viral enterotoxin.

Immune response

A serotype-specific humoral immune response is elicited after neonatal or primary rotavirus infection. However, during the first 2 years of life children are repeatedly infected with rotaviruses, leading to multiple serotype-specific, but also partially heterotypic, protection. The presence of rotavirus-specific secretory IgA coproantibodies seems to correlate best with protection against disease, although the exact correlates remain to be determined. There are also rotavirus-specific cytotoxic T-cell responses whose exact role in immunity is unknown.

Enteric adenoviruses

Structure and classification

Adenoviruses are non-enveloped icosahedral viruses possessing a genome of linear double-stranded DNA approximately 35 000 bp in size. Their capsid is between 70 and 80 nm in diameter and consists of 240 hexons and 12 pentons that stand out as projecting fibres at the corners of the icosahedral virus particle. Human adenoviruses occur in 51 distinct serotypes, ordered in six different subgroups (A–F). Those adenoviruses regularly associated with gastroenteritis are classified as subgroup F, serotypes 40 and 41. Adenoviruses of different groups (causing respiratory tract infections) are also found in the gut, but are not regularly associated with diarrhoea.

Replication

Adenoviruses attach to susceptible cells via the fibre proteins, and enter via receptor-mediated endocytosis. Phased early and late gene transcription of the viral DNA in the cellular nucleus is followed by translation and morphogenesis in the cytoplasm, and numerous particles are released after cell death. The early protein E1A is a potent blocker of apoptosis and of interferon-a and -b expression. Late adenovirus gene expression blocks cellular DNA expression. Some adenoviruses seem to decrease the expression of MHC class 1 antigens on the surface of infected cells, thus reducing susceptibility to adenovirus-specific cytotoxic T cells. There is a serotype-specific humoral immune response providing homotypic protection.

Small, round structured viruses (SRSV)

Structure and classification

These viruses were first recognized as the cause of gastroenteritis during outbreaks in Norwalk, Ohio, in the late 1960s. Norwalk virus (**NV**) particles are spherical and measure 27 to 35 nm in diameter. NV and Norwalk-like viruses (**NLVs**) are all members of the Caliciviridae family. Their 7.7-kb genome consists of single-stranded RNA of positive polarity. Cup-shaped depressions on the surface of virions have given the name to this viral family (Latin *calix* = goblet, cup). Phylogenetic trees of full-length sequences of caliciviral cDNA demonstrate at least three genogroups (Norwalk virus representing genogroup 1; Lorsdale virus, genogroup 2; and Manchester virus, genogroup 3). The SRSVs and typical caliciviruses that infect humans are shown in Table 2. Genogroup 3 viruses have been termed 'classical' caliciviruses as their structure is better preserved on electron microscopy, and the genomes of some show greater homology with several animal caliciviruses which are possible sources of human infection.

Replication

Details of the replication of human caliciviruses can only be deduced from those of animal caliciviruses as there is no reproducible *in vitro* cell-culture system for the human SRSVs. The viruses seem to interact with species-specific receptors and a single protein precursor is co- and post-translationally cleaved in a way similar to that observed in enteroviruses.

Immune response

Although SRSV infections elicit human immune responses, they do not seem to give full protection against subsequent infection. On the contrary, higher pre-existing antibody levels seem to predispose to more severe illness upon reinfection.

Astroviruses

Structure and classification

Astroviruses are members of the newly defined family Astroviridae. They possess a 6.8-kb genome of single-stranded RNA of positive polarity. So far, eight different serotypes have been distinguished, which correlate well with major differences in genome sequences (that is, genotypes).

Replication

Human astroviruses grow well in particular cell cultures. After viral absorption to unidentified cellular receptors and uncoating in the cytoplasm, full-length and subgenomic RNAs are made. These direct the production of protein precursors which are post-translationally cleaved. Replication takes place purely in the cytoplasm.

Illness

The onset of acute viral gastroenteritis follows a short incubation period of 1 to 2 days. It is sudden, with watery diarrhoea lasting between 4 and 7 days, vomiting, and varying degrees of dehydration. Over one-third of children with rotavirus infection have a fever of more than 39 °C. Fewer have a high fever after infection with SRSVs. The duration of diarrhoea after infection with SRSVs is as a rule shorter (1–2 days) than after infection with rotaviruses or enteric adenoviruses (4–7 days). Disease due to SRSV infection may be accompanied by abdominal cramps, headache, and myalgia. In rotavirus infection all degrees of severity are seen. Inapparent infections are not infrequent, particularly in neonates where the infection is caused by so-called 'nursery strains'. It is not clear whether the asymptomatic nature of rotavirus infection in neonates is due to infection with particular strains or depends on the presence of maternal antibodies that provide partial protection. Rotavirus infections are frequently accompanied by respiratory symptoms; but there is no strong evidence that rotavirus replicates in the respiratory tract. In immunodeficient children, chronic gut infections with rotaviruses, adenoviruses, and astroviruses have been observed, accompanied by virus shedding over weeks and even months.

Diagnosis

The diagnosis of rotavirus, astrovirus, and enteric adenovirus infections is relatively easy as large numbers of particles are shed during the acute phase of the illness. In contrast, SRSVs replicate for a shorter period and are shed at lower concentrations. Diagnosis is by electron microscopy of negatively stained specimen

suspensions ('Catch-all method'), by passive particle agglutination tests, virus-specific enzyme-linked immunosorbent assays (**ELISAs**), and more recently by viral genome detection using the polymerase chain reaction (**PCR**) (for adenoviruses) and reverse transcription-PCR (**RT-PCR**) (for rotaviruses, caliciviruses, and astroviruses). The morphological appearances of the main viruses pathogenic for humans are shown in Fig. 2. PCRs are extremely sensitive diagnostic tools, allowing both viral detection and typing. Aliquots of PCR amplicons can also be sequenced and the information used to establish phylogenetic trees. Such trees are becoming increasingly important not only for virus classification but also for epidemiological studies and surveillance (see below).

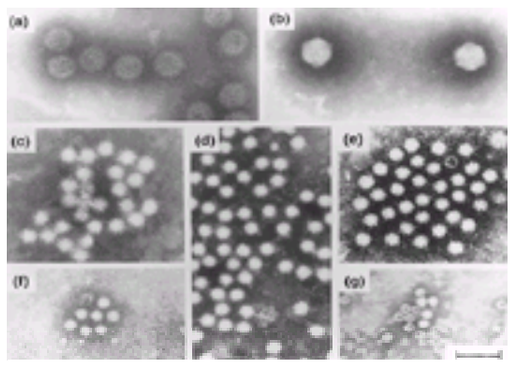

Fig. 2 Electron micrographs of (a) rotavirus, (b) enteric adenovirus, (c) SRSV, (d) calicivirus, (e) astrovirus, (f) enterovirus, and (g) parvovirus. Negative staining with 3 per cent phosphotungstate, pH 6.3; bar represents 100 nm (Figures by courtesy of Dr J. Kurtz, Oxford Public Health Laboratory (astroviruses) and Dr J. Gray, Clinical Microbiology and Public Health Laboratory Cambridge (all other viruses)). Reproduced from *Principles and practice of clinical virology,* 4th edn (eds. A. Zuckerman, J. Banatvala, J. Pattison), p. 236. J. Wiley and Sons, Chichester, 2000, with permission of the publisher.

Treatment

Treatment is mainly by oral rehydration or, in more severe cases, intravenous rehydration. In severe rotavirus infections, treatment with oral immunoglobulins has shown to affect the duration of diarrhoea and virus shedding. This is not, however, a routine treatment. Otherwise treatment is symptomatic, but the use of antimobility drugs (codeine phosphate, diphenoxylate, coperamide) in children is not advised. Specific antiviral agents have been tested in animal models of rotavirus infections but are not used for human treatment.

Epidemiology

Rotaviruses

Rotavirus infections occur endemically worldwide and cause over 800 000 deaths annually in children below the age of 2 years, mainly in developing countries. Therefore development of vaccine candidates has been a major goal since the early 1980s (see below). The epidemiology of rotaviruses is complex. Besides children, elderly patients and patients with immunodeficiencies can be affected. There is a strict winter peak of rotavirus infections in temperate climates, but infections occur year round in tropical and subtropical regions. Transmission is by the faeco-oral route. Nosocomial infections on infant hospital wards occur and are difficult to eliminate. Group A rotaviruses of different G and P types are found to cocirculate in various populations within the same geographical location, and the relative incidence of different types changes over time. Various surveys have shown that usually more than 90 per cent of cocirculating strains in temperate climates are types G1 to G4 and occur in combination with different P types as types G1P1A[8], G2P1B[4], G3P1A[8], and G4P1A[8]. Other G types may also be represented, seen particularly in tropical and subtropical areas but increasingly in temperate climates. For instance, G9 strains have caused outbreaks with increasing frequency in the United States and Europe. Most mammalian, as well as avian species, harbour a large diversity of rotaviruses and may act as a reservoir for human infections. An animal source is suspected for many of the more unusual human group A rotavirus isolates, and possibly for group B rotavirus isolates. The latter caused outbreaks in children and adults in China during the 1980s and have recently been isolated from patients with diarrhoea in Calcutta. Group C rotaviruses are associated with small outbreaks in humans.

Small, round structured viruses

Age-related seroprevalence studies of SRSVs have recently shown that infection is much more frequent and occurs from younger ages onwards than previously thought. Approximately 50 per cent of infants have been infected by the age of 2 years. The rate of inapparent infection is high, particularly in the young. In contrast to rotavirus infections, SRSVs cause outbreaks of acute gastroenteritis, mostly due to contamination of food or water. Contaminated oysters and green salads are often implicated as sources of infection. Outbreaks occur in older children and adults in recreational camps, hospitals, nursing homes, schools, cafeterias, hotels, cruise ships, at banquets, etc. SRSV outbreaks occur worldwide throughout the year, in contrast to the regular winter peaks of rotavirus infections in temperate climates. The viruses are highly infectious and spread rapidly. Transmission is spread by the faeco-oral route and also by projectile vomiting, scattering viruses into the environment by aerosol. At least seven different genotypes cocirculate, which have been confirmed as serotypes. Type 1 and 2 viruses are most frequent.

Astroviruses

Infections with astroviruses occur in infants and the elderly as endemic infections, but they can also cause food-borne outbreaks of diarrhoea. There are at least eight genotypes, correlating well with known serotypes which cocirculate. Serotype 1 is most frequently found, followed by serotypes 2 to 4 at intermediate and serotypes 5 to 7 at low frequencies. Seroprevalence studies have indicated that infection by more than one serotype is not unusual.

Vaccine development

Vaccines have been found to be the best individual and also population-based tools to restrict infection with epidemic viruses. Of the gastroenteritis-inducing viruses, vaccine development has only been intensely directed towards rotaviruses. After many trials with variable success, a live-attenuated, rhesus rotavirus (**RRV**)-based, human reassortant vaccine eliciting immunity to human rotavirus strains G1 to G4 has recently been found to confer significant protection (70–80 per cent) from severe disease including dehydration, whereas protection from infection alone was only moderate (40–50 per cent). This vaccine was recommended by the Advisory Committee on Immunization Practices (**ACIP**) in the United States in 1998. However, after 1.5 million doses had been used, the rare complication of gut intussusception was found to be apparently significantly associated with vaccination. In 1999 the ACIP withdrew the recommendation, and the vaccine has been taken off the market by the manufacturer. Studies of the epidemiological findings and possible mechanisms of pathogenesis are underway.

As a result, other approaches to immunization against rotavirus infection have gained interest, such as the use of virus-like particles obtained from baculovirus-recombinant coexpressed rotavirus proteins, enhancement of rotavirus immunogenicity by microencapsidation, DNA-based candidate vaccines, and possibly 'edible vaccines'.

No vaccines against other viruses causing gastroenteritis in humans have been developed so far. For NV-like viruses this is unlikely to happen, as long-term immunity does not usually seem to follow natural infection.

Outbreak control

Nosocomial rotavirus outbreaks among paediatric populations (on hospital wards and in day-care centres) are common. There have been numerous reports of outbreaks of diarrhoea and vomiting occurring in adults and children due to infections with Norwalk-like viruses, acquired from banquets, travel on cruise ships, cafeterias, schools, hotels, fast-food restaurants, etc.

Outbreak control measures should focus on the interruption of person-to-person transmission and the removal of common sources of infection (food, water, etc.) in

conjunction with measures to improve environmental hygiene (by food-handlers, etc.).

Further reading

Ball JM, *et al.* (1996). Age-dependent diarrhoea induced by a rotaviral nonstructural glycoprotein. *Science* **272**, 101–4.

Desselberger U (1992). Reoviruses. In: Greenwood D, Slack R, Peutherer J, eds. *Medical microbiology*, 14th edn, pp 619–33. Churchill-Livingstone, Edinburgh.

Desselberger U (2000). Viruses causing gastroenteritis. In: Zuckerman A, Banatvala J, Pattison J, eds. *Principles and practice of clinical virology*, 4th edn, pp 235–52. Wiley, Chichester.

Estes MK (1996). Rotaviruses and their replication. In: Fields BN, *et al.*, eds. *Fields virology*, 3rd edn, pp 1625–55. Lippincott-Raven, Philadelphia.

Kapikian AZ, Estes MK, Chanock RM (1996). Norwalk group of viruses. In: Fields BN, *et al.*, eds. *Fields virology*, 3rd edn, pp 783–810. Lippincott-Raven, Philadelphia.

Matsui, SM, Greenberg HB (1996). Astroviruses. In: Fields BN, *et al.*, eds. *Fields virology*, 3rd edn, pp 811–24. Lippincott-Raven, Philadelphia.

Offit PA (1994). Rotaviruses. Immunological determinants of protection against infection and disease. *Advances in Virus Research* **44**, 161–202.

Shenk T (1996). Adenoviridae: the viruses and their replication. In: Fields BN, *et al.*, eds. *Fields virology*, 3rd edn, pp 2111–2148. Lippincott-Raven, Philadelphia.

7.10.9 Rhabdoviruses: rabies and rabies-related viruses

M. J. Warrell and D. A. Warrell

Virology

The Rhabdoviridae are a family of more than 100 rod- or bullet-shaped RNA viruses found in vertebrates, insects, and plants (Fig. 1). Two genera infect animals: *Vesiculovirus* and *Lyssavirus*. Vesicular stomatitis virus is a *Vesiculovirus* of cattle and horses, which occasionally causes an influenza-like illness in farmers or laboratory workers. The genus *Lyssavirus* comprises seven genotypes: rabies and six genotypes of rabies-related viruses.

Fig. 1 Rhabdoviruses. Virion of rabies virus. (Electron micrograph by courtesy of Mr C. J. Smale and Dr Joan Crick.)

The rabies virion is approximately 180 × 75 nm. Its core is a single strand of negative non-segmented RNA, associated with a nucleoprotein, a phosphoprotein, and an RNA polymerase to form a helical ribonucleoprotein complex (RNP). This is enveloped in a matrix protein covered by a coat of glycoprotein (G) and host cell-derived lipid. The G protein forms numerous spikes or knobs, 10 nm long, and its composition determines viral virulence.

The virus is readily inactivated by ultraviolet light, drying, boiling, most organic lipid solvents including more than 45 per cent ethanol, soap solution, detergents, hypochlorite, and glutaraldehyde solutions.

Typing by means of monoclonal antibodies or genetic sequencing techniques allows the identification of strains of rabies and rabies-related viruses from different geographical areas and vector species, revealing the diversity of rabies virus strains.

Epidemiology

Rabies is a zoonosis that remains endemic in most parts of the world (Fig. 2). Currently, the following countries are rabies free: Iceland, Norway, Sweden, Finland, Switzerland, Portugal, Italy, Greece, Cyprus and other Mediterranean islands, Singapore, Sabah, Sarawak, Bali, New Guinea, New Zealand, Antarctica, Oceania, Hong Kong islands (but not the New Territories), Japan, South Korea, Taiwan, and Caribbean islands with the notable exceptions of Cuba, the Dominican Republic, Grenada, Haiti, Trinidad, and Tobago. Some other countries have no indigenous rabies but infected animals cross land borders.

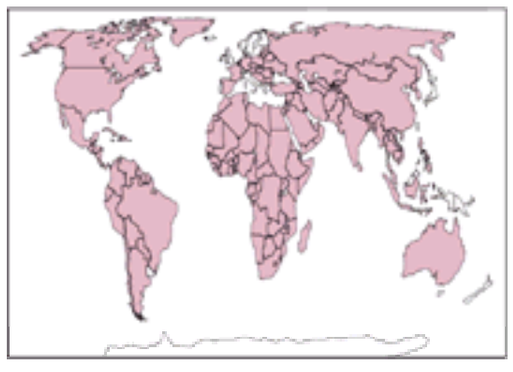

Fig. 2 Global distribution of rabies. Rabies-free areas shown in white.

Primarily an infection of wild mammals, rabies is spread by bites and rarely by inhalation of aerosols in bat caves and by ingestion of infected prey. The ecology of rabies virus can be divided into urban and sylvatic phases, which overlap to a varying extent in different countries.

In a particular area, transmission in the sylvatic phase tends to occur predominantly within a single species in separate ecological compartments. Each vector has a

separate virus strain and a distinctive method of transmission. The wild-mammal reservoir species varies in different geographical areas: in the United States, striped skunks (*Mephitis mephitis*) and to a lesser extent spotted skunks (*Spilogale putorius*) in the central States and California; raccoons in the east; the grey fox (*Urocyon cineroargenteus*) in central areas and red fox (*Vulpes vulpes*) in the east; coyotes in the south; and in the arctic, the fox *Alopex lagopus*.

Bat rabies in the Americas is all due to genotype 1 virus, whereas in the rest of the world bats have rabies-related lyssaviruses. In North America insectivorous bats are the vectors, including the Mexican free-tailed bat (*Tadarida brasiliensis mexicana*), the red bat (*Lasiurus borealis*), the big brown bat (*Eptesicus fuscus*), and the silver-haired bat (*Lasionycteris noctivagans*), whose virus is the main cause of human rabies infections in the United States (see below). Bat infection has been found in every state except Alaska and Hawaii.

The three species of true vampire bats, *Desmodus rotundus*, *Diaemus youngi*, and *Diphylla ecaudata* (Desmodontinae), occur from sea level to over 3500 m but usually below 1500 m, only in Mexico, Central and South America, and some Caribbean Islands (Fig. 3). The common vampire bat, *D. rotundus* is the main vector of vampire bat rabies in Trinidad, Mexico, and Central and South America. Carnivorous bats of the family Megadermatidae, such as the Indian 'vampire' (*Megaderma lyra*), are usually responsible for the myth that vampires occur outside this area. In Latin America vampire bat-transmitted paralytic rabies (derriengue) has locally serious economic consequences. In Brazil 50 000 head of cattle are estimated to die annually.

Fig. 3 Distribution of the three species of true vampire bats (Desmodontinae).

In Grenada and Puerto Rico mongooses (*Herpestes auropunctatus)* are vectors of sylvatic rabies; in most of Africa and Asia, wolves, jackals, and small carnivores of the families Mustelidae and Viverridae (e.g. the yellow mongoose *Cynictis penicillata* in South Africa and the palm civet *Paradoxurus hermaphroditus* in Indonesia); and in Europe, foxes, wolves, raccoon dogs (*Nyctereutes procyonoides*), and insectivorous bats (see rabies-related viruses). There are reports of rabies virus being isolated from wild rodents in many countries including the Russian Federation, Germany, Egypt, Nigeria, Thailand, and the United States, but the significance of this finding is uncertain.

Humans are occasionally infected by wild mammals, but domestic dogs and cats, the principal vectors of urban rabies, are responsible for more than 90 per cent of human cases worldwide. Domestic dogs are the principal reservoir in many parts of Africa and Asia, and in urban areas elsewhere. Rabies control programmes can reduce the risk of rabies in domestic animals to such an extent that wild animals, for example, bats in the United States, become the principal vectors of infection to humans.

Cyclical epizootics of rabies, such as the fox epizootic in Europe, result from an uncontrolled increase in the population of the key reservoir species. This epizootic started in Poland at the end of the Second World War. Initially it advanced at a rate of about 40 km a year across France, but recently has retreated. Although the fox is one of the most susceptible species to rabies, about 3 per cent of animals survive the infection and become immune. In the Caribbean island of Grenada, almost half of the mongooses have serum neutralizing antibody against rabies. Seropositive raccoons, bats, and very occasionally dogs have also been found.

Incidence of human rabies

The true incidence of human rabies throughout the world is not reflected in official figures, such as those reported by the World Health Organization. In 1998, in India alone, the estimated mortality was 30 000, in Bangladesh, it was 2000 ($1.6/10^5$ population) and in Sri Lanka, 110. The true mortalities were probably considerably higher. High mortalities also occur in Nepal, Pakistan, and Ethiopia, but there are very few data from Africa. In the United States in 10 years since 1990, 27 human cases occurred, seven of whom were infected by dogs (two from Asia). Twenty (74 per cent) patients were infected by bats, and 15 (75 per cent) of the bat infections were due to a rabies strain associated with silver-haired (*Lasionycteris noctivagans*) and pipistrelle bats (*Pipistrellus subflavus*). The Russian Federation reported 11 deaths from rabies in 1999. Rabies was eradicated from Britain by 1903 but in November 2002 a man died of European bat lyssavirus infection (see below) in Scotland.

Transmission

Virus can penetrate broken skin and intact mucosa. Humans are usually infected when virus-laden saliva is inoculated through the skin by the bite of a rabid dog or other mammal (Fig. 4). Saliva from a rabid animal can infect if the skin is already broken, by the animal's claws for example. In the United States infective contact with bats may be unnoticed: only one of 20 patients had a history of a bat bite. Animals can be infected through the gastrointestinal tract, but there is no evidence that this happens in humans.

Fig. 4 Child bitten on the face by a rabid dog. This wound carries a high risk of rabies with a short incubation period. (Copyright D.A. Warrell.)

Inhalation of aerosolized virus created by bats' infected nasal secretions may be an important method of transmission among cave-dwelling bats. In Texas, two men died of rabies after visiting caves inhabited by millions of Mexican free-tailed bats, many of which were rabid. Two laboratory workers in the United States developed rabies after inhaling fixed strains of rabies virus during the preparation of vaccines. The accidental use of vaccine in which the virus was not inactivated has led to fixed virus rabies (*rage de laboratoire*): for example, in Fortaleza, Brazil in 1960.

Transmission of rabies from one person to another has been proved in six patients who received infected corneal grafts (Fig. 5). The donors had died of undiagnosed neurological diseases. Twenty-two to 39 days after transplantation, the recipients developed retro-orbital headache on the side of the graft and died of rabies. In a seventh recipient, in Morocco, rabies was prevented by vigorous postexposure prophylaxis. Other infections spread by corneal grafts are Creutzfeldt–Jacob disease and cryptococcosis. Considering that the saliva, respiratory secretions, and tears of patients with rabies contain virus it is surprising that the disease has not been

spread to relatives and nurses.

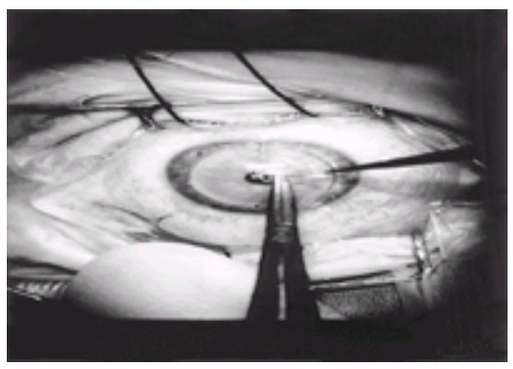

Fig. 5 Corneal transplant graft (by courtesy of Professor A. Bron).

Transplacental infection has been observed in animals but only once reported in humans. A number of women with rabies encephalitis have given birth to healthy babies. The transmission of rabies from mother to suckling infant via the breast milk has been suspected in at least one human case and is well known in animals.

Pathogenesis

The mechanism by which the highly neurotropic rabies virus enters the nervous system, travels into the brain, and out again to many organs is intriguing. The virus may replicate locally in muscle cells or attach directly to the nerve endings. Its preferential attachment to nicotinic acetylcholine receptors at motor end-plates is blocked by a-bungarotoxin, whose structure has a homologous sequence with rabies glycoprotein. Once inside peripheral nerves, virus travels in a retrograde direction within the axoplasm. This progression can be blocked experimentally by local anaesthetics, metabolic inhibitors, and nerve section. No virions are detectable in the axons. A current hypothesis is that the virus travels in an incomplete form, perhaps as naked ribonucleoprotein complexes, at this stage. Rabies virus is experimentally inaccessible to antibodies while inside the peripheral nerve.

On reaching the central nervous system, there is massive viral replication on membranes within neurones and direct transmission of virus from cell to cell occurs across synaptic junctions. There is neuronal dysfunction without significant pathological change. Centrifugal spread of virus from the central nervous system in the axoplasm of somatic and autonomic efferent nerves deposits virus in many tissues including skeletal and cardiac muscle, adrenal medulla, where infection may be clinically significant, and in kidney, retina, cornea, pancreas, and nerve twiglets in the hair follicles (see below). At this stage, productive viral replication occurs, with budding from outer cell membranes in the salivary and other glands. This is the means of the further transmission of rabies by bites to other mammals. In humans, virus is also delivered to the lacrimal glands, taste buds, and respiratory tract. Viraemia has very rarely been detected in animals and is not thought to be involved in pathogenesis or spread.

Immunology

Immunological response to rabies infection in humans

There is no detectable immune response until encephalitic symptoms develop, suggesting that rabies virus avoids or suppresses the immune system. Neutralizing and other antibodies become detectable in serum after about 7 days and in cerebrospinal fluid a little later. They may rise to high levels in patients whose lives are prolonged by intensive care. A small amount of rabies-specific IgM is sometimes detectable, but is not useful as a means of diagnosis.

There is little evidence of a lymphocyte-mediated immune response to rabies encephalitis. A pleocytosis appears in only 60 per cent of patients, with a mean leucocyte count of 75×10^3/mm. Peripheral-blood lymphocyte transformation has been shown in a few patients with furious rabies, but not in those with paralytic disease. Experimentally, in fatal rabies there is suppression of the cytotoxic T-lymphocyte response to unrelated viral antigens and a T-cell response is associated with survival in mice.

Interferon is induced by rabies infection, but appears to be at a very low level in human patients. In animals, latent infections can be reactivated by corticosteroids and stress. This provides a possible explanation for occasional reports of long incubation periods.

Rabies nucleoprotein is a weak superantigen, as it directly stimulates CD4 type 2 T lymphocytes bearing Vb8 T-cell receptors without intermediary antigen-presenting cells. The development of paralytic rabies in mice is dependent on the presence of the specific T-cell receptors. In humans the effect might be an inefficent polyclonal antibody response, preventing a specific immune response and eventually resulting in anergy.

Immunological response to rabies vaccination

Neutralizing antibody may be detectable as early as 7 days after the start of primary immunization. The surface glycoprotein of the virus induces neutralizing antibody, which protects against subsequent challenge with rabies virus in animals. The titre of these antibodies shows a better correlation with protection than any other measure of immune response.

Rabies nucleoprotein antigens also stimulate protective immunity in animals, through non-neutralizing antibody, helper T lymphocytes, and interferon-a induction. This protection is effective against a variety of rabies and rabies-related strains, unlike the glycoprotein-mediated immunity.

In human vaccinees, peripheral-blood lymphocyte transformation occurs in response to a variety of rabies and rabies-related virus antigens. The role of helper and cytotoxic T lymphocytes in protection against disease is unclear.

Neutralizing antibody is undoubtedly protective in the early stages after inoculation of virus, but in experimental animals the 'early death phenomenon' is due to a very low level of rabies antibody which accelerates the terminal phase of the encephalitis.

A low level of interferon may be induced briefly after the first dose of rabies vaccine. In animals, interferon induced by viruses or synthetic inducers, or the administration of exogenous interferon, was effective postexposure prophylaxis against rabies.

Rabies in animals

Any warm-blooded animal (mammal or bird) can be infected with rabies. In dogs the incubation period ranges from 5 days to 14 months, but is usually between 3 and 12 weeks. The first symptom, as in many humans, is intense irritation at the site of the infection. Despite the popular idea of the 'mad' rabid dog, only 25 per cent develop furious rabies. Clinical features include an early and marked change in behaviour, dysphagia, ptosis, altered bark, paralysis of the jaw, neck, and hind limbs (Fig. 6), hypersalivation, congested conjunctivas, pruritus, shivering, trembling, snapping at imaginary objects, pica, and extreme restlessness, causing the animal to wander miles from home. Dogs with furious rabies attack inanimate objects, often seriously injuring their mouths in the process. Virus may be excreted in the saliva 3 days before symptoms appear, and the animal usually dies within the next 7 days. This is the basis for the traditional 10-day observation period for dogs that have bitten humans. There are rare reports, from India, Ethiopia, and elsewhere, of chronic excretion of virus in the saliva of apparently healthy dogs. *Oulou fato* is a clinical variant of canine rabies with apparently reduced virulence for humans, seen in West Africa 50 years ago.

Fig. 6 Dog with paralytic rabies showing paralysis of the limbs and hypersalivation. (Copyright D.A. Warrell.)

Rabid foxes lose their fear of humans and the majority develop the paralytic form of the disease. An extreme degree of furious rabies is seen in 75 per cent of infected cats. Cattle usually develop paralytic symptoms, with dysphagia, hypersalivation, groaning, trembling, colic, diarrhoea, tenesmus, and rectal prolapse. Most other domestic ungulates develop paralytic symptoms. Horses often show furious features with sexual excitement. Most wild animals, like foxes, lose their fear of humans and may appear tame. Rabid skunks, raccoons, badgers, martens, and mongooses may become very aggressive. Dysphagia and inability to drink is common in rabid animals, but they do not exhibit hydrophobia.

Clinical features in humans

The incubation period ranges from 4 days to many years, but it is between 20 and 90 days in three-quarters of cases. It tends to be shorter after bites on the face (average 35 days) than after those on the limbs (average 52 days).

Prodromal symptoms

In many patients, the first symptom is itching, pain, or paraesthesia at the site of the healed bite wound (Fig. 7). Non-specific prodromal symptoms include fever, chills, malaise, weakness, tiredness, headache, photophobia, myalgia, anxiety, depression, irritability, and symptoms of upper respiratory tract and gastrointestinal infections. Subsequently, symptoms of either furious or paralytic rabies will develop, depending on whether the spinal cord or brain are predominantly infected.

Fig. 7 This man developed intense itching in the left leg, provoking scratching and excoriation, 6 weeks after being bitten in that limb by a mad dog. He died with furious rabies a few days later (by courtesy of Professor Sornchai Looaresuwan).

Furious rabies

Furious rabies is the more commonly diagnosed form. Most patients have the diagnostic symptom of hydrophobia: a combination of inspiratory muscle spasm, with or without painful laryngopharyngeal spasm, associated with terror (Fig. 8). Initially provoked by attempts to drink water, this reflex can be excited by a variety of stimuli including a draught of air ('aerophobia'), water splashed on the skin, irritation of the respiratory tract, or ultimately, the sight, sound, or even mention of water. The inspiratory spasm is violent and jerky. The neck and back are extended, the arms thrown up, and the episode may end in generalized convulsions with cardiac or respiratory arrest.

Fig. 8 Hydrophobic spasm in a 14-year-old Nigerian boy with furious rabies. Note the violent contraction of inspiratory muscles: sternomastoids and diaphragm (depressing xiphisternum). (Copyright D.A. Warrell.)

Patients experience hyperaesthesia and generalized arousal, during which they become wild, hallucinated, fugitive, and aggressive, alternating with lucid intervals. Despite these symptoms, attributable to brainstem encephalitis, neurological examination may prove surprisingly normal. Abnormalities include meningism, cranial-nerve lesions (especially III, VI, VII, IX, X, XI, and XII), upper motor-neurone lesions, fasciculation, and involuntary movements. Disturbances of the hypothalamus or autonomic nervous system cause hypersalivation (Fig. 9), lacrimation (Fig. 10), sweating, hypertension or hypotension, hyperthermia or hypothermia, inappropriate secretion of antidiuretic hormone, or diabetes insipidus and, rarely, priapism with spontaneous orgasms. Hypersexuality suggests similar aetiology to Klüver-Bucy syndrome.

Fig. 9 Hypersalivation in a Thai woman with furious rabies. (Copyright D.A. Warrell.)

Fig. 10 Lacrimation in a Thai patient with furious rabies. (Copyright D.A. Warrell.)

Without supportive treatment, about one-third of the patients will die during a hydrophobic spasm in the first few days. The rest lapse into coma and generalized flaccid paralysis, and rarely survive for more than a week without intensive care.

Paralytic or dumb rabies

This is the clinical pattern in less than a fifth of human cases except in the case of vampire bat-transmitted rabies (Fig. 11), which is invariably paralytic. The largest reported outbreak was in Trinidad between 1925 and 1935, when there were 89 human cases; others have been described from Mexico, Guyana, Brazil, Peru, Bolivia, and Argentina. The paralytic form of rabies was also seen in patients with postvaccinal rabies, in the two patients who inhaled fixed virus, and is said to be more likely to develop in patients who have received antirabies vaccine. After the usual prodromal symptoms, especially fever, headache, and local paraesthesiae, flaccid paralysis develops, usually in the bitten limb, and ascends symmetrically or asymmetrically with pain and fasciculation in the affected muscles and mild sensory disturbances. Paraplegia and sphincter involvement then develop, and finally fatal paralysis of deglutitive and respiratory muscles. Hydrophobia is unusual, but may be represented by a few pharyngeal spasms in the terminal phase of the illness. Even without intensive care, patients with paralytic rabies have survived for up to 30 days.

Fig. 11 Vampire bat bite inflicted on the ear of a sleeping child in Tapirái, São Paulo, Brazil (by courtesy of Dr João Luiz Costa Cardoso, São Paulo).

Other manifestations and complications

Respiratory system

Asphyxiation and respiratory arrest may complicate the hydrophobic spasms or generalized convulsions of furious rabies and the bulbar and respiratory paralysis of dumb rabies. Bronchopneumonia is an expected complication and a primary rabies pneumonitis may occur. Various abnormal patterns of respiration have been described, including cluster and apneustic breathing. There are some similarities to respiratory myoclonus. Pneumothorax may complicate inspiratory spasms.

Cardiovascular system

A variety of dangerous cardiac arrhythmias has been reported, including supraventricular tachycardias, sinus bradycardia, atrioventricular block, and sinus arrest. Hypotension, pulmonary oedema, and congestive cardiac failure are attributable to myocarditis.

Nervous system

Raised intracranial pressure resulting from cerebral oedema or internal hydrocephalus has been reported in a few cases, but spinal-fluid opening pressure is usually normal and papilloedema rarely found. Evidence of diffuse axonal neuropathy is consistent with histological appearances of degeneration of peripheral nerve ganglia and axons.

Gastrointestinal system

'Stress' ulcers and Mallory–Weiss syndrome are possible explanations for the haematemesis often reported in rabies.

Differential diagnosis

Rabies should be suspected whenever a patient develops severe neurological symptoms after being bitten by a mammal in a rabies endemic area. Some patients fail to remember that they have been bitten. Hydrophobia is pathognomonic of rabies and is unlikely to be mimicked accurately by the hysteric. Inspiratory spasms with

associated emotional response are produced by asking the patient to swallow accumulated saliva or by directing a draught of air on to their face. Patients are sometimes misreferred to otolaryngologists or psychiatrists.

Tetanus, which can also follow an animal bite, is similar to rabies in some respects, especially the pharyngeal form of cephalic tetanus ('hydrophobic tetanus'). It is distinguished by its shorter incubation period (usually less than 15 days), the presence of trismus, the persistence of muscle rigidity between spasms, the absence of meningoencephalitis (cerebrospinal fluid is universally normal), and a better prognosis. Hydrophobia does not occur in other encephalitides; the combination of intense brainstem encephalitis and furious behaviour in a conscious patient would be most unlikely except in rabies. Delirium tremens and some plant toxins (e.g. *Datura fastuosa*) and drugs (phenothiazines and amphetamines) may enter the differential diagnosis.

Paralytic rabies can be confused with other causes of ascending (Landry-type) paralysis. Postvaccinal encephalomyelitis (see below) usually develops within 2 weeks of the first dose of the older types of rabies vaccines. In poliomyelitis, objective sensory disturbances are absent and fever rarely persists after paralysis has developed. Examination of cerebrospinal fluid may help to distinguish acute inflammatory polyneuropathy (Guillain–Barré syndrome). *Herpesvirus simiae* (B virus) encephalomyelitis is transmitted by monkey bites, but the incubation period (3 to 4 days) is usually shorter than in rabies. Vesicles may be found in the monkey's mouth and at the site of the bite, and a diagnosis can be confirmed virologically.

Pathology

The brain, spinal cord, and peripheral nerves show ganglion-cell degeneration, perineural and perivascular mononuclear cell infiltration, neuronophagia, and glial nodules. Inflammatory changes are most marked in the midbrain and medulla (Fig. 12) in furious rabies and in the spinal cord in paralytic rabies. The diagnostic Negri bodies (Fig. 13 and Plate 1) are eosinophilic, intracytoplasmic inclusions predominantly consisting of masses of viral ribonucleoprotein, with a basophilic inner body, containing fragments of cellular organelles including ribosomes and occasional virions. They can be demonstrated by haematoxylin and eosin or Schleiften's stains in histological sections of grey matter in up to 75 per cent of human cases, especially in hippocampal pyramidal cells and cerebellar Purkinje cells.

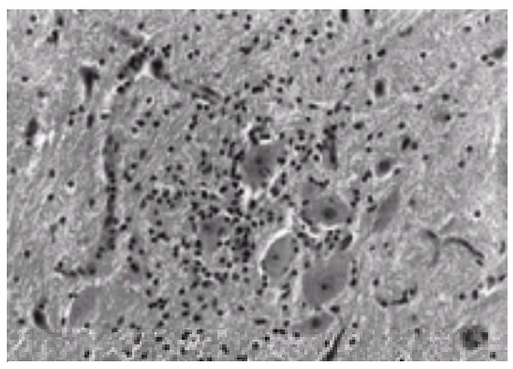

Fig. 12 Inflammatory cells around neurones in the central medulla (para-ambigualis region) of a patient who died of rabies encephalitis (× 400). (Reproduced by courtesy of Dr P. Lewis, London.)

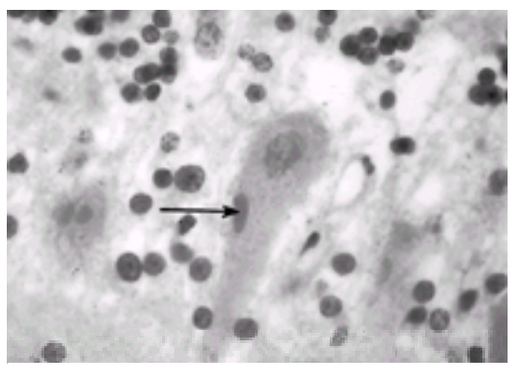

Fig. 13 Street virus in human cerebellar Purkinje cells as seen with the light microscope. Several Negri bodies can be seen (one is arrowed) (× 615). (By courtesy of the Armed Forces Institute of Pathology 73–12330.) (See also Plate 1.)

In view of the appalling prognosis of rabies encephalitis, neuronolysis is often surprisingly mild and patchy, and death can occur without any inflammatory response. Vascular lesions such as thrombosis and haemorrhage have also been described. The brainstem, limbic system, amygdaloid nuclei and hypothalamus appear to be most severely affected. Outside the nervous system, there is focal degeneration of salivary and lacrimal glands, pancreas, adrenal medulla, and lymph nodes. An interstitial myocarditis, with round-cell infiltration, is found in about 25 per cent of cases.

Laboratory diagnosis

A suspect rabid animal that might have infected a patient should be killed and the brain examined without delay. Rabies antigen detection by a direct immunofluorescent antibody test on acetone-fixed brain impression smears is usually used or alternatively, if no fluorescent microscope is available, by rapid enzyme immunodiagnosis. Virus isolation takes up to 3 weeks by intracerebral inoculation of mice, or about 4 days in murine neuroblastoma cell culture.

In humans, rabies can be confirmed early in the illness by demonstration of viral antigen by a direct immunofluorescent antibody test in nerve twiglets in skin biopsies (Fig. 14). This rapid method is positive in 60 to 100 per cent of cases, and no false-positive results have been reported. Antigen can also be detected in brain biopsies but tests on corneal impression smears are usually falsely negative. The polymerase chain reaction is being used increasingly to identify rabies antigen in secretions.

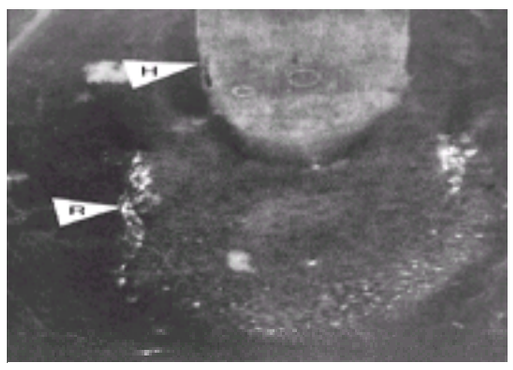

Fig. 14 Diagnosis of human rabies during life. Vertical section through a hair follicle and shaft (H) showing fluorescence of nerve cells around the follicle (R) indicating the presence of rabies antigen (× 250). (Copyright M.J. Warrell.)

During the first week of illness virus may be isolated from saliva, brain, cerebrospinal fluid, and very rarely urine. Rabies antibodies are not usually detectable in serum or cerebrospinal fluid before the eighth day of illness in unvaccinated patients. Antibody may leak into the cerebrospinal fluid in patients with postvaccinal encephalomyelitis, but a very high titre suggests a diagnosis of rabies. A specific IgM test has not proved useful diagnostically.

Treatment

Patients with rabies must be sedated heavily and given adequate analgesia to relieve their pain and terror. If intensive care is undertaken, the aim is to prevent complications such as cardiac arrhythmias, cardiac and respiratory failure, raised intracranial pressure, convulsions, fluid and electrolyte disturbances including diabetes insipidus and inappropriate secretion of antidiuretic hormone, and hyperpyrexia. Antiserum, antiviral agents, inter-feron-a, corticosteroid, and other immunosuppressants have proved useless.

Prognosis

Rabies was formerly regarded as a universally fatal disease, but there are reports of five cases of recovery or prolonged survival, following intensive care. Two patients who recovered had had nervous-tissue vaccine treatment, and the diagnoses were made serologically. Three patients were given pre- or postexposure tissue-culture vaccine and they never recovered from profound neurological deficits.

Despite intensive care, the prognosis of rabies encephalomyelitis is still virtually hopeless. At the time of the bite, however, before virus has invaded the nervous system, correct cleaning of the wound (see below) and the use of optimum postexposure immunization reduce the risk of rabies developing from about 35 to 65 per cent in untreated cases to near zero. The risk varies with the biting species and the site and severity of the bites. It is highest following head bites by proved rabid wolves, when the mortality in unvaccinated people may exceed 80 per cent.

Prevention and control

In countries where rabies is endemic

Control strategy is based on gathering information about the prevalence and host range of rabies in wild and domestic animals. This requires laboratory facilities for confirming the diagnosis. Domestic animals can be protected by yearly vaccination. Dogs are muzzled and kept off the streets; strays are eliminated. People should be discouraged from keeping wild carnivores, such as skunks, raccoons, coatimundis, and mongooses, as pets. Unnecessary contact with mammals should be avoided (e.g. stroking stray dogs, exploring bat-infested caves). Reduction of wild-animal reservoir populations may be attempted, but this is difficult to achieve and likely to cause ecological chaos. An alternative approach is to attempt the vaccination of key reservoir species by using live oral vaccines distributed in bait. An oral vaccinia-recombinant rabies glycoprotein vaccine has successfully controlled fox rabies in Western Europe, and is also used in North American raccoons and coyotes. Rabies is most likely to be controlled or eradicated where the principal reservoir is the domestic dog, as in nineteenth-century Britain, Malaysia, and Japan.

Education and publicity about rabies is always needed. Clinics and dispensaries must be adequately supplied with vaccine and antiserum to provide postexposure prophylaxis, but difficulties over expense, supply, and preservation often preclude this. In dog-rabies endemic areas, pre-exposure vaccination is advisable but rarely used.

In countries where rabies is not endemic

Importation of potential vectors, especially domestic dogs and cats and wild bats and carnivores, should be strictly controlled and, where feasible, imported mammals should be vaccinated against rabies and kept in quarantine for an adequate period.

Postexposure prophylaxis may be required for people who were exposed to the risk of rabies while abroad. Travellers should be educated to seek local medical help if they are bitten, scratched, or licked by animals. Many travellers wait until they return to their homeland, sometimes weeks or months after the bite, before asking for medical advice.

Pre-exposure immunization regimens

Pre-exposure vaccination is the most effective form of rabies prevention. No rabies deaths have been reported in anyone who had pre-exposure vaccine followed by postexposure booster doses. In rabies-free areas it is needed by those who handle imported animals before and during quarantine in kennels, zoos, and laboratories; those who work with rabies virus in laboratories; and those who are resident in or intend to travel to dog rabies-endemic areas. In endemic areas others particularly at risk in certain areas include veterinarians, dog-catchers, cave explorers, naturalists, and animal collectors.

A course of three doses of tissue-culture rabies vaccine (see below) is given on days 0, 3, and 28 (or 21) intramuscularly or 0.1 ml intradermally into the deltoid or the anterolateral thigh in children, preferably into the same limb. If chloroquine is being taken for malaria prophylaxis, or in other cases of suspected immunosuppression, the intramuscular route must be used. If sharing an ampoule for intradermal injections, a sterile needle and syringe must be used for each patient. The neutralizing antibody response is enhanced and prolonged by a booster dose after 1 year. A prompt secondary response to booster injections then occurs after many years. Repeated booster doses are only needed if the risk of infection is high.

The antibody response is so predictable that it need not normally be checked, unless there is immunosuppression. An antibody level above 0.5 IU/ml indicates immunity, and serological monitoring can prevent unnecessary booster doses for rabies laboratory staff and others at continuous risk.

Postexposure prophylaxis

The decision to give postexposure treatment depends on the precise geographical location of the exposure; when it occurred; its severity—whether it was a bite or lick on broken skin; the nature, appearance, behaviour, and fate of the biting animal and if possible, whether it had been vaccinated against rabies within the last year. This information may allow proper assessment of risk; but if there is any doubt the patient should be given full postexposure prophylaxis, even if the bite is several months old.

The aim is to neutralize inoculated virus before it can enter the nervous system. Wound cleaning and active and passive immunization must be implemented as soon as possible.

Wound cleaning

This is effective in killing virus in superficial wounds, but is often neglected. First aid consists of scrubbing the wound with soap and water for several minutes. Foreign material should be removed and a viricidal agent such as povidone iodine, or 40 to 70 per cent alcohol, should be applied liberally. Quarternary ammonium compounds, such as benzalkonium chloride, are inactivated by soap and so are no longer recommended. Hospital treatment of wounds involves thorough exploration, debridement, and irrigation of deep wounds, if necessary under local or general anaesthetic. Suturing should be avoided or delayed and the wound left without occlusive dressings. Attention should be given to tetanus prophylaxis and the range of bacterial and other pathogens, particularly associated with mammal bites. Most of the bacteria are sensitive to amoxicillin/clavulanic acid, cephoxitin, or tetracycline.

Specific prophylaxis

This consists of active and passive immunization. The indications are given in Table 1.

Active immunization

Tissue-culture vaccines

Human diploid cell vaccine (**HDCV**) (Imovax rabies™ Aventis Pasteur), purified chick embryo cell vaccine (Rabipur/Rab Avert™ Chiron Behring), and purified vero cell vaccine (**PVRV**) (Verorab™Aventis Pasteur) are now the tissue-culture vaccines of choice.

The intramuscular postexposure regimen of these vaccines is 5 × 1 ml doses into the deltoid on days 0, 3, 7, 14, and 28, although for PVRV each dose is only 0.5 ml. An economical eight-site intradermal regimen can be used with vaccines which have an intramuscular dose of 1 ml. On day 0, eight intradermal injections of 0.1 ml are given (deltoids, suprascapular, lower-quadrant abdominal wall, and thighs) using a whole ampoule in a Mantoux-type syringe. On day 7, four intradermal injections of 0.1 ml are given (deltoids and thighs), and single intradermal doses of 0.1 ml are given on days 28 and 91. Advantages of the eight-site method are: fewer hospital attendances; a rapid induction of neutralizing antibody, making it the treatment of choice when no rabies immunoglobulin is available; a wide margin of safety; using a whole ampoule on day 0, which avoids sharing ampoules of vaccine between patients during the emergency treatment; and finally, giving a large antigenic stimulus on day 0 gives the best chance of survival to patients who are 'low responders' to the vaccine and to those who fail to return on time for subsequent doses.

The two-site intradermal regimen was designed for use with PVRV. A dose of 0.1 ml for PVRV and 0.2 ml for the other two vaccines is given intradermally at two sites (deltoids) on days 0, 3, and 7 and a one site on days 28 and 90. Both of these intradermal regimens use a similar total amount of vaccine per course, about 40 per cent of that of the intramuscular regimen, but the antibody response following the eight-site method is significantly earlier and higher than that after the two-site regimen.

Side-effects of tissue-culture vaccines are mild and transient: local itching, redness or pain at the site of injection, influenza-like symptoms, and occasionally a rash. More serious allergic reactions include rare, type I immediate hypersensitivity during primary courses. Type III immune-complex hypersensitivity was reported in 6 per cent of those receiving booster doses of HDCV in the United States and consisted of urticaria, rash, angio-oedema and arthralgia 3 to 13 days after injection, but none has been fatal. A few cases of polyneuritis, Guillain–Barré syndrome, or local limb weakness have been reported in patients receiving tissue-culture vaccines. These events are very rare and no more frequent than for other commonly used virus vaccines.

Nervous-tissue vaccines

Semple vaccine, a 5 per cent sheep or goat brain suspension, and suckling mouse brain (Fuenzalida) vaccine are still used in Asia and the latter also in Africa and South America. The abdomen is often used as a suitable target for the daily subcutaneous injections. These vaccines produce neurological reactions, including postvaccinal encephalomyelitis.

Postvaccinal encephalomyelitis

Neurological reactions to nervous-tissue vaccines occur in up to 1 in 220 courses of Semple vaccine, with a 3 per cent mortality, and are an allergic response to myelin and related neural proteins in the vaccine. In Latin America, neuroparalytic reactions complicated 1/7865 to 1/27 000 courses of suckling mouse brain vaccine with a 22 per cent mortality. Most reactions to Semple vaccine affect the central nervous system, whereas at least 70 per cent of those following suckling mouse brain vaccine involved the peripheral nervous system.

The incubation period ranges from 3 to 35 days after the first injection of vaccine, but it is usually between 7 and 14 days. Clinical forms include a rapidly reversible mononeuritis multiplex involving particularly the cranial, radial, brachial, and sciatic nerves; a dorsolumbar transverse myelitis with fever, paralysis, and sensory loss in the lower limbs, with sphincter involvement, loss of tendon reflexes, extensor plantar responses, and severe girdle and thoracic pain; an ascending paralysis (Landry type), which ends in fatal bulbar paralysis in a third of cases; and meningoencephalitic and meningoencephalomyelitic reactions. The overall mortality of these reactions is 15 to 20 per cent. Most survivors make a complete recovery in 2 to 3 weeks, but a few are left with permanent neurological sequelae.

A moderate lymphocyte pleocytosis and elevated cerebrospinal fluid protein is usual. Pathological changes consist of swelling and chromatolysis of neurones with extensive perivascular demyelination and lymphocytic infiltration in the spinal cord. These features resemble experimental allergic encephalitis, postvaccinal encephalomyelitis after vaccinia vaccine, postinfectious encephalomyelitis, and acute multiple sclerosis. Corticosteroids, for example, prednisolone at 40 to 60 mg/day, are thought to be helpful, and the use of cyclophosphamide has been suggested. Vaccination should be stopped as soon as symptoms appear and the course continued with a tissue-culture vaccine.

Passive immunization

Rabies immune globulin (**RIG**) has proved valuable in protection, presumably by neutralizing rabies virus during the first week after initial vaccination, before neutralizing antibody has appeared, and it enhances the T-lymphocyte response to vaccine experimentally. Its use is recommended at the start of all primary postexposure courses of rabies vaccine, but it is vital following severe bites (on the head, neck, hands, and multiple or deep bites).

The dose of human RIG is 20 IU/kg body weight, and 40 IU/kg for equine RIG. Serum sickness has not been reported after human RIG treatment, but hypersensitivity reactions to equine RIG occur in 1 to 6 per cent of those treated; however, these are not reliably predicted by a previous intradermal test. RIG must be given even if the test is positive, and the skin test is unnecessary. Adrenaline should always be available in case of reactions.

The RIG is infiltrated into the tissues around the bite wound, and any remaining is injected intramuscularly into the thigh, not the buttock. If RIG is given hours or days before the first dose of vaccine, the immune response will be impaired. RIG is prohibitively expensive and is not available or affordable to more than 95 per cent of patients receiving postexposure treatment in developing countries.

Postexposure prophylaxis in people who have received previous vaccination

If a complete pre- or postexposure course of a modern potent tissue-culture vaccine has been given, or if the neutralizing antibody level has been over 0.5 IU/ml, only two intramuscular doses of tissue-culture vaccine should be given on days 0 and 3. The first dose of the vaccine can be divided between four or eight intradermal sites on day 0. Passive immunization is not required. Otherwise, full postexposure treatment must be given.

Failures of postexposure prophylaxis

Deaths from rabies have occurred despite vaccine treatment. These may be attributable to the use of low-potency nervous-tissue vaccines, delay in starting vaccination, an incomplete vaccine course, omission of passive immunization, failure to infiltrate RIG around the wound, injection of vaccine into the buttock, or decreased immune responsiveness of the vaccinee. So far, in the few cases in which the virus strain could be typed, tissue-culture vaccine failures could not be attributed to failure of neutralization of the particular infecting strain of virus by antibody induced by the vaccine. However, vaccine protection against rabies-related viruses may be less efficient than against genotype 1 rabies viruses (see below).

A reduced or delayed immune response to vaccine can sometimes be predicted. If treatment is started late (e.g. more than 2 days after exposure), no RIG is available for severe bites, the patient is immunocompromised, or a rabies-related virus infection is suspected, the immune stimulus can be enhanced either by doubling the initial dose of vaccine, or by dividing the first dose of tissue-culture vaccine between eight sites intradermally, as for the economical eight-site regimen (see above).

Rabies-related viruses known to infect humans

Mokola virus, Duvenhage virus, European bat lyssavirus, and Australian bat lyssavirus are rabies-related viruses that have been proved to infect humans. Three genotypes are only found in Africa. Genotype 2, Lagos bat virus has not been detected in humans. Mokola virus (genotype 3) has been isolated from shrews (*Crocidura* spp.) in Nigeria and Cameroon, and mainly from cats in South Africa, Zimbabwe, and Ethiopia. It was isolated from a child with meningitis who recovered, and from another with fatal encephalitis. Duvenhage virus (genotype 4) caused a fatal illness, with clinical features identical to furious rabies, in a South African of that name who was bitten by a bat and had then received a full course of rabies vaccine. The rabies fluorescent antibody test was negative in the Duvenhage case and

weakly positive in the Mokola cases.

Rabid bats had been found occasionally in Europe since 1954. In 1985, a woman was bitten by a rabid, insectivorous bat in Denmark, and an extensive search revealed many rabid bats there and across Europe. The European bat lyssavirus (**EBL**) group comprises genotype 5 (**EBL 1**) and genotype 6 (**EBL 2**), which are each subdivided into phylogenetically distinct groups a and b. The vector species of EBL 1 is *Eptesicus serotinus.* Type 1a is found in Russia, Poland, Germany, Denmark, and the Netherlands; type 1b in France, the Netherlands, and Spain. Two Russian girls died of rabies following bat bites. Myotis bats harbour EBL 2: EBL 2a in the Netherlands and the United Kingdom (three isolates including the human fatality in 2002), and EBL 2b in Switzerland. A fatal human case in Finland was due to EBL 2b.

The identification of a lyssavirus in fruit bats (genus *Pteropus*) in eastern Australia was an unexpected finding in 1996. This Australian bat lyssavirus (**ABL**) (genotype 7) has since caused a fatal rabies-like encephalitis in two women who had handled bats.

The G protein of rabies vaccine strains is very similar to that of all genotype 1 viruses, but shows a variable degree of antigenic homology to rabies-related viruses. Tissue-culture rabies vaccines have not protected animals against challenge with Mokola virus and their effect against Duvenhage virus is uncertain. Protection against EBL viruses may be slightly less efficient, but ABL is closely related to genotype 1 strains and so protection should be undiminished.

Further reading

Baer GM *et al*. (1988). Research towards rabies prevention. *Reviews of Infectious Diseases* **10** (Suppl. 4), S1–815.

Centers for Disease Control (1999). Human rabies prevention—United States, 1999. Recommendations of the Advisory Committee on Immunization Practices. *Morbidity and Mortality Weekly Report* **48**(Suppl) RR –1.

Dietzschold B, Morimoto K, Hooper DC (2001). Mechanisms of virus-induced neuronal damage and the clearance of viruses from the CNS. *Current Topics in Microbiology and Immunology* **253**, 145–55.

Helmick CG, Tauxe RV, Vernon AA (1987). Is there a risk to contacts of patients with rabies? *Reviews of Infectious Diseases* **9**, 511–18.

Jackson AC, Wunner WH, eds. (2002). *Rabies.* Academic Press, San Diego.

Smith JS (1996). New aspects of rabies with emphasis on epidemiology, diagnosis and prevention of the diease in the United States. *Clinical Microbiology Reviews* **9**, 166–76.

Warrell DA *et al.* (1976). Pathophysiologic studies in human rabies. *American Journal of Medicine* **60**, 180–90.

Warrell MJ, Warrell DA (1995). Rhabdovirus infections of humans. In: Porterfield JS, ed. *Exotic viral infections*, pp 343–83. Chapman & Hall, London.

Warrell MJ *et al.* (1985). Economical multiple-site intradermal immunisation with human diploid-cell-strain vaccine is effective for post-exposure rabies prophylaxis. *Lancet* **i**, 1059–62.

WHO (1997). WHO Recommendations on rabies post-exposure treatment and the correct technique of intradermal immunization against rabies. WHO/EMC/ZOO.96.6.

7.10.10 Colorado tick fever and other arthropod-borne reoviruses

M. J. Warrell and D. A. Warrell

Within the large family of Reoviridae human pathogens are found in four genera: *Reovirus*, *Rotavirus*, and two arthropod-borne genera *Coltivirus* and *Orbivirus*. Colorado tick fever is a group A coltivirus, together with Eyach virus, from Europe. Group B comprises coltiviruses from South-east Asia and Indonesia and Banna virus from China. This group may be reclassified as the *Seadornavirus* genus. Banna virus has been isolated from patients with encephalitis. The four pathogenic orbivirus serogroups are Kemerovo, Changuinola, Orungo, and Lebombo.

Colorado tick fever

The virus responsible for Colorado tick fever or 'mountain fever' is an 80-nm, double-shelled particle, covered with capsomeres. The icosahedral core contains 12 segments of double-stranded, negative-sense RNA. The virus has the ability to infect human erythrocytes and this may also occur with the other colti- and orbiviruses.

Colorado tick fever is a zoonosis involving hard (ixodid) ticks (principally *Dermacentor andersoni*, but also *D. occidentalis*, *D. variabilis*, *D. parumapertus*, *D. albipictus*, etc.) and wild mammals including porcupines, coyotes, squirrels, chipmunks, deer, mice, and other rodents. Ticks pass Colorado tick fever virus trans-stadially, but not transovarially.

Epidemiology

Colorado tick fever is acquired from tick bites in western and north-western parts of the United States (including California), British Columbia, and Alberta. Very rarely, it has been caused by an infected blood transfusion. Several hundred cases are reported each year in the United States, but the true incidence is thought to be at least 10 times higher than that. Hikers and campers are at special risk in rodent and tick-infested terrain. The prevalence of antibody to Colorado tick fever among shepherds was 32 per cent. The highest incidence is from May to July when ticks are most active. Infection usually confers lasting immunity.

Clinical features

In adults the infection is nearly always mild, but in children it is occasionally severe or even fatal. Three to six days after the tick bite (extreme range, 1 to 19 days) there is a sudden fever with rigors, generalized aches, myalgia, headache, and backache. In half the patients there is a biphasic fever. A maculopapular or petechial rash appears in about 10 per cent of cases and gastrointestinal symptoms in 20 per cent. Laboratory findings include leucopenia with relative lymphocytosis, occasional thrombocytopenia, and mild lymphocyte pleocytosis.

The illness usually resolves in about 10 to 14 days, but convalescence may be prolonged. Severe manifestations include meningism and drowsiness, sometimes associated with gastrointestinal symptoms, spontaneous bleeding, thrombocytopenia, and disseminated intravascular coagulation. Late, possibly immunological effects, include myocarditis, pericarditis, pleurisy, arthritis, and epididymitis. Colorado tick fever infection may precipitate abortion, but the transplacental infection and teratogenic effects reported in mice have not been observed in man.

Diagnosis

Viral antigen may be detected in erythrocytes by immunofluorescence 1 to 120 days after the start of symptoms. Erythrocyte precursors are infected in the marrow, but their survival is apparently not affected. Virus can be isolated from the blood, and if there is central nervous system involvement, the cerebrospinal fluid. Colorado tick fever virus produces a cytopathic effect on several cell lines, but intracerebral injection of ground blood clot, or preferably washed erythrocytes, into suckling mice is more sensitive for diagnostic isolation. An indirect fluorescent antibody test can provide early serodiagnosis, but acute infections can be diagnosed by polymerase chain reaction detection of antigen. Neutralizing antibody and specific IgM enzyme immunoassays become positive after 14 to 21 days and the IgM disappears after 45 days.

Differential diagnosis

Many other tick-borne acute febrile illnesses, some with rashes and nervous system involvement, can be acquired in the endemic area for Colorado tick fever. These include Rocky Mountain spotted fever, tularaemia, Lyme disease, and relapsing fever. Tick paralysis caused by *D. andersoni* and other ixodid ticks presents as a poliomyelitis-like, ascending, flaccid paralysis that is unlikely to be mistaken for the meningitic or encephalitic syndromes of Colorado tick fever.

Treatment

The symptomatic treatment of fever and pain should exclude salicylates in case of thrombocytopenia. Tribavirin (ribavirin) inhibits the replication of Colorado tick fever virus experimentally but its use in humans has not been reported.

Orbivirus serogroups

Kemerovo

The orbivirus serogroup Kemerovo contains three viruses isolated from *Ixodes* and *Hyalomma* ticks in Russia and Central Europe. They cause enign febrile illnesses and, occasionally, meningitis or encephalitis in spring and early summer, when ticks are active. Rodents and birds are involved in the zoonotic cycle. Oklahoma tick fever is another Kemerovo virus rarely causing febrile illness in the United States.

Changuinola

There is a single report of human febrile illness with the orbivirus Changuinola in Panama. The virus has been isolated from phlebotomine flies and mammals in that area.

Orungo

Orungo virus is found mainly in West Africa but also in Uganda and the Central African Republic. Up to 75 per cent of some populations are seropositive. The clinical effects are unknown but fever and diarrhoea occur in some people, perhaps with encephalitis, as in experimental mice. There is no rash or jaundice. It is transmitted

by *Anopheles*, *Aedes*, and other mosquitoes. Monkeys, sheep, and cattle may be infected.

Lebombo

This reovirus was isolated from one febrile child in Nigeria. Lebombo is also found in mosquitoes and rodents.

Further reading

Brown SE, Knudson DL (1995). Coltivirus infections. In: Porterfield JS, ed. *Exotic viral infections*, pp 329–42. Chapman & Hall, London.

Burgdorfer W (1977). Tick-borne diseases in the United States: Rocky Mountain spotted fever and Colorado tick fever. A review. *Acta Tropica* **34**, 103–26.

Libikova H *et al.* (1978). Orbiviruses of the Kemerovo complex and neurological diseases. *Medical Microbiology and Immunology* **166**, 255–63.

Monath TP, Guirakhoo F (1996). Orbiviruses and coltiviruses. In: Fields BN I>et al eds. *Fields virology*, 3rd edn, Vol 2, pp 1735–66. Lippincott-Raven, Philadelphia.

7.10.11 Alphaviruses

L. R. Petersen and D. J. Gubler

Introduction

The genus Alphavirus of the family Togaviridae comprises 27 registered viruses, 16 of which cause human infection (Table 1). Alphaviruses are lipid-enveloped virions with a diameter of 50 to 70 nm whose genome is a molecule of single-stranded, positive-sense RNA approximately 12 000 nucleotides in length. Most alphaviruses are maintained in nature in complex transmission cycles between wild or domestic animals and one or more mosquito species. Humans become infected from infective mosquitoes that take a bloodmeal. Patients develop high viraemias with some alphaviruses and may contribute to the transmission cycle by infecting mosquitoes. The epidemiology and geographic distribution of the alphaviruses depend on several factors, including the requirements for and presence of suitable amplifying hosts, the presence and feeding behaviour of a suitable arthropod vector, and the frequency of exposure of non-immune reservoir hosts and humans to infected vectors. Alphavirus infections are not communicable.

Most infections in humans are asymptomatic, but alphaviruses can cause a spectrum of clinical illness ranging from non-specific febrile illness, often with rash, myalgia, or arthralgia, to frank encephalitis and death (Table 1). No specific therapy is available. Vaccines for some alphaviruses are used in animals, although none have been licensed for humans.

Laboratory diagnosis

Alphavirus infections are diagnosed serologically by detection of immunoglobulin M (IgM) and G (IgG) responses. All alphaviruses have common antigenic determinants that result in cross-reactions in immunodiagnostic tests. Neutralization tests may be necessary for serological confirmation in areas where multiple alphaviruses are endemic/enzootic. Isolation of virus from acute-phase serum is possible with some alphaviruses, but they are seldom recovered from the central nervous system, including cerebrospinal fluid, except from fatal cases. Virological diagnosis may also be made using polymerase chain reaction and immunohistochemistry on tissue samples.

Specific alphavirus infections

Chikungunya

Aetiology and epidemiology

Chikungunya virus is found in Africa and Asia and is transmitted primarily by *Aedes* mosquitoes. Non-human primates such as monkeys and baboons may be the primary maintenance hosts in sylvan settings in Africa. In urban settings in Africa and Asia, the virus is transmitted from human to human via *Aedes aegypti* mosquitoes. Explosive urban epidemics occur during the rainy season. One epidemic, in Madras in India, caused an estimated 300 000 cases. Serosurveys have shown antibody prevalences greater than 90 per cent in some areas of Africa.

Clinical characteristics

The sudden onset of fever and crippling arthralgia follows an incubation period of 2 to 3 days (range 1 to 12). The fever may remit for 1 to 2 days and then recur ('saddle back' fever). Arthralgias are polyarticular, migratory, and mostly involve the small joints. Papular or maculopapular skin rashes, typically on the trunk and limbs, occur, usually during the second to fifth day of illness. Most infections are probably asymptomatic. Arthralgia may last several months; a few patients may have symptoms 5 years after infection. Children are less likely to present with arthralgia and rash, and more likely to have headache, injected pharynx, and gastrointestinal symptoms.

Diagnosis

Leucopenia is the only likely laboratory finding. Viraemia usually is present in the first 48 h of illness. Haemagglutinin-inhibition and IgM antibodies will be present in nearly all patients by the seventh day of illness. IgM antibodies detectable in serum by IgM antibody capture enzyme-linked immunosorbent assay (MAC-ELISA) may persist 6 months after infection.

Prevention, control, and treatment

Prevention and control can only be achieved by reducing *Ae. aegypti* populations in the large urban centres of the tropics and by avoiding mosquito bites. No licensed vaccines currently exist. There is no specific treatment. Anti-inflammatory drugs may relieve arthralgia. Chloroquine phosphate may be helpful for refractory arthralgias.

Eastern equine encephalitis

Aetiology and epidemiology

The virus is widely distributed throughout North, Central, and South America and the Caribbean; however, little is known about the epidemiology of eastern equine encephalitis outside North America. In the United States, human infections are usually sporadic and small outbreaks occur each summer, mostly along the Atlantic and Gulf Coasts. In recent years, one to 14 cases have been reported annually. In North America, wild birds and *Culiseta melanura* mosquitoes maintain the virus.

Clinical characteristics

The incubation period exceeds 1 week and the onset is abrupt with high fever. About 2 per cent of infected adults and 6 per cent of children develop encephalitis. Eastern equine encephalitis is the most severe of the arboviral encephalitides, with a mortality of 50 to 75 per cent. Symptoms and signs include dizziness, decreasing level of consciousness, tremors, seizures, and focal neurological signs. Death can occur within 3 to 5 days of onset. Sequelae, common in non-fatal encephalitis, include convulsions, paralysis, and mental retardation. Illness due to eastern equine encephalitis in South America appears to be less severe.

Diagnosis

Cerebrospinal fluid pressure may be raised, with slightly increased protein, normal sugar, and up to 2000 cells/mm ^{3}. IgM antibodies are readily detected in serum or cerebrospinal fluid by ELISA. Paired serum samples can be tested by haemagglutinin inhibition, ELISA, or neutralization tests. Horse or pheasant deaths and the proximity to swamps provide clues to the diagnosis.

Prevention control and treatment

Prevention depends on the avoidance of mosquito bites and mosquito control in suburban areas. Inactivated vaccines have been used successfully in horses, and an inactivated vaccine has been used experimentally in laboratory workers and others at high risk of exposure. No specific treatment is available.

Ross River virus

Aetiology and epidemiology

This virus causes 'epidemic polyarthritis' in Australia, south-western Pacific islands, and Fiji. *Ae. vigilax* is an important vector in Australia and *Ae. scutellaris* complex mosquitoes in some south Pacific islands, although the virus has been isolated from more than 30 mosquito species. An epidemic in various Pacific islands in 1979 to 1980 affected more than 50 000 people. An average of 4800 cases is reported annually from Australia. Explosive outbreaks and viraemias in humans suggest virus transmission from human to human by certain mosquitoes.

Clinical characteristics

The illness begins suddenly with arthralgia in the small joints of the hands and feet. A maculopapular rash occurs in about half of patients within 2 days of onset and is most prominent on the trunk and limbs, but can cover the entire body. The rash may progress to small vesicles. Myalgia, headache, anorexia, nausea, and tenosynovitis are common, but the temperature is only slightly elevated. The arthralgia may be prolonged. Symptomatic infection is rare in children.

Diagnosis

Isolation of virus from serum is possible for the first few days of illness. IgM antibodies will be detected by MAC-ELISA within 5 to 10 days of onset. Complement fixation, haemaglutinin inhibition, and neutralization tests may be useful, particularly when paired serum samples are available.

Prevention, control, and treatment

Avoidance of mosquito bites and peridomestic mosquito control can effectively reduce the risk of infection. No specific treatment is available. Non-steroidal anti-inflammatory drugs may relieve symptoms.

Venezuelan equine encephalitis complex

Aetiology and epidemiology

Six subtypes (I–VI) within the Venezuelan equine encephalitis virus complex have been identified. Five antigenic variants exist within subtype I (IAB, IC, ID, IE, IF). These subtypes and variants are classified as epizootic or enzootic, based on their apparent virulence and epidemiology. Epizootic variants of subtype I (IAB and IC) cause equine epizootics and are associated with more severe human disease. Enzootic strains (ID-F, II (Everglades), III (Mucambo, Tonate, Paramana), IV (Pixuna), V (Cabassou), VI (unnamed)) do not cause epizootics in horses, but may produce sporadic disease in man. Large epizootics (IAB and IC) have occurred in equines in northern countries of South America and Central America, sometimes reaching the United States. In 1969 to 1972, a massive epizootic extending from Ecuador to Texas killed more than 200 000 horses and caused several thousand human infections. In 1995, a large epizootic, which began in Venezuela and spread to Colombia, affected thousands of horses and caused approximately 90 000 human infections. Epizootic strains are carried by a wide variety of mosquitoes including *Aedes*, *Mansonia*, and *Psorophora* spp. Horses are the principal amplifying hosts during epizootics but are not amplifying hosts for enzootic transmission. Enzootic strains are maintained in a cycle involving *Culex* (*Melanoconion*) mosquitoes and rodents.

Clinical characteristics (epizootic virus infections)

After an incubation period of 1 to 6 days, there is a brief febrile illness of sudden onset, characterized by malaise, nausea or vomiting, headache, and myalgia. Acute symptoms last 2 to 5 days; generalized asthenia up to 3 weeks. Among those with clinical illness, less than 0.5 per cent of adults and less than 4 per cent of children develop encephalitis. Nausea and vomiting, nuchal rigidity, ataxia, convulsions, paralysis, and death may occur. Long-term sequelae following encephalitis are uncommon.

Diagnosis (epizootic virus infections)

A marked leucopenia is universal, often accompanied by neutropenia and thrombocytopenia, with moderate lymphocytosis in the cerebrospinal fluid. Virus isolation from serum or throat swab is possible within the first few days of illness. Paired sera can be tested by HI and neutralizing tests. Specific IgM can be detected by MAC-ELISA in the second week of illness.

Prevention, control, and treatment

Equine immunization is effective in controlling epizootic disease. Venezuelan equine encephalitis is highly infectious by the aerosol route; many laboratory infections have occurred. Live attenuated and inactivated vaccines have been used in laboratory workers. People in affected areas should avoid mosquito bites. No specific treatment is available.

Western equine encephalitis

Aetiology and epidemiology

This is a complex of closely related viruses found in North and South America, but human disease is rare outside North America and Brazil. Summer outbreaks may be precipitated by flooding, which increases breeding of *Culex* mosquitoes (particularly *Culex tarsalis* in the western United States). Large outbreaks of western equine encephalitis in humans and horses occurred in the western United States in the 1950s and 1960s; however, a declining horse population, equine vaccination, and improved vector control have reduced the reported number of human cases to zero, in most years during the last decade.

Clinical characteristics

The ratio of apparent to inapparent infection in adults is less than 1 in 1000; however, this ratio increases to 1:1 in infants under 1 year of age. Following an incubation period of about 7 days, headache, vomiting, stiff neck, and backache are typical; restlessness and irritability are seen in children. Weakness and hyporeflexia are common. Convulsions occur in 90 per cent of affected infants and 40 per cent of children between 1 and 4 years, but are rare in adults. Recovery in 5 to 10 days is common, but convalescence may be protracted. Although rare in adults and older children, sequelae are common in infants, with half of those with encephalitis left with convulsions, and/or severe motor or intellectual deficits. The case fatality rate is 3 to 7 per cent.

Diagnosis

Clinical laboratory findings in western equine encephalitis are often not remarkable. IgM antibodies are readily detected in serum by ELISA. Paired sera can be tested by HI, IgG ELISA, or neutralization tests. Virus can occasionally be isolated from serum or cerebrospinal fluid.

Prevention, control, and treatment

Prevention of western equine encephalitis relies on mosquito control and the avoidance of mosquito bites. Vaccine is available for horses. An inactivated vaccine has been used for laboratory staff and others at high risk of exposure. No specific treatment is available.

Other alphavirus infections

Barmah Forest virus

Since its first recognition as a cause of human disease in 1988, the geographical distribution of Barmah Forest virus has expanded recently in Australia. It causes sporadic disease and epidemics, with up to 300 serologically confirmed cases. The disease resembles that of Ross River virus infection, although the rash tends to be more florid and true arthritis is less common. The illness is prolonged in some patients. Little is known about the ecology of Barmah Forest virus, although outbreaks have coincided with Ross River virus outbreaks, and the virus has been identified in the same mosquito species.

Mayaro virus

Mayaro virus has been isolated from humans and various mosquito species (mostly *Haemagogus*) in Trinidad, Brazil, Bolivia, French Guiana, Surinam, Peru, and Venezuela. Several outbreaks have been identified, most recently in Venezuela in 2000. The disease is characterized by an abrupt onset with fever, chills, headache, myalgia, and arthralgia, mostly in the small joints of the extremities. A maculopapular rash may occur 2 to 5 days after defervescence. Arthralgia may persist for several months.

O'nyong-nyong virus

From 1959 to 1962, this virus caused epidemics in Uganda, Kenya, Tanzania, and Malawi, involving approximately 2 million people. The virus was isolated in 1978 from *Anopheles funestus* mosquitoes in Kenya after a long period of no apparent o'nyong-nyong virus activity. In 1996 to 1997, an outbreak occurred in Uganda. O'nyong-nyong is closely related to chikungunya and produces a similar illness, although fever is less pronounced and lymphadenopathy is more common. *Anopheles funestus* and *A. gambiae* transmit the virus.

Sindbis

Sindbis virus is widely distributed in Africa, India, tropical Asia, Australia, and northern Europe but is only rarely associated with human disease. The clinical features include fever, rash, arthralgia, myalgia, malaise, and headache. The fever, if present, is not high. The maculopapular rash progresses from trunk to extremities and vesicles can occur on the palms and soles. Virus has been isolated from vesicle fluid.

In northern Europe, symptomatic disease is recognized from Sweden, through Finland, to the former Karelian SSR, where it is known as Ockelbo disease, Pogosta disease, or Karelian fever, respectively. Prominent rheumatic complaints, sometimes persisting for several years, have been noted in Europe and South Africa. The virus has been isolated most often from ornithophilic *Culex* mosquito species. High antibody prevalences in Africa suggest that human exposure is common. Several outbreaks have been noted.

Further reading

Gubler DJ, Roehrig JT (1998). Arboviruses (Togaviridae and Flaviviridae). In: Collier L, *et al*., eds. *Topley and Wilson's microbiology and microbial infections*, 9th edn, Vol. 1, Virology, Ch. 29, pp. 579–600. Arnold, London.

Johnston RE, Peters CJ (1996). Alphaviruses. In: Fields BN, Knipe DM, Howley PM, *et al*., eds. *Fields virology*, 3rd edn, pp. 843–98. Lippincott-Raven, Philadelphia.

7.10.12 Rubella

P. A. Tookey and S. Logan

Introduction

Rubella infection usually causes a mild exanthematous disease of little clinical significance. However, in early pregnancy, infection may result in multiple congenital abnormalities. As a result of the widespread use of rubella vaccine, congenital rubella is now uncommon in most countries with developed health services.

The enveloped RNA virus of rubella is classified in its own genus, *Rubivirus*, within the family Togaviridae. Little variation has been detected among rubella isolates.

Epidemiology

Humans are the only known host for rubella virus. In temperate zones the infection is seen predominantly in spring and early summer. Before the introduction of rubella vaccine, rubella was endemic in virtually all countries. Epidemics were superimposed on the endemic infection every 4 to 9 years, and pandemics every 10 to 30 years. In most populations, in the absence of a mass immunization programme, around 10 to 20 per cent of women reach child-bearing age still susceptible to rubella infection. Infection is rare in infancy, incidence rises slowly in early childhood and then rapidly, peaking between 5 and 9 years of age.

Postnatally acquired infection

The rash usually begins on the face and spreads to the trunk and then the extremities; the pink maculopapular lesions are initially discrete but later tend to coalesce. The suboccipital and posterior cervical lymph nodes are characteristically enlarged. Mild fever, sore throat, coryza, cough, and conjunctivitis may be present; symptoms are usually mild and last 3 to 7 days. There may be a prodrome with malaise and fever, especially in adults. There is no specific treatment.

Arthralgia is a common complication in older patients, but frank arthritis is unusual; both are normally transient but recurrent or persistent symptoms have been reported, mainly in women. Less common complications include purpura, thrombocytopenia, postinfectious encephalitis, transverse myelitis, and the Guillain-Barré syndrome.

Rubella is clinically indistinguishable from a number of other infections and at least half of all infections are clinically inapparent or non-specific; a history of clinically diagnosed rubella infection is unreliable.

The incubation period is 14 to 21 days. The exact mode of transmission is uncertain but airborne spread by the respiratory route is likely and close contact is usually necessary for transmission. Individuals are infectious from about 5 to 7 days before to 3 to 5 days after the start of symptoms. Infectivity is highest immediately before, and on the first day of, symptoms. Congenitally infected infants shed large amounts of virus from the oropharynx and may be a source of infection for many months.

Infection usually produces lifelong immunity but reinfection has occasionally been reported.

Congenital infection

Congenital rubella is typically associated with cataracts, cardiac anomalies, and sensorineural hearing loss. The teratogenic effects may result in a wide range of defects (Table 1), but sensorineural hearing loss alone or combined with other abnormalities is most common. The earlier the stage of pregnancy at which infection occurred, the more likely it is that the child will have severe, multiple problems. The risk of damage following primary maternal infection in the first 10 weeks of pregnancy is around 90 per cent; this drops rapidly thereafter, and after 16 weeks' gestation even sensorineural hearing loss and growth retardation are rare; no abnormalities have been demonstrated following serologically confirmed maternal infection after 18 weeks' gestation.

Some defects, particularly sensorineural hearing loss, may not develop or reveal themselves until later infancy or childhood. Other reported late-onset problems include diabetes mellitus, thyroid dysfunction, and possibly autism and other behavioural and psychiatric disorders. A rare progressive rubella panencephalitis has been reported.

Maternal rubella infection is not always transmitted to the fetus. Transmission is most likely during the first trimester, the rate then declines until the last few weeks of pregnancy when it rises again. Most prospective studies of the risk to the fetus have been carried out on women with symptoms, but asymptomatic primary infection is thought to carry a similar risk. The risk of transmission following maternal reinfection in the first trimester is estimated to be about 8 per cent, but the likelihood of damage is low. Symptomatic maternal reinfection is rare, but in these circumstances viraemia is more likely and the risk to the fetus may be greater.

The diagnosis of congenital rubella infection is relatively easy if suspected early, but more difficult to confirm later. Virus can be isolated or detected by polymerase chain reaction (PCR) from multiple sites including the oropharynx, urine, and conjunctival fluid during the first months of life; viral shedding occasionally persists for years, but only about 10 per cent of infants are still shedding virus at 12 months. The presence of rubella IgM antibody in early infancy is virtually diagnostic of congenital infection because acquired infection is rare at this age. The presence of IgG antibody alone is not diagnostic since it is likely to indicate passively transferred maternal antibody, but persistence of IgG beyond 6 months is strongly suggestive of congenital infection. When abnormalities present late, a presumptive diagnosis can be made based on a compatible clinical picture and the presence or persistence of rubella IgG antibodies in a young child who has not yet been vaccinated.

Management of rubella-like illness during pregnancy

Routine antenatal rubella testing is not designed to identify infection in pregnancy, and specific diagnostic investigations are needed. Pregnant women with a rubella-like rash should be investigated simultaneously for rubella and parvovirus B19, since they are clinically indistinguishable. Even women previously reported to be immune should be investigated in case of laboratory error or reinfection. Blood should be tested for IgG and IgM antibodies, with a repeat test after 2 weeks if the results are equivocal. Rising IgG or detectable IgM antibody indicates recent infection. Investigations must be done in consultation with a virologist, who should be aware of the date and type of contact, stage of pregnancy, and history of previous immunization and testing.

Vaccination

Three strains of live, attenuated rubella vaccine were licensed in 1969. The RA27/3 strain is commonly used. Protective antibody levels are produced in around 95 per cent of recipients; protection is probably lifelong in most individuals.

In children, rubella vaccine causes few side-effects. Low-grade fever and rash are occasionally reported, and transient arthralgia has been seen in about 3 per cent of vaccinees; there have also been rare reports of myositis and vasculitis. Joint symptoms are common in adult women, affecting up to 40 per cent of vaccinees. They are less frequent and less severe than following naturally acquired rubella infection. Symptoms are generally mild and transient but a handful of cases of apparently

recurrent or persistent arthritis after rubella immunization have been described.

Different vaccine strategies have been pursued. In some countries, including the United States, there was mass immunization of all children early in their second year in an attempt to eliminate rubella from the community. This strategy not only protects those who are vaccinated but also reduces the risk of infection in susceptible pregnant women. However, if there is low vaccine uptake in childhood the spread of wild virus is slowed down but not eliminated: the effect is to push up the peak age of incidence of infection, which could, paradoxically, increase the number of congenital rubella cases.

In the United Kingdom, concern about low vaccine uptake and the duration of vaccine-induced immunity led initially to the adoption of a selective strategy of immunizing schoolgirls after the age of peak incidence. Routine testing and immunization of nursing and other staff who might be in contact with pregnant women was introduced, as well as antenatal screening for rubella susceptibility, with postpartum vaccination of susceptible women to protect subsequent pregnancies. The continued circulation of wild virus ensured that most women were protected by natural immunity acquired during childhood, with most of the remainder being covered by the schoolgirl and adult immunization programmes.

A third method was implemented in Sweden in 1982, when measles, mumps, and rubella (**MMR**) vaccine was introduced for all 1-year olds with a second dose at 12 years in order to reach those who did not receive vaccine previously or failed to respond, and to boost antibody status in the rest.

In 1988, in the United Kingdom, reassuring data on the persistence of immunity after vaccination and high vaccine uptake levels led to the introduction of mass immunization for all children in the second year of life with MMR; in 1996 this was supplemented by a preschool booster, and the schoolgirl vaccination programme was abandoned. Antenatal screening for rubella susceptibility continues, but the delivery of postpartum vaccination is not routinely monitored. Uptake of MMR by the age of 24 months averaged about 92 per cent between 1988 and 1992, but declined to about 88 per cent by 2000. Although the circulation of wild virus has dropped to very low levels since MMR was introduced, if the decline in vaccine uptake is not reversed it is possible that outbreaks of rubella could occur once again, putting susceptible pregnant women at risk.

Vaccination has led to dramatic declines in the numbers of susceptible pregnant women, rubella-associated terminations, and children born with congenital rubella. In the United Kingdom less than five congenitally infected infants were reported on average each year between 1990 and 2000, compared with 58.5 per year in the 1970s (when diagnostic methods and case ascertainment were less efficient). Terminations of pregnancy for rubella disease or contact averaged 612 a year in England and Wales during the 1970s, but during the 1990s the annual average was less than 10.

The strategy to be adopted in any country seeking to control congenital rubella by vaccination must depend on the projected uptake of vaccination and the long-term prospects for continuing the programme. An important element should be the screening and immunization of susceptible health personnel, particularly those in contact with pregnant women.

Vaccination in pregnancy

There have been persistent concerns that the vaccine virus might be teratogenic if given during pregnancy. Although vaccinees cannot infect other susceptible individuals, the virus can cross the placenta. Data pooled from studies of children born to several hundred women inadvertently vaccinated up to 3 months before conception or during pregnancy show less than 3 per cent with serological evidence of congenital infection, and no reported case of abnormalities attributable to congenital rubella. Over 80 of these infants were born to women vaccinated in the month of conception, probably the period of greatest vulnerability. These data suggest that the likely maximum theoretical risk is less than 5 per cent.

Further reading

Banatvala JE, Best JM (1998). Rubella. In: Mahy BWJ, Collier L, eds. *Topley and Wilson's microbiology and microbial infections: virology*, 9th edn, pp 551–77. Arnold, London.

Cooper LZ, Preblud SR, Alford CA (1995). Rubella. In: Remington JS, Klein JO, eds. *Infectious diseases of the fetus and newborn infant*, 4th edn, pp 268–311. W.B. Saunders, Philadelphia.

Miller E (1990). Rubella infection in pregnancy. In: Chamberlain G, ed. *Modern antenatal care of the fetus*, pp 247–70. Blackwell Scientific, Oxford.

Miller E *et al*. (1997). The epidemiology of rubella in England and Wales before and after the 1994 measles and rubella vaccination campaign. Fourth joint report from the PHLS and National Congenital Rubella Surveillance Programme. *Communicable Disease Report* **7**, R26–32.

7.10.13 Flaviviruses

L. R. Petersen and D. J. Gubler

Introduction

The genus Flavivirus of the family Flaviviridae comprises 68 registered viruses; 35 of which cause natural human infection (Table 1). Flaviviruses are small (37–50 nm), spherical particles whose genome is a molecule of single-stranded, positive-sense RNA approximately 11 000 nucleotides in length. Based on epidemiological and phylogenetic characteristics, the flaviviruses are classified into three groups: those that are mosquito-borne, tick-borne, and those in which no arthropod vector has been demonstrated. All flaviviruses of human importance belong to the first two groups; the last group contains a few viruses found in vertebrates.

Most flaviviruses are maintained in nature in complex transmission cycles between wild or domestic animals and one or more haematophagous arthropod vectors. Humans become infected from infected arthropod vectors that take a bloodmeal, but for most of the flaviviruses, humans do not usually develop high viraemias and are not thought to contribute to the transmission cycle. However, some flaviviruses of world-wide importance, including the dengue and yellow fever viruses, do produce high-level viraemias in humans and can be maintained in urban settings through a mosquito–human–mosquito transmission cycle.

The epidemiology and geographical distribution of the flaviviruses depend on several factors, including the presence of suitable amplifying hosts, the presence and feeding behaviour of a suitable arthropod vector, and the frequency of exposure of non-immune reservoir hosts and humans to infected vectors. Globalization of trade and travel, human population growth, and neglect of mosquito control programmes have produced conditions conducive for increasing incidence and geographic expansion of the flaviviruses. A recent dramatic example is the introduction and subsequent spread of the West Nile virus in the western hemisphere. Flavivirus infections are not communicable.

Flavivirus infection in humans can result in asymptomatic infection or a spectrum of clinical illness ranging from non-specific febrile illness, fever with rash or arthralgia or both, haemorrhagic fever, hepatitis, encephalitis, and death. The same virus can cause a variety of syndromes, and often the majority of those infected are asymptomatic. Although no specific therapy is available, prompt supportive treatment and proper management may substantially reduce mortality from some flavivirus infections. Ribavirin has been shown to have antiviral activity against several RNA viruses, but it has low *in vitro* and *in vivo* activity against flaviviruses.

Laboratory diagnosis

All flaviviruses have common group epitopes on the envelope protein that result in extensive cross-reactions in serological tests. The specificity of antibody detected should therefore be confirmed by cross-neutralization tests in areas where multiple flaviviruses are endemic/enzootic.

The IgM antibody capture enzyme-linked immunosorbent assay (MAC-ELISA) is widely used for diagnosis of flaviviruses. IgM antibody is usually detectable 5 to 8 days after infection. Because detectable IgM antibody persists for one or more months after infection with most flaviviruses, its presence is not confirmatory of current infection. Therefore, people with detectable IgM antibody are considered presumptive cases. Confirmatory laboratory diagnosis of most flaviviruses requires isolation of the virus, detection of specific viral RNA or specific antigen in a clinical sample, or virus-positive immunohistochemistry in autopsy tissues. A four-fold or greater rise in specific neutralizing antibody is confirmatory in some infections.

Important mosquito-borne flavivirus infections

Dengue and dengue haemorrhagic fever

Aetiology and epidemiology

There are four closely related, but serologically distinct dengue viruses, called DEN-1, DEN-2, DEN-3, and DEN-4. Since there is only transient, weak cross-protection among the four serotypes, persons living in an area of endemic dengue can be infected with three, and probably four, dengue serotypes during their lifetime.

Dengue fever is the most widespread and has the highest incidence of all the flaviviruses; with an estimated 50 to 100 million infections, and 200 000 to 500 000 cases of dengue haemorrhagic fever per year throughout most tropical regions of the world depending on epidemic activity (Fig. 1). The case fatality rate of dengue haemorrhagic fever cases averages 5 per cent. Over 2.5 billion people live in areas where dengue is endemic. The transmission cycle of most importance is the *Aedes aegypti*–human–*Ae. aegypti* cycle in large urban centres of the tropics. Multiple virus serotypes often cocirculate within the same city (hyperendemicity), causing periodic epidemics, especially in south-east Asia.

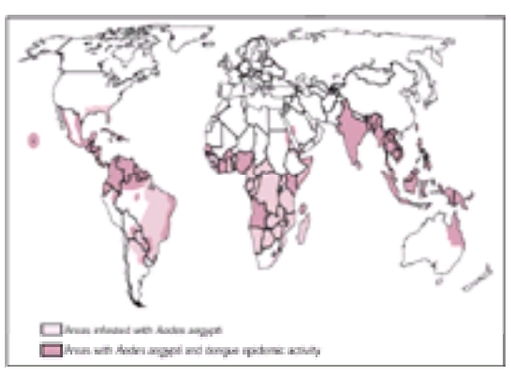

Fig. 1 The geographical distribution of dengue fever.

Humans are infected with dengue viruses by the bite of an infective mosquito. *Ae. aegypti*, the principal vector, is a highly domesticated tropical mosquito that lays its eggs in artificial water containers commonly found in and around homes. The adult mosquitoes rest indoors and prefer to feed on humans during daylight hours, with peak biting activity in the early morning or late afternoon. The adult female mosquitoes are nervous feeders and, if their feeding is interrupted, will return to the same person or different persons to continue feeding. Thus, during a single blood meal, several persons may become infected, making *Ae. aegypti* a highly efficient epidemic vector.

Clinical characteristics

Dengue virus infection in humans causes a spectrum of illness ranging from inapparent or mild febrile illness to severe and fatal haemorrhagic disease. The incubation period averages 4 to 7 days, with a range of 3 to 14 days. There are two major clinical syndromes: classic dengue fever and dengue haemorrhagic fever.

Dengue fever

Classic dengue fever, primarily a disease of older children and adults, is characterized by sudden onset of fever, frontal headache, retro-orbital pain, lumbrosacral pain, severe malaise, myalgias, bone pain, nausea, and vomiting. Patients may be anorectic, have altered taste sensation, and have a mild sore throat. Initial temperature may rise to 41°C and fever may last 2 to 7 days; a relative bradycardia may be present. Up to half the patients may have a transient rash or skin mottling in early illness. Defervescence occurs between days 3 and 6. A second, non-pruritic skin rash, varying in form from scarlatiniform to maculopapular and usually lasting an average of 2 to 3 days, may appear on the trunk, spreading to the face and extremities, sparing palms and soles, and often resolving with desquamation. The fever may rise again (saddleback pattern) with the presence of the rash.

Other findings associated with dengue fever include generalized lymphadenopathy and haemorrhagic manifestations. Skin haemorrhages, including petechiae and purpura, are the most common, along with gum bleeding, epistaxis, menorrhagia, and gastrointestinal haemorrhage. Haematuria is uncommon and jaundice is rare. Clinical laboratory findings include a neutropenia followed by a lymphocytosis, often marked by atypical lymphocytes. Liver enzyme levels may be mildly elevated, but in some patients serum alanine aminotransferase and aspartate aminotransferase levels reach 500 to 1000 U/l. Thrombocytopenia is not uncommon.

Classic dengue fever is rarely fatal, but weakness and depression may last several weeks following the acute illness, particularly in adults. There are no permanent sequelae.

Dengue haemorrhagic fever

During the last 50 years, dengue haemorrhagic fever has been recorded in south-east Asia, where all four dengue serotypes have circulated, but the first epidemic of dengue haemorrhagic fever in the Americas appeared in 1981 in Cuba, associated with the arrival of a new Asian virus of different genotype.

Pathogenesis

The characteristic development of dengue haemorrhagic fever during a heterotypic dengue infection, is presumed to result from pre-existing cross-reactive immunity. Subneutralizing levels of dengue antibodies enhance the infectivity of the virus by increasing the efficiency of binding and uptake of virus–antibody complexes through Fc receptors on blood monocyte or tissue macrophage cells, the chief site of replication in humans. This phenomenon of antibody-dependent enhancement is demonstrable *in vitro.* Cellular components of the primed immune response are probably also involved, and the resulting complement activation and release of cytokines and other mediators cause capillary leakage and coagulopathy, the hallmark of dengue haemorrhagic fever. The contribution of viral factors and variations in the host's immune response to the development of dengue haemorrhagic fever is unknown.

Clinical features

Although dengue haemorrhagic fever can occur in adults, it is primarily a disease of children under the age of 15 years. It is associated with all four virus serotypes. Dengue haemorrhagic fever is characterized by sudden onset of fever, usually lasting 2 to 7 days, and non-specific signs and symptoms. During this acute phase, it is difficult to distinguish dengue haemorrhagic fever from dengue fever and other illnesses found in tropical areas. However, at the critical time of defervescence, signs of circulatory failure or haemorrhagic manifestations may occur, the most common being petechiae, purpuric lesions, and ecchymoses. Epistaxis, bleeding gums, gastrointestinal haemorrhage, and haematuria occur less commonly. In its most severe form, dengue shock syndrome, occurring in approximately one-third of dengue haemorrhagic fever cases, patients experience hypotension, narrowing of the pulse pressure (£ 20 mmHg), and circulatory failure. Patients may complain of abdominal pain shortly before the onset of shock.

Physical findings associated with dengue shock syndrome include cool, blotchy, and congested skin, cyanosis, diaphoresis, tachypnia, and oliguria. Hepatomegaly, pulmonary effusion, and oedema are common, as is leucopenia. By definition, patients with dengue haemorrhagic fever/dengue shock syndrome must have objective evidence of plasma leakage such as a haemoconcentration (haematocrit elevated by 20 per cent), and thrombocytopenia with a platelet count of 100 000/mm ^{3} or less. Liver enzymes in serum may be elevated. The duration of shock is usually short. The mortality rate in aggressively treated patients is 1 per cent or less.

Differential diagnosis and diagnosis

The differential diagnosis during the acute phase of illness includes influenza, measles, rubella, typhoid, leptospirosis, rickettsia, malaria, and other arboviral infections with rash. Other viral haemorrhagic fevers and meningococcaemia should be considered in patients with haemorrhagic manifestations.

A definitive diagnosis depends on isolating the virus, detecting viral antigen or RNA in serum or tissues, or a rising titre of specific antibodies. The MAC-ELISA is the most widely used serological test for dengue diagnosis. Because antidengue IgM antibodies persist for several months, and because not all patients have detectable IgM antibodies 6 to 10 days after onset, diagnosis based on a single IgM antibody MAC-ELISA result should be considered provisional.

Prevention and control

Patients should be protected from mosquitoes and people should avoid mosquito bites in areas infested with *Ae. aegypti.* There are currently no licensed vaccines. Prevention and control can only be achieved by controlling *Ae. aegypti* in the large urban centres of the tropics. The most effective way to achieve this is via larval mosquito control by elimination of breeding sites in open stagnant water around communities where transmission is endemic/ epidemic.

Treatment

Supportive care includes intensive monitoring of vital signs and haematocrit. If signs of shock appear, prompt replacement of plasma volume and correction of metabolic acidosis, electrolyte imbalance, and hypoglycaemia are essential. After 1 or 2 days, the capillary leakage ceases and resorption of extravasated fluid begins. Care must then be taken not to induce pulmonary oedema by excessive intravenous fluids. Treatment with corticosteroids does not reduce mortality in children with dengue shock syndrome. Aspirin (salicylic acid) should be avoided.

Japanese encephalitis

Aetiology and epidemiology

Japanese encephalitis virus is the type species of the Japanese encephalitis serocomplex which includes several antigenically related viruses, including St Louis encephalitis, Rocio, West Nile, Koutango, Usuto, Murray Valley encephalitis, Kunjin, Alfuy, Stratford, and Kokobera viruses. Sequence analysis of the structural proteins suggests there are several genotypes of Japanese encephalitis, in distinct geographical areas.

Japanese encephalitis has a widespread distribution throughout Asia which has expanded in the past 20 years, with outbreaks in the Pacific, Australia, Nepal, and

Western India (Fig. 2). It is the most important cause of arboviral encephalitis, with about 45 000 cases reported annually. The highest incidence is in temperate and subtropical regions of China, northern Thailand, Nepal, and India. The virus is maintained in a cycle involving culicine mosquitoes and water birds and is transmitted to humans by *Culex* mosquitoes, primarily species of the *Cx. tritaeniorhynchus* complex, which breed in rice fields. Pigs are the primary amplifying host in the peridomestic environment. Epidemics occur in late summer in temperate regions and throughout the year in some tropical areas of Asia. Children have the highest attack rates, because of cumulative herd immunity with age.

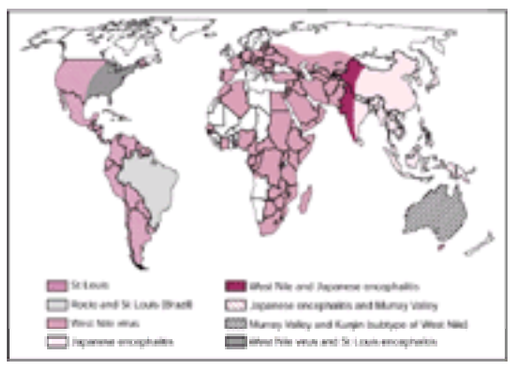

Fig. 2 The geographical distribution of the Japanese encephalitis serocomplex of the family Flaviviridae.

Clinical characteristics (see also Chapter 24.14.2)

Only about one in 250 infections result in symptomatic infection, which ranges from a febrile illness with headache, to aseptic meningitis, to encephalitis, and death. After an incubation period of 6 to 16 days, illness usually begins with a prodrome lasting several days followed by abrupt onset of high fever, change in mental status, nausea and vomiting, headache, and seizures, which occur in more than three-quarters of all paediatric patients. Generalized weakness and changes in tone, especially hypertonia and hyper-reflexia are common, but focal motor deficits, including cranial nerve palsies, paresis, hemiplegia, or tetraplegia, may also occur. Respiratory dysregulation, coma, abnormal plantar reflexes, and prolonged convulsions are associated with a poor prognosis.

Laboratory examination often reveals a moderate, peripheral leucocytosis and mild anaemia. Hyponatraemia, reflecting inappropriate antidiuretic hormone secretion, is common. Cerebrospinal fluid pressure is usually normal, pleocytosis ranges from a few to several hundred cells per mm ^{3}, and cerebrospinal fluid protein is moderately elevated in about half the cases.

Five to 30 per cent of cases are fatal; young children are more likely to die, and if they survive, are more likely to have residual neurological defects. Overall, up to 70 per cent of survivors have residual neurological abnormalities including parkinsonism, paralysis, behavioural changes, and psychological deficits. Evidence suggests that infection fails to clear in some patients, with clinical relapse several months after resolution of the acute illness. The clinical effects of congenital infection are unknown. Spontaneous abortions of women infected in the first and second trimesters have been reported.

Diagnosis

The differential diagnosis includes other viral encephalitides including arboviruses, herpes, and enteroviral infections, cerebral malaria, and bacterial infections. Epidemiological features such as place of residence or travel, season, and occurrence of other cases in the community provide clues to the diagnosis. Patients with encephalitis are rarely viraemic. Specific IgM can be detected in cerebrospinal fluid, serum, or both in nearly all patients by the seventh day after onset. Confirmation can be obtained by demonstrating four-fold or greater changes in specific IgM or neutralizing antibody titre.

Prevention and control

A formalin-inactivated mouse brain vaccine is used widely in Japan, Korea, Taiwan, Thailand, and other countries in Asia for childhood immunization and is licensed in Britain, the United States, and other developed countries to protect travellers. Hypersensitivity reactions to this vaccine, including generalized urticaria and angioedema, have occurred within minutes to as long as 2 weeks following vaccination at a rate of 1 to 104 per 10 000. Tissue culture based-vaccines (inactivated and live-attenuated) have been used in China. The risk to travellers to endemic areas during the transmission season can reach 1 per 5000 per month of exposure; risk for most short-term travellers may be less than 1 per million. In general, vaccine should be offered to people spending a month or more in endemic areas during the transmission season, especially if travel includes rural areas. Water and crop management and animal husbandry have been used to decrease human exposure to mosquito bites in the peridomestic environment.

Treatment

No specific therapy is available, but supportive treatment can reduce morbidity and mortality. One uncontrolled study showed a beneficial effect of a-interferon. Dexamethasone did not prevent death caused by oedema-induced increases in intracranial pressure in patients with severe encephalitis.

St Louis encephalitis

Aetiology and epidemiology

St Louis encephalitis virus is prevalent throughout the western hemisphere from Canada to Argentina (Fig. 1). The natural transmission cycle involves wild birds and *Culex* mosquitoes. Although clinical illness has been sporadically reported throughout much of this region, the highest incidence occurs in North America during epidemics. Fewer than 100 human cases are generally reported annually; epidemics with hundreds of cases have occurred in North America every 10 to 20 years.

Clinical characteristics

The ratio of infection to clinical illness is high, ranging from 800:1 in children under 10 years to 85:1 in persons over 60 years. Illness ranges from fever with headache, to aseptic meningitis, to encephalitis, and death. Advanced age is the strongest risk factor for both symptomatic disease and severity of encephalitis. After an incubation period of 4 to 21 days, the typical presentation of encephalitis is fever, headache, chills, nausea, and dysuria. Within 1 to 4 days, central nervous system signs appear and may include meningism, tremor, abnormal reflexes, ataxia, cranial nerve palsies, convulsions (especially in children), stupor, and coma. Complications include bronchopneumonia, sepsis, stress ulcer, and pulmonary embolism. Recovery is usually complete, except that 10 to 25 per cent of very young infants have residual mental deficits, personality changes, muscle weakness, and paralysis. Underlying diseases such as hypertension, diabetes, and alcoholism affect the outcome. The case fatality rate is about 6 per cent overall, but is only 1 per cent of those under 5 years, as the disease is generally milder in children. Short-lived sequalae of nervousness, memory impairment, and headache occur uncommonly in older children and adults.

The peripheral leucocyte count, serum transaminases, and creatine phosphokinase may be elevated. Hyponatraemia due to the syndrome of inappropropriate antidiuretic hormone secretion may be noted in up to one-third of patients. The cerebrospinal fluid contains fewer than 500 cells/mm ^{3}, principally leucocytes.

Diagnosis

The differential diagnosis includes other viral encephalitides such as arboviruses, herpes, and enterovirus, as well as other bacterial and fungal infections of the central nervous system. Epidemiological features (residence, season of the year, and occurrence of other cases in the community) provide diagnostic clues. Because of serological cross-reactivity with other flaviviruses, positive serum samples should be subjected to cross-neutralization tests. From fatal cases, virus may be isolated

from brain tissue or demonstrated by immunofluorescence. Virus has not been isolated from the blood during the acute phase of illness.

Prevention and control

No vaccine is available. Prevention is aimed at personal protection from mosquito bites and mosquito abatement.

Treatment

Treatment is supportive; no specific therapy is available.

West Nile encephalitis (see also Chapter 24.14.2)

Aetiology and epidemiology

West Nile virus is enzootic in Africa, the Middle East, and western Asia, and has caused periodic outbreaks in humans and horses in southern and central Europe (Fig. 2). In 1999, the virus was first detected in the New World during an outbreak in New York City. By 2001 the virus' distribution expanded to the entire eastern half of the United States and southern Ontario. Continued geographic expansion throughout the Americas is likely.

Phylogenic studies indicate two viral lineages: lineage one includes most strains isolated in recent outbreaks in Europe, the Middle East, and North America; lineage two includes many of the strains enzootic in Africa. Kunjin virus (see below) is a variant of West Nile virus and fits within lineage one. West Nile virus is maintained in a cycle involving culicine mosquitoes and wild birds. Bird migration may be important for transporting the virus geographically, particularly to temperate areas.

From the 1950s to the 1970s, epidemics, rarely associated with severe neurological disease and death, occurred in Israel, France, and Africa. No epidemic activity was then reported until the mid-1990s when epidemics associated with severe neurological disease and death in humans and/or equines were recorded in Algeria, Morocco, Tunisia, Italy, Romania, Israel, southern Russia, France, and the United States. Several of these epidemics have included hundreds of cases. Outbreaks typically occur in late summer and early autumn in temperate regions.

Clinical characteristics

Illness ranges from fever with headache, to aseptic meningitis, to encephalitis, and death; but most infections are asymptomatic. Investigations in Romania and the United States showed that less than one per cent of those infected developed meningitis or encephalitis requiring hospitalization. Approximately 20 per cent of those infected developed a systemic febrile illness. Advanced age was the most important risk factor for developing both encephalitis and death. The incubation period is 3 to 7 days (range 3–15 days). Mild illness presents as a dengue-like illness with fever, headache, backache, myalgia, and anorexia that lasts 3 to 6 days. A roseolar or maculopapular rash occurs in about half the patients and lasts up to a week without scaling. Generalized lymphadenopathy has also been reported as a common finding. In recent outbreaks, however, only about one in five patients had a rash and lymphadenopathy was uncommon. Disorientation, disturbed consciousness, and generalized weakness are predominant signs in persons with encephalitis. Ataxia, extrapyramidal signs, hypotonia, hyper-reflexia, coma, and seizures may be present. Motor weakness may be profound, suggesting Guillain–Barré syndrome. Other serious, non-neurological, rare complications include myocarditis, pancreatitis, and fulminant hepatitis.

Recovery is usually complete, but among those with severe encephalitis, up to half of survivors have residual neurological deficits. Case fatality rates in recent large outbreaks have ranged from 5 to 14 per cent. The peripheral leucocyte count, serum transaminases, and alkaline phosphatase may be elevated. Thrombocytopenia and anaemia may be present. Hyponatraemia can occur in up to three-quarters of patients. The cerebrospinal fluid usually contains up to 2000 cells/mm ^{3}, mostly leucocytes, and moderately elevated protein.

Diagnosis

The differential diagnosis includes other viral encephalitides including arboviruses, herpes, and enterovirus, as well as other bacterial and fungal infections of the central nervous system. Guillain–Barré syndrome should be considered in patients with profound muscle weakness. Epidemiological features (residence, season of the year, and occurrence of other cases in the community) provide diagnostic clues. Because of serological cross-reactivity with other flaviviruses, positive samples should be tested by cross-neutralization. Virus isolation or nucleic acid amplification tests of acute-phase blood, cerebrospinal fluid, or brain biopsy samples may provide a definitive diagnosis, but will not detect virus in all patients.

Prevention and control

No vaccine is available. Prevention is aimed at surveillance, mosquito abatement, and taking personal precautions to avoid mosquito bites.

Treatment

Treatment is supportive; no specific therapy is available.

Yellow fever

Aetiology and epidemiology

Yellow fever was first described in the seventeenth century and was one of the great plagues of mankind for over 400 years. In 1900, mosquito transmission and the viral aetiology were proven. The virus was isolated in 1927 and a vaccine developed in 1937. The virus is present in tropical America and Africa, but does not occur in Asia. Epidemics still occur, especially in West Africa. Between 1986 and 1991, a series of outbreaks in Nigeria caused an estimated 100 000 cases (although only about 5000 were officially reported), with attack rates in affected areas of 30/1000 and case-fatality rates exceeding 20 per cent. In South America, the disease affects up to 300 people annually, principally young men working in forest areas exposed to *Haemagogus* mosquitoes breeding in tree holes (jungle yellow fever). Disease in unvaccinated travellers is rare; however, since 1996 four travellers died in the US and Europe of infection acquired in South America and Africa.

Yellow fever virus has two cycles of transmission: jungle yellow fever and urban yellow fever. The forest or jungle transmission cycle involves canopy-dwelling mosquitoes and monkeys. The urban cycle involves humans as the vertebrate host and *Ae. aegypti* as the principal vector. In the past 30 years, *Ae. aegypti* has reinvaded Central and South America and a small outbreak of urban yellow fever occurred in Bolivia in 1998. The American tropics currently have the highest risk of urban epidemics of yellow fever in over 50 years. Epidemics in Africa often occur in moist savannah regions, involving forest (sylvatic) or peridomestic *Aedes* mosquitoes and humans as viraemic hosts. In dry areas and urban centres, epidemic transmission occurs where water-storage practices breed domestic *Ae. aegypti*. Several hundred thousand people are infected annually; outbreaks are frequent.

Clinical features

Approximately 1 in 20 infections results in clinical disease with jaundice. In its classical form, disease occurs abruptly after an incubation period of 3 to 6 days. The initial phase ('period of infection') is characterized by viraemia, fever, chills, headache, lumbosacral pain, myalgia, nausea, and prostration. On examination, the patient may have a relative bradycardia, and conjunctival injection. Within several days, the patient may recover transiently ('period of remission'), only to relapse ('period of intoxication') with jaundice, albuminuria, oliguria, haemorrhagic manifestations (especially 'black vomit' haematemesis), delirium, stupor, metabolic acidosis, and shock. The prognosis in such cases is poor; 20 to 50 per cent die during the second week of illness.

Clinical laboratory tests reveal leucopenia, thrombocytopenia, hepatic dysfunction, and renal failure. The bleeding diathesis is caused by decreased synthesis of clotting factors and may be complicated by disseminated intravascular coagulation. Pathological findings include midzonal hepatic necrosis and eosinophilic degeneration of hepatocytes (Councilman bodies), possibly representing apoptosis, and acute renal tubular necrosis. Focal myocarditis, and brain swelling and

petechial haemorrhages contribute to pathogenesis. Recovery is complete, without postnecrotic hepatic cirrhosis.

Diagnosis

Exposure and travel history provide important clues to aetiology. The differential diagnosis includes viral hepatitis, leptospirosis, rickettsial infections, dengue haemorrhagic fever, Rift Valley fever, Ebola, and Crimean–Congo haemorrhagic fever. Serological cross reactions with other flaviviruses may complicate serology. Postmortem histopathological examination of the liver is diagnostic, with or without immunocytochemical staining for viral antigen. Liver biopsy should never be performed on living patients, as it may precipitate haemorrhage.

Treatment

Treatment is symptomatic. Intensive care requires prompt awareness and treatment of acidosis, shock, and metabolic imbalance. Patients with renal failure may require dialysis.

Prevention and control

The live, attenuated 17D vaccine, delivered as a single 0.5-ml subcutaneous dose, is highly effective, and has minimal side-effects. Immunity is probably life long, but for travel certification, revaccination is recommended every 10 years. People with documented egg allergy should not be immunized or should be skin tested with the vaccine. The vaccine must not be given to children under 6 months of age, in whom there is a risk of postvaccinal encephalitis, and it is best to delay vaccination until 9 months of age. On theoretical grounds, immunosuppressed patients (including those with clinical AIDS) should not be immunized. The immune response in HIV infected persons is impaired. No evidence of clinical congenital infection has been found. Immunization during pregnancy is contraindicated, but, if inadvertently performed, recipients should be reassured and followed. The immune response in pregnancy was found to be impaired. Fatal infection following vaccination with the 17D strain has been rarely reported. Effective control has been achieved by controlling the principal urban mosquito vector, *Ae. aegypti*. This approach would also prevent epidemic dengue fever/dengue haemorrhagic fever in tropical urban centres.

Other mosquito-borne infections

Kunjin

Genomic sequencing indicates that Kunjin virus is a variant of West Nile virus. It is found over most of tropical Australia and Queensland and has a similar transmission cycle involving birds and *Culex* mosquitoes. Infection is usually asymptomatic, but occasional cases of encephalitis have been reported. Infections are generally milder than with Murray Valley encephalitis and are not life threatening. Kunjin virus infections that are non-encephalopathic usually present with fever, often with polyarthralgia. Cases occur sporadically, with only nine reported from 1990 to 1998. Treatment is supportive; there is no vaccine.

Murray Valley encephalitis

Murray Valley encephalitis is enzootic in New Guinea, in northern Western Australia, and in the Northern Territory, and possibly in northern Queensland (Fig. 1). The virus has a transmission cycle involving birds and *Culex* mosquitoes and is transmitted to man by *Cx. annulirostris* mosquitoes from the end of March to early June. Only one in 1000 to 2000 infections results in clinical illness; of those that have neurological disease, approximately one-third are fatal and a quarter have residual neurological deficits. Clinical illness resembles Japanese encephalitis. Children and the elderly are at the highest risk. In 1974, the largest recorded epidemic involved 58 cases and 10 deaths; since then, sporadic cases have been identified. Serological diagnosis is complicated by the presence of the closely related Kunjin virus, which also causes encephalitis. Treatment is supportive; there is no vaccine.

Rocio encephalitis

Rocio virus is antigenically related to St Louis encephalitis virus and is known only in Brazil. Epidemics in 1975 to 1976 caused 871 cases, principally among young adult, male, agricultural workers and fisherman. Since then, only sporadic illness has been reported. The probable vector is *Aedes scapularis* and wild birds are amplifying hosts. The clinical disease is typical viral encephalitis, with a 4 per cent mortality and 20 per cent of patients have neuropsychiatric sequealae. Virus is not recoverable from blood, but post-mortem diagnosis may be made by virus isolation from brain tissue. Treatment is supportive; there is no vaccine.

Tick-borne infections of the central nervous system

Tick-borne encephalitis

Aetiology and epidemiology

There are two subtypes of tick-borne encephalitis virus, eastern and western, which differ only slightly in viral protein structure. These viruses, along with the louping ill, Powassan, Kyasanur Forest disease, and Omsk haemorrhagic fever viruses, belong to the tick-borne encephalitis antigenic complex. The disease caused by the eastern subtype is also known as Russian spring/summer encephalitis and Russian epidemic encephalitis; and the western subtype as FSME (Fruehsommer-Meningo-Enzephalitis), early-summer encephalitis, and Kumlinge's disease. The geographic distribution of disease is determined by that of their tick vectors: *Ixodes persulcatus* for the eastern subtype causing human disease principally in the Far East, the Urals, and western Siberia; and *Ixodes ricinus* for the western subtype which occurs at highest incidence in eastern and central Europe, Moldavia, the Ukraine, and Byelorus, with smaller numbers of cases from western Europe, the Balkans, and Scandinavia.

Infections occur during the period of tick activity from April to November. Tick-borne encephalitis is largely a rural infection; occupational and vocational pursuits favouring tick exposure are risk factors. Human infection and outbreaks following consumption of raw milk or cheese from asymptomatic goats, or more rarely, sheep or cows, have been described. Hundreds to thousands of cases occur annually, with reported attack rates up to 200/100 000 residents in Latvia, the Urals, and Western Siberia. Aerosolized virus has caused laboratory infections.

Clinical features

Most human infections are subclinical. The illness produced by each subtype is generally similar but that produced by the eastern subtype carries a worse prognosis. The incubation period is 7 to 14 days, with a range of 2 to 28 days. Incubation periods of 3 to 4 days follow milk-borne exposure. The Western subtype typically produces a biphasic illness. The first phase is a non-specific, influenza-like, febrile illness lasting 2 to 7 days followed by an afebrile and relatively asymptomatic period lasting 2 to 10 days. Flushing, conjunctival haemorrhage, nausea, vomiting, dizziness, and myalgia are common findings. Approximately one-third of patients then develop higher fevers with aseptic meningitis or meningoencephalitis. The eastern subtype usually progresses without an asymptomatic phase. Signs and symptoms of meningoencephalitis or meningoencephaloymelitis include somnolence, coma, asymmetrical parasesis of the cranial nerves, tremors of the extremities, nystagmus, severe pain in the extremities, and flaccid paralysis of the neck and upper extremities.

Permanent paralysis develops in 2 to 10 per cent of patients with the western subtype, 10 to 25 per cent with the eastern. Corresponding case fatality rates are 0.5 to 2.0 per cent and 5 to 20 per cent for the western and eastern subtypes, respectively.

Laboratory findings include neutrophilia, although neutropenia, thrombocytopenia, and elevated liver enzyme levels may occur early. The cerebrospinal fluid white blood count is usually below 500/mm^3, primarily of mononuclear cells.

Diagnosis

The differential diagnosis is similar to Japanese encephalitis; the pattern of flaccid paralysis may be confused with poliomyelitis. A history of bite by small ixodid ticks is elicited in fewer than half the cases. Specific diagnosis is made by virus isolation from blood or cerebrospinal fluid during the first week of illness, or by serological

tests, including IgM enzyme immunoassay and neutralization test.

Treatment

Treatment is supportive.

Prevention and control

Effective inactivated vaccines are available in Europe in formulations for adults and children. Mass vaccination in Austria produced a dramatic decline in disease incidence. Vaccines appear to produce equal protection against the eastern and western strains. Use of a vaccine, available in the United Kingdom, should be seriously considered for tourists planning camping or extensive outdoor activities during the tick transmission season in enzootic areas, particularly in Russia and Central Europe. Commercial hyperimmune globulin preparations are available for use after tick exposure or to provide short-term pre-exposure prophylaxis.

Louping ill

This is a disease of veterinary importance, causing neurological illness in sheep and to a lesser extent in cows, horses, farmed deer, sheep-dogs, and pigs. The virus, isolated in 1931, is a member of the tick-borne encephalitis complex, and is transmitted by *Ixodes ricinus*. Louping ill occurs in the hill country along the western coast of Scotland and northern England, Ireland, and Norway. Natural infections resulting in human disease have been rare, but laboratory infections are not uncommon. Ten naturally occurring cases have been documented, including a veterinarian, abattoir workers, and farmers. Some of these cases were attributable to contact with sheep blood. The human disease is typically aseptic meningitis or encephalitis; no fatal infections have occurred. Avoidance of tick bite in enzootic areas is recommended. The licensed tick-borne encephalitis vaccine may be protective.

Powassan encephalitis

The virus was first isolated from the brain of a fatal case in Powassan, Ontario in 1958. Since then, approximately 20 human cases have been recognized in eastern Canada and the eastern United States, primarily in children, with a case-fatality rate of 10 per cent and a high incidence of residual neurological dysfunction. Serological surveys indicate an antibody prevalence of 1 to 4 per cent. The distribution of the virus in North America is considerably wider than indicated by human cases, and the diagnosis should be suspected in any case of summer–autumn encephalitis. The virus is transmitted between *Ixodes scapularis* (ricinus complex) ticks and rodents. The clinical features are those of viral encephalitis, with localizing neurological signs and convulsions. There is no specific treatment or vaccine.

Tick-borne haemorrhagic fever

Kyasunur Forest disease

Aetiology and epidemiology

This virus is a member of the tick-borne encephalitis antigenic complex. The virus has been isolated from humans, monkeys, and ticks since it was first recognized in 1957 during an outbreak of haemorrhagic fever affecting wild monkeys in Karnataka (then Mysore) State, India. Several hundred cases are reported annually, principally among people working in the forest in Karnataka State. In 1983, 1555 cases, including 150 deaths, occurred. The peak seasonal incidence is between February and May. The virus is transmitted between immature ixodid ticks (*Haemaphysalis spinigera*) and small mammals (rodents, porcupines), passes to the adult stage during moulting, and is spread to man and wild monkeys by adult ticks.

In 1995, a subtype of Kyasanur Forest disease virus was isolated from patients in Jeddah, Saudi Arabia, with clinical symptoms ranging from febrile illness to fatal haemorrhagic disease. Human infections were associated with handling meat or drinking unpasturized camel's milk. The virus seems to be associated with sheep and camels, and to be transmitted by ticks.

Clinical characteristics

After an incubation period of 2 to 7 days, there is abrupt onset of fever, chills, headache, myalgia, abdominal pain, nausea, vomiting, and diarrhoea. Physical signs include bradycardia, lymphadenopathy, and haemorrhagic manifestations. Hypotension is frequently noted during the end of the acute stage. Fatal cases develop shock and pulmonary oedema. A biphasic illness is not uncommon, with resolution of the first phase in 5 to 12 days, and return of the fever and signs of meningoencephalitis after an interval of 1 to 3 weeks. Localizing neurological signs are infrequent, and residual defects are rare. Convalescence is prolonged. Laboratory abnormalities include leucopenia, thrombocytopenia, and elevated serum transaminases during the acute phase.

Diagnosis

Diagnosis is by virus isolation from blood collected during the first week after onset or by serological tests. Virus isolation should be conducted under biosafety level four conditions.

Prevention and control

Tick bites should be avoided in endemic areas. Proper care should be taken when handling raw meat, and camel's milk should be pasteurized. A formalin-inactivated virus is available in India.

Treatment

Treatment is supportive; specific therapy is not available.

Omsk haemorrhagic fever

This disease was first recognized in 1945 in western Siberia. Cases were frequent between 1945 and 1949, with morbidity rates of 500 to 1400/100 000, but subsequently have been rare, mainly occurring among residents of rural areas working in the fields. The virus is a member of the tick-borne encephalitis complex. Human infections are acquired by tick bite or contact with infected muskrats. The disease is characterized by abrupt onset of fever, headache, myalgia, facial flushing, conjunctival suffusion, minor haemorrhagic manifestations, and leucopenia. Recovery occurs in the second week, and the case fatality rate is low (0.5–3 per cent). The differential diagnosis includes tularaemia, rickettsial infection, and leptospirosis. Specific diagnosis is made by virus isolation from blood during the acute phase or by serological tests. Only a few laboratories outside Russia with biocontainment level 4 facilities are capable of providing laboratory assistance. Tick-borne encephalitis vaccines may cross-protect against Omsk haemorrhagic fever.

Further reading

Centers for Disease Control and Prevention (1993). Inactivated Japanese encephalitis virus vaccine. Recommendations of the Advisory Committee on Immunization Practices (ACIP). *Morbidity and Mortality Weekly Reports* **42**, 1–15.

Dumpis U, Crook D, Oski (1999). Tick-borne encephalitis. *Clinical Infectious Diseases* **28**, 882–90.

Gubler DJ (1998). Dengue and dengue hemorrhagic fever. *Clinical Microbiology Reviews* **11**, 480–96.

Gubler DJ, Roehrig JT (1998). Arboviruses (Togaviridae and Flaviviridae). In: Collier L, *et al*. *Topley and Wilson's microbiology and microbial infections*, 9th edn, Vol. 1, Virology, ch. 29, pp. 579–600. Arnold, London

Monath TP, Heinz FX (1996). Flaviviruses. In: Fields BN, Knipe DM, Howley PM, eds. *Fields virology*, pp. 961–1034. Lippincott-Raven, Philadelphia.

Solomon T, Dung NM, Kneen R, Gainsborough M, Vaughn DW, Khanh VT (2000). Japanese encephalitis. *Journal of Neurology, Neurosurgery and Psychiatry* **68**, 405–15.

7.10.14 Bunyaviridae

J. W. LeDuc and J. S. Porterfield

The family Bunyaviridae currently contains around 300 viruses, and is divided into five genera (see Table 1). The family name, and that of the genus *Bunyavirus*, is derived from the type species, Bunyamwera virus, which was isolated in Uganda from *Aedes* mosquitoes. The other genera are also named after viruses: the genus *Hantavirus* after Hantaan virus, the causative agent of Korean haemorrhagic fever; the genus *Nairovirus* after Nairobi sheep disease virus; the genus *Phlebovirus* after phlebotomus or sandfly fever virus; and the genus *Tospovirus* after tomato spotted-wilt virus. All members of the family share certain structural, biochemical, and genetic properties, such as a spherical, enveloped virion 80 to 120 nm in diameter (see Fig. 1), and a genome of single-stranded, negative-sense RNA divided into three segments. Members of different genera vary substantially in their biological and biochemical properties and in the details of their mechanisms of replication. Bunyaviruses, nairoviruses, and phleboviruses, which together make up the greater part of the family, are all arthropod-borne animal viruses, or arboviruses; these circulate in nature in a wide variety of different vertebrate hosts and are biologically transmitted between vertebrates and to humans by the bites of blood-sucking arthropods, principally mosquitoes for bunyaviruses, sandflies for phleboviruses, and ticks for nairoviruses. By contrast, hantaviruses are not arboviruses, but are zoonotic agents infecting rodents and other small mammals, which may spread to humans if they are in close contact with their infected excreta, and tospoviruses are arthropod-transmitted plant viruses of no known medical importance.

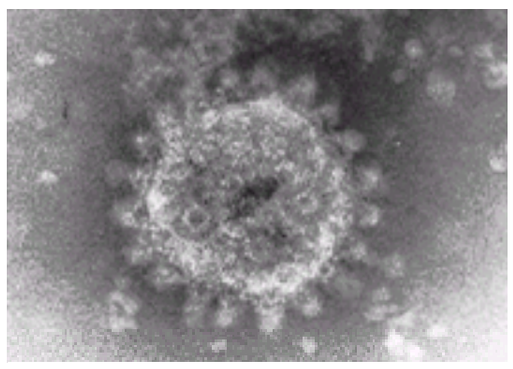

Fig. 1 Electron micrograph of Crimean–Congo haemorrhagic fever virus (× 400 000) (by courtesy of Dr D.S. Ellis).

Viruses within the larger genera are further subdivided into serogroups of more closely related members, there being at least 18 serogroups within the bunyaviruses, and seven within the nairoviruses (see Table 1). Of over 60 Bunyaviridae that are known to infect man, the type species and those that cause major human diseases are shown in bold type in Table 1 and are described in more detail in the following sections. Table 2 lists the continental distribution of the remaining viruses that cause only minor human infections and also indicates the principal arthropod vector of each virus. The habitats in which the different viruses and their vectors occur range from arctic to tropical, with every intermediate form. The enzootic cycles by which arboviruses are maintained in nature are very imperfectly understood; most viruses undergo alternate cycles of replication in vertebrate and invertebrate hosts, but transovarial and trans-stadial transmission within some mosquitoes, ticks, and phlebotomine flies, and venereal transmission from vertically infected male mosquitoes to uninfected females are also known to occur. Most arboviruses have a narrow host range, occur within a limited area, and are transmitted by specific vectors to a limited number of vertebrate hosts, but some viruses infect a wider host range, are transmitted by more than one type of vector, and may occur in more than a single continent. It is of interest that, for different members of the family, tick transmission predominates in Asia, but is unknown in South or Central America, and although some Bunyaviridae have been isolated in Australia, none is known to infect man in that continent. Further epidemiological details can be found in more specialized publications.

Following viral entry, whether through the skin after the bite of an infected arthropod or by another route, virus replicates in draining lymph nodes, which are frequently enlarged, and a viraemia follows. Symptoms develop when virus lodges in other sites and undergoes further replication cycles. Appropriate virucidal agents and methods include bleach, phenolic disinfectants and detergents, autoclaving or boiling, and the use of g-irradiation. Various enzymes such as nucleases will also inactivate these viruses. For human pathogens with the ability to spread by the aerosol route, biosafety level 3 (hantaviruses, Oropouche virus, others) or 4 (Crimean–Congo haemorrhagic fever virus only) is recommended. Added precautions are necessary when handling hantavirus-infected animals and virus concentrates.

Genus *Bunyavirus*

Much of our knowledge about the family as a whole derives from intensive studies on the type species, Bunyamwera virus, and a few other members of this large genus. The three-segmented genome permits reassortment when two closely related viruses infect the same cell, either in nature or under controlled laboratory conditions. Such studies have been used to establish the genomic control of viral proteins, and to analyse the basis of virulence for both vertebrate and invertebrate hosts. Two bunyaviruses, Akabane and Aino viruses in the Simbu serogroup, are notable for their ability to produce congenital deformities in sheep, goats, and cattle in Japan, Australia, Africa, and in the Middle East. However, there is as yet no evidence that any member of the genus or family produces teratogenic effects in man, although there is concern that Oropouche virus, an important Simbu serogroup pathogen of northern South America, may be a threat to pregnant women.

Bunyamwera virus

Symptoms

A mild, febrile illness, usually with headache, joint and back pains, sometimes associated with a rash, and occasionally with mild involvement of the central nervous system. Serological surveys indicated that infection of man is widespread in sub-Saharan Africa but most infections are unrecognized. Laboratory infections have been recorded. Garissa virus, recently isolated from haemorrhagic fever patients during outbreak investigations in Kenya and Somalia, contains L and S genome segments virtually identical to Bunyamwera virus, but with the M segment coming from a virus most closely related to Cache Valley virus. Neither Bunyamwera nor

Cache Valley virus is known to cause haemorrhagic disease in humans.

Treatment and prognosis

No treatment is necessary and the prognosis is excellent.

California encephalitis virus, Inkoo, Jamestown Canyon, La Crosse, Tahyna, and snowshoe hare viruses

The viruses named above, and perhaps others currently unrecognized, are responsible for the clinical condition known as California encephalitis. The viruses are widely distributed in nature throughout many parts of North America, Europe, and Eurasia. Most recognized human infections in the United States are due to La Crosse virus and are reported from Ohio, Wisconsin, Minnesota, and West Virginia; in 1999 a total of 70 cases was reported from nine states. The great majority of these occur in children, more often males than females, although Jamestown Canyon virus is unusual in that more adults are involved. There is nearly always a history of outdoor exposure in areas where woodland mosquitoes are prevalent. The incubation period is 5 to 10 days. Most cases of La Crosse encephalitis are relatively mild with headache, fever, and vomiting, progressing to lethargy, behavioural changes, and occasional brief seizures, followed by improvement. Severe cases (10 to 20 per cent) have an abrupt onset of fever and headache, disorientation, and seizures during the first 24 h of illness, sometimes progressing to coma and requiring intensive supportive care. Overall, symptomatic children suffer seizures in about 50 per cent of cases, status epilepticus in 10 to 15 per cent, and mortality approaches 1 per cent. Residual seizures occur in 6 to 13 per cent, persistent hemiparesis in about 1 per cent, and cognitive dysfunction in a small but poorly defined percentage of cases.

In Europe, Tahyna virus is widely distributed in Austria, the former Czechoslovakia, France, Germany, Italy, Norway, Romania, the former Yugoslavia, and the former USSR. Antibody rates can exceed 95 per cent in certain parts of the former Czechoslovakia, and are around 50 per cent in the Rhone valley in France and the Danube basin near Vienna; however, overt disease is seldom recognized. Inkoo virus is prevalent in Finland and in neighbouring regions of Russia, with the great majority of adult Lapps having antibodies; emerging information suggests that small children may have signs of central nervous involvement during acute infection. Antibodies reactive with California serogroup viruses have also been found in human sera collected in Sri Lanka, China, and in the far northern latitudes of Eurasia where a number of California serogroup viruses have been isolated from mosquitoes, some related to Inkoo and Tahyna viruses, but others to snowshoe hare virus. In another Russian study of some 50 persons, mainly 14 to 30 years of age, with infections caused by California serogroup viruses, about two-thirds had an influenza-like illness without central nervous involvement, while the remaining third had aseptic meningitis.

Control, treatment, and prognosis

Measures to limit mosquito breeding, particularly of *Aedes triseriatus*, are useful in endemic regions. No vaccines are available, and there is no specific treatment, although the fluid and electrolyte balance must be maintained, and anticonvulsive drugs may be required to control seizures. Intravenous ribavirin has been used to treat severe La Crosse encephalitis; however, more comprehensive clinical trials are needed.

Oropouche virus

Symptoms

Prior to 1961, Oropouche virus was known to have caused only a mild fever in a single forest worker in Trinidad, but that year it was responsible for a substantial epidemic in the Belem area of northern Brazil, with some 7000 individuals affected. Over the ensuing 40 years, massive epidemics of febrile illness have been recorded throughout the Amazon Basin, with perhaps as many as 200 000 persons infected. Symptoms include headache, generalized body pains, back pains, prostration, and moderately high fever (40°C). Rash occasionally accompanies infection, as does meningitis or meningismus. Illness lasts from 2 to 5 days, occasionally with protracted convalescence. No fatalities have been reported.

Control, treatment, and prognosis

No vaccine is available. Transmission is probably by the biting midge *Culicoides paraensis* and outbreaks appear to be a long-term consequence of agricultural development of the Amazon Basin. Accumulated organic waste from cacao and banana production provide ideal breeding sites for *Culicoides*, leading to massive populations and subsequent epidemic Oropouche disease. Thus, measures to reduce *Culicoides* breeding may be of benefit. Treatment is supportive, and the prognosis is good, although convalescence may be protracted.

Genus *Hantavirus*

Haemorrhagic fever with renal syndrome

The genus *Hantavirus* takes its name from Hantaan virus, the cause of Korean haemorrhagic fever in Korea. The name Hantaan in turn is from the Hantaan River near the demilitarized zone between North and South Korea, where the virus was first recovered from its rodent host, *Apodemus agrarius*. Hantaan virus was only isolated in 1976, although the clinical diseases it and related hantaviruses cause in the Eurasian continent have been known much longer under many different synonyms: epidemic haemorrhagic fever, Korean haemorrhagic fever, nephropathia epidemica, with haemorrhagic fever with renal syndrome preferred. Four distinct viruses are responsible for most recognized haemorrhagic fever with renal syndrome: Hantaan virus, found primarily in Asia; Dobrava virus, found in an enclave of disease in the Balkan region and sparsely elsewhere in Europe; Puumala virus, found in Scandinavia, western Russia, and much of Europe; and Seoul virus, probably globally distributed wherever *Rattus norvegicus* populations exist uncontrolled. Hantaan and Dobrava viruses cause severe, life-threatening disease with mortality of about 5 per cent, reaching as high as 30 per cent in select populations. Puumala virus infections are less severe, although patients still require admission to hospital, with death in less than 1 per cent of admitted cases. Seoul virus is thought to be the least severe of the pathogenic strains of hantaviruses, although it too has been associated with human deaths.

Each hantavirus is specifically associated with a particular rodent host in nature: Hantaan virus with the striped field mouse, *Apodemus agrarius*; Dobrava virus suspected with the yellow-necked mouse, *Apodemus flavicollis*; Puumala virus with the bank vole, *Clethrionomys glareolus*; and Seoul virus with the Norway rat, *Rattus norvegicus*. Human infection is from aerosols of infectious rodent excreta, or rarely by rodent bite, and is occupationally associated. Most disease is seen among adult men in rural environments. Those with occupations at greatest risk include farmers, woodcutters, shepherds, and especially the military in the field. Most hantavirus disease is markedly seasonal, with peak incidence seen in the late autumn and early winter, although the Balkan form is found most commonly during summer months in Greece and adjacent countries.

Symptoms

Incubation period for hantaviruses is rather variable, and may approach 2 months in some cases, but is generally 12 to 16 days. Severe disease, as typically associated with Hantaan or Dobrava virus infections in Asia or the Balkans, is characterized by five phases:

1. febrile, of 3- to 7-day duration;
2. hypotensive, lasting from a few hours to 3 days;
3. oliguric, from 3 to 7 days;
4. diuretic, from a few days to weeks;
5. a prolonged convalescence.

Characteristic signs and symptoms of the febrile phase include fever, malaise, headache, myalgia, back pain, abdominal pain, nausea and vomiting, facial flushing, petechias, and conjunctival haemorrhage (Fig. 2). The hypotensive phase is characterized by nausea, vomiting, tachycardia, hypotension, blurred vision, haemorrhagic signs, and shock, with approximately one-third of the deaths occurring during this phase. In the oliguric phase, nausea and vomiting may persist, and blood pressure may rise; kidney failure presents, which may include frank anuria; and about one-third of the cases may experience severe haemorrhage as epistaxis, gastrointestinal, cutaneous, or bleeding at other sites. Nearly one-half of deaths occur during the oliguric phase. In the diuretic phase, urine output increases to

several litres per day. Convalescence is protracted and may require months before full strength and function are regained.

Fig. 2 Patient with acute Korean haemorrhagic fever, showing extensive conjunctival haemorrhages (by courtesy of Professor H.W. Lee).

Less severe forms of the disease may skip phases, or spend less time in each phase. The milder forms of haemorrhagic fever with renal syndrome, such as nephropathia epidemica due to Puumala virus, follow a similar, but less severe course, with abrupt onset of fever of 38 to 40 °C, headache, malaise, backache, and generalized abdominal pain. Back or loin pain is especially common. Signs of renal failure are usually not as pronounced, and the need for renal dialysis varies. Transient blurred vision occurs in about 10 per cent of cases. Infection due to Seoul virus follows a similar course, but may present with more evidence of liver involvement. There is no evidence of person-to-person transmission.

Treatment and prognosis

Admission to hospital, avoidance of trauma and unnecessary movement, close observation, and careful supportive care are essential to patient survival. Treatment is phase specific, with special attention to fluid balance and volume, and control of hypotension and shock. Dialysis may be required in cases of acute renal failure. Specific antiviral therapy using ribavirin has been shown to be efficacious if started early in disease. Recovery is protracted, but until now considered complete and without permanent complications. Recent evidence, however, suggests that persons previously infected with Seoul virus may be at increased risk of chronic renal disease, hypertension, or stroke.

Hantavirus pulmonary syndrome

Recently, a new hantavirus disease, Hantavirus pulmonary syndrome, was reported from the United States and soon recognized in several South American countries as well. More than half the originally identified cases died in 1993, but mortality rates have declined to 20 to 40 per cent as clinical experience is gained. Cases were reported predominantly from the western United States and Canada, and more recently from Argentina, Chile, Brazil, and other South American countries. Sin Nombre virus was first associated with hantavirus pulmonary syndrome, but many additional hantaviruses have now been recognized as likely causes of this syndrome (Table 1). As Old World hantaviruses are generally associated with specific microtine rodents, so each American hantavirus appears to be associated with a specific sigmodontine host. A chain of apparent human-to-human transmission of Andes virus occurred during an outbreak in southern Argentina, including transmission to medical staff, suggesting that application of universal precautions when treating suspected cases of Hantavirus pulmonary syndrome may be warranted.

Symptoms

Hantavirus pulmonary syndrome is unusual in that symptoms are primarily those of acute unexplained adult respiratory distress syndrome, rather than renal disease. A nondescript prodrome of fever, myalgia, and malaise may last 4 to 6 days, with nausea, vomiting, and abdominal pain, often accompanied by dizziness. On admission, physical examination of patients with confirmed cases reveals fever (> 38 °C), tachycardia (> 100/min), tachypnoea (> 20/min), and often hypotension (systolic < 100 mmHg), and rales. Laboratory findings include hypoxia, leucocytosis, haemoconcentration, thrombocytopenia, atypical lymphocytosis, elevated serum lactate dehydrogenase and glutamic pyruvic transaminase, and prolonged prothrombin time (> 14 s). Chest radiography is helpful in diagnosis, noting progression from subtle interstitial findings to frank bilateral alveolar oedema; pleural effusions are usually present (Fig. 3). Thrombocytopenia and haemoconcentration are independent statistical predictors of Hantavirus pulmonary syndrome, although not present in every patient. Disease progresses rapidly once the lungs begin to fill, with death commonly seen in 24 to 48 h after admission, or sooner, due to hypoxia and/or circulatory failure. Hypotension and shock may occur independently in patients whose hypoxaemia is medically controlled.

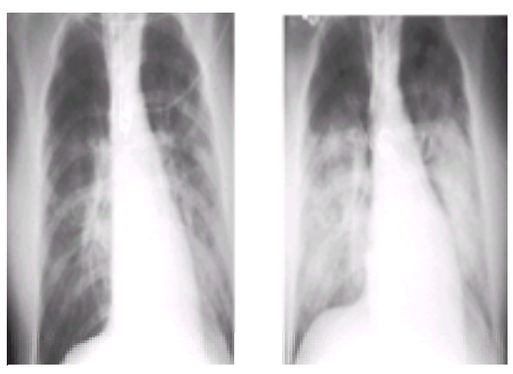

Fig. 3 Chest radiograph of patient with early Hantavirus pulmonary syndrome (L), and same patient 24 h later (R) showing development of bilateral perihilar alveolar oedema (by courtesy of Dr Loren Ketai).

Treatment and prognosis

Treatment is supportive, with careful management of hypoxia, fluid balance, and shock. About two-thirds of patients will require intubation and mechanical ventilation. Fluid loss into the lungs leads to haemoconcentration, but infusion of fluids exacerbates pulmonary oedema; consequently fluids should be administered with caution and careful monitoring. Limited experience in open label trials suggests that intravenous ribavirin does not have a marked effect on the course of Hantavirus pulmonary syndrome, perhaps because of the speed with which the disease progresses.

Control

Prevention involves avoidance of infected rodents, either through efficient rodent control programmes in urban settings for Seoul virus, or maintenance of clean campsites so that waste food is not allowed to accumulate and attract rodents. Vaccine development is under way, and nationally approved inactivated vaccines reported to be safe and efficacious against Hantaviruses are available for use in Asia.

Genus *Nairovirus*

The genus *Nairovirus* is named after Nairobi sheep disease, an acute, haemorrhagic gastroenteritis affecting sheep and goats in East Africa, with transmission by the

sheep tick, *Rhipicephalus appendiculatus*. In addition to the type species, which has caused laboratory infections, the genus also includes several other viruses known to infect man, of which the most important is Crimean-Congo haemorrhagic fever virus. Other nairoviruses causing less important human infections are Ganjam virus, almost indistinguishable from Nairobi sheep disease virus but first isolated in India from *Haemaphysalis intermedia* ticks collected from healthy goats; Hazara virus, recovered from *Ixodes redkorzevi* ticks collected from the vole *Alticola roylei*, in a subarctic habitat at an altitude of 3660 m in the Khaghan valley of Hazara district, Pakistan; Dugbe virus, isolated in Nigeria from *Amblyomma variegatum* ticks collected from healthy cattle; and Soldado virus, repeatedly isolated from a variety of bird ticks but recently linked to a mild illness in man.

Crimean–Congo haemorrhagic fever virus

Crimean haemorrhagic fever was first recognized as a cause of an acute, febrile, haemorrhagic disease affecting man in the Crimean region of the former USSR, transmitted by ticks and carrying a mortality of 15 to 30 per cent. In Africa, Congo virus was first isolated in the then Belgian Congo (now Democratic Republic of the Congo) from the blood of a local 13-year-old boy, and it caused a moderately severe laboratory infection; related viruses were isolated in Uganda, where more laboratory infections occurred, one of which ended fatally after a severe haematemesis. In Asia, a virus indistinguishable from Congo virus was isolated from pools of ticks collected from a variety of wild and domestic animals in Western Pakistan. It was later demonstrated that Crimean haemorrhagic fever virus was serologically indistinguishable from Congo virus, hence the use of the term Crimean–Congo haemorrhagic fever virus. Different strains of this virus have been associated with outbreaks of severe and sometimes fatal disease in the Crimea, Rostov, and Astrakhan regions of the former USSR, in Albania, Bulgaria, and Yugoslavia, in East, West, and South Africa, in Iran, Iraq, and in Western Pakistan, and in China. Most infections are acquired by tick bites, but infections have occurred in both hospital and laboratory environments. In South Africa an association with wild birds has been reported.

The incubation period is about 1 week. The onset of fever is usually sudden, and fever is usually continuous, although occasionally remittent or biphasic. Signs and symptoms include fever, headache, nausea, vomiting, joint pains, backache, photophobia, together with circulatory disorders, thrombocytopenia, and leucopenia. Haemorrhagic manifestations are common, with bleeding from nasal, gastric, intestinal, uterine, and renal membranes (Fig. 4). Patients may present with acute abdominal pain, mimicking an acute surgical emergency, and operating-theatre staff have become infected and have died through contact with infected blood or secretions exposed at operation. The mortality is about 15 to 30 per cent, but may be as high as 40 to 80 per cent in hospital or nosocomial outbreaks. Transient hair loss has been reported.

Fig. 4 Patient with Crimean–Congo haemorrhagic fever showing extensive ecchymoses on the arms and thorax (by courtesy of Professor D.I.H. Simpson).

Control, treatment, and prognosis

No vaccine is available. Avoidance of tick bites may reduce the risk of infection. In hospital outbreaks, meticulous attention to the containment of infected secretions is essential and barrier nursing should be used. There may be neurological involvement, which usually indicates a poor prognosis. Those patients who recover may be left with a polyneuritis that persists for months, but eventual recovery is to be expected.

Genus *Phlebovirus*

At least nine different phleboviruses are known to infect man (see Table 1). Pappataci fever, sandfly fever, or Phlebotomus fever was recognized as a clinical entity in the Mediterranean area during the nineteenth century, and the association with *Phlebotomus papatasi* sandflies was clearly demonstrated by showing that filtrates of human blood would reproduce the disease in human volunteers. For many years it was thought that man was the only vertebrate host, but antibody studies indicate that gerbils, cattle, and sheep may also be infected. Naples virus was isolated by American investigators from human serum collected during an outbreak of sandfly fever in Naples, and the Sicilian virus was isolated by the same workers from American troops with a similar fever in Palermo, Sicily. The two viruses have many common properties, but they are serologically quite distinct. Sandfly fever is widespread throughout the Mediterranean area, and also occurs in Egypt, Greece, Iran, Turkey, the former Yugoslavia, Bangladesh, India, Pakistan, and the southern states of the former USSR. Toscana virus, serologically related to the Naples virus, is found in countries bordering the Mediterranean; it is notable for its ability to infect the central nervous system, especially in central Italy where it is thought to be responsible for at least 80 per cent of acute summertime infections of the central nervous system in children. The viruses that cause classic sandfly fever do not occur in the New World, but in South and Central America a similar clinical condition follows infection with Alenquer, Candiru, Chagres, and Punta Toro viruses.

Rift Valley fever has long been known as a disease of domestic animals, mainly sheep, in East Africa, which occasionally spreads to farm workers and others handling infected animals. The infection is endemic, but seldom recognized, in many wild game animals in Africa. Molecular studies have established that Rift Valley fever virus is very similar to sandfly fever viruses and Punta Toro virus in having an ambisense replication mechanism; this property distinguishes the genus *Phlebovirus* from other genera within the family. In its biological properties, Rift Valley fever virus differs from the sandfly fever viruses, Punta Toro viruses, and most other members of the genus, in being normally transmitted by mosquitoes rather than sandflies. When it was recognized that the tick-transmitted Uukuniemi and Zaliv-Terpeniya viruses also shared an ambisense replication strategy, these viruses were removed from their earlier classification in the genus Uukuvirus and were redesignated to the genus *Phlebovirus*. The only evidence that Uukuniemi virus can infect man is the finding of specific antibodies in some human sera collected in Estonia and in the former Czechoslovakia. Zaliv-Terpeniya virus was isolated from bird ticks collected on an island in the Sea of Okhotsk, Sakhalin region, and there is some evidence that it may be pathogenic to man.

Sandfly fever, Naples, and Sicilian viruses

Symptoms

After an incubation period of 2 to 6 days, there is an abrupt onset of fever, chills, nausea and vomiting, epigastric pain, and often severe, generalized headache leading to incapacitating prostration. Fever of 38 to 40 °C usually resolves after 2 to 3 days, but may be biphasic and persist for a week. There is no rash, but small haemorrhages into the skin and mucous membranes may be seen. Photophobia and eye pain are not uncommon, lymphadenopathy is often seen, and the liver may be tender, although jaundice is rare. The disease is self-limiting, with complete recovery. No deaths have been attributed to either sandfly fever, Naples, or Sicilian viruses.

Rift Valley fever virus

Following its initial isolation in 1930 as the agent of enzootic hepatitis of domestic animals in Kenya, Rift Valley fever virus was recognized as the cause of sporadic human infections in East, Central, and West Africa, with a particular capacity to infect those handling the virus in the laboratory. In East and Central Africa the virus has been isolated from a variety of mosquito species and recent studies have shown that the virus is capable of persisting in mosquito eggs during the dry season, emerging when larvae hatch in the rainy season. From 1951 to 1956 there were severe epizootics in lambs in southern Africa, and many human cases occurred. Further human cases with several deaths were seen in South Africa in 1975, and a major outbreak occurred in East Africa following El Niño flooding in 1997 to 1998, apparently seeding a 'virgin soil' outbreak in Saudi Arabia and Yemen in 2000.

.In the Central African Republic in 1969 a virus isolated from *Mansonia africana* mosquitoes and named Zinga virus was associated with several cases of haemorrhagic fever; Zinga virus was later shown to be a strain of Rift Valley fever virus. In West Africa, Rift Valley fever virus was isolated from mosquitoes in Nigeria and from bats in Guinea, but despite the presence of antibodies in human sera collected in Nigeria and Senegal, human disease was unrecognized until 1987 when a substantial epidemic occurred in Mauritania, with further epidemics in following years. In 1977 the virus spread, apparently for the first time, into Egypt, producing a major epizootic in domestic animals, principally sheep and goats, but also cattle, and causing some 600 human deaths within a period of 3 months. The virus has been detected periodically since then in Egypt, and the principal vector seems to be the mosquito, *Culex pipiens*. It is of interest that both the Egyptian and the Mauritanian epidemics appear to be linked to major ecological changes following the construction of the Aswan Dam on the Nile and dams on the Senegal River.

Symptoms

After an incubation period of 3 to 6 days there is an abrupt onset of fever, shivering, nausea and vomiting, epigastric pain, and often severe, generalized headache. The fever may be biphasic, with temperatures between 38 and 40 °C, and may remain elevated for at least a week. There is no rash, but small haemorrhages into the skin and mucous membranes may be seen. Photophobia and eye pains are not uncommon; there may be conjunctival inflammation, and a central serous retinitis, leading to central scotoma and sometimes to retinal detachment. The fundus may show macular exudates that are slow to disappear. There is often a lymphadenopathy, and although the liver is frequently involved and may be tender, jaundice is rare, but appears to have been more common during the recent outbreaks in Mauritania. Convalescence may be protracted, but is usually uncomplicated; however, a small percentage of patients may suffer severe complications such as haemorrhagic fever, encephalitis, or eye lesions. Haemorrhagic disease presents as above, but progresses with petechial, mucous membrane, and gastrointestinal haemorrhagic, jaundice with severe liver and renal dysfunction often progressing to disseminated intravascular coagulation, hepatorenal syndrome, and may end in death. Patients with encephalitis typically recover from acute febrile disease only to present within a few days to 2 weeks later with headache, meningismus, confusion, and fever, often leading to residua or ending in death. Ocular complications are characterized by rapid onset of decreased visual acuity due to retinal haemorrhage, exudates, and macular oedema. These are also seen after apparent recovery from the initial disease. About half of these patients suffer some degree of permanent vision loss. Deaths from Rift Valley fever were rare before the 1977 outbreak in Egypt, but the Mauritanian epidemics in which at least 25 persons died with jaundice and haemorrhagic manifestations, and the recent East African and Arabian Peninsula outbreaks with several hundred suspect fatalities clearly establishes it as a life-threatening infection.

Control, treatment, and prognosis

Veterinary vaccines have been used for a number of years, and formalin-inactivated vaccines have also had limited use for the prevention of disease in laboratory workers and others exposed to high risk of infection. Improved vaccines based on molecular techniques are under development. Although there are no reports of nosocomial transmission, barrier nursing would be a sensible precaution.

Unassigned viruses and viruses causing only minor disease in man

The great majority of the viruses listed in Table 2 cause only a mild, febrile illness, but the following show certain additional features.

Bhanja virus (unassigned)

This virus was first isolated from *Haemaphysalis intermedia* ticks collected from healthy goats in India, but has since been isolated in Sri Lanka, in Africa, and in Europe. Infection of goats is widespread in Italy and in the former Yugoslavia, where there have been several reported human cases, including some with severe neurological disease, and at least two deaths. Laboratory infections have also occurred.

Bwamba virus (*Bunyavirus*)

This was first isolated in Uganda in 1941 and is very widespread throughout sub-Saharan Africa. More than 75 per cent of adult human sera collected in Nigeria and over 95 per cent of human sera collected in Uganda and Tanzania have antibodies against Bwamba virus. The original cases showed fever, headache, generalized body pains, and conjunctivitis, but no rash, although a rash has been described in the Central African Republic. No fatalities have been reported.

Nyando virus (*Bunyavirus*)

This virus was first isolated from mosquitoes in Kenya; it has since been isolated from man in the Central African Republic, where it caused fever, myalgia, and encephalitis.

Tataguine virus (unassigned)

This causes fever, rash, and joint pains in at least five African countries (Cameroon, Central African Republic, Ethiopia, Nigeria, and Senegal).

Wanowrie virus (unassigned)

This virus was first isolated in India from *Hyalomma marginatum* ticks collected from sheep. It has also been isolated in Egypt and Iran, and in Sri Lanka, where it was recovered from the brain of a 17-year-old girl who died following a 2-day fever with abdominal pain and vomiting.

Further reading

Bartelloni PJ, Tesh RB (1976). Clinical and serologic responses of volunteers infected with phlebotomus fever virus (Sicilian type). *American Journal of Tropical Medicine and Hygiene* **25**, 456–62.

Calisher CH, Thompson WH, eds (1983). *California serogroup viruses. Progress in Clinical and Biological Research*, Vol 123. Alan R. Liss, New York.

Hooper JW, Li D (2001). Vaccines against hantaviruses. *Current Topics in Microbiology and Immunology* **256**, 171–91.

LeDuc JW (1995). Hantavirus infections. In: Porterfield JS, ed. *Exotic viral infections*, pp 261–84. Chapman & Hall, London.

LeDuc JW, Pinheiro FP (1989). Oropouche fever. In: Monath TP, ed. *The arboviruses: epidemiology and ecology*, Vol 4, pp 1–14. CRC Press, Boca Raton, Florida.

Lee HW, Calisher C, Schmaljohn, CS (1999). *Manual of hemorrhagic fever with renal syndrome and hantavirus pulmonary syndrome.* WHO Collaborating Center for Virus Reference and Research (Hantaviruses), Seoul, Korea.

Monath TP, ed. (1989). *The arboviruses: epidemiology and ecology*. CRC Press, Boca Raton, Florida.

Peters CJ (1997). Emergence of Rift Valley fever. In: Saluzzo JF, Dodet B, eds. *Factors in the emergence of arbovirus diseases*, pp 253–64. Elsevier, Paris.

Peters CJ (1998). Hantavirus pulmonary syndrome in the Americas. *Emerging Infections* **2**, 17–64.

Peters CJ, LeDuc JW (1991). Bunyaviridae: bunyaviruses, phleboviruses, and related viruses. In: Belshe RB, ed.. *Textbook of human virology*, 2nd edn, pp 571–614. Mosby Year Book, St. Louis.

Pinheiro FP *et al.* (1981). Oropouche virus. I. A review of clinical, epidemiological, and ecological findings. *American Journal of Tropical Medicine and Hygiene* **30**, 149–60.

Saluzzo JF, Dodet B, eds (1999). *Factors in the emergence and control of rodent-borne viral disease (hantaviral and arenal diseases).* Elsevier, Paris.

Swanepoel R (1995). Nairovirus infections. In: Porterfield JS, ed. *Exotic viral infections*, pp 285–93.Chapman & Hall, London.

Tesh RB (1989). Phlebotomus fevers. In: Monath TP, ed. *The arboviruses: epidemiology and ecology*, Vol 4, pp 15–27. CRC Press, Boca Raton, Florida.

7.10.15

Arenaviruses

Susan Fisher-Hoch and Joseph McCormick

General considerations

Ecology and epidemiology

Arenaviruses infect rodents in the New and Old Worlds. There are at least 15 arenaviruses but only five produce significant human disease (Fig. 1); Lassa, Junin, Machupo, Guanarito, Sabia and lymphocytic choriomeningitis virus (LCMV). Arenaviruses have coevolved over very long periods of time with their natural rodent host so that the distribution of a given virus is restricted to its host range. Further, division into Old World (LCMV, Lassa) and New World (Tacaribe complex) viruses based on geographic distribution and antigenic typing is endorsed by nucleocapsid sequencing data.

Fig. 1 World map showing the approximate distribution of arenaviruses.

Rodents normally experience silent, lifelong infection, with perinatal transmission. Lifelong viruria is the primary source of contamination of the environment. Human infection results from intrusion into the rodents' ecological niche or infestation of poor housing. The virus infects through skin lesions and possibly mucosae or rodent-urine-contaminated dust aerosol. Human-to-human spread is common for Lassa fever both in community and hospital settings, but is apparently rare with the other pathogenic arenaviruses. Disease may be severe and haemorrhagic and all the pathogenic arenaviruses require Biosafety Level 4 (BSL4), except LCMV (BSL3).

Virology

By electron microscopy, host-cell ribosomes included in the virion resemble grains of sand ('arena' = sand in Latin) (Fig. 2). They are enveloped, pleomorphic, membrane viruses 50 to 300 nm (mean 110–130 nm) in diameter with a virion density in sucrose of 1.17 g/cm^3. They contain two segments of single-stranded RNA, tightly associated with a nucleocapsid protein of 65 000 to 72 000 molecular weight. The large strand of ambisense RNA, of molecular weight 2.0 to 3.2 × 10 ^{6}, codes for the viral polymerase and a zinc finger protein. The small single strand of ambisense RNA, molecular weight about 1.1 to 1.6 × 10 ^{6}, encodes the glycoprotein precursor and the nucleoprotein. The genome is enclosed in a membrane bearing two glycosylated proteins of molecular weights 34 000 to 44 000 (G1) and 54 000 to 72 000 (G2), derived from glycoprotein precursor by post-translational cleavage. Antigenic cross-reactivity is conserved at least at one epitope across all known arenaviruses.

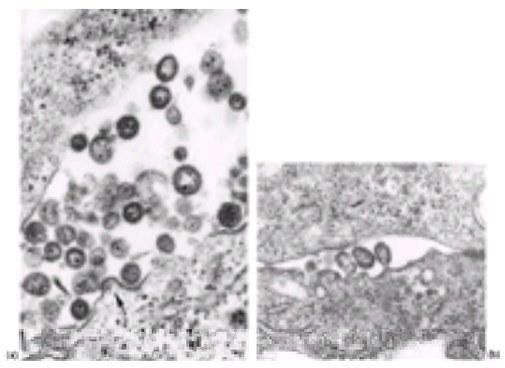

Fig. 2 Electronmicrographs of arenaviruses. (a) Machupo virus in tissue culture. Arrow shows virus budding through cell membrane. (b) Lassa virus budding from cell membrane.

Old World arenaviruses

Lassa fever

Epidemiology

Distribution and ecology

Lassa fever was first described in West Africa in the 1950s, but not isolated until 1969. It occurs from Northern Nigeria to Guinea, an area with a population of perhaps 180 million. Its only known reservoir is *Mastomys natalensis*, one of Africa's commonest rodents. In southern Africa, a related *Mastomys* carries the closely related Mopeia virus which can infect humans but is unable to cause significant clinical disease. *Mastomys* are highly commensal with man. In some areas, 50 per

cent of domestic rodents may be *Mastomys*, averaging 2.4 animals per house, but they have limited movement within a village. Prevalence of Lassa virus infection is variable in *Mastomys* and tends to cluster in houses, so that infection is endemic but focal.

Lassa fever is the haemorrhagic fever most likely to occur in travellers returning to developed countries. In the year 2000, at least four cases were imported into Europe. Cases in foreigners in 2000 have been seen during peacekeeping efforts in Sierra Leone. The rebels' stronghold is at the centre of the Lassa fever endemic area.

Epidemiology

Lassa fever is the arenavirus which affects the largest number of humans. Over 200 000 infections occur annually, and several thousand deaths. All age groups and sexes are affected. Antibody prevalence increases with age suggesting that most virus transmission to humans takes place in and around the home. Antibody prevalence ranges from 4 to 6 per cent in Guinea, 15 to 20 per cent in Nigeria, and up to 60 per cent in some villages in Sierra Leone. Seroconversion to Lassa virus positive ranged from 5 to 22 per cent/year among seronegative Sierra Leonean villagers. Disease to infection ratios range from 9 to 26 per cent in Sierra Leone, and the proportion of febrile illness associated with seroconversion to Lassa virus from 5 to 14 per cent. Five to 8 per cent of infected people may be hospitalized, of whom 17 per cent may die if untreated. However, the case fatality for all Lassa virus infections (hospitalized and non-hospitalized, symptomatic and asymptomatic) may be as low as 2 per cent. In endemic areas, Lassa fever may account for 10 to 16 per cent of all adult medical admissions and about 30 per cent of adult deaths among medical admissions.

Transmission

Direct contact between virus contaminated articles and surfaces and cuts and scratches on bare hands and feet may be the most important and consistent mode of transmission in endemic areas. The sporadic pattern of human infection in the community does not suggest aerosol transmission. Nosocomial spread in hospitals is associated with inadequate disinfection and direct contact with infected blood and contaminated needles. Increasing and indiscriminate use of routine intravenous therapy in West African hospitals, along with inadequate needle and syringe care, led to large-scale epidemics. A prospective study in a hospital in an endemic area showed that simple but rigorous barrier nursing techniques can reduce the frequency of infection in hospital personnel handling Lassa fever patients. In London, none of 159 unprotected hospital contacts of a severely ill Lassa fever patient was infected. In developed countries, (United Kingdom and United States) in 1990, there were no secondary infections among 907 documented hospital contacts of infected patients (188 classified as high risk).

Risk factors

Rodent-to-human infection is strongly associated with indiscriminate food storage, and practices such as catching, cooking, and eating rodents. Person-to-person spread is common in households. Risk of infection in villages is associated with direct contact, nursing care, or sexual contact with someone during the incubation, acute, or convalescent phases of illness.

Clinical features

Incubation period and prodrome

Following an incubation period of 7 to 18 days, symptoms begin insidiously with fever, weakness, malaise, severe, usually frontal, headache, and a painful sore throat (Fig. 3). More than 50 per cent of patients then develop joint and lumbar pain and 60 per cent or more develop a non-productive cough. Severe retrosternal chest pain, nausea with vomiting or diarrhoea, and abdominal pain are also common.

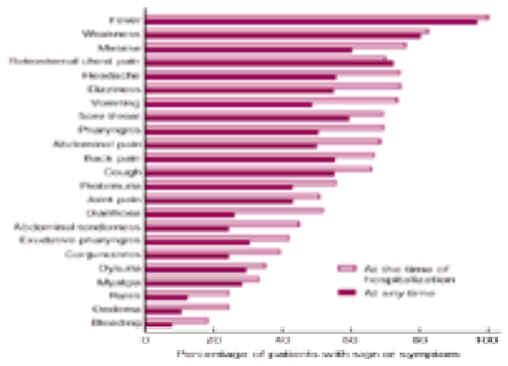

Fig. 3 Frequency of signs and symptoms of Lassa fever.

Respiratory and pulse rate and temperature are elevated and blood pressure may be low. There is no characteristic rash; petechiae and ecchymoses are not seen. About a third of patients will have conjunctivitis. More than two-thirds have pharyngitis, half with exudates, diffusely inflamed and swollen posterior pharynx and tonsils, but few ulcers or petechiae. The abdomen is tender in 50 per cent of patients. Neurological signs in the early stages are limited to a fine tremor, most marked in the lips and tongue.

Severe disease

Up to a third of hospitalized patients progress to a prostrating illness 6 to 8 days after onset of fever, usually with persistent vomiting and diarrhoea. Patients are often dehydrated with elevated haematocrit. Proteinuria occurs in two-thirds of patients, with moderately elevated blood urea nitrogen. About half of Lassa fever patients have diffuse abdominal tenderness without localizing signs or loss of bowel sounds. The severe retrosternal or epigastric pain seen in many patients may be due to pleural or pericardial involvement. Bleeding is seen in only 15 to 20 per cent of patients, restricted to mucosal surfaces, conjunctivitis, gastrointestinal, or genital tracts. Severe pulmonary oedema and adult respiratory distress syndrome is common in fatal cases with gross head and neck oedema, stridor, and hypovolaemic shock (Fig. 4).

Fig. 4 Acute Lassa fever in a patient showing facial oedema, decerebrate posturing, and respiratory distress. This patient survived.

Over 70 per cent of patients have abnormal electrocardiograms (non-specific ST-segment and T-wave abnormalities, ST-segment elevation, generalized low voltage complexes, and changes reflecting electrolyte disturbance), but none correlates with clinical or other measures of disease severity or outcome. There is no clinical evidence of myocarditis. Neurological signs are infrequent, but carry a poor prognosis, progressing from confusion to severe encephalopathy with or without general seizures, but without focal signs. Cerebrospinal fluid is usually normal, apart from a few lymphocytes, and low titres of virus relative to serum. Pneumonitis and pleural and pericardial rubs develop in early convalescence in about 20 per cent of hospitalized patients, sometimes associated with congestive cardiac failure.

Laboratory measurements

A normal mean white blood cell count on admission to hospital ($6 \times 10^9/1$) may mask early lymphopenia with relative or absolute neutrophilia, as high as $30 \times 10^9/l$. Thrombocytopenia is moderate, even in severely ill patients, but platelet function is markedly depressed. Serum aspartate aminotransferase levels greater than 150 U/l are associated with a case fatality of 50 per cent (Fig. 5). The ratio of aspartate aminotransferase to alanine aminotransferase is as high as 11:1. Prothrombin times, glucose, and bilirubin levels are nearly normal, excluding biochemical hepatic failure.

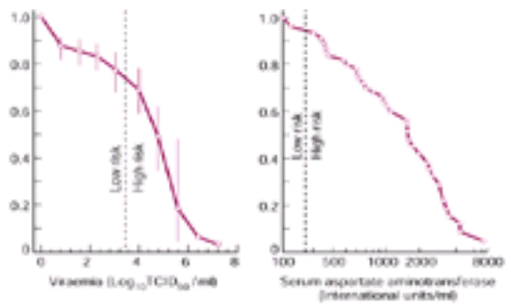

Fig. 5 Cumulative survival in Lassa fever related to serum viraemia and aspartate aminotransferase (AST) levels.

Viraemia of greater than 3 $\log_{10}$ $TCID_{50}$/ml is associated with increasing case fatality (Fig. 5). High virus titres occur in liver, ovary, pancreas, uterus, and placenta, but there is no histological evidence of organ failure. High viraemia and aspartate aminotransferase together carry a risk of death of nearly 80 per cent.

Mortality

The overall death rate for all infections may be as low as 2 per cent, but is about 16 per cent in hospitalized patients. In Nigeria, much higher death rates have been observed. The case fatality may be over 30 per cent in the third trimester of pregnancy, and 50 per cent in patients with haemorrhage.

Complications and sequelae

Nearly 30 per cent of patients develop uni- or bilateral-deafness beginning during convalescence. About half show a near or complete recovery after 3 to 4 months, but the other half remain permanently deaf. Many patients also exhibit transient cerebellar signs during convalescence, particularly tremors and ataxia. Infrequent complications are uveitis, pericarditis, orchitis, pleural effusion, ascites, and acute adrenal insufficiency. Renal and hepatic failure are not seen.

Lassa fever in pregnancy

Lassa fever is a common cause of maternal mortality in parts of West Africa. In the third trimester there is a two to three-fold risk of maternal death from infection. Very high levels of virus replication have been found in placental tissue. Mortality was reduced four-fold in women spontaneously or therapeutically aborted (OR for fatality with pregnancy intact 5.5). Fetal loss is as much as 87 per cent, and does not seem to vary by trimester. Lassa virus is present in the breast milk of infected mothers, and neonates are therefore at risk of congenital, intrapartum, and puerperal infection.

Lassa fever in children

Lassa fever is common in children, but may be difficult to diagnose. In very young babies marked oedema has been reported. In older children the disease may manifest as diarrhoea, pneumonia, or an unexplained prolonged fever. In a hospital in Sierra Leone, 21 per cent of paediatric admissions had Lassa fever, with 12 per cent fatality. An outpatient study found 24/435 children (6 per cent) with evidence of previous Lassa infection and 68 who seroconverted (16 per cent).

Differential diagnosis

The diagnosis of Lassa fever must be considered if there is a history of potential exposure to rodents or patients in or from West Africa. Cerebral malaria or septicaemia must be excluded but double infections can occur, particularly with malaria. In West Africa, fever with pharyngitis, proteinuria, and retrosternal chest pain had a predictive value for Lassa fever of 81 per cent and a specificity of 89 per cent. Likewise a triad of pharyngitis, retrosternal chest pain, and proteinuria (in a febrile patient) correctly predicted Lassa fever 80 per cent of the time (sensitivity 50 per cent). Bleeding and sore throat had specificity for fatal outcome of 90 per cent (sensitivity only 36 per cent). In contrast, vomiting and sore throat had a specificity of only 47 per cent for fatal disease (sensitivity 89 per cent).

Pathogenic processes

Pathogenesis

Arenaviruses invade through broken skin to lymph nodes, progressing to generalized multiorgan infection, affecting especially the reticuloendothelial system. In Lassa fever, severe disease manifestations are intractable hypovolaemic shock, and/or severe central nervous system involvement, bleeding, and oedema of the face. There is endothelial and platelet dysfunction, despite adequate numbers of circulating platelets.

Pathology

There is focal necrosis of the liver, but damage is insufficient to cause fatal hepatic failure. A substantial macrophage response is seen with little, if any, lymphocytic inflammatory response. Moderate splenic necrosis primarily involves the marginal zone of the periarteriolar lymphocytic sheath. Diffuse focal adrenocortical cellular necrosis has been less frequently observed.

Immunology

The immunological response is complex. While neutralizing antibodies are associated with clearance of viraemia due to the South American arenaviruses, Lassa and lymphocytic choriomeningitis viruses appear to depend primarily on cytotoxic T-cell responses. There appears to be a brisk B-cell response with a classic primary IgG and IgM antibody response to Lassa virus early in the illness. By the sixth day of illness IgG antibodies are found in 46 per cent of patients and IgM antibody in 59 per cent. By the sixteenth day of illness both are found in 100 per cent of patients, but this does not coincide with virus clearance. A high viraemia and high IgG and IgM titres often coexist in both humans and primates, and virus may persist in the serum and urine of humans for several months after infection, and probably in occult sites, such as renal tissue, for years.

In a minority of patients, some low titre serum neutralizing activity may be observed several months after resolution of the disease. Controlled clinical trials with human

convalescent plasma containing high titred antibodies have shown no protective effect. Thus the clearance of Lassa virus appears to be independent of antibody formation. In non-human primates, killed vaccine and vaccinia-vectored vaccines expressing the nucleoprotein elicited high level antibodies to Lassa virus but no protection, while vaccinia-vectored glycoprotein vaccine elicited little antibody, but nevertheless provided protection. Reinfection following natural Lassa infection may occur in humans, apparently without clinical disease.

Laboratory procedures

Diagnosis

Laboratory diagnosis is by isolation of virus from serum, demonstration of a four-fold rise in antibody titre, or high-titre IgG antibody with virus-specific IgM antibody in association with compatible clinical disease, or more recently by detection of viral sequences by reverse transcriptase polymerase chain reaction (RT-PCR), or by detection of viral proteins using an ELISA system.

Virus isolation may be accomplished easily from serum or tissues in cell cultures, but should be performed in BSL4 laboratory facilities. Specimens should be drawn into a vacuum tube to minimize risk of infection. Virus has also been isolated from breast milk, spinal fluid, pleural and pericardial transudate, and from autopsy material. Virus excretion in urine is intermittent but may persist. Antigen detection in conjunctival scrapings, buffy coat preparations, cells from pharyngeal aspirates, and urinary sediment have not been successful. ELISA tests have been notoriously difficult to calibrate in the field, particularly in West Africa. Newer serological tests include recombinant immunoblot techniques. Because of its superior sensitivity and specificity, where possible the ELISA should be used. Specific IgM, if detected early in disease, indicates acute infection. Antigen detection by ELISA and ELISA for IgM have recently been shown to have a specificity of 90 per cent and a sensitivity of 88 per cent compared with RT-PCR, but IgG was only detected in 16 per cent of the cases. A rising titre, with or without IgM, can be very useful in acute diagnosis. The RT-PCR or hybridization is much more rapid, and is as sensitive as virus isolation, and more sensitive than antigen detection by ELISA, and is the assay of choice for reference laboratories when a rapid diagnosis is needed. No commercial diagnostic reagents are currently available. BSL4 reference laboratories do not produce non-infectious reagents for use by other laboratories.

Containment and disinfection

Lassa virus is robust, and withstands drying. Blood from severely ill patients may contain 10^9 infectious units per ml, and rodent urine may contain 10^4 infectious units/ml. However, it can be inactivated by heat, detergents, chlorine, formalin, and UV radiation (including sunshine). Antigenic properties are best conserved by inactivation with gamma irradiation. Disinfection by washing with 0.5 per cent phenol in detergent (for example Lysol), 0.5 per cent hypochlorite solution, formaldehyde, or paracetic acid is recommended.

Patient management

Supportive care

Fluid, electrolyte, respiratory, and osmotic imbalances should be corrected, and full intensive care support, including mechanical ventilation offered if required. However, even vigorous support may be insufficient to prevent fatal progression of advanced disease. Pregnant patients often present with absent fetal movements, and survival may depend on aggressive obstetric intervention.

Antiviral therapy

Ribavirin (tribavarin) is effective but must be given early in disease. Ribavirin is given by intravenous infusion as a 2-g loading dose followed by 1 g every 6 h for 4 days then 0.5 g every 8 h for 6 more days. Toxicity is confined to mild reversible anaemia. A five to ten-fold decrease in the case–fatality ratio was demonstrated in patients treated with ribavirin compared with untreated patients when therapy was given within the first 6 days of illness. Patients with high aspartate aminotransferase and viraemia, risk factors for unfavourable outcome (Fig. 5), who were treated within the first 6 days of illness suffered a 5 to 9 per cent case fatality. Those with the same risk factors receiving treatment more than 6 days after the onset of illness had a 26 to 47 per cent fatality, compared with 52 to 78 per cent untreated. Ribavirin is contraindicated in early pregnancy because of potential terato-genicity, but the fetus rarely survives the infection. It has been used in late pregnancy (>24 weeks) but its efficacy has not yet been determined.

Prevention

Hospital containment

The key to prevention of human-to-human transmission is isolation of febrile patients and rigorous use of gloves and disinfection. Complete support, including intensive care, or surgery, should be provided. Patient isolators should not be used since they induce loss of manual dexterity and inhibit intensive care and communication. Recommendations issued for patient management and handling of clinical specimens from AIDS patients are adequate for containment of Lassa fever.

Contacts

High risk is associated with percutaneous or mucosal contact with blood or body fluids. Medium risk contacts (unprotected contact with blood or body fluids) may safely be observed for development of persistent high fever for 3 weeks from the last date of contact by daily temperature measurement and telephone reporting. The practice of following up airline passengers and other low risk contacts (no unprotected physical contact with patient or body fluids) is unnecessary.

Prophylaxis

Oral ribavirin should be offered to high-risk contacts as soon as possible after exposure to Lassa virus (600 mg orally, four times a day for 10 days).

Vaccine

A vaccinia virus recombinant expressing Lassa glycoprotein (G) protected primates from infection, but a nucleoprotein expressing recombinant did not. Immunization of primates with inactivated (gamma irradiated) whole Lassa virus resulted in brisk antibody responses to both proteins, but all animals died with serum virus titres equal to unvaccinated controls. In humans, the presence of antibody to neither glycoprotein nor nucleoprotein at the time of hospital admission was associated with survival, or even severity of disease.

Following Lassa virus challenge, high survival rates are seen after vaccination with vaccines expressing G, despite low or undetectable antibody levels preinfection. The duration of the interval between vaccination and challenge, and the challenge dose, affected the efficacy of vaccines. Almost all surviving, asymptomatic animals experienced viraemia, even those vaccinated with Mopeia virus (essentially a live, attenuated Lassa virus), consistent with the hypothesis that virus replication is controlled by CTL responses and not antibody responses. These data show that the G gene is necessary and sufficient to protect primates against a large parenteral challenge dose (Table 1).

Persistence

Lassa virus may persist at low titre for a limited time in primates, but virus is sequestered, and transmission is unlikely. Viraemia in patients is quickly controlled. However, virus may be detected intermittently in human urine for up to 60 days.

Control

Control of rodents would prevent Lassa fever in West Africa. Improvement of housing and food storage might reduce the domestic rodent population, but such

changes are not easily made without massive improvement in the local economy.

Lymphocytic choriomeningitis virus

Epidemiology

Distribution and ecology

Lymphocytic choriomeningitis (LCMV) virus was isolated in 1933 from the cerebrospinal fluid of a patient with suspected St Louis encephalitis. It is widely distributed in its natural host, *Mus musculus*. Transmission from feral rodents to humans rarely results in clusters of disease, and person-to-person spread has not been demonstrated. The virus has acquired scientific importance in laboratory studies of immunological tolerance and virus-induced immunopathological disease.

Epidemiology

The prevalence of antibody to LCMV virus in the general population is between 1 and 5 per cent. Sporadic disease in rural areas occurs during the colder months but in urban areas with large rodent populations the epidemiology may be different. In Baltimore, 4.7 per cent of those attending a sexually transmitted disease clinic had antibodies to LCMV.

Risk factors and transmission

Although feral mice cause sporadic infection, the most common sources of human infection are pet or laboratory rodents particularly hamsters, white mice, or nude mice. Laboratory outbreaks of human disease have followed contact with infected animals or virus, and aerosol transmission may have occurred.

Clinical features

Incubation period and disease

The incubation period is about 1 to 3 weeks. About 35 per cent of infections are asymptomatic, and a further 50 per cent have a febrile illnesses without significant central nervous system manifestations. About 15 per cent have lymphocytic choriomeningitis (disease to infection ratio 1:3). Illness begins with fever, malaise, weakness, myalgia, and headache, often severe, retro-orbital, and associated with photophobia. Anorexia, nausea, dizziness, and myalgia are common. Only two deaths have been reported, but there is prolonged convalescence.

Laboratory findings

Leucopenia and mild thrombocytopenia are common, and cerebrospinal fluid from patients with meningeal signs contains several hundred white cells, predominantly lymphocytes (>80 per cent), with slightly increased protein and occasionally low sugar levels. Virus is found in spinal fluid during acute disease.

Complications and sequelae

About one-third of patients with central nervous system manifestations will develop encephalopathy while the rest exhibit aseptic meningitis. An interstitial pneumonia, alopecia, orchitis, and transient arthritis of the hands have been described. Convalescence is prolonged with persistent fatigue, somnolence, and dizziness, and deafness. Neurological sequelae are unusual.

Differential diagnosis

LCMV should be considered in patients with fever with persistent meningeal signs, particularly if a history of rodent contact is elicited.

Pathogenic processes

Pathology

There are no published descriptions of the pathology of LCMV infection in humans. In one report of a fatal case there was perivascular macrophage infiltration in multiple areas of the brain. Antigen was observed in the meninges and cortical cells by IFA. In animal studies, the leptomeninges show dense infiltration with lymphocytic cells, with little involvement of the brain parenchyma.

Immunology

Antibody to LCMV appears in the first week of illness, with titres peaking at 40 to 60 days. In the natural host, the mouse, the immunology of natural and experimental infection has been extensively studied, but extrapolation to human disease needs to be made with caution.

Laboratory diagnosis

Virus may be cultured from cerebrospinal fluid or detected using RT-PCR during the acute phase of disease. IgG and IgM antibody may be detected in serum.

Patient management

There is no standard treatment for LCMV infection. Ribavirin is effective *in vitro*, but penetration into the cerebrospinal fluid is poor, however as the disease is severely debilitating its use should be considered.

Prevention

Laboratory outbreaks continue to occur, particularly through handling of persistently infected mice. The virus is a major laboratory hazard, and care must be taken to avoid infection. Exposure is too infrequent for there to be a market for a vaccine.

New World arenaviruses

The New World arenaviruses causing human disease are Junin (Argentine haemorrhagic fever), Machupo (Bolivian haemorrhagic fever), Guanarito (Venezuelan haemorrhagic fever), and Sabia (Brazilian haemorrhagic fever). The most important rodent hosts are the South American genera *Calomys*, *Sigmodon*, and *Oryzomys*. Together these are known as the South American haemorrhagic fevers. All are endemic in geographically limited areas, but new, related viruses are emerging in other yet unaffected areas. Between June 1999 and May 2000 three patients, two from southern California and a third from the San Francisco Bay area, died of an acute febrile illness with lymphopenia, thrombocytopenia, and acute respiratory distress syndrome. Two had liver failure and haemorrhagic manifestations. RNA fragments detected by PCR from all three patients shared 87 per cent identity with a recently described arenavirus from New Mexico, Whitewater Arroyo virus.

Argentine haemorrhagic fever

Epidemiology

Distribution and ecology

Argentine haemorrhagic fever was first recognized in the 1950s in the fertile farmland of north-western Buenos Aires Province in Argentina, and Junin virus was first isolated in 1958. The major rodent hosts are *Calomys musculinus* and *Calomys laucha* which, unlike *Mastomys* or *Mus*, are affected by the virus, with up to 50 per cent fatality in infected suckling animals, and stunted growth in many others. These are agrarian rodents, and most human cases are male agricultural workers, particularly harvesters of sugar cane.

Virology

Monoclonal antibody studies show Junin to be most closely related to Machupo and Tacaribe viruses, otherwise cross-reactivity with other New World arenaviruses is restricted to the nucleoprotein. Junin viruses comprise three clades depending on geographical origin, with the live attenuated Argentine haemorrhagic fever vaccine strain, Candid 1, a fourth, separate clade. No particular viral epitopes have been associated with varying severity or clinical forms of the human disease.

Epidemiology

About 21 000 cases have been reported since the early 1960s, averaging about 360 a year with wide annual fluctuations. Peak incidence is during summer and early autumn. The disease appeared to spread to new areas as incidence in the earlier affected areas decreased, possibly because of the virus' effect on rodent populations. Overall human antibody prevalence is about 12 per cent, with predominance in male agriculture workers, and about 30 per cent had no history of typical illness, (disease to infection ratio 2:3). The recent introduction of a live attenuated vaccine has dramatically reduced the incidence.

Transmission

The major routes of virus transmission to humans is probably through virus-infected dust, and mechanical harvesters are traditionally cited. Whether infection is through contamination of cuts and abrasions or mucosae or by aerosol is unclear. There is no recorded person-to-person spread. Recent studies have shown that the major host species, *Calomys musculinus*, is most frequently captured from roadsides and fence lines.

Clinical features

Incubation period and prodrome

After an incubation period of about 12 days, there is insidious onset of malaise, high fever, severe myalgia, anorexia, lumbar pain, epigastric pain and abdominal tenderness, conjunctivitis and retro-orbital pain, often with photophobia, and constipation. Nausea and vomiting frequently occur after 2 or 3 days of illness. There is no lymphadenopathy or splenomegaly, sore throat or cough, but there is high fever, marked erythema of the face, neck, and thorax, and conjunctivitis. Respiratory symptoms are uncommon. Petechiae appear by the forth or fifth days of the illness. There may be a pharyngeal enanthem, but pharyngitis is uncommon.

Severe disease

The infection either resolves after about 6 days or progresses to severe disease. In contrast to Lassa fever, South American haemorrhagic fevers are associated with haemorrhagic manifestations in nearly half of the patients (gingival haemorrhages, epistaxis, metrorrhagia, petechiae, ecchymoses, purpura, melaena, and haematuria). Severe cases have nausea, vomiting, intense proteinuria, microscopic haematuria, oliguria, and uraemia. Fatal cases develop hypotensive shock, hypothermia, and pulmonary oedema. Renal failure has been reported but glomerular filtration rates, renal plasma flow, and creatinine clearance are usually normal. There is some electrocardiographic evidence of myocarditis. Fifty per cent of Argentine haemorrhagic fever and Bolivian haemorrhagic fever patients also have neurological symptoms during the second stage of illness, such as tremors of the hands and tongue, progressing in some patients to delirium, oculogyrus, and strabismus. Meningeal signs and cerebrospinal fluid abnormalities are rare.

Laboratory findings

A low white blood cell count and a platelet count are invariable. Bleeding and clot retraction times are concomitantly prolonged. Though reductions of levels of Factors II, V, VII, VIII, and X and of fibrinogen are observed, alterations in clotting functions are usually minor. Disseminated intravascular coagulation is not a significant feature, despite some reports of the presence of fibrinogen degradation products and absence of fibrinolysis. Proteinuria is common and microscopic haematuria also occurs. Liver and renal function tests are only mildly abnormal. Virus titres in serum are not as high as in Lassa fever, but the infection is also apparently pantropic.

In a febrile patient, the combination of a platelet count of less than 100 000/mm ^{2} and a white blood cell count of less than 2500/mm ^{2} has a sensitivity and specificity of 87 per cent and 88 per cent respectively. The combination of a platelet count of less than 100 000/mm ^{2} and a white cell count of less than 4000/mm^2 has a sensitivity of 100 per cent and a specificity of 71 per cent. These criteria are now recommended for use in screening patients for potential therapy with immune plasma or ribavirin.

Mortality

Mortality is about 16 per cent in laboratory-confirmed, hospitalized patients with untreated Argentine haemorrhagic fever. There are no estimates of overall mortality in populations.

Complications and sequelae

A late neurological syndrome in about 10 per cent of cases, consisting mainly of cerebellar signs, is associated with treatment using high titre antiserum. It begins between 4 and 6 weeks after onset of acute illness and lasts less than a week. It is characterized by fever, headache, ataxia, and intention tremors, and a mild cerebrospinal fluid pleocytosis with anti-Junin virus antibody in the cerebrospinal fluid. Most patients recover within 3 months. Mild permanent damage to acoustic centres has been detected in a small group of patients. Argentine haemorrhagic fever is reported to be severe in pregnancy.

Pathogenic processes

Pathogenesis

Despite the different degrees of bleeding, there are sufficient similarities between the course of disease in Argentine haemorrhagic fever, Bolivian haemorrhagic fever, and Lassa fever to speculate that they share pathophysiological pathways. Organ function, other than the endothelial system, appears to remain intact, and the critical period of shock is brief, lasting only 24 to 48 h. Hepatitis is mild and renal function is well maintained. Bleeding is more pronounced but is not the cause of shock and death. Capillary leakage is significant, but the dramatic head and neck oedema characteristic of severe Lassa fever is absent. Proteinuria is significant, and dehydration important. Though petechiae suggest endothelial damage, no clear evidence of virus replication in endothelium has been demonstrated. Persistent hypovolaemic shock despite intravascular volume expanders suggest that this is due to leakage of fluid into extravascular spaces. Adult respiratory distress syndrome is not described, but tissue oedema is frequent and pulmonary oedema may follow vigorous intravenous fluid replacement.

Other observations include high levels of interferon in severely ill patients, and a decrease in complement. More recently, proinflammatory cytokines, namely interferon-a and tumour necrosis factor-a (TNF-a) and interleukins, IL6, IL8, and IL10, have been variably reported. A platelet inhibitor, which may be interferon-a, similar to that described in Lassa fever, has inhibitory effects on thrombin-induced aggregation and ^{14}C serotonin release. Raised G-CSF levels correlated with TNF-a and disease severity.

Pathology

There are large areas of intra-alveolar or bronchial haemorrhage, petechiae on organ surfaces, and ulcerations of the digestive tract, although bleeding is not

massive. Large areas of intra-alveolar or bronchial haemorrhage are often seen with no evidence of inflammatory process. Pneumonia with necrotizing bronchitis or pulmonary emboli is observed in half of the cases. Haemorrhage and a lymphocytic infiltrate have been observed in the pericardium, occasionally with interstitial myocarditis. Lymph nodes are enlarged and congested with reticular cell hyperplasia. Splenic haemorrhage is common, and medullary congestion with pericapsular and pelvic haemorrhages are frequently seen. Adrenal necrosis has not been reported. Renal damage occurs in about half of the fatal cases, and consists of severe structural damage in the distal tubular cells and collecting ducts with relative sparing of the glomeruli and proximal tubules. There is no evidence of direct viral central nervous system infection. Microscopically, there is mild oedema of the vascular walls, with capillary swelling and perivascular haemorrhage associated with viral antigen but no immunoglobulins. Electronmicroscopy shows intracytoplasmic and intranuclear inclusions and marked, non-specific cellular damage in all organs examined.

Immunology

In striking contrast to Lassa fever, the antibody response to Junin virus is effective in clearing virus during acute disease, and may also be sufficient to protect against infection. Neutralizing antibody may be detectable at the time the patient begins to recover from the acute illness, and the therapeutic efficacy of immune plasma in patients with Junin infection is directly associated with the titre of neutralizing antibody in the plasma given. This neutralizing antibody is directed towards the surface glycoproteins. Nevertheless, like Lassa, Junin virus may persist. In vaccine studies in Argentina, some people do not produce measurable neutralizing antibodies but do mount a Junin-virus-specific lymphocytic proliferative response. It is probable that both humoral and cellular immunity are important in limiting virus replication and thus in recovery and protection.

Differential diagnosis

Argentine haemorrhagic fever should be considered in patients in the endemic area, particularly male agricultural workers, who present with fever of unknown origin and a bleeding diathesis. No cases have been reported outside Argentina.

Laboratory diagnosis

Antibodies measured by IFA may be positive by the end of the second week of illness. Neutralizing and complement fixing antibody to Junin are usually detectable 3 to 4 weeks after onset. ELISA systems for antibody and antigen are described with sensitivity and specificity of 99.2 per cent and 98.8 per cent, respectively, but reagents are not generally available. Virus may also be cultured from serum, but this must be performed in BSL4 conditions. For acute diagnosis, RT-PCR on whole blood samples is now the method of choice with sensitivity of 98 per cent and specificity 76 per cent.

Patient management

Specific treatment

In contrast to Lassa fever, convalescent-phase plasma has been shown to be highly successful in Argentine haemorrhagic fever, reducing the mortality from 16 per cent to 1 per cent in patients treated in the first 8 days of illness. Viraemia is reduced within 24 h of treatment, and clinical symptoms and haematological alterations are less severe than in control cases receiving non-immune plasma. Efficacy is directly related to the concentration of neutralizing antibodies. Delayed treatment is less successful. Availability of appropriately screened plasma may be a problem. Ribavirin is effective in experimentally infected primates, and its therapeutic use late in disease is being explored. The late neurological syndrome of Argentine haemorrhagic fever may be associated with therapy, particularly very high titre immune plasma.

Prevention

The human–rodent encounter resulting in Argentine haemorrhagic fever occurs during crop harvests, and there are no means of controlling feral rodents. A successful live attenuated vaccine, Candid 1, for Argentine haemorrhagic fever has now undergone Phase III studies, and is in use in the endemic area of Argentina, where is has almost eliminated the disease. The vaccine has proved safe in large-scale trials, and has a protective efficacy of 84 per cent.

Bolivian haemorrhagic fever

Epidemiology

Bolivian haemorrhagic fever is caused by Machupo virus, first isolated in 1965, and is limited to a portion of the department of Beni in Bolivia. The only known reservoir is *Callomys callosus*, found in the highest density at the borders of tropical grassland and forest, in the eastern Bolivian plains, northern Paraguay, and adjacent areas of western Brazil. Infected rodents develop haemolytic anaemia and splenomegaly, with up to 50 per cent fatality among infected suckling animals, and stunted growth in many others. The virus renders *Calomys callosus* essentially sterile with the young dying *in utero*. Transmission from rodent to rodent is horizontal, not vertical, and is believed to occur through contaminated saliva and urine.

By 1962, more than 1000 cases had been identified in a confined area of two provinces. The largest known epidemic of Bolivian haemorrhagic fever, involving several hundred cases, followed a marked and unusual increase in the *Calomys* population in homes in the town of San Joaquin in 1963 and 1964. This seems to have been a unique event, and there have been virtually no cases until 1994, when there was an outbreak in north-eastern Bolivia. Since all ages and both sexes are affected, it can be assumed that most patients were infected in their homes. Person-to-person spread is rarely reported.

Virology

The sequence of the nucleocapsid protein of Machupo virus shows close relatedness to Junin and Tacaribe viruses. This, together with previous demonstrations of antigenic similarity and cross-protection, suggest that vaccines developed against Argentine haemorrhagic fever might also be effective against the Bolivian disease.

Clinical features

The incubation period, clinical disease, and pathology of Bolivian haemorrhagic fever closely resemble Argentine haemorrhagic fever. Initial symptoms include headache, fever, arthralgia, and myalgia. In the later stages of this illness, patients may develop haemorrhagic manifestations including subconjunctival haemorrhage, epistaxis, haematemesis, melena, and haematuria, as well as neurological signs including tremor, seizures, and coma. Case fatality in the 1960s was 22 per cent. Neurological sequelae are observed in experimentally infected primates. Diagnosis is made in the same way as for Argentine haemorrhagic fever. Machupo virus also induces a humoral immune response, which may include neutralizing antibody.

Treatment

During the 1960s, convalescent-phase immune plasma from survivors of Bolivian haemorrhagic fever was used. However, there is now a paucity of survivors of Bolivian haemorrhagic fever who can donate immune plasma, and there is no active program for collection and storage of Bolivian haemorrhagic fever immune plasma. In 1994, intravenous ribavirin was offered to two patients who both recovered without sequelae, but Machupo virus infection was only confirmed in one.

Prevention

The ideal method of prevention for these rodent-borne diseases is to prevent contact between rodents and humans. The effectiveness of this was admirably shown in the outbreaks of Bolivian haemorrhagic fever in the 1960s when rodent control programmes in the villages were highly successful in eliminating the epidemic situation. The Candid 1 vaccine used in Argentina has been proposed for use against infection with this virus.

Venezuelan haemorrhagic fever

Epidemiology

Guanarito virus, the aetiological agent of Venezuelan haemorrhagic fever, was first isolated in 1991. Person-to-person transmission is not reported, and is unlikely since there is low frequency of infection in family contacts and none in exposed hospital workers. That all ages and sexes are infected suggests transmission occurs in and around houses. Disease is endemic, without seasonal variation. There are no data on prevalence and risk factors for infection have not been identified. The cotton rat, *Sigmodon alstoni*, is now thought to be the principal rodent reservoir of Guanarito virus. Despite intensive surveillance, Venezuelan haemorrhagic fever has been detected in only the small region of western Venezuela where the first outbreak was seen.

Virology

Morphology and antigenic properties of Guanarito show it to be a new member of the Tacaribe complex with which it cross reacts broadly. Phylogenetic analysis of the nucleocapsid gene open reading frame showed that Guanarito virus is a genetically distinct arenavirus, with 32 per cent nucleotide sequence divergence ranging from 30 per cent (Junin) to 45 per cent (LCMV). This sequence region is a probable antigenic domain (amino acids 55–63) shared among all arenaviruses. Phylogenetic trees of rodent isolates delineate nine distinct Guanarito genotypes, most of which are restricted to discrete geographical regions. Human disease is not associated with a particular genotype or host rodent.

Clinical features

Little information is available but hospitalized patients with severe disease are described as febrile with prostration, headache, arthralgia, cough, sore throat, nausea/vomiting, and diarrhoea. Haemorrhage is manifest as epistaxis, bleeding gums, menorrhagia, and melaena. On physical examination, patients are toxic and usually dehydrated, with pharyngitis, conjunctivitis, cervical lymphadenopathy, facial oedema, or petechiae. Thrombocytopenia and neutropenia are common. The case fatality in 15 patients was over 60 per cent, but surveys suggest that overall mortality to infection ratio is much lower. Post mortem pathology included: pulmonary oedema with diffuse haemorrhages in the parenchyma and sub pleura; focal hepatic haemorrhages; cardiomegaly epicardial haemorrhages, splenic and renal swelling; and widespread bleeding into cavities. Like Argentine haemorrhagic fever and Bolivian haemorrhagic fever, antibodies to Guanarito virus appear later in the illness. The infection is likely to respond to ribavirin therapy, although no data are available.

Sabia virus

Sabia virus emerged in 1990 when it was isolated from a fatal case in São Paulo, Brazil. Subsequently, it caused two laboratory-acquired infections. Its natural distribution and host are still unknown. One incident involving a human exposure occurred in the Yale Arbovirus Research Unit on August 8, 1994 when a senior scientist was exposed to Sabia virus while purifying the virus from a large volume of tissue culture fluid. The patient treated himself immediately with ribavirin, and made a rapid and full recovery.

Molecular studies confirm that Sabia virus is distinct from all other members of the arenaviridae and shares a progenitor with Junin, Machupo, Tacaribe, and Guanarito viruses. It has a unique, predicted, three stem–loop structure in the S RNA intergenic region.

Further reading

Bowen MD, Peters CJ, Nichol ST (1997). Phylogenetic analysis of the Arenaviridae: patterns of virus evolution and evidence for cospeciation between arenaviruses and their rodent hosts. Molecular and Phylogenetic *Evolution* **8**, 301–16.

Enria DA, Briggiler AM, Fernandez NJ, Levis SC, Maiztegui JI (1984). Importance of dose of neutralising antibodies in treatment of Argentine haemorrhagic fever with immune plasma. *Lancet* **2**, 255–6.

Fisher-Hoch SP, Hutwagner L, Brown B, McCormick JB (2000). Effective vaccine for Lassa fever. *Journal of Virology* **74**, 6777–83.

Holmes GP, McCormick JB, Trock SC, *et al*. (1990). Lassa fever in the United States. Investigation of a case and new guidelines for management [see comments]. *New England Journal of Medicine* **323**, 1120–3.

Jahrling PB, Hesse RA, Eddy GA, Johnson KM, Callis RT, Stephen E (1980). Lassa virus infection of rhesus monkeys: pathogenesis and treatment with ribavirin. *Journal of Infectious Diseases* **141**, 580–9.

Johnson KM, McCormick JB, Webb PA, Smith ES, Elliott LH, King IJ (1987). Clinical virology of Lassa fever in hospitalized patients. *Journal of Infectious Diseases* **155**, 456–64.

Maiztegui JI, McKee KT Jr, Barrera Oro JG, *et al.* (1998). Protective efficacy of a live attenuated vaccine against Argentine hemorrhagic fever. AHF Study Group. *Journal of Infectious Diseases* **177**, 277–83.

McCormick JB, King IJ, Webb PA, *et al.* (1987). A case-control study of the clinical diagnosis and course of Lassa fever. *Journal of Infectious Diseases* **155**, 445–55.

McCormick JB, King IJ, Webb PA, *et al.* (1986). Lassa fever. Effective therapy with ribavirin. *New England Journal of Medicine* **314**, 20–6.

McCormick JB, Webb PA, Krebs JW, Johnson KM, Smith ES (1987). A prospective study of the epidemiology and ecology of Lassa fever. *Journal of Infectious Diseases* **155**, 437–44.

7.10.16 Filoviruses

Susan Fisher-Hoch and Joseph McCormick

Primary human infections with filoviruses are exceedingly rare. The first appearance of these viruses was in Marburg in 1967 when laboratory, medical, and animal care personnel exposed to tissues and blood from African Green monkeys were infected. A unique virus isolated from these patients had a strange, looped and branched filamentous form, hence filovirus. In 1976 and 1979, epidemics of a haemorrhagic disease with very high mortality in northern Zaire and in southern Sudan were found to be due to two strains of a related, yet distinct filovirus, named Ebola virus. Over the next 10 years rare, sporadic cases of filovirus infections in Africa were the only continuing evidence of the existence of these viruses. Their natural host and ecology remained elusive. In 1989, a filovirus was isolated near Washington, DC, from dying cynomolgus monkeys shipped to the United States from the Philippines. Since 1990, both Ebola and Marburg viruses have re-emerged in Central Africa causing several devastating epidemics.

Virology

Nucleotide sequence analyses now places the filovirus family in the order Mononegavirales. Filoviruses are among the largest known viruses, with highly variable length (up to 14 000 nm). They undergo rapid, lytic replication in the cytoplasm of a wide range of host cells. The virions are of uniform 80-nm diameter, with a helical nucleocapsid, consisting of a central axis, 20 to 30 nm in diameter, surrounded by a helical capsid, 40 to 50 nm in diameter, with 5-nm cross-striations. A host-cell membrane-derived layer with 10-nm projections in regular array surrounds the nucleocapsid.

The virions contain a single negative-strand RNA genome ranging from 4 to 4.5 $\times 10^{6}$ daltons. The RNA is a template for at least seven polypeptides, a nucleoprotein (N), a glycoprotein (G), a polymerase (L), and four other undesignated proteins (VP40, VP35, VP30, and VP24), two of which are associated with the nucleocapsid. The surface glycoprotein is heavily glycosylated. An abundant, but poorly glycosylated protein, VP40, and the nucleoprotein (N) are associated with the nucleocapsid. There is apparently close identity at the glycoprotein level among Asian filoviruses, but not African filoviruses.

Epidemiology

In 1967, epidemiological investigations revealed that 20/29 persons with blood contact with Marburg infected monkeys became infected, and four of 13 exposed to tissue culture. Five of the secondary cases resulted from person-to-person contact at home or in hospital. About 400 to 600 animals originating from four shipments reached Europe from Uganda over a 3-week period. Data on concurrent Belgrade enzootics showed an unusually high mortality characterized by ongoing transmission during 6 weeks quarantine; 46/99 animals died from a first shipment, and 20 and 30 from another two. No evidence could be found of epizootics in Uganda, but later some indirect, controversial information emerged that there had at that time of the outbreak been excess deaths in monkey colonies in islands near Lake Kyoga, north of Lake Victoria. Since then there were three isolated, primary human Marburg infections and two secondary cases in tourists or expatriate residents; one a traveller in Zimbabwe and two others from the Mount Elgon region of western Kenya, not far from the shores of Lake Victoria. Extensive epidemiological investigations in Zimbabwe and on Mount Elgon revealed no clues of the origin of these infections. In May 1999, an outbreak of Marburg virus disease occurred in Watsa and Durba in eastern Democratic Republic of Congo. There were an estimated 76 cases with 52 deaths in miners and their families. The common risk factor in miners was illegally entering and working in an officially closed gold mine in an area with major rebel fighting and in which investigations proved difficult and dangerous. Since then, sporadic reports suggest that the suspect outbreak may be ongoing.

Nearly a decade after the Marburg outbreak, simultaneous outbreaks of another lethal haemorrhagic fever struck in northern Zaire (Republic of Congo) and Sudan in 1976. Two more filoviruses were isolated, Ebola (Zaire) and Ebola (Sudan); 280 deaths among 318 probable or confirmed cases were identified in Zaire (case fatality 88 per cent). The index case may have been a recent traveller in the northern Equateur region of Zaire who visited the clinic of a mission hospital in Yambuku. The subsequent nine cases, however, had all received treatment for other diseases at the hospital. The major risk factor was receiving an injection at this hospital. Eleven of the 17 staff members of the hospital died, and the outbreak only terminated when the hospital was closed. There was subsequent dissemination in surrounding villages to people caring for sick relatives, attending childbirth, or through sexual intercourse. The following year, 1977, a single fatal case was identified in Tandala, also in northern Zaire. Also in 1976, an outbreak of a similar disease occurred in southern Sudan, with the index case in a single cotton weaving factory. There were 151 deaths among 284 cases identified (case fatality about 53 per cent). The focus of the infection was in the town of Nzara where the factory was located, and spread was to close relatives. The epidemic was augmented by high levels of transmission at nearby Maridi hospital following transfer of one of the Nzara patients, and further cases transferred to Juba and Khartoum. There were 203 cases in Maridi, 93 of which were probably infected in the hospital and 105 in the community. Forty one staff members died, and at the height of the epidemic all wards contained patients with overt haemorrhage. In 1979, there was a similar outbreak in Sudan when 22 of 34 infections (65 per cent) were fatal. Though closely related, the viruses from Zaire and Sudan were found to be distinct. The two virus strains isolated in Sudan in 1976 and 1979, however, are identical.

A large outbreak of Ebola virus disease caused by the Zaire strain occurred in 1995 in Kikwit, Democratic Republic of Congo, resulting in 315 cases with 81 per cent case fatality. Eighty cases (25 per cent) occurred among health-care workers, and the epidemic centred again around the hospital with secondary spread in the community. The outbreak was terminated by the initiation of barrier-nursing techniques, health education efforts, and rapid identification of cases

In 1994 and 1995, 49 patients with haemorrhagic symptoms were hospitalized in north-eastern Gabon. Two other epidemics (spring and fall 1996) occurred in the same province, one of which was the result of contact of a number of young people with the carcass of a dead chimpanzee. This chimpanzee was later cooked and eaten. Infection was associated with handling the carcass or meat of the dead animal, but was not associated with eating cooked meat. A single case, infected with a closely related strain, occurred in a veterinarian working in Côte d'Ivoire. In 2001 further outbreaks occurred in Gabon.

Sudan virus Ebola haemorrhagic fever re-emerged in Uganda in August 2000 in a widespread epidemic which only terminated in January 2001; 425 presumptive cases were recorded from three districts in an estimated population of 1.8 million. The first cases came apparently from rebel areas, and no information is available on the source of this outbreak. There were 224 deaths (case fatality 53 per cent); 29 health-care workers were infected. Infection of 14/22 health-care workers after establishing isolation wards required reinforcement of infection control measures. Two distant focal outbreaks in Uganda were initiated by movement of infected contacts.

In 1989 and early 1990, a filovirus closely related to Ebola virus was isolated from cynomolgus monkeys recently imported from the Philippines (in quarantine facilities in Reston, Virginia, in Texas and in Pennsylvania) into the United States. No link with Africa or African animals could be identified and this must be considered at present an Asian filovirus. Pathogenicity for cynomolgus monkeys was uncertain because of a high rate of concurrent infection with Simian haemorrhagic fever virus (SHFV), a DNA virus which is a known, severe simian pathogen unrelated to the Filoviridae. Evidence for ongoing epizootics and transmission was found in the Philippine export facilities which had provided the monkeys. Further importations of infected monkeys into Italy and the United States have occurred since 1990.

In the Philippines, there was no illness in any individuals associated with infected monkeys, and no association between seropositivity and other risk factors, such as bites, scratches, or eating monkey meat. In the facility at Reston, Virginia, five animal handlers had a high level of daily exposure to infected and dying animals, and four of these developed antibodies. One cut his finger while performing a necropsy on an infected animal. Daily monitoring of this individual revealed transient

viraemia and seroconversion, but neither he nor his colleagues had any illness attributable to filovirus infection.

Ecology

Transmission from the unknown natural reservoir to humans is rare. Searches for evidence of virus infection in many species of animals captured in central African countries have failed to provide any clues. Bats remain highly suspect, since they were indirectly implicated in the Kitum cave cases of Marburg disease and were also present in the sugar cane-processing factory where the index cases of both 1970s Sudan outbreaks worked.

Transmission and risk factors

Person-to-person spread has been the major mode of transmission in epidemics, with contact with patients ill with Ebola is the most important factor. Other risk factors are contaminated needles, blood or secretions, preparation of a body for burial, or, occasionally, sexual contact. Epidemiological studies do not suggest spread through casual contact or by aerosol transmission. The mode of acquisition of primary infection is totally unknown.

The most significant risk factor for the monkeys infected in the epizootic in the Philippines was being an occupant of a gang cage (six-fold increase of risk, $p < 0.001$, OR 5.96, 95 per cent CI 2.87–12.38). Ebola (Reston) has been identified at high titre in respiratory secretions in monkeys, and respiratory transmission at close quarters may be a factor in epizootics with this virus.

Laboratory infections and bioterrorism

The outbreak of Marburg virus in 1967 was in individuals handling fresh monkey tissues or contaminated equipment without gloves or other protective clothing. Otherwise there has only been one reported laboratory-acquired infection (needlestick) with Ebola virus in 1976. Because of its lethal potential, Ebola has been a candidate for biological warfare. Little information is available, but it has been handled extensively in biological research, and further accidental infections are said to have occurred with the death of one scientist in a laboratory in the former Soviet bloc. The key to safe laboratory handling of this virus is extreme care in avoiding accidental inoculation. Ebola has been named as a candidate for biowarfare or bioterrorism, but without extensive biological modification it is unsuitable for dissemination in this way.

Disease

Marburg and Ebola diseases are clinically indistinguishable. The incubation period is 3 to 10 days, shorter with needle transmission. The illness-to-infection ratio is high, though it is clear that asymptomatic infections occur. In contrast Ebola (Reston) virus is uniformly asymptomatic.

Onset is abrupt with fever, severe headache, myalgia, arthragia, conjunctivitis, and extreme malaise. Sore throat is often associated with severe swelling and dysphagia, but no exudative pharyngitis. A papular, eventually desquamating rash may occur. In non-human primates, petechiae are striking. Abdominal pain and cramping followed by diarrhoea and vomiting develop on the second or third day of illness. Jaundice is not a feature. There is invariably biochemical evidence of hepatic disease with elevated aspartate transaminase (AST) levels maximal by day seven of illness. Bilirubin is not elevated, and alanine transaminase (ALT) is disproportionately low. Bleeding begins about the fifth day of illness, most commonly from the mucous membranes. Death is associated with hypovolaemic shock and severe bleeding. Infection in pregnancy results in high maternal fatality and virtually 100 per cent fetal death. Central nervous system involvement has led to hemiplegia and disorientation, and sometimes frank psychosis. Even in convalescence, patients show prolonged weakness, severe weight loss, and in a few survivors serious but reversible personality changes are recorded, namely confusion, anxiety, and aggressive behaviour. Blindness has been recorded as a sequel.

Ten of the 29 known primary Marburg infections died (35 per cent). No fatalities occurred among the 10 secondary cases, overall mortality was 10/39 (25.6 per cent). The mortality ratios during the two epidemics of Ebola disease in Sudan were 55 and 65 per cent, while that during the Zaire epidemic in 1976 was 88 per cent. In the Kikwit epidemic of 1995, mortality was 81 per cent, and in Uganda in 2000, 53 per cent, though it was much higher in children (80 per cent).

Animal models

The monkey has been the most successful animal used for the study of Marburg and Ebola viruses, and has been used extensively for the study of pathogenesis of filovirus infection.. The ability of any of the viruses to kill guinea pigs is variable. Ebola (Zaire) kills guinea pigs consistently after several adaptive passages, the Sudan variant and Marburg virus do not. Only the Zaire virus is lethal for suckling mice.

Pathogenesis and immunopathogenesis

High titres of virus are found in serum and tissues taken at autopsy, and particles may be seen in large numbers with some obvious tropism for reticuloendotheial cells. The most profound physiological alteration, invariably associated with death, is hypotensive shock. Fatal infection is marked by absent specific IgG and barely detectable IgM, whereas in survivors early and increasing levels of Ebola-specific IgG against viral nucleoprotein (NP) and 40-kDa viral protein (VP40) is followed by activation of cytotoxic T cells. In fatal cases, DNA fragmentation in blood leucocytes and levels of 41/7 nuclear matrix protein in plasma indicate that massive intravascular apoptosis proceeds during the 5 last days of life. In survivors, upregulation of FasL, perforin, CD28, and IFNg messenger (m)RNA in peripheral blood mononuclear cells coincide with clearance of circulating viral antigen. In survivors there is also early activation of T cells, evidenced by mRNA patterns in peripheral blood mononuclear cells and marked release of IFNg in plasma. It is clear that events very early in Ebola virus infection determine control of viral replication, apoptosis of immune cells and possibly other cells, and recovery or death.

Bleeding is prominent, manifest as petechiae, uncontrolled bleeding from venepuncture sites and haemorrhagic effusions. Thrombocytopenia is invariable but bleeding is not usually of sufficient volume to account for the shock, nor is it associated with solid evidence of disseminated intravascular coagulation (DIC) in the small number of animals or humans studied so far, although much has been written about DIC in Ebola and Marburg disease. As in Lassa fever, platelet dysfunction has been described in experimentally infected primates, in which there is a decline in *in vitro* platelet aggregation beginning 1 to 3 days prior to the onset of bleeding and shock and progressing to virtually no aggregation at death. Liver enzymes (AST and ALT) are raised, but the rise in AST is disproportionately higher than ALT, as was described in the early Marburg cases.

At autopsy both Marburg and Ebola infected humans and primates show widespread haemorrhagic diathesis into skin, membranes, and soft tissue. There is focal necrosis in liver, lymph nodes, ovaries, and testis. Most prominent are eosinophilic inclusion bodies in hepatocytes (Councilman-like), without significant inflammatory response.

Several individuals in direct contact with blood or infectious secretions during two outbreaks in Gabon did not develop symptoms, but seroconverted with IgM and IgG between 2 and 4 weeks following exposure. Acute Ebola infection was confirmed by detection of viral genomic (negative-stranded) RNA in peripheral blood mononuclear cells for 2 weeks after exposure, together with positive-stranded viral RNA, indicating viral replication. These individuals mounted an early, strong inflammatory response, with high levels of IL-1a, IL-6, TNFa, MCP-1, and MIP-1a/b, but without evidence of an immediate T-cell response. This unexpected observation suggests that the early inflammatory response is able, in some individuals, to control viral replication and disease.

Diagnosis

Care should be taken in both drawing and handling blood specimens since virus titre may be extremely high, and the virus is stable for long periods even at room temperature. Sera may be safely handled for immunological tests by inactivating with gamma irradiation, or, if this is unavailable, heating for 60°C for 30 min. High or rising titre filovirus-specific IgG is diagnostic as is the presence of IgM by IFA. An antigen detection ELISA system has been found to be of considerable use in monitoring epidemics and epizootics. Virus may be isolated and identified within 2 to 3 days if suitable containment facilities are available. Polymerase chain reaction (PCR) assays are available for acute diagnosis.

Immunofluorescent antibody tests (IFA) used for serological studies is unreliable at low-titre or in the absence of a history of clinical disease. Antibody, sometimes with high prevalence, has been reported in monkeys and humans from many geographic locations, including unlikely populations such as Cona Indians from Central

America and Alaskans. Newer generation tests using recombinant antigens appear to have reduced the number of non-specific reactions whilst retaining sensitivity.

Patient management

Fluid, electrolyte, respiratory and osmotic imbalances should be managed carefully. Patients may require full intensive care support, including mechanical ventilation, along with blood, plasma, or platelet replacement. The maintenance of intravascular volume is a particular challenge but every effort is justified since the crisis is short lived, and complete recovery can be expected in survivors. There is no specific therapy, and the value of immune plasma is unproven.

Control

Since the reservoir(s) of the viruses are not known, no specific precautions can be identified which would avoid infections from the natural source of the viruses. Interruption of person-to-person spread of the virus is essential to control. Early institution of safe and orderly care of the ill should be set up with effective surveillance of high-risk contacts and prompt isolation of further cases.

Vaccine

Recent vaccine candidates have been described in animal studies, including a DNA vaccine which protected guinea pigs and monkeys. Human vaccines are not available at the time of writing, but may become so shortly.

Further reading

Baize S, Leroy EM, Georges-Courbot MC, *et al*. (1999). Defective humoral responses and extensive intravascular apoptosis are associated with fatal outcome in Ebola virus-infected patients [In Process Citation]. *Nature Medicine* **5**, 423–6.

Baron RC, McCormick JB, Zubeir OA (1983). Ebola virus disease in southern Sudan: hospital dissemination and intrafamilial spread. *Bulletin of the World Health Organanization* **61**, 997–1003.

Ebola hemorrhagic fever: lessons from Kikwit, Democratic Republic of the Congo (1999). *Journal of Infectious Diseases* **179** (Suppl. 1).

Leroy EM, Baize S, Volchkov VE, *et al*. (2000). Human asymptomatic Ebola infection and strong inflammatory response. *Lancet* **355**, 2210–5.

MacDonald R (2000). Ebola virus claims more lives in Uganda. *British Medical Journal* **321**, 1037.

Report of a WHO/ International Study Team (1978). Ebola haemorrhagic fever in Sudan, 1976. *Bulletin of the World Health Organanization* **56**, 247–70.

Report of a WHO/ International Study Team (1978). Ebola haemorrhagic fever in Zaire, 1976. *Bulletin of the World Health Organanization* **56**, 271–93.

Sullivan NJ, Sanchez A, Rollin PE, Yang ZY, Nabel GJ (2000). Development of a preventive vaccine for Ebola virus infection in primates. *Nature* **408**, 605–9.

7.10.17

Papovaviruses

K. V. Shah

General description

Papovaviruses are small, spherical, non-enveloped, doubled-stranded DNA viruses that multiply in the nucleus. Viruses of the papovavirus family infect a wide variety of species including man and are largely host specific. They fall naturally into two subfamilies, papillomaviruses (wart viruses) and polyomaviruses. Papillomaviruses and polyomaviruses differ in many significant ways. The genetic information of papillomaviruses is carried on only one DNA strand but that of polyomaviruses is distributed over both strands. Papillomaviruses infect surface epithelia and produce disease at these sites. Polyomaviruses are carried by viraemia, after initial multiplication at the site of entry, to affect internal organs such as the kidney and the brain. Viruses of both subfamilies produce experimental tumours in laboratory animals but only papillomaviruses are related to naturally occurring cancers. Within each subfamily the viruses are immunologically related and share nucleotide sequences but the two subfamilies are distinct.

More than 100 human papillomaviruses have been recognized, 35 types infecting mucous membranes (genital and respiratory tracts, and the oral cavity). Human papillomaviruses are the aetiological agents of skin warts, genital warts, respiratory papillomatosis, and papillomas at other mucosal sites (e.g. mouth, eye). Infection with some genital tract human papillomaviruses causes cervical cancer, one of the most common female malignancies in the world. Human papillomaviruses contribute to cancers at other sites.

Two polyomaviruses, BK virus and JC virus, infect man. JC virus is the aetiological agent of progressive multifocal leucoencephalopathy, a fatal demyelinating disease of immunodeficient people. Because of the emergence of AIDS, it is now more frequent and is found in younger people. BK virus is associated with haemorrhagic cystitis in bone marrow transplant recipients, and with renal failure in renal transplant recipients.

Human papillomaviruses (HPVs)

Human papillomaviruses cannot be propagated in tissue culture and require nucleic acid hybridization assays for their identification. Their double-stranded circular genome contains about 8000 base pairs, divided into an early region, necessary for transformation, a late region, encoding for capsid proteins, and a regulatory region, containing control elements (Fig. 1). Open reading frames of the viral genome are located on one strand: E1 to E8 in the early region and L1 and L2 in the late region. The functions assigned to the different open reading frames are listed in Table 1.

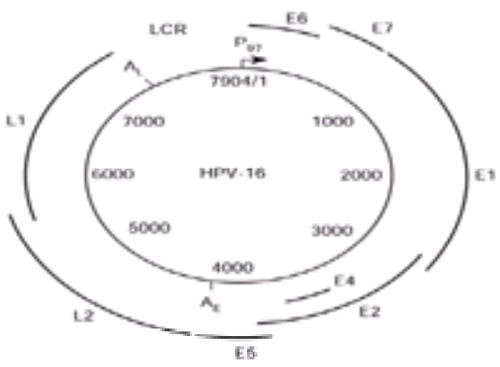

Fig. 1 Genomic map of HPV-16. On the inner circle, P97 represents the transcriptional promoter and A_E and A_L designate early and late polyadenylation sites. The location of the early-region open reading frames (E1–E8), the late-region open reading frames (L1, L2), and of the long control or regulatory region (LCR) are shown. (Reproduced from Shah and Howley (1990), with permission.)

Human papillomaviruses only infect humans. They show a marked degree of cellular tropism. Mucosal human papillomaviruses do not readily infect cutaneous epithelia and cutaneous human papillomaviruses are rarely present on mucous membranes. Infection is initiated when, after minor trauma (e.g. during sexual intercourse or after minor skin abrasions), the basal cells of the epithelium come in contact with infectious virus particles. The virus stimulates the proliferation of basal cells. The early-region open reading frames are expressed in all layers of the infected epithelium, but expression of the late-region open reading frames and synthesis of viral particles occur only in the upper differentiating and keratinizing layers.

Important disease associations and characteristics of mucosal HPVs are listed in Table 2; the genital tract is the reservoir for all but a few mucosal human papillomaviruses and genital human papillomavirus infections constitute the most common viral sexually transmitted disease. Genital human papillomaviruses may sometimes infect non-genital mucosal sites, for example, the respiratory tract, the mouth, and the conjunctiva. Transmission of genital tract HPV types 6 and 11 from an infected mother to the baby at birth results in juvenile onset recurrent respiratory papillomatosis. Infection with two types, HPV-13 and HPV-32, appears to be confined to the mouth.

Table 3 lists disease associations of cutaneous HPVs, transmitted by direct contact with infected tissue or by contact with a contaminated object.

Anogenital warts*

Anogenital warts (condylomas) are the most commonly recognized clinical manifestations of genital HPV infections. More than 90 per cent of condlyomas result from infections with HPV-6 and HPV-11. It is estimated that in the United States there are more than a million annual consultations with private physicians for anogenital warts.

Epidemiology

Genital and anal warts are most common between the ages of 16 and 24 years. They are transmitted by direct sexual contact to 60 per cent of sexual partners of

people with genital warts. Rarely, genital lesions are secondary to common warts on non-genital areas. Both anogenital warts and laryngeal papillomatosis may occur in children whose mothers had vulval warts at the time of delivery. Anogenital warts in children can also be due to close but non-sexual contact within a family or can be secondary to common skin warts, but in many cases sexual abuse by an infected adult is responsible.

Clinical features

The incubation period is between 3 weeks and 8 months (mean 2.8 months). In men, condylomata acuminata (exophytic condylomas) most often appear on areas exposed to coital trauma—the glans penis, coronal sulcus, prepuce, and terminal urethra. The soft fleshy vascular tumours are usually multiple and may coalesce into large masses (Fig. 2). Sessile or papular warts are more likely to occur on dry areas such as the shaft of the penis (Fig. 3). The raised pink or grey lesions, 0.5 to 3 mm in diameter, may occur alone or with exophytic condylomas. Subclinical HPV lesions (flat condylomas) are identiified by examining the genitalia with magnification after the application of 5 per cent aqueous acetic acid solution. The affected areas are slightly raised and shiny white (acetowhite), with a rough surface (Fig. 4). Flat condylomas affect the same areas as exophytic condylomas.

Fig. 2 Condylomata acuminata (exophtic condylomas) of the penis.

Fig. 3 Sessile (papular) warts of the penis.

Fig. 4 Subclinical HPV lesions (flat condylomas) of the penis after application of 5 per cent aqueous acetic acid.

Perianal warts are usually exophytic and in moist conditions around the anus may reach a large size. In 50 per cent of cases, condylomas also appear in the anal canal (Fig. 5). Areas of acetowhite epithelium indicative of subclinical HPV infection may be associated with perianal warts or occur alone.

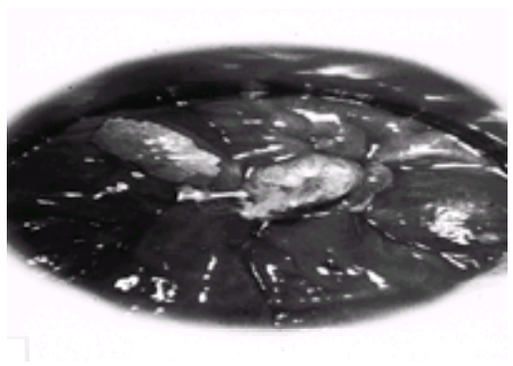

Fig. 5 Condylomata acynubate of the anal canal in an anoreceptive homosexual man.

In women, exophytic condylomas are the most common HPV lesions (Fig. 6). They appear at the fourchette and adjacent areas, and may spread to the rest of the vulva, the perineum, anus, vagina, and cervix. Multiple sessile warts may affect the labia and perineum. Subclinical HPV infection presents as slightly raised acetowhite lesions: the fissuring of these may cause dyspareunia. About 15 per cent of women with vulval warts have exophytic condylomas on the cervix. Subclinial infection is more common, and consists of acetowhite lesions with punctation due to capillary loops, which can be identified by colposcopy (Fig. 7). Large, exophytic vulval condylomas may develop during pregnancy and may become so large that they compromise delivery. Most regress postpartum.

Fig. 6 Condylomata acuminata of the vulva.

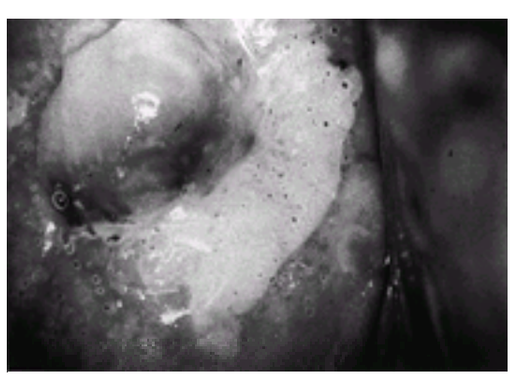

Fig. 7 Subclinical HPV infection of the cervix.

Intraepithelial neoplasia comprises Bowen's disease, bowenoid papulois, and carcinoma *in situ*. They may be associated with genital warts but contain sequences of HPV-16 or HPV-18 and may become malignant.

Diagnosis and management

Genital warts must be distinguished from Fordyce's spots, fibroepithelial polyps, molluscum contagiosum, and the papillar lesions of secondary syphilis. Intraepithelial neoplasia may be difficult to distinguish; lesions that appear atypical or respond poorly to treatment must be biopsied early.

Associated sexually transmitted diseases must be excluded. Sexual partners should be examined. Intraepithelial neoplasia must be excluded. Cervical cytological examination should always be done on women with vulval warts and on female partners of men with penile warts.

No specific antiviral treatment is available. The application of podophyllin or other cytotoxic agents, such as 5-fluorouracil and tricloracetic acid, is often unsuccessful. Warts may be destroyed with cryotherapy by liquid nitrogen or a nitrous oxide cryoprobe, electrocautery, electrodessication, and scissor excision. Interferons have been used in the treatment of persistent anogenital warts. A topical cream, which can be self-administered and is immunomodulatory, has become available recently for the treatment of genital warts.

Cervical cancer (see Chapter 21.5)

Human papillomavirus DNAs are recovered from more than 90 per cent of cases of invasive cervical cancer and squamous intraepithelial lesions of the cervix, which precede invasive cancer. The viral genome is present in the tumour cells of primary as well as metastatic cervical cancer. The progression from low-grade squamous intraepithelial lesions to invasive cancer may take more than 10 years; human papillomaviruses are found throughout this disease process. The viruses are recovered much less frequently from cytologically normal women of comparable age. In prospective studies of women with normal cervical cytology, the presence of HPV is a strong risk factor for the subsequent development of squamous intraepithelial lesions.

Certain HPV types are preferentially associated with invasive cancers. From their distribution in normal individuals and in preinvasive and invasive cervical disease, genital tract HPVs have been categorized as 'high-risk', 'intermediate-risk', and 'low-risk' types (Fig. 8; Table 2). HPV-16 and HPV-18 are the predominant viruses in invasive cancers and account for 40 to 60 per cent and 5 to 20 per cent, respectively, of HPV-positive cancers in different studies. About a dozen additional types of HPV are found in small proportions of invasive cancers. The 'low-risk' HPVs are almost never detected in invasive cervical cancers.

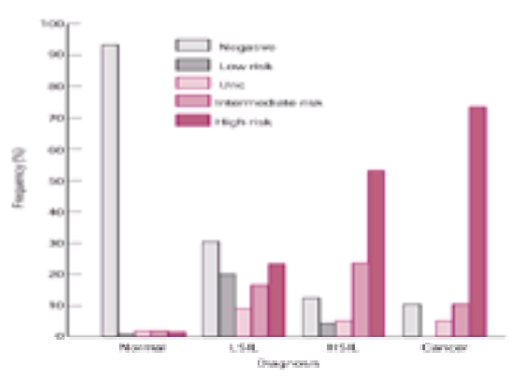

Fig. 8 Distribution of HPV types in normal women and in preinvasive (low-grade and high-grade squamous intraepithelial lesions (SILs)) and invasive cancer. In each diagnostic category, specimens are grouped as containing high-risk, intermediate-risk, and low-risk HPV types (see Fig. 8; Table 2), or as containing unclassified HPVs (Unc), or as negative (Neg). (Modified and reproduced from Lörincz *et al.* (1992), with permission.)

Comparisons of different HPV types for their ability to transform human keratinocytes *in vitro* show that HPV-16 and HPV-18, those most clearly associated with naturally occurring cervical cancers, also have the greatest oncogenic potential in laboratory studies. The transforming functions of HPVs are localized to open reading frames E6 and E7; these are the frames consistently expressed in naturally occurring HPV-positive cancers. The viral genome is integrated into the cellular DNA in most cervical cancers. The break in the circular viral genome that is required for integration occurs most frequently in the E1/E2 region and results in an enhanced expression of the transforming E6 and E7 open reading frames. The transforming HPV proteins E6 and E7 interact with cellular tumour-suppressor proteins p53 and Rb, respectively. It is likely that the oncogenic effect of HPVs is mediated partly by their ability to inactivate the tumour-suppressor proteins which normally regulate the cell cycle.

Epidemiology

Human papillomavirus infections of the genital tract are extremely common in sexually active populations. In young sexually active women, prevalence of HPV

infection as measured by the detection of HPV DNA in genital tract specimens by the sensitive polymerase chain reaction may be greater than 40 per cent. The prevalence decreases with increasing age. Most of these infections are found in women with normal cervical cytology and undoubtedly resolve without leaving a trace. Only a small proportion of infections progress to squamous intraepithelial lesions and then to invasive cancer. The cofactors that might be required for this progression are not conclusively identified, but smoking, use of oral contraceptives, parity, presence of other sexually transmitted diseases, and diet are incriminated to some degree, in some studies. Human immunodeficiency virus infection, and associated immunosuppression, leads to a much higher prevalence, and longer persistence, of HPV infections and to greater incidence of squamous intraepithelial lesions.

Prevention and control of cervical cancer

Screening for cervical cytological abnormalities by Pap smear and treatment of preinvasive and invasive cancers identified by screening have been credited with the decrease in incidence of cervical cancer and mortality due to the disease that has been observed in many developed countries over the last 40 to 50 years. The recognition that HPVs are linked aetiologically to cervical cancers has led to the exploration of HPV-based strategies for prevention and control of cervical cancer.

Clinical management

Women who have cytological abnormalities which are low grade or of uncertain significance may benefit from an HPV diagnosis. The presence of cancer-associated HPVs (high risk plus intermediate risk) would indicate a need for closer monitoring and colposcopy; HPV-negative women would be monitored routinely.

Prophylactic vaccines

Tests in rabbits, cattle, and dogs show that immunization of these animals with conformationally correct L1 capsid protein of their respective papillomaviruses protects them against papillomavirus-induced disease. Vaccines based on HPV L1 proteins have been formulated and tested in human volunteers to evaluate their safety. It is anticipated that the efficacy of these vaccines will soon be tested in clinical trials.

Therapeutic vaccines

Human papillomavirus-associated cancers express HPV E6 and E7 proteins in their tumour cells. Candidate therapeutic vaccines targeted to these proteins are being developed for the treatment of high-grade squamous intraepithelial lesions and invasive cancer.

Cancers at other lower anogenital tract sites

Human papillomavirus infections are very common on the vulva, vagina, penis, perineum, and anus. Synchronous neoplasia at multiple sites in the female lower genital tract is almost always associated with HPVs, especially HPV-16. Carcinoma of the vulva is aetiologically heterogeneous. Vulval cancers occurring in younger women are associated with HPVs but the typical squamous cell carcinoma of the vulva in older women is not. Neoplasia of the anal canal, seen frequently in HIV-seropositive homosexual men, is strongly associated with HPVs.

Cancer of the oropharynx

Some pharyngeal cancers, especially tonsillar cancers, appear to be associated with high-risk HPVs, most often HPV-16. The HPV-positive cancers are characterized by more frequent basaloid pathology, less frequent p53 mutations, and better prognosis, than HPV-negative cancers.

Respiratory papillomatosis

This rare disease is most common in children under the age of 5 years. It may become life threatening if it obstructs the airways. Papillomatosis usually involves the vocal cords and presents with hoarseness or voice change. Papillomas may recur after surgical removal.

HPV-6 and HPV-11, genital tract HPVs that are responsible for most of the exophytic genital warts also cause respiratory papillomatosis. Infants are infected during passage through the birth canal. In adults, transmission may occur by sexual contact. Respiratory papillomas very rarely progress to invasive cancer. Irradiation of papillomas with X-rays (a practice now discontinued) increases the risk of malignancy.

Caesarean delivery for mothers who are found to have genital warts or are infected with HPV-6 or HPV-11 would reduce the risk of juvenile onset respiratory papillomatosis, but it is not generally recommended because of the small risk of disease following perinatal infection. Interferon therapy is not very effective in the treatment of respiratory papillomas.

Human papillomaviruses in the oral cavity

The genital tract HPVs, especially HPV-6 and HPV-11, may infect the oral cavity (Table 2) and are readily recovered from oral lesions diagnosed histologically as condylomas or warty lesions. Focal epithelial hyperplasia of the mouth is distributed worldwide but is highly prevalent in indigenous populations of Central and South America and of Alaska and Greenland; it is aetiologically associated with HPV-13 and HPV-32. These two types are found exclusively in the oral cavity.

Skin warts (see Chapter 23.1)

Skin warts and verrucas may occur anywhere on the skin and are morphologically diverse. They are most common in older children and young adults. Except in the rare condition known as epidermodysplasia verruciformis (see below), they almost never become malignant. Most regress within 2 years. Specific HPV types are strongly associated with specific types of warts (Table 3).

Epidermodysplasia verruciformis

This is a rare, lifelong disease in which a patient has extensive warty involvement of the skin that cannot be resolved. It generally begins in infancy or childhood with multiple, disseminated, polymorphic wart-like lesions on the face, trunk and extremities that tend to become confluent. The warts are either flat or reddish-brown macular plaques that resemble pityriasis versicolor. In about a third of the cases, foci of malignant transformation occur in macular plaques in areas of the skin exposed to sunlight. The tumours are slow growing and rarely metastasize.

Epidermodysplasia verruciformis is often familial. Patients sometimes have a history of parental consanguinity. The pattern of inheritance is suggestive of an X-linked recessive disease resulting in an immunological inability to resolve the infection. The flat warts yield the same HPV types as those of normal individuals but a very large number of HPVs that are seldom encountered in normal individuals are recovered from the macular plaques (Table 3). It is unclear how patients with epidermodysplasia verruciformis become infected with these particular papillomaviruses. The factors that contribute to the occurrence of carcinoma in these patients therefore include a genetic defect, infection with specific HPVs, for example, HPV-5 and HPV-8, and exposure of the affected area to sunlight.

Non-melanoma skin cancers

HPV sequences have been recovered frequently from normal skin, from psoriatic lesions, and from non-melanoma skin cancers of normal and immunosuppressed populations. The sequences represent cutaneous HPV types, epidermodysplasia verruciformis (EV)-associated HPVs and many novel HPV sequences. It appears that the normal skin is seeded with many HPV types but it is not clear to what extent they contribute to the development of non-melanoma skin cancers.

Human polyomaviruses

In 1971 BK virus was isolated from the urine of a renal transplant recipient and JC virus was recovered from the brain of a patient with progressive multifocal leucoencephalopathy. The viruses have a double-stranded DNA genome of about 5000 base pairs, which is divided into an early region encoding viral T proteins, a

late region encoding viral capsid proteins, and a non-coding regulatory region. The early and late regions are transcribed from different strands of the viral DNA. Although BK and JC viruses are homologous for 75 per cent of their nucleotide sequence, the infections are readily distinguishable by conventional tests.

Infection occurs in childhood and is largely subclinical. Most children acquire antibodies to BK virus by the age of 10; infection with JC virus occurs at a later age. Early acquisition of antibodies suggests that infection occurs by the respiratory route. Both viruses establish latent, often lifelong, infection in the kidney and are occasionally excreted in the urine of normal people. Reactivation in immunodeficient people is responsible for most associated illnesses. The viruses are reactivated in pregnancy but without any apparent harm to the mother or the newborn.

BK virus-associated illnesses

Reactivation of BK virus in renal transplant recipients may cause ureteric obstruction, a late and uncommon complication of transplantation. Reactivated BK virus infection in patients with renal transplants has recently been linked and renal dysfunction and graft rejection. In bone-marrow transplant recipients receiving allogeneic marrow, late onset haemorrhagic cystitis and BK viruria are strongly correlated. Primary BK virus infection may be responsible for an occasional case of cystitis in normal children. A case of fatal tubulointerstitial nephritis in an immunodeficient child was ascribed to primary BK virus infection. Reports of the virus in pancreatic islet cell tumours and in brain tumours are unconfirmed.

Progressive multifocal leucoencephalopathy (see also Chapter 24.14.2)

JC virus causes progressive multifocal leucoencephalopathy, a subacute demyelinating disease of the central nervous system occurring in individuals with impaired cell-mediated immunity. Until recently, it was a rare disease found mainly in older patients with lymphoproliferative disorders or chronic diseases. In the past decade, it has been seen much more frequently and the majority of cases are in younger patients, as a complication in 1 to 2 per cent of AIDS cases. It has also been recognized in children who have inherited immunodeficiency diseases or have AIDS.

The key pathogenetic event in the leucoencephalopathy is the cytocidal JC virus infection of oligodendrocytes, which are responsible for the production and maintenance of myelin. This leads to foci of demyelination that tend to coalesce and eventually involve large areas of the brain. Infected oligodendrocytes, containing large inclusion-bearing nuclei filled with abundant virus particles, surround the foci of demyelination (Fig. 9). Enlarged astrocytes often show bizarre nuclear changes but are mostly virus negative. They are found within the foci of demyelination. JC virus is disseminated haematogenously to the central nervous system, probably through virus-infected B lymphocytes. The brain may be seeded with JC virus either at the time of primary infection or when the virus is reactivated in times of immunological impairment.

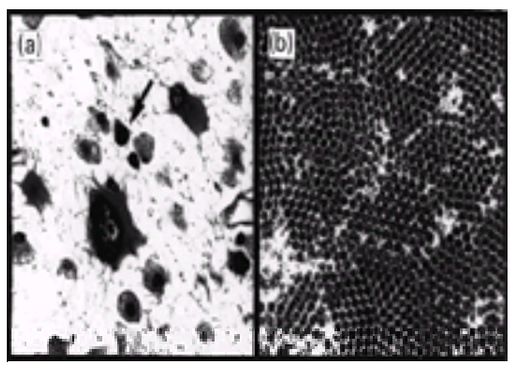

Fig. 9 A lesion of progressive multifocal leucoencephalopathy showing oligodendrocytes with enlarged, deeply staining nuclei (arrow) and giant astrocytes (left). A crystalloid array of JC virus particles in an infected oligodendrocyte nucleus (right). (Reproduced from Shah (1992), with permission.)

Progressive multifocal leucoencephalopathy starts insidiously. Early signs and symptoms indicate the presence of multifocal asymmetrical lesions in the brain and involve impairment of vision and speech, and mental deterioration. The disease is usually relentlessly progressive and fatal within 3 to 6 months but rarely it can become stabilized with survival for many years. Computed tomography and magnetic resonance imaging have been successfully used for diagnosis. Treatment with cytosine arabinoside and the presence of an inflammatory response in the brain have been associated with the few relatively successful outcomes.

---

*Includes material from *The Oxford Textbook of Medicine*, 3rd edn, pp 3366–9 (Chapter 21.7, Genital warts, J. D. Oriel).

Further reading

Binet I *et al.* (1999). Polyomavirus disease under new immunosuppressive drugs. A cause of renal graft dysfunction and graft loss. *Transplantation* **67**, 918–22. Describes BK virus nephrology in renal transplant recipients.

Cuzick J *et al.* (1999). HPV testing in primary screening of older women. *British Journal of Cancer* **81**, 554–8.

Greenlee JE (1998). Progressive multifocal leukoencephalopathy—progress made and lessons relearned. *New England Journal of Medicine* **338**, 1378–80.

IARC (International Agency for Research on Cancer) (1995). *Monograph on the evaluation of carcinogenic risks to humans volume 64, Human papillomaviruses*. IARC, Lyon. Systematic literature review of HPV–cancer link.

Koutsky L (1997). Epidemiology of genital human papillomavirus infection. *American Journal of Medicine* **102**, 3–8.

Lörincz AT *et al.* (1992). Human papillomavirus infection of the cervix: relative risk associations of 15 common anogenital types. *Obstetrics and Gynecology* **79**, 328–7.

Shah KV (1992). Polyomavirus, infection and immunity. In: Roitt IM, ed. *Encyclopedia of immunology*, pp 1256–8. Academic Press, New York.

Shah KV, Howley PM (1996). Papillomaviruses. In: Fields BN *et al*., eds. *Virology*, 3rd edn, pp 2077–109. Lippincott-Raven, Philadelphia.

Tindle RW, ed (1999). *Vaccines for human papillomavirus infection and anogenital disease.* RG Landes, Austin, TX. Multiauthored book discussing HPV vaccine candidates and strategies.

Weber T, Major EO (1997). Progressive multifocal leukoencephalopathy: molecular biology, pathogenesis and clinical impact. *Intervirology* **40**, 98–111.

7.10.18 Parvovirus b19

*B. J. Cohen**

Viruses of the subfamily Parvovirinae infect vertebrates

Introduction

Parvoviruses (family *Parvoviridae*) are widespread in nature causing disease in many animal species. They are small (23 nm), icosahederal, non-enveloped viruses (Fig. 1) containing a single-stranded DNA genome. The *erythrovirus* genus, which replicates only in nucleated red blood cell precursors includes human parvovirus B19 (B19 virus), the only member of the family *Parvoviridae* known to cause disease in humans. B19 virus was discovered in 1975 by chance as an asymptomatic infection in blood donors being screened for hepatitis B antigen.

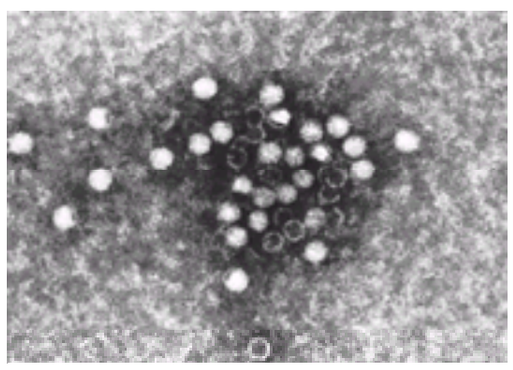

Fig. 1 Immune electron micrograph of parvovirus particles in serum from a case of aplastic crisis. Some particles are penetrated by the negative stain and appear 'empty'; other particles resist the stain and appear 'full'.

Epidemiology

B19 infection is usually spread by respiratory droplets. Contamination of hands and surfaces may also contribute. More rarely, it is bloodborne, either across the placenta or by transfusion of contaminated blood components. Infection is most common in children between 6 and 10 years. By 20 years of age, 60 to 70 per cent of the population have been infected. Susceptible adults remain at risk of infection, often following exposure to B19 virus in their own children. Epidemics occur every 4 or 5 years with peaks of infection in winter and spring.

Clinical features of parvovirus B19 infection

At least a third of B19 infections in children and adults are asymptomatic or present as non-specific febrile illness.

Erythema infectiosum

The most common specific clinical manifestation is erythema infectiosum, an erythematous rash illness (fifth disease) of childhood. The rash has an incubation period of 17 to 22 days (Fig. 2 and Plate 1) and classically the illness begins with mild fever and lassitude followed by the facial erythema referred to as 'slapped cheek disease' (Fig. 3). Subsequently the rash spreads to the trunk and limbs where it has a lacy or reticular appearance and tends to fade and recrudesce for a week or so after its initial appearance. School outbreaks are common during epidemic periods. Sporadic cases in children and adults may be misdiagnosed as rubella, streptococcal infection, or allergy. Occasionally, B19 infection presents as a purpuric rash.

Fig. 2 Slapped cheek' rash of erythema infectiosum: note circumoral pallor. (By courtesy of Dr Ken Mutton.) (See also Plate 1.)

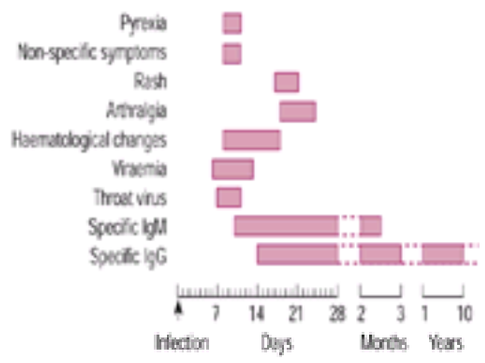

Fig. 3 Sequence of events following intranasal inoculation of volunteers with parvovirus B19.

Acute arthropathy

More than 80 per cent of adults with B19 infection (especially women) present with painful or swollen joints. An acute-onset, symmetrical polyarthritis specifically affects the small joints of the hands and feet. It usually resolves within a few weeks. In about 20 per cent of adult females with B19 infection, joint symptoms persist for more than 2 months and may resemble rheumatoid arthritis. Rheumatoid factor is usually absent and there is no erosive joint disease. No association with rheumatoid arthritis has been confirmed.

Infection in pregnant women

About 10 per cent of B19 infections in the first 20 weeks of gestation end in spontaneous abortion, a rate of fetal loss about 10 times greater than that in unaffected pregnancies. Embryopathy usually presents as hydrops fetalis 4 to 6 weeks after a maternal infection, which may be symptomatic or clinically silent. In epidemic years, 10 to 20 per cent of cases of non-immunological hydrops fetalis are associated with B19 infection. Fetal anaemia due to B19 infection may be treated with *in utero* blood transfusions. Surviving infants have no evidence of congenital disease or malformation.

Transient aplastic crisis

Interruption of erythropoiesis caused by B19 is transient and insufficient to cause clinically significant anaemia in individuals with normal red cell lifespan and function. In those with a shortened red cell lifespan, such as patients with sickle cell anaemia, B19 infection can rapidly lead to a more profound anaemia termed an aplastic crisis. In the acute phase there is erythroid aplasia and the absence of reticulocytes in peripheral blood and, in recovery, reticulocytosis and the appearance of giant pronormoblasts in the bone marrow. B19-induced aplastic crisis has also been recorded in patients with hereditary spherocytosis, b-thalassaemia intermedia, pyruvate kinase deficiency, and other red cell disorders.

Chronic anaemia in immunocompromised patients

Patients with congenital immunodeficiency, HIV infection, acute lymphatic leukaemia, or immunocompromise following organ transplantation fail to produce neutralizing antibody to B19 virus and infection becomes chronic. This results in persistent anaemia and patients may become transfusion dependent.

Prevention and therapy

A recombinant DNA-derived vaccine is being developed. To minimize spread by infected blood components, blood donor screening has been proposed, but its cost-effectiveness is unknown. The testing of plasma pools, however, is likely to become mandatory and should result in the reduction of viral load, if not the complete removal, of B19 virus from blood products. Early recognition of B19 infection in hospital patients is important for prevention of nosocomial transmission. Severe infections in immunocompromised patients are treated with high-dose intravenous normal immunoglobulin (400 mg/kg body weight for 5 or 10 days).

Laboratory diagnosis

Detection of B19 virus, by polymerase chain reaction (PCR) for example, is important for diagnosis in patients presenting in the viraemic phase of infection, including those with aplastic crisis, immunocompromise, and fetal infection. In most cases, however, the presenting symptoms of rash and arthropathy are postviraemic phenomena (Fig. 3) and the diagnosis is most commonly confirmed by detecting B19-specific IgM. The IgM response persists for 2 to 3 months after acute infection. Thereafter, B19-specific IgG remains as the sole marker of infection in the past and indicates immunity. Reinfection occurs only in the immunocompromised patient.

---

*Professor J.R. Pattison kindly wrote on Parvoviruses in the third edition of the *Oxford Textbook of Medicine*. Some of his text and Figures have been incorporated in this chapter and we are pleased to acknowledge his contribution.

Further reading

Brown KE, Young NS, Liu JM (1994). Molecular, cellular and clinical aspects of parvovirus B19 infection. *Critical Reviews in Oncology/Hematology* **16**, 1–31.

Hall SM (1990). Parvovirus B19 and pregnancy. *Reviews in Medical Microbiology* **1**, 160–7.

Prowse C, Ludlam CA, Yap PL (1997). Human parvovirus B19 and blood products. *Vox Sanguinis* **72**, 1–10.

7.10.19 Hepatitis viruses (including ttv)

N. V. Naoumov

Viral hepatitis is an ancient disease which remains an important health problem worldwide. The archetypal viral hepatitis, yellow fever (Chapter 7.10.13), is not included within this group. Over the last 30 years five major hepatitis viruses have been identified—A, B, C, D, and E (Table 1). These unrelated human viruses, different in their genome organization, biology, and epidemiology, are similar in their hepatotropism. Ten to fifteen per cent of cases of viral hepatitis are considered as non-A to E hepatitis: their aetiology remains unknown. The search for new hepatitis agents led to the identification of hepatitis G virus (HGV or GB virus-C) and TT virus. Both have been detected in a high proportion of the general population, but their pathogenic role is uncertain. The search for new agents responsible for the small proportion of patients with cryptogenic hepatitis continues. Details of symptomatology, management, and prevention of viral hepatitis are given in Section 14.20.

Hepatitis A virus (HAV)

HAV particles were detected by immune electron microscopy in 1973 in stool samples of patients with hepatitis A. The virus is classified in the genus *Hepatovirus* within the Picornaviridae family. The genome of HAV is a single-stranded, linear RNA of approximately 7500 nucleotides (Table 1). This includes a 5' non-translated region (**5'NTR**) of approximately 740 nucleotides, followed by a single long open reading frame (**ORF**) encoding a polyprotein of 2200 amino acids and a short 3' non-translated segment. After translation, the HAV polyprotein undergoes multiple cleavages by a virally encoded enzyme—3C protease. The polyprotein is considered to contain three functionally separate domains. At the aminoterminal end is domain P1 that includes the major structural polypeptides of HAV in the following sequence—VP2, VP3, and VP1. A fourth very small polypeptide, VP4, presumed to be involved in the HAV capsid formation, is located at the extreme aminoterminal end of the polyprotein. These four structural polypeptides assemble into a viral capsid containing 60 copies of each. How the viral RNA is incorporated into the virion is unknown, but both empty and RNA-containing capsids have been observed in most virus preparations. The other P2 and P3 domains of the viral polyprotein include at least six separate proteins involved in viral replication. These include 2B and 2C helicase, 3A and 3B proteins, 3C (the viral protease), and 3D (an RNA-dependent RNA polymerase).

Hepatocytes are the predominant site of HAV replication *in vivo*. Recent data indicate that HAV may also replicate within the epithelial cells of the gastrointestinal tract. However, the mechanism by which HAV reaches the liver remains unknown. Maximal HAV replication in hepatocytes occurs before serum aminotransferases increase. The virus is excreted via the biliary system into the faeces where it can be found in high concentrations around 1 to 2 weeks before the start of clinical symptoms. Viraemia is present from the earliest phase of infection. It results from HAV replication within hepatocytes. HAV differs from other picornaviruses in its non-cytolytic replication. Liver injury is immune mediated by natural killer cells, virus-specific CD8+ cytotoxic T lymphocytes, and non-specific inflammatory cells recruited to the liver. When clinical symptoms appear there is a humoral immune response and antibodies to structural HAV proteins (anti-HAV) are detectable in the serum. Initially these are mainly IgM antibodies (IgM anti-HAV) that usually persist for approximately 6 months. During convalescence, anti-HAV of IgG class become the predominant antibodies. They remain detectable indefinitely, representing protective immunity to HAV.

Hepatitis B virus (HBV)

In 1965, Blumberg and colleagues identified the surface antigen (**HBsAg**) of HBV, initially termed 'Australia antigen', and in 1970 the complete virion (a 42 nm particle) was detected by Dane and colleagues, using electron microscopy. The genome of HBV, the smallest DNA virus, contains only 3200 nucleotides (Table 1). One of the DNA strands, the 'minus' strand, is an almost complete circle containing four overlapping reading frames: precore/core, polymerase, envelope, and X genes (Fig. 1). The other ('plus') strand is shorter and varies in length. HBV belongs to the hepadnavirus family that includes similar hepatotropic DNA viruses specific for woodchucks, ground squirrels, and Pekin ducks.

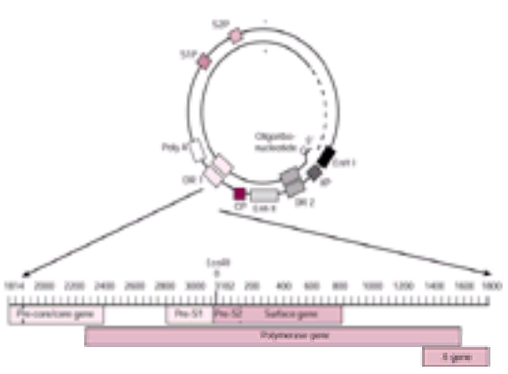

Fig. 1 Schematic representation of hepatitis B virus genome. CP, core promoter; S1P, preS1 promoter; S2P, pre-S2 promoter; XP, X gene promoter; EnhI, enhancer I; Enh II, enhancer II; DR1 and DR2, direct repeat 1 and 2; EcoRI, restriction site for EcoRI enzyme used as a starting point for numbering.

Genome organization

The envelope ORF contains three start codons separating the pre-S1, pre-S2, and S sequences. The surface gene encodes the major envelope protein (HBsAg) of 226 amino acids. The translation product of the pre-S2 and S gene is the middle envelope protein and the product of pre-S1, pre-S2, and S gene is the large envelope protein. In addition to the complete virion, many more non-infectious, 22 nm, spherical and filamentous subviral particles are produced in infected hepatocytes. HBsAg and the middle envelope protein are present in all viral and subviral particles, while the large protein is present in the virions and in some subviral filaments. The domain which binds to a specific HBV receptor (still not defined) on the plasma membrane of hepatocytes resides within the pre-S1 region.

The precore/core ORF has two start codons encoding two closely related proteins. Translation from the preC-start codon produces a precursor molecule, designated precore protein. In the endoplasmic reticulum, this protein undergoes two proteolytic steps at the amino- and carboxy-terminal ends. The resultant polypeptide is secreted from hepatocytes as hepatitis B e antigen (**HBeAg**). This is a non-structural protein, which is not essential for viral replication. However, detection of HBeAg in serum is a good marker of HBV replication. Translation from the C-start codon yields the nucleocapsid protein (**HBcAg**) of 183 amino acids. In the cytoplasm of hepatocytes HBcAg assembles spontaneously into nucleocapsid particles. HBeAg and HBcAg share about 90 per cent of the amino acids but differ substantially in their conformation. The polymerase ORF encodes the HBV polymerase protein with 832 amino acids. It has three functional domains—terminal protein, reverse

transcriptase, and RNAase H activity. The X ORF encodes a protein with 154 amino acids. The X protein is not essential for the replication of hepadnaviruses, but is believed to contribute to HBV-related hepatocarcinogenesis. It functions as a transactivator of cellular and other viral genes.

Seven different genotypes of HBV (A, B, C, D, E, F, and G) have been determined. The variations involve approximately 10 per cent of the genome. Genotype A is predominant in Central and Northern Europe, genotype D in the Mediterannean basin, genotypes B and C in Asia, and genotype E in Africa.

Viral replication

Following HBV entry into hepatocytes, the nucleocapsid is transported to the nucleus (Fig. 2). Cellular enzymes repair the open circular HBV DNA into covalently closed circular DNA (**cccDNA**), which serves as a template for synthesis of pregenomic and messenger RNAs. Viral DNA does not integrate into the host genome as part of the normal replication cycle. The pregenomic RNA is transported to the cytoplasm and serves as mRNA for translation of new core and polymerase proteins. When these three components (pregenomic RNA, core and polymerase proteins) reach sufficient quantities, they assemble into nucleocapsid particles. The polymerase protein is directly involved in the pregenomic RNA encapsidation. Inside the particles the pregenomic RNA is reverse transcribed into DNA 'minus' strand, while the RNA template is simultaneously degraded by RNAaseH. Finally, the 'plus' strand is produced, which completes a new partially double-stranded HBV DNA. Some of the newly synthesized nucleocapsids with HBV DNA are transported back to the nucleus, which maintains a stable pool of cccDNA. Others are enveloped and leave the cell as new virions. Hepadnavirus replication differs from that of retroviruses. Integration into the host genome is not obligatory during replication and functional mRNAs are produced from several internal promoters of the circular DNA genome.

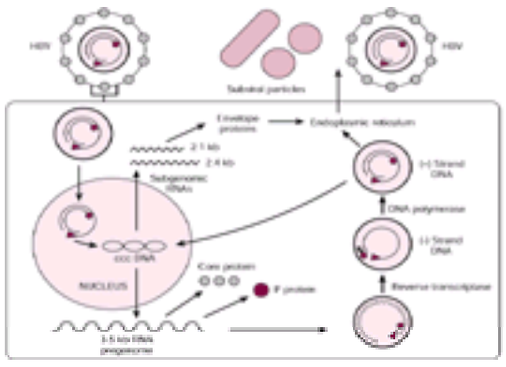

Fig. 2 Replicative cycle of hepatitis B virus.

Host immune response and pathogenesis

HBV is a non-cytopathic virus. The virus-specific cellular immune response is the main determinant of the outcome of infection. Both HLA class I- and class II-restricted T-cell responses are strong, polyclonal, and directed to multiple viral antigens in patients with acute self-limiting hepatitis B. Despite clearance of serum HBsAg, HBV DNA remains detectable by polymerase chain reaction (**PCR**) in most cases, and HBV-specific CD4+ and CD8+ T-cell reactivity has been demonstrated 10 to 20 years after acute infection. Cytokines released from these cells, especially interferon-g, exert non-cytolytic control on HBV replication without causing cell death. Thus, eradication of HBV may be rare, but the effective immune response controls HBV DNA expression and there is no liver disease. Patients with chronic HBV infection (seropositive for HBsAg) show weak virus-specific T-cell reactivity, which is the dominant cause for HBV persistence. This ineffective response, together with antigen non-specific inflammatory cells, recruited at the site of inflammation, results in progressive liver damage. During the course of chronic HBV infection, spontaneous reactivation of hepatitis may occur, associated with enhanced immune reactivity.

The humoral immune response involves antibodies directed to different HBV antigens (Table 1). It is clinically significant for: (i) diagnosis—the antibody profile in the serum, together with the result of HBsAg and HBeAg, is used to define the phase of HBV infection; (ii) prophylaxis—the development and the level of the protective antibody (anti-HBs) is used to monitor the response to vaccination; (iii) pathogenesis—the humoral immune response contributes to viral elimination from the circulation by forming immune complexes. In some cases, tissue deposition of antigen–antibody complexes is responsible for extrahepatic pathology such as glomerulonephritis, polyarteritis nodosa, arthritis, and skin changes.

Evolution of chronic HBV infection

The changes in HBV–host interactions over time and associated liver disease define three consecutive phases, particularly after vertical transmission of HBV. The early 'immunotolerant' phase is characterized by high levels of virus replication. HBeAg and HBV DNA are readily detectable in serum, while there is minimal liver inflammation. Over the years this is followed by a phase with enhanced immune reactivity to the virus, as reflected by hepatic inflammation and elevated serum aminotransferases. Serum HBeAg is still positive and serum HBV DNA level is usually lower. Some patients will progress spontaneously to the next 'non-replicative' phase, manifested by seroconversion to anti-HBe, undetectable HBV DNA (by conventional techniques), and resolution of hepatic inflammation. Persistent virus replication leads to the emergence of mutations in the HBV genome. Three groups of HBV mutants have direct clinical significance. First, surface escape HBV mutants can emerge in recipients of active and passive immunization. They contain mutations within the immunodominant 'a' determinant of the HBsAg, which abrogate the neutralizing effect of anti-HBs. The second group includes HBV mutants with impaired translation of HBeAg—HBe minus mutants. The most frequent is a G to A mutation at position 1896, which results in a precore stop codon. The third group includes mutations in the polymerase gene that may emerge during treatment with nucleoside analogues, such as lamivudine or famciclovir. The most typical is a lamivudine-associated mutation in the conserved YMDD (tyrosine–methionine–aspartate–aspartate) region of the polymerase, leading to substition of methionine with valine or isoleucine.

Hepatitis C virus (HCV, see also Chapter 7.10.20)

HCV was identified in 1989 and was shown to be the main aetiological agent of parenterally transmitted non-A, non-B hepatitis. For the first time a virus was discovered and characterized by molecular techniques without being seen or grown in culture. The virus has been placed in a separate genus of the Flaviviridae family (Table 1). The lack of an *in vitro* system supporting HCV replication has been a major limitation to the understanding of its biology and the development of antiviral compounds. Synthesis of a full-length cDNA clone, capable of generating infectious RNA transcripts, is an important step to establish a tissue culture system.

Genome organization

The HCV genome is a single-stranded RNA containing a single ORF encoding a polyprotein of 3010 to 3033 amino acids (Fig. 3). Both at the 5' and the 3' ends, it has a non-translated region (NTR). The 5'NTR consists of 341 nucleotides and is the most conserved region of the genome. It forms multiple stem–loop structures, important for ribosome entry and presumably for viral RNA replication. The 3'NTR comprises several regions, including a highly conserved sequence of 98 nucleotides at the 3' terminus, thought to be required for initiation of replication. Because HCV RNA does not replicate via a DNA intermediate, it does not integrate into the host genome. The HCV polyprotein undergoes proteolytic processing in the cytoplasm of infected cells resulting in 10 mature proteins from core to NS5B (Fig. 3). The putative nucleocapsid or core protein is conserved and highly immunogenic, containing several B- and T-cell epitopes. The envelope glycoproteins (E1 and E2) are believed to form the outer spikes of the viral envelope. The HCV E2 protein binds to the major extracellular loop of human CD81 molecule. CD81 is a cell surface protein, expressed on various cells including hepatocytes. It may act as a receptor for HCV. The first 27 amino acids at the N-terminus of the E2 region (between amino acids 384 and 410 of the polyprotein), which show a very high degree of variation, is termed hypervariable region I (**HVRI**). E2 is part of the virus envelope; it contains neutralizing epitopes, one of which appears to be in the HVRI; the immune pressure on this protein leads the selection of escape mutants. The NS3 serine protease domain and the NS4A protein form a complex that is essential for efficient polyprotein cleavage. Specific inhibition of the proteolytic activities of virally encoded proteases is regarded as a promising strategy for inhibiting HCV replication. The non-structural protein NS5B possesses RNA polymerase activity. This enzyme is also essential for HCV replication and is another important target for antiviral drug development.

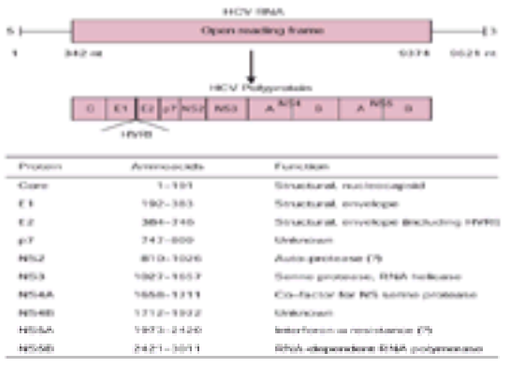

Fig. 3 Hepatitis C virus genome and proteins.

Genome variation and quasispecies

HCV is exceptionally heterogeneous. The NS5B replicase is an error-prone enzyme with no proof-reading activity. During HCV replication this generates many mutant strains which are selected on the basis of their fitness. Based on a phylogenetic analysis of the HCV core, E1, and NS5 regions, six major HCV genotypes (from 1 to 6) have been defined, with a further division into subtypes (1a, 1b, 2a, 2b, etc.). Isolates of type 1 are widely present throughout the world. Genotype 1a and 1b are predominant in North and South America and Europe. Genotypes 2 and 3 are widely distributed in many countries, but are rare in Africa. Genotype 4 is predominant in north and central Africa, especially Egypt, while genotype 5 is most frequent in southern Africa. Genotype 6 is responsible for many HCV infections in Hong Kong and Vietnam. HCV genotypes 7, 8, and 9 have been identified only in Vietnam where they represent almost 20 per cent of all HCV infections. Despite the substantial genomic variations between different HCV genotypes, both clinical and virological data show no significant phenotypic differences in the severity of liver damage or the potential to cause hepatocellular carcinoma. Genotype 1b responds less well than genotypes 2 or 3 to antiviral treatment.

In an individual host, the HCV population is a mixture of closely related, but heterogeneous, RNA sequences centred around one dominant viral sequence. The heterogeneous isolates in a single patient are termed 'quasispecies'. This is commonly based on the genomic variability within the HVRI region. Viral diversity increases during chronic HCV infection as a result of immune escape from antibodies directed to this hypervariable region.

Host immune response and pathogenesis

Clinical evidence suggests that HCV is not cytopathic for infected cells. A most striking feature of HCV is the high rate of chronic infection. The immune response is believed to play a central role in viral clearance and pathogenesis. Neutralizing antibodies are produced during HCV infection. However, they are isolate specific and are usually effective only against HCV strains present before the appearance of the corresponding antibodies. Although antibodies to core and non-structural proteins are detectable in the serum, no specific antibody profile has been established as a predictor of outcome. The titre of antibodies to E1 and E2 proteins correlate with viraemia. The presence of strong and multispecific T-helper cell responses to HCV results in viral clearance. In patients with chronic HCV infection both the CD4+ T-helper cell and the cytotoxic T-lymphocyte responses are much weaker than during acute, resolving infection. Although HCV-specific cytotoxic T lymphocytes have been detected in peripheral and intrahepatic lympocytes, they seem functionally impaired as they are unable to clear the virus. HCV may escape immune elimination through peripheral tolerance, exhaustion of T-cell response by a high viral load, viral inhibition of antigen presentation, and viral mutations abrogating or antagonizing antigen recognition by virus-specific T cells. Further studies are needed to clarify these possibilities and to provide a scientific basis for new therapeutic concepts and the development of effective vaccine.

Hepatitis D virus (HDV)

HDV is a defective virus that causes acute and chronic liver disease only in association with hepatitis B virus. This unique pathogen was discovered in 1977 by M. Rizzetto in liver biopsies from patients with hepatitis B. HDV particles contain the viral RNA nucleocapsid, which is hepatitis †-antigen (**HDAg**) and an outer envelope (HBsAg) provided by the helper virus HBV. The HDV genome is a single-stranded, circular RNA (Table 1). It is the smallest known animal virus genome. Because of a high degree of internal complementarity, 70 per cent of the nucleotides are base-paired. This gives it an unusual, rod-like structure. HDV RNA replicates via RNA-directed RNA synthesis by transcription of genomic RNA to a complementary antigenomic †-RNA that serves as a template for subsequent genomic RNA synthesis. HDV produces a single protein, HDAg, which is encoded by the antigenomic RNA. RNA editing of the antigenomic RNA allows the virus to make two forms of HDAg—'small (S)' (195 amino acids) and 'large (L)' (214 amino acids). Both forms are present in the virions and have different functions in the HDV replicative cycle. HDAg-S facilitates HDV RNA replication, while HDAg-L inhibits replication and is required for assembly of the virion. Although the formation of †-virions requires the helper fuction of HBV, the replication of HDV RNA within the cell can occur without HBV.

Three distinct HDV genotypes have been recognized. Genotype I, the most widespread, has been identified in North America, Europe, Africa, and Asia. It is associated with a broad spectrum of chronic liver disease. Genotype II is found only in east Asia and seems to cause mild hepatitis-†. Genotype III is found exclusively in northern parts of South America and is associated with particularly severe hepatitis.

Host immune response and pathogenesis

HDV can infect either simultaneously with HBV (coinfection) or as a superinfection of a chronic carrier of HBsAg. Because HDV requires the helper function of HBV, the duration of †-infection is determined by the duration of HBsAg positivity. Like antibodies to HBV nucleocapsid (anti-HBc), antibodies to HDAg are not protective. Chronic HDV infection is accompanied by high titres of IgG anti-HD. High serum levels of IgM anti-HD indicate acute †-infection or exacerbation of chronic hepatitis D. The roles of cellular immune responses to HDAg, HBV antigens, or both in the immunopathogenesis of hepatitis D is uncertain. The lack of liver pathology in transgenic mice expressing HDV and data from experimental infections suggest that HDV is not cytopathic. This is supported by the experience with patients undergoing liver transplantation for HDV cirrhosis. Although HDV always recurs in the graft, necroinflammation is absent unless HBV recurs as well. The presence of microvesicular steatosis in severe hepatitis D indicates a possible direct cytopathic effect in some circumstances.

Hepatitis E virus (HEV)

HEV was first identified in 1983 by immune electron microscopy of the faeces of patients and is now recognized as the agent responsible for enterically transmitted non-A, non-B hepatitis. The virus is classified in the Caliciviridae family. Without a cell culture system, studies on HEV have required experimental transmission to susceptible non-human primates, such as cynomolgous macaques. The HEV genome is a single-stranded, polyadenylated RNA of approximately 7500 nucleotides containing three open reading frames (Table 1). ORF1 encodes non-structural proteins involved in virus replication—helicase and RNA-dependent RNA polymerase. ORF2, comprising approximately 2000 nucleotides, codes for the major structural proteins. ORF3 has 328 nucleotides and also appears to code for a structural protein. The genomic organization of HEV is different from HAV and HCV because the structural and non-structural proteins are coded by discontinuous, partially overlapping ORFs. Non-structural proteins are encoded at the 5' rather than at the 3' end of the genome. Unlike HAV, HEV infection may be zoonotic. HEV RNA has been found in the faeces of wild pigs. Serological evidence of infection has been found in pigs, cattle, and sheep in endemic regions. Sequence analyses have identified two major genotypes of HEV (isolates from Burma and Mexico), which show 25 per cent nucleotide variability. The amino acid variability ranges from 1 to 5 per cent among different HEV isolates from Asia to 14 per cent between the Mexican and Asian isolates. A new genotype was recently isolated from a patient in the United States.

The primary site of HEV replication is not fully understood. Following intravenous HEV inoculation in experimental models, the elevation of serum aminotransferases occurs after 24 to 38 days. Expression of HEV antigens has been detected in the cytoplasm of hepatocytes as early as 7 to 10 days after inoculation. Experimental data indicate that during an initial phase with high HEV replication, the virus may be released from hepatocytes into bile before serum 'liver' enzymes increase and there are morphological changes in the liver. Virus shedding ceases when serum aminotransferases return to normal. HEV RNA is detectable by reverse transcriptase polymerase chain reaction (**RT-PCR**) in the serum of virtually all patients within 2 weeks of the start of hepatitis. Prolonged viraemia (4 to 16 weeks) has also been reported. Detection of anti-HEV by enzyme immunoassay involving recombinant HEV antigens or synthetic peptides is the most frequently used method for diagnosis

and for epidemiological studies.

The humoral immune response develops gradually in parallel with the rise in serum alanine aminotransferase. The serum level of anti-HEV IgM peaks around the time of peak enzyme levels and is detectable for 5 to 6 months. Although the IgG anti-HEV response persists for several years after the acute hepatitis, the natural history of protective immunity to HEV is not fully established. In contrast to HAV, hepatitis E shows an unusually high attack rate among adults, suggesting that immunity to HEV, if acquired in childhood, may wane.

GB virus-C (GBV-C) or hepatitis G virus (HGV)

The genome of GBV-C was identified in 1995 by molecular hybridization in the serum of a patient with the initials GB. Separately, another group of investigators identified the genome of a new RNA virus, named hepatitis G virus. The comparison of HGV and GBV-C genomes revealed high homology, both at nucleotide (86 per cent) and amino acid level (100 per cent). In is now accepted that they represent two isolates of the same virus. GBV-C/HGV is an RNA virus with a single ORF encoding a polyprotein of approximately 3000 amino acids (Table 1). Together with another two RNA viruses, GBV-A and GBV-B, it belongs to the Flaviviridae family. These three viruses show various similarities with HCV. Specific features of the GBV-C/HGV genome include absence of core gene (nucleocapsid); long 5'- and 3'NTR and lack of poly(A) tail. Unlike HCV, this virus has a very conserved E2 region. Longitudinal studies have shown that GBV-C/HGV can establish chronic infection with RNA persistence in serum for up to 15 years. Some patients clear the virus spontaneously and develop anti-E2 reactivity, which is used as a marker of past infection. Anti-E2 also seems to confer protective immunity. A large body of evidence suggests that GBV-C/HGV does not cause liver disease.

TT virus (TTV)

TTV was identified in 1997 by investigators in Japan. By applying the methodology used for the identification of GBV-C, they detected the genome of a new DNA virus in the serum of a patient with cryptogenic post-transfusion hepatitis. The patient's initials (TT) prompted the name of this new virus and a causative role for acute and chronic hepatitis was suggested. Subsequent studies revealed that the TTV genome is circular, single-stranded DNA of approximately 3850 nucleotides (Table 1). Two partial ORFs have been predicted, but TTV proteins have not been expressed so far. It is suggested that TTV belongs to a new family—Circinoviridae. TTV DNA has been detected in non-human primates and farm animals. The primary site of TTV replication is unknown. TTV DNA is present in the liver and in all fractions of peripheral blood mononuclear cells, although TTV RNA transcripts are detectable only in liver tissue. Unlike other DNA viruses, TTV shows remarkable genomic variability. The phylogenetic analysis demonstrates the presence of many genotypes although there is no internationally agreed classification yet. The TTV population in one patient could comprise several genotypes.

TTV infection is highly prevalent worldwide (for instance up to 92 per cent of healthy subjects in Japan). Initially, the virus was thought to be transmitted parenterally, although its prevalence in the general population indicates the importance of non-parenteral routes as well. Prevalence increases with age in paediatric and adult age groups.

It is uncertain whether TTV is pathogenic. Analysis of liver histology in patients with TTV infection and longitudinal studies, as well as experimental TTV inoculation in chimpanzees, demonstrate that this virus does not cause hepatitis. So far, TTV is an example of a human virus with no clear disease association.

Further reading

Cerny A, Chisari FV (1999). Pathogenesis of chronic hepatitis C: immunological features of hepatic injury and viral persistence. *Hepatology* **30**, 595–601.

Hadziyannis SJ (1997). Epidemiology of G/GBV-C infection. In: Boyer JL, Ockner RK, eds. *Progress in liver diseases*, Vol XIV, pp 219–45. WB Saunders, Philadelphia.

Lau JYN, Wright TL (1993). Molecular virology and pathogenesis of hepatitis B. *Lancet* **342**, 1335–40.

Major ME, Feinstone SM (1997). The molecular virology of hepatitis C. *Hepatology* **25**, 1527–38.

Rizzetto M (1983). The delta agent. *Hepatology* **3**, 729–37.

Torre F, Naoumov NV (1998). Clinical implications of mutations in the hepatitis B virus genome. *European Journal of Clinical Investigation* **28**, 604–14.

Wilson RA, ed. (1997). *Viral hepatitis. diagnosis, treatment, prevention*. Marcel Dekker, New York.

Zuckerman AJ, Thomas HC, eds (1998). *Viral hepatitis. Scientific basis and clinical management*, 2nd edn. Churchill Livingstone, Edinburgh.

7.10.20 Hepatitis C virus

D. L. Thomas

Introduction

By the mid 1970s, it was apparent that both acute and chronic hepatitis could be caused by something other than hepatitis A virus (**HAV**) or hepatitis B virus (**HBV**). This condition, called non-A, non-B hepatitis, was assumed to be a viral infection since it was reproduced in chimpanzees inoculated with blood from affected persons, even after passage through 90 nm filters. However, hepatitis C virus (**HCV**) was not discovered until the late 1980s when a portion of viral RNA was cloned, and the resulting antigen shown to react with sera from persons with non-A, non-B hepatitis. It is now clear that HCV causes most cases of bloodborne non-A, non-B hepatitis.

Aetiology (see Chapter 7.10.19)

Epidemiology

Prevalence of infection

An estimated 170 million people are infected with HCV worldwide. In economically developed nations, HCV infection is found typically in 1 to 2 per cent of the general population. A 10-fold higher HCV prevalence has been found in Egypt and in some regions of Japan, Taiwan, and Italy. In these highly-endemic regions a sharp decrease in prevalence is often found in those less than 30 to 40 years of age, a cohort effect that probably reflects discontinuation of a practice that once contributed to widespread infection. HCV infection occurs in 50 to 90 per cent of persons injecting illicit drugs, more than 90 per cent of patients with haemophilia transfused with clotting factors before they were inactivated, 10 to 50 per cent of patients on haemodialysis, 5 to 20 per cent of patients attending sexually transmitted disease clinics, and 1 to 3 per cent of health care workers. HCV infection is common in people with other bloodborne infections, such as HBV and HIV.

Transmission

Biological basis

Studies with molecular clones demonstrate that infection will occur if sufficient numbers of complete HCV RNA transcripts reach the liver. HCV RNA has been detected in blood, saliva, seminal fluid, and tears, and intravenously injected blood is clearly infectious. In addition, in one instance, a chimpanzee was infected by intravenous injection of saliva. It is not known if other body fluids contain enough intact virions to be infectious when administered percutaneously or if infection can be sustained when virions contact cells present in mucous membranes.

Percutaneous transmission

Nosocomial transmission

The principal route of HCV transmission worldwide is percutaneous exposure to HCV-containing blood. Transfusion of contaminated blood once accounted for 20 per cent of HCV infection in the United States. HCV has also been transmitted by intravenous administration of contaminated immunoglobulin and clotting factors, including several well publicized outbreaks in the United States and Europe. However, the incidence of HCV transmission through administration of blood and blood components has decreased dramatically in regions of the world where donors are screened for HCV antibody and viral deactivation procedures are now used for immunoglobulin and clotting factor products.

Although the incidence in economically developed nations is now very low, patient-to-patient HCV transmission has been documented following percutaneous medical procedures such as colonoscopy with biopsy and use of intravenous infusion devices. In such instances, a common source of transmission can be detected by higher than expected identity in RNA sequences from various persons, as has repeatedly been shown in haemodialysis centres. Nosocomial HCV transmission probably requires a breach in infection control policies, although this may be difficult to recognize retrospectively.

In economically developing nations, most HCV transmission occurs through medical treatments, both by modern and folk practices. In Egypt, where 50 per cent of persons more than 40 years of age are infected, HCV was transmitted through a widespread national campaign of injections to eradicate schistosomiasis. When this practice was discontinued, there was a sharp decrease in HCV prevalence in persons born thereafter. Elsewhere in the world, HCV transmission occurs where education and resources are insufficient to ensure sterilization of devices used for medical injections, scarification rituals, and other percutaneous practices. Misperceptions regarding the benefit of injections appear to be especially important.

Drug use

In some regions of the world, percutaneous exposure to contaminated needles and other drug-use implements is the dominant mode of HCV transmission. HCV infection, which often occurs within months of starting to abuse drugs by injection, is found in 50 to 90 per cent of people admitting drug use worldwide. There are conflicting data as to whether HCV can be transmitted by intranasal use of cocaine.

Sexual transmission

Transmission of HCV by intercourse has not been proven, but some data suggest that it occurs, albeit uncommonly. In some HCV-infected individuals, the only potential exposure that can be detected is sex with another infected person, and HCV infection occurs more often than expected in persons with multiple sexual partners. In the families of HCV-infected people, sexual partners are the only members whose risk of infection is increased. The viral nucleotide sequences often suggest a common source. However, most long-term sexual partners of people with hepatitis C are not infected, and in those who are infected, it is impossible to exclude exposures other than intercourse. In the few studies in which direct comparisons are possible, the prevalences of HBV and HIV in sexual partners are 5- to 10-fold higher than for HCV.

Most authorities do not recommend that people in monogamous relationships use condoms to prevent HCV transmission. However, many encourage HCV-infected people to discuss the risk of transmission with their sexual partners and encourage them to be screened.

Mother-to-infant transmission

HCV infection occurs in 2 to 8 per cent of infants born to HIV-infected mothers. This risk increases if the mother is also HIV infected or if the maternal level of HCV RNA is high. Because of passive transfer of maternal antibodies, the diagnosis of HCV infection in the child must be based on detection of HCV RNA or persistence of antibodies 18 months or more after birth. There is no conclusive evidence that HCV is transmitted by breast feeding, and only a single study that suggests the risk of perinatal infection is reduced by elective caesarian delivery.

Natural history and pathogenesis

Viral persistence

HCV RNA can be detected in blood within weeks of exposure and, for approximately 85 per cent of individuals, remains detectable indefinitely (Fig. 1). Most persistently-infected people have intermittent elevations in serum liver enzymes such as alanine aminotransferase (**ALT**) and after 10 to 20 years, 2 to 20 per cent develop cirrhosis. Within 5 years, approximately 20 per cent of those who develop cirrhosis will have a life-threatening complication, such as ascites, variceal bleeding, hepatic encephalopathy, or hepatocellular carcinoma.

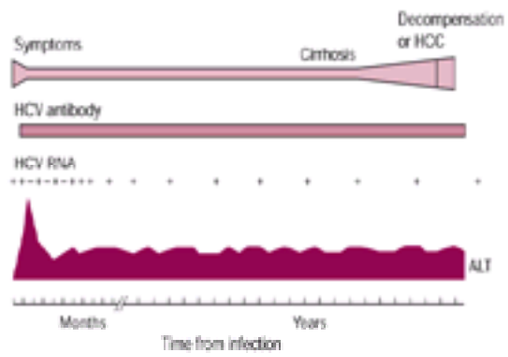

TABLE cellSpacing=0 cellPadding=0 align=left border=0 hspace="10" vspace="5">

Fig. 1 Natural history of HCV infection.

The incidence of cirrhosis is higher in persons infected at older ages and those who ingest alcohol, especially more than 50 g/day (or the equivalent of three alcoholic drinks). The effect of smaller amounts of alcohol is not known. HIV and HBV infections also appear to increase the incidence of HCV-related cirrhosis. Neither HCV genotype nor HCB RNA level are strong determinants of disease progression.

The pathogenesis of cirrhosis is poorly understood. A few HLA alleles have been associated with cirrhosis and more vigorous cytotoxic T-lymphocyte responses have been associated with severe liver disease. Conversely, immunosuppression from sources as diverse as HIV infection, agammaglobulinaemia, and steroid use appears to increase the incidence of cirrhosis.

Viral clearance

In approximately 15 per cent of people, HCV RNA can no longer be detected in blood one or more years after exposure, although HCV antibody and T-lymphocyte responses may remain. Long-term sequelae like cirrhosis and hepatocellular cancer do not appear to occur in those with viral clearance.

The mechanisms of viral clearance are not known. Both humoral and cellular immune responses are detectable to multiple HCV antigens within months of exposure, but occur even in those with persistent infection. Over time, individual variants may be eliminated from the HCV quasispecies only to be replaced by others that are sufficiently different to escape immune effectors. Those who clear HCV infection also tend to have strong T-cell proliferation responses to HCV antigens and a T_{H1} cytokine phenotype. The importance of these findings in viral clearance remains to be shown.

Clinical features

Acute infection

Acute HCV infection is indistinguishable from other forms of acute viral hepatitis, causing malaise, nausea, and right upper quadrant pain, followed by dark urine and jaundice. These symptoms occur in approximately 20 per cent of acutely infected adults, less frequently and typically with less severity than for hepatitis A or hepatitis B. Fulminant hepatic failure is rare.

HCV RNA is detectable before symptoms occur, but the level of viraemia varies in the first 6 months and can be transiently undetectable even in those who ultimately have persistent infection. Serum levels of liver enzymes such as ALT rise more than 10 times normal, then decline and, for those with persistent HCV infection, fluctuate indefinitely. The serum bilirubin may also be elevated for weeks after symptoms are first noted, but ultimately returns to a normal level. HCV antibody can usually be detected within a month of symptoms and within 8 weeks of exposure.

Chronic infection

The 85 per cent of people who develop persistent infection can be differentiated from those with viral clearance by repeated testing for HCV RNA in blood for 12 or more months. Other tests, such as serum ALT levels and the quantity of serum HCV RNA, do not reliably predict the outcome.

Fatigue and malaise may herald the onset of cirrhosis, which is suggested by thrombocytopenia, neutropenia, hypoprothrombinaemia, and hypoalbuminaemia. These haematological indicators of cirrhosis develop as late findings and imply a bad prognosis. Liver enzymes fluctuate throughout the course of HCV infection with little correlation to symptoms or the long-term outcome. Cirrhosis can occur even in the 20 to 40 per cent of patients who have repeatedly normal ALT levels. The levels of HCV RNA and HCV genotype likewise are poor predictors of disease. Probably the only reliable marker of disease progression is the liver biopsy (see below).

Extrahepatic manifestations

These include mixed cryoglobulinaemic vasculitis and membranoproliferative glomerulonephritis (see Chapter 20.7.8). Diagnosis of cryoglobulin-related vasculitis is based on the clinical syndrome as HCV-infected people commonly have cryoglobulins detectable in their serum. HCV infection is commonly associated with sporadic

porphyria cutanea tarda, and less commonly, with Sjögren's syndrome, lichen planus, idiopathic pulmonary fibrosis, and Mooren's corneal ulcers.

Pathology

The histopathological features of acute HCV infection are less severe than with the other hepatitis viruses. Mononuclear (mostly lymphocytic) inflammation is present throughout the lobule. Sinusoidal lining cells are activated and fat can be seen. Over time, the level of inflammation varies and fibrosis can occur, beginning in the portal areas and, in some cases, extending as septae beteween portal zones. Fibrous bands that bridge portal triads and formation of nodules denotes cirrhosis. The Knodell system quantifies the degree of periportal necrosis (0 to 10), intralobular necrosis (0 to 4), and portal inflammation (0 to 4) along with the stage of disease or fibrosis score (0 to 4). Although the histological findings fluctuate, this information remains the most important predictor of disease outcome and is often used to ascertain which individuals would benefit from treatment.

Liver cancer

Each year, an estimated 1 to 4 per cent of people with HCV-associated cirrhosis will develop hepatocellular cancer. The pathogenesis is unknown. The highest incidences are reported in Japan and Italy. In China and Korea, HBV infection is a more common cause of hepatocellular carcinoma. Serum a-fetoprotein levels and hepatic ultrasound are used for screening in persons with cirrhosis.

Diagnosis

Serological testing

HCV infection is usually diagnosed by testing for HCV antibodies in serum with an enzyme immunoassay that includes recombinant HCV proteins. Second and later generations of these antibody assays are highly sensitive screening tools (Table 1). Problems arise in acute infection as antibody development can be delayed for several months after exposure and in those with compromised antibody production (e.g. haemodialysis and agammaglobulinaemia). Uncommonly, false-negative enzyme results have been reported in persons on haemodialysis and, less commonly, HIV-positive people.

A positive HCV antibody test needs further evaluation. In low-risk screening (e.g. volunteer blood donation) an immunoblot assay can be used to detect antibodies to a variety of recombinant antigens. Reactions to more than one antigen strongly suggests infection. Where HCV infection is expected, HCV RNA testing is a more expedient confirmation approach, providing both an independent assessment of infection and indication of whether the infection has cleared or is ongoing.

HCV RNA testing

HCV RNA can be detected and quantified by a number of amplification techniques including reverse transcription polymerase chain reaction (**RT-PCR**). The reliability of HCV RNA assays has been questioned and the values of different quantitative tests are difficult to compare, although an international standard has been advanced. HCV genotype can be assessed by phylogenetic analysis of nucleotide sequences or detection of subtype-specific point mutations in RT-PCR amplified RNA.

Treatment

Interferon-a

Interferon-a induces expression of multiple genes that have antiviral and antiproliferative activity including those encoding RNAase L, 2'-5' oligo-adenylate synthase, M protein, and protein kinase R.

Almost half of the patients receiving interferon-a-2b (3 million units subcutaneously three times a week for 6 months) have a normal serum ALT and undetectable HCV RNA by the end of treatment (end of treatment response). However, many relapse and 6 months after completion of therapy, fewer than 20 per cent still have a normal ALT level and undetectable HCV RNA (sustained response). Longer treatment reduces the number of relapses, but overall sustained (6 months after treatment) response rates remain low. Higher interferon doses and daily administration accelerate the pace of viral clearance but do not consistently improve sustained virological response rates.

Interferon-a therapy causes a number of adverse reactions. Flu-like symptoms occur within 6 h of the first dose but generally diminish in 1 to 2 weeks. Fatigue, depression, and other mood disturbances may be severe, especially if there is a history of such problems in the past. Hair thinning and thyroid abnormalities may occur. Bone marrow suppression is common including neutropenia, thrombocytopenia, and anaemia. Interferon-a cannot be used safely in pregnancy.

Interferon-a therapy has been associated with improvements in quality-of-life indices and reductions in the incidence of hepatocellular cancer and cirrhosis. Although uncommon, sustained virological responses are durable; 5 years later more than 90 per cent of sustained responders still have normal serum ALT levels and no HCV RNA in blood or liver.

Other interferon formulations, including a recombinant consensus interferon, interferon-a-2a, and interferon-a-2b-n $_1$ (lymphoblastoid interferon) have similar efficacy and adverse effects. Interferon-a has been covalently linked to polyethylene glycol (pegylated interferon), resulting in a longer half-life (weekly dosing), higher sustained serum levels, and improved HCV clearance.

Interferon and ribavirin

Ribavirin is a guanosine analogue that has broad antiviral activity but may affect HCV by inducing a shift toward a T $_{H1}$ immune response. Used orally alone, ribavirin returns the level of serum ALT to normal in some individuals but does not substantially change HCV RNA levels. However, in combination with interferon-a-2b, 1000 to 1200 mg of oral ribavirin daily improves the sustained virological response rates both for people who have never been treated and those who initially responded to interferon but then relapsed (but not those who never responded). As initial treatment, approximately one-third of patients treated with interferon and ribavirin have a sustained virological response. Responses to interferon-a and ribavirin vary according to pretreatment characteristics, especially HCV genotype (Fig. 2). It is likely that pegylated interferon-a and ribavirin will be the most effective therapy available in 2001.

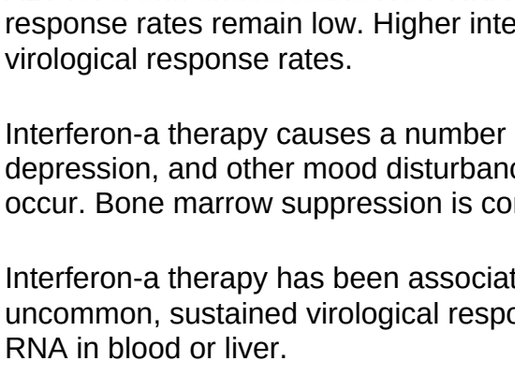

Fig. 2 Sustained virological response rates according to regimen and pretreatment factors.

Adverse reactions to ribavirin and interferon-a are similar to those with interferon-a alone, but ribavirin causes haemolytic anaemia in many patients. Ribavirin is

teratogenic; pregnancy must be prevented during and for up to 1 year after administration, whichever sex is being treated.

Other therapies

Oral amantadine and herbal products such as milk thistle have been used to treat HCV infection. New drugs are expected that interfere with the viral protease, helicase, or replicase, or with viral translation.

Prevention

Primary prevention

HCV transmission is preventable by reducing percutaneous exposures, while ensuring the safety of those that are medically or culturally necessary. HCV-infected people should not allow others to come into contact with their blood, especially by sharing razors or dental devices. HCV is not transmitted by typical household exposures (hugging, kissing, sharing eating utensils or food). Counselling may be needed to prevent unwarranted ostracism.

No vaccination has been licensed to prevent HCV transmission. Since HCV reinfection has been demonstrated even with an autologous inoculum, it is difficult to induce immunity that protects against infection. However, it may be possible to reduce viral persistence by vaccination.

Secondary prevention

Once infection occurs, the incidence of cirrhosis and hepatocellular cancer can be reduced by medical treatment and elimination (or reduction) of alcohol ingestion. Because of drug toxicity and expense, interferon-a use is rare except among selected residents of economically developed nations. More accessible treatments are needed to prevent development of disease worldwide.

Postexposure prevention

Although administration of HCV antibody-containing immunoglobulin may increase the time to development of infection, infection is not usually prevented, and immunoglobulin preparations available in many countries no longer contain HCV antibodies. Therefore, most authorities recommend that persons exposed to HCV do not receive immunoglobulin. Exposed persons should be monitored (for example, at 2 and 6 months) for development of HCV antibodies and possibly HCV RNA, since treatment may be more effective if provided within the first year of infection. There are no interventions available to prevent HCV transmission from a mother to her infant; in particular, current medical treatments (interferon and ribavirin) are contraindicated in pregnancy, immunoglobulin administration is not advised, caesarian section is not routinely indicated, and breast feeding should not be discouraged.

Further reading

Alter MJ *et al*. (1998). Hepatitis C. *Infectious Disease Clinics of North America* **12**, 13–26. [A review of the epidemiology of hepatitis C infection from an international expert.]

Bukh J, Miller RH Purcell RH (1995) Genetic heterogeneity of hepatitis C virus: quasispecies and genotypes. *Seminars in Liver Disease* **15**, 41–63. [A comprehensive evalution of the genetic complexity of hepatitis C by the pioneers in the field.]

Centers for Disease Control and Prevention (1998). Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. *Morbidity and Mortality Weekly Report* **47** (No. RR-19), 1–39. [A thorough review of the epidemiology and management of hepatitis C infection.]

Chang K-M, Rehermann B, Chisari HV (1997). Immunopathology of hepatitis C. *Springer Seminars in Immunopathology* **19**, 57–68. [A review of the immunology of hepatitis C infection by a group that has contributed substantially to the field.]

Davis GL, Nelson DR, Royes GR (1999). Future options for the management of hepatitis C. *Seminars in Liver Disease* **19**(Suppl. 1), 103–12. [A review of the manamgent of hepatitis C infection by leading experts.]

Pawlotsky JM *et al*. (1998). What strategy should be used for diagnosis of hepatitis C virus infection in clinical laboratories? *Hepatology* **27**, 1700–2. [A review of the approach to diagnosis of hepatitis C infection by an international expert.]

Seef LB (1997). Natural history of hepatitis C. *Hepatology* **26**, 21S–28S. [A review of the natural history of hepatitis C by an international expert.]

Thomas DL, Lemon SM (2000). Hepatitis C. In: Mandell GL, Bennett JE, Dolin R, eds. *Principles and practices of infectious diseases*, 5th edn, pp 1736–60. [A well-referenced review of hepatitis C.]

7.10.21 HIV and AIDS

G. A. Luzzi, T. E. A. Peto, R. A. Weiss, and C. P. Conlon

Introduction

The acquired immunodeficiency syndrome (**AIDS**) was first recognized in 1981 in the United States, when several cases of *Pneumocystis carinii* pneumonia and Kaposi's sarcoma were reported in homosexual men in New York and California. The variety of unusual infections and other conditions declared a new form of cellular immunodeficiency. Soon after, the syndrome was reported in injecting drug users, haemophiliacs, and recipients of blood transfusions. Early epidemiological data suggested that the cause was a sexually transmissible bloodborne infective agent. During 1983, in France, a new retrovirus was isolated from a patient with persistent generalized lymphadenopathy. Initially referred to as 'lymphadenopathy-associated virus' (**LAV**) or 'human T-lymphotropic virus III' (**HTLV-III**), it was renamed 'human immunodeficiency virus' (**HIV**) in 1986.

At the time of its discovery, HIV was already widespread, the earliest infections probably having occurred before the 1950s. The recognition of heterosexual intercourse as the most common means of HIV transmission worldwide followed the investigation of epidemics in Africa and the Caribbean. Infected mothers could pass the virus on to their fetus or neonate, establishing vertical transmission as another important route of HIV infection.

In 1986 a second retrovirus causing AIDS, HIV-2, was identified in West Africa. It is largely confined to this region, while HIV-1 is the cause of the world pandemic of AIDS. Over the past 5 years there have been advances in the understanding of the pathogenesis of HIV, in clinical monitoring, and in therapy. Table 1 lists the milestones in the history of HIV and AIDS (acquired immunodeficiency syndrome).

Epidemiology

The global HIV-1 pandemic has affected developing countries in particular. Despite under-reporting, the World Health Organization (**WHO**) estimated that by the end of 1998 over 10 million people had died of HIV, and over 30 million people were alive and infected worldwide, of whom 90 per cent were living in sub-Saharan Africa, South and South-East Asia, and Central and South America (Fig. 1).

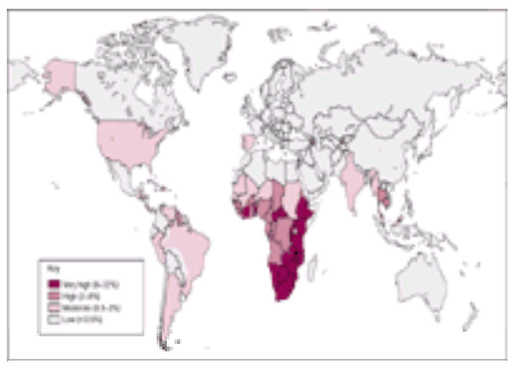

Fig. 1 World distribution of HIV (UNAIDS/WHO, 1998). (Reproduced from *Report on the global HIV/AIDS epidemic*, June 1998. (1998). UNAIDS/WHO, with permission.)

The numbers of people infected with HIV must be distinguished from cases of AIDS, which follows an asymptomatic period of about 10 years and may be influenced by interventions such as antiretroviral therapy. Worldwide, the WHO estimated a 9 per cent increase in new infections in 1997 compared with 1996.

In North America, western Europe, and Australasia the epidemic began in the late 1970s and early 1980s among homosexual men and injecting drug users. However, in these regions the proportion attributable to heterosexual transmission has increased. The estimated incidence of AIDS in western Europe rose every year between 1985 and 1994, stabilized in 1995, and fell by 10 per cent in 1996 and by over 20 per cent in 1997. A similar trend has been observed in North America. Cases attributed to injecting drug use form the largest proportion of diagnosed cases of AIDS in Europe. Large epidemics of HIV have been reported in injecting drug users in several countries of the former Soviet Union.

Some two-thirds of all cases are found in sub-Saharan Africa, where HIV transmission is predominantly heterosexual and perinatal. The estimated overall prevalence there is 7 to 8 per cent, rising to 20 to 30 per cent in some countries such as Zambia and Zimbabwe, where AIDS has curtailed population growth. Because of the predominant heterosexual transmission, the overall male-to-female ratio in Africa is approximately 1:1 compared with 9:1 in North America and western Europe.

In Africa, predicted rates of AIDS and new HIV infection were expected to plateau by 2000 and then to fall gradually, whereas trends suggest a continuing rise in South and South-East Asia, where the emergence of epidemic HIV occurred later. A rapid rise in incidence occurred in Thailand and India in the late 1980s, initially among intravenous drug users and prostitutes and then through heterosexual spread; the WHO estimates that 3 to 5 million people have been infected in India alone. Rapid spread and major epidemics of HIV have also been reported in China, Cambodia, Burma (Myanmar), and Vietnam.

High rates of transmission of HIV continue in developing countries because of the lack of awareness, poverty, high rates of other sexually transmitted infections, and higher risk behaviour such as the use of prostitutes and injecting drug use.

HIV-2 is endemic in parts of West Africa and is increasingly prevalent in Angola, Mozambique, France, and Portugal. In other parts of the world the prevalence is very low, although it is present in India. The clinical features of HIV-2 are similar to those of HIV-1, but some patients with HIV-2, for unknown reasons, appear to progress much more slowly.

HIVs may be regarded as zoonoses: HIV-1 is derived from a simian immunodeficiency virus in the chimpanzee (*Pan troglodytes troglodytes*), and the animal reservoir for HIV-2 is the sooty mangabey monkey (*Cercocebus atys*). Variation of HIV-1 RNA sequences has been identified, leading to a classification of 11 sequence subtypes (or clades), A to K, of the main group M, and N (new) and O (outlier) as two quite distinct groups in west central Africa. The subtypes have varying geographical distributions. For instance, subtypes A and D are found in central Africa, B in North America and Europe, and E in Thailand. Study of the genetic and geographical divergence of subtypes has shed light on the emergence and global spread of HIV.

Cellular biology of HIV

The viral replication cycle

HIV-1 (Fig. 2) and HIV-2 belong to the lentivirus subfamily of retroviruses. Retrovirus implies a 'backwards' step in biological information during viral replication attributable to its enzyme, reverse transcriptase. As with all retroviruses, the viral genes in infectious particles are carried as RNA, but upon infection of the host cell, reverse transcriptase catalyses the synthesis of a double-stranded DNA viral genome (Fig. 3). Insertion of the DNA genome into the chromosomal DNA of the infected cell is effected by viral integrase. The integrated provirus may remain latent, particularly in resting lymphocytes. In actively infected cells, however, RNA transcripts and proteins are synthesized, leading to the formation of new virus particles.

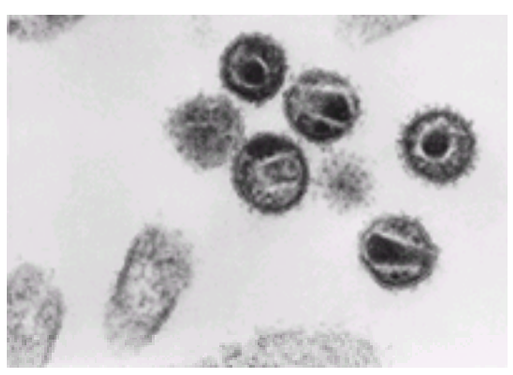

Fig. 2 Electron micrograph of HIV-1. (Reproduced by courtesy of H. Gelderblom.)

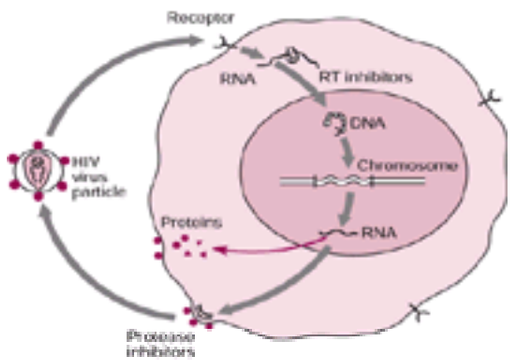

Fig. 3 Replicative cycle of HIV.

The core proteins derived from the *gag* and *pol* genes are made as large polypeptides that are then cleaved into smaller components representing the enzymes and building blocks of the virus. This cleavage is achieved by the viral protease. The unique reverse transcriptase and protease are targets of antiretroviral therapy (see Antiretroviral therapy below). Reverse transcriptase inhibitors such as zidovudine and lamivudine affect an early step in HIV replication, whereas the protease inhibitors, such as saquinavir or indinavir, block a late stage of virus assembly (Fig. 3). Compounds that inhibit any stage of HIV replication, without being too toxic to the infected person, are potential antiviral drugs. Agents have been developed to block viral entry (fusion inhibitors); in future, the integrase and viral RNA processing may become therapeutic targets.

HIV genes and proteins

Although regarded as a complex retrovirus, HIV has only nine genes (Fig. 4). The three structural genes are *gag*, *pol*, and *env*, encoding the core proteins p19, p24, and p17, the enzymes (protease, reverse transcriptase, and integrase), and the envelope glycoproteins (gp120 and gp41), respectively. The major regulatory genes *tat* and *rev* encode proteins that are not assembled into the virus but are essential for replication in the cell. The Tat protein acts in positive feedback to enhance transcription of viral RNA from the DNA provirus, while the Rev protein helps the efficient transport of viral RNA from the nucleus to the cytoplasm. Either of these proteins could be a suitable target for antiviral therapy, particularly Tat, because the synthesis of all the other viral proteins depends on its activity.

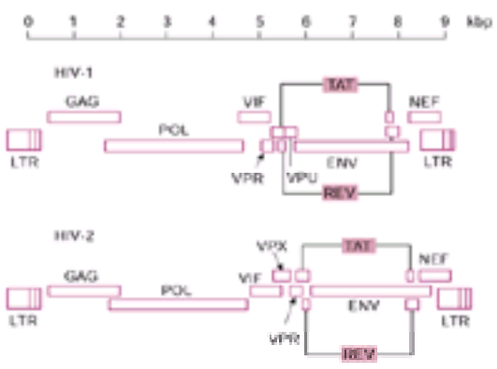

Fig. 4 HIV genome map.

The functions of the four accessory genes of HIV are less well understood. *Vif* encodes a protein assembled in virus particles that appears necessary for the infectivity at a stage soon after entry, possibly by facilitating disassembly of the virion to allow reverse transcription. Nef also effects an early postentry function; it is not needed by laboratory-adapted HIV strains or if virus enters via endosomal vesicles rather than fusing with the outer cell membrane. It also downregulates surface expression of the primary cell-surface receptor for HIV, the CD4 antigen, by drawing CD4 into clathrin-coated pits. Vpu similarly interacts with CD4, promoting its degradation by directing it to the ubiquitin–proteasome pathway. Vpr has dual functions; first, it directs the preintegration complex of the virus, containing the newly synthesized DNA, into the nucleus so that it can integrate into chromosomal DNA; second, it blocks cell proliferation in the G2 phase of the cell cycle, thereby enhancing the amount of viral progeny released per cell.

Unlike HIV-1, HIV-2 and the simian immunodeficiency virus (**SIV**) lack *vpu*, but have an alternative gene, *vpx*. HIV-2 Vpr leads the viral genome into the cell nucleus but does not arrest the cell cycle. These proteins presumably recognize cellular proteins and some of these interactions are species-specific. Thus the Vpr and Vif proteins in SIV of African green monkeys do not function in human cells, while the equivalent proteins of SIV from sooty mangabey monkeys work well in human cells. This could explain why sooty mangabey SIV was able to infect humans and become HIV-2, whereas the more widespread African green monkey SIV has not led to a zoonosis. Another difference is that HIV-1 incorporates the cellular protein cyclophylin A (the target of the drug ciclosporin A) into virus particles, where it may co-operate with Vif and is required for steps early in the infection. In contrast, HIV-2 does not contain cyclophylin A and replicates well without it.

HIV receptors and cellular tropism

CD4 is the cell-surface receptor for HIV; it is expressed on T-helper lymphocytes, the cells that become depleted in AIDS. CD4 is also expressed (to a lesser extent but sufficient to permit infection) on macrophages, Langerhans dendritic cells in mucous membranes, and brain microglial cells. These are the other target cells for HIV infection. CD4 is necessary to initiate HIV infection but is not sufficient to allow the virus to fuse with host-cell membranes: another cellular component or co-receptor is required.

Different substrains of HIV, even those isolated from the same patient, exhibit specific tropisms for different cell types in culture. All isolates can infect primary CD4 lymphocytes, but only some infect macrophages while others can infect cell lines established from CD4+ leukaemic cells. Macrophage-tropic strains predominate early in the course of HIV infection, and may be more transmissible from person to person. They do not cause CD4 lymphocytes to fuse together in culture and hence are referred to as non-syncytium inducing (**NSI**) strains. In contrast, many HIV isolates established from late-stage infection rapidly adapt in culture to infect T-cell lines and are syncytium-inducing (**SI**). Approximately 50 per cent of patients with AIDS develop SI strains in addition to NSI strains. The differences in cellular tropism and SI/NSI phenotype occur in all HIV subtypes or clades, which appear to reflect geographical variation of HIV rather than specific biological properties of the virus.

The complex cellular tropism of HIV has been explained by the discovery that different members of the chemokine receptor family act as co-receptors to CD4 for HIV entry into cells. Chemokines are chemoattractant, locally acting hormones or cytokines that bind to one or more receptors which are structurally related to olfactory and neurotransmitter receptors. Following binding to the CD4 receptor, primary NSI strains use CCR5, the chemokine receptor for macrophage-inhibitory proteins (**MIP-1a**, **MIP-1b**) and RANTES. In contrast, the SI strains of HIV use the CXCR4 co-receptor, the receptor for another chemokine, stromal-derived factor-1 (**SDF-1**). Other receptors such as CCR3 (the receptor of eotaxin) can be used by some NSI strains.

High levels of MIP-1a or -b in the blood correlate with relative resistance to HIV infection. Some exposed yet uninfected individuals are homozygous for an inherited defect of the CCR5 receptor involving a 32 base-pair deletion in the *CCR5* gene. This mutation has a high frequency in Caucasian people but is not found in African and Asian populations. Individuals who are homozygous for the deletion are healthy, indicating that the CCR5 receptor is not essential for the development of immune competence, probably because MIP-1 and RANTES can also bind to alternative receptors. However, homozygotes are genetically resistant to infection by NSI strains of HIV and the few homozygotes with A32 deletions who are HIV-positive appear to have been infected with SI strains that utilize CXCR4 instead. Other, more subtle, mutations in the promoter region of the *CCR5* gene allowing only low levels of co-receptor expression may confer relative resistance to HIV infection and also, if infection occurs, slower progression to AIDS.

The outer envelope glycoprotein, gp120, is the molecule on HIV that binds to CD4 and subsequently to the co-receptor. Gp120 is anchored to the viral envelope via gp41, the viral protein that is thought to effect membrane fusion. The gp120–gp41 is present in the viral envelope as a trimeric complex. SI strains have a gp120–gp41 structure that is less stable than NSI strains, readily undergoing conformational change on binding to CD4. This property makes SI strains more sensitive to neutralization by gp120 antibodies and also to inactivation by soluble forms of recombinant CD4, which were once seen as promising therapeutic agents. NSI strains, however, are more resistant. Mutations in the V3 loop of gp120 can convert NSI strains to SI strains. These mutations arise naturally during progression to AIDS and may allow HIV to switch to infect different cell types via new co-receptors.

The natural chemokines act as competitive inhibitors of HIV entry; certain chemically modified chemokines and chemical analogues act as strong HIV inhibitors without triggering the downstream signalling of the receptor. This has led to a new class of potential anti-HIV drugs, called 'co-receptor inhibitors'.

Diagnosis of HIV infection

Acute infection is accompanied by the development of serum antibodies to the core and surface proteins of the virus, usually within 2 to 6 weeks. Most seroconversions occur within 3 months of infection, and very rarely up to 6 months. Routine diagnostic tests, if negative, should be repeated 3 months after any possible exposure. Where there has been a high risk of transmission, additional tests that detect HIV directly (detection of viral RNA or DNA by polymerase chain

reaction, **PCR**) should be used, and may confirm HIV infection before antibodies become detectable.

Following seroconversion, antibody to envelope protein persists indefinitely in the serum and forms a highly specific test for HIV infection. In general, one or more sensitive enzyme immunoassay tests that detect HIV-1 and HIV-2 antibodies are used as the initial screening tests. Positive screening tests are confirmed by additional tests to confirm the presence of HIV antibodies.

Pretest discussion and counselling

Where possible, patients should understand the implications of being tested for HIV and should give informed consent before the test is done. This is especially important for asymptomatic people. Awareness of being HIV-positive allows the use of effective prophylaxis against the major opportunistic infections, and highly active antiretroviral drugs. It should also encourage behavioural change to reduce the risk of transmission to sexual partners, and may benefit children exposed to perinatal infection. However, early diagnosis may cause distress and disruption of domestic, social, and professional lives, although the infected person may be free from symptoms for many years. HIV-positive people may find it difficult to obtain life or medical insurance, obtain work, buy a house, and travel abroad.

Where HIV is relevant to the investigation of a patient's symptoms, it is in their interest to be tested so that appropriate treatment for an opportunistic condition, antiretroviral therapy, and prophylaxis can be provided. Where the patient is too ill to give consent, testing may be justifiable on these grounds. A high level of confidentiality must be maintained; disclosure of HIV-positive status should generally be allowed only in the medical interests of the patient and with their knowledge and consent.

Clinical presentation and features

Acute HIV syndrome

Between 2 and 6 weeks after exposure to HIV, 50 to 70 per cent of those infected develop a transient, often mild, non-specific illness (sometimes called primary infection or seroconversion illness) similar to infectious mononucleosis, with fever, malaise, myalgia, lymphadenopathy, and pharyngitis. However, unlike infectious mononucleosis over 50 per cent of people develop a rash, typically erythematous, maculopapular, and affecting the face and trunk. Other rashes and patterns of distribution, and oral and genital ulcers have also been reported. The illness begins abruptly and usually lasts for 1 to 2 weeks, but may be more protracted. Neurological complications include acute encephalitis, lymphocytic meningitis, and peripheral neuropathy. Severe or long-lasting illness and neurological involvement are associated with accelerated progression to AIDS and a bad prognosis, which may be influenced by early antiretroviral therapy.

Diagnosis requires a high index of suspicion. Acute HIV infection is a time of high viraemia (typically 10^5 to 10^6 viral particles/ml) during which antibodies to HIV may initially be absent (Fig. 5). Serological tests often need to be repeated at intervals to establish the diagnosis. Rapid diagnosis during the early stages of acute infection may be provided by detecting HIV viraemia using tests for HIV RNA or proviral cDNA (by PCR). A transient decrease in CD4 lymphocytes is usual during primary illness. Occasionally this may be substantial and associated with opportunistic infections such as oral or oesophageal candidiasis, and rarely pneumocystis pneumonia.

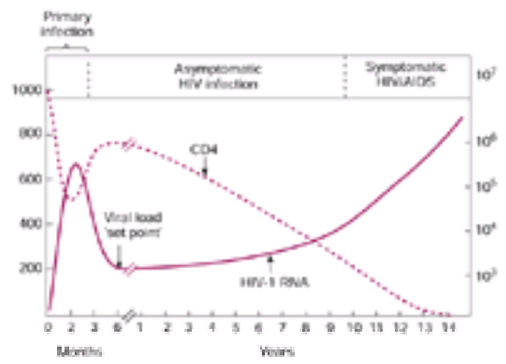

Fig. 5 Schematic representation of typical changes in CD4 lymphocyte count (left axis, per mm^3) and plasma HIV-1 RNA (right axis, copies/ml) with time, during the natural history of HIV infection.

Aggressive therapy of acute HIV infection with antiretroviral drugs does not eradicate the infection but, on theoretical grounds, may alter the natural history. After acute infection, the viral load becomes relatively stable after 6 to 9 months (Fig. 5). The plasma HIV RNA level at this virological steady state or 'set point' is of prognostic importance; therefore, treatment of the initial viraemic illness may lower the risk of progression. A placebo-controlled trial of zidovudine monotherapy during acute HIV infection showed a short-term benefit, but whether long-term outcomes are better compared with deferred treatment is not known. There are also concerns about the long-term toxicity of antiretroviral drugs. Current guidelines generally recommend considering treatment with highly active antiretroviral therapy, ideally within a clinical trial. The optimal duration of therapy for acute HIV infection is unknown.

Early HIV infection

Following the acute syndrome or subclinical seroconversion, there usually follows an asymptomatic period lasting an average of 10 years without antiretroviral therapy. Although a time of clinical latency, there is intense viral turnover: 10^9 to 10^{10} viral particles are replaced daily and the half-life of circulating CD4 lymphocytes is substantially reduced.

During the asymptomatic period, physical examination may be normal, but about one-third of patients have persistent generalized lymphadenopathy. The enlarged nodes, caused by a non-specific follicular hyperplasia, are usually symmetrical, mobile, and non-tender. The cervical and axillary nodes are most commonly affected. Nodes that are markedly asymmetrical, painful, or rapidly enlarging should be biopsied to exclude tumours such as lymphoma and opportunistic infections such as tuberculosis.

Symptoms of progressive HIV infection can be prevented by highly active antiretroviral treatment (see Management of HIV and prevention of complications, below). In the absence of treatment, patients often develop minor opportunistic conditions affecting the skin and mucous membranes. These are also common throughout the later stages of HIV disease. They include a range of infections: fungal (e.g. tinea, *Pityrosporum*), viral (e.g. warts, molluscum contagiosum, herpes simplex, herpes zoster), and bacterial (e.g. folliculitis, impetigo); and also eczema, seborrhoeic dermatitis, and psoriasis.

Drug rashes may occur at all stages of HIV, and particularly in late disease. Reactions to co-trimoxazole occur in up to 30 per cent of patients. They are most common when high doses are used in the treatment of pneumocystis pneumonia. Dapsone, clindamycin, b-lactam antibiotics, pentamidine, and nevirapine are commonly associated with drug rash.

Oral hairy leucoplakia usually appears as corrugated greyish-white lesions on the lateral borders of the tongue in homosexual men (Fig. 6). The condition is symptomless and non-progressive, but acts as a useful clue to HIV seropositivity. Epstein–Barr virus DNA has been demonstrated in these lesions.

Fig. 6 Oral hairy leucoplakia.

One of the characteristic clinical presentations of HIV disease is a sore mouth and throat due to oropharyngeal candidiasis (oral thrush) (Fig. 7). This sign of worsening immunodeficiency may be recurrent. Topical antifungals (amphotericin lozenges or nystatin suspension) are usually effective in the early stages, but later oral azole antifungals (ketoconazole, fluconazole, or itraconazole) are needed. *Candida albicans* is usually responsible, but other species (e.g. *C. glabrata*) may be implicated.

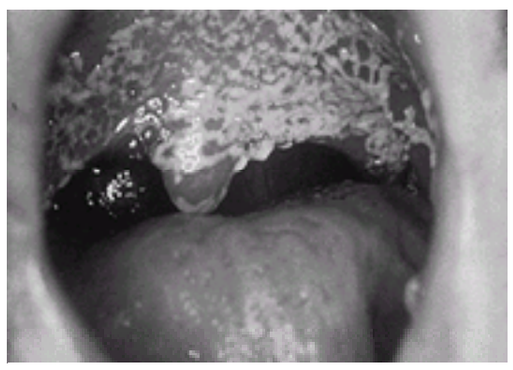

Fig. 7 Oral candidiasis.

There is an increased incidence of periodontal disease in those with HIV. Necrotizing (ulcerative) gingivitis and periodontitis may require extensive debridement and antimicrobials. Recurrent oropharyngeal aphthous ulceration is common and may be painful. Recurrent ulcers may occur in the oesophagus and other parts of the gastrointestinal tract. They usually respond to local or systemic corticosteroid therapy. Resistant cases may respond to thalidomide.

Later in the course of infection, intermittent or persistent non-specific constitutional symptoms may develop, which include lethargy, anorexia, diarrhoea, weight loss, fever, and night sweats. These symptoms may presage severe opportunistic infections or tumours.

Progression to AIDS

Various staging systems for HIV infection and case definitions of AIDS have been used since 1982 and modified by increased understanding of the pathogenesis and natural history. The 1987 Centers for Disease Control (**CDC**) definition listed a range of specific diseases indicative of AIDS. In 1993, an expanded definition was introduced in the United States that included additional AIDS indicator diseases, people with proven HIV infection, and a CD4 lymphocyte count of less than 200/mm ^{3} (0.2×10^{9}/l), irrespective of clinical manifestations. This last criterion has not been adopted in Europe.

The value of making a distinction between AIDS (as defined) and HIV infection at other stages is questionable, especially in industrialized countries. AIDS-defining illnesses were essential for surveillance when HIV status was frequently unknown, the natural history of HIV infection was poorly understood (the proportion developing opportunistic complications was uncertain), and disease-modifying drugs were not available. However, effective prevention of many of the opportunistic infections has led to an increase in the proportion of symptomatic patients who do not fulfil the criteria for AIDS. Highly active antiretroviral therapy often improves the clinical condition and survival even when started after progression to AIDS. These factors have undermined the epidemiological value and prognostic importance of a strict AIDS case definition. It is probably more useful to consider progressive HIV disease as a continuous spectrum.

However, clinical criteria to identify symptomatic HIV disease and AIDS are needed in developing countries, where laboratory confirmation of HIV seropositivity and AIDS-defining diseases is not possible. The WHO has, therefore, adopted clinical case definitions for AIDS surveillance in resource-poor countries, based on clinical manifestations with or without laboratory confirmation of HIV infection.

Non-progression

While the average time between infection with HIV and the development of AIDS is about 10 years, approximately 20 per cent of patients progress rapidly to AIDS within 5 years and 10 to 15 per cent remain clinically well for 15 to 20 years. Long-term healthy survivors are often called non-progressors, and to an extent this subgroup represents simply the tail end of a normal distribution of progression rates. Cohort studies have demonstrated that most apparent non-progressors are slow progressors, in whom a gradual decline in the CD4 lymphocyte count and increments in HIV viral load can be demonstrated. Although several investigators have reported virological, genetic, and cellular and humoral immunological factors that may be associated with non-progression, limitations in study design have made it difficult to identify what was responsible. A mutation in the gene for the macrophage chemokine receptor CCR5 is associated with non-progression in the heterozygous state; homozygotes have high-level resistance to HIV infection (see Cellular biology, above).

Management of HIV and prevention of complications

Impact of highly active antiretroviral therapy

Although a decline in the number of cases of AIDS and mortality from HIV was reported from the United States and Europe before the advent of protease inhibitors in 1996, the subsequent marked reductions in morbidity and mortality are mostly attributable to antiretroviral regimens that include the newer potent agents (protease inhibitors or non-nucleoside reverse transcriptase inhibitors) in combination with nucleoside drugs. Among 1255 HIV-positive patients attending HIV clinics in eight cities in the United States, mortality declined from 29.4 per 100 person-years in 1995 to 8.8 per 100 person-years in the second quarter of 1997. The incidence of pneumocystis pneumonia, disseminated *M. avium* complex (**MAC**) infection, and cytomegalovirus (**CMV**) retinitis declined dramatically. The mortality of patients with CD4 counts below 100/mm^{3} fell for the first time in 1996, at a time when protease inhibitors were increasingly being included in treatment regimens. A decline in the incidence of opportunistic infections, notably oral candidiasis, toxoplasmosis, cryptosporidiosis, and cryptococcal meningitis, was reported from the United States and Europe.

In Europe, the expected survival 10 to 15 years after seroconversion was shown to have risen substantially after the introduction of highly active antiretroviral treatment (Fig. 8). For instance, in the 35 to 44 years' age group, survival 10 years after acquiring HIV in the era of highly active antiretroviral treatment (1997–98) was estimated to be 83 per cent, compared with 43 per cent for those infected between 1986 and 1996.

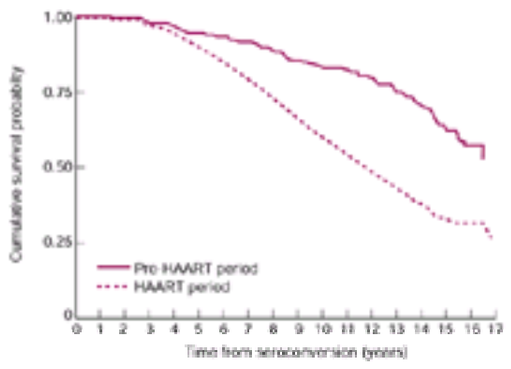

Fig. 8 Estimated proportions of individuals surviving from HIV-1 seroconversion in 1986–96 (pre-HAART* period) and 1997-98 (HAART* period). *HAART, highly active antiretroviral therapy. (CASCADE collaboration, *Lancet* (2000), **355**, 1158.)

Whether antiretroviral drugs will ever eradicate HIV and bring about a 'cure' is regarded as unlikely. Although HIV may be undetectable in plasma for many months, a long-lived reservoir of infectious virus can be recovered from latently infected (resting) memory CD4 lymphocytes. Since the half-life of this cell population is about 6 months, many years of effective antiretroviral treatment would be needed to clear virus from this reservoir. Other compartments exist that are relatively inaccessible to drugs—for instance, in the central nervous system, retina and testes—and unless viral replication can be successfully prevented at such sites there is also the risk of reinfection of compartments previously cleared by therapy.

General management

Ideally, HIV infection should be identified at the asymptomatic stage. Clinical and laboratory monitoring can detect waning immunity and the risk of disease progression, prompting antiretroviral therapy and prophylaxis against infections such as pneumocystis pneumonia. Serological screening detects past or current infections such as toxoplasma, CMV, hepatitis B and C, and syphilis, which may be reactivated or progress during immunosuppression. Clinic visits provide an opportunity for discussion of such issues as safer sex. Many problems can be managed by a primary care physician. Routine dental care is needed. Clinical and laboratory monitoring, the management of late complications, and the prescription and monitoring of antiretroviral drugs require specialist supervision.

Monitoring

Monitoring involves regular clinical assessment and prognostic laboratory tests. Oral candidiasis, or physical signs such as asymptomatic cutaneous Kaposi's sarcoma are of prognostic importance. CD4 lymphocyte count and quantitative estimation of HIV RNA in the blood plasma (viral load) are the two laboratory markers that have the best prognostic value.

The CD4 lymphocyte (T-helper cell) count is a reliable indicator of HIV-related immune impairment. CD4 counts, normal at or above 600/mm ^{3}, vary considerably, even in the absence of HIV infection. A fall in the CD4 lymphocyte count to below 200/mm ^{3} is associated with a risk of opportunistic infections of about 80 per cent over 3 years without antiretroviral treatment. However, progression is variable and a minority remain well for several years with stable low CD4 counts. This variability is explained partly by differences in HIV viral load. The level of CD4 lymphopenia generally determines the spectrum potential of infections (Table 2). For instance, whereas oral and oesophageal candidiasis and pneumocystis pneumonia are frequent at CD4 counts of 100 to 200/mm ^{3}, disseminated MAC infection and CMV retinitis are rarely seen until the CD4 count is below 50/mm ^{3}.

The prognostic value of measuring HIV RNA in plasma was reported from the United States in 1996. In HIV-positive men in a subgroup of the Multicenter AIDS Cohort Study, only 8 per cent with less than 5000 copies of HIV RNA/ml progressed to AIDS over 5 years, whereas 62 per cent with viral loads above 35 000 developed AIDS. For a given level of CD4 lymphocytes, variations in viral load predict the risk of progression. The most useful prognostic information is therefore derived from the CD4 count and viral load taken together (Fig. 9).

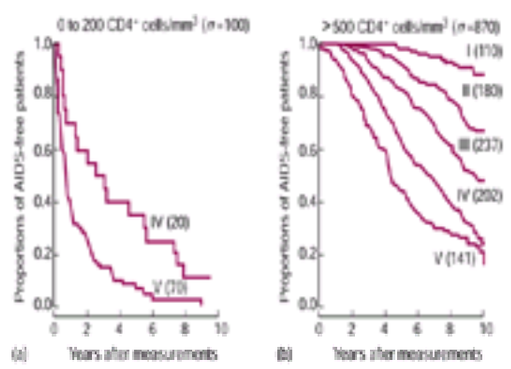

Fig. 9 Curves showing AIDS-free survival with time among groups with different baseline CD4 lymphocyte counts, according to HIV-1 RNA category. The five categories were (copies/ml): I, 500 or less; II, 501 to 3000; III, 3001 to 10 000; IV, 10 001 to 30 000; and V, above 30 000. (Sample sizes are shown in brackets).

In industrialized countries, HIV viral load measurements have become widely available. Techniques include reverse transcription followed by amplification by the polymerase chain reaction (**RT-PCR**), branched DNA (**bDNA**) signal amplification, and nucleic acid sequence-based amplification (**NASBA**). Highly sensitive tests with very low detection limits (about 20 copies/ml) are increasingly used.

Antiretroviral therapy

Nucleoside analogues

Knowledge of the viral lifecycle (Fig. 3) led to the development of a number of antiretroviral compounds with clinically useful activity against HIV (Table 3). The forerunner of these was zidovudine (AZT or ZDV), first shown to be active against HIV *in vitro* in 1985. Zidovudine, a nucleoside analogue that inhibits HIV reverse transcriptase, slowed down the rate of disease progression over a 12-month period in patients with AIDS and improved short-term survival, well being, body weight, and neurological features. However, clinical progression associated with viral resistance to the drug was observed after a year or two of therapy. When early treatment with zidovudine was compared to deferred zidovudine, there was no difference in survival or disease progression after 3 years.

The clinical failure of monotherapy prompted combination therapy in an attempt to reduce the development of drug resistance. Double nucleoside combinations proved superior to zidovudine monotherapy, especially in patients without prior exposure to zidovudine. Treatment with at least three drugs is more effective and has become the standard of care. In general, two nucleoside drugs are used with either a non-nucleoside reverse transcriptase inhibitor or a protease inhibitor. A combination of three nucleoside analogues (zidovudine, lamivudine, and abacavir) can also be used, and is available as a single tablet taken twice daily (Table 4).

Non-nucleoside reverse transcriptase inhibitors

The prototype of the class is nevirapine, a potent and selective inhibitor of HIV reverse transcriptase. When nevirapine is given alone, resistance develops rapidly and

this drug is of limited effectiveness in double therapy or when added to failing regimens. However, in antiretroviral-naive patients without AIDS (CD4 200 to 600/mm ^{3}), over a half of patients treated with nevirapine plus two nucleosides (zidovudine and didanosine) had undetectable plasma HIV RNA after 1 year of therapy, compared with 12 per cent for zidovudine/didanosine only. Efavirenz and delavirdine (which is not licensed for use in the United Kingdom) are other non-nucleoside reverse transcriptase inhibitors with similar properties to nevirapine.

Protease inhibitors

The HIV-encoded protease (or proteinase) is required for the production of mature infectious viral particles. This enzyme cleaves a number of structural proteins and enzymes from the polyprotein precursors produced by translation of the *gag* and *gag–pol* genes. Inhibitors of HIV protease act synergistically with nucleoside drugs and are potent inhibitors of HIV replication.

Protease inhibitors have a greater effect on HIV viral load and CD4 counts than nucleoside reverse transcriptase inhibitors, especially when used in triple therapy.

Indinavir, in combination with two nucleoside analogues (zidovudine/lamivudine or stavudine/lamivudine) produced good results in a large controlled trial with clinical endpoints (ACTG 320). Compared to double therapy (two nucleosides), the triple combination reduced the proportion of patients who progressed to AIDS or death from 11 to 6 per cent over about 38 weeks. The responses of CD4 cells and plasma HIV RNA paralleled the clinical results. Similar results were reported for combinations that involved other protease inhibitors, saquinavir, ritonavir, and nelfinavir. Ritonavir, in low dosage, may be included to boost blood levels of other protease inhibitors (especially saquinavir, indinavir, and a newer drug, lopinavir) by competitive inhibition of their hepatic metabolism. Combinations of non-nucleoside drugs and protease inhibitors are also being evaluated.

Other drugs

Fusion inhibitors, such as T-20, stop the HIV glycoprotein gp41 from effecting fusion of the viral and cellular membranes, and thereby prevent HIV entry into host cells. Compounds that inhibit HIV integrase and prevent proviral DNA integration into the host cell genome are also being identified. Hydroxyurea, not in itself an antiviral compound, is sometimes used in combination with nucleoside reverse transcriptase inhibitors; *in vitro* studies suggest that it acts synergistically by reducing the intracellular substrate for making DNA and thereby increasing the efficiency of chain termination. Another adjunctive agent under investigation is interleukin-2 (given subcutaneously), which raises CD4 lymphocyte counts substantially when used in combination with antiretrovirals. Influenza-like side-effects are prominent and therapy is very expensive. Its long-term efficacy is currently unknown and being studied in a large trial (ESPRIT).

General points on HIV therapy

There is a plethora of results from clinical trials of antiretroviral drugs, but several large, randomized controlled trials have made the greatest impact. Comparison between trials may be difficult because of differences in the clinical stage of HIV disease in those enrolled, CD4 counts at entry, previous antiretroviral experience, duration of treatment, and in the drug regimens used. Many trials measure surrogate endpoints, especially HIV viral load reduction and changes in CD4 lymphocyte count. It is assumed that these reflect clinical effectiveness. However, trials conducted over periods of less than 1 year may not predict longer term results. The value of such short-term studies based on surrogate markers is to identify treatments that should be evaluated in large, well-designed controlled trials that measure clinical endpoints (progression of HIV disease or death) ideally over several years. HIV trials may be stopped prematurely when significant differences in clinical outcomes are demonstrable between study arms, but before longer term benefits can be assessed. In fact, the long-term efficacy of currently recommended anti-HIV treatment regimens remains unknown.

When to start treatment

The optimum time to start antiretroviral therapy is not known, and no trials have adequately addressed this question. Data from several clinical cohorts suggest that patients who start treatment when the CD4 count is below 200/mm^3 have an increased mortality when compared with those starting at higher CD4 levels. Currently, there is no clear evidence for an advantage in starting treatment at any given range of CD4 count above 200/mm ^{3}. Therefore, recent guidelines generally recommend starting before the CD4 count drops to below 200/mm^3, or if the patient develops symptomatic HIV disease. Asymptomatic patients with CD4 counts in the range of 200 to 350/mm^3 whose CD4 counts are falling rapidly or who have a high viral load should be monitored more intensively, and earlier intervention may be considered.

What to start with

Highly active antiretroviral regimens consist of at least three drugs, usually a backbone of two nucleosides with either a non-nucleoside reverse transcriptase inhibitor or a protease inhibitor (see Table 4). As discussed above, for pharmacokinetic reasons, two protease inhibitors (one of which is low-dose ritonavir) may be used, and a triple nucleoside regimen is also available. The best starting regimen(s), and how treatment should subsequently be sequenced, have not been determined. No regimen or sequencing strategy has been shown to be clinically superior in controlled trials. Several factors should be taken into consideration when selecting initial therapy, including potential drug interactions, toxicity, and the likelihood of adherence. HIV viral load and CD4 count should be checked after 2 to 3 months. The aim of initial treatment is to achieve a reduction in viral load to undetectable levels (ideally <50 copies/ml) within 6 to 9 months of starting treatment. Whether initial regimens that include more than three drugs are clinically superior in the longer term is being studied in ongoing trials such as Initio.

Changing therapy

Recommendations for changing the treatment regimen are based on theoretical considerations. The principal reasons are treatment failure, toxicity, and poor adherence. There is no agreed definition for treatment failure. Patients whose viraemia was initially suppressed, and whose viral load subsequently rises, should be considered for changing to a completely new regimen of at least three drugs. This may be guided by a resistance test (see Drug resistance, below). However the optimal point at which the switch should be made remains to be defined. Poor absorption of protease inhibitors may sometimes cause treatment failure related to low blood levels, without development of resistance; measurement of blood levels may be useful in selected cases. In cases of drug toxicity (for instance, a severe rash), if the responsible agent is identified then a single drug substitution can be made. If adherence is poor or likely to be the cause of treatment failure, changing to a combination that is simpler to take should be considered, for instance based on once or twice daily dosage and low pill burden (see Patient adherence, below).

'Salvage' therapy

Salvage therapy is generally defined as treatment following exposure to multiple antiretroviral drugs. In this situation, numerous drug-resistance mutations are usually present and the likelihood of achieving sustained viral suppression below the detection level is much lower than for patients who have limited or no previous antiretroviral exposure. This is especially true if drugs from all three major classes have previously been used. Studies using clinical endpoints suggest that declines in viral load correlate with improvements in clinical outcome, even if suppression to below the detection limit is not achieved. Several factors may be considered when selecting a treatment regimen in these circumstances, including drugs or drug classes to which the patient has not been exposed, drugs with a lower likelihood of resistance that can be recycled, inclusion of new drugs or a new class of drug such as nucleotides, and results of tests for viral resistance. Whether 'mega' antiretroviral treatment using five or more drugs for salvage is superior to standard triple therapy is being examined in the OPTIMA trial in the United Kingdom, the United States, and Canada.

Patient adherence

A substantial proportion of HIV patients do not follow treatment recommendations. Reasons for non-adherence include poor communication, the complexity of drug regimens and number of tablets, disruption of life (including timing and food restrictions), side-effects, concerns about long-term effects, and lack of confidence in non-curative treatments of indefinite duration. Adherence to treatment requires a high level of understanding and motivation in the patient. This is of particular concern in HIV therapy because of the risk of developing drug-resistance mutations during suboptimal therapy. The recent development of simplified regimens (for example, once or twice daily dosage; reduced pill burden) has helped.

Drug resistance

Viral resistance is a major factor in treatment failure. There is evidence that resistant mutants arise spontaneously even in the absence of antiretroviral therapy. This

tendency is greatest when HIV viraemia is high, and lowest when HIV replication is completely suppressed by a potent drug combination.

Extensive genotypic variation of HIV occurs because of very high viral turnover and transcription errors by the reverse transcriptase enzyme, so that all possible single-point mutations are likely to occur frequently. While mutations causing resistance to single agents may be present before antiretroviral treatment, on statistical grounds it is unlikely that specific combinations of multiple mutations will be present. However, multiple mutations do develop during antiretroviral therapy with more than one drug when viral replication is at a high level. Therefore, controlling viral replication with a highly potent treatment regimen limits the appearance of resistant HIV mutants.

Genotypic and phenotypic assays have been developed to test for drug resistance in HIV isolates. Genotypic assays that identify codon mutations correlating with *in vivo* resistance to antiretrovirals are relatively easy to perform and inexpensive. Phenotypic assays that measure the ability of the virus to grow in increasing concentrations of drugs are time consuming and expensive, but provide more direct evidence of resistance to a particular drug. The clinical and prognostic value of resistance assays is being evaluated, and, increasingly, they are used in the selection of drug regimens and investigation of treatment failure. Interpretation of resistance patterns is increasingly difficult as the number of drugs and mutations involved increases.

Resistance mutations to antiretroviral agents are identifiable in up to 15 per cent of recent seroconverters in the United States, and transmitted drug resistance is increasing in Europe. Resistance mutations may be lost in the absence of therapy, so the clinical significance of the transmission of a virus carrying resistance mutations to one or multiple drugs is currently uncertain.

Drug toxicity and interactions

Adverse reactions to antiretroviral agents are relatively common; treatment may have to be stopped. Minor gastrointestinal disturbances (nausea, vomiting, diarrhoea), rashes, and headache are common, but some adverse reactions are serious. Drug interactions must be considered when prescribing antiretroviral drugs, especially in late HIV disease. Antiretroviral agents may interact with each other and with other drugs. Ritonavir, a potent inhibitor of cytochrome P-450, is especially prone to raising blood levels of other drugs and should not be given with most antiarrhythmics, anxiolytics, and antihistamines. Caution is required with several analgesics, anticonvulsants, and other categories of medication.

Metabolic complications, especially mitochondrial toxicity and disturbances of lipid and glucose metabolism, have emerged as important adverse effects of antiretroviral therapy. Mitochondrial toxicity is especially associated with nucleoside drugs (such as didanosine and stavudine) and may result in neuropathy, myopathy, pancreatitis, hepatic steatosis, and lactic acidaemia. Mild degrees of lactic acidaemia cause non-specific symptoms including malaise and gastrointestinal disturbance of gradual onset; progression can lead to fatal lactic acidosis. Nucleoside drugs are thought to cause mitochondrial dysfunction by inhibiting mitochondrial DNA polymerase-g.

A syndrome of lipodystrophy (loss of fat from face and limbs), truncal fat accumulation, hyperlipidaemia, and insulin resistance has been associated with protease inhibitors but has also been described with other antiretroviral drugs, and may not respond to changing the drug regimen. Whether the effects on lipid metabolism increase the risk of ischaemic vascular diseases is currently uncertain.

Considerable immune restoration seems to occur after treatment with potent antiretroviral regimens in patients with low CD4 lymphocyte counts. New disease manifestations have been reported. Disseminated MAC may cause widespread lymphadenopathy when antiretroviral therapy is started. Several new manifestations of CMV ocular disease have been reported, including vitritis, cystoid macular oedema (previously seen in HIV-negative patients following withdrawal of iatrogenic immunosuppression), and epiretinal membrane formation. These new disease manifestations are likely to be immune recovery phenomena.

Treatment interruptions

In general, once treatment is started it is continued indefinitely. There has been recent interest in whether interrupting treatment (in supervised or structured treatment interruptions) can be beneficial. In theory, such interruptions might enhance immune responses, reduce long-term toxicity, or reduce resistant virus by allowing repopulation with wild-type virus. The effect of treatment interruptions on clinical outcomes is being studied in clinical trials.

Late complications and their management

Pneumocystis carinii pneumonia

P. carinii pneumonia, one of the hallmarks of AIDS, is now less common because of primary prophylaxis and antiretroviral therapy. Some 85 per cent of cases occur in patients with CD4 lymphocyte counts below 200/mm^3, and mostly at counts below 100/mm^3. Symptoms—typically: increasing shortness of breath, dry cough, and fever—usually develop subacutely over a few weeks. Malaise, fatigue, weight loss, and chest pains or tightness may occur. Chest signs are usually minor (crackles) or absent. The characteristic chest radiograph shows bilateral perihilar interstitial shadowing (Fig. 10), but may be normal. Other appearances include localized infiltrates or consolidation, upper lobe shadows resembling tuberculosis, nodular lesions, and pneumothorax; effusions are very rare. The arterial oxygen saturation is less than 95 per cent at rest or falls after exercise.

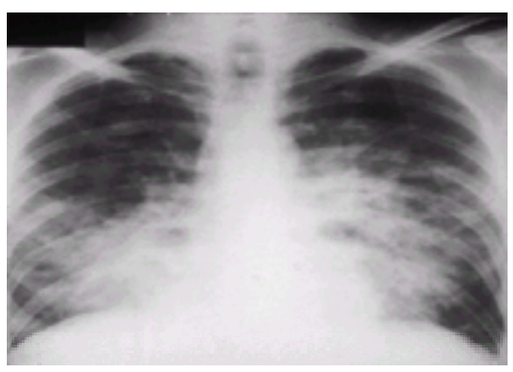

Fig. 10 Chest radiograph: *Pneumocystis carinii* pneumonia.

A foamy intra-alveolar exudate containing abundant *P. carinii* develops, which is associated with an interstitial inflammatory infiltrate and progressive impairment of lung function. The diagnosis can sometimes be confirmed by microscopy of sputum, which is induced by nebulized saline in isolated, properly ventilated rooms to reduce the risk of tuberculosis transmission (see Multidrug-resistant tuberculosis, below). *P. carinii* cysts and trophozoites are visualized by the use of special stains. If the result is negative, fibreoptic bronchoscopy with bronchial lavage may be indicated (Fig. 11); other causes of lung disease or coexistent infection may also be diagnosed by this technique, including tuberculosis, fungal infections, and Kaposi's sarcoma. Immunofluorescence using monoclonal antibodies, or DNA amplification by PCR, may improve diagnostic sensitivity when compared with conventional staining techniques, but these methods are not yet used routinely. In a minority of patients with *P. carinii* pneumonia the diagnosis is not confirmed.

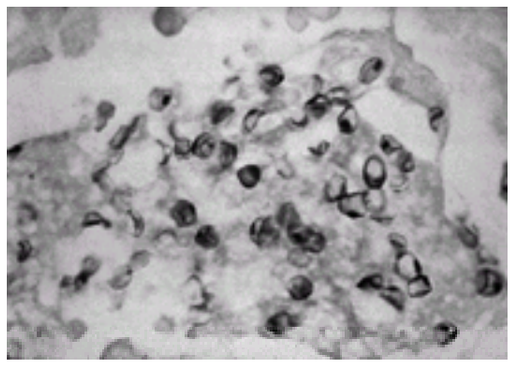

Fig. 11 *Pneumocystis carinii* cysts in bronchoalveolar lavage aspirate.

High-dose co-trimoxazole (120 mg/kg daily in divided doses) for 3 weeks is the first-line treatment for pneumocystis pneumonia. Oral therapy is often adequate, but in moderate and severe cases the drug should be given intravenously. The drug can be given orally if fever, symptoms, and oxygenation have improved after 10 days. Adverse reactions to co-trimoxazole—especially neutropenia, anaemia, rash, and fever—occur in up to 40 per cent of patients, usually after 6 to 14 days. Intravenous pentamidine (4 mg/kg per day) is the second-line choice for patients who do not tolerate co-trimoxazole.

Patients intolerant of co-trimoxazole and pentamidine may be treated with clindamycin plus primaquine or dapsone plus trimethoprim. These regimens have only been evaluated in patients with mild to moderate pneumocystis pneumonia, as has atovaquone, an antiprotozoal drug that is active against *P. carinii.* Although slightly less effective than co-trimoxazole, atovaquone causes fewer adverse effects.

In patients with moderate or severe pneumocystis pneumonia, high-dose corticosteroids reduce morbidity and mortality. If the arterial oxygen tension (PaO_2) is less than 9.3 kPa or the alveolar–arterial oxygen gradient is greater than 4.7 kPa, oxygen and intravenous methylprednisolone or oral prednisolone should be given for 5 to 10 days. Patients who develop respiratory failure may require ventilatory support. After treatment for pneumocystis pneumonia has been completed, secondary prophylaxis should be given to prevent recurrence. This can be discontinued if there is a good response to antiretroviral treatment, with a rise in the CD4 count sustained above 200/mm^3.

Bacterial pneumonia

The risk of bacterial pneumonia is increased in HIV, especially if the CD4 lymphocyte count is below 200/mm^3. The most common cause is *Streptococcus pneumoniae*; *Haemophilus influenzae* and *Moraxella catarrhalis* are relatively common, and *Staphylococcus aureus*, *Klebsiella* spp., and other Gram-negative rods are important causes in advanced HIV disease. Rarer causes include *Nocardia* spp. and *Rhodococcus equi.* The presentation may be atypical, and radiological appearances frequently include diffuse infiltrates that resemble pneumocystis pneumonia, as well as more typical segmental or lobar patterns. Cavitation with abscess formation, pleural effusion, and empyema may occur. HIV predisposes to recurrent invasive pneumococcal infections with bacteraemia; recurrent bacterial pneumonia in a 12-month period is an AIDS-defining condition. Chronic lung damage with bronchiectasis and colonization by *Pseudomonas aeruginosa* have been reported.

Other pulmonary complications

Disseminated fungal infections, including *Cryptococcus* spp., may involve the lungs. In endemic areas histoplasmosis, coccidioidomycosis, and disseminated *Penicillium marneffei* infection need to be considered (see Other disseminated infections, below). Invasive *Aspergillus fumigatus* infections may occur in patients with advanced HIV disease who have additional risk factors such as severe neutropenia. Patients usually have severe systemic illness. The radiographic appearances in all these fungal infections are usually non-specific. Bronchoalveolar lavage may be needed for diagnosis. HIV-associated lymphocytic interstitial pneumonitis causes diffuse abnormalities, usually in children but occasionally in adults. Bronchiolitis obliterans-organizing pneumonia is a steroid-responsive cause of lung infiltrates, probably a tissue response to various underlying conditions, which has also been reported in HIV and may be confused with pneumocystis pneumonia.

Tuberculosis

The interaction between HIV and tuberculosis was recognized early in the HIV epidemic. Studies in Central Africa in the mid-1980s showed that more than 60 per cent of newly diagnosed tuberculosis patients were HIV-positive at a time when the background seroprevalence of HIV in the population was much lower. Intravenous drug users were shown to have an increased risk of developing active tuberculosis if they were HIV-positive. After decades of progressive decline in the incidence of tuberculosis in the United States, notifications increased during the mid-1980s, soon after the emergence of the HIV epidemic. A similar trend was subsequently observed in western Europe. Globally, tuberculosis remains the most frequent life-threatening opportunistic infection in AIDS.

Most cases of tuberculosis in HIV-positive individuals represent reactivation of dormant bacilli. However, molecular typing of isolates of *Mycobacterium tuberculosis* by restriction fragment length polymorphism (**RFLP**) analysis suggests that up to 40 per cent are new infections. The WHO estimates that one-third of the world's HIV-positive population is co-infected with tuberculosis. In communities where *M. tuberculosis* is a common endemic organism, those who are immunosuppressed by HIV have an increased risk of relapsing or contracting new infections. Where the background prevalence of tuberculosis is low, the disease is uncommon in HIV-positive patients unless they become exposed, for instance through travel. Testing for HIV should be considered in patients presenting with active tuberculosis, and tuberculosis should be considered as a cause of unexplained symptoms in patients with HIV.

Active tuberculosis may occur at any time during the course of HIV infection. In early-stage HIV, it is more likely to present with the typical clinical features: subacute history of cough, fever, and weight loss, upper lobe cavitary disease and/or pleural disease on chest radiographs, and a positive skin test to tuberculin. In late-stage HIV, infected patients are more likely to present atypically with unusual chest findings, extrapulmonary involvement, and cutaneous anergy. The chest radiograph may be normal in up to 40 per cent of cases. Sputum smears should be examined for acid-fast bacilli. Blood cultures may be positive for *M. tuberculosis*.

Studies in Zambia have shown that, compared with HIV-negative patients, HIV-positive individuals with tuberculosis are less likely to be sputum-positive on microscopy, show less cavitation and more involvement of the lower lobes, and are more likely to relapse after completion of therapy and to die prematurely. Patients with advanced HIV infection are more likely to develop extrapulmonary tuberculosis involving lymph nodes, pericardium, liver, bone marrow, or meninges.

The standard 6-month regimen of three or four antituberculosis drugs (isoniazid, rifampicin, pyrazinamide, and ethambutol) is generally effective in patients with HIV, unless there is resistance to one or more of these first-line drugs. The drug regimen may need to be adjusted when *in vitro* sensitivity results are known. For fully sensitive organisms, after 2 months on three or four drugs, isoniazid and rifampicin should be continued for a further 4 months. Patients with pulmonary tuberculosis should be isolated initially. Contact tracing is important; HIV-positive contacts are at particular risk. Tuberculin testing is used to determine whether contacts should take isoniazid chemoprophylaxis.

Up to 20 per cent of patients with HIV experience adverse reactions to antituberculosis drugs. In HIV-positive patients with tuberculosis in Africa, the sulpha-based drug thiacetazone has been associated with serious skin reactions, including toxic epidermal necrolysis and fatal cases of Stevens–Johnson syndrome. Whereas response rates for conventional short-course tuberculosis treatment in industrialized countries are similar to those achieved in HIV-negative patients, in resource-poor countries and where compliance is less easily achieved, cure rates are lower and there is a risk that resistance will develop. Several countries have adopted 'directly observed therapy' to address this problem.

Multidrug-resistant tuberculosis

Over 15 outbreaks of multidrug-resistant tuberculosis (**MDRTB**) have been reported since the late 1980s. MDRTB isolates are resistant to at least two first-line antituberculosis drugs, most commonly isoniazid and rifampicin, and are often resistant to several agents. Most have occurred in HIV units in hospitals, but there have been outbreaks in prisons, drug treatment centres, and nursing homes. Most documented outbreaks have been in the United States. Elsewhere, over 200 people were involved in Buenos Aires, Argentina, and another outbreak affected over 100 people in Lisbon, Portugal. In MDRTB outbreaks, healthcare workers may become

infected. Initially, the mortality among HIV-positive patients was very high (up to 93 per cent), but more recently the outcome has improved because of more rapid diagnosis and treatment with at least four drugs to which the *M. tuberculosis* isolate is sensitive *in vitro*. To prevent outbreaks of MDRTB, special precautions are required when HIV-positive patients with possible tuberculosis are admitted to hospitals. Diagnosis must not be delayed, appropriate treatment must be started without delay, and drug resistance identified. Precautions include the isolation of patients in negative-pressure rooms, use of respiratory protection for staff, and special care during certain procedures such as bronchoscopy or nebulized pentamidine administration. With effective treatment, patients rapidly become non-infectious, but precautions need to be continued until the sputum is repeatedly smear-negative.

Mycobacterium avium complex

Patients with advanced HIV infection and CD4 lymphocyte counts below 50/mm^3 are at high risk of disseminated *M. avium* complex (**MAC**) infection, particularly in industrialized countries where it is reported to develop in up to 40 per cent of patients with AIDS. *M. avium* is a ubiquitous environmental organism of low pathogenicity that can be isolated from domestic water supplies. Infection is likely to be through the gastrointestinal tract. MAC infection becomes widely disseminated in those with advanced HIV and causes fever, night sweats, weight loss, diarrhoea, abdominal pain, anaemia, disturbed liver function, and reduced overall survival. The organism can usually be cultured from blood or bone marrow, or may be recognized as acid-fast bacilli in tissue biopsies (for example from lymph node, small bowel, or liver). It is unclear why the diagnosis is uncommon in underdeveloped countries; high mortality from other opportunistic infections at earlier stages of immunosuppression may be partly responsible.

MAC infection is intrinsically resistant to most first-line antituberculosis drugs. The optimal regimen has not been determined, and although clinical benefit and microbiological response is often achieved, survival benefit has been difficult to prove. Comparative trials suggest that initial therapy should be with two or three drugs: clarithromycin or azithromycin and ethambutol should be used, and additional rifabutin or a quinolone (e.g. ciprofloxacin) considered. In severely ill patients intravenous amikacin may be useful as the third agent. Lifelong treatment may be required to prevent relapse; but if immunity is restored by highly active antiretroviral therapy it may prove possible to cure MAC infection.

Other non-tuberculosis mycobacteria

Other mycobacteria, notably *M. kansasii*, *M. genavense*, and *M. celatum*, may cause opportunistic infections in those with HIV. *M. genavense*, which colonizes pet birds, was discovered in European patients with HIV and causes fever, diarrhoea, and severe weight loss. HIV does not seem to affect the incidence or natural history of leprosy (*M. leprae*).

Oesophageal candidiasis

Oesophagitis presents with retrosternal pain on swallowing, and in patients with HIV is most commonly caused by *Candida albicans.* Oesophageal candidiasis indicates advanced immunosuppression and is an AIDS-defining condition. The diagnosis should be suspected in a patient with oral candida and dysphagia, and may be supported by barium swallow or confirmed by endoscopy and biopsy. Treatment is with oral azole antifungal agents. Fluconazole may be more effective than ketoconazole. It may recur and in patients with severe immunosuppression, candida may become resistant to prolonged azole treatment. Resistance tends to develop gradually and can be monitored by *in vitro* testing. Such patients require treatment or continuous suppression with high doses of fluconazole (which is better tolerated than high doses of ketoconazole or itraconazole) or intermittent treatment with intravenous amphotericin. Azole-resistant oro-oesophageal candidiasis has become much less common since the advent of highly active antiretroviral therapy.

The differential diagnosis of oesophageal candidiasis includes oesophagitis caused by cytomegalovirus (**CMV**) or herpes simplex virus (**HSV**), which require specific antiviral therapy, and aphthous ulceration, which may respond to oral prednisolone or thalidomide.

HIV and the nervous system

The nervous system is a major site of involvement for direct and indirect complications of HIV at all stages of infection. All parts of the nervous system may be affected. In advanced HIV, opportunistic infections and tumours (lymphoma), and tissue damage caused by HIV replication in the brain and spinal cord, are important and relatively common during progressive HIV disease.

Cerebral toxoplasmosis

Cerebral infection with the intracellular protozoan *Toxoplasma gondii* is the most frequent infection of the central nervous system in AIDS when the CD4 lymphocyte count is below 200/mm^3. It usually results from reactivation of toxoplasma cysts in the brain, leading to the formation of focal lesions that are typically multiple but may be single. Symptoms develop subacutely and include focal neurological disturbance, headache, confusion, fever, and convulsions. On CT scanning the lesions appear as ring-enhancing masses with surrounding oedema (Fig. 12). Magnetic resonance imaging (**MRI**) is more sensitive and frequently detects lesions not visible on the computed tomography (**CT**) scan. Serum antibodies to *Toxoplasma* spp. are usually detectable; their absence makes the diagnosis unlikely but does not exclude it. Detection of toxoplasma DNA in cerebrospinal fluid by PCR is being evaluated as a diagnostic test. The principal differential diagnosis is cerebral lymphoma; other causes of focal brain lesions in AIDS include cryptococcoma, cerebral abscess (including infection with *Nocardia* spp.), tuberculoma, progressive multifocal leucoencephalopathy, and neurosyphilis. Brain biopsy is required for a definitive diagnosis, but is rarely performed. As toxoplasmosis is by far the most common treatable cause of focal cerebral lesions in HIV, it is standard practice to treat for this and only consider biopsy if there is no clinical improvement in 7 to 10 days.

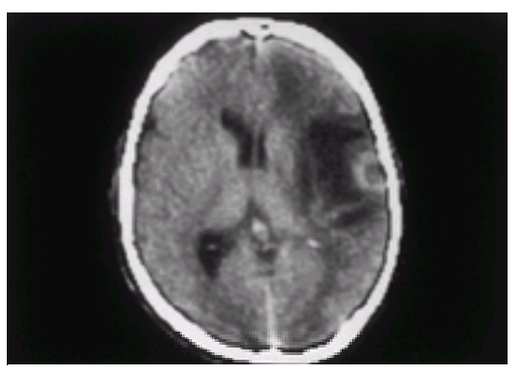

Fig. 12 Cerebral toxoplasmosis: ring enhancement and surrounding cerebral oedema (CT scan with contrast).

The condition responds well if treatment is started early; a combination of sulfadiazine at 4 to 6 g/day and pyrimethamine at 50 to 75 mg/day is the treatment of choice. More than 40 per cent of patients experience adverse effects, especially rash and nephrotoxicity caused by sulfadiazine. The haematological toxicity of pyrimethamine may be reduced by adding folinic acid (10 mg/day). If sulpha drugs are not tolerated, clindamycin with pyrimethamine is an effective alternative. Highly active retroviral treatment should also be started. Corticosteroids may be used to reduce cerebral oedema in patients with large lesions and serious mass effects, but this is controversial.

Treatment is usually given for 3 to 6 weeks, and in the absence of effective antiretroviral treatment relapse is common after stopping. In these circumstances, lifelong maintenance treatment is usually required using pyrimethamine (25–50 mg/day) with a sulpha drug or clindamycin. However, these can be discontinued if antiretroviral treatment leads to sustained immunological recovery.

Cryptococcal meningitis

Although infection of the central nervous system with *Cryptococcus neoformans* can occur in the absence of immunodeficiency, it most commonly arises in association with HIV infection. Before the widespread use of azole antifungals for mucosal candidiasis it accounted for 5 to 10 per cent of opportunistic infections in patients with AIDS. The presentation is usually subacute and may be subtle and non-specific with headache, vomiting, and mild fever, and few neurological signs. Less frequently, psychiatric disturbance, convulsions, cranial nerve palsies, truncal ataxia, or focal intracerebral lesions may occur. Neck stiffness is unusual. The diagnosis is made by identifying cryptococci in the cerebrospinal fluid by India ink staining, detection of cryptococcal antigen in the cerebrospinal fluid (uniformly positive), and culture. Cryptococcal antigen is also usually detectable in serum. *C. neoformans* in patients with AIDS causes minimal inflammation so the white cell count of the cerebrospinal fluid is often only mildly raised and the protein and glucose levels of the cerebrospinal fluid may be normal.

A randomized, controlled trial showed that the combination of amphotericin B and 5-flucytosine was superior to amphotericin B alone or fluconazole alone for the treatment of cryptococcal meningitis. Amphotericin B and 5-flucytosine together lead to more rapid sterilization of the cerebrospinal fluid but are not as well tolerated as fluconazole. Most patients should be given the combination, but milder cases may be treated with fluconazole alone. Resistance of cryptococcus to fluconazole is very rare. Itraconazole can be effective, but is not generally recommended. Adverse reactions to amphotericin are frequent, especially fever, myalgia, renal impairment, and electrolyte disturbances. Close monitoring is required. Lipid formulations of amphotericin are reserved for patients intolerant of conventional formulation. Raised intracranial pressure is associated with clinical deterioration and the risk of blindness: repeated lumbar punctures, ventricular shunting, or acetazolamide therapy may be required.

Without secondary prophylaxis, cryptococcal meningitis relapses in 50 to 80 per cent of patients with HIV in the absence of antiretroviral treatment. Oral fluconazole (200 mg/day) is effective for lifelong maintenance. This can be discontinued if antiretroviral treatment leads to sustained immunological recovery.

Progressive multifocal leucoencephalopathy

Progressive multifocal leucoencephalopathy is a progressive demyelinating condition of advanced HIV disease caused by JC virus, a polyomavirus cytopathic for oligodendroglia. It presents with focal neurological deficits, personality changes, or ataxia; headache and mass effects are absent. Brain MRI, the investigation of choice, usually shows multiple white-matter lesions. JC virus is detectable in cerebrospinal fluid by PCR, but this is not usually necessary for diagnosis. There is no specific treatment. Survival of less than 6 months is usual, but progression may sometimes be halted or reversed by highly active antiretroviral therapy. Cidofovir is active against JC virus and is being evaluated in patients. The other human polyomavirus, BK virus, is a very rare cause of encephalitis and interstitial nephropathy in AIDS.

HIV encephalopathy

HIV can infect the nervous system directly, leading to a variety of clinical problems. Most patients dying of AIDS show histological evidence of brain involvement including neurone loss. A smaller number (up to 10 per cent) develop the cognitive, behavioural, and motor abnormalities of dementia. In the early stages, there is impairment of concentration and memory and mood changes mimicking depression; gradual progression leads to intellectual incapacity and motor disability so that patients cannot care for themselves. Neurological signs include slow movement, incoordination, motor weakness, hyperreflexia, and extensor plantar responses; brain imaging shows reduced grey matter volume in the cortex and basal ganglia. Ultimately, a nearly vegetative condition develops with virtual mutism, inability to walk, and incontinence. These patients die within 2 years. Antiretroviral treatment can prevent, and in the earlier stages reverse, AIDS dementia.

Other psychological/psychiatric problems include anxiety, panic attacks, and depression. Psychotherapy may be helpful. Antidepressants may be needed in severe cases. Acute psychosis is rare. Dystonic reactions to various drugs, such as metoclopramide, are more common in patients with HIV.

In the late stages of HIV disease, the differential diagnosis of HIV dementia includes cytomegalovirus (CMV) encephalitis. This usually presents with rapidly progressive confusion and dementia, impaired consciousness, fever, cranial nerve lesions, and convulsions. MRI shows necrotizing periventriculitis; protein levels in cerebrospinal fluid may be elevated and CMV DNA is detectable in the cerebrospinal fluid by PCR. Ganciclovir and other anti-CMV agents may reduce progression.

Peripheral neuropathy and myelopathy

Peripheral neuropathy can occur at any stage of HIV infection, even at seroconversion, but is most common in advanced disease, when 10 to 15 per cent of patients have a distal symmetrical sensorimotor neuropathy of axonal type causing pain and paraesthesias that may limit walking and, less often, distal weakness and atrophy. Mononeuritis multiplex and acute inflammatory demyelinating polyneuropathy resembling the Guillain–Barré syndrome are also described, generally at an earlier stage. Drugs used in patients with HIV, including stavudine, didanosine, and vincristine, may cause or exacerbate peripheral neuropathy. HIV-related autonomic neuropathy may cause postural hypotension, diarrhoea, impotence, impaired sweating, and bladder symptoms. CMV infection in patients with AIDS presents with a lumbosacral polyradiculopathy causing sacral paraesthesias and numbness, lower limb weakness, and urinary retention that may progress to flaccid paraparesis if untreated.

HIV may involve the spinal cord directly causing a vacuolar myelopathy. This usually presents with bilateral leg weakness and sensory symptoms, usually paraesthesias, and may progress to spastic paraparesis, ataxia, and incontinence.

Ocular disease

Cytomegalovirus retinitis (see also Section 25)

Without antiretroviral therapy, up to 30 per cent of patients with AIDS (and CD4 lymphocyte count below 50/mm ^{3}) develop reactivation of CMV in the form of a destructive and blinding retinitis. This is rare in other types of immunosuppression. It usually presents with blurring of vision, scotomas, floaters, or flashing lights. The characteristic retinal changes are patches of irregular retinal pallor, caused by oedema and necrosis, and haemorrhages in a perivascular distribution (Fig. 13). The retinitis usually starts peripherally and progresses rapidly to involve the macula and whole retina, leading to blindness. Complications include retinal detachment, branch retinal artery occlusion, persistent iritis, and cataract. CMV retinitis should not be confused with cotton-wool spots (HIV retinopathy)—small, pale retinal lesions without haemorrhages that commonly occur in patients with HIV. These are benign and often come and go.

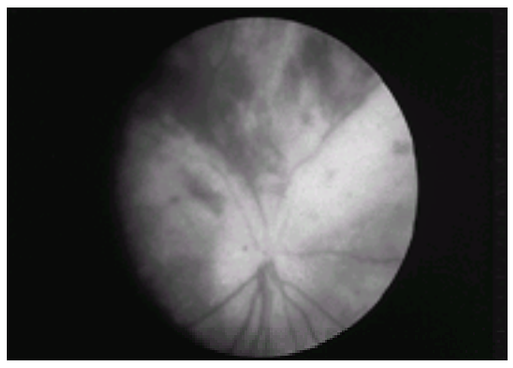

Fig. 13 Cytomegalovirus retinitis.

The diagnosis of CMV retinitis is clinical, based on the characteristic retinal appearance (see Section 25). CMV viraemia may be detectable by PCR and high or rising CMV viral load is associated with an increased risk of developing retinitis and other CMV disease. Anti-CMV drugs (ganciclovir, foscarnet, cidofovir) are virustatic; before the availability of highly active antiretroviral drug combinations, the aim of treatment was to stop progression rather than to cure disease. First-line treatment is with intravenous ganciclovir, which may cause severe neutropenia and thrombocytopenia that are dose-limiting in about 10 per cent of patients. Foscarnet

(phosphonoformate) is a relatively toxic second-line agent that causes dose-limiting reversible renal impairment and symptoms of hypocalcaemia in about 20 per cent of patients. Ganciclovir can also be given as a slow-release intraocular implant, but this may allow CMV to develop at other sites including the other eye.

For maintenance therapy, oral ganciclovir may be adequate, convenient, and well tolerated, although there is a greater risk of disease progression than with daily intravenous infusions of ganciclovir or foscarnet, and the eyes must be examined frequently. Cidofovir is more active against CMV than the other anti-CMV drugs. It can be given by intermittent intravenous infusion, initially weekly and then every 2 weeks. Whereas ganciclovir and foscarnet require a central venous catheter, cidofovir may be given in short infusions through a peripheral vein because of its prolonged antiviral effect. However, cidofovir is relatively toxic, causing irreversible nephrotoxicity, neutropenia, and peripheral neuropathy in over one-third of patients.

With the advent of highly active antiretroviral therapy, CMV retinitis is much less common in developed countries. Sustained suppression of HIV viral load and improvement in immune status can allow discontinuation of maintenance treatment. New manifestations of ocular CMV, such as vitritis, have been reported in patients treated with highly active antiretroviral therapy (see Impact of highly active antiretroviral therapy, below).

Other ocular syndromes

Acute retinal necrosis is a rare condition originally reported in reactivation of varicella zoster virus in otherwise healthy adults. In patients with advanced HIV infection it is usually preceded by dermatomal herpes zoster and typically presents with blurring of vision and pain in the affected eye. Progressive necrotizing retinitis leads to visual deterioration that may be associated with uveitis. An outer retinal necrosis syndrome with little ocular inflammation also occurs in patients with AIDS. There is a high risk of visual loss and retinal detachment. Both eyes may be affected. Suspected acute retinal necrosis should be treated with intravenous aciclovir.

Acute toxoplasma choroidoretinitis may resemble CMV retinitis, but the retinal scarring that follows treatment is distinctive. The disease is more common in countries such as Brazil and France where the background prevalence of toxoplasmosis is much higher than in the United Kingdom. Choroidoretinitis is also a rare complication of histoplasmosis and cryptococcosis.

HIV-related tumours

Kaposi's sarcoma

Kaposi's sarcoma characteristically presents as multiple, purplish nodular skin lesions (Fig. 14). Lesions start as small, pink, deep purple, or brown macules, and develop into nodules or plaques that may ulcerate. They also occur on mucosal surfaces, most commonly on the hard palate. Local or regional oedema and lymph node enlargement may occur. Mucocutaneous lesions are cosmetically and psychologically important but are rarely of clinical importance (Fig. 15). However, visceral disease, which most commonly affects the lungs and gastrointestinal tract, is an important cause of morbidity and even mortality. Lung lesions cause dyspnoea, cough, or haemoptysis, and gut involvement may cause abdominal pain, bleeding, or a rare protein-losing enteropathy. Extensive visceral involvement can cause constitutional symptoms such as fevers, night sweats, and weight loss. Kaposi's sarcoma rarely affects the central nervous system.

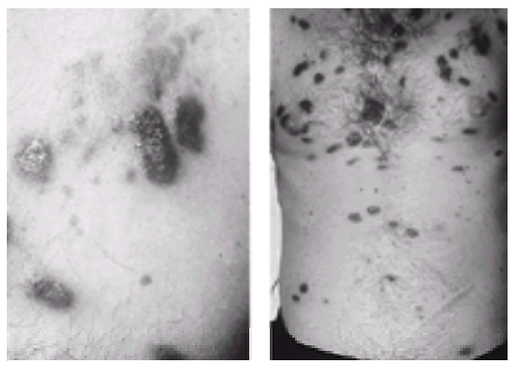

Fig. 14 Cutaneous Kaposi's sarcoma.

Fig. 15 Kaposi's sarcoma of the palate in a patient with HIV infection (copyright D. A. Warrell).

In industrialized countries, Kaposi's sarcoma is over 2000 times more common in HIV-infected individuals than in the general population. Classic Kaposi's sarcoma in HIV-negative individuals occurs in middle-aged and elderly men of Eastern European or Mediterranean origin. Endemic Kaposi's sarcoma in Africa has been known for decades. It is predominantly a disease of older men that has a fairly indolent course. HIV-related Kaposi's sarcoma, on the other hand, is a more aggressive disease and occurs mostly in those people who have acquired HIV via a sexual route, namely homosexual and bisexual men and in younger African men and women. The epidemic of Kaposi's sarcoma in Central and East Africa exactly mirrors the HIV epidemic in these regions. Kaposi's sarcoma is rare in intravenous drug users and very rare in recipients of blood products, including those with haemophilia. These epidemiological features suggested a sexually transmissible aetiological agent.

In 1994, a new herpesvirus, human herpesvirus-8 (**HHV-8**), was found in HIV-related Kaposi's sarcoma and was soon detected in the lesions of all forms of Kaposi's sarcoma. Seroepidemiological studies show that HHV-8 is common only in certain geographical regions, corresponding to where Kaposi's sarcoma was endemic before the era of HIV. HHV-8 is detectable in saliva but less often in semen. This may explain why both sexual and other routes of transmission occur. For instance in Africa, where HHV-8 infection is common, it is transmitted perinatally from mother to child.

Kaposi's sarcoma lesions are characterized by proliferating spindle cells, possibly of endothelial origin, thin-walled slit-like vascular spaces, infiltration by lymphocytes and plasma cells, and extravasated red cells. Multiple lesions appear synchronously in widely dispersed areas. Recent work has suggested a monoclonal origin for Kaposi's sarcoma lesions, but they may be reactive proliferative rather than truly cancerous. HHV-8 is detectable in spindle cells and flat endothelial cells lining the vascular spaces of Kaposi's sarcoma lesions. It is likely that the virus triggers the release of cellular and virus-encoded cytokines that promote the proliferation of spindle cells.

Highly active antiretroviral therapy has led to a dramatic reduction in the frequency and mortality of Kaposi's sarcoma in developed countries. In early Kaposi's sarcoma the progression is often halted or reversed by starting antiretroviral treatment alone. Otherwise, cutaneous lesions may be left untreated or treated with local radiotherapy, cryotherapy, or intralesional vinblastine. Widespread skin or visceral disease is usually treated by systemic chemotherapy, with single or multiple-agent regimens of vincristine, vinblastine, bleomycin, etoposide, and anthracyclines. The combination of vincristine and bleomycin is effective in 50 per cent of patients and well tolerated, but responses are usually short-lived. Liposomal preparations of anthracyclines (such as daunorubicin) are more effective and better tolerated, and are

now the treatment of choice. Treatment of disseminated Kaposi's sarcoma has not been considered to be curative, but remissions may be induced by a combination of highly active antiretroviral treatment and systemic chemotherapy.

Non-Hodgkin's lymphoma

Non-Hodgkin's lymphoma develops in 3 to 10 per cent of HIV-positive patients, an incidence 60 to 100 times higher than in the general population. Most tumours are extranodal and, histologically, 60 per cent are large-cell B-cell lymphomas; 30 per cent are Burkitt's type and the rest are of T-cell or non-B-, non-T-cell origin. Some 50 per cent are associated with Epstein–Barr virus (**EBV**) infection and are more aggressive with a shorter survival. A minority of HIV-related lymphomas are associated with HHV-8. They present as body-cavity lymphomas, causing pleural or peritoneal effusions (primary effusion lymphoma). Patients on highly active antiretroviral therapy have a reduced risk of developing Non-Hodgkin's lymphoma, and consequently the incidence of HIV-related lymphomas in developed countries has declined in recent years.

HIV-associated lymphoma outside the central nervous system may respond well to standard lymphoma chemotherapy regimens, in addition to highly active antiretroviral treatment. Response is better in those who are less immunosuppressed (CD4 above 200/mm ^{3} and no previous AIDS diagnosis). Opportunistic infections cause many deaths during chemotherapy. Lower dose or less toxic chemotherapy protocols are sometimes advocated for patients with more advanced HIV disease.

The central nervous system is a common site of HIV-associated non-Hodgkin's lymphoma, which is nearly always associated with EBV and sometimes with HHV-8 as well. Patients usually present with the symptoms and signs of a space-occupying cerebral tumour. Detection of EBV DNA in the cerebrospinal fluid may help to distinguish these lymphomas from cerebral toxoplasmosis. In the absence of antiretroviral therapy, neither chemotherapy nor radiotherapy have much impact, and the median survival after diagnosis is very poor, at about 3 months. However, this may be substantially prolonged in patients on highly active antiretroviral treatment.

Other tumours in AIDS

Some studies have reported an increased frequency of Hodgkin's disease in patients with HIV, particularly of the mixed cellularity type. Disseminated disease with a poor prognosis seems to be more frequent than for HIV-negative Hodgkin's disease. Castleman's disease (angiofollicular lymph node hyperplasia) is a lymphoproliferative condition that may be HHV-8 related and, in the multicentric form, is also associated with HIV. There is an increased incidence of squamous-cell carcinoma of the conjunctiva in patients with HIV infection, especially in Africa. HIV-infected women suffer a higher incidence of cervical intraepithelial neoplasia (**CIN**) and predisposition to cervical carcinoma; cervical cancer has been designated an AIDS-defining condition. The incidence of vulval intraepithelial neoplasia (**VIN**) is also increased by HIV infection. The incidence of squamous-cell anal carcinoma is increased in homosexual men, but the risk does not seem to be greatly magnified by HIV while the risk of anal intraepithelial neoplasia (**AIN**), a precursor of anal carcinoma, is significantly increased. The development of CIN, VIN and AIN may be related to co-infection with oncogenic types of human papillomavirus, especially type 16.

Common syndromes

Fever of unknown cause

Fever is rarely attributable to HIV infection *per se,* and patients should be fully investigated for other causes. There is increased susceptibility to pyogenic infections as well as opportunistic infections and tumours. Unlike fever of unknown origin (FUO) in other patients, however, most cases of FUO in HIV-positive patients are caused by an infection. Cultures of blood, urine, and faeces should be obtained, and chest radiography done. If sputum is available it should be stained and cultured for tuberculosis. In patients with intravenous lines and devices, catheter-related bacteraemia is an important cause. Occult infections (including sinusitis and dental sepsis) and drug-related fever should always be considered. Rarely, infection with *Pneumocystis* spp. and *Cryptococcus* spp. can present as fever without their typical focal signs, and dissemination to other sites (such as skin, fundi) may occur. Disseminated leishmaniasis is an important cause in those who have visited an endemic area (see Other disseminated infections, below).

In advanced disease, the most common causes of persistent high-swinging fevers are disseminated MAC infection (see *Mycobacterium avium* complex, above) and non-Hodgkin's lymphoma (see Non-Hodgkin's lymphoma, above). Imaging techniques, in particular CT scanning of the abdomen, are essential to find a suitable site for biopsy to diagnose lymphoma.

In some cases, the cause of fever is not found; in early disease the fever may resolve spontaneously. In advanced disease, fever and sweats may continue intermittently for many months. Symptomatic treatment includes non-steroidal anti-inflammatory drugs and low-dose prednisolone; therapeutic trials of anti-MAC treatment may be justified.

Breathlessness

Appropriate management of breathlessness is important because the most common cause, pneumocystis pneumonia, can be rapidly progressive. The differential diagnosis is broad, and includes bacterial pneumonia, pneumothorax, pulmonary Kaposi's sarcoma, other tumours, fungal infections, asthma, and heart failure. Routine investigations should include chest radiography, blood oxygen saturation, and peak flow measurement. Blood from febrile patients should be sent for culture. Sputum should be obtained, if necessary by induction with nebulized saline, and stained for the presence of *Pneumocystis* spp., mycobacteria, and fungi. Bronchoalveolar lavage should be considered early as patients sometimes progress quickly and become too ill for bronchoscopy without the support of mechanical ventilation.

Empirical treatment should cover *Pneumocystis* spp., *S. pneumoniae*, and *H. influenzae* by combining high-dose co-trimoxazole with a suitable broad-spectrum antimicrobial such as cefotaxime. A macrolide such as erythromycin or clarithromycin may be added if atypical pneumonia (such as mycoplasma) is suspected. If the clinical suspicion of pneumocystis is high, corticosteroids should be included if the patient is hypoxaemic (see *Pneumocystis carinii* pneumonia, above). Continuous positive airway pressure (**CPAP**) or mechanical ventilation may be needed in severe cases to allow diagnosis and time for patients to respond to treatment. If the chest radiograph shows diffuse bilateral infiltration and bronchoalveolar lavage fails to reveal any pathogen, presumptive treatment for pneumocystis pneumonia should be continued. If no diagnosis is made and deterioration occurs despite empirical treatment, open-lung biopsy should be considered to establish the diagnosis, but the prognosis is generally poor.

Diarrhoea

Chronic diarrhoea is a common problem in patients with advanced HIV infection, particularly in the tropics, and may be associated with weight loss and malabsorption. No cause other than HIV can be identified in at least half of cases; an HIV enteropathy characterized by partial villous atrophy has been described. The most common opportunistic cause is infection by the protozoan *Cryptosporidium parvum*, which causes a self-limiting gastroenteritis in those who are non-immunosuppressed. In HIV-positive patients diarrhoea may be protracted and severe, with marked fluid and electrolyte losses. The diagnosis is made by finding cryptosporidial oocysts in the stool using a modified acid-fast stain. Symptoms of cryptosporidiosis are often intermittent, as is excretion of the oocysts, so multiple stool specimens may need to be examined. Therapy with highly active antiretroviral drugs is the most important step in treatment. Symptomatic treatment with antidiarrhoeal drugs (such as loperamide), fluid and electrolyte replacement, and nutritional support are required. Octreotide may be useful in the most severe cases.

The coccidian protozoa *Isospora belli* and *Cyclospora cayetanensis* are important but less common causes of HIV-related chronic diarrhoea, diagnosed by the presence of sporocysts in the stool. Isosporiasis may respond to treatment with co-trimoxazole but the relapse rate after stopping treatment is 50 per cent, and similar experience is reported for cyclosporiasis. Microsporidia such as *Enterocytozoon bieneusi* and *Encephalitozoon intestinalis* are intracellular pathogens that may cause diarrhoea in advanced HIV disease. Special diagnostic staining methods applied to stool samples or electron microscopy of a rectal biopsy are needed to make the diagnosis. Albendazole may be effective, and treatment with highly active antiretroviral therapy may induce remission of intestinal microsporidiosis.

Giardia duodenalis and *Entamoeba histolytica* cysts are more commonly found in the faeces of homosexual men than heterosexual men, but these protozoa usually do not cause special problems in those with HIV. Bacterial infections with enteric pathogens such as salmonella, shigella, campylobacter, and *Clostridium difficile* do not lead to chronic diarrhoea but may take longer to clear. Salmonella infections may cause disseminated infection with recurrent bacteraemia that recurs even after prolonged antimicrobial therapy.

With advanced disease, MAC infection may cause diarrhoea among other symptoms. CMV causes a colitis that typically presents with abdominal pain and

tenderness, fever, and bloody diarrhoea. Rectal or colonic biopsy may confirm the diagnosis by identifying the characteristic nuclear inclusion bodies. HIV-related autonomic neuropathy is a rare cause of diarrhoea that is often most troublesome at night; anticholinergic drugs may help in addition to antidiarrhoeal agents. Diarrhoea in HIV-positive patients is a frequent side-effect of medications such as antibiotics and antiretroviral drugs.

HIV wasting syndrome

Weight loss is one of the most distressing features of progressive HIV infection. Its course fluctuates even in advanced disease; frequently it is attributable to specific complications such as diarrhoeal diseases, lymphoma, or disseminated MAC infection. It may also progress with no cause identified other than advanced HIV infection. In developing countries, such as those in sub-Saharan Africa, the wasting syndrome is characteristic evidence of AIDS and has been called 'slim disease'. Despite severe weight loss, patients may remain well for many months or even years. Numerous therapeutic approaches have been tried, mostly with disappointing results, including oral nutritional support, enteral feeding, total parenteral nutrition, and trials of growth hormone and thalidomide. Anabolic steroids such as nandrolone and stanozolol may reverse HIV-related wasting, but they may cause serious side-effects and their use has not been widely adopted. Highly active antiretroviral therapy is important in the prevention and reversal of HIV-related weight loss, but the fat redistribution associated with some antiretroviral agents may cause face and limb wasting (see Drug toxicity and interactions, below).

Miscellaneous conditions

Bacillary angiomatosis

Disseminated infection with *Bartonella henselae*, the principal agent of cat-scratch disease, is the cause of bacillary angiomatosis, an HIV-associated condition that typically causes multiple subcutaneous vascular lesions, fever, liver lesions (bacillary peliosis hepatis), and osteolytic bone lesions. The skin lesions are usually purplish nodules that may be mistaken for Kaposi's sarcoma, but the histology is distinct, acute neutrophilic inflammation and capillary proliferation, and clusters of bacilli revealed by modified silver staining. The organism may be cultured from blood. A similar syndrome in HIV-positive patients can be caused by the agent of trench fever, *Bartonella quintana.* Bacillary angiomatosis usually responds to treatment with a macrolide antibiotic. Cats and cat fleas form a reservoir for *B. henselae*, and patients who develop bacillary angiomatosis frequently have a history of contact with cats.

Other disseminated infections

In regions where invasive fungal infections are endemic (such as *Histoplasma capsulatum* in the Mississippi river region, *Coccidioides immitis* in the southern United States, and *Penicillium marneffei* in South-East Asia) or where there is a relevant travel history, disseminated fungal infection should be considered in HIV-positive patients presenting with fever, weight loss, anaemia, pulmonary infiltrates, lymphadenopathy, and hepatosplenomegaly. Papular skin lesions may be seen in disseminated histoplasmosis and *P. marneffei* infection. Similar lesions resembling giant molluscum (see Skin conditions in advanced HIV, below) may occur with disseminated cryptococcosis. Blood or bone marrow cultures or direct identification by the use of special stains on tissue obtained from skin lesions, bone marrow, or liver are required for diagnosis. Initial therapy is generally with intravenous amphotericin; itraconazole (for histoplasmosis and *P. marneffei*) or fluconazole (for coccidioidomycosis) may be adequate for subsequent maintenance treatment.

HIV-associated disseminated leishmaniasis is mostly reported from the Mediterranean littoral, South America, and Africa. It is caused by dissemination of *Leishmania* spp., protozoan parasites transmitted by sandflies. A high index of clinical suspicion is required because although the classic features are fever, weight loss, anaemia, and hepatosplenomegaly, a high proportion of patients have fever alone. Most cases can be diagnosed by bone marrow examination. Treatment is usually with the organic antimonial compound sodium stibogluconate, given parenterally.

Other visceral disease

Cryptosporidium (see Diarrhoea, above) and CMV may cause a sclerosing cholangitis-like syndrome with irregular dilatations and stenoses of the biliary tree (demonstrable by endoscopic retrograde cholangiography), abnormal liver blood tests, and occasionally jaundice. CMV is frequently identified histologically in the pancreas at autopsy, but its role in the development of clinical pancreatitis is unproved. A characteristic nephropathy (HIV-related glomerulosclerosis), primary pulmonary hypertension, and cardiomyopathy (see Section 15.16) are well described in patients with AIDS.

Haematological conditions

Thrombocytopenia is relatively common (5 to 15 per cent) in HIV infection and is associated with antiplatelet antibodies; symptomatic thrombocytopenia is uncommon but more likely in the later stages of HIV infection. Life-threatening bleeding is rare. Thrombocytopenia is not a marker for HIV progression and spontaneous remissions are frequent. When treatment is required, the principles and response are similar to those that apply in the treatment of HIV-negative immune thrombocytopenia, and include the use of prednisolone, intravenous immunoglobulin, and splenectomy. Thrombocytopenia also frequently responds to antiretroviral therapy using combinations that include zidovudine, which improves platelet production.

Anaemia is common in patients with advanced HIV infection, and is frequently related to medications (such as zidovudine). Human (B19) parvovirus infection is an important reversible cause of chronic anaemia in HIV infection. Bone marrow biopsy typically shows an absence of erythroid development with occasional giant pronormoblasts, and B19 parvovirus is detected by PCR. The anaemia may respond to treatment with intravenous immunoglobulin.

Mild neutropenia is common in HIV-positive patients at all stages of infection, and may be partly responsible for the increased risk of pyogenic bacterial infections; however, profound neutropenia (below 0.5×10^9/l) is rare. Antineutrophil antibodies may be present. Drugs (such as co-trimoxazole, ganciclovir, antiretrovirals) may increase the incidence and severity of neutropenia. In selected HIV-positive patients with refractory or life-threatening bacterial or fungal infection and severe neutropenia, the addition of recombinant human granulocyte colony-stimulating factor to the treatment regimen may improve the outcome.

Skin conditions in advanced HIV

In the later stages of HIV infection a number of infections have atypical cutaneous manifestations. These include giant molluscum contagiosum, characterized by large flesh-coloured non-tender umbilicated lesions often affecting the face in homosexual men. In advanced HIV disease, genital herpes simplex infection may cause painful chronic genital or anal ulcers that can become resistant to aciclovir and related compounds; intravenous foscarnet or cidofovir are effective. Aciclovir-resistant varicella zoster virus also occurs in AIDS; and reactivation of varicella zoster virus can take an unusual form, with a subacute course and dissemination causing scattered vesicular lesions in the absence of dermatomal zoster. CMV is a cause of chronic perianal ulceration that can be treated with ganciclovir. Atypical cutaneous presentations of syphilis may occur at any stage of HIV infection. In Asia, the varied skin manifestations of *P. marneffei* infection are familiar.

Children and HIV

Most paediatric infections result from the vertical transmission of HIV, although some children may be infected by blood products. The risk of vertical transmission is increased during advanced maternal HIV disease, if delivery is by the vaginal route, and if the baby is breast fed (see Vertical transmission, below). Diagnosis is important during the first year of life because about 20 per cent of HIV-infected children progress rapidly to AIDS during that time; however, a special diagnostic approach is needed before 18 months of age, because over this period uninfected children may have maternal HIV antibody. Techniques for virus detection (for example, HIV DNA by PCR) allow confirmation of HIV infection in 95 per cent of infected infants by 1 month of age. The sensitivity of these virological assays increases over the first few months, so if negative at 3 to 6 weeks, the tests should be repeated at 3 to 6 months; if HIV is not detected at this time, loss of HIV antibody should be confirmed at 15 to 18 months before the child is assumed to be HIV-negative.

HIV-infected children should be managed by paediatricians with experience in HIV care, usually in specialized units. About 10 per cent die in infancy, and progression to AIDS subsequently occurs at the rate of about 5 per cent per year. In recent European series, 40 per cent of children had developed AIDS before the age of 5 years and 25 per cent had died. The commonest AIDS diagnosis in infancy is pneumocystis pneumonia. The CD4 lymphocyte count is less valuable for monitoring than in adults, particularly in very young children; consequently prophylaxis against *Pneumocystis* spp. is usually given regardless of the CD4 count during the first year. In older children, the principles of monitoring are similar to those in adults, using clinical status, CD4 counts, and viral load estimation by plasma HIV-1 RNA measurement. The CD4 percentage (percentage of total lymphocytes) and CD4:CD8 cell ratio vary less with age and are more useful than absolute CD4 counts in

children under the age of 5 years.

In children, clinical conditions reasonably predictive of HIV infection include persistent oral candida, parotid swelling, and recurrent or frequent serious bacterial infections including pneumonia. Failure to thrive, diarrhoea, fever, lymphadenopathy, and hepatosplenomegaly are more common in HIV-infected infants but are non-specific and less predictive. HIV dementia, and other neurological and developmental problems are associated with a poor prognosis. HIV-related lymphocytic interstitial pneumonitis (**LIP**) is almost confined to children and characterized by progressive widespread reticulonodular shadowing on chest radiography. LIP develops gradually and may be asymptomatic; cough, breathlessness, clubbing, secondary bacterial infections, and bronchiectasis occur in severe cases and may be treated with oral prednisolone.

Principles of antiretroviral treatment are similar in children and adults. Clinical trials are in progress to determine optimal antiretroviral combinations, when to start treatment, and the tolerability of the newer drugs in all the major categories. Triple-therapy regimens are well tolerated in children and may produce sustained elevations in CD4 lymphocyte counts, but adherence is particularly difficult. As HIV-infected children grow older, the number of adolescents with perinatally acquired HIV is increasing, raising the need for advice on reducing the risk of sexual transmission.

Prevention of opportunistic infections (see Table 5)

The risk of developing an opportunistic infection rises greatly once the peripheral CD4 lymphocyte count falls consistently below 200/mm ^{3}. It is standard practice to introduce low-dose co-trimoxazole prophylaxis for pneumocystis pneumonia at this stage. This also reduces the risk of cerebral toxoplasmosis and may prevent bacterial pneumonia.

The risk of developing active tuberculosis in HIV-positive American intravenous drug users with positive tuberculin skin tests has been shown to be about 8 per cent per year and can be reduced by taking isoniazid for a year. In developing countries, in particular, the risk of active tuberculosis in HIV-positive individuals is high and isoniazid alone or in combination with rifampicin can reduce the risk, but the feasibility and cost-effectiveness of this approach in resource-poor countries require further evaluation. BCG vaccination does not appear to be protective in HIV.

Primary prophylaxis may prevent other conditions, such as CMV retinitis, cryptococcal meningitis, and histoplasmosis, but because of the relatively low incidence and lack of predictors of risk for these conditions, it is not cost-effective. Before the advent of highly active antiretroviral therapy, after treatment of an opportunistic infection the underlying tendency to the infection usually remained. Thus in early studies, following an episode of pneumocystis pneumonia, patients had a 50 per cent chance of a further episode within a year. Secondary prophylaxis with co-trimoxazole proved effective. Secondary prophylaxis for *Pneumocystis* spp. and other opportunistic infections, including MAC and CMV, can now usually be discontinued if there is a good response to antiretroviral treatment, with CD4 counts sustained above 200/mm^3 and low plasma levels of HIV RNA.

Simple measures, other than drugs, may reduce the risk of some infections. Avoiding undercooked eggs and poultry may reduce the risk of disseminated salmonella infection and adequate boiling of drinking water can prevent cryptosporidiosis. Stopping cigarette smoking may reduce the risk of bacterial chest infections

Prevention of HIV transmission

Sexual transmission

Sexual transmission accounts for most new cases of HIV infection. Education to alter behaviour and reduce the risk of HIV infection is an important part of HIV control programmes. The benefits of 'safer sex' should be publicized; condom promotion in Thailand has made an impact on HIV transmission rates. The presence of other sexually transmitted infections, especially those causing genital ulcers, facilitates HIV transmission. Accordingly, studies in Tanzania and elsewhere have demonstrated that programmes to prevent and treat sexually transmitted infections can reduce the incidence of new HIV infections.

Vertical transmission

As the number of women infected with HIV increases, the problem of vertical transmission of the virus assumes greater importance. In developed countries, the risk of transmission of HIV from a seropositive pregnant woman to her child is about 15 per cent, but this figure may be as high as 30 per cent in sub-Saharan Africa and other parts of the tropics. Although infection of the fetus can occur at any time during pregnancy and has been associated with breast feeding, most infections occur during labour. Zidovudine reduced the risk from 25 to 8 per cent, when given to women during late pregnancy and labour and to the neonate for 6 weeks. Simpler, cheaper regimens have been shown to be effective. During vaginal delivery, intrapartum interventions such as fetal blood sampling and use of fetal scalp clips should be avoided. Elective caesarean section further reduces vertical transmission (see British HIV Association quidelines in the Further reading list). Breast feeding should be avoided, if possible, but is still generally recommended in developing countries where the risks of bottle feeding probably outweigh the risks of breast feeding. Increasingly, the routine offer of HIV testing is being incorporated into antenatal care in developed countries where antiretroviral treatment is available.

Blood products

Screening of blood products began as soon as testing for HIV became available, and heat treatment for factor VIII concentrate was also introduced. These measures dramatically reduced the risk of virus transmission by blood and blood products in industrialized countries. However, there may still be a problem in developing countries where screening is not efficient, or where the background seroprevalence of potential donors is so high that HIV-infected blood may be screened as negative when donated by an individual in the 'window period' immediately after initial infection (see Diagnosis of HIV infection, above).

Injecting drug use

Needle-exchange programmes and the prescription of controlled drugs to registered addicts may reduce the incidence of new HIV infections in injecting drug users. Major problems still exist in countries such as India and Russia, where injecting drug use is becoming more common and education about the risk and the availability of clean needles is very limited.

Occupational exposure and postexposure prophylaxis

Based on data from more than 3000 occupational exposures to HIV, the average risk of HIV infection after needlestick injury or other percutaneous exposure was calculated to be 0.3 per cent (about 1 in 325). The risk following mucous membrane exposure has been estimated to be around 0.1 per cent. The risk of transmission is greatest for deep injuries; if there is visible blood on the device; during procedures involving direct cannulation of blood vessels; or if the source patient has advanced HIV disease. A small retrospective case-control study demonstrated an 80 per cent reduction in the likelihood of seroconversion in healthcare workers who took zidovudine soon after percutaneous exposure to HIV. In view of the greater activity of antiretroviral drug combinations but without direct evidence, it is currently recommended that high-risk occupational exposures to HIV are treated as soon as possible with two nucleoside inhibitors and a protease inhibitor (such as zidovudine, lamivudine, and nelfinavir) for 1 month. Nevirapine is not currently recommended in postexposure prophylaxis regimens because of a relatively high rate of adverse reactions. In the management of occupational exposure to HIV, a careful risk assessment should be done and information provided. If the risk of HIV transmission is identified, antiretroviral therapy should be offered and started promptly to maximize the chance of success. There is a theoretical argument for taking antiretroviral drugs after high-risk sexual exposure to HIV; at present there is no consensus on the appropriateness of postexposure treatment in this context, and no clinical or cost-effectiveness data are available.

Vaccine development

The high degree of viral variation and immune escape present difficulties for the development of an effective HIV vaccine. None the less, group-specific neutralizing antibodies and cross-reacting T-cell clones have been identified, and there is evidence from female prostitutes repeatedly exposed to HIV that certain individuals can develop specific T-cell responses without persistent infection. These individuals may be protected from infection when exposed to live virus.

To date, non-infectious killed whole virus or recombinant subunit vaccines have not been successful to date in protecting chimpanzees from HIV infection, or macaques from SIV infection and disease. Certain live attenuated strains of SIV, with deletion mutations in *nef* and other regulatory genes, initially appeared to

protect adult monkeys from challenge with virulent SIV strains, but more recently were reported to cause AIDS.

Human testing of candidate HIV vaccines, including a vaccine made from tiny recombinant fragments of gp120, the surface glycoprotein of HIV that binds to host-cell CD4 receptors, has so far not been successful. Several new approaches are being examined, which may prove more effective in inducing protective humoral and killer T-cell-mediated immunity. These include DNA vaccines, consisting of pieces of HIV DNA incorporated into harmless plasmid DNA from bacteria, and the use of live vectors (for example, poxviruses such as canary pox and modified vaccinia) to deliver portions of the HIV envelope. Researchers in Oxford are investigating a strategy ('prime–boost') using a DNA vaccine followed by boosting with a modified vaccinia vector vaccine. This approach is also being evaluated for therapeutic vaccination in HIV-positive patients with suppressed viraemia who are being treated with antiretroviral agents, to determine if vaccination will allow interruption of treatment without loss of virological control. Effective vaccination is likely to hold the greatest promise for controlling HIV infection in the future.

Further reading

Basic science

Clapham PR, Weiss RA (1997). Immunodeficiency viruses: spoilt for choice of co-receptors. *Nature* **388**, 230–1.

Emerman M, Malim MH (1998). HIV-1 regulatory/accessory genes: keys to unravelling viral and host cell biology. *Science* **280**, 1880–4.

Esparza J (2001). An HIV vaccine: how and when? *Bulletin of the World Health Organization* **79**, 1133–7.

Ho DD, *et al.* (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123–6.

Levy JA (1998). *HIV and the pathogenesis of AIDS*, 2nd edn. ASM Press, Washington DC.

Wyatt R, Sodroski J (1998). The HIV-1 envelope glycoproteins: fusogens, antigens and immunogens. *Science* **280**, 1884–8.

Clinical trials

Concorde Coordinating Committee (1994). Concorde: MRC/ANRS randomised double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. *Lancet* **343**, 871–81.

Delta Coordinating Committee (1996). Delta: a randomised double-blind controlled trial comparing combinations of zidovudine plus didanosine or zalcitabine with zidovudine alone in HIV-infected individuals. *Lancet* **348**, 283–91.

Hammer SM, *et al.* (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine* **337**, 725–33.

Epidemiology

Cascade collaboration (2000). Survival after introduction of HAART in people with known duration of HIV-1 infection. *Lancet* **355**, 1158–9.

Collaborative Group on AIDS Incubation and HIV Survival (2000). Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis. *Lancet* **355**, 1131–7.

Palella FJ Jr, *et al.* (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine* **338**, 853–60.

Treatment

Carr A, Cooper DA (2000). Adverse effects of antiretroviral therapy. *Lancet* **356**, 1423–30.

Flexner C (1998). Drug therapy: HIV-protease inhibitors. *New England Journal of Medicine* **338**, 1281–92.

Perrin L, Telenti A (1998). HIV treatment failure: testing for HIV resistance in clinical practice. *Science* **280**, 1871–3.

On-line resources

http://www.aidsmap.com/ [UK national guidelines (British HIV Association, regularly updated)]

http://www.hivatis.org/ [US HIV Treatment Guidelines Library (regularly updated), includes: 2001 USPHS/IDSA Guidelines for the prevention of opportunistic infections in persons infected with human immunodeficiency virus]

hivinsite.ucsf.edu/cochrane [Cochrane Collaborative Review Group on HIV Infection and AIDS]

http://www.unaids.org/ [Joint United Nations programme on HIV/AIDS]

7.10.22 HIV in the developing world

Charles F. Gilks

Introduction

The first definite evidence of human infection with HIV dates from a blood sample taken from an unidentified African man in Leopoldville (now Kinshasa) in the Congo in 1959. What happened before then is conjecture but it seems clear that HIV was originally an African primate virus, and that by the time it had spread to North America in the mid- to late 1970s the epidemic in sub-Saharan Africa was already well established. Public health surveillance is poor across Africa, and the disease went essentially unrecognized until AIDS had been identified as a new clinical entity in the United States. It is an uncomfortable truth that the main impetus to discover the cause of AIDS and produce reliable diagnostic tests, and the subsequent search for effective therapies and vaccines, has been the size and scale of the HIV epidemic in the West rather than the developing world.

Unfortunately, massive Western involvement with the disease has not always been directly beneficial. It has taken far too long to appreciate that HIV/AIDS is not the same disease in resource-poor countries as it is in rich North American or European cities; and that different clinical interventions and preventive approaches are necessary. Few textbooks deal with HIV/AIDS as anything but a disease of affluent communities which have access to high-technology medicine and expensive, state-of-the-art therapy. Many policy makers view the epidemic, and possible responses, in the impossible-to-replicate context of the costs to an industrialized health service—rather than what is necessary and can be achieved even with limited resources.

Aetiology

HIV has two variants, type 1 and type 2; each has different groups and subtypes or clades. HIV-1 group M is the cause of the pandemic: in industrialized countries subtype B predominates. In Africa two rare groups N and O have recently been identified; and whilst within group M the whole alphabet of subtypes A to J exist, subtype B is uncommon. In other parts of the developing world, because of founder effects, there is less HIV-1 heterogeneity. HIV-2 is largely restricted to West Africa although localized foci exist elsewhere. Dual infection can occur and prior infection with one type does not appear to generate useful protection against the other.

Most commercial diagnostic kits identify all types and variants of HIV. Typing and subtyping requires special resources and is not routinely carried out. HIV-2 appears to have important biological differences from HIV-1: it is less efficiently transmitted both sexually and vertically; disease progression is slower but it results in the same spectrum of related diseases. Less is known about the biological attributes of HIV-1 subtypes, although disease progression may be faster with some types (for instance subtype A compared with D in Uganda). Immune responses to infection may be both type subtype-specific, which has considerably complicated vaccine development. This chapter deals only with HIV-1 infection.

Epidemiology

Distribution

All parts of the world have reported HIV infection (Table 1). Current estimates (end 1999) suggest that 50 million people have been infected, of whom over 16 million have already died. Sub-Saharan Africa has borne the brunt of the epidemic: although constituting about 10 per cent of global population, 70 per cent of people living with HIV and 85 per cent of deaths come from the region. Ominously, the virus is continuing to spread rapidly in highly populous regions of India and South-East Asia. India has recorded 4 million cases, the largest number. In parts of east and central Africa, prevalence is stabilizing, sometimes at rates in excess of 30 per cent, and the 'HIV endemic' is emerging. In Uganda and Thailand, recent declines in HIV prevalence have been ascribed to behavioural change and successful control programmes. However, there is no developing country where the epidemic is in steady state. The burden of HIV/AIDS disease will continue to grow and the full impact of the epidemic on development and civil society will not be felt for many years.

Transmission

In developing countries, HIV is spread predominantly through heterosexual intercourse. Women become infected an average of 5 to 10 years earlier than men and in Africa more women than men are now infected. Several socio-economic and biological cofactors enhance transmission. In many cities there are large pools of migrant labour, women have few marketable skills, and there is often a preponderance of single men—all of which encourages the sex trade. Poor people have limited resources to devote to safer sex. Early sexual debut and frequent partner changes are common in some communities. Sexually transmitted infections facilitate transmission; in resource-poor societies treatment may often be delayed, incorrect, or inadequate.

Mother-to-child transmission is a direct result of the adult heterosexual epidemic. Infants can be infected transplacentally, during the birth process, or through breast feeding. Without intervention, overall risks are about 30 per cent. In high-prevalence regions up to 10 per cent of young children may be HIV infected. Those who escape infection may become orphans. The problem of AIDS orphans, particularly in urban centres, is a growing crisis. No simple solutions are emerging.

Intravenous drug use is widespread in certain areas and can be the vehicle for explosive transmission—as has occurred in Thailand, Vietnam, north-east India, and the Russian Federation. Needles and syringes may be shared by many individuals and tragically the epidemic may only be recognized when HIV is already well-established. Transmission through infected blood is important where transfusion is used indiscriminately, where inadequate provision is made for screening and quality control, where contaminated equipment is reused, and where professional donors and commercial blood banks resist supervision and regulation. China in particular is facing up to the consequence of poorly supervised and regulated blood transfusion practices.

Surveillance and disease burden

Accurate surveillance is essential to monitor the evolution of the HIV epidemic. Surveillance is often inadequate because of budgetary constraints, limited public health capacity, and occasionally lack of political will. Few countries can measure HIV incidence rates reliably. Antenatal clinic attenders have been the main focus for recording prevalence. At-risk or core transmission groups such as attenders of sexually transmitted disease (**STD**) clinics, and prostitutes are also important groups to

monitor. Many countries have only limited and incomplete national prevalence data. The absence of such basic information has hindered the effectiveness of AIDS control programmes across the developing world.

Surveillance for clinical disease has focused on AIDS, using a clinical case definition and monitoring HIV prevalence in patients with tuberculosis. Given the importance of non-AIDS disease and death and very poor AIDS case reporting across the developing world, the continued promotion of AIDS surveillance must be questioned. Assessing HIV prevalence in specific patient groups, such as those admitted to hospital with acute conditions, will generate more accurate and useful information for health planners.

In high-prevalence countries the burden of disease caused by HIV/AIDS is far broader than appreciated—consequences of fear and denial in the community and inadequate disease surveillance (Fig. 1). Failure to recognize the true impact of HIV has led to fragmented care responses, which serve inadequately the needs of communities affected by HIV. Stigmatization is maintained and potential benefits of enhancing prevention, by linking it with effective care packages, are not realized.

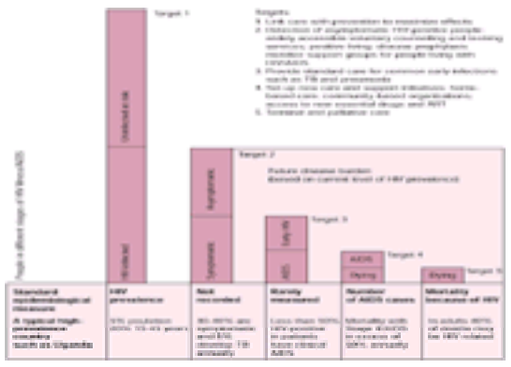

Fig. 1 The HIV/AIDS care burden in a community.

Prevention and control

Control strategies relate to mode of transmission. All require appropriate health information and education, to inform rather than scare and to promote behaviour change. Unfortunately, the background in many developing countries continues to be one of hostility, fear, and denial, in which those infected are discriminated against and stigmatized. Until there is a more rational and accepting attitude to HIV/AIDS, it is difficult to see how interventions can be fully effective. The United Nations Programme on HIV/AIDS (UNAIDS) addresses this by advocating the view that people with HIV/AIDS are not the problem but are part of the solution.

Strategies to reduce sexual transmission concentrate on the universal message of safer sex, condom usage, and improved treatment of sexually transmitted diseases. When properly used, condoms are effective barriers to all sexually transmitted pathogens—not just HIV. Condoms can be distributed free, or sold by local traders who buy at discounted wholesale prices, a process known as social marketing. This encourages the condom market and accepts than clients are more likely to use condoms that have been bought rather than given away. In some countries this has greatly expanded the use of condoms. In other countries, male reluctance is an impediment to widespread usage.

The rationale for improving STD treatment to control HIV is based on the increased risks that any STD, inflammatory or ulcerative, generates for person-to-person transmission of HIV. Initial enthusiasm for this was based on a single, randomized, controlled trial in East Africa (Mwanza, Tanzania) which documented a 42 per cent (95 per cent confidence interval 21–58) reduction in transmission. A larger study in Rakai, Uganda, showed no effect of mass community-wide STD treatment. One further randomized clinical trial is still underway in Uganda. It may be that differences relate to the stage of the epidemic. Improved STD treatment may have most impact where the virus is still epidemic rather than endemic, and where sexual contact with high-risk core transmitter groups is more important that transmission from spouse or regular partner.

In the West, mother-to-child transmission can be virtually eliminated with comprehensive antenatal counselling and HIV testing, antiretroviral therapy, elective caesarian section, and avoidance of breast feeding. In high-prevalence developing countries, where need is greatest, the financial resources and capacity to implement such interventions are woefully lacking. The promise of simple, cheap therapy with nevirapine reduces the cost barrier and has created a wide-ranging international debate with calls for charitable donations or concerted bilateral aid. However, reducing mother-to-child transmission may not be considered a leading public health priority in developing countries, particularly without robust policies addressing the care of AIDS orphans.

With intravenous drug users, education programmes promoting risk reduction, particularly avoiding the sharing of needles or syringes, have been effective in cities such as Bangkok. Implementing needle exchange programmes is usually a political, not public health, issue. Spread through infected blood has been the easiest route to control. Widespread provision of facilities and equipment for screening and, more recently, the advent of rapid single-use HIV tests have minimized the number of infected units transfused. Quality control remains the biggest single issue in maintaining the integrity of the blood supply. Establishing clear and appropriate guidelines for transfusion has reduced dramatically the numbers of units given, especially to anaemic children with malaria.

Clinical features

Natural history

Most people in the developing world who are immunosuppressed through HIV are poor. They are usually forced to live in unhygienic environments with inadequate sewerage and limited clean water, in overcrowded dwellings in close proximity to other sick and healthy people. There is substantial exposure to airborne and water-borne pathogens ensuring that rates of tuberculosis, pneumococcal disease, salmonellosis, and cryptosporidium are higher than in industrialized countries. With intense pathogen exposure, disease early on in the natural history of HIV is far more common (Fig. 2).

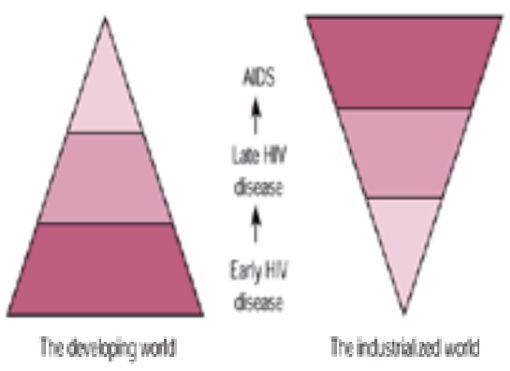

Fig. 2 The burden of disease caused by HIV infection in the developing world and the industralized world.

Poverty also limits the ability to maintain a healthy lifestyle and markedly reduces access to health care. Health-seeking behaviour is delayed and presentation with advanced disease is common. The quality of care available is often substandard. Health centres and hospitals may be rundown, poorly staffed, and lacking in even the most basic of drugs. Diagnostic capacity can be limited with few or no laboratory services provided. It seems obvious that survival will be severely compromised by

such disadvantages.

Surprisingly, despite the scale of the HIV/AIDS epidemic in the tropics, there are few cohort studies in which the natural history of disease has been well characterized. In one seroincident cohort in rural Uganda, initial disease progression was similar to that in Western pretherapy cohorts, and suggested median survival of about 9 years. Several studies document poor survival, a year or less, following the onset of a major AIDS-defining illness. There are no data to suggest that reduced survival with HIV/AIDS in Africa relates overall to more rapid disease progression. Rather, early death from readily preventable or treatable infections shortens survival. Perhaps, once severe intercurrent disease has developed, viral replication may be enhanced causing more rapid disease progression. To date this has not been documented.

Clinical staging

A clinical staging classification has been developed by the World Health Organization for use in clinics without access to CD4 counts or viral load measurement. Patients with asymptomatic disease or minor clinical problems are classified as stage 1 and 2, respectively. The importance of high-grade virulent pathogens early on in the disease process is encompassed; pulmonary tuberculosis (now classified as an AIDS-defining disease in the United States) and severe bacterial infections are stage 3 problems. Opportunistic infections characteristic of AIDS are classified as stage 4 events. The staging system has recently been validated against CD4 counts in Uganda and is useful for predicting survival.

Specific clinical problems

Tuberculosis

HIV-related tuberculosis is of supreme importance. HIV renders the individual highly susceptible to primary infection or reinfection and is a potent reactivator of latent tuberculosis. About 75 per cent of patients have thoracic disease. Classic pulmonary tuberculosis with positive sputum smears is seen in the early stages of HIV, pleural effusions are common, and hilar or paratracheal glands are frequently enlarged. With more advanced immunosuppression, extensive pulmonary disease with negative sputum smears develops. Extrapulmonary disease is common and dissemination with mycobacteraemia is increasingly recognized. Clinical features suggest both acute primary disease and reactivation are occurring.

Diagnosis can be difficult when only radiology and microscopy are available, particularly in smear-negative, disseminated, or extrapulmonary disease. Patients with early HIV disease usually respond well to standard therapy. National treatment guidelines should be followed and all cases notified to the tuberculosis control services. Thiacetazone is associated with high rates of hypersensitivity and should be avoided. In Africa, 25 to 30 per cent of patients may die during or soon after completing appropriate therapy, from community-acquired bacterial infections rather than poor therapeutic response. Treatment failure because of drug resistance is not yet a major problem in Africa but will be elsewhere. Tuberculosis frequently recurs through relapse and reinfection; rifampicin-based therapy may reduce relapse rates.

HIV-associated tuberculosis is preventable. The challenge is how to implement chemoprophylaxis without compromising existing tuberculosis control programmes or promoting drug resistance. One study in Abidjan showed that cotrimoxazole prophylaxis significantly improved survival of patients with tuberculosis/HIV, largely by reducing intercurrent bacterial infections. Studies are underway elsewhere to see how universally valid such an approach may be, given high rates of cotrimoxazole resistance across the developing world.

Pneumococcal infection

Streptococcus pneumoniae infection frequently develops with HIV. Lobar pneumonia accounts for about two-thirds of cases; acute sinusitis, occult bacteraemia, meningitis, pericarditis, skin sepsis, and pyomyositis also occur. Recurrent disease, primarily reinfection, is extremely common with rates of 25 to 30 per cent documented in East Africa.

Most patients with pneumonia are bacteraemic and Gram-positive diplococci are abundant in sputum. Patients respond well to standard penicillin therapy; ampicillin, chloramphenicol, and erythromycin are also effective. Lack of response suggests a second pathogen, often tuberculosis. Mortality rates are consistently higher in seropositive than seronegative patients. In Nairobi, penicillin resistance appears more frequent in pneumococci isolated from HIV-infected patients. Penicillin-resistant pneumococci are widespread; the worry is that HIV will facilitate further spread and evolution.

Polysaccharide pneumococcal vaccine is recommended (without efficacy data) in the United States and Europe. A randomized controlled trial in Uganda showed no efficacy but significantly higher rates of pneumonia in vaccine recipients. It should not be recommended in developing countries. The role of the new conjugate vaccines remains to be established. Studies in Abidjan showed reduced rates of pneumonia as well as other HIV-related problems with cotrimoxazole prophylaxis. Compliance remains a problem with long-term chemoprophylaxis.

Non-typhi salmonellas

Systemic salmonellosis is highly associated with HIV. *S. typhimurium* and *S. enteritidis* are most frequent; other non-typhi salmonellas are less important and HIV does not significantly predispose to *S. typhi* infection. Most patients present with bacteraemia and an enteric fever-like illness clinically indistinguishable from classical typhoid fever. Non-typhi salmonella coinfection also develops in patients with tuberculosis or pneumonia. Diagnosis requires blood culture. With inadequate microbiology facilities, few cases are identified and there is poor recognition of the general importance of non-typhi salmonellas or indeed other Gram-negative bacteraemic infections.

Patients with bacteraemia respond poorly if therapy is delayed, the diagnosis missed, or the organism is resistant. Antimicrobial resistance is common, especially to the widely used broad-spectrum antibiotics. Chloramphenicol, ampicillin and gentamycin, quinolones or third-generation cephalosporins are appropriate; use depends on cost and availability. Most centres treat for 2 to 3 weeks although optimal duration is not established. Relapse and reinfection both occur. Relapse usually develops within 6 to 8 weeks of stopping therapy and close follow-up is indicated. Quinolone maintenance therapy is effective but too costly for most patients. There are many environmental sources and no obvious ways to prevent exposure. Cotrimoxazole prophylaxis may prevent bacteraemia caused by non-typhi salmonellas.

Disseminated fungal infections

Cryptococcus neoformans is the most common disseminated fungal infection associated with HIV. There can be wide regional variations in incidence. In Abidjan evidence of infection was seen in only 2.5 per cent of autopsies. One cohort in Uganda diagnosed cryptococcosis at or around death in 25 per cent of patients. Most patients have subacute or chronic meningitis; some present with fever, non-specific pulmonary symptoms, and fungaemia; and cutaneous nodules occasionally develop. Diagnosis can be made by culture, antigen detection, or cerebrospinal fluid microscopy using Indian ink. Both fluconazole and amphotericin B are effective but lifelong maintenance therapy is required. Antifungal drugs are expensive and not usually considered essential drugs in low-income countries. Most poor families cannot afford therapy. Significant price reductions have recently been announced. Whether fluconazole prophylaxis is effective and implementable remains to be established.

Penicillium marneffei is an important HIV-related systemic mycosis in parts of South-East Asia. *Histoplasma duboisii* has been reported in HIV-positive patients in Central Africa and *H. capsulatum* in Central America and the Caribbean. Coccidioidomycosis and paracoccidioidomycosis have been associated with HIV infection in South America. Several other fungal diseases ecologically or geographically restricted to parts of the tropics may yet emerge as HIV-associated pathogens.

Chronic diarrhoea

Chronic diarrhoea with wasting, called 'slim' across Africa, is the most recognizable and obvious manifestation of AIDS in developing countries. Around 20 per cent of patients in a Ugandan cohort had slim at death. Rates elsewhere are not established.

Many potential pathogens have been associated. Cryptosporidium is probably the most important; prevalence of cyst excretion in African case series ranges from 5 to 48 per cent. In some areas *Isospora belli* is common; with improved diagnosis microsporidia are increasingly recognized. Other protozoan and helminth parasites are

not important. Enteric bacteria can be isolated in perhaps 15 per cent of cases but their pathogenic role is uncertain; enteric viruses are not thought important. Disseminated tuberculosis is probably an agonal endstage infection. As in the West, only about 50 per cent of patients with slim fully investigated will have a putative pathogen identified.

There is no effective therapy for cryptosporidium. Isosporiasis clears with cotrimoxazole but maintenance therapy is necessary. Some microsporidia may respond to albendazole. Imodium or codeine phosphate can initially relieve symptoms; whether benefits are maintained with long-term treatment is not known. Enteric pathogens are ubiquitous and it is difficult to avoid exposure. Hygiene is important but, in practice, is not always easy to implement or affordable. Cryptosporidium cysts can withstand chlorination and it is necessary to bring water to the boil.

The endemic tropical parasitic diseases

For years it has been unclear whether HIV interacts significantly with malaria. As with other intercurrent infections, HIV viral load rises with acute malaria. There is now mounting evidence from pregnant women and adults that HIV-related immunosuppression reduces protective antimalaria responses. The prevalence of placental parasitaemia is twice as high in HIV-positive compared to HIV-negative multigravidas, and this significantly increases the risks of mother-to-child HIV transmission. In endemic malaria areas, adults with HIV infection experience increased rates of clinical malaria with higher parasite densities. Whether severe complicated disease is more common and response to therapy is compromised as malaria-specific immune responses are lost remains to be established. In adults who have no background immunity to malaria, it also appears that HIV is a significant risk factor for severe and complicated disease and death with epidemic *P.falciparum* infection. There are very few data on malaria in HIV-infected children or on the interactions of HIV with *P.vivax*.

Leishmaniasis is well described in patients with AIDS from southern Europe following acute exposure or reactivation of latent infection. There are few case reports from endemic tropical areas. HIV prevalence is low in many remote rural foci, the disease is difficult to diagnose, and patients may die before being immunosuppressed enough to develop disseminated leishmaniasis. Any significant interaction would be of great importance in the kala-azar belt of India.

Unusual manifestations of Chagas' disease have been linked with HIV. There is no clear association with African trypanosomiasis and neither pathogenic amoebas nor giardia are exacerbated by HIV infection. A few cases of *Strongyloides stercoralis* hyperinfection have been described; despite being AIDS defining, it is very uncommon. To date, there are no data suggesting important interactions between HIV and schistosomiasis (except perhaps to reduce egg excretion), other flukes, hookworm, ascaris, filariasis, or any of the cestode/tapeworm infections.

Malignancies

Kaposi's sarcoma and HHV8 are endemic in sub-Saharan Africa. However, Kaposi's sarcoma develops in less than 5 per cent of African patients with AIDS, perhaps related to different patterns of HHV8 exposure. HIV has markedly altered Kaposi's sarcoma epidemiology. Women are more likely to be affected although there is still a male preponderance. Lesions are more extensive, frequently involve the mucosa, and often the viscera. Disease is more aggressive and can progress rapidly. Cytotoxic therapy is expensive and not widely available in developing countries. A relatively cheap regimen using actinomycin D and vincristine has proved effective in Zambia. There are few data on Kaposi's sarcoma and HIV from other regions.

Conjunctival squamous cell carcinoma has been recorderd in East Africa. Lymphomas are uncommon and occur much less frequently than in industrialized countries. No impact on cervical carcinoma has been reported from the tropics.

Opportunistic infections

Many African studies have looked carefully for the classic Western AIDS-defining opportunistic infections. *Pneumocystis carinii* is diagnosed by bronchoscopy or at autopsy in perhaps 2 per cent of patients with chronic lung disease. *Mycobacterium avium* has been isolated from the environment but mycobacteraemic disease is rare; disseminated tuberculosis is far more common. Toxoplasmosis or active cytomegalovirus infections are rarely diagnosed or clinically manifest, although they may be seen at autopsy. The most likely explanation for their rarity is that few poor patients survive long enough to develop profound immunosuppression. In developing countries only the urban elites have such a 'Western' pattern of HIV/AIDS disease.

Future challenges

The critical issue remains to reduce the number of new infections, currently over 16 000 per day. More effective linkage of prevention with care and more attention to adolescent groups are priority areas for action. AIDS orphans are a growing and highly emotive problem. In high-prevalence countries strategies for coping with the profound change in the burden of disease are needed urgently. A growing threat is the increasing toll that HIV/AIDS is having on trained health-care personnel, who themselves are having to cope with the rising tide of HIV/AIDS disease. Continued advances in antiretroviral therapy highlight the inequity in access to care and poor survival in most developing countries. It is difficult to see how any of these challenges can be addressed without substantial increases in health care budgets and commitments from Western industrialized countries to reduce the global inequity in HIV/AIDS care. The new Global Fund may have a huge impact, if health systems are improved along with financing drug purchase..

Further reading

Gilks CF *et al.*, eds (1998). *Care and support for people with HIV/AIDS in resource-poor settings.* Health and Population Occasional Paper. Department for International Development, London.

Harries AD, Maher D (1996). *TB/HIV, a clinical manual.* WHO/TB/96.200. World Health Organization, Geneva.

Kaldor JM, ed. (1998). *AIDS in Asia and the Pacific*, 2nd edn. Rapid Science Publishers, London.

Laga M, ed. (1997). *AIDS in Africa*, 2nd edn. Rapid Science Publishers, London.

World Bank Policy Research Report (1997). *Confronting AIDS: public priorities in a global epidemic.* Oxford University Press, New York.

UNAIDS (Joint United Nations Programme on HIV/AIDS) and WHO (World Health Organization) (1999). *Global summary of the HIV/AIDS epidemic*. UNAIDS/99.53E—WHO/CDS/CSR/EDC/99.9. World Health Organization, Geneva.

7.10.23HTLV-I and II and associated diseases

C. R. M. Bangham, M. Osame, and S. Nightingale

HTLV-I

Originally isolated from a patient with a cutaneous lymphoma, the human T-cell lymphotropic virus type I (**HTLV-I**) was the first pathogenic retrovirus to be discovered in humans. In contrast to the human immunodeficiency virus (**HIV**), HTLV-I causes disease in only about 5 per cent of infected people. However, the virus is of special interest and importance because it is associated with two different types of disease: adult T-cell leukaemia/lymphoma (**ATL**) and a range of chronic inflammatory conditions, of which the most commonly diagnosed is HTLV-I-associated myelopathy, also known as tropical spastic paraparesis (**HAM/TSP**). A suggested association with multiple sclerosis has been refuted.

HTLV-I is estimated to infect between 10 and 20 million people worldwide. There are large endemic areas in Central and West Africa, southern Japan, the Caribbean and South America, and smaller foci in the aboriginal populations of Australia, Papua New Guinea, and northern Japan. In Europe and North America the virus is found chiefly in immigrants from these endemic areas and in some communities of intravenous drug users. Within the endemic areas, the distribution of HTLV-I is characteristically uneven; the seroprevalence can vary between 1 and 20 per cent of adults in neighbouring towns.

There are three important modes of transmission: parental and neonatal infection from a seropositive mother, in which breast feeding is a significant factor; sexual transmission, particularly from males to females; and transmission by infected blood, either by transfusion or by sharing of needles among drug users. Transmission of the virus depends on transfer of cells from infected people, because there is little free virus in the serum. Blood for transfusion is now routinely screened for HTLV-I in several countries, including Japan, the United States, and Brazil.

HTLV-I is known as a complex retrovirus: in addition to the three genes present in other typical replication-competent exogenous retroviruses (*gag*, *pol*, and *env*) it encodes at least two further proteins—Tax, which stimulates transcription of the proviral genome, and Rex, which controls the splicing of HTLV-I mRNA. There are closely related leukaemia viruses in monkeys and cattle.

Although it can infect a wide variety of cell types *in vitro*, HTLV-I appears to replicate efficiently only in CD4+ (helper) T cells; these are the cells that are transformed in adult T-cell leukaemia/lymphoma. The cellular receptor for HTLV-I has not been identified, although the gene is known to lie on chromosome 17. Certain cell-surface adhesion molecules (ICAM-1, ICAM-3, and VCAM) also facilitate entry of HTLV-I into the cell.

Antibodies against the Gag protein are the first to appear after infection, and they predominate in the first 2 months. Thereafter, anti-envelope antibodies predominate, and about half of infected individuals subsequently produce antibodies to the Tax protein. Diagnosis of HTLV-I infection depends on the detection of specific antibodies by particle agglutination or enzyme-linked immunosorbent assay (ELISA), and confirmation by polymerase chain reaction (PCR) or western blot assay.

Most people infected with HTLV-I mount a vigorous, chronically activated, cytotoxic T-lymphocyte response to the virus, mostly directed against the viral Tax protein. This cellular response appears to play an important part in reducing the viral burden and the risk of the associated inflammatory diseases such as HAM/TSP.

HTLV-I associated myelopathy (HAM/TSP) and other inflammatory diseases associated with HTLV-I

The association between HTLV-I and HAM/TSP, formerly known as Jamaican neuropathy, was discovered in the Caribbean and in Japan in the mid-1980s. HTLV-I has since shown to be associated with uveitis, and a lymphocytic alveolitis (usually subclinical). There are also less certain associations with chronic infective dermatitis, polymyositis, arthritis, sicca syndrome, and a motor neurone disease-like disorder.

The prevalence of HAM/TSP is between 0.1 and 2 per cent of individuals infected with HTLV-I: about two-thirds of patients are female. Other known risk factors for HAM/TSP include a high proviral load of HTLV-I (Fig. 1) and the *HLA DRB1*0101* gene ('*HLA DR1*'). Possession of the *HLA A*02* gene is associated with a reduced proviral load and a reduced risk of HAM/TSP in Japan.

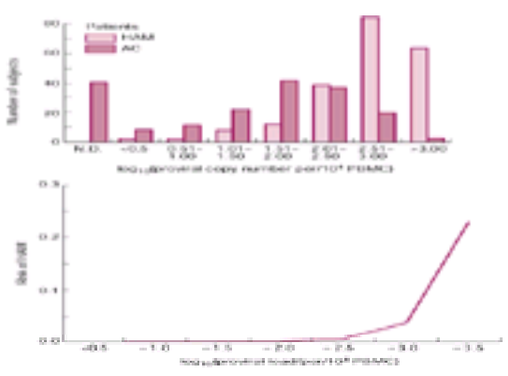

Fig. 1 Upper panel: distribution of HTLV-I provirus load in 202 patients with HAM/TSP (HAM) and 200 asymptomatic HTLV-I carriers (AC) in Japan. The median proviral load in HAM/TSP patients was 16 times greater than asymptomatic carriers. N.D., not detected. Lower panel: the risk of HAM/TSP in people infected with HTLV-I rises rapidly when the proviral load exceeds one copy per 100 peripheral blood mononuclear cells (PBMC). (Adapted from Fig. 2 and Fig. 3 in Bangham *et al*. (1999).)

Clinical features

HAM/TSP is characterized by a spastic paraparesis that is slowly progressive, or in some cases static after initial progression, and anti-HTLV-I antibody positivity in serum and cerebrospinal fluid. Almost all patients show spasticity and/or hyperreflexia of the lower extremities, initially presenting as gait and urinary disturbances. Presentation is commonly with low back pain, weakness of the lower extremities and a poorly defined (mild) sensory affectation, and rarely, with cerebellar ataxia.

Patients with a younger age of onset (less than 15 years) tend to have short stature and slow progression of the disease, while patients with an older age of onset (more than 61 years) show faster progression regardless of the mode of transmission.

Aside from HTLV-I antibody positivity, other essential laboratory findings in the cerebrospinal fluid include lymphocytic pleocytosis and increased neopterin levels. In more than two-thirds of patients, magnetic resonance imaging shows high signals of T_2-weighted spin echo in the white matter of the brain indicating atrophy of the spinal cord.

Autopsy reveals severe involvement of the thoracic spinal cord with mononuclear infiltration, marked myelin and axonal destruction, and astrocytic gliosis.

Several clinical trials have shown transient beneficial effects from corticosteroids, a-interferon, azathioprine, high-dose vitamin C, pentoxifyllene, danazol, and plasmapheresis. The nucleoside analogues zidovudine and lamivudine can reduce the provirus load of HTLV-I, but the clinical benefit is unknown.

Adult T-cell leukaemia/lymphoma (ATL)

HTLV-I infection carries a 5 per cent lifetime risk of ATL; the interval between infection and disease is frequently over 20 years. The disease is slightly commoner in males (1.2:1), and there is evidence of familial clustering of cases. In highly endemic areas it is an important cause of malignant disease: in Kyushu, Japan, ATL accounts for 75 per cent of non-Hodgkin's lymphomas.

The mean age at onset of ATL is about 60 years in Japan, and 40 years in the Caribbean and Brazil; the reason for this difference is not known.

The clinical features of ATL are those of a non-Hodgkin's lymphoma: malaise, fever, lymphadenopathy, hepatosplenomegaly, jaundice, drowsiness, weight loss, and opportunistic infections. Features particularly associated with ATL are skin involvement (nodules, plaques, or a generalized papular rash) and thirst. Laboratory findings include hypercalcaemia and high serum concentrations of lactate dehydrogenase and the soluble interleukin 2 (**IL-2**) receptor. The leukaemic cells are almost invariably CD4+, and are usually CD25+ (IL-2 receptor+). The transformed T cell has a characteristic appearance: the nucleus has several lobules, giving rise to the epithet 'flower cell'. Morphologically similar cells are found in small numbers in the peripheral blood in some asymptomatic carriers of the virus. When the proportion of abnormal cells is high, and there is a lymphocytosis, there is a greatly increased risk of development of ATL. However, in some cases the atypical cells regress spontaneously.

Southern blot analysis indicates the presence of oligoclonal or monoclonal proliferation of CD4+ cells carrying the HTLV-I provirus in their cellular DNA. Typically there is a progression from polyclonal to oligoclonal to monoclonal proliferation in the CD4+ population, accompanied by a progression to increasing IL-2-independence of cellular growth.

ATL is classified into clinical subtypes with different courses and prognoses. Intermediate states between lymphocytosis and frank ATL are called 'smouldering' or 'pre-' ATL. Initially response to standard chemotherapeutic regimes is commonly followed by early relapse with refractoriness to further chemotherapy after 2 to 6 months. Combination of interferon-a and zidovudine can prolong life expectancy by between 6 months and 2 years. The mean survival times (untreated) for acute, lymphomatous, and chronic ('smouldering') ATL in Japan are 6.2, 10.2, and 24.3 months, respectively.

HTLV-II

HTLV-II, a retrovirus closely related to HTLV-I, was first isolated from the tissue of a patient with an atypical form of hairy cell leukaemia. The virus occurs sporadically in West Africa. Among intravenous drug abusers in Europe and North America, infection with HTLV-II is as common as HTLV-I. HTLV-II is common n several native groups throughout the Americas.

HTLV-II is not associated with the typical B-cell form of hairy cell leukaemia, or with other haematological malignancies. A paralytic syndrome similar to HAM/TSP has been reported in individuals seropositive for HTLV-II. Paralysis may be flaccid rather than spastic. However, HTLV-II aetiology is not certain.

Further reading

Bangham CRM *et al*. (1999). Genetic control and dynamics of the cellular immune response to the human T-cell leukaemia virus HTLV-I. *Philosophical Transactions of the Royal Society of London Series B* **354**, 691–700.

Bangham CRM (2000). HTLV-I infections. *Journal of Clinical Pathology* **53**, 581–6.

Fields BN *et al.* (1996). *Fields virology*, 3rd edn. Lippincott-Raven, Philadelphia.

Nakagawa M *et al*. (1995). HTLV-I-associated myelopathy: analysis of 213 patients based on clinical features and laboratory findings. *Journal of Neurovirology* **1**, 50–61.

Uchiyama T (1997). Human T cell leukaemia virus type 1 (HTLV-I) and HTLV-I-associated diseases. *Annual Review of Immunology* **15**, 15–37.

7.10.24Viruses and cancer

R. A. Weiss

Introduction

Viruses are important in cancer for three main reasons: first, as a cause for about 15 per cent of the worldwide cancer burden; second, for the discovery and characterization of oncogenes and tumour suppressor genes; and third, as vectors for the delivery of gene therapy and immunotherapy.

Viruses as aetiological agents of cancer

Table 1 lists the viruses implicated in human cancer. In most but not all cases the viral genome is present in the malignant cells; the exceptions appear to be those that promote cancer indirectly, such as human immunodeficiency virus (**HIV**) and hepatitis C virus (**HCV**).

Oncogenic viruses establish persistent, lifelong infections, so that the event of infection may be far removed from the event of malignancy. Moreover, cancer is often a rare outcome of virus infection, and other cofactors play a part in viral carcinogenesis. For example, Epstein–Barr virus (**EBV**) is a ubiquitous infection yet children's Burkitt's lymphoma occurs only in areas of holoendemic malarial infection, whereas undifferentiated nasopharyngeal carcinoma occurs mainly in southern Chinese populations. Aflatoxin may act with hepatitis B virus (**HBV**) to cause liver cancer, and in hereditary epidermodysplasia verruciformis the ultraviolet radiation acts with human papillomavirus (**HPV**) strains to cause skin cancer.

Kaposi's sarcoma is a tumour that occurs much more frequently in immunodeficient patients. Its relative risk in recipients of organ transplants is about 400, and in persons with AIDS about 20 000. HIV probably contributes to Kaposi's sarcoma indirectly through immune suppression, although the Tat protein of HIV may also play a role. The primary cause of all forms of Kaposi's sarcoma is the recently discovered human herpes virus 8 (**HHV-8** or KSHV). This virus is also causally linked to primary effusion lymphoma and plasmablastic multicentric Castleman's disease.

Oncogenic viruses belong to many virus families, which have different routes of transmission. Some, like hepatitis B virus, are frequently acquired perinatally or through subsequent exposure to blood. With human T-cell lymphotropic virus type I (**HTLV-I**) the main route of transmission is vertical through infected cells in breast milk. Sexual transmission is common to HIV, HTLV-I (with a male to female bias), HBV, and HPV. Oncogenic viruses do not appear to be transmitted by the respiratory route, except adenoviruses, or via arthropod vectors, except some veterinary cases, such as bovine leucosis virus. Whereas EBV (transmitted through saliva) occurs worldwide, HBV, HTLV-I, and HHV-8 have a high prevalence mainly in those population groups in which the associated cancers occur.

Certain common human viruses are highly oncogenic in experimental animals but are not linked to human cancer, namely the polyomaviruses BK and JC, and the adenoviruses. Human adenovirus types 2 and 12 readily cause sarcomas and carcinomas in hamsters and other rodents. The viral genomes persist non-productively in the animal tumours and express early genes. It is surprising, then, that there is no epidemiological evidence linking adenovirus or BK infection with human cancer. There is some concern that a simian relative of BK virus, SV40, may be linked with mesothelioma, osteosarcoma, and ependymoma in humans, but these findings remain controversial.

Mechanisms of viral carcinogenesis

Physical and chemical carcinogens are usually mutagens. They cause DNA mutations in specific genes that contribute to the eventual malignant phenotype of the cancer. Oncogenes were first discovered in animal retroviruses, such as the Rous sarcoma virus of chickens, and are now known to originate from cellular genes. Most retroviruses do not carry oncogenes but the DNA provirus integrates into chromosomal DNA and can activate adjacent cellular oncogenes. Oncogene activation by retroviruses is comparable with activation by chromosomal translocation.

The mechanism of cell transformation by HTLV-I is different from that of the majority of animal retroviruses. HTLV-I encodes a viral protein, Tax, which is essential to promote full viral gene transcription. Tax acts as a transcriptional activator by associating with host nuclear proteins which activate expression of the viral genome. However, Tax also up-regulates certain cellular genes such as the interleukin-2 receptor. HTLV-I 'immortalizes' CD4+ T lymphocytes in culture, rather as EBV immortalizes B lymphocytes, but this is only one step in the pathway to malignancy. HTLV-I leukaemia does not become manifest until 40 or more years after infection, and in only 5 per cent of infected people.

Cell transformation by DNA viruses is best understood for polyomaviruses and adenoviruses. The transforming genes of these viruses are expressed early in the infection cycle and prevent tumour suppressor protein function. Adenovirus proteins E1A and E1B and BK T-antigen bind to p53 and Rb and block their normal interaction in the cell cycle. Thus instead of mutating these cellular tumour suppressor genes, the DNA tumour viruses block the normal function of their proteins, which similarly results in unregulated cell proliferation. The HHV-8 genome carries several oncogenes including a homologue of cyclin D2, which inactivates Rb by a different mechanism, phosphorylation.

To cause tumours, most oncogenic viruses persist in the tumour cells, often by integrating into chromosomal DNA. Oncogenic herpesviruses do not integrate but are maintained episomally. EBNA-1 is required for episomal replication of EBV (and LANA for HHV-8), while other nuclear and latent membrane proteins are responsible for the transformed cell phenotype. With HBV, integrated copies are found in many liver carcinoma lines, but a requirement for integration has not been unequivocally shown. HBV expresses transactivating functions from the *X* gene so its transformation may resemble that of HTLV-I. Some viruses might exert an oncogenic effect without persisting in the cells destined to become the malignant clone, by causing mutations in host DNA, thus acting in a 'hit-and-run' manner like other mutagens.

Indirect carcinogenic effects are those in which damage to tissues by viruses may allow clones of premalignant cells to proliferate that would not otherwise do so. HCV and possibly HBV might do this by destroying normal liver cells, resulting in a much greater rate of liver cell regeneration. HIV could be regarded as a special case of indirect viral carcinogenesis, promoting tumour development by destroying helper T-cell immunity to other viruses. The cancers elevated in AIDS are also seen in immunosuppressed transplant recipients, for example non-Hodgkin's lymphoma and Kaposi's sarcoma, and themselves have a viral aetiology.

Treatment and prevention

Oncogenesis is multifactorial, requiring several sequential events before a patient presents with a fully malignant tumour. Yet if a virus plays a crucial role in oncogenesis, its elimination should prevent that type of cancer. Currently, there is no special approach to the treatment of cancers that have a viral aetiology. Among the lymphoid malignancies, some respond well to radiotherapy or chemotherapy, such as Hodgkin's disease, whereas others seldom show remission, such as adult T-cell leukaemia. Cancers that express viral antigens should be responsive to immunotherapy. If immunosuppression promotes their presentation, they should be susceptible to immune attack. For tumours in which viral proteins are required for the maintenance of the malignant state, those proteins are potential molecular targets, as drugs that block them might spare normal cellular functions.

Prevention is preferable to cure and offers the greatest promise of reducing cancer mortality due to viruses. Prevention can be accomplished by three strategies: (i) early screening for tumours, (ii) screening for the virus with prevention of transmission, and (iii) immunization. Early screening is exemplified by cervical smears and, in China, for elevated serum IgA levels to EBV antigens for incipient nasopharyngeal carcinoma. Screening to prevent iatrogenic transmission via blood and blood products is routinely employed in many countries for potentially oncogenic viruses such as HBV, HCV, HIV, and HTLV-I. In Kyushu, Japan, where infection was

endemic, HTLV-I is being steadily eradicated through a policy of antenatal screening to prevent milk transmission.

Prevention of cancer by immunization against oncogenic viruses is likely to have a major impact on world cancer mortality in the twenty-first century. Currently, the only proven, mass-produced vaccine against a human oncogenic virus is the HBV vaccine based on surface antigen. Indeed, it is the first efficacious recombinant subunit vaccine against any virus. Other vaccines under development that are likely to be successful within the next decade are for EBV, HPV, and HTLV-I. Intensive research is also being undertaken on vaccines for HIV and HCV. There are likely to be some obstacles on the route to successful immunization, as HIV is extraordinarily variable, and even the relatively stable HBV shows evidence of immune-escape mutants in the face of vaccination programmes. Nevertheless, immunization against oncogenic viruses is likely to become a most effective cancer prevention strategy.

Viruses as therapeutic agents

Viruses may be put to use in the fight against cancer. First, some cytopathic viruses preferentially replicate in proliferating cells and destroy them, such as parvoviruses and mutant adenoviruses. Second, viruses as foreign antigens may aid the recognition of cancer cells by the host's immune system. Although the mechanism is ill understood, 'xenogenization' of tumour cells by virus infection can, in some cases, enhance immune control of non-infected cells of the same tumour. Third, viruses are favoured vectors for immunization and for gene therapy, by restoring tumour suppressor functions, by enhancing immune responses through the expression of antigens or cytokines, and by locally delivering genes for enzymes that convert inert prodrugs into active, chemotherapeutic agents.

Further reading

Arrand JR, Harper DR, eds (1998). *Viruses and human cancer*. BIOS, Oxford.

Boshoff CH, Weiss RA, eds (1999). Human herpes virus 8. *Seminars in Cancer Biology* **9**.

Dalgleish AG, Weiss RA, eds (1999). *HIV and the new viruses*. Academic Press, London.

Goedert JJ, ed. (2000). *Infectious causes of cancer. Targets for intervention*. Humana Press, Paterson, New Jersey.

Newton R, Beral V, Weiss RA, eds (1999). Infections and human cancer. *Cancer Surveys* **33**.

7.10.25 Orf

N. Jones

Aetiology

Orf virus, a member of the *Parapox* genus, normally causes ecthyma contagiosum or 'scabby mouth' (contagious pustular dermatitis) in sheep and goats. Orf virions are ovoid (approximate size, 260 × 160 nm), with tubular threadlike structures criss-crossing the surface of the virions, visible by negatively stained electron microscopy. The orf virus genome is double stranded DNA of 135 kbp. The viral genome encodes a polypeptide homologous to IL-10, inhibitors of interferon, IL-2 and GM-CSF, and a vascular endothelial growth factor. These contribute to the dermal lesions characterized by capillary proliferation and dilatation.

Other Parapox viruses infect cattle, deer, and seals.

Epidemiology

The disease affects mainly young lambs, who contract the infection from one to another, or possibly from persistence of the virus in the pastures (the virus can remain viable for long periods in dried scabs from lesions). Human disease is usually occupational, following contact with infected sheep. It is not uncommon in shepherds, veterinary surgeons, and farmers (Fig. 1). One attack normally confers immunity and human to human spread has not been recorded.

Fig. 1 Orf, typical lesions on a farmer's hand.

Clinical features

In sheep, papules and vesicles appear on the lips (Fig. 2) and gradually heal with no scarring over 4 weeks. In humans, after an incubation period of 5 to 6 days, a small, red, firm papule enlarges to form a flat-topped haemorrhagic pustule or bulla; the centre may be crusted. The lesion is usually 2 to 3 cm in diameter, but may be as large as 5 cm. Lesions are solitary or few in number and commonly occur on the hands and forearms, occasionally the face. Lymphangitis or regional lymphadenopathy are not uncommon. Slight fever and malaise can occur. Recovery is usually complete in 6 weeks and is spontaneous.

Fig. 2 Contagious pustular dermatitis ('orf') in a lamb.

Large fungating granuloma or tumour-like lesions have been reported, especially in association with haematological malignancy.

Erythema multiforme occasionally develops, typically 10 to14 days after the onset of orf (Fig. 3). Rarely, bullous pemphigoid has been reported in association with orf.

Fig. 3 Erythema multiforme complicating orf of the left middle finger in a veterinary student.

Diagnosis

The characteristic lesion in a person exposed to sheep and lambs allows a clinical diagnosis. This can be confirmed in the laboratory by electron microscopy of a biopsy of the orf lesion or by PCR. The virus can also be isolated in cell culture, but this is rarely performed. Histopathological examination of biopsy specimens shows a proliferation of keratinocytes with cellular swelling and balloon degeneration and B type cytoplasmic inclusion bodies.

Treatment

Secondary infection should be treated if it occurs. Large lesions can be removed surgically, but recurrence can occur in the immunocompromised. Cidofovir cream has been used successfuly to treat giant or persistent lesions in immunosuppressed patients.

Further reading

Gill, M.J., Arlette, J., Buchan, K.A., *et al*. (1990). Human orf. *Archives of Dermatology*, **126**, 356–8.

Groves, R.W., Wilson-Jones, E., and MacDonald, D.M. (1991). Human orf and milkers nodule: a clinicopathological study. *Journal of the American Academy of Dermatology*, **25**, 706–11.

Imlach W, *et al.* (2002). Orf virus-encoded interleukin-10 stimulates the proliferation of murine mast cells and inhibits cytokine synthesis in murine peritoneal macrophages. *Journal of General Virology* **83**, 1049–58.

Torfason EG, Gunadottir S (2002). Polymerase chain reaction for laboratory diagnosis of orf virus infections. *Journal of Clinical Virology* **24**, 29–84.

7.10.26 Molluscum contagiosum

N. Jones

Molluscum contagiosum is a benign skin tumour, caused by a poxvirus, which mostly affects children and young adults. It is exclusively a human disease.

Aetiology

Molluscum contagiosum virus (MCV), a large double-stranded DNA virus, is a member of the Poxviridae, genus *Molluscipox*. MCV has not been transmitted to laboratory animals and there is no *in vitro* cultivation system currently available. Restriction endonuclease analysis of the genome has identified three types, MCV I, MCV II, and MCV III. The majority of infections seem to be due to MCV 1. The virus encodes MC 159L protein which causes abnormal proliferation of epithelium by inhibiting TNF and apoptosis-inducing factors.

Epidemiology

Molluscum contagiosum is common and disease usually follows contact with an infected individual or contaminated object. In tropical countries, the infection tends to occur in younger children (1–4 years) than in temperate climates (10–12 years). Sexual transmission accounts for a second incidence peak in young adults. Unusually widespread lesions have been reported in HIV disease, sarcoidosis, and those taking immunosuppressive therapy.

Clinical features

The incubation varies from 14 days to 6 months. The individual lesion is a shining, pearly, hemispherical, firm, umbilicated papule with a central depression. Lesions can occur singly (Fig. 1) but are commonly multiple (Fig. 2). The lesions gradually grow to a diameter of 5 to 10 mm over 6 to 12 weeks. Occasionally one very large lesion can develop (> 10 mm), or a plaque of very small lesions (agimate form). Most cases persist for 6 to 9 months, occasionally as long as 5 years, following which spontaneous resolution occurs.

Fig. 1 Single lesion on eyelid.

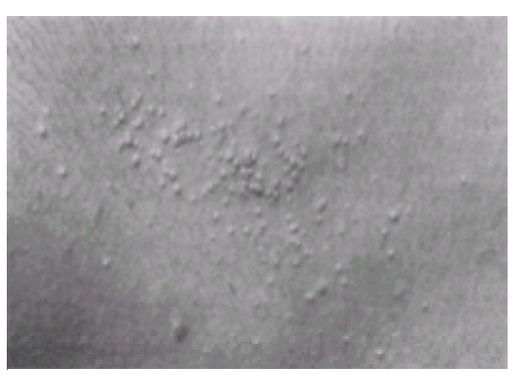

Fig. 2 Molluscum contagiosum: groups of papules characterized by a central punctum.

Lesions are commonly seen on the neck, trunk, or axilla, although any part of the skin can be affected. Lesions are rare on the palms, soles, and mucous membranes. Sexually acquired infections normally result in anogenital lesions. In about 10 per cent of cases, especially where there is a history of atopy, a patchy dermatitis develops around the lesions. In the HIV patient molluscum can be widespread, but particularly involves the face, neck, and around and inside the mouth in male homosexuals. Lesions may become large and atypical, and are mistaken for basal cell carcinomas or other skin tumours. The disease is often unremitting with increasing severity, especially when HIV is advanced.

Diagnosis

The diagnosis is usually clinical, but histological and electron microscopic examination of a curetted lesion establishes the diagnosis. The differential diagnosis can include lepromatous leprosy and, in HIV-seropositive patients, disseminated cutaneous histoplasmosis or cryptococcosis.

Treatment (see also Chapter 23.1)

Advice on prevention of spread of the infection to others should be given, such as avoidance of swimming pools, contact sports, or shared towels, until the lesions have resolved. Treatment may not be necessary, and depends on the site and number of the lesions and the age of the patient.

Cryotherapy (with liquid nitrogen) is effective and should be repeated at 3–4-weekly intervals. Other techniques include diathermy or curettage. In children the application of local anaesthetic cream prior to the procedure may be necessary.

Topical agents such as phenol (10–20 per cent solution), salicylic acid (15–20 per cent), silver nitrate, trichloroacetic acid, lactic acid, tretinoin, and cantharidin are

used. The agent can be delivered to the inside of the lesion using the sharpened end of a wooden applicator stick.

In severe cases associated with HIV, 5 per cent imiquimod cream has proved effective. Recently, the antiviral agent, cidofovir (intravenously or topically), has been successfully used to treat molluscum contagiosum.

Further reading

Birthistle, K. and Carrington, D. (1997). Molluscum contagiosum virus. *Journal of Infection*, **34**, 21–8.

Garvey TL, *et al*.(2002). Binding of FADD and Caspase-8 to *Molluscum contagiosum* virus MC 159v-FLIP is not sufficient for its antiapoptotic function. *Journal of Virology* **76**, 697–706.

Husar K, Skerlev M (2002). *Molluscum contagiosum* from infancy to maturity. *Clinics in Dermatology* **20**, 170–2.

Meadows, K.P., Tyring, S.K., Pavia, A.T., and Rallis, T.M. (1997). Resolution of recalcitrant molluscum contagiosum virus lesions in human immunodeficiency virus infected patients treated with cidofovir. *Archives of Dermatology*, **133** (8), 987–90.

Schwartz, J.J. and Myskowski, P.L. (1992). Molluscum contagiosum in patients with human immunodeficiency virus infection. A review of twenty seven patients. *Journal of the American Academy of Dermatology*, **27**, 583–8.

7.11.1 Diphtheria

Delia B. Bethell and Tran Tinh Hien

Introduction

Diphtheria is an acute infection of the upper respiratory tract, and occasionally of other mucous membranes or skin, usually caused by *Corynebacterium diphtheriae*. The disease was probably known to the Greeks and Romans. While virtually eliminated from most developed countries by mass immunization, diphtheria remains a threat in poorer countries. In the last decade there has been a resurgence in parts of the former Soviet Union.

Bacteriology

Morphology

C. diphtheriae are pleomorphic, Gram-positive rods or clubs. Adjacent cells lie at different angles to each other—'Chinese letters'. The presence of metachromatic granules, usually two or three per cell, is characteristic but not exclusive; these stain deep blue with Neisser's methylene blue or greyish-black with Albert's stain. Virulent and non-virulent *C. diphtheriae* cannot be distinguished by their morphological appearances.

Culture

C. diphtheriae does not grow well on ordinary agar, but prefers media containing blood or serum. Selective media, such as Loeffler's serum medium and blood tellurite agar, are necessary for its isolation from other respiratory flora, although other corynebacteria can also grow on these media. Biochemical tests are used to identify *C. diphtheriae*. There are three colonial types of *C. diphtheriae*: *gravis*, *intermedius*, and *mitis*. There is no good association between colonial appearance and disease severity.

Toxin production

The clinical manifestations of diphtheria are caused by an exotoxin produced by virulent corynebacteria. The structural gene of the toxin, *tox*, is carried by a lysogenic corynebacteriophage. The phage can pass from toxigenic to non-toxigenic strains; this may be important in outbreaks when harmless strains in carriers' throats may become toxigenic. TOX gene expression is regulated by the *C. diphtheria*-encoded, iron-activated, repressor DtxR; hence iron starvation leads to increased toxin production.

Production of toxin may be assessed using gel precipitation (Elek's test) or guinea-pig inoculation. More recently enzyme immunoassays have been developed.

Pathogenesis

Pseudomembrane formation

This results from an inflammatory reaction to the presence of multiplying toxigenic *C. diphtheriae*. Fluid and leucocytes move from dilated blood vessels to surround necrotic epithelial cells. The fluid clots to enmesh these dead cells, as well as leucocytes, diphtheria bacilli, cellular debris, and occasionally small blood vessels. The latter explains why the pseudomembrane is adherent to underlying tissues and often bleeds when it is pulled away.

Action of toxin

C. diphtheriae does not usually pass beyond the pseudomembrane site; it is the toxin that causes the severe complications of diphtheria. Diphtheria toxin is a 535-residue, 62-kDa exotoxin. It consists of two factors: spreading factor B attaches via its receptor (heparin-binding, epidermal growth factor-like precursor (**HB-EGF-LP**)) to the cell membrane allowing lethal factor A to enter the cell. Factor A catalyses the NAD+-dependent ADP-ribosylation of eukaryotic elongation factor 2, preventing protein synthesis. Locally the toxin causes tissue necrosis, leading to formation of the typical pseudomembrane and, when absorbed into the bloodstream, systemic complications. Diphtheria toxin affects all human cells, but the most profound effects are seen in the myocardium, peripheral nerves, and kidneys. Delivery of a single molecule of factor A to the cytosol of a eukaryotic cell will kill it.

Effects on the heart

Common changes are fatty degeneration of cardiac muscle (myocarditis) and infiltration of the interstitium with leucocytes, which may affect the conduction fibres. Parenchymal necrosis is rare. Generally the heart can recover completely from these effects, although severe fibrosis and scarring may lead to death in late convalescence. Mural endocarditis may cause embolism leading to cerebral infarction and hemiplegia. Valvular endocarditis is extremely uncommon. Neuritic changes may be seen in the nerves to the heart during the late paralytic stage of the disease.

Effects on nerves

Diphtheria toxin causes demyelination and degeneration of both sensory and motor nerves. It affects the nerves to the eye, palate, pharynx, larynx, heart, and limb

muscles. It is unclear whether the toxin can cross the blood–brain barrier and cause central lesions.

Effects on other organs

Non-specific changes in the kidneys, adrenals, liver, and spleen may be seen.

Epidemiology

Man is the only known reservoir for *C. diphtheriae*. Spread is usually via respiratory droplets or direct contact with respiratory secretions or exudate from skin lesions. Cutaneous diphtheria is more contagious than respiratory diphtheria: skin infections are the main reservoir of *C. diphtheriae* in environments of poverty, overcrowding, poor hygiene, frequent and slowly healing traumatization of unprotected skin, and insect bites. Fomites and dust are not important means of transmission, although *C. diphtheriae* may resist drying and has been isolated from dust on the floor of a ward. Diphtheria has been spread by contaminated milk. *C. diphtheriae* is killed y pasteurization and by most common disinfectants. Patients may become carriers of the infection and continue to harbour the organism for weeks or months, or even for a lifetime.

There is no given level of circulating antibody indicating protection or susceptibility to infection. The Schick test is used to assess the antibody response to diphtheria toxin. A measured amount of toxin is injected into the forearm causing a red reaction (positive) unless the patient has a sufficient antibody response to prevent it (negative). A Schick-negative person is very unlikely to have clinically significant diphtheria, while a Schick-positive person may have an attack of any severity. Neonates are very often Schick-negative, protected by maternal antibody, but become Schick-positive around 6 months of age. *C. diphtheriae* tends to die out in a highly immunized population, and children may grow to adult life without encountering the bacillus. In areas of the world that lack an effective immunization programme children generally meet *C. diphtheriae* early, maybe becoming a faucial, nasal, or aural carrier, and young children may suffer severe or fatal attacks of diphtheria.

In the 1990s a diphtheria epidemic gripped parts of the former Soviet Union. Economic hardship, crowding due to large urban migration, low vaccination coverage, and poor primary vaccination practices due to failing health systems have contributed. This has led to large numbers of susceptible children as well as an increase in susceptible adults as immunity was not maintained by periodic boosters. Prior to the vaccine era, most people acquired natural lifelong immunity during childhood through their exposure to *C. diphtheriae.* Serological studies in several countries indicate that 20 to 50 per cent of adults over the age of 20 years are susceptible to diphtheria, with a significant trend of decreasing immunity with increasing age. This potential risk assumes a particular significance in today's international travel.

Clinical features

After an incubation period of 2 to 5 days, diphtheria presents in a variety of forms depending upon the location of the pseudomembrane—anterior nasal, faucial, tracheolaryngeal, malignant, and cutaneous.

Anterior nasal

This is usually unilateral and relatively mild unless it coexists with other forms. It is relatively common in infancy. There is a nasal discharge, initially watery, then purulent and blood-stained. The nostril may be sore or crusted and a thin pseudomembrane can sometimes be seen within the nostril itself.

Faucial

This is the commonest form of diphtheria. Malaise, sore throat, and moderate fever develop gradually. At the onset of symptoms only a small, yellow-grey spot of pseudomembrane may be present on one or both tonsils, easily mistaken for other types of tonsillitis. The surrounding areas are dull and inflamed. Over the next few days the pseudomembrane enlarges and may extend to cover the uvula, soft palate, oropharynx, nasopharynx, or larynx. There is tender cervical lymphadenopathy, nausea, vomiting, and painful dysphagia. The pseudomembrane becomes greenish-black and eventually sloughs off.

Tracheolaryngeal

Some 85 per cent of tracheolaryngeal presentations are secondary to faucial diphtheria, but occasionally there may be no pharyngeal pseudomembrane. Initial symptoms include moderate fever, hoarseness, and a non-productive cough. Over the next day, as the pseudomembrane and associated oedema spread, the child becomes increasingly dyspnoeic with severe chest recession and cyanosis and asphyxiation unless the obstruction is relieved. Tracheostomy brings instant relief if the obstruction is confined to the larynx and upper trachea. In a minority of cases the pseudomembrane also involves the bronchi and bronchioles and tracheostomy has little effect.

Malignant

The onset is rapid, with high fever, tachycardia, hypotension, and cyanosis. Pseudomembrane spreads from the tonsils to cover much of the nasopharynx. It has a thick edge and as this advances the earlier parts become necrotic and foul smelling. There is gross cervical lymphadenopathy. Individual lymph nodes are difficult to feel because of surrounding oedema; this is the characteristic 'bull neck' of malignant diphtheria. The patient may bleed from the mouth, nose, or skin. Cardiac involvement with heart block occurs within a few days. Acute renal failure may ensue. Survival is unlikely.

Cutaneous

In contrast to many faucial infections, cutaneous diphtheria is usually chronic but mild. The morphological features of individual lesions can be extremely variable as *C. diphtheriae* can colonize any pre-existing skin lesion (such as impetigo, scabies, surgical wounds, or insect bites) without altering their picture. However, the ulcerative form is the most frequent and typical. Initially vesicular or pustular, filled with straw-coloured fluid, it soon breaks down to leave a punched-out ulcer several millimetres to a few centimetres across. Common sites are the lower legs, feet, and hands. During the first 1 to 2 weeks it is painful and may be covered with a dark pseudomembrane. After this separates a haemorrhagic base is seen, sometimes with a serous or serosanguinous exudate. The surrounding tissue is oedematous and pink or purple in colour. Spontaneous healing to leave a depressed scar usually takes 2 to 3 months, sometimes much longer.

Systemic complications, such as myocarditis, are rare. Occasionally, the affected limb becomes paralysed.

Other sites

A mild conjunctivitis may accompany faucial diphtheria. Occasionally, pseudomembrane forms in the lower conjunctiva and spreads over the cornea causing considerable damage. Dysphagia may indicate that pseudomembrane has spread from the tonsils to the oesophagus. Other parts of the gastrointestinal tract are not usually affected, but melaena with colicky abdominal pain is described. Diphtheria may spread by fingers from the throat to vulva or penis causing localized sores. *C. diphtheriae* occasionally invades the vagina and cervix, allowing the absorption of toxin. In one patient, pseudomembrane was found on the wall of the bladder at operation; peripheral neuritis and fatal heart failure ensued. Endocarditis is rare, but at least one reported case recovered following antimicrobial treatment.

Other corynebacteria

C. ulcerans produces two toxins, one of which seems to be the same as diphtheria toxin. It may cause membranous tonsillitis but toxic manifestations are rare. However, at least one fatality due to *C. ulcerans* has been reported. *C. ulcerans* has been spread to humans in cows' milk.

C. pseudodiphtheriticum is commonly present in the flora of the upper respiratory tract. It is non-toxigenic, but can cause exudative pharyngitis with a pseudomembrane identical to that produced by *C. diphtheriae*. More commonly it causes endocarditis in patients with anatomical abnormalities or infections of the lungs, trachea, or bronchi in immunosuppressed patients or those with pre-existing respiratory disease.

C. xerosis has been isolated from the blood of patients with endocarditis and from prosthetic valves at operation. *C. haemolyticum* has caused outbreaks of tonsillitis

with or without a maculopapular rash.

Complications

Patients surviving acute diphtheria may develop one or more complication. These result from delayed effects of the toxin following haematogenous spread. The risk and severity of complications correlates directly with the extent of the pseudomembrane and the delay in administration of antitoxin.

Cardiovascular

Approximately 10 per cent of patients with diphtheria will develop myocarditis. Some two-thirds of patients with severe infection will have some evidence of cardiac involvement. The frequency of cardiac involvement in laryngeal and malignant diphtheria is three- to eightfold higher compared with faucial diphtheria, and two- to threefold higher if antitoxin is given more than 48 h after onset of the disease.

Cardiac toxicity usually appears after the first week of illness, but in malignant forms can occur after just a few days. Patients complain of upper abdominal pain and may vomit. They become very lethargic and tired. Examination reveals a rapid thready pulse with hypotension. At this stage profound shock may lead to death. In less severe cases, congestive cardiac failure may develop, with a displaced apex beat, gallop rhythm, and murmurs audible over all areas of the heart. Profound bradycardia may result from heart block. The liver enlarges and oliguria develops. Most deaths from diphtheria occur at this stage. If the patient survives myocarditis, complete recovery is likely.

Electrocardiography (**ECG**) is the best way to demonstrate cardiac involvement. The most common abnormalities are T-wave inversion in one or more chest leads and prolonged QTc intervals. There may be right or left axis deviation, bundle-branch block, or heart block. Very occasionally, atrial fibrillation or tachyarrhythmias are seen. Many more bursts of arrhythmias can be demonstrated if 24-hour ECG monitoring is performed. Numerous ectopic beats have been recorded in patients who lacked other manifestations of cardiac involvement.

Neurological

Neurological complications usually appear weeks after the onset of the disease, when the patient appears to be recovering. Palatal paralysis is common and may be seen from the third week onwards. The patient develops a nasal voice and regurgitates fluids through the nose. This usually resolves within a week or so. A little later there may be blurred vision from paralysis of accommodation, or a transient squint from external rectus paralysis. About the sixth or seventh week more sinister paralyses may develop affecting muscles to the pharynx, larynx, chest and limbs. The nerves to the heart may be affected causing tachycardia and dysrhythmias. Patients may become profoundly hypotonic over a few hours and die from respiratory arrest. However, if intensive-care facilities and skilled staff are available, the patient should be able to make a complete recovery over the following weeks or months.

Diagnosis

In areas where diphtheria is relatively common it should be suspected in any child with exudate in the throat. If the exudate is thick and discoloured the child should be given antitoxin. Clinical diagnosis is much more difficult where diphtheria is rare. The differential diagnosis includes infectious mononucleosis, streptococcal or viral tonsillitis, peritonsillar abscess, oral thrush, and leukaemia and other blood dyscrasias. The bull-neck of malignant diphtheria may be mistaken for mumps. In adults, secondary syphilis can sometimes cause a glairy (resembling egg-white) exudate on the tonsils, and may be accompanied by rash and laryngitis.

Direct smears of infected areas of the throat are often used for diagnostic purposes, but are only of value in experienced hands. Confirmation of the diagnosis depends on culture and identification of *C. diphtheriae* from infected sites. Atypical corynebacteria can be classified only in a reference laboratory.

Treatment

Antitoxin is the mainstay of treatment, but to be effective it must be given before the toxin has reached tissues such as the heart and kidneys, preferably within 48 h of the onset of symptoms. This means that it must be given before bacteriological confirmation. Dosage depends on the site of primary infection, the extent of pseudomembrane, and the delay between the onset of symptoms and antitoxin administration. Between 20 000 and 40 000 units are given for faucial diphtheria of less than 48 h duration or for cutaneous infection; 40 000 to 80 000 units for faucial in excess of 48 h or for laryngeal infection; 80 000 to 100 000 units for malignant diphtheria. For doses over 40 000 units a portion is given intramuscularly followed by the bulk of the dose intravenously after an interval of 30 min to 2 h. Anaphylaxis can occur following antitoxin administration, and adrenaline (epinephrine) should always be available.

Antibiotics are given to eradicate the organism and prevent further toxin production. Benzylpenicillin (penicillin G) 150 000 to 250 000 units/kg per day (90–150 mg/kg/day) is given intravenously in four to six divided doses in children aged 1 month to 12 years. In adults the dose of benzylpenicillin is 12 million to 20 million units/day (7.2–12 g/day) in four to six divided doses. Oral penicillin V is substituted when the patient is able to swallow. Erythromycin may be used for penicillin-sensitive individuals, but a recent study suggests it may not be as effective in eradicating carriage. Antibiotic therapy should continue for 10 to 14 days.

Facilities for urgent tracheostomy should always be available in case of respiratory obstruction. Indications include increasingly laboured breathing and agitation. This procedure will be life-saving in many cases. Most tracheostomies can be closed after just a few days. Steroids may be used in conjunction with tracheostomy to reduce airway swelling, but there have been no controlled trials to support their use. Steroids are of no benefit in preventing myocarditis or neuritis.

Patients with signs or symptoms of cardiac involvement need to be managed in intensive-care units. Oxygen should be given. Temporary cardiac pacing is useful in patients with heart block, but is of doubtful value in cases of malignant diphtheria. An isoprenaline infusion may buy valuable time while the patient is transferred to a centre with facilities for pacing. Digoxin has been used in congestive cardiac failure. It has been suggested that carnitine may prevent some cases of myocarditis.

There is no specific treatment for neuritis. The severest cases will need mechanical ventilation and intragastric or intravenous feeding. With skilled nursing care full recovery can be expected.

Prevention

Diphtheria is a devastating but preventable disease. Its resurgence in Eastern Europe has highlighted the importance of vaccination. Experience to date suggests that a large gap in the immunity of adults poses an outbreak risk, but is probably not sufficient to sustain a large diphtheria epidemic. However, an immunity gap in adults coupled with the presence of large numbers of susceptible children and adolescents creates the potential for an extensive epidemic. Population migration may lead to massive introduction and spread of toxigenic strains of *C. diphtheriae*.

In industrialized countries, infants, children, and adolescents can be effectively immunized using a six-dose schedule: three primary doses of **DTP** (adsorbed diphtheria–tetanus–pertussis) are given in infancy (in the United Kingdom at 2, 3, and 4 months); a first booster dose with DTP vaccine at the end of the second year; a second booster dose with **DT** (adsorbed diphtheria-tetanus) (or DTP) at school entry; and a third booster dose with **Td** (adsorbed tetanus/low-dose diphtheria for adults) at school leaving. Protection against diphtheria may be inadequate if only a single booster of TD or Td vaccine is given at 4 to 10 years of age following the primary doses.

In developing countries, the immunization of infants with a primary series of three doses of DTP was introduced in the late 1970s. By 1995 the coverage of infants was 81 per cent. Where diphtheria is endemic this should be sufficient to prevent an epidemic of diphtheria, as natural mechanisms such as frequent skin infections caused by *C. diphtheriae* probably contribute to maintaining immunity. One or two DT or DTP booster doses may need to be added to the routine schedule in areas at increased risk of diphtheria.

Reduction in the *C. diphtheriae* reservoir due to the large-scale immunization of children means that adults in industrialized countries are no longer immune through natural exposure. Repeated doses of diphtheria toxoid are needed to maintain immunity in the adult population. A lower dose of toxoid is used in older children and adults because of a tendency for more severe adverse effects. Some industrialized countries schedule routine booster doses of Td for every 10 years, but this

strategy is difficult to monitor. Adults in developing countries do not require routine immunization.

Aggressive action is needed in the event of a diphtheria outbreak. Groups at risk should be immunized, there should be prompt diagnosis and management of cases, and identification of close contacts should be made so that the spread of infection can be halted. A single dose of DTP should be used for children under 3 years of age, DT for children aged 3 to 7 years, and Td vaccine for all persons aged over 7 years. Additional doses of vaccine will be needed in non-immunized individuals.

Further reading

Bonnet JM, Begg NT (1999). Control of diphtheria: guidance for consultants in communicable disease control. *Communicable Disease and Public Health* **2**, 242–9.

Eskola J, Lumio J, Vuopio-Varkila J (1998). Resurgent diphtheria—are we safe? *British Medical Bulletin* **54**, 635–45.

Galazka AM, Robertson SE (1996). Immunization against diphtheria with special emphasis on immunization of adults. *Vaccine* **14**, 845–57.

Hofler W (1991). Cutaneous diphtheria. *International Journal of Dermatology* **30**, 845–7.

Public Health Laboratory Service website. www.phls.co.uk/facts/diphtheria/dip.htm [Information on UK notifications and vaccine uptake.]

Rakhmanova G, *et al*. (1996). Diphtheria outbreak in St. Petersburg: clinical characteristics of 1,860 adult patients. *Scandinavian Journal of Infectious Diseases* **28**, 37–40.

Vitek CR, Wharton M (1998). Diphtheria in the former Soviet Union: reemergence of a pandemic disease. *Emerging Infectious Diseases* **4**, 539–50.

WHO (1998). Diphtheria. *Bulletin of the World Health Organization* **78**(Suppl 2), 129–30. [A concise summary of the global problem.]

7.11.2 Streptococci and enterococci

S. J. Eykyn

The term *Streptococcus* was first used by Billroth in 1874 to describe chain-forming cocci seen in infected wounds. They were also seen in the blood in puerperal sepsis by Pasteur (1879). In 1884, Rosenbach defined these streptococci as *Streptococcus pyogenes.* This organism remains one of the most important human pathogens. The genus *Streptococcus* contains numerous other species of varying degrees of pathogenicity for humans and animals. *Streptococcus faecalis* and *S. faecium* were split from the genus *Streptococcus* in 1984 and became *Enterococcus* spp. and numerous other species have since been included in this genus. The nutritionally-exacting streptococci *S. adjacens* and *S. defectivus* have also been assigned to a new genus, *Abiotrophia,* to which the newly described species *A. elegans* has been added.

Classification

Traditionally, classification of streptococci has relied on serological reactions, particularly Lancefield grouping based on cell wall carbohydrates, and haemolytic activity on blood agar, which has led to rather unsatisfactory streptococcal taxonomy. Genetic analysis has now enabled the subdivision of the species of *Streptococcus* into six clusters or groups as follows: pyogenic streptococci, milleri or anginosus group, mitis group, salivarius group, mutans group, and bovis group. Since the medically important members of the mitis, salivarius, and mutans groups are all oral streptococci, and of clinical relevance predominantly in endocarditis, they will be considered together.

The pyogenic streptococci

The pyogenic streptococci include the major human pathogen *S. pyogenes* (Lancefield group A), group B streptococci (*S. agalactiae*), and groups C and G streptococci. These organisms are b-haemolytic on blood agar.

Streptococcus pyogenes (b-haemolytic group A)

Since the beginning of the last century, and long before the introduction of antibiotics, infections with *S. pyogenes* declined in incidence and severity until, in the 1980s, highly virulent streptococci appeared causing very severe infections often in otherwise healthy people. Such cases occurred not only in the United Kingdom but in most of the developed world. *S. pyogenes* infection is usually community-acquired but may be acquired in hospital, where the most serious infections are postoperative.

Carriage

Although *S. pyogenes* is an invasive organism, it lives on epithelial surfaces (asymptomatic carriage) usually in the nose and throat; carriage can also be anal, vaginal, and on the scalp. Pharyngeal carriage rates are usually much higher in healthy children (5 to 20 per cent) than in adults (0.5 per cent) and also vary with season, year, and geographical location; they are also higher in crowded living conditions. *S. pyogenes* can persist for months after acute pharyngitis, though in decreased numbers. Survival in the environment is poor and *S. pyogenes* can only survive on skin squames and dust for a limited period and in low numbers.

Pathogenicity, virulence, and typing

S. pyogenes is an extracellular pathogen and produces virulence factors that enable it to avoid host defences and spread in tissues. The main virulence factor is the M protein; streptococci rich in M protein resist phagocytosis by polymorphs. Immunity to *S. pyogenes* infection is associated with the development of opsonic antibodies to antiphagocytic epitopes of M protein; it is usually type specific and lasts for many years, perhaps indefinitely. M protein was first described in the 1920s by Rebecca Lancefield; over 100 M types have now been differentiated. Lancefield also developed the supplementary T typing system which distinguishes 26 serotypes of a trypsin-resistant surface protein (T antigen), most of which can be expressed by several different M types. Certain M types also produce a serum opacity factor (OF+). These typing systems are still widely used in epidemiological studies to distinguish between strains of *S. pyogenes*. Recent studies have shown considerable genetic diversity in *S. pyogenes*, and horizontal transfer and recombination of virulent genes have played a major role. This finding is likely to be relevant to the emergence of new, unusually virulent clones of the organism.

In addition to M protein, lipoteichoic acid, important in the host–bacterial interaction, is expressed on the surface of the organism and is the adhesin that binds the organism to fibronectin on the surface of the oral epithelial cell membranes and initiates the colonization that precedes infection. *S. pyogenes* has a hyaluronate capsule which, like M protein, is also antiphagocytic, and is an additional virulence factor. The extent of encapsulation varies and colonies with prominent capsules are very mucoid on blood agar. Strains of *S. pyogenes* that are both rich in M protein and heavily encapsulated are readily transmitted from person to person, and tend to produce severe infections.

S. pyogenes produces many extracellular substances, several of which are important in the pathogenesis of infection. The most familiar are streptolysin O, deoxyribonuclease (DNAase) B, and hyaluronidase as serum antibodies to these provide retrospective confirmation of recent streptococcal infection. Other extracellular products include DNAases A, C, and D, streptolysin S, proteinase, streptokinase, and the substances previously known as erythrogenic toxins. These toxins have now been designated streptococcal pyrogenic exotoxins (**SPE**) -A, -B, -C, and possibly -D. SPE-A, and possibly others, is coded by a phage gene. These toxins, known as superantigens, have diverse effects on the host. In addition to the rash of scarlet fever, they cause fever, changes in the blood–brain barrier, organ damage, and lethal shock in animals. They have profound effects on the immune system including increasing susceptibility to endotoxic shock, blockade of the reticuloendothelial system, and alterations in T-cell function.

When *S. pyogenes* enters the body, either through the upper respiratory tract mucosa or a break in the skin, a local lesion may occur or there may be spread along tissue planes or lymphatics. The M protein is not toxic in itself but protects the streptococcus from phagocytosis and antibodies to the M protein are opsonic. In some two-thirds of patients with serious invasive disease, who may present with fever, shock, and renal impairment, the portal of entry is the skin, and infection of soft tissue is apparent, but in others the site of infection may not be evident.

Infections caused by *S. pyogenes*

S. pyogenes causes a variety of illnesses ranging from very common, usually mild, conditions such as pharyngitis and impetigo, through common, temporarily disabling cellulitis, to less common, puerperal sepsis and very severe infections such as type II necrotizing fasciitis, bacteraemia, and toxic shock. It is also associated with the non-suppurative sequelae of acute rheumatic fever and acute glomerulonephritis.

Streptococcal pharyngitis

Streptococcal pharyngitis or tonsillitis is one of the commonest bacterial infections in children from 5 to 15 years, but all ages are susceptible. The incubation period, at least in outbreaks, is short (1 to 3 days) and the onset of the infection marked by the abrupt onset of sore throat and pain on swallowing with malaise, fever, and headache. The signs are redness and oedema of the pharynx, enlarged red tonsils with spots of white exudate, and enlarged tender anterior cervical lymph glands. Nausea, vomiting, and abdominal pain are common in children, and in infants and preschool children there may be few definite signs of pharyngitis but fever, nasal discharge, enlarged cervical lymph glands, and otitis media occur.

Suppurative complications

Direct extension of streptococcal pharyngitis can give rise to acute sinusitis or otitis media and other suppurative complications include peritonsillar abscess (quinsy) and retropharyngeal abscess, which often contain oral flora including anaerobes with or without *S. pyogenes*, and suppurative cervical lymphadenitis.

Scarlet fever

Scarlet fever results from infection with a strain of *S. pyogenes* that produces SPE (erythrogenic toxin). It is usually associated with streptococcal pharyngitis but may follow streptococcal infections at other sites and occurs with invasive disease. Scarlet fever rarely follows streptococcal pyoderma. Most cases occur in school-age children and the rash must be distinguished from viral exanthems, Kawasaki disease, and staphylococcal toxic shock syndrome. The rash, which generally appears on the second day of clinical illness, is usually a diffuse erythema, symmetrical, and blanches on pressure. It is seen most often on the neck, chest, folds of the axilla, and groin. Occlusion of sweat glands gives the skin a 'sandpaper' texture, a useful sign in dark-skinned patients. The face appears flushed with circumoral pallor. There are small red haemorrhagic spots on the palate and the tongue is initially covered with a white fur through which red papillas appear ('strawberry tongue') and then, usually after the rash develops, the white fur peels off leaving a raw red papillate surface ('raspberry tongue'). The rash persists for several days and later (up to 3 weeks) peeling may occur, usually on the tips of the fingers, toes, or ears and less often over the trunk and limbs. A similar rash may develop as a reaction to streptokinase thrombolytic therapy.

Streptococcal perianal infection (cellulitis)

This is a superficial, well-demarcated rash spreading out from the anus in young children, usually boys, associated with itching, rectal pain on defaecation, and blood-stained stools. *S. pyogenes* is isolated from perianal cultures and usually also from pretreatment throat swabs.

Streptococcal vulvovaginitis

Vulvovaginitis in prepubertal girls is often caused by *S. pyogenes* and presents with serosanguinous discharge and erythema of the labia and vaginal orifice. As with perianal infections, *S. pyogenes* is usually also found in the throat. In both streptococcal perianal infection and vulvovaginitis, more than one child in the family may be affected and nasopharyngeal carriage is likely in both infected and uninfected children.

Streptococcal skin and soft tissue infections

Pyoderma/impetigo

Almost any purulent lesion of the skin can yield *S. pyogenes*, sometimes with *Staphylococcus aureus.* Such lesions include impetigo, infected cuts and lacerations, insect bites, scabies, intertrigo, and ecthyma. *S. pyogenes* also often causes secondary infection in varicella, occasionally with resultant bacteraemia. The term pyoderma is used synonymously with impetigo for discrete, purulent, apparently primary infections of the skin that are prevalent in many parts of the world, especially in children. These lesions are initially papules, then vesicular with surrounding erythema, and finally pustules with crusting exudate; they may be localized to one part of the body or generalized. Outbreaks of impetigo can occur among adults subject to skin trauma, such as rugby football players (scrumpox), and streptococcal infection of cuts on the hands and forearms are an occupational hazard for workers in the meat trade. Ecthyma is an ulcerated form of impetigo in which ulceration extends into the dermis.

Invasive streptococcal infections of skin and soft tissues

Erysipelas

This is an acute inflammation of the skin with lymphatic involvement. The streptococci are localized in the dermis and hypodermis. It usually affects the face, particularly in the elderly, but may occur elsewhere. It may be bilateral (Plate 1) and is sometimes recurrent. There is usually a history of sore throat, but the mode of spread to the skin is unknown. It is usually accompanied by fever, rigors, and toxicity. The cutaneous lesion begins as a localized area of erythema and swelling and then spreads with rapidly advancing raised red margins that are well demarcated from adjacent normal tissue. Facial erysipelas begins over the bridge of the nose and spreads over the cheeks. Vesicles and bullas appear, which become crusted when they rupture. There is marked oedema and the eyes are often closed. When the infection resolves it is often followed by desquamation. Intense local allergic reactions to topical agents, such as cosmetics, may cause confusion.

Cellulitis (Plate 2)

Cellulitis is commonly caused by streptococci and *Staphylococcus aureus.* This is an acute spreading inflammation of the skin and subcutaneous tissues with local pain swelling and erythema. Fever, rigors, and malaise may precede by a few hours the appearance of the skin lesion and associated lymphangitis and tender lymphodenopathy. Streptococcal cellulitis differs from erysipelas in that the lesion is not raised and the demarcation between affected and unaffected skin is indistinct. It may result from infection of burns, mild trauma, or surgical wounds. When this involves the leg, fungal infection of the feet is often present and predisposes to streptococcal invasion. After the first episode, there is a tendency for recurrence in the same area. Intravenous drug users are also at risk of streptococcal cellulitis associated with skin and tissue infection and septic thrombophlebitis.

(Type II) necrotizing fasciitis (streptococcal gangrene)

This infection, described by Meleney in 1924, involves the deep subcutaneous tissues and fascia (and occasionally muscle as well) with extensive, rapidly spreading necrosis and gangrene of the skin and underlying structures. It is generally community-acquired, usually involving the arm or leg, but may also occur after surgery, even quite minor. Some victims are diabetic, but the majority were previously healthy. Risk factors, providing a portal of entry, include surgery, trauma, childbirth, intravenous drug abuse, and chickenpox. Blunt trauma and muscle strain which may generate a haematoma and use of non-steroidal anti-inflammatory agents are also implicated. The infection begins at the site of trivial or even inapparent trauma with redness, swelling, fever, and rapidly escalating focal pain followed by purple discoloration and the development of bullae, often haemorrhagic. Bacteraemia is often present and within days skin necrosis occurs followed by extensive sloughing. The patient is profoundly ill and the disease has a high case fatality of 30 to 70 per cent. Features of streptococcal toxic shock syndrome are associated in many cases. The United Kingdom media memorably dubbed *S. pyogenes* the 'flesh-eater' in reports of a cluster of cases of necrotizing fasciitis in 1994. Treatment involves early intravenous antibiotics (clindamycin has several theoretical advantages over penicillin), urgent surgical débridement of necrotic tissue, and intensive care to support failing organs and systems (e.g. cardiovascular and renal). Benefits of immunoglobulin are anecdotal.

Streptococcal toxic shock syndrome

This syndrome was described in 1987 in patients with severe *S. pyogenes* infection and clinical features remarkably similar to those of the staphylococcal toxic shock syndrome described a decade earlier. Neither are likely to be new diseases. Definitions of streptococcal toxic shock syndrome vary. Some limit the definition to cases of shock and multiorgan failure where there is a rash or desquamation, whilst others include all cases of shock and its non-specific sequelae such as coagulopathy, uraemia, or jaundice, irrespective of skin lesions. Streptococcal toxic shock syndrome is usually associated with necrotizing fasciitis or myositis. It can occur at all ages and many of those affected are young and previously healthy. Most cases have been community-acquired, though it can be acquired in hospital. M1 has been the predominant serotype in many countries, though others, especially 2, 3, 12, and 28, have also been implicated. Most strains produce SPE-A. Interestingly there is an amino acid homology of 50 per cent and immunological cross-reactivity between SPE-A and staphylococcal enterotoxins B and C, which together with staphylococcal TSS toxin-1 are relevant in non-menstrual staphylococcal toxic shock syndrome.

Streptococcal bacteraemia

In parallel with the increase in serious *S. pyogenes* infections there has been an increase in bacteraemic infections, both community- and hospital-acquired (usually postoperative) (Plate 3). While many patients have an underlying disease, most often malignancy, immunosuppression, or diabetes, others are previously healthy adults between 20 and 50 years old. The portal of entry is usually the skin. The mortality is higher in patients with underlying disease.

Puerperal and neonatal infection

Historically *S. pyogenes* has always been an important cause of puerperal sepsis ('childbed fever'), but in the postantibiotic era it was rarely encountered in obstetric practice until the 1980s when sporadic cases occurred, some with streptococcal toxic shock syndrome, and some women have died. These infections follow abortion or delivery when streptococci (usually colonizing the patient herself) invade the endometrium, lymphatics, and bloodstream. They can be devastatingly severe and present with non-specific signs such as restlessness and gastrointestinal upset that may not immediately suggest sepsis. Fever may be absent resulting in further diagnostic confusion. The streptococcal infection not only involves the uterus and adnexa but sometimes distant sites such as joints as well. It can also affect the baby causing serious neonatal infection including meningitis. Instrumentation in the presence of asymptomatic vaginal or anorectal carriage of *S. pyogenes* can result in severe infection.

Other infections

S. pyogenes can, though rarely does, cause pneumonia (usually associated with viral infection or pulmonary disease), osteomyelitis, septic arthritis, meningitis, pericarditis (Plate 4), endophthalmitis, and endocarditis.

Laboratory diagnosis of *S. pyogenes* infection

S. pyogenes is easy to culture in the laboratory and usually grows on blood agar in 24 hours. Throat swabs must be taken before antibiotics are given or the chance of recovery is slim. Kits for the detection of the group A antigen directly from throat swabs are available and give few false-positive reactions, but they are seldom used in the United Kingdom. Even trivial skin lesions are worth swabbing (if necessary with a moistened swab) and a search for such lesions often pays dividends. Swabs from the surface of cellulitis and erysipelas rarely yield streptococci and although they may be recovered from specimens obtained by aspiration, in practice this is seldom done. Blood cultures should be done in any patient who is ill whether febrile or not. Serological confirmation of infection with *S. pyogenes* when the organism has not been isolated can be obtained by the detection of raised antibodies to its extracellular products. Most laboratories tend to use two or more tests. Interpretation requires knowledge of the level of titres in the community for those without a history of recent streptococcal infection. In the United Kingdom the upper limit of titres in teenagers and young adults without such a history is antistreptolysin O (ASO) 200, antideoxyribonuclease B (ADB) 240, and antihyaluronidase (AHT) 128.

Management and antibiotic treatment of *S. pyogenes* infection

Remarkably, *S. pyogenes* remains exquisitely sensitive to penicillin and this is the antibiotic of choice for treatment, parenterally for severe infections and orally otherwise. Conventionally, 10 days treatment is recommended for pharyngeal infections to eradicate the organism and prevent acute rheumatic fever. In practice, compliance with this regimen is poor as once the symptoms abate there is a natural reluctance to continue the antibiotic. Treatment of patients allergic to penicillin is most often with erythromycin or the newer macrolides (azithromycin and clarithromycin), but some 3 to 5 per cent of strains are erythromycin resistant. *S. pyogenes* is also sensitive to cephalosporins. Topical agents such as mupirocin and fusidic acid are useful in addition to systemic antibiotic treatment in impetigo and other skin lesions. Patients with streptococcal toxic shock syndrome will need intensive care and many require inotropic support, ventilation, and haemodialysis. Urgent surgical intervention is needed for necrotizing fasciitis and myositis. Clindamycin (in addition to penicillin) has been recommended for patients with established invasive streptococcal infections since this drug stops the metabolic activity of the streptococci and thus halts further production of toxin. This is specially relevant in type II necrotizing fasciitis/myositis and streptococcal toxic shock syndrome. Intravenous immunoglobulin has also been used in an attempt to neutralize the streptococcal toxins, but reports of its effects are inconclusive. Prevention of recurrent cellulitis of the lower legs involves meticulous foot hygiene with treatment of 'athlete's foot' fungi and reduction in skin carriage using topical mupirocin. Oedematous limbs can benefit from elastic stockings. Antibiotic prophylaxis may be required in cases of frequent recurrence refractory to these measures. Lastly it should be remembered that *S. pyogenes* is readily transmitted from person to person and thus appropriate infection control precautions should be taken until swabs show the organism has been eradicated.

b-Haemolytic groups C and G streptococci

These streptococci are sometimes referred to as 'large colony-forming group C and G streptococci' to distinguish them from the small colony-forming strains of streptococci with the same Lancefield antigens that belong to the anginosus or milleri group (see below). Groups C and G streptococci are closely related genetically. They are most conveniently regarded as 'pyogenes-like' as the infections they cause are similar to those caused by *S. pyogenes* though these streptococci tend to be less virulent than *S. pyogenes*. Infections with these streptococci are less common than *S. pyogenes* infections. Although post-streptococcal glomerulonephritis has been associated with pharyngitis caused by both groups C and G streptococci, acute rheumatic fever has not. Group C streptococci are less frequently encountered in human infections than group G and most group C infections are caused by *S. equisimilis*; those caused by *S. zooepidemicus* have an animal source. Group G streptococci are frequently isolated from leg ulcers and pressure sores, usually with other bacteria. In such patients cellulitis and systemic upset are rare and the organisms are just colonizing the lesions. They, like *S. pyogenes*, can cause cellulitis in lymphoedematous limbs.

b-Haemolytic group B streptococci (*S. agalactiae*)

The group B streptococcus has been known for over a century as a cause of bovine mastitis and in the 1930s it was recognized as a vaginal commensal, an occasional cause of puerperal fever, and an uncommon cause of invasive disease in adults. Not until the 1960s was it realized that the group B streptococcus was an important neonatal pathogen, and some 20 years later it had replaced *Escherichia coli* as the predominant neonatal pathogen.

Carriage

Group B streptococci can be recovered from various sites in healthy adults but vaginal carriage has been most extensively investigated. Swabs from the lower vagina are more often positive than cervical swabs and carriage rates of 3 to over 40 per cent have been reported. Higher rates have been obtained with selective media and enrichment techniques. Carriage also increases with sexual activity and is highest in women attending genitourinary clinics. The urethra, vagina, perineum, and anorectal region have all been suggested as the prime site of carriage. Some 5 to 10 per cent of normal adults carry group B streptococci in the throat, independent of urogenital and anorectal carriage.

Pathogenicity, virulence, and typing

The chief determinant of virulence appears to be the capsular polysaccharide, and most human strains carry one of six sialic acid-containing polysaccharides that surround the cell wall. In addition, a protein antigen (c, X, or R) may be carried. Certain combinations are common; serotypes III or III/R form one-quarter of all isolates from superficial sites on women, but three-quarters of all group B streptococci causing meningitis in infants. They are also the commonest serotypes found in adult (non-pregnant) infections. The type polysaccharide, like the M protein of *S. pyogenes*, inhibits phagocytosis. Colonization of the mucous membranes of the neonate results from vertical transmission of the organism from the mother either *in utero* by the ascending route or at delivery. The rate of vertical transmission in neonates

born to mothers colonized with group B streptococci is about 50 per cent, but the incidence of symptomatic infection in neonates born to colonized mothers is only about 1 to 2 per cent. It is much higher in preterm infants. Nosocomial colonization of neonates can also occur. In most cases of adult infections (other than in pregnant women) the source of the infection is unknown.

Infections caused by group B streptococci

These are commonly neonatal or puerperal infections, but group B streptococci also cause infection in the non-pregnant adult.

Neonatal infection

The frequency of neonatal infection (bacteraemia, meningitis, or both) has been variously quoted as between 0.3 and 5.4 cases/1000 live births, but these figures have wide confidence limits. Two fairly distinct clinical patterns of disease predominate, but the spectrum is wide and includes impetigo neonatorum, septic arthritis, osteomyelitis, pneumonitis, peritonitis, pyelonephritis, facial cellulitis, conjunctivitis, and endophthalmitis.

Early-onset disease

Symptoms develop within the first 5 days of life with a mean of 20 h, though they can present at birth suggesting an intrauterine onset of infection. Early-onset disease is most often a bacteraemia with no identifiable focus of infection, but can also be pneumonia or, infrequently, meningitis. The presenting signs include lethargy, poor feeding, jaundice, grunting respirations, pallor, and hypotension and they are common to all types of disease. Respiratory symptoms are nearly always present. The only reliable way of detecting meningitis is by lumbar puncture. Mortality rates are high in low birth-weight babies. In addition to positive blood cultures, the infecting strain can be found in the mother's vagina and cultured from 'screening' sites on the baby; these include ear, throat, and nasogastric aspirate.

Late-onset disease

This usually presents between 7 days and 3 months after birth, often in previously healthy babies born after a normal labour who are admitted unwell from home. The pathogenesis is less clear than in cases of early-onset disease and only about half the cases are associated with mucosal colonization during delivery. Most babies have meningitis and concomitant bacteraemia and present with non-specific symptoms such as lethargy, poor feeding, irritability, and fever. Neurological sequelae are common among survivors.

Puerperal infection

Puerperal infection with Group B streptococci usually occurs within 24 to 48 h of delivery or abortion. The source of the organism is always the vagina and infection is more likely when there has been premature rupture of the membranes and chorioamnionitis. Most infections are endometritis with fever and uterine tenderness sometimes associated with retained products of conception, but group B streptococci can also cause wound infection after caesarean section. Bacteraemia is common. Other bacteria, both aerobes and anaerobes, are sometimes isolated from the genital tract and wounds in addition to the group B streptococcus. Very rarely the streptococcus may spread to other sites in puerperal women.

Infection in non-pregnant adults

The prominence given to group B streptococci as neonatal and puerperal pathogens has tended to overshadow their importance in non-pregnant women and men in whom they cause significant morbidity and mortality. Most infections are community-acquired, occur in the middle aged and elderly, and are as common in males as females. Many, though by no means all, patients with group B streptococcal infection have underlying diseases, particularly diabetes and myeloma. Skin and soft tissue infections are especially common in patients with diabetes. Occasional urinary tract infections occur, in men as well as women. Bacteraemic infections serve to emphasize the virulence of group B streptococci, and they have increased in incidence, or perhaps have been increasingly recognized, since the early 1990s. Community-acquired group B streptococcal bacteraemia is similar in many respects to that caused by *Staphylococcus aureus* since common clinical manifestations include endocarditis, vertebral osteomyelitis, septic arthritis, endophthalmitis, and meningitis. As with staphylococcal infections, some bacteraemic patients have more than one metastatic focus of infection, which can lead to diagnostic confusion.

Laboratory diagnosis of group B streptococcal infection

Group B streptococci are readily isolated from any clinical specimen in the laboratory and easily identified by Lancefield grouping. The group B antigen is not shared by any other streptococcus. Importantly the antigen can be reliably detected in fluids such as blood, urine, or cerebrospinal fluid by latex particle agglutination enabling a rapid diagnosis.

Treatment of group B streptococcal infection

Group B streptococci are sensitive to penicillin and this is the antibiotic of choice for treatment. They are rather less sensitive to penicillin than *S. pyogenes* with minimum inhibitory concentrations some four- to 10-fold higher. For this reason penicillin is sometimes combined with gentamicin for meningitis and other serious infections, though this is not of proven benefit. Certainly, the maximum recommended dose of parenteral penicillin should be given whether combined with gentamicin or not. Penicillin allergy is not likely to be an issue in neonates; adults with meningitis can be treated with chloramphenicol. Most group B streptococci are sensitive to erythromycin and they are sensitive to cephalosporins.

Prevention of neonatal infection with group B streptococci

During the 1990s the incidence of disease caused by mother-to-child transmission of group B streptococci in the United States fell by two-thirds as a result of the increased use of intrapartum penicillin in women at high risk of transmitting the infection, an intervention largely brought about by parental pressure. The American authorities recommend either prenatal screening or a risk-based strategy to identify women to receive intrapartum antibiotics. Similar recommendations are to be introduced in the United Kingdom. Any protocol for prophylactic penicillin based on the isolation of group B streptococci in late pregnancy would present difficulties in a busy obstetric unit and culture methods may also fail to detect the organism unless vaginal and rectal swabs are cultured in selective broth media. Maternal colonization with group B streptococci can be identified rapidly and reliably by polymerase chain reaction assay, but this is unlikely to be adopted as a routine round-the-clock service. An effective vaccine is an alternative approach, as yet unavailable.

Streptococci of the anginosus or milleri group

This group of streptococci has been a source of considerable taxonomic confusion, partly the result of a lack of international consensus on nomenclature, but also because of a lack of reliable phenotypic differences between taxa within the group. Most clinicians are familiar with the organism they know as '*Streptococcus milleri*'. There are three species of milleri streptococci, *S. anginosus*, *S. constellatus*, and *S. intermedius*, but despite increasing awareness of the clinical significance of the *milleri* group little is known about the association between individual species and specific sites of isolation and diseases. These streptococci are found in large numbers in the normal flora of the upper respiratory tract, gastrointestinal tract, and genital tract, and are commonly isolated from a range of pyogenic infections, sometimes in pure culture, but often with other organisms, particularly anaerobes. These infections include dental abscesses, intra-abdominal abscesses (especially of the liver), subphrenic abscesses, lung abscesses and empyema, and brain abscesses. Such is the propensity of these organisms to cause deep-seated abscesses that isolation of a milleri streptococcus from a blood culture should prompt investigations to detect such a focus. Milleri streptococci are also commonly isolated from inflamed appendices and postappendicectomy wound infection. Unlike other viridans and non-haemolytic streptococci, milleri streptococci seldom cause endocarditis. They form minute colonies on blood agar and are preferentially anaerobic on primary isolation. They may be a-, b-, or non-haemolytic. Some have the Lancefield antigens A, C, G, or F. All group F streptococci are milleri group whereas not all milleri streptococci are group F. Another useful clue to their identity in the laboratory is the distinct caramel smell of many strains on blood agar, the result of the diacetyl metabolite. Most strains are very sensitive to penicillin.

The mitis, salivarius, and mutans groups of streptococci (oral/viridans streptococci)

This group of usually a-haemolytic (viridans) streptococci includes *S. pneumoniae* and those oral streptococci (*S. mitis*, *S. oralis*, *S. sanguis*, *S. gordonii*, and rarely, *S.*

salivarius) that are the commonest cause of infective endocarditis of oral or dental origin. These streptococci occasionally cause bacteraemia in neutropenic patients who sometimes have detectable mouth lesions and neonatal infection as they are found as part of the normal vaginal flora.

The bovis group of streptococci

Although this group comprises at least three species, *S. bovis* is the main species of medical importance. *S. bovis* is similar to the enterococci in that it bears the Lancefield group D antigen and is a gastrointestinal commensal, but unlike the enterococci, it is sensitive to penicillin. It can be misidentified in the laboratory either as an oral streptococcus or as an enterococcus. Most patients with *S. bovis* bacteraemia will have endocarditis and it is seldom isolated from other sites. It is important to recognize *S. bovis* in a blood culture as the organism is associated with colonic pathology, and patients should be specifically investigated for this.

Nutritionally variant organisms previously classified as streptococci, now *Abiotrophia* spp.

These organisms, which occasionally cause endocarditis, require pyridoxal or thiol group supplementation for growth in the laboratory and tend to form satellite colonies round *Staphylococcus aureus*. Although most blood culture media will support their growth, successful subculture requires supplementation or cross-streaking of the plates with *S. aureus* to provide the necessary growth factors. The *Abiotrophia* include three species, *S. adjacens*, *S. defectivus*, and the recently described *A. elegans*. They are less susceptible to penicillin than other streptococci.

Streptococcus suis

This streptococcus, which can be misidentified in the laboratory as *S. bovis* or an enterococcus as it reacts with group D antiserum, is an important pathogen of young pigs causing meningitis, septicaemia, arthritis, pneumonia, and endocarditis and is also carried in the pharynx of healthy pigs. *S. suis* type II (also referred to as group R streptococci) is not only the most invasive type in pigs, it can cause serious infection—mainly septicaemia and meningitis, but also septic arthritis, pneumonia, and endophthalmitis—in humans, in whom it is an occupational disease of pig farmers, abattoir workers, and factory workers handling pig meat (see Chapter 24.14.1). The streptococcus probably enters the bloodstream via skin abrasions that are common in the above occupations. *S. suis* type II meningitis results in deafness in about half of those affected.

Enterococci

Enterococci are Lancefield group D, Gram-positive cocci that can grow and survive in extreme cultural conditions, and are also more resistant to antibiotics than streptococci. They form part of the normal gut flora of humans and animals. Overall, the commonest clinical isolates of enterococci are *Enterococcus faecalis*, but the more antibiotic-resistant species *E. faecium* is increasingly encountered in hospitals. Nosocomial isolates of enterococci have dramatically increased in the 1990s. Other species, including *E. casseliflavus*, *E. durans*, and *E. avium*, are occasionally isolated. In most cases it is unnecessary to determine the species of enterococci in a clinical laboratory but sometimes differentiation between *E. faecalis* and *E. faecium* is helpful, for instance in epidemiological studies and in endocarditis because of their different antibiotic susceptibilities.

Infections caused by enterococci

Enterococci are an increasingly important cause of nosocomial infection and colonization, possibly the result of the large-scale use of antibiotics such as cephalosporins and quinolones to which they are inherently resistant. They occasionally cause community-acquired urinary tract infections but the most important community-acquired infection is endocarditis, which is increasing in incidence. This infection is almost always caused by *E. faecalis.* Any patient admitted from the community with *E. faecalis* in blood cultures should be assumed to have endocarditis until proved otherwise. Enterococci are predominantly hospital pathogens and cause urinary infection, particularly after instrumentation, intra-abdominal infections, wound infections (usually with other organisms), infections associated with intravascular devices and dialysis, and occasionally endocarditis.

Antibiotic sensitivity and treatment

Enterococci are not only intrinsically resistant to many antibiotics, they show a remarkable ability to acquire new mechanisms of resistance. This allows them to survive in environments in which large amounts of antibiotics are used and also has important therapeutic consequences, particularly for the treatment of endocarditis and other serious infections. Fortunately many patients from whom enterococci are isolated do not require antibiotic treatment. Sensitive enterococci cannot be killed by ampicillin/amoxycillin alone, though combination with an aminoglycoside is bactericidal (synergy); but many strains now exhibit high-level gentamicin resistance and for them the combination is not bactericidal. *E. faecium* is almost always resistant to ampicillin/amoxycillin and *E. faecalis* is occasionally. The first published report of vancomycin-resistant enterococci (**VRE**) was in 1988 from a London hospital outbreak, though such strains had been recognized a year before in Paris. Most strains of VRE in the London outbreak were *E. faecium* and overall most VRE are *E. faecium*. There are four recognized phenotypes of vancomycin resistance; the first isolates of VRE were highly resistant to vancomycin and teicoplanin and exhibit what is known as the VanA resistance phenotype. Since then, levels of resistance to teicoplanin in this phenotype have been more varied. Most VanA enterococci are *E. faecium*, but this phenotype also occurs in *E. faecalis* and occasionally in other species. The VanB phenotype is associated with low-level vancomycin resistance and sensitivity to teicoplanin and is found in both *E. faecalis* and *E. faecium.* Both VanA and VanB are acquired traits. The VanC phenotype is an intrinsic property of *E. casseliflavus* and *E. gallinarum* and these species have low-level resistance to vancomycin but are sensitive to teicoplanin. A fourth phenotype, VanD, has been described in a single strain of *E. faecium*. Vancomycin-resistant *E. faecium*, though not vancomycin-resistant *E. faecalis*, are sensitive to quinupristin/dalfopristin (Synercid) and all VRE are sensitive to the oxazolidinone Linezolid.

The antibiotic susceptibilities of the enterococci outlined above serve to emphasize that these bacteria are the most antibiotic-resistant Gram-positive bacteria now encountered in hospital practice. Fortunately many, perhaps most, of the patients from whom they are isolated do not require antibiotic treatment at all, but for those who do, the effective treatment of serious infection caused by enterococci and particularly antibiotic-resistant strains requires microbiological expertise.

Further reading

Bisno AL, Stevens DL (2000). *Streptococcus pyogenes* (including streptococcal toxic shock syndrome and necrotizing fasciitis). In: Mandell GL, Bennett JE, Dolin R, eds. *Principles and practice of infectious diseases*, pp 2101–17. Churchill Livingstone, New York.

Colman G *et al*. (1993). The serotypes of *Streptococcus pyogenes* present in Britain during 1980 to 1990 and their association with disease. *Journal of Medical Microbiology* **39**, 165–78.

Edwards MS, Baker CJ (2000). *Streptococcus agalactiae* (Group B streptococcus). In: Mandell GL, Bennett JE, Dolin R, eds. *Principles and practice of infectious diseases*, pp 2156–67. Churchill Livingstone, New York.

Jacobs JA (1997). The '*Streptococcus milleri*' group: *Streptococcus anginosus*, *Streptococcus constellatus* and *Streptococcus intermedius*. *Reviews in Medical Microbiology* **8**, 73–80.

Katz AR, Morens D (1992). Severe streptococcal infections in historical perspective. *Clinical Infectious Diseases* **14**, 298–307.

Murray BE (1990). The life and times of the *Enterococcus*. *Clinical Microbiological Reviews* **3**, 46–65.

Stevens DL (1992). Invasive Group A streptococcus infections. *Clinical Infectious Diseases* **14**, 2–13.

Stevens D (1995). Streptococcal toxic shock syndrome: spectrum of disease, pathogenesis and new concepts of treatment. *Emergencies in Infectious Disease* **1**, 69–78.

Woodford N (1998). Glycopeptide-resistant enterococci: a decade of experience. *Journal of Medical Microbiology* **47**, 849–62.

7.11.3Pneumococcal diseases

Keith P. Klugman and Brian M. Greenwood

Streptococcus pneumoniae (the pneumococcus) causes a considerable burden of vaccine-preventable disease. It is a leading cause of bacterial meningitis, pneumonia, otitis media, and sinusitis. The global HIV pandemic has greatly increased the burden of pneumococcal disease in both children and adults and the dissemination of a number of multiresistant pneumococcal clones has complicated the management of this disease. In the first decade of the twenty-first century it is likely that the introduction of pneumococcal conjugate vaccines will reduce the burden of pneumococcal disease in children and may also contribute to interrupting the transmission of antibiotic-resistant strains.

History and biology of the pathogen

Streptococcus pneumoniae is a Gram-positive, lanceolate-shaped diplococcus that was isolated independently by Sternberg and Pasteur in 1881. They had inoculated human sputum into rabbits. The first demonstration of the pathogen as a cause of pneumonia was made by Friedlander in 1883. The sensitivity of the pathogen to ethylhydrocupreine (optochin) was noted in the early 1900s and the use of this agent to treat experimental pneumococcal disease was one of the first examples of antibacterial chemotherapy. The emergence of resistance following treatment was noted in that study in humans and its use was abandoned due to side-effects, including temporary blindness. The multiple serotypes of the pneumococcus are due to 90 distinct capsular polysaccharides. The pneumococcus has played a role in biology beyond that of the description of its virulence factors. The discovery of DNA as the transforming principle was based on the transformation of pneumococcal serotypes.

Adherence and pathogenesis

Newborn infants are free of pneumococcal colonization, and infections follow colonization. Colonization occurs rapidly in developing countries and, in such communities, most infants are nasopharyngeal carriers before the age of 6 months. Early colonization is associated with an increased incidence of otitis media. It is probable that multiple serotypes of pneumococci are carried simultaneously and that current methods of detection fail to identify subdominant strains. The duration of carriage varies by serotype and there is some evidence that the risk of invasive disease is greatest at the time of acquisition of a new serotype.

Pneumococci bind to specific galactose receptors on nasopharyngeal epithelial cells and pneumocytes. The bacteria undergo phase variation into transparent and opaque phenotypes. The transparent phenotype is better able to adhere to epithelial cells in the nasopharynx and adherence is enhanced by interleukin-1 (IL-1) and by tumour necrosis factor-a(TNF-a). The basis for the invasion of colonizing pneumococci is not clearly understood although preceding viral infections, such as influenza or respiratory syncytial virus infection, may be important. Influenza virus enhances the adhesion of pneumococci to respiratory cells and the binding of pneumococci to platelet activating factor is associated with invasion of activated cells. The binding of transparent phenotype pneumococci to the PAF receptor is mediated by a phosphorylcholine ligand. Other surface receptors such as pneumococcal surface adhesin A (PsaA), pneumococcal surface protein C (PspC), and choline-binding protein A (CbpA) also play a role in adhesion.

Once invasion has occurred, other components of the bacterium such as the phosphorylcholine moiety of techoeic acid C polysaccharide in the cell wall contribute to the induction of a marked acute inflammatory response. Cytokines, such as tumour necrosis factor and interleukin 1, play an important part in the pathogenesis of this inflammatory process. Reduction of the inflammatory response of animals with experimentally induced pneumococcal meningitis with drugs such as corticosteroids increases their survival, but it is not known whether this is also the case in humans (see below). Virulence may also be enhanced by pneumococcal surface protein A (PspA), and by the production of bacterial enzymes such as hyaluronidase, neuraminidase, and pneumolysin. The direct neurotoxicity of nitrous oxide may also be important in the neurological damage of pneumococcal meningitis. The pathway of pneumococcal infection is illustrated in Fig. 1 and the pathogenesis of infection is summarized in Fig. 2.

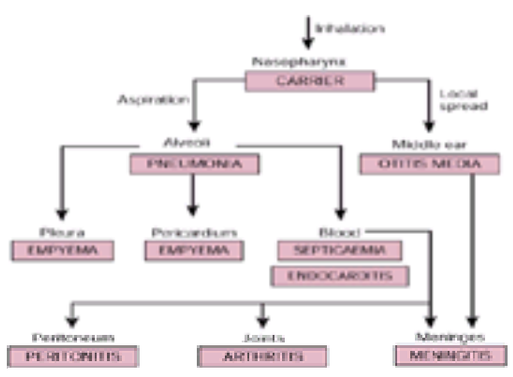

Fig. 1 The pathway of pneumococcal infection.

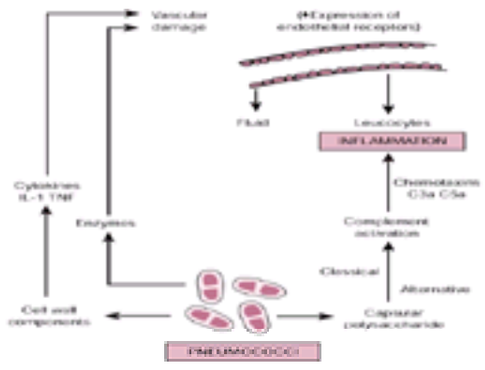

Fig. 2 The pathogenesis of pneumococcal infection.

Antibiotic resistance

There was the emergence of a global pandemic of antibiotic resistance in the pneumococcus during the 1990s. The use of antibiotics selects resistant pneumococci at the national, provincial, hospital, and individual level. The identification of penicillin resistance in the pneumococcus was first made by Hansman and Bullen in 1967, although resistance to macrolides and tetracycline had already been described. Multiresistant pneumococci were found first in South Africa in 1978. These strains were also fully resistant to penicillin (minimum inhibitory concentration greater than 1 µg/ml). The selection of multiresistant strains is complex in that the differential ability of each of the antibiotic classes to select the multiresistant pathogen is not clearly understood. There are data to suggest that trimethoprim sulphamethoxazole may select for multiresistant strains more successfully than b-lactams, and reduction in the use of macrolides and trimethoprim sulphamethoxazole has been associated with a reduction in the prevalence of multiresistant pneumococci.

The nasopharyngeal carriage rate of antibiotic-resistant strains has been shown to predict closely the susceptibility of strains isolated from blood and cerebrospinal fluid. Pneumococcal nasopharyngeal carriage is seasonal with increased rates in winter. There is evidence that HIV infection increases the carriage rate of pneumococci in adults. Carriage can be reduced by the administration of conjugate pneumococcal vaccine. There is some evidence that the same total dose of antibiotic given in low daily doses for a longer period of time may increase the risk of selection of resistant strains, compared to a shorter course in a higher dose.

While b-lactam and multiple resistance in the pneumococcus now have a global distribution, recent emerging problems include very high levels of macrolide resistance in China and fluoroquinolone resistance in Canada and Hong Kong.

Mechanisms of antibiotic resistance

The molecular basis of penicillin resistance in the pneumococcus is the creation of mosaic genes of the penicillin-binding proteins (PBPs) 2X, 2B, and 1A by the transformation and horizontal transfer of DNA from related streptococcal species. The interaction of these altered gene products in the construction of the pneumococcal cell wall is not yet understood but resistant pneumococci make altered cell walls. In certain genetic backgrounds, alterations in the MurM protein may contribute to very high penicillin MICs of 8 to 16 µg/ml.Tetracycline resistance is caused by the acquisition of the tetracycline resistance determinant *tet*M or rarely *tet*O. Macrolide resistance is due to the acquisition of the methylating enzyme encoded by *erm*AM, by the acquisition of an efflux mechanism encoded by the *mef*E gene, by mutations in 23S RNA, or by mutations in the L4 riboprotein. Chloramphenicol resistance is due to the acquisition of the *cat* gene encoding chloramphenicol acetyltranferase. The acquisition of this gene has been shown to be associated with the integration and linearization of a plasmid into the pneumococcal genome. Rifampicin resistance is due to mutations in the *rpo*B gene, which may occur in one of two domains. Additional mechanisms of rifampicin resistance remain to be described. Trimethoprim resistance is caused by an altered *dhfr* chromosomal gene and is mediated by a single base mutation at position 100. Sulphonamide resistance is due to expansions in a small area of the *dhps* gene. While pneumococcal tolerance to vancomycin and other antibiotics has been described to be due to a deletion in the DNA encoding a two-component signalling system, resistance to vancomycin has not yet been described in this pathogen. Fluoroquinolone resistance is mediated by single base mutations in the *par*C gene and high levels of resistance may be obtained by a combination of mutations in the *par*C and *gyr*A genes. Mutations in the *par*E and *gyr*B genes may also contribute to pneumococcal resistance to the fluoroquinolones. The expression of a multidrug inhibitor protein *pmr*A can also confer low levels of fluoroquinolone resistance in the pneumococcus.

Serotype distribution

The distribution of pneumococcal serotypes causing invasive disease differs between children and adults. The so-called paediatric serotypes are serotypes 6A, 6B, 9V, 14, 19A, 19F, and 23F. All these serotypes commonly colonize young infants and most strains multiresistant to antibiotics belong to these serotypes. Serotypes 1 and 5, which are important invasive serotypes in developing countries, are rarely carried in the nasopharynx and thus are less commonly resistant to antibiotics—it is hypothesized that most selection for antibiotic resistance takes place in the nasopharynx. Serotype 1 is a relatively uncommon cause of invasive disease in adults in the United States but remains important in the adult population in most other parts of the world. The serotype 3 capsule is distinguished by its mucoid appearance and encapsulated type 3 pneumococci can often be distinguished from other pneumococci by visual inspection of an agar plate. There is evidence that the serotype distribution of pneumococci causing invasive disease in HIV-infected adults is different to that in adults with no HIV infection. The paediatric serotypes are more common in HIV-infected adults suggesting that these patients may have lost immunity to paediatric serotypes. Antimicrobial resistance has contributed to an increase in the proportion of paediatric serotypes causing invasive disease in both children and adults.

The capsular type is determined by a set of capsular genes. There are common genes in the regions of DNA flanking the capsular genes allowing homologous recombination to occur, and capsular switching of pneumococcal clones has been documented. In some parts of the world (especially Papua New Guinea) children are colonized at a young age with adult serotypes. The reason for the unusual distribution of serotypes in these children is not known. The distribution of serotypes in pneumococci isolated from the nasopharynx cannot be used to predict accurately the serotypes causing invasive disease because some of the most important invasive strains (notably serotypes 1 and 5) are, as mentioned above, rarely carried.

The global burden of pneumococcal disease

The global burden of pneumococcal disease is not known with certainty. However, the World Health Organization estimates that 1 million of the estimated 5 million deaths from pneumonia that occur in young children each year are caused by the pneumococcus. The incidence of pneumococcal bacteraemia in adults over 65 years of age is about 50/100 000 persons per year. The incidence of invasive pneumococcal disease is greatly increased in both adults and children by HIV infection. In South Africa, the incidence of invasive pneumococcal disease in children under the age of 2 years is 1844/100 000 per year.

Risk factors

Predisposing factors to invasive pneumococcal disease are listed in Table 1.

Diagnosis

The mainstay of the diagnosis of invasive pneumococcal disease is the isolation of the organism from a sterile site. Ten to 30 per cent of patients with pneumococcal pneumonia have a positive blood culture. Patients with suspected pneumococcal pneumonia should have an adequate volume of blood cultured because an association has been found in clinical practice between the isolation rate and the volume of blood cultured. Culture of cerebrospinal fluid remains the gold standard for the identification of *Streptococcus pneumoniae* as a cause of meningitis. Gram stain of cerebrospinal fluid and detection of capsular antigen are both useful diagnostic tests, especially when antibiotics have already been given. The definitive identification of a pneumococcal aetiology of otitis media or sinusitis requires the culture of the organism from the middle ear following tympanocentesis or the sinus following sinus puncture.

While a negative pneumococcal culture from the nasopharynx reduces the likelihood of a pneumococcal aetiology of otitis media, there is little clinical use associated

with a positive nasopharyngeal culture. Nasopharyngeal cultures can be used to predict the susceptibility of invasive isolates, but they are not of much diagnostic use in individual patients. The identification of Gram-positive, lanceolate-shaped diplococci in a sputum sample of good quality remains a useful diagnostic aid for pneumococcal pneumonia. In children the most sensitive method of diagnosing pneumococcal pneumonia is lung puncture. While this procedure carries a minimal morbidity in experienced hands, it is not widely used in clinical practice. Some promising data have recently been reported for the direct urinary detection of pneumococcal capsular polysaccharide antigens in adults. The polymerase chain reaction has not yielded a useful clinical advantage when used on blood, compared with blood culture as the gold standard. However, PCR can be used on cerebrospinal fluid to make a rapid diagnosis and can even be used directly to identify penicillin-resistant pneumococcal meningitis.

The microbiological identification of pneumococci rests on the demonstration of typical draftsman colonies producing an a-haemolytic reaction on blood agar. Pneumococci are generally optochin sensitive and are always bile soluble. Agglutination with polyvalent serum can be used to confirm their identity.

Susceptibility testing

Disc susceptibility to a 1µg oxacillin disc is the best predictor of penicillin susceptibility of the pneumococcus. Pneumococci are defined as susceptible to penicillin when they are inhibited by a minimum inhibitory concentration (**MIC**) of less than 0.1 µg/ml. A MIC of 0.1 to 1 µg/ml defines intermediate susceptibility and greater than 1 µg/ml resistance to penicillin. Resistance to amoxicillin is defined as an MIC greater than 2 µg/ml. Susceptibility to the third-generation cephalosporins, cefotaxime or ceftriaxone is defined by a MIC less than 1 ug/ml. Intermediate susceptibility and resistance to these agents are defined by MICs of 1 µg/ml and more than 1 µg/ml, respectively. The differentiation of fully resistant from intermediately penicillin-resistant pneumococci and the identification of cephalosporin-resistant pneumococci cannot be done by disc testing. The definitive method for susceptibility testing of pneumococci is by an agar dilution or micro-broth dilution method. However, many laboratories use the E test to identify penicillin- and cephalosporin-resistant pneumococci. Susceptibility testing to trimethoprim sulphamethoxazole requires the use of lysed blood. Susceptibility to most other agents is predicted by routine disc methods, although there are currently no phenotypic methods available to detect first-step mutants that lead to fluoroquinolone resistance.

Clinical features

The clinical features of pneumococcal infection are described most conveniently in relation to the main clinical syndromes that can be caused by pneumococci. However, these syndromes are not mutually exclusive and many patients with pneumococcal disease have more than one clinical manifestation of the infection, for example pneumonia and meningitis.

Pneumonia

Pneumonia is one of the most common manifestations of pneumococcal disease.

Symptoms

In a typical case of pneumococcal pneumonia, the onset of illness is sudden, although there may be a history of a recent upper respiratory tract infection. Fever is usually the first symptom and it is frequently accompanied by rigors. The patient feels ill, anorexic, and weak. Headache and myalgia may be severe.

Chest pain usually appears during the course of illness although it may not be present on initial presentation. The pain, which results from involvement of the parietal pleura, is sharp and stabbing and is aggravated by deep inspiration or coughing. The patient may try to obtain relief by splinting the affected side of his chest with his hands or lying on the affected side. If the diaphragmatic pleura is involved, pain may be referred to the shoulder or to the abdomen.

Cough may be absent at the onset of the illness but, in most patients, it becomes a prominent symptom. Cough is initially non-productive and painful. Subsequently it becomes productive of a blood-tinged, 'rusty' sputum. Finally, the sputum becomes frankly purulent.

Among young children and elderly people, pneumococcal pneumonia may present less dramatically. The mothers of young children with pneumonia usually give a history of fever and cough and the mother may have noticed that the child has rapid respiration. Elderly patients and those who are immunocompromised may have little or no fever and few respiratory symptoms. In such patients, general malaise and confusion may be the presenting symptoms. The classic features of pneumococcal pneumonia may also be modified by prior antibiotic treatment.

Physical signs

Adult patients with lobar pneumonia are usually febrile and toxaemic. The rectal temperature may be as high as 40°C. Oral temperature may be lower because of hyperventilation. When the patient is first examined, no abnormal physical signs may be detected in the respiratory system. Later the classic signs of lobar consolidation may appear. The patient's breathing becomes rapid and distressed, and the nostrils may dilate on inspiration. Cyanosis may be present as a result of diminished alveolar ventilation or shunting of desaturated blood through the consolidated lung. Chest movement is diminished on the affected side. A dull note is obtained on percussion over the affected lobe. On auscultation, bronchial breathing is sometimes detected, fine crepitations are frequent, and a pleural rub may be heard.

General examination usually shows tachycardia and an atypical systolic murmur may be detected, as in any patient with a high fever. Examination of the abdomen may show some distension or, when the diaphragmatic pleura has been involved, upper abdominal tenderness and guarding. Jaundice may be present. The patient, especially if elderly, may be confused.

The classic signs of lobar consolidation are found infrequently in infants with pneumococcal pneumonia, although some auscultatory abnormalities, such as crepitations, can usually be detected. In young children, the most prominent features of pneumococcal pneumonia are usually a raised respiratory rate, chest-wall indrawing (Fig. 3), and nasal flaring.

Fig. 3 Severe lower-chest indrawing in a child with pneumococcal pneumonia (by courtesy of Dr Alice Greenwood).

Investigations

A polymorphonuclear neutrophil leucocytosis is usually present; a white-cell count of 15×10^9/l or more is found in about three-quarters of cases and counts as high as 40×10^9/1 may occur. A low white-cell count is associated with a poor prognosis. There may be a reticulocytosis. Both conjugated and unconjugated bilirubin levels are raised in jaundiced patients, and serum transaminases may be elevated. The Po_2 is often diminished, and measurement of the degree of hypoxaemia gives an

important indication of the severity of the infection, but the PCO_2 is normal unless terminal respiratory failure occurs.

The sputum of untreated patients usually shows large numbers of Gram-positive diplococci, together with polymorphonuclear neutrophils, and culture is frequently positive for pneumococci. However, in industrialized countries, where many patients have received partial treatment before presentation at hospital, sputum microscopy is positive in only about one-quarter of patients and culture positive in only about a half. Blood culture is positive in 10 to 30 per cent of patients.

Radiographs of the chest usually show homogeneous opacification of the affected part of the lung, but may appear normal on first presentation. Posteroanterior and lateral views are required to make an accurate diagnosis of the site of the infection. A small pleural effusion can be seen in some patients. Pneumococci can cause either segmental or lobar consolidation, or patchy shadowing. The latter is encountered more frequently in children. The lower lobes are affected more frequently than the upper. In about one-third of patients, more than one lobe is involved.

Differential diagnosis

The initial febrile phase of acute pneumococcal pneumonia cannot be differentiated from that of any other acute febrile illness. Once the characteristic respiratory symptoms and signs have appeared, a diagnosis of acute pneumonia can usually be made on clinical grounds, but chest signs may be absent when the patient is first seen. In developing countries, most cases of pneumococcal pneumonia in young children are diagnosed and treated by paramedical primary health-care workers, who may have only limited diagnostic skills. For this reason the World Health Organization (**WHO**) has devised a simple diagnostic scheme, based predominantly on measurement of the respiratory rate and on observation of lower chest-wall indrawing, to help primary health-care workers determine which children with acute respiratory-tract infections probably have pneumonia and require antibiotic treatment (Table 2). This scheme has played an important part in the rationalization of the management of acute respiratory infections in developing countries, but other severe infections, including malaria, can give rise to cough and a raised respiratory rate in young children, thus fulfilling the diagnostic criteria for pneumonia. For this reason an integrated approach to the management of childhood illness by primary health-care workers is now advocated by WHO and UNICEF. An algorithm has been developed which gives guidance on diagnosis and management; it is now being used in many developing countries.

Two important pulmonary conditions that may be confused with acute bacterial pneumonia in adult patients are infarction and atelectasis. Rigors and a high fever favour a diagnosis of pneumonia as opposed to one of infarction; a very sudden onset of symptoms and frank haemoptysis favour a diagnosis of infarction. Pulmonary atelectasis, resulting from the aspiration of mucus, may give rise to symptoms and signs that are very similar to those of pneumonia. Fever and signs of toxaemia are usually less marked in patients with atelectasis than in those with pneumonia unless the collapsed area of lung has become infected. In elderly patients, heart failure with atypical pulmonary oedema may sometimes mimic pneumococcal pneumonia.

Occasionally, subdiaphragmatic lesions such as cholecystitis, a subphrenic abscess, or an amoebic liver abscess cause a clinical picture that mimics that of lower-lobe pneumonia. Conversely, lower-lobe pneumonia, by producing abdominal pain and guarding, may suggest the diagnosis of an acute abdominal condition such as a perforated peptic ulcer, acute cholecystitis, or appendicitis.

Pneumococcal pneumonia can usually be differentiated from viral pneumonias or pneumonia caused by *Mycoplasma pneumoniae* because of its sudden onset, associated severe toxaemia, and accompanying polymorphonuclear neutrophil leucocytosis, but differentiation from other forms of acute bacterial pneumonia cannot be made without the aid of microbiological investigations. Klebsiella pneumonia, staphylococcal pneumonia, and legionnaires' disease may all produce a similar clinical picture. Confusion, signs of multisystem damage, lymphopenia, or a low serum sodium should raise the possibility of legionnaires' disease. In HIV infected patients the differential diagnosis also includes patients infected with *Pneumocystis carinii* and mycobacterial species.

Course and prognosis

Untreated patients who survive long enough to make specific anticapsular polysaccharide antibody recover spontaneously by crisis, or by a more gradual lysis, 7 to 10 days after the onset of their illness. Without treatment the mortality of acute pneumococcal pneumonia is high, especially when bacteraemia is present. Among patients treated promptly with antibiotics, overall mortality is about 5 per cent, but mortality remains as high as 30 per cent in patients with bacteraemia despite antibiotic treatment. Mortality is highest among the elderly and the very young, and among those with an associated underlying illness, such as cirrhosis, alcoholism, or heart disease. HIV infection probably increases mortality in children but this has not been found in all studies. Infection with certain pneumococcal serotypes, involvement of more than one lobe of the lung, bacteraemia, leucopenia, and jaundice are all bad prognostic signs. Most deaths from treated pneumococcal pneumonia occur within the first few days of admission to hospital. It is often difficult to establish an exact cause of death in such patients —peripheral circulatory collapse, cardiac arrhythmias, and respiratory failure are some of the contributory factors.

Complications of pneumococcal lobar pneumonia result from local or lymphatic spread of bacteria to adjacent pleura or pericardium, producing pleural or pericardial effusions, or from bacteraemic spread to meninges and other distant foci. The likelihood of one of these infective complications developing is reduced, but not completely abolished, by prompt treatment with antibiotics. Pneumococcal pneumonia may precipitate congestive cardiac failure in elderly patients and can precipitate acute dilatation of the stomach or paralytic ileus. Herpes labialis is a common accompaniment of the infection.

Pleural effusion and empyema

A large pleural effusion or an empyema develops during treatment in a small percentage (2 to 5 per cent) of patients with established pneumococcal pneumonia. Other patients present with the clinical features of pleural effusion without any preceding symptoms of pneumonia.

Symptoms

Some patients give a history suggestive of a previous parenchymatous lung infection. A history of days or weeks of fever, malaise, anorexia, and marked weight loss is often obtained. Fever may be hectic and accompanied by rigors and episodes of profuse sweating. Patients with a large pleural effusion are breathless and they may complain of dull pain on the affected side. A productive cough is unusual unless a bronchopleural fistula is present.

Physical signs

General examination shows persistent fever and tachycardia. The patient may look toxaemic and there may be signs of recent weight loss. Examination of the chest usually shows the characteristic signs of a pleural effusion—diminished chest movement, dullness of percussion, and diminished breath sounds over the accumulated fluid. The chest wall overlying an empyema may be tender.

Investigations

A persistent polymorphonuclear neutrophil leucocytosis is nearly always present. Radiographs or ultrasonography may be very helpful in localizing a loculated effusion. On aspiration, turbid fluid or thick pus is obtained, which contains pneumococci and degenerate white cells. If antibiotics have been given it may not be possible to culture pneumococci, but pneumococcal antigen can usually be detected by immunological assays.

Differential diagnosis

Association of persistent pyrexia and leucocytosis with abnormal chest signs indicates a chronic pulmonary infection. Absence of copious, purulent sputum differentiates the condition from a lung abscess. Differentiation from tuberculosis may be difficult on clinical grounds alone. Diagnosis of an empyema is confirmed by the aspiration of pus from the pleural cavity. Repeated needling with a wide-bored needle, preferably under ultrasound control, may be needed to find a loculated empyema. Pleural biopsy may provide diagnostic histology.

Course and prognosis

Untreated, an empyema may rupture through the chest wall (empyema necessitas) or rupture into a bronchus causing a bronchopleural fistula. Even when pus is

aspirated and healing achieved, subsequent fibrosis and calcification may seriously restrict expansion of the underlying lung.

Pericardial effusion and empyema

Pneumococci may spread from an infected lower lobe to produce pericarditis. Pericarditis is clinically silent in some patients; in other patients it is manifest only as a transient pericardial rub or as an abnormal electrocardiogram. However, occasionally pericardial involvement is the dominant feature of a pneumococcal infection. Only a proportion of such patients give a history suggestive of an initial acute respiratory-tract infection.

Symptoms

Patients with a pneumococcal pericardial empyema usually give a history of several days, or even weeks, of persistent fever, malaise, anorexia, and weight loss. They may complain of dull or pleuritic central chest pain and they may have noted swelling of the ankles or of the abdomen.

Physical signs

Many patients with a pneumococcal pericardial empyema are critically ill by the time that they reach hospital. They are febrile and toxaemic. There may be signs of severe pericardial tamponade—a rapid, small-volume pulse, pulsus paradoxus, a low blood pressure, elevation of the jugular venous pressure with a further increase during inspiration, and peripheral oedema and ascites. Percussion of the chest may show some enlargement of the area of cardiac dullness but this is an unreliable clinical sign. The heart sounds are usually faint and, in some patients, a pericardial rub is heard.

Investigations

A peripheral blood polymorphonuclear neutrophil leucocytosis is present and blood culture may be positive for pneumococci. A chest radiograph may show globular enlargement of the heart and there may be radiological evidence of an associated lung infection. An ultrasonographic examination may help to define the best site for drainage. The electrocardiogram shows low-voltage potentials and S-T elevation or depression may be present. On aspiration of the pericardium, turbid fluid or thick pus is obtained from which pneumococci can be isolated or in which pneumococcal antigen can be detected.

Differential diagnosis

Detection of the signs of pericardial tamponade in a patient who is febrile and toxaemic should suggest a diagnosis of pericardial empyema. The condition may be confused with tuberculous constrictive pericarditis, but patients with the latter condition usually have a longer history than patients with a pneumococcal pericardial empyema and are less toxic. Staphylococci and, rarely, other pyogenic bacteria can produce a similar clinical picture to that of pneumococcal pericardial empyema. Diagnosis of a pericardial empyema is confirmed by ultrasound and by pericardial aspiration. A pneumococcal pericardial empyema is a medical emergency and, following ultrasonographic examination if this is available, pericardial aspiration should be undertaken, if necessary at more than one site, as soon as this diagnosis is seriously suspected.

Course and prognosis

Pneumococcal pericardial empyema is a serious condition with a high mortality, even in treated patients. Patients who survive the initial episode may develop constrictive pericarditis within weeks or months of their acute illness.

Otitis media

Otitis media is probably the most common form of pneumococcal infection. The condition is seen most frequently in young children but it may also affect adults.

Symptoms

The onset of an attack of acute otitis media is sudden, although there may be a history of a recent upper respiratory-tract infection. Fever and severe pain in the ear are the usual presenting complaints in adults and older children, and patients may complain of deafness and tinnitus. Fever, crying, and extreme irritability are the usual features of the condition in young children, in whom febrile convulsions may occur.

Physical signs

On examination of the affected ear, the tympanic membrane is seen to be red and swollen, and it may bulge outwards into the external ear. If perforation has occurred, the external ear may be full of pus and a ragged hole may be seen in the tympanic membrane. The affected ear is usually partially deaf. In children, meningism may be present; this must be differentiated from meningitis by lumbar puncture.

Laboratory findings

A polymorphonuclear neutrophil leucocytosis is usually found. If the drum has ruptured, pneumococci may be found in the purulent discharge present in the external ear but contaminants are likely to be present also.

Differential diagnosis

A clinical diagnosis of otitis media is rarely difficult provided that the ears of all febrile and irritable children are examined carefully. A tympanogram usually shows a characteristic pattern. The aetiology of the condition can be established by examination of fluid obtained from the middle ear with a fine needle. This technique, widely practised in some countries but not in others, may become increasingly useful as determination of the antibiotic sensitivity pattern of pneumococci becomes an essential requirement for optimum treatment of pneumococcal infections.

Course and prognosis

Prompt treatment is usually followed by a rapid and complete resolution of the infection. However, some patients, especially those in whom rupture of the drum has occurred, are left with partial conductive deafness. When untreated, pneumococcal otitis media can give rise to a chronic discharging ear requiring prolonged and complicated treatment. Spread of the infection posteriorly may result in acute mastoiditis, and spread of the infection upwards can cause pneumococcal meningitis and/or a cerebral abscess.

Pneumococcal meningitis (see also Chapter 24.14.1)

Pneumococcal meningitis may follow damage to the base of the skull, and it can occur as a complication of pneumococcal otitis media or pneumococcal pneumonia. However, many patients with this condition, the proportion varying from series to series, present with the clinical features of acute pyogenic meningitis and have no features to suggest a primary focus of pneumococcal infection.

Symptoms

Fever and headache are the usual presenting symptoms of pneumococcal meningitis. Headache usually comes on gradually over a few hours; it is generalized and may be very severe. Nausea, backache, and photophobia may develop, and convulsions may occur. Confusion may be the most prominent symptom in elderly patients, and failure to feed the first symptom in infants.

Physical signs

Patients with pneumococcal meningitis are febrile and toxaemic. Neck stiffness and a positive Kernig's sign are usually found in adults and in older children. Impairment of consciousness is often present, which varies in severity from drowsiness and confusion to deep coma. Bradycardia and hypertension may indicate the presence of raised intracranial pressure, but papilloedema is rarely seen. Bulging of the anterior fontanelle may be present in infants. Cranial nerve palsies, most frequently of the VIth or of the IIIrd cranial nerve, may be found on presentation and, occasionally, other peripheral localizing neurological signs are present.

An associated pneumococcal lesion, such as otitis media or pneumonia, may be detected. Petechiae are rarely seen. Herpes labialis may be present.

Laboratory findings

A peripheral blood polymorphonuclear neutrophil leucocytosis is usually found and a positive blood culture may be obtained.

Examination of the cerebrospinal fluid shows a turbid fluid, which usually contains an increased number of cells and many bacteria. Most of the leucocytes are polymorphonuclear neutrophils. Cerebrospinal fluid bacterial counts are often very high in patients with pneumococcal meningitis, on average 10 times higher than in patients with meningococcal meningitis. Leucocytes are present in only small numbers in the cerebrospinal fluid of some patients; in such instances the fluid may still be turbid because of the presence of numerous bacteria. The protein level in cerebrospinal fluid is increased and its glucose level decreased below that of blood. Gram stain and culture are usually positive for pneumococci.

Differential diagnosis

It is not usually difficult to establish a clinical diagnosis of pyogenic meningitis in adults and older children. However, problems may arise in the very young and in the very old; signs of meningeal irritation may be absent in both these groups of patients. Fever and irritability may be the only clinical signs of pneumococcal meningitis in an infant. The appearance of confusion may be the only sign indicating involvement of the meninges in an elderly patient with pneumococcal pneumonia. An adverse change in the psychological or neurological state of an elderly patient with pneumonia is an indication for lumbar puncture.

On clinical grounds, pneumococcal meningitis cannot be differentiated with certainty from other forms of pyogenic meningitis. An associated ear infection or pneumonia favours the diagnosis of pneumococcal infection but is not diagnostic. If petechiae are found, meningococcal meningitis is more likely. Bacteriological diagnosis of pneumococcal meningitis is confirmed by examination of the cerebrospinal fluid.

Course and prognosis

The prognosis of patients with pneumococcal meningitis is poor. Many patients make no response to treatment, their conscious level deteriorates progressively, and they die within the first 24 to 48 h after their admission to hospital. Other patients make some initial response to treatment but then relapse, their conscious level deteriorates, and new neurological signs appear. This deterioration may be due to the collection of pus in the extradural space or brain but, more usually, follows a vascular occlusion. Patients who deteriorate after an initial clinical improvement must be fully investigated to exclude the presence of a space-occupying lesion. The clinical course of survivors of the early phase of pneumococcal meningitis is often stormy, being complicated by conditions such as bedsores, pneumonia, and venous thrombosis. It has been estimated that over one-half of survivors from pneumococcal meninigitis are left with some intellectual impairment or residual neurological disability such as deafness or partial hemiplegia. Small children who survive may develop hydrocephalus. Relapses may occur when treatment is stopped. Mortality figures for pneumococcal meningitis vary from series to series but in industrialized countries the true mortality from pneumococcal meningitis is probably around 30 per cent. In developing countries, mortality figures of around 50 per cent have been found consistently. Impairment of consciousness on admission to hospital, associated pneumonia, a low white-cell count, and a high bacterial count in the cerebrospinal fluid are all poor prognostic features. Death is almost inevitable in patients who are in deep coma at the time they are admitted to hospital.

Why the prognosis of pneumococcal meningitis is so poor is uncertain. Although there is little difference in the clinical features of patients with pneumococcal or meningococcal meningitis on presentation at hospital, death is at least five times more likely in a patient with pneumococcal meningitis than in a patient with meningococcal meningitis regardless of the level of patient care available. Vascular damage, rapid multiplication of bacteria, and defective leucocyte function have all been suggested as possible causes for the poor outcome of patients with pneumococcal meningitis, but the reasons for the very poor prognosis of patients with this condition remain a mystery.

Other clinical syndromes

The pneumococcus is an important cause of bacterial sinusitis. The likelihood of a bacterial aetiology of acute sinusitis is increased if the duration of symptoms exceeds 7 days.

Acute, fulminating septicaemia is a rare form of pneumococcal infection and is encountered most frequently in patients without a spleen or in those who are immunocompromised in some way. A sudden onset of fever, peripheral circulatory collapse, and bleeding (purpura fulminans) are the usual presenting features of this condition, which is indistinguishable from other forms of overwhelming bacterial septicaemia. Leucopenia is usually found. Bleeding is due to disseminated intravascular coagulation. The mortality of this condition is very high, even when treatment is started promptly. A milder form of bacteraemia is sometimes encountered in children who present with fever or febrile convulsions without any obvious focus of pneumococcal infection (occult bacteraemia).

Acute endocarditis may complicate pneumococcal septicaemia but this condition is now encountered only rarely. Healthy heart valves, especially the aortic valve, may be attacked and rupture of the aortic valve may occur, producing severe aortic incompetence. Emboli derived from cardiac vegetations may reach the brain and other organs. Progressions of the cardiac lesions may be very rapid and the prognosis of this condition is poor. Valve replacement may be necessary for patients who survive the initial episode.

During the course of pneumococcal septicaemia, with or without endocarditis, bacteria may reach many sites where they can multiply to produce a purulent lesion. Pneumococcal arthritis, verterbral osteomyelitis, ophthalmitis, and orchitis may be produced in this way. The pneumococcus has been rarely associated with the toxic shock syndrome and with the haemolytic–uraemic syndrome.

Pneumococcal peritonitis is an uncommon condition that is encountered most frequently in patients with the nephrotic syndrome or cirrhosis of the liver, conditions frequently resulting in ascites and generalized impairment in immunity. The condition has been described also in healthy young girls, perhaps as a complication of pelvic infection, and occasionally in neonates. The condition is characterized by a sudden onset of fever, and abdominal pain and tenderness. The ascitic fluid is turbid and contains polymorphonuclear neutrophils and pneumococci. The general features of an acute infection may not be so obvious in patients with cirrhosis, and peritonitis must be considered as a possible diagnosis in any patient with this condition whose clinical state shows a sudden deterioration. The prognosis of pneumococcal peritonitis is poor in patients with a serious underlying illness.

Treatment

In the era of the antibiotic-resistant pneumococcus, the appropriate treatment of pneumococcal disease is determined by pharmocodynamic principles. These suggest that successful b-lactam therapy of invasive pneumococcal disease requires drug levels at the site of infection that exceed the MIC of the organism for at least 50 per cent of the dosing interval. The same principle applies to therapy with macrolides. Fluoroquinolones, azalides, and aminoglycosides exert a concentration-dependent killing effect on pneumococci. The best predictor of an appropriate outcome is a measure that includes both the peak concentration and time above the MIC. This is best described by the area under the drug concentration curve over 24 h (**AUC24**). The ratio of the AUC_{24} to the MIC (AUIC) predicts optimal efficacy at a value of [3]125. The application of these principles probably applies to all invasive diseases caused by pneumococci. Their application to the treatment of otitis media has, however, been best studied to date.

Pneumonia

Clinical studies have shown that the drug level most predictive of outcome in pneumonia is the serum concentration. Measurement of these concentrations suggests that b-lactam therapy of pneumonia will successfully treat penicillin-resistant pneumococcal pneumonia. When the drug is given in high dose intravenously is it likely that pneumonia caused by pneumococci with MICs up to 4 µg/ml will respond. A number of studies in both adults and children have shown that the most important predictors of outcome in the management of pneumococcal pneumonia are severity of disease and the presence of underlying disease. b-Lactam susceptibility does not affect outcome when adequate doses of intravenous b-lactam drugs are used to treat the infection. The optimal drugs for treating pneumococcal pneumonia are penicillin or amoxicillin. When there is a high index of suspicion of a pneumococcal aetiology in a patient with pneumonia, intravenous management with penicillin, ampicillin, or amoxycillin is appropriate (amoxycillin has higher antipneumococcal activity than ampicillin). When the aetiology of pneumonia is unclear, empirical management requires broader-spectrum cephalosporin therapy such as treatment with cefuroxime. Cefotaxime or ceftriaxone are more active against pneumococci and ought to be effective against cephalosporin-resistant pneumococcal pneumonia caused by pneumococci with cephalosporin MICs of 1 to 2 µg/ml. The clinical relevance of macrolide resistance remains unclear, but it is likely that macrolide treatment of pneumonia caused by pneumococci with MICs in a range of 1 to 2 µg/ml (*mef*E-mediated resistance) will respond to intravenous macrolide therapy. Pharmocodynamic principles suggest that higher levels of macrolide resistance are likely to result in clinical failure. Newer fluoroquinolones are under development for the management of pneumococcal pneumonia. This class of agent has, until recently, had marginal activity against the pneumococcus, but newer agents with enhanced antipneumococcal activity may be useful for the management of highly penicillin-resistant pneumococcal pneumonia in adults. There is currently no indication for the addition of vancomycin to the management of patients with pneumococcal pneumonia. There are few data on the appropriate antibiotic for oral management of antibiotic-resistant pneumococcal pneumonia and it is in this situation that very high doses of amoxicillin or the new fluoroquinolones may have their most important role.

Otitis media

While the use of antibiotics for the treatment of otitis media remains controversial, pneumococcal otitis media will resolve in a minority of cases only with appropriate antibiotic therapy. Bacterial eradication from middle ear fluid correlates well with clinical response in the management of otitis media. Oral cephalosporins with poor antipneumococcal activity are inferior to amoxicillin in their ability to eradicate bacteria from the middle ear. The choice of appropriate antibiotic therapy for the management of otitis media requires evaluation of local studies in which initial and follow-up tympanocentesis has been performed. Studies with clinical endpoints lack the ability to differentiate between highly active and poorly active agents. Current guidelines suggest that high doses of amoxicillin (90 mg/kg.day) represent the best available oral therapy for pneumococcal otitis media. Patients in whom this therapy has failed require tympanocentesis to document the cause of the failure. Should a highly penicillin-resistant pneumococcus be isolated, the appropriate therapy is intravenous or intramuscular ceftriaxone for 3 days.

Meningitis

The mainstays of therapy of pneumococcal meningitis in developing countries, namely penicillin and/or chloramphenicol, can no longer be relied upon to treat this disease, given the global epidemic of b-lactam-resistant strains. In a study from South Africa in which the usefulness of this combination was tested against penicillin-resistant, but chloramphenicol-susceptible, strains there was a poorer outcome in patients with penicillin-resistant disease compared with those with penicillin-susceptible pneumococcal meningitis. Pharmocodynamic principles suggest that penicillin and chloramphenicol are an inadequate form of therapy for even intermediately penicillin-resistant pneumococcal meningitis. The drugs of choice for the management of pneumococcal meningitis are thus cefotaxime (300 mg/kg.day divided into 3 or 4 doses) or ceftriaxone (100 mg/kg.day divided into 2 doses). Adults should receive full doses of antibiotic. In communities where, for financial reasons, only limited amounts of cephalosporins are available, patients with pneumococcal meningitis should be given a high priority for treatment with these drugs because of the severity of this condition. In some parts of the world, strains with intermediate or full resistance to cefotaxime or ceftriaxone have emerged. In these countries the appropriate initial empiric management of meningitis includes the addition of vancomycin (60 mg/kg.day divided into 4 doses). Amongst the other available cephalosporins, ceftazidime should not be used for the management of pneumococcal meningitis. Cefepime has activity similar to that of cefotaxime and ceftriaxone, and may be used. Cefpirome has enhanced antipneumococcal activity but there are no clinical studies of the efficacy of this agent in the management of meningitis. Amongst the carbapenems, imipenem when used for meningitis is associated with an increased incidence of seizures. Two studies have demonstrated that patients treated with meropenem were not at increased risk of seizures on therapy compared with patients treated with cefotaxime. While newer fluoroquinolones are under clinical trial for the management of meningitis, there are insufficient data to recommend their use at this time.

The use of dexamethasone immediately prior to or simultaneously with the administration of cefotaxime, or ceftriaxone, appears to improve the outcome of pneumococcal meningitis in children. The use of this agent in adult pneumococcal meningitis is unproven and its use in conjunction with penicillin and/or chloramphenicol is controversial.

Other infections

There is little published information on the clinical impact of antibiotic resistance on the management of pneumococcal infections other than pneumonia, meningitis, and otitis media. The pharmocodynamic principles outlined above may be useful in guiding empiric therapy for these conditions.The principles of treatment for pneumococcal sinusitis are the same as those of otitis media. While pneumococci are intrinsically less susceptible than viridans streptococci to aminoglycosides, the addition of an aminoglycoside may have a synergistic effect on the bacterial killing rate in the treatment of pneumococcal endocarditis.

The minimum duration of therapy for invasive pneumococcal infections is under review, but the current usual duration is shown in Table 3.

Chemoprophylaxis

Children at particularly high risk of pneumococcal disease (such as those with sickle cell disease or nephrotic syndrome, or post-splenectomy) should receive regular oral penicillin prophylaxis for the first 5 years of life. The value of prophylaxis after this age is unproven, but should be considered in those who may not have responded to vaccine (such as patients with recurrent bacteraemia). Vaccination of high-risk children should be given at 2 years of age with the 23 valent vaccine. Conjugate vaccine is recommended in infancy, followed by a booter with the 23 valent vaccine at 2 years of age.

Patients who have undergone splenectomy should be educated about the risks of bacteraemia, and have prompt initiation of antibiotic therapy for febrile episodes.

Immunity and vaccines

The basis of immunity to invasive pneumococcal disease is thought to be the development of serotype-specific capsular antibodies of the IgG_2 subclass. These antibodies stimulate opsonophagocytosis, a process that is facilitated by the binding of complement. Previous studies of specific capsular polysaccharide antibody levels in adults have used a radioimmunoassay which detected both specific antibody and antibody to C polysaccharide. The use of C polysaccharide absorption in ELISA assays has helped to make these antibody assays more specific. The IgG response to capsular polysaccharide is poorly developed in young infants and the affinity and opsononophagocytic activity of these antibodies is reduced in older patients and patients infected with HIV.

The first attempt to develop a multivalent pneumococcal vaccine was made by Sir Spencer Lister at the South African Institute for Medical Research in 1917. Robert Austrian pioneered the development of multivalent capsular polysaccharide vaccines in the 1970s when these vaccines were shown to be effective in reducing the incidence of invasive pneumococcal disease in otherwise healthy adult gold miners. The vaccine also reduced the incidence of lobar pneumonia in miners. There is indirect evidence of the efficacy of pneumococcal polysaccharide vaccine against bacteraemia in high-risk groups of adults who have received this vaccine. The most compelling evidence comes from indirect cohort studies. There is less evidence of the efficacy of the vaccine in immunocompromised patients. Protection against pneumonia in the elderly has not been demonstrated in prospective randomized trials. Current indications for the use of the vaccine in the United States are listed in Table 4. Use of pneumococcal polysaccharide vaccines has generally been low in adults in industrialized countries, although in some there has been a recent improvement. Following the results of a trial in Uganda, which showed an increased incidence of invasive pneumococcal disease in HIV-positive subjects who had received polysaccharide vaccine, no firm recommendation about the use of this vaccine in patients who are not on antiretroviral therapy can be made before further studies are performed.

Protein vaccines are an attractive option for developing vaccines that are not serotype specific. The most promising vaccines are based on PspA and PsaA proteins which are conserved among pneumococci of various serotypes. These vaccines are in early clinical trial in humans.

In contrast to polysaccharide or protein vaccines, pneumococcal conjugate vaccines contain polysaccharide chemically linked to protein. These conjugate vaccines are immunogenic in young infants and induce immunological memory. They also reduce nasopharyngeal carriage of pneumococci of vaccine serotype, although in

some studies this has been accompanied by an increase in the carriage rate of pneumococci of non-vaccine serotype. This replacement may be a function either of exogenous infection with non-vaccine strains or simply the eradication of the dominant strain by the vaccine thus unmasking subdominant strains. Pneumococcal conjugate vaccines have been shown to reduce the carriage of antibiotic-resistant pneumococci.

A number of efficacy trials of these vaccines are under way. Results from a trial conducted in California recently became available; over 90 per cent efficacy against invasive pneumococcal disease was found and significant protection, although less marked, was obtained against radiographic pneumonia and recurrent otitis media. Because of serotype replacement the overall impact on pneumococcal otitis media was reduced further. In another study, undertaken in Finland, a conjugate vaccine reduced by 57 per cent the incidence of serotype-specific otitis media. These vaccines thus have the potential greatly to reduce mortality and morbidity from invasive pneumococcal disease and pneumonia in young infants and their availability in the first decade of the new millennium may be one of the most important public health interventions of this decade.

Further reading

Arason VA *et al*. (1996). Do antimicrobials increase the carriage rate of penicillin resistant pnuemococci in children? Cross-sectional prevalence study. *British Medical Journal* **313**, 387–91.

Chen DK *et al*. (1999). Decreased susceptibility of *Streptococcus pneumoniae* to fluoroquinolones in Canada. *New England Journal of Medicine* **341**, 233–9.

Dowson CG *et al*. (1989). Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae. Proceedings of the National Academy of Sciences, USA* **86**, 8842–6.

Friedland IR, Klugman KP (1992). Failure of chloramphenicol therapy and penicillin-resistant pneumococcal meningitis. *Lancet* **339**, 405–8.

Gillespie SH, Balakrishnan I (2000). Pathogenesis of pneumococcal infection. *Journal of Medical Microbiology* **49**, 1057 – 67.

Greenwood BM (1999). The epidemiology of pneumococcal infection in children in the developing world. *Proceedings of the Royal Society of London, Series B* **354**, 777–85.

Musher DM (1992). Infections caused by *Streptococcus pneumoniae*: the clinical spectrum pathogenesis, immunity and treatment. *Clinical Infectious Diseases* **14**, 801–9.

Pallares T *et al*. (1995). Resistance to penicillin and cephalosporin and mortality from severe pneumococcal pneumoniae in Barcelona, Spain. *New England Journal of Medicine* **333**, 474–80.

Watson DA *et al*. (1993). A brief history of the pneumococcus in biomedical research: a panoply of scientific discovery. *Clinical Infectious Diseases* **17**, 913–24.

7.11.4 Staphylococci

S. J. Eykyn

Although both Koch and Pasteur made observations on coccal organisms, the polyglot Scottish surgeon Alexander Ogston first associated cluster-forming cocci with abscesses. He presented his findings in German to the Surgical Congress in Berlin in 1880 and his classic paper *Über Abscesse* was published the same year. The Professor of Greek at Aberdeen University suggested the name 'staphylococcus' for the organism (*staphyle*—bunch of grapes; *kokkos*—berry) to distinguish it from the chain-forming streptococci. Rosenbach divided the genus *Staphylococcus* into *Staphylococcus aureus* and *S. albus*. Ogston's coccus was of course *S. aureus*.

Taxonomy

Staphylococci are Gram-positive cluster-forming cocci. There are some 32 recognized species of the genus *Staphylococcus* but only about half are of human origin (Table 1). Staphylococci are skin commensals of mammals and birds and some species, particularly *S. aureus*, are important human pathogens. In the clinical laboratory *S. aureus* is distinguished from other staphylococci by its ability to coagulate plasma. The slide coagulase test detects cell-associated clumping factor (bound coagulase), which reacts with fibrinogen to cause aggregation of the organisms. Commercial kits are used for this test and some also detect protein A, present in most strains of *S. aureus*. Occasional strains do not produce clumping factor or protein A, and certain other species of staphylococci produce clumping factor, hence the gold standard for the identification of *S. aureus* in the laboratory is the tube coagulase test in which staphylococci are mixed with plasma in a test tube. This detects extracellular coagulase (free coagulase), which activates prothrombin and initiates clot formation. The slide coagulase test is used to screen organisms, whereas the tube test is confirmatory and of more taxonomic significance. Other useful screening tests for *S. aureus* are the detection of DNAase activity and fermentation of mannitol, but neither is as reliable as the tube coagulase test.

Many clinical laboratories report any coagulase-negative staphylococcus other than *S. saprophyticus* as *S. epidermidis* without formal speciation and some still refer to these bacteria as '*Staph. albus*'. Availability of commercial identification kits has enabled speciation of coagulase-negative staphylococci in the routine laboratory though this is seldom undertaken routinely. Most clinical isolates are *S. epidermidis* (*sensu stricto*) or *S. saprophyticus* but several other species such as *S. lugdunensis* can occasionally be important pathogens.

Typing

Epidemiological studies of *S. aureus* infection, and increasingly these concern methicillin-resistant strains (**MRSAs**), require typing methods to distinguish between epidemic and endemic strains. Typing can also confirm the correlation between specific staphylococcal infections and a particular type of the organism. In many countries including the United Kingdom such studies still rely on bacteriophage typing. Phage typing has been organized internationally since 1953. The basic international set of phages consists of 23 phages (Table 2). There are four major phage groups—I, II, III, and V—and staphylococci may be lysed by a single phage from one group, more than one phage from a single group, or by phages from more than one group. The internationally recognized gold standard for discriminating between strains of *S. aureus* is pulse field gel electrophoresis (**PFGE**), a DNA finger printing technique. It is neither necessary nor feasible to use PFGE for typing all strains of *S. aureus* as most can be satisfactorily differentiated by phage typing. Epidemic MRSAs are designated specific numerical types (1 to 17) as well as phage types. Phage typing is much less satisfactory for coagulase-negative staphylococci and has largely been abandoned. Epidemiological studies on these bacteria are seldom required in clinical practice but could be done by PFGE.

Staphylococcus aureus

Pathogenicity

S. aureus produces a remarkable variety of extracellular substances that include: general toxic agents such as catalase, hyaluronidase, lipase, and membrane-damaging toxins that may be involved in the pathogenesis of local or systemic inflammation; and specific toxins such as enterotoxins and epidermolytic toxins that mediate particular non-suppurative diseases.

Membrane-damaging toxins

S. aureus produces five toxins that disrupt cell membranes—a-, b-, g-, and †- toxins and leucocidin. Many of these toxins disrupt red cell membranes producing haemolysis. The most extensively studied is a-toxin, which is formed by most strains and produces impressive biological effects; it is cytotoxic and necrotizing, kills leucocytes, lyses platelets, releases catecholamines, and causes renal cortical necrosis yet remarkably its specific role in staphylococcal infection in humans has yet to be defined.

Enterotoxins

There are now 11 staphylococcal enterotoxins—types A, B, D, E, G, H, I, and J which show major antigenic differences and type C which is subdivided into C1, C2, and C3 on the basis of minor antigenic differences. Enterotoxins G, H, I, and J were only described recently and have not yet been confirmed as emetic in humans. Enterotoxin F is identical to toxic shock syndrome toxin 1 and is now known as TSST-1. About 40 per cent of *S. aureus* produce enterotoxin, sometimes of more than one type. Staphylococcal food poisoning results from the ingestion of foods containing preformed enterotoxin. Most outbreaks in the United Kingdom are caused by enterotoxin A with or without D (see Table 3). Staphylococcal enterotoxins have a range of biological activities in addition to their ability to induce vomiting; they are pyrogenic, mitogenic, and can produce thrombocytopenia and hypotension.

P>**Epidermolytic toxins**

These toxins cause intraepidermal splitting and are responsible for the scalded skin syndrome and the blistering of impetigo. The production of epidermolytic toxin is particularly associated with (though not confined to) *S. aureus* of phage group II. There are two epidermolytic toxins, ETA which is heat stable and under

chromosomal control, and ETB which is heat labile and plasmid-mediated, and most phage group II staphylococci produce ETA or both ETA and ETB.

Toxic shock syndrome toxin (TSST-1)

This toxin is responsible for the toxic shock syndrome (**TSS**). Most cases of menstrually associated TSS are mediated by TSST-1 produced by *S. aureus* of phage group I, usually phage 29 or 52; TSS not associated with menstruation can occur with strains producing TSST-1, but also with phage group V strains that produce enterotoxin B.

Carriage

S. aureus is part of the normal flora in some individuals; about 25 per cent of people 'carry' the organism permanently, a similar proportion never do, and the rest do so intermittently. Common carriage sites are the nose, axilla, perineum, and toe webs. Nasal carriage rates vary from 10 to 40 per cent in normal adults outside hospital, but higher rates are often found in patients in hospital, particularly those who have been in hospital for several weeks. High carriage rates are also found in those with skin diseases such as eczema, those with insulin-dependent diabetes, patients on chronic haemodialysis or chronic ambulatory peritoneal dialysis, intravenous drug users, and HIV-positive patients. Some carriers, designated 'shedders', disperse large numbers of staphylococci into the environment on skin squamas. The carrier state is highly relevant to the epidemiology of *S. aureus* infection as to whether or not this complicates surgery or trauma and the source of the *S. aureus* in most patients who develop staphylococcal infection is endogenous.

Host factors in *S. aureus* infection

Intact skin and mucous membranes are important defences against staphylococcal infection. Wounds, whether traumatic or surgical, frequently become colonized with *S. aureus*, which may result in localized infection or in dissemination via the bloodstream to distant sites. Sometimes trivial, even unrecognized, skin trauma precedes such haematogenous spread. Burns and skin diseases are also important portals of entry for staphylococci. Certain viral infections such as influenza damage the respiratory epithelium and allow secondary staphylococcal invasion. Foreign material including intravascular devices, arteriovenous shunts, and vascular and orthopaedic prostheses is also relevant to the pathogenesis and perpetuation of staphylococcal infection.

Once *S. aureus* gains access to the tissues, polymorphs are the most important line of defence. Phagocytosis involves chemotaxis, opsonization, and intracellular killing. Chemotactic defects occur, for example, in Job syndrome (in which patients with recurrent eczema suffer from repeated skin infections and cold abscesses with *S. aureus*) and also in certain other rare syndromes. Opsonic defects tend to predispose to a variety of pyogenic infections, including, but not specifically, *S. aureus* infection, but *S. aureus* is a major pathogen in chronic granulomatous disease producing local and metastatic abscesses. In this disease, intracellular killing by the polymorphs is defective.

Susceptibility of *S. aureus* to antibiotics and antiseptics

Resistance to antibiotics is not a marker for virulence in *S. aureus* and strains that are sensitive to all antistaphylococcal antibiotics, including penicillin, can cause severe community-acquired infections. However, *S. aureus* has a record of rapid and successful development of resistance to antibiotics. Most isolates, whether acquired in the community or in hospital, produce penicillinase (b-lactamase) and are thus resistant to penicillin itself and related compounds including ampicillin and amoxycillin. Staphylococcal b-lactamase has a negligible effect on methicillin, cloxacillin, and flucloxacillin, which were sequentially introduced specifically for the treatment of staphylococcal infection. Methicillin-resistant strains of *S. aureus* (MRSAs) were detected soon after the introduction of methicillin in 1960, and reports of their isolation increased until 1971 when they accounted for some 5 per cent of strains submitted to the Staphylococcus Reference Laboratory of the Central Public Health Laboratory in the United Kingdom. MRSAs then diminished in frequency in the United Kingdom, possibly as a result of increased prescribing of aminoglycosides, but there was a resurgence in the early 1980s and for some years now virtually all hospitals have patients who are colonized or infected with MRSAs. MRSAs are usually, but not always, resistant to a variety of other antibiotics in addition to methicillin, and are resistant to all cephalosporins. *S. aureus* other than MRSA, whether penicillinase producing or not, are sensitive to many cephalosporins, though the newer third-generation cephalosporins, such as ceftazidime, are much less active than cefuroxime and cefotaxime. The incidence of erythromycin resistance relates to its use and varies from about 5 to 20 per cent. Gentamicin resistance is unusual, except in MRSA. Resistance of *S. aureus* to fusidic acid is uncommon, but most cultures contain a few resistant mutants and a fully resistant population can emerge, particularly after topical use. Since 1996 MRSAs of reduced sensitivity to vancomycin have been reported from Europe, Asia, and the United States. Such strains are referred to by the acronyms 'VISA' (vancomycin-intermediate *S. aureus*) and 'GISA' (glycopeptide-intermediate *S. aureus*); GISA is the more appropriate as these strains are of intermediate resistance to both the glycopeptide drugs vancomycin and teicoplanin, but the term glycopeptide is less familiar to clinicians. The emergence of such resistance is of concern as glycopeptides are extensively used for MRSA infections. However VISA/GISA and other MRSAs are sensitive to quinupristin/dalfopristin (Synercid) and the oxazolidinone Linezolid Rifampicin is highly active against *S. aureus*, but as with fusidic acid, minority populations of resistant cells are found and resistance may emerge during treatment.

The topical antibiotic mupirocin (Bactroban) is active against many *S. aureus* although unfortunately in some hospitals mupirocin resistance is common thus limiting the use of mupirocin to eradicate MRSA from the nose and other superficial sites. Most disinfectants and antiseptics inhibit or kill *S. aureus.* Chlorhexidine, hexachlorophane, triclosan, and iodine-containing compounds such as povidone iodine are all used for skin disinfection and when used correctly are highly effective in removing staphylococci from the skin.

Prevention of spread of *S. aureus*

Although any strain of *S. aureus* can spread between people whether patients or staff, measures to control spread are now largely directed at the nosocomial spread of MRSAs. MRSAs have caused innumerable hospital outbreaks in many countries; sometimes these outbreaks have involved colonization rather than infection, but severe infection is increasingly encountered and the MRSA has become the scourge of elective surgery in some hospitals. Several distinct strains have caused epidemics and the prevalent type varies: in the United Kingdom overall type 15 is most common, but in London the incidence of types 15 and 16 is equal. Epidemic strains not only spread readily but can cause severe invasive infection. Colonization with MRSA is a notoriously recalcitrant problem in hospitals, particularly on elderly care units, and is also an increasing problem in nursing and residential homes. Eradication of MRSA from the nose and from some surface lesions is readily achieved with topical agents such as mupirocin, but it is virtually impossible to eradicate MRSA from the throat, sputum, or from sites associated with a foreign body such as an indwelling catheter or tracheostomy unless these are removed. Repeated careful handwashing by all staff in contact with patients is the single most effective measure in preventing spread of staphylococci.

Clinical manifestations

S. aureus usually causes localized infection, sometimes with local spread, but this may result in bacteraemia and dissemination of the infection. Certain staphylococcal syndromes are produced by extracellular toxins rather than local invasion and will be considered separately.

Localized infections

Infection of the skin and its appendages

These infections often arise in association with hair follicles, and minor trauma, maceration, and skin diseases also predispose to them. Folliculitis is a superficial infection of the hair follicle commonly caused by *S. aureus.* Boils (furuncles) are deep-seated infections around a hair follicle usually on the neck, axilla, buttock, or thigh, often recurrent, and sometimes involving more than one member of a family. When several adjacent hair follicles are involved a carbuncle develops, usually on the back of the neck, with multiple draining sinuses and systemic disturbance. Although boils are very common, carbuncles are now rarely seen. Impetigo is a blistering skin lesion with crusting exudate affecting exposed areas (often the face) usually in children. Epidermolytic toxin is associated with these infections. Most acute paronychias are caused by *S. aureus*. Mastitis and breast abscess in the puerperium are caused by *S. aureus* as are many non-puerperal breast abscesses. Newborn babies commonly suffer from staphylococcal infection, with septic spots, 'sticky umbilicus', 'sticky eye', and occasionally breast abscess, as well as the much rarer toxin-mediated staphylococcal diseases. Styes, purulent infections of the glands of the eyelid, are caused by *S. aureus*.

Wound infection

S. aureus is the commonest cause of wound infection after surgery or trauma that does not involve the mucous membranes with their rich anaerobic commensal flora.

Staphylococcal wound infection varies from minimal erythema and serous discharge, through small abscesses often in relation to sutures, to marked cellulitis, deep pus, and wound dehiscence with pain and systemic disturbance. It is of particular concern after operations involving prosthetic material such as joint or vascular prostheses or heart valves as the infection can extend from the wound to infect the prosthesis with disastrous results.

Ear, nose, and throat infection

Staphylococcal infection of the hair follicles or sebaceous glands in the outer external auditory canal causes acute localized otitis externa with severe pain and itching. Acute otitis media and sinusitis are seldom caused by *S. aureus*. Although *S. aureus* is commonly grown from throat swabs, it behaves as a commensal at this site, and such patients have usually been taking antibiotics.

Pleuropulmonary infection

Staphylococcal pneumonia arises either from aspiration or by haematogenous spread with metastatic seeding of the lung. Aspiration pneumonia generally complicates pre-existing lung disease or viral respiratory disease, usually influenza. In children, other viral infections of the respiratory tract, including severe measles in developing countries, may be followed by secondary bacterial infection with staphylococci. *S. aureus* from carriage sites presumably reaches the damaged lung tissue via the trachea and bronchi. In contrast to aspiration pneumonia, haematogenous staphylococcal pneumonia characteristically affects a previously normal lung. There may sometimes be an identifiable local infection, often of the skin and usually trivial, that has resulted in haematogenous seeding or there may be evidence of release of infected thrombi via the venous system as in tricuspid endocarditis or occasionally when there is an infected intravascular device. Staphylococci can usually be isolated from the blood in haematogenous pneumonia, though seldom in aspiration pneumonia. Whatever its pathogenesis, *S. aureus* pneumonia is a severe disease. When secondary to influenza, it may occur without an obvious influenza-like prodromal illness and with alarming suddenness (Fig. 1(a) and Fig. 1(b)). It is usually complicated by abscess formation, empyema, and in children, by pneumatoceles and pyopneumothorax, but the radiological findings at initial presentation vary from local consolidation, to multiple patchy infiltrates, and abscess formation may or may not be detected.

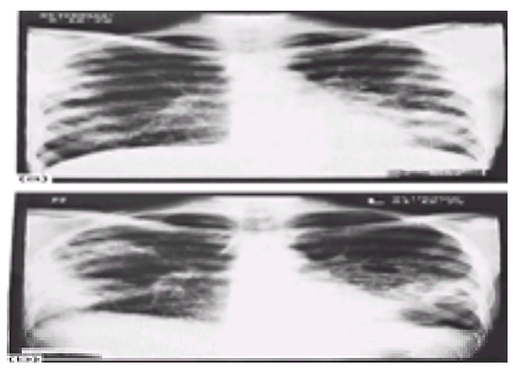

Fig. 1 Chest radiograph of a 24-year-old man with severe staphylococcal pneumonia. (a) on admission and (b) 13 days later. The patient was also suffering from influenza B.

Urinary tract infection

S. aureus urinary tract infection is uncommon and unlikely to occur in patients with a normal urinary tract except in staphylococcal bacteraemia with microabscesses in the kidney as may occur in staphylococcal endocarditis. *S. aureus* urinary infection sometimes occurs in patients with abnormal bladder function but in association with instrumentation or catheterization, presumably from previous urethral colonization with the strain.

Bacteraemia, septicaemia, and metastatic (haematogenous) infection

Bacteraemia in the strict sense means bacteria in the blood, that is, a positive blood culture. This may or may not be symptomatic in the patient. A symptomatic bacteraemia is referred to as a septicaemia. In fact, *S. aureus* in the blood is almost always symptomatic and thus strictly a septicaemia, but the terms tend to be used interchangeably adding to the confusion. Bacteraemia will be used here.

Community-acquired bacteraemia

In most patients who acquire *S. aureus* bacteraemia in the community the staphylococcus has entered the bloodstream from a carrier site or from a trivial unnoticed abrasion and seldom from a defined local lesion. Such bacteraemia then results in serious deep-seated infection. Such bacteraemias have been called 'primary' and are usually much more severe than those secondary to a defined focus of infection. Primary bacteraemia can occur at any age, and often in a previously healthy individual. Such patients may, but often do not, present with initial infection at a specific site. An ill patient with a community-acquired *S. aureus* bacteraemia and no detectable focus of infection will generally have endocarditis (see below).

Hospital-acquired bacteraemia

Nosocomial *S. aureus* bacteraemia usually results from an infected intravascular access site and this is as likely to be a peripheral cannula as a central catheter. There will often be obvious infection at the insertion site, but sometimes only minimal local signs with severe systemic disturbance. Such bacteraemias can result in metastatic infection involving bone, joint, lung, or heart valve. These disastrous iatrogenic sequelae of intravenous access are increasing. Nosocomial bacteraemia also sometimes occurs with severe wound sepsis.

Endocarditis

S. aureus endocarditis is a devastating illness often occurring in a previously healthy individual, but an asymptomatic left-sided valvular abnormality such as a bicuspid aortic valve or mitral leaflet prolapse is sometimes present. The infection typically presents as an influenza-like illness, often with gastrointestinal disturbance. Meningism is seen in about 25 per cent of cases and polymorphs, though seldom organisms, are detected in the cerebrospinal fluid. There may be systemic emboli. Valvular insufficiency can develop within days, sometimes hours, of admission. Staphylococcal endocarditis is a rapidly destructive disease, justifiably called malignant endocarditis by Osler.

The skin manifestations can be mistaken for meningococcal infection (Fig. 2(a) and Fig. 2(b)). Figure 3(a) and Figure 3(b) show another fatal staphylococcal infection in a previously healthy 54-year-old man who was febrile and confused at presentation and hemiplegic within 48 h. Blood and CSF (which contained 5000 polymorphs) grew *S. aureus*.

Fig. 2 Meningococcaemic-like infection in a 22-year-old man who died from an aortic root abscess from *S. aureus* endocarditis on a bicuspid aortic value. A false-positive meningococccaemic latex agglutination test on the cerebrospinal fluid (which contained 1500 polymorphs, but no organisms) taken on admission further increased the clinical confusion.

Fig. 3 Meningococcaemic-like disease. (a) Hand and (b) foot of a man with primary staphylococcal bacteraemia and meningitis who had disseminated intravascular coagulation.

Staphylococcal endocarditis is occasionally complicated by splenic abscess (Fig. 4). Intravenous drug users are at particular risk of staphylococcal endocarditis, but unless the affected individual has a previous valvular abnormality, the infection is likely to involve the tricuspid valve and present with fever, malaise, and respiratory signs that result from septic pulmonary emboli.

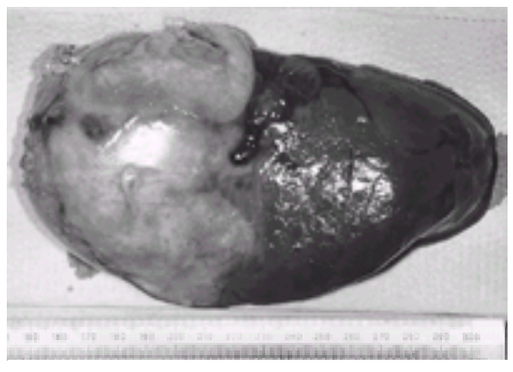

Fig. 4 Splenic abscess complicating staphylococcal endocarditis.

Bone and joint infections

S. aureus is the commonest cause of acute bone and joint infection. These infections can result from a 'primary' bacteraemia, but also from a contiguous focus of infection after trauma or surgery, especially that involving prosthetic implants. The overall incidence of acute haematogenous osteomyelitis has decreased but there has also been a change in its localization. Osteomyelitis of the long bones seen primarily in children, particularly boys, is now uncommon, but vertebral osteomyelitis has increased or is increasingly recognized. Patients with staphylococcal vertebral osteomyelitis are usually middle aged or elderly. This shift in the localization of the infection has not been explained. Vertebral osteomyelitis can be a notoriously difficult diagnosis, repeatedly referred to in the literature as a 'diagnostic pitfall'. Pain, not always localized to the spine is the only consistent feature. Fever is sometimes absent but at least if present should initiate a blood culture, which is likely to be positive whatever the temperature. Any patient with backache, a high C-reactive protein and erythrocyte sedimentation rate, and *S. aureus* in the blood should be assumed to have vertebral infection.

Staphylococcal septic arthritis may occur in previously normal or abnormal joints and at any age. It may involve one or more joints and multiple infection can occur in patients with previous joint pathology such as rheumatoid arthritis, when it may be difficult to diagnose.

Renal cortical abscess (carbuncle) and perinephric abscess

These metastatic staphylococcal infections are rare and usually cause diagnostic confusion. A renal cortical abscess, also known as a carbuncle, is a multilocular abscess involving the renal parenchyma, the result of the coalescence of cortical microabscesses from haematogenous seeding of the kidney from a previous infection, typically a boil, with *S. aureus.* The patient complains of fevers and loin pain but urinary symptoms are usually absent and unless the abscess communicates with the excretory system, the urine contains neither pus cells nor *S. aureus*. Although *S. aureus* is the commonest pathogen in renal carbuncle, perinephric abscesses—those external to the renal capsule but within the perinephric fascia—are more commonly caused by Gram-negative aerobes such as *Escherichia coli* or *Proteus* spp. than staphylococci. A renal carbuncle may rupture into the perinephric space producing a perinephric abscess. Again, urine cultures are unlikely to be positive and the signs are similar to those of a renal carbuncle.

Pyomyositis

See Chapter 24.22.6 for discussion.

Infections mediated by toxins of *S. aureus*

Staphylococcal food poisoning

This syndrome, characterized by vomiting, nausea, abdominal cramps, and diarrhoea, is caused by the ingestion of staphylococcal enterotoxin preformed in the food. The onset occurs within hours of ingestion of food contaminated during its preparation by an individual infected with, or shedding, an enterotoxin-producing staphylococcus. Unrefrigerated protein-rich foods containing meat or milk are likely to support the growth of staphylococci and the subsequent production of heat-stable enterotoxin. Only about 5 per cent of outbreaks of bacterial food poisoning reported to the Communicable Disease Surveillance Centre for which an aetiological agent is identified are caused by *S. aureus*. The diagnosis can be confirmed by culturing incriminated food, any skin lesions, the nose of food handlers, and the stools of the victims. In most outbreaks, both the organism and its toxin can be defined, but occasionally, enterotoxin alone is demonstrated in the food.

Staphylococcal scalded skin syndrome

This rare disease, originally known as Ritter's disease when it was first seen in infants in the late nineteenth century, is more commonly seen in children (Fig. 5) than adults (Fig. 6). It is characterized by the sudden onset of extensive erythema followed by bullous desquamation of large areas of skin. It is caused by the epidermolytic toxins of *S. aureus*. The disease of scalded skin syndrome must be distinguished from a similar clinical entity unassociated with *S. aureus*, that of toxic

epidermal necrolysis (Lyell's syndrome) which occurs in older children and adults and results from drug hypersensitivity. Histologically the two diseases can be readily distinguished as in scalded skin syndrome there is intraepithelial splitting at the level of the stratum granulosum and in toxic epidermal necrolysis there is total epidermal loss with separation at the dermal–epidermal junction.

Fig. 5 Staphylococcal scalded skin syndrome in a child. (Reproduced by courtesy of Professor W.C. Noble.)

Fig. 6 Staphylococcal scalded skin syndrome in an adult.

Staphylococcal toxic shock syndrome (TSS)

This syndrome (Fig. 7) of high fever, mental confusion, erythroderma, diarrhoea, hypotension, and renal failure was first defined in children in 1978, but had been described 50 years earlier and thought to be staphylococcal scarlet fever. In the late 1970s there was an epidemic of toxic shock syndrome in women associated with menstruation and tampon use, initially, and predominantly, in the United States, but later, though in far fewer numbers, in other countries. Toxic shock syndrome has also been described in women who were not menstruating and in men in association with a wide variety of conditions and operations including burns in children. Toxic shock syndrome may be fatal and a mortality rate of around 5 per cent was reported during the 'tampon epidemic'. Since the syndrome is mediated by toxin, the mainstay of treatment is supportive. Antistaphylococcal antibiotics should be given to eradicate *S. aureus* from the local site. Bacteraemia has rarely been reported in toxic shock syndrome. The staphylococci isolated are usually resistant only to penicillin.

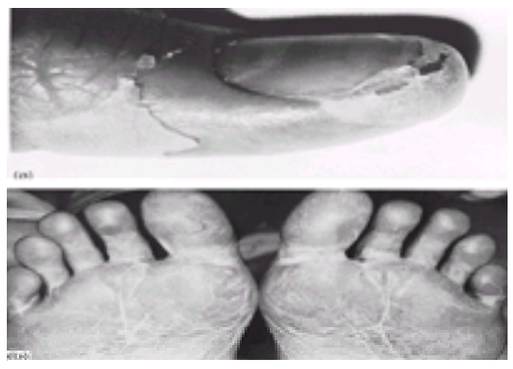

Fig. 7 Toxic shock syndrome. Desquamation of (a) hand and (b) feet in a girl with tampon-associated disease. (Reproduced by courtesy of Dr D.C. Shanson.)

Laboratory diagnosis of *S. aureus* infection

S. aureus is readily isolated in the laboratory. A Gram-stained film may enable a rapid diagnosis of a staphylococcal aetiology; the characteristic clumps of Gram-positive cocci, often intracellular as well as extracellular and sometimes of variable size, are readily identifiable. The diagnosis can be confirmed by culture within 18 to 24 h. Staphylococcal bacteraemia is readily detected by routine blood culture methods. The isolation of *S. aureus* from blood is almost always indicative of a genuine bacteraemia and the organism should only be dismissed as a contaminant if the patient has extensive skin disease such as eczema, rarely otherwise.

Treatment

Drainage of any pus is an essential prerequisite of the management of *S. aureus* infection. This may occur spontaneously or with only minor surgical intervention in most superficial infections such as boils, paronychias, styes, and stitch abscesses. Deep abscesses in wounds or organs and osteomyelitis that has progressed to the point of pus formation require definitive surgical drainage. Infections associated with intravascular devices or other prosthetic material seldom resolve with antibiotics and removal of the foreign material is usually required.

Antibiotics are indicated if the patient is systemically unwell or the infection is spreading and sometimes when given early in the course of a potentially localizing pyogenic infection may arrest its progress. They are of no benefit in staphylococcal food poisoning but should be given in scalded skin syndrome and toxic shock syndrome to eradicate toxin-producing *S. aureus*. The initial choice of agent for staphylococcal infection before sensitivities become available depends on whether the staphylococcus was acquired in the community or in hospital. For community-acquired infections a b-lactamase-resistant penicillin such as flucloxacillin or cephalosporin such as cefuroxime will be appropriate initial treatment and probably also definitive therapy. Penicillin is suitable only if the strain does not produce b-lactamase, and should never be used for the initial 'blind' treatment. Similar constraints apply to ampicillin and amoxicillin. Alternative agents to b-lactams for community-acquired infection when the patient is hypersensitive to penicillin include the macrolides erythromycin, clarithromycin, and azithromycin. Fusidic acid is an excellent antistaphylococcal agent although resistance may arise during treatment, especially when the organism cannot readily be eradicated. When the infection is acquired in hospital and the patient is unwell, in most hospitals it should be assumed to be an MRSA until cultures prove otherwise. The only agents with reliable activity against MRSA are vancomycin, teicoplanin, the combination drug quinupristin/dalfopristin (Synercid), and the oxazolidinone Linezolid. Most staphylococcal infection is satisfactorily treated with a single antibiotic. Combination therapy is often used for serious infections such as endocarditis and bone or joint infection but there is minimal evidence of any advantage over a single agent.

The length of the antibiotic course to treat staphylococcal infection is unknown, but for serious infections such as endocarditis, bone and joint infections, and pneumonia several weeks' treatment may be needed. For most other infections, antibiotics should be given until there is clinical improvement or for about 48 h after fever has resolved. Most patients are treated for too long inviting side-effects. Blood cultures that are persistently positive for *S. aureus* despite appropriate antibiotic therapy are seldom an indication for changing the antibiotics, but rather for an assessment of the need for intervention, for example to remove an infected intravascular device, excise an infected heart valve, or aspirate and wash out an infected joint. Topical antibiotics and antiseptics are useful for the treatment of staphylococcal skin infections.

Coagulase-negative staphylococci

Coagulase-negative staphylococci are the commonest contaminants in the laboratory, particularly, though not exclusively, in blood cultures, but they can also be important pathogens whose incidence continues to increase. The availability of commercial kits for their speciation has served to emphasize that they cannot be regarded as a homogeneous entity; the different species vary not only in their incidence in clinical infections but also in the type and severity of disease produced. Most infections with coagulase-negative staphylococci are hospital acquired, but certain species cause severe community-acquired infection.

Pathogenicity

Coagulase-negative staphylococci (usually *S. epidermidis*) that cause infections associated with prosthetic devices and intravascular catheters produce an exopolysaccharide ('slime') which is important in enabling these organisms to adhere to plastic material and probably also in their resistance to phagocytosis, other host defences, and to antimicrobial action. Coagulase-negative staphylococci from clinical infections produce a variety of potential toxins including haemolysins, cytotoxins, deoxyribonuclease, fibrinolysin, proteinase, and lipase-esterase, similar to those produced by *S. aureus*, and infection caused by some species, especially *S. lugdunensis* and *S. simulans*, mimics that caused by *S. aureus*.

Carriage

Coagulase-negative staphylococci, together with coryneforms, comprise most of the human skin flora. Many different species are found on the skin but the commonest is *S. epidermidis*; *S. hominis* and *S. haemolyticus* are also common. Distribution of species varies on different skin areas: the predominant species on the head and trunk is *S. epidermidis*, on the arms and legs it is *S. hominis*, and as its name suggests *S. capitis* is found mainly on the head. There are also geographical variations.

Host factors in coagulase-negative staphylococcal infection

Most infection with coagulase-negative staphylococci is associated with prosthetic material both in immunocompromised and non-immunocompromised patients. Infection of intravascular catheters arises via the catheter access site or the catheter hub from frequent disconnections. Prosthetic material can also become infected at implantation.

Antibiotic susceptibility

Hospital-acquired coagulase-negative staphylococci, particularly *S. epidermidis* and *S. haemolyticus*, are usually multiply resistant. Most are resistant to methicillin (and thus to cephalosporins), and many to gentamicin and erythromycin. Thus the usual nosocomial strain of coagulase-negative staphylococcus has an antibiotic susceptibility pattern akin to many MRSAs. Rare resistance to vancomycin and teicoplanin has been reported, initially in *S. haemolyticus*. In marked contrast to hospital-acquired infections those acquired in the community are usually caused by very sensitive strains; many are sensitive to penicillin.

Infections caused by coagulase-negative staphylococci

Most infections caused by coagulase-negative staphylococci are acquired in hospital and are increasingly common. They usually arise in association with an intravascular or prosthetic device or implant. Infection with more than one strain may occur in nosocomial infections.

Community-acquired infections though rare are probably increasing or increasingly recognized and are usually severe. In most community-acquired infections, repeated isolation of the same coagulase-negative staphylcoccus from the blood is essential for the diagnosis, and true bacteraemia must be distinguished from contamination.

Intravascular devices

There has been a marked increase in infection of intravascular devices with coagulase-negative staphylococci, particularly in neonates and immunocompromised patients, and they are the commonest bacteria in such infections. The degree of systemic disturbance varies, and this should determine the approach to treatment. In contrast to infections of intravascular devices caused by *S. aureus* for which it is usually necessary to remove the device to control the infection, with those caused by coagulase-negative staphylococci the catheter can often be left *in situ* and the infections controlled with antibiotics. If this fails, the device must be removed. Very occasionally, as with *S. aureus*, persistent bacteraemia can result in metastatic seeding of a heart valve or vertebral body.

Cerebrospinal fluid shunts

Coagulase-negative staphylococci, predominantly *S. epidermidis*, are the commonest cause of infection of cerebrospinal fluid shunts and these infections can present weeks, months, or years after the shunt insertion. They also cause infection of cerebrospinal fluid reservoirs used for chemotherapy. Signs of meningitis may be absent and usual findings include low-grade fever, malaise, and shunt malfunction. Serum antibodies to *S. epidermidis* can be used to monitor treatment and detect relapse. Treatment may require removal of the shunt and antibiotics, usually vancomycin with rifampicin, are best given intraventricularly. Occasionally, glomerulonephritis ('shunt nephritis') occurs in patients with colonized shunts as a result of deposition of immune complexes on the basement membranes of the glomeruli.

Peritonitis associated with continuous ambulatory peritoneal dialysis

Coagulase-negative staphylococci, predominantly *S. epidermidis*, are the commonest cause of peritonitis associated with continuous ambulatory peritoneal dialysis. The bacteria probably gain access to the peritoneum as a result of manipulation of the catheter connections. Patients have abdominal pain, occasionally nausea, diarrhoea, and fever, and abundant polymorphs in the dialysate in which Gram-positive cocci, usually scanty and intracellular, may be detected on a Gram-stained smear. The antibiotic sensitivities of infecting strains vary and since treatment must always be started before this information is available, vancomycin (preferably intraperitoneally) is the drug of choice.

Endocarditis

Coagulase-negative staphylococci can infect native or prosthetic heart valves. Nosocomial native valve infections with coagulase-negative staphylococci (usually *S. epidermidis*) generally result from infected intravascular devices; the affected valve may or may not have been previously abnormal. Nosocomial prosthetic valve endocarditis can be acquired in the theatre at the time of valve replacement surgery or shortly thereafter and presents within weeks or more often months of surgery ('early-onset'). In many series, coagulase-negative staphylococci are the commonest cause of early-onset prosthetic valve endocarditis. Prosthetic infection can also be acquired from an infected intravascular device. Community-acquired coagulase-negative staphylococcal endocarditis, which usually involves native valves, is increasingly recognized and most patients will have a pre-existing cardiac abnormality. The organisms must derive from the patient's skin, but predisposing skin lesions are seldom detected. The infection often mimics *S. aureus* endocarditis with rapidly destructive valvular disease, neurological manifestations, and concomitant vertebral osteomyelitis. The commonest pathogen is *S. epidermidis*, but there are increasing reports of other species, particularly *S. lugdunensis*, which seems to be especially virulent. These community-acquired strains are frequently penicillin sensitive.

Urinary tract infection

Coagulase-negative staphylococci are urinary pathogens both in the community and in hospital but different species are involved. In the community the curiously named *S. saprophyticus* is an important urinary pathogen in sexually active women, second only to *Escherichia coli*. It commonly produces cystitis, but may cause upper urinary tract infection and has been isolated from infected calculi. *S. saprophyticus* is a skin commensal but is not normally found colonizing the urethra, though it has been isolated from the rectal flora of women. Most strains are readily recognized in the laboratory by their resistance to novobiocin. They are sensitive to a wide range of antibiotics. Some nosocomial urinary tract infections are also caused by coagulase-negative staphylococci, predominantly *S. epidermidis*. These infections, usually after urological surgery, are seldom accompanied by pyuria, and may clear spontaneously on removal of the catheter. Nosocomial urinary isolates of coagulase-negative staphylococci are often multiply resistant.

Other infections

Coagulase-negative staphylococci are increasingly isolated from the blood of neonates and immunocompromised, neutropenic patients. Distinguishing true bacteraemia from contamination can be difficult. In many cases bacteraemia is related to the presence of an intravascular catheter (see above). In premature neonates, colonization of the respiratory tract occurs and respiratory infection can result. Infection of prosthetic joints and vascular prostheses is sometimes caused by coagulase-negative staphylococci. The organisms are introduced at the time of the surgery, although the clinical signs of infection may not become evident for weeks or months. Attempts to treat such infections with antibiotics generally fail and removal of the prosthesis is required. Coagulase-negative staphylococci are also the commonest cause of postoperative endophthalmitis after intraocular surgery.

Laboratory diagnosis

The laboratory diagnosis of much infection with coagulase-negative staphylococci poses greater difficulties than the diagnosis of *S. aureus* infection. The clinician is advised to enlist the help of a competent microbiologist when assessing the validity of culture results of invasive specimens growing coagulase-negative staphylococci. A further problem with these organisms occurs with the use of broth enrichment cultures for specimens of excised tissue; a single contaminating staphylococcus will multiply in liquid medium, thereby misleading unwary clinicians.

Treatment

As well as the prescribing of antibiotics, an integral part of the successful treatment of infections with coagulase-negative staphylococci is a critical clinical assessment of the need for removal of any prosthetic material with which so many infections are associated. So many nosocomial infections are caused by resistant strains that the only reliable initial therapy is vancomycin or teicoplanin. The length of treatment in most instances is somewhat arbitrary and the same principles apply to infections with these organisms as to those with *S. aureus.* Community-acquired infections, usually endocarditis, can often be treated with b-lactam antibiotics, sometimes with penicillin. As with serious *S. aureus* infections, combination therapy is often used.

Further reading

Archer GL (2000). *Staphylococcus epidermidis* and other coagulase-negative staphylococci. In: Mandell GL, Bennett JE, Dolin R, eds. *Principles and practice of infectious diseases*, pp 2092–100. Churchill Livingstone, New York. [Comprehensive chapter with large number of references.]

Chesney PJ *et al.* (1984). The disease spectrum, epidemiology, and etiology of toxic-shock syndrome. *Annual Review of Microbiology* **38**, 315–38.

Espersen F *et al.* (1991). Changing pattern of bone and joint infections due to *Staphylococcus aureus*: study of cases of bacteraemia in Denmark, 1959 to 1988. *Reviews of Infectious Diseases* **13**, 347–58.

Etienne J, Eykyn SJ (1990). Increase in native valve endocarditis caused by coagulase negative staphylococci: an Anglo-French clinical and microbiological study. *British Heart Journal* **64**, 381–4.

Vandenesch F *et al.* (1993). Endocarditis due to *Staphylococcus lugdunensis*: report of 11 cases and review. *Clinical Infectious Diseases* **17**, 871–6.

Waldvogel FA (2000). *Staphylococcus aureus* (including staphylcoccal toxic shock). In: Mandell GL, Bennett JE, Dolin R, eds. *Principles and practice of infectious diseases*, pp 2069–92. Churchill Livingstone, New York. [Very comprehensive chapter with large number of references.]

7.11.5 Meningococcal infections

P. Brandtzaeg

Neisseria meningitidis infection remains a major public health problem worldwide by causing clusters or epidemics of meningitis and acute lethal sepsis. Case fatality has gradually declined from 80 to 90 per cent to approximately 10 per cent but has remained at this level since the introduction of antimicrobial chemotherapy in 1937.

Bacterium

Neisseria meningitidis is an obligate human Gram-negative diplococcus normally located in the mucous membrane of the upper respiratory tract. Invasive isolates from blood or cerebrospinal fluid are encapsulated and express pili. Capsular polysaccharides that inhibit phagocytosis and bacterial adhesion are divided into 12 different serogroups (A, B, C, 29-E, H, I, K, L, W-135, X, Y, and Z). Serogroups A, B, and C account for more than 90 per cent of all invasive isolates. Less than 10 per cent of clinical isolates are from serogroups W-135 and Y.

The bacterial cell wall consists of an outer lipid bilayer containing lipopolysaccharides (endotoxin) and outer membrane proteins, a thin peptidoglycan layer, and the cytoplasmic membrane. Lipid A is a glycolipid that anchors the lipopolysaccharide to the lipid membrane. It is the major inflammatory (toxic) component of *N. meningitidis*. It can activate a variety of cells via CD14–toll-like receptor-4 interaction or indirectly through activation of blood coagulation, fibrinolysis, kallikrein–kinin, and complement systems. During growth, meningococci release outer membrane vesicles containing large amounts of lipopolysaccharides.

Outer membrane proteins are classified according to electrophoretic mobility into five major classes. Por A (class 1 protein) and Por B (class 2 or 3 proteins) are cation- and anion-selective porins, respectively. Surface exposed loops of Por B and Por A define serotype and serosubtype, respectively. Loops 1 and 4 in Por A are

major epitopes inducing bactericidal and opsonophagocytic antibodies when exposed to the human immune system.

Meningococci are fastidious bacteria that readily autolyse. They grow well on blood agar, supplemented chocolate agar, trypticar soy base, Mueller–Hinton agar, and selective GC-medium. Optimal growth occurs at 35 to 37°C in a humid atmosphere with 5 to 10 per cent carbon dioxide. The convex colonies (diameter: 1 to 4 mm) are transparent, non-pigmented, and non-haemolytic. They produce cytochrome oxidase and ferment glucose and maltose, but not lactose and sucrose, to acid without gas formation.

Practical handling of clinical specimens

Blood culture (10 ml for adults, 2 to 4 ml for infants/children) and swabs from the nasopharynx and the tonsils are collected immediately. Media for blood culture and transportation of swabs should be optimal for recovery of meningococci. Cerebrospinal fluid is best cultured by direct plating of 0.1 ml on supplemented chocolate agar or a similar medium, incubated at 35 to 37°C in 5 to10 per cent carbon dioxide. If direct plating is impossible or delayed, the sample should be stored at +4°C to +20°C, preferably at refrigerator temperature. Recovery of live meningococci may increase if some drops of the cerebrospinal fluid are stored on a sterile swab in transport medium or injected into blood culture medium and incubated at 35 to 37°C.

Direct visualization of *N. meningitidis* in clinical specimens

Intra- and extracellular diplococci can be observed in the cerebrospinal fluid, peripheral blood buffy coat (fulminant septicaemia), and biopsies of haemorrhagic skin lesions using Gram or acridine orange stains.

Polymerase chain reaction

Using primers that recognize various DNA sequences coding for different genes in *N. meningitidis*, it is possible to detect and classify meningococci in cerebrospinal fluid and blood without positive cultures.

Epidemiology

Industrialized countries

Infection presents as single cases or in small clusters. The incidence is usually 1 to 3 per 100 000 inhabitants per year. Strains belonging to specific clonal complexes may cause a hyperendemic situation characterized by a much higher incidence than usually observed (4 to 30 per 100 000 per year). This epidemiological situation may last for more than a decade in defined geographical areas before slowly declining. Serogroup A has disappeared as a cause of significant epidemics. Outbreaks in Finland in the 1970s and in New Zealand in the 1980s were exceptions.

Developing countries

Large-scale epidemics are confined to developing countries, primarily in sub-Saharan Africa. The incidence approaches 10 to 25 per 100 000 inhabitants per year. During epidemic peaks in Africa, as many as 500 to 1000 per 100 000 inhabitants may contract meningococcal infections. Serogroup A and to lesser extent serogroup C dominate the isolates of large epidemics.

The meningitis belt in sub-Saharan Africa

The area stretches from the Gambia in the west to Ethiopia in the east including Senegal, Guinea, Mali, Burkina Faso, Ghana, Togo, Benin, Nigeria, Niger, Chad, Cameroon, The Central African Republic, and Sudan. Mainly serogroup A strains belonging to a few clonal complexes cause the increased attack rate. In some of these countries large-scale epidemics occur every 8 to 12 years.

Season

In temperate climates most cases occur during the winter and early spring. In the sub-Saharan African meningitis belt the incidence increases from the middle and reaches its maximum at the end of the dry season (harmattan). New cases decline rapidly after the start of the rainy season.

Preceding infections

Influenza A predisposes to invasive meningococcal infections. Mycoplasma infections and rubella have been associated with outbreaks.

Age distribution

Cases are seen in all age groups. However, most occur from 0 to 4 years with a smaller peak from 13 to 20 years. During epidemics the median age appears to increase. Complement-deficient patients may contract the infection when they are older than others.

Genetic diversity

N. meningitidis can exchange and incorporate DNA from other *Neisseria* or closely related species. It reveals more genetic diversity than many other human pathogens. However, strains from certain clonal complexes may persist for many decades over wide areas, retaining their pathogenicity. Strains from seven clonal complexes have predominated since the late 1960s.

Predisposing factors for invasive disease Table 1)

Lack of protective antibodies

Antibodies against serogroups A, C, W-135, and Y capsule polysaccharides are bactericidal and confer protection at concentrations of 1 to 2 µg/ml of serum. Serogroup B polysaccharide induces a weak, transient IgM response that is not protective. Bactericidal and opsonophagocytic antibodies recognizing surface-exposed epitopes of the outer membrane protein, in particular Por A, developing after infection, are important for protection. Antilipopolysaccharide antibodies, recognizing commonly shared epitopes among virulent and non-pathogenic *Neisseria* and closely related species, presumably play a role in protection.

Defects in the complement system

Defects in the complement system can increase susceptibility to meningococcal infections up to 6000 times. The commonest are absent or malfunctioning properdin or late complement components C5 to C9. Properdin defects predispose to rapidly progressing, overwhelming septicaemia. Defects in C5 to C9 are associated with recurrent meningococcal infections. Previous studies suggested that the case fatality rate was reduced in those with late complement defects. However, a recent Dutch study found case fatality rates of 16 per cent and 32 per cent among patients with late complement defects and properdin defects, respectively—a difference that was not statistically significant, compared with 7.7 per cent in the general population. Defects in the classical pathway do not predispose to meningococcal infection. Complement defects are rare. They play a minor role in the development of serogroup A, B, and C systemic meningococcal disease, but are over-represented in patients with the uncommon serogroups W-135, X, Y, and Z.

Defects in the mannan-binding lectin

Mannan-binding lectin is a calcium-dependent, opsonizing, acute-phase protein. Mutations in codons 54, 57, and 52 in the mannan-binding lectin gene result in low

serum levels. Defects in this protein have been associated with 1/3 of all meningococcal cases in England and Ireland.

Polymorphism of Fcg-receptor II and Fcg-receptor III

Polymorphisms of Fcg-receptor II (Fcg-RIIa, CD32) and Fcg-receptor III (Fcg-RIIIb, CD16) on phagocytic cells are associated with reduced binding of antibodies. They are over-represented in patients with defects in the late complement components and in children with fulminant meningococcal sepsis. Fcg-RIIa receptors where arginine has replaced histidine at position 131 are associated with reduced binding of IgG2 subclass (antipolysaccharide) antibodies.

Nasopharyngeal colonization

Upper respiratory tract mucosa is the natural habitat of *N. meningitidis*. It is spread from person to person by droplets and direct mucosal contact. Most colonizing meningococci are non-pathogenic, genetically and phenotypically different from virulent invasive strains. Only a small minority of those colonized with virulent strains will develop invasive disease. Colonization is asymptomatic. It induces local and systemic immune responses within 1 to 2 weeks.

Carriage

Cross-sectional studies in England and Norway in the 1980s and 1990s indicated that approximately 10 per cent of the population harboured meningococci in the upper respiratory tract. However, only 1 per cent of the healthy normal population carried strains from typical virulent clones prevalent at the time. The acquisition rate leading to carriage appears to be independent of season, whereas invasive meningococcal infections peak in the winter and early spring in temperate countries.

The carriage rate in England is low (2 to 3 per cent) in the first 4 years of life, rising in children 10 to 14 years of age (9 to 10 per cent), reaching a maximum among young adults of 15 to 19 years (20 to 25 per cent), and than gradually declining to less than 15 per cent in persons above 25 years. It increases in closed or semi-closed communities and is particularly high in military camps where strains change frequently. In university communities with bar and catering facilities the carriage rate is high. Smoking increases the carriage rate.

Reservoir of virulent meningococci

Healthy adults carrying virulent strains of *N. meningitidis* are the main reservoir. Household members and kissing contacts of a patient harbour virulent strains more often than the average population. In industrialized countries, infants and children are usually infected by a local adult carrier. Spread from patients to medical staff is very uncommon. In Africa, children may more commonly infect each other with serogroup A strains.

Invasive infection

Most patients appear to develop invasive disease 2 to 4 days after acquiring the virulent strain in the upper respiratory tract, but some are carriers for up to 7 weeks before invasive infection develops. *N. meningitidis* adheres to specific structures on the epithelial cells in the nasopharynx and on the tonsils. During a period of adaptation and proliferation, meningococci presumably alter various surface structures (lipopolysaccharides, pili, outer membrane proteins) by phase variation before starting transepithelial migration. They reach submucosal tissue and via capillaries gain access to the circulation.

The initial bacteraemic phase

Bacteraemia is a prerequisite for systemic meningococcal infection. Meningococci may be eliminated from the blood by phagocytosis of opsonized bacteria and lysis induced by bactericidal antibodies and complement. Persistent bacteraemia allows meningeal invasion.

Bacterial proliferation and an inflammatory response may occur predominantly in either the subarachnoid space, causing meningitis, or in the circulation, causing meningococcaemia with or without shock.

The rash

Haemorrhagic skin lesions are the hallmark of systemic meningococcal disease, occurring in 70 to 80 per cent of all cases in industrialized countries. They appear as red or bluish petechiae. These lesions are larger and more irregular in size than the petechiae of thrombocytopenic purpura. Each lesion represents a local nidus of meningococci within the endothelial cells, thrombus formation, and extravasation of erythrocytes. The petechial rash indicates meningococcaemia, not necessarily severe sepsis. However, in fulminant meningococcal septicaemia the haemorrhagic lesions are larger (ecchymoses) with a propensity to locate on extremities (Plate 1). Some patients develop relatively large, non-specific, maculopapular lesions, with or without haemorrhagic lesions, at an early stage (Plate 2, Plate 3). The petechial lesions are difficult to discover on dark skin but may be observed in the conjunctivae.

Clinical presentations

Initial symptoms of systemic meningococcal infection are attributable to meningococcaemia. This may persist as a low-grade bacteraemia or develop into septic shock and multiple organ failure in a few hours. Most commonly the patient develops meningococcaemia without circulatory impairment which gradually evolves to meningitis within 12 to 72 hours. Occasionally, patients develop meningitis and persistent shock simultaneously. Based on easily recognizable clinical symptoms, meningococcal infections can be classified as (i) meningitis without shock, (ii) shock without meningitis, (iii) meningitis and shock, and (iv) meningococcaemia without shock or meningitis. Each clinical presentation is associated with a distinct pathophysiological background and prognosis (Table 2).

Distinct meningitis without persistent shock

Meningism dominates the clinical presentation. The onset is often insidious. The patients, particularly children, may complain of general malaise, nausea, and headache. They vomit and become febrile. The temperature may fluctuate and can be normal at times. Many patients are initially diagnosed as 'gastric flu', gastroenteritis, or upper respiratory tract infection. Gradually, the symptoms of meningitis dominate the clinical picture. The patient complains of headache, vomits, develops nuchal and back rigidity, photophobia, and in more advanced cases altered consciousness. Kernig's and Brudinski's signs become positive. Many patients are lethargic, some are agitated. The blood pressure is normal or slightly elevated by stress. Occasionally it is low but can be restored to normal by infusion of a limited volume of fluid. In untreated cases brain oedema develops, the intracranial pressure rises, and the central circulation is increasingly compromised. Finally, herniation of the cerebellum occurs with arrest of the brain circulation. The case fatality rate is usually less than 5 per cent.

Meningococcal meningitis without persistent shock accounts for more than 50 per cent of all cases of systemic meningococcal infections in industrialized countries and an even higher proportion of cases reaching hospitals in developing countries. The combination of multiple petechiae and symptoms of meningitis supports a diagnosis of meningococcal meningitis.

Pathophysiological background

In untreated patients, *N. meningitidis* can be cultivated from cerebrospinal fluid and blood. The concentration of lipopolysaccharides, reflecting the microbial proliferation, is 100 to 1000 times higher in cerebrospinal fluid than plasma. Levels of bioactive inflammatory mediators such as tumour necrosis factor-a, interleukins 1b, 6, 8, 10, and 12, and soluble receptors of these interleukins are much higher in cerebrospinal fluid than plasma. Plasma proteins, mainly albumin, leak into the cerebrospinal fluid. The influx of mainly neutrophils causes the pleocytosis. The glucose level of the cerebrospinal fluid is reduced mainly as a result of increased central glucose consumption rather than the pleocytosis.

Laboratory findings

The erythrocyte sedimentation rate, C-reactive protein, and leucocyte count in the peripheral blood are markedly elevated with increased numbers of band forms. Sodium, potassium, calcium, and magnesium ions, pH, renal, hepatic, and coagulation parameters are usually within normal range. Cerebrospinal fluid shows a

marked pleocytosis (more than 100 × 10^6 leucocytes/l), with increased levels of protein and decreased level of glucose. Intra- and extracellular Gram-negative diplococci can be detected by direct microscopy.

Persistent septic shock without distinct meningitis

Fulminant meningococcal septicaemia (Waterhouse–Friderichsen syndrome) is characterized by persistent circulatory failure and severe coagulopathy leading to thrombosis and extensive haemorrhage of the skin, thrombosis and gangrene of the extremities, and impaired renal, adrenal, and pulmonary function.

Symptoms develop very rapidly. Six to 12 hours after recognizing their first symptoms the patients are often desperately ill. Initially, they complain of 'flu-like' symptoms, such as fever, aching muscle, prostration, abdominal pain, and nausea. The temperature rises rapidly, commonly to between 39.0 and 41.5°C, but occasionally lower. Diarrhoea is quite common during the first few hours. The patient appears worryingly sick to relatives. Before the appearance of petechiae and ecchymoses the symptoms are often misinterpreted as influenza or acute gastroenteritis.

The haemorrhagic skin lesions are first seen as bluish petechiae, which rapidly increase in size and number. They are distributed all over the body but are often more pronounced and detected earliest on the extremities. Occasionally they are seen on the conjunctivae and other mucous membranes.

The circulation is severely impaired. The extremities are often cold and cyanotic. The blood pressure is low despite tachycardia. The tissue perfusion remains inadequate despite extensive fluid and pressor therapy. Initially, the circulation is hyperdynamic, but gradually becomes hypodynamic from persistent vasodilatation and gradually reduced myocardial performance. The heart becomes dilated with a reduced ejection fraction.

Patients usually lack nuchal and back rigidity. Kernig's sign is negative. Despite impaired circulation, many patients remain awake and alert on hospital admission, being able to communicate their complaints. They hyperventilate to compensate for the pronounced metabolic acidosis. Urine output gradually dwindles. They may develop acute respiratory distress syndrome (ARDS), i.e. pulmonary oedema after fluid volume repletion.

Circulatory collapse dominates the clinical picture during the first 48 to 96 h. Death is usually within 48 h. Later, ARDS, renal failure, and the consequences of the diffuse thrombosis of the extremities and the skin dominate the picture. The case fatality rate ranges from 29 to 53 per cent.

Rapidly evolving symptoms with fever, circulatory shock, and extensive skin haemorrhages in a person without a history of splenectomy makes the diagnosis of fulminant meningococcal septicaemia likely. The same clinical picture is, however, observed in cases of overwhelming *Streptococcus pneumoniae, Haemophilus influenzae, Streptococcus pyogenes*, and *Capnocytophaga canimorsus* infections and with viral haemorrhagic fevers (Plate 4).

Pathophysiological background

There is very rapid microbial proliferation in the circulation, generating large amounts of bacterial lipopolysaccharides in a few hours, but with limited or no bacterial growth in the subarachnoid space. The levels of lipopolysaccharides in the plasma predict the development of persistent septic shock, multiple organ failure, and death. Among 100 Norwegian patients with systemic meningococcal disease, admission levels of lipopolysaccharides in the plasma of less than 1000 pg/ml (10 endotoxin units/ml) were associated with no mortality caused by septic shock, rising to 100 per cent mortality in patients with more than 15 000 pg/ml (150 endotoxin units/ml), that is, 1.2 log higher (Fig. 1).

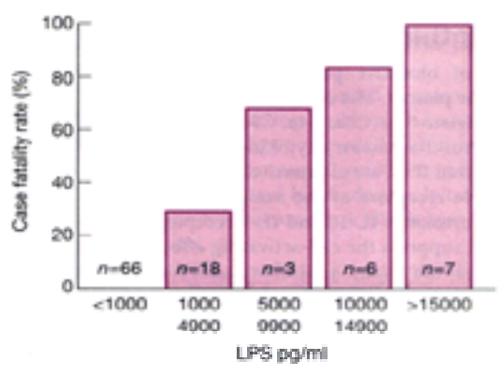

Fig. 1 The relationship between the levels of bacterial lipopolysaccharides in plasma and case fatality rate related to the development of septic shock and multiple organ failure in 100 Norwegian patients with systemic meningococcal disease.

Coagulopathy

Coagulation is activated primarily via the extrinsic (tissue factor–FVIIa) pathway. In patients with fulminant meningococcal septicaemia, there are increased levels of tissue factor in monocytes and on microparticles released from monocytes. The platelets disappear rapidly and remain at a low level for many days due to extensive consumption and a presumably altered endothelial surface. The activation of the coagulation system, as measured by formation of fibrin, is gradually reduced after antibiotic and fluid therapy is initiated (Table 3).

Inhibited fibrinolysis

Concurrent with activation of coagulation, fibrinolysis is inhibited by high levels of plasminogen activator inhibitor 1 (**PAI-1**), released from activated endothelial cells and platelets. High levels of PAI-1 are associated with development of persistent septic shock and a fatal outcome. Allelic variations in the promoter region of the PAI-1 gene enhance production and are associated with an increased risk of dying.

Thrombus formation

Thrombosis occurs particularly in the skin, kidneys, adrenals, muscles, peripheral extremities, and to some extent in the lungs. Levels of the natural coagulation inhibitors antithrombin III and protein C decline due to consumption, whereas tissue factor pathway inhibitor rises. Low levels of protein C are associated with diffuse thrombosis and necrosis of the skin even in non-septic patients.

Pro- and anti-inflammatory mediators

A multitude of bioactive pro- and anti-inflammatory mediators are released into the plasma. The complement and the kallikrein–kinin systems generate anaphylatoxins (C3a, C4a, C5a) and bradykinin, which are potent vasodilators. Proinflammatory cytokines, notably tumour necrosis factor-a, IL-1b, IL-6, and IL-8, are all massively upregulated. Concomitantly, high levels of soluble receptors of the same cytokines are released. The anti-inflammatory cytokines IL-10 and IL-1 receptor antagonist are present in high levels and suppress the cell-activating effect of the bacterial lipopolysaccharides and the many proinflammatory cytokines. It is uncertain whether inducible nitric oxide synthase, augmenting the production of nitric oxide in the endothelial cells, plays a role in meningococcal septic shock.

The subarachnoid space

Microbial proliferation is limited or absent although meningococci may be cultured from cerebrospinal fluid in as many as 50 per cent of untreated cases. The inflammatory response is very limited with a leucocyte count usually in the range of 10 to 100 × 10^6/l and normal contents of protein and glucose.

Laboratory findings

The erythrocyte sedimentation rate and C-reactive protein are only moderately elevated on admission, rising to high levels within 48 h. The leucocyte count is usually low with a marked shift to young band forms of neutrophils. There is evidence of a partly compensated metabolic acidosis with decreased levels of pH and PCO_2. Creatinine and urea are elevated, serum glucose is variable (high, normal, or low), potassium, calcium, and magnesium are low. Potassium rises with the renal failure. Serum aspartate aminotransferase and alanine aminotransferase are slightly elevated, whereas g-glutamyl transferase remains normal. Creatine kinase rises within 1 to 3 days, indicating rhabdomyolysis. Prothrombin, activated partial thromboplastin, and thrombin times are prolonged. The levels of platelets, fibrinogen, coagulation factors VII, X, and V, and prothrombin are low. Antithrombin III and protein C are low whereas tissue factor pathway inhibitor is elevated. Fibrin(ogen) degradation products, thrombin–antithrombin complexes, and PAI-1 are elevated. Lumbar puncture should be avoided in view of the bleeding diathesis.

Distinct meningitis and persistent shock

There are meningeal and circulatory symptoms. Usually the symptoms from the inflamed meninges dominate the picture. On admission there are classic signs and symptoms of meningitis such as headache and nausea, nuchal and back rigidity, and a positive Kernig's sign. The blood pressure remains low despite fluid volume repletion.

Circulating levels of bacterial lipopolysaccharide and inflammatory mediators are lower than in patients with fulminant septicaemia, and case fatality is lower. However, it is higher than in patients with meningitis without compromised circulation.

Meningococcaemia without distinct meningitis and persistent shock

The patient is febrile and usually presents with a rash but without symptoms of persistent septic shock or meningitis. It is a composite group of patients. Many of these patients have been admitted to hospital early, 12 to 24 h after their first symptoms. The case fatality rate is zero. Left untreated they might have developed symptoms of meningitis or fulminant shock.

Transient benign meningococcaemia

These patients develop fever and often an uncharacteristic rash, but no meningism. They are diagnosed as most likely having a viral infection and receive no antibiotic. When the blood culture results are known, the symptoms have disappeared spontaneously, usually within 1 to 3 days. This syndrome may occur in infants and young children.

Subacute meningococcaemia

A few patients develop fever, an uncharacteristic maculopapular rash, general malaise, and arthralgia but no signs of meningitis or shock. They feel uncomfortable but are not severely ill. Meningococci are isolated from blood cultures. Untreated the symptoms may last for days to several weeks but disappear within 1 to 2 days after penicillin therapy is initiated.

Chronic meningococcaemia

The patient develops undulating fever, arthralgia, and maculopapular rash. At times the symptoms may disappear completely. The symptoms may last for months. Blood cultures are sometimes repeatedly negative. Patients are often treated with corticosteroids because an underlying autoimmune disease is suspected. The fever disappears temporarily before reappearing. At this stage meningococci may well be isolated from blood cultures. Antibiotic treatment clears the symptoms within a few days.

Other organ manifestations

Pericarditis

The pericardium is seeded during meningococcaemia. Subsequent inflammation and exudate may lead to cardiac tamponade if left untreated. The patient is febrile, nauseated, and may complain of epigastric pain. The condition is often misdiagnosed as an acute abdominal condition. Blood cultures may be negative. *N. meningitidis* can be cultured and seen in aspirated pus by direct microscopy. Treatment consists of evacuating the pus and benzylpenicillin. The condition should be followed daily by ultrasound examination. Serogroup C organisms have been particularly implicated in these cases.

Arthritis

Acute meningococcal arthritis is an uncommon clinical manifestation of a preceding, often low-grade, meningococcaemia. It is usually located to one, or more rarely, several large joints. If the characteristic petechial rash is absent, isolation of meningococci from blood or joint cultures is necessary for a correct diagnosis. Arthritis caused by *Neisseria gonorrhoeae* is considerably more common than primary meningococcal arthritis. The symptoms disappear rapidly after penicillin treatment without long-term complications.

Arthritis induced by immune complexes

This is more common than the meningococcal arthritis. One or several large joints become swollen and painful. The symptoms usually develop at the end of the first week of treatment. Blood and joint cultures are negative. The temperature and inflammatory markers may rise after an initial decline. The symptoms disappear gradually after some days of treatment with non-steroidal anti-inflammatory drugs. Extended antibiotic therapy is not necessary.

Cutaneous vasculitis and episcleritis

This appears simultaneously with the immune complex arthritis and is commonly observed in sub-Saharan Africa. The vasculitis causes multiple blisters that readily rupture leading to multiple superficial skin ulcers.

Ocular infections

Conjunctivitis or panophthalmitis may precede other symptoms of invasive meningococcal infection. They are primarily observed in infants and children. The patient develops a red eye which in the case of panophthalmitis becomes painful with impaired vision. Local formation of microthrombi may complicate the infection.

Pneumonia

Strains belonging to serogroup Y and W-135 or more rarely other serogroups may cause pneumonia in adults and children. The diagnosis depends on detecting meningococci in a representative specimen from the low respiratory tract. It cannot be differentiated from pneumonia caused by other agents on the clinical symptoms alone.

Treatment

Prehospital antibiotic treatment

Health authorities in many countries advise general practitioners to start prehospital antibiotic treatment (i.e. benzylpenicillin) in suspected cases of meningococcal

infection. The doses in Table 4 rapidly lead to bactericidal concentrations in plasma.

The penicillin is injected laterally in one or both thighs in infants and children. The primary goal is to stop rapid growth of meningococci in the circulation before the intravascular inflammation becomes irreversible or causes grave sequelae. The patients most likely to benefit from this strategy, if applied early enough, are those who are distant from the hospital and have rapidly evolving symptoms leading to a compromised circulation and extensive haemorrhagic skin lesions. Retrospective studies in England suggest that prehospital penicillin treatment reduces case fatality.

Initial evaluation in hospital

The patients should be regarded as emergency cases. The main clinical presentation and severity should be evaluated immediately. A variety of prognostic scores have been developed. The Glasgow Meningococcal Septicaemia Prognostic Score is the one most commonly used. Scores can be used to select patients for intensive care treatment. They should never be used to justify withholding treatment as they often overestimate case fatality.

Antibiotic treatment

Adequate doses of benzylpenicillin, cefotaxime, ceftriaxone, or chloramphenicol effectively stop further proliferation of *N. meningitidis* in the circulation, cerebrospinal fluid, and other extravascular sites. Induction of an explosive release of bacterial lipopolysaccharides leading to a Jarisch–Herxheimer reaction has never been documented in patients receiving antibiotics for meningococcal infection. Plasma levels of lipopolysaccharides and the levels of important inflammatory mediators decline immediately after treatment with antibiotics is initiated in these patients (Table 5).

Benzylpenicillin, chloramphenicol, cefotaxime, ceftriaxone, and meropenem are bactericidal to *N. meningitidis*. Benzylpenicillin remains the drug of choice in most countries. It is effective, cheap, and non-toxic in high doses as long as renal function is normal. High doses are necessary since it penetrates the cerebrospinal fluid relatively poorly. In patients with fulminant septicaemia and severe renal dysfunction the doses should be reduced after 24 to 48 h.

Strains whose sensitivity to penicillin is reduced because of altered penicillin-binding protein 2 are an increasing problem. In most industrialized countries they account for less than 5 per cent of all meningococcal isolates, but the frequency is higher in Mediterranean countries, particularly Spain. Patients infected with these strains have been adequately treated with benzylpenicillin as long as dosage is adequate. Penicillinase-producing meningococci remain extremely rare.

Chloramphenicol is a good alternative in patients hypersensitive to b-lactam antibiotics. In developing countries it is the best and cheapest alternative to benzylpenicillin. Meningococcal strains resistant to chloramphenicol have recently emerged in France.

In many industrialized countries cefotaxime or ceftriaxone is combined with vancomycin as empirical treatment of bacterial meningitis until the aetiological agent has been identified. Cefotoxime and ceftriaxone are highly effective antibiotics that penetrate the blood–brain barrier better than benzylpenicillin. Meningococci remain fully sensitive to both drugs. Meropenem is a carbapenem highly active against *N. meningitidis*, *H. influenzae*, and *S. pneumoniae*. It does not induce seizures as observed with the imipenem–cilastatin combination.

In each country the health authorities and microbiological laboratories should recommend the optimal and affordable drug regimen.

Antibiotic treatment should be initiated promptly. Immediately after the first clinical evaluation and collection of the necessary samples for microbiological diagnosis, therapy should start. If there are contraindications to lumbar puncture or if it is delayed until after brain imaging, antibiotic treatment should be started immediately. Five days of treatment is adequate to eradicate sensitive meningococci.

Supportive treatment

Patients with persistent shock should be given extensive volume replacement, whereas patients with meningitis should receive a moderate amount of fluid. All patients should be monitored closely to detect early signs of a deteriorating circulation, renal and pulmonary failure, or increasing intracranial pressure.

Volume treatment

Patients with persistent hypotension and signs of inadequate peripheral circulation require massive fluid volume repletion. The extensive capillary leak syndrome increases the volume required. Children and adults may require an infused volume that is one to several times their circulating blood volume in the first 24 h. The optimal solution has not yet been defined. Colloids are often combined with crystalloid. Albumin and fresh frozen plasma were previously extensively used. However, the use of albumin in septic shock is controversial, expensive, and was not supported in a recently published meta-analysis. In many countries the use of fresh frozen plasma is no longer recommended because of the risk of transmitting pathogens, especially HIV.

Patients presenting with distinct signs of meningitis without shock should receive the basic daily requirement of fluid supplemented with extra volume for dehydration and loss due to vomiting and fever. Excessive hydration should be avoided since it may precipitate irreversible brain oedema and cerebellar herniation. In patients with persistent shock and meningitis, treatment of shock is the priority.

Inotropic support

If initial volume repletion fails to improve the circulation, inotropic support should be added. Dopamine, dobutamine, noradrenaline, and adrenaline are used. Most physicians start with dopamine at 2 to 10 µg/kg.min, or dobutamine at 1 to 10µg/kg.min. Ideally, patients should be infused through a central line.

Corticosteroid therapy

The use of corticosteroids in meningococcal septic shock is controversial. Methylprednisolone in pharmacological doses did not increase 28-day survival in two large series of patients with septic shock of various causes. Adrenal haemorrhage is common in patients with fulminant meningococcal septicaemia. It is also present in surviving patients. In most fatal cases of meningococcal infection, plasma cortisol levels are normal or high. Few patients have low levels. However, serial measurements of plasma cortisol and adrenal stimulation tests suggest a relative deficiency. Corticosteroids are not recommended routinely unless a deficiency is documented.

The benefit of dexamethasone in meningococcal meningitis is controversial. Efficacy has never been evaluated in double-blind, randomized, controlled clinical trials involving enough patients with meningococcal meningitis to allow a firm conclusion. In an open randomized study in Egypt involving 267 patients, dexamethasone injected every 12 h for 3 days did not improve the outcome. Corticosteroid treatment has been associated with relapse of the meningitis in patients who had otherwise been adequately treated. At present, dexamethasone is not recommended for routine use in patients with meningococcal meningitis.

Ventilatory support

Patients receiving volume treatment for profound shock are in danger of developing ARDS from capillary leak syndrome and volume overload. Increasing oxygen demand, decreased pulmonary compliance, and the appearance of diffuse infiltrates on chest radiograph indicate the development of ARDS. Some paediatricians advocate elective intubation and mechanical ventilation if more than 40 ml/kg per 24 h resuscitation fluid is needed to combat the septic shock, even if the oxygenation is normal.

Renal support

Patients with persistent septic shock and coagulopathy develop renal dysfunction from acute proximal tubular necrosis. Thrombosis in the small peritubular vessels, in glomeruli, and myoglobinaemia may contribute to the renal dysfunction. Serum creatinine and urea are elevated on admission and continue to increase for many days without adequate treatment. Hyperkalaemia, which may develop during the first 24 to 48 h, is an immediate threat. If possible, continuous haemofiltration should be

used. Peritoneal dialysis, although less effective, is an alternative to continuous haemofiltration.

Treatment of disseminated intravascular coagulation

P>The first priority is to stop further bacterial proliferation with antibiotics. In the 1970s heparin was extensively used. Two small controlled trials did not document any survival benefit in patients receiving heparin. Infusion of a continuous low-dose unfractionated heparin (10 to 15 IU/kg.h) has recently been advocated as supplement to treatment with concentrated protein C. The antithrombin III levels should be kept above 35 to 40 IU/ml.

Infusion of the natural anticoagulant protein C (loading dose: 100 IU/kg, followed by 15 IU/kg.h for days to keep the plasma concentration between 0.8 and 1.2 IU/ml) may possibly limit thrombus formation, skin necrosis, and the need for amputation. If used it should be started early. In the few uncontrolled studies that have been published, several patients treated with protein C concentrate still needed amputation. Randomized controlled trials have not been carried out.

Routine transfusion of platelets is controversial. In patients with life-threatening bleeding, massive platelet transfusion may be life saving. However, it may also aggravate thrombus formation by increasing levels of PAI-1 released from the platelets.

Fibrinolysis

To overcome inhibition by PAI-1, recombinant human tissue plasminogen activator (0.25 to 0.5 mg/kg in 1.5 to 4 h) has been infused to enhance fibrinolysis. Dramatic improvement was observed in some children. Recombinant human tissue plasminogen activator increases the risk of an intracerebral haemorrhage. If used, it should be started early. Efficacy has never been evaluated in a randomized controlled trial.

Plasmapheresis and blood exchange

Plasmapheresis or exchange blood transfusion have been tried, to remove pathologically activated plasma and leucocytes; 50 ml plasma/kg body weight has been exchanged with fresh plasma. These techniques do not increase the clearance of bacterial lipopolysaccharide substantially. Results suggest improved survival but adequate control groups are lacking. Even desperately ill patients have tolerated the procedures.

Extracorporal membrane oxygenation

A limited number of children have been treated with extracorporal membrane oxygenation in a few centres with apparently good results. However, equally good results have been achieved in another paediatric intensive care unit without using the procedure, suggesting that the experience of the intensive care unit is more important than the procedure *per se*.

Neutralization of bacterial lipopolysaccharides

Three different antiendotoxin principles, the anti-J5 serum, the human monoclonal IgM (HA-1A) antibody, and the recombinant bactericidal/permeability increasing protein (BPI_{21}) have been evaluated in randomized, double-blind, controlled clinical trials. None increased survival significantly. However, fewer patients treated with BPI_{21} required multiple severe amputations and more patients had a functional outcome similar to that before illness 60 days after treatment. None of the principles are presently commerically available.

Antimediator therapy

Strategies to neutralize tumour necrosis factor-a, IL-1, bradykinin, platelet-activating factor, and prostaglandins in patients with septic shock have not increased the 28-day survival rate. They have not been specifically evaluated in meningococcal septic shock.

Sequelae

Meningitis

Neurogenic deafness occurs in 1 to 10 per cent of the patients. It develops at an early stage and is usually irreversible. Reversible paresis of brain nerves IV, VI, or VII is occasionally observed. Epilepsy, hydrocephalus, and diffuse brain damage are at present rare complications in industrialized countries.

Persistent headache, altered sleep pattern, concentration difficulties, irritability, and neurasthenia may persist in 5 to 8 per cent of all patients.

Shock and coagulopathy

Most long-term complications are related to development of gangrene of the extremities requiring amputation and necrotic skin lesions requiring extensive grafting. The renal failure is usually reversible. Permanent adrenal insufficiency develops very rarely in survivors. Acute respiratory distress syndrome may lead to permanent pulmonary fibrosis and reduced function.

Vaccination

Capsule polysaccharide vaccine (A, C, Y, and W)

The serogroup A polysaccharide vaccine is immunogenic from 6 months of age. Infants vaccinated at 3 and 7 months develop higher antibody levels than do infants vaccinated only at 7 months, suggesting a booster effect. The serogroup C polysaccharide vaccine induces a short-lived immune response at 3 months but normal immune response in children above 18 months. No booster response is present. Revaccination may reduce the antibody level. When vaccination is required for serogroup A infection, infants of less than 24 months should receive two doses with at least a 1-month interval, whereas those above 2 years should receive one dose. For serogroup C infection, one dose should be given from 18 months. In children with malaria, the immune response is reduced. An antibody level of 1 to 2 µg/ml appears to be necessary for protection.

Indications for vaccination

Routine immunization with the A, C, Y, and W polysaccharide vaccine is advocated for people with documented deficiencies in the late complement components and properdin.

Non-outbreak situation

Indications for vaccination with A or C polysaccharide vaccine are, according to Peltola: close contacts of an index case, travellers to high-risk areas, military recruits, persons with asplenia, and alcoholics.

Outbreak situation

Vaccination has been recommended if two or more are attacked by the same strain in a school class or day-care centre, the attack rate exceeds 10 cases/100 000 population per 3 months, or the attack exceeds 1/1000 with 3 or more cases in a closed group setting.

Epidemic situation

An advocated threshold for mass vaccination is 15 cases/100 000 population per week for 2 consecutive weeks caused by the same strain. A steadily increasing number of cases and an increase in the median age of the patients indicate an epidemic.

Conjungate polysaccharide protein vaccine

Serogroup C polysaccharide conjugated to a protein carrier induces a significant booster response in infants vaccinated at 2, 3, and 4 months of age. The same has been shown for toddlers. The United Kingdom is the first country to start mass vaccination with serogroup C conjugate vaccine of infants, children, and adolescents owing to the increasing number of serogroup C cases.

Outer membrane vesicle vaccine

Since the capsule polysaccharide of serogroup B strains induces a short-lived IgM but no lasting IgG response, several groups have developed an outer membrane vesicle vaccine. The protection rate after two doses is lower (57 to 80 per cent) than for the polysaccharide A, C, Y, and W vaccines and is relatively strain specific. The protection rate in children below 4 years of age is much lower than for adults. Three doses induce a significantly better immune response than two doses given with a 6-week interval. Only one vaccine is available for sale.

Secondary prophylaxis

Antibiotic prophylaxis

Household contacts of an index case have a 100 to 1000 times increased relative risk for developing meningococcal infections. Usually the second case occurs within 2 weeks of the index case if no eradication treatment is given. However, there is doubt about the effectiveness of eradication treatment when the causative strain belongs to serogroup B.

Health authorities in most countries advise that close contacts have eradication treatment with rifampicin at 10 mg/kg, maximum dose 600 mg every 12 h for 48 h. Recently, 500 mg of ciprofloxacin or 400 mg of ofloxacin as a single dose has replaced rifampicin for adults in many countries. Pregnant women should receive 250 mg, and children of less than 12 years 125 mg, of ceftriaxone as one intramuscular injection.

Future prospects

The development of effective and affordable conjugate vaccines covering serogroups A and C will be a major step forward. They will cover the age group 2 months to 2 years where protection is most required and pave the way for routine vaccination. Development of a serogroup B vaccine with documented effect for infants and children is urgently needed.

Further reading

Brandtzaeg P (1996). Systemic meningococcal disease: clinical pictures and pathophysiological background. *Reviews in Medical Microbiology* **7**, 63–72.

Cartwright K, ed. (1995). *Meningococcal disease*. Wiley, Chichester.

Caugant DA (1998). Population genetics and molecular epidemiology of *Neisseria meningitidis*. *Acta Pathologica, Microbiologica et Immunologica Scandinavica* **106**, 505–25.

Girgis NI *et al.* (1989). Dexamethasone treatment for bacterial meningitis in children and adults. *Pediatric Infectious Disease Journal* **8**, 848–51.

Oppenheim BA (1997). Antibiotic resistance in *Neisseria meningitidis*. *Clinical Infectious Diseases* **24** (Suppl. 1), 98–101.

Peltola H (1998). Meningococcal vaccines. Current status and future possibilities. *Drugs* **55**, 347–66.

Pollard AJ *et al*. (1999). Emergency management of meningococcal disease. *Archives of Disease in Childhood* **80**, 290–6.

Van Deuren M, Brandtzaeg P, van der Meer JWM (2000). Update on meningococcal disease, with special emphasis on pathogenesis and clinical management. *Clinical Microbiology Review* **13**, 144–66.

7.11.6 Neisseria gonorrhoeae

D. Barlow and C. Ison

Neisseria gonorrhoeae, the gonococcus, has changed in three important ways since the advent of effective treatment: sensitivity to antibiotics has decreased (and continues to do so); its symptom-producing capabilities have lessened; and its incubation period has lengthened. A study from the United Kingdom in the early nineties gave a mean incubation period of 5.6 days and a median of 8.6 days in men.

Pathogenesis

N. gonorrhoeae is a particularly successful pathogen, with mechanisms that evade host defences and cause repeated infection. The major antigens of the outer membrane (**OM**) of the gonococcus that are exposed to the immune response are pili, lipo-oligosaccharide (**LOS**), and three major OM proteins, Por, Opa, and Rmp. *N. gonorrhoeae* primarily colonizes columnar epithelium of the lower genital tract and only occasionally progresses to the upper genital tract or invades to cause systemic disease. Successful colonization requires attachment and invasion of the epithelial layer to avoid being swept away by cervical secretions in women or urine in men. *N. gonorrhoeae* expresses receptors on the cell surface for transferrin or lactoferrin from which iron is released, unlike many other bacteria that produce soluble sideropores. Lack of iron can be a growth-limiting factor. For invasion to occur, gonococci must resist the bactericidal activity of serum. *In vivo* gonococci are serum resistant as a result of sialylation of LOS. *In vitro* most strains revert to serum sensitive, although a few remain resistant suggesting an additional unidentified mechanism.

Pili, Opa, and LOS have the ability to alter the surface-exposed part of the molecule and hence present a new antigen to the immune system. In the gonococcus this alteration occurs at a frequency higher than the normal mutation rate and is known as antigenic variation. On each encounter between the organism and the host, the gonococcus presents a range of immunologically distinct proteins which are not recognized by the host. The interaction of these bacterial receptors with the host cell is complex and the host-cell receptors are currently being unravelled, ranging from complex carbohydrates and glycosamines to lipoproteins and glycoproteins.

Epidemiology

Gonorrhoea is almost exclusively transmitted by sexual activity and, like HIV infection, is not evenly distributed amongst the sexually active population. The highest incidence is found in young (teenaged women, men in their twenties), urban, socio-economically deprived persons and ethnic minorities. Figure 1 shows the number of cases in England up to 2000, the drop in reported cases since 1974 being reversed in the early nineties. Use of antimicrobials has less effect on endemic levels than might be expected, although it reduces the incidence of complications. Gonorrhoea facilitates HIV transmission, producing an increase in detectable virus in urethral secretions when infection is present; this is reversed following antibiotic treatment.

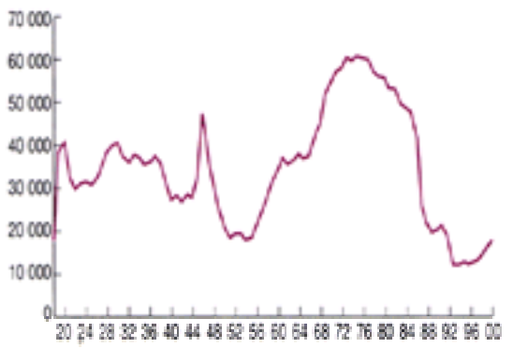

Fig. 1 Reported cases of gonorrhoea in England between 1918 and 2000 (Department of Health).

The incidence and prevalence of gonorrhoea serve as useful surrogates for unsafe sexual behaviour since diagnosis is swift and accurate and the infection can be treated and reacquired repeatedly. The incidence of neonatal gonococcal ophthalmia and the prevalence of antenatal infection measure the success, or otherwise, of a control programme. By both criteria, gonorrhoea is not a serious problem in the United Kingdom.

The infectivity of the gonococcus is probably higher from male to female and may reach 80 per cent. Condoms, when used invariably and throughout sexual contact, prevent transmission of gonorrhoea.

Symptoms, signs, and complications

Gonorrhoea in women

Because of the lack of specific symptoms there is no meaningful incubation period for uncomplicated gonorrhoea in women . The sites most commonly affected are cervix (90 per cent), urethra (75 per cent), rectum (40 per cent), and oropharynx (5 to 15 per cent). Signs at all these sites are unhelpful—the 'cervicitis' ascribed to gonorrhoea being found in other conditions and in healthly individuals. Symptoms likewise are absent or non-specific, including alteration in vaginal discharge or, rarely, mild dysuria. Women with gonorrhoea depend on notification by a partner or development of complications to alert them to the possibility of infection. Spread of the gonococcus to infect the endometrium, fallopian tubes, and pelvic adnexas is the most common complication (5 per cent of infections) and occurs at or soon after the menstrual period, probably resulting from retrograde flow of menses. Coincidental infection with *Chlamydia trachomatis* is common enough to warrant treatment of both organisms. Gonococcal infection of Bartholin's, Skene's, or periurethral glands is rare in the United Kingdom.

Disseminated gonococcal infection is four or five times more common in women than men, a reflection of women's lack of genital symptoms. Almost always caused by penicillin-sensitive organisms, disseminated gonococcal infection is a comparatively benign bacteraemia affecting joints and skin. The shoulder and knee are most commonly affected followed by the wrist, elbow, and small joints of the hands and feet, often with an associated tenosynovitis. The pathognomonic, painless, skin

lesions, usually 4 to 10 in number, evolve through vesicular, pustular, and haemorrhagic stages before healing (Plate 1, Plate 2). Erythema nodosum-like lesions have been described. The constitutional upset tends to be minimal and the white cell count and erythrocyte sedimentation rate are not greatly raised. Response to antibiotic treatment is rapid but joints may need to be aspirated. Blood or joint fluid culture may yield gonococci but the quickest diagnosis comes from anogenital and throat culture.

Perihepatitis, the FitzHugh–Curtis syndrome, more frequently appears with *C. trachomatis* than *N. gonorrhoeae*. Right hypochondrial pain, referred to the shoulder, occasionally with pleural effusion and rub, results in referral to a surgeon or a general, rather than geniturinary, physician.

Gonorrhoea in men

Affected sites are the urethra and oropharynx, and the rectum in homosexual men. Rectal and throat infections tend to be silent while discharge is the commonest urethral symptom. When the infection is fully developed, the discharge is white/yellow/green and profuse, staining the underwear. Differential diagnosis includes foreign body and, unusually, non-gonococcal urethritis. Scanty mucoid or mucopurulent discharge is seen in early infection. Discomfort on urination no longer seems to be severe or as common as before, with one large United Kingdom study eliciting dysuria as a presenting symptom in only 50 per cent of men. Asymptomatic patients (less than 10 per cent) include presymptomatic, post-symptomatic, and unobservant men. Urethral gonorrhoea acquired following fellatio is increasingly seen in gay men practising 'safe' sex and may be passed on as rectal gonorrhoea to a regular partner. Infection spreading to the epididymis and testis is more often due to *C. trachomatis* than the gonococcus, and other complications—tysonitis, prostatitis, periurethral abscess, or infection of the median raphe—constitute very few cases in the United Kingdom.

Diagnosis

Microscopy

Even in genitourinary medicine departments, the majority of patients will not have gonorrhoea. Much time and effort is spent excluding rather than diagnosing the disease. Investigations therefore need high sensitivity. The diagnosis of gonococcal infection is easier in men than in women.

Microscopy of a suitably stained specimen is the first line in diagnosis. The organisms must be Gram-negative, intracellular (within the cytoplasm of a leucocyte), and diplococci (GNID) (Plate 3). In samples from the male urethra, microscopy is highly sensitive (identifying 98 per cent of positives) and highly specific (less than 1 per cent will be found on culture to be *Neisseria meningitidis* or other species).

Microscopy of stained samples from rectum, cervix, and female urethra although much less sensitive, with identification of only 55 per cent or less of true positives, should still be performed since it has the advantage, where positive, of enabling immediate diagnosis and treatment. Because of the preponderance of other neisserias in the oropharynx, microscopy of samples from this site is not helpful.

Routine culture of samples from the male urethra provides an important means of quality control of the laboratory service, so crucial for diagnosis in women. It also enables assessment of antibiotic sensitivities and other characteristics of the organism for epidemiological and management purposes.

Laboratory detection of *N. gonorrhoeae*

Isolation of the causative organism, *N. gonorrhoeae*, has been regarded as the gold standard for the diagnosis of gonorrhoea for many years. However, the application of molecular techniques, such as polymerase chain reaction or the ligase chain reaction, are being more widely used.

Isolation and identification of *N. gonorrhoeae*

N. gonorrhoeae is fastidious in its growth requirements. It needs an enriched medium, such as Thayer–Martin or Modified New York City, which consist of GC agar base supplemented with a source of iron (lysed horse blood) and essential amino acids and glucose (IsoVitaleX or Vitox), and incubation in moist conditions with 5 to 7 per cent carbon dioxide at 37°C. Good specimen collection and efficient transport to the laboratory are crucial to successful isolation.

Specimens are taken using disposable loops or swabs and inoculated in the clinic or sent to the laboratory in transport medium. The isolation rates vary little if the specimen is dealt with rapidly. Isolation is enhanced by the addition of antibiotics to the medium to suppress other organisms that colonize the anogenital tract. While occasional problems can arise due to antibiotic-sensitive strains, a selective medium is essential and non-selective media should only be used as an adjunct not as a replacement.

After incubation, colonies that are oxidase positive (presence of cytochrome c oxidase), Gram negative, and cocci are considered to be *Neisseria* spp. In many parts of the world this would be considered presumptive identification of *N. gonorrhoeae*. However, in the industrialized world confirmation of identity as *N. gonorrhoeae* is usual. Historically, this has been achieved using carbohydrate utilization tests; *N. gonorrhoeae* differs from other species in that it alone produces acid from glucose. Identification kits combining carbohydrate utilization and enzyme profiles are commonly used. An alternative approach is to use immunological reagents; two reagents are available which contain antibodies raised to epitopes on the two types of the major outer membrane protein, Por or PI, linked to fluorescein (GC Microtrak, Syva Company, USA) and to staphylococcal Protein A (Phadebact Monoclonal GC OMNI test, Boule AB Sweden). These sensitive and specific reagents can identify colonies direct from the primary isolation medium and a result can be obtained on the same day as the organism is isolated. Correct identification of *N. gonorrhoeae* is most important in cases of sexual or child abuse. In such instances it is sensible to use more than one of the identification tests available to confirm an isolate as *N. gonorrhoeae*.

Molecular detection of *N. gonorrhoeae*

Antigen detection assays for *N. gonorrhoeae*, both immunological and molecular, have been largely unsuccessful because they offer little advantage over the Gram stain and culture and they cannot provide an organism for susceptibility testing. However, the sensitivity and specificity of detection assays by DNA amplification, polymerase chain reaction (**PCR**) and ligase chain reaction (**LCR**), now appear to be equal or superior to conventional techniques and may be less affected by suboptimal handling or transport. These assays may also be useful on non-invasive specimens such as urine or self-taken swabs. Currently there are no molecular tests available for determining antibiotic susceptibility, but the sequence of the appropriate resistance genes or mutations is known and could be detected using PCR or LCR. Assays that offer the combination of detection of *N. gonorrhoeae* and susceptibility to antibiotics are likely to be available in the near future.

Typing

Typing is useful for studying reinfection, treatment failure, coinfection, and to show correlations with pathogenicity and antimicrobial susceptibility patterns. Auxotyping is the determination of nutritional requirement. A large number of auxotypes have been described but in most studies three or four types predominate: non-requiring (NR) or prototrophic (Proto), proline-requiring (Pro), arginine-requiring (Arg), and those requiring arginine, hypoxanthine, and uracil (AHU). Serotyping, using a panel of monoclonal antibodies, divides strains into 24 IA serovars and 32 IB serovars. Auxotyping and serotyping are often used in combination to produce auxotype/serovar (A/S) classes giving greater discrimination. A variety of genotypic methods have also been used, from plasmid analysis, which is poorly discriminatory, to pulse-field gel electrophoresis or DNA sequencing, which are highly discriminatory.

Antibiotic resistance

Penicillin has been used as first-line therapy for gonorrhoea for many years. *N. gonorrhoeae* is inherently sensitive to most antibiotics such as penicillin, but with increased usage both chromosomally-mediated and plasmid-mediated resistance has developed. Resistance is most prevalent in the developing world where the incidence of gonorrhoea is high and appropriate antibiotics are often unavailable or misused. However, in the industrialized world these strains are often imported and then spread by the indigenous population. In 1989 the World Health Organization issued new guidelines for the treatment of gonorrhoea stating that penicillin should only be used as first-line treatment if the gonococcal population is known to be sensitive. If resistance is high or the susceptibility of the gonococcal population is unknown, alternative treatment is recommended: ciprofloxacin (a quinolone), ceftriaxone (a third-generation cephalosporin), or spectinomycin (a macrolide). Of these

antibiotics, ciprofloxacin is used increasingly is the United Kingdom because it is administered orally and is highly effective and inexpensive.

Chromosomally-mediated resistance

Decreased susceptibility to penicillin was detected as early as 1958, but this could be overcome by increasing the dose of penicillin and by adding probenecid. It was not until the 1970s that strains begun to appear with minimum inhibitory concentrations (**MICs**) to penicillin of greater than 1.0 mg/l and posed a therapeutic problem. Chromosomal resistance to penicillin in *N. gonorrhoeae* is the result of the additive effects of mutations at multiple loci, *penA*, *mtr*, and *penB*, the products of which reduce the permeability of the cell wall to penicillin.

Resistance to the alternative therapies—ceftriaxone, spectinomycin, and ciprofloxacin—has begun to emerge. Therapeutic failure to ceftriaxone has not yet been documented, but the loci responsible for chromosomal resistance to penicillin also confer decreased susceptibility to the earlier cephalosporins. If this type of resistance to penicillin continues to increase and is treated inappropriately, resistance to cephalosporins could emerge. Therapeutic resistance to spectinomycin has been reported sporadically, is high level, and due to a mutation on the chromosome that affects ribosomal binding. Spectinomycin has been an extremely useful antibiotic in treating resistant gonorrhoea and may be important in the future if mechanisms of resistance continue to evolve to newer antibiotics. Ciprofloxacin is now a popular alternative for the treatment of gonorrhoea because it is highly effective in a single oral dose of 500 mg. However, high-level resistance, resulting in therapeutic failure, has emerged in strains primarily originating from the western Pacific with mutations in the DNA gyrase gene, *gyrA*, and the topoisomerase IV gene, *parC*. The level of resistance may be enhanced by additional mutations in the *gyrB* gene or in changes in cell wall permeability, possibly due to efflux mechanisms. Surveillance of gonococcal isolates in the Western world should prolong the life of this useful antimicrobial agent.

Plasmid-mediated resistance

N. gonorrhoeae exhibiting plasmid-mediated resistance to penicillin were first described in 1976. Simultaneous reports appeared of two strains, one from Africa carrying a plasmid of 3.2 megadaltons (MDa) and the second from the Far East carrying a plasmid of 4.4 MDa. Both plasmids encode for the TEM-1 type b-lactamase (penicillinase). The smaller plasmid of 3.2 MDa has a deletion from the 4.4 MDa plasmid in a non-functional region. Penicillinase-producing *N. gonorrhoeae* carrying the 3.2 MDa and 4.4 MDa plasmids have now disseminated worldwide, although their prevalence is greatest in countries of the developing world. Penicillinase-producing *N. gonorrhoeae* carrying plasmids of differing size (2.9, 3.0, and 4.8 MDa) have been described more recently but have not spread in the same manner.

In 1985 plasmid-mediated resistance to tetracycline was first detected. It is high-level (MIC ³ 16 mg/l) and is due to the acquisition of the *tetM* determinant by the conjugative plasmid of *N. gonorrhoeae* resulting in a plasmid of 25.2 MDa. Strains carrying this plasmid are known as tetracycline-resistant *N. gonorrhoeae*. Tetracycline is not the treatment of choice for gonorrhoea but is commonly used, particularly in African countries, because it is inexpensive and available.

Susceptibility testing

The primary aim of susceptibility testing of *N. gonorrhoeae* is to predict therapeutic failure. However, it is also important to monitor drifts in susceptibility and to detect the emergence of resistant strains to the main first-line therapies. There is much controversy over the correct method for achieving this for gonococci. Determination of zones of inhibition around antibiotic-containing discs has been the method chosen by most clinical laboratories, but gonococci vary in their growth patterns and this method can be difficult to control and interpret. In recent years the breakpoint agar dilution technique, which uses one or two concentrations of antibiotic to estimate the MIC and categorize strains into susceptible, reduced susceptibility, and resistant, has been used increasingly. Determination of the full MIC is not necessary for most laboratories and is best performed by reference centres.

Plasmid-mediated resistance to penicillin can be easily detected using the chromogenic cephalosporin (nitrocefin) test, which can be performed direct from the primary isolation plate. Plasmid-mediated resistance to tetracycline can be detected using either the absence of a zone of inhibition around a 10 µg tetracycline disc or presence of growth on GC agar containing 10 mg/l of tetracycline. In a similar manner, high-level resistance to ciprofloxacin can be detected by screening for isolates that can grow on agar containing 1 mg/l ciprofloxacin.

Treatment

Treatment of uncomplicated gonorrhoea in both sexes is ideally by a single dose of antibiotics, the choice of which will depend on where the infection was acquired and from whom. In the Far East and Africa, a high percentage of strains will have chromosomal and/or plasmid-associated resistance, whereas organisms in the United Kingdom, unless imported, are still largely sensitive to penicillin, and with doses less than those required, say, in America. Standard treatment should cure at least 95 per cent of presenting cases of gonorrhoea and, in the United Kingdom, 2 or 3 g of amoxycillin or ampicillin, with 1 g of probenecid achieves this aim. Alternatively, 500 mg of ciprofloxacin has the advantage of higher cure rates in oropharyngeal infection. Treatment with 250 mg of ceftriaxone, 500 mg of spectinomycin, or 500 mg of cefotaxime, all intramuscularly, is suitable for infections acquired outside the United Kingdom, cefotaxime being particularly useful for organisms with both plasmid and high chromosomal resistance such as those found in the Philippines.

Many physicians add 1 g of azithromycin, or 100 mg of doxycycline twice daily for 1 week, for possible coincidental chlamydial infection.

American and British guidelines suggest that gonococcal pelvic infection or perihepatitis be treated with parenteral antibiotics although the evidence for this is not strong. All are agreed that antichlamydial therapy should be included. A single intramuscular dose of 250 mg of ceftriaxone or 2 g of cefoxitin, followed by 100 mg of doxycycline and 400 mg of metronidazole, both twice daily for 2 weeks, is recommended. For infection acquired in the United Kingdom, with no foreign connections, cure should occur with any of the standard single-dose treatments followed by doxycycline with metronidazole for 2 weeks, as above.

Gonococcal epididymo-orchitis can be treated with 500 mg of ciprofloxacin followed by 100 mg of doxycycline twice daily (or 2 g of erythromycin stearate daily in divided doses) for at least 2 weeks. A scrotal support eases symptoms.

Tracing of contacts of all cases of gonorrhoea and exclusion of other sexually transmitted infections must be undertaken.

Further reading

Bignell C (2000). European guidelines for the management of gonorrhoea. *International Journal of Sexually Transmitted Diseases and AIDS* **12** (Suppl 3), 27–9 and http://www.mssvd.org.uk/.

Centers for Disease Control (1998). Sexually transmitted disease treatment guidelines 1988. *Morbidity and Mortality Weekly Report* **47**, 1–111.

Hook EW, Handsfield HH (1999). Gonococcal infections in the adult. In: Holmes KK *et al*., eds. *Sexually transmitted diseases*, 3rd edn, pp. 451–6. McGraw-Hill, New York.

Ison CA (1996). Antimicrobial agents and gonorrhoea: therapeutic choice, resistance and susceptibility testing. *Genitourinary Medicine* **72**, 253–7.

Ison CA (1998). Gonorrhoea. In: Woodford N, Johnson AP, eds. *Methods in molecular medicine*, Vol 15. *Molecular bacteriology: protocols and clinical applications*, pp 293–308. Humana Press, New Jersey.

Nassif X *et al.* (1999). Interactions of pathogenic neisseria with host cells. Is it possible to assemble the puzzle? *Molecular Biology* **32**, 1124–32.

Taylor-Robinson D, Thomas B, Ison C (1999). Diagnostic procedures in genitourinary medicine: practical laboratory aspects. In: Barton SE, Hay PE, eds. *Handbook of genitourinary medicine*, pp. 19–48. Arnold, London.

7.11.7 Enterobacteria, campylobacter, and miscellaneous food-poisoning bacteria

G. T. Keusch and M. B. Skirrow

Humans are colonized by a huge number of micro-organisms, prominent among which are the Enterobacteriaceae, a large grouping of small, facultatively anaerobic, Gram-negative bacilli capable of residence in the gastrointestinal tract, and therefore often grouped together as the enterobacteria. However, Enterobacteriaceae are not just commensals in the intestinal flora; they may also be important causes of disease, both locally in the gut and at times invasively in the blood and elsewhere in the body. This chapter begins with a short general description of enterobacteria and the infections they cause outside the intestinal tract. It then focuses on these organisms as enteric pathogens. In the latter role the enterobacteria most often cause diarrhoea, which remains a major cause of morbidity in advanced industrial economies and mortality in developing countries, especially among children.

Although *Salmonella typhi* and *S. paratyphi*, the causes of typhoid and paratyphoid fever, are enterobacteria, their special attributes are described in detail in Chapter 7.11.8. Likewise, *Yersinia* infections are described in Chapter 7.11.17. However, enteritis due to Gram-negative non-Enterobacteriaceae, namely *Campylobacter*, *Aeromonas*, *Plesiomonas* spp., *Vibrio parahaemolyticus*, and other non-cholera vibrios are included here. Descriptions of food poisoning due to the Gram-positive bacteria *Clostridium botulinum* and *C. perfringens* are provided in Chapter 7.11.21. *Bacillus cereus*, and *Staphylococcus aureus* food poisoning are presented here. An overview of infections of the intestinal tract is given in Chapter 14.17.

The enterobacteria

Definition and general description

Strictly speaking, the term 'enterobacteria' applies to members of the large family Enterobacteriaceae that are found in the intestinal tract of humans and animals, or are associated with plants and soil. Classical taxonomy based on biochemical and immunological criteria has resulted in a family that includes a number of major tribes with widely varying properties, including 30 genera and at least 120 species. Within these various tribes, the genera most likely to be encountered in medical practice are *Salmonella, Shigella, Escherichia, Klebsiella, Enterobacter, Citrobacter, Serratia, Hafnia, Edwardsiella, Erwinia, Kluyvera, Proteus, Providencia, Morganella*, and *Yersinia.* However, this classical schema is likely to be altered as DNA homology becomes the basis of microbial classification. *Escherichia coli*, *Salmonella* spp., and *Shigella* spp. are the principal enteric pathogens among these groups. The others may cause enteric disease but are more commonly the cause of systemic infection, usually through a specific portal of entry, especially in immunologically compromised hosts.

Enterobacteriaceae are Gram-negative, oxidase-negative, non-spore-forming straight rods, most of which are motile by means of peritrichous flagella. Although they are aerobes, many are capable of growth under anaerobic conditions. Microbiologists divide the group into those capable of mixed-acid fermentation resulting in the production of acetate from pyruvate (such as *Escherichia coli, Salmonella,* and *Shigella*) and those that produce butanediol as the end-product of fermentation (*Serratia, Enterobacter,* and *Erwinia*). They are easy to cultivate in the laboratory on simple bacteriological media (most can grow using D-glucose as the sole source of carbon, producing acid), indeed they often outgrow and mask the presence of more fastidious bacteria. However, they are vulnerable to environmental stress, such as heating to 60 °C for 20 min, and desiccation. Many are acid-sensitive, but some are resistant to acid pH, which may serve as a virulence factor in the gastrointestinal tract.

The clinical microbiology laboratory takes advantage of the inability of *Salmonella* and *Shigella* spp. to ferment lactose by screening cultures for the presence of non-lactose-fermenting organisms. This is a useful initial and simple distinguishing feature, because virtually all other enterobacteria—with the exception of *Proteus, Providencia*, and *Morganella*—ferment lactose freely. By including lactose and a pH indicator in a culture medium, the colonies of non-lactose fermenters stand out from the lactose fermenters, which produce acid and change the colour of the included pH-sensitive indicator dye. This makes it simple to pick the pale non-lactose-fermenting colonies for further study. The lactose-fermenting enterobacteria are commonly grouped together as 'coliforms', a clinically convenient term that has little logic to commend it. It is particularly unfortunate that this rubric may serve to hide significant pathogens, as the notation is often interpreted to mean harmless organisms resembling ordinary *E. coli*.

Microbial structure and antigenicity

The Enterobacteriaceae are typical Gram-negative rods, in that they possess a complex cell wall containing three major layers: (1) the inner cytoplasmic membrane; (2) an intermediate region made up of peptidoglycan; and (3) an outer cell membrane, which is itself composed of an inner phospholipid–protein layer and an outer covering of lipopolysaccharide (**LPS**). In addition, some members of the group possess an outermost antigenic carbohydrate capsule, while others possess one or more flagella and hence are motile. The cytoplasmic membrane functions, as it does for all micro-organisms, to regulate the transport of metabolites, sugars, amino acids, small peptides, and ions into and out of the microbial cell. Peptidoglycan, a long-chain linear polymer and structural element common to Gram-positive and Gram-negative bacteria, covers this membrane. It is composed of alternating *N*-acetylmuramic acid and *N*-acetylglucosamine residues with a pentapeptide side chain terminating in a D-alanyl- d-alanine dipeptide. This dipeptide is the site for the formation of crosslinking peptide bonds between adjacent linear aminosugar chains, providing structural stability to the bacterial cell. Crosslinking is catalysed by transpeptidase enzymes, which are the targets of action of the b-lactam antibiotics, such as the penicillins and cephalosporins.

A major difference in the peptidoglycan layer between Gram-positive and Gram-negative bacteria is the greater thickness of the structure in Gram-positive bacteria. The region between the peptidoglycan layer and the outer cell membrane is known as the periplasmic space. Critical cell functions also take place within this space, for example the assembly and modification of microbial proteins and antigens that are inserted in, or excreted through, the outer layer. The latter is a complex structure commonly referred to as endotoxin because it contains lipid A, the endotoxic moiety of LPS, linked to a common-core carbohydrate structure. A highly variable polymer of sugar residues, the O-specific oligosaccharide chain, is displayed on the outermost microbial surface, and contains critical, specific, heat-stable antigenic carbohydrate determinants designated O-antigens. Oligosaccharides, even when composed of just a few sugar residues, are well suited to express immunologically specific and recognizable antigens, because they permit a large number of small stereospecific changes that can be distinguished by antigen-specific antibodies. It is this structural feature of the O-specific oligosaccharides that permits the separation of many different O-antigen serotypes within a species—160 in *E. coli* alone. In the case of motile strains, flagellar proteins are also expressed well, so additional heat-labile protein (H) antigens are identifiable. Finally, some Enterobacteriaceae produce an outer capsular layer possessing yet more (carbohydrate) antigens designated K antigens. K antigens may cover and mask the underlying O-antigens and obscure the identity of the organism unless first removed by boiling the culture.

Taken together, these three types of antigen form the basis of a useful system for the identification and differentiation of enterobacteria in the laboratory, originally devised by Kaufmann and White. Because these antigens induce the formation of specific antibodies, which may be employed as immunoreagents that are specific for particular microbial structures, they turn out to be highly useful for the serological diagnosis of specific infections. When there is a rise in antibody to the O, H, and, if present, K antigens, which represent the 'signature' of particular members of the group, a serological diagnosis can be made. However, adding to the complexity, a single organism can express multiple O, H, and K antigens and individual antigens; moreover, O-specific oligosaccharides may be shared by two or more specific organisms within a genus or across species. Therefore immunological identification of individual organisms and serological diagnosis of an infection is based on the pattern of antigens detected, or on a significant rise in antibody titre to a particular battery of antigens. This serological 'fingerprint' is often supplemented by biological

and biochemical information for the purposes of identification. Capsular antigens, when present, can add additional information, for example the Vi antigen of *Salmonella typhi*, which is shared with just one organism, *Citrobacter freundii*. Serological identification of specific strains of enterobacteria is of enormous utility for epidemiological investigations.

Extraintestinal infections caused by enterobacteria and related organisms

Before proceeding to specific diarrhoeal diseases caused by enterobacteria, a brief account of their infective role elsewhere in the body is provided. It is convenient to include other non-fastidious Gram-negative bacilli that behave clinically in a similar manner to enterobacteria and which are often found in mixed infections with them. Chief among these is *Pseudomonas aeruginosa*, which is notorious as a 'hospital' organism and as a cause of opportunistic and sometimes fatal systemic infection in debilitated or immunosuppressed patients. *Ps. aeruginosa* is naturally resistant to most of the commonly used antimicrobials and also to many antiseptics. Indeed certain pseudomonads, notably *Ps. cepacia*, are capable of growth in antiseptic solutions stored at dilute working strengths in hospital wards. Other common opportunistic Gram-negative bacilli are found in the genera *Alcaligenes, Acinetobacter, Aeromonas, Flavobacterium*, and *Chromobacterium*.

Gram-negative sepsis occurs when these organisms reach the bloodstream and result in clinical symptoms. The incidence of Gram-negative sepsis steadily increased from the 1970s to the 1990s, particularly in hospital patients. The widespread use (and abuse) of broad-spectrum antibiotics, which were generally more active against Gram-positive bacteria, is one reason, but another is the increasing proportion of susceptible patients being treated: more elderly patients; more receiving immunosuppressive or cytotoxic therapy; more with catheters, pacemakers, and prostheses that provide favourable sites for infection; and more undergoing more complex and adventurous surgery. Enterobacteria, pseudomonads, and similar Gram-negative bacilli easily acquire resistance to antimicrobials, which helps them to colonize the hospital environment. Such organisms quickly replace the normal sensitive bowel flora of patients admitted to hospital, particularly if antibiotics are being given. In this way the patient's anus becomes the gateway to colonization and infection elsewhere in the body. Interestingly, during the past decade Gram-negative bacteraemia diminished relative to Gram-positive bacteraemia. Today, almost 50 per cent of bloodstream isolates are Gram-positive (Table 1). This reflects acquired resistance among the Gram-positives and the use of newer broad-spectrum drugs with better coverage of the Gram-negative bacteria, which are often given empirically for the treatment of presumptive infection without a specific known aetiology.

It is against this background that one must consider Gram-negative sepsis, and it is not surprising that these organisms turn up in a wide variety of clinical material. They are frequently found colonizing wounds, sinuses, ulcers, burns, and chronically discharging ears—in fact wherever the body integument is broken. In many cases they are probably of little consequence, but their presence deep in a wound may cause harm by consuming oxygen and enhancing the growth of anaerobes. This is a common finding in foot infections in diabetic patients, for example. Distinguishing between simple colonization and infections of consequence can be challenging. This is one reason why practice of the specialty of infectious diseases requires both knowledge and thoughtful analysis—reflex administration of antimicrobials can be harmful to the health of the patient.

Specific types of infection

Urinary tract infection (Chapter 20.12)

The urinary tract is the most common site for genuine infection as opposed to simple colonization. Such infections range from a simple cystitis to pyelonephritis and pyonephrosis. Most infections are caused by *E. coli*, but resistant strains of *Klebsiella*, *Enterobacter*, *Proteus*, and *Ps. aeruginosa* are more likely to be the infecting agents in patients with complications: those with indwelling catheters; those who have undergone genitourinary surgery; and those with recurrent episodes of urinary tract infection treated with many courses of antimicrobial therapy. Septicaemia may arise from such infections, particularly after surgery (Chapter 7.5); even the simple removal of a urethral catheter from an infected individual may cause bacteraemia. *E. coli* strains causing parenteral infection usually belong to one of only 12 or so serogroups. Proteus infections in male children should alert a clinician to the possibility of a congenital abnormality of, for example, a urethral valve.

Sepsis linked to the intestinal tract

In these infections coliform bacteria are usually found with other bowel flora such as *Bacteroides* spp. and microaerophilic streptococci. Peritonitis secondary to a perforated bowel, intraperitoneal abscess (for example, pelvic, subphrenic, retrocaecal), cholecystitis, cholangitis, and liver abscess are examples of such infections. Coliforms and other bowel flora may also cause remote focal infections such as endocarditis or cerebral abscess.

Respiratory tract infections (Chapter 17.5.2)

Coliform bacteria seldom cause significant respiratory tract infections, but they are often isolated from sputum samples due to a tendency to colonize the mouths and upper respiratory tract of ill patients, particularly babies, the elderly, and debilitated patients in the intensive-care unit. Such colonization is significantly encouraged by antimicrobial chemotherapy, but this is not essential. The presence of colonizing coliforms in sputum is therefore commonly of no immediate clinical consequence, although true pneumonia may result. A classical example is caused by *Klebsiella pneumoniae* subspecies *pneumoniae* (Friedlander's bacillus), which has a characteristic morphology in the Gram stain of sputum (a fat Gram-negative rod with a large capsule) and a particular radiological appearance in the lung (sagging fissure, presumably due to the weight of capsular polysaccharide present). However, this organism accounts for only a tiny proportion of all bacterial pneumonias; most of the klebsiellae isolated from sputum are of the common aerogenes type and are not endowed with the thick capsule that characterizes *K. pneumoniae*. Patients with bronchiectasis or cystic fibrosis are especially prone to chronic bronchial superinfection, notably with capsulated 'mucoid' variants of *Ps. aeruginosa*.

K. ozaenae and *K. rhinoscleromatis* are associated with the uncommon nasal diseases ozaena and rhinoscleroma (Chapter 7.11.9). In the tropics, *Chromobacterium violaceum* occasionally causes a potentially fatal, rapidly progressive, septicaemic illness, with pneumonia and multiple abscess formation.

Neonatal infections

Newborn babies are especially liable to suffer from serious Gram-negative infections. This may be partially due to the immaturity of some host defence mechanisms, for example the complement system. Gram-negative septicaemia, which usually arises from an infected umbilicus, invariably involves the meninges as an incipient, if not overt, meningitis (Chapter 24.15.1). Fortunately such events are rare, but coliforms are the most common cause of neonatal meningitis, a fact that contrasts sharply with the scarcity of coliform meningitis after the age of 1 month. *E. coli* is usually responsible, but any of the enterobacteria may be involved. *Proteus mirabilis* infections, which take the form of a meningoencephalitis, are particularly severe, and occasionally this common organism, for reasons that are not understood, has caused disastrous outbreaks in hospital nurseries.

Specific enterobacterial infections of the gut

Escherichia coli infections

We begin with *E. coli* because the genus illustrates the vast range of pathogenic mechanisms available to the enterobacteria, together with the varied pathophysiology and clinical manifestations that ensue from infection. Most *E. coli* do not cause human illness, but are merely commensals that colonize the lower intestine from the terminal ileum to the anus. It is this property that has given rise to the term 'coliforms'. However, some have acquired virulence factors that have placed them among the leading causes of diarrhoea, particularly in the developing world. The concept that *E. coli* might be capable of causing enteritis is not a recent one. In 1895 the German paediatrician, Escherich, suspected that certain strains of '*Bacterium coli*' caused infantile diarrhoea, but attempts to differentiate pathogenic from non-pathogenic strains were unsuccessful, mainly because adequate serological classification was not then available. In the past 25 years, the combined approach using epidemiology, clinical research, and molecular biology has identified at least five distinct groups of *E. coli* that cause disease when they colonize the intestine of non-immune subjects, typically young children, or less commonly adults with no prior contact with these organisms (Table 2).

Enterovirulent *E. coli* possess a number of specific virulence genes, in addition to sharing the general properties of *E. coli* encoded in the genome of K-12 strains (the minimal genome necessary to be classified as an *E. coli* species). These are often concentrated in islands of DNA known as 'pathogenicity islands', which are segments of DNA that differ significantly in base composition from the backbone *E. coli* K-12 genome, signifying that they have most likely been imported from other bacteria by DNA transfer, infecting phage, or via plasmids. It is the cluster of virulence genes present in strains of *E. coli* that give each the capacity to cause specific types of intestinal disease. These clusters of virulence genes are often associated with specific O and H serotypes, which thus may serve as surrogate and putative markers for pathogenic strains; although as discussed below, this is not always sufficient to define which *E. coli* are the pathogens (see Table 2). The five groups of

enterovirulent *E. coli* are now described.

Enteropathogenic E. coli (EPEC)

EPEC strains are most important as a cause of endemic diarrhoea in developing countries, where they primarily infect children between the ages of 6 and 18 months. In developed countries, EPEC infection has declined to low levels over the last several decades. Some EPEC can infect adults, often in the context of traveller's diarrhoea. Their discovery goes back to the early 1940s when certain *E. coli*, recognized by the production of a distinct odour when grown on agar plates, were associated with severe outbreaks of diarrhoea in neonatal nurseries. When the Kaufmann–White scheme for serotyping *E. coli* was subsequently developed, these original strains, designated EPEC, were found to be from serogroups O111 and O55. Over the years, additional serotypes have been added to the EPEC group (see Table 2), and the epidemiology has changed from that of a strong association with neonatal units to that of a watery diarrhoea in young infants and children between 6 months and 3 years of age. Hospital microbiology units were capable of serotyping the 12 most frequent O-antigen types associated with these infections using commercial kits. Hospital clinical laboratories could then report back that a potentially enteropathogenic serotype of *E. coli* was isolated.

The modern era in the study and definition of EPEC was ushered in by the first successful human experimental infections with EPEC strains in 1978. In this experiment, two of three EPEC strains (an O127 and an O142) previously isolated from neonatal diarrhoea outbreaks caused clinical illness in adult United States volunteers, whereas the third (an O128) as well as a classical 'normal flora' strain (HS) were clinically benign. This indicated that it was possible to distinguish between virulent and avirulent EPEC and explain the finding that had puzzled clinical investigators and microbiologists for quite some time, namely that EPEC serotypes were frequently isolated from clinically well individuals. The use of animal models and *in vitro* experiments with cultured human cells soon demonstrated that virulent EPEC caused a unique pathological lesion. This consisted of the close attachment of organisms to the host intestinal-cell membrane and a change in the structure of the microvillus, which effaced and mounded up to form a platform-like pedestal upon which the attached organism was found. This characteristic change was therefore designated the 'attaching and effacing' (**A/E**) lesion, and was soon shown to occur *in situ* in human patients from whom biopsy tissue was available.

Clinical features

After an incubation period of a few days, the duration being inversely related to the inoculum size, the onset of diarrhoea is abrupt or gradual, with a tendency for cases with an abrupt onset to be more ill than the others. The stools become loose and green, then orange-coloured, and eventually watery. Vomiting is common in more severely affected children and it may even be projectile. The combination of watery diarrhoea and vomiting quickly leads to dehydration. The child is at first irritable, may have convulsions, and the temperature rises to 39 to 40 °C. In the absence of prompt fluid replacement, dehydration and metabolic disturbances may become irreversible, with the result that the child becomes apathetic, hypotensive, hypoglycaemic, and dies. Yet the disease may be mild, and marked only by the passage of a few loose stools without vomiting or general illness; this is the usual pattern in healthy children in developed countries. Occasionally, especially in poorly nourished infants in developing countries, the loose stools persist for days or even weeks, but beyond this time it becomes increasingly likely that other factors are involved, and in these subjects enteroaggregative *E. coli* may be identified.

Enteroaggregative E. coli (EaggEC)

EaggEC are typically associated with infantile diarrhoea, and seem to be more commonly present in cases of persistent diarrhoea lasting more than 14 days. It is not clear whether these organisms are the cause of these persistent episodes, or whether they are simply able to colonize the intestine damaged by another cause of diarrhoea. Persistent episodes are associated with significant nutritional deficits, and in developing countries EaggEC are a major cause of diarrhoeal mortality in young infants. When investigators first began to classify isolates of *E. coli* associated with diarrhoea by determining the pattern of adherence to certain cells in tissue culture, three distinct types were identified: those attaching in discrete packets; those adhering diffusely to the whole perimeter of the cell; and those that appeared to autoaggregate, stacking upon one another like bricks in a wall. Initially, this phenotypic characteristic, and then molecular methods that were used to identify associated and putative virulence genes, helped to define the epidemiology of EaggEC. The precise pathogenesis of EaggEC infection remains to be defined. A heat-stable toxin designated **EAST** (enteroaggregative stable toxin) has been identified in many isolates and appears to be a marker of virulence.

Enterotoxigenic E. coli (ETEC)

ETEC are common causes of diarrhoeal disease at any age. Because they produce toxins related to cholera toxin, ETEC can result in a clinical illness that resembles cholera. For most of these infections, the source is usually contaminated food or water. The inoculum size is generally high, and therefore contaminated unrefrigerated food can be an excellent vehicle in which a small inoculum can multiply to sufficient numbers to cause disease.

In studying patients with clinical cholera from whom *Vibrio cholerae* could not be isolated, investigators working in Calcutta in the mid-1960s found certain *E. coli* serotypes which, to their great surprise, caused fluid secretion in animal diarrhoea models. Two major classes of enterotoxins have since been identified: heat-labile proteins called labile toxin (**LT**), of which several subtypes have been identified; and small heat-stable peptide toxins (**ST**), which possess multiple disulphide bonds that enhance their resistance to heat inactivation, of which several subtypes have also been identified. LT acts in a manner similar to cholera toxin, catalysing the ADP-ribosylation of adenylate cyclase, the host enzyme involved in the production of cyclic-AMP. This product is an intracellular signal which, in intestinal epithelial cells, leads to a reduction in sodium absorption and increase in chloride secretion. The accumulation of excess NaCl in the intestinal lumen results in the movement of water into the lumen to maintain isosmolarity, which results in diarrhoea when the volume exceeds the absorptive capacity of the gut. It was a surprise, however, when ST was found to activate the particulate guanylate cyclase of intestinal epithelial cells and to increase intracellular cyclic-GMP, which also increased the lumenal accumulation of NaCl to cause diarrhoea, because in most other systems the effects of cAMP and cGMP tend to offset one another, providing the basis for an effective feedback control system.

ETEC also colonize the small intestine by means of adherence factors termed 'colonization-factor antigens' (**CFAs**). These are plasmid-encoded proteins expressed on the surface of the bacterium, either on a pilus or as a surface-displayed antigen. Some *E. coli* express more than one CFA antigen. These adhesins are used by the organism to attach to host cells via specific binding-to-host-cell receptors. These events are essential for virulence, for without the adherence mechanism, ETEC would just pass through the small bowel instead of intensely colonizing the proximal small bowel, a feature essential to the production of enterotoxins such as LT and ST.

The main features of ETEC infection are diarrhoea and vomiting, but proportionally more older children are affected compared with EPEC infection, and the fluid losses usually result in mild to severe dehydration. There is nothing particularly distinctive about this presentation, and it is difficult to distinguish between ETEC and rotavirus diarrhoea in these young children.

Enteroinvasive E. coli (EIEC)

Enteroinvasive *E. coli* were first identified as a cause of bloody diarrhoea in an outbreak in the United States that was traced to contaminated imported French Camembert cheese. In subsequent studies, these organisms have been identified in a low percentage of diarrhoeal illnesses in children under the age of 5 years in developing countries. These infections are rarely bloody or dysenteric, although some serotypes have been associated with a shigella-like illness . These serotypes have been shown to possess genes similar to those of *Shigella*, conferring the property of invasion of epithelial cells (this is described more fully below in the section on *Shigella*). Suffice it here to say that the products of these genes induce normally non-phagocytic cells, such as intestinal epithelial cells, to ingest the organisms within a phagocytic vacuole. Intracellular multiplication of the organisms is associated with cell damage, possibly by an apoptotic mechanism, and altered host physiology leading to diarrhoea.

In some geographical locations, EIEC are identified in about 5 per cent of watery diarrhoea episodes. They may cause disease in adults as well as children. Foodborne outbreaks have also occurred in industrialized nations, sometimes due to the importation of food from other industrialized nations tainted with the pathogenic organisms. EIEC are not as well adapted as pathogens as are *Shigella* spp., which require fewer bacteria to cause illness and generally result in more severe symptoms and complications. Clinically, EIEC infection is indistinguishable from most other causes of watery diarrhoea.

Enterohaemorrhagic E. coli (EHEC)

These organisms appear to be an evolutionary development of EPEC, as they possess the genetic determinants for the A/E lesion, engage the same signal-transduction pathways, and produce the characteristic pathological changes in the gut mucosa as EPEC. They were first identified in the United States during 1982 associated with outbreaks of bloody diarrhoea traced to contaminated hamburgers from fast-food restaurants. A serotype of *E. coli*, O157:H7, previously

unknown as a cause of human illness, was isolated from these patients, who had a distinctive haemorrhagic colitis. This haemorrhagic colitis represents the most severe end of the spectrum of *E. coli* infections. It has taken the better part of the past two decades to determine that the colitis is not the only manifestation of infection with EHEC, that a number of different *E. coli* serotypes other than O157:H7 can be implicated, and that a common characteristic of the group is their ability to produce shiga toxins. Hence these organisms are now more commonly designated 'shiga-toxin-producing *E. coli*' or **STEC**. (It should be noted that in some parts of Canada, the United Kingdom, and Europe, shiga toxin from *E. coli* is known as verotoxin (**VT**), because Vero cells were used to characterize its properties. Those who use the VT terminology refer to the group as 'verotoxin-producing *E. coli*' or **VTEC**. (However, the term 'STEC' is more correct, as it is named for the gene designation for the prototype shiga toxin from *Shigella dysenteriae* type 1.) There is an important epidemiological distinction between the terms 'EHEC' and 'STEC'. The former refers to STEC associated with a distinctive clinical syndrome, haemorrhagic colitis, most commonly due to serotype O157:H7. Yet, other STEC can produce a range of diarrhoeal illnesses that do not fit this description. Thus, all EHEC are STEC, but only some STEC are EHEC, and STEC is a more comprehensive term.

Epidemiology

STEC are found in cattle, and occasionally in other farm animals, in which they appear to be part of the normal flora. Ground hamburger meat prepared in large lots at slaughterhouses, then quick-frozen for distribution and later cooking, have been implicated in outbreaks of human infection. Under these conditions, one carcass contaminated with STEC from its faeces can contaminate meat prepared from a number of carcasses. Freezing preserves the organisms, and cooking the meat rare allows bacteria in the interior of a hamburger patty to survive. Disease results because the required inoculum size is only between 50 and 100 organisms. Hamburgers prepared in supermarkets have become a major source for sporadic cases or small outbreaks associated with picnics, school meals, church barbecues, and other similar small gatherings. Salami, sausage, and raw milk have also been implicated in outbreaks. A huge outbreak affecting over 10 000 school-age children occurred in Japan during 1996 caused by contaminated prepared school lunches.

Organisms can also be disseminated from farm animals to ground water and adjacent crops; lettuce, alfalfa sprouts, apple cider, and unpasteurized apple juice have been vehicles of infection. In fact, non-beef foods have become increasingly important sources of STEC, accounting for approximately 50 per cent of all cases in the United States. Direct infection from contact with animals, notably in children on school visits to farms, is another form of transmission. Person-to-person transmission is also well documented, which reflects the small inoculum size needed to cause infection.

Clinical features

Perhaps as a consequence of the small inocula of STEC, the incubation period is often as long as 5 to 7 days. The initial watery stools become blood-tinged and then grossly bloody over the course of a day or two, and there are abdominal cramps and tenderness. This is due to a diffuse inflammatory colitis with vascular leaks, rather than ulceration as in shigellosis. Lesser degrees of colon involvement lead to milder symptoms with less blood in the stool, with mild infections remaining as a watery diarrhoea. Patients usually improve clinically in 5 to 7 days. However, at about this time, particularly in infections with O157:H7, microangiopathic haemolytic anaemia, thrombocytopenia, and oliguric renal failure—the haemolytic-uraemic syndrome (**HUS**)—develop in 5 to 10 per cent of patients. In some patients, these manifestations are mild and self-limited; in others, rapidly developing hypertension may lead to haemorrhagic strokes and death in the acute phase. In still others, management of renal failure becomes a major clinical problem, requiring peritoneal or even haemodialysis before improvement occurs. In a small, but unknown, proportion of patients, manifestations of chronic renal damage occur a decade or more after the initial episode. Because HUS is also associated with shiga-toxin-producing *Shigella dysenteriae* type 1, it is apparent that this toxin is a significant pathogenetic factor.

Laboratory diagnosis of pathogenic E. coli

Although identifying an organism as an *E. coli* is simple, diagnosis of the different *E. coli* strains causing intestinal disease is both hard and easy. It is hard because, with the exception of 0157:H7 STEC strains (see below), there are no simple screening tests for their identification. It is easy because the virulence genes that characterize the different groups can be readily identified by polymerase chain reaction (**PCR**) and other genotyping methods; the problem is that PCR is not yet suitable for routine use in the clinical laboratory. Serotyping, as noted previously, is not specific enough, even for the EPEC strains that have classically been identified by this method. There are tissue culture methods that detect some virulence properties, such as cell-adherence patterns, or the ability to polymerize actin and reorganize cytoskeletal microfilaments and the microvillus surface, but routine clinical laboratories do not perform these tests. Modernization of the laboratory is imperative to enable the full diagnosis of *E. coli* infections, but fiscal and other considerations are likely to limit the rapidity with which this can be achieved.

The identification of 0157:H7 STEC is more straightforward. Most *E. coli* O157:H7 do not ferment sorbitol and can therefore be detected on sorbitol–MacConkey (SMAC) agar. Other STEC can be detected by commercially available enzyme-linked immunosorbent assay (**ELISA**) tests for toxin production; however, most clinical laboratories do not seek these organisms.

The detection of systemically invading *E. coli*, for example as the cause of sepsis and circulatory shock, is not a problem for the laboratory as *E. coli* are not 'normal flora' except in the colon. Sampling normally sterile sites, such as the bloodstream, readily yields the organisms, unless the patients have been given antibiotics in advance. The organisms are not fastidious, they grow rapidly, and are easily identified and tested for antimicrobial sensitivity within 48 h.

Treatment and prevention of pathogenic E. coli

With the exception of STEC infection, the main danger to an infant with *E. coli* gastroenteritis is dehydration; thus, the most urgent need is to replace fluid and electrolyte losses. Infants may require parenteral fluids, particularly with ETEC infection, but oral rehydration fluids are generally sufficient.

With STEC infection, dehydration is not the prime concern as the fluid losses are typically not severe. It is the systemic complications—the microangiopathic haemolytic anaemia and renal failure (HUS)—that are the clinical challenge. The use of antimicrobial therapy, even in STEC infection, is neither generally necessary nor advocated. While some believe the early use of antimicrobials to treat STEC will prevent the late complications, there is sufficient evidence to the contrary to pause for consideration. Certain antimicrobials induce shiga-toxin production and may predispose to HUS or increase its severity, and many believe antibiotic treatment to be contraindicated. There is no definitive, controlled clinical-trial evidence to resolve this controversy. Antimotility agents do not diminish fluid losses, so much as they prolong the interval between stooling. There is some evidence that antimotility agents can exacerbate illness and increase its severity by prolonging the contact between the pathogen and the gut mucosa, particularly in young children, and therefore they are generally considered both ineffective and potentially harmful. Measures taken to prevent infection are the same as those for shigellosis, notably good personal hygiene, especially among medical staff caring for these patients; the risk among neonatal infants is particularly high.

Shigella infections

But for history, the genus *Shigella* would be another type of *E. coli*. This is because the identification of the prototype organism of the genus occurred in 1896 rather than in 1996 when it became known that *Shigella* and *Escherichia* could not be distinguished by DNA hybridization. This bacterium, ultimately named *Shigella dysenteriae* type 1, honours the Japanese microbiologist, Kiyoshi Shiga, who isolated it from patients with dysentery during a particularly severe epidemic in Japan. Shiga proved its aetiological significance by demonstrating a rise in agglutinating antibodies during convalescence. The epidemic affected at least 100 000 individuals, with a mortality rate close to 25 per cent. After World War I, this species declined in prevalence, and was only rarely isolated. However, in 1968 it re-emerged as the cause of a widespread epidemic of dysentery in Central America and Mexico. Early in the epidemic, mortality rates reached levels similar to those in Japan 60 years before, partly because the organism was initially not grown from stool and partly because the presence of amoebic cysts was interpreted to mean that the disease was amoebic dysentery. As a result, the use of emetine (a highly toxic drug) to treat victims contributed to the high mortality. Once better media for the isolation of this organism were employed, the true cause of the outbreak was identified, proper antibiotic treatment was given, and deaths were reduced. With time, strains with multiple antibiotic resistance emerged, but the epidemic had already waned.

Soon after the publication of Shiga's work, Flexner, Sonne, and Boyd described related groups of organisms. By 1938, four groups of dysentery bacilli could be differentiated according to their biochemical reactions and antigenic structure (Table 3). Shigellae, like the salmonellae, are non-lactose fermenters (albeit, *S. sonnei* does ferment lactose after 24 h in culture), but they are unusual among the enterobacteria in lacking flagella (non-motile) and, with one minor exception, they are anaerogenic—they do not produce gas from sugars. *S. sonnei* and *S. boydii* are the least pathogenic species and usually cause minor illness. *S. flexneri* is of intermediate pathogenicity.

Epidemiology

S. dysenteriae type 1 remains the most virulent of the shigellae. It is the principal species involved in major epidemics. It thrives under conditions of poverty, overcrowding, and squalor, particularly where there are no proper means of sewage disposal. However, the same can be said for the more requently encountered *S. flexneri* and *S. sonnei*. It is not clear, therefore, why *S. dysenteriae* disappeared after the First World War, to be replaced by *S. flexneri*, or why *S. flexneri* diminished in prevalence in industrialized nations after the Second World War, where it was replaced by *S. sonnei*. Nor is it clear why *S. dysenteriae* type 1 reappeared in Mexico and Central America during 1968, in the Indian subcontinent during 1975, or in Central Africa during 1985, where it is now an endemic cause of dysentery. Shigellosis differs from other common diarrhoeal diseases of the developing world in that it affects older children and adults rather than targeting young infants.

Sources and transmission

Unlike salmonellae, shigellae are only found naturally in humans and occasionally certain non-human primates, which probably acquire infection from humans. Shigellosis is the most communicable of all bacterial infections of the gut; in adult human-volunteer challenge studies, dysentery has been produced by as few as 10 to 100 bacteria. This is partly because it resists acid pH and is able to survive passage through the stomach. Not surprisingly, the principal route of transmission is person to person by the direct faecal–oral spread of bacteria, mainly via the fingers. Apart from the obvious contaminating action of finger-to-mouth contact after touching or scratching the anal area, or not washing one's hands after defaecation, infective doses of bacteria can also be transferred to food or water. These can be ideal vehicles for transmission because shigellae do not have to multiply in food to cause infection. Thus the greatest risk is from foods that are most handled during preparation, such as salads, sandwiches, and fruit.

The global market in foods has resulted in many new ways of transporting and transferring shigellae. For example, an outbreak in Europe was caused by injecting watermelons with contaminated water in North Africa in order to increase their market weight. Lettuce fertilized with human faeces ('night-soil') in Mexico has transmitted shigellosis in the United States, facilitated in some instances by the shredding of lettuce for distribution to fast-food restaurants. It is not clear how much shigellosis is transmitted by contact with fomites, such as lavatory seats, flushing handles, taps, door knobs, roller towels, and other objects in the toilet. Cool, dark, damp conditions favour the prolonged survival of shigellae deposited on hard surfaces in this way, and under such conditions organisms have been shown to survive for at least 17 days on wooden lavatory seats.

Occasionally, large outbreaks have arisen through the faecal pollution of municipal water supplies. In countries lacking flushing toilets and sewerage systems, flooding has coincided with simultaneous increases in dysentery as flood waters wash infected human faeces deposited in fields into well water or other sources of drinking water. Flies can also transmit infection from exposed human faeces to food, and in such settings fly control reduces the incidence of infection.

Pathogenesis

The cardinal pathogenic feature of shigellae is their ability to invade and multiply in epithelial cells. Invasion occurs by a process analogous to phagocytosis, mediated by a set of genes present in a large-virulence plasmid. A number of different mutations in these genes will impair or eliminate this property and result in attenuation of virulence. *S. dysenteriae* type 1 also produces a powerful exotoxin (shiga toxin), which is associated with the haemolytic-uraemic syndrome (HUS). As described above, certain *E. coli* (STEC) that are also associated with HUS produce structurally and functionally related toxins. In experimentally infected rhesus monkeys, colonization initially takes place in the jejunum and upper ileum, giving rise to secretory diarrhoea. This may be due to the action of one of two shigella enterotoxins (Shet-1 and Shet-2), distinct from shiga toxin, but not all diarrhoea-causing shigellae produce these proteins. Because a huge dose of organisms (>10^{10} bacteria) is required to cause symptomatic infection in these animals, this finding may not be representative of human infection.

The characteristic pathology produced by shigellae is an acute, locally invasive colitis. This ranges in severity from mild inflammation of the mucous membrane of the rectum and sigmoid colon, typical of *S. sonnei* infections, to severe, necrotizing lesions affecting the whole colon and sometimes the terminal ileum, such as are seen in the worst forms of *S. dysenteriae* type 1 infection. Shigellae penetrate and multiply in the submucosa and within the epithelial cells of the colon, close to the enteric vasculature, and yet bacteraemia is rare, though less rare in patients with malnutrition. In severe cases the colon may be so damaged that it is confused with a global ulcerative colitis.

Local mucosal changes consist of oedema, capillary engorgement, and neutrophil infiltration. Small haemorrhages are common and the submucous veins may be engorged or thrombosed. The mucous membrane becomes intensely red and blood-stained mucus may be present. In the most severe forms of the disease, areas of mucosa undergo coagulation necrosis, which appear as thickened, semi-rigid, greyish patches. These eventually separate to leave raw, ulcerated areas. Haemorrhage and perforation may also result from such lesions. Extensive lesions lead to considerable protein loss, which adds to the severe debility that accompanies these infections. Extensive production of inflammatory cytokines is responsible for these mucosal responses.

Clinical features

The incubation period is usually between 2 and 3 days, but exceptionally it may be as long as a week. The illness usually starts with fever, abdominal colic, and watery diarrhoea. In many *S. sonnei* infections these are the only features and there is spontaneous resolution. In the more severe forms of shigellosis, diarrhoea and fever is accompanied by headache, anorexia, myalgia, and malaise. After 1 to 3 days the diarrhoea becomes bloody, and in some cases it may progress further to dysentery, characterized by the very frequent passage of small amounts of blood-stained mucus ('red-currant jelly') and pus, with abdominal cramps and tenesmus—the classic dysenteric syndrome. In severe forms of the disease this sequence is telescoped so that bloody, mucoid stools are passed virtually from the outset. The patient becomes toxic and restless, the pulse rapid and feeble, and there is a risk of death from hyponatraemia, hypoglycaemia (Fig. 1), septic shock due to polymicrobial bacteraemia with other coliforms, or renal failure and hypertension associated with acute HUS. Recovery in such cases is invariably slow, and occasionally patients continue with chronic or relapsing infection resembling ulcerative colitis. Indeed, 50 years ago some experts believed shigellae were the cause of ulcerative colitis. Exacerbation of haemorrhoids and rectal prolapse may result from rectal oedema and straining at stool. Shigellae are usually excreted in the faeces for a few weeks after the illness. Malnourished individuals, particularly young children, may excrete the organisms for months.

Fig. 1 Bangladeshi child with pouting and upward deviation of the eyes associated with profound hypoglycaemia complicating shigellosis. (Courtesy of RE Phillips)

Children may show striking meningism, which, in the presence of fever and headache, can be misleading if it occurs before the onset of diarrhoea. Shigellosis in children may also be associated with appendicitis and occasionally with intussusception in infants. The catabolic response, protein-losing enteropathy, and anorexia that occur and persist in shigella infections in children in developing countries can lead to acute protein-energy malnutrition (kwashiorkor). This is associated with frequent intercurrent infections and high mortality rates. Shigellae are rare causes of vaginitis in children and, as this focal infection can develop without any obvious history of diarrhoea, it can easily pass unrecognized.

Reactive arthritis or full Reiter's syndrome, associated with the HLA-B27 haplotype, purulent keratoconjunctivitis, and neuritis are uncommon late complications of infection with any of the shigellae.

Laboratory diagnosis

Isolation in culture remains the standard method for detecting shigellae. Faecal samples rather than rectal swabs should be submitted. Shigellae are delicate bacteria, so it is necessary to plate samples rapidly or inoculate a buffered transport medium if there is to be any delay before a stool can be delivered to the laboratory and processed. Some bacteriological media are more inhibitory to shigellae than others. For example, Salmonella–Shigella (**SS**) agar is moderately inhibitory, whereas Hektoen enteric and xylose–lysine–deoxycholate agar are less so. Multiple agars should be used to maximize the chance of isolation.

Antimicrobial chemotherapy

Although antimicrobial therapy is seldom needed for *S. sonnei* and other mild self-limited forms of shigellosis, it is the mainstay of treatment for severe shigellosis, especially for *S. dysenteriae* type 1 and *S. flexneri* infections. Laboratory identification should always be sought, as well as antimicrobial susceptibility data. Strains showing multiple antimicrobial resistances are common, especially in developing countries. Suppression of the normal microbial flora by inappropriate antimicrobial therapy exacerbates infection.

Co-trimoxazole and ampicillin have for many years been the drugs of choice for shigellosis, but most strains are now resistant; many are also resistant to tetracycline and chloramphenicol. Nalidixic acid is effective and cheap, but high resistance rates have arisen where the drug has been used intensively. Resistance to ciprofloxacin, which is 100 times more active against shigellae than nalidixic acid, is currently uncommon. A single dose of 1 g is effective in adults infected with shigellae other than *S. dysenteriae* type 1, but 3 to 5 days of treatment is usually required for the latter species. The treatment of children with antibiotic-resistant shigella infection is difficult, as quinolones are potentially toxic for the young. A short course of a fluoroquinolone may be necessary to treat severely affected children, but parenteral ceftriaxone (50 mg/kg per day for 5 days) and pivmecillinam are alternatives that have been used successfully.

Prevention and control

The safe disposal of excreta, provision of purified water, and control of flies are fundamental to the control of shigellosis. Where these are lacking, the incidence can still be reduced by the promotion of personal and domestic hygiene, notably hand washing after defaecation and before handling food. Unpurified water can be made safe by boiling or by the addition of hypochlorite tablets; salads and fruit can be disinfected by soaking in water containing 80 parts per million of free chlorine from household bleach. Breast feeding substantially increases resistance to infection in children. Oral vaccines for use in developing countries are being developed.

Outbreaks of *S. sonnei* dysentery in schools are difficult to control, but measures should be aimed at preventing spread by the hands. Supervised washing and disinfection of hands after defaecation, and frequent disinfection of lavatory seats, taps, and door knobs are effective if rigorously applied. Only disposable hand towels should be provided in this setting. It is impracticable to detect and exclude all children excreting shigellae, and therefore children suffering from diarrhoea, regardless of whether or not *S. sonnei* is isolated from stool, should be restricted from school until recovered.

Food handlers suffering from diarrhoea or dysentery should be excluded from work until they have produced at least three consecutive negative stool samples taken not less than 24 h apart, and at least 2 days after the cessation of any antimicrobial chemotherapy.

Salmonella infections

The genus *Salmonella* is a large complex group of organisms that continues to challenge the ability of taxonomists to classify them. After many years, and many schemes, the genus *Salmonella* is now considered to comprise a single species designated *S. enterica*. All of the more than 2400 individually distinguishable strains (based on their possession of sets of microbiological, biochemical, and serological properties), such as *S. choleraesuis, S. typhi, S. paratyphi* A or B, *S. typhimurium*, *S. dublin,* and *S. enteritidis*, are now considered to be serovars of *S. enterica*. This leads to a complex nomenclature in which the particular strain previously known as *S. enteritidis* is now called *S. enterica* serovar enteritidis. While this is now taxonomically correct, it is clinically awkward and clinicians will no doubt continue to use the old nomenclature. *S. enterica* is divided into seven subspecies, which largely correspond to the old 'subgenera'. Members of subspecies I are predominantly parasites of warm-blooded animals and include almost all the salmonellae pathogenic for humans. The other subspecies include organisms found mainly in cold-blooded animals or the environment. Subspecies IIIa and IIIb (the 'Arizona' group) are a group of organisms that, in contrast to subspecies-I organisms, ferment lactose. This makes their recognition in the laboratory more difficult because they initially resemble normal flora coliforms.

Salmonella in subspecies I fall into three epidemiological groups:

1. those highly host-adapted to humans, such as *S. typhi* and *paratyphi* A, which cause the distinctive clinical syndromes typhoid and paratyphoid fever;
2. those that are highly host-adapted to animals but which cause no human illness (e.g. *S. pullorum* in chickens);
3. a large group that are not particularly host-adapted, such as *S. typhimurium, S. dublin, S. heidelberg*.

It is in the third group that the food-poisoning salmonellae are found. Many of them are named after the city or place where they were first identified. While there are a huge number of distinguishable strains included in this group, only a few are commonly found causing human illness. Identification of uncommon serovars can be extremely useful in determining their source and mode of transmission.

Epidemiology

Salmonellae are one of the major causes of foodborne illness throughout the world, but particularly in industrialized countries. This is because many processed foods contain animal products likely to be contaminated with salmonellae, such as poultry and egg products. The latter have been particularly effective vehicles for spreading *S. enteritidis* originating in egg-laying hens (see below). It is remarkable how many processed foods contain potential sources of *Salmonella*. The socioeconomic cost of salmonella infection is enormous. In the United States it is estimated to be $1.4 billion per annum. The cost of mounting an outbreak investigation may seem high, but it is trivial in relation to the savings that can be made in medical and social costs by rapid interventions.

Animal and food sources

The food-poisoning salmonellae are enzootic in a wide range of vertebrates, unlike the typhoid and paratyphoid bacilli, which are highly host-adapted to humans. Infection may be acquired from direct contact with infected animals. On the farm, the source is often scouring calves; and in the home, family pets—even terrapins and turtles. But in general, animals are more important as a source of infection through the food chain. In developed countries, intensive animal husbandry and mass production methods encourage the spread of salmonellae. In the example of poultry, modern mechanized plucking and eviscerating methods, which are capable of processing 5000 birds an hour, can readily lead to gross cross-contamination.

Animal feeds are often the portal of entry for new *Salmonella* serotypes. The appearance and spread of *S. agona* in Britain and the United States in the early 1970s was traced to its introduction via Peruvian fish meal used in poultry and pig feed. Similarly, in the late 1970s, *S. hadar*, formerly unknown in Britain, became well established in turkey stocks after its introduction in feedstuffs from abroad. The prevalence of serotypes is constantly changing. Since 1985, the incidence of *S. enteritidis* infection has risen to unprecedented heights in Europe (mainly phage type 4) and North America (mainly phage types 8 and 13a), and probably elsewhere. The main source is poultry, for which these strains are more than usually invasive, causing oviduct infection and contamination of fresh eggs through vertical transmission, a new epidemiological dimension that had far-reaching political consequences.

The net result of the enzootic state is that raw meat and animal products, especially poultry and eggs, are commonly contaminated with salmonellae of one or another serovar. The consequences are not always as serious as they might appear, for healthy adults are able to deal with small inocula and, in general, clinical symptoms are unlikely unless multiplication of the inoculum is allowed to occur in the food before consumption. Thus correct handling, preparation, and storage of food can prevent clinical infection. Unfortunately, this ideal is not always attained. Failure to handle raw meats separately from cooked foods leads to cross-contamination, and incomplete thawing of large frozen carcasses, such as turkeys, results in inadequate cooking and the multiplication of surviving bacteria. Raw milk is often contaminated at source and is a regular cause of infection in those unwise enough to drink it. Faults in food processing plants can lead to widespread outbreaks corresponding to the distribution of the product. Major incidents have been caused by failure of heat treatment, or contamination of a food after heat treatment. The list of foods implicated is long, but a few examples are liquid egg, dried egg, dried-milk infant food, desiccated coconut, bean sprouts, chocolate, and meat pies topped up

after cooking with jelly from contaminated dispensing machines.

Human sources

Infection is not readily transmitted from person to person because of the relatively high inoculum required, a consequence of the acid susceptibility of these organisms. Infants, old people, and patients living in closed communities such as nursing homes and institutions for retarded or mentally ill individuals, in whom it may be difficult to maintain high levels of sanitation, are at particular risk. Salmonellae can be especially troublesome in hospital maternity units.

The importance of a *Salmonella*-excreting food handler as a source of infection has been exaggerated. These individuals are more likely to be the victim of handling contaminated animal products at work than a source of infection. It is rarely necessary to suspend an otherwise healthy food handler from duty until clear of salmonellae, providing the stools are formed and good standards of hygiene are maintained. However, it is mandatory to suspend food handlers with diarrhoea, whatever the apparent cause.

Pathogenesis

The infective dose is governed by many factors. As noted above, the inoculum required to cause infection is several hundred thousand organisms, as determined in experimental infections in human volunteers, even when the organism is administered in buffered solutions. There are important exceptions. The inoculum is lowered when taken in a food meal, which buffers gastric acidity and protects the organisms in their journey through the stomach to the small bowel. In fact, as few as 50 bacteria contained in certain high-fat foods, notably chocolate, cheese, and salami, can cause illness. Anything that reduces gastric acidity, such as atrophic gastritis, treatment with H_2-receptor blocking agents, and previous gastric surgery, also lowers the infective dose. Broad-spectrum antibiotics increase susceptibility by suppressing the normal competitive microflora of the gut. The newborn are especially susceptible before the gut becomes colonized with the normal intestinal flora.

The distal small intestine is the main site of infection, but the colon can also be affected. Salmonellae provide chemical signals to intestinal epithelial cells, leading to bacterial uptake within vesicles and translocation across the cytoplasm to the lamina propria where the organisms multiply and invade the bloodstream. Whereas circulating bacteria are generally contained and quickly eliminated, bacteria in the mucosa result in an acute inflammatory response with polymorphonuclear leucocytic infiltration of the submucosa. Flattening or loss of secretory epithelium occurs adjacent to these inflamed areas. Inflammatory cells are usually present in the stools, which provides a diagnostic clue to the invasive process. The mechanism by which salmonellae cause tissue damage and fluid secretion is not well understood. Production of inflammatory products that alter electrolyte and fluid transport (for example, prostaglandins and enterotoxins), and described in some strains, are considered to be responsible.

Although bacteraemia probably occurs in most infections at the outset, positive cultures are obtained in only a few per cent of all laboratory diagnosed infections. Yet, certain strains are associated with high rates of bacteraemia, for example *S. cholerae-suis* (75 per cent), *S. dublin* (25 per cent), and *S. virchow* phage-type 19 (5.5 per cent). Some of these so infected patients suffer a typhoidal illness, or even the severe manifestations of septic shock, and focal infection may arise in almost any organ of the body. *S. cholerae-suis* has a particular predilection for the aorta, where it can cause life-threatening mycotic aneurysms.

Clinical features

There are broadly six major clinical manifestations of salmonella infection, the first of which is asymptomatic infection (Table 4). Most people ingesting salmonellae never become ill. In every food-poisoning outbreak, despite the likelihood of a high inoculum in the suspect food, there are unaffected persons who excrete the organism in their faeces. Others suffer a typical attack of acute febrile enteritis lasting 2 or 3 days, and there are usually a few who suffer a more severe, prolonged attack. The proportion of people who become ill is determined by the extent of contamination of the food and the characteristics of the infecting strain.

The incubation period is usually 24 h, but it may range from 6 to 48 h, depending on the size of the infecting dose. The onset is abrupt, with malaise, nausea, headache, abdominal pain, and diarrhoea. Some patients vomit, but seldom more than once or twice. Shivering and fever is common in those who are more than mildly affected. Occasionally there is severe diarrhoea, with fluid, green, offensive stools that may contain mucus and blood. Dehydration, with cramps, oliguria, and uraemia may occur, most likely among those at the extremes of age who are at risk of a fatal outcome. Patients whose distal colon and rectum are severely affected occasionally develop tenesmus with the passage of small, bloody dysenteric stools, and there may be tenderness over the sigmoid colon. Salmonella enterocolitis may trigger off an attack of non-specific colitis, acute appendicitis in the young, or mesenteric thrombosis in the elderly. Bacteraemia early in the course of infection can lead to focal infection in certain organ systems (see below).

Reactive arthritis is an occasional late sequel of infection. Estimates of its incidence range from 1.2 to 7.3 per cent of all infections. Patients with the HLA-B27 haplotype have a strong predisposition for this complication, in whom it sometimes becomes a chronic, destructive arthritis.

Convalescent excretion

Most patients continue to excrete small numbers of salmonellae in their faeces for a few weeks after infection—about 4 to 8 weeks for adults, and 8 to 24 weeks for infants. The number of organisms present is usually low, but excretors have been found with 10^5 to 10^7 organisms per gram of faeces. Carriage of salmonellae, other than typhoid or paratyphoid bacilli, for more than 6 months is rare.

Focal infection

Focal infections are often difficult to diagnose because they may first manifest themselves long after an episode of enteritis—or the original bowel infection may even have been silent. Focal infections have a tendency to chronicity and can mimic tuberculosis, particularly in cases of osteomyelitis of a vertebra or paravertebral abscess. Salmonella osteomyelitis and arthritis are strongly associated with sickle-cell disease, where bony infarcts can be infected during the asymptomatic bacteraemic phase of salmonella enteritis. Salmonella abscesses may develop in virtually any site: the liver, gallbladder, spleen, psoas muscle, uterus (after septic abortion), and the peritoneal cavity (for example, subphrenic, pelvic) are the most common. Patients with a deep-seated salmonella infection who remain untreated suffer high mortality.

Laboratory diagnosis

Definitive diagnosis depends on the isolation of the infecting organism, for salmonella enteritis cannot be distinguished from other forms of enteritis on clinical grounds alone. The isolation methods for salmonellae allow the detection of small numbers of organisms, even when greatly outnumbered by other bacteria. While most specific serovars can only be identified in reference laboratories possessing a full set of serotyping antisera, most clinical laboratories are able to narrow identification down to a short-list by the use of restricted sets of commercially available antisera. Some reference laboratories offer strain identification of *S. typhimurium, S. enteritidis*, and other common serovars by phage-typing techniques. Patients produce antibody to their infecting strain during convalescence, but this is seldom of diagnostic value.

Antimicrobial chemotherapy

The mainstay for the treatment of salmonella gastroenteritis is fluid replacement. This can usually be accomplished by the use of oral rehydration solutions, which are discussed more fully in Chapter 14.17. In the past, antimicrobials were not recommended for treating this infection since there was no apparent shortening of the clinical illness and there was a tendency for carriage to persist longer. None the less, more aggressive treatment should be considered for certain severely affected patients or in particular situations. Examples are patients with bloody diarrhoea and those with an underlying illness such as sickle-cell disease or AIDS, which predispose patients to more severe, focally invasive, and sometimes fatal infections.

The ever-increasing rates of antimicrobial resistance among salmonellae to chloramphenicol, co-trimoxazole, tetracyclines, and ampicillin mean that ciprofloxacin and other related 4-quinolones are now the agents of choice. At times, because the 4-quinolones rapidly sterilize the stool and stop transmission, treatment may be warranted during an institutional outbreak in order to reduce the excretion and spread of salmonellae. Ciprofloxacin has also been used with remarkable success to eradicate salmonellae from chronic carriers (including those with *S. typhi*) after other treatments have failed, probably because it is concentrated in bile and mucus. Treatment is essential for invasive and focal disease, which requires full dosage of the drug for several weeks, as the serovars causing invasive infection may be

more virulent than the run-of-the-mill serovars causing diarrhoea. Resistance to ciprofloxacin (and other quinolones) developing during treatment has been reported, thus emphasizing the need for close laboratory control. The drug is not recommended for use in children because of a concern that it may cause cartilage damage, although there is little evidence for this from human studies.

Prevention and control

The correct hygienic preparation, handling, and storage of food dramatically reduces the transmission of salmonella infection, but lapses are inevitable at home and in restaurants. The most common fault is the failure to appreciate that even the briefest contact between a raw animal product and other foods can transfer an inoculum to the latter and initiate a salmonella outbreak; strict separation of the two is a fundamental food-safety prerequisite. Ideally, animal products should be salmonella-free, but this is far from the case. Methods of animal husbandry, slaughtering, processing, marketing, and policies for the safe disposal of animal and human waste are all reflected in the incidence of human infection.

Examples of control measures in animals are the compulsory heat treatment of imported and recycled animal feeds, and the competitive exclusion of salmonellae from chicks by dosing with normal gut flora. There should be severe restriction on the use of antimicrobial agents in animal rearing, particularly those drugs that are especially valuable for treating human disease, such as the fluoroquinolones. Unfortunately, economic considerations of animal husbandry seem to count more than human health, and it has proven difficult to restrict the use of antibiotics in animal feed. Because of market globalization, restrictions in one country may not affect the movement of drug-resistant organisms elsewhere. Terminal disinfection of poultry carcasses by irradiation would eliminate all pathogens including salmonellae, but irradiation has gained a bad public image and it may be a while before this will change. However, just as there was initial opposition to the pasteurization of milk (a measure of unquestioned public health value and of no risk), irradiation may be more widely used in the future to reduce foodborne infections.

Campylobacter infections

The name Campylobacter (Greek, curved rod) was coined by French workers in 1963 for a group of small, curved or spiral, Gram-negative bacteria formerly classified as vibrios. They now form part of a unique superfamily of mainly spiral bacteria that includes *Campylobacter*, *Arcobacter*, and *Helicobacter*. Campylobacters have a single flagellum at one or both poles of the bacterial cell, giving them a characteristic, rapid darting motility (Fig. 2). The type species, *Campylobacter fetus* (originally *Vibrio fetus*), was first isolated in England in 1906 from aborted sheep fetuses. In the ensuing years it became clear that this species was a major cause of infectious abortion in cattle and sheep, but it was not until the 1970s that *C. jejuni* and *C. coli* were recognized as a common cause of enteritis in humans. The reason they escaped detection for so long was that special methods are required for their isolation from faeces. *C. fetus* occasionally infects humans, but only as an uncommon opportunist causing systemic infection, sometimes with diarrhoea, in patients with immune deficiency or a serious underlying disease. Exceptionally, *C. fetus*, as well as *C. jejuni* and *C. coli*, cause human fetal infection and septic abortion.

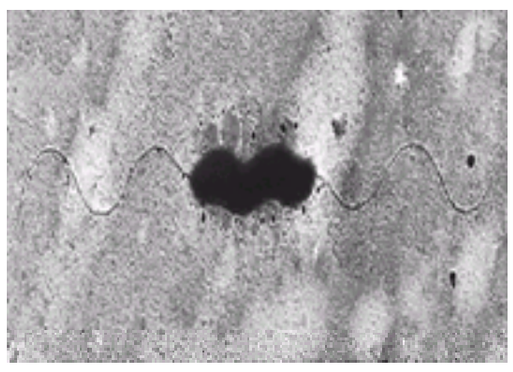

Fig. 2 Electron micrograph of a campylobacter (6650 ×). (By courtesy of Mr DR Purdham.)

Campylobacter enteritis

In industrialized countries almost all campylobacter enteritis is caused by *C. jejuni* and *C. coli*. These differ from most other *Campylobacter* species in having the high optimum growth temperature of 42 °C, in keeping with their adaptation to birds and other animals. In most regions about 90 per cent of infections are caused by *C. jejuni*. Many hundreds of strains exist, which are primarily defined serologically according to two classes of antigen: heat-stable O (lipopolysaccharide) antigens, and heat-labile surface and flagellar protein antigens. Serotypes can be further subdivided into phage types, and even finer discrimination can be attained by DNA analysis. In the United Kingdom, 473 serophage types were found among 9600 routine human isolates.

Several other campylobacters and related bacteria are associated with infection of the human intestinal tract. *C. lari* accounts for about 0.1 per cent of all cases of campylobacter enteritis. A subgroup of *C. jejuni* (*C. jejuni* subsp. *doylei*) and *C. upsaliensis*, *C. hyointestinalis*, and *Arcobacter butzleri* are scarce in industrialized countries, but more frequent in children in developing countries; although *A. butzleri* was implicated in an outbreak of abdominal pain, without diarrhoea, in an Italian school. *Helicobacter cinaedi* and *H. fennelliae* are associated with proctitis in homosexual men.

Epidemiology

Campylobacter enteritis is the most common bacterial infection of the gut in industrialized countries. Some 55 000 laboratory isolations per annum are currently reported in the United Kingdom, representing an annual incidence of 100/100 000, but the true incidence is likely to be at least ten times this figure. Incidences are similar in the United States, where the total number of cases is estimated to be 2.4 million per year, with 50 to 150 deaths. The economic burden of the disease runs to millions of dollars annually. In temperate zones there is a remarkably consistent and unexplained peak of incidence in early summer. In developed countries campylobacter enteritis affects people of all ages, especially young adults, but in developing countries it is almost entirely confined to children below the age of 2 to 3 years, after which they are immune through repeated exposure to infection.

Like salmonellosis, campylobacter enteritis is a zoonosis. Campylobacters are found in a wide variety of warm-blooded animals, especially birds, in which they form part of the normal intestinal flora. Pigs are the main host of *C. coli*. A few infections are acquired by direct contact with infected animals, either occupationally (farmers, slaughtermen, poultry processors) or domestically (typically contact with a puppy or kitten with campylobacter diarrhoea), but most are acquired indirectly via contaminated meat, milk, or water. Normal cooking destroys campylobacters, but the consumption of raw or barbecued meats, especially poultry, carries a distinct risk of infection.

Broiler chickens are the most prolific source of campylobacters. Retailed chickens are almost universally contaminated (frozen ones less so than fresh ones), so self-infection when handling them in the kitchen, or cross-contamination to other foods, readily occurs if good hygienic practice is not observed. Campylobacters do not multiply in food like salmonellae, but the infective dose is small enough for food to act as a passive vehicle, just as it does for shigellae. Foodborne infection therefore tends to be sporadic, or in small family outbreaks, rather than in the form of explosive outbreaks. Yet, major outbreaks of campylobacter enteritis affecting 3000 people at a time have been caused by the consumption of raw milk or contaminated municipally supplied water. The ubiquitous nature of campylobacters makes it difficult to pinpoint the sources of sporadic infections. There are probably many routes of infection yet to be discovered. For example, in certain areas of Britain during early summer, infections are caused by the consumption of milk contaminated by wild birds (magpies and jackdaws) pecking the foil caps of doorstep-delivered milk bottles.

Campylobacters do not survive well on inanimate objects, which is probably why person-to-person infectivity is low. Secondary cases are unusual in common-source outbreaks. Food handlers who are healthy excreters with formed stools are a negligible risk to others. The only human sources of consequence are toddlers with campylobacter diarrhoea and infected mothers at term, who may infect their babies during labour.

Pathology

Infection starts in the upper ileum and progresses distally to affect the terminal ileum and colon. The spiral configuration and motility of campylobacters enables them to penetrate, migrate, and colonize the mucus covering the intestinal epithelium in a way that conventional bacteria cannot. Histology shows an acute inflammatory response, with crypt abscess formation in the mucosa indistinguishable from that caused by salmonellae or shigellae. This, and the presence of mesenteric adenitis, suggest that campylobacters invade the mucosa. Bacteraemia is detected in only 0.1 to 0.2 per cent of infections; however, this figure probably underestimates the true incidence, as blood cultures are seldom taken from patients with diarrhoea early in the disease. Many strains produce a cholera-like enterotoxin and/or cytotoxins *in vitro*, but their role in the pathogenesis of the disease is unclear.

Specific antibodies to the infecting strain appear in patients' blood from about the fifth day of illness and remain detectable for several months.

Clinical features

After an incubation period of between 2 and 7 days (mean 3 days) the illness starts either with abdominal pain and diarrhoea, or with a prodromal period of fever, headache, and other influenza-like symptoms that precedes the diarrhoea by a few hours to a few days. A fever of 40 °C or more is not unusual and may be associated with convulsions in children and delirium in adults. Vomiting is not a conspicuous feature of the disease, except in infants, but nausea is common. Abdominal pain tends to be particularly severe and can be of a type and severity that suggests acute appendicitis (see below). Inflammatory cellular exudate can usually be detected microscopically in the stools and frank blood may appear after a day or two. Severe diarrhoea seldom lasts for more than 2 or 3 days, but loose stools and abdominal pain may persist for a while and patients feel 'washed out' and wretched. A brief relapse occurs in 10 to 15 per cent of patients. Death is rare and usually due to some associated disorder. Chronic disease or long-term carriage of campylobacters has not been recorded in normal subjects, but it has in patients with immune deficiency, such as hypogammaglobulinaemia or AIDS. The stools of most patients are culture-negative after about 5 weeks.

Misleading presentations and complications

Suspected appendicitis, particularly in older children and young adults, is the main reason for the referral of patients with campylobacter enteritis to hospital. If laparotomy is performed, the usual findings are an inflamed, oedematous ileum and enlarged, fleshy, mesenteric lymph nodes. Occasionally there is genuine appendicitis. In uncomplicated infection, abdominal tenderness may be present, but not the true signs of acute peritonitis.

Some patients present with the symptoms and sigmoidoscopic appearances of acute ulcerative colitis. The danger here is that they might be given steroids rather than antibiotics. On the other hand, campylobacter infection can exacerbate pre-existing ulcerative colitis and treatment must be given for both conditions.

Campylobacter biliary tract infection and cholecystitis are uncommon complications of infection and there are a few reports of pancreatitis and hepatitis. Other rare acute-stage complications are gastrointestinal haemorrhage, haemolytic-uraemic syndrome, glomerulonephritis, and rashes in the form of urticaria or erythema nodosum. Maternal infection and septic abortion is another rare complication that may arise without the mother having had obvious diarrhoea. Another consequence of maternal infection is neonatal infection occurring during labour. Infants may pass blood-stained stools and have symptoms that mimic intussusception. Outbreaks within neonatal units have been described, and some neonates have developed campylobacter meningitis, albeit of a relatively benign nature.

Reactive arthritis is a late complication arising 1 to 3 weeks after the onset of illness. It affects about 1 per cent of patients, or more if the frequency of the HLA-B27 haplotype in the population is high. Clinically, it is no different from the reactive arthritis following salmonella or other bacterial diarrhoeas.

Guillain-Barré syndrome

The link between campylobacter infection and Guillain-Barré syndrome (**GBS**), or postinfective polyneuropathy, was not recognized for several years. In fact, campylobacter enteritis is now the most frequently identified antecedent event in GBS (26–41 per cent of cases); moreover, patients with campylobacter-associated GBS have a worse prognosis than others. It is certainly the most distressing and dangerous of the regular complications of campylobacter enteritis. Like reactive arthritis, it arises 1 to 3 weeks after the onset of diarrhoea. Campylobacter infection is also associated with the Miller–Fisher variant of GBS, in which cranial nerves are affected, and the so-called Chinese paralytic syndrome. The demyelination of nerve sheaths that occurs in GBS is thought to be caused by an immunological crossreaction between parts of the lipopolysaccharide in the cell wall of certain *C. jejuni* strains, which initiates an autoimmune reaction. GBS and related diseases are described in Chapter 24.19.

Laboratory diagnosis

Diagnosis depends on the isolation of campylobacters from faeces, as the disease cannot be distinguished clinically from other forms of bacterial diarrhoea. The isolation of campylobacters is not difficult, but it does require special selective media and microaerobic atmospheric conditions. Campylobacters are labile bacteria, so faecal samples held for more than a few hours should be refrigerated. Faeces should be placed in transport medium if delays of more than a day are anticipated. A laboratory result is normally available in 48 h. It is conventional for laboratories to report the presence (or absence) of campylobacters without specifying whether the species is *C. jejuni* or *C. coli,* as the distinction is immaterial for clinical purposes. However, it could be important if an outbreak is suspected.

Methods for detecting campylobacters by DNA probes with PCR amplification work well, but they are currently too complex for use in routine clinical laboratories. In circumstances where a quick answer is required, such as a patient with suspected appendicitis or ulcerative colitis, the direct microscopy of faeces can be helpful—campylobacters are usually abundant in the acute stages of infection and their typical morphology and motility make them easily recognizable by a trained eye.

A retrospective serological diagnosis can be made in patients who have a late complication, such as reactive arthritis or Guillain-Barré syndrome, and in whom cultures are negative or were not performed.

Treatment

Oral rehydration and electrolyte replacement is all that is required for most patients with campylobacter enteritis. Antimicrobial therapy is of limited value because patients are usually recovering by the time a bacteriological diagnosis is made, yet it is effective if given early in the disease. Antimicrobials should be reserved for patients who are acutely ill or have complications. The choice is then between erythromycin (or another macrolide) if campylobacters are known or suspected to be the cause of illness, or ciprofloxacin given empirically if the cause of enteritis is unknown. Resistance to erythromycin rarely exceeds 5 per cent of strains in most countries (there are notable exceptions) and this figure has not changed substantially in years. Suitable dosage regimens are erythromycin stearate 500 mg twice daily for 5 days for adults, and erythromycin ethyl succinate 40 mg/kg per day for children. By contrast, resistance to ciprofloxacin and other fluoroquinolones has risen sharply in recent years; examples of current recent resistance rates are: United Kingdom, 20 per cent; United States, 24 per cent; Spain, 70 per cent. A major factor in the acquisition of quinolone resistance in *C. jejuni* is believed to be the extensive use of enrofloxacin in poultry. The traditional dosage of ciprofloxacin is 500 mg taken orally twice daily for 5 days, but shorter courses are probably effective.

For patients with life-threatening septicaemic infections, gentamicin or imipenem are the agents of choice. They are highly active against campylobacters and resistance is almost unknown. It should be noted that campylobacters are naturally resistant to most cephalosporins, and so must be considered as possible infecting agents in febrile patients who do not respond to these broad-spectrum antibiotics.

Prevention and control

As with any infection transmitted by the faecal–oral route, the safe disposal of sewage and the purification of water supplies are fundamental control measures. However, because campylobacter enteritis is a zoonosis, there is much more to its control. Prevention must be aimed at minimizing infection in food-producing animals and preventing their arrival and survival in food. Many of the measures taken to prevent salmonellosis, such as the pasteurization of milk, apply to campylobacters. Much could be done to reduce infection in broiler chickens, which are a major source of infection, but this is a complex matter that will cost money and require changes in attitude in the industry and the public. Good hygienic practice in the preparation and handling of chickens and other raw meats removes the risk of infection, but the public is largely ignorant of this and needs to be educated.

Miscellaneous food-poisoning bacteria

Although most infections of the intestinal tract are transmitted via food or water, only a limited number are typically classified under the heading of food poisoning. Of these, the minority are in fact due to microbial 'poisons' (or toxins) present in the food or water source. This convention is odd, as it groups both infections and toxin ingestions as a subset of foodborne infections. Yet the use of the term 'food poisoning' seems to be firmly fixed in the literature. For example, a Medline search for 'food poisoning' in February 2002 reveals 9086 citations, the vast majority of which refer to non-typhoidal *Salmonella* spp., enterotoxin-producing *Staphylococcus aureus*, *Clostridium* spp., *Bacillus* spp., or *Listeria monocytogenes*, as well as 1046 papers on mushroom and 588 papers on fish toxin (ciguatera) poisoning.

'Food poisoning' is most often recognized in the context of an outbreak involving a number of individuals exposed to the same food or water source. It is the number of cases that calls attention to the event and leads to the investigation that establishes a confirmed or probable aetiology. Because individual and sporadic cases are seldom investigated, particularly when symptoms are transient and insufficiently severe to lead to hospitalization or death, they are neither classified nor tabulated and remain epidemiologically invisible. However, the total number of such individuals affected in a given year probably greatly exceeds the total number of individuals involved in outbreaks.

Classification and differential diagnosis of food poisoning

Food poisoning episodes can be classified into three principal syndrome groups:

1. watery diarrhoea;
2. primarily vomiting and/or cramps; and
3. neurological symptoms.

A limited differential diagnosis (a listing of the possible causes of a clinical presentation) of food poisoning episodes can be constructed in this manner. Alternatively, a presumptive differential diagnosis can be developed by classifying the episode according to a critical epidemiological feature, for example the time elapsed between the ingestion of the presumed causative food or water and the appearance of symptoms, whatever nature they may take. In general, symptoms of food poisoning episodes occur in three time periods: within 4 h; between 8 and 16 h; 24 h or more. These periods correspond to distinct sets of common aetiologies. Therefore, simple outbreak epidemiology to determine when the suspect meal occurred in relation to the onset of symptoms can itself limit the differential diagnosis (Table 5). A third way to organize thinking about food poisoning agents is to separate them according to microbiological and taxonomic considerations, such as their Gram-stain reaction. There are certain advantages to each approach, but in clinical practice the most accurate diagnosis results from the integration of all three sources of information.

Non-cholera vibrios and vibrio-like organisms

Vibrio parahaemolyticus

This marine organism was first associated with human illness in Japan in 1963, since when it has come to be recognized as the most common cause of food poisoning in that country. Seafood is the main source of the organism, and the high incidence of infection in the Far East is doubtless due to the popularity of eating raw fish in the region. *V. parahaemolyticus* is most plentiful in warm waters, but it has been isolated from North Atlantic and Pacific coastlines, and cases of food poisoning have been reported from most continents, usually after the consumption of crabs, prawns, or raw oysters. Cooked shellfish may become contaminated after cooking, for a few bacteria picked up from a working surface contaminated with the raw product can multiply at atmospheric temperatures. This is one reason why the incidence in temperate regions is highest in summer.

After ingestion, *V. parahaemolyticus* multiplies in the gut to produce an enterotoxin, which causes watery diarrhoea and sometimes vomiting lasting for 1 or 2 days. The incubation period is usually 10 to 20 h (range 4–96 h). Despite excretion of the bacteria in enormous numbers in diarrhoeal stools, victims are not a significant source of infection. Tetracycline shortens the period of excretion (seldom more than 10 days), but antibiotics are not justified for such a short illness.

As marine vibrios fail to grow properly on routine media that are not supplemented with extra salt, it is essential for clinicians to notify the laboratory if *V. parahaemolyticus* is suspected. A positive Kanagawa test (b-haemolysis on medium containing human blood) is an indication of pathogenicity of the isolated strain.

Other non-cholera vibrios

Several other aquatic vibrios are capable of causing human gastroenteritis. A study in the southern United States showed that over half of the cases were linked to the consumption of raw oysters. Apart from *V. parahaemolyticus*, the species isolated were *V. mimicus*, *V. fluvialis*, *V. hollisae*, *V. vulnificus*, and *V. alginolyticus.* Half of the patients had fever and one-quarter had bloody stools. *V. vulnificus* is better known as a cause of severe, often fatal, septicaemic wound infection, which may arise from eating raw oysters or through damaged skin. *V. alginolyticus* also causes wound infection, but more typically otitis externa in swimmers. *Photobacterium damsela* (formerly *V. damsela*) causes a septicaemic infection like *V. vulnificus*. Again, as most of these bacteria are halophilic, it is essential to notify the laboratory of their possible presence so that high salt-containing media can be inoculated.

Aeromonas and Plesiomonas spp.

Aeromonads are ubiquitous in water, soil, and cold-blooded animals; some are major pathogens of fish. Their status as human pathogens is unclear, but it seems that some strains, mainly *Aeromonas hydrophila* (others are *A. sobria* and *A. caviae*) are capable of causing diarrhoea. Aeromonads are more frequent in hot climates, so most aeromonas infections are encountered in travellers visiting tropical and subtropical regions. Persistent aeromonas-associated diarrhoea with blood and mucus mimicking ulcerative colitis has been described in patients in Western Australia. These patients were treated successfully with trimethoprim. Most aeromonas infections are thought to be waterborne, but the absence of common-source outbreaks suggests their status as enteric pathogens is low.

Plesiomonas shigelloides is another aquatic vibrio-like organism that is occasionally associated with diarrhoea, usually of a mild nature. It has been implicated in outbreaks of diarrhoea in the Far East and has been isolated from sporadic cases throughout the world. Fewer than 100 cases a year are reported in the United Kingdom.

Gram-positive bacterial food poisoning

A limited number of Gram-positive bacteria are frequent causes of food poisoning in the strict sense of the term, namely the ingestion of preformed toxins in a meal that results in clinical illness. Gram-negative organisms do not cause 'food poisoning' in this manner even though they may produce toxins that are involved in the pathogenesis of clinical disease. The major Gram-positive causes are several species of rod-shaped bacilli (*Clostridium* spp., *Bacillus* spp.) and certain *Staphylococcus aureus* strains.

Clostridium botulinum

See Chapter 7.11.21 for further discussion.

Clostridium perfringens

See Chapter 7.11.21 for further discussion.

Bacillus cereus

This is another Gram-positive spore-forming bacillus like the clostridia, but it exhibits aerobic rather than anaerobic metabolism. It is normally and widely present in soil, hay, trees, and other plants and is frequently present in both raw and processed foods. As a spore former it is resistant to heat, chlorine, and other chemicals

used to eliminate microbial contamination, and humans are constantly exposed to this organism. While this provides an epidemiological edge for the organism, it is not a common cause of human disease, accounting for just 1 per cent of foodborne illness in Europe.

The virulence properties of *B. cereus* are not well studied. Some strains produce a chemically unique cyclic toxin, cereulide, and one or more enterotoxins acting on the intestinal mucosa by uncertain mechanisms. Cereulide is elaborated in food sources and thus causes a true food poisoning when preformed toxin is ingested in a food meal. Cereulide consists of three repeats of four different amino acids that have been modified so that every other residue and half the amino acids are a-hydroxy acid derivatives in ester, rather than amide, linkage. The cyclic structure provides heat stability, so that once elaborated by the organism the toxin is neither denatured nor detoxified by cooking. Systemically administered cereulide is a mitochondrial toxin that uncouples oxidative phosphorylation. While this may explain rare fatal cases of *B. cereus* food poisoning associated with liver failure, it does not offer any ready explanation for the self-limited emetic syndrome typical of human cases.

B. cereus also produces several distinctive enterotoxins that are suspected to cause watery diarrhoea by uncertain mechanisms. Spores induced to germinate in food by heating then develop into the replicating vegetative bacillary form in the intestinal tract where the enterotoxins are synthesized. These toxins are heterotrimers comprising three distinct polypeptides of an aggregate molecular weight greater than 100 000. A 41-kDa single peptide chain enterotoxin is present in some strains.

Food poisoning with the emetic toxin results in the rapid onset of profuse vomiting within hours of ingesting contaminated food. Most commonly this is boiled or fried rice prepared in Chinese restaurants in large amounts, stored at room temperature, and reheated before serving. Diarrhoea and abdominal pain due to enterotoxin production in the gastrointestinal tract is a longer 8- to 16-h incubation illness resembling *C. perfringens* food poisoning. It is transmitted by a number of food vehicles, including meats, stews, sauces, and dairy products. In some more recent outbreaks, the related organism *B. thuringensis*, producing similar enterotoxins, has been identified. This species is more commonly associated with its ability to produce a different toxin that kills insect larvae and is used for pest control in agriculture. However, were these strains of *B. thuringensis* to acquire the genes for enterotoxicity in nature, its use to control pest insects would be seriously compromised. The gene for this protein has been cloned into some genetically modified crops to endow them with insect-resistance properties, which, although a controversial strategy for pest control, would not carry the same risk.

Specific diagnosis depends on the isolation of high numbers of organisms from food, diarrhoea, or vomitus. Typically, however, the high temperature that induces germination and cereulide production kills the bacteria in the emetic form. Cell culture and ELISA tests for toxin are available in reference laboratories. Management of clinical cases is supportive, as symptoms are short-lived.

Listeria monocytogenes

See Chapter 7.11.34 for further discussion.

Staphylococcus aureus (see also Chapter 7.11.4)

Enterotoxin-producing strains of *Staph. aureus* are a classical cause of another true food poisoning, which results from the ingestion of preformed staphylococcal toxins produced in contaminated foods. The source of the bacteria is usually a food handler or preparer, as these organisms are ubiquitous colonizers of the skin, nasal, oral, and rectal mucous membranes. Fingers then readily transmit the bacteria from these sites to food in preparation. Such food handlers often have minor skin infections, such as boils, paronychia, impetigo, or an infected cut or skin abrasion, which facilitates the contamination of the food. The clinical illness is predominantly a short-incubation vomiting syndrome, with watery diarrhoea and abdominal cramps as less prominent symptoms. It is very common and is undoubtedly underdiagnosed.

Many foods can serve as a vehicle for the growth of *Staph. aureus*, including sliced meats, custards and cream pastries, potato and salads containing mayonnaise, and various dairy products. Organisms grow and elaborate toxins if the food is stored unrefrigerated or kept warm for serving. Staphylococcal food poisoning is not more common as relatively few isolates are toxin producers and because the ubiquitous coagulase-negative staphylococci are toxin-negative. There are eight serologically identifiable, small, structurally related, 22- to 28-kDa linear polypeptide toxins, designated staphylotoxin enterotoxins (**SE**) A to E and G to I (SEA, SEB, etc.). Another distinctive toxin, formerly identified as staphylococcal enterotoxin F, is uniquely able to translocate across mucosal surfaces and is now known as toxic-shock syndrome toxin-1 (**TSST-1**). The enterotoxin terminology is engrained in use, however inappropriate this now seems as these peptides are not prominent causes of diarrhoea. Rather they appear to target the stomach where they activate gastric emetic responses by, as yet, uncertain mechanisms. For example, it is not known what the gastric mucosal receptors for staphylococcal enterotoxins are, or even whether small quantities are absorbed to act centrally in the brain. Ultimately, however, it is the activation of medullary emetic centres in the brainstem via vagal or sympathetic-nerve transmitted signals that causes symptoms. Pathology in primates reveals inflammatory changes of the gastric mucosa, and to a lesser extent in the proximal jejunum. Brush-border alterations and mucopurulent exudates are also present. Some investigators theorize that the toxins induce mast-cell degranulation via direct binding to these cells, whereas others suggest that neuropeptides are first released from sensory nerves and secondarily stimulate mast cells to release inflammatory mediators.

Diagnosis is dependent on the isolation of high numbers of toxin-producing *Staph. aureus* from the suspect food. Stool culture is not diagnostic because *Staph. aureus* can often be found in stool in small numbers. None the less, stool culture can be useful epidemiologically if the same phage type of toxin-producing *Staph. aureus* is recovered from the food and the patient and other possible aetiologies are not found. If a food handler is suspected of being the source, culture of skin lesions for enterotoxin-producing *Staph. aureus* is also epidemiologically helpful. Treatment is supportive, as the symptoms are short-lived, generally just a matter of hours.

Summary

It is clear that Enterobacteriaceae constitute a clinically important group of organisms, primarily involved in intestinal infection, with a number of systemic syndromes due to microbial penetration of the intestinal mucosa. In addition, there are a number of foodborne illnesses due to Gram-positive organisms that overlap the spectrum of enteric disease caused by the Enterobacteriaceae. There has been considerable progress in understanding basic microbiology, epidemiology, clinical manifestations, and treatment and prevention strategies of all these organisms. Beyond the scope of this chapter, and therefore not discussed, are the general measures that are or could be taken to reduce the contamination of food and water sources of infection. These range from basic sanitation and safe food handling—for example, Hazard Analysis Critical Control Point (**HACCP**) systems to control microbial contamination of food in processing plants, or proper cooking and refrigeration of food in the home—to the use of large-scale, gamma-irradiation of food to diminish microbial counts, or other potential innovative methods to reduce the contamination of food and water sources with disease-causing micro-organisms. However, the physician must be familiar with all of these aspects of Enterobacteriaceae infections in order to properly diagnose and treat disease when prevention fails.

Further reading

Acheson DW, Kane AV, Keusch GT (2000). Shiga toxins. *Methods in Molecular Biology* **145**, 41–63.

Altekruse SF, *et al.* (2000). Vibrio gastroenteritis in the US Gulf of Mexico region: the role of raw oysters. *Epidemiology and Infection* **124**, 489–95.

Bennish ML, *et al.* (1990). Hypoglycemia during diarrhea in childhood. Prevalence, pathophysiology and outcome. *New England Journal of Medicine* **322**, 1357–63.

Blaser MJ, *et al.*, eds. (2002). *Infections of the gastrointestinal tract*. Lippincott, Williams and Wilkins, Philadelphia.

Brunder W, Karch H (2000). Genome plasticity in Enterobacteriaceae. *International Journal of Medical Microbiology* **290**, 153–65.

Crane JK (1999). Preformed bacterial toxins. *Clinics in Laboratory Medicine* **3**, 583–99.

Dinges MM, Orwin PM, Schlievert PM (2000). Exotoxins of *Staphylococcus aureus*. *Clinical Microbiology Reviews* **13**, 16–34.

Donnenberg MS (2000). Pathogenic strategies of enteric bacteria. *Nature* **406**, 768–74.

Donnenberg MS, Whittam TS (2001). Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic Escherichia coli. *Journal of Clinical Investigation* **107**, 539–48.

Dooley JSG, Roberts TA (2000). Control of vegetative micro-organisms in foods. *British Medical Bulletin* **56**, 142–57.

Farkas J (1998). Irradiation as a method for decontaminating food. *International Journal of Food Microbiology* **44**, 189–204.

Fierer J, Swancutt M (2000). Non-typhoid Salmonella: a review. *Current Clinical Topics in Infectious Diseases* **20**, 134–57.

Fleckenstein JM, Kopecko DJ (2001). Breaching the mucosal barrier by stealth: an emerging pathogenic mechanism for enteroadherent bacterial pathogens. *Journal of Clinical Investigation* **107**, 27–30.

Godaly G, *et al.* (2000). Innate defences and resistance to Gram negative mucosal infection. *Advances in Experimental Medical Biology* **485**, 9–24.

Granum PE, Lund T (1997). *Bacillus cereus* and its food poisoning toxins. *FEMS Microbiology Letters* **157**, 223–8.

Janda JM, Abbott SL (1999). Unusual food-borne pathogens. *Listeria monocytogenes, Aeromonas, Plesiomonas,* and *Edwardsiella* species. *Clinical and Laboratory Medicine* **19**, 553–82.

Keusch GT, Bennish ML (1998). Shigellosis. In: Evans AS, Brachman, PS, eds. *Bacterial infections of humans*, pp 631–56. Plenum Press, New York.

Nachamkin I, Blaser MJ, eds. (2000). Campylobacter, 2nd edn. ASM Press, Washington DC.

O'Hara CM, Brenner FW, Miller JM (2000). Classification, identification, and clinical significance of Proteus, Providencia, and Morganella. *Clinical Microbiology Reviews* **13**, 534–46.

Roberts JA (2000). Economic aspects of food-borne outbreaks and their control. *British Medical Bulletin* **56**, 133–41.

Sahly H, Podschun R, Ullmann U (2000). Klebsiella infections in the immunocompromised host. *Advances in Experimental Medicine and Biology* **479**, 237–49.

Sansonetti PJ (2001). Rupture, invasion and inflammatory destruction of the intestinal barrier by Shigella, making sense of prokaryote-eukaryote cross-talks. *FEMS Microbiology Reviews* **1**, 3–14.

Schimpff SC (1993). Gram-negative bacteremia. *Support Care Cancer* **1**, 5–18.

Skirrow MB, Blaser MJ (2002). *Campylobacter jejuni*. In: Blaser MJ, *et al.*, eds. *Infections of the gastrointestinal tract*. Lippincott, Williams and Wilkins, Philadelphia.

Tauxe R (1997). Emerging foodborne diseases: an evolving public health challenge. *Emerging Infectious Diseases* **3**, 425–34.

Thorpe CM, *et al.* (1999). Shiga toxins stimulate secretion of IL-8 from intestinal epithelial cells by altering regulation of cell processes. *Infection and Immunity* **67**, 5985–93.

Threlfall EJ, *et al.* (2000). The emergence and spread of antibiotic resistance in food-borne bacteria. *International Journal of Food Microbiology* **62**, 1–5.

Vallance BA, Finlay BB (2000). Exploitation of host cells by enteropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences, USA* **97**, 8799–806.

7.11.8Typhoid and paratyphoid fevers

J. Richens and C. Parry

Typhoid

Typhoid and paratyphoid, types A, B, and C (collectively known as enteric fevers) make up the group of salmonelloses whose main host is human. Their clinical features resemble other salmonelloses, ranging from gastroenteritis (more common with paratyphoid) to the septicaemic illness of severe typhoid.

Epidemiology

Worldwide, 15 to 30 million cases of typhoid occur each year with half a million deaths. In affluent countries, typhoid is seen in travellers or when food or water safety measures fail; with antibiotic treatment death is rare. High rates of transmission are seen in sub-Saharan Africa, the Indian subcontinent, central Asia, Vietnam, and Indonesia where annual incidence reaches 100 to 1000 cases per 100 000 population and up to 1 per cent of the population may carry *S. typhi*. In these countries, transmission has been exacerbated by antibiotic resistance. Peaks of transmission occur in dry weather or at the onset of rains. Case-fatality rates have exceeded 10 per cent in hospitalized patients in Indonesia and Papua New Guinea.

Pathogenesis

Aetiology

Salmonella enterica serovar *typhi* (*S. typhi*) is a Gram-negative bacillus capable of surviving in hostile environments and proliferating dangerously within dairy products, processed meats, and shellfish. Three antigens have been exploited for serodiagnosis; the somatic oligosaccharide O antigen (9 and 12), the protein flagellar H-d antigen, and the polysaccharide envelope Vi antigen which confers virulence by masking the O antigen from immunological attack. Antibiotic resistance is conferred by R plasmids, usually of the incompatibility group IncH-1 (chloramphenicol, amoxicillin, co-trimoxazole), and mutations in the chromosomal *gyrA* gene (fluoroquinolones). Many of the genes that give *S. typhi* its ability to survive in hostile extracellular and intracellular environments have been identified. The genome of *S. typhi* has a remarkable plasticity compared to other bacteria, which allows recombination of homologous rRNA operons as well as insertion of non-homologous DNA. The recent sequencing of an isolate of *S. typhi* will shed further light on the pathogenicity of this organism.

Transmission

Sources of typhoid transmission are excreting chronic or convalescent carriers and the acutely infected. Transmission occurs through contamination by carriers of food or water by effluents containing infected urine or faeces. 'Typhoid Mary' was a cook who infected 53 people early last century. The Aberdeen outbreak in 1964 was traced to a leaking corned beef tin which had been cooled with contaminated river water. Transmission of typhoid has also been attributed to flies, laboratory mishaps, unsterile instruments, and anal intercourse.

Infective dose

Hornick demonstrated that 10^7 organisms of Quailes strain of *S. typhi* infected 50 per cent of experimental subjects. Susceptibility is increased by antacids or vagotomy. The virulence of *S. typhi* varies. Infection may lead to acute disease, transient symptoms, or to a symptomless carrier state.

Multiplication and dissemination

Bacteria pass from the gut through the cytoplasm of enterocytes and M cells overlying lymphoid tissue (Peyer's patches) of the small intestine to reach the lamina propria from which they are conveyed to the mesenteric nodes, before reaching the blood stream via the thoracic duct. During a transient primary bacteraemia the organism is seeded to reticuloendothelial sites where intracellular multiplication occurs throughout a 7 to 14-day incubation period. A second bacteraemia follows, accompanied by symptoms as the infection spreads throughout liver, gallbladder, spleen, Peyer's patches, and bone marrow. Multiplication of *S. typhi* occurs mainly in macrophages. Concentrated sites of infection in reticuloendothelial tissues, known as typhoid nodules, are characterized by infiltrates of lymphocytes and macrophages. At post mortem, hypertrophy of lymphoid tissue is often visible within liver, spleen, mesenteric nodes, and Peyer's patches. Ulceration of Peyer's patches is seen where the inflammatory process has resulted in ischaemia and necrosis (Fig. 1).

Fig. 1 Typhoid at autopsy, showing transmural ulceration of Peyer's patches in the distal ileum.

Endotoxin plays a central role in stimulating the release of cytokines such as tumour necrosis factor and interleukins 1 and 6 from macrophages and neutrophils, by activating the complement cascade and upregulating the adhesive capacity of neutrophils and endothelial cells. These processes inflict inflammatory damage through the release of neutrophil proteases, free oxygen radicals, and arachidonic acid metabolites. Unlike in meningitis and malaria, no correlation between levels of tumour necrosis factor and clinical outcome has been demonstrated in typhoid. Levels of circulating tumour necrosis factor receptors are increased and the capacity of whole blood to produce proinflammatory cytokines following stimulation is reduced in patients with severe typhoid.

Immune response

In patients there is a cell-mediated immune response lasting about 16 weeks, a mucosal immune response lasting for up to 48 weeks, and persistent circulating anti-O and -H agglutinins for up to 2 years. The predominance of clinical typhoid among children and young adults in endemic areas suggests a degree of acquired immunity. Only 25 per cent of volunteers given a standard inoculum of *S. typhi* 20 months after an initial infection developed clinical illness. Prolonged elevation of Vi antibody occurs in typhoid carriers. Immunodeficiency reduces the ability to clear *Salmonella* infections.

Clinical features

Typhoid is predominantly an infection of children and young adults, affecting both sexes equally. The incubation period ranges from 3 to 60 days, but most infections occur 7 to 14 days after exposure.

The main focus of typhoid is in the small bowel, but systemic symptoms often overshadow abdominal symptoms. The predominant symptom is the fever which rises gradually to a high plateau of 39 to 40°C, and shows little diurnal variation. Rigors are uncommon, except in late or complicated typhoid or in patients treated with antipyretics.

Most patients complain of headache and malaise. Constipation is a frequent early symptom. Most patients will experience diarrhoea and typhoid can present as an acute gastroenteritis. Severe diarrhoea or colitis has been reported in HIV-infected patients. Bloody diarrhoea may be seen. The abdominal pain is usually diffuse and poorly localized but occasionally sufficiently intense in the right iliac fossa to suggest appendicitis. Nausea and vomiting are infrequent in uncomplicated typhoid but are seen with abdominal distension in severe cases. Other early symptoms include cough, sore throat, and epistaxes.

In developing countries, patients with typhoid in its second to fourth week present with accelerating weight loss, weakness, altered mental state, intestinal haemorrhage and perforation, refractory hypotension, pneumonia, nephritis, and acute psychosis. Those infected with multidrug resistant *S. typhi* may suffer more severe disease.

Physical examination is often unremarkable apart from fever. Careful examination may reveal splenomegaly, hepatomegaly, or rose spots. Tachycardia is common although temperature pulse-dissociation (relative bradycardia) is considered characteristic. Hypotension has important impli-cations (see below, Severe typhoid). A coated tongue is often observed. The lenticular rose spots, appear at the end of the first week. They form a sparse collection of maculopapular lesions on the abdominal skin, which blanch with pressure and fade after 2 or 3 days. Osler found them in 90 per cent of whites and 20 per cent of black skins. The rash may extend on to the trunk and arms. Melanesian typhoid patients develop purpuric macules that do not blanch (Plate 1).

Adventitious lung sounds, especially scattered wheezes, are common and may suggest pneumonia. These findings with a normal chest radiograph and high fever should prompt consideration of typhoid.

Abdominal examination may reveal the typhoid rash, distension, or a diffuse tenderness, occasionally localized to the area of the terminal ileum. Intra-abdominal inflammation sometimes provokes retention of urine. A moderate, soft, tender hepatosplenomegaly eventually develops in most patients but it less likely to be found early.

Patients with advanced illness may display the 'typhoid' facies (Fig. 2), a thin, flushed face with a staring, apathetic expression. Mental apathy may progress to an agitated delirium, frequently accompanied by tremor of the hands, tremulous speech, and gait ataxia. If the patient's condition deteriorates further the features described in the writings of Louis and Osler make their appearance—muttering delirium, twitchings of the fingers and wrists (subsultus tendinum), agitated plucking at the bedclothes (carphology), and a staring, unrousable stupor (coma vigil).

Fig. 2 Typhoid facies: 18-year-old male with severe typhoid.

Typhoid in children

Typhoid can develop in neonates born to infected mothers. The disease tends to take a milder course in children but case-fatality rates are higher in under-fives. The main differences, compared to adults, are a greater frequency of diarrhoea and vomiting, jaundice, febrile convulsions, nephritis (3 per cent in one series), or typhoid meningitis. Community-based studies in Chile and India have shown that unrecognized *S. typhi* and *paratyphi* bacteraemia can behave like a mild respiratory illness in very young children. Relative bradycardia is of greater diagnostic significance for typhoid in febrile children.

Diagnosis

A secure diagnosis of typhoid rests on the isolation of *S. typhi*. Many viral, bacterial, and protozoal infections as well as non-infectious conditions characterized by fever, such as lymphoproliferative disorders and vasculitides, resemble typhoid. Typhoid should always be considered when suspected malaria has not been confirmed or has not responded to antimalarial therapy.

Culture

S. typhi can be isolated from blood, bone marrow, stool, urine, bile, cerebrospinal fluid, and rose spots. Bone marrow gives the highest yield, including those exposed to antibiotics, but yields only marginally more than blood. For bone marrow culture the fine needle technique described by Hedley can be recommended (Hedley *et al.* (1982). *Lancet* **ii,** 415–16). Most clinicians culture blood, stool, and sometimes urine.

The median number of bacteria present in the blood of children are higher than adults and decline with increasing duration of illness. In mild typhoid, the number of bacteria may be as low as one colony forming unit per ml. Successful culture from blood can be achieved in 80 per cent of patients but depends on taking a generous volume of blood and using the correct volume of blood to broth (1:10). Automated continuously monitored culture systems (e.g. Bactec and Bact/Alert systems) can accelerate the culture from blood. Culture of bile obtained from an overnight duodenal string capsule gives a similar yield to blood and offers additional means to isolate *S. typhi* from children or from carriers. Rose spots, when present, can give a positive culture in 70 per cent of patients.

The number of organisms recoverable from faeces increases through the illness. Rectal swabs are less satisfactory than faecal samples. The results must be interpreted with caution in areas with many carriers. Isolation from urine is more common in areas endemic for schistosomiasis.

Serology

The use of a tube or slide agglutination test (the Widal test) to diagnose typhoid is cheaper and simpler than culture but fraught with pitfalls. The demonstration of a four-fold rise in titre of antibodies to *S. typhi* suggests typhoid but is too slow to help clinical decision-making and is not observed in all patients. Single measurements of antibody titres has been found useful in populations where accurate, up-to-date information about the predictive value of the test at specific cut-off points is available. False positive serological tests are obtained from persons with previous infection, infection with cross-reacting organisms, or following vaccination.

Other tests for typhoid

Many other tests for the detection of antibodies, *S. typhi* antigens, and salmonella DNA in body fluids have been described: these include passive haemagglutination, latex agglutination, counterimmune electrophoresis, radioimmunoassay, enzyme immunoassay, indirect fluorescent antibody tests, monoclonal antibodies, IgM capture, DNA probes, and PCR. Few have so far been adopted for routine use.

Other laboratory findings in typhoid

A mild normochromic anaemia, mild thrombocytopenia, and an increased erythrocyte sedimentation rate are common. The frequency of true leucopenia has been overstated in the past; most patients have a total white-cell count within the normal range. Leucocytosis suggests either perforation or another diagnosis. Laboratory evidence of mild disseminated intravascular coagulation is common but rarely of clinical significance. Common biochemical findings include hyponatraemia, hypokalaemia, and elevation of liver enzymes. The urine often contains some protein and white cells.

Management

The aims of management are to eliminate the infection swiftly with antibiotics, to restore fluid and nutritional deficits, and to monitor the patient for dangerous complications.

Antibiotics (see Table 1 for doses)

Effective antibiotic therapy in typhoid reduces mortality and complications and shortens the illness. Chloramphenicol was the first antibiotic found to be effective and the standard against which subsequent antibiotics have been measured. Ampicillin, amoxicillin, and co-trimoxazole have been shown to have comparable efficacy to chloramphenicol while having less toxicity. In many areas these drugs are no longer used because of the spread of multidrug resistant (MDR) strains of *S. typhi*. New antibiotics active against MDR *S. typhi* have emerged. Most active are the fluoroquinolones but resistance is again emerging. Other useful antibiotics are the extended-spectrum cephalosporins and azithromycin.

Most physicians start with a fluoroquinolone—ofloxacin, ciprofloxacin, fleroxacin, or pefloxacin. Treatment can be completed in a week or less with minimal toxicity In an analysis of 19 randomized trials of fluoroquinolones in the treatment of 788 patients with culture-confirmed enteric fever (>95 per cent *S. typhi* infection), the fever clearance was 2.5 to 5.2 days with a pooled cure rate of 97.3 (95 per cent CI, 96–98 per cent). Over half the studies reported no relapses and only one carrier (0.2 per cent) was detected among 591 patients followed up. Response rates in endemic areas may be better than those of non-immune travellers. For immunocompromised patients treatment may need to be extended for weeks or months. Questions remain about the safety of fluoroquinolones in children and during pregnancy. Careful follow-up studies of children in Vietnam following fluoroquinolone therapy have shown no toxicity and there is a growing consensus that the advantages of therapy outweigh the dangers. Ampicillin or amoxicillin is considered to be the safest drug to use in pregnancy with typhoid but should not be used in preference to a fluoroquinolone in patients likely to have MDR typhoid.

Strains of *S. typhi* with reduced susceptibility to fluoroquinolones are common in Asia and can be identified by being resistant to nalidixic acid. Patients infected with these strains may require longer courses of fluoroquinolones at the maximum dose or they may be treated with extended spectrum cephalosporins (ceftriaxone or cefixime). Azithromycin has recently shown to be effective in mild to moderate typhoid but currently cannot be recommended for severe disease.

Supportive care

Cooling is preferred to antipyretics for relief of fever. Simple analgesics may be used to relieve headache but note that paracetamol has been reported to lengthen the half-life of chloramphenicol five-fold. Most patients can eat and drink normally. Special diets do not protect the bowel from perforation. Daily assessment of the patient's mental and circulatory status are required plus examination of the abdomen for signs of impending perforation. Severely ill patients require intensive care with parenteral fluids, intravenous steroids (see below), inotropic support, and sedation.

Complications

Table 2 lists complications of typhoid. Most are rare and only likely to be encountered in patients who present with untreated disease lasting 2 or more weeks. Occasionally, a complication dominates the clinical picture and deflects attention from the underlying diagnosis of typhoid.

Severe typhoid

Studies from Indonesia and Papua New Guinea have revealed an important subgroup of patients with mental confusion or shock (defined as a systolic blood pressure of less than 90 mmHg in adults or less than 80 mmHg in children), with evidence of decreased skin, cerebral, or renal perfusion, who have a 50 per cent fatality and account for most typhoid deaths. In one study in Jakarta, high doses of dexamethasone substantially reduced the mortality of such severe cases. The criteria for severe typhoid were marked mental confusion or shock. In adults, dexamethasone, 3 mg/kg infused intravenously over half an hour, followed by eight doses of 1 mg/kg 6-hourly, resulted in a 10 per cent case-fatality compared to 55.6 per cent in controls.

Intestinal haemorrhage and perforation

Perforation of ileal ulcers occurs in less than 5 per cent of typhoid patients. The development of acute abdominal signs is often gradual, making diagnosis difficult. Severely ill patients display only restlessness, hypotension, and tachycardia. A chest radiograph may show free gas under the diaphragm. Ultrasonography is useful for demonstrating and aspirating faeculent fluid in the peritoneal cavity. To manage perforation start nasogastric suction, administer fluids to correct hypotension, and proceed to surgery promptly. Simple closure of perforations is adequate but experienced surgeons use procedures to bypass the worst-affected sections of the ileum in order to reduce postoperative morbidity. Closure of perforations should be accompanied by vigorous peritoneal toilet. Metronidazole or clindamycin should be added to the therapy of fluoroquinolone-treated patients. Metronidazole and aminoglycosides are recommended for patients receiving chloramphenicol, ampicillin, or co-trimoxazole. The survival of patients undergoing surgery for perforation is generally 70 to 75 per cent, but reaches 97 per cent in the best series. This compares with survival rates of around 30 per cent in conservatively managed patients.

Silent bleeding may be signalled by sudden collapse of a patient or a steadily falling haematocrit. Severe bleeding is sometimes seen in advanced typhoid. It is rarely fatal. Most bleeding episodes are self-limiting. A few require transfusion. In exceptional circumstances, surgery or intra-arterial vasopressin have been to halt haemorrhage.

Relapse

Relapse in typhoid is a second episode of fever, usually milder than the first, occurring a week or two after the recovery from the first episode. Isolates from relapsing patients usually have identical antibiotic susceptibility to those identified during the first episode. Relapse rates of 10 per cent have been described in untreated typhoid and chloramphenicol-treated patients. Relapse is managed with a similar or abbreviated course of the same therapy used in the initial episode.

Carriers

Many patients excrete *S. typhi* in their stools or urine for some days after starting antibiotic treatment. Convalescent carriers excrete for periods up to 3 months. Patients still excreting at 3 months are unlikely to cease and at 1 year meet the formal definition of 'chronic carrier'. Amongst carriers detected by screening, 25 per cent give no history of acute typhoid. Faecal carriage is more frequent in individuals with gallbladder disease and is most common in women over 40; in the Far East there is an association with opisthorchiasis. Urinary carriage is associated with schistosomiasis and nephrolithiasis. Acute typhoid in carriers has been reported. There is an increased risk of carcinoma of the gallbladder.

Patients discharged after treatment for typhoid with six negative stool and three negative urine specimens and negative Vi serology are considered free of infection. Most patients with positive stools at the completion of treatment excrete temporarily and can be safely followed up. Antibiotic eradication of carriage is advised in those still excreting at 3 months, or earlier in those at particular risk of communicating infection to others. The patient with a persistently elevated or rising Vi antibody titre is likely to be a carrier. Repeated checks of urine and faeces should be made and consideration given to obtaining bile cultures if these are negative. In Egypt, demonstration of H antibody in urine has been useful in identifying carriers.

Eradication of carriage requires prolonged, high-dose antibiotics (Table 1). Ampicillin, amoxicillin, and co-trimoxazole have been used with some success. More recently, good results have been reported with fluoroquinolones. Cholecystectomy and nephrectomy, once used to eliminate carriage (and not without operative mortality), are hard to justify on public health grounds alone, but can be considered if antibiotic methods fail and there are additional indications for operation. The success rates of surgery are increased by giving antibiotics as well.

Prevention

The elimination of typhoid from industrialized countries can be attributed to the provision of safe drinking water, safe disposal of sewage, legal enforcement of high standards of food hygiene, programmes to detect, monitor, and treat chronic carriers, and prompt investigation and intervention when these safeguards are breached. The tools of outbreak investigation are phage typing of isolates, DNA fingerprinting using pulse field gel electrophoresis or ribotyping, registers of known carriers and their phage types, and sewer swabs used to trace isolates back to their source.

Measures for individual protection are to kill *S. typhi* in water by heating to 57°C, iodination or chlorination, care with uncooked or reheated food, and immunization. Patients and convalescents with typhoid should be advised to wash their hands after using the toilet and before preparing food and to use separate towels.

Vaccines

The greatest need for typhoid vaccination is among children in endemic areas, especially where antibiotic resistance is increasing, and among laboratory workers handling *S. typhi*. In practice, vaccines are given mostly to travellers to endemic areas. A recent meta-analysis has suggested that whole cell vaccines (which are no longer widely available) are the most effective although side-effects are prominent. The most convenient is the Vi vaccine, as a single 25-µg intramuscular dose, giving 70 to 80 per cent protection for 3 years. An alternative is the live attenuated oral Ty21a vaccine which gives 65 to 70 per cent protection for 3 to 7 years. This vaccine can cause abdominal symptoms. Effectiveness can be reduced by mefloquine and antibiotics and it should not be given to immunosuppressed persons. Typhoid vaccines do not protect against paratyphoid infection and the protection afforded by vaccination can be overcome by large inocula of bacteria. Efficacy figures derive largely from trials conducted in partly immune populations and overestimate benefit in persons without prior exposure. The risks of typhoid among travellers are low (105–118 cases per million travellers to India) and the precise efficacy of currently recommended doses in previously unexposed adults remains unknown. A number of new vaccines are currently being evaluated, notably a Vi conjugate vaccine undergoing phase III clinical trials in Asia. For full details of typhoid vaccination readers should consult specialist texts.

Paratyphoid fever

Paratyphoid, type B has the widest distribution and resembles typhoid most closely. Paratyphoid A occurs chiefly in Asia and Africa and paratyphoid C in Asia and the Middle East. Paratyphoid A and C are more likely to present with a gastroenteritic than a typhoidal type of illness. *S.* Para*typhi* causes more asymptomatic infections than *S. typhi*. Outbreaks of paratyphoid are much more often food-borne than water-borne, probably because larger inocula are needed to establish infection. Paratyphoid has a shorter incubation period (4–5 days), shorter duration, and lower incidence of complications, including relapse and long-term carriage. Deaths are rare. The skin lesions of paratyphoid are larger, more numerous, and more extensive than those of typhoid. The management of paratyphoid is the same as that of typhoid. Paratyphoid organisms may display multidrug resistance as in *S. typhi*. Eradication of carriage with quinolones has been less successful in paratyphoid than in typhoid.

Further reading

Butler T, Knight J, Nath SK, Speelman P, Roy SK, Azad MAK (1985). Typhoid fever complicated by intestinal perforation: a persisting fatal disease requiring surgical management. *Reviews of Infectious Diseases* **7**, 244–56.

Christie AB (1987). Typhoid and paratyphoid fevers. In: Christie AB, ed. *Infectious diseases: epidemiology and clinical practice*, 4th edn, Vol. 1, pp. 100–64. Churchill Livingstone, Edinburgh. [An outstanding, detailed, and generously referenced monograph on typhoid.]

Engels EA *et al.* (1998). Typhoid fever vaccines: a meta-analysis of studies on efficacy and toxicity. *British Medical Journal* **316**, 110–16.

Forsyth JRL (1998). Typhoid and paratyphoid. In: Smith GR, Easmon CSF, eds. *Topley and Wilson's principles of bacteriology, virology and immunity*, 9th edn, Vol. 3, pp. 459–78. Arnold, London. [A useful chapter covering microbiological aspects of typhoid in depth.]

Hoffman SL *et al.* (1984). Reduction of mortality in chloramphenicol-treated severe typhoid fever by high-dose dexamethasone. *New England Journal of Medicine* **310**, 82–8.

Information concerning the *S. typhi* genome sequence can be accessed through the Sanger Centre web site, http://www.sanger.uk/Projects/Microbes

7.11.9 Intracellular klebsiella infections

J. Richens

Rhinoscleroma

Rhinoscleroma or scleroma in an infection of the upper airways characterized by inflammatory growths and caused by *Klebsiella pneumoniae*, subspecies *rhinoscleromatis.* Small endemic foci have been described in Africa (especially Egypt and Uganda), Siberia, Turkestan, the Middle East, the Indian subcontinent, China, the Philippines, Indonesia, and Papua New Guinea. There are many foci in South and Central America; it remains common in Guatemala where it has been identified in terracotta Maya heads of AD 300 to 600. The disease has retreated in Eastern and Central Europe where it was first described by Hebra and Kaposi in 1870.

Aetiology

K. rhinoscleromatis can be isolated from about 60 per cent of patients and is seen as intracellular inclusions in material taken from lesions. Patients show high titres of antibody which react with this organism and with the inclusions seen in sections.

Pathogenesis

Transmission is believed to occur from person to person in endemic areas. No incubation period has been defined. Initially patients infected with this organism may develop an atrophic rhinitis with squamous metaplasia, hyperkeratosis, and atrophy. The most characteristic phase of the disease is the nodular stage during which a granulomatous reaction to the organisms within macrophages leads to the development of bulky masses within any part of the respiratory tract from nares to tracheal bifurcation. The process can extend into and destroy neighbouring soft tissues, cartilage, bone, and skin. Histology shows a dense infiltrate of plasma cells among which are seen large foamy histiocytes (Mikulicz cells) containing Gram-negative bacteria and Russell bodies which are thought to be effete plasma cells. Patients with late-stage disease are liable to develop fibrosis and strictures.

Clinical features

Rhinoscleroma runs a slow, fluctuating course over several years, progressing through atrophic, nodular, and fibrotic stages. Systemic symptoms are not seen. The usual presentations are with nasal obstruction and bleeding and nasal deformity (splaying of the lower nose, often with a visible growth extending down to the upper lip known as Hebra nose) (Fig. 1). Some patients present with ozaena, which is an atrophic rhinitis accompanied by a foul smell and formation of crusts within the nose. Patients with tracheal involvement may present with stridor. With the help of sinus endoscopy and newer imaging techniques it is not unusual to find evidence of spread into the sinuses, orbits, cranial cavity, middle ear, and regional lymph nodes.

Fig. 1 Rhinoscleroma in a 30-year-old man from Papua New Guinea causing characteristic nasal splaying (Hebra nose) and obstruction of the left nostril. (Reproduced from Cooke R (1987). *Colour atlas of anatomical pathology*, p. 31. Churchill Livingstone, Edinburgh, with permission.)

Diagnosis

The diagnosis is usually made by demonstrating intracellular organisms in Giemsa or silver-stained sections taken from typical lesions combined with culture. Haemagglutination tests for Klebsiella capsular antigen III have high sensitivity and specificity. CT scanning and endoscopic techniques provide useful ways to define the extent of the disease.

Treatment

Rhinoscleroma is usually managed by ear, nose, and throat specialists using a combination of antibiotic therapy and surgery for obstructing lesions. Atrophic rhinitis may benefit from nasal lavage with saline. Treatment with ciprofloxacin, 250 mg twice daily for 4 weeks, appears to be substantially superior to previously used antibiotic regimens (rifampicin, streptomycin, tetracyclines, ampicillin and co-trimoxazole). The efficacy of fluoroquinolones may derive from their excellent intracellular penetration. For the same reason, azithromycin would be a logical choice for rhinoscleroma, particularly in view of its excellent results in donovanosis, which is caused by a very closely related intracellular Klebsiella infection. Debulking operations may be needed if there is obstructing nasal and tracheal disease and tracheostomy may be required as a temporary measure. Reconstructive surgery may be needed to deal with late fibrotic stenosis.

Donovanosis (granuloma inguinale)

Donovanosis is a sexually transmitted infection characterized by ulcers of the anogenital and inguinal areas. The name of the causative organism has recently been changed from *Calymmatobacterium granulomatis* to *Klebsiella granulomatis.* The disease is also known by the names granuloma inguinale and granuloma venereum, but should not be confused with lymphogranuloma venereum. The intracellular Gram-negative bacteria found within lesions (Donovan bodies) were first described by

the same Charles Donovan who found protozoal inclusions in visceral leishmaniasis (Leishman–Donovan bodies).

Aetiology

Recent research has indicated that it is possible to isolate an unusual Gram-negative bacillus in HEp-2 cells or human peripheral blood mononuclear cells from patients with characteristic lesions. This organism will not grow on conventional solid media. Previously named *Donovania* and subsequently *Calymmatobacterium*, it has now been classed as *Klebsiella granulomatis* on the basis of close DNA homology with other *Klebsiella* species. *Klebsiella granulomatis* shows morphological identity with Donovan bodies observed within clinical lesions of donovanosis and patients with characteristic lesions have high levels of antibody that react equally with Donovan bodies and with *K. granulomatis*. *K. granulomatis* is pathogenic only to man. Experimental transmission has been reported with lesion material, but to date not with a pure culture of this organism. Donovanosis shows a close macroscopic and microscopic similarity to rhinoscleroma which produces granulomatous lesions of the upper airways containing intracellular clusters of the closely related organism, *Klebsiella rhinoscleromatis*.

Epidemiology

Donovanosis is found in small endemic foci. The best known of these are in Papua New Guinea, India, southern Africa, Brazil, and among Australian aborigines. Smaller foci have been described in the Caribbean region and China. Where endemicity is greatest it is unusual for donovanosis to account for more than 20 per cent of genital ulcers. In most parts of the world donovanosis seems to be retreating, raising hopes of eventual eradication. Where it occurs, donovanosis appears particularly linked to poverty, poor hygiene, and prostitution. Dark-skinned persons appear to have greater susceptibility. Infectivity is believed to be low and sexual partners often remain free of infection despite prolonged exposure. The highest rates reported in partners have been 50 per cent. In the past, epidemics of donovanosis have occurred in New Guinea where they were linked to ritual homosexual and heterosexual promiscuity. The predilection of lesions for the anogenital region of sexually active adults and the frequent association with other sexually transmitted infections point to most transmission being sexual. Goldberg has put forward arguments for non-sexual transmission of an opportunistic rectal pathogen based on a single questionable isolation of the causative organism from faeces. Perinatal transmission has been observed in a few cases.

Pathogenesis

Transmission requires direct contact with an infected lesion and is thought not to occur through intact skin. The organism has a special tropism for dermal macrophages, in which it is able to avoid damage by lysosomal enzymes and toxic oxygen metabolites. The response to infection is characterized by vigorous granulomatous inflammation that damages the skin and subcutaneous tissues. Extension of the infection is predominantly a local process of spreading ulceration. The frequent inguinal lesions are probably seeded by lymphatic spread but, in general, involvement of lymphatics and lymph nodes in donovanosis is much less prominent than in lymphogranuloma venereum. Haematogenous dissemination and spread to the upper genital tract of women occur exceptionally and demonstrate the organism's ability to survive in deeper tissues. Lesions in women tend to be more extensive and may progress rapidly during pregnancy.

Clinical features

The best estimates of the incubation period range from 3 to 40 days. The early lesion is most common on the distal penis in men and near the introitus in women. Starting as a non-specific papule, the early lesion soon becomes an ulcer displaying a deep red colour, contact bleeding, low levels of pain and tenderness unless secondary infection is present, and a well-defined, rolled edge. Frequently lesions take the form of hypertrophic masses that pout outwards from the surrounding skin. Lesions are often accompanied by local oedema, particularly in women. Atypical lesions include: dry, warty, hypertrophic lesions with a cobblestone appearance; painful, excavated ulcers; and lesions with an ill-defined edge showing diffuse subcutaneous infiltration. Chronic lesions tend to expand gradually along skin folds and across to apposed skin surfaces forming a large, continuous area of ulceration, with a characteristic serpiginous outline (Fig. 2). Inguinal lesions are common, especially in men. They start as firm, subcutaneous swellings and often go on to ulcerate. The term 'pseudobubo' was originally coined to describe a subcutaneous inguinal abscess in donovanosis (rare) but tends to be used now to describe the more common ulcerating inguinal lesions. Primary lesions of the cervix are notorious for simulating carcinoma of the cervix. The uterus, fallopian tubes, ovaries, and adnexas may all be involved, simulating other forms of pelvic inflammatory disease with abscess formation or simulating malignancy with development of a frozen pelvis, large, hard masses, or hydronephrosis. Anal lesions in women commonly spread directly from the introitus; in men they are associated with anal intercourse. Involvement of the rectum very seldom occurs.

Fig. 2 Characteristic serpiginous ulcer in female patients with long-standing donovanosis.

Extragenital lesions of donovanosis occur most often in and around the mouth and sometimes on the neck. Haematogenous dissemination of donovanosis is associated especially with the trauma to an infected uterine cervix during pregnancy. The manifestations include lytic bone lesions, psoas, and perinephric abscesses. Spread to liver, spleen, and lung occurs exceptionally. Lesions in infants tend to involve the ears and regional lymph nodes.

Complications of donovanosis include extensive scar formation, lymphoedema of the genitalia, penile autoamputation, and the development of squamous carcinoma in active or healed lesions. Secondary infection with fusospirochaetal organisms can cause rapid, extensive, and sometimes fatal tissue destruction.

Diagnosis

Donovanosis is traditionally diagnosed by demonstrating the presence of Donovan bodies lying within histiocytes in material taken from a typical lesion (Fig. 3). The number of Donovan bodies present varies considerably so that sometimes a swab or scraping is sufficient to make a diagnosis whilst at other times a careful search must be made of biopsy material. Donovan bodies show well with Giemsa, Leishman, and Wright stains but poorly with haematoxylin and eosin. Histology typically shows a heavy plasma cell infiltrate and epithelial hyperplasia in addition to histiocytes containing Donovan bodies. The use of tissue culture, serological tests, and polymerase chain reaction to diagnose donovanosis have all been described recently, but are not yet commercially available. All patients should be offered screening for other sexually transmitted infections, especially syphilis and HIV. Donovanosis often causes diagnostic confusion when encountered outside endemic areas. Common misdiagnoses are squamous carcinoma of the cervix, vulva, or penis, secondary syphilis, and conditions that produce genital lymphoedema such as filariasis and lymphogranuloma venereum.

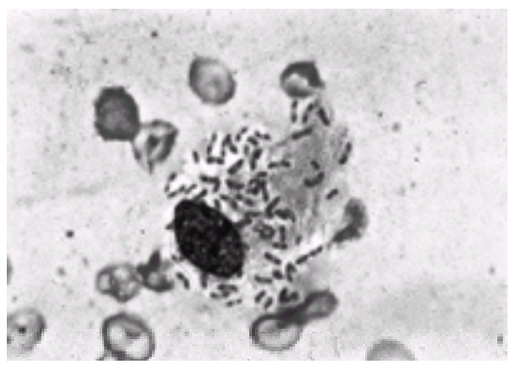

Fig. 3 Donovan bodies: Giemsa-stained smear from donovanosis lesion demonstrating the characteristic 'closed safety pin' appearance of encapsulated organisms within a large histiocyte.

Treatment and prevention

Recently published guidelines for the management of donovanosis recommend the use of azithromycin, ceftriaxone, ciprofloxacin, doxycyline, erythromycin, or co-trimoxazole. No randomized comparative trials have been conducted. Antibiotics are given at standard dosage until lesions have re-epithelialized. Expert opinion suggests that azithromycin gives the best results at a daily dose of 500 mg or weekly doses of 1 g. Antibiotic susceptibility testing is not currently feasible. Treatment failure with older antibiotics, such as doxycyline and co-trimoxazole, is well documented in individual cases. Erythromycin is safe and gives good results in pregnant women. Women in labour found to have untreated lesions of the cervix should be delivered by caesarian section to reduce known risks of haematogenous dissemination and transmission to the neonate. Patients with genital deformity may benefit from plastic surgical procedures. Partners of patients should be examined and treated if infected. A week of epidemiological treatment may be offered to healthy contacts to abort incubating infections. The main hopes for the control of donovanosis lie in strengthening services for patients with sexually transmitted infections in endemic areas, the use of newer antibiotics such as azithromycin coupled with health education and condom promotion. Eradication is currently being attempted in Australia.

Further reading

Borgstein J, Sada E, Cortes R (1993). Ciprofloxacin for rhinoscleroma and ozena. *Lancet* **342**, 122.

Bowden F, Savage J (1998). Is the eradication of donovanosis possible in Australia. *Australia and New Zealand Journal of Public Health* **22**, 7–8.

Carter JS *et al.* (1999). Phylogenetic evidence for reclassification of *Calymmatobacterium granulomatis* as *Klebsiella granulomatis* comb. nov. *International Journal of Systematic Bacteriology* **49**, 1695–1700.

Gamea AM (1990). Role of endoscopy in diagnosing scleroma in its uncommon sites. *Journal of Laryngology and Otology* **104**, 619–21.

Maher AI *et al*. (1990). Rhinoscleroma management by carbon dioxide surgical laser. *Laryngoscope* **100**, 783–8.

Meyer PR *et al*. (1983). Scleroma (rhinoscleroma). A histologic immunohistochemical study with bacteriologic correlates. *Archives of Pathology and Laboratory Medicine* **107**, 377–83.

Paul C *et al.* (1993). Infection due to *Klebsiella rhinoscleromatis* in two patients infected with human immunodeficiency virus. *Clinical Infectious Disease* **16**, 441–20.

Richens J (1992). The diagnosis and treatment of donovanosis (granuloma inguinale). *Genitourinary Medicine* **32**, 441–52.

Sehgal VN, Prasad AL(1986). Donovanosis. Current concepts. *International Journal of Dermatology* **24**, 8–16.

Ssali CLK (1975). The management of rhinoscleroma. *Journal of Laryngology and Otology* **89**, 91–9.

7.11.10Anaerobic bacteria

S. J. Eykyn

Definition of an anaerobe

The definition of an anaerobe is not entirely straightforward microbiologically. Anaerobes vary in their tolerance of oxygen and some strains will grow only in a very low concentration while others are relatively aerotolerant. In practice in the routine microbiology laboratory, bacteria that fail to grow on the surface of solid medium in 10 per cent CO_2 in air are classified as anaerobes. Confusion sometimes arises with organisms that while preferentially anaerobic (that is, they usually grow only on the anaerobic plate on primary isolation from a clinical specimen) are actually microaerophilic or capnophilic; these include *Actinomyces* which are often erroneously referred to as anaerobes. In this case the confusion is compounded as the clinical infection of actinomycosis (see Chapter 7.11.26) is usually caused by both *Actinomyces* and anaerobes. Preferentially anaerobic bacteria also include the 'milleri' group of streptococci which can readily be mistaken for anaerobic streptococci.

Incidence of anaerobic infection

Anaerobic infections are common, even if not always recognized as such, and may affect any tissue or organ and thus present to most clinicians regardless of specialty. Postoperative anaerobic sepsis was dramatically reduced when the prophylactic use of highly effective antianaerobic antibiotics was introduced in the 1970s. Although anaerobic bacteria were extensively studied in Europe in the late 19th and early 20th centuries, they were then largely ignored for many years. Anaerobes (with the exception of *Clostridium perfringens* and the occasional *Bacteroides fragilis*) were seldom isolated in clinical laboratories until the mid-1970s when an 'anaerobic renaissance' was initiated by American researchers and the anaerobes were 'rediscovered' as common and important human pathogens. This coincided with the advent of highly effective antianaerobic antimicrobials. Since then enormous advances have been made in the isolation, taxonomy, clinical diagnosis, management, and prevention of anaerobic infection.

Taxonomy

The classification and characterization of many anaerobic bacteria presents considerable difficulties and only dedicated anaerobists can hope (or need) to be abreast of current taxonomy. The many synonyms for some of these organisms bear witness to these difficulties; Finegold, for example, quoted over 50 for the organism now classified as *Fusobacterium necrophorum*. Such taxonomic confusion is further compounded by the many reports that refer to any Gram-negative anaerobic bacillus as a 'Bacteroides' and to those resistant to penicillin and ampicillin as *B. fragilis.* The use of genetic techniques has resulted in the reclassification of many anaerobes. The genus *Bacteroides* is limited to the *Bacteroides fragilis* group. The saccharolytic species previously included in the *B. melaninogenicus–oralis–ruminicola* group have been assigned to the new genus *Prevotella* which includes both pigmented and non-pigmented species. The asaccharolytic, pigmented, Gram-negative rods are now in the new genus *Porphyromonas*. Other taxonomic changes have affected the anaerobic Gram-positive cocci which have almost all become *Peptostreptococcus*. As these taxonomic changes affect some clinically important anaerobes, the new nomenclature will be used in this chapter. Table 1 lists some of the clinically important anaerobes and their old and new names where appropriate. Of the many hundreds of anaerobic species, only a small number are likely to be relevant to clinical practice and specifically reported by clinical microbiology laboratories.

Anaerobic commensal flora of man

The commensal flora of man is largely anaerobic; anaerobes are found on all the mucosal surfaces and on the skin.

Skin

It is surprising that although the skin is constantly exposed to the air, it still supports a considerable anaerobic microflora, predominantly 'anaerobic diphtheroids' that is the propionibacteria, including the lipolytic species *Propionibacterium acnes* associated with acne.

Mouth

Anaerobes are found in the tonsillar crypts, tongue crypts, gingival crevices, and dental plaque. Although some anaerobic species are found in young infants, the variety and number of anaerobes increases markedly with the eruption of the teeth. Predominant members of the oral anaerobic flora include *Prevotella*, *Fusobacterium*, *Peptostreptococcus*, *Veillonella*, and various anaerobic Gram-positive bacilli. The group *B. fragilis* group of anaerobes are rarely found in the mouth

and *Porphyromonas* only in small numbers if at all.

Intestine

The stomach and upper small intestine are normally sterile or contain transient, small numbers of anaerobic organisms derived from food, saliva, and nasopharyngeal secretions. The terminal ileum resembles the colon with a vast and diverse anaerobic flora which is established by the second year of life. Anaerobes account for about 99 per cent of the bacterial faecal mass and *Bacteroides* spp. are the commonest species. *B. vulgatus* and *B. thetaiotaomicron* are more frequently encountered than *B. fragilis*. Clostridia are also found in large numbers. Many hundred different species of anaerobe are found in the colon.

Genitourinary tract

The normal flora of the vagina is predominantly anaerobic, mostly lactobacilli, but also small numbers of *Prevotella*, fusobacteria, and peptostreptococci are found. The urethral flora consists of similar anaerobes.

Pathogenesis

The anaerobic bacteria that cause human infection are almost always derived endogenously from the host's own commensal flora. Exceptions to this include: bite and punch injuries, in which the anaerobic oral flora of assailant or victim is involved; animal and human bites (animal oral flora has large numbers of anaerobes as well as aerobic bacteria specific to animals such as *Pasteurella*); and neonatal sepsis in which the maternal vaginal anaerobes cause infection in the new-born baby. Many clostridia are found not only as normal gastrointestinal flora in man and animals but also in the soil. Clostridia are sporing anaerobes and cause infection in man in two distinct ways: firstly certain species produce potent toxins and these cause specific toxin-mediated infections that will be considered elsewhere; secondly, Clostridia, including sometimes *Clostridium perfringens,* often occur with non-sporing anaerobes in a variety of anaerobic infections in which they do not exert their toxic potential, and their presence or absence has no effect on the course of the disease. Most anaerobic infections are polymicrobial with not only several anaerobic species involved but usually aerobic species as well. The anaerobic component of these mixed infections seems to be the more important. Predisposing factors include disruption of normally intact cutaneous or mucosal barriers, tissue injury and necrosis, impaired blood supply, and obstruction. Virulence factors are also involved and include adhesins, capsules, lipopolysaccharide, hydrolytic and other enzymes, soluble metabolites, and growth factors. Precise virulence determinants for most anaerobic infections have not been established.

Adhesins

Surface attachment structures such as fimbriae have been described in some strains of *B. fragilis* and in other anaerobic species and may enable adherence to epithelial cells, an important factor in the initiation of colonization and infection.

Capsules

Capsule formation has been described in *B. fragilis* and some other anaerobes. Capsules confer resistance to phagocytosis, inhibit the migration of macrophages, and potentiate abscess formation.

Lipopolysaccharide

The lipid A component of the *B. fragilis* lipopolysaccharide differs chemically in certain respects from lipid A and this may account for its low endotoxic activity. *F. necrophorum* and *F. nucleatum* interestingly have conven-tional endotoxic lipopolysaccharide.

Enzymes

Most anaerobic pathogens produce numerous enzymes, including immunoglobulin proteases, enzymes capable of inactivating plasma proteins important in the initiation and control of the inflammatory response, and enzymes that degrade tissue components.

Diagnosis of anaerobic infection

Clinical

A working knowledge of the nature and whereabouts of the normal human commensal anaerobic flora is invaluable to the clinician as anaerobic infection frequently arises in association with this. Putrid discharge characterizes some, though not all, anaerobic infections and this results from the metabolic products of the bacteria. No other group of organisms can produce pus with such a foul, nauseating smell. Anaerobic infections, particularly necrotizing infections, are sometimes associated with cellulitis and gas formation. The former may be mistaken for streptococcal cellulitis and the latter for clostridial gas gangrene but anaerobic gangrene with gas formation generally causes far less toxaemia and prostration than clostridial infection in which the patient is alarmingly ill. Nor is the formation of gas in tissues confined to anaerobes, as aerobes are occasionally also involved. Another useful clue to the presence of anaerobes in a specimen is a report from the laboratory of 'sterile pus' despite the presence of organisms on a Gram-stained film. Lastly, in any patient who is receiving antibiotics inactive against anaerobes such as aminoglycosides, and still remains septic, an anaerobic infection should be considered.

Collection and transport of specimens for anaerobic bacteriology

All anaerobic bacteria are sensitive to oxygen but they vary in their aerotolerance. *B. fragilis* and *C. perfringens* will tolerate 2 to 4 per cent oxygen but fusobacteria and some peptostreptococci are much more sensitive to oxygen, hence more difficult to grow in the laboratory, and less likely to survive the journey from patient to culture medium. Until the renewed interest in anaerobes in the 1970s few laboratories ever isolated the more fastidious species. The best specimens for the isolation of anaerobes are aspirates, pus (in a universal container), or excised tissue and, although rapid delivery of specimens to the laboratory is desirable, in practice, anaerobes (even fastidious species) survive well in pus and tissue. Swabs are less satisfactory but are often all that is available, and for them a transport medium should be used. Complex commercial systems for the collection and transport of specimens for anaerobic bacteriology have been devised but are expensive and unlikely to appeal to clinicians. Many clinical specimens will be routinely cultured for anaerobes and no specific directive from the clinician will be required. One exception is expectorated sputum and clinicians should be aware that the microbiological diagnosis of anaerobic pleuropulmonary infection is best made from an invasive specimen.

Laboratory

The putrid smell of the pus in many anaerobic infections has been mentioned, and even swabs in such cases will be noticeably foul when processed. The Gram-stained smear of anaerobic discharge is often diagnostic to the experienced microscopist as it characteristically contains a variety of different bacteria, Gram-negative and Gram-positive rods and cocci. Filamentous or spindle-shaped Gram-negative rods (often hard to see) confirm the presence of fusobacteria. Successful culture of anaerobes requires fresh media and a reliable anaerobic atmosphere with 10 per cent carbon dioxide. Most laboratories now have either special anaerobic cabinets or automated systems using anaerobic jars. Relatively aerotolerant species will often grow in 24 to 48 h but fastidious anaerobes require undisturbed anaerobiosis for much longer (3–5 days) and if inoculated anaerobic culture plates are left out on the bench in the laboratory, such anaerobes will die. Even with the availability of commercial identification systems the definitive identification of many anaerobes is a technically demanding process and taxonomic exactitude has minimal appeal to clinicians. In clinical practice it is usually sufficient to recognize the *B. fragilis* group and Clostridia but it is clearly important that a limited number of laboratories (increasingly reference or research laboratories) retain sufficient skill to advise on the more unusual species and to define the patterns of infection associated with different sites.

Clinical spectrum of anaerobic infection

Infections of the head and neck

Acute necrotizing ulcerative gingivitis

This condition, also known as Vincent's disease, Vincent's angina, Vincent's gingivostomatitis, trench mouth, and fusospirochaetosis, affects the gingiva and buccal mucosa and was one of the earliest anaerobic infections described. The characteristic symptoms of painful bleeding gums, sometimes with a pseudomembrane, and foul breath readily suggest the diagnosis which can be confirmed with a Gram-stained smear in which large numbers of spirochaetes, fusiform, and other bacteria are seen.

Dental sepsis

The anaerobic oral commensal flora is found (with aerobic and microaerophilic bacteria) in periodontal infection, dental abscesses, and in postoperative infections associated with maxillofacial surgery.

Infections of the neck and jaw

These unusual necrotizing infections are frequently anaerobic and may be accompanied by marked cellulitis and oedema and cause respiratory embarrassment. Ludwig's angina is infection involving the main anterior compartment of the neck, the submandibular space. The source of the infection is usually the lower molar teeth, but it can arise from tonsillar infection as in the patient shown in Fig. 1. These infections spread via the fascial planes and may involve the chest with mediastinal abscess and empyema formation.

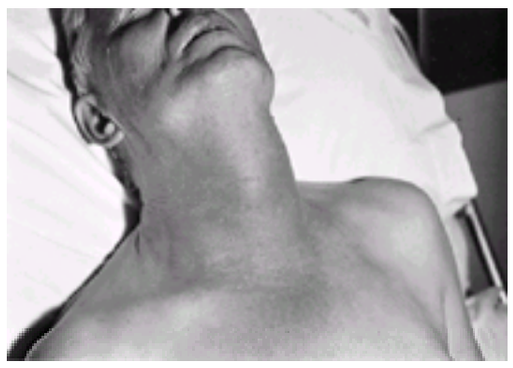

Fig. 1 Spreading cellulitis of the neck resulting from tonsillar sepsis (fatal)—'anaerobic neck'.

Ear, nose, and throat infections

Anaerobes are frequently isolated from tonsillar tissue in recurrent streptococcal tonsillitis and are also involved in peritonsillar abscesses (quinsy). They are commonly found in chronic infection of the sinuses, middle ear, and mastoid. Chronic sinus infection occasionally results in acute orbital cellulitis.

Infections of the central nervous system

Anaerobic bacteria are the major pathogens in cerebral abscesses other than those that follow surgery or trauma. Otogenic cerebral abscesses are most common and involve the temporal lobe or cerebellum. *B. fragilis* is usually isolated and aerobes, particularly *Proteus* spp. are often present. Frontal lobe abscesses of sinusitic or dental origin are usually caused by *S. milleri* group although oral anaerobes may also be found.

Pleuropulmonary infection

Anaerobic pleuropulmonary infection usually results from oropharyngeal aspiration but also occasionally from haematogenous seeding, particularly by fusobacteria (see necrobacillosis). Anaerobic pleuropulmonary infections include aspiration pneumonia, necrotizing pneumonitis, lung abscess, and empyema, as well as infection secondary to bronchiectasis and bronchial carcinoma. The anaerobes involved in these infections are the oropharygeal commensals. Patients with an anaerobic lung abscess will usually admit to the revolting taste (as well as smell) of their sputum. Definitive bacteriological diagnosis of anaerobic pleuropulmonary infection usually requires culture of an invasive specimen obtained either by bronchoscopy or percutaneous transthoracic aspiration. Expectorated sputum is rarely suitable. Specimens should preferably be obtained before antibiotics are given.

Intra-abdominal infections

These infections are usually associated with intra-abdominal pathology such as perforated gastric or duodenal ulcers, appendicitis, diverticulitis, inflammatory bowel disease, or malignancy and produce peritonitis or abscesses. Most are polymicrobial and the predominant anaerobes are those of the *B. fragilis* group. Before the advent of effective antianaerobic prophylaxis for intestinal surgery, anaerobic postoperative wound infection, abscess formation, and even septicaemia were commonly seen on surgical wards.

Hepatic and biliary tract infection

Hepatic abscesses are rare but likely to be caused by anaerobic bacteria (usually fusobacteria and *B. fragilis*) as well as by *S. milleri* group. They result from biliary tract infection, haematogenous spread from an intestinal source or direct extension of contiguous infection. Anaerobes are found in the bile in obstructive disease with stasis, and may cause cholangitis in patients who have had previous enterobiliary anastomoses.

Infections of the female genital tract and neonatal infection

Anaerobic bacteria cause bacterial vaginosis, tubo-ovarian sepsis, Bartholin's abscess, endometritis, septic abortion, and infection associated with intrauterine contraceptive devices. Vaginal hysterectomy carries a high risk of postoperative anaerobic infection, but wound infection after abdominal hysterectomy is uncommon and likely to be caused by *S. aureus.* Prolonged rupture of the membranes is associated with anaerobic infection and foul smelling liquor is often noted. Anaerobes, of vaginal origin, can be cultured from the liquor, the placenta, and the nasogastric aspirate, ear, and other surface swabs of the baby, which may develop anaerobic pneumonitis.

Infections of the male genitalia and prostate

The commensal anaerobic flora of the urethra is found in balanoposthitis, whose foul odour is well known to genitourinary physicians. Anaerobes also cause secondary infection of penile lesions. Scrotal abscesses are usually caused by anaerobes unless they follow acute epididymo-orchitis. Anaerobic scrotal abscesses which are often recurrent arise either *de novo*, and probably result from secondary infection of blocked apocrine glands or after surgery to the genitalia or urethra.

The eponymous term Fournier's gangrene was originally used at the end of the 19th century to describe necrotizing infections involving the penoscrotum and perineum that occurred in young, previously healthy men; these infections were almost certainly caused by *S. pyogenes*. Since then the term has been used for anaerobic (synergistic) necrotizing infections of the scrotum and perineum that sometimes also involve the thighs and abdominal wall. These infections are characterized by sudden intense pain and swelling with foul discharge and gas in the tissues as well as marked systemic disturbance and occur in middle aged or elderly men, particularly diabetics and alcoholics. There is a cutaneous, anorectal, or genitourinary source for the anaerobes.

Acute prostatic abscesses are rare but are sometimes caused by anaerobes. Anaerobes may also be relevant in chronic prostatitis, and can sometimes be cultured from prostatic secretions.

Infection of the urinary tract

Anaerobic urinary infection is very rare, so much so that urine is not routinely cultured anaerobically. Anaerobes can be recovered from the urine when there are abnormalities within the urinary tract such as vesicocolic fistulae, tumours, pyonephrosis, or perinephric abscess, and sometimes from ileal conduit specimens.

Bone and joint infection

Anaerobes are uncommon pathogens in acute haematogenous osteomyelitis and septic arthritis. Acute anaerobic osteomyelitis affecting long bones is likely to be caused by fusobacteria, whereas vertebral osteomyelitis, a infection occurring mainly in elderly patients, is likely to be caused by *B. fragilis*. Anaerobic septic arthritis usually occurs in patients with rheumatoid arthritis or other joint pathology and is also likely to be caused by *B. fragilis*. It can also result from bite and punch injuries to the hand in which the pathogens are oral bacteria, both anaerobic and aerobic. Anaerobes are sometimes isolated (with aerobes) in chronic osteomyelitis.

Skin and soft tissue infection

Diabetic foot ulcers

These often grow anaerobes, and the infections may be associated with underlying chronic osteomyelitis and sometimes with cellulitis, necrotizing fasciitis, and gas formation.

Venous ulcers

Anaerobes, particularly peptostreptococci, are often isolated from venous ulcers but are secondary invaders and are not relevant to the aetiology or perpetuation of the ulcer.

Decubitus ulcers

These are frequently infected with anaerobes, particularly *B. fragilis*, and anaerobic bacteraemia may occasionally result.

Sebaceous cysts

Anaerobes, especially peptostreptococci, are often isolated from infected sebaceous cysts.

Axillary abscess and hidradenitis suppurativa

Most axillary abscesses are caused by *S. aureus*, but some are anaerobic. Anaerobic abscesses are recurrent and more indolent than staphylococcal abscesses. Recurrences can result in hidradenitis suppurativa (Fig. 2). Anaerobic axillary abscesses and hidradenitis suppurativa result from apocrine blockage and infection is secondary. Hidradenitis suppurativa is not confined to the axilla but can affect the perineum, groins, buttocks, and back. Patients afflicted with this condition complain bitterly of the foul smell of their lesions.

Fig. 2 Hidradenitis suppurativa of axilla.

Perirectal abscess

These abscesses are frequently caused by anaerobes and when associated with an underlying fistula yield gut-specific anaerobes of the *B. fragilis* group and coliforms. Perirectal abscesses without a fistula are usually also caused by anaerobes but not gut-specific anaerobes; they may result from infection of blocked apocrine glands.

Breast abscess

Breast abscesses are usually assumed to be staphylococcal but in the non-puerperal woman are as likely to be anaerobic. Anaerobic breast abscesses are secondary infections of an underlying blocked duct, and are usually recurrent, subareolar, and associated with inverted nipples.

Human and animal bites

Human bites have been mentioned with reference to infection of the joints of the hand, but they may involve other parts of the body. Animal bites can also give rise to anaerobic infection but are more likely to become infected with *Pasteurella multocida* (see Chapter 7.11.17).

Paronychia

Paronychia can be caused by anaerobes, usually with aerobes. The anaerobes are oral commensals and are probably transferred to the fingers by licking or biting. Anaerobic paronychias are usually less acute than those caused by *S. aureus* or *Streptococcus pyogenes*.

Synergistic necrotizing infections

Anaerobic bacteria, usually with aerobes, cause a range of 'synergistic' infections. These infections can involve skin, fascia, and sometimes muscle and affect many areas of the body, occurring either spontaneously or after trauma or surgery (Fig. 3 and see Fournier's gangrene above).

Fig. 3 Necrotizing fasciitis involving perineum, buttock, and thigh 3 weeks after gastrectomy for carcinoma.

Bacteraemia and endocarditis

Anaerobic infection at any site, but particularly intra-abdominal infection, can cause bacteraemia sometimes with shock but anaerobes only account for less than 5 per cent of positive blood cultures, with the *B. fragilis* group most common. Anaerobes are also found in polymicrobial bacteraemia. Anaerobic endocarditis is very rare.

Fusobacterial bacteraemia, necrobacillosis, and Lemierre's postanginal septicaemia

Although most anaerobic infections are polymicrobial with not only several anaerobic species but also several aerobic species frequently isolated, fusobacteria, that is *F. necrophorum*, *F. nucleatum* and possibly other species, can be sole pathogens and produce severe infections. Their virulence is probably attributable to their lipopolysaccharide which is similar to that of Gram-negative aerobic bacteria. Although these serious infections are rare, they were well-described in the preantibiotic era. They are now being constantly 'rediscovered' by different clinicians and microbiologists, each convinced that they are describing a new disease. The species most often isolated from septicaemic disease is *F. necrophorum* and it is to this species that the term necrobacillosis refers.

Necrobacillosis

The earliest reports of necrobacillosis in man were of zoonotic skin infections acquired from animals with local infection with *F. necrophorum* usually in mixed culture, but in 1930 two fatal cases that presented 'hitherto undescribed clinical and pathological features of systemic infection' were described: a girl of 19 who died of lung abscesses, septic arthritis of the hip, and jaundice six days after a sore throat with rigors, and a man of 64 who died of a retropharyngeal abscess with gangrene and extension into the peritracheal and subcutaneous tissues. The former case is the 'postanginal septicaemia' later described by Lemierre (see below). The latter sounds like necrotizing fasciitis. Further clarification of the entity of necrobacillosis was provided in 1955 by Alston who recognized four different types of infection caused by *F. necrophorum*: those involving the skin and subcutaneous tissues, a large group where the infection started with a sore throat or otitis media, a third group associated with the female genital tract, the alimentary tract, or the urinary tract, and a fourth with empyema. Pyaemia and abscesses were very common in the last three groups. Alston's second group corresponds to Lemierre's postanginal septicaemia although Lemierre considered septicaemias arising from otitis media and mastoiditis to be a separate group. Since Alston's study, there have been numerous sporadic case reports of necrobacillosis but quite large series were published from the United Kingdom in 1989 and from Denmark in 1998. The term necrobacillosis is best used for any septicaemic infection with *F. necrophorum,* and postanginal septicaemic infection designated Lemierre's disease since this is a distinct clinical entity.

Lemierre's postanginal septicaemia (Lemierre's disease)

This unique manifestation of necrobacillosis occurs in previously healthy young people, usually adolescents or in their twenties. Lemierre suggested that it affected both sexes equally but recent series found a male predominance. There is an antecedent sore throat, often severe, and sometimes acute tonsillitis. Painful cervical lymphadenopathy is usual and septic jugular thrombophlebitis can occur. Within days, sometimes only hours, of the onset of sore throat, rigors develop with marked systemic upset and often impaired renal and hepatic function. Metastatic spread is characteristic, most commonly involving the lung, but also bone, joint, liver, brain, and heart valves. The 'pneumonia' is often severe and extensive, and cavitation of the septic infarcts and empyema may occur. Unless the relevance of the antecedent sore throat is appreciated, the diagnosis will be missed. There are occasional reports of coincidental Epstein–Barr virus infection in Lemierre's disease and viral infection might act as a trigger for fusobacterial invasion. Figure 4 shows the chest radiograph taken on admission to hospital of a 21-year-old heating engineer with rigors and severe shortness of breath about a week after a sore throat. He was thought to have Legionnaire's disease hence given erythromycin (to which fusobacteria are usually resistant); *F. necrophorum* was isolated from blood cultures. Although *F. necrophorum* is very sensitive to both penicillin and metronidazole, the infection responds only very slowly to antibiotic treatment, a reflection of the innate virulence of the organism.

Fig. 4 Chest radiograph taken on admission to hospital of a 21-year-old heating engineer who had developed rigors and severe shortness of breath about a week after a sore throat. He was thought to have possible Legionnaire's disease, hence given erythromycin (to which fusobacteria are usually resistant); *F. necrophorum* was isolated from blood cultures.

Sensitivity of anaerobic bacteria to antimicrobial agents

The susceptibility of most anaerobic bacteria to antimicrobial agents is remarkably uniform. Intrinsic resistance is often predictable and acquired resistance uncommon.

Metronidazole (and other nitroimidazoles)

Metronidazole is unique amongst the antimicrobial agents that are active against anaerobic bacteria as it is only active against anaerobes, with no activity against aerobes. Although it has been used to treat anaerobic infections for nearly 40 years, most clinically important anaerobes remain sensitive. There is little to choose between the activity of the different nitroimidazoles.

b-Lactam antibiotics

Contrary to popular belief, many anaerobes are still very sensitive to penicillin including many strains of *Prevotella*, *Porphyromonas,* and fusobacteria as well as clostridia, peptostreptococci, and spirochaetes. The *B. fragilis* group are almost uniformly resistant to penicillin, and resistance is also increasing amongst *Prevotella* and *Porphyromonas.* These penicillin-resistant anaerobes are also resistant to ampicillin, amoxycillin, piperacillin, ticarcillin, and most cephalosporins; cephamycins

such as cefoxitin and carbapenems such as imipenem and meropenem have some useful activity. The addition of the b-lactam inhibitor clavulanic acid to amoxicillin, ticarcillin, and piperacillin renders the *B. fragilis* group susceptible to these antibiotics.

Other agents

Most anaerobes are sensitive to clindamycin, and the antianaerobic activity of clindamycin is similar to that of metronidazole. Chloramphenicol is also highly active against anaerobes. Other agents with useful activity include erythromycin (though not against most fusobacteria), co-trimoxazole, and tetracyclines. The glycopeptides vancomycin and teicoplanin, whilst inactive against most Gram-negative anaerobes, possess useful activity against clostridia and peptostreptococci. Anaerobic bacteria are resistant to aminoglycosides and quinolones.

Treatment of anaerobic infections

Surgical intervention, particularly drainage of pus and excision of necrotic tissue, is of paramount importance in anaerobic infections and in many cases this will be all that is required to treat the infection. Indeed failure to carry out effective surgery will often result not only in the persistence of the infection but also in extension of this whatever antibiotic is given. Since most anaerobic infections are mixed with aerobes, it may be necessary to treat both groups of organisms. For anaerobic infections other than those of the *B. fragilis* group there is a wide choice of agent, but few clinicians think of anaerobes in distinct groups, and it is easier to recommend overall anaerobic cover which is best provided by metronidazole.

Prevention of anaerobic infection

Antibiotic prophylaxis for operations likely to be followed by postoperative anaerobic wound infection did not become routine until the mid 1970s but since then many trials bear witness to the efficacy of such prophylaxis in surgery involving sites with an anaerobic commensal flora and the putrid wound infections so familiar to gastrointestinal surgeons in the past are now rare. The patient in Fig. 5 featured in the trial of intravenous metronidazole versus placebo (saline) in elective colorectal surgery that took place at St Thomas' Hospital in 1976; he received saline! Such a trial would be quite unethical today. Most prophylactic regimens include cover for both aerobes and anaerobes. Antianaerobic prophylaxis is given for many different types of surgery, but particularly for that involving the gastrointestinal tract, genital tract, and upper respiratory tract. Such prophylaxis should be perioperative, intravenous, and of short duration (1–3 doses). There are many possible regimens but cefuroxime and metronidazole are widely used in the United Kingdom.

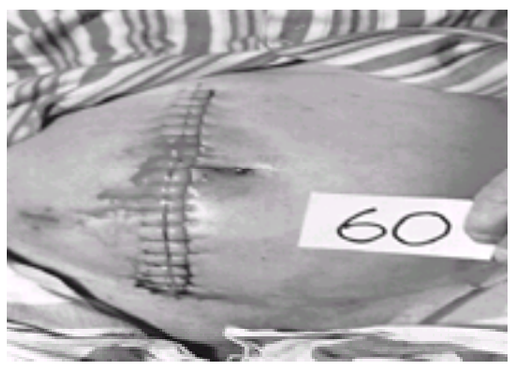

Fig. 5 Wound sepsis following elective colorectal surgery.

Further reading

Alston JM (1955). Necrobacillosis in Great Britain. *British Medical Journal* **ii**,1524–28. Old paper providing insight into various clinical presentations of fusobacterial septicaemia.

Eykyn SJ (1989). Necrobacillosis. *Scandinavian Journal of Infectious Diseases* **62** (Suppl.), 41–6.

Finegold SM, George WL, eds (1989). *Anaerobic infections in humans*. Academic Press, New York.

Hagelskjær LH, Prag J, Malczynski J, Kristensen JH (1998). Incidence and clinical epidemiology of necrobacillosis, including Lemierre's syndrome, in Denmark 1990–1995. *European Journal of Clinical Microbiology and Infectious Diseases* **17**, 561–5.

Lemierre A (1936). On certain septicaemias due to anaerobic organisms. *Lancet* **i**, 701–3. This paper contains the classic description of postanginal septicaemia.

Unattributed (1984). International symposium on anaerobic bacteria and their role in disease. *Reviews of Infectious Diseases* **6** (Suppl.1).

7.11.11

Cholera

Michael L. Bennish

Introduction

Cholera has caused millions of deaths during seven pandemics affecting all six inhabited continents over the past 200 years. Cholera can cause massive diarrhoea, dehydration, and death in healthy persons within 12 h of the onset of illness. Epidemics only occur where hygiene and social conditions are poor. This has made cholera a metaphor for death and decay in both the public ('cholera zoll der treppen'—'may the cholera strike you' is the traditional Yiddish curse) and the literary (*Love in the time of cholera* by Gabriel Garcia Marquez) imaginations. Cholera continues to cause tens of thousands of death annually in poor countries, despite ample evidence that the provision of clean water prevents disease and that inexpensive and simple-to-administer fluid therapy prevents death in those infected. That this is so is an indictment of our continuing neglect of global public health.

Aetiology

Cholera is caused by infection with one of two serogroups of *Vibrio cholerae*—O1 or O139—having been identified as the causative agent of cholera by Robert Koch in 1883. *V. cholerae* O1 and O139 are facultatively anaerobic, motile, curved Gram-negative rods that contain polar flagella and grow best in media containing increased concentrations (5 to 15 mmol/l) of sodium chloride when compared to most pathogenic micro-organisms. Hence their predilection for brackish environments—such as the Ganges Delta in the Indian subcontinent, the historic home of cholera. In addition to serogroup (determined by the somatic antigen type, of which more than 150 have been identified), *V. cholerae* serogroup O1 can be further divided into two biogroups, classical and El Tor (determined by their phenotypic characteristics) and three serotypes—Inaba, Ogawa, and Hikojima. Differences in virulence and epidemiological pattern have been described by biogroup (classical being more virulent than El Tor) but not for serotype.

Before 1992, all cholera was caused by infection with the O1 serogroup, the only serogroup then known to produce cholera toxin. In 1992, a new cholera toxin-producing serogroup—O139—was identified. This serogroup arose from the horizontal transfer of genes encoding the O139 lipopolysaccharide into a toxigenic El Tor *V. cholerae* strain. After explosive epidemics following its emergence (immunity to serogroup O1 did not protect against infection with O139) the new O139 serotype is now largely restricted to the Indian subcontinent. Despite fears that it would cause a new pandemic, it has not become endemic or epidemic elsewhere.

V. cholerae virulence genes have a number of recently defined mechanisms for horizontal transfer. The genes for cholera toxin are encoded on a filamentous phage, the receptor for which is the toxin co-regulated pilus. The latter is itself encoded by a lysogenic inovirus. In addition to being an example of evolutionary co-adaptation, the mobility of these virulence elements raises the concern of additional pathogenic strains arising.

Epidemiology

As convincingly demonstrated by the seminal epidemiologist John Snow in 1855, cholera occurs where clean drinking water is not available. *V. cholerae* resides in brackish surface water (perhaps in association with zooplankton), and initial infections during outbreaks most often stem from drinking such water. Once infection is established in a community, subsequent infections may occur by dissemination from infected individuals via contaminated food, or from drinking water newly contaminated by faecal pollution from a *V. cholerae*-infected individual. Since *V. cholerae* infections are more often asymptomatic than symptomatic, asymptomatic people may play a role in transmission. Chronic carriers are rare and do not play an important role in transmission. Because of the high inoculum required, infection rarely occurs directly from person-to-person without contaminated water or food as an intermediary vehicle. *V. cholerae* may remain viable on food (and multiply under favourable conditions) for days.

Although cholera has been present in the Indian subcontinent for centuries, the worldwide spread of cholera in modern times has been categorized as having occurred during seven pandemics. The first pandemic was recorded as starting in 1817, when Western observers became aware of the spread of cholera outside the Indian subcontinent. The current seventh pandemic, caused by an El Tor strain, started during 1961 in Sulawesi, Indonesia. Classical strains are now confined to Bangladesh.

Currently, cholera is periodically epidemic in many parts of Asia, Africa, and in Latin America, where it returned in 1991 after an absence of almost 100 years. Cholera is endemic and seasonally epidemic in the Ganges Delta, including most of Bangladesh and West Bengal, India. In 2000, 56 countries reported 137 071 cases of cholera and 4908 cholera deaths to the World Health Organization: 87 per cent of cases were in Africa. These figures are gross underestimates, both because of incomplete ascertainment and incomplete reporting. For instance, Bangladesh, where hundreds of thousands of cases of cholera occur annually, does not report cases to the World Heath Organization, presumably because of concerns about the effect on food exports.

Epidemics can be particularly severe in refugee camps, where crowding is common, and hygiene and clean water are often absent. Some 12 000 persons died in 3 weeks from a cholera outbreak in Rwandan refugee camps in Goma, Zaire.

Sporadic cases of cholera occur in the Gulf Coast region of the United States, Naples Bay in Italy, and other estuaries where *V. cholerae* lives. Infection in these areas is often linked to eating raw seafood. Because of good sanitation, epidemics no longer occur in industrialized countries following these sporadic cases. Because the infectious dose of *V. cholerae* is very high (10^9 or greater in normal hosts) infections in travellers with normal gastric acid secretion (*V. cholerae* are acid-labile—hypochlorhydria reduces the infective dose 3- or more fold) are exceedingly rare, and occur only when there are gross errors in standard hygienic practices.

Where cholera is endemic it disproportionately affects children, as many adults are immune. During epidemics in non-endemic regions, adults and children share similar risks of disease. Patients with blood group O are also thought to be at a moderately (20–100 per cent) increased risk of contracting the disease.

Pathogenesis

Depending on the inoculum and the host response, ingestion of *V. cholerae* can fail to establish infection, cause infection but not illness, or can result in disease. The incubation period between the ingestion of *V. cholerae* and the emergence of symptoms ranges from 12 to 72 h.

V. cholerae O1 or O139 cause disease by colonizing the small bowel and producing an enterotoxin—cholera toxin—that causes a massive secretion of electrolytes and water into the gut lumen. This results in the profound diarrhoea that is the hallmark of cholera. The toxonosis is the only gut derangement during *V. cholerae* infection. *V. cholerae* organisms do not invade the gut epithelium (with the exception of the antigen-processing M cells), there is no histological change in the mucosa

during infection, and no inflammatory response.

Two virulence factors have been demonstrated in all cholera causing *V. cholerae* O1 or O139: toxin co-regulated pilus and cholera toxin. Toxin co-regulated pili are essential for colonization of the intestinal epithelium brush border, an essential step for proliferation in the intestinal milieu. Cholera toxin is composed of two subunits, A and B. The monomeric A subunit is non-covalently linked to a pentameric B subunit. The B subunit binds to a glycolipid receptor present on enterocytes (and many other eukaryotic cells), ganglioside GM_1. This binding becomes rapidly irreversible at body temperature. The A subunit is the active moiety, being internalized in the cell following proteolytic cleavage into two polypeptide chains (A_1 and A_2).

The A_1 subunit alters concentrations of cyclic AMP, an important intracellular messenger. It does this by catalysing the ADP-ribosylation of a protein—G_s—that upregulates adenylate cyclase activity. The latter mediates the transformation of ATP to cyclic AMP. Cyclic AMP in turn activates a protein kinase that causes protein phosphorylation, which affects ion channels and ion movement. There is increased Cl^- secretion from intestinal crypt cells into the gut lumen (which drags water with it by changing the osmotic gradient) and decreased NaCl-coupled water absorption in the villus cells.

Virulence genes in *V. cholerae* exist in two clusters—the CTX element, containing the toxin genes, and the TCP pathogenicity island, where genes coding for the toxin co-regulated pilus reside. Expression of these genes, and a number of other putative virulence factors in these gene clusters, are controlled by the transmembrane ToxR regulatory protein. ToxR directly affects transcription of the genes coding for toxin, and indirectly controls transcription of other virulence genes by initiating a cascading system of regulatory factors. Expression of ToxR occurs in response to environmental factors present in the gut lumen. This co-ordinated expression of virulence factors is required for pathogenesis, and also perhaps for survival of *V. cholerae* in the intestinal lumen, giving the *V. cholerae* that possess virulence factors a selective advantage.

An increased secretion of fluid and electrolytes in crypt cells, and decreased absorption in villus cells, results in isotonic fluid accumulation in the small-bowel lumen. The rate of fluid loss is greatest in the jejunum, where fluid losses of 11 ml/cm of jejunum per hour may occur. Diarrhoea results when the amount of fluid produced exceeds the colon's absorptive capacity (approximately 6 litres/day). Because water and ions are lost in equal proportion, the resulting dehydration affects all compartments—intracellular, extracellular, and intravascular—equally.

Clinical and laboratory features

Signs and symptoms

Cholera is one of the most distinctive of clinical illnesses, and its severe form is immediately recognizable. Dr William O'Shaugnessy's description of a patient with cholera in Sunderland in 1831, during the second pandemic of cholera, captures the clinical features of disease as well as any description in the ensuing 170 years:

> On the floor...lay a girl of slender make and juvenile height, but with the face of a superannuated hag. She uttered no moan, gave no expression of pain, but she languidly flung herself from side to side...her eyes were sunk deep into her sockets, as though they had been driven an inch behind their natural position; her mouth was squared; her features flattened; her eyelids black; her fingers shrunk, bent and inky in their hue. All pulse was gone at the wrist, and a tenacious sweat moistened her bosom.

Vomiting and watery diarrhoea are the initial signs of cholera. Diarrhoea may be modest at first—and consist of faecal matter and watery stool. In the majority of infected persons the illness will not advance beyond this stage, and the disease will not be distinguishable from other more common causes of diarrhoea, such as that caused by enterotoxigenic *Escherichia coli*.

In some patients, the diarrhoea becomes profound—exceeding 200 ml/kg body weight per day. In these patients the stool will become 'rice-watery' in character—in other words, it resembles the opaque white water discarded after rice has been washed—it will not contain fecal matter, and is not malodorous. Diarrhoea is painless and patients are often incontinent of stool. In the absence of antimicrobial treatment, the total stool volume during the illness can exceed total body weight.

Vomiting—following by retching as the stomach contents are emptied—almost always occurs in patients with severe diarrhoea. The vomiting tends to abate after the first 24 h of illness. The cause of the vomiting has not been well established—being attributed to both direct effects of *V. cholerae* on intestinal motility, and to acid–base disturbances.

In the absence of effective fluid replacement, dehydration and prostration occurs in those patients with high rates of fluid loss. The features of dehydration in cholera are unmistakable. They include markedly diminished skin turgor, reflecting diminished interstitial fluid. Alterations in skin turgor are demonstrated most graphically by pulling the abdominal subcutaneous tissues between the thumb and forefinger. In the dehydrated patient the tissues will tent (Fig. 1). Other signs of dehydration in patients with cholera include wrinkled fingers ('washerwoman's hands'), sunken eyes, dry mucous membranes, tachypnoea, altered consciousness (apprehensiveness, lethargy, stupor), diminished urine volume, and tachycardia with a diminished or absent radial pulse. Blood pressure is often not recordable.

Fig. 1 Young girl with severe dehydration as fluid infusion is begun. In this picture the sunken eyes and lassitude of severe dehydration can be appreciated, as can the abdominal skin tenting following the assertive pinching of the abdominal subcutaneous tissues.

For purposes of management, dehydration is divided into three categories—none or mild, moderate or some, and severe—based upon the presence and severity of clinical findings (Table 1).

Abdominal cramping can occur, presumably because of gut distension. Cramping of the extremities is a common symptom. Carpopedal spasm and tetany may occur because of alterations in calcium homeostasis resulting from rapid changes in the acid–base status. The presence of coma may indicate severe hypoglycaemia.

Laboratory features

Laboratory abnormalities in patients with cholera result from the intravascular volume contraction and resultant prerenal azotaemia. Creatinine and blood urea nitrogen values are elevated as a result of the prerenal failure, and the packed cell volume and serum protein concentration are elevated as a result of haemoconcentration. Patients are acidaemic because of bicarbonate loss in the stool and lactic acidosis from volume contraction and hypoperfusion. There is no inflammatory response, although leucocyte numbers may be mildly increased because of the haemoconcentration.

Laboratory diagnosis

Cholera is a clinical diagnosis, especially because most cases of cholera occur in locations where laboratory facilities are not readily available. Because the treatment of any severely dehydrating diarrhoea is the same—fluid replacement—identification of the pathogen is not essential for patient management. As severe dehydration

is rare in adults, in the right epidemiological setting the existence of adults with severe dehydration should alert the clinician and public health authorities to the presence of cholera.

A definitive diagnosis is made by isolating *V. cholerae* from stool or rectal swab samples on selective media, and then using sera to identify the pathogenic serogroups 01 and 0139 and Ogawa or Inaba serotypes using slide agglutination tests. Since routine enteric media are not appropriate for the identification of *V. cholerae*, a more-selective media such as thiosulphate–citrate–bile salts–sucrose agar (**TCBS**) should be used. Because TCBS is expensive, and not always available, some laboratories in developing countries rely on less selective and efficient media, such as gelatin, meat extract, or MacConkey agar.

Specimens for dispatch to a laboratory distant from the site of patient care should be placed in a transport media: Cary–Blair medium is the most effective because of its high pH and ready commercial availability. Alkaline peptone water can be used when the time required for transport to laboratory is 6 h or less. When patients are in the same facility as the laboratory, plating should be done at the bedside. Because *V. cholerae* are excreted in such high numbers in stool (>10^7 organisms/ml of stool) enrichment of samples before plating is not routinely required.

Serological tests (vibriocidal or antitoxin antibodies) are useful only for retrospective epidemiological studies.

Treatment

This section will focus on treatment under conditions in which most cholera patients present—clinics or hospitals in developing countries with few resources. Only 35 cases of cholera were reported in Europe during 2000, and 9 cases in North America, and so the chances of physicians caring for patients with cholera in industrialized countries are remote.

Initial intravenous therapy

Treatment of the severely dehydrated patient with cholera is a medical emergency. With appropriate therapy, no patient with cholera who reaches a treatment facility alive should die; without adequate therapy, the death rate may be as high as 50 per cent.

The cornerstone of the treatment of cholera patients is rapid replacement of the fluid deficit. Estimates of the degree of dehydration should be made using the categories listed in Table 1, and the corresponding fluid deficit replaced in 2 to 4 h. Patients with severe dehydration should have their fluid volume replaced using intravenous fluid. The composition of the fluid used should closely resemble that lost in the cholera stool (Table 2). Such fluids have been developed in areas where cholera is common (Dhaka solution and Peru polyelectrolyte solution). Ringer's lactate is the commercially available intravenous solution that most closely meets these requirements. In the absence of an appropriate solution, the emphasis should still be on volume replacement. The maxim 'the dumbest kidney is smarter than the smartest intern' is appropriate here; if the intravascular volume (and renal perfusion) is restored, the kidney will achieve, albeit more slowly, electrolyte and acid–base homeostasis.

Deaths due to cholera usually occur because of the failure to realize the extent of fluid requirements in severely dehydrated patients (7 to 10 litres required during the first 2–4 h for a 75-kg severely dehydrated individual) and the need for close monitoring and continued high-volume fluid replacement after initial rehydration (Table 3). Patients who are not closely monitored can quickly again become dehydrated—and this time not in the high-visibility area of the triage or admission desk, but in the far corner of a rehydration tent set up for delivering care during an epidemic. Patients should be monitored every 1 to 2 h during the first 24 h of illness. Stool and urine volume should be collected and quantified. This can be most expeditiously achieved using a cholera cot—a simple cot with a plastic sheet and a hole in the middle so that stools and urine drain into a calibrated bucket. Rectal catheters can also be used for stool collection. Fluid replacement should then be adjusted to match continued fluid losses.

Although large-volume intravenous replacements are best accomplished using large-bore needles, any fluid replacement is better than none: even small-calibre needles can be used to initiate therapy. If an intravenous line cannot be established, oral rehydration solutions should be given by mouth if the patient is alert, or otherwise using a nasogastric tube.

Antimicrobial therapy

Antimicrobial therapy can halve the duration and volume of diarrhoea. All patients requiring intravenous therapy, or admission to clinic or hospital, should receive an antimicrobial agent. Almost all patients given an effective antimicrobial drug can be discharged within 24 h of admission, in contrast to 72 h or more if they are left untreated. Especially during epidemics, this reduction in hospital stay, and the associated reduction in demand for intravenous and oral fluids, can be critical for an effective response by already strained healthcare services.

Single-dose therapy—of which there are a number of options—is the preferred regimen, especially in epidemic settings (Table 4). Resistance in *V. cholerae* is not predictable, and hence the need to obtain isolates for susceptibility testing during an outbreak. Since outbreak strains are usually clonal, empirical therapy can be based upon a limited number of isolates. Resistance to fluoroquinolones and azithromycin has not been reported, and if these drugs are available at generic prices they, along with doxycycline (if the strain is susceptible), are the drugs of choice.

Oral rehydration

Oral rehydration fluids containing glucose and salts were developed following the observation that, although cholera toxin poisons the neutral sodium-chloride absorption channels in the intestinal mucosa, the glucose-mediated cotransport of sodium (and water) remains intact. Oral rehydration fluids should be given immediately after the onset of diarrhoea in an effort to prevent the development of dehydration. They should also be used in severely dehydrated patients following rehydration. Most patients can be managed with oral fluids alone within 12 to 24 h of admission. Provision of oral rehydration fluids is an inexact art, but the amount provided should be somewhat more than the volume of stool lost.

The most readily available oral rehydration salts are the sachets distributed by UNICEF and the WHO containing 90 mmol of sodium for reconstitution in water. Homemade solutions—using sucrose and salt, or cereal and salt—are also effective. Oral rehydration solution can be drunk from a cup by adults or fed to young children and infants using a spoon (Fig. 2).

Fig. 2 Mother providing oral rehydration solution to her reluctant child using the recommended method of a cup and spoon. In the background note the cholera cots and the buckets placed underneath the cutouts in the cot ('poop-chutes') that allow stool to be measured. Cots are covered with a plastic liner that is changed daily.

Complications

Rapid correction of the acidosis in patients with cholera may reduce ionized calcium concentrations, resulting in tetany. If tetany occurs, the rehydration solution used should be changed to normal saline for a brief period. Rapid correction of the acidosis can also result in a drop in the serum potassium concentration, but this is rarely symptomatic.

Renal failure is rare. Most patients thought to have renal failure actually have inadequate fluid replacement. Even with rapid rehydration, most children and adults produce no urine during the first 4 h of treatment, and only a median of 1 ml/kg body weight in the next 24 h. Patients with diminished urine output should be followed, and their creatinine level measured if possible. Additional fluids should be given, and the patients followed closely for oedema and other signs of fluid overload.

Hypoglycaemia occurs because of a failure of gluconeogenesis in stressed children; in Bangladesh the rate of severe hypoglycaemia in children with dehydrating cholera was 0.5 per cent. Thus dextrose-containing solutions, such as Ringer's lactate with 5 per cent dextrose, are preferred for rehydrating patients with cholera. In the absence of such a solution, children should have their blood glucose measured with a glucometer, or be given a bolus of glucose if they are in a state of altered consciousness.

If the patient develops a fever, or appears septic, the most likely cause is contamination of the infusate and/or the intravenous apparatus. The treatment is to replace both. Adherence to infection-control techniques is an important part of the management of patients, especially in epidemic situations.

Prevention and future research

The provision of clean water is the primary means of preventing cholera. In the absence of potable water, chlorine or iodine can be added to drinking water. Alum potash has also been reported to be effective. All sterilizing agents are ineffectual in water that has a high turbidity from suspended organic material. Food is best eaten cooked, and not from street vendors.

There are currently three vaccines available for cholera: a parenterally administered vaccine containing whole cells killed by phenol; and two oral vaccines: a killed, whole-cell recombinant B-subunit toxin vaccine; and a live, attenuated, *V. cholerae* vaccine strain—CVD 103—that does not express the cholera-toxin A subunit. The parenteral vaccine is of uncertain efficacy and is toxic because of its lipopolysaccharide content; the two oral vaccines have minimal toxicity and provide limited protection for short periods to persons living in endemic areas. The oral vaccines are not widely available in developing countries, and are not licensed in the United States. These vaccines are of limited utility; travellers are at a miniscule risk of contracting cholera, and the limited long-term efficacy of these vaccines makes them inappropriate for routine use in developing countries. They may have more use during epidemics; but epidemics are likely to have run their course before supplies can be mobilized and immunity induced.

There is a need to develop a vaccine that provides long-duration, high-level protection against cholera. Perhaps the most pressing research need is for a better understanding of how resources can be mobilized to provide clean water and sanitation to the billions of persons who currently lack it.

Further reading

Dhar U, *et al*. (1996). Clinical features, antimicrobial susceptibility and toxin production in *Vibrio cholerae* O139 infection: comparison with *V. cholerae* O1 infection. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **90,** 402–5.

Faruque SM, Albert MJ, Mekalanos JJ (1998). Epidemiology, genetics, and ecology of toxigenic *Vibrio cholerae*. *Microbiology and Molecular Biology Reviews* **62**(4)**,** 1301–14.

Field M, *et al*. (1972). Effect of cholera enterotoxin on ion transport across isolated ileal mucosa. *Journal of Clinical Investigation* **51,** 796–804.

Heidelberg JF, *et al*. (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406,** 477–83.

Hirschhorn N, *et al*. (1968). Decrease in net stool output in cholera during intestinal perfusion with glucose-containing solutions. *New England Journal of Medicine* **279,** 176–81.

Khan WA, *et al*. (1996). Randomised controlled comparison of single-dose ciprofloxacin and doxycycline for cholera caused by *Vibrio cholerae* 01 or 0139. *Lancet* **348,** 296–300.

Ryan ET, Calderwood SB (2000). Cholera vaccines. *Clinical Infectious Diseases* **31,** 561–5.

Waldor MK, Mekalanos JJ (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272,** 1910–14.

7.11.12 Haemophilus influenzae

E. R. Moxon

General

Haemophilus influenzae, a Gram-negative bacterium, is a commensal and potential pathogen that resides in the nasopharynx, the conjunctivae, and occasionally the genital tract of humans. Carriage of one or more strains for periods of days to months is common and most carriers are, and remain, healthy. However, *H. influenzae* is pathogenic and can result in two distinct patterns of disease (Table 1). First, there are infections in which there is invasion of the bloodstream and dissemination to distant sites, for example the meninges or synovial joints. These are usually caused by encapsulated type b strains and occur typically in infants. Second, there are infections that occur as a result of contiguous spread of *H. influenzae* within the respiratory tract, for example otitis media, sinusitis, and pneumonia. These are usually, but not invariably, caused by unencapsulated or non-typeable (**NT**) strains and are relatively common in children; however, they also occur in adults.

Epidemiology, pathogenesis, and immunology

Humans are the sole reservoir of *H. influenzae*; person-to-person spread is therefore crucial to the survival of the species. Transmission occurs by airborne droplets, or by direct contagion with secretions. The age of acquisition is extremely variable. In socio-economically deprived countries, most children are densely colonized with *H. influenzae* immediately after birth, whereas acquisition may be delayed for several weeks in infants living in, for example, Europe or the United States. Most of the colonizing strains are unencapsulated or so-called non-typeable (NT) organisms, but in 3 to 5 per cent of people, the *H. influenzae* express one of six, antigenically distinct polysaccharide capsules, designated a to f, the basis of the major typing system. Carriage of several different strains concurrently has been well described. Over time, phenotypic changes in major surface antigens, such as outer membrane proteins, occur in response to host immune selection pressures. The factors influencing acquisition and colonization of *H. influenzae* include a variety of surface adhesins, including pili, the production of IgA1 proteases, the inhibition of host clearance mechanisms by the inhibitory effect of cell wall glycopeptides, and the production of both local and serum antibodies. Colonization is a permissive event in the pathogenesis of disease and the importance of the type b capsule as a crucial factor in systemic, bacteraemic infections has been well established in animal models. Capsule impedes the clearance of organisms by phagocytes and complement-mediated killing. The core sugars of lipopolysaccharide also play an important role in promoting survival of *H. influenzae* and another key component, endotoxin (lipid A), is critical in mediating the damage to tissues, such as inflammation and breakdown of the blood–meningeal barrier. Prior viral infections such as influenza potentiate infection and appear to facilitate both contiguous spread within the respiratory tract—as in otitis media, sinusitis, or lower respiratory tract infection—–and the probability of dissemination into the blood.

Serum antibodies to type b capsule mediate protective immunity against systemic infections in humans. The serum of newborn babies and young infants, up until the age of 3 months, generally has sufficient amounts of passively acquired maternal antibodies to afford protection. Thereafter, the natural decline of maternally derived antibodies is followed by a period lasting until the age of 2 to 4 years when the levels of antibody are absent, or inadequate to provide protection.

In contrast to systemic type b infections where deficiencies in opsonophagocytic mechanisms are paramount, impairment of non-specific host defence mechanisms (e.g. impaired ciliary clearance) is the most obvious feature of those who have disease caused by NT *H. influenzae*. Other predisposing factors include smoking, viral infections, immunodeficiency, or chronic lung disease such as cystic fibrosis.

Haemophilus influenzae type b

Meningitis

Despite the availability of antibiotics, and more recently the highly effective conjugate vaccines, type b meningitis remains the commonest cause of purulent meningitis in early childhood worldwide, and the cause of many deaths and permanent central nervous system damage in survivors. The majority of the cases occur in young children aged less than 5 years, the peak incidence being from about the age of 3 months to 2 years. Reported risk factors include male sex, black rather than white race, absence of breast feeding, socio-economic deprivation, winter months, siblings (often asymptomatic carriers), and attendance at day-care or preschool nurseries.

Typically, meningitis presents after a few hours or days of antecedent symptoms, most commonly those of an upper respiratory tract infection in a young child; an associated or preceding otitis media is common. The most common symptoms and signs are fever, lethargy, vomiting, neck stiffness, and altered nervous system function, ranging from irritability to coma, but young babies may be afebrile and have few symptoms or signs. Raised intracranial pressure produces headache and vomiting and may cause a bulging fontanelle in young infants. Seizures are common in children; subdural effusions are present in about 33 per cent of children and occur most frequently in young infants. The key to diagnosis is to perform a lumbar puncture and examine the cerebrospinal fluid. This typically reveals inflammatory cells, raised protein and lowered glucose concentrations, and there are often organisms that can be seen by microscopy after staining with Gram's stain or methylene blue.

If diagnosis and treatment are prompt, more than 95 per cent of patients with *H. influenzae* meningitis will survive, but about 8 per cent of survivors have serious central nervous system sequelae, the commonest being sensorineural deafness. Before the advent of effective vaccines, this was said to be the most important cause of acquired mental handicap in the United States.

Epiglottitis

Acute respiratory obstruction, caused by a cellulitis of the epiglottis and aryepiglottic folds, usually occurs as a fulminating, life-threatening infection. Sore throat, fever, and dyspnoea progress rapidly to dysphagia, pooling of oral secretions, and drooling of saliva from the mouth. The child is toxic, restless, anxious or lethargic, and adopts a sitting position with an extended neck and protruding chin in an effort to to minimize airway obstruction. The voice and cry are muffled and the child may be reluctant to talk. Stridor is often absent; if present it is soft and wheezy. Cough is unusual. In the absence of adequate treatment death commonly occurs within a few hours. The course may be less dramatic with a prodromal illness of sore throat and hoarseness from one to several days preceding the onset of acute symptoms. The characteristic findings are that the epiglottis is red and swollen, obstructing the pharynx at the base of the tongue. Lateral radiographs reveal the swollen 'thumb-shaped' epiglottis. Examination of the larynx should be attempted only where there are facilities for immediate intubation/tracheotomy, since fatal respiratory obstruction may occur abruptly. The most important aspect of management of acute epiglottitis is the provision of an adequate airway and ventilation in addition to

antibiotic treatment.

Pneumonia and empyema

Lower respiratory tract infections occur most often in children aged less than 5 years and present as lobar pneumonia, often with pleural involvement. In many of the poorer countries of the world, such as New Guinea, *H. influenzae* type b pneumonia is second only to that caused by the pneumococcus and in these regions is a greater public health problem than *H. influenzae* meningitis. Type b pneumonia is also well recognized in adults as a primary cause of pneumonia, especially in alcoholics.

Cellulitis

This important infection occurs in young children who present with hectic fever and a raised, warm, tender area of distinctive reddish-blue hue, most often located on one cheek or in the periorbital region, that evolves over a few hours.

Septic arthritis

H. influenzae type b is one of the commonest causes of septic arthritis in children of less than 2 years of age. Typically, there is involvement of a single large, weight-bearing joint, usually without osteomyelitis. Response to drainage and appropriate systemic antibiotics is usually dramatic and apparently curative, but long-term follow-up is important since residual joint dysfunction occurs in a proportion of children.

Treatment of diseases caused by type b strains

Prior to the availability of antibiotic treatment, *H. influenzae* meningitis was invariably fatal. With the introduction of chloramphenicol in 1950, survival rates of 95 per cent or more have been possible. Overall, chloramphenicol remains an excellent drug for treating *H. influenzae* meningitis, but occasional isolates show resistance. Chloramphenicol carries a dose-related, reversible bone marrow toxicity, but this is rarely clinically a problem and can be completely avoided if blood levels are monitored. Idiosyncratic bone marrow aplasia has been reported but is extremely rare. Ampicillin, formerly considered an ideal treatment for *H. influenzae* meningitis, is no longer favoured because of the relatively high prevalence of resistant (b-lactamase producing) strains. The treatment of choice is parenteral third-generation cephalosporins such as ceftriaxone or cefotaxime; these have been shown to be highly effective as initial treatment of suspected bacterial meningitis. Cefuroxime is less effective.

Young children in the same household as a patient with invasive type b disease are at significantly increased risk of secondary invasive infection by *H. influenzae* type b. Rifampicin given orally once daily for 4 days is effective in eradicating nasopharyngeal carriage, and is recommended for all household contacts (children and adults).

Experimental and clinical studies support the administration of corticosteroids to reduce the incidence of neurological sequelae, especially sensorineural deafness. The presumed mechanism is the reduction of inflam-mation that results from release of bacterial cell wall fragments. Dexamethasone therapy (0.6. µg/kg.day) intravenously in four divided doses for 4 days is recommended for children older than 2 months of age.

Active immunization

In the 1940s, serum antibodies specific for the type b capsule were used as treatment of type b infection. Efforts to develop a vaccine for active immunization using purified type b capsule did not begin until the 1970s. However, by the 1980s, it was clear that this vaccine did not protect children aged less than 2 years old. Further research was directed towards developing conjugate vaccines in which type b capsule is covalently linked to a carrier protein, such as tetanus toxoid. Several commercially manufactured conjugate vaccines have been licensed and all have proved to be very safe and capable of affording high levels of protection to children immunized as early as 2 months of age. In the United Kingdom, conjugate vaccines have been given to infants as part of the routine immunization schedule since 1992 and their protective efficacy is more than 95 per cent.

Diseases caused by non-typeable *H. influenzae*

Pneumonia

Non-typeable strains are an important cause of pneumonia in children and adults, especially the elderly, and in those with established lung disease, such as chronic bronchitis. In many countries where adverse socio-economic circumstances are prevalent, acute lower respiratory tract infections in infants caused by NT *H. influenzae* represent an uncertain but probably major cause of morbidity and mortality.

It has been recognized for many years that exacerbations of chronic bronchitis correlate with an increase in the production of purulent sputum from which NT *H. influenzae* strains are cultured. Such episodes are often precipitated by prior viral infection. Progressive lung damage in conditions such as chronic bronchitis, cystic fibrosis, and hypogammaglobulinaemia is thought to result from heightened and protracted inflammatory response to a variety of bacteria, including NT *H. influenzae*, in people whose respiratory tract lacks the appropriate clearance mechanisms.

Maternal and neonatal sepsis

NT *H. influenzae* are a well-documented cause of tubo-ovarian abscess or chronic salpingitis. More ominously, the infants born to such mothers, often prematurely, may develop life-threatening neonatal septicaemia, meningitis, and a form of acute respiratory distress syndrome that is indistinguishable from that caused by group B streptococci.

Acute otitis media and sinusitis

H. influenzae accounts for about one-fifth of all cases of acute bacterial otitis media. More than 90 per cent of the organisms isolated from middle ear fluid are NT strains. Although such episodes occur at any age, they are most common in children aged 6 months to 5 years. Since more than two-thirds of children have one or more episodes of otitis media by the age of 3 years, a conservative estimate would indicate that more than 100 000 cases of *H. influenzae* otitis media occur each year in the United Kingdom. NT strains are also a common cause of sinusitis in both adults and children.

Conjunctivitis

H. influenzae is an important cause of purulent conjunctivitis. Most are NT strains that were formerly considered to be sufficiently distinctive to be referred to as *H. aegyptius*. Interest in these strains was heightened when, in 1984, an apparently new and serious disease was described in Brazilian children who developed a life-threatening infection known as Brazilian purpuric fever. Its peak age incidence is 1 to 4 years; purulent conjunctivitis, high fever, vomiting, purpura, vascular collapse, and a high mortality are characteristic.

Other infections

All of the diseases that are commonly caused by type b strains can, on rare occasions, be caused by strains of capsular serotypes a, c, d, e, and f as well as NT strains. A number of other unusual infections have been described including: endocarditis, pericarditis, peritonitis, and epididymo-orchitis. Two other closely related species, *H. parainfluenzae* and *H. aphrophilus*, are also causes of disease, such as endocarditis.

Treatment

Serious infections caused by NT strains such as meningitis, lower respiratory tract infections, tubal abscess, and neonatal sepsis require systemic treatment with third-generation b-lactams or co-trimoxazole. Chloramphenicol is highly effective but blood levels need to be monitored carefully, especially in young infants, because

of potential toxicity. Sinusitis and otitis media are often treated effectively with oral amoxycillin, but augmentin would be preferable, given the relatively high incidence of strains producing b-lactamase. Oral co-trimoxazole would be an equally sound or alternative choice for trimethoprim-susceptible strains. The use of antibiotics as pro-phylaxis or treatment of exacerbations of chronic bronchitis is controversial, but many advocate their use either to reduce the number of haemophili in the lower respiratory tract or to eradicate them. Drugs of the tetracycline group are effective, but are contraindicated in pregnancy, patients with impaired renal function, or children less than 10 years of age; amoxycillin and co-trimoxazole have also proved useful.

Passive immunization

People with increased susceptibility to infection with *H. influenzae*, but particularly NT strains, may have a deficiency of antibody synthesis. They benefit from immunoglobulin preparations administered either intramuscularly or intravenously. This form of immunoglobulin replacement undoubtedly decreases the incidence of systemic infections and the number of episodes of both upper and lower respiratory tract infections caused by NT *H. influenzae*.

Further reading

Booy R *et al*. (1997). Surveillance of vaccine failures following primary immunisation of infants with Hib conjugate vaccine: Evidence for protection without boosting. *Lancet* **349**, 1197–1202.

Hoiseth SK (1991). The genus *Haemophilus*. In: Balows A *et al*., eds. *The prokaryotes, a handbook on the biology of bacteria: ecophysiology, isolation, identification, applications.* Springer-Verlag, New York.

Moxon ER, Murphy TF (1999). *Haemophilus influenzae*. In: Mandell GL, Bennett JE, Dolin R, eds. *Mandell, Douglas, and Bennett's principles and practice of infectious diseases*, 5th edn, pp 2369–78. Churchill Livingstone, Philadelphia.

Murphy TF, Apicella MA (1987). Non-typeable *Haemophilus influenzae*: A review of clinical aspects, surface antigens, and the human immune response to infection. *Reviews of Infectious Diseases* **9**, 1–15

Turk DC (1982). Clinical importance of *Haemophilus influenzae*. In: Sell SH, Wright PF, eds, Haemophilus influenzae, *epidemiology, immunology and prevention of disease*, pp 30–3. Elsevier Biomedical, New York.

7.11.13 Haemophilus ducreyi and chancroid

Allan R. Ronald

Introduction

Genital ulcer disease is the presenting feature of sexually transmitted diseases in about 5 per cent of patients in Western societies; in the developing world, 10 to 50 per cent of patients with sexually transmitted diseases present with genital ulcer disease. In the West, genital herpes and primary syphilis are the commonest aetiological agents. In the developing world *Haemophilus ducreyi* accounts for most genital ulcer disease. Granuloma inguinale and lymphogranuloma venereum are occasionally imported from remaining foci in sub-Saharan Africa and Asia.

The epidemiological association between genital ulcer disease, particularly chancroid, and the risk of transmission of HIV-1 and HIV-2 has increased interest in its control.

Soft chancre was differentiated from the hard indurated chancre of syphilis by Ricord in 1838. In 1889, the Neapolitan physician Ducreyi identified short-chaining streptobacillary rods in exudate from ulcers following inoculation with chancroid pus.

Aetiology

Haemophilus ducreyi is a faintly bipolar staining Gram-negative rod. Due to extracellular linkage of the bacteria, the organism forms chains and demonstrates a 'school of fish' arrangement. *H. ducreyi* colonies appear after 48 h of incubation, are yellow-grey, dome-shaped, and variable in size and opacity. Colonies are cohesive and can be nudged intact with a straight wire.

Epidemiology

Chancroid is endemic in eastern and southern Africa, India, and the Caribbean where the annual incidence in adult males can exceed 1/1000. It occurs sporadically in industrialized countries, most frequently at major ports of entry. During the last two decades, there have been over 20 discrete outbreaks in North America. Prostitutes are the usual reservoir for dissemination of *H. ducreyi* and the male to female ratio usually is 5:1 or higher. Male circumcision decreases susceptibility to infection by about threefold.

Asymptomatic carriage has no proven role in the spread of *H. ducreyi*. In one study of men with culture-positive chancroid, all source contacts had genital ulcers. *H. ducreyi* is rarely transmitted non-sexually. Chancroid lesions on the fingers or breasts reflect direct contact from a genital lesion on the sexual partner or autoinoculation.

Pathogenesis and pathology

After an incubation period of 3 to 10 days, an inflammatory papule develops which ulcerates. Bacterial virulence factors include a haemolysin and a cytolethal distending toxin that interferes with intracellular signalling. Both humeral and cell-mediated responses to *H. ducreyi* occur, but their role in preventing or modifying infection is unknown. On histological examination, perivascular and interstitial mononuclear cell infiltrates predominate with occasional giant cell granulomas. Endothelial disruption with neutrophil invasion occurs superficially.

Clinical features

Chancroid begins as a tender papule which ulcerates. It is painful, rarely indurated, irregular, and sharply demarcated, usually with no surrounding inflammation. The ulcer base is uneven with a greyish-yellow exudate which bleeds readily. About 50 per cent of men and most women have multiple ulcers.

Numerous variants of chancroid occur including giant rapidly spreading ulcers, dwarf chancroid resembling herpes, follicular chancroid that mimics pyogenic infection, transient ulceration associated with lymphadenitis similar to lymphogranuloma venereum, a painless single ulcer similar to primary syphilis, and raised indurated 'beefy' lesions not unlike granuloma inguinale. In the absence of laboratory investigation, in men as many as 25 per cent and in women at least 50 per cent of ulcers could be attributed on clinical surmise to aetiological agents other than *H. ducreyi*. The index of suspicion for chancroid increases where the disease is highly prevalent.

Chancroid occurs anywhere on the genitalia. However, in uncircumcised men, over 50 per cent of ulcers are on the prepuce. The coronal sulcus is a common site with a circle of ulcers surrounding the entire sulcal circumference. Contact lesions are common on adjacent cutaneous surfaces. In women lesions occur in decreasing frequency on the fourchette, labia majora and labia minora, perianal area, and medial aspects of the thighs. Cervical and vaginal ulcers are uncommon.

Inguinal lymphadenopathy appears in about 40 per cent of men and 20 per cent of women within 7 to 10 days of ulceration. The lymph nodes are discrete, very tender, and often bilateral. If untreated, lymphadenitis progresses to a suppurative bubo which may form an inguinal abscess. Abscesses can penetrate deeply into the groin.

Laboratory diagnosis

Definitive diagnosis of chancroid requires culture of *H. ducreyi*. The Gram stain is not sufficiently sensitive or specific to diagnose *H. ducreyi* infection. No serological test is available. Diagnostic nucleic acid probes are under investigation. The sensitivity of *H. ducreyi* culture is in the range of 50 to 80 per cent. However, specificity is high as asymptomatic carriage of *H. ducreyi* is rare. Two or more sexually transmitted pathogens are present in 10 to 15 per cent of patients presenting with genital ulcers; *H. ducreyi* may be cultured concomitantly with either *Herpes simplex* or *Treponema pallidum*. Although the classic features of syphilis and chancroid appear to place them at opposite ends of a spectrum of genital ulceration, in about 20 per cent of patients the presentations are indistinguishable.

Whenever possible, exudate from the ulcer or bubo should be inoculated directly on to the primary selective media. Organisms will survive longer on a swab at 4°C than at room temperature. *H. ducreyi* grows well on gonococcal agar with added vancomycin (3 mg/l) to inhibit growth of Gram-positive bacteria, a vitamin supplement, and 0.25 per cent activated charcoal. Cultures for *H. ducreyi* should be incubated at 33°C in 5 per cent carbon dioxide and maximum humidity. A candle extinction jar with a moist paper towel is adequate. Distinct colonies appear within 72 h. *H. ducreyi* is identified by its Gram stain and its ability to use nitrate, a positive oxidase test, and a requirement for X factor.

Agar dilution tests with *H. ducreyi* correlate with the clinical response. Plasmid-mediated resistance, as in *Neisseria gonorrhoeae* and *Haemophilus influenzae*, encodes for b-lactamase production; other plasmids enable sulphonamide, tetracycline, and chloramphenicol, kanamycin, and streptomycin resistance. These plasmids have spread rapidly.

Fortunately, all isolates remain susceptible to the third-generation cephalosporins, the fluoroquinolones, and the macrolides.

Treatment

In the absence of specific treatment, chancroid is a prolonged illness with slow resolution and frequent recurrence. Genital ulcers and inguinal abscesses have been reported to persist for years.

Circumcision, cleanliness, and saline soaks were used prior to the sulphonamides. Ampicillin, streptomycin, and tetracycline were each shown to be equivalent treatment regimens with a mean time to complete healing of 10 days. Trimethoprim/sulphonamide combinations became standard therapy, but the emergence and rapid spread of trimethoprim resistance has thwarted its continuing use.

Other treatment regimens include ceftriaxone (a single dose of 250 mg intramuscularly), ciprofloxacin (a single oral dose of 500 mg), fleroxacin (a single dose of 400 mg), erythromycin (250 mg three times a day for 7 days), and azithromycin (a single oral dose of 1 g). All cure over 95 per cent of HIV-seronegative men with chancroid. Patients with chancroid concurrently infected with HIV are more likely to fail to respond to treatment with b-lactam antibiotics.

Epidemiological associations between HIV-1 and*Haemophilus ducreyi*

Chancroid is a risk factor for the heterosexual spread of HIV-1 and HIV-2. Chancroid in women increases the risk of acquisition of HIV-1 following heterosexual contact with HIV-1-infected men by four- to eightfold. The presence of chancroid in HIV-1-infected individuals increases the shedding of HIV-1 and the probability that partners will become HIV-1 infected.

Prevention and control

The control of chancroid can reduce heterosexual transmission of HIV substantially, perhaps by 30 per cent or more, in societies where both pathogens are being spread, particularly from prostitutes to their clients. Effective control of chancroid has been achieved on numerous occasions by treating men with ulcers and their sexual contacts. Most women who are source contacts of men with chancroid have few symptoms, despite the presence of ulcers, and so contact tracing is essential. The use of condoms by clients dramatically reduces the acquisition of chancroid from prostitutes.

Chancroid control is an essential cost-effective intervention to slow the transmission of HIV-1.

Further reading

Cameron DW *et al*. (1989). Female to male transmission of human immunodeficiency virus type 1: risk factors for seroconversion in men. *Lancet* **ii**, 403–7.

Coqtes-Bratti X *et al.* (1999). The cytolethal distending toxin from the chancroid bacteria *Haemophilus ducreyi* induces cell-cycle arrest in the G2 phase. *Journal of Clinical Investigation* **103**, 107–15.

Martin DH *et al*. (1995). Comparison of azithromycin and ceftriaxone for the treatment of chancroid. *Clinical Infectious Diseases* **21**, 409–14.

Ndinya-Achola JO *et al*. (1996). Presumptive specific clinical diagnosis of genital ulcer disease (GUD) in a primary health care setting in Nairobi. *International Journal of AIDS and STD* **7**, 201–5.

Trees DL, Morse SA (1995). Chancroid and *Haemophilus ducreyi*: an update. *Clinical Microbiology Reviews* **8**, 357–75.

7.11.14 Bordetella

Calvin C. Linnemann, Jr

Bacteria of the genus *Bordetella* are primarily pathogens of the respiratory tract of humans and animals because they can adhere to ciliated epithelial cells. The whooping cough syndrome or pertussis is characterized by paroxysmal coughing, an inspiratory whoop, and lymphocytosis. *B. pertussis*, *B. parapertussis*, and *B. bronchiseptica* can cause disease in man. *B. holmesii* has also been recovered from patients with whooping cough. Misattribution of pertussis to viral infection resulted from the difficulty in isolating *B. pertussis* and frequent coinfection with adenoviruses.

Bordetella infections should be suspected in patients with persistent lower respiratory tract infection and paroxysmal coughing, with or without an inspiratory whoop, or those with any respiratory symptoms after close contact with a documented infection. Most bordetella infections will go unrecognized because the symptoms are indistinguishable from other respiratory tract infections, and because appropriate diagnostic tests are usually done only in patients with typical pertussis. *B. bronchiseptica*, a common pathogen in animals, should be considered in animal handlers with respiratory tract infections.

The causative agent

Bordetella are small, aerobic, Gram-negative coccobacillary organisms. *B. pertussis* are slow growing and are inhibited by a variety of media constituents such as fatty acids that must be inactivated if culture is to be effective (see below). *B. parapertussis* and *B. bronchiseptica* are less fastidious and faster growing.

Bordetella pertussis adheres to ciliated epithelial cells in the respiratory tract. Attachment is followed by ciliostasis and subsequent loss of the ciliated cells. Biologically active components include filamentous haemagglutinin, fimbrias, pertactin, pertussis toxin (lymphocytosis-promoting factor), adenylate cyclase, and tracheal cytotoxin. Acellular vaccines, containing only selected components such as the filamentous haemagglutinin and pertussis toxin, protect against severe symptomatic infection. *B. pertussis* is non-invasive, usually remaining on the surface of the respiratory tract, but *B. parapertussis*, *B. bronchiseptica*, and *B. holmesii* bacteraemias have been reported.

Epidemiology

Humans are the only known reservoir of *B. pertussis* and *B. parapertussis*, whereas *B. bronchiseptica* is found in other mammals. *B. pertussis* is transmitted by droplets from symptomatic patients. Asymptomatic infections are not important in the spread of disease, and there are no chronic carriers. It is assumed that the transmission of *B. parapertussis* is similar to that of *B. pertussis.* Humans and other mammals may be reservoirs of *B. bronchiseptica*.

Before vaccine was available, epidemics of *B. pertussis* spread through schools, and were carried by the schoolchildren to their homes. Secondary attack rates in susceptible children were 25 to 50 per cent in schools, and 70 to 100 per cent in homes, reflecting the intense and prolonged exposure at home. Most children developed symptoms. Mild infections or reinfections in adults caring for sick children were known as 'grandmother's cough' or 'nurse's cough'.

In the vaccine era, major epidemics have disappeared in most developed countries. Mortality from *B. pertussis* was decreasing before the introduction of vaccine, but not the number of cases. In the United States and Canada, where effective vaccines have been widely used, the incidence of pertussis has decreased to 1 to 3/100 000 per year (Fig. 1). Results were less dramatic in the United Kingdom, related, perhaps, to early problems with vaccine efficacy and lower levels of vaccine usage. Pertussis did decrease in the United Kingdom, and the resurgence of *B. pertussis* in the late 1970s, following a decrease in vaccine usage, demonstrated the efficacy of vaccine.

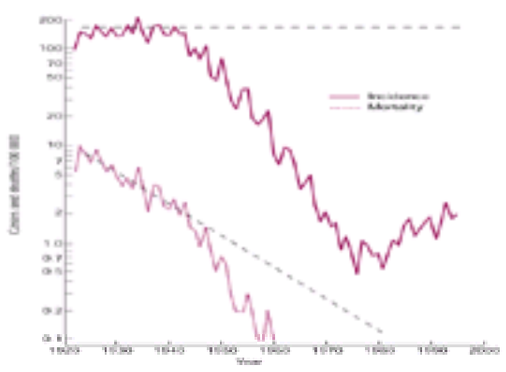

Fig. 1 The effect of pertussis vaccine on the incidence and mortality of pertussis in the United States. The lines superimposed on the graph indicate the trends prior to the vaccine and as projected if vaccine had not been introduced.

In a highly vaccinated population, older children and adults make up a larger proportion of cases and may play a more important part in the transmission of disease. Before vaccine was available, the source of infection could be identified in most cases as another child. It is now more difficult to trace the source, but in very young infants an adult family member frequently appears to be the source. Hospital epidemics have also demonstrated the part adults play in transmission. Doctors and nurses may acquire infection from a patient and then transmit it to other hospital staff and to patients.

The epidemiology of *B. parapertussis* has not been modified by vaccine usage. It is widespread in many countries, but it is seldom recognized because of the mildness of the disease. In Denmark epidemics occur every 4 years, alternating with epidemics of *B. pertussis*.

Clinical manifestations

In *B. pertussis* infection, non-specific upper respiratory symptoms, malaise, anorexia, and sometimes a low-grade fever begin 7 to 10 days after infection. This 'catarrhal stage' is indistinguishable from other mild respiratory infections. Towards the end of this stage a dry, hacking cough appears and progresses. Older, presumed partially immune, patients may not progress further. After 1 to 2 weeks, the paroxysmal stage begins and continues for several weeks. The cough is now paroxysmal. Prolonged coughing episodes may be followed by the characteristic 'whoop', produced by forced inspiration through a partially closed glottis. In severe cases, paroxysms of coughing are followed by vomiting, and may be associated with epistaxis, petechias, conjunctival or scleral haemorrhages, haemorrhagic myringitis, or periorbital oedema. Young infants may not have the whoop. Their paroxysms of coughing may be followed by cyanosis and apnoea. Fever is uncommon at this stage in uncomplicated infections. The convalescent stage begins after 2 to 4 weeks, with gradually resolving paroxysms of coughing. Patients may cough for weeks to months. Their whooping may be exacerbated by subsequent viral respiratory infections.

Leucocytosis with lymphocytosis appears toward the end of the catarrhal stage and continues in the paroxysmal stage. Lymphocytosis is most marked when coughing

is worst. There is a proportional increase in both T and B lymphocytes. Lymphocytosis may not occur in very young infants, older children, and adults.

Fever suggests a complicating bacterial infection. Otitis media and pneumonia are the most common. Atelectasis results from bronchial obstruction by the thick mucus. Bronchiectasis is uncommon. High pressures caused by paroxysmal coughing contribute to pulmonary, haemorrhagic, and gastrointestinal complications. These include mediastinal and subcutaneous emphysema, pneumothorax, inguinal hernias, and rectal prolapse. The extremely rare neurological complications include convulsions, paralysis, coma, blindness, deafness, and movement disorders.

B. parapertussis infections are clinically milder. Twenty per cent or less of children will develop the whooping cough syndrome. *B. bronchiseptica* rarely causes whooping cough. It is usually a non-pathogen in the respiratory tract but may cause bronchitis. In immunosuppressed patients, *B. bronchiseptica* can cause sinusitis, tracheobronchitis, and pneumonia. Bacteraemia, endocarditis, peritonitis, and meningitis have also been reported. *B. bronchiseptica* and *B. pertussis* have been reported in HIV-infected patients. *B. holmesii* septicaemia has been reported in compromised hosts.

Diagnosis

Definitive diagnosis is by isolation of the organism. Fluorescent antibody staining of material obtained by nasopharyngeal swabs from patients with *B. pertussis* infections provides only presumptive diagnosis. Polymerase chain reaction can be used for diagnosis, but assays are expensive and not standardized. Antibody responses can be measured in acute and convalescent sera by enzyme immunoassay.

The cough plate technique has been replaced by the nasopharyngeal culture technique. A wire calcium alginate swab is passed through the nose until it touches the posterior nasopharynx, allowed to remain for a few seconds, and removed. Cotton swabs may be used if the cotton has been shown to be non-bacteriostatic for *B. pertussis*. The swabs are streaked on to Bordet–Gengou agar plates, both with and without an antibiotic such as cephalexin. Multiple cultures increase recovery of *B. pertussis.* The organism has not been recovered from blood or other sites. Cultures must be held for 6 days before being discarded.

B. parapertussis, *B. bronchiseptica*, and *B. holmesii* can also be recovered on Bordet–Gengou medium and they also grow on routine media used for recovery of Gram-negative bacteria. These three organisms have been recovered from blood cultures, and *B. bronchiseptica* has been cultured from urine.

Treatment

Most patients can be managed at home. Very young children may need good nursing care in hospital. Cough medicines are useless as is passive immunization with available immunoglobulin preparations. Salbutamol and steroids may be useful. Sedation of young children is potentially dangerous but is practised by some paediatricians.

B. pertussis, *B. parapertussis*, and *B. holmesii* are sensitive to erythromycin, tetracycline, chloramphenicol, and trimethoprim–sulphamethoxazole. Early treatment, during the catarrhal stage, shortens the clinical illness. Treatment started at the paroxysmal stage is much less effective. The best results are achieved by treating symptomatic contacts of patients with diagnosed infections. Despite limited clinical benefit, patients in the paroxysmal stage should be treated to render them non-infectious. Erythromycin is the drug of choice: 40 to 50 mg/kg per day for children, 1.5 to 2 g per day for adults, for 14 days. Nasopharyngeal cultures become negative in the first few days of treatment, but the erythromycin should be continued to prevent bacteriological relapses (Fig. 2). Trimethoprim–sulphamethoxazole has been used in children who do not tolerate erythromycin, although its efficacy has not been proved. The newer macrolides, clarithromycin and azithromycin, at 10 mg/kg per day for 5 to 7 days may be as effective as erythromycin.

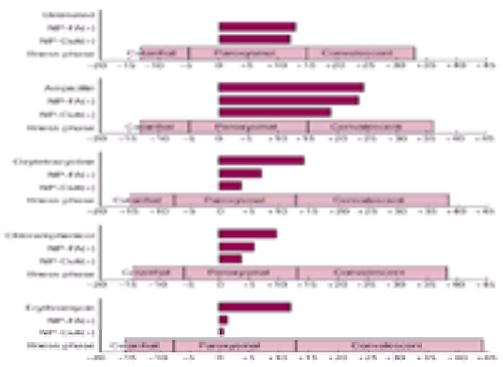

Fig. 2 Duration of excretion of *B. pertussis* as detected by fluorescent antibody staining and culture, and the effect of antimicrobial treatment. The graphs compare (group means) patients treated with antimicrobial agents with untreated control patients. (Reproduced from Bass JW *et al.* (1969), *Journal of Pediatrics* **75**, 768, with permission.)

B. bronchiseptica is sensitive to tetracycline and chloramphenicol but not to erythromycin. Antipseudomonal penicillins and aminoglycosides have been proved successful in serious infections.

Prevention

Patients with *B. pertussis* should avoid close contact with susceptible individuals to prevent droplet transmission. Untreated patients remain contagious for weeks. Communicability decreases rapidly after starting erythromycin. Nasopharyngeal cultures become negative within 48 to 72 h. Patients admitted to hospital are usually isolated for the first 5 days of treatment.

Chemoprophylaxis with erythromycin may be effective. Close contacts of patients with *B. pertussis* infection should be treated with erythromycin. Vaccination should continue according to routine schedules. Some recommend that, in addition to erythromycin, a booster dose of vaccine should be given to preschool children who have not received a booster within 6 months. Lower risk exposures, such as those occurring outside the home or day-care centre, require erythromycin only if respiratory symptoms develop.

Vaccination

Whole-cell pertussis vaccine prevents disease, but frequently causes local reactions, with or without fever, and rare neurological complications. Serious reactions occur less frequently after vaccination than with clinical disease. The killed whole-bacterial preparation is given with diphtheria and tetanus toxoids. An effective immunizing schedule includes three injections at 1- to 2-month intervals beginning at 6 to 12 weeks of age, and a fourth dose given 6 to 12 months after the third. A booster dose is given before entry to school. Acellular vaccines, containing three or four bacterial components—including pertussis toxoid, filamentous haemagglutinin, pertactin, and fimbriae—are replacing whole-cell vaccines, using the same immunization schedules.

Vaccination is usually restricted to children less than 7 years of age because of the local reactions, but immunity is neither complete nor lifelong. Protection may last only 12 years. Duration of immunity after acellular vaccines is unknown. Re-exposure to *B. pertussis* may induce continuing immunity in previously vaccinated patients. In future, acellular vaccines may be deployed in older children and adults. There are no vaccines generally available for *B. parapertussis* or *B. bronchiseptica*.

Further reading

Hoppe JE (1999). Update on respiratory infection caused by *Bordetella parapertussis*. *Pediatric Infectious Disease Journal* **18**, 375–81.

Linnemann CC Jr (1979). Host–parasite interactions in pertussis. In: Manclark C, Hill J, eds. *International symposium on* pertussis, pp 3–18. US Government Printing Office, Washington, DC.

Muller FM, Hoppe JE, Wirsing von Konig CH (1997). Laboratory diagnosis of pertussis: state of the art in 1997. *Journal of Clinical Microbiology* **35**, 2435–43.

Thomas MG (1989). Epidemiology of pertussis. *Review of Infectious Diseases* **11**, 255–62.

Woolfrey BF, Moody JA (1991). Human infections associated with *Bordetella bronchiseptica. Clinical Microbiology Reviews* **4**, 243–55.

7.11.15 Melioidosis and glanders

D. A. B. Dance

Melioidosis

Definition and aetiology

Melioidosis is an infection of humans or animals caused by the saprophytic bacterium *Burkholderia* (previously *Pseudomonas*) *pseudomallei*. It is an ovoid, oxidase-positive, motile, Gram-negative bacillus that often exhibits 'safety-pin' bipolarity. It grows well on most standard culture media, producing wrinkled and smooth colony types and giving off a sweet, earthy smell. It has often been overlooked or discarded as a contaminant by bacteriologists unfamiliar with its characteristics. A pattern of resistance to aminoglycosides and polymyxins with susceptibility to co-amoxiclav is a useful clue to its identity.

Epidemiology

Distribution

Melioidosis is endemic throughout south and south-east Asia and northern Australia. Its incidence varies within these regions in both place and time, being particularly high in north-east Thailand during heavy monsoon years. It is likely to be underdiagnosed unless good laboratory facilities are available. Sporadic cases have been reported from sub-Saharan Africa, Central and South America, the Caribbean, and Iran. A unique outbreak occurred in France during the mid-1970s.

Reservoir and transmission

B. pseudomallei is an environmental saprophyte found in soil and surface water, particularly rice paddy, in endemic areas. A closely related, arabinose-assimilating, avirulent soil organism, *Burkholderia thailandensis*, has recently been recognized and may contribute to the high seropositivity rate in endemic areas (up to 80 per cent by the age of 4 years in north-east Thailand). Humans and animals are probably usually infected through contaminated scratches and abrasions or occasionally aspiration of fresh water, although a specific episode of exposure is rarely identified. Iatrogenic infections have also been described. Other possible modes of acquisition include inhalation and ingestion. Direct transmission from infected humans or animals is extremely rare.

Descriptive epidemiology and risk factors

Melioidosis is a disease of people in regular contact with soil and water, such as rice farmers. It has a bimodal age distribution, with a peak incidence between the ages of 40 and 60 years. Males outnumber females by 3:2 in Thailand but more in Australia and Singapore. The disease is markedly seasonal, some 80 per cent of cases presenting during the rainy season in north-east Thailand, when *B. pseudomallei* accounts for almost 20 per cent of cases of community-acquired septicaemia. Most such infections are probably recently acquired, although periods of latency as long as 29 years have been described, which is highly unusual for a bacterial infection. The proportion of seropositive people who are latently infected is unknown.

Pathogenesis

The outcome of contact with *B. pseudomallei* depends on the size of the inoculum, the virulence of the infecting strain, and the host response. Massive exposure will overwhelm a normal immune system, but most infections are self-limiting, resulting merely in asymptomatic seroconversion.

Host response

Clinically apparent melioidosis is an opportunistic disease, over 70 per cent of patients having underlying predisposition to infection. Diabetes mellitus is particularly strongly associated with melioidosis, but pre-existing renal disease and thalassaemia are also significant independent risk factors, and other reported associations include alcoholism and cirrhosis, malignant disease, immunosuppressive and steroid therapy, and pregnancy. In animal models, a T-helper type 1 immune response confers relative resistance to infection, and g-interferon plays a crucial role in protecting against overwhelming sepsis. However, an overexuberant host response may also be damaging, as serum levels of several cytokines have been associated with a fatal outcome in human melioidosis. Humoral immunity may also play a role in defence, since animals may be passively protected by antibodies to lipopolysaccharide and flagellin, and levels of antilipopolysaccharide II correlate with survival in human melioidosis.

Virulence factors

Many putative virulence factors have been described in *B. pseudomallei* and transposon mutagenesis is proving useful in identifying their relative importance. For example, mutants deficient in one of the two forms of lipopolysaccharide produced by *B. pseudomallei*, lipopolysaccharide II, lose the natural resistance to complement-mediated bacteriolysis of the species and are 10 to 100 times less virulent in animal models than their parent strains. Other characteristics which may contribute to the pathogenicity of *B. pseudomallei* include: the ability to enter and survive in eukaryotic cells; the secretion of various extracellular enzymes (e.g. protease, lecithinase, and lipase); peptide, protein, and glycolipid toxins; extracellular polysaccharide; pili; a siderophore (malleobactin); and acid phosphatase.

Clinical features

B. pseudomallei may cause acute, chronic, localized, or disseminated infections. A 'flu-like' illness associated with seroconversion has been reported from Australia. Latent infections (see above) usually relapse at times of intercurrent stress ('Vietnam time bomb').

Septicaemic melioidosis

Sixty per cent of cases of culture-positive melioidosis have positive blood cultures. Most are clinically septicaemic; some have a more typhoid-like presentation. There is usually a short history (median 6 days; range 1 day to 2 months) of high fever and rigors. Approximately half the patients have evidence of a primary focus of infection, usually pulmonary or cutaneous. Confusion and stupor, jaundice, and diarrhoea may be prominent features. Initial investigations typically reveal anaemia,

neutrophil leucocytosis, coagulopathy, and evidence of renal and hepatic impairment. Such patients often deteriorate rapidly, developing widespread metastatic abscesses, particularly in the lungs, liver, and spleen, and metabolic acidosis with Kussmaul's breathing. Once septic shock has supervened, case fatality approaches 95 per cent, many patients dying within 48 h of admission. Other poor prognostic features include absence of fever, leucopenia, azotaemia, and abnormal liver function tests.

If the patient survives this acute phase, multiple foci of dissemination become prominent. Cutaneous pustules or subcutaneous abscesses occur in approximately 10 per cent of cases and an abnormal chest radiograph is found in 80 per cent of patients, the most common pattern being widespread nodular shadowing ('bloodborne pneumonia'; Fig. 1). Other common sites for secondary lesions include the liver, spleen, kidneys, prostate, bones, and joints. Involvement of the central nervous system may also occur.

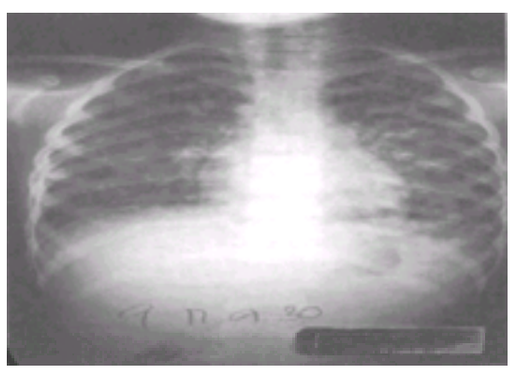

Fig. 1 Septicaemic melioidosis: widespread nodular shadowing—'bloodborne pneumonia' (by courtesy of Professor Sornchai Looareesuwan).

Localized melioidosis

The lung is the most frequent site. There is subacute, cavitating pneumonia accompanied by profound weight loss, often confused with tuberculosis (Fig. 2). Relative sparing of the apexes and infrequent hilar adenopathy may help to distinguish melioidosis from tuberculosis. There is a predilection for the upper lobes. Complications include pneumothorax, empyema, purulent pericarditis, and progression to septicaemia.

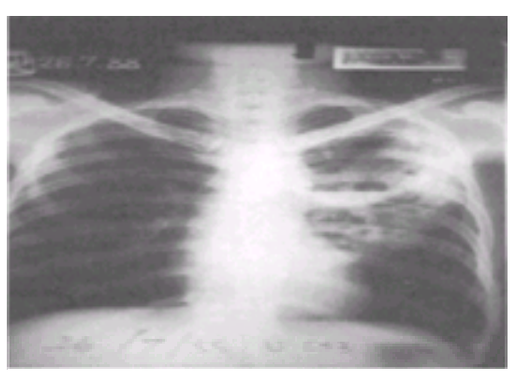

Fig. 2 Necrotizing *B. pseudomallei* pneumonia with central cavitation and fluid level in a rice farmer in north-east Thailand being treated with corticosteroids for nephrotic syndrome. Such patients are often misdiagnosed as having smear-negative tuberculosis, but fail to respond to antituberculous chemotherapy. (By courtesy of Professor Sornchai Looareesuwan.)

Acute suppurative parotitis is a characteristic manifestation of melioidosis in children, accounting for approximately one-third of childhood cases in north-east Thailand (Fig. 3). The reason is unknown. Most cases are unilateral and result in parotid abscesses requiring surgical drainage. They may rupture spontaneously into the auditory canal. Facial nerve palsy and septicaemia are rare complications.

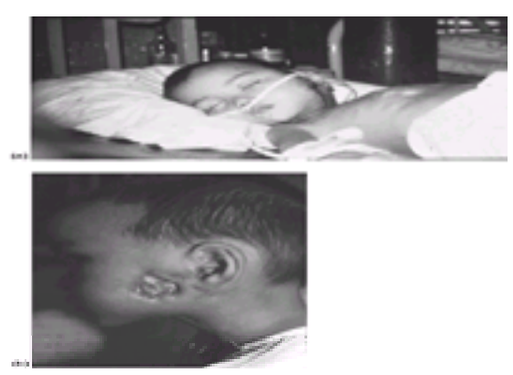

Fig. 3 (a) Acute suppurative parotitis—a common manifestation of childhood melioidosis in north-east Thailand. This child had parotid abscesses that required drainage despite having already ruptured into the auditory canal, extensive overlying ulceration, facial nerve palsy, and septicaemia. (b) Ulceration over healing parotid abscess.

Other sites of localized infection include cutaneous and subcutaneous abscesses, lymphadenitis, osteomyelitis and septic arthritis, liver and/or splenic abscesses, cystitis, pyelonephritis, prostatic abscesses, epididymo-orchitis, keratitis, and rarely, brain abscesses.

Pathology

B. pseudomallei is pyogenic, causing localized abscesses or granulomas, depending on the duration of the lesion. The presence of 'globi' of Gram-negative bacilli within macrophages and giant cells may give a clue to the aetiology.

Laboratory diagnosis

The diagnosis should be considered in any patient who has ever visited an endemic area and presents with septicaemia, abscesses, or chronic suppuration, particularly if there is evidence of an underlying disease such as diabetes mellitus. Specific diagnosis depends on the detection of *B. pseudomallei* or of corresponding antibodies. The laboratory should always be warned if melioidosis is suspected, both to enable appropriate methods and media to be employed, and to alert staff to the risk of infection (containment level 3 organism).

P>**Microscopy and culture**

Gram staining of smears of pus or secretions may reveal bipolar or unevenly staining Gram-negative rods, but this is neither specific nor sensitive. The most useful rapid diagnostic technique is immunofluorescent microscopy of smears. The mainstay of diagnosis is isolation and identification of *B. pseudomallei* from blood, pus, urine, sputum, or other specimens. The organism is not difficult to grow, although special selective media will increase the isolation rate from sites with a normal flora.

Serodiagnosis

Several tests for antibodies to *B. pseudomallei* have been described. The indirect haemagglutination test remains the most widely available. Assays that detect IgG give similar results. These tests are useful in patients from non-endemic areas in whom a single indirect haemagglutination titre in excess of 1:40 is highly suggestive of melioidosis. In populations continually exposed to *B. pseudomallei*, the high background seropositivity reduces the predictive value of the tests, and in such patients only a rising or very high titre suggests active melioidosis. Assays that detect specific IgM correlate better with disease activity but are not widely available.

Antigen and nucleic acid detection

Numerous antigen detection and polymerase chain reaction systems have been described, but all have problems of specificity and sensitivity.

Treatment

Treatment—general

Patients with septicaemic melioidosis require intensive supportive treatment, ideally in an intensive care unit. Particular attention should be paid to correction of volume depletion and septic shock, respiratory and renal failure, and hyperglycaemia or ketoacidosis. Abscesses should be drained whenever possible.

Antibiotic susceptibility

B. pseudomallei is intrinsically resistant to many antibiotics, including aminoglycosides and early b-lactams. Failure to respond to these agents may suggest the diagnosis of melioidosis.

Antibiotic therapy–acute phase

Five, randomized, controlled studies of regimens for the treatment of acute severe melioidosis have now been published. Two showed that regimens containing b-lactam approximately halved the mortality compared with conventional chloramphenicol plus doxycycline plus co-trimoxazole. Imipenem/cilastatin had a lower treatment failure rate than ceftazidime, which itself had a lower failure rate than co-amoxiclav. b-Lactam/co-trimoxazole combinations may have lower mortality and relapse rates than b-lactams alone, but larger studies are needed to confirm this. Unfortunately, all these b-lactams are too expensive to be practical in most endemic countries. The regimens used for acute treatment are listed in Table 1.

Antibiotic therapy—maintenance phase or mild disease

Long courses of oral antibiotics are needed to prevent relapse. Less than 12 weeks of treatment is inadequate, and the usual recommendation is 20 weeks. The conventional combination regimen was associated with a lower relapse rate than co-amoxiclav. The latter is preferable in children and pregnant women because of the risks of toxicity. Fluoroquinolones and doxycycline alone give unacceptable results, but co-trimoxazole alone warrants further evaluation. Regimens are given in Table 2.

Prognosis

In septicaemic melioidosis, the level of bacteraemia correlates with outcome. Even with optimal treatment, case fatality from acute severe melioidosis is high (25 to 40 per cent). Often survivors remain chronically ill both from the disease itself and the underlying conditions. At least 5 per cent of patients will still relapse despite long courses of antibiotics, particularly if compliance is poor. Antibiotic resistance may develop during treatment. Long-term follow-up should therefore be arranged. Monitoring of IgM titres or C-reactive protein may help early detection of relapse.

Prevention and control

No vaccines are available for human use. The only preventive measure is avoidance of exposure to the organism in soil for high-risk groups (e.g. people with diabetes). Both ciprofloxacin and doxycycline confer partial protection when given prophylactically to animals. Although person-to-person spread is rare, isolation of patients is recommended.

Glanders

Definition and aetiology

Glanders is a disease of horses caused by *Burkholderia mallei*, which may occasionally be transmitted to humans or other animals. Traditionally, glanders, a systemic respiratory tract disease, has been distinguished from farcy, a cutaneous infection.

In the early 1900s, equine glanders occurred worldwide. Over 200 000 horses were destroyed because of glanders during the First World War. However, no naturally acquired case has been reported in the United States or the United Kingdom since 1938.

It is thought still to occur in the Middle East, Africa, and Asia. Human infection, always uncommon, is confined to those in close contact with horses.

B. mallei is closely related to *B. pseudomallei* both taxonomically and antigenically, but it grows less luxuriantly in culture and is non-motile.

Clinical features

Glanders resembles melioidosis. Manifestations include septicaemia, wound infection, ulceration, lymphangitis with abscesses along the course of lymphatic drainage ('farcy buds'), ulceration of the respiratory (especially nasal) mucosa, polyarthritis, pneumonia and lung abscesses, nodular abscesses in any site, particularly muscle and subcutaneous tissue, and a widespread pustular rash.

Laboratory diagnosis

Diagnosis hinges on a history of contact with horses in an endemic area or laboratory exposure, and either the isolation of *B. mallei* or detection of specific antibodies. Like *B. pseudomallei* it requires handling in a containment level 3 laboratory.

Treatment

In vitro susceptibility is similar to that of *B. pseudomallei*, and so glanders should respond to the regimens used for melioidosis.

Further reading

Melioidosis

Chaowagul W *et al.* (1999). A comparison of chloramphenicol, trimethoprim–sulfamethoxazole, and doxycycline with doxycycline alone as maintenance therapy for melioidosis. *Clinical Infectious Diseases* **29**, 375–80.

Chong VFH, Fan YF (1996). The radiology of melioidosis. *Australasian Radiology* **40**, 244–9.

Dance DAB (1991). Melioidosis: the tip of the iceberg? *Clinical Microbiology Reviews* **4,** 52–60.

Simpson AJ *et al.* (1999). Comparison of imipenem and ceftazidime as therapy for severe melioidosis. *Clinical Infectious Diseases* **29**, 381–7.

Woods DE *et al.* (1999). Current studies on the pathogenesis of melioidosis. *Microbes and Infection* **2**, 157–62.

Glanders

Howe C. (1950). Glanders. In: *Oxford system of medicine*, Vol. 5, pp. 185–202. Oxford University Press.

7.11.16

Plague

T. Butler

History

Plague may have caused more deaths than most other diseases and warfare combined; it was estimated to have killed a quarter of Europe's population in the Middle Ages. The present pandemic of plague began in China in the 1860s and was spread by rats transported on ships to California and to ports in South America, Africa, and Asia. The genus of the plague bacillus is called *Yersinia* because Alexandre Yersin (1863 to 1943) went to Hong Kong in 1894 and successfully isolated the causative organism in pure culture. Urban plague transmitted by rats was brought under control in most affected cities, but the infection was transferred to sylvatic rodents, allowing it to become entrenched in rural areas of these countries. In the 1960s and 1970s, Vietnam during its war became the leading country for plague, reporting more than 10 000 cases a year. In 1994 in Surat, India, an outbreak of primary pneumonic plague was reported. There were hundreds of suspected cases, with 50 deaths, and thousands fled the city. However, none of the cases were confirmed by sputum culture. In 1997, Madagascar experienced an epidemic of pneumonic plague in which 8 out of 18 infected persons died.

Bacteriology

Yersinia pestis (formerly *Pasteurella pestis*), the cause of plague, is an aerobic, Gram-negative bacillus of the family Enterobacteriaceae. It is readily identified by its failure to ferment lactose on MacConkey agar, an alkaline slant and acid butt in triple-sugar–iron agar, and negative reactions for citrate utilization, urease, and indole. *Y. pestis* is virulent because it carries a 45 MDa plasmid that encodes for V and W antigens, which confer a requirement for calcium to grow at 37°C. Additionally, it produces lipopolysaccharide endotoxin and a capsular envelope containing the antiphagocytic principle fraction I antigen.

Epidemiology

From 1990 to 1996, there were 16 000 cases of plague and 1214 deaths (7.6 per cent) reported to the World Health Organization. The countries that reported more than 100 cases were, in the order from greatest number to least: Tanzania, Madagascar, Vietnam, Congo, Peru, India, Myanmar, Zimbabwe, China, and Uganda. In the United States, all the 64 plague cases occurred in the south-western states of New Mexico, Arizona, Colorado, Utah, and California. Most of the American cases occur during the months of May to October, when people are outdoors coming into contact with rodents and their fleas. Each endemic region has a specific season when plague tends to occur.

Plague is a zoonotic infection transmitted among animal reservoirs by flea bites and ingestion of animal tissues. The major animal reservoirs are urban and domestic rats as well as rural field rodents including ground squirrels and prairie dogs. The oriental rat flea *Xenopsylla cheopis* is the most efficient vector. When bitten by a rodent flea humans become an accidental host and play no role in disease transmission except in rare epidemics of pneumonic plague. Epizootics usually accompany human cases and can cause large die-offs of susceptible rodent species. Human plague affects both sexes and children of all ages depending on their exposure to rodent fleas. Risk factors for acquiring plague include contact with rodents or carnivores and presence of refuges or food sources for wild rodents near the home.

Pathogenesis

Bacteria are inoculated into the skin by a flea bite and migrate to regional lymph nodes, where they multiply during an incubation period of 2 to 8 days. Inflamed lymph nodes called buboes show polymorphonuclear leucocytes, destruction of normal architecture, haemorrhagic necrosis, and dense concentrations of extracellular plague bacilli. Bacteraemia occurs and results in purulent, necrotic, and haemorrhagic lesions in many organs.

Clinical manifestations

Bubonic plague

The most common presentation is acute lymphadenitis called bubonic plague (Table 1). The people of plague endemic regions know the disease and have local names, such as *dich hach* in Vietnamese, that conjure up the horror of recalled fatalities during previous seasons. Patients are affected by the sudden onset of fever, chills, weakness, and headache. Usually, at the same time, after a few hours, or on the next day, they notice the bubo, which is signalled by intense pain in one anatomical region of lymph nodes, usually the groin, axilla, or neck. A swelling evolves in this area, which is so tender that the patients typically avoid any motion that might provoke discomfort. For example, if the bubo is in the femoral area, the patient will characteristically flex, abduct, and externally rotate the hip to relieve pressure on the area and will walk with a limp. When the bubo is in an axilla, the patient will abduct the shoulder or hold the arm in a splint. When a bubo is in the neck, patients will tilt their head to the opposite side.

The buboes are oval swellings that vary from 1 to 10 cm in length and elevate the overlying skin, which may appear stretched or erythematous. They may appear either as a smooth, uniform, ovoid mass or as an irregular cluster of several nodes with intervening and surrounding oedema (Fig. 1). There is warmth of the overlying skin and an underlying tender, firm, non-fluctuant mass. Occasionally, there is a large area of oedema extending from the bubo into the region drained by the affected lymph nodes. Although infections other than plague can produce acute lymphadenitis, plague is virtually unique for the suddeness of onset of the disease and fulminant clinical course that can produce death in 2 to 4 days after the onset of symptoms. The bubo of plague is also distinctive for the usual absence of a detectable skin lesion or ascending lymphangitis in its anatomical region.

Fig. 1 A right femoral bubo consists of an enlarged, tender lymph node with surrounding oedema.

The patients are typically prostrate and lethargic, and often exhibit restlessness or agitation. Occasionally, they are delirious with high fever, and seizures are common in children. Temperature is usually elevated in the range 38.5 to 40.0°C, and the pulse rate is increased to 110 to 140/min. Blood pressure is characteristically low, around 100/60 mmHg, and may be unobtainable if shock ensues. The liver and spleen are often palpable and tender.

About one-quarter of patients in Vietnam showed varied skin lesions including pustules, vesicles, eshars, or papules in the anatomical region that is lymphatically drained by the affected lymph nodes, and they presumably represent sites of the flea bites (Fig. 2). Purpuric lesions may develop and become necrotic, resulting in gangrene of distal extremities, the probable basis of the epithet 'Black Death' attributed to plague through the ages.

Fig. 2 A right axillary bubo was accompanied by a purulent ulcer on the abdomen, which was the presumed site of the flea bite.

Other plague syndromes

Less common presentations may accompany the bubo or occur without a bubo. Septicaemic plague refers to bacteremia without a bubo. Pneumonic plague occurs as a secondary pneumonia due to bacteremic spread in about 10 per cent of patients with bubonic plague. Person-to-person spread of pneumonia by a coughing patient is less common, and a few cases of inhalation pneumonia have occurred in persons who handled sick cats. Bacterial meningitis is a rare complication of plague. Acute pharyngitis may occur.

Laboratory findings

The white blood-cell count is typically elevated in the range of 10 000 to 20 000 cells/mm ^{3}, with a predominance of immature and mature neutrophils. Occasionally, some patients, especially children, may develop myelocytic leukaemoid reactions with white cell counts as high as 100 000/mm ^{3}. Blood platelets may be normal or low in the early stages of bubonic plague. Although patients with plague rarely develop a generalized bleeding tendency from profound thrombocytopenia, disseminated intravascular coagulation is common in this infection. Liver function tests, including serum aminotransferases and bilirubin, are frequently abnormally high. Renal function tests may be abnormal in hypotensive patients.

Diagnosis

Plague should be suspected in febrile patients who have been exposed to rodents or other mammals in the known endemic areas of the world. A bacteriological diagnosis is readily made by Gram stain and culture of a bubo aspirate. The aspirate is obtained by inserting a 20-gauge needle on a 10-ml syringe containing 1 ml of sterile saline into the bubo and withdrawing it several times until the saline becomes blood tinged. Because the bubo does not contain liquid pus, it may be necessary to inject some of the saline and immediately reaspirate it. The Gram stain will reveal polymorphonuclear leucocytes and Gram-negative coccobacilli and bacilli ranging from 1 to 2 µm in length (Fig. 3). Smears of blood, sputum, or spinal fluid can be handled similarly (Fig. 4).

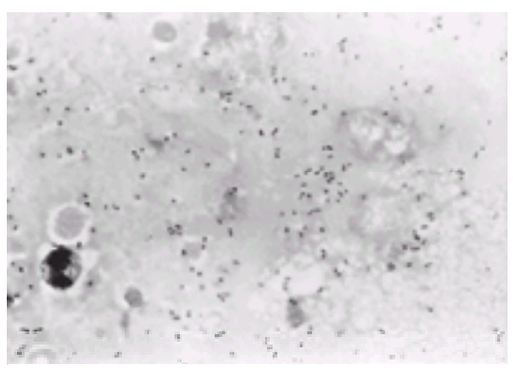

Fig. 3 Bubo aspirate shows bipolar bacilli stained with methylene blue (Wayson's stain).

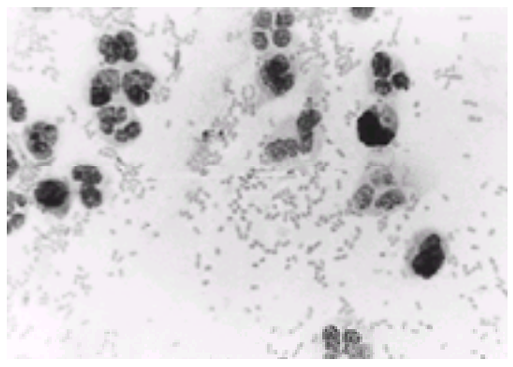

Fig. 4 Gram stain of spinal fluid in plague meningitis shows numerous Gram-negative bacilli.

The aspirate, blood, and other appropriate fluids should be inoculated on to blood and MacConkey agar plates and into infusion broth for bacteriological identification. At some reference laboratories, a serological test, the passive hemagglutination test or an ELISA utilizing fraction I of *Y. pestis*, is available for testing acute- and convalescent-phase serum. A fourfold or greater increase in titre or a single titre of 1:16 or higher is presumptive evidence of plague.

Treatment and prevention

Antimicrobials

Untreated plague has an estimated mortality rate of more than 50 per cent. Therefore, the early institution of effective antimicrobial therapy is mandatory following appropriate cultures. In 1948, streptomycin was identified as the drug of choice for the treatment of plague by reducing the mortality rate to less than 5 per cent. Streptomycin should be given intramuscularly in two divided doses daily, totalling 30 mg/kg body weight per day for 10 days. Most patients improve rapidly and become afebrile in about 3 days. The 10-day course of streptomycin is recommended to prevent relapses because viable bacteria have been isolated from buboes of patients with plague during convalescence.

When an oral drug is preferred, tetracycline is a satisfactory alternative. It is given orally in a dose of 2 to 4 g/day in four divided doses for 10 days. Tetracycline is contraindicated in children younger than 7 years of age and in pregnant women because it stains developing teeth. It is also contraindicated in renal failure. As an alternative drug that is especially suitable for meningitis, chloramphenicol can be given intravenously as a loading dose of 25 mg/kg of body weight followed by 60 mg/kg of body weight per day in four divided doses. After clinical improvement, chloramphenicol can be continued orally in a dose of 30 mg/kg to complete a total course of 10 days. There is no rationale for using multiple antibiotics to treat plague.

Other antimicrobial drugs have been used in plague or in experimental animal infections with varying success. These include sulphonamides, trimethoprim–sulphamethoxazole, kanamycin, gentamicin, ampicillin, cephalosporins, and fluoroquinolones. These drugs either are less effective than streptomycin or have not been subjected to adequate clinical studies and, therefore, should not be routinely chosen. An isolate from a 16-year-old boy in Madagascar in 1995 was resistant to streptomycin, tetracycline, chloramphenicol, and sulphonamide but was susceptible to trimethoprim–sulphamethoxazole. He recovered after receiving trimethoprim–sulphamethoxazole. Other than this case, antibiotic resistance in *Y. pestis* from humans has never been reported, nor has resistance emerged during antibiotic therapy.

Supportive therapy

Intravenous 0.9 per cent saline solution should be given to most patients for the first few days of the illness or until improvement occurs. Patients in shock will require additional quantities of fluid, with haemodynamic monitoring and use of vasopressors. The buboes usually recede without local therapy. Occasionally, however, they may enlarge or become fluctuant during the first week of treatment, requiring incision and drainage.

Precautions and prevention

Patients with plague who are promptly treated present no health hazard to other people. Those with a cough or other signs of pneumonia must be placed in respiratory isolation for at least 48 h after starting therapy or until the sputum culture is negative. The bubo aspirate and blood must be handled with gloves and with care to avoid aerosolization. Vaccines have been developed but at present are not available. Health departments advise personal protection against rodents and fleas, including living in rat-proof houses, wearing shoes and garments to cover the legs, and dusting houses with insecticide. For persons who report close contact with a coughing patient, prophylaxis with oral doxycycline or trimethoprim–sulfa-methoxazole is advised.

Further reading

Butler T (1994). *Yersinia infections*: Centennial of the discovery of the plague bacillus. *Clinical Infectious Diseases* **19**, 655–63.

Byrne WR *et al*. (1998). Antibiotic treatment of experimental pneumonic plague in mice. *Antimicrobial Agents and Chemotherapy* **42**, 675–81.

Campbell GL, Hughes JM (1995). Plague in India: a new warning from an old nemesis. *Annals of Internal Medicine* **122**, 151–3.

Chanteau S *et al*. (1998). F1 antigenaemia in bubonic plague patients, a marker of gravity and efficacy of therapy. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92**, 572–3.

Crook LD, Tempest B (1992). Plague. A clinical review of 27 cases. *Archives of Internal Medicine* **152**, 1253–6.

Galimand M *et al*. (1997). Multidrug resistance in *Yersinia pestis* mediated by a transferable plasmid. *New England Journal of Medicine* **337**, 677–80.

Ratsitorahina M *et al*. (2000). Epidemiological and diagnostic aspects of the outbreak of pneumonic plague in Madagascar. *Lancet* **355**, 111–13.

7.11.17 Yersinia, Pasteurella, and Francisella

*David Lalloo**

Yersiniosis

Definition

Yersiniosis is a disease caused by two species of enteric bacteria, *Yersinia enterocolitica* and *Yersinia pseudotuberculosis.* They cause a wide spectrum of clinical manifestations, which includes acute watery diarrhoea, acute mesenteric adenitis, extraintestinal infection, and bacteraemia. Postinfectious sequelae such as arthritis or erythema nodosum are also common.

The organisms

Yersinia sp. belong to the family Enterobacteriaceae. They are aerobic facultative and anaerobic, Gram-negative coccobacilli which grow on bile-containing media. There are three human pathogens within the genus, *Y. pseudotuberculosis*, *Y. enterocolitica*, and *Y. pestis*. The last, the causative organism of plague, is considered elsewhere (Chaper 7.11.16). Yersinia are usually non-lactose fermenting, catalase positive, and oxidase negative. The recovery of yersinia from stool samples can be improved by use of cefsulodin–irgasan–novobiocin (**CIN**) agar, cold enrichment, or potassium hydroxide pretreatment, but such manoeuvres are only occasionally necessary for clinical specimens. *Y. enterocolitica* can be divided into five biovars on the basis of biochemical reactions. The clinical and epidemiological significance of these biovars remains uncertain. Over 50 serogroups of *Y. enterocolitica* have been described, on the basis of O (somatic lipopolysaccharide) and H (flagellar) antigens. Six serotypes of *Y. pseudotuberculosis* have been described.

Virulence and pathogenicity

There are clear differences in virulence between strains of *Y. enterocolitica*; human disease is caused by a limited number of serotypes. A number of important factors have been identified. The possession of a 40 to 50 MDa plasmid is associated with virulence *in vitro* and the VW antigen complex and plasmid-encoded yersinia outer membrane proteins (YOPs) appear to be important in pathogenesis. The ability of yersinia to utilize exogenous iron by a number of mechanisms, including binding of exogenous siderophores, also appears to be an important factor in virulence; serious yersinia infections are much more common in patients with iron-overload syndromes. The vast majority of isolates produce enterotoxin, which is related to the enterotoxin of *Escherichia coli*. However, the role of enterotoxin in the production of diarrhoea remains uncertain and enterotoxin production does not correlate with other tests of virulence.

Most yersinia infections result from invasion via the gastrointestinal tract. Organisms adhere to the surface of the ileum and may invade the intestinal mucosa, via a bacterial outer membrane protein, invasin, which binds to a ligand on the cell surface. Bacteria multiply within intestinal epithelial cells and may reach Peyer's patches, where further multiplication occurs, with the potential for systemic spread.

Epidemiology

Y. enterocolitica causes infection throughout the world, but appears most common in the temperate regions, particularly northern Europe and North America. Both sporadic infections and outbreaks occur. Infection with serotypes 03 and 09 predominate in Europe whereas 08 and 03 are more commonly responsible for infection in North America. Yersiniosis is a zoonotic infection but usually causes foodborne illnesses. Animal reservoirs of *Y. enterocolitica* include pigs, rabbits, goats, cattle, horses, rodents, dogs, and cats. Animals may carry the organism asymptomatically in the oropharynx or gastrointestinal tract. The most important source of infection for man is the pig, although contact with household pets has also been implicated. Humans are infected via the faecal–oral route, usually after eating or drinking contaminated food or water; incompletely cooked pork is a major risk factor. Infection may also occur by person-to-person or direct animal-to-person contact and transmission through contaminated blood products has been reported. Infants and young children appear to be more susceptible to infection with *Y. enterocolitica* than adults. Most infections are sporadic but a number of specific outbreaks have been identified, following ingestion of contaminated foods such as pork chitterlings (intestines), water, or dairy products.

Infection with *Y. pseudotuberculosis* is less common, although cases are increasingly reported from Japan. Infection results from contact with both sylvatic and domestic animals and a number of birds. It most commonly affects patients aged between 5 and 20 years.

Clinical features

The usual incubation period for the acute manifestations of yersiniosis is 3 to 7 days. Common clinical syndromes are shown in Table 1. The most common manifestation of infection with *Y. enterocolitica* is an acute gastroenteritis, which particularly affects young children. Diarrhoea, fever, and abdominal pain may all be prominent. Stools contain mucus, leucocytes, and red blood cells; the organism can usually be detected on stool cultures. Clinically, the syndrome is indistinguishable from salmonella or campylobacter infection. Symptoms may last for up to 3 weeks and patients remain infectious over this period with continuous shedding of organism in the faeces. Rare complications include diffuse ulceration of the small intestine and colon, perforation, intussusception, toxic megacolon, cholangitis, and mesenteric vein thrombosis.

Older children more often develop mesenteric adenitis and terminal ileitis with either *Y. enterocolitica* or *Y. pseudotuberculosis* infection; this is the most common manifestation of *Y. pseudotuberculosis* infection. The presentation mimics appendicitis with fever, abdominal pain, right lower quadrant pain, and leucocytosis. Diarrhoea is unusual. Ultrasound and/or computed tomography may be helpful in demonstrating a normal appendix and enlarged mesenteric nodes. The infection is

usually self-limited. At laparotomy, enlarged mesenteric lymph nodes are found in the iliocaecal angle and there may be swelling of the terminal ileum or caecum. This presentation must be distinguished from acute appendicitis, or diseases causing terminal ileal disease such as Crohn's, tuberculosis, and rarely, neoplasia.

Y. enterocolitica may cause focal infection both in the absence of detectable bacteraemia and following bacteraemia. Isolated focal infection has been described in many sites, including the pharynx, skin and subcutaneous tissues, bones and joints, the conjunctiva, the renal tract, lungs, and peritoneum. *Y. enterocolitica* bacteraemia most often occurs in patients with chronic conditions such as diabetes, chronic liver disease, malignancy, and conditions causing immunosuppression. There is also a strong association with iron-overload syndromes or the treatment of iron overload (*Y. enterocolitica* is able to use exogenous iron chelators such as desferrioxamine to acquire iron itself). Over half of systemic bacteraemias are in patients with iron-overload syndromes; multiple hepatic abscesses may occur and the case fatality rate may reach 50 per cent in this population. Overall, the case fatality rates for *Y. enterocolitica* bacteraemia have ranged from 7.5 to 25 per cent over the last decade. Bacteraemia may lead to metastatic infections including endocarditis, intravenous line infection, meningitis, and septic arthritis.

Y. pseudotuberculosis bacteraemia is much less common, but is often associated with chronic illness. Case fatality rates are extremely high in the immunocompromised population. In Japan, *Y. pseudotuberculosis* infection has been associated with renal failure in young children.

Secondary, postinfective, complications are common following yersinia infection. In Scandinavia, they have been reported in up to 30 per cent of patients with *Y. enterocolitica* infection. A reactive polyarthropathy or erythema nodosum are the most common manifestations, classically occurring 1 to 2 weeks after an acute illness. Reiter's syndrome, glomeru-lonephritis, and myocarditis have also been described. The arthritis is polyarticular and asymmetrical, typically affecting the large joints of the lower limbs. There is a strong association with the possession of HLA B27. Synovial fluid culture is normally sterile although yersinia antigens can be found in the synovial tissue of patients. Symptoms of reactive polyarthritis may take several months to settle. The exact immunological mechanism of these postinfectious manifestations remains uncertain.

Diagnosis

Yersinia infection should be considered in anyone with fever and abdominal pain. A definitive diagnosis of yersiniosis may be made by culture of the organism from stool, lymph nodes, or blood depending upon the clinical presentation. However, isolation from stool may sometimes be slow because of the overgrowth of other faecal flora. Cold enrichment or CIN media may be used to optimize recovery from faecal samples.

A number of serological techniques, including tube agglutination assay, radioimmunoassays, and enzyme immunoassays, have been used to diagnose infection with yersinia. High titres in a previously healthy individual are suggestive of infection, but fourfold rises in titre are rarely found. Interpretation may be made difficult by cross-reactivity with *Brucella*, *Rickettsia*, and *Salmonella* spp. and possibly thyroid tissue antigens; some populations also have a high background prevalence of positive serology. Negative or minimal titres can occur following yersiniosis in infants or immunocompromised patients. Definitive diagnosis therefore depends upon the culture of the organism. However, serology is often the only way of diagnosing postinfectious complications as stool cultures may be negative by the time of appearance of symptoms such as arthritis. *Y. pseudotuberculosis* may be found in sterile site samples, but is rarely isolated from stool. Serology is often the only mode of diagnosis available; antigens cross-react with those of *Y. enterocolitica*.

Treatment

Antimicrobial therapy is not indicated in uncomplicated disease and treatment does not shorten the course or severity of enterocolitis. Localized infection, bacteraemia, and systemic disease, or enterocolitis in an immunocompromised patient, should be treated. *Y. enterocolitica* is resistant to most penicillins and first-generation cephalosporins due to the production of chromosomally encoded b-lactamases. Minimum inhibitory concentrations for amoxycillin/clavulanate combinations vary considerably; this drug should not be used for the treatment of infections. Aminoglycosides, chloramphenicol, tetracycline, and co-trimoxazole are all effective *in vitro* and have been used clinically with success. Third-generation cephalosporins are also effective, although in one recent study they were only successful in 85 per cent of cases, even when used in combination with other drugs. Fluoroquinolones have very good *in vitro* activity against yersinia and have been used successfully in clinical practice, but the optimal antibiotic therapy for the treatment of *Y. enterocolitica* has still to be determined. *Y. pseudotuberculosis* is sensitive to ampicillin and cephalosporins in addition to the drugs already discussed.

Pasteurella

Introduction

Pasteurella spp. are Gram-negative coccobacilli which cause a wide spectrum of disease in humans, ranging from local infection and abscesses to severe systemic infection. The majority of human infections are caused by *P. multocida*. They are most often acquired from contact with domestic animals.

The organism

Pasteurella spp. are small, non-motile, Gram-negative coccobacilli. They grow aerobically or as facultative anaerobes on standard media at 37°C; growth is enhanced by enrichment with carbon dioxide. They are oxidase and catalase positive and may stain bipolarly on Gram stain, sometimes being confused with *Haemophilus*, *Neisseria*, or *Acinetobacter* spp. *Pasteurella* spp. are a major veterinary pathogen, but only four species have been associated with human disease. The vast majority of human infections are caused by *P. multocida*, subspecies *septica* and subspecies *multocida*. Four capsular antigens and 15 somatic antigens of *P. multocida* have been identified. The capsule appears to be important in pathogenesis; heavily capsulate strains are resistant to phagocytosis.

Epidemiology

P. multocida is widely distributed as a nasopharyngeal or gastrointestinal commensal of animals and birds. The organism is carried by 70 to 90 per cent of cats and 50 to 70 per cent of dogs, but is also found in a large number of other domestic and wild animals. The organism can survive in water or soil for up to a month. *P. multocida* is a major animal pathogen, causing a number of different diseases, including fowl cholera and haemorrhagic septicaemia in wildstock.

Humans usually acquire infection with *P. multocida* from bites, scratches, or licks of dogs and cats, or close contact with these animals. However, in up to 15 per cent of cases, no known animal contact occurs. *Pasteurella* spp. can be isolated from 20 to 30 per cent of dog bite wounds and 50 per cent of cat bite wounds, although only a small proportion will become clinically infected. In clinically infected wounds, *Pasteurella* spp. are identified in 50 per cent of dog bites and 75 per cent of cat bites. Person-to-person spread has not been recorded, although *P. multocida* can occasionally be found in the nasopharynx of healthy humans exposed to animals.

Clinical features

P. multocida has been associated with a wide variety of different clinical presentations, outlined in Table 2. The vast majority of infections are due to animal bites which cause local soft tissue infection manifesting with remarkable rapidity. Symptoms and signs may develop within several hours of the bite. Local erythema, swelling, and purulent discharge are common; fever, lymphangitis, and local lymph node swelling may also occur. Soft tissue infections may also involve deeper tissues, causing abscesses, tenosynovitis, septic arthritis, or osteomyelitis.

The second most common site of isolation of *P. multocida* is the respiratory tract. Some of these patients have no history of animal contact, but over 90 per cent have chronic respiratory tract disease, particularly chronic obstructive pulmonary disease, bronchiectasis, or malignancy. Isolation of *Pasteurella* spp. may sometimes represent long-term colonization, but acute upper and lower respiratory tract infection does occur. Acute pneumonia, tracheobronchitis, empyema, and occasionally, lung abscess are the most commonly reported clinical syndromes. Most patients with pneumonia are elderly and bacteraemia occurs in 25 to 55 per cent of cases of respiratory infection with a reported case fatality rate of 29 per cent. There are no specific diagnostic features of *P. multocida* pneumonia; although lobar consolidation is the commonest chest radiograph appearance, multilobar and diffuse infiltrates also occur. Spread of Pasteurella infection from the upper respiratory tract may occasionally cause tonsillitis, sinusitis, pharyngitis, and epiglottitis.

Bacteraemia occurs in association with localized infections in many different sites. It is more common in patients with liver dysfunction. Bacteraemia is most often associated with meningitis (53 per cent of cases), respiratory disease, and septic arthritis (24 per cent). Endocarditis has been reported but appears to be relatively rare. Pasteurella meningitis affects mainly infants or the elderly. Septic arthritis normally affects already damaged joints. It is sometimes associated with bites distal to

the joint, but also occurs in patients with no trauma or even in some who have no pets. A number of different intra-abdominal infections have been reported; bacterial peritonitis is a particular problem in patients with liver disease.

Diagnosis

A history of animal exposure should always suggest the possibility of Pasteurella infection. *Pasteurella* spp. can be identified as small Gram-negative rods which may stain bipolarly and can be isolated from sputum, pus, blood, or cerebrospinal fluid. Differentiation from *Haemophilus*, *Acinetobacter*, and *Neisseria* spp. is important.

Prevention and treatment

The most important factor in avoiding Pasteurella infections is the adequate treatment of bites. Thorough cleaning and debridement of wounds is crucial. The role of prophylactic antibiotics is controversial; approximately 5 to 15 per cent of dog bites and up to 50 per cent of cat bites become infected. Most clinicians advocate prophylactic antibiotics for 'high-risk' bites, crush injuries, deep puncture wounds, and wounds to the hands. Patients who are immunosuppressed, who have asplenism, or have alcoholic liver disease should certainly be treated. One recent meta-analysis suggested that routine antibiotics reduced the incidence of infection with a number needed to treat (NNT) of 14 to prevent one infection. In view of the strong association of Pasteurella infection with domestic animal contact, some clinicians have suggested that patients who are immunocompromised or who have chronic disease such as cirrhosis should try to avoid contact with dogs or cats.

Penicillin is the treatment of choice for established infections, although occasional clinical isolates that produce b-lactamase have been reported. Oral agents with good *in vitro* activity include tetracyclines, amoxycillin, amoxycillin/clavulanate, co-trimoxazole, and most fluoroquinolones. Azithromycin appears to be the most effective macrolide. Erythromycin has poor activity and *P. multocida* is resistant to clindamycin and many first-generation cephalosporins. Penicillin, third-generation cephalosporins, particularly cefotaxime, and chloramphenicol have all been used successfully in severely ill patients admitted to hospital.

Although penicillin has good activity against *Pasteurella* spp., prophylactic antibiotics for bites need to cover other organisms commonly found in the oral flora of dogs and cats, for instance *Staphylococcus aureus*, other staphylococcal species, anaerobes, and *Capnocytophaga canimorsus*. Amoxycillin/clavulanate is the prophylactic drug of choice. Treatment options in penicillin-hypersensitive patients include tetracycline or a combination of clindamycin and a fluoroquinolone.

Tularaemia

Introduction

Tularaemia is a zoonotic, arthropod-, and water-borne disease caused by *Francisella tularensis*, a small Gram-negative bacterium that has a natural lifecycle in mammalian hosts and may be transmitted by ticks and biting flies. Human infections occur predominantly in the northern hemisphere causing several different clinical syndromes, ranging from the combination of a fever, cutaneous ulcer, and lymphadenopathy to severe systemic disease and pneumonia. Three biogroups of *F. tularensis* can be distinguished by their geographical distribution, epidemiology, and virulence.

Bacteriology

Francisella spp. are small, non-motile, pleomorphic, Gram-negative coccobacilli. The organism has a thin capsule and may stain only faintly or exhibit bipolar staining with Gram or Giemsa stains. *Francisella* spp. grow only on media that contains cystine or cysteine and are strict aerobes. Optimal growth occurs at 35°C with carbon dioxide enrichment. They are oxidase negative and weakly catalase positive. Several species exist within the genus, but the major human pathogen is *F. tularensis*. Several biogroups of *F. tularensis* can be distinguished by biochemical characteristics; these biogroups vary markedly in their geographical distribution, epidemiology, and virulence for humans. *F. tularensis* biogroup *tularensis* and *F. tularensis* biogroup *palearctica* cause the vast majority of human disease. *F. tularensis* biogroup *novicida* causes milder forms of tularaemia and *F. philmiragia* has occasionally been reported to cause infection in specific host groups.

Epidemiology

Although foci of disease occur throughout the world, the vast majority of cases have occurred in the northern hemisphere. The distribution of *F. tularensis* throughout Europe is shown in Fig. 1. A significant number of cases have also been reported from Japan. In North America, tularaemia has been reported from all states of the United States, Canada, and Mexico. However, the disease is now particularly associated with the Midwest in summer owing to transmission by tick bites and east of the Mississippi in winter from rabbit hunting. Transmission of the disease does not occur in the United Kingdom. Although tularaemia was extremely common in the United States, former Union of Soviet Socialist Republics, and Scandinavia in the middle of the last century, the recognition of occupational hazards and vaccination campaigns have reduced the incidence considerably. In the United States, the number of cases has declined from around 2000 per year in the post-war era to an average of 146 cases per year from 1990 to 1994, the last year for which tularaemia was a notifiable disease. Hundreds of thousands of cases occurred in the Union of Soviet Socialist Republics around the time of the Second World War, but the Russian Federation only reported 2019 cases in the 10 years from 1987 to 1997.

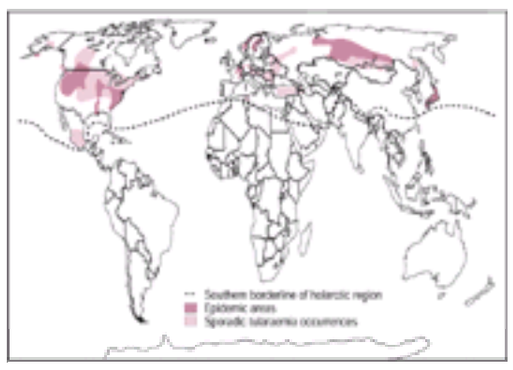

Fig. 1 Tularaemia foci in Europe.

Infection in humans is predominantly caused by two biogroups; *F. tularensis* biogroup *tularensis* (synonymous with type A or *nearctica*) and *F. tularensis* biogroup *palearctica* (synonymous with type B or *palearctica*). Ninety-five per cent of North American human infections are due to biogroup *tularensis.* In Europe, infection is by the less virulent biogroup *palearctica*, which has a greater variety of reservoirs and vectors than the North American biogroup.

In the United States, biogroup *tularensis* particularly infects ground squirrels, cottontail rabbits, hares, and jackrabbits, but also occurs in other wild and domestic animals. Human infections usually arise from contact with affected animals or following tick or fly bites. Hunting, skinning, or eating infected animals is a particular risk factor. Inoculation through the skin or airborne transmission may occur during preparation of carcasses and infection can also be acquired by ingestion of infected meat or contaminated water and through mammal bites. *F tularensis* biogroup *tularensis* may be transmitted by a wide range of arthropod vectors, including ticks of the genera *Dermacentor* and *Amblyomma*.

Biogroup *palearctica* is less common in the United States; strains of this type have been isolated from muskrats, in which they cause epizootics. In northern Europe, biogroup *palaearctica* infects a wide variety of mammals, mainly rodents, as well as hares. The main insect vectors are mosquitoes of the genus *Aëdes*, but ticks and biting flies also transmit the disease. Outbreaks of disease have often been reported. Human infections arise by a number of different mechanisms, apart from hunting and arthropod bites. Outbreaks have occurred from exposure to water contaminated by dead bodies or excreta of infected animals and from airborne dissemination of infection acquired by the inhalation of contaminated particles, such as dust from rodent-infested hay.

Pathogenicity

F. tularensis is primarily an intracellular pathogen that can multiply within mononuclear cells. Experimentally, as few as 50 organisms may cause infection through inoculation or inhalation; higher numbers are need to cause infection following ingestion. The organism spreads to regional lymph nodes from where it may be disseminated. Immunity is primarily cell mediated; focal necrosis and granulomas may be found in affected tissue. Little is known of the molecular basis of virulence, but loss of the capsule is associated with decreased virulence.

Clinical presentations of tularaemia (rabbit fever, deerfly fever, Ohara disease)

The clinical presentation of tularaemia depends on the route of transmission and virulence of the organism. Although clinical disease traditionally has been classified differently in the United States and Europe, manifestations of infection by *F. tularensis* biogroups *tularensis* and *palearctica* are essentially similar, except that biogroup *palaearctica* in Europe is clearly less virulent. The incubation period for tularaemia is usually between 3 and 5 days, but may be as long as 3 weeks. Most patients describe fever, chills, and prostration, and have a relapsing, protracted illness unless treated or vaccinated. A number of different discrete clinical syndromes (Table 3) have been described but there is considerable overlap between them.

Ulceroglandular tularaemia (Plate 1, Plate 2, Plate 3)

This is the most common presentation, particularly in North America. Patients present with sudden chills, fever, and often severe headache. An indurated and ragged ulcer evolves at the site of the initial entry of the organism; this is usually small and causes little pain. The ulcers are occasionally multiple and their site is related to the mode of contact. Local lymph nodes are tender, steadily increase in size, and frequently suppurate. If untreated, the ulcer may heal over several weeks leaving a scar.

Glandular tularaemia

Tularaemia may also present with lymphadenopathy without an obvious skin ulcer. Small lesions may be missed or have healed before presentation. Nodes may sometimes persist for weeks before the diagnosis is made. The differential diagnosis of glandular and ulceroglandular tularaemia includes pyogenic bacterial disease, cat-scratch disease, syphilis, mycobacterial infection, plague, and toxoplasmosis.

Oropharyngeal tularaemia (Plate 4)

This may occur due to oral contact with infected material and causes a pharyngitis, sometimes with ulceration, and enlargement of local lymph nodes. It may be more common in children than in adults. Pharyngitis and enlarged nodes may also occur in other forms of tularaemia. Pharyngeal tularaemia must be distinguished from other bacterial and viral causes of pharyngitis.

Oculoglandular tularaemia

This is a relatively rare form of tularaemia. The primary lesion is in the conjunctiva or cornea; infection usually occurs from splashing the face while cleaning infected animals, swimming in contaminated water, or from laboratory accidents. Conjunctivitis occurs along with chemosis and lid oedema. Unilateral preauricular lymphadenopathy is commonly observed. The differential diagnosis includes viral and bacterial causes of conjunctivitis including herpes simplex and syphilis.

Typhoidal tularaemia

This form of tularaemia occurs following any mode of acquisition of infection. Lymphadenopathy is not a feature. It appears to be more common in patients with pre-existing chronic illness. A febrile illness is associated with systemic symptoms which include fevers, chills, pharyngitis, myalgia, and gastrointestinal symptoms including watery diarrhoea. Meningism may occur. Patients may become severely ill and secondary pneumonic involvement occurs in over 40 per cent of patients.

Pulmonary tularaemia

Pulmonary disease may be primary, resulting from the inhalation of infected aerosols, or secondary from haematogenous spread. Pneumonia is commonly associated with typhoidal disease and also occurs in around a third of patients with ulceroglandular disease. Some patients have respiratory symptoms and signs of pneumonia, but a significant number of patients with tularaemia have asymptomatic radiological abnormalities. The commonest radiological finding is multiple parenchymal infiltrates in one lobe, but bilateral infiltrates and pleural effusions also occur. Hilar lymphadenopathy may be present. Examination of the sputum is rarely helpful. Most infiltrates clear rapidly on therapy. Tularaemia should be considered in any patient in an endemic area who presents with a community-acquired pneumonia which is resistant to standard therapy. The disease needs to be distinguished from all other causes of atypical pneumonia.

Secondary rashes have been reported following acute infection in all forms of tularaemia. Erythema nodosum has particularly been associated with pneumonic forms. *F. tularensis* biogroup *novicida* has low virulence for humans but may cause mild forms of tularaemia. *F. philmiragia* has been reported to cause severe infection in patients with chronic granulomatous disease or myeloproliferative disease and victims of near drowning.

Complications

Suppuration of lymph nodes is common in glandular forms, even after antibiotic treatment. Some patients may be unwell with malaise and fatigue for several months. In severe disease, impaired renal and hepatic function, elevated creatine kinase levels, and disseminated intravascular coagulation may occur. Severe disease is more common in patients with pre-existing illness and the elderly. Bacteraemia, renal impairment, pulmonary involvement, and elevated creatine kinase levels are all associated with a poorer prognosis. Overall case fatality rates for tularaemia in North America are approximately 2 to 3 per cent in patients treated with appropriate antibiotics. Deaths are rare in biogroup *palearctica* infections.

Diagnosis

In regions where *F. tularensis* is endemic, a provisional clinical diagnosis can often be made from the patient's exposure history and clinical signs. However, if there is no local ulcer, and if the patient has left the area in which the infection was acquired, patients who present with a persistent fever, lymphadenopathy, pneumonitis, or tonsillitis may be more difficult to diagnose. A detailed travel and epidemiological history may help. Routine laboratory investigations are rarely helpful, although sterile pyuria has been noted in up to 20 per cent of cases.

A definitive diagnosis can be made by isolation of the organism, but *F. tularensis* will not grow on routine plating and samples require inoculation on to supportive media. The organism can be isolated from blood, lymph nodes, wounds, and occasionally sputum, but even in optimum conditions, the organism grows slowly and prolonged culture may be needed. Clinicians should inform the laboratory if tularaemia is suspected; laboratory aerosols cause serious, occasionally fatal, laboratory-acquired infection. Swabs or aspirates from local lesions and lymph glands should be transported in approved containers. Immunofluorescence methods may be more sensitive for identification of organisms in smears and tissues, and reduce the risk to laboratory staff.

Traditionally, most diagnoses have been made serologically using various agglutination assays. A fourfold rise in titre, or a titre of 1:160 in a single sample by the end of the third week of illness, is considered to be diagnostic of tularaemia. Antibody may persist for years after infection. Cross-reactivity with *Brucella* and *Yersinia* antibodies occurs. Enzyme-linked immunosorbent assay (**ELISA**) techniques are more sensitive for the early detection of tularaemia and can detect class-specific immunoglobulins. Delayed hypersensitivity skin testing has been used for diagnosis, but antigens have not been standardized. The polymerase chain reaction (**PCR**) has been used for the detection of *F. tularensis* DNA in animal tissues and initial studies on clinical wound specimens suggest that the technique may be more sensitive than culture. *F. tularensis* DNA has also been detected in patients who are serologically negative but have other evidence of infection with *F. tularensis*. Further study is need to determine the role of PCR and improved ELISA techniques in the diagnosis of tularaemia.

Treatment

Aminoglycosides have been used most widely for the treatment of tularaemia because of their bactericidal activity against *F. tularensis.* Streptomycin is the drug of

choice and usually produces a dramatic clinical response. Gentamicin is an effective alternative, although relapses are more common than with streptomycin, occurring in 6 per cent of patients. Treatment is normally given for 10 to 14 days; shorter courses are associated with relapse. Tetracycline and chloramphenicol have been used successfully for the treatment of tularaemia, although relapse rates of 12 and 21 per cent, respectively, have been reported. Tetracyclines may be adequate for mild infections with biogroup *palaearctica* when given in high oral doses for 2 weeks.

Francisella spp. are resistant *in vitro* to most b-lactam antibiotics with the exception of third-generation cephalosporins. However, clinical experience with ceftriaxone has been disappointing. Quinolones have good *in vitro* activity against *F. tularensis* and successful clinical outcomes have been reported in a number of patients who have been treated with ciprofloxacin. Further study is required to define the role of these drugs in the treatment of tularaemia.

Prevention

Most important is reduction in human contact with infected animals and vectors in endemic areas. Protective clothing and insect repellent should be used when walking in endemic areas. Ticks should be sought out regularly and removed. Gloves should be worn when skinning or preparing rabbits; meat should be thoroughly cooked and sick animals should not be handled or eaten. Care should be taken in the laboratory to prevent transmission from potentially infective samples; *Francisella* spp. are category 3 pathogens and should be handled accordingly. Vaccines developed from live attenuated strains of biogroup *palearctica* are effective and should be considered for laboratory workers who regularly handle *F. tularensis* or for others with repeated occupational exposure. There is no evidence of efficacy of chemoprophylaxis for exposed individuals.

---

*This chapter is based on A. D. Pearson's account in the third edition of the *Oxford Textbook of Medicine* and the editors and author take pleasure in acknowledging his contribution.

Further reading

Adlam C, Rutter JM (1989). Pasteurella *and pasteurellosis.* Academic Press, London.

Cover TL, Aber RC (1989). *Yersinia enterocolitica. New England Journal of Medicine* **321**, 16–24.

Enderlin G *et al.* (1994). Streptomycin and alternative agents for the treatment of tularaemia. *Clinical Infectious Diseases* **19**, 42–7.

Evans ME *et al.* (1985). Tularaemia: a 30-year experience with 88 cases. *Medicine (Baltimore*) **64**, 251–69.

Gayraud M *et al.* (1993). Antibiotic treatment of *Yersinia enterocolitica* septicemia: a retrospective review of 43 cases. *Clinical Infectious Diseases* **17**, 405–10.

Gill V, Cunha BA (1997). Tularaemia pneumonia. *Seminars in Respiratory Infection* **12**, 61–7.

Koornhof HJ, Smego RA Jr, Nicol M (1999). Yersiniosis. II: The pathogenesis of *Yersinia* infections. *European Journal of Clinical Microbiology and Infectious Diseases* **18**, 87–112.

Larson JH (1979). The spectrum of clinical manifestations of infections with *Yersinia enterocolitica* and their pathogenesis. *Contributions to Microbiology and Immunology* **5**, 257–69.

Limaye AP, Hooper CJ (1999). Treatment of tularaemia with fluoroquinolones: two cases and review. *Clinical Infectious Diseases* **29**, 922–4.

Naktin J, Beavis KG (1999). *Yersinia enterocolitica* and *Yersinia pseudotuberculosis. Clinics in Laboratory Medicine* **19**, 523–36.

Smego RA, Frean J, Koornhof HJ (1999). Yersiniosis I: microbiological and clinicoepidemiological aspects of plague and non-plague *Yersinia* infections. *European Journal of Clinical Microbiology and Infectious Diseases* **18**, 1–15.

Talan DA *et al.* (1999). Bacteriologic analysis of infected dog and cat bites. *New England Journal of Medicine* **340**, 85–92.

Weber DJ *et al.* (1984). *Pasteurella multocida* infections. *Medicine* **63**, 133–56.

7.11.18

Anthrax

Thira Sirisanthana

Introduction

Anthrax is an acute bacterial infection caused by *Bacillus anthracis*. Herbivores are particularly susceptible to anthrax. They acquire the infection after coming into contact with soil-borne spores. Humans are infected when spores of *B. anthracis* enter the body through contact with infected animals or animal products, ingestion, or inhalation. The disease occurs in three clinical forms: cutaneous, gastrointestinal, and inhalation. Septicaemia and meningitis may occur from any of these primary foci. Other names for anthrax include malignant pustule, Siberian ulcer, charbon, malignant oedema, Milzbrand, and woolsorter's disease.

Anthrax has been known since antiquity. The fifth and sixth plagues described in the Bible are most likely outbreaks of anthrax in cattle and humans. Several distinguished scientists in the nineteenth century characterized the pathogenesis of the disease. In 1877, Robert Koch grew the organism in pure culture. He defined the stringent criteria needed to prove that the organism caused anthrax (Koch's postulates), then met them experimentally. Koch also discovered the spore stage that allows persistence of the organism in the soil. Louis Pasteur, in 1881, made a convincing field demonstration at Pouilly-le-Fort to show that vaccination of sheep, goats, and cows with heat-attenuated strain of *B. anthracis* prevented anthrax. In 1939, Sterne developed an animal vaccine that is a spore suspension of an avirulent, non-capsulated live strain. This is the animal vaccine still in use today.

Anthrax practically disappeared from North America, Western Europe, and Australia after the disease was eradicated in livestock following extensive vaccination programmes. However, it is still prevalent in developing countries, especially in Asia and Africa, where livestock are only poorly subjected to veterinary control, and where environmental conditions are favourable for an animal–soil–animal cycle to be established.

Recent interest in anthrax has been excited by fear of the use of anthrax spores as a biological weapon both in the battlefield and in a terrorist strike. An accident involving aerosolized anthrax spores at a Soviet military compound in 1979 (see below), and the revelation that Iraq had produced weapons containing anthrax spores during the 1991 Gulf War, confirm this fear.

Aetiology

B. anthracis is a large, non-motile, encapsulated, Gram-positive, aerobic, spore-forming bacillus that grows well in most nutrient media at 35°C. Spores are not produced in living animals. In the clinical laboratory, *B. anthracis* is recognized by its tendency to form very long chains of rods with elliptical central spores. The rectangular shape of the individual bacteria gives chains of *B. anthracis* a 'joint bamboo rod' appearance (Fig. 1 and Plate 1). On blood agar, *B. anthracis* forms non-haemolytic or weakly haemolytic greyish-white, rough colonies. In the presence of excess carbon dioxide, the organisms form capsules, and colonies are smooth and mucoid. The colonies produce a typical 'medusa head' or 'curled hair' appearance caused by chains of bacilli growing out from the edge of colonies. *B. anthracis* are pathogenic for small rodents. White mice, rabbits, and guinea pigs develop fatal infections after subcutaneous inoculation of very small numbers of the virulent organisms. The virulence factors of *B. anthracis* include a capsule that inhibits phagocytosis and three proteins collectively called anthrax toxin. The organism can be further identified by determination of susceptibility to bacillus phage g and by demonstration of species-specific antigens (including the capsule and the anthrax toxin). The spores of *B. anthracis* are very resistant and will resist dry heat at 140°C for 1 to 3 h or moist heat at 100°C for 5 min. They can persist in nature for many years. Boiling for 10 min, treatment with oxidizing agents such as potassium permanganate, or with formaldehyde will kill the spores. Most strains of *B. anthracis* are susceptible to penicillin. During growth inhibition by penicillin, the cell cylinders of *B. anthracis* tend to bulge, resulting in the classic 'string of pearls' reaction.

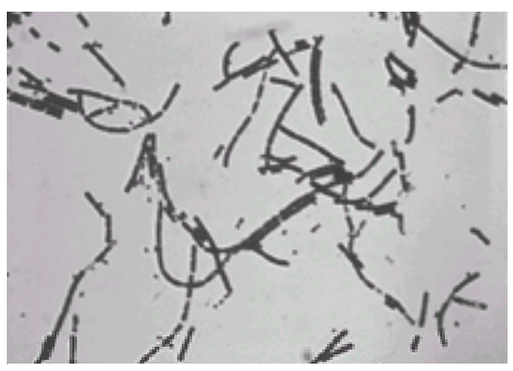

Fig. 1 Large Gram-positive bacilli in chains are typical of *B. anthracis*. An individual bacillus is 3 to 5 µm long and 1 to 1.25 µm wide with a flattened end. (See also Plate 1.)

Epidemiology

Anthrax is usually acquired through unrecognized breaks in skin or mucous membranes to which spores of *B. anthracis* gain access. The spores germinate to yield vegetative cells, which multiply and produce either localized or severe systemic infection depending on the animal species infected. Different species of animals are susceptible to varying degrees. Herbivores such as horses, sheep, goats, and cattle are most susceptible and develop fatal systemic disease. Dying herbivores have overwhelming bacteraemia and often bleed from the nose, mouth, and bowel, thereby contaminating soil with vegetative *B. anthracis*, which can sporulate and persist in the soil for a long time. The carcasses of infected animals provide additional potential foci of contamination. *B. anthracis* spores become part of the normal soil flora and can undergo bursts of local multiplication that increase the number of organisms in the soil. The cause of this local multiplication of anthrax bacilli is not known, but it is usually associated with major changes in the soil microenvironment such as those seen after abundant rainfall or a drought.

Human cases may occur in agriculture or industry. Agricultural cases result from direct contact with infected animals (herders, butchers, and slaughterhouse workers) and those who consume contaminated meat. Industrial cases involve those in contact with infected animal product such as hides, wool, goat's hair, or bone. All human cases are zoonotic in origin. No human-to-human transmission has been reported. The worldwide incidence of human anthrax is not known because many cases, especially in developing countries, do not receive medical attention and are not reported. However, one estimate puts the incidence at between 20 000 and 100 000 cases annually. Cutaneous anthrax typically follows skin exposure to infected animals or animal products. It is the most common form of the disease. Gastrointestinal anthrax follows ingestion of *B. anthracis*- contaminated food. Although rarely reported, gastrointestinal anthrax may not be uncommon in parts of Asia and Africa. Inhalation anthrax is a result of spore deposition in the lungs. Historically, woolsorters at industrial mills were at highest risk. Naturally occurring inhalation anthrax is

now a rare disease.

Recent important epidemics of human anthrax include those from Zimbabwe between 1978 and 1982, from Switzerland in 1991, and from Sverdlovsk in the former Soviet Union in 1979. The outbreak in Zimbabwe is the largest reported agricultural outbreak. There were more than 10 000 cases. Most patients had cutaneous infections, but some gastrointestinal cases were also reported. This happened after a cattle outbreak during the Rhodesian civil war that caused disruption to veterinary anthrax vaccination programmes. In the Swiss outbreak, 25 workers in one textile factory contracted the disease. Twenty-four cases had cutaneous and one inhalation anthrax. The factory had imported *B. anthracis*-contaminated goat hair from Pakistan. This outbreak shows that the potential for industrial transmission of *B. anthracis* still exists in the Western world. In the former Soviet Union, the health authorities initially attributed the outbreak in Sverdlovsk to meat from infected animals. It is now known that deaths were in fact due to inhalation anthrax, the result of an accidental aerosolized release of anthrax spores from a microbiology laboratory in the local military facility. There were at least 77 cases of anthrax and 66 deaths.

Pathogenesis

Pathological changes result from tissue invasion by *B. anthracis* and effects of its exotoxin. The organism is an extracellular pathogen that multiplies rapidly, invades the bloodstream and kills quickly. Virulent strains of *B. anthracis* possess two virulence factors: a capsule and a three-component protein exotoxin (anthrax toxin) that is made up of protective antigen (**PA**), oedema factor (**EF**), and lethal factor (**LF**). The capsule, which is composed of D-glutamic acid polypeptide, enhances virulence by making the organism resistant to phagocytosis. The genes encoding the anthrax capsule are carried on an extrachromosomal plasmid. A second plasmid carries the gene for the three proteins: PA, EF, and LF. PA is so named because it is the main protective constituent of anthrax vaccines. PA binds to a cell-surface receptor on the target cell and is cleaved by a protease into two fragments. After the cleavage, the larger membrane-bound fragment displays a binding site for EF and LF and mediates their entry into the target cell. EF, a calmodulin-dependent adenylate cyclase, increases intracellular cyclic adenosine monophosphate levels. The biological effects of EF include the formation of oedema characteristic of the disease. A mixture of EF and PA, known together as oedema toxin, also inhibits phagocytosis by polymorphonuclear leucocytes. The action of LF, believed to be a metalloproteinase, is less understood. Its exact intracellular target is not known. One recent study showed that LF inactivates an enzyme responsible for activating mitogen-activated protein kinase (**MAPK**). Thus, LF blocks the MAPK signal transduction pathway, an evolutionarily conserved pathway that controls cell proliferation and differentiation. Injection of a mixture of PA and LF, known together as lethal toxin, causes death in many species of experimental animals. Lethal toxin has been shown at high concentration to kill macrophages and at low concentration to cause macrophages to release tumour necrosis factor and interleukin 1.

When spores of *B. anthracis* are introduced subcutaneously, they germinate and multiply. The antiphagocytic capsule facilitates local spread. The oedema and lethal toxins impair leucocyte function and contribute to tissue necrosis, oedema, and relative absence of leucocytes in the skin lesion. The bacilli spread to the draining lymph node resulting in the typical findings of haemorrhagic, oedematous, and necrotic lymphadenitis. Gastrointestinal anthrax follows ingestion of contaminated and undercooked meat. Multiplication of the bacilli in the oropharynx and the draining lymph nodes causes the oropharyngeal ulcer and neck swelling. When the organisms are deposited in the duodenum, ileum, or caecum, they cause mucosal inflammation and ulcers. Transport of the bacteria to the mesenteric lymph nodes results in the development of haemorrhagic adenitis and ascites. Inhalation anthrax follows deposition of spore-bearing particles of 1 to 5 µm into alveolar spaces. They are phagocytosed by alveolar macrophages and transported to the tracheobronchial and mediastinal lymph nodes, where they germinate. Production of toxins leads to haemorrhagic, oedematous, and necrotic lymphadenitis and mediastinitis.

In all primary forms of anthrax, especially inhalation anthrax, the bacilli can spread through the blood causing septicaemia and at times haemorrhagic meningitis. Autopsies show numerous bacteria in blood vessels, lymph nodes, and other organs.

Clinical features

Cutaneous anthrax

The cutaneous lesion in anthrax is most often found on exposed areas of skin such as the face, neck, arms, or hands. The incubation period is 1 to 7 days, usually 2 to 5 days. Initially a small papule develops at the site of infection. During the next week, the lesion typically progresses through vesicular and pustular stages to the formation of an ulcer with a depressed black eschar (Fig. 2 and Plate 2). A striking degree of non-pitting oedema surrounding the lesion is typical. The early lesion may be pruritic. The fully developed lesion is painless. Small satellite vesicles may surround the original lesion. Lymphangitis and painful regional lymphadenitis is common. Associated systemic symptoms are usually mild, and the lesion heals after the eschar separates. In about 10 to 20 per cent of the patients, the disease progresses with massive local oedema, toxaemia, and bacteraemia, and a fatal outcome if untreated. Cutaneous anthrax should be considered when patients have painless ulcers associated with vesicles, oedema out of proportion to the size of the lesion, and have had contact with animals or animal products. The differential diagnosis includes staphylococcal skin infections, tularaemia, plague, cutaneous diphtheria, orf, rickettsial pox, and scrub typhus.

Fig. 2 Cutaneous anthrax lesion on the forearm on day 10 showing an ulcer with a depressed black eschar. (See also Plate 2.)

Gastrointestinal anthrax

Because gastrointestinal anthrax develops following consumption of contaminated meat, it can occur as familial clusters. The disease has been described in two forms. The incubation period is 2 to 5 days. Oropharyngeal anthrax follows deposition of the bacteria in the oropharynx. Patients present with fever, neck swelling, sore throat, and dysphagia. The neck swelling is caused by enlargement of the lymph nodes together with subcutaneous oedema as in diphtheria. The lymph node enlargement commonly involves the upper group of the jugular chain. There is an inflammatory lesion in the oral cavity or oropharynx. The lesion starts as an inflamed mucosa, progressing through necrosis and ulceration to the formation of a pseudomembrane covering the ulcer (Fig. 3 and Plate 3). In severe cases, the subcutaneous oedema extends to the anterior chest wall and axilla, with the overlying skin showing signs of inflammation. Toxaemia and death may follow. Oropharygeal anthrax should be considered in patients who present with fever, neck swelling, sore throat, and oropharyngeal ulcer and who give a history of eating raw or undercooked meat. The differential diagnosis includes diphtheria and peritonsillar abscess.

Fig. 3 Oropharyngeal anthrax on day 9 showing a pseudomembrane covering an ulcer. (See also Plate 3.)

In another form of gastrointestinal anthrax, the organisms are deposited in the duodenum, terminal ileum, or caecum. The onset is with fever, nausea, vomiting, and abdominal pain, followed by rapidly developing ascites and bloody diarrhoea. Haematemesis, melaena, haematochezia, and/or profuse watery diarrhoea may occur in some patients. In severe cases, toxaemia, shock, and death follow. It is difficult to make a diagnosis in the early stage, except in an epidemic setting.

Inhalation anthrax

Inhalation anthrax (woolsorter's disease) has been described as a two-stage illness. The incubation period is 1 to 5 days. It starts with malaise, myalgia, fever, and non-productive cough, symptoms similar to those of viral respiratory diseases. Signs of illness and laboratory results are non-specific. In some patients there is a transient improvement after 2 to 4 days. The second stage begins with severe respiratory distress, cyanosis, stridor, and profuse sweating. Subcutaneous oedema of the chest and neck may develop. A characteristic radiographic finding is symmetric mediastinal widening with or without pleural effusion. Blood culture will grow *B. anthracis*. Up to half of patients develop anthrax meningitis. Shock and death typically follow in less than 24 h. The initial phase of the disease is very difficult to diagnose in the absence of a known outbreak. Advanced disease may be suspected in the presence of a characteristically widened mediastinum despite otherwise normal chest radiographic findings. Inhalation anthrax must be distinguished from pneumonic plague.

Meningeal anthrax

Anthrax meningitis, frequently a consequence of overwhelming *B. anthracis* bacteraemia, may complicate any primary form of anthrax. Rarely, a case of anthrax meningitis has been reported in which the primary site was not identified. Within a few days of the primary lesion, the patient develops sudden onset of confusion, loss of consciousness, and focal neurological signs. The cerebrospinal fluid is haemorrhagic, purulent, or both. The disease is almost always fatal.

Laboratory diagnosis

Clinical specimens taken before antibiotic therapy is instituted will grow *B. anthracis* in culture. Gram stain of these specimens may show Gram-positive rods. These specimens include vesicular fluid (cutaneous anthrax), swabs from oropharyngeal lesions, ascitic fluid (gastrointestinal anthrax), and cerebrospinal fluid (meningeal anthrax). Severe cases of anthrax, especially the inhalation form, may have bacteraemia, but by the time blood cultures become positive many patients will have died. Because *Bacillus* species are frequent laboratory contaminants, most laboratories do not identify them further unless specifically asked to do so. Thus, it is important to notify the laboratory when *B. anthracis* infection is suspected. Laboratory confirmation can also be made by demonstration of *B. anthracis* in the clinical specimens by direct fluorescent antibody staining.

Serological tests are helpful in making a diagnosis, especially when prior antibiotics have eradicated the bacteria before cultures or smears were obtained. However, patients with severe disease, especially inhalation anthrax, die so quickly that these tests may not be helpful to the clinician. Useful tests include an enzyme-linked immunosorbent assay (ELISA), which detects antibodies to the capsular antigen and/or PA, and an electrophoretic immunotransblot test (Western blot), which detects antibodies to PA and/or LF. These tests are only available at national reference laboratories.

Treatment

Most strains of *B. anthracis* are susceptible to penicillin, which should be started as soon as possible. Mild cases of cutaneous anthrax may be treated with oral penicillin at the dose of 250 mg 6-hourly for 5 to 7 days. For extensive lesions, parenteral penicillin G, 2 million units 6-hourly, should be given until systemic toxicity subsides. The patient may then take oral penicillin for a total treatment period of 7 to 10 days. Ciprofloxacin, erythromycin, doxycycline, or chloramphenicol can be used in penicillin-sensitive patients. Antibiotics decrease systemic toxicity, but the skin lesions will continue to progress to the eschar phase. The skin lesion should be covered with a sterile dressing. Used dressings should be decontaminated. In gastrointestinal or inhalation anthrax or in anthrax meningitis, intravenous penicillin G, 4 million units 4-hourly, should be administered. Other drugs are intravenous ciprofloxacin (800 mg/day) or intravenous doxycycline (200 mg/day). Many patients will require intensive supportive care including vigilant monitoring and correction of electrolyte and acid–base disturbance, mechanical ventilation, and vasopressor administration.

Prognosis

A case fatality rate of 10 to 20 per cent has been reported for untreated cases of cutaneous anthrax. With appropriate antibiotic treatment, fatalities are rare. Almost all cases of inhalation anthrax and anthrax meningitis are fatal. Oropharyngeal anthrax causes death in about 15 per cent of the patients. The case fatality rate of the other form of gastrointestinal anthrax is not known.

Prevention and control

Control of anthrax in animals is essential to control of the disease in humans. Routine immunization of animals should be instituted in areas with continuing cases of animal anthrax. All cases of animal or human anthrax should be reported to the appropriate authorities. During an outbreak, suspected animals should be quarantined and infected herds sacrificed. Carcasses of animals that have succumbed to anthrax are buried intact or cremated to avoid sporulation and further contamination of the environment. Gastrointestinal anthrax can be prevented by public education about consumption of contaminated meat. Anthrax vaccines should be given to those at risk of acquiring the disease such as agricultural workers or veterinarians who have contact with potentially infected animals, laboratory workers who work with *B. anthracis*, and workers involved in the industrial processing of animal products. The incidence of industrial anthrax has been further decreased by educating workers about how anthrax is transmitted, improvement in industrial hygiene, better manufacturing equipment and environmental control, as well as the decline in using fibres of animal origin as raw material.

Current anthrax vaccines for human use include cell-free preparations consisting of alum-precipitated and aluminum hydroxide-adsorbed extracellular components (primarily PA) of uncapsulated *B. anthracis*, available in the United Kingdom and the United States, respectively. A live attenuated anthrax spore vaccine is available in countries of the former Soviet Union, but is not used elsewhere because of safety concerns. The current cell-free vaccines are manufactured from an undefined crude culture supernatant. They must be given several times to ensure protection and local reactions have been reported. These drawbacks and the potential use of *B. anthracis* as a biological weapon have stimulated efforts to develop improved vaccines. A minimally reactogenic, recombinant PA vaccine has been investigated. Other approaches, made possible by modern molecular biology technology, include cloning the PA gene into other bacteria or viruses and development of mutant avirulent strains of *B. anthracis.*

Further reading

Hanna P (1998). Anthrax pathogenesis and host response. *Current Topics in Microbiology and Immunology* **225**, 13–35. [A review on the pathogenesis of *Bacillus anthracis*.]

LaForce FM (1994). Anthrax. *Clinical Infectious Diseases* **19**, 1009–13. [A good review article on anthrax.]

Meselson M *et al*. (1994). The Sverdlovsk anthrax outbreak of 1979. *Science* **266**, 1202–8. [Description of the outbreak at a military facility at Sverdlovsk in the former Soviet Union.]

Pile JC *et al*. (1998). Anthrax as a potential biological warfare agent. *Archives of Internal Medicine* **158**, 429–34. [A good review article on anthrax in general and as a potential biological weapon.]

Sirisanthana T *et al*. (1988). Serological studies of patients with cutaneous and oral-oropharyngeal anthrax from northern Thailand. *American Journal of Tropical Medicine and Hygiene* **39**, 575–81. [Studies of serological diagnosis in patients with oropharyngeal and cutaneous anthrax.]

7.11.19

Brucellosis

M. Monir Madkour

Brucellosis is a common, classic zoonotic disease of worldwide distribution. It is transmitted to man from infected animal reservoirs. Human brucellosis may be caused by one of four species: *Brucella melitensis* (the most common cause) from goats, sheep, and camels; *B. abortus* from cattle; *B. suis* from pigs; and *B. canis* from dogs. Brucella organisms are small, non-encapsulated, non-motile, non-sporing, Gram-negative, aerobic bacilli, which are facultative intracellular parasites. They can survive for up to 8 weeks in unpasteurized, white, soft goat's cheese. They tend to die within 60 to 90 days in cheese that has undergone lactic acid fermentation during the period of maturing. Freezing milk or its products does not destroy the organism, but they are killed by boiling or pasteurization. Brucella organisms are shed in urine, stools, vaginal discharge, and products of conception. They remain viable in dried soil for up to 40 days and for longer if the soil is damp.

Epidemiology

There are only 17 countries in the world that are brucellosis free: Norway, Sweden, Finland, Denmark, Switzerland, the former Czechoslovakia, Romania, the United Kingdom including the Channel Islands, the Netherlands, Japan, Luxembourg, Cyprus, Bulgaria, Iceland, and the Virgin Islands of the United States. Canada and New Zealand are about to be declared brucellosis-free countries. However, the overall incidence of brucellosis in the world is increasing. With the ease of modern travel, patients may contract the disease while visiting endemic countries. The true global incidence of human brucellosis is difficult to determine because of the lack of essential statistics, disease reporting, and notification systems in many countries. Even in developed countries there are reports indicating that the incidence of human brucellosis is estimated to be 3 to 26 times higher than official figures.

The risk to public health

In endemic areas of developing countries, brucellosis affects predominantly males and younger age groups. Farm animals such as goats, sheep, camels, and cattle are kept in the backyards of houses and considered as pets. Childhood brucellosis indicates endemicity of the disease in that area. Serious human maternal morbidity and fetal loss through abortion, intrauterine death, and premature delivery, or active disease in neonates, are public health risks in endemic areas of developing countries. Where brucellosis is controlled in animals, human brucellosis is mostly an occupational disease, particularly among workers in meat-processing industries and in farmers, veterinarians, and laboratory workers.

Modes of transmission

In endemic areas, animal contact through inhalation of organisms is the most frequent cause of infection, and affects herdsmen, dairy-farm workers, and laboratory workers. *Brucella melitensis* is included among 10 types of biological warfare or bioterroristic agents which can be released as aerosols. Ingestion of untreated milk or its products, or raw meat, liver, or bone marrow, is a common route of infection through the gastrointestinal tract, particularly among those taking antacids. Penetration of intact or abraded skin is a common route of infection among abattoir workers in developing and developed countries. Accidental autoinoculation or conjunctival splashing of live brucella vaccine during animal vaccination are well-recognized routes of infection among veterinarians. Laboratory infections have been described. Transplacental transmission of infection from mother to fetus may occur. Brucella organisms have been isolated from human breast milk and nursing mothers may infect their infants through breast feeding. Sexual transmission in man is similar to that in animals, and has recently been reported, with isolation of the organisms from human semen. Other uncommon routes of transmission include blood transfusion and bone marrow transplants.

Pathogenesis

Polymorphonuclear cells and activated macrophages migrate to the site of entry of brucella organisms. During the early phase of invasion, extracellular killing is carried out by IgM and complement-mediated mechanisms.

However, brucella organisms can resist such killing. During invasion and phagocytosis, the organisms are killed inside macrophages by oxidative burst or oxygen-based killing using the myeloperoxide–hydrogen peroxide–halide system. The interaction between the organisms and macrophages will determine the severity and outcome of infection. Organisms surviving within or escaping from phagocytic cells multiply and reach the bloodstream via the lymphatics to enter body organs rich in reticuloendothelial cells. Other organs and tissues are also invaded through the bloodstream. Inflammatory responses with or without granulomas and caseation or even abscess formation may occur. The cytotoxic activity of natural killer cells, with decrease in the CD4+ and increase in the CD8+ lymphocyte subpopulations, is depressed in patients with active brucellosis. Cytokines including interleukin12 and tumour necrosis factor-a appear to play an important role in host defence against brucella infection.

Clinical features

The incubation period is about 1 to 3 weeks but may extend up to several months. Brucellosis is a disease of protean manifestations that may simulate other febrile illnesses. Its clinical features are not specific. In endemic areas, diagnosis is relatively easy. However, in non-endemic areas of developed countries, clinicians should remember brucellosis in the differential diagnosis of a febrile illness. A history of travel to endemic areas should be obtained, as well as the patient's occupational history. The clinical features of brucellosis largely depend on the species of the organism and may vary widely. *B. melitensis* has a high pathogenicity, producing more intense symptoms. The onset may be sudden (1 to 2 days) or gradual (1 week or more). It presents as a febrile illness, with or without localization to particular organs. Brucellosis is classified according to whether or not the disease is active (i.e. history, clinical features, and significantly raised brucella agglutinins with or without positive blood cultures) and whether or not there is localized infection. Evidence of active disease and the presence of localization have a significant impact on recommended treatment. Classification of brucellosis as acute, subacute, chronic, serological, bacteraemic, or mixed types serves no purpose in diagnosis and management. The term 'active brucellosis with/without localization' is recommended. The most frequent symptoms are given in Table 1. The fever has no distinctive pattern that could differentiate it from other febrile illnesses, despite the old name 'undulant fever'. It usually shows diurnal variation, being normal in the morning and

high in the afternoon and evening. Chills or rigors with profuse sweating may simulate malaria. Patients with brucellosis commonly look deceptively well and, less frequently, may look acutely ill. Physical signs may be lacking despite the multiplicity of symptoms, which may be labelled as psychological. The frequency of physical signs is shown in Table 2.

Localizations

Bones and joints

Reactive arthritis may occur in brucellosis. Septic arthritis may result from bloodborne spread to the synovium or from extension of brucellar osteomyelitis in a neighbouring long bone. Brucella spondylitis starts in the superior end-plate, an area of rich blood supply. The infection may either regress and heal or progress to involve the entire vertebra, disc space, and adjacent vertebrae. Early lesions are localized in the anterior aspect of the superior end-plate at the disc–vertebral junction, leading to a small area of bone destruction. Bone healing takes place at the same time, leading to sclerosis.

Arthritis is commonly polyarticular and migratory, affecting mainly the large joints including the knee, hip, sacroiliac, shoulder, sternoclavicular, wrist, ankle, and interphalangeal joints in decreasing order of frequency. Septic monoarthritis may lead to destruction of the affected joint if undiagnosed. Joints affected include the knee, hip, sternoclavicular, and sacroiliac joints and the shoulder. Spondylitis may involve single or, less frequently, multiple sites. The lumbar spine, particularly L4, is the most frequent site. The average age of onset of brucella spondylitis is 40 years; it is extremely rare during childhood.

Extraspinal brucella osteomyelitis is rare. Long bones, particularly the femur, tibia, humerus, or manubrium sterni, may be affected. Bursitis, tenosynovitis, and subcutaneous nodules may also occur. Unlike with septic arthritis and osteomyelitis due to other organisms, the peripheral white-cell count is normal and the erythrocyte sedimentation rate is normal or accelerated. The total white-cell count in synovial fluid ranges from 4000 to 40 000/mm ^{3} with 60 per cent polymorphonuclear cells. Glucose in synovial fluid may be reduced, but protein is usually raised and culture is positive in about 50 per cent of the cases.

Cardiovascular

These localizations may include endocarditis, myocarditis, pericarditis, aortic-root abscess, mycotic aneurysms, thrombophlebitis, and pulmonary embolism. The most frequent of these is endocarditis, which used to be the leading cause of death. The outcome is now more favourable with recent advances in diagnosis, cardiac surgery, and treatment. Brucella endocarditis usually occurs on a previously damaged valve or a congenital malformation, but can occur even on normal valves. The clinical features are similar to those caused by other organisms. Patients who live in endemic areas and have what has been labelled as 'sterile infective endocarditis' should have their blood culture extended for a period of up to 6 weeks.

Respiratory

Respiratory symptoms are common but, because they are usually mild, clinicians tend to overlook them. A flu-like illness with sore throat and mild dry cough is a common feature. Other rare foci of infection include hilar and paratracheal lymphadenopathy; pneumonia, with solitary or multiple nodular lung shadowing or even with abscess formation; soft-tissue miliary shadowing; pleural effusion; empyema; or mediastinitis.

Gastrointestinal

Gastrointestinal infections are usually mild and are rarely a presenting feature of the disease. They include tonsillitis, and hepatitis with mild jaundice (either non-specific or granulomatous with suppuration and abscess formation). Actual cirrhosis is rare. Deep jaundice is not a feature of brucellosis. Splenic enlargement with abscess formation is rarely reported. Mesenteric lymphadenopathy with abscess formation, cholecystitis, peritonitis, pancreatitis, and ulcerative colitis are described. The liver transaminases, alkaline phosphatase, and serum bilirubin may be mildly raised. The clinical and biochemical evidence of liver involvement is far less frequent than liver biopsies have indicated. The diagnostic significance of splenomegaly becomes doubtful in countries where malaria and bilharzia are also common.

Genitourinary

Genitourinary localizations may be the presenting feature of brucellosis. They include unilateral or bilateral epididymo-orchitis in children and in adults, prostatitis, seminal vesiculitis, dysmenorrhoea, amenorrhoea, tubo-ovarian abscesses, chronic salpingitis, and cervicitis. Acute nephritis or acute pyelonephritis-like features, renal calcifications, and calyceal deformities may occur. Renal granulomatous lesions with abscess formation, with caseation and necrosis may occur, as may cystitis and posterior urethritis.

Urine culture may be positive in about 50 per cent of patients with brucellosis. Brucella organisms have recently been isolated from human semen during investigation of possible sexual transmission.

Neurobrucellosis

Neurobrucellosis is uncommon but serious. Despite the multiplicity of symptoms, abnormal neurological findings may be lacking. They include meningoencephalitis, multiple cerebral or cerebellar abscesses, ruptured mycotic aneurysm, cranial nerve lesions, transient ischaemic attacks, hemiplegia, myelitis, radiculoneuropathy and neuritis, Guillain–Barré syndrome, a multiple sclerosis-like picture, paraplegia, sciatica, granulomatous myositis, and rhabdomyolysis.

The psychiatric features of brucellosis are no more severe than those caused by other infections. Neurobrucellosis may be caused by direct bloodborne invasion by brucella organisms, pressure from destructive spinal lesions, vasculitis, or an immune-related process. In meningoencephalitis the cerebrospinal fluid pressure is usually elevated and the fluid may look clear, turbid, or rarely, haemorrhagic; the protein, cells (predominantly lymphocytes), and oligoclonal immunoglobulin are raised, while glucose may be reduced or normal. Brucella organisms may be cultured from cerebrospinal fluid.

Pregnancy

In endemic areas the outcome of pregnancy in humans is similar to that in animals: normal delivery, abortion, intrauterine fetal death, premature delivery, or retention of the placenta and other products of conception.

Skin

Skin manifestations are uncommon. They include maculopapular eruptions and contact dermatitis, particularly among veterinarians and farmers assisting animal parturition. Other dermatological manifestations include erythema nodusum, purpura and petechias, chronic ulcerations, multiple cutaneous and subcutaneous abscesses, vasculitis, superficial thrombophlebitis, discharging sinuses, and rarely, pemphigus.

Ocular

Direct splashing of live brucella vaccine into the eyes may cause conjunctivitis. Keratitis, corneal ulcers, uveitis, retinopathies, subconjunctival and retinal haemorrhages, retinal detachment, and endogenous endophthalmitis with positive vitreous cultures are well documented. Neuro-ophthalmic complications of brucella meningitis may lead to papilloedema, papillitis, retrobulbar neuritis, optic atrophy, and ophthalmoplegia due to lesions on the IIIrd, IVth, and VIth cranial nerves.

Endocrine

Localization of brucella infection with or without abscess formation in the endocrine glands is commonly reported in the testicle and epididymis. Other endocrine gland localizations with or without abscess formation are well documented but rare. These include the thyroid, ovaries, mammary glands, the adrenals, and the prostate. Reported cases of endocrine gland localizations are commonly not associated with disturbed hormonal secretions, perhaps with the exception of the adrenals. The

syndrome of inappropriate secretion of antidiuretic hormone as well as raised serum calcium are reported in patients with active brucellosis.

Diagnosis

The diagnosis of brucellosis depends on the presence of clinical features and brucella agglutinins in a significantly raised titre. A positive blood or tissue culture is not always present. The organism's identity is confirmed by phage typing, DNA characterization, or metabolic profiling. Use of a carbon dioxide detection system (such as BACTEC; Becton Dickinson, Sparks, MD) for blood culture provides a more sensitive and rapid culture result than standard methods, with positivity usually apparent after only 2 to 5 days of incubation. Alternatively, extended incubation of blood cultures for up to 6 weeks (incubated at 37°C with and without an atmosphere of 10 per cent carbon dioxide) should be requested. Most authorities will consider an agglutination titre of 1/160 or higher to be significant in a symptomatic patient living in a non-endemic area. However, in endemic areas only titres of 1/320 to 1/640 or higher are considered significant. In endemic areas, otherwise asymptomatic individuals offering to donate blood may be found to have high brucella titres and should not be considered to be suffering from active brucellosis. Follow-up 2 to 4 weeks later is necessary in such individuals to exclude subclinical infection.

The presence of brucella antibodies in the patient's serum can be detected by the standard tube test, rose bengal plate test, 2-mercap-toethanol test, antihuman globulin test (Coomb's), radioimmunoassay, enzyme immunoassay, and polymerase chain reaction (**PCR**). The PCR is specific and highly sensitive for the detection of brucella agglutinins (the DNA used for the amplification is either phenol purified or comes directly from a suspension of brucella organisms). The antigens commonly used for serological screening are prepared from *B. abortus*, which cross-reacts with *B. melitensis* and *B. suis* antibodies as well. However, they do not cross-react with *B. canis* antibodies. To detect these antibodies, antigen prepared from *B. canis* organisms is needed, but they are not available commercially. A cross-reaction with tularaemia and cholera may occur. This can be distinguished by testing simultaneously for brucella, tularaemia, and cholera antibodies. Occasionally, brucella agglutination tests are negative in patients with positive tissue cultures. The prozone phenomenon is a false-negtive standard tube test caused by the presence of blocking antibodies in the a-globulin (IgG) and in the a $_2$-globulin (IgA) fractions. This phenomenon can be avoided by screening sera at low and high titres. An elevated IgM antibody indicates recent infection, while low titres indicate previous contact with the organism. An elevated IgG indicates active disease.

Haematological changes

The total white-cell count is usually normal and leucopenia with relative lymphocytosis does not always occur. Thrombocytopenia is less common and haematological features of disseminated intravascular coagulation are rare. The erythrocyte sedimentation rate is of no diagnostic value. Liver function tests, liver biopsies, and cerebrospinal and synovial fluid changes have been discussed under pathogenesis and localizations.

Treatment

Control and prevention of brucellosis should be directed primarily towards eradication of the disease in animals. The brucella organism is intracellular and therefore relatively inaccessible to antimicrobials. A combination of a tetracycline and an aminoglycoside remains the most effective regimen because of its synergistic effect. Oral doxycycline (100 mg, twice daily) is preferred to other tetracyclines (500 mg, 6-hourly) because of its rapid and complete absorption from the duodenum, longer half-life (18 h), and more efficient tissue penetration (it is more lipid soluble). Suitable aminoglycosides are streptomycin, netilmicin, or gentamicin. Streptomycin is given intramuscularly in a dose of 1 g/day for patients under 45 years of age and 0.5 to 0.75 g/day for older patients. The plasma trough concentration should be 1 to 2 µg/ml. Netilmicin, 4 to 6 mg/kg a day intramuscularly in two divided doses, can be used for outpatient treatment. The plasma trough concentration should be 2 to 4 µg/ml. Gentamicin is only used for patients in hospital as it is usually given as an intravenous infusion of 2 to 5 mg/kg daily, in divided doses, 8-hourly. The plasma trough should be 1 to 2 µg/ml. Combined therapy with a tetracycline and an aminoglycoside should be given for 1 month, followed by a tetracycline and rifampicin (600 to 900 mg/day as a single oral dose) or a tetracycline and co-trimoxazole (two tablets, 480 mg each, twice daily) for a further 1 to 2 months. This regimen has a relapse rate of 7 per cent. A three-drug regimen in combination with urgent surgical intervention is required in those with endocarditis, aortic root abscess, spondylitis, osteomyelitis, and abscesses in organs or other tissues. Neurobrucellosis without abscesses formation will require a three-drug regimen. The combination of doxycycline–netilmicin/gentamicin–rifampicin should be given for 4 weeks. A doxycycline–rifampicin combination should be continued for a further 4 to 8 weeks. Single daily dosing of netilmicin or gentamicin has been successfully used for other infections. Such dosing is being assessed at present for treatment of brucellosis and results are not yet available. Shorter periods of treatment have a higher relapse rate. Most patients with brucellosis are treated as outpatients and only those with localizations (e.g. endocarditis, neurobrucellosis, osteomyelitis, septic arthritis, and renal impairment), or who are pregnant or are infants, require admission to hospital.

Ciprofloxacin (750 mg, 12-hourly, orally) and other fluoroquinolones are synthetic broad-spectrum antibiotics with intracellular penetration used by some for treatment of brucellosis. There are reports of the development of resistance and cross-resistance with other quinolones and high relapse rates. Quinolones showed no synergism with other agents.

Children

Infants and children under 7 years of ages should be treated with a combination of rifampicin and co-trimoxazole for 2 to 3 months. However, in those with serious localizations in endemic areas where some discoloration of the teeth is of secondary importance, doxycycline can be used, in combination, as described above—doxycycline, 50 to 100 mg/day orally; gentamicin: infants aged up to 2 weeks, 3 mg/kg every 12 h; aged 2 weeks to 12 years, 2 mg/kg every 8 h intramuscularly or by slow intravenous injection or intravenous infusion; netilmicin: infants aged up to 1 week, 3 mg/kg every 12 h; aged over 1 week, 2.5 to 3 mg/kg every 8 h intramuscularly or by intravenous injections or infusions; rifampicin, 10 to 20 mg/kg a day, either orally or by slow intravenous injection as a single daily dose; co-tri-moxazole paediatric suspension (240 mg/ml) is given 12-hourly orally as follows: 6 weeks to 5 months of age, 120 mg; 6 months to 5 years, 240 mg; 6 to 12 years, 480 mg. Intravenous infusion: 54 mg/kg daily in two divided doses.

Renal impairment and pregnancy

Patients with renal impairment should be carefully monitored for serum concentration of aminoglycoside. If such monitoring is not available, then a doxycycline–rifampicin regimen should be administered. In pregnancy, co-trimoxazole–rifampicin for 8 to 12 weeks is the most suitable regimen.

Response to treatment

Patients become afebrile and other constitutional symptoms greatly improve within 4 to 14 days. The liver and spleen become impalpable within 2 to 4 weeks. Patients may experience an acute, intense flare-up of symptoms—the Jarisch–Herxheimer reaction—shortly after starting treatment, particularly with tetracyclines. This reaction is only transient and does not necessitate discontinuation of therapy. Follow-up of clinical, blood culture, and serological tests should be done every 3 to 6 months for 1 to 2 years.

Human vaccine

Human vaccine for brucellosis, used in the former Soviet Union, China, and France, was found to be effective in reducing markedly the rate of infection. Two injections, each of 1 mg of phenol-insoluble fraction, were given 2 weeks apart. It provides effective but short-lived immunity and should be repeated every 2 years. Vaccination is indicated in workers with an occupational risk of developing brucellosis. The outer membrane proteins (OMPs) are showing promise in experimental work on the development of a new vaccine.

Further reading

Banntyne RM *et al*. (1997). Rapid diagnosis of brucellar bacteraemia by using the BACTEC 9240 system. *Journal of Clinical Microbiology* **35**, 2673–4.

Berkowsky PB *et al*. (1997). Why should we be concerned about biological warfare? *Journal of the American Medical Association* **278**, 431–2.

Madkour MM (1989). *Brucellosis*, 1st edn. Butterworths, London.

Madkour MM (2001). *Madkour's brucellosis*, 2nd edn. Springer-Verlag, Heidelberg. [A complete monograph on brucellosis, enhanced by 216 figures of plain radiography, CT, and MRI modalities.]

Sharif HS *et al*. (1989). Brucellar and tuberculous spondylitis: comparative imaging features. *Radiology* **171**, 419–25.

Solera J *et al*. (1996). Treatment of human brucellosis with netilmicin and doxycycline. *Clinical Infectious Diseases* **22**, 441–5.

Tetanus

7.11.20

F. E. Udwadia

Tetanus is an acute, often fatal disease, resulting from the contamination of a wound by *Clostridium tetani*, a spore-forming, Gram-positive, motile, rod-shaped, obligate anaerobic organism. Under anaerobic conditions the vegetative form of the organism produces a powerful exotoxin, which on reaching the central nervous system causes the increased muscle tone and spasms that characterize the disease.

Epidemiology

The spores of *Cl. tetani* are ubiquitous, but the natural environment is soil, particularly cultivated soil rich in manure. Spores are commonly found in animal faeces, may also be detected in human faeces, and can occasionally be recovered from house dust or from the air of occupied buildings, slums, and even hospitals and operating theatres.

Tetanus is a killer disease chiefly afflicting the poor, uneducated, and underprivileged people of the world. It is thus widely prevalent in India, Bangladesh, Pakistan, parts of South-East Asia, Africa, the eastern Mediterranean region, and South America. In these countries where immunization programmes are inadequate, the disease is most common in the young and newborn and is more frequent in males than in females. In the 1980s, 1 million newborn babies died of tetanus every year. The annual worldwide mortality had declined to 480 000 by 1994 and to 277 400 in 1997. However, neonatal tetanus still accounts for 23 to 73 per cent of neonatal deaths in developing countries. An estimated 70 000 cases continue to occur in India every year. The decline in incidence of the disease in the nineties is largely due to the substantial increase in immunization coverage of pregnant women with a protective dose of tetanus toxoid. It is estimated that effective vaccination programmes have prevented 500 000 deaths in the South-East Asia region out of the 700 000 deaths prevented globally. Ninety per cent of the deaths prevented in South-East Asia are in India, Bangladesh, and Indonesia. In the West the disease is increasingly rare, fewer than 60 cases being reported annually from the United States between 1991 and 1994. The disease in the West is more frequent in people older than 60 years, in whom effects of immunization have worn off, in the unimmunized, impoverished, and in drug addicts.

Physiopathology

Under anaerobic conditions (e.g. presence of necrotic tissue, active infection, foreign body), the tetanus bacillus within a wound produces two toxins—tetanospasmin and tetanolysin. Only tetanospasmin has clinical effects. Tetanospasmin is a 150 kDa protein consisting of a heavy (100 kDa) chain and a light (50 kDa) chain joined by a single disulphide bond. The mechanism of spread of tetanospasmin is illustrated in Fig. 1. The released toxin spreads to underlying muscles and is bound by its heavy chain to receptors containing gangliosides on the neuronal membranes of presynaptic nerve terminals. The toxin is then internalized and transported intra-axonally and retrogradely within the peripheral nerves to cells of motor neurones of that segment of the cord supplying those muscles. The toxin almost always also enters and circulates in the bloodstream. It does not cross the blood–brain barrier, but by haematogenous spread binds to nerve terminals in muscles throughout the body. It is then transported retrogradely within numerous axonal pathways of all peripheral nerves to reach the a motor-neurone cell bodies of the whole spinal cord and brainstem. It thereby also reaches the sympathetic chain, the preganglionic sympathetic neurones in the lateral horns of the spinal cord, and the parasympathetic centres.

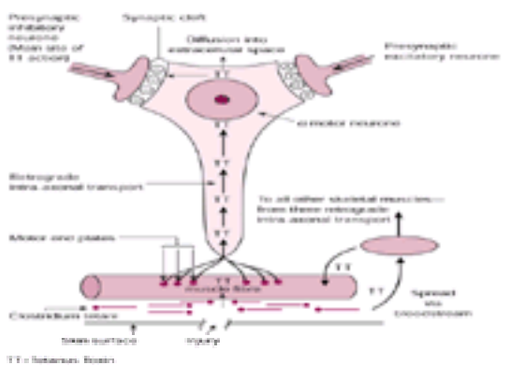

Fig. 1 Retrograde intra-axonal transport of tetanospasmin and its main site of action in the central nervous system.

After reaching the cell bodies in the spinal cord and brainstem, the toxin, by an unknown mechanism, passes retrogradely across the presynaptic cleft to bind to ganglioside receptors on presynaptic nerve terminals of inhibitory interneurones. The light chain of the toxin now acts to block the release of the inhibitory neurotransmitters, chiefly glycine and g-aminobutyric acid (**GABA**), from synaptic vesicles within nerve terminals of inhibiting neurones. This blockage releases motor and autonomic neurones from inhibitory control. The molecular mechanism behind this action is unknown. The toxin may well alter a calcium-dependant process necessary for neurotransmitter release. The uncontrolled excessive, disinhibited efferent discharge from motor neurones in the cord and brainstem to both agonist and antagonist muscles leads to widespread muscle rigidity and to reflex spasms characteristic of generalized tetanus. Muscles of the jaw, face, and head are involved first because the toxin has to travel along shorter axonal pathways to reach their controlling motor neurones in the brainstem. Muscles of the trunk and limbs are involved a little later because the toxin travels along longer axonal pathways to their controlling motor cells in the cord. Disinhibited autonomic discharge leads to disturbances in autonomic control, particularly to sympathetic overactivity with excessive catecholamines in the blood. Medullary centres and hypothalamic centres may also be affected by tetanus toxin. Myocardial dysfunction and disturbances in impulse conduction may occur.

When, rarely, tetanus toxin does not reach the bloodstream but spreads from the site of the wound along regional axonal pathways to motor neurones in a localized segment of the cord; local tetanus results. Rigidity and spasms are restricted to a group of muscles.

Tetanus toxin can also produce a peripheral neuromuscular blockade by preventing release of acetylcholine, similar to the effect of botulinum toxin. This peripheral paralytic effect is observed in cephalic tetanus.

Altered haemodynamics

Severe tetanus without complications is characterized by a high-output, hyperkinetic circulatory state with marked tachycardia, increased stroke-volume index, increased cardiac index, and a normal, left ventricular stroke-work index (Fig. 2). There is also an increase in the compliance of the vascular system due to arteriolar, capillary, and venous dilation, chiefly in skeletal muscle. These changes have been attributed to increased muscle contraction, increased sympathetic tone, and a rise in core temperature.

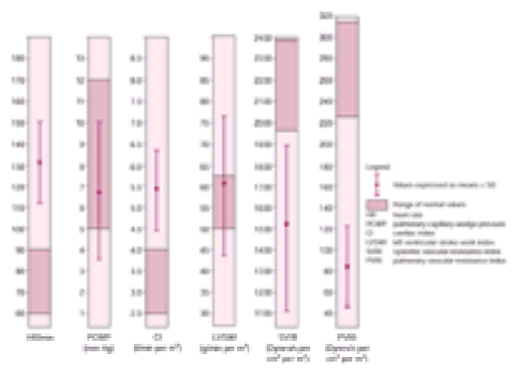

Fig. 2 Haemodynamic observations in 19 patients with severe uncomplicated tetanus.

Disturbances in the autonomic nervous system lead to marked cardiovascular instability with wide fluctuation in heart rate, systemic vascular resistance, and blood pressure.

Clinical features

Some 15 to 25 per cent of patients with tetanus have no evidence of recent wounds, for the disease can result from the most trivial of wounds. Contamination of the wound with garden soil or manure, or injury by rusty metals, are particularly dangerous. Tetanus can complicate burns, ulcers, gangrene, necrotic snake bites, frostbite, discharging middle-ear infections, septic abortions, childbirth, ritual scarification, and female circumcision. It can occur after intramuscular injections, particularly of drugs producing tissue necrosis (such as quinine), and after surgery. Tetanus neonatorum is most often due to non-sterile obstetric techniques, and in India to the dreadful practice of applying cow dung to the cut surface of the umbilical cord.

The clinical features of tetanus are rigidity, muscle spasms, and seizures. Severe tetanus is invariably associated with autonomic disturbances.

Muscle stiffness or rigidity

Stiffness of the masseters is often the first manifestation of the disease, resulting in difficulty in opening the mouth—trismus or lockjaw. Typically, stiffness extends to the muscles of the face, all skeletal muscles, and often involves muscles of swallowing, causing dysphagia. The facial expression in tetanus is diagnostic. The eyes appear partially closed, the forehead is furrowed, the corrugator muscle contracted, the nostrils flared, nasolabial folds prominent, and the lips pursed, thinned, and stretched, with the angles of the mouth extending outwards and often turned slightly down, producing a 'risus sardonicus' (Fig. 3 and Plate 1). This smile is perhaps more pathetic than sardonic. The expression is one of pain, anguish, and fear. Stiffness of the neck muscles results in retraction of the head. The muscles of the chest are stiff and the breathing movements are restricted. The abdomen often shows board-like rigidity. The arms and legs are often ramrod stiff and in children marked stiffness in the muscle of the back can lead to opisthotonos similar to that observed in meningitis (Fig. 4 and Plate 2).

Fig. 3 Facies in tetanus. (See also Plate 1.)

Fig. 4 Opisthotonos in severe tetanus during seizures. (See also Plate 2.)

Muscle spasms

Mild cases of tetanus exhibit only stiffness without spasms. Spasms or seizures are characterized by a marked reflex exaggeration of the underlying rigidity, producing tonic contraction of the stiff muscles. They are frequently brought on by touch but may also be triggered by visual, auditory, or emotional stimuli. Seizures vary in severity and frequency. They may be mild, infrequent, and brief (lasting a few seconds) or severe, protracted, painful, and spontaneous, the patient appearing to be in a state of perpetual convulsion. Severe protracted spasms render breathing impossible or shallow, irregular, and ineffective, so that the patient becomes very hypoxic and even cyanosed. Spasm of pharyngeal muscles prevents swallowing of saliva, so that pharyngeal secretions accumulate and are often aspirated into the lungs, causing atelectasis and aspiration pneumonia. Laryngeal spasm may occur by itself; it may accompany generalized spasms and can produce unexpected sudden death from asphyxia.

Patients with severe tetanus have fever, tachycardia, and often, an unstable cardiovascular system. Unless expertly managed, they usually die of respiratory complications, circulatory failure, or cardiac arrest.

Autonomic nervous system disturbances

In severe tetanus there is invariably involvement of the sympathetic and parasympathetic nervous systems. Features include tachycardia exceeding 150 beats/min, drenching sweats, frequent modest elevation in systolic and/or diastolic arterial blood pressure, increase in salivary and tracheobronchial secretions, and evidence of increased reflex vagal tone and activity.

Severity of tetanus

Grading the severity of tetanus is not just an academic exercise; it is useful both in prognosis and in the management of the disease. The criteria listed below are subjective and arbitrary but have stood the test of time in our unit.

Grade I (mild)

There is mild to moderate trismus, general spasticity, no respiratory embarrassment, no spasms, and little or no dysphagia.

Grade II (moderate)

There is moderate trismus, well-marked rigidity, mild to moderate short-lasting spasms, moderate respiratory embarrassment with tachypnoea in excess of 30 to 35/min, and mild dysphagia.

Grade III (severe)

There is severe trismus, generalized spasticity, reflex and often spontaneous prolonged spasms, respiratory embarrassment with tachypnoea in excess of 40/min, apnoeic spells, severe dysphagia, and tachycardia in excess of 120/min.

Grade IV (very severe)

The features are the same as grade III plus violent autonomic disturbances involving the cardiovascular system.

Cephalic tetanus

This occurs after an injury to the head and is confined to muscles innervated by the cranial nerves. It is characterized by unilateral facial palsy (Fig. 5 and Plate 3), trismus, facial stiffness of the unparalysed half, nuchal rigidity, pharyngeal spasms causing dysphagia, and frequent laryngeal spasms with danger of death from asphyxia. Rarely, facial palsy is bilateral. Paresis of the glossopharyngeal, vagus, and rarely of the oculomotor nerves may also occur. Cephalic tetanus may graduate to generalized tetanus.

Fig. 5 Brazilian patient with local tetanus confined to muscles innervated by the left VIIth cranial nerve and with trismus, showing the wound causing the infection. (By courtesy of Dr Pedro Pardal, Belém, Brazil.) (See also Plate 3.)

Tetanus neonatorum

The earliest symptom is a difficulty or inability to suckle and swallow owing to stiffness of muscles of the jaw and pharynx. There is increasing stiffness, with the classic tetanus facies (Fig. 6 and Plate 4), flexion at the elbows with the fists clenched and drawn to the thorax, extension of the knees with plantar flexion of the ankles and toes, and opisthotonos. Muscle spasms make breathing difficult; autonomic disturbances are frequent and death results from cardiorespiratory failure.

Fig. 6 Characteristic facies in neonatal tetanus. (See also Plate 4.)

Local tetanus

Rarely, rigidity and spasms may be localized to muscles adjacent to the wound or confined to a limb.

Natural history

The incubation period of tetanus (time between the injury and the first symptom) averages 7 to 10 days, but may range from 1 to 2 days to 2 months. The period of onset is the time between the first symptom and onset of spasms, and ranges from 1 to 7 days. The shorter the incubation period and the period of onset, the greater the severity of the disease. The disease peaks over 7 to 10 days, plateaus over the next 3 weeks, and then subsides over the next 2 weeks. Muscle stiffness and ankle clonus may persist for months after recovery. Severe tetanus is a markedly catabolic disease and significant weight loss is always observed even in patients who recover.

Complications

Complications in tetanus are frequent and numerous. The respiratory and cardiovascular systems are chiefly involved.

Respiratory

These include atelectasis, aspiration pneumonia, pneumonia, and bronchopneumonia. Bronchopulmonary infections are generally due to Gram-negative organisms—chiefly *Klebsiella* spp.and *Pseudomonas aeruginosa*. Prolonged laryngeal spasms if unrelieved can cause death. Severe hypoxia and respiratory failure due to incessant spasms are a certainty if these patients are not given curare-like drugs and ventilated. Episodes of severe unexplained tachypnoea and respiratory distress are probably central in origin. The acute respiratory distress syndrome can be caused by tetanus *per se* or may be due to complicating sepsis. Complications related to tracheostomy and ventilator support are also observed.

Cardiovascular and autonomic

Sustained tachycardia (more than 160/min), persistent hypotension, labile or sustained hypertension, and 'autonomic storms' complicate severe tetanus. Autonomic storms are due to wild fluctuations in sympathetic activity causing episodes of hypertension and tachycardia alternating within minutes or hours with hypotension and bradycardia. Such cardiovascular instability may be a forewarning of cardiac arrest and death. Increased vagal tone can cause bradycardia and sudden death, particularly during suctioning of tracheal secretions. Supraventricular arrhythmias, ventricular tachycardia, and infranodal conduction defects can also occur. Hyperthermia and rarely hypothermia suggest hypothalamic involvement.

Sudden death

Sudden cardiac arrest causing death remains the single most dreaded complication of moderate and severe tetanus. It may be related to cardiovascular instability due to fluctuating autonomic tone, excessive vagal activity, severe hypoxia, sudden hyperpyrexia, impaired infranodal conduction, massive pulmonary embolism, or no obvious reason.

Other complications

These are generally incidental to prolonged management of critically ill patients on ventilator support. They include: iatrogenic sepsis with multiple organ failure; gastrointestinal bleeds, ileus, or diarrhoea; renal insufficiency; fluid, electrolyte, and acid–base disturbances; fractures generally of one or more thoracic vertebrae during severe spasms; miscellaneous complications—bedsores, thrombophlebitis, rhabdomyolysis, peripheral neuropathy, corneal ulcers, anaemia, hypoproteinaemia, and deep vein thrombosis which may cause pulmonary embolism.

Diagnosis

Diagnosis is based solely on clinical features. Absence of a wound does not exclude tetanus. Trismus produced by tetanus should be distinguished from masseter spasm due to an alveolar or peritonsillar abscess. Dystonic reactions caused by phenothiazines and metoclopramide, spasms due to hypocalcaemic tetany, and seizures due to strychnine poisoning may superficially mimic tetanus. Meningitis and meningoencephalitis can also produce trismus, rigidity, seizures, and opisthotonus, but can be differentiated by a cerebrospinal examination, which is normal in tetanus. Cephalic tetanus can be mistaken for rabies because of severe dysphagia—however, hydrophobia never occurs in tetanus.

Mortality

Tetanus neonatorum carries a mortality of 60 to 80 per cent. Mortality in adult tetanus ranges from 20 to 60 per cent. It is higher in the older age group and in those with a short incubation period (under 4 days). A short period of onset (under 2 days) more reliably prognosticates severe disease. The introduction of critical care and ventilator support in severe tetanus has led to a drop in overall mortality from 30 to 12 per cent, and the mortality in fulminant tetanus from near 100 to 23 per cent, in our unit. In a good, well-equipped, critical care unit in Bombay, the mortality in severe tetanus in adults remains as low as 6 per cent.

Management

Mild tetanus (grade I) poses no serious problems. However, grade I (mild) tetanus can over a period of days graduate to grade II (moderate) or even grade III and grade IV (severe) tetanus. Such patients merit close observation. Wherever possible, all patients with tetanus should be admitted to an intensive care unit. Unfortunately this is not always feasible in poor developing countries. However, motivation and training for better patient care coupled with basic equipment for respiratory care and support can work wonders in reducing mortality even in the absence of full intensive care facilities.

The use of antiserum

Equine antiserum is generally available in poor countries; 10 000 units are given intravenously on admission after first doing a sensitivity test. However, fatal anaphylaxis can occur even when skin sensitivity is not observed. Human tetanus immunoglobulin is superior to equine antiserum, produces no hypersensitivity reactions, and if available should be given in preference to the equine antiserum in a dose of 5000 units intravenously. The intrathecal use of tetanus antiserum is best avoided as claims of its efficacy remain unproven. Local infiltration of 3000 units of antitoxin around an obvious wound is practised in some units, but its value is

uncertain. Antitoxin has no action on the toxin that has already been fixed to the nervous tissue; at best, it serves to neutralize newly liberated tetanus toxin. Antitoxin should be given before local manipulation of the wound, which is treated according to usual surgical principles with debridement of necrotic tissue and delayed primary suturing.

Antibiotics

Two mega-units of penicillin are given intravenously four times a day for 8 days. Though effective against *C. tetani in vitro*, its use in this disease is disappointing. Clindamycin or erythromycin are alternatives for penicillin-sensitive patients. Metronidazole at a dosage of 500 mg four times a day for 10 days is currently preferred. Other appropriate antibiotics may be necessary to counter secondary complicating infections.

Management strategies

Mild or grade I tetanus should be treated conservatively with the use of sedatives and muscle relaxants. Patients with grade II or moderate tetanus should, in addition to sedatives and relaxants, have a tracheostomy. Patients with grade III or IV (severe) tetanus require sedation, tracheostomy, and continuous ventilatory support after they have been paralysed with curare-like drugs, until spasms relent and recovery ensues.

Use of sedatives and muscle relaxants

The use of sedatives and muscle relaxants remains the cornerstone of management in grade I and grade II tetanus. The aim is to reduce rigidity and control spasms without significantly depressing respiration. Diazepam, a benzodiazepine and a GABA antagonist, is the drug of choice in most units. The dosage is 5 to 20 mg thrice daily in children and adults, and 2 mg thrice daily in neonates. In mild tetanus it is given orally; in moderately severe tetanus it is given in a slow intravenous infusion over 24 h. It is best not to exceed 120 to 150 mg/24 h in adults even in the presence of marked rigidity. Higher doses will inevitably depress respiration. Lorazepam with a longer duration of action and midazolam with a shorter half-life may also be used, but offer no advantage over diazepam. Chlorpromazine and phenobarbitone are second-line drugs that can be used in combination with diazepam when there is no alternative. The ideal sedative and muscle relaxant dosage schedule for each patient should be tailored to ensure continuous sedation at a level that ensures sleep, but that allows the patient to be aroused to obey commands. An objective guide, particularly in moderately severe tetanus, is relaxation of the abdominal muscles, which feel much less stiff to palpation.

Tracheostomy

Tracheostomy is mandatory for severe (grade III, IV) tetanus. Preferably, it should also be done in moderate tetanus simply because, even in an intensive care unit setting, the most important preventable cause of death is a sudden prolonged laryngeal spasm leading to asphyxia. The patient's inability to handle upper respiratory secretions in the presence of dysphagia, and the use of heavy sedation in many cases of moderately severe tetanus are both indications for elective tracheostomy.

Induced paralysis with ventilator support

Severe tetanus has a forbiddingly high mortality if management is conservative and confined to the use of high doses of sedatives and muscle relaxants. These patients require a tracheostomy, induced paralysis by curare-like drugs (pancuronium or vecuronium), and ventilatory support in an intensive care unit. In poor countries, this management strategy can also be implemented in the absence of good monitoring facilities provided the unit or ward has basic equipment for respiratory care and a trained medical and nursing staff. Results in such units may not be as good as in well-staffed and -equipped intensive care units but are still far superior to those observed with conservative management. Pancuronium is used in a dose of 2 to 4 mg intravenously every 30 min to 2 h; the dose of vecuronium is 0.1 mg/kg intravenously. Alternatively, either of these drugs may be given in a slow intravenous infusion, the dose being titrated to produce a degree of neuromuscular blockade and paralysis that allows efficient ventilator support. As the patient improves, pancuronium or vecuronium are given at longer intervals or the infusion rate is reduced. In fulminant tetanus with continuous spasms it may be almost impossible to induce complete paralysis. Twitches invariably break through within 30 min of the use of the drug, but these do not interfere with efficient ventilatory support. The average period of ventilatory support in severe tetanus is around 3 to 4 weeks, but may vary from 10 days to 6 weeks. Once spasms cease, the neuroparalytic drug is stopped; ventilator support is continued until such time as the patient is deemed fit to be weaned.

It is unnecessary, unwise, and probably dangerous in our opinion to use large doses of intravenous diazepam (a frequent practice in many units) in paralysed patients on ventilator support. A dose of 40 to 60 mg of diazepam intravenously over 24 h suffices to counter anxiety without dangerously depressing vital centres.

In poor countries the correct mode of management in severe tetanus is often constrained by a paucity of trained staff and material resources (particularly ventilators). Ethical considerations then restrict the use of ventilator support to: (i) patients with grade IV tetanus; (ii) patients with grade III tetanus whose spasms are uncontrolled on a conservative regime and whose $P(a)O_2$ is less than 55 mmHg on 6 to 8 litre/min of oxygen; and (iii) patients who develop any serious complication that in itself merits ventilatory support.

Treatment of autonomic circulatory disturbances

Intravenous b-blockers, heavy sedation, intravenous morphine sulphate, intravenous labetalol, intravenous infusion of magnesium sulphate, and intravenous clonidine have all been used to control autonomic storms in severe tetanus. These drugs do not alter the high mortality. It is best to rely on good overall critical care and efficient cardiorespiratory support and avoid drugs that strongly depress the central and autonomic nervous systems. Hypotensive spells are treated by a volume load and if this is ineffective or contraindicated, by inotropic support with dopamine or dobutamine to maintain a systolic pressure just above 100 mmHg. Hypertensive episodes with a systolic blood pressure in excess of 200 mmHg or a diastolic in excess of 100 mmHg are treated with a small oral dose of propranolol (10 mg) or 5 mg of sublingual nifedipine. Intravenous propranolol is dangerous and can cause sudden death, but a titrated infusion of esmolol (a b-blocker with a very short half-life) is useful in a hypertensive crisis. Bradyarrhythmias are treated with intravenous atropine, and persistent sinus tachycardia of more than 170/min with 40 mg of verapamil orally twice or thrice daily. In this situation it is best not to use more than 50 mg of diazepam ntravenously per day. Sedatives and drugs used at a dosage that strongly depresses the central or autonomic nervous system probably contribute to a high mortality by predisposing to cardiac arrest (particularly after sudden hypotensive spells and sudden bradyarrhythmias) and by preventing successful resuscitation. These management principles have achieved a very low mortality (6 per cent) in severe tetanus.

Treatment of other complications

Complications may involve almost every system in the body. These should be promptly recognized and treated.

Critical care and nursing

Good critical care and expert nursing play a vital role in reducing complications and preventing death. The following are of particular importance:

1. ensuring patency of the airway and care of the tracheostomy in scrupulous detail;
2. ensuring an adequate arterial oxygen saturation and oxygen transport to tissues through efficient cardiorespiratory support;
3. expert physiotherapy to the chest timed specifically during periods of drug-induced muscle relaxation;
4. maintaining fluid, electrolyte, and acid–base balance;
5. prevention, early detection, and prompt control of infection and sepsis with appropriate antibiotics;
6. supporting nutrition, if necessary by intravenous alimentation;
7. detection (by frequent monitoring) and treatment (by physical methods and by paracetamol) of hyperpyrexia—a surreptitious killer in tetanus.

Use of tetanus toxoid

Tetanus does not confer immunity. Active immunization is necessary and is achieved by giving the first dose of tetanus toxoid during convalescence and the next two doses at the recommended intervals.

Prevention

Tetanus is a preventable disease through active immunization with adsorbed tetanus toxoid (**ATT**) and by proper management of wounds.

Active immunization

The following regime is advocated. In infancy and childhood, three doses of triple vaccine (tetanus, diphtheria, pertussis) are given at monthly intervals, and a booster dose is given at 4 to 6 years of age. Unimmunized individuals older than 7 years should be given triple vaccine in three doses: the first and second 6 weeks apart, and the third dose 6 months after the second. Booster doses of ATT are advocated every 10 years, but this remains a practical impossibility in poor countries.

Immunization after minor, uninfected wounds

Passive immunization with equine or human tetanus antitoxin is not indicated for minor, clean wounds. Active immunization (with 0.5 ml ATT) is indicated if immunization status is unknown or more than 10 years have elapsed after the last dose of ATT. Under the above circumstances, especially in poor countries, ATT is also administered prior to emergency surgery, deliveries, and obstetric procedures.

Immunization after infected or major wounds

Passive immunization (250 to 500 units of human tetanus immunoglobulin or 5000 units of equine antitoxin, intramuscularly) is recommended in all individuals who are not immunized, partially immunized, or whose immunization status is unknown. Indication for administration of ATT is the same as for minor, uninfected wounds. However, it would be safer to use a booster dose of ATT even in a well-immunized individual if more than 5 years have elapsed since the last dose of ATT.

Prevention of tetanus neonatorum

Primary immunization of pregnant patients with two injections of ATT a month apart, preferably during the last two trimesters, together with education of nurses and midwives on sterile obstetric techniques, would further reduce the incidence of tetanus neonatorum in developing countries.

Further reading

Gupta SD, Keyl PM (1998). Effectiveness of prenatal tetanus toxoid immunization against neonatal tetanus in a rural area in India. *Pediatric Infectious Diseases Journal* **17**, 316–21.

Park K (1997). Tetanus. In: Banarsidas Bhanot, ed. *Park's textbook of preventive and social medicine*, 15th edn, pp 237–40. Jabalpur, India.

Sutton DN *et al*. (1990). Management of autonomic dysfunction in severe tetanus: the use of magnesium sulphate and clonidine. *Intensive Care Medicine* **16**, 75–80.

Udwadia FE (1994). *Tetanus*. Oxford University Press, Bombay.

Udwadia FE *et al*. (1987). Tetanus and its complications: intensive care and management experience in 150 Indian patients. *Epidemiology and Infection* **99**, 675–84.

7.11.21 Botulism, gas gangrene, and clostridial gastrointestinal infections

H. E. Larson

Botulism

Definition

Botulism is an acute, symmetrical, descending paralysis caused by a neurotoxin produced by *Clostridium botulinum*. Food contaminated by *C. botulinum* spores and elaborated toxin produces illness when ingested. Wound infections with *C. botulinum* or intestinal tract colonization in infants and adults occasionally cause botulism.

Occurrence

C. botulinum is ubiquitously distributed in soil and mud. The surfaces of potatoes, vegetables, and other foods are easily contaminated with spores, which survive brief heating at 100°C. The anaerobic conditions characteristic of canning, smoking, or fermentation facilitate clostridial growth and toxin release. Spores germinate in sausage or cheese if they are kept for extended periods at room temperature. An eighteenth century report associated paralytic illness with eating sausages, hence *botulus*, Latin for a sausage. Cases have been associated with fermented milk in Africa, cheese sauce on baked potatoes in North America, fermented stew in Japan, and imported fish in the United Kingdom.

Although past outbreaks typically involved small groups of people, home-canned peppers served in a restaurant caused two large outbreaks in the United States. Outbreaks caused by commercially processed foods are infrequent, but contamination of hazelnut purée added to commercially produced yoghurt caused 27 cases of botulism in Wales and north-west England in 1989, the largest recorded outbreak in the United Kingdom. Most of the contaminated cartons could not be accounted for, suggesting that the attack rate varied or that mild symptoms were not diagnosed as botulism. Commercially prepared, chopped garlic in soybean oil caused 36 cases dispersed over eight provinces and states in North America.

Some outbreaks involved only single contaminated items, such as in the Loch Maree episode in 1922 where eight people died after eating duck paste, the 1978 outbreak in Birmingham involving four people who ate tinned Alaskan salmon, and one case in 1989 following a meal on a commercial airliner. Uneviscerated fresh fish have been associated with botulism, usually where there have been deficiencies in refrigeration.

Purified botulinum toxin has recently come into therapeutic use. Toxin injections produce temporary muscle weakness in the treatment of strabismus, blepharospasm, torticollis, and for cosmetic purposes. Treatment doses are considered too small to account for systemic symptoms. Under experimental conditions aerosolized botulinum toxin causes illness in monkeys and the toxin has been mooted as an effective agent for biological warfare or terrorist activity. Botulinum toxin was loaded into SCUD missile warheads by Iraq during the Gulf War and stockpiled by the Aum Shinrikyo cult in Japan.

The toxin

There are seven serological types of botulinum toxin (A–G). Types A, B, and E account for nearly all human cases. Serotypes implicated in outbreaks of botulism parallel the geographical distribution of soil spores. Type E is nearly always associated with fish, but outbreaks caused by fish products involve types A and B. Spores of *C. botulinum* can survive up to 2 h of boiling (100°C), but are killed rapidly at autoclave temperatures (120°C).

C. botulinum toxin is heat labile and rapidly inactivated at ordinary cooking temperatures. It is a protein neurotoxin, and a dose as small as 0.1 µg has been estimated to cause death in a human being. The 150-kDa molecule is composed of two peptide chains connected by disulphide bonds. One chain binds to and penetrates the neurone, the other cleaves a protein essential for neurotransmitter release, reducing acetylcholine availability for impulse transmission. Toxin types A, C, and E hydrolyze a protein in the presynaptic membrane while types B, D, F, and G hydrolyze a protein in the synaptic vesicle.

Pathogenesis

Botulinum toxin is absorbed directly across mucous membranes. Locally acting toxin may produce some symptoms but cranial nerve paralysis results from blood stream distribution. Cranial nerves are preferentially affected because botulinum toxin binds more rapidly to sites where the cycles of depolarization and repolarization are frequent. Binding is irreversible and the toxin cannot thereafter be neutralized by antitoxin. Recovery occurs when nerve terminals sprout from the axon to form new motor end-plates.

Botulinum toxin blocks impulse transmission mediated by acetylcholine at myoneural junctions, at autonomic ganglia, and at parasympathetic nerve terminals. Transmission is blocked because the toxin prevents release of acetylcholine from the presynaptic membrane. Impulse conduction within peripheral nerves and muscle contraction are not affected. Synthesis of acetylcholine and impulse transmission within terminal nerve fibrils remain intact. On the other hand, the miniature end-plate potentials spontaneously generated by release of acetylcholine in a resting nerve decrease and eventually disappear in the presence of toxin. If a poisoned nerve is stimulated repetitively, temporary summation of acetylcholine release occurs, producing an augmented response.

History

The symptoms of botulism vary from mild fatigue to severe weakness and collapse leading to death within a day. Initially, nausea, vomiting, abdominal bloating, and dryness in the mouth and throat may suggest gastrointestinal tract illness. Diplopia, blurred vision, dizziness, unsteadiness on standing, and difficulty with speech or

swallowing are common early neurological symptoms. Subsequently, there is progression to weakness or paralysis in the limbs, and generalized weakness and lassitude. The dryness of the mouth and throat may become so severe as to cause pain. Eventually there may be difficulty holding up the head, constipation, urinary hesitancy, and problems in breathing. The incubation period is between 12 and 72 h. Patients with short incubation periods are likely to have ingested large amounts of toxin. However, individuals are known to have ingested large amounts of contaminated food without developing symptoms.

Physical examination

Negative findings in botulism are pertinent. Higher mental functions are preserved, although sometimes patients are drowsy. Sensation is intact. Fever is unusual. The mouth is dry and the tongue is furrowed. Lateral rectus weakness in the eyes produces internal strabismus. Failure of accommodation is common and the pupils may be fixed in mid position or dilated and unresponsive to light. Ptosis, weakness of other extraocular muscles, and inability to protrude the tongue or to raise the shoulders are other early findings. Weakness in the limbs is of the flaccid, lower motor neurone type and deep tendon reflexes are initially preserved. Facial muscles may be spared; gag and corneal reflexes are not lost.

Weakness of the respiratory muscles develops early in relation to other findings and deterioration can be rapid. Paralysis descends symmetrically from cranial nerves to upper extremities to respiratory muscles to the lower extremities in a proximal to distal pattern. Hypotension without compensatory tachycardia, intestinal ileus, and urinary retention are evidence of the widespread autonomic paralysis. Symptoms and signs can be confined to the autonomic nervous system.

Diagnosis

The diagnosis in the first case of an outbreak can be missed because cranial nerve symptoms and signs are ignored in what is apparently a gastrointestinal disturbance. The differential diagnosis usually lies between botulism and the descending form of acute inflammatory polyneuropathy or Guillain–Barré syndrome. There can be similarities in the clinical presentation and progression of symptoms in the two diseases. Patients with botulism have normal cerebrospinal fluid findings and respiratory weakness and failure develop early, prior to the presence of severe limb weakness. Patients with the Guillain–Barré syndrome have marked limb weakness prior to the development of respiratory failure. Sensation and mental status are preserved in botulism.

Other diagnoses that may be considered include diphtheria, intoxication with atropine or organophosphorus compounds, myasthenia gravis, cerebrovascular disease involving the brainstem and producing bulbar palsy, paralytic rabies, tick paralysis, and neurotoxic snake bite. Botulism is distinguished from polymyositis and periodic paralysis by its rapid progression and cranial nerve abnormalities. Sometimes patients with other types of poisoning are thought to have botulism, most often with an outbreak of staphylococcal food poisoning. Individuals with carbon monoxide poisoning have been mistakenly been thought to be poisoned by food, but they invariably have headaches and altered consciousness. Poisoning from chemicals or fish produces rapid onset of symptoms. Mushroom poisoning is characterized by severe abdominal pain.

The diagnosis of botulism can be confirmed by testing for botulinum toxin in the patient's serum, urine, stomach contents, or in the suspect food. Mice are inoculated intraperitoneally with 0.5 ml of sample, with and without mixing with polyvalent botulinum antitoxin, and observed for signs of botulism. Electromyography can be helpful in confirming a diagnosis of botulism. Single or low-frequency stimuli evoke muscle action potentials that are reduced in amplitude; tetanic or rapid stimuli produce an enhanced response. Nerve conduction velocities are normal. This result readily differentiates botulism from the Guillain–Barré syndrome. Patients with myasthenia gravis usually have muscle action potentials of normal or minimally decreased amplitude.

Treatment

The priorities in management are assessment of respiratory function followed by administration of antitoxin. Respiration should be monitored closely with a view to elective intubation since deterioration can occur rapidly. Prolonged respiratory support may be required. Profound hypotension can be secondary to hypoxaemia, acidosis, and accumulated fluid deficits or be a feature of the autonomic paralysis. Treat autonomic paralysis by expanding the intravascular volume using whole blood, protein, and/or saline while monitoring central venous pressure or by infusing low dose dopamine.

Trivalent (types A, B, and E) antitoxin has been shown to reduce case fatality and shorten the course of the illness. To be useful it must be given early, before free circulating toxin has bound to its peripheral targets and before the diagnosis can be confirmed by animal tests. Multivalent equine antitoxin is available from designated regional hospitals in the United Kingdom; half the dose is given intramuscularly and half intravenously. An intradermal 0.1-ml test dose is given, but most serum reactions are not predicted by this test. Human botulism immune plasma can be obtained from the Centers for Disease Control, Atlanta, Georgia, United States.

Many years ago it was shown that patients dying of botulism carried bacilli in their intestine. The discovery that clinical disease can result from toxin formed within the gastrointestinal tract of infants and adults makes antimicrobial treatment theoretically appealing. Gastric lavage, repeated high enemas, and cathartics have been given to attempt to remove unabsorbed toxin. Drugs capable of reversing neuromuscular blockade have been used to treat patients with botulism, but without any noticeable effect on respiratory muscle weakness or tidal volume.

The mortality from botulism in the early part of the twentieth century was 60 to 70 per cent, but this improved to 23 per cent for cases reported between 1960 and 1970 since the use of respiratory support. In a single, large outbreak in 1977 there were no deaths among 59 cases. Recovery from botulism depends upon the formation of new neuromuscular junctions; clinical improvement thus takes weeks to months. One severe case required respiratory support for 173 days with eventual recovery. Very prolonged fatigue and dyspnoea on exertion can be due to factors other than the neuromuscular blockade.

Wound botulism

Symptoms and signs of botulism can develop in people with injuries. Recognition may be complicated by the presence of fever from wound infection or gas gangrene, or by the absence of gastrointestinal symptoms. The diagnosis is confirmed by electromyography; botulinum toxin is detected in serum in only about half of the reported cases. The incubation period averages 7 days with a range of 4 to 17 days. Clinical findings and management are the same as for patients with food-borne botulism. Since 1991, wound botulism has increasingly become a complication of injection drug abuse; small abscesses at injection sites yield *C. botulinum.* An epidemic of wound botulism in the United States has been associated with the injection of black tar heroin. *C. botulinum* can be recovered from wounds in the absence of clinical botulism.

Infant botulism

Sporadically, cases of botulism are recognized in infants under 6 months of age. Previously healthy babies develop constipation, which progresses over 3 to 10 days to poor feeding, irritability, a hoarse cry, and weakness in head control. Examination shows a generally weak, hypotonic, afebrile infant. Abnormalities in eye movements and pupillary reactions are sometimes present and deep tendon reflexes are reduced or absent. There is considerable range in severity; respiratory failure can develop but most recover completely.

The diagnosis can be confirmed by finding *C. botulinum* and toxin in the faeces, and by electromyography. Botulinum toxin is not present in the serum. The disease is thought to follow ingestion of *C. botulinum* spores, which multiply in the infant's gastrointestinal tract and produce toxin. Excretion of *C. botulinum* and toxin may continue for as long as 3 months. Honey has been a source of spores for some cases. Other than supportive measures, no consistent pattern in treatment using antitoxin, antibiotics, cathartics, or enemas has been established.

Gas gangrene

Definition

Gas gangrene is a rapidly developing and spreading infection of muscle by toxin-producing clostridial species, especially *C. perfringens* (formerly known as *C. welchii*). It is accompanied by profound constitutional toxicity and is invariably fatal if untreated.

Aetiology

Although gas gangrene conjures up visions of battlefield injury, cases occur after civilian and iatrogenic trauma. Disease occurrence depends upon a conjunction of factors. Viable forms of clostridia must be present and the wound environment must be conducive to their growth. Proximity to faecal sources of bacteria is a risk factor, as in hip surgery, adrenaline injections into the buttock, and amputation of the leg for ischaemic vascular disease. Wound contamination with dirt, shrapnel, or bits of clothing reduces local oxygen concentrations. Similarly, wounds involving large muscle masses in the shoulder, hip, thigh, and calf, damage to major arteries, crush injuries, open fractures, and burns carry a higher risk. High-velocity missiles and impacts are regular features of modern injuries in both wartime and civilian life and such injuries produce extensive tissue damage.

The incidence of gas gangrene after trauma reflects the speed at which injured people can be evacuated and receive appropriate treatment. During the Vietnam and Falklands conflicts there were very few cases of gas gangrene among American and British wounded cared for by highly organized surgical teams. In comparison, when a jet airliner crashed in the Florida everglades, eight of the 77 injured survivors developed the disease.

Gas gangrene is caused by anaerobic, Gram-positive, spore-forming bacilli capable of producing potent exotoxins. Most cases are caused by *C. perfringens* type A, but some are due to *C. novyi* and a few to *C. septicum*. *C. histolyticum*, *C. sordelli*, and *C. fallax* cause few cases and not uncommonly more that one species is isolated. Clostridia are mainly saprophytes, occurring naturally in soil and in the gastrointestinal tracts of man and animals. Oxygen inhibits their growth and prevents toxin production. Possession of superoxide dismutase can permit the organisms to survive in the presence of small amounts of oxygen. Necrotic tissue, foreign bodies, and ischaemia in a wound reduce the locally available oxygen. Infrequently, gas gangrene occurs without preceding trauma. It can be a primary infection of the perineum or scrotum, or present in a limb, secondary to seeding from clostridial colonization of a colonic neoplasm. *C. septicum* is found in a higher percentage of these cases than where there is a history of trauma.

C. novyi and other clostridia cause soft tissue infections at injection sites in drug addicts. An epidemic of these infections was reported in Scotland, Ireland, England, and the United States in 2000 associated with hypotension, severe constitutional toxicity, and a high case fatality rate.

Toxins

The clostridia responsible for gas gangrene elaborate a wide range of toxin activities, with from four to more than 12 separate toxins described for *C. septicum*, *C. novyi*, and *C. perfringens*. The principal toxin of *C. perfringens* is a toxin; the toxic action has been shown to be due to an ability of the molecule to insert into and interact with a phospholipid membrane. Electron microscopy shows gaps of 7.5 to 18 nm appearing in the plasma membrane as early as 1 h. These plasma membrane defects increase with time and can be visualized adjacent to toxin molecules that have been labelled with ferritin. Toxin is not detected in the tissues or serum of patients with gas gangrene, possibly because the toxin binds rapidly and irreversibly.

History

The incubation period of gas gangrene is usually less than 4 days, often less than 24 h, and occasionally as short as 1 to 6 h. Pain is the most characteristic symptom. Patients describe this as severe or excruciating and sudden in onset. Evolution of symptoms and signs can be very rapid. Toxicity may prevent the patient from giving an adequate history.

Physical examination

Early on it may be difficult to account for the patient's pain by objective physical findings. Swelling, bluish discoloration, or darkening of the skin occurs at the affected site. The traumatic or surgical wound becomes oedematous and a thin, serous ooze emerges. Pain steadily increases in severity: the overlying skin becomes stretched and develops a brown or 'bronzed' discoloration. Haemorrhagic vesicles and finally areas of frank necrosis appear. A sweet odour from the wound has been described. In spite of the name, gas is not invariably present, especially early. Later, crepitus and exquisite tenderness are present in the wound.

Profound constitutional changes occur. Patients become sweaty and febrile, and though alert and oriented, are very distressed. The pulse is elevated out of proportion to the fever. Death may occur within 48 h. At operation, infected muscle appears dark red with purple discoloration; frank gangrene and liquefaction may be seen. Involved muscle does not contract after direct stimulation.

Clostridial myonecrosis must be distinguished from anaerobic cellulitis and from anaerobic streptococcal myositis. Anaerobic cellulitis occurs where putrefying anaerobic clostridia produce a purulent infection in traumatized muscle and other tissues. Streptococcal myositis is a spreading muscle infection with anaerobic streptococci and either *Streptococcus pyogenes* or *Staphylococcus aureus*. Neither is associated with the constitutional toxicity characteristic of gas gangrene and neither requires as radical excision. Diabetic patients develop gas gangrene due to ischaemic vascular disease. Numerous micro-organisms, both aerobic and anaerobic, produce gas in tissues.

Diagnosis

The diagnosis of gas gangrene has to be made on clinical grounds. Prompt recognition and treatment improves the prognosis. Sudden deterioration in a postoperative patient or following trauma requires examination of the wound and surrounding tissue. Cases of primary gas gangrene and cases following elective surgery may have a higher fatality because recognition is delayed. Gram stain of the wound discharge, of an aspirate, or of a needle biopsy may aid diagnosis. In gas gangrene there are many large, plump, Gram-positive bacilli, usually without spores. Few, if any, polymorphonuclear leucocytes are present. On the other hand, both anaerobic streptococcal myositis and anaerobic cellulitis show many leucocytes and the former is characterized by long chains of Gram-positive cocci.

CT scanning can detect gas deep in muscle, but the absence of gas does not exclude the diagnosis. Culture of clostridia does not confirm a diagnosis of gas gangrene, as simple colonization without clinical disease occurs in up to 30 per cent of wounds. Efforts to establish a portal of entry for cases of spontaneous, non-traumatic gas gangrene may improve the prognosis.

Treatment

Surgical removal of all affected muscle is essential. Although not substitutes for surgery, antimicrobials, hyperbaric oxygen, and administration of antitoxin have been thought to be helpful adjunctive therapies. Penicillin has been the drug of choice, but there is experimental evidence that clindamycin and metronidazole might be superior to penicillin, perhaps by inhibiting toxin production. This has led to the use of penicillin and clindamycin as combination therapy. Ceftriaxone or erythromycin are alternative choices for severely penicillin-allergic patients.

Hyperbaric oxygen is used to treat gas gangrene. An effect on mortality has never been shown by controlled trials, and comparable mortality rates have been achieved without using it. One hundred per cent oxygen is given at 303 kPa for 60 to 120 min, two to three times daily. Therapeutic administration of gas-gangrene antitoxin made from horse serum is controversial. Use during the Second World War reduced mortality but serum sickness and other allergic reactions occur. It is no longer produced in the United States. Shock, blood loss, dehydration, and septicaemia with micro-organisms such as *Escherichia coli* should be treated appropriately. *C. perfringens* septicaemia in association with gas gangrene is not common.

Prevention

The mortality of established disease still ranges between 11 and 31 percent. Prophylactic antibiotic treatment effectively eliminates this risk. A first generation cephalosporin is given intravenously before surgery and for three doses postoperatively. Metronidazole may be useful in patients who are hypersensitive to b-lactam antibiotics. Antibiotic levels can be detected in ischaemic tissues.

Traumatic wounds are treated to eliminate the conditions that allow gas-gangrene bacilli to grow. High-velocity missiles distribute energy radially from their path, producing more extensive tissue damage than missiles at low speeds or with a small mass. Wounds should be excised widely by resection back to healthy, viable muscle and skin. Closure is delayed for 5 to 6 days until it is certain that the wound is free of infection. Military surgeons usually give penicillin in high dosage over

this period. Experimentally, active immunization protects, but in man this requires the clear definition of risk categories.

Clostridial infections of the gastrointestinal tract

Pseudomembranous colitis

Definition

Pseudomembranous colitis is an acute exudative infection of the colon caused by *C. difficile*. The name derives from plaques of necrotic membrane that adhere to the mucosal surface in the clinically most severe form of the disease.

Aetiology

Pseudomembranous colitis was described as a clinical and pathological entity in 1893 with its clostridial aetiology becoming known in 1977. *C. difficile* is an anaerobic, spore-forming, bacillus found in the environment. Healthy adults are only rarely colonized with *C. difficile*. Antimicrobial treatment reduces resistance to intestinal colonization. Colonization and toxin production produce colitis. Because colonization and antimicrobial treatment may occur at different times, antibiotic-susceptible strains of *C. difficile* are able to produce disease. Resistance to colonization requires viable intestinal bacteria, but it is not known which species or combination of species determines this resistance. Usually resistance to colonization will spontaneously reconstitute itself unless an antimicrobial effect persists within the gut. Infants and young children can be asymptomatically colonized even in the absence of antimicrobial treatment.

Clinical history

The single most pertinent detail of the medical history is previous antimicrobial treatment. Direct questioning may be needed to elicit this history; antimicrobials may have been self-administered, taken for trivial complaints, or used as long as 3 or 4 weeks before the start of diarrhoea. Pseudomembranous colitis has been reported to follow the use of every antimicrobial in common medical practice, but its association with lincomycin, clindamycin, ampicillin, amoxacillin, and cephalosporins is strongest. It occasionally occurs in individuals with no history of antimicrobial treatment or as a complication of chronic colonic obstruction, carcinoma, leukaemia, or uraemia. Pseudomembranous colitis was identified as a pathological entity before any clinical use of antimicrobials. Community-acquired cases occur sporadically but case clustering in hospitals or nursing homes is not uncommon. The disease is more common in older patients but the typical syndrome has been described in people of all ages including infants.

Initial symptoms vary from mild, self-limiting diarrhoea to acute fulminating toxic megacolon. Illness can begin surreptitiously where persistent diarrhoea resists all efforts at symptomatic relief. Community-acquired cases tend to have a week or more of diarrhoea before seeking medical attention. Stools are described as watery or porridge-like, or patients may be obstipated. Other initial symptoms are sudden chills, fever, and signs of an abdominal catastrophe. Elderly patients may have diarrhoea that resolves and then recurs at intervals of one to several days. Severe abdominal pain is not common and a history of frank blood in the stools suggests a different type of colitis.

Physical examination

Elderly patients appear tired, toxic, and ill. Low fever, a dry furred tongue, and abdominal tenderness, sometimes with peritonism, are the most common clinical signs. Signs of dehydration may be present, but hypotension attributable to hypovolaemia is not common. Spiking temperatures may also be seen and a distended, tense, abdomen can suggest colonic obstruction. Reactive arthritis, IgA nephropathy, and hypoproteinaemia are potential complications of *C. difficile* colitis.

Diagnosis

Many patients show polymorphonuclear leucocytosis, sometimes with counts of 30 000/ul or more. Leucocytes are present in the faeces. Chemical findings in patients with prolonged diarrhoea include azotaemia and hypoalbuminaemia; the azotaemia may appear to be out of proportion to the dehydration. The presence of *C. difficile* toxin establishes a mechanism for the diarrhoea.

Sigmoidoscopy can be helpful in making an early diagnosis because the raised, mucoid to opaque yellow plaques (0.2–2 mm across) are diagnostic. If the mucosa appears normal, biopsy and multiple sectioning may reveal microscopic lesions. Some patients with *C. difficile* colitis do not have pseudomembranes, either because lesions are distributed unevenly in the colon or because the illness is mild. In these cases the diagnosis can only be confirmed by testing for toxin and *C. difficile*. Rarely, patients with pseudomembranes on sigmoidoscopy or rectal biopsy may fail to yield *C. difficile*. Usually confluent rather than focal mucosal necrosis is found. This appears to be the end result of several types of colonic mucosal injury, not specific to *C. difficile* infection.

The differential diagnosis of pseudomembranous colitis includes other forms of antimicrobial-associated colitis, diarrhoea due to *Salmonella*, *Shigella*, and *Campylobacter* species, intestinal amoebiasis, Crohn's disease, and non-specific ulcerative colitis. These can be differentiated by sigmoidoscopy and rectal biopsy, or by microscopy and culture of the faeces. Two-thirds or more of patients with simple antimicrobial-associated diarrhoea do not have infection with *C. difficile*. Often they complain of sudden abdominal pain and bloody diarrhoea that subsides within a day or two of stopping antimicrobial treatment. Occasionally, patients may be infected with *C. difficile* in addition to another micro-organism capable of causing diarrhoea. Infection with *C. difficile* may exacerbate symptoms in some patients with inflammatory bowel disease.

Treatment

Stopping the associated antimicrobial may allow *C. difficile* colitis to resolve spontaneously. If clinical circumstances dictate active treatment, the antimicrobial of choice is one to which *C. difficile* is susceptible and which is not absorbed following oral administration. Vancomycin is used in a dose of 125 mg every 6 h. Metronidazole, 250 mg four times a day, also appears to be effective, although it is absorbed. Some physicians regard it as less effective than vancomycin. Bacitracin may also be useful. Severe cases usually show improvement after 48 h of treatment and signs and symptoms rapidly return to normal. Failure to respond to vancomycin suggests that the diagnosis is incorrect or that an additional condition or complication may be present.

Patients who are dehydrated need fluid resuscitation. Cholestyramine resins bind *C. difficile* toxin *in vitro*, but have no effect on the clinical course of the colitis. Pseudomembranous colitis has been successfully treated by colectomy. However, the disease is completely reversible by appropriate antimicrobial treatment. In patients who are unable to take vancomycin orally, some physicians have attempted to instil it into the colon via a caecostomy tube; others combine intragastric vancomycin, intermittent clamping of the nasogastric tube, and parenteral metronidazole. *C. difficile* antitoxin is not available in the United Kingdom.

Any of the suggested antimicrobial treatment regimens for pseudomembranous colitis may be followed by relapse. The relapse illness can be clinically more severe than the original. There has never been any evidence that relapse is due to antimicrobial resistance and patients continue to respond to treatment with the original or an alternative drug. Patients relapse both because antimicrobial treatment may not completely clear them of *C. difficile* or because a new exposure to environmental strains has occurred. There is evidence that vancomycin and metronidazole themselves can reduce resistance to the infection; prolonged treatment may produce prolonged susceptibility. On the other hand, patients whose *C. difficile* colitis resolves without antimicrobial treatment usually do not relapse.

Occasional patients may have multiple relapses and many regimens have been suggested for their management. These include tapering doses of vancomycin, a *Lactobacillus* preparation three times a day, or cholestyramine three times a day after a therapeutic course of vancomycin. Cholestyramine can not be combined with vancomycin. In a patient recovering from multiple relapses, tapering vancomycin doses to once daily when diarrhoea stops, then to alternate days, then to progressively longer intervals, can prevent early relapse. Some patients with severe colitis or multiple relapses may continue to have diarrhoea without toxin in their stools. This resembles postdysenteric colitis where continued diarrhoea is due to lingering mucosal injury. Bowel rest with total parenteral nutrition can allow healing and recovery; continued treatment against *C. difficile* is not required. Normal flora may be reconstituted by giving a suspension of normal faeces as an enema.

It may be necessary under certain circumstances to continue an antimicrobial when a patient has developed pseudomembranous colitis. There is no evidence to suggest that concurrent therapy with vancomycin will not be successful, although clinical improvement occurs more slowly. It is reasonable to replace a drug commonly associated with pseudomembranous colitis by one which is not, such as a quinolone, aminoglycoside, tetracycline, or sulphonamide. Repeat treatment with

an inducing antimicrobial at some later time is not contraindicated in a patient who has recovered from pseudomembranous colitis.

Prevention

Clusters of cases of pseudomembranous colitis were reported before its infectious aetiology was understood. Now it is known that *C. difficile* may contaminate the environment of a patient, that patients acquire the organism, and that cross-infection is confirmed by strain typing. The chain of infection for isolated cases may be difficult to trace because spores can persist for months. Since patients receiving antimicrobial treatments are at risk, those with colitis ought to be nursed in barrier isolation. Patients with diarrhoea, especially those who are incontinent, are more important sources of cross-contamination than those with formed stools. Physical cleanliness, enteric precautions, confinement to a single room, and reduced use of the most frequent inducing antimicrobials are the approaches most often used to reduce institutional cross-infection. There is no proven value in retesting patients until they are free of toxin nor in treating asymptomatic toxin excretors.

Necrotizing enterocolitis

Definition

Necrotizing enterocolitis is a fulminating clinical illness characterized by extensive necrosis of the intestinal mucosa and wall. Terms such as darmbrand (Germany), enteritis necroticans, pig bel (Papua New Guinea), or gas gangrene of the bowel describe geographical variants. Cases occur sporadically in adults or as epidemics in all ages. Necrotizing enterocolitis occurs in infants, sometimes in clusters, but is not proven to be due to clostridial infection.

Aetiology

C. perfringens (*C. welchii*) is considered to be the cause. Sporadic cases usually yield *C. perfringens* type A. Gram stain of the necrotic mucosa and the bowel wall shows many Gram-positive bacilli. However, in the German and especially in the Papua New Guinea outbreaks, there is substantial evidence implicating *C. perfringens* type C. Type C produces large amounts of b-toxin, which has lethal and necrotizing effects. Papua New Guinea highlanders have a high prevalence of antibodies to b-toxin; antibodies are rare in people who live where the disease is uncommon. Patients with pig bel have rising levels of antibodies to b-toxin, and specific passive or active immunization has been shown to prevent disease. It is not clear whether exogenous human infection with these organisms occurs or whether the lesions are produced by the overgrowth of endogenous clostridia. Sweet potato, a local dietary staple, contains an inhibitor of trypsin. Combined with a low-protein diet this may impair the ability of the intestine to inactivate endogenously produced b-toxin. However, the methods used for roasting the pigs offer many opportunities for clostridial contamination.

History and physical examination

Sporadic cases, over 50 years of age or recovering from gastric surgery, are regularly reported from Scandinavia, Europe, the United States, Australia, and the Middle East. Alternatively, epidemic outbreaks as described in post-war Germany and among the highlanders of Papua New Guinea follow ingestion of contaminated food or a dramatic change in eating habits. Symptoms develop suddenly in someone who was previously well. There is severe abdominal pain, which is colicky at first and afterwards becomes continuous. Bloody diarrhoea and vomiting may occur. The patient may be extremely toxic and go into shock. On examination there is fever, with abdominal distension, localized or diffuse tenderness, and reduced bowel sounds. A tender mass may be palpated. Later, malabsorption or chronic partial obstruction may develop because of intestinal scarring.

Treatment and prevention

Patients with suspected pig bel should be treated with nasogastic suction and intravenous fluids. Pyrantel is given by mouth and the bowel rested by fasting. One megaunit of benzylpenicillin is given intravenously every 4 h and the patient observed for surgical complications. Mild cases recover without surgical intervention, but if surgical indications are present, the mortality ranges from 35 to 100 per cent. As pig bel continues to be a common disease in Papua New Guinea, consideration should be given to the use of a *C. perfringens* type C toxoid vaccine in local areas. Two doses spaced 3 to 4 months apart have been shown to prevent the disease.

Clostridium perfringens food poisoning

Occurrence and clinical findings

In the United Kingdom and the United States, food poisoning caused by *C. perfringens* is the third most common type of food-borne illness. Meat and poultry are responsible for at least 90 per cent of the outbreaks, which occur where food is prepared in large quantities. Two-thirds of the reported outbreaks are in schools, hospitals, factories, restaurants, or catering establishments, and in a typical outbreak 35 to 40 people are affected. An estimated 12 000 cases were associated with a single out-break in 1969.

The circumstances surrounding an outbreak repeat themselves with monotonous regularity. A meat dish is prepared by stewing, braising, boiling, or steaming and this is allowed to stand at ambient temperatures for a period of 4 to 24 h. The food is served cold or after desultory rewarming. Six to 12 h after eating the meal, the victims complain of crampy abdominal pain and then diarrhoea. Vomiting is unusual and fever inconsequential. Twelve to 24 h later the diarrhoea and pain have subsided. Fatal cases occur rarely; at autopsy they show severe enterocolitis.

Undoubtedly many cases of *C. perfringens* food poisoning occur at home but are not reported. Antibodies to the toxin mediating the symptoms are very common and it is likely that nearly everyone has experienced this disease once or more in their lifetime.

Aetiology

C. perfringens is an ubiquitous, sporulating anaerobe with an unparalleled virtuosity for production of biologically significant toxins. The clinical effects of infection with any particular strain may depend largely on its toxin-producing capacity. Strains associated with food poisoning have a number of special characteristics. They are type A, although their production of a-toxin is variable; they are often heat resistant. Eighty-six per cent of food-poisoning strains produce a specific, heat-labile enterotoxin. Toxin production *in vitro* is closely associated with sporulation rather than with the multiplication of vegetative cells. *In vivo*, toxin probably acts by damaging enterocyte membranes. Free enterotoxin has been detected in diarrhoeal stool after *C. perfringens* food poisoning, antibody to enterotoxin increases after such episodes, and ingestion of 8 to 12 mg of enterotoxin by volunteers produces abdominal pain and diarrhoea.

C. perfringens is a normal human faecal organism, is regularly found in the intestinal tract of domestic animals, often contaminates raw meat, and can be carried by flies. The distribution of enterotoxin-producing strains may be more restricted. However, surface contamination of meat with *C. perfringens* is common and subsequent rolling or grinding will distribute these organisms throughout. Heat-resistant strains survive at maximum temperatures of 100°C. Spores then germinate and multiply to 10^6 to 10^7 cells/g in the highly advantageous, anaerobic environment created when meat cools slowly or stands at ambient temperature. Reheating may not kill these cells; when ingested they multiply still further, sporulate, and release their toxin.

Enterotoxin-producing strains of *C. perfringens* may sometimes cause diarrhoea by means of overgrowth in the gut. Patients, usually elderly, begin to experience diarrhoea without known contact with contaminated food. The diarrhoea may be short lived or persist intermittently for several months. Colony counts of 10^8 to 10^{10}/g of faeces are associated with the presence of high titres of free toxin. Previous antimicrobial treatment may encourage the overgrowth and the same strain has been found to cross infect patients.

Further reading

Botulism

Cherington M (1998). Clinical spectrum of botulism. *Muscle and Nerve* **21**, 701–10.

Maselli RA (1998). Pathogenesis of human botulism. *Annals of the New York Academy of Sciences* **841**, 122–39.

Schreiner MS, Field B, Ruddy R (1991). Infant botulism: a review of 12 years' expderience at the Children's Hospital of Philadelphia. *Pediatrics* **87**, 159–65.

Hayes MT, Seto O, Ruoff KL (1997). Weekly clinicopathological exercises: Case 22-1997: A 58-year-old woman with multiple cranial neuropathies. *New England Journal of Medicine* **337**, 184–90.

Gas gangrene

Centers for Disease Control (2000). Update: *Clostridium novyi* and unexplained illness among injecting-drug users. *Morbidity and Mortality Weekly Report* **49**, 543–5.

Darke SG, King AM, Slack WK (1977). Gas gangrene and related infection: classification, clinical features and aetiology, management and mortality. A report of 88 cases. *British Journal of Surgery* **64**, 104–12.

Maclennan JD (1962). The histotoxic clostridial infections of man. *Bacteriology Reviews* **26**, 177–276.

Naylor CE, Eaton JT, Howells A, *et al*. (1998). Structure of the key toxin in gas gangrene. *Nature Structural Biology* **5**, 738–46.

Rood JI (1998). Virulence genes of *Clostridium perfringens. Annual Review of Microbiology* **52**, 333–60.

Shouler PJ (1983). The management of missile injuries. *Journal of the Royal Navy Medical Service* **69**, 80–4.

Gastrointestinal infections

Bartlett JG (1992). The 10 most common questions about *Clostridium difficile* and diarrhea/contis. *Infectious Diseases in Clinical Practice* **1**, 254–9.

Hobbs BC (1974). Clostridium welchii and Bacillus cereus infection and intoxication. *Postgraduate Medical Journal* **50**, 597–602.

Larson HE, Price AB, Honour P, Borriello SP (1978). *Clostridium difficile* and the aetiology of pseudomembranous colitis. *Lancet* **i**, 1063–6.

Lawrence GW, Murrell TGC, Walker PD (1979). Pigbel. *Papua New Guinea Medical Journal* **22,** 1–86.

7.11.22

Tuberculosis

Richard E. Chaisson and Jean Nachega

Introduction

Tuberculosis is one of the most important diseases in the history of humanity, and remains an extraordinary burden on human health today. Archaeological evidence demonstrates that tuberculosis was present in antiquity, and large epidemics of the disease emerged in Europe in the Middle Ages. While contemporary physicians consider tuberculosis to be one of the classical infectious diseases, recognition of the clinical manifestations of the disease has evolved over the past two millenia. The Greek term *phthisis* was used by Hippocrates to describe the wasting disease later known as tuberculosis. While the Greeks recognized various clinical manifestations of tuberculosis, understanding of the connection between the forms was limited. In the Middle Ages, the study of anatomy and the correlation of pathological findings with clinical syndromes led to a better understanding of the disease. The term 'tuberculosis' was introduced in the early nineteenth century, derived from the tubercles characterized in the study of pathological features of the disease.

The impact of tuberculosis on mankind cannot be overstated, as the disease has killed hundreds of millions of people over the centuries and has had economic and social effects perhaps unparalleled in the history of medicine. Between 1700 and 1950, tuberculosis was a great killer in the developed world, earning the sobriquet 'the captain of the men of death...' from John Bunyan, and 'the White Plague' from René and Jean Dubos. The inspiration that artists have drawn from tuberculosis, portrayed in literature, opera, and art, testifies not only to the importance of the disease within their contemporary societies, but also to the extent to which tuberculosis affected artists themselves. The annals of art are rife with those who succumbed to tuberculosis, including Keats, Chopin, the Brontë sisters, Robert Louis Stevenson, Poe, and many others.

The conquest of tuberculosis through the development of vaccines, drugs, and diagnostics was a principal goal of biomedical research in the nineteenth and twentieth centuries. The first description of the tubercle bacillus as the cause of tuberculosis by Robert Koch in 1882 was a scientific landmark. The postulates established by Koch for determining the microbial aetiology of disease have continuing influence today, and molecular correlates of those derived by Koch further strengthen the ingenuity of his thesis. The discovery of streptomycin by Schatz and Waksman in 1943 was a major triumph; both Koch and Waksman received the Nobel prize for their efforts. The development of additional antimicrobial agents against tuberculosis in the 1950s, 1960s, and 1970s and the evaluation of chemotherapy in elegant studies conducted by the British Medical Research Council, the United States Public Health Service, and the United States Veterans Administration led to a marked apathy about tuberculosis in the closing decades of the twentieth century.

Despite the availability of curative chemotherapy for more than half a century, however, tuberculosis continues to cause an enormous amount of suffering, disability, and mortality. In 1994, the World Health Assembly declared that tuberculosis was a global health crisis, and the situation has only grown more grave since then. Epidemics of HIV-related tuberculosis and multidrug-resistant disease have expanded in the past 5 years, and global control of tuberculosis is a remote possibility at present.

The unique biological properties of the causative organism, *Mycobacterium tuberculosis* complex, allow for a long incubation period between the time of infection and the development of symptoms. Latent tuberculosis infection can persist for decades prior to causing disease, or can persist for the lifetime of an infected person without ever causing clinically evident illness. Because latent infection creates a large reservoir of carriers of the infection, disease elimination is difficult to contemplate.

Aetiology

Tuberculosis is a granulomatous disease caused by organisms of the *M. tuberculosis* complex, including *M. tuberculosis*, *M. bovis*, and *M. africanum*, with *M. tuberculosis* greatly predominating. *M. tuberculosis* and the other mycobacteria are small, rod-shaped or curved bacilli in the Order Actinomycetales, Family Mycobacteriaceae, with a unique, thick cell wall composed of glycolipids and lipids. The lipid-rich coat of the mycobacteria renders these organisms resistant to acid decolorization following carbol-fuschin staining, hence the term 'acid-fast bacilli.' Classification of the mycobacteria was based for many years on the staining and growth properties described by Runyon, but this unwieldy system has been largely replaced with modern techniques that identify mycobacteria by specific DNA sequences and, to a lesser extent, biochemical assays. Mycobacteria are frequently considered according to the diseases they cause rather than their behaviour in the laboratory: *M. tuberculosis* complex causing tuberculosis; *M. leprae* the cause of leprosy; and the non-tuberculous mycobacteria, including rapid growers, associated with a wide range of manifestations, particularly in immunocompromised hosts.

The organisms of the *M. tuberculosis* complex are remarkably slow growing, with a generation time of between 20 and 24 h. The exceedingly slow intrinsic reproductive rate of *M. tuberculosis* contributes both to its behaviour as a pathogen and to difficulties in recovering the organism in culture. Moreover, *M. tuberculosis* is able to persist in a latent form within cells and granulomas for many years, and can reactivate to cause disease decades after infection is acquired. Tubercle bacilli are not known to form pores, but both typical bacilli and non-staining forms of the bacteria persist in cells and tissues, as evidenced by detection of DNA, years after infection is acquired and retain the capacity to replicate and produce clinical illness. These unique biological characteristics make the tubercle bacillus exceedingly difficult to combat and control.

Epidemiology

Despite the widely held belief that tuberculosis was waning during the 1980s, global tuberculosis incidence has been steady or increasing for several decades. In Western Europe and North America, the incidence of tuberculosis peaked in the 1700s and 1800s, then declined over a period of years prior to the development of chemotherapy. Improvements in hygiene and nutrition, along with reductions in household crowding, were credited with these trends. Following the introduction of curative treatment for tuberculosis in the era following Second World War, the incidence of disease fell even further, and tuberculosis deaths were greatly decreased. The success in controlling tuberculosis experienced in the Western nations was not replicated in developing countries, and increasing epidemics of the disease have been occurring in these areas. Ironically, progress in tuberculosis control in the Western nations led to neglect of public health programmes that were responsible for reductions in morbidity. As a consequence of inattention to control, the United States experienced a resurgence of tuberculosis between 1985 and 1992, with a 21 per cent increase in the annual number of reported cases during that time. In the United Kingdom, tuberculosis incidence has plateaued over the past decade, with an annual incidence of 11 cases per 100 000 population since 1991. Worldwide, tuberculosis continues to kill more than 2 million people per year, making it the second leading infectious cause of death after HIV infection. In fact, tuberculosis is a eading cause of death in AIDS, and HIV-related tuberculosis deaths are attributed to AIDS, not tuberculosis. If these deaths were attributed to tuberculosis, it would remain the leading infectious cause of death worldwide.

The World Health Organization estimates that 2 billion people, or one-third of the world's population, are infected with *M. tuberculosis*. From this seedbed of latent infection, about 8 million new cases of active disease arise each year, with a global incidence of approximately 160 cases per 100 000 population. The global

distribution of tuberculosis case rates is shown in Fig. 1. Disease due to *M. tuberculosis* is most common in developing nations, both in absolute numbers and incidence of new cases. Twenty-two countries account for 80 per cent of all tuberculosis, with India and China responsible for 23 and 17 per cent of cases, respectively. In general, the highest incidence of disease is found in the countries of sub-Saharan Africa, where HIV infection has contributed to extraordinary increases in case rates, while the greatest number of cases arise in the populous nations of Asia, which have moderately high rates of disease per capita. The global incidence of tuberculosis is increasing slightly, although population growth is resulting in higher numbers of cases each year. Declines in incidence in the developed world have been offset by increasing rates in the HIV-ravaged countries of Africa and by escalating incidence in Eastern Europe in the aftermath of the collapse of communism and its public health infrastructure.

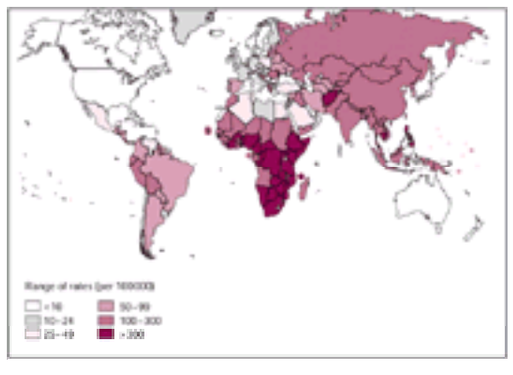

Fig. 1 Global tuberculosis incidence 1999.

Typically tuberculosis affects young adults, with peak incidence in those aged 25 to 44. The dynamics of tuberculosis within a particular country or region, however, reflects both historical trends in tuberculosis transmission and current risk factors and practices of disease control. In Western Europe, for example, tuberculosis is seen in two demographic groups: elderly native Europeans who were presumably infected many years ago and who experience reactivation of latent infections as they age or become immunocompromised, and younger immigrants from high-incidence countries in the developing world. In the United States, tuberculosis is seen in young adults who have immigrated from endemic areas and in those with HIV infection, whereas reactivation tuberculosis in the elderly is increasingly uncommon. In the developing world, tuberculosis most commonly occurs in young adults, with rapidly escalating rates in those with HIV infection. In all countries where tuberculosis is prevalent, young children who acquire tuberculosis from adults account for a small proportion of all cases. It is interesting that children between the ages of 5 and 15 have extremely low rates of tuberculosis, even in areas with a high disease burden.

The epidemiology of tuberculosis is a function of two distinct but related phenomena: the likelihood of becoming infected with *M. tuberculosis* and the probability of developing disease once infection has occurred. Risk factors for becoming infected relate to exposure to infectious individuals. Throughout the world, living with someone who has infectious tuberculosis is the most important risk factor for acquiring infection. The longer the duration of undiagnosed tuberculosis, the greater the severity of disease and probability of transmitting infection. The more intimate the contact, the greater the chance of becoming infected. Exposure to infectious individuals in other environments, including hospitals, prisons, and the workplace, is another important route of infection. In areas of the world where tuberculosis is frequent, exposure in the community is probably unavoidable. In low-prevalence countries community exposure is most likely to occur in distinct pockets of increased incidence, such as poorer areas of large cities or neighbourhoods with high HIV prevalence.

After *M. tuberculosis* infection is acquired, the risk of developing disease is dependent on host immunity. As discussed below, a number of conditions have been identified that increase the risk of active disease in a person with latent tuberculosis infection, most notably HIV infection. Strain differences in *M. tuberculosis* have not been associated with the risk of disease, although inoculum size is associated with the probability of becoming ill. Household contacts who are infected by patients with high sputum levels of acid-fast bacilli have a higher incidence of active disease than contacts of patients who have sputum smears negative for acid-fast bacilli. On the other hand, while there is some evidence that specific strains of *M. tuberculosis* may infect contacts more successfully than other strains, the risk of disease in those infected with these transmissable strains is not elevated.

Tuberculosis is a disease traditionally associated with specific population groups, notably the poor, alcohol and drug abusers, and more recently, those with HIV infection. The increased incidence of tuberculosis in impoverished populations is probably multifactorial, involving increased risk of infection (for example due to crowded living conditions and a higher background prevalence of disease in the community) and increased risk of developing disease after infection (for example due to malnutrition). Similar reasons may explain the higher rates of tuberculosis seen in alcohol and drug abusers, with suppression of host cellular immunity either directly or indirectly from substance abuse. The more recent association of tuberculosis and HIV infection is clearly related to development of cellular immunodeficiency in those with HIV, but in many settings those at highest risk for HIV infection are also more likely to be latently infected with *M. tuberculosis* than others.

The impact of HIV infection on the epidemiology of tuberculosis is striking. As will be discussed below, HIV infection is the most potent known biological risk factor for tuberculosis. The relative risk of tuberculosis in an HIV-infected person is 200- to 1000-fold greater than in someone without HIV infection. As a result of the extraordinary risk conferred by HIV infection, the majority of patients with tuberculosis in many sub-Saharan countries are HIV seropositive. In the United States and the United Kingdom, HIV infection accounts for a substantial proportion of tuberculosis cases in many cities. HIV infection is the unifying theme in many nosocomial outbreaks of tuberculosis, as infection is spread among immunocompromised patients receiving medical care at the same facility. It is increasingly apparent that control of tuberculosis will not be possible globally without control of HIV infection.

Another very important trend in tuberculosis epidemiology is the growing problem of drug-resistant tuberculosis. There are two categories: primary resistance, which is the presence of drug resistance in someone who has never had treatment for tuberculosis, and secondary resistance, the presence of resistance in a patient who has previously been treated for tuberculosis. Primary resistance results from acquiring an infection that is already drug resistant, while acquired resistance is the result of inappropriate therapy that selects for resistant mutants of *M. tuberculosis*. A global survey of resistance performed by the World Health Organization and the International Union Against Tuberculosis and Lung Disease found that the median prevalence of primary drug resistance was 10 per cent, and the median prevalence of acquired resistance was 36 per cent. Moreover, 'hot spots' of drug-resistant tuberculosis were identified on all continents, most notably in the former Soviet nations, where multidrug-resistant tuberculosis is identified in 10 to 20 per cent of all cases. Multidrug-resistant tuberculosis is exceedingly difficult to cure, and so failure to control its spread has ominous implications.

Pathogenesis

The development of active tuberculosis, like all infectious diseases, is a function of the quantity and virulence of the invading organism and the relative resistance or susceptibility of the host to the pathogen. Tubercle bacilli are transmitted between people by aerosols generated by coughing or otherwise expelling infectious pulmonary or laryngeal secretions into the air. *M. tuberculosis* bacilli excreted by this action are contained within droplet nuclei, extremely small particles (less than 1 μm) that remain airborne for long periods and are disseminated by diffusion and convection until they are deposited on surfaces, diluted, or inactivated by ultraviolet radiation. People breathing air into which droplet nuclei have been excreted are at risk of becoming infected if inhaled nuclei are deposited in their alveoli. Transmission of tuberculous infection by other routes, such as inoculation in laboratories and aerosolization of bacilli from tissues in hospitals, has been documented, but these are an insignificant means of spread. *M. bovis* can be acquired from contaminated milk from tuberculous cows, but modern animal husbandry practices and pasteurization of milk have virtually eliminated this mode of infection throughout most of the world.

The natural history of tuberculosis in humans is illustrated in Fig. 2. People who are in contact with someone with infectious tuberculosis may acquire infection, as described above. Factors that affect the likelihood of infection being transmitted include the severity of the disease in the index case (such as extent of radiographic abnormalities, cavitation, frequency of cough), the duration and closeness of exposure, and environmental factors such as humidity, ventilation, and ambient ultraviolet light. A number of studies in diverse locations and circumstances have shown that approximately 20 to 30 per cent of close contacts of a patient with untreated tuberculosis become infected with *M. tuberculosis*, as demonstrated by the development of a reactive tuberculin skin test.

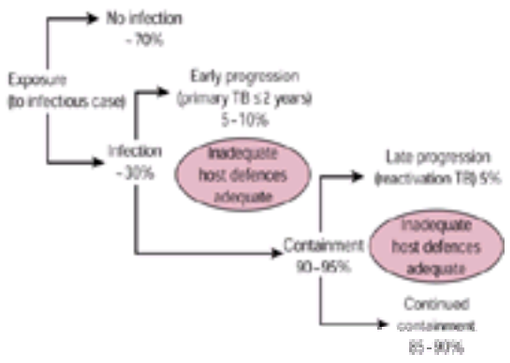

Fig. 2 Natural history of tuberculosis.

Deposition of tubercle bacilli in the alveoli results in a series of protective responses by the cellular immune system that forestall the development of disease in the majority of infected people. Alveolar macrophages ingest tubercle bacilli, which then multiply intracellularly and eventually cause cell lysis with release of organisms. Killing of *M. tuberculosis* within macrophages is prevented by inhibition of phagolysosome formation by the tubercle bacilli through a process that is not understood. Additional alveolar macrophages engulf progeny bacilli, resulting in further intracellular growth and cell death. Over a period of weeks, as tubercle bacilli proliferate within macrophages and are released, infection spreads to regional lymph nodes, elsewhere in the lungs, and systemically. Foci of tubercle bacilli can be established in multiple organs, including the lymph nodes, brain, kidneys, and bones. In most people, after several weeks, specific immunity is developed, with activated T lymphocytes mediating a T_{H1}-type response. Macrophages act as antigen-presenting cells, interacting with CD4 lymphocytes primed for *M. tuberculosis* antigens. Activated CD4 lymphocytes produce both IL-2, which promotes activation of additional T lymphocytes, and interferon-g, which binds with receptors on macrophages and promotes intracellular killing of organisms. Tumour necrosis factor-a production is induced in macrophages, and this too promotes killing of intracellular bacilli. The specific role of CD8 cells in the control of tuberculosis has not been fully elaborated, although there is evidence that cytotoxic T lymphocytes may play a role in containing a tuberculous infection. In addition, CD8 lymphocytes also produce interferon-g and participate in granuloma formation.

The classic immunological response to infection with tubercle bacilli is the walling off of viable bacilli in granulomas, collections of cells surrounding a focus of *M. tuberculosis*, usually within macrophages but sometimes extracellular organisms, that serve to contain the infection. Granulomas consist of macrophages, CD4 and CD8 lymphocytes, fibroblasts, giant cells, and epithelioid cells that produce an extracellular matrix of collagenous and fibrotic materials that are continually remodelled and can become calcified. A calcified granuloma at the initial site of infection in the lung is referred to as a Ghon complex, while the combination of a Ghon complex and a calcified regional lymph node is called a Ranke's complex.

The development of the cellular immune response to *M. tuberculosis* is accompanied by the development of delayed-type hypersensitivity to specific antigens from tubercle bacilli. While delayed-type hypersensitivity is distinct from the cell-mediated immunity that provides protection from disease, this sensitivity to tubercle-derived proteins has proved enormously useful for diagnosing tuberculosis infection. Use of purified protein derivatives (**PPD**) of tuberculin is the basis for estimating the prevalence of latent tuberculosis infection in populations. This is essential in studying the natural history of tuberculosis infection, and is frequently helpful in evaluating patients with suspected tuberculosis disease. The difference between delayed-type hypersensitivity and immunity to tuberculosis is illustrated by the observation that 80 to 90 per cent of patients with active disease, and therefore clearly not immune, have positive tuberculin tests.

For the majority of people acquiring a new tuberculous infection, the development of cell-mediated immunity to the organism is protective and holds the bacilli in check, although viability is usually maintained. A small minority will be unable to contain the infection and progress to active tuberculosis disease, often referred to as primary tuberculosis. Early progression of infection to disease is associated with immunosuppression, particularly with HIV infection, a higher inoculum of organisms, malnutrition, and perhaps, concomitant illness. While rates of active disease in young children who are contacts of infected individuals are no higher than for older contacts, young children with primary tuberculosis do develop more severe forms of tuberculosis than adults, including disseminated disease and tuberculous meningitis.

Those who successfully contain the organisms have a latent tuberculosis infection that may reactivate later in life. Studies of latent tuberculosis infection acquired in childhood or adolescence suggest a lifetime risk of reactivation of *M. tuberculosis* of about 10 per cent. Table 1 lists risk factors for reactivation of latent tuberculosis infection. The most potent is HIV infection, which increases the rate of reactivation by as much as 1000-fold. Immunosuppression from malignancy, cytotoxic therapy, corticosteroids, and other agents that alter cellular immune responses can also reactivate latent tuberculosis infection. Other potentiating factors include diabetes, endstage renal disease, injection drug use (independent of HIV infection), low body weight, gastrointestinal surgery, and silicosis. Cigarette smoking is associated with increased tuberculosis incidence (notably in India), as is alcohol abuse. Inhibitors of tumour necrosis factor-a used to treat rheumatoid arthritis or inflammatory bowel disease increase the risk of tuberculosis. Tuberculosis rates are usually higher in the elderly than in younger adults in developed countries, but this may represent a higher prevalence of latent infection in older cohorts, rather than immunological senescence.

Clinical features

Classification of tuberculosis infection and disease

Infection with *M. tuberculosis* can result in clinical manifestations ranging from asymptomatic carriage of latent bacilli to life-threatening pneumonia. Classification of the different stages of *M. tuberculosis* in humans by the American Thoracic Society (**ATS**) is shown in Table 2. This system is used more for public health purposes than for clinical management, but is useful because it reflects the natural history of *M. tuberculosis* and categorizes patients according to the type of evaluation and treatment they may need.

ATS Category 0 describes people with no history of tuberculosis exposure and a negative tuberculin skin test (if performed). Category 1 includes those people exposed to an infected individual but in whom no evidence of infection is found. This is a temporary category used during the evaluation of contacts of tuberculosis cases; repeat tuberculin testing several months after the exposure would result in these individuals being reclassified to another category. Category 2 is defined as latent tuberculosis infection without evidence of disease, and is based on a positive tuberculin skin test without clinical or radiographic signs of illness. Category 3 is confirmed, active tuberculosis disease requiring treatment. As discussed below, this category is further divided according to the site of disease and laboratory features, including results of acid-fast bacilli smears. Category 4 is defined as inactive tuberculosis. Patients in this category do not have clinical or laboratory evidence of active disease, but are known to have suffered previously from tuberculosis. This category includes those who have been treated and cured of active tuberculosis, as well as individuals who have spontaneously recovered from tuberculosis without treatment. Finally, category 5 refers to patients in whom tuberculosis is suspected, but who are still undergoing evaluation. Depending on the degree of suspicion of the diagnosis, such people might be started on presumptive therapy for tuberculosis pending the outcome of cultures and other laboratory assessments. Like category 1, it is a temporary category for patients undergoing evaluation. All are subsequently reclassified on the basis of diagnostic studies.

The clinical presentation of active tuberculosis is highly variable, depending on the site and extent of disease and the immune status of the host. Historically, active tuberculosis has been classified as 'primary' or 'post-primary' on the basis of both the presumed duration of infection and the clinical features of the disease. However, molecular epidemiological studies suggest that this classification may be unreliable. For example, the 'classic' presentation of reactivation tuberculosis has been seen in patients whose infection is clearly newly acquired, such as in nosocomial outbreaks where DNA fingerprinting confirms recent transmission. For practical purposes, tuberculosis is generally divided into pulmonary and extrapulmonary forms, with considerable clinical heterogeneity within these cate-gories.

Pulmonary tuberculosis

Pulmonary tuberculosis is usually a subacute respiratory infection with prominent constitutional symptoms. The most frequent symptoms of pulmonary tuberculosis are cough, fever, night sweats, and malaise. Cough in pulmonary tuberculosis is initially dry, but often progresses to become productive of sputum and, in some instances, haemoptysis. The sputum is generally yellow in colour, and is neither malodorous nor thick. Haemoptysis may occur acutely in patients with untreated tuberculosis, but is also a feature of treated tuberculosis; damage from prior tuberculosis may result in bronchiectasis or residual cavities that can either become superinfected or erode into blood vessels or airways, producing haemoptysis. Advanced tuberculosis may also present with bloody sputum. Rarely, the bleeding is

massive leading to shock, asphyxia, and death.

Chest pain is not a prominent symptom in pulmonary tuberculosis, although coughing may cause musculoskeletal pain. Patients with tuberculous pleurisy may experience pleuritic pain. Radicular chest pain may be associated with spinal tuberculosis. Dyspnoea alone may be a sign of extensive parenchymal destruction, large pleural effusions, endobronchial obstruction, or pneumothorax.

Patients with tuberculosis also experience loss of appetite and weight loss or cachexia, often out of proportion to their diminished intake of food. Elevations in tumour necrosis factor-a may be responsible. Mild symptoms include emotional lability, irritability, depression, and headache.

Most patients present after feeling unwell for weeks or months. In surveys of populations with high rates of disease and poor access to medical care, a history of cough for more than 3 weeks was strongly associated with a diagnosis of active tuberculosis. Untreated tuberculosis is associated with high mortality, but many patients may have persistent symptoms for years. A study of untreated pulmonary tuberculosis in the pre-therapy era found that after 5 years 50 per cent of patients had died, 25 per cent had spontaneously healed, and 25 per cent were chronically ill with pulmonary disease. A subset of patients have rapidly progressive disease, the so-called 'galloping consumption' of old. This is now most often seen in patients with HIV infection or other forms of severe immunosuppression. These patients have progressively severe pulmonary symptoms over a period of several weeks, often in the setting of disseminated disease. Failure to diagnose and treat these patients promptly may result in death.

Physical findings in pulmonary tuberculosis may be of limited usefulness in making a diagnosis. Fever is an irregular and unreliable feature in tuberculosis. While most patients complain of fevers prior to presentation, only one-half to three-quarters of patients with confirmed tuberculosis have a documented fever. Examination of the chest may reveal dullness to percussion and rales, although these findings are highly variable and non-specific. Signs of consolidation are usually absent. The classic post-tussive rales described in the last century are not often present and are not specific to tuberculosis. Patients with disseminated tuberculosis may have lymphadenopathy, hepatomegaly, or evidence of central nervous system involvement, but these are not generally seen in typical pulmonary tuberculosis. Clubbing and cyanosis are findings associated with prolonged and advanced pulmonary disease. Thus, the diagnosis of tuberculosis almost always rests on the patient's history and epidemiological characteristics, in conjunction with laboratory studies described below. The most important step in making a timely diagnosis of tuberculosis is to think of it in the first place.

Radiological evaluations play a critical role in the diagnosis of pulmonary tuberculosis. Disease due to *M. tuberculosis* can involve any portion of the lungs, and radiographic findings are usually only suggestive, not diagnostic, of tuberculosis. The typical radiological manifestations of pulmonary tuberculosis are upper lobe infiltrates that may show cavitation. *M. tuberculosis* exhibits a unique predilection for the upper zones of the lungs for reasons that are not well understood. Latent infection characteristically reactivates in the apical segments of the upper lobes, or the superior segments of the lower lobes. The infiltrates are often fibronodular and irregular, and may be diffuse and associated with volume loss. Cavities, when present, are rarely symmetrical and do not usually have air–fluid levels, such as those seen in pyogenic lung abscesses. Examples of the radiographic appearance of pulmonary tuberculosis are seen in Fig. 3.

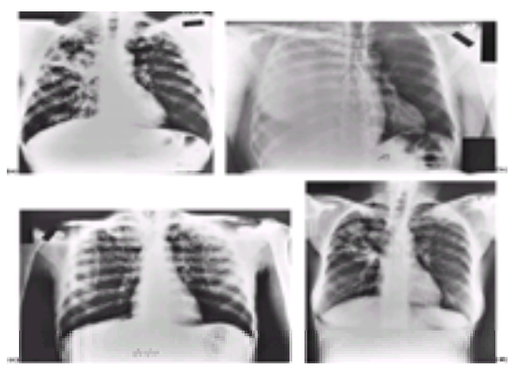

Fig. 3 Radiographic appearance of pulmonary tuberculosis. (a) Extensive tuberculosis with right upper lobe volume loss and multiple small cavities. This patient was the source of at least 14 secondary cases in contacts. (b) A 69-year-old man with right pleural tuberculosis. (c) Diffuse pulmonary nodules in an HIV-infected man with pulmonary tuberculosis. (d) Cavitary upper lobe disease in an HIV-infected woman.

The classic radiographic presentation described above is neither pathognomonic nor highly sensitive for pulmonary tuberculosis. A number of other lung infections, notably the pulmonary mycoses, can present with similar findings. More important, one-third to one-half of patients with pulmonary tuberculosis lack these classic radiographic findings. Lower lung zone infiltrates, mid-lung focal infiltrates, pulmonary nodules, and infiltrates with mediastinal or hilar adenopathy are also seen. In particular, HIV-infected patients with tuberculosis tend to present with 'atypical' findings, and up to 5 per cent of them may have a normal chest radiograph but sputum cultures that yield *M. tuberculosis*. The lack of typical radiographic features is not, therefore, grounds for rejecting the diagnosis in a patient with a history and symptoms compatible with tuberculosis.

Computed tomography (CT) is increasingly used to evaluate radiographic findings that are not readily explained after an initial assessment. They may reveal more extensive involvement than conventional radiographs, including multiple nodules, small cavities, and multilobar infiltrates.

The laboratory diagnosis of pulmonary tuberculosis relies on examination and culture of sputum or other respiratory tract specimens. Definitive diagnosis requires growth of *M. tuberculosis* from respiratory secre-tions, while a probable diagnosis can be based on typical clinical and radiographic findings with either sputum positive for acid-fast bacilli or other specimens, or typical histopathological findings on biopsy material. The specificity of these latter approaches depends on the prevalence of disease due to non-tuberculosis mycobacteria in the population.

Throughout most of the world, acid-fast staining of sputum is the sole test available to confirm the diagnosis of pulmonary tuberculosis. In developing countries, the positive predictive value of the sputum acid-fast smear is very high, as the likelihood of non-tuberculous mycobacterial disease is quite low. In industrialized countries, disease due to the non-tuberculous mycobacteria is relatively more common and reliance on smears without cultures is potentially misleading. Despite the best efforts of clinicians, a confirmed diagnosis of tuberculosis cannot be established in some patients who have the disease, and a response to presumptive therapy forms the basis for establishing the diagnosis.

Extrapulmonary tuberculosis

In the United States, extrapulmonary tuberculosis is defined as disease outside the lung parenchyma, and in the United Kingdom, as disease outside the lungs and pleura. This seemingly subtle distinction has considerable epidemiological significance, however, as pleural tuberculosis is the most common extrapulmonary site of disease in the United States.

During the initial seeding of infection with *M. tuberculosis*, described earlier, haematogenous dissemination of bacilli to a number of organs can occur. These localized infections, as in the lung, can progress into primary tuberculosis or become walled off in small granulomas where bacteria may remain dormant if they are not killed by cell-mediated immune responses. Extrapulmonary tuberculosis, therefore, can either be a presentation of primary or reactivation tuberculosis.

Extrapulmonary tuberculosis may be generalized or confined to a single organ. In otherwise immunocompetent adults, extrapulmonary tuberculosis is found in 15 to 20 per cent of all tuberculosis cases. In young children and immunosuppressed adults, rates of extrapulmonary disease are substantially higher. It is seen in more than half of patients with HIV-related tuberculosis and one-quarter of patients with tuberculosis under 15 years of age. Children less than 2 years old have high rates of miliary and meningeal disease.

The organs most frequently involved are listed in Table 3. To some extent the frequency with which specific organs are involved reflects the pathophysiology of the disease. Infection spreads from the lungs, the primary site of inoculation, by lymphatic and haematogenous routes, to the pleura, lymph nodes, kidneys and other genitourinary organs, bone, and central nervous system. Bacteraemia is transient and rarely detected except in patients with HIV infection and low CD4 lymphocyte

counts.

Both pulmonary and extrapulmonary disease are found in up to 50 per cent of patients with HIV-related tuberculosis, so it is important to consider the possibility of extrapulmonary pathology when pulmonary tuberculosis is diagnosed in an HIV-infected patient (and vice versa). Pulmonary involvement is seen in up to one-quarter of patients with tuberculous meningitis and less frequently with other sites of disease.

Pleural tuberculosis

This is the result of two distinct pathophysiological sequences, which present in strikingly different manners. Most pleural tuberculosis is associated with primary infection, and is the result of seeding of the visceral pleura with relatively small numbers of tubercle bacilli via direct extension from adjacent lung tissue. A large proportion of patients with this form of tuberculous pleurisy will have evident pulmonary disease, although findings may be subtle. The duration of symptoms is generally brief, usually several weeks. Patients complain of fever, chest pain, and non-productive cough. Other constitutional and respiratory symptoms may be present. Unlike pneumococcal pneumonia, which presents abruptly, tuberculous pleurisy has a more insidious onset.

The second form of pleural tuberculosis occurs when larger numbers of bacilli invade the pleural space and multiply, producing frank empyema. Tuberculous empyema is seen in older patients, almost all of whom have extensive pulmonary disease. Patients present with prolonged cough, chest pain, fever, cachexia, and night sweats. Pneumothorax is a common complication that may be associated with a more rapid disease course.

The radiographic picture in tuberculous pleurisy reflects the underlying pathophysiology of the disease. Patients with the primary type of pleurisy tend to have small, unilateral effusions, and up to half have visible parenchymal lesions on plain radiographs. In patients with tuberculous empyema, the effusions are larger, more likely to be loculated, and adjacent pulmonary involvement is often evident.

When pulmonary parenchymal involvement is manifest, sputum smears and cultures are likely to be diagnostic and pleural disease can be inferred from the pulmonary findings. When pulmonary findings are minimal, or the initial test results unrevealing, analysis of pleural fluid is essential. Acid-fast stains of pleural fluid are most often negative in patients with primary tuberculous pleurisy, as few organisms are present. Repeated sampling will show organisms in less than half of cases. Similarly, cultures may be negative. The pleural fluid is usually serous and exudative, with a protein concentration that is more than 50 per cent of the serum level, normal or low glucose, and a slightly acidic pH. The white blood cell count of the pleural fluid is usually in the range of 1000 to 10 000/µl, with a lymphocytic predominance. Lactate dehyrogenase and adenosine deaminase levels are generally elevated. These tests are non-specific and cannot reliably distinguish tuberculous pleurisy from other pleural diseases.

Percutaneous biopsy of the pleura reveals granulomatous inflammation in up to 80 per cent of cases, and cultures obtained at the time of biopsy are positive in over half of patients. If a first attempt fails to provide a diagnosis, a second biopsy may be successful. Compared with blind sampling with a percutaneous pleural needle, viewing biopsy targets by thorascopy improves the diagnostic yield.

Lymphatic tuberculosis

This can occur in any location, but classic scrofula, involving the cervical or supraclavicular chains, is the most common presentation. Mediastinal and hilar lymphatic tuberculosis is a feature both of primary and disseminated disease, but discovery of these lesions is usually incidental. Lymphatic tuberculosis is thought to result from drainage of bacilli in the lungs into supraclavicular and posterior cervical lymph node chains. In contrast, lymphatic disease caused by non-tuberculous mycobacteria more often involves anterior cervical, preauricular, or submandibular lymph nodes, suggesting acquisition through the oropharynx. In patients with HIV infection, many groups of lymph nodes may be involved, including axillary, inguinal, mesenteric, and retroperitoneal.

Symptoms in lymphatic tuberculosis are generally limited, unless the disease is disseminated. Painless swelling of a lymph node is the most common presentation. Constitutional symptoms are not prominent in most cases. Examination of the area may reveal several enlarged lymph nodes, as only about 20 per cent of patients have disease of a solitary node.

The diagnosis of lymphatic tuberculosis usually depends on cultures from affected nodes. Biopsies may show granulomatous changes and acid-fast bacilli. Such findings are non-specific, however, and cannot distinguish tuberculous from non-tuberculous lymphadenitis. As discussed elsewhere, the presence of a positive tuberculin skin test with typical biopsy findings is strongly suggestive of tuberculosis. If lymphatic tuberculosis is suspected, these findings warrant presumptive therapy.

Genitourinary tuberculosis

This encompasses a broad array of clinical entities, ranging from disease of the kidneys to endometrial, prostatic, and epididymal disease. The most common of these is renal tuberculosis, which results from haematogenous seeding of the renal cortex during the primary infection. The pathogenesis of other genitourinary sites is either from downstream extension of renal infection over time or from haematogenous seeding at the time of the initial acquisition of *M. tuberculosis*.

Renal tuberculosis is probably underdiagnosed because it is frequently asymptomatic. Many cases are diagnosed as a result of routine detection of sterile pyuria. The development of symptoms reflects a more advanced stage of disease, associated with considerable tissue destruction. When genitourinary tuberculosis is symptomatic, the most common complaints are localized, including urinary symptoms and flank pain. In men, tuberculosis can cause prostatitis and epididymitis, both of which can present with pain resulting from swelling. In women, genital tract tuberculosis may be symptomatic when it involves the ovaries and fallopian tubes; pelvic pain is also a feature of endometrial tuberculosis. However, menstrual abnormalities and infertility may be the only signs of genital disease.

The diagnosis of genitourinary tuberculosis depends on the anatomical site of the disease. Renal tuberculosis, suggested by sterile pyuria, is diagnosed by isolation of organisms in the urine. Early morning urine is more likely to grow *M. tuberculosis* than spot samples obtained at other times. In patients with symptoms of upper urinary tract illness, radiological studies are often helpful. The kidneys may appear calcified on abdominal radiographs. Intravenous pyelography may show distorted or dilated calyces or renal pelvis, papillary necrosis, cavitation or abscesses of the renal parenchyma, or intrarenal or ureteric obstructions. Use of renal ultrasound or CT scanning may be more sensitive for identifying the abnormalities of renal tuberculosis, but there is the greatest experience with contrast radiography. When tuberculosis of the bladder is suspected, cystoscopy with biopsy may lead to the identification of granulomas prior to identification of organisms by culture. Diagnosis of prostatic, testicular, or epididymal tuberculosis is usually accomplished with cultures obtained by fine needle aspiration or transurethral resection of the prostate. Cervical and endometrial tuberculosis can be diagnosed by biopsy with culture.

Tuberculous meningitis (see also Chapter 24.14.1)

This is the most common central nervous system manifestation of tuberculosis. It is much more likely to occur in children under the age of 5 and HIV-infected patients than in immunocompetent adults. Although meningitis accounts for only a small fraction of all cases of tuberculosis, it is a devastating form of the disease that is uniformly fatal if left untreated.

The pathogenesis of meningeal tuberculosis varies with the age and immunological status of the patient. Reactivation of microscopic granulomas in the meninges was found by Rich to cause diffuse meningeal infection. These foci of infection are probably implanted at the time of primary bacillaemia. When they rupture into the subarachnoid space they invoke an inflammatory response leading to tuberculous meningitis. Meningeal disease can also occur in conjunction with miliary disease, especially in children. Adults can acquire meningeal disease during bacillaemia of miliary disease, but this is not the usual pathogenesis of meningeal infection. Rarely, invasion into the spinal canal from a paraspinous or vertebral focus can be the source of central nervous system involvement.

Historically, the clinical spectrum of tuberculous meningitis has been categorized in three stages, defined by the British Medical Research Council in 1948. Stage 1 consists of a prodrome lasting for 1 to 3 months. Non-specific symptoms such as fever, malaise, and headache predominate. In this stage, patients are conscious and rational, but may have meningism. Focal neurological signs are absent and there are no signs of hydrocephalus. In stage 2 disease, single cranial nerve abnormalities, such as ptosis or facial paralysis appear, and paresis and focal seizures may occur. Kernig's and Brudzinski's signs and hyperactive deep tendon reflexes may be found. Prominent signs include altered cerebration, behavioural change, impaired cognitive ability, and increasing stupor. Headache and fever are common.

In stage 3, patients are comatose (Glasgow coma scale less than 8) or stuporous and often have multiple cranial nerve palsies and hemiplegia or paraplegia. By this stage, hydrocephalus is common and chronic inflammation in the enclosed space of the skull may result in intracranial hypertension. Seizures may be a prominent feature.

Fever, headache, changes in cerebration, and meningism are present in the majority of patients in most large studies, although no one single sign or symptom is reliably sensitive or specific. Children can be especially difficult to diagnose as symptoms such as fever, vomiting, drowsiness, or irritablity are commonly seen in many minor viral illnesses.

Transient tuberculous meningitis that presents as an aseptic meningitis and resolves without treatment has been described. Benign presentations of meningeal tuberculosis are exceedingly uncommon in clinical practice, and when the diagnosis is made, treatment is mandatory, even in the patient with seemingly trivial symptoms.

Diagnosis is often difficult and requires a high degree of suspicion. In disseminated disease, signs of tuberculosis in other organs, particularly the lungs, are often present. Between 25 and 50 per cent of patients with meningitis in most series also have radiographic evidence of pulmonary tuberculosis, either active or healed. The critical features of tuberculous meningitis, however, are found in the cerebrospinal fluid. Patients with tuberculous meningitis usually have elevated cerebrospinal fluid pressure. An exudative fluid with a mononuclear cell pleocytosis is characteristic. Cerebrospinal fluid is usually clear and the protein is generally in the range of 100 to 500 mg/dl. Hypoglycorrhachia is typical, with cerebrospinal fluid glucose less than 50 per cent of the serum value. The white blood cell count is rarely above1000/µl, and cell counts of below 500/µl are typical. In early meningitis the cells may be predominantly neutrophils, but mononuclear cells predominate in most instances. Acid-fast stains of concentrated cerebrospinal fluid are only positive in one-third or fewer of patients, and cultures are positive in only one-half, although repeated sampling increases the yield.

The disastrous consequences of failing to diagnose tuberculous meningitis, coupled with the low yield of acid-fast stains and cultures from cerebrospinal fluid, has prompted the development of additional tests for establishing a diagnosis. Adenosine deaminase was initially reported to be exceptionally accurate for tuberculous meningitis. Subsequent experience, however, has found it to be insufficiently specific to distinguish tuberculosis from a variety of other acute and chronic meningitides. A number of other tests based on identification of mycobacterial antigens or specific antibodies have been evaluated, but none has been found to be reliable. Nucleic acid amplification tests such as the polymerase chain reaction (**PCR**) have great appeal, but the sensitivity and specificity of available assays are only moderately good. Thus, the diagnosis of tuberculous meningitis often rests upon the astute judgment of a clinician with a high degree of suspicion based on epidemiological and clinical clues. Presumptive therapy is frequently necessary.

Central nervous system tuberculomas are an unusual manifestation and are seen in a small proportion of patients with tuberculous meningitis. Tuberculomas are the result of enlarging tubercles that extend into brain parenchyma rather than into the subarachnoid space. Patients with HIV infection appear to have an increased risk of tuberculomas of the central nervous system, but the disease is far less common than toxoplasmosis, even in areas where tuberculosis is highly prevalent. Tuberculomas of the central nervous system may appear with clinical features of meningitis or of intracranial mass lesions. In the absence of meningeal involvement, seizures or headaches may be the only symptoms. The diagnosis is suggested by brain imaging; MRI is more sensitive than CT scanning. Biopsy of the lesion is required for diagnosis, and material should be submitted for histopathological staining and culture.

Bone and joint tuberculosis

This may affect a number of areas, but vertebral tuberculosis (Pott's disease) is the most common form, accounting for almost one-half of cases. Haematogenous seeding of the anterior portion of vertebral bone during initial infection sets the stage for later development of Pott's disease. Infection grows initially within the anterior vertebral body, then may spread to the disc space and to paraspinous tissues. Destruction of the vertebral body causes wedging and eventual collapse. Patients usually complain of back pain, with constitutional symptoms less prominent. Neurological impairment is a late complication, but delays in diagnosis are common and many patients experience neurological sequelae. Imaging studies of the spine most often reveal anterior wedging, collapse of vertebrae, and paraspinous abscesses. The diagnosis is established with bone biopsy or curettage, or by culture of the drainage from a paraspinous abscess.

Miliary tuberculosis and disseminated tuberculosis

These terms are used interchangeably to describe widespread infection and absent or minimal host immune responses. The term 'miliary tuberculosis' is derived from the classic radiographic appearance of haematogenous tuberculosis, in which tiny pulmonary infiltrates with the appearance of millet seeds are distributed throughout the lungs. Miliary tuberculosis is a more common consequence of primary tuberculosis infection than reactivation, and is seen more frequently in children and immunocompromised adults. Primary miliary tuberculosis presents with fever and other constitutional symptoms over a period of several weeks. Clinical evaluation may reveal lymphadenopathy or splenomegaly, and laboratory tests may show only anaemia. The chest radiograph is initially normal but later develops the typical miliary pattern. Involvement of multiple organ systems is the rule, most often the liver, spleen, lymph nodes, central nervous system, and urinary tract. Patients with reactivation of latent infection who present with miliary disease may have a more fulminant course although, without treatment, progression to severe disease is the rule in all patients. The diagnosis is made on tissue biopsy and culture, as sputum smears are often negative, reflecting the small numbers of bacilli typically present in respiratory secretions.

Other forms of extrapulmonary tuberculosis

These less common sites of infection are diagnosed by a combination of clinical suspicion and the results of biopsies and cultures. Abdominal, ocular, adrenal, and cutaneous tuberculosis are rarely encountered in the modern era, even in immunocompromised patients.

Laboratory diagnosis

Evaluation of patients for *M. tuberculosis* infection or disease relies on both non-specific and specific tests. Imaging studies, body fluid chemistry and cell counts, and histochemical staining, as described above, are useful and important tests for the diagnosis of tuberculosis. Specific studies for identifying mycobacterial infections include the tuberculin skin test, acid-fast microscopy, and mycobacterial culture.

Tuberculin skin testing

Tuberculin skin testing involves the intradermal injection of purified proteins of *M. tuberculosis* (purified protein derivative, or PPD tuberculin) that provokes a cell-mediated delayed-type hypersensitivity reaction, which produces a zone of induration. Tuberculin originated with Robert Koch, who prepared a tubercle sensitin that he thought would cure tuberculosis. Administration of Koch's tuberculin, of course, did not cure the disease, and hypersensitivity reactions to the agent were sometimes severe or fatal, bringing Koch great discredit.

Fortunately, it was recognized that because tuberculin induced reactions in people who were infected with tuberculosis, the substance might prove a better diagnostic test than treatment. Over a period of years, refinements were made in the preparation of tuberculins, and in 1939 Seibert and Glenn produced the reference lot of tuberculin, called PPD-S, which has served as the international standard. Current tuberculin preparations are composed of a variety of small tuberculous proteins derived from culture filtrates and stabilized with a detergent (Tween) to prevent precipitation. The standard dose of tuberculin is 5 tuberculin units (**TU**) of PPD-S, equivalent to 0.1 mg of tuberculin in a volume of 0.1 ml. Commercial and other tuberculin products are standarized against PPD-S to ensure bioequivalence.

Tuberculin testing is used to identify individuals with *M. tuberculosis* infection, and the test cannot distinguish those who have disease from those with latent infection. Intradermal injection of tuberculin into an infected individual invokes a delayed-type hypersensitivity response. Specific T lymphocytes sensitized to tuberculous antigens from prior *M. tuberculosis* infection cause a local reaction at the site of injection. Inflammation, vasodilatation, and fibrin deposition at the site result in both erythema and induration of the skin, the key feature of a tuberculin response. The result of tuberculin testing is categorized according to the amount of induration measured.

Tuberculin skin testing should be done by the Mantoux method, as this is the only technique that has been standardized and extensively validated. An injection of 0.1 ml of PPD-S is given intradermally in the volar surface of the forearm using a tuberculin syringe and small-gauge needle, causing a small wheal. Injection subcutaneously will result in uninterpretable results. Multipuncture devices should not be used. The amount of induration should be measured 2 to 5 days after the

injection; measurements performed precisely 48 to 72 h later are not essential. The transverse diameter of induration should be measured in millimetres using a ruler. The edge of the induration can be seen and marked, or the margins can be detected using the ballpoint pen method, in which the pen is rolled over the skin with light pressure and its progress is stopped at the demarcation of the indurated area.

Criteria for the interpretation of tuberculin skin tests vary according to clinical and epidemiological circumstances. Cut-offs for positive tests developed by the American Thoracic Society and the Centers for Disease Control and Prevention are listed in Table 4. A cut-off of 5-mm induration is used for those at high risk of tuberculosis infection, or at high risk of disease if infected. This category includes close contacts of infectious individuals and patients with radiographic abnormalities consistent with tuberculosis. The rationale for the 5-mm cut-off in these patients is their high pretest probability of being infected. A 5-mm cut-off is also used for HIV-infected patients and those immunocompromised by corticosteroids or other agents. Failure to diagnose tuberculosis infection in these people could be calamitous, so a lower threshhold is used to maximize sensitivity. The use of control antigens such as *Candida* or tetanus toxoid to aid the interpretation of tuberculin tests in HIV-infected patients has been shown to be of no value and is not recommended.

A cut-off of 10-mm induration is used for people from populations with a high prevalence of tuberculosis or for people with conditions that increase the risk of developing active disease if infected. This would include immigrants from endemic areas, residents of some inner cities, and health care workers, as well as patients with diabetes, renal disease, silicosis, and other medical onditions associated with an elevated risk of reactivation of latent tuberculosis. Finally, a cut-off of 15 mm is used in people who have no risk factors for tuberculosis infection or disease. These people are unlikely to be tested.

Tuberculin skin testing is frustratingly crude and somewhat cumbersome, but has proved superior to numerous more 'modern' assays, including antibody tests, quantitative interferon-g detection assays, and other *in vitro* immunodiagnostics. Recently, the use of Elispot assays to detect antigen-recognizing T cells has shown promise as an alternative to tuberculin testing, although further validation is required.

The test does suffer, however, from limitations in both sensitivity and specificity. The 5-TU dose of tuberculin used diagnostically is based on studies in the 1940s that showed that 99 per cent of patients with chronic tuberculosis responded to this dose, while less than 20 per cent of persons without disease and no history of tuberculosis exposure had a response. Subsequent research suggested that the lack of specificity of tuberculin testing may be the result of cross-reactions due to exposure to non-tuberculous mycobacteria. Use of tuberculin derived from *M. avium intracellulare* (PPD-B), for example, induces larger reactions than PPD-S in healthy people from areas where this organism is widespread in the environment. Another important cause of non-specific reactions to tuberculin is vaccination with BCG (bacille Calmette–Guérin). While the reactogenicity of BCG vaccine differs according to the strain, immunization with BCG can produce falsely positive skin test results. Reactions induced by BCG tend to be smaller than true positive reactions, and wane over a period of several years. Studies in populations with high rates of BCG coverage indicate that tuberculin testing can still be used to predict those who are most likely to be infected with *M. tuberculosis*, even though precision is reduced because of cross-reactions.

False-negative tuberculin tests result from both errors in applying and interpreting the test and from anergy. Errors in injection of tuberculin are common, and inter-reader variability in measuring results is high. Fortunately, if there is doubt about the interpretation of a skin test, multiple readers can measure the result over a period of days, or the test can be repeated and reinterpreted. Specific anergy to tuberculin is seen in several situations. Approximately 10 to 20 per cent of patients with culture-confirmed pulmonary tuberculosis fail to respond to tuberculin as a result of anergy. These patients will often mount a response after their disease has been treated. HIV-infected patients have a high prevalence of anergy, both to tuberculin and other antigens. Only 10 to 40 per cent of patients with low CD4 counts and confirmed tuberculosis respond to tuberculin. Transient anergy is associated with acute viral infections such as measles, or live virus vaccinations, and other acute medical illnesses.

Microscopic staining

Microscopic staining of acid-fast bacilli is the method most widely used to diagnose tuberculosis throughout the world. Acid-fast staining is inexpensive, rapid, and technologically undemanding, making it an attractive technique for identifying mycobacterial infections. The waxy glycolipid matrix of the mycobacterial cell wall is resistant to acid-alcohol decolorization after staining with carbolfuchsin dyes, and red bacilli are visible after counterstaining. Both the Ziehl–Neelsen method (which requires heat fixation) and the Kinyoun method utilize methylene blue or malachite green counterstains, and have similar sensitivities for identifying acid-fast bacilli in clinical specimens.

The major limitation of acid-fast staining is that a relatively large number of bacilli must be present to be seen microscopically. Acid-fast smears are generally negative when there are fewer than 10 000 bacilli/ml of sputum, and many microscope fields need to be examined to identify bacilli even when there are 10 000 to 50 000 bacilli/ml. Thus, up to 50 per cent of patients with sputum cultures positive for *M. tuberculosis* have negative acid-fast smears. Where the sputum smear is the only test performed to confirm tuberculosis, a large number of smear-negative cases go unde-tected. This is a serious problem for patients without cavitary tuberculosis, who tend to have fewer bacilli in their sputum, including many HIV-infected patients with tuberculosis in developing countries.

Several techniques can be used to improve the yield of sputum smears. The most important method is enrichment of the specimen through concentration of the sputum. Centrifugation of sputum allows examination of the bacilli-rich pellet, which improves the sensitivity of smears substantially. Treatment of sputum with mucolytic agents is also helpful in identifying organisms by both smear and culture. Use of fluorochrome procedures to identify mycobacteria is more sensitive, but less specific, than acid-fast stains. Auramine O or auramine-rhodamine dyes are used on concentrated smears and examined under a fluorescence microscope. This technique allows much more rapid screening of slides that the traditional methods, but confirmation of positive results with Ziehl–Neelsen or Kinyoun staining is essential, as false-positive fluorochrome results are not uncommon.

The proper collection of specimens is also important for optimizing the results of microscopy and culture. Early morning sputum specimens tend to have a higher yield than specimens collected at other times, and overnight sputum collections provide even greater sensitivity. Morning gastric aspirates have a moderate yield for acid-fast bacilli in children, who generally have a difficult time producing sputum. Sputum induction with hypertonic saline is useful in evaluating patients with minimal or no sputum production, and the use of fibreptic bronchoscopy is often advocated for patients with negative sputum smears. In several series, however, the yield of post-bronchoscopy spontaneous sputum samples was higher than the bronchoalveolar lavage fluid. While the goal of sputum collection is to collect a pure lower respiratory tract sample, specimens that appear to consist primarily of upper respiratory tract or oral secretions often are smear or culture positive in patients with pulmonary tuberculosis.

Examination of multiple specimens increases the sensitivity of sputum microscopy for acid-fast bacilli. The first smear identifies 70 to 80 per cent of patients, the second another 10 to 15 per cent, and the third another 5 to 10 per cent. Review of additional specimens has little value.

In addition to the modest sensitivity of acid-fast staining, the specificity of this technique can also present problems. The morphological properties of the mycobacteria are sufficiently similar to make distinguishing *M. tuberculosis* from non-tuberculous mycobacteria impossible on the basis of acid-fast smears. This is not a serious concern where tuberculosis is common and non-tuberculous mycobacterial infections are unusual. However, in many industrialized countries, disease due to the non-tuberculous mycobacteria is relatively common compared with tuberculosis, and distinguishing these types of infections has important therapeutic and public health implications. Thus, while sputum microscopy is useful because of its rapidity and low cost, it should be supplemented with culture or other more sensitive and specific tests whenever feasible.

Culture, nucleic acid amplification, and susceptibility testing

Cultivation of *M. tuberculosis* in the laboratory is the gold standard for confirming the diagnosis of tuberculosis. A variety of media are available that support the growth of mycobacteria, including egg- and potato-based solid media and several broth-based media. The intrinsic growth rate of *M. tuberculosis* makes the recovery of the organism in culture a slow process. In traditional egg-based media such as Lowenstein–Jensen, growth of colonies of *M. tuberculosis* takes between 3 and 6 weeks, and 7H11 agar requires an average of 3 to 4 weeks to show colonies. Obviously, the glacial pace of these traditional culture systems interferes with optimal patient management, and more rapid techniques are required.

Several faster (not rapid) systems for detection of mycobacteria in culture have been commercially developed. The radiometric BACTEC system (BD Biosciences) utilizes ^{14}C palmitate in 7H12 broth to detect mycobacterial growth more quickly. Uptake and metabolism of the palmitate by mycobacteria releases $^{14}CO_2$ which is detected radiometrically. The relative amount of $^{14}CO_2$ produced is used to calculate a growth index, which is considerably more sensitive than visual inspection of colonies on agar. The average time to positive culture by BACTEC is 8 to 12 days, rather than the 3 to 4 weeks required with conventional media. The technology is

automated so that regular visual inspection of culture bottles is not required, but BACTEC systems are expensive and require radioisotopes.

The Septi-Chek system combines solid and broth media. The Mycobacterial Growth Indicator Tube (MGIT) is a broth-based system that uses automated fluorescence detection to monitor growth. Both systems are more rapid than conventional culture.

Many clinical laboratories use more than one culture system for mycobacteria, both to increase the overall recovery rate and to provide quality control. In addition, if one culture becomes contaminated, alternative cultures can still be utilized.

Preparation of specimens for mycobacterial culture follows the same steps as outlined for acid-fast smears. In addition, specimens being submitted for culture also require decontamination to prevent overgrowth by more rapidly multiplying bacteria. Sodium hydroxide and *N*-acetyl-L-cysteine are commonly used together for mucolysis and decontamination. By necessity, decontamination also inactivates more than 50 per cent of mycobacteria in a specimen, thereby reducing the potential yield of the culture. Failure to decontaminate, however, leads to bacterial overgrowth and uninterpretable results. Lack of growth as a result of over-decontamination and bacterial overgrowth resulting from under-decontamination emphasize the importance and utility of obtaining multiple specimens for culture, when possible. As with sputum smears, the yield of mycobacterial culture increases with evaluation of additional specimens.

After mycobacterial growth has been identified, speciation of the organism is required. Conventional techniques for identification involve characterization of colony morphology, pigmentation, rate of growth, and biochemical tests. Niacin reduction, nitrate reduction, and lack of catalase activity at elevated temperatures are all characteristic of *M. tuberculosis*. Species identification using these methods is time consuming and tedious, and further delays the diagnosis of tuberculosis.

The use of nucleic acid probes has dramatically simplified species identification of mycobacteria over the past decade. DNA probes that react with specific mycobacterial rRNA sequences to form DNA–RNA hybrids that can readily be detected by chemoluminescence are commercially available for *M. tuberculosis*, *M. avium* complex, *M. kansasii*, and *M. gordonae*. These tests can be performed within hours of detection of mycobacterial growth, and accelerate the diagnosis of specific pathogens. The sensitivity of these probes is approximately 90 to 95 per cent, depending on the species, with specificities approaching 100 per cent. Cultures that fail to respond to any of the DNA–RNA probes are almost always due to another mycobacterial species, but final identification depends on the traditional laborious biochemical techniques.

The difficulties of identifying mycobacteria in patient specimens accentuate the need for rapid and sensitive diagnostic methods for tuberculosis. If any infection seems suited to diagnosis by nucleic acid amplification assays, it would appear to be tuberculosis. Multiple studies of 'in-house' PCR assays for *M. tuberculosis* have shown modest sensitivity and specificity. PCR inhibitors in sputum have been a knotty problem in the molecular diagnosis of pulmonary tuberculosis, although the sensitivity has been lower than that of culture in non-respiratory specimens as well. Recently, several commercial nucleic acid amplification tests have been introduced or are nearing approval, including assays based on RT-PCR, transcription-mediated amplification, ligase chain reaction, and strand displacement amplification. All of these techniques use specific *M. tuberculosis* DNA sequences (most use the *M. tuberculosis* transposon *IS*6110) as targets for nucleic acid amplification. The great advantage of these assays is that they can provide results within 1 day of the collection of specimens. Their disadvantage is that they are uniformly less sensitive than culture, particularly in patients who have negative sputum smears. Early studies of these techniques have suggested that specificity was excellent overall but was reduced in smear-positive samples; further refinement in these assays has resulted in improved sensitivity and specificity.

Evaluation of nucleic acid amplification assays under field conditions has generally shown favourable results. When using these tests, however, clinicians must not forget fundamental clinical and epidemiological principles governing the diagnosis of tuberculosis: a negative test in a patient suspected of having tuberculosis should not exclude the diagnosis, nor should a positive test confirm it if clinical circumstances do not support the diagnosis. While both the positive and negative predictive values of nucleic acid amplification tests are high (70 to 90 per cent and over 90 per cent, respectively), misclassification of patients does occur, and it is important to use mycobacterial culture to validate the results of these rapid assays.

Drug susceptibility testing of *M. tuberculosis* isolates is essential for both clinical management and public health purposes. Susceptibility tests for the first-line antituberculosis drugs should be performed on at least one culture at the time of diagnosis for all patients. If the initial isolate is susceptible to the first-line agents, and treatment proceeds without incident, additional susceptibility tests are not required. Susceptibility testing should be performed for patients who relapse with tuberculosis and for patients whose treatment is a failure after 3 to 4 months of therapy.

Susceptibility testing for *M. tuberculosis* uses standard concentrations of antituberculosis drugs to measure inhibition of bacterial growth in culture. Drugs tested routinely include isoniazid, rifampicin, pyrazinamide, ethambutol and streptomycin. Testing of second-line antituberculosis drugs is only done when resistance to the first-line agents is documented or strongly suspected.

Susceptibility testing is generally performed on subcultures of the primary isolate, although direct inoculation of sputum or other specimens can be performed in the case of a strongly positive acid-fast bacilli smear. The standard method for measuring susceptibility to antituberculosis drugs is the proportions method. The organism is grown on agar plates in the presence of known concentrations of specific drugs. Growth on the plates is then compared with growth on control plates. By convention, if the test plate shows a colony count that is more than 1 per cent of the control value, the isolate is resistant. Laboratories will report the isolate as being susceptible or resistant to the concentration of the drug used in the assay.

An alternative method for susceptibility testing is the BACTEC system, in which culture bottles contain antituberculosis drugs. Growth indices are compared with control cultures to determine susceptibility. The BACTEC system provides results more quickly than the proportions method, is automated, but is more expensive. Other automated commercial culture systems have also been developed for determining drug susceptiblity.

The use of molecular methods to determine drug susceptibility is promising but not currently in routine use. Specific mutations in *M. tuberculosis* have been identified which confer resistance to antituberculosis drugs. For example, mutations in a small region of the *rpo*B gene of *M. tuberculosis* are responsible for more than 90 per cent of all rifampicin resistance. Sequencing of this portion of the genome using a variety of techniques has been shown to be feasible in research laboratories. Rapid identification of rifamicin resistance would be of enormous clinical benefit, as almost all rifampicin-resistant *M. tuberculosis* isolates are also resistant to isoniazid and, by definition, multidrug resistant. Molecular diagnosis of other types of resistance is more difficult, as the genetic basis of resistance to other drugs is either heterogenous or not completely understood.

Treatment of active tuberculosis

The treatment of tuberculosis requires the use of a combination of antimycobacterial drugs active against the strain of *M. tuberculosis* causing the patient's disease. The use of multiple agents is necessitated by the emergence of drug resistance when single agents are used. Mutations that confer resistance to antimycobacterial drugs arise spontaneously in wild-type populations of *M. tuberculosis* in frequencies ranging from 1 in 10^5 to 1 in 10^8 bacilli. When there are large numbers of organisms, such as are present during active pulmonary disease, a single agent will kill susceptible bacilli, but naturally drug-resistant mutants will survive and eventually emerge to cause drug-resistant disease. Since the mechanisms of resistance are genetically distinct and arise independently, multiple drug resistance within a single organism is exceedingly rare in nature. The use of two or more agents with different mechanisms of action assures that populations of drug-resistant bacilli are not selected for during therapy.

Drugs for tuberculosis are divided into first-line and second-line agents. First-line agents are widely available and used routinely in the treatment of tuberculosis, while second-line agents are generally less potent, more toxic, and less readily available. An exception to this is the fluoroquinolones, which appear to have moderately good antituberculosis activity and are widely available; their utility in tuberculosis, however, remains unstudied. Second-line drugs are reserved for the treatment of drug-resistant tuberculosis. Table 5 lists the first-line antituberculosis drugs, their activity in the treatment of tuberculosis, and common toxicities.

Regimens currently used for the treatment of tuberculosis are based in part on trials conducted by the British Medical Research Council over the past 30 years. By combining drugs that target both rapidly growing bacillary populations and slow-growing or semi-dormant organisms within cells, modern short-course chemotherapy can successfully cure drug-susceptible pulmonary tuberculosis in 6 months. The regimens recommended for treatment of drug-susceptible tuberculosis are shown in Table 6. Treatment of extrapulmonary tuberculosis is generally for the same duration as for pulmonary disease, with the exceptions of bone and joint and central nervous system tuberculosis, which are treated for 12 months. HIV-related tuberculosis is also treated for 6 months.

The dynamics of mycobacterial growth are such that treatment needs to be administered only once daily, and can be given as infrequently as twice per week. The long generation time of *M. tuberculosis* and a postantibiotic effect of antituberculosis drugs renders more frequent drug dosing unnecessary. The dosages for drugs

are listed in Table 7 according to the frequency with which they are administered.

Isoniazid is a key component of treatment because of its high bactericidal activity. Rifampicin is essential for short-course therapy because it is active against all populations of bacilli, both within and outside cells. Pyrazinamide is uniquely active during the first 2 months of therapy, but appears to have no activity thereafter. The addition of pyrazinamide to the treatment regimen allows the duration of therapy to be reduced from 9 to 6 months, however. Streptomycin has bactericidal activity against *M. tuberculosis*, and ethambutol has bacteriostatic activity at lower doses and bactericidal activity at high doses. These latter agents are given primarily to prevent the emergence of drug resistance, as they appear to add little activity to combination regimens against drug-susceptible tuberculosis.

Although antituberculosis therapy is remarkably well tolerated and almost always given to ambulant patients, important drug toxicities do exist. The most serious adverse drug reaction during tuberculosis treatment is liver toxicity, which may occur in up to 5 to 10 per cent of treated patients. Isoniazid, rifampicin, and pyrazinamide are all associated with liver toxicity. Use of these agents together increases the risk of a reaction. Isoniazid causes more hepatotoxicity than rifampicin or pyrazinamide, and is the agent most frequently implicated when reactions occur. Isoniazid can produce an idiosyncratic hepatocellular injury, manifested by elevated liver enzymes and clinical hepatitis. Elevation of transaminases does not always portend the development of hepatitis, but may serve as an important signal to anticipate clinical toxicity. The development of signs and symptoms of hepatitis, such as abdominal pain, nausea, vomiting, or jaundice, requires immediate discontinuation of isoniazid, as continuing treatment may result in death from hepatic failure. Risk factors for developing isoniazid hepatotoxicity include increasing age, chronic liver disease, alcohol abuse, daily dosing of isoniazid, and use of other hepatotoxic drugs, including rifampicin. In addition, people with a slow isoniazid-acetylation genotype are significantly more likely to develop hepatoxicity from the drug than intermediate or rapid acetylators.

Isoniazid interferes with metabolism of pyridoxine (vitamin B_6), which can result in a sensory neuropathy. Co-administration of pyridoxine with isoniazid abrogates this effect without compromising the antimicrobial activity.

Rifampicin also causes hepatotoxicity, although the characteristic picture of liver disturbances due to rifampicin is cholestasis. However, the incidence of hepatotoxicity when rifampicin is given with isoniazid is substantially greater than when isoniazid is given alone. Rifampicin predictably causes a discoloration of body fluids, resulting in orange-tinted tears, sweat, and urine. Haematological toxicity from rifampicin includes thrombocytopenia and anaemia. Higher doses of rifampicin may produce a hypersensitivity reaction, with fever, rash, and joint swelling. For this reason, doses of rifampicin are not escalated during intermittent therapy, whereas the intermittent dosages of the other drugs are increased to deliver weekly doses that are equivalent to daily dosing.

Pyrazinamide is often associated with arthralgias, and may precipitate gout. Pyrazinamide inhibits renal tubular excretion of uric acid, resulting in increased serum levels of uric acid. Frank gouty arthritis is relatively uncommon with pyrazinamide use, and its frequency is reduced with intermittent dosing. Routine use of allopurinol to prevent gout is not recommended.

The major toxicity of ethambutol is optic neuritis, which is common at doses above 30 mg/kg daily and unusual at doses below 25 mg/kg daily. Patients receiving ethambutol should have baseline tests of visual acuity and colour discrimination, with monthly monitoring while on treatment. Ethambutol use is discouraged in children under 7 years old because of their inability to report visual disturbances reliably. The incidence of optic neuritis with the doses of ethambutol typically used is so low that use in young children is only relatively contraindicated.

Streptomycin was a staple of antituberculosis therapy for many years, but its use has been greatly curbed in recent years. A number of studies have demonstrated that regimens containing isoniazid, rifampicin, and pyrazinamide are equally efficacious with or without streptomycin. Streptomycin is given by intramuscular injection, causing discomfort to patients and creating an infection risk for patients and health care workers. In addition, streptomycin can be ototoxic and nephrotoxic. Consequently, ethambutol has replaced streptomycin in many parts of the world.

Patients receiving therapy for tuberculosis require regular monitoring to assess compliance, clinical response, and adverse reactions. In the initial phase of therapy, monitoring by a nurse or other trained clinician at least weekly is recommended, and supervision of every dose of medication is suggested by the World Health Organization and other authorities (see below). Patients should be observed for clinical responses, including defervescence, improvement in cough and appetite, and weight gain. Improvement in these symptoms and signs may take several weeks, but usually occurs within 3 weeks after starting treatment. Failure to improve suggests that the patient is not adhering to treatment, has drug-resistant tuberculosis, or has another illness in addition to or instead of tuberculosis.

Treatment response should also be documented with repeated sputum smears and cultures and a follow-up chest radiograph after 2 to 3 months (for pulmonary tuberculosis). All patients should have a repeat sputum smear and culture after 2 months of therapy; those who are smear or culture positive at 2 months should have another at 3 months. Failure to convert sputum smears and cultures to negative with 3 months of therapy is associated with a high risk of treatment failure; patients who are still smear or culture positive at 4 months of treatment are considered to have experienced treatment failure and should be evaluated for drug-resistant disease. A culture at the end of therapy is recommended to document cure, while a radiograph at this time is not necessary.

Monitoring for drug toxicity is also required throughout therapy. At least monthly monitoring for symptoms and signs of liver toxicity is essential, and patients should be advised to stop therapy and seek care if evidence of hepatitis is noted. Routine liver enzyme monitoring is recommended primarily for patients with underlying liver disease or baseline abnormalities in liver enzymes. Patients with symptoms of hepatitis, of course, should have liver studies obtained. As noted above, monthly visual assessment is also recommended when ethambutol is given.

For more than 40 years, experts in tuberculosis have noted that the success of treatment depends largely on adherence to therapy. Poor adherence to therapy is responsible for treatment failures, early relapses, and the emergence of drug-resistant disease. Two major interventions to improve adherence and prevent bad outcomes are directly observed therapy (**DOT**) and the use of fixed-dose combination tablets. DOT was first promoted in the 1950s in India, and experience with DOT grew over the ensuing years. Intermittent dosing of tuberculosis therapy, along with the relatively short course of treatment, make supervision of treatment feasible in many situations. Ecological and programmatic studies of DOT programmes have shown that their introduction improves cure rates for tuberculosis, reduces non-compliance, and reduces the emergence of drug-resistant disease. Two observational studies have shown better survival of HIV-infected patients with tuberculosis who receive DOT.

On the other hand, two randomized trials of DOT in developing countries have not found improved treatment completion rates compared with self-administered treatment. These trials have been criticized for demonstrating only that even DOT can be carried out badly, but the lack of randomized studies documenting that DOT *per se* leads to improved outcomes is of some concern. The data from observational studies are compelling, however, and DOT is strongly encouraged by many experts and professional organizations.

The use of fixed-dose combination tablets is intended to reduce the risk of selecting for drug resistance, as opposed to improving adherence generally. By combining two, three, or four medications in the same tablet, depending on the regimen being used, the opportunity for patients to receive partial treatment that would select for drug resistance is avoided. The bioequivalence of fixed-dose combinations to individual medications has been established for some, but not all, of the combination products on the market.

The catastrophic state of global tuberculosis control led the World Health Organization (**WHO**) to develop the **DOTS** strategy (or directly observed therapy, short-course). This strategy is a series of policies related to national tuberculosis control practices. The five elements of the DOTS strategy are:

1. governmental commitment to tuberculosis control;
2. a reliable supply of tuberculosis drugs;
3. diagnosis of tuberculosis cases microscopically;
4. a registration system for tracking the outcomes of treatment; and
5. supervision (DOT) of at least the first 8 weeks of treatment.

The DOTS strategy has been extremely successful in focusing attention on serious problems in tuberculosis treatment and control, and implementation of the programme in a number of countries has produced remarkable improvements in clinical outcomes for patients with tuberculosis. There is strong evidence that the use of the DOTS strategy results in lower rates of drug-resistant tuberculosis. None the less, the WHO estimates that in 2000 only 25 per cent of patients with tuberculosis in the world were treated within a DOTS programme. Further expansion of the DOTS strategy and improvements in tuberculosis treatment programmes are clearly needed.

The treatment of drug-resistant tuberculosis is beyond the scope of this chapter. Patients with drug-resistant tuberculosis should be managed by a physician who is a tuberculosis expert. Supervised therapy is considered mandatory for patients with resistant tuberculosis. Physician mistakes remain one of the leading causes of the emergence of drug resistance, and the identification of a drug-resistant isolate of *M. tuberculosis* should result in immediate expert consultation.

Treatment of latent tuberculosis infection

Prevention of tuberculosis with isoniazid therapy was first documented in children in the mid-1950s. Subsequently, a number of controlled trials of isoniazid chemoprophylaxis were undertaken, and its efficacy firmly established. A meta-analysis of 11 placebo-controlled trials of isoniazid, involving more than 70 000 persons, found that treatment reduced tuberculosis incidence by 63 per cent. Among patients who adhered to more than 80 per cent of the isoniazid regimen, protection was 81 per cent. These studies also showed that isoniazid chemoprophylaxis reduced tuberculosis deaths by 72 per cent. The efficacy of isoniazid therapy in preventing tuberculosis in high-risk persons is incontrovertible.

Enthusiasm for isoniazid chemoprophylaxis was considerably dampened in the late 1960s and early 1970s when drug-related hepatotoxicity, including deaths, was observed. A number of studies based on decision analysis or modelling suggested that the risks of chemoprophylaxis might outweigh the benefits, and use of preventive therapy was curtailed or ignored in many settings. Because the risk of isoniazid-related hepatotoxicity increases with age, use of chemoprophylaxis in people over 35 years old was particularly discouraged. The resurgence of tuberculosis in the developed world, particularly HIV-related tuberculosis, and the uncontrolled global epidemic have renewed interest in the use of preventive therapy in high-risk individuals.

The use of preventive therapy for tuberculosis now focuses on high-risk groups of individuals who are either known or strongly suspected to be latently infected with *M. tuberuclosis*. The term 'treatment of latent tuberculosis infection' is now preferred, emphasizing that preventive treatment is really targeted at an established infection. The American Thoracic Society and the Centers for Disease Control and Prevention published guidelines in 2000 on screening for latent tuberculosis that stress the importance of targeting efforts on populations and patients who would benefit from treatment to prevent active disease. In the past, screening for tuberculosis infection has been unfocused and often directed at patients who, if found to be infected, would have little risk of progressing to active disease. The new guidelines propose that only people with a high risk of disease or high prior probability of latent tuberculosis be tested, and that treatment be offered to infected individuals regardless of age. Individuals who should be targeted for tuberculin testing are those listed in the first two columns of Table 4, that is, those in whom a positive test is considered as 5- or 10-mm or more induration. People without risk factors for tuberculosis (those in whom a positive test is 15 mm or more) should not be tested.

Treatment regimens for latent tuberculosis are listed in Table 8, along with the rating given to the regimen by the American Thoracic Society (ATS) and Centers for Disease Control and Prevention (CDC). Isoniazid remains a favoured drug for tuberculosis preventive therapy because of its well-documented efficacy, low cost, and relatively low toxicity. The optimal duration of isoniazid therapy for latent tuberculosis has been the subject of extensive debate in the past 20 years. The International Union Against Tuberculosis and Lung Disease conducted a landmark trial in Eastern Europe in the 1970s and 1980s that compared no treatment with 3, 6, or 12 months of isoniazid in adults with fibrotic changes on radiography. The results showed that, compared with placebo, 12 months of isoniazid reduced the incidence of tuberculosis by 75 per cent, compared with 66 per cent for 6 months and 20 per cent for 3 months. In addition, patients who completed the 12 months of therapy and were judged to be compliant experienced a 92 per cent reduction in tuberculosis risk, compared with a 69 per cent decrease for compliant patients completing a 6-month regimen. A meta-analysis by the Cochrane Collaborative found that 12 months of isoniazid was more effective than 6 months for prevention of tuberculosis. A recent analysis of varying durations of isoniazid therapy in Alaskan natives revealed that the effectiveness of isoniazid therapy was optimal after 9 months, and that further treatment conferred no additional benefit. The new ATS/CDC statement, therefore, recommends 9 months of isoniazid as the preferred regimen, with 6 months considered an alternative, but less effective, course of treatment.

Although isoniazid is a well tolerated drug, serious hepatotoxicity can occur in a small proportion of patients. Isoniazid may result in asymptomatic elevations in hepatic transaminase levels, but this does not always signal impending clinical toxicity. Hepatotoxicity is of concern when symptoms of hepatitis, including pain, nausea, vomiting, and jaundice, develop. Continuing isoniazid in the presence of symptoms may lead to death from fulminant hepatic necrosis and liver failure, with a case–fatality rate of 10 to15 per cent. Studies in the 1960s and 1970s found evidence of hepatotoxicity in 1 to 5 per cent of isoniazid recipients, with higher rates among older patients. More recent experience with isoniazid therapy that is closely monitored shows a risk of hepatotoxicity in the range of 0.1 to 0.3 per cent. Thus, appropriate patient screening and follow-up makes the use of isoniazid for treating latent infection markedly safer.

One of the most important new developments in the treatment of latent tuberculosis is the development of alternative regimens that shorten the duration of treatment. Based on studies in animal models of latent tuberculosis, rifampicin alone given for 3 to 4 months, or rifampicin and pyrazinamide given for 2 to 3 months, were felt to be potentially active regimens and were tested in clinical trials. A 3-month regimen of rifampicin alone was found to reduce the incidence of tuberculosis by about 65 per cent in men with silicosis, and was more effective than 6 months of isoniazid. Three studies of rifampicin and pyrazinamide for latent tuberculosis in HIV-infected, tuberculin-positive patients have been carried out. In each of these studies, the combination of rifampicin and pyrazinamide was as effective as 6 or 12 months of isoniazid. A meta-analysis of the studies found rifampicin with pyrazinamide was equivalent to isoniazid for preventing active tuberculosis, with an odds ratio of 1.0.

Rifampicin with pyrazinamide is generally well tolerated, but can be associated with serious hepatotoxicity. However, the use of rifampicin does pose the risk of important drug interactions. For example, reduction in methadone concentrations caused by rifampicin can precipitate narcotic withdrawal. Moreover, rifampicin can lower levels of protease inhibitors and non-nucleoside reverse transcriptase inhibitors used to treat HIV infection. Substitution of rifabutin for rifampicin in patients receiving anti-HIV drugs is based on the observation that rifabutin is equally as efficacious as rifampicin in the treatment of active tuberculosis.

Candidates for treatment of latent tuberculosis are listed in Table 4. Criteria for treatment include a positive tuberculin test according to the categories in Table 4, elevated risk for developing active tuberculosis if untreated, and exclusion of active tuberculosis by clinical evaluation and chest radiography. In addition, HIV-infected and other severely immunocompromised persons who are contacts of a patient with infectious tuberculosis should be treated for latent tuberculosis regardless of tuberculin skin test results.

Patients receiving treatment for latent tuberculosis should be monitored for drug toxicity, as well as to promote adherence to therapy. As in the treatment of active tuberculosis, patients receiving isoniazid should be warned about signs and symptoms of hepatotoxicity and advised to discontinue therapy and seek care if any of these occur. Patients with or at risk for chronic liver disease should have baseline liver enzymes obtained, with monthly monitoring if the results are abnormal. All patients should be clinically evaluated at least monthly to assess toxicity and those receiving rifampicin and pyrazinamide more often. Treatment of patients with mild transaminase elevations (3 times upper limits of normal or less) can proceed with regular clinical and laboratory monitoring. Higher elevations of transaminases, or the development of symptoms or signs of hepatitis, should be managed with discontinuation of therapy at least temporarily. Patients who complete therapy for latent tuberculosis do not require any periodic monitoring for tuberculosis subsequently.

Prevention of tuberculosis

Strategies to control tuberculosis are aimed at the prevention of the spread of *M. tuberculosis* infection and the development of clinical tuberculosis. The principal approaches employed toward this end are:

1. identification and treatment of infectious tuberculosis cases;
2. treatment of latent tuberculosis infection;
3. prevention of exposure to infectious particles in air, especially in hospitals and other institutions; and
4. vaccination.

Case identification and treatment reduces transmission by rendering patients with communicable tuberculosis non-infectious. Patients with pulmonary tuberculosis produce infectious aerosols that may transmit tubercle bacilli to contacts breathing the same air. When cases are identified and treated, infectiousness is rapidly eliminated. The duration of treatment required to prevent further transmission of infection is not known precisely, but experimental, clinical, and microbiological data suggest that the level of infectiousness is reduced enormously within several days of beginning effective treatment. The number of secondary infections generated by a patient with infectious tuberculosis varies greatly, depending on the duration of illness, the extent of pulmonary pathology, the amount of patient coughing, and the environment into which the patient expels infectious aerosols. Early diagnosis and treatment reduces the number of secondary infections, while delays can result in ongoing transmission to large numbers of contacts. Failure to retain patients in treatment until they are cured also contributes to spread of infection.

Treatment of latent tuberculosis infection has been discussed above (Table 8). The benefit of treating latent infection is not only to the individual patient, who does not fall ill with tuberculosis, but also accrues to the potential contacts of that patient, who might become secondarily infected were disease to develop. Targeting of high-risk groups for screening and treatment of latent tuberculosis thereby reduces tuberculosis incidence within communities. Groups that should be targeted for screening are listed in the first two columns of Table 4.

Control of exposure to infectious aerosols can have a major impact on the spread of tuberculosis. In the late 1980s and early 1990s, transmission of tuberculosis, including multidrug-resistant tuberculosis, was widespread in hospitals, shelters for the homeless, and correctional facilities in New York City. The congregation of large numbers of highly susceptible people, especially HIV-infected persons, in closed environments with individuals with untreated tuberculosis resulted in numerous microepidemics of both drug-susceptible and drug-resistant tuberculosis. Reversal of the resurgence of tuberculosis in New York at that time was attributable in large part to strengthening of infection control practices.

Tuberculosis infection control involves prompt identification and isolation of patients with suspected tuberculosis. The decision to isolate a patient in a hospital setting is a function of epidemiological and clinical factors. Patients with known risk factors for tuberculosis who present with symptoms and signs characteristic of pulmonary tuberculosis should be placed in respiratory isolation. Local epidemiological data should influence isolation practices. Where tuberculosis is prevalent, all HIV-infected patients with pneumonia may require isolation, whereas isolation is more selective and based on individual patient features in low-prevalence settings.

Respiratory isolation requires nursing the patient in a room with negative air pressure relative to adjoining areas. Ventilation to the room should provide at least six complete air changes per hour, and air should not be recirculated without filtering or irradiation. Patients should be instructed always to cover their mouths when coughing, and should wear surgical face masks when outside the room to reduce aerosol generation. Anyone entering the patient's room should wear an appropriate face mask or respirator to prevent inhalation of droplet nuclei with tubercle bacilli. Much debate has occurred in recent years in the United States about what constitutes appropriate protection for health care workers exposed to infectious tuberculosis. This debate is influenced as much by philosophy as by science, and will not be detailed here. Use of surgical masks for protection against tuberculosis is clearly inappropriate, even though these masks are useful when placed on patients to prevent creation of infectious aerosols. Tightly fitting face masks that filter out more than 99.7 per cent of particles greater than 0.5 µm in size (high efficiency particle air—HEPA—filters) are effective. Other devices, including positive air pressure respirators (PAPRs), are also effective.

Ultraviolet germicidal irradiation can be useful for reducing the number of infectious particles in ambient air in settings where ventilation alone is not sufficient. Ultraviolet light must be concentrated in areas of rooms where exposure to people will not occur, such as upper air zones, in order to prevent skin and ocular toxicity. Areas where ultraviolet lights are often used include bronchoscopy suites, inside air circulation ducts, in emergency rooms, and in shelters for the homeless.

Criteria for discontinuation of respiratory isolation are listed in Table 9. Guidelines for taking patients out of isolation in the hospital are strict and are intended to protect other vulnerable patients and hospital staff from any exposure to the disease. Respiratory isolation is not usually required or practical in the home setting, and patients with infectious tuberculosis do not need to be admitted to hospital solely for respiratory isolation. It is assumed that contacts in the home environment will already have significant exposure to tuberculosis by the time a diagnosis is made, and isolation of the patient affords no measurable benefit. Exceptions to this may include patients living in congregate living facilities or other special situations. The primary protective measures for contacts of infected individuals are a clinical evaluation to identify and evaluate symptoms of tuberculosis and tuberculin skin testing, with treatment of latent infection if present.

Vaccination against tuberculosis with BCG vaccine is widely administered throughout the world, but remains controversial. BCG is an attenuated live bacterial vaccine developed in the early twentieth century by Calmette and Guérin at the Institut Pasteur. After a series of uncontrolled and anecdotal assessments of the vaccine, a series of controlled trials of BCG was begun in the 1930s and continued through the 1990s. The efficacy of BCG has varied greatly in these studies, ranging from more than 80 per cent protection to complete lack of protection, with possibly increased risk in vaccine recipients. A meta-analysis of BCG trials performed in the early 1990s found that the weighted protective benefit of BCG was about 50 per cent for both the prevention of active tuberculosis disease and death.

In addition to the protective efficacy observed in trials of BCG, there is evidence that BCG diminishes haematogenous dissemination of primary tuberculosis infection and thereby reduces the incidence of miliary tuberculosis and tuberculous meningitis in children. It is primarily for this reason that BCG is included in the Expanded Program on Immunization of the WHO.

The current efficacy of BCG for preventing pulmonary tuberculosis is debated on the basis of several recent trials which have failed to show protection. A number of hypotheses have been proposed for the variation in efficacy reported in various studies, including differences in susceptibility within populations, environmental exposure to mycobacteria which masks vaccine effect, and attenuation of vaccine immunogenicity. The last explanation is very compelling and fits well with clinical trials data. Unlike most vaccines, BCG is not standardized and there is no seedlot of vaccine from which new batches are derived. BCG is grown in a number of laboratories around the world and has not been re-passaged in animals since it was derived from cattle a century ago. Multiple commercial and non-commercial BCG products are in use at present, and comparative genomic analysis demonstrates considerable genetic heterogeneity in these strains, with many gene deletions and polymorphisms. One analysis of BCG trials found that protective efficacy was reduced in studies using multiply-passaged vaccine strains. The evidence supports the hypothesis that BCG has become further attenuated over time and no longer promotes immunity to *M. tuberculosis* infection and disease in adults. This position has not been universally accepted, however, and BCG remains one of the most widely administered vaccines in the world, largely for its perceived effects on paediatric tuberculosis.

Areas for further research

Effective global tuberculosis control will require a co-ordinated set of clinical and public health strategies based on a thorough understanding of the epidemiology, pathogenesis, and therapy of infection with *M. tuberculosis*. It appears that the WHO DOTS strategy, which focuses on finding and effectively treating cases, is not sufficient to control or eliminate tuberculosis, particularly in countries with large HIV epidemics. Improved methods for the diagnosis and treatment of tuberculosis infection and disease, particularly drug-resistant tuberculosis, are urgently needed. Effective regimens for the treatment of multidrug-resistant tuberculosis, with both existing and new agents, need to be developed. A better understanding of the pathogenesis of and natural immunity to tuberculosis may contribute to the development of a more effective vaccine. The sequencing of the genome of *M. tuberculosis* promises to open the door to a new generation of research on tuberculosis and its control. Scientific progress alone, however, will be insufficient to combat tuberculosis worldwide. The willingness of societies and nations to pay for the deployment of the fruits of biomedical research, both past and future, to combat the disease where it is prevalent will be required for the conquest of tuberculosis.

Further reading

American Thoracic Society (1994). Treatment of tuberculosis and tuberculosis infection in adults and children. *American Journal of Respiratory and Critical Care Medicine* **149**, 1359–74.

American Thoracic Society/CDC (2000). Targeted tuberculin testing and treatment of latent tuberculosis infection. *American Journal of Respiratory and Critical Care Medicine* **161**, S221–47.

American Thoracic Society (2000). Diagnostic standards and classification of tuberculosis in adults and children. *American Journal of Respiratory and Critical Care Medicine* **161**, 1376–95.

Brudney K, Dobkin J (1991). Resurgent tuberculosis in New York City. Human immunodeficiency virus, homelessness, and the decline of tuberculosis control programs. *American Review of Respiratory Diseases* **144**, 745–9.

Chin DP *et al*. (1996). Reliability of anergy skin testing in persons with HIV infection. *American Journal of Respiratory and Critical Care Medicine* **153**, 1982–4.

Colditz GA *et al*. (1995). Efficacy of BCG vaccine in the prevention of tuberculosis. *Clinical Infectious Diseases* **20**, 126–35.

Comstock GW (1994). Field trials of tuberculosis vaccines: how could we have done them better? *Controlled Clinical Trials* **15**, 247–76.

Davies PDO, ed. (1998). *Clinical tuberculosis*, 2nd edn. Oxford University Press.

Dye C *et al.* (1999). Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. *Journal of the American Medical Association* **282**, 67–86.

Ellner JJ (1997). Review: The immune response in human tuberculosis: implications for tuberculosis control. *Journal of Infectious Diseases* **176**, 1351–9.

El-Sadr WM *et al*. (1998). Evaluation of an intensive intermittent-induction regimen and duration of short-course treatment for human immunodeficiency virus-related pulmonary tuberculosis. *Clinical*

Infectious Diseases **26**, 1148–58.

Ferebee SH (1970). Controlled chemoprophylaxis trials in tuberculosis: a general review. *Advances in Tuberculosis Research* **17**, 28–106.

Frieden TB *et al.* (1995). Tuberculosis in New York City—turning the tide. *New England Journal of Medicine* **333**, 229–33.

Fine PEM (1995). Variation in protection by BCG: implications of and for heterologous immunity. *Lancet* **346**, 1339–45.

Fox W, Ellard GA, Mitchison DA (1999). Studies on the treatment of tuberculosis undertaken by the British Medical Research Council tuberculosis units, 1946–1986, with relevant subsequent publications. *International Journal of Tuberculosis and Lung Diseases* **3**(Suppl 2), S231–79.

Graham NMH *et al*. (1992). Prevalence of tuberculin positivity and skin test anergy in HIV-1-seropositive and HIV-1-seronegative intravenous drug users. *Journal of the American Medical Association* **267**, 369–73.

Grzybowski S, Burnett G, Stylblo K (1975). Contacts of cases of active pulmonary tuberculosis. *Bulletin of the International Union Against Tuberculosis* **60**, 90–106.

Iseman MD (1993). Treatment of multidrug-resistant tuberculosis. *New England Journal of Medicine* **329**, 784–91.

Iseman MD (2000). *A clinician's guide to tuberculosis*. Lippincott Williams & Wilkins, Philadelphia.

Lalvani A, *et al*. (2001). Rapid detection of *Mycobacterium tuberculosis* infection by enumeration of antigen-specific T cells. *American Journal of Respiratory and Critical Care Medicine* **163**, 824–9.

Mahmoudi A, Iseman MD (1993). Pitfalls in the care of patients with tuberculosis: common errors and their association with the acquisition of drug resistance. *Journal of the American Medical Association* **270**, 65–8.

McKenna MT, McCray E, Onorato I (1995). The epidemiology of tuberculosis among foreign-born persons in the United States, 1986–93. *New England Journal of Medicine* **332**, 1071–6.

Murray CJL, Styblo K, Rouillon A (1990). Tuberculosis in developing countries: burden, intervention and cost. *Bulletin of the International Union Against Tuberculosis* **65**, 1–20.

Reichman LB, Hershfield ES, eds (2000). *Tuberculosis: a comprehensive international approach*, 2nd edn. Marcel Dekker, New York.

Reider HL, Snider DE, Cauthen GM (1990). Extrapulmonary tuberculosis in the United States. *American Review of Respiratory Diseases* **141**, 347–51.

Rom WN, Gary S, eds (1996). *Tuberculosis*. Little Brown, Boston.

Ryan F (1992). *The forgotten plague: how the battle against tuberculosis was won—and lost*. Little Brown, Boston.

Small PM *et al.* (1991). Treatment of tuberculosis in patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine* **324**, 289–94.

World Health Organization (1997). *Anti-tuberculosis drug resistance in the world. The WHO/IUATLD global project on drug resistance surveillance*, 1994–1997. World Health Organization, Geneva.

World Health Organization (2001). *Global tuberculosis control, WHO Report 2001*. World Health Organization, Geneva.

7.11.23 Disease caused by environmental mycobacteria

J. M. Grange and P. D. O. Davies

Introduction

In addition to the tubercle and leprosy bacilli, the genus *Mycobacterium* contains at least 60 species that exist naturally as environmental saprophytes and some of these occasionally cause opportunist disease in humans and animals. The environmental mycobacteria are divisible into two main groups, the slow and rapid growers, according to their rate of growth on subculture. Originally allocated to broad groups according to pigmentation and other cultural characteristics, almost all environmental mycobacteria are now readily identifiable at species level.

Most of the slow growers are able to cause human disease and the commonest pathogens are the closely related species *M. avium* and *M. intracellulare*, which are usually grouped together as the *M. avium* complex. With rare exceptions, the only pathogenic rapid growers are *M. chelonae* (including *M. abscessus*) and *M. fortuitum*. The principal pathogenic environmental mycobacteria are listed in Table 1.

The environmental mycobacteria cause two named diseases with characteristic features: swimming pool granuloma caused by *M. marinum* and Buruli ulcer caused by *M. ulcerans*. The other mycobacterioses are much less specific, often resembling tuberculosis, and require identification of the causative organism for diagnosis.

Ecology and epidemiology

The environmental mycobacteria are particularly associated with water and are found in swamps, ponds, rivers, and also colonize piped water supplies. They are readily transmissible to humans by drinking water, by inhalation of aerosols, or by traumatic inoculation. Infection of humans by environmental mycobacteria is widespread and common but overt disease is rare. In some regions such infection may be sufficient to cause cross-reactions on tuberculin testing and to modify the protective efficacy of subsequent BCG vaccination, thereby possibly explaining the diversity of protection seen in major BCG trials.

The incidence of overt disease due to environmental mycobacteria is related to the species and numbers of mycobacteria in the environment, the opportunities for infection, and the susceptibility of the human population. Person-to-person transmission of overt disease very rarely, if ever, occurs and the prevalence of such disease is unaffected by tuberculosis control measures designed to break the cycle of person-to-person transmission. In recent years there has been an increase in the incidence of disease due to environmental mycobacteria in many countries because of immunosuppression, notably due to HIV infection.

The types of environmental mycobacterial disease in humans

The environmental mycobacteria cause four main types of disease: chronic pulmonary, lymphadenitis, postinoculation, and disseminated.

Chronic pulmonary disease

This form of environmental mycobacterial disease usually occurs in patients with predisposing local lung lesions, including industrial dust disease, old tuberculous cavities, chronic obstructive pulmonary disease, cancer, cystic fibrosis, and bronchiectasis, or generalized autoimmune or immunosuppressive disorders. However, a substantial minority of cases occur in people who otherwise appear healthy. Most patients are middle aged or elderly and men are usually more frequently affected than women. Environmental mycobacterial infection is also commoner in people who smoke. In some areas of the United Kingdom the incidence of environmental mycobacterial disease in the middle aged and elderly white population exceeds that of tuberculosis.

The most frequent causes worldwide are the *M. avium* complex and *M. kansasii*. *M. xenopi* is more restricted geographically but frequently occurs in southern England while, for unknown reasons, *M. malmoense* is encountered as a pathogen with increasing frequency in many parts of Europe including northern England. Rarer causes include *M. scrofulaceum*, *M. szulgai*, and *M. chelonae*.

Clinical presentation

Symptoms develop insidiously over weeks or months and include cough, malaise, weight loss, and sweats, in a pattern similar to tuberculosis but more chronic.

There are no diagnostically reliable clinical and radiological differences between pulmonary environmental mycobacterial disease and tuberculosis and diagnosis therefore depends on the isolation and identification of the causative organism. In contrast to *M. tuberculosis*, environmental mycobacteria isolated from sputum may not be the primary cause of disease; they may be transitory contaminants of the pharynx or secondary saprophytes of diseased tissue. There are no absolute criteria for distinguishing between these possibilities, but at least two pure cultures from specimens taken at least 1 week apart from patients with compatible symptoms and radiological signs in whom other causes, including tuberculosis, have been rigorously excluded render the diagnosis very likely. In some cases, a diagnosis is made or confirmed by microbiological examination of washings, brushings, or biopsies obtained by fibre-optic bronchoscopy.

Lymphadenitis

This is principally a disease of young children, occurring most frequently in the second year of life and then declining in frequency up to the fifth year, after which it is seldom encountered. The risk is reduced by neonatal BCG vaccination. The disease usually affects the cervical lymph nodes but other nodes, such as axillary and inguinal, may be involved, especially in older patients. Lymphadenitis is caused by many mycobacterial species, the commonest cause being the *M. avium* complex and *M. scrofulaceum*. Most cases occur in otherwise healthy children with no obvious predisposing cause but some cases, particularly in older age groups, are associated with human immunodeficiency virus (**HIV**) infection.

In most cases without predisposing causes, a single node is involved and surgical excision, if technically possible, is curative. More limited treatment, such as incision and drainage, may lead to sinus formation and should be avoided. Disseminated disease may develop in a few children, particularly those with some form of congenital immune deficiency, and in HIV-positive people.

Postinoculation mycobacterioses

Buruli ulcer is thought to result from inoculation of the causative organism, *M. ulcerans*, into the skin, principally by spiky vegetation. This disease is described elsewhere.

The natural habitat of *M. marinum*, the cause of swimming pool granuloma or fish tank granuloma, is water: it enters cuts and abrasions acquired whilst indulging in aquatic activities such as swimming and tending to tropical fish-tanks. The cutaneous lesions are usually warty, although pustules and ulcers may develop. There may be 'sporotrichoid' spread of lesions along the draining lymphatics (Fig. 1). The lesions usually heal spontaneously after a few months, but chemotherapy (see below)

accelerates resolution. There have been occasional reports of tenosynovitis, carpal tunnel syndrome, osteomyelitis, and disseminated disease due to *M. marinum*.

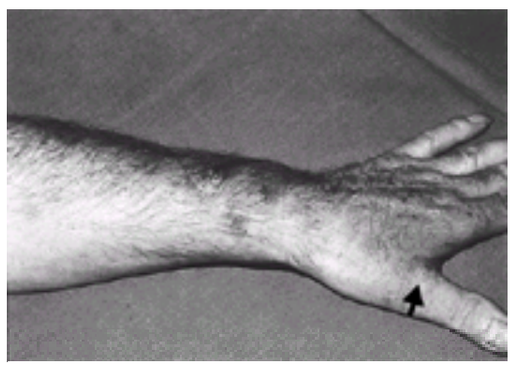

Fig. 1 *Mycobacterium marinum* infection. A small lesion at the base of the thumb (arrowed) and secondary lesions on the wrist and forearm due to 'sporotrichoid' spread (by courtesy of Dr G. Haase).

Most other cases of postinoculation disease are caused by the rapid growers *M. chelonae* and *M. fortuitum*. The most common lesions are postinjection abscesses, which may occur sporadically or in mini-epidemics due to the use of contaminated multidose vaccines or other injectable materials. Abscesses develop from 1 to 12 months after injection and may enlarge to 7 cm or more in diameter. They tend to be chronic and localized, but multiple abscesses with spreading cellulitis may develop in people with insulin-dependent diabetes. Localized abscesses usually respond well to excision or curretage, but chemotherapy (see below) may be required for multiple or spreading lesions.

Trauma to the cornea predisposes to infection by rapid growers *M. chelonae* and *M. fortuitum*. Treatment with topical amikacin and eythromycin may lead to temporary resolution but relapse is common, especially in cases due to *M. chelonae*, and corneal grafting is usually required.

More serious infections have followed accidental inoculation during surgical operations, especially when contaminated materials, including heart valve xenografts, have been inserted. Contamination during cardiac valve surgery has resulted in mycobacterial endocarditis with septicaemia and osteomyelitis of the sternum requiring extensive debridement.

Disseminated disease

Before HIV, disseminated disease due to environmental mycobacteria was very rare. Some cases, usually due to the *M. avium* complex or *M. chelonae*, occur in young people with congenital immune deficiencies (Fig. 2) and others, due principally to *M. chelonae*, occur in renal transplant recipients. *M. haemophilum* is a cause of multiple skin lesions in transplant recipients. As suggested by the name, this mycobacterium requires the addition of blood or other sources of iron in the medium for its *in vitro* cultivation.

Fig. 2 Ulcers of the lower lip as the initial manifestation of disseminated *Mycobacterium chelonae* infection in a 4-year-old girl with autosomal IgA deficiency (by courtesy of Dr K. Schopfer).

The situation changed dramatically after the advent of the HIV pandemic and disseminated environmental mycobacterial disease was reported in 30 to 50 per cent of patients with AIDS, particularly in the United States. For reasons that are not clear, the great majority of such cases, 90 per cent or more, are caused by the *M. avium* complex, usually those identifiable by DNA homology as *M. avium* rather than *M. intracellulare*. Some cases are due to *M. genevense*, a very slowly growing species which, like *M. avium*, has been isolated from diseased birds. The number of cases of disseminated AIDS-related environmental mycobacterial disease has declined in the wealthier nations following the introduction of highly active antiretroviral therapy (HAART). Although HIV infection is common in Africa, and *M. avium* is present in the environment, AIDS-related disease due to this species is, for unknown reasons, rare in that continent.

The mechanism of the establishment of this disease in humans is poorly understood. Some workers consider it to be the result of recent infection while others postulate that the disease emerges from dormant foci of infection in the lymphatic tissues of the alimentary or respiratory tracts acquired many years previously.

The symptoms—fever, night sweats, weight loss, those of anaemia and general malaise—are rather non-specific and may be caused by other AIDS-related infections. Involvement of the intestine may lead to malabsorption and chronic diarrhoea. The diagnosis of AIDS-related *M. avium* complex disease is made by culture of blood or of biopsies of liver, lymph nodes, or bone marrow. The bacilli may be isolated from faeces in disseminated disease, but they may also be present in the intestinal tract of healthy persons.

Treatment of established disease improves the quality of the remainder of the patient's life. Opinions differ as to the place for prophylactic therapy, but the introduction of highly active antiretroviral therapy makes this less relevant.

Therapy

This depends on the site and severity of the infection, the presence of predisposing conditions such as congenital or acquired immunosuppression, the species of mycobacterium, and the result of *in vitro* drug-susceptibility tests.

As indicated above, skin lesions may be cured by excision, curettage, or drainage. Surgical excision, when technically possible, is used to treat lymphadenitis and should be considered in cases of localized pulmonary lesions.

Most cases of pulmonary disease due to the *M. avium* complex and other slow-growing environmental mycobacteria respond to regimens containing rifampicin, ethambutol, and isoniazid. In contrast to tuberculosis, ethambutol appears to be more effective than isoniazid and should be continued for the full duration of therapy, provided that ocular toxicity does not occur. Treatment for 18 or 24 months produces up to 80 per cent cure rate in disease due to the *M. avium* complex, *M. xenopi*, and *M. malmoense*. Shorter regimens are effective for treatment of *M. kansasii* infections. Recommended regimens are summarized in Table 2. Surgery may be considered in certain cases where chemotherapy is ineffective.

There is evidence that the regimens based on the newer macrolides, as used to treat disseminated AIDS-related *M. avium* disease (see below), are effective in the treatment of pulmonary disease due to this complex in HIV-negative patients, but their suitability for the treatment of such diseases caused by other slow-growing species has not been established.

There have been no comparative trials of drug regimens for disease due to the rapidly growing species *M. chelonae* and *M. fortuitum*. Therapy is therefore based on anecdotal experience and the results of *in vitro* susceptibility tests. The duration of therapy depends on clinical response. Localized disease often responds to erythromycin with trimethoprim, while spreading or disseminated disease may require the addition of amikacin or a cephalosporin such as ceftriaxone. Limited experience indicates that the fluoroquinolones are effective against *M. fortuitum* and imipenem or meropenem against *M. chelonae*.

Skin lesions due to *M. marinum* respond to doxycycline or minocycline, or a combination of rifampicin and ethambutol.

The newer macrolides, clarithromycin and azithromycin, form the basis of therapy of disseminated infection, usually due to the *M. avium* complex in patients with AIDS. Commonly used regimens contain one of these together with rifabutin and ethambutol, but revision on the basis of *in vitro* drug susceptibility testing may be required. The duration of therapy depends on clinical and bacteriological response. At one time merely palliative, such regimens may be curative when combined with antiretroviral therapy. The place for prophylaxis, usually clarithromycin, is controversial and rendered less relevant by the advent of antiretroviral therapy.

Further reading

Banks J, Campbell IA (1998). Environmental mycobacteria. In: Davies PDO, ed. *Clinical tuberculosis*, 2nd edn, pp 521–33. Chapman & Hall Medical, London.

Collins CH *et al.* (1985). *Mycobacterium marinum* infections in man. *Journal of Hygiene (Cambridge)* **94**, 135–49.

Davies PDO, Ormerod LP (1999). Environmental mycobacteria. *Case presentations in clinical tuberculosis*, 259–75. Arnold Publishers, London.

Grange JM (1996). *Mycobacteria and human disease*, 2nd edn. Arnold Publishers, London.

Grange JM *et al.*. (1988). Inoculation mycobacterioses. *Clinical and Experimental Dermatology* **13**, 211–20.

Official Statement of the American Thoracic Society (1997). Diagnosis and treatment of disease caused by nontuberculous mycobacteria. *American Journal of Respiratory and Critical Care Medicine* **156**, S1–S25.

Subcommittee of the Joint Tuberculosis Committee of the British Thoracic Society (2000). Management of opportunist mycobacterial infections: Joint Tuberculosis Committee guidelines 1999. *Thorax* **55**, 210–18.

Wansborough-Jones MH, Banerjee D (1999). Non-tuberculous or atypical mycobacteria. In: James DG, Zumla A, eds. *The granulomatous disorders*, pp 189–204. Cambridge University Press, Cambridge.

7.11.24 Leprosy (Hansen's disease)

Diana N. J. Lockwood

Definition

Leprosy is a chronic granulomatous disease caused by *Mycobacterium leprae*. Its principal manifestations are anaesthetic skin lesions and peripheral neuropathy with peripheral nerve thickening. The clinical form is determined by the degree of cell-mediated immunity towards *M. leprae*. High levels of cell-mediated immunity with elimination of leprosy bacilli produces tuberculoid leprosy, whereas absent cell-mediated immunity results in lepromatous leprosy. Complications of leprosy result from nerve damage, immunological reactions, and bacillary infiltration. Nerve damage accompanying leprosy is a serious complication because it causes lifelong morbidity. Current antileprosy drugs are highly effective in killing bacilli but may not halt nerve damage. Patients with leprosy the world over are frequently stigmatized. Words such as 'leper' should be avoided and the disease should be referred to as Hansen's disease.

Aetiology

Leprosy is caused by *M. leprae*, an acid-fast intracellular organism not yet cultivated *in vitro*. It was first identified in the nodules of patients with lepromatous leprosy by Hansen in 1873. *M. leprae* preferentially parasitizes skin macrophages and peripheral nerve Schwann cells.

In vivo cultivation of *M. leprae*

M. leprae can be grown in the mouse footpad, but growth is slow, taking over 6 months to produce significant yields. The nine-banded armadillo is susceptible to *M. leprae* infection, and develops lepromatous disease. The armadillo and mouse models of *M. leprae* infection have been useful for producing *M. leprae* for biological studies and studying drug sensitivity patterns, respectively.

Biological characteristics

M. leprae is a stable, hardy organism, withstanding drying for up to 5 months. It has a doubling time of 12 days (compared with 20 min for *Escherichia coli*). The optimum growth temperature is 27°C to 30°C, consistent with the clinical observation of maximal *M. leprae* growth at cool superficial sites (skin, nasal mucosa, and peripheral nerves). *M. leprae* is a single species with isolates having similar biological characteristics and identical genotypes (using restriction fragment polymorphism analysis) irrespective of the type of leprosy, race, or geographical origin of the isolate.

Mycobacterial structure and metabolism

M. leprae possesses a complex cell wall comprising lipids and carbohydrates. It synthesizes a species-specific phenolic glycolipid and lipoarabinomannan. Antibody and T-cell screening has identified numerous protein antigens that are important immune targets.

M. leprae genome

M. leprae has a 3.27 Mb genome that displays extreme reductive evolution. Less than half the genome contains functional genes and many pseudogenes are present. One hundred and sixty-five genes are unique to *M. leprae* and functions can be attributed to 29 of them. Analysis of these unique proteins will be critical for developing new diagnostic tests. Comparison of biosynthetic pathways with *M. tuberculosis* is giving new insights into *M. leprae* metabolism. For lipolysis *M. leprae* has only two genes (*M. tuberculosis* has 22); *M. leprae* has also lost many genes for carbon catabolism and many carbon sources (e.g. acetate and galactose) are unavailable to it. This gene loss leaves *M. leprae* unable to respond to different environments and underlies the impossibility of growing the organism *in vitro*.

Epidemiology

Today, about 4 million people are disabled by leprosy. The much quoted figures of a fall in registered patients on treatment from 12 million in 1988 to 0.82 million in 1999 are misleading. Prevalence has fallen by means of effective antibiotic therapy and altered case definition. However, incidence remains stable at around 800 000 new cases annually with high rates of childhood cases. Intensive leprosy elimination campaigns held in 1998 and 1999 detected large numbers of new cases. A week-long campaign in Nepal found 11 696 new cases, doubling the national case load.

Geographical distribution

Seventy-seven per cent of patients with leprosy live in South-East Asia, 8.3 per cent in Africa, and 10 per cent in the Americas. India dominates the picture with 70 per cent of the world's leprosy cases; 86 per cent reside in six countries (India, Brazil, Indonesia, Myanmar, Madagascar, and Nepal). Leprosy has not always been a tropical disease; it was widespread in medieval Europe and was endemic in Norway until the early twentieth century. In North America, small foci of infection are still found in Texas and Louisiana. Nearly all new patients now seen in Europe and North America have acquired their infection abroad.

Risk factors

Leprosy is a chronic disease with a long incubation period. An average incubation time of 2 to 5 years has been calculated for tuberculoid cases and 8 to 12 years for lepromatous cases. American servicemen who developed leprosy after serving in the tropics presented up to 20 years after their presumed exposure. Age, sex, and household contact are important determinants of leprosy risk; incidence reaches a peak at 10 to 14 years; the excess of male cases is attributed to women's reluctance to present to health workers with skin lesions. Poor nutritional status is cited as predisposing to leprosy but no good evidence substantiates this. Improved socio-economic conditions, extended schooling, and good housing conditions reduce the risk of leprosy. Subclinical infection with *M. leprae* is probably common but the development of established disease is rare. Little work has been done on the early events in infection with *M. leprae* because there is no simple test that can establish whether an individual has encountered *M. leprae* and mounted a protective immune response.

HIV and leprosy

Studies from Malawi, Uganda, Mali, and South India have not found HIV infection to be a risk factor for leprosy. HIV/leprosy coinfected patients have typical skin lesions and typical leprosy histology and granuloma formation despite low circulating CD4 counts.

Transmission

The transmission of *M. leprae* is only partially understood. Untreated lepromatous patients discharge abundant organisms from their nasal mucosa into the environment. Studies in Indonesia and Ethiopia using polymerase chain reaction (PCR) primers to detect *M. leprae* DNA in nasal swabs have shown that up to 5 per cent of the population in leprosy endemic areas carry *M. leprae* DNA in their noses. The organism is then inhaled, multiplies on the inferior turbinates, and has a brief bacteraemic phase before binding to and entering Schwann cells and macrophages. The combination of an environmentally well-adapted organism, high carriage rates, and a long incubation period means that, even with effective antibiotics, transmission will continue for a long time.

Pathogenesis

Leprosy is a bacterial infection in which clinical features are determined by the host's immune response (Table 1).

The immune response to *M. leprae* and the leprosy spectrum

The Ridley–Jopling classification (Fig. 1) places patients on a spectrum of disease according to their clinical features, bacterial load, and histological and immunological responses. The two poles of the spectrum are tuberculoid (**TT**; paucibacillary) and lepromatous leprosy (**LL**; multibacillary). At the tuberculoid pole, well-expressed cell-mediated immunity effectively controls bacillary multiplication with the formation of organized epithelioid-cell granulomas; at the lepromatous pole there is cellular anergy towards *M. leprae* with abundant bacillary multiplication. Between these two poles is a continuum, varying from the patient with moderate cell-mediated immunity (borderline tuberculoid, **BT**) through borderline (**BB**) to the patient with little cellular response, borderline lepromatous (**BL**). The polar groups (TT, LL) are stable, but within the central groups (BT, BB, BL) the disease tends to downgrade to the lepromatous pole in the absence of treatment and upgrading towards the tuberculoid pole may occur during or after treatment.

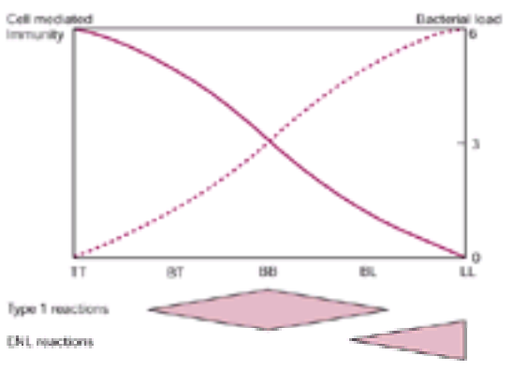

Fig. 1 The Ridley–Jopling spectrum of bacterial load, cell-mediated immunity, and reactions.

Both T cells and macrophages play important roles in the processing, recognition, and response to *M. leprae* antigens. In tuberculoid leprosy, *in vitro* tests of T-cell function such as lymphocyte transformation tests show a strong response to *M. leprae* protein antigens with the production of T_{H1}-type cytokines (interferon-g and interleukin 2, **IL-2**). Skin tests with lepromin, a heat-killed *M. leprae* preparation, are strongly positive. Staining of skin biopsies from tuberculoid lesions with T-cell markers shows highly organized granulomas composed predominantly of CD4 cells and macrophages with a peripheral mantle of CD8 cells. This strong cell-mediated immune response clears bacilli but with concomitant local tissue destruction, especially in nerves.

Patients with lepromatous leprosy have no cell-mediated immunity to *M. leprae* with a failure of the T-cell and macrophage response. Tests for lepromin are negative. This anergy is specific for *M. leprae*. Patients with LL disease respond to other mycobacteria such as *M. tuberculosis*, both *in vitro* and in skin tests. Identification of cell types in LL granulomas shows a disorganized mixture of macrophages and T cells, mainly CD8 cells. The T-cell failure may be due to clonal anergy or active suppression. Defects in cytokine production have been demonstrated; intralesional injections of recombinant IL-2 reconstitute the local immune response with elimination of *M. leprae* from macrophages. The T-cell cytokines that are produced are of the T_{H2} type. Macrophage defects described in LL disease include: defective antigen presentation and recognition, defective IL-1 production, a failure of macrophages to kill *M. leprae*, and a macrophage suppression of the T-cell response. Patients with lepromatous leprosy produce a range of autoantibodies that are both organ specific (against thyroid, nerve, testis, and gastric mucosa) and non-specific, such as rheumatoid factors, anti-DNA, cryoglobulins, and cardiolipin.

Bacterial load

In lepromatous leprosy, bacilli spread haematogenously to cool, superficial sites including eyes, upper respiratory mucosa, testes, small muscles, and bones of the hands, feet, and face as well as to peripheral nerves and skin. The heavy bacterial load causes structural damage at all these sites. In tuberculoid leprosy, bacilli are not readily found.

Leprosy reactions

Leprosy reactions are events superimposed on the Ridley–Jopling spectrum. Type 1 (reversal reactions) occur in borderline patients (BT, BB, BL) and are delayed hypersensitivity reactions caused by increased recognition of *M. leprae* antigens in skin and nerve sites. They are characterized by an increase in lymphocytes (CD4 and IL-2-producing cells) within lesions, severe oedema with disruption of the granuloma, and giant cell formation. There is local production of cytokines such as interferon-g and tumour necrosis factor-a. Type 1 reactions are probably associated with a switch from production of T_{H1}- to T_{H2}-type cytokines.

Type 2 reactions, erythema nodosum leprosum (**ENL**), are partly due to immune complex deposition and occur in BL and LL patients who produce antibodies and have a large antigen load. There is vasculitis with lesional immunoglobulin, complement, and polymorphs and circulating immune complexes. There is also enhanced T-cell activity with increased CD8 cells, increased circulating IL-2 receptors, and high levels of circulating tumour necrosis factor-a. After reaction, lepromatous patients revert to a state of immunological unresponsiveness.

Nerve damage

Nerve damage occurs in skin lesions and in peripheral nerve trunks. Myelinated and unmyelinated sensory fibres are affected early. In tuberculoid disease, epithelioid granulomas and perineural inflammation occur. In established lepromatous infection, almost all the cutaneous nerves and peripheral nerve trunks are involved. Bacilli are found in Schwann, perineurial, and endothelial cells. Extensive demyelination occurs and later wallerian degeneration. Despite large numbers of organisms in the nerve there is only a small inflammatory response, but ultimately the nerve becomes fibrotic and is hyalinized. The formation of small granulomas is characteristic of borderline leprosy; granulomatous regions may abut strands of normal looking but heavily bacillated Schwann cells. The combination of lepromatous bacillation and cell-mediated immunity produces widespread nerve damage in borderline leprosy.

Clinical features of leprosy

Patients commonly present with skin lesions, weakness or numbness due to a peripheral nerve lesion, or a burn or ulcer in an anaesthetic hand or foot. Borderline patients may present in reaction with nerve pain, sudden palsy, multiple new skin lesions, pain in the eye, or a systemic febrile illness.

The cardinal signs are:

1. typical skin lesions, anaesthetic at the tuberculoid end of the spectrum;
2. thickened peripheral nerves; and
3. acid-fast bacilli on skin smears or biopsy.

Presenting symptoms

Early lesions

The commonest early lesion is an area of numbness on the skin or a visible skin lesion. The classic early skin lesion, especially in surveys, is indeterminate leprosy, which is commonly found on the face, extensor surface of the limbs, buttocks, or trunk. Indeterminate lesions consist of one or more slightly hypopigmented or erythematous macules, a few centimetres in diameter, with poorly defined margins. Hair growth and nerve function are unimpaired. A biopsy may show the perineurovascular infiltrate and only scanty acid-fast bacilli. The indeterminate phase may last for months or years before resolving or developing into one of the determinate types of leprosy.

Skin

The commonest skin lesions are macules or plaques; papules and nodules are more rare. In lepromatous leprosy a diffuse infiltration of the skin often occurs. Lesions may be found anywhere although rarely in the axillas, perineum, or hairy scalp. Tuberculoid patients have few, hypopigmented lesions while lepromatous patients have numerous, sometimes confluent lesions. The few tuberculoid lesions are usually asymmetrical, more numerous lesions are likely to be distributed symmetrically.

Anaesthesia

Anaesthesia may occur in skin lesions when dermal nerves are involved or in the distribution of a large peripheral nerve. In skin lesions the small dermal sensory and autonomic nerve fibres supplying dermal and subcutaneous structures are damaged causing local sensory loss and loss of sweating within that area.

Peripheral neuropathy

Peripheral nerve trunks are vulnerable at sites where they are superficial or are in fibro-osseous tunnels. At these points a small increase in nerve diameter raises intraneural pressure causing neural compression and ischaemia. Damage to peripheral nerve trunks produces characteristic signs with dermatomal sensory loss and dysfunction of muscles supplied by that peripheral nerve. The sites of predilection for peripheral nerve involvement are ulnar (at the elbow), median (at the wrist), radial, radial cutaneous (at the wrist), common peroneal (at the knee), posterior tibial and sural nerves at the ankle, facial nerve as it crosses the zygomatic arch, and great auricular in the posterior triangle of the neck. All these nerves should be examined for enlargement and tenderness.

Tuberculoid leprosy (TT)

Infection is localized and asymmetrical. A typical tuberculoid skin lesion is a macule or plaque, single, erythematous or purple with raised and clear-cut edges sloping towards a flattened hypopigmented centre. The surface is anaesthetized, dry, and hairless. Sensory impairment may be difficult to demonstrate on the face, where there are abundant nerve endings. If one palpates the edge of the lesions, a thickened cutaneous nerve may be found. If peripheral nerve trunk involvement is present, only one nerve trunk is enlarged. No *M. leprae* are found in skin smears. True tuberculoid leprosy has a good prognosis, many infections resolve without treatment and peripheral nerve trunk damage is limited.

Borderline tuberculoid (BT) (Plate 1)

The skin lesions are similar to TT leprosy and there may be few or many lesions (Fig. 2). The margins are less well defined and there may be satellite lesions. Damage to peripheral nerves is widespread and severe, usually with several thickened nerve trunks. It is important to recognize BT leprosy because these patients are at risk of reversal reactions leading to rapid deterioration in nerve function with consequent deformities.

Fig. 2 Active tuberculoid annular lesions showing the sharp outer edge, thin, raised, erythematous, dry rim, and the broad, hypopigmented, dry centre with slight hair loss. The 'satellite' lesion at the lower outer edge indicates that this is borderline tuberculoid leprosy. As shown, biopsies and smears should be taken from the raised, active rim.

Borderline leprosy (BB)

BB disease is the most unstable part of the spectrum and patients usually downgrade towards lepromatous leprosy if they are not treated or upgrade towards tuberculoid leprosy as part of a reversal reaction. There are numerous skin lesions which may be macules, papules, or plaques and vary in size, shape, and distribution. The edges of the lesions may have streaming, irregular borders. Annular lesions with a broad, irregular edge and a sharply defined punched-out centre are characteristic of BB disease (Fig. 3). Nerve damage is variable.

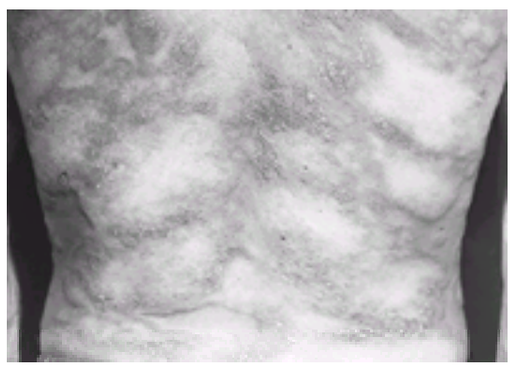

Fig. 3 Borderline annular lesions on the shoulder and back: the rim is broad, the edge irregular, and the 'punched-out' centre is hypopigmented and anaesthetized.

Borderline lepromatous leprosy (BL)

This is characterized by widespread, variable, asymmetrical skin lesions. There may be erythematous or hyperpigmented papules, succulent nodules or plaques, and sensation in the lesions may be normal. Peripheral nerve involvement is widespread. While patients with BL leprosy do not suffer the extreme consequences of bacillary multiplication that are seen in LL disease, they may experience either or both reversal and ENL reactions.

Lepromatous leprosy (LL) (Plate 2)

The patient with untreated polar lepromatous leprosy may be carrying 10^{11} leprosy bacilli. The onset of disease is frequently insidious, the earliest lesions being ill defined, shiny, hypopigmented or erythematous macules. Gradually the skin becomes infiltrated and thickened and nodules develop (Fig. 4); facial skin thickening causes the characteristic leonine facies (Fig. 5). Hair is lost, especially the lateral third of the eye brows (madarosis). Dermal nerves are destroyed leading to a progressive glove and stocking anaesthesia and sensory loss (light touch, pain, and temperature) which begins at the hands and feet and gradually extends to the whole body except for the axillas, groins, and scalp. Position sense is preserved. Sweating is lost, which is uncomfortable in the tropics as compensatory sweating occurs in the remaining intact areas. Damage to peripheral nerves is symmetrical and occurs late in disease. Infiltration of the corneal nerves causes anaesthesia, which predisposes to injury, secondary infection, and blindness.

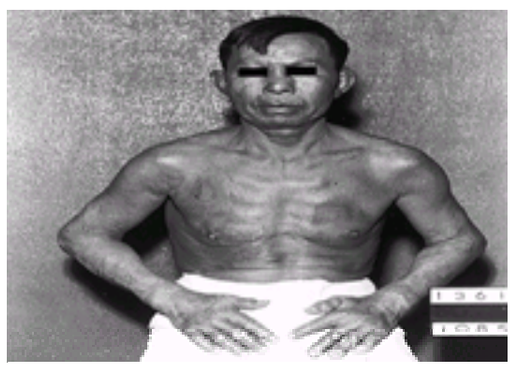

Fig. 4 Active, untreated lepromatous leprosy, showing generalized infiltration of the skin, swelling of fingers and lips, and thinning of eyebrows and eyelashes. The residual annular lesions visible in both pectoral regions indicate that this patient has 'downgraded' from borderline.

Fig. 5 Leonine facies in advanced untreated lepromatous leprosy, with gross thickening of the ear lobes. The skin of the trunk and limbs is infiltrated and mildly erythematous, and small papules are present on some knuckles.

Nasal symptoms can often be elicited early in the disease, and 80 per cent of patients with newly diagnosed lepromatous leprosy have hyperaemic or ulcerated nasal mucosa. Septal perforation may occur. There may be papules on the lips and nodules on the palate, uvula, tongue, and gums. Bone involvement is common, with absorption of the terminal phalanges and pencilling of the heads and shafts of the metatarsals. Testicular atrophy results from diffuse infiltration and the acute orchitis that occurs during ENL reactions. The consequent loss of testosterone leads to azoospermia and gynaecomastia. The extremities become oedematous. The skin of the legs becomes icthyotic and ulcerates easily.

Other forms of leprosy

There are several variant forms of leprosy. Pure neural leprosy occurs principally in India, where it is the presenting form for 10 per cent of patients. There is asymmetrical involvement of peripheral nerve trunks and no visible skin lesions. On nerve biopsy all types of leprosy have been found.

Histoid lesions are distinctive nodules occurring in lepromatous cases which have relapsed due to dapsone resistance or non-compliance with chemotherapy.

Lucio's leprosy is a form of lepromatous leprosy found only in Latin Americans, with a uniform, diffuse, shiny skin infiltration.

Eye disease in leprosy

Blindness due to leprosy, which occurs in at least 2.5 per cent of patients, is a devastating complication for a patient with anaesthesia of the hands and feet. Eye damage results from both nerve damage and bacillary invasion. Lagophthalmos results from paresis of the orbicularis oculi due to involvement of the zygomatic and temporal branches of the facial (VIIth) nerve. These superficial branches are frequently involved in borderline tuberculoid cases, particularly if there are facial skin lesions. In lepromatous disease, lagophthalmos occurs later and is usually bilateral. Damage to the ophthalmic branch of the trigeminal (Vth) nerve causes anaesthesia of the cornea and conjunctiva resulting in drying of the cornea and makes the cornea susceptible to trauma and ulceration. Lepromatous infiltration in corneal nerves produces punctate keratitis and corneal lepromas. Invasion of the iris and ciliary body makes them extremely susceptible to reactions.

Leprosy reactions

Type 1 (reversal reactions) (Plate 3, Plate 4)

These are characterized by acute neuritis and/or acutely inflamed skin lesions (Fig. 6). Nerves become tender with new loss of sensation or motor weakness. Existing skin lesions become erythematous or oedematous; new lesions may appear (Fig. 7). Occasionally oedema of the hands, face, or feet is the presenting symptom, but constitutional symptoms are unusual. Type 1 reactions occur in borderline patients—35 per cent of BL patients will experience a type 1 reaction. The commonest time for reactions is in the first 2 months after starting treatment and in the puerperium.

Fig. 6 Type 1 (reversal) reaction: this BL patient developed new, sharp-edged, well-defined, erythematous plaques with desquamating surfaces about 6 months after starting chemotherapy.

Fig. 7 Reversal-reaction plaque on the left cheek and ear: the edge of this BT lesion has become very sharply defined, more raised, and erythematous, dry, and scaly. Treatment with corticosteroids is imperative, as the patient is at grave risk of rapidly developing lagophthalmos due to associated involvement of branches of the facial nerve.

Type 2 (ENL reactions)

These occur in LL and BL patients. Before multidrug therapy some 50 per cent of LL patients experienced erythema nodosum leprosum (ENL) reactions, the clofazimine component of multidrug therapy has reduced this to 15 per cent. Attacks are acute and may recur over several years. ENL manifests most commonly as painful red nodules on the face and extensor surfaces of limbs (Fig. 8). The lesions may be superficial or deep, with suppuration or brawny induration when chronic. Acute lesions crop and desquamate, fading over several days. ENL is a systemic disorder producing fever and malaise and may be accompanied by uveitis, dactylitis, arthritis, neuritis, lymphadenitis, and orchitis.

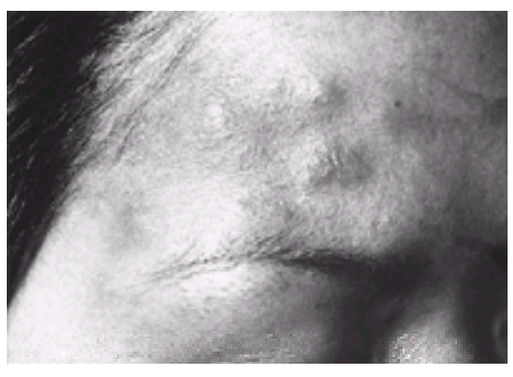

Fig. 8 Erythema nodosum leprosum (ENL) on the forehead of a patient with early lepromatous leprosy. The papules (and nodules) are firm and tender, with rather indefinite edges. In dark-skinned patients the ENL lesions are often easier to feel than to see, especially over the extensor surfaces of the arms and thighs.

Neuritis (Plate 5, Plate 6)

Silent neuropathy is an important form of nerve damage and presents as a functional neural deficit without a manifest acute or subacute neuritis. An Indian study following a cohort of 2608 patients found that 75 per cent of those developing deformity had no history of reactions. In Ethiopian and Bangladeshi cohort studies,

silent neuritis accounted for most neuritis. This emphasizes the importance of regular nerve function testing so that new deficits can be detected.

Diagnosis

The diagnosis is made on the clinical findings of one or more of the cardinal signs of leprosy and supported by the finding of acid-fast bacilli on slit skin smears. The whole body should be inspected in a good light otherwise lesions may be missed, particularly on the buttocks. Skin lesions should be tested for anaesthesia to light touch, pin prick, and temperature. The peripheral nerves should be palpated systematically examining for thickening and tenderness. Wherever possible the diagnosis should be supported by a skin biopsy, which is essential for accurate classification. Serology is not usually helpful diagnostically because antibodies to the species-specific glycolipid PGL-1 are present in 90 per cent of untreated lepromatous patients but only 40 to 50 per cent of paucibacillary patients and 5 to 10 per cent of healthy controls. Polymerase chain reaction for detecting *M.leprae* DNA has not proved sensitive or specific enough for diagnosis.

Outside leprosy endemic areas doctors frequently fail to consider the diagnosis of leprosy. Of new patients seen from 1995 to 1999 at The Hospital for Tropical Diseases, London, diagnosis had been delayed in over 80 per cent of cases. Patients had been misdiagnosed by dermatologists, neurologists, orthopaedic surgeons, and rheumatologists. A common problem was failure to consider leprosy as a cause of peripheral neuropathy in patients from leprosy endemic countries. These delays had serious consequences for patients; over half of them had nerve damage and disability.

Slit skin smears

The bacterial load is assessed by making a small incision through the epidermis, scraping dermal material, and smearing evenly on to a glass slide. At least six sites should be sampled (earlobes, eyebrows, edges of active lesions). The smears are then stained and acid-fast bacilli are counted. Scoring is done on a logarithmic scale per high-power field. A score of 1+ indicates 1 to 10 bacilli in 100 fields, 6+ over 1000 per field. Smears are useful for confirming the diagnosis and should be done annually to monitor response to treatment.

Differential diagnosis

Doctors should be aware of the normal range of skin colour and texture in their local population, and also of the common endemic skin diseases, such as onchocerciasis, that may coexist or mimic leprosy.

Skin

The variety of leprosy skin lesions means that a potentially wide range of skin conditions are in the differential diagnosis. At the tuberculoid end of the spectrum, anaesthesia differentiates leprosy from fungal infections, vitiligo, and eczema. At the lepromatous end the presence of acid-fast bacilli in smears differentiates leprosy nodules from onchocerciasis, Kaposi's sarcoma, and post-kala-azar dermal leishmaniasis.

Nerves

Peripheral nerve thickening is rarely seen except in leprosy. Hereditary sensory motor neuropathy type III is associated with palpable peripheral nerve hypertrophy. Amyloidosis, which can also complicate leprosy, causes thickening of peripheral nerves. Charcot–Marie–Tooth disease is an inherited neuropathy that causes distal atrophy and weakness. The causes of other polyneuropathies such as HIV, diabetes, alcoholism, vasculitides, and heavy metal poisoning should all be considered where appropriate.

Treatment

There are five main principles of treatment:

1. stop the infection with chemotherapy;
2. treat reactions;
3. educate the patient about leprosy;
4. prevent disability; and
5. support the patient socially and psychologically.

These objectives need the patient's co-operation and confidence. In endemic countries, this will usually be achieved through the leprosy outpatient clinic. In countries where leprosy is uncommon, or when the clinical or social situation is complicated, it is often best to admit the patient to an experienced unit. This permits careful assessment together with accurate evaluation of nerve and eye involvement, patient education, and initiation of treatment.

Chemotherapy

All patients with leprosy should be given an appropriate multidrug combination. In the hospital setting, where skin smears and skin biopsies can be combined with clinical data, patients can be classified into paucibacillary (skin smear-negative tuberculoid and BT) and multibacillary (skin smear-positive BT, all BB, BL, and LL). The first-line antileprosy drugs are rifampicin, clofazimine, and dapsone. Table 2 gives the drug combinations, doses, and duration of treatment. Patients with multibacillary disease and an initial bacterial index greater than 4 will need longer treatment and the duration should be guided by their clinical status and bacterial index.

Rifampicin is a potent bactericide for *M. leprae*. Four days after a single 600 mg dose, bacilli from a previously untreated patient with multibacillary disease were no longer viable in a mouse foot-pad test. It acts by inhibiting DNA-dependent RNA polymerase. Because *M. leprae* can develop resistance to rifampicin as a one-step process, this drug should always be given in combination with other antileprotics.

Dapsone (**DDS**, 4,4-diaminodiphenylsulphone) is weakly bactericidal. Oral absorption is good and it has a long half-life, averaging 28 h. It commonly causes mild haemolysis, but rarely anaemia. Glucose-6-phosphate dehydrogenase deficiency is seldom a problem. The 'DDS syndrome', which is occasionally seen in leprosy, begins 6 weeks after starting dapsone and manifests as exfoliative dermatitis associated with lymphadenopathy, hepatosplenomegaly, fever, and hepatitis.

Clofazimine is a red, fat-soluble, crystalline dye. The mechanism of its weakly bactericidal action against *M. leprae* remains unknown. The most troublesome side-effect is skin discoloration, ranging from red to purple-black, the degree depending on the drug dose and extent of leprous infiltration. The pigmentation usually fades within 6 to 12 months of stopping clofazimine, although traces of discoloration may remain for up to 4 years. Urine, sputum, and sweat may become pink. Clofazimine also produces a characteristic icthyosis on the shins and forearms.

New drugs bactericidal for *M. leprae* have been identified, notably the fluoroquinolones pefloxacin and ofloxacin, minocycline, and clarithromycin. These agents are now established second-line drugs and may replace dapsone and clofazimine. Minocycline causes a black pigmentation of skin lesions and so may not be an appropriate substitute for clofazimine if pigmentation is to be avoided.

Since the introduction of multidrug therapy more than 10 million patients have been treated successfully. Clinical improvement has been rapid and toxicity rare. The treatment duration has been shortened. Monthly supervision of the rifampicin component has been crucial to success. Other benefits are reduced deformity rates, increased compliance in control schemes, a halving of the annual case load, and reduction of the long-term (though not short-term) cost of control schemes. At the end of a 6-month treatment of borderline disease there may still be signs of inflammation, which should not be mistaken for active infection. The distinction between relapse and reaction may be difficult. World Health Organization (**WHO**) studies have reported a cumulative relapse rate of 1.07 per cent for paucibacillary leprosy and 0.77 per cent for multibacillary leprosy at 9 years after completion of multidrug therapy. *M. leprae* is such a slow-growing organism that relapse only occurs after many years. *M. leprae* isolates from relapsed patients who have received multidrug therapy are fully drug sensitive and patients can be retreated with the same regimen.

A single-dose triple-drug combination (rifampicin, ofloxacin, and minocycline) has been tested in India for patients with single skin lesions and improved 98 per cent of

patients. Although the study had major flaws and single-dose treatment is less effective than the conventional 6-month treatment for paucibacillary leprosy, it is an operationally attractive field regimen and has been recommended for use by the WHO.

Reactions may develop months or years after stopping chemotherapy, especially in BL or LL patients. It is therefore vital when discharging patients to warn them to return should new symptoms appear, especially in hands, feet, or eyes. Patients with reactions or physical or psychological complications will need long-term care.

Management of reactions

Awareness of the early symptoms of reversal reactions by both patient and physician is important because, if left untreated, severe nerve damage may develop. The peak time for reversal reactions is in the first 2 months of treatment. Patients should be warned about reactions because the sudden appearance of reactional lesions after starting treatment is distressing and undermines confidence. The treatment of reactions is aimed at controlling acute inflammation, easing pain, reversing nerve and eye damage, and reassuring the patient. Multidrug therapy should be continued.

Type 1 (reversal) reactions

Simple anti-inflammatory drugs are rarely sufficient to control symptoms. If there is any evidence of neuritis (nerve tenderness, new anaesthesia, and/or motor loss), corticosteroid treatment should be started. Prednisolone should be given, starting at 40 to 60 mg daily, reducing to 40 mg after a few days, and then by 5 mg every 2 to 4 weeks. Patients with BT leprosy in reaction commonly need 2 to 4 months of steroids while BL reactions may need 6 months.

Type 2 (ENL) reactions (Plate 7)

This is a difficult condition to treat and frequently requires treatment with high-dose steroids (80 mg daily, tapered down rapidly) or thalidomide. Since ENL frequently recurs, steroid dependency can easily develop. Thalidomide (400 mg daily) is superior to steroids in controlling ENL and is the drug of choice for young men with severe ENL. Women with severe ENL may benefit from thalidomide treatment. This is a difficult decision for the woman and her physician and needs careful discussion of the benefits and risks (phocomelia when thalidomide is taken in the first trimester). Women should use double contraception and report immediately if menstruation is delayed. Unfortunately, the problems with thalidomide mean that it is unavailable in several leprosy endemic countries despite its undoubted value. Clofazimine has a useful anti-inflammatory effect in ENL and can be used at 300 mg per day for several months. Low-grade chronic erythema nodosum, with iritis or neuritis, will require long-term suppression, preferably with thalidomide or clofazimine. Acute iridocyclitis is treated with 4-hourly instillation of 1 per cent hydrocortisone eye drops and 1 per cent atropine drops twice daily.

Neuritis

Silent neuritis should be treated similarly to reversal reactions—prednisolone in a dose of 40 mg daily and reducing slowly over a period of months.

Education of the patient

Stigmatization due to leprosy occurs worldwide. Patients are frightened of social ostracization, physical rejection, and the development of deformities. It is often useful to ask them about their fears so that these can be addressed. They should be reassured that having started treatment they are not infectious to family or friends. The importance of compliance with antibiotic therapy needs to be emphasized. The patient needs a careful explanation of the diagnosis, aetiology, and prognosis.

Prevention of disability

The morbidity and disability associated with leprosy is secondary to nerve damage. A major goal in prevention of disability is to create patient self-awareness so that damage is minimized. Monitoring sensation and muscle power in patient's hands, feet, and eyes should be part of the routine follow-up so that new nerve damage is detected early. The patient with an anaesthetized hand or foot needs to understand the importance of daily self-care, especially protection when doing potentially dangerous tasks and inspection for trauma. It is helpful to identify for each patient potentially dangerous situations, such as cooking, car repairs, or smoking. Soaking dry hands and feet followed by rubbing with oil keeps the skin moist and supple.

An anaesthetized foot needs the protection of an appropriate shoe. For anaesthesia alone, a well-fitting 'trainer' with firm soles and shock-absorbing inners will provide adequate protection. Once there is deformity, such as clawing, shoes must be made specially to ensure protection of pressure points and even weight distribution.

The patient should be taught to question the cause of an injury so that the risk can be avoided in the future. Plantar ulceration occurs secondary to increased pressure over bony prominences. Ulceration is treated by rest. Unlike ulcers in the feet of patients with diabetes or ischaemia, ulcers in leprosy heal if they are protected from weight-bearing. No weight-bearing is permitted until the ulcer has healed. Appropriate footwear should be provided to prevent recurrence.

Physiotherapy exercises should be taught to maximize function of weak muscles and prevent contracture. Contractures of hands and feet, foot drop, lagophthalmos, entropion, and ectropion are amenable to surgery.

Social, psychological, and economic rehabilitation

The social and cultural background of the patient determine the nature of many of the problems that may be encountered. The patient may have difficulty in coming to terms with leprosy. The community may reject the patient. Education, gainful employment, confidence from family, friends, and doctor, and plastic surgery to correct stigmatizing deformity all have a role to play.

Prognosis

The majority of patients—especially those who have no nerve damage at the time of diagnosis—do well on multidrug treatment with resolution of skin lesions. Left untreated, borderline patients will downgrade towards the lepromatous end of the spectrum and lepromatous patients will suffer the consequences of bacillary invasion. Borderline patients are at risk of developing type 1 reactions, which may result in devastating nerve damage. Treatment of the neuritis is currently unsatisfactory and patients with neuritis may develop permanent nerve damage despite corticosteroid treatment. It is not possible to predict which patients will develop reactions or nerve damage. Nerve damage and its complications may be severely disabling, especially when all four limbs and both eyes are affected.

Leprosy in women

Women with leprosy are in double jeopardy, not only may they develop postpartum nerve damage but they are at particular risk of social ostracization with rejection by spouses and family.

Pregnancy and leprosy

There is little good evidence that pregnancy causes new disease or relapse. However, there is a clear temporal association between parturition and the development of type 1 reactions and neuritis when cell-mediated immunity returns to prepregnancy levels. In an Ethiopian study, 42 per cent of pregnancies in BL patients were complicated by a type 1 reaction in the postpartum period. In the same cohort, LL patients experienced ENL reactions throughout pregnancy and lactation. ENL in pregnancy is associated with early loss of nerve function compared with non-pregnant individuals. Pregnant and newly delivered women should have regular neurological examination and steroid treatment instituted for neuritis. Rifampicin, dapsone, and clofazimine are safe during pregnancy. Clofazimine crosses the placenta and babies may be born with mild clofazimine pigmentation. Lepra reactions can be managed with the steroid regimens given above but with a more rapid reduction in dose. Women should be warned before becoming pregnant of the risk that their condition will deteriorate after delivery. Ideally pregnancies should be planned when leprosy is well controlled.

Prevention and control

The current strategy of leprosy control in endemic countries through vertical programmes providing case detection, treatment with WHO multidrug therapy, and contact examination and supported by case-finding campaigns, especially in schools, has been very successful. Effective treatment is not merely restricted to chemotherapy but also involves good case management with effective monitoring and supervision. An important secondary role of leprosy control programmes is the prevention of disabilities.

Vaccines against leprosy

The substantial cross-reactivity between bacille Calmette–Guérin (**BCG**) and *M. leprae* has been exploited in attempts to develop a vaccine against leprosy. Trials of BCG as a vaccine against leprosy in Uganda, New Guinea, Burma, and South India showed it to confer statistically significant but variable protection, ranging from 80 per cent in Uganda to 20 per cent in Burma. A case–control study in Venezuela showed BCG vaccination to give 56 per cent protection to the household contacts of patients with leprosy. Combining BCG and killed *M. leprae* has been tried, but in both a large population-based trial in Malawi and an immunoprophylactic trial in Venezuela there was no advantage for BCG plus *M. leprae* over BCG alone.

Areas of uncertainty/controversy

The optimum duration of treatment is a controversial area. The last WHO expert committee recommended that treatment for multibacillary cases could be reduced from 24 to 12 months. The classification of leprosy cases has been simplified. Cases are now classified by the number of lesions; slit skin smears are not mandatory. The decision to stop recommending slit skin smears was made because of the poor standard of smears in the field. A WHO sponsored trial comparing 12 months with 24 months treatment is in progress, but intake only started in 1992 and a 10-year follow-up is needed to assess relapse rates. Data from India show that patients with a high initial bacterial load (bacterial index greater than 4) treated with 2 years of rifampicin, clofazimine, and dapsone had a relapse rate of 8/100 person years, whereas patients treated to smear negativity had a relapse rate of 2/100 person years. The dilemma is that if skin smears are abandoned then those patients in need of longer treatment courses cannot be identified. These arguments illustrate the difficulty in providing sound evidence for policy decisions when a decade-long wait is needed.

The shortening of drug treatment for leprosy means that the vertical leprosy programmes are now treating far fewer patients. There is considerable discussion about how best to detect, treat, and prevent leprosy disability in the future. There are several possibilities—integration with tuberculosis programmes, dermatology programmes, or at health centre level. Whichever model is chosen as being locally appropriate, it should be remembered that treating patients with leprosy is a long-term enterprise involving patients, their families, and health workers.

Areas where further research is needed

The epidemiology of leprosy still poses unanswered questions. Why are 70 per cent of all patients with leprosy in India? Is this due to living conditions, genetic susceptibilty, or particular environmental conditions in India?

Early detection of cases is vital both at an individual and population level. It is now recognized that substantial nerve damage occurs before diagnosis. A test for early infection might help detect individual cases before nerve damage is established and before the spread of infection. Leprosy-specific peptides for skin tests have been generated and are being evaluated.

The medical management of reactions and nerve damage is currently limited to steroids. These are not effective for about 30 per cent cases. Thus trials of new immunosuppressants are urgently needed.

The WHO started the 1990s with the bold slogan of 'Eliminating leprosy as a public health problem by 2000'. This initiative galvanized leprosy control programmes worldwide, but the unique biology of *M. leprae* and its interaction with the human host meant that the target was unattainable. As the millennium approached the slogan was quietly dropped to the disappointment of many leprosy workers and governments. Leprosy is a bacterial disease with challenging immunological complications and will be a global and individual problem for many decades. It is unlikely to be eradicated until there is considerable improvement in general health, wealth, living conditions, and education.

Further reading

Britton WJ (1998). The management of leprosy reversal reactions. *Leprosy Review* **69**, 225–34. [A comprehensive review.]

Khanolkar-Young S *et al*. (1995). Tumour necrosis factor-a synthesis is associated with the skin and peripheral nerve pathology of leprosy reversal reactions. *Clinical and Experimental Immunology* **99**, 196–202. [Key paper demonstrating a tissue-damaging cytokine in leprosy nerves.]

Lockwood DNJ (1996). The management of erythema nodosum leprosum: current and future options. *Leprosy Review* **67**, 253–9.

Lockwood DNJ (1997). Rifampicin, ofloxacin, and minocycline (ROM) for single lesions in leprosy. What is the evidence? *Leprosy Review* **68**, 299–300. [A critical analysis of the trial report for single-dose treatment.]

Ponnighaus JM *et al*. (1992). Efficacy of BCG vaccine against leprosy and tuberculosis in northern Malawi. *Lancet* **339**, 636–9. [Demonstrates that adding killed *M. leprae* to BCG vaccine does not enhance protection against leprosy. This study also shows that BCG protects better against leprosy than tuberculosis in Africa.]

Sampaio EP *et al*. (1995). Cellular immune response to *Mycobacterium leprae* infection in human immunodeficiency virus infected individuals. *Infection and Immunity* **63**, 18848–54. [Key paper showing normal granuloma formation in patients coinfected with leprosy and HIV.]

Waters M (1998). Is it safe to shorten multidrug therapy for lepromatous (LL and BL) leprosy to 12 months? *Leprosy Review* **69**, 110–11. [Succinct review of the problems and uncertain data for short-course chemotherapy in leprosy.]

WHO expert committee on leprosy (1998). *WHO Technical Report Series.* World Health Organization, Geneva. [Summary of current recommendations for the management of leprosy in the field.]

Yamamura M *et al*. (1992). Cytokine patterns of immunologically mediated tissue damage. *Journal of Immunology* **149**, 1470–5. [Paper showing that lepromatous disease is associated with a T_{H2} pattern of cytokine production and TT with a T_{H1} pattern.]

7.11.25 Buruli ulcer: *Mycobacterium ulcerans* infection

Wayne M. Meyers and Françoise Portaels

Introduction

Buruli ulcer, also known as Bairnsdale or Searles' ulcer (Australia), and Kakerifu or Toro ulcer (Congo), is an indolent, necrotizing infection of the skin, subcutaneous tissue, and bone, caused by *Mycobacterium ulcerans*. After tuberculosis and leprosy, Buruli ulcer is the third most common mycobacterial disease, and is recognized by the WHO as a re-emerging infection.

Aetiology

In 1948 MacCallum and colleagues first isolated the causative agent from patients in Australia. *M. ulcerans* is a slow-growing, acid-fast bacillus which grows optimally at 32°C, and elaborates mycolactone, a cytotoxic and immunosuppressive polyketide. Putatively, this toxin is the primary virulence factor of *M. ulcerans*. Data from 16S rRNA sequences define four groups of *M. ulcerans*: African, American, Asian, and Australian strains.

Epidemiology and transmission

All endemic foci of Buruli ulcer are near rural freshwater wetlands, especially still or slow-moving water (ponds and swamps). All foci except those in southern Australia and northern Asia are tropical. Major endemic areas are Benin, Ghana, Ivory Coast, Nigeria, Congo, Gabon, Uganda, and adjacent countries. There are minor foci in South and Central America and south-east and northern Asia.

Documented environmental sources of *M. ulcerans* include water in irrigation systems and water bugs that dwell in the roots of aquatic plants in the bottom mud of swamps. In Australia, koala, possum, and naturalized alpaca contract the infection naturally.

Outbreaks of disease often follow environmental changes that promote flooding or alter water courses, such as deforestation or construction of dams and irrigation systems. Increases in farming populations near wetlands may contribute to the rapid re-emergence of Buruli ulcer in Africa. Approximately 75 per cent of all new infections are in children, who often play semi-naked in swampy terrain.

We postulate that humans become infected by traumatic introduction of the bacillus into the dermis or subcutis from the overlying *M. ulcerans*-contaminated skin surface. The trauma may be as slight as a hypodermic injection or as severe as a land-mine wound or snakebite. Biting insects (e.g. water bugs) may serve as mechanical vectors. Aerosols arising from the surface of ponds and swamps may disseminate *M. ulcerans*. Patient-to-patient transmission is rare.

Pathogenesis

No predisposing host factors are known. Once introduced, the small amount of mycolactone produced by a few *M. ulcerans* bacilli causes tissue necrosis and suppresses local immune responses ensuring survival of the bacillus in a nidus of nutrient necrotic tissue. The toxin targets subcutaneous fat cells so that necrosis can spread in and just superficial to fascial planes. *M. ulcerans* invades lymphatics and probably blood vessels, causing metastatic spread.

Clinical features

Clinical effects may be localized or disseminated. Except for those with massive lesions, patients are usually surprisingly well without specific systemic symptoms or abnormal laboratory findings.

Localized disease

Typically the initial cutaneous lesion is a single, firm, painless, non-tender, movable subcutaneous nodule up to 3 cm in diameter. Limbs are most frequently affected, often around joints. There is marked variation in the natural history of the disease, but nodules usually ulcerate within 1 to 3 months of inoculation. A whitish necrotic slough develops in the base of the ulcer and the surrounding skin is indurated and hyperpigmented. Ulcer borders are undermined, sometimes extending 15 to 20 cm or more (major ulcerative disease) (Fig. 1 and Plate 1). Some small (1 to 2 cm in diameter) ulcerated lesions with shallow undermining self-heal early without sequelae (minor ulcerative disease). Without treatment, major ulcerative lesions tend to become inactive, usually after months or years, and heal by scarring. Typically the scars are depressed and stellate, often causing disfiguring and crippling cicatricial contractions.

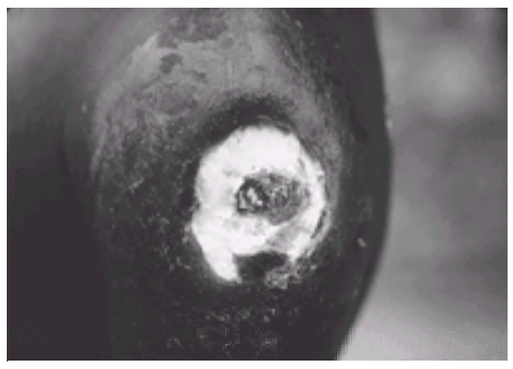

Fig. 1 Buruli ulcer on the left deltoid area in a 12-year-old Congolese boy who had received a hypodermic injection at this site 3 months previously. Note central necrotic slough in the base of the ulcer, and undermined edges. (See also Plate 1.)

Disseminated disease

Disseminated disease may pass only through the nodular stage or arise from localized major ulcerative lesions; however, following inoculation, the disease

sometimes disseminates directly and rapidly. These patients present with indurated plaques of varying size, sometimes covering an entire limb or vast areas of the trunk. Without treatment, such lesions will eventually slough, leaving a large ulcer with continuing extension of disease at the borders. Structures such as eyes, breasts, and genitalia may be damaged or lost.

While metastatic spread may arise from localized disease, patients with the highly bacilliferous disseminated cutaneous form are more prone to develop metastatic lesions. Spread may be to distant skin sites or to bone. Bones of the limbs are affected most frequently. *M. ulcerans* osteomyelitis is an increasing problem in many endemic areas, and often leads to amputations and other disabilities.

Differential clinical diagnosis

Diagnosis of the nodular stage is often perplexing. Differential diagnoses include bacterial, mycotic, and parasitic infections, inflammatory lesions, and tumours. Ulcers resembling Buruli ulcer include tropical phagedenic ulcer (malodorous and not undermined), venous stasis ulcer (not undermined), and bites by venomous snakes or spiders (history helpful).

Pathology

Optimal biopsy specimens contain the necrotic base of ulcers and undermined edge of lesions and subcutaneous tissue and fascia. Histopathological sections reveal a contiguous coagulation necrosis (non-caseating) of the deep dermis, panniculus, and fascia. Vasculitis and mineralization in these areas are common. Clumps of extracellular acid-fast bacilli are most plentiful in the base of the ulcer. Necrosis extends well beyond the location of the bacilli. Local and regional lymph nodes are often invaded and sometimes necrotic. In bone, the marrow is necrotic and contains acid-fast bacilli, and trabeculas are eroded. Development of delayed-type hypersensitivity granulomas heralds healing and eventual fibrosis.

Laboratory diagnosis

Smears stained by the Ziehl–Neelsen method from the ulcer base often reveal acid-fast bacilli in clumps. Cultures for *M. ulcerans* are often positive. Polymerase chain reaction is available for specific identification of *M. ulcerans*. Histopathological changes are characteristic.

Treatment

Wide surgical excision and skin grafting is the recommended treatment. Antimicrobial agents (e.g. rifampin and clarithromycin) should be administered before and after surgery to limit bacterial dissemination. Heating the lesion to 40°C is a useful adjunct. Oral antimycobacterial therapy without surgery may heal nodules and minor ulcerative lesions, but controlled trials are needed to establish efficacy. Physiotherapy is essential to prevent contraction deformities.

Prevention and control

Bacille Calmette-Guérin (BCG) vaccination provides short-lived protection. There are no practicable effective control measures for inhabitants of endemic areas. Tourists can avoid the wetlands in endemic countries.

Socio-economic impact

Patients are often disabled for life and require welfare services, often locally limited or unavailable. They require hospital stays of many months, taxing overburdened services.

Further reading

Asiedu K, Etuaful S (1998). Socioeconomic implications of Buruli ulcer in Ghana: a three-year review. *American Journal of Tropical Medicine and Hygiene* **59**, 1015–22. [Stresses the burden of Buruli ulcer as a chronic disease on the health-care delivery system of developing countries.]

George KM *et al*. (1999). Mycolactone: a polyketide toxin from *Mycobacterium ulcerans* required for virulence. *Science* **283**, 854–7. [Describes purification and characterization of the toxin of *M. ulcerans*, potentially opening new approaches to the treatment and prevention of Buruli ulcer.]

Meyers WM (1995). Mycobacterial infections of the skin. In: Doerr W, Seifert G, eds. *Tropical pathology*, pp 291–377. Springer-Verlag, Berlin. [Extensive coverage of the clinical and pathological features of Buruli ulcer.]

Van der Werf TS *et al*. (1999). *Mycobacterium ulcerans* infection. *Lancet* **354**, 1013–18. [A review of the current status of the epidemiology, diagnosis, and treatment of *M. ulcerans* infection in the world.]

7.11.26

Actinomycoses

K. P. Schaal

Definition

Actinomycoses are subacute to chronic, granulomatous as well as suppurative inflammatory diseases that tend to progress slowly and usually give rise to multiple abscesses and draining sinus tracts. Various fermentative (facultatively anaerobic or capnophilic) actinomycetes of the genera *Actinomyces* and *Propionibacterium*, but rarely also *Bifidobacterium*, may act as the principal causative agents of the disease. Because the term 'actinomycoses' denotes a polyaetiological inflammatory syndrome rather than a condition attributable to a single pathogenic actinomycete species, it should only be used in the plural.

Aetiology of human actinomycoses

Actinomyces israelii and *A. gerencseriae* are by far the most frequent and most characteristic pathogens aetiologically involved in the human form of the disease. *Propionibacterium propionicum*, *Actinomyces naeslundii*, *A. odontolyticus*, *A. viscosus*, *A. meyeri*, and *Bifidobacterium dentium* (formerly '*Actinomyces eriksonii*') are further potential but less common causes of actinomycotic infections, while *Actinomyces bovis* has been recovered solely from animals (Table 1).

Pathogenesis and pathology

Most of the fermentative actinomycetes pathogenic to man are found regularly and abundantly in the mouths of healthy adults. However, these microbes occur only sporadically or in low numbers in the digestive, respiratory, and genital tracts, as well as in the mouths of babies before teething and of adults without any natural teeth or tooth implants. Therefore, these actinomycetes may be considered facultatively pathogenic commensals of the human mucous membranes, which, apart from the very rare actinomycotic wound infections after human bites or fist fights, produce disease exclusively as endogenous pathogens.

For active invasion of the tissue, the classical pathogenic fermentative actinomycetes apparently require a negative redox potential, which may result either from insufficient blood supply (caused by circulatory or vascular diseases, crush injuries, or foreign bodies) or from the reducing and necrotizing capacity of other microbes in the lesion. Defective functions of the immune system do not specifically predispose to actinomycotic infections.

Synergistic polymicrobial infection

True actinomycoses are essentially always synergistic mixed infections, in which the actinomycetes act as the specific component, the so-called 'guiding organisms', which decide on the characteristic course and the late symptoms of the disease. The so-called concomitant microbes (Table 2), which may vary considerably in composition (about 100 aerobic and anaerobic species) and number (up to 10 per case) of species from case to case, are often responsible for the clinical picture at the beginning of the infection and for certain complications; they are also part of the resident or transient surface microflora of the mucous membranes of man.

Particularly pronounced synergistic interactions appear to exist between pathogenic fermentative actinomycetes, especially *Actinomyces israelii* and *A. gerencseriae*, and *Actinobacillus actinomycetemcomitans*. The latter organism, the name of which refers to its characteristic association with actinomycetes, may even sustain the inflammatory process under similar clinical symptoms after chemotherapeutic elimination of the causative actinomycete.

Histopathology

Initially, an inflammatory granulation tissue develops, which usually breaks down to form either an acute abscess or chronic multiple abscesses with proliferation of connective tissue. The pathognomonic sulphur granules are formed primarily in the infected tissue, but may also appear as free structures in abscess content or sinus discharge. They are then of the highest diagnostic importance.

Sulphur granules, which were originally designated 'Drusen' in Harz's first description of *Actinomyces bovis* in 1877, are macroscopically visible (up to 1 mm in diameter), yellowish, reddish to brownish particles, which exhibit a cauliflower-like appearance under the microscope at low magnifications (Fig. 1). They consist of a conglomerate of filamentous actinomycete microcolonies formed *in vivo* and surrounded by tissue reaction material, especially polymorphonuclear granulocytes. At high magnification, a Gram-stained smear of the completely crushed granule reveals the presence of clusters of Gram-positive interwoven branching filaments with radially arranged peripheral hyphae and of a variety of other Gram-positive and Gram-negative rods and cocci, which represent the concomitant flora. A club-shaped layer of hyaline material may be seen on the tips of peripheral filaments, which can aid in the differentiation of actinomycotic sulphur granules from macroscopically similar particles of various other microbial and non-microbial origins.

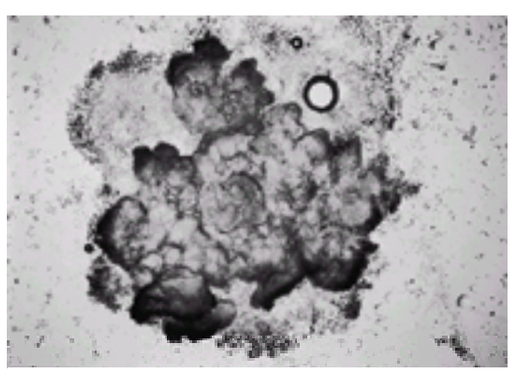

Fig. 1 Actinomycotic sulphur granule. Micrograph of a particle embedded in 1 per cent methylene-blue solution, original diameter 0.8 mm. Note the cauliflower-like structure in the centre of the particle and the partially dark-stained granulocytes in the periphery.

Clinical manifestations

The primary actinomycotic lesion usually develops in tissue adjacent to a mucous membrane at sites such as the cervicofacial, thoracic, and abdominal areas. The infection tends to progress slowly and to penetrate without regard to natural organ borders, or to spread haematogenously even to distant sites. Remission and exacerbation of symptoms with and without antimicrobial treatment is characteristic. As in other endogenous microbial diseases, the incubation period of actinomycoses is not defined.

Cervicofacial actinomycoses

In the vast majority of cases, actinomycotic lesions primarily involve the face or neck. Conditions predisposing to these cervicofacial infections include tooth extractions, fractures of the jaw, periodontal abscesses, foreign bodies penetrating the mucosal barrier (bone splinters, fish bones, awns of cereals), or suppurating tonsillar crypts.

Initially, the cervicofacial actinomycoses present either as an acute, usually odontogenic, abscess or cellulitis of the floor of the mouth, or as a slowly developing, chronic, hard, painless, reddish or livid swelling. Small acute actinomycotic abscesses may heal after surgical drainage alone. More often, however, the acute initial stage is followed by a subacute to chronic course if no specific antimicrobial treatment is given, thereby imitating the primarily chronic form, which is characterized by regression and cicatrization of central suppurative foci while the infection progresses peripherally producing hard, painless, livid infiltrations. These may lead to multiple, new areas of liquefaction, fistulae (Fig. 2), which often discharge pus containing sulphur granules, and multilocular cavities with poor healing and a tendency to recur after temporary regressions of the inflammatory symptoms.

Fig. 2 Primarily chronic cervicofacial actinomycosis with several draining sinus tracts and livid discoloration of the skin in a 42-year-old man.

With inappropriate or no treatment, cervicofacial actinomycoses extend slowly, even across organ borders, and may become life-threatening by invasion of the cranial cavity, the mediastinum, or the bloodstream.

Thoracic actinomycoses

Thoracic manifestations, which are much less common than the cervicofacial form (Table 3), usually develop after aspiration or inhalation of material from the mouth (dental plaque or calculus, tonsillar crypt contents) or a foreign body that contains or is contaminated with the causative agents. Occasionally, this form of disease may result from extension of an actinomycotic process of the neck, from an abdominal infection perforating the diaphragm, or from a distant focus by haematogenous spread.

Primary pulmonary actinomycoses present as bronchopneumonic infiltrations that may imitate tuberculosis or bronchial carcinoma radiographically, appearing as single dense or multiple spotted shadows in which cavitations may develop (Fig. 3). If not diagnosed and treated properly, pulmonary infections may extend through to the pleural cavity producing empyema, to the pericardium, or to the chest wall; they may even appear as a paravertebral (psoas) abscess tracking down to the groin.

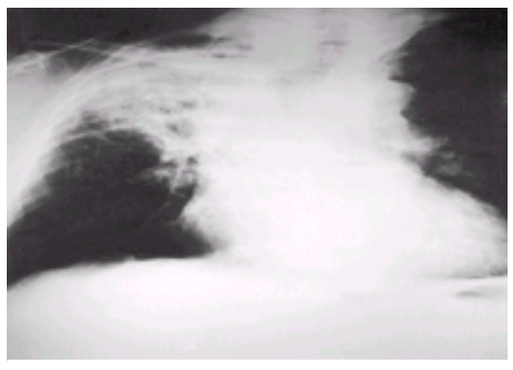

Fig. 3 Chest radiograph of pulmonary actinomycosis of the right upper lobe in a 62-year-old man. The disease was only diagnosed after a huge subcutaneous abscess had developed covering the whole right shoulder blade.

Abdominal actinomycoses

Actinomycoses of the abdomen and pelvis are rare (Table 3). They originate either from acute perforating gastrointestinal diseases (appendicitis, diverticulitis, various ulcerative diseases), from surgical or accidental trauma including injuries caused by ingested bone splinters or fish bones, or from inflammations of the female internal genital organs.

Women who wear intrauterine contraceptive devices or vaginal pessaries for long periods often show a characteristic colonization of the cervical canal and the uterine cavity by various fermentative actinomycetes and other anaerobes resembling the synergistic actinomycotic flora. However, this colonization only rarely results in an invasive actinomycotic process.

Most abdominal actinomycoses present as slowly growing tumours, which, in the absence of sinus tracts discharging pus with sulphur granules, are difficult to differentiate from malignant neoplasms such as colonic, rectal, ovarian, or cervical carcinomas. By direct extension, any abdominal tissue or organ may be involved including muscle, liver, spleen, kidney, fallopian tubes, ovaries, testes, bladder, or rectum. Haematogenous liver abscesses have been seen, especially associated with genital actinomycoses.

Actinomycotic infections of the central nervous system

Actinomycoses of the brain and the spinal cord are very rare. They may arise from direct extension of cervicofacial infections. Haematogenous spread is also

possible, particularly from primary lesions in the lungs or abdomen. The spinal canal may be directly involved from these sites. Brain abscess is much more common than meningitis.

Actinomycoses of the bone

Bone involvement is also very rare. It usually develops by direct extension from soft tissue infection resulting in a periostitis with new bone formation visible by radiography. If the bone itself is invaded, localized areas of bone destruction surrounded by increased bone density usually develop. Mandible, ribs, and spine are most frequently involved.

Cutaneous actinomycoses

Actinomycotic lesions of the skin are extremely rare. Usually, they originate from wounds that were contaminated with saliva or dental plaque following human bites or fist fights, but they may also result from haematogenous spread. Symptoms are similar to those of cervicofacial actinomycoses.

Diagnosis

Clinical symptoms are often misleading, especially in the early stages of the disease, histopathological appearances are unreliable, and diagnosis chiefly rests on bacteriological methods.

Radiography

In cervicofacial cases, radiography is useful only for detecting bone involvement. A pulmonary infiltrate associated with a proliferative lesion or destruction of ribs is highly suggestive of either actinomycosis or a tumour. Radiography may also help to locate the abdominal processes and to identify the involvement of organs such as liver, kidney, urinary bladder, or ureter. In general, however, radiographic changes are not diagnostic.

Laboratory diagnosis

Clinical chemistry and haematology

Small, localized actinomycotic lesions are not usually associated with abnormalities. In advanced cases, however, especially those in the thoracic or the abdominal area, a raised erythrocyte sedimentation rate and pronounced leucocytosis may be found. When the central nervous system is involved, a polymorphonuclear or mononuclear pleocytosis is commonly found. The protein content of the cerebrospinal fluid is frequently elevated and the sugar content moderately depressed.

Bacteriology

Pus specimens containing sulphur granules and occasionally looking like semolina should prompt the clinician to ask and the bacteriologist to look specifically for actinomycetes using suitable cultural techniques and other methods.

Pus, sinus discharge, bronchial secretions, granulation tissue, or biopsy materials are suitable specimens. Precautions must be taken to prevent contamination of the specimen by the indigenous mucosal flora. In cases of cervicofacial actinomycoses, pus should therefore be obtained only by transcutaneous puncture of the abscesses or by transcutaneous needle biopsy. When abscesses have already been incised, a sufficient amount of pus should be collected instead of using only a swab. Because sputum always contains oral actinomycetes, bronchial secretions should be obtained by transtracheal aspiration, or material should be collected by transthoracic percutaneous needle biopsy. Percutaneous puncture of suspected abscesses is often the only way of obtaining suitable specimens for diagnosing abdominal actinomycoses.

The transport of specimens to the bacteriological laboratory should be as fast as possible, preferably by messenger. Alternatively, a reducing transport medium such as one of the modifications of Stuart's medium should be used. The specimen should arrive in the laboratory within 24 h, although it has occasionally proved possible to isolate actinomycetes from samples that took 7 days or more to get to the diagnostic laboratory by post.

A quick and comparatively reliable tentative diagnosis is possible microscopically when sulphur granules are present (Fig. 1). The demonstration of concomitant bacteria in Gram-stained smears prepared from crushed granule material allows the differentiation of actinomycotic granules from similar particles produced by *Nocardia*, *Actinomadura*, or *Streptomyces* species.

Use of transparent culture media and careful microscopic examination of the cultures, preferably on Fortner plates, after at least 2, 7, and 14 days of incubation, enables a specialized laboratory to detect possible actinomycete colonies and to subculture them for identification. Isolation and definite identification to the species level may require a further 1 to 2 weeks. Techniques such as the application of gene probes or the polymerase chain reaction for detecting and identifying fermentative actinomycetes are not yet widely used.

Serological diagnosis

None of the routine serological methods has yet provided satisfactory results because sensitivity and specificity have been found to be too low.

Treatment

As the aetiology of human actinomycoses is always polymicrobial, the antibacterial drugs used for treatment should in principle cover both the causative actinomycetes and all of the concomitant bacteria. This usually requires the administration of drug combinations, in which aminopenicillins currently represent the therapeutic basis because they are slightly more active against the pathogenic actinomycetes than is penicillin G and because they are able to inhibit *Actinobacillus actinomycetemcomitans* which is usually resistant to narrow-spectrum penicillins. However, the presence of concomitant b-lactamase producers such as *Bacteroides fragilis*, *B. thetaiotaomicron*, or *Staphylococcus aureus* (b-lactamase producing) may impair the therapeutic efficacy of aminopenicillins and that of many other b-lactams so that the combination with a b-lactamase inhibitor is advisable or even necessary.

For cervicofacial actinomycoses, amoxicillin plus clavulanic acid has proved to be the treatment of choice. Three doses of 2.0 g amoxicillin plus 0.2 g clavulanic acid per day for 1 week and three doses of 1.1 g of the combination for an additional 7 days usually result in complete cure. Thoracic actinomycoses mostly respond to the same regimen. However, it is advisable to maintain doses of 2.2 g, three times per day, for 2 weeks, and to continue treatment for 3 to 4 weeks. Advanced pulmonary cases may require the addition of 2 g ampicillin, three times a day, in order to increase the tissue concentration of aminopenicillin and, depending on the composition of the concomitant flora, the use of an antimicrobial specifically active against resistant enterobacteriaceae; the application of drugs such as metronidazole or clindamycin against strict anaerobes is only necessary, as an adjunct to the aminopenicillins, in chronic cases with reduced blood supply.

Since in abdominal actinomycoses enterobacteriaceae and b-lactamase producing *Bacteroides* species are usually present and the correct diagnosis is mostly established late, suitable antimicrobial combinations for these cases are amoxicillin plus clavulanic acid plus metronidazole plus tobramycin (gentamicin) or ampicillin plus clindamycin plus an aminoglycoside. Imipenem might also be a good choice, but this drug has not yet been widely used for treating actinomycotic infections.

Neither clindamycin nor metronidazole should be used alone. Clindamycin is almost completely ineffective against *Actinobacillus actinomycetemcomitans* and metronidazole shows no activity at all against pathogenic actinomycetes. The use of further combinations, including additional aminoglycosides, cephalosporins, or b-lactamase-stable penicillins, may be necessary depending on the presence of unusual aerobic organisms. In patients allergic to penicillins, tetracyclines or possibly cephalosporins may be tried instead of aminopenicillins. Incision of abscesses and drainage of pus may still be necessary as an adjunct to the antimicrobial chemotherapy and may help to accelerate recovery and to decrease the risk of relapses.

Prognosis

The prognosis of cervicofacial and cutaneous actinomycotic infections is good provided that the diagnosis is established early and antimicrobial treatment is adequate. However, thoracic, abdominal, and systemic manifestations remain serious conditions that require all possible diagnostic and therapeutic efforts. Without proper treatment, the prognosis is grave.

Epidemiology

Actinomycoses are not transmissible and cannot be brought under control by vaccination or by measures that prevent spread. Sporadically, they occur worldwide. In Germany, the incidence of the disease was estimated to range from 1 in 40 000 (acute and chronic cases together) to 1 in 80 000 (chronic cases alone) per year, but appears to be decreasing in recent years.

Adult males are affected two to four times more frequently than are females by cervicofacial actinomycoses. Although actinomycoses may be found in patients of any age, men are predominantly affected between their 20th and 40th years and women in the second and third decade of their lives. Before puberty and in old age, actinomycoses occur sporadically in patients of both sexes.

Other diseases caused by fermentative actinomycetes

Fermentative actinomycetes play some part in dental caries and periodontal disease, but are clearly not the most important microbes contributing to these important health problems. Lacrimal canaliculitis with and without conjunctivitis is commonly caused by fermentative actinomycetes, in particular *Propionibacterium propionicum*, *Actinomyces viscosus*, or *A. israelii* and rarely by other actinomycete species. The concomitant flora, when present, is usually less complex than that of typical actinomycoses. Removal of the lacrimal concretions that are usually present and local application of antimicrobials always result in prompt cure.

Arcanobacterium pyogenes and *A. haemolyticum* (formerly '*Corynebacterium* (*Actinomyces*) *pyogenes*' and '*C. haemolyticum*') cause acute pharyngitis, urethritis, or cutaneous or subcutaneous suppurations. The recently described species *Actinomyces neuii* subspecies *neuii* and subspecies *anitratus*, *A. graevenitzii*, *A. europaeus*, *A. radingae*, *A. turicensis*, *A. funkei*, *A. radicidentis*, and *A. urogenitalis*, as well as *Arcanobacterium* (*Actinomyces*) *bernardiae* and *Actinobaculum schaalii* have been isolated from various clinical sources including abscesses and blood cultures, and may also be associated with mixed bacterial flora. *A. turicensis* and possibly *A. urogenitalis* seem to be particularly common in genital infections while *A. radingae* was found only in patients with skin-related pathologies, *A. europaeus* was detected in patients with urinary tract infections, and *A. radicidentis* was isolated from infected root canals of teeth.

Further reading

McNeil MM, Schaal KP (1998). Actinomycoses. In: Yu VL, Merigan TC Jr, Barriere SL, eds. *Antimicrobial therapy and vaccines*, pp 14–22. Williams and Wilkins, Baltimore.

Schaal KP (1986). Genus *Arachnia* Pine and Georg 1969, 269. In: Sneath PHA, Mair NS, Sharpe ME, Holt JG, eds. *Bergey's manual of systematic bacteriology*, Vol. 2, pp 1332–42. Williams and Wilkins, Baltimore.

Schaal KP (1986). Genus *Actinomyces* Harz 1877, 133. In: Sneath PHA, Mair NS, Sharpe ME, Holt JG, eds. *Bergey's manual of systematic bacteriology*, Vol. 2, pp 1383–418. Williams and Wilkins, Baltimore.

Schaal KP (1992). The genera *Actinomyces, Arcanobacterium*, and *Rothia*. In: Balows A, Trüper HG, Dworkin M, Harder W, Schleifer KH, eds. *The prokaryotes. A handbook on the biology of bacteria: ecophysiology, isolation, identification, applications*, 2nd edn, Vol. 1, pp 850–905. Springer, Berlin.

Schaal KP, Lee HJ (1992). Actinomycete infections in humans – a review. *Gene* **115**, 201–11.

Schaal KP, Pulverer G (1984). Epidemiologic, etiologic, diagnostic, and therapeutic aspects of endogenous actinomycete infections. In: Ortiz-Ortiz L, Bojalil LF, Yakoleff V, eds. *Biological, biochemical, and biomedical aspects of actinomycetes*, pp 13–32. Academic Press, Orlando.

7.11.27 Nocardiosis

R. J. Hay

Nocardiosis (nocardiasis) is the infection caused by *Nocardia* species, usually *Nocardia asteroides* but, less commonly, *N. brasiliensis*, *N. otitidiscaviarum*, and *N. transvaliensis*. The term is most commonly applied to systemic infection due to these organisms but can also be used to describe cutaneous disease that follows the implantation of infection. These organisms are also important causes of actinomycetoma, particularly in Mexico and Central America.

The nocardias are Gram-positive, filamentous, branching bacteria that ramify in infected tissues. They can also break up into bacillary forms and, in some conditions, aggregate into grains typical of mycetomas. These organisms are aerobic and partially acid fast. They grow readily on ordinary laboratory media.

Pathogenesis

Nocardia species are found in soil, particularly where there is decaying vegetation. They can also be isolated from the air and, in most cases, systemic infection is by the airborne route; rarely nocardiosis can be acquired after inoculation into the skin. The characteristic histopathological response to infection is the production of polymorphonuclear leucocyte abscesses without extensive fibrosis. Caseation and palisading granulomas are not generally seen. Metastases can occur in other organs. Dissemination of infection to the skin can occur in such systemic infections. By contrast, in primary cutaneous infections the lesion is usually localized to an abscess containing filaments at the site of inoculation and is accompanied by local lymphadenopathy. Mycetoma grain formation may occur in some of these infections that follow inoculation. It is not known why, in some patients, transcutaneous infection with nocardia results in the development of a mycetoma whereas in others a subcutaneous abscess containing filaments is formed. The tendency to develop into mycetomas appears to be more common with *N. brasiliensis* infections.

Epidemiology

Otherwise healthy patients may be infected by nocardia, although the frequency of subclinical exposure and sensitization in normal populations is unknown. However, the majority of patients with systemic nocardiosis are immunocompromised, most commonly with a condition that affects the expression of T-lymphocyte-mediated immune responses. The list of underlying conditions includes:

1. malignancies, including cancer and lymphoma;
2. AIDS and other immunodeficiency states such as chronic granulomatous disease;
3. solid-organ transplantation;
4. other conditions that require high doses of corticosteroids, such as collagen-vascular disease and rheumatoid arthritis; and
5. pre-existing pulmonary disease—alveolar proteinosis, in particular, seems to predispose to nocardiosis.

The usual site of primary infection is the lung and the disease may remain restricted to this site. It may also be disseminated to other organs, particularly to the brain and skin. Nocardiosis can occur at any age, although it is rare, particularly in childhood.

Clinical features

Primary cutaneous nocardiosis

This is an uncommon infection that appears to follow traumatic inoculation of organisms in a superficial abrasion. The usual primary lesion is a small nodule, ulcer, or abscess at the site of inoculation. There may be a small chain of secondary nodules (cf. sporotrichosis) along the course of a lymphatic and local lymphadenopathy is common. Some such cases resolve spontaneously. This form of disease is usually caused by *N. asteroides*.

Nocardia mycetoma

This is discussed in Chapter 7.12.1. *N. brasiliensis* is the usual cause.

Pulmonary nocardiosis

Pulmonary infection is seen in about 75 per cent of cases of systemic nocardiosis, even where there are disseminated lesions elsewhere. Symptoms of pulmonary nocardiosis are variable, with cough, fever, and leucocytosis. In otherwise healthy individuals the changes and signs may be very similar to pulmonary tuberculosis, whereas in the immunocompromised patient the lesions present as rapidly developing, single or multiple lung lesions. In patients with AIDS, symptoms are often minimal, even in the presence of extensive disease. These changes are reflected by the course of the disease. In some patients, progression is rapid, in others chronic.

Chest radiographs may show segmental or lobar infiltrates, cavitation, nodules, or diffuse miliary infiltrates. Calcification is not common. The infection may spread locally to involve adjacent structures such as the pleural space and diaphragm or may spread to other sites. Very occasionally, nocardias can be isolated from sputum of otherwise healthy patients. Whether this reflects the process of asymptomatic sensitization is not known. Most cases of pulmonary nocardiosis are caused by *N. asteroides*.

Disseminated nocardiosis

Haematogenous spread is common in the immunocompromised patient and may occur without evidence of pulmonary infection. The most common site for dissemination is the brain, where it presents with localized abscesses without meningeal involvement. The signs are those due to an intracerebral space-occupying lesion. Spread to other sites is less common, although dissemination to skin, liver, kidneys, and bone may occur.

The acute disseminated forms and those with involvement of the central nervous system have the worst prognosis. Continued therapy with corticosteroids also appears to have bad prognostic significance. Infection in patients with AIDS may not be recognized before death. Rapid diagnosis is therefore a key to successful management. By contrast, pulmonary infection in otherwise healthy patients is usually a chronic process and has to be distinguished from tuberculosis.

Laboratory diagnosis

The infection is often recognized initially by direct microscopy of pus, bronchial washings, or tissue. In Gram stains the organisms can be shown as fine, branching filaments, although distinction from other bacteria may be difficult if short, rod-like forms predominate. A modified acid-fast stain using weak acid can be used to

demonstrate filaments.

Nocardia species grow on ordinary media aerobically. Colonies may take 2 to 3 weeks to appear and cultures need prolonged incubation. Growth is generally more rapid on Lowenstein–Jensen medium.

Histopathological examination is useful in some cases. Filaments stain with modified acid-fast stains using an aqueous solution of a weak acid for decolorization, but can also be highlighted with the methenamine–silver stain (Grocott modification). The branching nature of the organism is best appreciated in histopathological material. Other pathogens such as *Pneumocystis* species may also be present in histopathological material.

Serological tests (usually counterimmunoelectrophoresis or enzyme immunoassay) can be obtained in reference centres and are generally used to monitor the progress of therapy rather than establish the diagnosis.

Therapy

The mainstays of therapy are sulphonamides such as sulphadiazine and sulphafurazole, given in doses of 4 to 6 g daily. Co-trimoxazole is also effective, particularly in pulmonary forms, although the ratio of the trimethoprim to sulphonamide components is not ideal for intracerebral infections. In many cases, drainage of abscesses may hasten recovery. Unfortunately, there have been no multicentre clinical studies aimed at reaching a consensus on the most appropriate therapy for this uncommon infection. Thus, much of the recommended drug therapy is derived from the personal experiences of few cases. It is, for instance, the general practice to use two antibiotics.

Other drugs that have been used include amikacin, ampicillin, and minocycline—although testing is necessary before using these. Experience of other drugs is similarly limited. For instance, ciprofloxacin, cefotaxime, and imipenem are all active *in vitro* but clinical experience with them is limited at present.

Clustering of cases may occur occasionally, suggesting exposure to a common source of infection. In two such episodes there had been extensive construction work in the vicinity of the hospital involved. At present, no methods of prevention are known, although the existence of more than two cases in a single or adjacent wards should alert clinicians to the possibility of environmentally acquired infection.

Further reading

Boiron P *et al.* (1992). Review of nocardial infections in France, 1987–1990. *European Journal of Clinical Microbiology and Infectious Diseases* **11**, 709–14.

Curry WA (1980). Human nocardiosis. *Archives of Internal Medicine* **140**, 818–24.

Georghiou PR, Blacklock ZM (1992). Infection with *Nocardia* species in Queensland. A review of 102 clinical isolates. *Medical Journal of Australia* **156**, 692–7.

Hay RJ (1983). Nocardial infections of the skin. *Journal of Hygiene* **91**, 385–91.

Houang ET *et al.* (1980). *Nocardia asteroides* infection—a transmissible disease. *Journal of Hospital Infection* **1**, 31–6.

Javaly K, Horowitz HW, Wormser GP (1992). Nocardiosis in patients with human immunodeficiency virus infection. Report of two cases and review of the literature. *Medicine* **71**, 128–38.

Sakai C, Takagi T, Satoh Y (1999). *Nocardia asteroides* pneumonia, subcutaneous abscess and meningitis in a patient with advanced malignant lymphoma: successful treatment based on *in vitro* antimicrobial susceptibility. *Internal Medicine* **38**, 683–6.

7.11.28 Rat-bite fevers

D. A. Warrell

Introduction

Rat bites are not uncommon, even in cities. Young children are often bitten while asleep. Patients with diabetic or leprous neuropathy are particularly vulnerable. Rodent bites can transmit lymphocytic choriomeningitis and other arenaviruses, rabies, leptospirosis, melioidosis, tularaemia, plague, murine typhus, trench fever, *Pasteurella multocida*, and the two rat-bite fevers caused by *Streptobacillus moniliformis* and *Spirillum minus*.

Streptobacillus moniliformis infection (streptobacillary rat-bite fever and Haverhill fever)

This organism is part of the normal pharyngeal flora of up to 50 per cent of wild and laboratory rats and can be recovered from the nasopharynx, middle ear, saliva, and urine. It can also cause severe disease in rodents: septicaemia, pneumonia, conjunctivitis, polyarthritis, and abortion. It has been isolated from rats, mice, guinea-pigs, gerbils, squirrels, and turkeys as well as animals that feed on rodents such as cats, dogs, pigs, ferrets, and weasels.

S. moniliformis derives its name from the filaments and chains with yeast-like swellings seen in mature cultures on solid media. It is a non-motile, pleomorphic, filamentous, Gram-negative rod, 1 to 5 µm long, and is microaerophilic. It can be grown in ordinary blood culture media, but thrives only when blood, serum, or ascitic fluid are added (for example, trypticase soy agar with 20 per cent horse or rabbit serum added under 8 per cent CO_2). In liquid media, 'puff ball' colonies appear in 1 to 6 days. In concentrations exceeding 0.0125 per cent, sodium polyanethol sulphonate ('Liquoid'), a laboratory anticoagulant often added to blood culture broths for isolating aerobic bacteria, inhibits the growth of *S. moniliformis*. In culture, L-phase variants occur spontaneously. These lack a cell wall and are therefore resistant to penicillin. The organism has been cultured from patients' bite wounds, blood, synovial and pericardial fluid, and from abscesses.

Epidemiology

The infection occurs worldwide in two forms. Rat-bite fever is caused by bites or scratches by rodents or their predators or mere contact with these mammals whether living or dead. In some countries, 10 per cent of those bitten by wild rats will be infected. Most rat-bite victims are children of poor families living in urban areas. The bite may not be suspected since many are inflicted while the patient is asleep. Laboratory staff who work with rats are also at special risk. Haverhill fever, named after a town in Massachusetts, follows ingestion of raw milk, food, or water contaminated by rats. An outbreak in a boarding school in England in 1983 affected 304 people, 43 per cent of the school's population, and was attributed to contamination of the water supply by rats.

Clinical features

After an incubation period, which is usually less than 10 days and often as short as 1 to 3 days, there is a sudden high fever with rigors, vomiting, severe headache, myalgia, and muscle tenderness. Evidence of the bite has usually disappeared by this stage. About 75 per cent of patients develop a rash between 1 and 8 days later. Discrete erythematous macules, 1 to 4 mm in diameter, appear symmetrically on the lateral and extensor surfaces and over the joints. They are often most marked on the hands and feet (palms and soles) with associated petechiae, but they also occur on the face. Papules, vesicles, and pustules with scabs have also been described. About half the patients develop an asymmetrical migratory polyarthralgia or arthritis, usually involving the knees, ankles, elbows, shoulders, and hips and often associated with effusions. Joint pains may be the dominant symptom in patients with rat-bite fever. Diarrhoea and loss of weight are described in young children. Fever and other symptoms subside in a few days in treated cases, but fever may persist for 1 to 2 weeks (or relapse over several months) and arthritis for many months in those untreated. Severe infections can lead to bronchitis, pneumonia, metastatic abscess formation (including cerebral abscess), myocarditis, pericarditis with effusion, subacute glomerulonephritis, interstitial nephritis, splenitis or splenic abscess, amnionitis, and anaemia. Infective endocarditis, usually with underlying rheumatic or other valve disease, has been described in 18 cases, one with human immunodeficiency virus (**HIV**) infection.

Haverhill fever (erythema arthriticum epidemicum) follows a similar clinical course after the patient has drunk unpasteurized milk or contaminated water. Vomiting, stomatitis, and upper respiratory tract symptoms such as sore throat are said to be more prominent than in rat-bite fever.

Diagnosis

Unlike *Spirillum minus* infection (sodoku), the incubation period is short, the bite wound heals permanently with little local lymphadenopathy, the rash is morbilliform or petechial, and arthritis is common.

The diagnosis can be confirmed by culturing the organism from blood, joint fluid, or pus. In patients with infective endocarditis the differential diagnosis of the slow-growing, microaerophilic organism will include *Haemophilus aphrophilus*, *Cardiobacterium hominis, Actinomyces actinomycetencomitans*, and *Eikenella corrodens*. A high or rising titre of agglutinins, complement-fixing or fluorescent antibodies, may be detected between 2 and 3 weeks. A peripheral leucocytosis of 10 000 to 30 000/µl is usual and false-positive serological tests for syphilis are found in 15 to 25 per cent of cases.

Treatment

Streptobacillus moniliformis is sensitive to penicillin and can be treated with procaine benzylpenicillin (adult dose 600 mg or 600 000 units) by intramuscular injection every 12 hours for 7 to 14 days, or by penicillin-V 2 g a day by mouth. Penicillin-resistant L-variants are susceptible to streptomycin, tetracycline, and probably erythromycin. For patients hypersensitive to penicillin, erythromycin, chloramphenicol, tetracycline, or cephalosporins can be used. Erythromycin was used successfully in the boarding-school outbreak of Haverhill fever in England in 1983.

Patients with endocarditis should be treated with intravenous benzylpenicillin, 4.8 to 14.4 g (8–24 000 000 units) each day for between 4 and 6 weeks, or 4.8 mega units of procaine benzylpenicillin daily by intramuscular injection for 4 weeks if the cultured organism has a sensitivity of 0.1 µg/ml. The addition of streptomycin improves bactericidal activity and eliminates L-forms.

Prognosis

The untreated case fatality was reported to be 10 to 13 per cent. However, the overall mortality in patients with endocarditis is about 50 per cent. Residual arthralgia,

persisting for as long as 10 years, has been described.

Spirillum minus infection (sodoku, sokosha)

Spirillum minus may be found in the blood of up to 25 per cent of apparently healthy rodents and in the eye discharge and mouths of rats with interstitial keratitis and conjunctivitis. *S. minus* is a relatively thick, tightly coiled, Gram-negative spirillum (**not** a spirochaete), between 2.5 and 5.0 µm long, with 2 to 6 (commonly 3) spirals, resembling campylobacters. It darts about under the power of its terminal flagella. Continuous culture on artificial media has not been achieved, but the organism can be demonstrated by inoculating material from the bite wound, regional lymph nodes, or blood intraperitoneally into mice or guinea-pigs. Organisms usually appear in the rodent's blood within 5 to 15 days of inoculation.

Epidemiology

Sodoku is found worldwide but is particularly common in Japan. It results from bites, scratches, or mere contact with rodents or their predators including dogs, cats, and pigs.

Clinical features

The initial bite wound usually heals without signs of local inflammation. After an incubation period of between 5 and 30 days, usually 7 days or more, there is sudden fever which, in untreated cases, reaches its height in 3 days and resolves by crisis after a further 3 days. Other acute symptoms include rigors, myalgia, and prostration. At the start of the illness the healed bite wound becomes inflamed and swollen; it may break down to become necrotic or suppurative. Regional lymph nodes are usually enlarged and tender. The exanthem often starts at the site of the bite and spreads from there. It consists of angry purplish or reddish-brown indurated papules, plaques, or macules with urticarial lesions. Arthralgia may be severe but there are no joint effusions. Severe manifestations including meningitis, cerebral abscess, encephalitis, endocarditis, myocarditis, myocardial abscess, pleural effusion, chorioamnionitis, subcutaneous abscesses, and involvement of liver, kidney, and other organs are seen in about 10 per cent of cases. Relapses of fever, rash, and other symptoms lasting 3 to 6 days may occur between remissions of a week or so for 2 to months and occasionally up to a year in untreated patients.

Diagnosis

Clinically, sodoku is distinguishable from streptobacillary rat-bite fever by its longer incubation period, by the marked reaction at the bite site with local lymphadenopathy at the start of symptoms, by the different rash (dark papular rather than morbilliform and petechial), and by the rarity of arthritis. The diagnosis can be confirmed by examining an aspirate from the bite wound, lymph nodes, exanthem, or blood (thick and thin films) by dark-field microscopy or by staining with Wright's or Giemsa stain. Spirochaetes can be detected in the blood, peritoneal fluid, or heart muscle of inoculated rodents but cannot be cultured on artificial media. No specific serological tests are available. False-positive serological tests for syphilis are found in 50 to 60 per cent of cases, and reactions with Proteus OXK are also common.

Differential diagnosis of rat-bite fevers

An acute, severe, febrile illness following a rat bite, or other contact with rodents or their predators, should raise the possibility of other rodent-related infections. These include: *Pasteurella multocida*, which produces local pain and erythema within a few hours of the bite; plague; tularaemia; leptospirosis; murine typhus; and arenaviruses such as lymphocytic choriomeningitis, Lassa fever (Africa), or Argentine, Bolivian, or Venezuelan haemorrhagic fevers (South America). Ingestion of raw milk should also raise the possibility of brucellosis.

Treatment

Penicillin is the drug of choice. For adults, procaine benzylpenicillin 600 mg (600 000 units) should be given every 12 hours for 7 to 14 days. Penicillin-V, 2 g/day by mouth, is also said to be effective. A Jarisch–Herxheimer reaction may complicate penicillin treatment.

Prognosis

Untreated case fatality is about 2 to 10 per cent.

Prevention of rat-bite fevers

These infections can be prevented by rodent control, by encouraging laboratory workers to wear gloves and use correct techniques when handling rodents, to clean all rodent bite wounds, and to take prophylactic penicillin when bitten. Haverhill fever is prevented by avoiding the con-sumption of raw milk, by monitoring water supplies (especially those not derived from the mains), and by controlling rat populations.

Further reading

McEvoy MB, Noah ND, Pilsworth R (1987). Outbreak of fever caused by *Streptobacillus moniliformis*. *Lancet* **ii**, 1361–3.

Raffin BJ, Freemark M (1979). Streptobacillary rat bite fever: a pediatric problem. *Pediatrics* **64**, 214–17.

Roughgarden JW (1965). Antimicrobial therapy of rat bite fever. A review. *Archives of Internal Medicine* **116**, 39–54.

Rupp ME (1992). *Streptobacillus moniliformis* endocarditis: case report and review. *Clinical Infectious Diseases* **14**, 769–72.

7.11.29 Lyme borreliosis

John Nowakowski, Robert B. Nadelman, and Gary P. Wormser

Lyme borreliosis (also called Lyme disease) is an infection caused by the spirochaete, *Borrelia burgdorferi*, which is transmitted to humans by the usually asymptomatic bite of certain ticks of the genus *Ixodes* (Plate 1). The entire chromosome of *Borrelia burgdorferi* and 11 of its plasmids have been sequenced. Ticks acquire this borrelial infection in a complex tick–vertebrate transmission cycle. The white-footed mouse is the most important reservoir for *B. burgdorferi* in North America, but in Europe a variety of small mammals and birds are involved, possibly reflecting the much more varied and complex ecology of the *Ixodes* ticks in Eurasia. White-tailed deer, an important host for adult *Ixodes* ticks, are not a reservoir for *B. burgdorferi*.

Lyme borreliosis occurs in north-eastern, mid-Atlantic, north-central, and far western regions of the United States, limited foci in Canada (mainly in eastern Ontario), and much of Europe and northern Asia. Migrating birds may play a role in the spread of ticks and *B. burgdorferi* to new geographical locations.

Lyme borreliosis occurs equally in males and females, and affects people of all ages. There is a bimodal age distribution with the highest rates in children 5 to 9 years old and adults more than 30 years old.

Clinical manifestations

The somewhat different manifestations of Lyme borreliosis in Eurasia compared with North America (Table 1) may be explained by the wider variety of genospecies of *B. burgdorferi*. Clinical features are similar in adults and children.

Erythema migrans (Plate 2)

Erythema migrans, the clinical hallmark of Lyme borreliosis, is recognized in approximately 90 per cent of patients with objective evidence of *B. burgdorferi* infection. Typically, erythema migrans begins as a red macule or papule at the site of a tick bite that occurred 7 to 10 days earlier. The rash expands over days to weeks. Central clearing may or may not be present. Secondary cutaneous lesions may develop after haematogenous spread of spirochaetes. Erythema migrans must be distinguished from local tick bite reactions **,**tinea, insect and spider bites, bacterial cellulitis, and plant dermatitis.

Systemic complaints in patients with erythema migrans are more common in the United States than in Europe, perhaps as a result of illness caused by a more virulent genospecies (*B. burgdorferi sensu stricto* rather than *B. afzelii*) or more frequent coinfection with other tickborne pathogens. Symptoms include fatigue, myalgia, arthralgia, headache, fever and/or chills, and stiff neck. Prominent respiratory and/or gastrointestinal complaints are so infrequent that their presence should suggest an alternative diagnosis or coinfection. The most common objective physical findings are regional lymphadenopathy and fever. Occasional cases of febrile viral-like illness without erythema migrans have been attributed to Lyme borreliosis.

Carditis

Typically cardiac disease develops within weeks to months after infection. It is usually manifested by fluctuating degrees of atrioventricular block which may cause the patient to complain of dizziness, palpitations, dyspnoea, chest pain, or syncope. Pericarditis with effusion is rarely observed. The incidence (as measured by ECG-confirmed heart block) has been observed to be low in both the United States (< 1 per cent) and Europe (< 4 per cent). *B. burgdorferi* has been recovered in culture from the myocardium of several European patients with congestive heart failure including two with acute myocarditis and one with chronic cardiomyopathy.

Neurological disease

The incidence of neurological Lyme disease in Europe (20 per cent) may be higher than in the United States (< 10 per cent). One explanation may be the greater neurotropism of *B. garinii* (a genospecies which has not been isolated in North America). The principal early neurological manifestations are cranial neuropathy (typically 7th nerve palsy), radiculopathy, and meningitis, which may occur alone or together. Late neurological manifestations are uncommon and include peripheral neuropathy, encephalopathy, and encephalomyelitis.

Antibiotics appear to hasten the resolution of meningitis but most studies are uncontrolled. The rate of resolution of motor dysfunction, which is fully reversible in the vast majority of cases, is not enhanced by antimicrobial therapy. Symptoms of encephalopathy and peripheral neuropathy improve or do not progress after treatment with antibiotics.

Rheumatological disease

Lyme arthritis is more frequently diagnosed in North America than in Europe. In a study of 55 untreated patients with erythema migrans diagnosed in the United States between 1977 and 1979, followed for a mean duration of 6 years, objective arthritis developed in more than half, occurring within 1 year for 90 per cent of patients. Without antibiotic treatment, intermittent attacks of migratory monoarthritis or asymmetric oligoarthritis occur, lasting a mean of 3 months (range 3 days to 11.5 months). The knee is affected at some point in almost all patients, but other large and (less often) small joints may be affected. Temporomandibular joint involvement occurred in 11 (39 per cent) of 28 patients with arthritis in one series. Although large effusions may occur, joint pain and erythema are often minimal. Baker's cysts may develop. Typically, synovial fluid analysis reveals a modestly elevated protein and white cell count (median 24 250 white cells/mm ^{3}; range 2100 to 72 250 white cells/mm^{3}) with a polymorphonuclear predominance and a normal glucose level. Synovitis lasting 1 year or more may ensue for a minority of American patients, sometimes associated with joint destruction. Although *B. burgdorferi* DNA can be detected by polymerase chain reaction (**PCR**) in the synovial fluid of up to 85 per cent of untreated patients with Lyme arthritis, *B. burgdorferi* has rarely been successfully cultured from joint fluid. In patients who receive antibiotics, the presence of *B. burgdorferi* can no longer be detected by PCR in repeat synovial fluid examinations.

Acrodermatitis chronica atrophicans

This develops insidiously on a distal extremity. It is a skin lesion that is swollen, bluish-red, and which ultimately atrophies. One-third of patients have an associated (usually sensory) polyneuropathy. *B. burgdorferi* has been recovered from skin biopsy specimens of acrodermatitis chronica atrophicans lesions of more than 10 years' duration. Since the usual causative agent, *B. afzelii*, does not occur in the United States, acrodermatitis chronica atrophicans is essentially a European disease.

Miscellaneous clinical manifestations

Borrelia lymphocytoma, principally caused by *B. afzelii* and *B. garinii*, is a tumour-like nodule which typically appears on the pinna of the earlobe or on the nipple or

areola of the breast. Lesions resolve spontaneously but disappear within a few weeks after antibiotics. This lesion does not occur in North America.

Direct involvement of the eye (e.g. uveitis, keratitis, vitritis, or optic neuritis) has been attributed to *B. burgdorferi* infection. However, since ophthalmological disorders have almost never been associated with the isolation of *B. burgdorferi* in culture, the actual pathogenesis in these cases is uncertain. Conjunctivitis, originally described in 11 per cent of patients with erythema migrans, was rare (< 5 per cent) in recent studies of culture-positive patients.

Case reports have suggested that adverse outcomes may be associated with pregnancies complicated by maternal Lyme borreliosis. The risk of transplacental transmission of *B. burgdorferi*, however, is probably minimal when appropriate antibiotics (Table 2) are given to pregnant women with Lyme borreliosis. There are no published data to support a congenital Lyme borreliosis syndrome.

Laboratory tests

Where Lyme borreliosis is endemic, diagnosis of erythema migrans is purely clinical. Laboratory testing is neither necessary nor recommended. However, culture is virtually 100 per cent specific and appears to be more sensitive (57 to 86 per cent) than serology (50 per cent in the United States and less than 50 per cent in Europe).

In patients with suspected extracutaneous Lyme borreliosis, serological testing is essential to support the diagnosis. Culture of *B. burgdorferi* has been a highly insensitive diagnostic technique for this group of patients, presumably because of inaccessibility of tissues which contain the organism.

A two-step approach to serological diagnosis has recently been proposed in the United States (and is being studied in Europe) to increase the specificity of a positive test. A positive or equivocal first-stage test (usually an enzyme-linked immunosorbent assay [ELISA] or indirect immunofluorescence assay [IFA]) is followed on the same serum sample by a second-stage test (immunoblot). Two-step testing, however, is not indicated for those with little or no clinical evidence of Lyme borreliosis because of a low positive predictive value. Since IgM and IgG antibodies to *B. burgdorferi* may persist in serum for years after clinical recovery, serology has no role in measuring response to treatment.

Patients with extracutaneous Lyme borreliosis almost always have diagnostic serum antibodies to *B. burgdorferi*, except for some patients with early 7th nerve palsy or occasional patients in whom antibodies to *B. burgdorferi* are present in cerebrospinal fluid only.

Coinfection

Ixodes scapularis (Plate 1) ticks are the vectors for several other infections which may be transmitted separately or simultaneously with *B. burgdorferi* such as *Babesia microti*, and the rickettsial agent that causes human granulocytic ehrlichiosis. In Europe, species of *Babesia* and *Ehrlichia* are present in *Ixodes ricinus* ticks, which are also vectors for a flavivirus causing tickborne encephalitis. Coinfection may alter the clinical presentation and response to treatment of Lyme borreliosis.

Treatment

Although most manifestations of Lyme borreliosis resolve spontaneously, antibiotics may speed the resolution of some manifestations and almost certainly will prevent the progression of disease. An approach to treatment is summarized in Table 2. Currently available quinolones, sulpha drugs, first-generation cephalosporins, rifampicin, and aminoglycosides have no appreciable activity against *B. burgdorferi* and should not be used. In addition, there is no evidence to support combination antimicrobial therapy, prolonged (> 1 month) or repeated courses of antibiotics, and 'pulse' or intermittent antibiotic therapy. Within 24 h after initiation of antibiotics, approximately 15 per cent of patients may develop transient intensified signs (e.g. rash and fever) and symptoms (e.g. arthralgias) consistent with a Jarisch–Herxheimer reaction. Treatment is symptomatic.

Most people treated for Lyme borreliosis have an excellent prognosis. Although some patients treated for erythema migrans in recent series continue to have a variety of mild non-specific complaints following antibiotic therapy, the development of objective extracutaneous disease after treatment is extremely rare. Lyme borreliosis may trigger a fibromyalgia syndrome that does not appear to respond to repeated courses of antibiotics, but may improve with symptomatic therapy. Patients with carditis and neurological disease also tend to do well, but may sometimes have residual deficits (e.g. mild 7th nerve palsy) after treatment. In patients with arthritis, clinical recovery typically occurs in conjunction with oral antibiotic therapy (often with a non-steroidal anti-inflammatory medication); occasionally such patients with subtle signs of neuroborreliosis who are treated with oral antibiotics may develop overt neuroborreliosis and require parenteral therapy. A small number of American patients with Lyme arthritis and the HLA DR4 haplotype, who continue to have synovial inflammation for months or even several years after the apparent eradication of *B. burgdorferi* from the joint following antibiotic therapy, have improved after synovectomy. An immunological mechanism rather than active infection may be responsible for the continued inflammatory response in these patients.

A sizeable number of American patients with a variety of complaints of uncertain aetiology, including pain and fatigue syndromes, have been labelled as having 'chronic Lyme disease' or 'post-Lyme syndrome'. This entity is controversial.

Prevention

This includes avoiding exposure by limiting outdoor activities in tick-infested locations, using tick repellents, tucking in clothing to decrease exposed skin surfaces, and frequent skin inspections for early detection and removal of ticks. Use of acaracides on property and construction of deer fences have also been proposed.

Antibiotic prophylaxis given after recognized *I. scapularis* tick bites has not been shown to be effective in reducing the low (< 5 per cent) risk of acquiring Lyme borreliosis after tick bites. Vaccination with a single recombinant outer surface protein A (**OspA**) preparation has been found to be safe and effective for preventing Lyme disease in the United States. The efficacy of this OspA vaccine may be related to the ability of OspA antibodies (ingested during the blood meal by the vector tick) to kill *B. burgdorferi* in the tick gut, thus preventing transmission of the spirochaete. A single antigen OspA vaccine is expected to be less effective in Eurasia where species of *Borrelia* are more heterogeneous and OspA is more variable.

Further reading

Aguero-Rosenfeld M *et al*. (1993). Serodiagnosis in early Lyme disease. *Journal of Clinical Microbiology* **31**, 390–5. [Serological response to *Borrelia burgdorferi* by ELISA and immunoblot in early Lyme disease.]

Barbour AG, Fish D (1993). The biological and social phenomenon of Lyme disease. *Science* **260**, 1610–16. [Description of the emergence of the disease in the United States and its importance to human and animal health.]

Dattwyler RJ *et al*. (1997). Ceftriaxone compared with doxycycline for the treatment of acute disseminated Lyme disease. *New England Journal of Medicine* **337**, 289–94. [Study demonstrating equivalent efficacy of oral doxycycline compared with intravenous ceftriaxone in patients with disseminated early Lyme disease.]

Nadelman RB, Wormser GP (1998). Lyme borreliosis. *Lancet* **352**, 557–65. [Comprehensive review of the disease.]

Steere AC, Schoen RT, Taylor E (1987). The clinical evolution of Lyme arthritis. *Annals of Internal Medicine* **107**, 725–31. [Description of the progression to Lyme arthritis in untreated patients with erythema chronicum migrans.]

Strle F *et al*. (1999). Comparison of culture-confirmed erythema migrans caused by *Borrelia burgdorferi sensu stricto* in New York State and by *Borrelia afzelii* in Slovenia. *Annals of Internal Medicine* **130**, 32–6. [Clinical comparison of the disease in the United States and Slovenia.]

7.11.30 Other borrelia infections

D. A. Warrell

The borreliae are large, loosely coiled, motile spirochaetes. *B. recurrentis* (Plate 1), first described by Obermeier in 1867, causes louse-borne relapsing fever; *B. duttonii* and a number of other species or groups of *Borrelia* cause tick-borne relapsing fever; and *B. burgdorferi* causes Lyme disease. *B. vincenti* (now renamed *Treponema vincentii*) was, with *Fusobacterium (Bacteroides) nucleatum (fusiforme)*, implicated in acute necrotizing ulcerative gingivitis and Vincent's angina but is now regarded as part of the normal oral flora.

Relapsing fevers

The borreliae that cause relapsing fevers are spirochaetes, 8 to 20 µm long and 0.2 to 0.6 µm thick with between 3 and 15 coils and, in some strains, 15 to 30 axial filaments or flagella. These motile organisms divide by transverse binary fission. Several species of *Borrelia* including *B. recurrentis* can be cultured in Kelly's BSKII artificial media. *Borrelia* spp. can also be cultured on chick chorioallantoic membrane and perpetuated in rodents and ticks. Plasmid DNA has been detected in at least three *Borrelia* species.

Epidemiology

Louse-borne (epidemic) relapsing fever (LBRF)

Humans are probably the only reservoir of LBRF. The vector is the human body louse, *Pediculus humanus* and, to a lesser extent, the head louse, *P. capitis*. *B. recurrentis*, ingested by the louse during a blood meal, multiplies in its body cavity. Under conditions of crowding and poor hygiene, lice move from person to person. When the host's body surface temperature deviates far from 37 °C, as a result of death, fever, or exposure, or if infested clothing is discarded, the louse is forced to find a new host. A new person is infected when the infected louse is crushed and its body haemolymph applied to mucous membranes, such as to the conjunctiva by rubbing the eye, or to abraded skin, or inoculated through intact skin by scratching. Transmission is possible by blood transfusion, needlestick injuries, or even, in medical personnel, by contamination of broken skin such as paronychia on the fingers, by infected patients' blood. Unlike the tick vectors or tick-borne relapsing fever, which are also reservoirs of the infection, lice cannot transmit the infection transovarially to their progeny.

Wars, famine, and other disasters and the resulting large numbers of refugees favour the spread of lice and epidemic louse-borne infections such as relapsing fever and typhus. The yellow plague in Europe in AD 550 and the famine fevers of the seventeenth and eighteenth centuries were probably LBRF. During the first half of the twentieth century there were an estimated 50 million cases worldwide with a 10 per cent mortality. Epidemics began in Europe, the Middle East, and northern Africa during 1903, 1923, and 1943. An endemic focus persists in the Horn of Africa. In Ethiopia there is an annual epidemic of thousands of cases coinciding with the cool, rainy season. Poor people with lice-infested clothes crowd together for shelter. Recent outbreaks have occurred in the Sudan, Somalia, West Africa, and Vietnam. Since there is no known animal reservoir, the infection must persist in humans between epidemics, in mild or asymptomatic form.

Tick-borne (endemic) relapsing fever (TBRF)

There is a close relationship between particular species of *Borrelia*, their soft tick vectors, and reservoirs (Argasidae genus *Ornithodoros*) and mammal reservoir species. In East and Central Africa, domestic ticks of the *O. moubata* complex transmit *B. duttonii* between humans, the only one of these infections that is not a zoonosis. In North, West, and East Africa and the Middle East small rodents have burrows in or near human dwellings, and borreliae of the *Crocidurae* group, may be transmitted to man by the rodent tick *O. sonrai* (formerly *O. erraticus sonrai*). In the Central and West United States and Mexico, *O. hermsi*, a parasite of chipmunks and other tree squirrels, transmits *B. hermsi* to humans especially to those individuals who sleep in tick-infested log cabins near the Grand Canyon, Arizona. Other important borreliae causing tick-borne relapsing fever (and their tick vectors) include: *B. hispanica* (*O. sonrai*) in Africa; *B. persica* (*O. tholozani* = *O. papillipes*) in the Middle East; *B. venezuelensis* (*O. venezuelensis*) in Central and South America; and *B. turicatae* (*O. turicatae*), *B. parkeri* (*O. parkeri*) and *B. mazzotti* (*O. mazzotti*) in North America. Tick-borne relapsing fever may result when night-feeding ticks have access to man.

TBRF has occurred in most continents except Australasia and the Pacific region. It is particularly common in West Africa, where a recent survey revealed a prevalence of 1 per cent among children (in western Senegal). Each year 1650 proven cases are treated at one health centre in Rwanda (6 per cent of all patients). Although cases are usually isolated and sporadic in North America, in 1968 a total of 11 out of a group of 42 boy scouts were infected while camping in rodent-infested cabins on Browne Mountain, Washington; and in 1973 there were 62 cases among people staying in the log cabins along the north rim of the Grand Canyon in Arizona. During the past 25 years, 280 cases of TBRF have been identified in the United States. In Colorado the incidence is increasing (23 confirmed cases since 1977). In Jordan between 1959 and 1969 there were 723 cases of TBRF with four deaths.

Spirochaetes enter the tick in its blood meal from infected humans or animals. Unlike *B. recurrentis*, they invade the tick's salivary and coccal glands and genital apparatus and so can be transmitted when the tick feeds on a new host and transovarially to the tick's progeny. Unlike lice, ticks are reservoirs of *Borrelia* spp. They infest the burrows, caves, tree stumps, and roughly built shacks that harbour their mammalian hosts—rodents, insectivores, lagomorphs, bats, and small carnivores. In western countries, TBRF is occasionally diagnosed in travellers, intravenous drug abusers, and recipients of blood transfusions.

Pathophysiology

The physiological changes during the spontaneous crisis and the Jarisch–Herxheimer reaction (**J-HR**) induced by antimicrobial treatment in LBRF are typical of an 'endotoxin reaction'. Endotoxin-like activity has been described for some spirochaetes: *B. burgdorferi*, *Treponema hyodysenteriae*, *B. vincenti* and *B. buccalis*, and *Leptospira canicola*, but not in *B. recurrentis*, *B. hispanica*, or *Treponema pallidum*. It is the outer-membrane, variable major lipoprotein (**VMP**) of *B. recurrentis* that stimulates monocytes to produce tumour necrosis factor (**TNF**) through NF-kappaB. In patients treated with antibiotics, symptoms of the severe J-HR are associated with a transient marked elevation in plasma concentrations of tumour necrosis factor-a (**TNF-a**), interleukin-6, interleukin-8, and interleukin-1b (Fig. 1). The stimulus for cytokine release is the phagocytosis of spirochaetes made susceptible by the action of penicillin. Benzylpenicillin attaches to penicillin-binding protein I in *B. hermsi* spirochaetes. Large surface blebs are produced and the damaged spirochaetes are phagocytosed rapidly by neutrophils in the blood and by the spleen. Complement may enhance phagocytosis of spirochaetes, especially in the non-immune host, but the complement system is not essential for elimination of spirochaetes whether or not specific immunoglobulins are present. *In vitro*, surface contact with spirochaetes induces mononuclear leucocytes to produce pyrogen and thromboplastin, which could be responsible for the fever and disseminated intravascular coagulation in LBRF. Kinins may be released during the J-HR of syphilis and LBRF. The marked peripheral leucopenia that develops during the reaction reflects sequestration, perhaps in the pulmonary blood vessels, rather than leucocyte destruction. Spirochaetes may be found in those organs that bear the brunt of the infection (liver, spleen (Plate 2), myocardium, and brain), but it is unclear how their pathological effects are produced. The petechial rash results from thrombocytopenia not vasculitis. The cardiorespiratory and metabolic disturbances in relapsing

fever are principally the result of persistent high fever, accentuated by the J-HR or spontaneous crisis.

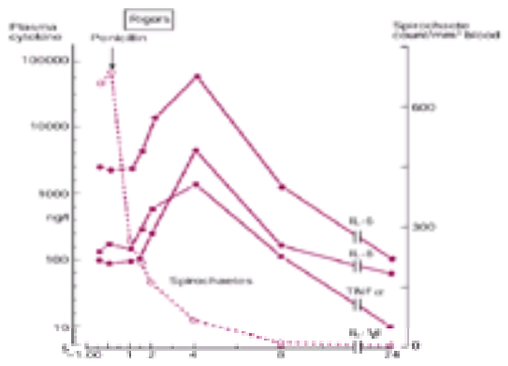

Fig. 1 Typical response in a patient treated with intravenous penicillin. Following penicillin, the number of spirochaetes fell abruptly; and circulating levels of TNF-a, IL-6, IL-8, and IL-1b started to rise after about 1 h, peaking at 4 h. This patient experienced sustained rigors as cytokine levels were increasing, which subsided before peak levels were achieved.

Immunological basis of the relapse phenomenon

Borrelia recurrentis exhibits antigenic variation of variable membrane proteins (**VMPs**), which are outer membrane lipoproteins. The organism has a repertoire of many VMPs but, at any one time, only one is expressed and is immunodominant. The expressed *VMP* gene is situated near the end of a linear plasmid and changes every 1 to 10 000 cell divisions. IgM is induced against the immunodominant VMP, leading to selection of borreliae of the next, emerging serotype. This explains the relapse phenomenon and the successive appearance of borreliae expressing different VMPs during the course of an untreated infection. These same VMPs are the principal TNF-a-inducing factors in LBRF. VMPs may differ in their potency as TNF inducers; they may also determine invasiveness of the borreliae (for example, into the central nervous system) and may affect virulence in other ways.

Pathology

The vast majority of spirochaetes are confined to the lumen of blood vessels, but tangled masses are also found in the characteristic splenic miliary abscesses (Fig. 2) (Plate 2) and infarcts as well as within the central nervous system adjacent to haemorrhages. Some strains of tick-borne borreliae can invade the CNS, aqueous humour, and other tissues. In LBRF, a perivascular histiocytic interstitial myocarditis, found in the majority of cases, may be responsible for conduction defects, arrhythmias, and myocardial failure resulting in sudden death. Splenic rupture with massive haemorrhage, cerebral haemorrhage (Plate 3), and hepatic failure are other causes of death. The liver shows hepatitis with patchy mid-zonal haemorrhages and necrosis (Fig. 3). There is meningitis and perisplenitis: most serosal cavities and surfaces of viscera are studded with petechial haemorrhages (Plate 4). Thrombi are occasionally found occluding small vessels, but the peripheral gangrene sometimes found in patients recovering from louse-borne typhus is not seen.

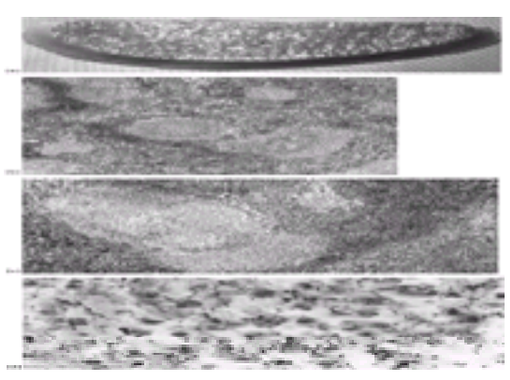

Fig. 2 Splenic miliary abscesses in louse-borne relapsing fever. (a) Section of spleen at autopsy (copyright DA Warrell). (b) Biliary microabscesses as seen under the microscope; 71 × (Armed Forces Institute of Pathology photograph, negative number 75–8838). (c) Microabscesses involve both follicles and extrafollicular areas of the spleen, the pale area of extrafollicular necrosis is clearly demarcated from the surrounding pulp; 145 × (Armed Forces Institute of Pathology photograph, negative number 77326). (d) Warthin–Starry stain showing tangled masses of spirochaetes at the periphery of an abscess; 2280 × (Armed Forces Institute of Pathology photograph, negative number 77317).

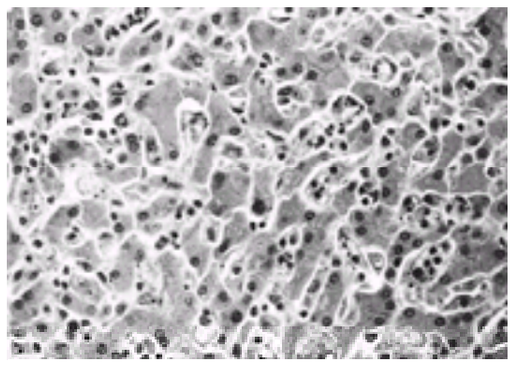

Fig. 3 Liver in louse-borne relapsing fever. Congestion with prominent Kupffer cells and lymphocytic and neutrophil infiltrate predominantly in the central and mid-zonal areas; 500 × (Armed Forces Institute of Pathology photograph, negative number 75–6523)

Clinical features

Poor, indigent, malnourished street-dwellers, beggars, and prisoners seem most likely to become infected, especially young men. Pregnant women appear to be specially susceptible to severe disease and abortions are frequent.

After an incubation period of 4 to 18 (average 7) days, the illness starts suddenly with rigors and a fever that mounts to nearly 40 °C in a few days. Early symptoms are headache, dizziness, nightmares, generalized aches and pains (especially affecting the lower back, knees, and elbows), anorexia, nausea, vomiting, and diarrhoea. Later there is upper abdominal pain, cough, and epistaxis. Patients are usually prostrated. Most are confused. Hepatic tenderness is the commonest sign (about 60 per cent). The liver is palpably enlarged in about 50 per cent of cases. Splenic tenderness and enlargement are slightly less common. Jaundice has been reported in between 10 and 80 per cent of cases. A petechial or ecchymotic rash is seen in between 10 and 60 per cent of cases: the lesions occur particularly on the trunk (Plate 5). Other sites of spontaneous bleeding include the nose in 25 per cent and less commonly the lungs, gastrointestinal tract, and conjunctivas (Plate 6) and retinas. Many patients have tender muscles. Meningism occurs in about 40 per cent of cases: other neurological features include cranial nerve lesions,

monoplegias, flaccid paraplegia, and focal convulsions attributable, perhaps, to cerebral haemorrhages.

Time course and relapses

In untreated cases of the louse-borne disease, the first attack of fever resolves by crisis in 4 to 10 (average 5) days, whereas the initial fever in tick-borne disease lasts only about 3 days. There follows an afebrile remission of 5 to 9 days, and then a series of up to five relapses in louse-borne disease and up to 13 in tick-borne disease (Fig. 4). No petechial rash occurs during the relapses, which are generally less severe than the initial attack but may be associated with iritis or iridocyclitis and severe epistaxis.

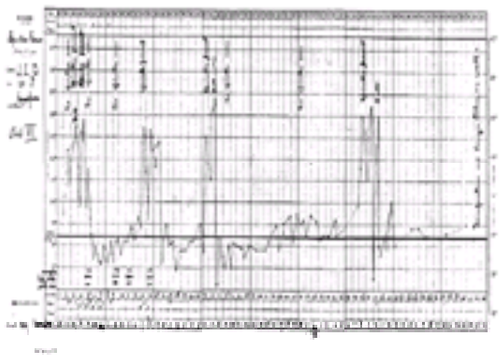

Fig. 4 Temperature chart of J Everett Dutton who, with JL Todd, discovered the transmission of tick-borne relapsing fever in the Congo. Dutton contracted tick-borne relapsing fever at the beginning of November 1904. He had relapses of fever and spirochaetaemia on the 7 and 16 December and the 8 January 1905. His death on 27 February 1905 has been attributed by some, but not by Todd, to relapsing fever (Dutton JE, Todd JL (1905). The nature of human tick-fever in the eastern part of the Congo Free State with notes on the distribution and bionomics of the tick. *Liverpool School of Tropical Medicine Memoir XVII*).

Differences between louse-borne and tick-borne relapsing fever

The tick-borne disease is generally milder and less drawn out. The incidence of some symptoms and signs in the two diseases appears strikingly different. For example, in some series of cases, only 7 per cent of patients with tick-borne relapsing fever were jaundiced and neurological signs were more common than in the louse-borne disease.

Severe manifestations

These include myocarditis that presents as acute pulmonary oedema, liver failure, and severe bleeding attributable to thrombocytopenia, liver damage, and disseminated intravascular coagulation. Dysentery, salmonellosis, typhoid, typhus, malaria, and tuberculosis have been described in association with relapsing fever.

The spontaneous crisis and Jarisch–Herxheimer reaction

Whether or not treatment is given, an attack of relapsing fever usually ends dramatically. About 1 h after intravenous tetracycline, or on about the fifth day of the untreated illness, the patient becomes restless and apprehensive and suddenly begins to have distressingly intense rigors that last between 10 and 30 min. The ensuing phenomena have features of a classical endotoxin reaction. During the initial chill phase, temperature, respiratory and pulse rates, and blood pressure rise sharply. Delirium, gastrointestinal symptoms, cough, and limb pains are associated. Some patients die of hyperpyrexia at the peak of fever. The flush phase, which lasts several hours, is characterized by profuse sweating, a fall in blood pressure, and a slow decline in temperature. Deaths during this phase follow intractable hypotension or the development of acute pulmonary oedema and are attributable to myocarditis. The classical J-HR is in syphilis. Milder reactions have been described in Lyme disease and leptospirosis (treated with penicillin), sodoku (arsenicals), *Brucella melitensis* (tetracycline), and even in meningococcal infections.

Laboratory findings

Spirochaete densities may exceed 500 000/mm^3 of blood. Other abnormalities include a moderate normochromic anaemia, neutrophil leucocytosis (with marked leucopenia during the spontaneous crisis or -HR), thrombocytopenia, mild coagulopathy, biochemical evidence of hepatocellular damage (raised levels of aminotransferases, alkaline phosphatase, direct and total bilirubin, low albumin) and mild renal impairment. The cerebrospinal fluid shows a polymorph/lymphocyte pleocytosis without visible spirochaetes.

ECG evidence of myocarditis includes prolonged Q–Tc, T-wave abnormalities, and ST-segment depression with transient acute right heart strain after the J-HR. Chest radiographs show pulmonary oedema in some cases.

Diagnosis

In febrile patients, spirochaetes can usually be demonstrated in thin or thick blood films stained with Giemsa or Wright's stain and counterstained for 10 to 30 min with 1 per cent crystal violet (Plate 1), by dark-field examination or the quantitative buffy-coat technique. Towards the end of the attack, during remissions, and particularly in children with tick-borne disease, spirochaetaemia may not be detectable. In these cases, blood or CSF can be injected intraperitoneally into young mice which will develop spirochaetaemia within 14 days. Serological methods are not generally used, but LBRF has been diagnosed by the detection of antibodies to glycerophosphodiesterase from *B. recurrentis*. The serum of patients with relapsing fever may give positive reactions with *Proteus* OXK, OX19, and OX2 and false-positive serological responses for syphilis in 5 to 10 per cent of cases.

Differential diagnosis

In a febrile patient with jaundice, petechial rash, bleeding, and hepatosplenomegaly, the differential diagnosis will include falciparum malaria, yellow fever and other viral haemorrhagic fevers, viral hepatitis, rickettsial infections (especially louse-borne typhus), and leptospirosis. The diagnosis can be quickly confirmed by examining a blood smear, but the possibility of a complicating infection, particularly typhoid, should not be forgotten.

Prognosis

The mortality in treated cases is less than 5 per cent. During major LBRF epidemics, mortalities of 40 per cent or higher have been reported. Deaths during relapses are most unusual: they occur only in the tick-borne disease.

Treatment

Antimicrobials

Although TBRF is usually milder than the louse-borne variety, it is more difficult to treat because spirochaetes persist in tissues, such as the central nervous system and eye, and produce relapses. Oral tetracycline, 500 mg every 6 h for 10 days is, however, effective. Oral erythromycin can be given to pregnant women (500 mg every 6 h for 10 days) and children (125–250 mg every 6 h for 10 days). In patients unable to swallow tablets, treatment can be initiated with 250 mg intravenous tetracycline hydrochloride or with 300 mg erythromycin lactobionate.

LBRF is readily cured with a single oral dose of 500 mg tetracycline or 500 mg erythromycin stearate. Few patients with severe louse-borne relapsing fever are able to

swallow the tablets without vomiting them up: a more reliable treatment is a single intravenous dose of 250 mg tetracycline hydrochloride or, for pregnant women and children, a single intravenous dose of 300 mg erythromycin lactobionate (children 10 mg/kg body weight). In mixed epidemics of LBRF and louse-borne typhus a single oral dose of 100 mg doxycycline has been effective.

Benzylpenicillin (300 000 units), procaine penicillin with benzylpenicillin (600 000 units), and procaine penicillin with aluminium monostearate (600 000 units), all by intramuscular injection, have been used; but they may fail to prevent relapses, and the long-acting preparations produce only slow clearance of spirochaetaemia. Chloramphenicol is effective in TBRF in a dose of 500 mg every 6 h for 10 days in adults, and 250 mg every 6 h for 10 days in older children; and in louse-borne relapsing fever in a single dose of 500 mg by mouth or intravenous injection in adults.

Jarisch–Herxheimer reaction

Antimicrobials have reduced the mortality of relapsing fevers from between 30 and 70 per cent to less than 5 per cent; however, drugs such as tetracycline, which generally rapidly eliminate spirochaetes from the blood and prevent relapses, usually induce a severe J-HR that may occasionally prove fatal. Clearly, in a disease with such a high natural mortality, treatment cannot be withheld, especially as severe spontaneous crises, which may also prove fatal, occur in a large proportion of louse-borne cases after the fifth day of fever. There is no evidence, however, that the shorter and more intense reaction following tetracycline is more dangerous than the more prolonged but apparently milder reaction following slow-release penicillin. Treatment with hydrocortisone, in doses up to 20 mg/kg, and paracetamol does not prevent the reaction but reduces peak temperatures, hastens the fall in temperature, and lessens the fall in blood pressure during the flush phase. Pretreatment with oral prednisolone can prevent the J-HR of early syphilis; but in LBRF, neither an oral dose of 3 mg/kg prednisolone given 18 h beforehand nor an infusion of 3.75 mg/kg betamethasone prevented the reaction to tetracycline treatment. However, meptazinol, an opioid antagonist with agonist properties, diminishes the reaction when given in a dose of 100 mg by intravenous injection. The discovery of an explosive release of TNFa, IL-6, and IL-8 just before the start of the J-HR prompted the testing of a polyclonal ovine Fab anti-TNFa antibody. Infused for 30 min before treatment with intramuscular penicillin, this antibody suppressed the J-HR.

Supportive treatment

Patients must be nursed in bed for at least 24 h after treatment to prevent postural hypotensive collapse and the precipitation of fatal cardiac arrhythmias. Hyperpyrexia should be prevented with antipyretics and vigorous fanning with tepid sponging. Although patients with acute LBRF have an expanded plasma volume, most are dehydrated and relatively hypovolaemic. Adults may need 4 or more litres of isotonic saline intravenously during the first 24 h. Infusion should be controlled by monitoring jugular venous, central venous, or pulmonary artery wedge pressures. Acute myocardial failure may develop, particularly during the flush phase of the J-HR or spontaneous crisis. This is signalled by a rise in central venous pressure above 15 cm H_2O; 1 mg digoxin should be given intravenously over 5 to 10 min. Because of the intense vasodilatation, diuretics may accentuate the circulatory failure by causing relative hypovolaemia. Oxygen should be given during the reaction, particularly in severe cases. Vitamin K should be given in all cases with prolonged prothrombin times. Heparin is not effective in controlling coagulopathy and should not be used. Complicating infections—typhoid, salmonellosis, bacillary dysentery, tuberculosis, typhus, and malaria—must be treated appropriately.

Control

No vaccines are available.

Delousing

Patients with LBRF are infectious until their louse-infested clothing is disinfected by heat, such as washing in water hotter than 60 °C, preferably followed by ironing. It is also recommended that infested people should wash their bodies with soap and a 1 per cent lysol (disinfectant) solution; however, most lice are attached to clothing not body hairs. These simple approaches are impracticable in epidemic situations and so insecticides are widely used for louse control. An insecticidal duster can be used to blow a 10 per cent DDT powder between the body and clothing. If DDT-resistant lice are present then dusts of 1 per cent malathion, 2 per cent temephos (Abate), 1 per cent propoxur, or 0.5 per cent permethrin can be used. Improved hygiene discourages lousiness, impregnation of clothing with a pyrethroid insecticide may give long-lasting protection against lice, and treated clothes may remain effective even after 6 to 8 washings. A study in Ethiopia demonstrated that treatment of cases of LBRF with antimicrobials was not effective in controlling an epidemic without the addition of vigorous delousing measures.

Tick control

Ticks should be searched for and removed. However, they usually feed for a short time and then detach and so are rarely found by the time the patient presents with tick-borne relapsing fever.

Ticks may be discouraged by insecticide-impregnated clothes or by applying repellents to the skin (for example, diethyltoluamide). Dwellings should be constructed with solid floors and walls to reduce tick infestation. Sleeping off the floor and under an insecticide-impregnated bed net can also reduce the risk of bites.

Tick control can be attempted by spraying buildings with insecticides such as pyrethroids, carbamates, and organophosphates and by reducing the numbers of rodent vectors.

Further reading

Bryceson ADM, *et al*. (1970). Louse-borne relapsing fever. A clinical and laboratory study of 62 cases in Ethiopia and a reconsideration of the literature. *Quarterly Journal of Medicine* **39**, 129–70.

Cutler SJ, *et al*. (1994). Successful *in-vitro* cultivation of *Borrelia duttonii* and its comparison with *Borrelia recurrentis*. *International Journal of Systematic Bacteriology* **49**, 1793–9.

Fekade D, *et al*. (1996). Prevention of the Jarisch–Herxheimer reactions by treatment with antibodies against tumor necrosis factor a. *New England Journal of Medicine* **335**, 311–15.

Felsenfeld O (1965). Borrelia, human relapsing fever and parasite–vector–host relationships. *Bacteriological Reviews* **29**, 46–74.

Felsenfeld O (1971). *Borrelia: strains, vectors, human and animal borreliosis*. WH Green, St Louis.

Negussie Y, *et al*. (1992). Detection of plasma tumor necrosis factor, interleukins-6 and -8 during the Jarisch–Herxheimer reaction of relapsing fever. *Journal of Experimental Medicine* **175**, 1207–12.

Scragg IG, Kwiatkowski D (2000). Structural characterization of the inflammatory moiety of a variable major lipoprotein of *Borrelis recurrentis*. *Journal of Biological Chemistry* **275**, 937–41.

Sundnes KO, Teklehaimanot A (1993). Epidemic of louse-borne relapsing fever in Ethiopia. *Lancet* **342**, 1213–15.

Trape JF, *et al*. (1991). Tick-borne borreliosis in West Africa. *Lancet* **337**, 473–5.

Udalova IA, *et al*. (2000). Direct evidence for involvement of NF-kappaB in transcriptional activation of tumor necrosis factor by a spirochetal lipoprotein. *Infection and Immunity* **68**, 5447–9.

Vidal V, *et al*. (1998). Variable major lipoprotein is a principal TNF-inducing factor of louse-borne relapsing fever. *Nature Medicine* **4**, 1416–20.

Warrell DA, *et al*. (1983). Pathophysiology and immunology of the Jarisch–Herxheimer like reaction in louse-borne relapsing fever: comparison of tetracycline and slow-release penicillin. *Journal of Infectious Diseases* **147**, 898–909.

Warrell DA, *et al*. (1970). Cardiorespiratory disturbance associated with infective fever in man: studies of Ethiopian louse-borne relapsing fever. *Clinical Science* **39**, 123–45.

7.11.31

Leptospirosis

George Watt

Leptospirosis is a worldwide zoonosis of the greatest public health importance in the tropics. Infection may be asymptomatic, but 5 to 15 per cent of cases are severe or fatal. Most cases go undiagnosed because symptoms and signs are often non-specific and serological confirmation is rarely available where most disease transmission occurs. Failure to diagnose leptospirosis is particularly unfortunate: severely ill patients often recover completely with prompt treatment but if therapy is delayed or not given, death or renal failure are likely to ensue.

Aetiology

The organism responsible is a tightly coiled spirochaete with an axial filament and hooked ends, 0.1 to 0.2 µm wide and 5 to 20 µm long. Leptospires are aerobic and travel with a corkscrew-like motion. Unstained organisms can be seen only by darkfield or phase-contrast microscopy. Silver staining is the method of choice for demonstrating leptospires in tissue specimens. The genus *Leptospira* contains two species: *Leptospira interrogans*, which is pathogenic, and *Leptospira biflexa*, which is saprophytic. Stable antigenic differences allow subclassification into serotypes, referred to in the literature as serovars (serovarieties). Antigens common to several serovars permit arrangement into broader serogroups. More than 200 serovars belonging to 23 serogroups have been identified for *L. interrogans*. Leptospirosis taxonomy is evolving and it has been proposed to establish five new species based on DNA relatedness.

Epidemiology

Measuring incidence by active surveillance confirms that leptospirosis is a surprisingly common disease. Antibody positivity rates of 37 per cent have been recorded in rural Belize and 23 per cent in Vietnam. More than 2527 human cases and 13 deaths were reported for the first 9 months of 1999 by the Ministry of Public Health in Thailand. Human leptospirosis is of significance in eastern and southern Europe, Australia, and New Zealand. In the United States, the disease is primarily of veterinary importance, with only 50 to 150 human cases reported annually.

Leptospires nest in the renal tubules of mammalian hosts and are shed in the urine. They can survive for several months in the environment under moist conditions, particularly in the presence of warmth (above 22 °C) and a neutral pH (pH 6.2 to 8.0). These conditions occur all year round in the tropics but only during the summer and autumn months in temperate climates. Roughly 160 animal species harbour organisms, but rodents are the most important reservoir. Carrier rates of over 50 per cent have been measured in Norway rats, which shed massive numbers of organisms for life without showing clinical illness. Some serovars appear to be preferentially adapted to select mammalian hosts. For example, the serovar icterohaemorrhagiae is primarily associated with the Norway rat, canicola with dogs, and pomona with swine and cattle. However, a particular host species may serve as a reservoir for one or more serovars and a particular serovar may be hosted by many different animal species.

The transmission of infection from animal to man usually occurs through contact with contaminated water or moist soil. Organisms enter man through abrasions of the skin or through the mucosal surface of the eye, mouth, nasopharynx, or oesophagus. Crowded Asian or Latin American cities that are flood-prone and have large rat populations provide ideal conditions for disease transmission. A outbreak in Nicaragua in 1995 and an urban epidemic in Salvador, Brazil in 1999 were associated with particularly heavy rains and flooding. Intense exposure to leptospires has been documented in rice, sugar cane, and rubber plantation workers. Less frequently, leptospirosis is acquired by direct contact with the blood, urine, or tissues of infected animals. Epidemiological patterns in the United States and United Kingdom have changed. Recreational exposure to fresh water (canoeing, sailing, water skiing) and animal contact at home have replaced occupational exposure as the chief source of disease.

Pathology and pathogenesis

Leptospires are disseminated by the blood and may be recovered from all organs within 48 h of entering the host. Leptospiraemia lasts from 4 to 7 days and ends when agglutinizing antibodies appear. Leptospires can persist for months in the kidneys and ocular tissue. Much of the pathogenesis of leptospirosis remains unexplained. There are only minor histopathological changes in the kidneys and livers of patients with marked functional impairment of these organs. Patients who survive severe leptospirosis have complete recovery of hepatic and renal function—consistent with the lack of structural damage to these organs.

Severely ill patients typically have marked leucocytosis but no leucocytic infiltrates in organs, a pattern produced by some toxins. Fatally infected animals and some human patients exhibit changes similar to those produced by the endotoxaemia of Gram-negative bacteraemia. An endotoxin-like substance is present in the cell wall of leptospires but lacks the ketodeoxyoctanoate of true endotoxin.

Kidney

Renal failure is the most common cause of death in leptospirosis. Leptospires are frequently found in human renal tissue, but their role in mediating kidney damage is unknown. Interstitial nephritis is found primarily in individuals who have survived until inflammation has had an opportunity to develop, but is frequently absent in patients with fulminant disease.

Impaired renal perfusion constitutes the fundamental nephropathic change. Oliguria is rapidly reversed by administration of intravenous fluid in many patients, suggesting that volume depletion is frequent. Hypovolaemia is multifactorial: insensible water loss, diarrhoea, vomiting, reduced fluid intake, and haemorrhage can all contribute. A defect in the kidney's ability to concentrate urine increases fluid loss while renal potassium wasting can lead to hypokalaemia. Widespread endothelial injury causes fluid to move from the intravascular to the extracellular space in some patients. Hypotension of cardiac origin is rare.

Liver

The pathogenesis of jaundice is unexplained; neither haemolytic anaemia nor hepatocellular necrosis are prominent features of leptospirosis. The most severe hepatic pathological changes are seen when organisms are difficult to demonstrate in tissue, suggesting subcellular toxic or metabolic insults.

Striated muscle

Myalgia is typical of early infection, and is presumably due to invasion of skeletal muscle by leptospires. Muscle biopsies in patients with early illness demonstrate vacuolation of the myofibrillar cytoplasm, loss of cellular detail, and fragmentation. Leptospiral antigen can be demonstrated by immunofluorescence within muscle tissue. Muscle pain resolves as antibody appears and organisms are cleared from the blood. Pathological changes are usually absent in muscle tissue from patients who have died, and myalgia is generally waning at the time of death.

Lungs

Localized or confluent haemorrhagic pneumonitis is the usual pulmonary finding, with petechial and ecchymotic haemorrhages noted throughout the lungs, pleura, and tracheobronchial tree. Early, life-threatening pulmonary haemorrhage has long been reported from Asia, and is now being increasingly recognized in Latin America. Necropsy findings include massive intra-alveolar haemorrhage with or without diffuse alveolar damage. Leptospires can be demonstrated in lung tissue.

Haemorrhage

A progressive severe haemorrhagic diathesis is a prominent feature of experimental leptospirosis. In humans, bleeding is generally restricted to the skin or mucosal surfaces, although occasionally massive gastrointestinal or pulmonary haemorrhage occurs. Coagulopathy and/or thrombocytopenia are common in leptospirosis but do not adequately explain bleeding. By exclusion, capillary damage is the postulated mechanism, and toxins have been suggested as the mediators of endothelial injury.

Meningitis

Organisms easily enter the cerebrospinal fluid during leptospiraemia, and this is thought to explain the high incidence of meningitis. However, signs of meningeal irritation are not due to the invasion of the meninges by leptospires, a process that elicits little reaction. Organisms are frequently isolated from cerebrospinal fluid that is otherwise normal and from individuals without clinically detectable involvement of the nervous system. Symptoms of meningitis coincide with the development of antibody and disappearance of leptospires from the blood and cerebrospinal fluid, suggesting an immunological mechanism. Pathological changes are minimal or absent, and the prognosis is excellent.

Heart

Focal haemorrhagic myocarditis has been reported, but hypovolaemia, electrolyte imbalance, and uraemia are more frequent causes of cardiac dysfunction. Minor electrocardiographic changes such as first-degree heart block are common and reversible.

Eye

The aqueous humour provides a protective environment for leptospires, which readily enter the anterior chamber of the eye during the leptospiraemic phase of the disease. There they can remain viable for months, despite the development of serum antibodies. Uveitis is common. Inflammation of the anterior uveal tract begins weeks or even months after the onset of disease and has been attributed to the persistence of organisms in the anterior chamber.

Clinical manifestations (see Table 1)

Subclinical infection is common and less than 10 per cent of symptomatic infections result in severe, icteric illness. Even relatively virulent serovars such as icterohaemorrhagiae lead more often to anicteric than to icteric disease. Old terms such as peapicker's disease, swineherd's disease, and canicola fever, which linked specific serotypes with distinct disease manifestations, are misleading and should be abandoned. The median incubation period is 10 days, with a range of 2 to 26 days. The duration of the incubation period has no prognostic significance. Once symptoms develop, they are said to follow a biphasic course: After an initial febrile illness, there is defervescence of fever and symptomatic improvement, followed by a second period of disease. However, a clear demarcation between the first and second stages is atypical of icteric leptospirosis and in mild cases the distinction can be unclear, or the second stage may never occur. The diagnostic usefulness of a history of a biphasic illness has been overemphasized. HIV coinfection does not seem to affect the clinical presentation of leptospirosis in the few coinfected patients described thus far.

Anicteric leptospirosis

Symptoms and signs

The disease typically begins with the abrupt onset of intense headache, fever, chills, and myalgia. Fever often exceeds 40 °C (103 °F) and is preceded by rigors. Muscle pain can be excruciating and occurs most commonly in the thighs, calves, lumbosacral region, and abdomen. Abdominal wall pain accompanied by palpation tenderness can mimic an acute surgical abdomen. Nausea, vomiting, diarrhoea, and sore throat are other frequent symptoms. Cough and chest pain figure prominently in reports of patients from Korea and China.

Conjunctival suffusion is a helpful diagnostic clue which usually appears 2 or 3 days after the onset of fever and involves the bulbar conjunctiva. Pus and serous secretions are absent, and there is no matting of the eyelashes and eyelids. Mild suffusion can easily be overlooked. Less common and less distinctive signs include pharyngeal injection, splenomegaly, hepatomegaly, lymphadenopathy, and skin lesions.

Within a week most patients become asymptomatic. After several days of apparent recovery, the illness resumes in some individuals. Manifestations of the second stage are more variable and mild than those of the initial illness and usually last 2 to 4 days. Leptospires disappear from the blood, cerebrospinal fluid, and tissues but appear in the urine. Serum antibody titres rise—hence the term 'immune' phase. Meningitis is the hallmark of this stage of leptospirosis. Pleocytosis of the cerebrospinal fluid can be demonstrated in 80 to 90 per cent of all patients during the second week of illness, although only about 50 per cent will have clinical signs and symptoms of meningitis. Meningeal signs can last several weeks but usually resolve within a day or two. Uveitis is a late manifestation of leptospirosis, generally seen 4 to 8 months after the illness has begun. The anterior uveal tract is most frequently affected, and pain, photophobia, and blurring of vision are the usual symptoms.

Laboratory findings

The white blood cell count varies but neutrophilia is usually found. Urinalysis may show proteinuria, pyuria, and microscopic haematuria. Enzyme markers of skeletal muscle damage, such as creatine kinase and aldolase, are elevated in the sera of 50 per cent of patients during the first week of illness. Chest radiographs from patients with pulmonary manifestations show a variety of abnormalities, but none is pathognomonic of leptospirosis. The most common finding is small, patchy, snowflake-like lesions in the periphery of the lung fields.

Icteric leptospirosis (Weil's disease)

This dramatic, life-threatening illness is characterized by jaundice, renal dysfunction, haemorrhagic manifestations, and a high mortality rate. Though jaundice is the hallmark of severe leptospirosis, fatalities do not occur because of liver failure. The degree of jaundice has no prognostic significance, but its presence or absence does—virtually all leptospirosis deaths occur in icteric patients. Icterus first appears between the fifth and ninth days of illness, reaches maximum intensity 4 or 5 days later, and continues for an average of 1 month. Hepatomegaly is found in the majority of patients and hepatic percussion tenderness is a reliable clinical marker of continuing disease activity. There is no residual liver dysfunction in survivors of Weil's disease, consistent with the absence of structural damage seen on pathological examination of this organ.

Bleeding is occasionally seen in anicteric cases but is most prevalent in severe disease. Purpura, petechiae, epistaxis, bleeding of the gums, and minor haemoptysis are the most common haemorrhagic manifestations, but deaths occur from subarachnoid haemorrhage and exsanguination from gastrointestinal bleeding.

Conjunctival haemorrhage is an extremely useful diagnostic finding, and when combined with scleral icterus and conjunctival suffusion, produces eye findings strongly suggestive of leptospirosis (see Fig. 1 and Plate 1). The frequency with which severe pulmonary haemorrhage complicates leptospirosis is variable, but is a cardinal feature of some outbreaks.

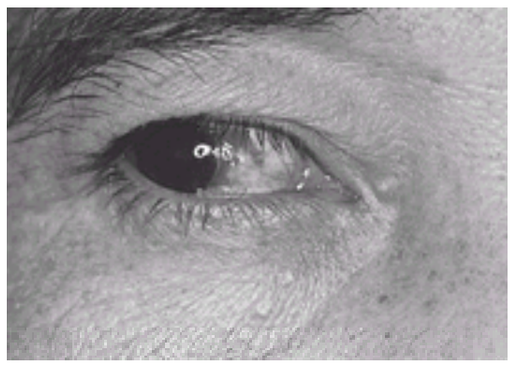

Fig. 1 Jaundice, haemorrhage, and conjunctival suffusion in acute leptospirosis. (See also Plate 1.)

Life-threatening renal failure is a complication of icteric disease, though all forms of leptospirosis may be associated with mild kidney dysfunction. Oliguria or anuria usually develop during the second week of illness, but may appear earlier. Complete anuria is a grave prognostic sign, often seen in patients who present late in the course of illness with frank uraemia and irreversible disease. Because renal failure develops very quickly in leptospirosis, symptoms and signs of uraemia are frequently encountered. Anorexia, vomiting, drowsiness, disorientation, and confusion are seen early and progress rapidly to convulsions, stupor, and coma in severe cases. Disturbances of consciousness in a patient with severe leptospirosis are usually due to uraemic encephalopathy, whereas in anicteric cases aseptic encephalitis is the usual cause. Renal function eventually returns to normal in survivors of Weil's disease, through detectable abnormalities may persist for several months.

Laboratory features of Weil's disease

Hyperbilirubinaemia results from increases in both conjugated (direct) and unconjugated (indirect) bilirubin, but elevations of the direct fraction predominate. Prolongations of the prothrombin time occur commonly but are easily corrected by the administration of vitamin K; modest elevations of serum alkaline phosphatase are typical. There is mild hepatocellular necrosis; greater than fivefold increases of transaminase (aminotransferase) levels are exceptional.

Jaundiced patients usually have leucocytosis in the range of 15 000 to 30 000 per mm ^{3}, and neutrophilia is constant. Anaemia is common and multifactorial; blood loss and azotaemia contribute frequently, intravascular haemolysis less often. Mild thrombocytopenia often occurs, but decreases in platelet count sufficient to be associated with bleeding are exceptional. The specific gravity of the urine is high. Hypokalaemia due to renal potassium wasting can occur.

Diagnosis

The diagnosis of leptospirosis is usually based on serology. The old commercially available microscopic slide agglutination test is being supplanted by a new generation of rapid serodiagnostic kits. Some enzyme-linked immunosorbent (EIA), agglutination, and immunofluorescent assays are very promising, although more data from prospective evaluations conducted in endemic areas are required. The need for practical, affordable diagnostic kits to be available in areas where leptospirosis is common cannot be overemphasized. The polymerase chain reaction and urine antigen detection are research tools which would be of greatest potential diagnostic value in patients who present early, before antibodies have reached detectable levels. The microscopic agglutination test is considered the serodiagnostic method of choice for leptospirosis, but its complexity limits its use to reference laboratories. Dilutions of patient sera are applied to a panel of live, pathogenic leptospires. The results are viewed under dark-field microscopy and expressed as the percentage of organisms cleared from the field by agglutination.

Isolation of leptospires from blood or cerebrospinal fluid is possible during the first 10 days of clinical illness, but specialized media are necessary. Serially diluted urine provides the highest yield. Unfortunately, culture results are only known 4 to 6 weeks later—too late to benefit hospitalized, severely ill patients.

Treatment

The approach to the patient with possible leptospirosis is summarized in Fig. 2. Placebo-controlled double-blind trials have proved that doxycycline benefits patients with early, mild leptospirosis, and that intravenous penicillin helps adults with severe, late disease. The outcome of severe, paediatric leptospirosis is also improved by penicillin therapy. Antibiotics should therefore be given to all patients with leptospirosis, regardless of age or when in their disease course they are seen. Doxycycline is given at doses of 100 mg orally twice a day for 1 week. Patients who are vomiting or are seriously ill require parenteral therapy. Intravenous penicillin G is administered as 1.5 million units every 6 h for 1 week. There is controversy regarding the occurrence of a Jarisch–Herxheimer reaction in leptospirosis. If present, it is much less prominent in leptospirosis than in other spiro-chaetal illnesses. The important practical consideration is that antibiotics should not be withheld because of the fear of a possible Jarisch–Herx-heimer reaction.

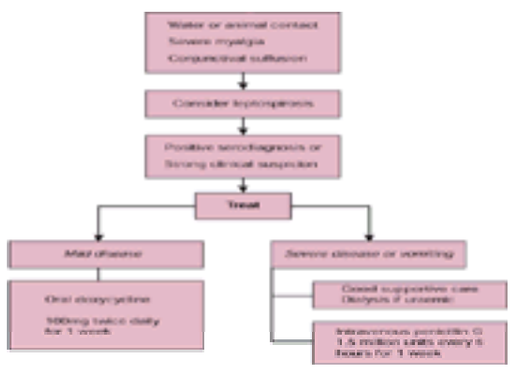

Fig. 2 Management of a febrile patient with possible leptospirosis.

The management of pulmonary haemorrhage often requires prompt intubation and mechanical ventilation. Respiratory support to maintain adequate tissue oxygenation is essential because in non-fatal cases complete recovery of pulmonary function can be achieved. Ensuring adequate renal perfusion prevents renal failure in the vast majority of oliguric individuals. Peritoneal dialysis is preferred to haemodialysis; frequent dialyses may be necessary because renal failure in leptospirosis is hypercatabolic.

Prevention

Doxycycline, 200 mg taken once a week, prevents infection by *L. interrogans*. Widespread use of doxycycline prophylaxis is not indicated, but it can benefit those who are at high risk for a short time, such as military personnel and certain agricultural workers.

Infection by leptospires confers only serovar-specific immunity; second attacks due to different serovars can occur. The efficacy and safety of human leptospiral

vaccines have yet to be conclusively demonstrated. Prevention of leptospirosis in the tropics is particularly difficult. The large animal reservoir of infection is impossible to eliminate, the occurrence of numerous serovars limits the usefulness of serovar-specific vaccine, and the wearing of protective clothing (e.g. rubber boots in rice fields) is both prohibitively expensive and impractical.

Prognosis

It is imperative to bring affordable tests to areas where leptospirosis is common because treatment (or lack of it) has a substantial impact on outcome. Atypical or mild cases are often confused with other entities such as aseptic meningitis, influenza, appendicitis, and gastroenteritis. Viral hepatitis is a common misdiagnosis in patients with Weil's disease. Leucocytosis, elevated serum bilirubin levels without marked transaminase elevations, and renal dysfunction are typical of leptospirosis but unusual in hepatitis. Malaria, typhoid fever, relapsing fever, scrub typhus, and Hantaan virus infection (haemorrhagic fever with renal syndrome) are important differential diagnoses in the tropics. Leptospirosis with prominent haemorrhagic manifestations is commonly misdiagnosed as dengue fever.

Further reading

Abdulkader RCRM *et al*. (1996). Peculiar electrolytic and hormonal abnormalities in acute renal failure due to leptospirosis. *American Journal of Tropical Medicine and Hygiene* **54**, 1–6.

Ko AI *et al*. (1999). Urban epidemic of severe leptospirosis in Brazil. *The Lancet* **354**, 820–5.

Marott PC *et al*. (1997). Outcome of leptospirosis in children. *American Journal of Tropical Medicine and Hygiene* **56**, 307–10.

Nicodemo AC *et al*. (1997). Lung lesions in human leptospirosis: microscopic, immunohistochemical, and ultrastructural features related to thrombocytopenia. *American Journal of Tropical Medicine and Hygiene* **56**, 181–7.

Sitprija V *et al*. (1980). Pathogenesis of renal disease in leptospirosis: clinical and experimental studies. *Kidney International* **17**, 827–36.

Watt G *et al*. (1988). Placebo controlled trial of intravenous penicillin for severe and late leptospirosis. *The Lancet* **1**, 433–5.

Zaki SR, Shieh WJ, the Epidemic Working Group. (1996). Leptospirosis associated with outbreak of acute febrile illness and pulmonary haemorrhage, Nicaragua. *The Lancet* **347**, 535–6.

7.11.32 Non-venereal endemic treponemoses: yaws, endemic syphilis (bejel), and pinta

P. L. Perine and D. A. Warrell

The endemic treponematoses are chronic, granulomatous diseases caused by spirochaetes belonging to the genus *Treponema*. Yaws occurs mainly in children living in rural areas in warm, humid climates in tropical countries. About 10 per cent of untreated cases develop late, disfiguring or crippling lesions of skin, bone, and cartilage.

Aetiology

Yaws is caused by *Treponema pallidum* ssp. *pertenue*, a spirochaete that is morphologically identical to *T. pallidum* ssp. *pallidum* (the cause of venereal syphilis) and *T. pallidum* spp. *endemicum* (the cause of non-venereal syphilis (bejel)), and to *T. carateum* (the cause of pinta). These treponemes share common antigens, so that infection by one species produces varying degrees of cross-immunity to the others. No serological test can differentiate the antibodies produced, and none of these organisms grows *in vitro*. The only means of differentiating yaws, syphilis, and non-venereal syphilis are their epidemiological characteristics and the pattern of infection produced in humans and experimentally infected laboratory animals (Table 1). Pathogenic treponemes can be differentiated at the molecular level by differences in the 5'- and 3'-flanking regions of their 15-kDa lipoprotein genes (*tpp*15).

The treponemes of yaws, syphilis, and pinta are fragile and readily killed by exposure to atmospheric oxygen, drying, mild detergents, or antiseptics. They prefer temperatures below 37 °C, which may explain their predilection for the skin and bones of the extremities. These organisms cannot penetrate intact skin, and gain entry to the body through small abrasions and lacerations.

Epidemiology (Table 1)

Yaws is transmitted by direct contact with an infectious lesion or by fingers contaminated with lesion exudate. It is enhanced by a crowded environment with poor sanitation and personal hygiene. The disease is usually acquired in childhood between the ages of 5 and 15. In endemic areas more than 80 per cent of the population are infected. In humid, warm environments the early lesion tends to proliferate and teems with spirochaetes, thus increasing the infectious reservoir; whereas in dry, arid climates or seasons the reverse is true.

There was a precipitous decrease in cases of yaws and other endemic treponematoses following mass penicillin-treatment campaigns during the 1950s and 1960s sponsored by the World Health Organization. An estimated 152 million people were examined and 46.1 million clinical cases, latent infections, and contacts were treated. The yaws reservoir was greatly reduced in West and Central Africa, Central and South America, and Oceania. However, over the past decade, yaws has been resurgent in the rural populations of Ecuador, the Ivory Coast, Ghana, Togo, Benin, Zaire, the Central African Republic, and Ethiopia in Africa, and in the island nations in the Pacific. Several nations initiated new campaigns of mass treatment during the 1980s.

Some African nations, such as Nigeria, previously rendered yaws-free by mass treatment campaigns, have also experienced a sharp rise in the incidence of venereal syphilis, perhaps representing a decline of herd immunity to yaws, and thereby to syphilis.

Endemic syphilis is transmitted by non-venereal contact among children. In contrast to yaws, transmission of infection by contaminated drinking vessels may be more common than by direct contact with infectious lesions. The disease tends to be familial, with spread of infection from children to adults rather than to the community in general. Endemic syphilis lesions are virtually indistinguishable from early yaws, and the two diseases may occur at different times in the same population but not in the same person. Venereal syphilis can be acquired by children through social contact with adults suffering from venereal syphilis, and then be spread by non-venereal, person-to-person contact if levels of sanitation and personal hygiene are low.

The Sahelian nations of Mauritania, Mali, Niger, Burkina Faso, and Senegal have reported dramatic increases in the number of cases of endemic syphilis. In Naimey (Niger), seroprevalence was 12 per cent among children under 5 years of age. The disease is also prevalent among the nomadic tribes of the Arabian peninsula, where late complications such as osteoperiostitis predominate.

Several variants of endemic syphilis are recognized by their geographical distribution: bejel of the Eastern Mediterranean, North Africa, and Niger; and njovera or dichuchwa of Africa; Bejel is the only type of endemic syphilis still prevalent. It is found in mainly seminomadic people such as the Tuareg living in the Saharan regions of Africa. Pinta is found only in remote parts of Central and South America, principally in the semiarid region of the Tepalcatepec Basin of southern Mexico and focal areas of Colombia, Peru, Ecuador, and Venezuela.

Pathogenesis

The lesion of yaws and the other treponematoses are due largely to the host's immune response to the treponeme. None of these treponemes carries or produces toxic substances. They have the ability to invade living cells without causing apparent injury. Cell destruction and tissue damage are probably due to the action of immune cells that injure normal tissue in the process of killing treponemes.

Host immunity reaches its highest level after several months of infection, just before disseminated lesions heal and latency begins. Thereafter the host is immune to reinfection and is not contagious, but since not all treponemes are killed, infectious lesions may reappear as immunity wanes over time. Most patients with yaws experience two or three infectious relapses during the first 5 years of infection.

In venereal, and possibly endemic, syphilis, infection is systemic and late lesions may develop in any organ or tissue of the body. In yaws, *T. pertenue* produce lesions only in skin and osseous tissue, although it is certain that periodically the organism spreads systemically; *T carateum* resides only in the skin. This peculiar tissue tropism is unexplained. It is probably an inherent property of the treponeme, acting in contact with climatic factors.

Clinical features

Like venereal syphilis, the clinical course of yaws and endemic syphilis have primary, secondary, and tertiary or late stages, separated by quiescent or latent periods.

The initial lesion in yaws usually appears on the extremities after an incubation period of 3 to 5 weeks. Characteristically it is a papule; a painless lesion that appears at the site of infection, enlarges, forming a raspberry-like ('framboesia'), vegetative lesion called a papilloma. The papilloma is round to oval, elevated and not indurated, ranging in size from 1 to 3 cm in diameter. The surface teems with spirochaetes and is often covered by a thin yellow crust, which is easily removed. The papilloma may ulcerate as it enlarges and becomes secondarily infected with other micro-organisms. Lymph nodes draining the initial lesion may enlarge and become tender, but systemic symptoms are rare.

Secondary or disseminated papillomas appear after 2 to 6 months, often without an intervening latent period, on the skin of moist areas such as the axillas, joint flexures, genitalia, and the gluteal cleft (Fig. 1). They also occur on the soles and palms and, because they are tender, may interfere with gait and use of the hands. Papillomas in different stages of development persist for 6 to 8 months and heal without scars unless they become secondarily infected. Despite the size and number

of lesions, children with generalized papillomas experience little discomfort or other constitutional symptoms.

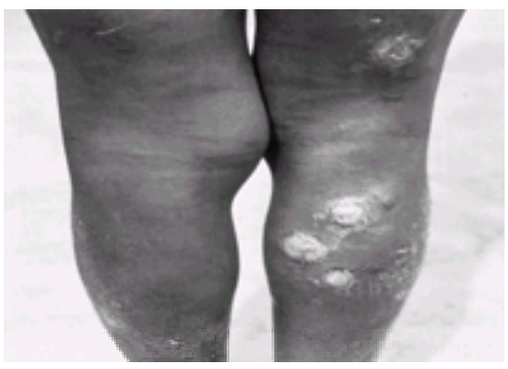

Fig. 1 Early ulceropapillomatous yaws. (Copyright PL Perine.)

Slightly raised, scaly, pigmented, macular yaws lesions measuring between 1 and 4 cm in diameter commonly occur when the climate is dry and arid. These lesions have the same distribution as papillomas and may appear together with lesions of different morphology in the same patient (maculopapular yaws).

The periosteum and osseous tissue of the bones of the extremities are frequently inflamed during early yaws, causing swelling, night-pain, and tenderness. There is dactylitis of the proximal phalanges. Painful osteoperiostitis of the legs, affecting mainly the tibias and fibulas, is especially common. Hypertrophic osteitis of the maxilla, either side of the bridge of the nose, can cause grotesque swellings ('goundo'). Scaly, tender, hyperkeratotic lesions of the palms and soles also occur and may be incapacitating. Hyperkeratotic and bone lesions are not contagious, and macular lesions are only minimally so.

One or more relapses of secondary-type lesions usually occur during the first 5 years of infection, each separated by a period of latency. Late yaws' lesions occur thereafter in about 10 per cent of untreated cases.

Late yaws' lesions are not infectious because they contain few treponemes. Cutaneous plaques produce atrophic scars; subcutaneous, granulomatous nodules erode skin and produce deep ulcers that destroy underlying tissue and disfigure. Hyperkeratotic palmar and plantar yaws are incapacitating and often prevent use of the hands, or the ability to walk normally. The weight is placed on the sides of the feet, which produces a gait much like that of a crab ('crab' yaws; Fig. 2).

Fig. 2 Planter papillomas with hyperkeratotic, macular, early plantar yaws ('crab yaws'); these lesions are painful. (Copyright PL Perine.)

The granulomas of late yaws have a histological appearance like the gummas of syphilis. These proliferative lesions may involve the palate and destroy the soft tissues of the nose, causing a terrible disfiguration called gangosa (Fig. 3). Gummatous periostitis of the skull, fingers, and long bones is erosive and often retards or stops growth. Active periostitis is occasionally found in young and middle-aged adults who had yaws in childhood.

Fig. 3 Gangosa (rhinopharyngitis mutilans) of endemic syphilis and yaws in an adolescent child. (Copyright PL Perine.)

The initial lesions of endemic syphilis usually appear at the mucocutaneous borders of the mouth or on the oral mucous membranes (mucous patches) as the result of transmission by contaminated drinking vessels. Late ulceronodules and osteoperiostitis are seen in late endemic syphilis, but cardiovascular and neurological complications are extremely rare.

In pinta, the initial papule appears on the skin of the extremities and enlarges slowly over a period of several weeks or months to form an erythematous plaque. Satellite papules form at the edge of the lesion and undergo a similar type of evolution. The plaques coalesce to form violaceous, pigmented plaques that, in several years, slowly depigment from lighter shades of blue to white, leaving atrophic depigmented scars.

Ulceronodular skin lesions of yaws and endemic syphilis resemble tropical ulcers. Yaws' lesions are not as painful, necrotic, nor as deep as tropical ulcers, which are usually singular and restricted to the lower one-third of the leg.

Plantar warts are frequently confused with plantar papillomas of yaws, and both conditions may occur in the same patient.

Diagnosis

The diagnosis of yaws is made by a combination of clinical assessment, of positive dark-ground examination of lesions, and of reactive serological tests for syphilis.

The diagnosis of early yaws, or endemic syphilis, is not difficult in endemic areas where the disease is familiar. The most difficult diagnostic problem arises when a

person who had yaws as a child emigrates to an area of the world where the disease never existed. Such a person usually has reactive serological tests for syphilis and may have a few atrophic scars suggestive of earlier infection. What are the chances that this patient has or has had venereal syphilis? Should he be treated for latent yaws or syphilis?

The patient's social and medical history should be carefully reviewed. Clinical findings suggestive of old yaws (scars, inactive tibial periostitis), and the absence of signs of congenital and venereal syphilis support the diagnosis of inactive or treated yaws.

If the patient has a reagin titre of less than 1:8 dilutions, they probably do not have active latent yaws or syphilis. If they received at least one therapeutic dose of long-acting penicillin in their native country during a yaws campaign, they require no further treatment. On the other hand, if the patient is a contact of a case of infectious venereal syphilis, they should be treated as being potentially infected with syphilis, because *T.p. pallidum* occasionally superinfects people who had yaws as children. If treatment is given, the patient should receive a certificate stating the drug and dosage used and the results of their serological tests to prevent unnecessary future treatment.

Treatment and prevention

Long-acting benzylpenicillin given by intramuscular injection is the recommended treatment for all the endemic treponematoses. The preparation used in previous mass treatment campaigns was penicillin aluminium monostearate (**PAM**), but benzathine penicillin is currently recommended because it is longer acting and more readily available than is PAM. Active infections and non-infectious cases should be given 1.2 mega units in a single intramuscular injection; children under 10 years of age receive 0.6 mega units. Patients allergic to penicillin may be given tetracycline or erythromycin, 500 mg by mouth four-times daily for 2 weeks; children under 10 years of age should be given erythromycin in dosages adjusted for their age. Treatment failures have been reported in Papua New Guinea.

Prevention of yaws in a community requires elimination of the reservoir of infection, often by treating the entire population with penicillin.

Further reading

Centurion-Lara A, *et al.* (1998). The flanking region sequences of the 15-kDa lipoprotein gene differentiate pathogenic treponemes. *Journal of Infectious Diseases* **177**, 1036–40.

Engelkens HJ, Vuzevski VD, Stolz E (1999). Non-venereal treponematoses in tropical countries. *Clinics in Dermatology* **17**, 105–6, 143–52.

Guthe T (1969). Clinical, serological and epidemiological features of framboesia tropica (yaws) and its control in rural communities. *Acta Dermatologica-Venerologia*, Stockholm, **49**, 343–68.

Hackett CJ, Loewenthal LJA (1960). *Differential diagnosis of yaws*. World Health Organization, Geneva.

Paris JL (2000). Treponemal infections in the pediatric population. *Clinics in Dermatology* **18**, 687–700.

Perine PL, *et al.* (1984). *Handbook of endemic treponematoses*. World Health Organization, Geneva.

Walker SL, Hay RJ (2000). Yaws—a review of the last 50 years. *International Journal of Dermatology* **39**, 258–60.

7.11.33

Syphilis

D. J. M. Wright and S. E. Jones

Definition

Venereal syphilis is a systemic, contagious disease of great chronicity, caused by *Treponema pallidum*, and capable of congenital transmission. The natural host is man. The incubation period is around 3 weeks, at the end of which a primary sore develops at the site of inoculation, usually on the genitalia, associated with regional lymphadenitis. In most patients, this is followed by the secondary bacteraemic stage characterized by a symmetrical rash, generalized lymphadenopathy, and other lesions. After a latent period of many years, in 40 per cent of cases a destructive late stage develops involving the skin, mucous membranes, skeleton, central nervous system, eyes, hearing, and, especially, the aorta. Any of these stages may be absent or inapparent. Venereal syphilis, unlike non-venereal syphilis, is distributed world-wide.

Bacterial taxonomy

T. pallidum is a bacterium which causes venereal syphilis and the non-venereal, endemic childhood syphilis, bejel and njovera. Other pathogenic treponemes include *T. pertenue* (yaws) and *T. carateum* (pinta). Although these spirochaetes produce distinct diseases, molecular techniques have not yet been able to demonstrate consistent differences in their genomic DNA.

There are a number of non-pathogenic treponemes (*T. denticola* etc.) in the mouth which are difficult to distinguish from *T. pallidum*. For that reason dark-field examination of samples from mouth lesions should be avoided because of the danger of misdiagnosis. Other treponemes of low pathogenicity (e.g. *T. balanitides*) reside in the genital tract and, together with fusiform bacilli, can under anaerobic conditions superinfect genital lesions producing 'fusospirochaetosis'.

The completed sequence of the 1.14 Mb genome of *T. pallidum* (website 1) revealed that this parasitic spirochaete had few sets of enzymes for basic metabolic processes—as expected, since *T. pallidum* has remained unculturable *in vitro*. Transporter systems (for amino acids, carbohydrates, and cations) comprise 5 per cent of the genome, and the lack of genes coding for protection from oxygen-derived free radicals indicates that *T. pallidum* will only survive in oxygen-depleted conditions.

T. pallidum is a delicate, motile spiral organism, 6 to 15 µm long and 0. 15 µm wide which renders it below the level of resolution of light microscopy and hence the need for dark-field or phase contrast illumination. It has an outer membrane, an electron-dense layer, and a cytoplasmic membrane. As with other bacteria, the cell wall has a trilaminar structure, the inner membrane constituting of a cytoplasmic membrane, while between the outer two layers there are axial filaments, structurally analogous to bacterial flagella, which wind around the axis of the organism and may be responsible for the motility of *T. pallidum*. All treponemes have not more than nine axial filaments, which distinguishes treponemes from borrelia (Table 1). *T. pallidum* has the unique ability of being able to bend in the middle to form a V-shape, if suspended in a medium of low viscosity. Motility does not necessarily indicate viability as mobile *T. pallidum* have been observed after they have been retained for 90 days in capillary tubes. *T. pallidum* may remain infective for up to a week in 'survival media' and, depending on the nature of the media, show limited multiplication; however, attempts at reproducible subculture of the microbe have to date been unsuccessful. Low concentrations of oxygen (between 3 and 5 per cent) may enhance survival. In practice, *T. pallidum* is propagated in rabbits.

Experimental inoculation of *T. pallidum* into man or animals shows that the organism divides every 30 to 32 h, suggesting approximate infective dose are between 10^6 and 10^7 organisms and an average incubation period of 3 weeks. Peptidoglycan in the inner layer of the bacterial membrane accounts for its susceptibility to penicillin. The phospholipids in the outer membrane, of which cardiolipin is the most prominent hapten, provides the antigenic basis for the synthetically substituted VDRL (Venereal Disease Reference Laboratory (test)) used as a serological test for syphilis. Flagella antigen reacts non-specifically with antibodies found in most known sera.

Origin of syphilis

Clinical differences between treponematoses have been explained as an adaptation of the organism to changing climatic factors, especially humidity and temperature, and with improvement in hygiene, the wearing of clothes, and less frequent intimate contacts between children. Yaws is found throughout the tropical belt, while pinta was forced to retreat into remote indigenous communities in South and Central America. Non-venereal childhood syphilis, such as bejel and similar conditions, was formerly found in more temperate climes including the Middle East, Yugoslavia, British Isles, Scandinavia, and South Africa. The lack of congenital transmission of these venereal treponematoses arose because they were essentially childhood infections and by the time these children were old enough to have their own offspring, the disease had become non-infectious. The treponeme causing venereal syphilis was perhaps an adaptation to people wearing clothes, when it was obliged to seek shelter in the protected, warm, and moist regions of the genitalia, so becoming sexually transmitted. It spread throughout the world as an adult disease and, because there appears to be no solid cross-immunity, may coexist with non-venereal treponematoses.

This adaptive theory fails to explain why *T. pallidum, sensu stricto*, unlike the other treponematoses, involves the central nervous system, aorta, and visceral organs and why, at the end of the fifteenth century, the virulent venereal form swept through Europe and Asia, eventually to become the milder modern syphilis. The alternative Columbian theory suggests that Columbus introduced this new disease from the Caribbean islands.

There are no definite descriptions of syphilis before this time. The finding of skeletons with long bone lesions compatible with syphilis, centuries before the fifteenth century epidemic of syphilis, remains speculative, though Boylston (reviewed by Morton (2001)) has reported syphilitic changes in the bones of pre-1450 skeletons found in Hull. The lack of such skeletons from America makes the relationship with the coincidental discovery of America and the advent of 'new world' syphilis even

more doubtful.

The recognition of the contagious nature of the disease was recorded in 1530 by Fracastro. Klebs ultimately proved the infectivity of syphilis by reproducing syphilitic lesions by inoculating of syphilitic tissue into rabbits. The use of a prolonged Giemsa stain, allowed Schaudinn finally to identify the treponemes in 1905.

Epidemiology (see also Chapter 21.1)

Transmission

Sexual transmission is the rule in adult patients. The untreated patient remains infective for 4 years after acquiring the infection. Asexual transmission by close contact with an open lesion of early acquired or congenital syphilis is rare, as is direct blood transfusion with blood from an infectious individual or contact with infected fomites. Congenital syphilis still remains a problem, except in Northern Europe.

Incidence

There has been a steady decline in the incidence of syphilis in the West since the 1850s, interrupted only by major wars. Since 1940, there has been a 99 per cent drop in admissions of general paresis of the insane and congenital syphilis in the United States, with similar trends in the United Kingdom and Europe. There has also been a sharp reduction in all other forms of late syphilis. Gumma have almost disappeared. Early syphilis has not declined to the same degree since the Second World War. In the United States, syphilis reached a low in 1956, with 6576 reported cases but by 1992 had risen to 83 902 acquired cases. The introduction of health measures is reflected in the recent fall in the number of cases. A similar pattern is also found in most European countries. During the 1970s and 1980s, there was a steady increase in the number of cases of early syphilis. In the Russian Federation, where there has been a breakdown in public health, up to 1 per cent of the population has been affected by syphilis, especially around St Petersburg and Moscow. In 1980, 58 per cent of cases of syphilis in the United Kingdom were in homosexuals.

Since that time, the United Kingdom and American rates have diverged. The appearance of AIDS and the national programmes for 'safe sex' has resulted in the annual number of cases of infectious syphilis falling to 337 in the United Kingdom in 1993, most of the recent infections being acquired heterosexually abroad. In the United States, however, despite the fall in the number of homosexual males infected, the number of cases of syphilis has continued to rise, especially in the underprivileged Afro-American and Hispanic community and among HIV-infected drug abusers. In 1992, there were approximately 34 000 cases of primary and secondary syphilis, and 3850 cases of congenital syphilis in children under the age of 2 years, compared with 1986 when only 57 such cases were recorded in the United States. The failure of the Public Health Service in the United States to cope is reflected in the resurgence of congenital syphilis. Following the public health initiative in the United States, started in 1997, the number of cases of primary and secondary syphilis reported for 2000 was 5979, while congenital syphilis had fallen to 529 cases. The comparative number of cases of syphilis seen in the genitourinary medicine clinics in England and Wales in 2000 was 328. In other parts of the world, notably in the Far East, infected prostitutes may play a central role in the spread of early syphilis. Estimates by the World Health Organization suggest that there are 10 to 20 million cases of syphilis each year.

The changing clinical presentation of syphilis

There is some clinical evidence that syphilis is becoming milder and less typical. This has been especially noted in neurosyphilis and the virtual disappearance of the gumma. The widespread use of antibiotics for unrelated conditions may be responsible. This is supported by finding that meningovascular syphilis has not shown the dramatic decrease of general paresis of the insane and tabes dorsalis, possibly because the last two conditions take many more years to develop, giving cumulative chances of antibiotics being given. It is also possible that the disease is tending to become milder and less typical as a result of 'natural' changes which appear to have started long before the antibiotic era. For whatever reasons, syphilis is apparently becoming clinically less clear-cut . Its exclusion by serology and other tests becomes more important in patients attending the dermatologist, neurologist, ophthalmologist, the ENT specialist, or cardiologist with conditions of uncertain pedigree.

The advent of AIDS has led to a re-examination of the progression and manifestations of concomitant syphilis. Although a variety of unusual syphilitic rashes, in particular more ulcerating multiple chancres and florid secondary rashes, have been described in association with HIV infections, all of them are recorded in the older literature. The suggestions that there might be an increase in syphilitic meningovascular relapse in patients with HIV infections, again may reflect the natural history of syphilis, since approximately 20 per cent of patients with early syphilis have a pleocytosis in the cerebrospinal fluid. If these patients are untreated, about one-fifth develop neurosyphilis. The high prevalence of syphilis in HIV patients leads to an apparent, rather than a real, increase in syphilis complications. Holtom, in California, found that in patients partially treated for syphilis with concomitant HIV infections, about 9 per cent developed a pleocytosis and 1 per cent then developed neurological disease. Unlike mycobacterial infections where unusual presentations of tuberculosis occur in patients with AIDS, the presentation of early neurosyphilis seems to be characteristic of the disease. This is possibly because the vasculitis of syphilis is not due to the cellular immune response but to the adherence of the spirochaete to the endothelial layer of the blood vessel. The blood vessel first becomes more permeable and subsequently there is proliferation of this layer. Simultaneously, the spirochaete induces a cellular infiltrate. These changes lead to endarteritis, which is the hallmark of the disease. What is true is that syphilitic relapses do not occur despite the potential for the persistence of spirochaetes (see below) and that benzathine penicillin G is less effective in eliminating spirochaetes (see below) in patients with altered immunity. Any persistence of cerebrospinal fluid pleocytosis in the neurosyphilitic patient with HIV after adequate penicillin treatment should lead to an investigation of causes of meningitis other than syphilis. However, doubt has been cast as to whether eradication of spirochaetes always occurs in the non-immunocompromised patient. Wilner and Brody found that one-third of the patients with a syphilitic encephalitis (general paresis of the insane) who had been followed for 30 years, developed neurological signs at the end of the period, despite having had 'adequate' penicillin therapy. Only 7 per cent of the non-syphilitic control group of demented patients, evinced new neurological signs. These were presumably due to cerebrovascular degenerative disease. Even in the pre-AIDS era, Rothenberg, when he reviewed the efficacy of penicillin in the treatment of neurosyphilis, found that it was not always effective. Lastly, laboratory experiments show that it is more likely that syphilis has a deleterious effect on the progression of AIDS rather than the reverse.

Sex and race

Early syphilis is less florid in women than men and is almost asymptomatic during pregnancy. Cardiovascular syphilis is at least twice as common in men than women, where it is more severe and appears earlier. Neurosyphilis is always more common in men than women. The reasons for these differences are not known.

Caucasians suffer more commonly from neurosyphilis than black Africans and they in turn are much more prone to develop cardiovascular syphilis than Caucasians.

Infectivity

The estimated figure for infectivity varies but is commonly assumed to be around 50 per cent for both homosexual and heterosexual patients. After a single exposure the figure is nearer 25 per cent.

Some control measures

The main reason for increased case identification is the more intensive use of serological tests for syphilis. Another valuable control measure is contact tracing, which varies greatly in different countries, but should be standard practice everywhere. Its use across international borders should be developed with proper safeguards to preserve confidentiality. This has proved especially useful in tracing networks of infection seen in the United Kingdom in recent years, such as the mini-outbreaks in Bristol (1999) and Manchester (2000). The application of social network analysis often traces more sexual contacts than direct partner notifications. Other measures which should prove valuable are the education of the young without inducing anxiety, information about sexually transmitted diseases on internet sex clubs, the education of doctors, and the encouragement of regular check-ups of high-risk individuals, such as homosexual men and prostitutes. A more controversial suggestion is to treat contacts of infectious syphilis epidemiologically in certain situations, for example promiscuous individuals, known defaulters, and those who may infect their regular consort if not treated. In England and Wales, during the years 1995 to 2000, 425 such contacts were so treated in genitourinary (STD) clinics (PHLS and DHSS Report, 2001). These measures can be expected to uncover up to 75 per cent of all cases of syphilis.

Persistence of treponemal forms

Persistence of *T. pallidum*-like forms in the cerebrospinal fluid, aqueous humour, lymph nodes, and other tissues in penicillin-treated patients with late or late latent

syphilis has been reported from several centres.

The natural course of untreated syphilis

T. pallidum penetrates the abraded skin and intact mucous membrane. Within hours it has disseminated via the bloodstream and lymphatics and is beyond any effective local treatment. The incubation period is traditionally given as 9 to 90 days but in practice it is around 3 weeks (range, 2 to 6 weeks). The time depends on the size of the inoculum, sexual practice, and hygienic measures. A single treponeme leads to the longest incubation period. The primary lesion develops at the site of contact and heals in 2 to 6 weeks. In a proportion of patients, a secondary stage appears 6 weeks after the primary lesion has healed but there may be an overlap of the healing primary and the onset of the secondary stage. In some cases, the period between these stages can be prolonged to several months. The main characteristic of the secondary stage is a generalized, symmetrical, painless, and non-irritating rash. In about 20 per cent of cases, infectious relapses occur during the following year (range, 1 to 4 years). In the rest, the latent symptomatic period follows and may persist for life in at least 60 per cent . In 30 to 40 per cent, a third, late destructive stage develops. Its more benign form involves only the skin, mucous membranes, and bones. In the serious form, the central nervous system, aorta, and other internal organs are affected. The major events are shown in Fig. 1 and Fig. 2.

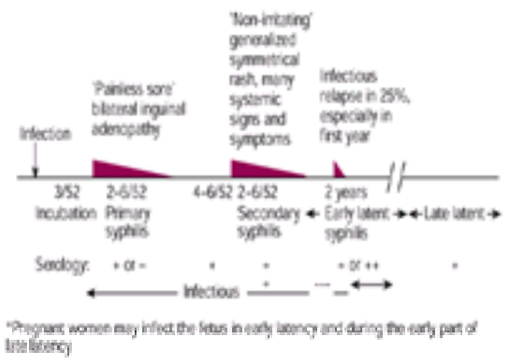

Fig. 1 The course of untreated, early-acquired syphilis.

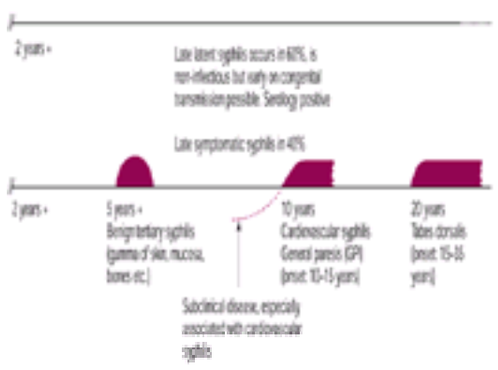

Fig. 2 The course of untreated, late-acquired syphilis. Asymptomatic neurosyphilis is present in 20 per cent and 20 per cent of these develop clinical neurosyphilis. Cardiovascular syphilis starts subclinically many years earlier and when clinically apparent, it is in fact in an advanced state. Prognosis: gumma heals spontaneously in a few years. Cardiovascular syphilis is usually fatal without treatment . Neurosyphilis: general paresis has a poor prognosis without treatment, meningovascular syphilis commonly responds well to penicillin, tabes progresses slowly but penicillin has no obvious influence. Overall mortality of untreated syphilis: 20 to 30 per cent.

Clinical features

Primary syphilis

The first sign is a small, painless papule which rapidly ulcerates. The ulcer (chancre) is usually solitary, round or oval, painless, and often indurated (Fig. 3). It is surrounded by a bright red margin. It is not usually secondarily infected, a feature of all open syphilitic lesions of any stage. *T. pallidum* can be demonstrated in the serum from the sore which is easily obtained after slightly abrading the base. In heterosexual men, the common sites are the coronal sulcus, the glans, and inner surface of the prepuce but may be found on the shaft of the penis and beyond. In homosexual men, the ulcer is usually present in the anal canal, less commonly in the mouth and genitalia. In women, most chancres occur on the vulva, the labia, and, more rarely, the cervix where they are liable to be overlooked.

Fig. 3 Large primary sore. Note the even shape and the absence of secondary infection.

Extragenital chancres usually involve the lips, where they become large and associated with some oedema. Other sites are the mouth, buttocks, and fingers. The regional lymph nodes are invariably enlarged a few days after the appearance of the chancre and with genital sores the lymph nodes are bilaterally involved. They are painless, discrete, firm, and not fixed to surrounding tissues.

Atypical primary sores are not uncommon and depend on the size of the inoculum and the immunological status of the patient; thus a small inoculum usually produces a small, atypical ulcer or papule and looks trivial. This may also be the case in patients who had previously treated syphilis and the lesion may be dark-field negative.

Histologically, the chancre shows perivascular infiltration with plasma cells and histiocytes, capillary proliferation and obliterative endarteritis and periarteritis. The affected lymph nodes contain numerous treponemes, a depletion of lymphocytes, follicular hyperplasia, and histiocytic infiltration. If *T. pallidum* cannot be recovered from the primary sore, it may possibly be demonstrated from a needle aspirate of the regional lymph node.

Differential diagnosis (see also Chapter 23.1)

All genital sores must be regarded as syphilitic until proven otherwise, especially when solitary and painless. The following must be differentiated:

1. Genital herpes (see Chapter 7.10.2), which is much more common than syphilis in either sex, produces a crop of painful or irritating vesicles which develop into shallow erosions. In the first attack there is also painful inguinal adenitis.
2. Traumatic sores are painful, irregular erosions which may become secondarily infected.
3. Erosive balanitis causes inflammatory, irregular erosions which may become purulent in the uncircumcized.
4. Fixed drug eruptions are macules or occasionally ulcers following various drugs, especially tetracyclines.
5. Chancroid is mostly seen in the tropics, presenting as painful, superficial, 'soft chancre', often multiple, with painful suppurative regional adenitis.

Other conditions include scabies, Behçet's syndrome, donovanosis, and lymphogranuloma venereum.

Secondary syphilis

The lesions are numerous, variable, and affect many systems. Inevitably there is a symmetrical, non-irritating rash and generalized, painless lymphadenopathy. Constitutional symptoms are mild or absent; they include headaches, which are often nocturnal, malaise, slight fever, and aches in joints and muscles. The rash is commonly macular, pale red, and sometimes so faint as to be appreciated only in tangential light . It may be papular and sometimes squamous (Fig. 4). Pustular and necrotic rashes are rarely seen in temperate climates but still occur in tropical regions. The later the secondary rash develops, the more exuberant it becomes. The distribution of the rash can be of great diagnostic help. It usually covers the trunk and proximal limbs, but when it is seen on the palms, soles, and the face, syphilis should always be high on the list of probable causes (Fig. 5 and Fig. 6). In warm and moist areas such as the perineum, female external genitalia, perianal region, axillae, and under pendulous breasts, the papules enlarge into pink or grey discs, the condylomata lata, which are highly infectious (Fig. 7). Mucous patches in the mouth and genitalia are painless greyish-white erosions forming circles and arcs ('snail-track ulcers'). They too are very infectious.

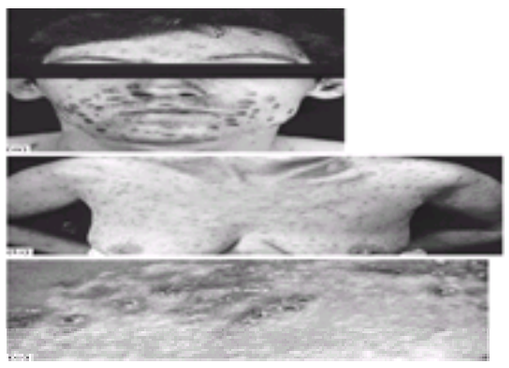

Fig. 4 (a and b) Secondary maculo-papular syphilitic rashes. (c) Late secondary early/tertiary papulosquamous lesions.

Fig. 5 Secondary papulosquamous rash of the soles.

Fig. 6 Secondary rash of the palms.

Fig. 7 Condylomata lata.

Meningism and headache, especially at night, are due to low-grade meningitis which can be confirmed by a raised cell count and raised protein in the cerebrospinal fluid. Less common lesions include alopecia and laryngitis. Syphilitic hepatitis is usually associated with a marked rise in serum alkaline phosphatase. The non-specific inflammatory changes in liver biopsy material are quite unlike those of viral hepatitis. Nephrotic syndrome may develop with glomerular immune-complex

deposits. Pain in the bones, often worse at night, is attributable to periostitis. Uveitis occurs in secondary and tertiary syphilis. In about one-fifth of patients, recurrent infectious episodes occur especially during the first year after the secondary stage. All these lesions disappear spontaneously leaving no trace.

Latent syphilis

By definition the patient is asymptomatic with normal cerebrospinal fluid findings but positive serology for syphilis. It is arbitrarily divided into early (<2 years) and late latent (>2 years) syphilis. Infectiousness does not stop with the advent of latency, as women may continue to give birth to congenitally infected infants during the early latent stage and for at least 2 years into the late latent stage. Approximately 60 per cent of patients remain latent for the rest of their lives, the only evidence of syphilis being positive serology, usually with a low titre. The rest develop clinical late syphilis but autopsy studies indicate that a higher proportion has subclinical infection, especially of the cardiovascular system.

Late syphilis (tertiary syphilis)

This includes late latent syphilis already referred to, benign tertiary syphilis, involvement of viscera, the central nervous system, and the aorta.

Pathogenesis of late benign syphilis

The gumma is a chronic, granulomatous lesion which is an intense inflammatory response to a few treponemes. Histologically, there is central necrosis with peripheral cellular infiltration of lymphocytes, mononuclear cells, and occasional giant cells with perivasculitis and obliterating endarteritis. *T. pallidum* is present and can be demonstrated by rabbit inoculation.

Clinical manifestations

Cutaneous gumma

Gummas are usually single but may be multiple or diffuse (Fig. 8). Clinically, it starts as a slowly progressive, painless nodule which becomes dull red and breaks down into one or several indolent punched-out ulcers. The base has a 'wash-leather' appearance, is remarkably free from secondary infection (Fig. 9), and often resembles other granulomatous conditions. It heals slowly from the centre which may become depigmented, whilst the periphery shows hyperpigmentation. Eventually, a paper-thin scar forms. Common sites are the face, legs, buttocks, upper trunk, and scalp.

Fig. 8 Multiple gummatous ulcers. This is a typical site.

Fig. 9 Single gumma. Note the punched-out ulcer and absence of secondary infection.

Mucosal gumma

These are most commonly seen in the oropharynx, involving the palate, pharynx, and the nasal septum. They tend to be destructive, causing perforation of the hard palate and the nasal septum and severe scarring of the pharynx and larynx. The most serious lesion is the diffuse gummatous infiltration of the tongue, leading first to a general swelling of the organ, then due to loss of papillae to a smooth red surface. After a while, the poor blood supply produces necrotic white patches on the dorsum of the tongue (Fig. 10). This leucoplakia has a strong tendency to become malignant. Penicillin has no effect on the progress of late syphilitic glossitis.

Fig. 10 Late syphilitic glossitis, early stage.

Late syphilis of bones

Osteoperiostitis of long bones such as the tibia and fibula causes thickening and irregularities which may be diffuse as in the 'sabre tibia' or localized as a

circumscribed bony swelling. Unlike most other syphilitic lesions, those of the bone are often painful, especially at night . Very rarely the process breaks through the skin producing a chronic 'syphilitic osteomyelitis'. Lesions of the palate, nasal septum, and the skull are destructive, leading to bone defects of the hard palate and nasal septum and multiple osteolytic lesions of the skull.

Differential diagnosis

Mucocutaneous gumma

Superficial skin lesions must be differentiated from fungal skin lesions, psoriasis, Kaposi's sarcoma, and iodide rashes. Deep gummas may resemble deep mycoses, sarcoidosis, tuberculosis, leprosy, donovanosis, lymphogranuloma venereum, reticulosis, and epithelioma of the skin. Serological tests for syphilis, which must include specific reactions such as the FTA–ABS (fluorescent treponemal antibody-absorbed test), prompt response to penicillin, and evidence of syphilis elsewhere clarify the diagnosis.

Late syphilis of bones

Primary and secondary carcinoma, Paget's disease, chronic osteomyelitis, tuberculosis, and leprosy should be considered. All forms of non-venereal syphilis, except pinta, give rise to similar lesions.

Visceral syphilis

This is not common and response to treatment is variable.

Liver

Multiple gummas of the liver give rise to irregular hepatomegaly ('hepar lobatum'), which may be asymptomatic. Symptoms may result from pressure on bile ducts or blood vessels or destruction of liver parenchyma.

Eyes

Uveitis, choroidoretinitis, or optic atrophy may sometimes be the sole feature of late syphilis. Uveitis can also develop during early syphilis, particularly in association with HIV, where bilateral uveitis is more commonly seen. Response of the late form to penicillin is poor. Optic atrophy is further discussed under neurosyphilis. The cerebrospinal fluid should be examined in all patients.

Stomach

Single or diffuse gummatous infiltrations of the stomach may respond to antisyphilitic treatment .

Lungs

Single or multiple gummas are rare and respond to treatment.

Testis

Gummatous infiltration and dense fibrosis may produce smooth, painless enlargement of a testis. Testicular sensation is lost .

For discussion of paroxysmal cold haemoglobinuria and neurosyphilis see Chapter 24.14.4.

Laboratory diagnosis of syphilis

Dark-field microscopy

The organism is seen in the wet preparation by dark-field microscopy in fluid taken from open lesions in early syphilis or needle aspirate from affected lymph nodes. In late lesions, the organism is not readily demonstrable by microscopy. Treponemes causing syphilis, non-venereal syphilis, yaws, and pinta cannot be differentiated morphologically. They give rise to the same serological reactions and are susceptible to penicillin.

Immunostaining directly with monoclonal antibody or by indirect fluorescent antibody staining of air-dried smears or tissue specimen has recently been reported to show numerous *T. pallidum* in a cutaneous gumma but that were not observed by dark-field microscopy or silver staining. Immunoperoxidase staining is useful in archival material.

Serology

Two classes of antibody tests are available:

a. those that measure non-specific antibodies (IgG and IgM) against lipoidal antigens (lipoidal antibody tests formerly called reagin tests);
b. those that measure specific antibodies stimulated by antigenic components of the treponeme and further divided into those stimulated by antigens found in pathogenic treponemes only and those shared with non-pathogenic treponemes.,

The non-specific cardiolipin antibodies act on lipoidal antigen which results from the action of *T. pallidum* on host tissue; it mirrors disease activity. The specific antigen is derived from *T. pallidum* and does not differentiate between past or present infection and is therefore of no value in assessing current activity. None of the tests distinguishes syphilis from other treponematoses. The interpretation of the tests is given in Table 2.

Animal inoculation

Animal inoculation with material from late lesions or from cases of 'persistent *T. pallidum*-like forms' is usually reserved for research purposes. PCR detection of treponemal nucleotides in amniotic or cerebrospinal fluid has proved no more sensitive than the detection of live spirochaetes by animal inoculation of these fluids.

Current serological tests for syphilis

The basis of these tests is shown in Table 3.

Lipoidal antigen tests

Rapid plasma reagin (RPR) test

This can be automated and is useful for screening purposes. It is the least technically demanding test (no microscope needed). It uses carbon-containing cardiolipin antigen and requires a minimal amount of blood. Filter paper or glass fibre discs can be used to post samples to laboratories; they are therefore mainly for use by primary health centres in outlying rural areas.

VDRL test

This is the preferred test and as it is used world-wide there is a good chance of it becoming standardized. It is simple and inexpensive. It is, above all, a quantitative test and as the titres reflect activity, it is of great value for this purpose. It may, for example, be the only evidence of reinfection in a patient with previous syphilis whose VDRL was either negative or weakly positive after treatment. A sharp, sustained rise of the titre, four-fold or higher, even in the absence of clinical signs is good evidence of active infection. False-positive VDRL is usually of a low titre (1:8 or less). The VDRL test becomes positive during the primary stage and rises to its maximum during the secondary stage (1:32 or more). After successful treatment, the titre declines (1:4 or less) and if treatment was given early in the disease it often becomes negative. An occasional, small and transient rise in titre (two tubes) is of no significance. The immunofluorescent test which detects cardiolipin F levels tends to be positive only in active infections.

Specific antitreponemal tests

T. pallidum haemagglutination (TPHA) test or T. pallidum particle agglutination test (TPPA)

This is a very valuable and simple test using an indirect haemagglutination method with red cells or by gelatin particles attached to sonicate of *T. pallidum* extract. It is almost as specific as the near-obsolescent TPI (treponemal immobilization test) reaction but is less sensitive than the FTA–ABS test. False-positive reactions occur in up to 2 per cent. The micromethod is particularly suitable for screening purposes. This test together with the VDRL it is probably the best combination for routine use. In cases of doubt, the FTA–ABS test is added. TPHA can be adapted for automation. There is no standard 'cut off' for the TPHA in the cerebrospinal fluid.

The FTA–ABS test

This uses the indirect fluorescent technique with killed *T. pallidum* as antigen. The organisms are fixed on a slide to which the serum is added. The antibody in the serum will unite with the treponemes and this can be made visible by adding antihuman globulin conjugated with a fluorescent stain which attaches itself and produces fluorescence of the treponemes seen by fluorescent microscopy. The test has been made more specific by absorbing the group antibodies with a sonicate derived from Reiter's treponemes. The test is then called FTA–ABS. The FTA–ABS is the most sensitive test available and is also specific. It becomes positive earlier during the primary stage of syphilis than other procedures. It is not suitable for assessing activity, as it persists long after successful treatment . When the routine serology includes the VDRL and TPHA tests, the FTA–ABS test should be added in cases of problem sera.

The FTA–ABS–IGM test

In the search for a specific test to differentiate active infection which has to be treated from adequately treated or 'burnt-out' inactive disease, the FTA–ABS–IGM is being evaluated to test for specific IgM which develops in the course of syphilis. This test sometimes gives false-positive reactions owing to the presence of anti-IgG globulins (e.g. rheumatoid factor) and false-negative reactions have also occurred; thus there are problems in the use of this test in the elderly. Its use in the rapid, early diagnosis of congenital syphilis was considered some years ago to be great advance. The basis for its use is that IgG is a small molecule and passes through the placenta; thus the baby may inherit maternal IgG but not necessarily the infection. Tests such as the VDRL may therefore be positive in the newborn by passive transfer and may take 3 months to disappear. IgM is a large molecule and does not pass through the placenta; thus if it is found in the neonate, it must be assumed that it has been produced by the infected infant.

Newer tests

Enzyme immunoassay (EIA) tests (e.g. CAPTIA IgG, IgM) have no practical advantage over current tests. Attempts with purified antigens from *T. pallidum* may be specific, although they tend to loose sensitivity, for example, those employing the TpN19 or Tp44.5 antigens. The problem is that early cases may not always have antibodies to a specific antigen. It would be better, therefore, to screen with as many antigens as possible. This is achieved by western blotting, which allows the separation by gel electrophoresis of all the major, stable antigens of the spirochaete, and enables a distinction to be made between the non-specific and specific antibody reactions with given bands. The major antigens are shown in Table 4.

This technique may have special application in seronegative syphilis (perhaps syphilis occurring with HIV infection) but not in primary syphilis, where up to half the cases show no reaction. The immunoblot may also be helpful in distinguishing false positive tests caused by borrelia infections.

Diagnosis of neurosyphilis by examination of the cerebrospinal fluid

The traditional tests include the VDRL, cell count, total protein globulin, and goldsol curve. The last two tests are now obsolete. The cerebrospinal fluid VDRL is unreliable as it can be negative in up to 50 per cent of samples from patients with active neurosyphilis. Cell counts exceeding 5/mm ^{2} (but usually not above 50/mm^2) and protein above 40 mg/ml (60 mg/ml in patients older than 65 years) are non-specific signs of inflammation. The specific FTA–ABS and TPHA tests in the cerebrospinal fluid may be positive due to passive transfer of serum IgG from adequately treated patients. If they are negative, active neurosyphilis can almost certainly be excluded.

Biological false-positive tests for syphilis

These concern mainly the cardiolipin tests and are classified as acute, as in drug addicts, or chronic if they occur in autoimmune disease (when they may precede the symptoms by years), leprosy, and in a small proportion of people over 70. The concurrence of HIV and syphilis in drug abusers makes the investigation of these false-positive reactions particularly important. There is no evidence for an increase in false-positive tests in patients with HIV. Particular mention should be made of the thrombotic antiphospholipid syndrome since the condition is associated with early miscarriage and cerebral thrombosis, manifestations which might be confused with syphilis. The confirmatory TPHA and FTA–ABS tests are always negative. A biological false-positive test may occur acutely in the cerebrospinal fluid in aseptic meningitis or in a seropositive patient, when a traumatic tap may give a false impression of a positive cardiolipin test in the cerebrospinal fluid, following transfer of plasma antibody.

The management of syphilis

Suggestions for drug treatment of syphilis are given in Table 5. It has been found that adherence to clinical guidelines is better maintained when treatment and follow-up is performed by a sexually transmitted disease clinics than by a non-institutional practitioner. As soon as a diagnosis of infectious syphilis has been made, the patient should be interviewed by a social worker regarding all sexual contacts. In the case of primary syphilis, this should cover the previous 3 months; in patients with secondary syphilis this should be extended to 1 year; and in patients with early latent syphilis to 2 years because of the possibility of infectious relapses during that period. The patient is warned against intercourse during treatment and for a further 2 weeks. Experience suggests that advice for longer abstinence will be disregarded in many cases and is almost certainly unnecessary.

If the patient gives no history of penicillin allergy, it is the first choice for the treatment of all stages of the disease. Penicillin is as effective now as it was more than 40 years ago when it was first introduced. Resistance to penicillin has not been described, perhaps related to the novel penicillin binding of the T $_P$47 (TPO971) protein. If there is penicillin allergy, the alternative drugs are tetracycline/doxycycline and erythromycin. The recent finding of a wild strain of *T. pallidum* resistant to erythromycin has led to an extensive investigation into the use of newer cephalosporins. Cephalosporins are effective but there is cross-allergy with penicillin in 5 to 7 per cent of patients.

The optimal dose or duration of treatment with penicillin, or the other drugs, has not been established and therefore a great variety of treatment schemes have been put forward, although the results appear to be similar, suggesting that a fair degree of variation is permissible. The general tendency is to treat with larger doses and over a longer period of time in the later stages of syphilis; some prefer to repeat the course. There is no convincing evidence that large, much extended, or repeated courses give any added benefit.

There is good experimental evidence that serum concentrations of penicillin should be at least 0.003 µg/ml, should be maintained for 8 to 10 days, and that troughs in the concentration should not exceed 15 h. Some physicians prefer a single injection of the long-acting benzathine penicillin (2.4 million units) for simplicity, but the

concentration reached is low and does not give a useful level in the cerebrospinal fluid; also the injection is quite painful. Others repeat this dose weekly, for 3 weeks. In patients with neurosyphilis and HIV infections, treponemes have been demonstrated in the cerebrospinal fluid after benzathine penicillin G treatments and in these patients the expected decline in VDRL cerebrospinal fluid titres after treatment occurred less often than in those without concurrent HIV infection. All treponemal infections are unaffected by sulphonamides, gentamicin, rifampicin, and quinolones in clinical dosage.

Procaine penicillin has several advantages over other penicillin preparations and is preferred by many. In some centres, the course is 1 million units/day for 10 days; in others it is given for 20 days though evidence that such a prolonged course gives better results is lacking. Procaine penicillin in 2 per cent aluminium monostearate (PAM) has a prolonged action and was used extensively by the World Health Organization in their mass campaign against non-venereal syphilis.

Penicillin reactions

All patients receiving penicillin injections should be kept in the clinic for 15 to 20 min as severe reactions needing immediate treatment will develop well within this period. An emergency tray to deal with anaphylactic penicillin reaction must be readily available wherever penicillin is given. It should contain ampoules of 1:1000 adrenaline (epinephrine) solution, syringes and needles, intravenous hydrocortisone, injectable antihistamine, aminophylline, an airway respirator (Ambu bag or Brooke's respirator), and oxygen with face mask or nasal catheter.

Prevention of penicillin reactions

Some 3 to 5 per cent of the population in the United Kingdom are allergic to penicillin and it is essential to enquire about this; if there is a history, penicillin must not be given. This fact should be displayed prominently on the cover of the medical notes and the patient told to inform any doctor who may wish to give this antibiotic. Careful history taking may, however, show that the 'allergy' to penicillin is doubtful, for example the rash antedated the giving of penicillin and may have been due to one of the childhood infections. It is quite common to be told that patients who apparently did have a penicillin reaction, had no problems when the antibiotic was inadvertently given subsequently as penicillin allergy is a transient phenomenon. In such cases we still prefer to avoid giving penicillin.

Clinical features

The most serious reaction is anaphylactic shock appearing immediately or within a minute or two after the injection. The more immediate the onset, the more severe the attack. The patient becomes unconscious, stops breathing, and becomes pulseless. Very rarely, the patient dies immediately. A fatal outcome is estimated to occur one or two times per 100 000 injections. In the more moderate reaction, the patient feels faint with acute anxiety and a feeling of impending death; there may be oedema of the face, possibly with an asthmatic attack, soon followed by urticaria. Arthralgia and some pyrexia may develop. The urticaria is liable to last 1 to 2 weeks.

The commonest form is the delayed reaction when urticaria appears days after injection or oral penicillin. Arthralgia and fever may develop. Sometimes a local reaction around the injection site is seen. It can be urticarial but is more commonly a painful red swelling and usually responds to rest . It is best to discontinue the course, as recurrences are otherwise common. In some patients a hysterical episode follows an injection and this may be due to procaine or possibly inadvertent intravenous injection. It passes off spontaneously.

Treatment of the anaphylactic reaction

The patient is laid flat with feet up and head down. Blood pressure and pulse are monitored throughout . Adrenaline 1:1000 (adult dose 0.5–1.0 ml) is given intramuscularly without delay. If bronchospasm develops, 250 mg aminophylline in 10 ml water is administered by slow intravenous injection. Intravenous hydrocortisone (100 mg) may also be tried and may be repeated. Some prefer intravenous antihistamine (chlorpheniramine injection 10–20 mg). Adrenaline, nevertheless, is the mainstay of treatment. If there is no response, the cardiac arrest team is summoned. If recovery is slow, the patient should be admitted as recurrences may occasionally occur. In any case, the patient must be kept under observation for several hours. Later, urticaria develops in most patients and prophylactic antihistamines by mouth are indicated.

Treatment of the delayed reaction

The leading feature is urticaria, possibly with oedema of the face, arthralgia, and some fever. Such patients respond to oral antihistamines such as chlorpheniramine 4 mg four times daily or terfenadine 60 mg twice daily until the condition is controlled. If it is very severe, prednisolone 10 mg four times daily may be added for a few days, reducing it as soon as possible. Penicillinase is not recommended as it may produce reactions of its own.

Procaine reaction

Two types of reaction are recognized:

1. The patient shows extreme anxiety with a feeling of impending death. Sometimes there are hallucinations, disorientation, and depersonalization. The reaction is self-limiting. It may be due to reduced procaine esterase leading to high procaine blood levels. Patients should be restrained and reassured.
2. The reaction is similar but associated with hyperventilation, hypertension, tachycardia, and vomiting. Rarely cardiovascular collapse has been reported but without fatalities. The reaction is thought to follow accidental intravenous administration of procaine penicillin leading to microemboli of the lungs and brain. Supportive treatment is usually sufficient.

Vasovagal attacks

These occur most commonly in young men following intramuscular injection or after having blood taken. The patient looks very pale and may faint . He may slump to the floor and occasionally go stiff and have jerky movements. The most important diagnostic sign is a slow pulse. Recovery is rapid once he is laid flat on a couch. There is a tendency for recurrence in the same individual under similar circumstances and this can usually be prevented by giving injections or taking blood whilst the patient is lying down.

Jarisch–Herxheimer reaction (see also Chapter 7.11.30)

This systemic reaction is believed to be due to the release of endotoxin-like substances when large numbers of *T. pallidum* are killed by antibiotics. It is mainly associated with early syphilis. The incidence of the reaction appears to be related to the total number of the organism in the body. The mechanism may not be straightforward as it is not a feature of neonatal syphilis or non-venereal syphilis in childhood. The reaction can be expected in 50 per cent of primary syphilis, 90 per cent of secondary syphilis, and in 25 per cent of early latent infection, but is very rare in late syphilis. It has been suggested that it is more often seen in patients with HIV.

The reaction begins 4 to 12 h after the first injection, lasts for a few hours or up to a day, and is not seen with subsequent treatment . There is malaise, slight to moderate pyrexia, a flush due to vasodilation, tachycardia, and leucocytosis, and existing lesions become more prominent. In some patients with early syphilis, a secondary rash may become visible which was absent before treatment. Rarely, syphilis may be suspected by the appearance of the febrile reaction of the Jarisch–Herxheimer, perhaps with a fleeting rash, when treating another infection with a treponemocidal antibiotic (e.g. penicillin in gonorrhoea).

In early syphilis the reaction is only a minor nuisance. In late syphilis it can on very rare occasions be more serious. Thus in neurosyphilis it may lead to epilepsy or a rapid. irreversible progression, and in general paresis it can cause exacerbation amounting to temporary psychosis. Sudden death has been reported in cardiovascular syphilis. In laryngeal gumma, local oedema may necessitate tracheotomy. In the later stages of pregnancy, fetal monitoring is advised.

It is customary to give corticosteroids in late symptomatic syphilis starting a day before the first penicillin injection and tailing it off the day after the first injection. This does not prevent the Jarisch–Herxheimer reaction but is said to ameliorate it . The analogous reactions in relapsing fever have been modified by meptazinol or pretreatment with infusions of polyclonal anti-TNF-a Fab with concomitant reduction in the plasma concentration of interleukin 6 and 8.

Follow-up

It is generally sufficient to perform blood tests 1, 3, 6, and 12 months after treatment of early syphilis. In late symptomatic syphilis, surveillance is for life. Patients with leucoplakia of the tongue should be checked every 3 months. In symptomatic cardiovascular syphilis regular radiological and clinical examination is essential to determine any change which might suggest the need for cardiac surgery. In neurosyphilis an annual review might be adequate.

In latent syphilis, if there is a satisfactory serological response, 2 to 3-yearly follow-up seems reasonable. The cerebrospinal fluid need not be examined in the non-immunocompromised patient, except in the presence of neurological signs. In early congenital syphilis, the observation time should be similar to that of early acquired syphilis. In late latent congenital syphilis, no further attendance is necessary unless symptoms of interstitial keratitis or other lesions not prevented by penicillin develop.

In high-risk patients such as male homosexuals and prostitutes a regular check-up every 3 months is advised. If such patients have had syphilis, the VDRL should have become negative or of a low titre after treatment. If the titre suddenly rises four-fold or more, reinfection must be assumed and treatment is indicated.

Prophylaxis

Treatment of asymptomatic contacts of early syphilis is recommended in the United States as there is a 50 per cent chance of infection. Such pre-emptive treatment is likely to reduce the spread of infection in the promiscuous or in those likely to infect their spouses or regular sexual partners. Use of condoms should be recommended. Various vaginal chemical spermicidal creams give a small degree of protection but are unreliable.

Further reading

Borisenko KK, Tikhonova LL, Renton AM (1999). Syphilis and other sexually transmitted infections in the Russian Federation. *International Journal of STD and AIDS* **10**, 665–8.

Byrne RE, Laska S, Bell M, Larson D, Phillips J, Todd J (1992). Evaluation of a *Treponema pallidum* western immunoblot assay as a confirmatory test for syphilis. *Journal of Clinical Microbiology* **30**, 115–22.

CDC (1998). Guidelines for treatment of sexually transmitted diseases. *Syphilis* **47**, (RRI) 28–48.

Egglestone SI, Turner AJL (2000). Serological diagnosis of syphilis. *Communicable Disease and Public Health* **3**, 158–62.

Grimble AS (1971). Venereal disease in the young patient: a perspective. *Guy's Hospital Reports* **120**, 323–6.

Haake DA (2000). Spirochaetal lipoproteins and pathogenesis. *Microbiology* **146**, 1491–504.

Kell P, McMorrow S, Smith A (2000). Management of syphilis in pregnancy. *Bulletin of Sexually Transmitted Infections and HIV* **4**, 9–12.

Luger AF, Schmidt BL, Kaulich M (2000). Significance of laboratory findings for the diagnosis of neurosyphilis. *International Journal of STD and AIDS* **11**, 224–34.

Morton RS (2001). 'The syphilis enigma', the riddle resolved. *Sexually ransmitted Infections* **77**, 322–4.

Norris SJ (1993). Polypeptides of *Treponema pallidum*: progress towards understanding their structural, functional and immunologic roles. *Microbiology Reviews* **57**, 750–79.

Oriel JD (1994). *Scars of Venus: a history of venereology*, pp. 1–181. Springer Verlag, London.

PHLS, DHSS and PS and the Scottish ISD (D) 5 Collaborative Group. Sexually transmitted infections in the UK. new episodes seen at genito-urinary medicine clinics, 1995 to 2000. *Public Health Laboratory Service*, London.

Subramanian G, Koonin EV, Aravind A (2000). Comparative genome analysis of the pathogenic spirochetes *Borrelia burgdorferi* and *Treponema pallidum*. *Infection and Immunity* **68**, 1633–48.

US Department of Health and Human Services, Public Health Services: Centers for Disease Control (1 993) (1992). *Sexually transmitted disease surveillance*, pp. 6–11, 139–148, (definitions: 185–187). Atlanta, GA.

Van Vorst Vader PC (1998). Syphilis management and treatment. *Dermatology Clinics* **16**, 699–711.

White RM (2000). Unravelling the Tuskegee study of untreated syphilis. *Archives of Internal Medicine* **160**, 585–98.

Wilner E and Brody JA (1968). Prognosis of general paresis after treatment. *Lancet* **2**, 1370–1.

World Health Organization (1993). *Draft recommendations for the management of sexually transmitted diseases*. WHO advisory group meeting, WHO/GPA/STD/93. 1, pp. 24–31. WHO, Geneva.

Young H (2000). Guidelines for serological testing for syphilis. *Sexually Transmitted Infections* **76**, 403–5.

Website addresses

Website 1 http://www.tigr.org/ The Institute for Genomic Research

Website 2 download by anonymous ftp at ftp://ncbi.nlm.nih.gov/pub/Koonin/Spirochetes

7.11.34 Listeriosis

P. J. Wilkinson

Listeriosis has been recognized since the 1920s as a systemic infection of man, domestic and farm animals, and rodents. Because the disease in rabbits and guinea pigs was characterized by a marked mononuclear leucocytosis, the causative Gram-positive bacillus was named *Bacterium monocytogenes*, then (in honour of Lord Lister) *Listerella* and finally *Listeria.* Human listeriosis was a relatively obscure disease until the 1980s, when a series of food-borne outbreaks awakened interest. Listeriosis remains a rare infection but carries a significant morbidity and mortality.

Listeria monocytogenes

*Listeria*spp. are non-sporing, facultatively anaerobic, Gram-positive bacilli that are ubiquitous in the environment and distributed world-wide. *L. monocytogenes* is the major pathogen, although occasional human infections with *L. ivanovii* and *L. seeligeri* have been reported. *L. ivanovii* and *L. innocua* can infect animals. *L. welshimeri* and *L. grayii* are not known to cause disease. Enrichment and selective methods are now well established for the isolation of listeria from food or the environment; immunoassays and nucleic acid amplification techniques have also been used. The ability to multiply at temperatures of 0 to 40° C and tolerate preserving agents makes listeria of particular concern if present in refrigerated foods that are consumed without further cooking.

Several typing methods are used to trace food sources, distinguish relapses from reinfections, and investigate outbreaks. Thirteen serovars are currently recognized, of which three (1/2a, 1/2b, and 4b) cause more than 90 per cent of human and animal infections. Phenotypic subtyping systems, based on patterns of lytic reactions with bacteriophages, bacteriocin (monocin) production, and multilocus enzyme electrophoresis have been enhanced by genotypic analysis, particularly pulsed field gel electrophoresis (PFGE).

Epidemiological associations

L. monocytogenes has been isolated from many foods, and the consumption of contaminated meat, milk, seafood, or vegetables is the principal route of infection. Outbreaks have been associated with coleslaw, raw fish, raw hot dogs, undercooked chicken, meat paté, pork rillettes, turkey franks, smoked fish or shellfish, and cheese and dairy products, particularly when pasteurization has been ineffective. The United Kingdom Department of Health advises pregnant women and immunocompromised persons not to eat soft ripened cheese (e.g. Brie, Camembert, and blue-vein types), all types of paté, and cook-chill meals and poultry unless thoroughly reheated until piping hot.

Direct transmission through contact with infected animals can give rise to primary cutaneous listeriosis, an occupational disease of farmers and veterinarians. Laboratory workers have acquired eye and skin infections from direct exposure to culture material. Nosocomial infection has spread between neonates in association with poor hand hygiene, close contact between infected patients and their mothers, and fomites such as rectal thermometers. Hospital outbreaks, which may have been food-borne, have also occurred in adult immunosuppressed patients.

Pathogenesis

Although listeria displays many characteristics of saprophytes, specific adaptations allow *L. monocytogenes* to become an intracellular pathogen where invasion and multiplication in both phagocytic and non-phagocytic cells occurs. CR3 complement receptors may be involved in the adhesion to phagocytes. Internalin, a listerial surface protein similar to the M protein of group A streptococci, plays a part in the initial stages of invasion of all cell types, as may p60, another cell surface protein with murein hydrolase activity. After internalization, *L. monocytogenes* becomes encapsulated in a vacuole, the membrane of which is dissolved by a thiol-activated haemolysin (listeriolysin O) and possibly also by phospholipase C. Having entered the host cell cytoplasm, the organisms grow, polymerize actin, acquire intracellular mobility, and spread to adjacent cells.

Clinical features

Although listeriosis is generally an opportunistic infection of the elderly, patients with severe underlying illness, pregnant women, newborn babies, and individuals without these risk factors can also become infected. The clinical presentation varies from a mild, influenza-like illness to fatal septicaemia and meningoencephalitis. The syndromes recognized include maternofetal and neonatal listeriosis, septicaemia, meningoencephalitis, cerebritis, gastroenteritis, and localized infections.

- In maternofetal listeriosis, the mother may develop a fever, headache, myalgia, and low back pain, associated with the bacteraemic phase of the disease. Transplacental infection causes amnionitis and usually leads to spontaneous septic abortion or to premature labour with the delivery of a severely infected fetus or baby.
- Neonatal listeriosis of early onset results from intrauterine infection and has a high mortality. The liquor is meconium-stained and the baby septic and jaundiced, with signs of purulent conjunctivitis, bronchopneumonia, meningitis, or encephalitis. Granulomas affect many organs, hence the term 'granulomatosis infantisepticum'. Late-onset disease, which develops several days to weeks in a baby who was initially healthy but subsequently develops meningitis, which may be acquired from the mother's genital tract or through cross-infection from an early-onset case.
- Septicaemia occurs mainly in adult patients with malignancies, in transplant recipients, and in immunosuppressed and elderly people. Most present with fever, hypotension, and shock but a third to a half develop meningitis, which is often then the presenting feature.
- Meningoencephalitis may start abruptly but, in adults, can also develop insidiously, with progressive focal neurological signs even in the absence of a brain abscess. Most patients have meningism, but fever may not be marked, particularly in elderly or immunosuppressed people. This infection should be considered in any patient with an acute brain-stem disorder associated with fever, particularly if there are no risk factors for cerebrovascular disease.
- Cerebritis is increasingly recognized, particularly in the immunosuppressed patient. Headache, fever, and varying degrees of paralysis can resemble a cerebrovascular accident. Rhomboencephalitis begins with a headache, fever, nausea, and vomiting followed in several days by symmetrical, progressive cranial nerve palsies, decreased consciousness, and cerebellar signs. Areas of uptake without ring enhancement may be shown by MRI or CT scan, and the cerebrospinal fluid shows few, if any, cells, and normal protein and sugar concentrations.
- Gastroenteritis with arthromyalgia, fever, diarrhoea, nausea, vomiting, and an incubation period of 1 to 3 days has recently been described in outbreaks of infection in immunocompetent adults who have ingested contaminated food. Because diagnostic laboratories do not usually culture diarrhoeal stools selectively for listeria, *L. monocytogenes* may be missed. Recent outbreaks have come to light when blood cultures from hospitalized patients have yielded the organism, or when serological testing of the blood of recently affected patients has shown antibody to listeria.
- Localized infections are rare, occur mainly in immunosuppressed people, and include abscesses, cholecystitis, endocarditis, endophthalmitis, osteomyelitis, septic arthritis, and peritonitis. They usually result from seeding during an initial bacteraemic phase, but focal skin and eye infection can also result from direct, occupational exposure.

Diagnosis

The microbiological diagnosis of invasive listeriosis is made by culture of the organism from meconium, nose or eye swabs, urine, cerebrospinal fluid, blood, tracheal aspirate, placental tissue, and/or lochia. Gram-positive bacilli may be seen in a stained smear. In listeria meningoencephalitis, the cerebrospinal fluid exudate is predominantly mononuclear and, if no bacteria are seen in a Gram-stained film, may be confused with viral meningitis; however, unlike viral meningitis, the cerebrospinal fluid protein is high and the glucose concentration low in relation to that in the peripheral blood. Tests for listeria antibodies in maternal and cord blood

samples do not contribute to the diagnosis of the acute infection. Selective culture techniques have considerably improved the isolation rate.

Antibiotic treatment

There are no controlled trials of antibiotic treatment for listeriosis. All strains are susceptible to ampicillin, which acts synergistically with aminoglycoside antibiotics, and high-dose intravenous ampicillin in combination with gentamicin remains the treatment of choice. Gentamicin is best avoided in pregnancy, when ampicillin may be used alone, or erythromycin if the patient is penicillin-allergic. Intravenous co-trimoxazole is the best second-line treatment for listeria meningoencephalitis. Vancomycin has been successfully used with gentamicin to treat bacteraemic illness, but does not cross the blood–brain barrier. Rifampicin and ciprofloxacin have not been evaluated in human listeriosis.

L. monocytogenes is inherently resistant to the cephalosporins and it is very important to be aware that treatment with this class of antibiotics is likely to fail. Since acute pyogenic meningitis is usually treated, until the pathogen is known, with ceftriaxone or cefotaxime, ampicillin should also be given with this initial treatment whenever listeriosis is a clinical possibility, unless the cerebrospinal fluid Gram-film shows good evidence of another bacterial cause, or the patient has unequivocal clinical features of meningococcal disease.

Intravenous ampicillin should be given in a daily dose of 200 to 300 mg/kg (neonates) or 6 to 12 g (adults) in three to four divided doses for 2 weeks (uncomplicated bacteraemia), 4 to 6 weeks (meningoencephalitis), or 6 to 8 weeks (endocarditis). Intravenous gentamicin should be given for the first 14 days in a dosage adjusted with the help of plasma concentration measurement. Focal listeriosis may be treated with ampicillin or amoxicillin, 3 to 6 g daily, until clinical resolution. In cases of genuine penicillin allergy, intravenous co-trimoxazole, 20 mg/kg per day (trimethoprim component) may be given in four divided doses. Alternatively, intravenous minocycline, which may have to be obtained specially from the manufacturer, can be used in a daily dose of 200 mg (adults) or 4 mg/kg (children), in combination with gentamicin.

Prognosis

Despite antibiotic therapy, the mortality of septicaemia and meningoencephalitis with *L. monocytogenes* remains high (20–50 per cent). There is significant long-term morbidity in the survivors. Efforts should therefore continue to be focused on the prevention of this infection by improvement in the microbiological safety of methods of food production and preparation and by the continued education of the public so that vulnerable people can avoid high-risk foods.

Further reading

Jones EM, MacGowan AP (1995). Antimicrobial chemotherapy of human infection due to *Listeria monocytogenes*. *European Journal of Clinical Microbiology and Infectious Diseases* **14**, 165–75.

McLauchlin J (1997). The pathogenicity of *Listeria monocytogenes*: a public health perspective. *Reviews in Medical Microbiology* **8**, 1–14.

McLauchlin J, Jones D (1998). Erysipelothrix and Listeria. In: Balows A and Duerden BI, eds. *Topley and Wilson's microbiology and microbial infections*, Vol. 2, pp. 683–703. Arnold, London.

McLauchlin J, Low JC (1994). Primary cutaneous listeriosis in adults: an occupational disease of veterinarians and farmers. *Veterinary Record* **135**, 615–17

Schlech WF III (1991). Listeriosis: epidemiology, virulence and the significance of contaminated foodstuffs. *Journal of Hospital Infection* **19**, 211–24.

Schlech WF III (1997). Listeria gastroenteritis—old syndrome, new pathogen. *New England Journal of Medicine* **336**, 130–2.

Salamina G *et al.* (1996). A foodborne outbreak of gastroenteritis involving *Listeria monocytogenes*. *Epidemiology and Infection* **117**, 429–36.

7.11.35 Legionellosis and legionnaires' disease

J. B. Kurtz and J. T. Macfarlane

In 1976 an outbreak of pneumonia occurred among American legionnaires who had attended a convention in a Philadelphia hotel. A total of 221 people developed pneumonia, 'legionnaires' disease', of whom 34 died. A newly identified organism, named after this outbreak, *Legionella pneumophila*, was responsible. Since then other outbreaks and sporadic cases have been recognized and 16 *L. pneumophila* serogroups, and other species of legionella besides *L. pneumophila* have been isolated from clinical and environmental samples. There are now at least 43 recognized species in the Legionellaceae family. Clinical illness caused by members of the family Legionellaceae is referred to as legionellosis. The pneumonia is called legionnaires' disease. Non-pneumonic legionellosis ('Pontiac fever') is a self-limiting, influenza-like illness, without radiographic changes in the lung, caused by many different legionella species. What determines the type of illness that will follow infection is unknown. Although, in a given outbreak, disease of both pneumonic and non-pneumonic types occurs, usually either one or other form predominates. *L. pneumophila* is responsible for over 80 per cent of legionellosis, and of the 16 serogroups serogroup 1 is the most frequently encountered in human infections. In some parts of Australia, however, *L. longbeachae* is the most frequently identified species causing legionnaires' disease. Other legionella species appear to be less pathogenic and are more frequently found as opportunist pathogens in immunocompromised people. Some have caused disease, others have only been isolated from the environment and have yet to be implicated as human pathogens.

The organism

The Legionellaceae are aerobic, non-sporing bacilli whose cell walls contain distinctive branched-chain fatty acids.

In the laboratory, legionellae are fastidious in their growth requirements and will not grow on standard bacteriological media. Aces buffered charcoal yeast-extract agar, pH 6.9, supplemented with L-cysteine, a-ketoglutarate, and iron, is a very satisfactory medium. On this medium, incubated at 35 to 37 °C, typical colonies usually appear in 3 to 5 days; occasional slow-growing strains require the plates to be incubated for 10 days. When isolates from a patient and a suspected environmental source (see below) have been obtained, an accurate comparison of the strains should be undertaken. Both genotypic (e.g. amplified fragment length polymorphism) and phenotypic (e.g. monoclonal antibody reaction pattern) methods of identification should be used in parallel to see whether the two isolates are indistinguishable or different.

Epidemiology

The natural habitat of legionellae is in freshwater streams, lakes, and thermal springs, moist soil, and mud. They have been found worldwide in waters with temperatures varying from 5 to 62 °C and pH values of 5.4 to 8.2. These organisms are inhibited by sodium chloride and are not found in sea water. In natural habitats they are found in only small numbers, forming part of the consortium of micro-organisms that makes up the biofilm. This includes amoebae and other protozoa, in certain of which legionellae multiply and later re-emerge. Inside these protozoa the bacteria form microcolonies, which are protected from adverse conditions (for example, in amoebic cysts from desiccation and up to 50 parts per million of free chlorine). This association enables the bacteria to survive and to disseminate widely in the natural environment.

Legionellae have been found in small numbers in water distribution systems, through which they can colonize man-made habitats, again as part of the biofilm from which they are shed into the water. Factors that encourage colonization and multiplication are temperature (20–45 °C) and stagnation. The most common sites in buildings in which legionellae have been found are hot-water calorifiers and storage tanks. Piped water, especially hot water from the calorifiers in large buildings and industrial complexes with long runs of pipework, is a potential source of infection. Other well-recognized sources include:

- recirculating water in air-conditioning and cooling systems;
- whirlpool spas and other warm-water baths;
- decorative fountains;
- nebulizers and humidifier reservoirs of hospital ventilation machines if topped up with contaminated tap water;
- potting compost for *L. longbeachae* serogroup 1 in Australia.

Dissemination of infection is by contaminated water droplets (aerosol), which are inhaled. In order to cause infection the droplets must be of a size (less than 5 µm diameter) that can reach the alveoli of the lungs. Taps and shower heads produce very localized aerosols, whereas the water droplets (drift) contained in the airstream released from a cooling tower may be carried a considerable distance and expose a greater number of people to risk. For example, in the 1976 Philadelphia outbreak, those infected in the street developed 'Broad Street pneumonia' and passers-by were infected in both the Stafford District General Hospital outbreak in 1985 (101 cases, 28 deaths) and near the BBC building, London, in 1988 (79 cases, 2 deaths). Person-to-person spread of legionellosis has never been recorded. Aspiration of contaminated water as might occur in hospital following an anaesthetic is also a well-recognized route of infection.

Although most studies of legionnaires' disease have been of outbreaks, sporadic cases account for about three-quarters of those reported in England and Wales. A source is only exceptionally found for them. Some of these sporadic events become part of an outbreak when other cases can be linked to them epidemiologically, as when patients from different geographical areas give a history of visiting a common site within the incubation period of their illnesses. An association with overseas travel was found in one-third of the sporadic cases in England and Wales for the years 1979 to 1986. Apart from travel, an analysis of sporadic (better called non-travel, non-outbreak) cases in Glasgow between 1978 and 1986 supported the hypothesis that cooling towers were the source of the infections. The relative risk was three times greater for people living within 500 m of a cooling tower compared with those living more than 1000 m away.

In temperate countries legionellosis has a seasonal pattern, most cases occurring in the summer and autumn. A multicentre British Thoracic Society study of community-acquired pneumonia requiring hospital admission in 1982 to 1983 showed that 2 per cent had legionnaires' disease. This suggests that about 1500 cases occur per year in Britain.

The susceptibility to infection of exposed people varies. For non-pneumonic legionellosis the attack rate is very high. In contrast, the attack rate for legionnaires' disease is about 1 per cent, although subclinical or mild infections can follow exposure, as indicated by serological surveys. For example, of the staff at the Stafford District General Hospital who were tested following the outbreak in 1985, 42 per cent had an antibody titre of 1 in 16 or greater.

Hospital-acquired legionellosis has been a particular problem. This is because of the complexity of the buildings and the difficulty of keeping the hot water hot (storage at 60 °C and 50 °C at the taps), either because of the length of pipework or for fear of scalding patients. Hospital patients, too, are a highly susceptible population and species other than *L. pneumophila* more frequently cause infections in these circumstances. In intensive care units, inhalation of air passed through contaminated humidifiers or aspiration of contaminated water are other potential sources of infection.

Clinical manifestations

Legionella pneumonia

Large studies have suggested that legionella infection is the cause of around 2 to 5 per cent of cases of community-acquired pneumonia admitted to hospital, although there is wider geographical and seasonal variation. Infection tends to lead to moderate or severe infection rather than mild illness, and most patients require hospital admission within 5 to 7 days of the start of symptoms.

The incubation period is usually 2 to 10 days, with a mean of 7 days; males are two to three times more frequently affected than females. Infection at the extremes of age is rare and the highest incidence is in 40- to 70-year-old people, with a mean age of 53 years. People particularly at risk include cigarette smokers, alcoholics, diabetics, and those with chronic illness or who are receiving corticosteroids or immunosuppressive therapy. Consequently, the type of patient who requires admission to hospital is particularly at risk from a nosocomial source.

Clinical features (Table 1)

Typically, the illness starts fairly abruptly with high fever, shivers, bad headache, and muscle pains. Upper respiratory tract symptoms, herpes labialis, and skin rashes are uncommon. The cough is usually dry initially but dyspnoea is common and the illness often progresses quickly. Sometimes there may be a history of a recent hotel holiday abroad or a stay in hospital, which can alert the clinician to the possible diagnosis.

The patient commonly looks toxic and ill, with a high fever over 39 °C. Confusion and delirium or diarrhoea can dominate the clinical picture, masking the true diagnosis of pneumonia. Focal neurological signs, particularly of a cerebellar type, are well described. Amnesia on recovery is common.

Laboratory findings

The total white count is usually only moderately raised, to 15 000 × 10^{6}/litre, often with a lymphopenia. Hyponatraemia, hypoalbuminaemia, and abnormality of liver function tests are detected in over half of the cases. Other non-specific features may include raised blood urea and muscle enzymes, hypoxaemia, haematuria, and proteinuria. Gram staining of sputum typically shows few pus cells and no predominant pathogen; initial blood and sputum cultures are negative unless dual infection is present.

Radiographic features

Radiographic shadowing is usually homogeneous. Characteristically, radiographic deterioration occurs with spread of shadows both within the same lung and to the opposite side. (Fig. 1).

Fig. 1 Chest radiograph of a 58-year-old man who returned from a Mediterranean hotel holiday with legionella pneumonia. There is extensive, bilateral, homogeneous consolidation. He required assisted ventilation for worsening respiratory failure.

Clearance of pulmonary shadows in survivors is particularly slow, with only two-thirds of radiographs being clear within 3 months and some taking more than 6 months to clear.

Complications

A wide variety of complications has been reported. The most important complication is acute respiratory failure requiring assisted ventilation, which occurs in up to 20 per cent of cases. Cardiac complications including pericardial and myocardial involvement are well recognized. A wide variety of neurological complications has been reported, leading to the suggestion of a specific neurotoxin. Acute, but usually reversible, renal failure may be seen in severe disease. There is anecdotal evidence that full clinical recovery may be very slow.

Pontiac fever

This is the acute non-pneumonic form of legionella infection and presents as a short-lived, self-limiting, influenza-like illness, dry cough, but no localizing signs in the chest (Fig. 2). The attack rate is extremely high, with an incubation period of usually 36 to 48 h. Investigations and chest radiograph are normal, and the illness improves spontaneously, usually within 5 days.

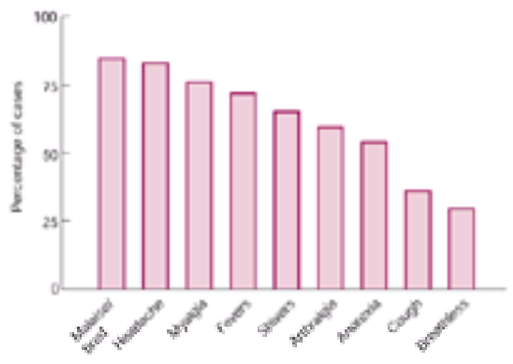

Fig. 2 Clinical features of 314 patients with non-pneumonic legionellosis (Pontiac fever). (Data adapted from Glick *et al.* (1978). *American Journal of Epidemiology* **107**, 149–60 and Goldberg *et al.* (1989). *Lancet* **i**, 316–18.)

Laboratory diagnosis

There is a range of laboratory procedures that can be used to diagnose legionellosis:

1. culture on a permissive medium, e.g. Aces buffered charcoal yeast-extract agar;

2. direct detection of bacteria or their nucleic acid;
3. urinary antigen detection;
4. serological response.

Suitable specimens from which legionellae can be isolated are expectorated sputum, endotracheal aspirates, bronchoalveolar lavage fluid, pleural aspirates, and lung. Isolation provides definite proof of infection, as colonization without infection has not been demonstrated. In addition it allows the causative strain to be typed and compared with those from the environment. A quicker diagnosis can be made by examining these samples directly for evidence of legionellae. With specific monoclonal antisera the bacteria can be visualized by immunofluorescence or immunoperoxidase techniques. Alternatively the use of the polymerase chain reaction can detect the bacterial nucleic acid directly in a specimen.

Soluble antigen is excreted in the urine for 1 to 3 weeks during the acute pneumonia and longer in immunocompromised infected patients. Tests to detect *L. pneumophila* serogroup 1 urinary antigen have a high specificity and sensitivity, making this assay the most valuable for the prompt diagnosis of legionnaires' disease. Urine antigen testing for legionella should be undertaken for any patient with severe community-acquired pneumonia.

Serology is the most widely used diagnostic approach. The major problem with serodiagnosis is the delay due to the slow production of antibodies. Only 20 per cent of patients with legionnaires' disease have diagnostic titres of antibody within 3 days of hospital admission, although about 40 per cent will have lesser but suggestive antibody levels by that time. Approximately 20 per cent of those infected appear not to respond serologically.

Although some heterologous cross-reacting antibodies may be produced, infections with legionellae other than *L. pneumophila* serogroup 1 do not necessarily give rise to antibody to the latter. Reference laboratories therefore use a battery of antigens to increase their ability to diagnose legionellosis. In Denmark, for example, 13 antigens are used and, in 1990, of 171 serologically diagnosed cases, 93 were *L. pneumophila* serogroup 1 to 6 while 78 were non-*L. pneumophila* legionellosis caused by other species. Occasionally patients with Q fever, leptospirosis, *Citrobacter freundii*, and more commonly with campylobacter infections make antibodies that cross-react with *L. pneumophila* serogroup 1. As diarrhoea can be an early feature of legionnaires' disease as well as the major consequence of campylobacter enteritis, it is important to culture stool samples and interpret with caution the legionella serology from such patients.

Differential diagnosis

Unfortunately there is no distinctive pattern that allows the early clinical differentiation of legionella infection from other, more common, causes of pneumonia. Epidemiological clues such as recent foreign travel can be valuable pointers, as well, of course, as knowledge of a local epidemic. Important clues suggesting legionella pneumonia in this context include high fever, confusion, multisystem involvement, absence of a predominant bacterial pathogen on sputum examination, and lack of response to b-lactam antibiotics.

Therapy

There are no clinical trials, and recommendations are based on retrospective case studies as well as *in vitro* and animal experiments. The most relevant factor is the ability of the antibiotic to penetrate intracellularly into alveolar macrophages where the legionella organism hides and divides. A macrolide such as erythromycin or clarythromycin is at present recommended as the drug of first choice, in dosages of 500 to 1000 mg every 6 h for erythromycin and 500 mg twice daily for clarithromycin, being given intravenously if required.

In vitro and animal experiments and clinical experience support the efficacy of rifampicin and fluoroquinolones. Rifampicin is often recommended as additional therapy to erythromycin, in a dose of 600 mg once or twice daily in patients with severe infection or who are deteriorating. Fluoroquinolones are preferred by some experts in immunocompromised patients. Anecdotal reports also support the use of doxycycline.

General supportive measures are particularly important, with attention to adequate hydration and correction of hypoxaemia with the early use of assisted ventilation for advancing respiratory failure.

Prognosis and mortality

The two most important factors affecting outcome include the prior health of the patient and appropriate, early therapy. The fatality rate in previously fit patients is low, of the order of 5 to 15 per cent, but in immunosuppressed individuals it can approach 75 per cent. The mortality is about 30 per cent in those requiring assisted ventilation.

Pathology

Legionellae are intracellular pathogens, both in protozoa in the environment and in animal hosts. Following infection, the bacteria are taken up by macrophages and internalized in macrophage endosomes. Legionellae block the development of the endosome to a phagolysosome which prevents the normal cellular killing mechanism of the ingested bacteria. Instead, legionellae multiply within the cell with a generation time of 1 to 2 h. When an intracellular microcolony has formed, the legionellae produce a pore-forming toxin which damages the cell's membrane and allows the bacteria to escape from the cell.

The lungs are usually the only organs affected in fatal cases and reveal lobar consolidation. Affected lung tissue shows a severe inflammatory response, with alveoli and terminal bronchioles distended by fibrin-rich debris, mononuclear inflammatory cells, and neutrophils. Organisms can be demonstrated within alveolar spaces by silver or immunofluorescence stains. In survivors alveolar and interstitial fibrotic changes can result.

Prevention

There are three aspects to consider in reducing the risk of legionellosis:

1. Measures to minimize colonization, growth, and release of legionellae into the atmosphere.
2. Physical or chemical treatment of water to kill the bacteria.
3. The protection of maintenance personnel who work on contaminated systems.

In Britain, particularly following the Stafford District General Hospital outbreak, a large number of publications aimed at minimizing the risk of legionellosis have appeared. In 1991 in Britain the Health and Safety Executive booklet HS(G)70 *The control of legionellosis including legionnaires' disease* was published, and it should be consulted for more details.

The most important principle to follow is to avoid holding water at temperatures between 20 and 45 °C, which is the range in which legionella multiplication occurs.

Further reading

Bhopal RS *et al.* (1991). Proximity of the home to a cooling tower and the risk of non-outbreak Legionnaires' disease. *British Medical Journal* **302**, 378–83.

Cunha BA (1998). Clinical features of legionnaires' disease. *Seminars in Respiratory Infections* **13**, 116–27.

Edelstein PH (1995). Antimicrobial chemotherapy for legionnaires' disease: a review. *Clinical Infectious Diseases* **21** (Suppl. 3), 5265–76.

Health and Safety Commission (2000). *Legionnaires' disease: the control of legionella bacteria in water systems. Approved code of practice and guidance L8.* HSE Books, Sudbury, Suffolk, UK.

Kwaik YA (1998). Fatal attraction of mammalian cells to *Legionella pneumophila. Molecular Microbiology* **30**, 689–95.

Ratcliffe RM *et al.* (1998). Sequence-based classification scheme for the genus Legionella targeting the *mip* gene. *Journal of Clinical Microbiology* **36**, 1560–7.

Woodhead MA, Macfarlane JT (1985). The protean manifestations of Legionnaires' disease. *Journal of the Royal College Physicians (London)* **19**, 224–30.

7.11.36 Rickettsial diseases including ehrlichioses

D. H. Walker

Introduction

Rickettsiae (Table 1) are obligate intracellular bacteria, which, during at least a part of their existence, occupy specific arthropods as their environmental niche. Rickettsiae are transmitted to man by their arthropod hosts and invade the endothelial cells of the blood vessel. In contrast, organisms of the genus *Ehrlichia* invade mainly phagocytes and do not cause primary vascular injury. Humans acquire *Coxiella burnetii* mainly by inhalation of aerosols from birth products of infected animals. The organisms proliferate within the acidic phagolysosome of host macrophages and cause an illness that ranges from acute atypical pneumonia to chronic endocarditis.

The public health importance of rickettsioses is underestimated because of difficulties with clinical diagnosis and lack of laboratory methods in many geographical areas. Active surveillance and serological surveys suggest that there is significant, unrecognized exposure to rickettsial organisms. It is particularly important to consider a rickettsial diagnosis when caring for the neglected poor of developing countries and travellers returning from areas endemic for murine typhus, scrub typhus, boutonneuse fever, African tick-bite fever, other spotted fevers, and Q fever. Rickettsiae infect previously healthy, active persons, and if undiagnosed, diagnosed late, or untreated, Rocky Mountain spotted fever, epidemic typhus, scrub typhus, Q fever endocarditis, boutonneuse fever (Plate 1, Plate 2, Plate 3), human ehrlichioses, and murine typhus are life threatening.

Many commonly prescribed antibiotics, including the penicillins, cephalosporins, and aminoglycosides, have no effect on the course of rickettsial diseases, but those antimicrobials active against rickettsial organisms can reduce morbidity and mortality.

Epidemics of louse-borne typhus fever have influenced the outcome of many wars between the 1500s and the 1920s. Wherever there are wars, famines, floods, and other massive disasters leading to widespread louse infestation of a population, the threat of epidemic typhus exists. Recent epidemics have occurred in Burundi, the economically devasted former USSR, and in extremely poor populations in the Andes.

Contemporary molecular analyses reveal that the spotted fever and typhus groups of the genus *Rickettsia* are very closely related to one another but not to *Orientia* (formerly *Rickettsia*)*tsutsugamushi*. They are relatively close relatives of *Ehrlichia* and the facultatively intracellular *Bartonella* and are evolutionarily distant from *Coxiella* and *Chlamydia*.

Vasculopathic rickettsial diseases of the spotted fever and typhus groups

Aetiological agents

These organisms measure approximately 0.3 by 1.0 µm and have a cell wall typical of Gram-negative bacteria.

Epidemiology

Seasonal incidence and geographical distribution are determined by the vector's activity. Spotted fever group rickettsiae are maintained in nature principally by transovarial and transstadial transmission in their tick or mite hosts. The most virulent rickettsiae are capable of killing their arthropod hosts (e.g. *R. prowazeki*) and require horizontal transmission to initiate epidemics of typhus fever. Reactivation of latent *R. prowazeki* infection in humans is the source for infection of lice that initiates epidemics of typhus fever.

Spotted fever group rickettsiae are transmitted to humans by secretion of infected tick saliva into the blood pooled in the site of the bite, and typhus group rickettsiae by infected louse or flea faeces deposited on human skin during arthropod feeding. Fluid or faeces of infected arthropods crushed between the fingers may enter a cutaneous wound or be rubbed into the conjunctiva.

Pathogenesis

Rickettsiae of some species of the spotted fever group frequently invade endothelial cells at the cutaneous portal of entry, proliferate, and cause a focus of dermal and epidermal necrosis, an eschar. Rickettsiae spread via the bloodstream to all parts of the body, where they infect endothelial cells lining the blood vessels. Typhus rickettsiae reach massive numbers intracellularly until the endothelial cell bursts. Spotted fever group rickettsiae are propelled through the cytosol by stimulating F-actin polymerization at one pole and spreading from cell to cell. Rickettsial lipopolysaccharides are non-endotoxic, and there is no evidence of any rickettsial exotoxin.

Host immune, inflammatory, and coagulation systems are activated with apparent overall benefit to the patient.

Progressive, disseminated infection and injury to endothelial cells cause increased vascular permeability, oedema, hypovolaemia, and signs and symptoms resulting from multifocal vascular injury in affected organs (Fig. 1). Infection of the pulmonary microcirculation and the resulting increased vascular permeability produce adult respiratory distress syndrome. Despite an interstitial myocarditis, myocardial function is usually preserved. Arrhythmias may result from vascular lesions affecting the conduction system. The vascular lesions in the brain are associated with coma and seizures in severe cases (Fig. 2). Multifocal infectious lesions in the dermis are the basis for the maculopapular, sometimes petechial, rash. Acute renal failure occurs in severe cases, usually as prerenal azotaemia or less frequently as acute tubular necrosis associated with severe hypotension.

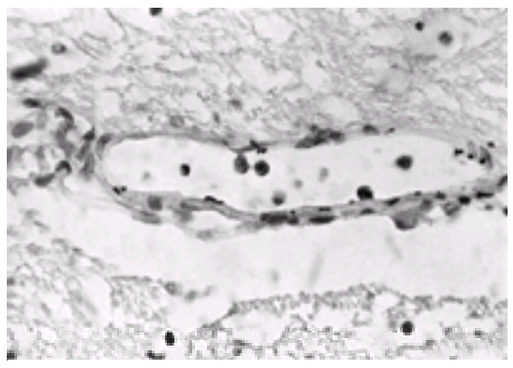

Fig. 1 Immunoperoxidase-stained *Rickettsia rickettsi* appear as dark bacilli in endothelial cells of a cerebral blood vessel with perivascular oedema but no host immune-cell infiltration.

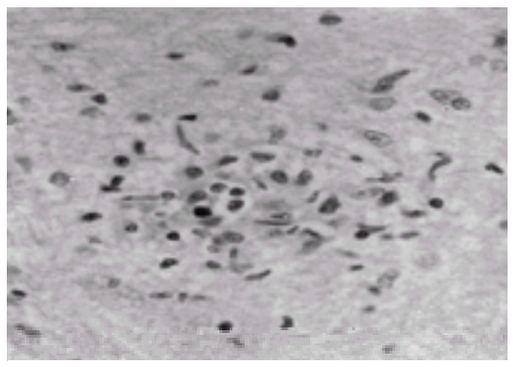

Fig. 2 Epidemic typhus fever. The typical lesion of rickettsial encephalitis is exemplified by the typhus nodule in the brain of a patient (death about 12th day) showing perivascular infiltration by macrophages and lymphocytes. (Reproduced from *Medical Clinics of North America* (1959),**43**, 1512, with permission.)

Clinical manifestations

The incubation period averages 1 week (range: 4 days to 2 weeks) after cutaneous inoculation. It is related inversely to the dose of inoculum.

Symptoms start with non-specific malaise, chills, fever, myalgia, and headache that is often severe, followed by anorexia, nausea, vomiting, abdominal pain, photophobia, and cough. A rash usually appears after 3 to 5 days of illness. Initially, it consists of macular or maculopapular lesions, 1 to 5 mm in diameter, that blanch on pressure. Later petechiae appear.

Pulmonary involvement causes cough, pulmonary oedema, radiographic infiltrates, hypoxaemia, dyspnoea, and pleural effusions in severe cases. Neurological manifestations consist of lethargy, progressing to confusion, delirium, stupor, ataxia, coma, focal neurological signs, and seizures. There may be a cerebrospinal fluid pleocytosis of 10 to 100 cells/µl with variable proportions of mononuclear and polymorphonuclear leucocytes, and/or an increased protein concentration.

Although serum aminotransferases and bilirubin may be elevated, jaundice is observed in fewer than 10 per cent of patients, and hepatic failure does not occur. The white blood-cell count is usually normal. The acute-phase reaction occurs in many patients. Hypoalbuminaemia is the result of leakage of this plasma protein into the interstitial space because of increased permeability of the microcirculation. Hyponatraemia is most often the result of the appropriate secretion of antidiuretic hormone in response to the hypovolaemic state.

Diagnosis

Differential diagnosis

Early, before the rash appears, the differential diagnosis includes influenza, typhoid fever, enteroviral infection, and infectious diseases suggested by geographical exposure (e.g. malaria, Lassa fever, louse-borne relapsing fever). Nausea, vomiting, and abdominal pain may suggest infectious enterocolitis. Prominent abdominal tenderness has occasionally led to the differential diagnosis of acute surgical abdomen and to exploratory laparotomy. Cough and abnormalities of physical and radiographic examination of the chest may suggest bronchitis or pneumonia. Fever, seizures, coma, neurological signs, and abnormalities of the cerebrospinal fluid may lead to consideration of meningitis and arboviral or herpes viral encephalitis. If an eschar is detected, the differential diagnosis may include cutaneous anthrax, tularaemia, syphilis, and chancroid. Once a rash has developed, differential diagnosis includes meningococcaemia, toxic-shock syndrome, leptospirosis, disseminated gonococcal infection, secondary syphilis, measles, rubella, enteroviral exanthem, infectious mononucleosis, dengue, filoviral or arenaviral haemorrhagic fevers, idiopathic or thrombotic thrombo-cytopenic purpura, and immune-complex vasculitides (e.g. systemic lupus erythematosus). It is important to enquire about exposure to ticks, fleas, mites, and lice and to consider the seasonal occurrence and geographical exposure, but people are frequently unaware of their exposure to arthropods, and cases may occur outside of the seasonal peak.

Laboratory diagnosis

Serological tests are useful in confirming the diagnosis in the convalescent stage, but seldom detect specific antibodies during the first week of illness. At present, the generally available serological assays are an indirect immunofluorescent antibody test, indirect immunoperoxidase antibody test, and dot enzyme immunoassay. These tests detect antibodies that are cross-reactive within the spotted fever or typhus group. The Weil–Felix test should be replaced because of poor sensitivity and specificity unless, as in some underdeveloped areas, nothing else is available. Isolation of the aetiological rickettsia, the definitive diagnosis of an infectious disesae, is seldom attempted because of the biohazard and technical challenges.

Detection of rickettsiae by immunohistochemistry in skin requires the presence of a rash to determine the site for biopsy and has sensitivity of approximately 70 per cent and a specificity of 100 per cent in the hands of an experienced microscopist. An approach that can be employed even during the period of illness before the onset of rash is immunofluorescent staining of rickettsiae in circulating endothelial cells captured by a monoclonal antibody fixed to immunomagnetic beads. Polymerase chain reaction has not been very successful in diagnosing Rocky Mountain spotted fever early in the course of illness, but has proved useful in murine typhus, epidemic typhus, Japanese spotted fever, boutonneuse fever, *R. felis* infection, and scrub typhus. Treatment should never be withheld while awaiting the results of laboratory tests.

Treatment

Spotted-fever and typhus-group rickettsioses respond favourably to treatment with doxycycline (200 mg/day for adults and children heavier than 45 kg, and 4.4 mg/kg body weight per day for smaller children), tetracycline (2 g/day in four divided doses for adults, and 25 mg/kg body weight per day in four divided doses for children), or chloramphenicol (2 g/day in four divided doses for adults, and 50 mg/kg body weight per day in four divided doses for children). Ciprofloxacin (200 mg intravenously every 12 h, or 750 mg orally every 12 h), ofloxacin (200 mg orally every 12 h), and pefloxacin (400 mg intravenously or orally every 12 h) have been used successfully to treat boutonneuse fever. Epidemic typhus fever has been treated effectively under field conditions with a single, 200 mg dose of doxycycline. Treatment is generally continued for 2 or 3 days after defervescence to avoid relapse of the infection.

Intravenously administered doxycycline or chloramphenicol is employed when oral treatment cannot be used because of vomiting or coma. Chloramphenicol and

josamycin (3 g/day for 8 days) have been used to treat rickettsioses during pregnancy when the tetracyclines are contraindicated.

Seizures should be treated with anticonvulsants. Renal failure is managed by haemodialysis, and hypoxaemia associated with interstitial pneumonitis and adult respiratory distress syndrome may require oxygen and mechanical ventilation.

Prevention

Immunization

Immunity to reinfection with spotted-fever or typhus-group rickettsiae is quite strong, although some patients with epidemic typhus fever will develop recrudescence of latent *R. prowazekii* infection many years after their acute infection.

Vaccines containing whole, killed organisms can reduce severity of illness and mortality, but a live, attenuated vaccine against *R. prowazekii* confers protection. However, some vaccine recipients develop mild typhus fever, and the vaccine strain may revert to a pathogenic state. Thus, there are at present no vaccines in general use against rickettsial diseases.

Vector control

Delousing reduces the spread of louse-borne epidemic typhus. Rodent control and insecticides decrease the incidence of murine typhus and rickettsialpox.

Regular daily or twice-daily inspection of the entire body, especially the scalp and groin, and prompt removal of ticks prevents inoculation of rickettsiae. Ticks are best removed by grasping their anterior parts firmly with pointed forceps flush with the skin and exerting steady traction until the intact tick is removed, frequently with a bit of attached skin. Care should be taken to avoid introduction of potentially infected tick fluids into the wound or mucous membranes.

Spotted fevers

Boutonneuse fever

Aetiology

The most prevalent spotted-fever rickettsiosis in Europe is boutonneuse fever, or Mediterranean spotted fever. *Rickettsia conorii* has been isolated in Spain, France, Italy, Croatia, Georgia, Russia, Ukraine, India, Pakistan, South Africa, Kenya, Somalia, and Ethiopia. *R. conorii* has more antigenic diversity than the other carefully analysed spotted-fever group rickettsia, *R. rickettsii*.

Epidemiology

The incidence of boutonneuse fever underwent a dramatic increase in Spain, France, Italy, and Portugal a decade or two ago. *R. conorii* is maintained transovarially in *Rhipicephalus sanguineus* and is transmitted to humans by tick bite. The peak incidence along the Mediterranean coast of southern Europe is in July and August when immature stages of the tick predominate.

Mortality rates of 1.4 to 5.6 per cent have been observed in patients admitted to hospital. In those who are elderly or have underlying diseases, suffer from alcoholism, or glucose-6-phosphate dehydrogenase deficiency, case fatality may be 33 per cent.

Pathogenesis

Reduction in the number of rickettsiae in the *tache noire* (black spot) or eschar at the site of the infective tick-bite eschar is associated with a perivascular influx of lymphocytes and macrophages. Autopsies of fatal cases of boutonneuse fever show systemic vascular infection and injury by *R. conorii*, with lesions in the brain, meninges, lungs, kidney, gastrointestinal tract, liver, pancreas, heart, spleen, and skin including sites of peripheral gangrene. Direct rickettsial injury of infected endothelial cells is the major pathogenic event. Hepatic biopsies show multifocal dead hepatocytes with a predominantly mononuclear cellular response.

Clinical manifestations

During the incubation period of boutonneuse fever, a red papule appears at the site of the tick bite and progresses to an eschar in approximately 70 per cent of cases, often associated with regional lymphadenopathy. The illness starts with fever, sometimes accompanied by headache and myalgias. The rash usually appears on the fourth day of illness as maculopapules, is petechial in 10 per cent of patients, and often involves the palms and soles. Other features include nausea, vomiting, cough, dyspnoea, conjunctivitis, stupor, meningismus, and hepatomegaly. Increased vascular permeability manifests as mild oedema, hypoalbuminaemia, and arterial hypotension. The white blood-cell count is usually normal. Platelet counts less than 100×10^{9}/l are detected in 12.5 per cent of the patients. Hyponatraemia of less than 130 mmol/l occurs in 23 per cent, and hypoproteinaemia is observed in 23 per cent of patients.

Serum urea and creatinine concentrations are elevated in 25 and 17 per cent of patients, respectively. Serum concentrations of aspartate and alanine aminotransferases are increased in 39 and 37 per cent, respectively, and serum bilirubin is greater than 20 µmol/l in 9 per cent. Severe features (6 per cent of patients) include cutaneous purpura and other haemorrhagic phenomena, neurological signs, altered mental status, respiratory symptoms, hypoxaemia, and acute renal failure.

Diagnosis

In the acute stage, diagnosis can be established by immunohistological demonstration of *R. conorii* in a biopsy of the *tache noire* or rash, or in circulating endothelial cells (see above). *R. conorii* can be isolated in cell culture. Serological methods include immunofluorescent antibody assay, latex agglutination test, indirect immunoperoxidase assay, dot enzyme immunoassay, and complement fixation test.

Treatment

See above.

Prevention

There is no vaccine to protect against *R. conorii*.

Rocky Mountain spotted fever

R. rickettsii is pathogenic for *Dermacentor* ticks, perhaps explaining why fewer than 1 in 1000 ticks in endemic areas contains this organism. *R. rickettsii*, the most virulent rickettsia, is also more invasive than other rickettsial species, causing infection not only of endothelial cells but also vascular smooth-muscle cells. Host factors also play a part in severity of illness. Fatality rates are higher in older patients, males, and black people. Fulminant Rocky Mountain spotted fever with death occurring within 5 days after onset is associated with haemolysis, particularly in black males with glucose-6-phosphate dehydrogenase deficiency.

Untreated, Rocky Mountain spotted fever has a 20 per cent fatality rate. In recent series, the death rate has been 5 per cent, with respiratory failure in 12 per cent, acute renal failure in 14 per cent, and anaemia requiring red-cell transfusion in 11 per cent. Thrombocytopenia occurs in 32 to 52 per cent of patients. Coma is a grave prognostic sign.

Early in the illness, nausea or vomiting occurs in 38 to 56 per cent of cases and abdominal pain in 30 to 34 per cent. The rash usually appears on the third day of illness, but may be delayed to or after day 6 in 20 per cent (Fig. 3). In 10 per cent of patients, a rash never appears. Petechiae occur in only 41 to 59 per cent of cases and appear late in the course, often only on or after day 6. The palms and soles are affected by the rash in 36 to 82 per cent with involvement often occurring after day 5 (Fig. 4).

Fig. 3 The early rash of Rocky Mountain spotted fever consists of pink macules in this 4-year-old boy on the fourth day of illness.

Fig. 4 Rocky Mountain spotted fever. Series showing haemorrhagic exanthem in a 4-year-old boy on about the eighth day of illness. Note oedema of face, hands, arms, and feet, and bleeding from mouth. Specific therapy with chloramphenicol resulted in complete recovery.

A history of tick exposure is obtained from only 60 per cent of patients. Reagents for indirect immunofluorescent antibody assay, latex agglutination test, and dot enzyme immunoassay for antibodies to *R. rickettsii* are commercially available.

Rickettsialpox

R. akari has been isolated in the United States, Ukraine, Croatia, and Korea. It is maintained by transovarian transmission in the gamasid mite *Liponyssoides sanguineus*, whose host is the domestic mouse, *Mus musculus*.

A cutaneous papule appears during the incubation period at the site where the mite has fed and evolves into an eschar over the next 2 to 7 days. About 10 days later, malaise, fever, chills, severe headache, and myalgia develop. A macular rash of discrete erythematous lesions, 2 to 3 mm in diameter, appears 2 to 6 days later and evolves into maculopapules, some of which develop central, deep-seated vesicles.

Other spotted-fever rickettsioses

R. sibirica, *R. australis*, *R. japonica*, *R. honei*, and *R. africae* differ antigenically in their surface proteins, DNA sequences, tick hosts, and known geographical distribution, but their clinical manifestations are similar to boutonneuse fever. The spotted-fever rickettsiosis of Flinders Island, Australia and Queensland tick typhus are clinically similar. Israeli spotted fever is a variant of boutonneuse fever in which eschar formation is usually lacking. *R. africae* is associated with a relatively less severe illness in which rash is less prevalent, but there are often multiple eschars. *Rickettsia slovaca*, previously considered as non-pathogenic, has recently been associated with clinical illness in Europe. *R. felis* is maintained in cat fleas and causes an emerging infectious disease.

In Sweden, *R. helvetica*, transmitted by *Ixodes ricinus* ticks, has been suggested to cause fatal chronic perimyocarditis. *R. conorii*, *R. typhi*, and *R. rickettsii* are also known to affect the microcirculation of the myocardium.

Typhus fevers

Murine typhus

Endemic flea-borne typhus fever caused by *R. typhi* is more prevalent in warm, coastal ports. It is maintained in a commensal cycle involving rat fleas, *Xenopsylla cheopis*, and rats, *Rattus rattus* and *Rattus norvegicus*. Rats are infected by *R. typhi* in flea faeces deposited on the skin. Fleas become infected for life after a rickettsaemic blood meal. Other species of fleas and other mammals can also maintain an infectious cycle of *R. typhi* (see Table 1).

A rash is detected in 80 per cent of fair-skinned and 20 per cent of black people. Other features include nausea (48 per cent), vomiting (40 per cent), abdominal pain (23 per cent), diarrhoea (26 per cent), cough (35 per cent), abnormal chest radiographs (23 per cent), thrombocytopenia (48 per cent), elevated serum aminotransferases (90 per cent), and central nervous abnormalities (8 per cent) including confusion, stupor, and hallucinations. Nearly 10 per cent of patients admitted to hospital are severely ill with acute renal failure, respiratory failure, or severe neurological abnormalities including seizures. Older age, delayed treatment, and initial treatment with sulphonamides are risk factors for severe disease. Case fatality is 1 to 2 per cent.

Epidemic typhus, recrudescent typhus, and sylvatic typhus

R. prowazekii causes epidemic louse-borne typhus fever, recrudescence of latent infection years after acute epidemic typhus, and zoonotic infection acquired from the ectoparasites of infected flying squirrels in North America. There is intense headache, prostration, continuous high fever, a macular rash usually appearing on the fourth or fifth day of illness, myalgia, and neurological abnormalities. Within 24 to 48 h of its appearance, the rash becomes petechial and does not blanch on pressure (Fig. 5). Its development is centrifugal from the trunk to the extremities. Other symptoms include cough, rales (71 per cent), nausea (30 per cent), abdominal pain (30 per cent), mental dullness (14 per cent), delirium (48 per cent), coma (6 per cent), seizures (1 per cent), and gangrene (3 per cent).

Fig. 5 Epidemic typhus fever. Typical truncal rash in louse-borne typhus on about the eighth day of illness showing many discrete haemorrhagic lesions.

The infection is now restricted to a few foci of sporadic occurrence in eastern Europe, central Africa, Ethiopia, southern Africa, Afghanistan, northern India, China, Mexico, Central America, and the Andes Mountains of South America. However, the danger of spread still exists as occurred recently during the war in Burundi where it is estimated that 100,000 cases occurred.

Recrudescent typhus (Brill–Zinsser disease) is the most important reservoir for initiation of epidemic louse-borne typhus in a susceptible population. Clinically it is milder than epidemic typhus (Fig. 6).

Fig. 6 Recrudescent typhus (Brill–Zinsser disease). Note the erythematous macular rash on the trunk. Illness is in an adult whose initial infection with typhus was 30 years earlier in Poland; the second attack was a week after appendicectomy and there was full recovery.

Ehrlichial diseases

Aetiological agents

Ehrlichiae are small, Gram-negative obligately intracellular bacteria that reside in a cytoplasmic vacuole and are transmitted by ticks. The four established human pathogens are *Ehrlichia chaffeensis*, *Anaplasma phagocytophila*, *E. ewingii*, and *Neorickettsia sennetsu*.

Ehrlichiae enter the host cell via phagocytosis, and they actively inhibit fusion of lysosomes with the phagosomes. They undergo binary fission to form clusters within a host vacuole. When stained by the Wright–Giemsa method, the cluster of organisms appears dark violet-blue and stippled and is called a morula from the Latin word for mulberry.

Epidemiology (see Table 1)

Ehrlichioses are maintained in cycles involving a mammalian host and a tick vector. *N. sennetsu* is a member of a genus which resides in flukes that parasitize fish and snails.

Most patients recall a recent tick bite. *E. chaffeensis* and *A. phagocytophila* infections peak between May and July in the United States, the season of greatest tick activity.

Human ehrlichioses

Haemopoietic cells are the primary targets of infection by *E. chaffeensis*, *A. phagocytophila*, and *E. ewingii*. Leucopenia and thrombocytopenia are probably caused by peripheral sequestration. Perivascular lymphohistiocytic infiltrations without vascular damage are observed in virtually any organ, including meninges. Hepatic involvement may include focal hepatocellular apoptosis and granulomas. Interstitial mononuclear pneumonitis has been observed, as well as diffuse alveolar damage.

Clinical severity ranges from asymptomatic seroconversion to fatal infection, very likely related to both host and microbial virulence factors. *E. chaffeensis* is the most pathogenic. Most patients have a fever, headache, chills, malaise, nausea, myalgias, and anorexia. Respiratory or renal insufficiency and abnormalities of the central nervous system have been reported. Pleocytosis, leucopenia, thrombocytopenia, and elevations in serum hepatic aminotransferases are the most often demonstrated clinical laboratory abnormalities. Overwhelming, often fatal, cases occur in immunosuppressed patients, particularly those with HIV-1 infection.

Differential diagnoses include Rocky Mountain spotted fever, meningococcaemia, bacterial sepsis, and infective endocarditis.

Human isolates of *E. chaffeensis* and *A. phagocytophila* have been obtained in cell culture of a dog histiocytoma cell line and a human promyelocytic leukemia cell line, respectively. Ehrlichiae, particularly *A. phagocytophila*, can sometimes be seen in peripheral white blood cells. The standard diagnostic test is indirect immunofluorescent antibody assay. A fourfold rise or fall in titre with a peak of 64 or greater is considered diagnostic. Human ehrlichiosis can also be diagnosed by detection of ehrlichial DNA amplified from the peripheral blood by polymerase chain reaction. Although these tick-borne human ehrlichioses were discovered only recently in the United States, human infections with *A. phagocytophila* have been reported in Europe.

Human ehrlichioses respond to treatment with doxycycline (200 mg/day in two divided doses) or tetracycline (25 mg/kg body weight per day in four divided doses).

Prevention is by avoiding tick bites.

Further reading

Bakken JS *et al*. (1996). Clinical and laboratory characteristics of human granulocytic ehrlichoisis. *Journal of the American Medical Association* **275**, 199–205.

Buller RS *et al*. (1999). *Ehrlichia ewingii*, a newly recognized agent of human ehrlichiosis. *New England Journal of Medicine* **341**, 148–55.

Dumler JS, Taylor JP, Walker DH (1991). Clinical and laboratory features of murine typhus in South Texas, 1980 through 1987. *Journal of the American Medical Association* **266**, 1365–70.

Elghetany MT, Walker DH (1999). Hemostatic changes in Rocky Mountain spotted fever and Mediterranean spotted fever. *American Journal of Clinical Pathology* **112**, 159–68.

Kass EM *et al*. (1994). Rickettsialpox in a New York City hospital, 1980 to 1989. *New England Journal of Medicine* **331**, 1612–17.

LaScola B, Raoult D (1997). Laboratory diagnosis of rickettsioses: current approaches to diagnosis of old and new rickettsial diseases. *Journal of Clinical Microbiology* **35**, 2715–27.

Lotric-Furlan S *et al*. (1998). Human granulocytic ehrlichiosis in Europe: clinical and laboratory findings for four patients from Slovenia. *Clinical Infectious Diseases* **27**, 424–8.

McDade JE, Newhouse VF (1986). Natural history of *Rickettsia rickettsii. Annual Review of Microbiology* **40**, 287–309.

Nilsson K, Lindquist O, Pahlson C (1999). Association of *Rickettsia helvetica* with chronic perimyocarditis in sudden cardiac death. *Lancet* **354**, 1169–73.

Perine PL *et al.* (1992). A clinico-epidemiological study of epidemic typhus in Africa. *Clinical Infectious Diseases* **14**, 1149–58.

Raoult D, Brouqui P (1999). *Rickettsiae and rickettsial disease at the turn of the third millenium*. Elsevier, Paris.

Raoult D *et al*. (1986). Mediterranean spotted fever: clinical, laboratory and epidemiological features of 199 cases. *American Journal of Tropical Medicine and Hygiene* **35**, 845–50.

Rikihisa Y (1991). The tribe *Ehrlichieae* and ehrlichial diseases. *Clinical Microbiology Reviews* **4**, 286–308.

Walker DH, Dumler JS (1996). Emergence of ehrlichiosis as human health problems. *Emerging Infectious Diseases* **2**, 18–29

Walker DH, Dumler JS (1997). Human monocytic and granulocytic ehrlichioses. Discovery and diagnosis of emerging tick-borne infections and the critical role of the pathologist. *Archives of Pathology and Laboratory Medicine* **121**, 785–91.

Walker DH, Fishbein DB (1991). Epidemiology of rickettsial diseases. *European Journal of Epidemiology* **7**, 237–45.

7.11.37 Scrub typhus

George Watt

Scrub typhus, or tsutsugamushi fever, is a zoonosis of rural Asia and the western Pacific islands. The causative organism, *Orientia* (formerly *Rickettsia*) *tsutsugamushi*, is transmitted to humans by the bite of a larval *Leptotrombidium* mite (chigger). An eschar and regional lymphadenopathy often develop at the site of infection, and may by followed by a systemic illness ranging in severity from inapparent to fatal. Many cases go undiagnosed, particularly those in which an eschar cannot be found. Rapid non-microscopic diagnostic tests are available and should enable more *O. tsutsugamushi* infections to be diagnosed.

Aetiology and epidemiology

Orientia tsutsugamushi has a different cell wall structure and genetic makeup from rickettsiae but looks like a rickettsia under light microscopy. The organism is an obligately intracellular Gram-negative bacterium. There are multiple serotypes of *O. tsutsugamushi*, and infection with one type confers only transient cross-immunity to another. Scrub typhus is a zoonosis. Larval mites (of the *Leptotrombidium deliense* group) usually feed on small rodents, particularly wild rats of the subgenus *Rattus*. Man becomes infected when he accidentally encroaches in a zone where there are infected mites. These zones are often made up of secondary or 'scrub' growth, hence the term scrub typhus. However, mite habitats as diverse as seashores, rice fields, and semideserts have been described. Infected chiggers are generally found in only very circumscribed foci within these zones. Large numbers of cases can occur when humans enter these so-called 'mite islands.' Disease transmission occurs when infected mites burrow into the skin, take a meal of tissue fluid, and inoculate the infectious organisms. Human to human transmission of scrub typhus via contaminated blood has never been documented. The endemic area forms a triangle bounded by northern Japan and southeastern Siberia to the north, Queensland, Australia, to the south and Pakistan to the west (Fig. 1). Disease transmission occurs in rural and suburban areas as well as in villages, but inhabitants of city centres are not at risk.

Fig. 1 Geographical distribution of scrub typhus.

Pathology and pathogenesis

Much remains unknown about the pathogenesis of scrub typhus, partly because most descriptions of severe cases pre-date advances made in immunohistology since the 1950s. Marked geographical variations in severity of the illness occur but determinants of severity are poorly characterized. Strains which differ in virulence, partial immunity, and regional differences in general health could affect disease presentation, but coinfection with the HIV-1 virus does not. Scrub typhus is a vasculitis, but clinical and pathological findings do not correlate closely. The host cell of *O. tsutsugamushi* in humans is thought to be the endothelial cell because of findings in experimental animals and by analogy with other rickettsial infections. However, in human liver infected with scrub typhus examined by electron microscopy, organisms predominate in Kupffer cells and hepatocytes rather than within endothelial cells (Fig. 2). *O. tsutsagamushi* is present in peripheral white blood cells of patients with scrub typhus. The HIV-1 viral load falls markedly in some AIDS patients who acquire acute *O. tsutsagamushi* infection. Some sera from HIV-seronegative patients with scrub typhus inhibit HIV replication *in vitro*.

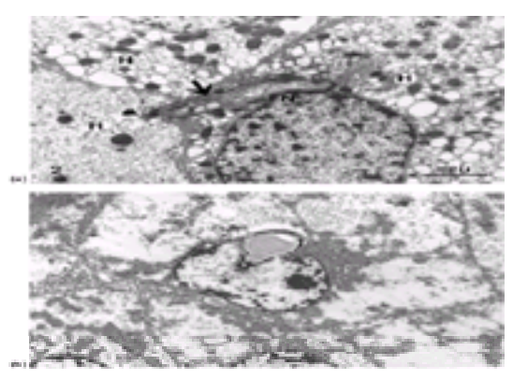

Fig. 2 *Orientia tsutsugamushi* in human liver visualized by electron microscopy (by courtesy of Dr Emsri Pongponratn). (a) Three hepatocytes (H) and a perinuclear scrub typhus organism (arrow) attached to the nuclear membrane. (b) *O. tsutsugamushi* piercing a hepatocyte nuclear membrane.

Clinical features

The painless chigger bite can occur on any part of the body, but is often in difficult to see in locations such as under the axilla or in the genital area. An eschar (Plate 1) forms at the bite site in about half of primary infections, but in a minority of secondary infections. The eschar begins as a small, painless papule which develops during the 6- to 18-day (median 10 days) incubation period. It enlarges, undergoes central necrosis, and acquires a blackened scab to form a lesion resembling a cigarette burn. Regional lymph nodes are enlarged and tender. The eschar is usually well developed by the time fever appears and is often healing by the time the patient presents to hospital.

Fever, headache, myalgia, and non-specific malaise are common symptoms. Hearing loss concurrent with fever is reported by as many as one-third of patients and is a useful diagnostic clue. Conjunctival suffusion and generalized lymphadenopathy are common, helpful physical signs. A transient macular rash may appear at the end of the first week of illness but is often difficult to see. The rash first appears on the trunk and becomes maculopapular as it spreads peripherally. Cough sometimes accompanied by infiltrates on the chest radiograph is one of the commonest presentations of *O. tsutsugamushi* infection. In severe cases, tachypnoea progresses to dyspnoea, the patient becomes cyanotic and full-blown adult respiratory distress syndrome may ensue. Apathy, confusion, and personality changes

frequently occur and only rarely progress to stupor, convulsions, and coma. Abnormalities resolve completely in non-fatal cases.

Diagnosis

The eschar is the single most useful diagnostic clue, and is pathognomonic when seen by a physician experienced in diagnosis of scrub typhus. Even typical eschars can be overlooked or misdiagnosed, however, and atypical presentations are common. Eschars in the genital area often lose their crust and can be confused with the ulcers of chancroid, syphilis, or lymphogranuloma venereum.

There is no constellation of laboratory test results which strongly suggests *O. tsutsugamushi* infection. Slight increases in the number of circulating white blood cells are common. Atypical lymphocytes and moderately elevated serum transaminase levels are not uncommon. Laboratory findings are chiefly useful to rule out other infections. A low white cell count and thrombocytopenia with a haemorrhagic rash suggest infection with dengue virus rather than *O. tsutsugamushi*. Raised serum creatinine and serum bilirubin levels with marked myalgia suggest leptospirosis rather than scrub typhus. Enteric fever rarely causes generalized lymphadenopathy or conjunctival suffusion.

The Weil–Felix test using the Proteus OX-K antigen is a commercially available serodiagnostic test which has been used for many years, but is insensitive. Immunofluorescent assay and the immunoperoxidase test are the confirmatory tests of choice but their complexity limits their use to a small number of reference centres. An accurate, rapid, dotblot immunoassay which does not require a microscope has been developed (Plate 2). Such kits would be of enormous benefit if they could be made affordable for use in rural tropical Asia where most scrub typhus cases occur.

Treatment

Prompt antibiotic therapy shortens the course of the disease and reduces mortality. Treatment must often be presumptive, but the benefits of avoiding severe scrub typhus by early antibiotic administration generally far outweigh the risks of a 1-week course of tetracycline—the treatment of choice. Either oral tetracycline 500 mg four times daily, or oral doxycycline 100 mg twice daily for 7 days are recommended. Oral chloramphenicol 500 mg four times a day is a cheaper alternative. Treatment for less than a week is initially curative, but may be followed by relapse. Parenteral doxycycline should be administered to patients who cannot swallow tablets or who are severely ill. A 7-day course of parenteral chloramphenicol (50–75 mg/kg/day) is an effective alternative in areas where parenteral formulations of tetracyclines are unavailable. Good supportive care and early detection of complications is important in severe cases if a good outcome is to be obtained.

Scrub typhus cases from northern Thailand which respond poorly to conventional therapy have been described, but neither the mechanism of resistance nor its geographical distribution have been defined. A controlled, blinded study demonstrated that patients treated with rifampin in northern Thailand became afebrile twice as quickly as did patients who received doxycycline. However, the optimum therapeutic regimen for the treatment of drug-resistant scrub typhus has not yet been determined. Therapy for pregnant women and children poses several problems. Chloramphenicol is best avoided during pregnancy and cannot be given to neonates; tetracycline is contraindicated in pregnancy and long courses administered to young children cause staining of the permanent teeth. Newer macrolide antibiotics appear to be effective for scrub typhus. Cases of both drug-sensitive and drug-resistant scrub typhus have been cured by azithromycin and three Japanese patients were treated successfully with clarithromycin. If their efficacy is confirmed, macrolides would be particularly useful for the treatment of infection during pregnancy and early childhood.

Prevention and control

Weekly doses of 200 mg of doxycycline can prevent *O. tsutsugamushi* infection. Chemoprophylaxis should be considered for non-immunes sent to an enzootic area to perform work which places them at high risk of acquiring scrub typhus. Soldiers and road construction crews are typical examples, but chemoprophylaxis should also be considered in high-risk travellers such as trekkers. Contact with chiggers can be reduced by applying repellant to the tops of boots, socks, and on the lower trousers and by not sitting or lying directly on the ground. Unfortunately these measures are frequently not practicable in those exposed occupationally. There is no vaccine for scrub typhus.

Prognosis

Scrub typhus was a dreaded disease in the preantibiotic era; case fatality rates reached as high as 50 per cent. Prompt antibiotic therapy generally prevents death, but up to 15 per cent of patients still die in northern Thailand. Deaths are attributable to a variety of factors including late presentation, delayed diagnosis, and drug resistance.

Further reading

Chayakul P, Panich V, Silpapojakul K (1988). Scrub typhus pneumonitis:an entity which is frequently missed. *Quarterly Journal of Medicine* **256**, 595–602.

Kantipong P *et al.* (1996). HIV infection does not influence the clinical severity of scrub typhus. *Clinical Infectious Diseases* **23**, 1168.

Olson JG *et al.* (1980). Prevention of scrub typhus. Prophylactic administration of doxycycline in a randomized double blind trial. *American Journal of Tropical Medicine and Hygiene* **29**, 989.

Pongponratn E *et al.* (1998). Electron microscopic examination of *Rickettsia tsutsugamushi*-infected human liver. *Tropical Medicine and International Health* **3**, 242–8.

Silpapojakul K *et al.* (1991). Scrub and murine typhus is children with obscure fever in the tropics. *International Journal of Systematic Bacteriology* **10**, 200–3.

Silpapojakul K *et al.* (1991). Rickettsial meningitis and encephalitis. *Archives of Internal Medicine* **151**, 1753–7.

Watt G *et al*. (1996). Scrub typhus infections poorly responsive to antibiotics in Northern Thailand. *Lancet* **348**, 86–9.

7.11.38*Coxiella burnetii* infections (Q fever)

T. J. Marrie

History

In August 1935, Dr Edward Holbrook Derrick, Director of the Laboratory of Microbiology and Pathology of the Queensland Health Department in Brisbane, Australia, was asked to investigate an outbreak of undiagnosed febrile illness among workers at the Cannon Hill abattoir. Derrick realized that he was dealing with a type of fever that had not been previously described—he named it Q (for query) fever. A couple of years later, Sir Frank Macfarlane Burnet in Australia and Herald Rea Cox in the United States isolated the micro-organism responsible for Q fever.

Coxiella burnetii (Fig. 1)

This micro-organism, the sole species of its genus, has a Gram-negative cell wall and measures 0.3 × 0.7 µm. It is an obligate phagolysosomal parasite of eukaryotes that sporulates, stains well by the Gimenez stain, and multiplies by transverse binary fission. *C. burnetii* undergoes phase variation akin to the smooth to rough transition in some enteric Gram-negative bacilli. In nature and laboratory animals it exists in the phase-I state. Repeated passage of phase-I virulent organisms in embryonated chicken eggs lead to the conversion to phase-II avirulent forms. Antibodies to phase-I antigens predominate in chronic Q fever, while phase-II antibodies are higher than phase I in acute Q fever.

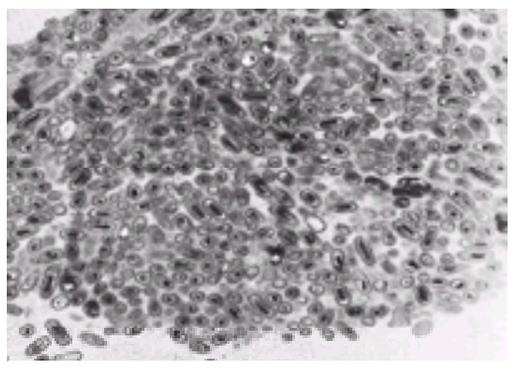

Fig. 1 Transmission electron micrograph showing *Coxiella burnetii* cells within a macrophage in the heart valve of a patient with Q fever endocarditis. The dark material in the centre of each cell is condensed DNA. 15 000 ×.

C. burnetii has survived for 586 days in tick faeces at room temperature, 160 days or more in water, in dried cheese made from contaminated milk for 30 to 40 days, and for up to 150 days in soil.

Epidemiology

Q fever is a zoonosis. There is an extensive wildlife and arthropod (mainly ticks) reservoir of *C. burnetii*. Domestic animals are infected through inhaling contaminated aerosols or by ingesting infected material. These animals rarely become ill but abortion and stillbirths may occur. *C. burnetii* localizes in the uterus and mammary glands of infected animals. During pregnancy there is reactivation of *C. burnetii* and it multiplies in the placenta, reaching 10^9 hamster infective doses per gram of tissue. The organisms are shed into the environment at the time of parturition. Man becomes infected after inhaling organisms aerosolized at the time of parturition, or later when organisms in dust are stirred up on a windy day. Infected cattle, sheep, goats, and cats are the animals primarily responsible for transmitting *C. burnetii* to man. There have been several outbreaks of Q fever in hospitals and research institutes due to the transportation of infected sheep to research laboratories. Some studies have suggested that ingestion of contaminated milk is a risk factor for the acquisition of Q fever; volunteers seroconverted but did not become ill after ingesting such milk.

Percutaneous infection, such as when an infected tick is crushed between the fingers, may occur but is rare. Transmission via a contaminated blood transfusion has rarely occurred.

Vertical transmission from mother to child has been infrequently reported. A 1988 review documents 23 cases of Q fever in pregnant women. These authors found that Q fever was present in 1 per 540 pregnancies in an area of endemic Q fever in Southern France.

Person-to-person transmission has been documented on a few occasions. To date, 45 countries on five continents have reported cases of Q fever. Q fever is estimated to cost $A1 million in Australia each year and results in the loss of more than 1700 weeks of work.

Clinical features

Man is the only animal known consistently to develop illness following infection with *C. burnetii*. There is an incubation period of about 2 weeks (range 2 to 29 days) following inhalation of *C. burnetii*. A dose–response effect has been demonstrated experimentally and clinically. *C. burnetii* is one of the most infectious agents known to man; a single micro-organism is able to initiate infection. The resulting illness in man can be divided into acute and chronic varieties.

Acute Q fever

Self-limiting febrile illness

The most common manifestation of acute Q fever is a self-limiting febrile illness.

Q fever pneumonia (Fig. 2 and Fig. 3)

This is the most commonly recognized manifestation of Q fever. There is often a seasonal distribution, most of the cases occurring between February and May. The onset is non-specific with fever, fatigue, and headache. The headache may be very severe, occasionally so severe that it prompts a lumbar puncture. A dry cough of mild to moderate intensity is present in 24 to 90 per cent of patients. About one-third have pleuritic chest pain. Nausea, vomiting, and diarrhoea occur in 10 to 30 per

cent of patients. Most cases of *C. burnetii* pneumonia are mild; however, about 10 per cent are severe enough to require admission to hospital; rarely, assisted ventilation is necessary. Death is rare in Q fever pneumonia and is usually due to comorbid illness. The white blood-cell count is usually normal, but is elevated in one-third of patients. Liver enzyme levels may be mildly elevated, at two to three times normal. Alkaline phosphatase is raised in up to 70 per cent of cases and 28 per cent are hyponatraemic. Reactive thrombocytosis is surprisingly common. Microscopic haematuria is a common finding.

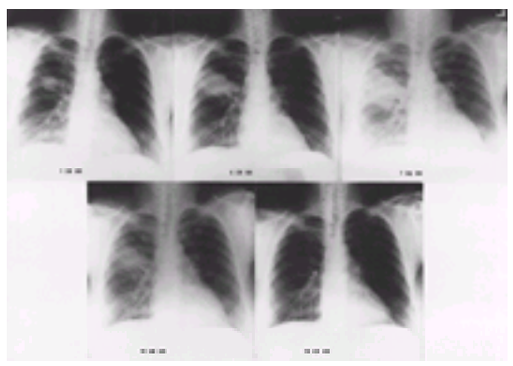

Fig. 2 Serial chest radiographs of a 35-year-old patient with Q fever pneumonia. The first radiograph (1 August 1989) shows a round opacity in the right upper lobe, which increases in size over the next 6 days. The pneumonia has completely cleared by 19 September 1989.

Fig. 3 Portable anteroposterior chest radiograph of a 72-year-old male with Q fever pneumonia. This radiographic picture is indistinguishable from pneumonia due to any other microbial agent.

The chest radiographic manifestations of Q fever pneumonia are usually indistinguishable from those of other bacterial pneumonias (Fig. 3). However, rounded opacities are suggestive of this infection (Fig. 2). Some investigators have reported delayed clearing of the pneumonia; however, in our experience resolution is usually complete within 3 weeks.

Hepatitis

The liver is probably involved in all patients with acute Q fever. There are three clinical pictures:

1. pyrexia of unknown origin with mild to moderate elevation of liver function tests;
2. a hepatitis-like picture—liver biopsy shows distinctive doughnut granulomas consisting of a granuloma with a central lipid vacuole and fibrin deposits;
3. 'incidental hepatitis'.

Neurological manifestations

Encephalitis, encephalomyelitis, toxic confusional states, optic neuritis, and demyelinating polyradiculoneuritis are uncommon manifestations of Q fever.

Rare manifestations

These include myocarditis, pericarditis, bone marrow necrosis, rhabdomyolysis, glomerulonephritis, lymphadenopathy, pancreatitis, mesenteric paniculitis, erythema nodosum, epididymitis, orchitis, priapism, and erythema annulare centrifugum.

Chronic fatigue may be a sequel of Q fever in some patients.

Chronic Q fever

The usual manifestation of chronic Q fever is that of culture-negative endocarditis. Some 70 per cent of these patients have fever and nearly all have abnormal native or prosthetic heart valves. Hepatomegaly and or splenomegaly occur in about half of these patients and one-third have marked clubbing of the digits. A purpuric rash due to immune complex-induced leucocytoclastic vasculitis and arterial embolism occurs in about 20 per cent of patients. Hyperglobulinaemia (up to 60 g/1) is common and is a useful clue to chronic Q fever in a patient with the clinical picture of culture-negative endocarditis.

Other manifestations of chronic Q fever include osteomyelitis, infection of aortic aneurysm, and infection of vascular prosthetic grafts.

The strains of *C. burnetii* that cause chronic Q fever do not differ from those that cause acute Q fever. Peripheral blood lymphocytes from patients with Q fever endocarditis are unresponsive to *C. burnetii* antigens *in vitro*, while responding normally to other antigens.

Diagnosis

A strong clinical suspicion based on the epidemiology and clinical features as outlined above is the cornerstone of the diagnosis of Q fever. This suspicion is confirmed by determining a fourfold or greater increase in antibody titre between acute and 2- to 3-week convalescent serum samples. A variety of serological tests are available: complement fixation, microimmunofluorescence, and enzyme immunoassay. The immunofluorescence antibody test is easiest to use. In acute Q fever the antibody titre to phase-II antigen is higher than that to phase-I antigen, while the reverse occurs in chronic Q fever. In chronic Q fever, antibody phase-I titres are extremely high, in the order of 1:8192 and higher. In acute Q fever, antibody titres to phase-I antigen are rarely in excess of 1:512, while peak antibody titres to phase-II antigen are between 1:1024 and 1:2048. The micro-organism can be isolated in embryonated eggs or in tissue culture; however, a biosafety level-3 laboratory is required. The polymerase chain reaction can be used to amplify *C. burnetii* DNA from tissues or other biological specimens.

Treatment

Acute Q fever is treated with a 2-week course of tetracycline or doxycycline. Chronic Q fever should be treated with two antimicrobial agents for at least 2 years. Some authorities recommend lifelong therapy for chronic Q fever. We use rifampicin, 300 mg twice a day, and ciprofloxacin, 750 mg twice a day, as agents of first choice. Rifampicin and doxycycline or tetracycline and trimethoprim–sulfamethoxazole have also been used to treat chronic Q fever. Another regimen for the treatment of

chronic Q fever is doxycycline 100 mg once daily and hydroxychloroquine 600 mg once daily to maintain a plasma level of between 0.8 and 1.2 µg/ml. This regimen is given for 18 months. Photosensitivity is a potential adverse reaction and patients should be warned to take preventive measures. In addition, an ophthalmologist must examine the optic fundus every 6 months for chloroquine accumulation. Antibody titres should be measured every 6 months for the first 2 years. A progressive decline in antibody titre reflects the successful treatment of chronic fever. Cardiac valve replacement may be necessary as part of the management of chronic Q fever.

Prevention

A formalin-inactivated *C. burnetii*, whole-cell vaccine is protective against infection and has a low rate of side-effects; 1 per cent of vaccinees developed an abscess at the inoculation site and another 1 per cent had a lump at this site 2 months after vaccination. The vaccine should be offered to those whose occupation places them at high risk for *C. burnetii* infection. Other measures to reduce Q fever infection are the use of only seronegative pregnant sheep in research facilities and the control of ectoparasites on livestock.

Further reading

Sawyer LA, Fishbein DB, McDade JE (1987). Q fever: current concepts. *Reviews of Infectious Diseases* **9**, 935–46.

7.11.39 Bartonelloses, excluding *Bartonella bacilliformis* infections

James G. Olson

Background

Bacteria belonging to the genus *Bartonella* cause human diseases, including verruga peruana (discussed elsewhere), cat-scratch disease, trench fever, and bacillary angiomatosis–peliosis. During the last two decades, the aetiology of cat-scratch disease was discovered, bacillary angiomatosis was recognized, and the relationship of these diseases to trench fever, an epidemic scourge of soldiers in the First World War, was demonstrated. Because of recent microbiological and genetic evidence, the aetiological agents of these diseases have been included in the genus *Bartonella*. Recent progress has lead to an improved understanding of the aetiology, diagnosis, and epidemiology of infections with *Bartonella*, and some improvements in patient care and prevention.

Cat-scratch disease

Introduction

Cat-scratch disease, in most patients, is an acute, self-limiting infection characterized by development of a papule at the site of inoculation by a cat, followed by regional adenopathy that may persist for 1 to 4 months. In a small percentage of patients, serious systemic complications may arise, including involvement of the central nervous system, liver, spleen, lung, bone, eyes, and skin.

Epidemiology

Cat-scratch disease occurs world-wide in all races, more often in males than in females. Most cases of cat-scratch disease occur in children, but the disease is rare in infants. Estimates of the proportion of cases occurring before the age of 18 years range from 55 to 87 per cent. In the United States, the estimated incidence of cat-scratch disease in ambulatory patients is nine cases/100 000 population. Some 0.8 cases/100 000 population are discharged from hospital with a diagnosis of cat-scratch disease. These data support earlier estimates and suggest that cat scratch disease affects about 24 000 people each year in the United States, resulting in approximately 2000 hospital admissions. Incidence in the United States is highest in humid southern states and lowest in arid western states. Most reported cases occur in the fall and winter, but patients can be infected during any season. About 90 per cent of patients have a history of exposure to cats. Cat-scratch disease is strongly associated with owning a kitten, particularly one with fleas, and the presence of a scratch or bite by a kitten. Although they remain asymptomatic, domestic cats serve as major persistent reservoirs for *B. henselae*. Blood samples cultured from pet and impounded cats suggest that more than 40 per cent of cats are bacteraemic. *B. henselae* was also detected in fleas taken from an infected cats by both direct culture and PCR. Cat fleas from infected cats readily transmit *B. henselae* to uninfected cats. *B. henselae* is readily transmitted to uninfected cats by the subcutaneous inoculation of infectious flea faeces. The cat flea certainly plays an indirect role in human disease by increasing the size of the feline reservoir, and a direct role by producing infectious faeces that are inoculated into the human via the scratch of the cat.

Aetiology

Serological, epidemiological, and molecular findings indicate that *B. henselae* is responsible for cat-scratch disease. *B. henselae* is a small, curved, pleomorphic, fastidious, Gram-negative rod that is oxidase and catalase negative, and X-factor dependent. Colony morphology is varied, ranging from small, dry, grey-white colonies to smooth, creamy-yellow colonies. Its slow-growing sensitivity to a broad range of commonly used antimicrobials (including ampicillin, tetracycline, trimethoprim–sulfamethoxazole, and aminoglycosides) does not always correlate with *in vivo* efficacy. It is most closely related to *B. quintana*, the louse-borne agent of trench fever. There are many newly described species in the genus *Bartonella* that have been recovered from animals, but only four, *B. henselae*, *B. quintana*, *B. elizabethae* (which was isolated from a single patient with endocarditis), and *B. bacilliformis* have been associated with human disease. *Afipia felis* was claimed to be the aetiological agent of cat-scratch disease but was probably a soil contaminant.

Pathology

Examination of the primary inoculation lesion demonstrates dermal necrosis with variable numbers of histiocytes and occasional multinucleate giant cells accompanied by scattered microabscesses with mixed inflammatory cells, including neutrophils, eosinophils, lymphocytes, and plasma cells. The epidermal changes are non-specific with parakeratosis, hyperkeratosis, oedema, and exocytosis of inflammatory cells.

Adenopathy

Early in the course of infection lymph nodes show reactive lymphoid follicular hyperplasia with initial minute microabscesses adjacent to the subcapsular sinus. As the disease progresses, characteristic histopathology is necrotizing granulomas with central microabscesses and palisading histiocytes. Most of the necrotic centres have a stellate configuration. Multinucleated giant cells in lymph nodes are either rare or absent. A perivascular neutrophilic infiltrate may be present. The Warthin–Starry or Steiner silver impregnation stains may reveal pleomorphic bacilli in clusters or short chains within the areas of central necrosis or around small vessels (Fig. 1). Although these histopathological features are characteristic of cat-scratch disease, they are not diagnostic and must be correlated with clinical findings and serological studies. Other infections, such as tularaemia, lymphogranuloma venereum, and fungal and mycobacterial infections, may have a similar histology.

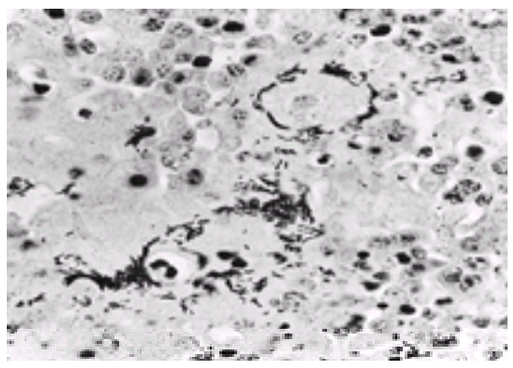

Fig. 1 Bacilli in tissue. Photomicrograph of Warthin–Starry silver impregnation stained section of an inguinal lymph node from a patient with a skin test positive for cat scratch disease. A vessel containing erythrocytes is cut in cross-section, bacilli are seen singly and in chains outlining the vessel (× 630). (Reproduced from Wear *et al*. (1983). *Science* **221**, 1403–5, Armed Forces Institute of Pathology negative 82–11271, with permission).

Clinical presentation

The typical course of cat-scratch disease begins with an erythematous papule or pustule at the inoculation site of a scratch or contact with a cat which usually persists for 1 to 3 weeks (Fig. 2). An inoculation site may be detected in over two-thirds of patients. Within 2 weeks, lymph nodes draining the site of inoculation become enlarged and tender. Lymphadenopathy occurs in more than 90 per cent of patients; it usually resolves spontaneously within a period of several months. In about 50 per cent, regional lymphadenitis is the only manifestation of the disease. Usually a single node or group of nodes is affected. The most common sites of lymphadenopathy are axillary, cervical, inguinal/femoral, and epitrochlear lymph nodes. Affected nodes are often tender and suppurate in about 10 per cent of the patients. Constitutional symptoms of fever, anorexia, malaise, and headache accompany the lymphadenitis in 75 per cent of patients, but in the vast majority these symptoms are mild. About one-third of patients complain of fever and one-quarter have malaise or fatigue. Other non-specific clinical features are headache, anorexia, weight loss, vomiting, sore throat, rashes (maculopapular and rarely erythema nodosum), and splenomegaly. Although considered to be a self-limiting illness, signs and symptoms of cat-scratch disease often persist for 2 to 4 months, and adenopathy for longer.

Fig. 2 Crusted erythematous papules at the site of a cat scratch above the umbilicus with bilateral inguinal lymphadenopathy, which developed 10 days later, in a 7-year-old boy (Copyright D.A. Warrell).

Atypical presentations occur in up to 15 per cent of patients. The most common, Parinaud's oculoglandular syndrome, was first described by Henri Parinaud in 1889. It is characterized by ocular granuloma or conjunctivitis with preauricular lymphadenopathy and fever. The affected eye is painless and non-pruritic and shows no evidence of discharge. Most patients recover spontaneously without any sequelae in 2 to 4 months. Other atypical manifestations include encephalopathy, aseptic meningitis, seizures, neuroretinitis, transverse myelitis, osteolytic lesions, hepatic and splenic granulomas, thrombocytopenic purpura, haemolytic anaemia, endocarditis, atypical pneumonia, pleural effusion, pulmonary nodules, breast mass, multiple granulomatous skin lesions, and recurrent adenopathy. In patients with central nervous system involvement, encephalopathy is the most commonly reported manifestation, occurring in 2 to 3 per cent of patients. Typically, 1 to 6 weeks after onset of lymphadenopathy, patients become abruptly confused and disoriented, rapidly progressing to coma. Cranial computed tomography is generally normal, and cerebrospinal fluid shows minimal pleocytosis or elevation of protein. Electroencephalography is frequently abnormal. Neurological recovery is almost always complete over 1 week, but persistent deficits have been reported. There have also been reports of patients presenting with recurrent fever, malaise, fatigue, and weight loss without obvious focal infection. The symptoms may persist for weeks to months before the diagnosis is made. Hepatic granulomas, osteomyelitis, and pulmonary involvement have also been reported as rare complications. All parts of the respiratory tract may be affected; bilateral hilar lymphadenopathy and primary atypical pneumonia have been reported. Severe manifestations have been described in an immunocompromised patient. Fatalities are extremely rare.

Diagnosis

The diagnosis of cat-scratch disease has evolved from a diagnosis by exclusion to one based on the laboratory confirmation of infection with the aetiological agent, *B. henselae.* The current case definition includes lymphadenopathy, with a serum IgG antibody titre more than 64 when tested by indirect immunofluorescence using *B. henselae* antigen; or PCR product specific for *B. henselae* as determined by RFLP or sequence analyses. Isolation of *B. henselae* is not practicable as viable bacteria are seldom present when the patient seeks medical care.

Differential diagnoses include lymphogranuloma venereum, syphilis, typical or atypical tuberculosis, other forms of bacterial adenitis, sporotrichosis, tularaemia, brucellosis, histoplasmosis, sarcoidosis, toxoplasmosis, infectious mononucleosis, and benign or malignant tumours.

Treatment

In the majority of patients, cat-scratch disease resolves spontaneously in 1 to 2 months. Azithromycin, rifampin, ciprofloxacin, trimethoprim–sulfamethoxazole, and gentamicin may benefit some patients. Antimicrobials should be considered for severe cases of cat-scratch disease but for uncomplicated cases of classical cat-scratch disease, treatment should be directed toward relief of discomfort. Application of moist soaks, local heat, analgesics, limitation of activity, and aspiration of suppuration may help to relieve the pain and resolve the inflammation. Aspiration is preferred to surgical drainage, which may lead to fistula formation or scarring. Spontaneous resolution of the infected node is common, but aspiration or surgical removal may be necessary. Healing is usually rapid. Treatment with erythromycin or doxycycline—either alone or in combination with rifampin—at standard doses but for longer duration (4 to 6 weeks) have been reported as effective and safe in both immunocompetent and immunosuppressed patients. Currently, the role of systemic steroid therapy, including in patients with neuroretinitis, is not clear.

Complications are uncommon and the prognosis is excellent. Recurrent attacks are rare and systemic sequelae are unusual.

Prevention

Isolation of patients is unnecessary. No vaccines are available, however, cat vaccines to protect cats from infection are under development. Treatments that prevent flea infestations in cats may be an effective means of preventing human infections. Cat owners should be encouraged to take their pets to routine veterinary visits and prevent ectoparasite infections, and to avoid cat scratches and bites. Cats implicated in transmission need not be destroyed.

Trench fever

Introduction

Trench fever is a febrile illness first described among British soldiers in 1915. From 1915 to 1918 it was thought to account from 40 to 60 per cent of all illnesses among soldiers. There were no deaths but much morbidity. By 1918 it was concluded that trench fever was an infectious disease and that the aetiological agent was transmitted by the human body louse. In 1961, *B. quintana* was isolated from the blood of a patient with trench fever and Koch's postulates for the causation of trench fever by *B. quintana* were fulfilled in 1969. Since the end of the Second World War, reports of trench fever have been rare but recent data suggest that cases may have escaped recognition; clusters of cases in homeless alcoholic men have been identified in the United States and France.

Epidemiology

Endemic foci of trench fever have been identified in Poland, the former Soviet Republics, Mexico, Bolivia, North Africa, Ethiopia, and Burundi, but its true incidence and geographical distribution are unknown.

B. quintana is transmitted by inoculation of contaminated louse faeces through a break in the skin from a louse bite or other injury. The incubation period is 7 to 30 days. It is not transmitted directly from person-to-person. The human body louse becomes infected by ingesting infected human blood.

In the 1980s, *B. quintana* re-emerged as an opportunistic pathogen among HIV-infected people in whom it causes bacillary angiomatosis, endocarditis, and bacteraemia. It has been isolated from AIDS patients in France and the United States. *B. henselae* is probably a more common cause of bacillary angiomatosis and bacteraemia among HIV-infected people.

More recently, *B. quintana* has been identified as a cause of invasive infection among HIV-seronegative, inner-city, homeless, alcoholics in Seattle, Washington, and Marseilles, France. A seroprevalence study conducted one year after the *B. quintana* outbreak among patients at a downtown Seattle clinic serving a primarily indigent and homeless population found that 20 per cent of patients had microimmunofluorescence titres at or greater than 64. Interpretation of these results is limited by the high cross-reactivity of the assay to *B. henselae*. The results suggest, however, that exposure to the organism was common in that population and that many cases of infection may have been asymptomatic or minimally symptomatic.

The mode of transmission of *B. quintana* among homeless persons is not well defined. Lice were detected on one patient at the time of presentation and five patients were reported to have been previously diagnosed with scabies. Transmission via a louse or other ectoparasitic vector is therefore a plausible hypothesis. Currently, the human body louse is the only known vector of *B. quintana*. No non-human vertebrate reservoir is known.

Clinical presentation

High fever is the most common clinical feature. Headache and myalgia are common prodromal symptoms. Fever starts acutely or insidiously and is often associated with headache, dizziness, and pains in the back, eyes, and legs, especially in the shins. Splenomegaly is common and a red macular rash (lesions 2–4 mm in diameter) may appear transiently. Complete recovery usually occurs within 5 to 6 weeks without antimicrobial therapy. Trench fever is not fatal but about half of the patients will have relapse of illness with fever and myalgia. Endocarditis has been described.

Four clinical patterns have been recognized:

1. asymptomatic or minimally symptomatic infection;
2. a single, acute, febrile attack lasting 3 to 4 days;
3. a periodic form, with multiple febrile paroxysms; and
4. a continuous form with weeks of fever.

Studies involving inoculation of humans with *B. quintana* from infected louse faeces suggest that the incubation period ranges from 5 to 20 days depending on the size of the inoculum. *B. quintana* may circulate in the blood for weeks after resolution of symptoms and infection may last as long as a year.

The clinical spectrum of infection among HIV-negative, alcoholic and homeless people has varied. Of these 13 cases of 'urban trench fever', five developed left-sided endocarditis which required valve replacement in four despite antibiotic therapy. One patient died 4 months after valve replacement surgery and one patient who had a concurrent positive blood culture for *Streptococcus pneumoniae* also died, presumably due to pneumococcal sepsis. Many of the patients with *B. quintana* bacteraemia presented with a subacute course of chronic fever, fatigue, and weight loss. Two patients had splenomegaly; however, other symptoms associated with classical trench fever, such headache, rash, and bone pain, were not reported.

Diagnosis

Bartonella quintana are slow growing bacteria which require special culture methods for isolation. The use of Isolator (lysis–centrifugation) tubes improves the yield from blood cultures. Specimens should be plated on enriched media (blood or chocolate agar) incubated at 35 to 37°C in 5 per cent CO_2 and high humidity and held for at least 4 weeks. The organism can also be isolated from blood using Bac Tec or resin-containing culture media if the contents are stained with acridine orange after 1 week of incubation. All stain-positive bottles are then subcultured onto enriched media and processed as above. Cocultivation of blood samples with endothelial cells has also been used for isolation of *Bartonella* species. With more widespread use of culture methods appropriate for the isolation of *Bartonella* species in clinical laboratories, the spectrum and apparent extent of infections due to this organism may be expanded.

Enzyme-linked immunosorbent (ELISA) and immunofluorescence (IFA) assays are available for serological diagnosis. Both exhibit substantial cross-reactivity between *Bartonella* species. The use of paired sera obtained four or more weeks apart is recommended. PCR amplification from infected tissues of DNA specific to *Bartonella* species has also been used.

Treatment

Optimal treatment has not been established. Erythromycin has been the drug of choice, although doxycycline, tetracycline, or azithromycin appear to be acceptable alternatives. At least 14 days of oral therapy is recommended for uncomplicated infection and for bacteraemia at least 4 weeks of therapy is indicated. Most of the few patients identified with *B. quintana* endocarditis have required cardiac valve replacement. Parenteral therapy for 2 to 3 months should therefore be considered for cases of suspected or confirmed *B. quintana* endocarditis.

Relapsing disease is well described, especially if therapy is terminated prematurely. In the Seattle outbreak one patient who was non-compliant with therapy had documented bacteraemia over an 8-week period. It is not known whether extended therapy will prevent relapses.

In immunocompetent hosts, infection with *B. quintana* is usually self-limited unless complicated by endocarditis. The disease is more severe in immunocompromised hosts and may progress to death.

Prevention

Control of the human body louse will prevent transmission of trench fever.

Bacillary angiomatosis–peliosis

Introduction

Bacillary angiomatosis was described in 1983 in an HIV-infected patient with fever and skin nodules. Since then it has been seen in many other HIV-infected patients

and in a few apparently immunocompetent individuals. Bacillary angiomatosis represents one aspect of a spectrum of infections due to the fastidious organisms *Bartonella quintana* and *B. henselae*. A similar disorder known as verruga peruana, caused by *B. bacilliformis*, is restricted to Peru and several neighbouring countries. More reliable diagnostic and identification methods are providing a better understanding of the ubiquitous distribution of both organisms and of the expanding spectrum and overlap of disease they cause. Like *B. bacilliformis*, *B. quintana* and *B. henselae* cause acute febrile illnesses (trench fever, Oroya fever), recurrent asymptomatic bacteraemias, skin lesions, and aseptic meningitis, while ocular involvement (Leber's stellate neuroretinis) appears to be limited to *B. henselae* infections. *B. elizabethiae* has been associated with endocarditis.

Epidemiology

Most cases of bacillary angiomatosis–peliosis have been reported from the United States but its incidence and global distribution are unknown. Epidemiological information is based on case reports, small case series, and a single case–control study. In the largest reported series of cases (*n* = 49), 45 (92 per cent) were HIV infected, one was HIV negative and immunodeficient, and three (6 per cent) were HIV negative and apparently immunocompetent.

Patients infected with *B. henselae* but not *B. quintana* were more likely than controls to own cats, to have been bitten or scratched by cats, to have been exposed to a household cat with fleas, and to have been bitten by cat fleas. The cat flea serves as an arthropod vector for *B. henselae.* Patients with *B. quintana* infections were more likely than controls to be homeless, to have a low annual income, and to be infested with head or body lice. The human body louse is the most likely the vector of *B. quintana.* Neither those infected with *B. henselae* or *B. quintana* were more likely than controls to be alcoholic or to use intravenous drugs.

Aetiology

B. quintana and *B. henselae* have recently been isolated from cutaneous lesions of bacillary angiomatosis. Their aetiological role is also supported by serological and molecular assay data. The morphological and staining characteristics, biochemical and antimicrobial sensitivity profiles, and phylogenetics of *B. quintana* are similar to those of *B. henselae* in cat-scratch disease.

Clinical presentation

Bacillary angiomatosis derives its name from the vascular proliferation and presence of numerous bacillary organisms in affected tissues. It has been reported to involve numerous tissues including skin, lymph node, muscle, bone, bone marrow, brain, liver, and spleen. Bacillary angiomatosis affecting the liver (also 'bacillary peliosis hepatis) and spleen has been referred to as 'bacillary peliosis'. Bacillary angiomatosis most commonly presents as single or clustered, reddish, papular lesions on the skin, but may also occur as brownish patches or subcutaneous nodules and may be confused with Kaposi's sarcoma or disseminated fungal infections, such as *Cryptococcus neoformans*, in the HIV-infected individual. Rarely, diffuse or isolated lymph node involvement can be seen without the characteristic rash. Systemic involvement may also occur, causing lytic bone lesions, peliosis hepatis, or disseminating to other visceral organ in more severe cases. Although descriptions of the disease were in patients with immune deregulation due to neoplastic processes, HIV-1 infection, or immunosuppressive therapy, bacillary angiomatosis has also been described in essentially immunocompetent individuals.

Diagnosis

Biopsy and histological examination of affected tissue is needed for diagnosis of bacillary angiomatosis. It is not possible to distinguish bacillary angiomatosis clinically from Kaposi's sarcoma or other diseases that may affect the skin, spleen, liver, and other tissues, especially in HIV-infected or other immunocompromised persons. Histological criteria for the diagnosis of bacillary angiomatosis include characteristic vascular proliferation on routine haematoxylin-and-eosin staining, and of demonstration of bacillary organisms by silver staining (Warthin–Starry, Steiner, or Dieterle) or electron microscopy.

B. henselae and *B. quintana* have been isolated from cutaneous lesions of bacillary angiomatosis after cultivation of tissue homogenates with endothelial cell monolayers, followed by plating of the supernatants on to solid agar. These organisms can also be isolated from the blood using a lysis–centrifugation method. Serological responses with an indirect immunofluorescence assay for antibodies to *B. quintana* and*B. henselae* may indicate recent infection and provide supporting evidence in a clinical syndrome compatible with diseases. While the duration of antibody responses is not known, IFA reactivity has been documented to last for over 1 year in several longitudinally followed cases. Tissue and blood for culture may also be useful—a positive culture is conclusive, but elusive. PCR of the 16s ribosomal subunit or the citrate synthase gene with restriction fragment length polymorphisms may also be employed.

Treatment

Most *Bartonella* infections are self limiting, particularly when associated with cat-scratch or isolated lymphadenopathy. However, with more disseminated disease such as bacillary angiomatosis, systemic antimicrobial therapy is necessary, particularly when it is an opportunistic infection in AIDS patients. Antimicrobial agents which achieve high intracellular concentrations, such as doxycycline, rifampin, erythromycin, and the macrolides, and possibly trimethoprim–sulfamethoxazol are the most effective in treating and clearing infection. In severe cases, combination therapy with doxycycline or a macrolide and rifampin have been used with success. Fluoroquinolones and cell-wall-active agents, including penicillins and cephalosporins, and aminoglycosides are ineffective. There is no clearly defined duration of therapy, although relapses have been seen when less than 4 weeks of antimicrobial treatment has been given both in HIV-infected and immunocompetent patients.

Prevention

Macrolide antibiotic (erythromycin, clarithromycin) prophylaxis is effective in preventing bacillary angiomatosis–peliosis. Both *B. henselae* and *B. quintana* can be transmitted by arthropod vectors to humans. Elimination of body louse infestations among human populations and cat flea infestations among domestic cats provide a potential means for preventing infections. The domestic cat is the zoonotic reservoir for *B. henselae*, and despite the fact that infected cats show no or mild clinical signs, effective clearance of infection may be achieved through the use of a variety of oral antibiotics.

Further reading

Bass JW, Freitas BC, Freitas AD, *et al*. (1998). Prospective randomized double blind placebo-controlled evaluation of azithromycin for treatment of cat-scratch disease. *Pediatric Infectious Disease Journal* **17**, 447–55.

Broqui P, Lascola B, Roux V, Raoult D (1999). Chronic *Bartonella quintana* bacteremia in homeless patients. *New England Journal of Medicine* **340**, 184–9.

Drancourt M, Mainardi JL, Brouqui P, *et al*. (1995). *Bartonella*(*Rochalimaea) quintana* endocarditis in three homeless men. *New England Journal of Medicine* **332**, 419–23.

Koehler JE, Quinn FD, Berger TG, LeBoit PE, Tappero JW (1992). Isolation of *Rochalimaea* species from cutaneous and osseous lesions of bacillary angiomatosis. *New England Journal of Medicine* **327**, 1625–31.

Margileth AM (1992). Antibiotic therapy for cat-scratch disease: clinical study of therapeutic outcome of 268 patients and a review of the literature. *Pediatric Infectious Disease Journal* **11**, 474–8.

Spach DH, Kanter AS, Dougherty MJ, *et al*. (1995). *Bartonella*(*Rochalimaea*) *quintana* bacteremia in inner-city patients with chronic alcoholism. *New England Journal of Medicine* **332**, 424–8.

Tappero JW, Koehler JE, Berger TG, *et al*. (1993). Bacillary angiomatosis and bacillary splenitis in immunocompetent adults. *Annals of Internal Medicine* **118**, 363–5.

Zangwill KM, Hamilton DH, Perkins BA, *et al*. (1993). Cat scratch disease in Connecticut. Epidemiology, risk factors, and evaluation of a new diagnostic test. *New England Journal of Medicine* **329**, 8–13.

7.11.39.1 *Bartonella bacilliformis* infection

A. Llanos Cuentas

Definition

Bartonellosis (Carrión's disease, verruga peruana, Oroya fever, Guaitará fever) is a non-contagious infectious disease that is endemic in the western Andes and inter-Andean valleys of Peru and occasionally has been reported in Colombia and Ecuador. The acute stage is characterized by infection of red blood cells leading to anaemia; in the late stage the patients develop dermal nodules, which are called 'verrugas'. This disease produces a temporary, reversible immunosuppression in the host, which explains why secondary opportunistic infections are common.

Aetiological agent

Barton, a Peruvian physician, described the causative organism in 1905. *Bartonella bacilliformis* is a small, motile, aerobic, Gram-negative bacillus that stains deep red or purple with Giemsa (Fig. 1). This facultative intracellular haemotrophic bacterium varies in morphology and quantity during various stages of the disease. In spite of being a pleomorphic organism, two essential types are distinguishable: bacilli or rod-shaped forms and coccoid forms. Rod-shaped forms predominate in the acute stage of the disease and coccoid in the convalescent stage. *B. bacillifomis* may infect red blood cells (Fig. 2), endothelial cells of capillaries, and sinusoidal lining cells. The organism is 2 to 3 µm long and 0.2 to 2.5 µm thick. In cultures, 1 to 10 flagella 3 to 10 µm long may originate from one end of the organism. Bartonella can be cultured in Columbia agar supplemented with 5 per cent defibrinated human blood or other supplemented media containing rabbit serum and haemoglobin at 28°C under aerobic conditions for up to 6 weeks.

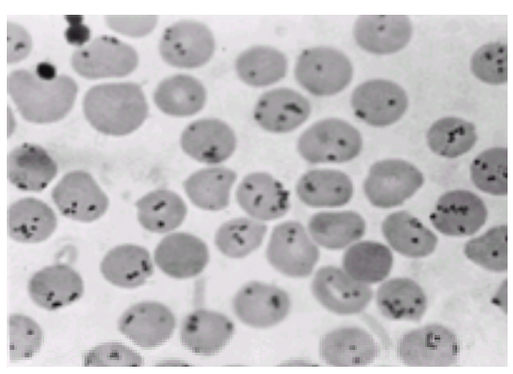

Fig. 1 Smear of peripheral blood with red blood cells parasitized by coccoid forms of *B. bacilliformis* (Wright's stain, × 1048). (Reproduced by courtesy of Professor Juan Takano Moron.)

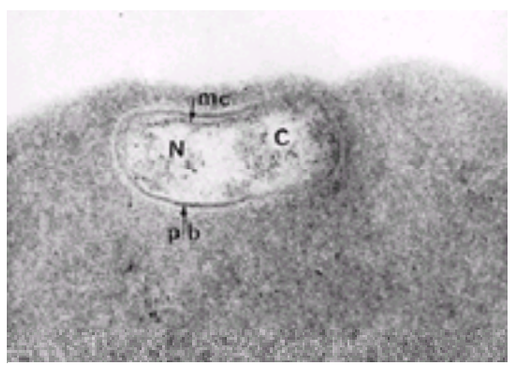

Fig. 2 Ultrastructure of coccobacillary form of *B. bacilliformis* in a red blood cell (× 31 915): mc, cell membrane; N, nucleus; C, cytoplasm; pb, bacterial cell wall. (Reproduced by courtesy of Professor Juan Takano Moron.)

Epidemiology

The disease has occurred since pre-Columbian times, as proved by artistic representations in pre-Inca potteries as well as lesions in a mummy. Bartonellosis is an endemic disease mainly in narrow river valleys and canyons usually in west Andean, and increasingly frequently in inter-Andean, valleys of the central and east Andes of Peru and high jungle areas (Fig. 3). Bartonellosis is a re-emergent disease that is extending its range in Peru. During the last two decades the disease has become endemic in new areas of Ayacucho, Huancavelica, Amazonas, Huanuco, Junín, and Cusco departments and important outbreaks have been reported in recent years. Outbreaks have been described in similar areas in Nariño, Colombia (in 1939) and in Loja and Chichipe areas, Ecuador (Fig. 4). It occurs between 500 and 3200 m above sea level. There are annual high and low transmission seasons and the transmission is greatest towards the end of the rainy season (March to May). Interepidemic periods occur every 10 to 15 years, influenced by climate, environmental, and ecological changes such as El Niño phenomenon.

Fig. 3 Endemic area for bartonellosis, Rimac Valley, Peru. (By courtesy of Professor David H. Molyneux, Liverpool.)

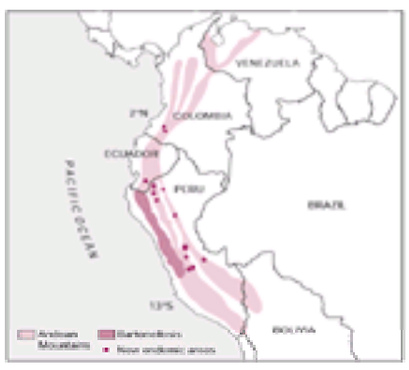

Fig. 4 Geographical distribution of bartonellosis. Coloured spots represent endemic areas of bartonellosis.

In recent years, other human infections by haemotrophic bacteria have been included in the genus *Bartonella: B*. (formerly *Rochalimaea*) *quintana*, the agent of trench fever, *B. henselae*, the major cause of cat-scratch disease, *B. elizabethae*, an aetiological agent of infective endocarditis, and *B*. (formerly *Rochalimaea*) *vinsonii*, until recently not considered to be a pathogen. In immunocompromised people, especially those with AIDS, *B. henselae* and *B. quintana* cause opportunistic infections, frequently manifested as cutaneous bacillary angiomatosis, resembling verruga peruana.

In endemic areas the disease appears in childhood and usually produces few symptoms. Outsiders generally develop acute severe forms of the disease (Oroya fever). Large epidemics have occurred when large groups of non-residents have entered endemic areas. In 1870 an epidemic involved workers building the railroad from Lima to Oroya (Fig. 5); the estimated mortality was 7000. Infection results from the bite of female sandflies of the genus *Lutzomyia*, the most important species being *L. verrucarum*. The vectors are closely associated with human dwellings and, because they are active during twilight hours, people acquire bartonellosis in the hours around sunrise and sunset. Although the reservoir is unknown, there is increasing evidence that humans are the major host. Asymptomatic infection by *B. bacilliformis* has been demonstrated in people of endemic areas. However, since bartonellosis can be acquired in several Andean areas uninhabited by humans, other reservoirs for the disease may exist. Some domestic animals, including horses, donkeys, mules, dogs, and cats, are susceptible and develop lesions similar to verrugas. *Bartonella*-like isolates have been obtained from a *Phyllotis* mouse.

Fig. 5 Puente verrugas' at an altitude of 1800 m above sea level near Lima, Peru. (By courtesy of Mr E. J. Perez.)

Pathogenesis

After inoculation of *B. bacilliformis* through a sandfly bite, the bacteria multiply in endothelial cells of small vessels, and phagocytic cells near the skin. Systemic invasion and multiplication in endothelial cells and red blood cells follows. In the most serious cases, 95 to 100 per cent of red cells are infected with numerous bacteria. The hallmark of the disease is the severe anaemia caused by massive infection of red blood cells and subsequent erythrophagocytosis. Several mechanisms contribute to anaemia: increased fragility, form and size alteration, and reduced half-life of infected and non-infected red cells. Some inhibition of haemoglobin synthesis, probably induced by toxic factors, has also been invoked, since production of red cells increases dramatically with reduction of bacteraemia. Erythrophagocytosis contributes to lymphadenopathy and hepatosplenomegaly. Blockade of the mononuclear phagocytic system and the presence of the circulating iron leads to superinfection, usually by enterobacteria, during the anaemia stage or early recovery from it. Transient depression of cellular immunity has been reported. During the anaemic phase, mild lymphopenia with a reduction of OKT4, a mild increase of OKT8, and decrease of the polyclonal stimulation of the lymphocytes occurs.

A few weeks to months after the acute illness has subsided, the cutaneous form 'verruga peruana' may develop (Fig. 6). The vascular skin lesion shows endothelial proliferation and histiocytic hyperplasia (the cells contain degenerate organisms; Fig. 7), with later fibrosis and necrosis. Electron microscopy of verrucous tissue shows *B. bacilliformis* in the interstitial tissue, indicating that the presence of the bacteria is important for this unusual vascular response to occur. Verruga peruana results from persistent infection, a probably insufficient immune response, and a peculiar vascular reaction, which could be caused by bacterial products acting as an angiogenic factor. In 1885, D.A. Carrión, a Peruvian medical student, linked both phases of the disease by self-experimentation and died.

Fig. 6 Histological section of miliary skin lesion of 'verruga peruana', a sessile or partly pedunculated molluscum-like lesion. (Armed Forces Institute of Pathology photograph negative no. 77355.)

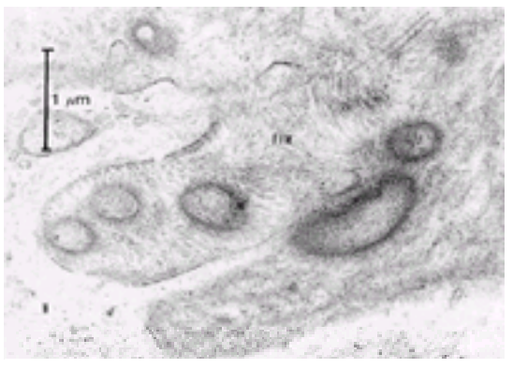

Fig. 7 Electron micrograph of a vascular skin lesion (verruga peruana) showing six *B. bacilliformis* in the fibrillar intertitial matrix (FIM). The cell wall, cell membrane, and internal structure of the bacteria can be seen. The clear cytoplasm of a histiocyte (H) can also be seen. (Reproduced by courtesy of Professor Sixto Pecevarren, Department of Pathology, UPCH and HNCH.)

Clinical features

The disease has two stages, anaemic and eruptive, with an asymptomatic intermediate period. After an incubation period of around 60 days (range 10 to 210 days), non-specific prodromal symptoms appear: onset is usually gradual with malaise, mild chills, fever, and headache. Occasionally, high fever may develop rapidly or build up over a few days. It is accompanied by sweating and rigors. Common symptoms include weakness, aching of the head, back, and extremities, prostration, and depression. The clinical picture is dominated by severe (haemolytic) anaemia: the patient rapidly become pale, dyspnoeic, and jaundiced. There may be hepatosplenomegaly, generalized lymphadenopathy, myocarditis, pericardial effusion, exudates, and retinal haemorrhages in the fundus; sometimes there is generalized oedema, a fine vesicular or petechial rash, and exceptionally, meningoencephalomyelitis. The duration of this state is variable (generally 2 to 4 weeks). In pregnant women the disease in this phase may cause abortion, fetal death, and be transmitted transplacentally; maternal death is common.

In the intermediate period the patients are asymptomatic and recover from the anaemia through great bone marrow activity. This pre-eruptive period varies from weeks to months.

In the eruptive stage, many nodular lesions of varying size appear on the face, trunk, and limbs during a period of one or more months and usually persist for 3 or 4 months. There is accompanying mild arthralgia, myalgia, and sometimes fever. The red or purplish skin lesions vary from papules a few millimetres across. Most often the eruption is miliary (miliary form) with many haemangioma-like lesions of the dermis (Plate 1). Nodular lesions (modular form) are larger but fewer and more prominent on the extensor surfaces or arms and legs (Plate 2). They are painless and prone to bleeding, secondary infection, and ulceration. The appearance may resemble haemangioma, granulome pyogenicum Kaposi's or fibrosarcoma, leprosy (hystioid form), or yaws. Occasionally, one to a few, large, deep-seated lesions that often ulcerate (mular form) develop. These tend to appear near joints, where they may be painful and limit motion. Apart from skin, the mucous membranes of the mouth, conjunctiva, and nose, serous cavities, and the gastrointestinal and genitourinary tracts may be involved. The eruptive phase tends to heal spontaneously, although the course is often prolonged. Inhabitants of endemic areas usually develop the eruptive stage as the sole manifestation of the disease.

The principal complication is superinfection, leading to septicaemia, which occurs at different stages of the disease but generally in the later part of the anaemic stage and during the intermediate stage. Formerly, *Salmonella typhi*,*S. typhimurium*,*S. dublin*,*S. anatum*,*S. enteritidis*, *Mycobacterium tuberculosis*, and *Enterobacter* spp. were the most frequent pathogens. Reactivation of toxoplasmosis, histoplasmosis, pneumocystosis, and sta-phylococcal infections are some of the other infections that are now frequent.

Diagnosis

Two elements must be considered: visiting or residence in an endemic area and a compatible clinical picture with demonstration of the bacteria in the blood film. Fluorescence antibody test, indirect haemagglutination, immunoblot, and enzyme-linked immunosorbent assay (ELISA) are the new serological tests, but they are not generally available.

Laboratory features

Bartonella can be isolated from the blood during the anaemic stage and sometimes during the eruptive stage. The enriched media may be positive in 4 to 28 days at 28 °C. As fever develops, intraerythocytic bacteria are visible in thick and thin films stained with Giemsa, Wright, or other variants of the Romanovsky stain. Organisms can also be seen and cultivated in the skin lesion of verrucous tissue. The haemolytic anaemia is Coombs' test negative. The blood picture is a macrocytic and hypochromic anaemia with polychromasia, anisocytosis, and poikilocytosis. The reticulocytosis is marked (average 11 per cent). The marrow is hyperactive and megaloblastic with erythrophagocytosis. The white cell count is not markedly elevated unless there is a secondary infection. Thrombocytopenia is quite common. After the crisis, the intracellular organisms ecome coccoid and later disappear, the white cell count rises, and there is lymphocytosis. Eosinophils, which are usually absent during the acute stage, reappear in differential counts of peripheral blood.

Prognosis and treatment

Death is usually during the anaemic phase, and in the preantibiotic era varied between 20 and 95 per cent. At present it varies between 1.1 and 2.4 per cent in endemic areas and around 9 per cent in patients admitted to hospital. During outbreaks, especially when the disease in not promptly recognized and treated, the mortality can reach 88 per cent. Alterations of consciousness (excitement, stupor, and coma) and progressive or focal neurological features, biochemical evidence of hepatic dysfunction (increased serum aspartate and alanine transaminases and alkaline phosphatase), pulmonary complications (non-cardiogenic pulmonary oedema), anasarca (severe hypoalbuminaemia), and pregnancy are associated with a higher mortality.

Chloramphenicol, penicillin, erythromycin, co-trimoxazole, tetracycline, and ciprofloxacin are dramatically effective, usually eliminating the fever in less than 48 h. Because of the common association with salmonellosis, chloramphenicol is the treatment of choice in a dose of 50 mg/kg per day for 10 days. An alternative is ciprofloxacin in a dose of 500 mg twice a day for 10 days. Supportive treatment includes transfusion of packed red cells and dexamethasone (if there is severe neurological involvement). Rifampicin (300 mg twice a day in adults or 10 mg/kg per day in children, orally for 14 to 21 days) is indicated for treatment of the verrucous form.

Prevention

When transmission is around dwellings, sandflies can be temporarily eliminated by spraying inside and outside with DDT or pyrethroids. Bites usually occur after dusk. They can be prevented by insect repellents, impregnation of clothes with pyrethroids, sleeping inside fine-meshed or impregnated nets, or by avoiding sleeping in highly endemic areas.

Further reading

Birtles RJ *et al.* (1999). Survey of *Bartonella* species infecting intradomicillary animals in the Huayllacallan valley, Ancash-Perú, a region endemic for human bartonellosis. *American Journal of Tropical Medicine and Hygiene* **60**, 799–805.

Gray GC *et al.* (1990). An epidemic of Oroya fever in the Peruvian Andes. *American Journal of Tropical Medicine and Hygiene* **42**, 215–21.

Maguiña C (1998). *Bartonellosis o enfermedad de Carrión*. AFA Editores Importadores SA Lima, Peru.

Maguiña C, Gotuzzo E (2000). Bartonellosis new and old. In: Emerging and re-emerging diseases in Latin America. *Infectious Diseases Clinics of North America* **14**, 1–22.

Walker DH, Guerra H, Maguiña C (1999). Bartonellosis. In: Guerrant RL, Walker DH, Weller PF, eds. *Tropical infectious diseases, principles, pathogens and practice*, pp 492–7. Churchill Livingstone, Philadelphia.

7.11.40 Chlamydial infections including lymphogranuloma venereum

D. Taylor-Robinson and D. C. W. Mabey

Trachoma is discernible as a cause of blindness in ancient Chinese and Egyptian writings. However, it was not until 1907 that L. Halberstaedter and S. von Prowazek first described intracytoplasmic inclusions in conjunctival scrapings from patients with trachoma and recognized the involvement of an infectious agent. In 1930, the first isolation of a chlamydial agent (*Chlamydia psittaci*) from psittacosis was made, that is about 27 years before the genomically and biologically different agent, *Chlamydia trachomatis*, was isolated in fertile hens' eggs from trachoma. The advent of the cell-culture technique in the late 1950s paved the way for the isolation of *C. trachomatis* in 1965 by this means and, together with immunological developments, made it possible to explore the nature, range, prevalence, and pathogenesis of clinical conditions associated with chlamydial infection.The complete sequencing of the chlamydial genome, accomplished recently, provides an even greater opportunity to define many unfathomed aspects of the biology and pathogenesis of chlamydial infections.

Classification

Chlamydial organisms, or chlamydiae, are ubiquitous pathogens infecting many species of mammals and birds. The genus *Chlamydia* comprises at least four species, of which three can infect humans. *C. trachomatis* is pathogenic for humans and causes ocular, genital, and systemic infections that affect millions of people worldwide. *C. pneumoniae* causes mainly human respiratory disease, has been associated with atherosclerosis, and equine and koala strains exist. *C. psittaci* infects birds and other animals, resulting in major economic losses and occasionally is transmitted to humans The fourth chlamydial species, *C. pecorum*, causes pneumonia, polyarthritis, encephalomyelitis, and diarrhoea in cattle and sheep.

Taxonomic reclassification based on 16S RNA analysis has been proposed recently. In this, the order Chlamydiales contains four families, the first of which, Chlamydiaceae, comprises two genera, namely *Chlamydia* (for example, *Chlamydia trachomatis*) and *Chlamydophila*(for example, *Chlamydophila pneumoniae*). This proposal has generated considerable debate.

Growth cycle, serovars, and protein profile

Chlamydiae probably evolved from host-independent, Gram-negative ancestors that contained peptidoglycan in their cell wall. The chlamydial envelope, like that of Gram-negative bacteria, has inner and outer membranes. Indeed, chlamydiae are best considered as bacteria that are specialized for an intracellular existence. The infectious elementary body is electron dense, deoxyribonucleic acid rich, and approximately 300 nm in diameter. Elementary bodies of *C. pneumoniae* often, but not always, have a wide periplasmic space and appear pear shaped, whereas those of the other chlamydial species have a narrow periplasmic space and are spherical. The elementary body begins its eukaryotic intracellular lifecycle by binding to the host cell and entering by 'parasite-specified' endocytosis. Inside the host cell, fusion of the chlamydia-containing endocytic vesicle with lysosomes is inhibited and the elementary body begins its unique developmental cycle. After about 10 h it has differentiated into the larger (800 to 1000 nm), non-infectious, metabolically active, reticulate body. This divides by binary fission and by 20 h there is the beginning of reorganization into a new generation of elementary bodies (Fig. 1), which reach maturation 20 to 30 h after entry into the cell. Their rapid accumulation within the endocytic vacuole precedes release from the cell between 30 and 48 h after the start of the cycle.

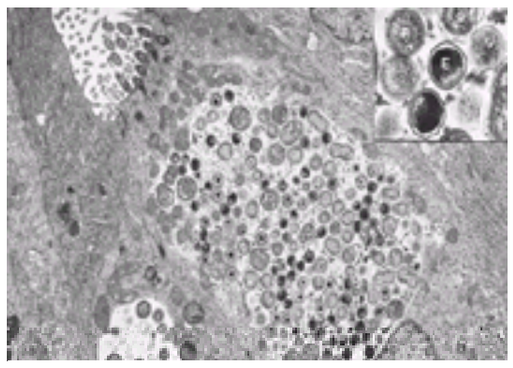

Fig. 1 Elementary bodies (E) and reticulate bodies (R) of *C. trachomatis* forming an inclusion in oviduct cell shown by transmission electron microscopy.

All species within the genus *Chlamydia* contain a common, heat-stable, lipopolysaccharide antigen, which is exposed on the surface of the reticulate body, but not on that of the elementary body. The major outer membrane protein (**MOMP**) is immunodominant in the elementary body and it contains epitopes that exhibit genus, species, and serovar specificity.The serovar-specific epitope is the basis of the microimmunofluorescence test by which *C. trachomatis* has been separated into 15 serovars: A, B, Ba, and C are responsible mainly for endemic trachoma, and D to K for oculogenital infections. Serovars L1, L2, and L3 of *C. trachomatis* cause the genital disease, lymphogranuloma venereum. At present, only one *C. pneumoniae* serovar has been identified, although minor geographical serovar variations have been described. The loosely defined *C. psittaci* species is likely to contain a wide variety of host-related serovars. Amino acid sequences of the MOMPs of all *C.*

trachomatis serovars are now known and epitope maps of different antigenic domains have been elucidated. It appears that the MOMP genes consist of five highly conserved regions punctuated by four short variable sequences. Serovar-specific epitopes have been demonstrated in variable sequence in I and II, while species-specific epitopes have been found in variable sequence IV. It is also probable that these variable sequences have some role in chlamydial pathogenesis. *C. trachomatis*, *C. psittaci*, and *C. pneumoniae* species have been compared and, although there is only 10 per cent DNA homology between each of them, MOMP gene analysis of the respective species reveals up to 65 per cent amino acid homology, indicating a probable common ancestor. A common chlamydial 57-kDa protein has been described and its possible role in disease pathogenesis is considered below.

Trachoma

Trachoma is a chronic keratoconjunctivitis believed to affect some 500 million people, of whom 7 million are blind and 10 million have some visual impairment. After cataract, it is the most common cause of blindness worldwide, but is now confined largely to developing countries. Trachoma is a disease of poverty rather than of hot climates, and in some respects the relation between genital and ocular chlamydial infections resembles that between syphilis and non-venereal treponematoses. In poor communities where hygienic standards are low, there is direct transfer of chlamydiae from eye to eye (compare with skin to skin for non-venereal treponematoses), and trachoma is endemic. As standards of hygiene improve, this mode of transmission is no longer possible and trachoma becomes less of a problem.

Clinical features

The active (inflammatory) stage of the disease, a follicular conjunctivitis, affects chiefly the subtarsal conjunctiva, but follicles occur elsewhere on the conjunctiva and at the limbus, where on resolution they leave characteristic shallow depressions known as Herbert's pits. New vessels (pannus) may be seen at this stage in the cornea, usually at the superior margin, and punctate keratitis may also be a feature. Since symptoms are mild or absent, the disease may not be suspected unless the upper eyelid is everted. Active trachoma affects mainly children in endemic areas. Among older children and adults in such areas, conjunctival fibrosis often develops as the follicles resolve and, if severe, it may distort the upper lid margin, turning it inward (entropion). Lashes rubbing against the globe (trichiasis) cause continuous discomfort and sometimes blindness due to corneal damage.

The World Health Organization (**WHO**) has proposed criteria for the clinical diagnosis of active trachoma and its potentially blinding sequelae and for grading their severity as follows:

1. trachomatous inflammation—follicular (**TF**): five or more follicles, each at least 0.5 mm in diameter, in the upper tarsal conjunctiva;
2. trachomatous inflammation—intense (**TI**): pronounced inflammatory thickening of the tarsal conjunctiva that obscures more than half the normal deep tarsal blood vessels;
3. trachomatous conjunctival scarring (**TS**): easily visible scarring in the tarsal conjunctiva;
4. trachomatous trichiasis (**TT**): at least one eyelash rubbing on the eyeball—evidence of recent removal of inturned eyelashes also graded as trichiasis; and
5. corneal opacity (**CO**): easily visible corneal opacity over the pupil, so dense that at least part of the pupil margin is blurred when viewed through the opacity.

Epidemiology

The reservoir of infection in endemic areas is the eye and possibly the nasopharynx of children with active disease. Active cases tend to cluster by household where there is prolonged intimate contact within the family. The higher prevalence of active disease and scarring in women than in men is probably due to the closer contact between women and children. *C. trachomatis* may be transferred from the eye of one individual to that of another via fingers, fomites, coughing, and sneezing, and by eye-seeking flies. Severe conjunctival scarring probably occurs only among individuals repeatedly exposed to reinfection. The availability and use of water are important determinants of the development of trachoma. When living conditions improve trachoma tends to disappear.

Diagnosis

In trachoma-endemic areas the diagnosis is generally made on clinical grounds, as most cases of follicular conjunctivitis are due to trachoma and laboratory facilities are usually lacking. Trachomatous follicles may be confused with the giant papillas of vernal conjunctivitis, in which pannus may also be seen. A number of viruses, notably adenoviruses, can cause follicular conjunctivitis. Intense cases of trachoma (TI), in which follicles may not be visible, should be distinguished from bacterial conjunctivitis. The diagnosis of trachomatous scarring is usually obvious, as few other conditions cause conjunctival scarring of the upper lid.

Laboratory diagnosis depends on the detection of *C. trachomatis*, which may be found in about 50 per cent of cases of active disease (TF or TI), but in only a minority of cases of scarring disease (TS). The microbiological diagnostic procedures are discussed later in this chapter.

Treatment

Inflammatory trachoma (TF and TI) responds to treatment with antimicrobial agents active against *C. trachomatis* (see Table 2). Until recently, the WHO has recommended 1 per cent topical tetracycline ointment, to be applied to both eyes daily for 6 weeks. This has proved impractical on a wide scale in trachoma-endemic communities. More recently, a single oral dose of azithromycin (20 mg/kg, to a maximum of 1 g) has been shown to be equally effective. When only individual cases are treated, reinfection is usually rapid; treatment of whole communities may reduce the rate of reinfection. Trichiasis and entropion require surgical correction. Several lid operations have been described, but few have been evaluated prospectively. Tarsal rotation is probably the operation of choice.

Prevention

The WHO has launched an initiative for the global elimination of blinding trachoma by the year 2020. The recommended strategy is based on the acronym 'SAFE': Surgery for trichiasis; Antibiotics for the treatment of inflammatory disease and the elimination of the reservoir of infection; promotion of Face washing, and Environmental improvement, to reduce fly populations, both of which are likely to reduce the rate of transmission of ocular *C. trachomatis* infection. In Mexico, children who washed their faces seven or more times per week were less likely to have trachoma than those who washed less often, and this intervention was also effective among rural villagers in Tanzania.

Genital tract infections

Infections of the genital tract due to *C. trachomatis* (Table 1) occur worldwide and, at least in developed countries, are much more common than gonococcal infections. The economic burden on health services due to genital chlamydial infections is enormous; for example, more than 3 billion dollars per year for pelvic inflammatory disease in the United States, based on 1994 incidence data. In Sweden, widespread and effective diagnostic testing, coupled with aggressive contact tracing and treatment, has greatly reduced genital chlamydial infections. This has not been achieved in other developed countries, but screening programmes are being or have been developed and implemented in some.

Non-gonococcal urethritis

C. trachomatis is detectable in the urethra of not more than 50 per cent of men with non-gonococcal urethritis and in up to 25 per cent of those with asymptomatic urethral infections. It is also likely that chlamydiae are a cause of some cases of chronic non-gonococcal urethritis

In women, there is no doubt that chlamydial urethral infection may cause urethritis but, in contrast to men, infection and inflammation are almost always asymptomatic. Thus, the dysuria and frequency of the urethral syndrome are rarely of chlamydial origin.

Prostatitis and epididymo-orchitis

There is no evidence that *C.trachomatis* causes acute symptomatic prostatitis. In chronic abacterial prostatitis diagnosed by the Stamey procedure, biopsy tissues taken transperineally to avoid the urethra have shown chronic inflammation, but chlamydiae have not been detected in them by culture and direct immunofluorescence techniques, although about 10 per cent have proved positive using polymerase chain-reaction technology. These largely negative observations, and the failure to

detect chlamydial antibody, suggest that chlamydiae are not often implicated directly in the chronic disease. However, the possibility cannot be excluded that a portion, at least, of chronic disease is chlamydial in origin, maintained perhaps by immunological means. A predominance of CD8 cells in the tissues is consistent with this notion.

C. trachomatis is responsible for epididymitis primarily in young men (35 years of age or less) in developed countries, being detected in at least one-third of epididymal aspirates. Furthermore, there is a strong correlation between IgM and IgG chlamydial antibodies, measured by microimmunofluorescence, and chlamydia-positive disease. In developing countries, although chlamydiae are important, *Neisseria gonorrhoeae* is still the major cause of acute epididymitis. In patients older than 35 years, an age boundary that, of course, is not strict, epididymo-orchitis tends to be caused by urinary-tract pathogens.

Convincing evidence that chlamydiae have been detected in the testes or that a previous chlamydial urethral infection or asymptomatic chlamydial infection causes male infertility has not been forthcoming.

Bartholinitis, vaginitis, and cervicitis

C. trachomatis has been weakly associated with bartholinitis, but is not regarded as a major cause. Chlamydiae are often detected more frequently in women with bacterial vaginosis than in those without this condition, but there is no evidence that they are causally associated or in any way contribute to the disease. It is apparent that the squamous epithelium of the vagina is not susceptible to chlamydial infection and that the cervix is the primary target for *C. trachomatis*. Indeed, it is a well known cause of mucopurulent/follicular cervicitis, although infection often may be asymptomatic. Women younger than 25 years, unmarried, using oral contraceptives, and who have signs of cervicitis are the most likely to have a chlamydial infection.

An association between cervical chlamydial infection and cervical intraepithelial neoplasia has been seen, but a causal link has not been established.

Pelvic inflammatory disease

Canalicular spread of chlamydiae to the upper genital tract leads to endometritis, which is often plasma-cell associated and sometimes intensely lymphoid. Further spread causes salpingitis, perihepatitis (the Curtis Fitz-Hugh syndrome), sometimes confused with acute cholecystitis in young women, in addition to periappendicitis and other abdominal complaints. Surgical termination of pregnancy or insertion or removal of an intrauterine contraceptive device may predispose to dissemination of the organisms.

Chlamydiae are the major cause of pelvic inflammatory disease in developed countries. Infertility is the outcome in about 10 per cent of such disease and may be the first indication of asymptomatic tubal disease. Fertility is influenced adversely by an increasing number and severity of upper genital tract chlamydial infections and infertility could result from endometritis, blocked or damaged tubes, or perhaps abnormalities of ovum transportation. Other consequences of salpingitis are ectopic pregnancy and chronic pelvic pain.

Other diseases associated with *C. trachomatis*

Adult paratrachoma (inclusion conjunctivitis) and otitis media

Adult chlamydial ophthalmia is distinguished from trachoma because it is caused by serovars D to K of *C. trachomatis* and commonly results from the accidental transfer of infected genital discharge to the eye. Chlamydiae can be detected in conjunctival specimens and in this respect the condition is different from the 'reactive' conjunctivitis seen in Reiter's syndrome (see below), where isolation from the conjunctiva is extremely unusual. Adult chlamydial ophthalmia usually presents as a unilateral follicular conjunctivitis of acute or subacute onset, the incubation period ranging from 2 to 21 days. The features are swollen lids, mucopurulent discharge, papillary hyperplasia due to congestion and neovascularization and later, follicular hypertrophy, and occasionally punctate keratitis. About one-third of patients have otitis media, complaining of blocked ears and hearing loss. The disease is generally benign and self-limited, but pannus formation and corneal scarring may occur unless systemic treatment is given. Patients and their sexual contacts should be investigated for the existence of genital chlamydial infections and managed accordingly.

Arthritis

Arthritis occurring with or soon after non-gonococcal urethritis is termed sexually acquired reactive arthritis (**SARA**); in about one-third of cases, conjunctivitis and other features characteristic of Reiter's syndrome are seen. At least one-third of cases of such disease are initiated by chlamydial infection and *C. trachomatis* elementary bodies and chlamydial DNA and antigen may be detected in the joints. *C. trachomatis* has also been associated in the same way with 'seronegative' arthritis in women. Viable chlamydiae have not been detected in the joints of patients with SARA and the pathogenesis of the disease is probably immunologically based (see below). Despite this, early tetracycline therapy is advocated by some investigators.

Immunocompromised states

C. trachomatis has been isolated from the lower respiratory tract of a few immunocompromised adults with pneumonia, some after renal transplantation, but its role has been obscured by the recovery of other agents from some. However, neither *C. trachomatis* nor *C. pneumoniae* is an important respiratory-tract pathogen in patients with AIDS. Nor does genital chlamydial disease seem to be more widely prevalent or severe in HIV-infected patients and hypogammaglobulinaemic patients do not appear to be especially prone to infection with any of the chlamydial species.

Neonatal infections

Although intrauterine chlamydial infection can occur, the major risk of infection to the infant is from passing through an infected cervix. The proportion of neonates exposed to infection depends, of course, on the prevalence of maternal cervical infection, which varies widely. However, between one-fifth and one-half of infants exposed to *C. trachomatis* serovars D to K infecting the cervix at the time of birth develop conjunctivitis, which occurs usually 1 to 3 weeks after birth. A mucopurulent discharge and occasionally pseudomembrane formation occur, but it is usually self-limited, resolution occurring without visual impairment. If complications do arise, however, they tend to be in untreated infants.

About half of the infants who develop conjunctivitis develop pneumonia, although the latter is not always preceded by conjunctivitis. A history of recent conjunctivitis and bulging eardrums is found in only about half of the cases. Chlamydial pneumonia occurs usually between the fourth and eleventh week of life, preceded by upper respiratory symptoms, and has an afebrile, protracted course in which there is tachypnoea and a prominent, stacatto cough. Hyperinflation of the lungs with bilateral, diffuse, and symmetrical interstitial infiltration and scattered areas of atelectasis are the radiographic findings. The occurrence of serum IgM antibody to *C. trachomatis* in infants with pneumonia is pathognomonic. Children so affected during infancy are more likely to develop obstructive lung disease and asthma than are those who have had pneumonia due to other causes.

The vagina and rectum also may be colonized by *C. trachomatis* at birth. However, vaginal colonization has not been associated with clinical disease, nor has there been evidence for chlamydial gastroenteritis in infants.

Lymphogranuloma venereum

This is a systemic, sexually transmitted disease caused by serovars L1, L2, L2a, and L3 of *C. trachomatis*. These chlamydiae are more invasive than the other serovars and cause disease primarily in lymphatic tissue. Although a small papule or necrotic genital lesion may be the first sign of infection, with the rectosigmoid colon also a primary site, the chlamydiae are soon carried to regional lymph nodes. These enlarge rapidly and inflammation of the capsule causes them to mat together. Multiple minute abscesses form in the parenchyma and in the absence of treatment they may coalesce and form sinus tracts, which rupture through the overlying skin. Scar tissue may obstruct lymphatic flow causing lymphoedema and elephantiasis of the genitalia and strictures, ulcers, and fistulas may develop.

Clinical features

Three stages of infection are usually recognized. After an incubation period of 3 to 21 days, a small, painless, papular, vesicular, or ulcerative lesion develops and disappears spontaneously within a few days without scarring. In men the lesion is on the penis, and in women most commonly on the fourchette, often going unnoticed, especially if it is in the rectum of homosexual men. Extragenital primary lesions on fingers or tongue are rare.

The secondary stage is conventionally separated into inguinal and genitoanorectal syndromes. The former is more common and is usually seen in men as an acute painful inguinal bubo. The lymphadenopathy is unilateral in two-thirds of cases, and rarely may be so extensive that the inguinal mass is cleaved by the inelastic Poupart's ligament—the almost pathognomonic 'groove sign' of the disease. Buboes are accompanied by fever, malaise, chills, arthralgia, and headache and about 75 per cent of them suppurate and form cutaneous draining sinus tracts. In women, the external and internal iliac lymph nodes and the sacral lymphatics are involved more often than are the inguinal lymph nodes. Signs include a hypertrophic suppurative cervicitis, backache, and adnexal tenderness. In both sexes, but more frequently in women, a genitoanorectal syndrome characterized by a haemorrhagic proctitis or proctocolitis may occur. Inflammation is limited to the rectosigmoid colon and is accompanied by fever, a mucopurulent or bloody anal discharge, tenesmus, and diarrhoea. Histopathological changes in such cases may mimic Crohn's disease. The process usually resolves spontaneously after several weeks but, rarely, anal, rectovaginal, rectovesical, and ischiorectal fistulas occur and, late in the disease, a rectal stricture. Rare manifestations of the secondary stage are acute meningo-encephalitis, synovitis, pneumonia, cardiac involvement, and follicular conjunctivitis, which is self-limited.

Lesions of the tertiary stage appear after a latent period of several years. They include genital elephantiasis, occurring predominantly in women as a sequel to the genitoanorectal syndrome and often accompanied by fistula formation, and rectal stricture, which is found almost exclusively in women or homosexual men. Gross ulceration and granulomatous hypertrophy of the vulva ('esthiomene') is very rare. Indeed, all late complications are rare today because of broad-spectrum antibiotics.

Epidemiology

Lymphogranuloma venereum is found worldwide, but its major incidence is limited to endemic foci in sub-Saharan Africa, South-East Asia, South America, and the Caribbean. All races are equally susceptible to infection, but the reported sex ratio is usually greater than 5:1 in favour of men because early disease is recognized much more easily in them. In North America and Europe, the disease is usually diagnosed in travellers, seamen, and military personnel returning from endemic areas, and in male homosexuals. The reservoir of infection is presumed to be asymptomatically infected women and male homosexuals.

Diagnosis

The differential diagnosis of lymphogranuloma venereum includes genital herpes, syphilis, chancroid, donovanosis, extrapulmonary tuberculosis, cat-scratch disease, plague, filariasis, lymphoma, and other malignant diseases. Of these, primary genital herpes can be the most difficult to distinguish. Lymphadenitis of the deep iliac nodes may mimic appendicitis or pelvic inflammatory disease and tuberculosis and certain parasitic and fungal infections of the genital tract cause lymphoedema and elephantiasis of the genitalia that may cause confusion.

The classic Frei skin test is no longer used for diagnosis. Staining of infected tissues to detect elementary bodies or inclusions (see later section on diagnosis) is not often used because the frequent bacterial contamination makes detection difficult. The use of cell culture is preferable, but only 25 to 40 per cent of patients with lymphogranuloma venereum have positive cultures of bubo aspirate, endourethral or endocervical scrapings, or of other infected material. The much more sensitive DNA amplification methods are being used with increasing frequency.

Of the serological tests (see later section on diagnosis), complement fixation is not specific for lymphogranuloma venereum. The microimmunofluorescence test is also not entirely specific, but is the method of choice and antibody titres of 1:1024 or more are not uncommon and can be regarded as diagnostic, particularly in a patient with typical signs and symptoms.

Treatment

Of the several antimicrobial drugs available, oral tetracycline is usually recommended (Table 2) although azithromycin is finding a place. Fever and bubo pain rapidly subside after antibiotic treatment is started, but buboes may take several weeks to resolve. Suppuration and rupture of buboes with sinus formation is usually prevented by antibiotic treatment. Surgical incision and drainage is neither necessary nor recommended. How long treatment needs to be continued to prevent relapse or progression of disease is debated but a minimum of 2 weeks is recommended. Fistulas, strictures, and elephantiasis may require plastic repair but surgery should not be attempted until the patient has had weeks or months of antimicrobial treatment to reduce inflammation and necrosis.

Chlamydia pneumoniae infections

It is interesting that the prototype strains of *C. pneumoniae* were isolated from conjunctival material collected in the mid-1960s from patients in trachoma-endemic areas. It was not until 1983, however, that a third *C. pneumoniae* strain was isolated, this time from the throat of a patient with acute pharyngitis. The two original isolates (TW-183 and IOL-207) were found to be identical serologically and distinct from *C. trachomatis* and *C. psittaci* and, in 1989, *C. pneumoniae* was defined as the third species of the genus *Chlamydia*. At present only one serovar of *C. pneumoniae* has been identified, although minor geographical serovar variations have been described.

Clinical features

Respiratory tract disease

At the outset of acute disease, pharyngitis is often present, more than 80 per cent of patients with lower respiratory-tract disease developing a sore throat. A cough may take some time to develop and fever is uncommon. Bronchitis is associated with some infections and in young adults about 5 per cent of primary sinusitis is associated with *C. pneumoniae*. Mild respiratory infections are probably frequent but, overall, pneumonia has been the most common feature; in mild cases, radiographs usually reveal a unilateral pneumonia, whereas in patients needing hospital care, bilateral pneumonia is quite common. This is often difficult to distinguish clinically from that caused by other micro-organisms, for example *Mycoplasma pneumoniae*.

Arthritis

An exaggerated synovial lymphocyte response to *C. pneumoniae* has been found in some adults with reactive arthritis and *C. pneumoniae* DNA and high titres of specific antibody have been detected in the joints of a few children with juvenile chronic arthritis, suggesting the possibility of a causal role.

Atherosclerosis

Patients with chronic coronary heart disease and acute myocardial infarction were noted first by Finnish investigators and later by others to have antibody to *C. pneumoniae* significantly more often than age-matched controls. The possibility that *C. pneumoniae* infection might be a risk factor for such disease was enhanced by detection of the organisms or their DNA in at least 40 per cent of atheromatous coronary and other major arteries of adults and subjects as young as 15 years. In addition, specific DNA has been found in at least 40 per cent of peripheral blood mononuclear cells, raising the possibility that they transmit the organisms to the arterial wall, atheromatous or otherwise, from the respiratory tract. Inoculation of mice and rabbits with *C. pneumoniae* has initiated or potentiated atherosclerotic-like changes, observations that are provocative but insufficient to determine the significance of the human findings.

Epidemiology

C. pneumoniae organisms are transmitted from person to person, apparently without any intermediate host. Serological evidence indicates that *C. pneumoniae* is widespread and endemic in many areas, although localized epidemics have been recorded in both military and civilian groups in Scandinavia, the United States, the United Kingdom, and elsewhere. *C. pneumoniae* probably causes many mild respiratory infections that were previously thought to be viral in origin and it is also likely that many infections labelled 'human psittacosis/ornithosis' in the past were in reality due to *C. pneumoniae*.

Chlamydia psittaci infections

The *C. psittaci* species comprises a diverse group of organisms that have been isolated from a variety of mammals and frogs and many avian species. Nine serovars have been proposed from mammals and seven from birds, and two biovars from koala bears. The relatively low degree of homology between serovars exhibited in DNA–DNA hybridization analyses signals the possibility of further speciation among organisms currently assigned to the species. This, in fact, has happened with the separation of strains causing pneumonia, polyarthritis, encephalomyelitis, and diarrhoea in cattle and sheep and the constitution of the fourth chlamydial species, *C. pecorum*.

The spectrum of animal diseases caused by *C. psittaci* species includes enteritis, abortion, sterility, pneumonia, and encephalitis, all of which cause economic loss. Occasionally, the organisms are transmitted to humans through contact with infected animals or birds or from contact with faecal materials from an infected source. Psittacosis may be a hazard to those who keep pet birds or who work in poultry processing plants, or in animal husbandry. Many birds are known to harbour the organisms, but psittacine species (parrots), poultry, and pigeons are probably the major sources of human infection.

Clinical features

Human respiratory infection with *C. psittaci* (psittacosis) is equally common in either sex. It is uncommon and mild in childhood and usually affects adults, particularly those in the 30- to 60-year age group. After an incubation period of 1 to 2 weeks, the clinical presentation can vary from a mild influenza-like illness to a fulminating toxic state with multiple organ involvement. The disease may be insidious in onset over a few days or start abruptly with high fever, rigors, and anorexia. Headache occurs in most, a cough, often dry, in over two-thirds, and arthralgia and myalgia in over one-third. Inspiratory crepitations are more common than classic signs of consolidation. Chest radiographs usually show patchy shadowing, most often in the lower lobes. Homogeneous lobar shadowing is less common, and miliary and nodular patterns even less so. Hilar lymphadenopathy has been reported in up to two-thirds of patients and a pleural reaction in more than half, but significant pleural effusions are infrequent. Extrapulmonary complications, mostly rare, include endocarditis, myocarditis, pericarditis, a toxic confusional state, encephalitis, meningitis, tender hepatomegaly, splenomegaly, pancreatitis, haemolysis, and disseminated intravascular coagulation. The advent of superior laboratory tests should not allow disease caused by *C. psittaci* to be confused with that due to *C. pneumoniae*, as occurred in the past.

Ovine *C. psittaci* strains have caused abortion, albeit rarely, in pregnant women, often farmers' wives, after exposure to sheep suffering from enzootic abortion during the lambing season. The feline keratoconjunctivitis agent, isolated from the genital tract of female cats, has caused follicular conjunctivitis in humans similar to that caused by *C. trachomatis* serovars D to K.

Diagnosis of chlamydial infections

The laboratory diagnosis of chlamydial infection depends on detection of the organisms or their antigens or DNA and to a much lesser extent on serology. The procedures mentioned, with some of their advantages and disadvantages, are summarized in Table 3. Certain swabs, for example those that are cotton tipped, are superior to others, and swabs provided in commercial enzyme immunoassay kits may be toxic if used for collecting specimens for culture. Examination of two or more consecutive swabs from patients rather than one improves the chlamydial detection rate and this may be achieved in women by pooling cervical and urethral specimens. However, in recent times attention has turned to the use of 'first-catch' urine specimens, which were ignored for years because they were found not to be suitable for chlamydial culture. Nevertheless, they are unquestionably valuable samples from both men and women, provided that the centrifuged deposits are tested by molecular methods. The same comment applies to the use of meatal samples in men and of vulvar/vaginal samples.

Culture and staining of chlamydia

The growth of chlamydiae about 40 years ago in cultured cells, rather than in embryonated eggs, revolutionized both their detection and chlamydial research. *C. pneumoniae* is particularly difficult to isolate and this may be facilitated by using a line of human lung cells. The method of detection used widely for *C. trachomatis* involves the centrifugation of specimens on to cycloheximide-treated McCoy cell monolayers. Inoculation of cell cultures is followed by incubation and staining with a fluorescent monoclonal antibody or with a vital dye, usually Giemsa, to detect inclusions; one blind passage may increase sensitivity. However, the cell-culture technique is no more than 70 per cent sensitive and is slow and labour intensive, drawbacks that have hastened the development of non-cultural methods .

Staining of epithelial cells in ocular and genital specimens with vital dyes was used first to detect chlamydial inclusions, but the method is insensitive and often non-specific. Papanicolaou-stained cervical smears provide an excellent example of these drawbacks. In contrast, detection of elementary bodies by using species-specific fluorescent monoclonal antibodies is rapid and, for *C. trachomatis* oculogenital infections, sensitivities ranging from 70 to 100 per cent and specificities from 80 to 100 per cent have been achieved. Skilled observers, capable of detecting a few elementary bodies, even one, provide values at the top of these ranges. However, the test is most suited for dealing with a few specimens and for confirming positive results obtained with other tests.

Enzyme immunoassays and DNA amplification techniques

The popularity of enzyme immunoassays that detect chlamydial antigens is due to their ease of use and not to their sensitivity. Indeed, it is rarely possible to detect small numbers of chlamydial organisms (less than 10) of whatever species. Since at least 30 per cent of genital swab specimens and a larger proportion of urine samples from women contain such small numbers, some chlamydia-positive patients are misdiagnosed. Despite this, immunoassays still occupy a diagnostic niche largely because of cost saving.

By enabling enormous amplification of a DNA sequence specific to the chlamydial species, polymerase chain reaction (PCR), ligase chain reaction (LCR), transcription mediated, and some other molecular assays have overcome the problem of poor sensitivity and may provide evidence for the existence of chlamydiae in chronic or treated disease when viable or intact organisms no longer exist. These sensitive assays have replaced culture as the 'gold standard' and have a place not only in research, but also in routine diagnosis and in promoting and maintaining effective screening programmes.

Serological tests

The complement fixation test tends not to distinguish between the chlamydial species and, therefore, is used infrequently. Most of the pertinent diagnostic information has come through the use of the microimmunofluorescence test by which class-specific antibodies (IgM, IgG, IgA, or secretory) may be measured. However, a fourfold or greater increase in the titre of antibody (IgM and/or IgG) is detected infrequently so that the value of serology in the diagnosis of chlamydial infections in individual patients is limited. A good correlation has been found between the presence of IgG and/or IgA antibody in tears and the isolation of *C. trachomatis* from the conjunctiva in endemic trachoma and adult ocular paratrachoma. In genital infections, serum antibodies are found frequently in the absence of a current chlamydial infection of the cervix, so that reliance cannot be put on a single serum or local IgA-specific antibody titre to denote a current infection. In pelvic inflammatory disease, especially in the Curtis Fitz-Hugh syndrome, antibody titres tend to be higher than in uncomplicated cervical infections. A very high IgG antibody titre, for example 512 or greater, suggests causation in pelvic disease, but high titres do not always correlate with detection of chlamydiae and are associated more with chronic or recurrent disease. However, specific *C. trachomatis* IgM antibody in babies with pneumonia is pathognomonic of chlamydia-induced disease.

In primary respiratory infections with *C. pneumoniae*, IgM antibody is considered to develop within a few weeks and IgG antibody by 2 months. In repeat infections, IgG but not IgM antibody develops more rapidly and to a greater titre than before. However, when only a single serum is available, it may be difficult to interpret information complicated by cross-reacting antibodies to the other species. Only in children is the finding of *C. pneumoniae* antibody in a single serum sample an assurance of infection with this species. Although the complement fixation test has been used in the past to diagnose lymphogranuloma venereum and psittacosis, as indicated above it is unwise to do so because of its lack of specificity.

Treatment of chlamydial infections

Chlamydiae are particularly sensitive to tetracyclines and macrolides, but also to a variety of other drugs. The rifampicins are probably more active than the tetracyclines *in vitro* but there is evidence for chlamydial resistance to the rifampicins which, in any case, are usually reserved for mycobacterial infections. Tetracycline resistance has been reported but is insufficiently widespread to cause a problem clinically. However, vigilance should be kept for resistant strains that might jeopardize clinical practice, particularly as the move away from cultural diagnostic procedures has made their detection less easy. Of the macrolides,

erythromycin is used most often, particularly to treat chlamydial infections in infants, young children, and in pregnant and lactating women. Azithromycin in a single dose has gained popularity because it seems to be effective and enhances compliance. Other alternatives, such as some of the quinolones, particularly ofloxacin, are effective but have not found regular use.

More detailed recommendations for dose and duration of antibiotic treatment are presented in Table 2. The principle of giving systemic treatment as well as topical to eradicate nasopharyngeal carriage in trachoma applies also in neonatal chlamydial conjunctivitis, where topical treatment provides no additional benefit. Oral erythromycin should be given to treat the conjunctivitis and to prevent the development of pneumonia. Azithromycin in a single oral dose (20 mg/kg) has been shown to be as effective as 6 weeks of topical tetracycline for active trachoma and may well be the drug of choice. Azithromycin as a single 1-g oral dose has also been shown to be effective in treating non-gonococcal urethritis. In complicated genital tract infections such as epididymo-orchitis and pelvic inflammatory disease, treatment will almost certainly be needed before a microbiological diagnosis can be established, following which additional broad-spectrum antibiotic cover may be required. In the case of *C. pneumoniae* and *C. psittaci* infections, treatment follows the same principles as for *C. trachomatis* infections. Finally, it should be kept in mind that treatment is likely to be most effective when given over a long rather than short time, suboptimal doses are avoided, compliance is strict, and when, in the case of genital infections, partners of patients are also treated.

Immune response and pathogenesis

The immune response to chlamydial infections may be protective or damaging, much of the pathology being immunologically mediated. The hallmark of chlamydial infection, whatever the anatomical site, is the lymphoid follicle. Follicles contain typical germinal centres, consisting predominantly of B lymphocytes, with T cells, mostly CD8 cells, in the parafollicular region. Between follicles the inflammatory infiltrate contains plasma cells, dendritic cells, macrophages, and polymorphonuclear leucocytes in addition to T and B lymphocytes. The late stage of chlamydial infection is characterized by fibrosis, seen typically in trachoma and pelvic inflammatory disease. T lymphocytes are also present and outnumber B cells and macrophages. Biopsies taken from patients with cicatricial trachoma and persisting inflammatory changes show a predominance of CD4 cells, but those from patients in whom inflammation has subsided contain mainly CD8 cells.

Repeated ocular infection by chlamydiae induces progressively worse disease with a diminished ability to isolate the organisms, features noted both naturally and experimentally. There is also experimental evidence that such events occur in the genital tract. For example, primary inoculation of the oviducts of pig-tailed macaques with *C. trachomatis* produced a self-limiting salpingitis with minimal residual damage, whereas repeated tubal inoculation caused hydrosalpinx formation with adnexal adhesions. In the cynomolgus monkey model, a similar exaggerated inflammatory response was effected by the genus-specific 57-kDa protein that has sequence homology with the GroEL heat-shock protein of *Escherichia coli*. It is thought that the damaging sequelae of chlamydial infections, such as scarring in trachoma, tubal adhesions following pelvic inflammatory disease, and reactive arthritis consequent to urethritis, may result from soluble mediators of inflammation in response to the 57-kDa protein. Thus, it is possible that interferon-g secreted by lymphocytes from immune subjects, together with other cytokines, particularly those that stimulate fibroblast activity, such as interleukin 1 and tumour necrosis factor-b, may play a part in the scarring process.

The epidemiology of trachoma suggests that protective immunity follows natural infection, as active disease is uncommon in adults in endemic areas, and *C. trachomatis* can rarely be isolated from them. Similarly, the chlamydial isolation rate for men with non-gonococcal urethritis is lower in those who have had previous episodes. Furthermore, women who have had cervical chlamydial infections accompanied by IgM and IgG antibodies to *C. trachomatis* are less likely to develop salpingitis than those without such antibodies. These observations and the results of animal experiments indicate that chlamydial infection of the eye, the genital tract, and also the respiratory tract provides moderate resistance to reinfection. Nevertheless, attempts to develop an effective vaccine against any of the chlamydial species have, so far, been unsuccessful, although DNA vaccines hold out some hope.

Further reading

Black CM (1997). Current methods of laboratory diagnosis of *Chlamydia trachomatis* infection. *Clinical Microbiology Reviews* **10**, 160–84.

Grayston JT *et al.* (1990). A new respiratory tract pathogen: *Chlamydia pneumoniae* strain TWAR. *Journal of Infectious Diseases* **161**, 618–25.

Mabey DCW, Bailey RL, Hutin YJF (1992). The epidemiology and pathogenesis of trachoma. *Review of Medical Microbiology* **3**, 1–8.

Perine PL, Stamm WE (1999). Lymphogranuloma venereum. In: Holmes KK *et al.*, eds. *Sexually transmitted diseases*, 3rd edn, pp 423–32. McGraw-Hill, New York.

Rasmussen SJ (1998). Chlamydial immunology. *Current Opinion in Infectious Diseases* **11**, 37–41.

Schachter J *et al*. (1999). Azithromycin in control of trachoma. *Lancet* **354**, 630–5.

Stephens RS *et al*. (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis. Science* **282**, 754–9.

Stephens RS, ed. (1999). *Chlamydia. Intracellular biology, pathogenesis and immunity*. American Society for Microbiology, Washington DC.

Taylor-Robinson D (1991). Genital chlamydial infections: clinical aspects, diagnosis, treatment and prevention. In: Harris JRW, Forster SM, eds. *Recent advances in sexually transmitted diseases and AIDS*, pp 219–62. Churchill Livingstone, Edinburgh.

Taylor-Robinson D (1997). Evaluation and comparison of tests to diagnose *Chlamydia trachomatis* genital infections. *Human Reproduction* **12**, 113–20.

Taylor-Robinson D, Thomas BJ (1998). *Chlamydia pneumoniae* in arteries: the facts, their interpretation, and future studies. *Journal of Clinical Pathology* **51**, 793–7.

Thylefors B *et al*. (1987). A simple system for the assessment of trachoma and its complications. *Bulletin of the World Health Organization* **65**, 477–83.

Zhang D-J *et al*. (1997). DNA vaccination with the major outer membrane protein gene induces acquired immunity to *Chlamydia trachomatis* (mouse pneumonitis) infection. *Journal of Infectious Diseases* **176**, 1035–40.

7.11.41

Mycoplasmas

D. Taylor-Robinson

Characteristics of mycoplasmas

Mycoplasmas, originally called pleuropneumonia-like organisms (**PPLO**), are the smallest free-living micro-organisms. They lack a rigid cell wall seen in other bacteria so that they are resistant to penicillins and other antimicrobials which act on this structure. Instead, they are bounded by a pliable unit membrane (Fig. 1), which encloses the cytoplasm, DNA, RNA, and other metabolic components necessary for propagation in cell-free media. Despite their general similarity, mycoplasmas comprise a heterogeneous group of micro-organisms that differ from one another in DNA composition, nutritional requirements, metabolic reactions, antigenic composition, and host specificity. Taxonomically, mycoplasmas are divided into four orders, namely the Mycoplasmatales, the Entomoplasmatales comprising those from insects and plants, the Acholeplasmatales, which do not require sterol for growth, and the oxygen-sensitive, strictly anaerobic Anaeroplasmatales, one genus of which also does not need sterol. The mycoplasmas isolated commonly from humans belong to the family Mycoplasmataceae within the order Mycoplasmatales. This family comprises the genus *Mycoplasma*, which contains organisms that metabolize glucose or arginine or both, but not urea, and the genus *Ureaplasma*, the organisms of which hydrolyse urea uniquely. The latter originally were termed T-strains or T-mycoplasmas because of the tiny (T) colonies they form on agar medium, in contrast to the larger characteristic 'fried-egg'-like colonies produced by most other mycoplasmas (Fig. 2).

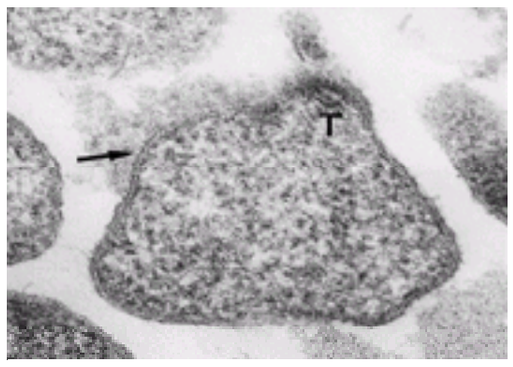

Fig. 1 Electron micrograph of *M. pulmonis* (murine origin), illustrating that the organism does not have a bacterial cell wall but has a trilaminar unit membrane (arrow); also note what appears to be a terminal structure (T). × 66 000.

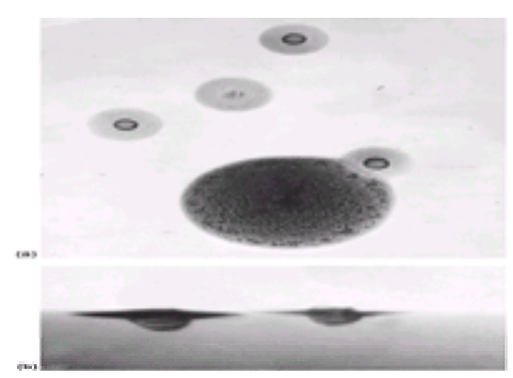

Fig. 2 (a) 'Fried-egg'-like mycoplasma colonies (one not well formed) and a larger bacterial colony. Transmission light microscopy, × 43. (b) Section through mycoplasma colonies illustrating growth in the depth of the agar. × 78.

The small size of the mycoplasma genome restricts their metabolic capabilities. Nevertheless, apart from their importance in humans, certain mycoplasma species are of economic importance because of the pneumonia, arthritis, keratoconjunctivitis, and mastitis they cause among livestock and poultry in Africa, Australia, and other parts of the world. Furthermore, a number of species are a laboratory nuisance as occult contaminants of cell cultures.

Occurrence of mycoplasmas in man

Sixteen mycoplasma species have been isolated from humans, but only 14 constitute the normal flora or behave as pathogens (Table 1). Most of them are found in the oropharynx. There is little information, as yet, about the distribution or significance of *M. penetrans*, *M. pirum*, and *M. spermatophilum*.

Respiratory infections

The relation between mycoplasmas and respiratory disease

M. pneumoniae is the most important mycoplasma found in the respiratory tract (see below), most of them behaving as commensals (Table 1). *M. genitalium* was found originally in the male genitourinary tract but was subsequently isolated from a few respiratory specimens, which also contained *M. pneumoniae*. However, the significance of *M. genitalium* in the respiratory tract remains to be determined. *M. fermentans* has been detected in the throat more often than hitherto because of the use of the polymerase chain reaction (**PCR**) (see later) and has been recovered from adults with an acute influenza-like illness, which sometimes deteriorates rapidly with development of an often fatal respiratory distress syndrome. *M. hominis*, on the other hand, shows little virulence. Despite it causing a mild exudative pharyngitis in adult male volunteers given large numbers of organisms orally, it has not been shown to cause naturally occurring sore throats in children or adults.

In the late 1930s, non-bacterial pneumonias were first recognized and brought under the heading of primary atypical pneumonia to distinguish them from typical lobar pneumonia. Gradually, primary atypical pneumonia was recognized to be aetiologically heterogeneous and, in one variety, cold agglutinins often developed. It was from this form of disease that an infectious agent was isolated in embryonated eggs. This micro-organism, the 'Eaton agent', produced pneumonia in cotton rats and hamsters, and was thought first to be a virus. However, this was seriously doubted when it was found to be affected by chlortetracycline and gold salts, and its mycoplasmal nature was established finally by cultivation on a cell-free agar medium. The agent was subsequently called *M. pneumoniae* and its ability to cause respiratory disease was established fully by studies based on isolation, serology, volunteer inoculation, and vaccine protection.

M. pneumoniae disease manifestations

M. pneumoniae produces a spectrum of effects ranging from inapparent infection and mild, afebrile, upper respiratory-tract disease to severe pneumonia. Clinical manifestations are often not sufficiently distinctive to permit an early definitive diagnosis of mycoplasmal pneumonia. Indeed, this shares the features of other non-bacterial pneumonias in that general symptoms, such as malaise and headache, often precede chest symptoms by 1 to 5 days, and radiographic examination frequently reveals evidence of pneumonia before physical signs, such as rales, become apparent. Usually, only one of the lower lobes is involved and the radiograph most often shows patchy opacities. About 20 per cent of patients suffer bilateral pneumonia, but pleurisy and pleural effusions are unusual. The course of the disease is variable but often protracted. Thus, cough, abnormal chest signs, and changes in the radiograph may persist for several weeks and relapse is a feature. The organisms may also persist in respiratory secretions despite antibiotic therapy, particularly in patients with hypogammaglobulinaemia, where excretion may continue for months or years rather than weeks. Although a few very severe infections have been reported, occurring usually in patients with immunodeficiency or sickle-cell anaemia, death has been rare. In children, infection has been characterized occasionally by a prolonged illness with paroxysmal cough followed by vomiting, thus simulating the features of whooping cough.

Extrapulmonary manifestations

Disease caused by *M. pneumoniae* is limited usually to the respiratory tract, but various extrapulmonary conditions may occur during the course of the respiratory illness or as a sequel to it. These complications and an estimation of their frequency are shown in Table 2. Whether any of them might be due to *M. genitalium* is a moot point. Haemolytic anaemia with crisis is brought about by the development and action of cold agglutinins (anti-I antibodies), the organisms apparently altering the I antigen on erythrocytes sufficiently to stimulate an autoimmune response. It is possible that some of the other clinical conditions, such as the neurological complications, may arise in a similar way. However, invasion of the central nervous system cannot be discounted as *M. pneumoniae* has been isolated from cerebrospinal fluid.

Microbiological diagnosis of *M. pneumoniae* infection

This depends on detection by culture or molecular methods and/or serology. The usual medium employed for isolation of *M. pneumoniae* consists of PPLO broth, 20 per cent horse serum, and 10 per cent (v/v) fresh yeast extract (25 per cent w/v). However, a more sensitive medium (SP4) is that used first for the isolation of spiroplasmas, comprising a conventional mycoplasma broth medium with fetal calf serum and a tissue-culture supplement. Either medium is supplemented with a broad-spectrum penicillin, and glucose, with phenol red as a pH indicator. The fluid medium, inoculated with sputum, throat washing, pharyngeal swab, or other specimen, is incubated at 37 °C and a colour change (red to yellow), which occurs usually within 4 to 21 days, signals the fermentation of glucose (Table 1), with production of acid, owing to multiplication of the organisms. This preliminary identification may be confirmed after subculturing to agar medium, usually by demonstrating inhibition of colony development around discs impregnated with specific antiserum (Fig. 3) or by immunofluorescence of colonies with an *M. pneumoniae*-specific antibody.

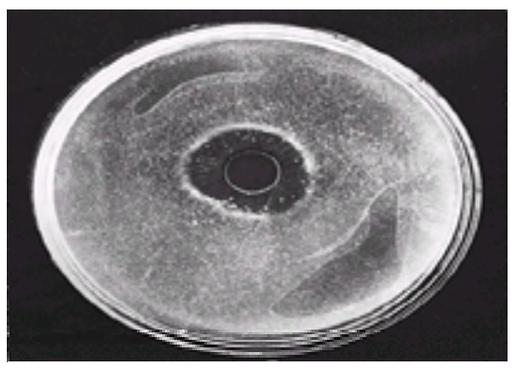

Fig. 3 Mycoplasma identification by agar growth inhibition. Colony development inhibited around a filter-paper disc impregnated with specific antiserum. Note also antibody–antigen precipitation at edge of inhibition zone.

Culture and identification are slow, but rapid PCR determination of *M. pneumoniae* positivity and then continued culture of only those specimens that are PCR-positive makes for a speedier diagnosis. However, this approach, or the use of the PCR to test specimens directly, is not routine, reliance still being placed on serology for diagnosis. Testing by complement fixation is undertaken in many laboratories and a fourfold or greater rise in antibody titre with a peak at about 3 to 4 weeks after the onset of disease is said to occur in about 80 per cent of cases and be indicative of a recent infection. A titre of 1:128 or greater in a single serum is suggestive but not proof of infection in the previous few weeks or months; a fourfold or greater fall in antibody titre, perhaps over 6 months, may be helpful but, sometimes, may be difficult to relate to a particular prior illness. It must be remembered that the complement-fixation test does not distinguish between *M. pneumoniae* and *M. genitalium* and the occasional occurrence of the latter in the respiratory tract may jeopardize a specific diagnosis of *M. pneumoniae* infection. More specific, perhaps, is the microimmunofluorescence test in which IgM antibody is sought; its presence provides some confidence in making an accurate diagnosis of a current infection or one within the previous few weeks. The same comment applies to a commercially available enzyme immunoassay specific for IgM. Cold agglutinins, detected by agglutination of O Rh-negative erythrocytes at 4 °C, develop in about half the patients and a titre of 1:128 or greater is suggestive of a recent *M. pneumoniae* infection. However, such agglutinins are occasionally induced by a number of other conditions, but the ability of *M. genitalium* to do so is unknown.

Epidemiology of *M. pneumoniae* infections

The consequence of infection depends on age, about a quarter of infections in persons 5 to 15 years old resulting in pneumonia, with about 7 per cent of infections in young adults doing so. Thereafter, pneumonia is even less frequent, but generally is more severe the older the patient.

Although *M. pneumoniae* causes inapparent and mild upper respiratory-tract infections more commonly than severe disease, it is responsible for only a small proportion of all upper respiratory-tract disease, most of it being of viral or streptococcal aetiology. It plays a relatively greater part in producing lower respiratory-tract disease together with *Chlamydia pneumoniae* and various other bacteria. Thus, in the United States, it has been calculated that in a large general population, the proportion of all pneumonias due to *M. pneumoniae* is about 15 to 20 per cent, and in certain populations, for example military recruits, it has been responsible for as much as 40 per cent of acute pneumonic illness.

M. pneumoniae infections have been reported from every country where appropriate diagnostic tests have been undertaken. Infection is endemic in most areas and occurs during all months of the year, with a predilection for late summer and early autumn. However, epidemic peaks have been observed about every 4 to 7 years in some countries. The incubation period ranges from 2 to 3 weeks and spread from person to person occurs slowly, usually where there is continual or repeated close contact, for example in a family.

Immunopathological factors in the development of *M. pneumoniae* pneumonia

Adherence of *M. pneumoniae* organisms to respiratory mucosal epithelial cells (Fig. 4) is a crucial factor in the pathogenesis of disease. After cytadsorption, mediated by P1, P30, and possibly up to seven other proteins on the surface of the organisms, immune mechanisms are important in the development of *M. pneumoniae* pneumonia in humans, which rarely causes death. Thus, the histopathological picture is derived mainly from infection in other animals. The pneumonic infiltrate is predominantly a peribronchiolar and perivascular accumulation of lymphocytes, most of which are thymus dependent (Fig. 5). The importance of cell-mediated immune mechanisms is indicated by the fact that immunosuppression in hamsters results in ablation of the pneumonia or a decrease in its severity. The development of a cell-mediated immune response to *M. pneumoniae* has been shown further by positive lymphocyte transformation, macrophage migration inhibition, and delayed-hypersensitivity skin tests. A polysaccharide–protein fraction of the organisms is involved in this response rather than the glycolipid that is the main antigenic determinant in complement-fixation and other serological reactions. The initial lymphocyte response is followed by a change in the character of the bronchiolar exudate, with polymorphonuclear leucocytes and macrophages predominating. The rather slow development of these events on primary infection contrasts with an accelerated and often more intense host response seen on reinfection. To at least some extent, therefore, the pneumonia caused by *M. pneumoniae* is an immunopathological process. Children of 2 to 5 years of age often possess mycoplasmacidal antibody, suggesting infection at an early age and the notion that the pneumonia that occurs in older persons is an immunological over-response to reinfection, the lung being infiltrated by previously sensitized lymphocytes.

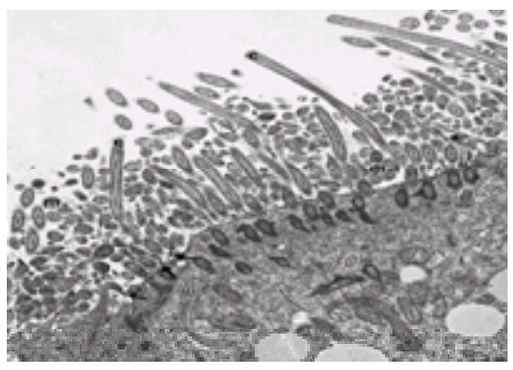

Fig. 4 Electron micrograph of ciliated epithelial cells in the tracheal mucosa of a hamster infected with *M. pneumoniae*. Note cilia (c) and individual organisms (m), some with specialized terminal structure oriented towards the membrane of the host cell (arrows). × 9880.

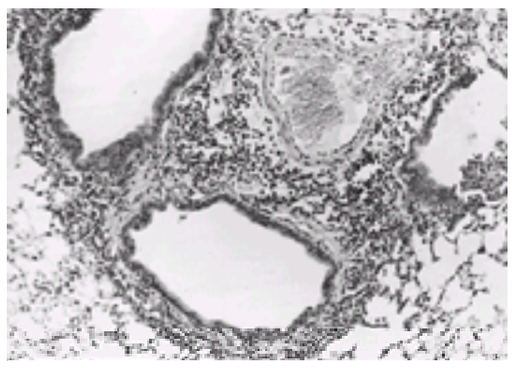

Fig. 5 Pneumonia 2 weeks after intranasal inoculation of a hamster with *M. pneumoniae*. Note peribronchiolar and perivascular infiltration of mononuclear cells, predominantly lymphocytes. Haematoxylin and eosin, × 98.

Treatment

M. pneumoniae, like other mycoplasmas, is sensitive to the tetracyclines and apparently more sensitive to erythromycin than the other mycoplasmas of human origin. It is also inhibited by the newer macrolides, such as clarithromycin and azithromycin, and the newer quinolones, such as sparfloxacin. The value of tetracyclines was shown first in a controlled trial of dimethylchlortetracycline in United States marine recruits, a dose of 300 mg three times daily for 6 days significantly reducing the duration of fever, pulmonary infiltration, and other signs and symptoms. Planned trials are the most favourable for determining the value of antimicrobials but in civilian practice they prove less effective, probably because disease is often well established before treatment is instituted. Despite this, treatment with an antimicrobial is worthwhile. For pregnant women and children it is advisable to use erythromycin rather than a tetracycline, and the former has sometimes proved more effective than a tetracycline in adults. Successful treatment of clinical disease, however, is not always accompanied by early eradication of the organisms from the respiratory tract, probably because they can become intracellular and the drugs only inhibit their multiplication and do not kill them. These are possible reasons for relapse in some patients and plausible ones for recommending a 2- to 3-week course of treatment. It is a moot point whether early treatment prevents some of the complications but, nevertheless, it should commence as soon as possible. As laboratory confirmation of *M. pneumoniae* infection sometimes may be slow, it would seem wise to start antimicrobial treatment on the basis of the clinical evidence and a suggestive antibody titre in a single serum sample despite the drawbacks, mentioned previously, of making a diagnosis serologically.

The true value of corticosteroids is in doubt, although, in conjunction with antimicrobials, they appear to have been helpful in patients with severe pneumonia and erythema multiforme.

Prevention

M. pneumoniae infection or disease may occur despite high titres of serum mycoplasmacidal antibody. Furthermore, mycoplasmal infection of the respiratory tract of laboratory animals may stimulate only a weak antibody response and yet induce greater resistance to reinfection and disease than parenteral inoculation with organisms that stimulate much higher titres of serum antibodies. Such observations have led to the belief that local immune factors are crucial in resistance. The correlation between the resistance of adult volunteers to *M. pneumoniae* disease and the presence of IgA antibody in respiratory secretions is consistent with this contention. This antibody could provide the first line of defence by preventing attachment of the organisms to respiratory epithelial cells.

So far as vaccination is concerned, the efficacy of formalin-inactivated *M. pneumoniae* vaccines in preventing pneumonia caused by this mycoplasma has ranged from 28 to 67 per cent in field trials. The failure of some killed *M. pneumoniae* vaccines to protect fully may have been due to poor antigenicity, but others induced serum antibody levels similar to those that develop after natural infection. This suggests that the relatively poor protection afforded by the killed vaccines may have been due to their inability to stimulate cell-mediated immunity and/or local antibody. With local antibodies in mind, live attenuated vaccines, particularly those based on temperature-sensitive mutants of *M. pneumoniae*, were developed. The organisms could multiply at the temperature of the upper, but not the lower, respiratory tract and some vaccines produced pulmonary infection in hamsters without causing pathological changes, and induced significant resistance to subsequent challenge with virulent wild strains of *M. pneumoniae*. However, because the same mutants produced some disease in human volunteers they were considered unacceptable for general human use. Recombinant DNA vaccines involving P1 and other proteins, or a recombinant vaccine developed by cloning a component of the *M. pneumoniae*

P1 gene into an adenovirus vector may offer greater success.

Chronic respiratory disease

As mycoplasmas of animals are frequently involved in chronic illnesses, the possible role of mycoplasmas in human chronic respiratory disease, particularly chronic bronchitis, is worthy of consideration.

M. pneumoniae frequently persists in the respiratory tract long after clinical recovery and occasionally the respiratory disease it causes has a protracted course. Furthermore, tracheobronchial clearance is very much reduced soon after infection and there is a tendency for slower clearance, in comparison with that in healthy subjects, even 1 year later. Despite this, there is no evidence that *M. pneumoniae* is a primary cause of chronic bronchitis, or that it is responsible for maintaining chronic disease other than by possibly causing some acute exacerbations. This is perhaps the case in some patients, experiencing an acute exacerbation of chronic bronchitis, from whom *M. pneumoniae* has been isolated and in whom a serological response has been seen.

There is no doubt that *M. salivarium*, *M. orale*, and perhaps other mycoplasmas present in the oropharynx of healthy persons spread to the lower respiratory tracts of some who suffer from chronic bronchitis. There is no substantial evidence that these mycoplasmas are a cause of acute exacerbations, but antibody responses to them occur in association with such exacerbations more frequently than at other times, which suggests that the organisms are more antigenic during exacerbations. It is tempting to think that this is probably due to increased mycoplasmal multiplication and participation in tissue damage brought about primarily by viruses and bacteria, and that in this way the mycoplasmas may play some part in perpetuating a chronic condition.

Mycoplasmas in the vagina are transmitted rarely to the infant *in utero*, but often during birth and *U. urealyticum* organisms (ureaplasmas), in particular, may be isolated from the throats and tracheal aspirates of some neonates. Ureaplasma-infected infants of very low birth-weight (under 1000 g), have died or have developed chronic lung disease twice as often as uninfected infants of similar birth weight or those of over 1000 g. *M. hominis* has also been implicated in pneumonia soon after birth, albeit even more rarely. However, whether these organisms are a cause of the disease in their own right or only together with the various bacteria that are associated with maternal bacterial vaginosis is unresolved. It remains to be seen whether *M. genitalium* might be involved, a possibility that exists because it has been detected in the vagina and cervix.

Genitourinary and related infections

Clinical conditions in which there is evidence strongly suggesting that mycoplasmas have an aetiological role, at least in part, will be considered in some detail. Other diseases in which the role of mycoplasmas is minimal, or the evidence for a mycoplasmal cause is weak and/or contentious, are mentioned briefly. All are summarized in Table 3.

Non-gonococcal urethritis and complications

There have been many studies concerned with the role of large-colony-forming mycoplasmas. It is clear that most of them cannot be considered as causes of non-gonococcal urethritis because they are isolated so rarely from the genitourinary tract either in health or disease (Table 1). However, *M. genitalium* (Fig. 6a) is associated strongly with acute non-gonococcal urethritis, having been detected by use of the PCR in about 25 per cent of such cases compared with a significantly smaller proportion (about 6 per cent) of healthy controls, and almost independently of *Chlamydia trachomatis.* Experimentally, it also causes urethritis in male chimpanzees and, like *M. pneumoniae*, adheres to and enters epithelial cells (Fig. 6b). Such a location could, at least, partially protect *M. genitalium* from antimicrobials and account for its occurrence in about one-quarter of men with persistent or recurrent non-gonococcal urethritis, in some of whom it may be a cause. Although *M. hominis* has been isolated from about 20 per cent of patients with acute non-gonococcal urethritis, it has not been implicated as a cause. Nevertheless, the fact that some cases are associated with bacterial vaginosis in sexual partners in whom *M. hominis* organisms are found in large numbers should not be overlooked. The role of ureaplasmas in non-gonococcal urethritis has been contentious for many years and is discussed below.

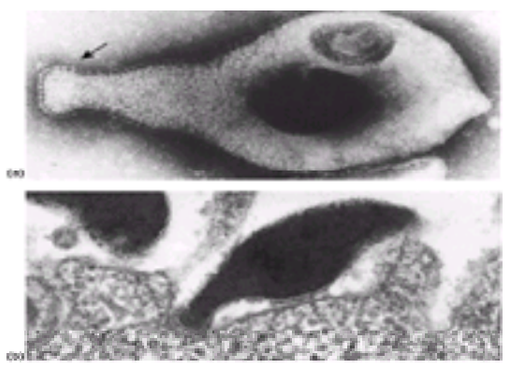

Fig. 6 (a) Electron micrograph of *M. genitalium*, negatively stained, to show flask-shaped appearance and terminal specialized structure (arrow). × 90 000. (Reproduced from Tully *et al.*, 1983, *International Journal of Systematic Bacteriology* **33**, 387, with permission.) (b) Electron micrograph of *M. genitalium* adhering to Vero cell by the terminal structure. × 60 000. (Reproduced from Tully *et al.*, 1983, *International Journal of Systematic Bacteriology* **33**, 387, with permission.)

The results of most studies, based on qualitative estimations, have failed to demonstrate a significant difference between the prevalence of ureaplasmas in men with or without acute non-gonococcal urethritis. However, if ureaplasmas are involved in the pathogenic process, it would be reasonable to expect them to be present in larger numbers in men with disease than if they were behaving only as commensals; a few workers have provided quantitative data to support this idea. In addition, some patients with hypogammaglobulinaemia develop a prolonged urethrocystitis in which persistent ureaplasmal infection seems to be responsible. In support of ureaplasmal pathogenicity in non-gonococcal urethritis are the results of several antimicrobial trials. For example, in one of these a larger proportion of patients responded to minocycline (active against ureaplasmas and *C. trachomatis*) than to rifampicin (active against *C. trachomatis* only), and those infected with ureaplasmas failed to respond to rifampicin significantly more often than those who were not infected. Although these findings might be difficult to explain if ureaplasmas had no involvement in non-gonococcal urethritis, the failure to take *M. genitalium* into account tends to jeopardize this conclusion. About 10 per cent of ureaplasmas are resistant to tetracyclines and the urethritis of patients infected by them sometimes responds only to treatment with antimicrobials, such as erythromycin, to which the organisms are susceptible.

Finally, some ureaplasma strains, unpassaged in the laboratory, produced a mild urethritis and an antibody response in male chimpanzees inoculated intraurethrally, the disease responding to tetracycline therapy. Furthermore, four investigators inoculated themselves intraurethrally and each developed urethritis. In one detailed study, two of them received cloned *U. urealyticum*, serotype 5, which had been isolated from a patient with acute non-gonococcal urethritis in whom no other potentially pathogenic micro-organisms could be detected, although *M. genitalium* was not sought at that time. Both developed symptoms and signs of urethritis which responded to treatment with minocycline. The results of the most recent volunteer experiment suggest that ureaplasmas may cause disease the first few times they gain access to the urethra but later insults result in colonization without disease, accounting perhaps for their frequent occurrence in the urethra of healthy men.

Ureaplasmas have been recovered from the urethra and directly from epididymal aspirate fluid, accompanied by a specific antibody response, in a patient with acute non-chlamydial, non-gonococcal epididymitis, and they may be a rare cause. *M. genitalium* has not been sought. Information suggesting that the prostate becomes infected during the course of an acute ureaplasmal infection of the urethra is scanty, although ureaplasmas have been isolated more frequently and in greater numbers from patients with acute urethroprostatitis than from controls, and most of those with more than 10^3 organisms in expressed prostatic fluid responded to tetracycline therapy. In contrast, ureaplasmas have not been found, and *M. genitalium* only rarely, in prostatic biopsy specimens from patients with chronic abacterial prostatitis, and, in most studies, *M. hominis* has not been associated with prostatitis of any kind.

Pyelonephritis

M. hominis has been isolated, sometimes in pure culture, from the upper urinary tract of almost 10 per cent of patients with acute pyelonephritis. In addition, antibody

to *M. hominis*, measured by indirect haemagglutination, has been demonstrated in serum and urine of some of these patients. In contrast, the mycoplasma has not been found in the upper urinary tract of patients with non-infectious urinary-tract diseases, nor has antibody been detected in their urine. These data have not been confirmed recently, but they suggest that *M. hominis* causes a few cases of acute pyelonephritis or acute exacerbations of chronic pyelonephritis and that ureaplasmas are involved less often if at all.

Pelvic inflammatory disease

Micro-organisms in the vagina and lower cervix may ascend to and cause inflammation of the fallopian tubes and adjacent pelvic structures. Like non-gonococcal urethritis, non-gonococcal pelvic inflammatory disease does not have a single cause and the possibility that infection by genital mycoplasmas might be one cause has engaged the attention of numerous investigators.

M. hominis has figured prominently among reports of the isolation of large-colony-forming mycoplasmas from inflamed fallopian tubes, tubo-ovarian abscesses, and pelvic abscesses or fluid. Swedish workers collected specimens by laparoscopy and found *M. hominis* in the tubes of about 10 per cent of women with salpingitis but not in those of women without signs of the disease, an observation that has found some support with investigators in the United Kingdom.

It is likely, but not certain, that this happens mostly when large numbers of *M. hominis* organisms occur in the vagina as a consequence of bacterial vaginosis. The same comment applies to hysterosalpingography, which may occasionally stimulate inflammation of the fallopian tubes in women who carry *M. hominis* in the lower genital tract.

Ureaplasmas have been studied less intensively, but they have been isolated directly from the fallopian tubes of a very small proportion of patients with acute salpingitis, from pelvic fluid, and from a tubo-ovarian abscess. *M. pneumoniae* is reported to have been isolated also from such an abscess and from the cervix. The significance of these findings is unclear but it would seem that ureaplasmas, at least, are of little importance . Examination of specimens for *M. genitalium* by PCR technology is needed to establish whether it is involved in pelvic inflammatory disease, which is possible since it has been found in the cervix.

Antibody to *M. hominis*, measured by the indirect haemagglutination technique, was found by Swedish workers in about half the patients they studied with salpingitis, but in only 10 per cent of healthy women. Furthermore, a significant rise or fall in antibody titre occurred during the course of disease in more than half of the women who had *M. hominis* in the lower genital tract. Others found that patients with gonococcal pelvic inflammatory disease responded serologically to *M. hominis* more often than those without such disease; they suggested that damage incurred by the gonococci was a factor in the serological response and questioned the primary role of *M. hominis*. However, a response to *M. hominis* has been seen quite often in women in whom gonococci are not the cause of pelvic inflammation. In addition, a significant antibody response to *M. genitalium* has been detected in about one-third of women with pelvic inflammatory disease in whom antibody responses to *M. hominis* and *C. trachomatis* could not be detected. In contrast, antibody responses to *U. urealyticum* have been detected rarely, consistent with the impression that ureaplasmas are of little importance in acute pelvic inflammatory disease.

Fallopian-tube organ cultures, in which the tissues are maintained in a condition similar to that *in vivo*, are particularly useful in assessing the effect of micro-organisms. In such cultures, gonococci destroy the epithelium, and *M. genitalium* causes some damage, whereas *M. hominis* organisms, although multiplying, produce no more than swelling of some of the cilia. No damage has been caused by ureaplasmas of human origin. This differential effect may be a true reflection of the pathogenic potential of these micro-organisms *in vivo*. However, failure to demonstrate damage does not mean necessarily that the organisms are avirulent, because organ-culture studies do not account for the role of the immune system in pathogenesis. Studies in intact animals may be helpful in elucidating this aspect. It is of interest, therefore, that the introduction of *M. hominis* into the oviducts of grivet monkeys and *M. genitalium* into the oviducts of these monkeys, as well as those of marmosets, resulted in a self-limited acute salpingitis and parametritis with an antibody response, whereas ureaplasmas had no effect.

These various data indicate that *M. hominis* and perhaps *M. genitalium* are likely to have a role in some cases of acute pelvic inflammatory disease.

Postabortal fever

The results of various studies suggest a role for *M. hominis* in fever after abortion (see Section 13), the mycoplasma having been isolated from the blood of about 10 per cent of women who had such fever but not from aborting women without fever, nor from normal pregnant women. However, the key point of whether the mycoplasma was isolated in pure culture was not always clear. A rise in the titre of antibody to *M. hominis* has been detected in half the women who become febrile but in only a small proportion of those who have abortions and remain afebrile. Thus, despite the caveat mentioned, on balance the evidence seems in favour of *M. hominis* causing some cases of postabortal fever, whereas there is none to suggest that ureaplasmas do likewise, possibly because studies have not been focused on this micro-organism.

Postpartum fever

Genital mycoplasmas found transiently in maternal blood a few minutes after normal vaginal delivery are not associated with postpartum fever. However, *M. hominis* isolated from the blood a day or more after delivery has been associated in 5 to 10 per cent of women who develop fever, often with an antibody response. As the organisms are seldom isolated from the blood of afebrile postpartum women, the inference is that *M. hominis* induces postpartum fever, assuming they are recovered in pure culture. Whether fever occurs predominantly in women who have bacterial vaginosis in pregnancy, in which there is proliferation of *M. hominis* together with various bacteria in the vagina, has not been resolved. Patients with postpartum or postabortal fever recover usually without antibiotic treatment.

Microbiological diagnosis of genitourinary and related infections

Swabs from the urethra or vagina provide a slightly more sensitive means of collecting specimens for mycoplasmal isolation than urine specimens. The basic medium is similar to that described for the isolation of *M. pneumoniae*, SP4 medium, mentioned previously, being best for *M. genitalium*. Advantage is taken of the metabolic activity of the mycoplasmas (Table 1) in order to detect their growth. Clinical material is added to separate vials of liquid medium containing phenol red and 0.1 per cent glucose, arginine, or urea. *M. genitalium* metabolizes glucose and changes the colour of the medium from red to yellow. *M. fermentans* does this also but, in addition, converts arginine to ammonia, as do *M. hominis* and *M. primatum*. Ureaplasmas possess a urease that breaks down urea to ammonia too. In each case, the pH of the medium increases and there is a colour change from yellow to red. Ureaplasmas change the colour usually within 1 to 2 days, while *M. genitalium* may take 50 days or longer. Indeed, it soon became clear that attempts to culture this mycoplasma failed and detection was salvaged only by the implementation of PCR technology. Introduction of specimens into Vero cell cultures with subculture to mycoplasmal medium when the PCR signifies multiplication is laborious but has been a successful strategy. A PCR assay is also useful for the detection of *M. fermentans* and *U. urealyticum.* Conventionally, however, when colour changes have occurred in liquid medium, subculture to agar medium results in the formation of colonies of about 200 to 300 µm diameter by most genital mycoplasmas; those of *M. genitalium* are usually smaller. Ureaplasma colonies are very small (15 to 60 µm) (Fig. 7(a)) and on medium containing manganous sulphate are brown in colour and, therefore, are detected more easily (Fig. 7(b)). On ordinary blood agar, *M. hominis*, but not ureaplasmas, produces non-haemolytic pinpoint colonies, the nature of which can be established by the methods outlined above. The metabolism-inhibition technique may be used to detect antibodies to *M. hominis* and the ureaplasmas. Antibody inhibits multiplication and metabolism of homologous organisms, thus preventing a change in colour of the medium due to organism multiplication. The indirect haemagglutination technique has been used to detect antibody to *M. hominis* and the microimmunofluorescence test to detect antibody to *M. genitalium*, but these and other tests, such as the enzyme immunoassay, tend to be used as research tools only.

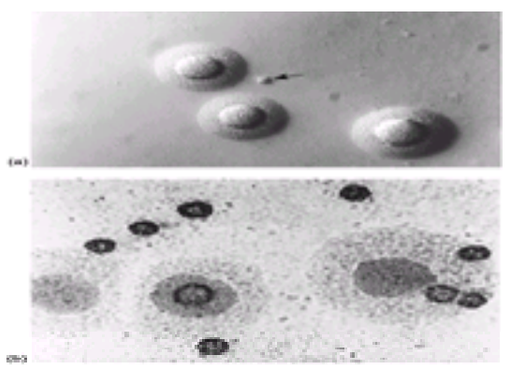

Fig. 7 (a) Colony of *Ureaplasma urealyticum* (15 µm diameter) (arrow) adjacent to colonies of *M. hominis* (90 µm diameter) grown from urethral exudate. Oblique light, × 68. (b) Dark colonies of *U. urealyticum* with colonies of *M. hominis* on agar containing manganous sulphate. × 136.

Joint infections

Rheumatoid arthritis

The fact that mycoplasmas cause several animal arthritides, and that gold salts inactivate mycoplasmas and have had a beneficial effect on rheumatoid arthritis, provided the impetus to search for mycoplasmas in the joints of persons suffering from this disease. However, numerous attempts over the second half of the twentieth century failed to detect mycoplasmas in rheumatoid joints or produced inconsistent or unrepeatable results, those mycoplasmas (*M. hyorhinis*,*M. hominis*) that were recovered by means of cell-culture techniques being nothing more than culture contaminants. The case made in the late 1960s and early 1970s, based on apparent isolation and immunological observations, that *M. fermentans* was important in rheumatoid arthritis was not substantiated in the next 30 years. However, renewed interest has been generated because this mycoplasma, and ureaplasmas, have been found by PCR technology in the joints of more than 20 per cent of patients with rheumatoid arthritis and other chronic inflammatory disorders, that is significantly more than in those with non-inflammatory disorders. The significance of the findings needs to be deter-mined.

M. pneumoniae and other mycoplasmal infections

A feature of the mycoplasmal arthritides of animals is that the mycoplasmas isolated from the joints are found often in the respiratory tract. The question arises, therefore, of whether the known human respiratory pathogenic mycoplasma, namely *M. pneumoniae*, causes arthritis. There is no doubt that respiratory infection is often accompanied by non-specific arthralgia or myalgia (Table 2) during the acute phase, and occasionally it leads to migratory polyarthritis affecting middle-sized joints in adults. A fourfold or greater rise in the titre of antibody to *M.pneumoniae* has been seen occasionally in juvenile chronic arthritis, but an aetiological association has not been demonstrated.

M. hominis has been isolated from septic joints, usually hip, that have developed in mothers after childbirth. The arthritis responds to tetracycline therapy and the diagnosis should be considered in a postpartum arthritis which is unaffected by penicillin.

Reiter's disease (see also Section 21)

Arthritis may occur soon after or concomitant with non-gonococcal urethritis (sexually acquired reactive arthritis; **SARA**) or the arthritis may be associated with conjunctivitis and urethritis (Reiter's disease). In defining the role of mycoplasmas, a difficulty is that the patients have often been treated with antimicrobials before microbiological investigation. However, the possible role of *M. genitalium* and ureaplasmas should not be ignored in view of the strong involvement of the former and the weaker involvement of the latter in uncomplicated non-gonococcal urethritis. *M. genitalium* has been detected in the synovial fluid of a patient with SARA, but further evidence is required to establish a causal link in the case of this mycoplasma, and ureaplasmas too. The latter have not been isolated from involved joints but synovial lymphocytes from some patients have been shown to proliferate *in vitro* in response to ureaplasmal antigens.

Arthritis in patients with hypogammaglobulinaemia

Arthritis of mycoplasmal aetiology (Fig. 8(a) and Fig. 8(b)) should be considered in patients with hypogammaglobulinaemia (see Section 5) who develop an abacterial septic arthritis. Thus, *M. pneumoniae* (together with *M. genitalium* in one instance), *M. hominis*, *M. salivarium*, and, in particular, ureaplasmas have been isolated from synovial fluids of at least two-fifths of these patients. Furthermore, in view of diminished immunity, vigilance should be kept for infection by mycoplasmas of non-human origin. The arthritis usually responds to tetracyclines or other antimicrobials to which the organisms are sensitive, an indication that they are a cause of the disease. Intravenous therapy may be required and administration of antiserum prepared specifically against the mycoplasma in question may be helpful in the few patients whose disease does not respond to antimicrobial therapy.

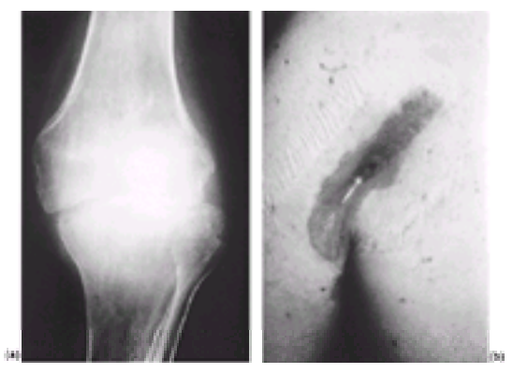

Fig. 8 (a) Damage to the knee joint of a hypogammaglobulinaemic patient caused by *U. urealyticum* infection. (b) Sinus connected with the shoulder joint of a patient with hypogammaglobulinaemia; ureaplasmas were isolated repeatedly from the sinus exudate. (By courtesy of A.D.B. Webster.)

Conditions of rare or equivocal mycoplasmal aetiology

The occurrence of *M. hominis* or ureaplasmas in cerebrospinal fluid from cases of neonatal meningitis or brain abscess is due presumably to infection *in utero* or to colonization at birth with subsequent invasion. This is a rare event but should be considered in cases of neonatal disease of the central nervous system in which the results of bacteriological staining and culture are negative.

M. hominis organisms, apart from inducing fever after abortion or childbirth, have been associated with fever attributed to burns and trauma, and have been implicated in some wound infections. Whether they have any role in the pathogenesis of bacterial vaginosis in which they occur in large numbers is difficult to define because they do so together with a variety of bacteria, which are found also in profusion. *M. hominis* has been associated with premature labour, but in view of this pregnancy outcome being associated strongly with bacterial vaginosis, the involvement of the mycoplasma would seem to be as part of the latter condition.

Data on ureaplasmas in urinary calculi suggest that they could be involved aetiologically, and these organisms have been associated also with infertility, chorioamnionitis, spontaneous abortion, and low birth weight. It is noteworthy, however, that like *M. hominis* organisms, ureaplasmas are found in larger numbers in the vagina of women with bacterial vaginosis than in those without disease. Thus, it cannot be ignored that, in the reproductive problems referred to, ureaplasmas may be involved, at least in part, as a component of bacterial vaginosis and to attribute the problems to ureaplasmas alone may sometimes be misleading.

Association of mycoplasmas with AIDS

The idea that mycoplasmas might act as a cofactor, enhancing human immunodeficiency virus type 1 (**HIV-1**) replication and accelerating progression to AIDS, was fuelled initially by studies *in vitro*. In these, treatment of HIV-infected cell cultures with tetracyclines or fluoroquinolones, active against mycoplasmas, inhibited cell killing without affecting virus replication. In other studies, certain mycoplasmas (*M. fermentans*, *A. laidlawi*i) enhanced cytopathic changes by HIV-1. Such *in vitro* observations, however, were preceded by attempts to identify a virus in Kaposi's sarcoma tissue which culminated in the recovery of a mycoplasma in cell culture, possibly a cell-culture contaminant. This was termed initially ' *M. incognitus*', but identified later as *M. fermentans*, and was found to be distributed widely in tissues

taken at autopsy from patients with AIDS. Subsequently, it was shown to be linked strongly with AIDS-associated nephropathy. In addition, some investigators, who used PCR technology, detected *M. fermentans* in peripheral blood monocytes, throat, and urine of 10, 23, and 8 per cent of HIV-seropositive patients, respectively, almost all of whom were homosexual men. They probably had the mycoplasma before acquiring the virus because the former was detected with similar frequency in samples taken from HIV-seronegative patients, a large proportion of whom were homosexual men, attending a sexually transmitted disease clinic. The interaction of mycoplasmas with the immune system could induce cytokines and so enhance HIV replication with increased loss of CD4+ cells, in this way the mycoplasma acting as a cofactor. Nevertheless, no association was found between infection by *M. fermentans* and the stage of the disease, the patients' CD4+ count, or the viral load. This does not eliminate the possibility that a mycoplasmal infection could influence the speed of disease progression. However, this seems unlikely as no significant difference has been found between the proportion of non-progressors, slow progressors, or rapid progressors of HIV-associated disease who have peripheral blood monocytes that are positive for *M. fermentans*.

A recently discovered mycoplasma of human origin, *M. penetrans*, was isolated from urine sediments of a small number of homosexual men infected with HIV-1, most of whom had AIDS. This mycoplasma avidly invades eukaryotic cells, and antibody to it, detected by an enzyme-linked immunosorbent assay, was found in the sera of 40 per cent of patients with AIDS, but in only a very small proportion of HIV-seronegative subjects, or subjects attending sexually transmitted disease clinics, and in none of a group of patients with other immune dysfunctions. However, the mycoplasma has not been found by PCR technology in the peripheral blood monocytes of a large number of HIV-positive homosexual men. Thus, despite its ability to penetrate cells and its apparent association with HIV infection and AIDS, there is no evidence that *M. penetrans* behaves as a cofactor in the development of AIDS.

Further reading

Maniloff J, ed. (1992). *Mycoplasmas. Molecular biology and pathogenesis*. American Society for Microbiology, Washington, DC.

Razin S, Tully JG, ed. (1996). *Molecular and diagnostic procedures in mycoplasmology*, Vol. 1,*Molecular characterization.* Academic Press, London.

Razin S, Yogev D, Naot Y (1998). Molecular biology and pathogenicity of mycoplasmas. *Microbiology and Molecular Biology Reviews* **62**, 1094–156.

Taylor-Robinson D (1989). Genital mycoplasma infections. In: Judson FN, ed. *Clinics in laboratory medicine. Sexually transmitted diseases*, Vol. 9, pp 501–23. Saunders, Philadelphia.

Taylor-Robinson D (1996). Infection due to species of *Mycoplasma* and *Ureaplasma*: an update. *Clinical Infectious Diseases* **23**, 671–84.

Taylor-Robinson D (1996). Mycoplasmas and their role in human respiratory tract disease. In: Myint S, Taylor-Robinson D, eds. *Viral and other infections of the human respiratory tract*, pp 319–39. Chapman & Hall, London.

Taylor-Robinson D, Bebear C (1997). Antibiotic susceptibilities of mycoplasmas and treatment of mycoplasmal infections. *Journal of Antimicrobial Chemotherapy* **40**, 622–30.

Taylor-Robinson D, Keat A (2001). How can a causal role for small bacteria in chronic inflammatory arthritis be established or refuted? *Annals of Rheumatic Diseases* **60**, 177–84.

Taylor-Robinson D, Gilroy CB, Jensen JS (2000). The biology of *Mycoplasma genitalium. Venereology* **13**, 119–27.

Tully JG, Razin S, ed. (1996). *Molecular and diagnostic procedures in mycoplasmology*, Vol. 2,*Diagnostic procedures.* Academic Press, London.

7.11.42Newly identified and lesser known bacteria

J. Paul

References

Specialists in clinical microbiology and infectious diseases aim, among other things, to detect and characterize novel pathogens and their disease associations and to refine our knowledge regarding the natural history and treatment of known infections. At the time of writing, at least 890 species of bacteria, with names which have standing in nomenclature, in 193 genera, plus a number of other less well characterized taxa, have been reported to be associated in some way with human disease (Table 1). The list includes a core assemblage of bacteria long known to be associated with infection, some of which are widely known (e.g. *Staphylococcus aureus*), others of which are of restricted geographical distribution (e.g. *Bartonella bacilliformis*, the agent of Oroya fever) or seldom encountered (e.g. *Erysipelothrix rhusiopathiae*, the cause of erysipeloid). Advances in laboratory methods have made it possible to associate novel pathogens with well-known clinical conditions (e.g. *Tropheryma whippelii* and *Arthrobacter* spp. with Whipple's disease; and *Bartonella henselae* and *Afipia felis* with cat scratch disease). Members of the human commensal flora from various body sites have been associated with localized infections (e.g. *Acidaminococcus fermentans*, *Bilophila wadsworthia*, and *Buttiauxella agrestis* from the faecal flora with abdominal sepsis; *Catonella morbi*, *Centipeda periodontii*, and *Cryptobacterium curtum* with periodontal disease). In such cases it is difficult to distinguish between aetiological agents and colonists of pre-existing disease foci. In addition to well-known zoonotic agents (e.g. *Mycobacterium bovis*, the agent of bovine tuberculosis), increasing numbers of species have been associated with contact with animals or animal products (e.g. *Capnocytophaga canimorsus* from dog bites). Some rarely encountered conditions have been associated with exposure to environmental organisms (e.g. humidifier fever following inhalation of *Parachlamydia acanthamoebae*; actinomycetoma following inoculation injury with *Actinomadura* spp.). Invasive medical procedures and devices and immunosuppressive therapies have allowed relatively non-pathogenic organisms (e.g. *Acinetobacter* spp. and *Staphylococcus epidermidis*) to cause infection, making it harder to distinguish between contaminants and isolates of clinical significance.

Clinicians cannot be expected universally to be familiar with more than a small proportion of the known potentially pathogenic bacteria. Hence, it is necessary to develop strategies that allow assessment of the likely significance of bacterial names encountered in the literature and laboratory reports. Table 1 lists genera in alphabetical order. Within genera, species are grouped according to associated clinical features, alongside which are given reported antimicrobial susceptibilities and treatments, concise notes, and selected references. The names used are those which have standing in nomenclature at the time of writing; that is to say, names that appear in the *Approved lists of bacterial names*(Skerman VBD, McGowan V, Sneath PHA (1989) American Society for Microbiology, Washington DC; amended edition), or the *Index of the bacterial and yeast nomenclature changes* (Moore WEC, Moore LVH (1992) American Society for Microbiology, Washington DC), or have been validated by publication in the *International Journal of Systematic Bacteriology*. Issues up to and including Part 4 of Volume 49 (October 1999) have been consulted. To check a taxon's standing in nomenclature, an extremely useful reference source is the *List of bacterial names with standing in nomenclature* (http://www.bacterio.cict.fr/). The use of correct names allows accurate communication between specialists, but name changes resulting from reclassification (e.g. the splitting of *Pseudomonas* into several genera) or from the correction of Latin (e.g. *Streptococcus sanguinis* instead of *S. sanguis*) may cause confusion. Table 1 includes some recently used synonyms (stated in parentheses) and CDC alphanumeric groups (e.g. CDC group DF-3) which await designation of scientific names. A useful list of such terms may be found in the *Summary of current nomenclature, taxonomy and classification* in *Clinical Infectious Diseases* 1999,**29**, 713–27. Names not validly published are stated in inverted commas (e.g. '*Flexispira rappini*'). An updated version of this chapter is available at (http://homepages.pavilion.co.uk/tetrix/).

The antimicrobial susceptibility and treatment notes are based on a wide range of sources and are no more than a rough guide. In the absence of well-established regimens for particular situations, treatment should take into account susceptibility data from the strain causing the infection and the monitoring of treatment response. Caution should be exercised in interpreting the significance of unusual isolates, especially from normally non-sterile sites.

References

1. Heath CH, *et al.* (1998). Vertebral osteomyelitis and discitis associated with *Abiotrophia adiacens* (nutritionally variant streptococcus) infection. *Australian and New Zealand Journal of Medicine* **28**, 663.

2. Biermann C, *et al.* (1999). Isolation of *Abiotrophia adiacens* from a brain abscess which developed in a patient after neurosurgery. *Journal of Clinical Microbiology* **37**, 769–71.

3. Namdari H, *et al.* (1999). *Abiotrophia* species as a cause of endophthalmitis following cataract extraction. *Journal of Clinical Microbiology* **37**, 1564–6.

4. Yabuuchi E, *et al.* (1998). Emendation of genus *Achromobacter* and *Achromobacter xylosoxidans* (Yabuuchi and Yano) and proposal of *Achromobacter ruhlandii* (Packer and Vishniac) comb. nov., *Achromobacter piechaudii* (Kiredjian *et al*.) comb. nov., and *Achromobacter xylosoxidans* subsp. *denitrificans* (Ruger and Tan) comb. nov. *Microbiology and Immunology* **42**, 429–38.

5. Rolston KVI, Messer M (1990). The *in-vitro* susceptibility of *Alcaligenes denitrificans* subsp. *xylosoxidans* to 40 antimicrobial agents. *Journal of Antimicrobial Chemotherapy* **26**, 857–60.

6. Peel MM, *et al.* (1988). *Alcaligenes piechaudii* from chronic ear discharge. *Journal of Clinical Microbiology* **26**, 1580–1.

7. Igra-Siegman Y, Chmel H, Cobbs C (1980). Clinical and laboratory characteristics of *Achromobacter xylosoxidans* infection. *Journal of Clinical Microbiology* **11**, 141–5.

8. Holmes B, Snell JJS, Lapage SP (1977). Strains of *Achromobacter xylosoxidans* from clinical material. *Journal of Clinical Pathology* **30**, 595–601.

9. Reverdy ME, *et al*. (1984). Nosocomial colonisation and infection by *Achromobacter xylosoxidans*. *Journal of Clinical Microbiology* **19**, 140–3.

10. Sugihara PT, *et al.* (1974). Isolation of *Acidaminococcus fermentans* and *Megasphaera elsdenii* from normal human feces. *Applied Microbiology* **27**, 274–5.

11. Chatterjee BD, Chakraborti CK (1995). Non-sporing anaerobes in certain surgical group of patients. *Journal of the Indian Medical Association* **93**, 333–5, 339.

12. Willems A, *et al.* (1990). *Acidovorax*, a new genus for *Pseudomonas facilis*, *Pseudomonas delafieldii* E Falsen (EF) group 13, EF group 16, and several clinical isolates, with the species *Acidovorax facilis* comb. nov., *Acidovorax delafieldii* comb. nov., and *Acidovorax temperans* sp. nov. *International Journal of Systematic Bacteriology* **40**, 384–98.

13. Rosenthal SL, Freundlich LF (1977). The clinical significance of *Acinetobacter* species. *Health Laboratory Science* **14**, 194–8.

14. French GL, *et al.* (1980). A hospital outbreak of antibiotic-resistant *Acinetobacter anitratus*: epidemiology and control. *Journal of Hospital Infection* **1**, 125–31.

15. Bouvet PJM, Grimont PAD (1986). Taxonomy of the genus *Acinetobacter* with recognition of *Acinetobacter baumanni* sp. nov., *Acinetobacter haemolyticus* sp. nov., *Acinetobacter johnsonii* sp. nov., and *Acinetobacter junii* sp. nov. and emended descriptions of *Acinetobacter calcoaceticus* and *Acinetobacter lwoffii*. *International Journal of Systematic Bacteriology* **36**, 238–40.

16. Haley S, *et al.* (1990). *Acinetobacter* sp. L-form infection of a cemented Charnley total hip replacement. *Journal of Clinical Pathology* **43**, 781.

17. Bergogne-Bérézine E, Joly-Guillou ML (1991). Hospital infection with *Acinetobacter* spp.: an increasing problem. *Journal of Hospital Infection* **18A**, 250–5.

18. Urban C, *et al*. (1993). Effect of sublactam on infections caused by imipenem-resistant *Acinetobacter calcoaceticus* biotype *anitratus*.*Journal of Infectious Diseases* **167**, 448–51.

19. Ellner JJ, *et al.* (1979). Infective endocarditis caused by slow-growing, fastidious, Gram-negative bacteria. *Medicine* (Baltimore)**58**, 145–58.

20. Kristinsson KG, Thorgeirsson G, Holbrook WP (1988). *Actinobacillus actinomycetemcomitans* and endocarditis.*Journal of Infectious Diseases* **157**, 599.

21. Peel MM, *et al.* (1991). *Actinobacillus* spp. and related bacteria in infected wounds of humans bitten by horses and sheep. *Journal of Clinical Microbiology* **29**, 2535–8.

22. Dibb WL, Digranes A, TØnjum S (1981). *Actinobacillus lignieresii* infection after a horse bite. *British Medical Journal* **283**, 583.

23. Wust J, *et al.* (1991). *Actinobacillus hominis* as a causative agent of septicemia in hepatic failure. *European Journal of Clinical Microbiology and Infectious Diseases* **10**, 693–4.

24. Marriott DJ, Brady LM (1983). *Pasteurella ureae* meningitis. *Medical Journal of Australia* **2**, 455–6.

25. Noble RC, Marek BJ, Overman SB (1987). Spontaneous peritonitis caused by *Pasteurella ureae*. *Journal of Clinical Microbiology* **25**, 442–4.

26. Lawson PA, *et al.* (1997). Characterization of some Actinomyces-like isolates from human clinical specimens: reclassification of *Actinomyces suis* (Soltys and Spratling) as *Actinobaculum suis* comb. nov. and description of *Actinobaculum schaalii* sp. nov. *International Journal of Systematic Bacteriology* **47**, 899–903.

27. Venugopal PV, Venugopal TV (1990). *Actinomadura madurae* mycetomas. *Australasian Journal of Dermatology* **31**, 33–6.

28. Smego RA Jr, Foglia G (1998). Actinomycosis. *Clinical Infectious Diseases* **26**, 1255–63.

29. Funke G, *et al.* (1997). *Actinomyces europaeus* sp. nov., isolated from human clinical specimens. *International Journal of Systematic Bacteriology* **47**, 687–92.

30. Colman G (1967). *Aerococcus*-like organisms isolated from human infections. *Journal of Clinical Pathology* **20**, 294–7.

31. Nathavitharana KA, *et al*. (1983). Acute meningitis in early childhood caused by *Aerococcus viridans*. *British Medical Journal* **286**, 1248.

32. Mercer NSG, *et al.* (1987). Medicinal leeches as sources of wound infection. *British Medical Journal* **294**, 937.

33. Gluski I, *et al.* (1992). A 15-year study of the role of *Aeromonas* spp. in gastroenteritis in hospitalised children.*Journal of Medical Microbiology* **37**, 315–18.

34. Joseph SW, *et al.* (1991). *Aeromonas jandaei* and *Aeromonas veronii* dual infection of a human wound following aquatic exposure. *Journal of Clinical Microbiology* **29**, 565–9.

35. Ong KR, Sordillo E, Frankel E (1991). Unusual case of *Aeromonas hydrophila* endocarditis. *Journal of Clinical Microbiology* **29**, 1056–7.

36. Janda JM, Duffey PS (1988). Mesophilic aeromonads in human disease: current taxonomy, laboratory identification, and infectious disease spectrum. *Reviews of Infectious Diseases* **10**, 980–97.

37. Flandry F, *et al.* (1989). Initial antibiotic therapy for alligator bites: characterization of the oral flora of *Alligator mississippiensis*. *Southern Medical Journal* **82**, 262–6.

38. Hickman-Brenner FW, *et al.* (1988). *Aeromonas schubertii*, a new mannitol-negative species found in human clinical specimens. *Journal of Clinical Microbiology* **26**, 1561–4.

39. Hickman-Brenner FW, *et al.* (1987). *Aeromonas veronii*, a new ornithine decarboxylase-positive species that may cause diarrhea. *Journal of Clinical Microbiology* **25**, 900–6.

40. Wolff RL, Wiseman SL, Kitchens CS (1980). *Aeromonas hydrophila* bacteremia in ambulatory immunocompromised hosts. *American Journal of Medicine* **68**, 238–40.

41. Young DF, Barr RJ (1981). *Aeromonas hydrophila* infection of the skin. *Archives of Dermatology* **117**, 244.

42. Champsaur H, *et al.* (1982). Cholera-like illness due to *Aeromonas sobria*. *Journal of Infectious Diseases* **145**, 248–54.

43. Motyl MR, McKinley G, Janda JM (1985). *In vitro* susceptibilities of *Aeromonas hydrophila*, *Aeromonas sobria*, and *Aeromonas caviae* to 22 antimicrobial agents. *Antimicrobial Agents and Chemotherapy* **28**, 151–3.

44. Brenner DJ, *et al*. (1991). Proposal of *Afipia* gen. nov., with *Afipia felis* sp. nov. (formerly the Cat Scratch Disease Bacillus), *Afipia clevelandensis* sp. nov. (formerly the Cleveland Clinic Foundation Strain), *Afipia broomeae* sp. nov., and three unnamed genospecies. *Journal of Clinical Microbiology* **29**, 2450–60.

45. Plotkin GR (1980). *Agrobacterium radiobacter* prosthetic valve endocarditis. *Annals of Internal Medicine* **93**, 839–40.

46. Hammerberg O, Bialowska-Hobrzanska H, Gopaul D (1991). Isolation of *Agrobacterium radiobacter* from a central venous catheter. *European Journal of Clinical Microbiology and Infectious Diseases* **10**, 450.

47. Freney J, *et al.* (1985). Septicemia caused by *Agrobacterium* sp. *Journal of Clinical Microbiology* **22**, 683–5.

48. Castagnola E, *et al.* (1997). Broviac catheter-related bacteraemias due to unusual pathogens in children with cancer: case reports with literature review. *Journal of Infection* **34**, 215–18.

49. Bizet J, Bizet C (1997). Strains of *Alcaligenes faecalis* from clinical material. *Journal of Infection* **35**, 167–9.

50. Hendolin PH, *et al.* (1999). High incidence of *Alloiococcus otitis* in otitis media with effusion. *Pediatric Infectious Disease Journal* **18**, 860–5.

51. Faden H, Dryja D (1989). Recovery of a unique bacterial organism in human middle ear fluid and its possible role in chronic otitis media. *Journal of Clinical Microbiology* **27**, 2488–91.

52. Aguirre M, Collins MD (1992). Development of a polymerase chain reaction-probe test for identification of *Alloiococcus otitis*. *Journal of Clinical Microbiology* **30**, 2177–80.

53. Lechevalier MP, *et al.* (1986). Two new genera of nocardioform actinomycetes: *Amycolata* gen. nov. and *Amycolatopsis* gen. nov. *International Journal of Systematic Bacteriology* **36**, 29–37.

54. Malnick H, *et al.* (1983). *Anaerobiospirillum* species isolated from humans with diarrhoea. *Clinical Pathology* **36**, 1097–101.

55. Goddard WW, Bennett SA, Parkinson C (1998). *Anaerobiospirillum succiniciproducens* septicaemia: important aspects of diagnosis and management. *Journal of Infection* **37**, 68–70.

56. Lee JI, Hampson DJ (1994). Genetic characterisation of intestinal spirochaetes and their association with disease. *Journal of Medical Microbiology* **40**, 365–71.

57. Tee W, *et al.* (1998) Three cases of *Anaerobiospirillum succiniciproducens* bacteremia confirmed by 16S rRNA gene sequencing. *Journal of Clinical Microbiology* **36**, 1209–13.

58. Malnick H (1997). *Anaerobiospirillum thomasii* sp. nov., an anaerobic spiral bacterium isolated from the feces of cats and dogs and from diarrheal feces of humans, and emendation of the genus *Anaerobiospirillum*. *International Journal of Systematic Bacteriology* **47**, 381–4.

59. Malnick H, *et al.* (1990). Description of a medium for isolating *Anaerobiospirillum* spp, a possible cause of zoonotic disease, from diarrheal feces and blood of humans and use of the medium in a survey of human, canine, and feline feces. *Journal of Clinical Microbiology* **28**, 1380–4.

60. Shah HN, Collins MD (1986). Reclassification of *Bacteroides furcosus* Veillon and Zuber (Hauduroy, Ehringer, Urbain, Guillot and Magrou) in a new genus *Anaerorhabdus*, as *Anaerorhabdus furcosus* comb. nov. *Systematic Applied Microbiology* **8**, 86–8.

61. Fell HWK, *et al.* (1977). *Corynebacterium haemolyticum* infections in Cambridgeshire. *Journal of Hygiene, Cambridge* **79**, 269–74.

62. Jobantputra RS, Swain CP (1975). Septicaemia due to *Corynebacterium haemolyticum*. *Journal of Clinical Pathology* **28**, 798–800.

63. Greenman JL (1987). *Corynebacterium hemolyticum* and pharyngitis. *Annals of Internal Medicine* **106**, 633.

64. Funke G, *et al.* (1997). Clinical microbiology of coryneform bacteria. *Clinical Microbiology Reviews* **10**, 125–59.

65. Lepargneur JP, *et al.* (1998). Urinary tract infection due to *Arcanobacterium bernardiae* in a patient with a urinary tract diversion. *European Journal of Clinical Microbiology and Infectious Diseases* **17**, 399–401.

66. Adderson EE, *et al.* (1998). Septic arthritis due to *Arcanobacterium bernardiae* in an immunocompromised patient. *Clinical Infectious Diseases* **27**, 211–12.

67. Ieven, M (1996). Severe infection due to *Actinomyces bernardiae*: case report. *Clinical Infectious Diseases* **22**, 157–8.

68. Na'was TE, *et al.* (1987). Comparison of biochemical, morphologic, and chemical characteristics of Centers for Disease Control fermentative coryneform groups 1, 2, and A-4. *Journal of Clinical Microbiology* **25**, 1354–8.

69. Lynch M, *et al.* (1998). *Actinomyces pyogenes* septic arthritis in a diabetic farmer. *Journal of Infection* **37**, 71–3.

70. Drancourt M, *et al.* (1993). Two cases of *Actinomyces pyogenes* infection in humans. *European Journal of Clinical Microbiology and Infectious Diseases* **12**, 55–7.

71. Vandamme P, *et al*. (1992). Outbreak of recurrent abdominal cramps associated with *Arcobacter butzleri* in an Italian school. *Journal of Clinical Microbiology* **30**, 2335–7.

72. Tee W, *et al.* (1988). *Campylobacter cryaerophila* isolated from a human. *Journal of Clinical Microbiology* **26**, 2469–73.

73. Vandamme P, *et al.* (1991). Revision of *Campylobacter*, *Helicobacter* and *Wolinella* taxonomy: emendation of generic descriptions and proposal of *Arcobacter* gen. nov. *International Journal of Systematic Bacteriology* **41**, 88–103.

74. Bodaghi B, *et al.* (1998). Whipple's syndrome (uveitis, B27-negative spondylarthropathy, meningitis, and lymphadenopathy) associated with *Arthrobacter* sp. infection. *Ophthalmology* **105**, 1891–6.

75. Hou XG, *et al.* (1998). Description of *Arthrobacter creatinolyticus* sp. nov., isolated from human urine. *International Journal of Systematic Bacteriology* **48**, 423–9.

76. Hsu C-L, *et al.* (1998). Septicaemia due to *Arthrobacter* species in a neutropenic patient with acute lymphoblastic leukemia. *Clinical Infectious Diseases* **27**, 1334–5.

77. Collins MD, Wallbanks S (1992). Comparative sequence analyses of the 16S rRNA genes of *Lactobacillus minutus*, *Lactobacillus rimae* and *Streptococcus parvulus*: proposal for the creation of a new genus *Atopobium*. *FEMS Microbiology Letters* **95**, 235–40.

78. Olsen I, *et al.* (1991). *Lactobacillus uli* sp. nov. and *Lactobacillus rimae* sp. nov. from the human gingival crevice and emended descriptions of *Lactobacillus minutus* and *Streptococcus parvulus*. *International Journal of Systematic Bacteriology* **41**, 261–6.

79. Rodriguez Jovita M, *et al.* (1999). Characterization of a novel *Atopobium* isolate from the human vagina: description of *Atopobium vaginae* sp. nov. *International Journal of Systematic Bacteriology* **49**, 1573–6.

80. Meijer-Severs GJ, *et al.* (1979). The presence of antibody-coated anaerobic bacteria in asymptomatic bacteriuria during pregnancy. *Journal of Infectious Diseases* **140**, 653–8.

81. Ihde DC, Armstrong D (1973). Clinical spectrum of infection due to bacillus species. *American Journal of Medicine* **55**, 839–45.

82. Slimans R, Rehm S, Shlaes DM. (1987). Serious infections caused by *Bacillus* species. *Medicine* (Baltimore)**66**, 218–23.

83. Weber DJ, *et al.* (1988). *In vitro* susceptibility of *Bacillus* spp. to selected antimicrobial agents. *Antimicrobial Agents and Chemotherapy* **32**, 642–5.

84. Isaacson P, *et al.* (1976). Pseudotumour of the lung caused by infection with *Bacillus sphaericus*. *Journal of Clinical Pathology* **29**, 806–11.

85. Samples JR, Buettner H (1983). Corneal ulcer caused by a biological insecticide (*Bacillus thuringiensis*). *American Journal of Ophthalmology* **95**, 258–60.

86. Samples JR, Buettner H (1983). Ocular infection caused by a biological insecticide. *Journal of Infectious Diseases* **148**, 614.

87. Reller LB (1973). Endocarditis caused by *Bacillus subtilis*. *American Journal of Clinical Pathology* **60**, 714–18.

88. de Carvalho CB, Moreira JL, Ferreira MC (1996). Epidemiology and antimicrobial resistance of *B. fragilis* group organisms isolated from clinical specimen and human intestinal microbiota. *Revista do Instituto de Medicina Tropical de Sao Paulo* **38**, 329–35.

89. Rasmussen BA, Bush K, Tally FP (1993). Antimicrobial resistance in *Bacteroides*. *Clinical Infectious Diseases* **16(Suppl** 4), 390–400.

90. Dauga C, *et al.* (1993). *Balneatrix alpica* gen. nov., sp. nov. a bacterium associated with pneumonia and meningitis in a spa therapy centre. *Research in Microbiology* **144**, 35–46.

91. Casalta JP, *et al.* (1989). Pneumonia and meningitis caused by a new nonfermentative unknown gram-negative bacterium. *Journal of Clinical Microbiology* **27**, 1446–8.

92. Ellis BA, *et al.* (1999). An outbreak of acute bartonellosis (Oroya fever) in the Urubamba region of Peru, 1998. *American Journal of Tropical Medicine and Hygiene* **61**, 344–9.

93. Daly JS, *et al.* (1993). *Rochalimaea elizabethae* sp. nov. isolated from a patient with endocarditis. *Journal of Clinical Microbiology* **31**, 872–81.

94. Regnery RL, *et al.* (1992). Serological response to '*Rochalimaea henselae*' antigen in suspected cat-scratch disease. *Lancet* **339**, 1443–5.

95. Relman DA, *et al.* (1990). The agent of bacillary angiomatosis: an approach to the identification of uncultured pathogens. *New England Journal of Medicine* **323**, 1573–80.

96. Koeler JE, *et al.* (1992). Isolation of *Rochalimaea* species from cutaneous and osseous lesions of bacillary angiomatosis. *New England Journal of Medicine* **327**, 1625–31.

97. Welch DF, *et al.* (1999). Isolation of a new subspecies, *Bartonella vinsonii* subsp. *arupensis*, from a cattle rancher: identity with isolates found in conjunction with *Borrelia burgdorferi* and *Babesia microti* among naturally infected mice. *Journal of Clinical Microbiology* **37**, 2598–601.

98. Reina J, Borrell N (1992). Leg abscess caused by *Weeksella zoohelcum* following a dog bite. *Clinical Infectious Diseases* **14**, 1162–3.

99. Ha GY, *et al.* (1999). Case of sepsis caused by *Bifidobacterium longum*. *Journal of Clinical Microbiology* **37**, 1227–8.

100. Brook I (1996). Isolation of non-sporing anaerobic rods from infections in children. *Journal of Medical Microbiology* **45**, 21–6.

101. Brook I, Frazier EH (1993). Significant recovery of nonsporulating anaerobic rods from clinical specimens. *Clinical Infectious Diseases* **16**, 476–80.

102. Bourne KA, *et al.* (1978). Bacteremia due to *Bifidobacterium*, *Eubacterium* or *Lactobacillus*; twenty-one cases and review of the literature. *Yale Journal of Biology and Medicine* **51**, 505–12.

103. Kasten MJ, Rosenblatt JE, Gustafson DR (1992). *Bilophila wadsworthia* bacteremia in two patients with hepatic abscesses. *Journal of Clinical Microbiology* **30**, 2502–3.

104. Summanen P, *et al.* (1989). *Bilophila wadsworthia*, gen. nov. and sp. nov., a unique gram-negative anaerobic rod recovered from appendicitis specimens and human faeces. *Journal of General Microbiology* **135**, 3405–11.

105. Wang G, *et al.* (1999). Molecular typing of *Borrelia burgdorferi sensu lato*: taxonomic, epidemiological, and clinical implications. *Clinical Microbiology Reviews* **12**, 633–53.

106. Fukunaga M, *et al.* (1996). Phylogenetic analysis of *Borrelia* species based on flagellin gene sequences and its application for molecular typing of Lyme disease borreliae. *International Journal of Systematic Bacteriology* **46**, 898–905.

107. Dworkin MS, *et al.* (1999). *Bordetella bronchiseptica* infection in human immunodeficiency virus-infected patients. *Clinical Infectious Diseases* **28**, 1095–9.

108. Vandamme P, *et al.* (1996). *Bordetella trematum* sp. nov., isolated from wounds and ear infections in humans, and reassessment of *Alcaligenes denitrificans* Rüger and Tan 1983. *International Journal of Systematic Bacteriology* **46**, 849–58.

109. Vandamme P, *et al.* (1995). *Bordetella hinzii* sp. nov., isolated from poultry and humans. *International Journal of Systematic Bacteriology* **45**, 37–45.

110. Weyant RS, *et al.* (1995). *Bordetella holmesii* sp. nov., a new gram-negative species associated with septicemia. *Journal of Clinical Microbiology* **33**, 1–7.

111. Bergfors E, *et al.* (1999). Parapertussis and pertussis: differences and similarities in incidence, clinical course, and antibody responses. *International Journal of Infectious Diseases* **3**, 140–6.

112. Mikosza AS, *et al.* (1999). PCR amplification from fixed tissue indicates frequent involvement of *Brachyspira aalborgi* in human intestinal spirochetosis. *Journal of Clinical Microbiology* **37**, 2093–8.

113. Lee JI, Hampson DJ (1994). Genetic characterisation of intestinal spirochaetes and their association with disease. *Journal of Medical Microbiology* **40**, 365–71.

114. Stanton TB, *et al.* (1997). Recognition of two new species of intestinal spirochetes: *Serpulina intermedia* sp. nov. and *Serpulina murdochii* sp. nov. *International Journal of Systematic Bacteriology* **47**, 1007–12.

115. Shida O, *et al.* (1996). Proposal for two new genera, *Brevibacillus* gen. nov. and *Aneurinibacillus* gen. nov. *International Journal of Systematic Bacteriology* **46**, 939–46.

116. Wen RR (1984). [A preliminary study of food poisoning by *Bacillus brevis* Migula]. *Chung Hua Yu Fang I Hsueh Tsa Chih* **18**, 168–9. [In Chinese]

117. Yabbara KF, Juffali F, Matossian RM (1977). *Bacillus laterosporus* endophthalmitis. *Archives of Ophthalmology* **95**, 2187–9.

118. Gruner E, *et al.* (1994). Human infections caused by *Brevibacterium casei*, formerly CDC groups B-1 and B-3. *Journal of Clinical Microbiology* **32**, 1511–18.

119. Funke G, Punter V, von Graevenit A (1996). Antimicrobial susceptibility patterns of some recently established coryneform bacteria. *Antimicrobial Agents and Chemotherapy* **40**, 2874–8.

120. Segers P, *et al.* (1994). Classification of *Pseudomonas diminuta* Leifson and Hugh 1954 and *Pseudomonas vesicularis* Büsing, Döll, and Freytag 1953 in *Brevundimonas* gen. nov. as *Brevundimonas diminuta* comb. nov. and *Brevundimonas vesicularis* comb. nov., respectively. *International Journal of Systematic Bacteriology* **44**, 499–510.

121. Lulu AR, *et al.* (1988). Human brucellosis in Kuwait: a prospective study of 400 cases. *Quarterly Journal of Medicine* **66**, 39–54.

122. Gessner AR, Mortensen JE (1990). Pathogenic factors of *Pseudomonas cepacia* isolates from patients with cystic fibrosis. *Journal of Medical Microbiology* **33**, 115–20.

123. Glass S, Govan JRW (1986). *Pseudomonas cepacia*—fatal pulmonary infection in a patient with cystic fibrosis. *Journal of Infection* **13**, 157–8.

124. Miller R, Pannell L, Ingalls MS (1948). Experimental chemotherapy in glanders and melioidosis. *American Journal of Hygiene* **47**, 205–13.

125. Dance DAB (1990). Melioidosis. *Reviews in Medical Microbiology* **1**, 143–50.

126. Dance DAB (1991). Melioidosis: the tip of the iceberg? *Clinical Microbiology Reviews* **4**, 52–60.

127. Dionisio D, *et al.* (1992). Appendicite: interazioni microbiche e nuovi patogeni. *Recenti Progressi in Medicina* **83**, 330–6.

128. Freney J, *et al.* (1988). Susceptibilities to antibiotics and antiseptics of new species of the family *Enterobacteriaceae. Antimicrobial Agents and Chemotherapy* **32**, 873–6.

129. Johnson CC, Finegold SM (1987). Uncommonly encountered, motile, anaerobic gram-negative bacilli associated with infection. *Reviews of Infectious Diseases* **9**, 1150–62.

130. Blazer MJ (1990). *Campylobacter* species. In: Mandell GL, Douglas RG, Bennett JE, eds. *Principles and practice of infectious diseases*, 3rd edn. Churchill Livingstone, New York.

131. Figura N, *et al*. (1993). Two cases of *Campylobacter mucosalis* enteritis in children. *Journal of Clinical Microbiology* **31**, 727–8.

132. Francioli P, *et al.* (1985). *Campylobacter fetus* subspecies *fetus* bacteremia. *Archives of Internal Medicine* **145**, 289–92.

133. Edmonds P, *et al*. (1987). *Campylobacter hyointestinalis* associated with human gastrointestinal disease in the United States. *Journal of Clinical Microbiology* **25**, 685–91.

134. Simon AE, Wilcox L (1987). Enteritis associated with *Campylobacter laridis*. *Journal of Clinical Microbiology* **25**, 10–12.

135. von Graevenitz A (1990). Revised nomenclature of *Campylobacter laridis*, *Enterobacter intermedium*, and '*Flavobacterium branchiophila*.' *International Journal of Systematic Bacteriology* **40**, 211.

136. Walmsley SL, Karmali MA (1989). Direct isolation of atypical thermophilic *Campylobacter* species from human feces on selective agar medium. *Journal of Clinical Microbiology* **27**, 668–70.

137. Gaudreau C, Lamothe F (1992). *Campylobacter upsalensis* isolated from a breast abscess. *Journal of Clinical Microbiology* **30**, 1354–6.

138. Decoster H, Snoeck J, Pattyn S (1992). *Capnocytophaga canimorsus* endocarditis. *European Heart Journal* **13**, 140–2.

139. Brenner DJ, *et al.* (1989). *Capnocytophaga canimorsus* sp. nov. (formerly CDC group DF-2), a cause of septicemia following dog bite, and *C. cynodegmi* sp. nov., a cause of localised wound infection following dog bite. *Journal of Clinical Microbiology* **5**, 231–5.

140. Anderson HK, Pedersen M (1992). Infective endocarditis with involvement of the tricuspid valve due to *Capnocytophaga canimorsus. European Journal of Clinical Microbiology and Infectious Diseases* **11**, 831–2.

141. Bilgrami S, *et al.* (1992). *Capnocytophaga* bacteremia in a patient with Hodgkin's disease following bone marrow transplantation: case report and review. *Clinical Infectious Diseases* **14**, 1045–9.

142. Sundqvist G (1992). Associations between microbial species in dental root canal infections. *Oral Microbiology and Immunology* **7**, 257–62.

143. Savage DD, *et al.* (1977). *Cardiobacterium hominis* endocarditis: description of two patients and characterisation of the organism. *Journal of Clinical Microbiology* **27**, 75–80.

144. Wormser GP, Bottone EJ (1983). *Cardiobacterium hominis*: review of microbiologic and clinical features. *Reviews of Infectious Diseases* **5**, 680–91.

145. Moore LVH, Moore WEC (1994). *Oribaculum catoniae* gen. nov., sp. nov.; *Catonella morbi* gen. nov., sp. nov.; *Hallella seregens* gen. nov., sp. nov., *Johnsonella ignava* gen. nov., sp. nov.; and *Dialister pneumosintes* gen. nov., comb. nov., nom. rev, anaerobic gram-negative bacilli from the human gingival crevice. *International Journal of Systematic Bacteriology* **44**, 187–92.

146. Gill VJ, Travis LB, Williams DY (1991). Clinical and microbiological observations on CDC group DF-3, a Gram-negative coccobacillus. *Journal of Clinical Microbiology* **29**, 1589–92.

147. Blum RN, *et al.* (1992). Clinical illness associated with isolation of dysgonic fermenter 3 from stool samples. *Journal of Clinical Microbiology* **30**, 396–400.

148. Aronson N, Zbick CJ (1988). Dysgonic fermenter 3 bacteremia in a neutropenic patient with acute lymphocytic leukemia. *Journal of Clinical Microbiology* **26**, 2213–15.

149. Bangsborg JM, Frederiksen W, Bruun B (1990). Dysgonic fermenter 3-associated abscess in a diabetic patient. *Journal of Infection* **20**, 237–40.

150. Farmer JJ III, *et al.* (1982). Bacteremia due to *Cedecea neteri* sp. nov. *Journal of Clinical Microbiology* **16**, 775–8.

151. Grimont PAD, *et al.* (1981). *Cedecea davisae* gen. nov., sp. nov. and *Cedecea lapagei* sp. nov., new *Enterobacteriaceae* from clinical specimens. *International Journal of Systematic Bacteriology* **31**, 317–26.

152. Funke G, Ramos CP, Collins MD (1995). Identification of some clinical strains of CDC coryneform group A-3 and A-4 bacteria as *Cellulomonas* species and proposal of *Cellulomonas hominis* sp. nov. for some group A-3 strains. *Journal of Clinical Microbiology* **33**, 2091–7.

153. Le Prowse C, McNeil MM, McCarty JM (1989). Catheter-related bacteremia caused by *Oerskovia turbata*. *Journal of Clinical Microbiology* **27**, 571–2.

154. Reller LB, *et al.* (1975). Bacterial endocarditis caused by *Oeskovia turbata*. *Annals of Internal Medicine* **83**, 664–6.

155. Lai CH, *et al.* (1983). *Centipeda periodontii* gen. nov., sp. nov., from human periodontal lesions. *International Journal of Systematic Bacteriology* **33**, 628–35.

156. Everett KDE, Bush RM, Andersen AA (1999). Emended description of the order *Chlamydiales*, proposal of *Parachlamydiaceae* fam. nov. and *Simkaniaceae* fam. nov., each containing one monotypic genus, revised taxonomy of the family *Chlamydiaceae*, including a new genus and five new species, and standards for the identification of organisms. *International Journal of Systematic Bacteriology* **49**, 415–40.

157. Tucker RE, Winter WG, Wilson HD (1979). Osteomyelitis associated with *Chromobacterium violaceum* sepsis: a case report. *Journal of Bone and Joint Surgery* **61**, 949–51.

158. Feldman RB (1984). *Chromobacterium violaceum* infection of the eye: a report of two cases. *Archives of Ophthalmology* **102**, 711–13.

159. Sorensen RU, Jacobns MR, Shurin SB (1985). *Chromobacterium violaceum* adenitis acquired in the northern United States as a complication of chronic granulomatous disease. *Pediatric Infectious Disease* **4**, 701–2.

160. Thong ML, Puthucheary SD, Lee EL (1981). *Flavobacterium meningosepticum* infection: an epidemiological study in a newborn nursery. *Journal of Clinical Pathology* **34**, 429–33.

161. Hsueh PR, *et al.* (1997). Increasing incidence of nosocomial *Chryseobacterium indologenes* infections in Taiwan. *European Journal of Clinical Microbiology and Infectious Diseases* **16**, 568–74.

162. Fraser SL, Jorgensen JH (1997). Reappraisal of the antimicrobial susceptibilities of *Chryseobacterium* and *Flavobacterium* species and methods for reliable susceptibility testing. *Antimicrobial Agents and Chemotherapy* **41**, 2738–41.

163. Doran TI (1999). The role of *Citrobacter* in clinical disease of children: review. *Clinical Infectious Diseases* **28**, 384–94.

164. Bittner J (1980). The clinical significance, taxonomy and special methodological problems of the pathogenic clostridia. *Infection* **8(Suppl**2), 117–22.

165. Kageyama A, Benno Y, Nakase T (1999). Phylogenetic and phenotypic evidence for the transfer of *Eubacterium aerofaciens* to the genus *Collinsella* as *Collinsella aerofaciens* gen. nov., comb. nov. *International Journal of Systematic Bacteriology* **49**, 557–65.

166. Atkinson BE, Smith DL, Lockwood WR (1975). *Pseudomonas testosteroni* septicemia. *Annals of Internal Medicine* **83**, 369–70.

167. Tamaoka J, Ha D-M, Komagata K (1987). Reclassification of *Pseudomonas acidovorans* den Dooren de Jong 1926 and *Pseudomonas testosteroni* Marcus and Talalay 1956 as *Comamonas acidovorans* comb. nov. and *Comamonas testosteroni* comb. nov., with an emended description of the genus *Comamonas*. *International Journal of Systematic Bacteriology* **37**, 52–9.

168. Lipsky BA, *et al.* (1982). Infections caused by non-diphtheria corynebacteria. *Reviews of Infectious Diseases* **4**, 1220–35.

169. Funke G, Lawson PA, Collins MD (1995). Heterogeneity within human-derived centers for disease control and prevention (CDC) coryneform group ANF-1-like bacteria and description of *Corynebacterium auris* sp. nov. *International Journal of Systematic Bacteriology* **45**, 735–9.

170. Vale JA, Scott GW (1977). *Corynebacterium bovis* as a cause of human disease. *Lancet* **2**, 682–4.

171. Philippon A, Bimet F (1990). *In vitro* susceptibility of *Corynebacterium* Group D2 and *Corynebacterium jeikeium* to twelve antibiotics. *European Journal of Clinical Microbiology and Infectious Diseases*.**9**, 892–5.

172. Gill VL, *et al.* (1981). Antibiotic-resistant group JK bacteria in hospitals. *Journal of Clinical Microbiology* **13**, 472–7.

173. Quinn JP, *et al.* (1984). Outbreak of JK diphtheroid infections associated with environmental contamination. *Journal of Clinical Microbiology* **19**, 668–71.

174. Messina OD, *et al.* (1989). *Corynebacterium kutscheri* septic arthritis. *Arthritis and Rheumatism* **32**, 1053.

175. Wilhelmus KR, Robinson NM, Jones DB (1979). *Bacterionema matruchotii* ocular infections. *American Journal of Ophthalmology* **87**, 143–7.

176. Funke G, Punter V, von Graevenitz A (1996). Antimicrobial susceptibility patterns of some recently established coryneform bacteria. *Antimicrobial Agents and Chemotherapy* **40**, 2874–8.

177. Barr JG, Murphy PG (1986). *Corynebacterium striatum*: an unusual organism isolated in pure culture from sputum. *Journal of Infection* **13**, 297–8.

178. Chomarat M, Breton P, Dubost J (1991). Osteomyelitis due to *Corynebacterium* group D2. *European Journal of Clinical Microbiology and Infectious Diseases* **10**, 43.

179. Soriano F, Ponte C (1992). A case of urinary tract infection caused by *Corynebacterium urealyticum* and coryneform group F1. *European Journal of Clinical Microbiology and Infectious Diseases* **11**, 626–8.

180. Soriano F, Fernandez-Roblas R (1988). Infections caused by antibiotic-resistant *Corynebacterium* group D2. *European Journal of Clinical Microbiology and Infectious Diseases* **7**, 337–41.

181. Porschen RK, Goodman Z, Rafai B (1977). Isolation of *Coryebacterium xerosis* from clinical specimens. *American Journal of Clinical Pathology* **68**, 290–3.

182. Liakim R, *et al.* (1983). *Corynebacterium xerosis* endocarditis. *Archives of Internal Medicine* **143**, 1995.

183. Krish G, *et al.* (1989). *Corynebacterium xerosis* as cause of vertebral osteomyelitis. *Journal of Clinical Microbiology* **27**, 2869–70.

184. Guillard F, Appelbaum PC, Sparrow FB (1980). Pyelonephritis and septicemia due to Gram-positive rods similar to *Corynebacterium* Group E (aerotolerant *Bifidobacterium adolescentis*). *Annals of Internal Medicine* **92**, 635–6.

185. Austin GE, Hill EO (1983). Endocarditis due to *Corynebacterium* CDC group G2.*Journal of Infectious Diseases* **147**, 1106.

186. Malanoski GJ, Parker R, Eliopoulos GM (1992). Antimicrobial susceptibilities of a *Corynebacterium* CDC group I1 strain isolated from a patient with endocarditis. *Southern Medical Journal* **80**, 923.

187. Tendler C, Bottone EJ (1989). *Corynebacterium aquaticum* urinary tract infection in a neonate and concepts regarding the role of the organism as a neonatal pathogen. *Journal of Clinical Microbiology* **27**, 343–5.

188. Golledge CL, Phillips G (1991). *Corynebacterium minutissimum* infection. *Journal of Infection* **23**, 73–6.

189. Colt HG, *et al. (*1991). Necrotizing tracheitis caused by *Corynebacterium pseudodiphtheriticum*: unique case and review. *Reviews of Infectious Diseases* **13**, 73–6.

190. Goldberger AC, Lipsky BA, Plorde JJ (1981). Suppurative granulomatous lymphadenitis caused by *Corynebacterium ovis (pseudotuberculosis)*. *American Journal of Clinical Pathology* **76**, 486–90.

191. Meers PD (1979). A case of classical diphtheria, and other infections due to *Corynebacterium ulcerans*. *Journal of Infection* **1**, 139–42.

192. Maurin M, Raoult D (1999). Q fever. *Clinical Microbiology Reviews* **12**, 518–53.

193. Nakazawa F, *et al.* (1999). *Cryptobacterium curtum* gen nov., sp. nov., a new genus of Gram-positive anaerobic rod isolatated from human oral cavities. *International Journal of Systematic Bacteriology* **49**, 1193–200.

194. Horowitz H, *et al.* (1990). Endocarditis associated with *Comamonas acidovorans*. *Journal of Clinical Microbiology* **28**, 143–5.

195. Bavbek M, *et al. (*1998). Cerebral *Dermabacter hominis* abscess. *Infection* **26**, 181–3.

196. Pal M (1995). Prevalence in India of *Dermatophilus congolensis* infection in clinical specimens from animals and humans. *Revue Scientifique et Technique* **14**, 857–63.

197. Gibson GR, Macfarlane GT, Cummings JH (1988). Occurrence of sulphate-reducing bacteria in human faeces and the relationship of dissimilatory sulphate reduction to methanogenesis in the large gut. *Journal of Applied Bacteriology* **65**, 103–11.

198. McDougall R, *et al. (*1997). Bacteremia caused by a recently described novel *Desulfovibrio* species. *Journal of Clinical Microbiology* **35**, 1805–8.

199. Tee W, *et al. (*1996). Probable new species of *Desulfovibrio* isolated from a pyogenic liver abscess. *Journal of Clinical Microbiology* **34**, 1760–4.

200. Porschen RK, Chan P (1977). Anaerobic vibrio-like organisms cultured from blood: *Desulfovibrio desulfuricans* and *Succinivibrio* species. *Journal of Clinical Microbiology* **5**, 444–7.

201. Willems A, Collins MD (1995). Phylogenetic placement of *Dialister pneumosintes* (formerly *Bacteroides pneumosintes*) within the *Sporomusa* subbranch of the *Clostridium* subphylum of the gram-positive bacteria. *International Journal of Systematic Bacteriology* **45**, 403–5.

202. Liu D, Yong WK (1997). Improved laboratory diagnosis of ovine footrot: an update. *The Veterinary Journal* **153**, 99–105.

203. Collins MD, *et al. (*1999). *Dolosicoccus paucivorans* gen. nov., sp. nov., isolated from human blood. *International Journal of Systematic Bacteriology* **49**, 1439–42.

204. Aguirre M, *et al. (*1993). Phenotypic and phylogenetic characterization of some *Gemella*-like organisms from human infections: description of *Dolosigranulum pigrum* gen. nov., sp. nov. *Journal of Applied Bacteriology* **75**, 608–12.

205. Miller PH, Facklam RR, Miller JM (1996). Atmospheric growth requirements for *Alloiococcus* species and related gram-positive cocci. *Journal of Clinical Microbiology* **34**, 1027–8.

206. Maskell R, Pead L (1990). A cluster of *Edwardsiella tarda* infection in a day-care center in Florida. *Journal of Infectious Diseases* **162**, 282.

207. Hargreaves JE, Lucey DR (1990). Life-threatening *Edwardsiella tarda* soft tissue infection associated with catfish puncture wound. *Journal of Infectious Diseases* **162**, 1416.

208. Janda JM, *et al.* (1991). Pathogenic properties of *Edwardsiella* species. *Journal of Clinical Microbiology* **29**, 1997–2001.

209. Murphey DK, Septimus EJ, Waagner DC (1990). Catfish-related injury and infection: report of two cases and review of the literature. *Clinical Infectious Diseases* **14**, 689–93.

210. Reger PJ, Mockler DF, Miller MA (1993). Comparison of antimicrobial susceptibility, beta-lactamase production, plasmid analysis and serum bactericidal activity in *Edwardsiella tarda* E *ictaluri* and *E. hoshinae*. *Journal of Medical Microbiology* **39**, 273–81.

211. Kageyama A, Benno Y, Nakase T (1999). Phylogenetic evidence for the transfer of *Eubacterium lentum* to the genus *Eggerthella* as *Eggerthella lenta* gen. nov., comb. nov. *International Journal of Systematic Bacteriology* **49**, 1725–32.

212. McDade JE (1990). Ehrlichiosis—disease of animals and humans. *Journal of Infectious Diseases* **161**, 609–17.

213. Anderson BE, *et al. (*1991). *Ehrlichia chaffeensis*, a new species associated with human ehrlichiosis. *Journal of Clinical Microbiology* **29**, 2838–42.

214. Dupon M, *et al. (*1991). Sacro-iliac joint infection caused by *Eikenella corrodens*. *European Journal of Clinical Microbiology and Infectious Diseases* **10**, 529–30.

215. Stoloff AL, Gillies ML (1986). Infections with *Eikenella corrodens* in a general hospital: a report of 33 cases. *Reviews of Infectious Diseases* **8**, 50–3.

216. Suwangol S, *et al. (*1983). Pathogenicity of *Eikenella corrodens* in humans. *Archives of Internal Medicine* **143**, 2265–8.

217. Pérez-Pomata MT, *et al. (*1992). Spleen abscess caused by *Eikenella corrodens*. *European Journal of Clinical Microbiology and Infectious Diseases* **11**, 162–3.

218. Dees SB, *et al. (*1986). Chemical characterization of *Flavobacterium odoratum*, *Flavobacterium breve*, and *Flavobacterium*-like groups IIe, IIh, and IIf. *Journal of Clinical Microbiology* **23**, 267–73.

219. Sanders WE Jr, Sanders CC (1997). *Enterobacter* spp.: pathogens poised to flourish at the turn of the century. *Clinical Microbiology Reviews* **10**, 220–41.

220. Morrison D, Woodford N, Cookson B (1997). Enterococci as emerging pathogens of humans. *Society for Applied Bacteriology Symposium Series* **26**, S89–99.

221. O'Hara CM, *et al. (*1998). First report of a human isolate of *Erwinia persicinus*. *Journal of Clinical Microbiology* **36**, 248–50.

222. MacGowan AP, Reeves DS (1991). Tricuspid valve infective endocarditis and pulmonary sepsis due to *Erysipelothrix rhusiopathiae* successfully treated with high doses of ciproflaxacin but complicated by gynaecomastia. *Journal of Infection* **22**, 100–1.

223. Venditti M, *et al. (*1990). Antimicrobial susceptibilities of *Erysipelothrix rhusiopathiae*. *Antimicrobial Agents and Chemotherapy* **34**, 2038–40.

224. Brook MG, Bannister BA (1993). Diarrhoea-causing *Escherichia coli. Digestive Diseases* **11**, 288–97.

225. Farmer JJ III, *et al. (*1985). *Escherichia fergusonii* and *Enterobacter taylorae*, two new species of *Enterobacteriaceae* isolated from clinical specimens. *Journal of Clinical Microbiology* **21**, 77–81.

226. Brenner DJ, *et al.* (1982). Atypical biogroups of *Escherichia coli* found in clinical specimens and description of *Escherichia hermanii* sp. nov. *Journal of Clinical Microbiology* **15**, 703–13.

227. Brenner DJ, *et al. (*1982). *Escherichia vulneris*: a new species of Enterobacteriaceae associated with human wounds. *Journal of Clinical Microbiology* **15**, 1133–40.

228. Sans MD, Crowder JG (1973). Subacute bacterial endocarditis caused by *Eubacterium aerofaciens*: report of a case. *American Journal of Clinical Pathology* **59**, 576–80.

229. Tew JG, *et al. (*1985). Serum antibody reactive with predominant organisms in the subgingival flora of young adults with generalized severe periodontitis. *Infection and Immunity* **49**, 487–93.

230. Devreese K, Claeys G, Verschraegen G (1992). Septicaemia with *Ewingella americana*. *Journal of Clinical Microbiology* **30**, 2746–7.

231. Collins MD, *et al. (*1997). Phenotypic and phylogenetic characterization of some *Globicatella*-like organisms from human sources: description of *Facklamia hominis* gen. nov., sp. nov. *International Journal of Systematic Bacteriology* **47**, 880–2.

232. Collins MD, *et al. (*1994). The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *International Journal of Systematic Bacteriology* **44**, 812–26.

233. Jalava J, Eerola E (1999). Phylogenetic analysis of *Fusobacterium alocis* and *Fusobacterium sulci* based on 16S rRNA gene sequences: proposal of *Filifactor alocis* (Cato, Moore and Morre) comb. nov. and *Eubacterium sulci* (Cato, Moore and Moore) comb. nov. *International Journal of Systematic Bacteriology* **49**, 1375–9.

234. Weir S, *et al. (*1999). Recurrent bacteremia caused by a '*Flexispira*'-like organism in a patient with X-linked (Bruton's) agammaglobulinemia. *Journal of Clinical Microbiology* **37**, 2439–45.

235. Sorlin P, *et al. (*1999). Recurrent '*Flexispira rappini*' bacteremia in an adult patient undergoing hemodialysis: case report. *Journal of Clinical Microbiology* **37**, 1319–23.

236. Hollis DG, *et al. (*1989). *Francisella philomiragia* comb. nov (formerly *Yersinia philomiragia*) and *Francisella tularensis* biogroup novicida (formerly *Francisella novicida*) associated with human disease. *Journal of Clinical Microbiology* **27**, 1601–8.

237. Evans ME, *et al. (*1985). Tularemia: a 30-year experience with 88 cases. *Medicine* (Baltimore)**64**, 251–69.

238. Moore-Gillon J, *et al. (*1984). Necrobacillosis: a forgotten disease. *British Medical Journal* **288**, 1526–7.

239. George WL, Kirby BD, Sutter VL (1981). Gram-negative anaerobic bacilli: their role in infection and patterns of susceptibility to antibiotic agents. II Little-known *Fusobacterium* species with miscellaneous genera. *Reviews of Infectious Diseases* **3**, 599–626.

240. Hillier SL (1993). Diagnostic microbiology of bacterial vaginosis. *American Journal of Obstetrics and Gynecology* **169**, 455–9.

241. Chatelain R, *et al. (*1982). Isolement de *Gemella haemolysans* dans trois cas d'endocardites bacteriennes. *Médecine et Maladies Infectieuses* **12**, 25–30.

242. Kilpper-Bälz R, Schleifer KH (1988). Transfer of *Streptococcus morbillorum* to the *Gemella* genus, *Gemella morbillorum* comb. nov. *International Journal of Systematic Bacteriology* **38**, 442–3.

243. Collins MD, *et al. (*1992). *Globicatella sanguis* gen. nov., sp. nov., a new gram-positive catalase-negative bacterium from human sources. *Journal of Applied Bacteriology* **73**, 433–7.

244. Drancourt M, *et al. (*1994). Brain abscess due to *Gordona terrae* in an immunocompromised child: case report and review of infections caused by *G. terrae*. *Clinical Infectious Diseases* **19**, 258–62.

245. Drancourt M, *et al. (*1997). *Gordona terrae* central nervous system infection in an immunocompetent patient. *Journal of Clinical Microbiology* **35**, 379–82.

246. Richet HM, *et al. (*1991). A cluster of *Rhodococcus (Gordona) Bronchialis* sternal-wound infections after coronary-artery bypass surgery. *New England Journal of Medicine* **10**, 104–9.

247. Riegel P, *et al. (*1996). Bacteremia due to *Gordona sputi* in an immunocompromised patient. *Journal of Clinical Microbiology* **34**, 2045–7.

248. Brenner DJ, *et al.* (1988). Biochemical, genetic, and epidemiologic characterization of *Haemophilus influenzae* biogroup aegyptius (*Haemophilus aegyptius*) strains associated with Brazilian purpuric fever. *Journal of Clinical Microbiology* **26**, 1524–34.

249. Brazilian Purpuric Fever Study Group (1987). *Haemophilus aegyptius* bacteremia in Brazilian purpuric fever. *Lancet* **2**, 761–3.

250. Bieger RC, Brewer NS, Washington JA II (1978). *Haemphilus aphrophilus*: a microbiologic and clinical review and report of 42 cases. *Medicine* (Baltimore)**57**, 345–55.

251. Goldberg R, Washington JA II (1978). The taxonomy and antimicrobial susceptibility of *Haemophilus* species in clinical specimens.*American Journal of Clinical Pathology* 70, 899–904.

252. Julander I, Lindberg AA, Swanbom M (1980). *Haemophilus parainfluenzae*: an uncommon cause of septicemia and endocarditis. *Scandinavian Journal of Infectious Diseases* **12**, 85–9.

253. Jones RN, Slepack J, Bigelow J (1976). Ampicillin-resistant *Haemophilis paraphrophilus* laryngo-epiglottitis. *Journal of Clinical Microbiology* **4**, 405–7.

254. Visvanathan K, Jones PD (1991). Ciprofloxacin treatment of *Haemophilus paraphrophilus* brain abscess. *Journal of Infection* **22**, 306–7.

255. Schmid GP (1999). Treatment of chancroid, 1997. *Clinical Infectious Diseases* **28(Suppl** 1), 14–20.

256. Jordens JZ, Slack MP (1995). *Haemophilus influenzae*: then and now. *European Journal of Clinical Microbiology and Infectious Diseases* **14**, 935–48.

257. Washington JA III, Birk RJ, Ritts RE (1971). Bacteriologic and epidemiologic characteristics of *Enterobacter hafniae* and *Enterobacter liquefaciens*. *Journal of Infectious Diseases* **124**, 379.

258. Stanley J, *et al. (*1993). *Helicobacter canis* sp. nov., a new species from dogs: an integrated study of phenotype and genotype. *Journal of General Microbiology* **139**, 2495–504.

259. Orlicek SL, Welch DF, Kuhls TL (1993). Septicemia caused by *Helicobacter cinaedi* in a neonate. *Journal of Clinical Microbiology* **31**, 569–71.

260. Vandamme P, *et al. (*1990). Identification of *Campylobacter cinaedi* isolated from blood and faeces of children and adult females. *Journal of Clinical Microbiology* **28**, 1016–20.

261. Totten PA, *et al.* (1985). *Campylobacter cinaedi* (sp. nov.) and *Campylobacter fennelliae* (sp. nov.): two new campylobacter species associated with enteric disease in homosexual men. *Journal of Infectious Diseases* **151**, 131–9.

262. Meining A, Kroher G, Stolte M (1998). Animal reservoirs in the transmission of *Helicobacter heilmannii*. Results of a questionnaire-based study. *Scandinavian Journal of Gastroenterology* **33**, 795–8.

263. Stanley J, *et al. (*1994). *Helicobacter pullorum* sp. nov.—genotype and phenotype of a new species isolated from poultry and from human patients with gastroenteritis. *Microbiology* **140**, 3441–9.

264. Marshall BJ (1986). *Campylobacter pyloridis* and gastritis. *Journal of Infectious Diseases* **153**, 650–7.

265. Chagla AH, *et al. (*1998). Breast abscess associated with *Helcococcus kunzii. Journal of Clinical Microbiology* **36**, 2377–9.

266. Peel MM, *et al. (*1997). *Helcococcus kunzii* as sole isolate from an infected sebaceous cyst. *Journal of Clinical Microbiology* **35**, 328–9.

267. Willems A, *et al. (*1997). Phenotypic and phylogenetic characterization of some *Eubacterium*-like isolates containing a novel type B wall murein from human feces: description of *Holdemania filiformis* gen. nov., sp. nov. *International Journal of Systematic Bacteriology* **47**, 1201–4.

268. Collins MD, *et al. (*1999). *Ignavigranum ruoffiae* sp. nov., isolated from human clinical specimens. *International Journal of Systematic Bacteriology* **49**, 97–101.

268a. Moore LV, Moore WE (1994). *Orbiculum catoniae* gen. nov., sp. nov.; *Catonella morbi* gen. nov., sp. nov.; *Hallella seregens* gen. nov., sp. nov.; *Johnsonella ignava* gen. nov., sp. nov.; and *Dialister pneumosintes* gen. nov., comb. nov., nom. rev.; anaerobic Gram-negative bacteria from the human gingival crevice. *International Journal of Systematic Bacteriology* **44**, 187–92.

269. Yagupsky P, *et al.* (1992). High prevalence of *Kingella kingae* in joint fluid from children with septic arthritis revealed by the BACTEC blood culture system. *Journal of Clinical Microbiology* **30**, 1278–81.

270. Goldman IS, *et al. (*1980). Infective endocarditis due to *Kingella denitrificans*. *Annals of Internal Medicine* **93**, 152–3.

271. Jenny DB, Letendre PW, Iverson G (1988). Endocarditis due to *Kingella* species. *Reviews of Infectious Diseases* **10**, 1065–6.

272. Namnyak SS, Quinn RJM, Ferguson JDM (1991). *Kingella kingae* meningitis in an infant. *Journal of Infection* **23**, 104–6.

273. Carter JS, *et al. (*1999). Phylogenetic evidence for reclassification of *Calymmatobacterium granulomatis* as *Klebsiella granulomatis* comb. nov. *International Journal of Systematic Bacteriology* **49**, 1695–700.

274. Chetoui H, *et al. (*1999). Epidemiological typing of extended-spectrum beta-lactamase-producing *Klebsiella pneumoniae* isolates by pulsed-field gel electrophoresis and antibiotic susceptibility patterns. *Research in Microbiology* **150**, 265–72.

275. Miller RH, *et al. (*1979). *Klebsiella rhinoscleromatis*: a clinical and pathogenic enigma. *Otolaryngology—Head and Neck Surgery* **87**, 212–21.

276. Farmer JJ III, *et al.* (1981). *Kluyvera*, a new (redefined) genus in the family *Enterobacteriaceae*: identification of *Kluyvera ascorbata* sp. nov. and *Kluyvera cryocrescens* sp. nov. in clinical specimens. *Journal of Clinical Microbiology* **13**, 919–33.

277. Sierra-Madero J, *et al. (*1990). *Kluyvera* mediastinitis following open-heart surgery: a case report. *Journal of Clinical Microbiology* **28**, 2848–9.

278. Stackenbrandt E, *et al. (*1995). Taxonomic dissection of the genus *Micrococcus*: *Kocuria* gen. nov., *Nesterenkonia* gen. nov., *Kytococcus* gen. nov., *Dermacoccus* gen. nov., and *Micrococcus* Cohn 1872 gen. emend. *International Journal of Systematic Bacteriology* **45**, 682–92.

279. Hickman-Brenner FW, *et al. (*1985). *Koserella trabulsii*, a new genus and species of *Enterobacteriaceae* formerly known as enteric group 45. *Journal of Clinical Microbiology* **21**, 39–42.

280. Elston HR (1961). *Kurthia bessonii* isolated from clinical material. *Journal of Pathology and Bacteriology* **81**, 245–7.

281. Pancoast SJ, *et al.* (1979). Endocarditis due to *Kurthia bessonii*. *Annals of Internal Medicine* **90**, 936–7.

282. Keddie RM, Shaw S (1986). Genus *Kurthia* Trevisan 1885, 92^{AL} Nom. cons. Opin. 13 Jud. Comm. 1954, 152. In: Sneath PHA, *et al.*, eds. *Bergey's manual of systematic bacteriology*, Vol. 2. Williams and Willkins, Baltimore, MD.

283. Chomarat M, Espinouse D (1991). *Lactobacillus rhamosus* septicemia in patients with prolonged aplasia receiving ceftazidime–vancomycin. *European Journal of Clinical Microbiology and Infectious Diseases* **10**, 44.

284. Rahman M (1982). Chest infection caused by *Lactobacillus casei* ss *rhamosus*. *British Medical Journal* **284**, 471–2.

285. Sussman JI, *et al. (*1986). Clinical manifestation and therapy of *Lactobacillus* endocarditis: report of a case and review of the literature. *Reviews of Infectious Diseases* **8**, 771–6.

286. Bantar CE, *et al. (*1991). Abscess caused by vancomycin-resistant *Lactobacillus confusus*. *Journal of Clinical Microbiology* **29**, 2063–4.

287. Elliott JA, *et al. (*1991). Differentiation of *Lactococcus lactis* and *Lactococcus garviae* from humans by comparison of whole-cell protein patterns. *Journal of Clinical Microbiology* **29**, 2731–4.

288. Rossmann SN, *et al. (*1998). Isolation of *Lautropia mirabilis* from oral cavities of human immunodeficiency virus-infected children. *Journal of Clinical Microbiology* **36**, 1756–60.

289. Tamura K, *et al. (*1986). *Leclercia adecatrboxylata* gen nov., comb. nov., formerly known as *Escherichia adecarboxylata*. *Current Microbiology* **13**, 179–82.

290. Benson RF, Fields BS (1998). Classification of the genus *Legionella*. *Seminars in Respiratory Infections* **13**, 90–9.

291. Hickman-Brenner F, *et al. (*1985). *Leminorella*, a new genus of Enterobacteriaceae: identification of *Leminorella grimontii* sp. nov. and *Leminorella richardii* sp. nov. found in clinical specimens. *Journal of Clinical Microbiology* **21**, 234–9.

292. Lecour H, *et al. (*1989). Human leptospirosis: a review of 50 cases. *Infection* **17**, 8–12.

293. Vemelen K, *et al. (*1996). Bacteraemia with *Leptotrichia buccalis*: report of a case and review of the literature. *Acta Clinica Belgica* **51**, 265–70.

294. Friedland IR, Snipelisky M, Khoosal M (1990). Meningitis in a neonate caused by *Leuconostoc* sp. *Journal of Clinical Microbiology* **28**, 2125–6.

295. Bernaldo de Quirós JCL, *et al. (*1991). *Leuconostoc* species as a cause of bacteremia: two case reports and a literature review. *European Journal of Clinical Microbiology and Infectious Diseases* **10**, 505–9.

296. Horowitz HW, Handwerger S, van Horn KG (1987). *Leuconostoc*, an emerging vancomycin-resistant pathogen. *Lancet* **1**, 1329–30.

297. Hof H, Nichterlein T, Kretschmar M (1997). Management of listeriosis. *Clinical Microbiology Reviews* **10**, 345–57.

298. Brancaccio M, Legendri GG (1979). *Megasphaera eldenii* endocarditis. *Journal of Clinical Microbiology* **10**, 72–4.

299. Kaye KM, Macone A, Kazanjian PH (1992). Catheter infection caused by *Methylobacterium* in immunocompromised hosts: report of three cases and review of the literature. *Clinical Infectious Diseases* **14**, 1010–14.

300. Gould FK, Venning MC, Ford M (1990). Successful treatment with chloramphenicol of *Pseudomonas mesophilica* peritonitis not resonding to aztreonam and gentamicin. *Journal of Antimicrobial*

Chemotherapy **26**, 458–9.

301. Rutherford PC, *et al. (*1988). Peritonitis caused by *Pseudomonas mesophilica* in a patient undergoing continuous ambulatory peritoneal dialysis. *Journal of Clinical Microbiology* **26**, 2441–3.

302. Smith SM, Eng RHK, Forrester C (1985). *Pseudomonas mesophilica* infections in humans. *Journal of Clinical Microbiology* **21**, 314–17.

303. Funke G, *et al. (*1997). Endophthalmitis due to *Microbacterium* species: case report and review of microbacterium infections. *Clinical Infectious Diseases* **24**, 713–16.

304. Funke G, *et al.* (1998). *Aureobacterium resistens* sp. nov., exhibiting vancomycin resistance and teicoplanin susceptibility. *FEMS Microbiology Letters* **158**, 89–93.

305. Saweljew P, *et al. (*1996). Case of fatal systemic infection with an *Aureobacterium* sp.: identification of isolate by 16S rRNA gene analysis. *Journal of Clinical Microbiology* **34**, 1540–1.

306. Nolte FS, *et al. (*1996). Vancomycin-resistant *Aureobacterium* species cellulitis and bacteremia in a patient with acute myelogenous leukemia. *Journal of Clinical Microbiology* **34**, 1992–4.

307. Hagiwara S, *et al. (*1995). [Hypersensitivity pneumonitis caused by a home humidifier]. *Nihon Kyobu Shikkan Gakkai Zasshi* **33**, 1024–9. [In Japanese]

308. Peces R, *et al.* (1997). Relapsing bacteraemia due to *Micrococcus luteus* in a haemodialysis patient with a Perm-Cath catheter. *Nephrology, Dialysis, Transplantation* **12**, 2428–9.

309. Shah HN, Collins MD (1982). Reclassification of *Bacteroides multiacidus* (Mitsuoka, Terada, Watanabe and Uchida) in a new genus *Mitsuokella*, as *Mitsuokella multiacidus* comb. nov. *Zentralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankhe und Hygiene. Abstract 1, Orig*.**C3**, 491–4.

310. Schwebke JR, *et al.* (1991). Identification of two new antigenic subgroups within the genus *Mobiluncus*. *Journal of Clinical Microbiology* **29**, 2204–8.

311. Glupczynski T, *et al* (1984). Isolation of *Mobiluncus* in four cases of extragenital infection in adult women. *European Journal of Clinical Microbiology* **3**, 433–5.

312. Spiegel CA (1987). Susceptibility of *Mobiluncus* species to 23 antimicrobial agents and 15 other compounds. *Antimicrobial Agents and Chemotherapy* **31**, 249–52.

313. Hickman-Brenner FW, *et al. (*1984). *Moellerella wisconsensis*, a new genus and species of *Enterobacteriaceae* found in human stool specimens. *Journal of Clinical Microbiology* **19**, 460–3.

314. Silverfarb PM, Lawe JE (1968). Endocarditis due to *Moraxella liquefaciens*. *Archives of Internal Medicine* **122**, 512–13.

315. Ebright JR, Lentino JR, Juni E (1982). Endophthalmitis caused by *Moraxella nonliquefaciens*. *American Journal of Clinical Pathology* **77**, 362–3.

316. BØvre K, Henriksen SD (1967). A new *Moraxella* species, *Moraxella osloensis*, and a revised description of *Moraxella nonliquefaciens*. *International Journal of Systematic Bacteriology* **17**, 127–35.

317. BØvre K, Fuglesang JE, Hagen N (1976). *Moraxella atlantae* sp. nov. and its distinction from *Moraxella phenylpyruvica*. *International Journal of Systematic Bacteriology* **26**, 511–21.

318. Percival A, *et al.* (1977). Pathogenicity of and beta-lactamase production by *Branhamella (Neisseria) catarrhalis*. *Lancet* **2**, 1175.

319. Salen PN, Eppes S (1997). *Morganella morganii*: a newly reported, rare cause of neonatal sepsis. *Academic Emergency Medicine* **4**, 711–14.

320. Falkinham JO 3rd (1996). Epidemiology of infection by nontuberculous mycobacteria. *Clinical Microbiology Reviews* **9**, 177–215.

321. Taylor-Robinson D, Bebear C (1997). Antibiotic susceptibilities of mycoplasmas and treatment of mycoplasmal infections. *Journal of Antimicrobial Chemotherapy* **40**, 622–30.

322. Vancanneyt M, *et al. (*1996). Reclassification of *Flavobacterium odoratum* (Stutzer 1929) strains to a new genus, *Myroides*, as *Myroides odoratus* comb. nov. and *Myroides odoratimimus* sp. nov. *International Journal of Systematic Bacteriology* **46**, 926–32.

323. Guidbourdenche M, Lambert T, Riou JY (1989). Isolation of *Neisseria canis* in mixed culture from a patient after a cat bite. *Journal of Clinical Microbiology* **27**, 1673–4.

324. Herbert DA, Ruskin J (1981). Are the 'non-pathogenic' neisseriae pathogenic? *American Journal of Clinical Pathology* **75**, 739–43.

325. Morla N, Guibourdenche M, Riou J-Y (1992). *Neiseria* spp. and AIDS. *Journal of Clinical Microbiology* **30**, 2290–4.

326. Wong JD, Janda JM (1992). *Neisseria* species, *Neisseria elongata* subsp. *nitroreductens*, with bacteremia, endocarditis, and osteomyelitis. *Journal of Clinical Microbiology* **30**, 719–20.

327. Berger SA, *et al.* (1988). Bartholin's gland abscess caused by *Neisseria sicca*. *Journal of Clinical Microbiology* **26**, 1589.

328. Gay RM, Sevier RE (1978). *Neisseria sicca* endocarditis: report of a case and review of the literature. *Journal of Clinical Microbiology* **8**, 729–32.

329. Lind I (1997). Antimicrobial resistance in *Neisseria gonorrhoeae*. *Clinical Infectious Diseases* **24(Suppl**1), 93–7.

330. Oppenheim BA (1997). Antibiotic resistance in *Neisseria meningitidis*. *Clinical Infectious Diseases* **24(Suppl**1), 98–101.

331. Boiron P, *et al. (*1998). *Nocardia*, nocardiosis and mycetoma. *Medical Mycology* **36(Suppl**1), 26–37.

332. Yassin AF, *et al. (*1997). Description of *Nocardiopsis synnemataformans* sp. nov., elevation of *Nocardiopsis alba* subsp. *prasina* to *Nocardiopsis prasina* comb. nov., and designation of *Nocardiopsis antarctica* and *Nocardiopsis alborubida* as later subjective synonyms of *Nocardiopsis dassonvillei*. *International Journal of Systematic Bacteriology* **47**, 983–8.

333. Holmes B, *et al.* (1988). *Ochrobactrum anthropi* gen. nov., sp. nov. from human clinical specimens and previously known as group Vd. *International Journal of Systematic Bacteriology* **38**, 406–16.

334. Moller LVM, *et al. (*1999). *Ochrobactrum intermedium* infection after liver transplantation. *Journal of Clinical Microbiology* **37**, 241–4.

335. Rihs JD, *et al.* (1990). *Oerskovia xanthineolytica* implicated in peritonitis associated with peritoneal dialysis: case report and review of *Oerskovia* infections in humans. *Journal of Clinical Microbiology* **28**, 1934–7.

336. Cruikshank SJ, Gawler AH, Shaldon G (1979). *Oerskovia* species: rare opportunistic pathogens. *Journal of Medical Microbiology* **12**, 513–15.

337. Truant AL, *et al. (*1992). *Oerskovia xanthinolytica* and methicillin-resistant *Staphylococcus aureus* in a patient with cirrhosis and variceal hemorrhage. *European Journal of Clinical Microbiology and Infectious Diseases* **11**, 950–1.

338. Kailath EJ, Goldstein E, Wagner FH (1988). Case report: meningitis caused by *Oeskovia xanthinolytica*. *American Journal of Medical Sciences* **295**, 216–17.

339. Hussain Z, *et al. (*1987). Endophthalmitis due to *Oeskovia xanthinolytica*. *Canadian Journal of Ophthalmology* **22**, 234–6.

340. Mesnard R, *et al. (*1992). Septic arthritis due to *Oligella urethralis*. *European Journal of Clinical Microbiology and Infectious Diseases* **11**, 195–6.

341. Rossau R, *et al. (*1987). *Oligella*, a new genus including *Oligella urethralis* comb. nov (formerly *Moraxella urethralis*) and *Oligella ureolytica* sp. nov (formerly CDC group IVe): relationship to *Taylorella equigenitalis* and related taxa. *International Journal of Systematic Bacteriology* **37**, 198–210.

342. Silpapojakul K (1997). Scrub typhus in the Western Pacific region. *Annals of the Academy of Medicine Singapore* **26**, 794–800.

343. Coudron PE, Payne JM, Markowitz SM (1991). Pneumonia and empyema infection associated with a *Bacillus* species that resembles *B. alvei*. *Journal of Clinical Microbiology* **29**, 1777–9.

344. Olenginski TP, Bush DC, Harrington TM (1991). Plant thorn synovitis: an uncommon cause of monoarthritis. *Seminars in Arthritis and Rheumatism* **21**, 40–6.

345. Sneath PHA, Stevens M (1990). *Actinobacillus rossii* sp. nov., *Actinobacillus seminis* sp. nov., nom. rev, *Pasteurella bettii* sp. nov., *Pasteurella lymphangitidis* sp. nov., *Pasteurella mairi* sp. nov., and *Pasteuerella trehalosi* sp. nov. *International Journal of Systematic Bacteriology* **40**, 148–53.

346. Johnson RH, Rumans LW (1977). Unusual infections caused by *Pasteurella multocida*. *Journal of the American Medical Association* **237**, 146–47.

347. Rogers BT, *et al. (*1973). Septicaemia due to *Pasteurella pneumotropica*. *Journal of Clinical Pathology* **26**, 396–8.

348. Pouëdras P, *et al.* (1993). *Pasteurella stomatis* infection following dog bite. *European Journal of Clinical Microbiology and Infectious Diseases* **12**, 65.

349. Yaneza AL, *et al.* (1991). *Pasteurella haemolytica* endocarditis. *Journal of Infection* **23**, 65–7.

350. Holst E, *et al.* (1992). Characterization and distribution of *Pasteurella* species recovered from infected humans. *Journal of Clinical Microbiology* **30**, 2984–7.

351. Mastro TD, *et al.* (1990). Vancomycin-resistant *Pediococcus acidilactici*: nine cases of bacteremia. *Journal of Infectious Diseases* **161**, 956–60.

352. Sire JM, *et al.* (1992). Septicaemia and hepatic abscess caused by *Pediococcus acidilactici*. *European Journal of Clinical Microbiology and Infectious Diseases* **11**, 623–5.

353. Sarma PS, Mohanty S (1998). *Pediococcus acidilactici* pneumonitis and bacteremia in a pregnant woman. *Journal of Clinical Microbiology* **36**, 2392–3.

354. Colman G, Efstratiou A (1987). Vancomycin-resistant leuconostocs, lactobacilli and now pediococci. *Journal of Hospital Infection* **2**, 1–3.

355. Petrini B, Welin-Berger T, Nord CE (1979). Anaerobic bacteria in late infections following orthopedic surgery. *Medical Microbiology and Immunology* (Berlin)**167**, 155–9.

356. Sklavounos A, *et al.* (1986). Anaerobic bacteria in dentoalveolar abscesses. *International Journal of Oral and Maxillofacial Surgery* **15**, 288–91.

357. Murdoch DA, Mitchelmore IJ, Tabaqchali S (1994). The clinical importance of gram-positive anaerobic cocci isolated at St. Bartholomew's Hospital, London, in 1987. *Journal of Medical Microbiology* **41**, 36–44.

358. Murdoch DA, *et al.* (1997). Description of three new species of the genus *Peptostreptococcus* from human clinical specimens: *Peptostreptococcus harei* sp. nov., *Peptostreptococcus ivorii*, sp. nov., and *Peptostreptococcus octavius* sp. nov. *International Journal of Systematic Bacteriology* **47**, 781–7.

359. Pelz K, Mutters R (1997). Taxonomic update and clinical significance of species within the genus *Peptostreptococcus*. *Clinical Infectious Diseases* **25**(Suppl), 94–7.

360. Coffey JA, *et al.* (1986). *Vibrio damsela*: another potentially virulent marine vibrio. *Journal of Infectious Diseases* **153**, 800–2.

361. Farmer JJ III, *et al* (1989). *Xenorhabdus luminescens* (DNA hybridization group 5) from human clinical specimens. *Journal of Clinical Microbiology* **27**, 1594–600.

362. Clark RB, *et al.* (1990). *In vitro* susceptibilities of *Plesiomonas shigelloides* to 24 antibiotics and antibiotic-b-lactamase-inhibitor combinations. *Antimicrobial Agents Chemotherapy* **34**, 159–60.

363. Brenden RA, Miller MA, Janda JM (1988). Clinical disease spectrum and pathogenic factors associated with *Plesiomonas shigelloides* in humans. *Clinical Infectious Diseases* **10**, 303–16.

364. Shah HN, Collins MD (1988). Proposal for reclassification of *Bacteroides asaccharolyticus*, *Bacteroides gingivalis*, and *Bacteroides endodontalis* in a new genus, *Porphyromonas*. *International Journal of Systematic Bacteriology* **38**, 128–31.

365. Shah HN, Collins DM (1990). *Prevotella*, a new genus to include *Bacteroides melaninogenicus* and related species formerly classified in the genus *Bacterioides*. *International Journal of Systematic Bacteriology* **40**, 205–8.

366. Flynn MJ, Li G, Slots J (1994). *Mitsuokella dentalis* in human periodontitis. *Oral Microbiology and Immunology* **9**, 248–50.

367. Riley TV, Ott AK (1981). Brain abscess due to *Arachnia propionica*. *British Medical Journal* **i**, 1035.

368. Brock DW, *et al.* (1973). Actinomycosis caused by *Arachnia propionica*. *American Journal of Clinical Pathology* **59**, 66–77.

369. Mobley HL, Belas R (1995). Swarming and pathogenicity of *Proteus mirabilis* in the urinary tract. *Trends in Microbiology* **3**, 280–4.

370. Hawkey PM (1984). *Providencia stuartii*: a review of a multiply antibiotic-resistant bacterium. *Journal of Antimicrobial Chemotherapy* **13**, 209–26.

371. Pallerono NJ (1984). Family 1. *Pseudomonadaceae*. In: Krieg NR, Holt JG, eds. *Bergey's manual of systematic bacteriology*, Vol. 1. Williams and Willkins, Baltimore, MD.

372. Woese CR (1987). Bacterial evolution. *Microbiological Reviews* **51**, 221–71.

373. Holmes B, *et al.* (1987). *Chryseomonas luteola* comb. nov. and *Flavimonas oryzihabitans* gen. nov. comb. nov. *Pseudomonas*-like species from human clinical specimens and formerly known respectively as groups Ve-1 and Ve-2. *International Journal of Systematic Bacteriology* **37**, 245–50.

374. Podbielski A, *et al.* (1990). *Flavimonas oryzihabitans* septicaemia in a T-cell leukaemic child: a case report and review of the literature. *Journal of Infection* **20**, 135–41.

375. Bendig JWA, *et al.* (1989). *Flavimonas oryzihabitans* (*Pseudomonas oryzihabitans*; CDC group Ve-2): an emerging pathogen in peritonitis related to continuous ambulatory peritoneal dialysis? *Journal of Clinical Microbiology* **27**, 217–18.

376. Levett PN, Garrett DA, Wickramasuriya T (1991). *Flavimonas oryzihabitans* as a cause of ocular infection. *European Journal of Clinical Microbiology and Infectious Diseases* **10**, 594–5.

377. Warwick S, *et al.* (1994). A phylogenetic analysis of the family I>Pseudonocardiaceae and the genera *Actinokineospora* and *Saccharothrix* with 16S rRNA sequences and a proposal to combine the genera*Amycolata* and *Pseudonocardia* in an emended genus *Pseudonocardia*.*International Journal of Systematic Bacteriology* **44**, 293–9.

378. Willems A, Collins MD (1996). Phylogenetic relationships of the genera *Acetobacterium* and *Eubacterium* sensu stricto and reclassification of *Eubacterium alactolyticum* as *Pseudoramibacter alactolyticus* gen. nov., comb. nov. *International Journal of Systematic Bacteriology* **46**, 1083–7.

379. Lloyd-Puryear M, *et al.* (1991). Meningitis caused by *Psychrobacter immobolis* in an infant. *Journal of Clinical Microbiology* **29**, 2041–2.

380. Gini GA (1990). Ocular infection caused by *Psychrobacter immobilis* acquired in a hospital. *Journal of Clinical Microbiology* **28**, 400–1.

381. Alballaa SR, *et al.* (1992). Urinary tract infection due to *Rahnella aquatilis* in a renal transplant patient. *Journal of Clinical Microbiology* **30**, 2948–50.

382. Goubau P, *et al.* (1988). Septicaemia caused by *Rahnella aquatilis* in an immunocompromised patient. *European Journal of Clinical Microbiology and Infectious Diseases* **7**, 697–9.

383. Lacey S, Want SV (1991). *Pseudomonas pickettii* infections in a paediatric oncology unit. *Journal of Hospital Infection* **17**, 45–51.

384. Fujita S, Yoshida T, Matsubara F (1981). *Pseudomonas pickettii* bacteremia. *Journal of Clinical Microbiology* **13**, 781–2.

385. Zapardiel J, *et al.* (1991). Peritonitis with CDC group IVc-2 bacteria in a patient on continuous ambulatory peritoneal dialysis. *European Journal of Clinical Microbiology and Infectious Diseases* **10**, 509–11.

386. Dan M, *et al.* (1986). Septicaemia caused by the Gram-negative bacteria CDC IVc-2 in an immunocompromised human. *Journal of Clinical Microbiology* **23**, 803.

387. Crowe HM, Brecher SM (1987). Septicaemia with CDC group IVc-2, an unusual gram-negative bacillus. *Journal of Clinical Microbiology* **25**, 2225–6.

388. Sane DC, Durack DT (1986). Infection with *Rhodococcus equi* in AIDS. *New England Journal of Medicine* **314**, 56–7.

389. Berg XX, *et al.* (1977). *Corynebacterium equi* infection complicating neoplastic disease. *American Journal of Clinical Pathology* **68**, 73–7.

390. Van Etta LL (1983). *Corynebacterium equi*: a review of twelve cases of human infection. *Reviews of Infectious Diseases* **5**, 1012–18.

391. Xu W, Raoult D (1998). Taxonomic relationships among spotted fever group rickettsiae as revealed by antigenic analysis with monoclonal antibodies. *Journal of Clinical Microbiology* **36**, 887–96.

392. Struthers M, Wong J, Janda JM (1996). An initial appraisal of the clinical significance of *Roseomonas* species associated with human infections. *Clinical Infectious Diseases* **23**, 729–33.

393. Broeren SA, Peel MM (1984). Endocarditis caused by *Rothia dentocariosa*. *Journal of Clinical Pathology* **37**, 1298–300.

394. Willems A, Collins MD (1995). Phylogenetic analysis of *Ruminococcus flavefaciens*, the type species of the genus *Ruminococcus*, does not support the reclassification of *Streptococcus hansenii* and *Peptostreptococcus productus* as ruminococci. *International Journal of Systematic Bacteriology* **45**, 572–5.

395. Nakatani S, *et al.* (1998). [A case report of epidural abscess due to anaerobic bacteria, producing a mass of gas]. *Rinsho Shinkeigaku* **38**, 224–7. [In Japanese]

396. Botha SJ, *et al. (*1993). Anaerobic bacteria in orofacial abscesses. *Journal of the Dental Association of South Africa* **48**, 445–9.

397. Threlfall J, Ward L, Old D (1999). Changing the nomenclature of salmonella. *Communicable Disease and Public Health* **2**, 156–7.

398. Bisiaux-Salauze B, *et al.* (1990). Bacteremias caused by *Selenomonas artemidis* and *Selenomonas infelix*. *Journal of Clinical Microbiology* **28**, 140–2.

399. Westh H, *et al.* (1991). Fatal septicaemia with *Selenomonas sputigena* and *Acinetobacter calcoaceticus*. A case report. *Acta Pathologica, Microbiologica et Immunologica Scandinavica (APMIS)* **99**, 75–7.

400. Pomeroy C, Shanholtzer CJ, Peterson LR (1987). *Selenomonas* bacteraemia: case report and review of the literature. *Journal of Infection* **15**, 237–42.

401. Yu VL (1979). *Serratia marcescens*: historical perspective and clinical review. *New England Journal of Medicine* **300**, 887–92.

402. Pfyffer GE (1991). *Serratia fonticola* as an infectious agent. *European Journal of Clinical Microbiology and Infectious Diseases* **11**, 199–200.

403. Zbinden R, Blass R (1988). *Serratia plymuthica* osteomyelitis following a motorcycle accident. *Journal of Clinical Microbiology* **26**, 1409–10.

404. Clark RB, Janda JM (1985). Isolation of *Serratia plymuthica* from a human burn site. *Journal of Clinical Microbiology* **21**, 656–7.

405. Horowitz HW, *et al. (*1987). *Serratia plymuthica* sepsis associated with infection of central venous catheter. *Journal of Clinical Microbiology* **25**, 1562–3.

406. Marne C, Pallarés R, Sitges-Sera A (1983). Isolation of *Pseudomonas putrefaciens* in intraabdominal sepsis. *Journal of Clinical Microbiology* **17**, 1173–4.

407. Laudat P, *et al. (*1983). *Pseudomonas putrefaciens* meningitis. *Journal of Infection* **7**, 281–3.

408. Kotloff KL, *et al. (*1999). Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. *Bulletin of the World Health Organization* **77**, 651–66.

409. Kahane S, *et al. (*1999). *Simkania negevensis* strain ZT: growth, antigenic and genome characteristics. *International Journal of Systematic Bacteriology* **49**, 815–20.

410. Wade WG, *et al. (*1999). The family *Coriobacteriaceae*: reclassification of *Eubacterium exiguum* (Poco *et al*. 1996) and *Peptostreptococcus heliotrinreducens* (Lanigan 1976) as *Slackia exigua* gen. nov., comb. nov. and *Slackia heliotrinireducens* gen. nov., comb. nov., and *Eubacterium lentum* (Prevot 1938) as *Eggerthella lenta* gen. nov., comb. nov. *International Journal of Systematic Bacteriology* **49**, 595–600.

411. Reina J, Borrell N, Figuerola J (1992). *Sphingobacterium multivorum* isolated from a patient with cystic fibrosis. *European Journal of Clinical Microbiology and Infectious Diseases* **11**, 81–2.

412. Holmes B, *et al. (*1983). *Flavobacterium thalpophilum*, a new species recovered from human clinical material. *International Journal of Systematic Bacteriology* **33**, 677–82.

413. Freney J, *et al. (*1987). Septicemia caused by *Sphingobacterium multivorum. Journal of Clinical Microbiology* **25**, 1126–8.

414. Southern PM, Kutscher AE (1981). *Pseudomonas paucimobilis* bacteremia. *Journal of Clinical Microbiology* **13**, 1070–3.

415. Bhatt KM, Mirza NB (1992). Rat bite fever: a case report of a Kenyan. *East African Medical Journal* **69**, 542–3.

416. Kloos WE, Bannerman TL (1994). Update on clinical significance of coagulase-negative staphylococci. *Clinical Microbiology Reviews* **7**, 117–40.

417. Zuravleff JJ, Yu VL (1982). Infections caused by *Pseudomonas maltophilia* with emphasis on bacteremia: case reports and a review of the literature. *Reviews of Infectious Diseases* **4**, 1236–46.

418. Drancourt M, Bollet C, Raoult D (1997). *Stenotrophomonas africana* sp. nov., an opportunistic human pathogen in Africa. *International Journal of Systematic Bacteriology* **47**, 160–3.

419. von Eiff C, Peters G (1998). *In vitro* activity of ciprofloxacin, ofloxacin, and levofloxacin against *Micrococcus* species and *Stomatococcus mucilaginosus* isolated from healthy subjects and neutropenic patients. *European Journal of Clinical Microbiology and Infectious Diseases* **17**, 890–2.

420. Condron PE, *et al. (*1987). Isolation of *Stomatococcus mucilaginosus* from drug user with endocarditis. *Journal of Clinical Microbiology* **25**, 1359–63.

421. Gruson D, *et al. (*1998). Severe infection caused by *Stomatococcus mucilaginosus* in a neutropenic patient: case report and review of the literature. *Hematology and Cell Therapy* **40**, 167–9.

422. Park MK, *et al.* (1997). Successful treatment of *Stomatococcus mucilaginosus* meningitis with intravenous vancomycin and intravenous ceftriaxone. *Clinical Infectious Diseases* **24**, 278.

423. Hagelskjaer L, Sorensen I, Randers E (1998). *Streptobacillus moniliformis* infection: 2 cases and a literature review. *Scandinavian Journal of Infectious Diseases* **30**, 309–11.

424. Akaike T, *et al. (*1988). *Streptococcus acidominimus* infections in a human. *Japanese Journal of Medicine* **27**, 317–20.

425. Schugk J, *et al. (*1997). A clinical study of beta-haemolytic groups A B, C and G streptococcal bacteremia in adults over an 8-year period. *Scandinavian Journal of Infectious Diseases* **29**, 233–8.

426. Weinstein MR, *et al.*, and *S. iniae* Study Group (1997). Invasive infections due to a fish pathogen, *Streptococcus iniae*. *New England Journal of Medicine* **337**, 589–94.

427. Bert F, *et al. (*1998). Clinical significance of bacteremia involving the ' *Streptococcus milleri*' group: 51 cases and review. *Clinical Infectious Diseases* **27**, 385–7.

428. Elliott PM, Williams H, Brooksby IA (1993). A case of infective endocarditis in a farmer caused by *Streptococcus equinus*. *European Heart Journal* **14**, 1292–3.

429. De Gheldre Y, *et al. (*1999). Identification of clinically relevant viridans streptococci by analysis of transfer DNA intergenic spacer length polymorphism. *International Journal of Systematic Bacteriology* **49**, 1591–8.

430. Crook DW, Spratt BG (1998). Multiple antibiotic resistance in *Streptococcus pneumoniae*. *British Medical Bulletin* **54**, 595–610.

431. Kohler W, *et al. (*1989). *Streptococcus suis* Typ 2 (R-Streptokokken) als Erreger von Berufskrankheiten. Bericht über eine Erkrankung und Literaturübersicht. *Zeitschrift fur die Gesamte Innere Medizin* **44**, 144–8.

432. Nasher MA, *et al. (*1989). *In vitro* studies of antibiotic sensitivities of *Streptomyces somaliensis*: a cause of human actinomycetoma. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **83**, 265–8.

433. Molitoris E, Wexler HM, Finegold SM (1997). Sources and antimicrobial susceptibilities of *Campylobacter gracilis* and *Sutterella wadsworthensis*. *Clinical Infectious Diseases* **25**(Suppl 2), 264–5.

434. Finegold SM, Jousimies-Somer H (1997). Recently described clinically important anaerobic bacteria: medical aspects. *Clinical Infectious Diseases* **25**(Suppl 2), 88–93.

435. Wexler HM, *et al. (*1996). *Sutterella wadsworthensis* gen. nov., sp. nov., bile-resistant microaerophilic *Campylobacter gracilis*-like clinical isolates. *International Journal of Systematic Bacteriology* **46**, 252–8.

436. Jenny DB, Letendre PW, Iverson G (1987). Endocarditis caused by *Kingella indologenes*. *Reviews of Infectious Diseases* **9**, 787.

437. Hollis DG, *et al. (*1981). *Tatumella ptyseos* gen. nov., sp. nov., a member of the family *Enterobacteriaceae* found in clinical specimens. *Journal of Clinical Microbiology* **14**, 79–88.

438. Farrow JA, *et al. (*1995). Phylogenetic evidence that the gram-negative nonsporulating bacterium *Tissierella (Bacteroides) praeacuta* is a member of the *Clostridium subphylum* of the gram-positive bacteria and description of *Tissierella creatinini* sp. nov. *International Journal of Systematic Bacteriology* **45**, 436–40.

439. McWhorter AC, *et al. (*1991). *Trabulsiella guamensis*, a new genus and species of the family *Enterobacteriaceae* that resembles *Salmonella* subgroups 4 and 5. *Journal of Clinical Microbiology* **29**, 1480–5.

440. Willis SG, *et al. (*1999). Identification of seven *Treponema* species in health- and disease-associated dental plaque by nested PCR. *Journal of Clinical Microbiology* **37**, 867–9.

441. Wyss C, *et al. (*1999). *Treponema lecithinolyticum* sp. nov., a small saccharolytic spirochaete with phospholipase A and C activities associated with periodontal diseases. *International Journal of*

Systematic Bacteriology **49**, 1329–39.

442. Koff AB, Rosen T (1993). Nonvenereal treponematoses: yaws, endemic syphilis, and pinta. *Journal of the American Academy of Dermatology* **29**, 519–38.

443. Wallace AL, Harris A, Allen JP (1967). Reiter treponeme. A review of the literature. *Bulletin of the World Health Organization* **36**(Suppl), 1–103.

444. Singh AE, Romanowski B (1999). Syphilis: review with emphasis on clinical, epidemiologic, and some biologic features. *Clinical Microbiology Reviews* **12**, 187–209.

445. Relman DA, *et al.* (1992). Identification of the uncultured bacillus of Whipple's disease. *New England Journal of Medicine* **327**, 293–301.

446. Granel F, *et al. (*1996). Cutaneous infection caused by *Tsukamurella paurometabolum. Clinical Infectious Diseases* **23**, 839–40.

447. Yassin AF, *et al. (*1996). *Tsukamurella pulmonis* sp. nov. *International Journal of Systematic Bacteriology* **46**, 429–36.

448. Chong Y, *et al. (*1997). *Tsukamurella inchonensis* bacteremia in a patient who ingested hydrochloric acid. *Clinical Infectious Diseases* **24**, 1267–8.

449. Shapiro CL, *et al. (*1992). *Tsukamurella paurometabolum*: a novel pathogen causing catheter-related bacteremia in patients with cancer. *Clinical Infectious Diseases* **14**, 200–3.

450. Funke G, *et al. (*1994). *Turicella otitidis* gen. nov., sp. nov., a coryneform bacterium isolated from patients with otitis media. *International Journal of Systematic Bacteriology* **44**, 270–3.

451. Hudson MM, Talbot MD (1997). *Ureaplasma urealyticum*. *International Journal of STD and AIDS* **8**, 546–51.

452. Teixeira LM, *et al. (*1997). Phenotypic and genotypic characterization of *Vagococcus fluvialis*, including strains isolated from human sources. *Journal of Clinical Microbiology* **35**, 2778–81.

453. Rogosa M (1984). Family I. *Veillonellaceae* Rogosa 1971, 232. In: Krieg NR, Holt JG, eds. *Bergey's manual of systematic bacteriology*, Vol. 1. Williams and Willkins, Baltimore, MD.

454. West PA (1989). The human pathogenic vibrios—a public health update with environmental perspectives. *Epidemiology and Infection* **103**, 1–34.

455. Pavia AT, *et al.* (1989). *Vibrio carchariae* infection after a shark bite. *Annals of Internal Medicine* **111**, 85–6.

456. Bode RB, *et al. (*1986). *Vibrio cincinnatiensis* causing meningitis: successful treatment in an adult. *Annals of Internal Medicine* **104**, 55–6.

457. Hickman-Brenner FW, *et al. (*1982). Identification of *Vibrio hollisae* sp. nov. from patients with diarrhea. *Journal of Clinical Microbiology* **15**, 395–400.

458. Jean-Jacques W, *et al. (*1981). *Vibrio metschnikovii* bacteremia in a patient with cholecystitis. *Journal of Clinical Microbiology* **14**, 711–12.

459. Bonner JR, *et al. (*1983). Spectrum of *Vibrio* infections in a Gulf coast community. *Annals of Internal Medicine* **99**, 464–9.

460. Levine WC, Griffin PM, Gulf Coast *Vibrio* Working Group (1993). *Vibrio* infections on the Gulf Coast: results of first year of regional surveillance. *Journal of Infectious Diseases* **167**, 479–83.

461. Boixeda D, *et al. (*1998). A case of spontaneous peritonitis caused by *Weeksella virosa. European Journal of Gastroenterology and Hepatology* **10**, 897–8.

462. Faber MD, *et al.* (1991). Response of *Weeksella virosa* peritonitis to imipenem/cilastin. *Advances in Peritoneal Dialysis* **7**, 133–4.

463. Li ZX, *et al. (*1990). First isolation of *Xanthomonas campestris* from the blood of a Chinese woman. *Chinese Medical Journal* **103**, 435–9.

464. Bercovier H, Mollaret HH (1984). Genus XIV *Yersinia* Van Loghem 1944, 15. In: Krieg NR, Holt JG, eds. *Bergey's manual of systematic bacteriology*, Vol. 1. Williams and Willkins, Baltimore, MD.

465. Kosako Y, Sakazaki R, Yoshizaki E (1984). *Yokenella regensburgei* gen. nov., sp. nov.: a new genus and species in the family *Enterobacteriaceae*. *Japanese Journal of Medical Science and Biology* **37**, 117–24.

7.12.1 Fungal infections

*R. J. Hay**

Introduction

Fungi are saprophytic or parasitic organisms that are normally assigned to a distinct Kingdom. As eukaryotes, they have the complex subcellular organization and highly organized genetic material seen in both animal and plant cells. The cell wall is a distinctive feature of fungi and has a complex skeleton based on mannan and glucan subunits. The arrangement and reproduction of individual cells is also characteristic. Most fungi form new cells terminally, which remain connected to form long, branching filaments or hyphae (the mould fungi). Some reproduce in a similar manner but each new cell separates from the parent by a process of budding (the yeast fungi). It is a feature of certain fungi to be yeast-like during one phase of their life history but hyphal at another, a phenomenon known as dimorphism. In culture, mould fungi usually form a cottony rowth on laboratory media while yeasts normally have a smooth, shiny appearance.

Fungi adversely affect humans in a number of ways. They cause disease indirectly by spoilage and destruction of food crops with subsequent malnutrition and starvation. Many of the common moulds produce and release spores, which may act as airborne allergens to produce asthma or hypersensitivity pneumonitis. Fungi elaborate complex metabolic by-products, some of which are useful to humans, such as the penicillins. However, others are toxic. Disease caused by the ingestion of fungal toxins includes both poisoning by eating certain mushrooms (mycetism) and damage caused by the ingestion of minute quantities of toxin (mycotoxicosis), for instance in contaminated grain. The contribution of the latter mechanism to human disease remains largely unexplored and, in addition, whether inhalation of toxic fungal spores may cause pathology. Finally, fungi may invade human tissue. Medical mycology is largely concerned with this last group. Invasive fungal diseases are normally divided into three groups: the superficial, subcutaneous, and deep mycoses. In superficial infections, such as ringworm or thrush, fungi are confined to the skin and mucous membranes. Extension deeper than the surface epithelium is rare. Subcutaneous infections are usually tropical: the main site of involvement is within subcutaneous tissue, although secondary invasion of adjacent structures such as bone or skin may occur. In deep or systemic infections, deep organs such as the lung, spleen, or brain are invaded. This classification of mycoses is based on the main 'sphere of involvement' by the causal organisms, but there are exceptions. For instance, brain involvement has been recorded in patients with chromoblastomycosis, which is normally a subcutaneous infection.

The fungi causing systemic mycoses are often classified in two groups: the opportunists and the endemic pathogens. The former cause disease in overtly compromised individuals. These contrast with the true pathogens, which cause infection in all subjects inhaling airborne spores.

---

*Dr M.A.H. Bayles prepared the chapter on Chromoblastomycosis for the third edition of this textbook. Much of her text has been included in this chapter and we acknowledge her contribution with grateful thanks.

Further reading—general

Ajello L, Hay RJ, eds (1997). *Mycology*. *Topley and Wilson's microbiology and microbial infections*, 9th edn, Vol 4. Arnold, London.

Kibbler CC, MacKenzie DWR, Odds FC (1996). *Principles and practice of clinical mycology*. John Wiley & Sons, Chichester.

Midgley G, Clayton YM, Hay RJ (1997). *Diagnosis in colour*. *Medical mycology*. Mosby-Wolfe, London.

Warnock DW, Richardson MD, ed. (1990). *Fungal infection in the compromised patient*. Wiley, Chichester.

Superficial fungal infections

The main superficial mycoses are the dermatophyte infections, superficial candidosis, and tinea versicolor (see Section 23). These are both common and widespread. Rare superficial infections include tinea nigra, and black or white piedra.

Dermatophyte infections (dermatophytoses)

Aetiology

The dermatophyte or ringworm infections are caused by a group of organisms capable of existing in keratinized tissue such as stratum corneum, nail, or hair. The mechanism of invasion is thought to be linked to production of extracellular enzymes, such as the three distinct keratinases produced by *Trichophyton mentagrophytes*, but other proteases may also be involved.

Epidemiology

Some dermatophyte fungi have a worldwide distribution; others are more restricted. The most common and most widely distributed is *Trichophyton rubrum*, which causes different types of infection in different parts of the world. It is commonly associated with athlete's foot (tinea pedis) in temperate areas as well as tinea corporis or tinea cruris in the tropics. This distinction is not based solely on climatic factors, as immigrants from tropical countries, particularly the Far East, may still have tinea corporis caused by *T. rubrum* when living in northern Europe. Certain dermatophytes are limited to defined areas. For instance, tinea imbricata caused by *Trichophyton concentricum*, is found in hot, humid areas of the Far East, Polynesia, and South America. Scalp ringworm tends to occur in well-defined endemic areas in Africa and elsewhere. In different regions, different species of dermatophytes may predominate. Thus, in North Africa, the most common cause of tinea capitis is *Trichophyton violaceum*; in southern parts of the continent, the major agents may be *Microsporum audouinii*, *Microsporum ferrugineum*, and *Trichophyton soudanense*. Not all dermatophyte infections are endemic and dominant species may disappear to be replaced by others. *M. audouinii*, once endemic and common in the United Kingdom, is now infrequent, probably because of improved treatment and detection of carriers. By contrast *Trichophyton tonsurans* is now established as a major cause of tinea capitis in urban areas in both the United Kingdom and the United States. Dermatophytes may be passed from person to person (anthropophilic infections), from animal to person (zoophilic), or soil to person (geophilic). Sources of zoophilic organisms in Europe include cats and dogs, cattle, hedgehogs, and small rodents. Rarer sources include horses, monkeys, and chickens. Lesions produced by zoophilic species may be highly inflammatory.

Factors governing the invasion of stratum corneum are largely unknown, but heat, humidity, and occlusion have all been implicated. Susceptibility to certain infection, such as tinea imbricata, may be genetically determined.

Clinical features

The clinical features of dermatophyte infections are best considered in relation to the site involved. Often the term tinea, followed by the Latin name of the appropriate part (such as *corporis*—body) is used to describe the clinical site of infection.

Tinea pedis

Scaling or maceration between the toes, particularly in the fourth interspace, is the most common form of dermatophytosis seen in temperate countries. Itching is variable, but may be severe. Sometimes blisters may form both between the toes and on the soles of the feet. The causative organisms are commonly *T. rubrum* and *Trichophyton interdigitale*, the latter being responsible for the vesicular forms. Similar appearances can be caused by *Candida albicans* and in the bacterial infection, erythrasma. Gram-negative bacterial infection causes erosive interdigital disease associated with discomfort.

'Dry type' infections of the soles and palms

These are normally caused by *T. rubrum*. Palms or soles have a dry, scaly appearance, which in the soles may encroach on to the lateral or dorsal surfaces of the foot. The palmar involvement is often unilateral, an important diagnostic feature (Plate 1). Nail invasion is often seen (see below). Itching is not prominent, and infections are usually chronic.

Tinea cruris

Infections of the groin, most often caused by *T. rubrum* or *Epidermophyton floccosum*, are relatively common. They occur in both tropical and temperate climates, although in the former the infection may spread to involve the whole waist area in both males and females. Tinea cruris in females is uncommon in Europe. An erythematous and scaly rash with a distinct margin extends from the groin to the upper thighs or scrotum. Itching may be severe. Coincident tinea pedis is common, and patients should be examined for this. The rash of crural erythrasma shows uniform scaling without a margin, whereas in candidosis, satellite pustules occur distal to the rim.

Onychomycosis (caused by dermatophytes)

Invasion of the nail plate is most often seen with *T. rubrum* infections. The plate is invaded distally and becomes thickened and friable with terminal loss of the nail plate. Onycholysis may be seen. More rarely, and most often with *T. interdigitale*, the dorsal surface of the plate is invaded, causing superficial white onychomycosis.

Tinea corporis (body ringworm)

Dermatophyte or ringworm infection on the trunk or limbs may produce the characteristic annular plaque with a raised edge and central clearing (Plate 2). Scaling and itching is variable. Lesions caused by zoophilic organisms may be highly inflammatory and in certain cases, particularly those caused by *Trichophyton verrucosum*, intense itching, oedema, and pustule formation (kerion) may develop. This reaction is seldom secondarily infected by bacteria but is a response to the fungus on hairy skin. Infections of the beard, tinea barbae, are often highly refractory to treatment. Facial dermatophyte infections may mimic a variety of non-fungal skin diseases, including acne, rosacea, and discoid lupus erythematosus. However, the underlying annular configuration can usually be distinguished. The term tinea incognito is used to describe such atypical lesions.

Tinea capitis (scalp ringworm)

In the United Kingdom as in the United States, the most common cause of scalp ringworm is *Trichophyton tonsurans*, an anthropophilic fungus which mainly occurs in inner cities, particularly in black Caribbean or African children. This has now replaced *Microsporum canis*, originating from an infected cat or dog, although this dermatophyte is dominant elsewhere. Scalp ringworm is mainly a disease of childhood, with rare infections occurring in adult women. Spontaneous clearance at puberty is the rule. *M. canis* causes an 'ectothrix' infection where spores form on the outside of the hair shaft and the scalp hair breaks above the skin surface. Scaling, itching, and loss of hair occur. Other causes of ectothrix infection include *M. audouinii*, which is becoming more common in Europe, and is still seen in the tropics. This infection can be spread from child to child and causes serious social handicap. The infection may occur in epidemic form, particularly in schools. By contrast, infections with *M. canis* are acquired from a primary animal source rather than by spread from human lesions. In endothrix infections where sporulation is within the hair shaft, scaling is less pronounced and hairs break at scalp level (black dot ringworm). Examples include *T. tonsurans* and *T. violaceum*, the latter being most prevalent in the Middle East, parts of Africa, and India, although it also is being recognized with increasing frequency in Europe.

Favus, now most often seen in isolated foci in the tropics, is a particularly chronic form of ringworm where hair shafts become surrounded by a necrotic crust or scutulum. Individual crusts coalesce to form a pale, unpleasant-smelling mat over parts of the scalp. Such infections may cause extensive and permanent hair loss.

Tinea imbricata (tokelau)

This infection is endemic in parts of the Far East, West Pacific, and Central and South America, and is caused by *Trichophyton concentricum*. In many cases the trunk is covered with scales laid down in concentric rings producing a 'ripple' effect. Alternatively, large, loose scales (tiled, Latin— *imbricata*) may form. The infection is often chronic, and may constitute a serious social handicap. There is some evidence that susceptibility of this disease in Papua New Guinea may be inherited as an autosomal recessive trait.

Infection in HIV and immunocompromised patients

While dermatophyte infections are no more common in the immunocompromised patient, they may differ clinically. In patients with HIV infections there may be (i) more tinea facei, (ii) more widespread and atypical skin lesions, and (iii) a distinct pattern of nail infection characterized by white discoloration spreading rapidly through the nail plate from the proximal nail fold.

Laboratory diagnosis

The mainstays of diagnosis are direct microscopy of skin scales mounted in potassium hydroxide (20 per cent) to demonstrate hyphae, and culture. Scalp hairs may also be examined in a similar way, and the site of arthrospore formation, inside or outside the shaft, determined. Fluorescent whitening agents (Calcofluor) or chlorazol black stain have been used to highlight fungi in scales. Further tests, such as the ability to penetrate hair, may be used to separate similar cultures. Identification of organisms is important, as it will indicate the source of infection in scalp ringworm, for example. When large numbers of children are involved, screening of scalp infections with a filtered ultraviolet (Wood's light) lamp is useful. Certain species, including *M. canis* and *M. audouinii*, cause infected hair to fluoresce with a vivid greenish light. Scalps can also be screened for infection by passing a sterile brush or scalp massager through the hair and plating this directly on to an agar plate.

Treatment

The treatment of dermatophyte infections depends to an extent on the nature and severity of infection. Topical therapy is reserved for circumscribed infections such as athlete's foot and tinea corporis, not involving hair or nail keratin. Scalp and nail infections, severe or widespread ringworm, and failures of topical therapy are usually treated orally with griseofulvin, itraconazole, or terbinafine.

Specific antifungal drugs in topical form are effective and well tolerated. The important compounds in this group are miconazole, clotrimazole, ketoconazole, and econazole, which are imidazole derivatives, undecenoic acid, and tolnaftate and the allylamine, terbinafine. Generally treatment is given for 7 to 30 days. They are all very similar in their clinical efficacy, but topical terbinafine is particularly rapid in foot infection (7 days or less). Adverse reactions are rare.

For oral therapy the main alternatives are terbinafine, itraconazole, or fluconazole. Terbinafine (250 mg daily) is rapidly effective in most forms of dermatophytosis that require oral therapy and also produces rapid responses in toe nail (12 weeks) and sole infections (2 to 4 weeks), without a high rate of relapse. Side-effects include headache and nausea, but loss of taste may also occur. Itraconazole is somewhat similar in its profile, but is given intermittently (200 mg twice daily for 7 days). This course is given once for sole infections but repeated three times at monthly intervals for toe nail infections, as pulsed therapy. Side-effects include nausea and abdominal discomfort. Fluconazole is also active and is given in a dose of 150 mg weekly; 300 mg may be necessary for toe nail infections. This side-effect profile is similar to itraconazole. All three drugs are extremely rare causes of hepatic toxicity. Griseofulvin is still used for tinea capitis in a dose of 10 to 20 mg/kg daily. Treatment should be continued for at least 6 weeks in tinea capitis. Side-effects are not common, but include headache, nausea, and urticaria. The drug can also precipitate acute intermittent porphyria and systemic lupus erythematosus in predisposed subjects.

Further reading—dermatophytosis

de Vroey C (1985). Epidemiology of ringworm (dermatophytosis). *Seminars in Dermatology* **4**, 185–200.

Hay RJ (1982). Chronic dermatophyte infections I. Clinical and mycological features. *British Journal of Dermatology* **106**, 1–6.

Hay RJ (1997). Fungal infections. In: Bos JD, ed. *Skin immune system (SIS)*, pp 593–604. CRC Press, Florida.

Hay RJ *et al.* (1996). Tinea capitis in south-east London—a new pattern of infection with public health implications. *British Journal of Dermatology* **135**, 955–8.

Torssander J *et al.* (1988). Dermatophytosis and HIV infection—study in homosexual men. *Acta Dermatologica et Venereologica* **68**, 53–9.

Scytalidium infections

The organisms, *Scytalidium dimidiatum* (*Hendersonula toruloidea*) and *Scytalidium hyalinum*, can cause a superficial scaly condition that resembles the 'dry type' of dermatophyte infection on the palms or soles. Nail plate destruction may also occur, the lateral border of the nail being the initial site of invasion. The disease has been seen in Europe, almost invariably in immigrants from the tropics, particularly the Caribbean, West Africa, and India or Pakistan. Its prevalence in the tropics is unknown, although in some surveys it has been shown to be relatively common. In skin scrapings the tortuous hyphae may resemble those of a dermatophyte, but the organisms do not grow on media containing cycloheximide, which is often incorporated into agar for routine dermatophyte isolation.

Treatment is difficult, but some improvement may follow the use of keratolytic compounds such as salicylic acid. Nail infections do not respond to terbinafine, griseofulvin, or azoles.

Further reading—*Scytalidium*

Hay RJ, Moore MK (1984). Clinical features of superficial fungal infections caused by *Hendersonula toruloidea* and *Scytalidium hyalinum*. *British Journal of Dermatology* **110**, 677–83.

Miscellaneous nail infections

Occasionally, fungi other than dermatophytes or *Scytalidium* species are isolated from dystrophic nails. These include *Scopulariopsis brevicaulis*, *Onychocola canadensis*, *Acremonium*, and *Fusarium* species, and certain types of *Aspergillus*. These infections are usually seen in the elderly. It is often difficult, particularly with *Aspergillus* species, to establish that the organism is playing a pathogenic role.

Pityriasis versicolor (tinea versicolor)

Aetiology

Pityriasis versicolor is a superficial infection caused by *Malassezia* species. Although most common in tropical countries, it has a worldwide distribution. Dermal penetration does not occur.

There are six species of *Malassezia* that can be found on normal skin, the commonest of which are *M. sympodialis* and *M. globosa*. In pityriasis versicolor there is transformation of yeast cells to produce hyphae. It is likely that the state of host immunity plays some part in pathogenesis and depression; for instance, endogenous or exogenous corticosteroids potentiate the disease in some individuals. However, it is also commonly seen in normal individuals, and climatic factors or sun exposure are believed to trigger the infection in many cases. There is no effective animal model for studies of this disease.

Epidemiology

Pityriasis versicolor is very common in the tropics, where it may be widespread on the body. Its incidence in temperate climates has increased over the last 20 to 30 years. It is not more common in HIV infected subjects.

Clinical features

The rash of pityriasis versicolor is asymptomatic or mildly pruritic. Its presents with scaling, confluent macules on the trunk, upper arms, or neck. These may be hypopigmented or hyperpigmented. In some individuals and in the tropics, other areas including face, forearms, and thighs may be involved.

The diagnosis is rarely confused with other complaints, although eczema or ringworm infections are sometimes considered. Patients are often anxious to exclude leprosy, but the two are unlikely to be mistaken. In vitiligo, depigmentation is complete and there is no scaling.

Laboratory diagnosis

P>The diagnosis is made by demonstration of the yeasts and hyphae of *Malassezia* in skin scales removed by scraping. Culture is difficult and unnecessary.

Treatment

Topical ketoconazole, miconazole, clotrimazole, or econazole are effective. Oral itraconazole may be used in recalcitrant cases. Alternatives include 2 per cent selenium sulphide or 20 per cent sodium hyposulphate lotions. Whatever the treatment, relapse is common.

Other *Malassezia*-associated conditions

Malassezia yeasts have been implicated in the pathogenesis of a number of other skin diseases such as seborrhoeic dermatitis and a form of itchy folliculitis, *Malassezia* folliculitis. The evidence connecting seborrhoeic dermatitis, one of the most common of skin diseases, and *Malassezia* is largely concerned with the response of antifungal drugs and the observation that improvements in the rash mirror disappearance of organisms from the skin. Severity of the skin condition does not appear to reflect the numbers of yeasts on the skin surface.

Further reading—*Malassezia*

Mathes BM, Douglas MC (1985). Seborrheic dermatitis in patients with acquired immunodeficiency syndrome. *Journal of the American Academy of Dermatology* **13**, 947–51.

Superficial candidosis (candidiasis)

Aetiology

Superficial candidosis is a term used to describe a group of infections of skin or mucous membranes caused by species of the genus *Candida*. They range in severity from oral thrush to chronic mucocutaneous candidosis, a chronic infection refractory to conventional antifungal treatment.

Candida albicans is the species most frequently involved. It is a saprophytic yeast often found as a commensal in the mouth and gastrointestinal tract, and is commonly present in the vagina. Several factors may influence the incidence of carriage. For instance, oral colonization is more common in hospital staff than in equivalent non-hospital subjects. Vaginal carriage is more common in pregnancy. Other factors (Table 1) are known that predispose to conversion from a commensal to a parasitic role with the causation of disease—candidosis. The list includes factors that influence host immunological response, such as carcinoma, AIDS, or cytotoxic therapy; those that disturb the population of other micro-organisms, such as antibiotics; and those that affect the character of the epithelium, such as dentures.

Other species of *Candida* may also cause superficial infections, but are less common. They include *C. glabrata*, *C. dubliniensis*, and *C. parapsilosis.* There is evidence that the first two species are more common now in oral infection in patients with HIV and *C. glabrata* in vaginal candidosis.

Epidemiology

Superficial *Candida* infections are seen in all countries.

Clinical features

There are a number of clinically distinct types of superficial infection caused by *Candida* species, as follows.

Oral candidosis (thrush)

Oral infection by *Candida* is fairly common, particularly in infancy and old age, or in association with antibiotic or cytotoxic therapy, or in diseases where the neutrophil or T-lymphocyte responses may be impaired. In the older age group, the wearing of dentures is a predisposing factor. The lesions present with discomfort both in the mouth and at the corners of the lips. The mouth and buccal mucosa show patchy or confluent, white adherent plaques; less commonly the mucosa and tongue are sore and glazed—erythematous candidosis. Angular cheilitis usually accompanies the oral lesions. In long-standing cases, the plaque may become hypertrophic, with oedema of the mucosal surfaces, or the mucosa may appear glazed and raw.

There is a significant correlation between leucoplakia and oral candidosis, and it has been suggested that the infection may lead to epithelial dysplasia.

The diagnosis is made by the demonstration of yeasts and hyphae of *Candida* in smears, and by culture.

Vaginal candidosis (thrush)

See Chapter 21.3 for further details.

Paronychia

Infection around the nail fold is seen in people whose occupations involve frequent wetting of the hands (such as cooks) or in those with eczema or psoriasis. The aetiology is complicated and there may be a mixture of bacterial infection and irritant or allergic contact dermatitis as well as *Candida* infection. The condition presents with painful, red swelling of the nail fold. Pus may be discharged. Secondary invasion of the lateral border of the nail plate by *Candida* may occur from this site.

Candida intertrigo

Infection of the moist folds of the skin in the groin or under the breasts causes itching and discomfort. The area becomes macerated and erythematous. Candida may contribute to this condition, but is certainly not the only factor. It may also superinfect the napkin area in infants. The presence of satellite pustules (see above) is a useful indicator of involvement by *Candida* in the disease process.

Direct invasion of toe-web folds by *Candida* closely resembles 'athlete's foot' caused by dermatophytes. A similar erosive infection may occur in the finger webs—interdigital candidosis—and is seen most commonly in the tropics.

Chronic superficial candidosis

Chronic *Candida* infections of the mouth, vagina, and nail present problems in management. Chronic oral candidosis, for instance, is associated with leucoplakia. Predisposing causes should be searched for. The most serious of this group of infections is chronic mucocutaneous candidosis, a rare condition in which chronic skin, nail, and mucosal infection coexist (Plate 3). A series of underlying genetic, endocrine (hypoparathyroidism, hypoadrenalism, or hypothyroidism), and immunological abnormalities has been found. Extensive human papilloma virus (wart) or dermatophyte infections may also be present in these patients, whose condition is normally diagnosed in childhood.

Oral candidosis is one of the earliest signs of untreated AIDS, occurring in a high proportion of patients. The appearances are similar to those seen with other groups,

although plaque formation may be very extensive. Oesophageal infection is common in this group.

Laboratory diagnosis

All these infections are diagnosed by microscopy and culture. When associated with the condition, *Candida* cells are always evident on microscopy. Culture establishes the specific identity and is important particularly where non- *albicans Candida* species may be involved.

Treatment

Two groups of drugs are effective in superficial candidosis. The polyenes such as nystatin and amphotericin B are topically active in many forms of candidosis. They are often less effective in oral candidosis in immunodeficient patients including those with AIDS. Likewise, topical azole drugs such as miconazole and clotrimazole are usually effective in superficial candidosis. For resistant cases, oral therapy with fluconazole, itraconazole, or ketoconazole may be necessary.

For vaginal infections, topical creams or vaginal preparations should be used—many requiring only a single treatment. Single-dose oral fluconazole is an alternative. In recalcitrant cases it may be necessary to use longer courses of fluconazole or itraconazole.

Further reading—candidosis

Bodey GP, ed. (1993). *Candidiasis. Pathogenesis, diagnosis and treatment*. Raven Press, New York.

Greenspan D, Greenspan JS (1987). Oral mucosal manifestations of AIDS. *Dermatologic Clinics* **5**, 733–7.

Torssander J *et al.* (1987). Oral *Candida albicans* in HIV infection. *Scandinavian Journal of Infection* **189**, 291–5.

Miscellaneous superficial mycoses

There are a number of relatively rare, superficial fungal infections such as tinea nigra, and black or white piedra. They never cause invasive disease, and are mainly confined to the tropics.

Tinea nigra

Tinea nigra is a superficial infection confined to the epidermis of the palms or soles, and more rarely elsewhere. The initial lesion is a dark macule without scaling, which resembles a brown stain on the skin and spreads slowly over the palmar or plantar surface. The disease is normally asymptomatic.

On scraping the skin, brown pigmented hyphae can be seen by direct microscopy, and the causative organism, *Phaeoanellomyces werneckii*, isolated.

The lesion responds to Whitfield's ointment.

Black piedra

Black piedra is a disease of the tropics in which small, dark nodules form on hair shafts in the scalp or, less commonly, elsewhere. There are no symptoms. Each nodule consists of a dense mat of hyphae containing the sexual spores (ascospores) of the fungus.

The diagnosis is made by direct microscopy of infected hair, and the isolation of *Piedraia hortae*. Treatment using formalin solution or amphotericin B lotion is usually effective.

White piedra

White piedra occurs in both temperate and tropical climates, and is rare. It produces pale nodules on the hair of the beard, groin, or scalp. The hair shaft may fracture. The nodule consists of hyphae, arthrospores (spores formed by fragmentation of hyphae), and blastospores (budding yeast cells). The organism *Trichosporon beigelii* can be readily cultured. The treatment is similar to that for black piedra.

The subcutaneous mycoses

Subcutaneous infections caused by fungi are rare, and are mainly seen in the tropics. The organisms gain entry via the skin; in mycetoma, organisms may be implanted subcutaneously via a thorn. The majority of the causative organisms in this group of infections can be isolated from vegetation or soil. Involvement of deep viscera is rare. Attempts to establish experimental infections that resemble the human diseases have been largely unsuccessful. A clearer understanding of the pathogenesis therefore awaits such a model system. These infections tend to be chronic, chemotherapy may be lengthy, and in the case of mycetoma, often unsuccessful.

Mycetoma (Madura foot)

Aetiology

Mycetoma is a chronic infection involving subcutaneous tissue, bone, and skin, in which colonies of infecting fungi or actinomycetes (grains) are found within a network of burrowing abscesses and sinuses (Plate 4).

A list of the more common organisms that cause mycetoma is shown in Table 2. The organisms are divided into two groups, the actinomycetomas and the eumycetomas, caused by actinomycetes and fungi, respectively. The size and colour of the grains (red, pale, or dark) are important clues to their identification. The organisms can be found in the natural environment, and some have even been identified in association with acacia thorns in an endemic area. The infection is initiated when an infected thorn is implanted in deep tissue. However, many years may elapse before the formation of a clinically apparent mycetoma.

Epidemiology

The disease is seen primarily in the tropics, although rare cases, apart from imported ones, may occur in temperate areas. Countries with the most reported cases include Sudan, India, Senegal, Mexico, and Venezuela. However, the disease is widely distributed in the tropics, particularly to the south and east of the Sahara Desert in Africa.

The pattern of prevalence of infections caused by certain organisms differs strikingly in different parts of the world. For instance, *Streptomyces somaliensis* is most common in the Sudan and Middle East. *Madurella grisea* is mainly found in the New World. Altogether about 60 per cent of reported infections are caused by actinomycetes, of which *Nocardia brasiliensis* is the most common (Chapter 7.11.27).

Clinical features

Early mycetomas may present with a circumscribed area of hard subcutaneous swelling (Plate 5). Later, sinus tracts open on to the skin surface and visible grains may be discharged, along with serosanguinous fluid (Plate 6). Bone erosion and destruction, leading to deformity, may occur. However, severe pain is rarely a problem. Local lymph node invasion may occur, but more widespread involvement is very rare.

Feet and lower legs are the areas most commonly involved, but the arms, buttocks, chest, and head may all be sites of infection. Mycetoma caused by *N. brasiliensis*

may occur in any site, but one favoured area is the chest wall.

The radiological features of mycetoma are cortical erosion, followed by the development of lytic deposits in bone. Periosteal proliferation and destruction, leading to deformity, may follow. MRI provides a clearer picture of bone involvement and may be positive earlier than radiography.

Laboratory diagnosis

The diagnosis is made by the demonstration and identification of grains obtained from the sinus openings by gentle pressure or curettage. If these measures are not successful, tissue should be obtained by deep surgical biopsy. Grains can be mounted in potassium hydroxide and examined microscopically. Those containing filaments of 3 to 4 µm in diameter or more are caused by true fungi (eumycetomas), and those with filaments of less than 1 µm by actinomycetes (actinomycetomas). These features can usually be distinguished by direct microscopy.

The morphology of grains fixed, sectioned, and stained with haematoxylin and eosin is typical. Special stains are less helpful. Grains can be used for culture, although several attempts at isolation may have to be made. Serology (such as immunodiffusion) can also be helpful, although the tests are not widely available.

Treatment

Actinomycetomas may respond to sulphones such as dapsone (50 to 100 mg daily) or sulphonamides such as sulphadiazine. The treatment of choice for many is long-term co-trimoxazole (two to three tablets twice daily) with an initial 2 to 3 months of streptomycin or rifampicin. Treatment may have to be continued for many months or years. Dapsone is an effective and cheaper alternative to co-trimoxazole. Extensive actinomycetomas may respond poorly and additional treatment with amikacin or fucidin may be necessary. The eumycetomas seldom respond to antifungal therapy. About 50 per cent of *Madurella mycetomatis* infections respond to ketoconazole. In other infections griseofulvin, amphotericin B, ketoconazole, and itraconazole have rarely produced remission or cure. A trial of therapy may be attempted, where the patient can be monitored closely in outpatient departments. Otherwise, radical surgery or amputation is usually necessary. Small, local excisions are rarely successful.

Mycetoma is slowly progressive and increasingly disabling. However, wider dissemination is very rare, and therefore cases are seldom fatal, except where the skull is involved. However, the deformity caused by the disease may be severely disabling.

Further reading—mycetoma

Hay RJ (1997). Granule forming pathogenic mould fungi. In: Ajello L, Hay RJ, eds. *Mycology. Topley and Wilson's Microbiology and Microbial Infections*, 9th edn, Vol 4, pp 487–98. Arnold, London.

Mahgoub ES (1976). Medical management of mycetoma. *WHO Bulletin* **54**, 303–10.

Chromoblastomycosis

Aetiology

Chromoblastomycosis, one of the intermediate subcutaneous mycoses, is a chronic granulomatous fungal infection characterized histologically by the presence of brown, spherical fungal cells known as sclerotic cells or fumagoid bodies. In most cases, the lesions are confined to the skin and subcutaneous tissues. In the past there has been great confusion over nomenclature of the aetiological agents of chromoblastomycosis. At present, five agents assigned to four genera are recognized as causing chromoblastomycosis. They are:

1. *Fonsecaea pedrosoi*, which occurs in high rainfall areas and is found worldwide;
2. *Cladophialophora carrionii*, the sole cause of chromoblastomycosis in arid areas;
3. *Phialophora verrucosa*, the first agent to be described;
4. *Fonsecaea compactum*, an uncommon cause and isolated only a few times;
5. *Rhinocladiella aquaspersa*, the rarest cause.

Sporadic cases caused by other dematiaceous fungi such as *Cladosporium trichoides* and *Taeniolélla boppii* have been reported from Uganda and Brazil.

Epidemiology

The principal endemic areas for chromoblastomycosis are the tropical and subtropical countries including Central and South America, Costa Rica, Africa, Japan, Australia, Malagasy, and Indonesia. Curiously, sporadic cases have been reported from Finland and Russia.

Although soil itself does not seem to be a particularly good substrate, the various agents of chromoblastomycosis occur as saprobic fungi in the environment and have been isolated from soil, decaying vegetation and rotting wood. Strains of *F. pedrosoi* and *P. verrucosa* have been isolated from the atmosphere but proved less virulent than those isolated from human lesions or organic material.

Infection occurs as a result of trauma, however minor, the fungi gaining entrance through a cut, abrasion, or thorn prick. Farmers and labourers in agricultural areas are most likely to be exposed to contaminated material. Although lesions on exposed areas may be accounted for in this way it was suggested by Wilson in 1958 that lesions on non-exposed areas may result from a previously unrecognized pulmonary focus. Bacquero later demonstrated the presence of *F. pedrosoi* in bronchial washings and subsequently proved their pathogenicity by inoculating those strains into normal skin of human volunteers and recovering the fungus from the ensuing skin lesions. Other methods of transmission have included metal particles from automobiles, and acupuncture. Person-to-person and animal-to-man transmission have not so far been reported. Chromoblastomycosis has rarely been reported in children and it may be that factors other than trauma and exposure to contaminated material are necessary for its development.

Pathogenesis

Host resistance and virulence of the organism are the two main factors associated with the pathogenesis of this disease. Chromoblastomycosis occurs mainly in healthy individuals. However, it has been found in patients where immunosuppression has occurred either from underlying disabling disease or from chemotherapy. Although the mechanism of granuloma formation is not well understood, it appears that lipids extracted from these fungi and cell-wall constituents may be responsible for this reaction.

Clinical features

The initial lesion of chromoblastomycosis is a small papule at the site of trauma, which gradually enlarges. Nodules and tumours develop, producing a malodorous discharge; eventually, over a period of years, a wide variety of morphological patterns may emerge including dry, hyperkeratotic plaques, verrucose lesions, and large, cauliflower-like masses. Extensive cicatricial plaques, surrounded by peripherally spreading vegetative lesions, may also be present. Evolution is slow and lesions usually involve the lower limb. However, any part of the body may be involved and the sites may be multiple.

Dissemination occurs by (i) surface spread, (ii) the lymphatics, the most common method, (iii) autoinoculation from scratching, and (iv) haematogenously, resulting in subcutaneous lesions at sites distant from the primary. Visceral metastases are known to occur and involvement of the central nervous system, respiratory system, larynx, and vocal chords has been recorded. Therapeutically, therefore, early diagnosis is important.

Complications of long-standing chromoblastomycosis include lymphoedema, flexion deformity of joints, and development of squamous carcinoma.

Diagnosis

Although the history and clinical presentation may suggest the diagnosis, the varied clinical presentation of chromoblastomycosis necessitates consideration of other granulomatous diseases such as sporotrichosis, cutaneous tuberculosis, Hansen's disease, blastomycosis, candidosis, leish-maniasis, paracoccidioidomycosis, rhinosporidiosis, tertiary syphilis, squamous carcinoma, and even psoriasis, sarcoidosis, and discoid lupus erythematosus.

Therefore, to establish a definitive diagnosis, histological and mycological investigations are essential. Diagnosis is confirmed by the presence of the characteristic brown, sclerotic bodies in histological sections. From both epidemiological and therapeutic points of view, culture is necessary as *F. pedrosoi* is the most difficult of the causative fungi to eradicate whereas *C. carrionii* responds rapidly to treatment.

Treatment

Small, single, localized lesions are satisfactorily eradicated by cryosurgery, but long-term follow-up is needed to assess accurately the success of this treatment. Thermotherapy has been found effective by some, again principally in the management of small, single lesions, but here the possibility of a burn must be borne in mind. Rapid spread of the disease has been associated with inadequate surgery, curettage, and electrodesiccation.

Oral monotherapy has been unsuccessful in some cases and drug resistance remains a problem. However itraconazole and terbinafine have both been reported as effective agents. A combination of 5-flucytosine with either thiabendazole or itraconazole may also be efficacious., particularly in long-standing disease

Whatever method of treatment is used, chromomycosis although clinically healed, should be followed-up for at least 2 years before its total eradication can be assumed.

Further reading—chromoblastomycosis

Bayles MAH (1989). Chromomycosis. In: Tropical fungal infections, *Baillière's clinical tropical medicine and communicable diseases*, Vol. 4, pp. 45–70. Baillière Tindall, London.

Bacquero GF, Lopez BP, Lescay BR (1961). Cromoblastomicosis experimental: cromoblastomicosis producida experimentalmente con cepas de *Hormodendrum pedrosoi* obtenida por lavado bronquial de enfermos que padecen la affecion. *Boletin de la Sociedad Cubana de Dermatologia y Sifilografia* **18,** 19–28.

Grigoriu D, Delacretaz J, Borelli D (1987). In *Medical mycology*, (English edn), pp. 333–42. Hans Huber, Toronto.

McGinnis MR, Ajello L, Schell WA (1985). Mycotic diseases: a proposed nomenclature. *International Journal of Dermatology* **24,** 9–15.

Silva CL, Ekizlerian SM (1985). Granulomatous reaction induced by lipids extracted from *Fonsecaea pedrosoi*, *Fonsecaea compactum*, *Cladosporium carrionii* and *Phialophora verrucosum. Journal of General Microbiology* **131**, 187–94.

Silva CL, Fazioli RA (1985). Role of the fungal cell wall in the granulomatous response of mice to the agents of chromomycosis. *Journal of Medical Microbiology* **20,** 299–305.

Wilson JW (1958). Importancia de las enfermedades fungosas en imunologia. *Boletin de la Sociedad Cubana de Dermatologia y Sifilografia* **15,** 115–24.

Sporotrichosis

Aetiology

The most common clinical form of sporotrichosis is a subcutaneous infection, which may spread proximally from its initial site in a series of nodules along the course of a lymphatic. More rarely, systemic involvement is seen, for example in the lung (see Systemic mycoses, below).

The causative organism *Sporothrix schenckii* can be found in soil, vegetation, or in association with plants or bark. People who develop the subcutaneous infection may have had contact with material that harbours the organism, such as moss or flowers (for example florists). It is assumed that the pathogen gains entry via an abrasion and in some endemic areas there is often a preceding history of a scratch or insect bite.

Epidemiology

Although sporotrichosis was once prevalent in Europe, particularly France, non-imported cases are now very rare in this area. However, the disease is seen in the United States, Mexico, Central and South America, and Africa. In the late 1930s, there was a remarkable epidemic of sporotrichosis in workers in the Witwatersrand gold mines. The source of infection was a large number of wooden pit props contaminated with the organism. Other, smaller 'epidemics' have been described in certain groups, such as Mexican pottery workers packing ceramics in straw. Normally, however, cases are sporadic in incidence. There are also 'hyperendemic' areas where there is an unexpectedly high incidence of this infection.

Systemic sporotrichosis is much rarer, and cases have mainly been described from the United States.

Clinical features

There are two main clinical types of subcutaneous sporotrichosis.

The first, the fixed type, presents with a solitary cutaneous ulcer or nodule. In this form of the disease, infection does not spread along lymphatics. It has been suggested that it is most common in children, and it has been described most frequently in Central and South America.

In the lymphangitic form, an initial nodule forms on a limb or extremity, such as a finger. This may break down and ulcerate. Subsequently, one or more secondary nodules develop along the draining lymphatic channel, which may ulcerate through the skin. Other variants include the psoriasiform or verrucous types or a superficial granuloma that resembles lupus vulgaris. These usually represent chronic infection.

Rarer forms include secondary spread via scratching, which may present with multiple widespread ulcers or multiple cutaneous lesions secondary to systemic disease. In HIV-positive individuals, widespread cutaneous lesions may develop.

Fixed-type sporotrichosis may resemble many other forms of cutaneous ulceration. However, in endemic areas a major source of confusion is cutaneous leishmaniasis. The lymphangitic variety may also resemble other infections, notably atypical mycobacterial infections, particularly fish-tank granuloma, or 'sporotrichoid' leishmaniasis.

Treatment

Some cases of sporotrichosis may heal spontaneously. However, treatment is usually advised to prevent scar formation. The cheapest treatment is potassium iodide, which is administered in a saturated aqueous solution. The starting dose is 0.5 to 1 ml, given three times daily, and this is increased drop by drop per dose to 3 to 6 ml, three times daily. The mixture is more palatable if given with milk. Treatment should be given for a month after clinical resolution. However, both itraconazole and terbinafine are also effective; minimal durations of treatment for these agents have not been defined.

Further reading—sporotrichosis

Bibler MR *et al.* (1986). Disseminated sporotrichosis in a patient with HIV infection after treatment for acquired factor VIII inhibitor. *Journal of the American Medical Association* **256**, 3125–6.

de Albornoz MCB (1989). Sporotrichosis. In: Hay RJ, ed. *Tropical fungal infections*, *Baillière's clinical tropical medicine and communicable diseases*, Vol. 4, pp. 71–96. Baillière Tindall, London.

Subcutaneous zygomycosis due to *Basidiobolus*

Subcutaneous zygomycosis is an infection primarily seen in children in Africa or the Far East (Indonesia). It is characterized by the development of localized woody swellings on the limbs or trunk. The swelling is rarely inflammatory, but has a well-defined leading edge, and is hard. Progression is slow. The causative organism *Basidiobolus haptosporus* can be cultured or demonstrated histologically in biopsy material. Although resolution has been recorded without treatment, therapy is normally given. Potassium iodide solution is the treatment of choice, and is given in as high a dose as possible (see Sporotrichosis, above). Itraconazole may also be effective.

Subcutaneous zygomycosis due to *Conidiobolus* (conidiobolomycosis or rhinoentomophthoromycosis)

Conidiobolomycosis is a similar infection confined to subcutaneous tissue and presenting with painless swelling. The infection is mainly seen in West Africa, but a case has been seen in the Caribbean. There are important differences from the subcutaneous zygomycosis caused by *Basidiobolus*. The disease is most common in young adults, and is confined to facial tissues around the nose, the forehead, and the upper lip. The initial site of infection is in the region of the inferior turbinate in the nose. The diagnosis is established by biopsy or culture. The causative organism is *Conidiobolus coronatus*. Treatment with itraconazole or ketoconazole is effective, but an alternative is high-dose potassium iodide. Relapse after treatment is common, and residual fibrosis may be severely disfiguring.

Lobo's disease (lobomycosis)

Lobo's disease is a subcutaneous infection. The organism, in tissue, appears to be a yeast. It has a tendency to form chains of four to six yeast cells with prominent nucleoli, joined by a narrow, intercellular bridge. However, the organism has never been cultured from human cases and can only be identified by biopsy and histology. The disease is seen in countries of South America around, and north of, the Amazon basin, and cases are also seen in Central America. Apart from humans the only other species affected are freshwater dolphins. Often, exposed sites (such as ear lobes) are invaded and small nodules containing the organisms develop. These may resemble keloids (Plate 7). More diffuse plaques may also be seen. Deep invasion has not been documented. The treatment is excision, and there is no effective chemotherapy.

Systemic mycoses

The systemic or deep visceral mycoses include some of the rare and more serious of the fungal infections. There are two main types of infection in this group, those caused by organisms which invade normal hosts, the endemic mycoses, and those which only cause disease in compromised patients, the opportunistic mycoses. The fungi associated with these two types of infection differ in their innate levels of pathogenicity, but an element of opportunism, depending on host susceptibility, is usually recognizable in all cases of systemic mycoses.

The endemic pathogens cause infections such as histoplasmosis or coccidioidomycosis. These diseases have well-defined endemic zones and the majority of those exposed remain symptomless but usually develop positive skin tests. However, in certain patients, chronic local or disseminated disease may occur. In the systemic infections caused by opportunistic fungi, there is usually a serious underlying abnormality in the patient affecting T lymphocytes (such as HIV) or neutrophils (such as cancer chemotherapy). Such infections are worldwide in occurrence: where tissue invasion occurs the mortality is high. Cryptococcosis, a systemic yeast infection, has features of both types of systemic disease and occurs in both normal and immunosuppressed subjects.

The systemic endemic infections are histoplasmosis, coccidioidomycosis, blastomycosis, paracoccidioidomycosis, and infections due to *Penicillium marneffei.* The significance of various laboratory tests in these infections is shown in Table 3.

Further reading—systemic mycoses

de Pauw BE, Meunier F (1999). The challenge of invasive fungal infection. *Chemotherapy* **45**(Suppl 1), 1–14.

Histoplasmosis (see also Section 17)

There are two forms of histoplasmosis. In both types, the organism is present in tissue in its yeast phase. In small-form or classic histoplasmosis, the diameter of the yeast cells is between 3 and 4 µm. Infections are most common in the United States, but sporadic cases are reported widely from the New World, Africa, and the Far East. By contrast, large-form or African histoplasmosis is most common in Central Africa, south of the Sahara and north of the Zambezi river. Yeast forms in infected tissue are much larger, 10 to 15 µm in diameter. Both infections are clinically distinct (see below), but cultural isolates are indistinguishable.

Histoplasmosis (classic or small-form histoplasmosis)

Aetiology

Histoplasmosis is a systemic infection caused by *Histoplasma capsulatum*. The main route of infection is pulmonary. The majority of those exposed are sensitized without overt signs of infection, but more rarely chronic pulmonary or disseminated forms of the disease are seen.

The organism, *H. capsulatum*, can be found in soil in endemic areas. Its growth is facilitated by the presence of bird excreta, for instance in old chicken houses, bird roosts, and barns. In tropical and some temperate areas, bat guano plays a similar role. Exposure to a suitable source, such as a cave containing bats, is often recorded in acute epidemic histoplasmosis (see below). It is rarely identified in more slowly evolving cases.

The condition of the host is important in determining the clinical course and manifestations of histoplasmosis. Slowly evolving (chronic), disseminated disease may occur in normal individuals. However, infants, elderly people, or those with untreated AIDS appear to be more likely to develop the more rapidly progressive forms of disseminated infection.

Epidemiology

The major endemic area, as shown by skin testing, is in the central region of the United States around the Ohio and Mississippi valley basins. Prevalence is highest in the states of Tennessee, Kentucky, and Ohio. Up to 95 per cent of those skin tested in certain parts of these areas have positive delayed reactions to intradermal histoplasmin (compare Mantoux test). Scattered cases of active disease, healed calcified foci in chest radiographs, and foci found at autopsy representing inactive histoplasmosis also provide evidence of spread within this area. However, the disease also occurs in other parts of the United States, Mexico, Central and South America, Africa, the Far East, and Australia. Outside the major endemic areas in the United States, human cases are less frequent, and much of the evidence of the endemicity comes from positive skin tests or the presence of the organism in selected sites, such as caves. Although there has been considerable discussion on the nature of soil factors responsible for the growth of *H. capsulatum*, the conditions limiting its occurrence to certain areas are largely unknown.

Clinical features

The clinical forms of histoplasmosis can be placed in several groups:

1. asymptomatic;
2. acute symptomatic pulmonary:
    i. acute epidemic,
    ii. acute reinfection;
3. chronic pulmonary;
4. disseminated (acute, subacute, and chronic); and
5. primary cutaneous (by inoculation).

Asymptomatic infection

Over 99 per cent of patients becoming infected in endemic areas record no overt symptoms but develop a positive skin test. The incidence of positive skin tests declines in individuals above the age of 60 years.

Acute (symptomatic) pulmonary histoplasmosis

Acute epidemic histoplasmosis

Groups of individuals exposed to a source of infection, for instance during cave exploration, or those who may have inhaled a large infecting dose, often develop a symptomatic illness 12 to 21 days after exposure. The main features are pyrexia, cough, chest pain, and malaise. Flitting arthralgia and, less commonly, erythema nodosum or multiforme may occur. The radiological appearances may be much more severe than would be supposed from the symptoms, and enlargement of hilar lymph nodes and diffuse or patchy consolidation suggesting pneumonitis may occur (Plate 6).

These patients develop precipitating or complement-fixing antibody, but this often follows the peak of illness. About 50 per cent of those with symptoms do not develop positive antibody responses. Likewise, skin test conversion is often too late to be of diagnostic value, and its use is normally contraindicated, as a single histoplasmin test may cause the development of false-positive serological results. Cultures are often negative. The symptoms and history of exposure to a suitable source, combined with a rising antibody titre, are often the best evidence of infection.

The majority of cases require no specific therapy apart from rest. Those with severe or prolonged symptoms or impaired gas exchange require intravenous amphotericin B or itraconazole. The lung lesions often heal to leave multiple scattered pulmonary calcifications.

Acute reinfection histoplasmosis

Massive acute exposure to *H. capsulatum* in sensitized individuals is believed by some physicians to cause a less severe infection associated with bilateral pulmonary infiltrates. The incubation period is shorter than with acute epidemic histoplasmosis, namely 5 to 10 days.

Chronic pulmonary histoplasmosis

Chronic pulmonary disease caused by *H. capsulatum* is mainly seen in the United States. It is more common in males and smokers, and there is often underlying pulmonary disease such as emphysema. Early cases may present with pyrexia and cough, but malaise and weight loss occur later. Lesions may heal initially, but relapse is common, leading to established consolidation and cavitation. The most common radiological appearance of early lesions is of unilateral, wedge-shaped, segmental shadows in the apical zones. Subsequently, the disease may become bilateral, with fibrosis and cavitation. In some cases, extensive and progressive destruction of lung tissue may occur.

Culture and serology are both helpful methods of diagnosis in this form of histoplasmosis, but repeated attempts may be required before positive results are obtained.

In early cases, resolution may occur on rest alone. However, relapse occurs in at least 25 per cent of cases, and these patients may require amphotericin B therapy or itraconazole. Although chemotherapy may virtually sterilize lesions, fibrosis persists and relapse may occur. Surgical excision or lobectomy is sometimes effective.

Solid lung tumours may persist after the primary infection. These may be single (coin lesions) or multiple, and have to be distinguished from carcinomas. The diagnosis is normally made at surgery, although the presence of calcification may give a clue to the nature of the lesion (histoplasmoma). The organisms can be demonstrated by histopathology, but they are seldom viable.

Disseminated histoplasmosis

There is considerable variation in the rate of progression of histoplasmosis that has spread beyond the initial focus in the lung. In rapid or acutely disseminated cases, widespread infiltration of reticuloendothelial cells of bone marrow, spleen, and liver may occur. Gastrointestinal lesions, endocarditis, and meningitis are less common, and meningitis is more usually associated with a slower course of disseminated disease. Infants, elderly people, or immunosuppressed patients are more susceptible to acute dissemination. The most prominent symptoms are fever and weight loss, with accompanying hepatosplenomegaly. Extensive purpura and bruising secondary to thrombocytopenia may occur. The blood picture may reflect marrow infiltration with organisms, leading to pancytopenia. Disseminated histoplasmosis is also seen in patients with AIDS. The clinical manifestations are not significantly different, although skin papules and ulcers have been reported in many; isolation of *Histoplasma* from blood has also been reported more frequently in these patients. Cultures, including sputum or bone marrow, should be taken. Serology is often positive, with high titres of complement-fixing antibodies occurring in some patients. However, new antigen detection systems in serum or urine provide a better means of confirming the diagnosis and monitoring treatment.

A much more slowly progressive form of disseminated histoplasmosis may present with fewer localized lesions, such as persistent oral ulcers, chronic laryngitis, or adrenal insufficiency. Granulomas, few of which contain organisms, can be found in the liver in some patients. Such cases may present up to 30 years after the patient has left an endemic area. Outside endemic areas this form is the most widely recognized presentation of histoplasmosis, occurring in Europeans, for instance, who have worked in Africa or the Far East.

The diagnosis of disseminated histoplasmosis is made on culture or biopsy of affected areas. Antibodies may only be positive in low titres and in all cases adrenal involvement should be looked for.

Treatment is required in all forms of disseminated histoplasmosis. Itraconazole is preferred by most physicians, although amphotericin B may be necessary in some patients. Oral ketoconazole is an alternative. In patients with AIDS who are acutely ill, the disease is often controlled by a short (2 week) course of amphotericin B and thereafter patients receive continuous itraconazole indefinitely.

Primary cutaneous histoplasmosis

Primary infection sometimes follows accidental inoculation of viable organisms in a laboratory or autopsy room. This type of infection is normally associated with a chancre at the site of inoculation and regional lymphadenopathy. The condition is self-limiting.

African histoplasmosis

Overt pulmonary involvement is rare in this form of histoplasmosis, and the normal portal of entry of the pathogen is not known. The most common presenting features are skin lesions (papules, nodules, abscesses, or ulcers) (Plate 8) or lytic bone deposits. Solitary or multiple foci may be present, and in the latter instances rapid progression and death may occur. In such cases, gastrointestinal and lung lesions may develop.

The diagnosis is normally made by culture, smear, or biopsy. The organism *H. capsulatum* var. *duboisii* is identical to that causing classic histoplasmosis in culture, but in lesions the yeast forms are considerably larger (10 to 15 µm).

While local excision of skin nodules has been reported to be curative, treatment with itraconazole, ketoconazole, or amphotericin B is usual. Some patients will respond to co-trimoxazole. A skeletal scan should be made to detect occult foci of infection.

Further reading—histoplasmosis

Ashford DA *et al.* (1999). Outbreak of histoplasmosis among cavers attending the National Speleological Society Annual Convention, Texas, 1994. *American Journal of Tropical Medicine and Hygiene*

60, 899–903.

Goodwin RA, Loyd JE, DesPrez RM (1981). Histoplasmosis in normal hosts. *Medicine* **60**, 231–66.

Khalil MA, Hassan AW, Gugnani HC (1998). African histoplasmosis: report of four cases from north-eastern Nigeria. *Mycoses* **41**, 293–5.

Mandell W, Goldberg DM, Neu HC (1986). Histoplasmosis in patients with the acquired immune deficiency syndrome (AIDS). *Annals of Internal Medicine* **111**, 655–9.

Blastomycosis (see also Section 23)

Blastomycosis (North American blastomycosis) caused by *Blastomyces dermatitidis* is a systemic fungal infection in which skin and lung involvement are common features.

The infective organism, *B. dermatitidis*, has only been isolated from the environment on rare occasions. Positive sites have included soil and rotten timbers. The organism infects humans and domestic animals, particularly the dog.

Epidemiology

Blastomycosis was originally thought to be confined to North America, where it occurs sporadically throughout the south and east-central area, and in areas of central Canada. 'Epidemics' of acute disease are rare, and where these occur a source of infection is rarely demonstrated. There is evidence that sources may include areas exposed to flooding.

More recently, cases have been found in Africa. Again, these are widely scattered from the north coast to the southern parts of the continent, and are rare in all areas. Patients with the disease have also been reported from the Middle East and Central Europe.

Clinical features

The clinical forms of blastomycosis differ from histoplasmosis in a number of important aspects. The existence of an asymptomatic form has not been proved conclusively, because there is no reliable skin test. Acute infections or infections in groups are rare, and the features are often similar to histoplasmosis (acute pulmonary). However, specific serological tests may be negative in 30 to 50 per cent of cases. The demonstration of the organisms in sputum and positive cultures are more reliable diagnostic criteria. Although some cases undoubtedly resolve without sequelae, some physicians advise chemotherapy, with a short course of amphotericin B in acute cases of blastomycosis.

Chronic pulmonary blastomycosis

Chronic consolidation or cavitation of the upper or mid-zones occur with chronic pulmonary infections. Fever, malaise, and cough with sputum are seen. Weight loss may be prominent. Culture is again the most reliable method of diagnosis.

The mainstays of treatment are itraconazole or amphotericin B.

Disseminated blastomycosis

Although generalized infiltration in skin, lungs, and liver may occur over a short period, leading to rapid death, signs of chronic extrapulmonary dissemination are more usual.

The skin is an area that is frequently involved (chronic cutaneous blastomycosis). The face or forearms and hands are common sites for skin lesions. These are slow, spreading, verrucose plaques with central scarring. The initial lesion is often a dermal nodule. Many such cases have underlying pulmonary consolidation, or cavities. The diagnosis is established by biopsy and culture. Bone deposits in the form of lytic lesions, and involvement of the genitourinary tract, particularly the epididymis, are also seen in chronic disseminated blastomycosis. Unlike tuberculosis, the kidneys are often spared.

In slowly progressive forms of blastomycosis, itraconazole (200 to 400 mg daily) has proved to be very effective. Alternatively, amphotericin B can be given intravenously and is indicated where there is rapidly progressive disease.

Further reading—blastomycosis

Emerson PA, Higgins E, Branfoot A (1984). North American blastomycosis in Africans. *British Journal of Diseases of the Chest* **78**, 286–91.

Sarosi GA, Davies SF (1979). Blastomycosis. *American Reviews of Respiratory Diseases* **120**, 911–38.

Coccidioidomycosis

See Chapter 7.12.3.

Paracoccidioidomycosis

See Chapter 7.12.4.

Systemic sporotrichosis

In addition to causing cutaneous disease, *Sporothrix schenckii* may be responsible for a systemic mycosis. The infection is rare and has been mainly reported from the United States. Involvement may be confined to a single site such as a lung or a joint, or it may be multifocal. Cavitation in the lung associated with weight loss and pyrexia is probably the most common variety of systemic sporotrichosis. Unlike cutaneous forms of the disease, systemic sporotrichosis responds poorly to potassium iodide, and amphotericin B is the treatment of choice.

Rare systemic infections

These include pulmonary invasion by *Geotrichum candidum* (geotrichosis) and adiaspiromycosis, a respiratory infection caused by *Emmonsia crescens* or *Emmonsia parva*. Isolated examples of human disease caused by fungi are consistently reported and almost always occur in the immunosuppressed host. In these patients many fungi that are normally saprophytes in the environment may invade and cause disease.

Systemic mycoses caused by opportunistic fungi

The opportunistic mycoses are a worldwide problem, although fortunately rare in most countries. In recent years they have been recognized more frequently with the increase in transplantations of organs such as heart or bone marrow and in the more effective but immunocompromising regimes of cancer chemotherapy. Opportunistic invasion by organisms such as *Candida* or zygomycetes (*Mucor*, *Absidia*) may also occur in cases of malnutrition. One of the recent trends in the management of the patient with neutropenia has been the emergence of new pathogens such as non-*albicans* species of *Candida* or other organisms such as *Fusarium*, *Trichosporon*, or *Hansenula* species.

The opportunists present particular problems in diagnosis and management. Because many of the organisms are normally saprophytic, it has to be positively established that they have assumed an invasive role. Mere isolation may not provide sufficient evidence and in some instances low titres of antibody may be present

even in normal hosts. The significance of various laboratory tests in these infections is shown in Table 3. Treatment is also difficult and it is important in most cases to attempt to reverse the process that led to the establishment of the infection.

Systemic candidosis

Aetiology

In addition to their role in superficial infections, yeasts of the genus *Candida* may also cause invasive systemic disease. The clinical forms described range from bloodstream isolation or candidaemia to disseminated invasive disease, sometimes with involvement of a single organ, site, or body cavity (deep focal candidosis) as may occur in peritonitis or meningitis. Urinary tract infections may also be caused by *Candida* species.

The factors underlying systemic *Candida* infections are shown in Table 4. All these factors are important in disrupting the balance by which *Candida* is maintained as a saprophyte. Intravenous or central venous pressure lines may serve as a portal of entry or as a nidus for circulating yeasts in a candidaemia. Antibiotic therapy may upset the balance by inhibiting a potentially competitive bacterial flora.

Candida albicans is the most common species involved but other species may be isolated, particularly in cases of endocarditis, for example *Candida parapsilosis*. *Candida tropicalis* has been implicated in infections of patients with neutropenia. These non- *albicans Candida* species are now more frequent causes of systemic infection and are important to recognize as their antifungal susceptibility may differ from *C. albicans*. Portals of entry include the gastrointestinal tract (common), skin, and urinary tract (rare). However, superficial candidosis or saprophytic colonization of mouth, skin, or airways may also occur in compromised patients and does not necessarily indicate systemic invasion.

Epidemiology

Systemic infections caused by *Candida* species are worldwide in distribution. However, they are particularly associated with a number of predisposing factors such as neutropenia, antibiotic usage, indwelling lines, and abdominal surgery.

Clinical features

Candidaemia

The isolation of *Candida* in blood culture may be linked to any of the factors listed in Table 4. Common predisposing features are the presence of intravenous lines, previous surgery (mainly gastrointestinal), antibiotic therapy, hepatic failure, or neutropenia. Patients develop a swinging fever and feel generally unwell. Clinical shock may occur.

Some such cases resolve following removal of predisposing factors, particularly the intravenous lines. Generally, however, all such patients receive treatment and a careful investigation should be made to identify the presence of established invasive disease. Other sites should be searched for evidence of infection; for example urine by culture or the presence of white cells. Signs of muscle invasion (tenderness) or metastatic skin nodules should be excluded (Plate 9). Other signs of invasion include the development of new cardiac murmurs or of soft, white, retinal plaques caused by *Candida*. Persistently positive blood cultures or serum *Candida* antigen levels or high antibody titres may also indicate possible deep invasion.

Disseminated candidosis

Although multiorgan invasive candidosis may follow candidaemia, at least 50 per cent of disseminated infections develop in patients without initially positive blood cultures. The features of some forms of invasive candidosis are listed above (under Candidaemia). Although *Candida* may be isolated from the sputum in these patients, there is rarely objective evidence of lung invasion. Moreover, there is no radiological appearance that is diagnostic of pulmonary candidosis and, indeed, chest radiographs may even appear normal. General localizing signs may be a late feature of disseminated candidosis.

Laboratory diagnosis of disseminated candidosis

The diagnosis may be made by culture and repeated attempts to isolate should be made where cultures are initially negative. Numerous techniques have been used to detect antibody or antigen in disseminated candidosis. However, in many patients, particularly those with neutropenia, it may not be possible to confirm the diagnosis using laboratory tests and treatment is often initiated on the basis of clinical suspicion (empirical therapy) as the risk of delaying antifungal therapy is great.

By themselves, positive cultures, particularly from sputum, or the presence of antibodies do not necessarily prove the existence of deep-seated candidosis. A positive isolation may simply indicate the presence of colonization and normal individuals may have low titres of antibody to *Candida*. If there is a readily accessible lesion from which to take a biopsy, such as a skin nodule or even a pulmonary infiltrate, this may provide the best evidence of invasion, although such procedures may carry their own risk (Plate 8).

Treatment of disseminated candidosis

Untreated disseminated candidosis is normally progressive and fatal. The signs must be separated from, for instance, bacterial septicaemia, which may coexist with the *Candida* infection.

The treatment of invasive candidosis is intravenous amphotericin B or intravenous or oral fluconazole given until there is a clinical and mycological response. This may take between 2 and 20 weeks depending on the site of infection and the underlying state of the patient. Fluconazole is usually used in infections where the patient is not neutropenic. Lipid-associated forms of amphotericin B are also useful and carry a lower risk of renal impairment. An alternative approach is to add flucytosine in doses of 150 to 200 mg/kg body weight daily to amphotericin B in serious infections or where cure may be hampered by poor penetration of amphotericin B, such as in the eye.

Deep focal candidosis

Candida infections in the peritoneum or meninges most often follow direct implantation after dialysis or surgery. Alternatively, secondary invasion from the middle ear or a perforated bowel is also possible. The signs and symptoms are similar to bacterial meningitis or peritonitis but *Candida* is isolated. Sometimes these infections clear spontaneously, but normally treatment is instituted with fluconazole, which penetrates areas such as peritoneum, or amphotericin B.

Candida endocarditis

Invasion of heart valves, mainly the mitral or aortic valves, most commonly follows homograft replacement, but it may occur also in patients with neutropenia or drug addicts. The symptoms are similar to bacterial endocarditis. However, *Candida* vegetations may reach considerable size. Embolic phenomena may involve obstruction of large vessels including the femoral artery or large cerebral vessels. The detection of large vegetations using an echo scanning device, particularly in cases with negative blood cultures, should raise the possibility of fungal endocarditis. Blood cultures are usually positive at some stage in the illness but repeated sampling may be necessary. High antibody titres are usually seen in such cases and serological tests are therefore of considerable value.

Untreated *Candida* endocarditis is uniformly fatal. There is also a high mortality associated with cases in which early surgical intervention is precipitated by impending heart failure. Normally, treatment consists of amphotericin B given intravenously and, where possible, valve replacement. There is no evidence to suggest that the addition of flucytosine to the regimen increases the effectiveness of treatment. However, the relapse rate is high and combination therapy may therefore be a reasonable approach on theoretical grounds.

Urinary tract candidosis

Candida species may be isolated from the urine, particularly in conditions associated with urinary stasis such as neurogenic bladder or where there is an indwelling catheter. Maturity-onset diabetes mellitus is another predisposing factor. There is no value in using the presence of pyuria or quantitative yeast-colony counts to assess the significance of infection. Treatment is normally given where there are symptoms such as dysuria or frequency or where there is a potential risk of invasion such as in immunosuppressed patients. Fluconazole is very useful in these patients as urinary levels are above inhibitory concentrations.

Further reading—opportunistic systemic mycoses

Krcmery V, Krupova I, Denning DW (1999). Invasive yeast infections other than *Candida* spp. in acute leukaemia. *Journal of Hospital Infection* **41**, 181–94.

Reiss E *et al.* (1998). Molecular diagnosis and epidemiology of fungal infections. *Journal of Medical Mycology* **36**(Suppl 1), 249–57.

Wingard JR (1999). Fungal infections after bone marrow transplant. *Biology of Blood and Marrow Transplantation* **5**, 55–68.

Aspergillosis (see Section 8.4)

Cryptococcosis

Aetiology

Cryptococcosis is a systemic infection caused by *Cryptococcus neoformans*. Its most common clinical feature is meningitis, but pulmonary, cutaneous, and widely disseminated forms of the infection are also recognized. There are two varieties of *C. neoformans* called *C. neoformans neoformans* and *C. neoformans gattii*. They differ in their geographical range and ecology. The *neoformans* variety is the most common in patients with AIDS.

C. neoformans neoformans is a yeast that can be isolated from the environment, although it is most often found in pigeon excreta. Its growth from soil appears to be enhanced by certain nitrogenous compounds, such as creatinine in the pigeon droppings. The birds are not infected, although their crops may be heavily colonized. Very large numbers of organisms (1×10^7 yeasts/g of droppings) may be found in densely populated urban areas. *C. neoformans gattii* has been detected in leaf and bark debris of certain eucalyptus species.

The portal of entry is usually the lung, from where the organism spreads to involve other organs or sites such as the meninges. Although many isolates from natural sources have small cells, one sequel to tissue invasion is the development of a large, mucoid capsule *in vivo*, a feature that may confer some protection to the organism. Infections with *C. neoformans* are seen in both normal and immunocompromised hosts. The main underlying processes are sarcoidosis, Hodgkin's disease, collagen disease, carcinoma, and the administration of systemic corticosteroid therapy, but AIDS is the commonest predisposition.

Epidemiology

Cryptococcosis has been recorded from most countries, although it is most prevalent in the United States and Australia. Before the AIDS epidemic in the United States approximately 50 per cent of cases were said to occur in normal persons. By contrast, in the United Kingdom, 85 per cent of cases were found in patients with underlying disorders. There is no skin-test reagent widely available, but some pilot studies in the United States suggest that workers exposed to the organism (for example in laboratories) are more likely than other groups to have a positive skin test without any overt sign of infection. It is probable, therefore, that there is an asymptomatic form of cryptococcosis (compare histoplasmosis). Additional evidence for the existence of subclinical infection is provided by the repeated isolation of *C. neoformans* in sputum from individuals without evidence of disease.

Clinical features

Pulmonary cryptococcosis

Acute or subacute respiratory disease caused by *C. neoformans* is seen in both HIV-positive and healthy individuals. The disease consists of a chest infection with fever and cough and scattered, often well-circumscribed, areas of pulmonary infiltration seen on radiographs. Pleural involvement can occur and sometimes massive pulmonary infiltrates may occur. Before the advent of the azoles, in some patients the whole process resolved without treatment, although it is probably advisable to give fluconazole to those with isolated pulmonary disease. More often, lung lesions accompany disseminated cryptococcosis or cryptococcal meningitis and the treatment is discussed below. The laboratory diagnosis is made by biopsy or culture. Isolated cryptococcal granulomas (cryptococcoma) may present as coin lesions and are removed surgically to exclude carcinoma. Once the correct diagnosis is made, many workers advise a short course of amphotericin B or fluconazole as there is a small risk of dissemination to other organs following surgery.

Disseminated cryptococcosis

The best-recognized form of extrapulmonary cryptococcosis is meningitis. This may present with signs of acute meningism. However, more usually the features are less specific. Pyrexia, headache, and mental changes such as confusion or drowsiness occur. The mental changes probably follow the development of hydrocephalus. Blurring of vision and papilloedema may also occur. Cranial nerve involvement is less common. Patients with AIDS often present with widely disseminated disease. The signs of meningeal involvement may be very subtle and the infection has often spread to other sites such as liver and spleen as well as skin.

The cerebrospinal fluid shows pleocytosis that is highly variable. Often there are excessive numbers of lymphocytes, but sometimes polymorphonuclear leucocytes abound. In some cases only small numbers of white cells (4 to 10/ml) are seen. Characteristically, but not invariably, the glucose concentration falls and protein rises. Cryptococci can be seen in some cases in an India ink or nigrosin preparation, which is used to highlight the capsule. A spun sediment is best for this purpose. The organism can also be cultured from the cerebrospinal fluid. The latex test for antigen is usually positive for cerebrospinal fluid, but on rare occasions this is negative. The antigen titre has both diagnostic and prognostic value. Initial high (> 100) titres are likely to correlate with relapse following therapy and with a poor prognosis. In patients with AIDS, antigen titres over 1:1000 convey poor prognosis and blood cultures are often positive. Extrameningeal disease should be looked for by sputum or urine culture and serology in patients presenting with meningitis.

Other sites

Cryptococci may disseminate to other sites including liver and spleen, kidney, skin, or bone. Infection in skin and bone are most often seen in patients with sarcoidisis. In every case, underlying deep disseminated lesions (such as meningitis) may be found. The methods of diagnosis and treatment are similar to those seen with meningitis. Only a small proportion of cases with solitary disseminated lesions of cryptococcosis, such as bone or skin, may have detectable antigen (15 to 30 per cent), and this may occur late in the course of therapy. In patients with AIDS the organism spreads widely involving bone marrow, liver, and spleen as well as other sites. Positive blood cultures are not uncommon. The serum antigen titres are often very high, for example over 500, and may not return to normal even during antifungal treatment.

It is important in all cases where cryptococcosis presents with lesions in an extrameningeal site to exclude occult meningitis by lumbar puncture.

Treatment

In the patient without AIDS the combination of flucytosine (150 to 180 mg/kg daily) and intravenous amphotericin B (0.3 to 0.6 mg/kg daily) is the most widely used treatment. It is possible to induce recovery with this approach and treatment is generally continued for at least 6 weeks or longer if necessary. The clinical response and antigen levels are useful for monitoring progress.

The situation is different in patients with AIDS because in patients not receiving combination antiretrovirals it is impossible to achieve complete recovery. The object of therapy is to induce the most rapid remission possible, followed by long-term suppressive therapy. There are various regimens used for induction of remission. The use of amphotericin B with or without flucytosine is favoured by many. This is given for 2 weeks and is then followed by indefinite treatment with fluconazole to prevent

relapse. Itraconazole is an alternative. In patients on highly active antiretroviral therapy (MAART) it may be possible to discontinue treatment but guidelines are awaited.

Further reading—cryptococcosis

Clark RA *et al.* (1990). Spectrum of *Cryptococcus neoformans* infection in 68 patients infected with acquired immunodeficiency virus. *Reviews of Infectious Diseases* **12**, 768–77.

Seaton A *et al.* (1996). Exposure to *Cryptococcus neoformans* var *gattii*—a seroepidemiological study. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **90**, 508–12.

Stevens DA (1990). Fungal infections in AIDS patients. *British Journal of Clinical Practice* **44**(Suppl 71), 11–22.

Invasive zygomycosis (mucormycosis, phycomycosis)

Aetiology

Invasive disease caused by mucor-like (zygomycete) fungi is rare. In the compromised host it may lead to paranasal destruction, necrotic lung or skin lesions, and disseminated disease.

The causative organisms commonly belong to three genera, *Absidia*, *Rhizopus*, and *Rhizomucor*. More rarely other organisms such as *Cunninghamella* or *Saksenaea* have been implicated. Most of the agents are associated with decaying vegetable matter and are common airborne moulds. The route of infection is highly variable: they may invade via the lungs, paranasal sinuses, gastrointestinal tract, or damaged skin. The predisposing illness may in some way determine the site of clinical invasion. Underlying factors include diabetic ketoacidosis (rhinocerebral involvement), leukaemia and immunosuppressive therapy (lung and disseminated infection), malnutrition (gastrointestinal infection), and burns or wounds (cutaneous invasion). These patterns are not always strictly followed.

Epidemiology

Invasive zygomycosis is rare but has a worldwide distribution. Its invasive nature, particularly the tendency to involve blood vessels and its selection of compromised hosts, distinguishes this form of infection from subcutaneous zygomycosis, which is also caused by zygomycete species.

Clinical features

The most characteristic features of this type of infection are the extensive necrosis and infarction that may follow blood vessel invasion leading to thrombosis. A similar type of invasion may occur with invasive aspergillosis, but is usually less prominent. Invasive zygomycosis follows a number of different patterns.

The infection may initially localize in one of several sites. The most common is in the paranasal sinuses and this is most often seen in diabetic patients with ketoacidosis. The patient presents with fever and unilateral facial pain. Subsequently, there may be facial swelling with nasal obstruction and proptosis. There may be invasion into the orbit leading to blindness, into the brain, and the palate. Palatal ulceration should be searched for. Widespread dissemination with infarction of major organs or limbs may occur subsequently. A similar pattern of invasion of surgical wounds or burns may occur and has on occasions been associated with contamination of dressing packs. Infections are initially localized causing extensive necrosis around the original wound. Gastrointestinal invasion may be heralded by perforation of viscera, and diarrhoea or haemorrhage.

Alternatively, a patient may present with established pulmonary or widespread dissemination. Such patients are usually leukaemic or are severely immunosuppressed. Neutropenia is often seen.

Once infection has spread beyond the original site, invasive zygomycosis is almost invariably fatal with or without treatment.

Laboratory diagnosis

The diagnosis is suggested by the combination of infection and extensive infarction, particularly if it occurs in any of the sites mentioned. The organisms may be difficult to culture even from biopsy and histology is often the quickest way of establishing the diagnosis. Serology is frequently negative.

Treatment

Treatment should be initiated as soon as possible and extensive surgical debridement combined with intravenous amphotericin B in maximum daily dosage offers the best chance of success. Local instillations of amphotericin B may also be used where appropriate (such as nasal sinuses). Some physicians also recommend anticoagulation with heparin to forestall thrombosis. Despite therapy, the mortality remains high. Liposomal amphotericin B also has been used with some success is cases of mucormycosis.

Further reading—zygomycosis

Nenoff P *et al.* (1998). Rhinocerebral zygomycosis following bone marrow transplantation in chronic myelogenous leukaemia. Report of a case and review of the literature. *Mycoses* **41**, 365–72.

Rhinosporidiosis

Rhinosporidiosis is an infection found in India, Sri Lanka, parts of East Africa, and South America. It is characterized by polypoid growth from the nose or conjunctiva. The causative organism can be demonstrated in tissue and consists of aggregates of large sporangia containing spores in various phases of development. However, they have never been successfully cultured and their fungal nature has only been assumed from their morphological appearance in histology.

The treatment is surgical excision.

Otomycosis and oculomycosis

External otitis is often multifactorial, but in some cases dense fungal colonization can contribute to the picture. In severe cases, the external ear may be plugged by a dense mat of mycelium. *Aspergillus* species are the most common organisms cultured, particularly *A. niger*, but *Candida*, *Penicillium*, and *Mucor* may all contribute. Intensive ear toilet may eradicate the infection without recourse to antifungal agents.

Infections of the eye, particularly the cornea, caused by fungi (oculomycosis) are rare. They often follow penetrating injuries to the globe or contamination of lacerations. An opacity develops within the cornea with associated pain and chemosis. An exudate is usually present in the aqueous humour. Prompt treatment with intensive topical instillation of drugs containing an antifungal drug such as miconazole or econazole is necessary every 2 to 4 h. Perforation of the eye may occur in advanced cases.

Approaches to management of fungal infections

Antifungal agents can be considered in four main groups: the polyenes, azoles, morpholines, and allylamines, and an assortment of drugs of specific activity that are not related.

The polyene antifungals are macrolide substances derived originally from species of *Streptomyces*. They include amphotericin B, natamycin, and nystatin. More recent additions to this group are partricin and mepartricin. Amphotericin B is the only one widely used as a parenterally administered drug. Nystatin and natamycin are purely topical. Amphotericin B is metabolized in the liver with low penetration of body cavities, cerebrospinal fluid, and urine. The polyenes have broad activity

against a wide range of fungi. The mode of action of the polyenes appears to involve inhibition of sterol synthesis in the fungal cell membrane.

The combination of an amphotericin B with a lipid, for instance a liposome, has been proposed as a means of reducing the nephrotoxicity of this drug. Three commercial lipid amphotericins are available: AmBisome (a true liposome), amphotericin B lipid complex—ABLC or Abelcet (a ribbon-like lipid binding amphotericin B), and amphotericin B colloidal dispersion (ABCD) (a dispersion of lipid discs).

The imidazoles are synthetic antifungal agents. They include miconazole, clotrimazole, econazole, isoconazole, ketoconazole, tioconazole, and bifonazole. The triazole series contains two potent oral agents, fluconazole and itraconazole. A third, voriconazole, is in clinical trial. Most are used topically except for ketoconazole (oral), itraconazole (oral), and miconazole (intravenous). These are metabolized in the liver and, like amphotericin B, affect fungal cell-membrane synthesis and penetrate cerebrospinal fluid and urine in low concentrations. The imidazoles have a broad spectrum of activity against many fungi, although neither miconazole nor ketoconazole are useful for aspergillus infections. By contrast, itraconazole is active *in vitro* against aspergilli. Fluconazole is less active against moulds and there are instances of both primary (*Candida krusei*, *C. glabrata*) and secondary resistance to this compound. The allylamines such as terbinafine are primarily active against superficial fungi, but *in vitro* appear to have fungicidal activity at low concentrations.

Other antifungal drugs include flucytosine, which is a synthetic pyrimidine analogue. Given either intravenously or orally it is mainly useful for chromomycosis and certain yeast infections. Drug resistance is a major problem with flucytosine, particularly with cryptococcus. The drug shows a number of modes of action including disruption of RNA transcription following uptake by the cell. Griseofulvin is derived from a species of *Penicillium*. It can be given orally and is only useful against dermatophytes. It is best absorbed when given with a meal and selectively accumulates in stratum corneum in concentrations approximately 10 times greater than serum levels. Griseofulvin acts by inhibiting intracellular microtubule formation. There are a large number of unrelated antifungal drugs, such as tolnaftate, haloprogin, and chlorphenesin, that are only used topically.

Management of superficial infections

Specific details of therapy are included under the separate diseases. Benzoic acid compound (Whitfield's ointment), which contains 2 per cent salicylic acid and 2 per cent benzoic acid, acts as a keratolytic agent by causing exfoliation of the superficial layers of the stratum corneum. Other topical agents with only eak antifungal activity include gentian violet (candidosis or dermatophytosis), Castellani's paint, which contains magenta and resorcinol (candidosis or dermatophytosis), and brilliant green (dermatophytosis). Two per cent selenium sulphide remains a highly effective method of treating pityriasis versicolor by application once daily for 2 weeks.

The more specific antifungals such as the polyenes, amphotericin B, nystatin, and natamycin (candidosis) or the imidazoles (candidosis, dermatophytosis, and pityriasis versicolor) are highly effective and probably quicker, although more expensive, than the keratolytics or dyes. Local irritancy can be a problem particularly with Whitfield's ointment, which is usually given as a half-strength preparation. Allergic contact dermatitis is rare but has been recorded from some imidazoles (miconazole, clotrimazole, tioconazole) and tolnaftate. Topical terbinafine is highly active in tinea pedis with cures being effected with less than 1 week of therapy.

Terbinafine or itraconazole are more effective in many forms of dermatophytosis requiring oral therapy than griseofulvin. In onychomycosis they are preferred. Terbinafine has occasional side-effects, mainly related to gastrointestinal intolerance, although it may also cause transient loss of taste. It is given in daily doses of 250 mg. Itraconazole is usually given in 'pulses', for example 200 mg twice daily for one week monthly. Itraconazole likewise can cause gastrointestinal discomfort and nausea. Both drugs rarely cause hepatic injury, with a frequency of less than 1 in 70 000–120 000. This is in contrast with ketoconazole, which also causes hepatitis but in around 1 in 8000 cases. Liver function tests should be monitored if ketoconazole is used extensively over any length of time. In high doses, ketoconazole may block human androgen biosynthesis causing side-effects such as gynaecomastia. Fluconazole is also effective in dermatophytosis and is given in weekly doses of 150 to 300 mg. Griseofulvin is still the principal treatment for tinea capitis (10 to 20 mg/kg per day).

In onychomycosis caused by dermatophytes both terbinafine and itraconazole lead to remission of toe-nail infections in only 3 months. Terbinafine is used on a daily basis, whereas itraconazole is given in a pulsed regimen, 200 mg twice daily for 1 week every month for 3 to 4 months. There is one study which shows better responses with terbinafine for toe-nail disease. Amorolfine, a morpholine drug, is used in the topical treatment of nail disease where there is less than complete involvement of the nails. It can be given together with other drugs, such as terbinafine.

Management of deep mycoses

There are very few drugs that are effective in systemic fungal infections and those that are used should always be accompanied by supportive measures and, if possible, an attempt to eliminate any predisposing conditions. For instance, if their condition permits, patients who have developed a candidaemia while a central venous line is in place should be managed by removal of the line. However, fluconazole is also usually given as well. In the patient with neutropenia, a positive blood culture would be regarded as evidence of dissemination and antifungal therapy would be required.

Amphotericin B is given intravenously in a 5 per cent dextrose infusion not containing additional drugs, if possible. A test dose of 1 to 5 mg is given over 2 h and this is followed by gradually increasing doses over the next 3 to 9 days to the normal maximum of 0.6 to 1.0 mg/kg body weight daily depending on the infection. In some cases this slow approach may help the patient to tolerate the drug better or may define the dose at which side-effects such as pyrexia start. In severely ill patients, half of the full dose may be given 4 h after a test dose of 5 mg, usually under hydrocortisone cover. The full dose is given 24 h later. Side-effects include thrombophlebitis, nausea, hypotension, and pyrexia. Renal clearance may fall in the initial period but this usually returns to normal after a temporary halt in therapy. More permanent renal tubular damage may follow a total dose of 4 g or more. Amphotericin B does not penetrate urine, cerebrospinal fluid, or peritoneal fluid in significant concentrations. Local instillations (such as the peritoneum) can be used, but can be highly irritant. Amphotericin B is normally given until clinical or mycological cure is induced. This is often difficult to judge accurately and in many of the mycoses caused by the systemic pathogens a course of at least 2 g is often used on an empirical basis. In the opportunistic infections, lower total doses are probably effective and the length of treatment should depend on the clinician's judgement.

This approach is not necessary with the lipid-associated amphotericin B formulations, which can be given without the slow build-up. The initial dose is usually 1 mg/kg but standard daily doses of 3 mg/kg are common. Patients are less likely to develop renal impairment although it can occur. There have been a few clinical trials comparing these formulations with amphotericin B and these show equal efficacy with less toxicity; however, these formulations are expensive. The main lipid-associated formulations are given above.

The azole drugs are also used in systemic mycoses. Fluconazole is given in systemic candidosis, urinary tract infections, and as a long-term suppressive, in addition to primary therapy, in cryptococcosis in patients with AIDS. Side-effects are uncommon, although it can cause nausea and vomiting. Fluconazole can be given orally or intravenously. It penetrates urine in effective concentrations. Its daily dosage varies from 100 to 200 mg for oropharyngeal infections to 600 to 800 mg for disseminated candidosis. It is highly active in *Candida* infections. It can also be used in some endemic mycoses such as histoplasmosis. Resistance to fluconazole has mainly been recorded with oropharyngeal candidosis, principally in HIV-positive patients, although it can occur with other *Candida* infections; *C. krusei* and *C. glabrata*, for instance, are often primarily resistant to this drug.

Itraconazole has been evaluated in a variety of systemic mycoses from aspergillosis to cryptococcosis. Its active range includes histoplasmosis, sporotrichosis, chromoblastomycosis, blastomycosis, coccidioidomycosis, and paracoccidioidomycosis. Itraconazole is used as an oral preparation, but a new intravenous formulation is now available. Oral absorption is often defective in individuals with AIDS and patients after bone marrow transplantion and in these groups the mean daily dosage is doubled (200 mg). An itraconazole suspension is also available for treatment of oral infections.

Flucytosine (5-fluorocytosine) is an effective oral and intravenous antifungal agent that is primarily active against yeasts such as *Candida* and *Cryptococcus*. It enters urine, cerebrospinal fluid, and peritoneal fluid. Its excretion is reduced in renal failure and the daily dose should be reduced accordingly and blood levels monitored. The main disadvantage of flucytosine is the development of either primary or secondary drug resistance in a significant number of isolates, and when given in toxic doses it may cause bone marrow depression. The serum level should not be allowed to rise above 100 to 120 µg/ml.

Combination amphotericin B and flucytosine therapy may offer an alternative but effective method of treatment. Theoretically, as the drugs synergize, the dose of amphotericin B may be reduced. In cryptococcal meningitis, combination therapy using a dose of 0.3 to 0.6 mg/kg body weight of amphotericin B with the normal dose of flucytosine is more effective at sterilizing the cerebrospinal fluid and preventing relapse. In other forms of systemic infection such as candidosis there is little evidence that it is more effective than amphotericin B alone, although this may be the case. Combinations of other drugs have not been evaluated *in vivo*.

The use of leucocyte growth factors has been reported to improve the recovery from fungal infections. The most effective combination has been a mixture of granulocyte and granulocyte–monocyte colony-stimulating factors. Further studies of these compounds in patients with neutropenia are warranted.

Further reading—therapy

Bohme A, Karthaus M, Hoelzer D (1999). Antifungal prophylaxis in neutropenic patients with hematologic malignancies: is there a real benefit? *Chemotherapy* **45**, 224–32.

Elweski B, ed. (1996). *Cutaneous fungal infections.* Marcel Dekker, New York.

Medoff G, Kobayashi GA (1980). The polyenes. In: Speller DCE, ed. *Antifungal chemotherapy*, pp 3–34. Wiley, Chichester.

Root RK, Dale DC (1999). Granulocyte colony-stimulating factor and granulocyte–macrophage colony-stimulating factor: comparisons and potential for use in the treatment of infections in non-neutropenic patients. *Journal of Infectious Diseases* **179**(Suppl 2), S342–52.

Vanden Bossche H *et al.* (1998). Antifungal drug resistance in pathogenic fungi. *Medical Mycology* **36**(Suppl 1), 119–28.

7.12.2 Cryptococcosis

William G. Powderly

Aetiology and epidemiology

Infection with the fungus *Cryptococcus neoformans* occurs mainly in patients with impaired cell-mediated immunity. It is the most common, systemic, fungal infection in patients infected with human immunodeficiency virus (HIV), and is also seen as a complication of solid organ transplantation, lymphoma, and corticosteroid therapy. *C. neoformans* is found world-wide as a soil organism; it is an encapsulated yeast measuring 4 to 6 µm with a surrounding polysaccharide capsule ranging in size from 1 to over 30 µm. Two varieties exist, distinguishable by serology—*C. neoformans* var. *neoformans* (serotypes A and D) and *C. neoformans* var. *gattii* (serotypes B and C). Virtually all HIV-associated infection is caused by *C. neoformans* var. *neoformans*. About 5 per cent of HIV-infected patients in the Western World develop disseminated cryptococcosis; the disease is more prevalent in sub-Saharan Africa and southeast Asia. *C. neoformans* var. *gattii* infection is more common in tropical and subtropical areas (Australia, New Guinea, the Philippines) in apparently immunocompetent people. It has rarely been reported in HIV-immunosuppressed patients.

The exact mechanism of infection is unknown. It is assumed that transmission occurs via inhalation of the organism leading to colonization of the airways and subsequent respiratory infection. Throughout the world, the excreta of birds such as pigeons is the richest environmental source of *C. neoformans* var. *neoformans*. The ecological association of *C. neoformans* var. *gattii* is with red river and forest river gum trees (*Eucalyptus camaldulensis* and *E. tereticonrnis*) and it has been suggested that infective basidiospores are released at flowering.

In the case of *C. neoformans* var. *neoformans*, the absence of an intact cell-mediated response results in ineffective clearance with subsequent dissemination. The polysaccharide capsule, composed mainly of glucuronoxylomannan, is thought to be its primary virulence factor. It is unclear whether cryptococcal infection in immunocompromised patients represents acute primary infection or reactivation of previously dormant disease.

Clinical features

The most common presentation of cryptococcosis is a subacute meningitis or meningoencephalitis with fever, malaise, headache, and altered behaviour and level of consciousness. Symptoms are usually present for 2 to 4 weeks before diagnosis. Classic meningeal symptoms and signs (such as neck stiffness or photophobia) occur in only about a quarter to a third of patients. Papilloedema and cranial nerve palsies (especially VI and VII) are common. Patients may present with encephalopathic symptoms such as lethargy, altered mentation, personality changes, and memory loss. Analysis of the cerebrospinal fluid (CSF) usually shows a mildly elevated serum protein, normal or slightly low glucose, and a lymphocytic pleocytosis. India ink staining of the CSF will usually reveal the yeast. Cryptococcal antigen is almost invariably detectable in the CSF. The opening pressure in the CSF is elevated in a majority of patients.

Infection with *C. neoformans* can involve sites other than the meninges. Isolated pulmonary disease has been well described and usually presents as a solitary nodule in the absence of other symptoms. Cryptococcal pneumonia also occurs. In immunocompromised patients, especially those with AIDS, subsequent dissemination is common but presentations such as cough or dyspnoea and abnormal chest radiographs may be the initial finding. Many patients have positive blood cultures. Skin involvement is common; several types of skin lesion have been described but the most common form is that resembling molluscum contagiosum. Osteolytic bone lesions and prostatic involvement have also been described.

In New Guinea, *C. neoformans* var. *gattii* is the commonest cause of chronic meningitis. Immunocompetent people are affected. Compared to *C. neoformans* var. *neoformans* meningitis in AIDS patients, victims of var. *gattii* have more aggressive retinal involvement with papilloedema and haemorrhagic papillitis in more than a half, leading to blindness in one-third of survivors.

Diagnosis

The latex agglutination test for cryptococcal polysaccharide antigen in the serum is highly sensitive and specific in the diagnosis of infection with *C. neoformans* and a positive serum cryptococcal antigen titre of greater than 1:8 is presumptive evidence of cryptococcal infection. Such patients should be evaluated for possible meningeal involvement. Culture of *C. neoformans* from any body site should also be regarded as significant and is an indication for further evaluation and initiation of therapy.

Treatment

Management of patients with cryptococcal infection depends on the extent of the disease and the immune status of the patient. The finding of a solitary pulmonary nodule in a normal host may not need treatment, provided patients have careful follow-up. Fluconazole, 200 to 400 mg/day can be given for 3 to 6 months in most patients with localized pulmonary disease. Extrapulmonary disease is generally managed in the same way as meningitis. In patients who are not known to be immunosuppressed, a search for underlying problems should be initiated. An HIV antibody test should be performed as cryptococcal meningitis may be the initial AIDS-defining event. Additionally, a CD4 lymphocyte count should be considered, as cryptococcal infection has been described as one of the manifestations of so-called 'isolated CD4 T-lymphocytopenia'.

Untreated, cryptococcal meningitis is fatal. In patients with AIDS, amphotericin B (0.7 mg/kg intravenously) given for 2 weeks followed by fluconazole 400 mg orally for a further 8 weeks is associated with the best outcome to date in prospective trials, with a mortality of less than 10 per cent and a mycological response of approximately 70 per cent. This regimen is also reasonable for treatment of meningitis in other circumstances. Concomitant use of flucytosine (100 mg/kg per day in four divided doses) with amphotericin B may be considered. In patients with AIDS, it does not improve immediate outcome but may decrease the risk of relapse. In other hosts, more prolonged use (4 to 6 weeks) of amphotericin B and flucytosine may be curative but is also toxic. The combination of fluconazole (400 to 800 mg/day) with flucytosine and liposomal formulations of amphotericin B are options for patients unable to tolerate the usual formulation of amphotericin B.

Clinical deterioration in patients with meningitis may be due to cerebral oedema, which may be diagnosed by a raised opening pressure of the CSF. All patients with cryptococcal meningitis should have the opening pressure measured when a lumbar puncture is performed, and if the opening pressure is high (>25 cm of water) pressure should be reduced by repeated lumbar punctures, a lumbar drain, or a shunt. In var. *gattii* meningitis, corticosteroid treatment is helpful in reducing intracranial pressure and reducing retinal damage.

Cryptococcal meningitis in AIDS requires life-long suppressive therapy unless the immunosuppression is reversed. In other immunocompromised patients, suppressive treatment for 6 to 12 months may be given. Fluconazole, 200 mg daily, is the suppressive treatment of choice. Fluconazole, in dosages ranging from 400 mg weekly to 200 mg daily, and itraconazole, 100 mg twice daily, are very effective in preventing invasive cryptococcal infections, especially in HIV-positive patients with CD4 counts less than 50 to 100 cells/mm^3. However, because of the relative infrequency of invasive fungal infections, antifungal prophylaxis does not prolong life and is not routinely recommended.

Further reading

Ellis DH, Pfeiffer TJ (1990). Ecology, lifecycle, and infections propagule of *Cryptococcus neoformans*. *Lancet* **36**, 923–5.

Graybill JR, *et al*. (2000). Diagnosis and management of increased intracranial pressure in patients with AIDS and cryptococcal meningitis. *Clinical Infectious Diseases* **30**, 47–54.

Lalloo D, Fisher D, Naraqi S, *et al*. (1994). Crytococcal meningitis (*C. neoformans* var *gattii*) leading to blindness in previously healthy Melanesian adults in Papua new Guinesa. *Quarterly Journal of Medicine* **87**, 343–9.

Mundy LM, Powderly WG (1997). Invasive fungal infections: Cryptococcosis. *Seminars in Respiratory and Critical Care Medicine* **18**, 249–57.

Van Der Horst CM, *et al*. (1997). Treatment of cryptococcal meningitis associated with the acquired immunodeficiency syndrome. *New England Journal of Medicine* **337**, 15–21.

7.12.3 Coccidioidomycosis

John R. Graybill

Aetiology and epidemiology

Coccidioides immitis was initially named by Gilchrest in 1986, because its round spherule form appears similar to coccidia, which are protozoans. Although traditionally associated with the southwest United States, the pathogen was initially discovered by Alejandro Posadas in Buenos Aires. Although the Argentine Pampas remains an important focus for *Coccidioides*, the desert southwest of the United States and northern Mexico is better known as the endemic zone. Fifty years ago, German prisoners of war who were held in Arizona developed coccidioidomycosis, and were mistakenly thought to be the victims of 'medical research'. This illness still plagues German airmen who train in Arizona.

Coccidioides immitis is dimorphic, and grows as a mycelium in desert soil where winters are mild. After spring rains, mycelia form myriads of small barrel-shaped arthroconidia. The mycelium disrupts readily, and the conidia are wafted for many kilometres. Earthquakes and desert sandstorms in California have caused large epidemics. Infection follows inhalation of just a few conidia. Over the course of several days the fungus converts to the pathognomonic spherule. This enlarges to as much as 60 µm in diameter, and is commonly filled with maturing endospores. Endospores are released after several days of growth; they are chemoattractive to neutrophils, which ingest but cannot kill them. Endospores enlarge into spherules and the growth cycle repeats itself. Although the spherule does not directly transmit disease, it is important to destroy all contaminated materials to prevent this most infectious of the endemic mycoses from converting back to the mycelium and causing secondary infections.

Clinical features

In no mycosis is the interplay of pathogen and host defences more important than coccidioidomycosis. The uncomplicated infection progresses through 3 to 6 weeks in the lungs, during which time protective cell-mediated immune responses develop. A strong immune response may cause arthralgias, fever, eosinophilia, and various rashes, including erythema multiforme or erythema nodosum ('desert fever'). A rise in IgM precipitin antibodies is diagnostic. Initial pulmonary infiltrates are later replaced by granulomas, which may condense to cicatrices or nodules. Cavitation of nodules may occur within weeks, or be a much later consequence of smouldering disease. The skin test to either spherulin or coccidioidin antigens commonly converts to positive. Illness resolves over weeks or months, to leave lifelong immunity. Low titres of IgG 'complement-fixation' antibody are generated 1 to 2 months after infection, and may persist for months.

Immunity and dissemination

Although immune suppression (by steroids or AIDS) is associated with dissemination, there are subtle factors of race (Blacks, American Indians, Filipinos) and gender (pregnancy) which also favour dissemination. The course of the disease may be strung out over years, with responses to treatment being followed by relapses. If the host is severely immune depressed, the course may evolve rapidly over weeks to persistent worsening pulmonary infiltrates and haematogenous dissemination to almost any tissue. Favourite locations are the bones (especially vertebral osteomyelitis), the skin (papular verrucous or proliferative), the lymph nodes, and the central nervous system. Coccidiodal meningitis presents insidiously (or rarely acutely after exposure) with headache, nausea, vomiting, seizures, and focal signs. Hydrocephalus and brain infarcts may develop. In general the association of skin or deep tissue abscesses draining pus with neutrophils and coccidioides indicates a poor host immune response, while granulomas showing spherules in Langerhans giant cells suggest better control of the organism.

Diagnosis

The IgG antibodies tend to rise to levels associated with the severity of disease (titres of $\geq$ 1:16 suggest worsening disease), and may remain elevated for many months. The erythrocyte sedimentation rate also rises. Coccidioidal meningitis is associated with lymphocytic and eosinophilic pleocytosis, hypoglychoracchia, and positive cerebrospinal fluid culture and/or serology for IgG.

The diagnosis of coccidioidomycosis may be made by culture or histopathology of tissues showing the characteristic spherule, or may be inferred from positive IgM or IgG serum (or IgG cerebrospinal fluid) antibody titres. The oganism is biphasic and converts readily to the mycelium in most culture media. Coccidioides is susceptible *in vitro* to most polyene and azole antifungals. Nevertheless, coccidioidomycosis is the most difficult of the endemic mycoses to treat.

Treatment

Clinical response is assessed using a scoring system developed by the Mycoses Study Group. This includes clinical symptoms and signs, cultures, radiographic changes, and serology. For non-meningeal disease, amphotericin B is reserved for those patients with the most fulminant courses, and even then there may be only a 70 per cent response rate. One troubling site of disease is vertebral osteomyelitis. This site is commonly refractory to medical therapy alone, and usually requires surgical stabilization of the spine for cure. For most patients the treatment of choice may be itraconazole, with a loading dose of 800 mg followed by 400 mg per day (capsules) until the illness resolves and then 9 to 12 months more for consolidation. A solution of itraconazole improves absorption but is less well tolerated than capsules. Resolution may require months or years, and occurs in fewer than 70 per cent of patients. Post-treatment relapses occur in 30 to 40 per cent of patients and may require repeated courses or higher doses. Fluconazole at 400 to 800 mg per day is an alternative for non-meningeal coccidioidomycosis, and is the drug of choice for meningeal coccidioidomycosis. Fluconazole allows highly toxic intrathecal amphotericin B to be avoided, but for coccidioidal meningitis it must be administered for the rest of the patient's life. More than two-thirds of patients relapse if fluconazole is stopped, even after many years of therapy. Cerebrospinal fluid abnormalities normalize very slowly on fluconazole; chemistries may improve more rapidly with intrathecal amphotericin B. However, amphotericin B causes arachnoiditis, and patients may even have cerebrovascular accidents complicating this therapy.

Two recently developed thiazoles, voniconazole and posaconazole, may be superior to itraconazole. Posaconazole has shown very rapid improvement in most patients but there are still relapses when treatment is dropped.

Further reading

Galgiani JN *et al.* (1993). Fluconazole therapy for coccidioidal meningitis. *Annals of Internal Medicine* **119**, 28–35.

Graybill JR *et al.* (1990). Itraconazole treatment of coccidiodomycosis. *American Journal of Medicine* **89**, 292.

Stevens DA (1995). Current concepts: coccidioidomycosis. *New England Journal of Medicine* **332**, 1077–82.

7.12.4

Paracoccidioidomycosis

M. A. S. Yasuda

Definition

Paracoccidioidomycosis is a systemic granulomatous disease caused by a dimorphic fungus, *Paracoccidioides brasiliensis*, that involves mainly the lungs, phagocytic mononuclear system, mucous membranes, skin, and adrenals.

History

The disease was first described in 1908 by Lutz, a Brazilian scientist. In 1912, Splendore classified the organism as a yeast of the genus *Zymonema* and in 1928, Almeida and Lacaz suggested the name Paracoccidioides. In 1930, Almeida named the fungus *Paracoccidioides brasiliensis*. Formerly the disease was known as South American blastomycosis, or Lutz–Splendore–Almeida' s disease. In 1977 it was renamed paracoccidioidomycosis.

Epidemiology

Paracoccidioidomycosis is the most common endemic human mycosis in Latin America and is geographically restricted to Central and South America, ranging from Mexico to Argentina. The disease is prevalent in Brazil, Colombia, Venezuela, Argentina, Uruguay, Paraguay, Guatemala, Equador, Peru, and Mexico. No cases have been reported in Chile, Belize, Nicaragua, Guyana, Surinam, or French Guyana. Imported cases have been recorded in the United States, Europe, and Asia.

Prevalence, inferred from the result of intradermal paracoccidioidin testing, ranges from 6 to 60.6 per cent among rural and urban populations of endemic and non-endemic areas. It is equally prevalent in both sexes.

The disease occurs mainly among 20- to 50-year-olds, who are agricultural workers or who have lived in rural endemic areas. The sex ratio of clinical cases is 10 or more males to each female among adults, while it is equally distributed among prepubescent boys and girls. This may be explained by the ability of oestrogens to inhibit the transformation of mycelium or conidia to yeast. Spouses of patients are rarely affected by the disease, which suggests that hormonal and genetic factors play a part in the distribution of this mycosis. Transmission from one person to another has not been shown.

Ecology

The geographical regions in which paracoccidioidomycosis is most commonly found are humid areas where the soil is more frequently acidic and the temperature ranges from 15 to 30 °C.

P. brasiliensis has been isolated from soil, animals such as armadillos and bats, dog food, and penguin faeces. It has also been isolated from the intestinal contents of bats. Efforts to maintain the fungus in bat intestines have been unsuccessful. The saprophytic habitat of *P. brasiliensis* has yet to be discovered.

Aetiology

Mycology

P. brasiliensis is a dimorphic fungus, which can be cultivated either as a mould or a yeast. When cultured at 25 °C it appears after 15 to 30 days as white colonies. When Sabouraud's dextrose agar is used the mycelium shows hyaline septate hyphae with branches.

P. brasiliensis also grows as a yeast in human and animal tissues (Fig. 1) and in cultures maintained at 37 °C. Colonies can be observed after 7 to 20 days. Under direct microscopy, yeast forms can be observed as oval or spherical cells with doubly refractile walls; the cells vary in size from buds of 2 to 10 µm in diameter to mature cells of 20 to 30 µm. Mother cells may produce 10 to 12 uniform or variably sized buds (Fig. 2), forming the characteristic 'pilot wheel' shape observed in biological samples or in infected tissues.

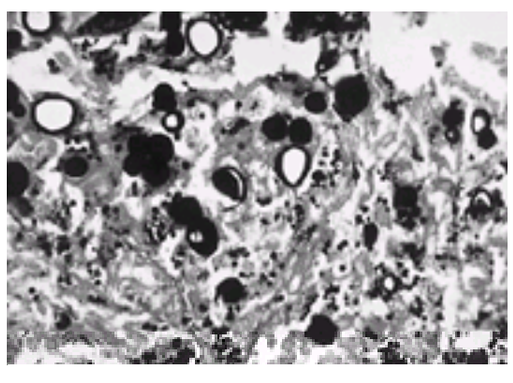

Fig. 1 Small and large yeast forms of *Paracoccidioides brasiliensis* in the lung of a transplant recipient. Methenamine silver stain.

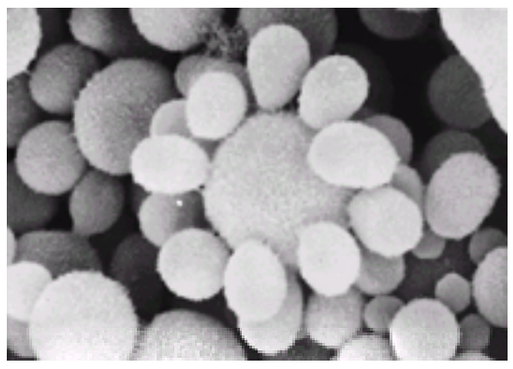

Fig. 2 Scanning electron micrograph of a multiple budding yeast cell of *Paracoccidioides brasiliensis*. (By courtesy of C. S. Lacaz.)

Genomic clones that encode a 70 kDa heat shock protein from this dimorphic fungus have been studied. A differential expression of this gene was observed between mycelial and yeast forms, with a higher level of expression in the yeast form.

Virulence

Virulence is defined as the ability to produce disseminated infection in experimental animals. Variation in the virulence of different fungal isolates has been documented but little is understood of the biochemical basis for these differences.

The presence of higher levels of a-1,3 glucan in virulent strains of *P. brasiliensis* compared with avirulent strains was initially related to virulence, but no correlation has been shown between glucans and virulence in experimentally induced infections.

Pathogenesis

Several experimental and clinicopathological observations provide evidence that the respiratory route is the main portal of entry and the lung is the primary site of infection.

The first fungus–host contact occurs through inhalation of airborne conidia. When mice are experimentally infected through the respiratory route, conidia have been observed in the alveoli soon after inoculation. Some 12 to 18 h after the exposure, yeast forms can be observed in the alveoli. There is an initial inflammatory response, which is mediated by polymorphonuclear cells, followed by granuloma formation.

The primary infective complex develops at the inoculation site and involves the surrounding lymphatic vessels and regional lymph nodes. The fungus spreads to other parts of the lung through peribronchial lymphatic vessels and drains into regional lymph nodes. Haematogenous dissemination to a variety of organs and tissues may occur at this time. The lesions usually undergo involution and the fungi remain dormant if the host's immune response can control their proliferation. A balanced host–fungus relationship is associated with the absence of symptoms, although in some children or young adults, acute disease may arise, primarily affecting the phagocytic mononuclear system. In adult life, previously quiescent lesions may become reactivated, especially in the lungs, leading to the adult or chronic form of the disease.

Pathology

The characteristic lesion is a granuloma containing *P. brasiliensis* cells. The infected tissue may exhibit a predominantly proliferative, granulomatous inflammatory response, and/or an exudative reaction, sometimes resulting in necrosis, with variable numbers of neutrophils and large numbers of extracellular yeast cells, leading to a chronic epithelioid granuloma.

Autopsy studies, mainly of adult patients, indicate that the organs most frequently involved are the lungs (42 to 96 per cent), adrenals (44 to 80 per cent), lymph nodes (28 to 72 per cent), pharynx/larynx (18 to 60 per cent), and skin/other mucosal surfaces (2.7 to 64 per cent).

Host–fungus interaction

Non-specific immune response

The influence of genetic factors on the individual susceptibility to this mycosis is suggested by the observation of higher rates of HLA phenotypes A9, B13, B40, and Cw3 among patients than in controls. In isogenic mice, resistance to *P. brasiliensis* is controlled by a single autosomal gene.

The ability of circulating human neutrophils obtained by bronchoalveolar washing to digest the yeast forms of fungi was impaired in severe cases, while this defect was absent in uninfected family members of patients.

Specific immune response

The relation of the severity of the human disease to deficient late hypersensitivity was established through intradermal testing for ubiquitous antigens and paracoccidioidin, or through lymphoblastic transformation tests to mitogens and to *P. brasiliensis* antigens, including the 43 kDa glycoprotein.

The different distribution of T-lymphocyte subpopulations according to the clinical form of the disease (decreased CD4 in chronic form and increased CD8 in acute form) suggests that different mechanisms might be involved in each form.

The deficient T-cell response is followed by a decreased ability of macrophages to control fungal multiplication and to kill the fungus. This capacity of murine pulmonary macrophages in intratracheal infection is increased *in vivo* and *in vitro* by treatment with interferon-g. Neutralization of endogenous interferon-g by monoclonal antibodies induced exacerbation of the pulmonary infection, earlier fungal dissemination to the liver and spleen, and impairment of the specific cellular immune response and increased levels of IgG1 and IgG2b specific antibodies.

In severe human disease there are decreased levels of T helper 1 type cytokines (interferon-g and interleukin 2 (**IL-2**)) and preserved T helper 2 type cytokines (IL-10 and IL-13 or IL-4). This pattern is associated with poor granuloma formation, spreading of the fungus and high levels of antibody production (immunoglobulins IgG1, IgG4, and IgE). The importance of late hypersensitivity in protection has been observed recently in patients receiving cytotoxic therapy for associated neoplasms and in those with AIDS.

Antibodies may enhance phagocytosis through opsonization of the fungus, but their role in resistance is not established.

Clinical features

The clinical picture ranges from an asymptomatic course to severe disseminated disease, which can lead to death. The incubation period is unknown except in a laboratory worker, who developed a skin lesion some days after an accidental inoculation. The disease has been reported in children 3 years of age or older who had lived for some years in the endemic area.

The following classification of clinical forms of paracoccidioidomycosis has been proposed:

1. paracoccidiodomycosis infection;
2. regressive (self-healing) paracoccidioidomycosis;
3. paracoccidioidomycosis disease;
    a. acute form (juvenile type): moderate or severe;
    b. chronic form (adult type): mild, moderate, or severe;
4. sequelae.

Localization in a particular tissue or organ and the degree of severity of the disease according to established criteria make this classification easily and uniformly applicable. General and nutritional debility and organ dysfunction (lung, brain, adrenals, bone marrow) indicate the severity of the disease.

Acute form (juvenile type)

Children, adolescents, and young adults (under 30 years old) are affected; males and females being afected in equal numbers. Only 1 to 20 per cent of the patients fall into this group. There is progression for 2 to 3 months or more, characterized by involvement of the phagocytic mononuclear system. Cervical, axillary, and inguinal nodes are the most commonly enlarged (Fig. 3). Nodes are initially hard but are sometimes fluctuant and drain pus rich in fungi. Less frequently, deep-seated lymph nodes may also be affected. When the hepatic perihilar lymph nodes are enlarged, they may produce symptoms of obstructive jaundice.

Fig. 3 Lymph node and skin involvement in a patient with the acute form of paracoccidioidomycosis. (Courtesy of C. S. Lacaz.)

The liver and spleen are usually moderately enlarged. Bones (clavicle, scapulae, ribs, skull, long, and flat bones) and, rarely, the bone marrow may be involved. Radiographs show lytic lesions without periosteal reaction. Involvement of the small bowel may be asymptomatic or produce abdominal pain, diarrhoea, constipation, and even intestinal obstruction. Radiological studies of the digestive tract reveal intestinal tract involvement in about 50 per cent of clinical cases.

Fever and weight loss are common. Multiple mucocutaneous lesions are more frequent in some geographical areas. High transient blood eosinophilia (up to 30 000/mm^3) has sometimes been described.

Clinical lung involvement is rarely described in this form of paracoccidioidomycosis. In some case reports either bronchopneumonia or primary complex-like disease was observed.

Chronic form

This form of the disease usually occurs in 30- to 50-year-old men who have worked in agricultural areas. The male:female ratio varies from 10:1 to 25:1. The evolution is insidious and in many cases clinically mild.

The organ most frequently involved is the lung, followed by skin and mucous membranes, mainly pharynx, larynx, and trachea. Lymph nodes and adrenals may be compromised. More than one organ or tissue is usually involved. Less frequently, intestine, spleen, bones, central nervous system (brain, cerebellum, meninges), eyes, genitourinary system, myocar-dium, pericardium, and arteries are involved.

The patients may be asymptomatic or complain of dyspnoea, cough, sometimes purulent sputum, and rarely haemoptysis. Fever is unusual. Physical examination is frequently normal or there may be scattered rales. In contrast, chest radiography commonly reveals bilateral, asymmetrical, reticulonodular infiltrates in the middle and lower parts of the lungs (Fig. 4). Apical cavities and pleural effusions are less frequently observed.

Fig. 4 Alveolar and interstitial infiltrates in both lungs in a patient with chronic paracoccidioidomycosis.

Cutaneous lesions include papules, pustules, ulcers, crusted ulcers, vegetations, tuberculoids, verrucoids, or acneiform lesions mainly on the face (Fig. 5) or limbs. Multiple, scattered lesions result from haematogenous dissemination. Subcutaneous cold abscesses, more commonly associated with bone lesions, can occur.

Fig. 5 Mucocutaneous lesions in a patient with chronic paracoccidioidomycosis. (By courtesy of C. S. Lacaz.)

Mucosal lesions are usually in the mouth and/or oropharynx, including the palate, uvula, and tonsils, or in the respiratory tract, involving mainly the larynx (vocal cords, glottis, and epiglottis) and trachea. Pain is usually intense, and may hamper mastication and swallowing. Hoarseness and dysphonia result from laryngeal lesions, and may lead to obstruction of the upper respiratory tract. Examination shows ulcerative, verrucous, vegetant, and infiltrative 'moriform' stomatitis, resembling a raspberry, with papules, vesicles, and haemorrhagic spots. The last is characteristic of this mycosis and appears as shallow ulcers, with a granular surface showing multiple, fine, haemorrhagic points.

Few lymph nodes may be involved, in contrast to the acute form of the disease.

Uni- or bilateral lesions in the adrenal glands have been found in about half of patients coming to autopsy. Partial adrenal insufficiency has been documented in about 40 per cent of the cases but only 7.4 per cent were symptomatic.

Concomitant tuberculosis is observed in about 10 to 15 per cent of cases of pulmonary paracoccidioidomycosis and has also been described in cases of lymph node involvement by *P. brasiliensis.* Carcinomas may arise in pulmonary or mucosal mycotic lesions.

Sequelae

Nowadays these constitute one of the most important problems in the management of paracoccidioidomycosis. Although fungal multiplication can been controlled by chemotherapy, impairment of vital functions might prove fatal.

Acute form

Lesions in the small intestine and mesenteric lymph nodes may fibrose causing lymphatic obstruction, intestinal malabsorption, or protein-losing enteropathy. A clinical picture of severe malnutrition and immunodeficiency has been reported (Fig. 6).

Fig. 6 Ascites, cachexia, and immunodeficiency due to malabsorption and protein-losing enteropathy as sequelae of acute paracoccidioidomycosis. (By courtesy of M. Shiroma.)

Chronic form

As the lesions usually tend to heal by fibrosis, sequelae such as microstomy and laryngeal, tracheal, or even bronchial stenosis may be observed. Corrective surgery is indicated.

Pulmonary emphysema, fibrosis, respiratory insufficiency, and, finally, cor pulmonale are frequent sequelae. Obstructive and restrictive patterns of ventilatory defect have been found in about 36 and 16 per cent of patients respectively. As many as 30 per cent of these patients may die as a result of respiratory or cardiorespiratory failure.

Diagnosis

Microbiological identification

Isolated or budding (single or multiple) mother cells are observed under direct microscopy in sputum, pus from lymph nodes, and material from the skin or mucous membrane lesions.

Specimens are cultured at 37 °C on blood, chocolate, or yeast extract agar. The colonies are produced after 7 days, usually in 10 to 20 days. Cultures can be maintained, at 25 °C, on Sabouraud's dextrose agar, where the colonies may be noticed after 15 to 30 days.

Histopathology

Silver or periodic acid-Schiff staining is required to detect the fungus on sputum. Diagnostic features are the variable size (1 to 30 µm) of the yeast cells, and their multiple budding. Proliferative or exudative reactions, as described in the section on pathology, may be observed.

Immunological test

Serological reactions

Immunodiffusion (Ouchterlony) and counterimmunoelectrophoresis are the best techniques initially. Sensitivities and specificities are as high as 95 per cent. Cross reactions are mainly with other deep mycoses such as histoplasmosis, aspergillosis, cryptococcosis, and candidiasis.

Complement fixation and indirect immunofluorescence are less reliable tests for diagnosis, but they can be employed in patients under treatment.

Recently, enzyme immunoassays employing *P. brasiliensis* antigens, including a 43 kDa glycoprotein have shown high sensitivity and specificity. Antibody titres tend to decrease about 3 to 6 months after starting specific therapy and to disappear after 9 months to 5 years or more.

Antigenaemia and antigenuria have been considered useful indications in patients presenting low levels of antibodies in the sera, both for diagnosis and follow-up after treatment, particularly in an immunocompromised host.

The correlation between immunological and histopathological findings and clinical forms is outlined in Table 1.

Therapy

Clinically active disease is treated for 3 to 6 months, followed by maintenance therapy with sulfamethoxipiridazine after the resolution of clinical signs and symptoms, continued for many months or until 1 to 2 years after antibody levels have fallen to normal.

Severe cases of acute or chronic disease should be treated with intravenous infusion of amphotericin B. The daily dose begins at 0.1 to 0.2 mg/kg, increasing up to 1.0 mg/kg. The total dose ranges from 1 to 3 g or more. Toxic reactions to amphotericin B include fever, chills, headache, anaemia, and nephrotoxicity characterized by tubular acidosis and potassium urinary excretion and resultant hypokalaemia and azotaemia. In most cases, these reactions can be controlled until the end of the course of therapy. Liposomal amphotericin has been used in severe cases of paracoccidioidomycosis, but this treatment was followed by relapses.

In milder cases, sulphonamides or imidazoles (ketoconazole 200 to 400 mg/day or itraconazole 100 to 200 mg/day) have been shown to be effective. In a randomized trial, sulphadiazine (150 mg/kg per day), itraconazole (50 to 100 mg/day), and ketoconazole (200 to 400 mg/day) were equally effective in patients with moderately severe disease. The combination of 160 mg of trimethoprim and 800 mg of sulfamethoxazole is also effective. Fluconazole has been used in a few cases and although it achieves high levels in cerebrospinal fluid, there is no conclusive experience in neuroparacoccidioidomycosis.

Prognosis

Even though the disease is easily controlled in the majority of cases, the course of treatment is long and in Brazil, for example, abandonment of treatment is the most important cause of therapeutic failure. Normalization of cellular specific responses, particularly of the skin test (paracoccidioidin) indicates a good prognosis.

Death may occur in severe acute or chronic cases and severe cases with sequelae.

Further reading

Bueno JP *et al*. (1997). IgG, IgM and IgA antibody response for the diagnosis and follow-up of paracoccidioidomycosis:comparison of counterimmunoelectrophoresis and complement fixation. *Journal of Medical and Veterinay Mycology* **35**, 213–17.

Calich VLG *et al*. (1985).Susceptibility and resistance of inbred mice to *P. brasiliensis*. *British Journal of Experimental Pathology* **66**, 585–94.

Restrepo A (1985). The ecology of *Paracoccidioides brasiliensis:* a puzzle still unsolved. *Journal of Medical Mycology* **23**, 323–34.

7.12.5 Pneumocystis carinii

Robert F. Miller and Ann E. Wakefield

Who gets *Pneumocystis carinii* pneumonia?

Most patients have abnormalities of T-lymphocyte function or numbers, but rarely *Pneumocystis carinii* pneumonia develops in patients with isolated B-cell defects and in individuals without evidence of immunosuppression. In non-HIV immunosuppressed individuals, glucocorticoid administration is an independent risk factor for development of *P. carinii* pneumonia irrespective of the type or intensity of immunosuppression or the nature of the underlying disease process. In HIV-infected individuals, those at greatest risk have CD4+ T lymphocyte counts less than 200 cells/μl. *P. carinii* pneumonia in HIV-infected patients in Europe, United States, and Australasia is now largely confined to patients who are unaware of their HIV serostatus at presentation or to those who are non-compliant with, or intolerant of, prophylaxis and antiretroviral therapy. The incidence of *P. carinii* pneumonia in HIV-infected individuals in Africa is lower than in the West.

Aetiology

Until recently, *P. carinii* was regarded taxonomically as a protozoan, based on its morphology and lack of response to antifungal agents such as amphotericin B. *P. carinii* pneumonia cannot be cultured *in vitro*, but molecular biological techniques demonstrate clearly that it is a fungus. *P. carinii* from different mammalian host species show antigenic, karyotypic, and genetic heterogeneity. Cross infection between host species has not been successful, suggesting host specificity and that *P. carinii* infection in man is not a zoonosis. In the human host, *P. carinii* shows lower levels of genetic diversity than occurs between *P. carinii* from different mammalian hosts. Over 30 genotypes of human type *P. carinii* have been described; some types are associated with a mild pneumonia, others with severe hypoxic pneumonia.

The demonstration of antibodies against *P. carinii* in the majority of healthy children and adults has been regarded previously as supportive of the hypothesis that *P. carinii* arises in an immunocompromised individual by reactivation of a childhood-acquired, symptomless, latent infection. However, this hypothesis is challenged by the failure to demonstrate *P. carinii* in bronchoalveolar lavage (BAL) fluid or necropsy lung tissue of immune competent individuals, and the observation that *P. carinii*-specific DNA is detectable only at low levels in less than 25 per cent of HIV-infected individuals with low CD4+ T lymphocyte counts presenting with respiratory episodes and diagnoses other than *P. carinii* pneumonia. Human *P. carinii* infection is now thought to arise from *de novo* infection from an exogenous source. The finding of different *P. carinii* genotypes in each episode in patients with recurrent *P. carinii* pneumonia supports the reinfection model.

Pathogenesis

After inhalation of *P. carinii*, the organism reaches the alveoli where the trophozoite form attaches to type 1 pneumocytes. In an immune-competent individual the organism is eliminated, in the immune-deficient host *P. carinii* pneumonia will develop.

The major surface glycoprotein of *P. carinii* binds macrophages and induces T-lymphocyte proliferation and increases secretion of IL-1 and -2 and TNF-a. Monocytes respond to major surface glycoprotein by releasing IL-8 and TNF-a. *P. carinii* induces changes in the quantity and quality of pulmonary surfactant: total cholesterol, glycerol, and phospholipase A-2 are increased while phospholipid is reduced.

Clinical presentation

This is non-specific. Patients typically present with progressive exertional dyspnoea, a non-productive cough, and fever of several days or weeks duration. Patients often report an inability to take in a deep breath, not due to pleural pain. Purulent sputum, haemoptysis, and pleural pain are atypical for *P. carinii* and suggest a bacterial or mycobacterial pathogen. In HIV-infected patients, the presentation is usually more insidious than in patients immunosuppressed by other causes, however in a small proportion of HIV-infected patients the disease course of *P. carinii* is fulminant with an interval of 7 days or less between onset of symptoms and progression to respiratory failure. Occasionally *P. carinii* may have an indolent presentation with respiratory symptoms worsening almost imperceptibly over many months. Rarely, *P. carinii* may present as pyrexia of undetermined origin.

Examination of the chest is usually normal; occasionally fine bibasal end-inspiratory crackles are heard. Signs of focal consolidation or pleural effusion suggest an alternative diagnosis.

Pathology

Within the lung, *P. carinii* infection is characterized by an eosinophilic, foamy intra-alveolar exudate, associated with a mild plasma cell interstitial pneumonitis. Morphologically, two forms of *P. carinii* may be identified: thick-walled cysts (6–7 μm diameter) which lie freely within the alveolar exudate are demonstrated by Grocott's methenamine silver, toluidine blue O, or cresyl violet stains. The exudate consists largely of thin-walled, irregularly shaped, single-nucleated trophozoites (2–5 μm diameter) which are shown by Geimsa stain but lack distinctive features. Rarely, interstitial fibrosis, diffuse alveolar damage, granulomatous inflammation, nodular and cavitary lesions, and pneumatocoele formation may occur.

Rarely *P. carinii* infection extends beyond the air spaces; extrapulmonary pneumocystosis involving liver, spleen, gut, or eye may occur and is strongly associated with use of nebulized pentamidine for prophylaxis or treatment.

Investigations

The chest radiograph may be normal in early or mild pneumonia. With more severe disease or later presentation, diffuse perihilar interstitial infiltrates are seen. These may progress to diffuse bilateral alveolar (air space) consolidation that mimics pulmonary oedema. In the late stages the lungs may be massively consolidated and almost airless. Radiographic deterioration from near normal at presentation to being markedly abnormal may occur over 48 h or less. Up to 20 per cent of chest radiographs are atypical, showing intrapulmonary nodules, cavitary lesions, lobar consolidation, pneumatocoeles, or hilar/mediastinal lymphadenopathy. Predominantly apical change may be seen in patients who develop *P. carinii* pneumonia having received *P. carinii* prophylaxis with nebulized pentamidine. All these typical and atypical radiographic appearances may also be seen in bacterial, mycobacterial, and fungal infection, and in non-specific pneumonitis and Kaposi's sarcoma.

With treatment and clinical recovery the chest radiograph in some individuals may remain abnormal for many months in the absence of symptoms. In others

postinfectious bronchiectasis or fibrosis occurs.

Arterial blood gases/oximetry

Less than 10 per cent of patients with *P. carinii* pneumonia have a normal PaO_2 and a normal PAO_2–PaO_2. These measures are sensitive though not specific for *P. carinii* pneumonia and may also occur in bacterial pneumonia, Kaposi's sarcoma, and tuberculosis.

Computed tomography

High resolution computed tomography scanning of the chest may be useful in the symptomatic patient with a normal or equivocal chest radiograph. Areas of ground glass shadowing indicate active pulmonary disease. These appearances may be caused by *P. carinii*, cytomegalovirus, or fungal pneumonia.

Induced sputum

Spontaneously expectorated sputum is inadequate for diagnosis of *P. carinii* pneumonia. Sputum induction by inhalation of ultrasonically nebulized hypertonic (3N) saline may provide a suitable sample. *P. carinii* is usually found in clear 'saliva-like' samples. Purulent samples suggest an alternative diagnosis. The sensitivity varies between 55 and 90 per cent and a negative result for *P. carinii* should prompt further diagnostic tests.

Bronchoscopy

Fibreoptic bronchoscopy with BAL has a sensitivity of more than 90 per cent for detection of *P. carinii.* Immunoflourescence staining increases the diagnostic yield compared to conventional histochemical staining. Transbronchial biopsies add very little to the diagnostic yield and are associated with a relatively high complication rate (pneumothorax in » 8 per cent). As *P. carinii* persists in the lung for many days after the start of antimicrobial therapy, bronchoscopy may be performed up to 1 week after commencing anti-*P. carinii* therapy without a reduction in diagnostic yield.

Molecular diagnostic tests

Detection of *P. carinii*-specific DNA by the polymerase chain reaction (PCR) on BAL fluid and induced sputum is superior to conventional histochemical methods. Detection of *P. carinii* DNA by PCR may also be achieved on oropharangeal samples obtained by gargling with normal saline; this technique compared to conventional staining of BAL fluid has a sensitivity of 89 per cent and a specificity of 94 per cent for *P. carinii*. These molecular techniques are not widely available.

Empirical therapy

Many centres in the United Kingdom and North America seek to confirm a diagnosis in every suspected case of *P. carinii* pneumonia. Other centres treat HIV-infected patients empirically who present with symptoms, chest radiographic abnormalities, and hypoxaemia, features typical of *P. carinii* pneumonia. Bronchoscopy is reserved for those who fail to respond to empirical therapy by day five or those who have atypical presentations. Both strategies are equally effective in clinical practice.

Treatment

It is important to stratify *P. carinii* pneumonia as mild (PaO_2 (on air) > 11.0 KPa, SaO_2 > 96 per cent) moderate (PaO_2= 8.0–11.0 KPa, SaO_2 = 91–96 per cent), or severe (PaO_2< 8.0 KPa, SaO_2 < 91 per cent) as some drugs are unproven or ineffective in severe disease.

First choice treatment is high-dose co-trimoxazole (sulfamethoxazole 100 mg/kg per day and trimethoprim 20 mg/kg per day) in two to four divided doses, orally or intravenously. In HIV-infected patients with *P. carinii* pneumonia 21 days are given, in those with other causes of immunosuppression 14 to 17 days are frequently given. In mild disease oral medication may be given throughout, in moderate/severe disease intravenous therapy is usually given for the first 7 to 10 days, then orally.

Other treatments in patients with severe disease include clindamycin 450 to 600 mg orally or intravenously, four times daily, with primaquine 15 mg once daily orally, or trimetrexate 45 mg/m^2 intravenously, daily, with folinic acid 20 mg/m^2 four times daily. Despite its toxicity, pentamidine 4 mg/kg daily, intravenously, may be used if other treatments have failed. In patients with mild or moderate disease, alternatives to co-trimoxazole include clindamycin with primaquine (doses as above), dapsone 100 mg orally once daily, with trimethoprim 20 mg/kg per day, or atovaquone 750 mg orally twice daily.

Adjuvant steroids

HIV infected patients with moderate/severe *P. carinii* pneumonia benefit from adjuvant glucocorticoids which reduce the risk of respiratory failure, need for mechanical ventilation, and risk of death. Many non-HIV infected patients with *P. carinii* pneumonia are already receiving glucocorticoids as part of their immunosuppression/chemotherapy and the benefits of adjunctive dose increases have not clearly been demonstrated. Adjunctive glucocorticoid regimens include prednisolone 40 mg twice daily orally for 5 days, then 40 mg once daily on day 6 to 10, 20 mg once daily on days 11 to 21 (or methylprednisolone intravenously at 75 per cent of these doses). An alternative regimen is methylprednisolone 1 g intravenously for 3 days, then 0.5 g intravenously on days 4 and 5, followed by prednisolone reducing from 40 mg orally once daily to zero over 10 days.

Adverse reactions

Adverse reactions to co-trimoxazole, which usually occur between day 6 and day 14 of treatment, are commoner in HIV-infected patients than in patients with other causes of immunosuppression. Anaemia and neutropenia (up to 40 per cent of patients), rash and fever (up to 30 per cent), and biochemical hepatitis (up to 15 per cent) are the most frequent adverse reactions. Coadministration of folic or folinic acid does not prevent or attenuate haematological toxicity and may be associated with increased therapeutic failure.

Glucose-6-phosphate dehydrogenase deficiency

Patients with glucose-6-phosphate dehydrogenase deficiency should not receive co-trimoxazole, dapsone, or primaquine.

Prophylaxis

HIV-infected patients are at increased risk of *P. carinii* pneumonia as the CD4+ lymphocyte count decreases. Primary prophylaxis (to prevent a first episode of *P. carinii* pneumonia) is given when the CD4 count falls below 200/µl or the CD4:total lymphocyte ratio is less than 1:5, to patients with HIV- associated constitutional features such as unexplained fever of 3 weeks' or more duration, or oral candida, irrespective of CD4 count, and to patients with other AIDS-defining diagnoses, for example Kaposi's sarcoma. Secondary prophylaxis is given after an episode of *P. carinii* pneumonia.

The first choice agent for primary and secondary prophylaxis is co-trimoxazole 960 mg daily. Lower doses, that is 960 mg three times weekly or 480 mg daily, may be equally effective and have fewer side effects. Co-trimoxazole may also protect against bacterial infections and reactivation of cerebral toxoplasmosis.

Adverse reactions, including rash with or without fever, occur in up to 20 per cent of patients receiving co-trimoxazole as prophylaxis. Desensitization may be attempted in those unable to tolerate co-trimoxazole; alternative less effective options include nebulized pentamidine 300 mg once per month via jet nebulizer (once per fortnight if the CD4 count is 50/µl or less), dapsone 100 mg daily with pyrimethamine 25 mg once weekly (pyrimethamine may also protect against cerebral toxoplasmosis), or atovaquone 750 mg twice daily.

Non-HIV infected patients with high attack rates of *P. carinii* pneumonia should receive prophylaxis (drug choice and doses as above). At risk groups include those with acute lymphoblastic leukaemia, severe combined immunodeficiency syndrome, Hodgkin's disease, rhabdomyosarcoma, primary and secondary central nervous

system tumours, Wegener's granulomatosis, and organ transplantation including allogenic bone marrow, renal, heart, heart/lung, and liver.

Areas of uncertainty/future research

The mode of transmission of *P. carinii* infection is unclear, but recent molecular data suggest that transmission from infected patients to susceptible immunocompromised individuals may occur. The drug target for sulfamethoxazole and dapsone is dihydropteroate synthase. The possibility that *P. carinii* may develop resistance to sulpha drugs is suggested by the finding of non-synonymous single nucleotide polymorphisms (which are associated with resistance in other organisms) in the dihydropteroate synthase gene of human *P. carinii* which occur more frequently in those who have received prophylaxis with co-trimoxazole or dapsone. The limited availability of other equally effective drugs for prophylaxis restricts the use of 'drug switching' as a strategy for preventing emergence of resistance of *P. carinii* to sulpha drugs.

Of new anti *P. carinii* drugs under development, and most promising are sordarin derivatives which target translation elongation factor-2 and inhibit fungal protein synthesis.

Further reading

Dei-Cas E, Cailliez JC, eds (1998). *Pneumocystis* and pneumocystosis: advances in *Pneumocystis* research. *FEMS Immunology and Medical Microbiology* **22**, 1–189. A summary of current knowledge about the molecular biology of the organism.

Miller RF (1999). Pneumocystis carinii infection in non-AIDS patients. *Current Opinion in Infectious Diseases* **12**, 371–7. Comprehensive review of non-AIDS-associated *Pneumocystis carinii* pneumonia.

Miller RF, Lipman MCI (1999). Pulmonary infections (AIDS). In: Albert R, Spiro S, Jett J, eds. *Comprehensive respiratory medicine*. Mosby, London, pp. 32.1–22. Comprehensive review of *Pneumocystis carinii* pneumonia and other infections in AIDS.

Miller RF, Lenoury J, Corbett EL, Felton JM, DeCock KM (1997). Pneumocystis carinii infection: current treatment and prevention. *Journal of Antimicrobial Chemotherapy* **77** (Suppl. B), 33–53. A comprehensive review of treatment and prophylaxis regimens for *Pneumocystis carinii* pneumonia.

7.12.6 Infection due to pencillium marneffei

Thira Sirisanthana

Introduction

Penicillium marneffei was first isolated from bamboo rats (*Rhizomys sinensis)* in Vietnam in 1956. The fungus is endemic in southeast Asia, northeast India, south China, Hong Kong, and Taiwan. Fewer than 40 cases of infection with *P. marneffei* were reported prior to the HIV epidemic. The incidence of disseminated *P. marneffei* infection has increased markedly over the past few years. This increase is mainly due to infection in patients already infected with HIV. The majority of patients have been reported from Thailand, Hong Kong, and Taiwan. Cases have also been reported in HIV-infected individuals from the United States, the United Kingdom, The Netherlands, Italy, France, Germany, Switzerland, Sweden, and Australia following visits to the ndemic region.

Aetiology

P. marneffei is the only dimorphic fungus of the genus *Penicillium*. The fungus grows in a mycelial phase at 25 °C on Sabouraud dextrose agar. Mould-to-yeast conversion is achieved by subculturing the fungus on to brain–heart-infusion agar and incubating at 37 °C. In its mycelial form, the colony is greyish white and downy. The colour of the colony may vary during differentiation. The reverse side becomes cerise to brownish red, as a soluble red pigment diffuses into the agar medium. Microscopic examination of the mycelial form shows structures typical of the genus *Penicillium*. Colonies of the yeast form of *P. marneffei* have a wrinkled or cerebriform surface. They are light tan to brown in colour. Microscopic examination of the yeast form reveals unicellular, pleomorphic, ellipsoidal-to-rectangular cells, about 2 µm by 6 µm in size, that divide by fission and not by budding.

Natural history

Many features of the natural reservoir, mode of transmission, and natural history of *P. marneffei* infection remain unknown. The fungus has been isolated from several species of bamboo rats in the endemic area. Since bamboo rats usually live near the forest and have limited contact with people, it is believed that both humans and bamboo rats become infected with *P. marneffei* from a common source, rather than the patients being infected by the rats. By analogy with other endemic systemic mycoses, such as histoplasmosis, it is likely that *P. marneffei* conidia are inhaled from a contaminated reservoir in the environment and subsequently disseminate from the lungs when the host experiences immunosuppression. The disease is more likely to occur in the rainy season, suggesting that there may be an expansion of the environmental reservoirs with favourable conditions for growth at this time.

In endemic areas it is likely that a certain proportion of the population is infected, but remains asymptomatic. Patients have been reported with long periods of asymptomatic infection before presentation with clinical *P. marneffei* infection. In other cases, clinical manifestation of *P. marneffei* infection occurred within weeks of exposure to the fungus.

Clinical features

Patients with *P. marneffei* infection commonly present with symptoms and signs of infection of the reticuloendothelial system. These include fever, chills, lymphadenopathy, hepatomegaly, and splenomegaly. Cough, dyspnoea, and lung crepitations may be present. Other manifestations are secondary to dissemination of the fungus via the bloodstream. Cutaneous and subcutaneous lesions are observed in up to two-thirds of patients. Arthritis and osteomyelitis are not uncommon. Cases with mesenteric lymphangitis, colitis, genital or oropharyngeal ulceration, retropharyngeal abscess, or pericarditis have been reported.

In HIV-infected patients, *P. marneffei* infection occurs late in the course of the disease. The patient's CD4+ cell count at presentation is usually below 50 cells per microlitre. HIV-infected patients with *P. marneffei* infection have a more acute onset and higher fever. They are more likely to have fungaemia and shock and their skin lesions are more numerous and tend to be papules with central necrotic umbilication. Patients who are not infected with HIVare more likely to have one or several subcutaneous nodules which may develop into abscesses and cause skin ulceration.

Biochemical and haematological laboratory findings are non-specific and include elevation of liver enzymes, anaemia, and leucocytosis. Chest radiographs may show diffuse interstitial, localized alveolar, or diffuse alveolar infiltrates. Cases with cavitary lesions or lung masses have been reported.

Diagnosis

Presumptive diagnosis can be made by microscopic examination of Wright's-stained samples of bone marrow aspirate, and/or touch smears of skin biopsy specimens, and/or lymph node biopsy specimens. Many intracellular and extracellular basophilic, spherical, oval, and elliptical yeast cells can be seen using this staining technique. Some of these cells had clear central septation, which is a characteristic feature of *P. marneffei*. The diagnosis is confirmed by histopathological section and/or by culturing the fungus from the blood, skin biopsy specimens, bone marrow, or lymph nodes. Cases of *P. marneffei* infection can clinically resemble tuberculosis, histoplasmosis, and cryptococcosis. Tests to detect antibodies or *P. marneffei* antigens have been developed. Clinical trials are needed to show their usefulness in the diagnosis of active *P. marneffei* infection and in predicting relapses. They may also be used to identify HIV-infected individuals who are infected with *P. marneffei* but who are still asymptomic. These persons may then benefit from pre-emptive treatment with an antifungal agent.

Treatment

P. marneffei infection is potentially fatal. The fungus is sensitive to ketoconazole, fluconazole, itraconazole, and amphotericin B. The recommended treatment is to give amphotericin B intravenously in a dose of 0.6 mg/kg/day for 2 weeks, followed by itraconazole 400 mg/day orally in two divided doses for the next 10 weeks. The majority of patients respond well, with resolution of fever and other signs of infection within the first 2 weeks. After initial treatment, HIV-infected patients should be given 200 mg/day of itraconazole orally as secondary prophylaxis for life.

Further reading

Deng Z *et al*. (1988). Infection caused by *Penicillium marneffei* in China and Southeast Asia: review of eighteen published cases and report of four more Chinese cases. *Review of Infectious Diseases* **10**, 640–52. A review of *Penicillium marneffei* infection in patients not infected with the human immunodeficiency virus.

Sirisanthana T, Supparatpinyo K (1998). Epidemiology and management of penicilliosis in human immunodeficiency virus-infected patients. *International Journal of Infectious Diseases* **3**, 48–53. A review of the epidemiology and management of penicilliosis.

Supparatpinyo K *et al.* (1994). Disseminated *Penicillium marneffei* infection in southeast Asia. *The Lancet* **344**, 110–13. A report of the clinical findings in patients with disseminated *Penicillium marneffei* infection.

Supparatpinyo K *et al.* (1998). A controlled trial of itraconazole to prevent relapse of *Penicillium marneffei* infection in patients infected with the human immunodeficiency virus. *New England Journal of Medicine* **339**, 1739–43. A report on the means to prevent relapse of *Penicillium marneffei* infection.

7.13.1 Amoebic infections

R. Knight

The amoebic species infecting humans belong to two very different groups. First, the obligate parasitic species of the gut that include the major pathogen *Entamoeba histolytica*, several non-pathogenic species including *E. dispar*, and a minor pathogen *Dientamoeba fragilis*. The second group are free-living, water and soil amoebae, which can become facultative tissue parasites. All motile feeding amoebae are called trophozoites; they move with pseudopodia and divide by binary fission. The hyaline external cytoplasm, the ectoplasm, is a contractile gel that surrounds the sol endoplasm containing numerous phagocytic and pinocytic vacuoles. Most species can form environmentally resistant cysts by rounding up and secreting a chitinous cyst wall.

Entamoeba histolytica infection

Biology and pathogenicity

Following ingestion of infective cysts a population of trophozoites becomes established in the caecum and proximal colon. Some degree of tissue invasion occurs in all subjects with at least low-titre seroconversion. Tissue invasion is frequently mild, self-limiting, and with minimal symptoms, but at the other end of the clinical spectrum it can lead to extensive destruction of the colonic mucosa. Invasive trophozoites have a characteristic morphology; they may reach 30 to 40 µm in diameter and are very active with apparently purposeful, unidirectional movements during which they become considerably elongated. Their most important diagnostic characteristic is the presence of host erythrocytes within the endoplasm, which otherwise appears clear and contains no bacteria. Trophozoites containing red blood cells are described as erythrocytophagous. Progression through tissues is by active movement, facilitated by secreted collagenase; leucocytes are drawn chemotactically towards the amoebae but most are rapidly destroyed on contact.

The transmissive cystic form of the parasite is derived entirely from a commensal population within the colonic lumen. Live commensal amoebae measure 10 to 20 µm in diameter, the endoplasm is granular and contains bacteria; the pseudopodia are blunt and movement is sluggish. Intestinal hurry from any cause, including the use of laxatives, can lead to the appearance of commensal trophozoites in the faeces. Cysts are spherical and measure 11 to 14 µm in diameter; when mature they contain four nuclei, several chromatoid bodies that are a ribosome store, and a glycogen vacuole.

Host factors may increase susceptibility to overt disease. Steroid therapy given systemically or locally into the rectum carries great risk, as may cytotoxic therapy. Severe bowel disease is particularly common in late pregnancy and the puerperium. Before puberty both sexes are equally susceptible to hepatic amoebiasis, but in adults this condition is at least seven times more common in males. Local disease can also favour tissue invasion; thus amoebic ulceration may be superimposed upon colonic and rectal cancers, or those of the uterine cervix. Colonic disease is favoured by concurrent *Trichuris* infection and intestinal schistosomiasis. Infection with human immunodeficiency virus appears to have little effect on outcome.

The taxonomic separation of *E. dispar* as a discrete non-pathogenic species from *E. histolytica* was formally made in 1993. Characterization of cultures by zymodeme, using isoenzyme electrophoresis of a small set of glycolytic enzymes, was the first convincing biochemical distinction between these two species, but many genomic differences have now been identified. All strains of *E. histolytica* are now regarded as pathogenic.

Epidemiology

The incidence of disease is particularly high in Mexico, South America, Natal, the west coast of Africa, and South-East Asia. In most temperate countries *E. histolytica* is now rare and nearly all amoebic disease seen in such countries will have been acquired elsewhere. Symptomless or convalescent carriers are the main source of infection; patients with dysentery normally pass only trophozoites in their stool, and are therefore non-infectious. Cysts remain viable in the environment for up to 2 months. The infection is eventually self-limiting and rarely exceeds 4 years. Tissue invasion can occur at any time during an infection, but is much more common during the first 4 months; the incubation period may be as short as 7 days.

The incidence of amoebiasis in a population is best estimated from seropositivity surveys. Surveys for cysts are of no value as differentiation from *E. dispar* is impossible. All the modes of faeco-oral transmission occur in amoebiasis; of special importance are the food handler and contaminated vegetables; transmission by flies and drinking water is less common. Drinking water can be contaminated in the home or at surface-water sources. Direct spread can produce outbreaks; it occurs within institutions for children and the mentally handicapped, and with contaminated colonic irrigation equipment. Household clustering is common; hand-fed infants are frequently infected from the fingers of their mother. Contamination of piped water supplies can lead to serious disease outbreaks as happened in the Chicago hotels epidemic in 1933. Nearly all *Entamoeba* infections among male homosexuals are due to *E. dispar*, *E. coli*, or *E. hartmanni*.

Pathology

The basic lesion is the result of cell lysis and tissue necrosis, which, by creating locally anoxic and acidic conditions, favours further penetration of the parasite; most amoebae are seen at the advancing edge of the lesion with little inflammatory cell response. In tissue sections amoebae stain indistinctly with haematoxylin and eosin but appear bright red with periodic acid–Schiff stain; iron haematoxylin is necessary to show nuclear detail. Cysts of *E. histolytica* are never seen in tissue.

Amoebic lesions of the gut are most common in the rectosigmoid and caecum but can occur anywhere in the large bowel; involvement may be patchy or continuous, less commonly the appendix or terminal ileum are affected. The initial lesions are either small, discrete erosions of the mucosa, or minute crypt lesions. Unrestrained, the lesions extend through the mucosa, across the muscularis mucosa, and into the submucosa, where they expand laterally to produce lesions that are typically flask shaped in cross-section. Further lateral spread of the submucosal lesions leads to their coalescence, and later, to denudation of overlying mucosa. The bowel wall may become appreciably thickened. Blood vessels involved in the disease may thrombose, bleed into the gut lumen or, in the case of portal-vein radicles, provide a vehicle for the dissemination of amoebae to the liver. In very severe lesions, and usually in association with toxic megacolon, there is an irreversible coagulative necrosis of the bowel wall.

Amoebomas are tumour-like lesions of the colonic wall measuring up to several centimetres in length; they are most common in the caecum and may be multiple. Histologically there is tissue oedema, with a mixed picture of healing and new areas of epithelial loss and tissue destruction; round-cell infiltration is patchy. Lesions may be annular and rarely an amoeboma initiates an intussusception; narrow, stricture-like amoebomas may occur in the anorectal region.

Amoebae reach the liver in the portal vein. Once initiated the amoebic lesion extends progressively in all directions to produce the liver-cell necrosis and liquefaction that constitute an 'amoebic liver abscess'. The lesions are well demarcated from surrounding liver tissue; untreated, nearly all will eventually extend into adjacent structures. Secondary bacterial infection is rare and usually follows rupture or aspiration.

Clinical manifestations

Invasive intestinal amoebiasis (Plate 1, Plate 2)

The clinical features show a wide spectrum from minimal changes in bowel habit to severe dysentery. Lesions may be limited to a small part of the large bowel or extend throughout its length. A relapsing course is common.

Amoebic colitis with dysentery

Dysentery, the passage of loose or diarrhoeal stools containing fresh blood, occurs when there is generalized colonic ulceration, or when more localized lesions occur in the rectum or rectosigmoid. Onset may be gradual, intermittent, or much less commonly, acute. Typically, constitutional upset is initially mild and the patient remains ambulant; mild or moderate abdominal pain is common, often colicky and maximal over affected parts of the gut. Tenesmus can occur but is rarely severe. Stools vary in consistency from semiformed to watery. They are foul-smelling and always contain visible blood; even when watery, faecal matter is nearly always present. Symptoms frequently wax and wane over a period of weeks or even months and such patients can become debilitated and wasted. In a few patients the disease runs a fulminating course. The most frequent physical sign is abdominal tenderness in one or both iliac fossas; but tenderness may be generalized. Affected gut may be palpably thickened. A low fever is common, but dehydration is uncommon. Abdominal distension occurs in the more severely ill patients, who sometimes pass relatively small amounts of stool.

When stool microscopy reveals no erythrocytophagous trophozoites, a careful proctoscopy or sigmoidoscopy should be done. The endoscopic appearances may be non-specific in early, acute, or very severe colitis; the findings are hyperaemia, contact bleeding, or confluent ulceration. In more chronic cases the presence of normal-looking intervening mucosa is highly suggestive of amoebiasis; early lesions are often elevated, with a pouting opening only 1 to 2 mm in diameter; later, ulcers may reach 1 cm or more in diameter, with an irregular outline and often a loosely adherent, yellowish or grey exudate. Mucosal scrapings or superficial biopsies taken at endoscopy should be examined immediately by wet-preparation microscopy.

Special forms of amoebic colitis

Fulminant colitis

This may arise *de novo*, for example in pregnant women or during steroid therapy, or it may evolve during a dysenteric illness. Patients show progressive abdominal distension, vomiting, and watery diarrhoea. Bowel sounds are absent and there may be little or no abdominal tenderness, guarding, or rigidity. Plain radiographs may reveal free peritoneal gas, together with acute gaseous dilatation of the colon; affected segments of bowel may appear relatively narrow and show visible musosal pathology. Barium enema and full sigmoidoscopy are contraindicated. Stools contain erythrocytophagous trophozoites.

Amoebic colitis without dysentery

When ulceration is limited to the caecum or ascending colon, or when early, mild, or localized lesions occur elsewhere in the colon there may be no dysenteric symptoms. Patients complain of change in bowel habit, blood-staining of the stool, flatulence, and colicky pain. Often the only physical sign is tenderness in the right iliac fossa, or elsewhere along the course of the colon. Some patients eventually go into complete remission; others progress to a dysenteric illness.

The most important diagnostic measure is repeated stool examination for erythrocytophagous amoebae; the finding of cysts or commensal trophozoites is of little diagnostic value, especially in endemic areas. Sigmoidoscopy is often normal when the distal bowel is not involved but colonoscopy may reveal typical lesions.

Amoeboma

These present as an abdominal mass, most frequently in the right iliac fossa. The lesion may be painful, tender, and associated with fever. Bowel habit is altered and some patients have intermittent dysentery, especially if lesions are multiple or distal. Evidence of partial or intermittent bowel obstruction may be present, particularly when lesions are distal and annular.

Localized perforation and amoebic appendicitis

Sudden perforation with peritonitis can occur from any deep amoebic ulcer; alternatively, leakage may lead to a pericolic abscess or retroperitoneal cellulitis. Amoebic appendicitis is an uncommon but important condition that occurs when amoebic lesions are confined to the appendix and caecum. The clinical presentation can resemble that of simple appendicitis, often with some clinical evidence of dysentery. If unrecognized at appendicectomy, the outcome can be disastrous with gut perforation; fresh smears should be made from the resected appendix, and examined immediately.

Rectal bleeding

Some patients with amoebiasis present with rectal bleeding, with or without tenesmus; this occurs particularly in children. Massive bleeding into the gut lumen can occur in any form of amoebic colitis but is rare.

Differential diagnosis

Amoebic colitis must be differentiated from other causes of infective colitis. High-volume diarrhoea, copious mucus, and severe tenesmus are all uncommon in amoebiasis. In temperate countries, non-specific ulcerative colitis and colorectal carcinoma create the greatest diagnostic problems. Parasitic conditions to be considered are intestinal schistosomiasis, heavy *Trichuris* infection, and balantidiasis. More chronic amoebic pathology may clinically resemble Crohn's disease, ileocaecal tuberculosis, diverticulitis, or anorectal lymphogranuloma venereum.

Hepatic amoebiasis

Less than half of all patients give any convincing history of dysentery and few have concurrent dysentery. In those with no dysenteric history the interval between presumed infection and presentation may be as short as 3 weeks, or as long as 15 years; for most it is between 8 weeks and a year.

The dominant symptoms are fever and sweating, liver or diaphragmatic pain, and weight loss. Onset of constitutional symptoms is often insidious; but pain may begin abruptly. Most patients seek medical help within 1 to 4 weeks. Fever is typically remittent, with a prominent evening rise, brief rigors, and very profuse sweating. Liver pain may be poorly localized initially and later become pleuritic, referred to the right shoulder tip, or localized to the abdominal wall. Within a few weeks, patients lose much weight and often become anaemic; a painful dry cough is common.

The most important clinical finding is liver enlargement (Fig. 1) with localized tenderness, which should be searched for in the right hypochondrium, the epigastrium, and along all the intercostal spaces overlying the liver. Liver pain, on compression or heavy digital percussion, is a less useful sign. Left-lobe lesions can present as an epigastric mass. Hepatomegaly may be difficult to detect by abdominal palpation when enlargement is mainly upwards, but bulging of the right chest wall may be noted, together with a raised upper level of liver dullness on percussion. Reduced breath sounds or crepitations may be heard at the right lung base.

Fig. 1 Amoebic liver abscess. Hepatic enlargement with focal tenderness in a Thai woman. (By courtesy of Professor S. Looareesuwan.)

Important radiological findings are a raised, or locally upward-bulging, right diaphragm (Fig. 2) with immobility on screening, areas of lung collapse or consolidation, and sometimes a pleural effusion. A neutrophil leucocytosis is almost invariable, the erythrocyte sedimentation rate is raised, and normochromic normocytic anaemia is common. 'Liver function tests' are frequently completely normal, or there may be a raised alkaline phosphatase; less commonly the serum transaminase or bilirubin is elevated. Liver scanning to demonstrate a filling defect is of great value; about 70 per cent of lesions are solitary, but multiple lesions are common in children and those with concurrent dysentery. Ultrasonographic scans and computed tomography are the most useful. Lesions appear round or oval, and are usually 4 to 10 cm in diameter at the time of presentation. On ultrasonography most are hypoechoic with well-defined walls without enhanced echoes. Even when concurrent dysentery is absent the stools are frequently, but not always, positive for *E. histolytica*. Colonoscopy may reveal unsuspected lesions.

Fig. 2 Amoebic liver abscess, radiographic changes: (a) elevated right diaphragm; (b) enormous abscess in the right lobe of the liver outlined with air (fluid level) after the aspiration of more than 1 litre of pus; (c) lateral view, same patient as (b). (By courtesy of Professor S. Looareesuwan.)

Complications

Most complications involve extension of hepatic lesions into adjacent structures: usually the right chest, the peritoneum, and the pericardium. Upward extension usually produces adhesions between the liver, the diaphragm, and the lung; in consequence, subphrenic rupture and amoebic empyema are rare, although a right serous pleural effusion is not uncommon. Untreated, the disease process advances upwards through lung tissue leading to hepatobronchial fistula and expectoration of brownish, necrotic liver tissue, the so-called 'anchovy sauce' sputum. Rupture into the peritoneum can occur at any time; it is sometimes the mode of presentation of an amoebic liver abscess, the cause of peritonitis being discovered only at laparotomy. Amoebic pericarditis usually results from upward extension of a left-lobe liver lesion. Initially patients have retrosternal pain, a pericardial friction rub, or a serous effusion; later rupture produces cardiac tamponade. The diagnosis is most difficult when an underlying liver abscess was not suspected.

Less commonly the lesion extends through the skin producing a sinus and cutaneous lesion. The gut, stomach, vena cava, spleen, and kidney are occasionally involved by direct spread. Bloodborne spread to the lung produces a lesion resembling an isolated pyogenic lung abscess. Amoebic brain abscesses due to *E. histolytica* are rare; most are discovered after death (Fig. 3). Jaundice occurs when a large lesion compresses the common bile duct or when multiple lesions compress several intrahepatic bile ducts. Rupture into a major bile duct can cause haemobilia. Portal-vein compression occasionally produces portal hypertension and congestive splenomegaly.

Fig. 3 Metastatic brain abscess in a patient with an amoebic liver abscess. (By courtesy of Professor S. Looareesuwan.)

Differential diagnosis

Amoebic serology and scanning have now greatly simplified diagnosis. However, a few patients, generally less than 5 per cent, are initially seronegative; scanning patterns may be atypical before lesions have liquefied. Pyogenic abscess, especially when cryptogenic, may be clinically indistinguishable and this condition is quite common in some Asian countries. Other conditions to be distinguished are primary and secondary carcinoma of the liver, lesions of the right lung base and right pleura, subphrenic abscess, cholecystitis, septic cholangitis including that resulting from aberrant *Ascaris* worms, and liver hydatid cysts.

Needle aspiration of the liver (Fig. 4) may be necessary for diagnostic or therapeutic purposes (see below). Suspected pyogenic abscess is the main indication for the former; blood cultures should also be taken. Typically the aspirate in hepatic amoebiasis is pinkish-brown ('anchory sauce') (Plate 3), odourless, and bacteriologically sterile; a thinner, malodorous, or frothy aspirate suggests bacterial infection. A therapeutic amoebicide trial is generally preferable to diagnostic needling of the liver.

Fig. 4 Diagnostic/therapeutic aspiration of 'anchovy sauce pus' from a patient with amoebic liver abscess. (Copyright Professor D.A. Warrell.)

Cutaneous and genital amoebiasis

Skin ulceration due to *E. histolytica* produces deep, painful, and foul-smelling lesions that spread rapidly. Secondary bacterial infection is common and may mask the amoebic pathology. Lesions are most frequent in the perianal area, but also occur at colostomy stomas, laparotomy scars, and at the site of skin rupture by a hepatic lesion.

Female genital involvement results from faecal contamination, the extension of perianal lesions, or by the formation of internal fistulas from the gut, which can involve the bladder. Lesions of the vulva and uterine cervix may resemble carcinoma. Male genital lesions follow rectal coitus, the lesion beginning as a balanoposthitis and progressing rapidly.

Laboratory diagnosis

Microscopy and culture

The identification of live erythrocytophagous trophozoites in temporary wet mounts is of prime importance because it confirms the diagnosis of invasive amoebic disease. Amoebae should be sought in dysenteric bowel-wall scrapings, the last portion of aspirate from a liver abscess (Fig. 5), sputum, and tissue smears from skin lesions. In non-dysenteric stools, flecks of pus, blood, or mucus should be looked for and examined. The amoebae remain active for about 30 min at room temperature and so recently voided stool samples should be examined without delay ('hot stool'). Other microscopical features of faeces in amoebic colitis are scanty or absent leucocytes (methylene blue staining), clumped or degenerating red cells, and sometimes Charcot–Leyden crystals. If wet preparations are not made, or are negative, a portion of the specimen should be preserved in polyvinyl alcohol or SAF (sodium acetate–acetic acid–formalin) fixative for later smear preparation; alternatively, drying faecal smears should be fixed in Schaudinn's solution. In either case fixed smears should be stained with Gomori trichrome or Heidenhain's iron haematoxylin.

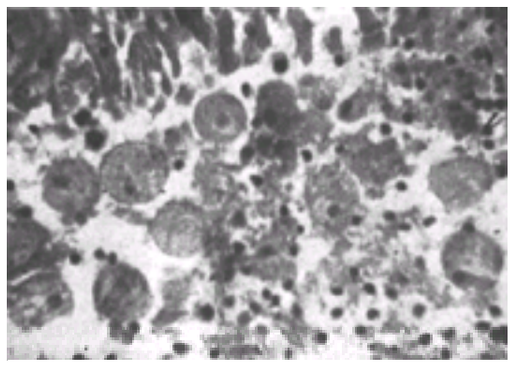

Fig. 5 Aspirate from amoebic liver abscess showing margin of hepatocytes and erythrocytophagous trophozoites of *E. histolytica*. (By courtesy of Professor S. Looareesuwan.)

Cysts and commensal trophozoites of *E. histolytica* found in wet faecal mounts are indistinguishable from those of *E. dispar*. The cysts of both species can be differentiated from the smaller *E. hartmanni* using an eyepiece micrometer. Direct mounts are made by emulsifying a small portions of stool in 1 per cent eosin, and in Lugol's iodine; however, the diagnostic sensitivity, per specimen, is only about 30 per cent. Concentration methods for cysts such as formol-ether sedimentation give a 70 per cent sensitivity per specimen. Cultivation of intestinal amoebae with bacterial associates in Robinson's medium is relatively easy; species identification requires immunofluorescent staining. Culture lysates provide material for zymodeme assay. Positive cultures from extraintestinal sites confirm invasive *E. histolytica*; amoebae are often difficult to find microscopically in liver aspirates.

Unless invasive trophozoites are found, differentiation of *E. histolytica* from *E. dispar* is only possible using zymodeme assay, or immunofluorescent staining of trophozoites in fixed faecal or amoebic culture smears.

Immunological tests

E. histolytica antigen can now be detected in faecal specimens and where this test is available it greatly simplifies diagnosis in both amoebic disease and in carriers; sensitivity and specificity of these tests is good. Assays for antigen in serum have also been used.

Many serodiagnostic methods have been applied to amoebiasis, most detectable antibody is IgG, with some IgM in active disease. However, seropositivity does not distinguish current and past tissue invasion. The more sensitive methods are indirect haemagglutination, enzyme immunoassay, and indirect immunofluorescence. Latex agglutination and gel-diffusion precipitation are also used, the former being commercially available as a slide test, taking only minutes to perform. Using sensitive tests, over 95 per cent of patients with liver abscess are seropositive, as are about 60 per cent of those with invasive bowel disease; patients with amoeboma are nearly all seropositive. All patients with tissue invasion eventually become seropositive. Titres decline after therapy but may remain positive for 2 years or more with the most sensitive tests.

Patient management

Chemotherapy

Nitroimidazoles are tissue amoebicides, and metronidazole for 5 days will be the first choice in most patients. The usual adult dose of metronidazole is 800 mg three times daily for 5 or 8 days; the paediatric dose is 35 to 50 mg/kg in three divided doses. An alternative is tinidazole, which has the advantage of a single daily dose of 2 g in adults and 50 to 60 mg/kg in children. A 5- or even a 3-day course may be sufficient for tissue amoebae but rates of parasitological cure may be low. When nitroimidazoles are contraindicated, or not available, erythromycin is useful in non-severe colitis.

The alkaloid emetine hydrochloride is a potent tissue amoebicide but has cumulative cardiotoxicity. Where appropriate nitroimidazoles are unavailable, as continues to be the case in many tropical contexts, this drug will continue to be life-saving, especially when a parenteral drug is needed. Emetine at 1 mg/kg daily (maximum 60

mg) by intramuscular injection for 5 days is usually sufficient; the synthetic derivative dehydoemetine hydrochloride is less toxic and more rapidly excreted in the urine, the daily intramuscular dose is 1.25 mg/kg (maximum 90 mg). Chloroquine is an effective alternative amoebicide in hepatic amoebiasis but is now little used; for adults a course of 150 mg of base twice daily is necessary.

Cutaneous and genital amoebiasis respond well to metronidazole, partly perhaps because these lesions often contain anaerobic bacteria. Amoebiasis at other sites is nearly always secondary to hepatic lesions and the chemotherapy will be the same. Metronidazole crosses the blood–brain barrier and should be used in the desperate situation of amoebic brain abscess due to *E. histolytica*.

Elimination of carrier state ('cyst'-passers)

All patients with *E. histolytica* infection treated with a tissue amoebicide should also be given diloxanide to eliminate all infection from the bowel and so prevent recurrence of tissue invasion or transmission to others. The dosage of diloxanide for adults is 500 mg three times daily for 10 days; the daily dose in children is 20 mg/kg in three divided doses.

Convalescent carriers should always be treated and also infected family contacts. Persons entering temperate countries from the tropics or new residents from such countries should be screened if there is a significant risk of infection; those with *E. histolytica* faecal antigen, or who are seropositive and have four-nucleated *Entamoeba* cysts in their stools, should be treated. In such cases, diloxanide is the drug of choice. Metronidazole is less effective, even using an 8-day course, and side-effects are troublesome. Unfortunately, cure rates with tinidazole are very low when followed up at 1 month.

Supportive and surgical management

Intestinal amoebiasis

Supportive management plays a major role in patients with complicated amoebic colitis, with emphasis on fluid and electrolyte replacement, gastric suction, and blood transfusion as necessary. Gut perforation in the context of extensive colitis carries a very poor prognosis; management may have to be medical. Parenteral metronidazole is invaluable in these contexts because of its activity against anaerobic bacteria in the peritoneum and bloodstream. Gentamicin plus a cephalosporin will normally be given as well.

Amoebomas respond well to metronidazole; a slow response should arouse suspicion that the amoebic lesion is superimposed upon other pathology, particularly a carcinoma. Surgical management is important in several situations. Acute colonic perforation in the absence of diffuse colitis, or ruptured amoebic appendicitis may be amenable to local repair. In the case of diffuse colitis, local repair, or end-to-end anastomosis, may not be possible because of the poor condition of the gut wall: temporary exteriorization with an ileostomy may be necessary. In fulminant colitis with multiple perforation the viability of the gut wall is uncertain and the only definitive option is total colectomy.

Hepatic amoebiasis

Parenteral metronidazole can be used in patients who undergo laparotomy. A favourable response to medical treatment alone can be expected in about 85 per cent of patients. Liver abscesses may rupture before, during, or after chemotherapy. Intra-abdominal rupture will always require laparotomy. Extension into the pleural or pericardial cavities necessitates drainage of these structures, together with aspiration of the liver lesion; pericardial drainage is most urgent when tamponade is present. Hepatopulmonary lesions generally require drainage of the liver lesion but medical treatment alone has been successful in some cases. Antimicrobials will always be needed when the abscess ruptures into the peritoneum or lung.

The most common management problem is slow response to the amoebicide. Patients whose pain and fever do not subside within 72 h are at significantly greater risk of rupture or therapeutic failure, and aspiration is generally to be recommended. A likely explanation of poor initial response is a tense lesion that restricts drug entry. Regular ultrasonographic monitoring is of great value as it will indicate the risk of rupture and guide the aspiration procedure. No change in lesion size on ultrasound can be expected during the first 2 weeks, although its outline may become clearer. Percutaneous aspiration with a wide-bore needle will be possible in most patients; if unsuccessful or anatomically contraindicated, then surgical help should be sought. Resolution time for small or moderate lesions is unaffected by aspiration. All patients with hepatic amoebiasis should be given a 10-day course of diloxanide to elimanate bowel infection.

Prognosis

Uncomplicated invasive intestinal disease and uncomplicated hepatic amoebiasis should normally have a mortality rate of less than 1 per cent. In complicated disease the mortality is much greater and may reach 40 per cent for amoebic peritonitis with multiple gut perforation. Prognosis is usually better n centres where the disease is common and more likely to be recognized early. Late diagnosis increases the probability of complicated disease and mortality rises accordingly.

Unless parasitological cure is achieved, and the gut completely freed of *E. histolytica*, clinical relapse is quite common, although probably limited by immunological responses. There is so far no evidence of naturally occurring strains of *E. histolytica* being resistant to normally used drugs. Hepatic scans show that nearly all liver abscesses completely disappear within 2 years; the median resolution time is 8 months. In secondarily infected lesions, bizarre hepatic calcification may be seen years afterwards. Healing of the bowel is remarkably rapid and complete; occasionally fibrous strictures persist after severe dysentery.

Prevention

Chlorination of water supplies does not destroy amoebic cysts, but adequate filtration will remove them. Regular stool screening of food handlers and domestic staff is of no value, but health education is important with encouragement to have a medical check if diarrhoea occurs.

Visitors to the tropics should not attempt chemoprophylaxis; in particular, long-term unsupervised use of hydroxyquinoline drugs must be strongly deprecated. Simple hygienic measures provide considerable protection. Boiling water for 5 min kills cysts. Routine examinations in temperate countries for returning visitors from the tropics or for new residents coming from such countries is of no value unless *E. histolytica* can be differentiated from *E. dispar*. Amoebic serology is particularly useful in those with gut symptoms or a history of dysentery.

Other parasitic gut amoebae including *Dientamoeba fragilis*

In addition to *E. histolytica* five species of *Entamoeba* infect humans, all have a nucleus with a small central endosome and abundant peripheral chromatin. *E. gingivalis* has no cystic stage and lives in the mouth within gingival pockets and tonsillar crypts. It is spread by kissing or more indirect oral contact. Its possible role in periodontal disease was formerly dismissed but there is now renewed interest following recognition of its high prevalence in individual lesions in people with this condition; it may act as a bacterial vector within the lesions. It has been found on intrauterine devices removed because of symptoms. Both in the uterus and in the mouth this amoeba occurs in association with the bacterium *Actinomyces israeli*.

The other *Entamoeba* species are non-pathogenic colonic commensals. *E. coli* has eight nuclei and is the commonest species in most surveys. *E. dispar* and *E. hartmanni* both have cysts with four nuclei; the former was previously known as 'non-pathogenic *E. histolytica*', and the latter as 'small race *E. histolytica*'; size is the only simple diagnostic criterion for *E. hartmanni*, its cysts are less than 10 µm in diameter. The global prevalence of *E. dispar* is about 10 per cent; even in the tropics the prevalence ratio of *E. dispar* to *E. histolytica* is often between 4:1 and 10:1. Lastly there is *E. polecki*, which is primarily a pig parasite; the cyst has one nucleus and an 'inclusion body'. Human infections are common in highland Papua New Guinea where humans and pigs may share a peridomestic environment; elsewhere it is rare.

Endolimax nana and *Iodamoeba buetchlii* both have nuclei with large endosomes and no visible peripheral chromatin. Cysts of the former are oval in shape with four nuclei; those of the latter are somewhat irregular in shape with a single nucleus and a large glycogen vacuole that stains prominently with iodine. Neither species is pathogenic.

Dientamoeba fragilis is overlooked in most parasitological laboratories and most reports are from developed countries. There is good evidence that it can cause

colonic inflammation; however, this is not severe and there is no ulceration or systemic spread. It has no cystic stage and unless this organism is specifically looked for it will be missed. In fixed stained smears, about 60 per cent of trophozoites have two nuclei; the endosome is large and lobulated without peripheral chromatin. Alternatively it may be identified in faeces or cultures using immunofluorescence with specific antibody. Transmission is believed to be nearly always within the eggs of the threadworm *Enterobius*. It causes a relatively mild diarrhoeal illness that may persist for several weeks, sometimes there is a superficial eosinophilic colitis. Blood eosinophilia is quite common and seropositivity is reported. This infection is common in some institutional contexts, but sudden outbreaks are not reported, presumably because of its mode of transmission. It is found within some resected appendices but a causal role is unlikely. Electron micrographs indicate that *D. fragilis* is an amoeboflagellate or a trichomonad rather than a true amoeba. The infection responds to metronidazole.

Free-living amoebae

Several species produce cytopathic changes in cultured cell monolayers and cerebral invasion after intranasal inoculation into mice and other animals. These amoebae are aerobic and their cytoplasm contains mitochondria, Golgi complexes, and a contractile vacuole; their natural food is bacteria. A shared feature is the very large central nuclear endosome, quite different from that of *E. histolytica*, from which differentiation may be necessary in tissue sections. Under dry conditions, trophozoites form resistant cysts that permit survival and also airborne dispersal; cysts can resist chlorination. Many species are thermophilic and they are one of the causes of 'humidifier fever', a form of extrinsic allergic alveolitis presenting with fever, cough, and sometimes progressive pulmonary fibrosis and dyspnoea. Some bacteria including *Legionella* and *Listeria* may live symbiotically within amoebae persisting within the phagosome, being resistant to lyzosomal enzymes. Surprisingly, *Legionella* can survive encystment: the amoebae provide a refuge for these bacteria when chlorination or other antibacterial measures are applied. The following three genera of free-living amoebae cause human infections.

Naegleria is an amoeboflagellate with two trophozoite forms. The amoeba moves rapidly with a single pseudopodium, it can transform into a non-feeding flagellate in hypotonic media and these free-swimming forms facilitate dispersal. Cysts are thin walled and spherical. It may be cultured aerobically on a confluent growth of *Escherichia coli*.

Acanthamoeba has no flagellate form. The small pseudopodia are multiple, thin, and spike-like (Fig. 6); they are called acanthopodia. Cysts are thick walled, angulated, and buoyant (Fig. 7); their dispersal may be wind-borne. It can be cultured on a confluent growth of *Escherichia coli.*

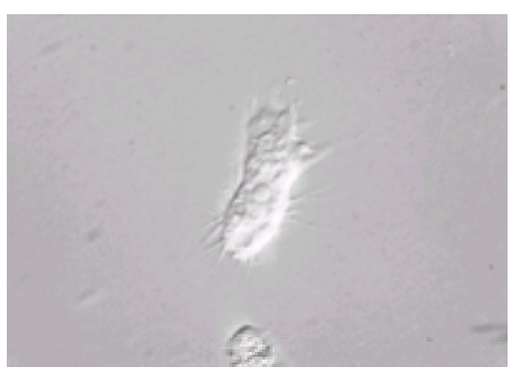

Fig. 6 Trophozoite of *Acanthamoeba* showing spike-like acanthopodia. (Copyright V. Zaman.)

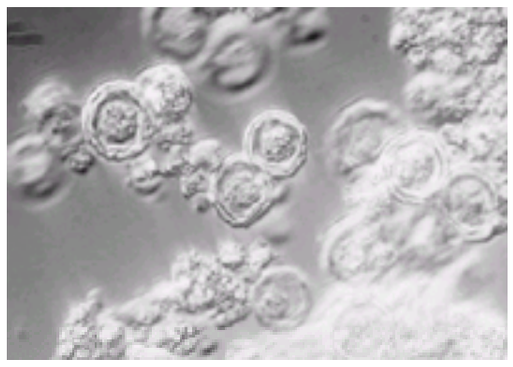

Fig. 7 Cysts of *Acanthamoeba.* (Copyright V. Zaman.)

Balamuthia is a leptomyxid amoeba; it shows little directional movement and has an irregular or branched shape. Cysts are thick walled and wrinkled. Human infections formerly attributed to *Hartmanella* are now all thought to be due to *B. mandrillaris*, a species described in 1993 from a mandrill baboon that died of meningoencephalitis in San Diego zoo. *Balamuthia* can only be cultured on tissue culture monolayers.

Primary amoebic meningoencephalitis due to *Naegleria fowleri*

Epidemiology and pathology

Nearly all patients give a history of swimming or diving in warm fresh water, or spa water, between 2 and 14 days before the illness began. Common-source outbreaks occur during warm summer months in temperate countries. Amoebic trophozoites cross the cribriform plate from the nasal mucosa to the olfactory bulbs and subarachnoid space. At autopsy the brain shows cerebral softening and damage to the olfactory bulbs; cysts are never formed in the tissues. So far only about 200 cases have been documented since the first human case was reported in 1965. However, some are missed clinically and discovered at autopsy, or in preserved pathological material. Specific antisera enable amoebae to be recognized by immunofluorescence staining.

Clinical features and diagnosis

Most patients are young adults and children. Initial nasal symptoms and headache are soon followed by fever, neck rigidity, coma, and later, convulsions; most die within a few days. Cerebrospinal fluid is often turbid, and bloodstained with high protein and low glucose levels and neutrophils. Amoebae must be urgently looked for in wet specimens using phase-contrast microscopy. Unless amoebae are seen, bacterial meningitis will be suspected; on Gram staining, amoebae appear as indistinct smudges. Fixed preparations stained with iron haematoxylin will show full details of nuclear structure. Confirmation is by culture at 37°C. Amphotericin B is the only effective drug. It should be given by daily intravenous infusion, and intrathecally, with the dosage regimens used for cryptococcal meningitis. So far, very few patients have survived, but this may partly be due to diagnostic delays.

Amoebic keratitis due to *Acanthamoeba*

Corneal lesions are painful and present as indolent and progressive ulcers leading eventually to perforation. Recognition may be in the context of lesions unresponsive to antibiotics or corticosteroids; differention must be made from herpes simplex. Inflammatory cells are mainly neutrophils. Infection may be by wind-borne cysts upon a damaged epithelium or from contact lenses. Solutions used to store, or wash, lenses can be contaminated by these amoebae, many of which are resistant to some antiseptics, especially as cysts.

Five species of *Acanthamoeba* are recognized to cause keratitis, the most common are *A. castellani* and *A. polyphaga.* Amoebae are found in corneal scrapings or histologically in corneal tissue but can be missed unless stained with iron haematoxylin or immunofluorescence using specific antisera. Cysts may be seen in tissue.

Cultures from fresh material should be at 30°C.

Lesions usually respond to local propamidine and neomycin, but the latter is not cysticidal; combinations of topical propamidine with chlorhexidine, or with polyhexamethylene, have recently been successful. Alternatives are topical miconazole and oral ketoconazole. Corneal grafting may be necessary. Wearers of contact lenses must take especial care to avoid contamination, especially when storage cases are used; raw tap water may contain *Acanthamoeba*. The most appropriate disinfectants are chlorhexidine and hydrogen peroxide.

Granulomatous amoebic encephalitis due to *Acanthamoeba*

Humans become infected by swallowing or inhaling cysts or amoebae; or these may contaminate wounds, or skin or mucosal ulcers. *Acanthamoeba* species are sometimes isolated from throat or nasal swabs, or from stool specimens.

Many patients have predisposing factors such as craniofacial trauma, vascular brain infarct, or a systemic disorder such as lymphoma, other malignancy, collagen disorder, alcoholism, or diabetes mellitus. Relatively acute cerebral lesions are described in a few patients with AIDS. Cerebral lesions arise haematogenously, by direct spread, or rarely from the nasal mucosa as with *Naegleria*. Pathologically, lesions resemble chronic bacterial brain abscesses or localized subacute haemorrhagic necrosis; involvement of the meninges is common. Some patients present with headache and meningism, others with evidence of a focal brain lesion. Primary lesions have been described from the lung, orbit and other cranial structures, and the gastric wall.

Unless these amoebae are found in wet-tissue preparations or cerebrospinal fluid, the diagnosis will be based upon histology. Cysts may be seen in tissue, but trophozoites may be missed unless stained with iron haematoxylin or immunofluorescence using specific antisera. Cultural diagnosis from fresh biopsies or cerebrospinal fluid is sometimes possible.

Survival of patients with this condition is very rarely reported. Total excision of cerebral lesions is occasionally possible. The drug sensitivities are poorly defined; a wide spectrum of resistance is common. Systemic amphotericin B or flucytosine will be the initial choice, but ketoconazole is an alternative.

Amoebic meningoencephalitis due to *Balamuthia mandrillaris* infection

Since 1990, when this condition was recognized in a non-human primate, more than 60 human cases have been reported, in the Americas, Europe, and Australia. Immunocompetent as well as immunocompromised patients may be infected. Exposure may be associated with contact with fresh water in pools. Cerebral lesions may be subacute and necrotizing, with prominent vasculitis, or chronic and granulomatous. Some patients have associated granulomatous facial lesions (Plate 4). Prolonged treatment with albendazole and itroconazole has proved effective in Peru. Diagnosis is made by finding amoebic trophozoites and cysts in infected tissue and by indirect immunofluorescence.

Further reading

Gut amoebae

Clark CG (1998). Amoebic disease: *Entamoeba dispar*, an organism reborn. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92**, 361–4.

Cuffari C, Oligny L, Seidman EG (1998). *Dientamoeba fragilis* masquarading as allergic colitis. *Journal of Paediatric Gastroenterology and Nutrition* **26**, 16–20.

Diamond LS, Clark CG (1993). A redescription of *Entamoeba histolytica* Schaudinn, 1903 (emended Walker 1911) separating it from *Entamoeba dispar* Brumpt, 1925. *Journal of Eukaryote Microbiology* **40**, 340–4.

Irusen EM *et al*. (1992). Asymptomatic intestinal colonization by pathogenic *Entamoeba histolytica* in amebic liver abscess: prevalence, response to therapy, and pathogenic potential. *Clinical Infectious Diseases* **14**, 889–93.

Jackson TF (1998). *Entamoeba histolytica* and *Entamoeba dispar* are distinct species; clinical, epidemiological and serological evidence. *International Journal of Parasitology* **28**, 181–6.

Martinez-Palomo A, ed. (1986). *Amoebiasis*. Elsevier, New York.

Ockert G (1990). Symptomatology, pathology, epidemiology, and diagnosis of *Dientamoeba fragilis*. In: Honigberg BM, ed. *Trichomonads parasitic in humans*, pp 395–410. Springer, New York.

Ravdin JI, ed. (1988). *Amebiasis. Human infection by* Entamoeba histolytica. Wiley, New York.

Ravdin JI (1995). Amebiasis. [Review.] *Journal of Infectious Diseases* **20**, 1453–64.

Ravdin JI, ed. (2000). *Amebiasis.* Imperial College Press, London.

Sachdev GK, Dhol P (1997). Colonic involvement in patients with amoebic liver abscess: endoscopic findings. *Gastrointestinal Endoscopy* **46**, 37–9.

Free-living amoebae

Carter RF (1972). Primary amoebic meningo-encephalitis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **66**, 193–208.

Denney CF *et al*. (1997). Amebic meningoencephalitis caused by *Balamuthia mandrillaris*: case report and review. *Clinical Infectious Diseases* **25**, 1354–8.

Harf C (1996). Amoebae in relationship with bacteria in their environment. In: Özcel MA, Alkan MZ, eds. *Parasitology for the 21st century*, pp 253–60. CAB International, Wallingford, UK.

Illingworth CD *et al*. (1995). *Acanthamoeba* keratitis: risk factors and outcome. *British Journal of Ophthalmology* **79**, 1078–82.

Visvesvara GS, *et al*. (1990). Leptomyxidameba, a new agent of amebic meningoencephalitis in humans and animals. *Journal of Clinical Microbiology* **28**, 2570–6.

Visvesvara GS, Schuster FL, Martinez AJ (1993). *Balamuthia mandrillaris*, N.G., N. Sp., agent of amebic meningoencephalitis in humans and other animals. *Journal of Eukaryote Microbiology* **40**, 504–14.

7.13.2 Malaria

*D. J. Bradley and D. A. Warrell**

Introduction

Malaria is the most important human parasitic disease globally, causing over 170 million clinical cases annually, of which over a million die, mostly in Africa. It has had large effects on the course of history and settlement in tropical regions. In recent years malaria has been subject to massive control efforts, with varying degrees of success but the disease has been resurgent for the last two decades. Resistance of falciparum malaria parasites to the main antimalarial drugs is now a serious problem in SE Asia. Malaria epidemics are an increasing problem. Malaria remains the dominant tropical vector-borne disease but, after decades of neglect, international interest in its control has recently revived.

Parasitology

There are over a hundred species of malarial parasite (*Plasmodium* spp.), but only four species have humans as their natural vertebrate host: *P. falciparum*, *P. malariae*, *P. vivax*, and *P. ovale* (Plate 1). Rare zoonotic infections have been recorded from non-human primate malarias such as *P. knowlesi, P. simium,* and *P. cynomolgi*.

Although each of the human malarias has distinguishing biological, morphological, and clinical characteristics (Table 1), their overall biology and lifecycles are similar (Fig. 1).

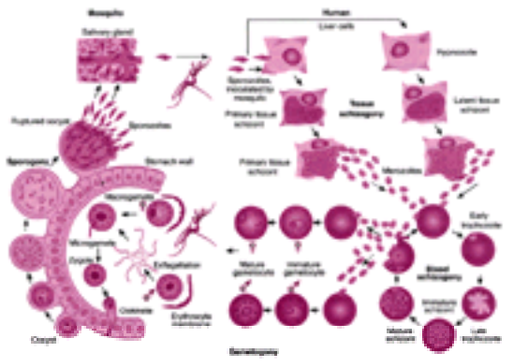

Fig. 1 Development cycle of *Plasmodium* spp. (redrawn by permission of F. Hoffman-la-Roche Ltd, Basel).

In both the mosquito and mammalian hosts the lifecycle of *Plasmodium* spp. has alternating stages of invasion and intracellular asexual division. The sexual stage, by facilitating the exchange of genetic information between different parasite strains or genotypes, assists in the generation of genetic diversity within the parasite population.

Infection is initiated when sporozoites from the salivary glands of a female *Anopheles* mosquito are inoculated during a blood meal into the human bloodstream. These organisms invade hepatic parenchymal cells within a few minutes, the process being largely complete within 30 min. Once inside the liver cell, two pathways of differentiation are possible.

In all species there is intracellular asexual multiplication. In addition, in *P. vivax* and *P. ovale* infections some parasites enter a cryptobiotic phase termed 'hypnozoites', which may lie dormant for months or even years before starting to divide and giving rise to late relapses. In *P. falciparum* and *P. malariae* infections there is no cryptobiotic phase and so relapses from the liver do not occur, although blood infections may persist for a few years in the case of *P. falciparum* or decades in the case of *P. malariae*.

The time required to complete the intrahepatic multiplication depends on the parasite species (Table 1). The products of the liver stage (extraerythrocytic merozoites) are liberated in their thousands into the bloodstream. Here they attach to and invade circulating erythrocytes. Inside the erythrocyte, asexual division begins and, over a period of 48 h (*P. falciparum, P. vivax, P. ovale*) or 72 (*P. malariae*), the parasites develop through a series of morphological changes from 'ring' forms to trophozoites and finally to schizonts containing daughter erythrocytic merozoites. These are liberated by red-cell lysis and immediately invade uninfected erythrocytes, producing a repetitive cycle of invasion and multiplication. Because the intraerythrocytic division cycle is usually fairly synchronous (particularly in *P. vivax* and *P. ovale* infections) and also tied to the diurnal cycle of the host, red-cell lysis and merozoite release occur at approximately the same time of day for a given individual. 'Malarial pyrogens' released at this time induce cytokine production (for example, tumour necrosis factor-a (**TNF-a**) and interleukin-1 (**IL-1**)) giving rise to the periodic 'agues' or paroxysms of fever that have long been a diagnostic feature of malaria infection. The asexual blood forms are the only forms of the parasite that give rise to clinical symptoms.

A small proportion of the merozoites within red cells develop into male and female gametocytes. Once mature, these gametocytes may return to the mosquito if ingested during a blood meal.

Inside the mosquito's midgut, male and female gametes are liberated from their host red cells and fuse to form a zygote which develops into an ookinete, able to penetrate the gut wall and form an oocyst. At this point a further series of asexual divisions takes place, giving rise to sporozoites that migrate to the insect's salivary glands to complete the lifecycle.

Genetics of the parasite

Variation has been found in isoenzyme types, antigenic markers, drug-resistance markers, and in the virulence of different isolates. Such genetic diversity has an important bearing on the disease, for an individual infection may consist of different parasite genotypes of varying drug resistance and exposure to a variety of antigenic types that may be needed before clinical immunity develops. During the intraerythrocytic cycle in the blood, the parasites are haploid. The diploid phase of the lifecycle occurs after gamete fusion in the mosquito where meiosis takes place .

Resistance to chloroquine and pyrimethamine results from mutations at unlinked loci. However, in the case of resistance to chloroquine alone, multiple mutations at independent loci may give rise to resistance. In some cases, resistance is stable in the absence of drug pressure. A locus on chromosome 7 segregates with the resistant phenotype.

Molecular biology

DNA from *P. falciparum* has proved to have an unusual base composition with an average A+T content of approximately 80 per cent. A very high proportion of genes, including 'housekeeping genes', contain large blocks of tandemly repeated amino acid sequences. It has been proposed that these sequences may act as immunological decoys by acting as T-independent antigens. The presence of multiple, low-affinity cross-reactivities between different repeats may serve to prevent the affinity maturation of specific B cells.

Comparison of rRNA sequences of different species of *Plasmodium* shows that *P. falciparum* is more closely related to avian malarias than to other mammalian species.

Proteins/antigens

The surface proteins of the sporozoite and sexual stages are dealt with in the section on vaccination below. Some of the molecules expressed on the surface of the merozoite, the invasive free-living form, are involved in the process of invasion of new erythrocytes. Other proteins bind specifically to the surface of normal red cells, but not to cells rendered refractory to invasion. Because red-cell invasion is an essential step in asexual parasite multiplication, understanding its molecular basis could lead to new forms of therapy.

Molecules on the surface of the infected red cell determine the adherence of infected cells to vascular endothelium and are a target of the protective immune response. Biochemical, immunochemical, and cell biological data reveal that a family of high molecular-weight molecules undergo a process of rapid, clonal, antigenic variation. The genes for this group of proteins (**PfEMP-1**, *Plasmodium falciparum* erythrocyte membrane protein-1), have recently been cloned.

Many other parasite-derived proteins are secreted into the host red cell, but do not find their way to the cell surface. Some interact specifically with the red-cell cytoskeleton modifying the host-cell environment in favour of the parasite.

In vitro culture

Since 1975 it has been possible to grow asexual forms of *P. falciparum* in long-term culture, using a suitable growth medium with uninfected red cells in an atmosphere of low oxygen and high carbon dioxide tension. The availability of large numbers of *P. falciparum* parasites, without recourse to patients or experimental non-human primates, has speeded up much of the basic research on malaria and has permitted the development of *in vitro* tests for sensitivity to certain antimalarial drugs. The complete development of the hepatic stage of *P. falciparum* has also been achieved *in vitro*. Gametocytes can also be produced from *in vitro* culture of asexual blood forms.

Biology of the mosquito vector

Human malarial parasites are transmitted only by Anopheles mosquitoes. There are many species with varying habits, breeding places, and effectiveness as malaria vectors. Anopheles can be distinguished from other adult mosquitoes by the way that the female, when taking a blood feed, inclines her whole body at an angle to her victim, while in the other, culicine mosquitoes, the body is parallel to the skin surface (Fig. 2). The culicine and anopheline larval stages are also distinguishable (Fig. 3).

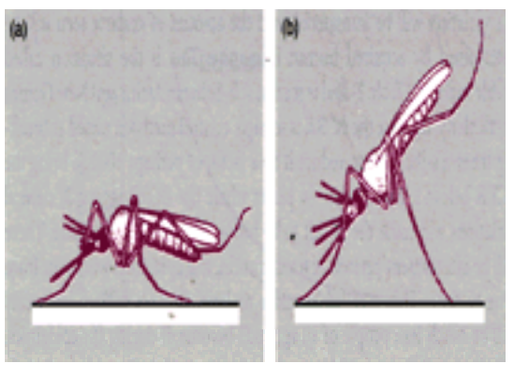

Fig. 2 Feeding posture of different types of adult mosquito. (a) Culicine, (b) anopheline.

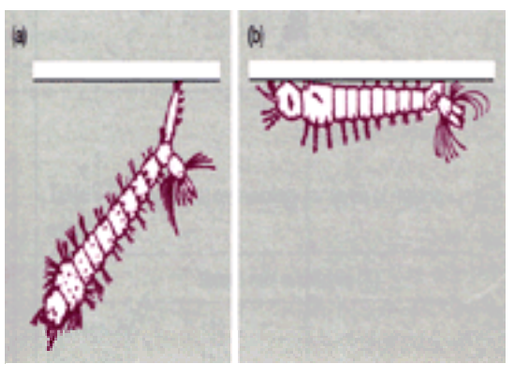

Fig. 3 Resting posture of different types of mosquito larva. (a) Culicine, (b) anopheline.

Since the female Anopheles needs a blood meal before egg laying, her adult life consists of finding a suitable blood meal, resting while it is digested, flying off to lay eggs at a suitable body of water, and then repeating this cycle every few days. The eggs, larvae, and pupae develop in water and the winged adults emerge. For ecological reasons, only a few species of Anopheles in a given locality are likely to be important malaria vectors, because to transmit malaria the mosquitoes need to be sufficiently abundant, to bite people rather than only some other vertebrate host, and to live long enough for ingested gametocytes to develop through to sporozoites. Identification of the main vector species in an area determines the design of specific control measures. Since most species are selective in their breeding sites, knowledge of the larval ecology permits engineering and other measures to be directed at the selective removal of the vector habitat, a process called 'species sanitation'.

The behaviour of the adult mosquito will dictate which insecticidal strategies are most likely to succeed. Anophelines vary in their preferred feeding and resting locations, though the majority bite in the evening and night. They may bite indoors (endophagic) or outside (exophagic). This determines whether the use of bednets and screened doors and windows will protect, or whether long sleeves and protective footwear when outside the house are more relevant. Of greater importance is where the female rests overnight to digest the blood meal. Endophilic mosquitoes, which rest on the inside walls of houses and in the roof, are thereby exposed to residual insecticides previously sprayed on the walls, whereas exophilic mosquitoes, resting outside houses, may escape the effects of insecticidal attack. The success of many antimalarial efforts has depended on the major vectors in several continents being endophilic, and failures of attempted eradication have sometimes resulted from exophilic vector species being present, as in many forested areas of SE Asia. Anopheline mosquitoes extend into temperate countries, and in the United Kingdom several indigenous species, notably *A. atroparvus*, were responsible for transmitting the historical English 'agues' (*P. vivax* and *P. malariae*).

Epidemiology

Malaria is widely distributed throughout the tropics (Fig. 4) except for the south-central Pacific islands from which anopheline mosquitoes are absent. *P. falciparum* is the predominant species in the highly endemic areas of Africa, New Guinea, and Haiti, while *P. vivax* is more common in Central America, North Africa, and southern and western Asia. Both species are prevalent in South America, the rest of Asia, and Oceania. *P. malariae* is widespread but often overlooked, and in West Africa *P. ovale* largely replaces *P. vivax,* to which the indigenous inhabitants are resistant.

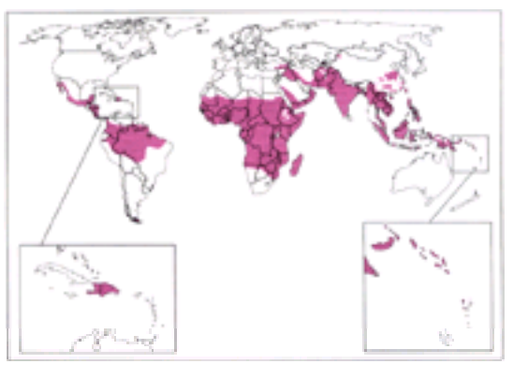

Fig. 4 Malarious areas of the world.

The epidemiological features of human malaria differ markedly even between endemic areas. At one extreme, as in tropical Africa, everyone is infected shortly after birth, parasitaemia is almost universal throughout childhood, and he brunt of mortality falls in early childhood; epidemics do not occur except at high altitude. By contrast, as in parts of India, malaria is an epidemic disease affecting all ages and causing temporary dislocation of community life due to the concurrent illness of the people. These differences result from differing levels of malaria transmission affecting the pattern of immunity in the human population, so that to understand even the clinical spectrum of malaria seen in patients from a given locality it is essential to understand the local epidemiological situation. The epidemiology of malaria is complex but relatively well understood. Attempts at control in recent years have changed the epidemiological pattern in many areas.

Climate and mosquito ecology are the primary determinants of malarial epidemiology. Once the biology of the relevant anopheline mosquito is understood, much of the complex epidemiology of malaria falls into place. There is some variation in susceptibility to malaria within the anophelines, so that *P. falciparum* from Africa may

fail to develop in some European anophelines even under optimal conditions, but usually in a given locality the indigenous anopheline mosquitoes will be capable of transmitting the local malaria strains, so that the importance of a vector species depends particularly on their behaviour and ecology.

The epidemiological pattern is determined by the density, human-biting habit, and longevity of the mosquito. Density is the number of vectors present in a locality relative to the human population. Malaria transmission will tend to be proportional to mosquito density, as might be expected. The human-biting habit combines two features: the frequency with which the female mosquito feeds and the choice of host. The human-biting frequency rises to as high as 0.5/day in *A. gambiae*, an African mosquito that feeds on alternate days and preferentially on people; while *A. culicifacies*, a vector in South Asia, may feed only every third day and as few as 10 per cent of its meals may be from people, giving a human-biting habit of 15-fold less. Because malaria transmission is proportional to the square of the human-biting habit, and as transmission involves both parasite uptake by bite and subsequent inoculation to human by a second bite, this factor has a large effect on malaria transmission.

Mosquito longevity has an even greater effect. The duration of the 'extrinsic cycle', the interval between when a mosquito ingests infective gametocytes and the first day on which sporozoites are present in the salivary glands ready for transmission, depends on the ambient temperature (Fig. 5), but it will rarely be less than 10 days. Only mosquitoes that become infected and then survive for longer than the duration of the extrinsic cycle (say 10 days) can pass on the infection. As mosquitoes of a given species have a relatively constant probability of dying during a day, regardless of their age, the longevity may be described by the probability of surviving through one day, and it varies greatly between mosquito species and environments. It will affect transmission very greatly indeed: if the chance of survival through one day is *p* and the duration of the extrinsic cycle *n* days, then transmission is proportional to *pn*, that is, something like the tenth power of *p*. Thus the most effective transmission of malaria will be by a long-lived mosquito that occurs at high density and frequently bites people. *A. gambiae* and *A. funestus* best fit this description well and are the predominant African malaria vectors.

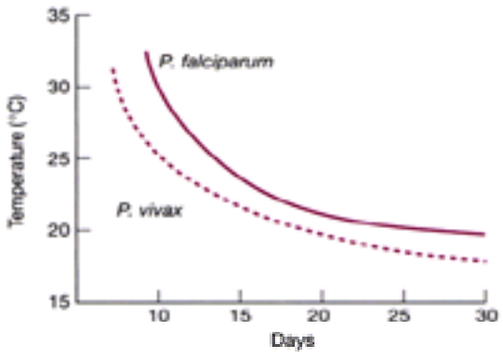

Fig. 5 The period of extrinsic development of *P. falciparum* and *P. vivax*.

Malaria transmission is most conveniently measured in terms of the basic case reproduction rate (**BCRR**). This is the average number of new cases of malaria that will result from one human case of malaria in a locality, assuming all the other people are non-immune and uninfected. The BCRR may vary from over 1000 in some areas of Africa to below 1. Where the BCRR is less than 1, the infection will not replace itself and the disease will die out. In the 'real world', the BCRR will vary considerably about a mean value. In areas with a very high BCRR everyone will become infected, the variation will be immaterial, and the amount of malaria seen will be determined by acquired human immunity. This is the situation called 'stable malaria' (Table 2) and is seen in sub-Saharan Africa and New Guinea particularly. Because the BCRR is so high, control methods aimed at breaking transmission have to reduce it by a factor of perhaps 1000 to bring the BCRR below 1. By contrast, in places where the BCRR is, say, 3, natural variations will cause the BCRR to be below 1 for much of the time. There will be intermittent periods of transmission, and epidemics will occur from time to time. This is called 'unstable malaria'. Because human immunity will be much less, people of all ages will become ill during the epidemics, but the transmission will be much easier to control. Unstable malaria is dramatically evident but kills fewer people than stable malaria, in which the brunt of the mortality falls on young children.

Even in stable malaria, seasonal variation may occur. In the African savannah, no mosquitoes may bite during the hot dry season and in more temperate zones it may be too cold for transmission for part of the year, but the annual peaks will be comparable, with all children infected each year. While the division between stable and unstable malaria is the most useful (Table 2), an earlier classification of areas by the parasite prevalence in children or by the proportion of children aged 2 to 9 years with enlarged spleens is often still used (Table 3). The prevalence of splenic enlargement gives a better cumulative picture of the amount of malaria than does the parasite prevalence, which is influenced by casual chemotherapy.

Under endemic conditions there is still a great deal of variation in risk to a non-immune visitor. At one extreme, in some parts of rural Africa an unprotected person is likely to be bitten on average by more than one infective mosquito nightly, whilst in a highly malarious part of India the corresponding rate is perhaps five times yearly or less. Yet both will be rightly perceived as highly malarious places by the local inhabitants.

The epidemiological background to clinical malaria is likely to change over time in most places due to environmental changes (often man-made, whether local or global), changing resistance of parasites to drugs, and the consequences of attempts at malaria control.

The widespread availability of chloroquine and other effective chemotherapeutic agents in endemic areas has resulted in the early treatment of a proportion of infections. This often leads to disparities between a high spleen rate in children and an artificially low parasite rate. With the increasing use of chemotherapy and of bed nets for personal protection, the acquisition of immunity is somewhat deferred. Under the most intense transmission, severe and often fatal anaemia in infants predominates. With rather less transmission, cerebral malaria in early childhood may be more apparent. Cerebral malaria will also be the main hazard in falciparum malaria epidemics and in non-immune individuals who contract it. Human migration is commonly associated with malaria epidemics, because population pressure in hilly areas drives the inhabitants down into malarious regions, or the aggregation of workers at new sites mixes infected people with those who are susceptible, or refugees may have impaired resistance to infection and public health measures may have collapsed. Migrants are commonly blamed for introducing malaria, but more usually they are non-immune individuals suffering from the disease acquired from the indigenous inhabitants.

Susceptibility to infection and innate resistance

People of West African origin are strikingly resistant to *P. vivax* infection. This correlates with the extreme rarity of the Duffy blood-group antigen alleles *Fya* and *Fyb*, which are receptors for penetration of the red cell by the merozoites.

Other genetic determinants affect the course and outcome of infection. Although *P. falciparum* is responsible for around 1 million deaths of African children annually, the mortality would be much greater but for a number of inherited resistance factors, and for the processes of acquired resistance discussed in the next section.

The high mortality associated with malaria is perhaps best illustrated by the way in which a number of otherwise disadvantageous genes have been selected in chronically exposed populations because of the resistance to malaria that they confer. In 1948, J.B.S. Haldane first suggested that heterozygotes of thalassaemia might be 'fitter than the normal [and] more resistant to attacks by the sporozoa that cause malaria'.

It has since become clear for several mutations affecting haemoglobin production or structure that these have reached their present frequencies by this selective mechanism. The best-known example is sickle-cell disease, due to a point mutation in position six of the b-globin chain. Here the mutant-gene frequency is stabilized because the enhanced survival of heterozygotes is counterbalanced by the lethal consequences of homozygosity in developing countries. Protection afforded to heterozygotes is seen most dramatically in case-control studies, which show that the relative risk of contracting severe malaria by heterozygotes and controls is about 1 to 10, respectively. Perhaps surprisingly, parasite rates and densities at the population level are very similar in normal and AS individuals, except in very young children, indicating that heterozygotes are resistant to disease rather than to infection.

Table 4 lists the genotypes for which there is either epidemiological or clinical evidence of selection by malaria. Despite this often clear evidence of protection, the

mechanisms involved are still controversial.

Acquired resistance

Those exposed to repeated malarial infection in endemic areas gradually acquire immunity in several stages, but it is rarely complete. Immunity is species-specific and largely strain-specific. The first change observed is a reduction in clinical symptoms and signs for a given level of parasitaemia. This is sometimes known as 'tolerance' but the mechanism is not understood.

Acquired resistance to the parasites takes months to develop and first affects the density of gametocytes in the peripheral blood. Subsequently, the density of asexual erythrocytic parasites, trophozoites, and schizonts falls and gradually reaches very low levels, so that under conditions of holoendemic transmission the prevalence of infection falls by half in those aged 15 years compared with children. Infected older children and adults from highly endemic areas often have very low-level, persistent, asymptomatic parasitaemias combined with relative resistance to superinfection.

It is clear that in a highly endemic area for *P. falciparum* there are several parasite strains circulating, and concepts of why resistance is so slowly acquired may either emphasize a balance in immune responses or the successive infection with various strains combined with a largely strain-specific response. The latter is favoured at present. Severe malaria in very young children is ascribed to multiple infections over a short time, and the cerebral malaria that predominates in slightly older children is possibly due to some more virulent strains.

Infants born to immune mothers are partially protected against severe malaria attacks by transplacental antibodies and those acquired from breast milk, for a few months, after which the infant suffers from severe malaria attacks with only gradual acquisition of resistance. Adult non-immune people, including visitors to the tropics from non-malarious areas, are equally susceptible to high mortality in their first few attacks, while women from an endemic area become more susceptible during pregnancy (second trimester), especially the first pregnancy. Splenectomy, for any reason, also increases susceptibility to malaria, which may have a fatal outcome.

Immunity is stage-specific, in that immunity to either sporozoite challenge or to gametocyte transmission does not protect against asexual parasites. It also has components that are specific for the parasite species, strain (genome), and antigenic variant within a strain. Thus protection against infection by sporozoites appears to be mediated largely by cytotoxic T cells, which can kill infected hepatocytes, although antibody to the repeat regions of the circumsporozoite protein may also have a role. Specific T and B cells, in addition to non-antigen-specific mechanisms, are involved in the control of asexual parasitaemia. The central role of antibody has been demonstrated by a variety of passive transfer experiments in people and experimental animals. Pooled immunoglobulin from highly immune donors is extremely effective in rapidly reducing parasitaemia, but not in the long-term control of infection. Maintaining parasite numbers below subclinical levels requires the involvement of T cells, as shown by adoptive transfer experiments in animals. Work in rodents suggests that early in infection, TH1 cells are critical but that later during the course of infection cells of a TH2 phenotype are more important. High levels of cytokines, such as TNF-a, during acute infection are a feature of severe malaria and are associated with a poor outcome. The acute response in non- or semi-immune individuals, while vital to the control of parasitaemia, may also contribute to the pathogenesis of disease by triggering a variety of non-specific effector mechanisms.

Clinical immunity takes many years and several infections to be effectively induced. It is also a non-sterilizing response, as immune adults are constantly and demonstrably reinfected. Several explanations have been put forward for these observations. Generalized, parasite-induced immunosuppression certainly does occur and may be clinically relevant in the response to certain non-malarial antigens such as meningococcal vaccine. It is not clear, however, whether overall it is any more severe than in other acute viral or bacterial infections. One area where it does seem to play a definite part is in the development of Burkitt's lymphoma, in which case it has been shown that individuals with acute *P. falciparum* infection have impaired T-cell control of endogenous, Epstein–Barr virus-infected, B-cell proliferation. Other explanations for the difficulty in inducing effective immunity have involved interference by the parasite in the development of specific responses. This could either be by the presence of important T-independent antigens, which induce a relatively poor response with no memory, or by the effect of crossreactive, tandem-repeat elements in inhibiting affinity maturation of specific B cells. Perhaps the most likely explanation is the extreme polymorphism or clonal variation of immunologically relevant antigens, such that the host requires exposure to a variety of 'strains' before a broadly effective response can develop. If the latter is true, then it presents formidable problems to vaccine development.

Malaria and HIV-immunosuppression

For at least 20 years, falciparum malaria and HIV-immunosuppression have coexisted in Africa, but the effects of their interaction on the clinical course of both diseases is still uncertain. Early studies that suggested an association between HIV and falciparum malaria in African children were vitiated by the greater risk of HIV-contaminated blood transfusions in children with severe malaria-related anaemia in this region. However, *P. falciparum* parasitaemia has been found to be more frequent among HIV-1-infected multigravidae than controls in malaria holoendemic areas of Central/Southern Africa, associated with increased perinatal mortality. Both HIV and malaria contribute to maternal anaemia and low birth-weight babies and, in Lusaka, Zambia, a dramatic increase in maternal morbidity over the last 20 years has been attributed to malaria, HIV-immunosuppression, AIDS-associated tuberculosis, and chronic respiratory illnesses. Recently, in a study of 613 adults with microscopically confirmed falciparum malaria in an area of unstable transmission in KwaZulu Natal, South Africa, the results suggested that underlying HIV infection might double the risk of severe malaria and increase the risk of fatal malaria by five- to sevenfold.

Molecular pathology

All the pathology associated with malaria infection is attributable to asexual parasite multiplication in the bloodstream. No adverse effects are caused by the quantitatively small degree of parasite invasion and multiplication within hepatocytes, nor by the presence of relatively small numbers of circulating gametocytes. The consequences to the host of the intraerythrocytic multiplication of parasites range from a variety of severe, but not life-threatening, symptoms common to all the species that infect humans, to the potentially lethal complications associated with acute *P. falciparum* infection and the chronic renal damage caused by some infections with *P. malariae*. The relative severity of *P. falciparum* infections, as well as the ability to culture this parasite *in vitro*, has meant that it has been the focus of most experimental effort.

It had been noted for centuries that malaria was characterized by periodic fevers. Once the causative organism was identified, it was clear that the bouts of fever generally followed the synchronous release of new merozoites into the bloodstream as each cycle of erythrocytic multiplication was completed. While it was assumed that the release of infected cell contents that occurs at this time was responsible for fever induction, it was not until very recently that it was proven that components of the infected cell such as the lipid, glycosyl phosphatidyl inositol anchor of a parasite membrane protein (perhaps **MSP-1**, merozoite surface protein-1) could directly induce the release of cytokines such as TNF-a and IL-1 from macrophages. Moreover, it was demonstrated that the older stages of parasites within erythrocytes were differentially sensitive to physiological increases in temperature, so that the effect of fever was both to limit parasite multiplication and to maintain synchronous development. Measurements of TNF-a in children suffering from severe malaria also demonstrated that very high levels of this cytokine were associated with a lethal outcome, although the correlation was not sufficient to be a useful prognostic indicator.

The principal life-threatening complications of *P. falciparum* in African children are cerebral malaria and severe anaemia often associated with metabolic acidosis and respiratory distress. The clinical picture in non-immune adults is more complex and can include single or multiple organ failure. Mechanisms responsible for severe malarial anaemia are poorly understood, but include parasite-induced dyserythropoiesis and accelerated red-cell clearance of both normal and infected cells by both immune and non-immune mechanisms (see below). However, the central event underlying the pathology of most other manifestations of severe falciparum malaria is the cytoadherence and resulting sequestration of infected cells, which is unique to this organism (Fig. 6 and Fig. 7 and Plate 2). Only the younger developmental stages of the parasite circulate, as the more mature forms adhere to specific receptors on venular endothelium. The distribution of infected cells found in tissue sections suggests that the chief sites of infected cell sequestration correlate with specific organ dysfunction. It is assumed, but not formally proven, that the reduction in, or obstruction of, local blood flow associated with the partial occlusion of small vessels with infected cells results in reduced perfusion and tissue damage. Some have suggested that the sequestered cells may induce the local release of a number of potentially toxic or pharmacologically active compounds (such as reactive oxygen species or nitric oxide) from macrophages, neutrophils, or endothelium, and that these may affect tissues locally.

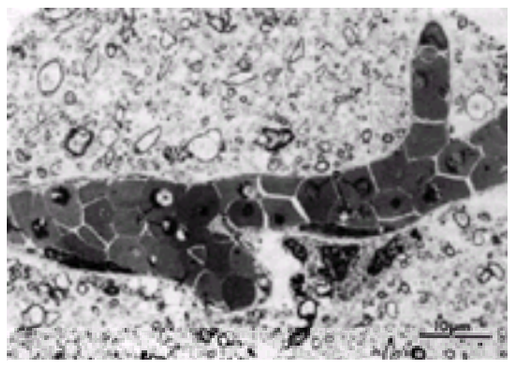

Fig. 6 Brain section of a patient who died of cerebral malaria, showing a blood vessel packed with red blood corpuscles, the majority of which were identifed as being infected by the presence of parasites (P), or at a higher magnification, the presence of knobs (by courtesy of Dr D. Ferguson, Oxford).

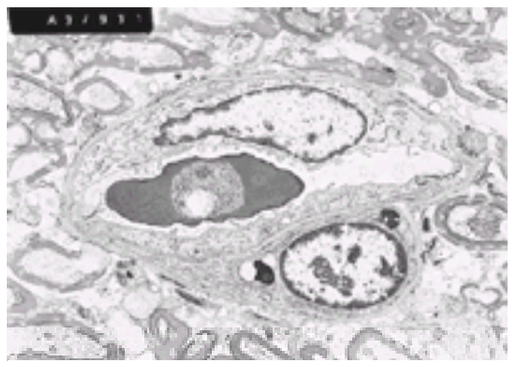

Fig. 7 Human cerebral malaria. Electron micrograph showing endothelial cell microvilli making contact with a parasitized erythrocyte via electron-dense strands (upper right). (Copyright N. Francis.)

While the detailed mechanisms by which sequestered cells result in the specific symptoms seen in cases of severe falciparum malaria remain largely unresolved, much more progress has been made in understanding the molecular interactions that lead to sequestration. Several endothelial receptors have been identified, including CD36 (formerly platelet glycoprotein IV), thrombospondin, intercellular adhesion molecule-1 (**ICAM-1**), and more recently, vascular cell-adhesion molecule (**VCAM**) and E-selectin (Table 5). Most field isolates bind to CD36 and thrombospondin and the majority bind to ICAM-1; to date there are insufficient data on VCAM and E-selectin. *In vitro* assays of purified proteins reveal great variability in the absolute levels of adhesion between isolates. No good correlation has yet emerged between the ability of parasites to bind to individual receptors and disease pattern. These studies are, however, fraught with difficulty and many more data are needed to resolve this point. The parasite molecules involved in adhesion are well characterized biochemically but have not yet been cloned, so that primary-sequence data are not available. They form a family of red-cell surface proteins that undergo clonal antigenic variation during a single infection and that also appear to be targets of a host-protective antibody response.

Some parasite isolates show two other properties: the rosetting of uninfected erythrocytes around red cells containing mature developmental forms of the parasite (Fig. 8) and autoagglutination of infected erythrocytes in the absence of immune serum. Rosetting has been linked to cerebral malaria in some, but not all, studies. It is presumed that the multicellular aggregates, if they occur *in vivo,* may exacerbate vascular obstruction caused by sequestration.

Fig. 8 Rosetting *in vitro*. The central parasitized erythrocyte shows many electron-dense protruberances (knobs) beneath its membrane (bar = 1 µm). (Copyright D. Ferguson.)

Despite the life-threatening nature of severe falciparum malaria, and the enormous number of childhood deaths that it causes in sub-Saharan Africa, the mortality rate of all malaria infections is extremely low. In holoendemic areas, infections in children are universal and constant, yet only a small proportion of those infected show clinical symptoms at any one time and only a fraction of these go on to develop severe illness. This is probably only partially explained by the known innate resistance factors and acquired immunity, and so it is likely that unidentified factors are also important in determining how far individual infections progress from parasitaemia to clinical illness and finally to severe disease.

Pathology

Brain

Only falciparum malaria causes cerebral pathology. At autopsy, the brain is sometimes oedematous but evidence of cerebral, cerebellar, or medullary herniation is rarely seen. The small blood vessels, including those of the leptomeninges, are congested with parasitized red blood cells containing malaria pigment (Fig. 6 and Fig. 7). This gives the surface of the brain its characteristic leaden or plum-coloured appearance and the cut surface a slatey-grey hue. Many of the parasites are schizonts and other mature forms. In larger vessels, parasitized red cells form a layer along the endothelium ('margination'). Up to 70 per cent of erythrocytes in the cerebral vessels are parasitized and these are more tightly packed than in other organs. The cerebrovascular endothelium shows pseudopodial projections, which may be in close apposition to electron-dense, knob-like protruberances on the surface of parasitized red cells (Fig. 7). Numerous petechial haemorrhages are seen in the white matter, resulting from haemorrhages from end arterioles, proximal to occlusive plugs of parasitized red cells and fibrin. Focal ring haemorrhages can be found centred on small subcortical vessels. Dürck's granulomas, small collections of microglial cells surrounding an area of demyelination, may develop at the site of these haemorrhages, but an inflammatory cell response is generally lacking.

Bone marrow

There is evidence of iron sequestration, erythrophagocytosis, dyserythropoiesis, and cytoadherence with plugging of sinusoids in the acute phase of falciparum malaria. Maturation defects are present in the marrow for at least 3 weeks after clearance of parasitaemia. Increased numbers of large, abnormal-looking megakaryocytes have been found in the marrow and the circulating platelets may also be enlarged, suggesting dyspoietic thrombopoiesis. Malaria pigment and

parasites can be found in monocytes and phagocytes in the marrow, even when they are not detectable in peripheral blood.

Liver

The liver is affected by all four species of human malaria parasites, but changes are most severe in falciparum malaria. The liver is enlarged and oedematous, and coloured brown, grey, or even black as a result of malaria pigment deposition. Hepatic sinusoids are dilatated, containing hypertrophied Kupffer cells and parasitized red cells that appear to obstruct the circulation. Parasitized and uninfected red cells are phagocytosed by Kupffer cells, endothelial cells, and sinusoidal macrophages. Small areas of centrilobular necrosis, which are occasionally seen in severe cases, may be attributable to shock or disseminated intravascular coagulation. Hepatocytes usually show only mild abnormalities but may be depleted of glycogen in some patients who are hypoglycaemic. Lymphocytic infiltration of portal tracts has been described (see also Tropical splenomegaly syndrome, below).

Gastrointestinal tract

Cytoadherent, sequestered, parasitized red cells may be found in the small and large bowel, especially in capillaries of the lamina propria and larger submucosal vessels. The bowel may appear congested, with mucosal ulceration and haemorrhage.

Kidney

Glomerular lesions range from the acute transient glomerular nephritis of falciparum malaria to the chronic lesions of quartan malarial nephrosis. In severe falciparum malaria, with or without 'blackwater fever', acute renal failure is associated with the histopathological changes of acute tubular necrosis. Parasitized red cells may be found in glomerular and peritubular capillaries, with fibrin thrombi and pigment-laden macrophages. Tubular pigment casts are prominent in cases of blackwater fever.

Lung

The lungs are oedematous in almost all patients dying of malaria. Pulmonary capillaries and venules are packed with inflammatory cells including neutrophils, plasma cells, and pigment-laden macrophages, and with parasitized red cells. The vascular endothelium is oedematous, causing narrowing of the capillary lumen, and there is interstitial oedema and hyaline-membrane formation. Secondary bronchopneumonia is a common finding.

Spleen

The spleen is large, engorged, and dark-red or greyish-black in colour. The red and white pulp is congested and hyperplastic, and the splenic cords and sinuses are filled with phagocytic cells containing pigment, parasitized red cells, and non-infected red cells. Tropical splenomegaly syndrome is described below.

Heart

There is no evidence of myocarditis. Subendocardial and epicardial petechial haemorrhages are unusual. The myocardial capillaries are congested with parasitized red cells, pigment-laden macrophages, lymphocytes, and plasma cells. However, the parasitized cells are not tightly packed and there is no evidence of cytoadherence.

Pathophysiology

Anaemia

This is attributable mainly to the destruction or phagocytosis of parasitized red cells, but other mechanisms contribute. The bone marrow shows dyserythropoietic changes. Initial iron sequestration and hypoferraemia may be explained by the very marked hyperferritinaemia, an acute-phase reaction. There is evidence of immune-mediated haemolysis in some populations. Erythrocyte survival is reduced even after the disappearance of parasitaemia. Increased splenic clearance of non-parasitized as well as parasitized red cells has been demonstrated.

Intravascular haemolysis occurs in patients whose erythrocytes are congenitally deficient in enzymes such as glucose 6-phosphate dehydrogenase (**G6PD**) in response to oxidant drugs such as primaquine. However, in classical blackwater fever, G6PD levels are, by definition, normal and the mechanism of haemolysis is unknown, although quinine-mediated haemolysis has been suspected.

Thrombocytopenia

This is attributable to sequestration in the spleen, failure of production by the marrow, and immune-mediated lysis.

Cerebral malaria

Mechanical obstruction to the microcirculation of the brain by cytoadherent, parasitized red cells, and perhaps 'rosettes' of uninfected red cells stuck around a parasitized red cells, is thought to be the principal mechanism leading to coma. Red blood cells infected with some strains of *P. falciparum* develop adhesive properties as they mature. Parasite-derived protein such as PfEMP-1 expressed on the surface of the parasitized red cell may act as a ligand that binds to receptors such as ICAM-1 on cerebral venular endothelium (Table 5). The expression of ICAM-1, and some other receptors involved in the cytoadherence of parasitized red cells, may be increased by TNF-a and other cytokines. Obstruction to cerebral blood flow could result in 'stagnant anoxaemia', leading to coma. In Thai adults with cerebral malaria, it was found that global cerebral blood flow was inappropriately low and there was evidence of cerebral anaerobic glycolysis with increased lactate concentrations in the cerebrospinal fluid. In African children with cerebral malaria, plasma concentrations of TNF-a, IL-1a, and other cytokines correlate closely with disease severity, as judged by parasitaemia, hypoglycaemia, case fatality, and the incidence of neurological sequelae. As well as enhancing cytoadherence, cytokines may have other effects on cerebral function, perhaps by releasing nitric oxide, which interferes with neurotransmission, or by leading to the generation of free oxygen radicals. Cytokines may also be responsible for fever, hypoglycaemia, coagulopathy, dyserythropoiesis, and leucocytosis in falciparum malaria.

In SE Asian adults, the opening pressure of cerebrospinal fluid at lumbar puncture was usually normal and cerebral oedema was demonstrable (by computed tomography (**CT**) scanning) during life in only a small minority, and usually as an agonal phenomenon. In these patients there was little evidence that brain swelling contributed to coma. However, in African children with cerebral malaria, intracranial pressure is usually elevated and there is evidence of brain swelling in the majority of those examined by CT scan. Ischaemic damage resulting from a critical reduction in cerebral perfusion pressure and other factors such as hypoglycaemia and status epilepticus are thought to be important in the mechanism of brain damage in these children.

Pulmonary oedema

This may develop in patients who have been overloaded with fluid in hospital and have elevated central venous and pulmonary-artery wedge pressures. More commonly, the clinical picture is of adult respiratory distress syndrome, with normal or low hydrostatic pressures in the pulmonary vascular bed. In these cases, the mechanism is likely to be increased pulmonary capillary permeability resulting from leucocyte products and cytokines. The histological appearances of neutrophil sequestration in the pulmonary capillaries, increased permeability, and hyaline membrane formation are consistent with this hypothesis.

Hypoglycaemia

This can be caused by cinchona alkaloids (quinine or quinidine), which are potent stimulators of insulin secretion by the pancreatic b-cells. The resulting reduction in hepatic gluconeogenesis and increased peripheral glucose uptake by tissues results in hypoglycaemia. In malaria, glucose consumption is increased by fever, infection, anaerobic glycolysis, and the metabolic demands of the malaria parasites. Glycogen reserves may be depleted, especially in children and pregnant women, as a result of fasting and 'accelerated starvation'. In African children with severe malaria, adult patients with severe disease, and pregnant women, hypoglycaemia develops spontaneously (without treatment with cinchona alkaloids) and is associated with appropriately low plasma insulin concentrations. Plasma lactate and alanine concentrations are elevated and ketone bodies are moderately increased. Counter-regulatory hormone levels are usually very high. The mechanism of

hypoglycaemia in these cases may be inhibition of hepatic gluconeogenesis by TNF-a and other cytokines.

Acute renal failure

Hypovolaemia, from dehydration, is responsible for acute renal failure in the majority of patients whose acute oliguria and renal dysfunction is reversible by fluid replacement. Hyperparasitaemia, jaundice, and haemoglobinuria are associated with a high risk of acute tubular necrosis. Renal cortical perfusion is reduced during the acute stage of the disease. Renal cortical necrosis must be rare, as survivors rarely show evidence of chronic renal impairment. Cytoadherence of parasitized red blood cells in the renal microvasculature, deposition of fibrin microthrombi, and prolonged hypotension ('algid malaria') may contribute to acute renal failure. Quartan malarial nephrosis is discussed below.

Hyponatraemia

In patients with relatively normal plasma osmolalities, hyponatraemia has been attributed to the inappropriate secretion of ADH triggered by fever or reduced effective plasma volume. However, the levels of ADH were appropriately high in Thai patients who were proved to be grossly hypovolaemic by carefully monitored fluid-repletion studies. Mild hyponatraemia is often attributable to intravenous therapy with 5 per cent dextrose alone in patients who are salt-depleted and dehydrated.

Hypovolaemia and 'shock' ('algid malaria')

This may result from hypovolaemia (dehydration and, rarely, haemorrhagic shock following splenic rupture or gastrointestinal haemorrhage) but is most often associated with a secondary Gram-negative bacteraemia. The source may be an intravenous cannula, urethral catheter, or aspiration pneumonia. Transient immunosuppression, impaired macrophage function, or 'blockade' of the reticuloendothelial system may increase the susceptibility of patients to severe secondary bacterial infections.

Clinical features

The pathogenic species of *Plasmodium* cause acute febrile illnesses characterized by periodic febrile paroxysms occurring every 48 or 72 h, with afebrile asymptomatic intervals and a tendency to recrudesce or relapse over periods of months or even years. The severity of the attack is determined by the species and strain, and hence the geographical origin, of the infecting parasite; on the age, genetic constitution, state of immunity, general health, and nutritional state of the patients, and on their use of antimalarial drugs.

Falciparum malaria ('malignant' tertian or subtertian malaria)

The shortest interval between an infecting mosquito bite and parasitaemia is 5 days, but this prepatent period is usually 9 to 10 days. The incubation period (the interval between infection and the first symptom) usually ranges from 7 to 14 days (mean 12 days) but may be prolonged further by immunity, chemoprophylaxis, or partial chemotherapy. In Europe and North America, 98 per cent of patients with imported falciparum malaria present within 3 months of arriving back from the malarious area. A few present up to 1 year later, but none after more than 4 years.

Several days of prodromal symptoms such as malaise, headache, myalgia, anorexia, and mild fever are interrupted by the first paroxysm. Suddenly the patient feels inexplicably cold (in a hot climate) and apprehensive. Mild shivering quickly turns into violent shaking with teeth-chattering. There is intense peripheral vasoconstriction and gooseflesh. Some patients vomit. The rapid increase in core temperature may trigger febrile convulsions in young children. The rigor lasts up to 1 h and is followed by a hot flush with throbbing headache, palpitations, tachypnoea, prostration, postural syncope, and further vomiting while the temperature reaches its peak. Finally, a drenching sweat breaks out and the fever defervesces over the next few hours. The exhausted patient sleeps. The whole paroxysm is over in 8 to 12 h, after which the patient may feel remarkably well. These symptoms are typical of a classical 'endotoxin reaction' produced by typhoid vaccine, infection with Gram-negative bacteria, or the release of TNF-a and other cytokines by other agents. Classical tertian or subtertian periodicity (48 and 36 h between fever spikes) is rarely seen with falciparum malaria. A high irregularly spiking, continuous or remittent fever, or daily (quotidian) paroxysm, is more usual. Other common symptoms are headache, backache, myalgias, dizziness, postural hypotension, nausea, dry cough, abdominal discomfort, diarrhoea, and vomiting. The non-immune patient with falciparum malaria usually looks severely ill, with 'typhoid' facies and, in dark-skinned races, a curious greenish complexion. Commonly, there is anaemia and a tinge of jaundice, with moderate tender enlargement of the spleen and liver. Useful negative findings are the lack of lymphadenopathy and rash (apart from herpes simplex 'cold sores') and focal signs.

Cerebral malaria and other severe manifestations and complications

The global case fatality of falciparum malaria is probably around 1 per cent or 1 to 3 million deaths per year. Cerebral malaria is the most important of the severe manifestations of *P. falciparum* infection, accounting for 80 per cent of these deaths. Patients who have been feverish and ill for a few days may have a generalized convulsion from which they do not recover consciousness, or their level of consciousness may decline gradually over several hours. High fever alone can impair cerebral function causing drowsiness, delirium, obtundation, confusion, irritability, psychosis, and, in children, febrile convulsions. The term 'cerebral malaria', implying encephalopathy specifically related to *P. falciparum* infection, should be restricted to patients in an unrousable coma (no appropriate verbal response and no purposive motor response to noxious stimuli—Glasgow Coma Scale £9/14) and evidence of acute *P. falciparum* infection, in whom other encephalopathies, including hypoglycaemia and transient postictal coma, have been excluded. Patients with cerebral malaria may have mild meningism but neck rigidity and photophobia are rare. Retinal haemorrhages (Plate 3) are present in about 15 per cent of African and SE Asian cases, but exudates are rare. (In Papua New Guinea these changes are not confined to patients with severe falciparum malaria.) Papilloedema is very rare (0.5 per cent of cases). Dysconjugate gaze is common. In adult patients the pupillary, corneal, oculocephalic, and oculovestibular reflexes are normal. Muscle tone and tendon reflexes are usually increased and there is ankle clonus. The plantar responses are extensor and abdominal reflexes are absent. In African children, brainstem reflexes may be abnormal and there may be neurological evidence of severe intracranial hypertension with rostrocaudal progression suggesting cerebral, cerebellar, and medullary herniation. Hypotonia is more common than in adults. In patients of all ages, abnormal flexor or extensor posturing (decerebrate or decorticate rigidity), associated with sustained upward deviation of the eyes (not the transient upward gaze of oculogyric crisis), pouting and stertorous breathing, is sometimes, but not always, associated with hypoglycaemia (Fig. 9(a) and Fig. 9(b)). About half of adult patients and more children have generalized convulsions. In children, seizures may be subtle and detectable only as twitching of the facial muscles, deviation with nystagmus of the eyes, irregularities of breathing, and sometimes posturing of one arm. Less than 5 per cent of adult survivors have persisting neurological sequelae; these include cranial-nerve lesions, extrapyramidal tremor, and transient paranoid psychosis. However, more than 10 per cent of African children who survive an attack of cerebral malaria suffer from sequelae such as hemiplegia, cortical blindness, epilepsy, ataxia, and mental retardation.

Fig. 9 (a) and (b) Extensor posturing (decerebrate rigidity) in a Thai woman with cerebral malaria and profound hypoglycaemia (Copyright D. A. Warrell).

Anaemia (see above) is an inevitable consequence of all but the mildest infections. It is most common and severe in pregnant women, children (Plate 4), and in

patients with high parasitaemia, schizontaemia, secondary bacterial infections, and renal failure.

Spontaneous bleeding, from the gums (Plate 5) and gastrointestinal tract, is seen in less than 5 per cent of adult patients with severe malaria. It is rare in children.

Jaundice (Plate 6) is common in adults but rare in children. Biochemical evidence of severe hepatic dysfunction is unusual. Hepatic failure suggests concomitant viral hepatitis or another diagnosis.

Hypoglycaemia is being increasingly recognized in patients with malaria. Pregnant women with severe or uncomplicated falciparum malaria and other patients with severe disease may become hypoglycaemic a few hours to 6 days after starting treatment with quinine or quinidine, even after the parasitaemia has cleared. Pregnant women and children with malaria, and other patients with hyperparasitaemia and complicating bacteraemias, may all become hypoglycaemic early in their illness and without quinine therapy. The symptoms and signs of hypoglycaemia—anxiety, tachycardia, breathlessness, feeling cold, confusion, sweating, light-headedness, restlessness, fetal bradycardia, other signs of fetal distress, coma, convulsions, and extensor posturing—may be misinterpreted as merely manifestations of malaria.

Hypotension and shock ('algid malaria') is seen in patients who develop pulmonary oedema, metabolic acidosis, complicating bacteraemias, and massive gastrointestinal haemorrhage. Mild supine hypotension with a marked postural drop in blood pressure is usually attributable to vasodilatation and relative hypovolaemia. Cardiac arrhythmias are rare but may be precipitated by rapid infusion or excessive doses of antimalarial drugs such as chloroquine, quinine, or quinidine. Patients with coronary insufficiency may develop angina during febrile crises of malaria. Patients with severe malaria sometimes develop complicating bacterial infections such as aspiration pneumonia, urinary-tract infections, infected bedsores, and phlebitis at intravenous drip sites.

Oliguria, with increased blood urea and serum creatinine concentrations, is seen in about one-third of patients with severe malaria. Although most of these patients respond to cautious rehydration, 10 per cent develop renal failure requiring dialysis.

In patients whose red blood cells are deficient in G6PD (and other enzymes), intravascular haemolysis and haemoglobinuria (Plate 7) may be precipitated by oxidant antimalarial drugs, especially primaquine, whether or not they have malaria. Classical blackwater fever is the association of haemoglobinuria with severe manifestations of falciparum malaria—including renal failure, hypotension and coma—in a non-immune patient who is not G6PD deficient.

Metabolic acidosis is seen in association with hyperparasitaemia, hypoglycaemia, and renal failure. Usually it results from lactic acidosis, even in patients with renal failure. In African children, respiratory distress manifested as deep (Kussmaul) breathing, associated with severe anaemia and metabolic acidosis, is emerging as a syndrome, and which carries a higher mortality than cerebral malaria.

Pulmonary oedema (Fig. 10) appears to be the terminal event in most fatal cases of falciparum malaria in adults. It may develop late in the clinical course as a result of fluid overload or in patients with severe disease in the absence of fluid overload. It may also appear suddenly after delivery in pregnant women who are in positive fluid balance. The earliest sign is an increase in respiratory rate. Pulmonary oedema may be difficult to differentiate from aspiration pneumonia, a common complication in comatose patients, and metabolic acidosis. Radiography may be needed to make this distinction with confidence. The patients who are not fluid-overloaded resemble those with adult respiratory distress syndrome with a normal jugular venous, central venous, or pulmonary-artery wedge pressure.

Fig. 10 Pulmonary oedema in a Vietnamese woman with cerebral malaria (copyright D. A. Warrell).

Cerebellar dysfunction

A rare presentation of falciparum malaria is cerebellar ataxia with unimpaired consciousness. Similar signs may be seen in patients recovering from cerebral malaria and, in Sri Lanka, delayed cerebellar ataxia has been described 3 to 4 weeks after an attack of fever attributable to falciparum malaria. Complete recovery is the rule.

Malarial psychosis

Acute psychiatric symptoms in patients with malaria may be attributable to their drug treatment, including antimalarial drugs such as chloroquine, mefloquine, and the obsolete mepacrine, and to exacerbation of pre-existing functional psychoses. However, in some patients, organic mental disturbances associated with malaria infection have been the presenting feature or, more often, have developed during convalescence after attacks of otherwise uncomplicated malaria or cerebral malaria. Depression, paranoia, delusions, and personality changes should probably be classified as brief reactive psychoses. These symptoms rarely last for more than a few days.

Vivax, ovale, and malariae malarias

The prepatent and incubation periods are given in Table 1. Some strains of *P. vivax*, especially those from temperate regions (*P. v. hibernans* from Russia, *P. v. multinucleatum* from China) may have very long incubation periods (250–637 days). Only about one-third of imported cases of vivax malaria present within a month of returning from the malarious area; 5 to 10 per cent will present more than a year later.

The 'benign' malarias cause paroxysmal, feverish symptoms no less hectic and distressing than those of falciparum malaria. Prodromal symptoms are said to be more severe with *P. malariae* infection. In untreated cases, the characteristic tertian (48–50 h) interval between fever spikes may be seen with *P. vivax* and *P. ovale* and the quartan (72 h) pattern in *P. malariae* infections.

This periodicity is established after several days of irregular fever. Vivax and ovale malarias have a persistent hepatic cycle, which may give rise to relapses every 2 to 3 months for 5 to 8 years in untreated cases. *P. malariae* does not relapse but a persisting, undetectable parasitaemia may cause recrudescences for more than 50 years.

Although symptoms may be severe and temporarily incapacitating, especially in non-immune individuals, the acute mortality is very low. For example, during the 1967 to 1969 Sri Lankan epidemic of predominantly vivax malaria, there were more than half a million cases with a case fatality of only 0.1 per cent. Only in immunocompromised, splenectomized or debilitated patients are the 'benign' malarias likely to prove life-threatening. However, acute pulmonary oedema has been documented in several cases of vivax malaria in non-immune travellers (Fig. 11).

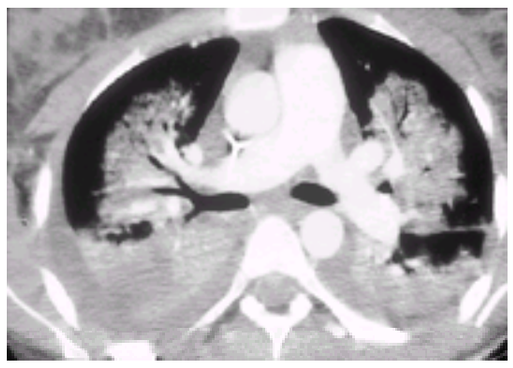

Fig. 11 Pulmonary oedema in a 20-year-old woman with vivax malaria and bilateral pleural effusions (copyright D. A. Warrell).

An important practical point is that indigenous West Africans are very rarely infected with *P. vivax*. Patients suffering from vivax malaria may become anaemic, thrombocytopenic, and mildly jaundiced with tender hepatosplenomegaly. Splenomegaly may be particularly gross in areas of *P. malariae* infection (see also Tropical splenomegaly syndrome, below). In debilitated patients with vivax malaria, anaemia rarely may be severe enough to be life-threatening. Splenic rupture, which carries a mortality of 80 per cent, is said to be more common with vivax than falciparum malaria. It results from acute, rapid enlargement of the spleen, with or without trauma; chronically enlarged spleens are less vulnerable. A ruptured spleen presents with abdominal pain and guarding, haemorrhagic shock (tachycardia, postural hypotension, and prostration), fever, and a rapidly falling haematocrit. These features may be misattributed to malaria itself.

Cerebral vivax malaria has occasionally been reported especially with the long incubation period strain in China. Mixed falciparum infection or another encephalopathy must be adequately excluded in such cases. The same strictures apply to cerebral *P. malariae* malaria, especially as this parasite coexists with *P. falciparum* throughout most of its range. The acute symptoms of ovale and malariae malarias may be as severe as those of vivax infection, but anaemia is less severe and the risk of splenic rupture is lower. *P. ovale* causes negligible mortality, but *P. malariae* causes many deaths from nephrotic syndrome (see below).

Malaria in pregnancy and the puerperium

Malaria is a major cause of maternal anaemia, death, abortion, stillbirth, premature delivery, low birth weight, and neonatal death in those areas of the tropics where malaria transmission is unstable and women of childbearing age have little acquired immunity. Even in some hyperendemic areas, clinical symptoms and parasitaemia are worse in primiparous than in multiparous women and other patients. In non-immune individuals, cerebral and other forms of severe falciparum malaria are more common in pregnancy. In the great epidemic of falciparum malaria in Sri Lanka during 1934 to 1935 the mortality among pregnant women was 13 per cent, twice that in non-pregnant women, and in Thailand, where malaria has been the most important cause of maternal mortality, cerebral malaria in late pregnancy had a mortality of 50 per cent. In some parts of Africa, one-quarter to one-half of all placentas are parasitized. The incidence is highest in primiparae. Changes in humoral and cell-mediated immunity in pregnancy do not explain this vulnerability, but it is clear that the placenta is a privileged site for parasite multiplication. An adhesion receptor for some strains (genotypes) of *P. falciparum*, chondroitin sulphate A, is expressed on the surface of the syncytiotrophoblast. This may explain sequestration in the placenta.

In most endemic regions, birth weights of neonates born to women with malaria are significantly less than those of controls. Fetal distress was observed in 6 out of 12 Thai women with malaria who were beyond the twenty-ninth week of pregnancy. Painless uterine contractions were detected in seven out of eight who were not in labour. This uterine activity subsided as the patients' temperatures were reduced by simple cooling.

Special risks to the mother of malaria during pregnancy are hyperpyrexia, hypoglycaemia, anaemia, cerebral malaria, and pulmonary oedema.

Severe anaemia, exacerbated by malaria, is an important complication of pregnancy in many tropical countries. Especially in communities where chronic hookworm anaemia is prevalent, high output anaemic cardiac failure may develop in late pregnancy.

Asymptomatic hypoglycaemia may occur in pregnant women with malaria before antimalarial treatment, and pregnant women with severe uncomplicated malaria are particularly vulnerable to quinine-induced hypoglycaemia (see above).

There is an increased risk of pulmonary oedema precipitated by fluid overload or by the sudden increase in peripheral resistance, or autotransfusion of hyperparasitaemic blood from the placenta, which occurs just after delivery (Fig. 10).

Prevention

Malaria is so dangerous in pregnancy that pregnant women who cannot leave the area of transmission must be given intermittent preventive treatment with sulfadoxine–pyrimethamine or antimalarial prophylaxis extending into the early puerperium. This is a most important part of antenatal care.

Congenital and neonatal malaria

Vertical transmission of malaria can be diagnosed by detecting parasitaemia in the neonate within 7 days of birth, or later if there is no possibility of postpartum, mosquito-borne infection. Save for a few discordant reports, most evidence from malarious parts of the world indicates that congenital malaria is rarely symptomatic, despite the high prevalence of placental infection. This confirms the adequacy of protection provided by IgG from the immune mother, which crosses the placenta to active immunization from exposure to soluble malarial antigens *in utero* and to the high proportion of fetal haemoglobin in the neonate, which retards parasite development. Congenital malaria is, however, much more common in infants born to non-immune mothers, and there is an increased incidence during malaria epidemics. It can cause stillbirth or perinatal death. All four species can produce congenital infection, but because of its very long persistence *P. malariae* causes a disproportionate number of cases in non-endemic countries. Fetal plasma quinine and chloroquine concentrations are about one-third of the simultaneous maternal levels. Thus, antimalarial concentrations that are adequate to cure the mother might result in subtherapeutic concentrations in the fetus. Quinine and chloroquine are excreted in breast milk, but the suckling neonate would receive only a few mg/day. Maternal hypoglycaemia, a common complication of malaria or its treatment with quinine, may produce marked fetal bradycardia and other signs of fetal distress.

Differential diagnosis

The clinical features of congenital malaria include fever, irritability, feeding problems, hepatosplenomegaly, anaemia, and jaundice. Unless parasites are found in a smear from a heel prick or cord blood, the patient may be misdiagnosed as having rhesus incompatibility or another congenital infection such as cytomegalovirus, herpes simplex, rubella, toxoplasmosis, or syphilis.

Transfusion malaria, 'needlestick', and nosocomial malaria

Malaria—like trypanosomiasis, Colorado tick fever, HIV, hepatitis viruses, and some other pathogens—can be transmitted in blood from apparently healthy donors. Exceptionally, donors may remain infective for up to 5 years with *P. falciparum* and *P. vivax,* 7 years with *P. ovale,* and 46 years with *P. malariae.* Because the infecting forms are erythrocytic (not sporozoites), no exoerythrocytic (hepatic) cycle will be established and so vivax and ovale malarias will not relapse. Theoretically, parasitaemia might be detectable immediately and hence the incubation period should be shorter than with mosquito-transmitted malaria. However, the incubation period tends to be longer because of the time needed to build up parasitaemias sufficient to cause symptoms. Mean incubation periods are 12 (range 7–29) days for *P. falciparum*, 12 (range 8–30) days for *P. vivax*, and 35 (range 6–106) days for *P. malariae*. Whole blood, packed cells (blood products), leucocyte or platelet concentrates, fresh plasma, marrow transplants, and haemodialysis have been responsible for transfusion malaria. As patients requiring transfusion are likely to be debilitated and may be immunosuppressed, and there may be a long delay before making the diagnosis because malaria is not suspected, unusually high parasitaemias may develop with *P. falciparum* and *P. malariae.* With *P. ovale* and *P. vivax* infections, the parasitaemia is usually limited to 2 per cent because only reticulocytes are invaded. Severe manifestations are common, mortality may be high, for example 8 out of 11 in a group of heroin addicts, and even acute *P. malariae*

infections may prove fatal.

Nosocomial malaria has resulted from contamination of saline used for flushing intravenous catheters, contrast medium, and intravenous drugs. Malaria has complicated parenteral drug abuse.

Prevention

Outside the malaria endemic area, donors who have been in the tropics during the previous 5 years should be screened for malarial antibodies (indirect fluorescent antibody) (see below). In the endemic area, recipients of blood transfusions can be given antimalarial chemotherapy, or at least should be watched carefully for evidence of infection.

Monkey malarias

Human erythrocytes can be infected with at least six species of simian plasmodia. There have been rare cases of natural infections or accidental laboratory infections by *P. brazilianum, P. cynomolgi, P. inui, P. knowlesi, P. schwetzi* and *P. simium*. Severe feverish and systemic symptoms have been described, but no cerebral or other severe complications. No patient has died. Parasitaemia may remain undetectable for 2 to 6 days after the start of symptoms. Periodicity is quotidian (*P. knowlesi*) or tertian (*P. simium* and *P. cynomolgi*). Infectivity and virulence may be enhanced by repeated passage in humans. Chloroquine is the treatment of choice.

Diagnosis

Malaria can present with a wide range of symptoms and signs, none of them diagnostic. **It must be excluded by repeated thick and thin blood smears in any patient with acute fever and an appropriate history of exposure.** Until malaria is confirmed or an alternative diagnosis emerges, smears should be repeated every 8 to 12 h. **However, if the patient is severely ill, or the symptoms persist or deteriorate, a therapeutic trial of antimalarial chemotherapy must not be delayed.** Antimalarial chemoprophylaxis should be stopped while the patient is under investigation for malaria, as this may make microscopical diagnosis more difficult. Patients should be asked about travel to malaria endemic countries during the previous year. The possibility of malaria must not be dismissed because the patient took prophylactic drugs, for none is completely protective. Short airport stopovers, even on the runway, or working in or living near an international airport, may allow exposure to an imported, infected mosquito. Transmission by blood transfusion, 'needlestick', or nosocomial infection should be borne in mind. Those who grew up in an endemic area will probably lose their immunity to disease after living for a few years in the temperate zone and become vulnerable when they return to their homeland on holiday. In malaria endemic regions, a large proportion of the immune population may have asymptomatic parasitaemia and it cannot be assumed that malaria is the cause of the patient's symptoms even if parasitaemia is detected. The diagnosis of malaria may be missed, even in the endemic zone, during an epidemic of some other infection (for example, meningitis, pneumonia, cholera).

Differential diagnosis (Table 6)

Malaria should be considered in the differential diagnosis of any acute febrile illness until it can be excluded by a definite lack of exposure, by repeated examination of blood smears, or by a therapeutic trial of antimalarial chemotherapy. In Europe and North America, imported malaria has been misdiagnosed as influenza, viral hepatitis, viral encephalitis, or travellers' diarrhoea, sometimes with fatal consequences. Cerebral malaria must be distinguished from other infective meningoencephalitides. Cerebrospinal fluid (**CSF**) examination will identify most of these infective causes (see Chapter 24.14.1). Abdominal reflexes are brisk in patients with psychotic stupor and hysteria but absent in cerebral malaria. Recognition of poisoning will depend largely on the history or the clinical circumstances. Overdose of antimalarial drugs (chloroquine and quinine) can be confused with cerebral malaria. Intravenous drug abusers are at risk both from severe malaria and drug overdose. Alcoholism may be confused with cerebral malaria, whether the patient presents simply as 'drunk', with delirium tremens, or encephalopathy.

Misdiagnosis of a viral haemorrhagic fever in a case of imported malaria is potentially dangerous, for patients may be placed in a high-containment unit where they may be denied basic investigations such as examination of a blood smear because of a fear of infection. Jaundice is a common feature of yellow fever, but not other viral haemorrhagic fevers.

Malaria in pregnancy may be confused with viral hepatitis, acute fatty liver with liver failure or eclampsia, and in the puerperium with puerperal sepsis or psychosis.

Laboratory diagnosis

Microscopy

It is most important to confirm the diagnosis by examining thick and thin blood films on several occasions (Plate 1). Parasites may be found in blood taken by venepuncture, finger-pulp or ear-lobe stabs, and from the umbilical cord and impression smears of the placenta. In fatal cases, cerebral malaria can be confirmed rapidly as the cause of death by making a smear from cerebral grey matter obtained by needle necropsy through the foramen magnum, superior orbital fissure, ethmoid sinus via the nose, or through a fontanelle in young children. Sometimes no parasites can be found in peripheral blood smears from patients with malaria, even in severe infections. This may be explained by partial antimalarial treatment or by sequestration of parasitized cells in deep vascular beds. In these cases, parasites or malarial pigment may be found in a bone marrow aspirate. Pigment may be seen in circulating neutrophils. A number of Romanowski stains, including Field's, Giemsa, Wright's, and Leishman's, are suitable for malaria diagnosis. The rapid Field's technique, which can yield a result in minutes, and Giemsa are recommended. Smears may be unsatisfactory because the slides are not clean; stains are unfiltered, old, or infected; the buffer pH is incorrect (it should be 7.0–7.4); drying is too slow, especially in a humid climate (producing heavily crenated erythrocytes); or the blood has been stored in anticoagulant causing lysis of parasitized erythrocytes. It is difficult to make a good smear if the patient is very anaemic. Common artefacts resembling malaria parasites are superimposed platelets, particles of stain and other debris, and pits in the slide. Other erythrocyte infections such as bartonellosis and babesiosis may be misdiagnosed as malaria. Parasites should be counted in relation to the total white-cell count (on thick films when the parasitaemia is relatively low) or erythrocytes (on thin films). An experienced microscopist can detect as few as 5 parasites/µl (0.0001 per cent parasitaemia) in a thick film and 200/µl (0.004 per cent parasitaemia) in a thin film.

Fluorescent microscopy

Becton-Dickinson's QBC (quantitative buffy coat) method involves spinning blood in special capillary tubes in which parasite DNA is stained with Acridine Orange and a small float presses the parasitized red blood cells against the wall of the tube where they can be viewed by ultraviolet microscopy. In expert hands, the sensitivity of this method can be as good as with conventional microscopy of thick blood films but species diagnosis is difficult, and the method is much more expensive.

Malarial antigen detection

Becton-Dickinson's 'Para Sight F' and ICT Diagnostics' 'ICT Malaria Pf' dipstick antigen-capture assays employ monoclonal antibody detecting *P. falciparum* histidine-rich protein-2 (**PfHRP-2**) antigen. These tests are rapid (taking about 20 min), sensitive, and specific for *P. falciparum*. A number of other, species-specific, antigen-detection methods are now marketed, such as OptiMAL (Flow Laboratories) which detects parasite lactate dehydrogenase.

Other methods

Enzyme and radioimmunoassays, DNA probes (using chemoluminescence for detection), and polymerase chain reaction (**PCR**) methods now approach the sensitivity of classical microscopy. They take much longer (up to 72 h), are much more expensive, and are unlikely to replace microscopy for routine diagnosis. However, some of these newer methods could be automated for screening blood donors or for use in epidemiological surveys and, in the case of PCR, identification of parasite strains as well as species is possible.

Serological techniques

Malarial antibodies can be detected by immunofluorescence, enzyme immunoassay, or haemagglutination for epidemiological surveys, for screening potential blood donors, and occasionally for providing evidence of recent infection in non-immune individuals. These tests are not useful in making an acute diagnosis of malaria. In

future, detection of protective antibodies will be important in assessing the response to malaria vaccines (see below).

Other laboratory investigations

Anaemia is usual, with evidence of haemolysis. Serum haptoglobins may be undetectable. The direct antiglobulin (Coombs') test is usually negative. Neutrophil leucocytosis is common in severe infections whether or not there is a complicating bacterial infection, but the white-cell count can also be normal or low. The presence of visible malarial pigment in more than 5 per cent of circulating neutrophils is associated with a bad prognosis. Thrombocytopenia is common in patients with *P. falciparum* and *P. vivax* infections; it does not correlate with severity. Prothrombin and partial thromboplastin times are prolonged in up to one-fifth of patients with cerebral malaria. Concentrations of plasma fibrinogen and other clotting factors are normal or increased, and serum levels of fibrin(ogen) degradation products are normal in most cases. Fewer than 10 per cent of patients with cerebral malaria have evidence of disseminated intravascular coagulation. However, antithrombin III concentrations are often moderately reduced and have prognostic significance. Total and direct (unconjugated) plasma bilirubin concentrations are usually increased, consistent with haemolysis, but in some patients with very high total bilirubin concentrations there is a predominance of conjugated bilirubin, indicating hepatocyte dysfunction. Some patients have cholestasis. Serum albumin concentrations are usually reduced, often grossly. Serum aminotransferases, 5'-nucleotidase, and especially lactic dehydrogenase are moderately elevated, but not into the range seen in viral hepatitis. Hyponatraemia is the most common electrolyte disturbance. Mild hypocalcaemia (after correction for hypoalbuminaemia) and hypophosphataemia have been described, especially when the patient has been given blood or a glucose infusion. Biochemical evidence of generalized rhabdomyolysis (elevated serum creatine kinase concentration, myoglobinaemia, and myoglobinuria) has been found in some patients. In about one-third of patients with cerebral malaria, the blood urea concentration is increased above 80 mg/dl (13 mmol/l) and serum creatinine above 2 mg/dl (176 µmol/l). Lactic acidosis occurs in severely ill patients, especially those with hypoglycaemia and renal failure. It may be suspected if there is a wide 'anion gap'. Blood glucose must be checked frequently, especially in children, pregnant women, and severely ill patients, even if the patient is not receiving quinine treatment and is fully conscious. A 'stix' method, with or without photometric quantification, is rapid and convenient. Microscopy and culture of cerebrospinal fluid is important in patients with cerebral malaria to exclude other treatable encephalopathies. In cerebral malaria the cerebrospinal fluid may contain up to 15 lymphocytes/µl and an increased protein concentration. Pleocytosis of up to 80 cells/µl, mainly leucocytes, may be found in patients who have had repeated generalized convulsions. The CSF glucose level will be low in hypoglycaemic patients and this result may be the first hint of hypoglycaemia. In view of the finding of cerebral compression and high opening pressures in many African children with cerebral malaria, some paediatricians prefer to delay lumbar puncture, while covering the possibility of bacterial meningoencephalitis with empirical antimicrobial treatment. Blood cultures should be performed in patients with a high white-cell count, shock, persistent fever, or an obvious focus of secondary bacterial infection. Gram-negative rod bacteria (*E. coli, Pseudomonas aeruginosa*, etc.) have been cultured from the blood of adult patients with 'algid' malaria. In Gambian children an association was found between malaria and non-typhoid salmonella septicaemia.

Urine should be examined by microscope and dipstix. Common abnormalities are proteinuria, microscopic haematuria, haemoglobinuria, and red-cell casts. The urine is literally black in patients with severe intravascular haemolysis. Urine specific gravity should be measured: the optical method is most convenient when urine output is small. Rapid measurement of plasma quinine or quinidine concentrations is possible in some hospitals. This is a valuable way of monitoring chemotherapy.

Treatment

Antimalarial drugs

Antimalarial drugs can be grouped as follows:

1. arylaminoalcohols, comprising quinoline methanols such as the cinchona alkaloids, quinine and quinidine (extracted from the bark of the cinchona tree), mefloquine, halofantrine, and lumefantrine;
2. 4-aminoquinolines, such as chloroquine and amodiaquine;
3. folate-synthesis inhibitors, including type 1 antifolate drugs, which compete for dihydropteroate synthase (e.g. sulphones and sulphonamides), and type 2 antifolate drugs, which inhibit malarial dihydrofolate reductase (e.g. the biguanides, proguanil and chlorproguanil, and the diaminopyrimidine, pyrimethamine);
4. 8-aminoquinolines, such as primaquine and tafenoquine (Etaquine, WR238,605);
5. antibiotics, such as tetracycline, doxycycline, clindamycin, azithromycin, and fluoroquinolones;
6. peroxides (sesquiterpene lactones)—artemisinin (qinghaosu) derivatives from the Chinese medicinal plant, *Artemisia annua,* and its semisynthetic analogues (artemether, arteether, artesunate, and artelinic acid); and
7. naphthoquinones, such as atovaquone (BW566C80).

The stages of the lifecycle sensitive to some of the principal antimalarial drugs are shown in Fig. 12. Among blood schizonticides, artemisinin derivatives can prevent the development of rings or trophozoites, but quinine and mefloquine cannot stop development before the stage of mature trophozoites, and pyrimethamine–sulphadoxine combinations do not prevent the development of schizonts.

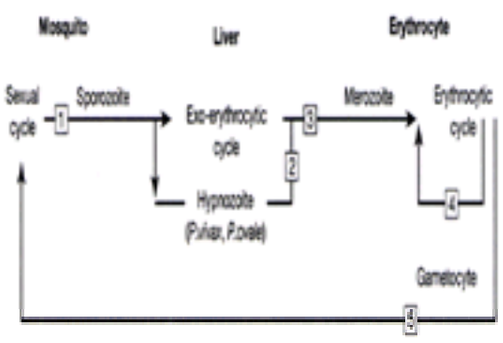

Fig. 12 Stage specificity of antimalarial drugs. 1. Sporontocidal (e.g. proguanil, pyrimethamine, atovaquone); 2. hypnozoitocidal (e.g. primaquine WR238,605); 3. tissue schizontocidal (e.g. proguanil, pyrimethamine); 4. blood schizontocidal (e.g. chloroquine, quinine, artemisinin); 5. gametocytocidal (e.g. primaquine, tafenoquine; chloroquine for *P. vivax, P. malariae,* and *P. ovale*).

Mechanism of action of antimalarial drugs

The mode of action of the antifolate drugs is well understood and described above. Chloroquine is concentrated in the parasite's lysosomes, where haemoglobin is digested, and may act by inhibiting the haempolymerase that converts toxic haemin into insoluble haemozoin (malarial pigment). Alternatively, the drug may interfere with parasite feeding by disrupting its food vacuole. Antimalarial antibiotics are all inhibitors of ribosomal protein synthesis and probably act on the parasite's mitochondria. In the case of artemisinin derivatives, iron within the parasite probably catalyses the cleavage of the endoperoxide bridge leading to the generation of free radicals, which then form covalent bonds with parasite proteins (alkylation). Naphthoquinones, such as atovaquone, act on the electron-transport chain in malarial mitochondria through their structural similarity to coenzyme Q. No satisfactory explanation of the mode of action of the other antimalarial drugs is yet available.

The alarming spread of drug resistance has prompted great experimental effort to reveal the mechanism of resistance. The observation that chloroquine resistance could be reversed *in vitro* by high concentrations of drugs such as the calcium-channel blocker verapamil, which in other situations could reverse the multidrug resistance (**mdr**) phenotype acquired by some tumour cells, focused attention on a malarial homologue of the human *mdr* gene. Recent work suggests involvement of the *P. falciparum mdr1* gene early in the development of resistance, but segregation of resistance in the cloned progeny of a cross showed that the product of a second, uncharacterized gene product from chromosome 7 was also required.

Chloroquine

Despite the widespread resistance of *P. falciparum* to this drug, and the recent emergence of chloroquine-resistant *P. vivax* in New Guinea and adjacent areas of Indonesia, chloroquine is still the most widely used antimalarial drug. It remains the treatment of choice for vivax, ovale, and malariae infections, and for

uncomplicated falciparum malaria acquired in the few areas where the parasite remains sensitive to this drug (Central America north-west of the Panama Canal, Haiti and the Dominican Republic, and parts of the Middle East). In the rest of the malaria-endemic region, the emergence of chloroquine resistance is having a devastating effect on malarial morbidity and mortality. For example, in Senegal, mortality form malaria in children under 5 years old, increased up to 11-fold between 1984 and 1995. After oral administration, chloroquine is rapidly and almost completely absorbed, peak plasma concentrations being reached in about 2 h. Absorption after intramuscular or subcutaneous injection is very rapid, which can produce dangerously high plasma concentrations unless small doses are given frequently. This probably explains the deaths of some children soon after they had received intramuscular injections of chloroquine. About half the absorbed dose is excreted unchanged by the kidney, the remainder being converted to active metabolites in the liver. Therapeutic blood concentrations persist for 6 to 10 days after a single dose and the terminal elimination half-time is 1 to 2 months. Plasma concentrations above about 250 ng/ml produce unpleasant symptoms such as dizziness, headache, diplopia, disturbed visual accommodation, dysphagia, nausea, and malaise. Chloroquine, even in small doses, may cause pruritus in dark-skinned races. Chloroquine may exacerbate epilepsy and photosensitive psoriasis. Cumulative, irreversible retinal toxicity from chloroquine has been reported after lifetime prophylactic doses of 50 to 100 g base (i.e. after 3–6 years of taking 300 mg of base per week), although this is most unusual. Chloroquine overdose is described in Chapter 8.1. Chloroquine is safe during pregnancy and lactation.

Amodiaquine, a 4-aminoquinoline that is structurally similar to chloroquine, retains activity against chloroquine-resistant strains of *P. falciparum* in some geographical areas. Unlike chloroquine, it is metabolized to a toxic quinoneimine that can produce a toxic hepatitis and potentially lethal agranulocytosis (which occurred in up to 1 in 2000 people taking amodiaquine prophylactically). Amodiaquine is still quite widely used, but, because of its risks and the limited therapeutic advantage over chloroquine, its use for prophylaxis and repeated treatment is now discouraged by the World Health Organization.

Quinine

The advent of chloroquine-resistant *P. falciparum* restored quinine to being the treatment of choice for falciparum malaria. Its antimalarial properties were discovered in Peru around 1600. Given by mouth it is rapidly and almost completely absorbed, producing peak plasma concentrations within 1 to 3 h. Some 20 per cent is excreted in the urine and the rest is metabolized in the liver. The elimination half-time in healthy people is approximately 11 h, and in patients with malaria approximately 16 h. Intravenous injection of quinine is dangerous as high plasma concentrations may result during the distribution phase, causing fatal hypotension or arrhythmias. However, quinine can be given safely if it is diluted and infused intravenously over 2 to 4 h. When intravenous infusion is not possible, but parenteral treatment is needed, quinine may be given by intramuscular injection divided between the anterior part of the thighs. For intramuscular injection, the stock solution of quinine dihydrochloride (300 mg/ml) should be diluted to 60 mg/ml. It is well absorbed from this site and complications are rare provided that strict sterile precautions are observed. Because most deaths from severe falciparum malaria occur within the first 96 h of starting treatment, it is important to achieve parasiticidal plasma concentrations of quinine as quickly as possible. This can be accomplished safely by giving a loading dose of twice the maintenance dose. A loading dose of 20 mg of the salt per kg of body weight and an 8- to 12-hourly maintenance dose of 10 mg/kg have proved safe and effective in children and adults in many tropical countries. The initial dose of quinine should not be reduced in patients who are severely ill with renal or hepatic impairment, but in these cases the maintenance dose should be reduced to between 3 and 5 mg/kg if parenteral treatment is required for longer than 48 h. Little is known about the optimal and safe quinine dosage in elderly and obese patients outside malaria endemic areas.

The minimum inhibitory concentration of quinine for *P. falciparum* in SE Asia and other areas of the tropics has risen steadily. Longer courses of quinine and in combination with other drugs, such as Fansidar, tetracycline, or clindamycin, have been required for complete cure. Recently, cases of RII and RIII resistance (failure to clear or failure to reduce parasitaemia in the first 7 days of treatment) to quinine have been documented in Thailand and Vietnam. Quinine should not be withheld or stopped in patients who are pregnant or haemolysing. In the doses used to treat malaria it does not stimulate uterine contraction or cause fetal distress. Hypoglycaemia is the most important complication of quinine treatment (see above). Plasma quinine concentrations above 5 mg/l produce a characteristic group of symptoms—'cinchonism'—transient high-tone deafness, giddiness, tinnitus, nausea, vomiting, tremors, blurred vision, and malaise. Rarely, quinine may give rise to haemolysis, thrombocytopenia, disseminated intravascular coagulation, hypersensitivity reactions, vasculitis, and granulomatous hepatitis. Blindness, deafness, and central nervous depression are commonly observed in patients who have attempted suicide by taking overdoses of quinine. These features are rarely seen in patients being treated for malaria, even though their plasma quinine concentrations may exceed 20 mg/l. This discrepancy may be explained by the increased binding of quinine to a-1 acid glycoprotein (orosomucoid) and to other acute-phase reactive serum proteins in patients with malaria.

Quinidine, the dextrorotatory stereoisomer of quinine, is more effective against resistant strains of *P. falciparum* but is more cardiotoxic than quinine. Because of its use for treating cardiac arrhythmias, it is more generally available (as quinidine gluconate injection) than parenteral quinine in continental Europe and North America, and in the United States has replaced quinine for the parenteral treatment of malaria. It must be infused slowly while the electrocardiogram and blood pressure are monitored. Infusion should be slowed if the blood pressure falls, the plasma concentration exceeds 22 µmol/1 (7 mg/ml), or if the Q–Tc interval increases by more than 25 per cent.

Mefloquine ('Lariam')

This synthetic drug is effective against some *P. falciparum* strains resistant to chloroquine, pyrimethamine–sulphonamide combinations, and quinine. It is too irritant to be given parenterally, but is well absorbed when given by mouth, reaching peak plasma concentrations in 6 to 24 h. The elimination half-time is 14 to 28 days. The drug can be given as a single dose but, to reduce the risk of vomiting and other gastrointestinal side-effects, the dose is best divided into two halves given 6 to 8 h apart. Gastrointestinal symptoms occur in 10 to 15 per cent of patients but are usually mild. Less frequent side-effects include nightmares and sleeping disturbances, dizziness, ataxia, sinus bradycardia, sinus arrhythmia, postural hypotension, and an 'acute brain syndrome' consisting of fatigue, asthenia, seizures, and psychosis. Mefloquine treatment should be avoided in pregnant women, especially during the first trimester, and pregnancy should be avoided within 3 months of stopping mefloquine. People taking b-blockers and those with a past history of epilepsy or psychiatric disease should also avoid the drug. Unfortunately, *in vitro* resistance to mefloquine and treatment failures have now been reported in SE Asia, Africa, and South America.

Halofantrine

This synthetic antimalarial compound is active against multiresistant, including mefloquine-resistant, *P. falciparum*, but is no longer recommended because of its cardiotoxicity.

Artemisinin

Artemisinin or qinghaosu (pronounced 'ching-how-soo') is the active principle of the Chinese medicinal herb *Artemisia annua*—family Compositae (sweet wormwood), which has been used as a treatment for fevers in China for more than 1000 years. The active principle was isolated in China during 1971 to 1972. It is a sesquiterpene lactone with an endoperoxide (trioxane) active group. It destroys young trophozoites as well as other blood stages of *P. falciparum,* including chloroquine-resistant strains, and clears parasitaemia more rapidly than any other antimalarial drug. Dihydroartemisinin, the active metabolite, is cleared rapidly. In severe falciparum malaria, most experience has been gained with intramuscular artemether, given in a loading dose of 3.2 mg/kg on the first day (as a single dose or divided, 12 h apart) followed by 1.6 mg/kg per day until the patient is able to take an oral drug such as mefloquine. The efficacy and safety of artemether was compared with quinine in a series of large randomized trials in children and adults with severe falciparum malaria in Africa, Asia, and Papua New Guinea. A meta-analysis of trials involving nearly 2000 patients confirmed it to be as effective as quinine, judged by case fatality and incidence of neurological sequelae, but it cleared parasitaemia more rapidly and was significantly superior in preventing 'adverse outcome' (either death or neurological sequelae). Artesunate, although inherently unstable in aqueous solution, can be made up with 5 per cent bicarbonate just before injection and given by intravenous or intramuscular injection (2 mg/kg on the first day followed by 1 mg/kg until the patient can take oral treatment). An extra dose of 1 mg/kg can be given 4 to 6 h after the initial loading dose in hyperparasitaemic patients. Suppository formulations of artemisinin have proved effective in severe falciparum malaria and should prove particularly valuable in treating children at peripheral levels of the health service. A combination of artemether and lumefantrine (Riamet, Co-artemether) is being marketed for the oral treatment of multiresistant falciparum malaria.

The severe neurotoxicity reported in animals given large doses of artemisinin has not been detected in any of the tens of thousands of human patients treated with these compounds.

Primaquine

This is the only readily available drug effective against exoerythrocytic (hepatic) forms of *P. vivax* and *P. ovale*, and is essential for the radical cure of these infections. It is also gametocytocidal for all species of malaria. Mass treatment of patients with *P. falciparum* infection could eliminate the sexual cycle in mosquitoes by sterilizing

gametocytes. Its elimination half-time is 7 h. The principal drawback of primaquine is that it causes haemolysis in patients with congenital deficiencies of erythrocyte enzymes, notably G6PD. However, severe intravascular haemolysis is unusual even in G6PD-deficient patients, except in certain areas of the world such as the Mediterranean (for example, Sardinia) and Sri Lanka. Primaquine can cross the placenta and cause severe haemolysis in a G6PD-deficient fetus, most commonly a boy. It is also excreted in breast milk. It should not be used during pregnancy or lactation in areas where G6PD deficiency is prevalent. Primaquine, like sulphonamides and sulphones (for instance, dapsone) can produce severe haemolysis and methaemoglobinaemia in patients with congenital deficiency of **NADH** (the reduced form of nicotinamide adenine dinucleotide) methaemoglobin reductase. The patient quickly develops dusky cyanosis, noticed first in the nail beds. In patients with G6PD deficiency, weekly dosage with 45 mg of primaquine is better tolerated than the usual daily dose of 15 mg. In the Solomon Islands, Indonesia, Thailand, and Papua New Guinea a total dose of 6.0 mg/kg (twice the usual dose) or even more may be needed to eliminate the primaquine-resistant Chesson-type strain of *P. vivax*. This is usually given as 15 mg base/day for 28 days. Tafenoquine (Etaquine), a new 8-aminoquinoline, is now in clinical trials. As a hypnozoiticide, it is over 10 times more active than primaquine, and is also a potent schizontocide.

Pyrimethamine–sulphonamide combinations (Fansidar, Metakelfin, etc.)

These synergistic combinations were once valuable in the treatment of chloroquine-resistant falciparum infections worldwide. A single adult dose of three Fansidar tablets (75 mg pyrimethamine, 1500 mg sulfadoxine) proved safe and effective, and is useful as an emergency standby for travellers out of the reach of medical facilities and as an adjunct to quinine in the treatment of *P. falciparum* infections in areas of increasing quinine resistance. However, in most of SE Asia, China, Oceanea, Latin America, and Africa already troubled by chloroquine resistance, resistance to pyrimethamine–sulphonamide combinations is also spreading. It results from mutations at residues 108, 51, 59, 16, and 164 of the parasite's dihydrofolate reductase gene. An intramuscular formulation has proved effective against *P. falciparum* in southern Africa. Pyrimethamine is a folate inhibitor and so may cause folic acid deficiency in pregnant women and others unless folinic acid supplements are given. The sulphonamide components of these combinations are potentially dangerous. In patients who are hypersensitive to sulphonamide they may cause systemic vasculitis, the Stevens–Johnson syndrome, or toxic epidermal necrolysis. In the United States the risk of fatal reactions has been calculated as 1 in 18 000–26 000 prophylactic courses. Aplastic anaemia and agranulocytosis can also occur. Both pyrimethamine and sulphonamide cross the placenta and are excreted in milk. In the fetus and neonate, sulphonamides can displace bilirubin from plasma protein-binding sites, thus causing kernicterus. For these reasons, pyrimethamine–sulphonamide combinations are not recommended for treatment during pregnancy or lactation unless no alternative drug is available, nor for prophylaxis at all.

P. vivax and *P. malariae* parasitaemias are generally cleared by all the drugs effective against *P. falciparum.* However, in some scattered areas, pyrimethamine–sulphonamide combinations may not be effective because of pyrimethamine resistance.

Chlorproguanil–dapsone ('lapdap')

This combination has been developed as an alternative to pyrimethamine–sulphonamide combinations (**PSD**) to replace chloroquine for the treatment of uncomplicated falciparum malaria in Africa. It has proved more effective than PSD in treating parasites with 108, 51, and 59 mutations, but should probably be further combined with an artemisinin to extend its useful therapeutic life.

Hydroxynaphthoquinones

Atovaquone (BW566C80) is marketed in combination with proguanil as 'Malarone' for the treatment and prevention of multiresistant *P. falciparum*. It inhibits the parasite's mitochondrial respiration by binding to the cytochrome bc, complex. The drug is poorly and variably absorbed, but bioavailability is greatly enhanced by a fatty meal. Its elimination half-life is between 50 and 70 h.

Antibiotics

Tetracycline, clindamycin, azithromycin, quinolones, and sulphonamides such as co-trimoxazole, have some antimalarial activity. Generally, they kill parasites too slowly to be used alone. In an emergency, in the absence of quinoline antimalarials, they could be used to treat malaria.

Practical antimalarial chemotherapy

Prescribing quinoline antimalarial drugs

The various salts of quinoline compounds contain greatly differing amounts of base. If the prescription fails to specify salt or base, or which particular salt is intended, serious problems can arise. Where possible, the dose of base should be prescribed. This is generally accepted for chloroquine, amodiaquine, mefloquine, and primaquine, but, in the case of quinine and quinidine, weights of salts are usually quoted. Conversions are given in Table 7.

Treatment of uncomplicated malaria (Table 8)

Chloroquine is the treatment of choice for *P. vivax, P. ovale, P. malariae*, and uncomplicated *P. falciparum* malarias in those geographical areas where this drug can still achieve a satisfactory clinical response. Chloroquine-resistant *P. vivax* has so far been reported only from New Guinea and adjacent islands of Indonesia. It should be treated by increasing the dose of oral chloroquine. Chloroquine resistant *P. falciparum* is very widespread.

Chloroquine is cheap, safe, and in the usual 3-day course well tolerated. However, despite the clinical improvement following chloroquine treatment, attributable to its anti-inflammatory action, its failure to eliminate parasitaemia and the subsequent recrudescences may eventually lead to the development of profound anaemia. Patients with *P. vivax* or *P. ovale* malarias who will not subsequently reside in malarious areas should be given a course of primaquine (or the new 8-aminoquinoline drug, tafenoquine) to destroy persistent exoerythrocytic stages (see Table 8) unless they are G6PD-deficient.

For the treatment of *P. falciparum* malaria in most parts of the malaria endemic area, chloroquine has been replaced by pyrimethamine–sulphonamide combinations such as 'Fansidar' and 'Metakelfin'. These have the great advantage of being single-dose treatments that are usually well tolerated. In Africa, chlorproguanil–lapudrine ('lapdap') is more effective than PSD against parasites with dihydrofolate reductase gene mutations. Quinine is an effective replacement for chloroquine in most areas where multidrug-resistant strains of *P. falciparum* are prevalent. However, it has the disadvantage of producing unpleasant symptoms. In some countries a short course (3–5 days) of quinine followed by a single dose of pyrimethamine–sulphonamide is still effective. Quinine has also been combined with antibiotics such as tetracycline and clindamycin to improve its efficacy. Mefloquine, given as a single dose, or in divided doses 6 to 8 h apart to reduce the risk of vomiting, was initially highly effective against multiresistant strains of falciparum malaria throughout the world. However, in some areas, notably in the border regions of Thailand, mefloquine resistance has developed rapidly and this drug is now used in combination with artemisinin derivatives such as artesunate. The newer combination drugs 'Malarone' and 'Co-artemether' are effective against multiresistant *P. falciparum*

Patients with uncomplicated malaria can usually be given antimalarial drugs by mouth. However, feverish patients may vomit the tablets. The risk of vomiting can be reduced if the patient lies down quietly for a while after taking an antipyretic such as paracetamol. The initial dose of antimalarial drug may have to be given by injection for those who vomit persistently.

Treatment of severe falciparum malaria (Table 9)

Appropriate chemotherapy should be started as soon as possible as there is a highly significant relationship between delay in chemotherapy and mortality. **In sick and deteriorating patients, a therapeutic trial is indicated even if initial smears have proved negative.** Whenever possible, the dosage should be calculated according to the patient's body weight. The parenteral administration of drugs is the rule for patients with severe and complicated falciparum malaria and in any patient who vomits and is unable to retain swallowed tablets. In the case of cinchona alkaloids, this is most safely and effectively achieved by infusing the drug, diluted in isotonic fluid, intravenously over a period of 2 to 4 h.

The therapeutic response must be carefully monitored by frequent clinical assessments, measurement of temperature, pulse, and blood pressure, and examination of blood films. Patients should be switched to oral treatment as soon as they are able to swallow and retain tablets. They must be watched carefully for signs of drug toxicity. In the case of cinchona alkaloids, the most common toxicity during antimalarial treatment is the development of hypoglycaemia. The blood sugar should,

therefore, be checked frequently.

General management

Patients with severe malaria should be transferred to the highest level of care available, preferably the intensive care unit. They must be nursed in bed because of their postural hypotension and because of the risk of splenic rupture were they to fall. Body temperatures above 38.5 °C are associated with febrile convulsions, especially in children, and between 39.5 and 42 °C with coma and permanent neurological sequelae. In pregnant women, hyperpyrexia contributes to fetal distress. Temperature should therefore be controlled by fanning, tepid sponging, a cooling blanket, or antipyretic drugs such as paracetamol (15 mg/kg in tablets by mouth, or powder washed down a nasogastric tube, or as suppositories). As the slight prolongation of parasitaemia associated with the use of paracetamol (and possibly other methods for controlling fever) is clinically insignificant, this possible disadvantage and theoretical arguments against lowering the temperature are outweighed by the symptomatic benefits. Pyrazolones such as metamizole sodium (Dipyrone) are widely used in tropical countries but carry an unacceptable risk of inducing agranulocytosis.

Cerebral malaria

Convulsions, vomiting, and aspiration pneumonia are common, so patients should be nursed in the lateral position with a rigid oral airway or endotracheal tube in place. They should be turned at least once every 2 h to avoid bed sores. Vital signs, Glasgow coma score, and occurrence of convulsions should be recorded frequently. Convulsions can be controlled with diazepam given by slow intravenous injection (adults 10 mg, children 0.15 mg/kg) or intrarectally (0.5–1.0 mg/kg), or with paraldehyde drawn in a glass syringe and given by intramuscular injection (0.1 ml/kg). Anaphylactic use of phenobarbital was associated with increased case fatality in a placebo-controlled study in African children and is not recommended. Stomach contents should be aspirated through a nasogastric tube to reduce the risk of aspiration pneumonia. Elective endotracheal intubation is indicated if coma deepens and the airway is jeopardized. Deepening coma with signs of cerebral herniation is an indication for CT or magnetic resonance imaging, or a trial of treatment to lower intracranial pressure, such as an intravenous infusion of mannitol (1.0–1.5 g/kg of a 10–20 per cent solution over 30 min) or mechanical hyperventilation to reduce the arterial P_{CO2} to below 4.0 kPa (30 mmHg).

A number of potentially harmful remedies of unproven value have been recommended for the treatment of cerebral malaria. Two double-blind trials of dexamethasone (2 mg/kg and 11 mg/kg intravenously over 48 h) in adults and children in Thailand and Indonesia showed no reduction in mortality but prolongation of coma and an increased incidence of infection and gastrointestinal bleeding. Low molecular-weight dextrans, osmotic agents, heparin, adrenaline (epinephrine), ciclosporin A, prostacyclin, and pentoxifylline (oxpentifylline), malarial hyperimmune globulin, and anti-TNF-a monoclonal antibodies have proved ineffective in the treatment of cerebral malaria. Most of these interventions were associated with serious side-effects.

Anaemia

Indications for transfusion—preferably with fresh, compatible whole blood or packed cells—include a low (less than 20 per cent or rapidly falling) haematocrit, severe bleeding or predicted blood loss (for example, imminent parturition or surgery), hyperparasitaemia, and failure to respond to conservative treatment with oxygen and plasma expanders. When the screening of transfused blood is inadequate and infections such as human immunodeficiency virus (**HIV**), human T-cell leukaemia virus-1 (**HTLV-1**), and hepatitis viruses are prevalent in the community, the criteria for blood transfusion must be even more rigorous. Exchange transfusion is a safe way of correcting the anaemia without precipitating pulmonary oedema in those who are fluid-overloaded or chronically and severely anaemic. The volume of transfused blood must be included in the fluid-balance chart. Diuretics such as furosamide (frusemide) can be given intravenously in a dose of 1 to 2 mg/kg body weight to promote diuresis during the transfusion, and in all cases transfusion must be cautious with frequent observations of the jugular or central venous pressure and auscultation for pulmonary crepitations. Survival of compatible donor red cells is greatly reduced during the acute and convalescent phases of falciparum malaria.

Disturbances of fluid and electrolyte balance

Fluid and electrolyte requirements must be assessed individually in patients with malaria. Circulatory overload with intravenous fluids or blood transfusion may precipitate fatal pulmonary oedema, but untreated hypovolaemia may lead to fatal shock, lactic acidosis, and renal failure. Hypovolaemia may result from salt and water depletion through fever, diarrhoea, vomiting, insensible losses, and poor intake. The state of hydration is assessed clinically from the skin turgor, peripheral circulation, postural change in blood pressure, peripheral venous filling, and jugular or central venous pressure. The history of recent urine output and measurement of urine volume and specific gravity may be useful. Adult patients with severe falciparum malaria usually require between 1000 and 3000 ml of intravenous fluid during the first 24 h of hospital admission. Fluid replacement should be controlled by observations of jugular, central venous, or pulmonary artery wedge pressures. Hyponatraemia (plasma sodium concentration 120–130 mmol/l) usually requires no treatment, but these patients should be cautiously rehydrated with isotonic saline if they are clinically dehydrated, have low central venous pressures, a high urinary specific gravity, and a low urine sodium concentration (below 25 mmol/l).

Renal failure

Patients with falling urine output and elevated blood urea nitrogen and serum creatinine concentrations can be treated conservatively at first, but established acute renal failure must be treated with haemofiltration or dialysis. Hypovolaemia is corrected by the cautious infusion of isotonic saline until the central venous pressure is in the range +5 to +15 cmH_2O. If urine output remains low after rehydration, increasing doses of slowly infused intravenous furosemide (frusemide) (up to a total dose of 1 g) and finally an intravenous infusion of dopamine (2.5–5 µg/kg per min) can be tried. If these measures fail to achieve a sustained increase in urine output, a strict fluid balance should be enforced with particular emphasis on fluid restriction. Indications for haemoperfusion/dialysis include a rapid increase in serum creatinine level, hyperkalaemia, fluid overload, metabolic acidosis, and clinical manifestations of uraemia (diarrhoea and vomiting, encephalopathy, gastrointestinal bleeding, and pericarditis). Haemofiltration is the most effective technique in malaria but haemodialysis or peritoneal dialysis are also effective. The initial doses of antimalarial drug should not be reduced in patients with renal failure but, after 48 h of parenteral treatment, the maintenance dose should be reduced by one-third to one-half.

Metabolic acidosis

This is usually attributable to lactic acidosis and is an important life-threatening complication, especially in anaemic children. It should be treated by improving perfusion and oxygenation by blood transfusion and correcting hypovolaemia, clearing the airway, increasing the inspired oxygen concentration, and by treating septicaemia, a frequently associated complication.

Pulmonary oedema

This must be prevented by propping the patient up at an angle of 45 degrees and controlling fluid intake so that the jugular or central venous pressure is kept below +5 cmH_2O. Those who develop pulmonary oedema should be propped upright and given oxygen to breathe. In a well-equipped intensive care unit, the judicious use of vasodilator drugs can be controlled by monitoring haemodynamic variables, fluid overload can be corrected by haemoperfusion, and oxygenation can be improved by mechanical ventilation with positive end-expiratory pressure.

Hypotension and 'shock' ('algid malaria')

This should be treated as for bacteraemic shock. The circulatory problems should be corrected with blood transfusion (for example, in anaemic children with respiratory distress and acidosis), plasma expanders, dopamine, and broad-spectrum antimicrobial treatment (such as gentamicin with ceftazidime or cefuroxime plus metronidazole) should be started immediately, bearing in mind that likely routes of infection include the urinary tract, lungs, and the gut. Other causes of shock in patients with malaria include dehydration, blood loss (for instance, following splenic rupture), and pulmonary oedema.

Hypoglycaemia

This may be asymptomatic, especially in pregnancy, and its clinical manifestations may be confused with those of malaria. Blood sugar must be checked every few hours, especially in patients being treated with cinchona alkaloids. Hypoglycaemia may arise despite continuous intravenous infusions of 5 or even 10 per cent dextrose. A therapeutic trial of dextrose (1 ml/kg by intravenous bolus injection) should be given if hypoglycaemia is proved or suspected. This should be followed by a continuous infusion of 10 per cent dextrose. Glucose may be given by nasogastric tube to unconscious patients or by peritoneal dialysis in those undergoing this

treatment for renal failure. Among agents that block insulin release, diazoxide was ineffective, but octreotide (Sandostatin), a synthetic somatostatin analogue, proved effective in some severe cases of quinine-induced hypoglycaemia.

Hyperparasitaemia and exchange blood transfusion

In non-immune patients, mortality increases with parasitaemia, exceeding 50 per cent with parasitaemias above 500 000/µl. Exchange transfusion reduces parasitaemia more rapidly than optimal chemotherapy alone, although this advantage will be less when artemisinins are used, and could have the additional benefit of removing harmful metabolites, toxins, cytokines and other mediators, and restoring normal red-cell mass, platelets, clotting, factors, albumin, etc. Potential dangers of the procedure include electrolyte disturbances (for example, hypocalcaemia), cardiovascular complications, and the introduction of infectious agents into the blood and through infection of intravascular lines. The use of exchange transfusion, haemopheresis, and plasmapheresis has been reported in more than 100 patients, the vast majority of whom survived. There was undoubtedly some reporting bias. Some patients showed clinical improvement, such as recovery of consciousness, and restoration of urine flow, soon after the procedure. A meta-analysis discovered no higher survival rate compared to chemotherapy alone and there have been a few recent reports of adult respiratory distress syndrome developing during the procedure. The efficacy of exchange transfusion is never likely to be put to the test of a randomized comparative study, but, where facilities allow and screening of donor blood is adequate, the procedure should be considered in non-immune patients who are severely ill, who have deteriorated on conventional treatment, and who have parasitaemias in excess of 10 per cent. The introduction of antimalarial agents, such as artemisinins, which clear parasitaemia very rapidly, may obviate the need for exchange transfusion.

Splenic rupture

Acute abdominal pain and tenderness with left shoulder-tip pain and shock in patients with vivax and falciparum malaria should suggest the possibility of splenic rupture, especially if there is a history of abdominal trauma. Free blood in the peritoneal cavity and a torn splenic capsule can be detected by ultrasound or CT and confirmed by needle aspiration of the peritoneal cavity, laparoscopy, or laparotomy. Conservative management with blood transfusion and close observation in an intensive care unit is sometimes successful but access to surgical help is essential in case there is a sudden deterioration.

Disseminated intravascular coagulation

Patients with evidence of a coagulopathy should be given vitamin K (adult dose 10 mg by slow intravenous injection). Prothrombin complex concentrates, cryoprecipitates, platelet transfusions, and fresh-frozen plasma should be considered.

Management of the pregnant woman with malaria

Malaria must be diagnosed and treated rapidly in pregnant women. Unwarranted fears of abortifacient and fetus-damaging effects of antimalarial drugs have led to the delay or even withdrawal of treatment, but experience since the nineteenth century has confirmed the safety of quinine in pregnancy. Chloroquine has been used extensively without ill effect to mother or fetus. However, pyrimethamine–sulphonamides, tetracycline, primaquine, and aspirin (but not paracetamol) are contraindicated in late pregnancy and mefloquine should be avoided if possible. In pregnant women, the total apparent volume of distribution of quinine is reduced and the drug is eliminated more rapidly. Initial dosage is the same as in non-pregnant patients, but in severe cases requiring prolonged parenteral treatment, the dose, but not the frequency of administration, should be reduced. The main danger of quinine in pregnancy is its stimulation of insulin secretion with resulting hypoglycaemia (see above). Blood glucose must be checked at least once a day in pregnant women with malaria, whether or not they are receiving quinine. Maternal fever should be reduced as soon as possible. Induction of labour, caesarean section, or speeding up of the second stage of labour with forceps or vacuum extractor should be considered in patients with severe falciparum malaria. Fluid balance is particularly critical in these patients. If possible, the central venous pressure should be monitored. Exchange transfusion of 1000 to 1500 ml of blood in late pregnancy proved an effective way of managing severe anaemia with high-output cardiac failure in Nigeria. Circulating volume could be reduced and the risk of postpartum pulmonary oedema lessened by replacing exfused blood with a smaller volume of packed cells.

Prognosis

The mortality of acute vivax, ovale, and malariae malarias is negligible. Strictly defined cerebral malaria has a mortality of about 10 to 15 per cent when medical facilities are good, and may be less than 5 per cent in Western intensive care units. Antecedent factors that predispose to severe falciparum malaria include the lack of acquired immunity or lapsed immunity, splenectomy, pregnancy, and immunosuppression. There is a strong correlation between the density of parasitaemia and disease severity. Severe clinical manifestations, such as impaired consciousness, retinal haemorrhages, renal failure, hypoglycaemia, haemoglobinuria, metabolic acidosis, and pulmonary oedema, carry a bad prognosis. The case fatality of pregnant women with cerebral malaria, especially primiparae in the third trimester, is approximately 10 times greater than in non-pregnant patients. The following laboratory findings carry a poor prognosis: peripheral schizontaemia, peripheral leucocytosis exceeding 12 000/µl, malarial pigment in >5 per cent of circulating neutrophils, high CSF lactate or low glucose, low plasma antithrombin III, serum creatinine exceeding 265 µmol/l, or a blood urea nitrogen of more than 21.4 mmol/l, haematocrit less than 20 per cent, blood glucose less than 2.2 mmol/l, and elevated serum enzyme concentrations (for example, aspartate and alanine aminotransferases, lactate dehydrogenase).

Chronic immunological complications of malaria

Quartan malarial nephrosis

In parts of East and West Africa, South America, India, South-East Asia, and Papua New Guinea, there is epidemiological evidence linking *P. malariae* infection to immune-complex glomerulonephritis, leading to nephrotic syndrome. Few of those exposed to repeated *P. malariae* infections develop nephrosis, suggesting that additional factors are involved. The histological changes, which are not entirely specific, are of a progressive focal and segmental glomerulosclerosis with fibrillary splitting or flaking of the capillary basement membrane, producing characteristic lacunae. Electron-dense deposits beneath the endothelium can be seen by electron microscopy. Immunofluorescence reveals glomerular deposits of immunoglobulins and C3, and *P. malariae* antigen, in about 25 per cent of cases. More than half the patients present by the age of 15 years with typical features of nephrotic syndrome. *P. malariae* is frequently found in blood smears and *P. malariae* antigen in renal biopsies in children but not in adults. The renal lesions may be perpetuated by autoimmune mechanisms. The pattern of immunofluorescent staining has some prognostic significance. Few patients respond to corticosteroids, but some are helped by azathioprine and cyclophosphamide, especially those whose renal biopsies show the coarse or mixed patterns of immunofluorescence. Antimalarial treatment is not effective. This condition could be prevented by antimalarial prophylaxis and has disappeared in countries such as Guyana during a period of malaria eradication.

Tropical splenomegaly syndrome (hyper-reactive malarial splenomegaly)

Transient splenomegaly is a feature of acute attacks of malaria in non-immune or partially immune patients, while progressive splenomegaly is seen in children resident in malarious areas during the process of their acquiring immunity to the infection. However, a separate entity has been described in Africa (especially Nigeria, Uganda, and Zambia), the Indian subcontinent (Bengal, Sri Lanka), South-East Asia (Vietnam, Thailand, and Indonesia), South America (Amazon region), Papua New Guinea, and the Middle East (Aden). The defining features are residence in a malarious area, chronic splenomegaly, elevated serum IgM and malarial antibody levels, hepatic sinusoidal lymphocytosis, and a clinical and immunological response to antimalarial prophylaxis. This condition is thought to result from an aberrant immunological response to repeated infection by any of the species of malaria parasite.

Pathophysiology

In Flores, Indonesia, *P. vivax* infection leads to the production of IgM lymphocytotoxic antibodies specific for the suppressor T lymphocytes, which normally regulate IgM production. The resulting disinhibition of B lymphocytes leads to their overproduction of IgM, forming macromolecular aggregates of IgM (cryoglobulins). The need to clear these aggregates stimulates the reticuloendothelial system and causes the progressive and eventually massive splenomegaly and hepatomegaly. The decrease in suppressor/cytotoxic (CD8) lymphocytes increases the helper:suppressor (CD4:CD8) ratio. Antimalarial chemoprophylaxis, by removing the antigenic stimulus provided by repeated malarial infections, allows the patient's immune system to return to normal. There are some differences between tropical splenomegaly syndrome in Africa, Flores, and Papua New Guinea. In Africa, but not in Flores or Papua New Guinea, there is a peripheral lymphocytosis resulting from an increase in B lymphocytes, and distinction from chronic lymphatic leukaemia may be difficult. In Ghana, clonal rearrangements of the JH region of the immunoglobulin gene were found in patients with tropical splenomegaly who failed to respond to proguanil chemoprophylaxis, suggesting that the syndrome may evolve into a malignant lymphoproliferative disorder. Some of these patients had features of splenic lymphoma with villous (hairy) lymphocytes. In Africa and Papua New Guinea, IgG levels

were significantly increased, but not in Flores. In Flores only the titres of *P. vivax* IgM antibodies were higher in patients with the splenomegaly syndrome, but in Papua New Guinea titres of *P. falciparum, P. vivax,* and *P. malariae* were increased, and in Africa *P. falciparum* and *P. malariae* are the species involved. The familial tendency of the tropical splenomegaly syndrome in Africa and Papua New Guinea suggests a genetic factor.

Clinical features

In malaria endemic areas, patients with tropical splenomegaly syndrome are distinguishable by their progressive splenic enlargement persisting beyond childhood. The spleen may be enormous, filling the left iliac fossa, extending across the midline and anteriorly, producing a visible mass with an obvious notch. The liver is usually enlarged, especially its left lobe. Symptoms attributable to the spleen include a vague dragging sensation and occasional episodes of severe pain with peritonism, suggesting perisplenitis or splenic infarction. Anaemia may become severe enough to cause the features of high-output cardiac failure. Acute haemolytic episodes are described. These patients are vulnerable to infections, especially of the skin and respiratory system, and most deaths are attributable to overwhelming infection. Chronic hypersplenic neutropenia or failure to mobilize neutrophils in response to acute bacterial infections may be the cause. In Papua New Guinea, 57 per cent of those with massive splenomegaly were dead within 7 years.

Patients with splenic lymphoma with villous lymphocytes (Ghana) had splenic discomfort, anorexia, and hepatosplenomegaly, with infiltration of the bone marrow with villous lymphocytes.

Laboratory findings

Severe chronic anaemia is the result of destruction and pooling in the spleen and dilution in an increased plasma volume. Thrombocytopenia may also be caused by splenic sequestration; it rarely causes bleeding. There is neutropenia and, in African patients, peripheral lymphocytosis and lymphocytic infiltration of the bone marrow. Serum IgM is greatly elevated.

The essential histopathological feature is lymphocytosis of the hepatic sinusoids with Kupffer-cell hyperplasia. In some cases, round-cell infiltration of the portal tracts is associated with fibrosis, leading to portal hypertension. In the spleen there is dilatation of the sinusoids, hyperplasia of the phagocytic cells with evident erythrophagocytosis, and infiltration with lymphocytes and plasma cells. No histopathological explanation has been found for the episodes of acute splenic pain.

In patients with splenic lymphoma and villous lymphocytes, more than 30 per cent of circulating lymphocytes are villous. These cells can be distinguished from hairy-cell leukaemia by their lack of CD25, CD11c, and tartrate-resistant acid phosphatase markers.

Differential diagnosis

Tropical splenomegaly syndrome must be distinguished from other causes of chronic, painless, massive splenomegaly, including leukaemias, lymphomas, myelofibrosis, thalassaemias, haemoglobinopathies, visceral leishmaniasis (by examination of bone marrow or splenic aspirates), and schistosomiasis (by liver biopsy, rectal snip, and stool examination). Lymphomas (especially chronic lymphatic leukaemia and follicular lymphoma—see above) and even leukaemias may develop in patients with tropical splenomegaly syndrome. Non-tropical idiopathic splenomegaly (normal serum IgM) and Felty's syndrome produce a similar histological picture in the liver. Many cases of splenomegaly in the tropics remain undiagnosed.

Treatment

Prolonged antimalarial chemoprophylaxis is the most important element of treatment. In Papua New Guinea, 70 per cent of the patients showed marked improvement after 12 months of chemotherapy. The choice of drug will depend on the local sensitivity of whichever species or group of species of malaria parasite are thought to be responsible for this syndrome (see Chemoprophylaxis below). The short- and long-term dangers of splenectomy rule out this procedure in the rural tropics. Similarly, splenic irradiation and antimitotic agents are dangerous and unnecessary. Folic acid may be needed. Diagnosis of patients with splenic lymphoma with villous lymphocytes (Ghana) is important as, in this condition, the risks of splenectomy are outweighed by the benefits.

Endemic Burkitt's lymphoma (see Chapter 7.10.3)

Endemic Burkitt's lymphoma, a tumour of the jaw, abdomen, and other areas that spreads to the bone marrow or meninges, is the most common type of childhood malignant disease in many parts of East and West Africa and Papua New Guinea. It has also been reported from Brazil, Malaysia, and the Middle East. Burkitt noticed that its distribution (by altitude, temperature, and rainfall) and even its seasonal incidence followed that of holoendemic falciparum malaria. Outside the malaria endemic area, Burkitt's lymphoma occurs sporadically. There is a suggestion that the B-cell line in Caucasian cases comes from lymphoid tissue, whereas in African cases it comes from the bone marrow. Epstein–Barr virus (**EBV**) produces a lifelong infection of B lymphocytes. In normal individuals this is controlled by specific, HLA-restricted, cytotoxic T cells, which recognize a virus-induced, lymphocyte-detected membrane antigen (**LYDMA**) on B cells. Immunosuppression, as in recipients of renal allografts, allows uncontrolled proliferation of the EBV-infected B-cell line, which may give rise to one of the three chromosomal translocations [t(8;14), t(2;8), t(8;22)] that activate the c-*myc* oncogene on chromosome 8 responsible for malignant transformation. Acute *P. falciparum* infection leads to a reduction in the numbers of suppressor T (CD8) lymphocytes and a decrease in the helper:suppressor (CD4:CD8) ratio, allowing proliferation and increased immunoglobulin secretion by EBV-infected B cells. No lymphocytotoxic antibody is found in acute plasma samples to explain the decrease in suppressor T cells. These tumours may grow so rapidly that massive local tissue destruction results in urate nephropathy and acute renal failure. Cyclophosphamide, vincristine, methotrexate, and prednisolone are used in chemotherapy, producing remissions in 80 to 90 per cent of patients and a long-term survival of 20 to 70 per cent. Breakdown of large tumours during the first week of chemotherapy may be so dramatic that the acute tumour lysis syndrome may be precipitated. This consists of metabolic acidosis, hyperuricaemia, hyperphosphaturia, hyperphosphataemia, hyperproteinaemia, and hyperkalaemia, which may result in fatal cardiac arrhythmia and acute uric-acid nephropathy with renal failure.

Malaria control

Malaria control relies on breaking the chain of transmission, often by attacks on the vector. As the insecticide resistance of mosquitoes and drug resistance of parasites increase, the environmental methods previously used to control anopheline breeding are being revived. The use at night of insecticide-treated bed nets (**ITNs**) has been a major innovation, combining personal protection with population protection (the mass effect) in some situations. No vaccine is yet available for operational use. There is currently a more balanced approach to malaria control than in the past, with emphasis on the early diagnosis and prompt treatment of infected people, selective and sustainable use of antivector measures, and epidemic control. The importance of malaria control has been acknowledged at the political level and available methods are being more energetically applied than for some decades, for example, in the WHO's 'Roll back malaria' programme.

Transmission control

Mosquitoes may be controlled in two ways: by removing, poisoning, or otherwise changing their larval habitats and so reducing their numbers; or by killing the adult mosquitoes by means of insecticides. These may be sprayed into the air for a transient effect or put on to the surfaces where mosquitoes rest to obtain a persistent or residual effect. Other methods may simply deter mosquitoes from biting people. Combination methods whereby insecticide is put on a mechanical barrier such as a bednet are currently much favoured and are discussed separately. For the future, there is also much interest in finding ways to transfect mosquitoes with genes to render them unable to transmit malaria, incorporating them into an infective agent that will spread through mosquito populations. Although killing the adult mosquitoes or their larvae will reduce mosquito numbers, and malaria transmission proportionately, residual insecticides have a greater effect on the survival of infected mosquitoes to the age at which they can pass on the infection, thereby reducing malaria transmission much more than might otherwise be expected.

Mosquito species are highly selective in their choice of larval habitat, and there are usually few major vector species in a given locality. The selective destruction of vector breeding sites (species sanitation), is a long-term method of mosquito control. Sites can be made unsuitable for vector breeding by drainage, changing the rate of water flow, and adding or removing shade, cutting emergent vegetation, and altering the margins of bodies of water. Near the sea, salinity changes may be relevant. For small reservoirs and irrigation canals, cyclical changes in water level by means of a large siphon may control larvae by alternately stranding and flushing. Intermittent drying out of irrigation channels may be of value. No generalizations are possible as, for example, water fluctuations that control vectors in the southern United States would increase breeding in sub-Saharan Africa. Enough local information is available to guide public-health engineering interventions in most endemic areas. As these and other measures against breeding reduce mosquito density, to which transmission is proportional, environmental control is most effective in areas of unstable malaria. Because costs of environmental measures are related to the area involved, and the resources and benefits are related to the human

population, environmental control is most likely to be feasible in areas of high population density. In cities it needs to extend beyond the periurban fringe where the poor are concentrated. Control of mosquitoes such as *A. gambiae,* which utilize temporary pools as small as hoof prints, is very difficult by environmental means without ruthless discipline.

Where habitats cannot be drained or rendered structurally unsuitable, chemical larvicides may be used. Diesel oil, at 40 1/hectare of water surface with or without the addition of insecticides, will prevent the larvae breathing when it is spread on the water surface with the addition of a spreading agent. In the correct formulation, 1 kg/hectare of Paris Green is effective, but 2 to 20 kg/hectare of temephos (Abate) granules is a safer alternative, usually needing to be repeated weekly or fortnightly.

The use of residual insecticides applied to walls and other indoor surfaces gives a far more persistent effect, so that **DDT** (dichlorodiphenyltrichloethane) at 2 g/m ^{2} will remain toxic to endophilic anophelines for 6 months or more on a non-absorbent wall material, as may l-cyhalothrin at a much lower dosage, while organophosphorus insecticides such as malathion, propoxur, and fenitrothion at the same dosage last about 3 months. This approach is a community one, requiring coverage of all houses and shelters, as it relies on killing the mosquito after it has fed. Where the aim is individual or family protection, a knock-down insecticide used before evening in a screened house is more relevant.

Prudent behaviour can greatly reduce the risk of an infective mosquito bite, especially for the visitor to an endemic area. As anophelines bite mostly in the evening, remaining in a screened area from dusk, wearing long sleeves and leg coverings, and sleeping beneath a mosquito net are of real, if underestimated, benefit. Recently, the use of bednets impregnated with synthetic pyrethroids such as permethrin or l-cyhalothrin has been found to give substantial malaria protection in endemic areas, reducing the number of clinical attacks even in areas of high transmission by 50 per cent, and where high coverage is achieved reducing the all-cause infant mortality rate by up to 27 per cent. The effect is due to a combination of reduced access of mosquitoes to people because of the net, a repellent and lethal effect of the insecticide on the mosquitoes trying to bite, and sometimes an effect on mosquito density so that even those outside the nets may get some protection. Nets are most effective when mosquito biting is concentrated late at night, and they can give good protection to babies in cots. The large-scale operational use of impregnated bednets in endemic areas is currently expanding, but as the net is a commodity rather than a health service, the best economic basis for sustainable high coverage by ITNs, and for their regular retreatments, are still being explored. ITNs appear to be one of the most hopeful means of control pending development of an operational vaccine.

As engineering methods are costly, though long lasting, and insecticides can be viewed as polluting the environment, other methods of mosquito control have been sought, with variable success. Genetic control of anophelines is not feasible at present; biological control is often a useful accessory method and usually relies on small fish, especially of such genera as *Gambusia* and *Lebistes* that preferentially feed on mosquito larvae. Species of fish that survive drying out of the habitat as eggs are now of interest. The micro-organism *Bacillus thuringiensis* is used in control, but it effectively functions as a biological insecticide because it produces a toxin.

Prevention of malaria in travellers

Advice to travellers

The prevention of malaria in travellers, particularly those usually resident in non-malarious areas but visiting endemic regions, is becoming increasingly difficult, owing to the spread of resistance to the commonly available antimalarial agents, which means that prevention cannot be completely successful. The four components of advice to travellers must therefore be: (1) to be aware of the risk; (2) to reduce exposure to being bitten by anopheline mosquitoes; (3) to take chemoprophylaxis where appropriate; (4) to seek immediate medical advice in the event of any fever or 'flu-like illness developing while in the area, or within 3 or more months of leaving it, and to consider malaria as a possibility regardless of the precautions taken. The first two of these are at least as important as the third in preventing mortality from malaria.

Preventive advice is subject to uncertainty. This is because unequivocal data on efficacy are often unavailable, published studies are conflicting, and the distribution of resistance to many prophylactics is not well mapped. The balance between the risk of malaria and the risk of side-effects involves value judgements on which experts differ. Moreover, prospective travellers consult several sources of advice, obtain different opinions, and compliance with any regimen thus falls. Published advice is usually by country (the World Health Organization annually produces the most useful list of risk areas) and is inevitably directed towards prophylaxis for the areas of greatest transmission. Consultation with someone who knows the country and the traveller's itinerary may well lead to good advice that differs and is more specific. Intelligent travellers need to be made aware of these issues but they also require clear advice that must include the general points discussed in the following paragraph.

For any traveller to an endemic area there is a risk of malaria. No prophylactic regimen will give total protection, but many will reduce the risk of a malaria attack substantially. In the event of a fever while travelling, or afterwards, malaria must be considered as a diagnosis. Strict compliance, even with a suboptimal prophylactic regimen, is more important than vacillation over finding the optimal one.

Prevention of mosquito bites

There are many additional ways to reduce the risk of malaria. Bednets impregnated with a pyrethroid insecticide (permethrin, deltamethrin, or l-cyhalothrin) should be used, properly tucked in, and without tears or other holes through which mosquitoes might enter. A well-screened bedroom and other accommodation, combined with use of a knock-down insecticide when the doors are closed, will give substantial protection. Clothes that deter mosquito bites, repellent sprays and soaps (containing *N*, *N*-diethyl-*m*-toluamide (**DEET**) or permethrin), and avoiding exposure to bites in the evenings will also help.

Chemoprophylaxis

Chloroquine and/or proguanil

In malarious areas from which chloroquine-resistant *P. falciparum* is absent, mainly in Western Asia, North Africa, and Central America, chloroquine 300 mg (base), usually two tablets taken once a week, will give good protection. However, since it acts as a suppressive of the blood forms of *Plasmodium* it will not prevent late attacks of *P. vivax* or *P. ovale.* Proguanil, 100 mg daily, or 200 mg daily in areas of intense transmission, will act as a true causal prophylactic but is poorly protective against *P. vivax* in these doses. The extremely low incidence of adverse side-effects from proguanil makes it acceptable to long-term residents in endemic areas. Chloroquine is suitable for up to 6 years of use, but beyond this proguanil may be substituted. Recommendations are summarized in Table 9 and Table 10.

By 1993, chloroquine-resistant *P. falciparum* had been reported from most malarious countries, and it constitutes a massive and increasing problem in sub-Saharan Africa and in SE Asia (where multiple drug resistance is common). Newer drugs and the more effective drug combinations for prophylaxis against chloroquine-resistant *P. falciparum* carry a significant risk of severe toxic side-effects that has to be balanced against the malaria risk, which varies greatly within countries, especially in Asia. Where the proportion of malaria resistant to chloroquine is low or the degree of resistance limited, the combination of chloroquine and proguanil ((b)1, Table 10) has the advantage of low toxicity and appears to be effective in many areas, including India and the rest of South Asia. These two drugs also have a good safety record in pregnant women and in young children. Long-term use of prophylactic chloroquine only carries a risk of retinopathy (probably very small) once the total cumulative dose exceeds 100 g of base (over 6 years at the standard prophylactic dose). Pruritis can be a problem in those with dark skins.

However, the combination of chloroquine and proguanil no longer provides adequate protection in sub-Saharan Africa where the malaria challenge in rural areas may exceed one infective bite per night and resistance is common, nor in SE Asia where there is a much lower transmission rate but a greater range of drugs to which *P. falciparum* is resistant.

Other prophylactic drugs

Other prophylactic regimens involve the use of mefloquine, doxycycline, and the combination of atovaquone and proguanil. Mefloquine was the most widely used of these three regimens and there are far more data on its use in malaria prevention than for the other two. Doxycycline was only recently licensed in the United Kingdom for the chemoprophylaxis of malaria, although it has been in use for the prevention of acne for many years; it has also been used as an antimalarial agent outside the license. There is much less experience with the combination of atovaquone and proguanil, though each of its component medicines has been used without high resistance and without high levels of side-effects. Trials of the efficacy of all three regimens demonstrate good protection against chloroquine-resistant falciparum malaria and the choice between them depends on the traveller, destination, and duration of travel. In the absence of specific resistance, mefloquine has a

prophylactic efficacy of around 90 per cent against falciparum malaria, and doxycycline is almost as effective in trials in Asia. However, the data on the combination of atovaquone and proguanil come mainly from studies of semi-immune people and are less extensive. Although the atovaquone/proguanil regimen gives a similar level of efficacy to the other two prophylactics, more information is needed: its particular advantage lies in the low level of serious adverse effects observed in studies to date.

Mefloquine ('Lariam')

Mefloquine has a long half-life and on a weekly dosage schedule the blood level rises to a plateau from about 7 weeks. The majority of the side-effects, which are the main problem with its use, are associated with the initial three doses of mefloquine. The drug should therefore be started 2½ weeks prior to departure for a malarious place, so that if side-effects are troublesome an alternative may be used. Although it is usual to avoid taking mefloquine for longer than a year, American experience suggests that no additional problems arise after 2 to 3 years. The main serious early side-effects of mefloquine are neuropsychiatric, and include anxiety, depression, delusions, fits, and psychotic attacks. The frequency of these is disputed. Airline passenger surveys have shown a frequency of 1:10 000, but experienced doctors in the United Kingdom assert a much higher frequency and further data are needed. As its safety during early pregnancy is uncertain, it is not recommended for those in the first trimester of pregnancy or those at risk of pregnancy during the 3 months after the end of chemoprophylaxis There is some evidence from SE Asia of an increased stillbirth rate in those taking it in later pregnancy, but the risk from malaria is also great in pregnancy. It is contraindicated in people with a history of epilepsy or psychiatric disease. Sporadic cases of mefloquine resistance are already reported from Africa, and on the border between Thailand and Cambodia up to 40 per cent of cases of falciparum malaria are mefloquine-resistant.

Doxycycline

Doxycycline has been shown to give good protection against drug-resistant falciparum malaria in trials in Oceania, and it is being increasingly used, especially for those with adverse reactions to, or who are unwilling to take, mefloquine. It should not be used in children or pregnant women. The main side-effects are photosensitization, which occurs in up to 3 per cent of users making it less than ideal for beach holidays in the tropics, a tendency to precipitate attacks of candidiasis in women (hence it is helpful for women to take doxycycline with a one-dose therapy for candidal infections), and the rare risk of *Corynebacterium difficile* diarrhoea. However, doxycycline is likely to reduce the risk of the commoner travellers' diarrhoeas; but gastrointestinal discomfort from the doxycycline itself is not uncommon. The drug is taken daily with food, taking care not to miss any days, but avoiding lying down too soon after taking it to avert a real risk of acute pain from ulceration of the lower oesophagus. It is best to start a few days before travel: this is to get accustomed to taking the daily medication rather than for pharmacological reasons.

Atovaquone–proguanil ('Malarone')

This combination, which has been successfully used for malaria treatment for several years, has now been licensed in the United Kingdom (and previously in the United States) and Europe for malaria prophylaxis. It appears to have two great advantages: the level and severity of adverse effects has so far been lower than for the mefloquine and doxycycline; and, in part, it acts as a causal prophylactic, attacking the pre-erythrocytic stages of the malarial parasites. As a consequence, it is continued for 7 days after leaving the malarious area, so that the chance of compliance with this shorter period is improved. There are two concerns over it at present: although it appears to afford comparable protection as the alternative drugs against falciparum malaria, the evidence in non-immune individuals is scanty, and it is uncertain how soon resistance to this drug combination will emerge. Although resistance to atovaquone alone readily occurs, resistance to the combination is a much rarer event. Malarone is extremely expensive. However, since it is taken daily the different overall regimen means that the cost for short visits is more comparable to the alternatives, but the cost rises greatly for longer visits (and it is currently licensed for up to 4 weeks abroad). There is no experience of its use in pregnancy and the licence for its use in Europe currently excludes pregnancy and childhood, though it is used for children in the United States where paediatric tablets are available.

Continuation of chemoprophylaxis after leaving the malarious area

All antimalarial agents except Malarone must be continued for 4 weeks after leaving the malarious area.

Choice of chemoprophylaxis (Table 10)

Where there is a substantial risk of chloroquine-resistant falciparum malaria, either mefloquine, doxycycline, or Malarone are appropriate, so providing a better range of protective options than a few years ago for healthy adults. However, of these only mefloquine is licensed for children, and none is ideal for pregnant women who are best advised to avoid such areas. Doses of prophylactic antimalarial drugs for children are given in Table 11.

Chemoprophylaxis in people with epilepsy

In patients with epilepsy, proguanil or atovaquone/proguanil or doxycycline do not increase the risk of fits and can be used for prophylaxis, depending on the particular geographical area and level of risk.

The fixed drug combination Maloprim (12.5 mg pyrimethamine and 100 mg dapsone per tablet), marketed as Deltaprim in parts of Africa, has been of value in patients with epilepsy and for others unable to take the first-line drugs. It is now hard to obtain, and it is important not to confuse Malarone and Maloprim. Maloprim alone gives poor protection against *P. vivax* and chloroquine may be given concurrently. The dose of Maloprim must not exceed one tablet a week or the incidence of the otherwise uncommon side-effect, agranulocytosis, rises. Methaemoglobinaemia occasionally occurs with Maloprim chemoprophylaxis.

Rejected chemoprophylactic drugs

The following drugs are unsuitable for chemoprophylaxis (but Fansidar has a role in treatment): amodiaquine because of the high risk of agranulocytosis; Fansidar (25 mg pyrimethamine and 100 mg sulfadoxine per tablet) because of the frequency of severe skin reactions; and pyrimethamine on its own, because it is ineffective in most malarial areas.

Risk of malaria and need to take chemoprophylaxis

The risk of malaria is much higher in sub-Saharan Africa than elsewhere and it would be folly not to take prophylactics, except where the altitude is too great for transmission to occur or in the non-endemic southern parts of the continent (see Fig. 4). In Asia, the risk is usually much lower. Visitors to the air-conditioned hotels of the larger cities of SE Asia do not need prophylaxis but elsewhere in Asia there may be urban malaria. Mefloquine does not protect adequately against malaria in SE Asia, and travellers to the areas of higher transmission will need regimens (c)2 or (c)3 (Table 10). Those residing for long periods in such areas may prefer to adopt vigilance and the early treatment of fevers, but awareness of the risk is essential. Freedom from malaria in Asia by travellers does not mean that they will escape infection in Africa!

Because no prophylactic is completely effective in chloroquine-resistant *P. falciparum* areas, travellers who may be in remote areas and away from prompt medical assistance should carry a therapeutic dose of Fansidar, Malarone, mefloquine, or Riamet/Co-artem ether. Resistance to Fansidar has been reported from many countries with highly chloroquine-resistant malaria. The prophylactic regimen used should be continued for the appropriate time, usually 4 weeks, after returning to a non-endemic area. Compliance is hard to achieve, but this will prevent most cases of imported malaria. However, no regimen is 100 per cent protective and whatever precautions are taken, the possibility of malaria must, however, be borne in mind by the traveller and pointed out to any medical adviser, whom he or she must seek in case of a fever.

Malarial vaccines

Difficulties facing the development of a malaria vaccine

No satisfactory vaccine has emerged from the many attempts, over the last 70 years, to immunize animals and humans against malaria. A major problem is the impracticability of producing large quantities of attenuated micro-organisms, the basis for most effective viral and bacterial vaccines. The alternative, a subunit vaccine, has proved much more difficult to produce. Other problems are attributable to biological features of the malaria parasite, selected by evolutionary pressure to enable it to persist long enough in the human host to be taken up by a mosquito and propagated. During the different stages of its lifecycle—in the bloodstream,

hepatocytes, and erythrocytes of the human host and in mosquitoes—*P. falciparum* expresses a variety of antigens. Antibodies elicited against sporozoites, the infective stage inoculated by the mosquito, will not recognize blood-stage antigens. A different set of immunizing antigens is therefore needed to target each stage of the lifecycle. The large and complex genome of *P. falciparum* (25–30 megabases with 5–6000 genes, many of them polymorphic, on 14 chromosomes) shows great diversity, and an attack of malaria may involve simultaneous infection with as many as eight different *P. falciparum* genomes. Antigenic variation of some parasite proteins, such as the high molecular weight PfEMP-1 on the surface of infected erythrocytes, enables *P. falciparum* to evade the host's immune response. Another problem facing the widespread use of a malaria vaccine is the variation in the innate genetic resistance of humans to the pathological effects of malaria infection, related, for example, to their MHC class I polymorphism (Table 4). The immune response to vaccines may also be determined genetically.

Pre-erythrocytic stage vaccines

Irradiation-attenuated sporozoites

The first successful attempt to immunize a human against malaria was reported by DF Clyde and his colleagues in 1973. Their technique was based on studies in mice infected with *P. berghei*, in which protective immunity had been achieved by repeated intravenous injections of live, irradiation-attenuated sporozoites or by exposure to bites by irradiated infected mosquitoes. In mice, protection was associated with the development of precipitating antibodies to circumsporozoite antigens. Over a period of 84 days, three healthy adult volunteers were exposed, on six occasions, to bites by *Anopheles stephensi* mosquitoes which had been irradiated after becoming heavily infected with sporozoites of the Burma (Thau.) strain of *P. falciparum*. After being challenged through bites by non-irradiated mosquitoes bearing the same strain of *P. falciparum*, 98 days after the start of immunization, two of them developed malaria but one remained uninfected. This uninfected man was exposed to bites by irradiated mosquitoes on five further occasions, after which he was challenged again on day 327. Again, he failed to develop parasitaemia but was not protected against an intravenous injection of blood-stage parasites of the same strain, illustrating the stage-specific nature of malarial immunity. Work over the next 20 years confirmed the principle that protective immunity could be induced in humans by bites of irradiated infected mosquitoes. Between 1989 and 1999 further studies were carried out of immunization by the bites of irradiated *P. falciparum*-infected mosquitoes. A group of 11 volunteers, immunized by receiving more than a thousand bites from irradiated mosquitoes harbouring infectious sporozoites of *P. falciparum* strain NF54 and the 3D7 clone of NF54, were protected against 33 out of 35 challenges through bites by non-irradiated infected mosquitoes. Protection lasted for at least 36–42 weeks and extended to parasitic strains from geographical areas different from the immunizing strains.

These studies of artificial infections by mosquito-borne attenuated sporozoites provided the first evidence that vaccination against malaria was possible. However, such a prolonged, intensive, and laborious process could never become a practicable way of immunizing even small groups of non-immune travellers, let alone endemic populations. The immunological mechanism of protection conferred by irradiated sporozoite immunization has been studied in the mouse model and in human volunteers. In mice, CD8+ T-cell recognition of infected hepatocytes, by targeting sporozoite proteins expressed within the cell, is the most important mechanism. Humoral antibodies to proteins on the sporozoite's surface and CD4+ T-cells may also play a role.

Effector T-cell vaccines

An encouraging recent development, based on these findings, has been the preparation and testing of effector T-cell vaccines targeting pre-erythrocytic stages of the lifecycle, in infected hepatocytes. Theoretically, these vaccines could prevent both blood-stage infection and transmission in malaria endemic areas. The two most productive strategies have been the use of protein-adjuvants (for example, in RTS,S/(SB)AS02 vaccine) and heterologous prime–boost immunization.

RTS,S/(SB)AS02 malaria vaccine

RTS,S is a fusion protein combining most of the circumsporozoite protein of *P. falciparum* with Hb_sAg with a complex adjuvant (AS02) capable of inducing strong antibody and CD4+ T-cell responses. This vaccine protected 50 per cent of volunteers challenged within 2 to 3 weeks of their last immunization, but after 6 months, only one in five was protected. Field trials in The Gambia showed an overall efficacy against infection during the whole surveillance period of 34 per cent (95 per cent confidence interval (CI), 8.0–53 per cent; $p = 0.014$). Efficacy was 71 per cent (46–85 per cent) during the first 9 weeks but there was no protection after that. A single booster vaccination led to similar protection during the next malaria season (efficacy, 47 per cent (3.8–71 per cent); $p = 0.037$). Protection correlated with a short-lived vaccine peptide-specific CD4+ T-cell response. It is hoped to improve this vaccine by modifying the adjuvant, by boosting with a vaccinia recombinant circumsporozoite protein, and by the addition of a blood-stage (MSP-1) antigen. Trials in Gambian children are underway.

Heterologous prime–boost immunization

AVS Hill and his colleagues have pioneered the strategy of priming with a DNA-based vaccine and boosting with a recombinant poxvirus. This is particularly effective in inducing CD8+ cytotoxic T lymphocytes and in enhancing TH1-type CD4+ T-cell responses, both of which are associated with protection. The DNA vaccine encodes a number of sporozoite epitopes together with the entire thrombospondin-related adhesion protein (**TRAP**), while the poxvirus recombinant consists of a highly attenuated vaccinia virus strain (Modified Vaccinia [Virus] Ankara, **MVA**)—which does not replicate in mammalian cells—containing the same malaria insert. Phase I and II studies in Oxford and The Gambia have confirmed the safety and immunogenicity of the regime and challenge studies are in progress. An even more promising regimen, based on mouse studies, consists of priming with a Fowlpox (Avipox FP9) recombinant, and boosting with the MVA recombinant.

(Table 11)

Contains same material by C. Newbold from previous editions.

Further reading

Bates I, *et al.* (1991). Use of immunoglobulin gene rearrangements to show clonal lymphoproliferation in hyper-reactive malarial splenomegaly. *Lancet* **337**, 505–7.

Beadle C, *et al.* (1994). Diagnosis of malaria by detection of *Plasmodium falciparum* HRP-2 antigen with a rapid dipstick antigen-capture assay. *Lancet* **343**, 564–8.

Berendt AR, *et al.* (1994). Molecular mechanisms of sequestration in malaria. *Parasitology* **108**, S19–28.

Bradley DJ, Bannister B (2001). Guidelines for malaria prevention in travellers from the United Kingdom for 2001. *Communicable Disease and Public Health* **4**, 84–101.

Hill AVS, Weatherall DJ (1998). Host genetic factors in resistance to malaria. In: Sherman IW, ed. *Malaria: parasite biology pathogenesis protection*, pp 445–55. ASM Press, Washington DC.

Hoffman SL, *et al.* (2002). Protection of humans against malaria by immunization with radiation-attenuated *Plasmodium falciparum* sporozoites. *Journal of Infectious Diseases* **185**, 1150–64.

Koch O, *et al* (2002). IFNGRI gene promotor polymorphisms and susceptibility to cerebral malaria. *Journal of Infectious Disease* **185**, 1684–7.

Kwiatkowski D, *et al* (1990). TNF concentration in fatal cerebral, non-fatal cerebral, and uncomplicated *Plasmodium falciparum* malaria. *Lancet* **336**, 1201–4.

MacPherson GG, *et al.* (1985). Human cerebral malaria: a quantitative ultrastructural analysis of parasitized erythrocyte sequestration. *American Journal of Pathology* **119**, 385–401.

Marsh K, *et al* (1995). Indicators of life-threatening malaria in African children: clinical spectrum and simplified prognostic criteria. *New England Journal of Medicine* **332**, 1399–404.

Miller LH (1994). Impact of malaria on genetic polymorphism and genetic diseases in Africans and African Americans. *Proceedings of the National Academy of Sciences (USA)* **91**, 2415–19.

Nardin EH, Nussenzweig RS (1993). T cell responses to pre-erythrocytic stages of malaria: role in protection and vaccine development against pre-erythrocytic stages. *Annual Reviews of Immunology* **11**, 687–727.

Ockenhouse CF (1993). The molecular basis for the cytoadherence of *Plasmodium falciparum*-infected erythrocytes to endothelium. *Seminars in Cell Biology* **4**, 297–303.

Riddle MS, *et al* (2002). Exchange transfusion as an adjunct therapy in severe *Plasmodium falciparum* malaria: a meta-analysis. *Clinical Infectious Diseases* **34**, 1192–8.

The Artemether–Quinine Meta-analysis Study Group (2001). A meta-analysis using individual patient data of trials comparing artemether with quinine in the treatment of severe falciparum malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **95**, 1–14.

Turner GDH, *et al* (1994). An immunohistochemical study of the pathology of fatal malaria. *American Journal of Pathology* **145**, 1057–69.

Warrell DA, Gilles HM (2002). *Essential malariology*, 4th edn. Arnold, London.

Warrell DA, *et al* (1982). Dexamethasone proves deleterious in cerebral malaria. A double-blind trial in 100 comatose patients. *New England Journal of Medicine* **306**, 313–19.

Wernsdorfer WH, McGregor IA (1988). *Malaria. Principles and practice of malariology*. Churchill Livingston, Edinburgh.

White NJ and Ho M (1992). The pathophysiology of malaria. *Advances in Parasitology* **31**, 83–173.

White NJ, *et al* (1983). Severe hypoglycemia and hyperinsulinaemia in falciparum malaria. *New England Journal of Medicine* **309**, 61–6.

World Health Organization (2000). Severe falciparum malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **94**(Suppl. 1), 51–90.

World Health Organization (2002). *International travel and health*. WHO, Geneva.

7.13.3 Babesia

P. Brasseur

Babesia are intraerythrocytic, tick-transmitted, protozoan parasites that infect a broad variety of wild and domestic animals including cattle, horses, dogs, and rodents. Human babesial infection may occur occasionally.

Epidemiology

Two species of *Babesia*, *B. microti* and *B. divergens*, are responsible for most human cases. More than 200 cases of *B. microti* infections have been reported since 1966 along the north-east coast of the United States, especially in Massachusetts including Nantucket Island, Martha's Vineyard, and Cape Cod. *B. microti* is transmitted by *Ixodes dammini* and the reservoir host of parasites is the common white-footed mouse, *Peromyscus leucopus*. The zoonotic *Borrelia burgdorferi* causing Lyme disease is also transmitted by *I. dammini*; coinfections are documented among residents of coastal New England, where the risk of both babesiosis and Lyme disease is highest in June when nymphal *I. dammini* are most abundant. *B. microti* babesiosis may occur in people with intact spleens as well as in asplenic subjects.

After the first description of a case in 1957, 28 additional cases have been documented in Europe. Seventy six per cent of cases were due to *B. divergens*, a common cattle pathogen transmitted by *Ixodes ricinus* and responsible for economic losses, by reducing weight gains and milk production. France, the British Isles, and Ireland account for more than 50 per cent of European cases. They usually occur between May and October, the season of activity of tick vectors such as *I. ricinus*, which often seems to be responsible for human transmission. Most patients were residents of rural areas such as farmers and foresters, or visitors such as campers and hikers. Splenectomized people are at highest risk, comprising 24 out of 29 patients with babesiosis including all 22 *B. divergens* cases. Although no transfusion-associated case has been reported in Europe, this route of transmission is possible because *B. divergens* survives in packed red blood cells for several weeks at 4°C. No case has been recorded among HIV-infected patients.

Pathogenesis

Ticks infected with *Babesia* inoculate parasites while feeding on a vertebrate, the parasites enter red blood cells directly and multiply by budding to form two or four parasites, rarely more, in about 8 to 10 h. These are released and will invade other erythrocytes. The spleen plays a major role in resistance to babesial infections, especially for *B. divergens* babesiosis.

Clinical features

Human *B. microti* babesiosis is characterized by a gradual onset of malaise, anorexia, and fatigue with subsequent development of fever, drenching sweats, and generalized myalgia appearing 1 to 4 weeks after a tick bite. Other clinical manifestations such as headache, shaking chills, nausea, depression, and hyperaesthesia have been observed less frequently. The only finding on clinical examination is occasional mild hepatosplenomegaly. Anaemia, thrombocytopenia, and generally a low or normal white blood cell count is observed. A mild to severe haemolytic anaemia is frequent. Lactate dehydrogenase, liver enzymes, and bilirubin levels may be increased. Parasites are found in the peripheral blood of 1 to 20 per cent of patients with intact spleens, but in up to 80 per cent of asplenic patients. Most patients infected with *B. microti* have no history of splenectomy, but splenectomized patients generally have a more severe illness. Babesiosis is more severe in people over 40 years old and in HIV-infected patients. The acute illness lasts from 1 to 4 weeks, but weakness and malaise often persist for several months. A low and asymptomatic parasitaemia may persist several weeks after recovery.

In Europe, babesial infections are usually more severe (in 76 per cent of cases, with 38 per cent mortality) than in North America. After an incubation period of 1 to 3 weeks, severe intravascular haemolysis begins suddenly, causing haemoglobinuria, severe anaemia, and jaundice, associated with non-periodic high fever (40 to 41°C), hypotension, shaking chills, intense sweats, headache, myalgia, lumbar pain, abdominal pain, vomiting, and diarrhoea (Plate 1). Peripheral blood *B. divergens* parasitaemia may vary from 5 to 80 per cent. Patients rapidly develop renal failure which may be associated with pulmonary oedema, coma, and death. In severe cases, haemoglobin falls to 7 to 8 g/dl, sometimes to 4 g/dl, in spite of blood transfusions. Plasma haemoglobin levels may exceed 4 g/dl, haptoglobin decreases dramatically, and bilirubin and liver enzymes are markedly elevated.

Diagnosis

Babesiosis should be suspected in any patient with fever and a history of tick bite from any area. Initially, *Plasmodium falciparum* infection may be suspected, but splenectomy, lack of recent travel to a malaria-endemic area, or blood transfusion should lead to suspicion of babesiosis. Diagnosis is based on discovery of parasites in Giemsa-stained thin blood smears. Although the variable morphology of the parasites may be confusing, *Babesia* species can be distinguished from malaria parasites by the absence of gametocytes and pigment in erythrocytes containing mature stages.

B. microti is characterized by multiple basket-shaped parasites. In some cases, parasitaemia is sparse and inoculation of patient's blood into hamsters may facilitate diagnosis. This method may detect parasitaemias as low as 300 parasites/ml. Amplification by polymerase chain reaction using species-specific primers may establish the diagnosis in 24 h with high specificity and sensitivity.

B. divergens infection is suspected if there are symptoms of intravascular haemolysis and renal failure. The presence of double piriform intraerythrocytic parasites or tetrads is typical of *B. divergens*, but annular, punctiform, and filamentous forms may also be encountered. Inoculation of infected blood in gerbils (*Meriones unguiculatus*) may confirm the diagnosis. Serological tests may be useful especially in *B. microti* infections, but correlation between antibody titres and severity is poor. Using an indirect immunofluorescent test, antibody titres rise during the first weeks and fall after 6 months. Serology is not used for rapid diagnosis of *B. divergens* infection.

Treatment and prevention

Chloroquine, sulphadiazine, pyrimethamine, co-trimoxazole, pentamidine, and berenil (diminazene aceturate) appear ineffective in completely eliminating *B. microti*. A combination of quinine and clindamycin is effective except in immunocompromised individuals, especially those with HIV. Quinine should be given orally in doses of 650 mg every 6 to 8 h daily and clindamycin intravenously at 1200 to 2400 mg in three or four divided doses daily for at least 7 to 10 days. In fulminating infection, exchange transfusion is recommended.

In Europe, babesiosis should be treated as a medical emergency. Immediate chemotherapy should reduce parasitaemia and prevent extensive haemolysis. Exchange transfusion should be considered at the first sign of *B. divergens* infection. Massive exchange transfusion (2 to 3 total blood volumes) followed by administration of intravenous clindamycin at a dose of 600 mg four times daily for at least 10 days has proved successful. Imidocarb, which is used to treat babesiosis in cattle, has been used successfully in two patients in Ireland, although this drug has not been approved for human use. Atovaquone is active *in vitro* and in gerbils, but has not yet been used in human *B. divergens* infections.

Further reading

Telford III SR *et al*. (1993). Babesial infections in human and wildlife. In: Kreier JP, ed. *Parasitic protozoa*, Vol. 5. Academic Press, New York.

Pruthi RK *et al.* (1995). Human babesiosis. *Mayo Clinic Proceedings* **70**, 853.

7.13.4 Toxoplasmosis

J. Couvreur and Ph. Thulliez

Parasitology, epidemiology, transmission

Toxoplasma gondii is a ubiquitous coccidian parasite. Its definitive host is the cat. It exists in three forms: (i) the oocyst, which is excreted with the cat faeces, can remain viable for months in the soil under certain conditions of temperature and humidity; (ii) the tachyzoite, which multiplies intracellularly (Fig. 1); and (iii) cysts, the result of this intracellular multiplication, which can persist as viable parasites in the brain and striated muscles throughout the life of the host.

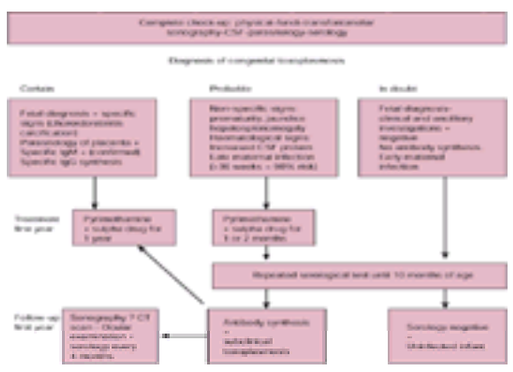

Fig. 1 Algorithm for neonates at risk of toxoplasmic fetopathy (mother infected during pregnancy).

The prevalence of toxoplasma antibodies is high in most populations. It depends mostly upon eating habits. Prevalence is well defined in fertile women—for instance 72 per cent seropositivity in Paris, 36 per cent in Stuttgart, 54 per cent in Padua, and 21 per cent in London.

Toxoplasmosis is usually acquired by ingestion of cysts. There are four stages: acute, subacute, chronic, and relapses. Organisms spread from the gut by lymphatics and the bloodstream, reaching every organ, where they multiply intracellularly (acute stage). Termination of this stage depends upon the development of both cellular and humoral immunity. In immunocompetent hosts, the parasite encysts and will persist without any inflammatory process as long as the cysts are not disrupted (chronic stage). If the host is or becomes immunocompromised, there is a tendency for the cysts to release bradyzoites and toxoplasma becomes an opportunistic agent. Congenital infection occurs through transplacental transmission of tachyzoites when a previously uninfected woman is infected during pregnancy.

Acute acquired toxoplasmosis

Acquired toxoplasmosis is usually subclinical. The typical presentation is lymphadenopathy affecting posterior cervical, suboccipital, retroauricular, or submental nodes. Supraclavicular, axillary, pectoral, epitrochlear, and inguinal localizations are less frequent. Lymphadenopathy is usually localized but it can be generalized. Nodes can be painful and tender for 1 or 2 weeks. They are rarely larger than walnuts, smooth, well defined, and mobile. They never suppurate; they can persist for months and even a year. Mesenteric lymphadenopathy has been observed. Other clinical signs and symptoms are fatigue for several weeks, headache, myalgias, low-grade fever for one or several weeks, and rarely, a transient rash. Ocular involvement can be observed. Hepatomegaly and splenomegaly are rare. Neurological signs and myocarditis are exceptional in immunocompetent patients.

The blood count shows a relative neutropenia with lymphocytosis. Atypical lymphocytes indistinguishable from those of infectious mononucleosis may be seen. Inversion of the CD4/CD8 ratio has been noted more often in clinical than in subclinical toxoplasmosis. Features suggesting toxoplasmosis rather than infectious mononucleosis are absence of pharyngitis, oral petechias, and splenomegaly and a less marked but more persistent lymphadenopathy. The histological pattern is characteristic when there are groups of epithelioid cells scattered throughout the node, or immature sinus histiocytosis and follicular hyperplasia with phagocytosis and nuclear debris. Inflammatory infiltrates sometimes extend into perinodal tissues and may be misinterpreted as lymphangioma, lymphoma, or sarcoidosis.

Toxoplasmosis of the central nervous system

In acquired toxoplasmosis, central nervous system involvement is observed most commonly in the immunodeficient patient. The selectivity of toxoplasma for brain tissue has been ascribed to low local immunity. In animal experiments, tachyzoites are demonstrable in brain 5 days after intraperitoneal inoculation, with ensuing perivascular inflammation with mononuclear cells. Tachyzoite-infected cells provoke multiple foci of micronecrosis. Mononuclear cells gather into microglial nodules associated with toxoplasma antigen. Cysts appear away from the inflammatory process. Intermediate appearances are observed between disseminated foci of microglial nodules, more or less numerous large necrotic areas, and large space-occupying masses.

Clinical features of central nervous damage are protean and may develop insidiously: generalized encephalitis with meningeal involvement and localizing signs with fever, headache, drowsiness progressing into coma, and death within a few days or weeks; encephalitis with low-grade meningeal involvement; 'pseudotumour cerebri' syndrome with transient intracranial hypertension; space-occupying mass mimicking a tumour or a brain abscess; multiple mass lesions; miscellaneous patterns—confusion, psychiatric features, seizures, and signs of brainstem or spinal cord injury. The above patterns can progress to death fulminantly within 2 weeks or persist for months or even years with or without therapy (chronic relapsing encephalitis).

The diagnosis of toxoplasmosis of the nervous system is often difficult. Clinical signs and results of imaging are not specific and can be misleading. Serological data are often perplexing. Biopsy is advocated whenever the diagnosis is uncertain. It is mandatory to look for an underlying disease or an immunodeficiency.

Ocular toxoplasmosis

Toxoplasma infection is the most common cause of retinochoroiditis and posterior uveitis. The focal necrotizing retinitis in its acute or subacute stage appears as cottonwool-like patchy areas of the fundus with vitreous exudate. The lesion heals within 3 to 6 weeks leaving a punched-out scar with central atrophy and a peripheral black pigmentation. The lesion can be peripheral or central, single or multiple. It may reach the size of the optic disc. Atypical presentations include retinal

detachment, haemorrhage, and optic nerve injury.

The natural history of ocular toxoplasmosis suggests that the first retinal lesion occurs more commonly during the subacute stage, weeks or months after the begining of the infection, than later during the chronic stage of the infection. It results from a previous colonization of the retina. The immediate cause is the rupture of a cyst. Its mechanism involves delayed sensitivity to toxoplasma antigens, and secondary proliferation of parasites.

Congenital toxoplasmosis is currently considered the major cause of ocular toxoplasmosis. Ninety per cent of retinochoroiditis discovered in infants and young children, and at least 20 per cent in adults, is attributable to congenital toxoplasmosis. It can be seen at birth or may occur much later, even in a previously normal retina, as is the case for 35 to 85 per cent of children with untreated congenital toxoplasmosis. There is a peak frequency of new lesions during puberty and adolescence. The common presenting signs are amblyopia or strabismus. There is some evidence that early treatment of even subclinical congenital toxoplasmosis decreases this risk. Ocular disease can complicate acquired toxoplasmosis more often than was previously considered. It can occur early following the acute stage of the infection, or after 2 years or more in one-third of cases. It is unilateral, with relapses in one-third of cases. It is generally isolated, being associated with neurological signs in no more than 10 per cent of cases, and most often without underlying immunodeficiency.

The diagnosis of ocular toxoplasmosis cannot be based on fundoscopic examination alone. The fact that it occurs mainly during the chronic stage of the infection while the antibody titre is low is a major problem. This can be solved by the demonstration of a local synthesis of antibodies in the aqueous humour.

Toxoplasmosis and immunodeficiency

Any patient with severe toxoplasmosis should be investigated for an immune defect, even subtle, and for an underlying disease, particularly AIDS. Conversely an immunodeficiency, either spontaneous or iatrogenic, can be complicated by severe toxoplasmosis. The last is generally related to chronic infection. The long persistence of cysts, particularly in brain, striated muscles, and myocardium is a well documented fact. In animals with chronic infection, corticoids, irradiation, or immunodeficiency can induce cyst rupture and proliferation of toxoplasma in the nervous tissue.

Among malignancies, the most common condition associated with cerebral toxoplasmosis is Hodgkin's disesase and less often lymphoproliferative disorders such as lymphosarcoma, non-Hodgkin's lymphoma, and angioimmunoblastic lymphadenopathy. All types of leukaemia are involved.

Severe toxoplasmosis can occur in organ transplant recipients. It is rare in renal, liver, and bone marrow transplantations but is more frequent in heart and heart–lung transplantations in which the risk can reach 57 per cent in mismatched transplantations of a serologically positive donor and negative recipient. This risk is mainly related to the infected heart-tissue transplant. It is increased by the use of steroids for graft rejection. Conversely, cyclosporin has an antiparasitic activity. The risk is reduced by antiparasitic treatment in all recipients.

Severe toxoplasmosis of the central nervous system, lungs, and heart appeared as a major problem in patients with AIDS. Toxoplasma encephalitis was observed in 25 to 80 per cent of patients with signs of cerebral injury. The risk of such an encephalitis was 6 to 12 per cent in toxoplasma seropositive patients and it was definitely increased by several factors: late stage of the disease; the presence of antibodies, particularly if the IgG titre was more than 150 IU; and a CD4 lymphocyte count of less than 200 per m. The risk of cerebral toxoplasmosis has been markedly reduced by the administration of co-trimoxazole for prevention of pneumocystis in these patients.

It is often difficult to prove that clinical manifestations encountered in immunocompromised patients are attributable to toxoplasmosis. Antibody titres may not be significantly elevated. Attempts to isolate toxoplasma from cerebrospinal fluid, from myocardial or cerebral biopsies, or from bronchoalveolar lavage may be necessary.

Congenital toxoplasmosis

Maternofetal transmission

Following seroconversion during pregnancy, 31 per cent of infants are infected, 2 per cent suffer intrauterine death, but 67 per cent are uninfected. These overall data vary according to the date of maternal infection: before pregnancy, 0 per cent; during the first month, 1 per cent; during the second and third month, 17 per cent; from the fourth to the sixth month, 45 per cent as an average; and later an increased risk up to 80 per cent during the ninth month. The date of the maternal infection is also important for the clinical pattern of the fetopathy: 83 per cent of the fetuses infected during the first trimester have clinical involvement, often severe, while clinical signs, mostly mild and ocular, are seen in only 12 per cent of the infants whose mothers were infected during the ninth month. The risk of fetopathy is reduced by more than 50 per cent if spiramycin is given to the mother. Very rare cases of transmission following chronic infection even years before pregnancy have been observed in immunocompromised mothers.

The placenta is the transmitting organ and placental infection is synonymous with fetopathy. If maternofetal transmission occurs early after maternal infection, fetopathy will be severe. If it is delayed, the fetus will be protected by passively transmitted maternal antibodies and toxoplasma will have a tendency to encyst in fetal tissues without causing serious early injury. Serological and clinical progression may thus be delayed for months after birth.

Clinical patterns

Five patterns can be identified in the protean presentation of congenital toxoplasmosis.

1. Systemic disease of the newborn baby with rash, jaundice, thrombocytopenic purpura, hepatosplenomegaly, pneumonia, progressive uveitis, high protein content of cerebrospinal fluid, cerebral ventricular dilation, and encephalomyelitis.
2. Neurological disease: hydrocephalus or microcephaly, microphthalmia, retinochoroiditis, and cerebral calcification. Hydrocephaly, always related to a stenosis of the duct of Sylvius, can be discovered *in utero* or several months after birth as well as in an infant initially considered as normal. Shunting is required.
3. Mild disease with isolated retinochoroiditis or mild cerebral calcification without any clinical signs of cerebral injury.
4. Subclinical infection. Prospective studies of women with acquired infection during pregnancy revealed that this is the most common pattern encountered in more than 70 per cent of the infected babies. The differentiation between subclinical toxoplasmosis and absence of infection is a common challenge for the paediatrician.
5. Relapses: flare-ups of retinochoroiditis can occur in infants, children, adolescents, or adults even in a previously intact retina in up to 85 per cent of cases (see Ocular toxoplasmosis). The possibility of late relapses in cerebral tissue is confirmed by the frequency of increased local synthesis of antibodies in the cerebrospinal fluid (see Laboratory diagnosis). Complete work-up, particularly examination of the cerebrospinal fluid is mandatory in any form of congenital toxoplasmosis even when subclinical.

Laboratory diagnosis

Serological methods are the main tools for diagnosis, but in the fetus and the immunocompromised patient the demonstration of parasites in body fluids and tissues is the preferred method of diagnosis.

Serology

In a non-immune pregnant woman who is tested repeatedly throughout pregnancy, seroconversion definitely proves the acquisition of infection. In the absence of seroconversion, the diagnosis of recent infection requires the demonstration of a significant rise of IgG antibody titre and the presence of IgM in serial samples obtained at 3-week intervals and tested in parallel. A stable IgG titre is consistent with an infection acquired at least 2 months before the first specimen was obtained. Since IgM antibodies may be detected for over a year after the infection, the use of complementary methods, based on acute-phase IgG antibodies, is necessary to rule out a recent infection, particularly in women who are evaluated late in pregnancy. The differential AC/HS agglutination test and the measurement of IgG avidity in enzyme immunoassay are suitable for this purpose.

In the newborn baby, the detection of specific IgM after 2 days of life or of specific IgA after 10 days is diagnostic of congenital infection. Synthesis of anti-toxoplasma

IgG antibodies can be demonstrated by comparing the ratio: specific IgG titre/total IgG on monthly serial samples. In the absence of infection, this ratio decreases as the infant produces IgG that does not contain toxoplasma-specific antibodies. If the ratio remains the same or increases, the diagnosis is proved. Synthesis of specific IgG, IgM, or IgA can also be demonstrated by immunoblotting or by using enzymze-linked immunofiltration assay. In some cases, the only marker of congenital infection is production of specific IgG which may be delayed for several months. Consequently, in infants born to women infected during pregnancy, serological tests must be repeated until specific IgG disappears within the first 12 months of life, before ruling out a congenital infection.

In HIV-infected patients, tests for determination of specific antibodies must be sensitive enough to avoid underdiagnosing a latent infection which should be considered for specific prophylaxis.

Intrathecal or intraocular production of specific antibodies can be determined by comparing the ratio of specific to total IgG in cerebrospinal fluid or aqueous humour with that of serum. A coefficient higher than 3 is considered positive.

Detection of *Toxoplasma gondii*

Parasites can be isolated from tissues or biological fluids by inoculation into mice or into cell cultures. Isolation from blood or cerebrospinal fluid indicates the presence of an acute infection. Conversely, positive isolation from muscle after enzymatic digestion, from brain, or from heart tissues is possible in old, chronic infections and does not prove a recent or progressing infection. Positive isolation from the placenta is indicative of congenital infection.

Polymerase chain reaction can be used to detect toxoplasma DNA in various clinical samples. It has proved reliable for diagnosis in immunocompromised patients. It is the method of choice for prenatal diagnosis of congenital infection on a single sample of amniotic fluid. The method is rapid and specific provided that carry-over contamination of the samples and contamination risks associated with handling steps are avoided. This is the most sensitive diagnostic method although all congenital infections cannot be identified prenatally because a delayed transmission of toxoplasma from the placenta to the fetus may occur after the date of the amniocentesis.

Treatment

Drugs

The combination of pyrimethamine and sulpha drugs is the mainstay of treatment. Pyrimethamine is given orally in a daily dose of 1 mg/kg or 50 mg in adults. The dosage of sulphadiazine, the sulpha drug currently used, is 50 to 100 mg/kg per day in infants, up to 2 to 6 g in adults in two to four divided doses. It is necessary to monitor weekly the antiparasitic treatment with blood counts because of the risk of bone marrow depression resulting in leucopenia with pyrimethamine or leucopenia and granulopenia with sulphadiazine. Folinic acid at a dose of 50 mg by oral or intramuscular route every 3 to 6 days can prevent the pyrimethamine side-effects.

The combination of 25 mg of pyrimethamine and 500 mg of sulphadoxine (Fansidar) may be given orally in a dose of one tablet per 20 kg every 7 to 10 days for months. Other drugs of interest are atovaquone, co-trimoxazole, or macrolides—spiramycin given in a daily oral dose of 3 g to infected pregnant women to prevent maternofetal transmission of the parasite or clarithromycin, clindamycin, and azithromycin in various combinations.

Indications

Acquired toxoplasmosis

Indications for treatment are marked systemic symptoms, and evidence of organ involvement. The pyrimethamine–sulphadiazine combination can be given for one or several weeks according to the clinical pattern.

Immunocompetent pregnant women

Seroconverters are given spiramycin throughout pregnancy. A positive *in utero* diagnosis of fetopathy warrants pyrimethamine–sulphadiazine treatment with written consent. Serial ultrasound examination of the fetus is mandatory to detect cerebral involvement as a guide to elective termination.

Congenital toxoplasmosis

Any case, even if subclinical, must be treated to control active disease and/or to prevent the risk of secondary retinochoroiditis. The combination pyrimethamine–sulphadiazine is given daily for 3 to 6 months according to the clinical data, followed by treatment three times a week until 1 year of age.

Cerebral toxoplasmosis in AIDS

In the acute stage of infection, the regimen of choice is 50 mg of pyrimethamine, 6 to 8 g sulphadiazine, and 20 mg folinic acid per day for 3 to 6 weeks. Maintenance therapy is necessary throughout life. The risk of toxoplasma encephalitis is markedly reduced by routine administration of co-trimoxazole.

Ocular toxoplasmosis

A flare-up requires emergency treatment with pyrimethamine–sulphadiazine together with steroids. The pyrimethamine–sulphadoxine combination can then be given for 6, 12, or more months if there is a tendency for repeated relapses.

Prevention

Any patient at risk should avoid contact with cats, as their faeces is potentially infectious (litter, soil, garden sand pits, vegetables), and eat meat that is well cooked or preserved by deep freeze.

Further reading

Couvreur J (1991). Foetopathie toxoplasmique: Traitement in utero par l'association pyrimethamine– sulfamides. *Archives Françaises de Pediatrie* **48**, 397–403.

Couvreur J, Desmonts G (1962). Congenital and maternal toxoplasmosis; a review of 300 congenital cases. *Developmental Medicine and Child Neurology* **4**, 519–30.

Couvreur J, Leport C (1998). *Toxoplasma gondii*. In: Yu VL, Meignan TC, Barriere NJS, eds. *Antimicrobial chemotherapy and vaccines*, pp 600–12. William & Wilkins, Baltimore.

Couvreur J *et al.* (1984). La production locale accrue d'anticorps dans le liquide cephalo-rachidien au cours de la toxoplasmose congénitale. *Annales de Pédiatrie (Paris)* **3**, 839–45.

Desmonts G, Couvreur J (1985). Congenital toxoplasmosis: a prospective study of 378 pregnancies. *New England Journal of Medicine* **318**, 271–5.

Hohlfeld P *et al*. (1994). Prenatal diagnosis of congenital toxoplasmosis with a polymerase-chain-reaction test on amniotic fluid. *New England Journal of Medicine* **331**, 695–9.

Leport C, Raffi F, Matheron S (1998). Treatment of central nervous system toxoplasmosis with pyrimethamine–sulfadiazine combination in 35 AIDS patients. Efficacy of long term continuous therapy. *American Journal of Medicine* **84**, 94–100.

McAuley J *et al*. (1994). Early and longitudinal evaluations of treated children and untreated historical patients with congenital toxoplasmosis; the Chicago Collaborative Treatment Trial. *Clinical Infectious Diseases* **18**, 38–72.

Remington JS, McLeod R, Thulliez P, Desmonts G (2001). Toxoplasmosis. In: Remington JS, Klein JO, eds. *Infectious diseases of the fetus and newborn infant*, 5th edn, pp. 205–346. Saunders, Philadelphia.

7.13.5 Cryptosporidium and cryptosporidiosis

D. P. Casemore and D. A. Warrell

Introduction

The cryptosporidia are obligate intracellular parasites of which primarily one species, *Cryptosporidium parvum*, is associated with infection in man, young livestock, and other mammalian species. First described in laboratory mice, by Tyzzer in 1912, *C. parvum* was first recognized as a cause of human infection in 1976. In the 1980s it emerged worldwide as a common cause of severe or life-threatening infection in severely immunocompromised patients, especially those with AIDS, and of acute, self-limiting gastroenteritis in otherwise healthy subjects, especially children.

Biology

Cryptosporidium spp. are members of the coccidia (phylum Apicomplexa) with oocysts that contain four sporozoites. The oocysts, an environmentally robust transmissible stage, are fully sporulated and infective when excreted. Cryptosporidia are monoxenous, that is, they complete their lifecycle in a single individual (Fig. 1). *C. parvum* is not tissue specific but shows a predilection for the lower ileum during the primary stages of infection.

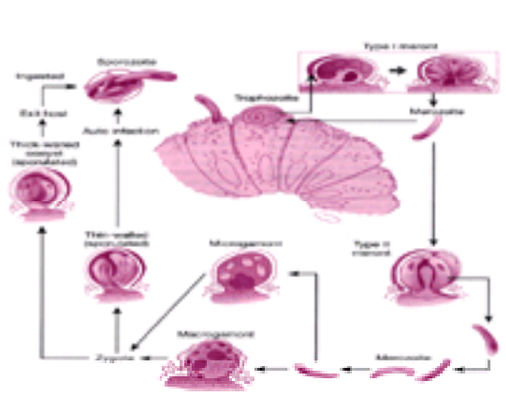

Fig. 1 Diagrammatic representation of the lifecycle of *C. parvum*. Following ingestion, the motile sporozoites are released, attach to cells, and develop into fixed trophozoites (uninucleate meronts) in an intracellular but extracytoplasmic location. These undergo schizogony (asexual multiple budding), the first-stage meronts producing eight merozoites, some of which recycle to form further type I meronts. Type II meronts produce four merozoites, which form gamonts (sexual stages) that mature as either macrogametes, or as microgamonts containing 16 motile microgametes. Most of the zygotes formed after fertilization develop into thick-walled, environmentally resistant, transmissible oocysts, which then sporulate, usually by the time they are excreted. Some have only a thin unit membrane, which ruptures to release the sporozoites *in situ* to produce an autoinfective cycle. (Adapted from a drawing by Kip Carter, University of Georgia, and shown by courtesy of Dr W.I. Current and CRC Press, Inc., Boca Raton.)

Following ingestion of oocysts, the motile sporozoites are released, through a suture in the oocyst wall, in the lumen of the small bowel. They quickly attach superficially to cells, rounding up to form fixed trophozoites (meronts). The initial site of infection is the brush border of enterocytes in the small bowel, but the parasite is able to infect other epithelial and parenchymal cells. The complex lifecycle includes both asexual and sexual stages of replication (Fig. 1 and Fig. 2). The endogenous (tissue) stages develop within a parasitophorous vacuole, the outer layer of which is derived from the host cell's outer membranes, in a unique intracellular but extracytoplasmic location.

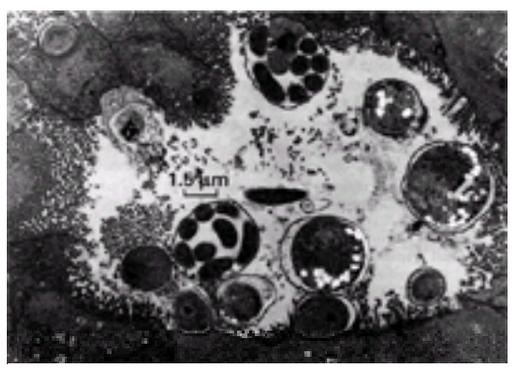

Fig. 2 Electron micrograph of a transverse section of small bowel of a mouse infected with *C. parvum*. The section shows numerous developmental stages: uninucleate meronts (trophozoites); type I meronts (schizonts) containing merozoites in which may be seen the darker granules of the apical complex organelles; the degenerate remains of a schizont and a free-swimming merozoite within the lumen; macrogamonts showing dark wall-forming granules and electron-lucent amylopectin (polysaccharide) food-storage granules. The parasitophorous vacuole can be clearly seen surrounding the parasite stages. Some of the intracellular

stages appear to be free within the lumen because of the plane of sectioning.

Molecular biology

Various genes and sequences have been studied and some have been cloned; chromosomes have been identified. Sequences coding for structural proteins such as actin and tubulin involved in parasite motility and attachment have been identified and may prove useful targets for the development of anticryptosporidial compounds. Nucleic acid sequences have differentiated recognized species and isolates within species, while some of the genotypes found may represent cryptic species. Differences between isolates of *C. parvum* from various sources show that those derived from animal hosts (genotype 2 or C) may, as expected, also be found in humans, but some isolates from humans appear to be relatively specific for humans (genotype 1 or H). Several other less common genotypes have been described, including from cats and dogs, the latter also having been found in a few immunocompromised humans. These findings have considerable public health significance and may also account for some of the variable responses seen in trials of specific anticryptosporidial therapy. Sensitive, genetically based probes are being applied to the detection of the parasite in clinical and environmental samples.

Epidemiology

C. parvum occurs worldwide and is common in humans and in livestock animals, especially lambs and calves, and has been reported from goats, horses, pigs, and farmed deer, as well as in mammalian wildlife. Prevalence in humans varies both geographically and temporally. Because of the diversity of host species, the epidemiology of the human infection is complex and involves both direct and indirect routes of transmission from animals to man (zoonotic transmission) and from person to person ('urban' cycle).

Direct zoonotic infection

Transmission from livestock is common, particularly in children, including those from urban homes and schools visiting educational farms and rural activity centres. Household pets are an infrequent source of infection in otherwise healthy subjects. Cryptosporidiosis is rarely seen in adults in rural areas, presumably as a result of frequent exposure and the development of immunity.

Urban transmission

Infection is common in children attending playgroups and day-care centres and outbreaks have been reported in the United Kingdom and the United States. This results mainly from direct (person-to-person) faecal–oral transmission, although the infection may be introduced in the first instance through zoonotic contact. Affected adults may acquire infection from young children in the home or occupationally. Infection may be transmitted sexually where this involves faecal exposure. Cryptosporidium is a cause of travellers' diarrhoea although apparently not as frequently as is the case with giardia.

Waterborne infection

In the United Kingdom, the United States, and elsewhere, there have been numerous well-documented outbreaks resulting from contamination of public drinking-water supplies. Some have been associated with human-specific isolates and thus are likely to have been the result of contamination of the supply by human sewage. Other outbreaks have been associated with the zoonotic type, while isolates from endemic (sporadic) cases, some of which will be waterborne, fall into both categories. Oocysts have been demonstrated widely in both raw and treated water and legislation has been introduced in the United Kingdom in an attempt to limit the latter.

Cryptosporidiosis associated with swimming pools has been reported from several countries, including the United Kingdom, America, and Australia, resulting from accidental faecal contamination from bathers, and may also be acquired from recreational use of natural waters.

Foodborne infection

Foods associated with infection include unpasteurized milk, sausage meat, and salad. Isolation of oocysts from such foods is generally extremely difficult. In the United States infection has resulted from the consumption of fresh-pressed apple juice.

Nosocomial infection

Transmission has been reported between health-care staff and patients and between patients, particularly the immunocompromised. Large numbers of oocysts may be present in patients' stools and in vomit; transmission via fomites occurs although this route of transmission is limited by the susceptibility of oocysts to desiccation. Poor hand-washing practice has been identified as an important factor. In an outbreak with high mortality in a ward of immunocompromised patients in Denmark, transmission was probably by patients' hands via a ward ice-making machine.

Demography

Age and sex distribution

In the United Kingdom approximately two-thirds of *Cryptosporidium*-positive samples are from children 1 to 10 years of age, with a secondary peak in adults under 45 years; the infection is uncommon in infants less than 1 year old and in the elderly. Distribution appears to be the same in both sexes. A relative increase in adult cases is often seen in waterborne outbreaks. In developing countries, infection is common in infants less than 1 year old and asymptomatic infection is common in older subjects.

Temporal distribution

In the United Kingdom there are peaks in the spring and in the autumn, which do not necessarily both occur in any one locality, nor recur year by year, which coincide generally with lambing and calving. The zoonotic genotype is more prevalent in humans in the spring, some of which may result from secondary spread. Similar seasonal peaks are seen in patients with AIDS. The human-specific type shows some increase in frequency later in the year and this may be associated with foreign travel.

Frequency of occurrence

Laboratory rates of detection in non-immunocompromised subjects average about 2 per cent (range: less than 1 to 5 per cent) in developed countries and about 8 per cent in developing countries (range: 2 to 30 per cent), about fourth in the list of pathogens detected in stools submitted to the laboratory. In the United Kingdom about 5000 to 6000 confirmed cases are reported annually, somewhat less frequent generally than giardiasis. Among young children in the United Kingdom, cryptosporidiosis is more common than salmonella infection and during peak periods detection rates may exceed 20 per cent.

Cryptosporidiosis is one of the most common causes of diarrhoea in patients with AIDS and in some studies prevalence has exceeded 50 per cent. The infection rate in patients with AIDS in the United Kingdom has been falling in recent years, which has been attributed to infection control advice and the use of multiple antiretroviral therapy. Infection rates are not generally increased for most other immunocompromised groups.

Clinical aspects

Pathology

Histopathology

There is mucosal involvement of the small bowel, other parts of the gastrointestinal tract, and sometimes beyond. Moderate to severe abnormalities of villous architecture occur, with stunting and fusion of villi and lengthening of crypts. There may be evidence of mild inflammation, with some cellular infiltration into the lamina propria.

The endogenous stages of the parasite in the luminal surface are generally inconspicuous and appear as small (2 to 8 µm) bodies, apparently superficially attached to the brush border, unevenly distributed over the apical cells and within the crypts of the villi (Fig. 1 and Fig. 2). Peaking and apoptosis of infected cells have been reported. There is usually little intracellular change at the ultrastructural level beyond the attachment zone of the parasite. Rectal biopsy may reveal mild, non-specific proctitis. Extensive and chronic involvement of the bile duct and gallbladder is seen in some patients with AIDS.

Immunological response

The particular immunodeficient conditions in which cryptosporidiosis has been reported to show increased severity or persistence suggest that both humoral and cellular factors have a role in limiting infection. An immune response has been demonstrated in the main immunoglobulin classes, although the initial IgG response may be poor. Serological diagnostic tests are, however, of little clinical value. Seroprevalence studies indicate that the infection is common, even in developed countries, and this may reflect water supply quality or other exposures.

Reports differ on the effect of breast feeding on incidence in infancy; some studies suggest a protective effect although protection from the environment by breast feeding may also be important.

Although functioning humoral and cellular immunity seem to be important in limiting or controlling infection, it currently appears that, in animal models, CD4+ and CD8+ T lymphocytes and interferon-g are especially important in this respect. In humans, CD4 cell counts of fewer than 200 cells/mm ^{3} probably indicate the need to take special care to avoid exposure to *Cryptosporidium*, and fewer than 100 cells/mm^{3} indicates a poor prognosis if infection occurs.

Possible pathogenic mechanisms

The watery diarrhoea is characteristic of non-inflammatory infection of the small bowel, especially that associated with toxin-producing organisms and enteric viruses. Several mechanisms have been suggested to explain the symptoms: reduction in absorptive capacity, particularly for water and electrolytes; increase in secretory capacity from crypt hypertrophy; osmotic effects from loss of brush-border enzymes (e.g. disaccharidases) resulting in malabsorption of sugars, increased osmolality of chyme, and subsequent microbial fermentation of sugars in the colon (which may account for the characteristic offensive smell); toxic activity has been described.

Clinical presentation in otherwise healthy (immunocompetent) people

Cryptosporidiosis in the immunocompetent person is a self-limiting, acute gastroenteritis with a variety of presenting symptoms. In cases where the time of exposure has been known the incubation period was about 5 to 7 days (range probably 2 to 14 days; wider limits have been suggested but are unlikely). There may be a prodrome of one to a few days, with malaise, abdominal pain, nausea, and loss of appetite. Gastrointestinal symptoms start suddenly, the stools being described as watery, greenish with mucus in some cases, without blood or pus, and very offensive. Patients may open their bowels more than 20 times a day but more usually 3 to 6 times. Other symptoms include colicky, abdominal pain, especially after meals, anorexia, nausea and vomiting, abdominal distension, and marked weight loss. 'Flu-like' systemic effects, including malaise, headache, myalgias, and fever, commonly occur. Gastrointestinal symptoms usually last about 7 to 14 days, but weakness, lethargy, mild abdominal pain, and intermittent loose bowels sometimes persist for up to a further month.

There is no evidence of transplacental transmission but infection during late pregnancy may cause metabolic disturbances in the mother, leading to the infant's failure to thrive. Failure to thrive has also been observed in older infants and children, and may be associated with persistent infection and enteropathy, especially in underdeveloped countries.

Reported sequelae include pancreatitis (associated with severe abdominal pain), toxic megacolon, and reactive arthritis. In immunocompetent patients, deaths are rarely attributable to cryptosporidiosis.

Clinical presentation in immunocompromised patients

Susceptibility to cryptosporidiosis and the severity of the disease is increased in patients who are immunocompromised as a result of AIDS, hypo- or agammaglobulinaemia, severe combined immunodeficiency, leukaemia, malignant disease, and bullous pemphigoid. Disease susceptibility and severity are also increased during immunosuppressive treatment with cyclophosphamide and corticosteroids as in patients undergoing bone marrow transplantation, and in children immunosuppressed by measles and chickenpox, especially where there is associated malnutrition. Infection in patients with leukaemia may be unusually severe and has sometimes proved fatal, particularly when associated with aplastic crisis, and may then require modification of chemotherapy to control the infection.

Symptoms of cryptosporidiosis are generally similar but often develop insidiously in immunocompromised patients. In those with late-stage AIDS with very low CD4 cell counts, or in some other profound deficiency states, diarrhoea may be frequent, profuse, and watery, like cholera. Patients may open their bowels frequently, passing up to 20 litres of infected fluid stool per day; persistent nausea and vomiting is usually associated with severe diarrhoea and suggests a poor prognosis. Associated symptoms include colicky, abdominal pain often associated with meals, severe weight loss, weakness, malaise, anorexia, and low-grade fever. Cryptosporidial infection in immunocompromised patients may involve the pharynx, oesophagus, stomach, duodenum, jejunum, ileum, appendix, colon, rectum, gallbladder, bile duct, pancreatic duct, and the bronchial tree. Cryptosporidial cholecystitis (presenting with severe right upper-quadrant abdominal pain), sclerosing cholangitis, pancreatitis, hepatitis, and respiratory-tract symptoms may occur, with or without diarrhoea. The clinical picture may include other features of HIV infection and there is often coinfection with other pathogens such as cytomegalovirus, *Pneumocystis carinii*, and *Toxoplasma*.

Patients with less severe impairment of immunity may experience resolution or a more chronic course, with less profuse diarrhoea, sometimes with remission and then recurrence, possibly associated with biliary tract involvement. Except in those patients whose immune suppression can be relieved by stopping immunosuppressant drugs, or, in the case of HIV, intensifying antiretroviral therapy, severe symptoms may persist until the patient dies. This is either as a result of dehydration, acid–base or electrolyte disturbances, and cachexia, from some other opportunistic infection or malignant disease, or a combination of these.

Laboratory investigations

In early acute cases the stools are usually watery, greenish with mucus in some cases, without blood or pus.

Peripheral leucocytosis and eosinophilia are found rarely. Serum electrolyte abnormalities will develop in patients who become severely dehydrated. In immunocompromised patients with cryptosporidial cholecystitis, serum alkaline phosphatase and g-glutamyl transpeptidase levels are raised, while aminotransferases and bilirubin levels may remain normal.

In patients with AIDS, common associated infections are with cytomegalovirus and *Isospora belli*. Mixed infection with *Campylobacter* and *Giardia* species may be found in immunocompetent patients.

In the bowel mucosa there is histological evidence of enterocyte damage, villous blunting, and inflammatory-cell infiltration of the lamina propria; cell peaking and apoptosis have been reported. Histopathological appearances of the affected biliary tract resembles primary sclerosing cholangitis. Radiographic abnormalities include dilatation of the small bowel, mucosal thickening, prominent mucosal folds, and abnormal motility, and in the biliary system, dilated distal biliary ducts, stenosis with an irregular lumen, and other changes reminiscent of primary sclerosing cholangitis.

Differential diagnosis

The absence of blood, pus, cells, or Charcot–Leyden crystals may distinguish cryptosporidiosis from some acute bacterial diarrhoeas and that associated with amoebiasis and isosporiasis. In immunocompetent patients, the symptoms of cryptosporidiosis resemble those of giardiasis or cyclosporiasis. Intense abdominal pain and cramps are generally more common in cryptosporidiosis, but bloating and weakness less common. In immunocompromised patients, especially in those with AIDS, isosporiasis is clinically indistinguishable, but can be diagnosed by finding the organisms in the stool, when Charcot–Leyden crystals may also be found. This infection responds to treatment with trimethoprim and sulphamethoxazole, as does cyclosporiasis.

Treatment of cryptosporidiosis

In immunocompetent patients, the illness is self-limiting, but they may become dehydrated and require intravenous fluids, electrolytes, and symptomatic treatment for their vomiting and diarrhoea.

Immunocompromised patients with persistent severe diarrhoea, malabsorption, and other complications may require prolonged palliative treatment. They should avoid excess milk, as lactose intolerance may develop. Parenteral feeding and fluid, electrolyte, and nutrient replacement may be needed. Antiperistaltic agents such as loperamide, diphenoxylate, or opiates may increase abdominal pain and bloating. Antiemetics may be needed for symptomatic relief. Temporary relief of biliary obstruction has been achieved by endoscopic papillotomy and of cholecystitis by cholecystectomy. Diarrhoea and vomiting may, however, prove intractable.

Some reports suggest possible activity with letrazuril/diclazuril, somatostatin, azidothymidine, diloxanide furoate, furazolidone, amprolium, the macrolides, roxithromycin, and nitazoxanide. Paromomomycin has also been suggested as an active agent although a very recent report indicates that it is no more effective than placebo for cryptosporidiosis in patients with advanced HIV infection. Zydovudine (Retrovir™) therapy may result in remission or amelioration of symptoms, as may treatment of coinfecting agents. Separating the effect of the drugs on copathogens or of fluctuations in immune competence, both spontaneous and drug-induced, may be difficult. Immunotherapy (e.g. with bovine colostrum, hyperimmune immunoglobulin, transfer factor, and interleukin 2) has been attempted, with variable results.

Laboratory detection and diagnosis

The characteristic endogenous stages (Fig. 1 and Fig. 2) may be found in histological sections, using light and electron microscopy, but diagnosis is usually by detection of oocysts in stools. (Plate 1, Plate 2, Plate 3, Plate 4, Plate 5, Plate 6, Plate 7, Plate 8, Plate 9 and Plate 10) Oocysts have also been found in vomit and sputum in some cases, especially those associated with AIDS. The oocysts of *C. parvum* are spherical or slightly ovoid, about 4 to 6 µm, and appear refractile in wet faecal preparations with a highly refractile inner body, the cytoplasmic residuum; the four sporozoites within may be distinguished with difficulty using special optical systems. Several conventional stains have been adapted for diagnostic purposes, such as the modified Ziehl–Neelsen method and phenol–auramine fluorescent stain. Immunofluorescent antibody and enzyme immunoassay methods, using monoclonal antibodies, are commercially available but are expensive. Standardization of approach to screening and of reporting is essential for epidemiological purposes. Ideally, all stool samples from cases of diarrhoea should be screened; restriction, where unavoidable, should be based on age group (see demography) and not on factors such as stool consistency. Concentration of stool specimens is not usually required for diagnosis in acute cases.

Fungal spores, yeasts, cysts of *Balantidium*, sporocysts of *Isospora*, and oocysts of *Clyclspora* may readily be mistaken for cryptosporidial oocysts.

Infectivity, resistance, and control

Infectivity

In studies using monkeys and lambs, the infective dose for *C. parvum* was fewer than 10 oocysts. Human volunteer studies in the United States, initially with the zoonotic genotype, suggest the minimum infective dose varies from fewer than 10 oocysts to more than 1000, varying with the isolate. Symptomatic reinfection was achieved in some subjects despite the presence of antibody. Studies with the human-specific genotype are now in progress.

Resistance and disinfection

Oocysts can survive for many months in a cool, moist environment but are highly susceptible to desiccation, prolonged freezing, and moderate heat (pasteurization temperatures). They are remarkably resistant to most disinfectants and antiseptics, including chlorine at concentrations far greater than those used in water treatment and even to glutaraldehyde under normal use conditions. Some disinfectants may be more effective if used at elevated temperature (37°C or higher). Oocysts are sensitive to 10 volume (3 per cent) hydrogen peroxide, and to appropriate levels of ozone and medium or high-pressure ultraviolet. The adequate disinfection of instruments such as endoscopes is difficult and prolonged immersion in disinfectant, preferably at elevated temperature and after thorough cleaning, is recommended. Recent studies suggest that a high concentration (200 p.p.m.) of chlorine dioxide is effective.

Control of transmission

Primary control is by limiting the opportunity for faecal–oral transmission, both direct and indirect. Symptom-free subjects not in contact with immunocompromised patients can normally be permitted to work if their hygiene is scrupulous. Spread via fomites is possible but this route is limited by the susceptibility of oocysts to desiccation. Patients with AIDS may be more susceptible to infection with uncommon species or genotypes including those normally associated with cats, dogs, and birds and advice may be needed to limit exposure.

Contamination of water supplies is inevitable from time to time, even in developed countries, and may be the source of some sporadic cases as well as outbreaks. When a public advisory notice is issued to boil water, raising the water just to boiling point is sufficient. In general, bottled water and water from point-of-use filters are unlikely to contain parasites but may carry an increased bacterial load, the health significance of which is uncertain for the immunocompromised. Patients with AIDS and others who are profoundly compromised should be advised never to drink water that has not been boiled or filtered through a suitable device. Users of filters should remember that these devices may concentrate potential pathogens and care is needed in replacing and disposing of filter elements.

Hospitals involved in the care of profoundly immunocompromised patients should be particularly vigilant in the management of patients with cryptosporidiosis. Long-term arrangements should be made for the provision of safe water for the immunocompromised to avoid difficulties when a notice to boil water is issued.

Further reading

Casemore DP (1991). Broadsheet No 128: The laboratory diagnosis of human cryptosporidiosis. *Journal of Clinical Pathology* **44**, 445–51.

Colford JM *et al*. (1996). Cryptosporidiosis among patients infected with the human immunodeficiency virus. *American Journal of Epidemiology* **144**, 903–9.

Coop RL, Wright SE, Casemore DP (1998). Cryptosporidiosis. In: Palmer SR, Soulsby Lord, Simpson DIH, eds. *Zoonoses—biology, clinical practice, and public health contro*l, pp 563–78. Oxford University Press, Oxford.

Current WL (1998). Cryptosporidiosis. In: Cox FEG, Kreier KP, Waklin D, eds. *Topley and Wilson's microbiology and microbial disease*, 9th edn, Vol 5, *Parasitology*, pp 329–47. Edward Arnold, London.

Fayer R, ed. (1997). *Cryptosporidium* and cryptosporidiosis. *CRC Press*, Boca Raton, FA.

Gasser RB, O'Donoghue P, eds (1999). Isolation, propagation and characterisation of *Cryptosporidium*. Invited review. *International Journal for Parasitology* **29**, 1379–413.

Meinhardt PL, Casemore DP, Miller KB (1996). Epidemiologic aspects of human cryptosporidiosis and the role of waterborne transmission. *Epidemiologic Reviews* **18**, 118–36.

7.13.6 Cyclospora

D. P. Casemore

Introduction

Cyclospora are coccidian or coccidian-like enteric parasites producing large, acid-fast oocysts (spore-like forms) initially recognized in faecal specimens examined by modified Ziehl–Neelsen stain for the detection of cryptosporidium. Isolates were first described variously as cryptosporidium-like bodies, fungal spores, and cyanobacteria (blue–green algae-like) associated with diarrhoeal illness. They have since been definitively identified as oocysts of a newly recognized protozoan parasite. They have been detected worldwide, most often in residents of, or travellers returning from, developing countries, and there is some evidence for foodborne and waterborne transmission.

Natural history

Cyclospora are apicomplexan protozoans which are widespread in nature but not previously described in man. Morphologically they resemble coccidia, but recent molecular evidence suggests that they may be more closely related to the *Eimeria*. They have a monoxenous life cycle similar to the cryptosporidia, with both sexual and asexual stages developing in the same host animal, resulting in the production of oocysts, the environmentally hardy transmissible stage. The species found in man, tentatively named *Cyclospora cayetanensis* new species, has an oocyst stage of about 8 to 10 µm in size which, when first excreted, unlike cryptosporidium, is unsporulated, with a characteristic morular inner structure (see below). During the extrinsic sporulation period, about 7 to 15 days depending on temperature, the oocysts develop two inner membrane-bound sporocysts, each containing two large sporozoites (1.2 × 9.0 µm). This species has not been found in other hosts, and experimental transmission to a variety of potential host species has been unsuccessful. It may thus, like many other species of coccidia, be restricted to a single host species. Oocysts of a morphologically similar cyclospora have been found in some primates but their precise identity is uncertain. Undefined cyclospora-like like bodies have been reported from dogs and some birds, although human isolates cannot be transmitted to these host species.

Epidemiology

Infection occurs in people of all ages; reported most commonly in young children from developing countries and among travellers to Nepal, Indonesia, Southern and Central America, and other underdeveloped areas, most of whom are adult. Sporadic cases have been identified in patients in the United States and the United Kingdom without a history of foreign travel. In recent years outbreaks have been identified in the United States associated with food consumption, particularly with raspberries imported from Guatemala, but also with some other fresh produce including mesclun (mixed lettuce leaves) and the herb basil. The precise mechanism of contamination is unclear, although contaminated water seems the most likely vehicle. The requirement for an extrinsic period of sporulation implies that transmission is likely to be indirect.

Clinical presentation

Enteric symptoms include watery diarrhoea, flatulence, bloating, dyspepsia, abdominal cramps, nausea, and vomiting: generalized symptoms include marked weight loss, malaise, and influenza-like symptoms. The infection tends to be protracted, lasting 14 days or more (range 1 to more than 60 days). Asymptomatic infection occurs in indigenous people in developing countries, probably reflecting endemicity and recurrent infection in the immune.

Diagnosis (Plate 1, Plate 2 and Plate 3)

Moderate numbers of oocysts are excreted in stools during the acute stage and variably thereafter; they can be detected by modified Ziehl–Neelsen staining, although the acid-fast staining is variable. The oocysts are 8 to 10 µm in size and have visible surface and internal structure. Phase contrast microscopy reveals the internal morula, a collection of refractile, membrane-bound spherical bodies, 1 to 2 µm in size, within an outer wall; fluorescence microscopy shows characteristic blue autofluorescence of oocysts. Stools stored at room temperature in 2.5 per cent potassium dichromate sporulate in about 7 to 15 days to show the two sporocysts; the sporozoites within cannot readily be seen.

The site of infection is primarily the small intestine. Endogenous stages may be detected intracellularly beneath the brush border of enterocytes in jejunal biopsy specimens, and possibly other tissues, by light and electron microscopy (Fig. 1). Histology shows altered mucosal architecture with shortening and widening of intestinal villi, diffuse oedema, mixed inflammatory cellular infiltrate, reactive hyperaemia with vascular dilatation, and capillary congestion. The parasite is found within a parasitophorous vacuole, midway between the nucleus and the cell membrane at the luminal side. Transmission electron microscopy reveals typical apicomplexan structures.

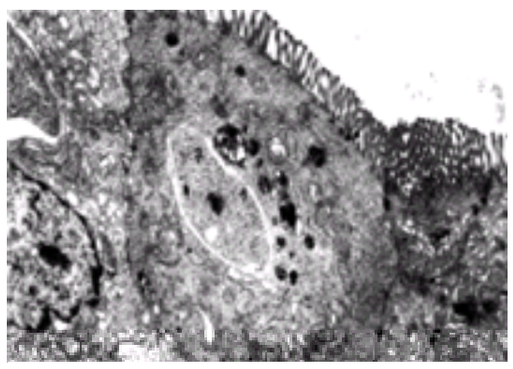

Fig. 1 Longitudinal section through jejunal biopsy specimen showing a single intracellular parasite by transmission electron microscopy. (© The Lancet Ltd and reproduced with permission from Bendall RP *et al. Lancet*, 1993; **341**: 590–2.)

Treatment

Cotrimoxazole (one tablet twice a day for 7 days) has proved to be effective in eradicating the infection.

Control

As the source of the parasite is currently unknown, specific recommendations to limit the reservoir cannot be made. Transmission is almost certainly primarily by an indirect faecal–oral route and hence can be limited by the usual hygienic precautions including avoidance of unboiled water, water-washed unpeeled fruit, salads,

uncooked vegetables, etc., in endemic areas. The parasite is difficult to remove from the surface interstices of fruit such as raspberries.

Further reading

Connor BA, Reidy J, Soave R (1999). Cyclosporiasis: clinical and histopathologic correlates. *Clinical Infectious Diseases* **28**, 11216–22.

Eberhard ML, Pieniazek NJ, Arrowood MJ (1997). Laboratory diagnosis of cyclospora infections. *Archives of Pathology and Laboratory Medicine* **121**, 792–7.

Herwaldt BL (2000). *Cyclospora cayetanensis*: a review focusing on the outbreaks of cyclosporiasis in the 1990s. *Clinical Infectious Diseases* **31**, 1040–57.

Sterling R, Ortega YR (1999). Cyclospora: an enigma worth unravelling. *Emerging Infectious Diseases* **5**, 48–53.

7.13.7 Sarcocystosis

V. Zaman

Humans can act as both the final and intermediate host of parasites belonging to the genus *Sarcocystis*. In their lifecycle there is an alternation of a sexual generation of the parasite in the intestinal tissues of a predator host (carnivores including snakes, omnivores, and scavenger animals) and an asexual generation in the tissues of a prey animal (herbivores and omnivores including rodents). The predator animals act as the final hosts and excrete oocysts or sporocysts in the faeces. The animal eaten by a predator acts as an intermediate host because cysts are present in the muscles and other tissues (Fig. 1).

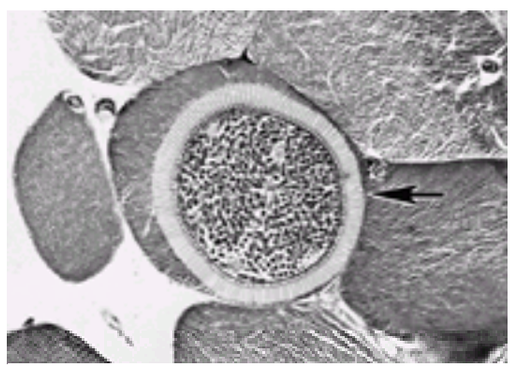

Fig. 1 Sarcocyst in muscle: the thickness of the cyst wall varies in different species; in this species a thick, striated wall is visible and the elongated structures inside the cyst are cystozoites (× 400).

Sarcocystis hominis (syn. *Isospora hominis*)

The intermediate host is cattle. Human infection results from eating uncooked beef. Prevalence in human populations is not known but the lifecycle has been studied in human volunteers.

Clinical aspects

Most patients who pass oocysts are asymptomatic and the development of the sporogonic stage in the human intestine is either non-pathogenic or only slightly pathogenic, resulting in mild gastrointestinal upset. However, the symptoms may vary, depending on the number of parasites ingested. Severe symptoms may occur after ingestion of heavily infected beef. This probably happened in six patients from Bangkok who developed symptoms suggestive of segmental necrotizing enteritis.

Diagnosis

This is based on the detection of oocysts or sporocysts in the faeces of infected individuals (Fig. 2). Sporocysts range in size from 13.6 to 16.4 µm by 8.3 to 10.6 µm. Occasionally, sporocysts may be seen attached in pairs and covered by a thin, transparent cyst wall (Fig. 3).

Fig. 2 *Sarcocystis hominis*: sporocyst with sporozoites; the residium (food store) can be seen at one end (× 1000).

Fig. 3 *Sarcocystis hominis*: sporocysts attached in a pair (× 1000).

Treatment

No chemotherapeutic agents are available. Prevention consists of not eating uncooked beef.

Sarcocystis suihominis

The lifecycle is similar to that of *S. hominis*, except that the intermediate host is the pig.

Clinical aspects

Human volunteers given infected tissues have experienced diarrhoea and mild fever.

As in the case of *S. hominis*, the intensity of symptoms probably varies with the size of the infective dose. If large amounts of heavily infected pork are ingested, symptoms could be quite severe. As this rarely happens, symptoms in most patients are mild or absent.

Diagnosis

This is based on the detection of oocysts or sporocysts in faeces; these are almost identical to those of *S. hominis*.

Treatment

No chemotherapeutic agents are available. Prevention consists of not eating raw pork.

Sarcocystis spp.

These produce sarcocystis in human muscles. There is probably more than one species involved. Infection is acquired by the ingestion of oocysts or sporocysts passed in the final hosts. The final hosts are unknown but could be carnivores, such as dogs or cats.

Clinical aspects

Most cases are asymptomatic. The infection is an incidental finding in muscle biopsies for other diseases or at autopsy. It appears that the cysts of some species are found only in skeletal muscles while others occur in cardiac and skeletal muscles. On the basis of morphology it is possible to differentiate the cysts into four types.

Diagnosis

In tissue sections, *Sarcocystis* can be diagnosed easily and it is generally not difficult to differentiate it from *Toxoplasma* tissue cysts. *Sarcocystis* has a distinct cyst wall and the cystozoites are larger. *Toxoplasma* cystozoites are positive to periodic acid-Schiff reagent, while *Sarcocystis* cystozoites are negative.

Treatment

None is available.

Further reading

Beaver PC, Gadgil RK, Morera P (1979). *Sarcocystis* in man: a review and report of five cases. *American Journal of Tropical Medicine and Hygiene* **28**, 819–44.

Bunyaratvej S, Bunyawongwiroj P, Nitiyanant P (1982). Human intestinal sarcosporidiosis: report of six cases. *American Journal of Tropical Medicine and Hygiene* **31**, 36–41.

Dubey JP, Speer CA, Fayer R (1989). *Sarcocystosis of animals and man*. CRC Press, Boca Raton, FL.

7.13.8 Giardiasis, balantidiasis, isosporiasis, and microsporidiosis

Martin F. Heyworth

Giardiasis

Aetiology

Giardia intestinalis (synonyms *Giardia lamblia* and *G. duodenalis*) colonizes the lumen of the small intestine. The parasite's lifecycle comprises two stages: motile trophozoites (Fig. 1) and thick-walled ellipsoidal cysts that are excreted in the faeces. *G. intestinalis* trophozoites are dorsoventrally flattened organisms with eight flagella, two nuclei, and a ventral adhesive disc that enables them to become attached to the luminal surface of intestinal epithelial cells. Trophozoites absorb nutrients in the small intestinal lumen and multiply in this anaerobic environment. New hosts become infected by ingesting *G. intestinalis* cysts; exposure of cysts to gastric acid leads to emergence of trophozoites from the cysts. Trophozoites encyst in the intestinal lumen, and the resulting cysts are excreted from the host.

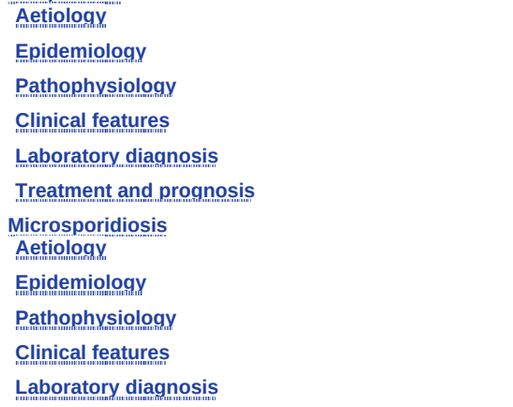

Fig. 1 Scanning electron micrograph of three *Giardia intestinalis* trophozoites on a jejunal biopsy specimen from a patient with giardiasis. The dorsal surfaces of two trophozoites are visible (D), and the ventral adhesive disc of the other trophozoite is shown (V). (Illustration by courtesy of Dr Robert L. Owen; modified from Carlson JR, Heyworth MF, Owen RL (1984). Giardiasis: Immunology, diagnosis and treatment. *Survey of Digestive Diseases* **2,** 210–23, S. Karger AG, Basel. Used by permission.)

Epidemiology

G. intestinalis infection is usually acquired by drinking water that contains cysts. Other modes of spread include direct faecal–oral transmission of cysts, as in day-care centres for small children, and occasional foodborne transmission of cysts. Waterborne giardiasis occurs as a result of drinking unfiltered, unboiled water from streams and lakes containing *G. intestinalis* cysts. Swimming in (and inadvertently drinking) such water is also a risk factor for giardiasis. Outbreaks of this infection have resulted from the unintended presence of *G. intestinalis* cysts in public water supplies.

Worldwide, many species of domestic, farm, and wild animal are hosts for *G. intestinalis*. Giardia cysts have been found in faecal specimens from cattle, sheep, horses, pigs, dogs, and cats. To what extent non-human mammals are sources of human giardiasis is, however, an unanswered question, which may be resolved by genotyping of giardia organisms isolated from various hosts.

Giardiasis occurs in temperate and tropical countries. Several genetically distinct, genetically stable, strains of *G. intestinalis* are known. Accordingly, some authors regard this species as a 'species complex', rather than a single species.

Immunodeficiency predisposes to the occurrence of severe and persistent giardiasis. Human immunodeficiency states that are associated with giardiasis include conditions that impair host antibody responses (notably, 'common variable' hypogammaglobulinaemia and X-linked immunoglobulin deficiency). Impairment of intestinal IgA production is a feature of these particular immunodeficiency diseases and may explain how they predispose to chronic giardiasis (via impaired production of antitrophozoite IgA). Some patients with common variable hypogammaglobulinaemia and chronic giardiasis have abnormally enlarged lymphoid follicles in the small intestine (nodular lymphoid hyperplasia), which contain numerous immature B lymphocytes that express IgM. These B lymphocytes appear to be developmentally arrested, such that they do not mature (as would normally be the case) into IgA-expressing B lymphocytes and IgA-secreting intestinal plasma cells.

Pathogenesis and pathophysiology

The mechanism(s) responsible for diarrhoea and malabsorption in giardiasis are not understood. In one study, the histological appearance of duodenal biopsies was reportedly normal (apart from the presence of trophozoites) in 96.3 per cent of patients with giardiasis (462/480 total patients); the other 3.7 per cent (18/480) had 'duodenitis' with 'mild villus shortening'. Shortening of microvilli on the luminal surface of intestinal epithelial cells has been observed in small intestinal biopsies from patients with giardiasis. Reduced activity of intestinal disaccharidases has been reported in giardia-infected human subjects and rodents. This functional enzyme deficiency could, conceivably, lead to osmotic diarrhoea (via the presence of undigested disaccharides in the intestinal lumen).

G. intestinalis trophozoites cultured in the presence of sodium glycocholate take up this bile salt from the culture medium. Uptake of bile salts by trophozoites in the intestinal lumen might, therefore, contribute to the fat malabsorption that occurs in some patients with giardiasis (by reducing the availability of bile salts for fat emulsification). In a study of neonatal rats infected with *G. intestinalis*, different strains of the parasite induced different degrees of small intestinal damage (as judged by alteration in villus length and in electrolyte absorption). This work raises the theoretical possibility that differences in severity of symptoms, among different patients with giardiasis, might reflect infection by distinct strains of *G. intestinalis* with different pathogenicities.

Study of immunity against giardia species has been more feasible in rodents than in human subjects. In mice, clearance of *Giardia muris* infection appears to be dependent on CD4+ (helper) T lymphocytes, and to follow the generation of an intestinal IgA response against the parasite. In human volunteers who were deliberately infected with *G. intestinalis*, a corresponding intestinal IgA response occurred. IgA directed against trophozoites binds to these organisms and may, conceivably, inhibit their attachment to the intestinal epithelium, such that they are susceptible to peristaltic expulsion from the host.

Clinical features

G. intestinalis infection can be asymptomatic (as shown by cyst excretion in the absence of symptoms), and can also cause various clinical problems. These include abdominal discomfort, tenderness, and distension, a sensation of 'fullness', nausea, anorexia, and watery diarrhoea. Other clinical features include 'heartburn', flatulence, steatorrhoea, and weight loss. In immunologically normal persons, untreated giardiasis typically lasts for several weeks, with symptoms that fluctuate in severity. Clinical sequelae that have occasionally been reported include megaloblastic anaemia resulting from impaired absorption of vitamin B_{12} or folic acid.

Laboratory diagnosis

In a patient suspected of having parasitic infection of the gastrointestinal tract (with one or more species of parasite that might include *G. intestinalis*), faecal light microscopy may be informative. If the patient has giardiasis, *G. intestinalis* cysts may be seen during this examination. Diagnostic sensitivity can be increased by immunofluorescence microscopy of faecal specimens incubated with a fluorescent monoclonal antibody directed against *G. intestinalis* cysts. If there is a strong suspicion of infection with *G. intestinalis*, in the absence of other species of gastrointestinal parasite (or if the aim is to check the effectiveness of treatment in clearing known giardiasis), immunoassay for *G. intestinalis* antigen(s) is the test of choice. This approach, which involves enzyme-linked immunoassay (**EIA**) of faecal specimen(s) with one of several commercially-available kits, is more objective and less labour intensive than immunofluorescence microscopy (which detects whole cysts). Various EIA kits for diagnosis of giardiasis have sensitivities in the range 88 to 100 per cent and specificities in the range 99 to 100 per cent. Commercially available EIA kits detect *G. intestinalis* cyst wall protein(s) in faecal specimens.

Immunocompetent persons with giardiasis develop serum antibodies against *G. intestinalis* trophozoites. Testing of human sera for such antibody is not useful for diagnosing current giardiasis in individual subjects, but can be informative in population studies examining the prevalence of this infection.

Treatment

Table 1 summarizes various drug regimens for treating giardiasis. Metronidazole resistance of *G. intestinalis* is an increasingly recognized problem.

Prevention

G. intestinalis cysts can be removed from water by filtration, for example using membrane filters with a pore diameter of less than 5 mm. Cysts in water are killed by boiling. Water intended for human consumption can be screened for *G. intestinalis* cysts by filtration, followed by immunofluorescence microscopy of particulate material retrieved on the filter(s), using a fluorescent anticyst monoclonal antibody. Experimental protocols for detecting giardia cysts in water, by amplification of giardia DNA using polymerase chain reaction (**PCR**), have also been described.

Controversies and future research

Efforts to confirm, or refute, the idea that intestinal antibody protects against giardia infections are justified. Studies with mice, kittens, and puppies have suggested that vaccination against giardia infections is feasible. Whether a vaccine against human giardiasis would have much practical utility, however, is an open question, even if it is biologically feasible to develop one. Interest has been expressed in sequencing the entire *G. intestinalis* genome.

Balantidiasis

Aetiology

This infection is caused by *Balantidium coli*, a ciliate protozoan that is the largest protozoan parasite of man. *B. coli* has a two-stage lifecycle comprising motile trophozoites that invade the colonic mucosa (Fig. 2) and non-motile cysts. Spread of the infection to new hosts occurs by ingestion of the parasite.

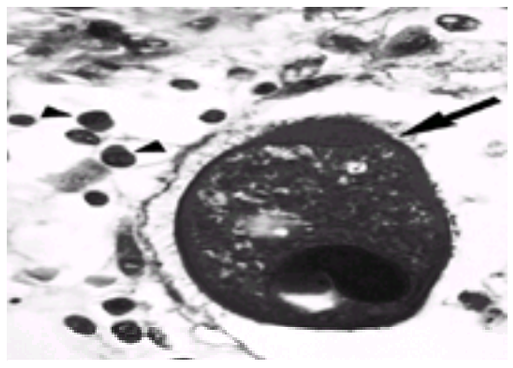

Fig. 2 Light micrograph of *Balantidium coli* trophozoite (arrow) in colonic tissue. Cilia are visible on the surface of the organism. Arrowheads indicate tissue plasma cells (× 705). (Modified from Neafie RC (1976). Balantidiasis. In: Binford CH, Connor DH, eds. *Pathology of tropical and extraordinary diseases*, Vol 1, pp 325–7. Armed Forces Institute of Pathology, Washington DC. Used by permission.)

Epidemiology

Balantidiasis occurs in temperate and tropical countries. There is circumstantial evidence that man can acquire the infection from animals. *B. coli* infection has been described in pigs and in many species of non-human primate. A high prevalence of the infection has been seen in human communities that live in close proximity to *B. coli*-infected pigs (for example, in New Guinea). Consequently, there has been speculation that pigs are a reservoir for spread of *B. coli* to man. However, balantidiasis has also occurred in human subjects who had no known contact with pigs or other animals. Clusters of cases of balantidiasis have been seen in

long-stay psychiatric hospitals. In India, *B. coli* cysts have been found in water available for either drinking or use in cooking.

Pathophysiology

B. coli trophozoites are invasive organisms that cause ulceration of the colonic mucosa. The mechanism(s) responsible for tissue invasion by these organisms are not known.

Clinical features

Human subjects with *B. coli* infection can be asymptomatic, or can develop diarrhoea with stools that are either watery or that consist of blood and mucus. In severe *B. coli* infection, patients can develop colonic perforation, peritonitis, gangrene of the appendix (resulting from the presence of *B. coli* in the appendiceal wall), and spread of the parasite to the liver or lungs. Balantidiasis is a rare cause of liver abscess. As is evident from the clinical features outlined above, balantidiasis may be clinically indistinguishable from amoebiasis, bacillary dysentery, and ulcerative colitis, and can be fatal.

Laboratory diagnosis

Balantidiasis can be diagnosed by microscopic examination of diarrhoeal stools, or of colonic mucus obtained at sigmoidoscopy. Examination may show motile trophozoites or, less frequently, cysts of *B. coli*. Histological examination of rectal biopsies may reveal *B. coli* trophozoites.

Treatment and prevention

Patients with balantidiasis have been treated empirically with various antimicrobial drugs. There is, however, little interpretable information about the effectiveness of such treatment, although eradication of *B. coli* has been reported in some individuals treated with metronidazole or tetracycline. Surgical intervention may be necessary in patients with liver abscess or clinical evidence of appendicitis or colonic perforation.

Prevention of balantidiasis involves avoidance of *B. coli* cyst ingestion (via filtration or boiling of drinking water, hand washing before handling food, and careful cleaning and cooking of food).

Isosporiasis

Aetiology

Isospora belli, the cause of isosporiasis, is a coccidian parasite of the human small intestine. Coccidia of the genus *Isospora* infect many species of vertebrate, and are relatively or absolutely host-specific. There is no evidence that, under natural conditions, *I. belli* infects any vertebrate species other than man, although this coccidian has been transmitted experimentally to gibbons.

I. belli oocysts are ellipsoidal structures that are excreted in the faeces of infected individuals (Fig. 3). Studies of isospora species that parasitize non-human hosts indicate that infection occurs via ingestion of oocysts, and that sporozoites (which emerge from oocysts) penetrate epithelial cells of the small intestine. Subsequent development of isospora species comprises: (i) an asexual pathway, with production of merozoites, which can infect additional epithelial cells; and (ii) a sexual pathway, in which fusion of gametes produces oocysts that are excreted from the host.

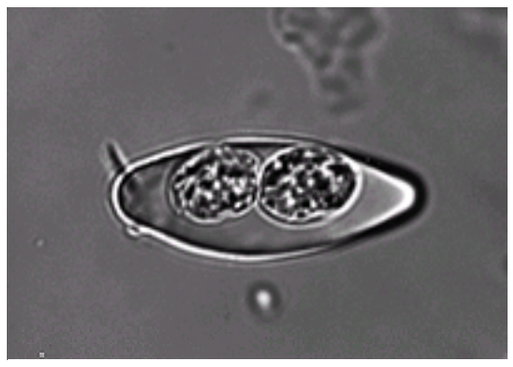

Fig. 3 Light micrograph of an *Isospora belli* oocyst (× 2500). (Illustration by courtesy of Dr William L. Current. From Garcia LS (2001). *Diagnostic medical parasitology*, 4th edn. ASM Press, Washington DC. Used by permission.)

Epidemiology

I. belli infection has been documented in immunosuppressed and, rarely, in immunocompetent individuals. Reported prevalence rates of this infection in patients with acquired immunodeficiency syndrome (**AIDS**) have been 1 per cent in Los Angeles, 8 per cent in Zambia, and 15 per cent in Haiti. Vehicle(s) for transmission of *I. belli* oocysts to human subjects have not been identified, but presumably include water and/or food.

Pathophysiology

Mechanism(s) responsible for the watery diarrhoea that occurs in isosporiasis are unknown. Presumably, the parasitization of epithelial cells in the small intestine contributes to the diarrhoea.

Clinical features

In patients infected with human immunodeficiency virus (**HIV**), *I. belli* infection is associated with chronic watery diarrhoea, abdominal cramps, nausea, fever, and weight loss. Severe dehydration can result from diarrhoea attributable to *I. belli* infection in HIV-infected patients. Reports of *I. belli* infection in immunocompetent persons are uncommon. In such individuals, however, symptoms ascribed to isosporiasis are similar to those that occur in AIDS-associated *I. belli* infection.

Rarely, extraintestinal *I. belli* infection has been described in patients with AIDS; in the relevant patients, tissues parasitized by *I. belli* have included gallbladder epithelium, liver, spleen, and mesenteric lymph nodes.

Laboratory diagnosis

Isosporiasis can be diagnosed by microscopic examination of faecal samples for *I. belli* oocysts. Although these structures are relatively large (approximately 20 to 30 mm in length), they are translucent and may be difficult to see in unstained samples. Their visibility is increased by incubation with carbol fuchsin, which stains oocyst internal structures red. An alternative approach is to examine faecal smears under ultraviolet light; with this type of illumination, *I. belli* oocysts show blue autofluorescence.

Treatment and prognosis

Because isosporiasis is diagnosed infrequently, most of the literature dealing with its treatment consists of anecdotal case reports. In the 1980s, oral

trimethoprim–sulphamethoxazole was found to be an effective drug combination for treating *I. belli*-induced diarrhoea, in a study of patients with AIDS and isosporiasis in Haiti. Recognition of adverse drug reactions to trimethoprim–sulphamethoxazole, and less than 100 per cent efficacy of this drug combination in treating isosporiasis, have prompted alternative therapeutic approaches. Diclazuril, albendazole–ornidazole, and pyrimethamine–sulphadiazine are three such alternatives that have shown anecdotal promise in treating isosporiasis associated with HIV infection.

In immunocompetent patients without HIV infection, isosporiasis can persist for weeks or months if untreated. The overall prognosis in patients with isosporiasis and HIV infection is determined by the HIV infection.

Microsporidiosis

Aetiology

Microsporidia are protozoa with features that are sufficiently distinctive for the organisms to be classified as a separate phylum (Microspora). They are obligate intracellular parasites of hosts that include insects, fish, and mammals. The lifecycle of microsporidia comprises an extracellular stage (spore) and stages that occur in the cytoplasm of host cells. Spores (Fig. 4) are shed into the environment by infected hosts, and infect other members of the host species. The spores induce infection by extruding a hollow tube, which penetrates a host cell and forms a channel for delivering sporoplasm (spore contents) into this cell. Replication of the parasite and subsequent production of spores occur in host cells.

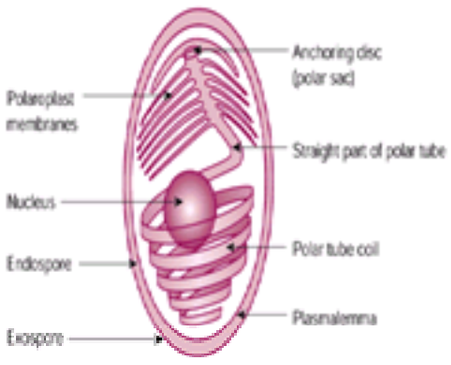

Fig. 4 Diagram of a microsporidian spore, showing internal structure. (Illustration by courtesy of Professor Elizabeth U. Canning. Modified from Canning EU and Hollister WS (1992). Human infections with microsporidia. *Reviews in Medical Microbiology* **3**, 35–42. Used by permission.)

Microsporidia that infect man are listed in Table 2. Several species listed in this table were unknown in any context before the 1990s. When it has been sought, *Enterocytozoon bieneusi* has been found in up to 30 per cent of patients with both diarrhoea and HIV infection. Authenticated human infections with microsporidia other than *E. bieneusi*, *Encephalitozoon intestinalis*, and *E. hellem*, are rare. Some of the microsporidian species listed in Table 2 have been found in one or two patients only. 'Microsporidium' (Table 2) is a non-taxonomic genus created for microsporidia of unclear identity.

Epidemiology

Before 1985, when *E. bieneusi* infection was first described, very few cases of human microsporidiosis had been reported. From the mid-1980s onwards, most of the documented clinical experience with microsporidiosis has occurred in patients with HIV infection. After its initial description, as an intestinal parasite in the HIV-infected population, *E. bieneusi* was reported in several HIV-negative, purportedly immunocompetent persons with diarrhoea. Similarly, human encephalitozoon infections have been reported most frequently in HIV-infected patients, but also occur in immunocompetent individuals. A serological survey revealed anti-encephalitozoon antibodies in sera from 8 per cent of 300 presumably healthy Dutch blood donors, and from 5 per cent of 276 pregnant French women.

Experimental work with animals suggests that human infection with some species of microsporidia occurs via ingestion of spores. Environmental sources of microsporidian spores that can infect human subjects appear to include water and, possibly, non-human hosts. In France, *E. bieneusi* DNA has been found in river water (by filtration and PCR amplification). Using a similar approach, DNA of *E. bieneusi* and of *Encephalitozoon intestinalis* has been detected in water in Arizona. Risk factors for *E. bieneusi* infection, in a population of HIV-infected patients surveyed in France, included swimming in a pool in the 12 months before the survey. In rural Mexican households, faecal excretion of encephalitozoon spores was associated with the use of unboiled water for drinking and for preparing food. Collectively, these observations suggest that *E. bieneusi* and encephalitozoon infections can be waterborne. Heavy parasitization of respiratory tract epithelial cells with *Encephalitozoon hellem*, in at least one HIV-infected patient examined at autopsy, raises the possibility that some microsporidian infections can be acquired by inhaling spores.

Some species of microsporidia listed in Table 2 are known to infect non-human hosts. For example, *E. hellem* infection has been described in budgerigars and parrots. Spores of *E. intestinalis* have been identified in faecal specimens from non-human mammals (dog, pig, goat, cow, and donkey), by immunofluorescence microscopy and PCR. *Enterocytozoon bieneusi* can infect pigs, and *Encephalitozoon cuniculi* (a rarely documented cause of microsporidiosis in HIV-infected patients) infects various hosts, including rabbits, dogs, and pigs. Genetically distinct strains of *E. bieneusi* and of *E. cuniculi* have been described; there is some evidence that the different strains show host selectivity.

Pathophysiology

In HIV-infected patients, diarrhoea is the clinical feature that has been most frequently associated with microsporidiosis. In particular, this symptom has been linked to infection with *Enterocytozoon bieneusi* and with *Encephalitozoon intestinalis*. The diarrhoea in these microsporidian infections presumably results from the presence of microsporidia in the small intestinal mucosa.

In mice at least, interferon-g contributes to protective immunity against *E. intestinalis* and *E. cuniculi* infections.

Clinical features

Clinical features of microsporidian infections reflect the anatomical site colonized by the microsporidia (Table 2). Besides watery diarrhoea, weight loss and fat malabsorption have been reported in HIV-infected patients with intestinal microsporidiosis. Microsporidian infection of the gallbladder has been described in occasional HIV-infected patients, who had acalculous cholecystitis (characterized by right upper abdominal pain, nausea, and vomiting), and who were treated by cholecystectomy. Symptoms of sinusitis, and cough and dyspnoea, have been reported in patients with microsporidian infection of the paranasal sinuses and respiratory tract. Symptomatic urethritis has been ascribed to microsporidian infection in occasional HIV-infected patients.

Microsporidian infection of the conjunctiva and corneal epithelium causes symptoms of keratoconjunctivitis (foreign body sensation in the eye, ocular discomfort and redness, photophobia, blurred vision, and sometimes reduced visual acuity). Microsporidian infections of the corneal stroma lead to reduced visual acuity, with or without corneal ulceration. Clinical features in patients with actual or presumed cerebral microsporidiosis have included headache, cognitive impairment, nausea, vomiting, and epileptic seizures. Symptoms of myositis (muscle pain, tenderness, weakness, and wasting) have been described in patients with microsporidian infection of skeletal muscles.

Laboratory diagnosis

Human microsporidian infections were originally documented by microscopic examination of tissue sections obtained by biopsy or at autopsy. The small size of microsporidia favoured the use of electron microscopy (rather than light microscopy) for diagnosing microsporidiosis in early studies. Because this is a

labour-intensive approach requiring equipment that may not be readily accessible, simpler diagnostic methods than electron microscopy were sought.

One such approach (which is non-invasive) involves examining faecal samples for microsporidian spores, which are ovoid in shape. The spores can be detected by using a number of stains, including crystal violet plus iodine and chromotrope 2R (leading to violet staining of the spores), optical brighteners such as Uvitex 2B and Calcofluor White M2R (which bind to chitin in the spores), and fluorescent antibodies directed against the spores. Spores can be seen by light microscopy (after staining with crystal violet/iodine/chromotrope 2R), or by fluorescence microscopy after incubation with optical brighteners (which lead to fluorescence of the spores) or fluorescent antibodies. Intestinal infection with *E. bieneusi* or *E. intestinalis* can be diagnosed by finding microsporidian spores in faecal samples. Spores of *E. bieneusi* are smaller (about 1.5 µm × 0.9 µm) than those of *E. intestinalis* (about 2.5 µm × 1.5 µm). In addition, microsporidian infection of the nasal mucosa and paranasal sinuses can be diagnosed by microscopic examination of nasal secretions for spores (after staining, as outlined above). Similarly, microsporidian spores can be found in urine and bile from patients with urinary tract and biliary tract microsporidiosis, respectively.

Approaches to diagnosis of microsporidian keratoconjunctivitis include examining conjunctival/corneal scrapings or biopsies for spores and (non-invasively) *in vivo* examination of the cornea with a scanning confocal microscope to look for spore-filled epithelial cells.

The commonest clinical situation that calls for efforts to diagnose microsporidiosis is the HIV-infected patient with diarrhoea. To supplement (or, perhaps, eventually replace) diagnostic methods that require microscopy, many authors have developed molecular methods for diagnosing microsporidiosis. Such approaches usually involve DNA extraction (for example, from faecal samples) and PCR amplification using primers specific for regions of gene(s) that encode(s) RNA in the small subunit of microsporidian ribosomes (SSU-rRNA). The DNA obtained by PCR amplification is detected by agarose gel electrophoresis.

Treatment and prognosis

HIV-infected patients with *E. bieneusi* infection and chronic diarrhoea have been treated with various drug regimens in an attempt to clear the *E. bieneusi* infection. To date, the most effective regimen for this purpose has been 'highly-active anti-retroviral therapy' (**HAART**), which involves simultaneous treatment with several drugs directed against HIV, including HIV protease inhibitor(s). When effective in HIV-positive, *E. bieneusi*-infected patients, HAART leads to reduction of HIV load, elevation of the circulating CD4+ T-lymphocyte count, clearance of *E. bieneusi* infection, and cessation of diarrhoea. Uncontrolled trials and anecdotal reports suggest that thalidomide, furazolidone, fumagillin, and atovaquone are potentially useful drugs for treating *E. bieneusi* infection.

Encephalitozoon infections can be treated effectively with albendazole. In a small controlled trial, HIV-infected patients with *E. intestinalis* infection were treated with either albendazole, 400 mg twice daily by mouth, or placebo. Albendazole treatment led to clearance of gastrointestinal *E. intestinalis* infection in this study. Uncontrolled trials and anecdotal case reports describe partial or complete resolution of symptoms (diarrhoea, sinusitis, and keratoconjunctivitis) in patients with *E. intestinalis*, *E. hellem*, or *E. cuniculi* infection following albendazole treatment. Pregnancy is a contraindication to albendazole treatment.

Microsporidial keratoconjunctivitis has been treated successfully with fumagillin eye drops in HIV-infected patients. HIV-negative patients with microsporidian infection of the corneal stroma have been treated by corneal transplantation, with results that have ranged from failure (opacification of the transplant) to apparent success, as judged by transparency of the graft 6 months after transplantation.

Individual patients infected with *Trachipleistophora hominis* or *Brachiola vesicularum* reportedly showed some clinical improvement after treatment with albendazole–sulphadiazine–pyrimethamine, or albendazole–itraconazole, respectively.

In HIV-infected patients with microsporidiosis, the overall prognosis is determined by the HIV infection.

Future research

Further efforts are warranted to identify environmental sources of microsporidia that infect human subjects. One unanswered question is the extent to which domestic water supplies contain viable spores of pathogenic microsporidia. Further work is also warranted on the prevalence of microsporidian infections in immunocompetent persons (including the extent to which these infections cause symptoms in the immunocompetent human population).

Further reading

Anonymous (1998). Drugs for parasitic infections. *The Medical Letter on Drugs and Therapeutics* **40**, 1–12. [Survey of drug treatment for parasitic diseases (including giardiasis, balantidiasis, isosporiasis, and microsporidiosis).]

Croft SL, Williams J, McGowan I (1997). Intestinal microsporidiosis. *Seminars in Gastrointestinal Disease* **8**, 45–55. [Review article focusing on *Enterocytozoon bieneusi* and *Encephalitozoon intestinalis* infections.]

Faubert G (2000). Immune response to *Giardia duodenalis*. *Clinical Microbiology Reviews* **13**, 35–54. [Review article that discusses host immune responses against giardia organisms and utility of immunoassays for diagnosing giardiasis.]

Franzen C, Müller A (1999). Molecular techniques for detection, species differentiation, and phylogenetic analysis of microsporidia. *Clinical Microbiology Reviews* **12**, 243–85. [Comprehensive review of human microsporidian infections, with particular reference to molecular biology of microsporidia.]

Garcia LS (1999). Flagellates and ciliates. *Clinics in Laboratory Medicine* **19**, 621–38. [Review of human giardiasis and balantidiasis.]

Heyworth MF (1996). *Giardia* infections. In: Paradise LJ, Bendinelli M, Friedman H, eds. *Enteric infections and immunity*, pp 227–38. Plenum Press, New York. [Brief survey of immunological and clinical aspects of giardia infections].

Lindsay DS, Dubey JP, Blagburn BL (1997). Biology of *Isospora* spp. from humans, nonhuman primates, and domestic animals. *Clinical Microbiology Reviews* **10**, 19–34. [Review of the genus *Isospora*, including a discussion of human infection with *Isospora belli*.]

Marshall MM *et al*. (1997). Waterborne protozoan pathogens. *Clinical Microbiology Reviews* **10**, 67–85. [Review of pathogenic protozoa that are known, or presumed, to be transmitted to previously uninfected hosts via water.]

Weiss LM, Keohane EM (1997). The uncommon gastrointestinal protozoa: Microsporidia, Blastocystis, Isospora, Dientamoeba, and Balantidium. *Current Clinical Topics in Infectious Diseases* **17**, 147–87. [Review of several protozoan species that infect the gastrointestinal tract.]

Weiss LM, Vossbrinck CR (1998). Microsporidiosis: molecular and diagnostic aspects. *Advances in Parasitology* **40**, 351–95. [Review article that discusses molecular biology of microsporidia and molecular approaches to the diagnosis of microsporidian infections.]

7.13.9*Blastocystis hominis* infection

R. Knight

This is an anaerobic protist of the caecum and colon, its taxonomic affinity has long been uncertain. Sequencing of ribosomal RNA genes indicates that it is a stramenopile, a group that includes certain unicellular, mostly flagellated, algae. The form commonly described in faeces, and also in cultures, is spherical, 5 to 8 µm in diameter, with a prominent central body surrounded by peripheral cytoplasm containing granules; electron microscopy reveals a nucleus, with a crescentic cap of heterochromatin, and mitochondria with tubular cristae (Fig. 1, Fig. 2, and Fig. 3). The organism grows readily in cultures with mixed bacteria but axenic cultures can be established; division is by binary fission and also probably by schizogony within the central body. Electron microscopy of faeces reveals a multivacuolar form with multiple small vesicles that can either form cysts or transform into the familiar form. Colonoscopy specimens have shown an amoeboid form that ingests bacteria with pseudopodia. Bizarre environmentally induced forms with huge vacuoles may develop in cultures. The common 'univacuolar form' was named *Blastocystis* by Brumpt in 1912, who considered it to be a yeast, although it was first described by Alexieff in 1911 as a protozoan cyst.

Fig. 1 *B. hominis* from culture showing binary fission; the cytoplasm is lying at the periphery. v, vacuole. Phase contrast, × 400.

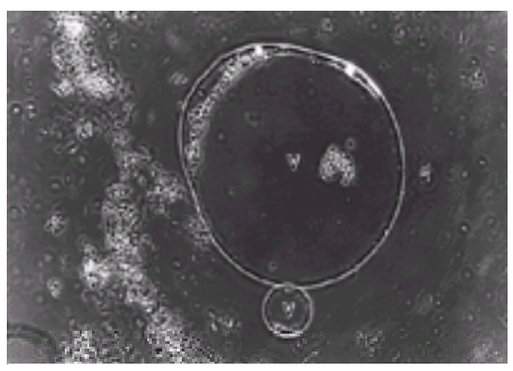

Fig. 2 *B. hominis* from culture showing the great variation in size. v, vacuole. Dark field, × 400.

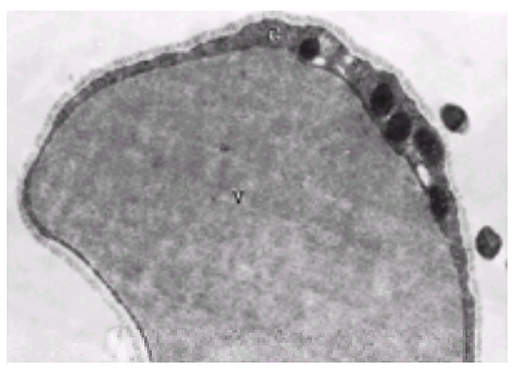

Fig. 3 *B. hominis.* Electron micrograph showing the peripheral cytoplasm (c) and the central vacuole (v); the inclusions in the cytoplasm are mitochondria. × 5000.

Epidemiology

Prevalence is high in many human populations in contexts of high faeco-oral transmission. This infection is associated with travel and institutions and may occur in outbreaks. Similar *Blastocystis* organisms of uncertain pathogenicity occur in birds, pigs, and monkeys. The recently recognized cysts are the transmissive stage.

Diagnosis

It is usually recognized in direct wet faecal smears or formol ether concentrates; wet mounts can be stained with iodine giving a brownish central body; or with toluidine blue. The organism is often numerous in symptomatic subjects. Permanent mounts stain well with trichrome. *Blastocystis* can resemble amoebic cysts but lacks their characteristic nuclei. In fixed smears stained specifically for *Cryptosporidium* there is no oocyst wall.

Clinical features and treatment

A diarrhoeal illness lasting 3 to 10 days is attributed to this organism. Sometimes symptoms continue for weeks or months. Associated features are abdominal bloating, flatulence, and anorexia. Symptoms are more prolonged in immunocompromised subjects. There is no association with irritable bowel syndrome. Illnesses are self-limiting in most people but infection can be eliminated with metronidazole or tinidazole; the organism is also sensitive to furazolidine and co-trimoxazole.

Evidence for pathogenicity

Serum antibody is detectable in symptomatic subjects; preliminary studies suggest *in vitro* cytotoxicity to tumour cell monolayers, and local lesions have been

produced in mice after intramuscular injection.

The situation with *Blastocystis* in humans may be similar to that of several anaerobic lumen-dwelling protozoa infecting vertebrates in which a self-limited non-invasive pathogenicity is followed by a longer carrier state. Such a relationship remains very difficult to prove or disprove, especially as there is genetic heterogeneity between *Blastocystis* isolates. Clinical response to metronidazole is hardly compelling evidence for pathogenicity since concurrent infection with other enteropathogens is common and this drug has a wide spectrum of activity, including an effect upon small bowel bacterial overgrowth. More well-documented outbreaks and cytopathic evidence are needed.

Further reading

Boreham PF, Stenzel DJ (1993). *Blastocystis* in humans and animals: morphology, biology, and epizootiology. *Advances in Parasitology* **32**, 1–70.

Moe KT *et al*. (1998). Cytopathic effect of *Blastocystis hominis* after intramuscular inoculation into laboratory mice. *Parasitology Research* **84**, 450–4.

Stenzel DJ, Boreham PF (1996). *Blastocystis hominis* revisited. *Clinical Microbiology Reviews* **9**, 563–84.

7.13.10 Human african trypanosomiasis

August Stich

Introduction

Sleeping sickness or human African trypanosomiasis (HAT) is caused by subspecies of the protozoan haemoflagellate *Trypanosoma brucei* transmitted to man and animals by tsetse flies (*Glossina* spp.). The distribution of the vector restricts sleeping sickness to the African continent between 14° north and 29° south (Fig. 1). Human disease occurs in two clinically and epidemiologically distinct forms, *gambiense* or West African and *rhodesiense* or East African sleeping sickness (Table 1). A third subspecies of the parasite, *T.b. brucei*, causes disease in cattle but is non-pathogenic in humans.

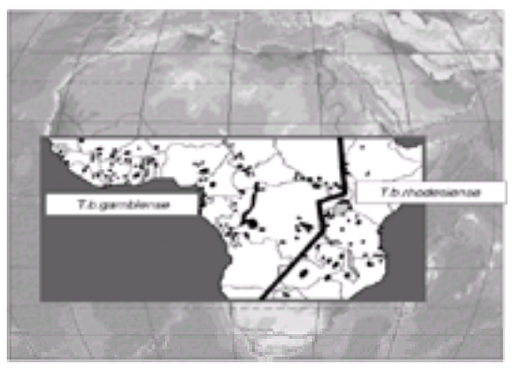

Fig. 1 The geographical distribution of human African trypanosomiasis.

First reports of the disease go back to the fourteenth century. In the past, the impact on health in Africa has been enormous. Many areas had long been rendered uninhabitable for people and livestock. During the first decades of the twentieth century, millions may have died in Central Africa around Lake Victoria and in the Congo basin (Fig. 2). The success of control programmes in the 1960s promised the disappearance of sleeping sickness as a public health problem. However, recent epidemics in the Democratic Republic of Congo, Northern Angola, Sudan, Uganda, and other countries confirm a major resurgence of HAT. According to current estimates by WHO, the achievements in sleeping sickness control during colonial times will be completely reversed in the near future.

Fig. 2 Sleeping sickness patients on an island in Lake Victoria; historical photograph taken during Robert Koch's research expedition to East Africa.

Today, 60 million people in some 40 African countries are exposed to the risk of HAT. Half a million are believed to be infected (almost all with *T.b. gambiense*). They are doomed if left untreated. For tourists and expatriates, sleeping sickness has always been a rare disease, but a recent cluster of cases in tourists to Tanzania re-emphasizes that it is also important in travel medicine.

Aetiology

In 1895, Sir David Bruce (1855–1931) suggested an association between trypanosomes and 'cattle fly fever', a major problem for livestock in southern Africa. In 1902, Robert M. Forde and Everett Dutton identified trypanosomes in the blood of a patient during a research expedition in The Gambia, and in 1903, Aldo Castellani isolated trypanosomes from the cerebrospinal fluid. In the same year, tsetse flies were identified as the vector.

Trypanosoma brucei (phylum Sacromastigophora, order Kinetoplastida) is an extracellular protozoal parasite. Like *Leishmania*, it possesses a centrally placed nucleus and a kinetoplast, a distinct organelle with extranuclear DNA. The kinetoplast is the insertion site of an undulating membrane, which extends over nearly the whole cell length and ends as a free flagellum.

The three subspecies of *Trypanosoma brucei* are indistinguishable morphologically. However, they differ considerably in their interaction with their mammalian host and the epidemiological pattern of the diseases they cause. Formerly, *T.b. gambiense* and *T.b. rhodesiense* isolates were characterized either by isoenzyme analysis or by animal inoculation. The advent of molecular techniques created expectations of more reliable tools for their differentiation. However, genomic characterization has revealed several more subdivisions rather than the three expected. Whereas West African isolates proved relatively homogeneous, East African isolates from humans and animals did not simply conform to what is still called *T.b. rhodesiense* and *T.b. brucei* but showed a complex relationship with evidence of sexual genetic exchange in the vector. Further molecular research may soon lead to a comprehensive phylogenetic tree and a deeper insight into trypanosomal evolution and biology.

Transmission

Although congenital, blood-borne, and mechanical transmission have been reported and may play an occasional role, the main mode of transmission is through the bite of infected tsetse flies (*Glossina* spp., order Diptera; Fig. 3 and Plate 1). These are biologically unique insects, which occur only in Africa in 31 distinct species and subspecies. Less than half are potential vectors of HAT. Their distinctive behaviour, ecology, and chosen habitat explain many epidemiological features of sleeping sickness. Tsetse flies can live for many months in the wild, but give birth to only about eight to 10 larvae per lifetime. Both sexes feed on blood. They are fastidious in requiring warm temperatures, shade, and humidity for resting and larviposition and so their distribution is highly localized. Recently, the mapping and monitoring of possible HAT transmission foci has become possible with the use of satellite imaging techniques.

Fig. 3 Adult tsetse fly (*Glossina morsitans*). (See also Plate 1.)

During the blood meal on an infected mammalian host, the tsetse fly takes up trypanosomes ('short-stumpy form') into its mid-gut, where they develop into procyclic forms and multiply. After about 2 weeks, they migrate to the salivary glands as epimastigotes where they finally develop into infective metacyclic forms. With the next blood meal, they are then injected into the new vertebrate host where they appear and multiply as 'long-slender' trypomastigotes.

Molecular and immunological aspects

The cyclic changes of the trypanosome into different developmental stages are accompanied by variations in morphology, metabolism, and antigenicity. Several unique metabolic pathways have been described in trypanosomes, distinct from their host and thus qualifying as potential drug targets.

The blood stream forms of *T. brucei* are covered with a dense coat of identical glycoproteins, numbering up to about 500 aminoacids per molecule. Being highly immunogenic, they stimulate the production of specific antibodies, mainly of the IgM subclass. Once the surface glycoproteins have been recognized by host antibodies, the parasitic cell will be attacked and destroyed through complement activation and cytokine release.

However, about 2 per cent of *T. brucei* in each new generation change the expression of their specific surface glycoprotein. The 'coat' will then be different in the new clone ('variant' surface glycoprotein: VSG). This phenotypic switch is done mainly by programmed DNA-rearrangements, moving a transcriptionally silent VSG gene into an active, telomerically located expression site. Within a trypanosome population, the potential repertoire of such different VSG copies seems to be virtually infinite.

Every new VSG copy is antigenically different, thus stimulating the production of a new IgM population. This antigenic variation is the major immune evasion strategy of the parasite, enabling the trypanosome to persist in its vertebrate host. It also reduces parasite load and prolongs the infection. But the inevitable outcome is immune exhaustion of the host, penetration of trypanosomes into immune-privileged sites such as the central nervous system, and finally death.

Clinical features

Sleeping sickness is a dreadful disease, causing great suffering to patients, their families, and the affected community. The infection often has an insidious onset, but *T. brucei*, whether the East or West African subspecies, will invariably kill if the patient is not treated in time. The natural course of HAT can be divided into different and distinct stages. Their recognition and differentiation is important for the management of the patient.

The trypanosomal chancre

Tsetse bites can be quite painful, usually leaving a small and self-healing mark. In the case of a trypanosomal infection, the local reaction can be quite pronounced and longer lasting. A small raised papule will develop after about 5 days. It increases rapidly in size, surrounded by an intense erythematous tissue reaction (Fig. 4 and Plate 2) with local oedema and regional lymphadenopathy. Although some chancres have a very angry appearance, they are not usually very painful unless they become ulcerated and superinfected. They heal without treatment after 2 to 4 weeks, leaving a permanent, hyperpigmented spot.

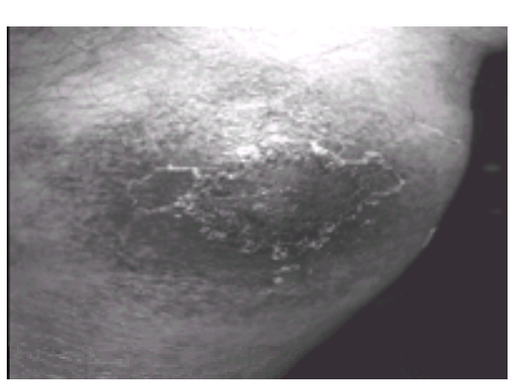

Fig. 4 Trypanosomal chancre on the shank of a missionary returning from the Congo. (See also Plate 2.)

Trypanosomal chancres occur in about half the cases of *T.b. rhodesiense*. In *T.b. gambiense*, they are much less common. They often go undetected in endemic populations.

Haemolymphatic stage (HAT stage I)

After local multiplication at the site of inoculation, the trypanosomes invade the haemolymphatic system, where they can be detected after 7 to 10 days. During this period of spread, they are exposed to vigorous defence mechanisms of the host, which they evade by antigenic variation. This continuous battle between antigenic switches and humoral defence results in a fluctuating parasitaemia with parasites frequently becoming undetectable, especially in *gambiense* HAT. The cyclic release of cytokines during periods of increased cell lysis results in intermittent, non-specific symptoms: fever, chills, rigors, headache, and joint pains. These can easily be misdiagnosed as malaria, viral infection, typhoid fever, or many other conditions. Hepatosplenomegaly and generalized lymphadenopathy are common, indicating activation and hyperplasia of the reticuloendothelial system.

A reliable sign, particularly in *T.b. gambiense* infection, is the enlargement of lymph nodes in the posterior triangle of the neck (Winterbottom's sign). Other typical signs are a fugitive patchy rash, a myxoedematous infiltration of connective tissue ('puffy face syndrome'), and an inconspicuous periostitis of the tibia with delayed hyperaesthesia (Kérandel's sign).

In *T.b. rhodesiense* infection, this haemolymphatic stage is very pronounced with severe symptoms, sometimes even resulting in early death through cardiac involvement (myocarditis). In the early stage of *T.b. gambiense* infection, symptoms are usually infrequent and mild. Febrile episodes become less severe as the disease progresses.

Meningoencephalitic stage (HAT stage II)

Within weeks in *T.b. rhodesiense* and months in *T.b. gambiense* infection, cerebral involvement will invariably follow; trypanosomes cross the blood–brain barrier. In children, HAT progresses even more rapidly towards this meningoencephalitic stage.

The onset of stage II is insidious. The exact time of central nervous system involvement cannot be determined clinically. Histologically, perivascular infiltration of inflammatory cells ('cuffing') and glial proliferation can be detected, resembling endarteritis. As the disease progresses, patients complain of increasing headache, and their families may detect a marked change in behaviour and personality. Neurological symptoms, which follow gradually, can be focal or generalized, depending on the site of cellular damage in the central nervous system. Convulsions are common, usually indicating a poor prognosis. Periods of confusion and agitation slowly evolve towards a stage of distinct perplexity when patients lose interest in their surroundings and their own situation. Sleep abnormalities result finally in a somnolent and comatose state. Progressive wasting and dehydration follows the inability to eat and drink.

There is no unique, clinical sign of late HAT, opening up a wide range of possible neurological and psychiatric differential diagnoses. However, the appearance of the patient, with apathy and the typical expressionless face, is a very characteristic sight in endemic areas (Fig. 5 and Plate 3).

Fig. 5 Patient with late-stage trypanosomiasis. (See also Plate 3.)

Diagnosis

HAT can never be diagnosed with certainty on clinical grounds alone. Definitive diagnosis requires the detection of the parasite in chancre aspirate, blood, lymph juice, or cerebrospinal fluid using various parasitological techniques. The methods for diagnosis are essentially the same for *gambiense* and *rhodesiense* HAT (Table 2).

Lymph node aspirate

Lymph node aspiration is widely used, especially for the diagnosis of *gambiense* HAT. Fluid of enlarged lymph nodes, preferably of the posterior triangle of the neck (Winterbottom's sign), is aspirated and examined immediately at 400 × magnification. Mobile trypanosomes can be detected for a few minutes between the numerous lymphocytes.

Wet preparation, thin, and thick blood film

During all stages of the disease, trypanosomes may appear in the blood where they can be detected in unstained wet or in stained preparations. The yield of detection is highest in the thick blood film, a technique widely used for the diagnosis of blood parasites such as *Plasmodia* or microfilaria. Giemsa or Field staining techniques are appropriate (Fig. 6 and Plate 4).

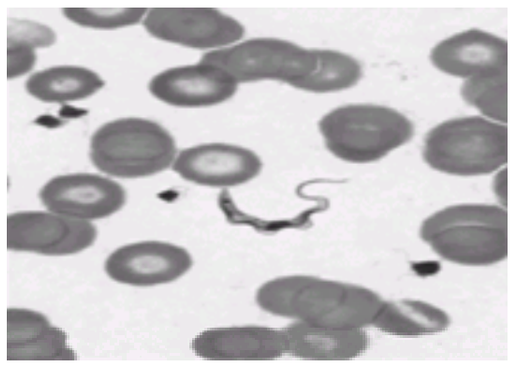

Fig. 6 Trypanosomes in thin human blood film (Giemsa stain, ×1000 magnification). (See also Plate 4.)

Especially in *gambiense* HAT, parasitaemia is usually scanty and fluctuating, often being undetectable. Repeated examinations on successive days are sometimes necessary until trypanosomes can be documented.

Concentration methods

To increase the sensitivity of blood examinations, various concentration assays have been developed. Trypanosomes tend to accumulate in the buffy coat layer after centrifugation of a blood sample. The best results in the field have been obtained with the m-AECT (mini anion exchange column technique), where trypanosomes are concentrated after passage through a cellulose column, and the QBC method (quantitative buffy coat), which was originally developed for the diagnosis of malaria.

Serological assays

Serology is a useful tool to detect antibodies against trypanosomiasis. Various test methods have been described and are now commercially available. They are mainly based on ELISA technique or immunofluorescence, but provide reliable results only in *gambiense* HAT.

For rapid screening under field conditions, the CATT (card agglutination test for trypanosomiasis) is an excellent tool in areas of *T.b. gambiense* infestation. It is easy to perform and delivers results within 5 min. A visible agglutination in the CATT suggests the existence of antibodies, but does not necessarily imply overt disease.

Non-specific laboratory findings

Anaemia and thrombocytopenia are caused by systemic effects of cytokine release, especially of TNF-a. Hypergammaglobinaemia can reach extreme levels as a result of polyclonal activation of immunoglobulins. IgM levels detected in HAT are among the highest observed in any infectious disease.

Diagnosis of stage II

Stage determination is crucial for the correct management of a patient. This cannot be done on clinical grounds alone. Therefore, a lumbar puncture for the examination of the cerebrospinal fluid has to be performed in every patient found positive for trypanosomes in blood or lymph aspirate. In addition, a lumbar puncture should be performed in all patients with the clinical suspicion of HAT even if peripheral examinations had proved negative. A minimum of 5 ml of cerebrospinal fluid is required to examine for:

- Leucocytes—cerebral involvement in HAT stage II is accompanied by pleocytosis, mostly lymphocytes, in the cerebrospinal fluid. By convention a number of more than five cells per mm^3 cerebrospinal fluid defines central nervous system involvement even if the patient does not (yet) have neurological symptoms. Pathognomonic for HAT is the appearance of activated plasma cells with eosinophilic inclusions in the cerebrospinal fluid, the morular cells of Mott (Fig. 7 and Plate 5).
- Trypanosomes—the chances of detecting trypanosomes in the cerebrospinal fluid increase with the level of pleocytosis and the technique used. The highest yield is obtained by cerebrospinal fluid double centrifugation.
- Protein—in patients with HAT, a level of 37 mg of protein per 100 ml cerebrospinal fluid (dye-binding protein assay) is highly suggestive of the advanced stage. Stage II HAT is characterized by an autochthonous production of IgM antibodies in the cerebrospinal fluid, which can be selectively detected if suitable laboratory facilities exist (e.g. latex IgM test).

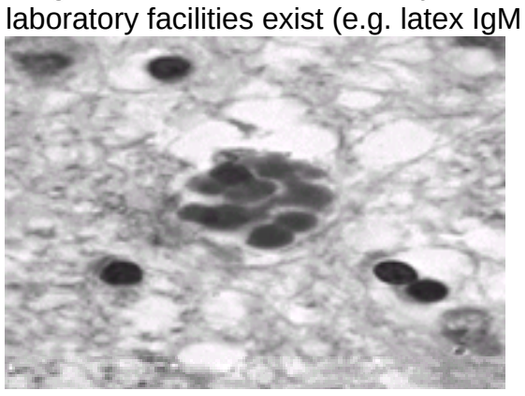

Fig. 7 Morular cell of Mott in a histological brain section of a stage II HAT patient (haematoxylin and eosin stain, ×1000 magnification). (See also Plate 5.)

Treatment

General considerations

HAT is curable, especially if the diagnosis is made in an early stage of the disease and treatment is administered correctly. In the stark reality of the African situation, however, there are many major obstacles to successful patient management:

1. Sleeping sickness is a disease of rural, remote places. The active foci of sleeping sickness are usually in far-away and insecure places, which are difficult to reach. Numerous treatment centres work under emergency conditions with extremely restricted resources. Many affected patients, without proper access to health care, are left unattended.
2. The diagnosis is difficult. Initial diagnosis and exact staging of trypanosomiasis requires sophisticated and dangerous methods, justified only in the hands of experienced personnel. Repetitive training programmes, constant supervision, and continuous quality control are necessary but in reality are rarely available.
3. The treatment of trypanosomiasis is extremely costly. Invariably it exceeds the locally available resources. However, external funding and a sustainable donor commitment for Africa is generally diminishing.
4. The treatment is complicated. Treatment of HAT is dangerous, prolonged, and usually requires hospitalization. Most patients with late-stage trypanosomiasis are severely ill and malnourished. Adverse drug reactions during treatment are difficult to assess due to concomitant pathologies. Their management requires considerable medical skill and good nursing care. Hospitals in rural Africa are often not sufficiently equipped to accomplish good patient care.
5. Many drugs are unavailable. Treatment of HAT is hampered by the limited availability of essential drugs on the international market. Many trypanosomicidal agents are on the verge of disappearance despite increasing demand. Many are no longer produced, as the affected patient populations cannot pay. The range of drugs is diminishing, and hardly any new treatments are in sight. This is especially worrying in view of the reported spread of drug resistance.
6. HAT treatment is not standardized. Trypanosomiasis treatment regimens vary considerably between countries and treatment centres. Results from different centres are comparable to only a very limited extent. Few properly conducted and sufficiently powerful clinical trials are available to evaluate duration, dosage, and possible combinations of drugs. There are few suitable research sites at major trypanosomiasis foci.

Stage I drugs

The treatment of HAT varies according to the trypanosome subspecies and the stage of the disease (Table 3).

Pentamidine

Since its introduction in 1937, pentamidine has become the drug of first choice for *gambiense* HAT stage I, achieving cure rates as high as 98 per cent. However, there are frequent failures in *rhodesiense* HAT. Lower rates of cellular pentamidine uptake in *T.b. rhodesiense* may explain these differences. Some cures of stage II infections have also been reported, but cerebrospinal fluid drug levels are usually not sufficiently high to guarantee a reliable trypanosomicidal effect in the central nervous system.

Pentamidine is usually given by deep intramuscular injection, manageable even in outpatients. If hospital care and reasonable monitoring conditions are available, an intravenous infusion, given in normal saline over 2 h, might be used instead. The main advantage of pentamidine over other drugs is the short treatment course and ease of administration. Adverse effects are related to the route of administration or its dose and are usually reversible (Table 4).

Pentamidine is also used as second-line therapy for visceral leishmaniasis and especially in the prophylaxis and treatment of opportunistic *Pneumocystis carinii*

pneumonia in AIDS. Since the advent of the HIV pandemic, the cost of pentamidine was increased more than tenfold by producers, making it unaffordable by health institutions in low-income countries. After an intervention by WHO, a limited amount of pentamidine is now made available for use in HAT at a subsidised rate.

Suramin

In the early twentieth century, the development of suramin, resulting from German research on the trypanosomicidal activity of various dyes ('Bayer 205'), was a major break-through in the field of tropical medicine. For the first time, African trypanosomiasis, at least in its early stages, became treatable.

Suramin is still used to treat stage I HAT, especially *rhodesiense*. Like pentamidine it does not reach therapeutic levels in cerebrospinal fluid. Suramin is injected intravenously after dilution in distilled water.

Adverse effects depend on nutritional status, concomitant illnesses (especially onchocerciasis) and the patient's clinical condition. Although life-threatening reactions have been described, serious adverse effects are rare and the drug remains one of the safest in trypanosomiasis treatment (Table 4).

Stage II drugs

Melarsoprol

Until the introduction of the arsenical compound melarsoprol in 1949, late stage trypanosomiasis was untreatable. Since then, it has remained the most widely used stage II antitrypanosomal drug both for *T.b. gambiense* and *rhodesiense* infections. It has saved many lives, but has a high rate of dangerous adverse effects. Increasing frequency of relapses and resistance has been reported in some parts of Uganda, Congo, and Angola.

Melarsoprol clears trypanosomes rapidly from the blood, lymph, and cerebrospinal fluid. Its toxicity usually restricts its use to late-stage disease. It is given by slow intravenous injection; extravascular leakage must be avoided.

A new, simpler regimen is based on recently acquired knowledge of the drug's pharmacokinetics (Table 4). The most important adverse effect is an acute encephalopathy, provoked around day 8 of the treatment course in 5 to 14 per cent of all patients. There is severe headache, convulsions, rapid neurological deterioration, or deepening of coma. Characteristically, the comatose patient's eyes remain open. Most probably, this is an immune-mediated reaction precipitated by release of parasite antigens in the first days of treatment. The overall case fatality ranges between 2 and 12 per cent, depending on the stage of disease and the quality of medical and nursing care. Simultaneous administration of glucocorticosteroids (prednisolone 1 mg/kg body weight; maximum 40 mg daily) reduces mortality, especially in cases with high cerebrospinal fluid pleocytosis. However, in areas where tuberculosis, amoebiasis, or strongyloidiasis are highly prevalent, corticosteroids have dangers of their own!

Eflornithine (DFMO)

Initially developed as antitumour agent, eflornithine was introduced in 1980 as an antitrypanosomal drug, in the hope that it might replace melarsoprol for treatment of stage II trypanosomiasis. However, exorbitant costs and limited availability have restricted its use to melarsoprol-refractory cases of *gambiense* sleeping sickness. *T.b. rhodesiense* isolates are normally much less sensitive.

It can be taken orally, but intravenous administration is preferred as it achieves a much higher bioavailability and success rate. Eflornithine should be administered slowly over a period of at least 30 min. Continuous 24-h administration is preferable if facilities allow.

The range of adverse reactions to eflornithine is wide, as with other cytotoxic drugs in cancer treatment. Their occurrence and intensity increase with the duration of treatment and the severity of the patient's general condition. Generally, all adverse effects of eflornithine are reversible (Table 4).

No pharmaceutical company has produced eflornithine for use against HAT since 1999, despite pressure by WHO. The discovery of its therapeutic effect in cosmetic creams against facial hair might help to restimulate production and thus have a beneficial 'spin-off effect' for HAT. In 2001 agreements were signed between WHO and two major drug producing companies which might help to assure a sufficient supply of eflornithine and other drugs essential for the treatment of HAT for the next few years.

Nifurtimox

Ten years after its introduction for the treatment of American trypanosomiasis in 1967, nifurtimox was found to be effective in the treatment of *gambiense* sleeping sickness. It has a place as second line treatment in melarsoprol-refractory cases or in a combination chemotherapy. Experience is limited to few cases treated on compassionate grounds. Prospective clinical trials in HAT are currently in progress.

Nifurtimox is generally not well tolerated, but adverse effects are usually not severe. They are dose-related and rapidly reversible after discontinuation of the drug (Table 4).

Combination treatments in HAT

Melarsoprol, eflornithine, and nifurtimox interfere with trypanothione synthesis and activity at different stages. There is also experimental evidence that combinations of suramin and stage II drugs might be beneficial. Therefore, by reducing the overall dosage of each individual component, drug combinations could perhaps reduce the frequency of serious side-effects, and the development of resistance, which are such common problems in the treatment of sleeping sickness.

Drug combination treatment of HAT is virtually confined to single-case reports. Properly conducted clinical trials are overdue.

Individual protection

HAT among tourists and occasional visitors of endemic areas is a rare event. Pentamidine or suramin chemoprophylaxis is historical, and can no longer be recommended. Long-sleeved, bright clothing and insecticide repellents are the best defence against attacking tsetse flies.

Prevention and control

In the past, tremendous efforts have been undertaken to control sleeping sickness as a threat to human lives and rural development. Control programmes are based on the five complementary pillars given in Table 5.

The most important strategy is active case finding. This requires mobile teams, which regularly visit villages in endemic areas. Mostly based on the results of CATT screening, patients, preferably in the early stage of the disease, are identified and treated. Gradually, the parasite reservoir is depleted. As *Glossina* is a relatively incompetent vector and susceptible to control measures such as insecticide application or trapping, the combination of various approaches can lead to a complete break of the transmission cycle. This was achieved in the past in many places. However, the recent resurgence of sleeping sickness in areas ridden by war and civil unrest, in combination with the decreasing availability of drugs on the international market and the general loss of interest in health in Africa, gives rise to the fear that HAT will soon be again uncontrollable and untreatable (Fig. 8).

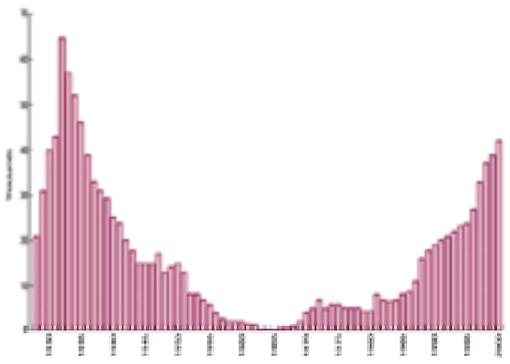

Fig. 8 Number of annually reported cases of HAT (source: WHO Report on Global Surveillance and Epidemic-prone Infectious Diseases); according to WHO, the actual patient numbers are about 10-fold higher.

Trypanosomiasis in the twenty-first century

There is hardly any other tropical disease which demonstrates more clearly the dichotomy characterizing our modern age. On one side, trypanosomes are kept in culture and studied extensively in numerous research laboratories. Their genome is sequenced, and many molecular, biochemical, and immunological phenomena have been discovered as a result of basic science research. General interest in this disease is usually restricted to research aspects only. On the other hand, diagnostic and especially therapeutic tools are increasingly unavailable, because the hundreds of thousands of infected people in Africa are not commercially viable consumers. The prospects for fighting trypanosomiasis look grim. African countries are less and less able to implement effective control programmes, because of political instability or financial incapacity. Global concern about the crisis of human trypanosomiasis in Africa is a question of scientific ethics and international solidarity.

Further reading

Bailey JW, Smith DH (1992). The use of the acridine orange QBC technique in the diagnosis of African trypanosomiasis. *Transactions of the Royal Society of Tropical Medicine Hygiene* **86,** 630.

Burri C, Nkunku S, Merolle A, Smith T, Blum J, Brun R (2000). Efficacy of new, concise schedule for melarsoprol in treatment of sleeping sickness caused by *Trypanosoma brucei gambiense*: a randomised trial. *Lancet* **355**, 1419–25.

Dumas M, Bouteille B, Buguet A, eds. (1999). *Progress in human african trypanosomiasis, sleeping sickness*. Springer-Verlag, France.

Keiser J, Stich A, Burri C (2001). New drugs for the treatment of human African trypanosomiasis: research and development. *Parasitology Today* **17**, 42–9.

Pepin J, Milord F, Guern C, Mpia B, Ethier L, Mansinsa D (1989). Trial of prednisolone for prevention of melarsoprol-induced encephalopathy in gambiense sleeping sickness. *Lancet* **i**, 1246–50.

Smith DH, Pepin J, Stich A (1998). Human African trypanosomiasis: an emerging public health crisis. *British Medical Bulletin* **54**, 341–55.

World Health Organization (1998). *Control and surveillance of african trypanosomiasis*. WHO Technical Report Series 881. WHO, Geneva.

7.13.11 Chagas' disease

M. A. Miles

A poeira de Curvelo
Não faz mal para ninguém não
Do pulmão lá ninguém morre
O que mata é o coração

The dust of Curvelo does not harm anybody
No-one dies there of lung disease
What kills is the heart

[From the poem 'O galo cantou na serra' by Luiz Claudio and Guimarães Rosa]

Introduction and aetiology

The Brazilian scientist, Carlos Chagas, discovered the disease that bears his name, and the entire lifecycle of the causative organism during a few months in 1907. Chagas first found the protozoan agent, *Trypanosoma cruzi*, in the gut of the large blood-sucking insect vector—the triatomine bug (Hemiptera: Reduviidae, subfamily Triatominae) (Fig. 1 and Plate 1). Later he returned to bug-infested houses and detected *T. cruzi* in the blood of sick children.

Fig. 1 Adult female triatomine bug (*Panstrongylus megistus*), with a single egg shown adjacent to the tip of the abdomen. (By courtesy of Dr T.V. Barrett.) (See also Plate 1.)

T. cruzi is a kinetoplastid protozoan. In addition to the nucleus, it has a second, microscopically visible, DNA-containing organelle—the kinetoplast. The main lifecycle stages (trypomastigote, amastigote, epimastigote) are distinguished by the position of the kinetoplast relative to the nucleus, and by the presence or absence of a free flagellum.

Vector-borne transmission of *T. cruzi* is by contamination of the mammal host with infected faeces of triatomine bugs, not by their bite. During or shortly after feeding, bugs release liquid faeces and urine on to the skin of the host. Infective forms (metacyclic trypomastigotes) penetrate mucous membranes or abraded skin. Inside the mammal, *T. cruzi* is primarily an intracellular parasite. Trypomastigotes enter non-phagocytic or phagocytic cells, in which they transform to ovoid or round amastigotes (no flagellum). Amastigotes multiply inside the cell by binary fission to produce a pseudocyst (Fig. 2 and Plate 2). After 5 days or more, the pseudocyst ruptures to release numerous new trypomastigotes, which reinvade cells or circulate in the blood. Multiplication may occur at the site of infection, but pseudocysts subsequently predominate in muscle, especially heart and smooth muscle. In the blood, trypomastigotes are small, often C-shaped, with a large terminal kinetoplast (Fig. 3 and Plate 3). Trypomastigotes do not multiply in the blood. Triatomine bugs become infected by taking a blood meal from an infected mammal; birds and reptiles are not susceptible to infection. Infection in the bug is confined to the alimentary tract, where *T. cruzi* multiplies by binary fission as epimastigotes (kinetoplast adjacent to the nucleus). Metacyclic trypomastigotes are produced in the hindgut and rectum of the bug. All stages of the *T. cruzi* ifecycle can be cultured *in vitro*. *T. cruzi* can also be transmitted by blood transfusion, organ transplantation, transplacentally, to the infant via breast milk (rarely), and orally through food contaminated by triatomine faeces and the raw meat of infected mammals. Sexual transmission has not been documented.

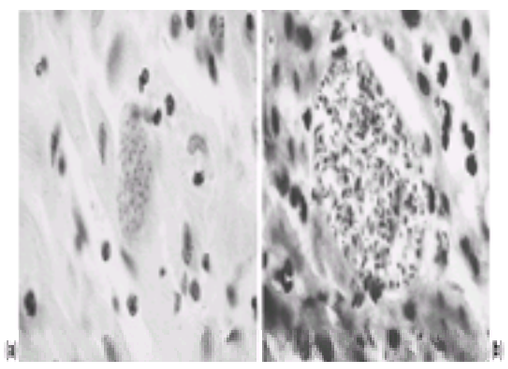

Fig. 2 (a) Pseudocyst of *Trypanosoma cruzi* in heart muscle. (By courtesy of J.E. Williams.) (b) Pseudocyst of *Trypanosoma cruzi* in umbilical cord, from a congenital case of Chagas' disease. (By courtesy of Dr Hipolito de Almeida.) (See also Plate 2.)

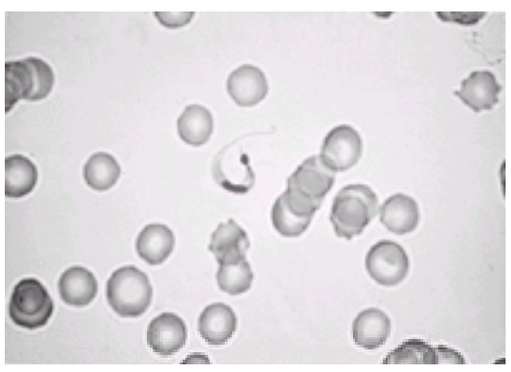

Fig. 3 *Trypanosoma cruzi* C-shaped trypomastigote in blood, note large posterior kinetoplast. (See also Plate 3.)

Epidemiology

T. cruzi is confined to the Americas, although closely related organisms of the same subgenus (*Schizotrypanum*) are cosmopolitan in bats. The vast majority of the 133 triatomine bug species are also restricted to the Americas. Their natural habitats are the refuges of mammals, birds, and reptiles, in trees, in burrows, and among rocks. All mammals are thought to be susceptible to *T. cruzi*, which has been reported from at least 150 mammal species. The opossum (*Didelphis* spp.) is the most common sylvatic host. A few triatomine species thrive as domestic colonies. More than 10 000 bugs have been found in a single house. *Triatoma infestans* is widespread in Southern Cone countries of South America (Argentina, Bolivia, Brazil, Chile, Paraguay, Uruguay, and in southern Peru). *Rhodnius prolixus* is the common vector in northern South America and Central America, with *Triatoma dimidiata* as secondary vector in that region. *Panstrongylus megistus* infests central and eastern Brazil, and *Triatoma brasiliensis* north-eastern Brazil. Animals that share human dwellings, such as guinea-pigs, dogs, cats, rats, and mice are domestic reservoirs of *T. cruzi* infection. Chickens, although not susceptible to *T. cruzi*, encourage bug infestation and can sustain large colonies.

Serological surveys suggest that up to 20 million people are infected with *T. cruzi* in South and Central America. In some communities seropositivity rates exceed 70 per cent. As expected from the precarious contaminative route of transmission, prevalence rises with age. Based on prevalence, it is estimated that up to 300 000 new infections might occur in Latin America each year. Less than 200 cases are known from the Amazon Basin, but that is because the local forest vectors do not colonize houses. For the same reason, autochthonous infection is very rare in the United States.

Initial acute infections are frequently asymptomatic or overlooked. It is thought that less than 10 per cent of acute infections in children are fatal. Morbidity due to Chagas' disease arises primarily from the chronic infection. Once acquired, infection is usually carried for life. Around 30 per cent of those infected will subsequently display electrocardiograph (**ECG**) abnormalities and chagasic cardiomyopathy, and a proportion of those have associated megaoesophagus or megacolon.

There are marked regional differences in the epidemiology of Chagas' disease. Mega syndromes are common in central and eastern Brazil but virtually unknown in northern South America and Central America. Molecular genetics' research has shown that *T. cruzi* is not a single entity, but a species with a vast subspecific heterogeneity. There are at least two radically distinct strain groups, TC1 and TC2. TC1 predominates in sylvatic and domestic transmission cycles, north of the Amazon. TC2 predominates in domestic transmission cycles in Southern Cone countries.

Pathogenesis and pathology

At the portal of entry, local multiplication of *T. cruzi* may lead to unilateral conjunctivitis or to a skin lesion (Fig. 4 and Plate 4). Unruptured pseudocysts in muscle apparently generate no inflammatory response. Pseudocyst rupture is followed by infiltration of lymphocytes, monocytes, and/or polymorphonuclear cells. Antigens released from pseudocysts may spread and be adsorbed on to adjacent uninfected cells. Such uninfected cells may be attacked by the immune response of the host, and destroyed. In this way expanded focal lesions may be produced. Postmortem histology on human hearts and experimental studies in dogs have demonstrated a clear association between ECG abnormalities and focal lesions in the conducting system of the heart. Much damage may occur in the acute phase of infection, particularly if pseudocysts are numerous. Postmortem histology has demonstrated that neurone loss is a feature of chagasic cardiopathy and of mega syndromes. Neurone loss may be exacerbated by further disease or age-related loss. Thus a threshold may be reached, often many years after the acute infection, at which organ function is perturbed. Further ECG abnormalities, aperistalsis, and organ enlargement may ensue. This 'neurogenic' pathogenesis has been linked to sudden death.

Fig. 4 Romaña's sign in acute Chagas' disease. (See also Plate 4.)

Pathological exposure of normal host sequestered antigens, or sharing of antigens between *T. cruzi* and its host, may precipitate autoimmune pathogenesis. Some chronic chagasic cardiomyopathy is said to display a renewed intense inflammatory response and a progressive diffuse myocarditis, and a slow decline in cardiac function.

The contribution of the lifelong infection to the pathogenesis of chronic Chagas' disease is controversial. After the initial acute phase, trypomastigotes are only detectable in the blood by sensitive indirect methods. Similarly, pseudocysts in the tissues are infrequent, but are detectable immunologically and by amplification of *T. cruzi* DNA.

T. cruzi infection is controlled primarily by a cell-mediated immune response, especially the TH1 arm of the immune response. Patients immunocompromised by **AIDS** (acquired immunodeficiency syndrome) have impaired TH1 responses. Thus **HIV** (human immunodeficiency virus)-positive patients chronically infected with *T. cruzi* may suffer reactivated acute Chagas' disease, with microscopically patent parasitaemia, and poor prognosis.

At the level of gross pathology, substantial megacardia may be seen. Thinning of the myocardium may be present, with focal aneurysms visible upon transillumination, especially at the apex of the left ventricle (Fig. 5) and thrombus in the right atrial appendage (Fig. 6). Apical aneurysm is considered to be a pathognomonic sign of chronic chagasic cardiomyopathy. Megaoesophagus (Fig. 7) and megacolon (see Fig. 9) may show enormous dilatation and thinning of the wall. Chagasic megaoesophagus is more frequent than chagasic megacolon, but both are known from single individuals and each is often accompanied by chagasic heart disease. Chagasic megaoesophagus may be a prelude to carcinoma.

Fig. 5 Apical aneurysm of the left ventricle in chronic Chagas' disease. (By courtesy of Dr J.S. de Oliveira.) (See also Plate 5.)

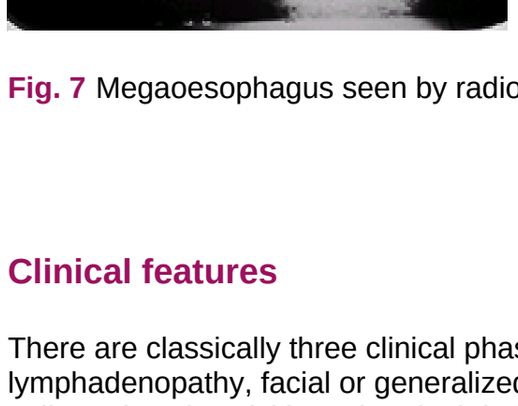

Fig. 6 Mural thrombus filling the right atrial appendage. (Copyright D.A. Warrell.) (See also Plate 6.)

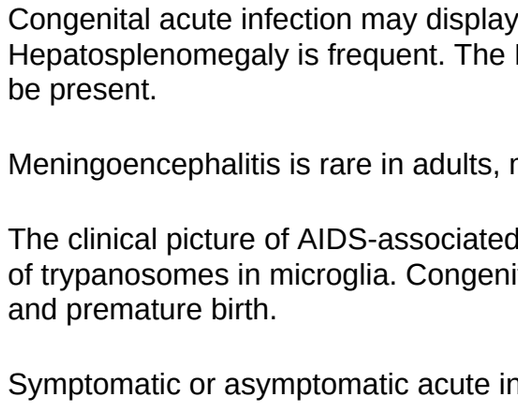

Fig. 7 Megaoesophagus seen by radiography in chronic Chagas' disease. (By courtesy of Dr J.S. de Oliveira.) (See also Plate 7.)

Clinical features

There are classically three clinical phases of Chagas' disease. Clinical signs in the acute phase may be fever, myalgia, headache, hepatosplenomegaly, generalized lymphadenopathy, facial or generalized oedema, rash, vomiting, diarrhoea, and anorexia. If *T. cruzi* has infected the eye, Romaña's sign may be present, with unilateral conjunctivitis and periophthalmic oedema (Fig. 4). If the portal of entry is the skin, an indurated oedematous cutaneous lesion (chagoma) may be seen. Regional lymphadenopathy may be present. Multiple chagomas may occasionally occur in acute-phase infections in infants. ECG abnormalities may include sinus tachycardia, increased P–R interval, T-wave changes, and low QRS voltage. The incubation period may be as short as 2 weeks, or as long as several months if infection is due to transfusion of contaminated blood. General lymphadenopathy and splenomegaly are frequent in blood transfusion-acquired infections.

Congenital acute infection may display fever, oedema, metastatic chagomas, neurological signs such as convulsions, tremors, and weak reflexes, and apnoea. Hepatosplenomegaly is frequent. The ECG is usually normal but low-voltage complexes, reduced T-wave height, and longer atrioventricular (**AV**) conduction time may be present.

Meningoencephalitis is rare in adults, more frequent in infants, common in immunocompromised patients, and carries a poor prognosis.

The clinical picture of AIDS-associated chagasic meningoencephalitis may be similar to toxoplasmosis. Haemorrhagic necrotic encephalitis is described in the nests of trypanosomes in microglia. Congenital infection may resemble toxoplasmosis, cytomegalic inclusion disease, or syphilis, with an increased likelihood of abortion and premature birth.

Symptomatic or asymptomatic acute infection may be followed by a symptom-free indeterminate phase of unpredictable length, which may be lifelong.

Chronic-phase symptoms may emerge in up to 30 per cent of patients recovering from the acute phase. Cardiac symptoms include arrhythmias, palpitations, chest pain, oedema, dizziness, syncope, and dyspnoea. The cardiac enlargement may be massive with chronic congestive cardiac failure, apical aneurysm (Fig. 5 and Plate 5), and thrombus in the right atrial appendage (Fig. 6 and Plate 6). The cardiac conducting system is involved, especially the sinus node, bundle of His and AV node, in which there is mononuclear and mast-cell infiltration, inflammation, and fibrosis. Characteristic ECG abnormalities are right bundle-branch block (**RBBB**) and left anterior hemiblock (**LAH**). AV conduction abnormalities, including AV block, may be present. Arrhythmias may include sinus bradycardia, sinoatrial block, ventricular tachycardia, primary T-wave changes, and abnormal Q-waves. The severity of heart disease is graded by the degree of disturbance. Sudden death is attributable, not to ruptured aneurysm but to arrhythmias often precipitated by exercise (e.g. on the football field). Radiography may reveal megacardia (Fig. 8). Signs of oesophageal involvement include loss of peristalsis, regurgitation, and dysphagia (Fig. 7 and Plate 7). Parotid enlargement may be associated. In megacolon there may be failure of defaecation, constipation, and faecaloma (Fig. 9 and Plate 8). Progressive dilatation of either organ can be graded clinically according to severity and may be detectable by radiography. Megaduodenum and megaureter are also described. The lymph nodes between the pulmonary trunk and the aorta are frequently enlarged.

Fig. 8 Chest radiograph showing gross cardiac enlargement in a Brazilian woman with chronic Chagas' disease. (Copyright D.A. Warrell.)

Fig. 9 Megacolon postmortem in chronic Chagas' disease. (By courtesy of Dr J.S. de Oliveira.) (See also Plate 8.)

A differential diagnosis requires distinction from other types of heart disease and ECG abnormality. RBBB and LAH are indicative, but a history of exposure to *T. cruzi* infection and laboratory diagnostic evidence must be considered (see below).

Laboratory diagnosis

A history of exposure to triatomine bugs, to potentially contaminated transfused blood, or a prolonged stay in endemic regions must be considered.

Motile trypomastigotes might be seen in unstained, wet blood preparations examined by microscopy. Nevertheless, parasitaemia is often scanty or undetectable by this method. The sensitivity of parasitological diagnosis may be enhanced by concentration methods, such as: microscopy of the centrifugation pellet from separated serum (Strout's method), microscopy of the haematocrit buffy coat layer, microscopy of Giemsa-stained thick films, or microscopy of the centrifugation sediment after lysis of red blood cells with 0.87 per cent ammonium chloride. All these tests may be negative if parasitaemia is low. Potentially infected blood must be handled with care, especially during haematocrit centrifugation, as a single trypomastigote can give rise to infection. Multiple blood cultures may also be performed, with a sensitive blood agar-based medium and physiological saline overlay. Even more sensitive than blood culture is xenodiagnosis, in which hungry fourth or fifth instar bugs from a triatomine colony, raised from bug eggs and fed only on birds, are allowed to feed on the patient. Bugs are applied in a plastic pot contained in a discrete black bag, which is tied beneath the patient's forearm. The bugs are dissected 20 to 25 days later. The hindgut and rectum are drawn out into a drop of sterile physiological saline, mixed with a blunt instrument (microspatula), and observed microscopically for motile epimastigotes and trypomastigotes. Dissection should be performed behind a small, Perspex safety screen or in a microbiological safety cabinet. *R. prolixus* is the most avid feeder for xenodiagnosis but may cause delayed hypersensitivity reactions in sensitized patients. The local vector should be used as the susceptibility of triatomine species varies with the strain of *T. cruzi*.

After the acute-phase infection, all the above methods of parasitological diagnosis will fail except xenodiagnosis, and possibly multiple blood cultures. Up to 50 per cent of patients in chronic phase may yield a positive xenodiagnosis, providing at least 10 triatomine bugs are used. Although polymerase chain reaction (**PCR**) amplification of *T. cruzi* DNA is sensitive and specific, it not available as a routine diagnostic test. Serum antibody is produced within a few days of *T. cruzi* infection and persists for life in untreated patients. There is an early IgM response, but it is not sustained at the high levels seen in African trypanosomiasis. Persistent IgG may be detected by the enzyme-linked immunosorbent assay (**ELISA**), by the indirect fluorescent-antibody test (**IFAT**), or by the indirect haemagglutination test (**IHAT**). Complement fixation, developed in 1913, is effective but now seldom used. Recombinant antigens are under trial but have not yet been adopted. Cross-reactions may occur, with visceral and mucocutaneous leishmaniasis, treponematoses, and possibly with other hyperimmune responses or autoimmune diseases. Serological assays must be standardized with negative and positive control sera, and by reference to experienced external reference centres to check reproducibility. Transplacentally acquired IgG may persist for up to 9 months in infants born of seropositive mothers. However, IgM specific seropositivity in such infants is an indicator of congenital infection.

Treatment

Proven acute cases must be treated promptly in an effort to minimize tissue damage and neurone loss. The synthetic oral nitrofuran, nifurtimox (LAMPIT ®) from Bayer was the first successful drug for the treatment of Chagas' disease but it is no longer readily available. Benznidazole (Rochagan ®) from Roche is now the sole first-line chemotherapy. An oral nitroimidazole, the adult dosage is 5 to 7 mg/kg for adults, in two divided doses, for 60 days; for children, 10 mg/kg also in two divided doses for 60 days. Adverse effects may demand interruption of treatment. These include rashes, fever, nausea, peripheral polyneuritis, leucopenia, and (rarely) agranulocytosis. Children tolerate treatment better than adults. Double or even higher doses have been used for immunocompromised patients, especially if meningoencephalitis is present. There is no guarantee that a full course of treatment will eliminate the infection. Although the value of drug treatment for chronic infections is still debated, it is favoured for children.

Chemotherapy is an important part of supportive treatment. In acute-phase heart failure, sodium intake is restricted and diuretics and digitalis may be indicated. Meningoencephalitis may require anticonvulsants, sedatives, and intravenous mannitol. Heart failure due to Chagas' disease may require vasodilatation (angiotensin-converting enzyme inhibitors) and maintenance of normal serum potassium levels; digitalis is a last resort because it may aggravate arrhythmias. A pacemaker may be fitted to improve bradycardia not responding to atropine, or for atrial fibrillation with a slow ventricular response that is not responsive to vagolytic drugs, or for complete AV block. Amiodarone has been suggested as the most useful drug to treat arrhythmias but it may still be aggravating. For ventricular extrasystoles lidocaine (lignocaine), mexiletine, propafenone, flecainide, and b-adrenoreceptor antagonists may be effective. Lidocaine may be used intravenously in emergencies. It is essential to consult detailed WHO expert reports and physicians with substantial experience in the management of chagasic heart disease.

Surgery is a vital part of case management for Chagas' disease. Resection of ventricular aneurysms has been suggested. Specialized surgery has been developed in Brazil for the treatment of megaoesophagus and megacolon. Early megaoesophagus may respond to balloon dilatation. The Heller–Vasconcelos operation, in which a portion of muscle at the junction of the oesophagus and stomach is removed, may alleviate megaoesophagus. Severe megaoesophagus requires replacement of the distal oesophagus, for example with a portion of jejenum. The modified Duhamel–Haddad operation has been considered the most successful surgery for correction of a megacolon: after resection, the colon is lowered through the retrorectal stump as a perineal colostomy. Subsequent suturing, under peridural anaesthesia, gives a wide junction between the colon and the rectal stump.

Prognosis, even in treated patients who show serological reversion, is unpredictable as the sequelae of damage due to the acute phase of Chagas' disease cannot be foreseen.

Prevention and control

There is no vaccine against Chagas' disease and no immunotherapy.

Chagas' disease flourishes on the back of poverty and in poor housing conditions. There are proven methods of controlling domestic triatomine bugs. These depend on insecticide spraying, health education, community support, and house improvement. Synthetic pyrethroids are the insecticides of choice, and several commercial sources are available. Vector control programmes consist of preparatory, attack, and vigilance phases. In the preparatory phase, the distribution of all dwellings must be mapped, the presence of infested houses assessed, and the attack and vigilance phases costed and planned. The attack phase involves spraying all houses and peridomestic buildings, irrespective of whether bugs have been found. During the vigilance phase, the community plays an essential role in reporting residual bug infestations, which elicit a rapid respraying response for the affected sites. Serology is vital for monitoring the success of control programmes. Children born after control programmes begin should be serologically negative beyond 9 months of age (to exclude transplacental transfer of IgG) except for infrequent cases of congenital transmission.

Blood donors in, or from, endemic areas should be screened serologically. If conditions demand the use of seropositive blood it can be decontaminated with crystal violet (250 mg per litre) and storage at 4 °C for at least 24 h. Potentially infected organ donors or recipients should be screened serologically. Seropositive

immunosuppressed recipients are likely to suffer reactivated acute-phase infection. Prophylactic chemotherapy with benznidazole may be effective.

The Southern Cone Programme launched a massive effort to eliminate *T. infestans* from Argentina, Bolivia, Brazil, Chile, Paraguay, Uruguay, and from southern Peru. Domestic infestation in Brazil has been reduced by 85 per cent. Uruguay and Chile are essentially free of vector-borne and blood-transfusion transmission. Substantial progress has also been made in the other participating countries. Similar international collaborations are planned for the Andean Pact countries and for Central America. Reinvasion of sylvatic bugs into domestic habitats may complicate vector control in some regions. A surveillance programme and rapid responses to new domestic triatomine populations are planned to protect the Amazon against domiciliation of vectors.

T. cruzi is of immense research interest. It is not entirely clear how the organism evades the host immune response. Furthermore, the pathogenesis of Chagas' disease is not fully understood. Molecular methods have radically changed our understanding of the epidemiology of *T. cruzi* infection. Molecular features unique to trypanosomatids (trypanosomes and leishmanias) make *T. cruzi* an attractive model for molecular biologists. Further research is required to produce a non-toxic, low-cost oral drug, which would eliminate the reservoir of infection in humans, and to clarify further the population genetics and epidemiological significance of diverse strains. The origins and evolution of the organism and its vectors are also of considerable academic interest.

Trypanosoma rangeli

The second human trypanosomiasis in the New World is due to *T. rangeli* infection. *T. rangeli* is also transmitted by triatomine bugs, in particular the genus *Rhodnius*. In *Rhodnius* spp., however, *T. rangeli* traverses the wall of the alimentary tract, infects the haemocoel, and reaches the salivary glands, in which the metacyclic infective trypomastigotes are produced. *T. rangeli* is thus transmitted by the bite of the triatomine bug and not by contamination with bug faeces. Although enzootic *T. rangeli* infection is widespread in Latin America, transmission to humans is virtually confined to areas in which *R. prolixus* is the domestic vector of *T. cruzi*. Co-infections of *T. cruzi* and *T. rangeli* may occur. The organism appears to be non-pathogenic in humans. *T. rangeli* can be pathogenic to *Rhodnius* spp. The importance of *T. rangeli* lies in the fact that it may confuse xenodiagnosis to detect *T. cruzi*. With care and experience, *T. rangeli* can be distinguished from *T. cruzi* either by its long slender epimastigotes (up to 80 µm in length), or by its smaller kinetoplast, or by its presence in the haemolymph or salivary glands of some xenodiagnosis bugs. The lifecycle in the mammalian host is uncertain, but *T. rangeli* is thought to divide in the peripheral blood. Trypomastigotes are rarely seen in human blood: they are much larger than *T. cruzi*, with a small subterminal kinetoplast. Antibodies to *T. cruzi* certainly crossreact strongly with *T. rangeli*. Based on experimental work in mice, *T. rangeli* infections are thought to induce very low crossreactive antibody titres to *T. cruzi*.

Further reading

Lent H, Wygodzinsky P (1979). Revision of the Triatominae (Hemiptera, Reduviidae) and their significance as vectors of Chagas disease. *Bulletin of the American Museum of Natural History*. **163**, 123–520. [An essential taxonomic monograph for all those interested in triatomine bugs, with keys for identification, but note that more species have since been described.]

Miles MA (1997). New World trypanosomiasis. In: Cox FEG, Kreier JP, Wakelin D, eds. *Topley and Wilson's microbiology and microbial infections*, pp. 283–302. London, Arnold. [A detailed account of the causative agent, the disease, and control efforts.]

Pan American Health Organization (1994). *Chagas disease and the nervous system*, Scientific publication No. 547. PAHO, Washington, DC. [An entire volume devoted to the interaction between *T. cruzi* and the nervous system.]

Raia AA (1983). *Manifestações Digestivas da Moléstia de Chagas*. Sarvier, São Paulo, Brasil. [For the surgeon, fascinating accounts of the development of lifesaving procedures, especially correction of megaoesophagus and megacolon (in Portuguese).]

World Health Organization (1991). *Control of Chagas disease*, Technical Report Series 811. WHO, Geneva. [Not strictly on control, but one of the best clinical reviews of Chagas' disease in the English language.]

7.13.12 Leishmaniasis

A. D. M. Bryceson

Leishmaniasis is caused by parasites of the genus *Leishmania*, which are transmitted by phlebotomine sandflies. The infection may be anthroponotic or zoonotic. In humans, the disease is usually either cutaneous or visceral. The most important variant is mucosal leishmaniasis of South and Central America. In certain places the disease is common and important, but there are few accurate statistics. The World Health Organization estimates 500 000 cases of visceral leishmaniasis and 1.5 to 2 million cases of cutaneous leishmaniasis annually, with 200 million people at risk of each disease.

Aetiological agent and lifecycle

In its vertebrate host the oval amastigote form of the parasite, which is 2 to 3 µm in diameter, is found in cells of the reticuloendothelial system (Fig. 1). In the sandfly or in culture medium it is in the elongated, motile, promastigote form with an anterior flagellum.

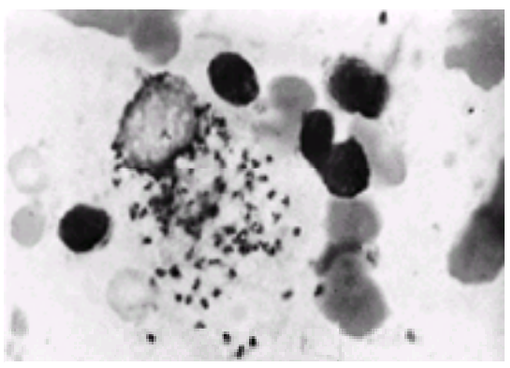

Fig. 1 Amastigotes of *Leishmania donovani* in a reticuloendothelial cell from the splenic aspirate of a patient with visceral leishmaniasis.

The most important species of *Leishmania* that cause disease in humans and their own reservoir hosts are shown in Table 1; isoenzyme patterns and DNA hybridization are used to distinguish species.

Sandflies require a precise microclimate that is provided in certain places in each endemic focus at particular seasons of the year. Transmission is often seasonal. Amastigotes are ingested from blood or tissues of the mammalian host by the female fly, and transform into promastigotes in the gut, rendering the fly infective after about 10 days.

Cutaneous leishmaniasis

Epidemiology (see Table 1)

The vectors of *Leishmania major* live in rodent burrows. Hunters, travellers, tourists, and dwellers at oases or in new settlements are affected. The disease may be sporadic or epidemic. The vectors of *L. tropica* live in crevices in buildings and walls. The disease may be endemic or epidemic. The vector of *L. aethiopica* bites people sleeping in their huts. The disease is endemic and most people are affected by early adulthood. *L. infantum* causes simple, self-healing skin lesions in some parts of southern Europe and North Africa. *L. donovani* causes post-kala-azar dermal leishmaniasis in India.

In the New World, transmission is usually in the forest. *L. brasiliensis*, the major cause of American cutaneous and mucosal leishmaniasis, is the most widely distributed of the New World species. Its vectors are highly anthropophilic and human infection is common. Periurban and urban foci of infection are increasing. Infection with *L. peruviana* occurs in high Andean valleys, where it may be locally common.

Pathogenesis and pathology

Leishmania inoculated by the sandfly invade and multiply in macrophages in the skin. The parasitized macrophage granuloma is infiltrated by lymphocytes and plasma cells. Piecemeal or focal necrosis destroys parasitized cells. The overlying epidermis shows hyperkeratosis, and ulcerates. In chronic lesions epithelioid cells and Langhans giant cells produce a picture similar to that of non-caseous tuberculosis. Rarely, the cellular immune response is suppressed and histology shows heavily parasitized macrophages with little or no lymphocytic infiltrate, characteristic of diffuse cutaneous leishmaniasis.

L. aethiopica, *L. mexicana*, and *L. brasiliensis* may invade cartilage. Cartilaginous lesions are extremely chronic. *L. brasiliensis*, and occasionally *L. panamensis* or *L. guyanensis*, may metastasize through the bloodstream to sites deep in the mucosa of the upper respiratory tract, where they may lie dormant. After months or years a lesion develops characterized by necrosis, vasculitis, and tissue destruction.

Immunity to a given species of *Leishmania* is usually lifelong. Second infections occur occasionally, especially in the elderly or immunosuppressed.

Clinical features

After an incubation period of a few days to several months an erythematous nodule develops at the site of the infected sandfly bite. A golden crust forms. The sore reaches its final size, usually 1 to 5 cm in diameter, over weeks or months. The crust may fall away leaving an ulcer with a raised edge (Fig. 2 and Plate 1). Satellite papules are common. After months or years the lesion starts to heal leaving a depressed, mottled scar. Secondary infection is unimportant. The lesion is not normally painful, but may disfigure or disable if scarring is severe or over a joint. Draining lymphatic vessels may be thickened or nodular.

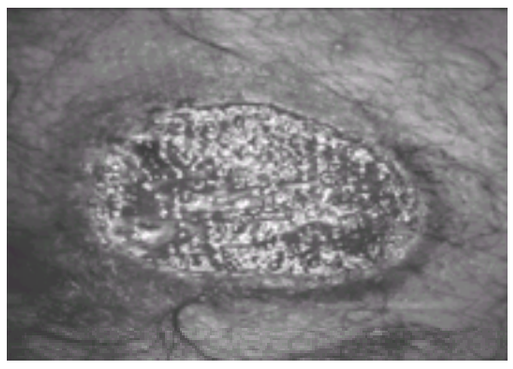

Fig. 2 Shallow ulcer with raised edge due to *L. brasiliensis* (copyright A.D.M. Bryceson). (See also Plate 1.)

There are many variations on this classical pattern. Sores due to *L. major* form and heal rapidly (mean 3–5 months) and may be inflamed and exudative: the so-called wet or rural sore. Sores due to *L. tropica* tend to be less inflamed and to heal more slowly (mean 10–14 months): the so-called dry or urban sore. Lesions due to *L. infantum* have an incubation period of many months, and may persist over several years. In *L. aethiopica* infections lesions are usually central on the face. Satellite papules accumulate to produce a slowly growing, shiny tumour or plaque that may not crust or ulcerate, taking 2 to 5 years to heal (Fig. 3); mucocutaneous leishmaniasis may develop, producing swelling of the lips and expansion and elongation of the nose.

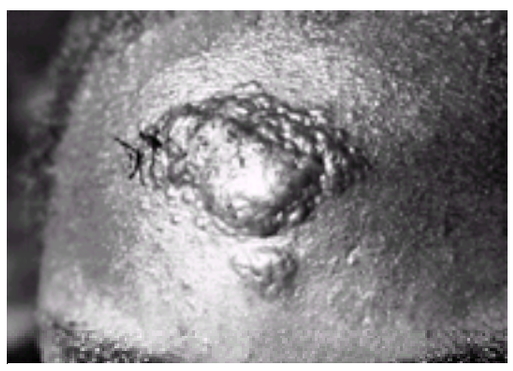

Fig. 3 Spreading nodular lesion, typical of *L. aethiopica*, Kenya.

L. brasiliensis often causes deep, spreading ulcers, which heal over 6 to 24 months. Up to 15 per cent of patients will relapse after spontaneous or therapeutic cure. *L. mexicana* lesions are commonly on the limbs or side of the face, and heal in 6 to 8 months. Sores on the pinna of the ear may invade the cartilage, persist for many years, and destroy the pinna.

Three forms of cutaneous leishmaniasis do not heal spontaneously: diffuse cutaneous leishmaniasis, leishmaniasis recidivans, and American mucosal leishmaniasis.

Diffuse cutaneous leishmaniasis

This occurs with *L. aethiopica* and *L. amazonensis* infections, but is rare. The primary nodule spreads locally without ulceration, and secondary blood-borne lesions appear on other sites in the skin, affecting especially the face and the cooler extensor surfaces of the limbs (Fig. 4). The eye, mucosae, viscera, and peripheral nerves are spared, in contrast with lepromatous leprosy with which it may be confused. The infection proceeds gradually over many years.

Fig. 4 Diffuse cutaneous leishmaniasis, caused by *L. aethiopica*, Ethiopia.

Leishmaniasis recidivans or lupoid leishmaniasis

This is a rare complication of *L. tropica* infection. The initial sore heals, but papules recrudesce in the edge of the scar and the lesion spreads slowly over many years (Fig. 5 and Plate 2).

Fig. 5 Lupoid or recidivans leishmaniasis in a citizen of Baghdad. (By courtesy of Dr Ahmed.) (See also Plate 2.)

American mucosal leishmaniasis, espundia

Up to 40 per cent of patients with untreated cutaneous ulcers due to *L. brasiliensis* may develop mucosal lesions, half of them within 2 years of the appearance of the original lesion, and 90 per cent within 10 years. About one in six patients gives no history of a previous skin lesion. In most cases the nasal mucosa is affected, and in one-third another site is also involved: the pharynx, palate, larynx, and upper lip, in order of frequency. The initial lesion is a nodule and the initial symptom is of nasal obstruction. It commonly presents as protuberant new growth of the nose or lips (Fig. 6 and Fig. 7 and Plate 3, Plate 4), or cicatrization which causes an elongated 'tapir' nose. Mucosal leishmaniasis is slowly destructive, the septum perforates, and eventually the whole nose and mouth may be destroyed. Death may result from secondary sepsis, starvation, or laryngeal obstruction.

Fig. 6 Swollen upper lip and nose due to mucosal leishmaniasis in Peru (copyright A.D.M. Bryceson). (See also Plate 3.)

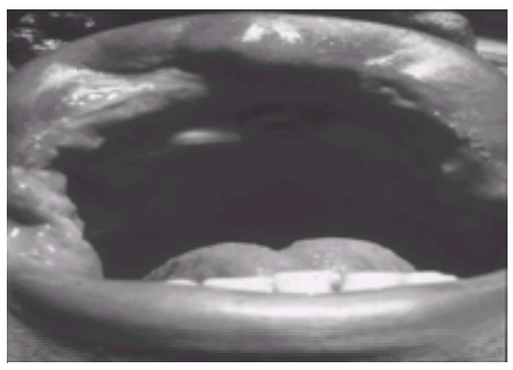

Fig. 7 Infiltration of lip and palate due to mucosal leishmaniasis in Peru (copyright A.D.M. Bryceson). (See also Plate 4.)

Laboratory findings

Parasitological diagnosis

Leishmania may normally be isolated from 80 per cent of sores during the first half of their natural course. The nodular part of the lesion is grasped firmly between the finger and thumb until it blanches. An incision a few millimetres long is made into the dermis with the point of a scalpel, which is used to scrape dermal tissue and juice. Material obtained may be used to inoculate special diphasic culture medium and to prepare smears for staining with Giemsa, Wright's, or Leishman's stain (Fig. 1). Biopsy material may be used to make impression smears, for culture and for histology. Diagnosis of mucosal leishmaniasis requires deep punch biopsy. Species diagnosis is desirable for American parasites, to assess the risk of mucosal leishmaniasis.

Immunological diagnosis

The leishmanin test is an intradermal test of delayed hypersensitivity which becomes positive in over 90 per cent of cases of self-healing forms of cutaneous leishmaniasis and mucosal leishmaniasis and is 95 per cent specific. Evaluation of a positive test must take into account naturally acquired positivity in the population at risk. Serology is unhelpful.

Treatment

Old World sores or those due to *L. mexicana*, *L. amazonensis*, and *L. peruviana* that are not troublesome may be left to heal naturally. But those that are disfiguring, potentially disabling, inconvenient, or around the ankle, where they heal slowly, should be treated either locally or systemically. Systemic treatment is required when there is risk that the sore may be due to *L. brasiliensis*, *L. panamensis*, or *L. guyanensis*, when the sore is too large or badly sited for local treatment, and for mucosal leishmaniasis, diffuse cutaneous leishmaniasis, and recidivans leishmaniasis.

Local treatment

Surgery, curettage, and cryotherapy are methods of removing small sores. Infiltration into the lesion with a pentavalent antimonial, twice weekly for 2 or 3 weeks, may be successful. Leishmanicidal ointments are under evaluation.

Systemic treatment (see Table 2 and Table 3 for dosage regimens)

All cutaneous species of *Leishmania* are sensitive to pentavalent antimonials in conventional dosage except *L. aethiopica*, when pentamidine or paromomycin may be used. Ketoconazole may be useful for *L. major* and *L. mexicana* infections. Patients with diffuse cutaneous leishmaniasis should be treated for at least 2 months longer than it takes to clear parasites from the skin, and relapses should be treated again promptly. Relapsed cases of mucosal leishmaniasis have usually become unresponsive to antimonials and should be treated with amphotericin B desoxycholate for at least 4 to 6 weeks or liposomal amphotericin B for 3 weeks. In addition they may require antibiotics for secondary sepsis, attention to nutrition, and later plastic surgery.

Visceral leishmaniasis

Epidemiology

Visceral leishmaniasis is found in four main zoogeographical zones (Table 1). Around the Mediterranean littoral, across the Middle East and central Asia, and in northern and eastern China human disease is endemic in many places. Children under 5 years of age are especially affected. In other places the disease is sporadic. Non-immune adults such as tourists, hunters, and soldiers are susceptible. The Ganges and Brahmaputra river valleys of India and Bangladesh are the home of epidemic visceral leishmaniasis, or kala-azar, which returns approximately every 15 to 20 years. The majority of cases are in young people under 15 years of age. In the interepidemic period the parasite survives in patients with post-kala-azar dermal leishmaniasis. Visceral leishmaniasis is endemic in parts of Sudan and Kenya. Older children and teenagers are most commonly affected. Sporadic cases also occur in nomads and visitors. An epidemic that began in southern Sudan in the late

1980s is still raging, and has caused over 100 000 deaths. It has been especially severe among refugees from the civil war.

In South America the disease is most common in northeastern Brazil, where older children are affected. Previously a rural disease, it is becoming increasingly important in towns.

Visceral leishmaniasis may be transmitted by blood transfusion from subclinical cases and appears unexpectedly in immunosuppressed patients, for example after renal transplantation, or as an opportunistic infection with HIV.

Pathogenesis and pathology

For every case of classical visceral leishmaniasis, there are about 30 subclinical infections that cause leishmanin positivity and lifelong immunity to *L. donovani.* Malnutrition predisposes to clinical disease. Established visceral infections are characterized by the failure of specific cell-mediated immunity. The leishmanin test is negative. The parasite multiplies freely in macrophages in the spleen, bone marrow, lymphoid tissues, and jejunal submucosa and Kupffer cells of the liver. Histology shows a variable degree of granuloma formation, and of interstitial inflammation in the liver that may lead to fibrosis. In the spleen especially there is massive reticuloendothelial hyperplasia and infiltration with plasma cells. Small splenic infarcts may develop.

Antibodies, polyclonal IgG, and immune complexes circulate at high concentration but rarely cause complications. About half of patients have mild malabsorption but seldom diarrhoea. Jaundice when present is usually due to intercurrent viral hepatitis. Spontaneous bleeding is unusual and is associated with hypoprothrombinaemia. Visceral leishmaniasis is characterized by anaemia, leucopenia, thrombocytopenia, and hypoalbuminaemia. The anaemia results mainly from shortened red-cell survival with destruction of cells in the spleen, together with splenic pooling and sequestration (hypersplenism). In young children, profound anaemia may develop rapidly as a result of severe haemolysis. Death is usually due to secondary infection.

Clinical features

The male/female ratio is between 3:1 and 4:1. The incubation period is usually 2 to 8 months. In endemic areas the onset is usually ill defined. The patient develops fever, discomfort from an enlarged spleen, abdominal swelling, weight loss, cough, or diarrhoea. Classically the fever spikes twice daily, usually without rigors, but daily, irregular, or undulant fevers are common. During an epidemic or in visitors to an epidemic area, the onset may be abrupt with high fever and rapid progression of illness with toxaemia, weakness, dyspnoea, and acute anaemia.

Physical examination of early cases may show only symptomless splenomegaly. Late cases are wasted with hair changes and pedal oedema typical of hypoalbuminaemia. Hyperpigmentation is characteristic of visceral leishmaniasis in India (kala-azar means black sickness). The spleen is huge, smooth, and non-tender unless there has been a recent infarct. The liver is moderately enlarged in one-third of cases. In Africa generalized lymphadenopathy is common.

Over months or years the patient becomes emaciated, with a distended abdomen (Fig. 8). Intercurrent infections are common, especially pneumococcal otitis, pneumonia, septicaemia, tuberculosis, measles, dysentery, other locally important infections, and rarely, cancrum oris. Untreated, 80 to 90 per cent of patients die.

Fig. 8 Visceral leishmaniasis in a Kenyan child. Note the wasting and massive enlargement of spleen and liver.

Post-kala-azar dermal leishmaniasis

Twenty per cent of Indian patients and 5 per cent of African patients develop a rash on the face and extensor surfaces of the arms and legs after recovery from visceral leishmaniasis. In India the rash begins after an interval of 1 or 2 years and progresses over many years: pale macules become erythematous plaques or nodules resembling lepromatous leprosy, and almost all the body surface may be involved (Fig. 9). In Africa the rash appears while the patient is still recovering, as discrete nodules which show a tuberculoid histology. It heals spontaneously within 6 months.

Fig. 9 Post-kala-azar dermal leishmaniasis in an Indian child, showing the typical hypopigmented macular rash. Note also the nodules on the lower lip.

Visceral leishmaniasis and AIDS

Visceral leishmaniasis may be associated with HIV infection and is an AIDS-defining illness in adults in southern Europe, where it is commonest among intravenous drug users. It may be due to reactivation of latent infection with *Leishmania* or to a recent infection. In Spain, over 50 per cent of adults with visceral leishmaniasis are HIV positive, and it is estimated that 9 per cent of HIV-infected individuals will acquire visceral leishmaniasis. The presentation may not be typical. Often the parasite is found by chance, for example in a rectal or skin biopsy taken for other purposes, or in bronchoscopic lavage. The bone marrow is teeming with parasites, but two-thirds of cases have no detectable antileishmanial antibodies. In 90 per cent of cases the CD4 count is less than 0.2×10^{6}/litre.

Laboratory diagnosis

Parasitological diagnosis

Leishmania may be isolated from reticuloendothelial tissue. Yields are of the order: spleen over 95 per cent, bone marrow or liver 85 per cent, African lymph node 65

per cent, and buffy coat 70 per cent. Bone marrow aspiration is most commonly used, but splenic aspiration is simple, painless, and safe if the prothrombin time is normal and the platelet count above 40 × 10^9/litre. Occasionally, the diagnosis is made accidentally on biopsy of bone marrow, liver, lymph node, or bowel mucosa. Antibodies are present in high titre. Indirect immunofluorescence is suitable for individual cases. Enzyme-linked immunosorbent assay or direct agglutination are the techniques of choice for field diagnosis. The leishmanin test is negative.

Other findings

There is normochromic, normocytic anaemia without reticulocytosis, and neutropenia, eosinopenia, and thrombocytopenia. Serum albumin is low (~20 g/litre) and globulin high (~70 g/litre), IgG and IgM being approximately thrice and twice the normal population values. Hepatic enzymes and prothrombin and partial thromboplastin times are usually normal.

Treatment

Chemotherapy (see Table 2 and Table 3 for dosage regimens)

Liposomal amphotericin B (AmBisome) by intravenous infusion is the best drug for visceral leishmaniasis. It is concentrated and retained in reticuloendothelial cells and is not toxic. All patients respond promptly, but HIV-coinfected patients relapse. At the moment it is far too costly for most countries where visceral leishmaniasis is endemic. Therefore, a pentavalent antimonial remains the drug of choice in most situations.

Sodium stibogluconate containing 100 mg antimony (Sb) per millilitre and meglumine antimoniate containing 85 mg Sb/ml, are of equal efficacy and toxicity. The drug is administered by intramuscular injection, which may be painful, or by intravenous injection through a fine-gauge needle, slowly or by infusion in 50 to 100 ml of 5 per cent dextrose over 20 min to reduce the risk of venous thrombosis. Treatment is given daily for 21 days. Usually the drug is well tolerated, but towards the end of treatment there may be malaise, anorexia, nausea, vomiting, and muscle pains. Should toxic effects develop, rest for 1 day and reduce each dose by 2 mg Sb/kg. Hepatic and pancreatic enzyme levels may rise and haemoglobin levels fall, but they return to normal when treatment is stopped. The electrocardiogram develops unimportant T-wave changes. At higher doses the corrected QT interval may be prolonged, heralding the development of a serious arrhythmia. If it is essential, for example during an epidemic, to give a shorter course of treatment, 10 mg Sb/kg may safely be given every 8 h for 10 days.

The aminoglycoside antibiotic paromomycin or aminosidine (IDA Pharmamed) is equally effective and well tolerated. It is given by intramuscular injection or intravenous infusion over 90 min.

In India, conventional amphotericin B desoxycholate is particularly effective. A new oral drug miltefosine is undergoing trials

Patients who are immunosuppressed as a result of HIV coinfection or immunosuppressive drugs respond slowly, require longer treatment, and are more liable to relapse than immunocompetent patients. Ideally, treatment of such patients should be monitored by splenic aspirate counts of parasites, and continued for 2 to 3 weeks beyond parasitological cure. Aminosidine is the drug of choice, as it is well tolerated and not prohibitively expensive. Renal function and hearing should be monitored. Clinical pancreatitis has been reported with the antimonials. Liposomal amphotericin B, although well tolerated, does not prevent relapse.

Supportive treatment

Intercurrent infection must be sought and treated, and nutritional deficiencies corrected. Blood transfusion is rarely needed.

Response to treatment

Fever, splenic size, haemoglobin, serum albumin, and body weight are useful monitors of progress. Proof of parasitological cure is not usually necessary. Reassessment at 6 weeks and 6 months will detect over 90 per cent of relapses. Relapse rates should be almost zero in Mediterranean and Indian disease and about 2 per cent in African disease. Relapsed patients are slower to respond, and run a 40 per cent chance of further relapse(s) and of becoming unresponsive to antimony. Primary resistance to antimonials is increasing in India where the first choice lies between aminosidine and amphotericin B desoxycholate.

Prevention and control of cutaneous and visceral leishmaniasis

Prevention is a matter of controlling reservoir hosts and sandfly vectors, or of avoiding bites by vectors. Successful control requires an accurate knowledge of transmission in each ecological focus.

In the Old World, urban cutaneous leishmaniasis is controlled by case-finding and treatment, better housing, and domestic spraying with residual insecticides, while rural leishmaniasis is controlled in the Middle East and North Africa by the destruction of gerbil colonies. Mediterranean and urban visceral leishmaniasis in South America may be controlled by the destruction or treatment of dogs. In India, mass campaigns to spray houses and cattle sheds are needed. In the interepidemic period, cases of post-kala-azar dermal leishmaniasis should be sought and treated.

Individuals may take precautions to prevent infection during the season of transmission by the use of insect repellent creams and fine mesh bed nets or chadors impregnated with permethrin.

Further reading

Alvar J *et al*. (1997). Leishmania and human immunodeficiency virus coinfection: the first 10 years. *Clinical Microbiology Review* **10**, 298–319.

Berman JD *et al*. (1998). Efficacy and safety of liposomal amphotericin B (AmBisome) for visceral leishmaniasis in endemic developing countries. *Bulletin of the World Health Organization* **76**, 25–32.

Grimaldi G Jr, Tesh RB, McMahon-Pratt D (1989). A review of the geographic distribution and epidemiology of leishmaniasis in the New World. *American Journal of Tropical Medicine and Hygiene* **41**, 687–725.

Jha TK *et al*. (1998). Randomised controlled trial of aminosidine (paromomycin) v sodium stibogluconate for treating visceral leishmaniasis in North Bihar, India [see comments]. *British Medical Journal* **316**, 1200–5.

Jha TK *et al*. (1999). Miltefosine, an oral agent, for the treatment of Indian visceral leishmaniasis [see comments]. *New England Journas of Medicine* **341**, 1795–800.

Seaman J, Mercer AJ, Sondorp E (1996). The epidemic of visceral leishmaniasis in western Upper Nile, southern Sudan: course and impact from 1984 to 1994. *International Journal of Epidemiology* **25**, 862–71.

WHO Expert Committee (1990). Control of the leishmaniases. *World Health Organization Technical Reports Series* **793**, 1–158.

7.13.13 Trichomoniasis

J. P. Ackers

Urogenital trichomoniasis is caused by infection with the protozoan *Trichomonas vaginalis*. About 170 million new cases each year may well make it the world's commonest non-viral sexually transmitted infection.

In clinical specimens or culture *T. vaginalis* is a motile, round or oval flagellate 10 to 13 µm long and 8 to 10 µm wide; fixed and stained it is about 25 per cent smaller (Fig. 1 and Plate 1). Diagnostic features include the jerky motility, undulating membrane, and microtubular rod (axostyle), which runs through the body and projects as a thin spine from the posterior end. In contact with vaginal epithelial cells *in vitro* the organism becomes extremely flattened and adherent. The lifecycle is simple; no resistant cysts are formed and there are no intermediate or reservoir hosts.

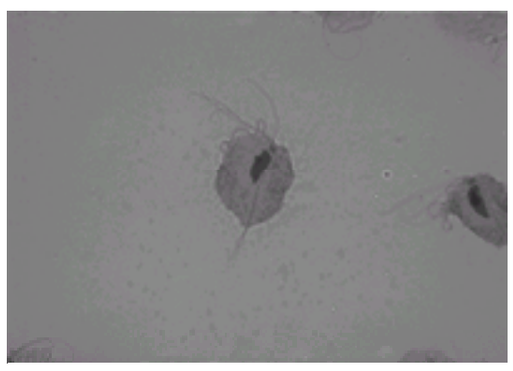

Fig. 1 Trichomonads, Giemsa stain, in vaginal secretions. (Copyright J.P. Ackers.) (See also Plate 1.)

Two other trichomonads—*T. tenax* and *Pentatrichomonas hominis*—are uncommon and probably harmless human parasites of the mouth and large bowel, respectively. All three species are site specific. Urogenital trichomoniasis is not due to contamination from other sites.

Epidemiology

Although it is often difficult to isolate the organism from male contacts of infected women, all epidemiological evidence suggests that the vast majority of infections are sexually acquired. *T. vaginalis* has been shown to survive for many hours at room temperature if kept damp so the theoretical possibility of non-venereal transmission exists. It is also known that a very small proportion of female babies of infected mothers will become infected during birth, but the infection is transient and trichomoniasis discovered in a child should immediately raise the suspicion of sexual abuse.

Very few studies have been made of genuinely unselected populations; the majority have examined either pregnant women or those attending sexually transmitted disease clinics. There are wide variations but most report that 10 to 25 per cent of patients are infected, although the full range is 0 to 63 per cent. Usually female cases outnumber males by 5 or 10 to 1. In several developed countries there has been a steady decline in the incidence of trichomoniasis in the past two decades, but this has not occurred in less-developed countries nor in deprived inner-city areas in industrialized nations. Human trichomoniasis is becoming a disease of the underprivileged.

Pathogenesis

In vitro, *T. vaginalis* has a well defined, contact-mediated, cytotoxic effect, but the relationship of this to pathogenesis *in vivo* is not known. The organism activates complement and attracts neutrophils; several together can kill the parasite, but their presence in large numbers may be responsible for much of the pathology observed. It seems likely that differences in clinical severity are due to both host and parasite factors.

Symptoms

In women trichomoniasis may present as anything from an asymptomatic infection (10 to 50 per cent of cases) to an acute inflammatory diseases with a copious and malodorous discharge; vulvovaginal soreness and irritation, dysuria, and dyspareunia are also frequently mentioned. The discharge may vary over time and, untreated, the infection may be spontaneously lost or persist for months or years. A recent study showed trichomoniasis significantly associated with symptoms of yellow vaginal discharge and vulvar itching and signs of colpitis macularis (strawberry cervix), purulent vaginal discharge, and vulval and vaginal erythema. Colpitis was seen frequently if colposcopy was undertaken, but hardly ever by naked-eye examination. Vaginal pH is usually elevated.

The majority of men with trichomoniasis are asymptomatic, but the parasite is responsible for a small but increasing proportion of cases of non-gonococcal urethritis.

Pathology

In women, *T. vaginalis* may be found in the vagina and the exterior cervix in over 95 per cent of infections, but is recovered from the endocervix in only 13 per cent. The urethra and Skene's glands are also commonly infected. In men the urethra is the most common site of infection, but the organism has also been recovered from epididymal aspirates. Dissemination beyond the lower urogenital tract is extremely rare even in severely immunocompromised patients.

Previously regarded as unpleasant but harmless, epidemiological studies have recently linked trichomoniasis in women with a modest increase in the risk of heterosexual HIV transmission and with adverse pregnancy outcome and have suggested that it might cause a few per cent of cervical neoplasias.

Laboratory diagnosis

The symptoms and signs are not sufficient to establish the diagnosis, which must be made by detecting the parasite. This is most frequently achieved by wet-film microscopic examination of vaginal (not endocervical) secretions, urethral scrapings, centrifuged urine sediment, or prostatic fluid. The specimen should be examined as soon as possible—a motile trichomonad is unmistakable. Sensitivity is moderate, 50 to 70 per cent in women but only 10 to 20 per cent in men.

Culture provides significantly greater sensitivity; media vary in efficiency but Diamond's TYM and the very convenient if rather expensive InPouch ® system are amongst the best.

Antigen detection, DNA probe, and polymerase chain reaction (**PCR**)-based tests have been developed; none has yet found widespread acceptance but the sensitivity

of the PCR-based methods offers new possibilities or making an accurate diagnosis on specimens obtained in less invasive ways including self-administered tampons.

Treatment

The 5-nitroimidazole drugs provided the first and so far only group of effective chemotherapeutic agents. Doses given here are for metronidazole and should be adjusted to give the equivalent amount of other compounds. Two regimens are used—the original one of 250 mg three times a day for 7 days, or a single 1.6 or 2 g dose. Cure rates in women are similar (about 95 per cent) with both regimens if male sexual partners are also treated, but appear to be lower with the single-dose regimen if they are not. Only the 7-day regimen has been extensively evaluated in males, where it is equally effective. Treatment failures with any of the 5-nitroimidazole drugs are rare, but a proportion is due to resistant parasites.

Further reading

Honigberg BM, ed. (1989). *Trichomonads parasitic in humans*. Springer-Verlag, New York.

Krieger JN (1995). Trichomoniasis in men: old issues and new data. *Sexually Transmitted Diseases* **22**, 83–96.

Petrin D *et al*. (1998). Clinical and microbiological aspects of *Trichomonas vaginalis*. *Clinical Microbiology Review* **11**, 300–17.

7.14.1Cutaneous filariasis

G. M. Burnham

Filarial infections of the skin and soft tissues

Filarial infections of man and animal are worldwide. Of the filarias which primarily affect the skin or subcutaneous tissues of man— *Onchocerca volvulus*, *Loa loa*, and *Mansonella streptocerca*— the burden imposed by *O. volvulus* is by far the greatest. *Loa loa* produces self-limited swellings of the extremities and the migrating adult worm may be seen subcutaneously. *Mansonella perstans* and *Mansonella ozzardi* cause minimal if any symptoms.

Onchocerciasis

Onchocerciasis, or river blindness, occurs in 34 countries in Africa, Latin America, and the Arabian Peninsula. An estimated 17.7 million people are infected, the vast majority in Africa. Infection has caused blindness in 270 000 and left another 500 000 with severe visual impairment. Besides eye changes, onchocerciasis has chronic systemic effects, causing extensive and disfiguring skin changes, musculoskeletal complaints, weight loss, changes to the immune system, and perhaps epilepsy and growth arrest as well. Of all the manifestations of onchocerciasis, skin lesions are the most common. These include acute and chronic itchy papular disease, and intensely pruritic lichenification. Lesions may be localized or widespread. In later stages, degenerative skin disease develops with a loss of elastic tissue, and extensive pigmentory changes.

The disease, endemic to some of the world's poorest areas, has great impact on the economic and social fabric of communities. A complex human–parasite tolerance allows people who host millions of parasites to continue daily existence. The discovery of ivermectin treatment has brought untold benefits to victims of the disease and to their communities.

Epidemiology

The microfilariae of *O. volvulus* were first observed by O'Neill in Ghana in 1875 in an intensely pruritic chronic skin condition called 'Craw-craw,' Leuckart described the adult worm 20 years later, and in 1923 Blacklock in Sierra Leone showed the blackfly, *Simulium damnosum*, to be the vector. Hissette in the Congo, and Robles in Guatemala linked blindness with onchocerciasis. Long before, Ghanians along the Red Volta river had associated the biting flies with skin lesions and blindness.

Vector control has now interrupted onchocerciasis transmission in the Volta river basin of West Africa, leaving the largest numbers of infected people in Nigeria, Cameroon, Chad, Ethiopia, Uganda, and the Congo. Most African foci are fairly stable, but in South America, foci continue to enlarge and new ones are found. Within foci, the disease may occur unevenly due to differences in both distribution of flies and exposure to bites. In the Americas, onchocerciasis is most common in the highland areas of Guatemala. Other countries with disease foci are Mexico, Venezuela, Colombia, Brazil, and Ecuador

In Africa, blindness was noted to be more common in savannah and woodland than rain forest areas, but people in forest areas had more depigmented skin disease. Parasite DNA probes have shown the existence of different strains or forms of the parasite, particularly in West Africa, although migration may now be blurring this geographical distribution. Other factors such as population density, genetic factors, transmission patterns, and perhaps nutrition may contribute to the risk of blindness. Onchocercal skin disease may reduce marital prospects (and dowry size), disrupt social relationships, and decrease the productivity of agricultural workers.

Experimental studies suggested considerable variation in the efficiency with which sibling species of *Simulium* flies transmitted forest and savannah strains of the parasite. This has given rise to the concept of vector–parasite complexes in which forest strains of parasites are preferentially transmitted by forest sibling species of flies and savannah strains by savannah sibling species. However, recent studies using polymerase chain reaction (**PCR**)-amplified *O. volvulus* larval DNA have questioned the importance of transmission complexes.

Parasitology

Larvae of *O. volvulus* enter the human during the blood meal taken by an infected female *Simulium* fly. Within 1 to 3 months larvae develop into male or female adult worms within palpable nodules commonly located over bony prominences of the thorax, pelvic girdle, or the knees (Fig. 1 and Plate 1). Nodules may be found on the head, particularly among children. These average 3 cm in diameter and are easily palpable, but some are deep, particularly around the pelvis.

Fig. 1 A 3-cm subcutaneous nodule. (See also Plate 1.)

A female worm may release 1300 to 1900 microfilariae per day for 9 to 11 years. From the nodules, these microfilariae find their way mainly to the skin and eye. In the skin they are found predominantly in the lymphatics of the subepidermis. In the eye, most are in the anterior chamber, but they are also found in the retina and optic nerve. When an infected human is bitten, anticoagulants from the *Simulium* fly create a pool of blood from which blood and microfilariae are ingested. Within the fly, those microfilariae that survive moult twice over the following 6 to 12 days to become infective larvae.

Microfilariae are about 250 to 300 µm in length and may live up to 2 years. They move easily through the skin and connective tissue ordinarily remaining within lymphatic vessels and provoking little reaction while alive. They have been seen in blood, urine, cerebrospinal fluid, and internal organs. One hundred million or more microfilariae may be present in heavily infected people. While live microfilariae are tolerated by their human hosts, dead and dying microfilariae may evoke intense inflammatory reactions which are responsible for the eye and skin damage. Tolerance of microfilariae may be regulated by MHC-encoded molecules.

Important species of *Simulium* are really complexes made up of sibling species, identifiable through banding patterns of their larval chromosomes. In Africa the main vectors are members of the *S. damnosum* complex or *sensu lato* (*s.l.*), which can fly long distances. The vector in areas of Uganda, Tanzania, Ethiopia, and the Congo are members of the *S. neavei* complex. In the Americas complexes of *S. ochraceum*, *S. metallicum*, and *S. exiguum* are the principal vectors, and these cover shorter distances. Some *Simulium* will bite humans almost exclusively while others are to varying degrees zoophilic.

Simulium develop in water courses varying from broad rivers to small streams, depending on the individual sibling species. Rapid flowing water provides the oxygenation needed for development of the immature stages. Most larvae and pupae develop on rocks or vegetation just below the water surface, but those of *S. neavei* develop on amphibious *Potamonautes* crabs. During this development period, the larvae are susceptible to insecticides.

Clinical features

Manifestations of onchocerciasis are almost entirely due to localized host inflammatory responses to dead or dying microfilariae. In a heavily infected person, 100 000 or more microfilariae die every day. The predominant immune response in onchocerciasis is antibody mediated, but with an important cellular component. Inflammatory responses may vary considerably between groups of people depending on length of exposure to antigens and the down-regulating activities by the host's immune system.

Eosinophils play an important role in the inflammatory responses. Cellular proteins derived from eosinophils are deposited on connective tissues throughout the dermis and are attached to elastic fibres causing skin changes.

In the eye, eosinophils are present in the anterior segment but lymphocytes and macrophages are more numerous. There is an activation of vascular endothelium, pericytes, and fibroblasts in people with chronic eye changes. Autoantibodies have been found to cells in the inner retina and to the retinal photoreceptors. The roll of these antibodies in causing retinal damage is uncertain. There is extensive evidence for a down-regulating of the immune response in chronically infected eye tissue by suppressor T cells and lymphocytes secreting interleukin 4.

Adult worms elaborate substances that inhibit the host's normal immune response. Exposure to filarial antigens *in utero* and through breast milk may induce an immune tolerance in residents of endemic areas. This could explain the difference in the disease patterns seen in people from non-endemic areas who become infected.

Among those coinfected with HIV, there is a lessened reactivity to *O. volvulus* antigens, but no difference in adverse reactions following ivermectin treatment.

Eye damage

The risks of visual impairment increase as prevalence and intensity of infection rises in a community. Microfilariae enter the cornea from the skin and conjunctiva, and a punctate keratitis develops around dead microfilariae which clears when inflammation settles. In those exposed to years of heavy infection, sclerosing keratitis and iridocyclitis are likely to develop, causing permanent visual impairment or blindness (Fig. 2).

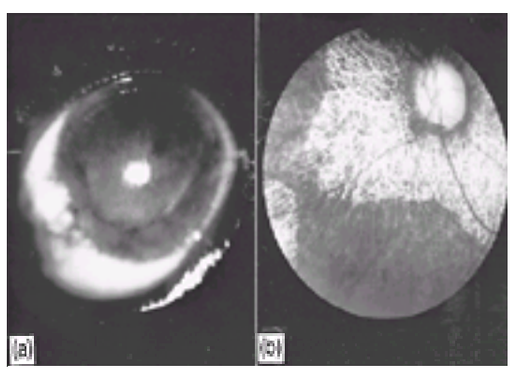

Fig. 2 (a) Sclerosing keratitis in a distorted eccentric pupil from anterior uveitis in a person blind from onchocerciasis. (b)Onchocerciasis producing a Hissette–Ridley fundus and optic atrophy in a person with central keyhole vision remaining.

The first sign of sclerosing keratitis appears as a haziness at the medial and lateral margins of the cornea. This is followed by a migration of pigment on to the cornea accompanied by a progressive ingrowth of vessels. Gradually the cornea becomes opacified. The central and superior areas are the last involved. Although eye lesions can be found wherever onchocerciasis occurs, in West Africa blindness is most common in savannah areas. Before control efforts began in Burkina Faso, 46 per cent of men and 35 per cent of women would eventually become blind.

Posterior segment lesions, which can coexist with anterior eye lesions, may be caused by inflammation around microfilariae entering the retina along the posterior ciliary vessels. Choroidoretinal lesions are commonly seen at the outer side of the macula or encircling the optic disc. Active optic neuritis is reported as an important cause of blindness in Nigeria. Optic atrophy has been reported to be present in 1 to 4 per cent of people with onchocerciasis in Cameroon and 6 to 9 per cent in northern Nigeria. Loss of peripheral vision is well recognized in onchocerciasis.

Skin disease

Of all the consequences of onchocerciasis, skin lesions are the most pervasive. Surveys of seven endemic sites in five African countries reported that between 40 and 50 per cent of adults had troublesome itching, which in some cases was so intense that people slept on their elbows and knees to minimize this symptom.

In its mildest form, onchocerciasis presents as itching with a localized maculopapular rash. These reactive lesions and itching may be evanescent, clearing completely without treatment in a few months. In other instances the papular lesions may become chronic and generalized, and accompanied by severe itching (Fig. 3 and Plate 2). Oedema and excoriations can be associated, and lesions may heal with hyperpigmentation. Particularly distressing are lichenified, hyperkeratotic lesions which may be widespread, and intensely itchy (Fig. 4 and Plate 3). A localized form of chronic papular dermatitis, often confined to one extremity, is known as *Sowdah*,

Arabic for dark. In this condition, first described from Yemen, there is an exceptionally strong IgG antibody response.

Fig. 3 Excoriated papular lesions of onchocerciasis with hyperpigmentation. (See also Plate 2.)

Fig. 4 Lichenified skin lesions with atrophy. (See also Plate 3.)

Light-skinned expatriates infected while visiting an endemic area may present a year or later with intensely itchy and red macular or maculopapular lesions. These may be confined to one area of the body or be more generalized, and may be associated with fever, muscle, joint pain, and sometimes oedema. Rash may sometimes persist for several months following ivermectin treatment.

In endemic areas, degenerative skin changes may develop in some people with long-standing infection. Elastic fibres are destroyed leaving the skin thinned with a wrinkled cigarette-paper appearance. The atrophied skin begins to sag, the most extreme state being 'hanging groin' with its apron-like skin folds. Depigmentation of the pretibial areas, or 'leopard skin', is a characteristic finding in older people living in endemic areas (Fig. 5 and Plate 4).

Fig. 5 Depigmented 'leopard skin'. (See also Plate 4.)

Other conditions associated with onchocerciasis

Both men and women with onchocerciasis weigh less than an uninfected cohort, and report more musculoskeletal pains. Evidence from Uganda and Burundi have suggested a possible association between epilepsy and onchocerciasis.

A peculiar pattern of growth arrest beginning between the age of 6 to 10 years was reported from a Ugandan onchocerciasis focus near Jinja in 1951. This Naklanaga syndrome, as it was called, now seems to have disappeared from there following elimination of onchocerciasis, but has been noted in western Uganda, and perhaps in Burundi.

Diagnosis

Finding microfilariae in skin snips is the time-honoured, though not very sensitive, method of diagnosis. Microfilariae lie close to the surface and are most plentiful in the iliac crest area, except in Latin America where they are more common in the shoulder and scapular areas. Using either a scalpel blade or a sclerocorneal punch, four to six snips (about 5 mg each) are taken under sterile conditions and immersed in normal saline. Microfilariae swimming free of the skin fragments can be counted easily with a dissecting microscope at 24 h or sooner. Examination of excised onchocercal nodules shows sections of adult worms. Enzyme immunoassay and PCR diagnostic methods have a high degree of sensitivity and specificity. Eosinophilia is common in onchocerciasis.

The Mazzotti test, in which people with onchocerciasis react with itching and a skin rash to 50 mg of diethylcarbamazine (DEC or Banocide), is seldom needed for diagnosis and is dangerous in heavy infections.

For community assessment, the prevalence of nodules in 30 to 50 males over the age of 20 years multiplied by 1.5 gives the approximate community prevalence of onchocerciasis. Where the prevalence of nodules is over 40 per cent the risk of blinding disease is high.

Treatment

The introduction of ivermectin for onchocerciasis in 1987 was one of the milestones of tropical disease treatment. Symptoms of onchocerciasis can be controlled effectively in individuals in a clinic or through mass treatment of endemic communities.

Ivermectin is derived from *Streptomyces avermitilis*. A single dose of 150 µg/kg clears microfilariae from the skin for several months. Annual treatment controls microfilarial counts and prevents progression of clinical findings, though in some locations it is given twice yearly. Treatment can be repeated if itching returns before

the next dose is due. In the absence of reinfection, treatment should probably be continued for 10 or more years, or until adult worms stop producing microfilariae. In Ghana, after 5 years of annual ivermectin, the number of microfilariae was reduced to 7 per cent of the pretreatment baseline count.

Limiting the numbers of microfilariae through annual treatment improves early and advanced anterior-segment eye lesions, halts development of optic nerve disease, and improves severe onchocercal skin lesions. Adverse reactions to ivermectin commonly consist of increased itching, swelling of the face or extremities, and headache and body pains. Hypotension has been reported rarely after treatment in heavily infected people. Bullas have been seen occasionally. The most pronounced adverse reactions occur after the first ivermectin treatment, decreasing after subsequent treatment cycles. Ivermectin has no adverse effects in uninfected people. Although ivermectin temporarily reduces the release of microfilariae by adult worms, it does not destroy the adults. Care should be exercised in treating people coinfected with *Loa loa*, particularly those with counts above 10 000 microfilariae/ml blood, as potentially fatal central nervous system events can occur.

Ivermectin acts primarily on parasite neurotransmitters producing paralysis. This action appears to be mediated by potentiation or direct opening of glutamate-gated chloride channels. Although some ivermectin resistance has developed in animal parasites, no drug resistance has been reported in humans .

Prevention and control

Methods have included insecticides added to rivers to interrupt *Simulium* breeding, mass distribution of ivermectin, and nodulectomy in an attempt to prevent blindness.

Vector control

Killing *Simulium* larvae by adding DDT to rivers eliminated onchocerciasis in Kenya and the Mabari forest of Uganda. In 1974 the Onchocerciasis Control Programme (OCP) was formed to control *Simulium* through the larviciding of rivers in the Volta basin of West Africa with ecologically suitable compounds. This highly successful vector control programme, later supplemented with ivermectin distribution, has now permitted tens of millions of people to live free of disease. Mass distribution of ivermectin is now the principal method for onchocerciasis control, though vector control may still be appropriate in some locations.

Ivermectin mass distribution

After the effectiveness of ivermectin had been shown, its manufacturers, Merck and Co., established the Mectizan® Donation Program to provide the drug free 'for as long as necessary to as many as necessary'. By mid-1998 over 100 million ivermectin treatments had been given in 33 of 34 endemic countries.

The goal of a control programme may be either complete eradication of the parasite reservoir or elimination of the public health and socio-economic consequences of continuing infection. In Guatemala, where high population coverage with 6-monthly treatment has reduced parasite transmission by 80 to 100 per cent after 3 years, eradication may ultimately be possible, and this could be true elsewhere in Latin America where sustained treatment is implemented.

The Onchocerciasis Elimination Program in the Americas (OEPA) and the African Programme for Onchocerciasis Control (APOC) have been formed with support by the World Bank and other United Nations agencies to eliminate the public consequences of infection. These programmes focus on regular mass administration of ivermectin through community-based distributors and mobile teams.

Because of the lifespan of adult worms, ivermectin distribution programmes must be sustained for a period of 15 years or more. In some places, the duration may have to be longer because of the difficulty in achieving good coverage, often because of insecurity.

Nodulectomy

A third form of onchocerciasis control has been the nodulectomy programmes of Mexico and Guatemala. For many years health workers have moved from village to village removing nodules, especially around the head. The evidence for this preventing blindness is not strong.

Areas needing further research

Although ivermectin brings great relief to the individual, and has a clear impact on the disease in mass distribution programmes, its does not kill adult worms. While symptoms and risks are controlled through annual treatment, the disease itself is not eradicated. A number of macrofilaricidal drugs, capable of killing adult worms, have been tested, but none has so far proved suitable for either individual or mass treatment. Diagnostic methods, although dramatically improved in recent years, are still not in a form suitable for practitioners in developing countries. Our basic knowledge of *O. volvulus* and the disease it causes still contains many gaps. These include a fuller understanding of the parasite and its relationship with the host, the nature of the systemic effects of *O. volvulus* infection, and better knowledge of the natural history of a disease which continues to affect millions worldwide.

Loa loa

Introduction

Loa loa is a filaria transmitted by the *Chrysops* fly in West Africa. The adult worm migrates beneath the skin, and sometimes across the eye, moving at about 1 cm/min. Periodically the infection causes sudden but transient localized inflammatory oedema known as Calabar swellings.

Parasitology

Larvae of *L. loa* burrow into the human skin during feeding of the *Chrysops* or 'mangrove fly' (*C. silacea* or *C. dimidiata*). In humans the parasites mature and live in the fascial layers. After a year or more, microfilariae are produced. Microfilariae are present in the blood during the day, when the *Chrysops* fly bites. Once taken up by the fly, microfilariae go through developmental stages in the fly's thoracic muscles. After 10 days the fly is able to infect a human, and can do so for another 5 days.

Epidemiology

Infection is most common around the Gulf of Guinea, particularly in Nigeria and Cameroon, but extends through Central Africa into Sudan, and Uganda, and south to Angola and the Congo (Fig. 6). Man is the only host, although a similar parasite is found in monkeys in the same areas. The fly lives in the rain forest canopy, and descends to bite humans, attracted perhaps by movement. Transmission may be most intense during the rainy season when flies are breeding on the muddy banks of forest streams.

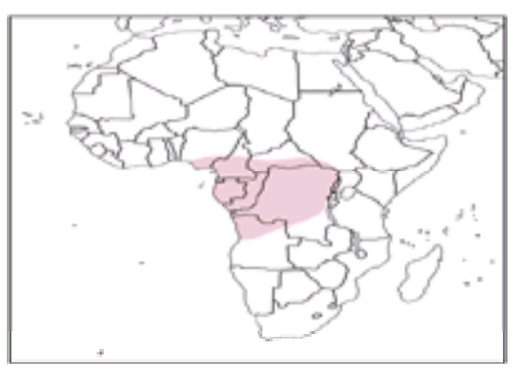

Fig. 6 Map of the approximate distribution of *Loa loa.*

Clinical features

The first clinical symptoms of loiasis may be delayed for several years after infection. Calabar swellings appear suddenly, most commonly in the forearms or wrists, and sometimes following heavy exercise. These oedematous lesions are red and itchy, and may be associated with fever and irritability. After a few hours, or 1 to 2 days at most, the affected part returns to normal. Swellings are not confined to the arms, but may be present in the face, breasts, or legs. They appear more commonly in hot seasons.

Calabar swellings are a hypersensitivity reaction to worm antigens which may be released in the process of migration or perhaps during the maturation of the worm. A high proportion of eosinophils are seen in peripheral blood smears, often exceeding 70 per cent.

A second common feature is the appearance of a migrating worm (Fig. 7 and Plate 5). This may be under the skin in any location, but is most dramatic when it crosses the eye ('eye worm', Fig. 8). Other than local irritation of the conjunctiva while the worm is passing, and the obvious concern of the host, there are no serious consequences.

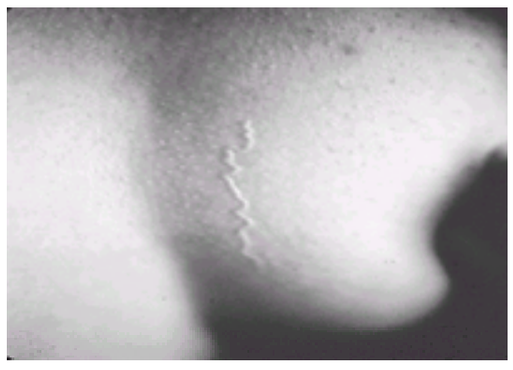

Fig. 7 Migrating *Loa loa. (See also Plate 5.)*

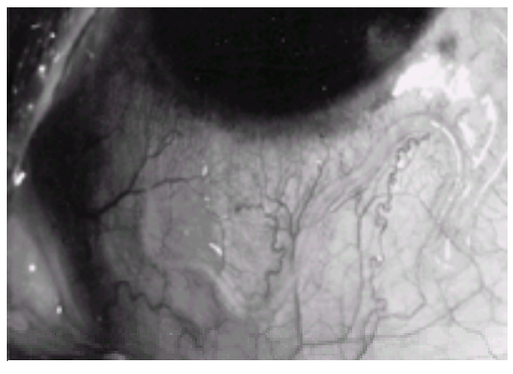

Fig. 8 *Loa loa* crossing the bulbar conjunctiva.

Rare but potentially serious consequences of *L. loa* are meningioencephalitis, renal disease, and endomyocardial fibrosis. The meningioencephalitis may occur spontaneously, though usually after treatment with diethycarbamazine or ivermectin. Recovery is common following supportive treatment, although fatalities have been reported. Those at most risk have microfilarial counts above 10 000/ml of blood. The renal and endocardial complications of loiasis may have an immune origin.

Laboratory diagnosis

Diagnosis has traditionally been by the finding of microfilariae in a daytime blood sample, or by a history or typical clinical findings. Use of more sensitive PCR methods has shown that many, even perhaps the majority of those infected, do not have microfilariae in their peripheral blood.

Treatment

The standard treatment has been diethycarbamazine (DEC), which kills microfilariae and many adult worms. The treatment is given as 50 mg on the first day, and the dose doubled each subsequent day until 2 to 3 mg/kg is reached (maximum 600 mg). This is then continued for up to 21 days. During treatment, fever, arthralgias, and itching can occur. Ivermectin at 200 µg/kg dramatically decreases the number of microfilariae and decreases some of the loiasis symptoms. As with diethycarbamazine, there is a risk of potentially fatal meningioencephalitis in those with high microfilarial counts. It might be prudent to initiate any ivermectin treatment at half dose, particularly in those with higher (more than 10 000 microfilariae/ml) parasite counts. Since many people with loiasis also have onchocerciasis, careful monitoring for severe eye and skin inflammation is important when giving diethycarbamazine. Treatment is unlikely to eradicate all adult worms, and in endemic areas reinfection is probable. Blood films for microfilariae or PCR examinations should be followed to indicate the need for retreatment.

Prevention

The best prevention is avoiding *Chrysops* fly bites. Having window screens on dwellings, wearing clothing to protect legs and forearms, and avoiding high biting areas can reduce risks.

The Mansonellas

Introduction

The mansonellas are a group of filarial infections common to many countries, and are of negligible clinical importance under most circumstances. Infection is transmitted by *Culicoides* midges.

Epidemiology

Mansonella (formerly *Dipetalonema*) *perstans* is found in much of tropical Africa as well as Trinidad and several parts of South America. Adult worms live free in the abdominal cavity, and microfilariae are found in the blood. *Mansonella ozzardi* is found in the West Indies and Central and South America. Microfilariae are found in the blood and skin. Adult worms have been found in the peritoneal cavity. In addition to *Culicoides*, *Simulium* flies have been reported to transmit *M. ozzardi* in the Amazon basis. *Mansonella* (formerly *Dipetalonema*) *streptocerca* is a common infection in West and Central Africa extending into western Uganda. Both microfilariae and adult worms are found in the skin, but without the nodules seen in onchocerciasis. Unless *M. streptocerca* microfilariae are differentiated parasitologically from those of *O. volvulus*, inappropriate mass treatment programmes for onchocerciasis could be implemented.

Clinical manifestions

Of the mansonellas, only *M. streptocerca* produces clear-cut symptoms, although even these can be confused with those of *O. volvulus* which may be a coinfection. Chronic papular lesions are commonly present, often associated with postinflammatory hyperpigmentation. Lichenification may occur less commonly. Hypopigmentation has been noted in areas of skin overlying the location of adult worms in the skin. In general these findings are not easily distinguishable from those of onchocerciasis.

M. perstans has been reported to produce Calabar-like swellings, and in Zimbabwe, central nervous system symptoms. *M. ozzardi* infections are generally without symptoms, though fever, arthralgias, headache, and itching have been associated in the Amazon area.

Diagnosis

A diagnosis is made by the finding of characteristic microfilariae in the blood or the skin. The microfilaria has a distinctive 'walking stick' shape to its tail, and four prominent nuclei in the tail, both of which distinguish it from the microfilaria of *O. volvulus.* Recently a PCR assay has been described for *M. streptocerca* and both QBC-fluorescence and ELISA methods for *M. perstans.* Eosinophilia is a characteristic finding.

Treatment

In asymptomatic persons no treatment is required. *M. streptocerca* responds well to ivermectin, often with mild reactions similar to those seen in onchocerciasis. Treatments of *M. perstans* with diethycarbamazine, and albendazole, have all been disappointing, though mebendazole given as 100 mg once or twice daily for 28 to 45 days has been reported to clear microfilariae. Ivermectin was able to lower microfilarial counts to 60 per cent of pretreatment values.

Further research

Little attention has been given to the mansonellas, ubiquitous in many places. A reliable, inexpensive field test kit for mass screening could help determine the extent of infection and any association with the ill-defined clinical symptoms often reported.

Further reading

Mectizan and onchocerciasis: a decade of accomplishment (1998). *Annals of Tropical Medicine and Parasitology* **92**(Suppl.), S1–174.

Alley ES *et al.*(1994). The impact of five years of annual ivermectin treatment on skin microfilarial loads in the onchocerciasis focus of Asubende, Ghana. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **88**, 581–84.

Brieger WR *et al*. (1998). The effects of ivermectin on onchocercal skin disease and severe itching: results of a multicentre trial. *Tropical Medicine and International Health* **3**, 951–61.

Chan CC *et al*.(1989). Immunopathology of ocular onchocerciasis. I. Inflammatory cells infiltrating the anterior segment. *Clinical Experimental Immunology* **77**, 367–73.

Cooper PJ *et al*. (1999). Eosinophil sequestration and activation are associated with the onset and severity of systemic adverse reactions following the treatment of onchocerciasis with ivermectin. *Journal of Infectious Diesases* **179**, 738–42.

Fischer P, Bamuhiiga J, Büttner DW (1997). Occurrence and diagnosis of *Mansonella streptocerca* in Uganda. *Acta Tropica* **63**, 43–55.

Garcia A *et al*. (1995). Longitudinal survey of *Loa loa* filariasis in southern Cameroon. *American Journal of Tropical Medicine and Hygiene* **52**, 370–5.

Mudroch ME *et al*. (1997). HKA-DQ alleles associate with cutaneous features of onchocerciasis. *Human immunology* **55**, 46–52.

Ottesen EA (1995). Immune responsiveness and the pathogenesis of human onchocerciasis. *Journal of Infectious Diseases* **171**, 659–71.

World Health Organization (1995). *Onchocerciasis and its control*. Geneva

Yameogo L *et al*. (1999). Pool screen polymerases chain reaction for estimating the prevalence of *Onchocerca volvulus* infection in *Simulium damnosum sensu lato*: results of a field trial in an area subject to successful vector control. *American Journal of Tropical Medicine and Hygiene* **60**, 124–8.

7.14.2 Lymphatic filariasis

R. Knight

Wuchereria bancrofti, *Brugia malayi*, and *Brugia timori* are mosquito-borne nematodes. They are important causes of morbidity in the tropics and subtropics between latitudes 41°N and 28°S in the Old World and 30°N and 30°S in the Americas (Fig. 1). Bancroftian filariasis due to *W. bancrofti* infects 110 million people; it was introduced into the Americas from Africa by the Atlantic slave trade. The two *Brugia* species infect about 13 million people in South and South-East Asia. Approximately 700 million people live in countries where these infections are endemic. *Brugia timori*, which was first described in 1964, has a very localized distribution but causes severe disease.

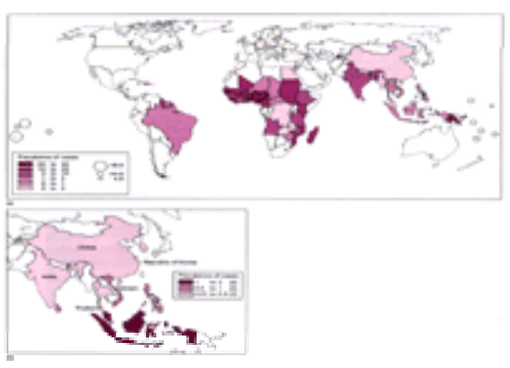

Fig. 1 Distribution of lymphatic filariasis; case prevalences (percentages) due to (a) *Wuchereria bancrofti* and (b) *Brugia* spp. The figures in parentheses indicate the number of countries. Circles denote Pacific Island prevalences. (By courtesy of E.Michael and D.A.P. Bundy.)

Aetiology—the biology of the parasite

Adult worms live in the larger lymphatic vessels and lymph nodes. They are smooth, creamy-white, and threadlike; females measure 80 to 100 mm in length, and males 40 mm; their lifespan is normally 2 to 5 years, but exceptionally much more. Mated females produce numerous microfilariae throughout their lives; these actively motile embryonic worms are sheathed by the remnants of the egg shell. They are 180 to 290 µm in length and 7 to 10 µm in diameter. Different species can be distinguished morphologically in stained films. Microfilariae migrate via the lymphatic system to the blood where they have a lifespan of up to 12 months (Plate 1). Their numbers in the peripheral blood vary during the day and night—a phenomenon known as periodicity; when not circulating they are sequestered in lung and reticuloendothelial capillaries. Maximal counts in the blood coincide with the biting cycle of the vector. The species and strain of parasite determine the periodicity; most common is nocturnally periodic with maximal counts between 22.00 and 02.00 h and virtual absence during the day. Alternatively, microfilariae may be present throughout the 24 h cycle with prominent peaks during the day or the night: diurnally and nocturnally subperiodic, respectively.

After uptake by the vector, microfilariae penetrate the gut and migrate to thoracic muscles where they mature over 9 to 15 days to infective third-stage larvae, which then migrate to the mosquito head and escape from the proboscis during a blood meal. Larval worms enter the puncture wound made by the vector, reach the peripheral lymphatic system, and move to larger lymph vessels below a lymph node. Sexual maturity and appearance of microfilariae in the blood usually takes 8 to 18 months, but sometimes only 3 months.

Epidemiology and transmission

In endemic areas microfilarial prevalence rates increase steadily from early childhood to reach a maximum in early adult life, when in highly endemic areas 10 to 30 per cent prevalences are not unusual; rates in males are generally higher, perhaps due to greater vector exposure. The cord blood of some infants shows microfilariae.

In some locations *Brugia malayi* is a zoonosis with an animal reservoir; elsewhere it is an anthroponosis with only a human source of infection.

Geographical distribution and mosquito vectors

W. bancrofti infection

Culex transmission

This vector breeds mostly in organically polluted water, usually in urban and suburban areas but also villages when there are suitable latrine and cesspit habitats. This is the commonest type of transmission and is increasing with urbanization; it occurs in India, Sri Lanka, Central and South America, some Caribbean Islands, urban and coastal villages in East Africa, Egypt, and parts of China. *Culex* bites at night, mostly on the legs, the microfilariae are nocturnally periodic. *Culex* is the most efficient vector and can maintain transmission at low microfilarial densities, making control difficult.

Anopheles transmission

The same vector species commonly transmit both filariasis and malaria. This occurs in East and West Africa, Papua New Guinea and Vanuatu, limited areas in South America, and parts of China. *Anopheles* bites nocturnally, mainly on the legs; microfilariae are nocturnally periodic.

Aedes transmission

This is limited to Southern Oceania, especially Fiji, Samoa, Tonga, the Cook Islands, and New Caledonia, but also patchily in Thailand, the Philippines, Vietnam, and the Nicobar Islands. *Aedes* feeds throughout the 24 h cycle with a daytime peak, and bites all over the body; the microfilariae are diurnally subperiodic.

B. malayi infection

Zoonotic *Mansonia* transmission in swamp forests

This occurs in Malaysia, Indonesia, and southern Thailand; monkeys and carnivores are reservoir hosts. *Mansonia* bites mainly by night but also by day, usually on legs below the knee; the microfilariae are nocturnally subperiodic.

Transmission in agricultural areas

In parts of Malaysia, Buru in Indonesia, and southern Thailand a mixed anthroponosis and zoonosis occurs in transitional zones with monkeys and cats as reservoirs, and both *Anopheles* and *Mansonia* as vectors. Microfilariae have periodicities intermediate between nocturnally periodic and nocturnally subperiodic.

In India (mainly Kerala), Malaysia, Sulawesi, southern Thailand, Vietnam, China, and Korea infection involves humans only with *Anopheles* as the main vector and *Mansonia* as the accessory vector; the microfilariae are nocturnally periodic.

B. timori infection

This is confined to Timor and other islands in the Lesser Sundas group in eastern Indonesia. *Anopheles barbirostris* is the vector. The microfilariae are nocturnally periodic.

Pathogenesis

Local immunological reactions to worm antigens provoke acute and subacute responses with oedema of lymphatic tissue and infiltration with eosinophils and monocytes. Antigens derive from moulting fluids of developing worms, excretory products, microfilariae trapped within the lymphatic system, and also dying worms including those killed by chemotherapy. Dead and disintegrating worms become surrounded by granulation tissue with giant cells and epithelioid cells. Stenosis or blockage of lymph vessels leads to distal dilatation with varicosities and valve incompetence. Prolonged or recurrent lymph stasis leads to accumulation of protein-rich interstitial fluid and fibroblast proliferation, dilated dermal lymphatics, and epithelial acanthosis and hyperkeratosis.

Determinants of pathology include duration of exposure, intensity of transmission, anatomical sites of infective mosquito bites, and the species and strain of parasite. Prenatal exposure to filarial antigen is of great importance and induces immunological tolerance. Residents in high-transmission areas often show patent microfilariaemia but little immunopathology. However, in many adults a later decline in microfilarial prevalence parallels increased host immunological reactivity and pathology. New residents and visitors show marked local reactivity to worms and often no blood microfilariae; the latter situation was well documented among American troops in the Pacific in 1942 to 1944 and French troops in former Indochina.

Clinical manifestations

Acute lymphatic filariasis

In endemic areas acute episodes are recurrent from the age of 10 years and most frequent 4 to 8 months after the peak of seasonal transmission. Episodes last several days or weeks; fever and malaise are common but blood eosinophilia is not marked. People leaving endemic areas cease to have acute episodes after 1 year although they may experience recurrent pain in previously affected tissues, especially after unusual exercise.

Filarial lymphadenitis and lymphangitis

Tender lymphadenopathy is most common in the inguinal and femoral nodes, but axillary and epitrochlear nodes are also affected. Tender retrograde lymphangitis typically spreads peripherally below the node.

Acute genital filariasis

This is uncommon in boys before puberty but common thereafter. The typical lesion is funiculitis with a tender fusiform or cylindrical swelling of the spermatic cord; epididymitis and orchitis are less common.

Filarial abscess and filarial fever

Affected nodes in the groin or elsewhere may break down producing an open ulcer that heals slowly leaving characteristic scars. Pelvic and retroperitoneal lymphadenitis can produce a febrile illness that is difficult to diagnose.

Chronic lymphatic filariasis

Lymphoedema and elephantiasis

Initially, transient pitting oedema occurs during inflammatory episodes in proximal nodes. Later, oedema persists between episodes becoming non-pitting distally. Eventually, brawny non-pitting oedema becomes permanent (Fig. 2). In patients with leg involvement epidermal thickening, papillomatosis, and fissuring are common (Fig. 3), and bacterial infection becomes an important complication.

Fig. 2 Chronic elephantiasis in a man in Belém, northern Brazil. Note the scars of unsuccessful surgery. (Copyright Pedro Pardal.)

Fig. 3 Chronic elephantiasis with epidermal thickening, fissuring, and papillomatosis in a man in north-east Nigeria. (Copyright D.A. Warrell.)

Chronic genital filariasis

Hydrocele is the commonest lesion and prevalence rates may reach 30 per cent in men over 35 years in highly endemic areas; many patients give a history of preceding episodes of funiculitis or epididymitis. The tunica vaginalis is often thickened. Nodular lesions of the spermatic cord and epididymis are common and the testis itself becomes enlarged and indurated. Lymphoceles occur on the cord. Dilated dermal lymphatics in the scrotal wall associated with atrophic epidermis produce lymph scrotum, the skin having a velvety appearance. Rupture of these lymphatics leads to weeping skin lesions and often secondary infection, occasionally complicated by Fournier's gangrene.

Lymphoedema of the scrotum is a late sequel (Fig. 4), often the testes are unaffected; penile lesions are rare. Vulval lymphoedema is under-recognized; it is associaeted with dilated retroperitoneal lymphatics and must be distinguished from lymphogranuloma venereum.

Fig. 4 Gross hydrocele in a patient with chronic filariasis. (By courtesy of the late P.E.C. Manson-Bahr.)

Chronic lymphadenitis and lymphangitis

Recurrent episodes of acute inflammation lead to persisting and sometimes massive lymph node enlargement. Thickened lymphatic cords may be palpable connecting the axillary and epitrochlear, or the femoral and popliteal nodes. Varicose lymph vessels may be visible in these areas. Lymph varices are fluctuant sacs of lymphatic tissue derived usually from the capsule of a node, hence the alternative term lymphadenocele. They partially empty when the part is raised; aspiration reveals lymph or occasionally chyle. They occur in the medial thigh, groin, axilla, and sometimes even the neck.

Chyluria and lymphuria

Dilated pelvic and retroperitoneal lymphatics may rupture into the urinary tract in the renal pelvis, ureter, or bladder. When there is lymph stasis above the cisterna chyli then small bowel chyle may reflux into the urine postprandially. Chyluria is often intermittent and blood stained (Fig. 5). Continued loss of protein and lipids in the urine may lead to weight loss and cachexia. Chyluria may eventually be self-limited.

Fig. 5 Chyluria and haematuria in a patient with chronic filariasis. (By courtesy of the late P.E.C. Manson-Bahr).

Non-lymphatic pathology

Tropical pulmonary eosinophilia

This presents as a subacute or chronic illness with cough, wheezing, and reticular or miliary pulmonary shadowing. Microfilariae are absent from the blood, but eosinophilia is marked and titres of filarial antibody are very high. Some patients have features of lymphadenopathic or genital filariasis, but many do not. Lung functional loss is restrictive. Response to antifilarial treatment is good but untreated the condition leads to pulmonary fibrosis and pulmonary hypertension. The syndrome is due to a heightened immunological response to dead microfilariae which may be found, in biopsies of lung and other tissue, surrounded by eosinophilic microabscesses. It occurs in most endemic areas, but is rare in Africa; it is commoner in men and rare in children; many patients are not long-term residents.

Filarial arthritis

Joint involvement is subacute and often recurrent with effusion; it usually affects the knee.

Filarial glomerulonephritis

The incidence of clinically significant disease is uncertain; it results from immune complex deposition on the glomerular basement membrane. Recurrent streptococcal infection associated with filarial lymphoedema is also implicated.

Diagnosis

Clinical

Many patients will have several clinical features that, together with history of preceding acute episodes, will be strongly suggestive diagnostically: manifestations such as varicose lymphatics, lymphadenocele, retrograde lymphangitis, and lymph scrotum are highly specific to filariasis. Genital lesions are rare in *Brugia* infections, which usually present with lymphoedema below the knee. In *B. timori* infections lymph node pathology in the legs is often severe, sometimes with skin ulceration. Upper limb and breast lesions are common in diurnally subperiodic *W. bancrofti* infections in the Pacific; but they do occur elsewhere with other strains of this parasite.

Parasitological

Microfilariae are typically found in blood films but also in aspirates from a lymph varix, hydrocele, lymphocele of the cord, or in urine. Blood should be taken to coincide with the expected microfilarial periodicity. Measured 10 or 20 µl volumes are used to prepare thick blood films stained by Giemsa. Counting chambers taking 100 µl of lysed blood can be used or larger volumes may be lysed and the spun deposit examined. Alternatively, 1 ml of lysed or unlysed blood is passed through a Millepore filter; the filter is then stained. Nocturnally periodic *W. bancrofti* microfilariae appear transiently in the blood 30 to 60 min after a 100 mg dose of diethylcarbamazine and this forms the basis of the 'provocation test'. Stained microfilariae can be identified by their sheaths, but these may be lost by *Brugia* parasites during staining; *B. timori* has distinctive sheath staining. The arrangement of nuclei at the caudal end allows species diagnosis; *B. malayi* has two subterminal nuclei separated by a space. The microfilariae of *Loa loa* also have sheaths and must be distinguished from those of species causing lymphatic filariasis.

Immunodiagnosis

Positive skin tests and filarial antibody are common in those exposed to infection and may be of value in visitors to an endemic area. Several tests for filarial antigen in serum are now available and a positive test indicates persisting adult worms. Antigen may be present in the absence of microfilariaemia.

Imaging of lymphatic vessels

Lymphangiography will delineate anatomical details of abnormal lymphatic tissues such as lymph varices and lymphatic connections to the urinary tract in chyluria. They are not usually diagnostic for filariasis. Scrotal ultrasound can show live worms—the 'filarial dance' sign.

Lymphoscintigraphy using technetium-labelled dextran or albumin is a less invasive and useful technique that can demonstrate lymphatic pathology. Abnormal dermal lymphatics occur in many asymptomatic infected persons in endemic areas but, so far, few local control subjects have been examined and comparisons with normal lymphatic studies in Western countries may not be justified.

Treatment

Individual chemotherapy

Diethylcarbamazine remains the treatment of choice. Adequate dosage will kill adult worms. Even a small single dose will clear blood microfilariae temporarily. Sensitivity reactions to filarial antigen, both local and systemic, are common in infected people and simulate some of the acute manifestations of the infection; they necessitate care and supervision in the initial stages, especially in *Brugia* infections. Treatment should be started at 1 mg/kg on the first day, increasing over 3 or more days to 6 mg/kg in divided doses; this dose then being continued for 21 days. Coinfection with *Loa loa* and *Onchocerca volvulus* must be excluded before diethylcarbamazine is given to avoid dangerous reactions.

Indications for curative treatment are acute manifestations with or without microfilariaemia, and chronic disease in patients who are either microfilaria positive or positive for filarial antigen identified serologically. Treatment often reduces the size of hydroceles but has little effect on chronic lymphoedema.

Surgical and supportive treatment

Acute manifestations of filariasis can mimic strangulated hernia and testicular torsion. Surgical treatment of filarial hydrocele is the same as that for non-filarial disease. Scrotal lymphoedema can be treated surgically, usually with preservation of the testes. Lymphosaphenous anastomosis is being used for leg elephantiasis; many other procedures have been used in the past, often with disappointing results (Fig. 2).

Bacterial infection is common in those with lymphoedema, especially when the skin is fissured, breached in an interdigital cleft, or when there is minor injury, ulcer, or insect bite. Early use of antibiotics and resting of the affected limb lessens the risk of increasing lymphoedema; supportive bandaging applied each morning or wearing elastic stockings reduces chronic oedema.

Filariasis at the community level

Surveys for lymphatic filariasis

These are carried out to assess the importance to public health and plan intervention programmes. Current clinical features together with history of acute features in the preceding 6 months are documented, and blood is taken to measure microfilarial density. Serum collected on such surveys has been the source of many immunopathological studies. The following disease groups are recognized, but the availabilty of tests for filarial antigen will add a new dimension to such surveys.

1. Asymptomatic without microfilariaemia—in highly endemic areas most of these people will have been exposed to infection; they are sometimes called 'endemic normals';
2. asymptomatic with microfilariaemia;
3. acute filariasis—many will show microfilariaemia, but this is absent in prepatent infections and in people with strong immunological responses, including visitors; and
4. chronic filariasis—in some geographical areas many subjects will be microfilaria negative, especially those with chronic lymphoedema; in other areas they are positive; negativity may be due to a decline in transmission over several years or to host immune responses.

Social and economic consequences of lymphatic filariasis

Surgical care of patients with hydrocele and other manifestations places a great burden on health care in highly endemic areas. In agricultural communities acute manifestations and episodes of secondary bacterial infection impair productivity. Social stigma is a major problem and may lead to divorce or make a woman unable to marry.

Vector control

These campaigns are targeted at the local vector. Larval *Aedes* breeding sites such a discarded tins, tyres, or coconut shells can be removed. *Culex* numbers can be reduced by improved sanitation, larvicides, and polystyrene beads applied to the water surface of latrines and cesspits. Bednets and repellants are universally applicable. Where *Anopheles* is the vector, malaria control can interrupt filariasis transmission as in Samoa, Vanuatu, and parts of southern China.

Population-based chemotherapy

Different dosage regimens of diethylcarbamazine have been used in many endemic areas; with annual or 6-monthly administration either to the whole population or to those found to be infected; medicated salt is an alternative. The main aim is to eliminate microfilariaemia and hence transmission; however, with repeated and higher doses many adult worms are eventually killed. Ivermectin offers an alternative method of reducing microfilaraemia. A single dose of 6 mg/kg of diethylcarbamazine is as effective as one 200 or 400 µg/kg dose of ivermectin. Both will reduce microfilariaemia to almost nil for 6 or 12 months. Sensitivity reactions are much commoner with diethylcarbamazine. Annual dosage with both of these drugs continued for 4 or 6 years—the lifespan of adult worms—should interrupt transmission. Albendazole is also effective as a microfilaricide and has some activity against adult worms; a 600 mg dose can replace either diethylcarbamazine or ivermectin in a two-drug annual regimen; diethylcarbamazine must not be used where onchocerciasis is co-endemic.

Further reading

Dreyer G *et al*. (1999). Acute attacks in the extremities of persons living in an area endemic for bancroftian filariasis: differentiation of two syndromes. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **93**, 413–7.

Freedman DO (1998). Immune dynamics in the pathogenesis of human lymphatic filariasis. *Parasitology Today* **14**, 229–34.

Freedman DO *et al*. (1994). Lymphoscintographic analysis of lymphatic abnormalities in symptomatic and asymptomatic human filariasis. *Journal of Infectious Diseases* **170**, 927–33.

Ismail MM *et al*. (1998). Efficacy of single dose combinations of albendazole, ivermectin and diethylcarbamazine for the treatment of bancroftian filariasis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92**, 94–7.

Michael E, Bundy DAP, Grenfell BT (1996). Re-assessing the global prevalence and distribution of lymphatic filariasis. *Parasitology* **112**, 409–28.

Michael E, Bundy DAP (1997). Global mapping of lymphatic filariasis. *Parasitology Today* **13**, 472–6.

Norões J *et al*. (1996). Occurrence of living adult *Wuchereria bancrofti* in the scrotal area of men with microfilariaemia. *Transactions of the Royal Society of Tropical Medicine* **90**, 55–6.

Nutman TB, ed. (2000). *Lymphatic filariasis*. Imperial College Press, London.

Southgate BA (1992). Intensity and efficiency of transmission and the development of microfilaraemia and disease: their relationship in lymphatic filariasis. *Journal of Tropical Medicine and Hygiene* **95**, 1–12.

Weil GT, Lammie PJ, Weiss N (1997). The ICT filariasis test: a rapid format antigen test for the diagnosis of bancroftian filariasis. *Parasitology Today* **13**, 401–4.

7.14.3Guinea worm disease: dracunculiasis

R. Knight

The clinical manifestations of Guinea worm and its surgical removal were known in antiquity. Attention was drawn to the seasonal occurrence of painful limb blisters that broke down to reveal a 'worm' in the floor of an ulcer. *Dracunculus medinensis* is the longest nematode infecting humans; in the Bible it is described as the 'fiery serpent'. It was the first human parasite to be shown to have an arthropod intermediate host: in 1869 the Russian naturalist Fedtschenko described the worm's early development in *Cyclops*—the 'water flea'. Recent attention is directed at eradication, for despite its complex lifecycle this can be achieved by public health measures alone.

Aetiology—the biology of the parasite Fig. 1)

Mature female worms, 70 to 120 cm in length, migrate along fascial planes and subcutaneous tissue to reach the skin, usually below the knee. Tissue damage caused by worm products produces a blister that soon ulcerates. Immersion of the affected part in water causes the worm to contract and expel numerous rhabditiform first-stage larvae from the uterus at the ruptured anterior end of the worm. Larvae swim vigorously in water for up to 7 days and some are ingested by predatory copepod crustaceans of the genus *Cyclops*. They penetrate the gut of the intermediate host and develop with two moults in the haemocele over a period of 14 days to become infective third-stage larvae. When water containing infected *Cyclops* is swallowed, the released infective larvae burrow though the wall of the duodenum to reach retroperitoneal tissue. After about 100 days the worms mate and the females begin their migration towards the limbs; the male worms die and may later calcify. Ten months after infection most female worms, containing fully formed larvae, have reached their destination; within the next month they will rupture through the skin to begin the cycle anew (Plate 1).

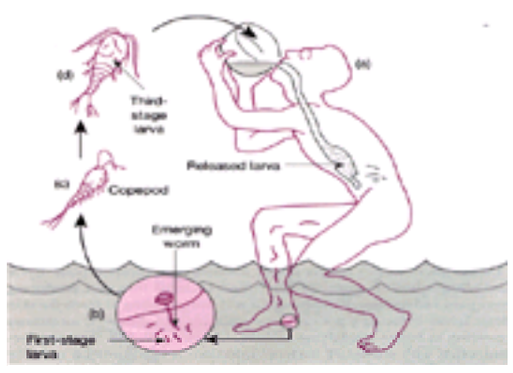

Fig. 1 Lifecycle of Guinea worm in man. (a) Copepods infected with third-stage larvae are ingested in drinking water; larvae are released in the intestine, migrate to the body cavity, mature, and mate. (b) Gravid female worms migrate to the limbs, cause a blister to form and release first-stage larvae into water. (c) First-stage larvae are ingested by copepods. (d) Larvae undergo two moults in the copepod and are infective after 2 weeks.

Epidemiology

Guinea worm transmission is predominantly rural with an annual cycle that often coincides with the planting or harvesting season. The seasonal morbidity causes great economic hardship. Water sources containing *Cyclops* are easily contaminated by infected people, including those seeking relief by immersion of their painful lesion. In semi-arid areas, transmission occurs in temporary ponds during the rainy season; in wetter areas, flooding and water turbidity limits transmission during the rains and infection occurs in shallow wells during the dry season. For practical purposes there is no zoonotic reservoir although infected dogs have been found in endemic areas and primates can be experimentally infected. Related species of *Dracunculus* are found in mink, raccoons, and otters in North America.

Geographic distribution

This infection was previously endemic over wide areas of the Middle East and the Indian subcontinent. Largely as a result of improved and protected water sources the infection disappeared from the Central Asian Republics between 1926 and 1933, from Iran in the 1970s, and Yemen and Saudi Arabia in the 1980s; India and Pakistan have recently become free of infection. It is now limited to the Sahel and Guinea savannah, between 2° and 18° north, in sub-Saharan Africa with most cases in southern Sudan, Niger, Nigeria, Mali, Burkina Faso, Chad, Ghana, and Uganda (Fig. 2). Formerly it was also present in the Americas having been introduced with the slave trade. By the 1880s it disappeared.

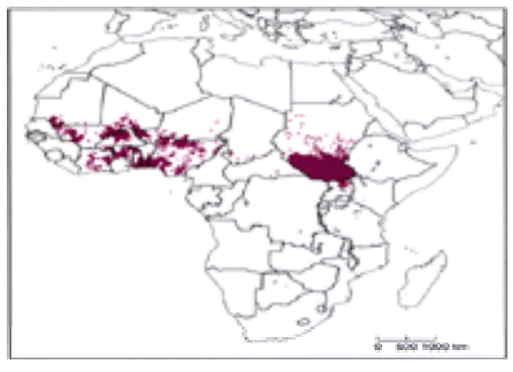

Fig. 2 Distribution of dracunculiasis – endemic villages in Africa. (Reproduced from Peries H, Cairncross S (1997). *Parasitology Today* **13**, 434, with permission: data from Joint WHO-UNICEF Programme on Mapping and Geographic Information Systems for Dracunculiasis Eradication (Health Map), WHO/CTD, Geneva.)

Clinical features

The blister is the first sign of infection in most patients (Plate 1). In others pre-emergent worms may be seen or felt under the dermis, some are actively motile. Allergic prodromal symptoms with urticaria, facial oedema, dyspnoea, and gastrointestinal manifestations may precede the blister by a few days; they disappear when the blister ruptures. Most patients have one or two worms each season, but up to 50 have been recorded. While most gravid worms emerge from a limb, other sites

include the trunk, scrotum, and vulva (Plate 3).

Uncomplicated cases resolve within 4 weeks; local complications derive from sensitization to worm products and bacterial infection producing severe pain and prolonged disability. Gravid worms failing to reach the skin release larvae inducing vigorous tissue reactions and abscesses, sometimes presenting as bubos, epididymo-orchitis, or acute arthritis. Joint involvement, often with secondary bacterial infection, is also common near the site of emergence: this leads to ankylosis and tendon contractures with deformities and permanent disability. Immature female worms may die before reaching the skin and become encapsulated by host tissue, where some calcify; they may also enter ectopic sites including the orbit, pericardium, and central nervous system. Mortality is usually less than 1 per cent. It results from systemic or local bacterial infection; tetanus is a significant risk when spores contaminate open lesions.

Diagnosis

Most patients in endemic areas recognize their condition. Worms release larvae on contact with water and these can be seen as a milky cloud. When the worm is not visible, ulcers may be irrigated with saline and the centrifuged deposit examined for larvae.

Patient management

Local treatment can be very painful and often must be repeated. Warm moist packs should be applied for several hours, followed by gentle massage along the tract of the worm towards the ulcer. Light traction is then applied to the worm; breakage must be avoided as this greatly aggravates the situation. Analgesics and antibacterial soaks are useful; oral antibiotics are often necessary. Between local treatments the lesion must be bandaged to reduce the risk of bacterial infection and contamination of water sources.

Pre-emergent worms can be surgically removed, a practice originating in India. A small incision is made adjacent to the worm near its mid-point, and a loop of worm is lifted out with a blunt curved probe (Plate 2). Massage is applied along the length of the worm towards the incision and by gentle traction the whole worm can usually be removed; in the event of breakage the worm ends should be ligated to minimize contact between host tissue and worm antigens. Deep abscesses require surgical treatment. Anthelmintics have no role in the treatment of Guinea worm.

Control and eradication

Several factors facilitate control: Guinea worm is recognized by local communities as a major health problem, there are no carriers beyond the annual cycle, and there is no animal reservoir. Provision of save water for drinking is the key to control; piped water supplies are unrealistic in most endemic areas, but covered tube wells or hand dug wells provided with parapets are appropriate. Additional measures are filtration of household water with finely woven cloth and the application of temephos (Abate) to ponds to kill copepods.

National programmes have played a major role in many endemic areas. Case detection surveys and health education can be integrated into existing primary health care systems. Unhygienic local treatments such as mud or leaf poultices and crude methods of worm extraction must be discouraged.

Several international health agencies took up the challenge of Guinea worm eradication in the mid-1980s with the target eradication date of 1995. Much has been achieved but the target was missed. Initial expensive hydrological programmes were later replaced by training of local cadres who could conduct health education, case detection, and management, partly independent of local health care services. In some areas private sector initiatives have been able to gain commercially from the publicity achieved by adopting control in a defined area.

There has been a decline of about 95 per cent in the incidence of Guinea worm in the last 15 years. The last stages of eradication will be the most difficult as vertical programmes then become inefficient. Unfortunately many of the major residual foci are in situations of civil disorder and mobile refugees; in others, lack of resources or an absence of democratic institutions will slow progress.

Further reading

Cairncross S *et al*. (1996). Community participation in the eradication of Guinea worm disease. *Acta Tropica* **61**, 121–36.

Hopkins DR *et al*. (1995). Eradication of dracunculiasis from Pakistan. *Lancet* **346**, 621–4.

Issakah-Tinorgah A *et al*. (1994). Lack of effect of ivermectin on prepatent Guinea-worm: a single-blind, placebo-controlled trial. *Transactions of the Royal Society of Tropical Medicine* **88**, 346–8.

Muller R (1971) *Dracunculus* and dracunculiasis. *Advances in Parasitology* **9**, 73–151.

Periès H, Cairncross S (1997). Global eradication of Guinea worm. *Parasitology Today* **13**, 431–7.

7.14.4 Strongyloidiasis, hookworm, and other gut strongyloid nematodes

R. Knight

Strongyloidiasis

The parasitic female *Strongyloides* worms are parthenogenetic. They measure 2 to 2.5 mm in length and normally live in tunnels between the enterocytes of the crypts of Lieberkühn in the duodenum and jejunum. In the external environment larvae may develop directly, through two moults, into infective larvae, in a manner similar to that of the hookworm (Fig. 1). Alternatively, they may follow the indirect cycle, develping into free-living male and female adult worms, about 1 mm in length, that produce a second generation of infective larvae. In either case the cycle is completed when infective filariform larvae penetrate the skin and are carried in the venous circulation to the lungs, from where they ascend the bronchi to be swallowed and so reach the upper small bowel, where they mature.

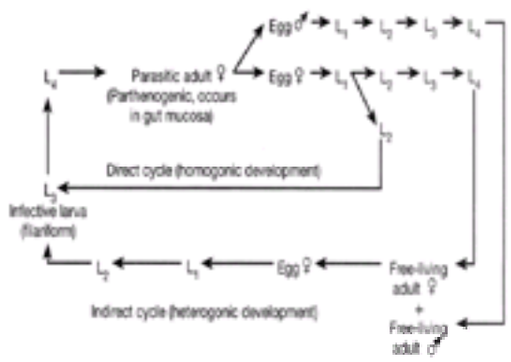

Fig. 1 Basic lifecycle in the genus *Strongyloides*; L_1, L_2, L_3, and L_4 are the larval stages. The indirect cycle occurs in the soil or faecal mass. Eggs of the parasitic female *S. stercoralis* hatch in the gut lumen and direct development may occur not only in the external environment but also on the perianal skin to produce external autoinfection, or in the gut lumen to produce internal autoinfection. The eggs of the parasitic female *S. fuelleborni* appear in the faeces and internal autoinfection is not possible.

Strongyloides stercoralis

Biology and epidemiology

Eggs hatch immediately on reaching the gut lumen, and the first-stage larvae (Fig. 2) then normally pass down the gut without moulting. Direct development in faecally contaminated soil takes 24 to 48 h; free-living adults mature in 72 to 96 h, and live for up to 10 days. Infective larvae can persist in the soil for 3 weeks. There is no second generation of free-living adults. Two types of autoinfection enable infection to persist in the host for long periods. In external autoinfection, infective larvae penetrate the perianal skin after rapid direct development on soiled skin. In internal autoinfection, larvae mature to the infective stage within the lumen of the gut and invade the mucosa of the small intestine or colon, they then pass via the gut lymphatics and portal vein to the lungs and back to the gut. In some patients, uncontrolled internal autoinfection leads to hyperinfection with massive worm loads and severe pathology.

Fig. 2 First-stage larvae of *S. stercoralis* in stool.

S. stercoralis is widely distributed in the tropics, where prevalence may be 5 to 10 per cent or higher in humid lowlands. It remains endemic in the southern United States, Japan, and in parts of southern Europe, for example, among Swiss and Italian horticulturalists. It also occurs in institutions when soil temperatures are high enough. Transmission among male homosexuals is very rare. Host risk factors are of great importance for internal autoinfection. Patients on steroid and cytotoxic therapy are at most risk, but also those with lymphomas and some other malignancies, hypochlorhydria, diabetic ketosis, hypogammaglobulinaemia, and malnutrition. Despite coprevalence with human immunodeficiency virus type 1 over much of its range, this viral infection does not predispose significantly to *S. stercoralis*

hyperinfection, except in patients with advanced AIDS. Servicemen in the Second World War became infected in Thailand and other parts of South-East Asia, mostly as prisoners of war. Many of these infections still persist and such people are at risk of hyperinfection if given steroids.

Pathology

In most persistent infections the parasite load is very low, evokes little pathological response, and the patient is free of symptoms. In some primary infections and when worm loads are higher there is villous blunting with oedema and cellular infiltration of the mucosa, leading to malabsorption and protein-losing enteropathy. In more severe infections and in hyperinfection the small-gut wall becomes oedematous and thickened with impaired motility, and the mesenteric lymph nodes are enlarged. In massive autoinfection there is patchy mucosal loss and some adult worms are found deep in the mucosa from where larvae may invade directly without entering the gut lumen (Plate 1). Invading infective larvae can produce a diffuse or haemorrhagic colitis; migrating or ectopic larvae may be found in any organ of the body. Rarely, adult female worms develop ectopically in the lungs, and these account for the occasional presence of eggs and rhabditiform larvae in sputum.

Clinical manifestations

Light persistent infections

Symptoms, if any, are usually intermittent, with episodes of upper abdominal pain, wheezy cough, and pruritus ani. Blood eosinophilia is common, and may be the only clinical finding. A pathognomonic sign is a rapidly migrating urticaria known as 'larva currens' that occurs on the buttocks, thighs, and lower trunk; it is a form of cutaneous larva migrans, arising from external autoinfection (Plate 2).

Moderate infections

Gut symptoms predominate, with diarrhoea and malabsorption. Weight loss and anorexia are prominent and not infrequently there is leg oedema. Pulmonary and skin lesions are not common. In primary infections a Loeffler's pneumonitis can occur, with high eosinophilia.

Hyperinfection

Diarrhoea is often severe, and sometimes bloody if there is colitis. Vomiting and abdominal distension may progress to pseudo-obstruction. Other manifestations are upper gastrointestinal bleeding, perforation, peritoneal and pleural effusions, pneumonitis (Fig. 3), and terminally, alveolar haemorrhages. Patients are often afebrile and without blood eosinophilia; they can deteriorate rapidly and develop Gram-negative septicaemia with shock, or meningitis, especially if they are immunosuppressed. Hypoglycaemia is a feature of autoinfection in malnourished children.

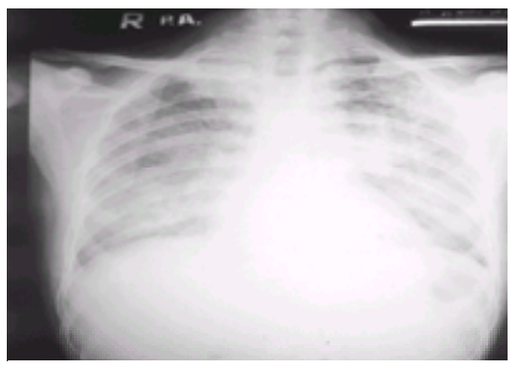

Fig. 3 Chest radiograph of a Thai patient who developed pneumonia as part of hyperinfection precipitated by corticosteroid treatment. (Copyright A.J.H. Simpson.)

Diagnosis

Rhabditiform larvae should be sought in the stool (Fig. 2). They may be scanty and numbers do not necessarily correlate with symptoms. Live larvae are seen in fresh, wet, preparations or Baermann concentrates. Agar-plate coprocultures give a result in 48 h, earlier than with conventional charcoal cultures. Formol-ether concentrates are useful, but sensitivity can be low. When stool specimens are not fresh, filariform *Strongyloides* larvae may be found. Duodenal aspiration is another useful technique. In hyperinfection, larvae may be found in sputum (Fig. 4) and in pleural, peritoneal, or cerebrospinal fluids.

Fig. 4 Gram stain of sputum from the patient whose chest radiograph is shown in Fig. 3, showing larvae of *S. stercoralis*. (Copyright A.J.H. Simpson.)

Serodiagnosis is useful, especially as a screening test in non-endemic areas. In heavy infections, small-bowel barium studies show segmental dilatation, narrowing, and abnormal motility; in hyperinfection, plain abdominal films may show fluid levels.

Treatment

Thiabendazole remains the drug of choice; 25 mg/kg is given twice daily (maximum 3 g/day), usually for 3 days. Intolerance is common and drug-induced hepatitis is reported. Treatment may fail in hyperinfection, which continues to have a high mortality. Such patients need supportive care and parenteral antimicrobials. Invermectin kills adult worms but not migrating tissue larvae; a single oral dose of 200 µg/kg, repeated after 1 week, or 200 µg/kg daily for 3 days are used, but experience in patients with hyperinfection remains limited. Albendazole is an alternative in non-urgent cases but cure rates are rather low.

Strongyloides fuelleborni

In this species eggs do not hatch in the gut lumen so there can be no internal autoinfection. The eggs are thin-walled and contain a larva. In Africa this parasite is common in non-human primates. In the forests of West and Central Africa, particularly Zaire, people are commonly infected, mainly from zoonotic sources. Elsewhere, for instance in Zambia and adjacent countries, there is person-to-person transmission. Infected volunteers have developed wheezing, upper abdominal pain, and loose stools, but symptomatology in natural human infections is poorly defined.

In Papua New Guinea a subspecies of this parasite, *S. f. kellyi*, is focally common in both children and adults. In a few communities a distinctive 'swollen belly syndrome' is associated with enormously high faecal egg counts and protein-losing enteropathy. Infants aged 2 weeks to 6 months are affected and show abdominal distension, diarrhoea, breathing difficulties, weight loss, hypoproteinaemia, and peripheral oedema; untreated, the mortality is high. There are no non-human primates in Papua New Guinea and no animal reservoir for this parasite is known. In infants, external autoinfection occurs when they are nursed in soiled string-bag cradles; transmammary transmission is suspected.

S. fuelleborni infection should be treated with thiabendazole. Supportive care including plasma infusion or blood transfusion, plus antibiotic cover, is needed for 'swollen belly syndrome' in Papua New Guinea.

Hookworm and other gut strongyloid nematodes

Adult strongyloid worms live attached to, or buried within, the bowel mucosa (Fig. 5). The ovoid eggs of all genera are similar in appearance, with thin, transparent shells containing a segmented embryo, commonly a 4-, 8- or 16-cell morula. Eggs hatch in the soil and development proceeds through three stages with two moults. The first and second larval stages feed upon bacteria. They are described as rhabditiform, because of resemblance to the soil nematode *Rhabditis*; the pharynx is short, muscular, and constricted in the posterior third, just anterior to a posterior bulb. The third stage, the infective filariform larva, does not feed and may retain the cuticle of the second stage; the pharynx is long and slender, without any constriction. In adults the buccal capsule and its oral armature, and the male copulatory bursa and spicules are used for species identification. Filariform larvae from cultures are generically distinct.

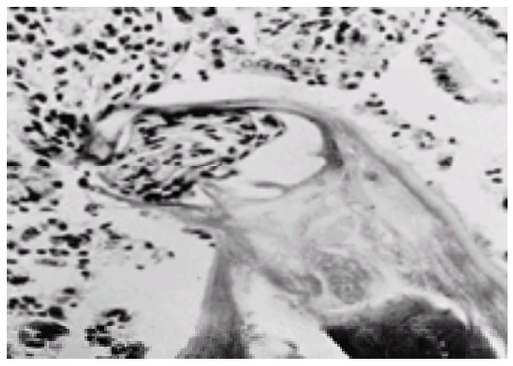

Fig. 5 Adult worm of *Necator americanus* showing relationship of its pharynx to a jejunal villus.

The hookworms

Aetiology—the biology of the parasites

Adult worm infections are due to *Ancylostoma duodenale* and *Necator americanus*. Several species that normally infect carnivores may accidentally infect humans and produce zoonotic cutaneous larva migrans or an eosinophilic enteritis in the case of *A. caninum*.

Adult worms measure 8 to 13 mm in length and taper at both ends (Plate 3). Anteriorly the worms are flexed dorsally, giving them their hooked appearance. They attach themselves to the jejunum by drawing mucosa into the buccal cavity (Fig. 5). A vigorous pharyngeal pump enables blood and tissue fluids to be ingested. Worms move frequently in response to host immunological responses. Females produce 5000 to 20 000 eggs per day, but output per worm declines as worm load rises. In the soil, development is temperature dependent. Under optimum conditions eggs hatch within 2 days and larvae develop to the infective stage in 5 days; they can persist in sandy soil for up to a month. Larvae penetrate host skin after soil contact, most commonly between the toes. After entry into dermal venules and lymphatics they are carried to the lung, ascend the bronchi and trachea, and after being swallowed, re-enter the gut where the final moult occurs. Eggs (Fig. 6) can appear in the faeces 50 to 60 days after cutaneous exposure.

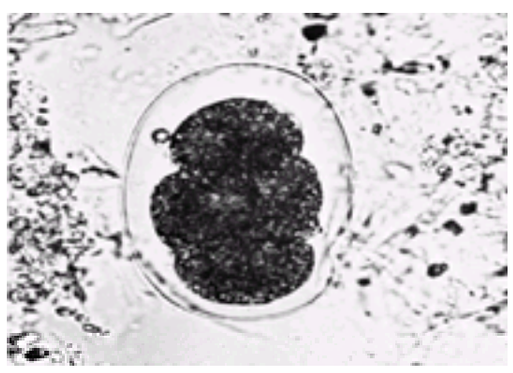

Fig. 6 Egg of *N. americanus*.

Epidemiology

N. americanus is found in the warm, moist tropics where transmission is sometimes perennial. Its introduction to the Americas dates from the transatlantic slave trade. It is a smaller worm than *A. duodenale* and the mouth is guarded by two cutting plates. The mean egg output per female is 8000 per day and the lifespan may exceed 5 years; transmission is exclusively by the percutaneous route.

A. duodenale is primarily a subtropical and temperate species with development in the soil at lower temperatures. It is widely distributed in North Africa, the Middle East, the Indian subcontinent, central and northern China, and in parts of Latin America. Formerly, it was endemic in southern Europe and Japan; it was responsible for 'miner's anaemia' in Cornish tin mines and the Gotthard tunnel. The mouth is guarded by two pairs of sharp teeth. Females produce 15 000 eggs per day. The lifespan is usually less than 2 years. In addition to the percutaneous route, larvae on vegetables can penetrate the buccal mucosa and undergo transpulmonary migration, or they can be swallowed and develop directly within the gut mucosa. Infection may also be transplacental; in China severe hookworm disease is reported in very young infants. Another lifecycle feature is arrested development when larval maturation is delayed at the third or fourth stage within skeletal muscle, or more commonly, in the gut mucosa. This postpones the onset of patent infection and is an adaptive mechanism to irregular or seasonal transmission.

In most populations where these parasites are endemic the prevalence and worm load both rise with age to reach a plateau in adults. Prevalence is highest in rural agricultural communities. Aridity and coolness at higher altitudes limit transmission, but irrigation schemes favour it by raising the water table. Children commonly acquire clinically significant infections between the ages of 5 and 10 years. Within communities, individuals differ greatly in worm load; behavioural factors are important but immune responses, including IgE antibody and eosinophils, limit the proportion of larval worms that mature to adults, the adult lifespan, and also female fecundity.

Pathogenesis of hookworm anaemia

Hookworms damage the mucosa mechanically and by the inflammatory response they evoke; bleeding continues at former attachment sites. Gut motility is affected, especially in primary infections and in children, and this may affect digestive and absorptive function. The major pathogenic mechanism is ingestion of plasma, interstitial fluid, and red cells by the adult worms. *A. duodenale* ingests about 0.15 ml of blood daily, and *N. americanus* 0.05 ml. Most red cells pass through the worm's gut and a proportion of the iron content, variously estimated at 10 to 50 per cent, is reabsorbed by the host. Because worm loads are commonly above 50, and may reach 500 or more, the cumulative effect can be serious. The main nutritional effects are iron deficiency and hypoproteinaemia. The rate at which blood loss leads to anaemia is determined by worm load, the duration of infection, iron stores, other blood loss, and dietary iron. Children, and pregnant or lactating women, with little reserve iron, can become anaemic in a few months; in a previously healthy adult male it can take 2 years or more. Loss of albumin into the gut may exceed the capacity of the liver to replace it; synthesis is depressed by low dietary protein, and by the anaemia. Hypoproteinaemia limits the normal, compensatory expansion of plasma volume that occurs in chronic anaemia. While the risk of pulmonary oedema is less, transition to a state of low cardiac output is made more likely.

Clinical features attributable to adult worms

In acute primary infections and in children, epigastric pain is common and may be associated with poor appetite and sometimes diarrhoea. Anorexia is an important mechanism leading to nutritional deficit in children. A few patients develop overt gut bleeding, and melaena is reported in transplacentally infected infants in China.

Most patients present with slowly progressive iron-deficiency anaemia, many have no gut symptoms. Exertional dyspnoea may begin at a haemoglobin level of 8 g/dl, but may not be noted until it falls to 5 g/dl. Palpitations, weakness, and faintness on exertion are common; and sometimes precordial pain or leg claudication. A puffy oedema of the face, arms, and hands is typical, and often unaccompanied by dependent oedema. Other features in heavy infections are mental apathy and depression, and in adults, amenorrhoea or impotence. Pica is common, especially in pregnancy.

Milder degrees of anaemia reduce physical work performance in adults. In children, growth and development may be slowed and cognitive impairment leads to reduced scholastic achievement.

Assessment of cardiovascular status is essential in anaemic patients, to differentiate a well-compensated, high-output state from a dangerous low-output one.

Clinical features attributable to larval worms: cutaneous larva migrans

Cutaneous lesions take the form of migrating, itchy, red, serpiginous papules, known as creeping eruption or cutaneous larva migrans (Plate 4). They commonly become vesiculated and excoriated, with bacterial pyoderma. *A. duodenale* or *N. americanus* cause 'ground itch' among estate workers and prominent lesions occur in experimental human infections; however, in many endemic areas they are unnoticed.

Zoonotic hookworms produce more vigorous lesions that may continue to move for several months. Most infections are due to *A. braziliense*, which is common in dogs throughout the tropics, subtropics, and warmer temperate regions. Less common are infections by two other dog parasites, *A. caninum* and *Uncinaria stenocephala*, and the cattle hookworm *Bunostomum phlebotomum*. Infections occur on sandy bathing beaches, in children's play areas, and by contact with pet sandboxes. Lesions are most common on the lower legs and buttocks, but also occur on the arms, hands, and face.

Wheezy cough due to pneumonitis is more common with *A. duodenale*; symptoms can continue for many months after one exposure, owing to remobilization of larvae arrested in muscle. Lung symptoms are most prominent in heavy primary infections.

In Queensland it is now recognized that larvae of the dog hookworm *A. caninum* can reach the gut to produce an eosinophilic enteritis with abdominal pain and vomiting; immature adult worms are found at laparotomy or colonoscopy.

Diagnosis

Stool microscopy will reveal eggs (Fig. 6), except in prepatent infections; examination of stool concentrates is rarely necessary. Faecal egg counts per gram of stool enable the intensity of infection to be estimated. The simplest method is a semiquantitative wet smear, using 2 mg of stool. For more precise results the use of the McMaster counting chamber is recommended. An alternative is the modified Kato technique, but this requires special care as hookworm eggs can overclear and become invisible. Isotope studies indicate that, with either species, 1000 eggs per gram is equivalent to 2.2 ml of blood loss per day. Culture to the infective larval stage, using the Harada Mori technique, will differentiate the two major species and the other genera of gut strongyloid nematodes.

Treatment

A single 400-mg dose of albendazole, or mebendazole at 100 mg twice daily for 3 days, are both very effective. Alternatives are pyrantel at 10 mg/kg daily for three or four doses, or bephenium at 5 g daily for three doses, the latter being less effective for *N. americanus.*

To replace iron reserves, oral ferrous sulphate will suffice in most patients, but several weeks of medication may be necessary. When compliance is doubted, consideration should be given to intramuscular iron or total-dose intravenous infusion of iron dextran.

Transfusion of packed or sedimented red cells may be necessary in pregnancy and when cardiac output is compromised. Frusemide may be necessary to cover the transfusion, but in other circumstances diuretics should be used with caution. Depletion of plasma volume in patients with hookworm anaemia together with hypoproteinaemia can compromise cardiac output. Even bed rest in formerly ambulant patients can lead to significant diuresis. Chemotherapy should generally be avoided in pregnancy.

Cutaneous lesions can be treated with thiabendazole at 25 mg/kg in two divided doses, for 2 days, and, if necessary, after 2 days' rest a further 5 days at the same dose. Alternatively, a single dose of ivermectin at 200 µg/kg may be given; this is more effective than a single dose of 400 mg albendazole.

Topical treatment avoids systemic effects (usually nausea). One 0.5-g tablet of thiabendazole can be ground up in 5 g of petroleum jelly or dimethylsulphoxide base and applied daily over the worm track for 5 days.

Control

Population-based measures are necessary when endemicity and morbidity are high. Latrines are generally beneficial, but can create foci for transmission when the water table is high. Provision of piped water reduces contact with soil polluted by promiscuous defaecation. Where human excreta is used as fertilizer, composting and chemical ovicides are needed. Cash-crop estates, plantations, and irrigation schemes should provide safe latrines and subsidized footwear.

Anthelmintic drugs can be deployed in several ways. Certain target groups such as agricultural and sewage workers, clinic outpatients with pallor, and anaemic blood donors can be treated empirically because of their likelihood of infection. Population chemotherapy, repeated twice yearly, aims to reduce both prevalence and mean worm load. It may include those with positive stool tests—selective chemotherapy; or whole communities—mass chemotherapy. The relative costs of drugs and diagnosis will change during the course of a programme. Single-dose medication is best and possible with mebendazole at a dose of 600 mg or albendazole at 400 mg.

Other gut strongyloids

Trichostrongylus spp.

These are common and economically important gut parasites of domestic ungulates. Infection is by ingestion of filariform larvae on vegetation. Development in the gut is direct, without lung migration. Adults are reddish-brown, 5 to 10 mm in length, and live with their anterior ends embedded in the jejunal mucosa where they feed on tissue fluid, not blood.

Human infection has been recorded with eight species; *T. colubriformis* and *T. orientalis* being the most important. Prevalence rates are highest in Iran, Iraq, Egypt, and Japan. Most infection is derived from sheep, goats, cattle, and camels, but in Iran *T. orientalis* is non-zoonotic. Worms cause mucosal damage, and loss of protein and some blood. Clinical features include abdominal pain, eosinophilia, and sometimes anaemia. The eggs are longer and narrower than those of hookworm, but larval culture is required for reliable differentiation. Infections respond to drugs used for hookworm.

Ternidens deminutus

Human infection is locally common in parts of Central and southern Africa. Infection is direct following oral ingestion of larvae; adult worms are 8 to 16 mm long. They live partly embedded in the colonic mucosa, where they produce superficial ulceration and cystic nodules. The worm is sometimes referred to as 'false hookworm', because of the similarity of its eggs; differentiation is important both clinically and epidemiologically. Non-human primates are infected in much of Africa, but most human infections are non-zoonotic. Infections respond to drugs used for hookworm.

Oesophagostomum spp.

These are important parasites of primates and ungulates: in veterinary practice they are known as the 'nodular worms'. Fourth-stage larvae and immature adults live in the colonic wall, often deeply situated or in the subserosa; lesions may become bacterially infected or perforate. Normally, adult worms return to the gut lumen. Most human infections are reported from forested parts of West and Central Africa. In some remote villages of north Togo and Ghana faecal surveys, using larval culture, have shown prevalences reaching 30 per cent. Most cases have presented surgically with masses, or abscesses, located in the caecum or other parts of the colon; or with bowel obstruction, or ectopic lesions in the peritoneum or abdominal wall. Clinically the lesions simulate carcinoma, tuberculosis, appendicitis, and amoeboma. Diagnosis in such cases has been histological.

The eggs resemble those of hookworm and are absent in prepatent surgical cases. Chemotherapy has been little studied in man, but albendazole has been successfully used. In Africa most human infections are with the monkey parasite *O. bifurcatum*, or *O. stephanostomum*, a parasite of anthropoid apes; in Asia, *O. aculeatum* is the likely cause. The ungulate species do not appear to infect humans.

Further reading

Adeneusi AA (1997). Cure by ivermectin of a chronic, persistent, intestinal strongyloidosis. *Acta Tropica* **66**, 163–7.

Ashford RW, Barnish G, Viney ME (1992). *Strongyloides fuelleborni kelly*i: infection and disease in Papua New Guinea. *Parasitology Today* **8**, 314–18.

Croese J *et al*. (1994). Human enteric infection with canine hookworms. *Annals of Internal Medicine* **120**, 369–74.

Grove DI, ed. (1989). *Strongyloidiasis: a major roundworm infection in man*. Taylor & Francis, London.

Grove DI (1996). Human strongyloidiasis. *Advances in Parasitology* **38**, 251–309.

Gulletta M *et al*. (1998). AIDS and strongyloidiasis. *International Journal of Sexually Transmitted Diseases and AIDS* **9**, 427–9.

Jongwutiwes J *et al*. (1999). Increased sensitivity of routine laboratory detection of *Strongyloides stercoralis* and hookworm by agar-plate culture. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **93**, 398–400.

Krepel HP *et al.* (1995). Reinfection patterns of *Oesophagostomum bifurcum* after anthelmintic treatment. *Tropical and Geographic Medicine* **47**, 160–3.

Mahmoud AAF (1996). Strongyloidiasis. *Clinical Infectious Diseases* **23**, 949–53.

Roche M, Layrisse M (1966). The nature and causes of hookworm anaemia. *American Journal of Tropical Medicine and Hygiene* **15**, 1029–102.

Schad GA, Warren KS, eds (1990). *Hookworm disease: current status and new directions*. Taylor & Francis, London.

7.14.5 Nematode infections of lesser importance

David I. Grove

Further reading

From time to time, a patient may be encountered who harbours an unusual nematode. Some of these organisms are free-living parasites and the patient has a spurious infection, usually as the result of ingestion of the worm or following the *in vitro* contamination of a clinical specimen such as faeces or urine. Other individuals may have true infections with worms being found either in the gastrointestinal tract or in the tissues. Many of these infections are with parasites of animals that are adapted poorly to the human host and are unable to complete their development in man. Thus, worms in varying stages of development including larvae, adults, and eggs may be found in specimens. Some parasites may be recovered from fluids and are viewed intact whereas others are seen only in histological sections. If there is uncertainty in identifying the worm in the former circumstance, help may often be obtained from a veterinary parasitologist who may be more used to dealing with the species concerned. In the latter instance, definitive diagnosis may be very difficult but excellent resources are available. [13,39] A summary of rarely reported nematodes is shown in Table 1. *Dirofilaria* species and unusual microfilariae in blood and tissues have been reviewed elsewhere. [40]

Nematodes found in the gastrointestinal tract may respond to a benzimidazole agent such as mebendazole (100 mg orally, twice daily, for up to 3 days) or albendazole (10 mg/kg orally, daily, for up to 1 week). Thiabendazole (25 mg/kg twice daily for several days) has been used traditionally orally for the treatment of systemic larval infections but its effectiveness is very variable; albendazole may be more active than thiabendazole and is absorbed better from the gut than mebendazole. If these drugs fail, ivermectin (0.15 mg/kg orally, daily, for several days) may be tried. Other drugs that have been used in these unusual nematode infections include levamisole and diethylcarbamazine. Unfortunately, some infections are refractory to all anthelmintics. Nevertheless, these worms generally cannot multiply in humans and the parasites will die spontaneously after months or years.

Further reading

1. Africa CM, Garcia EY (1936). A new nematode parasite (*Cheilospirura* sp.) of the eye of man in the Philippines. *Journal of the Philippine Islands Medical Association* **16**, 603–7.

2. Beaver PC, Jung RC, Cupp WE (1984). *Clinical parasitology*, 9th edn. Lea & Febiger, Philadelphia.

3. Beer RJ (1971). Experimental infection of man with pig whipworm. *British Medical Journal* **i**, 44.

4. Bhaibulaya M, Indrangarm S (1975). Man, as an accidental host of *Cyclodontostomum purvisi* (Adams, 1933), and the occurrence in rats in Thailand. *Southeast Asian Journal of Tropical Medicine and Public Health* **6**, 391–4.

5. Biocca E (1959). Infestazione umana prenatale da *Spirocerca lupi* (Rud. 1809). *Parassitologia* **1**, 137–42.

6. Boschetti A, Kasznica J (1995). Visceral larva migrans induced cardiac pseudotumor: a cause of sudden death in a child. *Journal of Forensic Science* **40**, 1097–9.

7. Boussinesq M *et al*. (1995). A new zoonosis of the cerebrospinal fluid of man probably caused by *Meningonema peruzzi*, a filaria of the central nervous system of Cercopithecidae. *Parasite* **2**, 173–6.

8. Buckley JJ (1933). *Necator suillis* as a human infection. *British Medical Journal* **i**, 699–700.

9. Burr WE, Brown, MF, Eberhard ML (1998). Zoonotic *Onchocerca* (Nematoda: Filarioidea) in the cornea of a Colorado resident. *Ophthalmology* **105**, 1494–7.

10. Calvopina M *et al*. (1998). Treatment of human lagochilascariasis with ivermectin: first case report from Ecuador. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92**, 223–4.

11. Chandler AC (1938). *Diploscapter coronata* as a facultative parasite of man, with a general review of vertebrate parasitism by rhabditoid worms. *Parasitology* **30**, 40–5.

12. Cheung WK *et al*. (1998). Conjunctivitis caused by *Thelazia callipaeda* infestation in a woman. *Journal of the Formosa Medical Association* **97**, 425–7.

13. Connor DH *et al.*, eds (1997). *Pathology of infectious diseases*, Vol 2, pp 1305–588. Appleton & Lange, Stamford.

14. Deardorff TL *et al.* (1986). Piscine adult nematode invading an open lesion in a human hand. *American Journal of Tropical Medicine and Hygiene* **35**, 827–30.

15. Dennett X *et al*. (1998). Polymyositis caused by a new genus of nematode. *Medical Journal of Australia* **168**, 226–7.

16. Doezie AM *et al*. (1996). *Thelazia californiensis* conjunctival infestation. *Ophthalmic Surgery Lasers* **27**, 716–19.

17. Eberhard ML *et al*. (1989). Intestinal perforation caused by larval *Eustrongylides* (Nematoda: Dioctophymatoidae) in New Jersey. *American Journal of Tropical Medicine and Hygiene* **40**, 648–50.

18. Evans AC, Markus MB, Steyne E (1990). A survey of the intestinal nematodes of bushmen in Namibia. *American Journal of Tropical Medicine and Hygiene* **42**, 243–7.

19. Fox AS *et al.* (1985). Fatal eosinophilic meningoencephalitis and visceral larva migrans caused by the raccoon ascarid *Baylascaris procyonis*. *New England Journal of Medicine* **312**, 1619–23.

20. Fülleborn F (1927). Durch Hakenwurmlarven des Hundes (*Uncinaria stenocephala*) beim Menschen erzeugte 'Creeping Eruption'. *Abhandlungen aus dem Gebiet der Auslandskunde, Hamburg Universität (Fetschrift Nocht)* **26**, 121–33.

21. Gardiner CH, Koh DS, Cardella TA (1981). *Micronema* in man: third fatal infection. *American Journal of Tropical Medicine and Hygiene* **30**, 586–9.

22. Goto Y *et al*. (1998). Creeping eruption caused by a larva of the suborder Spiruna type X. *British Journal of Dermatology* **139**, 315–18

23. Gutierrez Y, Cohen M, Machiaco CN (1989). *Dioctophyme* [*sic*] larva in the subcutaneous tissues of a woman in Ohio. *American Journal of Surgical Pathology* **13**, 800–2.

24. Jelinek T, Loscher T (1994). Human infection with *Gongylonema pulchrum*: a case report. *Tropical Medicine and Parasitology* **45**, 329–30.

25. Jones CC, Rosen T, Greenberg C (1991). Cutaneous larva migrans due to *Pelodera strongyloides*. *Cutis* **48**, 123–6.

26. Kagei N *et al*. (1992). A case of ileus caused by a spiruroid nematode. *International Journal for Parasitology* **22**, 839–41.

27. Kasimov GB (1941). (The first case of ostertagiasis in man). *Meditsinskaya Parazitologiya i Parazitarnye e Bolezn.* **10**, 121–3. In Russian. Abstracted in (1943). *Tropical Diseases Bulletin* **40**, 326.

28. Kates S, Wright KA, Wright, R (1973). A case of human infection with the cod nematode *Phocanema* sp. *American Journal of Tropical Medicine and Hygiene* **32**, 606–8.

29. Keller AE (1935). The occurrence of eggs of *Heterodera radicicola* in human feces. *Journal of Laboratory and Clinical Medicine*, **20**, 390–2.

30. Kenney M *et al*. (1975). A case of *Rictularia* infection of man in New York. *American Journal of Tropical Medicine and Hygiene* **24**, 596–9.

31. Kenney Y, Yermakov V (1980). Infection of man with *Trichuris vulpis*, the whipworm of dogs. *American Journal of Tropical Medicine and Hygiene* **29**, 1206–8.

32. Koyama, T *et al*. (1973). *Terranova* (Nematoda: Anisakidae) infection in man. II. Morphological features of *Terranova* sp. larva found in human stomach wall. *Japanese Journal of Parasitology* **21**, 257–61.

33. Le VH, Duong HM, Nguyen, LV (1963). Premier cas de capillariose cutanée humaine. *Bulletin de la Société de Pathologie Exotique* **56**, 121–6.

34. Little MD *et al*. (1983). *Ancylostoma* larva in a muscle fiber of man following cutaneous larva migrans. *American Journal of Tropical Medicine and Hygiene* **32**, 1285–8.

35. Mao SP (1991). Protozoan and helminth parasites of humans in mainland China. *International Journal for Parasitology* **21**, 347–51.

36. Maruyama H *et al*. (1996). An outbreak of visceral larva migrans due to *Ascaris suum* in Kyushu, Japan. *Lancet* **348**, 1766–7.

37. Nicolaides NJ *et al*. (1977). Nematode larvae (Spirurida: Physalopteridae) causing infarction of the bowel in an infant. *Pathology* **9**, 129–35.

38. Nosanchuk JS, Wade SE, Landolf M (1995). Case report of and description of parasite in *Mammomonogamus laryngeus* (human syngamosis) infection. *Journal of Clinical Microbiology* **33**, 998–1000.

39. Orihel TC, Ash LR (1995). *Parasites in human tissues*. American Society of Clinical Pathologists, Chicago.

40. Orihel TC, Eberhard ML (1998). Zoonotic filariasis. *Clinical Microbiology Reviews* **11**, 366–81.

41. Phills JA *et al*. (1972). Pulmonary infiltrates, asthma and eosinophilia due to *Ascaris suum* infestation in man. *New England Journal of Medicine* **286**, 965–70.

42. Pirisi M *et al.* (1995). Fatal human pulmonary infection caused by an *Angiostrongylus*-like nematode. *Clinical Infectious Diseases* **20**, 59–65.

43. Poinar GO Jr, Hoberg EP (1988). *Mermis nigrescens* (Mermithidae: Nematoda) recovered from the mouth of a child. *American Journal of Tropical Medicine and Hygiene* **39**, 478–9.

44. Prociv P, Croese J (1996). Human enteric infection with *Ancylostoma caninum*: hookworm reappraised in the light of a 'new' zoonosis. *Acta Tropica* **62**, 23–44.

45. Riley WA (1920). A mouse oxyurid, *Syphacia obvelata*, as a parasite of man. *Journal of Parasitology* **6**, 89–92.

46. Rosemberg S *et al*. (1986). Fatal encephalopathy due to *Lagochilascaris minor* infection. *American Journal of Tropical Medicine and Hygiene* **35**, 575–8.

47. Schaum E, Müller W (1967). Die Heterocheilidiasis. Eine Infektion des Menschen mit Larven von Fisch-Ascariden. *Deutsche medizinische Wochenschrift* **92**, 2230–3.

48. Sweet WC (1924). The intestinal parasites of man in Australia and its dependencies as found by the Australian Hookworm Campaign. *Medical Journal of Australia* **i**, 405–7.

49. Todd JC, Sanford AH (1943). *Clinical diagnosis by laboratory methods*. W.B. Saunders, Philadelphia.

50. Wittner M *et al*. (1989). Eustrongylidiasis—a parasitic infection acquired by eating sushi. *New England Journal of Medicine* **320**, 1124–6.

51.Yorke W, Maplestone RA (1926). *The nematode parasites of vertebrates*. J&A Churchill, London.

7.14.6 Other gut nematodes

V. Zaman

Ascariasis (roundworm)

Ascariasis is an infection caused by *Ascaris lumbricoides*. Normally, the adult worms are located in the small intestine. In unusual circumstances, such as fever, irritation due to drugs, anaesthesia, and bowel manipulation during surgery, the worms may migrate to ectopic sites where they may give rise to severe disease.

Geographical distribution

The distribution is cosmopolitan but the parasite occurs more frequently in moist and warm climates. In some rural tropical areas, the entire population may be infected. It is relatively more common in children, who also carry higher worm loads.

Morphology

The mature worm is cylindrical with tapering ends (Plate 1). It is creamy white to light brown in colour. The female measures 20 to 35 cm in length and 3 to 6 mm in breadth. The male measures 12 to 31 cm in length and 2 to 4 mm in breadth and has a curved tail. The head has three lips at the anterior end, which carry minute teeth or denticles along their margins. The lips can be closed or extended, allowing the worm to ingest food. In cross section, the worm reveals a thick cuticle, adjacent to which is the hypodermis which projects into the body cavity in the form of lateral cords (Fig. 1(a)). The somatic muscle cells are large and elongated and lie adjacent to the hypodermis. The worm is able to maintain its position in the small intestine by the activity of these muscles. If the somatic muscles are paralysed by anthelminthics, it is expelled by peristalsis.

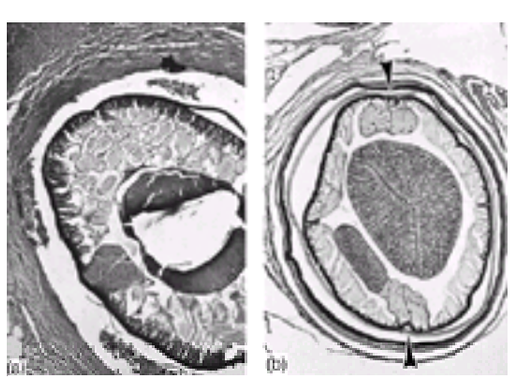

Fig. 1 (a) *Ascaris lumbricoides* in the bile duct (× 125). (b) Anisakis larva in cross section in the human stomach showing large bulbous lateral cords.

The fertilized eggs are ovoidal and measure 60 to 70 by 30 to 50 µm. When freshly passed they are not infective and contain a single cell. The larva becomes infective in the soil. The cell is surrounded by a thin vitelline membrane. Around the membrane is a thick, translucent shell, which in turn is surrounded by an irregular, albuminous coat (Fig. 2). The albuminous coat is sometimes lost or can be removed by chemical treatment, resulting in a decorticated egg. It was once assumed that the brown coloration of the egg was due to bile pigment, but tannins in the egg shell are probably responsible. The unfertilized eggs are 88 to 94 by 40 to 44 µm and

have disorganized contents. The larvae of *A. lumbricoides* may be seen in infected lungs and measure up to 2 mm in length, and 75 µm in diameter (Fig. 3). The larvae have a central intestine, paired excretory columns, and prominent lateral alae.

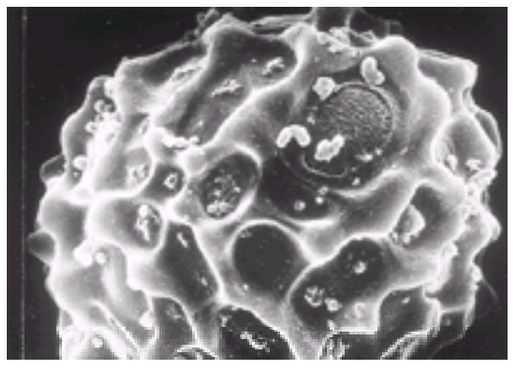

Fig. 2 Ascaris ovum seen by scanning electron micrography. (Copyright Viqar Zaman.)

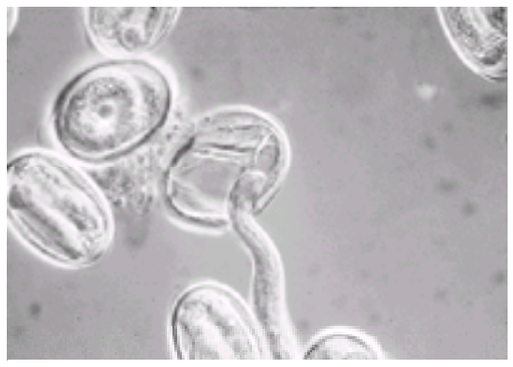

Fig. 3 Decorticated ova of Ascaris showing emergence of larvae. (Copyright Viqar Zaman.)

Lifecycle (see Fig. 4)

The gravid female produces 200 000 to 250 000 eggs daily. These take 3 or 4 weeks to develop into the infective stage, which is probably the third-stage rather than the second-stage larva as was previously thought. The eggs are resistant to chemicals and low temperatures and may remain viable for years in moist soil. On ingestion, the infective larva hatches out in the small intestine and penetrates the intestinal wall to enter the portal circulation. From the liver it is carried to the heart and via the pulmonary artery to the lungs. In the lungs, it breaks out of the capillaries into the alveoli and undergoes another moult to become a fourth-stage larva. From the lungs the larva moves up to the bronchi and then crawls over the epiglottis to enter the digestive tract. In the intestine, it moults again to become a sexually mature worm. The lifespan of an adult worm is approximately 1 year, after which it is spontaneously expelled. In hyperendemic areas, children are being continuously infected so that as some worms are being expelled, others are maturing to take their place.

Fig. 4 Lifecycle of *Ascaris lumbricoides*. (Adapted from Centers for Disease Control, Atlanta, Georgia, United States.)

Clinical aspects

In most cases, the infected person remains asymptomatic. However, there is much evidence to indicate that the presence of *Ascaris* causes nutritional problems and hinders the normal development of children. Occasionally, patients may develop fever, malaise, urticaria, intestinal colic, nausea, vomiting, diarrhoea, and central nervous system disorders.

The migration of larval *Ascaris* through the lungs may produce varying degrees of pneumonitis and bronchospasm known as Loeffler's syndrome (Plate 2). Chest radiographs may show diffuse mottling and increased prominence of peribronchial markings. There is generally high eosinophilia and the condition subsides after 7 to 10 days unless reinfection occurs. In areas where pig farming is common the larvae of *A. suum* (pig *Ascaris*) may also produce severe pneumonitis and bronchospasm.

Occasionally, ascariasis can cause severe, life-threatening disease. This could happen in the following situations.

1. When large numbers of worms get entangled to form a bolus and block the intestinal lumen producing signs and symptoms of acute intestinal obstruction.
2. When ectopic migration results in the entry of the worm into the appendix, common bile duct, or pancreatic duct. When the biliary tract is invaded, there is severe abdominal pain, often followed by suppurative cholangitis and multiple liver abscesses resulting from the disintegration of the trapped worm and secondary bacterial infection. Disintegration of the female worm releases a large number of eggs in the liver that can be recognized on histological examination.
3. When a worm blocks in the ampulla of Vater causing acute pancreatitis and pancreatic necrosis.

Diagnosis

This is usually made by detecting *Ascaris* eggs in the faeces. Sometimes, the patient brings developing or adult worms that have been passed in the faeces or have emerged from the anus or the nose in a sick child. They are roughly the same size as and have the appearance of earthworms. Occasionally, adult worms are outlined in the intestines during barium-meal examination.

Treatment

Whenever possible, all positive cases, irrespective of the worm load, should be treated, as even a few worms can undergo ectopic migration with dangerous

consequences.

Pyrantel pamoate

A single dose of 11 mg/kg body weight (maximum 1 g) is effective in curing over 90 per cent of cases of ascariasis. Side-effects are mild, if any, and the drug is well tolerated. It is also active against *Enterobius vermicularis* and hookworms (*Ancylostoma duodenale* and *Necator americanus*) but not against *Trichuris trichiura*. Its use in pregnant women and children of under 2 years is not recommended.

Mebendazole

This drug is a broad-spectrum anthelminthic with good host tolerance. It is active against *Ascaris*, *Trichuris*, *Enterobius*, and hookworms. This broad-spectrum activity is useful in endemic areas where multiple nematode infections are common. It is given as 100 mg twice daily for 3 days, for both adults and children of more than 2 years of age. Mebendazole should not be given to pregnant women, especially in the first trimester.

Albendazole

Like mebendazole this drug is also safe and effective against infections with *Ascaris*, *Trichuris*, *Enterobius*, and hookworms. The drug is given as a single oral dose of 400 mg for adults and children of more than 2 years of age. It should not be given to pregnant women.

Piperazine salts

These are old anthelminthics but still widely used because of their low cost and high degree of efficacy against *Ascaris* and *Enterobius*. The dose is 75 mg/kg (maximum of 3.5 g) given as a single dose daily for two consecutive days. It is not necessary to fast beforehand. Occasionally, symptoms involving the central nervous system, such as unsteadiness and vertigo, have been reported.

If signs of intestinal obstruction develop in a child living in an endemic area, ascariasis is a distinct possibility. The child should be admitted to hospital and prepared for surgical intervention. To start with the measures should be:

1. Decompression of the bowel through an intestinal tube with constant suction.
2. Rehydration and restoration of electrolyte balance by intravenous drip.
3. Antipyretics if fever is present.
4. Introduction of an appropriate dose of piperazine citrate through an intestinal tube. This induces flaccid paralysis of the worm.

In most cases this conservative therapy will relieve the obstruction and the child will rapidly recover. If, however, the signs of obstruction persist and the child's general condition worsens, laparotomy is required. Acute obstructive jaundice or pancreatitis due to obstruction of the common bile duct by *Ascaris* also requires urgent surgical intervention.

Prevention and control

As *Ascaris* eggs can survive in the soil for many years, prevention and control in endemic areas is difficult. Mass chemotherapy given at intervals of 6 months along with environmental sanitation can break the cycle. The prevalence of ascariasis and other soil-transmitted helminths is greatly reduced by improving housing. Infection is prevented by eating only cooked food and by avoiding green vegetables and salads in countries where human faeces are used as a fertilizer and where this parasite is endemic.

Anisakiasis

Anisakiasis is an infection caused by the larvae of nematodes belonging to the family Anisakidae.

Geographical distribution

The adult worms are commonly found in cetaceans (whales, dolphins, and porpoises) in many parts of the world. Human beings are infected when they eat raw or improperly cooked fish or squid. The incidence is highest in Japan, followed by Holland, Scandinavia, and countries along the Pacific coast of South America. A few cases have also been reported from California and the western United States.

Morphology

Complete speciation of the adults and larvae of this large group of parasites has not been done. In larval stages found in man the presence of large lateral cords which are bulbous in cross section is diagnostic of the family Anisakidae (Fig. 1(b)). Scanning electron microscopy shows the ventral side of the mouth to have a triangular boring tooth (Fig. 5).

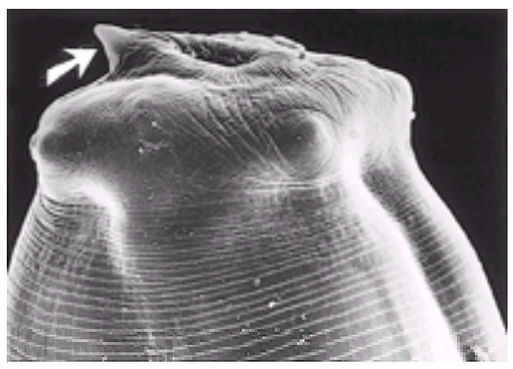

Fig. 5 Anisakis larva: third-stage larva of *A. simplex* showing the tip of the boring tooth (arrow) (× 400).

Lifecycle

Adults live in the lumen of the intestine of cetaceans. Eggs are passed in water and second-stage larvae are ingested by crustaceans, which are then ingested by fish or squid where they enter the muscles; cetaceans and man become infected by eating the fish or squid. In man, larvae do not develop to maturity but attach themselves to the mucosa of the stomach or intestine.

Clinical aspects

Most patients present with gastric symptoms that develop 4 to 24 h after eating infected fish. The symptoms are due to ulceration produced by the larvae as they burrow into the mucous membrane. There is epigastric pain, nausea, and vomiting, and sometimes haematemesis during the acute stage of the disease. If symptoms are mild and the patient is left untreated, the infection can become chronic with tumour formation. The small intestine may be involved, resulting in severe pain in the

lower abdomen, which may be misdiagnosed as appendicitis.

Diagnosis

Gastroscopy often reveals the lesion and the presence of larvae attached to the mucous membrane. Radiographs with a barium meal may show the presence of single or multiple ulcers and outline the worm. Serological tests are now available in some specialized centres.

Treatment

In acute infection an attempt should be made to remove all the larvae through a gastroscope. In chronic cases, surgical removal of the ulcerated areas or the tumour may be required. No effective chemotherapy is available. Prevention is by not eating raw fish and squid.

Capillariasis

Capillariasis is an infection by parasites belonging to the genus *Capillaria*. Two species infect man, *C. philippinensis*, which produces intestinal capillariasis, and *C. hepatica*, which produces hepatic capillariasis.

Geographical distribution

C. philippinensis has been described from the Philippines and Thailand. In the Philippines, the distribution of the disease includes the western and northern coastal areas of Luzon and the northeast of Mindanao. In Thailand, the infection is mostly sporadic and widely scattered. Infection has also been reported from Japan and recently from Dubai in the Middle East. *C. hepatica* is a rare human parasite but is commonly found in rodents in many parts of the world.

Morphology

Adult *C. philippinensis* are thin, small worms measuring 2.5 to 4.3 mm in length. They have a row of stichocytes at the anterior end, as in *Trichuris*. The eggs measure 36 to 45 µm in length and 19 to 21 µm in breadth, and have bipolar plugs, also like *Trichuris* (Fig. 6). However, unlike *Trichuris*, the eggs are not barrel shaped and the plugs do not protrude from the lateral ends. The adults of *C. hepatica* measure 52 to 104 mm in length and the anterior region contains the stichocytes. The eggs measure 48 to 66 by 28 to 36 µm and have bipolar plugs. The eggshell is thick and distinctly striated.

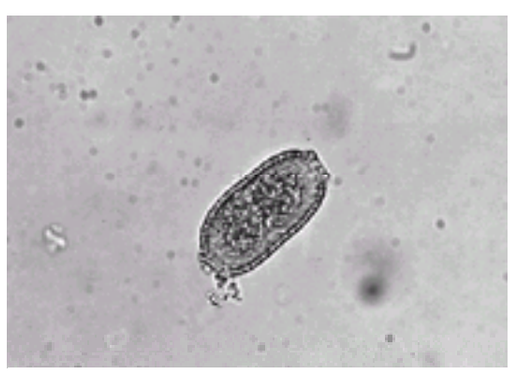

Fig. 6 *Capillaria philippinensis* egg (× 1400).

Lifecycle

The lifecycle of *C. philippinensis* has not been completely worked out. Humans are infected by eating freshwater fish and especially their succus entericus containing the infective larvae. Fish-eating birds act as a natural or reservoir host. Humans become involved accidentally in this natural fish–bird cycle. The main danger is of autoinfection, leading to very heavy worm loads.

C. hepatica is found in the liver of rodents and other mammals. The eggs are discharged in the liver tissue and remain there until the animal dies (Fig. 7). They eventually reach the soil by the decay of the carcass. Human beings are infected by accidentally swallowing embryonated eggs from the soil.

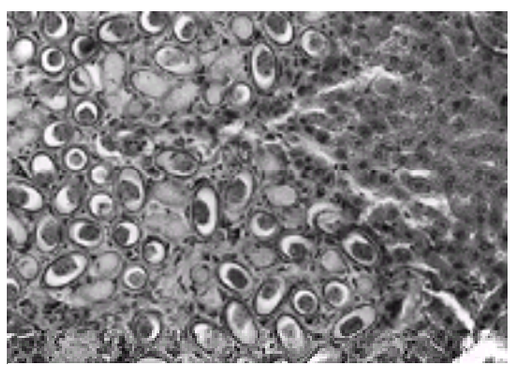

Fig. 7 *Capillaria hepatica* eggs in the liver (× 250).

Clinical aspects

C. philippinensis can produce severe and even fatal disease. Patients often present with abdominal pain, diarrhoea, and borborygmi. As the worm load increases through autoinfection, diarrhoea becomes more severe, with anorexia, nausea, and vomiting. Prolonged diarrhoea leads to cachexia and muscular wasting. There may also be signs of hypotension and cardiac failure. In untreated cases, the mortality rate approaches 20 per cent.

In *C. hepatica* infection, symptoms of visceral larva migrans may be present. The patient may have an enlarged tender liver with low-grade fever and eosinophilia.

Diagnosis

With *C. philippinensis*, diagnosis is made by finding the typical eggs in the faeces. Larvae or adult worms may also be present and repeated stool examination may be required in some cases. The parasite may also be found in jejunal aspirate or biopsy. With *C. hepatica*, diagnosis is made by identifying the parasite or eggs in a liver biopsy.

Treatment

All cases of *C. philippinensis* should be treated with mebendazole 200 mg, twice daily, until the symptoms subside and the eggs completely disappear from the faeces after repeated stool examination. This may take up to 20 days or more. Alternatively, albendazole 400 mg can be given daily for 10 days. Supportive measures to overcome malnutrition and diarrhoea will be required in severely ill patients. There is no specific treatment for *C. hepatica* infection. Infection with *C. philippinensis* is prevented by not eating raw fish in the endemic regions of Southeast Asia.

Trichinosis

Trichinosis is an infection caused by *Trichinella spiralis*.

Geographical distribution

The infection is endemic in many parts of the world where pork is consumed, including Central and Eastern Europe, Central, South, and North America, and parts of Africa and Asia. Infection is also endemic in the Arctic regions resulting from eating polar bear meat.

Morphology

The adult males are small nematodes measuring 1.4 to 1.6 mm. The female is viviparous and about twice as long as the male. The anterior part contains a row of glandular cells (stichocytes) as in *Trichuris trichiura*. The male worm lacks a spicule but has two copulatory papillae on the sides of the cloacal opening.

Lifecycle (see Fig. 8)

Human beings become infected by eating improperly cooked pork or pork products such as sausages. In some parts of the world, wild boars are heavily infected. After ingestion, the larvae are liberated in the small intestine and mature into adults. The female deposits larvae in the intestinal tissues from where they find their way into the bloodstream and then into the striated muscles of the body (Fig. 9). The most heavily parasitized muscles are the diaphragm, tongue, laryngeal, and abdominal muscles. After penetration, the larva undergoes three moults and coils into a spiral, which eventually becomes enclosed in a thick-walled cyst. In this form, the larva may remain viable for many years.

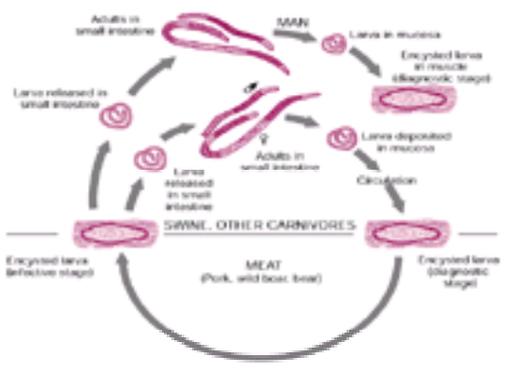

Fig. 8 Lifecycle of *Trichinella spiralis*. (Adapted from Centers for Disease Control, Atlanta, Georgia, United States.)

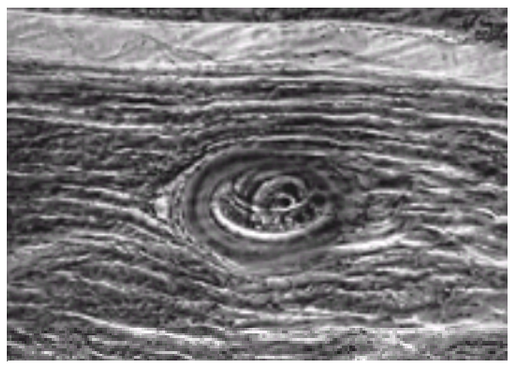

Fig. 9 *Trichinella spiralis* in human muscle (× 100).

Pigs become infected by eating infected scrap and garbage from slaughterhouses or farms. Occasionally, they may become infected by eating carcasses of infected rats.

Clinical aspects

Most people with light infections remain asymptomatic. In cases of heavy infection, there are three clinical stages.

The invasion stage

This is seen during the first week of infection and is due to juveniles and adults burrowing into the intestinal tissues. The patient complains of abdominal pain, nausea and vomiting, and diarrhoea of varying intensity. There may be fever, profuse sweating, and tachycardia.

The migration stage

This usually begins after the first week of infection. During this period, the larvae are liberated into the circulation by the gravid female and find their way to the muscles. Symptoms are attributable to toxic effects of the larvae and hypersensitivity reactions triggered by the liberation of parasite antigens. There is oedema of the face and periorbital tissues, fever, muscular tenderness, and hypereosinophilia. Complications involving the myocardium, lungs, and the central nervous system may occur due to the migrating larvae. However, the larvae do not encyst in the myocardium. A fine, macular skin rash, particularly on the trunk, may appear and last for a few days.

The encystment stage

This usually begins after the third week of infection. There is usually a gradual recovery from the symptoms. In a few cases with heavy infection, the symptoms may get worse and death may occur due to myocardiac failure, and respiratory and central nervous involvement. All serological tests become positive during this stage.

Diagnosis

This is based on a combination of clinical and epidemiological evidence. In a characteristic case, the patient will give a history of gastrointestinal disturbances (invasion stage) within 48 h of eating pork products, wild boar, or bear meat. If the patient presents in the later stage (migration stage), there is periorbital oedema,

myositis, irregular fever, and hyperosinophilia.

Among the various tests that become positive during the encystment stage, the two most commonly used are the skin test and the slide flocculation test. Recently, enzyme immunoassay has also been used. The skin test antigen is made from larvae and gives an immediate type reaction in positive cases. The test is very good for surveys but unsuitable for the detection of acute disease as it remains positive for many years after infection. The slide flocculation test is also prepared from larval antigen, which is attached to cholesterol particles. In a positive reaction, flocculation is seen under the microscope. The test remains positive for about 10 months after infection. Serum enzymes such as aminotransferases are elevated. Muscle biopsy is positive in approximately 90 per cent of clinically positive cases.

Treatment

The prognosis is good and most patients recover after the larvae have encysted. The mainstays of treatment are bedrest and salicylates. For myocarditis and severe myalgia, oral prednisone for 3 to 5 days (0.5–1.0 mg/kg per day) is useful and provides symptomatic relief. In experimental animals, thiabendazole is able to kill encysted larvae. In man, its efficacy against larvae is doubtful but a dose of 25 to 50 mg/kg per day for 2 to 5 days usually brings down the fever and eosinophilia. Mebendazole appears to be a good alternative to thiabendazole as it has fewer side-effects, and is given at a dosage of 300 mg daily for 7 days. A higher dose of 1000 mg daily for 10 to 14 days is recommended by some authorities to ensure complete killing of the larvae. Even with this high dosage, side-effects appear not to be serious, consisting of mild Jarisch–Herxheimer type reactions at the start of therapy. The manufacturer does not recommend giving mebendazole to children under 2 years of age or pregnant women.

Control and prevention

Trichinosis in the pig population can be greatly reduced or eliminated by hygienic rearing methods. Larvae in pork can be killed by freezing at – 18 °C for 24 h. Thorough cooking of pork is the best safeguard against infection in all endemic areas.

Enterobiasis (pinworm, threadworm)

Enterobiasis is an infection caused by *Enterobius vermicularis*.

Geographical distribution

This is one of the few parasites that is more prevalent in the temperate regions of the world than in the tropics. Children are more often involved than adults. It occurs in groups such as families living together, inmates of hostels, and in army camps.

Morphology

The male is approximately 5 mm long and 0.1 to 0.2 mm in diameter (Plate 3). The female is approximately 13 mm long and 0.3 to 0.5 mm in diameter. The gravid female has two distended uteri that practically fill the whole body. The male has a single spicule and a curved tail. The cervical alae of the cuticle allows easy recognition. The eggs are generally flattened on one side and measure approximately 50 to 60 µm in length and 20 to 30 µm in breadth. They have a thick, transparent shell. The eggs are unembryonated when passed but become infective within a few hours.

Lifecycle

The adults are mainly located in the caecal region (Fig. 10). The female deposits its eggs on the anus and perianal skin. Direct person-to-person infection occurs by inhalation and swallowing of the eggs. In addition, autoinfection occurs by contamination of fingers. It may occur as a sexually transmitted disease among male homosexuals. There is no visceral migration and the larva matures into an adult in the lumen of the intestinal tract. The lifecycle of the parasite is completed in about 6 weeks. Unlike *Ascaris* and *Trichuris* eggs, which need many days of development in soil before becoming embryonated, *Enterobius* eggs are embryonated when passed, hence there is rapid transmission from person to person.

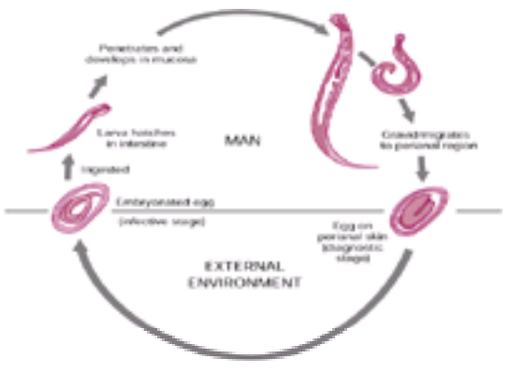

Fig. 10 Lifecycle of *Enterobius vermicularis*. (Adapted from Centers for Disease Control, Atlanta, Georgia, United States.)

Clinical aspects

The most common presenting symptom is pruritus ani. This can be very troublesome and occurs more often during the night. Persistent itching may lead to inflammation and secondary bacterial infection of the perianal region. Infected children may suffer from insomnia, emotional disturbance, anorexia, weight loss, and enuresis. Occasionally, adult worms may migrate, entering the female genital tract. Inside the uterus or the Fallopian tube they may get encapsulated and produce symptoms of salpingitis. In adolescents and children it is an important cause of vulvovaginitis. The parasite may also get lodged in the lumen of the appendix, leading to appendicitis (Fig. 11). The lifespan of the parasite is 3 to 6 weeks.

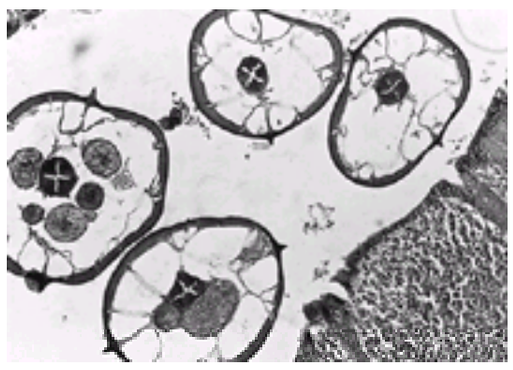

Fig. 11 *Enterobius vermicularis* in the lumen of the appendix (× 250).

Diagnosis

The eggs are not usually found in the faeces. They are most easily found around the anus, by swabbing or using cellulose adhesive tape. The anal examination for eggs should be done before defaecation or bathing. Sometimes intact worms are passed in the faeces and can be easily recognized by their size and shape.

Treatment

Attention to personal hygiene is an important part of treatment and prevention. The patient should be instructed to keep nails short and wash hands with soap and water after defaecating. The bed cover and sleeping garments should be changed every day and the floor in the bedroom kept clean. With these simple hygienic measures, infection will disappear on its own, due to the short lifespan of the parasite.

All the children and adults in a household should be treated at the same time. Many drugs are available to treat the infection. Piperazine citrate is given in a dose of 65 mg/kg for 7 days. The course is repeated after 2 weeks. Piperazine is contraindicated in renal and liver disease and epilepsy. Pyrantel pamoate is equally effective in a single dose of 11 mg/kg (maximum 1 g) and its side-effect profile is better than piperazine. The drug is repeated after 2 weeks. Mebendazole is effective in a single dose of 100 mg, repeated after 2 weeks. Alternately, albendazole in a single does of 400 mg can be given, repeated after 2 weeks. Both mebendazole and albandazole should not be given to pregnant women.

Trichuriasis (whip worm)

Trichuriasis is an infection caused by *Trichuris trichiura*.

Geographical distribution

It has a worldwide distribution and is the most common intestinal nematode in some tropical regions such as Southeast Asia.

Morphology

The adult male measures 30 to 45 mm and the female 35 to 50 mm in length. The parasite is commonly known as the whip worm because the anterior three-fifths is thin and elongated and the posterior two-fifths is bulbous and fleshy. One important feature of this group of worms is the possession of a thin, elongated oesophagus that is surrounded by gland cells known as stichocytes. The adults are mainly located in the caecum and produce barrel-shaped eggs, 22 to 50 µm long. At the lateral ends, they have transparent, blister-like plugs, which are single celled when freshly passed. In the soil the eggs become infective in about 3 weeks.

Lifecycle (see Fig. 12)

Infection occurs by the ingestion of the embryonated egg. The larva does not undergo visceral migration but penetrates the gut for a short period before returning to the lumen to mature into the adult stage. The worms attach themselves to the large intestine by threading their anterior end into the epithelium (Fig. 13). The posterior end hangs free in the lumen of the bowel. The whole period of development in the host takes about 3 months to complete.

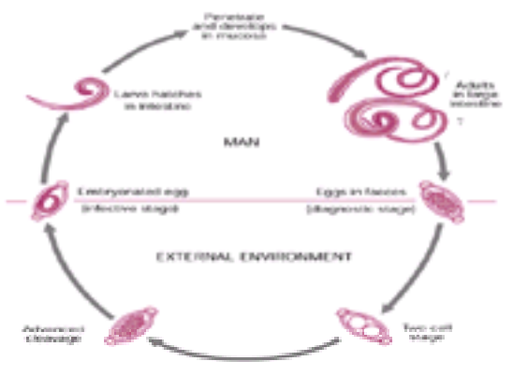

Fig. 12 Lifecycle of *Trichuris trichiura*. (Adapted from Centers for Disease Control, Atlanta, Georgia, United States.)

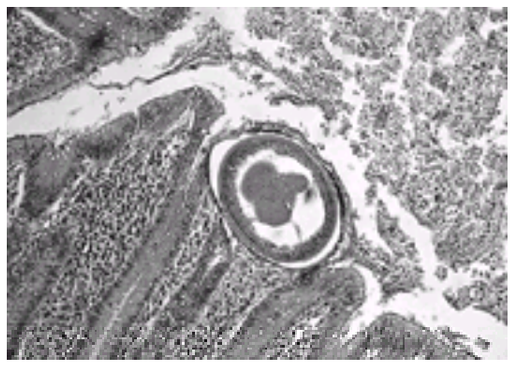

Fig. 13 *Trichuris trichiura*: anterior end embedded in the superficial layer of intestinal epithelium (× 250).

Clinical aspects

Light infections are generally asymptomatic. In heavy infections, there is colitis with the passage of blood and mucus in faeces. The clinical picture is often similar to that of amoebic dysentery. Heavy infection in children leads to anaemia that may cause oedema, and cardiac failure. There may be marked clubbing of the fingers. The anaemia is probably due to bleeding from the damaged and inflamed mucous membrane rather than sucking of blood by the parasite itself. In some cases, prolapse of the rectum occurs, probably due to constant irritation produced by the worms and the weakness of the levator ani muscle. Occasionally, the worm may lodge itself in the lumen of the appendix and cause acute appendicitis. The subtler form of trichuriasis is associated with long-term failure of children to grown in height, although they may not appear emaciated.

Diagnosis

This is based on finding characteristic barrel-shaped eggs in the faeces. Eosinophils and Charcot–Leyden crystals are often present. Sigmoidoscopy or proctoscopy may show worms attached to the mucous membrane and sometimes intact worms may be passed out in the faeces.

Treatment

Mebendazole

Mebendazole is the drug of choice and is given in a dose of 100 mg, twice daily for 3 days for both adults and children of not less than 2 years of age. Side-effects are few and it is well tolerated. During therapy, abnormal *Trichuris* eggs are produced as the drug interferes with embryogenesis of the worm. The drug should not be given to pregnant women, especially in the first trimester.

Albendazole

Albendazole is effective in a single dose of 400 mg for both adult and children of not less than 2 years of age. It should not be given the pregnant women. As with mebendazole, abnormal parasite eggs are produced during therapy.

Oxantel plus pyrantel pamoate

Oxantel is an analogue of pyrantel and a combination is effective against *Ascaris*, *Enterobius*, hookworms, and *Trichuris*. Dosage is 10 to 20 mg/kg body weight of each component as a single dose. In heavy infections, the drug may be repeated two or three times.

Preventive measures are the same as in ascariasis.

Further reading

Ascariasis

Crompton DWT, Nesheim MC, Pawlowski ZS, eds. (1989). *Ascariasis and its prevention*. Taylor and Francis, London.

Anisakiasis

Kliks MH (1986). Human anisakiasis: an update. *Journal of the American Medical Association* **255**, 2605.

Capillariasis

Cross JH, Basaca-Sevilla V (1983). Experimental transmission of *Capillaria philippinensis* to birds. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **77**, 511–14.

Trichinosis

Campbell WC (1983). Trichenella *and trichinosis*. Plenum, New York.

Trichuriasis

Cooper ES, Bundy DAP (1987). Trichuriasis. *Baillière's Clinical Tropical Medicine and Communicable Disease* **2**, 629–43.

7.14.7Toxocariasis and visceral larva migrans

V. Zaman

Visceral larva migrans is the name given to a syndrome characterized by hepatomegaly, fever, respiratory symptoms, and high eosinophilia. It is caused mainly by the migrating larvae of *Toxocara canis* and *T. cati*. Other parasites that may cause this syndrome are *Ancylostoma* spp., *Spirometra*, *Gnathostoma*, and *Alaria*.

Geographical distribution

Toxocarial infection occurs wherever there are large domestic dog and cat populations. Many cases have been reported from the United States and other Western countries. They result from close association between these animals and man. Surveys in Great Britain have shown that approximately 2 to 3 per cent of the population possess antibodies to these parasites.

Lifecycle

Infection with *T. canis* is maintained in the dog population by direct transmission from the soil containing embryonated eggs, from transplacental transmission from bitch to puppies, through the maternal milk to puppies, and by dogs eating infected meat containing larvae. These multiple routes of infection ensure that almost all puppies are born infected. Human beings become infected by ingesting embryonated eggs from the soil. Toddlers and young children are usually involved because of their habit of eating soil and dirt. The larvae hatch out in the small intestine and migrate to various organs of the body including the liver, lungs, eye, and brain. They do not mature into adult worms in man. After some time a granuloma forms around the larvae.

Morphology

The larva of *T. canis* is approximately 18 to 20 µm in diameter and that of *T. cati* is about 15 to 17 µm in diameter; they are otherwise indistinguishable. Both species have pointed lateral alae and two lateral excretory columns. The posterior region of the larva contains the intestinal tract. Larvae of *Toxocara* spp. are much smaller than these *Ascaris* spp. This may be the reason why *Ascaris* larvae are not widely dispersed in the tissues as they are unable to enter the small blood vessels. Eggs of *Toxocara* spp. are similar in size to those of *Ascaris* but their surface is pitted, making identification easy. Eggs are never seen in human faeces as adults mature only in dogs and cats.

In animals and man the migrating toxocarae larvae eventually became surrounded by a granuloma. Various cell types are involved in granuloma formation, including macrophages, lymphocytes, eosinophils, and fibroblasts. In a fully formed granuloma the larvae are surrounded by layers of fibrous tissue and inflammation subsides. It is now accepted that granuloma formation is caused by the immune system, especially T cells. It is an attempt by the host to limit disease, but it also favours the larva by ensuring its survival for a long time in a circumscribed area (Fig. 1 and Fig. 2).

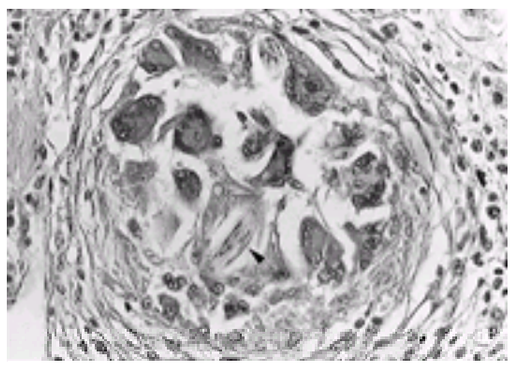

Fig. 1 Granuloma formation in an experimentally infected monkey showing a large number of giant cells and some fibroblastic reaction. The arrow marks the larva. Haematoxylin and eosin, × 400.

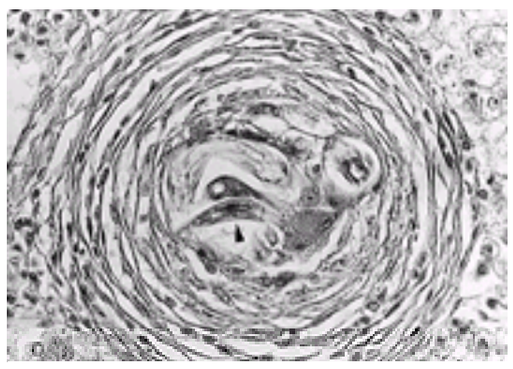

Fig. 2 Granuloma formation in the same animal as in Fig. 1 at a later stage, when the larva is completely surrounded by fibroblasts. Haematoxylin and eosin, × 400.

Clinical aspects

Toxocariasis may cause two different clinical pictures: classical visceral larva migrans syndrome and ocular toxocariasis.

Classical visceral larva migrans

This is seen most often in young children because of pica (dirt or soil eating). Most people remain asymptomatic. In a minority, symptoms consist of muscular pain, lassitude, anorexia, cough, and urticarial rashes. Physical signs include rhonchi, hepatomegaly, splenomegaly, and enlargement of the lymph glands. The acute phase generally lasts for 2 to 3 weeks followed by recovery. Sometimes the resolution of all the signs may take up to 18 months. Convulsions, paralysis, and other

neurological disorders result when larvae enter the central nervous system. Rarely the infection may end fatally if a massive dose of parasites has been ingested.

Ocular larva migrans

This is caused by granuloma formation in the eye. If this is near to the macula, impairment of vision or even blindness may result. Patients may present visual difficulty in one eye, with or without strabismus. As the generalized manifestation of visceral larva migrans may not be present, diagnosis is often difficult. Unlike in visceral larva migrans, eosinophilia may be absent. On fundoscopy a rounded swelling, often near the optic disc, may be detected.

Diagnosis

In classical visceral larva migrans there is leucocytosis with marked eosinophilia (20 to 80 per cent). In ocular larva migrans there may be no peripheral eosinophilia.

Serological tests are useful. An enzyme immunoassay using extracts of excretory–secretory products of *T. canis* larvae is positive in the majority of patients with visceral larva migrans. In ocular larva migrans the vitreous fluid is shown to have antibodies to the parasite.

Biopsy of the liver may show larvae with granulomas and eosinophilic infiltration. However, biopsy is rarely done because the chances of obtaining the appropriate specimen are remote, unless it is a massive infection. If facilities are available, laparoscopy may permit direct biopsy of a granuloma, which appears as a white dot on the surface of the liver.

Treatment

No antihelminthic drug is completely effective in killing the larvae and most patients recover without specific therapy. Diethylcarbamazine in a dosage of 6 mg/kg a day in three divided doses for 2 to 3 weeks has given equivocal results. Albendazole 400 mg twice a day for 3 to 5 days for both adults and children of more than 2 years of age, and mebendazole 100 to 200 mg twice a day for 5 days for both adults and children of more than 2 years of age may be of some value. The use of antihelminthics may sometimes provoke a greater inflammatory response, with worsening of the clinical picture, due to injury of the parasite. In severe cases, corticosteroids have been used with reports of improvement.

In ocular larva migrans, visible larvae can be photocoagulated by laser. Vitrectomy has been used in some cases, and local and intraocular steroids also appear to be of some value.

Prevention and control

Most cases of human toxocariasis are preventable by careful personal hygiene, not allowing children to play in environment likely to contain parasite eggs, and eliminating roundworms from pets, especially puppies and kittens. Unfortunately, people are often not aware that the infection can be transmitted from pets to humans. Efforts must be directed towards increasing this awareness.

Further reading

Schantz PM (1989). Toxocara larva migrans then and now. *American Journal of Tropical Medicine and Hygiene* **41** (Suppl.), 21–34.

Small KW *et al.* (1989). Surgical management of retinal retraction caused by toxocariasis. *American Journal of Ophthamology* **108**, 10–14.

7.14.8 Angiostrongyliasis

R. Knight

Human disease caused by two nematode species of the genus *Angiostrongylus* has been recognized comparatively recently. Both parasites normally infect rodents; molluscs are the primary intermediate hosts. They were initially thought to be rare and of limited distribution in humans, but these assumptions have proved incorrect. The epidemiology is complex because of multiple potential routes of transmission.

Angiostrongylus cantonensis

This is the rat lungworm; it causes cerebrospinal angiostrongyliasis. The first known case, reported in 1944, was a 15-year-old Taiwanese boy with meningoencephalitis in whose cerebrospinal fluid an immature adult worm was found. Detailed clinicopathological studies were made in 1962 during epidemics of eosinophilic meningitis in Tahiti.

Aetiology—the biology of the parasite

Adult worms live in the pulmonary arteries of rats; larvae from hatched eggs ascend the airways and so reach the faeces. Molluscs ingest these larvae and after two moults they are infective when eaten by a rodent. In the rat, infective larval worms migrate to the cerebral grey matter where they start to mature, move to the meninges, and then enter the venous sinuses and so reach the pulmonary arteries where maturation is completed. Infective larvae from a mollusc can also enter a second or third intermediate host in which they undergo no further development until they enter a mammalian host; such supernumerary hosts are termed paratenic or transport hosts, they are important sources of infection to humans.

Development in humans reaches the immature adult stage, measuring 11 to 15 mm in length. Nearly all will die in the superficial cortex, brainstem, and meninges causing vigorous tissue reactions; very few reach the lungs.

Epidemiology

Human infections occur throughout Oceania, especially Hawaii, Samoa, the Solomons, Papua New Guinea, Indonesia, the Philippines, and Northern Australia; and in south-east Asia, especially Thailand, Taiwan, and south Japan, but also India; a few infections are reported from Ivory Coast, Egypt, Madagascar, Cuba, the Caribbean, and South America. All ages can be affected and outbreaks have occurred after weddings and feasts; infections are often seasonal.

The principal rodent hosts are *Rattus rattus*, *R.norvegicus*, *R.alexandrinus*, and *R.exulans*. The prevalence in rats in endemic areas may be 40 per cent or more. The geographic spread and population increase of these peridomestic rodents has increased the zoonotic reservoir. Another factor leading to the increase in human infection has been the dispersal by human agency of the edible giant African snail *Achatina fulica* eastwards across the Indian Ocean and the Pacific, from Madagascar in 1800 to reach Hawai in 1936.

Many species of snail and slug act as intermediate hosts, and paratenic hosts include freshwater prawns, land and coconut crabs, frogs, and land planarians; in Thailand the yellow tree monitor is an important paratenic host. Common modes of transmission to humans include eating raw *Achatina* snails as a delicacy and for medicinal purposes; eating salads containing small undetected molluscs, their slime trails, or planarians; eating raw freshwater prawns and crabs; or drinking water containing tiny immature prawns, especially after heavy rains.

The modes of transmission differ geographically, by age and social group, and with time. In Thailand, *Pila* snails are a seasonal delicacy eaten by all the family, but young men take them raw with alcohol; another edible snail *Ampullaria canaliculatus* is infected in Taiwan and Japan. In the Ryuku islands of Japan, patients are usually infected by eating raw snails or toad liver for medicinal purposes.

Pathology

Inflammatory, granulomatous, and sometimes track-like lesions occur predominantly in the cortical grey matter and the meninges, but also in the brainstem and cerebellum; nerve roots and spinal cord may also be affected. Live worms are occasionally found at autopsy and dead worms are found in many of the lesions. The number of worms found varies greatly and may reach several hundred; worm tracks in the tissue and meninges are surrounded by a cuff of eosinophils; Charcot–Leyden crystal derived from eosinophils are numerous. Rarely adult worms have been found in human lung at autopsy. Ocular infection derives from worms that have migrated across the cribriform plate.

Clinical features

After an incubation period of 2 to 4 weeks the onset is acute with headache, intermittent at first, together with nausea and vomiting. There is constitutional upset and frequently meningism; fever is unusual. The illness is often self-limiting over a period of 4 weeks. Cranial nerve lesions include optic, abducens, and facial nerve damage. Less common are seizures, confusion, or radiculopathy with parasthesias, root pains, or weakness. Long tract signs and impaired consciousness are uncommon except in severe cases, but spinal cord damage can cause sphincter disturbance.

Ocular complications include retinal haemorrhages and larval worms in the vitreous or anterior chamber (Plate 1). Rarely migration to the lungs produces clinical evidence of pneumonitis. Numerous eosinophils occur in the cerebrospinal fluid and there is a blood eosinophilia.

Diagnosis

Lumbar puncture reveals high opening pressure with a clear or lightly turbid cerebrospinal fluid containing 500 to 2000 cells/mm of which 10 to over 90 per cent are eosinophils. Protein levels are high with normal glucose. Detailed examination at low power reveals larval or immature adult worms in up to 25 per cent of cases; they measure 5 to 15 mm in length. Cerebrospinal fluid changes may persist for up to 3 months. Computed tomography or magnetic resonance imaging may reveal focal cortical abnormalities. Serology using antigens from fourth-stage larvae is useful, but cross-reactions with other nematodes can cause difficulty. Techniques to detect worm antigen in cerebrospinal fluid and serum have also been developed.

Differential diagnosis is from other helminthic infections affecting the nervous system as eosinophils are otherwise rare in cerebrospinal fluid. A detailed geographic and dietary history is essential; conditions to be considered include gnathostomiasis, paragonimiasis, schistosomiasis, and neurocysticercosis. Confusion with

Gnathostoma spinigerum is a particular problem in Thailand, the latter more commonly causes long tract signs, bloody or xanthochromic cerebrospinal fluid, neck stiffness, and clouding of consciousness.

Treatment, prognosis, and control

Specific anthelmintic treatment is not useful and worm death aggravates the clinical condition. Headache can be relieved by repeated lumbar tap, analgesics, and sedatives. Steroids have been used in patients with focal neurological signs, but benefit is poorly substantiated. Larvae in the eye chambers should be removed surgically. Reported mortality rates vary from 0.5 to 30 per cent and depend mainly on the number of infective larvae ingested; some patients pass into coma after about 2 weeks and their prognosis is then very poor. Most patients improve in 2 to 4 weeks, but focal neurological deficits can persist longer; partial relapse after 2 months of illness may represent a reaction to dying worms. Some patients have relatively mild illnesses and can be discharged within a few days; during epidemics these patients may need only outpatient care.

Control measures include health education to limit ingestion of raw high-risk dietary items and unwashed salads. Warnings may be necessary regarding raw molluscs, amphibians, and reptiles used for medicinal purposes. Rodents in vegetable gardens and the peridomestic environment should be controlled.

Angiostrongylus costaricensis

This causes abdominal angiostrongyliasis. It was first recognized in Costa Rica in 1950 in surgical specimens simulating bowel malignancy. The parasite was described from such specimens in 1967 and the complete lifecycle in rodents was elucidated during the next 3 years.

Aetiology—the biology of the parasite

In both the rodent and human hosts the worms are located in the ileocaecal mesenteric arteries. The cotton rat *Sigmodon hispidus* is the principal reservoir host, but other species of rodents including the coatimundi are also involved, and even dogs and marmosets. In the rodent hosts, worm eggs embolize to gut wall capillaries and the hatched larvae pass into the gut lumen. Veronicellid slugs, especially *Vaginalus plebeius*, eat rodent faeces containing larvae and these develop with two moults in fibromuscular tissue of the mollusc into infective larvae over a period of 18 days. Infective larvae can persist in the slug for several months or be shed in slime trails. The prepatent period in rats eating infected slugs is 24 days.

In human infections the worms reach maturity but the embryonated eggs do not hatch.

Epidemiology

Infections occur especially in Costa Rica, Nicaragua, Guatemala, and Honduras, but also sporadically in the Americas from the United States to Argentina, and some Caribbean Islands. Recently, infections are being increasingly recognized from Brazil. Small veronicellid slugs are the main source of infection to humans; infection rates in these hosts can reach 85 per cent. Slugs may be unnoticed on fallen fruits or in salads, especially when small or chopped; their mucus also contains infective larvae. Many cases are in school children, but infants and older persons are also affected; an outbreak after eating mint is reported. Seropositivity in endemic areas suggests that there are unrecognized infections.

Pathology and clinical features

Lesions primarily affect small arteries; they produce subacute or chronic granulomatous inflammatory masses in the wall of the caecum (Fig. 1) and right colon; sometimes the predominant feature is ischaemic infarction. The finding of an adult nematode measuring 18 to 42 mm in length within a gut arterial vessel is diagnostic of this infection; eggs may be seen in vessels or in tissue where they are surrounded by eosinophil granulomas. Lesions also occur elsewhere in the colon or terminal ileum, in regional abdominal lymph nodes, or the omentum. Some larvae enter the hepatic artery and cause granulomatous or necrotic lesions in the liver; others enter testicular arteries causing similar lesions of the testis.

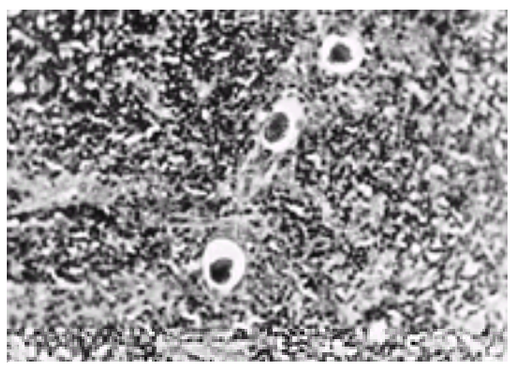

Fig. 1 Section of a human caecum showing three eggs of *A. costaricensis* with cellular infiltration mostly of eosinophils (by courtesy of Dr Pedro Morera, University of Costa Rica).

Clinically, most patients present with right-sided or right iliac fossa pain, with tenderness and sometimes a palpable mass in this region. Other features are eosinophilia, fever, diarrhoea, or rectal bleeding. Tender hepatomegaly with high blood eosinophilia occurs in some patients and sometimes focal necrotic lesions in the liver. Serious complications are bowel obstruction and perforation, and testicular infarction.

Diagnosis and treatment

Confirmation of diagnosis is usually made histologically on resected material. The condition can mimic appendicitis, bowel neoplasm, Meckel's diverticulitis, testicular torsion, or other surgical problems. Parasite eggs are not found in faeces, but serology using enzyme immunoassay or latex agglutination is useful. Contrast radiology reveals filling defects and altered motility of terminal ileum, caecum, or ascending colon. Laparoscopy can reveal the bowel and hepatic lesions; biopsy may be diagnostic.

The value of anthelmintic treatment remains unproven; thiabendazole or high doses of mebendazole have been used. Surgery is often necessary but can sometimes be deferred in uncomplicated cases when the diagnosis is strongly suspected, as spontaneous remission is common.

Preventive measures include washing and careful inspection of vegetables, and hand washing before meals by children and those preparing salads.

Further reading

Alicata JE (1991). The discovery of *Angiostrongylus cantonensis* as a cause of human eosinophilic meningitis. *Parasitology Today* **7**, 151–3.

Graeff-Teixeira C *et al*. (1997). Seroepidemiology of abdominal angiostrongyliasis: the standardization of an immunoenzymatic assay and prevalence of antibodies in two localities in southern Brazil. *Tropical Medicine and International Health* **2**, 254–60.

Kramer MH *et al*. (1998). First reported outbreak of abdominal angiostrongyliasis. *Clinical Infectious Diseases* **26**, 365–72.

Mackerras MJ, Sandars DF (1995). The life history of the rat lungworm, *Angiostrongylus cantonensis* (Chen). *Australian Journal of Zoology* **3**, 1–25.

Morera P (1996). Importance of abdominal angiostrongylosis in the Americas. In: Özcel MA, Alkan MZ, eds. *Parasitology for the 21st century*, pp 253–60. CAB International, Wallingford, UK.

Punyagupta S, Juttijudata P, Bunnag T (1975). Eosinophilic meningitis in Thailand. Clinical studies of 484 typical cases probably caused by *Angiostrongylus cantonensis*. *American Journal of Tropical Medicine and Hygiene* **24**, 921–31.

7.14.9 Gnathostomiasis

Pravan Suntharasamai

Gnathostomiasis is an extraintestinal infection with larval or immature nematodes of the genus *Gnathostoma* (order Spirurida). It is characterized by intermittent and migratory space-occupying lesions in the skin or the internal organs, resulting in inflammation and/or haemorrhage.

Aetiology

Four species of gnathostomes are known to infect man. Adult parasites live in the upper gastrointestinal tract of the definitive hosts: *Gnathostoma spinigerum*, the most common infection, in dogs, cats, and other mammals; *G. hispidum* and *G. doleresi* in pigs; and *G. nipponicum* in weasels. Larvae, hatched in water from ova shed with the host's faeces and ingested by cyclops, are eaten by freshwater fish, amphibians, reptiles, crustaceans, birds, or mammals. Third-stage larvae are found in the walls of the viscera and in the muscles of these second intermediate hosts. Unless they are eaten by definitive hosts, the parasites cannot developed into reproductive adults but they remain infectious to man and other paratenic hosts.

Consumption of the raw or undercooked flesh of second intermediate and paratenic hosts is the most common mode of transmission. Skin penetration after contact with infected material is less important. Prenatal transmission can occur, as larvae have been recovered in neonates as young as 3 days old.

Epidemiology

Isolated cases of *G. spinigerum* infections have been reported frequently in Thailand and Japan and sporadically in Australia, Bangladesh, Cameroon, China, Ecuador, India, Mexico, Southeast Asian countries, and Sri Lanka. Infections with the other three species have been reported from Japan.

Gnathostomiasis can present in places far away from these endemic areas due to migration of the latently infected human host or importation of the infective flesh of paratenic hosts. Consumption of a raw fish dish at a party can result in an outbreak.

Pathogenesis

The ingested larva penetrates the gut wall and migrates to the liver before wandering through almost any tissue except bone. Symptoms and signs vary according to the sites and sizes of the inflammatory or haemorrhagic lesions induced intermittently along the migratory route.

Clinical features

Nausea, vomiting, and epigastric pain may develop within 1 or 2 days after consumption of the infective food. Fever, pain in the right upper quadrant of the abdomen, chest pain, dry cough, and hypereosinophilia may develop within 1 to 2 weeks.

The primary invasive illness usually passes unnoticed and so the incubation period is not known in most cases. General health is scarcely impared and fever is uncommon. The illnesses can be categorized according to the affected organs as follows.

Cutaneous forms

Gnathostomal creeping eruption (Fig. 1)

This is rare in *G. spinigerum* infection but frequent with the other three species. The serpiginous track (Fig. 1) is similar to but bigger and more variable in depth than that caused by dog or cat hookworm larvae (see Chapter 7.14.9). A trail of subcutaneous haemorrhage is sometimes observed.

Fig. 1 Creeping eruption around the left thigh. (Reproduced from Bhaibulya and Charoenlarp (1983), with permission.)

Cutaneous migratory swelling

This most common manifestation of human gnathostomiasis is usually intermittent. The first swelling may develop 3 to 4 weeks after infection. Swelling can occur anywhere and may recur close to or distant from the original site (Fig. 2). It develops rapidly and usually lasts for about 1 to 2 weeks. Frequently it is extensive, involving the whole wrist or hand for example. Swelling of digits or plantar surfaces can be very painful and incapacitating. Itching is the main associated symptom. Regional lymphadenitis is usually absent. When swelling involves the eyelid, chemosis and conjunctival haemorrhage may be observed.

Fig. 2 Migratory swelling in a 23-year-old male. (a) At the eyelids for 5 days when seen on 5 June 1986. (b) At the right side of the upper lip on 9 June 1986 when the larva was picked out by needle puncture and squeezing.

The worms can escape spontaneously through the skin or the conjunctiva. The interval between episodes of swelling varies from a few days to a few months and rarely 1 to 2 years. Intermittent cutaneous migratory swelling can go on for more than 5 years.

Visceral forms

Visceral invasion as described below for *G. spinigerum* infection has not been reported in infections with other *Gnathostoma* species.

Spinocerebral gnathostomiasis

Involvement of the spinal cord commonly starts with intermittent, agonizing, shooting pains in a limb or a segment of trunk, followed by paraplegia with urinary retention and, rarely, quadriplegia. Sensation is correspondingly impaired and the Brown–Séquard syndrome is sometimes seen. A few patients with haematoma and inflammation due to brain invasion (Fig. 3) present with severe headache and vomiting, followed very quickly by coma, cranial nerve palsies, and/or hemiplegia. A rapidly advancing or changing pattern of neurological deficits is characteristic of the infection. Eosinophilic meningitis without focal neurological deficit or subarachnoid haemorrhage occasionally occurs.

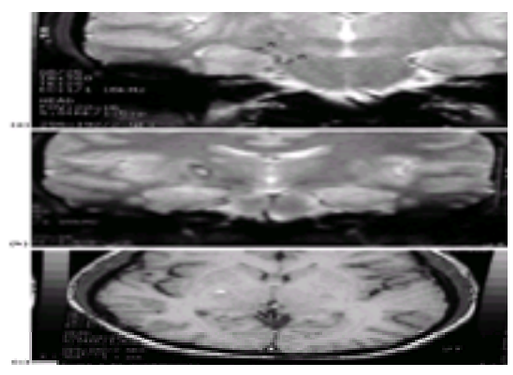

Fig. 3 Magnetic resonance image of a case of cerebral gnathostomiasis. (a), (b) Coronal T_2-weighted sections at the level of the brainstem and basal ganglion demonstrate a serpiginous mixed hyposignal with a hypersignal track (* *) along the right lateral upper pons and midbrain penetrating into the lentiform nucleus (arrow head). (c) Axial T_1-weighted section of the brain at the basal ganglion level demonstrates a small hypersignal T_1 subacute haemorrhage (short arrow) along the track of the parasite at the posterior limb of the internal capsule and adjacent serpiginous hyposignal T_1 track (long arrows). (By courtesy of Dr Jiraporn Laothamatas, Ramathibodi Hospital, Bangkok, Thailand.)

The cerebrospinal fluid can be bloody, xanthochromic, or slightly turbid with a minor increase in protein content. The proportion of eosinophils is higher than expected from haemorrhage *per se.*

Ocular gnathostomiasis

The gnathostome can be found in the anterior chamber (Fig. 4) and the vitreous. The parasite usually migrates through the sclera or the cornea. It can induce uveitis, iritis, intraocular haemorrhage, retinal detachment and scarring, and blindness.

Fig. 4 Gnathostome larva in the anterior chamber. (By courtesy of Dr Nesit Leelawong.)

Intra-abdominal and oral gnathostomiasis

These uncommon forms can present with intestinal obstruction or a painful intra-abdominal mass. Worms may emerge from the tongue, soft palate, gum, and cheek mucosa.

Pulmonary gnathostomiasis

Worms may be found in the sputum of patients with bronchitis, eosinophilis pneumonitis, pleurisy, pleural effusion, or pneumohaemothorax.

Genitourinary gnathostomiasis

The parasites have been recovered from blood-stained urethral discharge and urine, the glans penis, adnexal masses, and the cervix.

Auditory gnathostomiasis

The worms have been found in the external auditory canal in a patient with hearing loss and tinnitus, and penetrating the tympanic membrane in another patient.

Laboratory diagnosis

The diagnosis is definitive if the worm can be identified in sections of surgical specimens. The whole worm may be available if it emerges through the skin, in excretions and discharges, or from eye operations. Their sizes ranged from 0.34 × 2.2 mm to 1.0 × 16.25 mm. Their stage of development does not correlate with the duration of clinical illness. Infections with more than one worm are uncommon.

Enzyme immunoassay and Western blot tests are sometimes available in Thailand and Japan. Western blot is specific.

Blood eosinophilia (7 to 76 per cent) occurs irregularly in about 60 per cent of cases and therefore is not necessary for presumptive diagnosis.

Magnetic resonance imaging can show tortuous tracks and haemorrhage in cerebral gnathostomiasis (Fig. 3).

Differential diagnosis

Diagnosis of cutaneous forms is based on clinical characteristics, geographical and dietary history, and by excluding other causes. Differential diagnoses include contact dermatitis, angioedema and urticaria, Calabar swellings (caused by *Loa loa*), fascioliasis, paragonimiasis, sparganosis, dirofilariasis, and from non-infectious causes.

Gnathostomal aetiology is highly likely if rapidly advancing myelitis follows root pain, or if features of cerebral or subarachnoid haemorrhage occur in a person who is healthy apart from a history of cutaneous migratory swelling. Eosinophil pleocytosis is essential for the diagnosis, as is exclusion of eosinophilic meningoencephalitis caused by *Angiostronglus cantonensis*, *Baylisascaris procyonis*, and non-helminthic encephalomyelitis.

In intraocular infections, the larvae of *A. cantonensis* can be distinguished as they are thinner, longer, and folding. They usually appear in the eyeball 2 to 3 weeks after the manifestation of eosinophilic meningoencephalitis.

Visceral gnathostamiasis usually depends on identifying the worm in surgical specimens (at autopsy the worms may have migrated away from the site of the main pathological lesion), or in secretions such as sputum, urine, or vaginal discharge.

Treatment

Surgical removal is curative but advisable only in accessible areas such as the eye or skin. Blind exploration of subcutaneous tissues in areas of diffuse swelling is not productive.

Oral therapy with albendazole at an adult dosage of 400 mg twice daily for 2 to 3 weeks induces migration of the gnathostome to the skin. The worms are frequently recovered between days 2 and 14 of treatment by picking with a needle, excisional biopsy, or even by pinching with the patient's nails. However, the success rate is only 6 to 7 per cent. Recurrence of swelling in patients whose worms do not migrate to the skin is less frequent after albendazole treatment. Aminotransferases should be measured before this treatment even though hepatotoxicity at this dosage is usually mild and reversible.

Prognosis

Cerebral gnathostomiasis can be fatal and blindness is frequent after intraocular gnathostomiasis. Patients can be reassured that central nervous or intraocular involvement occur in less than 1 per cent of patients with cutaneous migratory swelling.

Prevention

All dishes that contain raw or poorly cooked flesh of animals in or imported from endemic areas must be avoided. Those who prepare potentially infected flesh should use gloves if prolonged exposure is likely.

Further reading

Bhaibulya M, Charoenlarp P (1983). Creeping eruption caused by *Gnathostoma spinigerum*. *Southeast Asian Journal of Tropical Medicine and Public Health* **14**, 226–8.

Inkatanuvat S *et al*. (1998). Changes of liver functions after albendazole treatment in human gnathostomiasis. *Journal of the Medical Association of Thailand*, **81**, 735–40.

Mijyazaki I (1991). *An illustrated book of helminthic zoonoses*. SEAMIC Publication No. 62, pp. 368–409. Southeast Asian Medical Information Centre, International Medical Foundation of Japan, Tokyo.

Rusnak JM and Lucey DR (1993). Clinical gnathostomiasis: case report and review of the English language literature. *Clinics in Infectious Diseases*, **16**, 33–50.

Suntharasamai P *et al*. (1992). Albendazole stimulates outward migration of *Gnathostoma spinigerum* to the dermis in man. *Southeast Asian Journal of Tropical Medicine and Public Health*, **23**, 716–22.

Swanson VL (1971). Gnathostomiasis. In: Marcial-Rojas RA, ed. *Pathology of protozoal and helminthic diseases with clinical correlation*, pp. 871–9. Williams and Wilkins, Baltimore.

7.15.1 Cystic hydatid disease (*echinococcus granulosus*)

Armando E. Gonzalez, Pedro L. Moro, and Hector H. Garcia

Introduction

Cystic hydatid disease is a zoonotic disease caused by infection with the larval stage (hydatid cyst) of the tapeworm *Echinococcus granulosus*. Hydatid cysts in liver and lung are frequent causes of human morbidity in endemic zones.

Aetiology

The lifecycle of *E. granulosus* requires two hosts. The adult tapeworm is found in the small intestine of the definitive host, usually dogs or other canids. It consists of only three to five proglottids, and measures between 3 and 7 mm long when fully mature. *E. granulosus* has remarkable biological potential; there may be as many as 40 000 worms in a heavily infected dog, each one of which sheds about 1000 eggs every 2 weeks. Dogs infected with *Echinococcus* tapeworms pass eggs in their faeces that contaminate the soil and vegetation and remain viable for long periods in cold humid places. Intermediate hosts (sheep, cattle, horses, pigs, and other mammals, including man) acquire hydatid disease by ingesting viable eggs of *E. granulosus*. Eggs hatch in the intestine freeing oncospheres which penetrate the intestinal mucosa and are transported by the blood and lymphatic systems to the liver, lungs, and other organs, where they develop into unilocular cysts.

Taxonomic studies have identified different strains of *Echinococcus*. Tapeworms developed from horse and sheep cysts are distinguishable, and *E. granulosus* from horse cysts is unlikely to infect humans. Wild cycles involving wolves with moose, caribou, and reindeer have been described in North America. In Africa, adult tapeworms have been identified in lions, hyenas, and jackals and cysts in antelopes and wild pigs.

Epidemiology

Hydatid disease is an important cause of human morbidity requiring costly surgical treatment. The infection is widely distributed in most parts of the world where sheep are raised and dogs are used to herd livestock. In the Americas most cases have been reported from Argentina, Chile, Uruguay, Peru, and southern Brazil. Recent studies in Peru have revealed prevalences of hydatid disease ranging from 5.7 to 8.9 per cent in highland villagers, and as high as 32 and 89 per cent in dogs and sheep, respectively. High prevalence of liver hydatid disease, with rates of up to 5.6 per cent, have also been reported in north-western Turkana in Kenya. *Echinococcus* is widespread in the Old World, particularly in Greece, Cyprus, Bulgaria, Lebanon, and Turkey. In the United States, most infections are seen in immigrants from endemic countries. However, sporadic autochthonous transmission is currently recognized in Alaska, California, Utah, Arizona, and New Mexico.

Communities at higher risk of infection include those where sheep are raised extensively and where dogs are used to care for large flocks of livestock. Known risk factors for infection include feeding dogs with raw offal and access of dogs to sheep that die in the field (Fig. 1). The risk of infection is also linked to poor hygiene and intimate contact with dogs. In north-western Turkana, dogs are allowed to stay within the house, and are used to clean up women's menses and lick vomit from faces and diarrhoea from the anal regions of their children.

Fig. 1 Epidemiological conditions for completion of the lifecycle of *Echinoccocus*: free dog waiting for sheep offal at a slaughterhouse.

Pathogenesis

The incubation period of human hydatid infections is highly variable and often prolonged for several years. Cysts have been reported to grow continuously between 1 and 5 cm per year. However, recent studies suggest that cyst growth is highly variable. Some cysts grow as much as 1 cm per year while other viable cysts showed no growth during 3 to 12 years of follow-up.

Most human infections remain asymptomatic; hydatid cysts are found incidentally at autopsy much more frequently than the reported local morbidity rates. The locality of the cysts, their size, and their condition determine the particular manifestations.

Clinical features

Hydatid cysts are most frequently seen in the liver (60 to 70 per cent) followed by the lungs (30 to 40 per cent). Signs of hepatic hydatid disease include hepatomegaly with or without the presence of a mass in the upper right quadrant. Obstructive jaundice, mild epigastric pain, indigestion, and nausea may occur occasionally. Hydatid cysts may become secondarily infected with bacteria presenting as a hepatic abscess. Features of lung involvement (Fig. 2) are cough, haemoptysis, dyspnoea, and fever. The ratio of liver to lung cysts may vary from one geographical region to another: a liver to lung ratio of 1.4:1 has been observed in Peru, in contrast to the 3:1 to 13:1 ratio reported in Argentina and Uruguay. Differences in *Echinococcus* strains may account for this variation. Brain cysts produce intracranial hypertension and epilepsy. Vertebral cysts compress the spinal cord causing paraplegia; bone cysts produce spontaneous fractures (Fig. 3 and Plate 1) and deformity. Sudden rupture of cysts in the peritoneal cavity may result in peritonitis, and rupture in the lungs may cause pneumothorax and empyema. Rupture may also cause allergic manifestations such as pruritus, oedema, dyspnoea, anaphylactic shock, and even death.

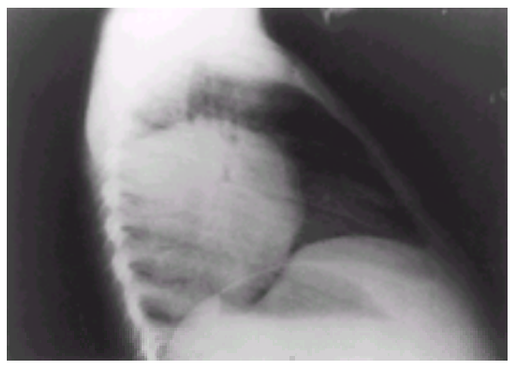

Fig. 2 Plain chest radiographs showing a lung hydatid cyst.

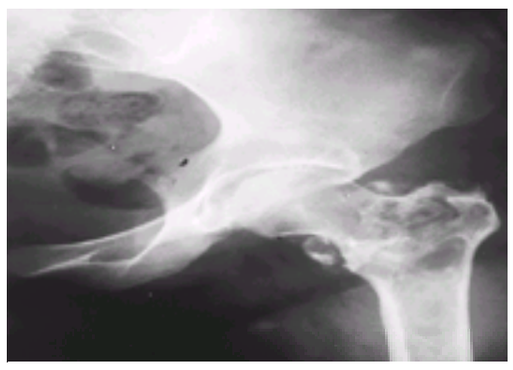

Fig. 3 Pathological fracture of the femur caused by hydatid infection (copyright D.A. Warrell).

Diagnosis

Clinical findings such as a space-occupying lesion and residence in an endemic region are suggestive of hydatid disease. Abdominal ultrasonography is an important aid to the diagnosis of abdominal cysts. Portable ultrasonography machines are used with good results in field surveys. Chest radiography is useful for diagnosis of lung cysts. CT scanning is very helpful, especially for diagnosis of non-typical lesions.

Serology

Efforts to develop sensitive and specific immunodiagnostic tests in humans have been relatively successful. A number of serological tests have been developed for diagnosis of hydatid disease, including an enzyme immunoassay, which identify antibodies against antigen B or components of this antigen. A Western blot assay based on the identification of three specific antigens of 8, 16, and 21 kDa is currently used. Major drawbacks in serological diagnosis are low sensitivity for detection of lung hydatid cysts and cross-reactivity with sera of patients with *Taenia solium* infection. In field surveys, serological tests should be used in combination with imaging techniques in order to detect most cases of hydatid disease.

Parasitological diagnosis

Although uncommon, this can be done from sputum samples of patients whose lung cysts have recently ruptured. Scolices have four spherical suckers and a rostellum with two rows of hooks.

Treatment

Surgery

Surgical removal of hydatid cysts remains the treatment of choice in many countries. The usual surgical approach involves injection of a protoscolicidal agent into the cyst, usually 20 per cent hypertonic saline solution or 90 per cent alcohol, followed by evacuation of the fluid, prior to surgical excision. Major risks of surgical treatment include accidental spillage of fluid and scolices into the peritoneal cavity, which may lead to anaphylaxis or secondary peritoneal hydatidosis. Recurrence rates following surgery may be as high as 30 per cent. Antihistamines are given as prophylxis and suction cones have been used to prevent spillage. The efficacy of these methods is uncertain.

Chemotherapy

Benzimidazole compounds have been shown to be effective against hydatid disease. Courses of albendazole in a dose of 10 to 15 mg/kg body weight per day for 28 days are interspersed with drug-free periods of 2 weeks. This regime cures approximately one-third of cases of liver hydatid disease and causes partial regression of cysts in another third of patients. However, many courses may be needed to achieve complete or partial cyst regression. Small liver or lung hydatid cysts should be treated with albendazole. Because of its high scolicidal activity, albendazole is recommended as a prophylactic agent 1 to 3 months prior to surgical intervention. Albendazole is indicated when surgery is contraindicated. Mebendazole may also be used, although it is less effective than albendazole. Albendazole, mebendazole, and other benzimidazole compounds should not be used in pregnant women because of their potentially teratogenic effects. Since benzimidazoles are potentially hepatotoxic, liver enzymes should be monitored before and during treatment.

Recent experimental studies in animals have shown that another benzimidazole compound, oxfendazole, has strong parasiticidal activity. Intermittent weekly therapy with oxfendazole was effective in sheep hydatid disease, suggesting the possibility that daily therapy as currently used with albendazole may not be needed. Future studies will explore the effect of oxfendazole in the treatment of human hydatid disease.

PAIR (percutaneous aspiration, injection, reaspiration)

PAIR consists of percutaneous puncture using sonographic guidance, aspiration of substantial amounts of the cyst fluid, and injection of a protoscolicidal agent, usually hypertonic saline for at least 15 min, followed by reaspiration of cyst contents. Albendazole should be administered before PAIR treatment, and antihistamines should be given to reduce the risk of allergic reactions if there is spillage of fluid. Good results have been reported with this procedure with no major complications. A recent study comparing the use of PAIR and surgical treatment for liver hydatid cysts found less complications and a shorter hospital stay in the PAIR-treated group.

Prevention and control

The earliest successful programme against echinococcosis was carried out in Iceland. It was based on a health educational campaign that eradicated the parasite. Control programmes have been aimed at educating dog owners to prevent their animals from having access to infected offal. This approach includes periodic treatment of sheepdogs with praziquantel (every 45 days), reduction in the dog population, close veterinary inspection of slaughterhouse facilities for the presence of dogs, and cremation of infected offal. Control programmes are in force in Argentina, Chile, and Uruguay. Partial success has been achieved in the first two countries. However, hydatid disease remains a serious problem in Uruguay. Control programmes in New Zealand and Tasmania have reduced the number of infected animals

and the incidence of human infection.

Serological tests such as the Western blot for diagnosis of sheep hydatidosis and the coproantigen ELISA for canine echinococcosis are potentially useful for measuring the burden of disease and monitoring control programmes in endemic regions. A recent major advance has been the development of a recombinant vaccine (EG95) which seems to confer 96 to 98 per cent protection against challenge infection. Recent trials in Australia and Argentina using this vaccine have reported that 86 per cent of immunized sheep were completely free of viable hydatid cysts when examined 1 year later. The number of viable cysts was reduced by 99.3 per cent. Although the results of these initial trials seem promising, further research is needed to assess the cost-benefit of using this vaccine.

Further reading

Allan JC *et al*. (1992). Coproantigen detection for immunodiagnosis of echinococcosis and taeniasis in dogs and humans. *Parasitology* **104**, 347–55.

Frider B, Larrieu E, Odriozola M (1999). Long-term outcome of asymptomatic liver hydatidosis. *Journal of Hepatology* **30**, 228–31.

Khuroo MS, Wani NA, Javid G (1997). Percutaneous drainage compared with surgery for hepatic hydatid cysts. *New England Journal of Medicine* **337**, 881–3.

Macpherson CNL *et al*. (1987). Portable ultrasound scanner versus serology in screening for hydatid cysts in a nomadic population. *Lancet* **ii**, 259–91.

Moro PL *et al*. (1997). Epidemiology of *Echinococcus granulosus* infection in the Central Andes of Peru. *Bulletin of the World Health Organization* **75**, 553–61.

Schantz PM, Williams JF, Posse CR (1973). Epidemiology of hydatid disease in southern Argentina. Comparison of morbidity indices, evaluation of immunodiagnostic tests, and factors affecting transmission in southern Rio Negro Province. *American Journal of Tropical Medicine and Hygiene* **22**, 629–41.

Thompson RCA, ed. (1986). *The biology of* Echinococcus *and hydatid disease*. George Allen and Unwin, London.

Verastegui M *et al*. (1992). Enzyme-linked immunoelectrotransfer blot test for the diagnosis of human hydatid disease. *Journal of Clinical Microbiology* **30**, 1557–61.

7.15.2Gut cestodes

R. Knight

Two groups of tapeworms infect man: the cyclophyllidean species which are covered in this chapter, and the pseudophyllidea (see Chapter 7.15.4).

The cyclophyllidean tapeworms maintain anchorage to the host small-gut mucosa by means of the scolex, a holdfast structure bearing a circlet of four suckers and usually a central evertible rostellum with one or more circlets of minute hooks (Fig. 1(a) and Fig. 1(b)). The rest of the body forms the strobila and consists of a chain of flattened proglottids, which bud behind the scolex. The worms change their site of attachment regularly, and are surprisingly motile. Gravid proglottids are lost from the end of the worm and are replaced by others that have matured as they pass down the strobila. Each proglottid possesses a complete set of hermaphroditic sex organs and marginal genital openings. Eggs accumulate in the uterus of gravid proglottids and only enter the faecal stream when the proglottids are disrupted. In many species the eggs enter the environment within intact proglottids. In either case the eggs are embryonated and contain a hexacanth embryo (onchosphere) that bears three pairs of hooks. The egg shells have two membranes; but in *Taenia* the outer is lost early and the inner forms the thick embryophore.

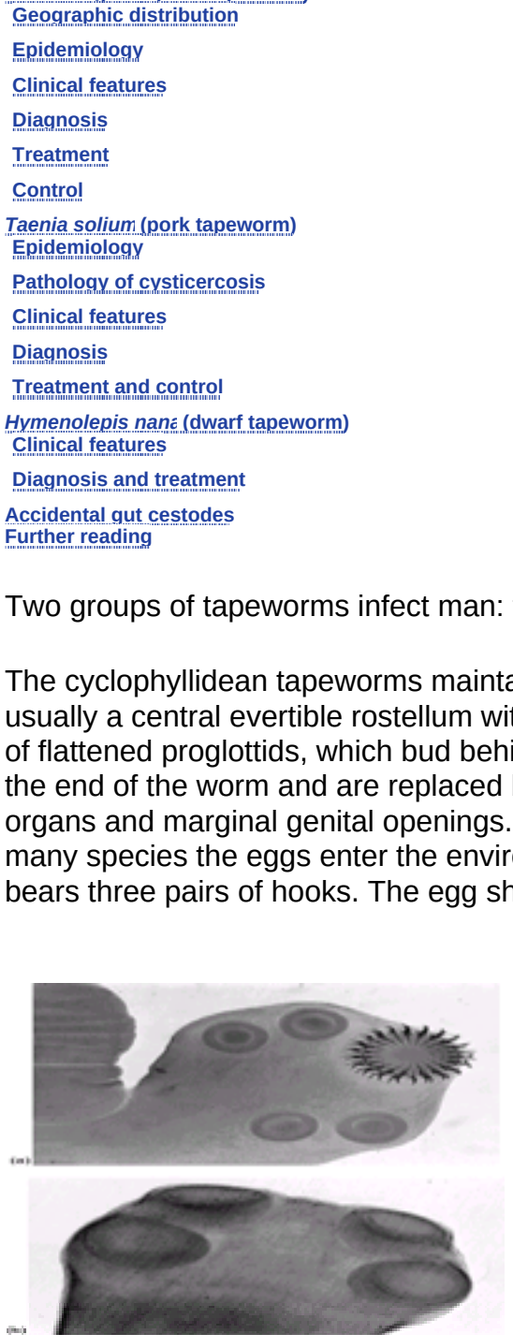

Fig. 1 (a) *Taenia solium* showing scolex with four suckers and a double row of hooks (× 250). (b) *Taenia saginata* showing scolex with four suckers and no hooks (× 250). (By courtesy of Professor V. Zaman.)

After ingestion by the intermediate host, eggs hatch and the released hexacanth embryos bore their way into the mucosa. The larval stages of the parasite are generally cystic with an invaginated embryonic scolex—the protoscolex. The cycle is completed when the larval stage, within the intermediate host or its tissues, is eaten by the definitive host; the protoscolex evaginates and attaches to the gut mucosa.

In three species, humans are an obligatory part of the lifecycle (Table 1; Fig. 2, Fig. 3, and Fig. 4), in the rest they are an accidental host (see Table 2). The two *Taenia* species cause anthropozoonoses because the cycle is maintained by an obligatory alternation between human and non-human hosts. Symptoms result from local hypersensitivity reactions to the worm and its scolex, and altered gut motility due to the physical mass of the worm. Patients often become aware of proglottids in their faeces. Some patients report poorly defined systemic symptoms, which may have an immunological basis. A blood eosinophilia up to 10 per cent can occur with any gut cestode.

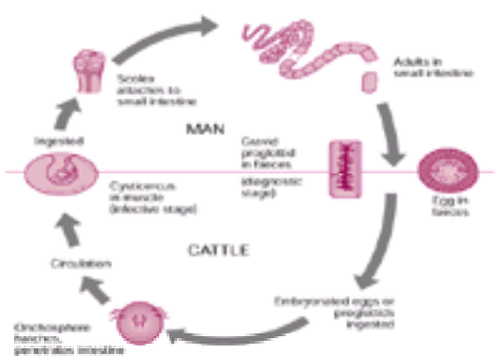

Fig. 2 Lifecycle of *Taenia saginata*. (Adapted by Professor V. Zaman from Centers for Disease Control, Atlanta, Georgia, USA.)

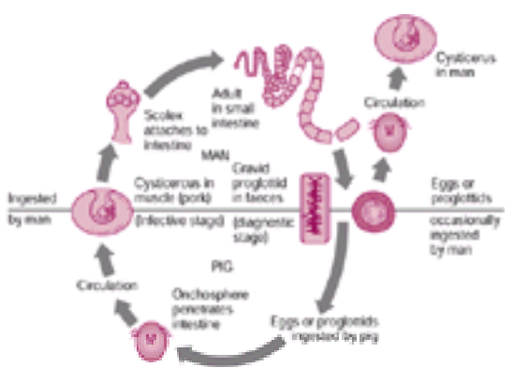

Fig. 3 Lifecycle of *Taenia solium*. (Adapted by Professor V. Zaman from Centers for Disease Control, Atlanta, Georgia, USA.)

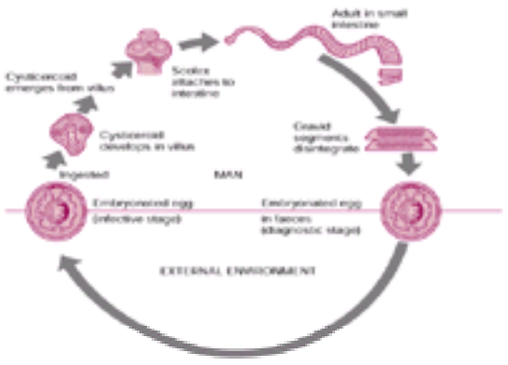

Fig. 4 Lifecycle of *Hymenolepsis nana*. (Adapted by Professor V. Zaman from Centers for Disease Control, Atlanta, Georgia, USA.)

Taenia saginata (beef tapeworm)

Geographic distribution

The beef tapeworm is prevalent where cattle have access to human faeces and where humans eat undercooked beef. The highest prevalence is in Africa, particularly in eastern and north-eastern parts; it is also common in many countries in the Middle East, South America, and South-East Asia. Prevalence is now very low in the United States, Canada, and Australia. It still persists endemically in Western Europe; but eastwards prevalence increases progressively across Europe and into the former USSR.

Epidemiology

Gravid proglottids are passed at defaecation, often in short chains; free eggs also occur in faeces. The whitish proglottids, approximately 2 to 3 cm long, are actively motile, elongating and contracting (Fig. 5). Viable eggs persist on pasture for many months and can survive most forms of sewage treatment. Cattle have access to human faeces on farms, at camp sites and recreation areas, and on railway lines. Infected herdsmen can initiate epizootics. Eggs may be dispersed by flies and dung beetles, and seabirds can ingest proglottids in estuarine waters and deposit them in their faeces on inland pastures.

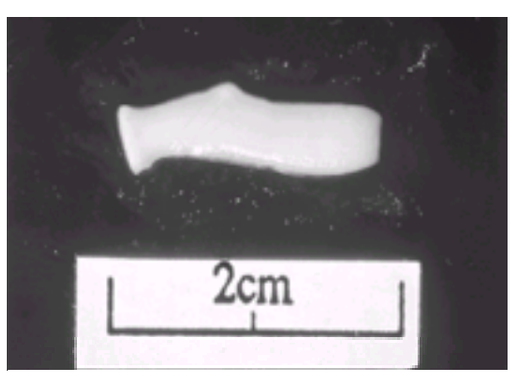

Fig. 5 Actively motile, contracting proglottid of *Taenia saginata* found by a patient in the stool. (Copyright D.A. Warrell.)

In cattle, cysticerci occur in striated muscle; they are whitish, ovoid, and measure 8 by 5 mm; they contain an invaginated protoscolex with no hooks. They become infective within 12 weeks and remain viable in the living host for 2 years; they are viable in stored, chilled meat for several weeks but are killed at –20°C within 1 week. The prepatent period in humans is 3 months and worms may live 30 years. Cattle develop protective immunity to new infection.

A subspecies *T. saginata asiatica* occurs in Taiwan, Korea, Indonesia, Thailand, and Burma. Infection follows ingestion of raw pig or wild boar liver; the protoscolex of the cysticercus bears hooks.

Clinical features

Most worms are solitary. Multiple worms are smaller, more common in high-transmission areas, and probably arise by simultaneous infection. Most patients are first aware of the worm by seeing proglottids on faeces (Fig. 5). Many will experience active worm migration through the anus, and this may induce an anxiety response. Many have no other symptoms, but others complain of nausea and upper abdominal pains, often relieved by food. A few patients eat to relieve symptoms. In children, impaired appetite can have nutritional consequences. Some patients have symptoms suggestive of hypoglycaemia, namely dizziness and sweating. Pruritus ani is common. The worm may be visible on small-bowel barium studies.

Proglottids have been found in a variety of surgical specimens including resected appendices, but a pathogenic role is usually difficult to establish. They occasionally obstruct the small intestine, pancreatic duct, or bile duct. After gut perforation they can occur in the peritoneum. Proglottids are recorded in the gallbladder, and eggs have been found in gallstones.

Diagnosis

The typical eggs may be found in faeces, but this is an insensitive method; perianal swabs are also useful. Eggs are indistinguishable from those of *T. solium*; patients should be asked to bring worm specimens. Unless the proglottid is fully gravid the number of uterine branches is an unreliable diagnostic character. A better morphological distinction is the presence of a vaginal sphincter; this is absent in *T. solium*. In human surveys in endemic areas a 24-h faecal collection after an anthelmintic will give the most reliable prevalence.

Treatment

A single morning dose of 2 g niclosamide is given to adults and older children on an empty stomach; the tablets should be chewed. Children of 2 to 6 years should receive 1 g, and those below 2 years, 500 mg. The alternative is praziquantel given in a single dose of 10 to 20 mg/kg after a light breakfast. After either drug the proximal part of the worm disintegrates in the gut and the scolex cannot be found. Failure of proglottids to reappear within 3 to 4 months indicates cure.

Control

This includes health education about raw beef, meat inspection, sanitation and hygiene on cattle farms, and proper sewage treatment and disposal. Mass treatments

of herd contacts, or whole adult populations, are the most effective short-term measures when endemicity is high. *T. saginata* causes great economic loss to the beef industry in some developing countries.

Taenia solium (pork tapeworm)

The clinical importance of the pork tapeworm relates mainly to cysticercosis, the occurrence of larval forms in human tissue (see Chapter 7.15.3). This arises when eggs hatch in the upper gut and humans become an accidental intermediate host. The source of such eggs is the faeces of persons infected with adult worms.

T. solium is generally less common than the beef tapeworm; it is now very rare in North America and Western Europe, but it remains common in much of sub-Saharan Africa, Mexico, South America, and in China, India, and other parts of Asia.

Epidemiology

In the pig muscle cysticerci produce 'measly pork' (Fig. 6). The cysts are most numerous in the tongue, masseter, heart, and diaphragm, but also occur in the brain. When eaten by humans in undercooked pork the worms mature in 5 to 12 weeks. The eggs have the same resistant qualities as *T.saginata*.

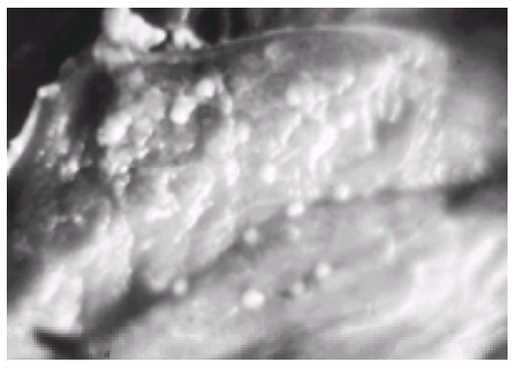

Fig. 6 'Measly pork' showing numerous cysts in the pig's muscle. (Copyright Sornchai Looareesuwan.)

Human cysticercosis is much more limited geographically than *T. solium* implying that internal autoinfection from disrupted proglottids is rare. Conditions favouring cysticercosis include poor personal hygiene, which facilitates external autoinfection and contaminated fingers among food handlers. Faecal pollution of the peridomestic environment, irrigation water, or cultivated vegetables is also important. In parts of Africa, tapeworm proglottids are used in traditional medicine. In the absence of these factors, cases of cysticercosis may be very sporadic even when *T. solium* is common. Cysticercosis is a major health problem in Mexico, some South American countries, and to a lesser extent in Africa and Asia. In 1969, *T. solium* was introduced from Bali into the highlands of Irian Jaya, New Guinea, where the disease is now of great importance.

Pathology of cysticercosis

Cysts occur especially in striated muscle, subcutaneous tissue, the nervous system, and the eye. Many remain clinically silent until the parasite dies after 3 to 5 years, when vigorous inflammatory and hypersensitivity reactions can occur; later lesions may calcify. In the brain, particularly in the subarachnoid and the ventricular system, atypical racemose cysts may occur. They appear as irregular or grape-like clusters of cysts that have no protoscolex; they can be mistaken pathologically for non-parasitic cysts.

Clinical features

Symptoms due to the adult worms are similar to those of *T. saginata* but are often milder and not associated with pruritus ani. The proglottids do not migrate actively *per anum*.

Diagnosis

Adult worm infection is detected as for *T. saginata*. Methods for detecting faecal antigen are available and have great potential use in epidemiological studies. Proglottid fragments can be identified using DNA probes.

Treatment and control

Adult worms are treated as for *T. saginata*. Vomiting must be avoided and an antiemetic is recommended, together with a purgative 2 h after the medication, which should be given after a light breakfast. It should be remembered that the faeces will be potentially highly infective for several days, both for the patient and attendants. Control measures are similar to those of *T. saginata* but local risk factors for human cysticercosis must receive special attention.

Hymenolepis nana (dwarf tapeworm)

The dwarf tapeworm is the most common cestode in man; it is also the smallest. When worm loads are high it causes more gut pathology than any other species. It is common in most developing and tropical countries. The lifecycle normally involves only humans (Fig. 4). Fully embryonated infective eggs are passed in the faeces; gravid proglottids normally disintegrate completely in the gut. Infection is commonly direct, but also by the other faecal–oral routes. Eggs hatch in the jejunum and the hexacanth embryo bores into a villus where it transforms into a cysticercoid larva. After 4 to 6 days it re-enters the gut, everts the scolex, and attaches to the mucosa; eggs appear in the faeces within 12 days. The lifespan is 3 months. The eggs are delicate and survive less than 10 days in the environment. Prevalence is usually much higher in children than adults; outbreaks can occur in families and institutions. External autoinfection is common in high-risk groups and enables high worm loads to build up. In addition, internal autoinfection occurs when there is gut stasis or retroperistalsis. Because of the importance of direct transmission, this infection may be common even in arid environments such as Western Australia.

Clinical features

In heavily infected people, especially children, up to 1000 or more worms may be present. Mucosal damage caused by both larval and adult worms leads to protein loss and sometimes malabsorption. Abdominal pains and anorexia are common.

Immunosuppressant or steroid therapy, particularly in patients with lymphoma, can lead to the development of bizarre cystic larval forms in the gut wall, mesenteric nodes, liver, and lungs. A similar condition can be produced in immunosuppressed mice.

Diagnosis and treatment

Eggs can be detected in faeces using concentration methods. Proglottids are rarely found in faeces, except after treatment.

Praziquantel in a single dose of 25 mg/kg is the most effective drug. If niclosamide is used, a 7-day course is needed to ensure that larval stages are killed when they re-enter the gut lumen. The dose on the first day is as for *T. saginata*; on the remaining days one-half of this dose is given. Relapses often result from persistence of

eggs in the patient's environment.

Accidental gut cestodes

Many species have been recorded in humans (see Table 2). All have arthropods as intermediate hosts, the larval cysticercoid stage being in the haemocele; the full lifecycles of some species are still uncertain. The normal definitive host becomes infected by eating the arthropod intentionally or accidentally. The means by which humans become infected is sometimes not clear, but fleas, small beetles, and mites are easily overlooked in food. *Dipylidium caninum* infection occurs in children who have groomed their pet. Infections with *Bertiella* are mostly in owners of pet monkeys, but oribatid mites are common in fallen fruit, especially mangoes. Children may eat insects deliberately, and this appears to be the mode of infection by *Raillietina siriraji* in Bangkok. Beetles are used for medicinal purposes in parts of Thailand and Malaysia, and this is the most likely route by which *Mathevotaenia* is acquired.

Hymenolepis nana fraterna is the murine strain of the human parasite and *H. diminuta* also infects rodents; both are rare in human beings. Both human and murine subspecies of *H. nana* will infect *Tribolium* beetles. In many of these species the eggs are in capsules that are released when the proglottid disintegrates in the gut, or more commonly, in the faecal mass. *Mesocestoides* is unique among these parasites in three respects: two intermediate hosts are required; the genital opening is medioventral rather than at the margin of the proglottid as in all other cyclophyllidean tapeworms; and larval worms can occur in man when mites are ingested.

Many patients will present because they have passed proglottids. *Dipylidium caninum* actively migrates out of the anus, like *T. saginata*. Faecal examinations of persons with abdominal complaints may reveal unusual eggs or egg capsules. Poorly defined systemic and allergic complaints are common. Treatment is as for *T. saginata*.

Recognition of these parasites is of epidemiological interest and may indicate potential transmission of other zoonotic pathogens. It is certain that all these parasites are underreported. Unusual proglottids or eggs should be preserved in formol saline and sent to a parasitologist.

Further reading

Chitchang S *et al.* (1985). Relationship between the severity of the symptom and the number of *Hymenolepis nana* after treatment. *Journal of the Medical Association of Thailand* **68**, 424–6.

Fan PC (1988). Taiwan *Taenia* and taeniasis. *Parasitology Today* **4**, 86–8.

Fan PC (1997). Annual economic loss caused by *Taenia saginata asiatica* taeniasis in three endemic areas of east Asia. *Southeast Asian Journal of Tropical Medicine and Public Health* **28**(Suppl 1), 217–21.

Fan PC *et al*. (1995). Morphological description of *Taenia saginata asiatica* (Cyclophyllidea: Taeniidae) from man in Asia. *Journal of Helminthology* **69**, 299–303.

Flisser A (1988). Neurocysticercosis in Mexico. *Parasitology Today* **4**, 131–7.

Flisser A *et al*. (1990). New approaches in the diagnosis of *Taenia solium* cysticercosis and taeniasis. *Annales de Parasitologie Humaine et Comparée* **65**(Suppl 1), 95–8.

Harrison LJ (1990). Differential diagnosis of *Taenia saginata* and *Taenia solium* with DNA probes. *Parasitology* **100**, 459–61.

Lucas SB *et al*. (1979). Aberrant forms of *Hymenolepis nana*: possible opportunistic infections in immunosuppressed patients. *Lancet* **ii**, 1372–3.

Mason PR, Patterson BA (1994). Epidemiology of *Hymenolepis nana* in primary school children in urban and rural communities in Zimbabwe. *Journal of Parasitology* **80**, 245–50.

Pawlowski Z, Schultz MG (1972). Taeniasis and cysticercosis (*Taenia saginata*). *Advances in Parasitology* **10**, 269–343.

Subianto DB, Tumada LR, Morgono SS (1978). Burns and epileptic fits associated with cysticercosis in mountain people of Irian Jaya. *Tropical and Geographic Medicine* **30**, 275–8.

7.15.3 Cysticercosis

Hector H. Garcia and Robert H. Gilman

Introduction

Known since the Hippocratic era, cysticercosis is the commonest helminthic infection of the human central nervous system. It is probable that the suspicion of its origins led some religions expressly to forbid the consumption of pork. Socio-economic improvements eradicated the infection in Europe and North America. However, endemic *Taenia solium* taeniasis/cysticercosis persists in most developing countries, where human cysticercosis is an important cause of epilepsy and other neurological morbidity, and porcine infections cause important economical losses to peasants.

Aetiology

Cysticercosis is infection with the larval stage (cysticercus) of *T. solium*, the pork tapeworm. In the lifecycle (Fig. 1) of this two-host zoonotic cestode, humans are the only definitive host and harbour the adult tapeworm, whereas pigs are intermediate hosts. The hermaphroditic adult *T. solium* inhabits the small intestine. Its head or scolex bears four suckers and a double crown of hooks, connected by a narrow neck to a large body (strobila) between 2 and 4 m long, composed of several hundred proglottids. Gravid proglottids, each containing 50 000 to 60 000 fertile eggs, detach from the distal end of the worm and are excreted in the faeces. The cycle is completed when pigs ingest stools contaminated with *T. solium* eggs. Once ingested by the pig, the invasive oncospheres in the eggs are liberated by the action of gastric acid and intestinal fluids and actively penetrate the bowel wall, enter the bloodstream, and are carried to the muscles and other tissues where they develop into larval cysts. When humans ingest undercooked pork containing cysticerci, the larvae evaginate in the small intestine, their scolices attach to the intestinal mucosa, and they begin forming proglottids. By accidentally ingesting *Taenia* eggs, humans may also act as intermediate hosts for *T. solium* and develop cysticercosis.

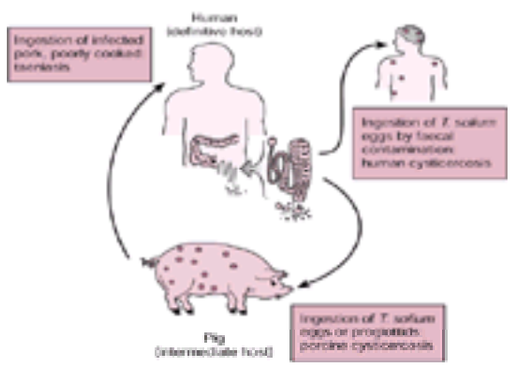

Fig. 1 Lifecycle of *T. solium*.

Epidemiology

The availability of neuroimaging studies and the subsequent development of specific serodiagnostic tests resulted in identification of neurocysticercosis as a frequent neurological disorder in Latin America, Africa, and Asia, where the prevalence of active epilepsy is almost twice that in Western countries. Neurocysticercosis is an emerging problem in industrialized countries, seen mainly in immigrants from endemic areas, some of whom may spread the infection as tapeworm carriers.

The main sources of human cysticercosis are ingestion of food contaminated with *T. solium* eggs and faecal–oral contamination in those carrying the tapeworm. Epidemiological studies suggest that almost every newly diagnosed patient with cysticercosis has been infected by someone in their close environment who is harbouring a *T. solium* and tends to dismiss the role of environment or water in transmission. Airborne transmission of *T. solium* eggs or internal autoinfection by regurgitation of proglottids into the stomach have been suggested but not proved.

Pathogenesis

Any organ may be infected but parasites survive more frequently in the nervous system, possibly because the immune response there is limited. Signs and symptoms are caused by perilesional inflammation and oedema, mass effect, or obstruction of cerebrospinal fluid circulation. Although complete development of cysts takes about 2 months, symptoms usually develop years after the initial infection. This clinically silent period and finding inflammation around cysts in symptomatic cases suggest that symptoms are due to inflammatory processes associated with death of the parasite rather than to the presence of the parasite itself.

Meningeal cysticerci elicit an intense inflammatory reaction causing thickening of basal leptomeninges. The optic chiasma and other cranial nerves are usually entrapped within this dense exudate, resulting in visual field defects and other cranial nerve dysfunctions. The foramens of Luschka and Magendie may be occluded by the thickened leptomeninges leading to hydrocephalus. Blood vessels may be affected by the inflammatory reaction. The walls of small penetrating arteries are invaded by inflammatory cells, leading to a proliferative endarteritis with occlusion of the lumen, and which may result in cerebral infarction.

Clinical features

Neurocysticercosis is a pleomorphic disease, whose manifestations vary with the number, size, and topography of the lesions and the intensity of the host's immune response to the parasites. Patients can be classified by the number and location of the cysticerci, and the presence or absence of associated inflammation or calcifications.

Epilepsy, the most common presentation of neurocysticercosis, is usually the primary or sole manifestation of the disease. Seizures occur in 50 to 80 per cent of patients with parenchymal brain cysts or calcifications but are less common in other forms of the disease. Other common focal signs include pyramidal tract signs, sensory deficits, signs of brainstem dysfunction, and involuntary movements. These manifestations usually follow a subacute or chronic course, making neurocysticercosis difficult to differentiate clinically from neoplasms or other infections of the central nervous system. Focal signs may occur abruptly in patients who develop a cerebral infarct as a complication of subarachnoid neurocysticercosis. Subarachnoid cysticerci may reach 10 cm or more in diameter ('giant' cysticercosis,

Fig. 2), and exert a mass effect.

Fig. 2 Giant cysticercotic cyst (brain CT).

Neurocysticercosis may present with increased intracranial pressure, usually from hydrocephalus secondary to cysticercotic arachnoiditis, granular ependymitis, or ventricular cysts. In these cases, intracranial hypertension develops subacutely and progresses slowly. An encephalitic picture may result from overwhelming inflammation around many parasitic cysts, a syndrome that occurs more frequently in younger people, especially women. In contrast, some patients may tolerate hundreds of intraparenchymal cysticerci with only minor symptoms.

Muscular pseudohypertrophy, a rare presentation, is caused by heavy cysticercal infection of skeletal muscles (Fig. 3) giving a 'Herculean' appearance. The few cases reported are all from India. Other apparent differences in clinical manifestations between Asia and Latin America include a high frequency of subcutaneous cysts and single degenerating brain lesions in Asia.

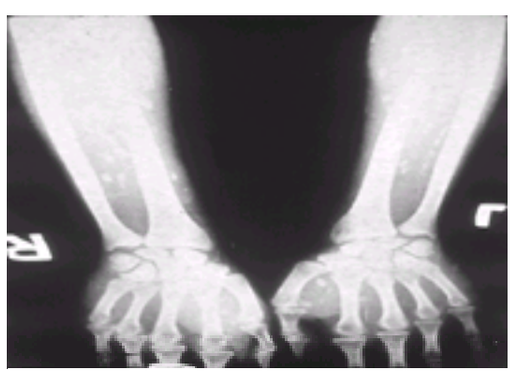

Fig. 3 Heavy cysticercal infection of skeletal muscles (copyright Sornchai Looareesuwan).

Pathology

The cysticerci are liquid-filled vesicles consisting of vesicular wall and scolex (Fig. 4). The vesicular wall is composed of an outer or cuticular layer, a middle or cellular layer with pseudoepithelial structure, and an inner or reticular layer. The invaginated scolex has a head or rostellum armed with suckers and hooks, and a rudimentary body or strobila that includes the spiral canal.

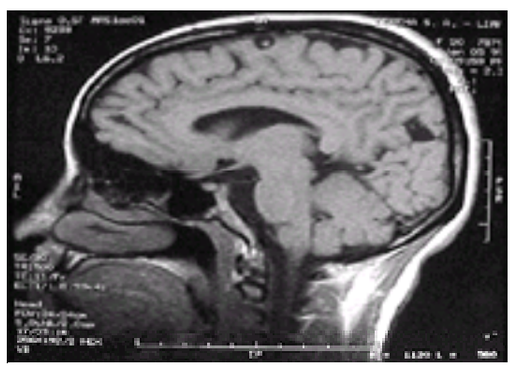

Fig. 4 Uncontrasted T_1 MR image showing two intraparenchymal cysticerci with visible scolices.

The macroscopic appearance of cysticerci varies in different locations within the central nervous system. Cysticerci within the brain parenchyma are usually small and tend to lodge in the cerebral cortex or basal ganglia. Subarachnoid cysts may be small if located in the depths of cortical sulci, or grow to 5 cm or more in the basal cisterns or sylvian fissures. Ventricular cysticerci are usually single, may or may not have a visible scolex, and may be attached to the choroid plexus or float freely in the ventricle. Spinal cysticerci are usually located in the subarachnoid space (rarely intramedullary). Their morphology is similar to cysts located within the brain.

Basal cysticerci may undergo a disproportionate growth of their membrane, with extension processes, resembling a brunch of grapes (racemose cysticercosis, Fig. 5). In these cases, the scolex is frequently unidentifiable even by microscopy.

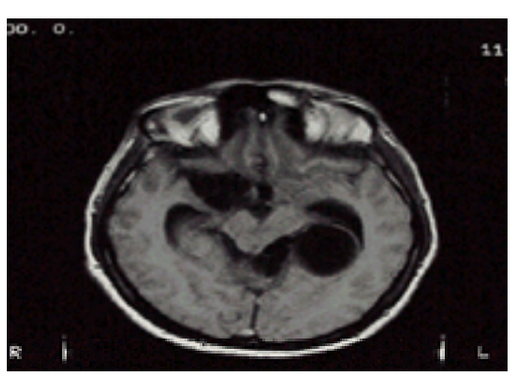

Fig. 5 Basal 'racemose' cysticercosis.

Viable, vesicular cysticerci elicit little inflammatory change in surrounding tissues because of active immune evasion mechanisms. The appearance of symptoms is interpreted as the result of immunological attack from the host, in a process of degeneration that ends with the death of the parasite. Inflammatory changes in the parasite membrane and increased density of cyst fluid mark the transition between four defined stages: viable, colloidal, granular nodular, and calcified cyst. Viable cysts may coexist with degenerating cysts or calcifications.

Laboratory/imaging diagnosis

The pleomorphism of neurocysticercosis makes it impossible to diagnose on clinical grounds alone. In endemic regions, late-onset seizures in otherwise healthy individuals are highly suggestive of neurocysticercosis. Most of these patients are normal on neurological examinations. Routine neuroimaging and serological studies are, therefore, mandatory. Finding cysticerci outside the central nervous system (eye, subcutaneous tissue, muscle) assists the diagnosis of neurocysticercosis. Muscular and subcutaneous cysticerci are far less common in American than in African or Asian patients with neurocysticercosis.

Neuroimaging

CT and MRI have drastically improved diagnostic accuracy by providing objective evidence about the topography of the lesions and the degree of the host inflammatory response to the parasite. Imaging findings in parenchymal neurocysticercosis depend on the stage of involution of cysticerci. Viable cysticerci appear as rounded cystic lesions on CT (Fig. 2), hypointense on MRI (Fig. 4), without associated enhancement, whereas degenerating parasites are seen as focal enhancing lesions surrounded by oedema, and calcifications as hyperdense dots or nodules (Fig. 6). Disappearance of cyst fluid signals the degenerative phase and calcified nodules the residual phase. Single or multiple ring-like or nodular enhancing lesions are non-specific and present a diagnostic challenge. Pyogenic brain abscesses, fungal abscesses, tuberculomas, toxoplasma abscesses, and primary or metastatic brain tumours may produce similar findings on CT or MRI.

Fig. 6 Calcified neurocysticercosis.

CT and MRI findings in subarachnoid neurocysticercosis are less specific. They include hydrocephalus, abnormal meningeal enhancement, and subarachnoid cysts. Cerebral angiography may show segmental narrowing or occlusion of major intracranial arteries in patients with cerebral infarcts secondary to parasitic vasculitis. In neurocysticercosis there is rarely fever or signs of meningeal irritation; glucose levels of cerebrospinal fluid are usually normal. MRI is generally better than CT for the diagnosis of neurocysticercosis, particularly in patients with basal lesions, brainstem or intraventricular cysts, and spinal lesions. MRI is, however, less sensitive than CT for the detection of small calcifications.

Immunological tests

Immunoblot (Western blot) is the best available serological test for *T. solium* antibodies. It is 98 per cent sensitive in cases with more than one active lesion, and 100 per cent specific. Its sensitivity may drop in patients with a single cyst. Other assays using unfractionated antigens (e.g. enzyme immunoassay) suffer from poor specificity but are more reliable when performed with cerebrospinal fluid than serum. Antigen-detection tests may provide a tool for serological monitoring of antiparasitic therapy. Although results of serology and imaging studies may be similar, they evaluate different aspects of the disease and may be discordant in some patients. Intestinal tapeworm carriers, naturally cured patients, or non-neurological infections may have normal brain images but be positive serologically. Those with only inactive lesions or a single cerebral lesion may be seronegative.

A proportion (about 10 to 15 per cent) of patients with neurocysticercosis are tapeworm carriers at the time of diagnosis, and in another 10 per cent or so a carrier can be detected in the household. Parasitological diagnosis is difficult: eggs and proglottids are shed only intermittently in stool and are usually missed by routine stool examination. Stool assays to detect parasite antigens are more sensitive than microscopy, but are not widely available. A recently described serological test for tapeworm carriers may improve detection.

A set of diagnostic criteria based on neuroimaging studies, serological tests, clinical presentation, and exposure history has been proposed by Del Brutto and colleagues. Besides absolute demonstration of the presence of the parasite, 'major' criteria (including typical findings on neuroimaging, demonstration of specific anticysticercal antibodies, or the presence of typical cigar-shaped calcifications in muscle) are combined with 'minor' criteria and epidemiological data to suggest a probable or possible diagnosis. Application of these criteria should improve the consistency of diagnosis.

Treatment

Because of the clinical and pathological pleomorphism of neurocysticercosis, precise assessment of the viability and size of cysts, the location of parasites, and the severity of the host's immune response is important before planning treatment.

Symptomatic treatment is very important. Seizures secondary to parenchymal neurocysticercosis can usually be controlled with anticonvulsants. However, the optimal length of anticonvulsant therapy in patients with neurocysticercosis has not been determined, and it is difficult to withdraw this treatment. Prognostic factors associated with recurrence of seizures include the development of parenchymal brain calcifications, and occurrence of recurrent seizures or multiple brain cysts before starting antiparasitic therapy.

Antiparasitic agents destroy viable cysts, although their long-term clinical benefit in seizures due to parenchymal neurocysticercosis has not been proved. Currently, albendazole is the drug of choice for antiparasitic treatment of cerebral cysticercosis (15 mg/kg.day for 7 days, with steroids), although a recently described single-day praziquantel regimen (75 to 100 mg/kg, in three doses at 2-h intervals, followed by steroids 6 h later) demonstrated similar cestocidal activity with few cysts. Longer courses may be required in patients with many lesions or subarachnoid cysticercosis. Transient worsening of neurological symptoms can be expected during antiparasitc therapy, secondary to the perilesional inflammatory reaction. There is no role for antiparasitic drugs in inactive neurocysticercosis (i.e. calcifications with or without enhancement on CT scan) since the parasites are dead.

Between the second and fifth day of antiparasitic therapy there is usually an exacerbation of neurological symptoms, attributed to local inflammation caused by the death of the larvae. For this reason, albendazole or praziquantel are generally given simultaneously with steroids in order to control the oedema and intracranial hypertension. Serum levels of praziquantel decrease when steroids are administered simultaneously, an effect that does not occur with albendazole. However, there is no evidence that cysticidal efficacy is decreased. Serum levels of phenytoin and carbamazepine may be lowered by simultaneous praziquantel administration. There are no data in patients receiving albendazole.

Some forms of neurocysticercosis should not be treated with antiparasitic agents. In patients with severe cysticercotic encephalitis these drugs may result in

worsening cerebral oedema and fatal herniation. In this case, the mainstay of therapy is high doses of corticosteroids to decrease the inflammatory response. In patients with both hydrocephalus and parenchymal brain cysts, antiparasitic drugs should be started only after placement of a ventricular shunt in case the intracranial pressure increases as a result of drug therapy. Antiparasitic drugs must be used with caution in patients with giant subarachnoid cysticerci. In such patients, concomitant steroid administration is mandatory to avoid cerebral infarction. Albendazole can successfully destroy ventricular cysts, but the surrounding inflammatory reaction may cause acute hydrocephalus if the cysts are located within the fourth ventricle or near the foramens of Monro and Luschka.

Surgery is limited to ventriculoperitoneal shunts to relieve obstructive hydrocephalus, and excision of single cysticerci (in the fourth ventricle or giant intraparenchymal cysts). However, shunts frequently dysfunction. The protracted course in these patients and their high mortality rates (up to 50 per cent in 2 years) is directly related to the number of surgical interventions required to change the shunts. Recently, neuroventriculoscopy has been employed as a less invasive option for resection of ventricular cysticerci.

Prognosis

Parenchymal cysticercosis has a good prognosis. Seizures usually subside in time without sequelae. In contrast, extraparenchymal cysticercosis and especially racemose cysticercosis have a poor prognosis, responding poorly to antiparasitic therapy, and leading to progressively deteriorating disease and death.

Prevention and control

Cysticercosis would not exist if pigs had no access to human faeces. However, this approach is hampered in endemic zones by the lack of sanitary facilities, veterinary inspection, and more importantly, because farmers tend to raise pigs under free-range conditions in order to reduce the cost of feeding them. Intervention programmes have concentrated on mass chemotherapy to eliminate human taeniasis, but their results have not been sustained. New tools for controls are oxfendazole, an effective and cheap single-dose therapy for porcine cysticercosis, and the candidate porcine vaccines under trial by several groups.

Monitoring the effect of an intervention requires suitable indicators. Human seroprevalence does not reflect changes in infection patterns because antibodies persist for years, even after successful treatment. Studies in Peru have shown that serological monitoring of porcine infection is a useful marker for both prevalence and changes in infection intensity over time. Similarly, the rate of infection in uninfected (sentinel) pigs over time can be used to estimate intensity of *T. solium* infection in the community. The prevalences of human and porcine infection are strongly correlated.

Areas of uncertainty/controversy

Although most cysts disappear after antiparasitic treatment, it remains uncertain whether this is associated with better control of seizures. Retrospective trials suggested that seizures were better controlled in treated patients. An open-label controlled trial failed to find a beneficial effect for albendazole or praziquantel in either clinical control or radiological evolution. However, the methodology of this study has been questioned, and data from double-blind randomized studies are not yet available.

Areas needing further research

A recent report suggests an association between brain calcifications secondary to cysticercosis and glial neoplasms. This has not yet been confirmed or rejected. Systematic long-term evaluation is needed to determine whether hydrocephalus is a late complication of antiparasitic therapy. The efficacy and costs of comprehensive human–porcine eradication programmes must be assessed.

Further reading

Corona T *et al*. (1996). Single-day praziquantel therapy for neurocysticercosis. *New England Journal of Medicine* **334**, 125.

Del Brutto OH (1997). Albendazole therapy for subarachnoid cysticerci: clinical and neuroimaging analysis of 17 patients. *Journal of Neurology, Neurosurgery and Psychiatry* **62**, 659–61.

Del Brutto OH *et al*. (2001). Proposed diagnostic criteria for neurocysticercosis. *Neurology* **57**, 177–83. [A guide to systematic diagnosis.]

Evans C *et al*. (1997). Controversies in the management of cysticercosis. *Emerging Infectious Diseases* **3**, 403–5.

Garcia HH, Martinez SM, eds (1999). Taenia solium *taeniasis/cysticercosis*, 2nd edn. Ed. Universo, Lima.

Garcia HH *et al*. (1993). Cysticercosis as a major cause of epilepsy in Perú. *Lancet* **341**, 197–200.

Garcia HH *et al*. (1997). Albendazole therapy for neurocysticercosis: a prospective double-blind trial comparing 7 versus 14 days of treatment. Cysticercosis Working Group in Peru. *Neurology* **48**, 1421–7.

Gonzalez AE *et al*. (1997). Treatment of porcine cysticercosis with oxfendazole: a dose–response trial. *Veterinary Record* **141**, 420–2.

White AC, Jr (1997). Neurocysticercosis: a common cause of neurologic disease worldwide. *Clinical Infectious Diseases* **24**, 101–13. [Comprehensive review.]

Wilkins PP *et al*. (1999). Development of a serologic assay to detect *Taenia solium* taeniasis. *American Journal of Tropical Medicine and Hygiene* **60**, 199–204.

7.15.4 Pseudophyllidean tapeworms: diphyllobothriasis and sparganosis

Seung-Yull Cho

Diphyllobothriasis

Diphyllobothriasis is a fish-borne infection of the intestine with tapeworms that belong to the genus *Diphyllobothrium*. The type species is *D. latum*.

Plerocercoid larvae of *D. latum* in fish can infect humans. In the intestine, the 1 cm long plerocercoid develops into a 5–6 m long adult, which produces a million eggs each day. In fresh water, a cycle is maintained—the egg embryonates to a coracidium, which becomes a procercoid larva in the copepod *Cyclops strenuus*, and then a plerocercoid in fish.

Human infections occur worldwide. The incidence is high in Siberia and in Baltic countries such as Finland. In Switzerland, the Lake Regions of North America, and in East Asia, cases of diphyllobothriasis are not uncommon. Humans may also be infected by zoonotic species of *Diphyllobothrium*. For instance, *D. yonagoense* and *D. nihonkaiense* in Japan and *D. pacificum* in Chile and Peru are intestinal parasites of seals while *D. ursi* and *D. dendriticum* in Alaska are parasites of bears and birds respectively. The habit of eating sliced raw fish, such as pike, burbot, perch, salmon, and other freshwater fish, creates the opportunity for infection. Prevention is achieved by freezing fish for 1 day at – 18 °C or lower.

Infection usually causes few symptoms. Abdominal discomfort, fatigue, diarrhoea, and urticaria may be the vague presenting symptoms. Vomiting up a tapeworm and intestinal obstruction due to a mass of worms occurs very rarely. A strip of gravid segments may pass out through the anus. Tapeworm pernicious anaemia may be associated with *D. latum* infection. In these patients, elimination of the tapeworm results in improvement of the anaemia. Clinically, haematological and neurological manifestations are the same as in pernicious anaemia.

Clinical symptoms are rarely responsible for raising the suspicion of diphyllobothriasis. The diagnosis can be confirmed by identifying eggs in the stool by microscopy. Discharged chains of gravid segments are also diagnostic. In endemic areas, all patients with pernicious anaemia should have their stools examined. Treatment is simple and effective. Niclosamide in a single adult dose of 2 g or praziquantel in a single dose of 10 mg/kg body weight are the drugs of choice.

Sparganosis

Sparganosis is a zoonotic infection caused by the larval tapeworm of *Spirometra mansoni* or *S. mansonoides.* The larvae invade a variety of human tissues (Fig. 1).

Fig. 1 A sparganum surgically removed from a subcutaneous mass.

The sparganum (plerocercoid) is a 1–30 cm long, white, slender tapeworm larva without round suckers. It is found in terrestrial vertebrates. Carnivorous mammals are infected with the adult stage in their small intestine. In fresh water, the egg embryonates, becoming a coracidium. The swimming coracidium is taken up by zooplankton, such as *Cyclops leuckarti*, and develops into a procercoid larva. When terrestrial vertebrates including man ingest the procercoid, it transforms into a tissue-invading sparganum.

Human sparganosis occurs sporadically worldwide. The procercoid larva in *Cyclops* can be inadvertently drunk in unfiltered water. Sparganum-infected frog, snake, poultry, or pork meat are important sources of human infection in endemic areas such as Japan, Korea, China, Vietnam, and Southeast Asian countries due to traditional habits. Some people believe that eating raw meat is a tonic or is beneficial for patients with tuberculosis. Rural people in Vietnam practise applying poultices of frog or snake skin to an inflamed eye. In this case a sparganum in the frog or snake skin can directly penetrate the conjunctiva.

Ingested larvae penetrate the intestinal wall and migrate systemically. The worm usually lodges in subcutaneous tissue or muscle of the chest or abdominal walls, breast, limbs, or scrotum. A lump may appear and then spontaneously disappear, only to reappear some weeks or months later at a site remote from the first. The sparganum secretes at least six different serine and cysteine proteases, which facilitate worm migration and evasion of the host's immune response. Orbital, chest, and abdominal cavities are involved. Sparganosis of the central nervous system is increasingly recognized.

A granuloma is formed along the tortuous migration track. Zones of necrosis and intense lymphohistiocytic reaction with eosinophilic infiltration surround the larva and its track. The disintegrated worm may be found in a granuloma, leaving behind calcareous corpuscles. Local bleeding and suppuration may complicate sparganosis. The sparganum can survive for more than 5 years. In general, one or only a few infect each patient.

Sparganum proliferum is an acephalic, branched, proliferating larva that is histologically similar to a non-proliferating sparganum. In very rare human infections the larvae are found in thousands in subcutaneous tissue and internal organs. The human infection has been found in Japan, the United States, and Venezuela. The biology of this larva is still unknown. The patient's serum reacts with sparganum antigen.

Diagnosis of sparganosis is rarely made preoperatively. Incidental recovery of the worm at surgery makes a definitive diagnosis. Preoperative diagnosis of cerebral sparganosis is made with high confidence when computed tomography or magnetic resonance imaging of the brain shows an enhancing nodule with changing shape or position in the sequential images, degeneration of white matter, and ventricular dilatation together with positive antibody tests in serum and cerebrospinal fluid (Fig. 2).

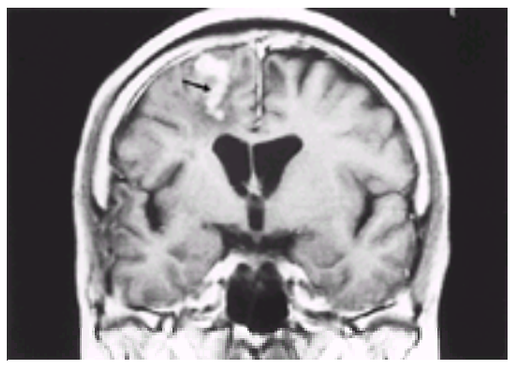

Fig. 2 A MRI finding of cerebral sparganosis. Coronal contrast-enhanced T_1-weighted image shows a tortuous curvilinear enhancing lesion (arrow) with surrounding low density of oedema and degeneration in right frontal lobe.

Excision of the mass or removal of the worm from the lesion is curative. Repeated surgery is necessary when the patient has multiple lesions. There are no drugs which are are known to be effective against sparganosis. The prognosis is excellent in almost all cases when treated surgically. However, all cases of *S. proliferum* infection have proved fatal.

Further reading

Kim DG *et al.* (1996). Cerebral sparganosis: clinical manifestations, treatment, and outcome. *Journal of Neurosurgery* **85**, 1066–71.

Moulinier R *et al.* (1982) Human proliferative sparganosis in Venezuela: report of a case. *American Journal of Tropical Medicine and Hygiene* **31**, 358–63.

Von Bonsdorff B (1977). *Diphyllobothriasis in man*. Academic Press, London.

7.16.1 Schistosomiasis

D. W. Dunne and B. J. Vennervald

Introduction

Schistosomiasis, also known as bilharzia, is caused by infection with parasitic trematode worms (flukes) of the genus *Schistosoma*. Disease is usually associated with chronic infections contracted by exposure to fresh water containing infective cercarial larvae that penetrate intact skin and develop into blood-dwelling worms. Most human infections are caused by one of three species, *S. mansoni*, *S. haematobium*, or *S. japonicum*. Two species, *S. intercalatum* and *S. mekongi*, are less important. Schistosomiasis is patchily distributed in parts of South America, Africa, the Middle East, China, and Southeast Asia. An estimated 600 million people in 74 countries are at risk of infection and some 200 million are infected. Of these, the World Health Organization estimates that 120 million may be symptomatic, while 20 million are suffering severe consequences of infection. Although simple diagnosis and effective drug treatment is available for individual uncomplicated cases, the world disease burden caused by these parasites has increased from the estimated 114 million human infections in 1947. Diagnosis and treatment are often not available to exposed rural populations, and drug-based control programmes are hampered by the continued susceptibility of treated individuals, particularly children, to reinfection. Human schistosomiasis is most often an insidious and chronic disease with a range of pathological manifestations involving the intestine and liver, or the urogenital tract. Mortality estimates are difficult, but 20 000 to 200 000 deaths may be directly associated with schistosomiasis each year.

Parasite lifecycle

The schistosome lifecycle requires two host species: a 'definitive' vertebrate host, in which adult male and female worms develop and sexual reproduction occurs, and an 'intermediate' freshwater snail host, in which the parasite multiplies asexually. Transmission between these hosts is achieved by two different free-swimming larval stages. For species that infect man, miracidia hatch from eggs excreted in the faeces or urine of the vertebrate host, and then seek out and infect snails. Cercariae are released from the snail and are able actively to penetrate intact human skin. Different schistosome species have their own, often very restricted, range of snail hosts. Schistosomiasis is thus closely associated with particular freshwater habitats, and its geographical distribution is restricted by the availability of particular snail species. *S. mansoni* and *S. haematobium* are confined to aquatic snails (genera *Biomphalaria* and *Bulinus* respectively) that inhabit ponds, lakes, irrigation canals, slow-flowing streams, and rivers. *S. japonicum* is transmitted by amphibious snails of the genus *Oncomelania* that, in addition to a variety of freshwater habitats, are also present in damp soil and vegetation, such as paddy fields. Schistosomes that infect man can also infect other mammals. This is important in the transmission of *S. japonicum*, a zoonotic infection in which cattle, water buffalo, pigs, dogs, and rodents can act as reservoir hosts of the human parasite. *S. mansoni* infects a narrower range of mammals and only a few rodent species and baboons have any potential to act as occasional reservoirs. In nature *S. haematobium* is essentially specific to man. The sites of maturation of the adult worms vary between schistosome species, affecting both the transmission of the infection and its clinical sequelae.

Once shed from freshwater snails, cercariae (Fig. 1) live for about 24 h, but their effective period of infectivity is probably shorter under field conditions. Cercarial behaviour and the timing of their release enhance their chance of contacting their vertebrate host of choice. Light and increasing temperature trigger the release of *S. mansoni* and *S. haematobium* cercariae during the day and their tails are used actively to maintain their position near the water surface. *S. japonicum* cercariae are shed late in the day and are closely associated with the meniscus, perhaps reflecting their wider host range, as species specific for rodents are shed at night. Contact with skin triggers adherence mechanisms and proteolytic enzymes and muscular movements allow penetration of the skin in minutes. Penetration initiates transformation into a schistosomular larva, with loss of the tail and of the protective outer glycocalyx layer, and the addition of an extra lipid bilayer to the surface membrane of the parasite's syncytial outer tegument. This tegument now forms the main parasite–host interface and so has physiological and immunological functions vital to long-term survival in the hostile environment of the bloodstream. These include uptake of nutrients, response to injury, and surface adsorption of host antigens to provide an immunological disguise.

Fig. 1 The infective larva (cerceria) of *Schistosoma mansoni*, length approximately 200 µm. The head region has characteristic suckers; the muscular forked tail propels the free-swimming larva, but is discarded during skin penetration. This larva will develop into an adult worm in a human host.

Newly transformed schistosomula remain in the epidermis for several days before migrating, via the bloodstream, lungs, and systemic circulation, to the hepatic portal system. Here the schistosomula mature and differentiate into adult worms, pair, and migrate against the portal blood flow to the small venules draining the genitourinary tract (*S. haematobium*) or the large and, to a lesser extent, small intestine (*S. mansoni*, *S. japonicum*, *S. intercalatum*, *S. mekongi)*. Male and female worms are 1 to 2 cm long and morphologically distinct. Paired worms remain permanently coupled, with the shorter, flatter, more muscular male gripping the female in its gynaecophoric canal (Fig. 2). Worms ingest blood cells into their blind-ending bifurcated gut, producing a haematin-like pigment that is regurgitated into the blood. Adult worms have average lifespans in man of 3 (*S. haematobium*) to 7 (*S. mansoni*) years, although active infections are reported in individuals who have left endemic areas more than 20 years previously. Female worms start to produce eggs between 5 and 12 weeks after infection, at rates of 300 (*S. mansoni*) to 3000 (*S. japonicum*) per day. A few days after an egg is laid, a single miracidium develops within the rigid eggshell, the shape and size of which is characteristic for each species. *S. mansoni* (Fig. 3) and *S. haematobium* eggs are ellipsoid, 65 by 150 µm, the former having a lateral spine and the latter a terminal spine. *S. japonicum* eggs are more spherical, 70 by 90 µm, with a small lateral knob that is not always apparent microscopically. Embryonated eggs pass from the venules into the gut or bladder lumen. This is facilitated by host immune responses to secreted egg antigens, as egg excretion is inhibited in immunosuppressed experimental hosts. The passage of the eggs causes tissue damage, as does the granulomatous reactions to eggs that fail to escape from the bloodstream and get swept into the liver by the

portal blood flow.

Fig. 2 Adult worms of *S. mansoni*. The shorter male encloses the female in its gynaecophoric canal, the characteristic haematin-like pigment can be seen in the female worm's gut.

Fig. 3 Egg of *S. mansoni* containing a fully developed miracidium and showing the characteristic lateral spine of this species.

Eggs deposited in fresh water rapidly hatch in response to osmotic changes, releasing the miracidium. This ciliated and actively swimming larva lives for about 6 h, and is able to detect chemically the proximity of snails, modifying its swimming behaviour as it approaches a potential host. The parasite actively penetrates the snail's tissues and transforms into a primary sporocyst. Asexual replication gives rise to daughter sporocysts that migrate to the snail's hepatopancreas where cercariae are asexually generated within each sporocyst. Thus, snails infected with a single miracidium release cercariae that are all of the same sex. Cercariae are first released from snails 3 to 6 weeks after infection, depending on parasite species and ambient temperature. Infected snails can shed hundreds of cercariae daily over several months.

Distribution (Fig. 4)

Schistosomiasis is associated with poor living conditions and inadequate sanitation and water supply. Its distribution has changed over the last 50 years. In some areas sustained control strategies have been successful. However, environmental changes, development of water resources, population increases, and migration, have led to its spread into previously non-endemic areas or areas with a low rate of infection. *S. japonicum* and *S. haematobium* have decreased, whereas *S. mansoni* has increased to become the most prevalent and widespread species. *S. japonicum* has been controlled effectively in many areas and is now endemic only in China, where it is much reduced, Indonesia, the Philippines, and Thailand. *S. mekongi* is found in Kampuchea and Laos, while *S. intercalatum* is found in 10 countries within the rainforest belt of central Africa. *S. mansoni* is present in most countries of sub-Saharan Africa, and in Madagascar, the Nile delta and valley, as well as Saudi Arabia, Yemen, Oman, Libya, northern and eastern Brazil, Surinam, Venezuela, and some Caribbean islands. *S. haematobium* is widespread in sub-Saharan Africa and Madagascar, and is more prevalent than *S. mansoni* in North Africa and the Middle East. Information on the geographical distribution of schistosomiasis is available from the World Health Organization website: http://www.who.int/ctd/html/schistoepidat.html.

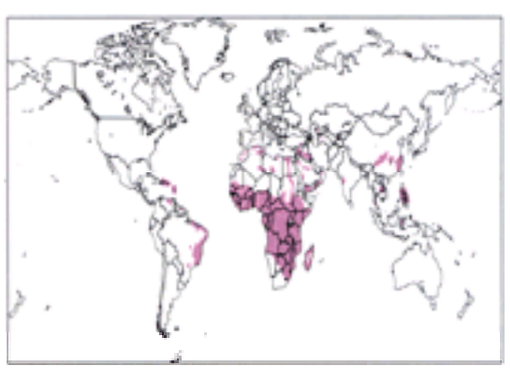

Fig. 4 Global distribution of the schistosomes that affect humans.

Clinical features

Stage of invasion: cercarial dermatitis or 'swimmer's itch'

When cercariae penetrate the skin they can cause a skin reaction, called cercarial dermatitis or 'swimmer's itch'. This is frequently seen after exposure to avian schistosomes, and is associated with the death of cercariae in the skin. It is seen both in areas endemic for human schistosomiasis and in non-endemic areas. In previously unexposed people, the invasion causes a transient immediate hypersensitivity reaction with intense itching. Within 12 to 24 h it is followed by a delayed reaction characterized by a small, red, pruritic, macular rash progressing to papules after 24 h. The rash may persist for up to 15 days and residual pigmentation may persist for months. Following repeated exposure, the signs and symptoms increase dramatically and start earlier. A similar reaction can be seen after re-exposure to human cercariae, predominantly *S. mansoni* and *S. japonicum.* Treatment, if needed, is symptomatic.

Stage of maturation: acute schistosomiasis or Katayama fever

The early stages of a primary infection can be associated with a severe systemic reaction that resembles serum sickness. This acute illness, called acute toxaemic schistosomiasis or Katayama fever, can occur following initial infection with any schistosome infecting humans, although it is more common in *S. japonicum* and *S. mansoni* infections. Acute schistosomiasis is most marked in primary infections in non-immune adults, but acute *S. japonicum* infection can occur in re-exposed individuals. Symptoms appear 2 to 6 weeks after exposure. The clinical picture resembles an acute pyrexial illness with fever as a prime characteristic. The patient feels ill, and may have rigors, sweating, headache, malaise, muscular aches, profound weakness, weight loss, and a non-productive irritating cough. Anorexia, nausea, abdominal pain, and diarrhoea can occur. Physical examination may reveal a generalized lymphadenopathy, an enlarged tender liver, and, sometimes, a

slightly enlarged spleen and an urticarial rash (Plate 1). Eosinophilia is almost always present. Patients may become confused or stuporose or present with visual impairment or papilloedema. Severe cerebral or spinal cord manifestations may occur, and this is an indication for urgent investigative measures. Even light infections may cause severe illness and the syndrome can, in rare cases, be fatal.

Differential diagnosis includes infections such as typhoid (leucopenia, no eosinophilia), brucellosis, malaria, infectious mononucleosis, miliary tuberculosis, leptospirosis, and other conditions with fever of unknown origin. Fever and eosinophilia occur in trichinosis, tropical eosinophilia, invasive ankylostomiasis, strongyloidiasis, visceral larva migrans, and infections with *Opisthorchis* and *Clonorchis* species.

Established infections

Urinary schistosomiasis (*Schistosoma haematobium*)

The signs and symptoms due to *S. haematobium* infection relate to the worms' predilection for the veins of the genitourinary tract, and result from deposition of eggs in the bladder, ureters, and to some extent the genital organs. In the phase of established infection two stages can be recognized:

a. An active stage mainly in children, adolescents, and younger adults with egg deposition in the urinary tract, egg excretion in the urine with proteinuria and macroscopic or microscopic haematuria.
b. A chronic stage in older patients with sparse or absent urinary egg excretion but the presence of urinary tract pathology.

In the active stage many patients will have minimal symptoms. The most frequently encountered complaint is a painless, characteristically terminal, haematuria, the prevalence and severity of which is related to the intensity of infection. In communities where *S. haematobium* is highly endemic, macroscopic haematuria among boys is considered a natural sign of puberty. Dysuria, frequency, and suprapubic discomfort or pain is associated with schistosomal cystitis and may continue throughout the course of active infection. Initially the eggs may give rise to an intense inflammatory response in the mucosa. This may cause ureteric obstruction leading to hydroureter and hydronephrosis. Cytoscopy reveals friable masses or polyps extending into the bladder, petechiae, and granulomas. These early inflammatory lesions, including the obstructive uropathy, are usually reversible after treatment with antischistosomal drugs. The bladder lesions and obstructive uropathy can be visualized by ultrasonography (Fig. 5).

Fig. 5 Bladder pseudopolyps as seen by ultrasound in *S. haematobium* infection. (Photograph by courtesy of Dr J. Richter, Heinrich-Heine-Universität Düsseldorf, Germany.)

As the infection progresses, the inflammatory component decreases, possibly due to modulation by the host immune response, and fibrosis increases. Various changes occur in the bladder including calcification, ulceration, and the development of papillomas. Cytoscopy reveals 'sandy patches' composed of large numbers of calcified eggs surrounded by fibrous tissue and an atrophic mucosal surface. The bladder lesions may lead to nocturia, precipitancy, retention of urine, dribbling, and incontinence. Calculus formation is common, as is secondary bacterial infection, usually due to *Escherichia coli*, *Pseudomonas*, *Klebsiella*, *Enterobacter*, or *Salmonella* species. The ureters are less commonly involved, but ureteric fibrosis can cause irreversible obstructive uropathy which can progress to uraemia. Bilateral ureteric involvement is common, although lesions may predominate on one side. Despite damage to the ureters, symptoms may be absent or minimal.

Egg deposition may also cause granulomas and lesions to develop in the genital organs, most commonly in the cervix and vagina in females and the seminal vessels in males. Dyspareunia, contact bleeding, and lower back pain may result in women, and perineal pain and painful ejaculation in males. Symptoms such as haematospermia and perineal discomfort have been described in travellers returning from Mali. In some of these patients, eggs have been demonstrated in seminal fluid but not in urine. The impact of genital lesions caused by *S. haematobium* infection on the spread of HIV needs to be elucidated. Although small numbers of *S. haematobium* eggs are frequently detected in faeces and rectal biopsies, intestinal symptoms are uncommon.

In some areas in Africa an association between *S. haematobium* infection and squamous cell carcinoma of the urinary bladder has been described. The aetiological significance of the parasite in the causation of this cancer is not proven, but is suggested by the finding that the prevalence of carcinoma of the bladder is correlated with intensity of *S. haematobium* infection. In the established stage *S. haematobium* should be distinguished from renal tuberculosis with haematuria, haemoglobinuria, and cancer of the urogenital tract.

Intestinal schistosomiasis

In most early *S. mansoni* and *S. japonicum* infections few, if any, minimal symptoms are apparent. Clinical features are generally encountered in those with high-intensity infections, and are diarrhoea, sometimes with blood or mucus in the stool, abdominal discomfort, and hypogastric pain or colicky cramps. Severe dysentery is rare, but can occur. The liver, especially the left lobe, may be enlarged and tender; the spleen may also be enlarged, but is usually soft. At this stage, the condition is entirely reversible by antischistosomal treatment, but the relative lack of symptoms may cause it to pass unnoticed until irreversible complications set in. Later stages present as intestinal or hepatosplenic disease. Intestinal schistosomiasis is associated with granuloma formation (Plate 2), inflammation, and fibrosis, primarily in the large intestine. Focal dense deposits of eggs of *S. mansoni* or *S. japonicum* in the large intestine can induce the formation of inflammatory polyps. The major clinical manifestation is intermittent diarrhoea with or without passage of blood or mucus, occasionally associated with protein-losing enteropathy and anaemia. Intestinal schistosomiasis in *S. japonicum* infection may also involve the stomach, with gastric bleeding and pyloric obstruction.

The differential diagnosis includes irritable bowel syndrome, amoebiasis, giardiasis, intestinal helminth infection, ulcerative colitis, Crohn's disease, and tuberculosis.

Hepatosplenic disease is the most severe chronic manifestation of *S. mansoni* and *S. japonicum* infection. The development of presinusoidal periportal fibrosis (clay pipe stem or Symmers' fibrosis) (Fig. 6) leads to portal hypertension, but hepatic function usually remains normal (Plate 3). Patients with periportal fibrosis may not excrete eggs in faeces. During the early stages the liver is enlarged, especially the left lobe; it is smooth, firm, and sometimes tender. Later, in many cases, it becomes small firm and nodular. The spleen is enlarged, often massively, due to passive congestion and reticuloendothelial hyperplasia (Fig. 7). The patient may be asymptomatic or may complain of a left hypochondrial mass with discomfort and anorexia. Anaemia may be present. Ascites, attributable both to the portal hypertension and to hypoalbuminaemia, may be seen, especially in *S. japonicum* infection. There may be reduced growth, infantilism, and amenorrhoea. Most importantly, 80 per cent of patients with hepatosplenic disease have oesophageal varices detectable by endoscopy. These patients may experience repeated bouts of haematemesis, melaena, or both. This is the most severe, potentially fatal, complication of hepatosplenic schistosomiasis, and death may result from massive loss of blood.

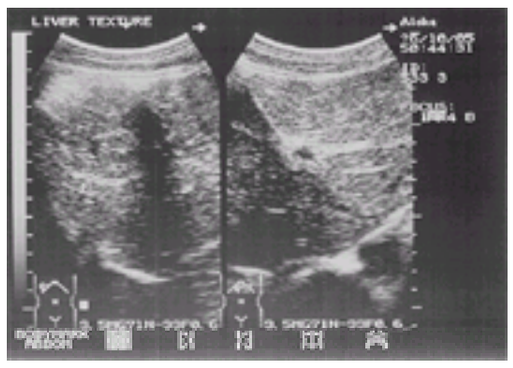

Fig. 6 Periportal fibrosis as seen by ultrasound in *S. mansoni* infection.

Fig. 7 Kenyan child with severe hepatosplenic schistosomiasis mansoni.

The differential diagnosis of hepatosplenic schistosomiasis includes kala-azar (visceral leishmaniasis), tropical splenomegaly syndrome associated with malaria, leukaemia, lymphoma, alcoholic, or viral cirrhosis, and some of the haemoglobinopathies. Some regression of periportal fibrosis may occur after specific antischistosomal therapy, as judged by ultrasonography examination of the liver, but in most individuals with periportal fibrosis and clinical manifestations of hepatosplenic disease, regression does not occur.

In comparison with *S. japonicum* and *S. mansoni* infections, clinical symptoms of disease in *S. intercalatum* infection are commonly mild or absent, and it is not regarded as a serious public health problem. Active infection is seen in children and adolescents and pathology is detected only in those with egg excretion exceeding 400 eggs per gram of faeces. The usual clinical presentation is one of diarrhoea, often with blood in the stool and lower abdominal pain or discomfort. *S. mekongi* infections are usually asymptomatic but may produce a clinical picture similar to that of *S. japonicum*, although the infections are usually milder. Hepatosplenomegaly can occur.

Other manifestations

Central nervous system manifestations

Central nervous system involvement in *S. mansoni* and *S. haematobium* infections most frequently affect the spinal cord following acute infection. This manifestation is not related to the intensity of infection. A myelopathy results from the inflammatory reaction, caused by the deposition of eggs around the spinal cord, and presents with ascending motor and sensory symptoms.The lesion is usually in the region of the canda equina. Although paraparesis is seen most commonly during acute schistosomiasis, it may also be a late stage complication of *S. mansoni* infection in endemic areas with high rates of transmission. Myelography, computed tomography, and magnetic resonance imaging are of diagnostic value. In acute cases lesions are seen on magnetic resonance imaging scans as a diffuse swelling of the lumbar cord with central softening or cyst formation.

The brain is the major site of central nervous system involvement in *S. japonicum* infections, with about 2 per cent of acutely infected patients experiencing symptoms that mimic acute encephalitis or a focal neurological process. Computed tomography shows multiple enhancing lesions. In chronic infections, patients may present with focal brain lesions that can resemble tumours and present as focal epilepsy. These lesions contain masses of eggs and granulomas. Uncontrolled studies suggest that treatment with a combination of antischistosomal drugs and glucocorticoids is effective.

Pulmonary manifestations

Deposition of eggs can also occur in the lungs. Granulomatous reactions and fibrosis develop in the pulmonary vasculature leading to pulmonary hypertension and/or cor pulmonale (Plate 4). This is normally seen secondary to hepatosplenic schistosomiasis in patients with portal fibrosis and portal hypertension, but pulmonary hypertension may also result from accumulation of *S. haematobium* eggs in the lungs. A syndrome of cough with multiple small radiographic lesions and eosinophilia has been described. Symptoms include fatigue, palpitations, dyspnoea, cough, and sometimes haemoptysis. Patients may progress to decompensation with congestive cardiac failure. In endemic areas schistosomiasis must always be considered as a possible cause of cor pulmonale.

Renal manifestations

Glomerulonephritis is a common occurrence in chronic *S. mansoni* infection in Brazil, especially in patients with hepatosplenic disease. Immunoglobulins, complement components, and schistosome antigens are deposited in the mesangial area. The condition is manifested clinically as proteinuria and/or nephrotic syndrome, sometimes with hypertension.

Miscellaneous manifestations

Patients infected with any of the three major schistosome species and subsequently infected with *Salmonella* may develop a prolonged intermittent febrile illness. Prolonged excretion of *Salmonella* in the urine and intermittent bacteraemia has been demonstrated in *S. haematobium* infection. Treatment for the *Salmonella* infection alone is often not effective without treatment of the underlying schistosome infection.

Diagnosis and investigations

Information about geographical area and history of exposure to potentially contaminated fresh water is important for diagnosis of schistosomiasis, especially in travellers. This can indicate the likelihood of infection and point to the schistosome species involved. A definitive diagnosis is made by the direct demonstration of schistosome eggs by microscopy of urine or stool samples, biopsies or, on rare occasions, secretions such as seminal fluid. In epidemiological studies it is usually important to obtain quantitative estimates of egg output to provide information about intensity of infection within a population.

Direct parasitological methods

In *S. haematobium* infection eggs can be detected in urine after filtration, sedimentation, or centrifugation followed by microscopy. Ideally, urine should be passed around midday and the terminal part of the stream examined. The most commonly used method in epidemiological studies in endemic areas is filtration of 10 to 20 ml

of urine using a syringe and a polycarbonate (Nucleopore®), polyamide (Nytrel®), or paper filter. Infection intensity is expressed as eggs per 10 ml of urine. This may not be sufficiently sensitive for detection of low-intensity infections in travellers. In such cases, diagnosis is often based on filtration of 24-h urine samples.

For *S. mansoni*, *S. japonicum*, *S. mekongi*, and *S. intercalatum* eggs in the faeces, sedimentation of the eggs followed by microscopy is a useful and simple technique. However, the Kato thick smear technique is the most widely used method in epidemiological studies. This is based on microscopic examination of a smear of a small but fixed amount of faecal sample (usually 20 to 50 mg). Coarse particles and fibrous material are first removed from the sample by passing it through a sieve. A fixed sample volume is obtained by the use of a template. This is placed on a microscope slide and squashed with either a piece of cellophane soaked in glycerol or a glass coverslip. After leaving the slide for 6 to 24 h to allow the preparation to clear, the eggs are counted and the level of infection expressed as eggs per gram of faeces. Unfortunately, watery or diarrhoeal stools cannot be processed this way, and low-intensity infections may not be detected, since only small faecal samples are examined and eggs may be clumped unevenly in the stool. Increased sensitivity is obtained by increasing the number of samples examined. For diagnosis of light infections in previously unexposed travellers, microscopic examination of a rectal tissue snip crushed between glass slides is often the most sensitive direct diagnostic method. This method can also be used for biopsies. The crushed tissue sample is far better than a sectioned biopsy for the detection and identification of eggs.

Other direct methods

Recently, sensitive enzyme immune assays have been developed to detect circulating schistosome antigens in serum or urine. These antigens, circulating anodic antigen and circulating cathodic antigen, are derived from the gut of the adult schistosomes. The assays have almost 100 per cent specificity and very high sensitivity, and are excellent epidemiological tools as they provide a direct estimate of worm burden and can be used to monitor the efficacy of chemotherapy. They are less well suited for diagnosis of light infections in travellers.

Indirect diagnostic techniques

In *S. haematobium* infections, chemical reagent strips for detection of microhaematuria are widely used in endemic areas as a diagnostic measure. The method can be used in areas of both high and low transmission and there is a consistent significant correlation between microhaematuria and intensity of infection. In intestinal schistosomiasis, blood may be found in the stools, but it is not as useful an indicator of infection. In urinary schistosomiasis, eosinophiluria, with high numbers of eosinophil granulocytes in the urine, is a characteristic finding. Recently, detection of the eosinophil granule protein **ECP** (eosinophil cationic protein) in urine has been used for the qualitative assessment of eosinophil infiltration of the bladder mucosa, and hence local inflammation. Measurement of ECP in urine has proved useful in following post-treatment resolution of urinary tract morbidity in endemic areas. Eosinophilia is often found in acutely infected travellers. In cases where eggs are difficult to find, eosinophilia plus a history of exposure may suggest the need for further examination for schistosomiasis including serodiagnosis.

Immunodiagnosis

In cases of suspected schistosomiasis in which eggs have not been detected, serology can be used to demonstrate specific antibodies. An indirect immunofluorescence test using sections of adult worms for detection of specific immunoglobulins (IgM and IgG) is widely used. For travellers, a positive antibody result combined with a history of exposure should lead to treatment. Serodiagnosis is not useful in endemic areas because of the high levels of specific antibodies found in naturally exposed populations.

Ultrasonography

Ultrasonography is non-invasive, portable, has no biological hazards for the patient, and can be used to either complement or replace many invasive diagnostic techniques. It is the technique of choice for grading schistosomal periportal fibrosis, portal hypertension, hydronephrosis, and urinary bladder lesions. A protocol for standardized investigations and methods of reporting has been produced by the World Health Organization. Ultrasonography is especially useful for monitoring decreases in morbidity after chemotherapy programmes.

Pathophysiology/pathogenesis

Schistosome eggs can be trapped in the tissues, often the walls of the intestines or, depending on species, the urinary bladder or ureters. The eggs of *S. mansoni* and *S. japonicum* are swept into the liver via the portal system, where they embolize into the portal radicles and give rise to vascular and granulomatous changes. Granulomatous pyelophlebitis and peripyelophlebitis is responsible for development of portal hypertension, while granulomata with subsequent fibrosis may be responsible for the periportal fibrosis. The characteristic lesion in the liver is a presinusoidal periportal fibrosis (Symmers' fibrosis). There is typically no bridging between the fibrous tracts, no nodule formation, and no hepatic cell damage. Increased portal pressure can result in the development of portosystemic collaterals and eggs may pass directly from the portal vein to the pulmonary circulation. Here the combination of vascular and granulomatous changes is responsible for pulmonary hypertension.

Treatment

Today the drug of choice is praziquantel, available as 600 mg tablets (e.g. Biltricide®, Distocide®). It is administered orally, normally in a single dose, and is effective against all schistosome species infecting man. It is also effective for most other trematode infections and against adult cestodes. The drug is safe and well tolerated. Drug dosages are shown in Table 1. Complete cure is achieved in up to 85 per cent of those treated, and egg counts are reduced by 95 per cent or more in others. In endemic areas, this level of efficacy is acceptable since very light residual infections do not lead to severe morbidity. In patients who are not cured by the initial treatment, the same dose can be repeated at weekly intervals for 2 weeks or on two successive days.

Although praziquantel has not been shown to be teratogenic, it is not recommended for use during pregnancy. Apart from this there are no contraindications. Any side-effects are generally mild, resolving spontaneously over a few hours and rarely requiring medication. Gastrointestinal side-effects include abdominal pain or discomfort and sometimes vomiting. They occur more frequently in individuals with high infection intensities. Urticarial skin reactions and periorbital oedema may occur in about 2 per cent of treated individuals. General side-effects including headache, dizziness, fever, and fatigue can also occur, but less frequently. As a general principle, all patients with acute schistosomiasis should be treated with praziquantel. It is disputed whether steroids should be added to specific drug treatments. A beneficial effect has been demonstrated in some studies where corticosteroids have been added to praziquantel treatment. Use of praziquantel for cerebral *S. japonicum* infections is safe and effective, resulting in rapid dissipation of cerebral oedema and resolution of cerebral masses. Chemotherapy is only part of the management of schistosomiasis-associated portal hypertension, since the main complications are due to obstructive pathology. Management of portal hypertension and prevention of bleeding from oesophageal varices is beyond the scope of this chapter. Praziquantel has largely replaced other drugs for treatment of schistosomiasis. However, metrifonate (Biarcil®) and oxamniquine (Mansil® (South America), Vansil® (Africa)) are still used sometimes.

Prognosis

Most infected people have few, if any, overt symptoms. Acute schistosomiasis can be fatal or can lead to severe residual damage to the nervous system if not treated, but responds well to antischistosomal therapy if started early. Early infections respond extremely well to treatment and the pathological lesions regress leaving little residual damage. However, in endemic areas individuals, particularly young children, are rapidly re-exposed and reinfected unless control measures are taken at the community level. Chronic infections with fibrosis respond less well to specific antischistosomal treatment, although some regression of hepatosplenic disease has been seen after treatment. The lifetime prognosis is worst in patients with severe hepatosplenic schistosomiasis and oesophageal varices. Previous episodes of haematemesis indicates a 70 per cent risk of rebleeding.

Transmission and epidemiology

Each successful cercarial penetration of human skin has the potential to give rise to a single male or female adult worm, but it is probable that many cercariae die naturally in the epidermis. People tend to accumulate worms with continued exposure to infection. However, human populations in endemic areas do not just continue to accumulate worms with age. Intensities of infection increase in children during their younger years (as estimated by numbers of excreted eggs), peaking around the age of 12 years, before falling to lower levels in adulthood (Fig. 8(a)). This is probably due to the death of older worms, which are not replaced at a similar rate in older people. This age–infection intensity profile is more pronounced if study populations are given chemotherapy to remove existing infections and then monitored for

levels of reinfection over several subsequent years. In these circumstances, it is clear that young children are much more susceptible to reinfection than older children or adults, and that a striking change in susceptibility to reinfection occurs after 12 years of age. The slower acquisition of worms in adulthood could be due to reduced exposure to infection or to age-dependent changes in innate resistance or acquired immunity. In many endemic areas children have more contact with water than adults, but careful observation of water-associated behaviour has shown that age profiles of water contact are variable between communities, whereas profiles of reinfection intensities are remarkably consistent (Fig. 8(b)). This suggests that host-related factors other than exposure influence susceptibility to reinfection. This has been most convincingly shown in fishing communities in areas with high *S. mansoni* transmission on Lake Albert, Uganda. Here occupational water contact results in adults having greater exposure to infection than their children, yet, within 12 months of treatment, it is the children under 12 years of age that suffer much higher reinfection intensities. Current research is focused on assessing the relative roles of innate resistance and acquired immunity in this age-dependent resistance and whether the onset of puberty or the length of time spent living in endemic areas might be important. For example, it is not known if this age-dependent resistance to infection holds true for travellers exposed to infection for the first time. Immune responses to schistosomes also differ between children and adults. Specific IgE and other characteristically T helper 2 type responses against the parasite are associated with resistance to reinfection. Whatever mechanisms underlie the contrasting susceptibilities of children and adults, continued exposure can be expected to result in reinfection, especially amongst younger children.

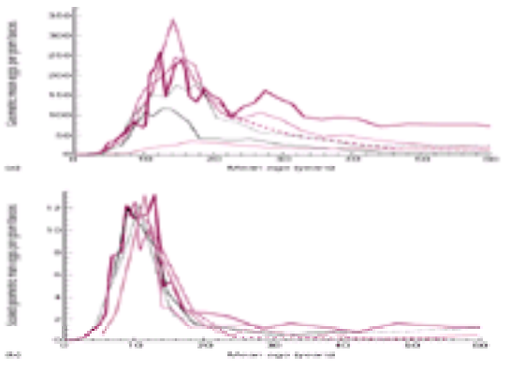

Fig. 8 (a) Age–intensity profiles of *S. mansoni* infection from six communities in Kenya. (Reproduced from Fulford *et al*. (1992) with permission.) (b) Age–reinfection intensity profiles of *S. mansoni* after chemotherapy in the same six communities in Kenya, assessed between 12 and 36 months after treatment. (By courtesy of AJC Fulford.)

Prevention and control

Despite the high risk of reinfection, chemotherapy is usually highly beneficial at both the individual and population levels, as those suffering high intensities of infection are at greatest risk of the more severe forms of schistosomiasis. Various chemotherapy-based control strategies can be employed depending on intensity of transmission and the available resources. In areas of high transmission, population-based chemotherapy can avoid the time and expense required for diagnosis and reduce the prevalence and severity of morbidity. Alternatively, schoolchildren can be targeted for treatment, as they invariably have the heaviest worm burdens and contribute most to on-going transmission. In areas of less intense transmission, treatment can be restricted to diagnosed cases. The provision of safe water supplies and sanitation, where it can be achieved, will make an important additional contribution. Mortality can be prevented and morbidity best controlled by a combination of health education, chemotherapy, provision of safe water supplies and sanitation, and, where appropriate, snail control. Health education should be aimed at improving practices of water use and preventing indiscriminate urination and defaecation. The role of molluscicides in control programmes depends on the local epidemiological and ecological circumstances and the resources available. Within the context of a larger concerted intervention, focal mollusciciding of major transmission sites can be useful. Eradication of host snail species is not usually feasible, although modification of the environment to eliminate snails has been successful in parts of China. In general, it has only been through sustained effort with integrated control strategies that disease control has been achieved. Schistosomiasis control strategies are guided by the Second Report of the WHO Expert Committee on the Control of Schistosomiasis (1993). Recognition that the available control methods, including effective chemotherapy, have failed to reduce the world burden of schistosomiasis has led to renewed efforts to develop an effective vaccine. Recombinant schistosome antigens have been partially successful in protecting experimental animals and several are progressing towards phase I and II human trials.

Further reading

Day JH *et al*. (1996). Schistosomiasis in travellers returning from sub-Saharan Africa. *British Medical Journal* **313**, 268–9. [A review on schistosomiasis in travellers with emphasis on most common symptoms and clinical findings.]

Fairley J (1991). *Bilharzia. A history of imperial tropical medicine*. Cambridge University Press, Cambridge. [An excellent and detailed history of schistosomiasis, including developments in research and control up until the 1970s.]

Feldmeier H, Poggensee G (1993). Diagnostic techniques in schistosomiasis control. A review. *Acta Tropica* **52**, 205–20. [A review of diagnostic techniques, also considering the constraints and drawbacks relating to the various diagnostic methods.]

Ferrari TC (1999). Spinal cord schistosomiasis. A report of 2 cases and review emphasising clinical aspects. *Medicine (Baltimore)* **78**, 176–90. [Review of 231 cases including clinical and treatment aspects.]

Jordan P, Webbe G, Sturrock RF, eds (1993). *Human schistosomiasis*. CAB International, Wallingford. [The definitive text on human schistosomiasis. Including: A comprehensive review of pathology and clinical aspects of *Schistosoma mansoni* infection by Lambertucci; of *S. haematobium* and *S. intercalatum* by Farid; and of *S. japonicum* and *S. japonicum*-like infections by Gang.]

Kabatereine NB *et al*. (1999). Adult resistance to schistosomiasis: age-dependence of reinfection remains constant in communities with diverse exposure patterns. *Parasitology* **118**, 101–6. [The demonstration that children are more susceptible to reinfection than adults.]

Mahmoud A, ed. (2001). *Tropical Medicine: Science and Practice*, Vol. 3 Schistosomiasis, Imperial College Press, London. [A recent book with reviews on various aspects of clinical and experimental schistosomiasis.]

Saconato H, Atallah A (1999). Interventions for treating schistosomiasis mansoni (Cochrane Review). In: *The Cochrane Library*, Issue 3. Update Software, Oxford.

Squires N (1999). Interventions for treating schistosomiasis haematobium (Cochrane Review) In: *The Cochrane Library*, Issue 3. Update Software, Oxford.

7.16.2Liver fluke infections

David I. Grove

Liver flukes, otherwise known as trematodes, are leaf-like hermaphroditic flatworms. The hepatobiliary system of humans is commonly infected by flukes of the genera *Clonorchis* and *Opisthorchis* and occasionally by other species (Table 1). In addition, *Eurytrema pancreaticum* has been found rarely in the pancreatic duct. These infections are usually diagnosed by finding eggs in the faeces. Unfortunately, eggs of many of these species cannot be differentiated from each other nor can they be distinguished reliably from the eggs of certain intestinal trematodes. In such cases, definitive diagnosis can only be made if adult worms are recovered from the stools after anthelmintic treatment, at surgery, or at autopsy; parasitological textbooks should be consulted for diagnostic details.

Clonorchiasis

Lifecycle

Clonorchis sinensis adult worms, 10 to 25 mm long by 3 to 5 mm wide, live in the bile ducts or occasionally the gallbladder attached to the mucosa. They produce eggs which are passed in the faeces (Fig. 1). The miracidium within the egg hatches after ingestion by a suitable species of aquatic snail; nine species belonging to the families Hydrobidae, Melanidae, Assimineidae, and Thiaridae are known to be susceptible but *Parafossarulus manchouricus* is perhaps the most common. The miracidia develop into sporocysts then in turn become rediae which produce larvae known as cercariae. After 6 to 8 weeks, the cercariae emerge from the snail and swim about in the water until they encounter certain freshwater fishes (over 100 species, mostly of the family Cyprinidae, i.e. carp, are susceptible). They attach to the surface of the fish, lose their tails, penetrate under the scales, encyst in the skin or flesh, and develop into infective metacercariae over several weeks. When raw or undercooked infected fish is eaten by humans, the metacercariae excyst in the stomach, enter the common bile duct through the ampulla of Vater, and ascend into the biliary passages where they mature in 1 month. Adult worms may live for up to 40 years.

Fig. 1 Egg of *Clonorchis sinensis*: this is identical with that of *Opisthorchis viverrini*. (By courtesy of Prayong Radomyos, Faculty of Tropical Medicine, Mahidol University, Bangkok.)

Epidemiology and control

Fish-eating mammals including humans, dogs, cats, and rats may be infected with *C. sinensis*. Human clonorchiasis is endemic in Japan, Korea, China, and Vietnam where the first and second intermediate hosts are found and where the population habitually consumes raw fish. In endemic areas, fish are kept in ponds and fertilized with human and animal faeces. Over 20 million people are thought to be infected in China. Control programmes include proper waste disposal, measures to control snail numbers, and mass treatment with praziquantel, but the most important is health education to discourage the habit of eating raw or undercooked fish.

Pathology

Pathological changes are related to the intensity and duration of infection. They are produced by mechanical irritation, toxin production, immunological responses, and secondary bacterial infection. Inspection of the cut surface of the liver often reveals dilated, thick-walled bile ducts with adult worms visible within their lumens. Adult flukes may be found in the gallbladder but they are usually killed by bile. Histologically, there is desquamation and hyperplasia of epithelial cells, formation of adenomatous tissue and proliferation of periductal connective tissue, and infiltration with eosinophils and mononuclear cells. This may be complicated by epithelial metaplasia then mucinous cholangiocarcinoma. Recurrent pyogenic cholangitis is a common complication and the worms and eggs act as a nidus for gallstone formation. Some patients have flukes in the pancreatic duct which may cause pancreatitis.

Clinical features

Most patients are asymptomatic and are diagnosed incidentally on stool examination. Symptoms are more common in older patients with heavy worm burdens. It is difficult to differentiate these symptoms from other conditions but they include right hypochondrial or epigastric pain or discomfort, lassitude, anorexia, and flatulence. Some patients complain of a peculiar, hot sensation on the skin of the abdomen or back. Cholangitis causes fever, right upper quadrant pain, and jaundice. Cholangiocarcinoma is associated with pain, jaundice, and weight loss.

Diagnosis

The diagnosis is suggested by finding eggs in faeces or in duodenal aspirates. They are yellow-brown, 25 to 35 µm long by 12 to 19 µm wide, and have a seated operculum with a small knob at the other end (Fig. 2). They cannot be differentiated from ova of *Opisthorchis* species. Furthermore, they are extremely difficult to differentiate from eggs of flukes in the family Heterophyidae (see Chapter 7.16.4), although the latter tend to have a smoother egg shell, a less prominent shoulder at

the operculum, and the knob may be absent. The diagnosis can only be confirmed by examination of adult flukes. Serological tests have been described but are not routinely used for individual patient diagnosis. Imaging techniques such as ultrasound or computed tomography may disclose adult worms in the gallbladder or bile ducts, which are often dilated and may contain sludge. Liver function tests may be abnormal, often with an obstructive picture.

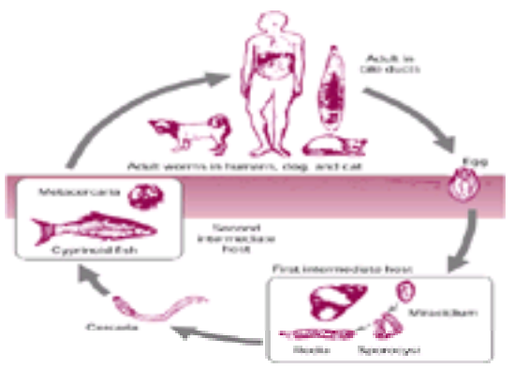

Fig. 2 Lifecycle of *Clonorchis sinensis* and *Opisthorchis* species.

Treatment

Praziquantel is the treatment of choice and in a dose of 25 mg/kg three times daily after meals for 2 days has a cure rate of close to 100 per cent; eggs should disappear from the stool within 1 week. Biliary tract abnormalities may reverse after treatment as this has been shown in opisthorchiasis. Triclabendazole may prove to be useful but there is insufficient documentation at present. Bacterial cholangitis is treated with antibiotic therapy such as a combination of amoxicillin, gentamicin, and metronidazole. Surgery may be required in some patients with obstructive jaundice.

Opisthorchiasis viverrini

This infection is very similar to clonorchiasis. The adult *Opisthorchis viverrini* is smaller than *C. sinensis*, measuring 7 to 12 mm by 2 to 3 mm, although there is some discrepancy in the literature over its size, perhaps reflecting different methods of preparation and fixation. It may live for over 10 years. The lifecycle is similar to that of *Clonorchis* with various species of the genus *Bithynia*, particularly *B. goniomphalus*, *B. funiculata*, and *B. siamensis* (= *laevis*), being the snail first intermediate host. Many species of carp serve as the second intermediate host. Humans, dogs, cats, and other fish-eating mammals are definitive hosts. This parasite is endemic in northern Thailand and adjacent Laos and Cambodia where 10 million people are estimated to be infected because of the popularity of chopped raw cyprinoid fish as a foodstuff.

The pathology and clinical features are similar to those induced by *C. sinensis.* The association with cholangiocarcinoma may be even more striking with this infection. The diagnosis is made as discussed under clonorchiasis. Praziquantel is the drug of choice; 25 mg/kg three times after meals for 1 day gives close to 100 per cent cure rate. Mebendazole (30 mg/kg daily) or albendazole (400 mg twice daily) may be effective if given for several weeks. Triclabendazole may prove to be useful but there is insufficient documentation at present. A control programme is underway in Thailand which includes detection and treatment of infected people together with intensive health education.

Opisthorchiasis felineus

This infection is very similar to clonorchiasis. The adult *Opisthorchis felineus* is morphologically very similar if not identical to *O. viverrini* (the two species have been distinguished by the pattern of flame cells in the cercariae). The lifecycle is similar with *Bithynia leachi* being the only known molluscan intermediate host. Many species of carp serve as the second intermediate host. Humans, dogs, cats, rats, foxes, seals, and other fish-eating mammals are definitive hosts. Infection is acquired by eating raw or undercooked fish; in Siberia, raw, slightly salted, and frozen fish is often consumed. This parasite is endemic particularly in Russia and adjacent countries but also in parts of southern Europe and eastern Asia with several million people probably being infected overall. Eggs are indistinguishable from those of *O. viverrini* and *C. sinensis*. The pathology, clinical features, diagnosis, and treatment are similar to *O. viverrini* and *C. sinensis* infections.

Fascioliasis

Lifecycle

Fascioliasis is due to infection with the sheep liver fluke, *Fasciola hepatica* or with *F. gigantica*. Adult *F. hepatica* flukes, 20 to 30 mm by 8 to 13 mm in size, live in the large bile ducts and produce eggs which are passed in the stools. The eggs require a period of 9 to 15 days for the miracidia to develop and hatch in water at 22 to 25°C, but remain viable for up to 9 months if kept moist and cool. The miracidia penetrate the tissues of various species of amphibious snails of the family Lymnaeidae and develop over the next 4 to 5 weeks through the stages of sporocyst, rediae, daughter rediae, and cercariae. The cercariae emerge from the snails and encyst on various kinds of aquatic vegetation to become metacercariae. A wide range of mammals is susceptible to infection, but sheep and cattle are the most important. Human infections are usually acquired by eating watercress or by drinking water contaminated with metacercariae. Metacercariae excyst in the duodenum, penetrate the intestinal wall, and pass into the peritoneal cavity. They then invade the liver capsule and migrate through the hepatic parenchyma to the bile ducts where they mature in about 3 to 4 months. The lifespan of these flukes is several years.

F. gigantica is large attaining a size of up to 7.5 cm. The eggs are difficult to distinguish from those of *F. hepatica* and the lifecycle of the two parasites is similar.

Epidemiology and control

Because of the wide range of susceptible definitive and intermediate hosts, the infection is geographically widespread. Human infections with *F. hepatica* have been reported from all continents. Fascioliasis gigantica is less frequent and has been seen in Africa and Asia. Infection is prevented by not eating fresh aquatic plants, particularly watercress (*Nasturtium officinale*) and by boiling drinking water. Veterinary control measures include elimination of the snail intermediate hosts by drainage of pastures and treatment with molluscicides and by eradication of infection from infected herds.

Pathology

In the early stages of infection, larvae migrating through the liver parenchyma may cause considerable destruction with necrosis, abscess formation, and haemorrhage. The number of tunnels lined by ragged walls of necrotic, bleeding, and inflamed liver tissue is proportional to the number of worms. In the chronic stages, the walls of the bile ducts become thickened by fibrous tissue and inflammatory infiltration, the epithelium becomes hyperplastic, and the bile ducts dilate. Occasionally the lumina of the bile ducts may become obliterated causing obstructive jaundice. These structural changes predispose to secondary bacterial infection which exacerbates the problem. Sclerosing cholangitis and biliary cirrhosis may follow prolonged heavy infection. There is no apparent association with cholangiocarcinoma.

Clinical features

Human fascioliasis is usually mild and related to the phase of infection. There are three phases.

1. In the migratory phase, symptoms usually begin about 1 month after infection. Patients may develop abdominal discomfort or pain (especially in the epigastrium and right upper quadrant), anorexia, nausea, vomiting, fever, headache, tender hepatomegaly, and urticaria. These initial symptoms may persist for several months.

2. The latent phase is asymptomatic and may last for months to years.
3. The obstructive phase is characterized by the recurrence or appearance for the first time of epigastric and right upper quadrant abdominal pain, biliary colic, anorexia, nausea, vomiting, tender hepatomegaly, fever, and jaundice. These features are frequently due to complicating bacterial cholangitis or cholecystitis and may be associated with bacteraemia.

Flukes occasionally migrate to other sites, especially the anterior abdominal wall. Acute oedematous nasopharyngitis ('halziun') may be an allergic response to larval flukes which attach to the pharyngeal wall after ingestion of infected, raw sheep or goat liver.

Diagnosis

In enzootic areas, early fascioliasis is suspected in patients with fever, tender hepatomegaly, and eosinophilia who give a history of consuming freshwater plants. If available, serological tests may be useful early in the illness before egg production begins. Liver biopsy may be helpful in some cases.

Chronic fascioliasis is diagnosed by finding the characteristic eggs in stools or fluid obtained by duodenal or biliary drainage. The eggs of *F. hepatica* and *F. gigantica* cannot be distinguished reliably from each other or from those of the intestinal fluke, *Fasciolopsis buski*; differentiation of these two infections requires identification of adult flukes. Liver function tests are often abnormal and may show an obstructive picture. Radiolucent shadows of flukes may be seen by cholangiography. Ultrasonography and computed tomography are useful in the demonstration of lesions in the liver and biliary tracts. If the patient has recently consumed liver, spurious infection (ingestion of eggs) should be ruled out by placing the patient on a liver-free diet for a few days and repeating the stool examination.

Treatment

The treatment of fascioliasis has been problematic. Success has been claimed for bithiniol and emetine but these drugs are not generally available. Chloroquine at 5 mg/kg per day orally for 3 weeks has limited effectiveness. Praziquantel, which is active against many trematodes, is often ineffective in fascioliasis but may be tried if other agents are not available. Recent studies have shown that triclabendazole in a single oral dose of 10 mg/kg is very effective although some patients require a second dose after a few weeks. This drug appears to have few side-effects. It is available in some countries but not others; further information can be sought from the manufacturer (Novartis, Basle, Switzerland). Flukes are evacuated through the intestinal tract. Another drug under investigation which shows promise is nitazoxanide administered in a dose of 500 mg orally twice daily for 6 days.

Dicrocoeliasis

Dicrocoelium dendriticum adult worms measuring 5 to 15 mm by 1.5 to 2.5 mm live in the biliary passages. Eggs passed in the stools are ingested by certain land snails (e.g. species of *Zebrina* and *Helicella*) in which they develop through two stages of sporocysts with the eventual production of cercariae. The snail leaves slime balls of cercariae on the ground and these are ingested by ants (*Formica* species) in which they develop into metacercariae.

This organism is primarily an infection of sheep, goats, deer, and other herbivores which ingest ants. Humans are rarely infected, usually by accident. Cases have been reported from Europe, Asia, and Africa. Spurious infections result from the consumption of raw, infected liver. Patients may be asymptomatic but may complain of dyspepsia, flatulence, and abdominal colic. The diagnosis is made by finding the eggs in faeces, bile, or duodenal fluid; they cannot be differentiated from those of *Eurytrema pancreaticum*. Definitive diagnosis is made by identification of adult worms. Treatment is with praziquantel at 25 mg/kg three times after meals for 1 day.

Metorchiasis

Many fish-eating mammals of North America serve as definitive hosts for *Metorchis conjunctus*. The aquatic snail *Amnicola limosa* is the first intermediate host; eggs are ingested, hatch into miracidia, and ultimately release cercariae. Metacercariae develop in the flesh of several species of freshwater fish. Ingested metacercariae hatch in the duodenum and migrate up the biliary tree.

A point source outbreak of this disease has been reported in 19 people who ate raw fish prepared from the white sucker (*Catostomus commersoni*) caught in a river north of Montreal. The illness was characterized by upper abdominal pain, low-grade fever, eosinophilia, and abnormal liver function tests. Ten days after ingestion of infected fish, eggs indistinguishable from those of *O. viverrini* were seen in the stools. The patients responded to treatment with praziquantel.

Further reading

Arjona R *et al*. (1995). Fascioliasis in developed countries: a review of classic and aberrant forms of the disease. *Medicine (Baltimore)* **74**, 13–23.

Bronstein AM, Zavoikin VD. Brief update on *Opisthorchis felineus* in Russia. http://www.cfound.to.it/html/bronste.htm .

Connor DH *et al*., eds (1997). *Pathology of infectious diseases*, Vol 2, pp 1305–588. Appleton & Lange, Stamford.

el-Karaksy H *et al*. (1999). Human fascioliasis in Egyptian children: successful treatment with triclabendazole. *Journal of Tropical Paediatrics* **45**, 135–8.

Jongsuksuntigul P, Imsomboon T (1998). Epidemiology of opisthorchiasis and national control program inThailand. *Southeast Asian Journal of Tropical Medicine and Public Health* **29**, 327–32.

Kino H *et al*. (1998). Epidemiology of clonorchiasis in Ninh Binh Province, Vietnam. *Southeast Asian Journal of Tropical Medicine and Public Health* **29**, 250–4.

MacLean JD *et al*. (1996) Common-source outbreak of acute infection due to the North American liver fluke *Metorchis conjunctus*. *Lancet* **347**, 154–8.

Pungpak S *et al*. (1997). *Opisthorchis viverrini* infection in Thailand: studies on the morbidity of the infection and resolution following praziquantel treatment. *American Journal of Tropical Medicine and Hygiene* **56**, 311–4

Rossignol JF, Abaza H, Friedman H (1998). Successful treatment of human fascioliasis with nitazoxanide. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92**, 103–4.

Watanapa P (1996). Cholangiocarcinoma in patients with opisthorchiasis. *British Journal of Surgery* **83**, 1062–4.

7.16.3Lung flukes (paragonimiasis)

Sirivan Vanijanonta

Lung fluke infection is caused by *Paragonimus* spp. At least 15 species cause disease in humans (Table 1). *Paragonimus westermani* is the most common and widespread, but *P. africanus*, *P. uterobilateralis* (West Africa), *P. ilokstuenensis* (China), and *P. peruvianus* (South America) are also causative. *P. heterotremus* (Thailand, Laos, Vietnam), *P. szechuanensis*, and *P. hueitungensis* also cause cutaneous paragonimiasis.

The adult flukes are reddish-brown and pea-shaped (Plate 1). They are 0. 8 to 1. 6 cm in length, 0. 4 to 0. 8 cm in width, and 0. 3 to 0. 5 cm thick with cuticular spines on the integument. Typically they are encapsulated in cysts adjacent to the bronchi. The eggs are golden brown and ovoid in shape (80–120 × 50–60 µm) (Plate 2).

Lifecycle

Adult flukes encyst in the lung. Ova are expelled through the bronchi and expectorated with sputum or swallowed and passed with faeces. They hatch in fresh water after a few weeks. The resulting miracidia then infect various species of freshwater snail in which they form sporocysts, rediae, and daughter rediae. Metacercariae develop in susceptible freshwater crabs and crayfish (Plate 3, Plate 4). Infection results from ingestion of viable metacercariae in raw or insufficiently cooked crabs and crayfish. Metacercariae excyst in the peritoneal cavity, where they grow and become young flukes. Most of these will then reach the lung by passing through the peritoneal cavity, diaphragm, and pleural cavity, before finally encysting in the lung parenchyma. Tunnels may be formed during their migration. Encysted flukes mature over a period of 6 to 8 weeks and eggs are produced in 10 to 12 weeks. The circuitous routes of migration allow young flukes to lodge and mature in ectopic locations. The reservoir hosts are wild and domestic felines that feed on crabs and crayfish. Freshwater snails that serve as the first intermediate hosts belong to the Thiaridae, Hydrobilidae, and Pleuroceridae families. The second intermediate hosts are the freshwater and brackish-water crabs *Eriocheir japonicus*, *Larnaudia beusekomae* (*Tiwaripotamon beusekomae*), and *Potamon smithiasis*, or crayfish of the genus *Cambaroides*, such as *C. japonicus* in Japan, and *C. similis*, *C. dauricus*, and *C. sckrenki* in China and Korea.

Epidemiology

Paragonimiasis is an important zoonosis. Human beings enter the lifecycle accidentally. However, in some areas human paragonimiasis may be common enough for person-to-person transmission to occur. Human infection is limited in its distribution to places where there are contributory factors that facilitate the lifecycle: reservoir hosts, suitable environment, first and second intermediate hosts, and permissive dietary habits. The three major foci of this disease are in Asia, Africa, and Central and South America. In Asia, endemic areas are to be found in China, Japan, Taiwan, Korea, The Philippines, Thailand, Laos, Vietnam, and Burma, in which the principal parasites are *P. westermani*, *P. skjabini*, and *P. heterotremus*. In Africa, the disease is endemic in eastern Nigeria, the Cameroons, the Congo valley, and the Republic of Congo. In Nigeria the dominant parasite is *P. uterobilateralis*, while in the Cameroons and the Republic of Congo, *P. africanus* predominates. *P. mexicanus*, *P. peruvianus*, and *P. caliensis* are causative agents in Mexico, Guatemala, Honduras, Costa Rica, Ecuador, Colombia, Peru, and Paraguay.

Transmission of *Paragonimus* spp. to man occurs mostly through ingestion of metacercariae in the second intermediate host. Paratenic hosts infected with immature worms also contribute to animal and human disease.

Pathology and pathogenesis

The pathogenesis of human paragonimiasis is unknown. In experimental animals the larval flukes penetrate the intestinal wall and reach the peritoneal cavity, then pass through the diaphragm and pleura to the lung. They cause irritation, acute inflammatory reactions, traumatic tracts, pressure effects, haemorrhage, and necrosis in affected tissues. Pathological findings in the pleural cavity include turbid and haemorrhagic fluid containing numerous pus cells and eosinophils. Acute, diffuse, fibrinoexudative peritonitis may also occur. Abscess cavities containing young flukes are then formed and become enclosed in a fibrous capsule. Mature cysts adjacent to the bronchial system may rupture into it and the cystic contents are then expectorated with sputum or swallowed and passed with faeces. Single or multiple cysts may occur, usually in the lower lobes of the lungs.

Extrapulmonary pathological changes may be caused by aberrant migratory flukes. Cysts, abscesses, and granulomas may be found in the abdominal viscera, subcutaneous tissue, muscles, genital organs, and the brain. *P. heterotremus* and *P. skjabini* also create migratory subcutaneous swellings.

Clinical manifestation

The clinical manifestations are divided into acute and chronic phases. The acute phase occurs after the consumption of an improperly cooked, infected crab or crayfish. The incubation period varies from a few days to weeks. The severity of symptoms usually correlates with the worm load. Invasion and migration by young flukes cause inflammatory and allergic responses such as fever, rashes, urticaria, abdominal pain and discomfort, and a feeling of tightness in the chest. Acute symptoms are rarely serious and patients progress to the chronic stage.

Chronic manifestations are classified as pulmonary and extrapulmonary.

Pulmonary paragonimiasis

The most remarkable clinical feature is a chronic, productive cough with jam-like, brownish-red sputum. Other symptoms include breathlessness, chest pain, unilateral or bilateral pleural effusions, and empyema. Occasionally patients may experience haemoptysis following heavy work or exertion, while pneumothorax occurs rarely.

Pulmonary paragonimiasis is an insidious and persistent lung disease. Patients have surprisingly good general health and usually show few abnormalities on physical examination. A minority of symptomatic patients have normal chest radiographs. Abnormal findings include linear infiltrations, exudative pneumonia, localized pleural effusion, and nodular or cystic lesions. These lesions are predominantly found in the basilar and peripheral regions of both lower lung fields. Cysts may be single or multiple; the most characteristic radiographic feature is a ring shadow with a crescent-shaped opacity along one side of the border resembling the corona phase of a solar eclipse. Other findings are pleural effusion, pleural thickening, and calcification. Long-standing, extensive lesions with fibroatelectasis resemble the lesions of chronic pulmonary tuberculosis.

Extrapulmonary paragonimiasis

Extrapulmonary paragonimiasis is caused by the aberrant migration of larval and young adult flukes to any organ. Migratory swelling of cutaneous or subcutaneous tissues may also occur.

Cerebral paragonimiasis

The clinical symptoms are similar to those of a cerebral space-occupying lesion and are related to the site of the lesion. However, one or more syndromes may be present. Epileptic seizures are common, and patients may develop mental disturbances of the schizoid and paranoid type. Increased intracranial pressure induces persistent intense headache, nausea, vomiting, papilloedema, diplopia, and loss of visual acuity. Patients with paragonimus cysts in the basal meninges will present with meningeal symptoms that include increased intracranial pressure, obstructive hydrocephalus, arterial thrombosis, and stroke. On rare occasions, patients may suffer from cerebellopontine-angle syndrome with tinnitus, progressive deafness, nystagmus, dysphagia, and hiccups.

Spinal-cord paragonimiasis

Spinal involvement produces progressive weakness, sensory impairment of the lower extremities, paralysis, and back pain.

Intra-abdominal paragonimiasis

Paragonimus spp. may create migratory tracts or pressure effects leading to necrosis of the spleen, liver, small and large intestinal wall, and cause non-specific abdominal signs and symptoms.

Subcutaneous paragonimiasis

P. skjabini, *P. westermani*, and *P. heterotremus* cause migratory subcutaneous nodules or asymptomatic subcutaneous nodule(s) at any part of the body.

Diagnosis

Pulmonary paragonimiasis should be excluded in any patient from an endemic area who presents with a chronic productive cough and jam-like, brownish-red or 'rusty' sputum. The definitive diagnosis is made by observing the characteristic ova in sputum, pleural effusion, or stool, or flukes in biopsy specimens. Expectoration of intact flukes has been reported. Other supportive evidence is obtained by chest radiographs, which show the characteristic shadows of single or multiple cysts in the lungs (Fig. 1) Computed tomography of the chest is also helpful (Fig. 2).

Fig. 1 Pulmonary paragonimiasis posteroanterior radiograph showing thick-walled cystic lesion in the right lower lobe and left lower lobe with pericystic fibrosis. (Copyright Professor Sirivan Vanijanonta.)

Fig. 2 Pulmonary paragonimiasis' CT scan, howing thick-walled lesion with pericystic fibrosis in the left upper lobe and a fibrocalcific lesion in the right upper lobe. (Copyright Professor Sirivan Vanijanonta.)

Serology is essential for the diagnosis of extrapulmonary paragonimiasis. Enzyme immunoassay, and dot enzyme immunoassay, and monoclonal antibody tests are highly sensitive and specific, as is counterimmunoelectrophoresis using adult or free metacercariae as a source of antigen. Other less sensitive but more specific tests include complement fixation and indirect haemagglutination. Intradermal skin tests have been used for epidemiological surveys.

Differential diagnosis

Pulmonary paragonimiasis should be differentiated from pulmonary tuberculosis, melioidosis, lung abscesses, and lung tumours.

Extrapulmonary paragonimiasis should be differentiated from other diseases that produce similar clinical manifestations in affected organs. For example, cerebral paragonimiasis should be differentiated from cerebral cysticercosis, hydatidosis, meningoencephalitis, brain abscesses, and tumours. Subcutaneous paragonimiasis may resemble gnathostomiasis, sparganosis, loiasis, or onchocerciasis.

Treatment

Specific

The drug of choice is praziquantel at a dosage of 75 mg/kg per day in three divided doses for 2 to 3 days. A cure rate of nearly 100 per cent has been reported in multicentre studies. Albendazole and tricarbendazole are also effective. The symptoms rapidly improve in a few days. Eggs disappear from the sputum in a few weeks. Radiological improvement takes months, depending on the extent and chronicity of the disease. Convulsions, seizures, coma, and behavioural changes may develop during treatment of cerebral paragonimiasis. As a result of parasite death, brain oedema and host–parasite interaction may cause increased intracranial

pressure. Therefore, treatment should proceed with caution and the dose adjusted if necessary. Dexamethasone cover has been suggested in some cases.

Symptomatic and supportive treatment

These treatments, including blood transfusion, bronchodilators, anticonvulsants, and analgesics, are also important.

Prognosis

Pulmonary paragonimiasis is rarely fatal and the lesions may calcify or completely resolve in a few years. Cerebral paragonimiasis may cause chronic morbidity such as epilepsy, mental changes, and neurological sequelae.

Prevention and control

Effective control measures are directed towards interruption of the lifecycle. However, control and eradication of intermediate hosts is impracticable; therefore, health education, changes in social and dietary customs, and the mass treatment of infected people in an endemic area are more effective for prevention and control.

Further reading

Calvopina M, *et al*. (1998). Treatment of human paragonimisis with tricarbendazole: clinical tolerance and drug efficacy. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92**, 566–9.

Chen GU, *et al*. (1986). Counterimmunoelectrophoresis in detecting antibodies in experimental paragonimiasis. *Chinese Journal of Zoonoses* **2**, 58.

Chung HL, *et al*. (1981). Recent progresses in studies of paragonimus and paragonimiasis control in China. *Chinese Medical Journal* **94**, 483–94.

Jun-ichi I (1987). Evaluation of ELISA for the diagnosis of paragonimiasis westermani. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **81**, 3–6.

Maleewong W (1997). Recent advance in the diagnosis of paragonimiasis. *Southeast Asian Journal of Tropical Medicine and Hygiene* **28**, 134–8.

Miyazaki I (1982). Paragonimiasis. In *CRC handbook series in zoonoses, Section C: Parasitic zoonoses*, Vol. III, pp. 143–64. Lea and Febiger, Philadelphia.

Miyazaki I, Harinasuta T (1966). The first case of human paragonimiasis caused by *Paragonimus heterotremus* (Chen et Hsia 1964). *Annals of Tropical Medicine and Parasitology* **60**, 509.

Pariyanonda S, *et al*. (1990). Serodiagnosis of human paragonimiasis caused by *Paragonimus heterotremus*. *Southeast Asia Journal of Tropical Medicine and Public Health* **21**, 103–7.

Queuche F, *et al*. (1997). Endemic area of paragonimiasis in Vietnam. *Sante* **7**, 155–9.

Vanijanonta S, Bunnag D, Harinasuta T (1984). *Paragonimus heterotremus* and other paragonimus spp. in Thailand: pathogenesis, clinical and treatment. *Drug Research* **34**, 1186–8.

Vanijanonta S, Bunnag D, Harinasuta T (1984). Radiological findings in pulmonary paragonimiasis heterotremus. *Southeast Asia Journal of Tropical Medicine and Public Health* **15**, 122–8.

Zhang YQ, *et al*. (1986). The significance of dot-ELISA in diagnosis of paragonimiasis. *Chinese Journal of Internal Medicine*, **25**, 679–81.

7.16.4 Intestinal trematode infections

David I. Grove

Intestinal trematode infections of humans other than intestinal schistosomiasis are widespread but are most common in Asia. This is a reflection of cultural factors, particularly the consumption of raw or undercooked vectors, most frequently freshwater fish and molluscs, but also water plants. More than 50 million people are estimated to harbour one or more species of these hermaphroditic flukes. In many instances, the extent of morbidity due to these infections is uncertain.

Diagnosis

The diagnosis of intestinal fluke infections is usually based upon recovery of eggs from stools. Unfortunately, ova from species within a given family often look very similar and it may be possible when using routine laboratory methods to identify an infection only to family level, such as a heterophyid or echinostomatid egg. Definitive identification relies upon recovery of adult worms after anthelmintic treatment. Identifying characteristics are provided in parasitology texts.

Treatment

Praziquantel has been shown to be effective with a number of these infections and is the drug of first choice. It is given in a dose of 20 mg/kg orally after a meal, perhaps repeated once or twice. Flukes are usually expelled the following day. The role of triclabendazole, for instance in a dose of 10 mg/kg orally, in the treatment of intestinal trematodiases is not yet clear. Other possibilities which are less likely to be effective include niclosamide at 150 mg/kg orally for 1 or 2 days and albendazole at a dose of 200 mg orally for 2 days.

Prevention

These fluke infections can be prevented by thoroughly cooking potentially infected foodstuffs.

Echinostomiasis

This term may be conveniently used to include all infections with flukes of the family Echinostomatidae. There are more than 30 genera in this family and so far 18 species have been reported to infect humans (Table 1). These species vary in size from 1 to 20 mm in length. Echinostomes live in the intestines of various birds and mammals. When eggs are passed in the stools and reach water, the miracidium develops, hatches, and enters a snail, the first intermediate host. It then develops through the stages of sporocyst, mother redia, and daughter redia to release cercariae. The cercariae in turn infect second intermediate hosts which include various species of snails, tadpoles, and fish or they encyst on vegetation. Humans are infected after ingestion of inadequately cooked food containing these metacercariae.

In humans, they live in the small bowel, particularly the jejunum, and attach to the mucosa where they may cause a variable amount of damage. Heavy worm loads may cause abdominal discomfort, flatulence, and diarrhoea. Eggs 80 to 150 by 50 to 75 µm in size are passed in the stools (Fig. 1). They are yellow-brown, ellipsoidal, thin-shelled, and operculate and contain an immature embryo; eggs of the various species cannot be reliably differentiated from each other or from those of the intestinal fluke *Fasciolopsis buski* or the liver flukes *Fasciola hepatica* and *F. gigantica*.

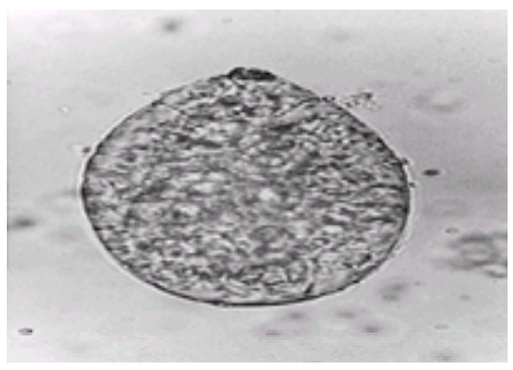

Fig. 1 Egg of *Echinostoma ilocanum* (by courtesy of P. Radomyos). All echinostome eggs look similar, as do those of *Fasciolopsis* and *Fasciola* species.

Fasciolopsiasis

This infection is caused by *Fasciolopsis buski* (see Table 3). The adult fluke (Fig. 2) is found in the small intestine of humans and pigs. When eggs are passed in the stools and reach water, the miracidium develops, hatches, and enters a snail, the first intermediate host; snail hosts include species of *Segmentina*, *Hippeutis*, and *Gyraulus*. The miracidium then develops through the stages of sporocyst and redia to release cercariae after 8 weeks or so. The cercariae swim out and encyst on water plants and develop into metacercariae over 4 weeks. Infection is acquired by ingestion of infected uncooked edible plants such as water caltrop (*Trapa bicornis*), water chestnut (*Eliocharis tuberosa*), water bamboo (*Zizania aquatica*), and watercress (*Neptunia oleracea*).

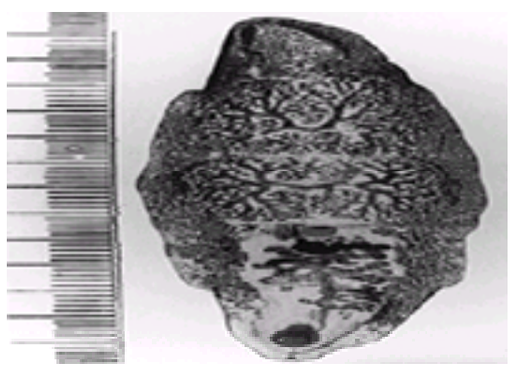

Fig. 2 Adult *Fasciolopsis buski*, 6.5 cm in length (by courtesy of P. Radomyos).

Fifty years ago it was estimated that 10 million people were infected with this parasite. The current prevalence is unknown. Fasciolopsiasis occurs most commonly in areas where people keep pigs and raise and eat freshwater lants.

The adult worms attach themselves to the mucosa of the upper small bowel where they may cause inflammation and erosion and provoke a mucous intestinal discharge. Light infections are generally asymptomatic but heavy worm burdens may be associated with anorexia, nausea, abdominal discomfort, and diarrhoea or even intestinal obstruction. Stools may be foul-smelling and contain undigested food. In severe cases, a protein-losing enteropathy is associated with ascites, generalized oedema, and prostration.

Eggs 130 to 140 by 80 to 85 µm in size are passed in the stools (Fig. 3). They are yellow-brown, ellipsoid, thin-shelled, and operculate and contain an immature embryo; they cannot be reliably differentiated from those of the intestinal echinostomes or of the liver flukes *Fasciola hepatica* and *F. gigantica*.

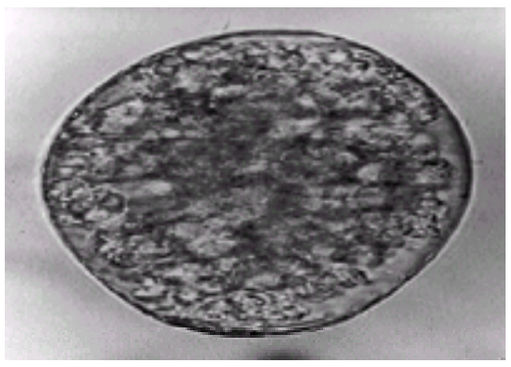

Fig. 3 Egg of *Fasciolopsis buski* (by courtesy of P. Radomyos). Note its similarity to ova of *Fasciola* species and echinostomes.

Heterophyiasis

This term may be conveniently used to include all infections with flukes of the family Heterophyidae although some infections are more precisely known by the generic name of the infecting organism, for instance metagonimiasis. These are small flukes, generally less than 1 to 2 mm in length. So far 37 species in this family have been reported to infect humans (Table 2). These infections are found in many places but are most common in Asia. *Metagonimus yokogawai* is believed to be the most common heterophyid infection.

Heterophyids live in the intestines of various mammals and birds. When eggs are passed in the stools, they contain a ciliated miracidium which hatches when ingested by a freshwater or brackish-water snail, the first intermediate host. Snails susceptible to *Heterophyes* include *Pirenella conica*, *Cerithidea cingulata*, and *Tympanotonus micropterus* while *Semisulcospira libertina* and *Thiara granifera* are host to *Metagonimus*. The miracidium then develops through the stages of sporocyst and one or two generations of rediae to release cercariae. The cercariae in turn infect various species of salmonoid and cyprinoid fish as the second intermediate hosts. These include mullet (e.g. *Mugil cephalus*) and minnow (*Gambusia* spp.) for *Heterophyes* species and carp (e.g. *Carassius carrasius*) and sweet fish (*Plecoglossus altivelis*) in the case of *Metagonimus* species. Humans are infected after ingestion of inadequately cooked fish containing metacercariae which mature in the flesh or scales of the fish.

The adult worms attach to or invade the mucosa of the upper small bowel where they may cause granulomatous inflammation and erosion. Light infections are generally asymptomatic but heavy worm burdens may be associated with anorexia, nausea, abdominal discomfort, and mucous diarrhoea. Occasionally ova deposited in the bowel wall enter blood vessels and embolize to other tissues. Eggs have been found in the heart and central nervous system. In cases of heterophyiasis described in the Philippines, cardiac failure was associated with subepicardial haemorrhages, myocardial damage caused by occlusion of vessels by ova, and eggs were stuck to a thickened, calcified mitral valve. Neurological features include focal cerebral disturbances and transverse myelitis.

Eggs 20 to 40 by 10 to 20 µm in size are passed in the stools (Fig. 4). They are yellow-brown, elongated, operculate, and contain a miracidium. Eggs of members of the family Heterophyidae cannot be reliably differentiated from each other. Furthermore, they are extremely difficult to differentiate from eggs of *Clonorchis sinensis* and *Opisthorchis* species although heterophyids tend to have a smoother egg shell, a less prominent shoulder at the operculum, and the abopercular knob may be absent.

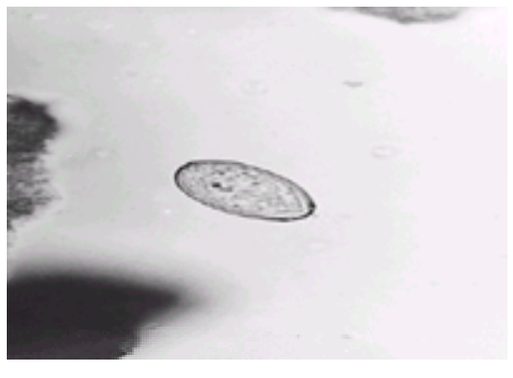

Fig. 4 Egg of *Metagonimus yokogawai* (by courtesy of P. Radomyos). All heterophyid eggs look similar, as do those of *Clonorchis sinensis* and *Opisthorchis viverrini.*

Other intestinal fluke infections

There are another dozen or so species of intestinal fluke belonging to various families that have been reported to infect humans (Table 3). As with other fluke infections, definitive diagnosis depends upon recovery of the adult worms; this is most commonly achieved by treatment with praziquantel. *Gastrodiscoides hominis* is unusual in that it attaches to the mucosa of the large bowel.

Alariasis

In North America, various species of the fluke *Alaria* are found in the intestines of wild carnivores such as wolves, foxes, bobcats, and skunks. The first intermediate hosts are snails and the second intermediate hosts are frogs and tadpoles. Cases of visceral larva migrans (sometimes fatal), ocular disease, and subcutaneous nodules due to *Alaria* mesocercariae have been described, usually following ingestion of undercooked frogs' legs. Other than surgical excision, no treatment has been described.

Further reading

Africa CM, De Leon W, Garcia EY (1940). Visceral complications in intestinal heterophydiasis of man. *Monographic series*, *Acta Medica Philippina*, No. 1 June.

Butcher AR *et al*. (1998). First report of the isolation of an adult worm of the genus *Brachylaima* (Digenea: Brachylaimidae) from the gastrointestinal tract of a human. *International Journal of*

Parasitology **28**, 607–10.

Chai JY *et al*. (1991). Intestinal trematodes infecting humans in Korea. *Southeast Asian Journal of Tropical Medicine and Public Health* **22**(Suppl), 163–70.

Chai JY *et al*. (1997). Two endemic foci of heterophyids and other intestinal fluke infections in southern and western coastal areas in Korea. *Korean Journal of Parasitology* **36**, 155–61.

Connor DH *et al*., eds (1997). *Pathology of infectious diseases*, Vol 2, pp 1305–588. Appleton & Lange, Stamford.

Cross JH, ed. (1991). *Emerging problems in food-borne parasitic zoonosis: impact on agriculture and public health*. Thai Watana Panich Press Co. Ltd.

Department of Parasitology, Chiang Mai University, Thailand. http:/www.medicine.cmu.ac.th/dept/parasite/official.p image.htm/trematodes

Hong SJ *et al*. (1996). One case of natural infection by *Heterophyopsis continua* and three other species of intestinal trematodes. *Korean Journal of Parasitology* **34**, 87–9.

Huffman JE, Fried B (1990). *Echinostoma* and echinostomiasis. *Advances in Parasitology* **29**, 215–69.

Kaewkes S *et al*. (1991). *Phaneropsulus spinicirrus* n. sp. (Digenea: Lecithodedriidae), a human parasite in Thailand. *Journal of Parasitology* **77**, 514–6.

McDonald HR *et al*. (1994). Two cases of intraocular infection with *Alaria* mesocercaria (Trematoda). *American Journal of Ophthalmology* **117**, 447–55.

Pungpak S *et al*. (1998). Treatment of *Opisthorchis viverrini* and intestinal fluke infections with praziquantel. *Southeast Asian Journal of Tropical Medicine and Public Health* **29**, 246–9.

Radomyos P, Bunnag D, Harinasuta T (1985) Report of *Episthmium caninum* (Verma, 1935) Yamaguti 1958 (Digenea: Echinostomatidae) in man. *Southeast Asian Journal of Tropical Medicine and Public Health* **16**, 508–11.

7.17 Non-venomous arthropods

J. Paul

Almost one million arthropod species have been described and it is likely that millions more await description. Most arthropods are of no medical importance. Medical problems they pose include envenoming, biting, transmission of infectious agents, allergy, infestation, and phobias. Arthropods may act as intermediate hosts of parasites and may cause nuisance by crawling over the skin, by making loud monotonous noises, or by invading dwellings. Most medically important arthropods are in the classes Insecta or Arachnida. Arthropod-related problems commonly present either as a particular clinical manifestation, such as bites or infestation, without an obviously visible causative agent, or as a problem visibly related to a specific kind of arthropod. Schemes of classification based on clinical manifestations and their likely causes and on the taxonomic arrangement of arthropods and their medical significance provide two useful approaches towards understanding arthropod-related problems.

Bites

Arthropod bites are common. They may be important because of the immediate physical discomfort of the bite, sensitization leading to pruritus, excoriation and secondary infection, other immunological phenomena including anaphylaxis, the transmission of infectious agents, and in exceptional circumstances blood loss. Reaction to bites varies with age, past exposure, and other factors which influence immune response. When the patient is able to associate bites with a particular kind of arthropod, management may be directed towards treatment of the bite if necessary (topical corticosteroids, systemic antihistamines), consideration of the risk of transmitted infection, and prevention of further bites (eradication of ectoparasites, change in behaviour to avoid exposure, repellents, special clothing, insecticide-impregnated bednets). It is often possible to associate bites with infesting ectoparasites, to arthropods which remain attached (ticks), and to predatory bloodsuckers which are highly visible (mosquitoes, midges, and black flies, when swarming) and which cause immediately painful bites (tsetse flies, some mosquitoes, tabanid flies). In is harder to ascribe a cause to bites from arthropods which bite at night or when the patient is asleep (some mosquitoes, sand flies, bedbugs, triatomine bugs) or from arthropods which are inconspicuous and which do not cause immediately painful bites (harvest mites, some fleas, some biting flies). Bites of larger arthropods typically have a central punctum and a surrounding area of inflammation and are pruritic. In cases of uncertainty it may be necessary to obtain a dermatological opinion to exclude other diagnoses, including organic disorders, artefact, and delusion.

Blood-sucking flies (Diptera)

Many flies are haematophagous (Table 1). Most blood-sucking flies are in the suborder Nematocera (mosquitoes, sand flies, black flies, biting midges) and the family Tabanidae of the suborder Brachycera (horse flies, clegs). The tsetse flies, *Glossina* spp., are in the suborder Cyclorrhapha. All blood-sucking flies are at least a nuisance: the bites are often painful and associated with sensitization. More importantly, biting flies may transmit infection. Mosquitoes (Culicidae) are vectors of filariasis and numerous viral diseases, including yellow fever and dengue. Mosquitoes of the genus *Anopheles* transmit malaria. Depending on species and location, mosquitoes bite at different times of the day. Mosquitoes need stagnant water for the development of their larval stages. Mosquitoes may be controlled by reducing their access to stagnant water and by application of insecticides to dwellings. Use of permethrin-impregnated bednets has been shown to reduce malaria transmission. Sand flies (Phlebotominae) are mainly tropical and subtropical in distribution and transmit leishmaniasis. In South America, sand flies of the genus *Lutzomyia* transmit *Bartonella bacilliformis*. Black flies (Simuliidae) occur worldwide but in Britain are rarely troublesome to humans except in certain localities, notably by the River Stour, Dorset. In Africa, simuliids transmit onchocerciasis, and in South America they are associated with the haemorrhagic syndrome of Altimira, but in Britain they are merely a nuisance (Blandford fly, *Simulium posticatum*). Black flies pierce the skin and suck blood from the edge of the puncture. The bites, oozing blood, have a characteristic appearance and may be associated with severe reaction by the host. Black fly larvae require fast-flowing water. Biting midges (Ceratopgonidae) are vectors of the filarial worms *Dipetalonema perstans* and *Mansonella ozzardi*. In Africa, tabanid flies transmit *Loa loa*. Tsetse flies, are vectors of African trypanosomiasis. When visiting locations where biting flies are troublesome, bites may be avoided to some extent by wearing clothing which covers the skin and by use of repellents.

True bugs (Hemiptera)

The two main groups of medically important Hemiptera are the bedbugs, *Cimex* spp., and the triatomine reduviid bugs, including *Rhodnius prolixus* and *Triatoma infestans*. The common bedbug *Cimex lectularius* (Plate 1) is cosmopolitan. The tropical bedbug *Cimex hemipterus* occurs in tropical and subtropical countries. There is no clear evidence to implicate bedbugs as vectors of disease. Bedbugs are nocturnal, hiding during the day and feeding at night. Although in some cases, bites may go unnoticed and there may be no allergic reaction, bedbugs may cause sleeplessness and the bites may cause pain and swelling (Fig. 1). Where a room is heavily infested, patients may complain of an unpleasant odour produced by the bugs. Bugs may be found by making special searches at night or by searching their hiding places during the day. Bedbugs superficially resemble lentils, being round and flat. Adults reach a length of about 5 mm. Nymphs pass through five instars to reach adulthood after about 4 months. Bedbugs can live for 6 months without feeding, becoming paper-thin. Bedbugs may be translocated in furniture and personal effects. Control relies on removal or steam cleaning of infested mattresses and treatment of infested rooms with insecticides. Related bugs which occasionally bite humans are the pigeon bug *Cimex columbarius*, the bat bug *Cimex pipistrelli*, and the martin bug, *Oeciacus hirundinis*. Infestation may be managed by restricting the access of host species to dwellings. In Britain, bats are protected under the Wildlife and Countryside Act.

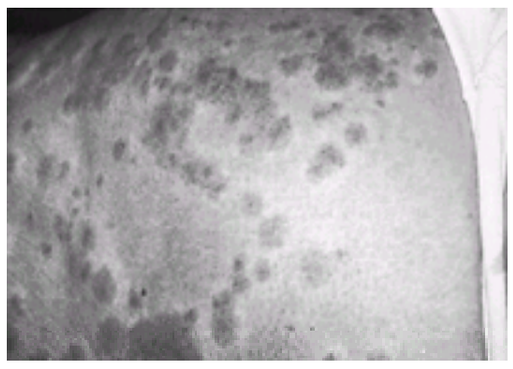

Fig. 1 Erythematous macules of bedbugs. (Reproduced by courtesy of D. Hill, Adelaide.)

Most of the 129 species of cone-nose bugs (family Reduviidae, subfamily Triatominae) occur in the Americas. Seven species occur in Asia and one species, *Triatoma rubrofasciata*, is cosmotropical. Many triatomines are obligate feeders on the blood of vertebrates. Triatomines transmit South American trypanosomiasis. Important vector species are *Rhodnius prolixus*, *Triatoma infestans*, *Triatoma brasiliensis*, *Triatoma dimidiata*, and *Panstrongylus megistus*. The bugs infest dwellings, hiding in crevices during the day and biting at night. Dwellings may be heavily infested: in Colombia, 11 403 specimens of *Rhodnius prolixus* were reported from one house, occupied by nine people, all seropositive for trypanosomiasis. As well as transmitting trypanosomiasis, triatomines may cause significant blood loss to occupants of infested buildings. Control depends on deinfestation of dwellings with insecticides and on the construction of buildings which offer few hiding places for the bugs.

Ticks (Ixodoidea)

Ticks bites are often recognized as such because ticks may remain attached to the skin for days. Hard ticks (Ixodidae) and soft ticks (Argasidae) occur worldwide. Stages of the lifecycle are egg, larva (six-legged), and nymph and adult (both eight-legged). Ticks attach and feed with a barbed hypostome and detach when engorged. Smaller stages and ticks in inconspicuous sites, such as on the perineum, may feed unobserved. Bites are usually painless but may result in local sensitization, secondary infection, and transmission of infectious agents, including numerous viruses, rickettsias, and Lyme disease (Table 2). Local reaction to bites may be confused with erythema migrans of Lyme disease, (which expands and typically develops a cyanosed centre). Ticks may be removed by gripping with forceps (or in the field, with finger and thumbnail), between the skin and the tick's head and pulling gently. One should avoid squeezing the tick. Careless removal may detach the hypostome, a potential source of secondary infection or inflammation. In Britain, ticks most often found on humans are the sheep tick *Ixodes ricinus* (a vector of Lyme disease), and the hedgehog tick *Ixodes hexagonus* (Plate 3). When visiting tick-infested places, bites may be avoided by tucking trousers into boots and by wearing light-coloured clothing which makes ticks highly visible. After visiting tick-infested habitats, searches of the body allow prompt removal of ticks which reduces the chance of disease transmission.

Harvest mites (Trombiculidae)

In Britain, larvae of the harvest mite *Neotrombicula autumnalis* are a common cause of bites in late summer. The mites are tiny and seldom noticed. They crawl rapidly on to the body, attach (often under tight-fitting clothes), inject proteolytic enzymes, feed on tissue fluid and detach, causing pruritic, sometimes bullous lesions hours later. Red bugs or chiggers (a term also applied to *Tunga penetrans*) are names given to trombiculids in the Americas. In Asia, trombolucids are vectors of scrub typhus. Where trombiculids are troublesome, tucking trousers into boots and application of diethyltoluamide and other repellents may be partially effective.

Accidental bites

Some arthropods which do not normally bite man can inflict painful but usually trivial bites when provoked by handling, as by children and entomologists: these include predatory true bugs such as the water boatman *Notonecta glauca* and the assasin bug *Reduvius personatus* in Britain and wheel bugs, *Arilus* spp., in the Americas; larger beetles (Coleoptera); dragonflies (Odonata); and bush-crickets (Orthoptera) such as the wartbiter *Decticus verrucivorus*. Spines used in defence by the great silver diving beetle *Hydrous piceus* and larger tropical grasshoppers of the subfamily Cyrtacanthridinae can cause penetrating injury when handled. Pincers of larger crabs and lobsters (Crustacea) can cause crushing injuries of digits and their spines may cause penetrating injury.

Infestation

Sites of infestation include the hair, body surface and immediate environment (ectoparasites: lice, fleas), the skin and subdermis (scabies, tungosis, dermal myiasis), wounds, tissues, and orifices (myiasis), and the gastrointestinal tract (myiasis, canthariasis). With ectoparasites, the main problems are related to their bites: diagnosis and management may be based on the identification of the ectoparasite. Delusory parasitosis is a condition in which the patient becomes convinced of infestation by parasites despite reassurance by the doctor and absence of clinical or laboratory evidence.

Scabies

The agent of human scabies, a chronic infestation, is the human scabies mite *Sarcoptes scabiei* var. *hominis*. Scabies mites adapted to other hosts, such as *Sarcoptes scabiei* var. *canis* cause a self-limiting pruritus in man. Clinical manifestations of scabies are caused by the adult female mite which burrows through the epidermis. The adult female is oval and about a third of a millimetre long (Fig. 2). The female lives for about a month, burrowing and ovipositing daily. The burrow may extend to a centimetre in length. Six-legged larvae hatch after a few days and moult to become eight-legged nymphs and later eight-legged adults. Adult males are smaller than females, do not burrow, and die after mating on the epidermis. Scabies is cosmopolitan in distribution. Prevalence rates vary but may be higher in conditions of overcrowing and following social disruption in wartime. Outbreaks may occur in nursing homes and in hospitals. Most cases must be acquired by close contact as the mites do not survive long away from the body. The main presenting symptom is pruritus which occurs with sensitization about a month after the onset of infestation. Symptoms may be worse at night and after a hot bath or shower. Burrows commonly occur in web spaces between the fingers and on the wrists but may be very widespread. There is often evidence of excoriation but the appearance of the skin is variable and may show secondary infection, eczematization, lichenification, and papulovesicles (Fig. 3). Careful examination may reveal burrows and mites. Diagnosis may be confirmed by microscopy of scrapings from affected areas, especially interdigital spaces, but many cases are atypical and a dermatological opinion may be required to exclude other causes (Fig. 4). Immunosuppressed patients, including transplant recipients and patients with AIDS, are prone to crusting or so-called Norwegian scabies in which crusting lesions of scales and mites accumulate over the hands, feet, and other sites such as the eyebrows, but the patient suffers relatively little discomfort. Such cases and presumably their fomites are highly contagious. Treatment of scabies is by topical application of acaricides. Malathion and permethrin are currently recommended, applied twice, one week apart, in the United Kingdom. Gamma benzene hexachloride is also effective. To prevent reinfection, close contacts should be treated simultaneously. During outbreaks, it may be necessary to treat whole cohorts of patients or healthcare teams. Ivermectin, by mouth, has been used to treat cases of crusting scabies in immunosuppressed patients. Occasionally, the mites *Dermanyssus gallinae* and *Ornithonyssus* spp. whose normal hosts are birds, bite humans, causing lesions which resemble scabies.

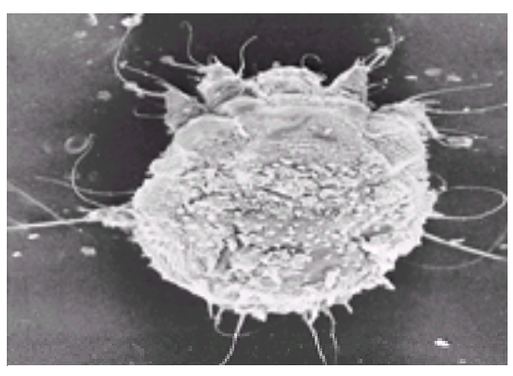

Fig. 2 Adult specimen of *Sarcoptes scabiei.* (Reproduced by courtesy of R.V. Southcott, Adelaide, South Australia.)

Fig. 3 Secondarily infected scabies in mother and child. (Reproduced with permission from Reeves and Maibach (1984). *Clinical dermatology illustrated: a regional approach*. ADIS Health Science Press, Australia.)

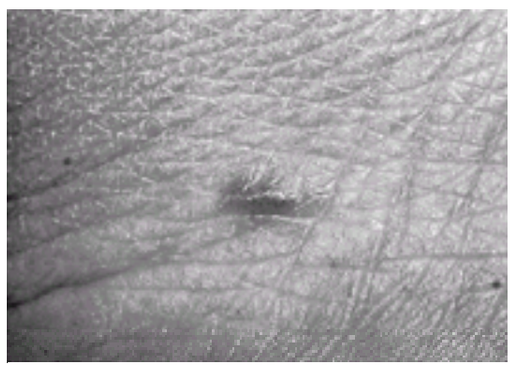

Fig. 4 Papulovesicular lesions of scabies.

Louse infestation

Lice are obligate parasites of animals. They bite using piercing mouthparts to feed on blood or tissue fluids. Three species, of cosmopolitan distribution, are associated with man: *Pthirus pubis*, the pubic louse, the body louse (or clothing louse) *Pediculus humanus*, and the head louse *Pediculus capitis*. Body and head lice are morphologically similar and are treated by some authors as subspecies or forms of *P. humanus*. Lice complete their lifecycle on their host. Adult females deposit eggs (nits) on hairshafts (pubic and head lice) or on clothing (body louse). Larvae hatch after about 1 week, begin to feed, and over the course of about 2 weeks undergo several moults before reaching adulthood. Adult females live for about a month and may lay about a hundred eggs. Egg cases remain where attached and may persist after successful treatment of infestation. Most infestations are probably acquired through close contact with an infested case but some cases may result from contact with clothing, bedclothes, or hairbrushes containing living lice or their eggs which may be attached to shed hairs. In addition to the aesthetic and social drawbacks of louse infestation, medical problems common to all three taxa, relate to sensitization of the host to louse antigens from bites and the resulting pruritus which may lead to excoriation and secondary infection. Louse bites have a central punctum and surrounding small red macule. Body lice may transmit a number of agents, including those of endemic typhus (*Rickettsia prowazeki*), trench fever (*Bartonella quintana*), and relapsing fever (*Borrelia recurrentis*). Louse infestation may be treated by topical application of pediculocides. Pediculocides should be used with caution in children and asthmatics.

Pubic lice (crab lice)

The lice attach themselves to pubic hairs. Rarely they may be found on eyebrows, eyelashes (phthirus palpebrarum), axillary, head, or chest hair. Eggs are deposited on hair shafts. Most infestations are probably acquired through sexual contact with an infested case. Children may acquire phthirosis at atypical sites through close contact with adults. Lice seldom stray from the body. Transmission is possible but unlikely without close contact with an infested case. The main symptom is pruritus, sometimes with excoriation and secondary infection. Bluish-grey patches (maculae caeruleae) may occur on the skin. Diagnosis is by observation of the lice, which may be difficult to find, or of eggs or egg cases attached to hairshafts. Adults are 1 to 2 mm long. The anterior legs are smaller than the other two pairs. The body is squat and crablike (body length, excluding head, about 1.2 times body width) (Fig. 5). Infestation may be treated by topical application of carbaryl or malathion to the whole body, repeated a week later to kill newly hatched larvae. (The original description contained a printing error (*Pthirus*) for phthirus (Greek for 'louse').)

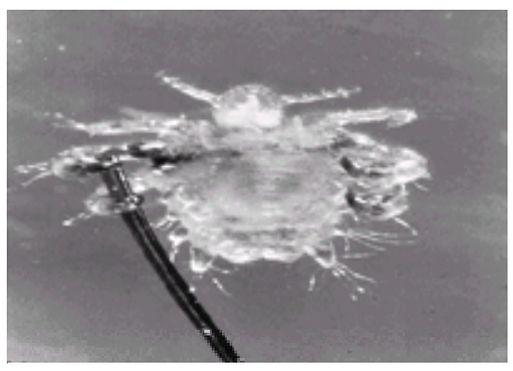

Fig. 5 Adult specimen of *Pthirus publis.* (Reproduced by courtesy of R.V. Southcott, Adelaide, South Australia.)

Head lice

Head lice infest the scalp and rarely other body sites. They lay their eggs at the base of hair shafts. Infestation is more common in children than in adults and more common in females than in males. Prevalence rates vary but may be very high in certain communities or institutions, such as schools. Prevalence rates may be high despite good standards of hygiene. Most cases occur probably as a result of close contact. The main symptom is pruritus which may be associated with excoriation, secondary infection, and lymphadenopathy. Diagnosis is by observation of lice, which generally remain close to the scalp, or of eggs or egg cases attached to hairs (Fig. 6). A fine comb (nit comb) may be used to collect material to make the diagnosis. Adults are 3 to 4 mm long. Infestation may be treated by application of pediculicide lotion to the scalp overnight and repeated a week later to destroy newly hatched larvae. Currently available pediculocides in the United Kingdom include malathion, permethrin, phenothrin, and carbaryl. Treatment failure with permethrin has been reported from many parts of the world. Compared with laboratory reference strains, lice collected from infestations failing to respond to permethrin have shown relative resistance to the agent. There is evidence that in Israel permethrin resistance may be due to mono-oxygenase plus nerve insensitivity resistance mechanisms. Malathion resistance has been reported and may be due to a

malathion-specific esterase. Regular and fastidious use of a nit comb may be used (on its own or in combination with a pediculocide) to treat infestation. There is much anecdotal evidence that combing can be effective. Combing avoids concerns of pediculocide toxicity and resistance, but a study in Wales showed combing to be less effective that chemical treatment. In institutions, co-ordinated treatment campaigns may be required to prevent reinfestation.

Fig. 6 Nits attached to hair. (Reproduced by courtesy of D. Hill, Adelaide.) Photograph from a patient with pediculosis showing several hair fibres with numerous egg cases attached.

Body lice

Body lice infest clothing and body hair. They lay their eggs on clothing, often along seams. Body lice are morphologically like head lice but they are slightly larger (Plate 4). Body louse infestation is associated with poor hygiene and social deprivation, as may occur in wartime. Transmission occurs as a result of close contact or through contact with infested clothing. Bites occur on the body, resulting in pruritus which may be associated with excoriation, eczematization, and secondary infection. Diagnosis is confirmed by finding lice, usually on the clothing. Infestation may be treated by topical application of carbaryl or malathion to the whole body, repeated a week later to kill newly hatched larvae. Hot washing of clothing will destroy adults and early stages.

Fleas (Siphonaptera)

Fleas are bloodsucking ectoparasites. There are thousands of species, adapted to various host animals. Adults are a few millimetres long, brown, laterally compressed, and typically very active. Adults move through the fur or under clothing but can survive in the environment for long periods without feeding. Eggs are dropped to the ground, where the larvae develop, feeding on organic matter. The pupa may remain in the environment for long periods before the adult emerges. Increasing standards of hygiene in developed countries have made the human flea, *Pulex irritans*, a rarity. Most flea bites in Britain are due to cat and dog fleas, *Ctenocephalides felis* (Plate 2) and *C. canis*, either through direct exposure to an infested animal or to an environment exposed to an infested animal, possibly months previously. Flea bites result in intense pruritus at the bite site. There is a central punctum and there may be bulla formation (Fig. 7). Flea bites often occur in groups. Although patients may not witness fleas, clues that bites have been caused by fleas include intense pruritus, the appearance of bites in small groups and a history of exposure to a flea-ridden animal or its domain. Troublesome bites may be treated with topical corticosteroids and systemic antihistamines. Prevention of bites is by good domestic hygiene and treatment of infested animals and environments with insecticides. Certain species of flea are vectors of a number of infectious diseases including plague and murine typhus.

Fig. 7 Fleabites. Erythematous macropapule with central bite point visible. (Reproduced by courtesy of D Hill, Adelaide.)

Tungosis

Tungosis is infestation by a flea, *Tunga penetrans*, the jigger, chigger, or chigoe (but popular names are shared with tromboculid mites). The gravid female, about 1 mm long, burrows into exposed skin (usually the foot) or under a toenail and swells to about 1 cm in diameter, causing local discomfort. Lesions may be enucleated surgically, the diagnosis being confirmed by histology. Local remedies in endemic areas (tropical Africa and the Americas) of shelling out fleas may leave cavities prone to secondary infection and tetanus. The wearing of footwear prevents infestation.

Myiasis

Myiasis is the infestation of living animals by the larvae of flies (Diptera). Useful schemes of classification of myiasis include those based on the anatomical site (dermal, subdermal, wound, nasopharyngeal, orbital, ophthalmic, aural, urogenital, pulmonary, intestinal) and on the species of fly involved. Myiasis caused by flies whose larvae are obligate parasites of living tissues may be termed specific or primary myiasis. Myiasis associated with larvae which feed on decaying organic matter may be termed opportunistic or secondary myiasis. Myiasis due to larvae which find there way into the body (especially the gastrointestinal tract) by chance may be called accidental myiasis. Of the many species listed as possible agents (Table 3), most are opportunists whose saprophagous larvae feed on decaying organic matter, which might include necrotic wound tissue. Opportunists usually confine themselves to dead tissue and may even benefit the healing process. There is no dipterous obligate intestinal parasite of man. Intestinal myiasis may be caused by coprophagous larvae which invade the rectum or by resilient maggots, such as those of the false stable fly *Muscina stabulans* and the cheese skipper *Piophila casei* which survive when swallowed in food and may cause intestinal disturbance and scarring. Intestinal myiasis may be spurious following diagnosis based on observation of rapidly hatching larvae on freshly passed faeces. Flies from several genera, notably *Fannia*, may cause urogenital myiasis. Scuttle flies (Phoridae) have been reported to cause pulmonary myiasis, possibly following inhalation of the gravid female fly. A small number of flies are obligate parasites of living tissues and a few species are closely associated with, but not specific to, humans. Many cases of myiasis are benign, self-limiting, and relatively harmless, but aural, nasopharyngeal, and malign wound myiasis are potentially lethal entities that may require removal of the larvae and possibly reconstructive surgery. Myiasis is diagnosed by observing dipterous larvae in a lesion. Identification of larvae may require entomological expertise but management of the patient, which depends on the type of lesion, may involve the removal of larvae, surgical exploration, debridement, or treatment of secondary infection and should be based on clinical assessment.

Dermal myiasis

Dermatobia hominis, the human bot fly, is a common cause of dermal myiasis in the American tropics. The female fly lays her eggs on biting arthropods such as mosquitoes. The eggs hatch when in contact with skin into which the larva burrows. The larval stage lasts about 10 weeks (Fig. 8), a boil with a small aperture forming as the larva grows. Such boils are not infrequently seen in Europeans returning from the neotropics. The larva may grow to more than a centimetre in length. An early symptom is sporadic pain caused by the spiny larva. Unless in an unusual anatomical site, such as close to the eye, infestation is generally harmless. Secondary

infection of the wound is the most common complication. Larvae may be removed through a simple incision. Remedies which include application of raw meat or glue to the lesion may not be successful. Squeezing may rupture the larva to evoke a local granulomatous reaction.

Fig. 8 Two third larval instars of the human bot fly (*Dermatobia hominis*) (approximately 13 mm long) extracted from a facial 'boil' in a European who had been visiting Guyana. (b) Larva of *Dermatobia hominis* initially infesting the scalp of a young child in Panama. The larva made a 4-mm hole in the anterior fontanelle and entered the frontal lobe of the brain. The child died of malaria. (Armed Forces Institute of Pathology photograph, neg. no. 50807.)

The tumbu fly, *Cordylobia anthropophaga*, is widespread in the Afrotropical region. The female oviposits on sand and also on drying clothes. Ironing destroys eggs. Contact with viable ova on clothing leads to infestation. The larvae (Plate 6) pierce the skin and grow rapidly. An uncomfortable boil forms which oozes serosanguinous fluid. Fever and lymphadenopathy may occur. Larvae reach maturity in about 10 days. Larvae may be removed through a simple incision, but with care it may be possible to express larvae following application of petroleum jelly.

The larvae of warble flies, *Hypoderma* spp., occasionally cause dermal myiasis in man. Larvae of horse bot flies, *Gasterophilus* spp., cannot complete their lifecycle in man but they can pierce human skin, where they wander for a week or so, causing intense itching (creeping eruption).

Wound myiasis

Many dipterous species are known to cause wound myiasis, but most of them are facultative feeders on necrotic tissue and are rarely destructive to the host, although the presence of maggots in a wound may cause distress. Debridement of nectrotic tissue will control such infestation. In contrast, under controlled conditions, clinicians may introduce maggots to promote healing.

Causes of malign myiasis include the New World screw-worm *Cochliomyia hominivorax*, in the Americas and the Old World screw-worm *Chrysomya bezziana* and Wohlfahrt's wound myiasis fly *Wohlfahrtia magnifica* in the Old World. Their larvae are obligate parasites of living tissue. Eggs are laid on wounds, in ears, and on mucous membranes. The larvae (Fig. 9) burrow in groups into healthy tissue, causing widespread destruction which may be mutilating or fatal (Fig. 10). Secondary bacterial infection or secondary wound myiasis may ensue. All species may cause nasopharyngeal, aural, orbital, genital, and malign wound myiasis. Infestation is best avoided by cleaning and dressing wounds as they occur. Treatment involves surgical removal of the larvae, debridement of affected tissue, and treatment of secondary infection. Reconstructive surgery may be required.

Fig. 9 Larvae of the New World screw-worm (*Cochliomyia hominivorax*) (approximately 8 mm long) extracted from the wound illustrated in Fig. 10. These were sent to the Natural History Museum, London, where they were identified. Larvae of the second myiasis species (*C. macellaria*) were also found in the sample and were probably collected from the edges of the wound. (By courtesy of Dr Martin JR. Hall, Medical and Veterinary Division, The Natural History Museum, London.)

Fig. 10 Fatal myiasis (New World screw-worm): historical illustration of a 50-year-old Honduran woman who complained of a small chronic ulcer on the right cheek; on admission to hospital she was found to have a huge ulcer exposing the bones of the face and forehead and destroying the tissues of the cheek and face, right eye, and orbit; more than 300 larvae were removed (see Fig. 9). (Harrison JHH (1908). A case of myiasis. *Journal of Tropical Medicine and Hygiene*, **XI**, 20.)

Ophthalmic myiasis

Nasal bot flies, *Oestrus* spp., naturally parasitize various herbivorous mammals. They are larviparous and drop their larvae into the nostrils of the host. Dropped into human eyes they cause a self-limiting conjunctivitis. Larvae of warble flies, *Hypoderma* spp., are more dangerous: they may burrow into the eye, resulting in pain, nausea, and much damage and must be surgically removed.

Canthariasis

Infestation of the body by beetles (Coleoptera) or their larvae is called canthariasis. Clinically, it may resemble myiasis but is much rarer. Larvae swallowed with food may dwell temporarilly in the intestines, causing discomfort and may be detected in excreta. Beetles occasionally invade orifices. In Sri Lanka, scarabid dung beetles have been reported to invade the rectum. A specimen of the ground beetle *Sciates sulcatus* was recovered from the vagina of a women complaining of vaginal

discharge who had visited Pakistan (Plate 5). In Israel, the dung beetle *Maladera matrida* has been reported to invade the external auditory canal. In Oman, two cases of invasion of the external auditory canal by the ground beetle *Crasydactylus punctatus* have been reported. In one case, the beetle reached the middle ear causing sensorineural hearing loss.

Allergy

A wide range of immunological responses to arthropod bites has been described, from local pruritus to anaphylaxis. The dead remains, cast skins (exuviae), and faeces of many arthropods include sensitizing agents. They may act as contact or inhalant allergens following domestic or occupational exposure, resulting in dermatitis, conjunctivitis, rhinitis, and asthma. Allergic patients may show specific IgE antibodies to a wide range of domestic pests including house flies, clothes moths, cockroaches, carpet beetles, *Anthrenus* spp., silverfish, *Ctenolepisma longicaudata*, and house dust mites *Dermatophagoides* spp. *Dermatophagoides* spp. are a common cause of allergy in Britain and exposure to cockroach allergens in household dust has been associated with asthma in the United States. Following mass emergence, the exuviae of mayflies (Ephemeroptera) and caddis flies (Trichoptera) may act as inhalant allergens. Entomologists who collect insects by sucking them into pooters may develop inhalant allergy to their subject of study. Larvae of the beetles *Tenebrio molitor* (mealworm) and *Alphitobius diaperinus* (lesser mealworm), which are reared for fish bait and animal food, have been associated with rhinoconjunctivitis, contact urticaria, and asthma. Beetles which infest stored grain, including *Tenebrio molitor*, *Tribolium confusum* (confused flour beetle), *Sitophilus* sp. (grain weevil), and *Alphitobius diaperinus* have been associated with occupational allergy in grain workers or bakers. Allergy has been associated with other beetles, including *Dermestes peruvianus* (hide beetle), *Gibbium psylloides* (mite beetle), and *Harmonia axyridis* (Asian ladybird). Insect allergy can be investigated by skin prick tests, measurement of allergen-specific serum IgE, and by monitoring respiratory function following allergen exposure.

Insects and hygiene

Synanthropic insects which feed or wander over faeces, wounds, and food may serve as passive vectors of bacterial and viral diseases. Such insects include pharaoh's ants, *Monomorium pharaonis*, flies, and cockroaches (Dictyoptera). Despite many reports of the isolation of pathogenic bacteria and viruses from these insects, there have been few epidemiological studies to define their importance as passive vectors but is generally accepted that the presence of these insects in hospitals should be monitored and controlled.

Flies

Many species of fly (especially of the suborder Cyclorrhapha), frequent human and animal food, wounds, eyes, and faeces. Such flies vomit and defaecate where they feed. Numerous pathogenic bacteria and viruses have been isolated from flies, suggesting that they may act as passive vectors of bacterial and viral diseases. A controlled study in The Gambia where fly control was associated with fewer new cases of trachoma, suggested that flies may act as vectors of the trachoma agent, *Chlamydia trachomatis*. In The Gambia *Musca sorbens* is the most common eye-visiting fly. In Pakistan, a controlled study showed fly control to be significantly associated with a reduction in incidence of childhood diarrhoeal illness. In Israel, fly control was associated with a reduction in cases of shigellosis. Flies may be controlled by using insecticides or fly traps in dwellings and latrines.

Pharaoh's ants

Pharaoh's ants, *Monomorium pharaonis* L., commonly infest hospitals, where they invade sterile packs and wound dressings. They are potential passive vectors: bacteria, including *Salmonella* spp. and *Staphylococcus* spp. have been isolated from these ants, which should, therefore, be controlled with insecticides.

Cockroaches

Cockroaches are omnivorous scavengers. A few of the 3500 described species have become cosmopolitan synanthropes. The main pest species are the common cockroach, *Blatta orientalis*, the American cockroach, *Periplaneta americana*, the German cockroach, *Blattella germanica*, and the banded cockroach, *Supella longipalpa*. Other species may be locally important, for example *Ectobius lapponicus*, described by Linnaeus as infesting dried fish in Lapland. The common pest species are mostly of tropical origin and require temperatures of 25 to 33 °C but *B. orientalis* will tolerate 20 °C. In cooler climates they are restricted to permanently heated areas and can occur in large numbers in hospitals and in sewers. Many pathogenic viruses, including poliomyelitis virus and Coxsackie A virus, and bacteria, including *Shigella* spp., have been isolated from cockroaches. There is evidence that cockroaches acted as vectors of hepatitis A during an outbreak in California and of *Salmonella typhimurium* on a paediatric ward in Belgium. Cockroaches are potential allergens, 7.5 per cent of healthy persons having a positive skin test in one study. Cockroaches wander over sleepers and are attracted to nasal and oral secretions. Herpes blattae is a dermatitis described from Réunion and attributed to cockroach allergy. Cockroaches sometimes wander into the ears and nostrils, where they become trapped or reluctant to leave. Lidocaine spray is reported to hasten the exit of such visitors.

Eye-frequenting moths and beetles

Some nocturnal moths of the families Pyralidae, Noctuidae, and Geometridae in Africa and Southeast Asia habitually feed on the lachrymal secretions of animals. They may visit human eyes, causing a certain amount of discomfort, and may transmit eye infections, including trachoma and viral conjunctivitis. They may also cause mechanical damage to the cornea. The moths stimulate the flow of secretions by vibrating and probing with their probosces. Implicated species include *Lobocraspis griseifulva*, *Arcyophora* spp., and *Filodes fulvidorsalis*. *Calyptra eustrigata* is a skin-piercing, blood-sucking noctuid moth from Malaya. Such Lepidoptera may be avoided by sleeping under a net. In Australia a beetle, *Orthoperus* sp., has been associated with corneal erosion.

Further reading

Alexander JO'D (1984). *Arthropods and the human skin*. Springer, Berlin.

Auerbach PS, ed. (1995). *Wilderness medicine: management of wilderness and environmental emergencies*. Mosby, St Louis, MO.

Baker AS (1999). *Mites and ticks of domestic animals: an identification guide and information source*. The Stationery Office, London.

Hope FW (1840). On insects and their larvae occasionally found in the human body. *Transactions of the Royal Entomological Society* **2**, 256–71.

James MT (1947). *The flies that cause myiasis in man*. United States Department of Agriculture, Washington.

Roberts DT, ed. (2000). *Lice and scabies: a health professional's guide to epidemiology and treatment*. Public Health Laboratory Service, London.

Rosenstreich DL *et al.* (1997). The role of cockroach allergy and exposure to cockroach allergen in causing morbidity among inner-city children with asthma. *New England Journal of Medicine* **336**, 1356–63.

Roth LM, Willis ER (1957). The medical and veterinary importance of cockroaches. *Smithsonian Miscellaneous Collection* **134**, 1–147.

Smith KGV, ed. (1973). *Insects and arthropods of medical importance*. British Museum (Natural History), London.

Zumpt F (1965). *Myiasis in man and animals in the Old World*. Butterworth, London.

7.18 Pentastomiasis (porocephalosis)

D. A. Warrell

The Pentastomida, pentastomes or 'tongue worms', inhabit the respiratory tracts of vertebrates where they feed on blood and other tissues. There are more than one hundred species, classified into two orders, Cephalobaenida (e.g. genus *Raillietiella*) and Porocephalida (e.g. genera *Linguatula*, *Armillifer*, *Leiperia*, and *Sebekia*). About ten species are known to be capable of causing zoonotic infections in humans. Pentastomida are probably arthropods, but they have also been classified as Branchiuran crustacea, annelids, and in a separate phylum. The name pentastome derives from their having two pairs of anterior hooks and a mouth, giving the impression of five stomata (Fig. 1). In humans, visceral pentastomiasis is most often caused by *Linguatula serrata* or *Armillifer armillatus* and nasopharyngeal pentastomiasis (halzoun or Marrara syndrome) by *L. serrata*.

Fig. 1 Adult pentastomid showing mouth (arrowed) and lateral hooks giving the appearance of five stomata. Scanning electron micrograph, 400 ×. (By courtesy of Professor Viqar Zaman.)

Aetiology

Linguatula species

Linguatula serrata occurs in Europe, the Middle East, Africa, and North, Central, and South America, but not Asia. The names 'Linguatula' and 'tongue worm' describe the flattened shape, particularly of the adult female. Their surface bears numerous annular grooves. Dogs, foxes, and wolves, the definitive hosts, harbour adults and nymphs in their upper respiratory tract and shed them in their nasal secretions, saliva, and faeces. Herbivorous animals ingest the ova which hatch in the lumen of the gut releasing larvae that burrow into the tissues and encyst. When these intermediate hosts are eaten by the definitive host, nymphs hatch from the cysts and migrate to the lungs and nasopharynx where they mature.

Clinical features

If humans ingest ova of *Linguatula* spp., cysts usually develop in the liver but do not cause symptoms unless they obstruct or compress, for example, the filtration angle of the anterior chamber of the eye (2nd- or 3rd-instar larvae), biliary, gastrointestinal or respiratory tracts, meninges, or brain. Ingestion of cysts containing third-stage larvae in raw liver and lymph-nodes of sheep, goats, cattle, and lagomorphs can result in acute nasopharyngitis, known as halzoun, Marrara syndrome, or nasopharyngeal pentostomiasis. This has been reported from the Middle East, especially Lebanon, Greece, and the Sudan. In the human stomach, larvae escape from the cysts and migrate up the oesophagus to the nasopharynx mucosa. Within a few hours of eating the infected viscera there is intense irritation of the upper respiratory and gastrointestinal tracts associated with coughing, sneezing, rhinorrhoea, retching, vomiting, lacrimation, haemoptysis, epistaxis, cervical lymphadenopathy, transient deafness, difficulty in speaking, dysphagia, wheezing, and dyspnoea. Larvae can be found in sputum and vomitus. Patients usually recover in 1 or 2 weeks, but deaths have resulted from acute upper airway obstruction. These features suggest a hypersensitivity reaction. Halzoun has also been attributed to flukes (*Fasciola hepatica*) and nematodes (*Mammomonogamus laryngeus*) ingested in raw sheep and goat liver and to leeches (*Limnatis nilotica* and *Dinobdella ferox*). Very rarely, larvae may develop into adults in the human nasal cavity.

Armillifer (*Porocephalus*) species

These are also annulated, non-segmented parasites (Fig. 2). The adults, up to 20 cm long, inhabit the respiratory and digestive tracts of snakes, especially those of the genera *Python*, *Boaedon*, *Naja*, and *Bitis*, and other vertebrates. Ova are shed in the snake's nasal secretions and are picked up by herbivorous mammals. Larvae encyst in the tissues of these intermediate hosts and will develop to the nymph stage if ingested by another animal, but develop to adults only in snakes. Humans may ingest ova by drinking water contaminated by snakes, or they may ingest living encysted larvae in raw snake meat eaten as part of *ju ju* rituals (West Africa), or from inadequately cooked snake (Temuan tribe of Malaysian aborigines and in Benin and other West African countries). Ingested eggs hatch in the gut releasing larvae which burrow into the tissues where they develop into nymphs. Hundreds of wriggling nymphs have been discovered beneath the visceral peritoneum at laparotomy.

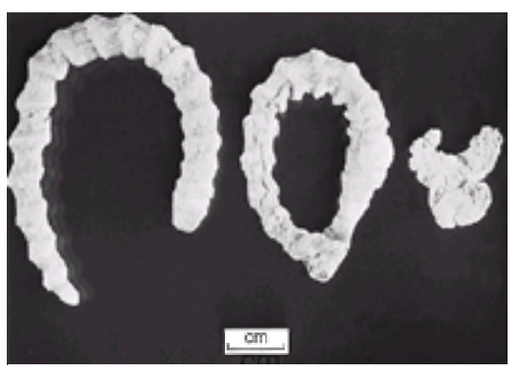

Fig. 2 *Armillifer armillatus*. Left: two adults found in the lungs of a Gaboon viper (*Bitis gabonica*). Right: calcified nymph from the mesentery of a Ghanaian patient. (By courtesy of Dr G. M. Ardran.)

Human infections with the larvae or nymphs of the following species of *Armillifer* have been reported:

- *A. agkistrodontis*—China (in the snake *Deinagkistrodon acutus*);
- *A. armillatus* (18–22 annular rings)—Africa: Egypt, Senegal, The Gambia, Ghana, Benin, Nigeria, Cameroon, Congo, and Zimbabwe;
- *A. grandis*—Congo;
- *A. moniliformis* (30 annular rings)—Malaysia, Philippines, Indonesia, Tibet, and Australia;
- *A. najae*—India.

Clinical features

The commonest evidence of infection by *Armillifer* spp. is the discovery of calcified nymphs (Fig. 2) on radiographs of the abdomen and chest (Fig. 3). These appear as discrete, crescent-shaped, soft tissue calcifications, 4 × 4 mm in size. In West Africa they are seen particularly in the right upper quadrant and are situated beneath the peritoneum covering the liver. In Ibadan, Nigeria, these shadows were seen in 2 per cent of adult males and 4 per cent of adult females. Hundreds of calcified encysted nymphs have been found in the peritoneum at laparotomy or at autopsy in the liver (Fig. 4), spleen, gut wall and lumen, lungs, cirrhosal cavities, central nervous system, eye, and elsewhere in 27 per cent of cases in The Congo (*A. armillatus* and *A. grandis*), in 6 to 13 per cent in Cameroon (*A. armillatus*), and in 45 per cent in Malaysian Orang Asli (*A. moniliformis*).

Fig. 3 Typical radiographic appearance of calcified nymphs of *Armillifer armillatus* in the abdominal cavity of a Ghanaian patient. (By courtesy of Dr G. M. Ardran.)

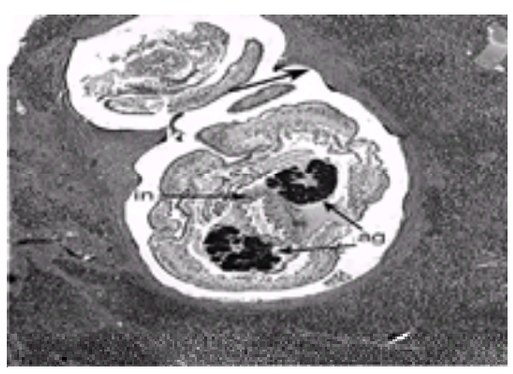

Fig. 4 Encysted nymph/larva of *Armillifer armillatus* in human liver. The outer layer of the parasite (arrowed) lines the cyst wall. Acidophilic glands (ag), intestine (in), 21 ×. (Armed Forces Institute of Pathology photograph, negative number 75–2703.)

Armillifer spp. infection is usually symptomless or causes vague abdominal pain. Serious inflammatory and obstructive effects have been described in the gut, lungs, biliary tract, pericardium, central nervous system, and anterior chamber of the eye. Severe acute reactions may be related to hypersensitivity or perhaps to massive infection such as might follow ingestion of a gravid female. Migration of large numbers of larvae from the gut into the tissue might produce abdominal pain and obstructive jaundice. A few fatal cases have been reported, including one patient who died from intestinal obstruction caused by nymphs of *A. armillatus*. It has been suggested (Nigeria and The Congo) that infection by *Armillifer* spp. might be associated with malignancy of the colon and elsewhere.

Other pentastomid infections

Human infections with *Leiperia cincinnalis* have been described in Africa, and subcutaneous infections by *Raillietiella gehyrae* and *R. hemidactyli* in Vietnam and by *Sebekia* species in Costa Rica. In Vietnam, infection with *Raillietiella* spp. follows the swallowing of small live lizards for medicinal purposes.

Diagnosis

The radiographical appearances of calcified pentastomid nymphs are distinctive (Fig. 3). Pentastomes may be discovered at surgery or autopsy. In the liver (Fig. 4), intestinal wall, mesentery, mesenteric lymph nodes, peritoneum, or lung, viable encysted larvae or granulomas containing necrotic pentastomes or their moulted cuticles may be identified. Initially, encysted larvae excite little or no tissue reaction, but the granulomas are surrounded by hyalinized or calcified fibrous tissue. In tissue sections, pentastomes can be distinguished from helminths. Antibodies to *Armillifer* spp. have been detected by fluorescence in infected patients.

Treatment

There is no specific treatment, although mebendazole has been suggested. Obstruction and compression should be relieved surgically. Hypersensitivity phenomena should be treated with adrenaline (epinephrine), antihistamines, and corticosteroids. Infections can be prevented if all meat is thoroughly cooked.

Other zoonoses transmitted from reptiles to humans

The most important of these is salmonellosis transmitted to humans by the faecal–oral route or by scratches and bites, from chelonians (tortoises, turtles, terrapins) and from snakes and lizards, especially iguanas. In Britain, 38 per cent of imported tortoises (*Testudo* species) contain *Salmonella* spp.; in the United States, where 8 million reptiles are kept as pets, 14 per cent of reported salmonellosis cases (280 000 per year) were attributable to pet terrapins, as were up to 17 per cent of cases of infant salmonellosis in Puerto Rico. The species include *S. typhimurium, S. muenchen, S. ealing, S. volta, S. alachua, S. stanley, S. marina, S. poona, S. pomona*, and *S. java*.

Other infections transmissible from reptiles to humans include *Arizona hinshawii* (in snake powder, Pulvo de Vibora, made from rattlesnakes), *Plesiomonas shigelloides, Edwardsiella tarda*, leptospirosis, Q fever, sparganosis, capillariasis, strongyloidiasis, mesocestoidiasis, and infestation with the mite *Ophionyssus natricis*. Potential zoonoses include *Mycobacteria, Pseudomonas*, other *Aeromonas* species, *Proteus* spp., and some togaviruses (such as western equine encephalitis in garter snakes in western North America) and herpesviruses.

Further reading

Drabick JJ (1987). Pentastomiasis. *Reviews of Infectious Diseases* **9**, 1087–94.

Fain A (1975). The Pentastomida parasitic in man. *Annales de la Societé Belge de Medecine Tropicale* **55**, 59–64.

Faisy C, *et al.* (1995). La porocéphalose, parasitose méconnue revue de la litérature à propos d'un cas congolais. *Médecine Tropicale (Marseille)* **55**, 258–62.

Haugerud RE (1989). Evolution in the pentastomids. *Parasitology Today* **5**, 126–32.

Lavarde V, Fornes P (1999). Lethal infection due to *Armillifer armillatus* (Porocephalida): a snake-related parasite. *Clinical Infectious Diseases* **29**, 1346–7.

Riley J (1986). The biology of pentastomids. *Advances in Parasitology* **25**, 45–128.

Self JT, Hopps HC, Williams AO (1975). Review. Pentastomiasis in Africans. *Tropical and Geographical Medicine* **27**, 1–13.

Warwick C, *et al.* (2001). Reptile-related salmonellosis. *Journal of the Royal Society of Medicine* **94**, 124–6.

Yagi H, *et al.* (1996). The Marrara syndrome: a hypersensitivity reaction of the upper respiratory tract and buccopharyngeal mucosa to nymphs of *Linguatula serrata*. *Acta Tropica* **16**, 127–34.

7.19 Chronic fatigue syndrome (postviral fatigue syndrome, neurasthenia, and myalgic encephalomyelitis)

Michael Sharpe

Historical introduction

Illness characterized by chronic fatigue but without a clear pathological basis has a long history. In the nineteenth century it attracted a diagnosis of 'neurasthenia' but by the early twentieth century this diagnosis fell into disuse. It is probable that since that time patients have continued to present to doctors with similar symptoms but have received other diagnoses such as depression, brucellosis, and Epstein–Barr virus infection. Apparent epidemics of fatiguing illness have been occasionally reported over the last 50 years and attributed to infection by various agents. One occurred among staff at the Royal Free Hospital, London in 1955 giving rise to the term myalgic encephalomyelitis. More recently, it has become accepted that the cause of many cases of chronic disabling fatigue is not an infection. The purely descriptive term chronic fatigue syndrome was consequently introduced in 1988. The cause and treatment of chronic fatigue syndrome remain controversial.

Introduction

The terms chronic fatigue syndrome, postviral fatigue syndrome, neurasthenia, and myalgic encephalomyelitis have all been used to describe an idiopathic syndrome characterized by disabling fatigue occurring chronically after minimal exertion. The current international consensus definition is shown in Table 1. It is now clear that there is considerable overlap between chronic fatigue syndrome and other 'functional' syndromes such as irritable bowel syndrome and fibromyalgia.

- What is fatigue? Fatigue is an imprecise term with many meanings. In clinical medicine, the symptom of fatigue implies a subjective feeling of lack of both energy and endurance.
- When is fatigue abnormal? Fatigue after exertion is a normal phenomenon. It is abnormal when it is disproportionate to the exertion, persistent, and associated with impaired function.
- Why 6 months? Six months is used to define chronicity for purposes of research. It excludes the short-lived fatigue that can follow any illness.
- Which other symptoms? Patients with chronic fatigue syndrome commonly complain of additional symptoms, which are described below under clinical features.
- What conditions must be excluded? Chronic fatigue may be a symptom of many, if not most, medical and psychiatric illnesses. The term chronic fatigue syndrome has been reserved for patients in whom the fatigue remains medically unexplained or 'idiopathic' after clinical assessment.

Aetiology

No specific aetiology has been identified. However, a number of psychological and biological factors have been identified that may play a role in the onset and/or perpetuation of illness. It is likely that different and perhaps multiple factors operate in different cases.

Infection

Patients frequently give a history suggestive of acute infection at the outset. Fatigue states lasting several months can follow infections such as Epstein–Barr virus and Q fever. Infection may therefore trigger chronic fatigue syndrome, but the available evidence does not support the hypothesis that chronic infection is the cause of chronic fatigue syndrome.

Immune dysfunction

Minor immune abnormalities have been detected in a proportion of patients. Cytokines can cause fatigue. However, no consistent immune abnormality or casual link of these with symptoms has been established in patients with chronic fatigue syndrome. The role of immune factors remains of interest but is unclear.

Stress and emotional disorder

More than half of the patients seen in hospital clinics who meet criteria for chronic fatigue syndrome also meet criteria for depressive and anxiety disorders. Many patients describe major life stresses. It has therefore been suggested that at least some cases of chronic fatigue syndrome are due to emotional disorder that is expressed somatically (somatization). This is undoubtedly true in some cases, but in many others these syndromes are either absent or are an inadequate explanation of the illness.

Psychological and behavioural factors

There is evidence that psychological and behavioural factors play a role in perpetuating chronic fatigue syndrome. They include misconceptions about the nature of

the illness, excessive avoidance of physical activities so that the person becomes inactive, repeated seeking of (ineffective) medical care, and failure to resolve continuing psychological and social problems.

Inactivity

Some patients with chronic fatigue syndrome are profoundly inactive. This may lead to muscle wasting, changes in the cardiovascular response to exertion, and postural hypotension. The consequent intolerance of activity may perpetuate the illness.

Dysfunction of the central nervous system

Abnormalities have been found in tests of cognitive function. There is also evidence for abnormalities in neuroendocrine tests and in functional neuroimaging. However, abnormalities in similar domains have been found in patients with depression and anxiety disorders. Consequently the specificity and clinical utility of such tests remains to be established.

Endocrine dysfunction

A slightly reduced 24-h cortisol excretion has been reported. Replacement therapy with low-dose hydrocortisone has been reported to produce short-term symptomatic relief. However, the balance of risks and benefits of this treatment is not yet established.

Idiopathic postural hypotension (neurally mediated hypotension)

A tendency to postural hypotension has been reported in a proportion of patients with chronic fatigue syndrome. At present the aetiological and treatment implications of this observation remain unclear.

Sleep disorder

Various sleep abnormalities have been found in patients with chronic fatigue syndrome. They include both specific sleep disorders such as sleep apnoea syndrome and non-specific abnormalities such as fragmentation of sleep. These may contribute to daytime fatigue in at least some cases.

Social and iatrogenic factors

P>Information about chronic fatigue syndrome or myalgic encephalomyelitis, whether from doctors, patient groups, or the media, that leads patients to see their illness as mysterious with a hopeless prognosis and best treated by rest is likely to be unhelpful.

Epidemiology

Prevalence

One-quarter of the general population complain of persistent fatigue. In contrast, recent population studies in the United Kingdom and United States suggest that only approximately 0.5 per cent could be regarded as having chronic fatigue syndrome. Most of these persons are aged between 20 and 40 with a predominance of females. The syndrome is also seen in children and adolescents.

Epidemics

Epidemics of a chronic fatigue-like syndrome have been described from various parts of the globe. This observation is compatible with, but does not establish, an infective cause. It remains unclear whether these were true epidemics and also whether the clinical picture reported is similar to that of cases of sporadic chronic fatigue syndrome.

Pathogenesis

A number of explanations have been proposed (see section on aetiology above). These are best considered in the categories of predisposing, precipitating and perpetuating factors:

- Predisposing factors: Certain individuals may be predisposed to develop chronic fatigue syndrome by virtue of genetics, personality, or other vulnerability.
- Precipitating factors: The condition may be precipitated by factors such as infection or psychological stress (life events).
- Perpetuating factors: For practical management the most important factors are those that perpetuate the illness and consequently act as barriers to recovery. Perpetuating factors include the modifiable psychological and behavioural factors described above as well as biological factors.

Clinical features

Main symptoms

The principal symptom of chronic fatigue syndrome is chronic mental and physical fatigue, tiredness, or exhaustion that is exacerbated by activity. Patients often report being able to perform activities for brief periods, but subsequently experiencing severe fatigue for hours or days thereafter.

Other symptoms

Other common symptoms include muscular pain, unrefreshing sleep, dizziness and breathlessness, headache, tender lymph glands, and symptoms of irritable bowel syndrome. Patients often describe day-to-day fluctuations in their symptoms, irrespective of activity. Periods of almost complete recovery may be followed by relapse, often sufficiently severe to make normal daily activity impossible. Depression and anxiety are common, and a proportion of patients suffer panic attacks. Patients and their relatives may hold strong beliefs about the nature and aetiology of their illness (see section on aetiology), and these may be of importance when planning management.

Physical signs

Physical examination is typically unremarkable. Complaints of fever and lymphadenopathy are generally not confirmed on examination. The presence of definite physical signs (such as objectively measured fever) should not be ascribed to the syndrome and alternative diagnoses should be sought.

Differential diagnosis

Almost any disease may present with unexplained fatigue. The differential diagnosis of idiopathic chronic fatigue syndrome is correspondingly large (see Table 2). The nature of the fatigue may offer useful clues. Muscular disease should be considered if the patient has objective weakness, no psychological symptoms, and a family history. If the patient's complaint of fatigue includes prominent sleepiness, a sleep disorder should be considered. In particular, prominent snoring and morning headaches in the obese patient raise the possibility of obstructive sleep apnoea.

It is important to also assess the mental state. Depression is suggested by fatigue that is worse in the morning and accompanied by loss of motivation, interest, and pleasure. Other symptoms of depression should be sought including sadness, loss of appetite and weight, and feelings of pessimism and failure. If there is evidence of depression it is essential to ask about suicide plans. Chronic anxiety is also associated with fatigue and may also give rise to many of the somatic symptoms of

chronic fatigue syndrome such as muscle pain, impaired concentration, and poor sleep.

Pathology

There is no established pathology other than muscular and other changes associated with inactivity. Reports of abnormal structural brain scans have not been confirmed.

Laboratory diagnosis

There are no diagnostic tests and no characteristic abnormalities on laboratory investigation. These are conducted purely to exclude other diseases. All patients should have a full blood count, erythrocyte sedimentation rate or C-reactive protein, basic biochemistry screen, urine analysis, and possibly thyroid function and antinuclear antibody tests. Further investigation depends on the clinical findings and differential diagnoses under consideration.

Treatment

General management

The doctor should listen to the patient's story and ask about his or her own understanding of the illness in all cases. It is usually also worth seeing the partner or relevant family members. This provides opportunities to correct misconceptions. It is especially important to explain that the illness is not progressive or life threatening (see below). A more positive and less sinister explanation of chronic fatigue syndrome as a 'reversible dysfunction of the central nervous system' with a cautiously optimistic prognosis may be offered. The adverse physiological and psychological effects of prolonged bed rest should be explained if necessary. The patient should be encouraged to adopt a consistent but gradually increasing level of activity, avoiding extremes of both inactivity and exertion. An evidence-based self-help book may be useful (such as *CSF/ME: the facts*; see Further reading). An initial hospital appointment that achieves all the above usually requires at least 45 min.

Drug therapy

A trial of an antidepressant drug may be considered, especially if there is evidence of depression. It is advisable to choose a non-sedating type and to start with a low dose. Low doses of antidepressant drugs may reduce anxiety, improve the quality of sleep, and reduce pain. If there is evidence that the patient has a depressive disorder it is important to give an adequate dose for an adequate period.

Psychological aspects of care

Many patients have ongoing difficulties in their work or personal relationships. Whether contributors to or consequences of the illness, they may need help to improve their ability to cope with these. For some patients long-term follow-up provides an opportunity to encourage them to persevere with rehabilitation and to minimize the risk of multiple referrals and resulting iatrogenic harm from conflicting advice, repeated medical investigation, and failed attempts at treatment.

Rehabilitation and cognitive behavioural treatment

A systematic review has found that rehabilitative psychological therapy (cognitive behavioural therapy) is more effective than conventional management, with significant improvement rates of 60 per cent and 25 per cent respectively at 12 months. Two trials have suggested that supervised and gradually increased physical activity alone may be effective for willing patients. If available, a general hospital liaison psychiatry or psychology service may be best placed to offer treatment in a setting acceptable to the patient. General rehabilitation services are of value for chronic severe disability.

Other treatments

Many other treatments have been proposed but none has been adequately evaluated. Patients should be discouraged from pursuing unproven treatments unless they are part of a clinical trial.

Prognosis

The prognosis for functional recovery is relatively good for patients seen in general practice. It is poor for those severe enough to be referred to hospital clinics. A long history, multiple symptoms, and entrenched belief that the illness is irreversible predict a particularly poor prognosis. Effective treatment can improve the patient's ability to function. There is no mortality associated with chronic fatigue syndrome other than suicide, which may reflect unrecognized depressive illness.

Prevention

As the cause is unknown there is no specific primary prevention. Secondary prevention is important as it is likely that good early management and avoidance of iatrogenesis will reduce the risk of chronicity.

Other aspects

Occupational, quality of life, and psychological aspects

Chronic fatigue syndrome can be associated with a markedly reduced quality of life. Occupational issues are often prominent. Occupational stress may be an aetiological factor and having chronic fatigue syndrome makes it difficult to sustain employment. The ambiguous status of chronic fatigue syndrome as an accepted disease makes it especially difficult for patients with this diagnosis to negotiate with employers, insurers, and other agencies.

Areas of controversy

Almost all aspects of chronic fatigue syndrome are controversial. Most controversy has centred on whether it is most appropriately regarded as a medical or a psychiatric syndrome. This debate should be seen in the context of the stigma associated with psychiatric illness.

Areas needing further research

Further work is needed into both aetiology and treatment. In particular a refinement of the definition of the syndrome to enable the identification of more homogeneous groups will be important in developing effective treatment.

Further reading

Campling F, Sharpe M (2000). *CFS/ME: the facts*. Oxford University Press, Oxford.

Fukuda K *et al*. (1994). Chronic fatigue syndrome: a comprehensive approach to its definition and management. *Annals of Internal Medicine* **121**, 953–9.

Komaroff AL, Buchwald DS (1998) Chronic fatigue syndrome: an update. *Annual Review of Medicine* **49**, 1–13.

Wessely S, Sharpe M, Hotopf M (1998). *Chronic fatigue and its syndromes*. Oxford University Press, Oxford.

Whiting P, Bagnall A, Sowden A, Cornell JE, Mulrow C, Ramirez G (2001). Interventions for the treatment and management of chronic fatigue syndrome: a systematic review. *Journal of the American Medical Association*, **286**, 1360–8.

7.20 Infection in the immunocompromised host

J. Cohen

One of the most distressing experiences in medicine is to see a patient cured of a serious underlying disease only to die as a result of a complication of the treatment. While the benefits of immunosuppression have been enormous, there is no doubt that unwanted effects, and in particular serious infection, have proved to be a major drawback.

Classification

The term 'immunocompromised host' has no formal definition, but embraces a group of overlapping conditions in which the ability to respond normally to an infective challenge is in some way impaired. Nevertheless, it is helpful to think of such patients falling into one of four broad groups (Fig. 1).

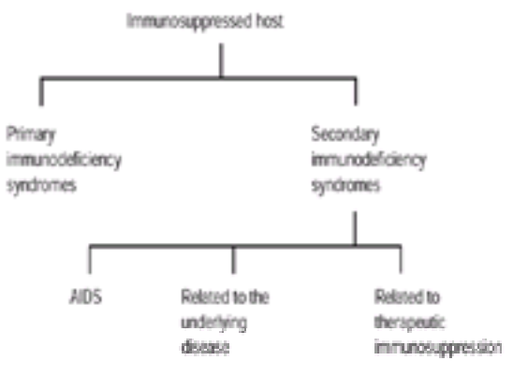

Fig. 1 A classification of the immunocompromised host.

Primary immunodeficiency syndromes

These are patients with congenital defects in immunity that render them more susceptible to infection. At the most extreme, children with severe combined immunodeficiency have virtually no functioning cellular or humoral immunity, and if unprotected will die from infection within a few months of birth. In contrast, some patients with chronic granulomatous disease, an inherited defect of neutrophil function, remain undiagnosed until early adult life. A complete description of the diagnosis and management of this group of disorders is given elsewhere.

Secondary immunodeficiency syndromes

AIDS

AIDS is a model for an acquired defect of immunity leading to an increased risk of infection. Although there are inevitably parallels with other groups of immunocompromised patients, there are particular issues both in the diagnosis and management of infection in AIDS that warrant separate discussion (see Chapter 7.10.21).

Infection related to the underlying condition

The notion of opportunistic infection in the immunocompromised host is most familiar with haematological malignancies or organ transplantation, discussed in detail below. Less obvious, but probably more numerous, are the many physiological conditions and other diseases associated with an increased incidence of infection (Table 1). These immune defects are usually mixed, and frequently poorly characterized, but the clinical problem is real enough. In malnutrition, for example, infection due to mycobacteria and *Salmonella* is more common, and *Pneumocystis carinii* pneumonia was first described in children with protein-calorie malnutrition. There is an extensive literature documenting multiple defects of host defence in association with alcohol abuse; clinically, this is reflected in an excess of lower respiratory tract infections with *Streptococcus pneumoniae*, *M. tuberculosis*, and *Klebsiella pneumoniae.* In Cushing's disease, the excess endogenous steroid production can result in a pattern of opportunistic infections that mirrors that seen in patients receiving steroid therapy (see below). Diabetes mellitus is good example of a disease that is frequently complicated by infection, typically with staphylococcal skin abscesses.

Patients who have had their spleen removed, or who have functional (or more rarely congenital) asplenia, are at increased risk of certain infections. The degree of risk is related to the underlying cause; overall, 4 to 12 per cent will suffer a serious infection, but this varies from 1.5 per cent following traumatic splenectomy to as high as 25 per cent in patients with thalassaemia. Serious infections are most common during the first 5 years following splenectomy, and particularly during the first year; recurrent infections occur in about 20 per cent of those affected. Approximately 50 per cent of infections are meningitis or bacteraemias, and most of the remainder are pneumonias (see Chapter 22.4.4).

In myeloma and chronic lymphocytic leukaemia the primary defect is hypogammaglobulinaemia, manifested clinically by an excess of bacterial infections, typically those caused by encapsulated organisms such as *Streptococcus pneumoniae* and *Haemophilus influenzae*. These patients (and others, such as those with rheumatoid arthritis, systemic lupus erythematosus, or polyarteritis nodosa) all have impaired immunity as a consequence of their underlying disease, but because they also commonly receive treatment with immunosuppressive drugs it can be very difficult to attribute cause and effect.

Infection complicating therapeutic immunosuppression

In addition to the well-recognized risk groups such as those with haematological malignancy or allograft recipients, infective complications of immunosuppression are now being recognized in a much broader range of patients. Conditions as diverse as severe skin disease, asthma, inflammatory bowel disease, and rheumatoid arthritis are routinely treated with drugs such as prednisolone (typically at doses of 5 to 25 mg/day), azathioprine, cyclosporin, and cyclophosphamide. These patients are not so profoundly immunosuppressed as the recipient of a bone marrow transplant, but they are certainly at risk of opportunistic infections. Any failure to recognize the risk may mean that diagnosis is delayed.

Immunosuppressed patients have multiple risk factors; a recipient of a bone marrow transplant may have been neutropenic, receiving corticosteroids and cyclosporin

A for management of graft-versus-host disease, and have an indwelling right atrial catheter for feeding purposes. Clearly each of these factors represents a substantial and very different type of risk factor for infection, and it is important to remember that, in such patients, multiple pathogens can cause disease simultaneously. The following sections describe the management of some of the common clinical syndromes that present as infection in immunosuppressed patients.

Common clinical syndromes

A general approach to management

Infections in immunosuppressed patients can progress with frightening rapidity; the early physical signs are often muted and the microbiology can be confusing. Patients need to be reviewed frequently and will often need empirical therapy, but this need not be totally 'blind'; a structured and informed assessment will generally allow a logical response to what are the most likely pathogens.

The history

This may reveal exposure to community-acquired infections such as varicella zoster or tuberculosis, which can be particularly severe in the immunocompromised patient. Note should be made of any past history of infection; bronchiectasis, for instance, can be very troublesome in transplant recipients. A detailed travel history is important; patients who have visited certain parts of the United States may have been exposed to the systemic mycoses such as histoplasmosis or coccidioidomycosis, which are unfamiliar to many clinicians. Visitors to Central America or the Far East, even many years ago, may have acquired an asymptomatic infection with the helminth *Strongyloides stercoralis*; immunosuppression can lead to overt disease (the hyperinfection syndrome) with a high mortality (see below).

Physical examination

This may be unhelpful; immunosuppressed patients often do not mount a good inflammatory response. Thus there may only be a low-grade fever, a thin serous exudate may suffice for pus, and mild abdominal tenderness be the only sign of peritonitis. Nevertheless, careful, and if necessary repeated, clinical examination is worthwhile, as signs of inflammation may become apparent only when immune function returns. Particular attention should be paid to the presence of new skin lesions. In neutropenic patients bacteraemias may be accompanied by striking embolic lesions; *Pseudomonas* infections (and less commonly *Klebsiella* and *Aeromonas*) can cause a focal necrotic cellulitis called ecthyma gangrenosum. Fungal infections present as indolent, locally invasive lesions; *Aspergillus* infections often have a black eschar. The perianal area and the insertion sites of indwelling right atrial catheters repay careful examination. Aspiration and/or biopsy of any new skin lesion in immunosuppressed patients is well worthwhile, since it may quickly point to an otherwise inapparent diagnosis. Lymphadenopathy is always important and provides the opportunity for aspiration or biopsy. It is more usually seen as a feature of subacute or chronic complication of immunosuppression. It may be due to direct infection (for example tuberculosis), or as a manifestation of a lymphoproliferative condition secondary to Epstein–Barr virus. Post-transplant lymphoproliferative disease is emerging as one of the major causes of late death following renal transplantation.

Underlying disease

This can provide valuable clues. Neutropenia is a major risk factor for infection, conferring susceptibility to bacteraemias, particularly with Gram-negative organisms such as *Escherichia* and *Pseudomonas*. A patient with an obstructing bronchial neoplasm may develop a lung abscess due to inadequate drainage. Corticosteroids are used widely; when given in doses exceeding 15 to 20 mg daily for long periods they increase susceptibility to infections with viruses, fungi, parasites, and bacteria such as *M. tuberculosis*—all organisms normally associated with cellular immune defences.

Duration of immunosuppression

This often has a profound effect on the type of infection that occurs. This effect is well illustrated by comparing the 'timetables' of infections in renal transplant recipients with these receiving bone marrow transplants (Fig. 2). In the first 6 weeks after a renal transplant bacterial infections predominate, typically complicating the surgery or urinary infections. Between 6 weeks and 6 months post-transplant the patient is most at risk from the 'classical' opportunistic infections; as time continues and the intensity of immunosuppression declines, typical community-acquired infections become more common. In bone marrow transplantation, the initial period of neutropenia is characterized by bacterial infections; later, when many patients receive high-dose steroids for graft-versus-host disease, cytomegalovirus and fungal infections (*Candida* and *Aspergillus*) develop.

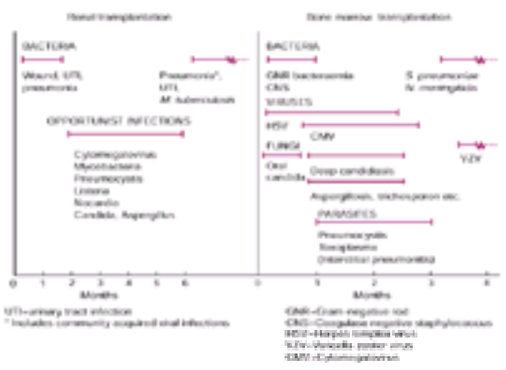

Fig. 2 A scheme representing a timetable for the development of infective complications in renal transplant and bone marrow transplant recipients.

Speed of progression

An assessment of this is helpful in both differential diagnosis and in deciding on empirical therapy. In neutropenic patients, the onset of fever is usually an indication for immediate empirical antibiotic therapy (see below). In contrast, the response to a fever and new pulmonary infiltrates in a patient 8 months after renal transplant will depend on the pace of the illness. Rapid deterioration over the space of a few hours will suggest a bacterial infection or a non-infectious cause, and will need urgent therapy; a more indolent presentation would point to a fungal or mycobacterial aetiology, and treatment can be delayed for a short period to try and establish the diagnosis.

Investigations

It is important that the diagnostic laboratories be made aware of the clinical problem, since handling of specimens from immunosuppressed patients—and interpretation of the results—will often differ substantially from routine procedures. In addition, routine screening of transplant recipients and donors should include serological tests for cytomegalovirus, hepatitis B, and HIV.

Pyrexia of unknown origin

In neutropenic patients, fever is often the first and only sign of bacteraemia, and prompt action is necessary. In this situation, a pyrexia of unknown origin is defined as a fever of over 38 °C sustained for 2 h, and not obviously due to an identifiable cause such as concomitant blood transfusion.

The risk of bacteraemia is directly related to the severity of neutropenia; the incidence is greatest at below 1×10^{9}/litre, and particularly when the count falls to less than 0.1×10^{9}/litre (Fig. 3). Some years ago, the commonest isolates were Gram-negative bacteria such as *E. coli* and *Klebsiella*, generally derived from the patient's gut flora, and *Ps. aeruginosa*, a common environmental pathogen, especially in hospitals. Gram-negative bacteraemia in neutropenic patients carried a very high mortality, and led to the introduction of a number of preventive strategies such as the use of prophylactic antibiotics and colony stimulating factors. Although these approaches have not been entirely successful, the incidence of Gram-negative bacteraemias has declined substantially, and in most units Gram-positive organisms,

notably coagulase negative staphylococci (*Staphylococcus epidermidis*) and viridans streptococci, are now the commonest isolates.

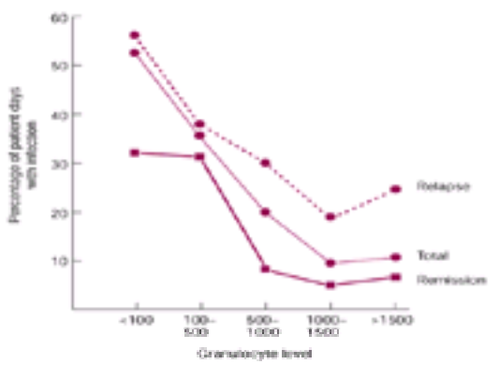

Fig. 3 The relationship between neutrophil count and the risk of invasive Gram-negative infection. (From Bodey G *et al*. (1966). *Annals of Internal Medicine* **64**, 328–40, with permission.)

The clinical features are entirely non-specific. Occasionally a focus will be suggested by erythema around the point of entry of an indwelling catheter, a finding often associated with staphylococcal infection. Septic shock is infrequent, although it can be associated with viridans streptococci; endocarditis is rare.

Blood cultures should be drawn before treatment is begun. Ideally two sets should be obtained, at least one of which should be from a peripheral vein (rather than an indwelling catheter), although this is not always possible. Culturing larger volumes of blood (for example 30 ml rather than the more conventional 10 ml) increases the yield. Appropriate samples must also be taken from other potential foci of infection. Nevertheless, it has been one of the enduring frustrations of this subject that even the most rigorous of microbiological investigations in the febrile neutropenic patient will yield only some 40 to 50 per cent of positive cultures. The explanation for this is unknown; some studies have suggested that it is due to endotoxaemia in the absence of bacteraemia, but the data are inconclusive. What is clear, however, is that treatment must begin before the results of the cultures are available; delay can kill.

The choice of the initial empirical antibiotic regimen for the febrile neutropenic patient has been the subject of intense investigation. The ideal regimen will be safe and have good bactericidal activity against all the common pathogens. No single regimen is perfect; much will depend on the availability (and cost) of antibiotics in a given institution, and on local patterns of antibiotic susceptibility. Well-validated regimens include the combination of an antipseudomonal penicillin plus an aminoglycoside (for example piperacillin plus gentamicin), or the use of single agents such as a third-generation cephalosporin (for example ceftazidime) or a penem such as meropenem. All these regimens are very active against the common Gram-negative organisms, but are relatively ineffective at treating Gram-positive bacteria such as coagulase negative staphylococci, nowadays a common problem. Unfortunately, the only drugs which are reliably active for these organisms are glycopeptides such as vancomycin, and some clinicians have advocated adding vancomycin to the initial empirical regimen. The disadvantage of this approach is the toxicity (and cost) of vancomycin, which may not be justified since unlike the Gram-negative infections, coagulase negative staphylococci rarely cause death. Several prospective clinical trials have concluded that vancomycin can usually be withheld until the results of blood cultures are known. An alternative strategy under investigation is to use a fourth-generation cephalosporin, such as cefepime, that has an improved spectrum of activity against Gram-positive bacteria while preserving good cover against Enterobacteriaceae and *Pseudomonas aeruginosa*.

In patients who respond to the initial regimen the treatment should be continued for at least 7 days, and ideally until the neutrophil count has returned to more than 0.5 × 10^9/litre. Sometimes this is not possible; the patient may have a persistent or unresponsive neutropenia (for example in aplastic anaemia or following bone marrow transplant). In these cases treatment is usually stopped cautiously after an arbitrary period such as 14 days; rebound bacteraemias will need further treatment.

A common problem is the patient who continues to have high swinging fevers in the absence of any obvious focus or positive microbiology. In this situation, deep fungal infection becomes more likely. The few clinical trials which have addressed this problem have concluded that persistent fever for 72 h should be treated by the addition of amphotericin B. For patients who are intolerant of amphotericin B or who develop nephrotoxicity, one of the liposomal formulations of the drug can be used; toxicity is greatly diminished and they are of at least equal efficacy, although they are extremely expensive.

Fever of unknown origin in the non-neutropenic immunosuppressed patient is a completely different problem. Fever in this situation is rarely immediately life-threatening, and since there is a wide differential diagnosis it is better to pursue the cause rather than embark on empirical therapy.

Fever and new pulmonary infiltrates

The development of fever and new pulmonary infiltrates is one of the most challenging clinical problems in this group of patients. Pneumonia is the commonest infective cause of death in immunocompromised patients. In the presence of diffuse airspace disease the mortality approaches 50 per cent irrespective of the underlying defect in host defence, although the epidemiology varies between different patient groups and with the intensity of the immunosuppression (Table 2).

The condition can progress extremely quickly, and conventional diagnostic procedures may be unhelpful. The list of possible causes is so daunting (Table 3) that clinicians can be tempted to use multiple empirical antimicrobial agents, sometimes to the patient's detriment. It is often not possible to 'guess' with any certainty the precise cause of the problem and multiple causes may be present. But considering the available information it may be possible to construct a 'short-list' to guide further investigation and treatment.

The initial evaluation should follow the approach outlined above, in particular making an assessment of the intensity of the immunosuppression and the speed of progression of the pulmonary disease. The main purpose is to determine the need for empirical therapy, either because the clinical picture is suggestive of a 'simple' bacterial pneumonia, or because of a potentially more serious, progressive cause of uncertain aetiology. Factors favouring bacterial aetiology include neutropenia, a rapidly developing clinical evolution (for example deterioration over a period of 12 h), progressive hypoxia, a sputum Gram stain showing a marked predominance of a single bacterial morphology (even in the absence of neutrophils), or a chest radiographic appearance that has worsened over a short period. High fever is not necessarily a part of this syndrome; it is important to emphasize that this rapidly evolving clinical picture is not inevitably due to infection. Non-infective causes such as acute lung haemorrhage or pulmonary oedema can present in an identical fashion, and the most appropriate therapy may be diuretics rather than antimicrobials. However, antimicrobials will often need to be given as well because of what has been termed 'infection-provoked relapse'. In immunologically mediated diseases such as systemic lupus erythematosus or antiglomerular basement membrane disease (Goodpasture's syndrome) infection can precipitate a relapse of the underlying disease. Thus, the development of fever and new pulmonary shadows in a patient with antiglomerular basement membrane disease may be primarily due to lung haemorrhage associated with a rise in antiglomerular basement membrane antibodies, but this in turn can be precipitated by an infection that need not necessarily be in the lung. Treatment must be directed both towards improving oxygenation and the underlying infection.

Blood cultures should always be obtained, and attempts made to obtain sputum. A chest radiograph and arterial blood gas analysis are essential. The initial treatment will be dictated by the clinical circumstances, but it is best to avoid a complex regimen to provide very broad spectrum cover. Rapid clinical deterioration is usually attributable to bacterial infections; if community acquired, a combination of an extended-spectrum cephalosporin plus erythromycin is appropriate. For hospital-acquired infections, a cephalosporin (combined with an aminoglycoside if there is strong evidence of *Pseudomonas* infection) is reasonable. Where staphylococcal infection is suspected, flucloxacillin plus an aminoglycoside should be used. Unusual ('opportunist') organisms such as mycobacteria, *Nocardia*, or cytomegalovirus rarely cause such a rapid clinical deterioration and are difficult to distinguish on clinical grounds alone. For these reasons, the addition of further empirical agents is usually not warranted.

In patients in whom immediate empirical therapy is not necessary, additional diagnostic procedures can be performed. These should include serological tests for atypical organisms, and examination of blood and urine for cytomegalovirus (tests for cytomegalovirus early antigen are very helpful). The chest radiograph should be repeated, but it is not as sensitive as arterial blood gas measurements, which should be done twice daily. The radiographic appearances are rarely sufficiently specific to suggest a precise diagnosis, although they can be suggestive. Thus a bilateral interstitial midzone infiltrate associated with marked hypoxia is typical of *Pneumocystis* pneumonia, and a pleural based infarct is suggestive of *Aspergillus*. However, there are pitfalls in relying on the radiographic appearance alone in

guiding the choice of therapy. First, no radiographic appearance is pathognomonic of any single pathological process; cytomegalovirus or pulmonary oedema can mimic *Pneumocystis* pneumonia, for example, and *Legionella* pneumonia cannot be distinguished from *Aspergillus*. Secondly, multiple agents can coexist, each requiring separate treatment. Other imaging techniques, such as computed tomography (**CT**), can often provide useful additional information on the extent of the process, and will sometimes point to the cause (for example the 'halo sign' associated with invasive aspergillosis).

It is often appropriate to try and make a specific diagnosis by obtaining material directly from the bronchial tree. In most cases the method of choice is bronchoscopy with bronchoalveolar lavage. This will provide adequate material without incurring a serious risk of bleeding (many such patients are thrombocytopenic) or pneumothorax. In most series, bronchial brush or transbronchial biopsy specimens produce only a marginal increase in the diagnostic yield, and are usually not done unless the clinical picture is suggestive of a non-infective process such as an infiltrating tumour. The highest yield is from open lung biopsy, but this should not usually be done as a first-line procedure.

Acute neurological syndromes

A large number of conventional and opportunist pathogens can lead to neurological infection in immunocompromised patients. Although there is some degree of overlap, the underlying defect in host defence is a often a good indicator of the likely cause (Table 4).

The clinical features may help suggest the diagnosis. Meningitic syndromes are more likely to be associated with conventional bacterial infections, listeriosis, and tuberculosis, as well as fungi such as *Cryptococcus* and *Candida*. In contrast, infections with *Toxoplasma*, *Aspergillus*, or *Nocardia* more commonly present as space-occupying lesions. Pure encephalitic syndromes are less common, but can occur with herpes simplex. Rhinocerebral mucormycosis is a progressive, destructive infection caused by *Mucor* and related moulds, that usually begins in the paranasal sinuses and spreads caudally to involve the orbits or the frontal lobes of the brain. It is seen particularly in patients with uncontrolled diabetes mellitus or as a complication of neutropenia.

Bacterial infections generally proceed rapidly, while fungi and parasites pursue a more indolent course. However, exceptions to this are common, and there is no substitute for obtaining a precise diagnosis. Examination of the skin (see below) and fundoscopy may be valuable. Retinitis is not usually a feature of systemic infection with toxoplasma or cytomegalovirus; in contrast, *Candida* endophthalmitis may be the only manifestation of deep-seated infection.

Examination of the cerebrospinal fluid is mandatory. A high index of suspicion is necessary, since the clinical features of meningitis are often muted in these patients. An unexplained low-grade fever and mild headache may be the only clues; frank meningism, photophobia, or focal neurological signs occur late. Examination of the cerebrospinal fluid should include direct microscopy (neutrophil, lymphocyte, or eosinophil pleocytosis) and culture for (myco)bacteria and fungi, a cryptococcal latex agglutination test, antigen tests for pneumococcus, and a search for specific antibody production (for example for coccidioidomycosis) or DNA sequences by the polymerase chain reaction (for example for herpes simplex and papovaviruses).

Some organisms are rarely seen by direct microscopy: Mycobacteria are seen in fewer than 10 per cent of cases, *Nocardia* and *Aspergillus* only very rarely. A predominance of lymphocytes suggests partially treated bacterial infection, tuberculosis, or a viral aetiology but not infection with *Listeria*, despite its name. A low cerebrospinal fluid glucose points to tuberculosis but is not specific. Sometimes the only abnormality is a modest elevation of the cerebrospinal fluid protein; this should never be ignored, even in the seeming absence of other features of neurological infection. Where appropriate, cytological examination of the cerebrospinal fluid should be done to exclude carcinomatous or leukaemic meningitis, which can mimic an acute infective presentation.

Certain neurological infections are often associated with pulmonary disease; these include *Legionella*, tuberculosis, *Aspergillus*, *Mucor*, and *Nocardia*. A brain CT scan, which should be contrast-enhanced, is valuable. Focal, usually enhancing, lesions are particularly associated with pyogenic abscesses and toxoplasmosis. Tuberculomas can appear as single lesions. Magnetic resonance imaging is better then CT scanning for detecting abnormalities of the brainstem (for example the basal meningitis associated with cryptococcal infection), and frequently reveals lesions in toxoplasmosis which are not seen on CT scans. It may pre-empt the need for brain biopsy when a diagnosis of progressive multifocal leucoencephalitis is considered.

Any new skin lesions should be biopsied. Nasal biopsy may reveal *Mucor*. An electroencephalogram is not helpful, unless herpes encephalitis is suspected. Brain biopsy is done rarely; it should not be considered unless empirical therapy has failed, or there is a real prospect of therapeutic benefit to the patient.

If the cerebrospinal fluid is non-diagnostic but bacterial infection cannot be excluded, empirical antibiotics should be given immediately. An extended spectrum cephalosporin such as cefotaxime is suitable. Serological tests for toxoplasmosis are not specific in this setting, and if the infection is suspected it is better to start empirical therapy with pyrimethamine and sulphadiazine. Cerebral aspergillosis and mucormycosis have a very poor prognosis; treatment should be begun with high-dose amphotericin B, and surgical debridement considered if possible. There is no effective treatment for progressive multifocal leucoencephalitis.

Acute gastrointestinal syndromes

The organisms associated with specific gastrointestinal syndromes in these patients are shown in Table 5.

Severe stomatitis is a common complaint in immunosuppressed patients. The three commonest causes: *Candida*, herpes simplex, and chemotherapy-induced mucositis are clinically indistinguishable and indeed can coexist and cause disease together. For these reasons, the diagnosis should always be confirmed by microscopy and culture. Herpetic stomatitis in particular can be atypical in these patients; the classical appearance of groups of small vesicles is unusual, and a more common presentation is ulceration, which can be extensive. In profoundly immunosuppressed patients such as bone marrow transplant recipients, oral candidiasis is very common, and in patients who are seropositive before transplant, reactivation of herpes simplex is almost universal. For these reasons, prophylaxis is usually given. Both herpes simplex and *Candida* can cause oesophagitis, generally (but not exclusively) as an extension of oral disease. Oesophagoscopy with brush cytology and/or biopsy is the investigation of choice. Proven oesophageal candidiasis should be regarded as 'invasive' disease and treated with systemic antifungals (amphotericin B or fluconazole).

A large number of organisms can cause acute diarrhoeal syndromes; in addition, non-infective conditions such as radiation enteritis, drugs, and graft-versus-host disease must be included in the differential diagnosis. There are no distinguishing clinical features of note, and diagnosis depends on microbiological examination of the faeces.

The diarrhoea caused by *Clostridium difficile* is usually due to a pseudomembranous colitis. However, patients with leukaemia or aplastic anaemia may develop neutropenic enterocolitis (previously called typhlitis), a fulminating invasive colitis characterized by diffuse dilation and oedema of the bowel walls, haemorrhage, ulceration, and a high mortality. Classically this has been associated with clostridial bacteraemia, in particular *C. septicum*, but other clostridia, including *C. difficile*, and even Gram-negative bacteria, can also be found.

Strongyloides stercoralis is a nematode that can be carried asymptomatically for many years after exposure (see Chapter 7.14.4). Strongyloidiasis has been recognized as a complication of HTLV-I infection, and also occurs secondary to immunosuppression (typically with high-dose steroids and in recipients of solid organ transplants). A rise in the worm burden results in the hyperinfection syndrome, which may present as pneumonitis or intermittent intestinal obstruction. Worms moving through the gut wall can carry with them enteric bacteria, resulting in polymicrobial bacteraemias and Gram-negative meningitis when the worms invade the cerebrospinal fluid.

Giardiasis is particularly associated with hypogammaglobulinaemia, and curiously is rarely seen in other groups. *Cryptosporidium*, *Microsporidia*, and *Isospora* are now well recognized causes of severe and sometimes chronic diarrhoea in AIDS patients, but may also occur in other less severely immunocompromised patients. Among the viruses, the most difficult problem is cytomegalovirus. Cytomegalovirus can cause a severe colitis, and in these cases ganciclovir is beneficial. The diagnosis should be confirmed by biopsy, but ultimately may depend on the result of a therapeutic trial since demonstration of the organism does not necessarily indicate that it is causing disease.

Mild abnormalities of liver function tests are a common accompaniment of many systemic infections, but hepatitis is a particular feature of both toxoplasmosis and cytomegalovirus infection. An increased prevalence of hepatitis B has been found in patients on chronic haemodialysis (10 per cent), and those with Hodgkin's disease (8 per cent) and lepromatous leprosy (20 per cent). The acute hepatitic episode is mild, often anicteric, and may pass unnoticed. However, persistent viral replication (hepatitis e antigenaemia) and the development of complications associated with chronic infection are more likely. Although it is likely that infection with the

other hepatitis viruses occurs in immunosuppressed patients there are as yet no detailed clinical or epidemiological data that define the problem.

A particular form of systemic candidiasis has been called chronic hepatosplenic candidiasis (although other organs can be involved, and the syndrome is better referred to as chronic systemic candidiasis). The patient presents with unremitting fever and occasionally abdominal pain; palpable hepatomegaly is unusual. Typically the neutrophil count has returned to normal after a recent course of chemotherapy for acute leukaemia; the liver function tests show a markedly raised alkaline phosphatase and there may be hyperbilirubinaemia, but microbiological investigations (including blood cultures) are negative. The diagnosis is made by ultrasonography or CT scan of the abdomen, which reveals multiple intrahepatic (or less commonly splenic) abscesses. Unfortunately biopsy of these lesions reveals histological or culture evidence of *Candida* in only about a third of cases. Treatment has been difficult; conventional therapy with amphotericin B (even with the addition of flucytosine) has often failed, but the use of a liposomal formulation of amphotericin B and fluconazole has produced encouraging results.

Further reading

Fishman JA, Rubin RH (1998). Infection in organ-transplant recipients. *New England Journal of Medicine* **338**, 1741–51.

Klastersky J (1995). *Infectious complications of cancer*. Kluwer Academic, Boston.

Meunier F (1995). *Invasive fungal infections in cancer patients. Baillière's clinical infectious diseases*. Baillière Tindall, London.

Pizzo PA (1999). Fever in immunocompromised patients. *New England Journal of Medicine* **341**, 893–900.

Rosenow EC, Wilson WR, Cockerill FR (1985). Pulmonary disease in the immunocompromised host I. *Mayo Clinic Proceedings* **60**, 473–87.

Rubin RH, Young LS (1994). *Clinical approach to infection in the compromised host*, 3rd edn. Plenum, New York.

Warnock DW, Richardson MD (1991). *Fungal infection in the compromised patient*, 2nd edn. Wiley, Chichester.

Wilson WR, Cockerill FR, Rosenow EC (1985). Pulmonary disease in the immunocompromised host II. *Mayo Clinic Proceedings* **60**, 610–31.

8.1 Poisoning by drugs and chemicals

A. T. Proudfoot and J. A. Vale

Introduction

In the minds of most people, not least those of doctors, the term poisoning suggests an acute event demanding immediate care and attention. This is often so, but poisoning may take other forms. The consequence is not always immediate even after a single dose—so-called acute poisoning. Prolonged uptake may result in accumulation, as with many heavy metals, and the damage may arise only after prolonged exposure—that is, chronic poisoning.

Exposure by oral, inhalational, cutaneous, or other routes, should not itself be equated with poisoning. Uptake is necessary for there to be a toxic effect, and even if this occurs, poisoning does not necessarily result, as the amount absorbed may be too small.

If true poisoning does occur, the ensuing clinical syndrome may be distinctive. For example, fixed dilated pupils, exaggerated tendon reflexes, extensor plantar responses, depressed respiration, and cardiac tachyarrhythmias suggest tricyclic antidepressant poisoning. Anaemia, constipation, colic, and motor nerve palsies are indicative of lead poisoning. However, with a whole range of psychotropic medications there may be only non-specific central nervous depression, respiratory impairment, and hypotension. In some instances, distinctive sequelae may not appear until many years have elapsed as, for example, with carcinoma of the

oesophagus following ingestion of corrosives or hepatic haemangiosarcoma from vinyl chloride exposure.

Poisoning may be accidental or deliberate. It is usually accidental in small children, but in adults it is invariably deliberate with parasuicidal, suicidal, or rarely, homicidal intent.

Thus, the medical approach to poisoning should never be confined to the poison and its effects. All the circumstances surrounding the episode must be taken into account, especially in cases where litigation may follow, for example in the event of an occupational mishap with a chemical. It is therefore important that the doctor concerned, having instituted any necessary life-saving measures, should take a careful history, retain all pertinent evidence such as suicide notes and biological specimens, and make a meticulous record of symptoms, signs, progress, and outcome.

Epidemiology

Few health-care professionals would deny that poisoning, accidental or deliberate, is a common problem in most countries throughout the world. Yet it is remarkably difficult to obtain reliable statistics on the morbidity or mortality it causes, even in countries with comparatively advanced systems for collection of population health data. In the developing world, about 600 000 deaths/year are attributed to deliberate self-harm, the majority from poisoning with pesticides such as organophosphorus insecticides. The following observations are based primarily on statistics from the United Kingdom and the United States but, wherever appropriate, observed variations in patterns of poisoning in other countries are noted.

Hospital admissions due to poisoning

Poisoning from accidental or deliberate ingestion or inhalation of drugs or chemicals is a common acute medical emergency requiring hospital admission. In the period 1957 to 1976, the annual number of hospital admissions due to acute poisoning in England and Wales rose steadily from less than 20 000 to more than 125 000. Since then, there has been a decline in the incidence of self-poisoning in England to approximately 100 000 admissions each year. Despite this decline, self-poisoning still accounts for more than 10 per cent of acute adult medical admissions in the United Kingdom. However, the true incidence of self poisoning may be as much as three times that of the hospital admission rate.

In the United Kingdom, with the exception of young children, females predominate in all age groups in those admitted to hospital because of acute poisoning and there is a marked preponderance in those aged 15 to 44 years. Many paediatric episodes are poisoning scares, rather than true poisonings, though this often only becomes clear in retrospect. The majority of adults who poison themselves are not suicidal.

In Western Europe and North America, drugs have always been the most common agents taken by adults and rank second only to household products as the substances most often ingested by children. In the United Kingdom, alcohol is taken in addition to the drug overdose in 60 per cent of males and 40 per cent of females, and at least one-third of self-poisoning episodes involve one or more drugs. Approximately two-thirds of adults ingest drugs that have been bought in retail outlets or prescribed for themselves or a close relative. Therefore, the pattern of pharmaceutical agents used for self-poisoning reflects prescribing habits (particularly for illnesses occurring in those aged 15 to 44 years) and common symptoms that are self-treated. Barbiturate and non-barbiturate hypnotics are now seldom encountered causes of poisoning, while use of other psychotropic agents such as the benzodiazepines, tricyclic antidepressants, and selective serotonin-specific reuptake inhibitors in overdose is now more frequent; analgesic poisoning also occurs much more commonly than previously.

Within Europe there are variations from country to country. In Finland, for example, alcohol, cardiovascular drugs, and psychotropics are the most common causes of poisoning. Outside Europe, and in developing countries in particular, the situation is often very different. In Sri Lanka, for example, agrochemicals account for nearly 60 per cent of all poisonings; such agents account for less than 1 per cent of hospital admissions for poisoning in England and Wales. In South Africa, the pattern of poisoning in the white population mirrors that in North America and Western Europe, whereas that observed in black South Africans is very different, with kerosene (paraffin) and traditional medicines accounting for the majority of hospital admissions (and deaths) attributable to poisoning. In countries where malaria is prevalent, poisoning by antimalarials is an additional cause of morbidity and mortality.

Deaths from poisoning

The number of deaths from poisoning is to some extent determined by the lethality of the agents involved. This, in turn, results in regional differences with rates that are often much higher in developing countries.

In contrast to the rise in the number of hospital admissions for poisoning in England and Wales, deaths from acute poisoning have decreased over the last 40 years and, since 1972, they have remained virtually constant at 4000 per annum. The lack of change over the last 30 years is particularly striking because this period follows the substitution of 'natural gas' for 'town gas', which led directly to a fall in carbon monoxide deaths from nearly 4000 in 1963 to just over 1000 per annum 10 years later.

Despite relatively little change in the overall mortality statistics for acute poisoning in England and Wales, there have been very substantial changes in the agents responsible. Deaths from barbiturate and non-barbiturate hypnotics have fallen, while those due to analgesics and psychotropic agents have risen. Deaths from carbon monoxide have shown a slow but steady increase since 1975, reaching just over 1500 in 1991. Increasing numbers of young men (14 to 24 years) in the United Kingdom are killing themselves. Self-poisoning with car exhaust fumes (containing carbon monoxide) is currently the most common means of doing so.

In England and Wales, about 80 per cent of individuals who die as a result of poisoning do so at home, the inpatient mortality being less than 1 per cent of all cases admitted to hospital. The age and sex distribution of deaths attributed to acute poisoning is significantly different from that for admissions—there are fewer patients in the age range 15 to 44 years and more in the older age categories where males predominate. A similar pattern of causes of death has been observed elsewhere in Western Europe and in North America, but in many developing agricultural countries, agrochemicals (cholinesterase inhibitors, paraquat, aluminium phosphide, and other pesticides) more commonly predominate. On a global scale, it has been estimated that pesticides account annually for 1 million serious unintentional poisonings and 2 million people admitted to hospital for suicide attempts, predominantly in developing countries.

Childhood poisoning

It is estimated that in the United Kingdom as many as 41 000 poison exposures occur in children aged 4 years or less each year. In the United States, the Toxic Exposure Surveillance System (TESS) of the American Association of Poison Control Centers for 2000 records 1 142 796 poison exposures in children less than 6 years of age.

Children aged less than 5 years are particularly active and exploratory and have a strong impulse to put things into their mouths. These characteristics predispose to accidental poisoning, which in Western Europe is particularly likely to occur when parents are inattentive or neglectful, as at times of family crises. The vast majority (80 to 85 per cent) of cases occur in the child's own home, and in many instances the substances involved are out of their usual storage place or have been put into some other container; grandparents, for example, may find it convenient to remove drugs from child-resistant closures or leave the caps off containers because they themselves have difficulty opening this type of packaging.

A child may also be poisoned by an adult who administers a toxic substance by mistake, and rarely a parent or carer may poison a child as a form of abuse, sometimes with fatal consequences. In addition, older (typically 10 to 16 years) emotionally disturbed children may deliberately poison themselves. Abuse of volatile substances is a continuing problem in adolescents.

Further reading

Casey P, Vale JA (1994). Deaths from pesticide poisoning in England and Wales: 1945–1989. *Human and Experimental Toxicology* **13**, 95–101.

Eddleston M (2000). Patterns and problems of deliberate self-poisoning in the developing world. *Quarterly Journal of Medicine* **93**, 715–31.

Fingerhut LA, Cox CS (1998). Poisoning mortality 1985–1995. *Public Health Report* **113**, 218–33.

Hawton K, Fagg J, Simkin S (1996). Deliberate self-poisoning and self-injury in children and adolescents under 16 years of age in Oxford, 1976–1993. *British Journal of Psychiatry* **169**, 202–8.

Hoppe-Roberts JM, Lloyd LM, Chyka PA (2000). Poisoning mortality in the United States: comparison of national mortality statistics and poison control center reports. *Annals of Emergency Medicine* **35**, 440–8.

Kasilo OMJ, Nhachi CFB (1992). A pattern of acute poisoning in children in urban Zimbabwe: ten years experience. *Human and Experimental Toxicology* **11**, 335–40.

Litovitz TL *et al.* (2001). 2000 Annual Report of the American Association of Poison Control Centers Toxic Exposure Surveillance System. *American Journal of Emergency Medicine* **19**, 337–95.

Owens D *et al.* (1994). Outcome of deliberate self-poisoning. An examination of risk factors for repetition. *British Journal of Psychiatry* **165**, 797–801.

Pickett W *et al.* (1998). Suicide mortality and pesticide use among Canadian farmers. *American Journal of Industrial Medicine* **34**, 364–72.

Shepherd G, Klein-Schwartz W (1998). Accidental and suicidal adolescent poisoning deaths in the United States, 1979–1994. *Archives of Pediatric and Adolescent Medicine* **152**, 1181–5.

Tay SY *et al.* (1998). Patients admitted to an intensive care unit for poisoning. *Annals of the Academy of Medicine of Singapore* **27**, 347–52.

Woolf AD, Lovejoy FH Jr (1993). Epidemiology of drug overdose in children. *Drug Safety* **9**, 291–308.

Diagnosis

Ideally, the diagnosis of acute poisoning requires that the doctor establish the chemical composition of the poison, the magnitude of the exposure, and the route of exposure (whether by ingestion, injection, inhalation, or skin contamination), so that the features likely to develop can be anticipated and the risk assessed. As in any other branch of medicine, diagnosis of acute poisoning is based on the patient's history and on a combination of circumstantial evidence, the findings on physical examination, and appropriate investigations when a history is not available. However, in acute poisoning, there are many obstacles to establishing the information required. Young children may not be able to give a history and adults are often unreliable while physical signs are rarely diagnostic. Similarly, circumstantial evidence may not be available, be only tentative or misleading, and laboratory diagnosis can never be fully comprehensive.

History

Since accidental poison exposure in childhood is most common between the ages of 9 months and 5 years, an unequivocal history is unlikely to be forthcoming from the victim but may be obtainable from older witnesses. Clearly, however, statements about amounts must be interpreted with caution since knowledge of the quantities in original containers is frequently inaccurate or unknown.

In contrast, 90 per cent or more of adults presenting with acute poisoning are conscious or only slightly drowsy and there would seem little reason why diagnosis of self-poisoning on the basis of the history should be difficult. Indeed, while a small number of patients adamantly deny having taken a poison, the majority usually admit to it without hesitation, although problems often arise in trying to establish precisely the nature and quantity of what has been taken. Comparison of patients' statements with the agents detected by laboratory analysis of blood or urine consistently reveals major differences in about half the cases. In consequence, patients are often thought to be deliberately untruthful. However, surprise at these findings may merely reveal a lack of medical insight into the circumstances under which self-poisoning occurs. It is commonly an impulsive act; the patient ingests the contents of the first bottle that comes to hand, often when under the influence of alcohol. Moreover, although about 60 per cent of episodes involve drugs prescribed for the victims or their relatives, like many other patients, they are often ignorant of their names.

If these considerations make it difficult to establish the nature of the poison, it is hardly surprising that they should make the amounts involved even more suspect. Few patients count the number of tablets they consume and it is impossible for patient or doctor to know what constitutes a 'handful', 'bottleful', or similar arbitrary quantity.

Circumstantial evidence

Circumstantial evidence becomes important in the diagnosis of acute poisoning when patients are either unable to give a history (for example young children, adults who have severe learning difficulties or who are demented, and unconscious patients) or are unwilling to do so. However, although circumstantial evidence may strongly suggest poisoning, it is seldom incontrovertible. It takes several forms.

Circumstances under which found

The mother may return to the kitchen or bathroom to find her child with some substance all over his hands, face, and clothing, or surrounded by pills, one of which he is eating. The assumption that more has been ingested may or may not be correct and the amount swallowed is a matter of speculation. Similarly, adults may be found unconscious with tablet particles around the mouth or on clothing as the only clue to diagnosis. More often, the presence of empty drug containers with occasional tablets or capsules nearby suggests the diagnosis. Less commonly, they are found unconscious or dead in some remote location. The lack of personal effects to indicate who they are or where they live may suggest a desire not to be identified and should arouse suspicion of drug overdosage. Self-poisoning is a common cause of coma in previously healthy young adults. Protestations by relatives that patients would never take overdoses are usually wrong.

Suicide notes

Suicide notes are reliable indicators of drug overdose in the absence of physical violence as a cause of coma. The note may specify what has been taken in addition to expressing despair, futility, worthlessness, and remorse.

Features

There are few symptoms or physical signs that cannot be attributed to one poison or another. However, a clinical feature rarely arises in isolation and clusters of features are of much greater diagnostic value. Those most commonly encountered in present-day practice are given in Table 1.

Conscious patients with abnormal behaviour, who may be experiencing auditory and visual hallucinations, may have ingested amphetamines, phencyclidine, LSD (lysergic acid diethylamide), 'magic' (psilocybin-containing) mushrooms, and drugs such as the older antihistamines and tricyclic antidepressants that have marked anticholinergic actions. Occasionally a patient with severe salicylate intoxication, who cannot give a history despite being conscious, is hyperventilating, sweating, flushed, and tachycardic, suggesting a diagnosis which can then be confirmed analytically.

Drowsiness, ataxia, dysarthria, and nystagmus are common after ingestion of benzodiazepines. Coma with hypotonia and hyporeflexia may follow, particularly if alcohol has also been taken. Hypotension, hypothermia, and respiratory depression are rare. All of these features, however, may occur after overdosage with outmoded drugs such as barbiturates, methaqualone, meprobamate, and ethchlorvynol that are still occasionally prescribed. In present-day clinical practice, tricyclic antidepressants remain among the most common central nervous system (**CNS**) depressants encountered in overdose. They cause hypertonia, hyperreflexia, extensor plantar responses, and dilated pupils. Sinus tachycardia and prolongation of the QRS interval on the electrocardiogram support a diagnosis of intoxication with these drugs. Hypotension and hypothermia are less common features. Tricyclic antidepressants and non-steroidal anti-inflammatory agents, particularly mefenamic acid, are the most common causes of seizures after drug overdosage. Coma with pinpoint pupils and a reduced respiratory rate is virtually diagnostic of overdosage with opioid analgesics and is an indication for a therapeutic trial of naloxone. Many patients with opioid poisoning will be habitual drug abusers and have venepuncture marks and evidence of venous tracking in the antecubital fossae. Alcohol may be smelt on the breath, as may solvents such as toluene, acetone, or xylene as the result of 'sniffing' glues, cleaning agents, or other preparations. Skin blisters occur in poisoning by many drugs (see below) but rarely in coma due to other causes. Burns around the lips or in the buccal cavity or pharynx indicate ingestion of corrosives, including paraquat.

Lateralizing neurological signs

Since most serious poisonings are associated with impairment of consciousness, neurological signs are particularly important. Lateralizing signs (unless they are attributable to a known neurological disease) virtually exclude a diagnosis of acute poisoning. Such findings have been recorded with barbiturate and phenytoin overdose but so rarely that the general rule is not significantly compromised. A possible exception is transient inequality of pupil size. This has been reported only rarely in acute poisoning but is not an uncommon finding in normal individuals (for instance due to Holmes–Adie pupils); clinical experience suggests that it occurs more frequently in poisoning than seems apparent from the literature.

Decerebrate and decorticate movements

Unconscious poisoned patients may respond to painful stimuli with flexor and extensor limb movements of the type seen in decorticate and decerebrate states. However, in poisoning, these signs do not indicate irreversible brain damage and patients showing them can be expected to recover fully. Hypoglycaemia must be excluded in these cases.

Strabismus, and internuclear and external ophthalmoplegia

A variety of ocular signs including strabismus, internuclear ophthalmoplegia, and total external opthalmoplegia, may be found in acutely poisoned patients. They are also features of Wernicke's encephalopathy in chronic alcohol abusers.

Strabismus has been described in poisoning with phenytoin, carbamazepine, and tricyclic antidepressants. Usually the optic axes diverge in the horizontal plane but in some patients there is additional vertical deviation. It is present transiently and only in patients who are unconscious. Dysconjugate, roving eye movements may also be seen if both eyes are observed for a period of time. It is important to know that such abnormalities occur so that they are not misattributed to intracranial vascular lesions or some other pathology requiring surgical intervention.

Dysconjugate eye movements may become apparent only when vestibulo-ocular reflexes are examined by caloric stimuli. Installation of ice-cold water into the external auditory meatus should make both eyes turn to the side irrigated and failure of one eye to deviate is evidence of internuclear ophthalmoplegia and a lesion of the medial longitudinal fasciculus. This has been reported in poisoning with a variety of drugs including tricyclic antidepressants, phenothiazines, benzodiazepines, barbiturates, and ethanol and can be detected in 10 per cent of cases if caloric tests are carried out. Both sides are usually affected but internuclear ophthalmoplegia also occurs on one side only in acute poisoning.

In some cases, cold-induced lateral eye movements are followed after an interval of 5 to 15 s by forced downward gaze lasting several minutes, but the suggestion that the latter may be diagnostic of drug-induced coma requires further study before acceptance.

It is widely accepted that absence of oculocervical (abnormal 'doll's eye' responses) and vestibulo-ocular responses indicates severe brainstem damage and the likelihood that the patient will not

survive. However, this is not the case in acute poisoning where these reflexes may be abolished in patients who subsequently make a full recovery.

Management

Antidotes and methods to enhance elimination are available for very few poisons. Management of most poisoned patients is based on what has been called 'an orderly if unspectacular regimen of supportive therapy'.

Immediate treatment

A small but important number of poisoned patients arrive at hospital with respiratory obstruction, ventilatory failure, or in cardiorespiratory arrest. In these cases, conventional resuscitation takes precedence over detailed assessment of the patient and attempts to obtain a history. The opioid antagonist, naloxone, can be of inestimable value in emergency treatment. It is safe and should be used whenever there is the slightest suspicion that an opioid may be involved. Its use may transform a desperate situation for the better within seconds and even if it is given inappropriately, it is unlikely to have adverse effects.

Unconscious patients need scrupulous attention to respiration, hypotension, hypothermia, and other complications if they are to survive. Expert nursing is as important as medical measures.

Airway

Establishing and maintaining an adequate airway is of paramount importance in the management of the unconscious poisoned patient. The tongue falling back, dental plates being dislodged, other foreign bodies, buccal secretions, vomitus, and flexion of the neck may obstruct the airway. In the first instance, the neck should be extended and the tongue and jaw held forward. Secretions in the oropharynx must be removed and an oropharyngeal airway should be inserted before turning the patient into a semiprone position. If the cough reflex is absent, an endotracheal tube should be inserted to prevent aspiration into the lungs and allow regular aspiration of bronchial secretions. It is then important to ensure that the inspired air is adequately warmed and humidified.

Ventilation

Once a clear airway has been established the adequacy of spontaneous ventilation should be assessed from the results of arterial blood gas and pH measurements. These should be carried out in all patients who are unconscious irrespective of the presence or absence of features suggesting inadequate gas exchange. Unconscious poisoned patients often have a mild, mixed respiratory and metabolic acidosis with carbon dioxide tensions at the upper limit of normal and oxygen tensions that fall with increasing depth of coma. Increasing the oxygen content of the inspired air is often sufficient to correct hypoxia. Patients with acute respiratory failure should have an endotracheal tube inserted to reduce the respiratory dead space and thereby increase alveolar ventilation. If this does not reduce carbon dioxide tensions, assisted ventilation is indicated. High-inspired oxygen concentrations are imperative in patients with carbon monoxide and cyanide poisoning and in pulmonary oedema resulting from inhalation of irritant gases.

Hypotension

Hypotension in acute poisoning can be due to a variety of factors including a relative reduction in the intravascular volume secondary to expansion of the venous capacitance bed, metabolic acidosis, arrhythmias, the cardiodepressant effects of some drugs, and blood or fluid loss into the gut. Correct management of individual cases obviously depends on accurate identification of the causes. Young patients are generally not at risk of cerebral or renal damage unless the systolic blood pressure falls below 80 mmHg, but in those over the age of 40 years it is preferable to keep the systolic blood pressure above 90 mmHg. Hypotension often responds to elevation of the foot of the bed by 15 cm and, if this is unsuccessful, a central venous line should be inserted and the intravascular volume expanded as necessary. Dobutamine 2.5 to 10 µg/kg.min or adrenaline (epinephrine) at 1 to 10 µg/kg.min are indicated if hypotension is resistant to these measures.

Arrhythmias

Although many poisons are potentially cardiotoxic, the incidence of serious cardiac arrhythmias in acute poisoning is very low. Tricyclic antidepressants, b-adrenoceptor blocking drugs, chloral hydrate, cardiac glycosides, amphetamines, cocaine, bronchodilators (particularly theophylline and its derivatives), and antimalarial drugs are the most likely causes. Cardiotoxicity usually occurs together with other features of severe poisoning including metabolic acidosis, hypoxia, convulsions, respiratory depression, and abnormalities of electrolyte balance that should be corrected before considering the use of antiarrhythmic drugs. The latter have narrow therapeutic ratios and their use may further impair myocardial function. In general, drug therapy should only be given for persistent, life-threatening arrhythmias associated with peripheral circulatory failure. The drug used must be selected from a knowledge of the pharmacology and toxicology of the poison involved and in such a way that it will not further compromise cardiac function. Lignocaine is probably the drug of choice for clinically important ventricular tachydysrhythmias since its half-life is short and the dose can be adjusted readily.

Convulsions

Convulsions are potentially life-threatening because they cause hypoxia and metabolic acidosis and may precipitate cardiac arrhythmias and arrest. Short isolated convulsions do not require treatment but those that are recurrent or protracted should be suppressed with diazepam intravenously 10 mg in an adult, repeated as necessary. This drug is highly effective in adequate doses and alternatives are seldom needed. However, it is important to remember that giving benzodiazepines in this way may potentiate the respiratory depressant effects of other poisons and further complicate management. The combination of convulsions, coma, and vomiting, which may occur with overdosage of theophylline derivatives, is particularly dangerous and in these circumstances it may be preferable to paralyse the patient, insert an endotracheal tube, and start assisted ventilation. However, although this ensures control of the airway and oxygenation, thus avoiding the risk of inhalation of gastric contents, it does not suppress seizure activity; cerebral function must therefore be monitored and parenteral anticonvulsants given as required.

Hypothermia

Any poison that depresses the central nervous system may impair temperature regulation and cause hypothermia, especially when discovery of the patient is delayed and environmental temperatures are low. This important complication may be missed unless temperature is recorded rectally using a low-reading thermometer. In severe cases, peripheral and core temperatures should be monitored. Treatment includes nursing the patient in a warm room (27 to 29°C) and a heat conserving 'space blanket'. Cold intravenous fluids should be avoided and bottles for use should be stored in the room or the lines should pass through a heating device.

Hyperthermia

Rarely, body temperature may increase to potentially fatal levels after overdosage with central nervous system stimulants such as cocaine, amphetamines, phencyclidine, monoamine oxidase inhibitors, butyrophenomes, and theophylline and its derivatives. In such cases, muscle tone is often grossly increased and convulsions and rhabdomyolysis are common. Cooling measures, including administration of chlorpromazine, may be indicated and dantrolene should be given to reduce muscle tone.

Acid–base abnormalities

Acid–base disturbances commonly accompany coma due to drugs. Acute respiratory acidosis is less common than might be expected but some elevation of arterial carbon dioxide tensions towards the upper limit of normal is usual. This, in combination with mild hypoxia in the deeper grades of coma, produces overall acidaemia. In general, acidosis should be prevented and managed by ensuring adequate ventilation, oxygenation, and tissue perfusion, and control of convulsions rather than by giving bicarbonate. However, a number of poisons, particularly methanol and ethylene glycol, cause life-threatening metabolic acidosis, which should be corrected by infusion of sodium bicarbonate.

Acute respiratory alkalosis, often in combination with a minor metabolic acidosis, is commonly found in acute salicylate poisoning. The metabolic component may require treatment if it is the dominant feature and is causing overall acidaemia. Respiratory alkalosis should not be treated.

Electrolyte abnormalities

Electrolyte abnormalities may result from acid–base disturbances or the direct effects of poisons. Massive tissue damage, usually rhabdomyolysis, may allow potassium to leak from cells leading to potentially lethal hyperkalaemia. Cardiac glycosides cause hyperkalaemia secondary to loss from cells due to inhibition of the membrane sodium–potassium pump while the reverse occurs with sympathomimetic drugs. Ingestion of potassium salts, even in sustained release formulations, may lead to hyperkalaemia and fatal arrhythmias. Oxalic acid and ethylene glycol (which is metabolized to oxalic acid) may cause hypocalcaemia by leading to the formation of insoluble calcium oxalate that is deposited in tissues. Similarly, ingestion of fluorides is also a possible cause of hypocalcaemia, but the amounts children tend to ingest in the form of tablets to prevent dental caries seldom cause serious problems.

Bladder care

Urinary retention is a common complication of acute poisoning, particularly with tricyclic antidepressants and other drugs that have marked anticholinergic actions. However, bladder catheterization is all too often an unthinking measure in unconscious poisoned patients. Coma *per se* is not an indication for catheterization in poisoned patients, the great majority of whom regain consciousness within 12 h. The bladder can usually be induced to empty reflexly (provided it is not allowed to become grossly overdistended) by applying gentle suprapubic pressure. Catheterization should be reserved for those patients in whom suprapubic pressure is insufficient to empty the bladder, and in those thought to be developing renal failure.

Skin, muscle, and nerve lesions

Skin blisters may be found after poisoning with a wide variety of drugs including barbiturates, tricyclic antidepressants, and benzodiazepines, and non-drug toxins. They often occur over bony prominences that have been subjected to pressure and less frequently at sites where two skin areas have been in contact, such as the the inner aspects of the knees. They should be managed as partial thickness burns. Rhabdomyolysis is a further possible result of immobility and may occur in combination with skin lesions or independently. Drug overdose is the most common non-traumatic cause of this condition and it may lead to acute renal failure and, rarely, to ischaemic muscle contractures and long-term disability. Similarly, peripheral nerves such as the radial, ulnar, and common peroneal may be damaged by direct pressure while the patient is unconscious or by being entrapped in fibrosing muscle after rhabdomyolysis.

Antidotes

Naloxone for opioid analgesics, oxygen for carbon monoxide, and possibly, flumazenil for benzodiazepines are the only antidotes commonly needed in the management of unconscious poisoned patients. *N*-Acetylcysteine is used frequently for paracetamol poisoning. Other antidotes of proven value are listed in Table 2. They are seldom required and although their use in correct circumstances may be lifesaving, some are toxic in their own right and the reader is recommended to seek further advice from a poisons information service. Antivenoms for bites and stings by venomous animals are discussed in Chapter 8.2.

Reduction of poison absorption

Prevention of absorption of poisons through the lungs obviously requires removal from the toxic atmosphere and occasionally removal of soiled clothing as well. The latter is also necessary when absorption is thought to have been percutaneous. In addition, the contaminated skin should be thoroughly washed with soap and water.

While it appears logical to assume that removal of unabsorbed drug from the gastrointestinal tract ('gut decontamination') will be beneficial, the efficacy of current methods remains unproven and efforts to remove small amounts of 'safe' drugs are clearly not worthwhile or appropriate. The two major international societies of clinical toxicology (American Academy of Clinical Toxicology and the European Association of Poisons Centres and Clinical Toxicologists) have produced Position Statements on each method. The Position Statements are summarized below.

Gastric lavage

Gastric lavage should not be employed routinely in the management of poisoned patients. In experimental studies, the amount of marker removed by gastric lavage was highly variable and diminished with time. There is no certain evidence that its use improves clinical outcome and it may cause significant morbidity. Gastric lavage should not be considered, therefore, unless a patient has ingested a potentially life-threatening amount of a poison and the procedure can be undertaken within 1 h of ingestion. Even then, clinical benefit has not been confirmed in controlled studies. Unless a patient is intubated, gastric lavage is contraindicated if airway protective reflexes are lost. It is also contraindicated if a hydrocarbon with high aspiration potential or a corrosive substance has been ingested.

Syrup of ipecacuanha

Syrup of ipecacuanha should not be administered routinely in the management of poisoned patients. In experimental studies the amount of marker removed by syrup of ipecacuanha was highly variable and diminished with time. There is no evidence from clinical studies that syrup of ipecacuanha improves the outcome of poisoned patients and its administration, even in children, should be abandoned. In particular, syrup of ipecacuanha should not be administered to a patient who has a decreased level or impending loss of consciousness as aspiration pneumonia might ensue. In addition, it should not be administered to a patient who has ingested a corrosive substance or hydrocarbon with high aspiration potential.

Single-dose activated charcoal

Single-dose activated charcoal should not be administered routinely in the management of poisoned patients. Based on volunteer studies, the effectiveness of activated charcoal decreases with time; the greatest benefit is within 1 h of ingestion. The administration of activated charcoal may be considered if a patient has ingested a potentially toxic amount of a poison (which is known to be adsorbed to charcoal) up to 1 h previously; there are insufficient data to support or exclude its use after 1 h of ingestion. However, there is no evidence that the administration of activated charcoal improves clinical outcome. Unless a patient has an intact or protected airway, the administration of charcoal is contraindicated.

Cathartics

The administration of a cathartic alone has no role in the management of the poisoned patient and is not recommended as a method of gut decontamination. Experimental data are conflicting regarding the use of cathartics in combination with activated charcoal. No clinical studies have been published to investigate the ability of a cathartic, with or without activated charcoal, to reduce the bioavailability of drugs or to improve the outcome of poisoned patients. Based on available data, the routine use of a cathartic in combination with activated charcoal is not endorsed. If a cathartic is used, it should be limited to a single dose in order to minimize adverse effects.

Whole bowel irrigation

Whole bowel irrigation should not be used routinely in the management of the poisoned patient. Although some volunteer studies have shown substantial decreases in the bioavailability of ingested drugs, no controlled clinical trials have been performed and there is no conclusive evidence that whole bowel irrigation improves the outcome of the poisoned patient. Based on volunteer studies, whole bowel irrigation may be considered for potentially toxic ingestions of sustained-release or enteric-coated drugs. There are insufficient data to support or exclude its use for potentially toxic ingestions of iron, lead, zinc, or packets of illicit drugs, but it remains a theoretical option for these ingestions. Whole bowel irrigation is contraindicated in patients with bowel obstruction, perforation, ileus, and in patients with haemodynamic instability or a compromised airway. Whole bowel irrigation should be used cautiously in debilitated patients, or in patients with medical conditions that may be further compromised by its use.

Methods to increase poison elimination

Once a poison has been absorbed and providing there is no antidote, it is reasonable to consider the use of treatments that might speed its elimination from the body. Formerly, forced diuresis, peritoneal and haemodialysis, charcoal haemoperfusion and, less commonly, plasmapheresis were the techniques employed most commonly. In recent years, however, it has been shown that multiple doses of oral activated charcoal given over many hours significantly shortened the plasma half-life of many drugs, at least in volunteers.

Forced diuresis

In the past, forced diuresis enjoyed extensive use in the treatment of acute poisoning if only because it did not require special equipment and could be instituted rapidly and in any hospital. However, there was considerable ignorance of its rationale and its use for many poisons was not justified. The efficacy of forced diuresis depends on the poison being excreted unchanged by the kidney or as an active metabolite. However, most drugs are either degraded by the liver to non-toxic metabolites or have such large volumes of distribution that there is insufficient active drug elimination in urine for forced diuresis to be of any clinical value as the amount removed is insignificant compared with that removed by hepatic metabolism. Urine pH is more important than urine flow and in recent years there has been a trend away from forcing a diuresis (i.e. infusing large volumes of fluid) to attempting to alter urine pH alone.

Urine alkalinization

Most drugs are partly reabsorbed from the urine as it flows through the renal tubules. Reabsorption is confined to unionized, lipid-soluble molecules. Increasing the concentration of ionized drug in the urine should reduce reabsorption and further enhance elimination. This is achieved by manipulating urine pH. Thus, rendering the urine alkaline enhances elimination of weakly acidic compounds such as salicylates, phenobarbital, chlorpromamide, fluoride, and phenoxyacetate herbicides such as 2,4-D and mecoprop.

In practice, inducing an alkaline urine is only used in cases of poisoning due to salicylates and phenoxyacetate herbicides as phenobarbital poisoning may be treated effectively with multiple-dose activated charcoal.

Before alkalinizing the urine, it is important to correct plasma volume depletion and electrolyte and metabolic abnormalities. Sodium bicarbonate, most conveniently administered as an 8.4 per cent solution (1 mmol bicarbonate/ml), is infused intravenously to ensure that the pH of the urine, which is measured by narrow-range indicator paper or a pH meter, is more than 7.5 and preferably close to 8.5.

As urine alkalinization is a metabolically invasive procedure requiring frequent biochemical monitoring, and medical and nursing expertise, it should be performed in a critical care area.

Acid diuresis

Although, theoretically, induction of an acid diuresis should increase the elimination of basic drugs such as amphetamines, there is seldom any need to use it and no evidence that it is of value in cases of poisoning.

Multiple doses of oral activated charcoal

Multiple doses of activated charcoal aid the elimination of some drugs from the circulation by interrupting their enterohepatic circulation and adsorbing that which diffuses into the intestinal juices. The rate of transfer of the latter is dependent upon the blood supply to the gut, the area of mucosa available for transfer, and the concentration gradient of the drug across the mucosa. The adsorptive capacity of charcoal is such that zero drug concentrations are present in luminal fluid and that the diffusion gradient remains as high as possible. The process has been termed 'gut dialysis' since, in effect, the intestinal mucosa is being used as a semipermeable membrane.

The American Academy of Clinical Toxicology and the European Association of Poisons Centres and Clinical Toxicologists have published a Position Statement on multiple-dose activated charcoal. This confirms that although many studies in animals and volunteers have demonstrated that multiple-dose activated charcoal increases drug elimination significantly, this therapy has not yet been shown in a controlled study in poisoned patients to reduce morbidity and mortality. Further studies are required to establish its role and the optimal dosage regimen of charcoal to be administered.

Based on experimental and clinical studies, multiple-dose activated charcoal should be considered only if a patient has ingested a life-threatening amount of carbamazepine, dapsone, phenobarbital, quinine, or theophylline. In all of these cases there are data to confirm enhanced elimination, though no controlled studies have demonstrated clinical benefit.

Although volunteer studies have demonstrated that multiple-dose activated charcoal increases the elimination of amitriptyline, dextropropoxyphene, digitoxin, digoxin, disopyramide, nadolol, phenylbutazone, phenytoin, piroxicam, and sotalol, there are insufficient clinical data to support or exclude the use of this therapy in patients poisoned with these drugs.

The use of multiple-dose charcoal in salicylate poisoning is controversial. One animal study and two of four volunteer studies did not demonstrate increased salicylate clearance with multiple-dose charcoal therapy. Data in poisoned patients are insufficient at present to recommend the use of multiple-dose charcoal therapy for salicylate poisoning.

Multiple-dose activated charcoal did not increase the elimination of astemizole, chlorpropamide, doxepin, imipramine, meprobamate, methotrexate, phenytoin, sodium valproate, tobramycin, and vancomycin in experimental and/or clinical studies.

Unless a patient has an intact or protected airway, the administration of multiple-dose activated charcoal is contraindicated. It should not be used in the presence of intestinal obstruction. The need for concurrent administration of cathartics remains unproven and is not recommended. In particular, cathartics should not be administered to young children because of their propensity to cause fluid and electrolyte imbalance.

In conclusion, based on experimental and clinical studies, multiple-dose activated charcoal should be considered only if a patient has ingested a life-threatening amount of carbamazepine, dapsone, phenobarbital, quinine, or theophylline.

Recommended adult doses of charcoal for this purpose are 50 to 100 g initially, followed by 50 g 4-hourly or 25 g 2-houly until charcoal appears in the faeces or recovery occurs.

Further reading

American Academy of Clinical Toxicology/European Association of Poisons Centres and Clinical Toxicologists (1997/1999). Position Statem*Journal of Toxicology—Clinical Toxicolog* **35**, 699–762; **37**, 731–51.

Dialysis

Haemodialysis in acute poisoning is indicated most commonly for the treatment of acute renal failure and only infrequently to increase the elimination of poisons. The rate of elimination across the dialysis membrane depends upon a number of variables including the molecular weight of the poison, the extent to which it is protein bound, the concentration gradient, and pH of blood and dialysate. Haemodialysis is of little value in patients who ingest poisons with large volumes of distribution (such as tricyclic antidepressants) because the plasma contains only a small proportion of the total amount of drug in the body. Haemodialysis is indicated in patients with severe clinical features and high plasma concentrations of salicylate, lithium, methanol, isopropanol, ethylene glycol, and ethanol. Peritoneal dialysis increases the elimination of poisons such as ethylene glycol and methanol but is much less efficient than haemodialysis.

Haemoperfusion

Haemoperfusion involves the passage of blood through an adsorbent material that retains the poison. Activated charcoal is the most popular adsorbent. Within 4 to 6 h haemoperfusion can reduce significantly the body burden of compounds with a low volume of distribution (less than 1 litre/kg). The technique effectively removes barbiturates, carbamazepine, disopyramide, ethchlorvynol, glutethamide, meprobamate, methaqualone, theophylline, and trichloroethanol derivatives. However, there is now evidence that multiple-dose activated charcoal is as effective as haemoperfusion in phenobarbital, carbamazepine, and theophylline poisoning, and is simpler to use. Furthermore, barbiturate and non-barbiturate hypnotics are now prescribed only rarely.

Acetone

Acetone is a clear liquid with a characteristic pungent odour and sweet taste, used widely in industrial and household products. Once absorbed either through the lungs, skin, or gut, acetone is exhaled unchanged or metabolized to carbon dioxide.

Clinical features

Acetone has an irritating effect on the mucous membranes of the eyes, nose, and throat. Intoxication results in headache, excitement, restlessness, chest tightness, incoherent speech, nausea, and vomiting. Occasionally, gastrointestinal bleeding, coma, and convulsions have been reported.

Treatment

If toxicity has followed inhalation, remove from exposure and give supportive treatment. After ingestion, gut decontamination is not useful.

Further reading

International Programme on Chemical Safety (1998). *Environmental Health Criteria 207. Acetone*. World Health Organization, Geneva.

Acids

Acids commonly involved in cases of poisoning include the inorganic acids such as hydrochloric, hydrofluoric, nitric, phosphoric, and sulphuric acids; and organic acids such as acetic, formic, lactic, and trichloroacetic acids. Car battery acid typically contains 28 per cent sulphuric acid. Proprietary cleaning agents and antirust compounds often contain a mixture of hydrochloric and phosphoric acids.

Clinical features

On the skin acids behave characteristically as corrosives leading to erythema and burns. In the eyes, intense pain and blepharospasm are common, and corneal burns may occur. When ingested, acids flow rapidly along the lesser curvature of the stomach to the prepyloric region where they pool because of spasm of the pylorus and antrum to cause almost instantaneous coagulative necrosis of one or more layers of the stomach. In many cases, acids spare the oesophagus because of rapid transit and resistant squamous epithelium.

There is immediate pain in the mouth, pharynx, and abdomen, intense thirst, vomiting, haematemesis, and diarrhoea. The pain and mucosal oedema cause dysphagia and drooling saliva. Gastric and oesophageal perforation may occur resulting in chemical peritonitis. Other effects include hoarseness, stridor, and respiratory distress secondary to laryngeal and epiglottic oedema, shock, metabolic acidosis, leucocytosis, acute tubular necrosis, renal failure, hypoxaemia, respiratory failure, intravascular coagulation, and haemolysis.

Hydrofluoric acid ingestion causes chelation of calcium, with resultant weakness, paraesthesiae, tetany, convulsions, cardiac arrhythmias, and disturbed coagulation.

Treatment

Acid burns to the skin should be irrigated liberally with water or saline. Dressings are applied as for a thermal burn. Skin grafting may be necessary.

After ocular exposure, the eye should be irrigated preferably with saline for 15 to 30 min. Topical local anaesthetic is usually required to relieve pain and to overcome blepharospasm. Ophthalmic advice should be sought.

After ingestion a clear airway should be established. Opioids are often necessary for analgesia. Dilution and/or neutralization is contraindicated. Urgent panendoscopy is needed and resection of necrotic tissue and surgical repair should be undertaken to ensure survival, particularly if inorganic acids have been ingested. Total parenteral nutrition is often required. Corticosteroids confer no benefit and may mask abdominal signs of perforation; antibiotics should be given for established infection only.

Acid ingestion may result in antral, pyloric, or jejunal strictures, achlorhydria, protein-losing enteropathy, and gastric carcinoma.

Further reading

Advisory Committee on Pesticides (1998). *Evaluation of fully approved or provisionally approved products*. *Evaluation number 174: sulphuric acid*. Advisory Committee on Pesticides, Pesticides Safety Directorate, York.

Boyce SH, Simpson KA (1996). Hydrochloric acid inhalation: who needs admission? *Journal of Accident and Emergency Medicine* **13**, 422–4.

Ochi K *et al.* (1996). Surgical treatment for caustic ingestion injury of the pharynx, larynx, and esophagus *Acta Otolaryngologica* **116**, 116–19.

Stiff G *et al.* (1996). Corrosive injuries of the oesophagus and stomach: experience in management at a regional paediatric centre. *Annals of the Royal College of Surgeons of Englanc* **78,** 119–23.

Alkalis

Those commonly encountered in cases of poisoning include drain, lavatory, and pipe cleaners (sodium hydroxide), dishwashing detergents (sodium carbonate, sodium silicate, sodium tripolyphosphate), denture cleaning tablets (sodium perborate, sodium phosphate, sodium carbonate), urinary glucose testing tablets (sodium hydroxide), water sterilizing tablets (sodium dichloroisocyanurate), alkaline batteries, and sodium hypochlorite (a bleaching agent).

Clinical features

The features of eye, skin, and laryngeal contamination with alkalis are similar to those produced by acids (above). When ingested, alkalis typically damage the oesophagus but usually spare the stomach. There is little immediate oral discomfort but subsequently a burning sensation develops in the mouth and pharynx, together with epigastric pain, vomiting, and diarrhoea. Oesophageal ulceration with or without perforation may occur with mediastinitis, pneumonitis, cardiac injury, and aorto-enteric fistula formation as secondary complications of perforation.

Treatment

The treatment of corrosive injuries caused by alkalis is largely the same as for those produced by acids.

Corticosteroids do not alter the incidence of stricture formation but may decrease the need for surgical repair of strictures if they are used in conjunction with either anterograde or retrograde oesophageal dilation. Methylprednisolone at a dose of 40 mg intravenously 8-hourly in adults or prednisolone at 2 mg/kg per day intravenously can be given, until oral intake is resumed, when an equivalent dosage of prednisolone is given orally and tapered off over a period of 3 to 6 weeks. A broad-spectrum antibiotic, such as amoxicillin, should be prescribed at the same time as the corticosteroid.

Alkali ingestion may result in stricture formation and there is a risk of malignancy. The mean latent period for development of carcinoma of the oesophagus following alkali ingestion is more than 40 years.

Further reading

Anderson KD, Rouse TM, Randolph JG (1990). A controlled trial of corticosteroids in children with corrosive injury of the esophagus. *New England Journal of Medicine* **323**, 637–40.

Davis AR *et al.* (1997). Topical steroid use in the treatment of ocular alkali burns. *British Journal of Ophthalmology* **81**, 732–4.

Gaudreault P *et al.* (1983). Predictability of esophageal injury from signs and symptoms: a study of caustic ingestion in 378 children. *Pediatrics* **71**, 767–70.

Keskin E *et al.* (1991). The effect of steroid treatment on corrosive oesophageal burns in children. *European Journal of Pediatric Surgery* **1,** 335–8.

Lee KAP, Opeskin K (1995). Fatal alkali burns. *Forensic Science International* **72,** 219–27.

Ochi K *et al.* (1996). Surgical treatment for caustic ingestion injury of the pharynx, larynx and oesophagus. *Acta Otolaryngologica* **116**, 116–19.

a-Chloralose

a-Chloralose is marketed as cereal baits containing 4 per cent rodenticide, while technical a-chloralose (about 90 per cent pure) is used against moles and is occasionally encountered in self-poisoning episodes. The toxic amount for an adult is approximately 1 g and for an infant, 20 mg/kg body weight.

Clinical features

Toxic amounts of a-chloralose cause severe CNS excitation with hypersalivation, increased muscle tone, hyperreflexia, opisthotonus, and convulsions. Rhabdomyolysis is a potential complication. Coma, generalized flaccidity, and respiratory depression may follow.

Treatment

No treatment is required for ingestion of a-chloralose baits. Supportive measures are necessary when large amounts of bait or the technical compound is involved. Gastric emptying should not be carried out since the stimulation may provoke seizures.

Further reading

Thomas HM, Simpson D, Prescott LF (1988). The toxic effects of a-chloralose. *Human Toxicology* **7,** 285–7.

Aluminium (aluminum)

Aluminium hydroxide is used as an antacid and as a phosphate binder in the management of chronic renal failure. Aluminium sulphate is employed in water purification and paper manufacture. Aluminium may be absorbed orally and by inhalation. More than 90 per cent of absorbed aluminium is bound to transferrin. Though some accumulates in brain tissue, most body aluminium is stored in bone and the liver. It is excreted mainly via the kidneys.

Clinical features

Acute poisoning

Ingestion of a significant quantity of soluble aluminium salts such as aluminium sulphate gives rise to burning in the mouth and throat, nausea, vomiting, diarrhoea, abdominal pain, hypotension, seizures, haemolysis, haematuria, and rarely, hepatorenal failure. Topical aluminium sulphate may be irritant to the skin and eyes. By contrast, insoluble aluminium salts, such as aluminium oxide, do not produce an acute toxic response.

Chronic poisoning

Inhalation of 'stamped aluminium powder' can cause a persistent cough and breathlessness due to lung fibrosis or occupational asthma. Workers involved in aluminium production may be at increased risk of developing lung cancer.

'Dialysis dementia' involves the accumulation of aluminium, mainly in the brain (‹Section 20).

Aluminium has also been implicated in Alzheimer's disease and may contribute to osteomalacia in renal osteodystrophy. It may cause contact allergy.

Treatment

Desferrioxamine (deferoxamine) forms a stable complex with aluminium and mobilizes aluminium primarily from bone with subsequent urinary elimination of the chelate. Theoretically 100 mg of desferrioxamine can bind 4.1 mg of aluminium. As desferrioxamine is absorbed poorly from the gastrointestinal tract, parenteral therapy is preferred.

The desferrioxamine chelate is dialysable and all published clinical studies of aluminium chelation using desferrioxamine have involved patients in renal failure undergoing either dialysis or haemofiltration. Haemofiltration is probably superior to haemodialysis in enhancing aluminium elimination using desferrioxamine. As the aluminium–desferrioxamine chelate concentration reaches a maximum 12 to 24 h postinfusion, desferrioxamine should be administered shortly before dialysis for maximum benefit.

There is clinical evidence that desferrioxamine can improve aluminium-induced encephalopathy, bone disease, and anaemia in patients on dialysis. Desferrioxamine should therefore be prescribed when features of dialysis encephalopathy are present, when there is an increased body aluminium load, and there is clinical evidence of aluminium-related bone disease. In addition, desferrioxamine should be considered in the presence of severe, transfusion-dependent anaemia even in the absence of characteristic clinical or analytical features of aluminium overload. It has also been proposed that desferrioxamine should be employed in the presence of an increased (greater than 60 µg/l) serum aluminium concentration.

Conventionally, desferrioxamine at 40 to 80 mg/kg has been administered once weekly as a subcutaneous or intravenous infusion, reduced to 20 to 60 mg/kg when treatment is required for several months. However, recent studies suggest that desferrioxamine 5 mg/kg once weekly is an adequate alternative regime.

Further reading

International Programme on Chemical Safety (1997). *Environmental Health Criteria 194. Aluminium.* World Health Organization, Geneva.

McCarthy JT, Milliner DS, Johnson WJ (1990). Clinical experience with desferrioxamine in dialysis patients with aluminium toxicity. *Quarterly Journal of Medicine* **74**, 257–76.

Aluminium and zinc phosphides

Aluminium and zinc phosphides react with moisture in the air and the gastrointestinal tract to produce phosphine (see below), a gas with a garlic-like odour.

Clinical features

Exposure to phosphine causes lacrimation, rhinorrhoea, cough, breathlessness, chest tightness, dizziness, nausea, and drowsiness. Pulmonary oedema may develop later. Ingestion of aluminium phosphide causes vomiting, epigastric pain, peripheral circulatory failure, severe metabolic acidosis, and renal failure in addition to many of the features induced by inhalation of phosphine.

Treatment

Treatment is symptomatic and supportive. Gastric lavage should be considered if the poison has been ingested within 1 h.

Further reading

Gupta S, Ahlawat SK (1995). Aluminium phosphide poisoning: a review. *Journal of Toxicology—Clinical Toxicology* **33**, 19–24.

Ammonia

Ammonia, a colourless gas with a strong irritating odour, is used in aqueous solution in industry and in the home.

Clinical features

Ammonia may be absorbed by inhalation, ingestion, or percutaneously. It irritates the eyes, upper respiratory tract, and pharynx. Exposed surfaces may develop chemical burns, blisters, thrombosis of surface vessels, and severe local oedema that may lead to respiratory obstruction and death, if the larynx and glottis are involved. Inhaled high concentrations may cause dyspnoea, pulmonary oedema, and persistent lung damage.

Treatment

The casualty should be removed from the contaminated area. The eyes should be irrigated with water or 0.9 per cent saline for 15 to 30 min and an ophthalmic opinion sought as permanent blindness may result. Pulmonary complications should be treated with humidified supplemental oxygen, bronchodilators, and if necessary, assisted ventilation with positive end-expiratory pressure. Although widely employed, there is no conclusive evidence that diuretics and corticosteroids alter the prognosis. Patients who survive for 24 h are likely to recover fully.

Further reading

De La Hoz RE, Schlueter DP, Rom WN (1996). Chronic lung disease secondary to ammonia inhalation injury. *American Journal of Industrial Medicine* **29**, 209–14.

Wibbenmeyer LA *et al.* (1999). Our chemical burn experience: exposing the dangers of anhydrous ammonia. *Journal of Burn Care and Rehabilitation* **20**, 226–31.

Amfetamines and ecstasy (MDMA)

Amfetamine, dexamfetamine, methamfetamine, and 'ecstasy' (3,4-methy-lenedioxymethamfetamine, MDMA) stimulate the central nervous system. Poisoning with them is usually the result of their use for pleasurable purposes rather than single massive doses.

Clinical features

These drugs cause increased alertness and self-confidence, euphoria, extrovert behaviour, increased talkativeness with rapid speech, lack of desire to eat or sleep, tremor, dilated pupils, tachycardia, and hypertension. More severe intoxication is associated with excitability, agitation, paranoid delusions, hallucinations with violent behaviour, hypertonia, and hyperreflexia. Convulsions, rhabdomyolysis, hyperthermia, and cardiac arrhythmias may also develop. In severe cases of MDMA poisoning, hyperthermia, disseminated intravascular coagulation, rhabdomyolysis, acute renal failure, and hyponatraemia are observed. Hepatic damage has also been reported. Rarely, poisoning due to amphetamines may result in intracerebral and subarachnoid haemorrhage and acute cardiomyopathy; these complications may be fatal. Hyperthyroxinaemia may be found in chronic amphetamine users.

Treatment

Gastric lavage should be considered if a substantial overdose has been ingested in the preceding 1 h. Sedation with diazepam, chlorpromazine, or droperidol may be required. b-Adrenoceptor blocking drugs will antagonize the peripheral sympathomimetic actions of amphetamines.

Further reading

Boot BP, McGregor LS, Hall W (2000). MDMA (ecstasy) neurotoxicity: assessing and communicating the risks. *Lancet* **355,** 1818–21.

Ernst T *et al.* (2000). Evidence for long-term neurotoxicity associated with methamphetamine abuse. ^{1}H MRS study. *Neurology* **54**, 1344–9.

Maurer HH *et al.* (2000). Toxicokinetics and analytical toxicology of amphetamine-derived designer drugs ('Ecstasy'). *Toxicology Letters* **112–13,** 133–42.

McGuire P (2000). Long term psychiatric and cognitive effects of MDMA use. *Toxicology Letters* **112–13,** 153–6.

Ricaurte GA *et al.* (2000). Toxicodynamics and long-term toxicity of the recreational drug, 3,4-methylenedioxymethamphetamine (MDMA, 'ecstasy') *Toxicology Letters* **112–13,** 143–6.

Angiotensin-converting enzyme (ACE) inhibitors

Clinical features

Anorexia, nausea, abdominal discomfort, headache, and paraesthesiae have been reported. In addition, hypotension (which may be mediated by the endogenous opioid system), sinus tachycardia, bronchospasm, and hyperkalaemia may develop. Fatalities have been reported.

Treatment

Gastric lavage or activated charcoal administration should be considered if the patient presents within 1 h of a substantial overdose. Supportive therapy should then be employed, including volume expansion with plasma expanders for hypotension. Naloxone in a dose of 0.8 to 1.2 mg may reverse ACE inhibitor-induced hypotension. Marked hyperkalaemia may require an intravenous infusion of glucose (50 g) and soluble insulin (15 units).

Further reading

Lip GYH, Ferner RE (1995). Poisoning with anti-hypertensive drugs: angiotensin converting enzyme inhibitors. *Journal of Human Hypertension* **9**, 711–15.

Antibacterial agents

Most patients develop no symptoms and require no treatment. Transient nausea, vomiting, and diarrhoea may occur. There have been rare reports of renal failure after overdosage with co-trimoxazole, pancreatitis with erythromycin, and haemorrhagic cystitis with amoxicillin.

Further reading

Berger TM *et al.* (1992). Acute pancreatitis in a 12 year old girl after an erythromycin overdose. *Pediatrics* **90,** 624–6.

Cohen H, Francisco DH (1994). Twelve-gram overdose of ciprofloxacin with mild symptomatology. *Annals of Pharmacotherapy* **28,** 805–6.

Jones DP *et al.* (1993). Acute renal failure following amoxycillin overdose. *Clinical Pediatrics* **32**, 735–9.

Anticholinergic substances

Anticholinergic substances are occasionally abused. Plants have been used for this purpose for many years.

A large number of drugs currently used in medicine also have anticholinergic properties and have been misused. They include antihistamines, particularly cyclizine, antiparkinsonian drugs such as benzhexol (trihexyphenidyl), benzatropine, orphenadrine, biperiden, and tricyclic antidepressants (see below).

Further reading

Burns MJ *et al.* (2000). A comparison of physostigmine and benzodiazepines for the treatment of anticholinergic poisoning. *Annals of Emergency Medicine* **35,** 374–81.

Ramirez M, Rivera E, Ereu C (1999). Fifteen cases of atropine poisoning after honey ingestion. *Veterinary and Human Toxicology* **41**, 19–20.

Thabet H *et al.* (1999). *Datura stramonium* poisoning in humans. *Veterinary and Human Toxicology* **41**, 320–21.

Anticoagulant rodenticides

Warfarin was widely used as a rodenticide until target species developed resistance to it. The newer anticoagulant rodenticides such as brodifacoum, bromodialone, chlorophacinone, coumatetralyl, difenacoum, diphacinone, and flocoumafen are more potent antagonists of vitamin K$_1$ than warfarin and reduce the synthesis of clotting factors II, VII, IX, and X. In the case of short-acting formulations, such as those containing warfarin, prolongation of the International Normalized**INR**) will be evident within 24 h and patients can remain anticoagulated for several days. In contrast, the ingestion of the more potent anticoagulant rodenticides may result in prolongation of the INR for weeks or months.

Clinical features

Gastrointestinal bleeding, haematuria, and bruising are the commonest features, though the most common site of fatal haemorrhage is intracranial.

Treatment

If the patient is not receiving an anticoagulant therapeutically, is not bleeding, and if the INR is more than 8, give vitamin K$_1$ (phytomenadione) in a dose of 5 mg slowly intravenously. If there is active bleeding give prothrombin complex concentrate at 50 units/kg (or fresh frozen plasma 15 ml/kg if the concentrate is not available) and vitamin K$_1$ 5 to 10 mg intravenously. Treatment may be required for several weeks in the case of the more potent anticoagulant rodenticides. The INR should be monitored for at least 2 weeks after stopping vitamin K$_1$ therapy. If the patient is receiving warfarin therapeutically see below under Warfarin for further discussion.

Further reading

Casner PR (1998). Superwarfarin toxicity. *American Journal of Therapeutics* **5**, 117–20.

McCarthy PT *et al.* (1997). Covert poisoning with difenacoum: clinical and toxicological observations. *Human and Experimental Toxicology* **16**, 166–70.

Antihistamines

First-generation antihistamines include brompheniramine, chlorpheniramine (chlorphenamine), cyclizine, diphenhydramine, mepyramine, methapyrilene, promethazine, and trimeprazine. Second-generation drugs include astemizole and terfenadine. The toxicity of the two groups varies.

Clinical features

The older antihistamines have anticholinergic actions and their effects are therefore similar to the tricyclic antidepressants (see below) although convulsions, coma, respiratory depression, arrhythmias (other than sinus tachycardia), and death are rare.

Astemizole and terfenadine lack anticholinergic actions but are cardiotoxic causing Q_c interval prolongation and ventricular tachycardia, including the *torsade de pointes* type. Associated giant U waves have been described. Terfenadine in overdose can cause convulsions.

Treatment

Gastric lavage may be undertaken or oral activated charcoal may be given if the patient presents less than 1 h after the ingestion of a substantial overdose of a first-generation antihistamine. The patient should be observed for about 12 h with cardiac monitoring if the Q–T interval is prolonged. Intravenous magnesium sulphate may abolish serious ventricular arrhythmias.

Further reading

June RA, Nasr I (1997). Torsades de pointes and terfenadine ingestion. *American Journal of Emergency Medicine* **15**, 542–3.

Zareba W *et al.* (1997). Electrocardiographic findings in patients with diphenhydramine overdose. *American Journal of Cardiology* **80,** 1168–73.

Antiparkinsonian drugs

Amantadine, benzhexol, and orphenadrine have anticholinergic effects in overdosage. Orphenadrine is probably the most toxic and has caused deaths.

Clinical features and treatment

The features of poisoning are similar to those of the tricyclic antidepressants and should be managed in the same way.

Further reading

Jones AL, Proudfoot AT (1997). The features and management of poisoning with drugs used to treat Parkinson's disease. *Quarterly Journal of Medicine* **91**, 613–16.

Antiseptics and disinfectants

Once these solutions commonly contained phenol but phenol has been replaced largely by small quantities of either chlorophenol or chloroxylenol which, although less toxic than phenol, can be hazardous if ingested in large quantities. More dangerous are isopropanol and ethanol (see below).

Clinical features

Ingestion of a substantial quantity results in a sensation of burning in the mouth and throat, followed by drowsiness, stupor, depression of respiration, and coma.

Treatment

Management is supportive. (See also ethanol and isopropanol.)

Further reading

Chan TYK (1994). Poisoning due to Savlon (cetrimide) liquid. *Human and Experimental Toxicology* **13**, 681–2.

Chan TYK, Critchley JAJH (1996). Pulmonary aspiration following Dettol poisoning: the scope for prevention. *Human and Experimental Toxicology* **15**, 843–6.

Arsenic

Arsenic forms both trivalent and pentavalent derivatives. Inorganic arsenical compounds may generate arsine gas (see below) when in contact with acids, reducing metals, sodium hydroxide, and aluminium. Some 90 per cent of an ingested dose of most inorganic arsenicals is absorbed. The half-life is in the range of 1 to 3 days. Excretion is predominantly in the urine. Soluble arsenical compounds can also be absorbed by inhalation, but skin absorption is generally poor. In exposed individuals high concentrations of arsenic are present in bone, hair, and nails.

Clinical features

Acute poisoning

This can follow accidental, suicidal, or deliberate ingestion, the toxicity being largely dependent on the water solubility of the ingested compound. Within 2 h of substantial ingestion of a soluble arsenical compound, severe haemorrhagic gastritis or gastroenteritis may ensue with collapse and death usually within 4 days. A metallic taste, salivation, muscular cramps, facial oedema, difficulty in swallowing, hepatorenal dysfunction, convulsions, and encephalopathy are reported. A peripheral neuropathy (predominantly sensory), striate leuconychia (Mee's lines), and hyperkeratotic, hyperpigmented skin lesions are common in those surviving a near-fatal ingestion. In moderate or severe arsenic poisoning investigations may show anaemia, leucopenia, thrombocytopenia, and disseminated intravascular coagulation. ECG abnormalities have been reported and include Q–T prolongation and ventricular arrhythmias.

Chronic poisoning

The ingestion of arsenic in contaminated drinking water (recently in Bangladesh) or 'tonics' has led to progressive weakness, anorexia, nausea, vomiting, stomatitis, colitis, increased salivation, epistaxis, bleeding gums, conjunctivitis, weight loss, and low-grade fever. Characteristically there is hyperkeratosis of the palms and soles of the feet, 'raindrop' pigmentation of the skin, and 'Mee's lines' on the nails. There is an increased risk of skin cancer (usually squamous cell epithelioma) in affected individuals. A symmetrical peripheral neuropathy is typical. Hearing loss, psychological impairment, and electroencephalogram changes have been reported. Other chronic effects include disturbances of liver function and ulceration and perforation of the nasal septum. Chronic exposure to arsenic has been linked to lung cancer.

Treatment

Traditionally, dimercaprol (British Anti-Lewisite, BAL) has been the recommended chelator. However, DMPS (unithiol) and DMSA (succimer) are preferable, if available, as they are more effective in reducing the arsenic content of tissues and, unlike dimercaprol, they do not cause accumulation of arsenic in the brain. DMSA and DMPS may be given orally in a dose of 30 mg/kg body weight daily for 5 days, whereas dimercaprol must be given by a deep intramuscular injection in a dose of 2.5 to 5 mg/kg 4-hourly for 2 days followed by 2.5 mg/kg intramuscularly twice daily for 1 to 2 weeks.

Further reading

Cullen NM, Wolf LR, St Clair D (1995). Pediatric arsenic ingestion. *American Journal of Emergency Medicine* **13**, 432–5.

Kingston RL, Hall S, Sioris L (1993). Clinical observations and medical outcome in 149 cases of arsenate ant killer ingestion. *Journal of Toxicology—Clinical Toxicology* **31**, 581–91.

Wong SS, Tan KC, Goh CL (1998). Cutaneous manifestations of chronic arsenicism: review of seventeen cases. *Journal of the American Academy of Dermatology* **38,** 179–85.

Arsine

Arsine is a colourless, non-irritating gas which binds with oxidized haemoglobin causing marked intravascular haemolysis. Haemoglobinuria and acute renal tubular necrosis then develop.

Clinical features

There is usually a delay of some 2 to 24 h after exposure before the onset of eadache, malaise, weakness, dizziness, breathlessness, migratory abdominal pain, fever, tachycardia, tachypnoea, nausea, and vomiting. A bronze skin colour is noted in some patients but most have the typical appearance of a jaundiced patient. Acute renal failure is observed by the third day after substantial exposure and the urine is dark red then brown before anuria ensues. Investigations will show leucocytosis, reticulocytosis, elevated

plasma haemoglobin, and haemoglobinuria.

Treatment

If haemolysis is severe, plasmapheresis or exchange transfusion should be undertaken and, if renal failure ensues, haemodialysis/filtration. Dimercaprol and other chelating agents are of no value.

Further reading

Romeo L *et al.* (1997). Acute arsine intoxication as a consequence of metal burnishing operations. *American Journal of Industrial Medicine* **32**, 211–16.

Barbiturates

Amylobarbital, butobarbital, cyclobarbital, heptabarbital, hexabarbital, pentobarbital, and secobarbital are regarded as being short- or medium-acting. The more lipid soluble, shorter-acting preparations are associated commonly with more serious poisoning than phenobarbital and barbital, which are much more water-soluble.

Clinical features

Impairment of consciousness, respiratory depression, hypotension, and hypothermia are typical and potentiated by alcohol and benzodiazepines. There are no specific neurological signs. Hypotonia and hyporeflexia are the rule and the plantar responses are either flexor or absent. Hypotension, skin blisters, and rhabdomyolysis may develop. During recovery from coma, with or without hypothermia, it is common to observe a peak of temperature that cannot be explained by infection. Most deaths result from respiratory complications.

Treatment

Gastric lavage may be considered if it can be undertaken within 1 h of overdose; supportive measures should be used as appropriate. Although charcoal haemoperfusion is very effective for severely poisoned patients, phenobarbital can be removed efficiently by multiple-dose oral activated charcoal.

Further reading

Hantson P *et al.* (1996). Severe hypoxia and hypothermia following barbiturate poisoning. *Intensive Care Medicine* **22**, 998–9.

Benzene

Benzene is a colourless, volatile liquid with a pleasant odour. It is an ingredient in many paints and varnish removers and some petrols (gasolines). About 10 per cent of inhaled benzene is excreted unchanged in the breath. The remainder is metabolized by mixed function oxidase enzymes predominantly in the liver, but also in the bone marrow, the target organ of benzene toxicity.

Clinical features

Acute exposure

Following inhalation or ingestion, euphoria, dizziness, weakness, headache, blurred vision, mucous membrane irritation, tremor, ataxia, chest tightness, respiratory depression, cardiac arrhythmias, coma, and convulsions have been reported. Direct skin contact with liquid benzene may produce marked irritation.

Chronic exposure

The toxic effects of chronic poisoning may not become apparent for months or years after initial contact and may develop after all exposure has ceased.

Anorexia, headache, drowsiness, nervousness, and irritability are well described. Anaemia (including aplastic anaemia), leucopenia, thrombocytopenia, pancytopenia, leukaemia, lymphomas, chromosomal abnormalities, and cerebral atrophy have been reported. Patients have recovered after as long as a year of almost complete absence of formation of new blood cells. A dry, scaly dermatitis may develop on prolonged or repeated skin exposure to liquid benzene.

Treatment

Following removal from the contaminated atmosphere, treatment should be directed towards symptomatic and supportive measures. Gastric lavage is hazardous as aspiration is likely to occur.

Further reading

Barbera N, Bulla G, Romano G (1998). A fatal case of benzene poisoning. *Journal of Forensic Sciences* **43**, 1250–1.

Ireland B *et al.* (1997). Cancer mortality among workers with benzene exposure. *Epidemiology* **8,** 318–20.

Benzodiazepines

These are widely used as tranquillizers, hypnotics, and sedatives.

Clinical features

Although many benzodiazepines have active metabolites that account for their sometimes prolonged sedative effects, they all share a remarkable safety when taken alone in overdose. However, there is individual variation in response; some otherwise healthy elderly people respond to an overdose with prolonged toxicity. Benzodiazepines potentiate the effects of other CNS depressants, particularly alcohol, tricyclic antidepressants, and barbiturates. Dizziness, drowsiness, ataxia, and slurred speech are the usual features while coma, respiratory depression, and hypotension are uncommon and usually mild. Flurazepam is most likely to cause significant CNS depression.

Treatment

Gastric lavage is unnecessary unless the overdose exceeds 30 therapeutic doses in an adult and the patient presents within 1 h. In severe poisoning, the specific benzodiazepine antagonist, flumazenil, may be indicated; 0.5 mg is given intravenously over 30 s and, if necessary, a further 0.5 mg over 30 s. Most patients will respond to a total dose of between 1 and 3 mg.

Further reading

Hojer J, Baehrendtz S, Gustafsson L (1989). Benzodiazepine poisoning: Experience of 702 admissions to an intensive care unit during a 14-year period *Journal of Internal Medicine* **226,** 117–22.

Weinbroum A *et al.* (1996). Use of flumazenil in the treatment of drug overdose: a double-blind and open clinical study in 110 patients. *Critical Care Medicine* **24**, 199–206.

Benzyl alcohol

Benzyl alcohol has been used as a preservative in intravascular flush solutions and in drug formulations. Benzyl alcohol is metabolized to benzoic acid that is then conjugated with glycine in the liver and excreted as hippuric acid. The immature liver's capacity to metabolize benzoic acid is limited and when exceeded leads to accumulation of this metabolite and metabolic acidosis.

Clinical features

In 1982, a syndrome consisting of metabolic acidosis, convulsions, neurological deterioration (due to intraventricular haemorrhage), gasping respirations, hepatic and renal abnormalities, cardiovascular collapse, and death was described in small premature infants between 2 and 14 days of age. The removal of benzyl alcohol solutions from neonatal units led to a considerable reduction both in morbidity and mortality and in particular there was a reduction in cases of kernicterus and intraventricular haemorrhage.

Further reading

Anderson CW *et al.* (1984). Benzyl alcohol poisoning in a premature newborn infant. *American Journal of Obstetrics and Gynecology* **130**, 344–6.

b-Adrenoceptor blocking drugs

b-Adrenoceptor blocking drugs antagonize the effects of endogenous catecholamines on the heart and other tissues by competitive inhibition at b-adrenoceptors. In overdose these drugs exhibit a marked negative inotropic action.

Clinical features

Sinus bradycardia may be the only feature following a small overdose, but if a substantial amount has been ingested, coma, convulsions (particularly with propranolol), profound bradycardia, and hypotension may occur. Other effects include drowsiness, delirium, hallucinations, low-output cardiac failure, and cardiorespiratory arrest (asystole or ventricular fibrillation). Bronchospasm and hypoglycaemia occur rarely.

First-degree heart block, intraventricular conduction defects, right and left bundle branch block, ST segment elevation, ventricular extrasystoles, and disappearance of the P wave may be noted on the electrocardiogram. Sotalol has been reported to cause Q–T interval prolongation and ventricular arrhythmias and asystole may follow severe overdose from any b-adrenoceptor blocking drug.

Treatment

A delay in treatment may be fatal in patients who are severely poisoned. The blood pressure and cardiac rhythm of the patient should be monitored immediately in an intensive care area and supportive measures implemented. Gastric lavage should be considered in adults who have ingested a substantial overdose less than 1 h previously; 0.6 to 1.2 mg of atropine intravenously may prevent vagal-induced cardiovascular collapse during this procedure.

Glucagon is the drug of choice for severe hypotension and should be given in a bolus dose of 50 to 150 µg/kg (typically 10 mg in an adult) over 1 min, followed by an infusion of 1 to 5 mg/h according to response.

Insertion of a temporary transvenous pacemaker wire, atropine, and isoprenaline 0.5 to 10 µg/min intravenously or other inotropic agents have been recommended but are probably less effective than glucagon. Occasionally, 5 to 10 mg of diazepam intravenously may be needed for convulsions. If bronchospasm supervenes, salbutamol (albuterol) by nebulizer, or aminophylline by intravenous infusion, should be employed. Hypoglycaemia should be corrected.

Further reading

Lip GYH, Ferner RE (1995). Poisoning with anti-hypertensive drugs: b-adrenoreceptor blocker drugs. *Journal of Human Hypertension* **9,** 213–21.

Taboulet P *et al.* (1993). Pathophysiology and management of self-poisoning with b-blockers. *Journal of Toxicology—Clinical Toxicology* **31,** 531–51.

b_2-Adrenoceptor stimulants

Poisoning with b_2-adrenoceptor stimulants, including fenoterol, pirbuterol, reprobuterol, rimiterol, salbutamol, and terbutaline, has followed deliberate and accidental ingestion of these drugs and may also result from confusion over the difference between oral and parenteral doses.

Clinical features

These include a feeling of excitement, hallucinations, and agitation, accompanied by palpitations, tachycardia, tremor, and peripheral vasodilation. More serious complications such as hypokalaemia, ventricular tachyarrhythmias, ECG changes of myocardial ischaemia, pulmonary oedema, convulsions, hyperglycaemia, and lactic acidosis are uncommon.

Treatment

Gastric lavage may be considered or activated charcoal administered if the patient presents within 1 h of a substantial overdose. Hypokalaemia should be corrected as soon as possible by the administration of an infusion of potassium at a rate of 40 to 60 mmol/h diluted in 1 litre 5 per cent dextrose. A non-selective b-blocker, such as propanolol 1 to 5 mg by slow intravenous injection will also reverse hypokalaemia induc$\bar{2}$-adrenoceptor stimulants. However, its use may exacerbate pre-existing chronic air flow obstruction. Methods to increase elimination have no role.

Further reading

Leikin JB *et al.* (1994). Hypokalemia after pediatric albuterol overdose: a case series. *American Journal of Emergency Medicine* **12**, 64–6.

Bismuth chelate (tripotassium dicitratobismuthate)

Although bismuth absorption from bismuth chelate is low after a therapeutic dose, a significant quantity may be absorbed after overdose. Renal toxicity is dose-dependent in animals and is directed primarily towards the tubular epithelial cells.

Clinical features

Self-poisoning with large doses of bismuth chelate has caused reversible renal failure up to 10 days after overdose and at least one death. During prolonged (and sometimes high-dose) therapy, bismuth-induced encephalopathy has been reported.

Treatment

If a patient presents within 1 h of a substantial overdose, gastric lavage should be considered. Dimercaprol can lower brain bismuth concentrations though there is no evidence that it can prevent nephrotoxicity. DMSA and DMPS may be effective oral alternatives.

Further reading

Akpolat I *et al.* (1996). Acute renal failure due to overdose of colloidal bismuth. *Nephrology, Dialysis, Transplantion* **11**, 1890–1.

Bleaches and lavatory leaners

Household bleach is normally a 3 to 6 per cent solution of sodium hypochlorite, whereas industrial bleaches contain more than 10 per cent. Some bleaches also contain sodium hydroxide. Household bleach may give rise to toxic gases such as chlorine if mixed with other cleaning agents in a lavatory bowl.

Clinical features

Ingestion may cause a burning sensation in the mouth, throat, and oesophagus, accompanied by a sensation of thirst, vomiting, and abdominal discomfort. Pharyngeal and laryngeal oedema and hypernatraemia may develop.

Treatment

When small quantities of household bleach have been ingested, liberal fluids by mouth are all that is required. Gastric lavage should only be considered if concentrated bleach has been swallowed less than 1 h previously. Endoscopy should be performed if industrial bleach has been ingested.

Inhalation of gases liberated by mixing bleach with other products may result in severe respiratory irritation and pulmonary oedema. Treat as for inhalation of chlorine.

Further reading

Hilbert G *et al.* (1997). Euro bleach: fatal hypernatremia due to 13.3 per cent sodium hypochlorite. *Journal of Toxicology—Clinical Toxicology* **35**, 635–6.

Kiristioglu I *et al.* (1999). Is it necessary to perform an endoscopy after the ingestion of liquid household bleach in children? *Acta Paediatrica* **88**, 233–4.

Butyrophenones

Benperidol, droperidol, haloperidol, and triperidol are used as antipsychotic and neuroleptic agents.

Clinical features

Overdosage may result in drowsiness and hypotension, but acute dystonic reactions are the most dramatic consequences. Neuroleptic malignant syndrome has also been reported.

Treatment

Treatment is supportive. Acute dystonic reactions should be treated with benztropine (benzatropine) 1 to 2 mg or procyclidine 5 to 10 mg intravenously for an adult.

Further reading

Yoshida I *et al.* (1993). Acute accidental overdosage of haloperidol in children. *Acta Paediatrica Scandinavica* **82**, 877–80.

Cadmium

Cadmium compounds are poorly absorbed orally but are well absorbed through the lungs. Cadmium is deposited in the liver and kidneys and very slowly excreted in the urine (half-life 10 to 30 years).

Clinical features

Acute poisoning

Inhalation of cadmium oxide fumes produced in welding or cutting has led to the development of severe lung damage and death. Often there are no initial symptoms but after some 4 to 10 h there is increasing respiratory distress. Chills and tremor accompany dyspnoea, cough, and chest pain. Severe pulmonary oedema may develop, or chemical pneumonitis in less severe cases. Recovery may be complicated by progressive pulmonary fibrosis.

The ingestion of cadmium salts (more than 3 mg/kg body weight) may lead to gastrointestinal disturbance which, in severe cases, may progress to circulatory ollapse, acute renal failure, pulmonary oedema, and death.

Chronic poisoning

Repeated exposure to cadmium leads to renal tubular dysfunction with glycosuria, aminoaciduria, and hypercalciuria, an increased incidence of renal stones and osteomalacia. Less common features include anosmia, anaemia, teeth discoloration, and neuropsychological impairment. Later, emphysema may develop.

Workers repeatedly exposed to high concentrations of cadmium have developed carcinoma of the prostate or lung.

Treatment

There is no clinical evidence that any currently available antidote chelates a substantial body burden of cadmium.

Further reading

Järup L *et al.* (1998). Health effects of cadmium exposure—a review of the literature and a risk estimate. *Scandinavian Journal of Work, Environment and Health*: **24**, 1–51.

Calcium-channel blockers

Calcium-channel blockers (amlodipine, diltiazem, felodipine, isradipine, nicardipine, nifedipine, nimodipine, verapamil) interfere with the inward transmembrane passage of calcium ions in myocardial cells, the cardiac conducting system, and vascular smooth muscle.

Clinical features

In overdose, calcium-channel blockers cause nausea, vomiting, dizziness, slurred speech, confusion, sinus bradycardia and tachycardia, prolonged atrioventricular conduction, atrioventricular dissociation, hypotension, pulmonary oedema, respiratory arrest, convulsions, coma, hyperglycaemia, and metabolic acidosis. Large overdoses carry a poor prognosis, particularly in patients with ischaemic heart disease and in those taking b-adrenoceptor blocking agents.

Treatment

Gastric lavage should be considered in all patients who present within 1 h of substantial overdose or, alternatively, 50 to 100 g of activated charcoal may be administered. Calcium gluconate in a dose of 10–20 ml of 10 per cent solution intravenously may reverse prolonged intracardiac conduction times but inotropic support with dobutamine at 2.5 to 10 µg/kg.min or isoprenaline at 0.5 to 10 µg/kg.min by intravenous infusion, will also be needed to maintain cardiac output in severe cases.

Further reading

Lip GYH, Ferner RE (1995). Poisoning with anti-hypertensive drugs: calcium antagonists. *Journal of Human Hypertension* **9**, 155–61.

Yuan TH *et al.* (1999). Insulin-glucose as adjunctive therapy for severe calcium channel antagonist poisoning. *Journal of Toxicology—Clinical Toxicology* **37**, 463–74.

Cannabis

Cannabis is obtained from the plan*Cannabis sativa* that contains many active substances of which the most important are the tetrahydrocannabinols.

Smoking is the common route of use of cannabis. It is occasionally ingested and, rarely, made into a 'tea' and injected intravenously.

Clinical features

Euphoria with drowsiness and distorted and heightened images, colours, and sounds are the usual effects of this compound. Tactile sensations may also be altered. A tachycardia is often present and heavy use may lead to conjunctival suffusion, hypotension, and ataxia. Higher doses induce auditory hallucinations, confusion, depersonalization, and panic. Some people find the distortion of perception pleasurable but novice users may panic and seek medical help. Long-term use may predispose to psychosis.

Intravenous injection of cannabis infusions causes serious illness. Within a few minutes, there is nausea, vomiting, and chills followed after an interval of 1 h or so by profuse watery diarrhoea, tachycardia, hypotension, and arthralgia. A marked neutrophil leucocytosis is often present and hypoglycaemia has been reported in some cases. There may also be transient renal failure.

Treatment

Most patients respond to reassurance. Sedation with intravenous diazepam may be required for those whose behaviour is disruptive or who are clearly very distressed. Those who have injected cannabis infusions should be treated supportively.

Further reading

Ashton CH (1999). Adverse effects of cannabis and cannabinoids. *British Journal of Anaesthesia* **83,** 637–49.

Fant RV *et al.* (1998). Acute and residual effects of marijuana in humans. *Pharmacology, Biochemistry and Behavior* **60,** 777–84.

Carbamate insecticides

Like many organophosphorus compounds, carbamate insecticides inhibit acetylcholinesterase. However, in comparison the duration of this effect is relatively short-lived since the carbamate–enzyme complex tends to dissociate spontaneously.

Clinical features

See organophosphorus insecticide poisoning (below).

Treatment

Symptomatic cases require atropine, but the use of oximes is usually unnecessary; rapid recovery within 24 h is the rule.

Further reading

Wagner SL (1997). Diagnosis and treatment of organophosphate and carbamate intoxication. *Occupational Medicine (Philadelphia)* **12,** 239–49.

Carbamazepine

Carbamazepine is structurally related to the tricyclic antidepressants and has similar anticholinergic actions. Overdosage may therefore result in a dry mouth, drowsiness, coma, and convulsions. Cardiotoxicity similar to that seen in tricyclic antidepressant poisoning also occurs but is uncommon. Relapse into coma has been described during the course of recovery, probably due to continuing drug absorption. Treatment should include administration of activated charcoal or gastric lavage, if appropriate, and supportive therapy. Diazepam may be required to treat convulsions. Carbamazepine elimination is hastened by giving multiple doses of activated charcoal.

Further reading

Apfelbaum JD *et al.* (1995). Cardiovascular effects of carbamazepine toxicity. *Annals of Emergency Medicine* **25**, 631–5.

Boldy DAR *et al.* (1987). Activated charcoal for carbamazepine poisoning. *Lancet* **i**, 1027.

Durelli L, Massazza U, Cavallo R (1989). Carbamazepine toxicity and poisoning. *Medical Toxicology and Adverse Drug Experience* **4**, 95–107.

Montoya-Cabrera MA *et al.* (1996). Carbamazepine poisoning in adolescent suicide attempters. Effectiveness of multiple-dose activated charcoal in enhancing carbamazepine elimination. *Archives of Medical Research* **27,** 485–9.

Carbon dioxide

Carbon dioxide is a colourless gas that is also available commercially as a solid for refrigeration purposes ('dry ice'). High concentrations may accumulate in wells, silos, manholes, and mines.

Clinical features

Dyspnoea, cough, headache, dizziness, sweating, restlessness, paraesthesiae, and sinus tachycardia are features after modest carbon dioxide exposure. Higher concentrations produce psychomotor agitation, myoclonic twitches, eye flickering, coma, and convulsions. Death occurs from acute cardiorespiratory depression.

Skin contact with 'dry ice' may result in frostbite and local blistering.

Treatment

The casualty should be removed from the contaminated environment. Thereafter, supportive care should be employed.

Further reading

Guillemin MP, Horisberger B (1994). Fatal intoxication due to an unexpected presence of carbon dioxide. *Annals of Occupational Hygiene* **38**, 951–7.

Carbon disulphide

Carbon disulphide is used as a fumigant for grain and as a solvent, particularly in the rayon industry. It is a clear, colourless, volatile liquid with an odour like that of decaying cabbage.

Clinical features

Acute exposure

Acute poisoning is rare. Absorption occurs through the skin as well as by inhalation. Carbon disulphide, due to its potent defatting activity, causes reddening, cracking, and peeling of the skin and a burn may occur if contact continues for several minutes. Splashes in the eye cause immediate and severe irritation. Acute inhalation may result in irritation of the mucous membranes, blurred vision, nausea and vomiting, headache, delirium, hallucinations, coma, tremor, convulsions, and cardiac and respiratory arrest.

Chronic exposure

There is an increased incidence of cardiovascular disease among workers exposed to carbon disulphide. In addition, sleep disturbances, fatique, anorexia, and weight loss are common complaints. Intellectual decline, depression, stereotyped behaviour, ocular changes, cerebellar and extrapyramidal signs, hepatic damage, and permanent impairment of reproductive performance have been described.

Treatment

Treatment involves removal from exposure, washing contaminated skin, irrigation of the eyes with water, and supportive measures. In the majority of cases, however, preventive measures to keep carbon disulphide concentrations in the workplace as low as possible are more important.

Further reading

Spyker DA, Gallanosa AG, Suratt PM (1982). Health effects of acute carbon disulfide exposure. *Journal of Toxicology—Clinical Toxicology* **19**, 87–93.

Carbon monoxide

Carbon monoxide is a tasteless, odourless, colourless, non-irritating gas produced by incomplete combustion of organic materials. Normal endogenous carbon monoxide production is sufficient to maintain a resting carboxyhaemoglobin level of 1 to 3 per cent in urban non-smokers and 5 to 6 per cent in smokers.

Common sources of carbon monoxide are car exhaust fumes (in the absence of a catalytic converter), improperly maintained and vented heating systems, and smoke from all types of fire. Carbon monoxide derived from domestic heating systems is a major cause of accidental death in the developing world. Inhalation of methylene chloride (found in paint strippers) may also lead to carbon monoxide poisoning.

Mechanisms of toxicity

Symptoms and signs that follow inhalation of carbon monoxide are the result of tissue hypoxia. The affinity of haemoglobin for carbon monoxide is approximately 240 times greater than that for oxygen. Carbon monoxide combines with haemoglobin to form carboxyhaemoglobin, reducing the total oxygen-carrying capacity of the blood. In addition, the oxygen dissociation curve shifts to the left and modifies oxygen-binding sites. As a result, the affinity of the remaining haem groups for oxygen is increased, the oxygen dissociation curve is distorted as well as being shifted and the resulting tissue hypoxia is thus far greater than that which would result from simple loss of oxygen-carrying capacity.

Carbon monoxide may also inhibit cellular respiration as a result of reversible binding to cytochrome oxidase. Brain lipid peroxidation mediated by carbon monoxide may play a role in the development of delayed neuropsychiatric sequelae.

Clinical features

The clinical features of carbon monoxide poisoning are summarized Table 3.

Acute exposure

The symptoms of mild to moderate exposure to carbon monoxide are non-specific and may even be mistaken for a viral illness and for this reason it is important that the diagnosis is always borne in mind. Elderly patients and those with pre-existing cardiorespiratory disease are at greater risk. A carboxyhaemoglobin concentration of less than 10 per cent is not normally associated with symptoms and 10 to 30 per cent carboxyhaemoglobin may result only in headache and mild exertional dyspnoea. Even low concentrations of carbon monoxide produce significant effects on cardiac function during exercise in subjects with coronary artery disease. Coma, convulsions, and cardiorespiratory arrest may be expected with carboxyhaemoglobin concentrations in excess of 60 per cent.

Delayed effects

Neuropsychiatric problems after recovery from carbon monoxide intoxication may develop insidiously over a number of weeks. They include intellectual deterioration, memory impairment, features of cerebral, cerebellar, and midbrain damage, parkinsonism, akinetic mutism, irritability, verbal aggressiveness, violence, impulsiveness, and moodiness.

Treatment

The patient should be removed from exposure and 100 per cent oxygen administered using a tightly fitting facemask. Endotracheal intubation and mechanical ventilation may be required in those who are unconscious. The administration of oxygen should be continued until the carboxyhaemoglobin concentration is less than 10 per cent.

Controlled studies of hyperbaric oxygen have not shown greater benefit than 100 per cent normobaric oxygen with elective ventilation.

General symptomatic and supportive measures will be required. Diazepam (5 to 10 mg intravenously) repeated as necessary is the agent of choice for the management of convulsions. The benefit of corticosteroids for the treatment of cerebral oedema has not been proved but mannitol may be useful.

Further reading

International Programme on Chemical Safety (1999). *Environmental Health Criteria 213. Carbon monoxide*, 2nd edn. World Health Organization, Geneva.

Thom SR *et al.* (1995). Delayed neuropsychologic sequelae after carbon monoxide poisoning; prevention by treatment with hyperbaric oxygen. *Annals of Emergency Medicine* **25**, 474–80.

Carbon tetrachloride (tetrachloromethane)

Carbon tetrachloride was once widely used as a dry-cleaning chemical, degreasing agent, and fire extinguisher but international regulations have now restricted it to laboratory and industrial usage and it is no longer manufactured in most developed countries.

Clinical features

Acute exposure

The immediate effects include nausea, vomiting, abdominal pain, and diarrhoea. High concentrations cause dizziness, confusion, coma, respiratory depression, hypotension, and occasionally convulsions. Death may follow from respiratory failure or ventricular fibrillation due to cardiac sensitization to circulating catecholamines. Hepatorenal damage supervenes after a delay of up to 2 weeks. Hepatic enzyme activities increase before jaundice and a tender swollen liver develop. Maximal liver damage probably occurs within 48 h of an acute exposure and may progress to fulminant hepatic failure. Acute renal tubular necrosis is common and may develop in the absence of hepatic dysfunction 1 to 7 days after exposure. Rarely, cerebellar dysfunction, cerebral haemorrhage, optic atrophy, and parkinsonism may occur. Alcohol and previous liver damage render the individual more susceptible.

A plain film of the abdomen may confirm ingestion has occurred (Fig. 1).

Fig. 1 Plain abdominal radiograph in a patient who ingested carbon tetrachloride.

Chronic exposure

Repeated exposure to low concentrations of carbon tetrachloride may also cause hepatic and renal damage. Hepatic cirrhosis and hepatoma may develop. Prolonged carbon tetrachloride exposure is associated with polyneuropathy, various visual disturbances, and anaemia, including fatal aplastic anaemia.

Treatment

After ingestion, gastric emptying is probably best avoided because of the risk of aspiration. If the patient presents within 12 h of expo*N*-acetylcysteine should be given as for paracetamol overdose (see below). Renal and liver failure should be managed conventionally.

Further reading

International Programme on Chemical Safety (1999). *Environmental Health Criteria 208. Carbon tetrachloride.* World Health Organization, Geneva.

Chlorates

Clinical features

Sodium chlorate and potassium chlorate are powerful oxidizing agents and are highly toxic if ingested. The early features include nausea, vomiting, diarrhoea, abdominal pain, and cyanosis secondary to methaemoglobinaemia. Intravascular haemolysis occurs causing hyperkalaemia, jaundice, and oliguric renal failure.

Treatment

Gastric lavage should be considered if the patient presents within 1 h of ingestion. Methaemoglobinaemia can be corrected by slow intravenous injection of methylthioninium chloride (methylene blue) 2 mg/kg body weight, although antidotal efficacy is reduced in the presence of haemolysis. Blood transfusion may be required. Plasma potassium concentrations should be monitored and reduced if necessary. Haemodialysis will remove chlorate and may also be required for the management of renal failure and hyperkalaemia. Plasmapheresis has also been employed since it will remove chlorate, circulating free haemoglobin, and red cell stroma and thus help to prevent the development of renal failure.

Chlorine

Chlorine is a greenish-yellow gas normally transported as a pressurized liquid. Exposure after spillage may be prolonged because gaseous chlorine is heavier than air, causing it to remain near ground level. Chlorine has a pungent odour that can usually be detected by smell at concentrations of less than 0.5 ppm.

Molecular chlorine, a strong oxidizing agent, is known to react with many functional groups in cell components, forms chloramines, oxidizes thiol radicals, reacts with tissue water to form hypochlorite and hydrochloric acid, and it may generate oxygen free radicals.

Clinical features

Symptoms begin within minutes and include irritation of the mucous membranes of the eyes, nose, and throat, followed by cough, breathlessness, expectoration of white sputum (which may be bloodstained), chest pain and tightness, abdominal pain, nausea, headache, dizziness, and palpitation due to ventricular ectopic beats. Laryngeal oedema may cause hoarseness of the voice and stridor, and cardiac arrest may occur secondary to hypoxia.

Restrictive as well as obstructive ventilatory defects arise in those who have inhaled sublethal amounts. Diffusion is impaired, leading to arterial hypoxaemia. In very severe cases, non-cardiogenic pulmonary oedema and respiratory failure may develop. Survival is usually followed by complete resolution of the pulmonary defects.

Some workers chronically exposed to the gas become anosmic.

Treatment

The first priority is to remove the casualty from exposure. Conjunctival skin burns should be treated as for acids (see above).

Patients with respiratory symptoms persisting beyond the period of exposure should be admitted to hospital in case they require bronchodilators and humidified oxygen. Some will require mechanical ventilation, particularly if non-cardiogenic pulmonary oedema develops. Frusemide has been reported to be of value. Corticosteroids and prophylactic antibiotics have not been shown to be of value. Correction of serious metabolic acidosis with intravenous sodium bicarbonate may be necessary.

Further reading

Mvros R, Dean BS, Krenzelok EP (1993). Home exposures to chlorine/chloramine gas: review of 216 cas*Southern Medical Journa* **86**, 654–7.

Schonhofer B, Voshaar T, Kohler D (1996). Long-term sequelae following accidental chlorine gas exposi*Respiratior* **63**, 155–9.

Chlorofluorocarbons (CFCs)

CFCs are derived by the partial or complete substitution of the hydrogen atoms in methane and ethane with chlorine and fluorine atoms. CFCs were developed as refrigerants some 60 years ago but have been used as propellants in aerosols, as blowing agents in foam insulation products, and as intermediates for plastics. The aerosol propellant market, which previously consumed half of the total production, is currently a minor application due to international restrictions imposed as a result of concerns that CFCs damage the ozone layer of the earth's atmosphere.

Clinical features

Acute exposure

Inhalation of CFCs may result in a tingling sensation, humming in the ears, apprehension, slurred speech, and decreased performance in psychological tests. Exposure to high concentrations may result in clinically significant arrhythmias, coma, and respiratory depression; fatalities have occurred particularly after CFC abuse (see volatile substance abuse).

Chronic exposure

A sensorimotor neuropathy developed in a laundry worker exposed for several years both to CFC 113 and tetrachloroethylene and in a refrigerator repair worker. However, epidemiological studies in exposed workers have not found evidence of a causal relationship.

Further reading

International Programme on Chemical Safety (199C*Environmental Health Criteria 113. Fully halogenated chlorofluorocarbo* World Health Organization, Geneva.

International Programme on Chemical Safety (1991*Environmental Health Criteria 126. Partially halogenated chlorofluorocarbons (Methane derivativ* World Health Organization, Geneva.

International Programme on Chemical Safety (1992*Environmental Health Criteria 139. Partially halogenated chlorofluorocarbons (Ethane derivative* World Health Organization, Geneva.

Chlorophenoxy herbicides

The chlorophenoxy herbicides (Table 4) include the substances popularly referred to as 'hormone' weedkillers and are used widely in agriculture and by the public. Most instances of serious poisoning have been due to deliberate ingestion but few cases have been reported. These herbicides are often coformulated with the chemically related herbicide, dicamba, which is of low toxicity, and ioxynil

and bromoxynil, which uncouple oxidative phosphorylation.

Clinical features

Ingestion causes burning in the mouth and throat, nausea, vomiting, and abdominal pain. The face may be flushed and there is often profuse sweating and fever. CNS depression leading to deep, prolonged coma, hyperventilation, metabolic acidosis, and pulmonary oedema may develop. ECG abnormalities and skeletal muscle damage leading to proximal myopathy have been reported.

Treatment

Gastric lavage should be considered if the patient presents within 1 h of overdose. Supportive measures should be employed as necessary. Alkalinization of the urine (see above) is indicated for severe poisoning since it considerably enhances elimination of 2,4-dichlorophenoxy acetic acid (2,4-D) and, to a lesser extent, mecoprop, particularly if combined with a high urine flow (>600 ml/h). Alkalinization of the urine probably does not have a beneficial effect on the elimination of other phenoxyacetates and does not affect that of ioxynil.

Further reading

Bradberry SM *et al.* (2000). Mechanisms of toxicity, clinical features and management of acute chlorophenoxy herbicide poisoning: A rev *Journal of Toxicology—Clinical Toxicolog* **38**, 111–22.

Chloroquine

Chloroquine overdose is probably the most common form of self-poisoning with drugs in Africa, the Far East, and West Pacific and is a growing problem in Europe.

Clinical features

Toxicity can result from doses greater than 1 g (about six tablets) in adults. Cardiac arrest is commonly the first clinical manifestation of poisoning, but hypotension usually precedes it and may progress to cardiogenic shock and pulmonary oedema. Electrocardiographic abnormalities, bradyarrhythmias, and tachyarrhythmias are common and are similar to those of quinine (see below). Visual disturbance, agitation, drowsiness, acute psychosis, dystonic reactions, seizures, and coma may ensue. Hypokalaemia is common and is due to potassium channel blockade.

Treatment

Gastric lavage or activated charcoal (50 to 100 g) should be considered if the patient presents within 1 h. Supportive measures should be employed and hypokalaemia corrected. There is no specific antidote. There is evidence that mechanical ventilation, the administration of adrenaline (epinephrine) (0.25 µg/kg. min) and high-doses of diazepam (1 mg/kg as a loading dose and 0.25 to 0.4 mg/kg.h maintenance) may reduce the mortality to 10 per cent in severe poisoning. Multiple-dose activated charcoal may enhance chloroquine elimination.

Further reading

Clemessy JL *et al.* (1996). Treatment of acute chloroquine poisoning: a 5-year experien *Critical Care Medicin* **24**, 1189–95.

McKenzie AG (1996). Intensive therapy for chloroquine poisoning—a review of 29 cas *South African Medical Journa* **86**, 597–9.

Chromium

There is no good evidence that chromium(II), chromium(III) and chromium(IV) compounds are dangerous but chromium(III) compounds have produced skin sensitization. Chromium(VI) is the most important toxicologically because it can cross cell membranes readily. In contrast, chromium(III) compounds are confined to the extracellular space.

Chromium is absorbed mainly by inhalation and, to a lesser extent, via the skin or gastrointestinal tract. Hexavalent chromium compounds are generally better absorbed than trivalent chromium compounds and, understandably, soluble chromium compounds such as sodium (VI) chromate are absorbed more readily than insoluble compounds such as chromium(III) oxide. Chromium is excreted via the kidney.

Clinical features

Acute poisoning

Inhaled soluble chromium(VI) compounds, such as sodium and potassium chromate and dichromate, are highly irritant to mucous membranes and may lead to inflammation of the nasal mucosa. Inhalation of chromium(VI) trioxide (chromic acid) causes cough, headache, chest pain, dyspnoea, and cyanosis.

Ingestion of highly water-soluble chromium(VI) compounds leads within minutes to nausea, vomiting, abdominal pain, diarrhoea, and a burning sensation in the mouth, throat, and stomach; gastrointestinal haemorrhage is a frequent complication.

Methaemoglobinaemia, haemolysis, and disseminated intravascular coagulation, and renal and hepatic failure have been reported.

Chromic acid splashes produce severe burns. Percutaneous absorption may lead to kidney and liver failure; fatalities have occurred.

Chronic poisoning

'Chrome ulcers' may develop after repeated topical exposure to chromium(VI) compounds. They have a well-defined circular margin with raised edges and a central cavity that may penetrate to bone and is filled with exudate or a tenacious crust. Chromium(VI) compounds are also skin sensitizers and contribute to the development of cement dermatitis and contact dermatitis from paint primer, tanned leather, tattoo pigments, and matches.

Inhalation of chromium(VI) compounds has led to atrophy, ulceration, and perforation of the nasal septum. Pharyngeal and laryngeal ulcers may also occur. Asthma may be precipitated by exposure to fumes. Lung fibrosis, bronchitis, emphysema, and proximal tubular damage result from occupational exposure. There is an increased risk of lung cancer.

Treatment

Ascorbic acid reduces chromium(VI) to the less toxic chromium(III). Topical 10 per cent ascorbic acid, as an ointment or in solution, has led to dramatic resolution of occupational chromium dermatitis but there is no clinical evidence that the systemic administration of ascorbic acid, or any other reducing agent, lessens morbidity or mortality in severe chromium poisoning. Topical preparations containing sodium calcium edetate may also afford some protection to the skin but there is no evidence that systemic chelation treatment is beneficial in chromium poisoning. Haemodialysis effectively removes chromium from the blood but the high tissue uptake limits the value of this treatment when used alone.

Further reading

Barceloux DG (1999). Chromium *Journal of Toxicology—Clinical Toxicolog* **37,** 173–94.

Bradberry SM, Vale JA (1999). Therapeutic review: is ascorbic acid of value in chromium poisoning and chromium derma *Journal of Toxicology—Clinical Toxicolog* **37,** 195–200.

Clomethiazole (chlormethiazole)

Clinical features

This hypnotic drug taken in overdose may cause coma, respiratory depression, reduced muscle tone, hypotension, and excessive salivation. The characteristic odour of clomethiazole is often detected on the breath.

Treatment

Treatment is supportive.

Clonidine

Clonidine exerts its hypotensive action by reduction of sympathetic tone mediated by a central effect on postsynaptic a_2-adrenoceptors in the medulla. Clonidine decreases heart rate, cardiac output, and total peripheral resistance. In the presence of high plasma clonidine concentrations, peripheral a_2-agonist activity predominates and accounts for those instances of vasoconstriction and hypertension reported following clonidine overdose.

Clinical features

Poisoning may be severe and life-threatening, particularly in children. Hypertension and severe vasoconstriction are unusual while bradycardia, hypotension, coma, and respiratory depression are common. Toxic effects last about 16 h, but may extend to several days in severe overdose.

Treatment

Gastric lavage should be considered or 50 to 100 g of activated charcoal administered if a patient presents within 1 h of a substantial overdose. Bradycardia is usually reversed by atropine in a dose of 0.6 to 2.4 mg intravenously. The use of a-adrenergic blocking drugs (tolazoline or phentolamine) has been advocated in severely poisoned patients but their action may be unpredictable. Severe hypotension should be treated with a plasma expander and then, if necessary, an inotropic agent such as, dobutamine 2.5 to 10 µg/kg.min may be given by intravenous infusion. The use of naloxone has been advocated but its benefit is inconsistent and it may produce hypertension. Sodium nitroprusside 0.5 to 8.0 µg/kg.min by intravenous infusion is the most effective agent for management of severe hypertension and peripheral vasoconstriction. Although forced diuresis has been employed in the treatment of clonidine poisoning, renal elimination is not increased.

Further reading

Erickson SJ, Duncan A (1998). Clonidine poisoning—an emerging problem: epidemiology, clinical features, management and preventative strategies. *Journal of Paediatrics and Child Health* **34,** 280–2.

Nichols MH, King WD, James LP (1997). Clonidine poisoning in Jefferson County, Alabama. *Annals of Emergency Medicine* **29**, 511–17.

Cobalt

Cobalt is a relatively rare element and usually exists in association with nickel, silver, lead, copper, and iron ores. It is used in steel alloys, in the manufacture of magnets, and in the hard metal industry as a binder for tungsten carbide. It is also an essential dietary trace element available as a component of vitamin B_{12} (cyanocobalamin).

Cobalt can be absorbed orally and by inhalation. Most absorbed cobalt is excreted within days but a small proportion is retained with a biological half-life of approximately 2 years. The normal body burden of cobalt is about 1.1 mg.

Clinical features

Acute poisoning

Cobalt salts are relatively non-toxic but their ingestion may lead to gastrointestinal disturbance.

Chronic poisoning

Occupational exposure to cobalt dust occurs mainly in the tungsten carbide industry and causes 'hard metal' pneumoconiosis with interstitial fibrosis. This usually develops after several years of exposure to high concentrations of cobalt and may prove fatal. There is also a higher incidence of bronchitis and emphysema amongst cobalt workers and occupational asthma has been reported.

Chronic occupational exposure also leads to anosmia, auditory nerve damage, visual disturbance, irritability, headache, memory deficit, weakness, peripheral neuropathy, gastrointestinal disturbance, and weight loss. There is no firm evidence that cobalt is carcinogenic and assessment of its cancer risk is often confounded by a simultaneous exposure to nickel and arsenic.

Chronic ingestion of cobalt causes polycythaemia, inhibits the iodination of tyrosine (and therefore can cause goitre), and leads to cardiomegaly, congestive cardiomyopathy, pericardial effusion, and hypertrichosis, most of which are reversible when exposure is discontinued.

Simultaneous allergies to nickel and to cobalt are frequent and there is some evidence for a mutual enhancing effect of contact sensitization to one metal in the presence of the other.

Treatment

If the patient presents early after ingestion of a cobalt salt, gastric lavage should be considered. In two studies, DMSA (succimer) significantly reduced mortality in mice poisoned with cobalt chloride but in another study, DTPA (calcium trisodium pentetate) was more effective than DMSA. No satisfactory human studies have yet been performed.

Further reading

Mucklow ES *et al.* (1990). Cobalt poisoning in a 6-year-old. *Lancet* **335**, 981.

Cocaine

In recent years there has been a considerable increase in the recreational use of cocaine. Cocaine is a powerful local anaesthetic and vasoconstrictor and may be abused by smoking, ingestion, injection, or by 'snorting' it intranasally. Users, body packers, and those who swallow the drug to avoid being found in possession of it ('stuffers') are at risk of overdose. 'Street' cocaine (cocaine hydrochloride) is sometimes dissolved in an alkaline solution from which the cocaine is extracted into ether or other solvent that is then evaporated to leave crystals of relatively pure ('freebase') cocaine. 'Crack' is another type of freebase cocaine made by heating cocaine hydrochloride with baking soda. The hard paste which crackles when smoked is also used widely. Other drugs such as ethanol, cannabis, and conventional hypnotics and sedatives are frequently taken with cocaine to reduce the intensity of its less pleasant effects.

Clinical features

The features of cocaine overdosage are similar to those of amfetamine. In addition to euphoria, it also has sympathomimetic effects including agitation, tachycardia, hypertension, sweating, and hallucinations. Prolonged convulsions with metabolic acidosis, hyperthermia, rhabdomyolysis, ventricular arrhythmias, and cardiorespiratory arrest may follow in the most severe cases. Less common features include dissection of the aorta, myocarditis, myocardial infarction, dilated cardiomyopathy, subarachnoid haemorrhage, cerebral haemorrhage, and cerebral vasculitis.

A number of rare complications of the method of use of cocaine have been reported. These include pulmonary oedema after intravenous injection of freebase cocaine and pneumomediastinum and pneumothorax after sniffing it. In addition, chronic 'snorting' has caused perforation of the nasal septum, cerebrospinal fluid rhinorrhoea due to thinning of the cribriform plate, and pulmonary granulomatosis.

Treatment

Users who are intoxicated may require sedation with diazepam to control agitation or convulsions. Measures to prevent further absorption are usually irrelevant. Active external cooling is required when body temperature exceeds 41°C. Myocardial ischaemia is best treated with intravenous glyceryl nitrite or a calcium channel blocker. Hypertension and tachycardia usually respond to sedation and cooling. b-Adrenoceptor blocking drugs are absolutely contraindicated because of the risk of precipitating paradoxical hypertension. Phentolamine in a dose of 2 to 5 mg intravenously or other vasodilator can be employed, if necessary. Accelerated idioventricular rhythm should not normally require treatment but ventricular fibrillation and asystole should be managed conventionally.

Further reading

Hatsukami DK, Fischman MW (1996). Crack cocaine and cocaine hydrochloride—are the differences myth or reality? *Journal of the American Medical Association* **276**, 1580–8.

Hollander JE (1996). Cocaine-associated myocardial infarction. *Journal of the Royal Society of Medicine* **89,** 443–7.

Kloner RA *et al.* (1992). The effects of acute and chronic cocaine use on the heart. *Circulation* **85,** 407–19.

Marzuk PM *et al.* (1995). Fatal injuries after cocaine use as a leading cause of death among young adults in New York city. *New England Journal of Medicine* **332**, 1753–7.

Rubin RB, Neugarten J (1992). Medical complications of cocaine: changes in pattern of use and spectrum of complications. *Journal of Toxicology—Clinical Toxicology* **30,** 1–12.

Co-phenotrope (Lomotil)

Co-phenotrope is a mixture of an opioid, diphenoxylate hydrochloride, and atropine.

Mechanism of toxicity

Gastric emptying is delayed and intestinal motility reduced. The onset of toxicity following an overdose may be delayed for up to 12 h.

Clinical features

Respiratory depression is the major complication of diphenoxylate poisoning. Vomiting, abdominal pain, drowsiness, and coma also occur. Even though co-phenotrope tablets incorporate only a small amount of atropine this is often toxic to children under 5 years and several deaths have been reported. Anticholinergic features are to be expected (see tricyclic antidepressants below).

Treatment

Repeated doses of naloxone may be necessary to reverse respiratory depression because of the long duration of action of diphenoxylate (see opiates and opioids below). Lavage may be appropriate in an adult presenting within 1 h of a substantial overdose before toxicity develops; activated charcoal (50 to 100 g) may also reduce absorption significantly if administered within that period.

Further reading

McCarron MM, Challoner KR, Thompson GA (1991). Diphenoxylate-atropine (Lomotil) overdose in children: An update (report of eight cases and review of the literature). *Pediatrics* **87,** 694–700.

Copper

Copper is used for pipes and roofing material, in alloys, and as a pigment. It is a component of several enzymes, including tyrosinase and cytochrome oxidase, and is essential for the utilization of iron. Copper sulphate is used as a fungicide, an algicide, and in some fertilizers.

Approximately one-third of an ingested copper salt is absorbed and in the blood 80 per cent is bound to caeruloplasmin. Most absorbed copper is deposited in the liver and eliminated mainly in bile.

Clinical features

Acute poisoning

Acute copper poisoning usually results from the ingestion of contaminated foods or from accidental or deliberate ingestion of copper salts. Following a substantial ingestion of a copper salt there is profuse vomiting with abdominal pain, diarrhoea, headache, dizziness, and a metallic taste. Gastrointestinal haemorrhage, haemolysis, and hepatorenal failure may ensue and fatalities have occurred. Body secretions may have a green or blue discoloration.

Occupational exposure to copper fumes (during refining or welding) or to copper-containing dust causes 'metal-fume fever' with upper respiratory tract symptoms, headache, fever, and myalgia.

Chronic poisoning

Chronic occupational copper poisoning causes general malaise, anorexia, nausea, vomiting, and hepatomegaly. Contact dermatitis, pulmonary granulomas, and pulmonary fibrosis have also been described. There is no convincing evidence that copper is carcinogenic in humans.

Treatment

Although vomiting occurs invariably following the ingestion of many copper salts, gastric lavage may be of value in reducing copper absorption if presentation is early. Blood copper levels correlate well with severity of intoxication, a concentration of less than 3 mg/l indicating mild to moderate poisoning and a concentration in excess of 8 mg/l severe intoxication. D-Penicillamine 25 mg/kg body weight daily until recovery enhances copper chelation in both acute and chronic copper poisoning. There is now animal evidence to suggest that *N*-acetylcysteine and DMPS (unithiol) are of similar efficacy.

Further reading

Barceloux DG (1999). Copper. *Journal of Toxicology—Clinical Toxicology* **37**, 217–30.

International Programme on Chemical Safety (1998). *Environmental Health Criteria 200. Copper*. World Health Organization, Geneva.

Cyanide

Hydrogen cyanide and its derivatives are used widely in industry and are released during the thermal decomposition of polyurethane foams. Cyanide poisoning may also result from the ingestion of the cyanogenic glycoside, amygdalin (vitamin B_{17}), which is found in the kernels of almonds, apples, apricots, cherries, peaches, plums, and other fruits.

Mechanisms of toxicity

Cyanide reversibly inhibits cellular enzymes which contain ferric iron, notably cytochrome oxidase a_3, so that electron transfer is blocked, the tricarboxylic acid cycle is paralysed, and cellular respiration ceases.

Clinical features

Acute exposure

The ingestion by an adult of 50 ml of (liquid) hydrogen cyanide or 200 to 300 mg of one of its salts is likely to prove fatal. Inhalation of hydrogen cyanide gas may produce symptoms within seconds and death within minutes.

Acute poisoning is characterized by dizziness, headache, palpitation, anxiety, a feeling of constriction in the chest, dyspnoea, pulmonary oedema, confusion, vertigo, ataxia, coma, and paralysis. Cardiovascular collapse, respiratory arrest, convulsions, and metabolic acidosis are seen in severe cases. Cyanosis may occur, and the

classic 'brick-red' colour of the skin is noted occasionally. There is sometimes an odour of bitter almonds on the breath, but the ability to detect it is genetically determined and some 40 per cent of the population are unable to do so.

Chronic exposure

Chronic exposure results predominantly in neurological damage that can include ataxia, peripheral neuropathies, amblyopia, optic atrophy, and nerve deafness.

Treatment

Cyanide poisoning is a medical emergency, although specific antidotal treatment may not always be necessary. Where appropriate, the patient should be removed from the source of exposure, contaminated clothing discarded, and the skin washed with soap and water. Gastric lavage should be considered if a cyanide salt has been ingested less than 1 h previously, but this procedure must not delay treatment if symptoms or signs of toxicity are present. It may be difficult to differentiate between the genuine fear and anxiety of a patient and the early symptoms of cyanide poisoning. However, a patient who has been exposed to hydrogen cyanide gas and who is conscious 30 min later is unlikely to require antidotal therapy.

Oxygen

The administration of oxygen is of paramount importance in the treatment of cyanide poisoning. It is believed to prevent inhibition of cytochrome oxidase a_3 and to accelerate its reactivation.

Dicobalt edetate

Cobalt compounds form stable inert complexes with cyanide. Dicobalt edetate (Kelocyanor), if available, is the treatment of choice for confirmed cyanide poisoning and should be given intravenously in a dose of 300 to 600 mg over 1 min, with a further 300 mg if recovery does not occur within 1 min. It should be administered only if the diagnosis is certain because, in the absence of cyanide, Kelocyanor may cause serious side-effects including vomiting, tachycardia, hypertension, chest pain, and facial and palpebral oedema as it contains free cobalt, which is responsible in part for its efficacy.

Sodium thiosulphate

Cyanide is detoxified by conversion to thiocyanate. Thiosulphate is required for this reaction. Sodium thiosulphate 12.5 g (25 ml of a 50 per cent solution) should be given by intravenous injection over 10 min. Experimental studies have shown that the coadministration of sodium nitrite enhances the antidotal benefit of sodium thiosulphate.

Sodium nitrite, 4-dimethylaminophenol (4-DMAP)

Another means of inactivating cyanide is to convert a portion of the body's haemoglobin to methaemoglobin, which binds cyanide. Although the affinity of cyanide for methaemoglobin is less than that of cytochrome oxidase, the presence of a large circulating methaemoglobin pool diminishes cyanide toxicity by binding cyanide ion before tissue penetration occurs. Methaemoglobinaemia may be induced by the administration of either sodium nitrite or 4-dimethylaminophenol (**4-DMAP**). 4-DMAP may produce unexpectedly high methaemoglobin concentrations and cause acute tubular necrosis and Heinz-body haemolytic anaemia. Nitrites may also mitigate cyanide toxicity by virtue of their vasodilator actions and improvement of tissue perfusion. Sodium nitrite 300 mg (10 ml of a 3 per cent solution) should be administered by intravenous injection over 3 min.

Inhalation of amyl nitrite was recommended in the past but it produces only low circulating concentrations of methaemoglobin.

Hydroxocobalamin

One mole of hydroxocobalamin inactives one mole of cyanide but, on a weight-for-weight basis, 50 times more hydroxocobalamin is needed than cyanide because hydroxocobalamin is a far larger molecule. Con-centrated formulations of hydoxocobalamin are not yet available in all countries. If available, give hydroxocobalamin in a dose of 5 g intravenously over 30 min. A second dose (5 g) may be required in severe cases of cyanide poisoning.

Conclusion

If dicobalt edetate or hydroxocobalamin are not available, a combination of sodium nitrite and sodium thiosulphate should be administered..

Further reading

Mueller M, Borland C (1997). Delayed cyanide poisoning following acetonitrile poisoning. *Postgraduate Medical Journal* **73**, 299–300.

Rosenow F *et al.* (1995). Neurological sequelae of cyanide intoxication—the patterns of clinical, magnetic resonance imaging and positron emission tomography findings. *Annals of Neurology* **38**, 825–8.

Yen D *et al.* (1995). The clinical experience of acute cyanide poisoning. *American Journal of Emergency Medicine* **13**, 524–8.

Dapsone

Dapsone is available formulated alone or in combination with pyrimethamine (as Maloprim).

Clinical features

Dapsone poisoning can be severe and result not only in methaemoglobinaemia but also in haemolysis, jaundice, drowsiness, coma, seizures, and metabolic acidosis.

Treatment

If presentation after overdose is within 1 h, gastric lavage should be considered or, alternatively, 50 to 100 g of activated charcoal may be administered. Administration of repeated doses of activated charcoal seems to have comparable efficacy to haemodialysis in increasing dapsone elimination. Methylthioninium chloride (methylene blue) at 1 to 2 mg/kg should be given intravenously over 5 min for severe methaemoglobinaemia.

Further reading

Ferguson AJ, Lavery GG (1997). Deliberate self-poisoning with dapsone—a case report and summary of relevant pharmacology and treatment. *Anaesthesia* **52**, 359–63.

Diethylene glycol

Diethylene glycol is used mainly in polyester resins and polyols, as a humectant in the tobacco industry, and as a solvent. It achieved notoriety in 1985 when it was discovered that for some years it had been added to some wines. Several pharmaceutical errors have also led to fatalities.

Mechanism of toxicity

Animal studies suggest that diethylene glycol is first oxidized by alcohol dehydrogenase to 2-hydroxyethoxyacetaldehyde and then to 2-hydroxyethoxyacetic acid.

Clinical features

Nausea, vomiting, and abdominal pain occur frequently and are followed by the development of jaundice and hepatomegaly, pulmonary oedema, metabolic acidosis, coma, and renal failure in most cases.

Treatment

Supportive measures to treat dehydration and to correct metabolic acidosis should be instituted promptly. Ethanol or fomepizole (4-methylpyrazole) should be administered to block diethylene glycol metabolism and dialysis should be employed if renal failure supervenes. A loading dose of 50 g of ethanol orally (conveniently given as 125 ml of gin, whisky, or vodka) should be administered followed by an intravenous infusion of 10 to 12 g ethanol/h to produce a blood ethanol concentration of 500 mg to 1 g/l. The infusion should be continued until diethylene glycol is no longer detectable in the blood. If dialysis is employed, the rate of ethanol administration will need to be increased to 17 to 22 g/h. The regimen for fomepizole is given in the section on ethylene glycol below.

Further reading

O'Brien KL *et al.* (1998). Epidemic of pediatric deaths from acute renal failure caused by diethylene glycol poisoning. *Journal of the American Medical Association* **279,** 1175–80.

Woolf AD (1998). The Haitian diethylene glycol poisoning tragedy—A dark wood revisited. *Journal of the American Medical Association* **279,** 1215–16.

Digoxin and digitoxin

Toxicity occurring during chronic administration of these cardiac glycosides is common. In contrast, acute poisoning from digoxin and digitoxin is infrequent, though the mortality may be as high as 20 per cent after a substantial overdose, particularly if digoxin-specific antibody fragments are not employed.

Clinical features

Nausea, vomiting, dizziness, anorexia, and drowsiness are common. Confusion, diarrhoea, visual disturbances, and hallucinations may also occur. Sinus bradycardia, often marked, is the earliest cardiotoxic effect and may be followed by supraventicular arrhythmias with or without heart block, ventricular premature beats, and ventricular tachycardia. Hyperkalaemia occurs due to inhibition of the Na^{+}-K^{+} ATPase pump. The diagnosis may be confirmed by measurement of the serum digoxin concentration.

Treatment

Gastric lavage should be considered in patients with a history of a substantial overdose less than 1 h previously. Alternatively, 50 to 100 g of activated charcoal may be administered to reduce absorption and repeated doses will also enhance elimination. Potassium supplements should not be given until the serum potassium concentration is known, as severe poisoning is commonly associated with hyperkalaemia that should be treated conventionally.

Sinus bradycardia, ventricular ectopics, atrioventricular block, and sinoatrial standstill or block are often reduced or abolished by atropine in a dose of 1.2 to 2.4 mg. Ventricular ectopics alone should not be treated unless cardiac output is impaired. Ventricular tachydysrhythmias may be treated with intravenous lignocaine, atenolol, phenytoin, or amiodarone; if clinically significant and persistent, digoxin-specific antibody fragments should be considered. Failure to achieve a satisfactory cardiac output by drug therapy in patients with bradycardia, atrioventricular block, or sinus arrest is an indication for insertion of a right ventricular pacing wire or, if available, the administration of digoxin-specific antibody fragments (6 to 8 mg/kg body weight is sufficient in most cases of poisoning. In very severe cases consult the product literature to calculate the optimal dose). An improvement in the patient's condition should occur within 20 to 40 min.

Forced diuresis, peritoneal dialysis, haemodialysis, and haemoperfusion do not significantly increase the elimination of the drug.

Further reading

Kinlay S, Buckley NA (1995). Magnesium sulfate in the treatment of ventricular arrhythmias due to digoxin toxicity. *Journal of Toxicology—Clinical Toxicology* **33**, 55–9.

Williamson KM *et al.* (1998). Digoxin toxicity: an evaluation in current clinical practice. *Archives of Internal Medicine* **158,** 2444–9.

Dishwashing liquids, fabric conditioners, and household detergents

Most of these products including carpet shampoo, dishwashing rinse aid for dishwashing machines, fabric washing powder and flakes, scouring liquids, creams, and powders contain surfactants that have both hydrophilic and lipophilic groups to allow fat-soluble substances to be dispersed in aqueous media.

There are three types of surfactants of differing toxicity: anionic surfactants, which have a negative electrical charge on the lipophilic groups; cationic surfactants, which have a positive charge; and non-ionic surfactants that have no charge.

Clinical features

Anionic detergents irritate the skin by removing natural oils and cause redness, soreness, and even a papular dermatitis. Ingestion may cause mild gastrointestinal irritation, nausea, vomiting, and diarrhoea. Non-ionic surfactants irritate the skin only slightly and appear to be completely harmless when ingested. Cationic surfactants (e.g. quarternary ammonium compounds) are much more toxic than the others but are rarely found in household cleaning materials.

Treatment

After ingestion of products containing either a non-ionic or anionic surfactant, liberal amounts of water or milk should be administered.

Further reading

Cornish LS, Parsons BJ, Dobbin MD (1996). Automatic dishwasher detergent poisoning: opportunities for prevention. *Australian and New Zealand Journal of Public Health* **20**, 278–83.

Disulfiram (Antabuse)

Mechanism of toxicity

Disulfiram and its main metabolite diethyldithiocarbamate inhibit the activity of a wide range of enzymes, particularly aldehyde dehydrogenase. Carbon disulphide, another metabolite, may account for some of the side-effects observed during disulfiram therapy.

Clinical features

Adult cases are likely to be alcoholics who have been taking disulfiram before the overdose and to be malnourished, factors that may explain the frequency of neuropsychiatric features. Sensorimotor neuropathy, flaccid tetraparesis, and encephalopathy have been described after overdose though these features may have been exacerbated by pre-existing malnourishment. Vomiting for several days, abdominal pain, and diarrhoea were reported in a patient who ingested 18 g of disulfiram.

Several cases of paediatric poisoning have been reported. Drowsiness, pyrexia, hypotonia, ataxia, uncontrollable and inappropriate arm movements, irritability,

speech difficulties, hallucinations, coma, and hyperreflexia were the major features.

Disulfiram–ethanol reaction

Nausea, vertigo, anxiety, blurred vision, hypotension, chest pain, palpitation, tachycardia, facial flushing, and throbbing headache are the usual features. Symptoms usually last for 3 to 4 days but may persist for 1 week. Occasionally the reaction is very severe with respiratory depression, cardiovascular collapse, cardiac arrhythmias, coma, cerebral oedema, hemiplegia, and convulsions; fatalities have been reported.

Further reading

Zorzon M *et al.* (1995). Acute encephalopathy and polyneuropathy after disulfiram intoxication. *Alcohol and Alcoholism* **30,** 629–31.

Diuretics

Most overdoses involving diuretics are minor, although inevitably some disturbance of fluid and electrolyte balance will result. When combined diuretic and potassium formulations are ingested, the potassium content is likely to pose the greater risk. More serious consequences are likely if a potassium-sparing diuretic has been ingested.

Clinical features

Symptoms and signs of toxicity include anorexia, nausea, vomiting, diarrhoea, profound diuresis, dehydration, and hypotension. In addition, dizziness, weakness, muscle cramps, tetany, and occasionally gastrointestinal bleeding may be seen. The electrolytic and metabolic disturbances that may be observed include hyponatraemia, hypoglycaemia or hyperglycaemia, hyperuricaemia, hypokalaemia, and metabolic alkalosis. Hyperkalaemia may develop following the ingestion of combined diuretic and potassium preparations and potassium-sparing diuretics, such as amiloride, spironolactone, or triamterene and small-bowel ulceration and stricture formation has followed poisoning due to diuretics with an enteric-coated core of potassium chloride.

Treatment

Symptomatic and supportive therapy should be employed with correction of fluid and electrolyte imbalance. Patients with hyperkalaemia may need a glucose and insulin infusion followed by oral or rectal administration of an ion-exchange resin.

Further reading

Lip GYH, Ferner RE (1995). Poisoning with anti-hypertensive drugs: diuretics and potassium supplements. *Journal of Human Hypertension* **9**, 295–301.

Ethanol

Ethanol is commonly ingested in beverages before, or concomitant with, the deliberate ingestion of other substances in overdose. It is also used as a solvent and is found in many cosmetic and antiseptic preparations. It is rapidly absorbed through the gastric and intestinal mucosas and approximately 95 per cent is oxidized to acetaldehyde and then to acetate; the remainder is excreted unchanged in the urine and to a lesser extent in the breath and through the skin.

Ethanol is a central nervous depressant that exacerbates the effects of other central nervous system depressants, in particular, hypnotic agents. The fatal dose of ethanol alone is between 300 and 500 ml of absolute alcohol, if this is ingested in less than 1 h.

Clinical features

The clinical features of ethanol intoxication are well known and are generally related to blood concentrations (Table 5).

Severe hypoglycaemia may accompany alcohol intoxication due to inhibition of gluconeogenesis. This occurs more commonly in children than in adults and typically occurs within 6 to 36 h of ingestion of a moderate to large amount of alcohol by either a previously malnourished individual or one who has fasted for the previous 24 h. The patient is often in coma and hypothermic but flushing, sweating, and tachycardia are frequently absent. Rarely lactic acidosis, ketoacidosis, and acute renal failure have been described.

Treatment

Hypoglycaemia is usually unresponsive to glucagon and therefore intravenous glucose in a dose of 50 ml of 50 per cent solution should be given. Treatment is supportive. Gastric lavage has not been shown to be of benefit in ethanol poisoning. The use of haemodialysis should be considered if the blood ethanol concentration exceeds 5000 mg/l and/or if metabolic acidosis is present.

Further reading

Ernst AA *et al.* (1996). Ethanol ingestion and related hypoglycemia in a pediatric and adolescent emergency department population. *Academic Emergency Medicine* **3**, 46–9.

Ethylene glycol (1,2-ethanediol)

Ethylene glycol has a variety of commercial applications and is commonly used as an antifreeze fluid in car radiators. Its sweet taste and ready availability have contributed to its popularity as a suicide agent and as a poor man's substitute for alcohol.

It is thought that the minimum lethal dose of ethylene glycol is about 100 ml for an adult, although recovery after treatment has been reported following the ingestion of up to 1 litre.

Mechanism of toxicity

Ethylene glycol itself appears to be non-toxic. Metabolism takes place in the liver and kidneys. Accumulation of metabolites including aldehydes, glycolate, oxalate, and lactic acid may explain toxicity.

Clinical features

The clinical features of ethylene glycol poisoning may be divided into three stages depending on the time after ingestion (Table 6). In addition, hypocalcaemia, severe metabolic acidosis, and calcium oxalate crystalluria are observed in severe cases. The severity of each stage and the progression from one stage to the next depends on the amount of ethylene glycol ingested.

Death may occur during any of the three stages. A serum ethylene glycol concentration in excess of 500 mg/l indicates severe poisoning.

Treatment

Early diagnosis and appropriate therapy significantly reduce the mortality from ethylene glycol poisoning. Gastric lavage should be considered if presentation occurs less than 1 h after ingestion. Supportive measures to combat shock, respiratory distress, hypocalcaemia, and metabolic acidosis should be instituted. Thereafter, treatment has two main aims. First, the competitive inhibition of ethylene glycol metabolism, using ethanol or fomepizole (4-methylpyrozole), and, second, the

increased elimination of the glycol from the body using dialysis. A loading dose of 50 g of ethanol (conveniently given as approximately 125 ml of gin, whisky, or vodka) should be administered followed by an intravenous infusion of 10 to 12 g of ethanol to provide blood ethanol concentrations of 500 mg to 1 g/l. The infusion should be continued until ethylene glycol is no longer detectable in the blood. If dialysis is also employed, the rate of ethanol administration will need to be increased (17 to 22 g/h). Alternatively, fomepizole at 15 mg/kg body weight can be administered over 30 min, followed by four 12-hourly bolus doses of 10 mg/kg until ethylene glycol concentrations are less than 200 mg/l. If treatment is needed for more than 48 h, the bolus dose should be increased to 15 mg/kg to compensate for the induction of fomepizole metabolism. If dialysis is employed, the frequency of dosing should be increased to 4-hourly during dialysis because fomepizole is dialysable.

Ethylene glycol, its aldehyde metabolites, and glycolate may be removed by either peritoneal or haemodialysis, though the latter is two to three times more efficient. Oxalate, however, is poorly dialysable. In addition, it may be necessary to treat the uraemic complications of ethylene glycol poisoning with dialysis and to use haemodialysis/ultrafiltration to correct the sodium overload that can result from the necessary, but sometimes overjudicious, correction of the metabolic acidosis with sodium bicarbonate. Dialysis should be continued until ethylene glycol is no longer detectable in the blood.

Further reading

Barceloux DG *et al.* (1999). American Academy of Clinical Toxicology practice guidelines on the treatment of ethylene glycol poisoning. *Journal of Toxicology—Clinical Toxicology* **37,** 537–60.

Glaser DS (1996). Utility of the serum osmol gap in the diagnosis of methanol or ethylene glycol ingestion. *Annals of Emergency Medicine* **27**, 343–6.

Jacobsen D, McMartin KE (1997). Antidotes for methanol and ethylene glycol poisoning. *Journal of Toxicology—Clinical Toxicology* **35**, 127–43.

Lewis LD, Smith BW, Mamourian AC (1997). Delayed sequelae after acute overdoses or poisonings: cranial neuropathy related to ethylene glycol ingestion. *Clinical Pharmacology and Therapeutics* **61**, 692–9.

Flecainide

Flecainide is a local anaesthetic-type antiarrhythmic drug that inhibits fast sodium channels of cardiac myocytes and markedly shortens action potential duration in the Purkinje system.

Clinical features

The features of overdose are predictable on the basis of the drug's known effects and include hypotension, bradycardia, intraventricular conduction abnormalities, atrioventricular block, and ventricular tachycardia. In severe cases convulsions and cardiorespiratory failure occur and fatalities have been reported.

Treatment

If the patient presents within 1 h of a substantial overdose, gastric lavage should be considered or, alternatively, activated charcoal (50 to 100 g) may be administered. Supportive measures should then be employed, as no specific antidote is available. Lignocaine has been found to be of value in controlling ventricular tachycardia after overdose. There is evidence from volunteer studies that acidification of the urine will increase flecainide elimination; haemodialysis and haemofiltration are of no benefit.

Folic acid

Overdosage with this vitamin does not cause toxic features. No treatment is necessary.

Formaldehyde

Formaldehyde is a flammable, colourless gas with a pungent odour. It is most commonly available commercially as a 30 to 50 per cent w/w aqueous solution and is an important raw material in the synthesis of organic compounds such as plastics and resins.

Metabolism

Formaldehyde is oxidized rapidly to formic acid then converted more slowly to carbon dioxide and water.

Clinical features

Acute exposure

Severe irritation of the mucous membranes of the eyes, nose, and upper airways occurs after minimal exposure to low (less than 5 ppm) formaldehyde concentrations, and tends to prevent higher exposure in even the most tolerant subjects. Substantial exposure may result in severe bronchospasm, pulmonary oedema, and death.

Formaldehyde solutions splashed into the eye have caused corneal damage and skin contamination has resulted in dermatitis. Spillage of phenol-formaldehyde resin on to the skin has produced extensive necrotic skin lesions, fever, hypertension, adult respiratory distress syndrome, proteinuria, and renal impairment. Ingestion of formaldehyde solution has resulted in severe corrosive damage to the buccal cavity and tonsils, oesophagus, and stomach with ulceration, necrosis, and subsequent fibrosis and contracture. Shock, metabolic acidosis (due in part to high formate concentrations), respiratory insufficiency, and renal impairment usually then ensue. Death may follow ingestion of less than 100 ml in an adult.

Treatment

Supportive measures, including the correction of acid–base disturbances, should be employed. Haemodialysis is only moderately effective in increasing formate elimination.

Further reading

Cohen N *et al.* (1989). Acute resin phenol-formaldehyde intoxication. A life threatening occupational hazard. *Human Toxicology* **8**, 247–50.

Glyphosate

Glyphosate-containing herbicides usually incorporate the isopropylamine salt together with a surfactant. The original surfactant was probably the main cause of toxicity but this is no longer present in currently marketed formulations.

Clinical features

The most prominent effects are on the alimentary tract with burning in the mouth and throat, nausea, vomiting, dysphagia, and diarrhoea being the main features. Upper gastrointestinal haemorrhage is a much less common complication. A polymorphic leucocytosis is usual. Hypotension, tachycardia, bradycardia, acute chemical pneumonitis, oliguria, haematuria, and metabolic acidosis may be seen in severe poisoning.

Treatment

Management is largely symptomatic and supportive. Intravenous fluids or blood transfusion may be required. Respiratory and renal failure should be managed

conventionally. The toxicokinetics of glyphosate in man are not known and rational use of elimination procedures is therefore not possible.

Further reading

Chang C-Y *et al*. (1999). Clinical impact of upper gastrointestinal tract injuries in glyphosate-surfactant oral intoxication. *Human and Experimental Toxicology* **18,** 475–8.

n-Hexane

n-Hexane is an extremely volatile liquid that is used as a solvent.

Clinical features

When ingested it causes nausea, dizziness, CNS excitation and then depression and, as a result, presents an acute aspiration hazard resulting in chemical pneumonitis and non-cardiogenic pulmonary oedema. Following inhalation, either inadvertently or deliberately, similar symptoms occur. The development of a progressive sensorimotor neuropathy is the principal hazard of chronic exposure.

Treatment

Treatment is supportive and symptomatic.

Further reading

International Programme on Chemical Safety (1991). *Environmental Health Criteria 122. n-Hexane*. World Health Organization, Geneva.

Household products

There is a commonly held belief that household products contain a wide range of highly toxic chemicals, and so the ingestion of these substances by children is a frequent cause for alarm in parents and doctors alike. So-called poisoning from household products is more often the result of accidental than deliberate ingestion, mostly involves young children, and is not usually serious. Even when the toxicity of a household product is high, the risk it poses is usually low, certainly when ingested accidentally. However, adults intent on suicide may, by deliberately swallowing massive quantities, succeed in killing themselves. Antiseptics and disinfectants, dishwashing liquids, fabric conditioners, detergents, bleaches and lavatory cleaners, lavatory sanitizers, and deodorants are dealt with elsewhere.

Further reading

Gad-Johannsen H, Mikkelsen JB, Larsen CF (1995). Poisoning with household chemicals in children. *Acta Paediatrica* **83,** 62–5.

H_2-receptor antagonists

H_2-receptor antagonists such as cimetidine, famotidine, nizatidine, and ranitidine are very widely prescribed but few cases of overdose have been reported.

Clinical features

Most patients remain asymptomatic. In a few, drowsiness, dryness of the mouth, slurred speech, dizziness, confusion, vomiting, and abdominal discomfort have been reported. Rarely, bradycardia, respiratory depression, and coma may result.

Treatment

Gut decontamination is unnecessary and supportive and symptomatic measures should be employed if features develop. Although forced diuresis has been employed in one case, no supporting evidence of efficacy was given.

Further reading

Krenzelok EP *et al.* (1987). Cimetidine toxicity: an assessment of 881 cases. *Annals of Emergency Medicine* **16**, 1217–22.

Hydrogen fluoride

Hydrogen fluoride is a corrosive, fuming, nearly colourless liquid (hydrofluoric acid) at atmospheric pressures and temperatures below 19°C; above 19°C it is gaseous. Hydrogen fluoride is very soluble in cold water and for this reason it fumes strongly in moist air. Aqueous solutions dissolve glass.

Mechanisms of toxicity

Fluoride directly inhibits many enzyme systems, including glycolytic enzymes, cholinesterases, and those in which magnesium and manganese are present. In addition, fluoride appears to have a direct toxic effect on nerve tissue and muscle; depression of vasomotor and smooth muscle tone may also occur.

Clinical features

Inhalation or ingestion of hydrogen fluoride causes severe corrosive damage similar to other acids (see above). Following absorption by whatever route, fluoride chelates calcium and lowers the serum ionized calcium concentration and causes weakness, paraesthesiae, tetany, and convulsions. Hypotension and cardiac arrhythmias, including ventricular fibrillation, may be observed. Central effects of fluoride include confusion and coma. Hepatic and renal failure may develop.

Skin contact with anhydrous hydrogen fluoride produces liquefactive necrosis and severe burns that are felt immediately. Concentrated aqueous solutions also cause an early sensation of pain, but more dilute solutions may give no warning of injury. If the solution is not removed promptly, penetration of the skin by fluoride ion may occur, leading to painful ulcers that heal only slowly.

Treatment

Following inhalation of hydrogen fluoride, the casualty should be removed immediately from the contaminated atmosphere. Further treatment is symptomatic and supportive. Mechanical ventilation with positive end-expiratory pressure may be needed to treat pulmonary oedema.

If hydrofluoric acid has been ingested, 10 to 20 g of soluble calcium tablets should be given by mouth, followed by an intravenous injection of 10 ml of 10 per cent calcium gluconate solution. Symptomatic and supportive measures should be employed thereafter.

Skin contact requires thorough washing of the affected area with copious quantities of water for 20 min, even if there is no apparent burn or pain. Skin burns should be coated repeatedly with 2.5 per cent calcium gluconate gel, but if the gel is unavailable, immersion of the skin in iced water until the pain subsides is often helpful. If the pain does not subside, 10 per cent calcium gluconate solution (up to 0.5 ml/cm^2) should be injected under the burn area, though calcium gluconate intra-arterially is more effective.

Further reading

Bentur Y *et al.* (1993). The role of calcium gluconate in the treatment of hydrofluoric acid eye burn. *Annals of Emergency Medicine* **22**, 1488–90.

Dunn BJ *et al.* (1996). Topical treatments for hydrofluoric acid dermal burns. *Journal of Occupational and Environmental Medicine* **38**, 507–14.

Henry JA, Hla KK (1992). Intravenous regional calcium gluconate perfusion for hydrofluoric acid burns. *Journal of Toxicology—Clinical Toxicology* **30,** 203–7.

Hydrogen sulphide

Hydrogen sulphide is a colourless gas that smells of rotten eggs, although high concentrations cause olfactory nerve paralysis. The gas is also found in mines and sewers and is liberated from decomposing fish (a hazard in fishing boats if the hold is filled with 'trash' fish used for making fish meal) and liquid manure systems.

Mechanisms of toxicity

It is now thought that the serious sequelae following exposure to high concentrations of hydrogen sulphide are due principally to inhibition of cytochrome oxidase a $_3$, in which respect it may be more potent than cyanide.

Clinical features

Exposure to low concentrations leads to blepharospasm, pain and redness of the eyes, blurred vision, and coloured haloes round lights. Headache, nausea, dizziness, drowsiness, sore throat, and cough may also occur. With exposure to higher concentrations, cyanosis, confusion, pulmonary oedema, coma, and convulsions are common. Six per cent of casualties die.

Treatment

The casualty should be moved to fresh air from the contaminated atmosphere by a rescuer who has donned breathing apparatus beforehand.

It has been shown in mice that the administration of sodium nitrite is superior to oxygen alone in the treatment of acute hydrogen sulphide poisoning. However, the mechanism of this benefit is disputed and the value of this treatment in humans has not been established.

Further reading

Guidotti TL (1996). Hydrogen sulphide. *Occupational Medicine* (Oxford) **46**, 367–71.

Milby TH, Baselt RC (1999). Hydrogen sulfide poisoning: clarification of some controversial issues. *American Journal of Industrial Medicine* **35,** 192–5.

Hypoglycaemic agents

Intentional overdose with insulin and oral hypoglycaemic agents is uncommon. However, deaths from insulin and sulphonylurea overdosage have been reported. Chlorpropamide and glyburide (available only in the United States) are the oral agents most commonly ingested. Chlorpropamide, because of its long half-life, may, in overdose, induce hypoglycaemia for a considerable period of time. In all cases of poisoning with hypoglycaemic agents prompt diagnosis and treatment are essential if death or cerebral damage from neuroglycopenia are to be prevented.

Clinical features

Features of overdosage include drowsiness, coma, twitching, convulsions, depressed limb reflexes, extensor plantar responses, hyperpnoea, pulmonary oedema, tachycardia, and circulatory failure. Hypoglycaemia is to be expected and hypokalaemia, cerebral oedema, and metabolic acidosis might occur. Neurogenic diabetes insipidus and persistent vegetative states are possible long-term complications. Cholestatic jaundice has been described as a late complication of chlorpropamide poisoning.

Treatment

The blood or plasma glucose concentration should be measured urgently and intravenous glucose given. Glucagon may be ineffective. If the blood sugar is normal, gastric lavage should be considered if the patient has presented within 1 h of the ingestion of an oral preparation.

Recurring hypoglycaemia is highly likely. A continuous infusion of glucose together with carbohydrate-rich meals is required in cases of severe insulin overdosage, though there may be difficulty in maintaining normoglycaemia. In the case of sulphonylurea overdosage, however, further glucose (although its administration may be unavoidable) only serves to increase already high circulating insulin concentrations. Diazoxide has therefore been recommended since it increases blood glucose concentrations and raises circulating catecholamine concentrations while blocking insulin release. The dose is 1.25 mg/kg body weight intravenously over 1 h, repeated at 6-hourly intervals if necessary.

Further reading

Palatnick W, Meatherall RC, Tenenbein M (1991). Clinical spectrum of sulfonylurea overdose and experience with diazoxide therapy. *Archives of Internal Medicine* **151,** 1859–62.

Roberge RJ, Martin TG, Delbridge TR (1993). Intentional massive insulin overdose: recognition and management. *Annals of Emergency Medicine* **22**, 228–34.

Iron

Most medicinal iron preparations are ferrous salts that must be oxidized to the ferric state before being absorbed and stored in the liver and reticuloendothelial system. Iron overdosage is much more common in preschool children than in adults. Toxic features are unlikely unless more than 60 mg of elemental iron/kg body weight has been ingested, probably because absorption is poor. Poisoning is therefore seldom severe and deaths are rare.

Mechanism of toxicity

Iron salts have complex actions, including direct corrosive effects on the upper gastrointestinal tract and potentially serious effects on the circulation; at a cellular level they tend to concentrate around mitochondrial cristae where they may act as an electron 'sink', thereby interfering with intermediary metabolism.

Clinical features

The course of iron poisoning is conventionally divided into four phases.

Phase 1

The first phase starts immediately after ingestion and lasts about 6 h. Nausea, vomiting, abdominal pain, and diarrhoea, all of which result from direct irritation of the gut, characterize it. The gastric and upper small bowel mucosas may be stained and impregnated with iron and become ulcerated, the severity of these changes decreasing with distance from the stomach. The disintegrating tablets may make the vomitus and stools grey or black in colour. Polymorphic leucocytosis and hyperglycaemia are common. Iron tablets in the upper gut may be visible in an abdominal radiograph, particularly if it is taken within 2 h of alleged ingestion.

A few patients develop haematemesis, hypotension, coma, and shock, which may be fatal.

Phase 2

This phase lasts from about 6 to 24 h after ingestion and is a period during which patients improve symptomatically. Indeed, most do not progress further.

Phase 3

During this phase, 12 to 48 h after ingestion, a small minority of patients deteriorate, often with profound shock, metabolic acidosis, and features which are due to acute renal tubular and hepatocellular necrosis. Liver failure and its complications develop and may be fatal. The extent of liver damage varies from almost complete necrosis in some areas to only periportal damage in others.

Phase 4

This is the period 2 to 6 weeks after ingestion. The features at this stage are those of high intestinal obstruction by a stricture formed at the site of corrosive damage to the mucosa, usually the pyloric antrum. Children are most likely to be affected.

Assessment of the severity of poisoning

The amount ingested is not reliable because of vomiting. Shock, coma, and acidosis indicate severe poisoning. Other clinical features are less useful. Emergency estimation of the serum iron concentration is essential. If the 4 to 6 h concentration exceeds the predicted normal iron-binding capacity (usually more than 90 µmol/l), free iron is circulating and treatment is needed. Measurement of the total iron-binding capacity in acute iron poisoning may give misleading results and is not recommended.

Treatment

Reducing absorption

Gastric lavage should be considered if more than 20 mg of elemental iron/kg body weight has been ingested in the previous 1 h. Addition of bicarbonate, phosphates, and desferrioxamine (deferoxamine) to lavage fluids does not reduce absorption further and may be dangerous. Whole bowel irrigation may have a role if a large amount (particularly of a slow release formulation) has been ingested and has already passed through the pylorus.

Severe poisoning with coma or shock

When coma or shock are present the specific iron-chelating agent desferrioxamine (deferoxamine) should be given without delay and without waiting for the result of the serum iron concentration. The dose is 15 mg/kg body weight/h intravenously and the total amount infused should not exceed 80 mg/kg in 24 h. Clinical improvement can be expected within 1 to 2 h, after which the rate of infusion may be reduced. There is no simple or readily available method of deciding when to stop desferrioxamine administration; the clinical state of the patient is probably the most appropriate guide. Desferrioxamine may also be given intramuscularly in a dose of 2 g for an adult and 1 g for a child.

Hypotension due to desferrioxamine-induced histamine release may develop if the recommended rate of administration is exceeded. Other adverse effects include hypersensitivity reactions and, rarely, anaphylaxis. Pulmonary oedema and adult respiratory distress syndrome attributed to desferrioxamine have been reported in patients given 15 mg/kg for 65 h and longer.

Poisoning without coma or shock

Routine administration of desferrioxamine cannot be recommended. Patients who are not in coma or shock but who have a serum iron concentration greater than 90 µmol/l should be given desferrioxamine.

Overdose without features

Patients who have not developed features of poisoning within 6 h have probably not ingested toxic amounts and therefore they do not require treatment. Those who present earlier than 6 h should be assessed as described above and treated accordingly.

Supportive measures

Only a small minority of patients will require supportive measures in addition to those described above. Attention to the airway and ventilation is obviously important if consciousness is impaired and fluid and electrolytes should be replaced as necessary. Blood transfusion may be required if there has been significant haemorrhage. Liver and renal function should be monitored and failure managed conventionally.

Overdosage in pregnancy

Overdosage with iron salts during pregnancy should be treated as under other circumstances. Limited evidence indicates that desferrioxamine is not fetotoxic or teratogenic and to withhold it when it is indicated may be fatal.

Further reading

Bosse GM (1995). Conservative management of patients with moderately elevated serum iron levels. *Journal of Toxicology—Clinical Toxicology* **33**, 135–40.

Chyka PA, Butler AY, Holley JE (1996). Serum iron concentrations and symptoms of acute iron poisoning in children. *Pharmacotherapy* **16**, 1053–8.

Tenenbein M (1996). Benefits of parenteral deferoxamine for acute iron poisoning. *Journal of Toxicology—Clinical Toxicology* **34**, 485–9.

Isoniazid

Poisoning with isoniazid is potentially very serious, but uncommon.

Mechanisms of toxicity

Isoniazid depresses brain concentrations of g-aminobutyric acid (GABA), thus leading to seizures.

Clinical features

The ingestion of 80 to 150 mg of isoniazid/kg body weight is likely to cause severe poisoning. Nausea, vomiting, slurred speech, dizziness, and visual hallucinations may develop. Stupor, coma, and convulsions follow rapidly and may be associated with hyperthermia, hyperreflexia, extensor plantar responses, and later, rhabdomyolysis. In addition, dilated pupils, sinus tachycardia, and urinary retention may be observed. In severe cases, hypotension, acute renal failure, and respiratory failure may ensure. Marked metabolic (lactic) acidosis is common. Less commonly, hyperglycaemia, ketoacidosis, glycosuria, and ketonuria are found.

Treatment

Supportive measures including the correction of metabolic acidosis should be instituted immediately if the patient is unconscious. If the airway can be protected, gastric lavage or the administration of 50 g of activated charcoal should be considered if presentation is less than 1 h after overdose. Pyridoxine (1 g for 1 g of isoniazid ingested) should be given intravenously to control convulsions. When the ingested dose of isoniazid is unknown, an initial intravenous dose of 5 g of pyridoxine should be given. Diazepam alone may be ineffective but the use of diazepam and pyridoxine is synergistic and both should be used in those with convulsions. Pyridoxine in a dose of 5 g may be repeated if convulsions persist (in one case 52 g of pyridoxine was given intravenously without ill effects).

Charcoal haemoperfusion is the most effective technique for elimination of isoniazid from the circulation but its use should rarely be necessary provided appropriate supportive measures and adequate and repeated doses of pyridoxine and diazepam are given.

Further reading

Blowey DL, Johnson D, Verjee Z (1995). Isoniazid-associated rhabdomyolysis. *American Journal of Emergency Medicine* **13**, 543–4.

Gurnani A *et al.* (1992). Acute izoniazid poisoning. *Anaesthesia* **47,** 781–3.

Wilcox WD, Hacker YE, Geller RJ (1996). Acute isoniazid overdose in a compliant adolescent patient. *Clinical Pediatrics* **35**, 213–14.

Isopropanol (isopropyl alcohol; 2-propanol)

Isopropanol is used as a sterilizing agent and as rubbing alcohol. It is also found in aftershave lotions, disinfectants, and window-cleaning solutions. Intoxication can result both from ingestion and skin absorption. Isopropanol is oxidized in the liver to acetone.

Clinical features

Features of toxicity include coma and respiratory depression, the odour of acetone on the breath, gastritis, haematemesis, hypotension, hypothermia, renal tubular necrosis, acute myopathy, and haemolytic anaemia; cardiac arrest has occurred. The development of hypotension is a poor prognostic feature.

Treatment

Gastric lavage should be considered if the patient presents less than 1 h after ingestion. In addition to supportive measures, haemodialysis should be employed in severely poisoned patients as it not only removes isopropanol but also shortens the duration of coma.

Further reading

Pappas AA *et al.* (1991). Isopropanol ingestion: report of six episodes with isopropanol and acetone serum concentration time data. *Journal of Toxicology—Clinical Toxicology* **29**, 11–21.

Lavatory sanitizers and deodorants

Solid lavatory sanitizer or deodorant blocks normally contain paradichlorobenzene. Ingestion may cause nausea, vomiting, diarrhoea, and abdominal pain. Symptomatic and supportive treatment is all that is required unless many grams have been ingested in which case gastric lavage should be considered if the patient presents within 1 h of ingestion.

Lead

Exposure to lead occurs in the reclamation of lead from scrap metal, in the demolition and flame-cutting of old railway bridges previously painted with lead paint, and in the manufacture of storage batteries and ceramics. Children with pica who chew on lead-painted railings in homes, or who eat contaminated soil, have developed lead poisoning. As a consequence of lead leaching out of the glazing material, poisoning has also been described in individuals who have consumed drinks from lead-glazed mugs. Ingestion of lead-based powders in paints and imported baby tonics and application of lead-containing cosmetics such as 'surma' to the face in Asian communities has resulted in lead intoxication. Rarely, lead acetate has been injected intravenously with suicidal intent. Tetraethyl lead, which is used as an anti-knock agent in leaded petrol, can be absorbed systemically by inhalation, ingestion, and through the skin. Transplacental transfer of lead from mother to fetus results in reduced viability of the fetus, reduced birth weight, and premature birth.

The Centers for Disease Control and Prevention in Atlanta have reduced the concentration of lead at which intervention is indicated from 150 µg/l as some adverse health effects have been observed in young children at a blood lead concentration of 100 µg/l.

Lead absorbed into the body is mainly (95 per cent) deposited in the bones and teeth. Of the lead in the blood, 99 per cent is associated with erythrocytes. As the body accumulates lead over many years and releases it into the urine only slowly, even small doses can in time lead to intoxication.

Clinical features

Mild intoxication may result in no more than lethargy and occasional abdominal discomfort, whereas abdominal pain (which is usually diffuse but may be colicky), vomiting, constipation, and encephalopathy develop in more severe cases. Lead colic was first described by Hippocrates and, on occasions, has been incorrectly managed surgically as a case of an acute abdomen. Encephalopathy (seizures, mania, delirium, coma) is more common in children than in adults. Classically, lead poisoning results in foot drop attributable to primary motor peripheral neuropathy; wrist drop occurs only as a late sign.

Renal effects include reversible renal tubular dysfunction causing glycosuria, aminoaciduria, and phosphaturia and irreversible interstitial fibrosis with progressive renal insufficiency leading to hypertension.

A bluish discoloration of the gum margins due to deposition of lead sulphide is observed occasionally.

Lead depresses the enzymes responsible for haem synthesis and shortens erythrocyte lifespan leading to microcytic or normocytic hypochromic anaemia. In severe intoxication haemolytic anaemia may occur. Basophilic stippling of erythrocytes is due to nuclear remnants. Lead blocks the conversion of †-aminolaevulinic acid to porphobilinogen leading to an increase in †-aminolaevulinic acid in blood and urine. Lead also inhibits ferrochelatase that results in elevated free erythrocyte protoporphyrin (**FEP**) concentrations. There is a concomitant increase in urinary coproporphyrins and FEP, commonly assayed as zinc protoporphyrin.

An elevated zinc protoporphyrin concentration (more than 350 µg/l) reaches a steady state in the blood only after the entire population of circulating erythrocytes has turned over (approximately 120 days). Consequently, it lags behind blood lead concentrations and is an indirect measure of long-term lead exposure. Moreover, zinc protoporphyrin is not a good screening test, as it is not sensitive at the lower levels of lead poisoning.

Medical surveillance

The current practice in the United Kingdom and some other European countries is to recommend stopping work with lead where a worker's blood lead concentration is shown to be above 700 µg/l although workers may be symptomatic below this concentration. In workers exposed to organic lead compounds, the urinary lead concentration (more than 150 µg/l) is a good indicator of exposure.

Treatment

Primary prevention aimed at eliminating lead hazards for children and workers must receive due public-health attention. The social dimension of the problem must also be recognized: simply giving children chelation therapy and then returning them to a contaminated home environment is of no value. Similarly, if an occupational source of lead exposure is implicated, a thorough evaluation of the workplace, other exposed workers, and the systems for handling lead at work is appropriate.

The decision to use chelation therapy is based not only on the blood lead concentration but also on the symptoms present and, if available, an estimate of the total body burden of lead using X-ray fluorescence. Sodium calcium edetate and DMSA (succimer) both increase lead excretion, though the former must be given intravenously and may result in increased uptake of lead into the brain. In severe acute lead poisoning, particularly of occupational origin, sodium calcium edetate 75 mg/kg body weight daily for 5 days provides rapid relief of symptoms with minimal risk of adverse effect; a second course may be given a week after the first. If hydration is maintained during chelation, proximal tubular damage is not usually observed. DMSA 30 mg/kg body weight orally for 5 days is an alternative though less efficient chelator than sodium calcium edetate.

Further reading

Centers for Disease Control and Prevention (1991). *Preventing lead poisoning in young children*, 4th edn. US Department of Health and Human Services, Washington DC.

Liftshitz M, Hashkanazi R, Phillip M (1997). The effect of 2,3 dimercaptosuccinic acid in the treatment of lead poisoning in adults. *Annals of Medicine* **29**, 83–5.

Lignocaine and related drugs

Lignocaine, mexiletine, and tocainide are sodium channel blockers. Intoxication with these agents, particularly lignocaine, occurs most often as a result of therapeutic overdose in intensive care areas or inadvertent intravenous administration during local anaesthesia. Topical absorption of lignocaine may result in systemic toxicity, particularly in children.

Clinical features

Poisoning induces nausea, vomiting, paraesthesias, tremor, drowsiness, dizziness, dysarthria, diplopia, nystagmus, ataxia, confusion, convulsions, and coma. Sinus bradycardia, heart block, and hypotension may develop in severe poisoning and cardiac arrest may ensue; mexiletine may also cause atrial fibrillation.

Treatment

Gastric lavage should be considered or 50 to 100 g of activated charcoal administered if an overdose has been ingested less than 1 h previously. Diazepam in a dose of 5 to 10 mg intravenously should be given for convulsions, if they are not short-lived, and atropine in a dose of 1 to 2 mg intravenously should be administered for sinus bradycardia. Inotropic support may become necessary if heart block or severe hypotension supervene. Pacing may be attempted but the ventricular response is usually poor. Tocainide elimination is increased significantly with haemodialysis.

Further reading

Denaro CP, Benowitz NL (1989). Poisoning due to class 1B antiarrhythmic drugs: Lignocaine, mexilitine and tocainide. *Medical Toxicology* **4**, 412–28.

Lindane

Lindane is an organchlorine pesticide.

Clinical features

The main toxic effects following ingestion are on the central nervous system with rapid loss of consciousness and the development of myoclonus, hypertonia, hyperreflexia, convulsions, and rhabdomyolysis. Metabolic acidosis, disseminated intravascular coagulation, renal tubular and hepatocellular necrosis, pancreatitis, and proximal myopathy have been reported.

Treatment

Treatment is symptomatic and supportive. Gastric lavage should be considered if lindane has been ingested less than 1 h previously. Metabolic acidosis and convulsions should be treated conventionally.

Further reading

Aks SE *et al.* (1995). Acute accidental lindane ingestion in toddlers. *Annals of Emergency Medicine* **26**, 647–51.

Fischer TF (1994). Lindane toxicity in a 24-year-old woman. *Annals of Emergency Medicine* **24**, 972–4.

Liquefied petroleum gas (LPG 'bottled gas')

Liquefied petroleum gas (LPG 'bottled gas') contains propane and butane (and sometimes propylene and butylene). Propane and butane may cause vertigo and drowsiness and, at high concentrations, may act as asphyxiants.

Further reading

Gray MY, Lazarus JH (1993). Butane inhalation and hemiparesis. *Journal of Toxicology—Clinical Toxicology* **31**, 483–5.

Lithium carbonate

The therapeutic index of lithium is low and toxicity is usually the result of therapeutic overdose rather than deliberate self-poisoning. However, individuals on or not on long-term treatment with the drug occasionally ingest single large doses.

Clinical features

Features of intoxication include thirst, polyuria, diarrhoea, and vomiting, and in more serious cases, impairment of consciousness, hypertonia, tremor, and convulsions; irreversible neurological damage may occur. Measurement of the serum lithium concentration confirms the diagnosis, therapeutic toxicity usually being associated with concentrations above 1.5 mmol/l. However, acute massive overdosage may produce much higher concentrations without causing toxic features, at least initially.

Treatment

Gastric lavage may be considered if the patient presents less than 1 h after a substantial overdose. Thereafter, treatment is supportive together with measures to enhance the rate of lithium elimination. The decision to enhance elimination is based on the severity of features and a serum lithium concentration greater than 3 mmol/l, particularly in patients receiving lithium chronically. Forced diuresis with 0.9 per cent sodium chloride is effective but its use is commonly complicated by hypernatraemia and increased plasma osmolality; the infusion of low-dose dopamine at 2.5 µg/kg.min may be an effective alternative. Peritoneal dialysis or haemodialysis may be needed if renal function is impaired and in severe poisoning; peritoneal dialysis is much less effective. However, the relatively slow movement of lithium ions across cell membranes limits the efficacy of all these techniques. It is easy to reduce serum lithium concentrations but they frequently rebound when

treatment is stopped and clinical improvement is much slower.

Further reading

Scharman EJ (1997). Methods used to decrease lithium absorption or enhance elimination. *Journal of Toxicology—Clinical Toxicology* **35**, 601–8.

Lysergic acid diethylamide (LSD)

As with cannabis, individuals intoxicated with LSD rarely present to medical services.

Clinical features

The ability of LSD to distort reality is well known. Visual hallucinations, distortion of images, agitation, excitement, dilated pupils, tachycardia, hypertension, hyperreflexia, tremor, and hyperthermia are common; auditory hallucinations are rare. Time seems to pass very slowly and behaviour may become disturbed with paranoid delusions. Flashbacks in which the effects of LSD may be re-experienced without further exposure to the drug occur in about 15 per cent of users for several years and are not explained.

Treatment

Most individuals will require little more than reassurance and sedation. Supportive measures are all that can be offered to those who are seriously ill.

Mefenamic acid

Clinical features

Overdose of mefenamic acid produces nausea, vomiting, and occasionally, bloody diarrhoea. Drowsiness, dizziness, and headaches are common and hyperreflexia, muscle twitching, convulsions, cardiorespiratory arrest, hypoprothombinaemia, and acute renal failure have been reported. In a study of 29 cases of mefenamic acid poisoning, convulsions were noted in 38 per cent of patients, although only rarely were they persistent.

Treatment

Gastric lavage or activated charcoal may be considered if the patient presents less than 1 h after overdose. Symptomatic and supportive measures should be employed and haemodialysis/filtration undertaken for renal failure.

Further reading

Turnbull AJ, Campbell P, Hughes JA (1988). Mefenamic acid nephropathy—acute renal failure in overdose. *British Medical Journal* **296**, 46.

Mercury

Mercury is the only metal which is liquid at room temperature. It exists in three forms, metallic (Hg°), mercurous (Hg_2^{2+}), and mercuric (Hg^{2+}). Metallic mercury is very volatile and when spilt has a large surface area so that high atmospheric concentrations may be produced in enclosed spaces, particularly when environmental temperatures are high. In addition to simple salts, such as chloride, nitrate, and sulphate, mercuric mercury forms organometallic compounds where mercury is covalently bound to carbon, such as methyl-, ethyl-, phenyl-, and methoxyethyl mercury.

Non-occupational mercury exposure occurs principally from dietary intake and to a minor extent from dental amalgam. Many foodstuffs contain small amounts of mercury.

The absorption of mercury depends on its chemical form. Inhaled mercury vapour is absorbed rapidly and oxidized to Hg^{2+} in erythrocytes and other tissues. Prior to oxidation, absorbed mercury vapour can cross the blood–brain barrier, but the divalent ion oxidation product serves to trap mercury in the brain. Mercury vapour is also absorbed via the skin, at an average rate of 0.24 ng/cm^2.min. Less than 1 per cent of an ingested dose of metallic mercury reaches the systemic circulation. Organic mercuric salts are better absorbed following ingestion than are inorganic mercuric salts. Organic mercury compounds cross the blood–brain barrier readily to accumulate in the brain. In contrast the kidney is the main storage organ for inorganic mercury compounds. *In vivo* mercury is bound to metallothionein, which serves a protective role since renal damage is caused only by the unbound metal. Mercury is excreted mainly in urine and faeces although a small amount of absorbed inorganic mercury is exhaled as mercury vapour. The half-life of most body mercury is 1 to 2 months but a small fraction has a half-life of several years.

Clinical features

Acute poisoning

Acute mercury vapour inhalation causes headache, nausea, cough, chest pain, bronchitis, and pneumonia. In a few individuals renal damage from such acute exposure may produce gross proteinuria or nephrotic syndrome. In addition, a fine tremor and neurobehavioural impairment occurs and peripheral nerve involvement has also been observed.

Ingestion of metallic mercury is usually without systemic effects as metallic mercury is poorly absorbed from the gastrointestinal tract. However, mercuric chloride or other inorganic mercuric salts cause an irritant gastroenteritis with corrosive ulceration, bloody diarrhoea, and abdominal cramps and may lead to circulatory collapse and shock. The ingestion of disc batteries containing mercuric oxide usually results in uneventful spontaneous passage through the gastrointestinal tract, but potentially toxic mercury levels may result if the battery opens in transit. Mercury-containing batteries have been withdrawn from the European Union. Mercurous compounds are less soluble, less corrosive, and less toxic than mercuric salts. Ingestion of mercurous chloride in teething powder has led to 'pink disease' or acrodynia in infants.

There are reports of deliberate intravenous or subcutaneous metallic mercury injection. Accidental injection also has occurred after injury from broken thermometers and, in the past, following gas analysis procedures using mercury as a syringe sealant. Intravascular mercury may result in pulmonary venous or peripheral arterial embolism. Subcutaneous mercury initiates a soft-tissue inflammatory reaction with granuloma formation. Signs of systemic mercury toxicity are rare following metallic mercury injection.

Chronic poisoning

Chronic poisoning from inorganic mercury compounds or mercury vapour is characterized by non-specific early symptoms such as anorexia, insomnia, abnormal sweating, headache, lassitude, increased excitability, tremor, gingivitis, hypersalivation, personality changes, and memory and intellectual deterioration. Glomerular and tubular damage may follow chronic exposure to mercury and renal tubular acidosis has been described in children.

Exposure to organic mercury compounds usually involves aromatic derivatives such as phenyl mercuric acetate and phenyl mercuric benzoate, or aliphatic compounds such as methylmercury and ethylmercury chloride. The main features of poisoning are paraesthesias of the lips, hands, and feet, ataxia, tremor, dysarthria, constriction of visual fields, deafness, and emotional and intellectual changes. There is often a latent period of several weeks between the last exposure and the development of symptoms.

Treatment

Even prompt removal from exposure to mercury vapour may not prevent the development of serious sequelae. Early intensive supportive measures are of paramount importance in the management of the severe gastrointestinal complications caused by the ingestion of mercuric salts, such as mercuric chloride. In these circumstances gastric lavage is best avoided as significant oesophageal erosions may be present.

Traditionally, dimercaprol (British Anti-Lewisite, BAL) has been used in the treatment of inorganic mercury poisoning, but it has to be administered intramuscularly and has adverse effects. Oral DMPS (unithiol) and DMSA (Succimer) in a dose of 30 mg/kg body weight daily have been shown to enhance mercury elimination significantly, protect against renal damage, and increase survival, at least in animal studies. In some of these experimental studies DMPS appears to be significantly better than DMSA in reducing the total body mercury burden, renal deposition of mercury, and mortality.

DMPS also appears to be of value in the treatment of acute methylmercury poisoning. Limited data suggest DMPS may improve the neurological features of chronic mercury poisoning.

Further reading

O'Carroll RE *et al.* (1995). The neuropsychiatric sequelae of mercury poisoning. The Mad Hatter's disease revisited. *British Journal of Psychiatry* **167**, 95–8.

Toet AE *et al.* (1994). Mercury kinetics in a case of severe mercuric chloride poisoning treated with dimercapto-1-propane sulphonate (DMPS). *Human and Experimental Toxicology* **13,** 11–16.

Torres-Alanis O, Garza-Ocanas L, Pineyro-Lopez A (1997). Intravenous self-administration of metallic mercury: report of a case with a 5 year follow-up. *Journal of Toxicology—Clinical Toxicology* **35**, 83–7.

Metaldehyde

Metaldehyde in the form of pellets is used widely for killing slugs and in some countries as a solid fuel.

Clinical features

Nausea, vomiting, abdominal pain, and diarrhoea often occur 1 to 3 h after ingestion of any amount, while more than 100 mg/kg body weight may cause hypertonia, convulsions, impairment of consciousness, and metabolic acidosis. Hepatic and renal tubular necrosis may become apparent after 2 to 3 days.

Treatment

Gastric lavage should be considered if more than 50 mg/kg has been ingested within 1 h. Treatment thereafter is supportive.

Further reading

Moody JP, Inglis FG (1992). Persistence of metaldehyde during acute molluscicide poisoning. *Human and Experimental Toxicology* **11**, 361–2.

Methanol (methyl alcohol)

Methanol is used widely as a solvent. It is also found in antifreeze solutions, paints, duplicating fluids, paint removers and varnishes, and shoe polishes. The ingestion of as little as 10 ml of pure methanol has caused permanent blindness and 30 ml is potentially fatal although individual susceptibility varies widely. Toxicity may also occur as a result of inhalation or percutaneous absorption.

Mechanisms of toxicity

In humans, methanol is metabolized by alcohol dehydrogenase and catalase enzyme systems to formaldehyde and formic acid (formate). The concentration of formate increases greatly and is accompanied by accumulation of hydrogen ions causing metabolic acidosis.

Clinical features

Ingested alone, methanol causes mild and transient inebriation and drowsiness. After a latent period of 8 to 36 h, nausea, vomiting, abdominal pain, headache, dizziness, and coma supervene. Blurred vision and diminished visual acuity may occur and the presence of dilated pupils, unreactive to light, suggests that permanent blindness is likely to ensue. A severe metabolic acidosis may develop and this may be accompanied by hyperglycaemia and raised serum amylase activity. A blood methanol concentration greater than 500 mg/l confirms serious poisoning. Mortality increases with the severity and duration of the metabolic acidosis. Survivors may show permanent neurological sequelae including blindness, rigidity, hypokinesis, and other parkinsonian-like signs; these features follow the development of optic neuropathy and necrosis of the putamen.

Treatment

Gastric lavage may be considered in patients who present less than 1 h after ingestion. Thereafter, the treatment of methanol poisoning is directed towards: first, the correction of metabolic acidosis; second, the inhibition of methanol oxidation; and third, the removal of circulating methanol and its toxic metabolites. Substantial quantities of bicarbonate (often as much as 2 mol) may be required and since this must be accompanied by sodium, hypernatraemia and hypervolaemia may result.

Ethanol and fomepizole (4-methylpyrazole) inhibit methanol oxidation. These antidotes should be given and monitored as for ethylene glycol. If admission plasma concentrations show that most of the methanol ingested has already been metabolized, ethanol or fomepizole administration will not be of benefit and ethanol might exacerbate the acidosis.

Dialysis is indicated when a patient has ingested more than 30 g of methanol, or develops metabolic acidosis, mental, visual, or fundoscopic abnormalities attributable to methanol, or a blood methanol concentration in excess of 500 mg/l. Folinic acid (30 mg intravenously 6-hourly) may protect against ocular toxicity by accelerating formate metabolism.

Further reading

Brent J *et al*. (2001). Fomepizole for the treatment of methanol poisoning. *New England Journal of Medicine* **344**, 44–90.

Jacobsen D, McMartin KE (1997). Antidotes for methanol and ethylene glycol poisoning. *Journal of Toxicology—Clinical Toxicology* **35**, 126–43.

Methyl bromide (bromomethane)

Methyl bromide is a colourless, odourless gas at ordinary temperatures and, therefore, dangerous concentrations may accumulate without warning. Methyl bromide has high penetrating power and is non-flammable and explosive; these features explain its increasing use as a disinfectant to fumigate soil, a wide range of commodities, grain, warehouses, and mills. Its high density causes it to settle at floor level.

Mechanism of toxicity

Methyl bromide is absorbed readily through the lungs and is excreted largely unchanged by the same route. The remainder is metabolized and inorganic bromide is excreted in the urine. The mechanism of toxicity is uncertain but methyl bromide appears to have an affinity for intracellular proteins, particularly those with sulphydryl

groups.

Clinical features

There is a latent period of up to 12 h before toxic symptoms occur. Symptoms include dizziness, headache, nausea, vomiting, abdominal pain, malaise, transient blurring of vision, diplopia, and breathlessness. In severe cases, coma, status epilepticus, tremor, ataxia, hyporeflexia, paraesthesiae, hallucinations, acute psychosis, and polyneuropathy may be found. Proteinuria, oliguria (due to renal tubular and cortical necrosis), and jaundice have been described.

Long-term exposure to methyl bromide may lead to a chronic polyneuropathy, lethargy, personality changes, intolerance of alcohol, dysarthria, and epilepsy.

Treatment

The casualty should be removed promptly from the contaminated atmosphere and undressed, as methyl bromide can penetrate clothing and rubber gloves. Contaminated skin should be washed with water. Treatment is supportive.

Further reading

Hustinx WNM *et al.* (1993). Systemic effects of inhalational methyl bromide poisoning: a study of nine cases occupationally exposed due to inadvertent spread during fumigation. *British Journal of Industrial Medicine* **50**, 155–9.

Zwaveling JH *et al.* (1987). Exposure of the skin to methyl bromide: A study of six cases occupationally exposed to high concentrations during fumigation. *Human Toxicology* **6**, 491–5.

Methylene chloride (dichloromethane)

Methylene chloride is a common ingredient in paint removers and is used as a solvent for plastic films and cements and also as a degreaser and aerosol propellant.

Mechanism of toxicity

Methylene chloride is metabolized to carbon dioxide and carbon monoxide. Carboxyhaemoglobin concentrations of 3 to 10 per cent (exceptionally 40 per cent) are attained.

Clinical features

Skin contact with liquid methylene chloride can be painful. Following inhalation, dizziness, tingling and numbness of the extremities, throbbing headache, nausea, irritability, fatigue, and stupor have been reported. Severe and prolonged exposure may lead to irritative conjunctivitis, lacrimation, respiratory depression, and death. Hepatorenal dysfunction and pulmonary oedema have also been described. In addition, if high concentrations of carboxyhaemoglobin are present, the features of acute carbon monoxide poisoning may occur, although these tend to be mild even in the presence of very high carboxyhaemoglobin concentrations.

Treatment

Prompt removal from exposure prior to death usually results in complete recovery. Thereafter, treatment is supportive and should include the use of supplemental oxygen.

Further reading

McDonald W, Olmedo M (1996). Accidental deaths following inhalation of methylene chloride. *Applied Occupational and Environmental Hygiene* **11**, 17–19.

Metoclopramide

Overdose causes acute dystonic reactions affecting the eyes, tongue, and neck.

Treatment

Gastric lavage and activated charcoal may be considered if the patient presents less than 1 h after a substantial overdose. Benztropine in a dose of 1 to 2 mg for an adult should be given intravenously if extrapyramidal features are present. Alternatively, 5 to 10 mg of diazepam intravenously is effective and has the additional advantage of alleviating anxiety and agitation.

Further reading

Miller LG, Jankovic J (1989). Metoclopramide-induced movement disorders. *Archives of Internal Medicine* **149**, 2486–92.

Monoamine-oxidase inhibitors (MAOIs)

Phenelzine and tranylcypromine are now used less frequently in the treatment of depression, and poisoning with them is correspondingly uncommon. A new type A MAOI, moclobemide, is now marketed.

Clinical features

The onset of features may be delayed for 12 to 24 h after acute overdose and are due principally to increased sympathetic activity. They include excitement, restlessness (which may be extreme), hyperpyrexia, hyperreflexia, convulsions, opisthotonos, rhabdomyolysis, and coma. Cardiovascular effects include sinus tachycardia and either hypotension or hypertension.

Treatment

Gastric lavage and activated charcoal may be considered if the patient presents less than 1 h after a substantial overdose. Treatment of overdose is essentially supportive and includes control of convulsions and marked excitement with drugs such as diazepam. Dantrolene may be used to treat hyperpyrexia and extreme restlessness. Hypotension should, in the first instance, be treated by fluid replacement to restore a normal circulating blood volume. The use of sympathomimetic drugs should clearly be avoided. Hypertension, which persists despite diazepam administration, should be treated by the administration of an a-adrenoceptor blocker, such as chlorpromazine.

Further reading

Iwersen S, Schmoldt A (1996). Three suicide attempts with moclobemide. *Journal of Toxicology—Clinical Toxicology* **34**, 223–5.

Lichtenwalner MR *et al.* (1995). Two fatalities involving phenelzine. *Journal of Analytical Toxicology* **19**, 265–6.

Natural gas (methane, ethane)

Natural gas contains methane and ethane. Methane and ethane are pharmacologically inert and can be tolerated in high concentrations without producing any toxic effects. Both gases, however, if present in very high concentration (greater than 80 per cent), may produce asphyxia in poorly ventilated areas, as a result of oxygen

deprivation. After removal from the asphyxia-inducing atmosphere, supplemental oxygen should be administered.

Nickel

Nickel is a ubiquitous trace metal and is mined in the form of sulphide ore. It is used primarily for producing stainless steel and other alloys. Nickel carbonyl, an intermediate compound in nickel purification, is used as a catalyst in the petroleum, plastic, and rubber industries. Nickel compounds have been divided into nickel carbonyl, soluble nickel salts (e.g. acetate, bromide, chloride, chloride hexahydrate, nitrate, sub-sulphide, sulphate), insoluble nickel compounds (e.g. arsenate, carbonate, hydroxide, oxide, phosphate) and metallic nickel. Nickel sulphate is used for electroplating and nickel hydroxide is a component of nickel–cadmium batteries.

Nickel can be absorbed both orally and by inhalation, and in the blood is transported bound principally to albumin. Nickel is concentrated in the kidneys, liver, and lungs and is excreted primarily in the urine.

Clinical features

Acute poisoning

Nickel carbonyl is a colourless, volatile liquid which when inhaled leads, within a few minutes, to dizziness, headache, vertigo, nausea, vomiting, cough, and dyspnoea. In many cases these symptoms disappear and there follows a symptom-free period lasting several days before the start of tachypnoea, dyspnoea, haemoptysis, cyanosis, chest pain, vomiting, tachycardia, weakness, and muscle fatigue. Paraesthesiae, diarrhoea, abdominal distension, delirium, and convulsions have also been reported. Death may occur 4 to 11 days after exposure from cardiorespiratory failure.

Urine nickel concentrations immediately following exposure to nickel carbonyl provide a guide as to the severity of exposure.

At high concentrations soluble nickel salts are primarily skin, gut, and eye irritants. Workers at an electroplating plant who drank water accidentally contaminated with nickel sulphate experienced nausea, vomiting, diarrhoea, abdominal pain, headache, cough, and breathlessness which persisted for up to 2 days. A 2-year-old child died 4 h after ingesting 15 g of nickel sulphate crystals.

Chronic poisoning

Chronic exposure to aerosols of nickel salts may lead to chronic rhinitis and sinusitis and, in rare cases, anosmia and perforation of the nasal septum. Inhaled nickel can produce a type I hypersensitivity respiratory reaction manifest as bronchial asthma with circulating IgE antibodies to nickel. Pulmonary eosinophilia (Loeffler's syndrome) due to a type III hypersensitivity reaction to nickel has also been described.

A significant increase in deaths from non-malignant respiratory disease or pneumoconiosis has also been observed in nickel refinery workers. There is evidence that occupational exposure to nickel may cause cancer of the lung and nasal sinuses.

Metallic nickel and nickel salts cause allergic contact dermatitis in up to 10 per cent of females and 1 per cent of males that is due to type IV delayed hypersensitivity.

Treatment

Severe acute nickel carbonyl poisoning requires intensive supportive care. Although chelation therapy with oral or parenteral diethyldithiocarbamate has been employed, its benefit remains uncertain. Ingestion of a substantial quantity of a nickel salt is likely to produce vomiting but if this does not occur, gastric lavage or activated charcoal may be considered if presentation is within 1 h.

Avoidance of exposure and symptomatic treatment of exacerbations with topical or systemic steroids, remain the mainstay of treatment of nickel allergy.

Further reading

Barceloux DG (1999). Nickel. *Journal of Toxicology–Clinical Toxicology* **37**, 239–58.

Bradberry SM, Vale JA (1999). Therapeutic review: do diethyldithiocarbamate and disulfiram have a role in acute nickel carbonyl poisoning? *Journal of Toxicology–Clinical Toxicology* **37**, 259–64.

Nitrates

Organic nitrates such as isosorbide mononitrate and isosorbide dinitrate relax smooth muscle cells and undergo extensive first-pass metabolism in the liver.

Clinical features

The symptoms and signs caused by nitrates in overdose are due primarily to *in vivo* conversion to nitrites causing excessive arteriolar and venous dilation. Headache and vomiting are common, accompanied by flushing of the skin and dizziness. Sinus tachycardia, severe orthostatic hypotension, and syncope may develop. Convulsions and coma may be seen in severely poisoned patients. Methaemoglobinaemia is seen very rarely with organic nitrates.

Treatment

Mild hypotension may be treated by placing the patient in a head-down position, but more severe hypotension will require plasma expanders or a vasopressor agent.

Further reading

Sanders P, Faunt J (1997). An unusual cause of cyanosis (isosorbide dinitrate induced methaemoglobinaemia). *Australian and New Zealand Journal of Medicine* **27,** 596.

Nitrogen dioxide

Combustion of fossil fuels yields nitric oxide and nitrogen dioxide (a largely insoluble, brown, mildly irritating gas). Fermentation of silage produces high concentrations of this gas within 2 days of filling the silo. It is also a by-product of many industrial processes.

Clinical features

The clinical features following acute exposure to high concentrations of nitrogen dioxide depend on the concentration and duration of exposure to the gas. Since nitrogen dioxide is only a mild upper respiratory tract irritant, modest acute exposure (less than 50 ppm) for a short time often produces no immediate symptoms, although throat irritation, cough, transient choking, tightness in the chest, and sweating have been observed. By contrast, exposure to a massive concentration of nitrogen dioxide such as that found in a silo can produce severe and immediate hypoxaemia, which may be fatal. In less severe cases, the onset of symptoms may be delayed for a few hours (typically 3 to 36 h) and the patient then develops dyspnoea, chest pain (which may be pleuritic), haemoptysis, tachycardia, headache, conjunctivitis, generalized weakness, and dizziness (which may be due to hypotension). Bronchiolitis obliterans may develop within 2 to 6 weeks.

Treatment

Bronchodilator and corticosteroid therapy is sufficient in most cases. Pulmonary oedema responds poorly to diuretics; corticosteroids and mechanical ventilation with

positive end-expiratory pressure offer the best hope of reducing the mortality.

Further reading

Karlson-Stiber C *et al.* (1996). Nitrogen dioxide pneumonitis in ice hockey players. *Journal of Internal Medicine* **239**, 451–6.

Opiates and opioids

Acute opioid overdose occurs commonly in 'addicts' in whom the presence of venepuncture marks and thrombosed veins in the arms and legs should prompt the diagnosis.

Clinical features

The cardinal signs of opioid overdose are pinpoint pupils, reduced respiratory rate (often accompanied by cyanosis), and coma. These depressant effects are increased by alcohol. Hypotension, due to peripheral vasodilation, occurs in less than 10 per cent of cases. Hypothermia and hypoglycaemia may also complicate the clinical picture of opioid poisoning. As many as 50 per cent of heroin overdose victims develop non-cardiogenic pulmonary oedema, the majority of whom, in turn, develop bacterial pneumonia.

Methadone poses particular problems because of its long half-life (up to 50 h).

Codeine, dextropropoxyphene, and pethidine cause increased muscle tone, twitching, and convulsions. Rhabdomyolysis and its complications have been reported in association with poisoning due to diamorphine, dihydrocodeine, dipipanone, methadone, and morphine poisoning.

Diphenoxylate is used as an antidiarrhoeal agent in conjunction with atropine, and paediatric poisoning due to the ingestion of this antidiarrhoeal preparation is not uncommon (see Co-phenotrope).

Management

Naloxone is used to reverse severe respiratory depression and coma due to opioid poisoning. The adult dose is 0.8 to 1.2 mg given intravenously or, less satisfactorily, intramuscularly; the dose in children is 5 to 10 µg/kg body weight. If the diagnosis of opioid poisoning is correct, the patient should improve within 1 min with an increase in respiratory rate, an improvement in the level of consciousness, and dilation of the pupils. In severe opioid poisoning, larger initial doses of naloxone (e.g. 2.4 mg) may be required to obtain the desired response. The duration of action of naloxone (1 to 4 h) is often less than that of the drug taken in overdose and, for this reason, careful observation of the patient is necessary. Repeated doses of naloxone should be given as required.

The respiratory depressant effects of pentazocine and buprenorphine are only partially reversed by naloxone. Assisted ventilation may be necessary.

Gastric lavage and the administration of activated charcoal may be of value if an opioid had been ingested in overdose less than 1 h previously.

The development of non-cardiogenic pulmonary oedema may necessitate the use of assisted ventilation. Antibiotics will be required to treat secondary bacterial infection. Hyperkalaemia and renal failure, as a result of rhabdomyolysis, should be treated conventionally.

Organophosphorus insecticides

Organophosphorus insecticides are among the most extensively used pesticides throughout the world. They vary widely in their toxicity and while some (the phosphates) are directly toxic, others (the phosphorothioates) need biotransformation to become active.

Mechanisms of toxicity

Organophosphorus insecticides inhibit acetylcholinesterase causing accumulation of acetylcholine at central and peripheral cholinergic nerve endings, including neuromuscular junctions.

Clinical features

The features of organophosphorus insecticide poisoning are dose related. Minor exposure may produce subclinical poisoning in which there is reduction of cholinesterase activity but no symptoms or signs. Poisoning is characterized by anxiety, restlessness, insomnia, nightmares, tiredness, dizziness, headache, and muscarinic features such as nausea, vomiting, abdominal colic, diarrhoea, tenesmus, sweating, hypersalivation, and chest tightness. Miosis may be present. Nicotinic effects follow with muscle fasciculation and flaccid paresis of limb muscles, respiratory muscles, and occasionally, various combinations of extraocular muscles. Respiratory failure ensues and is exacerbated by the development of pulmonary oedema and by the retention in the bronchi of large amounts of respiratory secretions. Consciousness is impaired in severe poisoning and convulsions may occur. Hyperglycaemia and glycosuria have been reported though ketonuria is absent. Though bradycardia would be expected from the mode of action of organophosphates, it is present in only about 20 per cent of cases. Rarely, complete heart block and arrhythmias occur.

Diagnosis

Diagnosis of organophosphorus insecticide poisoning is difficult in the absence of a history of exposure and requires a high index of suspicion. Gastroenteritis is a common erroneous diagnosis and the findings of glycosuria and hyperglycaemia may prompt consideration of diabetes mellitus and its complications. Miosis is an important diagnostic sign but is not invariable. Once raised, the diagnosis can be confirmed by demonstrating reduced plasma, but preferably erythrocyte, cholinesterase activity. However, the extent of reduction correlates only crudely with the severity of poisoning. In subclinical poisoning cholinesterase activity may be reduced by up to 50 per cent while mild, moderate, and severe poisoning are associated with reduction of cholinesterase activity to approximately 20 to 50 per cent, 10 to 20 per cent, and less than 10 per cent of normal, respectively.

Treatment

Subclinical poisoning does not require treatment other than appropriate measures to prevent further absorption of the poison. The patient should be kept under observation for about 24 h to ensure that delayed toxicity does not develop. The management of symptomatic organophosphorus insecticide poisoning involves supportive measures and judicious administration of antidotes. Soiled clothing should be removed and contaminated skin washed with soap and water to prevent further absorption. Gastric lavage should be considered if the insecticide has been ingested less than 1 h previously. A clear airway, effective removal of respiratory secretions, and correction of hypoxia are essential using endotracheal intubation and assisted ventilation if necessary. The early use of diazepam may reduce morbidity and mortality; 5 to 10 mg intravenously for an adult reduces anxiety and restlessness but larger doses may be required to control convulsions.

Atropine in a dose of 2 mg intravenously every 10 to 30 min for an adult, depending on the severity of poisoning, should be given to reduce bronchorrhoea and bronchospasm or until signs of atropinization (flushed dry skin, tachycardia, and dry mouth) develop. As much as 30 mg and occasionally much more may be required in the first 24 h in severe cases. Children should be given atropine at 0.02 mg/kg body weight but may require up to 0.05 mg/kg.

Pralidoxime reactivates phosphorylated acetylcholinesterase and should be given together with atropine to every symptomatic patient. The dose (of the mesylate and chloride salts) is 30 mg/kg body weight by slow intravenous injection. Improvement will usually be apparent within 30 min. Further bolus doses of pralidoxime may be required every 4 to 6 h. Alternatively, an infusion of pralidoxime at 8 to 10 mg/kg body weight per h may be administered. Monitoring of erythrocyte (not plasma) cholinesterase activity may be used together with clinical signs to guide the duration of therapy.

Complications

A small number of patients develop what has been called the intermediate syndrome that comprises cranial nerve and brainstem lesions and a proximal neuropathy starting 1 to 4 days after acute intoxication and persisting for 2 to 3 weeks. Respiratory failure secondary to muscle weakness is observed. The aetiology of this syndrome is uncertain but is probably due to inadequate oxime therapy.

A variety of longer-term complications may develop including tiredness, insomnia, inability to concentrate, depression, and irritability. A peripheral neuropathy starting 2 weeks after exposure and mainly affecting the lower limbs is also well recognized. Axonal degeneration of large myelinated motor and sensory fibres has been demonstrated and is thought to be caused by inhibition of neuropathy target esterase.

Further reading

Benson B, Tolo D, McIntyre M (1992). Is the intermediate syndrome in organophosphate poisoning the result of insufficient oxime therapy? *Journal of Toxicology—Clinical Toxicology* **30,** 347–9.

Committee on Toxicity of Chemicals in Food, Consumer Products and the Environment (1999). *Organophosphates*. Department of Health, London.

Okumura T *et al.* (1996). Report on 640 victims of the Tokyo subway sarin attack. *Annals of Emergency Medicine* **28**, 129–35.

Thiermann H *et al.* (1997). Cholinesterase status, pharmacokinetics and laboratory findings during obidoxime therapy in organophosphate poisoned patients. *Human and Experimental Toxicology* **16**, 473–80.

Oxicams

These include meloxicam, piroxicam, and tenoxicam.

Clinical features

The clinical features of oxicam overdose are summarized in Table 7.

Treatment

Treatment is symptomatic and supportive. Gastric lavage or activated charcoal may be considered if the patient presents within 1 h of a substantial overdose.

Paracetamol (acetaminophen)

Mechanism of toxicity

The toxicity of paracetamol is related to its metabolism. In therapeutic doses, 60 to 90 per cent is metabolized by conjugation to form paracetamol glucuronide and sulphate (Fig. 2). A much smaller amount (5 to 10 per cent) is oxidized by mixed function oxidase enzymes to form a highly reactive compound (*N*-acetyl-*p*-benzoquinoneimine, **NAPQI**) that is then immediately conjugated with glutathione and subsequently excreted as cysteine and mercapturate conjugates. Only 1 to 4 per cent of a therapeutic dose of the drug is excreted unchanged in urine.

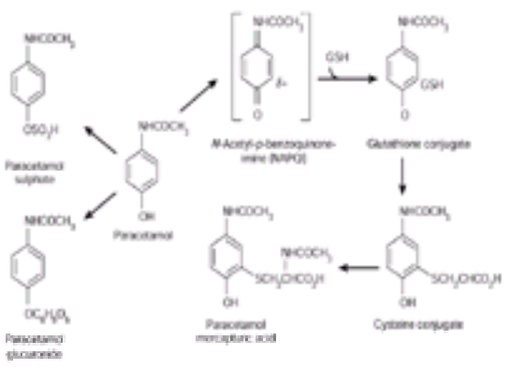

Fig. 2 Paracetamol metabolism in therapeutic dose and overdose.

In overdose, larger amounts of paracetamol are metabolized by oxidation because of saturation of the sulphate conjugation pathway. As a result, liver glutathione stores become depleted so that the liver is unable to 'deactivate' the toxic metabolite. The reactive metabolite has a high affinity for cell protein and binds to liver cell macromolecules. However, covalent binding of NAPQI to cell structure nucleophiles is not thought to be directly responsible for paracetamol-induced hepatic necrosis. NAPQI is believed to have two separate but complementary effects. First, it reacts with glutathione, thereby depleting the cell of its normal defence against oxidizing damage. Second, it is a potent oxidizing as well as arylating agent; it inactivates key sulphydryl groups in certain enzymes, particularly those controlling calcium homeostasis. Inhibition of membrane calcium translocase activity and impairment of microsomal calcium uptake leads to a marked increase in cytosolic calcium concentration, which causes depolymerization of microtubules and contraction of microfilaments, with consequent disruption of cellular architecture and function. The activity of the mixed function oxidase enzyme system and the size of liver glutathione stores may be modified by pharmacological means.

Paracetamol-induced renal damage probably results from a mechanism similar to that which is responsible for hepatotoxicity, that is by formation of NAPQI, although in the kidney this is generated by prostaglandin endoperoxide synthetase rather than by cytochrome P450-dependent mixed function oxidases.

Prediction of liver damage

In the early stages following ingestion of a paracetamol overdose, most patients have few symptoms and no physical signs. There is thus a need for some form of assessment that estimates the risk of liver damage at a time when the liver function tests are still normal. Details of the dose ingested may be used but, in many cases, the history is unreliable and, even when the dose is known for certain, it does not take account of early vomiting and individual variation in response to the drug. However, a single measurement of the plasma paracetamol concentration is an accurate predictor of liver damage provided that it is taken not earlier than 4 h after the overdose. Information gained from several studies has enabled the production of a graph which may be used for prediction of liver damage and which serves as a guide to the need for specific treatment (Fig. 3). Sixty per cent of patients whose plasma paracetamol concentration falls above a line drawn between 200 mg/l (1.32 mmol/l) at 4 h and 50 mg/l (0.33 mmol/l) at 12 h after the ingestion of the overdose are likely to sustain liver damage (serum alanine or aspartate aminotransferase, **ALT** or **AST**, greater than 1000 iu/l) unless specific protective treatment is given. When more than 12 h have lapsed after ingestion, the plasma paracetamol concentration is still of value and should be considered in conjunction with changes in the prothrombin time (see below) when assessing the prognosis of an individual patient.

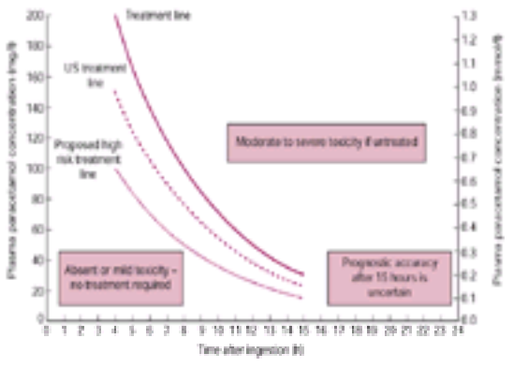

Fig. 3 Prediction of liver damage after paracetamol overdose.

There is, however, some variation in individual susceptibility to paracetamol-induced hepatotoxicity and patients with pre-existing liver disease, those with a high alcohol intake particularly if malnourished, and those receiving enzyme-inducing drugs should be considered to be at greater risk ('High risk group'; Fig. 3). Individuals with HIV-related disease also appear to be more susceptible to paracetamol-induced hepatic damage. It is uncommon for young children to develop paracetamol-induced liver or renal damage, probably because they ingest relatively small amounts in overdose.

Clinical features

As would be expected from the mechanism of toxicity, the severity of paracetamol poisoning is dose-related. An absorbed dose of 15 g (approximately 200 mg/kg) or more is potentially serious in most patients.

Following the ingestion of an overdose of paracetamol, patients usually remain asymptomatic for the first 24 h or, at most, develop anorexia, nausea, and vomiting. Liver damage is not usually detectable by routine liver function tests until at least 18 h after ingestion of the drug, and hepatic tenderness and abdominal pain are seldom exhibited before the second day. Maximum liver damage, as assessed by plasma ALT or AST activity or prothrombin time, occurs 72 to 96 h after ingestion. Hepatic failure, manifest by jaundice and encephalopathy, may then develop between the third and fifth day (Table 8) with the rate of clinical deterioration reflecting the severity of the overdose. More usually there is prolongation of the prothrombin time and a marked rise in aminotransferase activity without the development of fulminant hepatic failure. Renal failure due to acute tubular necrosis develops in about 25 per cent of patients with severe hepatic damage and in a few without evidence of serious disturbance of liver function.

Other features, including hypoglycaemia and hyperglycaemia, cardiac arrhythmias, pancreatitis, gastrointestinal haemorrhage, and cerebral oedema, may all occur with hepatic failure due to any cause and are not direct consequences of paracetamol toxicity.

There are two additional metabolic complications of paracetamol overdosage: metabolic acidosis and hypophosphataemia. Paracetamol can cause metabolic acidosis at two distinct periods after overdosage. Transient hyperlactataemia is frequently found within the first 15 h of ingestion of paracetamol in all but minor overdoses. This appears to be due to inhibition of mitochondrial respiration at the level of ubiquinone and increased lactate production, and may be associated with a metabolic acidosis. It is rarely of clinical consequence, although in very severe paracetamol poisoning (plasma paracetamol concentration more than 500 mg/l at 4 h after ingestion) the acidosis may be associated with coma. The second phase of hyperlactataemia and acidosis occurs in those patients who present late and go on to develop hepatic damage. In this instance decreased hepatic lactate clearance appears to be the major cause, compounded by poor peripheral perfusion and increased lactate production. The development of lactic acidosis consequent upon paracetamol-induced liver damage is associated with a poor prognosis.

Hypophosphataemia is a recognized complication of acute liver failure, including that due to paracetamol, and may contribute to morbidity and mortality by inducing mental confusion, irritability, coma, and abnormalities of platelet, white cell, and erythrocyte function. Phosphaturia appears to be the principal cause of hypophosphataemia in paracetamol poisoning; it may occur in the absence of fulminant hepatic failure and indicates paracetamol-induced renal tubular damage.

Prognostic factors

The overall mortality of paracetamol poisoning in untreated patients is only of the order of 5 per cent. The prothrombin time is usually the first liver function test to become abnormal and, for this reason, it is of particular value in assessing the prognosis of an individual patient. The more rapid the increase in prothrombin time, the worse the prognosis of the patient. A prothrombin time of more than 20 s at 24 h after ingestion indicates that significant hepatic damage has been sustained, and a peak prothrombin time of more than 180 s is associated with a chance of survival of less than 8 per cent.

Acid–base disturbances are also a good guide to prognosis. Systemic acidosis developing more than 24 h after overdose indicates a poor prognosis; patients with a blood pH below 7.30 at this time have only a 15 per cent chance of survival. In addition, a rise in the serum creatinine concentration is associated with poor survival; patients with a serum creatinine concentration above 300 µmol/l have only a 23 per cent chance of survival.

A study of prognostic indicators in paracetamol-induced fulminant hepatic failure treated conventionally compared the sensitivity (percentage of patients who died with a positive test), predictive accuracy (percentage of patients whose outcome was predicted accurately), positive predictive value (percentage of patients with a positive test who died), and specificity (percentage of survivors with a negative test) of measurement of factors V and VIII with conventional tests. (Factor V is vitamin K dependent and concentrations fall in liver failure; levels of factor VIII rise in patients with liver failure.) An admission pH below 7.30 with a serum creatinine concentration above 300 µmol/l and a prothrombin time above 100 s in grade III to IV encephalopathy has a sensitivity, predictive accuracy, positive prediction value, and specificity of 91, 86, 83, and 91, respectively. However, a factor VIII/V ratio above 30 had comparable values of 91, 95, 100, and 100.

Treatment

Gastric lavage and activated charcoal may be of value within 1 h of overdose. Parenteral fluid replacement should be given if nausea persists or vomiting occurs.

From knowledge of the mechanism of toxicity, it may be predicted that replenishment of glutathione stores would be of value. Two substances, methionine and *N*-acetylcysteine have emerged as effective protection agents, provided that they are administered within 8 to 10 h of ingestion of the overdose; thereafter, the protective effects decline rapidly (Table 9). Both substances act by replenishing cellular glutathione stores, though *N*-acetylcysteine may also repair oxidation damage caused by NAPQI either directly or, more probably, through the generation of cysteine and/or glutathione. It may also act as a source of sulphate and so 'unsaturate' sulphate conjugation. Methionine appears more effective when given orally than when administered intravenously. As oral *N*-acetylcysteine induces vomiting in most patients, the intravenous route is preferred. Some 6 per cent of patients treated with intravenous *N*-acetylcysteine develop rash, angio-oedema, and bronchospasm. These reactions are seldom serious and no fatalities have so far been reported in those receiving the regimen outlined in Table 9. An antihistamine such as chlorpheniramine or terfenadine should be given if such anaphylactoid reactions do not settle after discontinuing the infusion of *N*-acetylcysteine for 30 to 45 min.

Forced diuresis, dialysis, or haemoperfusion have no role to play in the management of paracetamol poisoning, though dialysis or haemofiltration will be required if acute renal failure supervenes.

Fortunately, only a minority of patients present more than 12 to 24 h after an overdose of paracetamol but, in these cases, the morbidity and mortality is greater and the correct treatment is that intended to prevent or support hepatic failure, though the use of *N*-acetylcysteine has been advocated. Ten per cent glucose solution should be administered to prevent the onset of hypoglycaemia. Although there is no evidence that correction of severe coagulation abnormalities (prothrombin time greater than 100 s) with fresh frozen plasma improves the prognosis, bleeding can be catastrophic if it does occur. Fresh frozen plasma will be required to cover the insertion of intracranial pressure monitoring apparatus, if employed. An H_2-receptor antagonist or proton pump inhibitor may reduce the risk of gastrointestinal bleeding from 'stress' ulceration/erosion. If acute renal failure supervenes, this should be managed conventionally.

Current evidence suggests that if fulminant hepatic failure does supervene, the use of intravenous *N*-acetylcysteine (see Table 9 for 20.25 h regimen; the 16 h

infusion is continued until recovery or death) will reduce morbidity and mortality. This beneficial effect is observed even after the onset of encephalopathy. In one study the survival rate in 25 patients with paracetamol-induced fulminant hepatic failure was 20 per cent, with an incidence of cerebral oedema and hypotension requiring inotropic support of 68 and 80 per cent, respectively. With *N*-acetylcysteine, the comparable figures in 25 matched patients were 48 per cent (survival rate), 40 per cent (cerebral oedema), and 48 per cent (hypotension), respectively.

Liver transplantation has been performed successfully in patients with paracetamol-induced fulminant hepatic failure, using criteria outlined above (see Prognostic factors) to identify those individuals who would otherwise be likely to die. However, there has been no formal study to compare the value of transplantation with *N*-acetylcysteine in fulminant hepatic failure.

Further reading

Keays P *et al.* (1991). Intravenous acetylcysteine in paracetamol induced fulminant hepatic failure: A prospective controlled trial. *British Medical Journal* **303**, 1026–9.

Makin AJ, Wendon J, Williams R (1995). A 7-year experience of severe acetaminophen-induced hepatotoxicity (1987–1993). *Gastroenterology* **109**, 1907–16.

Vale JA, Proudfoot AT (1995). Paracetamol (acetaminophen) poisoning. *Lancet* **346,** 547–52.

Paraffin oil (kerosene)

Paraffin oil has three physical properties accounting for its toxicity. Its low viscosity and surface tension allow it to spread rapidly throughout the lungs when aspirated after ingestion. Its low vapour pressure makes it unlikely to cause poisoning by inhalation.

Clinical features

Repeated local application to the skin results in dryness, dermatitis, and rarely, epidermal necrolysis. Pulmonary toxicity may occur within 1 h of ingestion and is characterized by pyrexia, cough, tachypnoea, tachycardia, basal crackles, and cyanosis. Non-segmental consolidation or collapse is seen radiologically. Pneumatocele formation, pneumothorax, pleural effusion, or pulmonary oedema may occur.

Paraffin ingestion causes respiratory symptoms, a burning sensation in the mouth and throat, vomiting, diarrhoea, abdominal pain, mild hepatomegaly with hepatic dysfunction, and in severe cases, atrial fibrillation and ventricular fibrillation.

Treatment

Gastric lavage and emesis should be avoided because of the increased risk of chemical pneumonitis, but lavage may be considered in those adults who ingest very large quantities of paraffin oil, if the airway can be protected and the procedure can be carried out within 1 h. There is no evidence that corticosteroids and antibiotics reduce morbidity or mortality; mechanical ventilation with positive end-expiratory pressure may be necessary in severe cases of aspiration.

Further reading

Baldachin BJ, Melmed RN (1964). Clinical and therapeutic aspects of kerosene poisoning: A series of 200 cases. *British Medical Journal* **2**, 28–30.

Nagi NA, Abdullah ZA (1995). Kerosene poisoning in children in Iraq. *Postgraduate Medical Journal* **71**, 419–22.

Paraquat and other bipyridyl herbicides

The bipyridilium herbicides include diquat, morfamquat, and paraquat, the last being the one most commonly encountered in clinical toxicology.

Clinical features

Occupational carelessness in handling paraquat has led to reversible changes in the fingernails and inhalation of spray may cause pain in the throat and epistaxis. Skin splashes that are promptly and thoroughly washed should not cause problems, but prolonged dermal exposure may cause burns and, very rarely, may enable enough paraquat to be absorbed to cause serious and fatal systemic poisoning. Splashes in the eye cause blepharospasm, lacrimation, and corneal ulceration.

Potentially lethal poisoning is most common after paraquat ingestion. Probably no more than 5 per cent of the ingested amount is absorbed but absorption is rapid, the volume of distribution is high, and there is energy-dependent accumulation in some organs (particularly the lungs). Elimination is mainly through the kidneys.

The features of toxicity are largely dependent on the amount of paraquat swallowed. Ingestion of 6 g or more of paraquat ion is likely to be fatal within 24 to 48 h, while 3 to 6 g is likely to lead to a more protracted, but still fatal, outcome. Nausea, vomiting, abdominal pain, and diarrhoea, rapidly followed by peripheral circulatory failure, metabolic acidosis, impaired consciousness, convulsions, and increasing breathlessness and cyanosis secondary to acute pneumonitis are the features of ingestion of large amounts of paraquat. With smaller amounts, the cardiovascular and CNS complications are not seen and alimentary features dominate the course of poisoning, particularly painful ulceration of the mouth, tongue, and throat, which makes it difficult to swallow, speak, and cough. Perforation of the oesophagus with subsequent mediastinitis has been reported. Mild jaundice may be seen and renal failure is usually severe. Breathlessness, tachypnoea, widespread crepitations, and central cyanosis may be present by 5 to 7 days after ingestion and progress relentlessly until the patient dies from hypoxia a few days later.

Ingestion of 1.5 to 2.0 g of paraquat causes nausea, vomiting and diarrhoea, mild renal tubular necrosis, and pain in the throat. Respiratory involvement may not be apparent until 10 to 21 days after ingestion, but may progress till the patient dies of respiratory failure as late as 5 or 6 weeks after taking the paraquat.

Diagnosis

The diagnosis of poisoning is usually made on the basis of the history and can be readily confirmed by a simple qualitative test on urine. This test is of particular value in accidental inhalation or ingestion of very small quantities and if performed on urine passed within 4 h of alleged ingestion; a negative test indicates that not enough has been taken to cause problems.

Treatment

There is no evidence that any form of intervention can alter the outcome of paraquat poisoning. It is traditional to empty the stomach despite the corrosive effects of the toxin, but administration of activated charcoal is probably more effective in reducing absorption.

Symptomatic measures including antiemetics, mouth washes, and analgesics are indicated and intravenous fluids may be necessary to replace gastrointestinal losses. Skin ulcers should be treated as burns. Unfortunately, currently available techniques for enhancing the elimination of poisons appear incapable of rapidly removing toxicologically significant quantities of paraquat. Equally there is no evidence that corticosteroids, drugs to prevent free radical formation, free radical scavengers, immunosuppressive agents, radiotherapy to the lungs, or lung transplantation reduce mortality. The prognosis in individual cases can be predicted from the plasma paraquat concentration related to the time from ingestion.

Further reading

Jones GM, Vale JA. Mechanisms of toxicity, clinical features and management of diquat poisoning: A review. *Journal of Toxicology—Clinical Toxicology* **38**, 123–8.

Pond SM (1990). Manifestations and management of paraquat poisoning. *Medical Journal of Australia* **152,** 256–9.

Petrol (gasoline)

Petrol is a complex mixture of hydrocarbons containing a small proportion of non-hydrocarbon additives.

Clinical features

Acute exposure

Following the inhalation of petrol, dizziness and irritation of the eyes, nose, and throat may occur within 5 min followed by euphoria, headache, and blurred vision. If inhalation continues, or if significant quantities of petrol are ingested, then excitement and depression of the nervous system occurs; incoordination, restlessness, excitement, confusion, disorientation, hallucinations, ataxia, nystagmus, tremor, delirium, coma, and convulsions may be seen. The inhalation of high concentrations of petrol may cause immediate death, probably from ventricular fibrillation or respiratory failure. Chemical pneumonitis may occur as in paraffin oil ingestion (see above) and the clinical features and management are then identical. In addition, intravascular haemolysis, hypofibrinogenaemia, and cardiorespiratory arrest have been reported together with, in one patient, epiglottitis so severe that near total airway obstruction resulted.

Chronic exposure

Men engaged in cleaning storage tanks and those who habitually sniff may develop both hydrocarbon and lead poisoning.

Treatment

Following removal from exposure, supportive measures provide the basis of treatment. Gastric lavage and emesis will increase the risk of aspiration and chemical pneumonitis. Lavage could be considered, with protection of the airway, if an adult ingested a large amount of petrol and the procedure could be performed within 1 h of ingestion.

Further reading

Caprino L, Togna GI (1998). Potential health effects of gasoline and its constituents: a review of current literature (1990–1997) on toxicological data. *Environmental Health Perspectives* **106,** 115–25.

Phenol

Phenol (carbolic acid) is recognizable by its odour and, distinctively, the pain to which it gives rise is much less than might be expected. This is due to its ability to damage afferent nerve endings.

Clinical features

If phenol is spilt on the skin, pain is followed promptly by numbness. The skin becomes blanched, and a dry opaque eschar forms over the burn. When the eschar sloughs off, a brown stain remains. Phenol penetrates intact skin rapidly and is well absorbed through the lungs. After ingestion, vomiting and abdominal pain occur. Systemic toxicity may follow exposure by any route. Features include coma, loss of vasoconstrictor tone, and hypothermia together with cardiac and respiratory depression. An initial phase of CNS stimulation, and rarely convulsions, has sometimes been observed in children. Phenol poisoning is associated with grey or black urine and although this is due in part to metabolites of phenol, Heinz body haemolytic anaemia as well as methaemoglobinaemia and hyperbilirubinaemia are recognized features. Renal complications are seen frequently.

Treatment

Gastric lavage may be considered if the patient presents less than 1 h after ingestion and severe oropharyngeal burns are not suspected. Skin and eye contamination, renal failure, and methaemoglobinaemia are managed appropriately.

Further reading

Christiansen RG, Klaman JS (1996). Successful treatment of phenol poisoning with charcoal hemoperfusion. *Veterinary and Human Toxicology* **38,** 27–8.

Phenothiazines

The phenothiazines block peripheral cholinergic and a-adrenergic receptors, reuptake of amines, and the effects of histamine and serotonin.

Clinical features

The features of overdose include impairment of consciousness, hypotension, and respiratory depression. Chlorpromazine, perphenazine, and promazine seem more prone to cause hypothermia and hypotension, while anticholinergic effects with tachycardia, ECG abnormalities, and arrhythmias are most common with overdosage of thioridazine and mesoridazine, and acute spasmodic torticollis, oculogyric crises, and orolingual dyskinesias with trifluperazine and prochlorperazine.

Treatment

Treatment is supportive and symptomatic. Benztropine in a dose of 2 mg intravenously in an adult is required occasionally to reverse spasmodic torticollis and oculogyric crises.

Phenylpropionic (arylpropionic) acid derivatives

These include fenbufen, fenoprofen, flurbiprofen, ibuprofen, cetoprofen, naproxen, and tiaprofenic acid.

Clinical features

Propionic acid derivative poisoning causes nausea, vomiting, abdominal pain, drowsiness, headache, tinnitus, ataxia, stupor, and rarely, coma and convulsions. Hyperventilation, bronchospasm, and hypotension occur but gastrointestinal haemorrhage and renal failure are rare. These and additional clinical features are summarized in Table 10.

Treatment

Management is supportive.

Phenytoin

Clinical features

Acute overdose of phenytoin results in nausea, vomiting, headache, tremor, cerebellar ataxia, nystagmus, and rarely, loss of consciousness.

Treatment

Gastric lavage may be considered if the patient presents within 1 h after a substantial overdose. Multiple-dose activated charcoal may increase elimination although this has not been confirmed.

Further reading

Evers ML, Izhar A, Aqil A (1997). Cardiac monitoring after phenytoin overdose. *Heart and Lung* **26,** 325–8.

Manto M, Preiser JC, Vincent JC (1996). Hypoglycemia associated with phenytoin intoxication. *Journal of Toxicology—Clinical Toxicology* **34**, 205–8.

Phosgene

Phosgene is a colourless gas, which is now used in the synthesis of isocyanates, polyurethane and polycarbonate resins, dyes, and is produced in fires.

Mechanism of toxicity

Phosgene reacts with glutathione. When glutathione stores become depleted beyond a critical level covalent binding occurs between phosgene and cell macromolecules with resultant hepatic and renal necrosis.

Clinical features

Exposure to phosgene causes irritation of the eyes, dryness or burning sensation in the throat, cough, chest pain, and nausea and vomiting. There is usually a latent period lasting between 30 min and 24 h (rarely, 72 h) during which the casualty suffers little discomfort and has no abnormal chest signs. Subsequently, pulmonary oedema develops due to increased capillary permeability; circulatory collapse may follow.

Treatment

Administration of *N*-acetylcysteine may confer some protection. Oxygen should be administered. Mechanical ventilation may be life-saving in severe cases.

Further reading

Wyatt JP, Allister CA (1995). Occupational phosgene poisoning: a case report and review. *Journal of Accident and Emergency Medicine* **12**, 212–13.

Phosphine

Phosphine is a colourless gas with a fish-like odour and is used as a fumigant against insects and rodents in stored grain, particularly in grain elevators and, increasingly, aboard ships. It is also used to treat silicon crystals in the semiconductor industry.

Clinical features

Fatigue, nausea, vomiting, diarrhoea, chest tightness, breathlessness, productive cough, dizziness, and headache are common features of acute phosphine exposure. Acute pulmonary oedema, hypertension, cardiac arrhy-thmias, and convulsions have been described in severe cases. Ataxia, intention tremor, and diplopia may be found on examination. Focal myocardial infiltration with necrosis, pulmonary oedema, and widespread small-vessel injury were found at autopsy in a child who died.

Treatment

The casualty should be removed from exposure as soon as possible. Thereafter, treatment is supportive and symptomatic. The value of steroids in preventing pulmonary damage (which may be delayed) has not been established.

Further reading

Schoonbroodt D *et al.* (1992). Acute phosphine poisoning? A case report and review. *Acta Clinica Belgica* **47**, 280–4.

Primaquine

Clinical features

Primaquine poisoning is rare, though toxicity is frequent if more than 60 mg is ingested in 1 day. Headache, nausea, abdominal pain, and methaemoglobinaemia may occur and haemolytic anaemia and leucopenia have been observed, especially in patients with glucose-6-phosphate dehydrogenase deficiency.

Treatment

Gut decontamination should be considered if presentation is within 1 h of exposure. If methaemoglobinaemia exceeds 40 per cent, methylthioninium (methylene blue) at 1 to 2 mg/kg body weight should be administered intravenously over 5 min. Antidotal efficacy is NADPH dependent and therefore reduced in the presence of G6PD deficiency

Further reading

Jaeger A *et al.* (1987). Clinical features and management of poisoning due to antimalaria drugs. *Medical Toxicology* **2,** 242–73.

Propylene glycol (1,2-propanediol)

Propylene glycol is used widely as a preservative as a vehicle for both oral and intravenous medications and in preparations used for treating burns.

Mechanism of toxicity

Propylene glycol is oxidized to lactic acid and pyruvate.

Clinical features

The ingestion of substantial quantities of propylene glycol or its administration to neonates or those in renal failure may cause convulsions, coma, cardiac arrhythmias, hepatorenal damage, intravascular haemolysis, metabolic acidosis, and increased serum osmolality.

Treatment

Gastric lavage should be considered if the patient presents within 1 h after ingestion. Metabolic acidosis, renal failure, and respiratory depression should be treated

conventionally. Haemodialysis removes propylene glycol efficiently.

Further reading

Levy ML *et al.* (1995). Propylene glycol toxicity following continuous etomidate infusion for the control of refractory cerebral edema. *Neurosurgery* **37**, 363–71.

Pyrethroids

Clinical features

Pyrethroids are best known for their ability to cause facial paraesthesiae following occupational exposure; these symptoms last only a few hours at most. Coma, convulsions, and pulmonary oedema may occur after substantial ingestion, percutaneous absorption, or inhalational exposure.

Treatment

Symptomatic and supportive measures should be employed and reassurance given that facial paraesthesiae will not be a long-term problem.

Further reading

Kühn K-H *et al.* (1999). Toxicokinetics of pyrethroids in humans: consequences for biological monitoring. *Bulletin of Environmental Contamination and Toxicology* **62,** 101–8.

Müller-Mohnssen H (1999). Chronic sequelae and irreversible injuries following acute pyrethroid intoxication. *Toxicology Letters* **107,** 161–76.

Wilkes MF (2000). Pyrethroid-induced paresthesia—a central or local toxic effect? *Journal of Toxicology—Clinical Toxicology* **38**, 103–5.

Pyridoxine

High doses (more than 2 to 3 g/day) have been used for the treatment of a variety of conditions including the premenstrual syndrome, carpal tunnel syndrome, schizophrenia, and hyperactivity in childhood.

Clinical features

Prolonged daily intake of 50 to 300 mg in women has been reported to cause headaches, irritability, tiredness, shooting pains, circumoral and limb paraesthesiae, numb extremities, clumsiness, and ataxia, indicating a sensory neuropathy.

Treatment

Improvement occurs within 2 months of stopping the drug. There is no specific treatment.

Quinidine and quinine

Though quinidine and quinine are optical isomers, quinine is more oculotoxic and quinidine more cardiotoxic in overdose. Poisoning with quinine is much more common than with quinidine and may be iatrogenic, suicidal, or from attempted abortion or adulterated heroin.

Clinical features

Doses as low as 2 g can be toxic in adults. Cinchonism (tinnitus, deafness, vertigo, nausea, headache, and diarrhoea) is common at plasma concentrations greater than 10 mg/l. In more serious poisoning, collapse with impairment of consciousness (due to ventricular arrhythmias), convulsions, rapid shallow breathing, hypotension, pulmonary oedema, and cardiorespiratory arrest may be observed. Ventricular tachycardia and fibrillation and depression of automaticity and intracardiac conduction are potentially lethal. Pulmonary oedema and acute renal failure have been described. About 40 per cent of patients develop ocular features, which may be unilateral, including blindness, contracted visual fields, scotomas, dilated pupils, blurred disc margins, macular oedema, arteriolar spasm, and late optic atrophy. Oculotoxicity is likely when plasma concentrations exceed 10 mg/l. Visual loss is permanent in about 50 per cent of cases. Hypoglycaemia resulting from insulin release is a common side-effect of quinine and quinidine, even when used in therapeutic doses.

Treatment

Gastric lavage may be indicated if the patient presents within 1 h of ingestion or, alternatively, 50 to 100 g of activated charcoal may be administered. Multiple doses of activated charcoal increase quinine clearance. Forced diuresis, charcoal haemoperfusion, haemodialysis, and stellate ganglion block are of no value. Electrolyte and acid–base disturbances and hypoglycaemia should be corrected. Bradyarrhythmias may respond to isoprenaline; overdrive pacing may be required if *torsade de pointes* occurs. Plasma expanders should be given for hypotension, but if the response is poor, an inotrope should be administered.

Further reading

Mackie MA, Davidson J, Clarke J (1997). Quinine—acute self-poisoning and ocular toxicity. *Scottish Medical Journal* **42**, 8–9.

Nordt SP, Clark RF (1998). Acute blindness after severe quinine poisoning. *American Journal of Emergency Medicine* **16**, 214–15.

Rifampicin

Clinical features

Poisoning with rifampicin results in the so-called 'red man syndrome' that can be fatal. The skin, and subsequently the sclerae, become yellow-orange in colour (the colour of a boiled lobster) and the skin discoloration may be removed by washing. These appearances are due to the intense colour of rifampicin and its metabolites that are distributed throughout the body. In addition, nausea, vomiting, abdominal pain, pruritis, a sensation of the skin burning, and convulsions have been observed. Less commonly, marked oedema of the forehead, cheeks, chin, and lips with associated eosinophilia has occurred. Elevations in serum activities of hepatic enzymes and bilirubin concentration have been noted. Sudden death has also been recorded in two patients, due probably to cardiorespiratory arrest.

Treatment

Treatment is supportive and symptomatic. Gastric lavage may be considered if the patient presents within 1 h of a substantial overdose.

Further reading

Holdiness MR (1989). A review of the red man syndrome and rifampicin overdose. *Medical Toxicology and Adverse Drug Experience* **4,** 444–51.

Salicylates

Despite the introduction of child-resistant packaging, the ingestion of aspirin by children still occurs, iatrogenic overdose is not uncommon, and moreover, aspirin remains the drug of choice for many adults who choose deliberately to poison themselves. Salicylate poisoning may also result from percutaneous absorption of

salicylic acid (used in keratolytic agents), and ingestion of methyl salicylate (oil of wintergreen).

Mechanisms of toxicity

In therapeutic doses, aspirin is absorbed rapidly from the stomach and small intestine, but in overdose, absorption may occur more slowly and the plasma salicylate concentration may continue to rise for up to 6 to 12 h.

The pharmacokinetics of elimination of aspirin are an important determinant in the development of salicylate toxicity. The biotransformation pathways concerned with the formation of salicyluric acid and salicyl-phenolic glucuronide (Fig. 4) are saturable, a fact which has the following clinical consequences: (i) the time needed to eliminate a given fraction of a dose increases with increasing dose; (ii) the steady-state plasma concentration of salicylate, particularly that of the pharmacologically active non-protein-bound fraction, increases more than proportionately with increasing dose; and (iii) as the metabolic pathways of elimination become saturated, renal excretion of salicylic acid becomes increasingly important, a pathway which is extremely sensitive to changes in urinary pH.

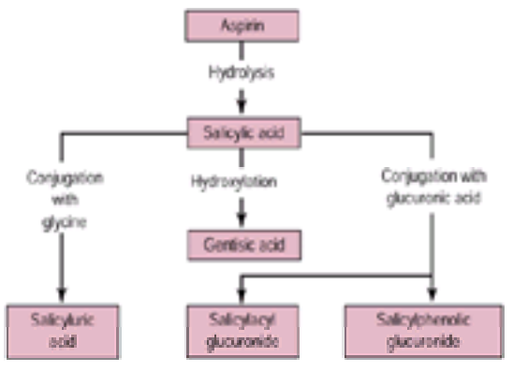

Fig. 4 The principal biotransformation pathways of aspirin.

When ingested in overdose, salicylates directly stimulate the respiratory centre to produce both increased depth and rate of respiration, thereby causing a respiratory alkalosis (Fig. 5). At least part of this effect on the respiratory centre has been shown to be due to local uncoupling of oxidative phosphorylation within the brainstem. In an attempt to compensate, bicarbonate, accompanied by sodium, potassium, and water, is excreted in the urine. Dehydration and hypokalaemia result, but more importantly, the loss of bicarbonate diminishes the buffering capacity of the body and allows an acidosis to develop more easily. A very high salicylate concentration in the brain depresses the respiratory centre and may further contribute to the development of acidaemia.

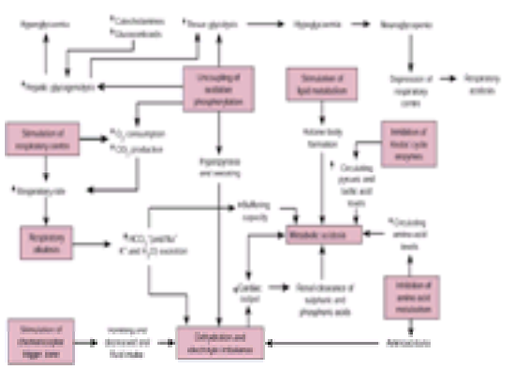

Fig. 5 The pathophysiology of salicylate poisoning.

Simultaneously, a variable degree of metabolic acidosis develops, not only because of the presence of salicylic acid itself, but also because of interference with carbohydrate, lipid, protein, and amino acid metabolism by salicylate ions (Fig. 5). Inhibition of citric acid cycle enzymes causes an increase in circulating lactic and pyruvic acids. Salicylates stimulate fat metabolism and cause increased production of the ketone bodies, b-hydroxybutyric acid, acetoacetic acid, and acetone. Dehydration and lack of food intake because of vomiting further contribute to the development of ketosis. Protein catabolism is accelerated, protein synthesis diminished, and aminotransferases (responsible for the interconversion of amino acids) inhibited. Increased circulating blood concentrations of amino acids result, together with aminoaciduria; this latter feature is further enhanced by inhibition of active tubular reabsorption of amino acids. The aminoaciduria increases the solute load on the kidneys and, thereby, increases water loss from the body.

A primary toxic effect of salicylates in overdose is uncoupling of oxidative phosphorylation (Fig. 5). ATP-dependent reactions are inhibited and oxygen utilization and carbon dioxide production increased. Energy normally used for the conversion of inorganic phosphate to ATP is dissipated as heat. Hyperpyrexia and sweating result causing further dehydration. Fluid loss is enhanced because salicylates stimulate the chemoreceptor trigger zone and induce nausea and vomiting and, thereby, diminish oral fluid intake. If dehydration is sufficiently marked, low cardiac output and oliguria will aggravate the metabolic acidosis already present which, if severe, can itself diminish cardiac output.

Glucose metabolism also suffers as a result of uncoupled oxidative phosphorylation because of increased tissue glycolysis and peripheral demand for glucose (Fig. 5). This is seen principally in skeletal muscle and may cause hypoglycaemia. The brain appears to be particularly sensitive to this effect and neuroglycopenia can occur in the presence of a normal blood sugar level when the rate of utilization exceeds the rate at which glucose can be supplied from the blood. Increased metabolism and peripheral demand for glucose activates hypothalmic centres resulting in increased adrenocortical stimulation and release of adrenaline. Increased glucose-6-phosphatase activity and hepatic glycogenolysis contribute to the hyperglycaemia that is sometimes seen following ingestion of large amounts of salicylate. Increased circulating adrenocorticosteroids exacerbate fluid and electrolyte imbalance.

Although rarely a practical problem, salicylate intoxication may be accompanied by hypoprothrombinaemia due to a warfarin-like action of salicylates on the physiologically important vitamin K_1 epoxide cycle. Vitamin K is converted to vitamin K 2,3-epoxide and then reconverted to vitamin K by a liver membrane reductase enzyme which is competitively inhibited by salicylates (and warfarin).

Clinical features and assessment of severity of salicylate intoxication

The dose of salicylate ingested and the age of the patient (see below) are the principal determinants of the severity of an overdose. The plasma salicylate concentration should be determined on admission, but it is important to repeat it 2 h later to ensure that the concentration is not rising. If the concentration has risen, the level should be repeated after a further 2 h. Generally speaking, plasma salicylate concentrations that lie between 300 and 500 mg/l some 6 h after ingestion of an overdose are associated with only mild toxicity, concentrations between 500 and 700 mg/l are associated with moderate toxicity, and concentrations in excess of 700 mg/l confirm severe poisoning.

Salicylate poisoning of any severity is associated with sweating, vomiting, epigastric pain, tinnitus, and deafness (Table 11).

Young children quickly develop metabolic acidosis following the ingestion of aspirin in overdose, but by the age of 12 years, the usual adult picture of a combined dominant respiratory alkalosis and mild metabolic acidosis is seen. Dehydration and electrolyte imbalance occur early. To some extent, the presence of an alkalaemia protects against serious salicylate toxicity because salicylate remains ionized and unable to penetrate cell membranes easily. Development of acidaemia allows

salicylates to penetrate tissues more readily and leads, in particular, to CNS toxicity characterized by excitement, tremor, delirium, convulsions, stupor, and coma. Very high plasma salicylate concentrations cause paralysis of the respiratory centre and cardiovascular collapse due to vasomotor depression.

Pulmonary oedema is seen occasionally in salicylate poisoning, and although this is often due to fluid overload as a result of treatment, it may be non-cardiac and occur in the presence of hypovolaemia. In these circumstances, the pulmonary oedema fluid has the same protein and electrolyte composition as plasma, suggesting increased pulmonary vascular permeability.

Although aspirin overdose may be complicated by inhibition of platelet aggregation and hypoprothrombinaemia, gastric erosions and gastrointestinal bleeding are rare following acute salicylate overdose.

Oliguria is sometimes seen in patients following the ingestion of salicylates in overdose. The most common cause is dehydration, but rarely, acute renal failure or inappropriate secretion of antidiuretic hormone may occur.

Whilst the urinary pH may be alkaline in the early stages of salicylate overdose, it soon becomes acid. Measurement of arterial blood gases, pH, and standard bicarbonate may show a respiratory alkalosis in the early stages of salicylate intoxication accompanied by the development of a metabolic acidosis. The plasma potassium concentration is often low; rarely, the blood sugar may be high.

Treatment

Gastric lavage and activated charcoal may be considered if the patient presents within 1 h of ingestion. Fluid and electrolyte replacement is particularly important and special attention should be paid to potassium supplementation. Severe metabolic acidosis requires at least partial correction with bicarbonate. Sedatives and respiratory depressant drugs should be avoided because they may hasten the development of acidaemia and CNS toxicity. Mild cases of salicylate poisoning may be managed with either oral or parenteral fluid and electrolyte replacement only.

Patients who exhibit marked symptoms or signs of salicylism and whose plasma salicylate concentration is in excess of 700 mg/l (or lower if acidaemia is present) should receive specific elimination therapy. Urine alkalinization is most often used for this purpose. The pH of the urine during this procedure is of far greater importance than the volume of urine excreted. The urinary pH should be in excess of 7.5 and should ideally lie between 8.0 and 8.5. Rarely, patients prove refractory to urine alkalinization, or this therapy may be contraindicated. Haemodialysis may then prove necessary to remove salicylate from the body. Haemodialysis is the treatment of choice for severely poisoned patients, particularly those with features of CNS toxicity and metabolic acidosis, and has the advantage that it enables simultaneous correction of the acid–base and fluid and electrolyte imbalances.

Pulmonary oedema occasionally complicates salicylate toxicity. Fluid overload should be excluded as far as possible, but if increased pulmonary vascular permeability is suspected, measurement of the pulmonary artery wedge pressure may be needed both for confirmation of the diagnosis and to monitor subsequent fluid administration. Positive and expiratory pressure ventilation appears to be beneficial in this form of pulmonary oedema.

Further reading

Chapman BJ, Proudfoot AT (1989). Adult salicylate poisoning: deaths and outcome in patients with high plasma salicylate concentrations. *Quarterly Journal of Medicine* **72,** 699–707.

Varela N *et al.* (1998). Salicylate toxicity in the older patient. *Journal of Clinical Rheumatology* **4,** 1–5.

Selective serotonin reuptake inhibitors (SSRIs)

Citalopram, fluoxetine, fluvoxamine, paroxetine, and sertraline are new antidepressants that inhibit serotonin reuptake (SSRIs). They lack the anticholinergic actions of tricyclic antidepressants.

Clinical features

Doses of up to 3.6 mg/kg body weight of fluoxetine and fluvoxamine do not appear to cause toxicity and even larger amounts are relatively safe unless potentiated by ethanol. Most patients will show no signs of toxicity but drowsiness, nausea, diarrhoea, and sinus tachycardia have been reported. Rarely, junctional bradycardia, seizures, and hypertension have been encountered and influenza-like symptoms may develop after a day or two.

Treatment

Supportive measures are all that are required. Activated charcoal may reduce absorption if administered within 1 h of overdose, but should only be considered if a substantial overdose has been ingested.

Further reading

Gross R *et al.* (1998). Generalized seizures caused by fluoxetine overdose. *American Journal of Emergency Medicine* **16,** 328–9.

Phillips S *et al.* (1997). Fluoxetine versus tricyclic antidepressants: a prospective multicenter study of antidepressant drug overdoses. *Journal of Emergency Medicine* **15,** 439–45.

Smoke

Smoke consists of a suspension of small particles in hot air and gases that are generated by thermal decomposition and combustion. The particles consist of carbon and are coated with combustion products such as organic acids and aldehydes. The gaseous phase of smoke has an extremely variable composition, depending on the materials involved in the fire. Carbon dioxide and carbon monoxide are always present and usually constitute major components.

Other toxic gases commonly contained in the gaseous phase, though not necessarily in high concentration, include acrolein, ammonia, chlorine, hydrogen bromide, hydrogen chloride, hydrogen cyanide, oxides of nitrogen, phosgene, phosphorus pentoxide, and sulphur dioxide.

Clinical features

The main effects are asphyxia and severe pulmonary irritation and oedema. Smaller particles, acids, and aldehydes cause lacrimation, burning of the throat, and nausea and vomiting when swallowed. Highly water-soluble gases (e.g. hydrogen chloride, sulphur dioxide) cause immediate irritation to the upper respiratory tract, whereas gases with low solubility (e.g. chlorine, nitrogen dioxide, phosgene) penetrate further into the lung and cause injury to the distal airways and alveoli.

Thus, features range from mild irritation of the eyes and upper airways to severe tracheobronchitis, bronchospasm, pulmonary oedema, and bronchopneumonia, which may result in pulmonary insufficiency and death. Laryngitis and laryngeal oedema can also occur and may progress to complete laryngeal obstruction over a period of several hours. Acute hypoxaemia may be associated with the occurrence of frequent ventricular premature beats; tissue hypoxia due to an elevated carboxyhaemoglobin concentration may lead to chest pain and cardiac arrhythmias in subjects with pre-existing ischaemic heart disease.

Treatment

Casualties should be removed from the smoke and resuscitated. Humidified oxygen should be administered, together with a nebulized bronchodilator such as salbutamol if bronchospasm is present. A carboxyhaemo-globin concentration should be obtained and arterial blood gases should be measured.

In the case of burning plastics, the possibility of cyanide poisoning should be considered. Early fibreoptic laryngoscopy or bronchoscopy may assist the diagnosis and

enable the severity of any subglottal injury to be determined. Evidence that corticosteroids protect against pulmonary injury is lacking.

Further reading

Barillo DJ, Goode R, Esch V (1994). Cyanide poisoning in victims of fire: analysis of 364 cases and review of the literature. *Journal of Burn Care and Rehabilitation* **15**, 46–57.

Baud FJ *et al.* (1991). Elevated blood cyanide concentrations in victims of smoke inhalation. *New England Journal of Medicine* **325**, 1761–6.

Hantson P *et al.* (1997). Early complications and value of initial clinical and paraclinical observations in victims of smoke inhalation without burns. *Chest* **111**, 671–5.

Orzel RA (1993). Toxicological aspects of fire smoke: polymer pyrolysis and combustion. *Occupational Medicine* **8**, 415–29.

Shusterman D *et al.* (1996). Predictors of carbon monoxide and hydrogen cyanide exposure in smoke inhalation patients. *Journal of Toxicology—Clinical Toxicology* **34**, 61–71.

Sodium chloride

Poisoning with sodium chloride is uncommon but has occurred accidentally (e.g. addition of salt instead of sugar to infant feeds), as a result of deliberate intent (e.g. as a form of child abuse), or iatrogenically (e.g. use of hypertonic saline in gastric lavage or too rapid administration of saline during treatment of hyponatraemia).

Mechanism of toxicity

An increase in plasma sodium will increase plasma osmolality causing a shift of water from the intracellular to the extracellular space. CNS cell dehydration results in distended cerebral vessels. Subarachnoid, subdural, and intravascular haemorrhages may follow; these changes may be aggravated by overzealous rehydration.

Clinical features

Poisoning with sodium chloride can induce vomiting, increased thirst, anorexia, fever, hypotonia, lethargy, dehydration, peripheral vasoconstriction, irritability, muscular rigidity, convulsions, and coma. Hypernatraemia, increased plasma and urine osmolality, hyperglycaemia, metabolic acidosis, and hypocalcaemia may ensue. In severe cases pulmonary oedema and congestive heart failure may develop.

Treatment

The aim is to lower the serum sodium concentration slowly so that cerebral oedema, pulmonary oedema, convulsions, and coma are not provoked. Five per cent dextrose followed by hypotonic saline solutions should be given intravenously, but in severe cases complicated by renal insufficiency, haemodialysis (peritoneal dialysis is less efficient) should be considered. Hyperglycaemia should not be corrected by insulin as this may induce cerebral oedema.

Further reading

Addleman M, Pollard A, Grossman RF (1985). Survival after severe hypernatremia due to salt ingestion by an adult. *American Journal of Medicine* **78**, 176–8.

Sodium nitroprusside

This vasodilator is converted *in vivo* to nitric oxide and cyanide. Accumulation of cyanide occurs if too high an infusion rate of nitroprusside is used.

Clinical features

Hypotension is the major side-effect of treatment and may be corrected by a change in infusion rate. Metabolic (lactic) acidosis is usually the first indication of cyanide toxicity. Thiocyanate accumulation may lead to anorexia, nausea, lethargy, fatigue, and psychosis.

Treatment

During prolonged infusions the blood cyanide and thiocyanate concentrations should be measured and should not exceed 1 mg/l and 100 mg/l, respectively. The risk of toxicity can be avoided by not exceeding the recommended infusion rates and/or by giving sodium thiosulphate or hydroxocobalamin intravenously. Cyanide toxicity should be treated conventionally (see above).

Further reading

Johanning RJ *et al.* (1995). A retrospective study of sodium nitroprusside use and assessment of the potential risk of cyanide poisoning. *Pharmacotherapy* **15**, 773–7.

Sodium valproate

Clinical features

Sodium valproate in overdose causes impairment of consciousness and respiration. In severe cases, abnormal liver function tests, hyperammonaemia, increased anion gap acidosis, hypocalcaemia, and hypernatraemia have been reported. In addition, optic nerve atrophy, cerebral oedema, non-cardiogenic pulmonary oedema, renal failure, and pancreatitis have been observed.

Treatment

Gastric lavage may be considered if the patient presents less than 1 h after the ingestion of a substantial overdose. Treatment is symptomatic and supportive.

Further reading

Andersen GA, Ritland S (1995). Life threatening intoxication with sodium valproate. *Journal of Toxicology—Clinical Toxicology* **33**, 279–84.

Strychnine

Poisoning with strychnine, an alkaloid from *Strychnos* species (*Loganiaceae*), is now rare. It is no longer used as a medicine but has a continuing role for killing rodents, moles, and other vermin, and and has been used to 'cut' illicit drugs, especially cocaine and heroin. Strychnine is readily absorbed from the gastrointestinal tract and nasal mucosa and is highly toxic.

Clinical features

Systemic effects occur within 30 min of ingestion or inhalation. The threshold for CNS stimulation is lowered with the result that any sensory stimulus may produce violent muscular spasms reminiscent of tetanus. Features include stiffness of the neck and facial muscles, producing trismus and risus sardonicus. Increased muscle tone, hyperreflexia, agitation, restlessness, and convulsions lead to profound lactic acidosis, rhabdomyolysis, and hyperthermia. Painful muscle spasms and convulsions may be provoked by touch, pain, and noise. Opisthotonos may develop. Death usually results from contracture of respiratory muscles. Strychnine blocks glycine-mediated activation of chloride channels, especially in Renshaw cells of the spinal cord, disinhibiting reflex activity and leading to tetanus-like spasms and muscular rigidity.

Treatment

The patient should be kept at rest. Any type of stimulation must be reduced to a minimum and gastric lavage avoided. Oral activated charcoal, in a dose of 50 to 100 g for an adult, adsorbs strychnine and may reduce its absorption if given within 1 h of ingestion.

In severe poisoning, supportive measures to establish and maintain a clear airway and adequate ventilation are of prime importance. Convulsions should be controlled with intravenous diazepam and muscular spasms with neuromuscular blockade (e.g. with pancuronium) and mechanical ventilation. This, together with supplemental oxygen, will help correct metabolic acidosis. Hyperthermia, rhabdomyolysis, and renal failure should be managed conventionally. Patients may recover without apparent sequelae.

Further reading

Boyd RE *et al*. (1983). Strychnine poisoning. Recovery from profound acidosis, hyperthermia and rhabdomyoysis. *American Journal of Medicine* **74**, 507–12

Oberpaur B *et al.* (1999). Strychnine poisoning: an uncommon intoxication in children. *Pediatric Emergency Care* **15**, 264–5.

Styrene (vinyl benzene)

Styrene is a colourless to yellow liquid with a pleasant, sweet odour at low concentrations. Styrene monomer is an important agent in the production of plastic and a stabilizing agent in a variety of products.

Mechanism of toxicity

Styrene is oxidized to styrene oxide that binds covalently to cellular macromolecules, probably due to depletion of glutathione.

Clinical features

Although inhalation is the most common route of exposure, absorption of styrene may also occur through the skin and gut. It is irritant to the eyes, skin, mucous membranes, and respiratory system accompanied by mucous secretion, a metallic taste, drowsiness, and vertigo. Higher concentrations may cause CNS depression.

Treatment

If acute exposure has occurred, the subject should be removed from further exposure, the skin washed, and the eyes irrigated. CNS effects should be treated symptomatically and supportively. Potentially, *N*-acetylcysteine could be of value in preventing hepatic damage.

Further reading

Pahwa R, Kalra J (1993). A critical review of the neurotoxicity of styrene in humans. *Veterinary and Human Toxicology* **35**, 516–20.

Sulphur dioxide

Sulphur dioxide is a colourless gas which has a pungent irritating odour. The combustion of fuels for heating and power generation results in environmental pollution from this cause. Sulphur dioxide is also employed in the manufacture of sulphuric acid and is a potential occupational problem in paper mills, steel works, and oil refineries.

Mechanism of toxicity

The irritant effects of sulphur dioxide are thought to be caused by the rapidity with which it forms sulphurous acid on contact with moist membranes.

Clinical features

Following exposure to sulphur dioxide, lacrimation, rhinorrhoea, cough, increased bronchial secretions, bronchoconstriction, and in severe cases, pulmonary oedema and respiratory arrest occur. Corneal burns can follow eye exposure and liquified sulphur dioxide can cause skin burns. Survivors of massive sulphur dioxide exposure have shown a chronic obstructive defect in serial pulmonary studies along with bronchial hyperactivity.

Treatment

After removal from exposure, admission to hospital for observation is mandatory in severe cases to ensure that delayed pulmonary oedema is treated effectively. Symptomatic and supportive measures should be given and, if necessary, mechanical ventilation with positive end-expiratory pressure should be undertaken if diuretics alone do not control pulmonary oedema; the role of corticosteroids is uncertain. The eyes and skin should be irrigated with water if exposure has occurred.

Further reading

International Programme on Chemical Safety (1979). *Environmental Health Criteria 8. Sulfur oxides and suspended particulate matter.* World Health Organization, Geneva.

Tetrachloroethylene (perchloroethylene)

Tetrachloroethylene is a colourless, non-flammable liquid with a chloroform-like odour. It is used widely as an industrial solvent, particularly for dry cleaning and degreasing. Poisoning may occur by inhalation or ingestion.

Mechanisms of toxicity

A considerable proportion of an inspired dose is exhaled unchanged and that retained is excreted only slowly (half-life approximately 144 h), mainly by metabolism to trichloroacetic acid, the major urinary metabolite.

Clinical features

Following inhalation or ingestion, there is depression of the central nervous system; nausea and vomiting may occur and persist for several hours. Irritation of the eyes, nose, and throat may occur. Hepatic and renal dysfunction may also develop and ventricular arrhythmias and non-cardiogenic pulmonary oedema have been reported.

Treatment

After removal from exposure, treatment is supportive and symptomatic.

Further reading

Garnier R *et al.* (1996). Coin-operated dry cleaning machines may be responsible for acute tetrachloroethylene poisoning: report of 26 cases including one death. *Journal of Toxicology—Clinical*

Toxicology **34**, 191–7.

Theophylline

Poisoning may complicate therapeutic use as well as being the result of deliberate self-poisoning. It is important to establish at an early stage the precise theophylline product involved in a poisoning incident since many of them are sustained-released formulations. As a consequence, peak plasma concentrations of the drug are frequently not attained until 6 to 12 h after overdosage and the onset of toxic features is correspondingly delayed.

Clinical features

Most symptomatic patients have concentrations in excess of 25 mg/l. Convulsions are seen more commonly when concentrations are greater than 50 mg/l.

Symptoms include nausea, vomiting, hyperventilation, haematemesis, abdominal pain, diarrhoea, sinus tachycardia, supraventricular and ventricular arrhythmias, hypotension, restlessness, irritability, headache, hyperreflexia, tremor, and convulsions. Hypokalaemia results predominantly from Na^{+}/K^{+} ATPase activation. A mixed respiratory alkalosis and metabolic acidosis is common.

Assessment of the severity of poisoning

The severity of theophylline intoxication is important in deciding management. Plasma potassium concentrations of less than 2.6 mmol/l, acidaemia, hypotension, seizures, and arrhythmias are indications for urgent measurement of plasma theophylline concentrations. Patients require close observation to detect the onset of delayed toxicity and it may be necessary to repeat measurement of the plasma theophylline concentration a few hours after admission.

Treatment

Gastric lavage may be considered in patients who have ingested a significant overdose of theophylline and who present within 1 h; administration of 50 to 100 g of activated charcoal (by nasogastric tube if necessary) is an alternative. Multiple doses of charcoal (e.g. 50 g 4-hourly) will enhance the systemic elimination of theophylline. Intractable vomiting may be alleviated by ondansetron in a dose of 8 mg intravenously in an adult. Gastrointestinal haemorrhage may require blood transfusion and a proton pump inhibitor should be given intravenously. Tachyarrhythmias may be induced by the rapid flux of potassium across cell membranes and early correction of hypokalaemia may prevent their development. The plasma potassium concentration should therefore be measured on admission and at houry intervals thereafter while the patient is symptomatic. Potassium supplements will be needed in almost all cases and doses of up to 60 mmol/h may be required at the outset. Non-selective b-adrenoceptor blocking drugs, such as propranolol, may also be useful to treat tachyarrhythmias and reverse hypokalaemia. Convulsions should be managed conventionally.

Although charcoal haemoperfusion increases theophylline clearance, it should be reserved for those in whom intractable vomiting or recurrent seizures make oral charcoal impracticable or hazardous.

Further reading

Minton NA, Henry JA (1996). Acute and chronic human toxicity of theophylline. *Human and Experimental Toxicology* **15**, 471–81.

Thyroxine

Acute overdosage with thyroid hormones is uncommon. Not surprisingly, thyroxine is the agent most commonly involved.

Clinical features

Probably only a small percentage of patients who ingest large amounts of thyroid hormones develop features of toxicity. Symptoms develop within a few hours with tri-iodothyronine (T_3) and after 3 to 6 days with thyroxine (T_4). They tend to resolve in about the same time as they take to develop. Mental confusion, agitation, irritability, and hyperactivity with sinus tachycardia, tachypnoea, pyrexia, and dilated pupils are common while atrial fibrillation, sweating, loose stools, and the ocular features of hyperthyroidism are rare. Convulsions developed in one child.

Treatment

Gastric lavage may be considered if more than 2 mg of thyroxine has been ingested within the preceding 1 h; activated charcoal is an alternative. Serum T_4 and T_3 concentrations should be measured in blood taken 6 to 12 h after ingestion since a normal result precludes the possibility of delayed toxicity and allows the patient to be discharged. Those with high T_4 concentrations should be reviewed for evidence of toxicity on the fourth or fifth day after ingestion. Patients who develop toxicity should be given propranolol for 5 days.

Further reading

Hack JB *et al.* (1999). Severe symptoms following massive intentional L-thyroxine ingestion. *Veterinary and Human Toxicology* **41**, 323–6.

Toluene

Toluene has much lower volatility and toxicity than benzene. It is used extensively as a solvent in the chemical, rubber, paint, glue, and pharmaceutical industries and as a thinner for inks, perfumes, and dyes.

Metabolism

Following inhalation or ingestion, toluene is oxidized to benzoic acid then to hippuric acid benzoylglucuronates that are excreted in the urine.

Clinical features

Acute poisoning results in euphoria, excitement, dizziness, confusion, increased lacrimation, headache, nervousness, nausea, tinnitus, ataxia, tremor, and coma.

Chronic exposure

Chronic poisoning may give rise to muscle weakness, abdominal pain and haematemesis, cerebellar abnormalities, optic neuropathy, peripheral neuropathy, altered mental state, dementia, hearing loss, hypokalaemia, and hepatorenal disease, including distal renal tubular acidosis and urinary calculi.

Treatment

If poisoning results from inhalation, whether accidental or intentional as in volatile substance abuse (see below), the patient should be removed from the ontaminated environment. Thereafter, treatment consists of symptomatic and supportive measures.

Further reading

Einav S *et al.* (1997). Bradycardia in toluene poisoning. *Journal of Toxicology—Clinical Toxicology* **35**, 295–8.

Hussain TF, Heidrenreich PA, Benowitz N (1996). Recurrent non-Q-wave myocardial infarction associated with toluene abuse. *American Heart Journal* **131**, 615–16.

1,1,1-Trichloroethane (methyl chloroform)

1,1,1-Trichloroethane is a colourless, non-flammable liquid of high volatility widely used as a solvent in industry, in the office (e.g. typewriter correction fluid), and at home (e.g. waterproofing aerosol products).

Mechanism of toxicity

1,1,1-Trichloroethane has low systemic toxicity because only small amounts of trichloroacetic acid and trichloroethanol are formed. Most of an inhaled dose is expired unchanged. Concomitant ingestion of ethanol is known to enhance toxicity.

Clinical features

Following inhalation of a sufficiently large dose, CNS depression occurs in proportion to the amount inhaled; hepatic and renal dysfunction may also result. Deaths have followed exposure to very high concentrations in unventilated tanks. In such cases death may either be due to CNS depression, culminating in respiratory arrest, or to fatal arrhythmias as a result of myocardial sensitization to circulating catecholamines in the presence of hypoxia. Inhalation of a weather-proofing aerosol containing 96.6 per cent 1,1,1-trichloroethane has been reported to give rise to transient shortness of breath, constricting chest pain, cough, and myalgia.

Treatment

The casualty should be removed from the contaminated environment. Thereafter treatment is symptomatic and supportive.

Further reading

House RA *et al.* (1996). Paresthesias and sensory neuropathy due to 1,1,1-trichloroethane. *Journal of Occupational and Environmental Medicine* **38,** 123–4.

Liss GM, House RA (1995). Toxic encephalopathy due to 1,1,1-trichloroethane. *American Journal of Industrial Medicine* **27,** 445–6.

Trichloroethylene

Trichloroethylene is a colourless, volatile liquid used widely as an industrial solvent, particularly in metal degreasing and extraction processes.

Mechanisms of toxicity

Trichloroethylene is absorbed readily from the gut and through the skin and lungs. Following inhalation, it is excreted unchanged in the breath and metabolized via chloral hydrate to trichloroethanol (by alcohol dehydrogenase) and trichloroacetic acid, which are excreted in the urine.

Clinical features

Following inhalation, ingestion, or dermal absorption, CNS depression occurs with nausea and vomiting, hepatic and renal dysfunction, and death. 'Degreaser's flush' (in which the skin on the face and arms becomes markedly reddened) may occur if ethanol is consumed shortly before or after exposure to trichloroethylene, as the metabolism of trichloroethylene is inhibited. Cranial nerve damage, cerebellar dysfunction, and convulsions have been described.

Treatment

Removal from exposure will reduce CNS depression, and thereafter, whether trichloroethylene has been inhaled, ingested, or absorbed through the skin, treatment is supportive and symptomatic.

Further reading

Szlatenyi CS, Wang RY (1996). Encephalopathy and cranial nerve palsies caused by intentional trichloroethylene inhalation. *American Journal of Emergency Medicine* **14**, 464–7.

Yoshida M *et al.* (1996). Concentrations of trichloroethylene and its metabolites in blood and urine after acute poisoning by ingestion. *Human and Experimental Toxicology* **15**, 254–8.

Tricyclic antidepressants

Tricyclic antidepressants have complex actions that account for the diverse nature of the features seen after overdosage. They block the reuptake of noradrenaline into peripheral and intracerebral neurones, thereby increasing the concentration of monoamines in these areas. They also have anticholinergic actions and class 1 antiarrhythmic (quinidine-like) activity.

Clinical features

Features of poisoning typically appear within 30 to 60 min after ingestion of an overdose and usually reach maximum intensity in 4 to 12 h. Drowsiness, sinus tachycardia, dry mouth, dilated pupils, urinary retention, increased reflexes, and extensor plantar responses are the most common features of mild poisoning. Severe intoxication leads to coma, often with divergent strabismus and convulsions. Plantar, oculocephalic ('doll's head'), and vestibulo-ocular reflexes may be temporarily abolished. Skin blisters and rhabdomyolysis may be present.

Sinus tachycardia is very common and the dose-related quinidine-like action decreases myocardial contractility and delays conduction producing a bizarre ECG; P–R and QRS intervals increase and the P waves diminish in amplitude and may be completely obscured by the preceding T wave. These changes, in conjunction with the increased heart rate, not infrequently make differentiation between ventricular tachycardia and supraventricular tachycardia with aberrant conduction difficult, if not impossible. Serious arrhythmias, particularly ventricular tachycardia, occur in only 4 per cent of cases. In severe cases the blood pressure and cardiac output fall; metabolic acidosis and cardiorespiratory depression are the major contributing factors to death.

Treatment

The great majority of patients poisoned with tricyclic antidepressants recover with supportive therapy alone. Potentially lethal complications such as convulsions and arrhythmias are most common within 6 h of overdosage. It is uncommon for coma to last for more than 24 h and most severely poisoned patients recover consciousness within 48 h.

Gastric lavage may be considered in adults when more than 250 mg of the drug has been ingested less than 1 h previously. Alternatively, 50 to 100 g of activated charcoal can be administered orally.

Management of tricyclic antidepressant-induced cardiotoxicity poses serious difficulties. In general, the natural inclination to use antiarrhythmic drugs to treat tachycardia and arrhythmias should be resisted. Attention to supportive measures, particularly adequate oxygenation, control of convulsions, and correction of acidosis will generally be more rewarding; 50 mmol of sodium bicarbonate intravenously over 20 min should be given even if there is no acidosis. Lignocaine in a dose of 50 to 100 mg intravenously may be tried cautiously if ventricular tachycardia is compromising cardiac output.

Physostigmine salicylate, a cholinesterase inhibitor, has no role. When benzodiazepines have been taken in overdose together with tricyclic antidepressants,

flumazenil may unmask the tricyclic antidepressant-induced seizure potential and should therefore be used with caution.

Forced diuresis and haemodialysis are of no value. Nor is there convincing evidence that charcoal haemoperfusion is effective.

Delirium with auditory and visual hallucinations is a frequent and troublesome complication during the recovery phase. Sedation with oral or intravenous diazepam may be required.

Further reading

Buckley NA *et al.* (1996). Interrater agreement in the measurement of QRS interval in tricyclic antidepressant overdose: implications for monitoring and research. *Annals of Emergency Medicine* **28,** 515–19.

Liebelt EL, Francis PD, Woolf AD (1995). ECG lead aVR versus QRS interval in predicting seizures and arrhythmias in acute tricyclic antidepressant toxicity. *Annals of Emergency Medicine* **26**, 195–201.

Taboulet P *et al.* (1995). Cardiovascular repercussions of seizures during cyclic antidepressant poisoning. *Journal of Toxicology—Clinical Toxicology* **33**, 205–11.

Vinyl chloride (monochloroethylene, chloroethene)

Vinyl chloride is a colourless, highly flammable, and explosive gas. It is usually handled as a liquid under pressure and in this form it polymerizes readily, at temperatures in the range of 40 to 70°C, to form polyvinyl chloride (PVC).

Mechanisms of toxicity

The main route of absorption of vinyl chloride is through the lungs, although some skin penetration does occur. Metabolism to a reactive metabolite appears to be necessary before toxic effects are seen. It has also been postulated that the various features of vinyl chloride disease may have an immunological basis.

Clinical features

Acute exposure

Acute exposure to vinyl chloride results in CNS depression, but concentrations need to be in excess of 10 000 ppm before this effect becomes noticeable. Exposure of volunteers to 20 000 ppm for 5 min caused dizziness, light-headedness, nausea, and dulling of vision.

Chronic exposure

Acro-osteolysis has been described in workers engaged in cleaning autoclaves by hand. The syndrome has three main components: (i) Raynaud's phenomenon, (ii) skin changes resembling scleroderma, and (iii) bony changes of the terminal phalanges of the fingers and sometimes the toes, radial and ulnar styloid processes, sacroiliac joints, and lower poles of patellas.

Angiosarcoma of the liver and hepatic fibrosis, often associated with splenomegaly and portal hypertension, have been reported in vinyl chloride workers.

Exposure to vinyl chloride may also be associated with the development of cancer of the liver and biliary tract and cancer of the brain, though the latter association has not been confirmed in some studies.

Treatment

Preventive measures, adopted worldwide, and designed to protect against angiosarcoma of the liver, should also prevent other adverse effects. Symptoms due to osteo-acrolysis do not improve significantly after removal from exposure, but radiographic improvement of the phalangeal lesions, with recalcification, has been demonstrated.

Further reading

Lewis R (1999). Vinyl chloride and polyvinyl chloride. *Occupational Medicine (Philadelphia)* **14,** 719–42.

Volatile substance abuse

Solvent abuse may be defined as the intentional inhalation of volatile organic chemicals other than conventional anaesthetic gases. These include organic solvents and vapours, hydrocarbon mixtures such as petrol (gasoline), and aerosol propellants.

Volatile substances are either 'bagged' (sprayed into a plastic bag and then inhaled until the subject passes out) or 'huffed' (sprayed on to a cloth held to the mouth). Glue is most often sniffed from a potato crisp bag and repeated abuse in this manner leads to the development of erythematous spots around the mouth and nose ('glue-sniffer's rash').

Clinical features

The clinical features of intoxication with volatile substances are similar to those of alcohol intoxication with initial CNS stimulation followed by depression. Other symptoms may include euphoria, blurring of vision, tinnitus, slurring of speech, ataxia, feelings of omnipotence, headache, abdominal pain, anorexia, nausea, vomiting, jaundice, chest pain, bronchospasm, impaired judgement, irritability, and excitement. Less often a delirious state is seen, with clouding of consciousness and hallucinations. Many chronic users report transient psychotic symptoms that often have an affective component. Convulsions, status epilepticus, and coma may occur. Those under the influence of volatile substances may carry out self-destructive and antisocial acts. Psychological dependence and tolerance may develop, but physical dependence is rare.

Unexplained listlessness, anorexia, and marked moodiness are suggestive of chronic abuse. Poor school adjustment and scholastic performance have been noted in chronic glue sniffers apparently due to lack of motivation.

'Glue sniffing'

Glues are volatile, semiliquid preparations that usually contain an aromatic hydrocarbon as the vehicle. The physical sequelae of prolonged glue sniffing include aplastic anaemia and acute hepatic and renal damage. Features of renal toxicity include proteinuria, haematuria, distal renal tubular acidosis, and recurrent urinary calculi. Irreversible neurological sequelae such as optic atrophy, encephalopathy, cerebellar degeneration, and equilibrium disorders have been reported in adults who are chronic abusers. Toluene inhalation may cause encephalopathy in children. Neurological damage may occur after 'sniffing' of less than 1 year's duration and symptoms may progress for up to 3 months after the habit has been abandoned. Glues containing *n*-hexane and toluene have been associated with the development of muscle weakness and atrophy and sensory impairment of either the 'glove and stocking' or sensorimotor type, with or without muscle atrophy.

A review of adults who had sniffed toluene indicated three major patterns of presentation: (i) muscle weakness, (ii) gastrointestinal complaints (abdominal pain, haematemesis), and (iii) neuropsychiatric disorders (altered mental status, cerebellar abnormalities, and peripheral neuropathy). In addition, hypokalaemia, hypophosphataemia, and hyperchloraemia were common. Rhabdomyolysis occurred in 40 per cent of cases. Cardiac and haematological toxicity due to toluene appears to be uncommon.

Petrol (gasoline) sniffing

Abusers of petrol have reported that 15 to 20 breaths of the vapour are sufficient to produce intoxication for 3 to 6 h.

The euphoria of mild intoxication may be accompanied by nausea and vomiting. After prolonged inhalation, or rapid inhalation of highly concentrated vapour, the 'sniffer' may experience a phase of violent excitement followed by loss of consciousness and coma. While unconscious, the subject may suffer convulsions and the pupils may become fixed and dilated or unequal. Nystagmus and conjugate deviation of the eyes may be observed. Cerebral and pulmonary oedema and renal and hepatic damage have been noted at autopsy. The greater danger from petrol 'sniffing' is related to the long-term effects of chronic exposure that include loss of appetite and loss of weight, neurasthenia, muscle weakness and cramps, and neuropsychological damage. Encephalopathy in petrol sniffers may also be due to tetraethyl lead.

Chlorinated hydrocarbon abuse

Inhalation of chlorinated hydrocarbons causes a sense of euphoria, and sometimes excitement, associated with headache, dizziness, nausea, vomiting, stupor, coma, and convulsions.

Aerosol inhalation

The most commonly abused aerosol propellants are the chlorofluorocarbons (CFCs). Several hundred teenagers have died from this cause. It is likely that the fatalities were due to cardiac arrhythmias.

Diagnosis and treatment

The clinical features described above and the circumstances in which patients are found usually point to the diagnosis, but confirmation may be obtained by detection of solvents in blood or metabolites in the urine.

Acute intoxication from volatile substance abuse is usually brief and self-limiting. If respiratory depression and cardiac arrhythmias supervene, they should be treated conventionally and renal and hepatic failure may require further supportive measures and dialysis.

Further reading

Brady WJ *et al.* (1994). Freon inhalation abuse presenting with ventricular fibrillation. *American Journal of Emergency Medicine* **12**, 533–6.

Cox MJ *et al.* (1996). Severe burn injury from recreational gasoline use. *American Journal of Emergency Medicine* **14**, 39–43.

Martinez JS *et al.* (1989). Renal tubular acidosis with an elevated anion gap in a 'glue sniffer'. *Human Toxicology* **8,** 139–40.

Steffee CH, Davis GJ, Nicol KK (1996). A whiff of death: fatal volatile solvent inhalation abuse. *Southern Medical Journal* **89,** 879–84.

Tenenbein M (1997). Leaded gasoline abuse: the role of tetraethyl lead. *Human and Experimental Toxicology* **16**, 217–22.

Warfarin

Warfarin toxicity is more likely to occur in the setting of therapeutic anticoagulation (as a result of a drug interaction) than as a consequence of acute overdose.

Clinical features

Epistaxis, gingival bleeding, spontaneous bruising, haematomas, haematuria, bilateral flank pain, rectal bleeding, and haemorrhage into any organ may occur. Spontaneous haemoperitoneum has been reported. Severe blood loss may result in hypovolaemic shock, coma, and death.

Treatment

If the intention is to continue anticoagulation

If the INR is less than 6.0 but more than 0.5 units above the target value, reduce the dose or stop warfarin; restart when the INR is less than 5.0. If the INR is 6.0 to 8.0 and the patient is not bleeding (or only minor bleeding is present), stop warfarin and restart when the INR is less than 5.0. If the INR is greater than 8.0 and the patient is not bleeding (or the bleeding is minor), stop warfarin and restart when the INR is less than 5.0. If there are other risk factors for bleeding, give 0.5 mg of vitamin K_1 intravenously slowly. If major bleeding occurs, give 5 mg of vitamin K_1 by slow intravenous injection together with prothrombin complex concentrate at 50 units/kg or, if the complex is not available, give fresh frozen plasma at 15 ml/kg.

If continued anticoagulation is unnecessary

If the INR is greater than 6.0, give 5 mg of vitamin K_1 and repeat if necessary. If active bleeding occurs, give prothrombin complex concentrate at 50 units/kg in addition or, if the complex is not available, give fresh frozen plasma at 15 ml/kg.

Further reading

Baglin T (1998). Management of warfarin (coumarin) overdose. *Blood Reviews* **12,** 91–8.

Xylenes

The three isomers of xylene, which possess similar properties, are used widely as solvents in paints, lacquers, pesticides, gums, resins, adhesives, and the paper-coating industry.

Metabolism

Xylene is oxidized and excreted in the urine either free or conjugated with glycine as methylhippuric acid. Ethanol inhibits xylene metabolism.

Clinical features

Following inhalation, dizziness, excitement, flushing of the face, eye irritation, drowsiness, incoordination, ataxia, tremor, confusion, coma, respiratory depression, and catecholamine-induced ventricular arrhythmias may occur. Hepatorenal damage also has been described. Immersion in liquid xylene may result in a burning feeling, erythema, and some scaling of the skin.

Treatment

Treatment is supportive.

Further reading

Ansari EA (1997). Ocular injury with xylene—a report of two cases. *Human and Experimental Toxicology* **16**, 273–5.

Hageman G *et al.* (1999). Parkinsonism, pyramidal signs, polyneuropathy, and cognitive decline after long-term occupational solvent exposure. *Journal of Neurology* **246,** 198–206.

Zinc

Zinc oxide fumes are emitted in any process involving molten zinc and are the most common cause of metal fume fever. Exposure to zinc chloride occurs in soldering, in the manufacture of dyes, paper, and deodorants, and on military exercises when it is used as a smoke screen.

Poisoning has followed the accidental or deliberate ingestion of elemental zinc and zinc chloride and fatal intoxication has followed inadvertent intravenous administration. Inhalation of zinc chloride and oxide may lead to nasopharyngeal and respiratory toxicity.

Zinc may be absorbed through broken skin when zinc oxide paste is used to treat wounds and burns.

Clinical features

Acute poisoning

Zinc sulphate ingestion causes gastrointestinal irritation, sometimes in association with headache and dizziness. Zinc chloride is highly corrosive and ingestion has led to erosive pharyngitis, oesophagitis, and haematemesis. Acute renal failure and pancreatitis have also been recorded after ingestion of zinc salts. Topical exposure to zinc chloride causes ulceration and dermatitis of the exposed skin. Zinc chloride is highly irritant to the eye.

Metal fume fever starts up to 24 h after exposure to zinc oxide fumes. It presents as an influenza-like illness with headache, fever, sweating, chest tightness and discomfort, and joint pains. Typically symptoms appear after the weekend. The illness usually has an excellent prognosis and the symptoms often improve towards the end of the working week as some short-term immunity from further symptoms develops.

In contrast to the relatively mild clinical course after zinc oxide inhalation, exposure to zinc chloride ammunition bombs (hexite) produces a chemical pneumonitis with marked dyspnoea, a productive cough, fever, chest pain, and cyanosis. The adult respiratory distress syndrome may ensue in the most severe cases; fatalities have been reported.

Chronic poisoning

Repeated topical exposure to zinc oxide may cause a papular folliculitis. Chronic excessive ingestion of zinc supplements (zinc sulphate) may induce reversible anaemia and leucopenia secondary to a relative copper deficiency.

Treatment

Symptomatic and supportive measures should be employed with treatment as for acids as appropriate (see above).

Further reading

Barceloux DG (1999). Zinc. *Journal of Toxicology—Clinical Toxicology* **37**, 279–92

Hantson P, Lievens M, Mahieu P (1996). Accidental ingestion of a zinc and copper sulfate preparation. *Journal of Toxicology—Clinical Toxicology* **34**, 725–30.

Hjortso E *et al.* (1988). ARDS after accidental inhalation of zinc chloride smoke. *Intensive Care Medicine* **14**, 17–24.

8.2 Injuries, envenoming, poisoning, and allergic reactions caused by animals

D. A. Warrell

Mechanical injuries caused by animals

Lions, tigers, leopards, jaguars, hyenas, wolves, dingoes, bears, elephants, hippopotamuses, buffaloes, rhinoceroses, musk oxen, wild pigs, and ostriches have mauled and killed humans. About 100 shark attacks are reported each year between latitudes 46 °N and 47 °S, half of which are fatal (Plate 1). Other fish capable of causing life-threatening mechanical trauma are barracuda, moray and conger eels, garfish, groupers, stingrays, and piranãs. Tiny Amazonian catfish (genus *Vandelia*; Spanish—'canero'; Portuguese—'candirú') are the only vertebrate parasites of humans. Attracted by urine, these 5-cm long fish may enter the urethra, vulva, or anus of swimmers causing pain, bleeding, and obstruction. Their spines make them difficult to remove. The 'electric eel' (*Electrophorus electricus*) of South American rivers can discharge up to 650 volts, 1 amp, 400 times per second, a shock capable of killing an adult. Marine torpedo rays can produce a dangerous 80-volt, high-amperage shock. Crocodiles (*Crocodilus niloticus*) kill about 1000 people each year in Africa, while in northern Australia 27 deaths from 60 attacks by the salt-water crocodile (*C. porosus*) have been reported since 1876. Giant pythons have attacked, killed, and even swallowed people in Indonesia (*Python reticulatus*) (Plate 2), Africa (*P. sebae*), and South America (*Eunectes murinus*). Collisions between vehicles and deer cause more than a 100 injuries each year in Kentucky and are also common in other areas. In the United States, injuries to horseback riders result in more than 46 000 visits to casualty departments each year and at least 20 deaths.

Bites by domestic dogs are common. An estimated 6 million dogs live in England and Wales. More than 200 000 patients bitten by dogs attend hospital each year. About 600 000 people are bitten by dogs each year in the United States. Other domestic animals which have caused severe injuries or deaths include camels, cattle, water buffalo, sheep, pigs, cats, and ferrets.

Treatment

First aid of severe injuries involves controlling bleeding, closing perforating injuries with pressure dressings, and rapid evacuation to hospital. All injuries inflicted by animals must be assumed to be infected by a range of organisms. Wounds may contain teeth and other foreign bodies and necrotic tissue, especially if treatment is delayed.

Wounds should be thoroughly cleaned with soap and water as soon as possible; suitable antiseptics include iodine and alcohol solutions. Prophylactic antimicrobials such as amoxicillin/clavulanic acid, doxycycline, or erythromycin have proved effective in dog- and cat-bite wounds and are indicated for multiple or severe wounds and bites on the face and hands.

Specific infections, such as tetanus, rabies, and herpes simiae virus must be considered and treated/prevented appropriately. Emergency surgery may be required, with: replacement of blood loss; attention to local mechanical complications such as fractures, tension pneumothorax, damage to large blood vessels, perforation of the bowel, and lacerations of other abdominal viscera; and thorough débridement or amputation of dead tissue with removal of foreign material, teeth, etc.; irrigation and drainage. Except for wounds on the head and neck, which can be sutured immediately, primary suturing should be delayed for 48 to 72 h, after which further débridement, suturing, or covering with split skin grafts should be considered. In these cases infection must be prevented by using a combination of penicillin, an aminoglycoside (such as gentamicin for 48 h), and metronidazole.

Venomous animals

For predation or defence, some animals inject venoms through fangs, chelicerae (venom jaws), stings, spines, hairs, nematocysts, and other specialized venom organs. 'Spitting' snakes, scorpions, and millipedes squirt venom on to absorbent mucous membranes. The flesh or skin of some animals contain poisons acquired through the food chain. Allergic reactions to injected venoms (for example, of *Hymenoptera* and cnidarians) and ingested poisons (for example, ciguatera) may create medical problems more commonly than their direct toxic effects.

Venomous mammals

Male duck-billed platypuses (*Ornithorhynchus anatinus*) have erectile venomous spurs on their hind limbs. These aquatic, egg-laying mammals of eastern Australia sting at least one person each year in Victoria, but only 17 cases have been reported since 1817. There is immediate, agonizing, persistent local pain, as well as prolonged local swelling, chronic pain on movement, hyperaesthesia, wasting, inflammation, and regional lymphadenopathy. These effects are not life-threatening in humans, but dogs have died of envenoming. In the absence of specific treatment, non-steroidal anti-inflammatory agents or corticosteroids have proved effective. The venom contains a C-type natriuretic peptide (which causes mast-cell degranulation), nerve growth factor, four defensin-like peptides, enzymes, and other peptides and proteins. Males of the echidna, the other egg-laying mammal, possess a similar but smaller venom apparatus. Several species of Insectivora produce venomous saliva conducted into bite wounds by curved and sometimes grooved lower incisors. Venomous species include the Hispaniolan and Cuban solenodons (*Solenodon paradoxus, S. (Apotogale) cubanus*), northern water shrew (*Neomys fodiens*), southern water shrew (*N. anomalus*), and the North American short-tailed shrew (*Blarina brevicauda*). Their bites can kill rodents and lagomorphs, but in humans the effect is local pain, swelling, and inflammation. The saliva of vampire bats (Desmodontinae) contains permeability increasing factors, a platelet inhibitor, draculin—an inhibitor of activated factors X and IX, and a plasminogen activator which is being developed as a thrombolytic drug.

Venomous snakes

Fewer than 200 species of venomous snake (families Colubridae, Atractaspididae, Elapidae, and Viperidae) have been responsible for severely envenoming humans resulting in death or permanent disability. Since it may be difficult to distinguish venomous from non-venomous species, unnecessary contact with snakes should be avoided and patients bitten by any species should be assessed carefully.

Distribution of venomous snakes

Free from venomous snakes are the Antarctic, most islands of the western Mediterranean, Atlantic, Caribbean, and Eastern Pacific (including Hawaii), Madagascar, New Caledonia, New Zealand, Ireland, Iceland, and the Atlantic Ocean, and Chile. Elsewhere, venomous snakes are widely distributed up to altitudes of more than 4000 m in the Himalayas (*Gloydius himalayanus*), within the Arctic Circle (*Vipera berus*), in the Indian and Pacific Oceans as far north as Siberia (*Pelamis platurus*), and in some freshwater lakes (*Hydrophis semperi*).

Classification

Medically important species have in their upper jaws one or more pairs of enlarged teeth (fangs) that inject venom into their victims through a groove or closed channel.

Colubridae

The short, immobile fangs are at the back of the maxilla (Fig. 1). Most familiar non-venomous snakes, such as the British grass snake and smooth snake, belong to this large family. However, some species have caused severe envenoming or death, including: three African species—the boomslang (*Dispholidus typus*) and the vine, twig, bird, or tree snake or Voëlslang (*Thelotornis kirtlandii* and *T. capensis*); the Japanese yamakagashi (*Rhabdophis tigrinus*); the Southeast Asian red-necked keelback (*R. subminiatus*); the Australasian brown tree snake (*Boiga irregularis*) introduced to Guam; and the South American green racer (*Philodryas olfersii*) (Fig. 1).

Fig. 1 Back fangs of the green racer (*Philodryas olfersii*), a South American colubrid snake. A case of fatal envenoming has been reported from Brazil. (Copyright D.A. Warrell.)

Atractaspididae

The African and Middle Eastern burrowing asps, stiletto snakes, burrowing or mole vipers or adders, strike sideways, impaling their victims on a long front fang

protruding through the partially closed mouth (Fig. 2). Several species, including *Atractaspis microlepidota, A. engaddensis,* and *A. irregularis* have killed humans.

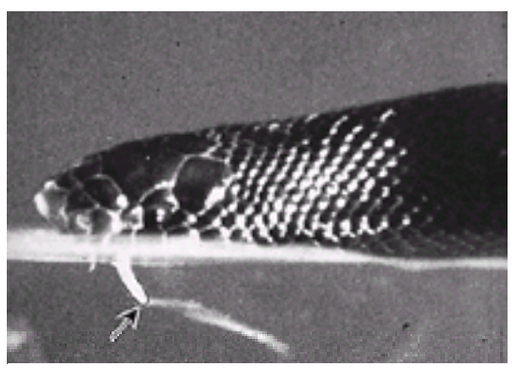

Fig. 2 Burrowing asp (*Atractaspis engaddensis*) showing side-stabbing action with one of the long fangs (arrow) (in this case through a membrane). (Photograph by courtesy of Dr I. Golani and Professor E. Kochva, Tel Aviv.)

Elapidae

This family includes cobras (Fig. 3), kraits, mambas, shield-nose snakes, coral snakes, garter snakes, venomous Australasian snakes (Fig. 4), and sea snakes (Fig. 5). The short, front fangs are immobile (Fig. 3(a) and Fig. 5). Several African and Asian species (rinkhals and spitting cobras) can eject their venom from the tips of the fangs as a fine spray for a distance of a few metres into the eyes of an enemy.

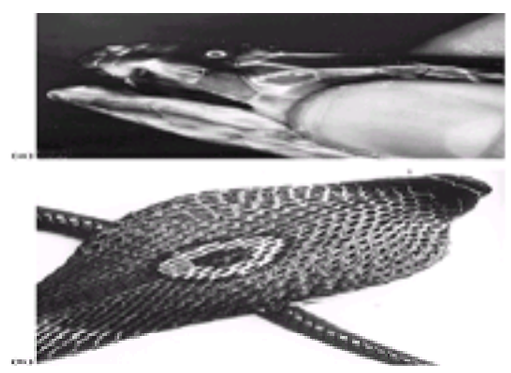

Fig. 3 Cobras. (a) Short front fang of the Sri Lankan cobra (*Naja naja*) a typical elapid snake. (b) Hood of monocellate Thai cobra (*Naja kaouthia*). (Copyright D.A.Warrell.)

Fig. 4 Papua New Guinean taipan (*Oxyuranus scutellatus canni*), an Australasian elapid snake. (Copyright D.A. Warrell.)

Fig. 5 Short front fangs of the laticaudine sea snake (sea krait) *Laticauda colubrina*. (Copyright DA Warrell.)

Viperidae

The front fangs are long, curved, and capable of a wide range of movement (Fig. 6). The subfamily Crotalinae comprises the American rattlesnakes (Fig. 7), moccasins, lance-headed vipers, and Asian pit vipers which possess a heat-sensitive pit organ behind the nostril (Plate 3). The Old World vipers and adders (subfamily *Viperinae*) have no pit organ.

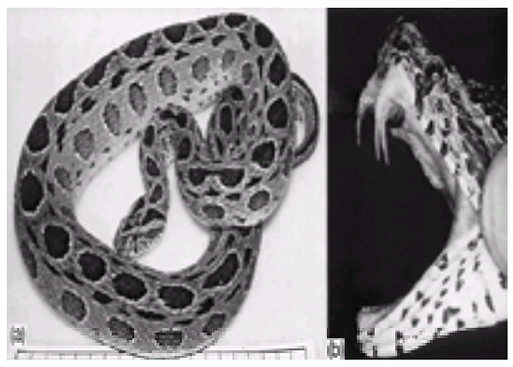

Fig. 6 Russell's vipers. (a) Thai Russell's viper (*Daboia russelii siamensis*), a typical viperine snake (scale in cm). (b) Long, hinged front fangs (reserve fang on the left side) in dental sheath. (Copyright D.A. Warrell.)

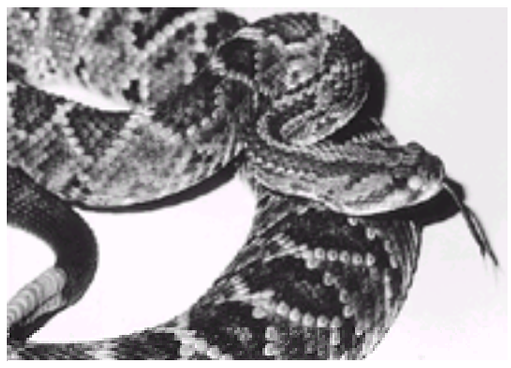

Fig. 7 South American tropical rattlesnake or cascabel (*Crotalus durissus cascavella*). (Copyright D.A. Warrell.)

Incidence and importance of snake bites

Snake bite is an important medical emergency in some parts of the rural tropics; its incidence is usually underestimated because most victims seek the help of traditional healers rather than practitioners of western-style medicine. In a rural population in Kenya, snake bites cause 6.7 deaths per 100 000 per year, 0.7 per cent of all deaths; it was found that 68 per cent of bitten people had sought treatment from traditional healers. In Africa, the saw-scaled or carpet viper (*Echis* spp.), puff adder (*Bitis arietans*), and spitting cobra (*Naja nigricollis, N. mossambica*, etc.) are the species of greatest medical importance. In the Benue Valley of NE Nigeria, *E. ocellatus* (Fig. 8) causes some 500 bites per 100 000 population per year, with a 12 per cent mortality. Vipers of the genus *Echis*, whose geographical range extends through the northern third of Africa, the Middle East, and Eastern Asia to India, are responsible for many bites and deaths. In India, the most important species are cobras (*Naja naja, N. oxiana, N. kaouthia*) (Fig. 3), common krait (*Bungarus caeruleus*), Russell's viper (*Daboia russelii*) (Fig. 6), and *E. carinatus*. An annual snake-bite mortality of 30 000 has been suggested. In Burdwan District, West Bengal, 8000 people are bitten and 800 die each year. In Southeast Asia, the Malayan pit viper (*Calloselasma rhodostoma*), *D. russelii*, green pit vipers (e.g. *Trimeresurus albolabris*) (Plate 3), and cobras (*N. kaouthia* and *N. siamensis*) cause most bites and deaths. In Burma, Russell's viper bite is the leading cause of acute renal failure and is responsible for most of the estimated 1000 snake-bite deaths each year. In Central and South America, medically important species include rattlesnakes (e.g. *C. durissus terrificus*) (Fig. 7) and the lance-headed vipers, *Bothrops atrox* ('barba amarilla' or 'fer de lance'), *B. asper* ('terciopelo'), and *B. jararaca* ('jararaca'). In the United States, 1370 bites were reported to Poisons Centers during 1995, but with only one death. Deaths are caused by rattlesnakes, especially eastern and western diamond-backs (*C. adamanteus* and *C. atrox*). In the Amami and Ryukyu islands of Japan, the habu (*T. flavoviridis*) inflicted an average of 610 bites with 5.6 deaths per year during the 1960s. In Britain, the adder or viper (*Vipera berus*) is the only venomous species (Fig. 9). More than 200 people are bitten each year but only 14 deaths have been reported since 1876, the last in 1975. In Sweden, this species causes between 150 and 200 hospital admissions each year: 44 deaths occurred between 1911 and 1978; and in Finland, 21 deaths in 25 years with an annual incidence of almost 200 bites. *V. aspis* causes most bites in France, while *V. ammodytes* is important in eastern Europe.

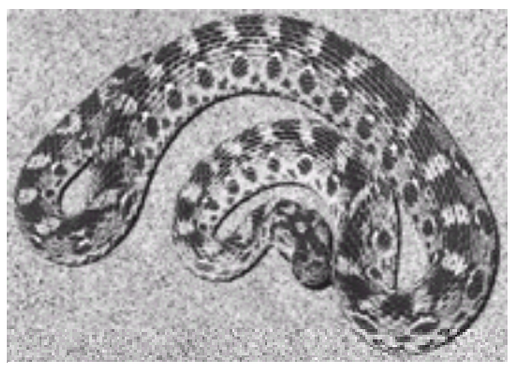

Fig. 8 Saw-scaled or carpet viper from West Africa (*Echis ocellatus*). (Copyright D.A. Warrell.)

Fig. 9 European adder or viper (*Vipera berus*), Britain's only venomous snake. This specimen is 50 cm long. (Copyright D.A. Warrell.)

Australia harbours the deadliest snakes in the world, judging by the lethal potency of their venoms. However, only 3 or 4 people die each year from a snake bite. The most important species are the eastern brown snake (*Pseudonaja textilis*), tiger snake (*Notechis scutatus*), taipan (*Oxyuranus scutellatus*) (Fig. 4), and death adder (*Acanthophis* spp.). The highest snake-bite mortalities, up to 24 per cent of all adult deaths, are recorded among the hunter–gatherer tribes of Brazil (Kashinawa), Venezuela (Yanomamo), Ecuador (Waorani), Tanzania (Hadza), and Papua New Guinea.

Epidemiology

Most snake bites are inflicted on the lower limbs of farmers, plantation workers, herdsmen, and hunters in rural areas. The snake is usually trodden on at night or in undergrowth. Some species such as the Asiatic kraits (*Bungarus* spp.) and African spitting cobras (*N. nigricollis*) enter human dwellings at night and may bite people who roll over on to them while sleeping on the floor. Snakes do not bite without provocation, but may strike if inadvertently trodden upon or touched. In Europe, North America, and Australia, snakes are increasingly popular 'macho' pets: in these countries many bites are inflicted on the hands of males who are picking up the snake. In the United States, 25 per cent of bites result from snakes being attacked or handled. Serious bites by back-fanged (colubrid) snakes usually occur only under these conditions. Seasonal peaks in the incidence of snake bite are associated with agricultural activities, such as ploughing before the annual rains in the West African Sahel and the rice harvest in Southeast Asia, or to fluctuations in the activity or population of venomous snakes. Severe flooding, by concentrating the human and snake populations, has given rise to epidemics of snake bite in Colombia, Pakistan, India, Bangladesh, Nepal, Burma, and Vietnam. Invasion of virgin jungle during construction of new highways and irrigation and hydroelectric schemes has led to an increased incidence of snake bite in Brazil and Sri Lanka. Snake bite or injection of snake venom has long been used for murder and suicide. Snake venoms or purified toxins have been used therapeutically.

Venom apparatus

The venom glands of Elapidae and Viperidae are situated behind the eye, surrounded by compressor muscles (Fig. 10). A venom duct opens within the sheath at the base of the fang and venom is conducted to its tip through a canal. In Colubridae, venom secreted by Duvernoy's gland tracks down grooves in the anterior surfaces of fangs at the posterior end of the maxilla (Fig. 1). The average dry weight of venom injected at a strike is approximately 60 mg in *N. naja*, 13 mg in *E. carinatus*, 63 mg in *D. russelii*, and 32 mg in *V. palaestinae*. The amount injected when a snake bites humans is very variable. A proportion of bites are 'dry bites', associated with negligible envenoming: more than 50 per cent of those bitten by Malayan pit vipers (*C. rhodostoma*) or Russell's vipers; less than 10 per cent bitten by *Echis* spp.; but more than 75 per cent bitten by common brown snakes in Australia (*Pseudonaja textilis*). The Palestine viper (*V. palaestinae*) expends only about one-tenth of the capacity of its venom gland at each consecutive strike, whereas *D. russelii* exhausts more than three-quarters of its supply at the first strike. There is no support for the popular belief that snakes are less dangerous after they have eaten.

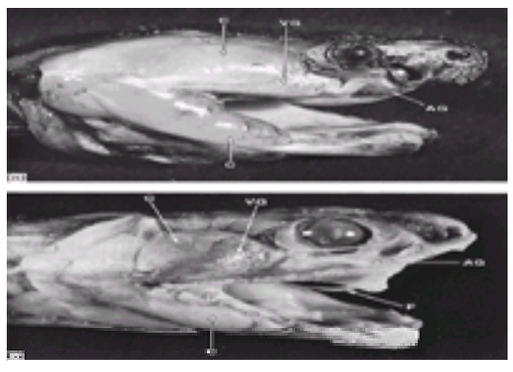

Fig. 10 Venom apparatus of viperine and crotaline snakes. (a) Venom gland of Palestine viper (*Vipera palaestinae*). (b) Venom gland of Western rattlesnake (*Crotalus viridis*). C, compressor glandulae muscle; VG, venom gland; AG, accessory gland; F, fang. (Dissection by Professor E. Kochva, reproduced from Gans, C. and Gans, K.A. (eds), 1978, *Biology of the reptilia*, vol 8. Academic Press, London, by permission.)

Venom properties

Snake venoms may contain 20 or more components. More than 90 per cent of the dry weight of venom is protein, in the form of enzymes, non-enzymatic polypeptide toxins, and non-toxic proteins such as nerve growth factor. Between 80 and 90 per cent of viperid and 25 to 70 per cent of elapid venom consists of enzymes, including digestive hydrolases, hyaluronidase, and activators or inactivators of physiological processes. Most venoms contain L-amino acid oxidase, phosphomono- and diesterases, 5'-nucleotidase, DNAase, NAD-nucleosidase, phospholipase A_2, and peptidases. Elapid venoms, in addition, contain acetylcholine esterase, phospholipase B, and glycerophosphatase, while viperid venoms have endopeptidase, arginine ester hydrolase, kininogenase, as well as thrombin-like, factor X, and prothrombin-activating enzymes. Phospholipase A_2 (lecithinase) is the most widespread and extensively studied of all venom enzymes. It damages mitochondria, red blood cells, leucocytes, platelets, peripheral nerve endings, skeletal muscle, vascular endothelium, and other membranes, produces presynaptic neurotoxic activity, opiate-like sedative effects, and the autopharmacological release of histamine. The acetylcholinesterase found in most elapid venoms does not contribute to their neurotoxicity. Hyaluronidase promotes the spread of venom through tissues. Proteolytic enzymes (endopeptidases or hydrolases) are responsible for local changes in vascular permeability leading to oedema, blistering, and bruising, and to necrosis. L-amino acid oxidase, which gives yellow snake venoms their colour, is a digestive enzyme.

Polypeptide toxins (neurotoxins)

Postsynaptic (a) neurotoxins such as a-bungarotoxin and cobrotoxin, contain about 60 to 62 or 66 to 74 amino acids. They bind to acetylcholine receptors at the motor endplate. Presynaptic (b) neurotoxins such as b-bungarotoxin, crotoxin, and taipoxin, contain about 120 to 140 amino-acid residues and a phospholipase A subunit. These release acetylcholine at the nerve endings at neuromuscular junctions and then damage the endings, preventing further release of transmitter.

Pharmacology

The neurotoxins of the Elapidae are rapidly absorbed into the bloodstream, whereas the much larger molecules of Viperidae venoms are taken up more slowly through the lymphatics. Venoms of the spitting cobras and rinkhals can be absorbed through the intact cornea, causing systemic envenoming and even death in animals. Envenoming after ingestion of snake venom has not been reported in humans. Most venoms are concentrated and bound in the kidney and some components are eliminated in the urine. Crotaline venoms are selectively bound in the lungs, concentrated in the liver, and excreted in bile, while polypeptide neurotoxins, such as a-bungarotoxin, are tightly bound at neuromuscular junctions. Most venom components do not cross the blood–brain barrier.

Pathophysiology

Swelling and bruising of the bitten limb result from increased vascular permeability induced by proteases, phospholipases, membrane-damaging polypeptide toxins, and endogenous autacoids released by the venom, such as histamine, 5-hydroxytryptamine, and kinins. Venoms of some of the North American rattlesnakes and viperine species cause a generalized increase in vascular permeability resulting in hypovolaemia, haemoconcentration, hypoalbuminaemia, albuminuria, serous effusions, pulmonary oedema, and, in the case of Burmese *D. russelii*, conjunctival and facial oedema (Fig. 11). Tissue necrosis near the site of the bite is caused by myotoxic and cytolytic factors: in some cases, ischaemia resulting from thrombosis, intracompartmental syndrome, or a tight tourniquet may contribute. Causes of hypotension and shock include hypovolaemia, vasodilatation, and myocardial dysfunction. Some venoms release vasodilating autacoids such as histamine and kinins. Venom of the Brazilian jararaca (*B. jararaca*) was found to activate bradykinin and, through a bradykinin-potentiating peptide, to prolong its hypotensive effect by inactivating the peptidyl dipeptidase responsible both for destroying bradykinin and for converting angiotensin I to angiotensin II. This observation led to the synthesis of angiotensin-converting enzyme (**ACE**) inhibitors. Bradykinin-potentiating and ACE-inhibiting peptides have also been found in a number of other crotaline venoms (genera *Bothrops* and *Agkistrodon*). To date, four sarafotoxins have been isolated from the venom of the Israeli burrowing asp (*Atractaspis engaddensis*) (Fig. 2). They show 60 per cent sequence homology with the endothelins, which are also 21-amino acid polypeptides. Sarafotoxins and endothelins are potent vasoconstrictors (including coronary arteries), delay atrioventricular conduction, and are positively inotropic.

Fig. 11 Gross bilateral conjunctival oedema (chemosis) in a Burmese rice farmer 48 h after being bitten by a Russell's viper. (Copyright D.A. Warrell.)

Snake venoms can cause haemostatic defects in a number of different ways: venom procoagulant enzymes, many of them serine proteases, activate the blood clotting cascade at various sites. Some Viperidae venoms contain thrombin-like fibrinogenases which remove fibrinopeptides from fibrinogen directly. Others activate endogenous plasminogen. Venoms may induce or inhibit platelet aggregation. Spontaneous systemic bleeding is caused by haemorrhagins, metalloendopeptidases, some with disintegrin-like and other domains, which damage vascular endothelium (Fig. 12). The combination of consumption coagulopathy, thrombocytopenia, and vessel wall damage can result in massively incontinent bleeding, a common cause of death after bites by Viperidae, Australasian Elapidae, and the few medically important Colubridae. Many venoms are haemolytic *in vitro*, but clinically significant intravascular haemolysis, apart from the microang-iopathic haemolysis associated with disseminated intravascular coagulation, is seen only after bites by *D. russelii* (Sri Lanka and India), and some *Bothrops* and colubrid species. Acute renal tubular necrosis may be caused by severe hypotension, disseminated intravascular coagulation (*D. russelii*), a direct nephrotoxic effect of the venom (*D. russelii*), and myoglobinuria secondary to generalized rhabdomyolysis (sea snakes, *D. russelii* in Sri Lanka and India, and tropical rattlesnakes). Neurotoxic polypeptides and phospholipases block neuromuscular transmission causing death through bulbar or respiratory paralysis.

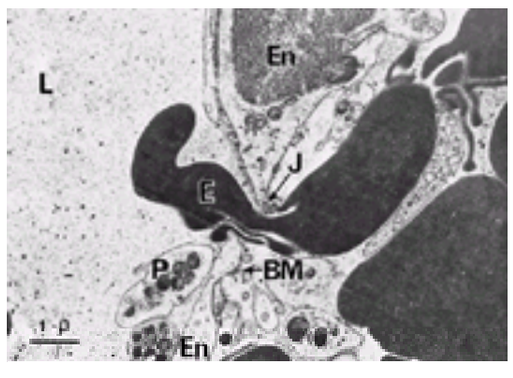

Fig. 12 Haemorrhagin activity. An erythrocyte (E) spurting through an open endothelial junction (J) between endothelial cells (En) from the lumen (L) of a rat mesenteric blood vessel, 5 min after exposure to habu (*Trimeresurus flavoviridis*) venom. Note extensive destruction of the basement membrane (Bm), and failure of the platelet (P) to undergo viscous metamorphosis. (By courtesy of Dr A. Ohsaka and Academic Press, reproduced from Ohsaka A, Suzuki K, Ohashi M (1975). *Microvascular Research* **10**, 208).

Clinical features

Fear and treatment effects as well as the venom contribute to the symptoms and signs of snake bite. Even patients who are not envenomed may feel flushed, dizzy, and breathless and may notice constriction of the chest, palpitations, sweating, and acroparaesthesiae. Tight tourniquets may produce congested and ischaemic limbs; local incisions at the site of the bite may cause bleeding, and sensory loss and herbal medicines often induce vomiting. The earliest symptoms directly attributable to the bite are local pain and bleeding from the fang punctures, followed by pain, tenderness, swelling and bruising extending up the limb, lymphangitis, and tender enlargement of regional lymph nodes. Early syncope, vomiting, colic, diarrhoea, angio-oedema, and wheezing may follow bites by some snakes (e.g. European Vipera, *D. russelii*, *Bothrops* spp., Australian elapids, and *Atractaspis engaddensis*). Nausea and vomiting is a common early symptom of systemic envenoming.

Bites by Colubridae (back-fanged snakes)

Severe envenoming causes repeated vomiting, colicky abdominal pain, headache, widespread systemic bleeding with extensive ecchymoses, incoagulable blood, intravascular haemolysis, and renal failure. Local swelling and bruising may be the only results of envenoming. The first symptoms of envenoming may be delayed for 24 to 72 h after the bite.

Bites by Atractaspididae (burrowing asps or stiletto snakes)

Local effects are pain, swelling, blistering, necrosis, and tender enlargement of local lymph nodes. Violent gastrointestinal symptoms (nausea, vomiting, and diarrhoea), anaphylaxis (dyspnoea, respiratory failure), and electrocardiogram (**ECG**) changes (atrioventricular block, ST, T-wave changes) have been described in patients envenomed by *A. engaddensis*.

Bites by Elapidae (cobras, kraits, mambas, African garter snakes, coral snakes, and Australasian snakes)

Bites by kraits, mambas, coral snakes, and some cobras (e.g. *N. haje*, *N. nivea* and *N. philippinensis*) produce minimal local effects, but the venoms of African spitting cobras (*N. nigricollis*, *N. mossambica*, etc.) and Asian cobras (*N. naja, N. kaouthia, N. sumatrana*, etc.) cause tender local swelling, blistering, and superficial necrosis which may be extensive (Fig. 13). 'Skip' lesions, separated by apparently normal areas of skin, are a typical finding (Fig. 14). However, elapid venoms are best known for their neurotoxic effects. Early symptoms, before there are objective neurological signs, include vomiting, 'heaviness' of the eyelids, blurred vision, paraesthesias around the mouth, hyperacusis, headache, dizziness, vertigo, hypersalivation, congested conjunctivas, and 'gooseflesh'. Paralysis is first detectable as ptosis and external ophthalmoplegia appearing as early as 15 min after the bite, but sometimes it is delayed for 10 h or even more than 24 h following death-adder (Acanthophis) bites. Later the face, palate, jaws, tongue, vocal cords, neck muscles, and muscles of deglutition may become paralysed (Fig. 15). The pupils are dilated. Respiratory failure may be precipitated by airway obstruction at this stage, or later after paralysis of intercostal muscles and the diaphragm. Neurotoxic effects are completely reversible, either acutely in response to antivenom or anticholinesterases (for example, following bites by Asian cobras, some Latin American coral snakes— *Micrurus* spp., and Australasian death adders—Acanthophis) or they may wear off spontaneously in 1 to 7 days.

Fig. 13 Extensive necrosis of skin and subcutaneous tissues in a Nigerian girl bitten 9 days previously on the elbow by a black-necked or spitting cobra (*Naja nigricollis*). (Copyright D.A. Warrell.)

Fig. 14 Sierra Leonian woman showing 'skip lesion' separated by an area of unaffected skin after envenoming by a black-necked spitting cobra. (Copyright D.A. Warrell.)

Fig. 15 Neurotoxic envenoming. Ptosis, ophthalmoplegia, and inability to open the mouth and protrude the tongue in a Sri Lankan patient envenomed by the common krait (*Bungarus caeruleus*). (Copyright D.A. Warrell.)

Envenoming by terrestrial Australasian elapids produces four main groups of symptoms: neurotoxicity (Fig. 16), haemostatic disturbances and, rarely, generalized rhabdomyolysis and renal failure. Painful regional lymph nodes are a useful sign of impending systemic envenoming, but local signs are usually mild, except after bites by the king brown or Mulga snake (*Pseudechis australis*). Early symptoms include vomiting, headache and syncopal attacks.

Fig. 16 Generalized paralysis, including ptosis, external ophthalmoplegia, inability to open the mouth, protrude the tongue, swallow or speak, and respiratory paralysis requiring mechanical ventilation in a Papua New Guinean man bitten 24 h previously by a taipan (*Oxyuranus scutellatus canni*). (Copyright D.A. Warrell.)

Patients 'spat' at by spitting elapids may develop venom ophthalmia. There is intense pain in the eye, blepharospasm, palpebral oedema, and leucorrhoea (Fig. 17(a)). Corneal erosions can be seen by slit-lamp or fluorescein examination in more than half of patients spat at by *N. nigricollis*. Rarely, venom is absorbed into the anterior chamber causing hypopyon and anterior uveitis. Secondary infection of corneal abrasions may lead to permanent blinding opacities or panophthalmitis (Fig. 17(b)).

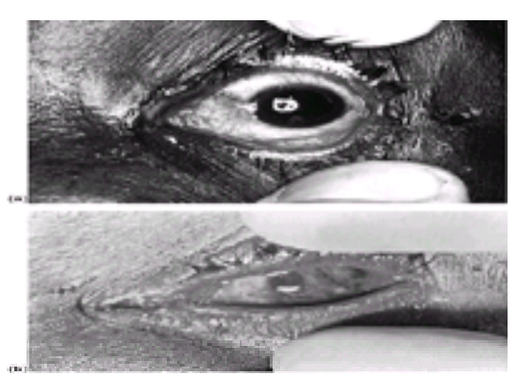

Fig. 17 Venom ophthalmia caused by the black-necked spitting cobra (*Naja nigricollis*). (a) Acute venom ophthalmia showing intense painful inflammation and

discharge. (b) In this case, the corneal injury was neglected and so secondary infection developed, necessitating enucleation of the eye. (Copyright D.A. Warrell.)

Bites by sea snakes and sea kraits

Patients envenomed by sea snakes notice headache, a thick feeling of the tongue, thirst, sweating, and vomiting. Between 30 min and 3.5 h after the bite there is generalized aching, stiffness, and tenderness of the muscles. Trismus is common. Later there is generalized flaccid paralysis. Myoglobinuria appears 3 to 8 h after the bite. Myoglobin and potassium released from damaged skeletal muscles can cause renal failure, while hyperkalaemia may precipitate cardiac arrest.

Bites by Viperidae (vipers, adders, rattlesnakes, lance-headed vipers, moccasins, and pit vipers)

Viper venoms usually produce more severe local effects than do those of other snakes. Swelling may become detectable within 15 min but is sometimes delayed for several hours. It spreads rapidly, sometimes involving the whole limb and adjacent trunk. There is associated pain and tenderness in regional lymph nodes, with bruising of overlying tissues and lymphangitic lines. Bruising, blistering, and necrosis may appear during the next few days (Fig. 18). Necrosis can be severe following bites by some rattlesnakes, lance-headed vipers (genus *Bothrops*), Asian pit vipers, and the large African *Bitis* species (puff adder, Gabon, and rhinoceros-horned vipers, etc.). When the envenomed tissue is contained in a tight fascial compartment such as the pulp space of digits or the anterior tibial compartment, ischaemia may result (Fig. 19). Absence of swelling 2 h after a viper bite suggests that there has been no envenoming. However, fatal envenoming by a few species can occur in the absence of local signs (for example, *C. d. terrificus, C. scutulatus*, and Burmese Russell's viper). Haemostatic abnormalities are characteristic of envenoming by Viperidae. Persistent bleeding from fang puncture wounds, venepuncture or injection sites, other new and partially healed wounds, and postpartum, indicates that the blood is incoagulable. Spontaneous systemic haemorrhage is most often detected in the gingival sulci (Plate 4). Epistaxis, haematemesis, cutaneous ecchymoses, haemoptysis, and subconjunctival, retroperitoneal, and intracranial haemorrhages (Fig. 20) are also reported. Patients envenomed by Burmese Russell's vipers may suffer haemorrhagic infarction of the anterior pituitary (Sheehan's syndrome) (Fig. 21). Hypotension and shock are common in patients bitten by North American rattlesnakes (e.g. *C. adamanteus, C. atrox*, and *C. scutulatus*), *Bothrops, Daboia*, and *Vipera* species (e.g. *V. palaestinae* and *V. berus*). The central venous pressure is usually low and the pulse rate rapid, suggesting hypovolaemia resulting from extravasation of fluid into the bitten limb. Patients envenomed by Burmese Russell's vipers and children envenomed by *Vipera berus* show evidence of generally increased vascular permeability. Direct myocardial involvement is suggested by an abnormal ECG or cardiac arrhythmia. Patients envenomed by some species of the genera *Vipera* and *Bothrops* and Australasian elapids may experience early transient and recurrent syncopal attacks associated with features of an autopharmacological or anaphylactic reaction, such as vomiting, sweating, colic, diarrhoea, shock, and angio-oedema. These symptoms may appear as early as 5 min or as late as many hours after the bite. Early collapse after bites by Australian brown snakes (genus *Pseudonaja*) and tiger snakes (genus *Notechis*) has been attributed to coronary and pulmonary thromboembolism. Renal failure is a common mode of death in patients envenomed by Viperidae. Victims of Russell's viper may become oliguric within a few hours of the bite and complain of loin pain suggesting renal ischaemia. Neurotoxicity, resembling that seen in patients bitten by Elapidae, is a feature of envenoming by a few species of Viperidae (e.g. *C. d. terrificus,* berg adder—*Bitis atropos* and other small *Bitis* species, and the Sri Lankan *D. russelii*). There is evidence of generalized rhabdomyolysis (Fig. 22), but progression to respiratory or generalized paralysis is unusual.

Fig. 18 Severe blistering in a Thai boy bitten on the leg by a Malayan pit viper (*Calloselasma rhodostoma*). (Copyright D.A. Warrell.)

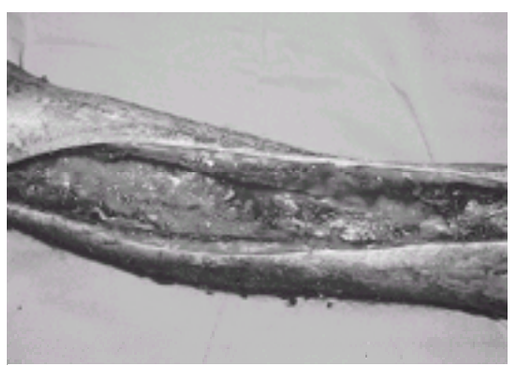

Fig. 19 Extensive necrosis of skin and muscle including the contents of the anterior tibial compartment in a patient bitten by a common lancehead (*Bothrops atrox*) in Brazil. (Copyright D.A. Warrell.)

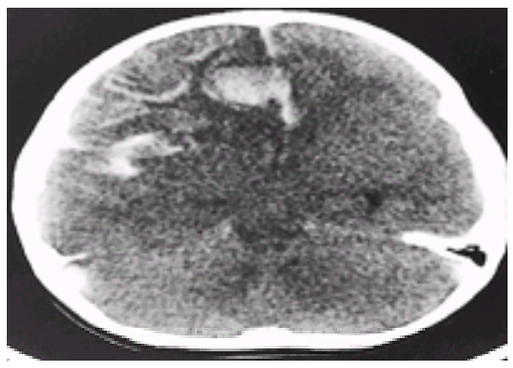

Fig. 20 CT scan showing intracranial haemorrhage in a child bitten by a common lancehead (*Bothrops atrox*) in Ecuador. The fluid level in the larger collection of blood indicates that the blood was incoagulable. (Copyright Hospital Vozandes Quito.)

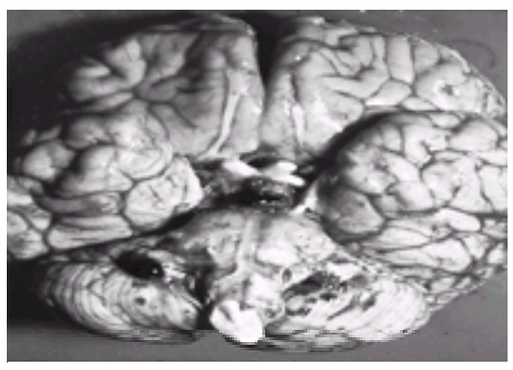

Fig. 22 Brazilian girl bitten 24 h previously by a tropical rattlesnake (*Crotalus durissus terrificus*). She has bilateral ptosis, paralysis of the facial muscles, and gross myoglobinuria resulting from generalized rhabdomyolysis. (Copyright D.A. Warrell.)

Envenoming by European vipers

The common viper or adder (*V. berus*) (Fig. 9), the only venomous snake found in Britain, occurs in England, Wales, Scotland, and northern Europe, extending into the Arctic Circle and through Asia as far east as Sakhalin Island and south to northern Korea. There are four other vipers that are widely distributed in mainland Europe: the nose-horned or sand viper (*V. ammodytes*) in the Balkans, Italy, Austria, and Romania; the asp viper (*V. aspis*) in France (south of Paris), Spain, Germany, Switzerland, and Italy; Lataste's viper (*V. latasti*) in Spain and Portugal, and Orsini's viper (*V. ursinii*) in south-eastern France, central Italy, and Eastern Europe. The Montpellier snake (*Malpolon monspessulanus*) is a large back-fanged colubrid snake whose bite can cause transient mild symptoms.

Clinical features of European viper bite

Pain usually develops quickly at the site of the bite and local swelling is evident within a few minutes, but is sometimes delayed for 30 min or longer. Local blisters containing blood are uncommon. Swelling and bruising may advance to involve the whole limb within 24 h, extend on to the trunk, and in children become generalized. A few cases of intracompartmental syndromes and necrosis have been described. Pain, tenderness, and enlargement of local lymph nodes is sometimes noticeable within hours. Marked lymphangitis and bruising of the affected limb appears within a day or two. Dramatic, early systemic symptoms may appear within 5 min of the bite or be delayed for many hours. They include retching, vomiting, abdominal colic, diarrhoea, incontinence of urine and faeces, sweating, vasoconstriction, tachycardia, shock, and angio-oedema of the face, lips, gums, tongue, throat, as well as epiglottis, urticaria, and bronchospasm. These symptoms may persist for as long as 48 h. Hypotension is the most important sign. It usually develops within 2 h, and may be transient (resolving spontaneously within 2 h), or persistent, recurrent, or progressive, and fatal. ECG changes include flattening or inversion of T waves, ST elevation, second-degree heart block, and cardiac brady- and tachyarrhythmias, atrial fibrillation, and myocardial infarction. Defibrinogenation (incoagulable blood) or milder degrees of coagulopathy and spontaneous bleeding into the gastrointestinal tract, lungs (Fig. 23), or urinary tract are uncommon. Other clinical features include fever, drowsiness, and, rarely, coma and seizures secondary to hypotension or cerebral oedema, respiratory distress/pulmonary oedema (in children), acute renal failure, cardiac arrest, intrauterine death, acute gastric dilatation, and paralytic ileus. Laboratory findings include a neutrophil leucocytosis (more than 20 000/µl in severe cases), thrombocytopenia, initial haemoconcentration and later anaemia resulting from extravasation into the bitten limb, and, rarely, haemolysis, elevation of serum creatine kinase, and metabolic acidosis. Deaths usually occur between 6 and 60 (average 34) h after the bite. Most adder bites cause only trivial symptoms, but patients must be assessed individually. Children may be severely envenomed: in a French series there were three deaths in a group of seven children aged between 2½ and 10 years. The dangers of adder bite should not be underestimated. The antivenom treatment of adder bite is discussed below.

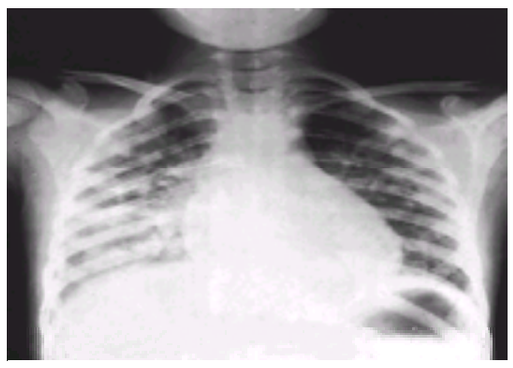

Fig. 23 Chest radiograph of a 9-year-old girl, 3 days after being bitten by *Vipera berus*, showing interstitial pulmonary bleeding. (By courtesy of Dr R. Pugh, Hull, and *The Practitioner*.)

Laboratory investigations

The peripheral neutrophil count is raised to 20 000 cells/µl or more in severely envenomed patients. Initial haemoconcentration, resulting from extravasation of plasma (*Crotalus* species and Burmese *D. russelii*), is followed by anaemia caused by bleeding or, more rarely, haemolysis. Thrombocytopenia is common following bites by pit vipers (e.g. *C. rhodostoma, Crotalus viridis helleri*) and some Viperidae (e.g. *Bitis arietans* and *D. russelii*), but is unusual after bites by *Echis* species. A useful test for venom-induced defibrinogenation is the 20-min whole-blood clotting test. A few millilitres of venous blood is placed in a new, clean, dry, glass test tube, left undisturbed for 20 min, and then tipped once to see if it has clotted or not. Incoagulable blood indicates systemic envenoming (consumption coagulopathy or anticoagulant) and may be diagnostic of a particular species (for example, *Echis* species in the northern third of Africa). Patients with generalized rhabdomyolysis show a steep rise in serum creatine kinase, myoglobin, and potassium levels. Black or brown urine suggests generalized rhabdomyolysis or intravascular haemolysis. Concentrations of serum enzymes, such as creatine kinase and aspartate aminotransferase, are moderately raised in patients with severe local envenoming, probably because of local muscle damage at the site of the bite. High concentrations suggest generalized rhabdomyolysis. Urine should be examined for blood/haemoglobin, myoglobin and protein, and for microscopic haematuria and red cell casts. Electrocardiographic abnormalities such as sinus bradycardia, ST–T changes, various

degrees of atrioventricular block, and hyperkalaemic changes may be seen.

Immunodiagnosis

Specific snake venom antigens have been detected in wound swabs, aspirates or biopsies, serum, urine, cerebrospinal fluid, and other body fluids. Of the various techniques for their detection, radioimmunoassay is probably the most sensitive and specific, but enzyme immunoassay (**EIA**) has been the most widely used. Under ideal conditions, relatively high venom antigen concentrations (wound swabs or aspirates) may be detected quickly enough (15–30 min) to allow the selection of the appropriate monospecific antivenom. A commercial test kit for Australian elapids is produced by CSL, Melbourne. For retrospective diagnosis, including forensic cases, tissue around the fang punctures, wound and blister aspirate, serum, and urine should be stored for EIA immunodiagnosis.

Management of snake bite

First aid

The patient should be reassured and moved to the nearest hospital or dispensary as quickly, as comfortably and passively as possible. The bitten limb should be immobilized with a splint or sling and all unnecessary movement discouraged.

Most traditional first aid methods are potentially harmful and should not be used. Local incisions and suction do not remove venom effectively and may introduce infection, damage tissues, and cause persistent bleeding. Vacuum extractors, potassium permanganate, and ice packs may potentiate local necrosis. Electric shocks may act as counterirritants but are dangerous and have not been proved beneficial. Tourniquets and compression bands are potentially dangerous as they can cause gangrene (Fig. 24), increased fibrinolysis, and bleeding in the occluded limb, peripheral nerve palsies, compartmental ischaemia, and intensification of local signs of envenoming.

Fig. 24 Gangrene of forearm in a Thai patient who applied a tight tourniquet above the elbow for several hours after being bitten by a Malayan pit viper. (Copyright D.A. Warrell.)

The pressure immobilization method developed by the late Struan Sutherland and his colleagues in Australia involves bandaging the entire bitten limb as tightly as for a sprained ankle, using a long crêpe bandage, starting at the toes or fingers, and incorporating a splint (Fig. 25). In animals, this method exerted a pressure of about 55 mmHg and was effective in preventing the systemic uptake of Australian elapid and some other venoms. Anecdotal experience supports the use of the method. Several reported patients have deteriorated after release of the pressure bandage. Prospective clinical studies are needed to assess the risks and benefits of this interesting technique. In the meantime, it is recommended for the first-aid of bites by Australasian elapids, sea snakes, and other elapid snakes whose venoms can have a rapid neurotoxic effect (e.g. kraits, coral snakes, and some cobras). The method is not recommended for bites by snakes whose venoms cause massive local swelling and necrosis as these effects may be accentuated by the bandage, which will increase pressure in fascial compartments, risking ischaemic necrosis.

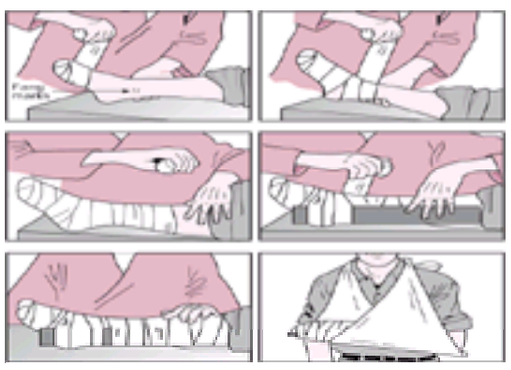

Fig. 25 Sutherland's pressure immobilization method for envenoming by neurotoxic species, such as Australasian elapids. (By courtesy of Australian Venom Research Unit, University of Melbourne.)

Pursuing and killing the snake is not recommended, but if the snake has been killed it should be taken with the patient to hospital, but it must not be handled as even a severed head can inject venom.

Patients being transported to hospital should lie on their left side to prevent aspiration of vomit. Persistent vomiting can be treated with chlorpromazine by intravenous injection (25–50 mg for adults, 1 mg/kg for children). Syncope, shock, angio-oedema, and other autonomic symptoms can be treated with 0.1 per cent adrenaline (epinephrine) by subcutaneous injection (0.5 ml for adults, 0.01 ml/kg for children) and an antihistamine such as chlorphenamine (chlorpheniramine) maleate by intravenous injection (10 mg for adults, 0.2 mg/kg for children). Patients with incoagulable blood will develop haematomas after intramuscular and subcutaneous injections, and so the intravenous route should be used whenever possible. Respiratory distress and cyanosis should be treated by clearing the airway, giving oxygen, and, if necessary, assisted ventilation. If the patient is unconscious and no femoral or carotid pulses can be detected, cardiopulmonary resuscitation must be started immediately.

Hospital treatment

Clinical assessment

In most cases of snake bite there are uncertainties about the species and the quantity and composition of venom injected, which can be resolved only by admitting the patient for at least 24 hours of observation. Local swelling is usually detectable within 15 min of pit viper envenoming and within 2 h of envenoming by most other vipers, but may not develop in patients bitten by some neurotoxic species such as kraits, coral snakes, and sea snakes. Fang marks are sometimes invisible. Tender enlargement of regional lymph nodes draining the bitten area is an early sign of envenoming by Viperidae, some Elapidae, and Australasian elapids. All the tooth sockets should be examined meticulously as this is usually the first site of spontaneous bleeding (Plate 4): other common sites are the nose, conjunctivas, skin, and gastrointestinal tract. Persistent bleeding from venepuncture sites and other wounds implies incoagulable blood. Hypotension and shock are important signs of hypovolaemia or cardiotoxicity, seen particularly in patients bitten by North American rattlesnakes and some Viperinae (e.g. *V. berus, D. russelii, V. palaestinae*). Ptosis is the earliest sign of neurotoxic envenoming (Fig. 15). Respiratory muscle power should be assessed objectively and repeatedly, for example by measuring vital capacity. Trismus and generalized muscle tenderness suggest rhabdomyolysis (sea snakes). If a procoagulant venom is suspected, the coagulability of whole

blood should be checked at the bedside using the 20-min, whole-blood clotting test.

Antivenom treatment

The most important decision is whether or not to give antivenom, the only specific treatment for envenoming. There is now abundant evidence that in patients with severe envenoming the benefits of this treatment far outweigh the risks of antivenom reactions (see below). Antivenom has reduced the mortality of systemic envenoming by *Echis ocellatus* in Nigeria from 20 to 3 per cent and by *C. d. terrificus* in Brazil from 74 to 12 per cent. Antivenoms are effective in reversing hypotension caused by *V. berus* envenoming and coagulopathies caused by *Bothrops* species, *D. russelii, C. rhodostoma, T. albolabris*, and *Oxyuranus scutellatus*. Antivenom, also known as antivenin, antivenene, and antisnakebite serum, is the partially purified immunoglobulin (whole IgG, $F(ab')_2$, or Fab fragments) of horses or sheep which have been immunized with venom.

General indications for antivenom

Antivenom is indicated if there are signs of systemic envenoming such as:

1. haemostatic abnormalities, for example spontaneous systemic bleeding, incoagulable blood, or thrombocytopenia;
2. neurotoxicity;
3. hypotension and shock, abnormal ECG, or other evidence of cardiovascular dysfunction;
4. generalized rhabdomyolysis.

Supporting evidence of severe envenoming is a neutrophil leucocytosis, elevated serum enzymes such as creatine kinase and aminotransferases, haemoconcentration, severe anaemia, myoglobinuria, haemoglobinuria, methaemoglobinuria, hypoxaemia, and acidosis.

In the absence of systemic envenoming, local swelling involving more than half the bitten limb, extensive blistering or bruising, bites on digits, and rapid progression of swelling are indications for antivenom, especially in patients bitten by species whose venoms are known to cause local necrosis (e.g. Viperidae, Asian cobras, and African spitting cobras).

Special indications for antivenom

Some developed countries can afford a wider range of indications.

United States and Canada

After bites by the most dangerous rattlesnakes (*C. atrox, C. adamanteus, C. viridis, C. horridus*, and *C. scutulatus*) antivenom therapy should be given early, even before systemic envenoming has become obvious. Rapid spread of a local swelling is considered an indication for antivenom, as is immediate pain or any other symptom or sign of envenoming after bites by coral snakes (*Micruroides euryxanthus*, *Micrurus fulvius*, and *M. tener*).

Australia

Antivenom should be given to any patient with proved or suspected snake bite if there are tender regional lymph nodes or any other evidence of systemic spread of venom, and in anyone effectively bitten by an identified highly venomous species.

Europe

(Adder—*Vipera berus*—and other European Vipera). Zagreb antivenom, or Protherics ViperaTAb (Table 1), is indicated to prevent morbidity and reduce the length of convalescence in patients with moderately severe envenoming, as well as to save the lives of severely envenomed patients. Indications are:

1. a fall in blood pressure (systolic to <80 mmHg, or by more than 50 mmHg from the normal or admission value) with or without signs of shock;
2. other signs of systemic envenoming (see above) including spontaneous bleeding, coagulopathy, pulmonary oedema or haemorrhage (shown by chest radiograph), ECG abnormalities, and a definite peripheral leucocytosis (more than 15 000/µl) and elevated serum creatine kinase;
3. severe local envenoming—swelling of more than half the bitten limb developing within 48 h of the bite—even in the absence of systemic envenoming;
4. in adults, swelling extending within 4 h beyond the wrist after bites on the hand or beyond the ankle after bites on the foot within 4 h of the bite.

Patients bitten by European Vipera who show any evidence of envenoming should be admitted to hospital for observation for at least 24 h. Antivenom should be given whenever there is evidence of systemic envenoming ((1) or (2) above), even if its appearance is delayed for several days after the bite.

Prediction of antivenom reactions

Skin and conjunctival tests do not predict early (anaphylactic) or late (serum sickness type) antivenom reactions and should not be used.

Contraindications to antivenom

Atopic patients and those who have reacted previously to equine antiserum are at increased risk of developing severe antivenom reactions. In such cases, antivenom should be given only if there is severe envenoming. Reactions may be prevented or ameliorated by pretreatment with subcutaneous adrenaline (epinephrine), antihistamine, and hydrocortisone, or a continuous intravenous infusion of 1:1 000 000 adrenaline while antivenom is being given. However, this prophylaxis is not safe enough to be recommended for routine use. Rapid desensitization is not recommended.

Selection and administration of antivenom

Antivenom should be given only if its stated range of specificity includes the species responsible for the bite. Opaque solutions should be discarded, as precipitation of protein indicates loss of activity and an increased risk of reactions. Expiry dates quoted on ampoules are often very conservative for commercial reasons. Liquid and lyophilized antivenoms stored below 8 °C usually retain most of their activity for 5 years or more. Monospecific (monovalent) antivenom is ideal if the biting species is known. Polyspecific (polyvalent) antivenoms are used in many countries because of the difficulty in identifying the species responsible for bites. Polyspecific antivenoms may be effective but a higher dose is required because they contain less of the specific neutralizing antibody per unit of immunoglobulin than monospecific antivenoms. Antivenoms may exhibit a range of paraspecific neutralizing activity. For example, the South African Institute for Medical Research's 'polyvalent antivenom', which is raised against the venoms of 10 species, has paraspecific activity against a further five species.

It is almost never too late to give antivenom while signs of systemic envenoming persist, but, ideally, it should be given as soon as it is indicated. Antivenom has proved effective up to 2 days after sea snake bites and, in patients still defibrinated, weeks after bites by Viperidae. In contrast, local envenoming is probably not reversible unless antivenom is given within a few hours of the bite. The intravenous route is the most effective. An infusion of antivenom diluted in approximately 5 ml of isotonic fluid/kg body weight is easier to control than an intravenous 'push' injection of undiluted antivenom given at the rate of about 4 ml/min, but there is no difference in the incidence or severity of antivenom reactions in patients treated by these two methods.

Dose of antivenom

Manufacturers' recommendations are based on mouse protection tests and may be very misleading. Few clinical trials have been performed to establish appropriate starting doses, and in most countries antivenom is used empirically. Many hospitals in the rural tropics give a standard dose of 1 to 2 ampoules to very patient who claims to have been bitten, irrespective of clinical severity. This practice squanders scarce, expensive antivenom and exposes non-envenomed patients to the risk of

reactions. Some suggested initial doses are given in Table 1. *Children must be given the same dose as adults.*

Response to antivenom

Often, there is marked symptomatic improvement soon after antivenom has been injected. In shocked patients, the blood pressure may rise and consciousness return (*C. rhodostoma, V. berus, Bitis arietans*). Neurotoxic signs may improve within 30 min (*Acanthophis* spp, *N. kaouthia*), but the response usually take several hours. Spontaneous systemic bleeding usually stops within 15 to 30 min and blood coagulability is restored within 6 h of antivenom treatment, provided a neutralizing dose has been given. More antivenom should be given if severe signs of envenoming persist after 1 to 2 h, or if blood coagulability is not restored within about 6 h. Systemic envenoming may recur hours or days after an initially good response to antivenom. This is explained by the continuing absorption of venom from the injection site after clearance of antivenom from the bloodstream. The apparent serum half-lives of antivenoms in envenomed patients range from 26 to 95 h. Envenomed patients should therefore be assessed daily for at least 3 or 4 days.

Antivenom reactions

Early (anaphylactic) reactions

These develop within 10 to 180 min of starting antivenom in between 3 and 84 per cent of patients. The incidence increases with dose and decreases when more highly refined antivenom is used, and when administration is by intramuscular rather than intravenous injection. The symptoms are itching, urticaria, cough, nausea, vomiting, other autonomic manifestations, fever, and tachycardia. Up to 40 per cent of patients with early reactions develop systemic anaphylaxis: hypotension, bronchospasm, and angio-oedema. Deaths are rare, but individual cases, such as the asthmatic boy who died from anaphylactic shock after receiving Pasteur antivenom in England in 1957, have been widely publicized and have led to an unreasonable rejection of antivenom treatment. Early antivenom reactions are unlikely to be type-I, IgE-mediated hypersensitivity reactions to equine serum protein. They result from complement activation by immune complexes or aggregates of IgG.

Pyrogenic reactions

Pyrogenic reactions result from contamination of the antivenom with endotoxin-like compounds. Fever, rigors, vasodilatation, and a fall in blood pressure develop 1 to 2 h after treatment. In children, febrile convulsions may be precipitated.

Late reactions

Late reactions of serum sickness type may develop between 5 and 24 (mean 7) days after antivenom therapy. The incidence of these reactions and the speed of their development increases with the dose of antivenom. Clinical features include fever, itching, urticaria, arthralgia (sometimes involving the temporomandibular joint), lymphadenopathy, periarticular swellings, mononeuritis multiplex, albuminuria, and, rarely, encephalopathy. This is a classical immune complex disease.

Treatment of antivenom reactions

Adrenaline (epinephrine) is the effective treatment for early reactions; 0.5 to 1.0 ml of 0.1 per cent (1 in 1000, 1 mg/ml) is given by intramuscular injection to adults (children 0.01 ml/kg) at the first signs of a reaction. The dose may be repeated if the reaction is not controlled. Patients with profound hypotension, severe bronchospasm, or laryngeal oedema may be given adrenaline by slow intravenous injection (0.5 mg diluted in 20 ml of isotonic saline over 10–15 min). A histamine anti-H_1 blocker, such as chlorphenamine (chlorpheniramine) maleate (10 mg for adults; 0.2 mg/kg for children) should be given by intravenous injection to combat the effects of histamine release during the reaction. Pyrogenic reactions are treated by cooling the patient and giving antipyretics. Late reactions respond to an oral antihistamine such as chlorphenamine (2 mg every 6 hours for adults; 0.25 mg/kg per day in divided doses for children) or to oral prednisolone (5 mg every 6 hours for 5 to 7 days for adults; 0.7 mg/kg per day in divided doses for children).

Supportive treatment

Neurotoxic envenoming

Bulbar and respiratory paralysis may lead to death from aspiration, airway obstruction, or respiratory failure. A clear airway must be maintained and, if respiratory distress develops, a cuffed endotracheal tube should be inserted or a tracheostomy performed. Provided they are adequately ventilated, patients with neurotoxic envenoming remain fully conscious with intact sensation. Patients have been effectively ventilated manually (by Ambu bag or anaesthetic bag), as in the 1952 poliomyelitis epidemic in Copenhagen, for 30 days and have recovered after 10 weeks of mechanical ventilation. Although artificial ventilation was first suggested for neurotoxic envenoming more than 100 years ago, patients continue to die because they are denied this simple procedure. Anticholinesterases have a variable but potentially useful effect in patients with neurotoxic envenoming, especially when postsynaptic neurotoxins are involved. The 'Tensilon test' should be performed in all cases of severe neurotoxic envenoming, as with suspected myasthenia gravis. Atropine sulphate (0.6 mg for adults; 50 µg/kg for children) or glycopyrronium is given by intravenous injection followed by an intravenous injection of edrophonium chloride (10 mg for adults; 0.25 mg/kg for children). Patients who respond convincingly can be maintained on neostigmine methyl sulphate (50–100 µg/kg) and atropine, every 4 hours by continuous infusion.

Hypotension and shock

If the central venous pressure is low or there is other clinical evidence of hypovolaemia, a plasma expander, preferably fresh whole blood or fresh-frozen plasma, should be infused. If there is evidence of increased capillary permeability (e.g. facial and conjunctival oedema, serous effusions, haemoconcentration, hypoalbuminaemia, etc.) it may be safer in the long term to rely on a selective vasoconstrictor such as dopamine (starting dose 2.5–5 µg/kg per min by intravenous infusion). Delayed hypotension developing about one week after bites by Burmese *D. russelii* may respond to intravenous hydrocortisone.

Oliguria and renal failure

Urine output, serum creatinine, urea, and electrolytes should be measured each day in patients with severe envenoming and in those bitten by species known to cause renal failure (e.g. *D. russelii, C.d. terrificus, Bothrops* spp., sea snakes). If urine output drops below 400 ml in 24 h, urethral and central venous catheters should be inserted. If urine flow fails to increase after cautious rehydration, diuretics should be tried (e.g. furosemide (frusemide)) by slow intravenous injection, 100 mg followed by 200 mg, and then mannitol. Dopamine (2.5 µg/kg per min by intravenous infusion) has proved effective in some patients bitten by Russell's vipers. If these measures are ineffective, the patient should be placed on strict fluid balance. Peritoneal or haemodialysis will usually be required. In Rangoon, the mortality of established renal failure following *D. russelii* envenoming has been reduced to less than 30 per cent by using peritoneal dialysis, usually for only 72 h.

Local infection at the site of the bite

Bites by some species (e.g. *Bothrops* spp., *C. rhodostoma*) are likely to be complicated by local infections caused by bacteria in the snake's venom or on its fangs. This should be prevented with penicillin, chloramphenicol, or erythromycin and a booster dose of tetanus toxoid, especially if the wound has been incised or tampered with in any way. An aminoglycoside such as gentamicin should be given for 48 h if there is evidence of local necrosis.

Management of local envenoming

Bullae are best left intact. The bitten limb should not be elevated as this increases the risk of intracompartmental ischaemia. Once definite signs of necrosis have appeared (blackened anaesthetic area with putrid odour or signs of sloughing), surgical debridement, immediate split-skin grafting, and broad-spectrum antibiotic cover are indicated.

Intracompartmental syndrome and fasciotomy

Increased pressure within tight fascial compartments such as the digital pulp spaces and anterior tibial compartment may cause ischaemia. This complication is most likely after bites by North American rattlesnakes, *Calloselasma rhodostoma*, *Trimeresurus flavoviridis, Bothrops* spp., and *Bitis arietans*. The signs are excessive pain,

weakness of the compartmental muscles and pain when they are passively stretched, hypoaesthesia of skin supplied by nerves running through the compartment, and obvious tenseness of the compartment. Detection of arterial pulses by palpation or Doppler does not exclude intracompartmental ischaemia. Intracompartmental pressures exceeding 45 mmHg carry a high risk of ischaemic necrosis. In these circumstances, fasciotomy may be justified, but it did not prove effective in saving envenomed muscle in experimental animals. Fasciotomy is contraindicated until blood coagulability has been restored. Early adequate antivenom treatment will prevent the development of intracompartmental syndromes in most cases. Necrosis of digits is especially common.

Haemostatic disturbances

Once specific antivenom has been given to neutralize venom procoagulants, restoration of coagulability and platelet function may be accelerated by giving fresh whole blood, fresh-frozen plasma, cryoprecipitates (containing fibrinogen, factor VIII, fibronectin, and some factors V and XIII) or platelet concentrates. Heparin has been used to treat a variety of snake bites, usually with disastrous results. Heparin did not prove beneficial in patients envenomed by *Echis ocellatus*.

Other drugs

Corticosteroids, antifibrinolytic agents (aprotinin–Trasylol and e-amino-caproic acid), antihistamines, trypsin, and a variety of traditional herbal remedies have all been used, but none has proved effective and many are potentially harmful.

Treatment of snake venom ophthalmia

When cobra venom is 'spat' into the eyes, first aid consists of irrigation with generous volumes of water or any other bland liquid which is available. Unless a corneal abrasion can be excluded by fluorescein staining or slit-lamp examination, treatment should be the same as for any corneal injury: a topical antimicrobial such as tetracycline or chloramphenicol should be applied. Instillation of antivenom is not recommended. A 0.1 per cent adrenaline eyedrop preparation relieves the pain.

Interval between bite and death

Exceptionally, patients may die 'within a few minutes' (reputedly after a bite by the king cobra, *Ophiophagus hannah*) or as long as 41 days (*E. carinatus*) after snake bite. However, most deaths occur about 8 h after cobra bites (*N. naja*), 18 h after krait bites (*Bungarus caeruleus*), 16 h after North American rattlesnake bites (*Crotalus* spp.), 3 days after *D. russelii* bites, and 5 days after *Echis* bites.

Prevention of snake bite

To reduce the risk of bites, snakes should never be disturbed, attacked, cornered, or handled, even if they are thought to be a harmless species or appear to be dead. Venomous species should never be kept as pets or as performing animals. In snake-infested areas, boots, socks, and long trousers should be worn for walks in undergrowth or deep sand, and a light should always be carried at night. Collecting firewood, dislodging logs and boulders with bare hands, pushing sticks or digits into burrows, holes, and crevices, climbing rocks and trees covered with dense foliage, and swimming in overgrown lakes and rivers are particularly hazardous activities. Unlit paths and gutters are especially dangerous after heavy rains. To prevent sea-snake bites, fishermen should not touch these animals when they are caught in nets or on lines, swimmers and divers should not aggravate them and should avoid wading in the sea, especially in muddy estuaries, in sand, or near coral reefs. It is futile and ecologically undesirable to attempt to exterminate venomous snakes. Various substances toxic to snakes, such as insecticides and methylbromide, have been used to keep human dwellings free of these animals. However, no effective but harmless snake repellent has been discovered.

Immunization against envenoming

The idea of inducing protective levels of antibodies against lethal venom components in high-risk populations by pre-exposure immunization has been tried, with inconclusive results, in Japan (*T. flavoviridis*) and considered in Burma (*D. russelii*). To be effective, high titres of a neutralizing antibody would have to be present at the time of the bite. The accelerated secondary response stimulated by envenoming would be too late to be useful.

Venomous lizards

Of the lizards, two species—the gila monster (*Heloderma suspectum*) and Mexican beaded lizard (*H. horridum*), up to 60 to 80 cm long—are venomous (Plate 5). They occur in the south-western United States, western Mexico, and Central America as far south as Guatemala. Venom from mandibular glands is conducted along grooved lower teeth. It contains lethal gila and horridum glycoprotein toxins, phospholipase A_2, and several fascinating bioactive peptides: helospectins I and II and helodermin (vasoactive intestinal peptide analogues) and extendins-3 and -4 (glucagon-like peptide-1 analogues). Humans are rarely bitten. The lizard clings on with a bulldog-like grip and may leave radiolucent teeth in the wound. There is immediate severe local pain and regional lymphadenopathy. Systemic symptoms include weakness, dizziness, hypotension, syncope, sweating, rigors, tinnitus, nausea, vomiting, leucocytosis, and ECG changes. There are no reliable reports of fatalities, but a patient envenomed by *H. suspectum* developed refractory hypotension resulting in myocardial infarction and renal failure and associated with coagulopathy. Specific antivenom is not available. Strong analgesia may be required. Hypotension should be treated with plasma expanders and, if persistent, with a pressor agent such as dopamine.

Poisonous amphibians

The moist skin of amphibians such as frogs, toads, newts, and salamanders is an accessory respiratory organ, which is protected from micro-organisms by highly toxic secretions containing amines, peptides, proteins, steroids, and alkaloids. Some compounds are synthesized *de novo*, while others are sequestered from prey such as ants, beetles, and millipedes. The bitter flavour and lethal effects of these secretions and the vivid warning coloration of many species defend them against predators. The skin of poison frogs (Dendrobatidae) of Central and South America secrete lipophilic alkaloids such as batrachotoxins (*Phyllobates* spp.), which activate sodium channels; histrionico toxins (*Dendrobates histrionicus*) (Plate 6), which block nicotinic receptors; pumiliotoxins (*D. pumilio*), which affect sodium channels; and epibatidine (*Epipedobates tricolor*), a powerful analgesic and nicotinic receptor agonist. Two Colombian tribes, the Embará and Noanamá Chocó, use the skin poisons of three species of Phyllobates to coat the tips of their blow-gun darts (Plate 7). Some toads can squirt venom from their parotid glands, this contains bufadienolides which affect membrane Na^+/K^+-ATPase. When licked or put in the mouth by dogs or children or when ingested as Chinese traditional medicines such as *Kyushin*, *Yixin Wan*, or the topical aphrodisiac *Ch'an-Su*, the poisons can cause fatal digoxin-like poisoning. Symptoms include hypersalivation, cyanosis, cardiac arrhythmias, and generalized convulsions. Antidigoxin antibodies ('Digibind') have some therapeutic effect.

The skin of three species of newts, genus *Taricha*, from the western United States, contains tarichatoxins identical to tetrodotoxin, which also occurs in some toads, frogs, fish, crustaceans, and octopuses (see below). Tetrodotoxin can be absorbed through the gastric mucosa, explaining the death of a man who swallowed a 20-cm long Oregon rough-skinned newt (*Taricha granulosa*). He developed paraesthesia of the lips, progressing to more generalized numbness and weakness, and had a cardiopulmonary arrest about 2 h after swallowing the newt.

Poisonous birds

The feathers, skin, and breast muscles of three species of pitohui or thickhead, passerine birds from New Guinea (genus *Pitohui*, Pachycephalidae) contain homobatrachotoxin, a potent steroidal alkaloid that activates sodium channels and was originally isolated from the skin of South American poison-dart frogs (*Phyllobates,* Dendrobatidae—see above). Poisonous pitohuis have an unpleasant peppery odour and their skin has a bitter flavour. Contact with their feathers causes numbness and burning of the tongue, lip or skin wounds, and sneezing. This may be a protective mechanism, and the striking 'warning' coloration of the hooded pitohui (*P. dichrous*) (Plate 8) may be the subject of Müllerian mimicry by less poisonous species. Judging by their reputation in Papua New Guinea, other species, including birds of paradise and the blue-capped Ifrita (*Ifrita kowaldi*), may also prove to have poisonous tissues.

Venomous fish

More than 100 species of fish can inflict dangerous stings on humans. Venom is injected through spines in front of the fins and tail and in the gill covers. The Indo-Pacific region and other tropical waters have the richest venomous fish fauna, but dangerous species such as sharks, chimaeras, and weevers also occur in temperate northern waters and a number of large rivers in South American, West Africa, and Southeast Asia are inhabited by freshwater stingrays (*Potamotrygon*

spp.). The following groups are capable of fatal envenoming: Squaliformes (sharks and dogfish), Rajiformes (stingrays and mantas), Siluroidei (catfish), Trachinidae (weevers), Scorpaenidae (scorpionfish and stonefish), and Uranoscopidae (stargazers and stonelifters). Venom glands are embedded in grooves in the spines or, in the case of stingrays, lie beneath a membrane covering the long barbed precaudal spine.

Incidence and epidemiology

Weeverfish are common around the British coast especially in Cornwall. Hundreds of stings occur in some years, with a peak incidence in August and September. A total of 58 cases were seen at one hospital at Pula on the Adriatic coast over a period of 13 years. It has been estimated that there are 1500 stings by rays and 300 stings by scorpionfish in the United States each year. Stings by venomous freshwater rays (*Potamotrygon hystrix, P. motoro*) are common in the Amazon region of Brazil and especially in Acré. In 4 years, 81 cases of stonefish (*Synanceja* spp.) sting were seen in Pulau Bukom Hospital near Singapore. Ornate, but aggressive and venomous members of the genera *Pterois* and *Dendrochirus* (lion, zebra, tiger, turkey, or red fire fish) (Plate 9), which are popular aquarium pets, often sting their owners on the fingers. Most fish stings are inflicted on the soles of the feet of people wading near the shore or in the vicinity of coral reefs. Venomous fish are effectively camouflaged (*Synanceja* spp.) or lie partly covered by sand. Stingrays lash their tails at the intruding limb and usually impale the ankle (Plate 9). Fatal fish stings are very rarely reported.

Venom composition

The instability of most fish venoms at normal ambient temperatures has made them difficult to study. Stingray and weeverfish venoms contain peptides, enzymes, and a variety of vasoactive compounds such as kinins, 5-hydroxytryptamine, histamine, and catecholamines. Pharmacological effects include local necrosis, direct actions on cardiac skeletal and smooth muscle resulting in ECG changes, hypotension, paralysis, and central nervous system depression.

Clinical features

Immediate sharp, agonizing pain is the dominant symptom. Hot, erythematous swelling extends up the stung limb and may persist with pain for several days and be complicated by necrosis (Plate 10) and secondary infection by marine *Vibrio* spp. (such as *V. vulnificus*) and other unusual bacteria, particularly if the spine remains embedded in the wound. Stingray spines, which are up to 30 cm long, can cause severe lacerating injuries especially to the lower legs, but if the victim inadvertently lies on the ray or falls onto it, the spine may penetrate the thoracic or abdominal cavities with fatal results.

Systemic effects are uncommon after weever stings (Trachinidae), but people stung by rays or Scorpaenidae (scorpion- and stone-fish) may develop nausea, vomiting, signs of autonomic nervous system stimulation; such as diarrhoea, sweating and hypersalivation; cardiac arrhythmias, hypotension, respiratory distress, neurological signs, and generalized convulsions. Patients have died within 1 h of being stung by *Synanceja verrucosa*.

Treatment

Pain is alleviated by immersing the stung limb in water which is uncomfortably hot (but less than 45 °C) yet not scalding. The 50 °C recommended by some authorities will cause a full thickness scald! Temperature can be assessed with the unstung limb. Addition of magnesium sulphate is unnecessary. Injection of a local anaesthetic is less effective even when applied as a ring block in the case of stung digits, but a local nerve block with 0.5 per cent of plain bupivacaine does seem to work. The venomous spine (which may be barbed), fragments of membrane, and other foreign material should be removed as soon as possible. Systemic effects must be treated symptomatically. An adequate airway should be established and cardiopulmonary resuscitation may be needed. Severe hypotension may respond to adrenaline (epinephrine), bradycardia to atropine. The Commonwealth Serum Laboratories (**CSL**) in Australia manufacture an antivenom specific for *Synanceja trachynis, S. verrucosa*, and *S horridus*. This has paraspecific activity against the venoms of the North American scorpionfish (*Scorpaena guttata*) and some other members of the Scorpaenidae. One ampoule (2 ml or 2000 units) is given intravenously for each two puncture marks found at the site of the sting. The dose is increased for patients with severe symptoms. Antibiotic treatment for secondary infections should take into account the range of possible marine pathogens. Ciprofloxacin covers *Vibrio* and *Aeromonas* spp.

Prevention

Fish stings can be prevented by employing a shuffling gait when wading, by avoiding handling living or dead fish, and by keeping clear of fish in the water, especially in the vicinity of tropical reefs. Footwear protects against most species except stingrays.

Poisoning by ingestion of aquatic animals

Acute gastrointestinal symptoms ('food poisoning') after eating seafood are usually caused by bacterial or viral infections such as *Vibrio parahaemolyticus* (crustaceans, especially shrimps), *V. cholerae* (crabs and molluscs), non-O group 1 *V. cholerae* (oysters), *V. vulnificus* (oysters), *Aeromonas hydrophila* (frozen oysters), *Plesiomonas shigelloides* (oysters, mussels, mackerel, cuttlefish), *Shigella* spp. (molluscs), *Campylobacter jejuni* (clams), *Salmonella typhi* (molluscs), hepatitis A virus (molluscs, especially clams, and oysters), Norwalk virus (clams and oysters), and astro- and caliciviruses (cockles and other molluscs). Botulism has been caused by eating smoked fish and canned salmon; and in Japan and elsewhere, fish and molluscs became contaminated with methyl mercury from industrial waste, causing severe neurological damage and fetal abnormalities ('Minamata disease').

Toxins in seafood may also give rise to gastrointestinal neurotoxic and histamine-like symptoms. Two main syndromes are described.

Gastrointestinal and neurotoxic syndromes

Nausea, vomiting, abdominal colic, tenesmus, and watery diarrhoea may precede neurotoxic symptoms of paraesthesia of the lips, buccal cavity, and extremities, distorted temperature perception (so that cold objects impart a burning sensation like dry ice), myalgia, progressive flaccid paralysis, dizziness, ataxia, cardiovascular disturbances, bradycardia, and rashes. Important causes of this syndrome are:

1. *Ciguatera fish poisonings*—Symptoms develop between 1 and 6 h (extreme range, minutes to 30 h) after eating fish such as groupers, snappers, parrot fish, mackerel, moray eels, barracudas, and jacks. These are warm-water shore or reef fish. The global incidence is thought to exceed 50 000 cases per year. Up to 1 per cent of the population may be affected each year (e.g. in Kiribati, Tokelau, and Tuvalu in the Pacific region) with a case fatality of 0.1 per cent. The toxins responsible are polyethers—such as ciguatoxin (activates Na^+ channels), maitotoxin (activates Ca^{2+} channels), and scaritoxin, ultimately derived along the food chain from benthic dinoflagellates such as *Gambierdiscus toxicus*. They are concentrated in the liver, viscera, and gonads, especially of large carnivorous fish. The increasing market for exotic fish from the Caribbean and elsewhere has led to cases of ciguatera in Britain. Gastrointestinal symptoms resolve within a few hours, but paraesthesiae and myalgia may persist for a week or even months. Similar symptoms (chelonitoxication) may follow ingestion of marine turtles in the Indo-Pacific area, but the case fatality is much higher.
2. *Tetrodotoxin poisoning*—Scaleless porcupine, sun, puffer, and toad fish (order: Tetraodonitiformes) may become highly poisonous at certain seasons, such as May to June, the spawning season in Japan. Tetrodotoxin, an aminoperhydroquinazoline, is one of the most potent non-protein toxins known. It produces neurotoxic and cardiotoxic effects by blocking voltage-gated sodium ion channels. It is found: concentrated in the ovaries, viscera, and skin of tetraodontiform fish; in the skin of newts (genus *Taricha*), frogs and toads (genera *Colostethus, Atelopus, Bracycephalus*), and salamanders; in the saliva of octopuses; from the digestive glands of several species of gastropod molluscs; in a starfish, flat worm (*Planorbis* spp.), and Nemertine worms in Japan; and is produced by some bacteria (*Pseudomonas* spp.).
3. Puffer fish ('fugu') is particularly popular in Japan where, despite stringent regulations, there are still cases of tetrodotoxin poisoning, with about four deaths each year. Neurotoxic symptoms develop within 10 to 45 min and death from respiratory paralysis usually occurs between 2 and 6 h after eating the fish. There may be no gastrointestinal symptoms. Erythema, petechiae, blistering, and desquamation may appear.
4. *Paralytic shellfish poisoning*—Bivalve molluscs, such as mussels, clams, oysters, cockles, and scallops (and also xanthid, coconut, and horseshoe crabs) may acquire tetrahydropurine neurotoxins such as saxitoxin from dinoflagellates (*Alexandrium* spp.). These may be sufficiently abundant between latitudes 30 °N and 30 °S during the warmer months of May to October to produce a 'red tide'. The dangerous season is signalled by the deaths of large numbers of fish and sea birds. Symptoms develop within 30 min of ingestion and may progress to fatal respiratory paralysis within 12 h in 8 per cent of cases. Milder gastrointestinal and neurotoxic symptoms (neurotoxic shellfish poisoning) without paralysis can follow the ingestion of molluscs contaminated by brevitoxins from *Gymnodinium breve*. These microalgae can also cause a 'red tide'. In the United Kingdom there have been several outbreaks of neurotoxic red-whelk (*Neptunea antiqua*) poisoning attributable to tetramine.

5. *Amnesic shellfish poisoning*—develops after ingestion of mussels and other molluscs contaminated with domoic acid from diatoms (*Pseudonitzschia* spp.). Gastroenteritis starts within 24 h of exposure and, in severe cases, within 48 h there is headache and coma followed by short-term memory loss.

Histamine-like syndrome (scombrotoxic poisoning)

The red flesh of scrombroid fish (tuna, mackerel, bonito, and skipjack) and of canned non-scrombroid fish, such as sardines and pilchards, may be decomposed by the action of bacteria such as *Proteus morgani* and *Klebsiella pneumoniae*, which decarboxylate muscle histidine into saurine, histamine, cadaverine, and other unidentified toxins: 100 g of spoiled fish may contain almost 1 g of histamine. Histamine absorbed from the gut is normally broken down by *N*-methyl-transferase and diamine oxidase (histaminase), but if the histamine concentration is very high, or the patient is taking a diamine oxidase inhibitor such as isoniazid (as antituberculosis chemotherapy), scombrotoxic poisoning may result. Toxic fish may produce a tingling or smarting sensation in the mouth when eaten. Within minutes or up to a few hours after ingestion, flushing, burning, sweating, urticaria, and pruritis may develop with headache, abdominal colic, nausea, vomiting, diarrhoea, bronchial asthma, giddiness, and hypotension.

Diagnosis and treatment

The differential diagnosis includes bacterial and viral food poisoning and allergic reactions.

No specific treatments or antidotes are available, but gastrointestinal contents should speedily be eliminated by emetics and purges. Activated charcoal adsorbs saxitoxin and other shellfish toxins. Mannitol has been advocated for ciguatera poisoning. Atropine is said to improve gastrointestinal symptoms and sinus bradycardia in patients with gastrointestinal and neurotoxic poisoning. Calcium gluconate may relieve mild neuromuscular symptoms. Oximes and anticholinesterases appear ineffective in ciguatera and tetrodotoxin poisoning, respectively. Patients who develop respiratory paralysis should be intubated and ventilated. The symptoms of scrombrotoxic poisoning can be alleviated with antihistamines and bronchodilators.

Prevention

Ciguatera toxin, tetrodotoxin, scombrotoxins and some other marine toxins are heat-stable so cooking does not prevent poisoning. Some toxins are fairly water-soluble and may be leached out by soaking, so water in which fish are cooked should not be drunk. In tropical areas, the flesh of fish should be separated as soon as possible from the head, skin, intestines, gonads, and other viscera which may contain high concentrations of toxin. All scaleless fish should be regarded as potentially tetrodotoxic and very large fish carry an increased risk of being ciguatera-toxic. Moray eels should never be eaten because of the high risk of unusually rapid and severe ciguatera fish poisoning. Scrombroid poisoning can be prevented by eating fresh fish or by freezing them as soon as possible after they are caught. Shellfish should not be eaten during the dangerous seasons and when there are red tides.

Poisoning by ingesting carp's gallbladder

In parts of the Far East, the raw bile and gallbladder of various species of freshwater carp (e.g. the grass carp *Ctenopharyngodon idellus*, 'plaa yeesok' *Probarbus jullieni*) are believed to have medicinal properties. Patients in China, Taiwan, Hong Kong, Japan, Thailand, and elsewhere have developed acute abdominal pain, vomiting, and watery diarrhoea 2 to 18 h after drinking the raw bile or eating the raw gallbladder of these fish. One patient developed flushing and dizziness. Hepatic and renal damage may develop, progressing to oliguric or non-oliguric acute renal failure (acute tubular necrosis). The hepatonephrotoxin has not been identified, but is heat-stable and may be derived from the carp's diet.

Venomous marine invertebrates

(Cnidarians (coelenterates): jellyfish, cubomedusoids, sea wasps, Portuguese-men-o'-war or bluebottles, hydroids, stinging corals, sea anemones, etc.)

The tentacles of cnidarians are armed with millions of nematocysts (stinging capsules). When triggered by contact or chemicals, stinging hairs are everted at enormous acceleration and force, penetrating the skin. These produce lines of painful irritant weals. Cnidarian venoms contain peptides and other vasoactive substances such as 5-hydroxyhistamine, histamine, prostaglandins, and kinins which cause immediate excruciating pain, inflammation, and urticaria.

Epidemiology

The most dangerous species, the box jellyfish, cubomedusoid, sea wasp, or indringa (*Chironex fleckeri*) of northern Australia, has caused more than 70 deaths since 1883. Most stings occur in December and January. Fatal jellyfish stings in the Indo-Pacific region—from India, north to the Philippines, and east to Bougainville Island—are attributable to *Chiropsalmus quadrumanus* and *C. quadrigatus*. Fatal stings have also been inflicted by the Portuguese man-o'-war (*Physalia* spp.) and the Chinese jellyfish *Stomolophus nomurai*. Many stings in northern Queensland are caused by *Carukia barnesi* and other tiny cubomedusoids (Irukandji stings). Hundreds of thousands of swimmers off the northern Adriatic coast were stung by a plague of *Pelagia noctiluca* during the summers of 1977 to 1979. Stings by the sea anemone, *Anemonia sulcata*, are also reported from the coasts of Slovenia and Croatia.

Clinical features

Nematocyst stings may leave a diagnostic pattern on the skin: *C. fleckeri* produces wide, striated brownish-purple weals (Plate 11), whereas *C. barnesi* causes a transient erythematous macule and the Portuguese man-o'-war (genus *Physalia*) produces chains of oval weals surrounded by erythema. Immediate severe pain is the commonest symptom. Chirodropids (genera *Chironex* and *Chiropsalmus*) cause the most severe systemic symptoms such as respiratory arrest, generalized convulsions, pulmonary oedema, and cardiac arrest within minutes of the accident. Other systemic effects include cough, nausea, vomiting, abdominal colic, diarrhoea, rigors, severe musculoskeletal pains, and profuse sweating. 'Irukandji' syndrome consists of severe musculoskeletal pain, anxiety, trembling, headache, piloerection, sweating, tachycardia, hypertension, and pulmonary oedema starting about 30 min after a sting by *C. barnesi* (and at least seven other species of tiny cubomedusoids) and persisting for hours. *Physalia* species can also cause severe systemic envenoming, including intravascular haemolysis, peripheral gangrene, and renal failure.

Treatment

Patients stung by jellyfish must be removed from the water as soon as possible to prevent drowning. The aim is to prevent a further discharge of nematocysts on fragments of tentacles stuck to the skin. Alcoholic solutions such as methylated spirits and suntan lotion, the traditional remedy, cause massive discharge of nematocysts. Commercial vinegar or a 3 to 10 per cent aqueous acetic acid solution are, however, effective against stings by *Chironex* spp. and other cubozoans, including Irukandji. Baking soda and water (50 per cent (w/v)) is effective for the widely distributed Atlantic genus, *Chrysaora*. Vinegar is not recommended for stings by *Chrysaora*, *Physalia*, or *Stomalophus* spp. Ice packs relieve the intense pain. Pressure immobilization may increase the amount of venom injected and is not recommended. Cardiopulmonary resuscitation has proved life-saving in several Australian patients who became cyanosed, comatose, and pulseless. A specific 'sea wasp' antivenom for *C. fleckeri* is manufactured in Australia. Treatment with verapamil is not recommended.

Prevention

Bathers, especially children, should keep out of the sea at times of the year when dangerous cnidarians are prevalent, especially when warning notices have been put up; or they should bathe in 'stinger-resistant' enclosures, but these do not exclude Irukandji. Wet or 'Lycra' suits, nylon stockings, and other clothing will protect against nematocyst stings.

Echinodermata (starfish and sea urchins)

These animals are protected by hard exoskeletons with numerous long, sharp projecting spines and grapples (globiferous pedicellariae) which can release venom when embedded in the skin. Severe pain and local swelling may result, and sometimes systemic effects such as syncope, numbness, generalized paralysis, aphonia, respiratory distress, cardiac arrhythmias, and even death. Embedded fragments of spines may lead to secondary infection and chronic granulomas or damage to

bones and joints.

Treatment

Hot water (see above) may relieve the pain. Skin penetrated by the spines, usually the soles of the feet, should be softened with 2 per cent salicylic acid ointment or acetone. The spines can then be squeezed out or removed surgically. No antivenoms are available.

Mollusca (cone shells and octopuses)

The 500 species of cone shells (genus *Conus*) are carnivorous marine snails that harpoon their prey (fish, polychaete worms, and other molluscs), implanting a radular tooth charged with venom containing a mixture of small (10–30 amino acid) peptide toxins. These include: conotoxins which block acetylcholine receptors and voltage-sensitive Ca^{2+} and Na^{+} channels; conantokins which are *N*-methyl-D-aspartate receptor antagonists with anticonvulsant activity; and conopressins which target vasopressin receptors. Cone shells are attractive and valuable collectors' items. People who pick them up may be stung. Symptoms of envenoming are nausea, vomiting, paraesthesia and numbness of the lips and site of sting, numbness, dizziness, ptosis, diplopia, dysarthria, dyspnoea, and loss of consciousness. In a series of 35 cases mostly stung by *Conus geographus* reported in Japan (1896–1996), 10 died within 2 to 5 h of the sting (Plate 12).

Several species of small octopus found in the Australian and West Pacific region (blue ringed octopuses—genus *Hapalochlaena*—Plate 13) can inject salivary tetrodotoxin when they bite swimmers with their powerful beaks. These bites are painful and cause local bleeding, swelling, and inflammation. Severe neurotoxic symptoms, and even fatal generalized paralysis, may develop within 15 min of the bite.

Treatment

No antivenoms are available. Cardiopulmonary resuscitation and mechanical ventilation may be required.

Venomous arthropods

(Hymenoptera (bees, wasps, yellowjackets, hornets, and ants))

The commonest and most severe stings from *Hymenoptera* spp. are caused by members of the families Apidae (e.g. honey bees, *Apis mellifera*, *A.. cerania*, bumble bees, etc.), Vespidae (e.g. wasps, genera *Polistes* and *Paravespula*), American yellowjackets and 'hornets' (genera *Vespula* and *Dolichovespula*) and European and Asian true hornets (genus *Vespa*), and Formicidae (e.g. fire ants, genus *Solenopsis*). Allergic reactions to single stings from *Hymenoptera* spp. are common, whereas toxic reactions resulting from many stings are rare, except in the Americas. Venom allergens include phospholipases A, hyaluronidase, acid phosphomonoesterases, and melittin (*A. mellifera*). Non-allergenic compounds include vasoactive amines such as histamine, 5-hydroxytryptamine, catecholamines and kinins, cholinesterase (in the venom of the common European wasp (*Paravespula germanica*), pheromones, and 2-methylpiperidine alkaloids in venoms from *Solenopsis* (fire ant) spp.

Epidemiology

Fewer than 10 people die from hymenopteran sting anaphylaxis in England and Wales each year, 2 or 3 per year in Australia, and in the United States there are between 40 and 50 deaths per year. The incidence of systemic reactions to stings by *Hymenoptera* spp. has been reported as 0.4 to 0.8 per cent in children. In an adult population in the United States, the prevalence of systemic allergic sting reactions was found to be 4 per cent; 20 per cent of this population showed evidence of venom hypersensitivity (skin tests or radioallergosorbent test, **RAST**). In Britain, most patients allergic to bee venom are beekeepers or their relatives. Since the escape of swarms of African honey bees (*A. m. scutellata*) in Brazil, in 1957, this aggressive strain has spread throughout Latin America and north as far as Las Vegas in the United States. About 30 deaths from mass attacks by these bees have been reported each year. Two species of fire ants, *Solenopsis richteri* and *S. invicta*, were imported into the United States from South America in 1918 and have now spread to 13 southern states where an estimated 2.5 million individuals are stung each month. The incidence of systemic allergic reactions is about 4 per 100 000 population per year, and there have been fatalities.

Clinical features

Toxic effects

In non-sensitized individuals, a sting, which, in the case of Vespidae and Apidae, introduces about 50 µg of venom, will rapidly produce a hot, red, painful swelling and weal a few centimetres in diameter, which persists for a few hours. These effects are dangerous only if the airway is obstructed, for example following stings on the tongue. As few as 30 stings can cause fatal systemic envenoming in children, but children and adults have survived more than 1000 stings by *A. mellifera*. In some patients, symptoms have suggested histamine toxicity: vasodilatation, hypotension, vomiting, diarrhoea, throbbing headache, coma, and bronchoconstriction. In Latin America, victims of attacks by *A. m. scutellata* have shown evidence of generalized rhabdomyolysis (grossly elevated serum creatine kinase, aminopeptidases, and myoglobin), intravascular haemolysis, hypercatechola-minaemia (hypertension, pulmonary oedema, myocardial damage), bleeding, hepatic dysfunction, and acute renal failure (Plate 14). In non-sensitized people, stings form *Solenopsis* spp. produces pain, itching, swelling, and erythema around a central weal which last a few hours, and later vesicles or pustules. In an unsensitized patient an estimated 10 000 *S. invicta* stings caused no systemic envenoming.

Allergic effects

Clinical suspicion of dangerous venom hypersensitivity arises when systemic symptoms follow a sting. Patients may die within minutes of the sting. Systemic symptoms include: a tingling scalp; itching of the palms, soles, axillas, and perineum, becoming generalized; flushing; dizziness; syncope; wheezing; abdominal colic (uterine colic in women), diarrhoea, incontinence of urine and faeces; tachycardia and visual disturbances; all developing within a few minutes of the sting. Over the next 15 to 20 min, urticaria, angio-oedema of the lips, gums, and tongue, a generalized redness of the skin with swelling, oedema of the glottis, profound hypotension, and coma may develop. A few patients develop serum sickness a week or more after the sting. Some patients with sting allergy have other evidence of an atopic disposition. Reactions are enhanced by b-blockers.

Diagnosis of anaphylaxis and venom hypersensitivity

Serum concentrations of mast-cell tryptase may be raised for up to 6 h after an anaphylactic attack, distinguishing this diagnosis from panic attacks and other causes of collapse. Type I hypersensitivity is confirmed by detecting venom-specific IgE in the serum using RAST or by prick skin tests. Patients who have suffered a systemic reaction, have a 50 to 60 per cent risk of a reaction to their next sting. Local reactions, even massive ones involving persistent swelling of the whole stung limb, in the absence of systemic symptoms do not predict a systemic reaction when stung again. Children who have generalized urticaria after a sting have only a 10 per cent chance of a systemic reaction when restung. Hypersensitivity to venom may be lost spontaneously, especially by children and young adults. The RAST test can be used for a postmortem diagnosis of Hymenoptera sting anaphylaxis.

Treatment

The barbed stings of Apidae remain embedded at the site of the sting and continue to inject venom, so they should be removed immediately by any means possible. Vespids can withdraw their stings and sting repeatedly. Wasp stings may become infected because these insects feed on rotting meat. Domestic meat tenderizer (papain) diluted roughly 1:5 with tap water is said to produce immediate relief of pain. Ice packs and aspirin are also effective. Systemic but not topical antihistamines can be used for more severe local reactions. Massive local reactions may require aspirin, non-steroidal anti-inflammatory agents, or even corticosteroids. **Systemic anaphylaxis** must be treated with 0.1 per cent (1:1000) adrenaline (epinephrine) (0.5–1 ml for adults; 0.01 mg/kg for children) given by intramuscular injection, or, if the patient is unconscious or pulseless, diluted 1:100 000, by intravenous injection. In rare cases, blood pressure fails to respond even to large doses of adrenaline and plasma expanders. These patients should be given cardiopulmonary resuscitation, selective bronchodilators such as salbutamol, pressor agents such as dopamine, and intravenous histamine H_1 and H_2 blockers such as chlorphenamine (chlor-pheniramine) maleate (10 mg for adults; 0.2 mg/kg for children) and cimetidine. Corticosteroids probably have no effect in acute anaphylaxis but may prevent relapses a few hours later. Patients who know they are hypersensitive should wear an identifying tag (such as provided by Medic-Alert in Britain) as they may be discovered unconscious after being stung. They should be trained to give themselves adrenaline using an 'EpiPen' or similar apparatus. Adrenaline delivered by a pressurized inhaler ('Medihaler-Epi') or squirted down the endotracheal tube

produces transient blood levels insufficient to combat anaphylaxis. Respiratory tract obstruction and shock are the main causes of death.

Severe envenoming from multiple stings by *Hymenoptera* spp. should be treated with adrenaline, intravenous antihistamines (doses as above), and corticosteroids. Intensive care is essential. Intravenous mannitol and bicarbonate may protect the kidneys from the damaging effects of myoglobinuria and haemoglobinaemia ('pigment nephropathy'), as in patients with the crush syndrome. Experimental antivenoms have been produced but are not yet commercially available. Exchange transfusion or plasmapheresis might be considered to remove venom in severe cases. Renal dialysis is often needed.

Prevention

Patients over the age of 25 years who have a history of systemic anaphylaxis following a sting and who have evidence of hypersensitivity (venom-specific IgE detectable in the serum or a positive skin test) should be considered for desensitization with purified venoms. This treatment proved significantly more effective than placebo or the previously used whole body extracts of *Hymenoptera* spp. in preventing anaphylactic reactions to sting challenge. Desensitization usually involves weekly visits to the hospital for at least 8 weeks for the administration of gradually increasing doses of venom. When protection has been demonstrated by the patient's ability to tolerate 100 µg of venom (equivalent to two stings) they are ready for maintenance therapy, usually 100 µg of venom every 4 to 8 weeks. A period of 2 to 5 years of maintenance desensitization is recommended, after which more than 90 per cent of subjects will remain protected against systemic reactions after stopping treatment. Desensitization is complicated by systemic reactions in 5 to 15 per cent of patients and by local reactions in 50 per cent.

Wasps are attracted by sweet things and meat in kitchens, greengrocers, orchards, and vineyards. Vespidae are attracted by brightly coloured floral patterns and perfumes. Hornets are attracted by light. Some species are so aggressive that their nests must be eradicated before their territory can be used for farming.

Venomous lepidoptera

The stinging hairs of some species of adult moths can cause contact dermatitis and urticaria ('lepidopterism'), while caterpillars can produce local or even systemic effects ('erucism'). Venomous lepidoptera are found in all parts of the world, but most cases of lepidopterism are reported from Middle and Southern America. Severe cutaneous urticating eruptions (Plate 15) can be caused by caterpillars of the genus *Megalopyge* (called 'pussmoths' in the southern United States) and by adult female moths of the genus *Hylesia* which have barbed setae ('flechettes') on their abdomens. Epidemics of stings by these moths have been described, especially from coastal areas of Brazil, Peru, Venezuela, and Mexico. Caterpillars of the genus *Lonomia* (*L. obliqua, L. achelous*) can inject through their bristles venom containing a plasminogen activator, plasmin-like enzyme, protease-activating factors V and XIII, and prothrombin, resulting in defibrinogenation, spontaneous bleeding into the skin, urinary tract, and even fatal intracranial haemorrhage (Plate 16). Laboratory findings in envenomed patients are decreased plasma fibrinogen, factor V, factor XIII, and plasminogen concentrations, as well as increased fibrin(ogen) degradation products and fibrinolytic activity but a normal platelet count.

In Pará State, Brazil, rubber tappers are frequently in contact with caterpillars (*Premolis semirufa*) whose stinging hairs can cause a disabling arthritis of the hands ('pararama').

Venomous coleoptera (beetles)

The most famous vesicating beetle is 'Spanish fly'—*Lytta vesicatoria* (Meloidae—blister beetles). Its venom contains cantharidin, which causes blistering when applied to the skin and priapism (hence its reputation as an aphrodisiac) and renal failure when given systemically or absorbed after eating the legs of frogs which have fed on meloid beetles.

'Nairobi eye' and similar blistering conditions in Australia and South-East Asia are caused by species of the genus *Paederus* (Staphylinidae) (Plate 17). A typical skin lesion (dermatitis linearis) consists of erythema, itching, and blistering caused by inadvertently crushing and smearing the beetle. Systemic symptoms such as fever, arthralgia, and vomiting may arise in severe cases. The active principle pederin is the most complex non-proteinaceous insect toxin known. Treatment is palliative. The toxin is easily spread to other sites such as the eye by fingers.

Scorpions (Scorpiones: Buthidae, Scorpionidae)

Species capable of inflicting fatal stings occur in North Africa and the Middle East (genus *Androctonus*, *Buthus*, and *Leiurus*); South Africa (*Parabuthus* spp.); India (*Mesobuthus tamulus*); North, Central and Southern America, Trinidad, and Tobago (genera *Tityus* (Plate 18) and *Centruroides*). Scorpion toxins target Na^+, K^+, Ca^{2+}, and Cl^- channels causing direct effects and the release of neurotransmitters such as acetylcholine and catecholamines.

Epidemiology

There have been no deaths from scorpion sting in the United States since 1968. However, between 1000 and 2000 deaths were reported each year in Mexico, but recently this number has declined dramatically. Case fatality is highest (formerly about 50 per cent) in children less than 4 years old. In Trinidad, stings by *Tityus trinitatis* are an occupational hazard of sugar cane and cocoa plantation workers. In the 1960s, there were 33 deaths in a group of 698 cases. Mortality was 25 per cent in children under 5 years compared to 0.25 per cent in adults. In Brazil, mortality was about 1 per cent in adults and 15 to 25 per cent in children less than 6 years old. In India, many people are stung by the red scorpion (*Mesobuthus tamulus*) with fatalities in adults and children.

Clinical features

Intense local pain is the commonest symptom. There may be slight local oedema and tender enlargement of regional lymph nodes, but stings by *Hemiscorpius lepturus* (Iran, Iraq, Pakistan, and Yemen) cause a local macular lesion which may blister and become necrotic resulting in sloughing of a large area of skin (Plate 19). Systemic symptoms may develop within minutes or be delayed for as much as 24 h. Symptoms vary depending on the species and geographical area. Scorpion venoms stimulate the release of acetylcholine and catecholamines, often resulting in initial cholinergic and later adrenergic symptoms. Early symptoms include vomiting, profuse sweating, piloerection, alternating brady- and tachycardia, abdominal colic, diarrhoea, loss of sphincter control, and priapism. Later, severe life-threatening cardiorespiratory effects may appear: hypertension, shock, tachy- and bradyarrhythmias, ECG changes (Fig. 26), and pulmonary oedema (Fig. 27) with or without evidence of myocardial dysfunction. Severe cardiovascular complications are particularly associated with stings by *Leiurus quinquestriatus, Mesobuthus tamulus*, and *Tityus* species. Stings by the North American *Centruroides sculpturatus* produce neurotoxic effects such as fasciculation, spasms, respiratory paralysis, and convulsions. Victims of the Trinidad black scorpion (*Tityus trinitatis*) develop severe abdominal pain with nausea, vomiting and haematemesis, hyperglycaemia, and biochemical evidence of acute pancreatitis.

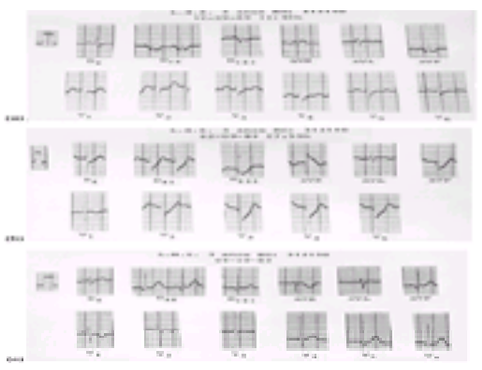

Fig. 26 Electrocardiographic changes caused by envenoming by the Brazilian scorpion (*Tityus serrulatus*). (By courtesy of Dr Carlos Amaral, Belo Horizonte, Brazil.) (a) ECG of 3-year-old girl 2 h after sting, showing sinus tachycardia and mild T-wave and ST-segment changes in the inferior and anterior leads. (b) Same patient 24 h after the sting, showing sinus tachycardia with marked ST down-sloping and ST displacement in the inferior and anterolateral leads. (c) ECG of the same patient 3 days after the sting, showing return to normal sinus rhythm and ST pattern.

Fig. 27 Pulmonary oedema caused by *Tityus serrulatus* envenoming. (By courtesy of Dr Carlos Amaral, Belo Horizonte, Brazil.) (a) Chest radiograph of a 9-year-old girl 24 h after sting, showing alveolar infiltrates in both lungs with air bronchograms. The heart size is increased. (b) Chest radiograph of 7-year-old girl 10 h after sting, showing bilateral pulmonary oedema with enlarged heart.

Treatment

Pain responds temporarily to local infiltration or ring block with local anaesthetic. Local injection of emetine or dehydroemetine is said to relieve the pain but may cause necrosis. Parenteral opiate analgesics, such as pethidine or morphine, may be required, but are said to be dangerous in victims of *C. sculpturatus*.

Antivenom is manufactured in a number of countries. Its use is strongly advocated in Africa and the Americas, but ancillary pharmacological treatment is regarded as being much more important in the Middle East and India. However, if specific antivenom is available it should be administered intravenously as soon as possible in patients with systemic envenoming and in young children stung by dangerous species, even before the development of these symptoms. Patients with cardiovascular symptoms benefit from vasodilator treatment with a-blockers (e.g. prazosin), calcium-channel blockers (e.g. nifedipine), or ACE inhibitors (e.g. captopril). Atropine should not be used except in cases of life-threatening sinus bradycardia. The use of cardiac glycosides and b-blockers is controversial. Anticonvulsants such as phenobarbital (phenobarbitone) are recommended for neurotoxic symptoms.

Prevention

Scorpions can be kept out of houses by including a row of ceramic tiles in the outside wall, making the doorsteps at least 20 cm high, and using residual insecticides, such as 1 per cent lindane or dieldrin powders.

Spiders (Araneae)

All but one family of this enormous order are venomous, but few species have proved dangerous to humans. Spiders bite with a pair of small fangs, the chelicerae, to which the venom glands are connected (Plate 20). There are four medically important genera of venomous spiders: *Loxosceles* cause necrotic araneism and *Latrodectus, Phoneutria*, and *Atrax* cause neurotoxic araneism.

Epidemiology

Loxosceles laeta (Plate 21) is widely distributed and causes many bites in people in Central and Southern America, especially in Chile, where the mortality ranges from 1 to 17 per cent. *L. reclusa*, the brown recluse spider, caused at least 200 bites and six deaths in the United States during the last century. More than 60 cases were reported from Texas between 1959 and 1962. Bites by *L. rufescens* have been reported from the Mediterranean region, North Africa, and Israel. Most bites from *Loxosceles* spp. occur in bedrooms while people are asleep or dressing, and in the United States a number of men were bitten on the genitalia while they sat on outdoor lavatories in which the spiders had spun their webs. *Latrodectus tredecemguttatus* (sometimes referred to, loosely, as 'tarantula') lives in fields in Mediterranean countries and has been responsible for a series of epidemics of bites. In Italy, 946 cases were reported between 1946 and 1951. *L. hasseltii*, the Australian redback spider (Plate 22), causes up to 340 bites each year in Australia and 20 deaths have been reported. The black widow spider (*L. mactans*) was responsible for 63 deaths in the United States between 1950 and 1959. Other species of *Latrodectus* are found in Latin America.

Phoneutria nigriventer (Plate 20 and Plate 23), the banana spider, causes bites and deaths in South American countries. These spiders have been imported to temperate countries on bunches of bananas, causing a few bites and deaths. The funnel web spiders (genus *Atrax*) are restricted to south-eastern Australia and Tasmania. The Sydney funnel web spider (*A. robustus*) occurs only within a 160-mile (256 km) radius of Sydney. The aggressive males of this species caused at least 13 deaths between 1927 and 1980. Members of the related genera *Hadronyche* and *Missulina* may be equally dangerous. In England, mild neurotoxic araneism has been described after bites by *Steatoda nobilis* (Theridiidae) and the woodlouse spider (*Dysdera crocata*).

Necrotic araneism

The bite itself is usually painless and unnoticed. Burning develops over several hours at the site of the bite, with swelling and development of a characteristic macular lesion (Plate 24(a)) which shows areas of red (vasodilatation), white (vasoconstriction), and blue (prenecrotic cyanosis). Eventually the lesion forms a blackened eschar (Plate 24(b) and Plate 24(c)) which sloughs in a few weeks, sometimes leaving a necrotic ulcer. The necrotic area may, rarely, cover an entire limb. Facial lesions may cause much oedema. Some 13 per cent of cases have systemic symptoms such as fever, headaches, scarlatiniform rash, jaundice, methaemoglobinaemia, and haemoglobinuria resulting from intravascular haemolysis. Renal failure may ensue. The average case fatality is about 5 per cent.

Neurotoxic araneism

The bite is very painful but local signs are minimal (*L. mactans*) or moderate (*L. hasselti*). After about 30 min there is painful regional lymphadenopathy, then headache and nausea and vomiting, with local sweating and piloerection ('gooseflesh') (Plate 25) and painful muscle spasms and tremors that may be severe enough to embarrass respiration. Local sweating and piloerection at the site of the bite is very suggestive of neurotoxic araneism. Other features include tachycardia, hypertension, restlessness, irritability, psychosis, priapism, and rhabdomyolysis. The 'facies latrodectismica' is a painful grimace caused by facial spasm and trismus associated with swollen eyelids, congested conjunctivas, flushing, and sweating (*L. tredecemguttatus*). Similar effects are seen in patients bitten by *Phoneutria* and *Atrax* spp.

First-aid treatment

In the case of bites by spiders with rapidly acting and potent venoms, such as *A. robustus* and *Hadronyche* species, firm crêpe bandaging and splinting of the bitten limb may delay venom spread until the patient reaches hospital.

Specific treatment

Antivenoms for a bite from *Latrodectus* spp. are made in Australia, the United States, Russia, Italy, Croatia, South, Africa and South America; for *Atrax* spp. bites in Australia; for *Loxosceles* spp. in Peru, Brazil, and Argentina; and for *Phoneutria* spp. in Brazil. Neurotoxic araneism seems more responsive to antivenom than does the necrotic type.

Supportive treatment

Oral dapsone (100 mg twice daily) is said to reduce the extent of necrotic lesions by inhibiting neutrophil degranulation. Calcium gluconate (10 ml of a 10 per cent solution, given by slow intravenous injection) relieves the pain of muscle spasms caused by the venom of *Latrodectus* spp. rapidly and more effectively than muscle relaxants such as diazepam or methocarbamol. Antihistamines, corticosteroids, a-blockers, and atropine have also been advocated. For necrotic araneism cause by *Loxosceles* spp. surgical debridement, corticosteroids, antihistamines, and hyperbaric oxygen all have their advocates, but there is no basis for recommending their use.

Ticks (Acari)

Taxonomy and epidemiology

Ticks, with mites, form the order Acari of the class *Arachnida*. Adult females of about 34 species of hard tick (family *Ixodidae*) and immature specimens of nine species of soft ticks (family *Argasidae*) have been implicated in human tick paralysis. The tick's saliva contains a neurotoxin which causes presynaptic neuromuscular block and decreased nerve-conduction velocity. The tick embeds itself in the skin with its barbed hypostome introducing the salivary toxin while it engorges with blood.

Although tick paralysis has been reported from all continents, most cases occur in western North America (*Dermacentor andersoni*), eastern United States (*D. variabilis*), and eastern Australia from north Queensland to Victoria (*Ixodes holocyclus*) known as the bush, scrub, paralysis-, or dog tick). In British Columbia there were 305 cases with a 10 per cent mortality between 1900 and 1968. About 120 cases have been reported in the United States, and in New South Wales there were at least 20 deaths between 1900 and 1945.

Clinical features

Ticks are picked up in the countryside or from domestic animals, particularly dogs, in the home. The majority of patients and almost all fatal cases are children. After the tick has been attached for about 5 or 6 days a progressive ascending, lower motor neurone paralysis develops with paraesthesiae. Often a child, who may have been irritable for the previous 24 h, falls on getting out of bed first thing in the morning and is found to be weak or ataxic. Paralysis increases over the next few days: death results from bulbar and respiratory paralysis and aspiration of stomach contents. Vomiting is a feature of the more acute course of *Ixodes holocyclus* envenoming.

This clinical picture is often misinterpreted as poliomyelitis. Other neurological conditions, including Guillain–Barré syndrome, paralytic rabies, Eaton–Lambert syndrome, myasthenia gravis, or botulism, may also be suspected. Diagnosis depends on finding the tick, which is likely to be concealed in a crevice, orifice, or hairy area of the body. The scalp is the commonest place. Fatal tick paralysis has been caused by a tick attached to the tympanic membrane.

Treatment

The tick must be detached without being squeezed. It can be painted with ether, chloroform, paraffin, petrol, or turpentine, or prised out between the partially separated tips of a pair of small, curved forceps. Following removal of the tick there is usually a rapid and complete recovery; but in Australia, patients have died after the tick has been detached. An antivenom, raised in dogs, is available in Australia, and, recently, rabbits have been used to produce an antitoxin against *I. holocyclus* saliva. This is recommended for severely affected or very young patients; 20 to 30 ml are given intravenously.

Centipedes (Chilopoda)

Many species of centipedes can inflict painful bites (Fig. 28), producing swelling, inflammation, and lymphangitis. Systemic effects such as vomiting, headache, cardiac arrhythmias, and convulsions are extremely rare and the risk of mortality was probably greatly exaggerated in the older literature. The most important genus is *Scolopendra* which is distributed throughout tropical countries. Local treatment is the same as for scorpion stings. No antivenom is available.

Fig. 28 Venom 'jaws' (modified limbs) of a Thai centipede (*Scolopendra* species). (Copyright DA Warrell.)

Millipedes (Diplopida)

Most species possess glands in each of their body segments which secrete, and in some cases squirt out, irritant liquids for defensive purposes. These contain hydrogen cyanide and a variety of aldehydes, esters, phenols, and quinonoids. Members of at least eight genera of millipedes have proved injurious to humans. Important genera are *Rhinocricus* (Caribbean), *Spirobolus* (Tanzania), *Spirostreptus* and *Iulus* (Indonesia), and *Polyceroconas* (Papua New Guinea). Children are particularly at risk when they handle or try to eat these large arthropods. When venom is squirted into the eye, intense conjunctivitis results and there may be corneal ulceration and even blindness. Skin lesions initially stain brown or purple, blister after a few days, and then peel. First aid is generous irrigation with water. Eye injuries should be treated as for snake venom ophthalmia (see above).

Leeches (Phylum Annelida, Class Hirudinea)

Leeches are blood-sucking, hermaphroditic, egg-laying annelids which have elongated annulated bodies. They attach themselves to leaves, rocks. or the host by a posterior sucker. To feed, the leech applies its anterior sucker containing the mouth armed with three radially arranged jaws which make a Y-shaped incision. Blood is sucked out by the action of the muscular pharynx. To prevent blood clotting, the saliva contains a histamine-like vasodilator and anticoagulants, such as: hirudin from the medicinal leech (*Hirudo medicinalis*), which inhibits thrombin and factor IXa; hementin from *Haementeria ghilianiı*, which is directly fibrinolytic; and hementerin from *H. depressa* (= *H. lutzı*), a plasminogen activator. Other enzymes include esterases, antitrypsin, antiplasmin, and antielastase. Recombinant hirudin is now produced as a therapeutic anticoagulant. The medicinal leech is still used by plastic surgeons to reduce haematomas under skin grafts; the wound may become infected with *Aeromonas hydrophila* which lives symbiotically in the leech's gut.

Two groups of leeches cause human morbidity and even mortality in tropical countries.

Land leeches

Species of the genera *Haemadipsa* and *Phyrobdella* are 1 to 8 cm long. They infest, often in enormous numbers, the damp, leafy floor and low vegetation of rainforests, choosing game trails and watering places. By standing on the posterior sucker and waving the anterior sucker, they can sense their prey with amazing efficiency. They drop on to the prey or pursue it with a looping or lashing motion. These leeches usually attach themselves to the lower legs or ankles and are adept at penetrating clothing, even long trousers tucked into socks and lace-up boots. The bite is usually painless and infested persons may not realize what has happened until they hear a squelching sound, notice that their feet are warm and wet, and see blood welling over the tops of their boots. Land leeches ingest about 1 ml of blood

in one hour and then drop off, but the wound continues to bleed for some time and forms a fragile clot.

Aquatic leeches

These species may be swallowed by those who drink stagnant water or even mountain stream water, or they may attack bathers, entering the mouth, nostrils, eyes, vulva, vagina, urethra, or anus. The enormous brightly coloured buffalo leech (*Hirudinaria manillensis*) of South-East Asia, is up to 16 cm long and can ingest 1 ml of blood in 10 min. *Limnatis nilotica* occurs around the Mediterranean, Middle East, and North Africa. *Myxobdella africana* occurs in East Africa. *Dinobdella ferox* (5 cm long) is found in Asia. Some aquatic leeches are very slow feeders and may remain attached for days or even weeks.

Clinical features

The main effect is blood loss, but other symptoms include pain caused by the bite, secondary infection, a residual itching, and phobia. Ingested aquatic leeches usually attach to the pharynx but may penetrate the bronchi or oesophagus. *H. manillensis* entering via the anus can reach the rectosigmoid junction of the bowel causing perforation and peritonitis. Patients with a leech in the pharynx often have a feeling of movement at the back of the throat with cough, hoarseness, stridor, breathlessness, epistaxis, haemoptysis, and haematemesis. Fatal upper airway obstruction may result. Bleeding may persist for up to a week after the leech has dropped off. In rural Thailand, vaginal bleeding in girls who have swum in ponds or canals is often attributable to infestation by aquatic leeches. Sexual abuse may be wrongly inferred if this diagnosis is not considered. Transmission of rinderpest and other viruses, leptospirosis, and *Trypanosoma cruzi* has been suggested but not proved. Secondary infection of medicinal leech bites by *Aeromonas hydrophila* has been described.

Treatment

Leeches will detach if a grain of salt, a lighted match or a cigarette, alcohol, turpentine, or vinegar are applied. Local bleeding can be stopped by applying a styptic, such as silver nitrate or a firm dressing. Aquatic leeches that have penetrated the respiratory, upper gastrointestinal, genitourinary tracts, or rectum must be removed by endoscope. Spraying with 30 per cent cocaine, 10 per cent tartaric acid, or dilute (1:10 000) adrenaline (epinephrine) makes the leech detach from the nasopharynx, larynx, trachea, or oesophagus, while irrigation with a concentrated salt solution may be effective in the genitourinary tract and rectum. Leeches should not be pulled off so roughly that the mouth parts are left in the wound as this will lead to a chronic infection. Antimicrobial treatment of secondary bacterial infections (e.g. of *Aeromonas hydrophila* with cefuroxime or a quinolone) may be required.

Prevention

This can be achieved by impregnating clothing, especially the bottoms of trousers and socks, with repellents such as dibutyl phthalate and diethyl toluamide. They may also be applied to the skin and the inside and outside of footwear. If these compounds are not available, invasion of footwear during jungle walks can be prevented, rather messily, by rolling a rope of tobacco in the tops of the socks and keeping the feet well soaked with water. Children should be discouraged from bathing in leech-infested waters and all drinking water should be boiled or filtered.

Further reading

Mechanical injuries caused by animals

Barss PG (1982). Injuries caused by garfish in Papua New Guinea. *British Medical Journal* **284**, 77–9.

Barss P, Ennis S (1988). Injuries caused by pigs in Papua New Guinea. *Medical Journal of Australia* **149**, 649–56.

Baxter DN, *et al.* (1984). The deleterious effects of dogs on human health. *Community Medicine* **6**, 29–36, 185–97, 198–203.

Middaugh JP (1987). Human injury from bear attacks in Alaska, 1900–1985. *Alaska Medicine* **29**, 121–6.

Wallet T (1978). *Shark attack and treatment of victims in South African waters.* Purnell, Cape Town.

Venomous mammals

Fenner PJ, Williamson JA, Myers D (1992). Platypus envenomation: a painful learning experience. *Medical Journal of Australia* **157**, 829–32.

Venomous snakes

Gopalakrishnakone P, ed (1994). *Sea snake toxinology*, pp 1–36. Singapore University Press,

Harvey AL, ed (1991). *Snake toxins. International encyclopedia of pharmacology and therapeutics*, Section 134. Pergamon Press, New York.

Meier J, White J, eds (1995). *Handbook of clinical toxicology of animal venoms and poisons.* CRC Press, Boca Raton, FL.

Reid HA (1976). Adder bites in Britain. *British Medical Journal* **2**, 153–6.

Reid HA, *et al*. (1963). Clinical effects of bites by Malayan viper (*Ancistrodon rhodostoma*). *Lancet* **i**, 617–21.

Russell FE (1980). *Snake venom poisoning.* Lippincott, Philadelphia, PA.

Sutherland SK, Tibballs J (2001). *Australian animal toxins. The creatures, their toxins and care of the poisoned patient*, 2nd edn. Oxford University Press, Melbourne.

Theakston RDG, Warrell DA (1991). Antivenoms: a list of hyperimmune sera currently available for the treatment of envenoming by bites and stings. *Toxicon* **29**, 1419–70.

Warrell DA (1990). Treatment of snake bite in the Asia-Pacific region: a personal view. In: Gopalakrishnakone P, Chou LM, eds. *Snakes of medical importance (Asia-Pacific region)*. Singapore University Press.

Warrell DA (2003). Epidemiology, clinical features and management of snakebites in Central and South America. In Campbell J, Lamar WW, Greene H, eds. *Venomous reptiles of the Americas*. Cornell University Press, Ithaca NY (in press).

Warrell DA, ed (1999). WHO/SEARO Guidelines for the clinical management of snake bites in the South East Asian region. *South East Asian Journal of Tropical Medicine and Public Health* **30**(Suppl 1), 1–85.

Venomous lizards

Bogert CM, Martin del Campo R (1956). The Gila monster and its allies. *Bulletin of the American Museum of Natural History* **109**, 1–238.

Preston CA (1989). Hypotension, myocardial infarction and coagulopathy following Gila monster bite. *Journal of Emergency Medicine* **7**, 37–40.

Russell FE, Bogert CM (1981). Gila monster, venom and bite—a review. *Toxicon* **19**, 341–59.

Poisonous amphibians

Brubacher JK, *et al*. (1996). Treatment of toad venom poisoning with digoxin-specific Fab fragments. *Chest* **110**, 1282–8.

Daly JW (1998). Thirty years of discovering arthropod alkaloids in amphibian skin. *Journal of Natural Products* **61**, 162–72.

Hitt M, Ettinger DD (1986). Toad toxicity. *New England Journal of Medicine* **314**, 1517–18.

Kwan T, Dino Paiusco A, Kohl L (1992). Digitalis toxicity caused by toad venom. *Chest* **102**, 949–50.

Myers CW, Daly JW, Malkin B (1978). A dangerously toxic new frog (Phyllobates) used by the Emberá Indians of Western Colombia with discussion of blowgun fabrication and dart poisons. *Bulletin of*

the American Museum of Natural History **161**(Art 2), 307–66.

Poisonous birds

Dumbacher JP, *et al*. (1992). Homobatracho-toxin in the genus *Pitohui*: chemical defense in birds? *Science* **258**, 799–800.

Venomous fish

Castex MN (1967). Fresh water venomous rays. In: Russell FE, Saunders PR, eds. *Animal toxins*, pp 167–76. Pergamon Press, Oxford.

Halstead BW (1988). *Poisonous and venomous marine animals of the world*, 2nd revised edition. Darwin Press, Princeton.

Lehane L, Rawlin GT (2000). Topically acquired bacterial zoonoses from fish: a review. *Medical Journal of Australia* **173**, 256–9.

Mareti, Z. (1973). Some epidemiological, clinical and therapeutic aspects of envenomation by weever fish sting. In: De Vries A, Kochva E, eds. *Toxins of animal and plant origin*, pp 1055–65. Gordon and Breach, New York.

Sutherland SK, Tibballs J (2001). *Australian animal toxins. The creatures, their toxins and care of the poisoned patient*, 2nd edn. Oxford University Press, Melbourne.

Williamson JA, *et al.*, eds (1996). *Venomous and poisonous marine animals: a medical and biological handbook.* University of New South Wales Press, Sydney.

Poisoning by ingestion of aquatic animals

Bagnis RA, *et al.* (1979). Clinical observations on 3009 cases of Ciguatera (fish poisoning) in the Southern Pacific. *American Journal of Tropical Medicine and Hygiene* **28**, 1067–73.

Daranas AH, Norte M, Fernández JJ (2001). Toxic marine micro-algae. *Toxicon* **39**, 1101–32.

Halstead BW (1988). *Poisonous and venomous marine animals of the world*, 2nd revised edition. Darwin Press, Princeton.

Trishnananda M, *et al.* (1966). Poisoning following the ingestion of the horseshoe crab (*Carcinoscorpius rotundicauda*): report of four cases in Thailand. *Journal of Tropical Medicine and Hygiene* **69**, 194–6.

Uragoda CG, Kottegoda SR (1977). Adverse reactions to isoniazid on ingestion of fish with a high histamine content. *Tubercle* **58**, 83–9.

Williamson JA, *et al.*, eds. (1996). *Venomous and poisonous marine animals: a medical and biological handbook.* University of New South Wales Press, Sydney.

World Health Organization (1984). *Aquatic (marine and fresh water) biotoxins.* Environmental Health Criteria, 37. WHO, Geneva.

Poisoning by ingestion of carp's gallbladder

Lin YF, Lin SH (1999). Simultaneous acute renal and hepatic failure after ingesting raw carp gall bladder. *Nephrology, Dialysis and Transplantation* **14**, 2011–12.

Venomous marine invertebrates

Beadnell CE, *et al.* (1992). Management of a major box jellyfish (*Chironex fleckeri*) sting. *Medical Journal of Australia* **156**, 655–8.

Halstead BW (1988). *Poisonous and venomous marine animals of the world*, 2nd revised edn. Darwin Press, Princeton.

Hartwick R, *et al.* (1980). Disarming the box-jellyfish. Nematocyst inhibition in *Chironex fleckeri. Medical Journal of Australia* **1**, 15–20.

Martin JC, Audley I (1990). Cardiac failure following Irukandji envenomation. *Medical Journal of Australia* **153**, 164–6.

Olivera BM, Cruz LJ (2001). Conotoxins, in retrospect. *Toxicon* **39**, 7–14.

Pereira PL, *et al.* (2000). Pressure immobilisation bandages in first-aid treatment of jelly fish envenomation: current recommendation reconsidered. *Medical Journal of Australia* **173**, 650–2.

Sutherland SK, Tibballs J (2001). *Australian animal toxins. The creatures, their toxins and care of the poisoned patient*, 2nd edn. Oxford University Press, Melbourne.

Williamson JA, *et al.*, eds. (1996). *Venomous and poisonous marine animals: a medical and biological handbook.* University of New South Wales Press, Sydney.

Venomous arthropods

British Society for Allergy and Clinical Immunology Working Party (1993). Position paper on allergen immunotherapy. *Clinical and Experimental Allergy* **23**(Suppl 3), 19–22.

de Shazo RD, Butcher BT, Banks WA (1990). Reactions to the stings of the imported fire ant. *New England Journal of Medicine* **323**, 462–6.

França FOS, *et al*. (1994). Severe and fatal mass attacks by 'killer' bees (Africanised honey bees— *Apis mellifera scutellata* in Brazil: clinicopathological studies with measurement of serum venom concentrations. *Quarterly Journal of Medicine* **87**, 269–82.

Hunt J Jr, *et al.* (1978). A controlled trial of immunotherapy in insect hypersensitivity. *New England Journal of Medicine* **2991**, 157–61.

McHugh SM, *et al*. (1995). Bee venom immunotherapy induces a shift in cytokine responses from a TH-2 to a TH-1 dominant pattern: comparison of rush and conventional immunotherapy. *Clinical and Experimental Allergy* **25**, 828–38.

Mueller UR (1990). *Insect sting allergy. Clinical picture, diagnosis and treatment*. Gustav Fischer, Stuttgart.

Piek T (1986). *Venoms of the* Hymenoptera. *Biochemical, pharmacological and behavioural aspects.* Academic Press, London.

Winston ML (1992). *Killer bees. The Africanized Honey Bee in the Americas.* Harvard University Press, Cambridge, Mass.

Venomous lepidoptera

Kelen EMA, Picarelli ZP, Duarte AC (1995). Hemorrhagic syndrome induced by contact with caterpillars of the genus *Lonomia* (Saturniidae, Hamileucinae). *Journal of Toxicology. Toxin Reviews* **14**, 283–308.

Venomous coleoptera (beetles)

Eisner T, *et al*. (1990). Systemic retention of ingested cantharidin by frogs. *Chemoecology* **1**, 57–62.

Frank JH, Kanamitsu K (1987). *Paederus sensulato* (Coleoptera: Staphylinidae): natural history and medical importance. *Journal of Medical Entomology* **24**, 1555–91.

Roberts JI, Tonking HD (1935). Notes on an East African vesicant beetle, *Pederus crebripunctatus* Epp. *Annals of Tropical Medicine and Parasitology* **29**, 415–20.

Southcott RV (1989). Injuries from Coleoptera. *Medical Journal of Australia* **151**, 654–9.

Scorpions

Amaral CFS, *et al*. (1991). Electrocardiographic, enzymatic and echocardiographic evidence of myocardial damage after *Tityus serrulatus* scorpion poisoning. *American Journal of Cardiology* **67**, 655–7.

Amaral CFS, De Rezende NA, Freire-Maia L (1993). Acute pulmonary edema after *Tityus serrulatus* scorpion sting in children. *American Journal of Cardiology* **71**, 242–5.

Bawaskar HS, Bawaskar PH (1992). Management of the cardiovascular manifestations of poisoning by the Indian red scorpion (*Mesobuthus tamulus*). *British Heart Journal* **68**, 478–80.

Bettini S, ed (1978). *Athropod venoms. Handbook of experimental pharmacology*, Vol 48, p 977. Springer-Verlag, Berlin.

Brownell P, Polis G, eds (2001). *Scorpion biology and research*. Oxford University Press, New York.

Fet V, *et al*., eds (2000). *Catalog of the scorpions of the world. (1758–1998)*. New York Entomological Society, New York.

Freire-Maia L, Campos JA, Amaral CFS (1996). Treatment of scorpion envenoming in Brazil. In: Bon C, Goyffon M, eds. *Envenomings and their treatments*, pp 301–10. Edition Fondation Marcel Mérieux, Lyon.

Gueron M, Ilia R, Sofer S (1992). The cardiovascular system after scorpion envenomation. A review. *Clinical Toxicology* **30**, 245–8.

Keegan HL (1980). *Scorpions of medical importance*. University Press of Mississippi, Jackson.

Polis GA, ed (1990). *The biology of scorpions*. Stanford University Press, Stanford, CA.

Radmanesh M (1990). Clinical study of *Hemiscorpion lepturus* in Iran. *Journal of Tropical Medicine and Hygiene* **93**, 327–32.

Waterman JA (1938). Some notes on scorpion poisoning in Trinidad. *Transactions of the Society of Tropical Medicine and Hygiene* **31**, 607–24.

Spiders

Clark RF, *et al.* (1992). Clinical presentation and treatment of black widow spider envenomation: a review of 163 cases. *Annals of Emergency Medicine* **21**, 782–7.

Mareti Z, Lebez D (1979). *Araneism with special reference to Europe.* Novit, Pula-Ljubjan, Yugoslavia.

Southcott RV (1976). Arachnidism and allied syndromes in the Australian region. *Records of the Adelaide Children's Hospital* **1**, 97–186.

Sutherland SK, Tibballs J (2001). *Australian animal toxins. The creatures, their toxins and care of the poisoned patient*, 2nd edn. Oxford University Press, Melbourne.

Warrell DA, *et al.* (1991). Neurotoxic envenoming by an immigrant spider (*Steatoda nobilis*) in southern England. *Toxicon* **29**, 1263–5.

Ticks

Gothe R, Kunze K, Hoogstraal H (1979). The mechanism of pathogenicity in the tick paralyses. *Journal of Medical Entomology* **16**, 357–69.

Murnaghan MF, O'Rourke FJ (1978). Tick paralysis. In: Bettini S, ed. *Arthropod venoms. Handbook of experimental pharmacology*, Vol 48, p 419. Springer-Verlag, Berlin.

Pearn J (1977). The clinical features of tick bite. *Medical Journal of Australia* **2**, 313.

Stone BF (1987). Toxicoses induced by ticks and reptiles in domestic animals. In: Harris JB, ed. *Natural toxins: animal, plant and microbial*, pp 56–71. Oxford University Press, Oxford.

Millipedes

Bettini S, ed (1978). *Arthropod venoms. Handbook of experimental pharmacology*, Vol 48, p 977. Springer-Verlag, Berlin.

Radford AJ (1975). Millipede burns in man. *Tropical and Geographical Medicine* **27**, 279–87.

Leeches

Adams SL (1988). The medicinal leech. A page from the Annelids of Internal Medicine. *Annals of Internal Medicine* **109**, 399–405.

Cundall DB (1986). Severe anaemia and death due to the pharyngeal leech *Myxobdella africana*. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **80**, 940–4.

Editorial (1992). Hirudins: return of the leech? *Lancet* **340**, 579–80.

Keegan HL (1963). Leeches as pests of man in the Pacific region. In: Keegan HL, McFarlane WR, eds. *Venomous and poisonous animals and noxious plants of the Pacific region*, pp 99–104. Pergamon Press, Oxford.

Sawyer RT (1986). *Leech biology and behaviour.* Oxford University Press, Oxford.

Snower DP, *et al.*, eds (1989). *Aeromonas hydrophila* infection associated with the use of medicinal leeches. *Journal of Clinical Microbiology* **27**, 1421–2.

Websites

WCH Clinical Toxinology Resource (covers venomous animals worldwide).

http://www.toxinology.com/.

Venomous Snake Systematics.

sbsweb.bangor.ac.uk/~ bss166/update.htm.

Australian Venom Unit.

www.pharmacology.unimelb.edn.au/avruweb/index.htm.

Shark attacks.

www.flmnh.ufl.edu/fish/Sharks/ISAF/ISAF.htm.

Arachnids.

www.ufsia.ac.be/Arachnology.

8.3Poisonous plants and fungi

M. R. Cooper, A. W. Johnson, and H. Persson

There are relatively few plants and fungi that cause serious poisoning, but some have a very wide geographical distribution: in this group are oleanders (*Nerium* and *Thevetia* spp.), monkshood (*Aconitum napellus*), thorn apple (*Datura* spp.), angels' trumpets (*Brugmansia* spp.), *Euphorbia* spp. and the death cap fungus (*Amanita phalloides*).

Most incidents with plants occur in small children and are entirely accidental. Adults may sometimes confuse edible and toxic plants, but are more often poisoned by deliberate ingestion of those with psychoactive properties; ingestion of toxic plants as a means of suicide occurs sporadically in most parts of the world but in Sri Lanka and South India yellow oleander (*Thevetia peruviana*) is a common agent of suicide. In many countries poisoning by fungi is a more prominent medical problem than poisoning by plants, and is mostly the result of misidentification.

The general principles for treatment of poisoning outlined elsewhere (see Chapter 8.1) also apply to plant and fungal poisoning. In most cases only simple measures (applied at home) are necessary, such as rinsing of the mouth and giving peroral fluids for dilution. However, if a highly toxic plant or fungus has been ingested in significant amounts, energetic gut decontamination is required, including gastric emptying and administration of activated charcoal, to which most of the toxins bind well. Symptomatic and supportive care should be given as necessary, and specific antidotes are available for a few types of poisoning.

Plant poisoning

Clinical effects can be used to classify plant poisoning. However, poisoning by some plants causes symptoms of more than one type, and the severity may vary according to the sensitivity of the individual.

Digestive tract irritation

Some plants have very irritant sap: that of cuckoo pint (*Arum maculatum*), dumb cane (*Dieffenbachia* spp.) (Plate 1), elephant's ear (*Philodendron* spp.), and black bryony (*Tamus communis*) contains calcium oxalate, and that of *Euphorbia* spp., mezereon, and spurge laurel (*Daphne* spp.) diterpene esters. These cause immediate soreness, reddening, and even blistering of the lips and mouth, salivation, and dysphagia.

With most plants, however, the first signs of poisoning are nausea, abdominal cramps, and vomiting. In many cases the vomiting may eliminate the poisonous substances and partly prevent the development of further toxic effects; elimination is also promoted by the associated diarrhoea. Toxic agents causing irritation of the stomach and intestines include anthraquinones found in *Aloe* spp. and purging buckthorn (*Rhamnus cathartica*), cytisine in *Laburnum* spp. (Plate 2), lectins in *Jatropha*, protoanemonin in *Anemone* spp. and *Helleborus* spp., saponins in horse chestnut (*Aesculus hippocastanum*) and ivy (*Hedera* spp.), and viscotoxins in mistletoe (*Viscum album*). Unidentified irritants are present in white bryony (*Bryonia dioica*), holly (*Ilex* spp.), *Lantana camara*, privet (*Ligustrum vulgare*), pokeweed (*Phytolacca americana*), and snowberry (*Symphoricarpos alba*). With other plants, some of which are much more toxic than those mentioned so far, the start of gastrointestinal symptoms is delayed for several hours (up to 2 days). The toxins responsible include colchicine in autumn crocus (*Colchicum autumnale*) and glory lily (*Gloriosa superba*), the very highly toxic lectins in jequirity beans (*Abrus precatorius*) (Plate 3), castor beans (*Ricinus communis*) (Plate 4), and false acacia (*Robinia pseudoacacia*), oxalic acid in some members of the Polygonaceae family, solanine in *Solanum* spp. (Plate 5), and unidentified agents, such as that in spindle (*Euonymus europaeus*).

After ingestion of plants containing irritant sap, rational procedures to follow are rinsing of the mouth and dilution by drinking fluids; the addition of activated charcoal may be beneficial. Gastric emptying is rarely indicated, unless large amounts or highly toxic plants have been ingested. Symptomatic care is given as required.

Cardiovascular disturbances

In addition to the well-known effects of the cardiac glycosides of foxglove (*Digitalis purpurea*) (Plate 6), there are various other plants that produce the digitalis-like effects seen in digoxin overdosage (see Chapter 15.5.1): lily of the valley (*Convallaria majalis*), bluebell (*Hyacinthoides non-scripta*), oleander (*Nerium oleander*) (Plate 7), and yellow oleander (*Thevetia peruviana*). Cardiovascular effects are also induced by aconitine in *Aconitum* spp. (Plate 8) and larkspurs (*Delphinium* spp.). Aconitine causes a burning sensation in the mouth and pharynx, gastrointestinal symptoms, paraesthesia, and, in particular, cardiac arrhythmias (mainly ventricular) that may prove fatal. Other examples of cardiovascular toxins are: andromedotoxins (grayanotoxins) in mountain laurel (*Kalmia latifolia*), *Menziesia* spp., *Pieris* spp., and *Rhododendron* spp., phoratoxin in American mistletoes (*Phoradendron* spp.), proveratrines in *Veratrum* spp., taxines in yews (*Taxus* spp.) (Plate 9), and veratrine in death camas (*Zigadenus* spp.). These cardiac glycosides also irritate the digestive tract.

Gastric emptying (avoid ipecacuanha) is performed and activated charcoal administered if the ingested dose and time since ingestion indicate its usefulness. In addition to symptomatic and supportive care, digoxin-specific antibodies (ovine Fab fragments such as 'Digibind' and 'DigiTab') have been used successfully in reversing life-threatening digitalis effects in plant poisoning, notably poisoning by *Thevetia peruviana* (Apocynaceae), a common agent of suicide in South Asia, and anecdotally in poisoning by other Apocynaceae and even toad skin toxins. In experimental animals, fructose-1,6-diphosphate has also proved effective. Ingestion of aconitine alkaloids also requires vigorous gastrointestinal decontamination, careful cardiac monitoring, and treatment of arrhythmias.

Central nervous effects

Some plants contain hallucinogenic compounds, for which they are smoked, chewed, eaten, or infused in water to make teas. Among the plant hallucinogens are tetrahydrocannabinols in cannabis (*Cannabis sativa*), alkaloids in khat (*Catha edulis*) (Plate 10), D-lysergic acid amide in morning glory (*Ipomoea* spp.), mescaline in peyote (*Lophophora williamsii*), myristicin in nutmeg (*Myristica fragrans*), and vincristine and vinblastine in periwinkle (*Vinca* spp.); others contain hyoscine (see atropine-like effects below).

Patients who have taken hallucinogens require reassurance and supportive treatment.

A number of plants contain convulsants, although actual clinical cases mainly involve Umbelliferae, such as cowbane (*Cicuta virosa*) and hemlock water dropwort (*Oenanthe crocata*) that contain cicutoxin and oenanthotoxin respectively. Other convulsants include hypoglycin A in ackee (*Blighia sapida*), coriamyrtin in *Coriaria*

myrtifolia, anthracenones in *Karwinskia humboldtiana*, tetranortriterpenes in chinaberry (*Melia azedarach*), alkaloids in moonseed (*Menispermum canadense*), podophylloresin in May apple (*Podophyllum peltatum*), and strychnine and brucine in nux vomica (*Strychnos nux-vomica*). Cicutoxin, one of the most potent convulsants known, may cause an extremely dramatic syndrome with gastroenteritis, increased secretions, and long-lasting intense episodes of generalized tonic-clonic convulsions, resulting in severe metabolic acidosis and multiple organ failure. Consumption of unripe ackee fruit (*Blighia sapida*) causes epidemic vomiting (Jamaican vomiting sickness) and a toxic hypoglycaemic syndrome which can result in fatal encephalopathy.

Treatment is symptomatic and supportive, aiming at control of convulsions and secondary effects.

Atropine-like effects

The tropane alkaloids atropine, hyoscine (scopolamine), and hyoscyamine present in some plants of the Solanaceae competitively inhibit the muscarinic effects of acetylcholine and block the parasympathetic nervous system. This will cause a classic anticholinergic syndrome. *Datura stramonium* is used as a hallucinogen in parts of Africa (for example in Niger where it is known as 'sobi-lobi') (Plate 18). The symptoms may vary depending on the proportion of the alkaloids in the different plants. Typical is a warm, dry-skinned and dry-mouthed, thirsty, anxious, excited, and hallucinating patient. Facial flushing, tachycardia, and mydriasis are also common. In serious cases, arrhythmias, urinary retention, psychosis, convulsions, coma, and respiratory failure may ensue; fatal outcome has been reported. The plants in this group include angels' trumpets (*Brugmansia* spp.) (Plate 17) and thorn apple (*Datura stramonium*), which are among the most common plants involved in poisoning cases worldwide, deadly nightshade (*Atropa belladonna*) (Plate 16) and henbane (*Hyoscyamus niger*). Unilateral dilatation of the pupil (gardener's mydriasis), usually without systemic effects, may be seen in gardeners who rub their eyes after handling members of the Solanaceae such as angels' trumpets (*Brugmansia* spp.).

Gastrointestinal decontamination is performed, if appropriate, considering ingested dose and time since ingestion. A quiet environment, with reassurance and good symptomatic and supportive care, is basic. In cases where typical central and peripheral anticholinergic effects are present, specific treatment with physostigmine, 1 to 2 mg intravenously in adults (children 0.02 to 0.04 mg/kg), may be useful in reversing hallucinations, delirium, and psychotic behaviour. Physostigmine should be withheld if cardiotoxic agents have been ingested or there is widening of the QRS complexes.

Nicotine-like effects

Nicotine in tobacco plants (*Nicotiana tabacum*) and other alkaloids with similar actions, such as coniine from hemlock (*Conium maculatum*) (Plate 15), used for judicial executions in ancient Athens, first stimulate and then paralyse all autonomic ganglia. Centrally, small doses cause respiratory stimulation, while large doses can lead to convulsions and arrest of respiration. Many other symptoms may occur but they are complex and unpredictable. Nicotine-like effects occur with gelsemine and related alkaloids in yellow jessamine (*Gelsemium sempervirens*), cytisine in *Laburnum* spp., and lobeline in *Lobelia* spp.

Gastrointestinal decontamination by gastric lavage (not ipecacuanha-induced emesis because of rapid onset of symptoms) and administration of activated charcoal are indicated at an early stage. In severe cases, extensive symptomatic and supportive care may be necessary.

Liver damage

Various plants in the Compositae (*Senecio* spp.), Leguminosae (*Crotalaria* spp.), and Boraginaceae (*Heliotropium* and *Symphytum* spp.) (Plate 11) as well as some in other families contain pyrrolizidine alkaloids, which principally affect the liver. Poisoning has occurred from contamination of cereal crops by these plants and their subsequent incorporation into bread, and from their use in herbal medicines and bush teas. These alkaloids can cause acute damage to the liver (veno-occlusive disease), which has occurred mainly in Jamaica, India, and Afghanistan. Symptoms, including nausea, abdominal pain and distension, hepatomegaly, and sometimes fever and vomiting, first appear a few days after ingestion. A chronic cirrhosis of the liver can occur in people ingesting small quantities of pyrrolizidine alkaloids over a long period.

Treatment can only be supportive as, once absorbed, there is no specific method of preventing the toxic effects of these alkaloids.

Kidney damage

This can occur after ingestion of plants rich in oxalates, for example the leaves of rhubarb (*Rheum rhabarbarum*) and docks and sorrels (*Rumex* spp.). Nephrotoxicity may also occur after ingestion of terpene-containing plants, such as *Daphne mezereum* and *Juniperus sabina*.

Fluid replacement is essential and generous hydration advisable to promote renal excretion of oxalates.

Hydrocyanic acid toxicity

Parts of some plants contain relatively high concentrations of cyanogenic glycosides. The most likely sources are the kernels of fruits of *Prunus* species (almonds, apricots, cherries, peaches, etc.) or of loquat (*Eriobotrya japonica*), a large number of apple pips (*Malus* spp.), the berries or leaves of the cherry laurel (*Prunus laurocerasus*), or inadequately prepared cassava (*Manihot esculenta*) (see below). The glycosides will release cyanide only when the kernels or other plant material are being chewed. This takes some time, so the onset of symptoms is not so rapid as after ingestion of inorganic cynanide compounds. In practice, cyanide poisoning from plants (with the exception of cassava) is a rare phenomenon.

When it is suspected that large amounts of cyanogenic glycosides have been ingested, gastrointestinal decontamination is indicated. If typical symptoms ensue, treatment is the same as for hydrocyanic acid poisoning from any other source (see Chapter 8.1).

Skin damage

The most common form of skin damage by plants is a non-allergic dermatitis that results from direct contact with various plants that contain irritants, for example the stinging hairs of the stinging nettle (*Urtica dioica*), the diterpene-containing latex of *Euphorbia* spp., or the calcium oxalate crystals of *Dieffenbachia* spp. and other members of the Araceae family. There are also allergic forms of dermatitis that result from hypersensitivity to plant allergens. The most common and severe forms of allergic contact dermatitis in the United States are caused by poison ivy (*Rhus radicans*) (Plate 12) and western poison oak (*Toxicodendron diversilobum*), and in the United Kingdom by a primula (*Primula obconica*), but such cutaneous hypersensitivity can occur to a very large number of vascular plants.

Contact with the sap of some plants followed by exposure to sunlight (ultraviolet radiation) can give rise to a characteristic skin reaction. Phototoxic psoralens (furanocoumarins) are present in various members of the Umbelliferae, including giant hogweed (*Heracleum mantegazzianum*) (Plate 13), other *Heracleum* spp., and parsnips (*Pastinaca sativa*), and also celery (*Apium graveolens*), especially when infected with the pink rot fungus (*Sclerotinia sclerotiorum*). Other psoralen-containing plants are rue (*Ruta graveolens*) (Plate 14) and the gas plant (*Dictamnus albus*). Typical lesions in dermatitis caused by psoralens are erythema, papules, vesicles, and enormous bullae localized to exposed areas of skin.

Treatment is as for a chemical burn. Psoralens can also induce hyperpigmentation that can last for several months.

Food plant toxicity

The following toxins are present in plants used regularly for food:

1. Lectins. These phytohaemagglutinins are glycoproteins, and beans (Leguminosae) are the main food source. The lectins are not readily digested by pepsin and have a strong affinity for the intestinal mucosa, where they prevent absorption of carbohydrates. Adequate cooking destroys the lectins, but eating raw or incompletely cooked beans can cause diarrhoea, while long-term exposure can lead to retarded growth and may even be fatal.
2. Cyanogens. Cyanogenic glycosides are present in some staple foods, such as cassava, sweet potato, and yam, all of which can be made safe to eat by adequate soaking, drying, or fermentation. Chronic poisoning may cause tropical ataxic neuropathy and spastic paraparesis ('konzo').
3. Alkaloids. Dangerously high quantities of these may develop in some plant foods, for example potato tubers that have sprouted or been stored in the light and

become green. The green colour is chlorophyll and is harmless, but under conditions where greening occurs the glycoalkaloid solanine and its derivatives will also have been produced; these exhibit anticholinesterase activity.

4. Oxalates. These accumulate in some members of the Polygonaceae, notably rhubarb, of which only the red leaf stalks should be eaten, and then only after cooking.
5. Polyphenols. These are a possible cause of upper digestive tract cancers that may develop after eating sorghum, or taking teas or alcoholic drinks containing high concentrations of tannins.

Some specific diseases attributable to plants are: lathyrism, a paralytic disease caused by a neurotoxic amino acid in chick peas (*Lathyrus sativus*); favism that occurs among natives of some Mediterranean and Middle Eastern countries who have a genetic deficiency of glucose-6-phosphate dehydrogenase and cannot digest broad (fava) beans (*Vicia faba*); and Jamaican vomiting sickness that results from eating the unripe fruits of the ackee (*Blighia sapida*). Other foods that cause poisoning are the seeds of *Mucuna pruriens*, flour ground from the cones of cycads (for example *Zamia* spp.) (Plate 19), and young fronds of bracken fern (*Pteridium aquilinum*).

Herbal medicines

It should not be assumed that herbal preparations are safe because they are 'natural', and some may have harmful effects that are not immediately obvious, for example carcinogenicity, hepatotoxicity, or teratogenicity. For example, plants of the genus *Aristolochia* are constituents or contaminants of some Chinese herbal medicines (such as 'Mu Tong' and 'Fang Ji') and can cause genotoxic carcinogenesis and interstitial nephritis. Adulteration of herbal medicines with cheaper ingredients, heavy metals, and orthodox drugs is not uncommon.

Fungal poisoning

Fungal poisoning (mycetismus) is a sporadic problem in many parts of the world. It is usually accidental, occasionally homicidal or suicidal, and may become epidemic when conventional food is scarce in times of war or famine. In 2000, outbreaks of mushroom poisoning involving 212 people with 17 deaths were reported from Voronezh and Volgograd in Russia.

The toxicity of fungi may vary with location, season, and in different years; for some mushroom toxins there may also be a variation in the susceptibility of individuals. Cases of fungal poisoning can be classified into those that develop signs of toxicity within a few hours of ingestion, or after a delay of 10 (occasionally 6) to 24 h or several days.

Symptoms that develop within a few hours

Gastroenteritis

Few of the chemical agents responsible for this type of poisoning have been identified. The fungi in this group include flat-capped psalliota (*Agaricus placomyces* or *A. praeclavesquamosus*), yellow-staining mushroom (*Agaricus xanthodermus*), honey fungus (*Armillaria mellea*), the devil's mushroom (*Boletus satanas*), *Chlorophyllum molybdites* (*Lepiota morgani*), *Entoloma* spp., *Hebeloma* spp., wax caps (*Hygrocybe* spp.), sulphur tuft (*Hypholoma fasciculare*), milk caps (*Lactarius* spp.), shaggy parasol (*Macrolepiota rhacodes*), broad-gilled agaric (*Megacollybia platyphylla*), Jack-o'lantern or copper trumpet (*Omphalotus olearius*), pink coral fungus (*Ramaria formosa*), brittle gills (*Russula* spp.), earthballs (*Scleroderma* spp.), and spotted tricholoma (*Tricholoma pardinum*).

Gastrointestinal symptoms usually start within a few hours and resolve fairly quickly, but are sometimes more intense and result in fluid and electrolyte imbalance.

Treatment is symptomatic and admission to hospital is seldom necessary, but may occasionally be required, especially in children and the elderly.

Cholinergic symptoms

The parasympathomimetic effects of muscarine, which is the toxic ingredient of certain *Inocybe* and *Clitocybe* species, include abdominal pain and diarrhoea, sweating, lacrimation, salivation, myosis, bronchorrhoea, and sometimes bronchospasm, bradycardia, and hypotension. Only trace amounts of muscarine occur in fly agaric (*Amanita muscaria*) (Plate 20) after which the toxin was named.

Treatment is specific and includes 1 to 2 mg of atropine in adults (children 0.02 mg/kg) that effectively antagonizes muscarinic overstimulation. The dose is repeated as required.

Mental confusion

This may result from poisoning by panther cap (*Amanita pantherina*), fly agaric (*Amanita muscaria*), and *Amanita strobiliformis*, whose active ingredients are isoxazoles (ibotenic acid and muscimol) that act as g-amino-butyric acid agonists. Anxiety, euphoria, visual disturbances, erratic behaviour, hallucinations, and gastrointestinal symptoms may occur, and peripheral anticholinergic symptoms are also frequently observed; the occasional appearance of muscarine symptoms in *Amanita muscaria* poisoning is a rare phenomenon.

Treatment is symptomatic and supportive. Diazepam, 5 to 10 mg in adults (children 0.1 to 0.2 mg/kg), repeated as required, can be given for sedation. In severe cases with evident psychotic behaviour, chlorpromazine or haloperidol may be useful.

Hallucinations

These are mainly visual and are the predominant feature in poisoning by psilocybin, a tryptamine derivative. Psilocybin and related substances are found in *Psilocybe* species and several other genera of (usually small) fungi, often collectively called 'magic mushrooms'; among these are some *Conocybe* spp., *Gymnopilus* spp., *Panaeolina foenisecii*, *Panaeolus* spp., *Pluteus* spp., and *Stropharia* spp. These fungi are eaten because of their psychoactive properties. The effects of the toxin are similar to those of LSD (stimulation of central and blockade of peripheral serotonin receptors). Other early symptoms are relaxation, euphoria, depersonalization, and altered sense of time and space. Less pleasurable, but equally typical symptoms, are vertigo, bizarre and terrifying hallucinations, anxiety, agitation, tachycardia, mydriasis, and flushing. Gastrointestinal symptoms, apart from nausea, are uncommon. Occasionally fever and seizures are observed in children. Sometimes frightening flashbacks occur after days or weeks.

Treatment is symptomatic, with reassurance and rest in a quiet environment. In cases of severe anxiety and agitation, diazepam may be given; in addition haloperidol or chlorpromazine is often necessary.

Alcohol-associated effects

When alcohol is taken at the same time as some fungi, notably an ink cap (*Coprinus atramentarius*) but also club foot (*Clitocybe clavipes*), or up to 5 days after they are eaten, a reaction similar to that induced by disulfiram (Antabuse) used for treating alcoholics (see Chapter 26.7.1) is produced. The skin becomes flushed, and sweating, mydriasis, nausea, anxiety, dyspnoea, and hypotension can occur as a result of the accumulation of acetaldehyde, because the toxin (named coprine in *C. atramentarius*) blocks the liver enzyme aldehyde dehydrogenase.

Treatment is symptomatic and supportive.

Symptoms with delayed onset

Severe gastroenteritis

Gastrointestinal symptoms that appear after a long latent period followed by liver damage, and sometimes also by kidney damage, are typical of poisoning by amatoxins, a group of cyclic octapeptides occurring in the death cap (*Amanita phalloides*) (Plate 21), which is by far the most common cause of death from the

ingestion of fungi. Amatoxins also occur in the destroying angel (*Amanita virosa*), the fool's mushroom (*Amanita verna*) and some *Galerina* and *Lepiota* spp. (in particular *Galerina marginata* and *Lepiota cristata*). These toxins inhibit transcription from DNA to mRNA by blocking RNA polymerase II (B). This results in deficient protein synthesis and subsequent cell death. The main target organs are the intestinal mucosa, liver, and kidneys.

Abdominal pain and vomiting, but in particular a severe, watery diarrhoea, start 10 to 24 h (more rarely at 6 h but normally around 12 h) after ingestion. Dehydration and exhaustion will ensue. Sometimes there is a transient recovery, after which the signs and symptoms of liver damage gradually appear from the second day. Hepatic failure may ensue, and this is the cause of death in fatal cases. Impaired kidney function is common initially because of dehydration; kidney dysfunction at a later stage is normally an indication of toxic renal damage and is a poor prognostic sign.

The mainstay of treatment is early admission to hospital, with rapid correction of fluid, electrolyte, and metabolic disturbances, adequate and preferably slightly increased diuresis for the first few days, and multiple oral doses of activated charcoal until 3 days after ingestion. Silibinin (also called silybin or silymarin), the main active constituent of milk thistle (*Silybum marianum*), reduces hepatic amatoxin uptake and, although its value in the clinical setting is controversial, it should be considered in patients who present after significant ingestion, as judged from their history and clinical findings. The dose of silibinin is 5 mg/kg intravenously over 1 h, followed by 20 mg/kg/24 h as a continuous infusion for 3 days after ingestion. Benzylpenicillin in large doses may be an alternative, if silibinin is not available. Haemodialysis or haemoperfusion is not indicated, unless the patient is admitted very early and before the onset of symptoms (very unusual) or in cases where there is pre-existing renal dysfunction. Liver transplantation may be the only therapeutic option.

Gastrointestinal and central nervous effects

The toxin gyromitrin, found in the false morel (*Gyromitra esculenta*), decomposes in the stomach to form hydrazines (in particular monomethylhydrazine). Hydrazines block pyridoxine synthesis, cause glutathione depletion in erythrocytes, and form free oxygen radicals that may bind to macromolecules in the liver. Gastrointestinal symptoms, associated with headache, vertigo, sweating, diplopia, nystagmus, ataxia, cramps, delirium, and sometimes coma, appear 6 to 24 h after ingestion or inhalation of cooking vapour. Severe poisoning may also cause hypoglycaemia, hepatic damage, and haemolysis; renal damage has also been reported.

Apart from symptomatic and supportive care, treatment includes administration of pyridoxine at 25 mg/kg over 30 min to prevent and control neurological symptoms; repeat doses may be required. It is advisable to administer a glucose infusion and to maintain adequate diuresis.

Kidney failure

Poisoning characterized by kidney dysfunction after a latent period of 2 to 4 days, in milder cases even up to 2 weeks, is typical of the group of brownish-orange *Cortinarius* spp. (*C. orellanus*, *C. speciosissimus*, and *C. splendens*) that contain orellanine, a highly nephrotoxic tetrahydroxylated di- *N*-oxidized bipyridine, and some less toxic derivatives such as orelline. During the latent period some patients may have diffuse gastrointestinal symptoms (often a day or more after the mushroom meal), but this is not the rule. This lack of early warning signs makes *Cortinarius* poisoning especially insidious, and patients are almost invariably admitted to hospital when the renal damage is already established and treatment procedures to prevent the toxic effects are no longer relevant. Typically, symptoms appear days after the meal, with fatigue, intense thirst, headache, chills, and abdominal, lumbar, and flank pain. After a transient polyuric phase, oliguria and anuria may follow. Renal function will recover in some cases, whereas in others kidney failure is permanent.

As these patients are normally admitted very late, management possibilities are restricted to monitoring of renal function and symptomatic care, including haemodialysis. In cases of persistent renal failure, kidney transplantation may be an alternative to chronic dialysis. If admitted during the first day, vigorous gut decontamination and haemoperfusion for poison elimination may be considered on theoretical grounds.

Allergic reactions

The roll-rim cap (*Paxillus involutus*) (Plate 22), previously eaten with impunity, may suddenly give rise to an immunohaemolytic anaemia in which decreased haemoglobin levels result in shock and even renal insufficiency. A respiratory allergy, called farmer's lung, results from inhaling spores of the mould *Faenia rectivergula* that is sometimes present on grain or hay. Spores of oyster mushrooms (*Pleurotus ostreatus*) or puffballs (*Lycoperdon* spp.) can also cause respiratory allergies. Skin allergies have been reported with the shiitake mushroom (*Lentinus elodes*).

Ergotism and mycotoxicoses

The ascomycete fungus *Claviceps purpurea* infects (Plate 23) cereal crops, in which its hard, purplish-black fruiting bodies or sclerotia (ergots) develop in the seed heads. Ergots contain several alkaloids that can cause disease when flour made from contaminated grain is eaten. There are two forms of ergotism ('St Anthony's fire'): vasoconstriction, sometimes leading to loss of extremities through gangrene (occurring after small quantities of contaminated food have been eaten over a long period) and a less common, acute form characterized by muscular tremors, convulsions, and hallucinations.

Other potent toxins (mycotoxins) are produced by some moulds (*Acremonium*, *Alternaria*, *Aspergillus*, *Fusarium*, *Penicillium*) that develop on growing crops, particularly if wet and harvested late, but occur mainly in cereal grains, rice, or nuts stored under damp, inadequately ventilated conditions. Intoxication may result from ingestion or inhalation. Aflatoxins (from *Aspergillus* spp.), ochratoxins (from *Aspergillus ochraceus* and *Penicillium verrucosum*), and trichothecenes and zearalenone (from *Fusarium* spp.) have been detected in foods and are suspected of causing immunosuppression, myelosuppression, hepatotoxicity, nephrotoxicity, neurotoxicity, dermatotoxicity, oestrogenicity, mutagenicity, teratogenicity, and carcinogenicity (liver, colon, alveolar cells). Ochratoxin A, which inhibits mitochondrial oxidative phosphorylation in the proximal tubule of the nephron, has been implicated as a possible cofactor in Balkan nephropathy. Specific mycotoxins can be detected at 0.1 ng/ml by radioimmunoassay and enzyme immunoassay.

Sources of information

A directory of poison information centres worldwide is kept by the International Programme on Chemical Safety (IPCS), WHO, Geneva. A computerized, image-based, identification system for poisonous plants and fungi has been developed by the Royal Botanic Gardens, Kew, in collaboration with the Medical Toxicology Unit of Guy's and St Thomas' Hospital Trust, London. The system, called *Poisonous Plants and Fungi in Britain and Ireland* is available on CD-ROM from the Royal Botanic Gardens, Kew. Similar systems are being developed for other countries.

Further reading

Bresinsky A, Besl H (1990). *A colour atlas of poisonous fungi. A handbook for pharmacists, doctors, and biologists*, transl. N.G. Bisset. Wolfe, London.

Cooper MR, Johnson AW (1998). *Poisonous plants and fungi in Britain. Animal and human poisoning*, 2nd edn. The Stationery Office, London.

Eddleston M *et al*. (2000). Anti-digoxin Fab fragments in cardiotoxicity induced by ingestion of yellow oleander: a randomised controlled trial. *The Lancet* **355**, 967–72.

Everist SL (1981). *Poisonous plants of Australia*, 2nd edn. Angus and Robertson, Sydney.

Frohne D, Pfänder HJ (1997). *Giftpflanzen*: *ein Handbuch für Apotheker, Ärzte, Toxikologen und Biologen*, 4th edn. Wissenschaftliche Verlagsgesellschaft, Stuttgart. (An English translation of the first edition, by N.G. Bisset was published as *A colour atlas of poisonous plants* by Wolfe, London, 1984.)

Lovell CR (1993). *Plants and the skin.* Blackwell Scientific, Oxford.

Malloy CD, Marr JS (1997). Mycotoxins and public health. *Journal of Public Health Management Practice* (3) 61–9.

Meda HA *et al*. (1999) Epidemic of fatal encephalopathy in preschool children in Burkina Faso and consumption of unripe ackee (*Blighia sapida*) fruit. *The Lancet* **353**, 536–40.

Parish RC, Doering PL (1986). Treatment of *Amanita* mushroom poisoning. A review. *Veterinary and Human Toxicology* **28**, 318–22.

Turner NJ, Szczawinski AF (1991). *Common poisonous plants and mushrooms of North America*. Timber Press, Portland, OR. (Reprinted 1995.)

Watt JM, Breyer-Brandwijk MG (1962). *The medicinal and poisonous plants of southern and eastern Africa*, 2nd edn. Livingstone, Edinburgh.

8.4.1 Occupational and environmental health and safety

J. M. Harrington with contributions from K. Gardiner, I. S. Foulds, T. C. Aw, E. L. Baker, and A. Spurgeon

General introduction

J. M. Harrington

Definition and scope

Occupational health is that area of public health concerned with managing the health of working people. Occupational health professionals include physicians, nurses, hygienists (scientists with experience in monitoring and more importantly controlling the exposure of working people to chemical, physical, and biological agents in their place of work), toxicologists, and—increasingly these days—psychologists to the assess the psychosocial aspects of work. Safety engineers are responsible for prevention and investigation of accidents at work (see under Occupational safety).

Occupational health issues can be promoted by these professionals, but for long-term success it is crucial that both managers and the workforce consider it an integral part of their working practices and philosophy. It is in everyone's interest that the workforce is 'happy, healthy, and here'!

History of occupational disease

Stone Age flint knappers were probably exposed to airborne silica dust during their work. However, life expectancy then was probably shorter than the pathogenesis of silicosis. Some industries like mining have always been hazardous. The ancient Egyptians recognized this by restricting such work to slaves and criminals. By the Middle Ages, the plight of the free miner had been recognized by Georgius Agricola (1494–1555) and Paracelsus (1493–1541). Agricola not only described the 'galloping consumption' of Carpathian silver miners but also proposed ways of reducing the dust in mines by improved ventilation.

The first authoritative treatise on occupational disease was written by Ramazzini (1633–1764). His book *De Morbis Artificium* is unsurpassed in its classic descriptions of many occupational diseases ranging from mercurialism in mirror workers to repetitive strain injury in clerical workers. The Industrial Revolution in Britain brought occupational diseases to the attention of Parliament, largely through the work of philanthropists like Robert Owen, Robert Peel, and Lord Shaftesbury. Early legislation to control the worst vicissitudes of factory labour was emasculated by Parliament but the process had begun. The First Act of 1802 (which, interestingly, introduced the concept of limiting the hours of work) was followed by others leading to the 1833 Act which saw the start of His/Her Majesty's Factory Inspectorate (HMFI)—an enforcing authority.

By the early twentieth century the toxic effects of arsenic, mercury, phosphorus, and lead were so common that notification of these diseases became required by law and compensation for ill health was granted. Clearly, working conditions in the Western world have improved greatly since then but the recent revelations about factory life in Eastern Europe as well as the working conditions for many in Third World countries demonstrate an important tenet of occupational health practice: that is, while occupational disease may be preventable, the continued—often necessary—use of hazardous materials and processes ensures that many such diseases cannot be eliminated, only controlled.

Occupational health services

The notion that employers should provide health care for workers is hardly new. During the fourteenth century, the Pope — not an employer in this connection! — decreed that prostitutes should be examined regularly for evidence of sexually transmitted disease. Whether the results were significant epidemiologically is not recorded. The first recognizable occupational health service in England began in the mid eighteenth century when the London (Quaker) Lead Company recognized the adverse effect of mining on workers and provided health and welfare services in northwest England. Since then, occupational health provision has expanded along different lines in different countries.

The International Labour Organization (Convention 161, 1985) urged members 'to develop progressively occupational health services for all workers.... The provision made should be adequate and appropriate to the specific needs of the undertaking.' Services are not universally available and interpretation of the requirements varies greatly between countries and employment sectors. Initially, most services arose from a mixture of philanthropy and self-interest; the theory being that the healthy worker was likely to be more productive. Present-day services range from total health care including primary care and hospital medicine (as in the states of Eastern Europe), to industry-specific systems concerned almost solely with the adverse health effects of a single industrial environment (as with the extractive, offshore, and chemical industries). Services in the United States and much of Europe may include general health promotion and education, but much inequity in health care exists between enterprises within the same country.

Recent increase in the provision of occupational health services has followed the enactment of effective health and safety legislation. The Health and Safety at Work etc. Act (1974) in the United Kingdom is an example. Some countries such as the Nordic countries, The Netherlands, and Australia, require the provision of occupational health services by law. Statutory provision of such services in the United Kingdom and the United States is limited to particular industrial sectors and specific occupational exposures such as ionizing radiation, heavy metals, fibrogenic materials, and carcinogens. In general, major Health and Safety Acts will 'enable' a variety of government departments to create legislation in the form of 'regulations' requiring action from employers in particular occupational and environmental circumstances. The increasing attention paid in the United Kingdom to public health issues has enhanced the chances of improving occupational health services, although, counterintuitively, many of the larger industrial enterprises are divesting themselves of an 'in-house' service.

The European Community, through 'directives', increasingly drives the occupational health and safety agenda in member states requiring them to modify or create

legislation in response.

The latest moves from Brussels have been to encourage the delivery of occupational health to all by 'competent' persons, although the mode of delivery—even the definition of competent—has been left to member states for interpretation.

An ideal 'menu' of such services depends on many variables including the industrial sector, existing legislation, management/employee collaboration, the availability of employment, and compensation for damage. Most importantly, the level of service provided should be based on a thorough risk assessment of the work processes in that organization and a clear and logical procedure of risk management.

To deliver even such a basic service will require multidisciplinary teams including trained physicians, hygienists, and nurses. Few companies have such services and many are too small even to contemplate such provision.

Developments will be tempered primarily by the economic climate, perceptions of what constitutes occupationally mediated disease, and political will. However, an exponential rise in legal action, insurance costs, and compensation will play a significant part in persuading management that competent occupational health services are an absolute requirement of profitable organizations. If this is to encompass the small and medium-sized enterprises then provision must come either from larger employers (such as the National Health Service) or from private providers. It is important to remember that in the United Kingdom, for example, companies employing fewer than 250 people account for 99 per cent of all businesses and 40 per cent of the working population.

Prevention

The prevention of occupational disease depends upon recognition of the condition as occupational, assessment of the level and duration of exposure and hence its possible effects, control of the problem at source, audit of the risk management procedures, and perhaps health surveillance of those exposed using suitable techniques. These procedures are dealt with later in the chapter.

Recognition of diseases as occupational may vary from country to country. The EC has proposed a harmonization of national lists of occupational disease for the purpose of compensation of workers.

Compensation for occupational diseases

In the early years of Western industrialized society, the chances of a worker winning compensation for an occupationally related disease or injury were slim, resting as they did on a successful common law suit against the employer for negligence. Such cases are still notoriously difficult to win. However, by the end of the nineteenth century, many countries in Europe as well as the United States had passed workman's compensation laws of one sort or another. Such schemes were usually restricted to specified diseases or occupations. For example, the 1897 Act in Britain was for accidents only, with six diseases added in the 1906 Act. Today the Industrial Injuries Scheme extends to 67 diseases.

The principles underlying such schemes are that they should be 'no fault', that the disease should be, with reasonable certainty, caused by work, and that the benefit claimed should offset job loss, wage earning deficit, disability, or provide death benefit to the next of kin. While the scheme in Britain has suffered considerable erosion over the last 10 years, similar schemes exist in one form or another in all EC member states with the exception of The Netherlands where compensation for ill health is of a more general nature. Advice on proposed additions to the list of compensatable diseases is made in the member states by government appointed advisory groups. In Britain, this group is the Industrial Injuries Advisory Council which reports to the Secretary of State for Work and Pensions.

New EC recommendations propose that national schemes be made uniform. In addition, the EC wishes to see the introduction of a concept to aid those who can prove they have an occupationally related disease which is not on the standard list—the so-called 'individual proof' system. In Britain such a dual system of specified agents and individual claim opportunities exists only for asthma. No specific agents are listed for dermatitis.

It is important that the clinician is aware of such schemes. If the disease and work exposure seem related, and are listed, the patient should be advised to claim for compensation.

Further reading

Adams RM (1999). *Occupational skin disease*, 3rd edn. WB Saunders, Philadelphia.

Cox RAF, ed (2000). *Fitness to work; the medical aspects*, 3rd edn. Oxford Medical Publications, Oxford.

Harrington JM *et al.* (1998). *Occupational health*, 4th edn. Blackwell Scientific, Oxford.

International Labour Office (1998). *Encylopaedia of occupational health and safety*, 4th edn. International Labour Office, Geneva.

Levy BS, Wegman DH (2000). *Occupational health*, 4th edn. Little, Brown, Boston.

Parkes WR, ed (1994). *Occupational lung disorders*, 3rd edn. Butterworth, London. (4th edn in preparation.)

Baxter P *et al.*, eds (2000). *Hunters' diseases of occupation*, 9th edn. Hodder and Stoughton, London

Health and Safety Executive (1998). *Help on work-related stress. A short guide*, INDG 281. Health and Safety Executive, London.

Assessment of the workforce and their environment

K. Gardiner and J. M. Harrington

Introduction

Occupational health comprises recognition, evaluation, and control. Control (including prevention) is the paramount element; there is no point in knowing what may happen or diagnosing what has happened if no remedial action is taken. For too long, society/industry has relied on occupational physicians to identify the failures in control by means of diagnosing occupationally related disease without giving sufficient weight to the need for control and therefore prevention.

As a result of this realization, current health and safety legislation is driven by risk assessment, i.e. those who generate the risk(s) are responsible for undertaking an assessment, the detail of which must be commensurate with the complexity of the situation and the ultimate risk. It follows from this that the quantified risk must be managed, and hence employers must now add risk management to their portfolio of activities.

Before embarking on any evaluation, especially where there is suspicion of occupational aetiology, great care needs to be exercised beforehand to determine what the legislation requires, what the known toxicological/health effects are, previous evidence from similar circumstances/environments, etc. This is the 'recognition' aspect, and highlights the necessity not only to be well informed but also competent (i.e. to be able to interpret and act upon the information appropriately).

Evaluation

One aspect of occupational medicine, distinct from most others, is that commonly one is dealing with groups rather than individual patients. In the main, this acts to complicate the evaluation as interindividual variance is added to the intrinsic intraindividual variance (for example, two people exposed to exactly the same quantity of toluene over the same time period will not excrete the same amount of urinary hippuric acid or excrete it at the same rate). In order to minimize any erroneous decisions and to maximize the usefulness of any information, evaluation of the workforce (and workplace) must be conducted in a systematic manner. This is often achieved by posing a number of simple questions.

The most fundamental of these questions is why? It is essential that the rationale for evaluating an individual or group is clear and justifiable. However, as occupational health professionals exist to eliminate ill health at work their priority must always be to eliminate problems in the workplace immediately rather than to delay while quantifying them. The law is prescriptive for only for a limited number of hazards (ionizing radiation, lead, asbestos).

Where large groups are involved, it may be necessary to evaluate a subset—this is the 'who' question. When there is no other over-riding need for selection (such as those working for extended periods (more than 8 h per day), those working at an elevated metabolic/breathing rate, those already unwell, or those likely to undertake unusual or unscheduled tasks (maintenance)), individuals

should be chosen randomly.

The timing of measurements (i.e. 'when') is often critical. Should health outcome measures be taken before, during, or after the putative exposure or is there some legislative requirement that dictates the timing? For example, urinary mandelic acid (for styrene exposure) is best collected at the end of a working shift whereas urinary trichloroacetic acid (for trichloroethylene exposure) is best measured at the end of the working week, this being determined, in part, by the differing half-lives of the substances concerned.

Usually, the issue of what to assess is self-evident, but, as with all other aspects of this systematic approach, answers to the questions are interlinked, certainly with the 'why' question. For example, is biological monitoring or biological effect monitoring required? The former is the detection of a chemical or its metabolite in a biological sample as a measure of exposure. The latter is measurement of a change in some biochemical or physiological variable to indicate the effect of the contaminant on the body.

Choice of an appropriate technique for making these measurements should be based where possible on sensitivity and specificity.

The technique should be both suitable and sufficient for its purpose. For example, in the diagnosis of occupational asthma suitable techniques include peak flow meters, simple spirometry (time–volume curves), and whole-body plethysmography, but the simplest of these may be adequate.

Remedial action

The law is clear: the employer must adapt the workplace to be a safe environment for the employees. The employees should not have to adapt themselves to the stresses and strains of their work. Despite this philosophy, it is common for the chain of control to start with ineffective pre-employment medicals whose aim is only to avoid exposing people to contaminants likely to exacerbate allergies or other diseases.

Control is often viewed as a physical alteration of the workplace (engineering) or alteration of the behaviour of the workforce (administrative). It is preferable to have designed the workplace appropriately in the first place. Unfortunately, most control is required remedially, after an unacceptable level of risk has been identified as a result of a formal risk assessment. Remedial action may involve:

- elimination—removal of the process;
- contaminant substitution—replacemnt of the contaminant with one carrying a lower risk;
- process modification (temperature, agitation, enclosure, etc.);
- substance modification (wavelength, form, size, etc.);
- isolation/segregation (time, distance, shielding);
- local extract ventilation;
- minimization of the duration and frequency of exposure;
- education and training and the use of personal protective equipment.

Occupational dermatology

I. S. Foulds

An occupational dermatosis is a pathological condition of the skin for which occupational exposure can be shown to be a major contributory factor. Legal definitions vary between countries. In the United Kingdom the majority are defined by Prescribed Disease D5 of the Department of Health and Social Security as 'non-infective dermatitis of external origin (including chrome ulceration of the skin but excluding dermatitis due to ionizing particles or electromagnetic radiation other than radiant heat)'. However, this definition may exclude newly described dermatoses, and in the case of occupationally acquired hypomelanosis a new category (Prescribed Disease C25) was necessary.

Incidence

The true incidence of occupational dermatoses is difficult to obtain, as in some countries occupational accidents and illnesses are not differentiated, and others fail to separate dermatitis from other skin diseases. Few statistics are derived from short absences from work or disease without disability, and most information is based on compensation paid.

Dermatitis, mostly due to irritant contact factors, accounts for 95 per cent of all occupational dermatoses. In the United Kingdom a million working days a year may be lost from such conditions. The diagnosis and management is discussed elsewhere (see Chapter 23.1).

Some managers and even some doctors believe that people with dermatitis are out to obtain whatever they can in the way of compensation. However, in times of high unemployment, many people are reluctant to seek help in case they are labelled as suffering from industrial dermatitis and are then transferred to less skilled work with the loss of bonuses or even their jobs. Once labelled in this way, they are unlikely to find alternative employment.

Many employers believe that they do not have a problem with dermatitis, but careful inspection of employees' skin shows that the condition may be present in up to a third of the workforce. Considerable time is lost due to dermatitis and suffering due to discomfort, depression, and social ostracism is impossible to quantify.

Employees' and employers' attitudes to dermatitis

As with many skin diseases there are many myths attached to dermatitis. The 'leper approach' still exists. Many people believe that dermatitis is infectious and can be passed on by towels and touching; this results in the affected individual becoming socially isolated. Under no circumstances should dermatitis ever be regarded as infectious. It can only be acquired by wear and tear occurring to the skin or by the development of an allergy to a substance that has been in contact with the skin. Affected individuals therefore should not have unnecessary restrictions placed on them.

When someone develops severe occupational dermatitis this is often apparent to their colleagues who, after discussion, may think that their own skin problems are also occupationally related. However, since skin diseases are common, individual problems may be falsely attributed to a work-related cause. With outbreaks of apparent dermatitis, great care is needed in handling the situation to identify the causes and differentiate those with occupationally related problems from those with other skin diseases. It may be necessary to ask for the help of a dermatologist with specialist knowledge if the confidence of the workforce is to be maintained.

Barrier creams

No cream provides a barrier to penetration of substances into the skin. In fact, in some situations they may actually enhance penetration. Although there are numerous formulations available, they can be divided into those suitable for dry or for wet work.

The main benefit they offer is from their bases such as lanolin which may help to improve the hydration (suppleness) of the skin, with the result that, when cleansers are used, less degreasing of the skin occurs. In theory, this may help to reduce irritant contact dermatitis from repeated hand washing. There is no evidence that barrier creams protect against sensitizers, and occasionally sensitization may occur to some of the constituents of the cream. The use of a barrier cream may give an employee a false sense of security and lead to increased skin abuse.

Skin cleaning

When substances remain on the skin after the working day, the risk of irritation or sensitization is increased. The most efficient skin cleaners, however, are often the most irritant of substances due to their solvent or detergent content. If cleaners are too mild for the task, workers will often use degreasing agents in the manufacturing process, for example solvents or paraffins, to obtain adequate cleaning. Although these substances will 'clean' they are potentially very irritant with repeated use. It is often not appropriate to provide one type for different jobs. Agents should be chosen that clean adequately in a short period of time without having too strong a degreasing effect.

After-work creams

There are many such creams, essentially moisturizers, which have the benefit of increasing the hydration of the skin following cleaning at the end of the day. They are of particular benefit in occupations where excessive drying of the skin may occur. Their use should be encouraged where hot air driers are used as these tend to dry the skin unduly.

Skin protection

In industrial situations the hands and forearms are most at risk. The use of proper gloves (along with gauntlets or arm bands to prevent powders entering under the cuff) coupled with a high standard of hygiene, can minimize contact and provide adequate protection. Where moving machinery is present, wearing gloves may pose a potential danger. Even when gloves are used, they may be taken off to undertake tasks requiring manual dexterity, resulting in contaminated hands being placed back inside the gloves.

A wide range of materials have been used to manufacture gloves including cotton, leather, nylon, glass fibre, acrylonitrile, rubber, neoprene, butyl rubber, viton, polyurethane, PVC, PVA, and Teflon. These confer individual benefits in specific occupational applications.

Non-dermatitic occupational dermatoses

These form a minority of occupational dermatoses and awareness of their existence is important for those involved in managing skin patients. The more important are briefly mentioned below:

1. Psoriasis (see Chapter 23.1) may erupt at sites of injury or as a response to friction in manual workers. When this occurs on the hands, it may be confused with dermatitis although vesicles (blisters) are not found in psoriatic skin.
2. Infections include anthrax (albeit rare in the United Kingdom), human papilloma virus warts, orf (a poxvirus), and fungal infections such as tinea pedis and cattle ringworm (*Trichophyton mentagrophytes*).
3. Chronic paronychia due to candida occurs in those exposed to wet work.
4. Acne may be caused by a variety of chemicals including oils, coal tar chlorphenols, and petroleum products. Chloracne is a particularly refractory form of acne caused by halogenated aromatic chemicals, which may also cause systemic toxicity. Mild cases may be difficult to differentiate from conventional acne but multiple comedones (blackheads) are found over the malar regions. Industrial accidents can result in exposure and subsequent symptoms: for example in 1976 an explosion in a chemical plant near Seveso, Italy resulted in the exposure of the local population to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin which caused severe systemic and cutaneous symptoms.

5. Vitiligo may be occupationally acquired. Several substituted phenols including *p*-tertiary-butyl phenol and monobenzyl ether of hydroquinone may cause hypomelanosis indistinguishable from vitiligo.
6. Scleroderma may be caused by exposure to vinyl chloride, and has been reported with trichlorethylene and organic solvents. Scleroderma-like lesions have also been reported in people exposed to epoxy resin fumes, and silicosis associated with scleroderma has been reported in miners.
7. A variety of chemicals may cause alteration of skin pigment; for example mercury and silver, which causes argyria.
8. Occupationally induced itching may occur in atopic individuals who exhibit dermographism due to histamine release from mast cells caused by fibres typically around 3 µm in size which can penetrate the skin. This occurs with exposure to glass fibres, ceramic fibres, and fibreglass.

Occupational cancer (see also Chapter 6.1)

J. M. Harrington

Background

Georgius Agricola's account in 1555 of the illnesses of Carpathian silver miners includes evidence of a rapidly progressive and fatal lung disorder. The fact that these mines are now known to contain uranium ore suggests that exposure to radon gas may well have been high enough to cause lung cancer in the miners. Nevertheless, it is Percival Pott's description in 1775 of an excess risk of scrotal cancer in postpubertal chimney sweeps that first raised the possibility of chemicals—particularly polynuclear aromatic hydrocarbons (PAH)—causing cancer. Confirmatory evidence from animal experiments did not arrive until 1915 and the first carcinogenic hydrocarbon was identified by Kennaway in 1924 as 1,2,5,6-dibenzanthracene.

While the polynuclear aromatic hydrocarbons were generating interest as skin carcinogens, clinical observations of dyestuff workers were suggesting a link between bladder cancer and aromatic amines. In 1895, Rehn described three cases in a group of 45 workers in Germany involved in the preparation of fuchsin. Further reports followed from other countries and the classic studies of Case and his colleagues in the 1950s showed that 2-naphthylamine and benzidine were human carcinogens in manufacturing and user industries. 2-napthylamine was a contaminant of the antioxidant used in tyre manufacture. Rehn discovered that an organ distant from the point of first contact could bear the main force of the carcinogenic effect if its exposure was most prolonged and most intense.

In the same year that Rehn made his discovery, Röntgen discovered X-rays and 3 years later, the Curies isolated radium. Unfortunately, knowledge of the carcinogenic properties of ionizing radiation came from the skin and bone marrow cancers suffered by these early pioneers with confirmatory animal data following soon after. The bone sarcomas noted in laboratory animals were followed by human evidence in the 1930s among the painters of luminous dials who used radium-235 and mesothorium. The inventor of the luminous paint, Dr von Sochocky, died of aplastic anaemia in 1928. Again in the 1930s, case reports were appearing of lung cancer (an unusual tumour in those days) in workers exposed to asbestos fibre. In asbestosis cases the incidence was reported as 18 per cent and reports of pleural mesothelioma followed a decade or so later.

Thus, within a century of Rehn's discovery, chemical carcinogenesis had become a well-recognized phenomenon with much of the evidence coming from occupational studies.

Diagnosis

Clinical acumen remains of paramount importance. It is clinicians who have played the major role in discovering new causes of cancer with confirmatory evidence coming from laboratory studies and epidemiological investigations (see Fig. 1).

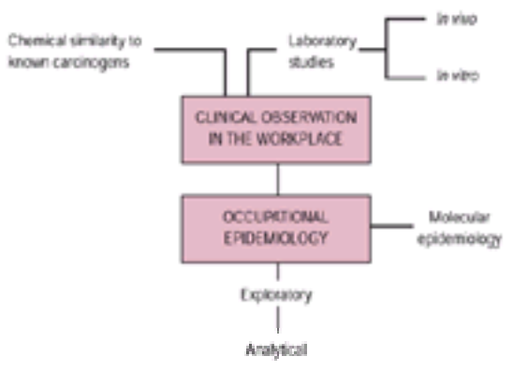

Fig. 1 The scientific basis for establishing occupational causality.

Such information is collated and interpreted by various national and international agencies. The most reliable source is the monograph series of the International Agency for Research on Cancer (**IARC**) based in Lyon. To date they have published over 70 such monographs with several updating supplements. When a patient is diagnosed as having cancer, it is important that the clinician should review the occupational history to consider an occupational cause. If this seems probable, enquiries should be made to see if state compensation is available for that cancer and that workplace exposure.

Attribution of cancer to occupational causes

The most widely accepted estimates of the proportion of all cancers attributable to occupational exposures is 4 per cent with a range of 2 to 8 per cent for a developed country like the United States. For the 20 per cent or so of the population in which occupationally related cancers are almost exclusively concentrated (manual workers aged 20 or over, in mining, agriculture, and industry broadly defined), perhaps as much as one lung or bladder cancer in every five may be attributed to workplace exposure.

In addition, it is necessary to consider other exposures which may interact with workplace exposures. These are particularly relevant when considering the relative effectiveness of removing or reducing exposure to one or more jointly acting agents. Few good studies have been completed on interaction, but there is good evidence for the multiplicative effects of cigarette smoking and asbestos exposure in the genesis of lung cancer. Besides asbestos, interactions have been demonstrated to be at least additive for tobacco consumption and exposure to arsenic and nickel compounds as well as ionizing radiation.

Table 1 and Table 2 show some important examples of occupational exposures causing cancer.

Polynuclear aromatic hydrocarbons

Polynuclear aromatic hydrocarbons are a large and complex group of compounds mainly generated during the incomplete combustion of carbonaceous products, of which coal and oil result in the most important occupational exposures. Cigarette smoke contains a number of these compounds and so it is often difficult to distinguish lifestyle from occupational factors. The site of action of these compounds is mainly the lung, skin, and bladder. The industries most prominently linked to such exposures are coke ovens, gas production, steel industries, aluminium refineries, iron and steel foundries as well as workers exposed to exhaust fumes from soot, pitch, tar, and petroleum products.

Aromatic amines

Aromatic amines are a group of chemically similar compounds which have particular importance as dyestuffs or antioxidants or as intermediates in dye production. Some are known human bladder carcinogens, a larger number are known animal carcinogens but for which human data are limited or lacking. The more potent carcinogens are now banned.

Metals and metalloids

The most important carcinogenic metals are compounds of arsenic, chromium, and nickel. Arsenic and its compounds cause lung and skin cancer, and these risks occur in the extraction of metalliferous ores (which are frequently contaminated by arsenic compounds) and in the now limited use of arsenic in pesticides and other industrial usages. Hexavalent chromium compounds used in the pigment and plating industries have been shown to cause lung cancer. Lung and nasal cancer are associated with the refining of nickel and the most likely causative agents are the oxidic and sulphidic nickel compounds. Other metals for which there is evidence of carcinogenicity include beryllium and cadmium.

Other organic compounds

Benzene is widely used in industry both as the base compound and as an important building block in the organic chemical industry. It causes leukaemia and aplastic anaemia. Vinyl chloride monomer, which is the starting point for synthesis of polyvinyl chloride, causes angiosarcoma of the liver. The most potent lung carcinogens are apparently the chloromethyl ethers which are used in ion exchange resins. Other suspect organics include acrylonitrile, acrylamide, butadiene, diethyl sulphate, epichlorhydrin, ethylene dibromide, formaldehyde, styrene oxide, tetrachloroethylene, and trichlorothylene.

Industrial processes

Some processes listed in Table 1 are linked to specific exposures, such as polynuclear aromatic hydrocarbons in aluminium production and radon in underground mining. For others such as boot and shoe manufacture, furniture making, and painting, the specific relevant exposures have not been identified.

Musculoskeletal disorders

T. C. Aw

Musculoskeletal disorders rank alongside stress as a major cause of workplace ill health. As musculoskeletal disorders are common in the general population, their suspected relationship to specific occupations or activities can often be difficult to confirm. A combination of occupational, psychological, personal, social, and home factors may be involved. The epidemic of repetitive strain injury in

Australia during the 1980s highlighted the complex interaction of symptoms, group perceptions, and employment.

Repetitive strain injury

This term was coined to describe pain and discomfort in the wrist and forearm associated with the performance of repeated tasks, mainly at work. Other terms used synonymously are cumulative trauma disorder, occupational overuse syndrome, and work-related upper limb disorder, with some terms restricting the definition to conditions in upper limbs caused by work. Unfortunately, there is no widely accepted case definition, and several disorders such as carpal tunnel syndrome, de Quervain's disease, epicondylitis, and tenosynovitis are often included in the entity. Clusters of cases have been recognized in occupational groups such as typists, telephonists, computer keyboard operators, musicians, cleaners, hairdressers, butchers, and assembly line workers. While the wrist and forearm are the anatomical regions most frequently affected, the symptoms are often very diffuse and the shoulder and neck may also be involved. Risk factors common to many cases are:

1. repeated movements at certain joints, often through the full range of movements, for example flexion and extension at the wrists;
2. constrained postures, sometimes at the extremes of the range of movements, and often against pressure or force.

Low back pain

This is possibly the commonest musculoskeletal condition experienced by people at work. Poor lifting and manual handling techniques and sitting for prolonged periods in the course of work activities (for example professional drivers) are contributory factors. Nurses, porters, and brick-layers are groups with a high prevalence of low back pain. The total cost of sickness absence, early retirement, and treatment for low back pain in many countries is considerable.

Occupations associated with musculoskeletal disorders

Several occupations are associated with a high risk of sustaining an occupationally related musculoskeletal injury. These include fractures and joint damage in construction workers, miners, and deep sea fishermen, avascular bone necrosis from decompression sickness in tunnellers and divers, acro-osteolysis in workers exposed to vinyl chloride monomer, and septic bone and joint lesions from brucellosis in meat process workers. The joints of coal-face workers, farmers, and professional dancers are subjected to frequent heavy impact loading leading to a high incidence of degenerative changes of the hip and knee.

Clinical investigations and management

Several investigations such as erythrocyte sedimentation rate, C-reactive protein, rheumatoid factor, and imaging may help to exclude an underlying medical condition.

The management of a case of occupational musculoskeletal disorder requires alleviation of symptoms and the consideration of modifications to the system of work or the layout of the workstation. Non-steroidal anti-inflammatory medication and physiotherapy can reduce pain, discomfort, and limitation of function. Localized areas of inflammation will respond to local injections of corticosteroids. Physical therapy procedures and occasionally surgery may be considered. There should be an evaluation of ergonomic factors in the design of equipment at the individual's workplace. This requires attention to the adequacy of the workstation, ease of access of the worker to tools, components, and other equipment, and the suitability of the general work environment—its lighting, temperature, humidity, and noise. Training in proper methods of lifting and manual handling is essential.

Musculoskeletal disorders affecting occupation

Patients with known arthritic disease, such as rheumatoid arthritis and systemic lupus erythematosis, need to be assessed for the effect that the condition may have on the performance of their work duties. The fluctuating nature of most forms of chronic arthritis makes precise predictions difficult. Physical disability can improve despite persistence of the disease. This is due both to the beneficial effects of treatment and to the patient's adaptation to the consequences of the disease.

Cardiovascular system

J. M. Harrington

Cardiovascular disease is the major cause of mortality and morbidity in industrialized countries (see Section 15). The association between personal risk factors and cardiovascular disease is well known, but less attention has been paid to occupational and environmental influences.

There is good evidence from the classical studies of London bus drivers and conductors that sedentary workers have a higher risk of ischaemic heart disease than those who are more active. There is some evidence linking job stress and heart disease. The Whitehall II studies suggest that control over one's job is an important factor in determining subsequent risk of myocardial infarction.

Exposure to chemicals such as carbon disulphide, chlorinated organic solvents, nitroglyecrine, and vinylchloride monomer may contribute to cardiovascular disease. The cardiovascular effects of exposure to heavy metals are largely secondary to their nephrotoxic effects. Toxic industrial gases produce their effect secondary to anoxia. Among physical agents, vibration is known to cause peripheral vascular disease and acute high exposure to noise is known to raise blood pressure. Workers on rotating shifts have an increased risk of ischaemic heart disease.

Genitourinary system

J. M. Harrington

The kidney plays a crucial role in the excretory and detoxification mechanisms of the body. Perhaps the most effective detoxification manoeuvre of the liver is to increase the polarity or acidity of the absorbed substance. This, in turn, increases the water solubility of the chemical and hence its renal excretion. The kidney, therefore, bears the brunt of many exposures to toxic chemicals. Some toxic substances do reach the kidney unchanged but most are metabolized to some extent or other. Some, such as cadmium, become sequestered in the renal cortex while others, such as the aromatic amines, are present in the bladder long enough and at a high enough concentration to induce malignant change in the transitional cell epithelium.

Sudden, severe exposures to some chemicals can cause acute nephropathy. Such compounds may damage the kidney directly due to their instrinsic nephrotoxicity or may induce secondary damage due to prerenal effects, such as the haemolysis following arsine exposure. Hypovolaemic shock can follow acute fluid loss or extreme heat, while post-traumatic renal failure can follow crush injuries or high-voltage electric shock, both of which cause muscle necrosis. Chronic lower-dose exposure leading to nephropathy is more commonly associated with metals or organic solvents.

The metals most commonly implicated in renal disease are mercury, cadmium, lead, and, perhaps, uranium (see Chapter 8.1). Mercury exposure resulting in acute tubular necrosis or the nephrotic syndrome is most unusual these days. Under present-day workplace exposures, the effects of inhaled mercury vapour or absorption of mercury salts are more likely to cause mild proteinuria and limited tubular dysfunction. Biological monitoring of mercury-exposed workers has had to become more sophisticated. Sensitive tests of urinary enzymes, such as *N*-acetyl-b-D-glucosamidase, are necessary to detect those subtle effects. Similarly, modern industrial exposures to cadmium rarely result in the proximal or distal tubular dysfunction or renal cortical damage that was more prevalent in the past. However, cadmium is only slowly leached from the renal cortex and remains a potentially serious long-term cumulative poison. The environmental cadmium contamination which caused widespread tubular dysfunction with hypercalciuria and osteomalacia in multiparous postmenopausal Japanese women (Itai-Itai disease) has not been described in the West. Lead nephropathy is also a rarity nowadays, but was not uncommon in the early part of this century. Lead is capable of causing damage to all parts of the nephron. More subtle tests of renal enzymes are now needed to assess the effects of lead exposure on the kidney. Soluble uranium compounds such as uranium hexafluoride have been shown to be potent nephrotoxins after acute accidental exposure but this problem is virtually unknown in a well-controlled modern facility.

Chlorinated aliphatic solvents such as carbon tetrachloride and chloroform can cause the hepatorenal syndrome. The renal damage is largely an effect on the proximal tubular epithelium which can lead to tubular necrosis and acute oliguric renal failure. The weight of evidence from case–control studies of workers exposed to solvent suggests an excess risk of glomerulonephritis. The mechanism is unclear, but a link with Goodpasture's syndrome suggests possible autoimmune damage to the glomerular basement membrane.

Although the prostate possesses the curious ability to concentrate (and excrete) heavy metals, little evidence exists of occupationally related prostatic disease. Earlier reports of a link between cadmium exposure and prostatic cancer have not been confirmed. Cancers of the urinary tract associated with occupational exposure to aromatic amines and polynuclear aromatic hydrocarbons were described earlier.

Gastrointestinal tract

T. C. Aw

The gastrointestinal tract acts as a semipermeable membrane through which ingested pollutants are absorbed. There are defence mechanisms that limit the damage to the gastrointestinal tract from such pollutants, and minimize their absorption. The mucous lining of the gut and diarrhoea and vomiting form part of these defence mechanisms. Hence, systemic absorption via the gastrointestinal tract plays a relatively minor role in causing occupational disease.

Acute gastroenteritis may follow the ingestion of chemicals such as soluble salts of heavy metals.

The liver is frequently at risk from occupational exposures, as it is the target organ for detoxification and metabolism of absorbed compounds. A wide variety of infectious and chemical agents cause different types of hepatocellular injury which may eventually lead to cirrhosis and liver failure (Table 3).

The haemopoietic system

T. C. Aw

The metal most frequently associated with bone marrow damage is inorganic lead. Lead causes anaemia by inhibiting the enzymes involved in haem synthesis and also by haemolysis. Determination of blood lead levels is used in the monitoring of lead-exposed workers. A diagnosis of lead poisoning is supported by symptoms of malaise, colic, and constipation, signs of anaemia and peripheral motor neuropathy (rare, usually only in severe cases), and laboratory evidence of basophilic stippling of erythrocytes, elevated blood lead, low haemoglobin, raised free erythrocyte protoporphyrin, and raised urinary †-aminolaevulinic acid. Indications of excessive lead absorption should lead to removal of the affected worker from further occupational exposure, with full investigation into the circumstances of exposure to lead at work and possibly elsewhere.

Massive intravascular haemolysis is caused by acute exposure to arsine, phosphine, and stibine. These are gases encountered in the smelting and refining of metals, galvanizing processes, and in

certain soldering procedures. They are formed from the reaction between nascent hydrogen and arsenic, phosphorus, and antimony respectively.

Occupational exposure to ionizing radiation can occur following the industrial use of radioactive sources to test the integrity of welds, in the health-care industry, and in nuclear power stations (see Chapter 8.5.9).

Benzene is encountered in the petroleum industry, and is used as a starter chemical for the production of other aromatic organic compounds. Its effects on the haemopoietic system include early platelet deficiency, mild haemolysis, and pancytopenia. Major effects are aplastic anaemia and lymphoid and myeloid leukaemia.

Methaemoglobinaemia can result from exposure to occupational and environmental agents such as nitrates, and nitro and amino derivatives of aromatic compounds. Specific examples are aniline, aminobenzene, nitrobenzene, and nitrates in the soil.

Infections

T. C. Aw

Certain industries are associated with an increased risk of occupational infections, and the health-care industry is possibly the best example of this. Working in tropical environments also exposes workers to the risk of tropical diseases. Occupational and environmental infections involve a range of organisms from viruses, rickettsiae, bacteria, and fungi to larger organisms such as parasites and insects. Occupations involved include farming, forestry, and sewage work.

In the health-care industry, blood-borne infections such as hepatitis B and human immunodeficiency virus (**HIV**) pose practical problems for the safety of staff during contact with infected patients, and the safety of patients during contact with infected staff. Worldwide, several dozen cases of HIV have been documented to have occurred in health-care workers following contact with infected blood or body fluids from patients. These have involved needlestick injuries, mainly from contaminated hollow-bore needles, or substantial blood contamination of damaged skin. Only two cases of HIV transmission from health-care workers to patients have been noted. One involved a dentist and the other an orthopaedic surgeon. With hepatitis B, the incidence of outbreaks from doctor to patient and vice versa is much more common. Until recently, transmission of hepatitis B from doctor to patient was thought to involve only those who were e-antigen positive carriers. There have been several cases to date involving surface-antigen positive but e-antigen negative carriers. This has led to a review of occupational procedures that may pose a risk to patients from such carriers. Protection of health-care staff from hepatitis B infection involves the intramuscular administration of yeast-derived genetically engineered vaccines (plasma-derived vaccines were previously used) to staff at risk. The vaccine does not protect against other hepatitis infections (hepatitis C, D, and G) which are similar to hepatitis B in their transmission and sequelae . Other occupational infections affecting staff in microbiological laboratories and other health-care workers include tuberculosis, salmonellosis, syphilis, and malaria.

The reproductive system

T. C. Aw

Occupational and environmental exposures may affect female and male reproductive systems and influence different stages in the outcome of pregnancy. Children can be affected by parental exposure to physical and chemical hazards.

The male reproductive system

Exposure to the pesticides 1,2-dibromo-3-chloropropane (**DBCP**) and chlordecone (Kepone), are among the best documented cases. Liquid DBCP was used as a nematocide for crops. Chlordecone was used as an insecticide. Gynaecomastia has been reported in male workers following prolonged contact with oestrogenic agents such as diethylstilbestrol, and in those involved in the preparation of oral contraceptive products.

Effects on the male reproductive system may also occur following exposure to heat (potentially a problem in workplaces such as foundries), microwave and ionizing radiation, cytotoxic drugs, animal growth promoters, fumigants such as ethylene dibromide, and heavy metals, for example lead, manganese, and mercury compounds.

The female reproductive system

Effects on the female reproductive system range from alterations in the menstrual cycle—both the amount of menstrual flow as well as the regularity of cycles—spontaneous abortions, and infertility. Rubella can be a problem for female health-care workers and those working in microbiological laboratories, especially if adequate procedures are not in place for occupational health vetting and immunization of this occupational group. Concerns about an increase in spontaneous abortions amongst female anaesthetists exposed to anaesthetic gases were investigated, and recent reviews of the findings suggest that the evidence is weak. Nevertheless, occupational exposure standards were set in the United Kingdom in 1996 for limiting workplace exposure to nitrous oxide, halothane, enflurane, and isoflurane. Similar concerns about working with visual display units (such as computer monitors) and spontaneous abortions led to several large-scale epidemiological studies. These studies indicated that the association could well have occurred by chance, given that spontaneous abortions are relatively common events and that there has been a rapid increase in use of visual display unitss in recent years.

Other occupational and environmental exposures that can affect the female reproductive system are ionizing radiation, cytotoxic drugs, lead, and ethylene oxide.

Other environmental effects

Environmental contamination with 'endocrine-disruptors' such as plasticizers has been thought to be responsible for an alteration to the sex ratio of fish populations. Pollution from similar organic compounds has been suggested to be contributing to the decline in mean sperm counts for male populations in several parts of Europe. These observations have led to proposals for testing chemicals for endocrine disrupting properties.

Neurological disorders

E. L. Baker

Neurological disorders caused by exposure to chemicals

Central nervous system effects

Exposure to chemicals, primarily encountered by workers in manufacturing, construction, and agricultural jobs, can cause transient and persistent effects on the central nervous system (Table 4). Transient central nervous system dysfunction is most commonly caused by exposure to volatile organic solvents, to organophosphate insecticides, or to carbon monoxide. In each instance, these substances, acting through different mechanisms, may cause central nervous system dysfunction ranging from acute intoxication manifested by light-headedness and dizziness to loss of consciousness and even death. Persistent central nervous system sequelae may occur following one exposure episode if exposure levels are high and the time of exposure is prolonged.

Persistent central nervous system dysfunction, manifesting as neurobehavioural performance deficits, has been reported following chronic exposure to moderate concentrations of various agents encountered in the workplace and occasionally in the environment. This syndrome, chronic toxic encephalopathy, consisting primarily of memory impairment, impaired psychomotor function, and mood disorders, has been seen following chronic exposure to lead, styrene, and certain organic solvents. In more severe cases the deficits persist, but do not progress, following cessation of exposure. If behavioural symptoms are present without evidence of abnormal neurobehavioural test performances (i.e. organic affective syndrome) reversal of these manifestations usually occurs following cessation of exposure.

Peripheral nervous system effects

Exposure to certain agents (Table 5) may cause either motor or sensorimotor polyneuropathy. Rarely, exposure to lead at high levels for long periods may cause upper extremity motor neuropathy, consisting of wrist extension weakness or wrist drop. Certain substances (for example acrylamide, hexacarbon solvents, and certain organophosphorus compounds) may act as axonal toxins causing a mixed sensorimotor polyneuropathy manifesting as symmetrical, distal sensory loss. Upon removal from exposure, the symptoms usually recede over a period of months with modest or no residual damage.

Other nervous system effects

A variety of other neurological effects have been reported following exposure to toxic agents in the environment (Table 6). In most cases, symptoms recede once exposure has ceased.

Neurological disorders caused by physical factors

Repetitive trauma disorders

See under Musculoskeletal disorders and Section 18.

Entrapment neuropathies

In certain occupations, sustained postures such as working overhead may cause muscular hypertrophy or other changes resulting in entrapment of nerve roots or individual nerves. Such conditions are diagnosed by obtaining a careful work history and are managed by modification of the work environment.

Vibration-induced neuropathy (see also Section 24)

Certain jobs involving the use of vibrating hand tools or pneumatic drills may be responsible for the occurrence of peripheral neuropathy. These disorders may originate from a combination of physical trauma to the nerve itself as well as damage to blood vessels which supply the nerves.

The role of psychology in occupational health

A. Spurgeon

There is growing recognition of the importance of psychological factors in a range of occupational health concerns. The following represents the main areas where psychology is now considered to have a significant contributory role.

Occupational stress

Reports of work-related stress have increased dramatically in recent years. A number of factors in the working environment have been identified as potential psychosocial hazards. These are usually categorized as in Table 7.

Prolonged exposure to one or more of these conditions may result in a range of symptoms of psychological distress such as feelings of anxiety, irritable or aggressive behaviour, lack of concentration, lack of confidence and an inability to make decisions, sleep disturbance, and fatigue. There may also be associated physical symptoms such as frequent headaches and nausea. Occupational stress is often identified as a result of the individual's inappropriate (maladaptive) coping strategies such as frequent short-term absences, alcohol and other substance abuse, poor time-keeping, and by uncharacteristically poor work performance. Effective management of occupational stress usually requires an integrated approach which includes attention to both the workplace and the individual and consists of intervention at three levels:

1. Primary intervention focuses on the identification of particular sources of stress in the working environment and the institution of measures to eliminate or reduce these.
2. Secondary intervention focuses on improving the coping skills of employees by the use of specific forms of stress management training (for example relaxation, conflict management, assertiveness, time management) and health promotional activities. These are particularly appropriate where workplace stressors are intrinsic to the particular occupation and therefore non-removable, for example the potential for aggressive confrontation with members of the public.
3. Tertiary intervention is concerned with counselling and rehabilitation of psychologically distressed individuals. The source of stress may often be multifactorial, and not therefore solely work-related, but it has an impact upon work performance and may be exacerbated by the demands of work

Neurobehavioural effects (see Sections 24 and 26)

The response to hazard exposure

A growing number of occupational health complaints are characterized by a lack of a firm diagnosis or clear occupational causation. These include, in particular, complaints which consist of a range of non-specific symptoms, typically headache, fatigue, nausea, depressed mood, cognitive confusion, and sometimes eye and nasal irritation. They are reported in diverse situations such as in air-conditioned offices (sick-building syndrome), proximity to low-frequency electromagnetic fields and perceived exposure to very low (often undetectable) levels of chemicals. In addition the prevalence of musculoskeletal complaints in many workplaces has been shown to be related to the presence of psychosocial hazards. At an individual level the reporting of such complaints, the decision to be absent from work, and the subsequent response to treatment and the duration of absence from work also appear to be strongly related to psychological factors. An important element is the structure of health beliefs and attitudes which the individual brings to the situation. This influences their response to real or perceived exposure to hazards by determining their selection of which information to attend to and their subsequent interpretation of that information. Specific examples of syndromes with these features which occur in both an occupational and wider community setting may be multiple chemical sensitivity and chronic fatigue syndrome. Current approaches to effective management of these and other similar conditions favour a 'biopsychosocial' approach which rejects the artificial distinction between a physically and a psychologically based complaint and treats both physical and psychological symptoms simultaneously.

Risk perception

Individual response to exposure to hazards is one aspect of the wider subject of risk perception. Public and individual perceptions of risk are invariably at odds with quantitative risk assessment based on real data. Risk perception and associated safety behaviour are known to depend on a range of attitudinal factors, notably whether exposure to the hazard is perceived to be voluntary or involuntary, whether the consequences are perceived to be short or long term, who is perceived to be responsible or to benefit, the perceived importance or vulnerability of those at risk, and the level of concern generated by the media and the activity of pressure groups. Modern approaches to risk assessment accept that subjective and objective evaluations of risk should both be taken into account when defining the tolerability of any risk to individuals or groups, either in the workplace or in the wider environment. In addition, communication about risk is unlikely to be effective if it does not take account of the existing attitudes and beliefs of the audience and the reasoning behind these. The application of this to the development of a positive safety culture is discussed elsewhere.

8.4. Occupational safety 2

Richard T. Booth

Introduction

Accidents at work occur under a variety of circumstances. A few serious accidents involve fires and explosions in chemical process plants, and transportation disasters. But the vast majority of accidents involve everyday events such as contact with moving machinery, falls, cuts from handling materials, slipping on or striking against objects, and injuries from hand tools. Road accidents at work are the main cause of occupational fatalities. Despite the diversity of accidents, it is now agreed that the causes of all these accidents and principles of safety management are common to all, and to the control of health hazards at work.

The regulation of occupational safety in the United Kingdom evolved along two largely independent pathways. The first was the development of legislative controls of safety, gradually embracing an increasing range of hazards and of industries and employers. The second pathway was the development of safety management.

The distinctive feature of industry- and hazard-based safety legislation from 1844 was that the regulators identified the hazards, implicitly assessed the risks (largely as a reaction to events), and prescribed 'blanket' rule-based control standards. Safety was driven by legislation, and a knowledge of the law was the foundation of workplace controls, as described in 1972 by the Robens' Report for the Committee on Safety and Health at Work.

In contrast, safety management began (in Great Britain at least) after World War I, promoted by the British Industrial Safety First Movement, later the Royal Society for the Prevention of Accidents. Industrial accident prevention was seen as being achieved by committees, workers' participation, the employment of safety officers, joint accident investigations, and the promotion of a positive safety culture.

Limitations of this approach were the tendency to distance safety management from operational management generally, the adoption of a reactive approach to prevention, seeing accident causation in very simple terms, and a predilection for panaceas, such as 'accident proneness'.

The first determined attempt to amalgamate these two pathways or 'traditions' of safety within a framework of law was proposed by the Robens' Report. The Committee recognized that much safety law was outdated, too detailed, prescriptive, and legalistic. Robens proposed the wholesale repeal of the then existing law and its replacement with 'goal-setting' legislation. The Robens' philosophy was partly realized by the Health and Safety at Work, etc. Act of 1974, but arguably only fully implemented in the early 1990s as a result the European Union (**EU**) Framework and other directives (heavily influenced in most cases by experience in Great Britain), together with the government's deregulation initiatives. Whereas traditional law drove the safety system, the contemporary doctrine is that the law should underpin good, and promote best, practice. To a large extent, the Health and Safety Commission and the Health and Safety Executive (**HSC**, **HSE**) together became not just the custodian of the law, but also the custodian of good practice (as evidenced by standards) and the prime mover of the safety system. The duty of employers was no longer slavishly to adhere to a set of statutory rules, but to develop their own safety organization and arrangements that could be shown to follow good practice, which employers' and employees' organizations had been involved in framing.

The last 30 years have therefore seen the recognition of safety as an integral part of company management. Risk assessment is accepted as the essential basis for prevention.

The size of the problem

During 1998 to 1999, 188 employees were killed at work in Great Britain. In addition, 65 self- employed people and 369 members of the public (including trespassers and suicides on the railways) received fatal injuries at work. The numbers of non-fatal accidents are less certain. Accidents at work in Great Britain should be reported to the relevant enforcing authority under the terms of the Reporting of Injuries, Diseases and Dangerous Occurrences Regulations 1995 (**RIDDOR**). However, the Department of Employment in its in 1992 Labour Force Survey suggested that only about 30 per cent of reportable accidents to employees were reported. Reporting levels vary sharply in different employment sectors; from 80 per cent in the energy sector to 17 per cent in agriculture. Only about 5 per cent of reportable accidents to the self-employed are reported.

In 1991, the HSE argued, from the evidence of reported accidents, that accident rates in Great Britain were lower than in other industrialized countries. But the findings of the Labour Force Survey suggest that the lower accident rates in Great Britain may be partly a consequence of serious under-reporting and also a result of differences in reporting criteria. Notably, the published Great Britain data do not include fatal and serious injury from occupational road accidents. These data are included in reports from many EU countries. Some further reservations about accident data as a measure of safety performance are mentioned later.

Table 1 shows the frequency, causes, and results of different kinds of accident. Accidents leading to fatal injuries can be markedly different from those leading to major injuries. Falls on the level cause most major injuries, but few fatalities. Falls from heights and injuries from moving/falling objects contribute substantially to both severity categories.

Table 2 shows the kinds of injuries associated with fatal and serious non-fatal accidents. Fractures (15 per cent), concussion and internal injuries (14 per cent), and injuries of more than one type (26 per cent) were the injuries most often associated with fatality. Nearly three-quarters of all major injuries are attributable to fractures. The nature of injury was unknown in 9.3 per cent of cases, presumably because these details were not given on the standard pro-forma but were included in supplementary documents that did not find their way into the statistics.

The evolution of safety management

Accident prevention requires the creation, and maintenance, of a safe working environment, and the promotion of safe behaviour—the avoidance of error—by those doing hazardous work. However, safety management has emphasized the prevention of repetitions of accidents that have already occurred, using information derived from detailed accident investigations. Accidents meriting investigation usually involve a casualty, and it is not surprising that the behaviour of those injured may dominate the minds of the investigators. Reactive prevention tends to be easier than proactive prevention. Assessing risks and devising preventive plans without the help of accident data is difficult: it involves weighing the probabilities of a wide range of unwanted outcomes, and preparing an integrated control plan to cope with all the detected hazards.

Key features of the traditional approach were the search for a primary accident cause, and the debate about whether the primary cause was an unsafe act or an

unsafe condition.

Most practical accident prevention involved the preparation of a safety rule designed to prevent a recurrence of the unsafe act, or a physical safeguard to obviate the unsafe condition existing immediately before the accident.

The causation debate, with its political overtones and desire to apportion blame, often missed three crucial inter-related, issues:

- The concept of a single primary cause for an accident is a bizarre simplification of a complex multicausal process. The term 'unsafe act' embraces a wide range of unintentional errors and intentionally risky behaviour 'violations'.
- The distinction between the contribution of unsafe conditions and unsafe acts in causation has masked the more important distinction between the relative contribution of conditions and behaviour in prevention, and the need for prevention plans to promote both safe conditions and safe behaviour.
- The argument has focused almost exclusively on the errors made by those people involved in the accidents, not the managers and engineers whose errors (remote in time and place from the accident) may have created a physical environment conducive to the risk of serious accidents.

The safety management approach proved inadequate to cope with major hazards created by rapidly developing technology. Preventive measures may be rendered obsolete by each technical advance. Rules and safeguards devised in the aftermath of accidents may later be perceived as overzealous and may conflict with the needs of both employers and employees to get the job done. Both parties may tacitly conspire to evade the safety rules or to defeat the physical safeguards. Measures taken to prevent one specific accident may conflict with the measures adopted to prevent a different accident, and with production-oriented rules. Rule books and legislation drawn may become incomprehensible and contradictory. At least two company rules may exist for any situation: the rule to get the job done in time, and a more demanding rule that may be invoked when things go wrong.

Proactive safety management

Accident investigations and prevention programmes must address the distinctive elements of the accident causation process described below.

H4>Multicausality

Few accidents are associated with a single cause. Rather, they happen as a result of a chance concatenation of many distinct causative factors, each one necessary but not sufficient to cause a final breakdown. The coverage of prevention plans should therefore seek to permeate all aspects of the organization's activities. Accident investigators should continue to seek out causative factors even when 'a familiar, abnormal event is found which is therefore accepted as explanation, and a cure is known'.

Active and latent failures

Active failures are errors that have an immediate adverse effect, while latent failures lie dormant in an organization becoming evident only when combined with local triggers. The triggers are the active failures: unsafe acts, and unsafe conditions such as mechanical failure.

Skill-, rule-, and knowledge-based errors, and violations

The standard framework for classifying error is the skill—rule—knowledge-based model.

Skill-based errors involve 'slips' or 'lapses' in highly practised and routine tasks. Knowledge-based errors are failures to create an adequate new rule to cope with a situation. Violations, or risk taking, are another category of error. Someone deliberately does something contrary to a rule, such as an approved operating procedure.

Hazard identification, risk assessment, and preventive action

These are the essential foundation of safety management—the avoidance of latent failures and of safe personal behaviour in the face of danger—the avoidance of active failures.

To create and maintain a safe working environment, and to work safely in a dangerous environment, people must have the knowledge and skills and must know the rules, and be motivated, to:

1. identify hazards;
2. assess accurately the priority and importance of the hazards (risk assessment);
3. recognize and accept personal responsibility for dealing with the hazards in an appropriate way;
4. have appropriate knowledge about what should be done (including specified rules);
5. have the skills to carry out the appropriate necessary sequence of preventive actions, including monitoring the adequacy of the actions, and taking further corrective action.

The organization should be aware of circumstances where managers, supervisors, and other personnel may:

- underestimate the magnitude of risks;
- overestimate their ability to assess and control risks;
- have an impaired ability to cope with risks.

The aims of safety management

The aim of safety management is not limited to hazard identification, control, and monitoring. Employers must plan for safety. Decisions have to be made about priorities for resource allocation, training needs, the appropriate risk-assessment methods to be adopted, the need to assess human reliability, and the choice of tolerable-risk criteria. Safety criteria should underpin every decision made by the enterprise and must be considered as an integral part of day-to-day decision-making. The employer must establish an organization and communications systems that helps to integrate safety within the management process, and which ensures that everyone is fully informed about safety issues, and ideally has had an opportunity to discuss them.

Key functions of safety management

1. *Policy and planning*: relies on the determination of safety goals, objectives, and priorities, and a programme of work designed to achieve the objectives, which is subject to measurement and review.
2. *Organization and communication*: involves establishing clear lines of responsibility and two-way communications at all levels.
3. *Hazard management*: depends on ensuring that hazards are identified, risks assessed, and control measures determined, implemented, and subject to measurement and review.
4. *Monitoring and review*: requires the establishment of whether the above steps are in place, in use, and work in practice.

The four key elements of safety management are underpinned by the requirements of the Management of Health and Safety at Work Regulations 1999.

Safety management and safety culture

These procedures and systems are necessary elements of an effective safety programme, but are not the whole story. There is a danger that an organization's safety policies, plans, and monitoring arrangements, although appearing well considered and comprehensive, may create an aura of respectability which disguises sullen

scepticism or false perceptions among influential people at management and shop-floor levels.

The concept of safety culture

The Health and Safety Commission has defined safety culture as follows:

> The safety culture of an organization is the product of individual and group values, attitudes, competencies, and patterns of behaviour that determine the commitment to, and the style and proficiency of, an organization's health and safety programmes.
>
> Organizations with a positive safety culture are characterized by communications founded on mutual trust, by shared perceptions of the importance of safety, and by confidence in the efficacy of preventive measures.

The Confederation of British Industry has reported the results of a survey of 'how companies manage health and safety'. The idea of the culture of an organization was incorporated in the report's title 'Developing a safety culture'. The dominant themes to emerge were:

1. the crucial importance of leadership and the commitment of the chief executive;
2. the executive safety role of line management;
3. involvement of all employees;
4. openness of communication; and
5. demonstration of care and concern for all those affected by the business.

The Health and Safety at Work Act requires all companies employing five or more people to prepare a health and safety policy. A written corporate statement on the safety policy and organization is a crucial element in the promotion and maintenance of a positive safety culture within the organization, and of high standards of safety awareness in the minds of both management and workforce.

The policy should embody a positive approach to the management of safety. It must be more than a one-off event written to fulfil the letter of the Act; its objective should be to establish the corporate attitude to safety and the necessary organization through which the safety objectives can be assured. This must be subject to regular, systematic review.

Monitoring safety performance

Many companies measure safety performance merely by counting the number of accidents, a belated and potentially misleading statistic. In many organizations there are not enough accidents to determine whether differences between sites or over time are due to real differences, or to chance. Company and national accident statistics are influenced by variations in the time that individuals choose to take off for a given injury severity.

The design of a health and safety monitoring system for an enterprise must address two crucial issues:

1. There is no single unambiguous measure of safety performance that is resistant to abuse. Accident data must be combined with other measures.
2. Monitoring should be designed to both check and promote compliance.

Performance measures should be designed to permeate every activity within the organization. A battery of distinctive tests should be incorporated into the safety programme, so that the limitations of one are balanced by the strengths of another.

Safety auditing

Proprietary safety auditing systems are often used to measure safety performance, and to identify aspects of safety management requiring improvement. Audits typically comprise a checklist with about 400 questions and a scoring system. These audits allow companies to compare their safety management procedures against objective criteria, but some companies may try to improve their audit score in ways that do not lead to real improvements in safety.

Safety training

In the past, too much training was of poor quality and of doubtful effectiveness, but training is an essential part of the company's safety arrangements. The Institution of Occupational Safety and Health have published a safety training policy that emphasizes the need for:

1. explicit training in organizations to promote and maintain a positive safety culture;
2. training of senior managers to be competent in strategic safety management;
3. training of managers and workers to be competent in hazard identification, risk assessment, and control;
4. training of trainers to be competent in safety training;
5. in-company evaluation of training effectiveness.

Further reading

Booth RT (2000). Challenges and opportunities facing the Institution of Occupational Safety and Health. *Journal of The Institution of Occupational Safety and Health* **4**, 7–21.

Committee on Safety and Health at Work (Robens Committee) (1972). *Safety and health at work*. HMSO, London.

Confederation of British Industry (1990). *Developing a safety culture*. CBI, London.

Department of Employment (1992). Health and safety statistics 1990–1991. *Employment Gazette* (*Occasional Supplement 3*) **100**, September.

Hale AR, Glendon AI (1987). *Individual behaviour in the control of danger*. Elsevier, Amsterdam.

Health and Safety Commission (1991). *Second report: human reliability assessment—a critical overview. ACSNI Study Group on human factors*. HMSO, London.

Health and Safety Commission (1993). *Third report: organising for safety—ACSNI study group on human factors*. HMSO, London.

Health and Safety Commission (2000). *Health and safety statistics 1999/2000*. HSE Books, Sudbury, Suffolk, UK.

Health and Safety Executive (1991). *Workplace health and safety in Europe*. HMSO, London.

Health and Safety Executive (1997). *Successful health and safety management*, Health and Safety Series booklet HS G 65. HSE Books, Sudbury, Suffolk, UK.

Heinrich HW (1969). *Industrial accident prevention*, 4th edn. McGraw Hill, New York.

Institution of Occupational Safety and Health (1992). *Institution policy statement on safety training*. IOSH, Leicester.

International Nuclear Safety Advisory Group (1988). *Basic safety principles for nuclear power plants*, Safety Series No 75-INSAG-3. International Atomic Energy Authority, Vienna.

International Nuclear Safety Advisory Group (1991). *Safety culture*, Safety Series No 75-INSAG-4. International Atomic Energy Authority, Vienna.

Rasmussen J (1987). Reasons, causes and human error. In: Rasmussen J, Duncan K, Leplat J, eds. *New technology and human error*, pp. 293–301. Wiley, Chichester, UK.

Reason JT (1990). *Human error*. Cambridge University Press.

Zohar D (1980). Safety climate in industrial organisations: theoretical and applied implications *Journal of Applied Psychology* **65**, 96–102.

8.5.1 Environmental extremes—heat

M. A. Stroud

Thermoregulation in the heat

Most of human evolution took place in Africa and hence all races are heat tolerant. Indeed, we try to maintain a near tropical microclimate against our skin, by using clothing to reduce heat loss to our surroundings. Our thermal balance is regulated by the hypothalamus which integrates information from skin temperature sensors with core temperature data from receptors in the walls of large blood vessels and in the brain. Rising temperatures trigger both behavioural and physiological responses.

Behavioural changes include reducing physical activity, altering clothing, and seeking shade or cool shelter. Cold drinks are also helpful. Although these responses seem simplistic, decisions may not be straightforward. If physical activity is low and water is in short supply, it is better to increase clothing cover and protect yourself from high radiant heat inputs. If activity must be continued and water is freely available, minimal clothing to permit maximal sweat evaporation is preferable. Immediate physiological responses involve vasodilatation of skin and subcutaneous blood vessels to enhance surface heat loss from radiation, conduction, and convection. The vasodilatation is triggered by a sympathetic cholinergic reflex in response to skin warming with additional direct effects of heat on arteriolar tone. In a resting individual, skin vasodilatation can maintain thermal equilibrium in environmental temperatures up to 32°C, but with higher temperatures or heat production from activity, core temperatures will rise. This will trigger sweating to promote evaporative cooling.

Heat acclimatization

Repeated heat exposure can increase our capacity to lose heat by about 20-fold. This is partly due to greater skin blood flow from increases in circulating volume and improved vasodilatory responses, but changes in sweating responses are more important. In the non-acclimatized, sweating is triggered by a rise in core temperature of about 1°C and maximum rates are limited to about 0.5 l/h. Following acclimatization, a 0.5°C core rise will trigger the response and rates may exceed 2.0 l/h. Acclimatization also leads to aldosterone-mediated reductions in sodium loss in both sweat and urine. The acclimatized individual therefore requires no sodium supplementation and giving supplements can delay the acclimatative process. Avoiding them altogether, however, does risk salt depletion in non-acclimatized persons during prolonged heat stress. Acclimatization develops swiftly and around 90 per cent of maximum heat tolerance is present after 7 to 10 days on which core temperature has risen by more than 1°C for more than 1 h. Physical exertion combined with heat makes changes even more rapid. After returning to cool environments, adaptation is lost in 20 to 40 days.

Susceptibility to heat-related illness

Although we are generally heat tolerant, heat-related illness is relatively common and a number of factors increase vulnerability. Above an environmental temperature of about 35°C, we tend to gain heat from our surroundings and this, along with metabolic heat production, can only be lost via evaporation of sweat. High humidity, hot environments are therefore the greatest threat. Acclimatization status has a marked influence on heat-related risks with the unacclimatized prone to hyperthermia and salt depletion. The fully acclimatized are vulnerable to dehydration from high sweat rates. Dehydration in itself limits sweating capacity and skin blood flow and hence increases risks. It can occur easily since thirst is a poor trigger for adequate drinking. Sweat rates can also exceed gut capacity for water absorption.

Prolonged physical activity can cause heat illness under quite modest environmental conditions. This is particularly common when individuals are obliged to wear insulative or non-vapour-permeable clothing. Military heat casualties are sometimes due to these factors but there have also been fatalities in soldiers who have been heat susceptible for no obvious cause. Such genetic or constitutional vulnerability should be suspected whenever a heat-related problem occurs following relatively modest heat stress. These individuals should be strongly advised to avoid similar circumstances in future. Obesity and poor physical fitness are further risk factors in the heat, as is diabetic autonomic dysfunction. The elderly are generally heat sensitive and, in addition, are prone to problems from the increased circulatory demands of vasodilatation. Drugs can also induce heat illness (see below).

Heat exhaustion

Most casualties in hot environments suffer from heat exhaustion. There is usually a history of prolonged heat stress followed by nausea, weakness, headache, thirst, and sometimes collapse. Patients appear dehydrated with a tachycardia and low blood pressure. If hyperthermic, the casualty should be complaining of feeling hot and should appear flushed and sweaty. The absence of these symptoms and signs, especially with a very high core temperature, suggests heat stroke. Heat exhaustion is ascribable to sodium and/or water depletion but discriminating between the predominant loss can be difficult. Sodium depletion tends to be greater if the casualty was poorly acclimatized and hence sweated relatively more sodium than water. Conversely, water depletion is more common in acclimatized individuals. Muscle cramps or whole body dehydration without marked changes in haematocrit or serum proteins are suggestive of excessive sodium loss, but serum sodium tends to be normal in such cases unless enthusiastic fluid replacement without salt has led to hyponatraemia. This sometimes occurs in runners following marathons in hot environments. In predominantly water-depleted heat exhaustion, haematocrit, serum proteins, and serum sodium tend to be high. Renal impairment occurs in either form of heat exhaustion and the treatment of both types often requires 5 to 10 l of oral or intravenous fluids in the first 24 h. Sodium supplementation is given as appropriate but if unsure about sodium status it is usually safer to provide some than to precipitate acute hyponatraemia.

Heat stroke

Mild heat stroke has occurred when a hot environment or high activity levels have led to pyrexia with cerebral disturbance. Core temperature is usually 38 to 41°C. The condition frequently follows heat exhaustion but temperature rise may have occurred rapidly with no time for salt or water depletion. Sufferers have headaches and may be either drowsy or irritable. They often hyperventilate. The great danger is progression to more severe heat stroke, in which core temperature reaches levels that cause irreversible denaturing of proteins. This usually occurs at above 41.5°C. Damage is widespread and particularly effects brain, liver, kidney, and muscle. Furthermore, the hypothalamic thermoregulatory centre may fail, switching off vasodilatation and sweating, and switching on cold defences inappropriately. Patients may therefore claim to feel cold and on examination may be shivering with a dry, vasoconstricted skin. A disastrous vicious cycle of increasing temperatures can then ensue.

Treatment for all heat stroke requires early recognition and rapid cooling. Tepid water and fan-assisted evaporation may be more effective than cold water immersion which can limit heat loss through intense peripheral vasoconstriction. Intraperitoneal fluids, paralysis, and ventilation may be needed and, in extremis, cooling by cardiac by-pass should be considered. Hyperkalaemia, hypocalcaemia, acidosis, rhabdomyolysis, disseminated intravascular coagulation, and hepatic or renal failure are all common complications. Ventricular fibrillation is a frequent terminal event. Even if apparently resuscitated and cooled successfully, a 12 to 24-h 'lucid interval' may precede major deterioration. Permanent neurological damage is common.

Drug-induced heat illness

Many drugs can cause mild degrees of pyrexia by inducing local or systemic inflammation or hypersensitivity. Some also increase susceptibility to environmental heat by inhibiting central thermoregulation (e.g. barbiturates and phenothiazines) or reducing sweating capacity (e.g. anticholinergics). Salicylate overdose can generate heat stroke by increasing metabolic heat production while impairing hypothalamic regulation. There are two types of heat-related drug reaction, however, which are

particularly dangerous.

Malignant hyperpyrexia

This is usually a dominantly inherited condition although different gene defects may effect families. Administration of a variety of anaesthetic agents, including halothane and suxamethonium, leads to rapid, massive heat production from generalized increases in skeletal muscle tone. Contraction is triggered at the muscle cell membrane and hence neuromuscular blocking agents are ineffective. Intravenous dantrolene, an inhibitor of muscle calcium flux, is helpful and can be used along with ventilation and cooling/supportive measures. Fatalities are common and it is therefore important to avoid risks whenever possible. In patients with a relevant personal or family history, in whom an anaesthetic is unavoidable, oral dantrolene should be given prior to the use of low-risk agents.

Neuroleptic malignant syndrome

This condition has similarities to malignant hyperpyrexia but is induced by idiosyncratic reactions to normal doses of antidopaminergic drugs including phenothiazines and butyrophenones. The onset is less rapid than malignant hyperpyrexia, occurring over a few days. The increased muscle tone is also induced presynaptically and hence neuromuscular blocking agents help. Some recreational drugs such as ecstasy may induce this type of response, although most cases of ecstasy-induced hyperthermia are probably cases of heat stroke induced by enthusiastic dancing with limited fluid intake in hot, humid environments.

Further reading

Hodgson P (1991). Malignant hyperthermia and the neuroleptic malignant syndrome. In: Swash M, Oxbury J, eds. *Clinical neurology*, pp. 1344–5. Churchill Livingstone, Edinburgh.

Hubbard RW, Armstrong LE (1988). The heat illnesses: biochemical, ultrastructural, and fluid-electrolyte considerations. In: Pandolf KB, Sawka MN, Gonzalez R, eds. *Human performance physiology and environmental medicine at terrestrial extremes*, pp. 305–59. Benchmark, Indianapolis.

Stroud MA (1993). Environmental temperature and physiological function. In: Ulijaszek SJ, Strickland SS, eds. *Seasonality and human ecology*, pp 38–53. Cambridge University Press.

8.5.2Environmental extremes—cold

M. A. Stroud

It has only been 10 000 to 15 000 years since ancestral humans dwelt exclusively in warm or hot climates. Humans are therefore poorly cold adapted and hypothermia occurs quite frequently even in temperate regions. With water immersion it may occur in the tropics. In truly cold areas, there is also the risk of non-freezing cold injury and frostbite. Nevertheless, behavioural changes allow us to operate safely in the coldest environments.

Thermoregulation in the cold

Core temperatures in the cold are usually maintained by adjustments in clothing and physical activity. The latter can increase heat production from a resting 100 watts to 1 to 2 kilowatts. This is very effective. While it takes highly specialized, multilayered clothing to keep warm while inactive in an environment of +5°C, clothing insulation equivalent to normal office dress (1 clo) will maintain core temperature at –20°C when working moderately hard.

Our limited physiological cold protection is under hypothalamic control. Falling surface and, to a lesser extent, core temperatures lead to decreased blood flow in the skin due to increased sympathetic adrenergic tone and direct cooling effects of cold on skin arterioles. This minimize surface heat loss. Unfortunately, vasoconstriction also leads to severe cooling of the hands and feet with problems of temporary skin numbness and muscle weakness, and risks of more permanent peripheral cold injury. It is often this peripheral cooling that limits our capacity to work in the cold.

Falling skin temperatures will also lead to higher resting muscle tone and shivering, especially when declining core temperature releases hypothalamic inhibition of shivering. These mechanisms can only increase resting heat production to around 500 watts and, unlike newborn infants and some other mammals, adult humans cannot add significant non-shivering heat production to this figure.

Effects of falling core temperature

Falling core temperature leads to progressive decline in function. At 34 to 36°C, hypothermic individuals are conscious of feeling cold and try to move around, add clothing, or seek shelter. Simultaneously, physiological defences are activated. With further falls, mental and physical problems increase with some individuals becoming withdrawn while others exhibit aggression or disinhibition. Once core temperatures reach 33 to 34°C, victims often stagger and become confused or drowsy. It is also around this point that 'paradoxical undressing' may occur. This phenomenon is well described and appears to be due to hypothalamic dysfunction with alteration of set-point temperature. Victims therefore think that they are hot and appropriate behavioural and physiological responses disappear. At core temperatures varying between 26°C and 32°C, coma will ensue and between 17 to 26°C cardiac output becomes inadequate to sustain life for prolonged periods. The risk of ventricular fibrillation is also high. Nevertheless, successful resuscitations of victims with core temperatures below 15°C have been reported.

Causes of hypothermia

A number of factors increase hypothermic risk. Wetting of skin or clothing extracts enormous amounts of heat and reduces insulation of garments. Complete immersion is particularly hazardous and globally, more than 100 000 people per year die of cold shock or inexorable hypothermia in the water. This far exceeds deaths from drowning without cold. Winds also increase environmental cooling and a still air temperature of +5°C equates to –50°C if wind speed is 40 km/h. Coupled with rain, these effects often contribute to hypothermic accidents amongst hill walkers and mountaineers, although in these cases fatigue may contribute. Prolonged exertion depletes muscle glycogen which reduces heat production capacity from both exercise and shivering. A low blood glucose also impairs hypothalamic temperature control.

Small, thin individuals cool easily due to increased surface-to-volume ratios. They also have reduced subcutaneous insulation and low heat producing mass. A fat individual can maintain core temperature at rest, even if mean skin temperature is 12°C, whereas a thin individual struggles to maintain thermal equilibrium with a skin temperature of 25°C. Sometimes, however, rapid cooling can have benefits. A small child in cold water may cool so rapidly that vagally triggered bradycardia and lowered brain metabolic demands may permit successful resuscitation after very prolonged immersion.

The elderly may also be small and thin and are at risk of so-called 'urban hypothermia'. Poverty, illness, immobility, malnutrition, and a less sensitive regulatory system may contribute but, in many cases, hypothermia on admission to hospital is secondary to other pathology, for example a stroke may have led to prolonged immobility in a cool environment. Drugs that impair consciousness or induce vasodilatation are risk factors, and alcohol is particularly hazardous. Alcoholics with no fixed abode and tendency for hypoglycaemia are frequent urban cold casualties.

Hypothermic illness

General management of the hypothermic casualty is similar to that for any comatose or semicomatose individual. Abnormalities in blood gases, pH, electrolytes, and glucose are common, and pancreatitis or rhabdomyolysis are recognized complications. Accurate measurement of core temperature is surprisingly difficult and axillary, tympanic, and oral temperatures can all be misleading. A low reading rectal thermometer is best. Hypothermia has one, very specific risk. Pronouncement of death is fraught with difficulty since profound bradycardia, minimal stroke volume, and marked respiratory depression occur. The old adage that you are 'never dead unless warm and dead' must be taken seriously.

A variety of rewarming methods are available. Warm blankets and hot drinks will suffice in many cases but, although used widely, metallized 'space blankets' are of no proven benefit. Warmed intravenous fluids are helpful and, in extreme cases, peritoneal warmed fluids or cardiac bypass can be used. Specialized equipment providing heated, humidified air also permits core rewarming. Hot baths are effective but difficult to utilize safely since a paradoxical fall in core temperature can occur as blood flow is rapidly restored to cold limbs. In general, if cooling was prolonged in onset or duration, rewarming must be undertaken with extreme caution. In critical cases, where rapid rewarming is needed, full resuscitation facilities must be available, although safe defibrillation in the presence of water is impossible.

Careful monitoring during rewarming is vital. Blood volumes are often low due to early cold-induced diuresis followed by the inability of hypothermic kidneys to retain salt and water. In immersion casualties, hydrostatic effects on the limbs may have promoted additional fluid loss and, if possible, these individuals must be kept recumbent throughout rescue and rewarming to minimize risks from extreme postural hypotension. Warming cell membranes are extremely unstable and uncontrollable fluxes in potassium and other electrolytes may occur although care must be taken in interpreting biochemical results from cold peripheral blood sampling.

Non-freezing cold injury

Local temperatures of less than12°C prevent normal membrane pumping and paralyse nerve and muscle conduction. If such cooling is prolonged, permanent damage may ensue. Immersion in cold water is particularly likely to cause this type of damage and soldiers in military campaigns are frequent victims of 'trench foot'.
Long-term damage is likely whenever an anaesthetic, paralysed, cold region becomes hot, red, painful, and swollen after rewarming, although this change may take several days. Degeneration of nerve and muscle can then follow leading to prolonged anaesthesia, muscle contractures, or inappropriate peripheral vascular control with intolerance to local heat or cold. There may be slow improvement over months or years.

Frostbite

Human tissues freeze at around –2°C. Ice forms outside cells but the remaining extracellular fluid becomes hyperosmolar and hence severe intracellular dehydration occurs. This denatures proteins. Vascular endothelial cells are particularly vulnerable and following rewarming small blood vessels may leak plasma and then become blocked by red cell sludge and clot. Additional ischaemic necrosis is then superimposed on the frost damage.

Frozen tissues appear hard and white and are anaesthetic. Rewarming leads to pain and swelling often accompanied by blistering. Deep freezing results in irreversible necrosis but appearances can be misleading and early amputation of digits should be avoided. If still frozen, rewarming is best achieved rapidly by using immersion in water at 40 to 42°C, although any thawing should be avoided if refreezing is likely. Once thawed, treatment is similar to that used for burns with prevention of infection paramount. Generous analgesia is required.

Further reading

Granberg PO (1997). Cold injury. In: Chant ADB, Barros D'Sa AAB, eds. *Emergency vascular practice*, pp. 119–34. Arnold, London.

Hamlet MP (1988). Human cold injuries. In: Pandolf KB, Sawka MN, Gonzalez R, eds. *Human performance physiology and environmental medicine at terrestrial extremes*, pp. 435–66. Benchmark, Indianapolis.

Stroud MA (1993). Environmental temperature and physiological function. In: Ulijaszek SJ, Strickland SS, eds. *Seasonality and human ecology*, pp. 38–53. Cambridge University Press.

8.5.3 Drowning

Peter J. Fenner

Introduction

Drowning is an important cause of accidental death and neurological damage, particularly in children, and is usually preventable. Drowning is listed in the Global Burden of Disease Study as the fourth most common injury worldwide (behind road-traffic accidents, self-inflicted injuries, and violence but ahead of war deaths), and is the twentieth most common cause of death worldwide, with numbers estimated at 504 000 each year.

Drowning is defined as death from suffocation by submersion in a liquid, usually freshwater or seawater. Near-drowning is survival, at least temporarily, from suffocation by submersion.

Mortality and morbidity

Acute hypoxia is the cause of the haemodynamic effects and death. Neurological morbidity in survivors of near-drowning includes difficulty with learning, memory, attention, and planning and cerebral palsy. A major study of childhood immersions has shown that approximately 68 per cent of survivors have no neurological defect, 29 per cent have some deficit, and 3 per cent will live in a permanent vegetative state. Economic costs to the community are immense.

Epidemiology

Bathtub and bucket drownings may involve infants and toddlers under the age of 12 months. Ten per cent of fatal bucket or tub immersions are the result of child abuse. Drowning rates in young children worldwide, many of whom are unsupervised, have decreased little despite preventive strategies such as fencing swimming pools. However, despite the lack of preventive strategies aimed at older children, drowing rates in older children have declined dramatically in the last decade. Ocean drownings are less common than freshwater drownings. Fewer children swim unsupervised in the ocean, and the preventive and rescue efforts of lifesaving and lifeguard associations guarding the beaches have proved effective, especially in Australia and the United States. Reasons for variations in rates of drowning include climate, the availability of beaches, lakes, and other natural and artificial water sources, employment of lifeguards at water parks, use of different kinds of vessels suitable for recreation (kayaks, personal watercraft, etc.), provision and use of lifejackets, and popularity of hobbies, pastimes, and professions associated with a risk of drowning (for example fishing). Recently a study provided proof of the principle that an unexplained drowning or near drowning may have a genetic basis.

Ethnicity

White American children aged 1 to 4 years have twice the drowning rate of African-American children of a similar age, these occurring mainly in residential swimming pools. Conversely, among children and young people aged 5 to 19 years, the rate of drowning in African-Americans is greater than for white Americans. In Australia deaths from drowning in Aboriginal children are more frequent than in whites.

Alcohol

Alcohol affects vision, balance, and movement and is a risk factor in drowning for both adolescent and adult swimmers, as well as for boat operators and passengers, who may fall overboard while intoxicated. Some 25 to 50 per cent of adult drowning victims may have had some exposure to alcohol at the time of rescue, resuscitation, or death.

Pathophysiology

Early animal studies in unanaesthetized dogs suggested that spontaneous respiratory efforts continue for around 60 s after immersion. Complete cardiac arrest supervenes after 4.5 min (mean 262 s). Aspiration is usual in drowning and near-drowning. It has been suggested that approximately 10 to 15 per cent of drowning victims do not aspirate water ('dry' drowning, possibly occurring from laryngeal spasm). However, further studies of these data by Modell suggest the conclusion may be without foundation and the effect may be due to cardiac standstill or other causes of sudden death; careful postmortem examination is required.

The haemodynamic effects following inspiration of liquid are the same irrespective of whether hypotonic, isotonic, or hypertonic solutions are involved, and are similar to anoxic controls. Both groups show a rapid fall in cardiac output, an increase in pulmonary capillary wedge pressure, central venous pressure, and pulmonary vascular resistance. Reduction in the dynamic compliance of the lungs is similar following inspiration of all types of solutions. However, aspiration of large volumes of hypertonic seawater draws fluid by osmosis from the circulation into the lung, resulting in fluid-filled, non-ventilated but perfused alveoli, incapable of normal gas exchange. Aspiration of large amounts of hypotonic fresh water may cause sufficient absorption of fluid into the circulation from the alveoli to cause both acute hypervolaemia and haemolysis, although within an hour redistribution of fluid and pulmonary oedema occurs, causing a decreased circulating blood volume. Although Modell suggested that 85 per cent of human drowning victims aspirate only 22 ml/kg of water or less, Conn estimated that water equivalent to about 10 per cent of body weight may be absorbed from the lungs during freshwater drowning.

The brain has a limited ability to maintain ATP anaerobically when cerebral blood flow is reduced. Without cerebral blood flow, the brain suffers irreparable damage within 4 to 6 min. There is some potential, however, for the restoration of activity for up to 60 min of total anoxia in hypothermia. Death or severe neurological impairment occurs after sumbersion of more than 5 to 10 min. Bystanders' estimates of sumbersion time are usually inaccurate.

Hypothermia

Submersion (not immersion) in ice-water with associated hypothermia is an important cause of near-drowning. Continuous aspiration of cold water results in rapid reduction in the core temperature while the circulation is intact. Such victims may survive, protected by hypothermia, with little or no neurological deficit after long submersion with extreme anoxia. In water at 16 °C with maximum submersion for 10 min or less, a good outcome could often be predicted.

Intact neurological survival occurred in a 6-year-old boy with a rectal temperature of 16.4 °C. after submersion for 65 min when the blood was rewarmed over 96 min in steps of 3 °C. In adults, success is less common. A notable exception was a 31-year-old man with a core temperature of 23 °C who had been asystolic for 80 min

and was warmed by cardiopulmonary bypass and recovered. In warm-water drownings, Frates was unable to show any statistically significant correlation between duration of submersion and survival.

Causes of drowning

Drowning occurs in many different situations including accidental immersion of people with poor or no swimming ability, with head and neck injuries, following cardiac and neurological emergencies (including epilepsy), as a result of impaired ability (including the effects of alcohol and drugs), metabolic disease (including hypoglycaemia), and even child abuse and murder. In countries with large coastlines or bathing beaches, drowning is common and is often caused by swimmers being caught in rip currents (large volumes of water returning back out to sea after onshore wave action): there is no such entity as the frequently suggested 'undertow'. Swimmers in difficulty may be able to shout for help but, contrary to public opinion, those drowning do not. Most drowning victims adopt a characteristic vertical position in the water—legs hang vertically, head tilted back for quick exhalation and inhalation before bobbing under water, with no time nor sufficient breath to call for help. After only 20 to 60 s, victims may submerge permanently.

Clinical features

Prognostic indicators

Success or failure of the resuscitation of the near-drowned depends on the promptness and adequacy of emergency resuscitation and subsequent respiratory intensive care. Up to 25 per cent of drowning victims presenting to the emergency department will die and a further 6 per cent have neurological sequelae.

Those with a normal chest radiograph on admission survive. PaO_2 may not relate to radiographic appearances. Although the cause and pathophysiological changes of pulmonary insufficiency vary depending on the type and volume of fluid aspirated, serum electrolytic and haemoglobin concentrations (or haematocrit) are unhelpful in predicting survival.

Cardiovascular status

This is a better guide to outcome than the neurological status. Mortality is high in victims with circulatory arrest on admission but victims with sinus rhythm, reactive pupils, and neurological responsiveness at the scene of immersion have good outcomes. Victims who are asystolic on arrival at hospital and remain comatose for more than 3 h have a poor outcome unless they are hypothermic. Rapid hypothermia from sudden submersion in cold water (see Chapter 8.5.2) carries a relatively good prognosis, compared with hypothermia after prolonged submersion and cardiac arrest.

Neurological status

Victims who are alert when medical help arrives have a survival rate approaching 100 per cent, whereas the prognosis in those who are comatose with fixed dilated pupils is poor. Victims with blunted consciousness have a survival rate of 87 per cent with no neurological defects, 2 per cent with minor defects, and 11 per cent die. Approximately 40 to 50 per cent of victims who are comatose on arrival have incapacitating brain damage. Those with no spontaneous limb movements and abnormal brainstem function 24 h after the accident have a poor neurological outcome.

A modified Glasgow coma score is helpful in evaluating neurological injury. A score of 5 or less predicts a mortality risk of over 80 per cent. Pupil reactions at the time of arrival differentiate survivors from fatalities but could not differentiate between those with minor or incapacitating neurological deficits. Fixed dilated pupils or total flaccidity are associated with a high mortality. Victims with any motor activity, even posturing or seizures in the immediate postresuscitation period, had a higher incidence of intact survival, but posturing movements persisting or recurring after 12 to 24 h, indicate a high probability of severe brain damage.

An abnormal computed tomography scan in the initial 36 h following an immersion incident is associated with a dismal prognosis. Magnetic resonance imaging with qualitative and quantitative magnetic resonance spectroscopy data may allow a more accurate prognosis.

The gravity of the early clinical state, the estimated duration of cardiorespiratory arrest, the severity of the hypothermia, seizures, and paroxysmal motion activity do not determine the severity of near-drowning encephalopathy. Early EEG patterns with moderate background activity, sleep patterns, response to auditory and painful stimulations, and numerous beta rhythms suggest a good outcome whereas bad outcomes are suggested by high-voltage, rhythmic delta waves, biphasic sharp waves, monotonous EEG, 'burst-suppression' pattern, and the absence of beta rhythms. Children without spontaneous movements and normal brainstem function 24 h after near-drowning suffer severe neurological deficits or death.

Treatment

The near-drowning victim must be treated immediately for ventilatory insufficiency, hypoxia, and the resulting acidosis. A successful outcome depends on early effective resuscitation at the scene and on competent intensive life support. Australian surf lifesaving teaching has shown that respiratory resuscitation in deep water can be effective.

Immediate

Lying the victim on his or her side for assessment of the airway and breathing will assist drainage of any excess water from the lungs. On-site cardiopulmonary resuscitation is necessary, with all victims having supplemental oxygen as soon as possible, preferably at the scene using positive airway pressure (bag, valve, mask). An oropharyngeal airway or endotracheal tube should be inserted in comatose victims if suitably qualified personnel are present. Pulse oximetry is helpful. Vomiting and regurgitation are a significant risk during early resuscitation. Respiratory and cardiopulmonary arrest may occur after an apparently successful rescue, mandating close, uninterrupted monitoring and the early administration of oxygen to all immersion victims.

On hospital arrival

On arrival at the hospital, after initial establishment of a clear airway and cardiocirculatory support, arterial blood gas tensions and pH should be measured. The pH of the blood will indicate whether a metabolic acidosis remains secondary to a significant period of hypoxia.

Mechanical ventilation may be necessary with positive end-expiratory pressure or continuous positive airway pressure. A central venous catheter, or pulmonary artery catheter in selected cases, helps to assess the effective circulating blood volume to guide fluid therapy.

In both freshwater and seawater aspiration, large volumes of intravenous colloid are usually needed while circulating blood volume and cardiac output are estimated. Failure of response to intravascular replacement with 20 ml/kg of colloid is an indication for starting inotropes. Steroid and prophylactic antibiotic therapy do not appear to increase the chance of survival.

Inpatient treatment

Extracorporeal membrane oxygenation has been used successfully for the treatment of adult respiratory distress syndrome secondary to near-drowning, although this addresses only the pulmonary not the cerebral injury. Patients with severe hypoxaemia may also have irreversible cerebral ischaemia.

If adult respiratory distress syndrome occurs, it is usually within 6 h of admission. There is evidence that alveolar epithelial barrier function is well preserved even after aspiration of large quantities of hypertonic salt water. Surfactant has been used with some success in refractory respiratory failure in near-drowning, but it is expensive.

The risk of secondary pneumonia is high, especially when mechanical ventilation has been used. Although prophylactic antibiotics are not recommended broad-spectrum antibiotics may be required. Mild reversible renal impairment (serum creatinine <0.30 mmol/l (3.4 mg/dl)) is usual but severe acute renal failure (ARF)

requiring dialysis can occur. It is recommended that any patient who presents after near-drowning or immersion should be assessed for potential ARF by serial estimations of serum creatinine, particularly when there is an increase in the initial serum creatinine, marked metabolic acidosis, an abnormal urinalysis, or marked lymphocytosis.

Prevention of drowning

Swimming pools and natural bodies of water present the greatest risk to young children. Preventive measures include public educational campaigns in the media, education and supervision by the parents, training in cardiopulmonary resuscitation, and better safety standards and safety devices, including pool fencing. The number of pool drownings in Brisbane (Australia) decreased after legislation made pool fencing compulsory. Strategies for the prevention of drowning should also consider hazards in rural areas.

The swimming and safety skills of young children may be improved by training. Education of the public is essential—in Australia only 17 per cent of surf rescues and resuscitations occurred within patrolled areas (up to 95 per cent successful resuscitation), while 55 per cent were saved and resuscitated outside patrolled areas (62 per cent successful); resuscitation success rates fell with increasing distance from patrolled areas. Some 5 per cent of all non-boating drownings in Australia were overseas tourists; 89 per cent of these were drowned in the ocean.

Eighty-nine per cent of children aged 35 to 59 months and 6 per cent of those younger than 3 years of age are sometimes bathed without adult supervision. An adult should supervise infants aged under 3 and all epileptic children in the bath.

Deaths from drowning in accidents involving boating and personalized water craft can be prevented by using lifejackets (personal flotation devices), but as many as 50 per cent of boaters do not use them. Efforts to increase their use should target adolescents, adults, and boating enthusiasts, especially those using motor boats. Specific measures tailored to prevent drowning associated with vessels capsizing and sinking in Alaska's commercial fishing industry have been successful.

More men than women drown in most age groups, probably reflecting an overestimation of their abilities, and more alcohol consumption. Middle-aged men dominate the group who die of cardiac events (mostly on the surface). Those who die of breath-holding hypoxia tend to be young males. Hyperventilation to increase breath-hold time is a dangerous practice that should be discouraged. Drownings are rare at supervised water parks, probably because of the large number of lifeguards on duty.

Further reading

Cummins P, Quan L. (1999). Trends in unintentional drowning. The role of alcohol and medical care. *Journal of the American Medical Association*, **281**: 2198–202.

DeNicola LK, Falk JL, Swanson ME, Gayle MO, Kissoon N. (1997). Submersion injuries in children and adults. *Critical Care Clinics*, **13**: 477–502.

Manolios N, Mackie I. (1988). Drowning and near-drowning on Australian beaches patrolled by lifesavers: a 10-year study, 1973–1983. *Medical Journal of Australia*, **148**:165–171.

Modell JH, Layon AJ. (1999). Drowning and near-drowning. In: Lungdren CEG, Miller JN, eds. *The lung at depth*. New York: Marcel Dekker Inc., 395–417.

Orlowski JP. (1987). Drowning, near-drowning, and ice-water submersion. *Pediatric Clinics of North America*, **34**: 75–92.

Pearn J. (1987). Drowning and near drowning. Paediatric emergencies. In: Black JA, ed. *Medical aspects of drowning in children*, 2nd edn. London: Butterworths.

Spicer ST, Quinn D, Nyi NN, Nankivell BJ, Hayes JM, Savdie E. (1999). Acute renal impairment after immersion and near-drowning. *Journal of the American Society of Nephrology*, **10**: 382–6.

8.5.4 Diseases of high terrestrial altitudes

D. Rennie

High altitude terrain and populations

Until the late nineteenth century, mountains were viewed by Europeans as dangerous, mysterious, hostile, and remote. Yet there is evidence that mountainous regions have, for many thousands of years, been the home of large and elaborate civilizations such as that of the Incas, which in the fifteenth century included Ecuador to the north and much of northern Chile and Argentina some 5000 km to the south, an empire of at least 12 million people. The Altiplano or high plateau of the Andes is still home to millions of descendants of these Incas, many of whom have never been below altitudes around 4000 m above sea level.

Temperature tends to determine the fauna and flora. Since, other things being equal, temperature falls with increasing altitude, the high altitude climate tends to be an arctic one. The snow line and tree line become lower with increasing distance from the equator and to live, work, hunt, and cultivate at altitudes above 3000 m is possible only within about 40° of the equator. This includes the Andes of Ecuador, Bolivia, Peru, and northern Chile, the Rocky Mountains in the United States, the high lands of east Africa, the Caucasus, the Pamirs, and the Himalayas, but does not include, for example, the European Alps.

The fall in temperature of some 1°C for every 150 m rise in altitude, irrespective of latitude, and the high winds increase the danger of cold injury. The low humidity contributes greatly to fluid loss and dehydration, as does the increased solar radiation, which may be very much exaggerated by reflection from the snow. These factors are, however, common to most arctic environments and are discussed elsewhere.

The fact that so many people of diverse races have been born and have lived at such altitudes, and the fact that, excluding Antarctica, about 2.5 per cent of the land lies above 3000 m, gives high altitude physiology and medicine an economic, political, and cultural relevance. As modern transport brings the highest mountains within range of the meanest purse, tourists, hikers, mountaineers, and downhill and cross-country skiers, as well as mining engineers, geologists, and surveyors, are flocking up into the hills. Every day in July and August about 3600 people visit the summit of Pike's Peak in Colorado (4300 m) and there is now even a 42-km marathon race up and down that mountain. In 1950, three expert Western climbers first reached the base of the ice fall below Mount Everest's Western Cwm, at 5300 m. Twenty-five years later, in a mere 4 weeks, well over 500 inexpert tourists did so.

The vast majority of mankind lives below 1000 m altitude and, though there is an exponential rise in the numbers of lowlanders going to high altitude, they cannot assume that the ascent can be made with impunity.

Hypoxia

Though the proportion of oxygen in the air (20.93 per cent) is the same at every altitude, the atmospheric pressure decreases with increasing altitude. The total atmospheric pressure at sea level (barometric pressure, PB varies but is usually around 760 torr (mmHg) (101 kPa) and that due to oxygen (the partial pressure of oxygen) is 20.93 per cent of this, that is 159 torr (21.2 kPa). In the lung, the air is rapidly saturated with water vapour at body temperature. At any altitude this is 47 torr (6.25 kPa) at 37°C. The actual combined pressures of gas taken into the lungs is therefore PB– 47 torr and the inspired oxygen tension ($P\mathrm{IO}_2$) is 20.93 per cent of this: 0.2093 × (PB–47). At sea level this is 0.2093 × (760–47) = 149 torr (19.8 kPa). At about 5500 m the atmospheric pressure is about half that at sea level and at the summit of Everest (8848 m) the atmospheric pressure is about 250 torr (33 kPa; one-third that at sea level) and the partial pressure of oxygen is 0.2093 × 250 = 52.3 torr (7 kPa). The partial pressure of water, however, reduces the $P\mathrm{IO}_2$ from 52.5 torr to 0.2093 × (250–47) = 42.5 torr (5.7 kPa). It is clear that the fraction of total inspired gas pressure due to water vapour, which at sea level is 6 per cent, increases with altitude. At the summit of Everest it will be nearly 19 per cent and at 19 200 m (63 000 feet), the total pressure of inspired gases would be a fatal 47 torr—fatal because all of it would be water vapour. Conversely, the proportion of inspired gas due to oxygen, which is 19.6 per cent at sea level, is reduced to 17 per cent at the summit of Everest.

Following a plane's sudden decompression or a rapid balloon ascent, for example, a resting man, just up from sea level, would lose consciousness in a matter of minutes at altitudes between 6400 and 7300 m ($P\mathrm{AO}_2$ = 24–15 torr; 3.2–2 kPa) and in seconds above 7300 m ($P\mathrm{AO}_2$ below 15 torr; 2 kPa), yet during 1 year (1978–9) a total of 14 men, on the three highest peaks in the world (Everest, K2, and Kanchenjunga) were not only fully conscious at rest more than 1000 m higher than this but were able to climb over difficult terrain in bad weather to their respective summits (8848 m, 8611 m, and 8598 m) all without the benefit of supplemental oxygen.

Acclimatization

The difference between the ineffectual newcomer on arrival at high altitudes and the resident is the sum of a myriad of physiological adjustments called 'acclimatization'. This seems to depend solely upon the length of exposure and the age when first exposed and has little to do with genetic factors. The time of greatest adjustment is in the hours and days after arrival at a higher altitude, but changes may continue for years.

The alterations in response to hypoxia affect every tissue. They may be summarized as a series of adjustments which boost oxygen supply to the mitochondria by keeping the partial pressure of oxygen in the tissue capillaries as high as possible, by decreasing the distance oxygen has to diffuse in the tissues, and by increasing the concentrations of respiratory enzymes. A few of the principal steps involved may be summarized as follows.

Ventilation

When $P\mathrm{AO}_2$ has fallen to 55 to 60 torr (7.3–7.9 kPa) at an altitude of about 2000 to 3000 m, the peripheral chemoreceptors are stimulated by hypoxia and ventilation is increased. This initial reaction is amplified over the next 3 or 4 days so that $P\mathrm{AO}_2$ levels which are only very slightly lower than normal (as one would find at, say, 1000 m altitude) now begin to stimulate ventilation. After a few weeks of sojourn, however, ventilation slowly decreases and this process continues for years, though ventilation is always higher in sojourners than in people who were born and have lived all their lives at altitude. With exercise, ventilation increases more at high altitude than at sea level and more in sojourners than in natives.

Driving respiration by hypoxia implies, from the point of view of carbon dioxide (CO_2), 'hyperventilation', since CO_2 is blown off and a respiratory alkalosis develops. The $P\mathrm{CO}_2$ falls in a linear manner with altitude, and the alkalosis is only partly compensated by a rise in urinary excretion of bicarbonate. The effect of the increased ventilation is rapidly to increase the alveolar oxygen pressure by reducing the oxygen gradient between ambient and alveolar air and by reducing $P\mathrm{ACO}_2$.

Pulmonary diffusion

Though no increase in pulmonary diffusing capacity occurs in sojourners at high altitude, natives have increased pulmonary diffusing capacity with a lowered alveolar–arterial oxygen gradient, associated with an increased capillary surface for diffusion. This is due to an opening up of pulmonary capillaries, and to polycythaemia which decreases the distance necessary for gaseous diffusion. In the newcomer to high altitude, exercise is accompanied by a marked fall in arterial oxygen saturation, in contrast to the unchanged values on exercise at sea level, and this may be partly due to a limitation in diffusion.

Circulation

Though there is an abrupt increase in cardiac output on ascent to high altitude, there follows a progressive decrease in stroke volume and maximal cardiac output is reduced at all levels of exercise including maximal exercise. There is no evidence for insufficient myocardial oxygenation and there is argument about whether or not the myocardium is actually depressed by the hypoxia. There is an immediate alteration in the distribution of blood flow. For example coronary and cutaneous flow both fall, cerebral and retinal flow increase, renal flow temporarily decreases, and then, with acclimatization, returns to normal.

The oxygen-carrying ability of the blood is considerably increased by the massive increase in red cell production, in total red cell mass, and, more importantly, in tissue capillarity. It is somewhat offset by the higher haematocrit which increases the blood viscosity and decreases the rate of flow. The shift in the oxy–haemoglobin dissociation curve to the right, which occurs on ascent and is due to an increase in red cell 2,3-diphosphoglycerate, favours unloading of oxygen to the tissues, but this particular adjustment is offset by the shift to the left caused by alkalosis.

Tissue adaptations

Apart from the very major role of increased tissue capillarity which reduces the average capillary–mitochondrial distance and so dramatically reduces the distance for diffusion, increased myoglobin facilitates oxygen diffusion and there is an increase both in mitochondrial density and in many enzymes of the respiratory pathway (e.g. in cytochrome oxidase).

Oxygen uptake

Though the oxygen uptake at rest is not diminished even at very high altitudes, above an altitude of 1500 m the maximal oxygen uptake, $\dot{V}O_{2max}$, falls about 10 per cent for each gain in altitude of 1000 m between 1500 and 6700 m and though it is improved by administration of pure oxygen, it does not return to normal until several days after descent. The cause of this drop in $\dot{V}O_{2max}$ has been debated. If, at any one altitude, the work load is increased, $\dot{V}O_2$ may reach a maximum but ventilation, already increased at high altitude, is still able to increase further and so does not seem to be the factor limiting $\dot{V}O_{2max}$.

At altitude, as opposed to at sea level, arterial oxygen saturation falls with increasing exertion, but at moderate altitudes rises again near maximal exertion. There is probably no increase in alveolar–arterial oxygen gradient and, because of the rise in haemoglobin, the amount of oxygen carried is kept up. Diffusion may be a little limited but diffusion and the blood's oxygen carrying capacity are probably not factors limiting $\dot{V}O_{2max}$. After a few days at high altitude, however, maximal heart rate and, particularly, maximal cardiac stroke volume are reduced, the cause being unclear, and so the inability of the heart to go on increasing cardiac output seems to be the reason for the fact that $\dot{V}O_{2max}$ declines progressively with increases in altitude.

Extreme altitudes

At extreme values, where every increment in height results in a precipitous fall in $\dot{V}O_{2max}$, the oxygen cost of the work of ventilation rises considerably and it assumes an even greater proportion of total oxygen cost when $\dot{V}O_{2max}$ is reduced to really low levels. Moreover, the maximal ventilation itself is reduced at such altitudes and, in addition, there is a diffusion defect within the lungs which becomes very marked on exercise. This, together with the steady decline in maximal cardiac output, causes maximal oxygen uptake to fall very precipitously at 8848 m where PB is around 250 torr almost to resting or basal levels of oxygen uptake, a figure of about 5 m/kg per min. The state of a climber at 8000 m who is comfortable resting in his tent but whose $\dot{V}O_{2max}$ is reached when he puts on his boots, for example, is perilous.

Alveolar gas samples taken on the summit of Everest (8848 m; barometric pressure 253 torr, 33.4 kPa) show an inspired PO_2 = 43 torr (5.68 kPa), and alveolar PO_2 = 35 torr (4.62 kPa). Estimated arterial gas values are: PO_2 = 28 torr (3.7 kPa); PCO_2 = 7.5 (0.99 kPa); pH = greater than 7.7; oxygen saturation = 70 per cent. A number of reports have suggested mild, possibly permanent, defects in cognition in climbers who have ascended to extreme high altitudes.

Illness due to altitude

Acute mountain sickness

For centuries it has been known that when lowland dwellers climb mountains some of them become ill. The illness, usually called acute mountain sickness, generally begins after a few hours, and is characterized by non-specific symptoms such as headache and vomiting. In the vast majority of people it is transient and trivial but in a few becomes progressive, severe, and may be fatal. It is to some extent relieved by breathing oxygen and it is cured by descent to sea level. Its cause is unknown.

Acute mountain sickness may be more than a temporary annoyance. It is of concern because it is probably a manifestation of mild, transitory cerebral oedema, and in a small proportion of cases this oedema may worsen and become clinically overt (high altitude cerebral oedema). Acute mountain sickness is also associated with, and made much worse by, the increased hypoxia consequent upon high altitude pulmonary oedema.

Symptoms and signs

Symptoms come on between 8 and 96 h after ascent and include, after an initial euphoria, lethargy, headache, fitful sleep with arousals, an increasingly severe headache, often occipital, nausea, vomiting, dizziness, and loss of balance. The sufferer lies groaning in his sleeping bag, refusing food, and dozing. After a day or two of rest, the symptoms disappear and are soon forgotten, yet they may recur on further ascent to a greater altitude.

The signs are few: an irritable, depressed, but usually fit person, often, because he is starved, smelling of ketones and vomit, and holding his aching head. The most useful diagnostic sign is ataxia. Occasionally the patient, especially if vomiting and taking diuretics, may be too dehydrated and hypotensive to stand up. At this stage crackles may be heard in more than a quarter of people examined.

In a few, the symptoms rapidly worsen with the onset of pulmonary oedema. Soon the breathlessness, even at rest, is extreme and may be accompanied by a dry cough. The patient, anxious and sometimes incoherent, becomes progressively more dyspnoeic and sometimes orthopnoeic. He is very cyanosed, with a mild pyrexia (38.3°C), a pronounced tachycardia, and sometimes mild hypotension. There are no signs of cardiac failure, but loud crackles can easily be heard all over the chest and frothy sputum, sometimes tinged with blood, wells out of the mouth and nose.

Pulmonary oedema, if very severe, decreases oxygenation and so tends to be accompanied by high altitude cerebral oedema but each syndrome may occur independently as features of acute mountain sickness. In cerebral oedema without pulmonary manifestations, the patient, having had progressively worsening symptoms of acute mountain sickness for several days, becomes incoherent, hallucinated, and too ataxic and drowsy to stand up or look after himself. He may be unable to get out of his tent, or he may do so and then fall into a snow drift and lie there. Soon he is stuporose and snoring stertorously. His sleeping bag may be wet with urine and in a few hours he is in deepening coma. There are rarely any localizing signs; mild bilateral papilloedema is characteristic.

Predisposing factors

Males and females are equally likely to develop the acute mountain sickness syndrome, and the incidence is inversely related to age. There is no good evidence that the incidence of acute mountain sickness relates to prior physical fitness as expressed by, say, maximal oxygen consumption, and much anecdotal evidence from

many experienced observers is that there is no such relationship. We now know, however, that statistically-speaking, the higher the vital capacity, the less the chance of acute mountain sickness and there is increasing evidence that susceptibility to acute mountain sickness is correlated with a poor ventilatory response to hypoxia, measured at sea level. Many other physiological functions have been measured at sea level but none usefully predict whether any one individual will develop acute mountain sickness. Factors related statistically to the occurrence of acute mountain sickness are severe exertion, pre-existing respiratory infection, starting an ascent at a higher altitude; previous acute mountain sickness, especially pulmonary oedema; and having a low vital capacity and a poor ventilatory response to hypoxia. Healthy, high-altitude dwellers seem to be at unusually high risk of pulmonary oedema after brief descent to lower altitudes followed by reascent.

Pathophysiology

On ascent to high altitudes, a great many factors affect cerebral oxygenation: an increase in the rate and depth of breathing; a fall in plasma volume and increase in haemoglobin; an increase in cerebral blood flow due to hypoxic cerebral vasodilatation (to some extent off set by hypocapnic cerebral vasoconstriction); and increases in brain oxygen extraction. The net result is that cerebral oxygen delivery and metabolism are both normal at the sort of altitudes most climbers are likely to experience. But this is not to say that all brain function is normal. There is now good evidence, for example, that at levels of hypoxia frequently met by climbers the synthesis of neurotransmitters such as serotonin is diminished, and this may explain the changes in mood and reversible defects in cognition that have been reported.

Patients with acute mountain sickness, whether this is due to failed or exaggerated physiological adjustments to hypoxia, tend to have ventilation that is inappropriately low for the altitude. They thus have lower blood oxygen and higher PCO_2 levels. Pulmonary oedema causes a vicious cycle of deepening hypoxia.

People with acute mountain sickness also have an antidiuresis and retain fluid. The skull is a rigid box, so small increases in volume lead to large increases in pressure, and recent work suggests that those people who have relatively large ratios of cerebrospinal fluid to brain volume may be able to displace the cerebrospinal fluid more easily and so be protected. There is movement of fluid into the cells, and considerable hypoxic cerebral vasodilatation. As Hackett has argued, neither is sufficient to cause high altitude cerebral oedema. The finding, using MRI, of oedema of the white matter in cerebral oedema of high altitude lends support to the concept that in acute mountain sickness there is vasogenic cerebral oedema. This is an oedema due to movement of fluid and proteins from inside the blood vessels across the blood–brain barrier, perhaps due to a failure of vascular autoregulation and transmittal of increased pressures to the capillaries. This fits with the observation that steroids, which are useful in high altitude cerebral oedema and in prevention of acute mountain sickness, are effective only in the vasogenic sort of cerebral oedema. Hypoxia stimulates the formation of inducible nitric oxide synthase (iNOS) which increases the permeability of the blood–brain barrier, and iNOS may be the cause of the oedema where cerebral capillary pressure is already raised. Autopsies in people who have died from cerebral oedema have shown, besides oedema, numerous tiny haemorrhages from the capillaries.

Moderate acute mountain sickness is associated with oedema of the brain, but exposure to high altitude causes increased brain volume (increased brain water, or blood volume, or both) in men whether they have acute mountain sickness or not. The reason why some individuals are susceptible to developing symptoms is not known, but there is increasing interest in the idea that susceptibility might depend on the ability of an individual's craniospinal anatomy to accommodate volume increases. Since the mean volume needed to raise intracranial pressure 10-fold is 26 ml for an adult, small individual differences in compliance on the part of the intraspinal space might explain differences in susceptibility. Evidence that mild acute mountain sickness is due to cerebral oedema is as yet lacking, though this is likely.

It is known that on ascent to high altitude there is a shift in blood from the systemic 'capacitance vessels' (veins) to the pulmonary circuit. Exercise and cold increase this shift, while phentolamine, an a-adrenergic blocker, decreases pulmonary vascular resistance and improves gas exchange in high altitude pulmonary oedema. In addition, there is a greater than normal hypoxic pulmonary arterial vasoconstriction with a rise in pulmonary artery pressure. Nifedipine, which reduces pulmonary arterial pressure, has been shown to be effective in both the prevention and treatment of high altitude pulmonary oedema. Neither of these processes causes the pulmonary oedema but both contribute to it. Cardiac function is not impaired. Pulmonary wedge pressures, reflecting average pulmonary venous pressures, are normal, as are atrial pressures.

The oedema fluid of people suffering from high altitude pulmonary oedema has been shown, using bronchopulmonary lavage, to be high in protein content, which is characteristic of a breakdown in vascular permeability, and not like that of 'high-pressure' cardiogenic pulmonary oedema. Recent work, however, suggests a more complex picture. Albumin escape rates into the alveolae may not be sufficiently increased in high altitude pulmonary oedema, and the increased alveolar protein might be due to water reabsorption from the alveolae. The oedema fluid also contains large numbers of cells, the vast majority of them macrophages as well as markers of chemotactic activity and inflammation. Various mediators of inflammation, such as E-selectin, are raised early in high altitude pulmonary oedema. This accords with the observation that upper respiratory infections seem to predispose to high altitude pulmonary oedema, but recent studies suggest that the high permeability leak is non-inflammatory.

Cases of high altitude pulmonary oedema after minimal exertion at modest altitudes, in people with congenital absence of the right pulmonary artery, suggest that pulmonary hypertension and shearing forces due to a massively increased circulation (in these cases through one lung) might together contribute to high altitude pulmonary oedema. It seems that in high altitude pulmonary oedema some lung capillary beds are closed off, whether by hypoxic vasoconstriction or by capillary wall oedema and thrombosis. The ventilation/perfusion mismatch increases hypoxia and there is diversion of the whole circulation through the remaining, widely patent capillaries, which are damaged both by the increased pressure and the shear stress of the large flow. As a consequence, they leak in a patchy fashion, causing the typical patchy appearance on radiographs of high altitude pulmonary oedema. Nitric oxide improves high altitude pulmonary oedema by reducing this ventilation/perfusion mismatch.

Incidence

Lacking rigorous, large-scale studies, we can say that acute mountain sickness has been reported below 1550 m, is uncommon below 3000 m, and that the higher the altitude, the faster the ascent, and the greater the exertion the higher the incidence (from fewer than 10 per cent to 60 per cent), and the worse the illness.

Most large studies suggest that some symptoms, for example a bad headache that is not relieved by 600 mg of aspirin, occur in over half of lowlanders going to above 4000 m and that over 10 per cent of climbers have more than one severe symptom at that altitude. High altitude pulmonary oedema occurs in about 5 per cent of people at that altitude, but symptomless crackles in the lung in almost a quarter. Cerebral oedema, which takes a few days to develop, is less common, occurring in perhaps 0.5 to 2 per cent.

Diagnosis

The differential diagnosis is small. For mild acute mountain sickness the non-specific symptoms mean that hypothermia, exhaustion, and dehydration should be considered. The effects of alcohol and marijuana may duplicate those of acute mountain sickness. For pulmonary and cerebral oedema, only infectious diseases (pneumonia or meningoencephalitis) are likely possibilities, though carbon monoxide poisoning due to poorly ventilated tents occurs. Periodic breathing with apnoeic phases is normal at high altitude and not by itself a cause for alarm, though it may be very marked in cerebral oedema. Fever is moderate in high altitude pulmonary oedema, which is accompanied by extreme tachycardia and cyanosis, without purulent sputum. In cerebral oedema there is no meningism and little if any pyrexia, nor evidence of ear infection or of localizing signs. There is a striking absence of any signs of cardiac decompensation (cardiac enlargement, raised neck veins). Giving oxygen—which is expensive, often not available, and has to be given in 4 to 5 l/min amounts—frequently produces equivocal results as everyone at high altitude, with almost any illness or none at all, feels temporarily better on oxygen.

All illnesses improve on descent, but the very dramatic recovery on descent in acute mountain sickness, and particularly in pulmonary and cerebral oedema, is so striking as to be diagnostic in itself. Numerous cases have been described of people unconscious and worsening at 4000 m who were fully conscious at 3000 m.

The physician who sees the patient on descent is, however, usually obliged to perform a battery of tests on the patient who, except in the case of very prolonged cerebral oedema, is recovering fast. The chest radiograph will show scattered fluffy patches of oedema; the electrocardiogram shows simple right ventricular hypertrophy and 'strain', and the cerebrospinal fluid will be normal.

Prophylaxis

Acute mountain sickness and its complications are the consequence of ascent that for any one individual is too high or too fast and are therefore easily and completely preventable by going up slowly. Everyone going to high altitudes has the responsibility to set a schedule for themselves that includes frequent rests, and

that allows them to descend at will. Acute mountain sickness affects the capacity to reason, so everyone must also take responsibility for their fellows, realizing that pulmonary or cerebral oedema can be rapidly fatal and that it is often simple to walk a sufferer down 1000 m, whereas a few hours delay can endanger the patient's life and that of scores of rescuers. Even now, communications are frequently faulty, planes and helicopters may not exist or may crash, or be unable to get near. Energetic propaganda advising slow ascent and rest days and nights has been shown to lower the incidence of acute mountain sickness.

Since insensible water loss is increased due to the increased exertion, the extra ventilation and the dry air, every effort must be made to avoid dehydration, by drinking enough to urinate clear, dilute urine. The climber should also eat plenty, especially carbohydrates. There is evidence that sleeping tablets, which further lower the already lowered oxygen saturation during sleep should be avoided.

Acetazolamide, 5 mg/kg per day, reduces the incidence and severity of acute mountain sickness when taken before and during ascent. It probably works by decreasing cerebrospinal fluid formation and, more importantly, stimulating nocturnal respiration by causing an intracellular acidosis in the respiratory centre so that periodic episodes of apnoea, which are very common at altitude, are abolished and blood oxygen saturation is never allowed to fall. Its principal side-effect is tingling in the hands, feet, and around the mouth.

Dexamethasone, (2 mg every 6 h or 4 mg every 12 h) a glucocorticoid effective in the management of cerebral oedema, has been shown to prevent acute mountain sickness, while reducing the retinal arterial dilation that follows exposure to hypoxia.

Treatment

Rest, aspirin for headache, and a descent of 500 m or so for a couple of nights usually suffice, though ataxia may require the patient be assisted down further. Acetazolamide (250 mg, 8-hourly), started 12h before ascent, increases oxygenation during sleep and relieves symptoms. Dexamethasone (4 mg 6-hourly) is highly effective. Diuretics are logical, given the water retention. The presence of pulmonary oedema or coma necessitates urgent removal down at least 1000 m, and careful observation. In addition to descent, oxygen, should be administered (if necessary through an endotracheal tube at a rate of 5–6 l/min). Immediate descent by even a small amount may save the lives of such critically ill patients, and reliance on air evacuation is dangerous.

In pulmonary oedema, sit the patient up if this makes him comfortable and give frusemide (20–40 mg orally) and morphine (15 mg intravenously). Both these drugs divert blood from the pulmonary to the systemic circuit. Frusemide causes a diuresis and morphine decreases anxiety and respiration rate.

The calcium blocker, nifedipine, prevents high altitude pulmonary oedema in susceptible subjects, improves oxygenation, and reduces the excessive rise in pulmonary artery pressure in pulmonary oedema. It works by improving the matching of perfusion to ventilation, by lessening the formation of oedema fluid, and by decreasing the inflammatory response. Its side-effects (hypotension and tachycardia) suggest that its use should be limited to established oedema. Further studies may confirm that inhaled nitric oxide is useful.

A portable, light-weight hyperbaric bag, made of impermeable nylon, has been developed. The victim lies inside while the air pressure is increased using a foot pump, by 104 torr. For example at 4200 m, where ambient barometric pressure is 449 torr, the pressure in the bag, 553 torr, would be equivalent to an altitude of 2544 m. This is therefore equivalent to taking the patient down 1656 m. Controlled clinical trials have shown the bag to be effective in acute mountain sickness and mild pulmonary oedema, and expeditions should consider taking such bags with them, though descent is still the best, quickest, and usually the safest therapy.

In cerebral oedema, intravenous dexamethasone in large doses (e.g. 8 mg) should be given to reduce the oedema. A cuffed endotracheal tube should be passed, the lungs kept inflated and the bladder kept empty with a Foley catheter (all these devices should form part of the kit of a responsible physician travelling at high altitudes).

Prognosis

This depends not merely on the altitude and terrain but on the speed with which early signs are noticed by the patient's colleagues and their determination and skill in evacuating him to lower altitudes immediately. Once pulmonary oedema and to the rarer but more dangerous cerebral oedema have developed, the patient probably has about a one-third chance of dying. If he lives, full recovery is the rule.

Retinal haemorrhage of high altitude

On ascent, there a considerable increase in cerebral and retinal blood flow. Ophthalmoscopy around 5000 to 6000 m has shown a 20 to 25 per cent increase in diameter of both arteries and veins, and striking increase in the tortuosity of these vessels (Fig. 1). Suffusion of the optic disc with blood due to capillary dilatation may delude the examiner into diagnosing papilloedema.

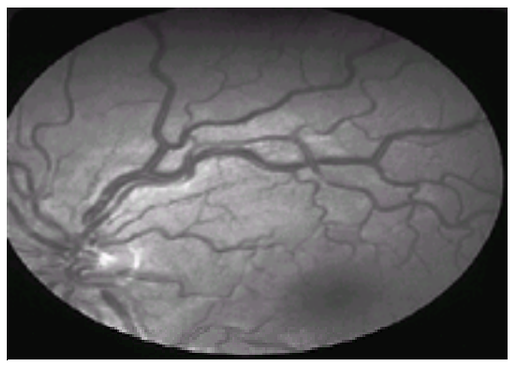

Fig. 1 The left optic fundus of a 26-year-old male climber photographed at 5900 m altitude. Both the veins and arteries are dilated and tortuous and there is hyperaemia of the optic disc as well as mild papilloedema. The climber had severe headache, nausea, and vomiting.

Studies have shown that one-third to one-half of symptomless climbers descending from between 7000 and 8000 m down to 5000 m altitude develop flame-shaped retinal haemorrhages (Fig. 2). The haemorrhages are usually near the optic disc, may be of any size, and are usually unnoticed and resolve on descent, unless they are at the fovea, where they may leave a permanent blind spot.

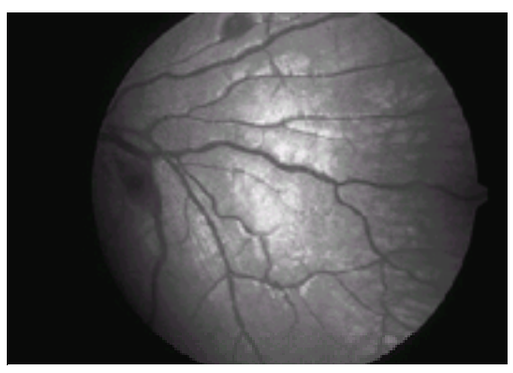

Fig. 2 The left optic fundus of a fit and symptomless 28-year-old male climber photographed at 5900 m altitude. The vessels are dilated, as in Fig. 1, but are less

tortuous. Retinal haemorrhages are shown.

They may be caused by transmission of high thoracic pressures developed during straining or prolonged coughing, and transmitted through a vasculature dilated because of hypoxia. It is probable that the finding of retinal haemorrhages in someone at high altitude is not by itself a reason to counsel descent.

Peripheral oedema

Swelling of hands, face, and ankles may occur in climbers on ascent to high altitude, just as on long-continued hiking at sea-level, and rarely there may be gross anasarca. The oedema is more common in women than in men though no association with menses has been found. It is associated with acute mountain sickness and pulmonary oedema but may occur by itself, and it is relieved by the diuresis that accompanies descent.

Other illnesses of high altitudes

People frequently ask their physicians whether their hearts will be all right at high altitude. The physiological facts are that the normal heart is not limited by heights up to the top of Everest. Unhappily, the doctor has no tests that are sensitive or specific enough to give an accurate prediction, unless the patient has symptoms. The doctor has to remind the prospective trekker that he or she is going on an adventure, and getting away from, among other things, good emergency medical treatment. The physician should be very hesitant to forbid any activity which may have great meaning to the patient, in the absence of actual, as opposed to theoretical, evidence that it is harmful.

On theoretical grounds, people with diseases which limit ventilation, diffusion of oxygen, circulation, and tissue adaptation, will fare badly at high altitudes. Often, however, there is little evidence and doctors should be cautious about the constraints they place on a patient's activities and frank about our ignorance.

Chronic mountain sickness (Monge's disease)

This is a clinical syndrome affecting a few very long-term residents at altitudes, and equivalent to the alveolar hypoventilation syndrome seen at sea level. It is to be distinguished from the persistent failure of a newcomer to adjust to the altitude which results in weeks or even months of acute mountain sickness. It usually occurs in men between the ages of 20 and 50, and the symptoms are of headache, dizziness, depression, irritability, and, most strikingly, drowsiness and even episodes of coma. The signs are those of polycythaemia that is excessive for the altitude (in the high 70s) and cyanosis with suffused, congested conjunctiva, ear lobes, cheeks, and lips, as well as clubbed finger nails. Signs of congestive heart failure may be present. Poor ventilation is reflected by a low oxygen saturation and arterial $P\mathrm{CO}_2$. The right heart may be enlarged on chest radiograph and there is evidence of right ventricular hypertrophy on ECG, itself a reflection of marked pulmonary hypertension. All the symptoms, signs, and physiological abnormalities are cured by descent to sea level, and sufferers should move there permanently.

Myocardial infarction

All idigenous populations studied at high altitudes have, compared with people in the West, low rates of coronary disease, perhaps because they exercise more, eat less (especially less salt), have lower serum lipids and have less hypertension. In the United States, the decline in mortality from arteriosclerotic disease with increasing altitude may be due to migration down to lower altitudes of people liable to coronary disease.

The physician at sea level is often asked by people whether it is safe for them to go up to 3000 or 4000 m altitude to ski or to climb. Logic dictates that as the myocardium has no oxygen reserves, as myocardial oxygen extraction even at sea level leaves no room for improvement, and as coronary flow is decreased, people who have poor coronary circulation will be at great risk when they ascend from sea level and start hard exercise. There are, however, no credible studies, with useful denominators, to guide the physician. A prudent physician might tell the patient not to ascend until he is fit to take exercise at sea level, but unless the patient has had a previous myocardial infarction, we have no evidence that an electrocardiogram, with or without exertion or breathing low oxygen mixtures simultaneously, has any predictive value at all, and the patients' previous history of fitness and of illness is likely to be far more useful.

Pulmonary emboli

Deep vein thrombosis tends to occur in climbers at very high altitudes partly because the erythropoietic stimulus never shuts off above an altitude of about 5500 m so that the blood is very viscous and partly because of dehydration and enforced inactivity due to storms. Pulmonary emboli, sometimes fatal, occur. Evacuation to lower altitudes and the administration of aspirin as an anticoagulant should be tried.

Sickle-cell anaemia

Since cells containing HbS sickle, when hypoxic, become sticky and rigid, it is not surprising that homozygous cases of HbS (sickle-cell anaemia) are at great danger from high altitude. In Denver, at 1609 m, people who are heterozygotes (sickle-cell trait) apparently lead normal lives, but cases of splenic infarction in people with sickle-cell trait have been described in men at between 3500 and 4500 m. One should be cautious in advising anyone with sickle-cell trait to exercise at altitudes above about 2000 m.

Further reading

Grissom CK, Elstad MR (1999). The pathophysiology of high altitude pulmonary edema. *Wilderness and Environmental Medicine* **10**, 88–92.

Hackett PH (1999). The cerebral etiology of high-altitude cerebral edema and acute mountain sickness. *Wilderness and Environmental Medicine* **10**, 97–109.

Hackett PH, Roach RC (1994). High altitude medicine and physiology. In: Auerbach PS, ed. *Management of wilderness and environmental emergencies*, 4th edn. C.V. Mosby, St Louis.

Hackett PH, Roach RC (2001). High altitude illness. *New England Journal of Medicine* **345**, 107–14.

8.5.5 Aerospace medicine

D. M. Denison and M. Bagshaw

Introduction

Aviation medicine concerns the welfare of humans in flight through the Earth's atmosphere, an oxygen-rich gas that shields the ground below from solar radiation above. Held to the Earth by gravity, compressed under its own weight, the atmosphere is denser close to the ground than further away. Long waves of infrared light penetrate it easily but heat the ground below. Hot ground reradiates some of this heat at shorter wavelengths which are absorbed by carbon dioxide and water vapour, making the air close to the ground much warmer than that higher up. Short waves of ultraviolet sunlight, absorbed by oxygen molecules early in their journey, create a belt of warm ozone at high altitudes. Some rays intercepted in the same region generate secondary rays that extend lower down. Very few reach the ground. At sea level the atmosphere exerts a pressure of about 760 mmHg (101 kPa); it is variably moist, has a temperature that ranges from –60 °C to +60 °C, and moves at wind speeds from 0 to 160 km/h. With increasing altitude, the temperature, pressure, and water content of the atmosphere fall and wind speeds increase. Some of these features are summarized in Fig. 1. On ascent conditions generally become more severe and more uniform, so turbulence decreases. Most flights through the atmosphere last for a day or less, but some balloon flights persist for longer. The main medical problems of flight through the atmosphere are hypoxia, confining people to small spaces, exposing them to high accelerations, and getting them back to the ground safely should something go wrong.

Fig. 1 Some physical features of the Earth's atmosphere, showing the variations in barometric pressure, air temperature, and ozone concentration with altitude. (NB: There is an international aviation safety convention that all altitudes are given in feet.) The shaded diagram on the left illustrates how the Earth's atmosphere is compressed under its own weight. The atmosphere absorbs much solar radiation.

Space medicine concerns the welfare of humans in flight through vacuums remote from the Earth, the protection of its atmosphere, and the pull of its gravity. Here the problems are those of very prolonged flight times, preserving the pressure and composition of a self-contained gaseous environment, much greater radiation hazards, coping with weightlessness, and engineering the safety of extravehicular excursions that may last several hours.

Atmospheric pressure

Total gas pressure falls with altitude in a regular, almost exponential way, halving every 18 000 ft (5500 m), as in Fig. 2. The oxygen content of the atmosphere (20.93 per cent) is constant to very high altitudes, so the same curve can be used to obtain the ambient oxygen pressure by rescaling the ordinate (also shown in Fig. 2). The oxygen pressure of physiological importance is that which exists in ambient air when it is warmed and wetted on entering the bronchial tree. This process necessarily raises water vapour pressure to about 47 mm Hg, regardless of the total gas pressure outside. The oxygen pressure in moist inspired gas (PIO_2) fully saturated with water vapour at 37 °C is given by the relationship:

$$PIO_2 = FIO_2 (PB - 47)$$

where FIO_2, the fractional concentration of oxygen in the inspirate, is 0.2093 when air is inspired.

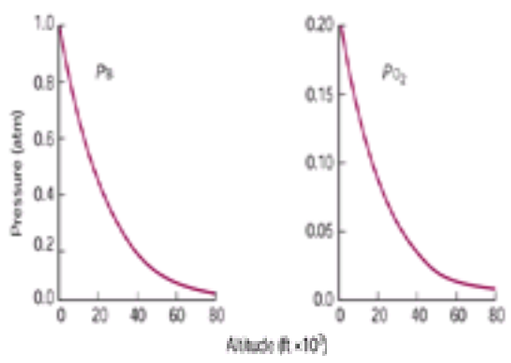

Fig. 2 The variations of barometric pressure (PB) and ambient oxygen pressure (PO_2) with altitude.

Atmospheric temperature

Atmospheric temperature drops more or less linearly with altitude, at about 2 °C/1000 ft (300 m), to the tropopause (40 000 ft (12 200 m)), is stable at –56 °C up to about 80 000 ft (24 400 m)and then rises to almost body temperature at about 150 000 ft (46 000 m), but by then air density is so low that its temperature is unimportant (cf. the dense cold air below).

Atmospheric ozone

Atmospheric ozone is formed by irradiation of diatomic oxygen molecules which dissociate into atoms. At very high altitudes the ultraviolet irradiation is so intense that all oxygen exists in the monatomic form. Lower down, some of the monatomic oxygen produced higher up combines with oxygen molecules to form the triatomic gas O_3 (ozone), at concentrations from 1 to 10 parts per million (**ppm**). The ozonosphere normally exists between 40 000 and 140 000 ft (12 200 and 42 700 m), i.e. from one-fifth to one-thirtieth of an atmosphere. Below 40 000 ft (12 200 m) the irradiation is normally too weak for significant amounts of ozone to form. Concentrations of 1 ppm at sea-level cause lung irritation. Ten times that concentration can cause fatal lung oedema. Although the ozonosphere is at much lower pressure, the ventilation systems of aircraft flying at very high altitudes can take in the ozone and compress it to partial pressures at which pulmonary irritation is a credible hazard.

Mechanical aspects

Propeller-driven aircraft need sufficient air to bite on but not enough to slow them down. They fly best at altitudes below 30 000 ft (9150 m), but if cruising above 10 000 ft (3000 m) they need pressurization to prevent cabin altitudes rising above about 8000 ft (2440 m). They use control surfaces and the resistance of the atmosphere to force changes in direction and ambient air to ventilate the pressurized cabins. Aircraft cabins are usually pressurized and ventilated with sufficient air from outside to dilute the carbon dioxide excreted by the occupants' lungs to an ambient partial pressure of 3 mmHg (0.4 kPa) or less. Many studies have shown that exposure to carbon dioxide at such levels increases lung ventilation slightly and appropriately, but has no other ill effects. Submariners have accepted for many years, with good evidence, a carbon dioxide partial pressure of 7 mmHg (0.9 kPa) as a safe upper limit for continuous exposure of healthy adults for 3 months. Because the ventilating air is dry, humidity in the cabin is low. This dries out the mouth and nose, which is marginally uncomfortable but has no other ill effects. Total water loss from this cause per person over a 10 h flight would be about 100 g.

Jet aircraft propel themselves by throwing a stream of hot gas behind them, but they need atmospheric oxygen to ignite the fuel that does this. They fly best at altitudes above 30 000 ft (9150 m) but below 65 000 ft (19 800 m). Generally, they use control surfaces and the resistance of the atmosphere to change direction and use ambient air to ventilate their pressurized cabins. However, some aircraft, such as 'jump jets', can also manoeuvre by changing their direction of thrust.

Rockets take an oxygen supply with them, and fly best in a vacuum. Space flights are commonly of long duration. There is no atmosphere for control surfaces to work against, or to be compressed to ventilate. Thus all navigation has to be achieved by directional motors and the pressure and composition of cabin atmospheres has to be maintained by on-board systems.

The atmosphere permits vehicles to travel through it at very high speeds. Usually these are achieved and lost so gradually that the changes of pace pass unnoticed, but so-called 'agile' aircraft can make continuous high-speed turns, sustained for a minute or more. The accelerations that such turns produce are proportional to the square of the vehicle's speed and inversely proportional to the radius of its turning circle (as shown in Fig. 3). Because the aircraft turns by applying its broad wing surface to the air, these accelerations are almost perpendicular to that surface and roughly parallel to the long axes of the people within. Most often, aircraft make 'head to the middle' turns, and blood and loosely tethered organs like the liver and heart fall towards the feet (positive '*g*' or $+G_z$ accelerations). Healthy unprotected adults, sitting erect in 'head to the middle' turns will 'black out' at about +3 *G*. Protective devices that raise intrathoracic pressure and apply counterpressure to the abdomen and legs during 'head to the middle' turns can maintain consciousness up to about +10 *G*. When aircraft make 'head-out' turns, (negative '*g*' or $–G_z$ accelerations), mobile organs and blood are forced towards the head, eventually leading to 'red-out'.

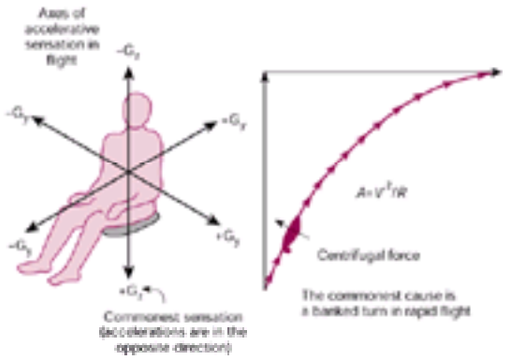

Fig. 3 The axes of acceleration in flight are labelled according to the sensations experienced by the aviator. Thus, when the pilot is accelerated upwards (+ G_z) the body fluids and tissues sare felt to sink towards the feet.

In brief, the important features of the atmosphere are that its temperature and pressure fall and radiation intensities rise with altitude. In addition there is a poisonous belt of ozone at high altitude. The atmosphere also permits vehicles to travel at high speeds and make sustained severe accelerations. The biochemical and mechanical hazards posed by these stresses are extremely challenging but on the whole they only affect professional aircrew. The main biochemical risk, shared by many patients at sea-level and all who fly, is hypoxia.

Biochemical hazards

Hypoxia

Oxygen has a dual role in most animal cells, as it is life-giving and extremely poisonous at the same time. In air, or dissolved in simple solution, it is benign and only ionized with difficulty. However, once an electron is successfully attached to an oxygen molecule it becomes a highly corrosive superoxide ion, forming a cascade of other very destructive oxygen radicals. This is an essential feature of oxygen toxicity, which is discussed in Chapter 8.5.6. Superoxide dismutase and various peroxidases have evolved to protect most cells from the effects of spontaneous formation of oxygen radicals by quenching the ions as rapidly as they appear.

More recently in evolution, other enzymes developed which harness this property of oxygen molecules in a controlled way. There are three sorts: oxidases, oxygenases, and hydroxylases. Quantitatively, cytochrome a_3 oxidase is the most important because, using oxygen as the ultimate electron sink, it allows many metabolic processes to proceed and at the same time unlocks and traps most of the energy the body needs (oxidative phosphorylation). However, several other oxidases that release this energy are unable to trap it. They are used to denature various unwanted products of metabolism.

Oxygenases introduce an oxygen molecule into organic molecules creating new compounds. Although these enzymes, of which there are many, only consume a small fraction of the body's total oxygen requirement, they are particularly important because they are responsible for production and dismemberment of many critical compounds such as the amine transmitters of the brain.

Hydroxylases insert one atom of oxygen and another of hydrogen into organic molecules. They too are responsible for many critical metabolic processes and for the denaturation of many drugs in the liver, kidney, and elsewhere.

These enzymes handle virtually all the oxygen uptake measured at the lips but differ in their affinity for oxygen, described by the Michaelis constant (for oxygen). This

constant (K_{mO2}) is that partial pressure of oxygen which, when all other factors are equal, just allows an oxygen-consuming reaction to proceed at half its maximum velocity. The major oxidase (cytochrome a_3), which is the cocatalyst of oxidative phosphorylation, has a very high oxygen affinity and thus a very low K_{mO2}, of 1 mmHg or less. That means this particular type of oxygen consumption, representing 80 to 90 per cent of the whole, can proceed full tilt down to very low levels of oxygen supply indeed. By contrast (Fig. 4), the other enzymes, which are quantitatively less important but qualitatively critical, have Michaelis constants for oxygen that vary from 5 to 250 mmHg. A fall in oxygen supply will influence these processes long before oxidative phosphorylation is affected and at times when overall oxygen consumption is diminished little if at all.

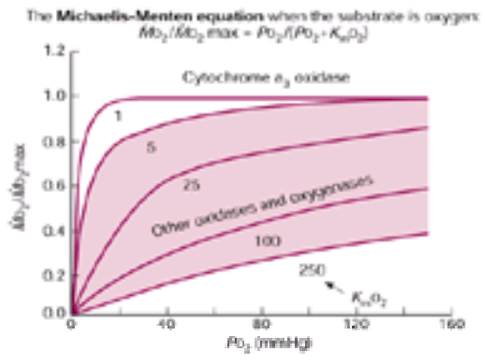

Fig. 4 Curves of oxygen uptake (O_2) as a fraction of the theoretical maximum (O_2max) against the partial pressure of oxygen (PO_2) for a family of oxygen-handling enzymes with Michaelis constants for oxygen (K_{mO2}) from 1 to 250 mmHg.

Although Fig. 2 describes how ambient oxygen pressure is related to altitude, it does not convey the measure of oxygen supply critical to humans, namely the pressure of oxygen to be found in the lungs. That pressure is determined by two equations (Fig. 5). The alveolar ventilation equation states that alveolar CO_2 pressure ($PACO_2$) depends only on CO_2 excretion (CO_2) and alveolar ventilation (Va), so:

$$PACO_2 = k(CO_2/V\text{a}).$$

The alveolar air equation states that since at any one time there is a fixed trading ratio between oxygen uptake and CO_2 excretion ($R = CO_2/O_2$), alveolar oxygen pressure (PAO_2) can be calculated from the moist inspired oxygen pressure (PIO_2*) and alveolar PCO_2, so:

$$PAO_2 = PIO_2^* - (PACO_2/R).$$

Progressive hypoxia leads to a mild hyperventilation (i.e. a rise in Va and fall in $PACO_2$). Knowing this it is possible to sketch a graph of alveolar oxygen pressure against altitude as in Fig. 6(a).

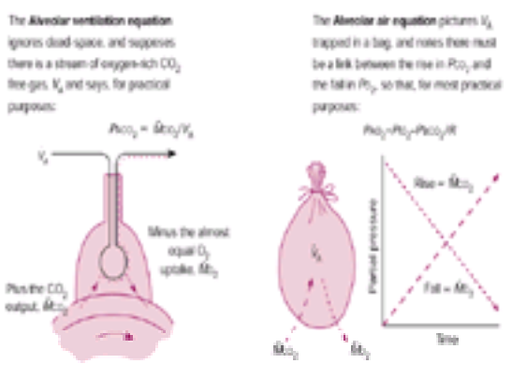

Fig. 5 Graphical representations of the alveolar ventilation and alveolar air equations.

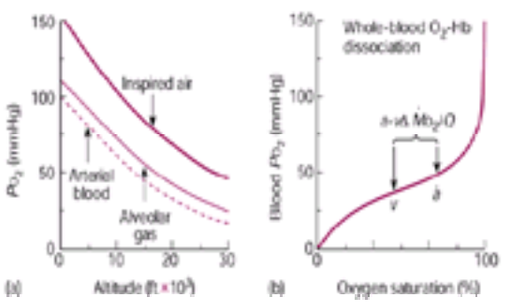

Fig. 6 (a) Variations in moist inspired, alveolar, and arterial oxygen pressure (PO_2) with altitude in normal men. (b) The conventional oxygen–haemoglobin dissociation curve of whole blood plotted to the same pressure scale as the left-hand graph, so that arterial O_2 content can be read directly (at the same horizontal level as the PO_2 curve. It also emphasizes that the arteriovenous oxygen content difference (a–v‡) is proportional to the ratio of oxygen uptake (MO_2) to local blood flow (Q).

When arterialized blood leaves a healthy lung it has an oxygen pressure some 10 mmHg less than that in the alveoli, due to uneven matching of ventilation to perfusion, some anatomical shunting, and an almost nominal obstacle to diffusion. In resting people, the alveolar–arterial oxygen gradient does not change much with altitude, although the relative importance of the factors contributing to it alter considerably; so subtracting a further 10 to 15 mmHg describes the relation between arterial oxygen pressure and altitude (also shown in Fig. 6).

The most important change is the loss of the head of pressure driving oxygen from the alveoli to blood, as the fall in alveolar PO_2 is much greater than that in mixed venous PO_2 (because of the shape of the oxygen dissociation curve). As a result the alveolar–venous gradient for oxygen diffusion is smaller and equilibration slower than at ground level.

People ascending to altitude in a matter of minutes, rather than over several days, make two adaptive responses to hypoxia: an increase in blood flow and the modest hyperventilation mentioned previously. These limit but do not abolish the effects of lack of oxygen. The consequences (Fig. 7) include loss of night vision, impairment of the ability to learn complex and then simple tasks, a deterioration in the performance of already learnt skills, a progressive loss of muscular power (aerobic capacity), and eventually loss of consciousness, convulsions, and death.

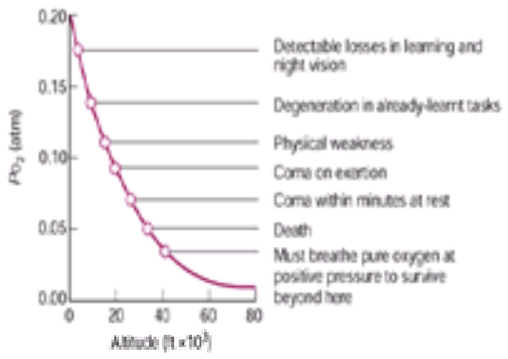

Fig. 7 A summary of the functional consequences of altitude hypoxia.

As Fig. 7 shows, people abruptly exposed to altitudes of 10 000 ft (3000 m) and above are mentally unreliable and physically weak. This altitude is taken as the ceiling above which it is mandatory to provide aviators with oxygen. To be safe, the ceiling that is actually used is almost always 8000 ft (2440 m), at which barometric pressure is 565 mmHg, arterial oxygen pressure is around 55 mmHg (i.e. sitting just at the top of the sloping part of the oxyhaemoglobin dissociation curve), and venous oxygen pressures have only fallen by 1 to 2 mmHg. It is the maximum cabin altitude that is generally permitted in civilian passenger aircraft. There is some evidence that, even at this altitude, people tire more quickly and learn more slowly than at ground level. Some aircraft have a lower maximum cabin altitude, but this requires a stronger and thus heavier cabin to contain the higher pressure, or engines producing a higher mass flow (such as the Concorde supersonic transport aircraft) which makes the aircraft less economical to run.

Two physiological features of altitude hypoxia are especially important in aviation. The first is a total lack of awareness that the mind is breaking down. This means that an affected individual cannot be relied upon to take corrective action, however well trained. It follows that protective equipment has to be designed to sense the hypoxia and come into operation automatically. The second feature, known as the time of useful consciousness, describes how rapidly consciousness is lost and thus dictates how quickly this equipment must respond.

The time of useful consciousness is the interval after the onset of hypoxia during which an aviator can be relied upon to act sensibly. This is a difficult characteristic to test, as sophisticated abilities need to be sampled in an adequate and time-consuming way at moments when the level of consciousness may be changing rapidly. Many studies have confirmed the general relation between this time interval and the altitude of sudden exposure, which is shown in Fig. 8(a). The time of useful consciousness diminishes from about 4 min at 25 000 ft (7620 m) to a minimum of roughly 15 s, which is reached at 35 000 (10 700 m) to 40 000 ft (12 200 m). This asymptote represents the sum of the 7 s or so required for blood to travel from the lungs to the brain and the equal time needed for the brain to consume the oxygen that is already dissolved in its substance.

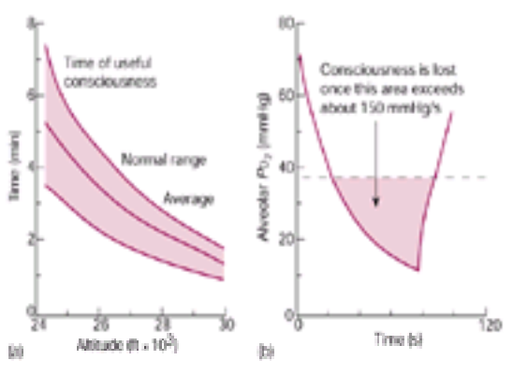

Fig. 8 (a) Variations in the time of useful consciousness with altitude. (b) One way of expressing the dose of hypoxia needed to bring about the loss of consciousness.

Some recent studies have defined the dose of hypoxia that can be suffered before useful consciousness is lost. In trained and healthy men breathing normally (i.e. with an alveolar PCO_2 of 35 to 40 mmHg), it is equivalent on a curve of alveolar PO_2 against time, to an area of 150 mmHg s, where PO_2 is less than 38 mmHg (Fig. 8(b)). However, this is sensitive to many other factors, of which the most important are the degree of hyperventilation and the acceleration that the person is exposed to at the time. Hyperventilation causes cerebral vasoconstriction, and a 'positive *g*' opposes the upward flow of blood to the brain (see below). Sometimes deterioration in consciousness is quickened by vasovagal syncope, but more often the heart is beating quite rapidly as consciousness is lost. Exertion also quickens loss of consciousness, as mentioned previously, because it forces blood to rush through the lungs, leaving insufficient time for oxygen equilibration.

The minimum cabin pressure of 565 mmHg (75.1 kPA) (8000 ft (2440 m)) in commercial passenger aircraft, is sufficiently low to bring a normal person's arterial PO_2 along the plateau of the oxyhaemoglobin dissociation curve until it is sitting just at the top of the steep part (Fig. 6). Because their blood is still fully saturated with oxygen they will not be cyanosed at this altitude. At ground level, many people with chest diseases have arterial oxygen pressures that are as low as 55 to 60 mmHg (or even lower, in which case they become cyanosed). As they ascend to 8000 ft (2440 m) their arterial PO_2 will fall further. If their hypoxaemia at ground level is due to a mismatch of ventilation to perfusion, as is usually the case, the drop in arterial PO_2 will not be as extensive as in normal people (about 40 mmHg), but if it is due to diffusion defect associated with desaturation on exertion, as in some fibrotic conditions, it may be greater. However, in either event, it can be reversed completely by the administration of oxygen, because 30 per cent oxygen at 8000 ft (2440 m) is equivalent to breathing air at ground level. The medical services of all the major airlines can provide a personal oxygen supply for any passenger if they are given notice beforehand. (It is worth checking the altitudes of the patient's destination and of any stopping point *en route* at the same time.)

Oxygen equipment and pressure cabins

Aircraft that fly below 10 000 ft (3000 m) do not need any oxygen equipment at all. Most of those that fly higher have reinforced cabins capable of holding a higher pressure inside them than out. These are of two sorts, the high-differential type, seen in passenger and transport aircraft generally, and the low-differential variety found in military high-performance aircraft. The former, holding a high transmural pressure, usually prevent pressure falling below 565 mmHg (8000 ft (2440 m)). They provide an environment in which oxygen equipment is not needed routinely and the occupants breathe cabin air. However, it is always possible that the pressure-cabin system can fail, allowing the pressure within to fall to the level of that outside. This fall can be limited by descent to a lower altitude, but it is not always practical to put the aircraft into a very steep dive, for reasons of structure or air traffic control. Similarly, it is not always practical to descend below 10 000 ft (3000 m) because, in mid-Atlantic for example, there may not be sufficient fuel for the vehicle to reach the nearest land through the dense air at the lower altitudes. For these reasons, if there is a cabin-pressure failure, an emergency oxygen supply is available for passengers and crew.

A high-differential cabin limits the vehicle's range and manoeuvrability. It also increases the risk of catastrophic damage if the fuselage is punctured. So, military high-performance aircraft are fitted with low-differential cabins. These usually prevent cabin pressure falling below 280 mmHg (37.2 kPA) (equivalent to a pressure altitude of 25 000 ft (7620 m)). That is the level at which decompression sickness becomes a serious hazard (see below). In such aircraft, oxygen equipment is needed routinely.

Sometimes military air crew have to escape in flight. The faster the aircraft is travelling the more difficult this is to do. All modern fighters are equipped with ejection seats to launch the crew into the air stream and get them clear of the tail. Usually, the ejected crew free-fall in their seats until below 10 000 ft (3000 m). This gets them through the cold hypoxic upper air as quickly as possible. A small seat- or suit-mounted emergency oxygen supply sees them safely through this stage. Then

their parachutes deploy automatically as they are released from their seats to get them safely to ground.

Mechanical effects of pressure change

In civilian passenger and transport aircraft the climb to cruise altitude takes about 30 min and involves a fall of about 200 mmHg (26.6 kPA) in cabin pressure (to the equivalent of 8000 ft (2440 m)). Descent to the ground takes much the same time. Body fluids and tissues generally are virtually incompressible and do not alter shape to any important extent when such pressures changes are applied. The same is true of cavities such as the lungs, gut, middle ear, and facial sinuses that contain air, provided that they can vent easily. Gas-containing spaces that cannot vent easily behave differently.

The thoracoabdominal wall can develop transmural pressures of +100 mmHg or so briefly, but is normally flaccid and has a transmural pressure of a few millimetres of mercury. Gas within will usually be at a pressure very close to that outside, and must follow Boyle's law. Ascent from ground level (760 mmHg) to 8000 ft (2440 m) (565 mmHg) will expand a given volume of trapped gas in a completely pliable container by about 35 per cent, which is equivalent to a radial increase of 10 per cent if it were in a sphere or 18 per cent in a cylinder of fixed length. In the abdomen this may cause slightly uncomfortable gut distension in healthy people but it is not an important problem.

The time constant of emptying of a lung that is kept full by continued decompression is less than 1 s in healthy people (peak expiratory flow is normally 1.5 lungfuls/s). Evidence from very rapid decompressions in fighter aircraft shows that healthy people at functional residual capacity can tolerate decompressions of about 200 mm Hg, from about 280 mm Hg to 80 to 120 mm Hg, which are complete in as little as 0.1 s. Even very diseased lungs can vent themselves over a minute or so. In consequence, the risk of lung rupture in normal flight is extremely rare (Fig. 9).

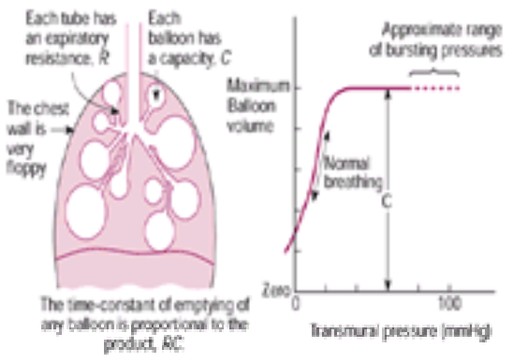

Fig. 9 A graphical summary of the factors determining lung rupture.

The cavity of the middle ear poses a separate problem since it vents easily but sometimes fails to fill because the lower part of the Eustachian tube behaves as a non-return valve, especially when it is inflamed. As a result, the cavity equilibrates quite easily on ascent but does not refill on descent, and the ear-drum bows inwards, causing pain that can be severe.

Altitude-induced decompression sickness

If ambient pressure falls quickly to less than half its original value, the gas dissolved in blood and tissue fluids may come out of solution precipitously, forming bubbles and obstructing flow in small blood vessels. The time symptoms take to develop varies widely between individuals and shortens markedly as the altitude of exposure rises. A guide to these times and variability is given in Fig. 10. Symptoms usually resolve quickly after a descent of a few thousand feet and rarely persist after descent to ground level, on oxygen. Should they persist, treatment should be along the lines detailed in Chapter 8.5.6. The risk continues to be significant in some military flights but, with the exception of passengers who have been scuba-diving very recently, it is not a hazard in commercial flights (see below).

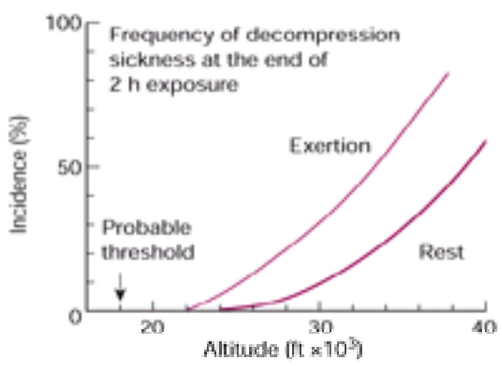

Fig. 10 The incidence of decompression sickness (percentage) at the end of 2 h of exposure to various altitudes in men at rest, or exerting themselves.

Medical problems

Most doctors at times have to consider whether their patients are fit to fly as passengers. No aircraft flight is without risk. These risks include mechanical failure of the aircraft, adverse weather conditions, pilot error, and medical incapacitation of aircrew. Commercial flying, however, is very safe. Worldwide, the accident rate is 1 in 2 million flying hours with some countries achieving 1 in 5 million flying hours. Medical conditions causing temporary or incomplete incapacitation of passengers are not uncommon in flight. The stress of modern of travel should not be underestimated. Passengers on international flights are required to check in at least 2 h before departure, airports often have very long walkways and it may be difficult to obtain help. Many passengers are elderly with heavy luggage, and flight legs may often take 12 h or more.

No significant mental or physical ill effects in healthy people can be attributed to the pressure or composition of cabin atmospheres in normal commercial passenger flight. Flights generate other concerns, namely those of jet-lag, ear and sinus problems, deep vein thrombosis, cross-infections, the safe transport of passengers who are unwell, and the medical fitness of aircrew.

Jet-lag or circadian dysrhythmia

When people are transported across two time zones or more, bodily rhythms, especially sleep/wake cycles, are disturbed and take several days to readjust, at the rate of 1 day per time zone crossed. As a result people feel sleepy when they need to be awake and vice versa. These effects are compounded by interrupted sleep and physical inactivity on prolonged flights. The combined effects may increase people's vulnerability to infections but at present there is no good evidence that this is so. What is clear is that jet-lag affects the ability to think straight for a while and may well interfere with the ability to drive. A number of studies are in progress.

Ear and sinus problems

As the pressure in the cabin changes, gases trapped in the body expand and contract. This can lead to a little abdominal discomfort on ascent and relief on descent, but matters little in healthy passengers. However, air trapped in the nasal sinuses and the cavity of the middle ear can cause pain, commonly more marked on

descent than ascent. Swallowing, jaw-wriggling, and nose-blowing manoeuvres usually sort things out, but passengers who know they have active ear or sinus problems should take decongestants with them and, if the problems are more than trivial, seek medical advice beforehand.

Cross-infection

Transmissible respiratory infections, such as tuberculosis, can be passed to fellow passengers in flight via person to person droplet spread The major obstacle to study is that passengers disperse widely and rapidly soon after landing and may not associate or report infections subsequent to flight. The risk of cross-infection in aircraft is reduced by passing recirculated cabin air through high-efficiency particulate filters which have an efficiency in excess of 96 per cent.

Deep vein thrombosis

There is a evidence that people obliged to sit in cramped positions for a long time, whether in underground bomb shelters, cars, trains, planes, or at home, may be more to liable to develop deep vein thrombosis. Ironically, this was first recognized explicitly during the airborne bombardment of London in 1940. The condition is notoriously difficult to diagnose. Again, because passengers disperse far and wide and may not link thromboembolic phenomena with recent journeys, the evidence relating the condition to flight is unsatisfactory. The mild hypoxia of commercial flight produces complex changes in clotting mechanisms. The topic is highly controversial. Good epidemiological studies are required. Passengers who have recently given birth or suffered abdominal or pelvic surgery, or are taking hormonal medication, are at increased risk and should take medical advice before travelling. The advice is usually simple: 'walk around as much as you can during the flight'.

Passengers who are unwell

Although the reduced total and oxygen pressures of the cabin atmosphere have little effect on healthy people, they can adversely affect passengers who are already unwell. Those at risk are people with definite heart or lung disease or recent chest or abdominal surgery. Almost all can be transported safely provided that the carrier knows beforehand. Usually the provision of 2 to 4 litres of oxygen per minute is sufficient for those with heart and lung problems. Passengers in late pregnancy or with a recent pneumothorax or very recent surgery need particular care.

Amongst the apparently well, those with recent fractures should note that air trapped beneath their plaster casts may expand so new casts should be split before travel. Patients who have recently undergone clinical procedures that introduce gas into the body (for example arthroscopies, air encephalograms, pulmonary needle biopsies) should not fly until it is known that the gas has been resorbed. Passengers should not scuba-dive in the 12 h before flight because otherwise they may suffer decompression illness *en route*. This should be increased to 24 h if the dive depth exceeded 30 ft (9 m). If an apparently fit passenger develops progressive limb pain in flight, oxygen should be given immediately and the passenger advised to stay at ground level for at least 12 h at the next landing point.

Methods of safe transport are well understood. The essence of the problem is passenger education. Passengers should be given accurate and easily understood information about the very slight risks of flying and about when they should seek medical advice or inform the carrier before flying. A number of carriers provide helpful information on their web sites (for example: http://www.britishairways.com/heath).

Medical fitness of aircrew

Commercial aircrew are responsible for very expensive vehicles and for many lives. In battle, military aircrew may be responsible for much more expensive vehicles and weapons and for very many more lives. Because of this and the great costs of training, military forces and airlines demand the highest physical and psychological standards of aircrew. The medical criteria set for entry to training are usually higher than those for aircrew who are already trained and experienced. Commercial and private aircrew need a current certificate that they are 'fit to fly', which can only be provided by authorized medical examiners. In Europe medical standards for commercial aircrew are imposed by the Joint Aviation Authorities in line with the standards set by the International Civil Aviation Authority.

Summary

Aerospace medicine is a subject that is largely understood. The major peer-review journal in the field is *Aviation, Space and Environmental Medicine*, published by the Aerospace Medical Association. 'Fitness to fly' certificates for commercial and military aircrew can only be provided by authorized medical examiners. Further epidemiological studies are needed on deep vein thrombosis, the effects of jet-lag, and the transmission of respiratory infections resulting from commercial flights. For the care of individual patients, airline medical services are a valuable source of advice and will provide in-flight supplies of oxygen if needed. Specialist organizations can supply medical and nursing support for patients who are too ill to travel alone. Surprisingly few patients are truly too unfit to fly.

Further reading

General

Campbell RD, Bagshaw M (1999). *Human performance and limitations in aviation.* Blackwell Science, Oxford.

Coker RK, ed (2001). Managing passengers with lung disease planning air travel: British Thoracic Society recommendations. *Thorax* (in press).

DeHart RL, Millet KC, Murphy J, eds (1996). *Fundamentals of aerospace medicine*. Williams and Wilkins, Philadelphia.

Ernsting J, Nicholson AN, Rainford DJ (1999). *Aviation medicine*, 3rd edn. Butterworth Heinemann, London.

House of Lords Inquiry (2000). *Air travel and health.* Her Majesty's Stationery Office, London.

Jagoda A, Pietrzak M (1997). Medical emergencies in commercial air travel. *Emergency Medicine Clinics of North America* **1**, 251–60.

Joint Aviation Authorities (1996). *Joint aviation requirements: flight crew medical requirements* (JAR/FCL/Part 3-Medical). Westwood Digital, Cheltenham.

Rosenberg CA, Pak F (1997). Emergencies in the air: problems, management and prevention. *Journal of Emergency Medicine* **15**, 159–64.

Thibault C (1997). Special Committee report: cabin air quality. *Aviation, Space and Environmental Medicine* **68**, 80–2.

Concerning the deep vein thrombosis controversy

Bendz B *et al*. (2000). Association between acute hypobaric hypoxia and activation of coagulation in human beings. *Lancet* **356**, 1657–8.

Ferrari E *et al*. (1999). Travel as a risk factor for venous thromboembolic disease. *Chest* **15**, 40–444.

Kraaijenhagen RA *et al*. (2000). Travel and the risk of venous thrombosis. *Lancet* **356**, 1492–3.

Miller A, ed (1997). Suspected acute pulmonary embolism: a practical approach; British Thoracic Society recommendations. *Thorax* **52** (Suppl. 4), 1–24.

8.5.6 Diving medicine

D. M. Denison and T. J. R. Francis

Introduction

Divers are exposed to many hazards. Frequently it is too late or impractical to give specific help once trouble occurs, so diving medicine is largely concerned with prevention. It depends upon a thorough understanding of the jobs that divers do and the risks that they run completing them. Almost every diving accident is due to a failure of education or equipment design.

Leaving a typical shore (Fig. 1), the sea bed falls with a slope of about 1 in 50 until it is 200 to 300 m deep. This shallow stretch is the continental shelf. The bed then angles more steeply (roughly 1 in 15), as the continental slope, to descend to vast flat expanses of soft mud, the abyssal plains, at depths of 3 to 6 km, interrupted by occasional mountains and chasms. The deepest point is just over 11 km below the surface. Water is almost incompressible. Ambient pressure rises linearly with depth by 1 atmosphere for every 10 m descent. Diving is confined largely to the continental shelves, i.e. to pressures of 1 to 30 atmospheres.

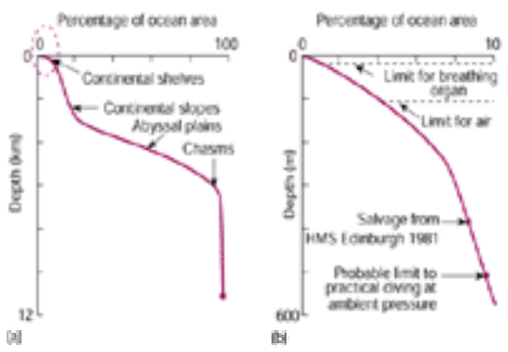

Fig. 1 (a) A cumulative depth versus area plot of the oceans. (b) A similar plot of the top 600 m, including the continental shelves.

Currents, arising from differences in water temperature and salinity, course across the abyssal plains and well up the continental slopes as mineral-rich streams to supply vegetable life in sunlit upper zones. Animals that feed on these plants or each other are concentrated so that 80 per cent of the biomass lies in the top 200 m, mainly close to the shore. Together, these sites form an area equal to that of Africa, infinitely more fertile and, as yet, virtually unfarmed.

Limitations to diving

At the surface, tidal currents are accelerated or slowed down by features of the shore but often exceed the speed at which people can swim (Fig. 2(a)). It may only be practical to dive in slack water, i.e. for an hour or two each day. Waves are often sufficiently high to prevent divers being launched or recovered safely (Fig. 2(b)). Tidal currents tunnelled along marine canyons, and springs of fresh water or falls of cold ocean water can carry divers in unexpected directions without them being aware.

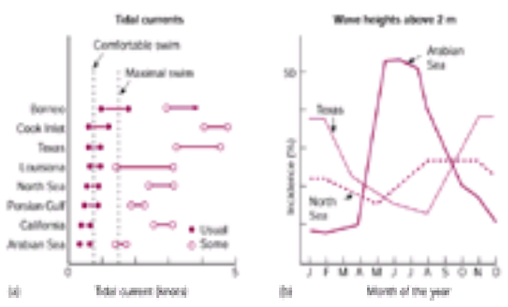

Fig. 2 (a) A plot of the usual and the not uncommonly seen tidal currents in eight diving sites around the world. (b) A plot of the percentage incidence of waves exceeding a height of 2 m at different times of the year in three of the diving sites.

Dawn arrives late and dusk comes early to the sea. Light that is not reflected at the surface is absorbed and scattered, halving its intensity with every 1 or 2 m of descent. It is effectively 'night' below 80 m. Most sports diving takes place in clear shallow waters at placid times of the year. Professional diving, for example harbour work, hull inspections and repairs, pipeline surveys, oil-rig work, and wreck salvage, occurs throughout the year, alongside or beneath large obstructions, in turbid waters where finding the task, let alone completing it, may be very demanding. Artificial illumination is often ineffective because of backscattering.

Underwater, binaural localization of sound is poor. Noises are transmitted almost five times as fast and many times more efficiently through water than air. Loss of air conduction raises auditory thresholds by 30 to 60 dB. Neoprene foam hoods that keep the head warm raise thresholds by a further 30 dB or so. Superior transmission of sound also increases susceptibility to blast injury.

All of the oceans except for the surface waters of tropical seas are too cold for individuals to remain in for long without insulation (Fig. 3). In air, people maintain body temperature at 37 °C with minimal effort when the air temperature is 18 to 24 °C, the zone of thermal neutrality. In water, this zone is high and narrow (35.0 to 35.5 °C). Loss of tactile discrimination and manual dexterity are major problems when working in cold water.

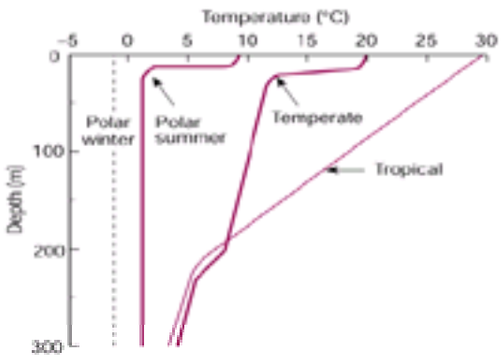

Fig. 3 Variations in sea temperature with site and depth. Note that water temperatures of less than 20 °C are too cold for unclothed individuals to stay in for very long.

Problems of simple immersion

On immersion blood is displaced upwards from the abdomen and legs, some 500 ml entering the chest, distending the large veins and the right atrium. Local stretch receptors signal an excess circulating volume and promote diuresis. On emmersion the displaced blood drops down from the chest revealing hypovolaemia.

Effort in water is lost in moving, making most tasks more tiring and less efficient than on land (Fig. 4). The maximum sustained thrust that swimmers can develop is about 5 kg, which is just enough for propulsion at 1 to 2 knots. On full inspiration an adult swimmer is about 2.5 kg positively buoyant and needs half of their maximum swim power to descend. A swimmer breathing out to residual volume will be about 2.5 kg negatively buoyant and will need half of their maximum swim power to ascend. The swimmer's almost weightless body can be displaced with ease. The body can be poised at will but body weight can no longer be used to apply leverage or torque, or to stay in place when a current is running.

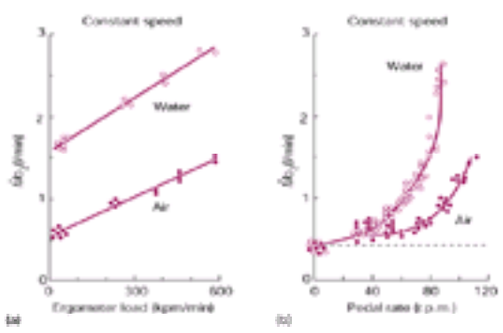

Fig. 4 A comparison of oxygen consumption (O_2) when pedalling a cycle ergometer in air and under water, (a) at a constant speed (60 rev/min) and (b) at a constant light load. Note the high cost of moving the limbs through water. Most people's aerobic capacity is about 2.5 litres O_2/min.

Considerations before diving—fitness to dive

Divers have a physically demanding job and often work in sites remote from medical aid. They must be strong and free of any active or latent condition that could erupt while they are away. Military and commercial divers must have annual medical examinations for 'fitness to dive', and sports divers are urged strongly to do likewise. The medical checks have three aims: to determine whether candidates are fit to swim in swift currents and rough waters at remote sites, to discover whether they are fit enough to rescue a fellow diver in trouble, and to find out if they have any specific problems that would make them a liability to themselves or others under water.

Divers are expected to be bodily fit and mentally stable and free of conditions such as epilepsy and ill-controlled diabetes or asthma. They should not be addicted to alcohol or any other drug and they should not have a history, or other evidence, of obstructive lung disease, ruptured eardrums, or aural surgery. Divers who are generally fit to dive should not be allowed to do so when they have chest, upper airway, or ear infections, or if they become obese. They should not dive while taking any medication that could impair their ability to think clearly or orientate themselves in space correctly.

Three unresolved issues in 'fitness to dive' assessments cause concern: the role of lung function tests, the significance of a probe-patent foramen ovale, and how to deal with a history of asthma.

The role of lung function tests

If a candidate runs several kilometres a day, is a good swimmer, was always good at games at school, and has no history of recent respiratory disease it is very likely that they will be fit enough to dive. The forced expired volume in 1 s (**FEV1**), which should be more than 75 per cent predicted, if multiplied by 35, measures the most air that the subject can process in 1 min. The forced vital capacity (**FVC**), which is a guide to the volume of useable lung, should also be more than 75 per cent predicted because there is good evidence that subjects with a low FVC and, by inference, lungs that are too small for their bodies, are more liable to lung rupture on rapid ascents. The FEV1/FVC ratio is not especially helpful. The expiratory flow rate that matters in diving is the peak expiratory flow (**PEF**)/FVC ratio which is a reasonable measure of the time constant of emptying when the lung is full. The peak expiratory flow rate should be at least one and a half predicted FVC per second.

Patent foramina ovale

On most ascents many bubbles can be detected in systemic venous blood. The bubbles are normally trapped unnoticed in pulmonary capillaries. One in four healthy people has a slightly patent foramen ovale that may allow bubbles to pass into the left heart causing cerebrospinal or coronary gas embolism. Four out of five of the victims of cerebrospinal gas embolism in diving have patent foramina, but it is unclear whether that is cause or effect. Systemic venous bubbles are common but arterial gas embolisms are rare. Bubbles have complicated inflammatory interactions with vessel walls, and any significant load arriving in the lung could trigger pulmonary vasoconstriction, raising right heart pressures and opening a previously 'closed' foramen. Present evidence does not justify excluding one-quarter of the population from diving, as gas embolism is rare.

A history of asthma

It has been the custom to bar candidates with a history of asthma from diving, because of the fear that they were more likely to rupture their lungs on fast ascents. However, childhood asthma often disappears and very many people with current mild asthma are known to have completed very many dives without ill effect. It seems there is a case for relaxing this requirement slightly, but three questions prevail: is the candidate fit enough to cope with a diving emergency, liable to

exercise-induced asthma, or likely to bronchoconstrict on the inhalation of salt water? Perhaps the most reasonable view at present is to allow a candidate with very mild asthma to dive, provided that they can demonstrate a stable and essentially normal spirometry over a 2-month period, that they do not have exercise-induced asthma, and that they do not constrict abnormally to a saline aerosol challenge.

Problems of descent

On leaving the surface, gas in the chest, abdomen, and clothing is subject to Boyle's law. Vertical movements may quickly become uncontrolled because of the positive feedback between depth and buoyancy. The deeper divers go, the denser they become and the more rapidly they fall. The higher they rise, the less dense they become and the faster they ascend. The chest wall can only maintain a pressure difference equivalent to 1 or 2 m of water, so the gas within the respiratory tract (lungs, upper airways, sinuses, the middle ear, and the Eustachian tubes) is virtually at the same pressure as the surrounding sea. The lung of a person breath-hold diving to 30 m (4 atmospheres absolute) is compressed from total lung capacity to residual volume, so they will need half of their aerobic capacity to ascend. Gas in clothing exaggerates these changes, leaving little margin for controlling unexpected ascent or descent.

Descent may lead to several mechanical problems that become obvious early in a dive (suit/facemask squeeze, lung squeeze, sinus squeeze, middle-ear squeeze, inner-ear squeeze, reversed ear syndrome, alternobaric vertigo, and caloric vertigo). These occur commonly, even on breath-hold dives.

The squeezes

On descent, rising ambient pressure may force the diver's face into the air-filled mask, uless the mask is vented, leading to facial oedema and subconjunctival haemmorhages. These will resolve spontaneously and need no treatment. Similarly a dry suit, particularly if poorly tailored, can pinch the skin resulting in linear wheals that are commonly distributed around the neck, axillae, and groins. Again, these require no active intervention but should not be confused with the cutaneous signs of decompression illness. Occasionally suit squeeze can be severe enough to limit a diver's movements in a deep-water emergency. Increasing ambient pressure compresses the lung according to Boyle's law but more blood is drawn into the chest to replace the vanishing lung so lung sqeeze is only serious at great depths in breath-hold dives. When gas in obstructed sinuses is compressed, sinus walls become oedematous and may bleed. The blood or clot is usually expelled by trapped gas expanding on ascent.

When gas trapped in the middle ear contracts it draws the ear drum inwards and the round and oval windows of the inner ear outwards. Difficulty in clearing the ears is the commonest problem in diving. Attempts to clear them by over-vigorous Valsalva manoeuvres have the opposite effects. If one ear clears much before the other disorientation may occur due to uneven stimulation (alternobaric vertigo). If the external auditory canals are unequally blocked at the start of a dive, cool water will enter one canal before the other leading to caloric vertigo. Drum ruptures in any of these conditions normally heal spontaneously. Diving should not be permitted until the drums have healed. Persistent ruptures require surgery.

Problems at the bottom of a dive

For more sustained dives, gas must be delivered to the diver at the same pressure as the surrounding water. It can be sent via an umbilical pipe from the surface, in which case it can flush through the helmet or face mask continuously—which is wasteful of gas but easily engineered—or it can supply a regulating valve that provides gas on demand only. Alternatively, the diver can take self-contained underwater breathing apparatus (scuba). This feeds a demand regulator and rarely lasts for more than 1 h. Professional divers often use a combination of all three systems, i.e. a surface-demand supply for routine use, with a helmet-flushing capability for occasional comfort or emergency use, and a small, back-mounted gas supply ready in case the surface supply fails. Sustained dives can cause biochemical problems (nitrogen narcosis, oxygen toxicity, and the high-pressure nervous syndrome). These influence the choice of the breathing mixture to be used.

Nitrogen narcosis

Air can be breathed safely down to depths of 50 m, although tests of sophisticated cerebral function show that there is already some impairment at 20 m. Below 50 m, mental deterioration becomes obvious, manifested by actions such as divers offering their mouthpiece to neighbouring fish. This occurs as further nitrogen dissolves in some parts of nerve membranes, making them thicker. Replacing the nitrogen with smaller helium molecules allows divers to go deeper. Changes in function develop within minutes and are rapidly reversible, because they depend purely on passive chemical solution. If divers breathe an oxygen–helium mix rather than air they can descend to the lowermost parts of the continental shelves (730 m) without narcosis. Nitrogen narcosis is a wholly preventable hazard of diving, which can be completely reversed within minutes by ascent and needs no treatment.

Oxygen toxicity

Oxygen becomes toxic to the lungs when alveolar oxygen pressure exceeds half an atmosphere, and it becomes toxic to the nervous system when the alveolar, and so arterial, oxygen pressure exceeds 2 atmospheres. Effects are due to complex chemical interactions, rather than physical solution, and so take time to develop and reverse. Oxygen damages the lung by irritating its endothelial and epithelial surfaces. The time taken for symptoms to appear depends upon the dose. It varies from several hours at half an atmosphere PO_2 to a few hours at 2 atmospheres. Above that pressure, the neurological sequelae overshadow the pulmonary damage that still occurs (Fig. 5).

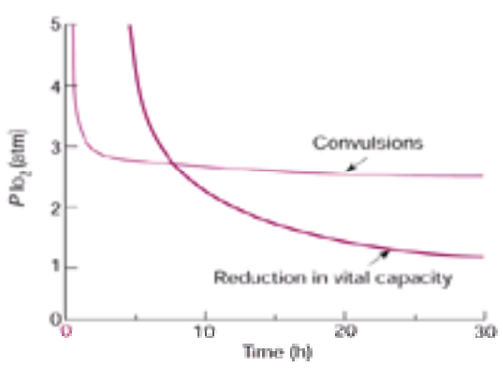

Fig. 5 Commonly observed pulmonary and central nervous O_2 toxicity versus time curves related to inspired PO_2 (PO_2) (constructed from the data of many workers).

High-pressure nervous syndrome

At great depths, people breathing oxygen–helium mixtures show neurological disturbances due to direct compression of nerve tissues, making parts of the membranes thinner. Nerve impulses travel more easily and convulsions result. Thickening the tissues with some nitrogen prevents this and allows divers to go deeper.

Problems of ascent

Return to the surface poses three hazards: expansion of gases trapped in sinuses, lung rupture, and decompression sickness.

Expansion of gas in sinuses

If gases trapped in sinuses cannot escape on ascent they will press on the walls and eventually burst them. Rupture of the ethmoid sinus is rare but feared because of the risk of cerebral infection.

Lung rupture

In a sustained dive the lungs contain enough gas to burst them on ascent unless they are adequately vented. The full lung has a bursting pressure of about 75 mmHg (1 m of seawater) and time constants of emptying that are close to 0.3 s. Divers are taught to exhale continuously whenever they ascend, and to ascend no faster than the bubbles they exhale. In such ascents the lungs have time to empty sufficiently and the risk of rupture is low.

Lung rupture often occurs in divers breath-holding on ascent or ascending too fast. Central tears lead to mediastinal emphysema. Peripheral tears cause pneumothorax. Gas may also enter the circulation as air emboli. Escaped gas expands as the ascent continues, making matters worse. The victim may lose consciousness immediately but otherwise notes dyspnoea, cough, or haemoptysis a few minutes later. Voice change and discomfort in the throat or behind the sternum may also be noticed. There may be surgical emphysema of the neck and upper chest, increased cardiac dullness or crepitus, and/or evidence of a pneumothorax. If air embolism has occurred there will be additional neurological signs. Uncomplicated pneumomediastinums, superficial emphysema, or non-tension pneumothorax are treated by oxygen alone. They should not be positive-pressure ventilated. Tension pneumothorax needs an immediate chest drain. Patients with neurological signs should be recompressed as soon as possible.

Decompression sickness occurs because during any dive extra inert gas, usually nitrogen or helium, goes into passive solution in tissues. On ascent this gas can come out of solution in an uncontrolled way, forming bubbles in the circulation and within tissues. As the ascent continues, these bubbles increase in size and number, blocking blood vessels and distorting or rupturing cells. On redescent the bubbles contract and are eventually resorbed. If the first or the subsequent ascent is slow enough, few if any bubbles are formed, the extra gas diffuses into the bloodstream and out of the lungs easily, and the diver reaches the surface unharmed.

A vast amount of experimental work has been done to determine the safe limits to 'no-stop' diving (Fig. 6) and the depth–time profiles that have to be followed on returning to the surface after any longer dive. Knowledge of safe practice is tabulated in a series of lengthy decompression schedules, which can be obtained from diving clubs or the helplines given below. After very long dives the ascent is very slow and can take several days.

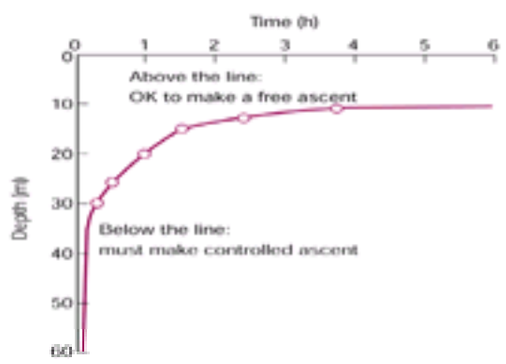

Fig. 6 The 'no-stop' diving curve that determines whether a dive has been shallow and brief enough for the diver to make a free ascent to the surface.

About 1 per cent of dives conducted to recommended schedules, and many badly conducted dives, lead to decompression sickness. The most common presentation in military and commercial diving is limb pain, commonly of the shoulders or elbows in divers and of the knees and hips in tunnel workers. Pain may present a few minutes after a dive or as much as 24 h after the diver has left the water, often as a dull and poorly localized ache of gradual onset. It is not usually made worse by movement of the joint, although weight bearing may make knee pain worse. Physical signs of inflammation are uncommon. Left untreated, the pain will regress and disappear over 2 or 3 days. Recompression commonly improves the pain quickly.

Although sports divers also experience limb pain, neurological symptoms are more prevalent. Sensory disturbance is common, with numbness and paraesthesiae being frequent manifestations, often patchily distributed with no clear dermatomal or peripheral nerve distribution. A particularly fulminant form commences with girdle pain accompanied by loss of sensation and power in the lower limbs. Involvement of the brain is common and may be subtle. Denial is a frequent feature. Any of the higher functions can be involved, including loss of short-term memory, altered affect, visual disturbance, and loss of consciousness. Involvement of the inner ear can present in an identical manner to inner ear squeeze.

Divers who display any manifestation of decompression illness within 24 h of a dive should be managed as if they have the condition. This involves instigating first aid measures and obtaining an informed opinion. The two mainstays of the first aid management are the provision of oxygen (as close to 100 per cent as possible) and rehydration. Within the United Kingdom, informed opinions can be obtained from the British Hyperbaric Association or the Institute of Naval Medicine via their 24-h helplines (0831 151523 and 02392 768026 respectively).

All but trivial cases of decompression illness should be recompressed as soon as possible. It is an effective treatment. The object is to reduce the size of existing bubbles and prevent the formation of new ones, before irreversible infarction and oedema occur.

Problems after the dive

Most ascents are associated with the appearance of many bubbles in systemic venous streams, but these do not usually cause symptoms and so such dives are considered to be 'safe'. Decompression tables that have been established experimentally define the boundaries beyond which more than 2 or 3 per cent of divers will experience symptoms of decompression illness. Such dives are 'unsafe'. There is no doubt that the neurological or other sequelae to unsafe dives may fail to resolve completely, but these are regarded as the consequences of unsafe practices. More recently, diving physicians have been asking whether 'safe' dives lead to insidious, cumulative damage. Although the damage is slight or subclinical, there is definite evidence now that they do. Autopsies on asymptomatic divers with no history of acute decompression illness have revealed that their brains and spinal cords contain considerably more microinfarcts than those of non-diving controls. More importantly, radiographs of the long bones of divers and caisson workers show increasing numbers of aseptic infarcts in a sizeable minority (up to 11 per cent). The incidence is higher in those with a history of overt decompression illness but also occurs in those without. Infarcts can occur after a single decompression, but their incidence rises with age, depth, and diving intensity. Those in the shafts of bone are asymptomatic, but those at juxta-articular surfaces can be severely disabling (dysbaric osteonecrosis). They are more common in caisson workers than divers, but are even seen in professional breath-hold divers, such as the Ama of Japan, in whom the dissolved gas burden must be light. The aetiology is unknown, but gas embolism is the favoured explanation.

Commercial diving, especially saturation diving, causes the lung's total and vital capacities to expand, its FEV1:FVC ratio to fall, and its pulmonary capillary blood volume, as judged by carbon monoxide transfer, to fall. The effects are slight but definite and may be cumulative. The expansions in lung volumes are attributed to the training effects of breathing compressed gases for long times. The fall in the FEV1:FVC ratio is mainly due to the rise in FVC but there are hints of additional small-airway damage. The fall in pulmonary capillary blood volume appears to be due to transient episodes of hyperoxia during saturation diving procedures, but may also be associated with the influx of bubbles from systemic veins on 'safe' decompressions.

It is also known that commercial divers develop a mild degree of high-tone deafness, currently attributed to the noise of gas flows within their helmets.

Codes of safe diving practice given in the Health and Safety Executive's Diving Operations at Work Regulations 1998 and the Compressed Air Regulations 1996 have been accepted nationally and internationally.

Conclusion

Diving is a sometimes very vigorous activity that demands a high degree of mental and physical fitness. It exposes people to several physical and chemical challenges that are reasonably well understood. Because it takes place remote from medical help, there is a strong emphasis on prevention of illness by following

safe practices. The practices have been developed empirically on the non-appearance of symptoms in brief trains of dives. There are now indications that these practices may not be quite as safe as first thought, but the cumulative effects are generally slight.

Further reading

Bennett PB, Elliott DH, eds (1993). *The physiology and medicine of diving*, 4th edn. Saunders, London.

Bove AA, ed (1997). *Diving medicine*, 3rd edn. WB Saunders, Philadelphia.

Edmonds C, Lowry C, Pennefather J, eds (1993) *Diving and sub-aquatic medicine*, 3rd edn. Butterworth-Heinemann, London.

Lundgren C, ed (1999). *The lung at depth.* Marcel Dekker, Basle.

8.5.7 Lightning and electrical injuries

Chris Andrews

Introduction

Lightning is a powerful force; it provides spectacular displays and has evoked an extensive mythology. The comparatively recent discovery and distribution of electricity have had an equally profound effect, and provide truth to the adage that 'electricity is a good servant and a bad master'.

Epidemiology

Lightning injury

The accepted case fatality of lightning shock is around 30 per cent. Estimated mortality is around 0.3 per million population in the United States each year, compared with 1.5 per million in Singapore, 0.3 per million in Australia, and fewer than 0.1 per million in the United Kingdom.

In the early part of the twentieth century, most individuals struck were outdoor workers (67 per cent) and outdoor recreationalists (28 per cent). The present corresponding figures are 45 per cent and 50 per cent, which may be explained by changes in social and work habits. Indoor strikes (for example by current conducted through communication or power apparatus) continue to account for about 5 per cent of these accidents.

Males are more often injured than females (1.67 males to 0.33 females); the age group most at risk are those aged 20–29 years. Risky situations include sheltering under trees, on open water, on tractors and in open fields, and playing golf. Regional differences correlate well with storm activity and population density in that area.

Electrical injury

Industrial and domestic situations are the two main areas in which electrical injury occurs.

Electrocution ranks fifth in the causes of workplace death, accounting for the death of 10 000 workers each year in the United Sates, with a further 10 million being injured. Most of the victims work for utility companies, followed by mining and construction workers. Contact with power lines causes 53 per cent of fatal shocks, and contact with power tools account for a further 22 per cent. The most dangerous times of day seem to be between 10 a.m. and 3 p.m. on Mondays, Tuesdays, and Thursdays. Most of the victims are trade and labouring staff; sales, clerical, and professional categories are at least risk. Metal ladders and antennas are particularly dangerous and can be hoisted easily into overhead power lines. Codes of safe practice are written accordingly.

In domestic situations, contact with overhead lines is again important. Faulty repair of equipment and faulty apparatus, wiring, and especially power and extension cords account for large numbers of deaths and injuries. Children are at particular risk. Death from domestic electric shock has shown a marked decrease with the introduction of residual current devices (known as RCDs). These detect a large proportion of potentially dangerous situations—they sense if current is diverted from the supply main to earth and interrupt it in a matter of milliseconds.

Mechanisms of injury

Lightning injury

Lightning injury may be sustained in four separate ways. First, a person may be struck directly. Secondly, a nearby object, such as a tree or a building, may be struck, and someone in direct contact with it may receive a shock. Thirdly, without direct contact an arc may 'jump' to a nearby person from the struck object, thereby generating a 'side flash'. Finally, as current disperses away from the base of a strike to ground, an individual may divert current flowing in the ground to themselves. This is termed shock due to increase in earth potential. Recently a fifth mechanism has been proposed — the transient flow of current due to corona and streamer formation.

Both cardiac and respiratory function cease instantaneously under lightning strike, the cardiac arrest being asystolic. Cardiac function restarts under local pacemaker control, but respiratory function does not recommence and secondary hypoxic cardiac arrest supervenes.

The major cranial orifices are portals of entry for lightning current, and from there pathways to the brainstem are short. Respiratory function is thought to be affected there, and thence conduction through the cerebrospinal fluid and then blood to the myocardium.

The QT prolongation resulting from lightning injury may predispose to episodic arrythmias. There is no evidence to that lightning inhibits body metabolism. Resuscitation is as urgent as with any other injury.

Electrical injury

With electric shock it is important to assess the points of entry and exit and the pathway of current through the body. Once the pathway has been determined, a locus for expected injury can be established and the flow of current can be estimated from the applied voltage divided by the impedance of the proposed pathway. Most impedance is in the skin barriers, and the impedance demonstrated is non-linear. There is an initial (contact) impedance which decreases as current flow continues. Impedance also varies with time since application, contact surface area, and frequency.

For currents with a frequency of 15 to 100 Hz, externally applied from hand to hand or hand to foot, relevant parameters include 0.5 mA as the threshold of perception and 10 mA for 'let go' current. A 50 per cent chance of fibrillation exists at 2000 mA conducted for 10 ms, or at the other extreme 100 mA conducted for 10 s. Direct internal application of less than 200 µA to the heart muscle may induce fibrillation.

Joule heating may account for tissue damage in the path of the current. It may be calculated from the power dissipation in the tissue—the square of the tissue current times its impedance, the former often being hard to estimate. The complex phenomenon of electroporation, where cell membranes are breached by the electrical

induction of unstable pores in the membrane, may alternatively lead to cell death.

Presentation of the injured person

Lightning injury

A witnessed strike offers the best chance of resuscitation. The victim is not dangerous to touch, and does not constitute a risk to the rescuer. Immediate cardiopulmonary resuscitation is paramount. It has been stated that:

> Any person found with linear burns and clothing exploded off should be treated as the victim of a lightning strike. Feathering burns are pathognomic of lightning injury and occur in no other type of injury. ...Another complex diagnostic of lightning injury includes linear or punctate burns, tympanic membrane rupture, confusion, and outdoor location...
>
> Cooper *et al*. (2000)

In assessing a lightning victim, the following features must be sought.

Cardiovascular and pulmonary consequences

Asystolic arrest is the main cardiac event in lightning injury. ECG signs may take many forms, with ischaemic and infarct forms. They almost invariably resolve completely over time. Alterations in QT interval and arrythmias of many kinds are seen. ECG changes may not occur until late in the course, and so are a poor diagnostic tool. Respiratory arrest is common. A person not suffering cardiopulmonary arrest is highly unlikely to die from lightning strike ($p < 0.0001$).

Neurological consequences

Direct neural injury may occur both centrally and peripherally. All forms of intracranial bleeding have been reported. Direct cerebral damage particularly affects the basal ganglia, cerebellum, and brainstem. Dural tears, scalp haematomata, and fractures are also seen. Seizures occur as a result of anoxia and injury.

Peripheral nerve injury, including autonomic injury, can give prolonged and long-lasting disability which often develops late. Other late features include spinal cord atrophic paralysis, cerebellar ataxia, inco-ordination, paraesthesiae, and aphasiae. Ongoing complex regional pain syndromes may be seen.

Keraunoparalysis and burns

Over 70 per cent of victims demonstrate keraunoparalysis. This is a syndrome of cold, pulseless, mottled, and asensory extremities. The syndrome resembles a compartment syndrome and occurs in the line of passage of the strike current. It resolves spontaneously within 24 h with no sequelae, and requires no surgical intervention.

Burns are of minor consequence in lightning injury, and again require little intervention. Entry and exit burns may be full thickness though small. Arborescent (feathering) burns resemble fern-like patterns on the skin. Their aetiology is unknown, but they fade within 24 h. Linear burns are due to the passage of hot plasma tongues over the skin. Eschar is simply allowed to separate without further treatment. Flash may be seen, like sunburn or welder's flash, from the profound radiation of the strike. Sheet burns resulting from efflux of hot plasma may be a variant of linear burning, since both seem to follow moisture and sweat lines. There may be contact burns from heated metal such as buckles and coins.

Eye, ear, and explosive injuries

The explosive force of the lightning insult blasts clothing apart, and may cause percussive injury to the lungs and abdominal viscera. Tympanic membranes are usually ruptured, perhaps from the explosive force of the strike. Percussive eye injury, particularly retinal, has been reported. Cataracts may develop much later.

Other injuries

Renal and haematological damage have occasionally been reported. In pregnant women the fetus is unlikely to survive. Menstrual and sexual difficulties have been reported.

Electrical injury

In contrast to lightning injury, victims may suffer prolonged attachment to the source of electrical current, making them dangerous to touch. Before resuscitation they must be removed from the current source, and this usually means interrupting the current flow at the supply point.

Burns are far more serious, and may merit intense surgical treatment. The likelihood of internal burning (remembering the possibility of electroporation) may require further surgical intervention. Cardiac and respiratory burns may also exist.

Cardiovascular consequences

Fibrillation is the most common cardiac abnormality following electrical injury. Cardiopulmonary resuscitation is urgently required. Electricity suppliers have standard first aid/resuscitation procedures. Cardiac dysfunction may persist for long periods, and ECG signs may not resolve.

Neurological and muscular consequences

Neural injury may be categorized into early and late syndromes, at cerebral, cord, and peripheral level.

Early tetanic muscular contraction locks the victim onto the electrical conductors. This tetany may compromise respiratory function. Neurological injury may be hard to distinguish from hypoxic and vasospastic injury. Similarly, neural injury is often hard to separate from ischaemic injury due to vessel spasm. Early and late generalized convulsions may occur. Pareses and paraesthesiae may develop, both early and late.

In the long term complex regional pain syndromes and other chronic pain syndromes must be considered.

Burns

Burns are often severe in electrical injury and merit much treatment effort. Arc and flame burns and contact burns from current entry and exit are seen. For example, tetanic gripping of the electrical conductor causes grasp burns to the hand.

Severe internal thermal or electroporation damage may occur. The management is largely surgical. Joints, ligaments, and tendons may be severely damaged by the heat generated, and osteonecrosis may be seen.

Other aspects

Widespread muscle damage generates myoglobin that must be cleared by the kidney with a severe risk of renal damage. Other metabolic and biochemical disturbances secondary to hypoxia may develop. Massive hyperkalaemia has implications for the use of depolarizing muscle relaxants.

Eye damage includes retinal damage, with punctation and detachment, and thermal damage to other media. During follow-up the possibility of ocular pareses and cataracts must be recognized.

After shock during pregnancy the prognosis for the fetus is poor. Non-focal injury is more likely in survivors.

Psychological consequences of electrical and lightning injuries

Although electrical and lightning injuries are fundamentally different in nature and management, their psychological sequelae are similar. Sequelae may be disabling. To a large extent psychological consequences of persisting pain and dysfunction cannot be separated from organic psychological consequences.

Emotional sequelae include depression, often with organic features. It is hard to separate this from the emotional reaction to injury and continuing disability. Aggression, anxiety, and phobic features are common. Marital disharmony commonly follows social withdrawal, disinterest, and a fatigue state. Loss of interest in sex and in relationships, together with a feeling of fault or guilt, may be associated. Sleep disturbance is common.

There is loss of short-term memory with impaired concentration, higher mental functions, and loss of identity and ability.

Treatment of the injuries

First, urgent and life-saving treatment must be administered. Secondly, there must be surveillance for delayed sequelae, and thirdly long-term monitoring for morbidity, including cataract formation and psychological problems.

Lightning injury

First the casualty is resuscitated and evacuated. Cardiopulmonary resuscitation is continued until medical emergency help is obtained. Ventilation and cardiac support may be required.

ECG monitoring must be used to detect subtle effects like QT prolongation. Associated trauma is treated.

In the long term, patients are observed for development of pain syndromes. Ocular and auditory function are monitored. Sensitivity to the psychological sequelae is required, and preventive interviewing may be useful.

Carbemazepine, gabapentin, clonazepam, flecainide, and mexilitine are useful to control neurally derived pain and resulting weakness. An antidepressant (see below) is a useful adjunct to this.

Electrical injury

Urgent life support is indicated. Ventilatory and inotropic support and correction of arrhythmias may be required.

For burns, progressive debridement and/or amputation may be needed. Renal damage should be prevented.

Associated trauma is treated. Ocular and auditory function are monitored and psychological disturbances are reviewed. In the long term surveillance is similar to lightning injury.

Psychological elements

In all cases the management of the psychological syndrome is paramount and may be the greatest determinant of long-term functional capability. Awareness of the impact of the injury on employment and relationships and social networks is fundamental. Cognitive and computer aids are being developed.

An antidepressant such as a selective serotonin reuptake inhibitor, possibly citalopram, paroxetine, venlafaxine, or a tricyclic such as clomipramine, may be useful.

Early and continuing neuropsychological assessment is desirable.

Controversy

The place of polaxamers in discovering the extent of electroporation and in delineating debridement levels is of great interest.

The mechanisms of the psychological disability remain to be elucidated. Victims are frequently 'written off' as malingering or simply depressed, when a more extensive syndrome exists. The useful duration of monitoring of otherwise asymptomatic people has not been determined.

Further reading

Andrews CJ (1996). Electric shock and lightning strike. In: Pearn J, ed *The science of first aid*. St John's Ambulance Press, Canberra.

Andrews CJ *et al.* (1992) *Lightning injuries: electrical, medical and legal aspects*, chapter 17, pp. 148–70. CRC Press, Boca Raton, FL.

Bridges J *et al.* (1985). *Electric shock safety criteria*. Pergamon, Oxford.

Cooper MA (1980). Lightning injuries: prognostic signs for death. *Annals of Emergency Medicine* **9**, 134.

Cooper MA, ed. (1995). *Seminars in Neurology* **15** (3, 4). Special issues on lightning and electrical injuries.

Cooper MA *et al*. (2000). Lightning injuries. In: Auerbach P, ed *Wilderness medicine*, 4th edn, chapter 3, pp. 73–111. Mosby, St Louis, MO.

Lee RC, Capelli-Schellpfeffer M, Kelley K (1994). Electrical injury. *Annals of the New York Academy of Science* **720**.

Lee RC, Cravalho EG, Burke JF (1992). *Electric trauma*. Cambridge University Press, Cambridge.

8.5.8 Podoconiosis

S. M. Evans, J. J. Powell, and R. P. H. Thompson

Introduction

Podoconiosis (from the Greek *podos* 'of the foot' and *konia* 'dust'), named and characterized by the late Dr Ernest W. Price, is a non-filarial form of elephantiasis affecting the feet and legs of barefoot agrarian communities. It is endemic in parts of Africa, Central and South America, northwest India, and Indonesia. The disease is an obstructive lymphopathy, caused by the penetration of fine particles of silica and aluminosilicates into the tissues of the foot.

Aetiology

Areas of high prevalence are typified by a reddish brown volcanic soil that has a high clay content and is extremely slippery after rain. A large proportion of the soil comprises particles of less than 2 µm in diameter, in which the aluminosilicate kaolinite predominates. Other environmental features of areas in which podoconiosis occurs include high altitude (1250 to 2500 m), average daytime temperatures of 20 °C, and a high hot-seasonal rainfall in excess of 1000 mm annually. In these conditions, weathering of the volcanic alkali basalt rock favours formation of a fertile clay soil, which attracts subsistence farming and a largely barefooted population. In adjacent, tropical, non-endemic, lowland areas, more complete chemical degradation of the parent rock reduces soil fertility and also the biological activity of the soil particles.

The greater exposure of men to soil particles during farming is probably responsible for their higher incidence of podoconiosis compared with that of women. In areas of high prevalence, for example in highland areas of East Africa, up to 70 adults in every 1000 are affected. There appears to be a familial tendency to develop the disease, and this may reflect an inherited inability of the immune system to process and detoxify ultrafine particles. The socio-economic consequences of belonging to such a family, in terms of marriage opportunity and employment, are considerable.

Microparticles

Using polarized light microscopy, birefringent particles are frequently seen in lymph node macrophages, but in tissue sections from the foot electron microscopy is needed. Price's X-ray microanalytical studies of the tissues of the foot and femoral nodes in areas with podoconiosis demonstrated the presence of ultrafine particles, consisting mainly of silicon and aluminium with smaller amounts of titanium and iron. Particles were found both within macrophage phagolysosomes (Fig. 1) and free within the tissue, and their elemental compositions matched those of the fine particles of the local soil. Morphological analysis identified kaolinite (an aluminosilicate), amorphous silica, titanium and iron oxides, with some quartz, exactly as in the local soil. Such mineral particles are also found in many areas of the world without podoconiosis so the exact reason for their pathogenicity can only be hypothesized. However, fine particles have been shown to be important aetiological agents in a number of human diseases, in particular chronic inflammatory lung conditions such as silicosis and asbestosis, and asthma.

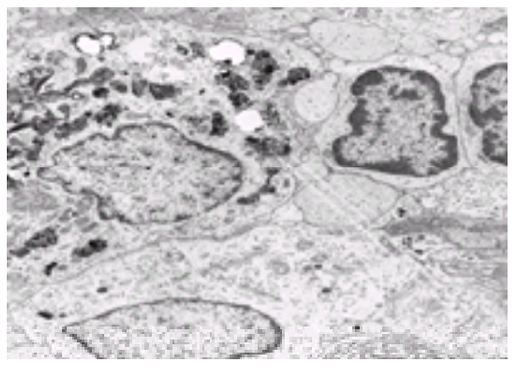

Fig. 1 Electron micrograph of a femoral lymph node of an elephantiasic person. Note the two macrophages on the left, the upper of which contains a number of clumps of microparticles. X-ray analysis showed that most of these were silica with varying amounts of aluminium, iron, and titanium (×7200).

In podoconiosis, the small size of the particles is likely to play an important role in dictating toxicity. Evidence for this comes from data showing that a significantly higher proportion of the soil in endemic areas consists of particles less than 5 µm in diameter than in neighbouring non-endemic areas. Indeed, particles released after digestion of diseased femoral lymph nodes were found to range in size from 0.3 to 6.0 µm. Transcutaneous entry of macromolecules and micron-sized particles is well recognized, but it may be facilitated in podoconiosis by the presence of larger, abrasive crystals of quartz in the silt fraction of the soil.

The volcanic origin of the rock is clearly important. Using thermoluminescence, specific defects in the crystal lattice, which are fixed in the rock as the hot volcanic lava undergoes rapid cooling, have been identified in microparticles from affected areas. Furthermore, newly fractured mineral crystals exhibit much higher reactivity than those that have been weathered. The microtopography of particle surfaces has been analysed by thermostimulated exoelectronic emission, which also identified a characteristic peak in soil fractions from endemic areas.

Another factor is persistence, since the particles are not easily dissolved or degraded in the tissues. In experiments with cultured murine macrophages, soil particles from endemic areas were toxic at a lower concentration than the corresponding soil fraction from non-endemic areas. Thus, the particles have a direct toxicity to immunocytes but this could be modified by interactions with other ions (such as iron) or biomolecules.

Once internalized, toxic mineral microparticles exert a range of effects, from mild irritation to extensive fibrosis and obstructive lymphopathy.

Pathology

The earliest histological feature of podoconiosis is seen in the dermis where collections of lymphocytes accumulate around channels of the superficial lymphatic plexus, sweat ducts, and sweat glands. A chronic inflammatory infiltrate is evident in the adventitia of the deeper lymphatic vessels, with subendothelial oedema gradually replaced by fibrosis. Similarly, lymph node capsules are thickened by fibrosis and the interfollicular sinuses become dilated and packed with lymphocytes. Macrophages are evident in the cortex of the node and frequently contain aggregations of particles, which often exhibit birefringence. With increasing severity of disease, fibrosis becomes more advanced, especially at the corticomedullary boundary.

Clinical appearances

The clinical manifestations of podoconiosis depend upon the degree to which fibrosis occurs in the dermis. Two main types are described, although most cases show

features of both types, between the two extremes.

In the soft or 'water-bag' type, there is little dermal fibrosis and minimal hyperkeratotic change. The skin is smooth and can be grasped between the fingers. Lymphoedema is marked and extends gradually from the dorsum of the foot to involve the whole lower leg. The swollen limb pits on pressure (Fig. 2(a)).

Fig. 2 Clinical types of elephantiasis. (a) The soft or 'water-bag' type. Swelling is readily reduced by compression or elevation. The skin is soft and can be pinched off the bones. (b) The hard or 'leathery' type. The skin is rough and fixed to deeper tissues. Nodulation is seen on the toes and dorsum of the foot.

By contrast, the hard or 'leathery' type is characterized by extensive dermal fibrosis, which may be more than 3 cm thick. The skin is fixed to deeper tissues, does not pit, and hyperkeratosis and hyperpigmentation are marked. In front of the ankle and behind the toes, there may be many folds of firm, rough skin, which are easily traumatized, and multiple nodules often develop (Fig. 2(b)).

A subgroup of the hard or 'leathery' type is the 'slipper' type, in which hyperkeratosis is distributed around the back of the heel, the borders of the sole, over the toes and on to the dorsum of the foot. Lymphatic obstruction is confined to the deep lymphatics of the plantar region so that the dermal fibrotic reaction occurs in the distribution of a slipper.

Natural history

Podoconiosis usually starts between the ages of 10 and 20 and progresses at a variable rate over months or years. The first symptoms are a burning sensation of the soles of the feet and itching of the lower legs, usually intermittent and unilateral at first, especially in bed at night or after exertion. Uncovering or raising the leg usually alleviates symptoms, but alcohol consumption, exercise, or menstruation may exacerbate them. The first swelling is usually evident just proximal to the first toe cleft, and spreads after a time to involve the rest of the dorsum of the foot, and later the whole lower leg. The chronic progression of the disease is marked by intensification of the burning discomfort and itching, and it may be punctuated by acute episodes of lymphangitis. Infected ulcers and cellulitis may complicate the clinical picture. Both legs are always involved, although there may be considerable asymmetry. Once established, the femoral lymph nodes are often enlarged and tender but they rarely suppurate.

Management

Price has emphasized the importance of obtaining the co-operation and understanding of patients in order to maximize the benefits of intervention.

The mineral load in the tissues cannot be reduced once it has been absorbed. Therefore, the main strategy is aimed at prevention. Provision of footwear is the most obvious protection, and it is highly effective, but matting to cover the bare ground of residential huts is also used. Occasionally, it is possible to arrange a change of employment or residence.

Anti-inflammatory analgesics, such as aspirin, are effective for burning discomfort. Reduction of oedema will improve itching, more easily accomplished in the early phases of the disease and in the soft or 'water-bag' type. Elevation of the foot of the bed and elasticated stockings may be sufficient initially, but a programme of treatment using compressive methods (compressive bandaging or intermittent compression machines) is needed in more advanced cases.

Finally, a variety of traditional drugs in tropical Africa, derived from plants, has been used in podoconiosis with the aim of reducing fibrosis and progression of the disease. Some of these contain benzopyrones, coumarin, and toxerutin, which may have antifibrogenic properties, but none has yet been convincing in the long-term control of limb swelling in this unpleasant and disfiguring disease.

Further reading

Blanke JH *et al.* (1983). Correlations between elephantiasis and thermoluminescence of volcanic soil. *Radiation Effects* **73**, 103–13.

Blundell G, Henderson WJ, Price EW (1989). Soil particles in the tissues of the foot in endemic elephantiasis of the lower legs. *Annals of Tropical Medicine and Parasitology* **83**, 381–5.

Davies JE, Townsend PD (1990). Exoemission of Ethiopian soils and the endemicity of non-filarial elephantiasis. *Radiation Protection Dosimetry* **4**, 185–8.

Harvey RJ, Powell JJ, Thompson RPH (1996). A review of the geochemical factors linked to podoconiosis. In: Appleton JD, Fuge R, McCall GJH, eds *Environmental geochemistry and health*, Geological Society special publication 113, pp 255–60. Geological Society, Bath.

Price EW (1990). *Podoconiosis: non-filarial elephantiasis*. Oxford University Press, Oxford.

Price EW, Plant DA (1990). The significance of particle size of soils as a risk factor in the etiology of podoconiosis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **84**, 885–6.

Spooner NT, Davies JE (1986). The possible role of soil particles in the aetiology of non-filarial endemic elephantiasis: a macrophage cytotoxicity assay. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **80**, 222–5.

8.5.9 Radiation

J. R. Harrison

There are two types of radiation: ionizing and non-ionizing. The former includes radiation from naturally occurring sources, such as radon gas or cosmic rays, and man-made sources, such as X-rays. The latter includes radiation from natural ultraviolet light and radiowaves and man-made emissions from radio transmitters and mobile phones. In all cases the electromagnetic radiation may cause biological damage by the deposition of energy in the body. They have the capacity to produce early acute effects and some have the ability to produce late effects, such as cancer. High doses can cause early death.

Ionizing radiation

Physicians are likely to become involved with radiation issues if people are over-exposed. For example, the theft of a caesium-137 radiotherapy source in Goiânia, Brazil led to 50 people being over-exposed, resulting in four deaths. In that incident large areas of land and property were contaminated. In the accident at the Chernobyl nuclear power plant 28 members of the workforce died of radiation-related injuries. The release of radioactivity also led to widespread contamination of houses, land, and foodstuffs. Experience shows that in these situations both individual and public health concerns dominate the agenda and place a burden on health professionals.

In the early twentieth century, due to a lack of knowledge about the harmful biological effects of ionizing radiation, many radiologists received high doses to their hands (often calibrating their machines by an erythema dose). Subsequently many died of radiation-induced cancers.

Atoms, radioactivity, and radiation

Some isotopes of elements are unstable and undergo radioactive decay. The time necessary to reach a stable form depends on the element and ranges from a few fractions of a second to several thousand years. The unit for radioactivity is the becquerel (Bq): 1 Bq equals one atomic disintegration per second. The average amount of natural potassium-40 (^{40}K) in every kilogram of the average person is 60 Bq. This means that about 15 million ^{40}K atoms disintegrate inside a person each hour.

Electromagnetic radiation (X-rays and gamma rays) or subatomic particles such as alpha and beta particles and neutrons can all interact with matter and tissues and cells. The different types of radiation penetrate matter, the extent of the penetration being determined by the size, charge, and energy of the particle/wave. Alpha particles are stopped by a thin piece of paper or the dead layer of the skin, while beta particles can penetrate the hand but will be stopped by a thin sheet of aluminium. X- and gamma rays penetrate the body and an aluminium sheet but are stopped by lead. Neutrons penetrate most materials but may be stopped by thick polythene or concrete (hydrogenous materials). These overall properties of radiation affect the degree of cellular damage following exposure and the methods needed for protection.

Ionizing radiation has sufficient energy to break chemical bonds and ionize atoms and molecules, producing ion pairs. These ions are charged and are capable of causing further ionization and energy deposition leading to physicochemical changes in cellular constituents. Some of these changes may be of no biological consequence and others may be repaired, but there is a finite probability that damage may cause cell death or irreparable damage to vital cell constituents such as the DNA.

Dose quantities

The absorbed dose is a measure of the mean energy absorbed by unit mass of tissue, and the absorbed dose in gray (Gy) is equal to the deposition of one joule (J) of energy in one kilogram (kg) of tissue. Overall, the greater the dose, the greater the likelihood of a biological effect being seen.

Various radiation and tissue weighting factors are used to convert the absorbed dose in gray to an effective dose in sievert (Sv). This system allows external and internal exposures to be combined into one dose—on the basis of equality of risk. Once a radionuclide is incorporated it will continue to expose surrounding tissues until it finally decays or is excreted.

Submultiples of the gray and sievert are commonly used, such as the milligray (mGy) and millisievert (mSv), which is one-thousandth of a sievert. For example, the world average individual dose received due to exposure to natural background radiation is about 2 mSv per year compared with the occupational dose limit of 20 mSv per year.

Acute radiation sickness

An acute exposure (measured in seconds, minutes, or hours) causes cell damage, the severity of which depends on the tissue type, or even death. It is characterized by a sudden sensation of anorexia or nausea and is soon followed by vomiting and sometimes diarrhoea. In the sensitive mucosal stem cells of the gastrointestinal tract, in particular in the stomach and small intestine, 5-hydroxytryptamine (**5-HT$_3$**) is released into the bloodstream which stimulates the nausea/vomiting centres in the brain and other 5-HT$_3$ receptors. There is a concomitant increase in bowel motility, which may be caused by bile salts acting on the damaged mucosa. The total spectrum of effects is dependent upon the dose but may be anorexia, nausea and vomiting, and diarrhoea. Radiation sickness extends beyond the early symptoms, as damage to other tissues, such as the bone marrow, causes lowering of blood cell counts leaving the body vulnerable to infection. High radiation doses can also lead to permanent sterility, serious damage to other organs, and to death (with or without medical treatment).

Many other symptoms may arise which depend on the dose, the dose rate, and the area of the body affected. These might include loss of hair, skin burns, or haemorrhages in the short term and increased incidence of cancer in the long term. These symptoms may vary due to individual susceptibility and because in most uncontrolled situations the nature of the exposure is non-uniform. In radiotherapy, where exposures are controlled and fractionated to enable normal sensitive tissues to recover between treatments, nausea and/or vomiting usually only occur where there is a high-dose total-body irradiation, for example in the ablation of bone marrow for subsequent bone marrow transplantation. It is normal to give antiemetics such as ondansetron (a 5-HT$_3$ antagonist) to reduce such side-effects.

At doses greater than 1 Gy significant reductions in blood cell counts follow depletion of the bone marrow, and can lead to reduced resistance to infections, haemorrhages, and anaemia. If there is significant direct exposure or surface contamination with radioactive materials skin burns may occur leading to further fluid loss and the danger of infection. The acute symptoms are sometimes grouped together and called the acute radiation syndrome. Combined injuries have a worse prognosis, and this is important in medical management. Without medical treatment an acute dose of approximately 4 Gy is likely to lead to death within 60 days in 50 per cent of those exposed. Doses in excess of 10 Gy are likely to result in earlier death, even with treatment. Similar doses over longer periods (days, weeks) may

cause a variety of symptoms, but death is unlikely as the cells and tissues have time to repair the damage.

The following figures are given as guidelines for adults: anorexia may be seen in 5 per cent of those at 0.4 Gy and 95 per cent at 3 Gy, nausea in 5 per cent at 0.5 Gy and 95 per cent at 4.5 Gy, vomiting in 5 per cent at 0.6 Gy and 100 per cent at 7 Gy, and diarrhoea in 5 per cent at 1 Gy and over 20 per cent at 8 Gy. If the time from exposure to onset of any of the above symptoms is less than 1 h the dose is likely to be more than 3 Gy, if more than 3 h the dose is likely to be more than 1 Gy, and if they last for more than 24 h the dose is likely to be more than 6 Gy. These general spectra can be helpful to physicians as an aid to medical triage before more refined estimates can be made.

Clinical treatment of radiation injuries

General

Experience from many radiation accidents has shown that the heterogeneous nature of the exposures tends to confound both the clinical and pathological picture so that estimating the scale of the radiation damage and exposure is difficult. The early treatment of conventional injuries is the main factor that determines survival in patients who have been accidentally irradiated.

Casualties on admission can be classified into four treatment categories: mild, moderate, severe, and lethal. Stating specific dose ranges for these categories is not possible, primarily because of the difficulty in converting an exposure to a meaningful tissue dose; however, equivalent whole-body dose ranges would be less than 2 Gy, 2 to 5 Gy, 5 to 10 Gy, and more than 10 Gy respectively. The most reliable prognostic guides for treatment in the early stage are the change in levels of blood cells, which relates to dose, and cytogenetics, which will give a good estimate of dose.

The degree of radiation-induced marrow aplasia (reversible or irreversible) may not be known for days because of the uncontrolled nature of the exposure and the likelihood that it was non-uniform and heterogeneous. However, assessment of the dose, while important, is secondary to the treatment.

Reliable triage and good clinical care based on comprehensive biological data will ensure the best chance of recovery provided that some critical stem cells survive the radiation exposure. There is a need for biological monitoring, such as absolute lymphocyte counts and cytogenetics, to assess the effective dose and the probability of survival.

Sepsis

The suppression of bone marrow activity will reduce the resistance to bacterial and viral infections. For febrile neutropenic casualties, blood, urine, and faecal samples for cultures need to be taken on admission. Patients must be started on broad-spectrum antimicrobials and these should be given until the patient is afebrile. Experience has shown that fungal lung infections can be the cause of late deaths when all the other radiation effects have been stabilized. Early use of antifungal agents and gammaglobulin for viral infections is essential.

Marrow aplasia

Following a suspected high radiation exposure, initial patient assessment should be based on dosimetric testing (biological, physical), daily full blood counts, viral titres (for cytomegalovirus and HIV for example), HLA subtyping for possible bone marrow transplantation, and the administration of haematopoietic growth factors (colony stimulating factors such as granulocyte or granulocyte/macrophage colony stimulating factors). Particular care is needed if the patient has other injuries, such as pulmonary infections, burns, or smoke inhalation.

Subsequent supportive therapy might include platelet transfusions, particularly if surgery is indicated for other injuries. Other treatments might include the use of thrombopoietic drugs. However, bone marrow transplants should only be given if an autologous donor is available.

Gastrointestinal

Vomiting should be controlled by use of effective antiemetics such as ondansetron. This relaxes the gut and reduces mechanical damage to the lining of the small intestine. Antiemetics also help to reduce fluid loss. Diarrhoea should be treated with fluids and electrolytes and efforts should be made to improve host defences by elemental diets, vitamins, and glutamine parenterally or intravenously.

Cutaneous

Remove all clothing as soon as possible, bathe very gently in lukewarm baths, use acetic acid or ion exchangers if surface contamination is thought to be soluble caesium, and only remove contamination mechanically from the soles of the feet or the palms of the hands.

Later treatments might include the use of topical creams, systemic acitretin, and gamma interferon. These have been found to be helpful in treating the chronic skin damage seen in the Chernobyl firefighters.

Combined injury

Radiation injury is not immediately life threatening; initial care should address the associated conventional injuries, for example thermal burns and wounds. After stabilization, radio-isotope decontamination should be performed before emergency surgery, definitive care, and treatment of radiation injuries. Collection of biological samples during the resuscitation stages will supplement the initial data collected during triage. Ideally, definitive care should immediately follow resuscitation.

Management of soft tissue wounds requires alternative ways to close wounds, for example biological wound coverings and skin grafts. Surgical correction of life-threatening and other major injuries should be carried out as soon as possible (within 36 to 48 h); elective procedures should be postponed until late in the convalescent period (45 to 60 days) following haematopoietic recovery. Treatment of thermal burns should include early excision of potentially septic tissue and closure of the wounds, preferably by skin grafting.

Counselling

Overexposure to ionizing radiations increases the risk of subsequent radiation-induced cancers. So it is important that over-exposed patients are counselled on the radiation risks in comparison with other risks in life. Experience from the Chernobyl accident has shown that many people have unwarranted fears from their radiation exposures due to poor communications and a lack of trust.

Non-ionizing radiation

Ultraviolet radiation

The main organs affected by ultraviolet radiation are the skin and the eye. There is evidence that ultraviolet radiation can also induce changes in the immune system but the significance of this to humans is not yet clear. The most serious health effects are the cutaneous malignancies.

The skin

Short-term effects

Short-term skin effects are sunburn, principally consisting of erythema and oedema, both of which may be very severe. In some people this sunburn is followed by increased production of melanin (a suntan). A suntan is not an indication of good health and only offers minimal protection against further exposure. It is a sign that

damaged skin is attempting to protect itself from further harm.

Long-term effects

The most serious long-term effect is the induction of cancer. The non-melanoma skin cancers are mainly basal cell carcinomas and squamous cell carcinomas. They are relatively common in white populations, although they are rarely fatal. The overall incidence is difficult to assess because of under-reporting. Malignant melanoma is the main cause of death from skin cancer, particularly in young people, although its incidence is lower than that of non-melanoma skin cancers. The risk of developing malignant melanoma has increased substantially in white populations for several decades, and the annual incidence in the United Kingdom is now approaching 10 new cases per 100 000. Chronic exposure to solar radiation causes photoageing of the skin, which is characterized by a leathery, wrinkled appearance and loss of elasticity. A small quantity of ultraviolet radiation is beneficial in terms of vitamin D synthesis in the skin.

The eye

Responses of the human eye to acute exposure to ultraviolet radiation include photokeratitis and photoconjunctivitis. Repeated exposure is also a major factor in the causation of non-malignant clinical lesions of the cornea and conjunctiva such as climatic droplet keratopathy, pterygium, and, probably, pinguecula. Epidemiological data on cataract formation in highly exposed people suggest that cumulative exposure to ultraviolet radiation is a principal causative factor in the development of, at least, cortical cataracts, although the extent to which this is an important risk factor for cataracts in the general population is unclear.

Immune responses

Biological studies have shown that exposure to ultraviolet radiation can suppress the normal antigen-specific immune response to some skin tumours and to various skin pathogens, although immunity acquired from prior infection is not affected. The significance for human health of ultraviolet radiation-induced immune suppression is not clearly established at present.

Radiofrequency electromagnetic waves

The possible hazards to humans of exposure to high-intensity radiofrequency electromagnetic waves are more controversial, but individuals chronically exposed to such radiation at high intensity have been reported as developing such conditions as anaemia, alopecia, or psychological disorders. In animal studies, changes in the high spontaneous rate of cancer in particular strains have been reported. The widespread use of radiofrequency microwaves (in mobile phones and ovens for home and restaurant cookery) has made their safety a more important question. With the exception of the danger from induction heating, with consequent damage due to thermal burns, there is no evidence that there is significant risk of danger to the general public from the use of such appliances. The National Radiological Protection Board has given advice that acceptable exposures should not result in a rise in body temperature of more than 0.5 °C as shown by skin and rectal temperature. Local tissue temperatures should not exceed 38 °C for the head, 39 °C for the trunk, or 40 °C for the limbs.

Magnetic fields

With the increasing use of magnetic resonance imaging interest has developed in the possible hazards to humans of exposure to strong and varying magnetic fields. Once again, the biological evidence is largely anecdotal and in part contradictory, and mainly involves transient psychological changes. The National Radiological Protection Board has recommended that for static magnetic fields the individual being imaged should not be exposed to a field greater than 2.5 tesla (T) to the whole or a substantial portion of the body. There is no evidence so far to suggest that the embryo is sensitive to magnetic fields and radiofrequency at the intensities encountered in clinical magnetic resonance imaging, but it is felt to be prudent to exclude pregnant women during the first trimester.

Summary

The man-made use of ionizing radiation in medicine and industry and for nuclear power has shown that high doses are lethal and early (deterministic) effects are visible and can be shorten life. Latent (stochastic) effects, particularly cancers, are well documented and risk estimates based on the study of occupationally and medically exposed cohorts are reasonably robust. Legislative dose limits will prevent deterministic effects and reduce the risk of stochastic effects to an acceptable level, although all doses should be kept as low as reasonably practicable. Information on the possible health effects of radiofrequencies and electromagnetic fields is less robust. However, the degree of risk acceptable to the public and workforce is likely to change.

Further reading

Berry RJ (1986). The radiologist as a guinea pig: radiation hazards to man as demonstrated in early radiologists, and their patients. *Journal of the Royal Society of Medicine* **79**, 506–9.

IAEA (International Atomic Energy Agency) (1988). *The radiological accident in Goiânia*. IAEA, Vienna.

ICRP (International Commission on Radiological Protection) (1991). *1990 Recommendations of the International Commission on Radiological Protection*, ICRP Publication 60.

International Atomic Energy Agency and World Health Organization (1998). *Diagnosis and treatment of radiation injuries*, Safety Report series no 2. IAEA, Vienna.

International Atomic Energy Agency and World Health Organization (1998). *Planning the medical response to radiological accidents*, Safety Report series no 4. IAEA, Vienna.

MacVittie TJ, Weiss JF, Browne D, eds (1996). *Advances in the treatment of radiation injuries*, Advances in the Biosciences vol. 94. Pergamon, Oxford.

Mettler FA, Upton AC (1995). *Medical effects of ionizing radiation*, 2nd edn. WB Saunders, Philadelphia.

NRPB (National Radiological Protection Board) (1999). Board atatement: advice on the 1998 ICNIRP guidelines for limiting exposure to time-varying electric, magnetic and electromagnetic fields (up to 300 Ghz). *Documents of the National Radiological Protection Board* **10**, no 2.

WHO (World Health Organization) (1996). *Health consequences of the Chernobyl accident*, WHO Scientific Report. WHO, Geneva.

8.5.10 Noise

R. McCaig and T. C. Aw

Introduction

Noise is one of the most prevalent hazards in the workplace, particularly where there is a large manufacturing sector. In developed countries exposure to noise remains a significant occupational health problem with well-established sources of exposure in engineering, printing, textiles, and other production processes being supplemented by those in entertainment and broadcasting, and, more recently, by the growth of telephone customer service centres, where workers are potentially exposed to noise from head sets. Although it has been known for over a century that excessive exposure to noise results in hearing loss, it is only within the last 20 to 30 years that comprehensive hearing conservation programmes have been adopted in industry. Noise-induced hearing loss, which is entirely preventable, remains a common occupational injury.

The vibration of objects in air propagates acoustic energy which impinges on the tympanic membrane. The subjective effect of this energy may be pleasant, as in speech or music, yet it retains the potential to be harmful at sufficient intensities. The term 'noise' generally refers to loud or harsh sound of any kind. The ear is sensitive to a very wide range of pressure change and, to accommodate this, a logarithmic scale is required. A reference pressure of 0.000 02 pascal (Pa), the threshold of hearing, is used and a rating in decibels is obtained from the formula:

$$\text{sound pressure} = 20 \log_{10} \text{measured pressurereference pressure decibel (dB).}$$

Thus a sound pressure of 0.2 Pa is equivalent to 80 dB, and one of 2.0 Pa to 100 dB.

A sound pressure level of 60 dB may be found in an office environment and levels between 90 dB and 100 dB in factories. Simple instruments are available to measure sound levels. However, the full characterization of a noise source requires more complex analysis, including the frequency spectrum. Most noise sources operate continuously but some produce intermittent peaks of exposure (impulse noise) arising from sources such as drop forging or the detonation of explosives. Separate criteria for the risk of damage apply to these two types of exposure.

In a normal young adult the ear is sensitive to a range of frequencies from about 20 Hz up to 20 kHz. The ear and brain together are not equally sensitive to all frequencies of sound, being markedly less sensitive at frequencies below 500 Hz and above 4 kHz. The more sensitive frequencies which lie in between 500 Hz and 4 kHz are those required for the perception of speech. Sound level meters contain filtering circuits which allow them to mimic the response of the human ear to sound. Different weightings are available, related to the subjective loudness of sounds. The one most commonly used is the A scale, reported as dB(A).

The effects of noise

Exposure to noise results in both physiological and pathological changes. An early response to acute noise exposure is an increase in blood pressure. More prolonged exposure to a sufficient intensity of noise causes a temporary shift in the threshold of hearing (temporary threshold shift). This may be experienced as a transient dullness of hearing and can be measured objectively. If exposure to noise is prolonged over months and years, a permanent threshold shift can result. Along with a reduction in the sensitivity of hearing, the person also experiences difficulty in differentiating similar sounds, particularly those of the consonants which are heard over a frequency range of 700 Hz to 4 kHz. With severe hearing loss there may be loudness recruitment; when sound intensity is increased beyond the now elevated threshold of hearing, a rapid and uncomfortable increase is experienced in the sound perceived. Noise-induced hearing loss may also be accompanied by tinnitus.

Excessive noise exposure causes damage to the outer hair cells of the cochlea, with breakage and disruption of the pattern of cilia on these cells which operate as local electromechanical amplifiers within the cochlea. When stimulated by oscillations of the basilar membrane at the appropriate frequency these cells contract and, by means of the cilia, pull down the overlying tectorial membrane. In doing so the inner end of the tectorial membrane is brought more closely into contact with the cilia of the inner hair cells, increasing their stimulation and that of the nerve fibres leaving the cells. Thus the outer hair cells assist frequency selectivity and amplification within the cochlea. Noise damage physically disrupts this amplification mechanism.

The 'non-auditory' effects attributed to noise include changes in both performance and health. Exposure to unwanted background noise, in particular speech, can disturb the performance of mental tasks. However, it has not been clearly established that environmental noise, for example around airports, results in an increase in psychiatric morbidity within the surrounding community. There is better evidence that environmental noise from road traffic is associated with an increase in cardiovascular morbidity but such an association has not been established in occupational settings.

Hearing loss

Permanent hearing loss is first detected as a reduction in hearing thresholds at the higher frequencies on the audiogram. Typically, this begins at 4 kHz and, with time, affects the adjacent frequencies creating a 'notch' pattern in the audiogram, described as the 4 kHz dip (Fig. 1). This sensitivity arises from the structure and configuration of the outer ear and the cochlea. With continuing exposure this hearing loss will extend to the lower frequencies and the thresholds will worsen (Fig. 2). Much of the damage is incurred during the initial years of exposure. Younger people may not be aware of hearing loss. It is in early middle age when the difference in hearing thresholds between those exposed to noise and the non-exposed becomes more obvious. Thereafter this difference is reduced, as presbyacusis (the reduction in hearing threshold with age) affects both groups.

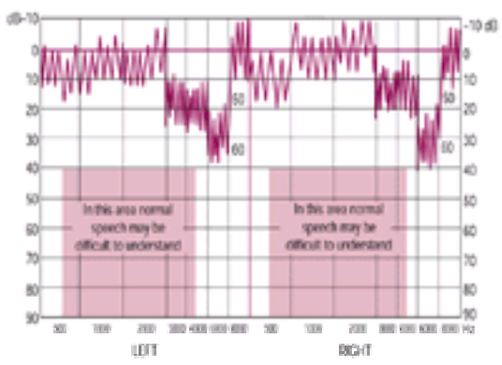

Fig. 1 Hearing loss after continuing noise exposure. Initially the lower frequencies are largely unaffected and there is recovery in the higher frequencies, appearing as a characteristic notching of the audiogram.

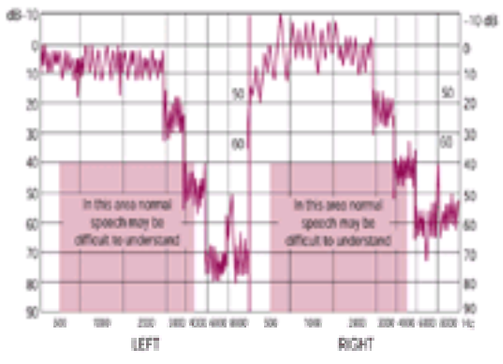

Fig. 2 With continuing noise exposure the extent of loss increases and extends into the lower frequencies.

The management of workers regularly exposed to noise should centre on the reduction of noise at source. A programme of screening audiometry should be introduced where daily personal exposure to noise (a time-weighted measure) is of the order of 85 to 90 dB(A), and in conformity with relevant national legislation and guidance. The technical requirements for industrial hearing conservation audiometry are specified in ISO 6189. Measures are normally taken to avoid temporary threshold shifts at the time of testing, with examinations repeated every 2 to 3 years.

Where abnormalities are detected, it is important to establish the history of occupational and leisure noise exposure, the use of hearing protection, exposure to ototoxic drugs and chemicals, and relevant past diseases and injuries, and tinnitus. Otoscopic examination, bone conduction audiometry, tympanometry to determine middle ear function, and evoked response audiometry may have a role in clinical assessment. Referral for a specialist opinion is relevant where pathology other than noise-induced hearing loss is suspected (such as the rare acoustic neuroma) and where it is thought that a hearing aid or tinnitus masker may be of help to the individual. In some countries noise-induced hearing loss results in the payment of compensation to workers and, where appropriate, individuals should be advised of their eligibility. Redress may also be available through the civil courts.

Hearing conservation

All workers exposed to noise should be included in a hearing conservation programme. This will include measures to control noise at source, the marking of remaining noisy areas, the use of hearing protectors, and the training of workers. In many countries this is required by national legislation. Where this is not the case, published damage risk criteria, for example ISO 1999, can be used to determine risk and the need for intervention. Control at source can be achieved, for example, by redesigning parts and machinery or by providing soundproof enclosures. This is always preferable to reliance on the use of hearing protectors, which must be fitted correctly and used continuously during the period of exposure to noise. This can be difficult to achieve in practice, particularly in the absence of good supervision. Hearing loss found on the audiogram can, however, be persuasive evidence to the worker of the need for improved hearing protection. Health professionals have a role in education in support of hearing conservation.

Further reading

Axelson A *et al.* (1996) *Scientific basis of noise-induced hearing loss*. Thieme, New York.

International Organization for Standardization (1983). *Acoustics—pure tone air conduction threshold audiometry for hearing conservation purposes*, ISO 6189. International Organization for Standardization, Geneva.

International Organization for Standardization (1990). *Acoustics—determination of occupational noise exposure and estimation of noise-induced hearing impairment*, ISO 1999. International Organization for Standardization, Geneva.

8.5.11 Vibration

Tar-Ching Aw and R. McCaig

Definition

Vibration is 'the mechanical oscillation of a surface around its reference point'. Workplace exposure to vibration results in local effects, mainly on the hands when the vibration is transmitted to the upper limbs, and as general systemic effects (mainly low back pain) when vibration is transmitted to the whole body. The clinical syndrome in the former has been termed 'vibration white finger' or 'hand–arm vibration syndrome'.

Exposure

Occupational exposure to whole-body vibration occurs in drivers of tractors, forklift trucks, mobile cranes, buses, lorries, and in helicopter pilots. The nature of the surface over which the vehicle is driven may be as important as the characteristics of the vehicle. Hand–arm vibration exposure occurs in factory workers involved in fettling, chipping, grinding, riveting, swaging, and using handheld pneumatic hammers, drills, chisels, and polishing and rotary tools. It also affects forestry, agricultural, and woodworkers using chain saws, miners drilling rock surfaces, and construction and road workers using drills and compactors. It has been estimated that around 3 per cent of the working population are occupationally exposed to sources of vibration.

Clinical effects

Whole-body vibration

Exposure to whole-body vibration causes physiological changes to the cardiovascular, respiratory, and musculoskeletal system. Clinical effects attributed to whole-body vibration include headache, motion sickness, sleep and visual disturbances, and urinary and abdominal complaints. The only effect with reasonable evidence of association with whole-body vibration is low back pain. In drivers, low back pain may occur as a result of vibration, poor posture within the vehicle cab, and/or other work duties such as loading and lifting.

Hand–arm transmitted vibration

This causes secondary Raynaud's phenomenon which manifests as frequent prominent episodic pallor of the digits, usually on exposure to cold. Patients frequently describe digital pallor in the morning, or following an outdoor activity such as fishing or gardening, especially in cold weather. The vascular changes may be accompanied by neurological and musculoskeletal effects which contribute to the disability experienced. Vascular and sensorineural effects may appear and progress independently. The latent period between initial exposure and development of symptoms is usually 5 to 10 years, although periods as short as 6 months or as long as 20 years have been noted depending on the intensity and duration of exposure. The sequence of colour change in the affected digits include pallor, a bluish hue due to cyanosis, and redness with spontaneous reversal of the vascular spasm.

Neurological effects include paraesthesia, reduced temperature perception, loss of manual dexterity, and pain. Severe tingling and discomfort often follow rapid warming of the hands. Loss of proprioception causing the inability to distinguish and hold small objects such as coins or to button up clothes contributes to physical and social disability. Musculoskeletal effects are not as well established, but muscle weakness, exostoses and cysts in the carpal bones, carpal tunnel syndrome, osteoarthritis, and Dupuytren's contracture have been associated with exposure to vibration.

Diagnosis

For a diagnosis of hand–arm vibration syndrome, the following criteria should be met:

a. there must be evidence of sufficient exposure to vibration;
b. episodic pallor of the digits and/or sensorineural effects should be confirmed;
c. other causes of Raynaud's phenomenon or sensory changes must be considered.

The presence of associated musculoskeletal features would add support to the diagnosis. Physical examination may show callosities on the hands, loss of sensation in the affected digits, and muscle weakness, although there may be no obvious abnormalities, especially in the early stages of the disease.

Various clinical and special tests have been used in the evaluation of patients with hand–arm vibration syndrome. These include digital blood pressure measurements, vibrotactile thresholds, sensory aesthesiometry, and cold provocation tests. The clinical and occupational history is of greater importance than the results of any of the various tests in the diagnosis of hand–arm vibration syndrome.

The differential diagnosis should include other causes for Raynaud's phenomenon (constitutional or secondary to rheumatoid arthritis, systemic lupus erythematosus, scleroderma, and other autoimmune disorders, cryoglobulinaemia, frostbite, and the thoracic outlet syndrome). Use of ergot, clonidine, and b-blockers, occupational exposure to vinyl chloride monomer, and heavy cigarette smoking may be contributory factors.

Management, treatment, and prevention

The severity of hand–arm vibration syndrome can be staged using the Stockholm Workshop Scale (Table 1). The grade of the disorder is indicated by the stage and the number of affected fingers on each hand; for example, '2L(2)/1R(3)' refers to two digits at stage 2 in the left hand; and three digits at stage 1 in the right hand.

Engineering controls can minimize the transmission of vibration from machinery to the body or hands. The patient may be able to continue in their job following such action. Where the condition is severe and the source of vibration cannot be eliminated, redeployment should be considered. In early cases redeployment may arrest or reverse the progression of symptoms. In severe cases the disease can progress regardless of removal from further exposure to vibration.

Advice to the patient includes avoidance or reduction of further exposure to vibration, use of appropriate gloves, keeping the body and hands warm, especially in cold weather, and cessation of cigarette smoking. Vasodilatory drugs such as tolazoline, inositol, and cyclandelate, and calcium antagonists such as verapamil and nifedipine, angiotensin-converting enzyme inhibitors, prostaglandins, and stanazolol have been tried with varying success.

A diagnosis of Raynaud's phenomenon should always include detailed inquiry into occupational exposure to vibration. The diagnosis of a case of hand–arm vibration syndrome should be viewed as a sentinel event warranting further investigation of the workplace to assess whether improvements in work practices can be implemented to prevent other cases from occurring.

Further reading

Griffin MJ (1997). Measurement, evaluation, and assessment of occupational exposures to hand-transmitted vibration. *Occupational and Environmental Medicine* **54**, 73–89.

Pelmear PL, Wasserman DE (1998). *Hand–arm vibration: a comprehensive guide for occupational health professionals*, 2nd edn. OEM Press, Beverly Farms, MA.

8.5.12 Disasters: earthquakes, volcanic eruptions, hurricanes, and floods

Peter J. Baxter

Deaths from natural disasters present a rising global trend. In the last half of the twentieth century, about 250 great natural disasters hit the planet, killing at least 1.4 million people and disrupting the lives of many millions more. At least 130 million people are affected each year by disasters and this number is rapidly increasing. Between 500 and 700 natural catastrophes are recorded every year and in the late 80s and 90s individual events entailed extremely large economic losses. Most human and economic losses have been from earthquakes, windstorms, and floods (Fig. 1). Global changes responsible for the worsening impact of disasters include the continuing, usually unplanned, expansion of populations into more and more exposed areas and the environmental degradation that is increasing the vulnerability of settlements, especially in heavily urbanized areas. This reckless development is going on throughout the world, even in areas of well-known risk.

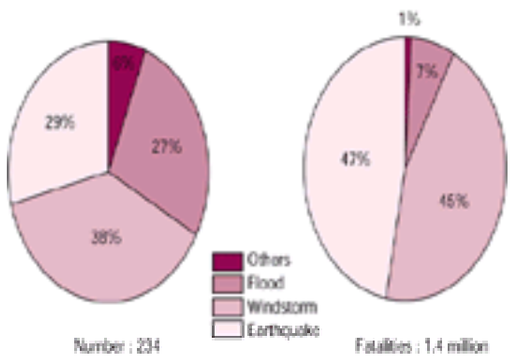

Fig. 1 Great natural catastrophes 1950 to 1999

Climate forecasts of an increase in the world's average temperature of 1.0 to 3.5°C and warming of the oceans could increase the potential for more intense hurricanes over wider areas and, in temperate regions, more severe wind storms and fluctuations in rainfall (floods and drought) and rising sea levels, increasing the frequency of severe coastal floods.

Predisaster measures

Disaster is the result of a vast ecological breakdown in the relation between humans and their environment, a serious and sudden (or slow, as in drought) event on such a scale that the stricken community needs extraordinary efforts to cope with it, often with outside help or international aid. Accurately forecasting the timing and size of natural disasters is rarely possible. This places major constraints on preventing loss of life by timely evacuation of people from the areas at risk. Disasters are quite different from major incidents in that normal lifelines and infrastructure usually break down. But despite their chaotic elements, disasters are amenable to scientific study and communities can develop planning and preparedness measures to reduce their impacts. Health workers have a key role in hazard management in risk assessment, predisaster planning and preparedness, as well as the emergency response.

Earthquakes

Many parts of the world are known to be at risk from devastating earthquakes, but it remains impossible to predict where and when a quake will strike. Most deaths and injuries are caused by collapsing buildings. Secondary causes such as fires can be important. Timber, masonry, reinforced concrete, and other types of buildings inflict injuries in different ways and to different degrees of severity when they collapse. In masonry buildings, an important cause of death is often suffocation from the weight and dust from the wall or roof material which buries the victims. The impacts of falling masonry cause crush injuries to the head and chest, external or internal haemorrhage, and chest compression (traumatic asphyxia). Little is know about the survival time of people trapped in collapsed buildings, but most victims die immediately or within 24 h, depending upon such factors as the severity of aftershocks, fire outbreaks, and rainfall. Rapid extrication of survivors and application of first aid by the uninjured immediately after the event could save up to 25 to 50 per cent of injured victims. The greatest demand for emergency medical services is within the first 24 h and the need for emergency treatment fades after 3 to 5 days. Causes of delayed death include dehydration, hypothermia, crush syndrome, and postoperative sepsis. Most of those requiring medical assistance suffer minor injuries such as lacerations and contusions.

A great earthquake struck Kobe City in Japan on 17 January 1995. It exposed serious flaws in the Japanese emergency services, as well as showing the need for comprehensive planning for medical care and welfare in earthquake-prone countries. The scale of disaster was beyond all expectations. The immediate victims included 5520 dead (10 per cent killed in fires) and 41 527 wounded. Four days later, 342 000 people (10 per cent of the total population of the region) were staying in shelters at a time of subzero temperatures. The plight of the elderly was not appreciated, as many of their carers were either dead or surviving as victims themselves. About 80 per cent of hospitals in the area were damaged, 7 per cent completely destroyed. Undamaged hospitals lacked the structural means to adapt rapidly to deal with overwhelming numbers of casualties requiring emergency care; the breakdown of supplies and water, electricity, and gas and failure of communications badly hampered other hospitals that were still functioning. The transfer of pharmaceuticals and medical equipment took longer than expected. The economic losses exceeded 100 billion US dollars, making it the costliest natural disaster in the world to date.

Volcanic eruptions

About 500 to 600 volcanoes around the world are known to be capable of eruptive activity and several major eruptions occur every year. The vast majority of volcanoes are explosive and unpredictable in behaviour, providing little opportunity for people to escape unless full evacuation measures are taken as soon as premonitory signs develop. This is in contrast to the less common lava flow eruptions, where people can normally escape by the time the lava flow heads towards them. Most deaths and injuries in explosive eruptions (such as engulfed ancient Pompeii) are caused by pyroclastic flows and surges, which are clouds of hot ash and gas that can travel at hurricane speeds. Survival is uncommon but victims will have extensive and severe skin burns, as well as inhalation injuries. The worst volcanic disaster in the twentieth century was at St Pierre, Martinique, in 1902, when 28 000 people were killed in a laterally directed pyroclastic surge. One of the most dangerous volcanoes in the world is Vesuvius, Italy, where uncontrolled building has resulted in over one million people living in an area which could be devastated by pyroclastic flows in a new eruption. Another major cause of death is lahars or wet flows of debris, either ash that has built up on the slopes of the volcano or unstable masses that are mobilized during the eruption or by rain, or rarely by release of water from a crater lake. The eruption of Ruiz del Nevada volcano in Colombia in 1984 triggered a lahar (mud flow) by rapid melting of the glacier at the summit, the melt waters rushing down valleys and mixing with debris as they went. Although adequate warning could have been given to the people below, lack of preparedness meant that the lahar engulfed towns including Armero, killing around 24 000 people. In one of the largest eruptions of the century, at Mount Pinatubo in the Philippines in 1991, 50 000 people were successfully evacuated from the threat of pyroclastic flows, but over 300 people died while sheltering in their homes from the collapse of roofs due to the accumulated weight of rain and ash.

The eruption of the Soufriere Hills volcano on the tiny Caribbean island of Montserrat began in July 1995 and gradually escalated, forcing the evacuation of thousands of people from their homes because of the threat of pyroclastic flows. By 1997, these flows had devastated the southern part of the island, evicting three-quarters of the population of 12 000 people. Air pollution from volcanic gases and ash has been a major consideration because of the close proximity of the

population to the volcano and the frequent eruption of fine, respirable ash containing hazardous amounts of the crystalline silica mineral, cristobalite, which can cause silicosis.

Hurricane and windstorm

Hurricanes are one of a broad class of extreme weather phenomena that include winter storms (snow, sleet, freezing rain, and freezes), thunderstorms (e.g. tornadoes, heavy rains, lightning, wind, and hail), extreme precipitation (e.g. flood and flash floods), and windstorms. Hurricanes (or typhoons as they are called in the Western Pacific) are tropical cyclones that form over warm oceans with ocean surface temperatures over 26°C. Once over land they soon run out of energy and rapidly abate, but can still cause flooding from heavy rain. Very high wind speeds, up to 250 km/h, are restricted to a relatively narrow track, usually no more than 150 km wide, within which localized gusts may even achieve tornado speeds and be extremely destructive. However, most deaths and injuries are not from the effect of wind on people (who normally remain inside for protection) or from building damage (building collapse or being struck by flying debris). Instead, deaths and injuries are very commonly the result of flooding from the sea surge as the hurricane strikes land, concurrent heavy rainfall (typically up to 60 cm, over a larger area and extending further inland than wind speed) and resulting landslides. Hurricanes lift the sea, forming a sea surge that typically rises 3 to 4 m above existing tides, with the wind generating waves on top of these.

Hurricane Andrew, which struck Florida in August 1992, was the most destructive natural disaster in the history of the United States and the third most intense storm to strike the United States in the twentieth century. Andrew developed sustained wind speeds above 250 km/h, even greater in localized tornado-like gusts. Mitigation measures (evacuation of three-quarters of a million people) held the death toll at 43. The most destructive effects of storm surge are on beaches and offshore islands, but in low-lying regions these can extend as far as 40 km inland. A dome of water lifted by the hurricane may be up to 100 km wide on open coasts. Such a surge drowned at least 200 000 people in Bangladesh in 1970. During a 2-week period in October/November 1998, Hurricane Mitch battered Central America. It unexpectedly slowed to a tropical storm and years of rainfall fell on Central America within a few hours. The result was about 13 000 slope failures and landslides, adding to the destruction of extensive flooding. Over 20 000 people were killed and over two million people left homeless.

Over 90 per cent of fatalities in hurricanes are drownings associated with storm surges or floods. Other causes of death include burial beneath houses collapsed by wind, penetrating trauma from broken glass or wood, blunt trauma from floating objects or debris, or entrapment in mudslides. The number of severe trauma patients among survivors is usually small; the greatest need in the postimpact phase is the provision of adequate shelter, water, food, and clothing, and sanitation. Most suffer from lacerations caused by flying glass or other debris, or minor trauma such as closed fractures and puncture wounds. After Hurricane Mitch the incidence of cholera and other water-borne diseases (leptospirosis, dengue, and malaria) increased, particularly in the urban areas.

Floods

In addition to the major losses of life that can be caused by hurricanes and their associated sea surges, floods mostly result from moderate to large events (rainfall, snow melt, high tides) occurring within the expected range of stream flow or tidal conditions. In Britain, as in many countries with low-lying coastal land, the hazard of coastal flooding from sea surges and high tides dominates over river flooding, although the latter is more frequent. Floods have been particularly devastating in China, with such notable historic disasters as the flooding of the Hung Ho in 1887 and 1931 claiming 900 000 and 3.7 million lives respectively. Over 40 million people are estimated to be affected yearly by flood. Flood warning and forecasting, combined with effective land management, community preparedness, and evacuation planning, are as essential as engineered river and coastal defences.

The primary cause of death from floods is drowning, but trauma from impacts with floating debris and hypothermia due to cold exposure is also important. The portion of survivors requiring emergency medical care is small, most injuries being minor, such as lacerations. This absence of victims with severe or multiple trauma is likely to reflect the long delay in reaching survivors, so they die from their injuries or from exposure before search and rescue teams can arrive. Increased morbidity and mortality in survivors who experience flooding was reported in the year after the East Coast Flood in 1953 and a river flood in Bristol in 1968; an increase in suicides and mental health problems were found in the aftermath of severe flooding caused by heavy rains in central Europe in July 1997.

Postdisaster relief

Myths surrounding postdisaster relief include:

- any kind of international assistance is needed;
- the affected population is too shocked and helpless to take responsibility for their own survival;
- natural disasters trigger secondary disasters through outbreaks of communicable diseases;
- life gets back to normal after a few weeks.

Most deaths in sudden-onset disasters happen long before outside aid arrives. However, relief teams have an important role in restoring roads and bridges, bringing in potable water, ensuring solid waste management, food protection, vector control, and sanitation. Attendances at medical facilities may return to normal within a few days of a disaster, and restoration of primary care then becomes a top priority rather than emergency treatment. Epidemiology has an important role in postdisaster assessment and health surveillance, particularly where relocation of large populations has occurred, as well as investigating the causes of mortality and morbidity in disasters, including mental ill health and long-term health sequelae.

Further reading

Coburn A, Spence R (1992). *Earthquake protection*. Wiley, Chichester.

Lumley JSP, Ryan JM, Baxter PJ, eds (1996). *Handbook of the medical care of catastrophes*. Royal Society of Medicine, London.

Noji EK, ed. (1997). *The public health consequences of disasters*. Oxford University Press, New York.

Pielke RA Jr, Pielke RA Sr (1997). *Hurricanes: their nature and impacts on society*. Wiley, Chichester.

Smith K, Ward R (1998). *Floods: physical processes and human impacts*. Wiley, Chichester.

9 Principles of clinical pharmacology and drug therapy

*Andrew Herxheimer**

Clinical pharmacology is the application of scientific principles to understanding the ways in which drugs behave and work in humans. Such understanding underlies rational drug therapy. Every decision to prescribe or use a drug comprises a series of separate decisions that are usually taken at great speed and often without enough thought. The main questions that must be decided are:

1. Does this patient's problem require a drug? What is likely to happen if no drug is used?
2. If a drug is needed, what type of drug action would be most appropriate?
3. Which class of drug will best provide the desired action?
4. Which particular drug in that class should be chosen, and in what pharmaceutical form?
5. What dose is to be used, how often, and at what times of the day or in what circumstances?
6. For how long is the drug to be used, initially and in the long term?
7. Will the cost of the drug matter to the patient?
8. How will the drug treatment fit into the overall treatment plan?
9. Will it be necessary to review the treatment? If so, when, how, and by whom?
10. What will the patient need to understand about treatment? Who will communicate this, and how?
11. What other help may the patient need to use the treatment optimally?

In a consultation all these questions cannot be considered from scratch. The physician must therefore answer as many of them beforehand as possible. This can be done with the help of appropriate therapeutic guidelines and a good formulary (which is really a specialized collection of guidelines concerning mainly one aspect of treatment, the choice of drug). Guidelines exist to provide thoroughly prepared decision paths for the common problems. The time and effort saved by using them enables physicians to focus on those aspects of an individual patient and problem that require individual adaptation of or departure from the relevant guideline. A guideline is like the trunk and large branches of an easily climbed tree—a decision tree. Once in the crown one is free to decide which small branch or leaf is the best. Many of the questions listed can only be answered satisfactorily after discussion between patient and doctor. For example, the patient's attitude to and feelings about taking medicines often determines whether or not to use a drug; the formulation chosen is influenced by the patient's needs or preference, for example difficulty in swallowing tablets or aversion to suppositories; the dosage regimen must be practicable for the patient. The best treatment decisions build on the patient's as well as the clinician's knowledge and experience. The clinician therefore has to find out about patients' experience of treatments—this should always be part of taking the history, but it is often forgotten.

The treatment history gives a necessary background for prescribing decisions because it can indicate how effective or ineffective previous treatment has been, and whether previous treatment has caused problems that should be avoided in the future. It can warn of possible interactions. It will show what the patient knows and believes about drugs and other treatments, as well as what he or she does not know about his/her treatment, and would like to know. Areas to ask about include:

- medicines currently being used, including dosage
- previous hospital treatment
- previous treatment from the family doctor
- 'alternative' treatments such as herbal or homeopathic medicines, acupuncture
- self-prescribed medicines
- past bad experiences with medicines, fears or worries about medicines.

Benefit and harm in prescribing

A doctor who prescribes a drug does so expecting that it will help the patient, and believing that any disbenefits, which may range from minor inconvenience or discomfort to a chance of serious harm, are acceptable. Such judgements are often referred to as assessments of the 'benefit/risk ratio', but this is a meaningless and confusing term. Risk means 'probability of harm', whereas benefit usually means 'amount of benefit'. Both benefit and harm have two dimensions, magnitude and probability, which must be clearly separated. Unfortunately we have no word for 'probability of benefit'.

The data that we have on benefits differ greatly from those used to estimate harm. Knowledge of the benefits to be obtained from a drug is almost always greater than knowledge of the harm that it may do. This is because the initial clinical research on a drug is done explicitly to assess the hoped-for benefits, which must occur in a high proportion of patients if the drug is to be useful, whereas harmful effects, especially unexpected ones, cannot be looked for in clinical trials—which are anyway too small to detect uncommon ones. Serious harm tends to affect relatively few patients and often is not suspected or discovered until a drug has been widely used for at least several years. As a result much less is known about the adverse effects of new drugs than about their benefits, and knowledge about harm accrues remarkably slowly in the first few years of a drug's life. If the drug is widely used, more of the iceberg gradually becomes visible. Because of the unknown risks of new drugs, it is better to choose older established drugs unless there is a clear reason for preferring a particular new drug.

Another dimension that must be considered is time: benefits are typically seen quickly, many adverse effects may not occur for years. That makes it even harder to weigh benefits against harm. Often they just cannot be compared. Two examples illustrate the problem.

First, in the case of pneumococcal pneumonia treated with penicillin, a cure is known to be likely and adverse effects are usually few and minor. However, if the patient is known to be allergic to penicillin, the risk of a serious adverse effect becomes too great and the benefit no longer outweighs the risk; in such a case another

antibiotic would be used. In this example the balance of benefit and harm is easily determined.

In contrast, consider the treatment with warfarin of a patient with atrial fibrillation and a history of bleeding peptic ulcer. Which outweighs the other: the potential benefit of preventing an embolic stroke by using the anticoagulant, or the risk of causing serious gastrointestinal bleeding? These two are not readily comparable, but a decision has to be made. It may be decided to use warfarin and hope to prevent serious bleeding with a proton pump inhibitor such as omeprazole, arguing that the degree of morbidity from an embolic stroke is likely to be more serious than that from an acute gastrointestinal bleed, with an approximately equal risk of mortality. But the patient may strongly dislike having to attend for regular monitoring of the anticoagulant effect (international normalized ratio, INR). There is no right answer.

The balance of benefit and harm from a particular treatment depends on four major factors:

1. the seriousness of the illness;
2. the effectiveness of the treatment;
3. the seriousness of possible adverse effects;
4. the risk of possible adverse effects.

If a disease is life threatening, if the drug to be used is highly effective and the only one available, and if the risk of serious adverse effects is negligible, the balance clearly favours the treatment. At the other end of the spectrum, if the disease is trivial, if the treatment is not very effective and more effective and safer alternatives exist, and if the risk of serious adverse effects is high, the likely harm outweighs the benefit. Table 1 illustrates this spectrum. Most cases lie somewhere between these two extremes.

The example of phenylbutazone, a highly effective non-steroidal anti-inflammatory drug, illustrates the principles involved. Use of the drug has a risk of marrow aplasia of between 1 in 30 000 and 1 in 100 000. While no other equally effective drugs existed, the benefit from phenylbutazone was considered great enough to outweigh the relatively high risk of marrow aplasia. However, once other equally effective and safer drugs became available, the risk of its adverse effects was seen to outweigh whatever benefit it gave, and its use was restricted.

Note that in making this decision to restrict the use of phenylbutazone the benefit was considered not in absolute terms but in relation to the severity of the disease and the benefit obtainable from other drugs. While phenylbutazone was more potent therapeutically than other available drugs it was considered highly beneficial, but when equally effective drugs arrived its net benefit was perceived to be less, although the therapeutic effect of the drug itself had not changed during that time. Phenylbutazone is still used today in some cases of severe ankylosing spondylitis that have responded poorly to other anti-inflammatory drugs. In those cases doctors and patients seem to believe that the benefit outweighs the risk of marrow aplasia.

Efficacy, effectiveness, efficiency of drugs

These terms need to be distinguished. Efficacy refers to any arbitrarily chosen effect, which may or may not be clinically relevant. It is best avoided in clinical contexts. Effectiveness means clinical effectiveness, and is defined as the likelihood and extent of the desired effects on the patient.

Efficiency is the ratio of effectiveness to cost: it is clearly more efficient to use the cheaper of two equally effective, equally safe, and equally reliable drugs.

The therapeutic index of a drug

The therapeutic index of a drug is the ratio of the dose at which adverse effects become important to that at which a therapeutic benefit is expected. This index does not have a numerical value, as precise estimates of these doses are not available and vary from person to person. Instead, drugs are divided into two categories, those with high and low therapeutic indices. For example, in the absence of hypersensitivity, penicillins have a high therapeutic index; very large doses can be given without fear of adverse effects. In contrast, digoxin has a low therapeutic index; it takes very little more to cause toxicity than to produce therapeutic benefit. Drugs with a low therapeutic index include the aminoglycoside antibiotics, anticoagulants, anticonvulsants, antihypertensives, some antiparasitic and antiviral drugs, cardiac glycosides, cytotoxic and immunosuppressant drugs, oral contraceptives, sympathomimetics, and drugs that act on the brain.

The benefit for a particular drug is more likely to outweigh the risk of harm if the drug has a high therapeutic index. In choosing between two equally effective drugs, one would choose the one with the higher therapeutic index.

Formularies

Formularies are lists of medicines for prescribers and pharmacists, intended to guide the choice and facilitate the dispensing of medicines. Many give details of the formulation and doses of drugs. Each formulary is produced primarily for a particular group, usually the prescribers in one country or region or institution, or even one practice. Most formularies are restrictive, that is they make a more or less narrow choice of medicines from all those available that could be used. This is typical for the formulary of a hospital, or of a health maintenance organization. A hospital formulary lists only the preparations that are stocked in the hospital pharmacy; a health maintenance organization formulary only those that the organization will pay for. The British National Formulary, probably the best known and most widely used formulary of all, differs in this respect, for it includes all medicines available for prescription in Britain, whether they are good choices and recommended or not. However, in every section concise and critical 'notes to facilitate the selection of suitable treatment' precede the list. These notes help prescribers everywhere (not only in Britain) to think, and add greatly to the value of the book, especially since it is updated twice a year and accessible electronically (http://bnf.org/).

The WHO 'Model list of essential drugs'

In many poor countries large sections of the population have no access to drugs or health care, and governments cannot afford to provide necessary drugs. To help them to use their limited funds in the best ways, the World Health Organization has since 1978 published a regularly updated 'Model list of essential drugs'. Essential drugs are defined as 'those that satisfy the health care needs of the majority of the population; they should therefore be available in adequate amounts and in the appropriate dosage forms'. The WHO list is called a model list because it cannot be used as it stands but must be adapted to the needs and priorities of individual countries. That is not easy, and WHO offers governments help with it and related work. Many developing countries now have an essential drugs programme.

The principles of clinical pharmacology

Drug therapy can be considered under four headings—pharmaceutical, pharmacokinetic, pharmacodynamic, and therapeutic—the four major processes of clinical pharmacology, each of which is associated with a question about drug therapy (Table 2, Fig. 1).

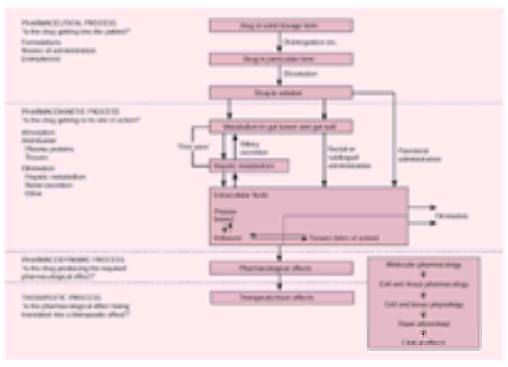

Fig. 1 The four processesses of clinical pharmacology in relation to drug therapy.

The pharmaceutical process

The pharmaceutical process concerns the question 'Is the drug getting into the patient from the appropriate formulation?'

Usually there is no need to worry about the formulation of a drug, beyond deciding whether to use the oral or parenteral route. For example, oral frusemide (furosemide) or bumetanide are prescribed as tablets. But sometimes the decision is harder. For example, in Britain six different oral lithium formulations exist, differing in strength and absorption characteristics. Glyceryl trinitrate is formulated for use orally, sublingually, buccally, transdermally, and as a spray. To understand the differences it is necessary to understand the concept of systemic availability.

Systemic availability

Systemic availability (commonly called bioavailability) is the proportion of administered drug that reaches the systemic circulation and is available for distribution to the site of action. The term usually refers to formulations given orally, but it can also refer to other routes of administration, such as intramuscular or transdermal.

When a drug is given intravenously it all enters the systemic circulation (systemic availability = 100 per cent).

When a drug is given orally in solution all the drug is immediately available for absorption. How much of it enters the systemic circulation will depend on the extent of absorption and the metabolic effect of the liver as it passes through for the first time (the so-called first-pass effect).

When a drug is given orally as a tablet or capsule the amount that enters the systemic circulation also depends on the properties of the formulation, including how fast the tablet disintegrates and how fast the drug particles dissolve in the intestinal fluid—the 'pharmaceutical availability'.

Prescribers rely on the pharmacist and pharmaceutical chemist to provide formulations of high stability and predictable pharmaceutical availability. When different formulations of the same drug are available the doctor may have to decide which to prescribe. This may be particularly important for drugs with a low therapeutic index, for which differences in pharmaceutical availability will matter. For example, if a patient is taking a formulation of low pharmaceutical availability that is producing a good therapeutic effect, switching to one of high availability can cause adverse effects.

Special drug formulations

Most drugs are given orally; oral formulations include elixirs, ordinary (quick-release) tablets and capsules, and modified-release formulations. However, drugs may be given by other routes, including sublingually, buccally, rectally, transdermally, by inhalation, and by injection intravenously, subcutaneously, intramuscularly, or locally.

Modified-release formulations

'Modified-release' formulations are oral formulations with some special release mechanism. Most are intended to prolong the duration of action of a drug and to smooth its effects by gradual release during the dosage interval. Examples include formulations of theophylline, nifedipine, diltiazem, and lithium. Prescriptions of these drugs should specify the exact formulation, as formulations differ in systemic availability.

Sublingual, buccal, rectal formulations

Drugs that are absorbed through the oral or rectal mucosa enter veins and pass into the systemic circulation intact, avoiding first-pass metabolism in the liver. For example, sublingual glyceryl trinitrate is effective in doses about 10 times less than those required by mouth and the effect is rapid. Rectal administration achieves a direct effect on the large bowel (for example corticosteroids in ulcerative colitis).

Transdermal formulations

Some drugs are well absorbed through the skin, and their transdermal administration via 'patches' allows controlled release of small amounts over many hours. Examples are glyceryl trinitrate in the long-term treatment of angina pectoris, transdermal hyoscine for travel sickness, estradiol as hormone replacement therapy, and nicotine for stopping smoking.

Inhalations

Inhaled formulations come in several forms, with different intentions. Sodium cromoglicate is in a powder for inhalation, designed to act locally on the bronchioles in the management of asthma. Salbutamol aerosol, on the other hand, is designed to produce bronchodilatation by a metered dose (100 µg) of droplets, small enough (2–5 µm) to reach the bronchioles. About 10 per cent of any inhaled drug reaches the bronchial tree, the rest being lost in the air, absorbed from the oropharynx, or swallowed. Nebulizers provide continuous administration of aerosolized drugs for short periods.

Subcutaneous, intramuscular, and local injections

The rate of absorption of insulin from the site of subcutaneous injection is controlled by its physical state (for example crystalline or non-crystalline), its zinc or protein content, and the nature and pH of the buffer in which it is suspended. Thus, soluble insulin has a rapid onset and short duration of action (about 6 h), while ultralente insulin, which has large crystals and a high zinc content, begins to act at about 7 h and acts for about 36 h.

Drugs are sometimes absorbed erratically from intramuscular injection sites; for phenytoin and diazepam the intramuscular route should be avoided if possible. Absorption may be retarded by the use of thick oils, which slow down diffusion of the drug from the site of injection and retard intramuscular absorption—for example vasopressin tannate in oil used in diabetes insipidus, and fluphenazine decanoate in oil in schizophrenia.

Some local anaesthetic injections contain adrenaline, which by vasoconstriction prevents rapid removal of the drug from the injection site.

Combination formulations in oral therapy

Combination products are seductive, but should be used only when at least two criteria are met:

- the frequency of administration of the two drugs is the same;
- the fixed doses in the combination product are therapeutically and optimally effective in most cases (i.e. when it is not necessary to alter the dose of one drug independently of the other).

Acceptable combination products include:

- Aspirin plus codeine (cocodaprin) or paracetamol plus dihydrocodeine (codydramol), pairs of drugs that have different analgesic actions (which summate) and different adverse effects (which do not).
- Levodopa plus a peripherally-acting dopa decarboxylase inhibitor (benserazide or carbidopa); the peripheral action of the decarboxylase inhibitor blocks peripheral metabolism of levodopa, which is free to enter the brain, where it is converted to the pharmacologically active product dopamine, producing the therapeutic effect in Parkinson's disease.
- Combined oral contraceptives, which contain an oestrogen and a progestogen.
- Ferrous sulphate plus folic acid, used to prevent anaemia in pregnancy.
- Coamoxiclav (amoxicillin plus clavulanic acid); the b-lactamase inhibitor, clavulanic acid, prevents the breakdown of amoxicillin by bacterial penicillinase, so

broadening its spectrum.

The patient's use of the medicine: compliance and concordance

Compliance is used as a technical term for the extent to which the patient follows a prescribed regimen. However, the word has unwelcome overtones, implying orders that the patient is expected to obey. The word 'concordance' has been introduced to supplant it, to make it clear that therapeutic decisions are best arrived at jointly by prescriber and patient agreeing on what to do. If both understand and accept the reasons for a particular choice, then the patient will be more committed to it than if a choice is imposed without explanation or discussion. This partly accounts for the wide variation in 'non-compliance', which in different studies has been found to be as low as 10 per cent and as high as 90 per cent. Compliance is affected not only by the behaviour of patient and doctor, but also by the nature of the treatment and the type of illness.

The nature of the treatment

Apart from the cost to the patient (which matters greatly in some countries), two main aspects of the treatment itself determine compliance, the complexity of the regimen and adverse effects. The complexity of the prescribed regimen involves two factors: the frequency of administration (the more often during the day patients have to take a drug the less likely they are to take it) and the number of drugs prescribed (the more drugs prescribed the less likely is overall compliance). It is difficult to take several medications at different dosages that involve different numbers of tablets at different times of the day.

If patients experience symptoms that they attribute to adverse drug effects (for example diarrhoea due to antibiotics or sedation due to anticonvulsants), then, unless they can be persuaded that the likely benefits of treatment outweigh the disadvantages, they will stop taking the medicine. Some patients, notably children and the elderly, have problems with certain formulations, for example sickly elixirs or large, dry, bitter tablets.

The type of illness

People who are severely mentally disturbed, for example patients with schizophrenia or manic-depressive psychosis, often take medicines unreliably.

Physical disability may cause difficulty even in patients who want to take their medicine. For example patients with rheumatoid arthritis who cannot reach the tablets or cannot remove the top of a child-proof container, cannot take them without help.

Sometimes a good response to treatment leads patients to stop. For example, patients with tuberculosis need long courses of several drugs to eradicate the infection; motivation may wane once the symptoms have resolved, risking reactivation and the emergence of resistant bacilli.

Some diseases may promote compliance. Patients with insulin-dependent diabetes easily become very ill quite quickly if they forget to take their insulin, and that is likely to make them comply, although they may not use it precisely as advised. Patients in whom a b-blocker or vasodilator has largely prevented anginal attacks will be conditioned to good compliance.

The patient's behaviour

People tend to forget to take medicines, or can't be bothered; they may feel no need for treatment (for example in asymptomatic hypertension); they may be unclear about the prescribing instructions; they may not want to feel dependent or be thought to be dependent on 'drugs'. There may be social or physical reasons why they cannot reach a pharmacist, financial difficulties, or everyday inconveniences in carrying and taking the medication. (See www.dipex.org/hypertension, for interviews in which patients talk about their tablets.)

The doctor's behaviour

The enthusiasm and confidence with which a treatment is prescribed, and the extent to which these attitudes are transmitted to the patient, may influence not only compliance but also the response to therapy. This is partly related to the placebo effect.

Methods of assessing compliance

It is important to assess compliance both in everyday practice and in clinical trials. The most obvious and usually the easiest approach is to ask the patient whether he or she has been taking the drugs, and whether there have been any problems. If the doctor is non-judgmental and indicates that difficulties are common, the patient is encouraged to be open.

Less directly one can ask to see the patient's tablets: this confirms that the prescription has been cashed. Counting the tablets left in the bottle is a guide to how many have been taken, but some may have been thrown away. Recording devices fitted in the caps of medication containers can record the frequency and exact timing of the opening of the container, and are useful in research.

When a patient is vague or untrustworthy, measurement of some compounds in the plasma or urine may give a good indication of compliance, but this does not allow for patients who take their treatment only on the day they visit the doctor. The compound measured may be the drug itself (for example phenytoin or lithium in the plasma, or salicylate in the urine) or a marker (for example riboflavin, which is easily detected in the urine).

Detecting the pharmacological effect of a drug can also give evidence of compliance; for example, the response of the heart rate for b-blockers, the prothrombin time for oral anticoagulants, and the reticulocyte count for haematinics. Failure to detect the pharmacological effect of a drug implies non-compliance or inadequate dosage. During long-term antibiotic therapy for urinary tract infection, antibacterial activity in the urine can be measured.

Obviously, if the desired therapeutic effect occurs the question of compliance is unimportant. However, a good therapeutic outcome may occur irrespective of the treatment used. It would be wrong to attribute a good outcome to the effect of a drug that the patient may not have taken; this is an important principle in clinical trials.

Methods of improving compliance

Compliance can be improved by supervised administration of the drug, by removing barriers to compliance, by simplifying the therapeutic regimen, and by educating the patient on the need to take the medicine, with reminders when possible.

Supervised administration

Administration of a drug by the doctor or nurse ensures compliance. This is possible in hospital or when only occasional administration is required (for example intramuscular injections of vitamin B_{12}, long-acting depot injections of neuroleptics in the treatment of schizophrenia, and supervised twice-weekly antituberculosis therapy). Sometimes a single dose of a drug can be given at the time of consultation, rather than a short course of tablets (for example an intramuscular antibiotic in the treatment of gonorrhoea). In some cases, a relative or other carer can give the drug at home.

Removing barriers to compliance

Compliance may be encouraged by prescribing pleasant-tasting elixirs rather than tablets for children and the elderly, and by using a drug or formulation that minimizes adverse effects.

Simplification of the therapeutic regimen

The therapeutic regimen can be simplified by reducing both the number of drugs a patient has to take and the frequency of administration. This can sometimes be

done with the help of modified-release or combination formulations.

Education and reminders

Educating the patient about why treatment is necessary (for example, treating hypertension or diabetes reduces the risk of serious complications) is time consuming but undoubtedly improves compliance. In the treatment of certain infections (as in patients with AIDS or tuberculosis and in typhoid carriers) the importance to the community should also be explained.

Even when patients are well motivated, reminders to take the treatment may improve compliance as it is easy to forget to take medication. Many medicines for long-term use are dispensed in a 'calendar pack', for example oral contraceptives.

Patients can be helped by a clearly written list of their current drugs, explaining the names, and labelling medicine bottles clearly. Information leaflets may also help to educate and remind them about their treatment.

The pharmacokinetic process

Pharmacokinetics concerns the question 'Is the drug getting to its site of action?' It comprises absorption and systemic availability, distribution, metabolism, and excretion.

Absorption and systemic availability

After an oral dose the amount of drug that reaches the systemic circulation depends on its absorption and how far it escapes metabolism in the gut, the liver, and the lungs. This is called its systemic availability (or bioavailability).

Systemic availability is defined in terms of the amount of administered drug that reaches the systemic circulation intact and the rate at which that happens. The rate of availability depends on pharmaceutical factors (see above) and gastrointestinal absorption, presystemic metabolism being relatively unimportant. But the extent of availability depends on both the extent of absorption and the extent of presystemic metabolism.

Figure 2 illustrates what is meant by the rate and extent of systemic availability. It shows three curves representing the theoretical plasma concentrations resulting over time after the same oral dose of three different formulations of the same drug. Each curve has three important attributes: the peak concentration (C_{max}), the time taken to reach the peak (t_{max}), and the total area under the curve (**AUC**). C_{max} and t_{max} are measures of the rate of availability, and the total AUC is a measure of its extent. In the three hypothetical cases the rates of availability clearly differ. For formulation I the systemic availability is rapid, perhaps too rapid, leading to potentially toxic plasma concentrations. Formulation II is less quickly available and plasma concentrations never reach the potentially toxic range. Formulation III is slowly available and plasma concentrations after a single dose are always subtherapeutic. However, in contrast to the rate of availability, the extent of availability of these three formulations, as assessed by the AUC, is the same in each case.

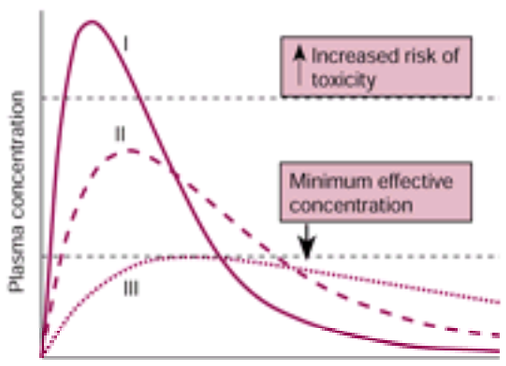

Fig. 2 The theoretical plasma concentrations resulting over time after the oral administration of three different formulations of the same dose of the same drug. The profile in each case depends on both the rate and the extent of systemic availability.

For drugs whose action may depend on the threshold plasma concentration achieved after a single dose (for example analgesics), such differences may be important. Thus, for the rapid relief of pain a soluble aspirin formulation, giving a curve like formulation II, would be preferable to an enteric-coated formulation, giving a curve like that of formulation III.

A curve of type I can be therapeutically useful if a very fast rate of absorption is needed to produce a quick therapeutic effect, for example sublingual glyceryl trinitrate to relieve an acute attack of angina pectoris, although it is also more likely to cause headache by dilating extracranial blood vessels.

For drugs whose action is related to a steady-state concentration during multiple dosing, the differences in rate of availability become less important and the chief consideration is the extent.

The rate of absorption

Gastrointestinal motility

Drugs are absorbed mainly in the upper small intestine, and altered gastric emptying therefore alters the rate. For example, in migraine the rate of absorption of analgesics is reduced because of reduced gastric motility, delaying the response to an oral analgesic. This delay can be reduced by giving metoclopramide, which hastens gastric emptying.

When a drug dissolves more slowly than the stomach empties, increased gastrointestinal motility may reduce both the rate and extent of absorption. Thus, in severe diarrhoea enteric-coated formulations may pass through the gut intact.

Malabsorption

Drug absorption is often impaired in patients with malabsorption, but not always. For example, the absorption of propranolol, cotrimoxazole, and cefalexin is increased in patients with coeliac disease. Digoxin, however, is less well absorbed from tablets in patients with coeliac disease, radiation-induced enteritis, and other gastrointestinal disease, and thyroxine absorption is impaired in coeliac disease.

Food

Food may alter the rate and extent of absorption of drugs. For example, eggs impair iron absorption, and milk (and any calcium, aluminium, magnesium, or ferrous salt) impairs tetracycline absorption by the formation of an insoluble chelate. But such effects are rarely important clinically.

First-pass metabolism

First-pass metabolism is metabolism that occurs before the drug enters the systemic circulation. This may happen in the gut lumen (for example with benzylpenicillin

or insulin), the gut wall (tyramine, chlorpromazine), the liver (the most important), and the lungs (various amines).

Many drugs undergo first-pass metabolism in the liver. For instance, propranolol is metabolized to 4-hydroxypropranolol, which is pharmacologically inactive.

When first-pass metabolism results in the formation of compounds with less pharmacological activity than the parent compound, the drug's efficacy is lower after oral than intravenous administration. In some cases (for example insulin) metabolism is so extensive that oral therapy is impossible. However, such a drug given sublingually, rectally, or transdermally, can bypass the liver (see above).

Distribution

Protein binding

Many drugs are bound to circulating proteins, usually albumin (acidic drugs), but also globulins (hormones), lipoproteins, and acid glycoproteins (basic drugs). Only non-protein-bound drug can bind to cellular receptors, pass across tissue membranes, and reach cellular enzymes, thus being distributed to other body tissues, metabolized, and excreted (for example by the kidney). Thus, changes in protein binding may sometimes cause changes in drug distribution. However, such changes matter only if the drug is more than 90 per cent bound in the plasma and is not widely distributed to body tissues. This is important mainly with phenytoin and warfarin.

The albumin binding of drugs may be changed in renal impairment (the explanation for this is unknown), hypoalbuminaemia (drug binding is reduced when the plasma albumin concentration falls below 25 g/l), the last trimester of pregnancy (during which protein binding is reduced partly because of hypoalbuminaemia), and displacement by other drugs.

The binding of drugs to a_1-acid glycoprotein is increased after trauma and surgery, and in inflammatory diseases and infections. For example, the binding of quinine is increased in malaria.

Tissue distribution

The extent of drug distribution to the tissues of the body varies widely. Some drugs enter only the body fluids, while others are bound extensively in tissues. The distribution of drugs to different tissues is influenced by plasma-protein binding, specific receptor sites in tissues (for example the binding of cardiac glycosides to Na^+,K^+-ATPase in cell membranes throughout the body), regional blood flow (well-perfused organs, such as the heart, kidneys, and liver, accumulate drugs more than poorly perfused organs, such as fat and bone), lipid solubility (non-polar relatively lipid-soluble drugs enter tissues more readily than polar compounds), active transport across cell membranes (for example the adrenergic neurone-blocking drugs), and the effects of other drugs (tricyclic antidepressants inhibit the active transport of the adrenergic neurone blockers, reducing their access to the site of action in the brain, and thus their efficacy).

In some diseases drug distribution is altered for unknown reasons. In renal failure, apart from its effect on protein binding, the distribution of some drugs (for example insulin and digoxin) may also be decreased. In cardiac failure the distribution of some antiarrhythmic drugs, such as disopyramide, is reduced. Obesity and malnutrition influence the distribution of drugs that are highly fat soluble (anaesthetics for example).

Metabolism

Most drugs are metabolized in the liver. Examples of other sites are: suxamethonium in the plasma; insulin and vitamin D in the kidneys; cytosine arabinoside, cyclophosphamide, and other cytotoxic drugs in many cells; and acetylcholine and other neurotransmitters at synapses and within nerves.

Drug metabolism occurs in two phases:

1. Phase I metabolism involves chemical alteration of the basic structure of the drug, for example by oxidation, reduction, or hydrolysis. Oxidation reactions are further subdivided according to whether or not they are effected by the cytochrome-linked mixed function oxidases. Examples of phase I reactions include the *N*-demethylation of diazepam to desmethyldiazepam, an active metabolite with a long duration of action, and the oxidation of ethanol to acetaldehyde.
2. Phase II metabolism involves conjugation, for example by sulphation, glucuronidation, methylation, or acetylation. Some drugs are conjugated without prior phase I transformation, while others undergo phase I metabolism before conjugation can occur. The end products of conjugation are compounds that are more water soluble and therefore more rapidly excreted by the kidneys. They are usually, though not always, pharmacologically inactive. Examples of phase II reactions are the glucuronidation of paracetamol, the *N*-acetylation of hydralazine and procainamide, and the methylation of desipramine to its active metabolite imipramine. A conjugated product may sometimes be further metabolized. For example, oestrogens are excreted via the bile, deconjugated in the gut by bacteria, and then reabsorbed.

The end-result of drug metabolism is inactivation, but during the process compounds with pharmacological activity may be formed. An inactive compound may be metabolized to an active one; inactive drugs administered for the effects of their active metabolites are called 'prodrugs'. Levodopa, for Parkinson's disease, can be regarded as a prodrug because it enters the brain and is there metabolized to dopamine. Usually a pharmacologically active compound is transformed to inactive compounds, but sometimes other active compounds are formed first. Examples of parent drugs with active metabolites include diamorphine (which is rapidly metabolized to morphine) and some benzodiazepines (such as diazepam, metabolized to desmethyldiazepam).

Some active compounds are metabolized to toxic compounds. Examples include pethidine (meperidine), the accumulation of whose toxic metabolite norpethidine limits the duration of therapy, and phenytoin, whose main metabolite may inhibit the further metabolism of phenytoin. The normally minor metabolic pathway by which paracetamol is metabolized to a toxic metabolite is enhanced in overdose because the usual detoxifying pathways are saturated (see Chapter 8.2.1).

The factors that affect hepatic drug metabolism are genetic (see pharmacogenetics below), other drugs (see drug interactions below), hepatic blood flow (for drugs that are rapidly cleared), liver disease (only important in extensive liver disease or when there is arteriovenous shunting), and age. The metabolism of some drugs is impaired in old people and in babies younger than about 6 months, particularly premature babies. In both cases this is due to reduced activity of the hepatic microsomal drug metabolizing enzymes. For example, in neonates uridine 5-diphosphate glucuronyl transferase, which conjugates chloramphenicol, is relatively inactive; neonates eliminate chloramphenicol slowly, and may suffer peripheral circulatory collapse (the 'grey syndrome') when given it in weight-related doses that do not harm adults.

Excretion

The kidney is the main route of drug excretion. Other, usually minor, routes include the lungs (important for paraldehyde), breast milk, sweat, tears, and genital secretions (alarming if the patient is not expecting the orange-red discoloration caused by rifampicin), saliva, and bile. Excretion in bile can lead to the reabsorption of some compounds in the gut, for example chloramphenicol (whose inactive metabolites are reactivated by hydrolysis in the gut), oestrogens (which bacteria in the gut deconjugate), morphine, rifampicin, and tetracyclines.

Renal excretion of drugs involves three main processes (Fig. 3): glomerular filtration, passive tubular reabsorption, and active tubular secretion. Thus:

Total renal clearance = Clearance by filtration + Clearance by secretion – Retention by reabsorption.

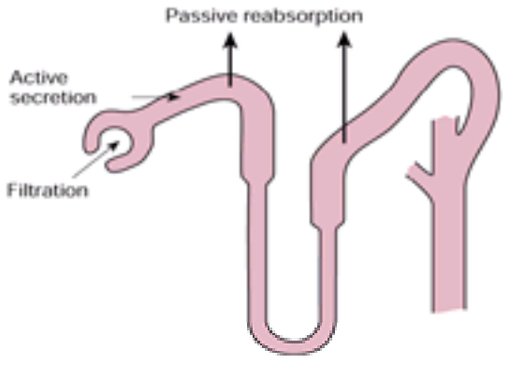

Fig. 3 A diagrammatic representation of a nephron, showing the sites of the three major processes of drug excretion via the kidney.

If a drug is mainly metabolized to inactive compounds, renal function will not greatly affect elimination of the active compound. However, if the drug or an active metabolite is excreted unchanged via the kidneys, changes in renal function will influence its elimination.

Glomerular filtration

All drugs are filtered at the renal glomerulus. The extent of filtration is directly proportional to the glomerular filtration rate (**GFR** = 120 ml/min) and to the fraction of unbound drug in the plasma (*fu*). Thus:

$$\text{Rate of clearance by filtration} = fu \times \text{GFR}.$$

If the total renal clearance of a drug is equal to *fu* × GFR then it is cleared principally by filtration. It may, of course, also be affected by the other two mechanisms, secretion and reabsorption, but in that case those effects must balance each other. Examples of drugs whose clearance is similar to the GFR (after correction for protein binding) are digoxin, gentamicin, procainamide, methotrexate, and ethambutol. As creatinine is cleared mainly by filtration, the creatinine clearance is useful in estimating the clearance rates of these drugs.

Passive tubular reabsorption

Drugs are passively reabsorbed by the renal tubules. The elimination of drugs with very low rates of renal clearance (i.e. approaching urine flow rate, or about 1–2 ml/min) will be significantly affected by changes in urine flow rate (because a doubling of flow rate will increase their rate of clearance by 1–2 ml/min, i.e. twofold). However, for weak acids and weak bases the main factor affecting passive reabsorption is the pH of the renal tubular fluid, because the extent of their ionization (and therefore of their passive reabsorption) depends on the pH when seen in relation to the p K_a of the drug. For example, in an alkaline urine weak acids with a p K_a below 7.5, such as aspirin, are more highly ionized, and therefore less well reabsorbed. The reverse is true for weak bases with a p K_a greater than 7.5, such as amphetamine, whose reabsorption is reduced, and whose clearance is therefore enhanced, by an acid urine. These principles are sometimes put to use in the treatment of overdose (see Section 8). Renal failure alters passive reabsorption indirectly, by changing urine flow rate and pH.

Active tubular secretion

If the renal clearance of a drug is greater than expected by filtration, it is also secreted actively by the proximal tubule. Penicillin is an example. Some drugs inhibit active tubular secretion, and this explains some drug interactions (see below).

Plasma half-life ($t_{1/2}$)

The half-life of a drug is the time it takes for the plasma concentration or amount of drug in the body to halve. In most cases it is constant, no matter what the starting concentration or amount. After one half-life, 50 per cent of the drug will be eliminated, after two half-lives 75 per cent (50 + 25 per cent), and so on. Thus, it takes between four and five half-lives for about 95 per cent of the drug to be eliminated. If the half-life of the drug is prolonged (as for digoxin in renal failure), elimination will take proportionately longer. The rate at which drug accumulates in the body during regular multiple dosing is determined by the half-life. Just as it takes four half-lives for 95 per cent of the drug to be eliminated after withdrawal, it takes four half-lives for about 95 per cent of steady state to be reached during regular administration (Fig. 4(a) and Fig. 5, bottom and middle curves).

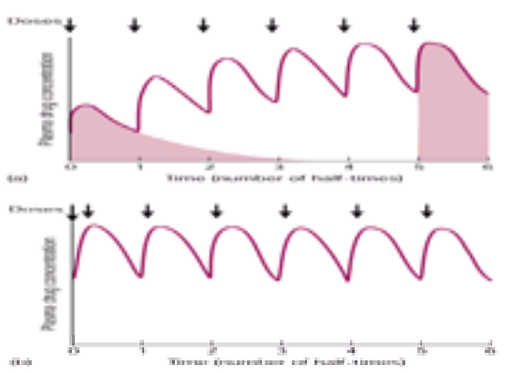

Fig. 4 (a) The theoretical plasma concentrations of a drug over time during its repeated oral administration. With repeated doses the drug accumulates to an eventual steady state. (b) If the correct loading dose is given a steady state can be achieved rapidly and then maintained by giving a smaller maintenance dose. In this example, because the drug is being given once every half-life the maintenance dose is half the loading dose.

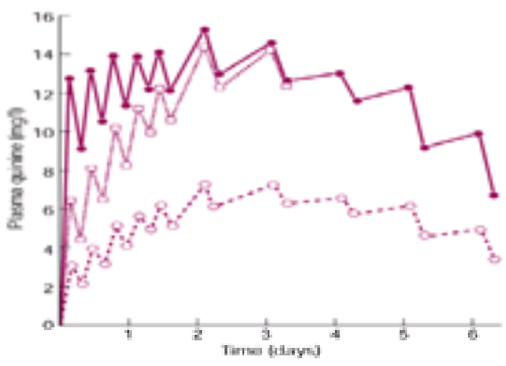

Fig. 5 Plasma quinine concentrations in cerebral malaria during repeated administration of 5 mg/kg every 8 h (bottom curve), repeated administration of 10 mg/kg every 8 h (middle curve), and repeated administration of 10 mg/kg every 8 h after a loading dose of 20 mg/kg (top curve). (From White NJ *et al*. (1983). Quinine loading dose in cerebral malaria. *American Journal of Tropical Medicine and Hygiene* **32**, 1–5, with permission.)

The delay in reaching a steady state can be overcome by giving a loading dose. When a drug has a half-life greater than 24 h (for example digoxin, 40 h; digitoxin 7 days; *S*-warfarin, 32 h), it takes several days or weeks of regular administration of the same daily dose before the steady-state plasma concentration or amount of drug in the body is reached. Such a delay may be unacceptable if the eventual steady-state plasma concentration is needed for a therapeutic effect. In such cases a loading dose can be given in order to boost the amount of drug in the body to the required level. Thereafter the regular maintenance dose is given to maintain the steady state (Fig. 4(b) and Fig. 5, top curve).

Non-linear kinetics

Although the pharmacokinetics of most drugs are linear, some are not. This means that some aspect of their kinetics becomes saturated in the therapeutic dosage range, so that a linear increase in dosage does not produce a proportionate increase in plasma concentration or effect. The prime example is phenytoin, whose enzymatic metabolism in the liver becomes saturated at dosages in the therapeutic range; plasma phenytoin concentrations therefore increase non-linearly with dosage.

The pharmacodynamic process

The pharmacodynamic process describes all those matters concerned with the pharmacological actions of a drug, whether they lead to therapeutic effects or adverse effects. The associated question is 'Is the drug producing the required pharmacological effect?'

Drugs produce their pharmacological effects in many different ways, often by stimulating or blocking a receptor (Table 3).

Actions via direct effects on receptors

Receptors are proteins situated either in cell membranes or within the cellular cytoplasm. For each type a specific group of ligands, drugs or endogenous substances exists that bind to the receptor and produce pharmacological effects. Ligands are of three types: agonists, antagonists, and partial agonists.

Agonists bind to a receptor and produce a response. For example, morphine is an agonist at μ opioid receptors, which mediate its analgesic action.

Antagonists (or blockers) prevent an agonist from binding to a receptor and thus prevent its effects. However, pure antagonists do not themselves have any pharmacological actions mediated by receptors. For example, naloxone is an opioid-receptor antagonist; when it binds to opioid receptors it prevents or reverses the effects of morphine and other opiates.

Unlike a full agonist, which is capable of producing a maximal response when it binds to enough receptors, a partial agonist cannot produce the maximal response of which the tissue is capable, even when it binds to the maximum number of available receptors. Thus, above a certain level of binding a partial agonist may bind to receptors without producing any further increase in effect. However, in so doing it may prevent the action of other agonists, and may thus behave as an antagonist. This mixture of actions is called partial agonism. For example, buprenorphine is a mixed opioid agonist/antagonist at μ receptors. If used in combination with a high dose of a pure agonist it will tend to oppose its action rather than supplement it.

Receptor subtypes

In some cases a receptor may have subtypes, for which certain ligands may have some selectivity. For example, of at least three subtypes of opioid receptors, μ and † are both involved in analgesia, gastrointestinal motility, and respiratory depression, and _ is involved in analgesia, sedation, and miosis. These receptors are variably distributed in the nervous system. Most opiates act at μ receptors, but none is completely selective and they may act at other subtypes.

Long-term effects of drugs at receptors

During long-term therapy the effects of a drug may be altered by adaptive responses, usually accompanied by either increases ('upregulation') or decreases ('downregulation') in receptor numbers. Such changes may explain both beneficial and adverse effects of drugs. Examples include:

- The therapeutic response to antidepressants, perhaps related to changes in receptors in the brain secondary to the actions of these drugs on neurotransmitter uptake. This could explain why the therapeutic response to antidepressants takes a few weeks.
- The way in which the response to levodopa in Parkinson's disease changes during long-term administration (producing, for example, the 'on–off' effect).
- Withdrawal syndromes that may occur because long-term changes become unopposed when the drug is withdrawn, for example after the long-term abuse of opiates, or the use of benzodiazepines in chronic anxiety.

Actions via direct effects on second messengers

When an agonist stimulates a membrane-bound receptor its effect is produced in one of two ways: either through a so-called second messenger (see Fig. 6) or by changing the activity of an ion channel linked to the receptor. Some drugs may act by affecting second messengers directly. For example, lithium may work by inhibiting phosphatidylinositol turnover at different sites. Theophylline and caffeine work by inhibiting the breakdown of cyclic AMP.

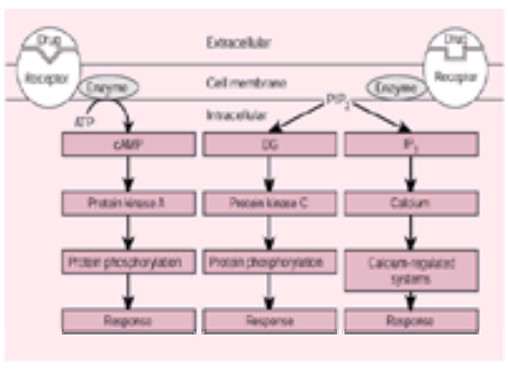

Fig. 6 A schematic representation of the second-messager systems that mediate the effects of drugs acting at receptors. PIP_2, phosphatidylinositol bisphosphate; DG, diacyl glycerol; IP_3, inositol triphosphate.

Actions via indirect alterations of the effects of endogenous agonists

Some drugs indirectly alter the effect of an endogenous agonist rather than acting via a receptor. Some oppose the physiological effects of others. For example, glucagon is a physiological antagonist of the actions of insulin and can be used to treat hypoglycaemia.

Increase in endogenous release

Some drugs enhance the action of endogenous agonists by increasing their release. For example, amphetamines increase the release of dopamine from nerve endings. Because amphetamines can cause a syndrome resembling schizophrenia, this action has led to the idea that schizophrenia may be related to excess

dopamine action in the brain.

Prevention of endogenous release

If a drug prevents the release of an endogenous agonist it will reduce its effects. For example, angiotensin-converting enzyme inhibitors prevent the formation of angiotensin II; this prevents the release of endogenous aldosterone, resulting in potassium retention.

Inhibition of endogenous reuptake

Some drugs inhibit the reuptake of endogenous agonists, enhancing their effects. For example, many antidepressants inhibit the reuptake by neurones of certain neurotransmitters, such as noradrenaline and 5-hydroxytryptamine.

Inhibition of endogenous metabolism

Some drugs potentiate endogenous agonists by inhibiting their metabolism. For example, vigabatrin inhibits the metabolism of g-aminobutyrate in the brain, enhancing its actions and suppressing seizures.

Actions via the inhibition of transport processes

Because cations (such as sodium, potassium, and calcium) and other substances (such as organic acids in the kidneys) have so many important roles in the maintenance of normal cellular function, inhibition of their transport is an important type of mechanism of drug action.

Diuretics

Most diuretics act by inhibiting sodium reabsorption in the renal tubules. The loop diuretics frusemide (furosemide)and bumetanide inhibit Na/K/Cl cotransport in the ascending limb of Henle's loop. The thiazide diuretics inhibit Na/Cl cotransport in the proximal segment of the distal convoluted tubule. The potassium-sparing diuretic amiloride inhibits sodium channels in the distal segment of the distal convoluted tubule.

Calcium antagonists

The calcium antagonists, such as verapamil, diltiazem, and the dihydropyridines (for example nifedipine), inhibit the transport of calcium via potential-operated calcium channels. The different calcium antagonists have different selectivities for calcium channels in different tissues, and have various actions. For example, verapamil has an antiarrhythmic action in the heart, and nifedipine a vasodilator action on peripheral arterioles.

Drugs acting on potassium channels

Potassium channels in cell membranes control the rate of efflux of potassium from the cells. Cellular activity is reduced by drugs that open K^+ channels and increased by drugs that close them.

Drugs that open potassium channels include vascular smooth muscle relaxants, such as minoxidil and hydralazine. Drugs that close them include the sulphonylureas, which increase the release of insulin from pancreatic b cells, and 3,4-diaminopyridine, which increases the release of acetylcholine at the neuromuscular junction.

Actions via enzyme inhibition

Some drugs act by directly inhibiting enzymes.

Neostigmine is a reversible cholinesterase inhibitor. It increases the concentration of acetylcholine at the muscle motor endplate, improving neuromuscular transmission. Xanthine and hypoxanthine are oxidized to uric acid by xanthine oxidase; this is inhibited by allopurinol, which therefore reduces the synthesis of uric acid in gout.

The monoamine oxidase inhibitors inhibit the metabolism of the monoamines 5-hydroxytryptamine, noradrenaline, and dopamine in the brain, and they presumably produce their antidepressant effect by this action. Just as drugs that act via receptors may be selective for a subtype of receptor, so monoamine oxidase inhibitors may be selective for one subtype of monoamine oxidase. For example, selegiline selectively inhibits monoamine oxidase type B: it inhibits the metabolism of dopamine in the brain and enhances the action of levodopa in parkinsonism. However, because gut monoamine oxidase is of type A, selegiline does not produce the 'cheese reaction' (due to tyramine and other amines, see below) that non-selective monoamine oxidase inhibitors do. In contrast, the antidepressant moclobemide selectively inhibits monoamine oxidase A; although it inhibits the metabolism of tyramine in the gut it does not inhibit its metabolism by monoamine oxidase B after absorption.

The cardiac glycosides act by inhibiting the Na/K pump, changing the disposition of sodium, which secondarily changes calcium disposition within cells.

Other drugs that act via enzyme inhibition include warfarin (vitamin K epoxide reductase), aspirin and other non-steroidal anti-inflammatory drugs (the enzymes involved in prostaglandin synthesis), angiotensin-converting enzyme inhibitors, disulfiram (alcohol dehydrogenase), some anticancer drugs such as cytarabine (DNA polymerase), and some anti-infective drugs (bacterial or viral enzymes; for example, trimethoprim inhibits bacterial dihydrofolate reductase, the quinolones inhibit bacterial DNA gyrase, and zidovudine and didanosine inhibit the reverse transcriptase of HIV).

Danazol and stanozolol are examples of drugs that inhibit an enzyme indirectly—they stimulate the production of an inhibitor of C1 esterase and are used to treat hereditary angio-oedema, in which there is reduced plasma activity of the inhibitor.

Actions via enzyme activation or direct enzymatic activity

Some drugs activate enzymes or are themselves enzymes.

The clotting and fibrinolytic factors are enzymes, and certain drugs that act on clotting and fibrinolysis do so by increasing their activity. Heparin acts by activating antithrombin III. The thrombolytic drugs streptokinase, alteplase, and anistreplase activate plasminogen.

Clotting factor deficiencies can be treated by replacing deficient enzymes of the clotting pathway, for example factor VIII in patients with haemophilia, and fresh frozen plasma in warfarin toxicity. Pancreatic enzymes are used in treating malabsorption in patients with chronic pancreatic insufficiency.

L-Asparaginase is an enzyme that hydrolyses asparagine, an essential amino acid in neoplastic cells. Some patients with acute lymphoblastic leukaemia may be helped by depriving their leukaemic cells of asparagine.

Actions via other miscellaneous effects

Chelating agents

Drugs that chelate metals can be used to hasten their removal from the body. Calcium sodium edetate chelates many divalent and trivalent metals and is used to treat poisoning, particularly with lead. Dimercaprol chelates some heavy metals and is used to treat mercury and gold poisoning. Desferrioxamine chelates iron and is used in treating iron poisoning and the iron overload that occurs with repeated blood transfusion (as in thalassaemia). Penicillamine chelates copper and is used in treating

hepatolenticular degeneration (Wilson's disease); it is also used to complex cystine and thus prevent renal damage in cystinuria.

Osmotic diuretics

Mannitol is freely filtered at the glomerulus but the renal tubules reabsorb only a little. It therefore increases the concentration of osmotically active particles in the tubular fluid and takes water with it.

Volatile general anaesthetics

General anaesthetics form a diverse group of agents, such as the halogenated hydrocarbons (halothane, trichloroethylene), and non-halogenated agents (nitrous oxide, ether, cyclopropane), which produce similar effects on the brain. Their main action is probably on the lipid matrix of the biological membrane, changing its biophysical properties, and so altering ion fluxes or other functions that influence neuronal excitability.

Replacement of vitamins and minerals

Some drugs are used simply to replace deficiencies, for example ferrous salts in iron deficiency anaemia and hydroxocobalamin (vitamin B_{12}) in vitamin B_{12} deficiency.

Stereoisomerism and drug action

Stereoisomerism (chirality) of organic compounds is due to asymmetry in one or more of their atoms (usually carbon), resulting in two structures (enantiomers) that cannot be superimposed on each other.

The terminology used to describe chiral compounds is complex, but in summary some are called *R* and *S* (from the Latin *rectus* = right and *sinister* = left), others D and L (from the Latin *dexter* = right and *laevus* = left), and yet others *d* and *l*. Examples of drug enantiomers are *R*-warfarin and *S*-warfarin, D-glucose (dextrose) and L-glucose (laevulose), and *d*-propran-olol and *l*-propranolol.

Of all synthetic drugs used in clinical practice about 40 per cent are chiral and about 90 per cent of those are marketed in the racemic form (i.e. as an equal mixture of the two enantiomers). Examples include *d*,*l*-propranolol and *R*,*S*-warfarin. Naproxen is one of the few examples of a synthetic compound that is marketed as a single enantiomer. In contrast, naturally occurring and semisynthetic compounds are almost all chiral and almost all are marketed as a single isomer. Examples include D-glucose (dextrose) and the naturally occurring amino acids (for example L-dopa).

Enantiomers often have differing actions. For example, *l*-propranolol is a b-blocker, while *d*-propranolol has membrane stabilizing activity like that of local anaesthetics; *l*-sotalol is a b-blocker, while *d*-sotalol has antiarrhythmic effects like those of amiodarone.

Sometimes the difference between enantiomers is a difference between therapeutic and adverse effects, as dramatically demonstrated by the example of thalidomide whose *R* enantiomer is hypnotic but whose adverse effects seem to be due to the *S* enantiomer.

In some cases, differences between enantiomers are limited to differences in potency. For example, *S*-warfarin and *R*-warfarin have the same anticoagulant actions, but the former is about five times more potent.

Sometimes enantiomeric differences tell us something about the mechanism of action of a drug. For example, *S*-timolol is a more potent b-blocker than *R*-timolol, but both are equally effective in reducing intraocular pressure in patients with glaucoma. This suggests that timolol lowers the intraocular pressure by a mechanism unrelated to b-blockade.

One enantiomer may be eliminated differently from the body than the other. For example, the half-lives of *S*-warfarin and *R*-warfarin are 32 h and 54 h, and the routes of metabolism are to 7-hydroxywarfarin for *S*-warfarin and to warfarin alcohols for *R*-warfarin. This is important in some drug interactions with warfarin, because some drugs (such as metronidazole) that inhibit the metabolism of warfarin, primarily affect the more potent enantiomer, *S*-warfarin.

Pure enantiomers of synthetic drugs are expensive to make, but new techniques may reduce the costs and lead to the emergence of more formulations of pure enantiomers for clinical use.

The therapeutic process

The question associated with the therapeutic process is 'Is the pharmacological effect being translated into a therapeutic effect?'

Translation of pharmacological effect into therapeutic effect during short-term therapy

The short-term therapeutic and toxic effects of drugs occur as a result of the pharmacological actions discussed above. However, the translation of molecular and cellular pharmacological effects into the therapeutic or toxic effect is not a simple process, but one that involves several translational stages at different pharmacological and physiological levels.

Take, for example, the action of salbutamol, a b_2-adrenoceptor agonist, in the treatment of asthma. Salbutamol stimulates bronchial b_2-adren-oceptors, and so increases the activity of adenylate cyclase; this is its pharmacological effect at the molecular level. The increase in adenylate cyclase activity leads to an increase in the intracellular concentration of cyclic AMP, a pharmacological effect at the cellular level. The increase in cyclic AMP in some way alters the function of bronchial smooth muscle cells, and results in an inhibition of the release of inflammatory mediators from bronchial mast cells, with effects on cell physiology. All this in turn results in bronchodilatation, an effect on tissue physiology. Bronchodilatation improves lung function, an effect on organ physiology. Finally, the patient can breathe more easily, the desired clinical effect.

This analysis of the short-term effects of drugs teaches us several things about drug action: how drug action may be modified; how therapeutic and adverse effects may be mediated via different pharmacological effects; the relation between the pharmacological effects of a drug and the rate of onset or duration of its action; and drug/disease interactions.

How drug action may be modified

It is often possible to modify the action of a drug beneficially or adversely. For example, one would expect a xanthine derivative, such as theophylline, which increases cellular cyclic AMP concentrations by the inhibition of phosphodiesterase, to potentiate the action of salbutamol at the stage at which it alters adenyl cyclase activity. This turns out to be both a beneficial and an adverse clinical interaction—beneficial because theophylline enhances the therapeutic action of salbutamol, adverse because it also enhances the hypokalaemia that salbutamol may cause by stimulating Na^+/K^+-ATPase.

Different pharmacological actions may mediate therapeutic and adverse or other effects

Some drugs have more than one molecular mechanism of action, and two different therapeutic effects of a drug may result from different actions. For example, tetracycline acts against bacteria by interfering with their protein synthesis, but in acne it helps by interfering with sebum production in sebaceous glands.

Similarly, a therapeutic effect may be brought about by one pharmacological action and an adverse effect by another. For example, the therapeutic effect of salbutamol results from its action on b_2-adrenoceptors, but it causes the unwanted effect of tachycardia by stimulating b_1-adrenoceptors.

Different therapeutic or adverse effects may be produced by the actions of a single drug on the same or a similar molecular mechanism in different tissues. For

example, the inhibition of b_2-adrenoceptors causes bronchoconstriction in the lungs of susceptible people and impairs glycogenolysis in the liver.

Peripheral or non-therapeutic effects of drugs can be clinically important in several ways. For example, aminoglycoside antibiotics can all damage the inner ear, leading to impaired hearing or loss of balance; these effects are quite separate from the antibiotic activity of these drugs but must be considered when choosing dosage regimens and mean that plasma concentration must be monitored.

The relation between the pharmacological actions of a drug and the rate of onset and duration of its effects

The rate of onset of action of a drug depends not only on its pharmacokinetics (i.e. the time it takes for the appropriate amount of drug to build up at the site of action; Fig. 4), but also on how long it takes for the full pharmacodynamic sequence of events to unroll. In the case of salbutamol the time between b_2-adrenoceptor stimulation and bronchodilatation (Fig. 7) is of the order of a few minutes. However, for other drugs the sequence of events takes much longer. For example, corticosteroids react with a receptor protein in cellular cytoplasm to form a steroid–receptor complex. This complex enters the cell nucleus, where it binds to chromatin and directs the genetic apparatus to transcribe RNA. This leads, for instance in liver cells, to the production of enzymes involved in gluconeogenesis and amino acid metabolism. Once the steroid has bound to its intracellular receptor it sets off a sequence of reactions that then has its own timescale, independent of the quantity of steroid, either in the blood or combined with the receptor. The induction of protein synthesis by RNA transcription takes several hours and each new protein has its own biological lifespan. This is why the effect takes hours or days to occur.

Similarly, the duration of action of a drug is related not only to the time it takes for the drug to be cleared from the body, but also to the duration of its pharmacological effects. For example, aspirin inhibits cyclo-oxygenase by acetylating a serine moiety at the active site of the enzyme. In platelets prostaglandin synthesis is thus inhibited for the lifetime of the platelet. So, although aspirin leaves the body within a few hours, its effect on platelets lasts for days.

Drug/disease interactions

Because of the complex links between the pharmacological effects of a drug and its therapeutic or adverse effects, the pathophysiology of the disease being treated, or of other incidental diseases, can variously alter the way in which the pharmacological effect is translated into a therapeutic effect.

The use of digoxin in cardiac failure is an example of the ways in which drug/disease interactions may influence therapy. Digoxin inhibits the activity of the membrane-bound Na/K pump. This is its pharmacological effect at the molecular level, which increases the intracellular concentration of sodium, and thereby alters the intracellular disposition of calcium (a pharmacological effect at the cellular level). This in turn leads to an alteration in the action potential of cardiac muscle (a physiological effect at the cellular level), which then causes an increase in the rate of contractility of the myocardial fibres (a physiological effect at the tissue level). Consequently cardiac output increases (a physiological effect at the level of the whole organ).

If this chain of events occurs without interruption the clinical result will be relief of the signs and symptoms of heart failure. However, drug/disease interactions can alter the therapeutic outcome at various stages in the chain. Thus, potassium depletion enhances the binding of digoxin to the Na/K pump and this increases the extent of inhibition of sodium transport, which can in turn cause digoxin toxicity. In hyperthyroidism the nature of the interaction between digoxin and the Na/K pump is altered, resulting in resistance to the inhibitory effects of digoxin. Increasing the dose may merely cause digoxin toxicity without ever producing a therapeutic effect. In patients with chronic cor pulmonale, digoxin may inhibit the Na/K pump and may even cause cardiac arrhythmias without ever increasing the rate of myocardial contractility. This may be because of tissue hypoxia and acidosis in cor pulmonale, which may also contribute to an increased risk of digoxin-induced cardiac arrhythmias without benefiting patients with acute myocardial infarction. In patients with hypertrophic obstructive cardiomyopathy, although digoxin increases the rate of myocardial contractility, this is not translated into an increase in cardiac output, as left ventricular outflow remains obstructed.

Thus even when it can be shown that a drug is having its expected action at a particular pharmacological or physiological level, it cannot automatically be assumed that it will have a consequent therapeutic effect.

Interactions with circadian phase

Most bodily functions vary in a circadian rhythm - for example sleep and wakefulness, urinary excretion, secretion of hormones—and drug effects are liable to differ at different phases of the rhythm. In some instances the difference is dramatic. For example melatonin taken late in the afternoon evening advances the circadian phase, and sleepiness comes earlier, whereas taken in the morning the same dose retards circadian phase. Bright light switches off the body's own melatonin release, darkness switches it on. So in the morning exogenous melatonin prolongs the effect of darkness, late in the day it switches on the response to darkness early. Another example is related to the secretion of cortisol by the adrenal cortex. A high plasma cortisol concentration inhibits corticotrophin release from the pituitary. Corticotrophin is secreted at night, and plasma cortisol peaks around 8 a.m., and about then the corticotrophin concentration is zero. Exogenous cortisol given at this time cannot inhibit corticotrophin release because none is occurring. But given in the evening it will completely inhibit corticotrophin release during the night. For this reason cortisol given once daily in the morning causes much less pituitary inhibition than the same dose given in the evening, or than doses spread over the 24 hours. Circadian therapeutics are complicated to study, and so far few practical examples are known. An important principle that is likely to find increasing application is that the 'hour of greatest therapeutic effect' will differ from the 'hour of greatest toxicity', since therapeutic and toxic effects are mediated by widely differing mechanisms. This could be exploited to increase the safety margin or tolerability of treatment, for example in cancer chemotherapy.

Translation of pharmacological effect into therapeutic effect during long-term therapy

During prolonged therapy adaptation may develop to the short-term pharmacological effects of the drug (Table 4), with several consequences.

Therapeutic effects through adaptation

In immunization, by adapting to an initial immunological challenge the immune system develops the ability to respond to a subsequent similar challenge, tetanus immunization for example.

Although tricyclic antidepressants rapidly inhibit reuptake of noradrenaline and 5-hydroxytryptamine in the brain, the therapeutic effect of these drugs takes 1 to 2 weeks to become evident. The brain adapts to the increased concentrations of noradrenaline and 5-hydroxytryptamine in the synaptic cleft in certain areas, where the sensitivity of responses to neurotransmitters is decreased by 'downregulation'; part of this adaptive effect could be the pharmacological action through which these drugs produce their therapeutic effects.

Tolerance: increasing ineffectiveness of therapy

Tolerance is a state of decreased responsiveness to a drug, resulting from previous exposure, either to the same drug or to one with similar short-term effects.

For example, it can develop to the vasoconstricting effects of ephedrine nosedrops, used to treat vasomotor rhinitis: as ephedrine acts by releasing noradrenaline from sympathetic nerve endings, when the noradrenaline is depleted the ephedrine can no longer be effective.

Patients who take long-term glyceryl trinitrate, particularly from transdermal patches, become tolerant and may not respond to its acute effects. To avoid this, a patch should be applied for no longer than 18 h. This effect probably reflects depletion of tissue sulphydryl groups by oxidation to disulphide groups.

Physiological tolerance by homoeostatic mechanisms

Secondary hyperaldosteronism is a physiological response to sodium loss produced by loop or thiazide diuretics. The enhanced potassium excretion that it causes may be obviated by using a potassium-sparing diuretic (for example amiloride) or the aldosterone antagonist spironolactone.

Another type of physiological tolerance occurs in patients given the diuretic acetazolamide. This is a powerful kaliuretic and can cause severe potassium depletion when it is first given. However, because it inhibits carbonic anhydrase activity in the kidney it causes bicarbonate depletion, and the resulting acidosis causes

retention of potassium. Thus, potassium depletion due to acetazolamide lasts for only a few days or weeks.

Metabolic tolerance

Metabolic tolerance results from faster metabolism of the drug. The commonest cause is induction of hepatic microsomal drug metabolizing enzymes by drugs such as barbiturates, phenytoin, or carbamazepine. In a particular case, the hepatic inactivation of the carbamazepine is increased after long-term exposure to carbamazepine itself (the phenomenon known as 'autoinduction'), and tolerance occurs spontaneously. Induction of drug metabolism can lead to adverse drug reactions (see below).

Withdrawal syndromes

A common, though not inevitable, outcome of an adaptive response to long-term drug use is the occurrence of a withdrawal response, which occurs either when the drug is withdrawn or when an antagonist is given, and which usually takes the form of some sort of adverse reaction.

A withdrawal syndrome occurs in opiate addicts when the opiate is withdrawn or when an antagonist, such as naloxone, is given. The symptoms consist of yawning, rhinorrhoea, and sweating, followed by shivering and goose flesh ('cold turkey'); later, nausea, vomiting, diarrhoea, and hypertension may occur. The acute syndrome subsides within a week, but the addict may be anxious and sleep badly for several weeks or months after. This syndrome can be avoided by introducing increasing doses of methadone as the opiate is withdrawn, as the later withdrawal of methadone, which has a much longer duration of action than morphine, may not result in this syndrome.

Delirium tremens may occur on withdrawal of alcohol from chronic alcoholics. This syndrome consists of disorientation and visual hallucinations. Withdrawal of benzodiazepines after long-term therapy may result in a disturbance of sleep pattern (rebound insomnia associated with abnormal sleep patterns), agitation, restlessness, and occasionally epileptic convulsions.

The risk of angina pectoris, myocardial infarction, and arrhythmias is increased in patients with ischaemic heart disease when b-adrenoceptor antagonists are withdrawn after long-term use. This may be due to an increase in the numbers of cardiac b-adrenoceptors, with increased sensitivity to the b-adrenergic effects of sympathetic stimulation.

Angina and even myocardial infarction have been reported in munitions workers who have become tolerant of the effects of nitroglycerine (glyceryl trinitrate) and have taken a break from work.

Long-term therapy with corticosteroids suppresses pituitary secretion of adrenocorticotrophic hormone, leading to adrenal cortical atrophy. When treatment is suddenly withdrawn, secretion of adrenocorticotrophic hormone by the pituitary may take several weeks or months to recover. Because the adrenal cortex has to increase in size again in order to become normally responsive to adrenocorticotrophic hormone the patient is at great risk of an addisonian crisis if stressed.

Adverse effects directly due to adaptation

Patients taking a neuroleptic drug (for example chlorpromazine, fluphenazine, or haloperidol) continuously for a long time commonly develop abnormal movements (known collectively as tardive dyskinesia). The face, mouth, and tongue are often affected, with stereotyped sucking and smacking of the lips, lateral jaw movements, and darting movements of the tongue. Occasionally more widespread dyskinesia may resemble choreoathetosis. The long-term blockade of brain dopamine function with neuroleptics is thought to lead to increased sensitivity to the effects of dopamine in certain areas of the brain, perhaps by increasing the number of dopamine receptors. Tardive dyskinesia may result from such increased sensitivity in extrapyramidal areas of the brain.

Adverse drug reactions

Adverse effects are unwanted effects of drugs, and they may be either toxic effects or side-effects. A toxic effect is an adverse effect that arises through an exaggeration of the pharmacological action that is responsible for the therapeutic effect of the drug (for example arrhythmias due to digoxin), and is therefore dose related. A side-effect is an adverse effect that arises through some pharmacological action other than that which produces the therapeutic effect (for example the anticholinergic effects of tricyclic antidepressants); such effects may or may not be dose related. The term 'adverse effects' covers all types of unwanted effects.

Incidence

The risks of adverse drug reactions have been variously estimated as follows:

- 10 to 20 per cent of hospital inpatients suffer an adverse reaction;
- 0.24 to 2.9 per cent of deaths in hospital inpatients are due to adverse reactions;
- 0.3 to 5.0 per cent of hospital admissions are due to adverse reactions.

Classification (see Table 5)

Dose-related adverse reactions

Dose-related adverse drug reactions are usually due to a pharmacokinetic or pharmacodynamic abnormality producing an exaggeration of a known pharmacological effect of the drug. The pharmacological effect that proves adverse may be the same as that which achieves the therapeutic effect (for example hypoglycaemia due to insulin), or due to another effect occurring in parallel (the anticholinergic action of tricyclic antidepressants, producing a dry mouth or urinary retention).

Dose-related adverse reactions may occur because of variations in the pharmaceutical, pharmacokinetic, or pharmacodynamic properties of a drug, often due to some disease or pharmacogenetic characteristic of the patient. These mechanisms are illustrated below.

Pharmaceutical defect

Adverse reactions can be caused by a contaminant, for example pyrogens or even bacteria in intravenous formulations, if quality control breaks down. If a febrile reaction occurs in a patient being given an infusion, the drip should be taken down and all its components sent for bacteriological investigation. The manufacturer should be urgently notified.

Out-of-date formulations may sometimes cause adverse reactions because of degradation products. For example, outdated tetracycline may cause Fanconi's syndrome, because it is degraded to anhydrotetracycline and epiandrotetracycline. The omission of the preservative citric acid from tetracycline formulations has reduced the risk of this effect, but has not removed it completely.

Pharmacokinetic variation

Normal individuals vary greatly in their rate of elimination of drugs. This variation is greatest for drugs cleared by hepatic metabolism and is determined by several factors, which may be genetic, environmental (diet, smoking, alcohol), or hepatic (blood flow and intrinsic drug-metabolizing capacity). On top of this normal variation specific pharmacogenetic or hepatic abnormalities may be associated with adverse reactions. In addition, renal and cardiac disease can change drug pharmacokinetics. Pharmacogenetics is discussed below.

The reserve of the liver parenchyma is large, and adverse reactions due to impaired hepatic metabolism are uncommon. Nevertheless, in patients with severe liver disease care must be taken, particularly with drugs with a low therapeutic index and those subject to extensive first-pass elimination. For example, hepatocellular dysfunction, as in severe hepatitis or advanced cirrhosis, may reduce the clearance of drugs for which the capacity of the liver is limited, phenytoin, theophylline, and warfarin for example. Portosystemic shunting in portal hypertension, associated with cirrhosis, reduces the clearance of drugs normally cleared rapidly by the liver, for

example morphine and other narcotic analgesics, propranolol, labetalol, and chlorpromazine.

Drugs that the kidneys excrete unchanged, or whose active metabolites are excreted, will accumulate in renal failure. Important examples include digoxin, lithium, tetracyclines, aminoglycoside antibiotics, and vancomycin.

Pharmacodynamic variation

People vary greatly in their pharmacodynamic responses, and that variability may be compounded by the effects of disease, as the following examples show.

Liver disease may influence pharmacodynamic responses to certain drugs by several mechanisms. Blood clotting may be impaired in cirrhosis and acute hepatitis because of reduced production of clotting factors; patients with oesophageal and gastric varices caused by portal hypertension in cirrhosis are also at risk of bleeding. Drugs that may impair haemostasis, or that may predispose to bleeding by causing gastric ulceration, should be avoided; these include anticoagulants and non-steroidal anti-inflammatory drugs (for example aspirin, indometacin, ibuprofen).

In patients with hepatic encephalopathy (hepatic coma or precoma), the brain is more sensitive to the effects of sedating drugs. It is therefore wise to avoid opioid and other narcotic analgesics and barbiturates; chlorpromazine dosage should be reduced. Chlormethiazole or a short-acting benzodiazepine may be used cautiously as a tranquillizer.

Diuretics used to treat ascites and peripheral oedema may precipitate hepatic encephalopathy, particularly if diuresis is too rapid.

In hepatic cirrhosis, certain drugs may exacerbate sodium and water retention; they include indometacin, corticosteroids, and carbamazepine.

The pharmacodynamic effects of some drugs may be altered by changes in fluid and electrolyte balance. For example, both hypokalaemia and hypercalcaemia potentiate the toxic effects of digoxin. The class I antiarrhythmic drugs, such as quinidine, procainamide, and disopyramide, may be more arrhythmogenic in hypokalaemic patients, and this combination particularly increases the risk of polymorphous ventricular tachycardia. Hypocalcaemia prolongs the action of muscle relaxants such as tubocurarine. Fluid depletion enhances the hypotensive effects of antihypertensive drugs.

Non-dose-related adverse reactions

Non-dose-related adverse drug reactions are caused by immunological and pharmacogenetic mechanisms. Allergic drug reactions are:

- Unrelated to the usual pharmacological effects of the drug.
- A delay often occurs between the first exposure to the drug and the subsequent adverse reaction.
- Very small doses of the drug may elicit the reaction once allergy is established.
- The reaction disappears on withdrawal; and the illness is often recognizable as a form of immunological reaction, for example rash, serum sickness, anaphylaxis, asthma, urticaria, angio-oedema.

Factors associated with an increased risk of allergic drug reactions include a history of allergic disorders (patients with a history of atopic disease and those with hereditary angio-oedema) and HLA status (the risk of nephrotoxicity from penicillamine is increased in patients with the HLA types B8 and DR3, and reduced in patients with HLA DR7; the risk of skin reactions with penicillamine is associated with HLA DRw6, and that of thrombocytopenia is associated with HLA DR4; patients with HLA DR4 also have a greater risk of the lupus-like syndrome (see below) when it is associated with hydralazine).

Drug allergy and its manifestations are classifiable according to the classification of hypersensitivity reactions, i.e. into four types, I to IV (see Section 5).

Type I reactions (anaphylaxis; immediate hypersensitivity)

In type I reactions the drug or metabolite interacts with IgE molecules fixed to cells, particularly tissue mast cells and basophil leucocytes. This triggers a process that leads to the release of pharmacological mediators (histamine, 5-hydroxytryptamine, kinins, and arachidonic acid derivatives), which cause the allergic response.

Clinically, type I reactions manifest as urticaria, rhinitis, bronchial asthma, angio-oedema, and anaphylactic shock. Drugs likely to cause anaphylactic shock include penicillins, streptomycin, local anaesthetics, and iodide-containing radiographic contrast media.

Type II reactions (cytotoxic reactions)

In type II reactions a circulating antibody of the IgG, IgM, or IgA class interacts with a hapten (drug) combined with a cell membrane constituent (protein), to form a hapten–protein/antigen–antibody complex. Complement is then activated and cell lysis occurs. Most examples are haematological: thrombocytopenia from quinidine or quinine ('gin and tonic purpura'), and occasionally rifampicin; 'immune' neutropenia, which can be difficult to distinguish from neutropenia occurring as a direct toxic effect on the bone marrow—phenylbutazone, carbimazole, tolbutamide, anticonvulsants, chlorpropamide, and metronidazole have all been incriminated; and the haemolytic anaemias that penicillins, cephalosporins, rifampicin, and quinidine can also produce by this mechanism.

Type III reactions (immune complex reactions)

In type III reactions, antibody (IgG) combines with antigen, that is the hapten–protein complex, in the circulation. The complex so formed is deposited in the tissues, complement is activated, and damage to capillary endothelium results.

Serum sickness, with fever, arthritis, enlarged lymph nodes, urticaria, and maculopapular rashes, is the typical drug reaction of this type. Penicillins, streptomycin, sulphonamides, and antithyroid drugs may cause it. Another type III reaction is the acute interstitial nephritis that may be caused by penicillins, some non-steroidal anti-inflammatory drugs, and some diuretics.

Type IV reactions (cell-mediated or delayed hypersensitivity reactions)

In type IV reactions, T lymphocytes are sensitized by a hapten–protein antigenic complex. When the lymphocytes meet the antigen an inflammatory response ensues. Examples are the contact dermatitis caused by local anaesthetic creams, antihistamine creams, and topical antibiotics and antifungal drugs. Rashes in response to sulphonamides and thiacetazone are more common in people infected with HIV.

Pseudoallergic reactions

'Pseudoallergy' (included here for convenience) is a term applied to reactions that resemble allergic reactions clinically but for which no immunological basis can be found. For example, asthma and rashes caused by aspirin are pseudoallergic reactions. In some asthmatics aspirin may trigger an attack of asthma. Aspirin-sensitive asthmatics are often sensitive to other salicylates and to other non-steroidal anti-inflammatory drugs, such as indometacin and ibuprofen. In addition, about half of aspirin-sensitive asthmatics are also sensitive to tartrazine (E102), a yellow dye used to colour some medicines and foodstuffs.

In some patients the administration of ampicillin or amoxicillin causes a maculopapular erythematous rash resembling the toxic erythema which can occur in penicillin hypersensitivity. However, the so-called ampicillin rash seems not to have an immunological origin. It can be distinguished from true penicillin hypersensitivity by its later onset after the first dose (typically 10–14 days compared with 7–10 days in penicillin hypersensitivity, though they overlap) and non-recurrence after re-exposure. Unlike penicillin hypersensitivity it carries no increased risk of a serious allergic response to other penicillins. An ampicillin rash occurs in about 1 per cent of the normal population, but in a much higher proportion of some groups of patients: it occurs almost invariably in patients with some viral infections (for example infectious mononucleosis, cytomegalovirus infection, measles), lymphomas, and leukaemias, and the risk is increased in patients taking allopurinol.

Clinical manifestations of allergic reactions

The immunological mechanistic approach does not always fit the clinical presentation, in which one is generally faced with some allergic syndrome.

Drug fever as an isolated phenomenon can occur with penicillins, phenytoin, hydralazine, and quinidine. Such fevers are usually of low grade and the patient is generally not very ill. The fever subsides within a few days of stopping the drug. With the penicillins it can sometimes be difficult to distinguish drug fever from a fever that persists because of resistant infection. Fever is also a manifestation of the neuroleptic malignant syndrome, a rare serious idiosyncratic adverse reaction of unknown cause, usually manifested by a sudden onset of fever, akinesia, rigidity, reduced consciousness, and autonomic disturbances including tachycardia and hypertension; it can kill, and demands emergency treatment by active cooling and intravenous dantrolene.

Rashes of several types may occur, including toxic erythema (due for example to antibiotics, sulphonamides, thiazide diuretics, frusemide (furosemide), sulphonylureas, phenylbutazone), urticaria (penicillins, codeine, dextrans, radiographic contrast media), erythema multiforme (penicillins, sulphonamides, barbiturates, phenylbutazone), erythema nodosum (sulphonamides, oral contraceptives), cutaneous vasculitis (sulphonamides, thiazide diuretics, allopurinol, indometacin, phenytoin), exfoliative dermatitis and erythroderma (gold salts, phenylbutazone, isoniazid, carbamazepine), photosensitivity (amiodarone, sulphonamides, thiazide diuretics, sulphonylureas, tetracyclines, phenothiazines, nalidixic acid), fixed eruptions (barbiturates, sulphonamides, tetracyclines), and toxic epidermal necrolysis (Lyell's syndrome) (phenytoin, sulphonamides, gold salts, tetracyclines, allopurinol).

Drugs may produce thrombocytopenic purpura (for example quinine, digitoxin, rifampicin) or purpura from capillary damage or fragility without thrombocytopenia (corticosteroids, thiazide diuretics).

A syndrome mimicking systemic lupus erythematosus, often involving joints and generally without renal involvement, may follow treatment with hydralazine, procainamide, phenytoin, or ethosuximide. Although this reaction is conveniently discussed here it is to some extent dose dependent, as the risk is increased at higher drug doses, and with hydralazine and procainamide it is more common among slow acetylators.

Asthma occurring as a pseudoallergic reaction to non-steroidal anti-inflammatory drugs and tartrazine has been noted above. Other adverse reactions in the lung include pneumonitis associated with the lupus-like syndrome (see above), pulmonary eosinophilia, and fibrosing alveolitis.

Jaundice may occur as an allergic response to some drugs through either cholestasis (for example with phenothiazines, erythromycin, and chlorpropamide) or generalized liver damage (for example with halothane, isoniazid and monoamine oxidase inhibitors).

Long-term effects causing adverse reactions

Some adverse effects during long-term therapy are related to both the duration of treatment and the dose.

Adaptive changes

These are the basis of some adverse reactions. Examples include the development of tolerance to and physical dependence on narcotic analgesics, and tardive dyskinesia in some patients receiving long-term neuroleptic therapy for schizophrenia.

Rebound phenomena

When adaptive changes occur during long-term therapy, sudden withdrawal of the drug may result in rebound reactions. Examples include the typical syndromes that occur after the sudden withdrawal of narcotic analgesics or of alcohol (delirium tremens). Sudden withdrawal of barbiturates may result in restlessness, mental confusion, and convulsions, and a similar syndrome in which anxiety features prominently may occur after the sudden withdrawal of benzodiazepines. Sleeplessness may also be a feature of the sudden withdrawal of these and a variety of other hypnotic drugs. Sudden withdrawal of some antihypertensive drugs may result in rebound hypertension; this was common with clonidine. Sudden withdrawal of b-adrenoceptor antagonists may result in rebound tachycardia and arrhythmia, sometimes precipitating myocardial ischaemia.

Sudden withdrawal of corticosteroids can cause acute adrenal insufficiency; after long-term administration withdrawal should be very slow.

Reversal of the effects of heparin with protamine sulphate may be associated with rebound hypercoagulability and an increased risk of thromboembolism. However, this risk may have to be taken when heparin overdosage has caused life-threatening bleeding. But the withdrawal of oral anticoagulants, such as warfarin, does not lead to rebound hypercoagulability.

Other long-term effects

Chloroquine may accumulate in the corneal epithelium (causing a keratopathy) and in the retina (causing a pigmentary retinopathy and blindness). The former occurs in over 90 per cent of patients on long-term therapy; the latter is less common but it is more serious. The risk increases with daily doses of over 4 mg/kg and in patients also taking probenecid.

Some of the long-term adverse effects of amiodarone are caused by the deposition in the tissues of lipofuchsin, including a neuropathy, pulmonary alveolitis, liver damage, microdeposits in the cornea, and skin phototoxicity.

Delayed effects causing adverse reactions

Carcinogenesis

The incidence of vaginal adenocarcinoma is increased in the daughters of women who took diethylstilbestrol during pregnancy (hoping to prevent threatened abortion). The incidence of uterine endometrial carcinoma is probably increased in women taking oestrogen replacement therapy for menopausal symptoms, and oral contraceptives increase the incidence of benign liver tumours. Anabolic steroids carry an increased risk of liver tumours.

Various anticancer drugs increase the risk of tumours. Examples include the increased risk of bladder cancer with long-term cyclophosphamide, and of non-lymphocytic leukaemias with alkylating agents such as melphalan, cyclophosphamide, and chlorambucil. Similarly, patients on immunosuppressive drug regimens, such as azathioprine with corticosteroids, have a greatly increased risk of developing lymphomas. This has mainly been noted after renal transplantation, but also in other patients.

Adverse reactions associated with reproduction

Some drugs impair fertility. For example, cytotoxic drugs can cause ovarian failure with amenorrhoea. Sperm production may be reversibly impaired by sulphasalazine, nitrofurantoin, monoamine oxidase inhibitors, and antimalarials, and irreversibly by cytotoxic drugs.

Teratogenesis

Teratogenesis occurs when a drug taken early in pregnancy causes a developmental abnormality in a fetus. The first trimester of pregnancy, and particularly the period from the second to the eighth weeks of gestation, the period of organogenesis, is the most critical, and during this time drugs may cause structural abnormalities. The brain is vulnerable throughout pregnancy.

For a drug to be teratogenic it must first pass across the placenta. The drugs that do this are those that have a low molecular weight, are poorly ionized at

physiological pH, and are very fat soluble. The few drugs that do not pass across the placenta illustrate these principles. For example, heparin is ionized and of high molecular weight and tubocurarine is ionized and relatively lipid insoluble; neither crosses the placenta. However, most drugs in the maternal circulation do reach the fetus to some extent.

If a drug is a known teratogen in humans or animals then the data sheet will say so. However, if a drug is not known to be teratogenic in humans, lack of evidence of teratogenicity in animals cannot be taken as evidence that the drug is not teratogenic in humans. Many new drugs are introduced with the advice that they should not be taken during pregnancy, simply for lack of evidence. Table 6 gives important examples of drugs to avoid during pregnancy.

Adverse effects on the fetus during the later stages of pregnancy

Some drugs that are not teratogenic may have adverse effects on the fetus if given later in pregnancy. Table 7 lists some important drugs that should be avoided or used with care during later pregnancy (some throughout the whole duration of pregnancy).

What should be done if a woman of childbearing potential is given a drug, and then finds out days or weeks later that she is pregnant? First, it is important to identify the drug and the exact time of exposure to it. If it is a known or a likely teratogen, the relation between the time of exposure and the likely time of conception should be determined. It is sometimes possible to identify the precise date of conception, but if it is not, one should try to estimate the gestational age by carefully documenting the recent menstrual history and, if necessary, date the pregnancy by ultrasound. If exposure to a known teratogen has occurred during the first 8 weeks of pregnancy, further investigation may be necessary to identify precise fetal abnormalities. For example, ultrasound can detect many structural abnormalities, and neural tube defects may be diagnosed by measuring serum and amniotic a-fetoprotein concentrations. In such circumstances any advice on termination of a pregnancy should be based on a consideration of the risk of fetal abnormality from both published information and investigation of the individual case.

Adverse reactions to drugs in breast milk

Some drugs can cause adverse effects in babies after ingestion in breast milk. These include drugs that are so extensively excreted in the milk as to cause dose-related adverse effects in the infant, and drugs that do not necessarily enter the milk in large amounts, but whose adverse effects are not dose related. The latter include drugs that may cause hypersensitivity reactions (for example penicillins and sulphonamides), and drugs that are hazardous to babies with glucose 6-phosphate dehydrogenase deficiency (for example nitrofurantoin and primaquine). Drugs to be avoided are listed, for example, in the British National Formulary and the Physician's Desk Reference. If the safety of a drug is in doubt it is best either to choose another drug or, if the drug must be used, to advise the mother to suspend breast feeding.

Methods used in detecting adverse reactions

During the period of clinical trial that a drug undergoes before its general release only the most frequent of adverse reactions will be detected, because so few patients are studied. We need methods for detecting adverse reactions as quickly as possible after marketing, for confirming that the events detected are truly adverse reactions, and for assessing their overall incidence, so that we can try to judge the balance of benefit and harm. Table 8 summarizes some methods of doing this.

Prevention of adverse effects: the role of consumers†

Ordinary people taking a medicine do not usually expect unwanted effects, even if they have received information about them from their doctor, the pharmacist, or most often in the leaflet with the medicine. Most people are optimists, and believe that a medicine prescribed by a doctor or sold in a pharmacy will not cause them any problem—though others might suffer some unpleasant effects.

The aim of information that patients get about adverse effects is to warn that such effects may occur, so that they can:

1. assess the potential disadvantages of the medicine, before deciding whether to take it ;
2. connect an adverse event with the taking of the medicine and take appropriate action.

What the information rarely, if ever, does is to explain in what specific ways adverse effects may be prevented or at least minimized. There are even many doctors who do not understand this very clearly. The necessary strategy has several components, each related to a particular category of adverse effects.

First come the dose-dependent effects known to occur in everybody if the dose is high enough. They can be prevented or minimized by keeping the dosage as low as possible. If harm is more likely with long continued use of the drug, then the duration of use should be limited, as for example with the neuroleptic antipsychotic drugs.

Second, the effects of drug interactions, which can be prevented by avoiding concurrent use of interacting drugs, or minimized by very careful monitoring if the drugs are essential.

Third are new adverse effects that have not yet been reported, and are a puzzling and unpleasant surprise for everyone concerned. These generally affect such a small proportion of patients that they have remained undetected in premarketing clinical trials. Such surprises are rarer after a drug has been widely used for several years. They can therefore be partially prevented by avoiding the use of new drugs whenever possible, at least during the first few years of their life. If there is a compelling reason for using a new drug, this should be explained to the patient and documented in the records. Furthermore, use of the new drug should be monitored more closely than would be necessary with an older and familiar drug. It is thus important that patients should know whenever the drug prescribed for them is new.

Fourth, adverse effects may result when a prescriber uses a drug with which he or she is insufficiently familiar. So when a doctor is not sure what to prescribe, and at last finds something to try in one of his or her compendia, the patient can reasonably ask how familiar the doctor is with that drug. If it is new to the doctor, then hesitation is appropriate and it may be reasonable to seek a second opinion from someone more experienced.

Fifth, some individuals are especially liable to suffer adverse effects. They may have a genetic abnormality—for example they may be poor metabolizers of a wide range of drugs, or they may lack normal plasma cholinesterase, leading them to suffer suxamethonium apnoea. The development of genetic testing in the coming years will greatly help to identify these people and so enable us to protect them from many kinds of drug-related harm. Meanwhile the best that can be done is always to take a careful medication history, asking about any previous bad experience with medicines, and where a genetic problem is suspected, also a family history. When a genetic abnormality has led to an adverse effect, the patient must be given detailed information about it so that he or she can warn any treating doctor in order to avoid future dangerous exposures.

It can be concluded from all this that all of us, as consumers who are liable to become patients at some time, need to understand a few fundamental principles:

1. All drugs have not only the intended beneficial effects, but can and often do cause harmful or unpleasant effects as well. This knowledge should influence the decision whether or not to accept medication that the doctor or pharmacist or anyone else proposes.
2. Harmful or unpleasant effects are more likely and are often worse with higher than lower doses. It is therefore important to use the lowest dosage that works, and to do so with the doctor's help.
3. Well known and established drugs are safer than new or relatively untried drugs. It is therefore best to avoid new drugs unless there is a convincing reason for using one.
4. If it is necessary to use more than one medicine at the same time, it is important to check that they are safe to use together, and to monitor their use with special care.
5. People can differ greatly in how drugs affect them. We should each learn from our own personal experiences with medicines, and make sure our doctors learn from them too in what they prescribe for us individually.

It will take much effort to educate consumers to the point where they can apply these principles in consultations with health professionals and in their use of

medicines. We as professionals must take the lead and teach them, at least by example.

Finding the lowest effective dose: two examples†

A retired economist's hypertension was well controlled by atenolol 25 mg/day, but she was bothered by her slow pulse rate, often below 60 beats/min, and asked her doctor if she could take a smaller dose. He said yes, why not.

The 25 mg tablets are not scored, but the patient succeeded in cutting them more or less in half, rejecting any fragments that looked too small. The blood pressure remained well controlled on half the previous dose, and the pulse rate rose modestly.

Another patient, a biologist with ischaemic heart disease, had a raised serum cholesterol, which a low-fat diet had reduced only slightly to 6.1 mmol/l. On simvastatin 10 mg/day this fell to 5.3 mmol/l, which was still above the target level of 4.5. Simvastatin is made only in 10, 20 and 40 mg tablets, so the doctor wanted to prescribe 20 mg/day. The patient felt that doubling the dose might be unnecessary, so the doctor suggested taking 15 mg/day or one and a half tablets. (That is easier to remember than taking one tablet (10 mg) and two (20 mg) on alternate days.) Although the tablets are not scored they have an oval shape and are easy to break in two. The dose of approximately 15 mg/day lowered the serum cholesterol to 4.4 mmol/l.

Biological variation is often neglected in clinical trials and in the dosage recommendations that are derived from them. New drugs are commonly marketed at higher doses than some patients need, because the company wants to be sure of achieving a therapeutic effect in the majority.The result is that the most sensitive patients get more than they need or than is good for them. It is worth asking patients about their experience of drugs—whether they have found small doses effective in the past, or whether they have been ineffective. Dose titration with the patient's help can on occasion improve treatment. And to facilitate this, all tablets should be scored—manufacturers please note.

Drug interactions

A drug interaction occurs when the effects of one drug (the object drug) are altered by the effects of another drug (the precipitant drug). Usually this results in an adverse drug reaction, but in a few cases a drug interaction may prove beneficial. Interactions form about 7 per cent of all adverse drug reactions; among the few patients who die from adverse drug reactions (about 4 per cent of all deaths) about a third are due to interactions.

Drugs likely to precipitate interactions include the following: drugs that are highly protein bound (aspirin, phenylbutazone, sulphonamides, and trichloracetic acid, a metabolite of chloral hydrate), as they are likely to displace object drugs from protein-binding sites; those that stimulate the metabolism of other drugs, including various anticonvulsants (phenytoin, carbamazepine, phenobarbital), rifampicin, griseofulvin. Those that inhibit the metabolism of other drugs, including allopurinol, chloramphenicol, cimetidine, metronidazole and other imidazoles (for example ketoconazole), monoamine oxidase inhibitors, azapropazone and related drugs, and quinolone antibiotics (for example ciprofloxacin); those that affect renal function and alter the renal clearance of object drugs (for example diuretics, probenecid).

The most likely object drugs in interactions are those that have a steep dose–response curve (i.e. drugs for which a small change in dose results in a relatively large change in therapeutic effect, important in interactions causing decreased efficacy of the object drug), and those that have a low therapeutic index (i.e. drugs for which the dose at which toxic effects start to occur is little more than the therapeutic dose, important in interactions causing toxic effects of the object drug). Drugs that fulfil these criteria include the aminoglycoside antibiotics, anticoagulants, anticonvulsants, antihypertensive drugs, cardiac glycosides, cytotoxic and immunosuppressant drugs, oral contraceptives, and drugs that act on the central nervous system.

Drug interactions can be classified by mechanism; Table 9 lists some important examples.

Pharmaceutical interactions

Pharmaceutical interactions are physicochemical, either of a drug with an intravenous infusion solution or of two drugs in the same solution. Such interactions result in the loss of activity of the object drug. Pharmaceutical interactions are too numerous to remember in detail, but they can be avoided by adhering to some simple principles. These include giving intravenous drugs by bolus injection if possible or via an infusion burette, by not adding drugs to infusion solutions other than dextrose or saline, and by avoiding mixing drugs in the same infusion solution, unless the mixture is known to be safe (for example potassium chloride with insulin).

Pharmacokinetic interactions

Pharmacokinetic interactions occur when the precipitant drug alters the absorption, distribution, or elimination (metabolism or excretion) of the object drug.

Absorption interactions

One drug can alter the absorption of another, but such effects are rarely important. Exceptions include the interactions of cholestyramine with warfarin and digitoxin, whose initial absorption and reabsorption after biliary excretion are reduced resulting in increased dosage requirements. Oestrogens are metabolized in the liver and some of their metabolites are excreted in the bile, deconjugated by bowel organisms, and reabsorbed. The effect of the combined oral contraceptive can therefore be reduced and pregnancy can result if a poorly absorbed antibiotic such as ampicillin prevents the bacterial deconjugation of the oestrogen.

There are two good examples of beneficial absorption interactions. Metoclopramide hastens gastric emptying and so hastens the absorption of analgesics used to treat an acute attack of migraine (Fig. 7). Charcoal binds certain drugs in the gut and thus prevents their initial absorption or their reabsorption after biliary excretion or intestinal secretion. This is valuable in the treatment of self-poisoning with drugs such as phenobarbital and tricyclic antidepressants.

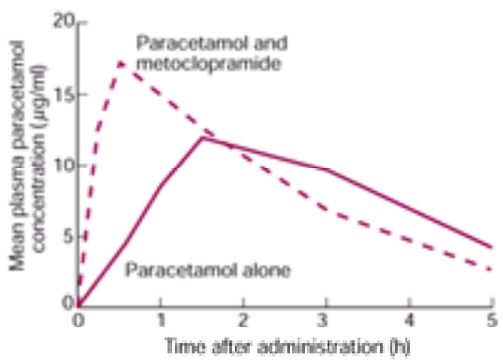

Fig. 7 The effect of metoclopramide on the absorption of paracetamol. When metoclopramide (10 mg intravenously) was given with paracetamol (1.5 g orally) the rate of paracetamol absorption was increased, as evidenced by a higher and earlier peak paracetamol plasma concentration. (Adapted from Nimmo J *et al.* (1973). Pharmacological modification of gastric emptying: effects of propantheline and metoclopramide on paracetamol absorption. *British Medical Journal* **i**, 587–9, with permission.)

Protein-binding displacement interactions

Displacement of one drug by another from its sites of binding to plasma proteins will cause an increase in the circulating concentration of unbound drug, and thus the potential for an increased effect of the displaced drug. Such interactions are important if the object drug is highly protein bound (greater than 90 per cent) and has a low apparent volume of distribution. The important drugs concerned are warfarin, phenytoin, and tolbutamide.

The most common precipitant drugs in protein-binding displacement interactions are sulphonamides, salicylates, and chloral hydrate and some of its congeners (because of their metabolite, trichloracetic acid). In addition, valproate specifically displaces phenytoin.

However, the importance of protein-binding displacement interactions has been exaggerated, and they are often clinically unimportant. The reason is that when drugs such as warfarin, phenytoin, and tolbutamide are displaced their rates of clearance increase in proportion to the degree of displacement. This means that the total concentration of drug in the plasma will fall after displacement, negating the initial effect. Thus, if the patient weathers the initial increase in unbound concentration of the object drug, the interaction will not matter.

Interactions through induction of metabolism

Certain drugs increase ('induce') drug metabolism by increasing the amount of endoplasmic reticulum in hepatocytes and by increasing the content of cytochrome P450 enzymes and cytochrome *c* reductase, which catalyse mainly oxidative reactions. Induction of the metabolism of an object drug in this way causes a reduction in its effects (resulting, for example, in epileptic fits while on phenytoin, or pregnancy while on an oral contraceptive). Drugs that induce drug metabolism include barbiturates, carbamazepine, griseofulvin, phenytoin, and rifampicin.

Interactions through inhibition of metabolism

Certain drugs inhibit drug metabolism. Interactions of this type fall into two categories: those in which the precipitant drug inhibits one or more particular cytochrome P450 enzymes and those involving other specific metabolic pathways.

Important examples of inhibition of drug metabolism by inhibition of oxidative reactions are: inhibition of warfarin metabolism by cimetidine (Fig. 8), metronidazole, and other imidazoles, chloramphenicol, norfloxacin, and other quinolones, phenylbutazone, azapropazone, sulfinpyrazone; inhibition of phenytoin metabolism by isoniazid; inhibition of theophylline metabolism by quinolone and macrolide antibiotics.

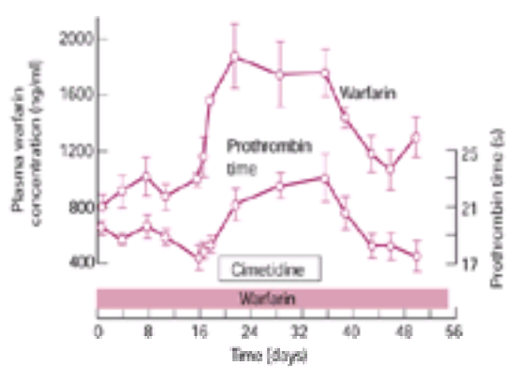

Fig. 8 An example of an interaction involving inhibition of drug oxidation. Plasma warfarin concentrations and the prothrombin time both rose after cimetidine (200 mg three times daily) was introduced in volunteers taking daily maintenance doses of warfarin. (Adapted from Serlin MJ *et al*. (1979). Cimetidine: interaction with oral coagulants in man. *The Lancet* **ii**, 317–19, with permission. ©*The Lancet*.)

The interaction of allopurinol with azathioprine and 6-mercaptopurine results from the effect of inhibition of a specific metabolic pathway. Both 6-mercaptopurine and azathioprine (which is metabolized to 6-mercapto-purine) are metabolized by xanthine oxidase, which allopurinol inhibits.

The interaction of monoamine oxidase inhibitors with dietary tyramine results in severe hypertension, which can kill. Inhibition of monoamine oxidase results in an increase in the noradrenaline content of sympathetic nerve endings. When tyramine is ingested monoamine oxidase normally inactivates it in the gut wall; however, when monoamine oxidase is inhibited tyramine passes through the gut wall and liver and reaches the systemic circulation. Tyramine releases noradrenaline from its increased stores in nerve endings and a hypertensive crisis results.

Excretion interactions

Most interactions involving drug excretion occur in the kidneys.

Inhibition of renal tubular secretion

Probenecid inhibits the tubular secretion of penicillins and cephalosporins, prolonging their therapeutic effects. Quinidine, verapamil, and amiodarone inhibit the tubular secretion of digoxin and salicylates inhibit the active secretion of methotrexate; in both cases toxic effects can occur.

Increased renal tubular reabsorption

Diuretics that inhibit renal tubular sodium reabsorption cause compensatory reabsorption of lithium with consequent toxicity.

Reduced renal tubular reabsorption

Changing the pH of the urine will alter the reabsorption of drugs that are subject to passive reabsorption. This is put to use in the treatment of overdose with salicylates and amphetamines (by alkalinizing or acidifying the urine respectively).

Pharmacodynamic interactions

Pharmacodynamic interactions occur when the precipitant drug alters the effect of the object drug at its site of action.

Interactions at the same site

Direct pharmacodynamic interactions occur when two drugs either act on the same site (antagonism or synergism) or act on two different sites with a similar end result. Many antagonistic interactions are therapeutically beneficial, including the reversal of the effects of opiates with naloxone and the reversal of the actions of warfarin by vitamin K. In contrast, synergistic interactions are often adverse. The effects of warfarin may be increased or decreased by changes in the affinity of warfarin for vitamin K epoxide reductase (clofibrate, D-thyroxine, anabolic steroids), alterations in the synthesis rate of clotting factors (anabolic steroids), changes in the activity of clotting factors (tetracyclines), or decreased availability of vitamin K secondary to decreased plasma lipids (D-thyroxine, anabolic steroids).

The effects of depolarizing skeletal muscle relaxants are potentiated by some antibiotics (for example aminoglycosides, polymixin B, and colistin) and by quinidine and quinine. These interactions are due to the curare-like effects of the precipitant drugs on the motor endplate of skeletal muscle.

Concurrent use of verapamil and a b-adrenoceptor antagonist increases the risks of cardiac arrhythmias and heart failure.

Interactions at different sites

Drugs that depress central nervous function may potentiate each other, whether or not they have effects on the same receptors. The most common example is that of alcohol with any centrally acting drug (Fig. 9).

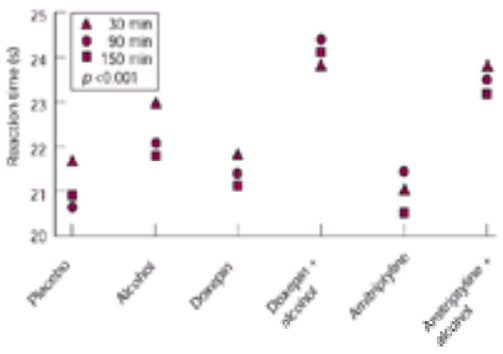

Fig. 9 A pharmacodynamic interaction. Neither doxepin (20 mg three times daily) nor amitriptyline (20 mg three times daily) alone altered the reaction time. Alcohol alone prolonged the reaction time slightly at all times of testing after drug administration (30, 90, and 150 min). However, the combination of alcohol with either doxepin or amitriptyline prolonged the reaction time by much more than one would expect from the separate effects of each component of the combination. (Adapted from Seppala T, *et al.* (1975). Effect of tricyclic antidepressants and alcohol in psychomotor skills related to driving. *Clinical Pharmacology and Therapeutics* **17**, 515–22, with permission.)

Other examples include the many combinations of cytotoxic drugs used to treat lymphomas and leukaemias, and the use of combinations of antibiotics in the treatment of some infections, even when only one organism is implicated (for example in infective endocarditis and tuberculosis). These are beneficial interactions.

Indirect pharmacodynamic interactions

In indirect pharmacodynamic interactions a pharmacological, therapeutic, or toxic effect of the precipitant drug in some way alters the therapeutic or toxic effect of the object drug, but the two effects are not themselves related and do not themselves interact.

Warfarin and other anticoagulants may be involved in indirect interactions if platelet aggregation is reduced (for example by salicylates, dipyridamole, sulfinpyrazone, mefenamic acid, and other non-steroidal anti-inflammatory drugs), when there is drug-induced thrombocytopenia, if a drug causes gastrointestinal ulceration (for example aspirin and other non-steroidal anti-inflammatory drugs), or if a drug causes enhanced fibrinolysis (for example the biguanides).

Alterations in fluid and electrolyte balance may secondarily alter the effects of some drugs. The effects of cardiac glycosides and the arrhythmogenic effects of some antiarrhythmic drugs (such as quinidine, procainamide, and phenytoin) are increased by potassium depletion (due to potassium-wasting diuretics, corticosteroids, and purgatives for example).

Pharmacogenetics

Pharmacogenetics is the study of the influence of heredity on the pharmacokinetics of drugs and pharmacodynamic responses to them.

Pharmacokinetic defects

The extent to which an individual metabolizes a drug is, at least in part, genetically determined. For example, monozygotic twins metabolize drugs similarly, while dizygotic twins often do not. For most drugs the variability in metabolism is unimodally distributed. However, for some the distribution is bimodal or trimodal, indicating that separate populations of people metabolize those drugs at different rates. The important pathways of drug metabolism subject to pharmacokinetic variability are acetylation, hydroxylation, and suxamethonium hydrolysis.

Acetylation

Some drugs are acetylated by the hepatic enzyme *N*-acetyltransferase, which is distributed bimodally. Fast acetylators have more *N*-acetyltransferase in the liver than slow acetylators; this is inherited as an autosomal dominant trait. The ratio of fast:slow acetylators varies among races, being for example 40:60 in Europe, 85:15 in Japan, and 95:5 in the Inuit. Drugs affected are isoniazid, hydralazine, procainamide, phenelzine, dapsone, and some sulphonamides. Slow acetylators require lower doses of isoniazid and hydralazine than fast acetylators in the treatment of tuberculosis and hypertension respectively. They are also more likely to develop the lupus erythematosus-like syndrome caused by isoniazid, hydralazine, and procainamide, and the peripheral neuropathy caused by isoniazid (which can be prevented or treated with pyridoxine). The interaction between isoniazid and phenytoin, in which isoniazid inhibits phenytoin metabolism, resulting in phenytoin toxicity, occurs more frequently among slow acetylators.

The acetylator status of an individual is easily assessed by giving a sulphonamide, such as sulphadimidine or sulphapyridine, orally and measuring the relative proportions of acetylated and total sulphonamide in a sample of urine passed 5 to 6 h later.

Oxidation

Certain varieties of oxidation are bimodally distributed, but in contrast to acetylation this is a heterogeneous group of defects and they are not all uniformly due to decreased amounts of enzyme. Individuals with impaired and normal oxidation are classified as poor and extensive metabolizers respectively. The main type of defect is the debrisoquine type, an autosomal recessive defect of cytochrome P450 of the type CYP2D6. It occurs in about 9 per cent of Caucasians, and has a lower prevalence in other racial types.

Other drugs that are affected include captopril, codeine, flecainide, metoprolol, nortriptyline, propafenone, and timolol. The dose-related adverse effects of these drugs (for example central nervous system toxicity with nortriptyline) are more likely in poor hydroxylators. Quinidine inhibits some oxidative reactions and may turn an extensive metabolizer of the debrisoquine type into a poor metabolizer.

The metabolism of mephenytoin is mediated by cytochrome CYP2C, whose activity is bimodally distributed. Poor metabolizers of proguanil (used to prevent malaria) do not convert it to the active form cycloguanil, and may fail to respond to treatment.

Sulphoxidation of penicillamine is polymorphic. Poor sulphoxidation is associated with a fourfold increase in the risk of adverse effects in rheumatoid arthritis. Adverse reactions to gold salts containing a thiol group may also be linked to poor sulphoxidation.

Disease associations with polymorphic metabolism

As some diseases may be related to the effects of environmental chemicals, it is of interest that polymorphic acetylation, hydroxylation, and sulphoxidation have other clinical associations. For example, the risks of bladder cancer may be increased in slow acetylators, of parkinsonism in poor debrisoquine hydroxylators, of bronchogenic carcinoma in extensive debrisoquine hydroxylators, and of primary biliary cirrhosis in poor sulphoxidizers.

Suxamethonium hydrolysis

Suxamethonium (succinylcholine) is metabolized in the plasma by the non-specific esterase pseudocholinesterase. Normally this happens quickly and neuromuscular blockade lasts only a few minutes. However, in some people the pseudocholinesterase is of abnormal affinity and amount, and metabolizes the suxamethonium only slowly, resulting in prolonged neuromuscular blockade. Three types of abnormalities of pseudocholinesterase occur, each inherited in autosomal recessive fashion, the dibucaine-resistant, fluoride-resistant, and 'silent' gene types.

Some individuals, perhaps 1 in 1000, have a two- or threefold higher concentration of pseudocholinesterase in the plasma and resist the effects of suxamethonium.

Pharmacodynamic defects

Some biochemical abnormalities make individuals peculiarly sensitive or resistant to the effects of certain drugs.

Red cell enzyme defects (see Section 22)

Unusual drug reactions may affect people whose erythrocytes are deficient in any one of three different but functionally related enzymes, glucose-6-phosphate dehydrogenase, glutathione reductase, and methaemoglobin reductase, which help to prevent the oxidation of various cell proteins. If such an erythrocyte is exposed to an oxidizing agent, haemolysis occurs, probably because of unopposed oxidation of sulphydryl groups in the cell membrane.

Porphyria (see Section 11)

The hepatic porphyrias, acute intermittent porphyria and porphyria cutanea tarda, are characterized by abnormalities of haem biosynthesis. Certain drugs may precipitate an attack of porphyria.

Malignant hyperthermia

This is a serious, potentially fatal, complication of general anaesthesia with halothane, methoxyflurane, and suxamethonium. It affects about 1 in 20 000 anaesthetized patients and is inherited in autosomal dominant fashion. The body temperature rises acutely to 40 to 41 °C, with muscle stiffness, tachycardia, sweating, cyanosis, and tachypnoea. Dantrolene, which decreases the amount of calcium released from sarcoplasmic reticulum, is effective.

Corticosteroid glaucoma

Intraocular pressure rises during daily use of corticosteroid eyedrops, and the rise is trimodally distributed, 65 per cent, 30 per cent, and 5 per cent of individuals having small, medium, and large increases in pressure. Those who have a large increase in pressure are at increased risk of glaucoma. Inheritance is autosomal recessive.

Vitamin D-resistant rickets

Three varieties of rickets are resistant to the effects of vitamin D (cholecalciferol): familial hypophosphataemic rickets, vitamin D dependency, and Fanconi's syndrome (see Section 12).

Coumarin resistance

Coumarin resistance is a rare defect in which 20 times the usual dose may be required to produce satisfactory anticoagulation. It has autosomal dominant inheritance; the mechanism may be resistance of the vitamin K epoxide reductase to inhibition by warfarin.

Monitoring drug therapy

Monitoring drug therapy usually involves trying to measure the clinical response directly. If this is difficult, or is not related directly in time to a dose of the drug, another measure of the pharmacological effect may be required. In some cases it may be necessary to resort to measurement of the plasma concentration of the drug.

Monitoring the therapeutic effects of drugs

Some events can be directly monitored in the individual patient, while some are monitored in a population. The latter can be applied to the individual only in terms of a statistical probability derived from the observed population variability. Examples of therapeutic events that can be monitored in the individual include frequency of seizures during anticonvulsant drug therapy, muscle power during treatment of myasthenia gravis, the frequency of attacks of angina pectoris, and body weight during diuretic therapy.

Preventive measures in medicine cannot be monitored in the individual and their effects must be gauged by population studies. Examples include the frequency of infections after immunization, the reduction of the risks of hypertension by diuretics, and the prevention of the complications of myocardial infarction by streptokinase and aspirin.

Monitoring the pharmacodynamic effects of drugs

In some circumstances the pharmacological effect of a drug can be carefully measured, followed sequentially, and used as a guide to drug therapy even though it may not be correlated precisely with the therapeutic effect. Examples include the effect of insulin on the blood glucose concentration in diabetes mellitus, anticoagulants on the prothrombin time, bronchodilators on FEV_1 and peak flow rate in bronchial asthma, and cancer chemotherapy on tumour markers.

Monitoring drug pharmacokinetics (plasma concentration measurement)

This is useful for a few drugs, namely those for which:

- there is difficulty in measuring or interpreting the clinical evidence of therapeutic or toxic effects.
- the relation between dose and plasma concentration is unpredictable.
- there is a good relation between plasma concentration and effect, which have a low therapeutic index, and which are not metabolized to active metabolites.

The drugs for which plasma concentrations are commonly and usefully measured are listed in Table 10.

Measurement may be useful when individualizing therapy (for example at the start of therapy when the relation between dose and plasma concentration in the individual is uncertain, when rapid changes in renal function alter the relation between dose and plasma concentration, or when another drug alters the relation between dose and plasma concentration), in the diagnosis of suspected toxicity, and in assessing compliance (see above).

Phenytoin

Plasma concentrations of phenytoin in the toxic range are quite well related to its acute toxic effects, but not to its long-term adverse effects, such as gingival hyperplasia, hirsutism and acne, and folate and vitamin D deficiencies. At low dosages it takes about 2 weeks of maintenance therapy to reach steady state after a change in dose, and the higher the plasma concentration the longer it takes (up to 3 weeks or longer in some patients). For this reason the dosage should not be changed too frequently. Provided the sample is not taken too soon after a dose (i.e. within 1 to 2 h), the time of sampling hardly matters for phenytoin, as plasma concentrations fluctuate little between doses.

Digoxin

Plasma digoxin concentrations correlate well with toxic effects but not with the therapeutic effect within the therapeutic dosage range. The time of blood sampling should be at least 6 h after the previous dose, and 12 h is the best time in patients taking once daily treatment. During regular maintenance dosage without a loading dose, steady state is reached after about 7 days (normal renal function) to 18 days (functionally anephric). The relation between dose and plasma digoxin concentrations is altered by renal impairment (which reduces the clearance of digoxin), for example in older people, and drug interactions (see above). Factors that alter the link between the concentrations and effects of digoxin, and which make it difficult to interpret the plasma concentration, include potassium depletion (which increases the effect of a given concentration of digitalis on the heart) and thyroid disease (hyperthyroidism decreases responsiveness and hypothyroidism increases it). Children younger than 6 months have lower plasma digoxin concentrations at a given dose than older children and adults, and they are also more resistant to the pharmacodynamic actions of digitalis; in them plasma digoxin concentrations cannot be clearly interpreted.

Lithium

Serum lithium concentrations correlate quite well with the therapeutic effect in the range 0.4 to 0.8 mmol/l. At 1.0 to 1.5 mmol/l the incidence of both acute toxicity and long-term adverse effects is increased. Concentrations above 1.5 mmol/l should be avoided. Blood samples should be taken at exactly 12 h after the previous dose, or as near to that as possible. It takes about 3 days for steady state to be reached during regular maintenance therapy, but patients vary widely, and in some it may take a week.

Monitoring of plasma lithium concentration is necessary for several reasons. Lithium is nephrotoxic and is excreted by the kidneys; toxicity is thus self-perpetuating because it causes renal damage, further retention of lithium, and further toxicity. Systemic availability varies from person to person, is altered by diarrhoea, and varies for different formulations. Changes in sodium balance alter the renal excretion of lithium; for example, renal sodium loss induced by diuretics leads to lithium retention.

Aminoglycoside antibiotics

The same principles apply to all the aminoglycoside antibiotics and we shall take gentamicin as an example. The relation between the plasma concentration of gentamicin and its therapeutic efficacy is complicated by the fact that different organisms have different sensitivities to the antibiotic. The toxic effects of gentamicin on the ears and kidneys are related to the 'peak' concentration (the highest concentration measured after a dose, usually occurring about 1 h after an intramuscular injection or the start of an intravenous infusion) and the 'trough' concentration (the concentration measured just before the next dose is due).

With standard regimens a peak plasma concentration of 5 to 9 mg/l is generally considered necessary, although when gentamicin is used together with benzylpenicillin to treat bacterial endocarditis, lower plasma gentamicin concentrations may be effective. Bacteriological measurement of *in vitro* inhibitory concentrations will help to guide therapy. Expert advice on dosage and target plasma concentrations should be obtained.

Theophylline

Plasma theophylline concentrations correlate well with therapeutic and toxic effects. Measurement is essential in any patient who has been taking oral theophylline and is to be given it intravenously.

Ciclosporin

Ciclosporin is generally measured in whole blood and the result of the assay may depend on whether the measurement technique is by immunoassay or high-performance liquid chromatography. The time to steady state is about 2 days and samples should be taken just before the next dose is due. Factors that alter the whole blood concentration of ciclosporin without a change in dose include reduced absorption (due to diarrhoea or reduced bile-salt production), reduced metabolism (due to liver disease or inhibition by drugs such as ketoconazole and cimetidine or grapefruit juice), and increased metabolism (due to enzyme-inducing drugs).

---

*Much has been taken from Aronson and White's excellent text in the third edition of the *Oxford Textbook of Medicine*.

†Text previously published in *Thérapie*.

†We are indebted to *Health Which?* (October, 2000), p.15. Consumers Association, London for permission to reproduce these examples.

Further reading

Clinical pharmacology

Laurence DR, Bennett PN, Brown MJ (1997). *Clinical pharmacology*, 8th edn. Churchill Livingstone, Edinburgh. [Lively and readable, with interesting and stimulating quotations and references.]

Ritter JM, Lewis LD, Mant TGK (1999). *A textbook of clinical pharmacology*, 4th edn. Arnold, London. [Well organized, illustrated with many clinical case vignettes.]

Pharmacological effects of drugs

Rang HP, Dale MM, Ritter JM (1999). *Pharmacology* 4th edn. Churchill Livingstone, Edinburgh.

Pharmacokinetics

Rowland M, Tozer TN (1994). *Clinical pharmacokinetics. Concepts and applications*, 3rd edn. Lea and Febiger, Philadelphia. [An excellent text that needs to be worked through systematically. Covers basic concepts, principles of kinetics as applied to drugs, therapeutic regimens, and individualization of therapy. Well illustrated with practical problems.]

Adverse effects of drugs

Dukes MNG, Aronson JK, eds (2000). *Meyler's side effects of drugs*, 14th edn. Elsevier, Amsterdam. [Adverse reactions to drugs discussed under the headings of the individual drugs or groups of drugs, arranged in chapters according to class of drug. Good indexes with separate listings for drugs and diseases. Supplemented by *Side effects of drugs Annuals.* Elsevier, Amsterdam, published annually since 1977. A special feature is minireviews of specific topics, distinguished from the main text typographically.]

Herxheimer A (1991). How much drug in the tablet? *Lancet* **337**, 346–8.

Inman WHW, ed (1986). *Monitoring for drug safety*, 2nd edn. MTP, Lancaster.

Monitoring drug therapy

Aronson JK, Hardman M, Reynolds DJM (1993). *ABC of monitoring drug therapy.* BMJ Publications, London. [Introduction to the principles of monitoring drug therapy by plasma drug concentration measurement, with separate monographs for each important drug.]

Useful website

The British National Formulary http://bnf.org/.

10.1 Diseases of overnourished societies and the need for dietary change

J. I. Mann and A. S. Truswell

Introduction

Nutrition issues at different stages of technical and economic development

Inappropriate nutrition contributes to illness and premature death in populations throughout the world. The nutritional problems of a country depend more upon the stage of technical and economic development than geographical location (Table 1). Until about 10 000 years ago our ancestors were hunter–gatherers. There are few contemporary hunter–gatherers left, but some of these have been studied, for example !Kung bushmen. Hunter–gatherers collected a wide range of plant foods, but also ate meat and fish. They ate little salt, alcohol, or milk (other than breast milk as infants), little cereal, and no refined sugar apart from honey. Studies of contemporary hunter–gatherers indicate that they do not become obese but may experience seasonal hunger. We infer also that malnutrition was uncommon unless illness or injury supervened. High blood pressure or coronary heart disease would have been rare and plasma cholesterol was probably low. Teeth were worn down by hard food and caries was rare. With prolonged lactation, births were spaced fairly widely. The weaning period of childhood was precarious but general nutritional health was good. *Homo sapiens* evolved as hunter–gatherers and it is unlikely there has been sufficient time to adapt physiologically to many modern foods.

Many contemporary people in the Third World are peasant agriculturists living a way of life similar to that of the rural population of Western Europe and North America until the industrial revolution. They tend to rely on the one crop with the best yield and are vulnerable to crop diseases or crop toxins and droughts. Milling and refining cereals increases the risk of malnutrition. Though some foods are stored, diet is seasonal. Malnutrition can occur from lack of essential nutrients in the staple food, for example vitamin A deficiency, pellagra, kwashiorkor, and iodine deficiency disorders. Hypertension occurs (salt is available) but coronary heart disease is rare.

Urban slum and periurban shanties are homes for an increasing proportion of the growing populations of Third World countries who are pouring into overcrowded, unsanitary accommodation in vast, polluted cities. Conditions are reminiscent of the slums of London, Manchester, and New York in the nineteenth century. These people have lost their contact with the land and food traditions and for them food is expensive. Mothers of young children have to go out to work. Breast feeding is impossible and it is very difficult to keep bottle feeds hygienic. Young children are most susceptible to diarrhoeal and other infectious diseases; these, with the mothers' absence and inadequate food, can lead to marasmus. Among adults, however, some become obese, others may be alcoholics.

In affluent societies, the prosperous people in the First World do not have to worry about the problems of getting food and keeping it uninfected. Food is cheap for them and they can eat their favourite food all year round. There is a multiplicity of nutrition advice and concerns with breakthroughs and scares, science, and pseudoscience about all food. The diet is high in fat and dense in energy. Obesity and its related disorders, coronary heart disease and hypertension (with its complications), are the principal causes of death. While sports are watched on television by millions, many ordinary citizens do not undertake any physical activity that promotes health. Obesity is unfashionable but difficult to avoid and increasing. Malnutrition occurs in the frail elderly and the sick but this malnutrition is usually subclinical and identified mainly by biochemical tests.

As mortality due to infectious disease is reduced by antibiotics and immunization in most Third World countries, and as their people are living longer and increasingly adopting Western foods and labour-saving techniques, non-communicable diseases are the major causes of death. These non-communicable diseases formerly affected only the ruling and merchant elite of developing countries but there is now a growing epidemic of obesity, diabetes, hypertension, and coronary heart disease. Health authorities in all but the least developed countries have the formidable task of providing education and food policies both to prevent malnutrition and at the same time to prevent overnutrition.

Epidemiological methods used to study nutrition-related diseases

The nutritional component of non-communicable diseases is more difficult to study than classical nutrition deficiency diseases because they develop slowly and are multifactorial. The dietary factor may be a 'risk factor' rather than a direct cause. However, there is now convincing evidence that dietary change can appreciably reduce the risk of some important non-communicable diseases. Very often the first clue to the association between a food or nutrient and a disease comes from observing striking differences in disease incidence between countries (or groups within a country) which correlate with differences in nutritional intake. Sometimes, dietary changes over time in a single country have been found to coincide with changes in disease rates. Such observations give rise to hypotheses about possible diet–disease links, rather than proof of causation because many potential causative factors may be confounded by parallel dietary changes.

Retrospective or case–control studies have sometimes been used as a rapid and inexpensive way of testing hypotheses. A series of people who have been diagnosed with, for example, cancer of the large bowel, are asked what they usually eat, or what they ate before they became ill. These 'cases' are compared with at least an equal number of 'controls', people without bowel cancer, matched for age, gender, and, if possible, social condition. Weaknesses of the method include the possibility that the disease may affect food habits, that cases cannot recall their diet accurately before the cancer was diagnosed, that controls may have some condition that affects dietary habits, or that food intakes are recorded from cases and controls in a different way ('bias'). Furthermore, it is conceivable (and for coronary heart disease and cancer, likely) that dietary factors may operate many years before the condition comes to light.

Prospective or cohort studies avoid the biases involved in asking people to recall past eating habits. Information about food intake and other characteristics are collected well before onset of the disease. Large numbers of people must therefore be interviewed and examined; they must be of an age at which bowel cancer (say) starts to be fairly common (i.e. middle aged) and in a population which has a fairly high rate of this disease. The healthy cohort thus examined and recorded is then followed up for five or more years. Eventually, a proportion will be diagnosed with bowel cancer and the original dietary details of those who develop cancer can be compared with the diets of the majority who have not developed the disease. Usually, a number of dietary and other environmental factors are found to be more, or less, frequent in those who develop the disease. These, then, are apparent risk factors, or protective factors. However, they are not necessarily the operative factor. Fruit consumption may appear to be protective but perhaps, in this cohort, smokers eat less fruit and smoking may be more directly related. This confounding has to be quantified by analysing the data to see the relationship of fruit to the disease at different levels of smoking.

Definitive proof that a dietary characteristic is a direct causative or protective factor may come from one or more randomized, controlled, prevention trials. These involve either the addition of a nutrient or other food component as a supplement to those in the experimental group and a placebo taken by the control group, or the prescription of a dietary regime to the experimental group while the controls continue to follow their usual diet. Disease (and death) outcomes in the two groups are compared. Such trials have the advantage of being able to prove causality as well as potential cost/benefit of the dietary change. However, they are costly to carry out because it is usually necessary to study large numbers of people over a prolonged period. Quite often a single trial does not, in itself, produce a definitive answer but by combining the results of all completed trials in a meta-analysis more meaningful answers are obtained.

Much research involving the role of diet in chronic, degenerative disease has centred around the effects of diet on modifying risk factors rather than the disease itself.

For many chronic diseases there are biochemical markers of risk. High plasma cholesterol, for example, is an important risk factor for coronary heart disease. Innumerable studies have examined the role of different nutrients and foods on plasma cholesterol or other risk factors. Such studies are cheaper and easier to undertake than population-based studies since far fewer people are studied over a relatively short period of time. They have helped to find which foods lower cholesterol and so should help protect against coronary heart disease. It is this information that has formed the basis of the public health messages which may have contributed to the decline in the incidence of coronary disease in most affluent societies over the last 20 years.

In this age of evidence-based medicine one would ideally like to see the results of randomized, controlled trials before offering dietary advice. However, this may be impossible to achieve. For example a clinical trial to demonstrate that a particular dietary manipulation will reduce the risk of cancer may require a trial of such magnitude and duration that it is impractical and a decision as to whether or not to recommend dietary change will need to be based on a lesser degree of evidence. Such evidence might include consistent, strong associations in longitudinal studies, biological plausibility, and corroborative experimental evidence. A dose–response relationship between the putative causal factor and the disease provides particularly strong evidence that the association is causal and that modification may reduce risk.

Tools of the trade for nutritionists

The methods for measuring food and drink intake are important tools for those who study diet–chronic disease relationships. Many studies have examined the association between changing patterns of food consumption at a national level and disease rates in various countries. Such studies do no more than provide clues for further research since the dietary data do not accurately reflect the consumption by individuals. Dietary intake of individuals is assessed by means of food frequency questionnaires, diet records, and recalls which all have different strengths and weaknesses. A weakness of most methods of assessing dietary intake is that some people, especially those who are overweight or obese, tend to underestimate their intake. Food composition tables are required to convert information gathered regarding food intake to consumption of energy and essential nutrients. Sometimes it is more reliable to assess the intake of a nutrient by measuring the level in the blood or urine or activity in the body than attempting to calculate intake from a diet record or food frequency questionnaire. Urinary sodium and iodine are examples of such biomarkers. Glutathione peroxidase activity provides a measure of assessing selenium status in those with a relatively low intake.

Coronary heart disease

Experimental, epidemiological, and clinical trial data provide strong evidence for the role of nutritional factors in the aetiology of coronary heart disease and the potential for dietary modification to reduce cardiovascular morbidity and mortality in the population as a whole, in individuals at high risk, and in those who have already experienced a cardiovascular event. Prospective studies suggest a wide range of foods and nutrients which may be involved (Table 2). Foods which increase the risk of coronary heart disease, when consumed in large amounts, probably do so because they are rich in saturated or *trans*-unsaturated fatty acids, and dietary cholesterol. 'Protective' foods contain several different nutrients which may reduce cardiovascular risk. Oily fish is rich in very long chain polyunsaturated fatty acids (eicosapentaenoic and docosahexaenoic acids; C20:5, *n*-3, C22:6,*n*-3). Fruit and vegetables are good sources of antioxidant nutrients, folate, and other biologically active substances. Nuts contain several potentially 'protective' fatty acids (oleic and linoleic acids; C18:1, C18:2) as well as vitamin E. Whole grain cereals are good sources of dietary fibre as well as some unsaturated oils. An example of the findings in one epidemiological study are given in Table 3. Each of the nutrients mentioned has an appropriately favourable or adverse effect on one or more of the cardiovascular risk factors (Table 4). The fact that some prospective studies suggest a protective effect of antioxidants derived from foods, whereas others have found particular benefit from antioxidant nutrient supplements, represents the single inconsistency in a very large body of data.

Clinical trials have shown that when diet is modified to facilitate appropriate changes in the nutrients mentioned above, levels of risk factors are altered in a favourable direction and cardiovascular events are reduced, even when the intervention is started in middle age with cardiovascular disease already present. The various trials have involved different dietary interventions so that formal meta-analysis is inappropriate, nevertheless it is possible to draw certain general conclusions regarding likely benefit from various dietary changes. Most have aimed for a reduction in plasma cholesterol by manipulation of dietary fat intake. For every 1 per cent reduction in plasma cholesterol a 2 to 3 per cent reduction in cardiovascular events occurs. Thus an 8 to 10 per cent reduction in cholesterol, which can be achieved by modifying the nature of dietary fat (replacing saturated fatty acids with mono- and *cis*-polyunsaturated fatty acids and cereals, vegetables, and fruit) will result in appreciable benefit. Trials which have examined potential benefits of dietary manipulations other than those designed to lower plasma cholesterol suggest that further clinical benefit might accrue from favourable changes in other risk factors. Consumption of oily fish two or more times per week, or a small amount of fish oil taken as a supplement, have been shown to reduce cardiovascular death in those with pre-existing coronary artery disease. While increased intakes of vegetables and fruit may confer a cardioprotective effect, there is at present no convincing evidence from clinical trials of benefit associated with the use of antioxidant nutrient supplements. The role of margarines rich in plant sterols and stanols, which may further lower dietary cholesterol by preventing absorption and reabsorption of dietary cholesterol, is yet to be established.

Community programmes, aiming to achieve dietary change along the lines indicated here, have been shown to reduce cardiovascular risk factors and one—the North Karelia Project in Finland—has shown that, in the intervention county, cardiovascular disease mortality decreased to a greater extent than might have been expected on the basis of experience in other Finnish provinces. The availability of appropriate food choices at reasonable cost is an essential component of any programme aimed at reducing cardiovascular risk, since rates are highest in those of the lowest socioeconomic status. While those at the highest personal risk are likely to show the greatest individual benefit from dietary and lifestyle changes, national coronary heart disease rates will only be reduced if changes are made by the population at large. The main purpose of such recommendations is to reduce the risk of morbidity and mortality from coronary heart disease in those who are in the prime of life. Even greater reduction in morbidity and mortality and an improvement in life expectancy may occur in succeeding generations who will have reduced lifetime exposure to risk factors related to lifestyle.

Hypertension

Over 30 years ago, Dahl drew attention to the correlation between salt intake and prevalence of hypertension in populations. This and other similar studies were flawed by methodological difficulties associated with measuring salt intake and blood pressure. The more recent Intersalt Study used standardized blood pressure measurements and 24-h urinary sodium excretion, the best available method of assessing intake. The study, which involved 52 centres in 32 countries, suggested that a difference in sodium intake of around 100 mmol/day over a 30-year period might be expected to result in differences of approximately 10 mmHg in systolic blood pressure and 6 mmHg in diastolic blood pressure. Meta-analyses of observational data and intervention trials show broadly comparable results. In one such meta-analysis, of 68 crossover trials and 10 randomized controlled trials of dietary salt reduction, it appears that a reduction of daily sodium intake of about 50 mmol (about 3 g salt) would, after a few weeks, lower systolic blood pressure by an average of 5 mmHg and by 7 mmHg in those with high blood pressure; diastolic blood pressure would be lowered by half as much. It is estimated that such a reduction in salt intake by a whole Western population would reduce the incidence of stroke by 26 per cent and of coronary heart disease by 15 per cent. Reduction also in the amount of salt added to processed foods would lower blood pressure by twice as much and could prevent as many as 70 000 deaths per year in Britain. The heterogeneity in the response of individuals to sodium restriction suggests the possible existence of a group of hyper-responders but there is as yet no clear indication as to how such individuals might be defined.

Several mechanisms have been suggested to explain the association between salt intake and blood pressure, including reduced urinary sodium excretion and fluid retention by some individuals, increased sympathetic nervous system activity and impaired baroreflex function, and alterations of ion transport in vascular smooth muscle.

Obese people have higher blood pressures than non-obese people and if they lose weight blood pressure falls even if the usual salt intake is maintained on the restricted diet. An Australian trial showed, in a clinical trial setting, that weight reduction (maximum loss of 7.4 kg) compared favourably with metoprolol in the treatment of mild hypertension, and diet was associated with an improved plasma lipid profile not seen on the drug.

In epidemiological studies, blood pressure increases progressively when reported alcohol intake increases above three drinks per day. Several intervention studies have shown that reduction of alcohol intake can produce an appreciable reduction in blood pressure amongst hypertensive heavy drinkers. For example one study showed that replacing standard beer (5 per cent alcohol) with a reduced alcohol beer (0.9 per cent alcohol) produced a reduction in alcohol intake from 450 to 64 ml/week and a significant fall in blood pressure.

High intakes of potassium and calcium and a vegetarian diet have also been shown to be associated with reduced levels of blood pressure. However, it was previously believed that these were relatively unimportant compared with dietary sodium and overweight. The recent DASH (Dietary Approaches to Stop Hypertension) trial has shown a remarkably powerful blood pressure-lowering effect in association with a diet rich in fruit, vegetables, and low-fat dairy products. In hypertensive individuals, systolic and diastolic blood pressures were lowered by 11 and 6 mmHg respectively, compared with the control group. Smaller changes were

seen in the normotensive group, 3 and 9 mmHg for systolic and diastolic pressures respectively. These findings require confirmation.

Diabetes mellitus

Diet undoubtedly plays an important aetiological role in type 2 diabetes mellitus. As early as 1920, Himsworth suggested that excessive intake of energy, deficiency of dietary carbohydrate, and possibly an excessive intake of fat might increase the risk of diabetes. His conclusions were largely based on the improved glucose tolerance observed in non-diabetic individuals eating a high carbohydrate compared with a high fat/low carbohydrate diet. He also noted the association between high carbohydrate intakes and reduced mortality from diabetes and the fact that newly diagnosed diabetic patients were found to have a high intake of total energy (especially energy derived from fatty foods) before the onset of symptoms than non-diabetic controls. Although Himsworth's findings are based on studies which may not fulfil the criteria required of modern nutritional epidemiology, the conclusions have largely been confirmed by subsequent epidemiological and clinical studies.

Epidemiological surveys and longitudinal studies in many countries confirm the striking association between increasing degrees of obesity and the risk of developing type 2 diabetes mellitus. The association is strongest in those with central (android) obesity and those with a family history of the condition, reflecting the importance of genetic determinants (see also Chapter 10.5). While energy intake in excess of requirements is now universally accepted as an important risk determinant, controversy has raged with regard to the extent to which macronutrient distribution and some micronutrients might be involved in the aetiology of type 2 diabetes. Epidemiological evidence suggests that the condition is uncommon in people eating a range of 'traditional diets' high in fresh fruit, vegetables, and cereals, and therefore high in starches and non-starch polysaccharides and low in fat. Diabetes prevalence seems to increase rapidly when traditional lifestyles are exchanged for the Western way of life, particularly when such transitions occur over a short time span. Such changes have been clearly demonstrated in Micronesians, Polynesians, American Indians, and Aboriginal Australians, as well as Chinese and Asian immigrants to Mauritius. A sharp rise in the frequency of type 2 diabetes mellitus also emerges from studies of Asian Indian immigrants to Fiji, South Africa, and Britain, and of Chinese in Singapore, Taiwan, and Hong Kong. The change from traditional to a Western diet is usually accompanied by a reduction in physical activity. As lifestyle change starts to occur in China and India, by far the most populous countries, type 2 diabetes mellitus may create an enormous public health problem. Indeed, it seems reasonable to claim that the disease has already reached epidemic proportions in many indigenous populations amongst whom it is difficult to disentangle the relative importance of genetic and environmental factors. By contrast, North Western Europeans and Anglo-Celts have a relatively low susceptibility to diabetes.

It is not yet clear whether any single attribute of the Western way of life is particularly important in increasing the risk of diabetes. Excess sucrose has been largely exonerated as an important dietary factor in the aetiology of type 2 diabetes, except perhaps when, alongside a high intake of fat, it contributes towards an excessive energy intake. A high intake of saturated fatty acids undoubtedly further decreases insulin sensitivity, an underlying abnormality in type 2 diabetes. One large prospective study of health professionals in the United States has found that a high intake of low glycaemic index foods (i.e. predominantly carbohydrate-containing foods producing a relatively low glycaemic excursion after ingestion when compared with a comparable amount of glucose) tends to protect against type 2 diabetes and that the effect is independent of other individual dietary attributes. A low level of physical activity appears to be an important predisposing factor for obesity and hence type 2 diabetes. Thus, it seems most likely that a combination of factors is responsible.

Australian Aboriginals who have adopted a Western diet have high rates of diabetes. A study in Australia showed that reverting to their traditional lifestyle is associated with a marked improvement in several indices of carbohydrate metabolism. Many studies in affluent societies have shown that weight reduction in overweight people, those with impaired glucose tolerance, and people with diabetes can often result in normal, or near normal, blood glucose levels without the need for oral hypoglycaemic therapy. Furthermore, diets high in soluble forms of non-starch polysaccharides and in which low glycaemic index carbohydrate-containing foods predominate can improve glycaemic control in those with diagnosed diabetes, independent of an effect on body mass. However, in people who are not overweight, sufficient improvement to reduce the need for drug therapy and achieve even near normal blood glucose concentrations is seen only with extreme dietary change (i.e. diets consisting largely of raw and unprocessed foods and exceptionally low in fat). Intervention studies in people with impaired glucose tolerance suggest that risk of progression to diabetes can be reduced and that this results chiefly from weight loss rather than modification of individual nutrients. However, the largest intervention study involving advice to modify diet and increase physical activity carried out so far has been in progress for 6 years, and there is a suggestion that the condition may relapse with time, possibly as a consequence of reverting to previous lifestyle habits. Thus, while it appears that reducing the level of obesity in the population at large is the single measure most likely to reduce overall rates of type 2 diabetes mellitus in high-risk groups, the extent to which risk reduction can be achieved in practice remains to be established.

Although diet is important in the management of type 1 diabetes, nutritional factors do not appear to have contributed to the aetiology of the disease to the same extent as for type 2 diabetes mellitus. Genetic and other environmental factors are believed to be more important. Recent studies have suggested, however, that infants who have been breastfed may have a reduced risk of type 1 diabetes mellitus in later life and this observation could be linked with immune mechanisms known to be associated with this condition.

Cancers

The development of cancer involves several stages and occurs over a long period of time. Nutritional factors may operate at one or more of these stages. During the first stage of initiation, the DNA of the healthy cell is damaged by chance mutation or a carcinogen. During the promotion (second) stage, the 'initiated' cells may be exposed to promoters, environmental factors, which create conditions which favour their growth over that of normal cells. This phase tends to be prolonged and may be delayed or accelerated by environmental factors. Genetic factors also operate. Ultimately, preneoplastic cells are formed which differ in appearance and function from normal cells. During the final stage of progression, additional mutations tend to occur leading to the transformation of preneoplastic to neoplastic cells. Nutritional factors may act as carcinogens or promoters. Given this long natural history of the disease process, it is hardly surprising that few data from intervention trials are available and data relating dietary factors to various cancers are derived from epidemiological associations and animal experiments. Nevertheless Doll and Peto have estimated that about one-third of all cancers in Western countries may be attributed to diet. The dietary and nutritional factors which may play a role in human cancer are listed in Table 5.

Breast cancer is a common cancer amongst women in the Western world. There is a strong, positive correlation between fat intake and breast cancer rates in different countries and a pooled analysis of case–control studies suggest an increase in relative risk of breast cancer associated with a high fat intake in postmenopausal but not premenopausal women. Animals given chemical carcinogens develop fewer tumours if they are given a low fat diet. However, prospective studies have not confirmed higher rates of breast cancer in women with high fat intakes. Thus, it is not absolutely clear whether fat intake is a risk factor *per se* or simply a marker for some other important environmental factor operating in countries with high rates of breast cancer. Furthermore, high fat intakes may be associated with obesity, high circulating endogenous oestrogens, and early menarche, all of which have been clearly established as risk factors. Plant oestrogens (isoflavones from soy products, lignans in wholegrain cereals and vegetables) may also be protective by increasing the length of the menstrual cycle, thus reducing the number of menstrual cycles and reducing exposure to peak hormone levels.

Cigarette smoking is unquestionably the most important cause of lung cancer, one of the most common cancers. However, vegetables and fruit have consistently been shown to have a protective effect in more than 30 case–control and prospective studies. A large clinical trial in which heavy smokers were given supplements of a-tocopherol, b-carotene (presumed to be the protective factors in vegetables and fruit), or placebo, showed no clinical benefit but with hindsight the results were not surprising. It seems most unlikely that the carcinogenic effects of prolonged, heavy cigarette smoking can be reversed by vitamin supplements given over a short period of time.

Colorectal cancer is the second most common cause of cancer amongst males. A low intake of vegetables and non-starch polysaccharide appear to be the most consistent dietary factors associated with increased risk. Non-starch polysaccharide escapes digestion in the small intestine and is fermented in the large bowel by the colonic microbial flora. Short chain fatty acids are produced, one of which, butyrate, is an antiproliferative agent. Non-starch polysaccharide may further reduce the risk of large bowel cancer by stimulating bacterial growth and increasing the biomass as well as the unfermented non-starch polysaccharide binding with water, so leading to an increase in stool weight, dilution of colonic contents, and faster transit time through the large bowel. High fat diets are associated with increased concentration of bile acids in the colon. These are converted to secondary bile acids, of which deoxycholic acid is a known promoter of large bowel cancer. A 4-year intervention trial has shown that a low fat diet (fat providing less than 25 per cent total energy) plus 25 g wheat bran daily can reduce the frequency of large adenomas in the large bowel but this has not been confirmed in two other comparable studies. There is insufficient evidence to reliably incriminate another potentially deleterious factor: high intake of red meat. However, it is possible that the method used for cooking meat may be important. Increased intakes of heterocyclic amines may form on the surface of cooked meat, especially during grilling, frying, or cooking on a barbecue. It has also been suggested that *n*-3 polyunsaturated fatty acids may be protective because they slow down the rate of cell division of mucosal colonic cells.

Cancer of the prostate is the third most common cancer in males, with a seventy-fold variation in incidence, the highest rates (in the United States, Europe, and

Australasia) being associated with the highest intakes of fat. In contrast with most other sites, there appears to be no protective effect of vegetables and fruit.

Cancer at all other sites is less frequent, but nutritional causes appear to be involved for some. Alcohol appears to play an important role in cancers of the mouth, pharynx, oesophagus, and liver. Obesity is an important factor in endometrial cancer and may have some role in postmenopausal breast cancer in women and bowel cancer in men. A high intake of salted foods may be involved in cancer of the stomach and vegetables and fruit appear to be protective against cancers at most sites.

Obesity (see also Chapter 10.5)

Obesity is the most obvious and important nutritional disease in affluent societies, its comorbidities including type 2 diabetes, coronary heart disease, hypertension, stroke, gallstones, osteoarthritis, some cancers, and obstructive sleep apnoea. Obese people may also be disadvantaged in terms of social, economic, and psychological effects. The psychological well being of children may be particularly affected. Most of the adverse consequences of obesity are appreciably reduced by weight loss, though gallstone formation may not be reduced. While the genetic component of obesity is acknowledged, the dramatic increase in prevalence in virtually all westernized countries in recent years provides ample evidence of the overwhelming importance of environmental factors. Inactivity is unquestionably an important cause but increased consumption of readily available 'convenience' and other energy-dense foods contribute to an energy intake in excess of expenditure. It seems unlikely that the epidemic of obesity will be reversed unless the nutritional environment in which we live is altered by creating more opportunities for physical activity and improving availability of food choices, and providing appropriate health education.

Diverticular disease of the colon

The first suggestion that deficiency of non-starch polysaccharides in the diet may be implicated in the aetiology of diverticular disease of the colon came from striking geographical variations in its prevalence and the documented increase in disease rates in several European countries since the 1920s. These variations and trends in rates are certainly compatible with a causative link with low non-starch polysaccharide diets but could also be explained by several alternative dietary and other environmental influences. The best-documented evidence comes from comparisons of asymptomatic groups of vegetarians and meat eaters who volunteered to have a barium meal. Radiological diverticular disease was found more frequently amongst non-vegetarians (33 per cent) than vegetarians (12 per cent) who had appreciably higher intakes of non-starch polysaccharide. Furthermore, when comparing individuals with and without diverticular disease in both the vegetarian and non-vegetarian groups, those with diverticular disease had appreciably lower intakes of non-starch polysaccharide than those with no evidence of diverticulae following barium meals. Animal experiments provide support (for example rats given a diet low in non-starch polysaccharide have been shown to develop diverticulae, as do rabbits fed with white bread, sugar, and vitamins, and given prostigmine). An increase in non-starch polysaccharide intake is widely recommended to patients with symptomatic diverticular disease, a treatment justified by the findings of some (but not all) controlled clinical trials.

Plausible theories concerning pathogenesis have been suggested; small, hard faeces, undoubtedly seen with a diet deficient in non-starch polysaccharide, are associated with narrowing of the colon and the formation of closed segments in which pressure increases. Additional work is needed by colonic muscles to provide the pressure to move the more solid faeces, producing muscular hypertrophy in addition to the diverticula at sites of weakness where blood vessels penetrate the muscular coat.

Dental caries

Dental caries was exceptionally rare among young people in ancient Britain. Surveys over the past 15 years have suggested that as many as 80 per cent of 5-year-olds require treatment for dental caries and about 10 per cent of all children enter school with more than half their teeth seriously decayed. Some 5 per cent of the adult population in England and Wales and 15 per cent of that in Scotland are edentulous by the age of 30 years. Several strands of evidence suggest a nutritional cause. Amongst the indigenous population of many countries where unrefined foods form the bulk of the diet (e.g. China, Uganda) dental caries once had a very low prevalence. Within a few years of the addition of sugar and other refined foods the frequency showed a rapid increase. A similar change has been shown experimentally in monkeys. In a classical experiment carried out in a Swedish mental hospital, volunteers given toffee apples, chocolate, and caramel in addition to their controlled diet had a thirteen-fold greater number of tooth surfaces becoming carious each year, compared with those eating the controlled diet alone. While frequency, timing, and amount of extrinsic sugars may be important in the aetiology there is no doubt that fluoride in the water or in tooth paste can profoundly reduce the risk of dental caries.

Constipation and the irritable bowel syndrome

Nine-nine per cent of a large population sample studied in Britain reported that they defecated at least three times per week but perceived constipation is a frequent complaint. Approximately 3 per cent of all prescriptions written in the National Health Survey (in the United Kingdom) were for purgatives and laxatives, at a cost of around £4 000 000, and many times this amount must have been spent in buying these preparations over the counter. In another survey, 6 per cent of people aged between 18 and 80 years described straining when passing stools. No data are available concerning the frequency of passing small stools. There seems little doubt that constipation is uncommon in populations with a high intake of non-starch polysaccharide. In rural Africa, stool weights are frequently around 500 g daily, and bowel transit times around 40 h. In Britain, stool weights in non-vegetarians are more usually around 100 g (with a very wide range), whereas in vegetarians the average stool weight is over 200 g. Factors other than non-starch polysaccharide might be involved but British vegetarians and non-vegetarians with high average daily intakes of non-starch polysaccharide have transit times of less than 75 h and rarely report constipation, whereas those with lower intakes have transit times ranging from 20 to 124 h and frequently complain of constipation. There is no doubt that increasing the non-starch polysaccharide content of the diet (especially that derived from cereals) relieves the symptoms of constipation, an observation now confirmed by controlled clinical trials. There is no direct evidence of a causal link between a diet low in non-starch polysaccharide and the irritable bowel syndrome, but diets rich in non-starch polysaccharide are widely recommended in its treatment and are believed to be of value, even in the absence of formal clinical trials.

Osteoporosis

Osteoporosis is an important cause of morbidity amongst the elderly, especially in women, and the incidence of osteoporotic fractures is increasing steadily as people are living longer. In 1990, there were an estimated 1.66 million hip fractures world-wide. By the year 2025, it is projected that there will be 1.16 million hip fractures in men and 2.78 million in women due to osteoporosis. The aetiology of osteoporosis is complex; women have a lower peak bone mass than men and then lose bone rapidly after the menopause in association with a decline in oestrogens. Women lose approximately half their trabecular bone and a third of their cortical bone, while men lose a third of their trabecular bone and a fifth of their cortical bone. Genetic factors influence peak bone mass and bone loss and these may operate by some of the well-known risk factors: strong family history of osteoporosis, short stature, early menopause, and white or Asian race. However, there are also clearly established environmental factors, including leanness (genetic factors probably operate in addition), multiparity, cigarette smoking, and excessive alcohol intake. The role of dietary calcium has been uncertain but there is now convincing evidence that the best way of avoiding osteoporotic fractures in later life is to achieve optimal skeletal mass for one's genetic potential and to retain this as long as possible. The best means of doing so is by ensuring lifelong adequate consumption and maximum absorption and retention of calcium. The need for substantial amounts of dietary calcium, taken in conjunction with physiological amounts of vitamin D, is particularly important during the periods of growth, pregnancy, lactation, and in the postmenopausal years.

Other diseases

Gallstones, appendicitis, haemorrhoids, varicose veins, and hiatus hernia all occur frequently in developed countries and rarely in developing countries but the evidence linking these diseases to a nutritional cause in tenuous. Gallstones are undoubtedly associated with obesity. Both gallstones and appendicitis are more common in non-vegetarians than vegetarians and there are some rather indirect data suggesting an association with diets high in sugars and deficient in non-starch polysaccharides. The addition of bran to the diet can make bile less saturated and experimentally induced gallstones in animals tend to be reduced if foods rich in non-starch polysaccharides are given. Data from the United Kingdom and South Africa taken together provide interesting information concerning appendicitis; appendicitis rates were compared in two matched South African Caucasian groups, the privileged group living in university halls of residence and the other living in the establishments for the indigent where the diets contained more fibre. Annual rates were 7.8 and 1.8 per thousand, respectively. Of course, factors other than diets might explain this but the rates were strikingly similar to those found in an almost identical study in Bristol (7.6 per thousand in a public school and 0.8 per thousand in an orphanage).

The case for dietary change

Nutrition research often generates results which may be translated by researchers, self-styled 'experts', or the media into potentially confusing and conflicting messages. It is therefore critically important for governments, who develop food and nutrition policies, for doctors, others involved in health and nutrition education, and for consumers to have authoritative recommendations which represent consensus opinions of nutrition scientists. Terminology regarding such recommendations has been confusing, but the most recent British Government publication suggests a new approach. Dietary reference values are intended for policy makers and health professionals who recommend diets for individuals. These include population averages (sometimes also recommended individual minimum and maximum intakes) for the macronutrients, average requirements for total energy, and reference nutrient intakes (previously referred to as recommended dietary intakes or allowances) for vitamins and minerals. The latter values are defined as levels of intake which will satisfy the requirements of the great majority in any given population. Thus, the reference nutrient intake is set at two (estimated) standard deviations above the average of all individual requirements for each nutrient. While this approach helps to ensure nutritional adequacy, the situation has become more complex with the suggestion that some nutrients might confer protection against certain chronic diseases when intakes exceed the reference nutrient intake. Thus, it is now considered necessary to set an upper safe level beyond which toxic effects may occur. These dietary reference values will be largely meaningless to the population at large. For the general public, dietary guidelines have been developed to translate dietary reference values into practical advice.

The dietary reference values for macronutrients (Table 6) are based to a considerable extent on the evidence-based data which suggest that alteration of dietary fat intake from that typical of most Western countries is likely to reduce population and individual risk of coronary heart disease. The emphasis is on reduction of saturated and *trans* unsaturated fatty acids and the recommendation represents a substantial reduction from present levels of intake. The recommended intake of *cis*-monounsaturated acids is similar to that at present but adherence to the recommendation for saturated fat will inevitably lead to an alteration in source, from animal to vegetable, since at present much of both mono- and saturated fatty acids are derived from the same animal sources. The total intake of *cis*-polyunsaturated fatty acids has been set at a level slightly higher than at present but the proportions of *n*-3 and *n*-6 are not specified. However, a recent European consensus group has suggested that those not eating fish at least once a week should consider obtaining 200 mg of very long chain *n*-3 polyunsaturated fatty acid daily from other sources. In view of the growing evidence that plant *n*-3 polyunsaturated fatty acid might independently reduce coronary heart disease risk, an intake of 2 g/day or 1 per cent total energy from a-linolenic acid is also recommended. The group advised against the use of *n*-3 to *n*-6 ratios but favoured separate recommendations for a-linolenic acid, marine *n*-3 polyunsaturated fatty acid, and linoleic acid.

Only about 50 g of carbohydrate daily is required to avoid ketosis and at the other end of the range of intake many populations maintain an adequate nutritional status when carbohydrate provides up to 80 per cent total energy. However, most Western societies are unaccustomed to a high carbohydrate intake and are reluctant to accept substantial increases. Therefore, a modest increase in total carbohydrate is advised with a limitation on the use of non-milk extrinsic sugars because of their role in dental caries, especially when consumed frequently. Furthermore, foods high in extrinsic sugars are energy-dense and contain relatively few essential nutrients. On the other hand, it is considered appropriate (Table 6) that intrinsic sugars (that is those incorporated into the cellular structure of foods), milk sugars, and starches should provide the balance of dietary energy not provided by protein, fat, and non-milk extrinsic sugars, that is, an average of 37 per cent total dietary energy. Foods containing these macronutrients are usually good sources of essential micronutrients as well as dietary fibre (non-starch polysaccharide), an increase in which is also recommended because of the convincing evidence for benefit in terms of bowel function and possible benefits regarding a range of diseases of the large intestine, diabetes, and cardiovascular disease.

It seems conceivable that these recommendations regarding carbohydrate may be modified in the future. It is becoming increasingly clear that the health benefits associated with carbohydrate-containing foods are principally derived from carbohydrates which are not digested and absorbed in the small intestine but which enter the large bowel where they undergo fermentation or act as stool-bulkers. The process of fermentation leads to the production of fatty acids which are useful not only as fuel sources but by their antiproliferative effect may also reduce the risk of colon cancer. Carbohydrates which undergo fermentation include oligosaccharides, resistant starch, and certain non-starch polysaccharides such as gum, pectins, and mucilages. Those which remain intact and act as stool bulkers include cellulose, hemicellulose, and lignin. The suggestion that complex carbohydrates may confer health benefits is becoming outmoded since many forms of starch are rapidly digested and absorbed as sugars. Thus the extent and speed with which carbohydrates are digested and absorbed in the small intestine provide the best indicator for health-promoting, carbohydrate-rich foods. The most useful, simple indicator of this is the glycaemic index (Fig. 1). A recent World Health Organization Expert Consultation has suggested that this index might guide the choice of carbohydrate-containing foods, the lower glycaemic index foods being the preferred food choices (see Table 7).

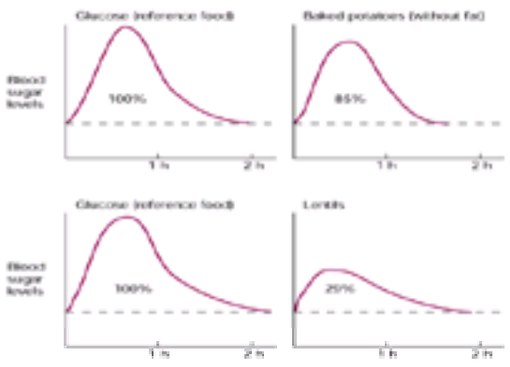

Fig. 1 Calculation of the glycaemic index of a food. The area under the blood glucose curve is calculated geometrically using the fasting level as baseline and expressed as a percentage of that seen after consumption of 50 g pure glucose. In practice, 8 to 12 people are tested and the mean of the group is the glycaemic index of the test food.

The glycaemic index concept does have limitations. It is important to emphasise that it should not be used for carbohydrate-containing foods which are also high in fat. These will usually have a low glycaemic index but are obviously not regarded as the best choices. Sucrose and sucrose-rich foods will also have a relatively low glycaemic index because of their fructose content. While the detrimental effects of sucrose have been overstated in the past, sucrose confers no health benefits other than acting as an energy source. Given the need to reduce energy dense foods in many countries sucrose-rich foods are not generally encouraged.

While appropriate distribution of macronutrients might be expected to reduce cardiovascular risk, improve bowel function, and probably also reduce the risk of certain cancers and other diseases of the large bowel, it is imperative to consider the extent to which diet can influence obesity and its comorbidities, especially type 2 diabetes which accounts for a public health problem of enormous magnitude. Increasing carbohydrate-containing foods rich in non-starch polysaccharide at the expense of saturated fat is likely to enhance satiety and so help to reduce excessive energy intake. However, increasing energy output by increasing physical activity is an equally essential component of public health measures designed to stem the tide of the obesity epidemic occurring in most affluent societies and many developing countries.

Reference nutrient intakes (adequate for most individuals) are provided for vitamins and minerals; selected examples are shown in Table 8. Clinical vitamin deficiencies, discussed in detail in Chapter 10.3 are uncommon in affluent societies except in at-risk subgroups within populations. For example immigrants who have migrated from sunny, tropical to high latitude, cloudy countries may be at risk of vitamin D deficiency; strict vegetarians (who consume no animal or dairy products) may become deficient in vitamin B_{12}, and disadvantaged groups (especially the very young, pregnant and lactating women, and the elderly) may have generally inadequate intakes.

On the other hand, inappropriate intakes of certain minerals are fairly common. Many groups are particularly vulnerable to iron deficiency, due to high physiological requirements (infants and toddlers, adolescents, pregnant women), high losses (menstruating women), or poor absorption (the elderly and people consuming foods high in inhibitors of absorption, such as fibre and tannin in tea). Vegetarians are also at increased risk of iron deficiency even when total intake of iron appears to be adequate since non-haem iron from plant foods is less bioavailable than haem iron from animal sources. Bioavailability is enhanced by the consumption, at the same time, of vitamin C. Iodine and selenium are deficient in soils in various parts of the world. Clinical selenium deficiency has only been reported from China though the consequences of lesser degrees of selenium deficiency have yet to be established with certainty, especially in regions where soils are known to be deficient (e.g. the South Island of New Zealand). Endemic iodine deficiency is widespread, especially in the Himalayas and the Andes, and clinical deficiency states are largely avoided by the use of iodized salt and sanitizers containing iodine used by the dairy industry. Interestingly, in New Zealand where goitre due to iodine deficiency had virtually been eliminated, mild iodine deficiency appears to be reoccurring possibly as a result of reduced use of iodized salt and the introduction by the dairy industry of

alternative sanitizers. Young women often have insufficient calcium to help achieve peak bone mass, and older women may have an inadequate intake to help reduce an age-related bone loss.

Excessive intakes of sodium, to such an extent that it probably contributes to hypertension and its consequences, are common throughout the Western world. Targets for reduction may be more important than reference nutrient intakes for sodium. An intake of 100 mmol/day (2.3 g/day), a level currently exceeded in most countries, might be an appropriate maximum.

Folate is probably the best example of a nutrient for which intakes greater than those recommended have been shown to confer benefit. Intakes of 400 µg or more per day (reference nutrient intake 200 µg/day) can appreciably reduce the risk of neural tube defects. Furthermore, the potential for folate to reduce cardiovascular risk by a reduction in plasma homocysteine concentrations may provide further justification for higher recommended intakes if the trials underway at present demonstrate reduced cardiovascular risk. There is evidence that higher than recommended intakes of several antioxidant nutrients reduces subsequent cardiovascular risk and possibly also the risk of certain cancers. However, there are insufficient data regarding optimum intakes to contemplate firm recommendations at present.

Dietary assessment methods are not sufficiently sensitive to determine accurately intake of several important nutrients so that in order to assess either adequacy or excess in an individual or group of individuals for clinical or research purposes it is necessary to use biomarkers. Intakes of iodine and sodium are assessed by measuring amounts in 24-h urine collections. Measurement of folate concentration in the serum or red blood cells provide a good estimate of intake since the amounts in fruit and vegetables vary enormously and are also dependent upon shelf-life and method of preparation. Fatty acid composition of serum or red cell membrane provides an indication of the nature of dietary fat intake. For some nutrients which are not always readily bioavailable adequacy of intake must be assessed by alternative means. In the case of iron, measurement of ferritin in the blood (indicating iron stores) is a more useful indicator of iron status than dietary intake (see also Chapter 22.4.4).

Substantial changes in what have become traditional eating habits of many affluent societies are required in order to achieve the advised changes in distribution of macronutrients and recommended intake of all essential micronutrients. A multifocal approach is necessary if there is to be a real chance of achieving dietary change. At the policy making and government level, there needs to be a serious commitment to enabling the population as a whole to make appropriate food choices. Fatty cuts of meat, high-fat products (e.g. meat pies), and convenience foods (e.g. fish and chips, burgers) are relatively inexpensive and therefore frequently eaten by those of lower socio-economic status who have the highest rates of coronary heart disease. Policies are required which ensure that more appropriate food choices are available at reasonable cost. This is not easy to achieve in many Western countries where farmers may have considerable political influence and subsidies may be available for some high-fat dairy products such as butter and cheese. Governments and intergovernmental agencies also have the responsibility for ensuring that food labels and health claims are accurate, interpretable, and likely to facilitate health-promoting food choices, a particularly important issue given the increased consumption of packaged food.

Dietary guidelines are necessary to provide clear directions to individuals and families who wish to aim for a healthy diet pattern. These guidelines vary slightly from country to country though some are almost universal (see Box 1). Others are less consistent (see Box 2). The public also need education regarding food groups and the nutrients they contain, the interpretation of food labels, the meaning of health claims, and the methods of food preparation. The increased use of convenience and packaged food has meant that many people no longer possess basic cooking skills. They also need (and want) to know the merits and demerits of obtaining certain essential micronutrients by taking supplements or fortified food products rather than conventional foods.

Box 1 Dietary guidelines for which there is almost complete agreement

1. Eat a nutritionally adequate diet composed of a variety of foods.
2. Eat less fat, particularly saturated fat.
3. Adjust energy balance for body weight control—less energy intake, more exercise.
4. Eat more wholegrain cereals, vegetables and fruits.
5. Reduce salt intake.
6. Drink alcohol in moderation, if you do drink.

Box 2 Additional dietary guidelines in some countries

1. Recommendation regarding sugar and sugary foods may vary from 'no increase' to 'decrease'.
2. Drink plenty of fluids each day.
3. Make sure you get enough calcium or milk.
4. Eat foods containing iron.
5. Do not eat too much protein.
6. Limit caffeine intake.
7. Drink fluoridated water.
8. Preserve the nutritive value of food (by good food preparation).
9. Eat three good meals a day.

Doctors are frequently asked to give nutritional advice but may lack the necessary expertise. Dietitians, nutritionists, and appropriately trained practice nurses, play an invaluable role in providing the public with practical advice to facilitate changes from the typical Western diet as well as providing instruction regarding therapeutic diets for those with diseases requiring specific diet therapy. The enormous potential for dietary change to reduce the effects of a wide range of diseases should encourage physicians to approach the nutritional management of their patients with enthusiasm.

Further reading

COMA–CHD Panel on Dietary Reference Values of the Committee on Medical Aspects of Food Policy (1991). *Report on Health and Social Subjects: 41. Dietary Reference Values for Food Energy and Nutrients for the United Kingdom*. HMSO, London.

de Deckere EAM, Korver O, Verschuren PM, Katan MB (1998). Health aspects of fish and n-3 polyunsaturated fatty acids from plant and marine origin. *European Journal Clinical Nutrition* **52**, 749–53.

Department of Health (1994). *Report on Health and Social Subjects: 46. Nutritional aspects of cardiovascular disease. Report of the Cardiovascular Review Group Committee on Medical Aspects of Food Policy*. HMSO, London.

Mann J, Truswell AS, eds. (2002). *Essentials of Human Nutrition*, 2nd edn. Oxford University Press.

Truswell AS (1999). *ABC of Nutrition*, 3rd edn. BMJ Books, London.

World Heath Organization/Food and Agriculture Organization (1998). *Carbohydrates in Human Nutrition*. FAO Food and Nutrition Paper 66. Report of a Joint FAO/WHO Expert Consultation. Rome.

10.2 Nutrition: biochemical background

Keith N. Frayn

Introduction: flux of 'energy substrates'

Food intake is sporadic: for most people it occurs in three major boluses each day. Energy expenditure, however, is continuous, with variations during the day that bear no resemblance to the pattern of energy intake, except that over some reasonable period of time (a week or more) the two will, in most people, match almost exactly. Therefore the body has developed complex systems that direct nutrients into storage pools when they are in excess, and that regulate the 'mobilization' of nutrients from these pools as they are needed. The situation is analogous to the petrol (gas) tank of a motor car and the throttle that regulates fuel oxidation, except that in the motor car there is just one fuel and just one engine: in humans there are three major nutrients and a variety of tissues and organs, each of which may have its own preferences for fuels, that vary with time. The fact that we can carry on our daily lives without thinking about whether to store or mobilize fuels, and which to use, attests to the remarkable efficiency of these control systems.

The body's principal macronutrient stores are listed in Table 1 and are related to daily fluxes in the body.

The regulation of macronutrient flux

The need for the co-ordinated control of nutrient storage and mobilization, to flux between tissues and along the many metabolic pathways, is met by a complex series of control mechanisms. These may be viewed on several levels.

The simplest involves the effects of substrate concentration, and is dependent upon the kinetic properties of enzymes and transport proteins. An illustration is the uptake of glucose by the liver. The facilitated carrier responsible for entry of glucose into the hepatocyte, GLUT2, has a high Michaelis constant, K_m (around 10–20 mmol/l), for glucose and a high capacity. The first enzyme of glucose metabolism in the hepatocyte, hexokinase IV (glucokinase), has similar properties. Therefore, flux of glucose into the hepatocyte is dependent upon the extracellular to intracellular concentration gradient. When glucose is absorbed from the small intestine and transported to the liver in the portal vein the hepatocyte is subject to an external glucose concentration of perhaps 10 mmol/l, and acts as a 'sink' to remove glucose and buffer fluctuations in delivery of glucose to the systemic circulation. In contrast, the glucose transporter in the brain, GLUT3, has a low K_m (around 1–3 mmol/l), and glucose utilization is therefore almost constant while the plasma glucose concentration varies within the normal physiological range. Within tissues the same may be true. The aminotransferases (transaminases) that catalyse the removal of amino groups from amino acids, the first step in their catabolism, have a high K_m and high capacity, and again flux is driven by concentration. Hans Krebs showed that this was one explanation for the old observation that an increase in protein intake is followed by rapid oxidation of the excess amino acids.

The next level involves more specific interaction of nutrients, or pathway intermediates, with enzymes, usually through allosteric effects (binding of the effector alters the conformation of the enzyme and hence its catalytic properties). There are many examples in the metabolism of carbohydrate, fat, and protein. The enzyme 6-phosphofructo-1-kinase (which converts fructose-6-phosphate to fructose-1,6-bisphosphate in the pathway of glycolysis) is a good example, subject to allosteric regulation by many compounds that relate to the 'energy status' of the cell. For instance, it is activated by AMP (indicating energy shortage) and inhibited by ATP. Such mechanisms undoubtedly provide important 'fine tuning' of flux along various pathways entirely in accord with the modern view that regulation of flux does not reside in certain 'rate-limiting steps' but is distributed amongst many steps along a pathway.

These mechanisms operate essentially within one tissue. However, the co-ordination of nutrient metabolism requires considerable interaction between tissues and organs. This co-ordination is largely brought about by the hormonal and nervous systems. Certain hormones play a particularly important role in regulation of macronutrient flux (Table 2). The role of the nervous system in metabolic regulation is often difficult to assess. For instance, whilst the effects of adrenaline are properly regarded as hormonal, liberation of noradrenaline from sympathetic nerve endings in tissues may bring about identical effects and can be difficult to distinguish. The somatic nervous system (motor neurones innervating skeletal muscle) has clear effects, for example stimulation of breakdown of muscle glycogen linked to muscle contraction. The autonomic nervous system probably plays multiple roles, but some are indirect: for example regulation of blood flow and cardiac output, thus affecting delivery of substrate to tissues, and regulation of the secretion of pancreatic hormones.

The effects of hormones are mediated in many ways, but may be divided into acute effects (usually acting within seconds or minutes), often brought about through reversible (de)phosphorylation of enzymes, and longer-term effects (hours or days), brought about by regulation of gene expression. The former are usually exerted through binding to cell surface receptors linked to a variety of second messenger systems, the latter through nuclear receptors (for example for glucocorticoids and thyroid hormones). However, the distinction is not absolute: insulin, for example, brings about both acute and longer-term effects through binding to the same cell surface receptor.

A further level of co-ordination is through the effects of nutrients themselves, or important cellular components such as cholesterol, upon gene expression. This can be seen as a longer-term mechanism to ensure that metabolism is appropriate to the diet being ingested and the lifestyle followed. A variety of nutrient response elements are being discovered in the promoter regions of genes for enzymes concerned with substrate metabolism. Particular examples are the carbohydrate response element (upregulating expression of genes for glucose metabolism such as pyruvate kinase in the glycolysis pathway), the sterol response element (cellular sterols downregulate expression of the low-density lipoprotein receptor and the enzymes of cholesterol biosynthesis) and response elements for fatty acid derivatives. Fatty acids affect gene expression through a family of transcription factors known as the peroxisome proliferator activated receptors which act in turn upon the peroxisome proliferator response elements. These were discovered through the effects of a diverse group of compounds that promote proliferation of peroxisomes in the rodent liver. Now they are recognized as factors that mediate the response to dietary fatty acids. They are summarized in Table 3.

Carbohydrate metabolism in the postabsorptive and postprandial states

The postabsorptive state

In the overnight-fasted (postabsorptive) state, no glucose enters the plasma from the small intestine. Glucose in the plasma turns over at about 2 mg/kg body weight/min (200 g/24 h). About one-half of this will be consumed by the brain. Of the remainder, a considerable proportion will be utilized by blood cells and peripheral tissues by anaerobic glycolysis, thus returning lactate to the liver for reconversion to glucose (Fig. 1). This is the Cori cycle.

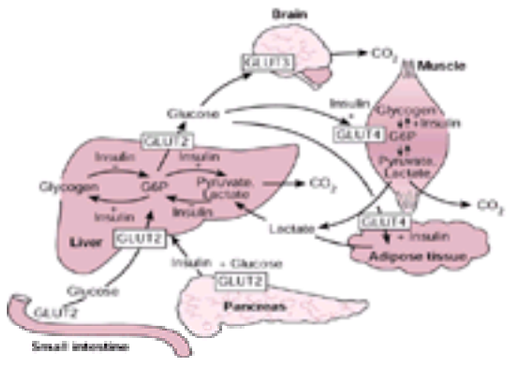

Fig. 1 Overview of carbohydrate metabolism. Pathways in the liver shown as regulated by insulin are probably controlled by the insulin/glucagon ratio (high in the fed state, low in fasting). In muscle, contraction is an important stimulus for glycogen breakdown and glycolysis. Adrenaline also contributes. Not shown is the significant glucose uptake by red blood cells and other glycolytic tissues, returning lactate to the liver. GLUT2 is the high- K_m non-insulin regulated glucose transporter (i.e. the glucose flux is determined by concentration), GLUT3 is the low- K_m brain glucose transporter (the glucose flux is relatively independent of concentration within the normal range), and GLUT4 the insulin-regulated glucose transporter. G6P is glucose 6-phosphate.

Glucose is produced by hepatocytes from glycogen breakdown and from gluconeogenesis. Net glycogen breakdown is stimulated by the relatively low insulin/glucagon ratio after overnight fasting. The major substrates for gluconeogenesis are lactate and pyruvate, released from blood cells and peripheral tissues, together with alanine and glycerol. The pathway of gluconeogenesis predominates over that of glycolysis, again because of the relatively low insulin/glucagon ratio.

Glucose metabolism following a meal

When a meal enters the system, this pattern of metabolism changes rapidly. About 12 g of free glucose are present in the circulation and extravascular space. Typically, a single meal will provide about 100 g of glucose, entering the circulation over perhaps 60 min. In order to minimize variations in plasma glucose concentration, co-ordinated mechanisms come into play to suppress the production of endogenous glucose and to increase the rate of removal of glucose from the circulation.

Much of the incoming glucose may be taken up by hepatocytes as described earlier, but some enters the systemic circulation, from where it can stimulate pancreatic insulin secretion (and somewhat suppress glucagon secretion, although this is not so marked an effect). Insulin is liberated into the portal vein. Thus, the liver is exposed at this time to high concentrations of glucose (from the small intestine) and insulin. The net effect is to reverse net glycogenolysis, so that net glycogen synthesis begins. In addition gluconeogenesis is suppressed and glycolysis favoured (Fig. 1). Hepatocyte glucose output is therefore rapidly suppressed and converted to a net uptake of glucose. At the same time, utilization of glucose by insulin-sensitive peripheral tissues such as skeletal muscle and adipose tissue is increased. The main mechanism of this short-term change is the recruitment of the insulin-regulated glucose transporter GLUT4 to the cell membrane from an intracellular pool. However, the reduction in concentration of plasma non-esterified fatty acids (see below) will also remove inhibition of glucose uptake caused by fatty acid oxidation. Within muscle, glycolysis and glycogen synthesis will be stimulated by insulin. In adipose tissue, increased glucose uptake provides glycerol-3-phosphate (formed from glycolysis) for esterification of fatty acids (see below). Thus, insulin is the key regulator of the rapid changes that occur in glucose metabolism in the postprandial state: it brings about glucose storage as glycogen, and promotes the utilization of glucose at the expense of fatty acids.

Fat metabolism in the postabsorptive and postprandial states

Forms of fat in the circulation (see Chapter 11.6)

Fatty acids circulate in various forms: as non-esterified fatty acids, as triacylglycerol (triglyceride) fatty acids, esterified to glycerol in phospholipids, and esterified to cholesterol as cholesteryl esters. The first two are involved in 'energy metabolism'. The main carriers of triacylglycerol in the circulation are the triacylglycerol-rich lipoproproteins, chylomicrons secreted from the small intestine and transporting dietary fat, and very low-density lipoprotein particles secreted from the liver, transporting endogenous triacylglycerol. In the postabsorptive state, chylomicron triacylglycerol secretion will be virtually zero. Secretion of very low-density lipoprotein is a means of exporting fat from the liver to peripheral tissues. In these tissues it is hydrolysed by the enzyme lipoprotein lipase situated in the capillaries of skeletal muscle, adipose tissue, mammary glands, and other tissues that use fatty acids. Lipoprotein lipase acts on the circulating triacylglycerol-rich particles to liberate fatty acids which may diffuse into the parenchymal cells (muscle fibres, adipocytes, etc.). Lipoprotein lipase in skeletal muscle is downregulated by insulin, whereas that in adipose tissue is upregulated by insulin. In the postabsorptive state muscle lipoprotein lipase is likely to predominate as the site of removal of triacylglycerol from the very low-density lipoprotein particles. The fatty acids can then be used as an oxidative fuel by the muscle. In this process very low-density lipoprotein particles lose their triacylglycerol core and become relatively enriched with cholesterol and phospholipids. After several cycles through such capillary beds, they are reduced to simple particles with a core of cholesteryl ester and an outer phospholipid shell: in fact, low-density lipoprotein particles, the main carrier of cholesterol in the circulation.

Non-esterified fatty acids and 'energy transport'

Fat is mobilized from adipose tissue stores in the form of non-esterified fatty acids (Fig. 2). The adipocyte has a central droplet of triacylglycerol, which is hydrolysed by the intracellular enzyme hormone-sensitive lipase, releasing glycerol and non-esterified fatty acids. These fatty acids are liberated into the plasma bound to albumin for transport to other tissues, including liver and skeletal muscle. Hormone-sensitive lipase is stimulated by catecholamines but powerfully suppressed by insulin, each exerting control over reversible (de)phosphorylation (see Fig. 2) (insulin leads to dephosphorylation and deactivation). Thus, it is active in the postabsorptive state when there is a call upon the body's fat stores. It is also activated during exercise, mainly by catecholamine stimulation. The turnover of non-esterified fatty acids in the plasma is rapid. They are the major oxidative fuel in muscle after overnight fast (glucose only supplies around 5 to 10 per cent of the oxidative fuel for skeletal muscle in this state), and in the liver they are both a fuel for oxidation and a substrate for synthesis of triacylglycerol that will be exported as very low-density lipoprotein. A typical concentration of non-esterified fatty acids in the plasma after overnight fast is 500 µmol/l, one-tenth that of glucose, but because of their rapid turnover and their larger molecular mass fatty acids account for about twice the 'energy turnover' of glucose in the circulation.

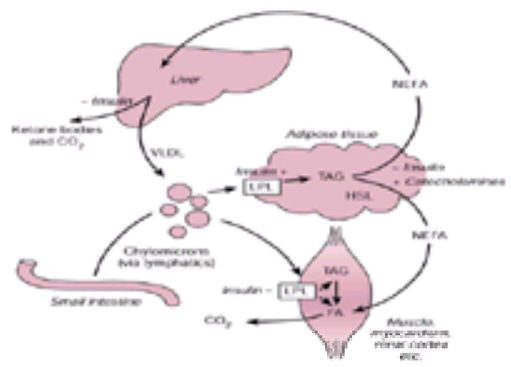

Fig. 2 Overview of fat metabolism. Triacylglycerol (TAG) enters the circulation from the diet in the form of chylomicrons. Fatty acids are taken up by tissues through the action of the enzyme lipoprotein lipase (LPL). Adipose tissue TAG is the major store. It is mobilized in times of energy demand by the enzyme hormone-sensitive lipase (HSL), liberating non-esterified fatty acids (NEFA) into the circulation, from where they may be taken up by a number of tissues and used for synthesis of new TAG and for oxidation. Major points of hormonal regulation are shown (italic).

Disposition of dietary fat

Dietary fat is almost entirely (typically 95 per cent or more) in the form of triacylglycerol. A typical meal might contain 30 to 40 g of fat. The typical plasma triacylglycerol concentration in a healthy subject is 1 mmol/l, confined to the vascular space; this means that about 3 g of triacylglycerol is present in the circulation. Therefore, as in the case of glucose, the amount in a meal could overwhelm the system unless co-ordinated mechanisms come into play to ensure its rapid dispersion.

Dietary triacylglycerol is digested in the stomach and small intestine and packaged by the enterocytes of the duodenum and proximal ileum into chylomicrons, which enter the circulation via the lymphatics (Fig. 2). Therefore, unlike other nutrients absorbed from the small intestine, they bypass the liver on first passage. The chylomicrons also carry other lipid constituents of food, including cholesterol and fat-soluble vitamins. In the circulation their fate is similar to that of very low-density lipoprotein particles, although the tissue-specific regulation of lipoprotein lipase ensures that adipose tissue (where lipoprotein lipase is upregulated by insulin) is a major site of clearance of their triacylglycerol. The pathway of triacylglycerol synthesis in adipocytes, as in the liver, is stimulated by insulin. Therefore, there is a short and energy efficient pathway for storage of dietary fatty acids in adipose tissue (Fig. 2). The half-life of chylomicron triacylglycerol in the circulation is about 5 min. After hydrolysis of most of the triacylglycerol, the remnant particles are removed by receptors in the liver and other tissues. Thus dietary cholesterol which remains in the remnant particles along with most fat-soluble vitamins is transported mainly to the liver.

Provided that a meal contains carbohydrate or protein, stimulation of insulin secretion will rapidly suppress the mobilization of adipose tissue fat stores, and concentrations of non-esterified fatty acids in the plasma will fall after a meal. Therefore utilization of fatty acids by tissues such as skeletal muscle and liver will be reduced simply by lack of availability. As noted above, this reduces competition for oxidation in muscle, further increasing glucose utilization. In liver, the lack of non-esterified fatty acids is likely to decrease the secretion of very low-density lipoprotein triacylglycerol. Insulin appears also to suppress very low-density lipoprotein triacylglycerol secretion directly. This is somewhat controversial, and the effects of insulin may be different in the acute, postprandial situation from the situation of prolonged hyperinsulinaemia (as in insulin resistance). Within the liver, insulin powerfully stimulates esterification of fatty acids (for triacylglycerol synthesis) at the expense of oxidation of fatty acids (see below), so the suppressive effect of insulin on very low-density lipoprotein triacylglycerol secretion can only be short-term. Nevertheless, it seems an exact parallel with the suppression of hepatic glucose output by insulin.

Inter-relationships between carbohydrate and fat metabolism

Links between carbohydrate and fat

In mammals, fat cannot be converted to glucose in a net sense (for each molecule of acetyl CoA entering the tricarboxylic acid cycle, two molecules of CO_2 are produced). Glucose can, however, be converted to fat: acetyl CoA produced by pyruvate dehydrogenase leaves the mitochondrion (being transported across the mitochondrial membrane as citrate, then 'liberated' by the cytosolic enzyme ATP-citrate lyase), and is then a substrate for the pathway of *de novo* lipogenesis, which begins with the enzyme acetyl CoA carboxylase (forming malonyl CoA). At one time it was believed that *de novo* lipogenesis was a major route for laying down storage fat. Although this may be true in small rodents under some conditions, many measurements made in recent years have confirmed that this pathway makes a quantitatively small contribution to triacylglycerol synthesis in humans. Instead, it seems that almost all the triacylglycerol that we deposit in adipose tissue arises from dietary fatty acids, taken up from circulating triacylglycerol-rich lipoproteins by the lipoprotein lipase pathway (see Chapter 4.3).

The lack of quantitatively significant interconversion of carbohydrate and fat has led to the suggestion that we may view carbohydrate and fat balance as independent. This view is entirely erroneous. Despite the lack of interconversion, carbohydrate balance strongly influences fat balance, and vice versa. These influences occur at a number of levels.

At a chronic level, ingestion of a high-carbohydrate diet will induce enzymes of fat synthesis and downregulate enzymes of fatty acid oxidation, through insulin- and carbohydrate-response elements in the promoter regions of the genes in question. On a whole-body but more acute level there are clear hormonal effects. Principal amongst these is carbohydrate-induced insulin secretion. Insulin, as outlined above (see Chapter 4.3), powerfully suppresses the release of non-esterified fatty acids from adipose tissue. Therefore, when carbohydrate is readily available, fat stores are conserved.

The glucose–fatty acid cycle

Beyond this, there are specific cellular mechanisms that regulate the relative oxidation of carbohydrate and fat. These probably operate in a number of tissues although they have been most studied in skeletal and heart muscle and in liver. In 1963 Philip Randle and colleagues described the glucose–fatty acid cycle, which encompasses one aspect of this mutual relationship between carbohydrate and fat oxidation. It is summarized in Fig. 3. The concept was based upon observations that availability of fatty acids reduced the oxidation of glucose in skeletal and cardiac muscle. The mechanism involves generation of acetyl CoA from b-oxidation of fatty acids. A high ratio of acetyl CoA to free CoA within the mitochondrion inhibits pyruvate dehydrogenase. In addition, export of citrate to the cytoplasm would inhibit phosphofructokinase, and consequent accumulation of glucose-6-phosphate would inhibit uptake of hexokinase and glucose. The precise mechanism has since been disputed, but the basic observation has been confirmed many times. The glucose–fatty acid cycle describes the normal interplay between fat and carbohydrate oxidation, but has also been invoked (following a suggestion made in the original paper in 1963) to explain pathological situations involving excess availability of fat and insulin resistance of glucose uptake (for example type 2 diabetes and obesity).

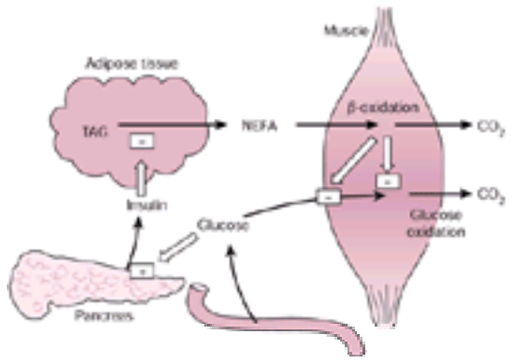

Fig. 3 The glucose–fatty acid cycle. When glucose and insulin concentrations are high, release of non-esterified fatty acid (NEFA) from adipose tissue is suppressed, and glucose utlization predominates in insulin-sensitive tissues such as skeletal muscle. In the fasting state (glucose and insulin concentrations are low) NEFA utilization predominates, reinforced by inhibitory effects of the products of b-oxidation of fatty acids on glucose uptake and oxidation. This may have pathological significance in that states in which NEFA concentrations tend to be high (e.g. type 2 diabetes) will be associated with resistance of glucose utilization to insulin.

Glucose and the regulation of fatty acid oxidation

An additional mechanism was first described in 1977 by Denis McGarry and Daniel Foster. They were following up a longstanding observation that the generation of ketone bodies by the liver (a reflection of b-oxidation of fatty acids) was suppressed by insulin or in livers from carbohydrate-replete animals. They showed that malonyl CoA, the first committed intermediate in the pathway of *de novo* lipogenesis (produced by acetyl CoA carboxylase; see above), strongly inhibited fatty acid oxidation. This inhibition is mediated via the enzyme carnitine palmitoyltransferase-1 (also called carnitine acyltransferase-1) in the mitochondrial membrane. Carnitine palmitoyltransferase-1 is responsible for the transport of fatty acids from the cytoplasm to the mitochondrion for b-oxidation. Acetyl CoA carboxylase is activated by insulin (both by increased gene transcription and by reversible dephosphorylation). Hence in a carbohydrate-replete state malonyl CoA will be formed and fatty acid oxidation inhibited (Fig. 4).

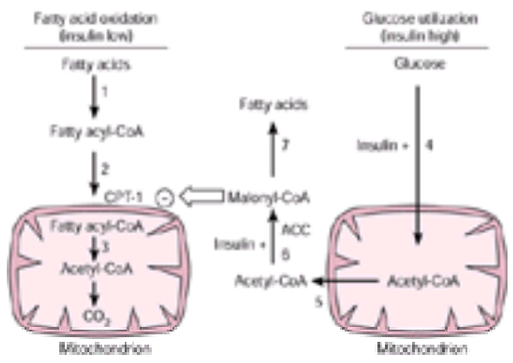

Fig. 4 The inhibition of fatty acid oxidation when glucose and insulin concentrations are high. Pathways are: 1, fatty acid 'activation' (fatty acyl CoA synthase); 2, transfer of acyl CoA into the mitochondrion via carnitine palmitoyltransferase-1 (CPT-1); 3, b-oxidation; 4, glycolysis (cytosolic) followed by pyruvate dehydrogenase (mitochondrial); 5, transfer of acetyl CoA to the cytosol via the citrate shuttle (details not shown); 6, acetyl CoA carboxylase (ACC) to form malonyl CoA, which is a powerful inhibitor of CPT-1; 7, fatty acid synthesis (not present in all tissues in which this mechanism operates, for example skeletal muscle).

This is now recognized as a widespread regulatory mechanism. There are two isoforms of acetyl CoA carboxylase. Acetyl CoA carboxylase 1, expressed in lipogenic tissues such as liver and adipose tissue, appears to be involved in *de novo* fatty acid synthesis. Acetyl CoA carboxylase 2 is expressed more in tissues oxidizing fatty acids such as heart and skeletal muscle and is thought to produce malonyl CoA for regulatory rather than synthetic purposes. In fact, in muscle the pathway of fatty acid synthesis does not occur (fatty acid synthase is not expressed). Muscle carnitine palmitoyltransferase-1 is more sensitive to inhibition by malonyl CoA than is the liver enzyme. The ability of glucose to inhibit the oxidation of fatty acids in muscle has recently been clearly demonstrated *in vivo*, and has been termed the 'reverse glucose–fatty acid cycle'.

Protein and amino acid metabolism and their regulation

Since there are 20 different amino acids incorporated into protein, and a variety of other amino acids that have important biological rules, it is essential for this purpose to generalize somewhat about amino acid and protein metabolism.

The body pools of protein and amino acids, and their turnover, are summarized in Fig. 5. Insulin exerts a net anabolic role on body protein, mainly in skeletal muscle, whereas thyroid hormones and cortisol are generally 'catabolic'. Anabolism is also stimulated by anabolic steroids, by physical training, and during growth by the insulin-like growth factors.

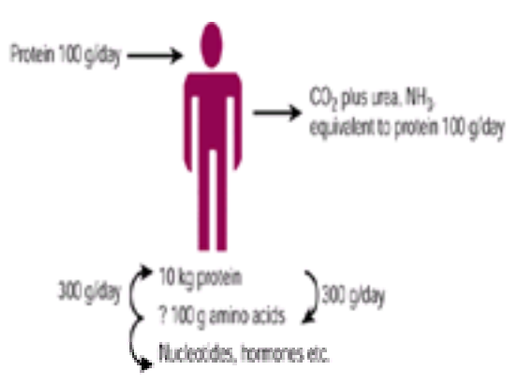

Fig. 5 The body pools of protein and amino acids, and their turnover. Figures are approximate.

Dietary protein, digested in the small intestine and absorbed as free amino acids and short peptides, enters the portal vein. In the enterocytes of the small intestine some amino acids, especially glutamine, are removed for use as an oxidative fuel. The remaining products of digestion next meet the liver where further preferential extraction takes place. Amino acid oxidation is, under most circumstances, the major oxidative pathway in the liver. About 60 per cent of incoming amino acids may be directed into immediate oxidation. The rate of hepatic protein synthesis is also high, and since much of the protein is secreted (for example albumin), this represents a net loss of amino acids from the liver (perhaps a further 20 per cent of the incoming amino acids). The remaining mixture of amino acids, around 20 per cent of those absorbed, enters the systemic circulation. This mixture is enriched in the branched chain amino acids leucine, isoleucine, and valine, which have a special role in muscle.

Urea synthesis takes place only in the liver. (The pathway is present in the brain but this is not a significant site of blood urea production.) Therefore, amino acids released from proteolysis in peripheral tissues must be transported, or must transfer their amino nitrogen, to the liver. This results in considerable interaction between the pathways of amino acid, carbohydrate, and fat metabolism. Measurements of arteriovenous differences across muscle and adipose tissue show that the release of the two amino acids alanine and glutamine predominates. Since glutamine carries two nitrogens it is, under most circumstances, the predominant carrier of nitrogen. Arteriovenous difference measurements across the splanchnic bed (by catheterization of the hepatic vein) show an almost identical pattern for uptake: removal of alanine and glutamine far exceeds that of other amino acids. Therefore amino acids in tissues including muscle and adipose tissue must transfer their amino nitrogen to alanine (by transamination with pyruvate) and glutamine (formed from glutamate, itself arising by transamination with 2-oxoglutarate). The aminotransferases (transaminases) bring about this transfer. It is important that the 2-oxoacid acceptors, pyruvate and 2-oxoglutarate, are common metabolic intermediates and thus readily available.

Much of the alanine released from skeletal muscle comes from transamination of pyruvate formed in glycolysis. Within the liver the amino group can be transferred further, for example to oxaloacetate, forming aspartate which is one of the immediate donors of nitrogen to the urea cycle. The pyruvate thus formed may be a substrate for gluconeogenesis, producing glucose that can be recycled to peripheral tissues. This metabolic cycle has been called the glucose–alanine cycle. It closely parallels the Cori cycle (see above).

The other route of entry of nitrogen into the urea cycle is ammonia. In peripheral tissues ammonia may be formed by the oxidative deamination of glutamate, catalysed by glutamate dehydrogenase. This reaction, in combination with the aminotransferases, can be seen to capture amino nitrogen from a number of amino acids. However, blood ammonia concentrations are very low (it is highly toxic) and instead it seems to be 'fixed' in the amido group of glutamine by the enzyme glutamine synthetase. In the liver the ammonia required for the urea cycle may be formed from the amido nitrogen of glutamine, removed by the enzyme glutaminase, or by the oxidative deamination of glutamate. There is also a supply of ammonia from the small intestine.

An important aspect of the large store of muscle protein is that it represents a potential source of synthesis of new glucose during fasting. In that situation, while the brain continues to require glucose for oxidation, and as glycogen reserves are depleted, new glucose can only be formed from glycerol, released in adipose tissue lipolysis, and from amino acids. The pathways described above are for the transfer of nitrogen, but not necessarily of carbon, to the liver. To explain the latter, we have to invoke pathways whereby amino acid carbon can also be exported. Amino acids whose 2-oxoacid can enter the tricarboxylic acid cycle may generate pyruvate (which can also accept amino nitrogen to become alanine). Pairs of amino acids can provide all the carbons necessary for glutamine synthesis.

Further reading

Frayn KN (1996). *Metabolic regulation: a human perspective*. Portland Press, London.

Friedman JM, Halaas JL (1998). Leptin and the regulation of body weight in mammals. *Nature* **395**, 763–70.

Havel RJ (1997). Postprandial lipid metabolism: an overview. *Proceedings of the Nutrition Society* **56**, 659–66.

McGarry JD (1979). Lilly Lecture 1978. New perspectives in the regulation of ketogenesis. *Diabetes* **28**, 517–23.

Nonogaki K (2000). New insights into sympathetic regulation of glucose and fat metabolism. *Diabetologia* **43**, 533–49.

Nordlie RC, Foster JD, Lange AJ (1999). Regulation of glucose production by the liver. *Annual Review of Nutrition* **19**, 379–406.

O'Brien RM, Granner DK (1996). Regulation of gene expression by insulin. *Physiological Reviews* **76**, 1109–61.

Randle PJ *et al*. (1963). The glucose-fatty acid cycle. Its role in insulin sensitivity and the metabolic disturbances of diabetes mellitus. *The Lancet* **1**, 785–9.

Sidossis LS, Wolfe RR (1996). Glucose and insulin-induced inhibition of fatty acid oxidation: the glucose-fatty acid cycle reversed. *American Journal of Physiology* **33**, E733–E738.

Zammit VA (1999). The malonyl-CoA-long-chain acyl-CoA axis in the maintenance of mammalian cell function. *Biochemical Journal* **343**, 505–15.

Zierler K (1999). Whole body glucose metabolism. *American Journal of Physiology* **276**, E409–E426.

10.3 Vitamins and trace elements

M. Eastwood

Vitamins are diverse organic compounds that the body cannot synthesize; they are required in small amounts to contribute to health.

Historical background

Our understanding of the role of vitamins comes from clinical observations, nutritional experiments in animals, and studies using purified preparations of the active principle used to treat deficiency states. Early pioneers differentiated dietary deficiency from infection and other causes of disease. Scurvy was once the scourge of mariners and explorers, but the clinical trials of Lind (1753), confirmed by Captain Cook on his voyages, showed the benefits of citrus fruits. Many years later, Holst and Froelich (1907) produced scurvy in guinea-pigs by dietary deprivation.

Rickets arose in sun-starved urban slums, and Trousseau noted the beneficial effects of cod liver oil (1860). Ejikmann and Grijns fed chickens the same diet as their patients who had beriberi (1897); neuropathy in the chickens resolved when the diet contained whole-grain, rather than polished, rice. In the 1900s Gowland Hopkins, the discoverer of vitamins, described a fat-soluble, essential growth accessory food factor A in milk. This was differentiated from water-soluble accessory food factor B by McCollum and Davis. Mellanby treated rickets in puppies with a fat-soluble food factor D. Lucy Wills described the megaloblastic anaemia of pregnancy in 1931, which is now known to be caused by a deficiency of folic acid.

In 1894 Atwater published a table of food composition and dietary standards for the United States. The first *USA Recommended Daily Allowances* was published in 1941. Food rationing in Britain during the Second World War was a triumph for the science of applied nutrition. The natural development of this work was the emergence of recommended daily intakes that recognized the differing requirements of the young and growing, pregnant and lactating, middle-aged and old, and ill; with this, developed the concept of the optimal intake for optimal nutritional status. The isolation and chemical synthesis of the vitamins and their active principles provided formidable challenges to scientists, rewarded by eight Nobel prizes in Medicine and Physiology and four in Chemistry. In 1976, Linus Pauling advocated gram intakes of ascorbic acid as a prevention against the common cold. Pauling thus founded 'orthomolecular medicine'.

The traditional classification of vitamins into water and lipid soluble and by their associated deficiency conditions becomes less useful as their biochemical roles are better understood. An inadequate dietary vitamin intake will result in specific cellular failure and even death. There is a dose–response relationship with vitamin intake from the physiological through the pharmacological to the toxic. Recommendations for vitamin intake for different ages, needs, and communities are based on dietary intake, bioavailability, steady-state concentrations in plasma and tissue at defined intakes, urine excretion, adverse effects, biochemical and molecular function, and freedom from deficiency. With increasing intake, either orally or as an infusion, a vitamin is distributed through the body fluids and tissues until the saturation point is exceeded. The prescription of vitamins parenterally, bypassing the absorptive processes, also has dosage implications.

The carotenoids illustrate the complexity of vitamins in physiology, pharmacology, and as toxins. Carotenoids are used by archaebacteria to reinforce cell membranes, their long, rigid carbon backbone acting as a rivet across the membrane. The polyene chain of between 9 and 11 double bonds serves to harvest light energy in plants, and, as the pigment retinal, is a visual pigment in animals. The linear system of conjugated C=C bonds make for a high reducing and antioxidant potential. Carotenoids act as the coloration in plants and to protect egg proteins against the enzymatic activity of proteases. Retinoic acid in animals and abscisic acid in plants act as hormones. When retinoids and carotenoids were used in lung cancer chemoprevention trials, the incidence of cancer increased, ascribed possibly to the increase in the oxidized products of b-carotene. A mix of vitamins and antioxidants might prevent such oxidation. Such a mix can be found in fruit and vegetables, which emphasizes the benefit of a good diet containing five portions (60 to 150 g) of fruit and vegetables a day.

Vitamin C (ascorbic acid)

Ascorbic acid is a simple sugar that reversibly oxidizes to dehydroascorbate. Dietary sources are fresh fruit and fruit juices, especially blackcurrants, guavas, green leafy vegetables, and fresh milk. Ascorbic acid is readily oxidized during cooking—a process accelerated by traces of copper in alkaline solution. Since humans, guinea-pigs, the Indian fruit-eating bat, the red vented bulbul, and some birds are unable to synthesize ascorbic acid, it thus represents a vitamin in these species. Ascorbic acid

- is a water-soluble, non-specific radical-trapping antioxidant and reducing agent which is present in all tissues;
- acts synergistically with vitamin E;
- is involved in copper-containing hydroxylase enzymes (e.g. proline and lysine hydroxylase), and is important in collagen metabolism and 2-oxoglutarate-linked iron hydroxylases;
- is involved in carnitine biosynthesis, which is necessary for fatty acid transport from the cytosol into mitochondria; and
- is involved in the synthesis of hormones (e.g. epinephrine (adrenaline)).

Ascorbic acid is rapidly absorbed from the small intestine. The plasma concentration (5 per cent as dehydroascorbate) and dietary intake are in a sigmoidal relationship (80 µmol/l on 100 mg/day) which plateaus at 1000 mg/day intake. The body pool size is 900 mg (5 mmole in the normal adult): approximately 3 per cent, irrespective of pool size, is degraded each day and excreted in the urine as free ascorbic acid, dehydroascorbate, or diketogulonate. High tissue concentrations at birth steadily decline with increasing age.

A shortfall in vitamin C intake without clinical scurvy may be associated with a reduction in the body's water-soluble antioxidant capacity, the consequences of which are still being debated. The young and elderly are particularly at risk of scurvy, which results from an inadequate intake of ascorbic acid. Clinical scurvy appears after

4 weeks on an ascorbic acid-deficient diet when the body pool is less than 300 mg (1.7 mmole). There is a failure of connective tissue collagen synthesis, and cartilage, bone, and dentine growth are all compromised. Tissues bleed readily, and heal poorly due to defective intracellular linkages between the endothelial cells and capillary basement tissue. Individuals are initially lethargic and irritable. Characteristic livid-coloured, spongy, bleeding gingivitis of the gums develops, and scurvy buds appear in the papillae between the teeth. Large or microscopic haemorrhages occur in the gums, as well as in the eyes (especially bulbar conjunctiva), subcutaneous tissues, synovia of joints, and beneath the periosteum of bones. Perifollicular bleeding occurs in the dependent parts of the body, later becoming more generalized. Fatal haemorrhages may also occur in the brain or heart muscle. Keratin-like material heaps on the surface of hair follicles, through which a deformed corkscrew hair projects. Other signs include dependent oedema, oliguria, depression, megaloblastic or normoblastic anaemia, and superinfection. Infants present with irritability, tender legs, and pseudoparalysis. Scurvy buds, but not gingivitis, occur in edentulous infants. Large subperiosteal haemorrhages develop over the long bones, especially the femur.

Diagnosis requires an awareness of the condition, a careful dietary history, and a clinical examination. The patient can improve on hospital diet alone. Ascorbic acid is measured in plasma or whole blood. While leucocyte or buffy-coat vitamin C concentrations reflect tissue concentrations, this is complicated in disease by different leucocytes types which vary in number and ascorbic acid content.

Requirements

The optimum dietary intake of ascorbic acid has yet to be defined. The body can be saturated with 1 g/day for 5 days. Recommendations for dietary intake range from 40 to 200 mg per day. An upper limit of intake is recommended at 1 to 2 g/day, based upon body saturation figures rather than toxicity. During pregnancy and lactation, intake should be between 100 and 200 mg daily. The ascorbic acid content of breast milk varies between 30 and 80 mg/l which provides 25 mg per day, clinical scurvy has not been observed in fully breast-fed infants. The ascorbic acid intake of the elderly is generally adequate but should be monitored. The pharmacological use of ascorbic acid is extensive and imaginative.

Biotin

Biotin contains a ureido group in a five-membered ring fused with a tetrahydrothiophene ring with a five-carbon side chain terminating in a carboxyl group. Dietary biotin is found in yeast, bacteria, liver, kidney, egg yolks, cooked cereals, pulses, nuts, chocolates, and some vegetables. Biotin is a cofactor for the acetyl-coenzyme A (**CoA**), propionyl CoA, and pyruvate carboxylase systems involved in the incorporation of bicarbonate as a carboxyl group into substrates in fatty acid synthesis and gluconeogenesis. Biotin is absorbed from the upper gastrointestinal tract. Raw egg white contains the glycoprotein avidin (molecular weight, 68 000) which binds biotin with a high affinity and prevents biotin absorption. Biotin is transported in plasma to the liver for storage. It is metabolized before excretion in the urine as biotin, bisnorbiotin, and biotin sulphoxide.

Biotin deficiency results in fatigue, depression, sleepiness, nausea, loss of appetite, muscle pain, hyperaesthesia and paraesthesia, alopecia, dermatitis, conjunctivitis, smooth tongue, and dry skin. Individuals receiving treatment for epilepsy are at risk of biotin deficiency. There are no indications that excess biotin is toxic. Body stores are measured by plasma biotin concentrations, and lymphocyte propionyl CoA carboxylase and its activation index (ratio of enzyme activity incubated with and without biotin) or urinary 3-hydroxy isovalerate.

Requirements

The dietary requirement of biotin is not known with certainty. The average intake of a British adult ranges from 10 to 70 µg/day, which is seen to be adequate. In infants, preterm to 5 years, an intake between 5 and 25 µg/day is suggested.

Niacin: nicotinic acid and nicotinamide

Niacin is a B vitamin. It occurs in food as nicotinic acid, as a pyridine nucleotide coenzyme derivative (NAD and NADP), as an amide, nicotinamide (niacinamide), or as a nicotinoyl ester, niacytin in maize. Nicotinic acid is found in meat, poultry, fish, wholemeal cereals, pulses, and coffee. Nicotinic acid is synthesized from tryptophan, catalysed by kynureninase and kynurenine hydroxylase which are vitamin B_6 and riboflavin dependent. A deficiency of either vitamin B_6 or riboflavin may aggravate niacin deficiency. Some 60 mg of dietary tryptophan generates 1 mg of nicotinic acid. The nicotinic acid equivalent is the dietary nicotinic acid content plus 1/60th of the dietary tryptophan.

Nicotinamide is a component of the coenzymes nicotinamide adenine dinucleotide (**NAD+**) and nicotinamide adenine dinucleotide phosphate (**NADP+**, also known as triphosphopyridine nucleotide (**TPN+**)). NAD coenzymes are biological carriers of reducing equivalents, that is to say electrons, during metabolic oxidation to NADH+. $NADH^+$ acts as a true coenzyme in enzymes involved in epimerization, aldolization, and elimination.

The various conjugated forms of niacin are hydrolysed and absorbed in the upper gastrointestinal mucosa as the free acid. Niacytin from maize is neither hydrolysed nor absorbed. In Central America, maize is eaten as tortillas, in which lime water hydrolyses the nicotinoyl ester component. Niacin and tryptophan deficiencies occur in poor populations dependent upon maize for their protein intake. Zein, the principal maize protein, is deficient in tryptophan. Protein energy malnutrition, dietary amino acid imbalances (for example, an excess leucine intake), anaemia, and other vitamin deficiencies worsen the problem. Dietary fortification with other proteins is necessary.

An inadequate dietary intake of 1- to 2-months' duration leads to significant tissue depletion and pellagra, which if untreated progresses to death. There is no apparent tissue storage of this vitamin. The vitamin is excreted in the urine as nicotinuric acid, nicotinamide- *N*-oxide and 5'-methylnicotinamide. Pellagra is characterized by dermatitis, diarrhoea, and dementia; the disease is chronic with a seasonal periodicity. Individuals suffer weight and stamina loss, which is worsened by secondary bacterial and parasitic infections. Erythematous dermatitis is symmetrically distributed on skin exposed to sunlight and mechanical irritation. In chronic pellagra the skin looks sunburnt. In severe cases, gastrointestinal disturbances occur with diarrhoea. There can be glossitis, angular stomatitis, cheilosis, and an inflamed tongue, and secondary infection of the mouth. Mild mental changes include anxiety and irritability progressing to manic depressive illness. There may be paraesthesia in the lower limbs, with loss of vibration sense and proprioception leading to ataxia and spasticity.

Pellagra can occur in alcoholism, malabsorption syndromes, and Hartnup disease. Mild cases rapidly improve on treatment with niacin or a suitable dietary protein supplementation. Oral nicotinamide (100 mg every 4 h) results in symptom resolution within 24 h, although mental symptoms, especially dementia, may be unresponsive to treatment. Nicotinic acid may cause unpleasant flushing and burning sensations. A supplementary diet containing good-quality protein and all of the vitamins is important. Nicotinic acid, but not nicotinamide, is used therapeutically for hyperlipidaemias at a dose of 2 to 6 g/day, but it can be hepatotoxic.

Microbiological methods provide the most sensitive means of measuring niacin and nicotinamide in serum, urine, and food. Alternatively, urinary nicotinamide, *n*-methyl nicotinamide (**NMN**), is measured over a defined time or as a ratio of creatine in urine. Urinary *N*-methyl-2-piridone-5-carboxamide (2-pyridone) excretion is a more sensitive measurement in borderline cases of nicotinamide deficiency, and is expressed in relation to urinary creatinine excretion.

Requirements

These are related to dietary tryptophan intake. A protein intake of between 60 and 85 g/day contains approximately 13 mg tryptophan/g, equivalent to between 13 and 17 mg/day of niacin. The recommended intake as niacin equivalents is 6.6 mg (54 µmole)/4185 J (4.185 J = 1 kCal). Hormonal changes during pregnancy affect tryptophan metabolism, so that 30 mg of tryptophan is equivalent to 1 mg of dietary niacin. Breast milk should provide not less than 3.5 mg preformed niacin/4185 J. Mature human milk provides preformed niacin (2.7 mg/l), therefore an increment of 2 mg/day niacin for the nursing mother is suggested.

Pantothenic acid

Pantothenic acid is the dimethyl derivative of butyric acid joined by a peptide linkage to a-alanine. The active form, 4'-phosphopantetheine, is present in all tissues. Pantothenic acid is thermo- and acid-labile. It is widely available in animal-based foods, especially liver, although cereals and legumes are also sources. 4'-Phosphopantetheine is a constituent of CoA and acyl carrier protein. CoA plays a central role in intermediary metabolism, fatty-acid b-oxidation, sterol synthesis, and other acetylation processes. Acyl carrier protein is an acyl carrier in the synthesis of lipids. Pantothenic acid is found in food as the coenzyme CoA or acyl carrier

protein form and is hydrolysed by a pancreatic enzyme before absorption. Urinary excretion is in the free acid form.

Pantothenic acid is not stored in the body. Although no specific deficiency syndrome has been recorded, pantothenic acid deficiency may occur as part of the overall problem in people who are severely malnourished. No toxic intakes have been recorded. There is no biochemical method for measuring pantothenic acid status in humans.

Requirements

Most human diets provide 3 to 10 mg of pantothenic acid per day. A safe and adequate intake is between 3 and 7 mg per day, including during pregnancy and lactation. Infants require 1.7 mg/day; human milk provides 2.6 mg/day. Infant formula milk should contain at least 2 mg/litre.

Riboflavin

Riboflavin is a substituted alloxazine ring linked to ribotol, an alcohol derived from the pentose sugar ribose. It is light sensitive. Dietary sources are liver, milk, cheese, eggs, some green vegetables, and beer. Other sources are yeast extracts (for example, Marmite) and meat extracts (for example, Bovril). Riboflavin in the diet exists in either the free form or the phosphorylated coenzyme form.

Riboflavin links with phosphoric acid as flavin mononucleotide (or riboflavin–5'-phosphate) (**FMN**), which with adenosine monophosphate (**AMP**) forms flavin adenine dinucleotide (**FAD**): the prosthetic groups of the flavoprotein enzymes. Flavoproteins are involved in redox processes involving the hydrogen-transfer chain in the mitochondria and the production of ATP. FAD/FMN acts as a coenzyme in oxidation/reduction reactions, electron transport, oxidative phosphorylation (for example, succinic dehydrogenases), and fatty-acid b-oxidation.

Riboflavin is absorbed from the upper gastrointestinal tract, there is no specific storage tissue, and it is excreted in the urine either free or in small amounts of hydroxylated products. Chronic infection can affect urinary riboflavin excretion. A deficiency of riboflavin causes cheilosis, angular stomatitis, superficial interstitial keratosis of the cornea, and nasolabial seborrhoea. Riboflavin deficiency may impair iron absorption. No toxic effects have been shown for riboflavin.

Riboflavin status can be estimated from the urinary riboflavin:creatinine ratio, which is insensitive at low intakes. The erythrocyte glutathione reductase activation coefficient (**EGRAC**) measures tissue saturation and long-term riboflavin status.

Requirements

Adults, including the elderly, need between 1 and 1.5 mg of riboflavin per day. The average riboflavin content of breast milk in Britain is approximately 0.3 mg/l, which is dependent upon maternal intake. Intakes should increase by 0.3 mg/day during pregnancy, and 0.5 mg/day during lactation. Recommended intakes for children range from 0.4 mg/day for infants up to 3 months of age and 1.0 mg/day thereafter.

Thiamin

Thiamin hydrochloride consists of a substituted pyrimidine ring linked by a methylene group to a sulphur-containing thiazole ring. All animal and plant tissues contain thiamin, usually in the phosphorylated form. The important sources are plant seeds and cereal germ, nuts, peas, beans, pulses, and yeast. Losses occur with cooking and alkaline pH. Thiamin diphosphate is the coenzyme in a-ketoacids and decarboxylation reactions involved in the oxidative decarboxylation of pyruvic acid to acetyl-CoA, and the transketolase reaction in the hexose monophosphate shunt. Thiamin may also play a role in neural excitation mechanisms.

Absorption is from the upper gastrointestinal tract, followed by phosphorylation to the active diphosphate form. There is no body store and the only reserve is the vitamin functionally bound to enzymes. Multiple endproducts are excreted in the urine. Beriberi is caused by dietary thiamin deficiency, a disease that was endemic in the East as a result of the ingestion of polished rice which is deficient in the vitamin. Carbohydrate metabolism is impaired by a deficiency of thiamin pyrophosphate, a coenzyme necessary for the decarboxylation of pyruvate to acetyl-CoA. Pyruvic and lactic acid accumulate in the body. The clinical presentations of thiamin deficiency are:

- wet beriberi (high-output cardiac failure);
- dry beriberi (polyneuropathy);
- infantile beriberi;
- neuropathy and cardiomyopathy in chronic alcoholism; and the
- Wernicke–Korsakoff syndrome.

Initially the symptoms are of non-specific malaise and evidence of early cardiac failure and neuropathy. Wet beriberi is characterized by left- and right-sided high-output cardiac failure, cardiomegaly, hypotension, rapid deterioration, and death. Dry beriberi is a polyneuropathy affecting motor and sensory nerves. Initially there is paraesthesia progressing to painful muscle wasting and polyneuritis. Total sensory loss occurs and patients become immobile and emaciated; they are at a high risk for the development of Wernicke–Korsakoff's encephalopathy. Infantile beriberi occurs in breast-fed infants of thiamin-deficient mothers, usually when they are between 2- and 5-months old. In the acute form, the child is restless, distressed with evidence of high output cardiac failure; convulsions may develop and the child becomes comatose. In the chronic form, the child is fretful, sleeps poorly, and the muscles may be flaccid. Cardiac failure, gastrointestinal symptoms, and sudden death are common. Alcoholic neuropathy presents with a sensory and motor neuropathy sometimes complicated by cardiomyopathy. Sensory nerve dysfunction includes paraesthesia and severe nerve pain. Motor nerve lesions are of both upper and lower motor neurone type. A patient with Wernicke–Korsakoff syndrome is disorientated and apathetic. Nystagmus, ataxia, and confabulation are not infrequent consequences of lesions in the brainstem, diencephalon, and cerebellum.

Treatment is with intramuscular thiamin 25 mg twice daily for 3 days, and thereafter 10 mg two or three times daily. The reversal of the wet type of beriberi is rapid. Improvement is slow for dry beriberi, especially for the neurological abnormalities. In infantile beriberi, both mother (10 mg thiamin twice daily) and the infant (thiamin intramuscularly 10 to 20 mg/day for 3 days, and thereafter 5 to 10 mg twice daily) are treated. Beriberi can be prevented by eating thiamin-containing foods, unmilled or thiamin-fortified rice, or thiamin supplements.

Long-term intakes in excess of 50 mg per kg body weight/day are toxic, leading to headaches, irritability, insomnia, rapid pulse, weakness, contact dermatitis, pruritus, and even death. Thiamin status can be measured by urinary thiamin and the thiamin:creatinine ratio with or without a loading dose, or by the reactivation of the cofactor-depleted red-cell enzyme transketolase *in vitro*.

Requirements

Thiamin requirements are related to energy and carbohydrate metabolism. The average requirement for adults, normal pregnancy or lactation, and children is 0.4 mg/4185 J and not less than 0.8 mg/day for adults with a supplement of 0.6 mg/4185 J during pregnancy and lactation. Human breast milk contains the equivalent of 0.3 mg/4185 J. The elderly may require 1 mg/day.

Vitamin B_6 (pyridoxine)

Pyridoxine is to be found in several forms: pyridoxal, pyridoxamine, and pyridoxine. This family of vitamin B_6 compounds is found in many foods: cereals, meat (particularly liver), fruits, and leafy and other vegetables. The free form is common in plants, the phosphorylated form, pyridoxamine phosphate, in animal tissues. Pyridoxal-5'-phosphate is a coenzyme and plays a major role in the intermediary metabolism of amino acids, in a-decarboxylation, aldolization, and transamination reactions. Pyridoxal-5'-phosphate acts as a coenzyme with glycogen phosphorylase in muscle, and has a role in the actions on hormones which modulate gene expression. Vitamin B_6 is absorbed in the free form and phosphorylated for use in enzymes. There is no specific storage in tissues and it is excreted in urine largely as 4-pyridoxic acid.

Primary dietary deficiency has not been reported in adults, largely because of the wide availability of the vitamin. A biochemical deficiency occurs in alcoholics. Patients taking isoniazid may develop a pyridoxine deficiency and neuropathy. Women taking massive supplements of between 2 and 7 g/day, a dose which has been recommended for alleviating premenstrual symptoms, may develop a sensory neuropathy.

There is no single marker sensitive at all levels of dietary intake. Biochemical markers include plasma pyridoxal phosphate concentrations, red-cell transaminase activation, and the urinary excretion of vitamin B_6 degradation products. Metabolic loading tests also measure vitamin B_6 status, including the tryptophan- and methionine-load tests.

Requirements

The total body pool of vitamin B_6 is 15 µmole (4 mg)/kg, 80 per cent of which is in muscle, with a half-life of 33 days. Adults, pregnant and lactating women, and the elderly require a daily intake of 13 µg/g of protein (approximately 4 mg/day). The vitamin B_6 content of human breast milk is low at between 40 and 100 µg/l (or 3–8 µg vitamin B_6/g protein). Infants under 3 months of age require 6 µg/g protein, increasing to up to 13 µg/g protein at 7 to 10 years.

Folate (folic acid, folacin)

Sources of folate include liver, yeast extract, and green leafy vegetables. Folates are derivatives of folic acid (pteroylglutamic acid) including the folylpolyglutamate found in foods. Folic acid is a pterin ring (2-amino, 4-hydroxypteridine) attached to *p*-aminobenzoic acid conjugated to L-glutamic acid (**PteGlu**). Variants include:

- di-(7,8-tetrahydrofolic acid) (**DHF**) and tetra-(5,6,7,8-tetrahydrofolic acid) (**THF**) reduced forms of the pteridine ring;
- one-carbon substitution (methyl, formyl, methenyl, methylene, or formimino) at positions N5 or 10: 5-formyl-THF, 10-formyl-THF, 5-formimino-THF, 5,10-methenyl-THF, 5,10-methylene-THF, and 5-methyl-THF;
- a chain of 4–6 glutamates attached to the L-glutamate.

Folic acid gives and receives 1-carbon groups on the N5 or N10 position in nucleic acid and amino acid biosynthetic reactions. Most dietary folate is in the polyglutamyl form, which is hydrolysed to monoglutamate before being absorbed from the duodenum. A brush-border glutamyl carboxypeptidase is inhibited by alcohol, which is of relevance in alcoholic folic deficiencies. Folate bound to a milk protein is absorbed from the ileum. Folic acid is stored in the liver. Plasma folates are mainly 5-methyl-THF monoglutamate. Within cells, 5-methyl–THF is converted to THF polyglutamates, the main cellular forms of folic acid.

Folate polyglutamates do not readily cross cell membranes. The polyglutamate form has two functions: storage; and as a coenzyme for normal 1-carbon metabolism (for which it is the most efficient coenzyme). The 5-methyl group is transferred to homocysteine (creating methionine and THF); the enzyme is the vitamin B_{12}-dependent methionine synthase, wherein methionine, folic acid, and vitamin B_{12} interlink.

Reactions in which folate is involved include:

- methylation of amino acids;
- serine reversibly interconverting with glycine;
- methionine interconverting with homocysteine (methionine is the precursor of *S*-adenosyl-L-methionine (**SAM**), a methyl donor in the methylation of lipids, hormones, DNA, cell division, and proteins); and
- thymidine and purine synthesis.

Folic acid deficiency may arise:

- as a dietary deficiency;
- in malabsorption syndromes;
- where there are excessive demands, as with increased cell proliferation (for example, in leukaemias and haemolytic anaemias);
- where drugs interfere with folic acid metabolism; and
- in the rare inborn errors of folic metabolism.

Folic acid deficiency is an important cause of megaloblastic anaemia, never to be confused with vitamin B_{12} deficiency.

Neural tube defects (**NTDs**) are congenital deformities of the spinal cord and brain: spina bifida, anencephaly, encephalocele, and iniencephaly. Folic acid is involved in the aetiology of NTDs. The precise mechanism is unclear, but there may be an underlying genetic predisposition involving a variant of 5,10-methyl-THF reductase. Closure of the neural tube occurs early in pregnancy thereby making aetiological studies difficult. The results of recommendations to take folic acid supplements prophylactically are encouraging.

Epidemiological studies suggest an increased risk of vascular disease associated with hyperhomocysteinaemia. Homocysteine is reversibly methylated to methionine, a step which involves folate, vitamin B_{12}, and vitamin B6. Supplementation of these vitamins has been proposed to reduce the putative dangers of hyperhomocysteinaemia.

Folate status is measured by the folate concentration in serum and red cells. Red-cell folate levels reflect body stores. A coincidental measurement of serum vitamin B_{12} is important.

Requirements

Children and adults, including the elderly, need a folate intake of 200 µg/day. Women planning a pregnancy should increase their intake of folic acid to 0.4 mg/day by capsule supplement. If there has been a previous NTD-affected pregnancy then 5 mg of folic acid/day preconception is suggested. A problem is that some pregnancies are unplanned. Dietary supplementation is impractical. Total folic acid excretion in breast milk averages 40 µg/day, and an additional maternal intake of 60 µg per day is required.

Vitamin B_{12}

Vitamin B_{12} is a cobalt-containing corrinoid, four linked pyrrole rings (corrin) co-ordinating with a central cobalt atom. The cobinamides necessary for human well being are methylcobalamin, adenosylcobalamin, hydroxycobalamin, and cyanocobalamin. Micro-organisms, including colonic flora, synthesize cobalamin. Yeast is a source of cobalamin, primarily as adenosyl- and hydroxocobalamin. Methyl cobalamin is found in egg yolk, cheese, and cow's milk. A vegan diet carries a risk of vitamin B_{12} deficiency.

The reactions requiring vitamin B_{12} include:

- isomerization of methylmalonyl-CoA to succinyl-CoA (methyl malonic acid concentrations increase in cases of vitamin B_{12} deficiency);
- methyltransferase reactions: for example, homocysteine to methionine, i.e. the transfer of a methyl group from 5-methyl-TFH to homocysteine which converts homocysteine to methionine.

Vitamin B_{12} has an important role in the maintenance of myelin. Deoxyadenosyl B_{12} is essential for propionyl-CoA reactions by transmutation of methylmalonyl-CoA to succinyl-CoA.

Vitamin B_{12} binds to food proteins and is released by saliva, acid pH, and pepsin, depending upon the mode of cooking and type of food protein. Vitamin B_{12} at stomach pH forms complexes with glycoproteins, transcobalamin, haptocorrin, and intrinsic factor. In the duodenum, cobalamin is released by pancreatic enzymes and alkaline pH and binds solely to intrinsic factor. The vitamin B_{12}/intrinsic factor complex is absorbed from the ileum through a specific receptor. Cobalamin is released from intrinsic factor, converted (80 per cent to methyl and also adenosyl and hydroxy forms) and carried in the blood by transcobalamins I, II, and III. Of these, transcobalamin II releases vitamin B_{12} to the tissues to be stored in the adenosyl form. The total body cobalamin content in adults is between 2 and 5 mg, most of which is stored in the liver. Turnover is 0.1 per cent of the body pool each day. There is efficient conservation by the kidneys and the enterohepatic circulation. Most vitamin B_{12} is excreted in urine and small amounts in faeces, including unabsorbed bacterially synthesized vitamin B_{12}. The relationship between dietary intake and serum concentrations is not linear because the body stores of vitamin B_{12} are largely in the liver.

Vitamin B_{12} deficiency results in megaloblastic anaemia and neurological disorders, especially in the posterolateral columns of the spinal cord. Causes of deficiency are dietary (vegans), lack of intrinsic factor (Addison's pernicious anaemia, gastric resection), intestinal colonization by bacteria and parasites (for example, tapeworms), and ileal resection. The megaloblastic anaemia is due to a lack of vitamin B_{12} for methionine synthase, insufficient methionine regeneration, and 5,10-methylene-THF and deficient thymidylate synthesis. In vitamin B_{12} deficiency, odd number 15- and 17-carbon fatty acids and branch-chain fatty acids are synthesized and incorporated into an unstable myelin nerve sheath. An inability to regenerate methionine from homocysteine, for the *S*-adenosylmethionine generation necessary for myelin proteins, leads to demyelination.

The efficiency of vitamin B_{12} absorption is measured by the Schilling test or by a whole-body scanner.

Requirements

A dietary intake of between 1 and 2 µg/day for adults of all ages is protective. In pregnancy there is a compensatory increased absorption so that 1.5 µg/day is sufficient. During lactation an increment of 0.5 µg/day should ensure an adequate supply in breast milk (0.2 to 1.0 µg/l). The requirement for infants is of the order of 0.1 µg/day and for children aged 3 to 10 years, 0.5 to 1.0 µg/day. Vitamin B_{12} has extremely low toxicity and as much as 3 mg/day may be taken.

Vitamin A

The vitamin A family (retinoids) are related to the plant pigment carotene. Dietary vitamin A comes in two forms: from animal sources, preformed, vitamin A fatty-acid esters; or from plant, provitamin A carotenoids. b-Carotene, a provitamin A carotenoid, consists of two retinol molecules. Retinol is vitamin A alcohol, a hydrocarbon chain with a betaionone ring at one end and an alcohol group at the other, usually esterified with a fatty acid (retinyl esters) as the *all-trans* stereoisomer. The *cis* configuration isomer (11 or 13 position) is less potent. Retinol can be oxidized to an aldehyde (retinal) or acid (retinoic acid). Retinol and carotene are readily oxidized and are protected by vitamin E.

Retinol is present in dairy products, liver, and fatty fish liver oils. Carotenes are found predominantly in green vegetables as well as in yellow and red fruits and vegetables. Vitamin A is essential for the maintenance of epithelial tissue, visual function, and the immune system. Most actions of vitamin A in development, differentiation, and metabolism are mediated through retinoid receptors of the nuclear steroid receptor family of proteins that bind retinoic acid and regulate gene expression.

The photopigment rhodopsin is formed by the protein opsin and 11-*cis* retinal. A photon of light converts the *cis* retinal to the *trans* form, which reversibly dissociates from the opsin and is seen as light. Retinyl esters are hydrolysed by pancreatic hydrolases and the enteric mucosa. b-Carotene is cleaved to two retinols by b-carotene 15,15'-deoxygenase, and carotenoids oxidatively cleave to retinal and apocarotenoids. Retinal is reduced to retinol, esterified with long-chain fatty acids, and transported to the liver as retinyl long-chain fatty esters in chylomicrons through the lymph. Retinol is a major liver storage form and circulates to tissues, bound to retinol binding protein. There is an enterohepatic circulation of retinoids.

Vitamin A deficiency is a major worldwide cause of blindness, due to a poor dietary intake of green vegetables, fruit, and dairy produce. Malabsorption is a less common cause. Vitamin A deficiency results in reduced rhodopsin in the retinal rods resulting in loss of vision. Xerophthalmia causes blindness in 500 000 young children each year, especially amongst bottle-fed infants and breast-fed infants with vitamin A-deficient mothers. Protein calorie malnutrition compounds the problem during weaning. Vitamin A deficiency aggravates damage from other causes of keratoconjunctivitis, for example measles. Epithelial surfaces undergo squamous metaplasia, followed by corneal ulceration and irreversible visual damage. Clinical forms are:

- conjunctival xerosis (Bitots's spots are white plaques of thickened conjunctival epithelium indicative of vitamin deficiency in the young);
- corneal xerosis;
- keratomalacia, leading to blindness;
- night blindness, an early symptom with or without xerophthalmia;
- xerophthalmia fundi; and
- corneal scars.

Vitamin A is important in epithelial metabolism. Deficiency leads to epithelial metaplasia and inappropriately keratinized epithelium in the mucous membranes of the respiratory, gastrointestinal, and genitourinary tract. Sebaceous glands become blocked thereby causing follicular keratosis.

Prophylaxis demands education in the eating of dark-green vegetables. Where xerophthalmia is endemic, vitamin A is given prophylactically in capsule form or by food fortification. Frank deficiency is treated by high-potency Vitamin A: 200 000 IU for 2 days, and a third dosage at least 2 weeks later. Thereafter, improved diet and supplementation is obligatory. Corneal ulceration is treated by antibiotics, and the response is rapid. Vitamin A deficiency is a putative risk factor for childhood morbidity and death, especially for the underweight and premature infant.

b-Carotene is not toxic, but high intakes lead to a yellow appearance sparing the eyes (hypercarotenaemia). Polar bear's liver, rich in retinol, is toxic and ingestion can cause drowsiness, headache, vomiting, and excess peeling of the skin. Large intakes of retinol are teratogenic. Pregnant women should be careful not to exceed the recommended intake of vitamin A in the first trimester.

Plasma retinol is an insensitive indicator of vitamin A status. The relative-dose-response (**RDR** test), which measures retinol transport by the retinol binding protein, is used as a functional test for calculating retinol stores. The concentration of plasma carotenoids reflects short- to medium-term intakes. The following are equipotent to 1 µg of *all-trans*-retinol equivalents/day: 3.33 IU vitamin A, 3.5 nmol retinol or retinyl ester, or 6 µg *all-trans*-b-carotene.

Requirements

Adults require 500 µg retinol equivalents/day; infants, 250 to 350 µg retinol equivalents/day; children, 350 µg retinol equivalents/day; pregnancy, an increment of 100 µg retinol equivalents/day, particularly during the third trimester. Lactation requires an increment of 300 µg retinol equivalents/day. Breast milk vitamin A concentration should exceed 1.5 mmol/l.

Vitamin D

The vitamin D family of sterols includes vitamin D_3 (cholecalciferol) and vitamin D_2 (ergocalciferol). Cholecalciferol is to be regarded as a hormone rather than a vitamin, and is produced by the ultraviolet irradiation of dietary 7-dehydrocholesterol (provitamin D_3) in the skin. The extent of exposure to sunlight determines production. Cholecalciferol is also available in fatty fish (for example, cod), eggs, and chicken liver. Ergocalciferol is also the result of the exposure of ergosterol (provitamin D_2) to ultraviolet light. Ergocalciferol differs from cholecalciferol in having an extra methyl group at C-24 and a double bond at C-22,23.
1,25-dihydroxycholecalciferol vitamin D (**1,25(OH)2D**) regulates calcium and phosphate absorption, metabolism, and export into the bloodstream. Such regulation is through steroid:thyroid hormone nuclear receptors. $1,25(OH)_2D$ is also a developmental hormone inhibiting proliferation and promoting differentiated function in cells.

Dietary vitamin D is absorbed in the small intestine, as a lipid, transported to the liver bound to a-globulin (*trans*-calciferol) in chylomicrons. Both vitamin D_2 and vitamin D_3 are inactive. They are converted in the liver by a P-450 inducable microsomal enzyme into 25-hydroxy vitamin D (**25(OH)D**), which has modest biological activity, before plasma transport on a specific globulin. The active form of vitamin D is 1,25(OH)2D formed in the kidney by a mitochondrial hydroxylase acting on 25(OH)D. The half-life of $1,25(OH)_2D$ is less than 24 h. All forms of vitamin D are stored in fat. Vitamin D is 24-hydroxylated and, as the glucuronide, is excreted in bile into an enterohepatic circulation.

The prime consequence of vitamin D deficiency is rickets, caused by a failure to mineralize the bony skeleton. The epiphyseal cartilage replacement is defective, leading to an overgrowth of subperiosteal osteoid tissue and poor mineralization of the bone matrix resulting in soft bones. The type of bony abnormalities depends on the age of onset and the weight-bearing bones involved. The appearance of the rachitic child is deceptive, the child may appear quite well or be restless with hypotonic muscles and twisted limbs' postures. The abdomen is swollen and the child suffers from diarrhoea, respiratory infection, and delayed tooth development. In the infant, the commonest abnormality is enlargement of the end of the radius and the costochondrial junction—termed the ricketic rosary. Later there is bossing of the frontal and parietal bones and delayed closure of the anterior fontanel; so-called 'pigeon chest' can result, which is an undue prominence of the sternum and a transverse depression from the costal margins towards the axillae. All are due to pressure on the soft bones when the child is supine. In the walking child, the weight-bearing bones bend, kyphosis of the spine and bowing of the lower ends of the femur, tibia, and fibula result. Pelvic deformity can make delivery difficult in subsequent pregnancies. Tetanic spasm can occur, whereby spasm of the hands, feet, and vocal chord result in high-pitched cries and breathing problems.

Diagnosis is based on the clinical appearance and measurements of plasma alkaline phosphatase (although interpretation of the results is difficult in the growing child) and plasma 25(OH)D levels. There are several risk factors:

- inadequate exposure to sunlight (dependent on latitude, in the United Kingdom 30- to 90-min exposure of the face and legs per day will restore 25(OH)D concentrations;
- strict vegetarianism;
- a vitamin D-deficient mother breast feeding her baby;
- high melanin content in the skin, which screens the metabolically active skin sites; and
- malabsorption.

Prevention of vitamin D deficiency requires a supplement of 10 µg of vitamin D daily or regular exposure to sunlight in well-nourished individuals.

Osteomalacia, may present with muscular weakness and a waddling gait. Bone pain, tetany, and spontaneous fractures may develop. Radiographic features include bone rarification, pseudofractures, and Looser's zones at points of compression stress. Renal disease, from many causes, may be associated with impaired renal synthesis of $1,25(OH)_2D$. A failure of 25-hydroxylation of vitamin D can occur in hepatic disease. Dietary vitamin D intake may be important in immunity to tuberculosis, and 25-OH vitamin D deficiency may contribute to the occurrence of tuberculosis. Hypervitaminosis occurs during infant supplementation and replacement therapy. Plasma calcium concentrations increase with tetany, ECG changes with resultant convulsions, and occasionally death. Vitamin D in milligram amounts is poisonous, and is used as a rodenticide.

The best measure of the vitamin status in humans is the plasma 25(OH)D concentration.

Requirements

No minimum dietary intake has been identified for adults exposed to ample sunlight. However, 10 µg/day vitamin D is recommended for those with poor sun exposure. Breast-milk vitamin D concentrations are low (0.25–1.25 µg/litre) and are reduced in the winter. Vitamin D intakes are a problem for the 6- to 12-month-old baby dependent upon modestly fortified weaning foods. The diet thereafter expands and the plasma 25(OH)D concentrations are usually satisfactory. Pregnant and lactating women should receive supplementary vitamin D at 10 µg/day. Where the elderly are insufficiently exposed to the summer sun, their stores may be reduced and a supplement of 10 µg/day vitamin D is recommended.

Vitamin K

Vitamin K, a naphthoquinone, occurs in two forms in human nutrition: vitamins K_1 and K_2. Vitamin K_1 of plant origin is a phytylmenaquinone (also known as phylloquinone or phytylmenadione) and consists of 2-methyl-1,4-naphthoquinone (menadione or menaquinone) attached to a 20-carbon phytyl side chain. Vitamin K_2 is one of several homologues produced by bacteria with 4 to 13 isoprenyl units in the side chain (menaquinone-4 to -13). Vitamin K_1 is present in fresh green vegetables (for example, broccoli, lettuce, cabbage, and spinach) and beef. Vitamin K is involved in the synthesis of proteins central to blood coagulation, namely prothrombin and factors VII, IX, and X. Vitamin K is necessary for the post-translational carboxylation of glutamic acid in the coagulation proteins. g-Carboxyglutamate allows the binding of calcium and phospholipids in the formation of thrombin.

Vitamin K is absorbed as a lipid, and is transported from the intestine in the blood in chylomicrons as b-lipoproteins. Vitamin K_2 of bacterial origin is absorbed from the colon. When there is vitamin K deficiency, the blood clotting time is prolonged and factor VII, IX, and X activities are reduced. Deficiency is uncommon in adults. However, in infants deficiency results from a sterile intestinal tract and the inadequate vitamin K content of human and cow's milk. The problem is compounded by the immature liver of the infant being slow to synthesize prothrombin. Acquired deficiencies occur as a result of any cause of lipid malabsorption and after bowel sterilization with broad-spectrum antimicrobial agents. Vitamin K deficiency may also result from the regular ingestion of liquid paraffin, since the vitamin partitions preferentially into this non-absorbed, non-polar hydrocarbon oil and is excreted rather than absorbed. Liquid paraffin is still used in many parts of the world as a regular aperient, principally by the elderly.

Natural vitamin K preparations are free from toxic effects. There are naturally occurring vitamin K antagonists—for instance, spoilt sweet clover produces a dicoumarol that prolongs the prothrombin time of the cow, thereby causing a bleeding condition. Drugs designed to prolong prothrombin time were developed as a result of this observation. Vitamin K deficiency can be detected by the 'prothrombin time' test which measures prolongation of clotting time.

Requirements

The children and adult dietary requirements of phylloquinone are between 0.5 and 1.0 µg/kg body weight per day. Vitamin K in human breast milk is as the phylloquinone and the concentration varies between 1 and 10 µg/l. An adequate intake for breast-fed infants is 8.5 µg phylloquinone/day. Vitamin K is given as supplements in malabsorption syndromes and prophylactically for haemorrhagic disease of the newborn.

Vitamin E

The vitamin E family consists of fat-soluble biologically active tocopherols and tocotrienols. The tocopherols are the most potent, their activity depending upon the position and number of methyl substitutions. a-Tocopherol, is the most potent; g-tocopherol and g-tocotrienol have activity of 48 and 20 per cent, respectively, when compared to a-toco-pherol.

Vegetable oils—wheat germ, sunflower seed, cottonseed, safflower, palm, rape seed, and other oils are abundant sources of vitamin E. The free-radical scavenging properties of vitamin E are a function of the fused chroman ring system, the phytyl side chain facilitates entry into the hydrophobic environment of the membrane. Ascorbic acid may reduce tocopheroxyl radicals formed by the scavenging of free radicals during metabolism. This enables a molecule of tocopherol to scavenge many radicals. The absorption of the vitamin is incomplete and varies between 20 per cent and 80 per cent. Vitamin E enters the systemic circulation in chylomicrons and very low-density lipoproteins (**VLDL**) and coincidentally protect the polyunsaturated fatty acids (**PUFA**) which are also transported. Lipoprotein lipase controls uptake by the liver or transfer to other lipoproteins. The normal lipoprotein concentrations of vitamin E as a-tocopherol range from 11 to 37 µmol/l. a-Tocopherol forms 90 per cent of the vitamin E found in tissues, including all cell membranes where it inhibits the non-enzymatic oxidation of PUFA by molecular oxygen. Biochemical deficiency may occur as a result of gastrointestinal malabsorption and in premature infants. Vitamin E deficiency has been implicated in peripheral neuropathy associated with malabsorption syndromes and often accounts for the acanthocytosis, retinitis pigmentosa, and neurological features in Bassen-Kornzweig disease (abetalipoproteinaemia). Patients with fat malabsorption due to cystic fibrosis, coeliac disease, prolonged cholestasis, and after massive small intestinal resections

are particularly at risk from vitamin E deficiency; supplements prevent and may slowly ameliorate the neurological deficit. Homozygosity for inactivating mutations in the a-tocopherol transfer protein is a cause of isolated vitamin E deficiency and ataxia which closely resembles the spinocerebellar ataxia of Friedreich's disease. Patients with Friedreich's ataxia in the absence of frataxin mutations should be investigated for vitamin E deficiency and defects in the tocopherol transfer protein; vitamin E supplements may slowly improve this condition. Longstanding profound deficiency of vitamin E is associated with progressive spinocerebellar ataxia, visual loss due to retinitis pigmentosa, haemolysis (with red cell acanthocytosis), upward visual gaze palsies, dementia, and muscle weakness. If detected, vigorous long-term vitamin E supplementation is indicated as well as attention to the primary cause of the deficiency. There appears to be no adverse effects from large doses of vitamin E up to 3200 mg/day.

Vitamin E status can be measured from the plasma tocopherol concentration, or expressed as a ratio of total blood lipids or vitamin E:cholesterol. A functional test of vitamin E status is the hydrogen peroxide haemolysis test (erythrocyte stress test).

Requirements

The average intake in Britain is 6 mg/day, most of which is derived from fats, oils, and cereals. The dietary requirement is determined by the PUFA content of membranes and tissues and the PUFA content of the diet. The relationship between PUFA intake and vitamin E requirements is not a simple linear relationship. Intakes of 4 mg and 3 mg of a-tocopherol equivalents per day, respectively, for men and women have been regarded as adequate, but may be too low. Alternatively, and better, would be 0.4 mg a-tocopherol equivalents per gram of dietary PUFA/day, thereby increasing the recommendation to 7 mg. This formula might also be used for infants. Human breast milk contains 10 mg of a-tocopherol equivalents/l in colostrum, reducing to 3.2 mg/l at 12 days and thereafter.

Trace elements

Trace elements are important, since they:

- act as cofactors in enzyme oxidation–reduction reactions;
- maintain the specific configuration of proteins;
- are incorporated into the structure of hormones; and
- play a structural and catalytic role in gene expression and transcriptional regulation of genes.

Trace elements are required in small amounts. The trace elements contained in soil and drinking water vary from area to area, which determines the variation in intake seen in different communities. The amount and chemistry of dietary constituents eaten with the trace elements affects the absorption efficiency of the essential elements. The absorption of calcium and trace metals (for example, zinc) can be inhibited by dietary phytate. Copper absorption is reduced by competitive interactions affecting its solubility. A mild degree of iron depletion increases not only iron absorption but also lead, zinc, cadmium, cobalt, and manganese.

Individual trace elements

[The following abbreviations are used below: At wt, atomic weight; val, valency; iso, natural isotope; Abund, natural abundance in earth's crust as a per cent of the total.]

Cobalt

At wt: 59; val: 2, 3; iso: 59; abund: 0.0018 per cent

Sources are wholemeal flour and seafoods. Cobalt's role is as a component of vitamin B12. Uncomplexed cobalt can be absorbed and subsequently excreted in urine. Intakes of cobalt are approximately 5 µmole/day and the total body content is 1.5 mg (7.5 µmole). In Quebec, a cobalt-containing beer improver (15 µmol of cobalt/l) proved to be toxic; its best customers developed severe cardiomyopathy.

Chromium

At wt: 52; val: 2, 3; iso: 50, 52, 53, 54; abund 0.033 per cent.

Chromium is present in most foods especially wheat germ, molasses, green beans, and broccoli. Dietary intakes vary between 5 and 100 µg/day, but absorption is meagre at 1 per cent. The plasma concentration of chromium is 0.3 µg/ml bound to transferrin, and it is excreted in urine. Deficiency increases the risk of type II diabetes mellitus and cardiovascular disease. The recommended adult intake is 0.5 µmole/day and between 2 and 19 nmole/kg per day for children and adolescents. The adult body contains 100 to 200 µmole. The chromium content of human milk is 0.06 to 1.56 ng/ml. Chromium in high dosage is well tolerated.

Copper

At wt: 64; val: 1, 2; iso: 63, 65; abund: 0.010 per cent.

Good sources of copper include green vegetables, fish, oysters, and liver. Copper is required by the immune, nervous, and cardiovascular systems for skeletal development, iron metabolism, red-cell formation, and enzymes including cytochrome oxidase and superoxide dismutase. Between 35 and 70 per cent of ingested copper is actively absorbed, but this is affected by age and the chemistry of accompanying food. Copper is concentrated in the liver, excreted in bile, and lost in faeces. In plasma, copper is bound to caeruloplasmin and albumin. The total amount of copper in an adult is approximately 2 mmole (50 to 120 mg). Copper accumulates in the fetal liver for early extrauterine life, premature birth results in depleted copper stores. Copper deficiency in adults has not been reported, but patients taking large doses of the chelator penicillamine for cystinuria and as a second-line agent in rheumatoid arthritis, are at risk. Poisoning with copper presents with haemolysis and brain and hepatocellular damage. The requirement for adults (including during pregnancy) is 1.5 to 3 mg/day, children range 0.6 to 1.5 mg/day. An infant requires 0.4 to 0.6 mg/day.

Iodine

At wt: 127; val: 1; iso: 127; abund: 6×10^{-6} per cent.

Iodine is present in modest amounts in most food and drinking water. Seafood, milk, and meat are good sources. Iodine is required for thyroxine 3,5,3',5'-tetraiodothyronine (T4) and 3,5,3'-tri-iodothyronine (T3). Selenium and zinc are important in the conversion of T4 to the active T3, catalysed by selenium-dependent iodothyronine deiodinase. Dietary iodine from food and water is absorbed as inorganic iodide and transported to the thyroid gland. The body content of iodine is between 20 and 50 mg (160 to 400 µmole).

Goitre results from iodine deficiency and is endemic in mountainous areas. Some 800 million people are at risk of iodine deficiency, of whom 190 million may develop goitres and more than 3 million are cretinous. These populations have an iodine intake less than 25 µg/day—the required intake being 80 to 150 µg/day. Iodine replacement is essential for these populations and may be added to food, salt, or water or by the direct administration of iodine. Goitre may also arise through eating plants containing goitrogens (for example, thiocyanate in cassava, maize, bamboo shoots, etc.) Maternal iodine deficiency is associated with perinatal death, stillbirths, spontaneous abortions, endemic cretinism, and congenital abnormalities. Thyroxine is essential for brain development during the first 2 years of life. A modest increase in the incidence of hyperthyroidism occurs following the ingestion of iodized salt preparations in individuals over 40 years of age.

Thyroid hormone and urinary iodine measurements reflect iodine status. Adults require 140 µg/day, babies 40 µg/day, infants 50 µg/day, and children 50 to 140 µg/day. Intake should increase to 175 µg/day during pregnancy and 200 µg/day with lactation. Breast milk contains 44 to 93 µg/litre, an adequate iodine intake.

Magnesium

At wt: 24; val: 2; iso: 24, 25, 26; abund: 1.94 per cent

Magnesium is present in most foods, particularly chlorophyll-containing vegetables. Magnesium is absorbed from the distal intestine and excretion is through the kidneys. Magnesium homeostasis includes reabsorption of endogenous magnesium from enteric secretions. The plasma concentration varies between 0.6 and 1.0 mmol/l, and adult whole-body magnesium is 1 mole or 25 g—two-thirds in bone with phosphate and bicarbonate, the remainder being complexed with ATP. Magnesium is a cofactor for cocarboxylase and is involved in the replication and transcription of DNA and translation of RNA. Magnesium deficiency is manifested by progressive muscle weakness, failure to thrive, neuromuscular dysfunction, arrhythmias, hallucinations, positive Trousseau's sign, coma, and death. Malabsorption, alcohol abuse, and diuretics are causes of a low serum magnesium levels.

Urinary magnesium is an approximate measure of dietary intake. Typically, a diet in the United Kingdom contains between 8 and 17 mmole of magnesium/day (200 to 400 mg). Adults (including during pregnancy), require between 8 and 10.3 mmole/day (200 to 250 mg). Human breast milk contains 0.12 mmole (2.8 mg)/litre. The lactating mother should increase her magnesium intake by 2.0 mmole/day (50 mg). Babies require 30 mg/day, infants 75 mg/day, and children 80 to 200 mg/day

Manganese

At wt: 55; val: 2, 4; iso: 55; abund: 0.085 per cent.

Tea, cereals, legumes, and leafy vegetables are good sources of manganese. Manganese is a cofactor and enzyme activator. Manganese absorption is only 3 to 4 per cent efficient. Calcium, phosphorous, fibre, and phytate interact with and reduce manganese absorption. The plasma concentration is between 1 and 2 µg/g bound to transferrin, the body pool contains 0.3 mmole, and excretion is in bile. Manganese deficiency in man has not been reported. The average intake in Britain is 2 to 5 mg/day, half from tea. Safe intakes for adults are more than 1.5 mg (25 µmole)/day and for children and infants more than 16 µg (0.3 µmole)/kg per day. Breast milk contains 15 µg/litre.

Molybdenum

At wt: 96; val: 2, 3, 4; iso: 92, 94, 95, 96, 97, 98, 100; abund: 7×10^{-4} per cent

Important dietary sources are wheat flour and its germ, legumes, and meat. Molybdenum is a cofactor for oxidases important in the metabolism of DNA and sulphites. Intestinal absorption efficiency is high at 40 to 100 per cent. Plasma concentration is 1 µg/100 ml, and molybdenum is bound to protein. Storage is in the liver and excretion in urine. There are no reports of molybdenum deficiency in man. Gout has been attributed to high molybdenum intakes of 10 to 15 mg/day. Adults require between 75 and 200 µg of molybdenum/day (0.75 to 4 µmole/day). Breast-fed infants require 1.0 µg/kg per day. Babies require 15 to 30 µg/day, infants 20 to 40 µg/day, and children 40 to 100 µg/day.

Nickel

At wt: 59; val: 2, 3; iso: 58, 60, 61, 62, 64; abund: 0.018 per cent

It has not been established whether nickel is essential in humans. Absorption is 3 to 6 per cent of the dietary intake. Plasma concentrations are between 2 and 4 µg/100 ml, some of which is bound to albumin. Nickel is excreted in urine. Nickel deficiency might result in depressed growth and haemopoesis.Requirements are unknown, but intakes in the United Kingdom are about 140 µg/day (2.4 µmole/day).

Phosphorus

At wt: 31; val: 3, 5; iso: 31; abund: 0.12 per cent.

Phosphorus is present in all natural foods, the usual diet in Britain providing 1.5 g of phosphorus daily. Phosphorus is an important physiological component: with calcium, of the bony skeleton; of adenosine triphosphate (ATP) in oxidative phosphorylation; in nucleic acids through phosphorylation of nucleotides; and in enzyme control through phosphorylation by protein kinases. Phosphorus is absorbed as free inorganic phosphorus from the diet (controlled by 1,25(OH)$_2$ vitamin D) at both the brush-border and basolateral membranes. The plasma concentration of phosphorus is between 0.8 and 1.4 mmol/l and it is excreted in both urine and faeces. The bony skeleton contains 80 per cent of the body content of phosphorus as the calcium salt, 19 to 29 mmole (600 to 900 g). Recommended phosphorus requirements are equimolar to calcium.

Selenium

At wt: 79; val: 2, 4; iso: 74, 76, 77, 78, 80, 82; abund: 8×10^{-5} per cent

Selenium is found in food as selenoamino acids, selenoproteins, and as selenide, selenite, or selenate. The main sources of selenium are cereals, meat, and fish. Selenium is a cofactor of certain enzymes (for example, iodothyronine deiodinase and glutathione peroxidases). Both selenium and vitamin E are important in stabilizing lipid membranes by inhibiting oxidative damage. Absorption is efficient at 35 to 85 per cent. The plasma concentration of selenium is between 7 and 30 µg/100 ml protein bound. Excretion is in urine and possibly bile. In New Zealand, Venezuela, and China both the soil and population can be deficient in selenium. In China, selenium deficiency is associated with the childhood cardiomyopathy, Keshan's disease.

In certain areas of Tibet and Nepal, Kashin-Beck disease, a degenerative arthropathy associated with selenium deficiency, is endemic. Iodine deficiency, resulting in hypothyroidism and goitre, is an associated risk factor for Kashin-Beck disease even in individuals with normal serum selenium concentrations. The precise relationship between iodine deficiency and selenium status in areas where selenium deficiency is endemic remains to be established—especially in patients with Kashin-Beck disease.

Urinary selenium output, red-cell selenium levels, or glutathione peroxidase activity are markers of recent and medium-term dietary intake. Adults require 75 µg/day (0.9 µmole/day). Fertility requires an adequate selenium intake, but the pregnant woman has no additional dietary selenium requirements. Lactation requires an increase in dietary intake of 15 µg/day (0.2 µmole/day). Breast-fed infants should receive approximately 10 µg/day (0.1 µmole/day) and children 15 to 30 µg/day (0.2 to 0.4 µmole/day). Breast milk contains 20 to 60 µg/litre of selenium.

Silicon

At wt: 28; val: 4; iso: 28, 29, 30; abund: 25.8 per cent

Cereal grains and other sources of dietary fibre are important sources of silicon. The role of silicon in human nutrition may be important in cartilage and connective tissue as the human aorta, trachea, lungs, and tendons are rich in silicon. Silicic acid is readily absorbed. The body storage pool is approximately 3 g (1 mole) in a 60-kg man; the plasma monosilicic acid concentration is 500 µg/100 ml. The dietary requirements of silicon are unknown.

Sulphur

At wt: 32; val: 2, 4; iso: 32, 33, 34, 36; abund: 0.048 per cent

Sulphur occurs in: proteoglycans; dermatan, chondroitin and keratin sulphate; glutathione; and coenzymes including coenzyme A. Cysteine, methionine, and disulphide crosslinkage are important in proteins, and sulphate is involved in detoxification processes. Sulphur is absorbed as amino acids, which are subsequently desulphated, and excreted in urine as sulphates. Dietary intake is of the order of 0.7 mg (22 µmole)/day. The dietary requirements of sulphur are unknown.

Zinc

At wt: 65; val: 2; iso: 64, 66, 67, 68, 70; abund: 0.02 per cent

Dietary sources are meats, whole grains, legumes, and oysters. Zinc is required as a cofactor in DNA synthesis, cell division, and protein synthesis. Zinc-finger 'proteins' are important as gene transcriptional regulators. It has long been believed that zinc is important for wound healing. Approximately 20 per cent of dietary zinc is absorbed complexed with amino acids, phosphates, and organic acids. Phytates and oxalates form insoluble complexes which inhibit absorption. The normal plasma concentration of zinc is between 80 and 110 µg/100 ml, complexed with albumin. The adult body content of zinc is over 2 g (30 mmole). Bone, the prostate, semen, and the choroid of the eye all contain high concentrations of zinc. Loss of zinc from the body is in faeces.

Zinc is widely available in foods so dietary deficiency is rare, and only reported in malabsorption and in male dwarfs in Iran. The clinical signs are growth retardation, hypogonadism, bullous-pustular dermatitis, paronychia, lethargy, hepatosplenomegaly, and iron deficiency anaemia which responds to zinc supplements (15 mg three times/day). Zinc deficiency has been reported as a feature in a number of diseases, including the florid deficiency state and skin condition acrodermatitis enteropathica, an autosomal recessive trait leading to selective impairment of zinc uptake by the upper small intestinal mucosa. Similar signs may be observed in patients with severe malabsorption due to Crohn's disease and other intestinal disorders, especially those associated with a loss of inflammatory cells in the gut lumen. Excessive zinc can lead to nausea, vomiting, and fever.

The plasma concentration of unhaemolysed zinc is a measure of a person's current zinc status. Zinc in the red-blood cells and hair gives a long-term assessment of zinc status. Adults of all ages (and during pregnancy and lactation) require between 12 and 15 mg of zinc (110 to 145 µmole)/day. Infants need between 4 and 5 mg/day; however, human milk is not a rich source of zinc (2–3 mg/litre) and the infant depends very much on the stores obtained during its last 3 months of interuterine life. Children require 10 to 15 mg/day

Further reading

General

Eastwood M (1997). *Principles of human nutrition*. Aspen, Gaithersburg, MD.

Panel on Dietary Reference Values of the Committee on Medical Aspects of Food Policy (1991). *Report on Health and Social Subjects 41. Dietary reference values for food energy and nutrients for the United Kingdom*. HMSO, London.

Powers HJ (1997). Vitamin requirements for term infants: considerations for infant formulae. *Nutrition Research Reviews* **10**, 1–33.

Sadler MJ, Strain JJ, Caballero B, eds. (1999). *Encylopedia of human nutrition*. Academic Press, San Diego, CA.

Ziegler EE, Filer LJ Jr, eds. (1996). *Present knowledge in nutrition*, 7th edn. ILSI Press, Washington DC.

Carotenoids as background examples

Cox DN *et al*. (1998). Take five, a nutrition education intervention to increase fruit and vegetables intakes. *British Journal of Nutrition* **80**, 123–31.

Vershinin A (1999). Biological functions of carotenoids—diversity and evolution. *Biofactors* **10**, 99–104.

Wang XD, Russell RM (1999). Procarcinogenic and anticarcinogenic effect of b-carotene. *Nutritional Reviews* **57**, 263–72.

Vitamin C

Benzie IFF (1999). Vitamin C: prospective functional markers for defining optimal nutritional status. *Proceedings of the Nutrition Society* **58**, 469–76.

Sauberlich HE (1994). Pharmacology of vitamin C. *Annual Review of Nutrition* **14**, 371–91.

Thiamin, biotin, and pantothenic acid

Bender DA (1999). Optimum nutrition: thiamin, biotin and pantothenate. *Proceedings of the Nutrition Society* **58**, 427–33.

Folate

Butterworth CE Jr, Bendich A (1996). Folic acid and the prevention of birth defects. *Annual Review of Nutrition* **16**, 73–97.

McNulty H (1997). Folate requirements for women *Proceedings of the Nutrition Society* **56**, 291–303.

Scott JM (1999). Folate and vitamin B12. *Proceedings of the Nutrition Society* **58**, 441–8.

Selhub J (1999). Homocysteine metabolism. *Annual Review of Nutrition* **19**, 217–46.

Vitamin B_{12}

Scott JM (1999). Folate and vitamin B12. *Proceedings of the Nutrition Society* **58**, 441–8.

Vitamin A

McClaren DS (1980). *Nutritional ophthalmology*. Academic Press, London.

Semba RD (1997). Vitamin A and human immunodeficiency virus disease. *Proceedings of the Nutrition Society* **56**, 459–69.

Thurnham DI, Northrop-Clewes CA (1999). Optimal nutrition: vitamins and the carotenoids. *Proceedings of the Nutrition Society* **58**, 449–57.

Vitamin D

Nutritional aspects of bone; a symposium (1997). *Proceedings of the Nutrition Society* **56**, 903–87.

Wilkinson RJ *et al*. (2000). Influence of vitamin D deficiency and vitamin D receptor polymorphisms on tuberculosis among Gujarati Asians in west London: a case control study. *Lancet* **355**, 618–21.

Vitamin E

Gabsi S, *et al.* (2001). Effect of vitamin E supplementation in patients with ataxia with vitamin E deficiency. *European Journal of Neurology* **8**, 477–81.

Halliwell B (1996). Antioxidants in human health and disease. *Annual Review of Nutrition* **16**, 33–50.

Moreno-Reyes R, *et al.* (1998). Kashin-Beck osteoarthropathy in rural Tibet in relation to selenium status. *New England Journal of Medicine* **339**, 1112–20.

Morrisey PA, Sheehy PJA (1999). Optimal nutrition: vitamin E. *Proceedings of the Nutrition Society* **58**, 459–68.

Stevenson VL, Hardie RJ (2001). Acanthocytosis and neurological disorders. *Journal of Neurology* **248**, 87–94.

Traber MG, Sies H (1996). Vitamin E in humans: demand and delivery. *Annual Review of Nutrition* **16**, 321–47.

Trace elements

Arthur JR, Beckett GJ, Mitchell JH (1999). The interactions between selenium and iodine deficiencies in man and animals. *Nutrition Research Reviews* **12**, 55–73.

Cousins RJ (1994). Metal elements and gene expression. *Annual Review of Nutrition* **14**, 449–69.

Failla ML (1999). Considerations for determining 'optimal nutrition' for copper, zinc, manganese and molybdenum. *Proceedings of the Nutrition Society* **58**, 497–505.

Goyer RA (1997). Toxic and essential metal interactions. *Annual Review of Nutrition* **17**, 37–50.

Lukaski HC (1999). Chromium as a supplement. *Annual Review of Nutrition* **19**, 279–302.

10.4 Severe malnutrition

Alan A. Jackson

Introduction

Severe malnutrition reflects a society which is not able to meet its basic needs for health care and survival. It is the consequence of a range of factors which together characterize underdevelopment, poverty and deprivation, an insanitary environment, frequent infections, and food which is poor in quality or limited in availability. A series of vicious cycles operate within individuals and across generations, limiting the ability of vulnerable groups, families, and individuals to cope with the harsh realities of a hostile environment, either through the exigencies of nature or a human unwillingness to share the available resources with greater equity. Across the globe, severe malnutrition is a common condition during childhood and is most prevalent amongst the poorest in developing countries, but it is also found with uncomfortable frequency amongst the most deprived of every society, including those in Europe and North America. It is a frequent aspect of clinical medicine in patients who, for any reason, have a loss of appetite or a reduction in food intake. The same principles of management and care apply wherever the problem is found.

Malnutrition at any age impairs the ability to perform and function. Children with severe malnutrition are at risk of serious, life-threatening problems which require urgent attention. More insidiously, malnutrition during childhood stunts development and leaves a scar which remains for the rest of that person's life. This lost potential can express itself as an increased risk of ill health, as impaired intellectual development leading to poor school performance, or in limited physical development leading to poorer work performance. Once part of an individual's potential for development has been lost, the clinical and social implications tend to be cumulative. On a global scale, the sum total of the loss of individual capability represents a fundamental brake on aspirations for social and economic development. Throughout the world, 54 per cent of childhood deaths can be attributed directly or indirectly to malnutrition, amounting to 11.6 million 1995. This burden is not spread evenly between different parts of the world. For the worst affected countries, as many as 50 per cent of children under 5 have malnutrition which is severe enough to threaten life, most frequently in sub-Saharan Africa and Southeast Asia. This is representative of the day-to-day situation, and is not a peculiarity of special emergencies.

Notwithstanding the large number of children with severe malnutrition, over the past 20 years there has been a shift to the right of the curve for the distribution of the height and weight of children, indicating a general improvement and marked success for specific interventions. Thus, change is possible and when suitable measures are put in place sustained improvement can be achieved. However, there is absolutely no basis for complacency, as recent figures suggest a slowing down, or even a reversal, of this improvement. This may relate to an inability to control infections effectively, with tuberculosis, malaria, and diarrhoea continuing to play a major role and the HIV epidemic making a significant contribution. The world's population continues to increase, so an improvement in percentage terms does not necessarily mean a decrease in the absolute numbers of malnourished people across the globe.

Severe malnutrition is a late stage in a process where an individual has had inadequate access to sufficient energy and nutrients for a period of time. During this time the function of the body changes and there are important ways in which severely malnourished children are different from normal children. They do not respond to medical treatment in the way that would be expected, and if they are treated in the same way as normal children they will very likely die. Based upon best practice mortality would be expected to be around 5 per cent, but in many centres case mortality has remained unchanged for 50 years, around 50 per cent, with four major errors in care occurring in about 80 per cent of centres (Table 1).

The World Health Organization (**WHO**) has produced guidelines on effective treatment in 10 steps (Fig. 1). During the early period of care the order in which different aspects of treatment are carried out is critical for a successful outcome. A central feature is that as a first step the body's cellular machinery has to be repaired if function is to be restored. Silent infections are common. There have been unusual losses of nutrients from the body which cannot be corrected adequately on a standard diet. The damaged systems of the body are not able to cope with excess energy or further stress. Effective treatment requires the ordered correction of the underlying problems before any attempts are made to correct the tissue deficits.

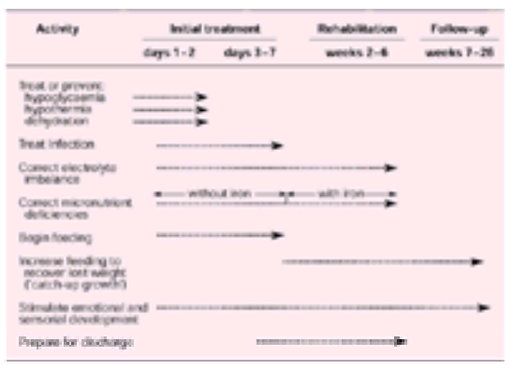

Fig. 1 WHO recommendations for the 10-step approach to the management of severe malnutrition.

Clinical syndromes

Severe malnutrition can present with an array of clinical symptoms and signs, which depend upon the duration of the illness, the extent of coinfection, the particular pattern of nutrient deficiencies and metabolic disturbances, and other associated complications such as diarrhoea and vomiting with attendant disturbances in fluid and electrolytes (Table 2). All descriptions of the condition emphasize one or other feature of the presentation. The archetypal descriptive terms for childhood malnutrition, kwashiorkor, marasmus, or marasmic kwashiorkor, were originally used to characterize clinical syndromes.

The first description of the kwashiorkor syndrome emphasized the development, location, and timing of the skin lesion, with progression from friable hyperpigmented skin, which stripped to reveal hypopigmented skin which ulcerated easily to provide a ready portal for infection—lesions distinct from pellagra. Other features such as abnormal affect and hepatomegaly were noted, but were less remarkable. Placing emphasis upon variability in clinical presentation has made comparison difficult and encouraged the idea that the underlying pathophysiology, and hence its treatment, differs in important ways between locations. This has diverted attention from

similarities in the fundamental changes which take place across the range of clinical presentations. Identifying differences in clinical presentation for this complex disorder are less important than thought previously for successful management.

The function of the body is controlled through the integration of many systems. A fault in any one has implications for the function of all the others. Thus, there is the need for adequate amounts of energy, energy-generating nutrients (carbohydrate, lipid, and protein), minerals, and a range of micronutrients for the body to function effectively in a harmonized way. Lack of any component, or an imbalance, leads to deranged handling of others. By adopting an agreed classification relevant comparisons have been drawn and it is clear that the range of clinical features represent varying manifestations of a clinical disorder with the interaction of qualitative and quantitative factors. The quantitative change results from an inadequate intake of food and leads to a wasting syndrome, classically marasmus, with the progressive loss of tissue, especially marked for subcutaneous fat and muscle. The result is a thin appearance, with pinched features, thin arms and legs, and a scaphoid abdomen. Qualitative changes are usually associated with unusual losses of nutrients from the body, for example through diarrhoea or infection, reordered metabolism to deal with metabolic stress, or the toxic effects of a range of noxious exposures. The end result of this process is likely to be the loss of cellular integrity and control, leading to oedema.

Classification

Severe malnutrition is defined as severe wasting (less than 70 per cent, or a score of less than –3 standard deviations, weight for height, compared with the National Centre for Health Statistics/World Health Organization (NCHS/WHO) reference), or the presence of oedema of both feet, or clinical signs of severe malnutrition, reflecting that there are quantitative and qualitative changes taking place (Table 3). The more severely malnourished an individual, the greater the risk of complications and the risk of an adverse outcome is related to the severity of the weight deficit or the extent to which normal function is deranged. An effective classification differentiates those at greatest risk, guides suitable interventions, and helps determine the extent to which interventions have successfully corrected the problem. These changes can all be marked either quantitatively or qualitatively.

Quantitative measures indicate the extent to which the expected pattern of growth in height and weight has not been achieved: low height for age, low weight for height, and low weight for age. Low height for age (shortness or stunting) is indicative of longer-term malnutrition or poor health. Low weight for height (thinness or wasting) implies recent or continuing current severe weight loss. Low weight for age (insufficient weight relative to age) implies stunting and/or wasting. Weight is more easily measured than height and assessing weight for age is the simplest way of excluding severe malnutrition in the absence of oedema. This has been the basis of a widely used system of classification—the Gomez classification. Weight for age is influenced by both height for age and weight for height. Where deprivation is common there is a high prevalence of low height for age, and weight for age is more strongly influenced by stunting than by wasting, and requires broader public health approaches for its alleviation being unlikely to respond in the short term to aggressive clinical intervention. The prevalence of stunting starts to increase at around 3 months of age, and the process of stunting slows down at around 3 years of age, after which mean heights run parallel to the reference. Weight for height has the advantage that it can be used when age is not known reliably and suggests recent severe weight loss, indicating those children who are most likely to benefit from immediate aggressive nutritional intervention and support. The rate at which weight improves is used to assess progress during recovery, and success of care is indicated by the achievement of a weight which is appropriate for the individual's height. The Waterlow classification usefully identified the relationship between height and weight and suitable cut-off points for intervention.

Qualitative criteria are more difficult, because of their variability and uncertainty about whether they mark any particular pathophysiological process. It has been agreed that the presence of pitting oedema is the archetype of qualitative change, identified as kwashiorkor in the Wellcome classification and now called oedematous malnutrition. In milder forms oedema might be restricted to the limbs, but embraces the entire body in more severe forms. Obtaining a reliable measure of body weight is difficult in the presence of oedema, because of the uncertain contribution of oedematous fluid. The overall appearance might be of a child who superficially appears full, but has evident wasting below the oedema when examined carefully with the clothes removed. Poor appetite or anorexia might be very common, but is not used as a diagnostic criterion. Multiple infection is common and often silent, so that specific sites of infection may be difficult to identify or localize. A high index of suspicion is required for the presence of silent infections, which should be presumed to be present. Infection is not part of the diagnostic criteria.

Natural history and clinical presentation

Inadequate nutrition slows the pace of growth and development and the greater the severity of the limitation or insult, or the longer its duration, the greater the difference between the achieved development and that expected. The stress of an insult of greater severity evokes a metabolic response which is associated with a loss of body weight and a reordering of function, so that resources and effort devoted to growth and development are diverted to maintain the integrity of the individual. The nutritional health of the infant is critically determined by how well prepared the mother was to carry the pregnancy, and the effectiveness with which breast feeding is established and maintained. During pregnancy and for the first months of life the infant is totally dependent upon the mother for its nutrient supply. During early pregnancy there is the elaboration and maturation of function in the fetus. The last trimester is of critical importance as it is when the fetus accumulates effective reserves of nutrients, helping survival and facilitating development during the first year of life. The fetus accumulates reserves of energy, as subcutaneous lipid, and of minerals and vitamins, such as iron, zinc, copper, vitamin A, riboflavin, or pyridoxine, in liver and muscle. At birth the relative protection of the intrauterine environment is replaced by the many hazards of the external world. Gastrointestinal and respiratory infection are amongst the serious dangers to survival and breast feeding provides effective protection from both. Even in affluent societies breast feeding provides the infant with a level of protection against ill health that identifies effective breast feeding as a singularly important feature in any rational policy in public health nutrition. There is a massive increase in the risk of ill health for infants who are not breast fed during the early months of life. This risk is magnified enormously for infants exposed to unsanitary environments with limited access to health care. Anything which limits the growth of the fetus, impairs its development, or causes it to be delivered early will limit its ability to cope with extrauterine life, and increase the risk of problems, infections, and malnutrition. There is enhanced mother–infant bonding and emotional development with breast feeding, and other special benefits include the remarkable bioavailability of energy and nutrients, the presence of non-nutritional factors, protective factors, and growth factors.

Screening: identification and prevention

Malnutrition is a preventable condition and the early identification of those at risk and the implementation of interventions which correct underlying problems and prevent further deterioration is central to strategies for effective care. Early growth failure can be detected by regular weighing, as an integral part of immunization and other health programmes. A series of plotted weights is most valuable and intervention is required for those whose weight crosses two growth centiles on successive measurements. If measurements are only available for a single time point, then height for age and weight for height provide an indication of any past or ongoing growth failure. Advice and demonstration of best practice in child care and feeding may be sufficient to correct a mild degree of growth failure, but persistent or more severe growth failure require closer investigation to exclude underlying problems. Poor anthropometry, with a history of poor appetite and weight loss, should always be taken very seriously and pursued until a cause has been identified and corrected. Severe malnutrition is a medical emergency.

Childhood malnutrition is a clinical problem for the individual, but is also a symptom of ineffective public health policy. Targeted interventions should address the immediate needs of the child, but should also embrace broader considerations. For the child, there is the need to effectively immunize against infection, recognize and treat infection in a timely way, and ensure an effective period of nutritional support following infection. For the family, there is the need to enhance the child rearing skills of the parents, create a stimulating environment, acquire and practice simple skills in hygiene and food preparation, and strengthen family dynamics and coping strategies. For the community, there is the need to improve the economic base of households, increase food purchasing power, increase food security or household food availability, and to treat specific nutrient deficiencies. Sound hygienic practices have to be strengthened at the group or household level, and where necessary the amount and quality of water and the safe and effective removal of solid waste improved. Each activity can exert a beneficial effect on growth and development. Any one might be relatively easy to introduce, but the real difficulty is to ensure that all are sustained. The need is for a fundamental change in the health culture and the creation of a framework of behaviour in which development activities become rooted and take place as a matter of course. A failure to establish and maintain an effective system of health care leads to a progressive deterioration in the clinical state of the most vulnerable infants leading eventually to severe malnutrition.

Aetiology and pathophysiology

Children may become malnourished simply because there is not enough food available, but sick malnourished individuals have no appetite for food. It seems paradoxical that a child who has obviously lost weight and needs to eat may refuse food even when it is readily available. If food is forced there is the possibility that the child will become worse, or even die. In managing severe malnutrition appetite is one of the most important symptoms. A loss of appetite is an important protective mechanism against consuming food which is likely to stress the systems of the body. In experimental studies there are two major biological reasons why appetite is lost: a deficiency of a specific nutrient and infection. Severe malnutrition is a disorder which results from the interaction of three distinct but related processes, each of

which appears to be related directly to the food consumed, but none of which can be easily understood simply by a consideration of food:

- reductive adaptation
- inflammatory and immune responses
- specific nutrient deficiencies.

Food helps meet the many needs for normal function, growth, and development in childhood, but also the ability to cope with environmental challenge. A diet which is adequate but marginal under normal circumstances is inadequate for the increased demands during recovery from frequent intercurrent illness with the double burden of the need for catch-up growth and to make good the unusual losses of nutrients during the infective episode itself. The time available for successful convalescence, before the next bout of infection, is too short to adequately make up the deficit.

Reductive adaptation: failure to meet the usual demands of the body for macronutrients

Reductive adaptation takes place when the demands of the body for energy and nutrients are not adequately met by the dietary intake. The general features are similar regardless of the basis for the inadequate intake. It is a general response to preserve essential function, but carries a cost. Normal metabolism takes place within a highly regulated environment, through the control and integration of exchange and turnover amongst cells and tissues. For the cellular machinery of the body to remain functionally intact and operationally effective requires a constant supply of energy and other nutrients. An estimated one-third of resting energy expenditure may be consumed through the synthesis and degradation of macromolecules such as protein, and a further third is associated with the movement of material across membranes, for example through the pumping activity of the sodium/potassium pump, Na^+/K^+-ATPase. These processes represent the internal work of the body at cellular level and underlie the functioning of all the organs and tissues. They take place continuously and the total activity can be measured as energy expenditure. As food consumption is intermittent the processes are independent of the immediate food intake. However, a sustained lack of food leads to progressive impairment of the cellular machinery as damage due to the wear and tear of normal use can no longer be replaced effectively.

Structure

When food consumption is significantly reduced, metabolic processes continue to enable the body to function, and the energy to support these processes is derived from reserves within the body. The body is in negative energy balance, and tissue mass cannot be maintained, leading to loss in weight. The losses are uneven between tissues, with major losses in subcutaneous fat and muscle and relative preservation of the metabolically more active visceral tissues. One important consequence is that heat generated by muscle is reduced and at the same time insulation in the skin is impaired leading to greater heat loss. The altered body composition underlies all anthropometric methods which are used to assess nutritional status. In addition to the changes in mass, efficiencies in the utilization of energy have to be found.

Function

Efficiencies are achieved by reducing the amount of work carried out by the body. External work is reduced by decreasing physical activity. Internal work is reduced by decreasing cellular metabolic activity, with subsequent effects upon tissue function. Significant efficiencies might be achieved for the major energy-consuming processes such as membrane pumping, protein turnover, and cellular replication. Fundamental to maintaining the chemical environment of cells is the relative distribution of potassium in the intracellular space and sodium in the extracellular space. As potassium tends to leak out of the cell and sodium tends to leak into the cell, for the cell membrane to maintain the effective partitioning of electrolytes requires that sodium is pumped out of the cell in exchange for potassium, consuming ATP. The cell membrane tends to become more 'leaky' in malnutrition as its lipid composition changes, and the Na^+/K^+-ATPase is downregulated as one way in which to reduce energy expenditure. Therefore compared with normal, all people with malnutrition have reduced intracellular potassium and increased intracellular sodium, hence decreased total body potassium and increased total body sodium, which is not identified necessarily on standard biochemical tests. The ability to maintain protein synthesis is fundamental but energetically expensive; energetic efficiency requires a reduction in protein synthesis, which is not divided equally amongst tissues. Liver normally accounts for about 25 per cent of protein synthesis, with the synthesis of nutrient transport proteins playing a critical role in the delivery of lipid, minerals, and vitamins to the other tissues. Reduced synthesis of nutrient transport proteins may save energy, but at the cost of reduced delivery to peripheral tissues; thus for example limited synthesis of apolipoproteins limits the delivery of lipid to peripheral tissues and enhances the accumulation of lipid in liver. Cellular replication is energetically demanding, requiring the ready availability of all nutrients. A reduction in cellular replication provides efficiencies in energy and nutrient use but impairs the function of systems critically dependent upon cellular replication: the skin, gastrointestinal tract, respiratory tract, and the immune system.

Functional and metabolic cost of reductive adaptation

The function of the cells in all tissues is affected by reductive adaptation. With relative protection of more vital functions, the cost is a reduction in those functions which are not immediately vital but which provide the functional reserve capability which enables the metabolic flexibility to respond to a changed internal environment or a challenge from the external environment. As a consequence, changes which would be readily managed in the normal state present a metabolic stress in the reductively adapted state. What would normally be a modest challenge can induce a major metabolic perturbation. Reductive adaptation represents the loss of reserve capacity, which leads to increased metabolic brittleness and vulnerability. The cellular machinery is no longer capable of responding effectively to the usual challenges. There is a change in the function of all systems.

Gastrointestinal tract

There is loss of mucosa and submucosal tissues, loss of gastric acidity, and a reduced capacity for digestion and absorption. This leads to impaired bioavailability of nutrients from food, decreased transit time, and predisposition to small bacterial overgrowth. An impaired ability to repair and maintain the integrity of the endothelium predisposes to bacterial translocation and overexposure to endotoxins.

Skin

The skin wastes, loses its ability to maintain heat, and readily becomes breached and infected.

Immune system

There is increased exposure to pathogens and a decreased capacity to respond (inflammation and immune response see below).

Liver

There is downregulation of synthetic and excretory processes. The reduced functional reserve makes it more difficult to maintain glucose homeostasis in the face of increased bacterial exposure. Intermediary metabolism is impaired and transport proteins for the delivery of lipid, vitamins, and minerals to other parts of the body are reduced. The formation of clotting factors is impaired. Reduced bile and bile salt formation affect digestion. Metabolism and clearance of drugs, toxins, and xenobiotics is also reduced.

Cardiovascular system

A reduction in the functional reserve of the heart, slower pulse, and increased circulation time make heart failure more likely if excess fluid is given intravenously. There is poor circulatory control with a tendency to reduced intravascular volume with an expanded interstitial fluid space.

Iron is an integral part of haemoglobin in red cells, involved in the transport of oxygen from the lungs to the tissues. The mass of red cells is related to the amount of oxygen which has to be transported, which in turn relates to the mass of active lean tissue. As part of reductive adaptation there is a decrease in the lean tissue of the body with an associated decrease in the red cell mass. The iron which is released from haemoglobin is not required immediately for the formation of more red cells. The level of iron in the body is controlled by the rate at which it is absorbed from the gastrointestinal tract, as once in the body there are no recognized mechanisms through which iron can be lost. Therefore, the iron released from red cells cannot be excreted and is placed into storage. Free iron is highly reactive and acts as a focus for uncontrolled excess generation of free radicals, thereby damaging other cellular components. Excess iron is stored in the liver, bound to ferritin. A demand

for ferritin synthesis is energetically expensive and diverts amino acids from the formation of other proteins. As part of reductive adaptation, the ability to effectively sequester iron in a chemically quiescent state is impaired.

Renal

There is decreased functional capacity of the kidney, with an impaired ability to concentrate, dilute, or acidify urine.

Muscle

Muscle mass is reduced and muscle function impaired by reduced potassium, which together lead to reduced generation of heat.

Brain

Brain function is relatively well preserved. Nevertheless there is blunting of higher functions with decreased mentation, apathy and depression, and impaired control of hormonal and integrative responses. There is a decrease in activity, poor work performance, and a decrease in discretionary activities, which together contribute to a slowing of learning.

Infection: the inflammatory and immune responses

Survival in a potentially hostile world requires effective non-specific and specific defence mechanisms. Non-specific physical barriers (skin and mucous membranes) and chemical protection (gastric acidity, secretions such as tears, mucins, etc.) depend upon cellular replication, which is less well maintained during reductive adaptation and even minor damage leads to a breach which is not repaired. Local damage with bacterial invasion usually elicits local inflammation, a systemic or acute phase response, and a specific immune response. The mounting, co-ordination and regulation of an effective response requires energy, increased cellular replication, and protein synthesis. The changes in hormones and cytokines associated with reductive adaptation impair the establishment and control of normal inflammatory and immune responses. The localized signs of tissue damage or infection—enlarged lymph nodes, enlargement of the spleen or liver, and the normal features of the acute phase response (fever, rapid pulse, and respiration)—are blunted or lost in malnutrition, making diagnosis more difficult.

Loss of appetite is a central feature of a more severe acute phase response, as the body raids its own tissues for the nutrients it requires to satisfy this unusual demand. There is a shift from the usual pattern of protein synthesis with less emphasis on growth. As muscle wastes the amino acids are made available for the synthesis of proteins for the immune system and the liver shifts from synthesizing large amounts of nutrient transport proteins to the formation of acute phase response proteins which limit cell damage and help repair. Albumin synthesis is inhibited, and it is redistributed to the third space leading to a reduced plasma albumin concentration. The low albumin which is frequently seen in malnourished people reflects the presence of an ongoing infection rather a dietary deficiency of protein. Correcting the problem requires that the underlying infection be effectively treated, not that dietary protein be increased. The cells of the inflammatory and immune systems increase their utilization of glucose, with increased gluconeogenesis from amino acids. A feature of the acute phase response is a profound change in the handling of micronutrients. There is a block in the absorption of iron. Net tissue breakdown releases components for which there is no immediate use. The circulating concentrations may be reduced (iron and zinc which are sequestered in the liver), or increased (copper), and there may be increased losses from the body in urine or stools (zinc and vitamin A). In childhood, diarrhoea is a frequent accompaniment of infection, which adds an excessive loss of nutrients from the body, especially potassium, magnesium, zinc, and vitamin A.

Specific nutrient deficiencies

Deficiency of specific nutrients is the most difficult aspect of severe malnutrition to manage effectively. Whereas in classical deficiency states, inadequate dietary intake is usually the major underlying cause, in severe malnutrition it is the failure to correct excessive losses of nutrients which leads to major imbalances. Major losses of intracellular nutrients can be difficult to identify with reliability for three reasons:

1. Losses of intracellular content may not be readily identified using standard biochemical tests on blood, for example potassium.
2. Bone acts as a very effective buffer for many nutrients and therefore severe total body depletion can develop without obvious biochemical change or loss of function, for example magnesium.
3. During an inflammatory response, redistribution of nutrients within the body makes standard tests for nutrient deficiency very difficult to interpret, for example vitamin A, zinc, or iron.

Infection causes an unbalanced loss of nutrients, which may be obvious in association with diarrhoea and vomiting, but may be more subtle as in the increased urinary losses of vitamin A and zinc which are an integral feature of the acute phase response. For an individual consuming a diet which is marginal in one or other nutrient, increased losses may make the critical difference to achieving balance, which cannot be restored unless additional nutrients are provided during the convalescent period. All cellular functions are likely to be affected to a greater or lesser degree by specific deficiencies, but one process which is of special importance is the ability to cope with 'free radicals' or oxidative induced cell damage.

Antioxidant protection

In severe malnutrition there is a major imbalance between the potential for damage induced by free radicals and protective antioxidant systems. Infection, oxidative burst, and free iron all contribute to an increased potential for damage. Mortality is greatest in those with an obvious impairment of the antioxidant defences. Children with oedematous malnutrition have severely reduced concentrations of glutathione in blood and mortality is highest in those with impaired activity of glutathione peroxidase. Although the pattern varies with location, deficiencies of micronutrients are common and result in impaired cell function and membrane damage. The many layers of antioxidant protection which are specific for each compartment of the cell provide a measure of safety. However, the system is potentially vulnerable to deficiencies or limitations in multiple micornutrients, for example niacin, folate, thiamine, riboflavin, cobalamin, ascorbic acid, carotenoids, tocopherol, selenium, zinc, copper, magnesium. A deficiency might not be readily identifiable, either clinically or biochemically, and a high index of suspicion is required.

Oedema

Oedema reflects an inability to maintain the correct distribution of fluid in the intracellular space, the vascular space, and the interstitial space, and is a final common pathway representing a loss of metabolic control. Incorrect approaches to the management of oedema—the use of diuretics or the use of high protein diets—are amongst the commonest reasons for increased mortality. The rationale behind the incorrect approach to management presumes that oedema is simply the consequence of hypoalbuminaemia, itself the result of inadequate dietary protein. There are profound perturbations of protein metabolism in kwashiorkor, but these are due to concurrent infection, loss of appetite, and increased losses of nitrogen in stools rather than a diet deficient in protein. A low plasma albumin usually indicates an acute phase response to an unrecognized infection. Treatment with a high-protein diet or infusions of albumin does not correct the oedema, but does increase mortality. A low plasma concentration of albumin might contribute to formation of oedema, but is seldom the sole or primary cause. Although diuretics exert a direct effect on cell membranes, giving a diuretic is less likely to be effective if the intravascular space is reduced. Diuretics which lead to increased urinary losses of potassium make the underlying problem of a deficiency of body potassium even worse.

The normal distribution of water between the different body compartments is tightly controlled through a number of interlinked factors. Disruption of one or more of these factors may lead to the development of oedema, and will need to be corrected for the oedema to be effectively cleared (Table 4).

Potassium deficiency leads to retention of sodium. Altered membrane structure and reduced activity of Na^{+}/K^{+}-ATPase allows intracellular potassium to fall and intracellular sodium to rise. All malnourished individuals should be presumed to be deficient in potassium and to have excess intracellular sodium, regardless of the composition of the plasma measured on routine biochemistry. Indeed plasma sodium concentrations might be low and it is tempting to give extra sodium, which is absolutely the wrong thing to do. There is more than enough sodium in the body, but it is in the wrong place. A direct approach which seeks to correct the disordered biochemistry is less likely to succeed than an approach which recognizes that the fundamental problem is that of disordered cellular function. Similar factors lead to cellular damage in any severely undernourished person, and by treating the malnutrition and repairing the metabolic machinery of the cells of the body, oedema will be effectively treated. What is required is generous supplements of potassium and correction of the underlying membrane dysfunction, which enables fluid and electrolyte balance to be restored. There is a close metabolic interdependence of potassium and magnesium, both of which are readily lost from the body in diarrhoea. It is extremely difficult to correct potassium deficiency in the presence of an associated magnesium deficiency, or to correct a magnesium deficiency in the

face of a potassium deficiency. They have to be corrected together.

Principles of care

Phases of treatment: the 10 steps (Fig. 1, Table 5)

One of the important reasons why mortality from malnutrition has not been reduced in many centres is because the primary objective of treatment has been to try and correct the obvious weight deficit. In attempting to replace the lost tissue as soon as possible, generous intakes of food have been provided, encouraged and even forced. If appetite is poor, or anorexia is a feature, then generous forcefeeding by nasogastric tube has been used. This can be very dangerous. The 10-step approach to treating malnutrition clearly identifies that treatment must be divided into different phases: the cellular machinery has to be repaired before it can be used to enable tissue growth.

Two clinical features which are directly related to specific nutrient deficiencies and are particularly difficult to manage are oedema and persistent diarrhoea. Any specific nutrient deficiency impairs cellular function and increases the risk of infection. Infection increases nutrient losses through tissue wasting as an intrinsic feature of the acute phase reaction and as vomitus or diarrhoea. Increased generation of free radicals is part of the body's attempts to deal with infecting organisms, and deficiencies of specific micronutrients directly impair the ability to cope with free radical generation. Even if an individual recovers from an infection, nutrients which have been depleted from the body are not easily replaced. This has two important effects. Firstly, the individual is deficient in a specific nutrient and carries the specific and general features of the deficiency, importantly loss of appetite. Secondly, if the deficiency is severe it may be very difficult for it to be corrected by consuming a normal diet without the addition of specific nutrient supplements. Under this circumstance, poor appetite, persistent reductive adaptation, and continued risk of further infection is maintained.

If energy is provided in excess of the requirements for maintenance, there are few ways in which it can be excreted or handled metabolically. Any significant excess is deposited as new tissue, either as cells or as cells filled with fat. There is a considerable underlying drive to form new cells, but in addition to energy this requires the availability of all the nutrients contained within the cell structure. When specific deficiencies have not been corrected individual nutrients are limiting for cell formation and it is not possible to handle the excess energy through the formation of new tissue. The excess energy creates a very serious metabolic upset (see recovery syndrome below). Therefore, during the period when nutrient deficiencies are being corrected and infections treated it is important to give sufficient energy to cover the needs of the body, but not so much that the body is forced to make new tissue. This is the basis for identifying the different phases of treatment: first to repair the machinery and gain control of metabolism by providing only enough energy to satisfy the needs for maintenance, but not enough to drive growth. Managing reductive adaptation, specific nutrient deficiencies, infection, and free radical-induced membrane and cellular damage lie at the heart of the problems associated with immediate care during the resuscitation period.

A loss of appetite is an important protective mechanism limiting food consumption which is likely to stress the systems of the body. In experimental studies there are two major biological reasons why appetite is lost: a deficiency of a specific nutrient and infection. Hence, the loss of appetite is a cardinal sign of an underlying metabolic problem which is ongoing. If the problem is identified and corrected, then appetite is restored very quickly. Most malnourished children have a profound loss of appetite due to a combination of infection and deficiencies of specific nutrients which interact to make the problem worse. Correcting the loss of appetite is central to effective care. The restoration of appetite marks the restoration of metabolic control and is a key component of therapy and a marker of progress. Once the emergency treatment required to resuscitate the child has been completed, the emphasis of care is to treat the underlying problems which are associated with a loss of appetite.

Resuscitation

Severely malnourished children present a medical emergency because of two sets of problems: the deadly triad of infection, hypothermia, and hypoglycaemia, and marked fluid and electrolyte disturbances (Table 5).

The deadly triad: hypoglycaemia, hypothermia, and infection

Brain cells are absolutely dependent upon a regular supply of glucose and oxygen to maintain the availability of ATP. Death occurs within 5 min if the supply of either is impaired, through poor circulation, reduced respiration, or a low blood glucose. The glucose required is either made in the liver or taken in the diet. Reductive adaptation limits the capacity for glucose formation and delivery and a regular dietary supply is required if blood concentrations are to be maintained. The availability of glucose for the brain can be impaired if there is competition from other tissues or functions, for example in order to maintain body temperature or to deal with infection. Malnourished individuals generate less heat and have reduced thermal insulation and therefore cool rapidly when exposed. Any attempt to generate more heat consumes glucose and other energy-providing fuels. A normal effective response to infection is a burst of activity in white blood cells which place heavy demands on available glucose, competing with the brain and leading to hypoglycaemia, and increasing the rate of heat loss leading to hypothermia. Therefore, the triad of hypoglycaemia, hypothermia, and infection indicate a very serious situation in which the body is no longer able to adequately maintain the supply of glucose to support essential functions. The treatment is to increase the supply by giving oral or intravenous glucose, reducing competing demands through decreasing the amount of heat lost, and by effectively treating infections. To deliver glucose and oxygen to the brain effectively requires an adequate circulation, which is compromised with intravascular dehydration. The correction of dehydration is closely associated with the correction of electrolyte imbalances, with energy homeostasis, and with normal cellular function. Care has to be taken to ensure that each is corrected in concert with the other to ensure that imbalances do not arise. All malnourished individuals are deficient in potassium and carry excess sodium.

Specific micronutrients: vitamin A, zinc, and iron

Iron is highly reactive chemically, and fulfils many important functions related to the generation of energy for normal cellular function. High reactivity, if not adequately controlled, carries the potential for cell damage. Red cell mass reduces in malnutrition as the lean body mass decreases. The iron is not used for further haemoglobin formation and cannot be excreted so has to be stored innocuously, as any which is unbound is liable to increase oxidative cell damage. In severe malnutrition there is increased stored iron and free iron. The available iron is not used for haemoglobin formation and giving iron supplements to treat anaemia simply adds to the load, stresses the system further, and increases mortality, especially in the presence of infection such as malaria. Initially, it is more important to repair and restore the capacity to cope with free radicals by improving vitamin and trace element status. Later, when the acute problems have been resolved, the iron will be removed from storage and used to form new tissue. As stored iron is used up, supplemental iron will have to be provided to keep pace with the rate of tissue demand.

Blindness and other eye signs of overt vitamin A deficiency are common in many parts of the world. Less obvious changes lead to impaired integrity of mucosal surfaces in the gastrointestinal and respiratory tract, lowering resistance to gastroenteritis and respiratory infections. During infection vitamin A is lost from the body, severe deficiency may develop rapidly, and the eye signs often deteriorate during early treatment. In areas where vitamin A deficiency is common, a large dose of vitamin A given very early in the treatment is an urgent necessity.

Zinc is required for the function of a wide range of enzymes, and a deficiency has widespread effects. A shortage of zinc impairs the replication of cells such as the gut mucosa, leading to further mucosal damage and increased diarrhoea. Zinc deficiency leads to diarrhoea and diarrhoea leads to zinc deficiency. Similar changes take place in damaged skin leading to ulcerated skin which is readily damaged with mild trauma.

Persistent diarrhoea

Many malnourished children have diarrhoea which can take time to settle. The diarrhoea may be infective in origin or have an infective component, due to viruses, bacteria, fungi, or helminths. However, diarrhoea which has persisted for any time will also have an element due to specific nutrient deficiencies (zinc and vitamin A) or chemical injury (bile salt deconjugation). With continued diarrhoea there are ongoing losses of nutrients. Few bacteria exist in the healthy small intestine, but small bowel overgrowth develops readily in malnutrition, due to a combination of gastric achlorhydria, reduced motility (potassium and magnesium deficiency), leading to bile salt deconjugation, damaged mucosa, and bacterial translocation. For the bowel to repair and re-establish its resistance requires adequate nutrients, most especially zinc, vitamin A, and folates. Thus the effective treatment of chronic diarrhoea requires a three-pronged approach: correction of potassium deficiency, treatment of bacterial overgrowth (with metronidazole), and effective repletion of specific nutrient deficiencies (such as zinc, vitamin A, and folate).

Management

The objectives of the resuscitation phase are to stabilize vital functions, by giving oxygen, supporting respiratory and cardiac function, and correcting fluid imbalance, to ensure that adequate amounts of glucose are delivered to the brain. Body temperature must be maintained by maintaining glucose supply to the system, limiting heat loss through the skin, and starting to control infection. As the capacity for the body to carry out metabolic functions is impaired, external support has to be supplied regularly on a 24-h cycle. The regular intake of small amounts over 24 h (especially at night) is a very effective way of achieving this (Table 5). All infections must be treated. Specific nutrient deficiencies must be corrected, but no iron or extra sodium should be provided. The metabolic state must be controlled by limiting the intake of energy and protein to that required to maintain body weight, and ensuring that there is no excess (see below). These steps will enable the repair of the metabolic machinery and allow cellular function to move towards normal. The response to a successful intervention will be a return of appetite; the patient will feel better and smile.

Recovery syndrome

Limited availability of one or more nutrients leads to competition between all cells for the little available. Some nutrients become relatively more deficient, upsetting the balanced function between tissues and the clinical signs of a deficiency become more obvious. There is a similar explanation for why the clinical signs of a deficiency are not always apparent, even though the body might be particularly deficient. During reductive adaptation, the demand for nutrients is decreased, and the signs of a deficiency are masked. Signs of deficiency become exposed in rapidly dividing tissues, when the demand for nutrients is greatest. Vitamin A and zinc are examples, but the same principles apply to many other nutrients, especially the B vitamins. The recovery or refeeding syndrome develops when individuals who have undergone reductive adaptation are suddenly provided with a relative excess of food. Excess energy drives metabolism while specific nutrient deficiencies are inadequately corrected and the metabolic machinery is still compromised. The syndrome may vary in its details, but consists of left- and right-sided heart failure associated with an overloaded circulation. This may progress to vascular collapse with abdominal distension as the circulating vascular volume is poured into the bowel as a profound secretory diarrhoea. The first sign of the onset of the recovery syndrome is an increase in pulse and respiratory rate. If food continues to be consumed at the same rate the load on the heart will progress to heart failure. This is a medical emergency, and it is vitally important that the food intake is reduced or stopped. If the changes are identified early and are relatively mild, then food intake should be reduced. If the condition has advanced and is severe then it may be necessary to stop all food for 12 to 24 h. The problem will then resolve.

Replacing lost weight

The ultimate objective of treatment is to replace the lost tissue. Cellular hypertrophy and hyperplasia are critically dependent upon and limited by the available energy and nutrients. For tissue of average composition the formation of 1 g tissue requires 20 kJ of energy. A normal 1-year-old infant gains 1 g/kg body weight/day, but for catch-up weight gain during recovery from malnutrition weight it is possible to form tissue at up to 20 g/kg/day, by consuming an additional 400 kJ/kg/day. Achieving this requires an energy-dense diet which is consumed throughout the 24 h of the day. Energy is necessary but not sufficient for new tissue formation. The nutrients needed for the formation of cell membranes and protoplasm are required in adequate amounts and suitable proportions. As the lean body mass grows it has an increased need for oxygen, and the red blood cell mass increases. Iron is taken out of storage to form new red cells, and eventually these stores are depleted with the need to add supplemental iron to the diet. There is an increased demand for amino acids to meet the needs of new tissue formation. It is safe to allow quite large intakes of protein. As the amino acids are deposited in tissue and do not accumulate in the free form there is no risk of toxicity. However, to meet the pattern of amino acids required by the body will require the endogenous biosynthesis of relatively large amounts of the 'non-essential' amino acids in the body, which in itself will require the generous availability of minerals and vitamins.

Important general aspects of care

The physical care which is provided to correct the biochemical, metabolic, and infective problems is critical for success. However, there is also the need to address the broader needs of the child for healthy development. In part this is provided by creating a warm, caring environment, in part by suitably structured activities which provide an appropriate level of stimulation to encourage brain function to recover and develop.

All aspects of care need skill and sympathy. The severely malnourished child is desperately sick and must be nursed as a critically ill child with minimum physical disturbance. With correct treatment, progress can be very rapid, and it is desirable to involve the parents and siblings, to encourage and demonstrate preferred childcare practices. This will facilitate the transfer between hospital and home, and make it more likely that the practices become embedded. Less seriously ill children can be effectively managed as outpatients, using the same principles and approach to the management decisions.

Further reading

Calder PC, Jackson AA (2000). Undernutrition, infection and immune function. *Nutrition Research Review* **13**, 3–29.

De Onis M, Blossner M (1997). *WHO global database on child growth and malnutrition*. WHO, Geneva.

Khanum S, Ashworth A, Huttly SRA (1994). Controlled trial of three approaches to the treatment of severe malnutrition. *The Lancet* **344**, 1728–32.

Scholfield C, Ashworth A (1996). Why have mortality rates for severe malnutrition remained so high? *Bulletin of the World Health Organization* **74**, 223–9.

Waterlow JC (1992). *Protein-energy malnutrition*. Edward Arnold, London.

WHO (1995). Physical status: the use and interpretation of anthropometry. *Report of a WHO Expert Committee*, WHO Technical Report Series no.854. WHO, Geneva.

WHO (1999). *Management of severe malnutrition: a manual for physicians and senior health workers*. WHO, Geneva.

WHO/UNICEF (2000). *Management of the child with a serious infection or severe malnutrition: guidelines for care at the first-referral level in developing countries*. WHO, Geneva.

10.5 Obesity

Peter G. Kopelman and Stephen O'Rahilly

Historical introduction

Obesity is now so common that it may replace undernutrition and infectious disease as one of the most significant contributors to global ill health. Modern research shows that obesity represents an important and defined medical disorder. Investigations into the genetic contribution to weight gain and the intra-abdominal distribution of fat (central obesity) have identified certain ethnic groups and susceptible families who are specifically at risk. There is, moreover, increasing awareness that overweight and obesity are key factors in the development of other chronic diseases, in particular type 2 diabetes and coronary heart disease, to which they contribute to the high mortality. Obesity can no longer be dismissed as a cosmetic problem affecting certain individuals: it is, in effect, an epidemic that requires effective measures for its prevention and management.

Definition of overweight and obesity

In clinical practice body fat is most commonly and simply estimated by using a formula that combines weight and height. The underlying assumption is that most variation in weight for persons of the same height is due to fat mass. The formula most frequently used in epidemiological studies is the body mass index (**BMI**), which is the weight in kilograms divided by the square of the height in metres. BMI is strongly correlated with densitometric measurements of fat mass adjusted for height in middle-aged adults. The main limitation of the BMI is that it does not distinguish fat mass from lean mass.

Measurements of body circumference are important because excess visceral (intra-abdominal) fat is itself a potential risk factor for chronic diseases, independent of total adiposity. Waist circumference and the ratio of waist circumference to hip circumference are practical measures for assessing upper body fat distribution, although neither provides a precise estimate of visceral fat. Measurement of skinfold thickness with callipers provides a more precise assessment of body fat, especially if taken at multiple sites. Skinfolds are useful in the estimation of fatness in children, for whom standards have been published. However, these measurements are more difficult to make in adults (particularly in the very obese), are subject to considerable variation between observers, require accurate callipers, and do not provide any information on abdominal and intramuscular fat. In general they are not superior to simpler measures of height and weight.

Measurement of bioimpedance is based on the principle that lean mass conducts current better than fat mass because it is primarily an electrolyte solution. A measurement of the resistance to a weak current (impedance) applied across the extremities provides an estimate of body fat when combined with height and weight and an empirically derived equation. Although impedance devices are simple and practical to use, they neither measure fat nor predict biological outcomes more accurately than the simpler anthropometric measurements. Table 1 lists the methods used to characterize obesity.

Defining a 'healthy weight' for a particular society is not straightforward. There are methodological difficulties that derive from a definition based on total mortality rates. People frequently lose weight as a consequence of illness, unrecognized at the time of survey, that is ultimately fatal. This gives an appearance of a higher mortality among those with lower weights (reverse causation). The effect can be minimized by either excluding persons with diagnoses that might affect weight and/or those who report recent weight loss, or excluding those who die during the first years of follow-up. A second concern is the confounding factors that may distort the association between body weight and mortality: cigarette smoking is of particular importance. The Nurses Health Study, which prospectively studied 116 000 women in the United States during a 17-year period, shows a U-shaped relationship between mortality and BMI in an overall age-adjusted analysis. However, the relationship becomes a simple positive association when reverse causation is accounted for and the analysis is limited to those who have never smoked.

There is a close relationship between BMI and the incidence of several chronic conditions caused by excess fat (Fig. 1)—type 2 diabetes, hypertension, coronary heart disease, and cholelithiasis. This relationship is approximately linear for a range of BMIs less than 30: American women with a BMI of 26 have a twofold increased risk of coronary heart disease compared with women with a BMI of less than 21, and an eightfold increased risk of developing type 2 diabetes. The equivalent figures for American men are a 1.5-fold increase and a fourfold increase. The risk of hypertension is doubled in both men and women with a BMI of 26. All risks are greatly increased for those subjects with a BMI greater than 29, independent of gender.

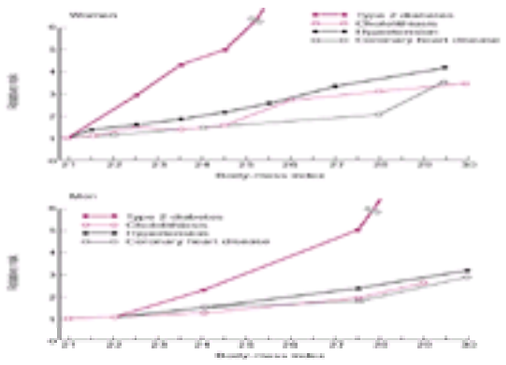

Fig. 1 Relation between body mass index (BMI) up to 30 and the relative risk of type 2 diabetes, hypertension, coronary heart disease, and cholelithiasis. Panel A shows these relations for women, initially 30 to 55 years of age, who were followed up for 18 years. Panel B shows the same relations for men, initially 40 to 65 years of age, who were followed up for 10 years. (From Willett WC, Dietz WH, Colditz GA (1999) *New England Journal of Medicine* **341**, 427–33, with permission of the authors and the Editor.)

Body mass index can be used to estimate the prevalence of obesity within a population and the associated risks. It does not, however, account for the wide variation in the nature of obesity between different individuals and populations. A World Health Organization Expert Committee has proposed a classification of overweight and obesity that applies to both men and women and to all adult age groups (Table 2).

Waist circumference correlates with measures of risk for coronary heart disease such as hypertension or blood lipid concentrations. The choice of cut-off points on the waist circumference continuum involves a reciprocity between sensitivity and specificity similar to that for BMI. An expert panel has suggested increased risks if the waist circumference is more than 102 cm for men and more than 89 cm in women. However, lower cut-off points are associated with a two- to threefold increase in the relative risk of type 2 diabetes. Gender-specific cut-off points for waist circumference may guide the interpretation of values for adults: proposed cut-off levels are shown in Table 3. Level 1 is designed to alert clinicians to potential risk while level 2 should initiate therapeutic action. It must be emphasized that these figures reflect knowledge acquired largely from epidemiological studies in developed countries. Preliminary information from developing nations indicates that lower cut-off levels for both BMI and waist circumference are appropriate for certain populations, who are at particular risk from comparatively modest degrees of overweight.

Epidemiology of overweight and obesity

The value of estimating prevalence and secular trends in overweight and obesity for identifying those at risk cannot be overemphasized. Conservative estimates of the economic costs of obesity in developed countries are between 2 and 7 per cent of the total health costs: this confirms that obesity represents a significant portion of expenditure in national health care budgets.

The range of BMI varies significantly according to the stage of economic transition and associated industrialization of a country: the shift from dietary deficit to one of dietary excess. As the proportion of the population with a low BMI decreases, there is an almost reciprocal increase in the proportion of the population with a BMI above 25. This indicates the tendency for a population-wide shift as socio-economic conditions improve so that overweight replaces thinness. Importantly, changes in adult prevalence of overweight and obesity are reflected by an even more striking changes in childhood and adolescence in both industrialized and developing countries. The early onset of obesity leads to an increased likelihood of obesity in later life, as well as an increased prevalence of obesity-related disorders.

Obesity, defined as a BMI of more than 30, is a common condition in Europe and the United States. The most comprehensive information in Europe comes from the data collected between 1983 and 1986 for the MONICA study. On average 15 per cent of men and 22 per cent of women were found to be obese, with overweight being much more common among women than men. More than half the adult population between 35 and 65 years of age in Europe were either overweight or obese. In England and Wales the most recent health survey has confirmed an increase in the prevalence of obesity in adults from 6 per cent of men and 8 per cent of women in 1980 to 17 per cent of men and 21 per cent of women in 1998. National surveys in the United States have shown a marked increase in the prevalence of obesity over time. The striking increase in prevalence between 1980 and 1994 confirms that population-wide increases in overweight and obesity may occur over short periods. The most recent data from the United States, derived from the third NHANES (1988 to 1994), shows about 20 per cent of American men and about 25 per cent of American women to be obese. Subanalysis shows African-American women and other minority populations to have particularly high rates of obesity. Obesity is likewise prevalent in Latin America and is a particular problem in the Caribbean.

The increasing prevalence of obesity is not confined to Europe and the Americas. In Southeast Asia a dramatic rise is being seen in all populations. Obesity is now more prevalent in Malaysia than undernutrition in both urban and rural communities, but the most striking figures come from the Pacific region. In urban Samoa the prevalence of obesity is estimated as greater than 75 per cent of adult women and 60 per cent of adult men. High prevalence rates are also observed in the Middle East.

Aetiology of human obesity

At its simplest level the 'cause' of obesity is already known. By definition it results from an imbalance between energy intake and energy expenditure. In any individual, excessive caloric intake or low energy expenditure, or both, may explain the development of obesity. A third factor, so-called 'nutrient partitioning', a term reflecting an individual's propensity to store excess energy as fat rather than lean tissue, may contribute but will not be discussed further here. Multiple studies over several decades have attempted to clarify the relative roles of energy intake, energy expenditure, and nutrient partitioning in human obesity. The lack of consensus, despite intensive efforts, reflects a number of problems inherent in the study of the obese state. These include the heterogeneity of human obesity and errors in measurements of energy intake and expenditure, combined with the minor degrees of energy imbalance required to accumulate fat tissue over time, and the confounding effect of obesity *per se* on the measurement of physiological parameters.

Role of altered energy intake in the pathogenesis of obesity

It is surprising that no direct correlation has been reported between increasing prevalence of obesity and increased energy intake in developed nations given the ready availability of highly palatable foods. Underreporting of food intake confounds the understanding of the role of energy intake in the aetiology of obesity. Underreporting is an almost invariant feature of obesity, with comparisons of energy intake and expenditure in obese subject showing a consistent shortfall in self-reported food intake of approximately 30 per cent of the energy requirements. It is thus extremely difficult to obtain reliable data on the energy intake of free-living obese subjects.

Role of altered energy expenditure in the development and maintenance of obesity

The principal components of energy expenditure are resting metabolic rate, which represents the energy cost of maintaining physiological homeostasis, and physical activity, being the most variable component and representing 20 to 50 per cent of total energy expenditure. Resting energy expenditure is readily measured by indirect calorimetry, and the use of water labelled with two stable isotopes allows accurate measurements of total energy expenditure over a 10- to 20-day period. Studies of age- and sex-matched pairs of lean and obese women following an imposed activity schedule clearly demonstrate consistently higher energy expenditure in obese subjects compared with their lean pairs. The measurement of energy expenditure within the home, using doubly labelled water, shows comparable values between obese and lean subjects when corrected for different body sizes. Thus, a defect in metabolic mechanisms controlling energy expenditure has not been identified in human obesity. However, studying subjects when they are already obese is always confounded by the energy cost of the increased fat and associated lean mass, and corrections made for these may be insufficiently precise.

Prospective studies of energy expenditure in groups of individuals are more informative. Longitudinal studies of Pima Indians in Arizona suggest that the risk of a weight gain of 10 kg during a 4-year follow up is sevenfold higher in those in the lowest tertile of relative resting metabolic rate compared with those in the highest tertile. Nevertheless, even in a population which is predisposed to obesity, resting metabolic rate only predicts 40 per cent of the weight gain. No association has been observed between resting metabolic rate and 10-year weight gain in a Dutch population, and results from other studies have also questioned the validity of such

an association.

In developed countries there is a relationship between low levels of physical activity and obesity. A longitudinal Finnish study found that those reporting physical exercise three or more times each week had on average lost weight since a preceding survey. By contrast, those who undertook little physical activity gained weight and had twice the risk of gaining 5 kg or more. Among children in the United States, the relative risk of obesity is 5.3 times greater for children who watch television for five hours or more each day compared with those children who watch for less than two hours, even after correcting for a wide range of socio-economic variables.

In the United Kingdom, a study combining data on energy intake and physical activity in relation to the secular increase in adult obesity showed no relationship between total energy intake or fat consumption and the prevalence of obesity, but a close relationship between proxy measures of physical activity (television viewing and car ownership).

Environmental and cultural factors in obesity

The evidence for the critical role of environmental factors in the development of obesity comes from migrant studies and the 'Westernization' of diet and lifestyles in developing countries. The dramatic increase in age-standardized prevalence of obesity (> 60 per cent in men and women) in the Naurians in Micronesia and Polynesians in Western Samoa is closely paralleled by alterations in diet and lifestyle. A marked change in BMI is frequently witnessed in migrant studies, where populations with a common genetic heritage live under new and different environmental circumstances. Pima Indians living in the United States are on average 25 kg heavier than Pima Indians living in Mexico.

In men and women the prevalence of overweight and obesity increases with age until 50 to 60 years; it is particularly apparent between the ages of 20 and 40 years. There are large, usually unexplained, variations between ethnic groups—this is particularly apparent in women in the United States with the rapidity of change occurring with increasing affluence of particular groups (22 per cent of Cauasian women are obese, 30 per cent of African-American women, and 34 per cent of Mexican-American women). In industrialized countries, a higher prevalence of overweight and obesity is observed in those with lower educational attainments and low income, although the reverse may be seen in developing countries. There is a tendency for overweight to increase after marriage and with increasing parity. Dietary intake and physical activity are crucially important factors in increasingly affluent societies.

The analysis of the prevalence of obesity by socio-economic status in England and Wales demonstrates a strong social class gradient, especially in women, ranging from 10.7 per cent in social class 1 (high) to 25 per cent in social class V (low). Interestingly, this is accompanied by marked differences in measures of physical activity, with social classes IV and V spending significantly more time watching television and being more likely to define themselves as inactive than those in social class 1.

Genetic factors in obesity

Evidence from twin, adoption, and family studies shows unequivocally that inherited factors contribute importantly to interindividual differences in fat mass. Studies of identical twins raised separately are particularly striking; they demonstrate a much closer correlation of BMI with biological rather than adoptive family members. The identification of genetic variants influencing human fat mass is a cherished goal of obesity research; it is only recently that the first mutations causing human obesity have been found. In general, mutations have been found in those rare children with extreme obesity and clear evidence for monogenic inheritance. The discovery of these defects has been dependent on rapid advances in mouse genetics. The seminal murine discovery is that of the novel adipocyte hormone leptin, which is deficient in the obese ob/ob mouse. Administration of recombinant leptin by injection restores ob/ob mice to normal body weight and corrects the associated metabolic and reproductive malfunction.

The discovery of leptin and the identification of its receptor, which is highly expressed in the hypothalamus, has advanced the understanding of molecular mechanisms within the hypothalamus that regulate appetite and energy expenditure. A particularly important element of this system is the group of neurones in the arcuate nucleus expressing pro-opiomelanocortin and the melanocortin 4 receptor, to which peptides derived from pro-opiomelanocortin bind. Thus far, five different genetic defects causing monogenic human obesity have been identified. These include mutations in leptin and the leptin receptor and two defects involving the pro-opiomelanocortin system, mutant pro-opiomelanocortin and mutant melanocortin 4 receptor. A single obese subject with mutations in a prohormone convertase enzyme shows major defects in pro-opiomelanocortin processing that may account for the obese phenotype (Table 4). These discoveries have shown for the first time that rare forms of severe human obesity may occur as a result of unitary molecular lesions within the appetite control pathway making such cases largely resistant to voluntary or imposed dietary restrictions.

Genetic studies in the more common forms of obesity have yet to have the same tangible success that has been seen with monogenic subtypes. Nevertheless, rapid progress has been made in the identification of chromosomal loci containing genes conferring susceptibilty to obesity. In particular, a region of chromosome 2 has been reported in independent studies of different racial groups to influence obesity-related phenotypes. It is of note that this region contains the pro-opiomelanocortin gene. Given the speed of progress of the Human Genome Project, it may not be long before this locus and other variants responsible for obesity syndromes are identified.

Gene/environment interaction

Obesity represents a heterogeneous group of conditions with multiple causes. Body weight is determined by an interaction between genetic, environmental, and psychosocial factors acting through the physiological mediators of energy intake and expenditure. Fatness runs in families, but the influence of the genotype on the aetiology of obesity may be attenuated or exacerbated by non-genetic factors. The genetic influences appear to operate through susceptibility genes. Such genes increase the risk of developing a characteristic but are not essential for its expression or, by themselves, sufficient to explain the development of a disease The susceptible gene hypothesis is supported by findings from twin studies in which pairs of twins were exposed to periods of positive and negative energy balance. The differences in the rate of weight gain, the proportion of weight gained, and the site of fat deposition showed greater similarity within pairs than between pairs. This suggests that differences in genetic susceptibility within a population determine those who are most likely to become obese in any given set of environmental circumstances. Implicit to the susceptible gene hypothesis is the role of environmental factors that unmask latent tendencies to develop obesity.

Pathophysiology of obesity

Obesity causes or exacerbates many disorders, both independently and in association with other diseases. In particular, it is associated with the development of diabetes mellitus, coronary heart disease, an increased incidence of certain forms of cancer, obstructive sleep apnoea, and osteoarthritis of large and small joints. The Build and Blood Pressure Study has shown that the adverse effects of excess weight tend to be delayed, sometimes for 10 years or longer. Life insurance data and epidemiological studies confirm that increasing degrees of overweight and obesity are important predictors of decreased longevity. In the Framingham Heart Study, the risk of death within 26 years increased by 1 per cent for each extra pound (0.45 kg) increase in weight between the ages of 30 and 42 years, and by 2 per cent between the ages of 50 and 62 years. Despite this evidence many clinicians consider obesity to be a self-inflicted condition of little medical significance.

Obesity and type 2 diabetes mellitus

Obesity is accompanied by an elevated fasting level of plasma insulin and an exaggerated insulin response to an oral glucose load. Overall fatness and the distribution of body fat influence the metabolism of glucose through independent but additive mechanisms. Increasing upper body obesity is accompanied by a progressive increase in the response of glucose and insulin to an oral glucose challenge, and there is a positive correlation between increasing upper body obesity and resistance to the effects of insulin. Posthepatic insulin delivery is increased in upper body obesity, leading to more marked peripheral insulin concentrations which, in turn, lead to peripheral resistance to insulin.

Differences in the ability of insulin to suppress lipolysis, and of catecholamines to stimulate lipolysis, also vary according to fat distribution. These factors contribute to an exaggerated release of free fatty acids from abdominal adipocytes into the portal system. Free fatty acids have a deleterious effect on uptake of insulin by the liver and contribute to the increased hepatic gluconeogenesis and hepatic glucose release observed in upper body obesity. Insensitivity to insulin is not confined to adipocytes, and is accentuated by the resistance of skeletal muscle to insulin.

The elevation in plasma free fatty acids, particularly postprandially when their concentration is usually suppressed by insulin, leads to an inappropriate maintenance of glucose production and an impairment in the use of glucose by the liver (impaired glucose tolerance). Reduced hepatic clearance of insulin leads to increased

peripheral (systemic) insulin concentrations and to a further downregulation of insulin receptors.

In the initial phases of this process, the pancreas can respond by maintaining a state of compensatory hyperinsulinaemia. With ever increasing plasma concentrations of free fatty acids, the insulin-resistant individual cannot maintain this state of compensatory hyperinsulinaemia, and hyperglycaemia prevails. Hyperinsulinaemia and insulin resistance contribute to the characteristic alterations in the profile of plasma lipids which are associated with obesity: elevated fasting concentration of triglycerides in the plasma, reduced high-density lipoprotein (**HDL**) cholesterol, marginal elevations in the concentration of cholesterol and low-density lipoprotein (**LDL**) cholesterol, and an increase in the number of ApoB-carrying lipoproteins.

Prospective population studies confirm a close association between increasing body fatness and type 2 diabetes. In the Nurses Cohort Study, BMI was the dominant predictor of the risk of diabetes after adjustment for age. In this study the risk of diabetes was increased fivefold for those with a BMI of 25, 28-fold for those with a BMI of 30 and 93-fold for those with a BMI of 35 or greater. Women who gained 8 to 10.9 kg in weight during the period of study had a 2.7-fold increased risk of diabetes compared with women of stable weight. Similarly, the risk of diabetes in men increases for all BMI levels of 24 or above. The distribution of fat tissue is also independently associated with diabetes: a waist circumference of more than 102 cm (40 in) increases the risk of diabetes 3.5-fold even after controlling for the BMI.

Fetal nutrition

Recent evidence suggests that undernutrition of the fetus during intrauterine development determines the later onset of obesity, hypertension, and type 2 diabetes, independent of genetic factors. Such a phenomenon suggests the possibility of long-term programming of genetic expression as a consequence of altered intrauterine growth: Barker has proposed that an adverse nutritional environment *in utero* causes defects in the development of body organs, leading to a 'programmed' susceptibility that interacts with later diet and environmental stresses to cause overt disease many decades later. In support of this hypothesis is the finding of an inverse relationship between birthweight and systolic blood pressure and type 2 diabetes in both men and women in later life, with the highest mean systolic blood pressures and blood glucose concentrations being observed in those with the lowest birthweight and highest current weight.

Cardiovascular function in obesity

The effects of increased body fatness on cardiovascular function are predictable. Total body oxygen consumption is increased due to an increase in lean tissue mass as well as the oxidative demands of metabolically active adipose tissue, and this is accompanied by an absolute increase in cardiac output. However, the values are within the normal range when they are normalized to body surface area. The total blood volume in obesity is increased in proportion to body weight, such that obesity can be regarded as a state of expanded volume. This increase in blood volume contributes to an increase in the left ventricular preload and an increase in resting cardiac output. The increased demand for cardiac output is achieved by an increase in stroke volume while the heart rate remains comparatively unchanged. The obesity-related increase in stroke volume results from an increase in diastolic filling of the left ventricle. The volume expansion and increase in cardiac output lead to structural changes in the heart. The increase in left ventricular filling results in an increase in the size of the left ventricular cavity and an increase in wall stress. As left ventricular dilatation is accompanied by myocardial hypertrophy, the ratio between the radius of the ventricular cavity and wall thickness is preserved. This thickening of the wall with dilatation results in eccentric hypertrophy. The mass of the left ventricle increases directly in proportion to BMI or the degree of overweight. The blood pressure is a function of cardiac output and the vascular resistance against which the blood is pumped—systemic vascular resistance. An elevated cardiac output is common with moderate obesity, but not all obese patients are hypertensive. However, in those subjects where systemic resistance is increased, the combination of hypertension and obesity results in an increase in the dimensions of the ventricular wall disproportionate to the chamber radius, which eventually leads to concentric hypertrophy.

The cardiovascular adaptation to the increased intravascular volume of obesity may not completely restore normal haemodynamic function. Marked systolic dysfunction occurs when the ventricle can no longer adapt to volume overload. Dilatation of the left ventricular cavity radius reduces ventricular contractility. Despite an elevation in cardiac output, obese individuals have been shown to have depressed myocardial contractility proportional to excess weight. With left ventricular hypertrophy, reduced ventricular compliance alters the ability of the chamber to accommodate an increased volume during diastole and this results in diastolic dysfunction. A combination of systolic and diastolic dysfunction progresses to clinically significant heart failure. Body weight, independent of several traditional risk factors, was directly related to the development of congestive cardiac failure in the Framingham Heart Study.

In addition to congestive cardiac failure, the presence of hypertrophy of the left ventricle has been associated with a greater risk of morbidity and mortality from coronary heart disease and sudden death as well as arrhythmia. In the Framingham Heart Study, the 26-year incidence of coronary heart disease in women and men was related proportionately to excess weight. The incidence of coronary heart disease increased by a factor of 2.4 in obese women and a factor of two in obese men under the age of 50 years. The independent risk of coronary heart disease attributed to obesity in multivariate analysis may reflect other important mediators such as upper body fat, altered blood flow and haemostasis, hyperinsulinaemia, or sleep apnoea.

Sleep-breathing abnormalities in obesity

An increased amount of fat in the chest wall and abdomen has a predictable effect on the mechanical properties of the chest and the diaphragm and leads to an alteration in respiratory excursions during inspiration and expiration, reduced lung volume, and mismatched regional ventilation. The increased mass of fat additionally decreases compliance of the respiratory system as a whole. All of these changes are significantly exaggerated when an obese person lies flat. The mass loading effect of fat requires the respiratory muscles to exert increased force to overcome the excessive elastic recoil and there is an associated increase in the elastic work of breathing. The changes in respiratory function related to obesity are most important during sleep.

During rapid eye movement sleep, there are decreases in voluntary muscle tone with reduced arterial oxygen saturation and a rise in carbon dioxide in all individuals, but these changes are especially marked in obese subjects. Irregular respiration and occasional apnoeic episodes often occur in lean people during rapid eye movement sleep but obesity, with its influence on respiratory mechanics, increases their frequency and may result in severe hypoxia and cardiac arrhythmias. Studies of obese men and women have demonstrated that the obstruction occurs in the larynx and is associated with loss of tone of the pharyngeal and glossal muscles, in particular the genioglossus muscle. Relaxation of the genioglossus allows the base of the tongue to fall back against the posterior pharyngeal wall, occluding the pharynx. This results in a temporary cessation of breathing (apnoea) and transient hypoxia. It is not uncommon to observe saturation values as low as 6.5 kPa during rapid eye movement sleep in some obese subjects while their awake arterial blood gases are normal.

A few obese patients suffer a marked depression in both hypercapnic and hypoxic respiratory drives accompanied by abnormal and irregular patterns of breathing during sleep and (eventually) in the waking state. Characteristically, such individuals show frequent and prolonged episodes of sleep apnoea—sleep is disturbed with frequent awakening related to the resumption of breathing following an apnoeic episode. Daytime somnolence soon intervenes accompanied by persistent hypoxia/hypercapnia, pulmonary hypertension (superimposed upon an increased circulatory volume), and right-sided cardiac failure. Such changes constitute the clinical manifestation of the obesity-hypoventilation syndrome (formerly known as the Pickwickian syndrome).

In the Swedish Obese Subjects study (SOS), which examined 3034 subjects with a BMI over 35, over half of men and one-third of women reported snoring and apnoea. In contrast, 15.5 per cent of Swedish men of comparable age were self-reported habitual snorers.

An increased risk of myocardial infarction and stroke has been reported in sleep apnoea. Snoring is a strong risk factor for sleep-related strokes, while symptoms of sleep apnoea increase the risk for cerebral infarction.

Other complications associated with obesity

Gallbladder disease is the most common digestive disease in obese individuals; it has a progressive and linear risk from a BMI of 20 upwards. Liver abnormalities are described in obesity mainly due to fatty infiltration but, on occasions, associated with fibrosis and/or cirrhosis. Certain forms of cancer are more common in obese subjects: colorectal and prostate in obese men, carcinoma of the gallbladder, breast, and endometrium in obese women. Osteoarthritis frequently accompanies obesity, while bone density tends to be increased in obese subjects. Obesity in women is also associated with menstrual irregularity and infertility; obesity may be, but is not always, associated with the polycystic ovary syndrome. Table 5 lists the morbidity associated with increasing body weight.

Clinical assessment

Clinical setting

The usual principles for a medical consultation are applicable to the assessment of an overweight patient. The consultation room should preferably be properly equipped with larger than average chairs, access for wheelchairs for patients with mobility problems and medical equipment of appropriate size (examination couch, blood pressure cuff, weighing scales, stadiometer, and tape measure).

Historical background

Table 6 outlines the areas of medical history that should be investigated. The history of weight gain should be described in detail to identify possible causes and to assess the patient's insight and understanding of the factors causing weight gain. It is also useful to distinguish obesity which began in childhood from that occurring later in life either in relation to specific physiological 'critical periods' or illness. A number of syndromes are associated with the onset of obesity in childhood, but the longevity of the history and the associated clinical features generally make such cases obvious (Table 4). Disease involving the hypothalamus can often be distinguished from 'spontaneous' or 'simple' obesity by a shorter duration of weight gain and specific symptoms related to associated endocrine disturbances. The identification of the single gene disorders involving leptin and its signalling pathways are somewhat more difficult to distinguish from simple obesity, but extreme weight gain from early childhood, a positive family history, and the associated clinical features described in Table 4 are all characteristic. The most common single gene disorder causing obesity, melanocortin 4 receptor deficiency, is problematic as there are no pathognomonic features, but the diagnosis should be considered in cases of early onset familial obesity, usually with a clear dominant inheritance.

The measurement of serum leptin is not recommended as a routine examination, but in cases of severe early onset obesity this should be undertaken, since, although it is rare, congenital leptin deficiency is a potentially treatable disorder.

Clinical examination

An outline of a scheme for clinical examination is given in Table 7. Height should be measured accurately using a stadiometer and weight measured by accurate scales calibrated against known weights. Fat distribution is assessed by measurement of the waist circumference and is used to refine an assessment of risk for patients with a BMI of 25 to 34.9. Waist circumference is taken as the midpoint between the lower rib margin and the iliac crest. An examination of the skin is important: thin, atrophic skin is a feature of excess corticosteroids; acanthosis nigricans (pigmented 'velvety' skin creases, especially in the axillae) suggests insulin resistance; severe hirsutism in women may indicate polycystic ovary syndrome. A neck circumference of more than 43 cm (17 in) indicates a likelihood of obstructive sleep apnoea, while abnormal external gonadal status accompanied by intellectual impairment may suggest a rare genetic syndrome.

Assessment of risk

An assessment of an obese patient's absolute risk status requires an assessment of associated disease conditions (established coronary heart disease, other atherosclerotic diseases, type 2 diabetes, and sleep apnoea), other obesity-associated diseases such as gynaecological abnormalities, osteoarthritis, gallstones, and stress incontinence, and cardiovascular risk factors. These will include cigarette smoking, hypertension, high-risk LDL cholesterol (> 4 mmol/litre), low HDL cholesterol (< 1 mmol/litre), impaired fasting blood glucose levels, and a family history of premature coronary heart disease. Patients can be classified as being of high absolute risk if they have three of these risk factors and will usually require therapeutic intervention.

In the obese patient who smokes, cessation of smoking is a major goal of risk management. An obstacle to cessation of smoking is the attendant weight gain. The weight gained on stopping smoking is less likely to impair health than is continued smoking. For this reason, cessation of smoking should be advocated at the same time as measures to prevent weight gain.

Assessment of motivation to lose weight

Not all patients are prepared for weight reduction despite a referral to a medical practitioner. It is often useful to confirm that a patient understands the need for weight loss and is prepared to follow medical advice to achieve and maintain an agreed weight goal.

Treatment

The recommendation to treat overweight and obesity is based on evidence that relates obesity to increased mortality and the results from randomized controlled trials which demonstrate that weight loss reduces the risk of disease. Professional, governmental, and other bodies have drawn up guidelines for obesity management. These strategies for providing care to the obese patient provide useful and evidence-based guidance for clinical management.

Aims for a weight loss programme

Any treatment programme for overweight and obese patients should place equal importance on the problem of weight reduction and the maintenance of the lowered weight. Obesity may not respond to conventional methods of treatment such as a low-calorie diet: its management frequently requires an individually tailored approach.

Goals of weight loss

The success or failure of a treatment programme may be judged by an arbitrarily chosen target weight or percentage weight loss. After an initial period of relatively rapid reduction of weight, an average continuing loss of anything up to 1 kg per week should be considered acceptable. Assessment of success must take account of the age of the patient, the initial degree of obesity, the presence of indicators of associated risk or complications, and previous attempts at weight control. Weight loss goals for overweight and obese patients should be tailored to the individual. A weight loss of 5 per cent of the initial body weight will result in some improvement, while a loss of 10 per cent is of major benefit with clinically useful changes such as a lowered blood pressure, reduction in levels of plasma total cholesterol and triglycerides, an increase in HDL cholesterol, and a significant improvement in diabetic control (Table 8). The primary goal of treatment is a 10 per cent reduction from the initial weight; successful weight loss should be regarded as a loss of more than 5 per cent of the initial weight with the consequent amelioration of risk factors; very successful weight loss would be a loss of more than 20 per cent in obese patients. Weight loss should be approached incrementally with new goals for weight loss negotiated with the patient once the original target has been achieved. Goals for older patients (more than 65 years) will be different from those for younger patients—data suggest that a population becomes heavier with age whereas the risk from obesity does not increase proportionately. In some patients, particularly older patients, prevention of further weight gain may be more appropriate than actual weight loss.

Dietary treatment of obesity

Control of diet is the focus of management of overweight and obese patients, and its primary importance must be emphasized. Long-term changes in food choices, eating behaviour, and lifestyle are needed, rather than a temporary restriction of specific foods. The treatment should be nutritionally sound and aim to promote a healthier diet while moderating energy intake and increasing physical activity. Such a treatment may require a period of supervision for at least 6 months. A review of 48 randomized control trials shows that an average weight loss of 8 per cent of the initial body weight can be obtained over 3 to 12 months with a low-calorie diet, and that this weight loss effects a decrease in abdominal fat.

The weight-reducing dietary regimen tailored to an individual's need should initially provide a 600 kcal/day (2.5 MJ/day) energy deficit, based on estimated initial maintenance energy. The diet may best be achieved by a reduction in overall fat intake and will mean, for example, an energy prescription of approximately 1500 kcal/day (6.27 MJ/day) for a moderately active woman of average height aged between 31 and 60 years, and approximately 1800 to 2000 kcal/day (7.52 to 8.36 MJ) for a man of similar age and activity. For overweight patients with a BMI in the range of 27 to 35, a decrease of 300 to 500 kcal/day (1.25 to 2.1 MJ/day) will result in weight loss of about 0.25 to 0.5 kg/week. For more severely obese patients, with a BMI greater than 35, deficits of 500 to 1000 kcal/day (2.1 to 4.2 MJ/day) will lead to weight loss of 0.5 to 1 kg/week. After 6 months, the rate of weight loss usually declines and the weight remains constant because of less expenditure of energy at the lower weight. A further adjustment of calorie intake will be indicated at this stage.

Very low calorie diets

The use of very low calorie diets should only be considered after the failure of determined attempts to lose weight with conventional restriction of normal diets. Their use should follow all of the recommendations from the Committee on Medical Aspects of Food Policy, in particular that such preparations must provide a minimum of 400 kcal (1.7 MJ) per day for women and 500 kcal (2.1 MJ) per day for men. It must be recognized that these diets do not alter eating habits or weight loss beneficially in the longer term. Very low calorie diets may occasionally be useful in the hospital setting for rapid weight loss before surgery. Evidence from randomized trials confirms that over the longer term (more than a year) weight loss following very low calorie diets is no different from that obtained with a low-calorie diet.

Behaviour management

Behavioural interventions seek to alter an individual's lifestyle. Behavioural weight control programmes encourage patients to become more aware of their eating and physical activity and focus on changing the lifestyle and environmental factors that influence their behaviour. All dietary regimens should ideally be linked to behavioural therapy: such therapy may be used by self-help groups. The key difference between behavioural methods and other forms of treatment for obesity is that they lay particular emphasis on personal responsibility for initiating and maintaining treatment rather than the imposition of external authority.

In most trials, behavioural intervention has produced consistent short-term weight loss. Recent studies suggest that a focus on calorie restriction and reduced fat intake, as part of the behavioural approach, is more successful than calorie restriction alone. Studies of methods that seek to avoid circumstances that induce excessive eating are inconclusive, whereas behavioural programmes which provide appropriate foods may be of value. Evidence from randomized controlled trials confirms that behavioural strategies reinforce changes in diet and physical activity in obese adults to produce weight loss in the range of 10 per cent over 4 months to 1 year. Longer-term follow-up shows a return to baseline weight in the absence of continuing behavioural intervention. Trial evidence suggests that behavioural therapy, when used in combination with other weight loss methods, induces further short-term (up to a year) weight loss and that extended treatment programmes improve long-term maintenance of weight.

Exercise and physical activity

When physical activity or exercise alone is used in the treatment of obesity, weight losses are modest and average 2 to 3 kg. This weight loss, although small, exceeds that predicted if direct energy expenditure calculations are performed. For any given weight loss, the loss of fat-free mass is less in exercising versus non-exercising subjects: this is important because fat-free mass is the best predictor of resting metabolic rate which is the largest contributor to total daily energy expenditure. A review of randomized controlled trials provides strong evidence that physical activity alone in obese adults results in modest weight loss and increased cardiovascular fitness.

Regular exercise results in reduction in blood pressure, both in association with or independently of weight loss, and an improvement in atherogenic lipid profiles. A reduction in plasma triglycerides and low-density lipoprotein cholesterol and elevation of high-density lipoprotein cholesterol has been reported with exercise and physical training in obese patients. Exercise also has beneficial effects on glucose metabolism and the sensitivity of skeletal muscle to insulin. However, persuading an obese person to participate in long-term exercise programmes, and to maintain exercise as part of daily routine, is not easy. It is not necessary for the obese patient to increase maximal oxygen uptake by strenuous exercise to derive benefit from exercise: metabolic evidence of improvement in fitness is achieved with less vigorous exercise such as walking increased distances and swimming. The risks from exercise are small, provided it is introduced gradually and pre-existing conditions such as osteoarthritis and ischaemic heart disease are taken into account. The results from randomized controlled trials suggest that a combination of diet and exercise generally produces more weight loss than diet alone, including decreased abdominal fat. More importantly, subjects who exercise adhere to the prescribed diet better than those who do not exercise. One of the most consistent findings in randomized controlled trials of the effect of exercise is the maintenance of weight loss for 2 years.

Drug treatment

The criteria applied to the use of an antiobesity drug should be similar to those applied to the treatment of other relapsing disorders. Many drugs have been advocated over the years as treatment for obesity. Some of these compounds are effective, but many are ineffective.The use of drugs in the management of obesity is bedevilled by the limitations of the available published scientific evidence. It is therefore important that doctors who use these drugs make themselves fully familiar with either the primary literature for any drug, or an authoritative summary document.

Indications for antiobesity drug treatment

It may be appropriate to consider drug treatment if after at least 3 months of supervised diet, exercise, and behavioural management, or at a subsequent review, a patient's BMI is equal to or greater than 30 and weight loss is less than 10 per cent of the presenting weight. In certain clinical circumstances it may also be appropriate to consider antiobesity drug treatment for those patients with established comorbidities whose BMI is 27 or greater, if this is permitted by the drug's licence.

The initiation of drug treatment will depend on the clinician's judgement about the risks to an individual from continuing obesity: drug treatment may be particularly appropriate for patients with comorbid risk factors or complications from their obesity. A drug should not be considered ineffective because weight loss has stopped, provided that the lowered weight is maintained. However, continuation of the drug should depend on the balance between the health benefits of maintained weight and the potential adverse effects of the drug.

Types of drug treatment for obesity

There are currently two categories of antiobesity drugs—those which act on the gastrointestinal system (pancreatic lipase inhibitors) and those which act on the central nervous system to suppress appetite.

Drugs acting on the gastrointestinal system (pancreatic lipase inhibitors)

Orlistat inhibits pancreatic and gastric lipases thereby decreasing the hydrolyisis of ingested triglycerides. It produces a dose-dependent reduction in absorption of dietary fat that is near maximum at a dose of 120 mg three times daily. These actions lead to weight loss in obese subjects. Adverse effects of Orlistat are predominantly related to malabsorption of fat. These include loose or liquid stools, faecal urgency, and oily discharge; they can be associated with malabsorption of fat-soluble vitamins. As the consumption of a high-fat meal will inevitably lead to severe gastrointestinal symptoms, it is possible that some of the weight loss with Orlistat treatment results from an 'antabuse effect', enforcing behavioural change; Orlistat is not itself systemically absorbed.

Centrally acting antiobesity drugs

Drugs which act on the central nervous system can be divided into three groups: those acting via serotoninergic (5-hydroxytryptamine) pathways, for example fenfluramines, those acting via noradrenergic pathways, for example phentermine, and those acting via serotoninergic and noradrenergic pathways, for example sibutramine.

- Drugs acting on serotoninergic pathways: The two drugs from this category, fenfluramine and dexfenfluramine, principally act by releasing serotonin from synapses in the central nervous system; they have only a modest action on inhibiting the reuptake of serotonin into nerve terminals. Because the fenfluramines have been withdrawn they will not be considered further.
- Drugs acting on catecholamine pathways: Phentermine is a phenylethylamine derivative with minor sympathomimetic and stimulant properties whose antiobesity action is due to suppression of appetite. Phentermine has also been withdrawn and will not be considered.
- Drugs acting on noradrenergic and serotoninergic pathways: Sibutramine promotes a sense of satiety through its central action as an inhibitor of serotonin and noradrenaline reuptake. It may also have an enhancing effect on thermogenesis through stimulation of peripheral noradrenergic receptors. Sibutramine is well absorbed following oral ingestion and undergoes first-pass metabolism in the liver to produce two active metabolites that have long elimination half-lives. This enables sibutramine to be given on a single daily basis at a starting dose of 10 mg. Adverse effects include nausea, dry mouth, rhinitis, and constipation. The noradrenergic actions of the drug may cause an increase in blood pressure and heart rate in some patients, or prevent the expected fall in these parameters

with weight loss. The drug should be used with caution in hypertensive patients.

Drugs not appropriate for the treatment of obesity

There is no published evidence to suggest that bulk forming agents (e.g. methyl cellulose) have any beneficial long-term action for weight reduction. Diuretics, human chorionic gonadotrophin, amphetamines, dexamphetamines, and thyroxine are not treatments for obesity and should never be used to achieve weight loss. Under no circumstance should thyroxine be prescribed for obesity in the absence of biochemically proven hypothyroidism. Metformin and acarbose may be useful in the management of the obese non-insulin-dependent diabetic patient: they have no proven efficacy for obesity alone and are not licensed for such use.

Prescribing antiobesity drug treatment

A review of randomized controlled trials provides good evidence that pharmacological therapy combined with diet, lifestyle modification, and physical activity results in weight loss in obese adults that is significantly greater than placebo when the drugs are used for 6 months to 2 years. Experience of the use of antiobesity drugs gained during 12- to 24-month randomized controlled trials indicates that approximately 30 to 40 per cent of the actively treated patients respond as judged by a 5 to 10 per cent reduction in body weight maintained over 12 months. The weight loss occurs in the 'responder' group within 12 weeks. This indicates a suitable time period when 'responders' to drug treatment can be identified and a decision taken to continue the medication. Continual assessment of drug therapy for efficacy and safety is essential. If the drug is effective in helping a patient to lose and/or maintain weight loss, and there are no serious side-effects, it may be continued. If not, it should be discontinued. Once a weight loss target has been achieved, there should be an opportunity for renegotiation of a new target, if indicated, and/or long-term monitoring with reinforcement. Combination therapy of two drugs is contraindicated because of lack of evidence of safety.

Figure 2 summarizes recommendations for the appropriate use of an antiobesity drug.

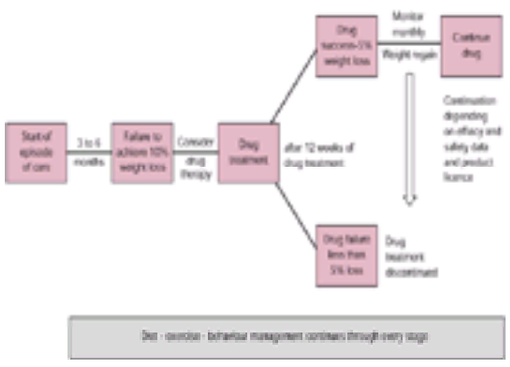

Fig. 2 Suggestions for the practical management of antiobesity drug treatment.

A system of regular medical audits should be a prerequisite of a weight management programme with a record of results and audit action.

Surgical treatment of obesity

There are two operative procedures currently used for the surgical treatment of obesity: gastric restriction and gastric bypass operations.

Gastric restriction involves the creation of a small capacity compartment (less than 20 ml) by either a combination of vertical stapling and a constrictive band opening or a circumgastric band pinching off a small proximal pouch. A modification of the latter procedure is an inflatable circumgastric band attached to a subcutaneous reservoir which allows access by a hypodermic syringe to inject or withdraw fluid thereby tightening or enlarging the band width.

Gastric bypass is performed by stapling shut a vertically oriented pouch of less than 20 ml and connecting this pouch to the jejunum transected 50 cm from the ligament of Treitz (Roux-en-Y gastric bypass). Published evidence confirms that this procedure produces greater weight loss but is accompanied by more frequent adverse effects including 'dumping'.

Most surgical procedures used in the treatment of obesity have been performed laparoscopically which reduces the requirement for sedating pain medication and facilitates prompt postoperative mobilization.

The initial findings from the Swedish Obese Subjects study of severely obese subjects (those with a BMI of more than 40) indicate that weight loss of approximately 30 kg over 2 years is associated with a 60 per cent reduction in plasma insulin, a 25 per cent decrease in plasma glucose and triglycerides, and a 10 per cent reduction in blood pressure. Furthermore, this degree of weight loss resulted in a 14-fold reduction in the risk of developing diabetes and a three- to fourfold risk reduction for the development of hypertension, hypertriglyceridaemia, and low HDL cholesterol levels. Poor health-related quality of life was dramatically improved after gastric restriction surgery, while only minor fluctuations in health-related quality of life were observed in subjects treated by conventional dietary methods. The positive changes in health-related quality of life at 2 years were related to the magnitude of weight loss: the greater the weight loss, the greater the improvement in health-related quality of life.

Randomized controlled trials confirm that surgery for obesity is an option for carefully selected patients with clinically severe obesity (BMI ³ 40 or BMI ³ 35 with comorbid conditions) when less invasive methods of weight loss have failed, and the patient is at high risk for obesity-associated morbidity and mortality. The nature of the surgical procedures necessitates long-term hospital follow-up for such patients.

Weight maintenance

In most patients obesity results not from an inability to lose weight but a profound difficulty in maintaining a lowered weight. A programme to enable the individual to maintain their lowered weight must follow any successful weight loss. A combination of appropriate eating, physical activity, and reinforcement of behavioural methods is the most successful in the long term. Physicians and others can reinforce the importance of this approach but the ultimate responsibility for following such advice must lie with the patient.

Management of obesity during pregnancy

In pregnancy, a weight gain of 12 kg in women of normal weight is associated with the best outcome. By contrast, in women who start pregnancy with a BMI of more than 28, the lowest perinatal mortality is seen with a weight gain of only 4 kg. In pregnancy the aim is thus to limit total weight gain in obese women to 4 kg: the mother-to-be can achieve this by following a nutritionally balanced eating programme prescribed by a registered dietitian, throughout pregnancy.

Management of obesity in childhood

The management of overweight and obesity in children follows the same principles as for adults except that the use of medication is not recommended. Dietary intervention to reduce calorie intake needs to be tempered by a necessity to provide adequate micronutrients such as iron and calcium. Early treatment of childhood obesity has the advantage that it provides the opportunity for the child to 'grow' into their weight. In other words, height may continue to increase for many years and the child maintaining his/her weight will achieve a more favourable body habitus. Any form of dietary restriction must be combined with a regular exercise programme—the design of the activity programme needs to be tailored to the skills of the child and may include walking, swimming, and jogging.

Prevention

The two priority areas for public health strategies aimed at preventing obesity are increasing physical activity and improving the quality of the available diet within a community. However, such strategies must address the need to improve the population's understanding of the nature of obesity and its management and reduce exposure to an environment which promotes obesity. Achievement of these aims requires the involvement of individuals, their families, health professionals, health services, and a commitment from all sectors of the community.

Further reading

Barsh GS, Farooqi IS, O'Rahilly S (2000). Genetics of body weight regulation: applications and opportunities. *Nature* **404**, 644–51.

Chan JM *et al*.(1994). Obesity, fat distribution and weight gain as risk factors for clinical diabetes in men. *Diabetes Care* **17**, 961–9.

Clinical Guidelines, National Heart, Lung and Blood Institute Web site: http://www.nhlbi.nih.gov/nhbli/cardio/obes/prof/guidelns/ob_gdlns.htm

Royal College of Physicians of London (1998). *Clinical management of overweight and obese patients with particular reference to the use of drugs*. Royal College of Physicians, London.

Farooqi IS *et al*. (1999). Effects of recombinant leptin therapy in a child with congenital leptin deficiency. *New England Journal of Medicine* **341**, 879–84.

Flegal KM, Carroll MD, Kuczmarski RJ, Johnson CL (1998). Overweight and obesity in the United States: prevalence and trends, 1960–1994. *International Journal of Obesity* **22**, 39–47.

Glenny A-M *et al*. (1997). The treatment and prevention of obesity: a systematic review of the literature. *International Journal of Obesity* **21**, 715–37.

Grunstein RR (1998). Pulmonary function, sleep apnoea and obesity. In: Kopelman PG, Stock MJ, eds. *Clinical obesity*, pp 248–89. Blackwell Science, Oxford.

Hubert HB *et al.* (1983). Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham heart study. *Circulation* **67**, 968–77.

Kopelman PG (2000). Obesity as a medical problem. *Nature* **404**, 635–643.

Kral J (1998). Surgical treatment of obesity. In: Kopelman PG, Stock MJ, eds. *Clinical obesity*, pp 545–63 Blackwell Science, Oxford.

Lew EA (1985). Mortality and weight: insured lives and the American Cancer Study. *Annals of Internal Medicine*, **103**, 1024–9.

Manson JE *et al*. (1995). Body weight and mortality among women. *New England Journal of Medicine* **333**, 677–85.

Prentice AM, Jebb SA (1995). Obesity in Britain: gluttony or sloth? *British Medical Journal* **311**, 437–9.

Ravussin E *et al*. (1988). Reduced rate of expenditure as a risk factor for body weight. *New England Journal of Medicine* **318**, 467–72.

Willett WC, Dietz WH, Colditz GA (1999). Guidelines for healthy weight. *New England Journal of Medicine* **341**, 427–33.

World Health Organization (1997). *Obesity: preventing and managing the global epidemic.* WHO, Geneva.

World Health Organization Expert Committee (1995). *Physical status: the use and interpretation of anthropometry.* WHO Technical Report Series no. 854. WHO, Geneva.

World Health Organization MONICA Project (1988). Geographical variation in the major risk factors of coronary heart disease in men and women aged 35–64 years. *World Health Statistics Quarterly* **41**, 115–40.

Zhang Y *et al*. (1994). Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–32.

10.6 Special nutritional problems and the use of enteral and parenteral nutrition

M. Elia

Anorexia is a common consequence of disease. Undernutrition may increase morbidity after elective surgery and accidental injury, produce non-specific symptoms such as lethargy, depression, and fatigue, reduce tolerance to cytotoxic drugs or radiotherapy, and prolong the hospital stay.

It is possible to administer sufficient nutrients to individuals at risk of malnutrition, including those with gastrointestinal failure and severe burns, and those who are unconscious. Artificial nutritional support may be given in nursing homes and even at home so that it has become an important aspect of treatment.

The benefits of nutritional support depend on the severity of the disease and on nutritional status; this assessment of clinical and nutritional status is the first important step in rationalizing its use.

Nutritional screening and assessment

A well-structured history may provide useful information about possible undernutrition and the likelihood of specific nutrient deficiencies. Detecting malnutrition in routine clinical practice should be simple, reliable, and reproducible. Criteria for detecting malnutrition include the following: a body mass index of less than 18.5 kg/m ^{2} (chronic protein–energy undernutrition) or less than 20 kg/m ^{2} with a history of weight loss; unintentional weight loss of more than 10 per cent of body weight during the preceeding 6 months; and associated reduction in food intake/reduced appetite. The history may also reveal underlying psychosocial problems (loneliness, bereavement, isolation, alcoholism) and physical disabilities (active disease; for example painful mouth conditions, difficulties with eating and swallowing, inability to self-care) that are likely to have contributed to the weight loss. It is obvious that these underlying problems will also have to be addressed as part of the management.

The clinical history may indicate specific nutrient deficiencies: blood loss leads to iron deficiency; previous gastric surgery to vitamin B_{12} or iron deficiency; coeliac disease may lead to folic acid and iron deficiency; and intestinal resections or fistulae associated with large amounts of intestinal effluents lead to deficiency of many nutrients including sodium, magnesium, and zinc. However, it is often difficult to establish the diagnosis of specific nutrient deficiencies from the history alone.

A dietary history may alert the clinician to major reductions in dietary intake and may indicate the likelihood of specific nutrient deficiencies. For example: anaemia in a vegan may be due to an inadequate intake of vitamin B_{12}; a diet poor in fruit and vegetables may predispose to vitamin C deficiency; a diet poor in fish and margarine (which is normally supplemented with vitamin D) may predispose to rickets or osteomalacia, especially in housebound individuals not exposed to sunlight.

Clinical examination

The relationship of height to weight provides a useful indication of nutritional status. In children an inadequate intake frequently results in growth retardation. Centile charts are useful for this assessment, particularly if sequential measurements are made. In adults an inadequate intake leads to wasting. The body mass index (weight (in kg)/height^2 (in m^2)) can give some indication of the extent of depletion (Fig. 1), as can skinfold thickness (measured with a caliper), arm circumference, and arm muscle area (calculated from the arm circumference and skinfold thickness at the level at which the circumference was measured) (Fig. 2). However, the range of normality is large, and it is possible for subjects to lose a substantial amount of body weight, adipose tissue, or muscle mass and still exhibit normal indices (Fig. 1 and Fig. 2), or even above the normal range (for example obese individuals). It is therefore more useful to undertake sequential rather than single measurements.

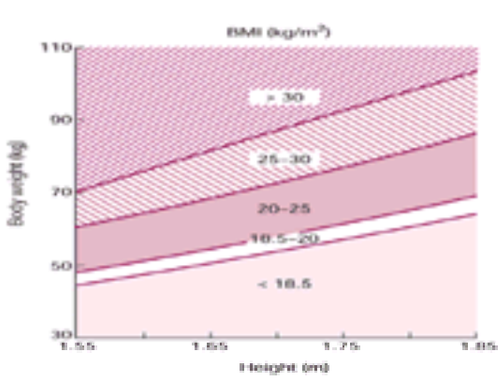

Fig. 1 Ranges of body mass index (weight (in kg)/height^2(in m^2)): less than 17, moderate to severe chronic protein–energy malnutrition (severe if less than 15); 17 to 20, chronic protein–energy malnutrition but some normal subjects; 20 to 25, desirable (some authorities 19 to 25); 25 to 30, mildly overweight (grade 1 obesity); 30 to 40, grade II obesity; over 40, grade III obesity.

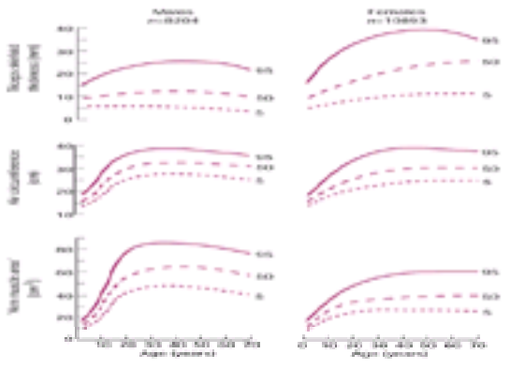

Fig. 2 Percentile curves of triceps skinfold thickness and the estimated mid-upper arm cross-sectional muscle are of American citizens. Based on the United States Health and Nutrition Survey (Frisancho 1981, 1984).

An overall impression of nutritional status (subjective global assessment) may emerge from a combination of history and clinical examination. The examination should assess muscle wasting, the amount of subcutaneous fat, and the presence of oedema. Many other signs may be useful, especially on examining the limbs and integument (Table 1): chelosis or stomatitis may be the result of deficiency of vitamin B complex (but angular stomatitis is common in elderly subjects and may also result from badly fitting dentures); atrophic glossitis may result from folate or vitamin B_{12} deficiency (this may also result from antibiotic therapy); koilonychia may result from severe iron deficiency; tetany may occur in patients with vitamin D deficiency, or hypomagnesaemia; and a rash (acrodermatitis enteropathica) may occur in zinc deficiency. In alcoholics the presence of Wernicke's encephalopathy, Korsakoff's psychosis, and peripheral neuropathy point to the likelihood of thiamin deficiency. Easy bruising may be a sign of scurvy or vitamin K deficiency.

Since the signs and symptoms of many nutrient deficiencies are non-specific, they are often present in a severe form before they can be detected clinically. Deficiencies are usually present in combinations rather than in isolation. The key is to remember to include nutritional causes in the differential diagnosis.

Preventing and treating malnutrition and the indications for artificial nutritional support

Malnutrition in a variety of clinical conditions has been linked to poor outcome. This, together with the high prevalence of malnutrition reported in hospitals, has contributed to enthusiasm for artificial nutritional support. However, there are many simple things that the clinician can do to improve nutritional state. Nausea may be helped by an antiemetic. A person with dysphagia due to an oesophageal stricture may be helped by the provision of sloppy or liquid meals rather than solid foods. In contrast, patients with neurological disorders of swallowing may benefit from more viscous liquids. Pain causes anorexia, and its relief may improve appetite. Dedicating time to feeding weak and elderly patients may do much to improve their nutrition or at least prevent malnutrition. This task may be undertaken by nurses, health care assistants, or relatives.

When intake is inadequate, oral supplements may be tried. If these fail, enteral or parenteral nutrition may have to be used (see below for indications). In some patients it is immediately obvious that artificial nutritional support is necessary. This applies to patients who are unconscious (and likely to remain unconscious for a long period), and those who are unable to swallow or have intestinal failure. It also may apply to patients subjected to major surgery, for example oesophagogastrectomy, who are routinely prevented from eating for a week or more until the anastomosis has adequately healed, and patients receiving aggressive chemotherapy for haematological malignancies (bone marrow transplantation). These latter patients typically develop severe inflammation of the mouth and other parts of the gastrointestinal tract, so that artificial nutritional support (enteral or parenteral nutrition) may be required.

Oral or enteral nutrition should be used where possible because it is simpler, cheaper, and more physiologically acceptable than parenteral nutrition. In addition, enteral nutrition appears to be better than parenteral nutrition in maintaining the integrity of the 'gut barrier', which prevents bacteria and associated endotoxins from entering the systemic circulation. General and specific recommendations about the use of enteral tube feeding are given in Table 2. Some of these guidelines also apply to parenteral nutrition but only when the gut is not available for feeding. Well-recognized indications are prolonged gastrointestinal failure in the form of ileus, peritonitis, severe and recurrent pancreatitis, high intestinal fistulae, short bowel syndrome, or severe inflammatory disease of the intestine, for example severe mucositis following cytotoxic therapy, or Crohn's disease complicated by fistulae. The use of parenteral nutrition in the postoperative period is discussed separately.

Infusion of nutrients into peripheral veins has often been associated with rapid development of phlebitis and venous occlusion. These complications may be reduced by infusing solutions of lower osmolarity (a larger volume, less glucose, and more fat—since fat emulsions have an osmolarity close to that of blood), through suitable fine-bore cannulae. Small doses of heparin and/or corticosteroids and vasodilatory glycerin trinitrate skin patches may prevent venous occlusion. However, peripheral parenteral venous feeding has obvious limitations for patients with poor peripheral venous access and for those requiring prolonged infusions of hypertonic solutions.

Nutritional requirements

Protein and energy

Recommendations about nutrient intake depend on disease activity and nutritional state. Figure 3 shows the effects of increasing nitrogen intake on the nitrogen balance in subjects who are close to energy equilibrium. Normal individuals in energy balance achieve nitrogen balance at a mean intake of 0.1 g N/kg. The World Health Organization recommends a minimum of 0.12 g N/kg, to take into account the variability between individuals (+2 standard deviations). A greater intake produces little improvement in nitrogen balance in normal subjects. In contrast, depleted individuals, particularly those without associated inflammatory or infective disease, achieve a progressively greater positive balance as more nitrogen is taken in (Fig. 3).

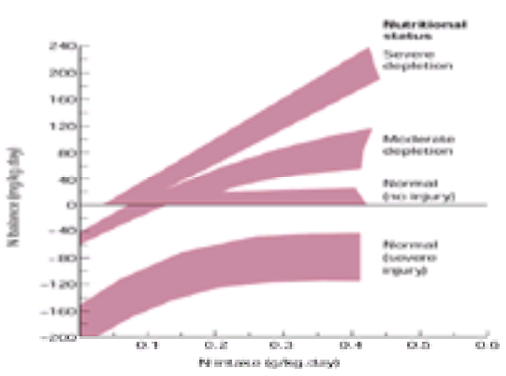

Fig. 3 Relationship between nitrogen intake (1 g N = 6.25 g protein) and nitrogen balance in subjects who are receiving sufficient energy to be close to energy balance (see text).

Negative nitrogen balance in catabolic states is due to a combination of the disease itself, which enhances both net muscle proteolysis and liver gluconeogenesis, immobility, and the effect of drugs such as steroids. The catabolism is usually greatest within the first few days of injury but in patients suffering from burns it may continue for weeks. In well-nourished patients (such as those with sepsis, trauma, or burns) who are close to energy balance, an increase in nutritional intake results in improved nutritional balance. However, as Fig. 3 indicates, the relationship between intake and nitrogen balance is disturbed so that the more severe the injury, the greater the catabolism. In practice, many patients become malnourished following a severe catabolic injury, so that the response to nutrient intake is intermediate

between malnutrition uncomplicated by disease and severe injury uncomplicated by malnutrition.

From these different responses (Fig. 3) some general recommendations emerge for nitrogen intake. (Fig. 4). The recommended energy intake also varies with the clinical state. In well-nourished individuals who are likely to receive nutritional support for long periods, it is wise to aim for energy balance. In the depleted patient, it is desirable to achieve a positive energy balance (as well as a positive nitrogen balance), whereas in obese individuals, loss of adiposity (while limiting the loss of lean tissue) is desirable.

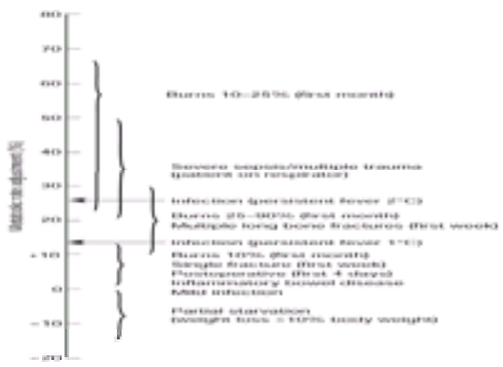

Fig. 4 Guidelines for estimating the approximate energy and nitrogen (N) requirements (1 g N = 6.25 g protein) for an adult patient receiving artificial nutritional support (based on Elia 1994). Energy: 1, Estimate the approximate basal metabolic rate of a normal adult with the same weight as the patient from standard reference tables. 2, Adjust the basal metabolic rate for disease according to the accompanying nomogram. 3, Add a combined factor for activity and thermogenesis (mainly diet induced)—bed-bound plus10 per cent; bed-bound, mobile/sitting, plus 15 to 20 per cent; mobile on ward, plus 25 per cent. 4, If an increase in energy stores is required add 400 to 800 kcal/day (1680 to 3360 kJ/day). If a decrease in energy stores is required energy intake can be reduced. Protein: 1, Consult the nonogram to determine the degree of hypermetabolism—0 per cent, 0.17 g N/kg/day ('normal' protein requirement); 5 to 25 per cent, 0.2 g N/kg/day; more than 25 per cent, 0.20 to 0.25 g N/kg/day. 2, If patient is depleted can increase to 0.25 (0.2 to 0.3) g N/kg/day For obese individuals with a body mass index of 30 to 40 kg/m ^{2} use about 75 per cent of the value estimated from body weight and for those with a body mass index over 40 kg/m^2 use about 60 to 65 per cent of the value estimated from body weight

In calculating energy requirements it is also important to remember that reduced physical activity often offsets any increase in basal metabolic rate due to disease. Even in ambulatory patients receiving parenteral nutrition in hospital and at home, total 24-h energy expenditure has often been found to be 1700 to 2300 kcal/day (7140 to 9660 kJ/day).

Energy requirement varies substantially with body size, nutritional status, and disease activity. Approximate estimates for requirements are given in Fig. 4. Most hospitalized adult patients generally require 1500 to 2500 kcal (6300 to 10 500 kJ)and between 10 and 15 g N/day. In prescribing artificial feeding it is often possible to approximate the requirements to the contents of three or four standard regimens, to monitor the patient's progress, and to adjust accordingly. The intake of energy can sometimes be reduced in the acute phase of disease, when there is metabolic instability and substantial intolerance to nutrients such as glucose, or in obese individuals. The intake of protein may be reduced in hepatic encephalopathy, and in renal failure if the aim is to reduce the uraemia and either to avoid dialysis or to increase the interval between dialyses. However, protein requirements are frequently administered and the patient is dialysed according to needs. Fluid restriction may limit the administration of nutrients and a compromise between the general clinical and nutrient needs may be necessary.

Fluid

Usually 1.5 to 3.0 litres of fluid are given to adults receiving enteral (around 1.5–2.0 litres) or parenteral nutrition (around 1.5–3.0 litres), but different intakes may be required. Particularly high fluid and electrocyte requirements are necessary in patients with large-output intestinal fistulae. Usually the fluid allocated to nutritional support is restricted, partly because the patient may already be receiving fluid for other purposes (for example infusions of drugs and blood), and partly because the clinical conditions demand it. Low fluid intake may be necessary in oedematous patients, those with renal hepatic or cardiac failure, and patients who have suffered a recent head injury. After head injury fluid restriction is frequently recommended to limit brain oedema, which can adversely affect clinical outcome; however, some head-injured subjects may develop diabetes insipidus, when extra fluid is required. Patients with burns also have high fluid requirements. On the other hand, many acute diseases (such as various forms of trauma or injury) are associated with impaired renal excretion of water and salt loads, and therefore fluid balance has to be monitored carefully. In the intensive care unit, fluid intake is often modified according to measurements of central venous pressure. Major changes in body weight from one day to the next can also help in the assessment of fluid requirement since these predominantly reflect changes in fluid balance.

Minerals and trace elements

For many adult patients receiving artificial nutritional support, the recommended intake of sodium and potassium is 50 to 100 mmol/day. Chloride intake is usually similar to that of sodium, and phosphate is usually prescribed at a dose of 20 to 40 mmol/day. However, since artificial nutrition is used in a wide range of patients with different disorders it is not surprising that mineral requirements vary considerably. For example, sodium restriction may be necessary in patients with renal, hepatic, and cardiac failure who are prone to fluid retention. Additional sodium is required in patients with increased gastrointestinal effluents. The additional requirements of sodium and potassium can be predicted from effluent composition as indicated in Table 3. Note that the loss of 1 litre of gastrointestinal fluid may more than double the sodium requirements while affecting potassium requirements to a much smaller extent. However, more potassium may be required in patients with excessive renal losses. Potassium requirements may double in patients receiving amphotericin B, which is often given to recipients of a bone marrow transplant. Adequate amounts of potassium and phosphate, which are predominantly intracellular ions, are also necessary during repletion of lean tissue. Indeed, deficiencies of these substances can exacerbate negative nitrogen balance.

The recommended intakes of calcium, magnesium, and many trace elements are quite different for oral/enteral nutrition compared with intravenous nutrition (Table 4). This is because the gut only absorbs a proportion of these nutrients, sometimes less than 10 per cent (for example in the case of chromium). Prolonged intravenous administration of trace elements at the dose recommended orally may prove to be toxic. One of the functions of the gut is to limit the uptake of potentially toxic substances that may be present in excess in the diet or in the gut. The gut is the most important organ regulating the availability to the body of some trace elements, as in the case of iron and manganese for which there is little capacity for disposal by other organs such as the kidney or liver. However, the gut does not appear to be so important in the regulation of fluorine, iodine, or selenium status.

The requirements of trace elements in various diseases are not clearly established, although patients with intestinal fluid losses may have substantially greater requirements for zinc (see footnote to Table 4).

Vitamins

In contrast to the intravenous recommendations for trace elements, which are generally lower (sometimes by severalfold) than those given orally, the reverse is true for vitamins (Table 5). This is partly because the vitamins are generally absorbed to a much greater extent than most trace elements, and partly because their requirement probably increases in many diseases. Although the vitamin requirements in particular diseases, especially infective and active inflammatory ones, are not well established, the requirements for some vitamins may be considerably greater than in health. In addition, some vitamins may degrade during the preparation and storage of parenteral nutrition solutions. For example vitamin A, riboflavin, and vitamin K are photosensitive, and vitamin C may degrade in the presence of trace elements and oxygen. Thiamin can degrade in the presence of sulphite, which is used as a preservative, and vitamin A palminate may be adsorbed on some plastic storage bags or administration sets. It should be also be remembered that some patients are depleted of vitamins prior to the initiation of therapy, and therefore extra intake is necessary to replete the stores.

Manufactured enteral feeds have a long shelf-life and contain trace elements and vitamins. Parenteral feeds have a shorter shelf-life and vitamins and trace elements are often added shortly before use because of the concern about stability (see above). Vitamin K need not be added routinely in parenteral feeds and particular care must be taken in patients who are on anticoagulants. Sufficient quantities of this vitamin are normally synthesized in the gut but a weekly intramuscular dose is often

recommended, especially in those receiving antibiotics that affect the metabolism of intestinal bacteria. More frequent doses may be given to patients with liver disease who have a coagulation problem and are at risk of gastrointestinal bleeding.

Although deficiencies are not likely to develop for months or possibly years (vitamin A, vitamin D, vitamin B_{12}), it is usual practice to administer a mixture of trace elements and vitamins from the outset even if nutritional support is only likely to be required for a few weeks. This is because the stores of some (mainly water-soluble vitamins such as thiamin and riboflavin) are very small. Furthermore, some patients are malnourished at presentation so that the stores of other trace elements or vitamins may already be depleted.

Complications of artificial nutritional support: prevention, treatment, and monitoring

A summary of the complications that may be encountered during parenteral and enteral nutrition is given in Table 6.

Parenteral nutrition

Mechanical

Complications related to the insertion of a central venous catheter, usually into the subclavian vein, are not common, although pneumothorax may occur in 2 to 3 per cent of cases (the frequency depends on the expertise of the person involved in the procedure). The insertion of the catheter is carried out under aseptic conditions and the position of the catheter tip is confirmed radiologically after insertion. Radiography also helps exclude other complications such as pneumothorax.

Many of the risks associated with parenteral nutrition can be reduced by following appropriate protocols. For example, laying the patient head down while changing feeds and checking the position of locks prevents air embolism. The use of strict aseptic techniques is essential to prevent catheter-related sepsis (see below).

Occlusion of the catheter may result from reflux of blood into the catheter, but it may also result from coagulation of the feed, especially when all-in-one solutions which include lipid are infused. The incidence of catheter blockage depends on how long the catheter is used for, the diameter of the catheter, the type of catheter (soft polyurethane and Teflon catheters are said to have a lower risk of thrombosis than rigid polyethylene catheters), and the type of feed administered (all-in-one mixtures tend to cause line blockage more readily). In patients receiving cyclic nocturnal feeding, flushing the catheter with heparin (50 IU/ml) at the end of feeding reduces the risk of thrombosis. Some recommend routine inclusion of heparin (2–3 IU/ml) in the parenteral nutrition solution to prevent both catheter blockage and local venous thrombosis. Several methods may be tried to unblock an occluded catheter. Gentle suction may remove the clot. The clot may be lysed by inserting a solution of urokinase (5000 U/ml) for about 1 h. Alcohol (50 per cent) may be used in a similar way to dissolve lipid-associated occlusions. Insertion of hydrochloric acid (1 M) into the catheter is another potentially effective method.

Infections

Catheter-related infection is an important complication. The infecting organisms are typically derived from the skin (for example *Staphylococcus aureus*, *Staphylococcus epidermidis*), although a variety of other organisms from the systemic circulation including Gram-negative organisms and fungi may seed on to the catheter tip, especially when it is associated with a fibrin clot. Catheter-related sepsis can largely be avoided by the use of aseptic techniques during insertion of the catheter and during the change of feeds, and by avoiding the use of the central venous catheter for purposes other than feeding, such as blood sampling or the administration of drugs and blood.

When catheter-related sepsis is strongly suspected and the patient is unwell and deteriorating, the catheter should be removed. However, it should be remembered that most episodes of pyrexia are not due to catheter-related sepsis, and the skill of the physician/surgeon in diagnosing alternative causes (such as wound infection, pyrexia of trauma, urinary tract infection in patients with urinary catheters, pulmonary embolism) can prevent unnecessary removal of central venous catheters and the hazards that accompany recannulation. Blood cultures taken both from the central venous line and a peripheral vein and a swab from the catheter enteral site may help to identify the type of organism causing sepsis, and the likelihood that it is related to the catheter.

The procedure of tunnelling the line under the skin from its site of insertion (typically the subclavian vein) to the anterior chest wall makes dressing and care of the catheter easier, and the location is often more comfortable for the patient. However, there is little evidence that the use of a tunnelling procedure reduces the incidence of catheter-related sepsis.

Multilumen catheters are sometimes required for multiple uses (sampling and administration of blood, and use for infusion of parenteral solutions including drug therapy). They have the advantage of convenience, especially in patients with limited peripheral venous access. Reports suggest that such catheters become infected more frequently than single-lumen catheters used solely for parenteral nutrition but this is not surprising because multilumen catheters are used in patients who have more severe disease.

Metabolic complications

Fluid and electrolyte abnormalities are common during parenteral nutrition. This is largely because the underlying condition may result in excess fluid and electrolyte losses (postoperative nasogastric losses, intestinal fistula, etc.) or retention (renal, cardiac, and hepatic failure). Drug therapy may also affect acid–base fluid and electrolyte status. Clinicians involved with the nutritional support of the patient must liaise closely with those involved with other aspects of management. The nutrition team has an important role, partly because it can make the necessary daily adjustments of fluid and electrolytes in the parenteral nutrition solution and partly because it can make adjustments to other minerals or micronutrients that are not administered routinely to ill patients, for example additions of zinc and magnesium in patients with persistent loss of intestinal effluent. Magnesium deficiency may lead to neuromuscular excitability and tetany. It may also produce hypocalcaemia which is not corrected by calcium administration. Zinc deficiency may impair wound healing and produce severe dermatitis.

Hyperglycaemia is common in patients receiving parenteral nutrition in hospital. This is largely because glucose intolerance is frequently associated with severe disease. It can be managed by reducing the intake of glucose (with or without an increase in lipid intake) or by administering insulin, either as a constant infusion or by intermittent subcutaneous or intramuscular injections, at a dose determined by blood glucose concentrations. Particularly high glucose concentrations may occur if the rate of infusion of nutrients is not adequately regulated. Without the use of an infusion pump the rate of infusion may increase severalfold to cause severe hyperglycaemia, hyperosmolarity, headaches, vomiting, and an impaired level of consciousness.

Abnormal liver-related tests are frequently observed in patients receiving parenteral nutrition in hospital, for example increased activities of glutamate oxaloacetate transaminase, serum glutamic pyruvic transaminase, and alkaline phosphatase. These frequently reflect the underlying disease (such as sepsis, malignancy, inflammatory bowel disease, pre-existing liver disease) but other factors may be involved: infusion of lipid or excess glucose (leading to hepatic steatosis); bacterial overgrowth in the intestine; and biliary sludge and even gallstones. The prolonged absence of oral intake during parenteral nutrition fails to stimulate normal gallbladder contraction and this is probably responsible for the development of biliary sludge. Detailed investigation of abnormal liver function (biochemical tests, ultrasound scans, and sometimes liver biopsy) may, on occasion, be necessary to discover the underlying pathology.

Deficiencies of phosphate and essential fatty acids have both been reported during parenteral nutrition, arising from their lack of inclusion in parenteral nutrition solutions. Phosphate deficiency can cause muscle weakness and impair the utilization of protein. It also causes hypercalcaemia and, in the long term, bone disease. Deficiency of essential fatty acids produces alopecia, thrombocytopenia, anaemia, and a skin rash as early as 6 weeks after starting intravenous nutritional support without fat. Biochemically, it is diagnosed by an increase in the triene to tetraene ratio (> 0.4) since, in the absence of linoleic acid, oleic acid is metabolized to eicosatrienoic acid. The condition is more likely to develop in patients receiving continuous rather than intermittent parenteral nutrition, because essential fatty acids from the endogenous lipid stores are continually prevented from being released by hyperinsulinaemia. The deficiency syndrome is rapidly reversed by administering an intravenous lipid. Regular application to the skin of oils containing essential fatty acids allows sufficient absorption of fatty acids to treat or prevent this syndrome.

In patients intolerant of lipid (for example patients with hyperlipidaemia and some patients with renal or hepatic disease or diabetes) hypertriglyceridaemia results. This may affect the assays of a number of standard biochemical tests and dilute other plasma constituents (thereby causing pseudohyponatraemia). Visual inspection of plasma for lipid several hours after cessation of the lipid infusion can alert the clinician to this problem. Measurement of plasma triglycerides provides a more

accurate assessment.

Lipid infusion has been implicated in affecting the function of some organs. For example, hepatic steatosis may cause abnormal liver function tests, and lung deposition in patients with respiratory distress can impair pulmonary function by reducing the permeability of the lung to gases.

Excessive administration of glucose may also have adverse respiratory effects. This is because glucose produces 30 per cent more CO_2 per MJ than fat, and an even greater amount of CO_2 per MJ when there is net lipogenesis from carbohydrate. Furthermore, excessive administration of glucose increases energy expenditure (dietary-induced thermogenesis) to a greater extent than fat. In patients with impaired pulmonary function this may precipitate respiratory failure, or impair weaning of a patient with respiratory failure from a respirator. However, with the typical amount of glucose infused in most patients this is a theoretical rather than a practical occurrence and the use of high-fat regimens is generally not indicated.

Metabolic bone disease (mainly osteoporosis) is associated with long-term parenteral nutrition (usually home parenteral nutrition). Several factors may contribute, including corticosteroid therapy, the underlying disease, and immobility. Excess amino acid intake and heparin have also been implicated. Aluminium toxicity has been implicated as a cause of a painful metabolic bone disorder.

Trace element and vitamin deficiencies may occur in patients on long-term parenteral nutrition. Usually this is due to the prescription of insufficient amounts, but excessive losses of intestinal effluents may also be responsible. Deficiencies of several trace elements have been described, for example copper, zinc, iron, selenium, as well as case reports of chromium and molybdenum deficiencies. Several vitamin deficiencies have also been described, including biotin deficiency (eczematous dermatitis, hair loss, depression, anorexia) which is rare under normal circumstances, and night blindness due to vitamin A deficiency. These should be uncommon if appropriate protocols are followed.

Enteral nutrition

The complications most associated with enteral feeding in hospitalized patients include nausea or vomiting (10–20 per cent), abnormal bloating and cramps, diarrhoea (5–30 per cent), and constipation (but mainly in long-term feeders at home). Delayed gastric emptying is a feature of many conditions including postoperative abdominal surgery, head injury, and severe sepsis. Gastric stasis may lead to accumulation of feed in the stomach, so that eventually the patient develops nausea and vomiting.

In unconscious patients and those with an impaired swallowing reflex, vomiting may lead to aspiration pneumonia, which is one of the most serious complications of enteral nutrition. Gastro-oesophageal regurgitation may also predispose to aspiration pneumonia. Some of the effects of poor gastric emptying can be prevented by administering the feed directly into the small intestine. A nasogastric tube may be placed in the small intestine under radiographic control, endoscopically, or during surgery. In those at risk of gastric stasis/regurgitation, gastric pooling can be checked by intermittent aspiration through tubes with a sufficiently wide bore. Continuous infusion of feed into the stomach can prevent the sudden gastric disturbance associated with bolus feeding. The use of an infusion pump to control delivery of feed into the stomach can prevent gastric flooding associated with inadequate manual flow control systems. The use of metoclopramide or erythromycin, which stimulate gastric emptying, may be beneficial. Erythromycin acts as a motilin receptor agonist.

Despite the frequency of diarrhoea in patients receiving enteral tube feeding in hospital, the mechanism is not entirely understood. However, it is often associated with antibiotic therapy. Lactose intolerance has also been implicated but most enteral feeds are free of lactose. Rapid delivery of nutrients into the gastrointestinal tract may lead to diarrhoea, especially if the delivery is postpyloric. Here the protective effect of the pylorus in regulating delivery of nutrients into the small intestine is bypassed. Other factors have been implicated, including bacterial contamination of enteral diets, an underlying gastrointestinal disease, the use of laxatives, lack of dietary fibre, and neuroendocrine reflexes whereby the administration of feed in the stomach or upper small intestine cause secretion of fluid in the small and large bowel. Diarrhoea may be prevented by taking care not to contaminate the feed with bacteria, controlling the rate of feed infusion, and/or treating the underlying condition. Drugs such as codeine phosphate or loperamide may help to control the symptoms.

Constipation may also complicate long-term enteral nutrition, particularly in elderly, inactive subjects. Lack of fibre has been implicated, but the constipation remains despite fibre supplements.

Another potential problem is regurgitation of feed in patients with impaired gastro-oesophageal function, such as elderly people with a hiatus hernia who have an impaired swallowing reflex. This may lead to aspiration pneumonia which can be avoided by administering the feed with the upper part of the body elevated to an angle of about 30°; in those with a high risk of aspiration, it is best to administer the feed directly into the small intestine (jejunostomy feeding).

Several metabolic disturbances have been described during enteral feeding: hyperglycaemia in glucose-intolerant subjects; rebound hypoglycaemia after sudden withdrawal of feed; disturbances in plasma potassium, depending on the patient's renal and gastrointestinal function and the potassium content of the feed; hypophosphataemia during refeeding. Refeeding malnourished subjects may produce hypophosphataemia and hypokalaemia as lean tissues containing these eletrolytes are accreted. New legislative regulations implemented in the European Community in 1999, which demand that the micronutrient to energy ratio is within a specific range, should reduce the incidence of trace element, mineral and vitamin deficiencies.

The complications of enteral nutrition at home are similar to those in hospital. However, lack of enteral access (due to tube blockage or dislodgement) can be an important problem, because it may lead to dehydration, especially in those with swallowing-related problems who are prevented from drinking. Flushing of the tubes with water at the end of each feeding period can prevent tube blockage. A blocked tube may be unblocked by flushing it with water in the first instance, followed by a warm solution of sodium bicarbonate or by digesting the coagulated feed with pancreatic enzymes. Fizzy cola drinks may also be effective in unblocking tubes. If a gastrostomy or enterostomy tube has been dislodged, it is important to replace it quickly because the stoma may rapidly close up and make further access difficult.

General monitoring of patients on artificial nutritional support

Careful observations should be made of the patient shortly after the start of enteral or parenteral nutrition to assess feed tolerance. In patients receiving parenteral nutrition, urine should be analysed every 4 h to check for glycosuria, and blood glucose should be measured in these patients and others with suspected glucose intolerance. The development of glycosuria in previously stable patients without glycosuria may indicate the development of a complication such as infection before it has been diagnosed clinically in hospital. Routine measurements of temperature, pulse, blood pressure, and fluid balance are also essential. Changes in daily weight are the best indices of day-to-day changes in fluid balance. In the longer term, changes in weight usually indicate changes in lean and adipose tissue in response to the support provided.

The frequency with which other investigations are carried out depends on the patient and the underlying condition (see also above). In patients with large losses of gastrointestinal fluids or those on long-term parenteral nutrition (especially when there is little or no enteral or oral intake) an assessment of trace element and/or vitamin status is often necessary. The catheter site should be inspected regularly as a possible source of infection, and fresh dressings applied according to standard protocols.

In patients receiving enteral tube feeding, it is important to assess feed tolerance by ensuring that gastric pooling does not occur, especially in those at risk of gastric stasis and those with an impaired swallowing reflex. The development of diarrhoea should be investigated promptly so that appropriate action (such as adjustment of infusion rate or eradication of specific gastrointestinal pathogens) can be taken.

Nutritional aspects of specific conditions

The reader is referred to other sections for information on nutritional and fluid and electrolyte aspects of various clinical conditions, for example acute and chronic renal failure, malabsorption syndrome, hyperlipidaemia, diabetes mellitus, cystic fibrosis, short bowel syndrome, and enterocutaneous fistulae. Nutritional aspects of other conditions that have recently gained prominence are discussed here.

Acquired immune deficiency syndrome (AIDS)

AIDS has important nutritional consequences. Early studies suggested that weight loss at death was commonly over 20 per cent of body weight, and sometimes more

than 40 per cent of body weight. Severe weight loss occurs in low-income countries but in more developed countries where protease inhibitors have been extensively used, survival has been prolonged and the frequency of severe undernutrition has been reduced. These changes have been associated with the development of a syndrome that includes abnormal central fat distribution, hyperlipidaemia, and a tendency to lactic acidosis. In the long term, the risks of malnutrition may be replaced by an increased risk of cardiovascular disease. Important clinical problems evident in the era before antiprotease therapy remain relevant to patients in high-income countries, especially at presentation, as well as those in countries where protease inhibitors are not widely used.

Acute weight loss is usually a consequence of acute infections, whereas more chronic loss is usually associated with an enteropathy. There are many direct causes of weight loss. Food intake is usually decreased during acute infections, but during recovery it may be normal or even greater than normal. In advanced AIDS decreased food intake is often associated with opportunistic infections in the mouth, pharynx, and oesophagus, which may cause pain and dysphagia. Associated malignancy (Kaposi's sarcoma and non-Hodgkin's lymphoma) may also lead to anorexia. Furthermore, antifungal and antiviral drugs can cause nausea, vomiting, and anorexia, and the use of chemotherapeutic agents for malignancy can produce stomatitis, pharyngitis, and oesophagitis, which can make swallowing painful and distressing. The enteropathy can be caused by many pathogens and can lead to steatorrhoea, fluid and electrolyte disturbances, and trace element and vitamin deficiencies. Neurological disease is also often associated with malnutrition. Dysphagia and coma obviously lead to reduced nutrient intake and malnutrition, and 'dementia', which may affect half the patients with advanced AIDS, may make dietary assessment and management particularly difficult.

Treatment of the underlying condition, such as mouth, throat, and other systemic infections which cause dysphagia, or systemic infections which cause anorexia, can do much to improve nutritional status. There is substantial anabolic potential between acute infective episodes, and the rapid diagnosis and treatment of such infections produce better nutritional results than delayed treatment.

Nutritional assessment should begin at the outset and changes in body weight in relation to the disease process should be closely monitored. Psychological evaluation and social counselling should not be neglected. Anxiety, apathy, and depression are common and may lead to self-neglect, irregular food intake, and deterioration of nutritional status. Support should begin with general nutritional advice about diet, but it may need to progress to the use of supplements and, occasionally, enteral tube feeding or parenteral nutrition.

Surveys have shown that many patients with AIDS take supplements of vitamins and trace elements. In some cases, ingestion of a large excess of micronutrients leads to toxicity.

Burns

Severe burns provide one of the most powerful catabolic stimuli. The injury response may persist for weeks or months during the period of wound healing. It is not surprising that the nutritional requirements of such patients are greater than for most other catabolic states (see Fig. 4). Inappropriate nutritional support to metabolically unstable patients (burned or non-burned subjects) may cause severe metabolic disturbances: thus before aggressive nutritional support is started, shock and acid–base disturbances should be at least stabilized.

With minor burns a normal oral intake, with or without supplements, is all that is required. However, with severe burns, anorexia is frequently severe and prolonged, and artificial nutritional support is usually required.

Feeding through a nasogastric tube is often well tolerated. However, the use of enteral tube feeding may be restricted early after burns because of gastric stasis or ileus. In most patients gastrointestinal motility improves within 2 to 4 days, so that more nutrients can be administered. Nevertheless, it is important to aspirate the gastric contents intermittently, especially in the early phase after burns, to ensure that there is no gastric pooling. The fluid and electrolyte requirements are often considered in association with nutritional needs, and frequently extra fluid is drunk or given through the enteral feeding tube or an intravenous line, to match the increased fluid losses from the skin surface. Some units routinely provide extra micronutrients, but the extent to which vitamin and trace element requirements are increased after burns is poorly defined.

In patients with severe burns, and those with other associated problems such as intra-abdominal injuries and sepsis, smoke inhalation, and multiorgan failure resulting in artificial ventilation, ileus often persists. In others, the gastrointestinal tract may not tolerate sufficient enteral nutrition to meet the increased requirements of the patient. Some patients, particularly those with severe burns, may thus need both enteral and parenteral nutrition. The enteral feeding is to be encouraged because it may maintain gut integrity, prevent bacterial translocation from the gut into the systemic circulation, and speed up the transition from parenteral nutrition to normal oral intake.

Head injury

Head injury provides another important indication for artificial nutritional support, especially since many patients have an impaired level of consciousness for prolonged periods.

Head injury frequently coexists with other major injuries, with the result that there is a severe hypermetabolic and catabolic response (for example, the negative nitrogen balance may be greater than 20 g N/day in the first week after injury even when a limited intake of nutrients is provided).

Immobility contributes to the loss of muscle bulk and weight loss can occur very rapidly. One of the main constraints to the use of enteral tube feeding is delayed gastric emptying, which frequently lasts for more than 10 days after severe head injury. This may be due to the injury alone but it may also be due to associated abdominal trauma. The use of traditional nasogastric feeding techniques in such patients often makes it difficult to provide the full nutritional requirements within 1 to 2 weeks. The impaired level of consciousness and poor or absent swallowing reflex also means that gastric pooling and aspiration are not uncommon. Parenteral nutrition has been used routinely in some centres but more recently increasing emphasis has been placed on enteral feeding, partly because it may prevent mucosal atrophy and bacterial translocation from the gut, and partly because parenteral nutrition is associated with a number of potentially serious complications, such as catheter-related sepsis (see above). Therefore, attempts have been made to improve methods of delivering nutrients enterally. This includes the use of pump-assisted delivery and the use of metoclopramide and erythromycin to stimulate gastric emptying. Attempts have also been made to introduce feeds directly into the small intestine, either endoscopically or under radiological control. Bypassing the stomach can allow sufficient delivery of feed to be achieved, but displacement of the tube back into the stomach may occur. Another way of dealing with the problem of gastric stasis is to place jejunostomy tubes, using a laparoscopic percutaneous procedure. Preliminary results of feed tolerance in non-randomized studies of injury are encouraging but further work is needed, particularly since the tubes may be regurgitated back into the stomach.

In patients in whom enteral feeding is not successful, for example when there is poor gastric emptying, difficulties in placing a nasojejunal tube, or severe feed-induced diarrhoea, parenteral nutrition is indicated.

Acute pancreatitis

In patients with acute pancreatitis, malnutrition may arise because of anorexia, prolonged ileus, and catabolic complications such as necrotizing pancreatitis (in around 15 per cent of patients) and pancreatitic infection (in around 4 per cent of patients). There is also concern that enteral feeding will stimulate pancreatic exocrine function and precipitate further attacks of pancreatitis or delay recovery from the initial attack. Therefore, parenteral nutrition was routinely advocated for severe and complicated cases of acute pancreatitis. Parenteral nutrition continues to have an important role in the management of acute pancreatitis, especially when it is complicated by enteric fistulae or when enteral nutrition cannot be used to administer sufficient nutrients to meet needs. However, the recent trend is to use less parenteral nutrition and more delivery of nutrients directly into the jejunum. This trend has come about because there is no evidence that parenteral nutrition improves clinical outcome compared with no nutritional support in patients with mild to moderate pancreatitis. There is also no evidence that jejunal feeding adversely affects outcome by stimulating exocrine pancreatic secretions. Indeed, three recent randomized controlled trials suggest that early nasojejunal feeding (2 days after the onset of pancreatitis) has advantages over parenteral nutrition with respect to the severity of subsequent disease (Ranson criteria, in a study of predominantly mild acute pancreatitis, and other indices of the inflammatory response in a study of more severe pancreatitis), and septic complications (in studies of severe pancreatitis). Enteral tube feeding is more physiological and cheaper. Although clinical attitudes are changing, it is necessary to confirm by large randomized controlled studies that jejunal feeding attenuates the inflammatory response and improves clinical outcome to a greater extent than parenteral nutrition. In practice it is also necessary to ensure that there are adequate facilities for placing nasogastric tubes (for example radiologically or endoscopically), and for dealing with tubes that coil back into the stomach.

Transplantation

Artificial nutritional support is variably required for patients requiring a transplanted organ. With renal transplantation, where the gastrointestinal tract is not affected unless surgical complications occur, it is usual to eat normally shortly after the surgery.

Patients with liver or heart plus lung transplants are usually artificially ventilated in the intensive care unit after transplantation. Artificial nutritional support is often given during this period, especially in those with pre-existing malnutrition or postoperative complications. In those with substantial malnutrition, attempts should be made to improve nutritional status before the procedure.

Artificial nutritional support may also be required following treatment with cytotoxic drugs or radiotherapy. In bone marrow transplantation the use of aggressive cytotoxic therapy or radiotherapy may result in inflammation of the mucous membranes of the gastrointestinal tract from mouth to rectum beginning a few days after cytotoxic therapy. Swallowing then becomes painful and diarrhoea a problem. The severity of this mucositis may limit the dose of cytotoxic drugs used. A nasogastric tube is often uncomfortable to such patients, and bleeding may occur from friction with the inflamed mucosa, especially in those with thrombocytopenia, and parenteral nutrition is often required.

There is much research interest in the use of specific nutrients or bioactive substances to protect the mucosa of the gut from damage by cytotoxic drugs or radiotherapy but these have not found their place in routine clinical practice. Graft-versus-host reactions may also be accompanied by gastrointestinal symptoms. Severe, prolonged, watery diarrhoea, amounting to several litres a day, presents a serious problem. Parenteral nutrition is usually necessary after intestinal transplantation, partly because the pre-existing bowel disease produces malnutrition (many patients receiving bowel transplants are on long-term parenteral nutrition prior to the transplant), and partly because it is necessary to ensure that sufficient time has been allowed after transplant surgery to ensure that anastomoses have adequately healed and that mucosal integrity has recovered. Fluid electrolyte and trace elements may need to be given in increased amounts to balance the increased loss associated with gastrointestinal effluents. The protein and energy requirements of patients receiving transplants can be calculated according to the scheme indicated in Fig. 4.

Perioperative nutrition

Although malnutrition may be present before to elective surgery, it is more likely to occur postoperatively, especially in those with complications. Many studies have been undertaken to assess whether perioperative artificial nutritional support reduces the complication rate after surgery. They have produced conflicting results, with some suggesting improvements, others no significant effect, and in yet others an increase in infective complications (such as catheter-related sepsis) when parenteral nutrition is used. These conflicting results have occurred at least partly because of the multiple other factors that affect the outcome of surgery (age, sex, severity of disease, skill of the surgeon, nursing care, presence or absence of a nutrition team, and preoperative nutritional status). The results are also affected by the type of nutritional support provided. Preoperative enteral tube feeding has been reported to improve clinical outcome, but few studies of this exist. Postoperative oral supplements and supplementary enteral tube feeding have been have been reported to improve clinical outcome in malnourished patients with a fractured hip or femur. Postoperative supplements and enteral tube feeding in more well-nourished individuals has produced clinical benefit in only some studies. Perioperative parenteral nutrition has been more controversial.

The largest multicentre trial of perioperative parenteral nutrition (The Veteran Affairs Total Parenteral Nutrition Co-operative Study Group) involved several hundred patients undergoing elective abdominal and thoracic surgery. The routine administration of parenteral nutrition (from up to 2 weeks before surgery and at least 3 days after surgery) provided no overall benefit. However, in the subgroup of patients who were severely malnourished, perioperative nutritional support decreased the non-infective complications from 42 per cent to between 5 and 23 per cent (depending on the method used to define malnutrition). This and other studies emphasize the importance of patient selection for nutritional support. It is obvious, for instance, that other patient groups, such as those with prolonged ileus, massive bowel resection, fistulae, or recurrent severe pancreatitis, are likely to benefit from parenteral nutrition. Meticulous fluid and electrolyte balance is of major importance (see Table 3) in those with fistulae, nasogastric aspirates, or multiorgan failure.

Patients with prolonged anorexia and intra-abdominal sepsis are also likely to benefit from nutritional support. Parenteral nutrition may be required for long periods, for instance in the patient in whom the fluid output from a fistula gradually decreases in the absence of oral food intake, or after massive intestinal resection (short bowel syndrome) when intestinal adaptation may take weeks or months. In some patients with the short bowel syndrome (less than 25 cm of the small intestine remaining) such support will be required indefinitely.

Home nutritional support

One of the most important recent developments in clinical nutrition is the use of artificial nutrition in the community. In many developed countries there is now more tube feeding taking place in the community than in hospital. Furthermore, the increasing emphasis on home enteral tube feeding is likely to continue. Home parenteral nutrition is also increasing, but it is used there much less than in hospitals. In low-income countries the trend towards home care has been limited by the lack of an adequate infrastructure and organization, which are necessary to train, discharge, and monitor patients, and to ensure that the feeds and accessories are delivered to the home regularly and reliably.

Artificial nutritional support at home has a number of advantages over treatment in hospital. Patients frequently feel more comfortable in the familiar home environment, where many of them not only care for themselves but also for other family members. Affected adults may frequently go to work and children attend school. The treatment of patients at home is cheaper and frees beds for the use of other patients. Treatment at home does involve a major commitment on the part of the patient or carer and serious complications may arise.

Indications

Since most malnutrition occurs in the community, it is necessary to identify patients who are likely to benefit from simple dietary measures, the use of mixed macronutrient supplements, and home parenteral nutrition or home enteral tube feeding. The type of treatment depends on whether the gastrointestinal tract is available for the digestion and absorption of nutrients. In the case of artificial nutritional support, it is necessary to ensure that the patients or carers are able to perform the necessary tasks to a sufficiently high standard.

Food and oral supplements

When malnourished patients do not respond adequately to simple dietary advice (for example frequent ingestion of appetizing high energy density meals or snacks), mixed solid or liquid macronutrients can be given. A systematic review of 84 trials of oral nutritional supplements in the community concluded that the supplements were much more likely to produce benefit in patients with a body mass index of less than 20.0 kg/m^2 than in those with a body mass index of more than 20 kg/m^2. Furthermore, in patients with a body mass index of less than 20 kg/m^2 the supplements largely added to oral intake, whereas in those with a body mass index of more than 20 kg/m^2 they largely replaced oral food intake. The benefits varied according to the disease. In patients with chronic obstructive airways disease the supplements increased muscle strength, walking distance, and well being. In children with cystic fibrosis they improved growth, and in elderly subjects they reduced the number of falls and increased activities of daily living. In patients with HIV infection they improved immune function tests. In general improvements were not noted unless body weight increased by more than 5 per cent. Supplements were found to be of little or no value in patients who had a body mass index of more than 20 kg/m^2.

Home parenteral nutrition

Intestinal failure due to Crohn's disease (with or without fistulae) is the most common indication in the United Kingdom. The short bowel syndrome, motility disorders (such as scleroderma, and pseudo-obstruction), congenital bowel disease, and radiation enteritis are also important indications. Home parenteral nutrition has also been used in patients with malignancy (usually those with intestinal obstruction) and in patients with AIDS who are unable to tolerate enteral nutrition.

Home enteral tube feeding

The principal indications are those associated with swallowing difficulties. These may be obstructive (such as malignancy of the upper gastrointestinal tract) or non-obstructive and due to neurological disorders (for example strokes, Parkinson's disease, motor neurone disease, multiple sclerosis, and primary muscle diseases that affect swallowing). In children, although a neurological disorder of swallowing is an important indication, chronic anorexia leading to failure to thrive is perhaps more common (due to, for example, congenital malformations, severe cystic fibrosis, inborn errors of metabolism, and some gastrointestinal disorders such as Crohn's disease).

Age distribution

Many patients receiving enteral nutrition at home are elderly: this is not surprising since conditions such as stroke, motor neurone disease, and oesophageal malignancy typically occur in the elderly. Home tube feeding is also relatively common in children, mainly in those aged less than 10 years. This age distribution has particular implications for home care. Many elderly people are unable to care for themselves because of weakness, immobility, arthritis, poor eyesight or hearing, etc., and therefore a carer, usually a family member, has to be identified. Similarly, carers are frequently required for children on home enteral nutrition.

The age distribution of patients on home parenteral nutrition is different, partly because the indication is often Crohn's disease, which occurs predominantly in subjects aged 20 to 50 years.

Management

The principles of artificial nutritional support at home are similar to those in hospital. Management begins with nutritional screening, so that patients with malnutrition or at high risk of malnutrition are identified, as well as the underlying risk factors (see section on nutritional screening and assessment). Nutritional goals are set, the underlying psychosocial and physical problems (active disease) are addressed, and nutritional counselling and support are provided. Normal food is encouraged whenever possible, but when this is inadequate supplements may be used. This type of nutritional support is normally initiated in the community, whereas home enteral and parenteral nutrition is usually initiated in specialist centres. Special considerations apply to patients receiving home enteral and parenteral nutrition.

Psychological evaluation

The thought that artificial nutritional support may have to be given at home for months or years may surprise some patients and they may find it difficult to accept the concept. Nevertheless, given appropriate support, as the patient or carer gains confidence their fears and anxieties frequently subside. It is always essential to involve family members (or the carer) as well as the patient. Contact with patients who are already on home treatment may do much to reassure. School-age children may also have particular difficulties in coming to terms with this form of therapy, but careful counselling and the use of nocturnal feeding alone frequently allows them to adjust, attend school, and lead a reasonably normal life.

Training

The training should be supervised only by those experienced in the field. Despite pressure for hospital beds, discharge should not take place until it is clear that the patient is adequately skilled and appropriate arrangements have been made at home.

Patients (or carers of the patients) requiring artificial nutrition at home should learn the basic principles of nutritional support and of asepsis. They should know how to program the infusion pump that delivers the nutrients, how to add solutions to the feed (if required), how to connect and disconnect the feeds to the catheter/tube, how to change dressings, how to recognize problems associated with feeding, such as a blocked catheter and infection, and how to recognize hyper/hypoglycaemia, measure blood glucose, and screen for glycosuria. Training is often helped by audiovisual aids and written instruction, and reinforced by repeated practice. A trial of home nutritional support over a weekend may be a useful way of assessing the patient's ability to adjust and cope.

Evaluation of the home environment

An assessment of the home environment is essential before discharge. There must be space available for the storage of feeds and accessories. A refrigerator is usually necessary for storage of parenteral nutrition solutions and drugs. Modifications to the home can be made to allow routine activities to be carried out more efficiently. A wheelchair or other extra equipment may be very useful for some patients, such as a bed harness for very disabled patients and adjustable V-shaped boards for children with cystic fibrosis who require physiotherapy and postural drainage of lung secretions.

Financial arrangements

Financial arrangements for home artificial nutrition clearly vary from country to country and in different parts of the same country. A clear statement about finances should be made prior to discharge. In many countries patients and/or carers are entitled to some sort of financial or other support. These should be made known to those entitled to receive them.

Written instructions and follow-up arrangements

Prior to discharge all patients should have written instructions for the routine procedures for home nutritional support and how to recognize and act when complications arise. Patients/carers should also have the telephone number of the appropriate health professional to contact in an emergency, on a 24-h basis.

Outcome

The outcome of home nutritional support varies considerably, depending on the underlying condition and the initial selection criteria.

Home parenteral nutrition

The most extensive analysis of outcome of patients on home parenteral nutrition has come from North America (Fig. 5). For a variety of conditions (congenital bowel disease, Crohn's disease, motility disorders of the gut, and radiation enteritis), there is a substantial mortality in the first 2 years (10–30 per cent), but few deaths occur after this period. The mortality is usually due to the underlying condition, although a few deaths arise from complications of parenteral nutrition. Mortality in patients receiving home parenteral nutrition for AIDS and malignancy has been reported to be high (Fig. 5). For example, in one survey the mortality of AIDS patients receiving parenteral nutrition in the pre-antiprotease era was as high as 93 per cent in a year. This mortality in AIDS and malignancy clearly depends on when nutritional support is started in relation to the stage and severity of the disease. Such a high mortality dictates that such treatment should only be offered to those in whom there is good reason to expect a substantial consequent improvement in the quality of life.

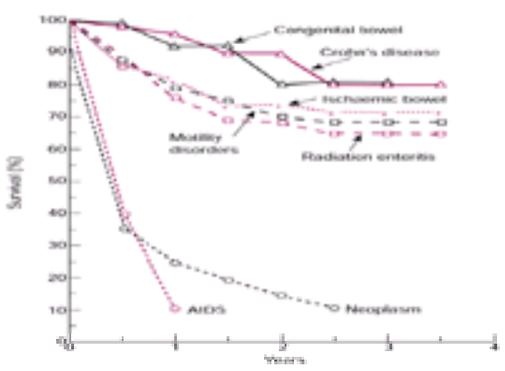

Fig. 5 Survival of different groups of patients receiving home parenteral nutrition (from Howard *et al.* 1991).

In that context in general, the British Artificial Nutrition Survey has revealed that 71 per cent of patients on home parenteral nutrition undertook full normal activity, 13 per cent limited activity, and the remaining 16 per cent were housebound. Seventy-one per cent were independent, 19 per cent required some help or total help, and only 10 per cent of patients were severely disabled and heavily dependent on others.

Home enteral tube feeding

The British Artificial Nutrition Survey has provided outcome data on patients receiving home enteral and parenteral nutrition (Table 7). Only 62 per cent of patients on home enteral tube feeding were continuing at 1 year. Twenty-two per cent died whilst receiving tube feeding, but the total mortality is likely to have been greater because some patients almost certainly died after stopping tube feeding. However, there was great variability in the outcome, which was related to age, and both type and stage of disease.

Despite the high mortality, many family members feel that the quality of life of affected individuals frequently improves, or does not deteriorate. They appreciate and enjoy being with the affected family member, even in the face of a poor prognosis. In a survey of children receiving home tube feeding, predominantly for failure to thrive, parents reported greater freedom because they did not have to spend so much time trying to feed their children as previously. They also reported that their children were happier and more active than before tube feeding.

Monitoring

Patients on home nutritional support should be seen at regular intervals, at a frequency determined by the needs of the patient or carer and the stability of the clinical condition. It is necessary to assess whether the initial goals have been met, and if not a decision has to be made whether new goals should be set or the management changed in an attempt to meet the initial goals. Functional outcome measures, such as walking distance in patients with respiratory or motor disabilities, or eating capabilities, are particularly important. For example, patients with swallowing problems may improve and return to normal oral feeding. If swallowing function is not assessed intermittently in certain patient groups (for example those with strokes) home enteral tube feeding may continue unnecessarily. Nutritional assessment is often based on changes in growth or body weight (taking into consideration the presence or absence of oedema), but changes in skinfold thicknesses and arm muscle circumference and a variety of biochemical parameters may also be used (for example urinary creatinine in accurately collected 24-h urine specimens is a reasonable index of muscle mass).

The need for blood counts, urea and electrolytes, and liver function tests varies, depending on the clinical situation. They are more often required for those taking parenteral nutrition. Measurements of trace elements and vitamins are sensible for patients receiving long-term artificial nutrition, particularly those receiving parenteral nutrition.

Ethical considerations

Legal and ethical considerations about indications and when to start and terminate home nutritional support vary from country to country. Furthermore, the law prohibits discontinuation of nutritional support in some states in America but not in others. Particularly difficult ethical issues concern patients with dementia who stop eating for unknown reasons, and those who are unconscious or have very severe disabilities that are likely to deteriorate.

The role of nutrition teams in the management of patients in hospital and at home

Nutritional support is required for a wide range of patients distributed in different wards. A nutrition team, consisting of a clinician, a specialist nurse, dietitian, pharmacist, and possibly others such as a chemical pathologist and bacteriologist, can advise, supervise, and co-ordinate the management of patients, and maintain high standards of care. It has been shown repeatedly that such teams minimize the incidence of complications associated with nutritional support, avoid unnecessary nutritional support, and reduce wastage of feeds. The reduction in the incidence of catheter-related sepsis alone, from a rate of 25 per cent to 3 per cent has substantial economic implications, since one episode of catheter-related sepsis in a British hospital is currently estimated to cost between £1500 and £5000. Nutrition teams also have an important role to play in training patients for home support, providing advice, co-ordinating the supply of feeds and other equipment to the home, and supervising follow-up.

Further reading

American Medical Association (1979). Guidelines for essential trace element preparations for parenteral use. A statement by the nutrition advisory group. *Journal of the American Medical Association* **241**, 2051–4.

American Medical Association Department of Foods and Nutrition (1979). Multivitamin preparations for parenteral use. *Journal of Enteral and Parenteral Nutrition* **3**, 258–62.

American Society of Parenteral and Enteral Nutrition Board of Directors (1987). Guidelines for the use of enteral nutrition in the adult patient. *Journal of Parenteral and Enteral Nutrition* **11**, 435–9.

Elia M (1993). Artificial nutritional support in clinical practice in Britain. *Journal of the Royal College of Physicians* **27**, 1–15.

Elia M (1994). Home enteral nutrition: general aspect and a comparison between the United States and Britain. *Nutrition* **10**, 1–9.

Elia M, Jebb SA (1994). Nutrition. *Medicine International* **22**, 381–420.

Elia M *et al.* (2001). *Trends in artificial nutrition in the UK during 1996-2000*. British Association for Parenteral and Enteral Nutrition (BAPEN), Maidenhead, UK.

Frisancho AR (1981). New norms of upper limb fat and muscle areas for assessment of nutritional status. *American Journal of Clinical Nutrition* **34**, 2540–5.

Frisancho AR (1984). New standards of weight and body composition by frame size and height for the assessment of nutritional status of adults and the elderly. *American Journal of Clinical Nutrition* **40**, 808–19.

Green CJ (1999). Existence, causes and consequences of disease-related malnutrition in the hospital and the community, and clinical and financial benefits of nutritional intervention. *Clinical Nutrition* **18** (Suppl. 2), 3–28.

Howard L *et al.* (1991). Four years of North American Registry home parenteral nutrition outcome data and their implications for patient management. *Journal of Parenteral and Enteral Nutrition* **15**, 384–91.

Matarese LE, Gottlschlich MM (1998). *Contemporary nutrition support practice: a clinical guide*. WB Saunders, Philadephia.

National Research Council (1989). *Recommended dietary allowances*, 10th edn. National Academy Press, Washington, DC.

Rombeau JL and Rolandelli RH (1997). *Enteral and tube feeding*. WB Saunders, Philadelphia.

Shenkin A, Wretlind A (1977). Complete intravenous nutrition including amino acids, glucose and lipids. In: Richards JJ, Kinney JM, eds. *Nutritional aspects of care in the critically ill*, pp 345–65. Churchill Livingstone, Edinburgh.

Stratton R, Elia M (1999). A critical systematic analysis of the use of oral nutritional supplements in the community. *Clinical Nutrition* **18** (Suppl. 2), 29–84.

Taylor S, Goodinson-McLaren S (1992). *Nutritional support: a team approach*. Wolfe Publishing, London.

The Veteran Affairs Total Parenteral Nutrition Co-operative Study Group (1991). Peri-operative total parenteral nutrition in surgical patients. *New England Journal of Medicine* **325**, 525–32.

Wilcock H, Armstrong J, Cottee S (1992). Artificial nutrition in a health district with particular reference to tube feeding. *Health Trends* **23**, 93–100.

11.1 The inborn errors of metabolism: general aspects

Richard W. E. Watts

There are around three to four thousand known unifactorially inherited diseases, that is familial diseases, the inheritance of which can be described as being autosomal recessive, autosomal dominant, sex-linked recessive, or sex-linked dominant (Mendelian inheritance). The inborn errors of metabolism are those inherited diseases in which the phenotype includes a characteristic constellation of chemical abnormalities that can be ascribed to an alteration in the catalytic activity of a single specific enzyme. There are unifactorially inherited diseases in which the current techniques are too insensitive for a chemical abnormality to be identified, so that the syndrome has to be defined in clinical, gross structural, and/or pathological terms; further study may bring these into the category of inborn errors of metabolism.

Almost all the unifactorially inherited diseases arise from mutations in the nuclear genome which spans about three billion base pairs of deoxyribonucleic acid (**DNA**). A few mitochondrial proteins have their structures encoded in the mitochondrial DNA (**mtDNA**). This genetic information is transmitted only through the female line and the category of inborn errors of metabolism includes this group. Both the nuclear and the mitochondrially inherited diseases stem from single mutations within a cistron (the functional unit of DNA) which directs the synthesis of a single specific polypeptide chain. The molecular changes in the enzyme protein may affect the primary, secondary, tertiary, or quaternary structure, decreasing, increasing, or abolishing its catalytic activity. Some mutations affect the function of an activator protein, others reduce the binding of hormones and paracrine factors to cell surfaces and/or subcellular structures, and some derange the migration of proteins within cells; another group impairs the transport of metabolites across cellular and subcellular membranes (Table 1). Most intracellular enzymes are located in the cytosol where they are correctly orientated in relation to one another, sometimes as macromolecular complexes, and to their substrates. Some are bound to cellular and subcellular membranes and a minority are located in anatomically defined subcellular structures or organelles: the mitochondria, lysosomes, and peroxisomes.

Mitochondrial diseases

The mitochondrial genome is a circular double strand containing 16.5 kilobases of DNA. It encodes 13 of the respiratory chain enzymes the remainder of which, about 60, are encoded in the nuclear DNA. Abnormal mitochondrial function impairs the supply of energy for biochemical work in all tissues and therefore has wide-ranging effects. Each mitochondrion also contains 24 RNA genes that participate in intramitochondrial protein synthesis. Transcription and translation of mtDNA are regulated by the nucleus through the non-coding D-loop region of the mitochondrial genome. Human cells contain about a thousand copies of mtDNA, but the individual mitochondria in a cell may not all carry a given specific mutation and different cells carry different proportions of mutated mitochondria (heteroplasmy). The proportion of mutant mtDNA must exceed a critical level before the mitochondrial respiratory chain disease declares itself. This variability, as well as tissue-specific differences in dependence on oxidative metabolism, explains, at least partially, why some tissues are preferentially affected in patients with mtDNA diseases. Postmitotic tissues (e.g. neurones, muscle, endocrine tissues) have high levels of mutated mtDNA and are often clinically affected, whereas rapidly dividing tissues (e.g. bone marrow) are less often clinically affected. Differences in the proportions of mutated and non-mutated mtDNA between and within family members also contribute to the wide phenotypic range encountered in the mitochondrial diseases. The spermatozoal cytoplasm, including its mitochondria, is entirely lost at fertilization and, for this reason, mitochondrial diseases are only transmitted through the female line. Clinically affected women rarely transmit a mtDNA deletion to their children. However, a woman with a heteroplasmic mtDNA point mutation or duplication may transmit a variable amount of mutated mtDNA to her progeny. The number of mtDNA molecules in each oocyte is reduced and then amplified to a total of about 10^5 during early development of the oocyte; this, presumably random, process contributes to the different amounts of mutated mtDNA in different children in the same family. Women whose gametes contain high concentrations of mtDNA are more likely to have clinically affected children than mothers with lower levels of mtDNA. The general clinical manifestations of the mitochondrial diseases are shown in Table 2 and specific examples of mitochondrial diseases are given in Table 3.

Peroxisomal diseases

Some enzymes that are encoded in the nuclear DNA are specifically expressed in peroxisomes, to which they are imported soon after translation. Mutations in the relevant genes result in the diseases listed in Table 4.

Lysosomal storage diseases

Lysosomes are subcellular organelles containing hydrolases with low optimum pH values ('acid hydrolases') which catalyse the degradation of macromolecules. The macromolecules are either derived from the metabolic turnover of structural cellular components or have entered the cell by endocytosis. The products of this macromolecular degradation process leave the lysosomes by specific efflux processes.

In most of the lysosomal storage diseases an inborn error of metabolism affects a specific lysosomal enzyme so that either undegraded or partially degraded macromolecules accumulate in the lysosomes. The engorged lysosomes distort the internal architecture of the cell, disturb its function, and inhibit the activities of other lysosomal enzymes so that macromolecules other than those related to the primary enzyme deficiency also accumulate.

Cystinosis (cystine storage disease) and Salla disease (*N*-acetylneuraminic (sialic) acid storage disease) are due to metabolic lesions involving the specific efflux processes whereby these two low molecular weight products of macromolecule metabolism (cystine and sialic acid respectively) leave the lysosome (Table 1).

Lysosomal enzymes are glycoproteins which are subject to exocytosis and re-uptake by endocytosis. Their protein moieties are synthesized on the rough endoplasmic reticulum and the oligosaccharide side chains are added in the Golgi apparatus. The addition of a terminal mannose-6-phosphate residue recognition marker is necessary if the enzyme molecule is to be correctly routed into the lysosomes, and if it is to be available for receptor mediated re-uptake from the interstitial fluids. The types of lysosomal storage diseases and the nature of their metabolic defects together with examples of each group are presented in Table 5.

Heterogeneity in the inborn errors of metabolism

The individual inborn errors of metabolism are defined on the basis of the phenotype, including the specific enzyme lesion, and by their pattern of inheritance. Close study of any particular inborn error of metabolism reveals unexpected heterogeneity. This is due to:

- Multiple allelism.
- Mutations at different gene loci affecting the structure of different polypeptide chains in a single enzyme protein.
- Mutations at different gene loci affecting different proteins with similar catalytic functions.
- Differences in the overall genetic background against which the single mutation acts.
- Environmental factors.

Clinical pointers towards a diagnosis of an inborn error of metabolism

Although the symptoms of metabolic disease may appear vague and protean, and an inherited disease cannot be diagnosed in the absence of an appropriate family history, some clinical settings suggest the presence of an inborn error of metabolism (Table 6). In taking the family history special inquiries should be made about affected siblings, possible parental consanguinity, paternity, miscarriages, perinatal deaths, abortions, about the sexes of possibly affected relatives and their placement on the maternal or paternal side of the family, the ages at death of relatives, as well as the ethnic and geographical origins of the parents.

General approaches to the treatment of inborn errors of metabolism

The treatments available for the individual inborn errors of metabolism cover a wide range and may need to be specially developed for individual patients. However, the principles involved can be broadly classified as in Table 7. Palliative surgical and other measures may be needed to deal with specific complications (for example corneal grafting to restore vision in patients with corneal clouding due to one of the mucopolysacchridoses). Consideration should also be given to meeting the educational and social needs of these patients as well as to optimizing their overall clinical state and correcting the biochemical parameters. The successful management of patients with inborn errors of metabolism requires a multidisciplinary approach which utilizes the special skills of dieticians, social workers, educationalists, and occupational therapists as well as those of physicians, surgeons, biochemists, and geneticists. It is particularly important to plan for the handover of specialist care from the paediatrician to the most appropriate adult physician when follow-up in a paediatric department becomes inappropriate. The perfect outcome is to achieve a physically and mentally normal adult who is capable of begetting normal children. Unfortunately the nature of many of the inborn errors of metabolism mitigates against the attainment of this ideal so that treatment has to aim at optimizing the child's potential in all its physical, mental, and social aspects. Treatment and support also have to be extended to the parents and siblings who, if not overtly affected themselves, may be carriers of the abnormal gene concerned and require appropriate advice about the genetic and other aspects of the disease.

The ability to clone human genes into bacteria and yeasts which can then produce large amounts of the human gene product is widening the horizons for future treatment by enzyme administration. The development of macrophage-targeted b-glucocerebosidase enzyme replacement therapy for Gaucher disease (glucosylceramidase deficiency) type I is a notable recent development in this field and is now regarded as the definitive line of treatment. Attempts to utilize transplanted fibroblasts and amniotic cells as a source for enzyme replacement therapy have not been successful. Bone marrow transplantation has been used for the treatment of two groups of inherited metabolic disorders: those in which it is desired to replace a particular type of non-functioning bone marrow cell by its normally functioning counterpart and those in which an attempt has been made to utilize the fact that the bone marrow produces 50 to 100 g of polymorphonuclear leucocytes per day and that these cells exocytose (release) their lysosomal enzymes for endocytic uptake by enzyme-deficient cells in the body tissues generally. This strategy has been more successful with the first group of diseases, which includes disorders of neutrophil function (e.g. cyclic neutropenia), functional abnormalities of lymphocytes, and osteopetrosis. The beneficial effect on the last disease is due to the introduction of normal osteoclast precursors. The results in the second group of diseases, namely those in which the white cell lineage derived from the transplanted bone marrow is used to supply normal enzyme to enzyme-deficient tissues, for example Hurler disease (mucopolysaccharidosis IH) and metachromatic leucodystrophy, have been less successful particularly in terms of neurological function. Haemopoietic stem cells have been implanted into the fetus *in utero* to correct severe congenital immunodeficiency but this has not, so far, been applied to diseases without immunodeficiency. This procedure takes advantage of the immunological tolerance of the fetus. The possibility of using liposomes and resealed erythrocyte envelopes as carriers of therapeutic enzymes is also being explored.

Liver transplantation is used as a sophisticated form of enzyme replacement therapy in some inborn errors of metabolism where this organ is the specific site of the metabolic lesion. Liver transplantation has the advantage that the enzyme is introduced in the correct organ, the correct cell with its correct subcellular location, and correctly orientated with respect to its substrate and other enzymes with which it must act in concert. Liver transplantation can also be regarded as a form of gene replacement therapy in that the donor liver contains the normal gene which will direct the synthesis of a normal enzyme protein. Prenatal transplantation of fetal liver stem cells has potential in the treatment of some inborn errors of metabolism. Successful engraftment at the 12th to 24th week postfertilization with partial correction of the metabolic defect has been demonstrated in b-thalassemia. Treatment by gene replacement, using retroviral vectors and gene constructs to introduce the desired DNA sequence into the patient's explanted haemopoietic stem cell genome, these genetically corrected cells being cultured and then returned to the patient's circulation, may have some potential in diseases where expression of the metabolic lesion in the haemopoietic system determines the phenotype, or in those situations where genetically corrected migratory cells of haemopoietic origin can deliver normal enzyme to the enzyme-deficient tissues. Although somatic cell gene therapy possibly using viral vectors and/or gene constructs to introduce the desired DNA sequences into other cell types is currently being investigated extensively in *in vitro* model systems and in animal models of some human inborn errors of metabolism using, for example, hepatocytes, none of these have reached application in clinical practice. For example, the possibility of using herpes simplex virus type 1 as a means of introducing corrected genes into the nervous system is being explored. Another approach is to use either resealed erythrocyte cell membranes or liposomes as carriers of therapeutic enzymes. Thus, although there are some prospects of correcting some enzyme defects in the somatic cell genome, the correction of defects in the germline seems remote although the development of advanced *in vitro* fertilization techniques, preimplantation DNA analysis, gene transfer, insertion or conversion, and embryo implantation procedures may render this judgement premature. Ethical objections to human germline modifications are also being raised, and could lead to this work being discontinued.

Screening for inborn errors of metabolism

The realization that very early diagnosis is essential in order to achieve good results in the treatment of some inborn errors of metabolism, such as phenylketonuria and galactosaemia, has stimulated interest in the possibility of examining either whole populations or selected groups of predisposed individuals for the biochemical differences which characterize particular inherited metabolic diseases. Diagnosis is needed at a stage which is not only presymptomatic but which precedes the onset of self-perpetuating secondary pathological changes.

Screening for inborn errors of metabolism may be either non-selective (whole population) or selective. The latter, which includes carrier detection studies, aims to cover a part of the population. This may be defined on clinical, genetic, ethnic, or geographical grounds. Phenylketonuria and congenital hypothyroidism are the only members of this group of disorders for which neonatal whole-population screening is generally practised, although the inclusion of galactosaemia, cystic fibrosis, and congenital adrenal hyperplasia (21-hydroxylase deficiency) has been proposed. Whole-population screening should only be established for treatable or preventable diseases, and the consistency of the association of the proposed biochemical or other marker and the serious clinical phenotype must have been proved beyond any doubt. There must be a reliable and robust analytical method suitable for use with a sample of blood or urine which can be obtained without distressing either the parents or the baby. The possibility that metabolic screening will bring to light previously unrecognized variants, which are either mild and do not require treatment, or which by virtue of a fundamentally different biochemical lesion will resist the currently established therapies, has to be borne in mind. Phenylketonuria illustrates these problems. Here, beside classical phenylketonuria, whole-population screening has identified both the clinically unimportant essential (mild) hyperphenylalaninaemia, and the devastatingly serious, but treatable, inborn errors of tetrahydrobiopterin synthesis which produce the 'malignant' hyperphenylalaninaemia syndrome. It is also possible that in some cases immediate postnatal screening and treatment may be too late to prevent minor manifestations of the disease, (e.g. in congenital hypothyroidism).

The incidence of disease which merits whole-population screening should be at least similar to that of phenylketonuria in Caucasians (between 1 in 6000 and 1 in 12 000). Cystic fibrosis has an incidence of 1 in 2500 (gene frequency l in 25) in Caucasians and would merit neonatal whole-population screening on this basis. Molecular genetic approaches are potentially useful. If the disease is not too genetically heterogeneous and when the full range of possible causative mutations is known the specific mutation could be sought directly. Otherwise, after DNA amplification the mutational change in the DNA structure could be detected either by the presence of a restriction endonuclease site or by probing with another primer that hybridizes with only one of the alleles. An appreciable proportion of individuals classified as being homozygotes on the basis of classical genetic analysis prove to be double heterozygotes, that is they carry two different mutations in the same gene. The number of inborn metabolic errors in which the affected individuals and the heterozygous carriers can be identified by molecular genetic analysis is increasing rapidly. It includes such numerically important diseases as sickle cell anaemia, b-thalassaemia, haemophilia, Duchenne muscular dystrophy, cystic fibrosis, and phenylketonuria, as well as rarer but devastating conditions such as the Lesch–Nyhan syndrome.

Prenatal diagnosis

The procedures used in prenatal diagnosis are:

- Direct examination of the fetus by ultrasonography and fetoscopy.
- Chemical analysis of amniotic fluid.
- Biochemical and cytological analysis of cultured amniotic cells (amniocytes) obtained by amniocentesis at the fifteenth to sixteenth week of pregnancy.
- DNA analysis on uncultured amniocytes.
- Karyotypic enzymological and DNA analysis of chorionic villi obtained by biopsy at the eighth to tenth week of pregnancy.

- Biochemical studies on tissue obtained by biopsying the fetus *in utero*.

Carrier state diagnosis

Carriers are either individuals carrying the gene for a recessive disorder, which does not express itself in the heterozygous state (e.g. phenylketonuria), or those who carry the gene for a dominant disorder, that is one which does express itself in the heterozygous state, but in which symptoms occur in later life (e.g. Huntington's chorea (Huntington's disease)).

The general approaches to carrier state diagnosis are:

- The detection of minor clinical, radiological, and clinicopathological abnormalities.
- The demonstration of levels of enzyme activity in tissue (e.g. leucocytes or cultured fibroblasts) which are intermediate between those observed in individuals homozygous for the abnormal and the normal forms of the enzyme respectively (the observed level of activity may not be exactly 50 per cent of the normal value).
- The demonstration of intermediate levels of a characteristic metabolite in an accessible body fluid.
- The demonstration of mosaicism with respect to the product of the mutant gene on the X chromosome in the case of sex-linked recessive disorders.
- Direct gene analysis using either a specific gene probe or a linked restriction fragment length polymorphism.

The ability to recognize asymptomatic carriers of serious recessive diseases and presymptomatic individuals in the case of dominant disorders raises major ethical and social issues with respect to the psychological impact that this information will have on the affected individuals and their families. This is especially so with the clinically normal carriers of a crippling, lethal, and untreatable disease such as Huntington's chorea.

In vitro fertilization and the inborn errors of metabolism

The human embryo produced by *in vitro* fertilization can be biopsied at a very early stage of development (i.e. at the eight-cell stage). A single cell is removed and examined for the DNA mutation responsible for the disease which the parents are known to be carrying. This enables only fertilized ova which do not carry the mutant gene to be implanted.

Animal genetic models of inborn errors of metabolism in man

Animal genetic models of the inborn errors of metabolism can be useful in the early stages of investigating new approaches to treatment before attempting to transfer these to man. It is also possible to investigate the pathophysiology of the diseases at different stages of their evolution more easily, rapidly, and predictably than if one has to rely entirely on the *ad hoc* availability of clinical and pathological material. However, there are obvious limitations when cognitive and behavioural abnormalities are part of the clinical phenotype.

Further reading

Bax BE *et al*. (1999). Survival of human carrier erythrocytes *in vivo*. *Clinical Science* **96**, 171–8 [Important advance in method for possible enzyme replacement.]

Billings PR, Hubbard R, Newman SA (1999). Human germ line modification: a dissent. *Lancet* **353**, 1873–5. [A critical review.]

Brooks DA (1997). Protein processing: a role in the pathophysiology of genetic disease. *FEBS Letters* **409**, 115–20. [A full review of the field.]

Chan L, Teng BB, Lau P (1996). Apolipoprotein B mRNA editing protein: a tool for dissecting lipoprotein metabolism and a potential therapeutic gene for hypercholesterolaemia. *Zeitschrift für Gastroenterologie* **34** (Suppl. 3), 31–2. [Example of a current tool.]

Chinnery PF, Turnbull DM (1999). Mitochondrial DNA and disease. *Lancet* **354** (Suppl. 1), S17–S21. [Comprehensive review.]

Cox TM (2001). Gaucher's disease – an exemplary monogenic disorder. *Quarterly Journal of Medicine* **94**, 399–402. [A fully up-to-date review of clinical, biochemical, and therapeutic aspects.]

Eisensmith RC, Woo SLC (1996). Gene therapy for phenylketonuria. *European Journal of Pediatrics* **155** (Suppl. I), S16–S19. [Review of situation in a disease which is a prototype for future research.]

Graeber MB, Muller U (1998). Recent developments in the molecular genetics of mitochondrial disorders. *Journal of the Neurological Sciences* **153**, 251–63. [Review of a currently expanding field.]

Haskins M (1996). Bone marrow transplantation therapy for metabolic diseases: animal models as predictors of success in *in utero* approaches. *Bone Marrow Transplantation* **18** (Suppl. 3), S25–S27. [A short review.]

Hegele RA (1997). Small genetic effects in complex diseases: a review of regulatory sequence variants in dyslipoproteinemia and atherosclerosis. *Clinical Biochemistry* **30**, 183–8.

Khanna A *et al*. (1999). Liver transplantation for metabolic liver disease. *Surgical Clinics of North America* **79**, 153–62. [Review concentrating on general principles as exemplified by hereditary haemochromatosis and Wilson's disease.]

Lachmann RH, Efstathiou S (1999). Use of herpes simplex virus type I for transgene expression within the nervous system. *Clinical Science* **96**, 533–41. [An example of a modern approach.]

Leonard JV, Schapira AHV (2000). Mitochondrial respiratory chain disorders. *Lancet* **355**, 299–304 and 389–94.

Leonard JV, Grünewald B, Clayton P (2001). Diversity of congenital disorders of glycosylation. *Lancet* **357**, 1382–3.

Lowenstein PR *et al*. (1998). Gene therapy for inherited neurological disorders: towards therapeutic intervention in the Lesch–Nyham syndrome. *Progress in Brain Research* **117**, 485–501. [Reviews strategies for gene therapy in neurological diseases as exemplified by work on the Lesch–Nyhan syndrome.]

Sandig V, Strauss M (1996). Liver-directed gene transfer and application to therapy. *Journal of Molecular Medicine* **74**, 205–12. [Review article.]

Smith AE (1999). Gene therapy—where are we? *Lancet* **354** (Suppl. I), S1–S3. [Critical appraisal of the subject.]

Surbek DV *et al*. (1997). Intrauterine transplantation of hematopoietic stem cells for therapy of genetic diseases. *Zeitschrift für Geburtshilfe und Neonatologie* **201**, 158–70. [Report of position at time of writing.]

Touraine JL (1996). *In utero* transplantation of fetal liver stem cells into human fetuses. *Journal of Haematotherapy* **5**, 195–9. [A potentially important therapeutic area.]

Vogler C *et al*. (1998). Murine mucopolysaccharidosis VII: the impact of therapies on the clinical course and pathology in a murine model of lysosomal disease. *Journal of Inherited Metabolic Disease* **21**, 575–86.

Watts RWE, Gibbs DA (1986). Animal genetic models of some inborn errors of metabolism which occur in man. In: Watts RWE, Gibbs DA, eds. *Lysosomal storage diseases: biochemical and clinical aspects*, pp.235-6. Taylor and Francis, London.

Winchester B (1999). Outlook for screening for sphingolipidoses. *Lancet* **354**, 879–88.

11.2 Inborn errors of amino acid and organic acid metabolism

P. J. Lee and D. P. Brenton

History

Following the early insights of Garrod and pioneers such as FØlling, it was the use of paper chromatography by Dent and the automated column chromatography of Moore and Stein which led to modern developments in the field of inborn errors of metabolism. The laboratory contributed more discoveries with the advent of gas–liquid chromatography and later mass spectroscopy. More recently the rise of genetics and molecular biology have revolutionized the field and now tandem mass spectroscopy is proving a powerful tool in screening and diagnosis. The inborn errors of metabolism has been a spectacularly developing field for several decades and provides vivid examples of the successful application of molecular cell biology to the diagnosis and treatment of human disease.

Introduction

An overview of amino acid metabolism and genetic defects

Humans depend upon dietary protein as a source of amino acids; some amino acids cannot be synthesized in the human body, and all are used very economically. Stool nitrogen losses are only about 1 g/day and bacterial protein accounts for much of this. Renal conservation of amino acids is extremely effective, with low clearance values (Table 1). Amino acids taken in excess of requirement are not stored but are used for energy. After the removal of the amino group for conversion to ammonia and urea (Fig. 1), the carbon skeletons degrade to major metabolic intermediates such as acetyl coenzyme A, acetoacetyl coenzyme A, pyruvate, or to citric acid cycle intermediates (Fig. 2) via individual amino acid pathways. Amino acids are referred to as glucogenic when their carbon skeletons degrade to intermediates used in gluconeogenesis and ketogenic when their degradation products can form ketone bodies. Degradative enzymes frequently have important coenzymes and inherited defects of catabolism may be due to defects of the apoenzymes or their vitamin coenzymes. Table 2 shows one biochemical classification of the genetic defects of amino acid metabolism. A clinical classification would be more practical but is difficult because of the non-specific nature of many clinical features, for example mental retardation.

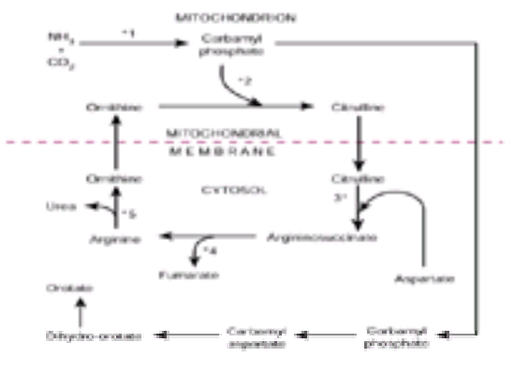

Fig. 1 The urea cycle functions partly in the mitochondrion and partly in the cytosol. Carbamyl phosphate, if it accumulates, may be diverted to orotic acid synthesis.

Asterisked enzymes are: 1, carbamyl phosphate synthetase; 2, ornithine transcarbamylase; 3, argininosuccinate synthetase; 4, argininosuccinate lyase; and 5, arginase.

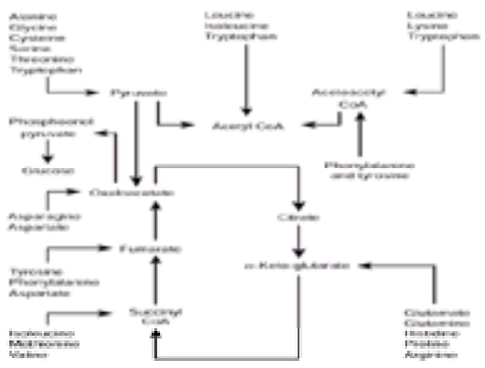

Fig. 2 Amino acids as a source of energy. The multiple entry points to the citric acid cycle for the metabolites of carbon chain catabolism.

Nitrogen balance and dietary treatment

Some biochemical defects such as homocystinuria respond well to vitamin (coenzyme) supplementation and others are treated by diet. Generalized moderate protein restriction is a usual approach to urea cycle defects and one or two other diseases, but very specific restriction of one or two amino acids applies crucially to a small number of essential amino acid disorders. Thirty years ago Rose and colleagues defined the eight essential amino acids in adults (Table 3) and the minimum daily requirements for sustaining nitrogen balance. This also requires an additional intake of 'non-essential' amino acid nitrogen and an adequate calorie intake. Histidine and taurine may be essential in the neonate. Dietary restriction can be used to treat specific metabolic defects of essential amino acids but is unlikely to be successful for disorders of non-essential amino acid metabolism.

Almost all ingested protein in the infant (recommended intake about 2 g/kg/day) is utilized in the synthesis of new protein for growth. This persistent anabolic state, however, is easily upset by intercurrent infection, starvation, trauma, or surgery, with a rapid swing to a catabolic state and negative nitrogen balance. The amino acids released from protein hydrolysis increase the load on urea formation and their normal pathways of intermediary metabolism. This renders infants and young children prone to frequent clinical illness with some amino acid disorders. In adults an intake of natural protein of 60 to 80 g/day is probably about twice that needed to maintain nitrogen balance and health. Adults are less prone to become catabolic than infants but the same circumstances may nevertheless precipitate it. In addition periods of particular risk include late adolescence when the growth spurt ceases and the postpartum period. Insulin and glucose can be used to reverse a catabolic state and produce positive nitrogen balance in some inborn errors.

Amino acid transport defects

General pathophysiology

Historically, the renal tubular aspects of amino acid transport have been of major importance following Dent's (1948) successful introduction into clinical practice of paper chromatography for the analysis of urinary amino acids. Table 4 sets out a classification of amino aciduria. Normal renal clearance values for the amino acids are given in Table 1. A general account of amino acid transport would need to cover not only the renal tubule but the intestinal mucosa, the placenta, the blood–brain barrier, cell membranes in a host of tissues, and intracellular membranes. No attempt is made here to address the generality of transport issues.

The generalized amino acidurias

The Fanconi syndrome

General aspects

There are four components to the Fanconi syndrome:

1. characteristic low-molecular-weight proteinuria, e.g. a $_1$-microglobulin, b$_2$-microglobulin, b$_1$-glycoprotein, and retinol binding protein;
2. tubular transport defects;
3. metabolic bone disease, rickets, or osteomalacia;
4. slow loss of glomerular function.

Glycosuria, generalized amino aciduria, and phosphaturia are a classic triad. The conservation of sodium, potassium, bicarbonate, and urate is impaired and the plasma concentrations of the last three decreased. Many examples of the Fanconi syndrome are not primarily disorders of amino acid metabolism (cystinosis is an exception) but the effects of exogenous or endogenous toxins (e.g. galactose-1-phosphate) which accumulate in other genetic defects (Table 5). The Fanconi–Bickel syndrome is a distinct genetic entity.

Maleic acid (maleate) has been used to produce experimental models of the Fanconi syndrome, as have 4-pentenoate and succinyl acetone (see below). Experimentally, maleate lowers intracellular concentrations of amino acids and sugars predominantly by increasing efflux. Maleate affects mitochondrial oxidation processes, impairs 1a-hydroxylation of 25-hydroxycholecalciferol and may directly affect cell membranes. It has still not proved possible to be sure whether the Fanconi syndrome should be regarded as a disorder of proximal or distal tubules, or both, whether efflux from cell to lumen is more important than reabsorption defects, and whether all causes of the syndrome act through a final undefined common mechanism. A central role for impaired energy production is suggested by new reports of tubular defects in mitochondrial disorders, for example cytochrome c oxidase deficiency and the Kearns–Sayre syndrome.

The Fanconi–Bickel syndrome

This is a rare autosomal recessive disorder caused by mutations in a glucose transporter gene expressed in kidney, liver, intestine, and pancreas. It is associated with hepatomegaly, glycogen storage, fasting hypoglycaemia, short stature, and proximal tubular nephropathy.

The dominantly inherited Fanconi syndrome

This disorder, of unknown cause, characteristically presents in the second to fourth decade and slowly evolves into late adult life when renal failure may be advanced (Fig. 3). The clinical presentation is commonly with rickets or osteomalacia, which require treatment with calcitriol. Potassium, sodium bicarbonate, and phosphate supplements may also be needed.

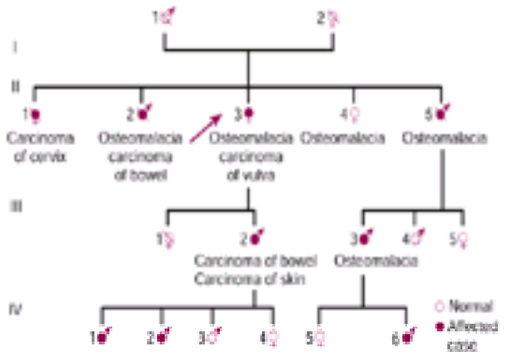

Fig. 3 Pedigree of dominantly inherited Fanconi syndrome. (From Brenton *et al*. (1981) with permission.)

The oculocerebrorenal syndrome of Lowe

This is an X-linked disease characterized by dwarfism, severe mental retardation, and blindness secondary to cataracts, microphthalmos, and glaucoma. The tubular defect includes proteinuria, rickets but not usually glycosuria, and an amino aciduria with relative sparing of the branched chain amino acids. The *OCRC1* gene is on the long arm of the X chromosome and codes for inositol polyphosphate-5-phosphatase.

Cystinosis

Clinical

Cystinosis results from defective carrier-mediated transport of cystine through the lysosomal membrane, which may rupture due to cystine crystallization in hexagonal or rectangular forms causing cell damage. In the proximal renal tubule this leads to the Fanconi syndrome. In the severe infantile form clinical presentation occurs after a few months of life with polyuria, thirst, salt and water depletion, hypokalaemia, and proximal renal tubular acidosis. Poor feeding and failure to thrive are characteristic. Hypophosphataemia and impaired 1-hydroxylation of 25-hydroxycholecalciferol contribute to florid rickets.

Photophobia develops with the accumulation of cystine crystals in the cornea and retinopathy. Hypothyroidism is common and renal failure develops leading to death by 10 years of age. Growth is invariably impaired even before kidney transplantation and the concomitant steroid immunosuppression. Sexual development is late. Intelligence is normal in early life. In transplanted patients retinopathy and visual loss may progress and central nervous system changes may occur. Cystine crystals are not seen here, but tissue cystine concentrations are very elevated. Cortical atrophy and memory defects occur in some older patients. frank neurological features are now more commonly described in survivors after renal transplantation, and may respond to treatment. The spectrum of organ defects is likely to widen in long-term postrenal transplant survivors.

Variant forms

A benign adult form presents with photophobia due to corneal crystals. There may also be crystals in the bone marrow and leucocytes but the kidney is spared and life expectancy is normal. An intermediate form is like the classic infantile form but presents in late childhood or early adult life. Renal involvement and renal failure occur.

Biochemistry

It is probable that all tissues accumulate cystine, but not equally, and some (e.g. muscle and brain) never seem to develop crystals. Crystals occur in the tissues with the highest cystine concentrations, increasing with age to values several hundred times normal. Cultured fibroblasts and leucocytes have values of 50 to 100 times normal, but cultured lymphoid cells are only four to five times normal. Leucocyte cystine content is higher in the intermediate than the benign form, and highest in the severe classic infantile form. The intralysosomal cystine originates from proteins catabolized within the lysosome and extracellular cystine transported into the cell. Cystine egress from the lysosome is defective. The carrier is not shared by other amino acids, which have other lysosomal transport systems. Cystine loaded renal tubules have severely compromised ATP production due to a deficient intracellular phosphate concentration.

Diagnosis

This is based on the clinical features, features of the Fanconi syndrome, and the presence of cystine crystals. In the cornea these can be seen with a hand lens or a slit lamp in an older child but in infancy they are best seen in bone marrow aspirates (Fig. 4) fixed in alcohol and examined under polarized light. Analysis of peripheral leucocytes for their cystine content is possible in only a few laboratories.

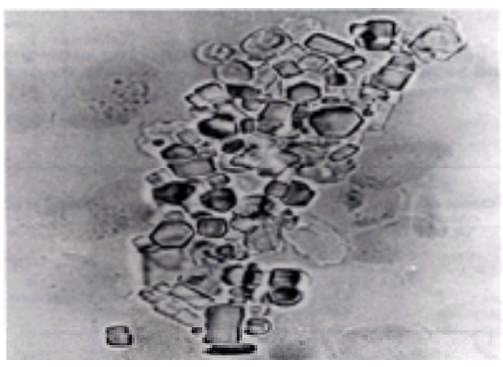

Fig. 4 Cystine crystals in the marrow of a child with cystinosis (× 2200, partially polarized light). (By courtesy of Dr B. Lake, The Hospital for Sick Children, Great Ormond Street, London and with the permission of Heinemann Medical Books.)

Genetics

The disease is autosomal recessive. The incidence is about 1 in 200 000 live births. A higher incidence has been reported from parts of France. Heterozygotes are clinically normal but have raised leucocyte cystine concentrations. Patients have mutations in a gene (*CTNS*) on the short arm of chromosome 17 encoding a lysosomal membrane protein cystinosin. Over 30 mutations in the *CTNS* gene have been recognized in nephropathic cystinosis and others in variant forms.

Prenatal diagnosis

This has been successfully achieved using cultured amniocytes and measuring ^{35}S cystine uptake, or by direct analysis of chorionic villus samples for cystine content. Mutation analysis is now possible.

Treatment

Renal losses of salt, bicarbonate, and potassium may require initial intravenous replacement, but oral supplements including phosphate suffice later although the need for them may be substantial. Phosphate alone may not heal the rickets without the addition of calcitriol. Oral cysteamine, or phosphocysteamine, given in divided doses, depletes leucocyte cystine and gives improved growth and preservation of renal function. Cysteamine eye drops have been used in very young children to clear corneal crystals. The role of cysteamine in preventing the consequences of cystine accumulation in non-renal tissues after transplantation is under study. Dialysis and/or renal transplantation are required for renal failure. Transplanted kidneys do not accumulate cystine. Thyroxine is needed for hypothyroidism. Growth hormone treatment increases height but has been reported to hasten the need for renal replacement. others have not found this, and cysteamine treatment in early childhood improves growth anyway. Plasma carnitine concentrations are often low and can be increased to normal by the use of supplements but this may not help any muscular weakness.

Specific amino acidurias

The recognition of genetic disorders characterized by the excretion of a specific group of amino acids has stimulated research into amino acid transport. Major clinical problems are found in cystinuria and lysinuric protein intolerance.

Cystinuria

Clinical

Cystine stone formation in the kidneys and its attendant complications of pain, haematuria, renal obstruction, and infection is the classic clinical presentation. Only 1 to 2 per cent of all renal stones in adult life are cystine stones but the proportion is higher in childhood. The stones may have grown to large staghorn calculi before diagnosis. They are radio-opaque.

Biochemistry

Cystine has a solubility of 400 mg/litre at neutral pH and excretion varies from 400 to 1200 mg/day in affected individuals, with increased excretion of lysine (up to 2 g/day), ornithine, and arginine, and impaired intestinal absorption of the free amino acids. All are absorbed as dipeptides in combination with another amino acid outside the group. There is no deficiency of any amino acid and no urea cycle defect. The faecal and urinary excretion of diamines such as putrescine and cadaverine result from the action of intestinal bacteria on unabsorbed lysine and arginine. Experimental work indicates that one renal transport defect in cystinuria affects a low-K_m system in the brush border shared by the four amino acids. Other transport systems for cystine and the dibasic amino acids exist. Cystine excretion can exceed the glomerular filtration rate, implying the possibility of tubular secretion.

Diagnosis

Diagnosis requires an amino acid chromatogram and quantitation of cystine excretion. Calcium-containing stones have been observed in cystinuria—possibly because infection predisposes to deposition of calcium salts on small cystine deposits. Confusion is most likely when stone analysis is used for diagnosis without a chromatogram.

Genetics

Three subtypes of cystinuria were identified 30 years ago from studies of amino acid excretion and intestinal absorption. A gene on chromosome 2p with over 20 described mutations probably provides the basis for type I cystinuria with high cystine excretion and a high risk of stone formation. A second cystinuria locus on chromosome 19q may be responsible for types II and III cystinuria. Combinations of alleles at these loci probably explain the different subtypes of cystinuria and the confusing family histories. For example type I cystinuria heterozygotes have normal cystine excretion and the disease is always clearly recessive. However, type II heterozygotes excrete substantial amounts of cystine and the pedigrees can appear dominant.

Prenatal diagnosis

This has not been described.

Treatment

As the relationships between genotype, cystine excretion, and risk of stone formation especially in childhood become clearer so will the recommendations become clearer. The daily fluid intake must not be less than 3 litre/day in adults and this must include 500 ml before retiring to bed with a nocturnal rise to pass urine and drink a further 500 ml. Keeping the urine dilute over the 24-h period is the difficult part, but may be sufficient treatment for those without stones.

Reduced protein intake diminishes cystine excretion but this is not much used in treatment. Cystine is much more soluble at alkaline pH (> 7.5). Use of sodium bicarbonate is limited by the large doses (6 g/day or more) needed to raise urine pH significantly. High sodium intakes are contraindicated in hypertension or renal failure. In addition, alkaline urine may dispose to the precipitation of calcium salts. However, high fluid intake with potassium citrate supplements is recommended by some in childhood if cystine excretion is high.

Penicillamine treatment produces the much more soluble disulphide—half cystine and half penicillamine and an overall reduction of cystine excretion greater than can be accounted for by disulphide formation. The effective dose (1 to 3 g/day) should reduce the free cystine excretion to around 200 mg/day if stones are to dissolve. It is usual to start at a dose of 125 mg/day and increase over several weeks to full dose. The unwanted side-effects include blood dyscrasias, rash with arthralgia, fever, and lymphadenopathy. A syndrome with skin lesions resembling pseudoxanthoma elasticum, elastosis perforans serpinginosa, may complicate long-term penicillamine use, as may pyridoxine deficiency. This latter effect is prevented by coadministration of 25 to 50 mg pyridoxine daily. Patients on penicillamine need blood counts every 2 weeks initially and then monthly. Regular urinalysis is needed. Proteinuria is common and above 2 g/day may necessitate stopping penicillamine, as do blood dyscrasias or other severe reactions. Penicillamine is a helpful preventive treatment in patients with recurrent stone formation at lower doses. Large doses are reserved for trying to dissolve large calculi, which may take 1 to 2 years. It is usually well tolerated in cystinuria.

a-mercaptopropionylglycine is an alternative to penicillamine to which it has structural similiarities. It has been used less than penicillamine but should be considered in patients showing the serious toxic effects of penicillamine. Captopril is a sulphydryl compound which forms a disulphide with cystine. decreased cystine excretion related to treatment with captopril does occur but no therapeutic use has yet been established for it. Similarly, decreasing sodium intake and excretion reduces cystine excretion but a therapeutic role has not been accepted.

Cystine stones are not easily broken by lithotripsy, but it may still be helpful. Percutaneous removal may have its place for smaller stones, particularly in those who cannot take penicillamine and who are unable to regulate their drinking adequately.

Lysinuric protein intolerance

Clinical

Defective ornithine, lysine, and arginine transport affect the renal tubule and intestine with only minor defects of cystine transport. Stones do not form. At weaning, vomiting and diarrhoea begin. There is nutritional deficiency. failure to thrive, poor appetite, and poor growth are common. Occasional intermittent hyperammonaemic encephalopathy occurs. Osteoporosis is an important part of the clinical picture, with vertebral collapse. Interstitial lung disease causes breathlessness, cough, fever, and reduced arterial PO_2. Intellect is normal or mildly impaired. Pregnancy is associated with haemorrhage during labour. Immunological abnormalities have been reported.

Biochemistry

Plasma concentrations of arginine, ornithine, and lysine are low but citrulline, alanine, and glutamine are increased. Renal clearance values for lysine are 20 to 30 times normal and renal losses may be up to 1 g/day. Less marked increases of orthinine and arginine excretion are found but cystine increases are minor.

Plasma lysine values fail to rise after oral lysine loads or the ingestion of lysyl peptides. Intracellular peptide hydrolysis liberates lysine, which cannot be transported across the basolateral membrane, the site of the transport defect. There is also evidence of a transport defect in cultured fibroblasts but not in red cells. A deficiency of intramitochondrial ornithine due to a transport defect across the mitochondrial membrane may impair the urea cycle, causing hyperammonaemia and orotic aciduria (see below).

Genetics

The disease is an autosomal recessive with a relatively high incidence in Finland (1 in 60 000) compared with the rest of the world. The gene has been ocalized to chromosome 14q coding for a permease-related protein.

Prenatal diagnosis

This is possible with molecular techniques.

Treatment

Hyperammonaemia can be largely prevented by a low-protein diet. However, adequate calorie intake is difficult to sustain in infancy and appetite often remains poor. Protein restriction does not correct lysine deficiency and oral lysine supplementation causes diarrhoea. Oral citrulline (2.5 to 8.5 g/day), absorbed via a different transport system, corrects ornithine and arginine deficiency and lowers plasma ammonia by priming the urea cycle. Acute hyperammonaemic crises are managed with intravenous glucose and intravenous or oral sodium benzoate or phenylbutyrate (see below). Citrulline treatment should be maintained but intravenous citrulline is not readily available. Intravenous ornithine and arginine have been tried.

e-*N*-acetyl lysine has been used in the attempt to overcome lysine deficiency, which may be a factor in the osteoporosis and other problems. Plasma lysine concentrations rise but there is no agreement on its use, and cost and availability are a problem.

The cause of the serious interstitial pneumonia is not clear. It has not apparently responded to antibiotics given for the possibility of pneumocystis infection. Successful treatment with prednisolone has been reported.

Neutral amino aciduria: the Hartnup syndrome

This is an autosomal recessive disorder of neutral amino acid transport across the luminal brush border membrane of kidney and intestine. It does not involve cystine and the basic amino acids, the acidic acids, glycine, or the iminoacids (see Fig. 3). Clinical effects may include a light-sensitive rash on exposed skin, cerebellar ataxia, and mental disturbance, but patients with this disorder frequently remain normal. Affected individuals may respond to nicotinamide, but this does not change the amino acid transport defect. The relative deficiency of nicotinamide is attributed to the losses of the precursor amino acid tryptophan and its impaired intestinal absorption. Bacterial action on unabsorbed tryptophan generates indoles, which appear in the stools and urine and are characteristic of the disorder.

Familial renal iminoglycinuria

The excretion of glycine, proline, and hydroxyproline is raised in the Fanconi syndrome and in the inborn errors of proline or hydroxyproline metabolism when plasma concentrations of these amino acids are raised. Transient raised excretion of the three amino acids is usual in neonates, which reflects the ontogeny of one shared transport system. Genetic iminoglycinuria is an autosomal recessive defect of another transport system. The evidence supports several allelic mutations in the genetic defect with some heterozygotes having raised glycine excretion and some normal amino acid excretion. Familial iminoglycinuria is the consequence of a well worked out transport defect which is clinically harmless.

The g-glutamyl cycle

A possible amino acid transport system

Six enzyme-catalysed reactions link the steps for the synthesis of glutathione and its metabolism (Fig. 5). Glutathione is believed to be transported to the cell membrane, where its antioxidant properties may be important in preventing lipid peroxidation. Tissues with low g-glutamyl transpeptidase levels in the cell membrane transport glutathione into the body fluids and circulation. Some is filtered at the glomerulus. g-glutamyl transpeptidase, bound to the cell membrane of transport epithelia such as the choroid plexus, ciliary body, nephron, and jejunum, has been assigned a role in the membrane transport of amino acids via the formation of g-glutamyl amino acid peptides, which is quite different from free amino acid transport. The peptides are cleaved by g-glutamyl cyclotransferase to free the transported amino acid and the g-glutamyl moiety, which cyclizes to 5-oxoproline (pyroglutamic acid). Cystine is among the amino acids transported in this way and one function of the cycle may be to conserve cystine and, indirectly, cysteine. There is no suggestion of any defect in the cycle in cystinuria. The inherited defects of the g-glutamyl cycle are summarized in Table 6. Some of the links between biochemical defects and clinical manifestations are tentative.

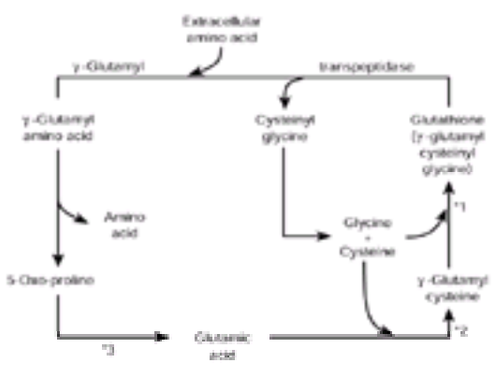

Fig. 5 The g-glutamyl cycle synthesizes glutathione and may play a role in amino acid transport. Asterisked enzymes are: 1, glutathione synthetase; 2, g-glutamyl cysteine synthetase; and 3, 5-oxoprolinase.

Defects of the urea cycle

Amino acids taken in excess of synthetic need are catabolized and the amino group converted to urea. Hyperammonaemia is one of the major metabolic abnormalities in urea cycle defects but is not unique to them (Table 7).

The formation of urea

Nearly all waste nitrogen disposal—10 to 12 g/day—is in the form of urea synthesized in the liver from ammonium ions (NH_4^+) and the a-amino nitrogen of aspartic acid (see Fig. 1). The ammonium nitrogen is incorporated into the first committed synthetic step to urea formation—the production of carbamyl phosphate for which

N-acetyl glutamine is believed to be regulatory. The a-amino nitrogen of aspartic acid comes from many amino acids during their transamination reactions with oxaloacetic acid. It is incorporated during the formation of argininosuccinic acid. Ornithine nitrogen is not incorporated into urea. Bicarbonate provides the carbon moiety of urea but this is not generally regarded as important in acid–base balance.

The source of ammonium ions (NH_4^+) for the generation of carbamyl phosphate is less clear. Glutamine synthesized in skeletal muscle is extensively taken up by the intestine. Glutamine nitrogen is released into the portal blood as alanine, ammonium ions, and citrulline. Apart from these urea precursors, ammonium ions are released into the renal vein by the action of renal glutaminase on glutamine. The generation of ammonium ions within the liver had been attributed to the deamination of glutamate by glutamate dehydrogenase. Transamination reactions involving glutamate are probably more important in linking glutamate to the urea cycle. Within the liver a number of other amino acids are deaminated and may be a source of ammonium for urea synthesis.

The extrahepatic urea cycle enzymes

The urea cycle synthesizes arginine but it has been noted that hepatic transplantation for urea cycle defects does not correct previously low plasma concentrations of citrulline and arginine. The intestine can also synthesize citrulline with the mitochondrial parts of the cycle. Other tissues contain only some of the urea cycle enzymes. Citrulline transported to a variety of tissues with the cytosolic components of the cycle can be used to synthesize arginine via argininosuccinic acid. This extrahepatic synthesis of arginine may be crucial to the body's needs.

The inherited defects of the urea cycle

Four of five inherited defects of the urea cycle (see Fig. 1) have common clinical features but arginase deficiency is different. Quite separately the activity of carbamyl phosphate synthetase can be impaired by a rare genetic defect in *N*-acetylglutamine formation which is not considered here. Ornithine transcarbamylase deficiency is the most common of the defects.

Clinical features of carbamyl phosphate synthetase deficiency, ornithine transcarbamylase deficiency, argininosuccinic acid synthetase deficiency, and argininosuccinic acid lyase deficiency

The neonatal presentation of these conditions is identical. After a brief normal period of 24 to 72 h, poor feeding, lethargy, and vomiting precede the descent to unresponsiveness and hyperammonaemic coma. Argininosuccinic acid lyase deficiency may be less acute and severe than carbamyl phosphate synthetase deficiency, ornithine transcarbamylase deficiency, or argininosuccinic acid synthetase deficiency because argininosuccinic acid excreted at the glomerular filtration rate (there being no tubular reabsorption) is a means of nitrogen excretion, and hyperammonaemia tends to be less severe. In males, ornithine transcarbamylase deficiency is usually fatal, but survival in the other conditions is more likely. Survivors may suffer intellectual impairment and other neurological damage. Only one of the four is X linked (ornithine transcarbamylase deficiency) and female carriers may sometimes present clinically in the neonatal period, presumably because of preponderant inactivation of the X chromosome with a normal gene.

Later presentations come in two broad clinical forms. Mental retardation and epilepsy without any clear neonatal history are well described in argininosuccinic acid lyase deficiency and also in carbamyl phosphate synthetase deficiency, ornithine transcarbamylase deficiency, and argininosuccinic acid synthetase deficiency. Children with argininosuccinic acid lyase deficiency may also show the hair defect of trichorrhexis nodosa, which is not shared by the other urea cycle defects. Another late presentation is with intermittent encephalopathy. This is seen in females who are carriers for ornithine transcarbamylase deficiency, including presentation in the puerperium after a symptomless pregnancy, and males hemizygous for ornithine transcarbamylase deficiency with less severe mutations who have presented in late childhood or the teenage years. Death has been recorded in these late onset encephalopathies. Carbamyl phosphate synthetase deficiency and argininosuccinic acid synthetase deficiency may also present in this way.

Clinical features of arginase deficiency

There is a progressive spastic quadriparesis, most marked in the legs, with psychomotor retardation, epilepsy, and poor growth. Obvious manifestations present in early childhood. Hyperammonaemic coma occurs but hyperammonaemia is less marked than in the other disorders.

Biochemistry

These defects are summarized in Table 8. Hyperammonaemia is preceded by raised plasma alanine and glutamine concentrations and may be accompanied by a rise in transaminases and prolongation of the prothrombin time. The raised excretion of orotic acid in some defects is caused by the accumulation of carbamyl phosphate, which is directed to pyrimidine synthesis (see Fig. 1).

Experimental hyperammonaemia in primates initially causes decreased activity, lethargy, and vomiting, and then hyperventilation and respiratory alkalosis, which have also been recorded in humans. Seizures and coma follow with progressive rise of intracranial pressure and cerebral oedema. The astrocytes, which occupy one-quarter to one-third of brain volume, exhibit marked swelling and mitochondrial change. High astrocyte glutamine concentrations may act osmotically to cause cerebral oedema. Many metabolic changes in hyperammonaemia are secondary to cerebral oedema. Glutamine concentrations ten times normal have been recorded in the cerebrospinal fluid in ornithine transcarbamylase deficiency and argininosuccinic acid lyase deficiency. Other amino acid abnormalities in the cerebrospinal fluid have been described in arginase deficiency. An early effect of hyperammonaemia on amino acid transport across the blood–brain barrier has been described, with tryptophan transport being regarded as particularly important.

Diagnosis

The biochemical defects are diagnostically important (see Table 8). Carbamyl phosphate synthetase deficiency can only be diagnosed when hyperammonaemia is not associated with the biochemical changes of the other urea cycle defects, although a low plasma citrulline value gives a clue. Other causes of hyperammonaemia must be excluded (see Table 7), which requires urinary organic acid analysis, consideration of Reye's syndrome, and acute valproate encephalopathy. Confirmatory enzyme assays on liver biopsy samples may be needed in carbamyl phosphate synthetase deficiency and ornithine transcarbamylase deficiency. Liver function and clotting tests should be checked.

Genetics

With the exception of ornithine transcarbamylase deficiency, the diseases are autosomal recessive. The gene for carbamyl phosphate synthetase is on the short arm of chromosome 2. Inherited deficiency is rare, with 14 mutations described. The enzyme protein may be targeted to the mitochondria by a leader peptide and the mature enzyme constitutes a relatively high proportion of mitochondrial protein. The gene for ornithine transcarbamylase is on the short arm of the X chromosome and its product targeted to mitochondria in a manner similar to carbamyl phosphate synthetase. Functional catalytic trimers form within the mitochondrial matrix. Ornithine transcarbamylase deficiency is associated with a variety of gene defects—insertions, deletions, and point mutations; about 140 have been described.

Argininosuccinic acid synthetase, argininosuccinic acid lyase, and arginase are cytoplasmic enzymes. Argininosuccinic acid synthetase catalyses the synthesis of argininosuccinic acid from citrulline and aspartic acid, requires adenosine triphosphate and magnesium ions, and functions as a tetramer of about 185 000 Da. The gene is on the long arm of chromosome 9. Argininosuccinic acid lyase, which cleaves argininosuccinic acid, functions as a tetramer of about 173 000 Da and the coding gene is on the short arm of chromosome 7. About 12 mutations have been described. Fibroblast studies of argininosuccinic acid lyase indicate that crossreacting material is usually present and correlates poorly with residual enyzyme activity. There are multiple complementation groups and, by implication, multiple alleles at the structural gene locus.

Hepatic arginase, a trimer of submit size around 35 000 Da, cleaving arginine to urea and ornithine, has a locus on the long arm of chromosome 6. A separate mitochondrial arginase is present in kidney.

Antenatal diagnosis

A restriction fragment length polymorphism has been helpful in diagnosis of carbamyl phosphate synthetase deficiency, with fetal liver biopsy and enzyme assay the

only alternatives. Antenatal diagnosis in ornithine transcarbamylase deficiency is complex. If the mother is known to be a carrier from pedigree analysis or biochemical testing, three approaches are possible:

1. If the mutation is known within the family then direct examination of the fetal genotype is possible using appropriate probes, but this occurs in a minority of cases.
2. If a restriction fragment polymorphism is linked to the mutant gene in the family then this approach may be possible.
3. If no such information is available then sexing the fetus followed by fetal liver biopsy and enzyme assay in the male is the only approach left.

Antenatal diagnosis for argininosuccinic acid synthetase deficiency is also difficult. The enzyme can be assayed in amniocytes and placental villus material, but it is more reliable to culture amniocytes with radioactive citrulline and measure the incorporation of the radioactive products into cell protein. Amniotic fluid citrulline concentrations may help. Molecular analysis is possible in argininosuccinic acid lyase deficiency. analysis of amniotic fluid for argininosuccinic acid or enzyme assay on cultured amniocytes have been used successfully. Arginase deficiency has been detected on fetal red cells and a number of mutations have now been identified.

Heterozygote detection in ornithine transcarbamylase deficiency

Because of its X-linked inheritance carrier detection is particularly important. Pedigree analysis including DNA studies where necessary, or investigation of frank symptomatic episodes may settle the issue. The symptomless female can be a problem, however. Protein loading with serial measurements of plasma ammonia and urinary orotic acid may reveal the biochemical defect but may also cause serious symptoms. Allopurinol causes a greater excretion of orotic acid and orotidine in carrier females than in normals and forms the basis of an acceptable safe test of heterozygosity. It may fail to identify some carriers.

Treatment and prognosis

The management of acute encephalopathy involves reducing the need to synthesize urea. Dietary protein is stopped and endogenous protein breakdown suppressed by a high oral carbohydrate intake or using intravenous 10 to 20 per cent dextrose and insulin if needed to control blood glucose concentrations. The blood ammonia is lowered in the neonatal period by peritoneal dialysis or haemodialysis (more effective). Slower methods useful in carbamyl phosphate synthetase deficiency and ornithine transcarbamylase deficiency include the use of intravenous or oral sodium benzoate, which is excreted as its glycine conjugate hippuric acid, so raising nitrogen excretion. The use of sodium phenylbutyrate, which is excreted as phenyl acetylglutamine, is more effective. Serious toxicity from either benzoate or phenylbutyrate overdose is possible. in argininosuccinic acid synthetase and argininosuccinic acid lyase deficiencies, oral or intravenous arginine is an urgent and important therapy to remedy deficiency. In argininosuccinic acid lyase deficiency in particular, plasma ammonia levels fall when arginine is administered. The prognosis for severe neonatal illness is poor (see above) especially if plasma ammonia concentrations are over 1000 µmol/litre.

Maintenance treatment of all urea cycle defects (including arginase deficiency) between encephalopathic episodes involves protein restriction to the minimum required for growth and development and supplementation with arginine in argininosuccinic acid synthetase deficiency and argininosuccinic acid lyase deficiency. The continuous use of oral sodium benzoate or sodium phenylbutyrate in carbamyl phosphate synthetase deficiency and ornithine transcarbamylase deficiency may be needed to maintain low plasma ammonia concentrations. Late onset forms of the urea cycle diseases carry a better prognosis, but arginase deficiency seems relentlessly progressive. Babies with argininosuccinic acid lyase deficiency picked up by neonatal screening but who have not developed early clinical illness are reported to develop with normal IQ on large arginine supplements and a low protein intake. Others do less well and urea cycle defects generally have a poor prognosis.

Valproate should be avoided in the treatment of seizures in urea cycle defects and ornithine transcarbamylase carriers because it may precipitate coma.

Liver transplantation has sometimes been carried out for urea cycle defects. Selecting patients and balancing the risks is extremely difficult. Gene transfer therapy has been attempted but the problems of suitably safe vectors and stable expression remain.

The disorders of carbon chain metabolism

The classification in Table 2 is a useful approach, but many different catabolic pathways and associated clinical abnormalities necessitate separate consideration of individual amino acids (or groups of them) with their relevant vitamin coenzymes. Pyridoxine, because of its central and varied roles, is considered separately below. The relatively 'non-specific' nature of some biochemical abnormalities is stressed again.

Pyridoxine

Pyridoxal phosphate is the coenzyme in amino acid transaminations, decarboxylations, and deaminations. Considerable molecular detail of its role in transamination has been worked out. It is also the coenzyme in the synthesis and breakdown of cystathionine in the trans-sulphuration pathway. The normal dietary pyridoxine intake is 2 to 3 mg/day but a number of diseases respond to doses of 10 to 500 mg/day. These include deficiencies of ornithine aminotransferase, cystathionine b synthase, cystathionase, hyperoxaluria due to peroxisomal glyoxylate aminotransferase deficiency, and some neonates with seizures considered due to defective glutamine decarboxylase, the enzyme which generates g-aminobutyric acid (see later).

'Non-specific' biochemical defects

The multiple causes of hyperammonaemia have been listed (see Table 7). Elevations of plasma glycine may also be non-specific and not necessarily a result of primary enzyme defects in glycine metabolism. Increases of glutamine and alanine in plasma are common in the early stages of ammonia accumulation. Hypoglycaemia is frequent in the organic acidurias as well as in specific defects of gluconeogenesis or glycogen metabolism. Alanine concentrations rise in lactic acidosis.

Defects of ornithine metabolism

P>Ornithine is a non-protein amino acid upon which the synthesis of urea takes place (see Fig. 1) and which is regenerated once the urea moiety is split off. It is also produced when arginine reacts with glycine to produce guanidinoacetate, the precursor of creatine. Ornithine-†-amino transferase produces glutamic semialdehyde, which cyclizes to pyrroline-5-carboxyllic acid, and is also produced from proline. The decarboxylation of ornithine produces the diamine putrescine.

Deficiency of ornithine-†-aminotransferase: gyrate atrophy

Clinical

The major abnormality is an atrophy of choroid and retina, beginning as a small yellowish spot and increasing to a circular lesion edged with pigment giving an 'atypical retinitis pigmentosa' appearance. Children present with myopia and decreased night vision progressing to blindness in middle life. Cataracts also develop but optic discs, cornea, and iris remain normal. A few patients develop mild proximal muscle weakness. Microscopic abnormalities of skeletal muscle fibres are found. Magnetic resonance imaging shows changes in the central nervous system, but the longer-term clinical implications are uncertain.

Biochemistry

Plasma ornithine values range from 400 to 1000 µmol/litre (normal 75 µmol/litre) with high concentrations in cerebrospinal fluid and aqueous humour. 400 to 900 mg/day is excreted with increased amounts of arginine and lysine (competitive inhibition of reabsorption).

The activity of ornithine-†-aminotransferase is low in liver and skeletal muscle. Most affected patients have less than 1 per cent of normal activity in fibroblasts. Some have values up to 5 to 6 per cent and some enzyme-deficient lines show marked increase of activity with very high concentrations of pyridoxal phosphate.

Diagnosis

The clinical picture and the amino acid defects are adequate means of diagnosis. Enzyme assays can be used to confirm it.

Genetics

It is an autosomal recessive with the highest incidence in Finland, (where it may be as high as 1 in 50 000). There are several mutants, as evidenced by complementation studies. The gene has been mapped to chromosome 10q. Two pseudogenes exist on the X chromosome. Different mis-sense mutations have been described in pyridoxine responsive and non-responsive forms. Splicing defects have also been described. Over 50 mutations have been described in gyrate atrophy.

Treatment

Despite encouraging therapeutic studies on a mouse model there are no reports of clinical improvement in humans but deterioration may be slower in patients whose plasma ornithine levels fall with pyridoxine treatment (500 mg/day or less). Low-arginine diets may reduce plasma ornithine concentrations as do large doses of lysine given to augment renal ornithine excretion. Creatine has been given and has been reported to improve muscle histology, but ocular deterioration continues. Local proline deficiency in the retina has been suggested as a cause of the retinal degeneration. Proline supplementation does not stop disease progression. The best approach if patients do not respond to pyridoxine maybe a combination of diet and high lysine doses. Studies on siblings in affected families indicate that the development of retinal changes is at least delayed by control of the plasma ornithine concentration.

Hyperornithinaemia with hyperammonaemia and homocitrillinuria

Clinical

Hyperornithinaemia with hyperammonaemia and homocitrillinuria is referred to as the **HHH** syndrome. Intermittent hyperammonaemic encephalopathy with vomiting, drowsiness, and coma may date back to infancy, or patients may present much later. Impairment of IQ from low normal to more severe retardation, with epilepsy and frank neurological features, is another form of presentation. Growth tends to be poor. Chorioretinal atrophy has been reported in one patient but to date has not been commonly seen.

Biochemistry

Intermittent hyperammonaemia, with plasma ornithine values three to ten times normal and increased excretion of orotic acid are believed to result from impaired transport of ornithine into the mitochondria which leads to the accumulation of carbamylphosphate. This increases orotic acid formation and the production of homocitrulline by the transcarbamoylation of lysine.

Genetics

It is an autosomal recessive. A gene for an ornithine transporter across the mitrochondial membrane (*ORNT1*) has been mapped to chromosome 13q. It has been reported that three mutant alleles in this gene account for a high proportion of HHH patients in North America.

Treatment

Moderate protein reduction (1 g/kg/day) reduces plasma ammonia and ornithine concentration. Ornithine supplementation may then lower plasma ammonia further by raising intracellular ornithine concentrations, which may induce entry of more ornithine into the mitochondria. In siblings presenting as adults, treatment with citrulline and sodium phenylbutyrate has decreased plasma ammonia, increased plasma ornithine, and relieved episodic confusional episodes. The outcome of treatment in the longer term is not known.

Defects of phenylalanine metabolism

The importance of tetrahydrobiopterin

The hyperphenylalaninaemias are a group of disorders characterized by defective hydroxylation of phenylalanine to tyrosine and plasma phenylalanine values above the normal fasting range of 40 to 80 µmol/litre. Tetrahydrobiopterin is the required coenzyme for this hydroxylation and high phenylalanine values may be due to defects in the apoenzyme or the generation of tetrahydrobiopterin.

An adult phenylalanine intake is about 3 to 4 g/day, one-quarter of which is incorporated into protein and three-quarters hydroxylated to tyrosine (Fig. 6). Adults need about 1 g/day, but in classic severe phenylketonuria health is maintained on half this. Transamination to phenylpyruvic acid and decarboxylation to phenylethylamine assume much greater importance in phenylketonuria because they occur only at elevated phenylalanine concentrations.

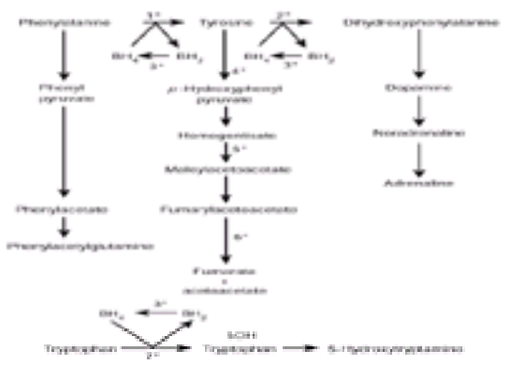

Fig. 6 The metabolism of phenylalanine and tyrosine and the role of tetrahydrobiopterin. The asterisked enzymes are: 1, phenylalanine hydroxylase; 2, tyrosine hydroxylase; 3, dihydrobiopterin reductase; 4, tyrosine aminotransferase; 5, homogentisic acid oxidase; 6, fumaryl acetoacetate hydrolyase; and 7, tryptophan hydroxylase.

Classic phenylketonuria

Clinical

Phenylalanine values are higher than 1200 µmol/litre (sometimes much higher). Untreated, phenylketonuria almost invariably causes severe mental retardation, with IQ values only occasionally above 60, and most often well below. a few patients have normal IQ values despite the biochemical defect; some female patients have been discovered only because of abnormalities in their offspring (see below). Brain phenylalanine concentrations measured by magnetic resonance spectroscopy have been lower than expected in some of these patients probably accounting for the preservation of IQ. Both microcephaly and epilepsy are common. About one in 20 untreated patients develop neurological problems in adult life, usually spastic paraparesis but sometimes extrapyramidal features. Pigmentary deficiency in the iris and hair are features of the untreated disease and so is eczema.

Milder variants

Mutations with greater residual enzyme activity produce phenylalanine values of 300 to 1200 µmol/litre. Those over 480 µmol/litre should be treated: some were not

with variable outcome for IQ.

Biochemistry

Plasma phenylalanine concentrations are elevated to 20 to 60 times, being highest in babies. Phenyl pyruvic acid which is converted to phenyl lactic acid, phenylacetic acid, and phenylacetyl glutamine accumulates with phenylethylamine. The ketone phenylpyruvic acid in the urine gives the disease its name and a green colour in the ferric chloride test. The defective enzyme phenylalanine hydroxylase, which requires tetrahydrobiopterin as a cofactor, has been found only in the liver in humans. It has never been found in the brain of any species. Phenylalanine hydroxylase may be tetrameric or trimeric with units of molecular weight between 50 000 and 60 000.

Pathology

The pathology of phenylketonuria is not clear. Phenylalanine itself is probably the damaging agent but there is controversy about the mechanism: relative tyrosine deficiency may also be important, reflected in the pigment deficiency and changes in neurotransmitters. High phenylalanine concentrations are associated with impaired brain growth and probably fewer nerve cells. Phenylalanine inhibits an enzyme important in sulphation of myelin intermediates and myelin formation is abnormal. In animal experiments high phenylalanine concentrations reduce transport of other amino acids at the blood–brain barrier and at the placenta. In addition, many *in vitro* biochemical processes (e.g. protein synthesis) are impaired by high phenylalanine concentrations. Patients with classic phenylketonuria also have low concentrations of homovanillic acid and 5-hydroxyindoleacetic acid in the cerebrospinal fluid, indicative of possible deficiency of the neurotransmitters dopamine, noradrenaline, and 5-hydroxytryptophan. Dietary treatment restores normal concentrations in the cerebrospinal fluid.

Diagnosis

All newborns in the United Kingdom should be screened for raised phenylalanine values between the sixth and tenth day of life, either by Guthrie's bacterial inhibition assay, chromatography, or tandem mass spectometry. Phenylalanine values greater than 240 µmol/litre are rechecked and, if confirmed, are investigated. Raised phenylalanine values are seen in the important variants due to defects in tetrahydrobiopterin synthesis and these must be excluded as they require specific treatment. Transient neonatal hyperphenylalaninaemia is probably less common now that cows' milk, with its relatively high protein content, is used less in infancy, but it must be distinguished from permanent forms. Liver disease must be excluded.

Genetics

The disease is autosomal recessive, with an incidence in Western countries of 1 in 8000 to 12 000 live births. It is rare in Finland and Japan. One in 50 people carry a mutant gene. These include splicing mutations, deletions, and mis-sense mutations. The location on chromosome 12 has been confirmed. The majority of patients are compound heterozygotes rather than being homozygous for a single mutation. Residual enzyme activity in liver biopsies has correlated fairly well with *in vivo* studies on the conversion of deuterated phenylalanine to tyrosine and there is growing information on which genotypes cause the most severe functional defects in the enzyme. Over 400 mutations have been described. The contribution which other genes (e.g. for amino acid transport into the central nervous system) may make to the disease manifestations may, however, become clearer.

Antenatal diagnosis

Restriction fragment polymorphisms in linkage disequilibrium with these mutations have been useful in some families for antenatal diagnosis. Patient demand for antenatal diagnosis has been relatively low.

Treatment

Natural protein intake is reduced to provide just what is necessary for growth and development while keeping the plasma phenylalanine between 120 and 360 µmol/litre using the Guthrie test or other technique for regular monitoring. These are lower phenylalanine values than were once recommended because outcome in terms of IQ is closely related to the control of abnormally high values. Persistently low values may also adversely affect outcome. Despite normal or near normal IQ results, more subtle neuropsychological defects have been described in well-treated phenylketonuria patients and may be very important scholastically.

In infancy, milk restriction with supplements is relatively easy. Later it is necessary to introduce other foods on an exchange basis using tables that define the weight of the food containing 1 g of protein (roughly 50 mg phenylalanine). Fruits and some vegetables very low in protein are allowed freely. Adults with classic phenylketonuria tolerate only three to four exchanges, which provide about the same amount of phenylalanine as the free foods. These diets are supplemented with phenylalanine-free amino acid mixtures, minerals, and vitamins. Specially produced low-protein products make the diet more palatable.

Regression of IQ when diets were stopped in later childhood has led to continuation of dietary treatment into the teenage years. Patients generally have not suffered when diets have stopped at 15 or 16 years of age. However, there is no follow-up of a substantial number with respect to IQ change who have been off diet for 20 years or more, and there is concern about possible neurological deterioration.

High plasma phenylalanine concentrations may produce a pharmacological impairment of mental function revealed by psychological tests in short-term studies, which improve when concentrations fall. Long-term damage to intellect or neurological function is another issue. A small number of patients who were not on diet in adult life have developed spastic paraparesis, epilepsy, or extrapyramidal features. These may improve on diet. All these have cerebral changes on magnetic resonance imaging, as do an appreciable proportion of those off diet without neurological manifestations. The imaging changes also improve on diet regardless of whether there were clinical manifestations or not. Together with the known neurotransmitter defects there is a genuine concern for the long-term welfare of patients. Diet for life is restricting and costs £7000 to £8000 annually for the diet alone. There is an urgent need for more information.

Maternal phenylketonuria

The retrospective review of Lenke and Levy in 1980 did much to emphasize the adverse fetal effects of maternal hyperphenylalaninaemia (Table 9). Experience in other centres with large clinics broadly supports these figures. Microcephaly and congenital heart disease in the offspring of mothers returning to diet at the seventh or eighth week emphasizes the need for preconception diet. This is the best policy. Starting dietary measures very early in the first trimester (5 to 6 weeks) lowers the incidence of impaired brain development, but an increased risk certainly remains to the face and heart.

The ratio of fetal to maternal phenylalanine plasma levels is around 1.5 to 1.7 because of active placental transport. Maternal values should be controlled at between 100 and 250 µmol/litre, which requires very careful monitoring twice weekly. Some values will rise above this in the critical first trimester when tolerance is very low and nausea restricts calorie intake. Dietary tolerance in the mother increases from about week 18 due to increased requirement for growth by the fetus and uterus, but also probably because phenylalanine hydroxylase in the fetal liver can be detected early in the second trimester (Fig. 7). There is already clear evidence that lower maternal phenylalanine values result in neonates of higher birth weight and larger head circumference. Dietary control before conception prevents congenital heart disease.

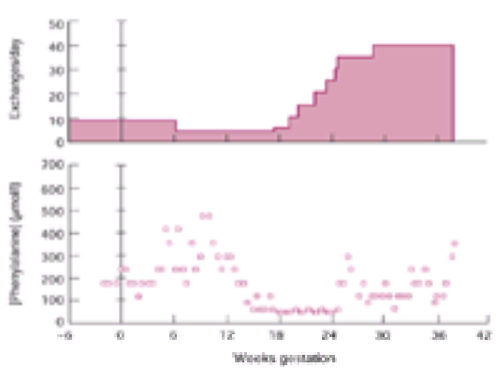

Fig. 7 Diet for a phenylketonuric mother illustrating the marked rise in phenylalanine tolerance in the second half of the pregnancy. (From Fernandes *et al*. (1990) with permission of Springer.)

Defects of biopterin metabolism

In the hydroxylation of phenylalanine the cofactor tetrahydrobiopterin is consumed and must be regenerated. A deficiency of tetrahydrobiopterin adversely affects the function not only of phenylalanine hydroxylase, but also of tyrosine hydroxylase and tryptophan hydroxylase (Fig. 7). Tyrosine hydroxylation is needed for the synthesis of noradrenaline and dopamine, and tryptophan hydroxylation for the production of 5-hydroxytryptophan. Tetrahydrobiopterin is therefore crucial to the production of neurotransmitters. The supply of this coenzyme is impaired in several enzyme defects. All produce hyperphenylalaninaemia, which may not be marked, and all produce progressive neurological disability despite a low-phenylalanine diet. About 1 to 2 per cent of newborns with abnormally raised phenylalanine values have a deficiency of tetrahydrobiopterin.

Dihydropteridine reductase deficiency

Clinical

Progressive neurological deterioration occurs with psychomotor retardation, epilepsy, pyramidal, and extrapyramidal features, especially the latter. Calcification occurs in the cerebral hemispheres.

Biochemistry

Plasma phenylalanine values are elevated. The enzyme dihydropteridine reductase is a dimer or tetramer of four units, each 25 000 Da. It has a wide tissue distribution.

Diagnosis

The most reliable test is an enzyme assay on red cells. It can be carried out on dried blood spots. Oral loading tests with tetrahydrobiopterin may be useful as the plasma phenylalanine may then fall, but as it is not regenerated when the enzyme is deficient the results may be equivocal. Urinary biopterin analyses are needed in the differential diagnosis of these defects.

Genetics and prenatal diagnosis

The disease is an autosomal recessive and the enzyme assay can be carried out on cultured anmiocytes. There are crossreacting material-positive and -negative forms. The gene is on chromosome 4p encoding a protein of 244 amino acids functioning as a homodimer with over 20 described mutations.

Treatment

A low-phenylalanine diet is combined with the administration of L-dopa, 5-hydroxytryptophan, and, in some cases, folinic acid. Early treatment has been reported to give good results. Monitoring of neurotransmitters and folate in the cerebrospinal fluid may help treatment.

Guanosine triphosphate cyclohydrolase deficiency and 6-pyruvoyltetrahydrobiopterin synthase deficiency

The clinical features are similar to those of dihydropteridine reductase deficiency. Intermittent hyperthermia has been described. All urinary biopterin and neopterin values are low in the cyclohydrolase deficiency whereas 6-pyruvoyltetrahydrobiopterin deficiency has high neopterin values and low biopterin values. Tetrahydrobiopterin is used in treatment because, in the presence of dihydropteridine reductase, it can be regenerated from dihydrobiopterin. However, the clinical outcome is not assured and there is concern that tetrahydrobiopterin does not easily enter the central nervous system. Treatment, therefore, is also being attempted with low-phenylalanine diet, L-dopa, and in addition, 5-hydroxytryptophan. From reports on Saudi Arabian families with a high incidence of 6-pyruvoyltetrahydrobiopterin synthase deficiency, tetrahydrobiopterin is said to produce a good outcome if started very early in life.

Disorders of tyrosine metabolism

The steps in tyrosine metabolism starting with the rate-limiting step—the conversion to *p*-hydroxyphenyl pyruvic acid by tyrosine amino-transferase—are outlined in Fig. 6. They are the means of production of the catecholamines, dopamine, and the principal pigments of hair and skin. Diagnosing a specific disorder of tyrosine metabolism needs consideration of the non-specific elevations of plasma tyrosine and methionine seen in liver disorders of various aetiologies and the frequency of transient neonatal tyrosinaemia.

Neonatal tyrosinaemia

An increase of plasma tyrosine concentration and excretion of tyrosine and phenolic acids was commonly seen in premature infants given cows' milk feeds. Lower-protein infant feeds approximating to breast milk have reduced the incidence greatly. Transient deficiency of *p*-hydroxyphenylpyruvate oxidase is considered the unproven cause and appears to be harmless. It responds to reducing any high protein intake and sometimes to ascorbic acid. A repeat tyrosine measurement is indicated to exclude other persistent causes of a raised tyrosine.

Tyrosinaemia type I

Clinical

An acute presentation occurs in the early weeks of life with failure to thrive, vomiting, hepatomegaly, fever, oedema, and epistaxis. Death from hepatic failure occurs within the first year. A milder more chronic presentation is compatible with survival for several years with chronic liver disease, a renal tubular Fanconi syndrome with hypophosphataemic rickets, and sometimes abdominal pain and neuropathy suggestive of acute porphyria (see below). Hypertrophic obstructive cardiomyopathy has been described. One-third of patients progress to hepatocellular carcinoma of the liver.

Biochemistry

Deficiency of fumarylacetoacetate hydrolyase (see Fig. 6) is the cause. A raised plasma tyrosine (and often a raised methionine) result. Succinyl acetone is excreted, formed from fumarylacetoacetate, which also inhibits porphobilinogen synthesis so that †-amino laevulinic acid increases in the urine. Human fumarylacetoacetate hydrolyase is a dimer with a monomer molecular weight of 43 000. Activity is found in liver, kidney, fibroblasts, lymphocytes, and amniocytes.

Diagnosis

Raised plasma tyrosine, succinyl acetone, and †-aminolaevulinic acid excretion and a Fanconi syndrome are the biochemical markers. Fumarylacetoacetate hydrolyase can be assayed in lymphocytes or fibroblasts. It is non-specifically depressed in the liver in a variety of liver diseases. A pseudodeficiency gene in the general population causes low '*in vitro*' assay results for fumarylacetoacetate hydrolyase but no clinical illness. Untreated plasma tyrosine values in proven tyrosinaemia type I may be normal, creating another diagnostic problem. Liver function tests are abnormal.

Genetics

The disease is an autosomal recessive. The acute neonatal form lacks immunologically detectable enzyme protein in contrast to the more chronic form. The fumarylacetoacetate hydrolyase gene has been localized to chromosome 15 and a variety of mutations identified.

Prenatal diagnosis and carrier detection

The measurement of succinyl acetone in amniotic fluid and fumarylacetoacetate hydrolyase in cultured amniocytes or chorionic villus samples forms the basis of prenatal diagnosis. In approximately 5 per cent of families one parent carries both a true mutant allele and the pseudogene, which lowers the parental enzyme activity into the homozygous disease state and causes confusion in prenatal diagnosis. The pseudogene also makes the detection of carriers less certain. Where the mutation is known molecular prenatal diagnosis should be possible and preferable.

Treatment

Restricted intake of tyrosine and phenylalanine may reduce the excretion of succinyl acetone and produce regression of the Fanconi tubular defects. Rickets may require treatment however. The liver disease is not cured. The risk of hepatocellular carcinoma remains. Therapeutic trials are in progress using a metabolic inhibitor, NTBC, which blocks the pathway before homogentisic acid thus reducing the production of toxic metabolites. The results are encouraging, with over 200 patients under follow-up and a greatly reduced incidence of liver damage and hepatic carcinoma since NTBC was introduced in 1991.

Liver transplantation remains the treatment of choice for some who do not respond to NTBC which may also improve renal function, although some succinyl acetone continues to be excreted. transplant timing is immensely problematic. Neither a-fetoprotein nor ultrasound are totally reliable at detecting early malignant change. After liver transplantation the future is uncertain. Chronic renal failure has occurred.

Tyrosinaemia type II

Clinical

Corneal erosions and dendritric ulcers may form within a few months of birth with later scarring, nystagmus, and glaucoma. Corneal transplants can be valuable. The skin lesions may begin after the eye lesions with blistering, painful palms and soles, and hyperkeratosis. Tongue changes have been described. Mental retardation is an inconstant feature but language defects may be more common with possible impaired co-ordination and self-mutilation. The pathology is considered secondary to the deposition of tyrosine crystals in cells precipitating an inflammatory response.

Biochemistry

Tyrosine aminotransferase, which is deficient, catalyses the formation of *p*-hydroxyphenylpyruvic acid (see Fig. 8) and requires pyridoxal phosphate and a-ketobutyrate. It is a liver enzyme, absent from brain, heart, and kidney, with a subunit size of 49 000 which forms dimers. The enzyme is synthesized rapidly, induced by steroids, and has a short half-life. The gene has been mapped to chromosome 16.

Plasma tyrosine values reach 20 times normal (normal 40 to 100 µmol/litre) in younger patients and 10 times normal in others. There is increased excretion of tyrosine, *N*-acetyl tyrosine, and tyramine; there is no Fanconi syndrome. Excreted phenolic acids come from phenylalanine or tyrosine metabolized at high concentrations by other enzymes.

Diagnosis

The clinical features and amino acid analyses are usually sufficient.

Treatment

A low-tyrosine, low-phenylalanine diet has been used to produce rapid improvement of skin and eye manifestations. There is little information on the neurological results of treatment and little on the degree of dietary control needed to sustain clinical improvement.

Alcaptonuria

Clinical

Presentation in infancy occurs only if discoloration of the urine is noticed. It is usually normal when passed, but darkens on standing (more rapidly at alkaline pH) to deep brown or almost black. Back pain begins in the second and third decade with increasing stiffness due to intervertebral disc degeneration. Involvement of the hips, knees, and shoulders follows. Greyish discoloration of cartilage is seen in the pinna, and pigment is deposited in the sclera. Abnormal pigmentation is seen in the heart valves and pigmented stones are common in the prostate. Discoloration of cartilage, tendons, and ligaments is more orange when seen microscopically (ochronosis). The prognosis for the joints is poor. By the fifth decade the lumbar spine is likely to be rigid and other joints will be seriously affected.

Pathology

The pigment is assumed to be a polymer derived from homogentisic acid after enzymatic conversion to the corresponding quinone (homogentisic acid polyphenol oxidase). Virchow described the internally pigmented cartilages including the larynx, tracheal rings, and ribs. The joint cartilages become thinned and fragmented. The intervertebral discs calcify.

Biochemistry

Homogentisic acid oxidase contains ferrous iron and several –SH groups. Molecular oxygen is consumed in splitting the ring to convert homogentisic acid to maleylacetoacetic acid. Homogentisic acid produces a false positive for glucose in the 'Clinitest' reaction but the reaction mixture quickly darkens because of the alkaline pH. There is no reaction with glucose in standard dipstick tests for glucose. Affected individuals excrete 4 to 8 g of homogentisic acid per day.

Diagnosis

In the presence of the clinical symptoms simple urine tests virtually make the diagnosis secure. The homogentisic acid can be demonstrated on thin-layer chromatography and quantitated by gas–liquid chromatography or high-pressure liquid chromatography.

Genetics

It is an autosomal recessive with an incidence of only 1 in 200 000 but small populations of very high incidence exist, especially in the former Czechoslovakia and the Dominican Republic. The gene has been localized to chromosome 3q and a variety of mutations described.

Antenatal diagnosis

This has not been required but is theoretically possible.

Treatment

The amount of homogentisic acid produced is decreased by a low-protein diet. It is very probable that specifically designed low-phenylalanine and low-tyrosine diets would lower the production still further. There seems to be no demand for such a restricting diet to deal with an arthritis which begins only in adult life and progresses slowly over many years. Ascorbic acid may slow the rate of oxidation of homogentisic acid to pigment precursors but there are no data on its clinical usefulness. Theoretically NTBC may be beneficial, but its current high cost would discourage trials and the longer-term toxicity not known.

Albinism

Tyrosinase deficiency in melanocytes prevents the conversion of *p*-hydroxyphenylalanine to dihydroxyphenylalanine and thence to dopaquinone, the precursor for pigment formation in the skin, the iris, the fundus, and the inner ear. The absence of pigment is the characteristic of the group of disorders referred to together as albinism. It is a complex group of ten or more types. The manifestations are primarily in the skin and eye.

The three main types are compared in Table 10. However, two points worth noting are: oculocutaneous albinism may also occur in association with a bleeding tendency—the Hermansky Pudlak syndrome—and in association with the leucocyte killing defect—the Chédiak–Higashi syndrome. Ocular albinism, too, in some genetic forms, occurs in association with nerve deafness.

Oculocutaneous albinism is characterized by structural optic tract defects. All the fibres at the optic chiasma cross over so there are no ipsilateral fibres and no binocular vision. The geniculate bodies and the radiation onwards to the cortex are also structurally abnormal. The inner ear lacks pigment that is normally said to be protective against noise trauma. The predisposition to squamous carcinoma of the skin is important. Further details are given in Table 10.

Disorders of sulphur amino acid metabolism

The trans-sulphuration pathway transfers the sulphur of methionine to serine to produce cysteine (Fig. 8). Methionine adenosyltransferase, with widely distributed isoenzyme forms, produces *S*-adenosylmethionine, the donor in a variety of methylation reactions. In creatine formation alone adult males may utilize more methyl groups than provided by dietary methionine. *S*-adenosyl homocysteine is cleaved to homocysteine, the sulphhydryl compound which exists in reversible equilibrium with its disulphide homocystine. Half of the homocysteine formed goes through the trans-sulphuration pathway and the other half takes a methyl group from betaine (betaine methyltransferase) or 5-methyltetrahydrofolic acid (methionine synthase). The latter is a cobalamin-dependent enzyme which is functionally impaired in defects of vitamin B_{12} metabolism. The remethylation of homocysteine is also impaired if the activity of the reductase that generates 5-methyltetrahydrofolate is inadequate.

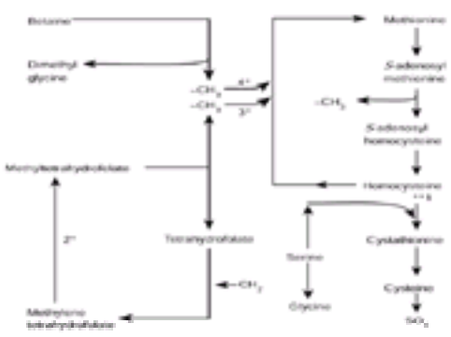

Fig. 8 The trans-sulphuration pathway from methionine to cysteine is shown on the right and the remethylation of homocysteine on the left. Asterisked enzymes are: 1, cystathionine synthase; 2, methylene tetrahydrofolate reductase, 3, methionine synthase; and 4, betaine methyltransferase.

When accumulation of homocystine results from defects of homocysteine remethylation plasma methionine concentrations are low. They are high when homocystine accumulates from impaired activity of cystathionine synthase, which forms the thioether cystathionine, an intermediate subsequently cleaved to produce the sulphydryl compound cysteine. Further metabolism of cysteine produces inorganic sulphate for excretion.

Cystathionine b synthase deficiency (homocystinuria)

Clinical

The classic clinical features in the older child and adult are mental retardation, lens dislocation, a thrombotic tendency, and skeletal abnormalities. Mental retardation, affecting two-thirds of patients, is sometimes gross but more commonly IQ values are around 65. Others are in the normal range with a few high values. Patients responsive to pyridoxine (vitamin B_6) (see below) have generally higher IQ values than non-responsive patients. Seizures affect about one-fifth and a few patients show extrapyramidal features, sometimes with severe involuntary movements. Psychiatric disturbances have been described but an increased frequency of schizophrenia is unproven.

Lens dislocation is acquired, usually in the preschool years, but later dislocation is well recognized especially in pyridoxine-responsive patients, and a few have not developed it even in adult life. Monocular and binocular blindness has been relatively frequent due to secondary glaucoma, staphyloma formation, buphthalmos, and retinal detachment.

The skeletal abnormalities include osteoporosis and spontaneous crush vertebral fractures. The common abnormalities seen in Marfan's syndrome—high arched palate, pectus excavatum or carinatum, genu valgum, pes cavus or planus, scoliosis—are all well recognized in homocystinuria. Arachnodactyly is less common and the fingers not infrequently (and elbows occasionally) show mild flexion contractures. Skeletal disproportion with a crown pubis length less than the pubis heel length is usual (Fig. 9).

Fig. 9 Child with cystathionine synthase deficiency. Note the kyphosis and short trunk.

Pathology

Thromboembolism is a major cause of morbidity and the main cause of the relatively high premature mortality. Thromboses have been described in a wide variety of arteries and veins: cerebral, coronary, mesenteric, renal, and peripheral. About 50 per cent are in peripheral veins with associated pulmonary emboli in many. Postoperative and postpartum thrombotic risks are high. Premature atheromatous vascular degeneration has been described, as has arterial aneurysm formation.

Homocysteine may interfere with crosslinking in collagen. Degeneration of zonular fibres around the lens causes the lens dislocation but these fibres are not collagen. Recent work on fibrillin in Marfan's syndrome suggests that defects in this protein may be important in cystathionine b synthase deficiency. There is still no accepted explanation for the relationship of homocystine/homocysteine to endothelial damage, platelet abnormalities, thromboses, and vascular change. Heterozygotes for the enzyme defect may be disposed to premature vascular disease and thrombosis. Finally, although the cerebral hemispheres normally have a high concentration of cystathionine, which is reduced in cystathionine b synthase deficiency, this is not considered a cause of the mental deficiency, and neither does diffuse vascular disease seem relevant to this problem.

Biochemistry

Elevated plasma methionine values between 100 and 500 µmol/litre (sometimes higher) are seen with plasma homocystine values of 50 to 200 µmol/litre (Fig. 8). A mixed disulphide (half homocysteine, half cysteine) is always present at concentrations somewhat below homocystine. Total homocysteine measured by high-performance liquid chromatography is used by some laboratories for diagnosis and monitoring treatment. This includes both homocysteine moieties of homocystine, the homocysteine moiety of the mixed disulphide, and the homocysteine bound to plasma proteins. The urinary excretion of homocystine is usually 250 to 1000 µmol/day, which accounts for only about 10 to 20 per cent of ingested methionine sulphur. The active cystathionine b synthase apoenzyme, which requires pyridoxal phosphate, is a tetramer of 63 kDa units found predominantly in liver but also in brain and intestinal mucosa. Much lower levels of activity can be found in cultured fibroblasts and stimulated lymphocytes. Residual hepatic activity of 1 to 2 per cent occurs in affected patients, this may increase two- to fourfold in pyridoxine-responsive cases. In some patients higher residual activities up to 9 to 10 per cent have been found. Heterozygotes have 25 to 45 per cent of normal activity. *In vitro* responsiveness to pyridoxal phosphate can also be detected in cultured fibroblasts.

Diagnosis

The urine gives a positive nitroprusside test (it is also positive in cystinuria). The amino acid defects are diagnostic if the plasma is deproteinized promptly to minimize binding of homocystine to protein. Plasma methionine concentrations are usually well above the normal values of 15 to 30 µmol/litre and homocystine is present in plasma and urine.

Genetics

The disease is an autosomal recessive with a birth incidence of about 1 in 40 000. The gene is on chromosome 21 with over 50 mutations already described.

Antenatal diagnosis

This has so far rested on enzyme assays on cultured amniotic cells. It is likely that work on the mutant gene will supersede this.

Treatment

Oral pyridoxine may rapidly reduce methionine and homocystine to near normal values. It should be the first treatment to try using 150 to 300 mg/day in the older child or adult and reducing the dose if a response is achieved. Very large sustained doses (1000 mg/day or more) in adults cause peripheral neuropathy. A very low-protein diet with a system of exchanges is appropriate for those not responding to pyridoxine and requires a methionine-free amino acid supplement, minerals, and vitamins. Biochemical control may only be achieved in older children and adults on natural protein intakes of 5 to 10 g/day. Cystine supplementation of diets should be considered in patients partially responsive to pyridoxine. Both folic acid (5 to 10 mg/day) and betaine (up to 12 g/day) can further reduce plasma homocystine levels but may produce large elevations of plasma methionine. Low red cell folate values occur and even megaloblastic anaemia. Low serum vitamin B_{12} values have also been found. The relationships between homocystine, the mixed disulphide, and total homocyteine values are not linear, making target values for treatment difficult to establish. Effective treatment lowers the incidence of vascular events.

Defects of homocysteine remethylation

Two defects have been described: a deficiency of methylene tetrahydrofolate reductase and a deficiency of methionine synthase (methyltetrahydrofolate homocysteine methyltransferase). The latter requires methylcobalamin as coenzyme.

Methylene tetrahydrofolate reductase deficiency

Clinical

Neurological features predominate with psychomotor retardation, seizures, abnormalities of gait, and psychiatric disturbance. Presentation occurs from early to late childhood. The risk of vascular disease is high.

Pathology

At autopsy dilated ventricles and low brain weight have been seen; thromboses may be present in arteries and veins. Demyelination occurs and the changes may resemble the classic findings of subacute combined degeneration seen in vitamin B_{12} deficiency. Calcification of the basal ganglia occurs.

Biochemistry

Plasma methionine concentrations are below normal and plasma homocystine concentrations in the range 20 to 200 µmol/litre with an excretion of 15 to 600 µmol/day.

Diagnosis

Homocystine is easily missed at low concentrations but is the important clue. The enzyme can be assayed in liver or fibroblasts.

Genetics and prenatal diagnosis

It is an autosomal recessive and enzyme assays on cultured amniocytes have been used for prenatal diagnosis. Several mutations have already been described.

Treatment

Betaine in large doses lowers plasma homocystine and raises plasma methionine. Other treatments tried alone or in combination include folinic acid, vitamin B_{12}, pyridoxine, and methionine. Some have suggested a 'cocktail' of all these treatments. It is difficult to be sure of clinical success.

Methionine synthase deficiency

The enzyme transfers a methyl group from methyltetrahydrofolate to homocysteine. Methyl cobalamin is the required coenzyme. This metabolic step may be impaired

by an apoenzyme defect or defects in cobalamin metabolism, some of which limit only the formation of methyl cobalamin. Other cobalamin defects are considered under methyl malonic acidaemia.

Clinical

The characteristic findings are developmental delay and megaloblastic anaemia, but the onset may be in later in childhood with dementia and spasticity. Retinal degeneration, cardiac defects, and haemolysis have been described.

Biochemistry and diagnosis

The findings include low plasma methionine and raised homocystine in plasma and urine. Methylmalonic acid should be measured in urine to exclude other cobalamin defects (see methylmalonic aciduria). Methione synthase can be assayed in liver or fibroblasts and antenatal diagnosis has been carried out on cultured amniocytes.

Treatment

This may involve large doses of hydroxocobalamin with betaine and possibly folinic acid.

Other defects of sulphur amino acid metabolism

Among several known defects, cystathioninuria due to cystathionase deficiency is probably clinically harmless. Cystathionine in excess of 1 g/day may be excreted at clearance values close to the glomerular filtration rate.

Methionine adenosyl transferase deficiency causes raised plasma methionine levels (up to 1200 µmol/litre; normal 15 to 30 µmol/litre) which seems to be harmless. The enzyme defect is partial.

Neither of these defects is considered further but sulphite oxidase deficiency is clinically important.

Sulphite oxidase deficiency

Most cases are due to abnormalities of the molybdenum cofactor, which therefore also affects the action of xanthine oxidase and aldehyde oxidase.

Clinical

Lens dislocation occurs, with severe neurological abnormalities, delayed psychomotor development, and xanthinuria. The neurological defects include seizures and axial hypotonia with increased limb tone. The disease is fatal.

Biochemistry

Sulphite concentrations are raised and sulphite is excreted in the urine. Direct reaction in the body between sulphite and cysteine yields sulphocysteine. Plasma urate levels are low and urine xanthine is increased when the disease is due to cofactor abnormalities but not if the defect is in the apoenzyme of sulphite oxidase.

Diagnosis

There is a dipstick test for sulphite which must be applied to fresh urine. *S*-sulphocysteine can be detected on an amino acid analyser. Sulphite oxidase can be measured in fibroblasts or liver.

Genetics and prenatal diagnosis

It is an autosomal recessive disorder. Prenatal diagnosis has been carried out on cultured amniocytes by enzyme assay.

Treatment

No effective treatment is known. Some damage may be prenatal. Measures that could be considered include diets low in methionine and cystine. Penicillamine might lower sulphite concentrations by binding with it. The nature of the molybdenum-containing cofactor is not well enough understood to be a useful therapeutic approach.

Defects of glycine metabolism

Folate and activated 1-carbon units

Tetrahydrofolate carries 1-carbon units—methyl, methylene, methenyl, formyl, or forminino—bonded to the N-5 or N-10 nitrogen atoms and the units are interconvertible. One-carbon units are donated from the tetrahydrofolate derivatives in a variety of syntheses. New 1-carbon units are accepted by tetrahydrofolate in degradative reactions, of which the most important is the conversion of serine to glycine. As serine can be formed from 3-phosphoglycerate, carbohydrates are the ultimate source of 1-carbon units (Fig. 10).

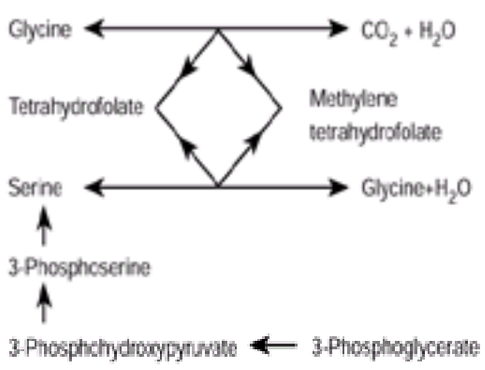

Fig. 10 Reversible glycine cleavage to carbon dioxide and water is illustrated together with reversible interconversion of serine and glycine. These reactions also serve to generate 1-carbon units. 3-phosphoglycerate (glycolysis) is the ultimate source.

The glycine cleavage system

This system, which generates methylene tetrahydrofolate from carbon-2 of glycine, and carbon dioxide from carbon-1, consists of four mitochondrial proteins. The P protein is a decarboxylase requiring pyridoxal phosphate. The heat-resistant H protein contains lipoic acid and carries the aminomethyl moiety. Both proteins are needed to generate carbon dioxide from the carbon-1 of glycine. The T protein requires tetrahydrofolate and produces methylene tetrahydrofolate from carbon-2 of glycine. The fourth protein (L protein) is needed to transfer hydrogen from the lipoic acid moiety of the H protein to nictotinamide adenine diphosphate. Reversal of the

sequence synthesizes glycine. Glycine can be converted to glyoxylate and to †-aminolaevulinic acid for porphyrin synthesis.

Non-ketotic hyperglycinaemia

Clinical

Twenty four to 48 h after birth, lethargy, convulsions, anorexia, poor feeding, and vomiting progress to coma and unresponsiveness. Apnoea may require ventilation at least temporarily. The mortality at this stage is high. Intellectual development does not occur in survivors, seizures persist, and tendon reflexes are increased. Microcephaly, poor head control, profound etardation, and a picture of spastic cerebral palsy result. Hiccupping *in utero* maybe recognized retrospectively.

There is a later childhood form presenting with spastic paraparesis, clonus, and extensor plantar responses with modestly raised plasma and cerebrospinal fluid glycine values. Optic atrophy with cerebellar signs has also been described.

Biochemistry

The defect is in the glycine cleavage system with plasma glycine values of 600 to 1200 µmol/litre. Normal values for cerebrospinal fluid levels of glycine are around 4 to 5 µmol/litre, the cerebrospinal fluid plasma ratio being around 0.02. Cerebrospinal fluid values are greatly increased in patients, raising the cerebrospinal fluid:plasma ratio to between 0.07 and 0.30. Large quantities of glycine appear in the urine, but this is not accompanied by proline or hydroxyproline.

Diagnosis

This rests on the analysis of glycine concentrations in plasma and cerebrospinal fluid. Activity of the glycine cleavage system can be measured on liver biopsies and in a few laboratories in leucocytes.

Genetics

The variant forms are autosomal recessives. The P protein is absent in classic phenotypes. T protein defects have been found in different phenotypes and H protein defects in later onset degenerative forms. Hyperglycinaemia seems to be commoner in Japan and Finland. Different mutations in these two populations affect the P protein.

Antenatal diagnosis

The enzyme system is unstable and not present in fibroblasts or cultured amniotic cells. Chorionic villi are being used for enzyme assay in prenatal diagnosis combined with amniotic fluid glycine:serine ratios. Increasing information on causative mutations will facilitate antenatal diagnosis.

Treatment

This is very unsatisfactory. Some damage to the central nervous system may be prenatal. Plasma glycine levels can be lowered by exchange transfusion or peritoneal dialysis but without clinical improvement. Low-protein diets have only a limited effect on decreasing plasma glycine concentrations. Supplying 1-carbon units in the form of methionine or *N*-formyltetrahydrofolate has not helped. The combination of sodium benzoate to increase glycine excretion and diazepines, which compete for inhibitoryglycine receptors in the central nervous system, has lowered plasma and cerebrospinal fluid levels of glycine and reduced seizures without clearly improving prognosis. Glycine is also a coagonist at the excitatory *N*-methyl-D-aspartate (NMDA) receptor blockage which has been attempted with several agents. Success has been absent or very limited. Imipramine may warrant further trial.

Defects in branched chain amino acid (leucine, isoleucine, and valine) metabolism

These essential amino acids, with a branched carbon chain structure, collectively make up 10 to 15 per cent of animal protein and are catabolized by transamination to the corresponding keto acids, 2-keto-isocaproic, 2-keto-3-methylvaleric, and 2-keto-isovaleric acids (Fig. 11). In all tissues except the liver aminotransferase activity exceeds a-ketodehydrogenase activity. Peripheral tissues, notably muscle, predominantly transaminate but the keto acids are largely transported back to the liver for subsequent metabolism.

Fig. 11 Branched chain amino acid metabolism. Transamination produces the keto acids (top) all of which are metabolized by the branched chain a-ketodehydrogenase complex (asterisked) 1. 2, Propionyl coenzyme A carboxylase; and 3, methylmalonyl coenzyme A mutase.

Branched chain a-ketodehydrogenase: the role of thiamine

The oxidative decarboxylation of branched chain keto acids is analogous to the oxidative decarboxylation of pyruvate and a-ketoglutarate to acetyl coenzyme A and succinyl coenzyme A, respectively. All are three-subunit mitochondrial enzymes, the first part of which, E_1, uses thiamine pyrophosphate as a coenzyme. The thiamine moiety is crucial to the decarboxylase function of branched chain a-ketodehydrogenase (E_1) and the release of carbon dioxide. Branched chain a-ketodehydrogenase (E_2) is the core protein of the complex, the acyl transferase that generates acyl coenzyme A while its lipoate moiety is reduced. The third part (E_3) regenerates the oxidized lipoate and is actually shared by all three dehydrogenase complexes. Branched chain a-ketodehydrogenase (E_1) is active in a dephosphorylated form and inactivated by phosphorylation, which provides a control mechanism. Branched chain ketoaciduria (maple syrup urine disease) arises from defects in the branched chain a-ketodehydrogenase complex. Some patients have a thiamine responsive form of this disease (see below).

Branched chain ketoaciduria

Clinical

In the classic disease the baby is well for 2 to 3 days and then poor feeding and sleepiness progress to coma and apnoea. Vomiting is inconstant. The mortality is high and survivors show dystonia, psychomotor retardation, spastic quadriplegia, and other neurological abnormalities.

Milder forms of the disease are described, sometimes with later presentation and intermittent forms where patients may be biochemically normal between attacks but succumb during intercurrent infection or illness or excessive protein intake.

Pathology

Myelin abnormalities that occur in patients dying of branched chain ketoaciduria are also found in Poll-Hereford calves with the same genetic defect, and in other experimental animal models.

Biochemistry

In the acute stage, hypoglycaemia and hyperammonaemia may occur. Leucine values may be as high as 4000 to 5000 µmol/litre. Isoleucine and valine are also much increased in plasma and urine (see Table 1 for normal values.) The three keto acids cause mild metabolic acidosis and the sweetish smell of maple syrup in urine. Residual enzyme activity in fibroblasts is 1 to 2 per cent for the classic severe disease but 20 to 40 per cent of normal in mild variants.

Diagnosis

The plasma amino acids and urine keto acids are diagnostic. Diagnosis before 6 days of age carries a better prognosis than later diagnosis with patients discovered by neonatal screening doing best of all.

Genetics

This is an autosomal recessive disorder. Screening is possible by bacterial assay but the disease is too rare to justify the cost. The incidence is about 1 in 120 000 in Europe but 1 in 200 000 in most of the United States, although an incidence of more than 1 in 1000 has been recorded in a Mennonite community. As the E_1 component of the branched chain a-ketodehydrogenase is subdivided further into E_{1a} and E_{1b} at least four genes code for the complex, plus two genes for the controlling phosphatase and kinase. Enzyme assays and immunological and complementation studies have already revealed genetic defects in E_{1a}, E_{1b} and E_2 in different families. The Mennonite mutation is an asparagine substitution for tyrosine in the E_{1a} subunit.

Prenatal diagnosis

This has been based on enzyme assays in cultured amniocytes or chorionic villus samples.

Treatment

A high calorie intake, given parenterally as 10 to 20 per cent dextrose if necessary, is needed to suppress nitrogen catabolism in the acutely ill. An amino acid mixture excluding leucine, isoleucine, and valine can be introduced by nasogastric tube to provide 2 g protein/kg/day. Normal protein sources (milk, etc.) are omitted until the branched chain amino acid concentrations fall towards normal. Both exchange transfusion and peritoneal dialysis have been used to speed biochemical recovery but haemofiltration is thought to be better. Hypoglycaemia, sepsis, and hypotension need intensive care and monitoring. Dietary treatment is lifelong but needs frequent adjustments. The aim is to keep plasma leucine, isoleucine, and valine concentrations close to their normal values (see Table 1). Coma carries a poor prognosis for subsequent development and function of the central nervous system. The incidence of impaired intellect and neurological handicap is high and special schooling will be necessary.

Responsiveness to thiamine(10 to 20 mg/day) has also been described in a few patients. It is claimed that large doses up to 500 mg/day improve some cases of classic branched chain ketoaciduria. *In vitro* evidence indicates that the E_{1a} subunit is stabilized by thiamine supplements, which may saturate all subunits. An increase in enzyme activity has even been described in normal subjects on thiamine treatment.

Other defects of branched chain amino acid metabolism

Rare cases of defective deamination have been described causing isolated hypervalinaemia or hyperleucinaemia–isoleucinaemia, indicating either separate amino transferases in humans or different mutations affecting different substrate binding sites in a common enzyme.

The organic acidaemias in branched chain amino acid metabolism

The catabolic steps outlined in Fig. 11 illustrate the formation of isovaleric acid, propionic acid, and methylmalonic acid, each of which accumulates in one of the three more common organic acidaemias. In the further metabolism of two of these acids there are important vitamin coenzymes—biotin for priopionyl coenzyme A carboxylase and cobalamin for methylmalonyl coenzyme A mutase. Biotin metabolism is considered under multiple carboxylase deficiency later and cobalamin metabolism immediately below. A range of other organic acidaemias have been described after discovery by gas–liquid chromatography with mass spectroscopy; their diagnosis may beome more frequent with the wider diagnostic use of tandem mass spectroscopy. They have been reported only rarely to date and are not considered further.

Vitamin B_{12} metabolism

Vitamin B_{12} has a complex metabolism but is required in only two metabolic steps—the remethylation of homocysteine to methionine and the conversion of methylmalonyl coenzyme A to succinyl coenzyme A. An outline of cobalamin metabolism in the body is shown in Fig. 12. In the cytosol hydroxocobalamin may become the coenzyme methyl cobalamin, which is required by methionine synthase, or be transported into the mitochondria to be metabolized to adenosyl cobalamin, the coenzyme of methylmalonyl coenzyme A mutase.

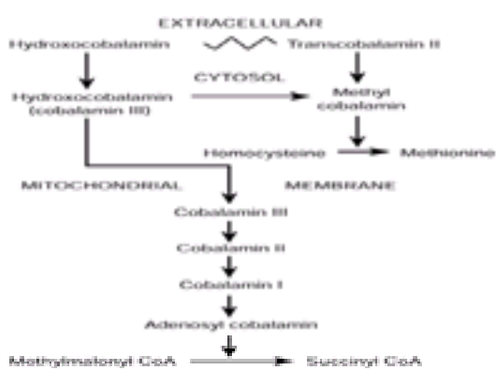

Fig. 12 Naturally occurring cobalamin is converted in the cytosol to methyl cobalamin, or adenosyl cobalamin is eventually formed by successive valency reductions of the cobalt moiety within the mitchondria.

Isovaleric, propionic, and methylmalonic acidaemias

Clinical

One to several days after a normal pregnancy and delivery the child stops feeding. Respiratory problems ensue with varying tonal change, both axial hypotonia and episodes of generalized hypertonia and myoclonic jerking. Apnoea, coma, and death supervene. Characteristically the child is acidotic, possibly ketotic, and non-specific increases of ammonia and glycine may occur. Both hypoglycaemia and hyperglycaemia have been described, the latter causing confusion with diabetic

ketoacidosis. Hypocalcaemia is also found. Early mortality is high and patients are often difficult to treat. Survivors have recurrent episodes of decompensation. There is an abnormal body odour likened to sweaty feet in isovaleric aciduria.

A more chronic form of these diseases is recognized, with anorexia, failure to thrive, psychomotor retardation, hypotonia, and weakness. Cardiomyopathy has been reported as a late complication. Damage to the basal ganglia with movement disorders is common and chronic renal failure may develop in survivors with methylmalonic aciduria.

The intermittent clinical forms present as recurrent attacks of encephalopathy and ataxia with normality between attacks. Changes in blood glucose may again be confusing (see above). Acute attacks may be followed by neurological abnormalities of a pyramidal or extrapyramidal nature. Leucopenia and thrombocytopenia sometimes occur.

Biochemistry

Isovaleric acidaemia is due to a deficiency of isovaleryl coenzyme A dehydrogenase and is characterized by the excretion in the urine of isovaleric acid, isovalerylglycine, 3-hydroxy isovaleric acid, and isovalerylcarnitine.

Isolated propionic acidaemia is due to a deficiency of the apoenzyme for propionyl coenzyme A carboxylase, a biotin-requiring enzyme. The enzyme converts proprionyl coenzyme A to methylmalonyl coenzyme A. Characteristically, plasma and urine propionate values are raised with the formation of methylcitrate from the condensation of propionyl coenzyme A with oxaloacetate (Fig. 13). Propionylcarnitine excretion is increased.

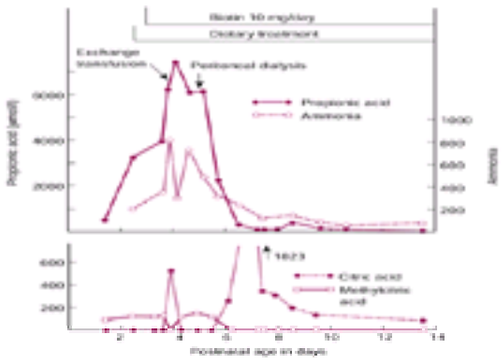

Fig. 13 Neonatal propionic acidaemia with hyperammonaemia, raised plasma methylcitrate levels, and low levels of citrate (µmol/litre). Treated by diet, exchange transfusion, and peritoneal dialysis. (From Brenton and Krywawych, unpublished data.)

Methylmalonic acidaemia is due to deficient activity of methylmalonyl coenzyme A mutase, the enzyme converting methylmalonyl coenzyme A to succinyl coenzyme A, which requires adenosyl cobalamin. Two apoenzyme defects are described, one with virtually zero activity and one with residual activity of 2 to 75 per cent of normal. Two genetic defects in the formation of adenosyl cobalamin have been described. One affects the formation of both adenosyl and methyl cobalamin, resulting in methylmalonic aciduria and homocystinuria. The other affects only adenosyl cobalamin, and only methylmalonic aciduria occurs. Patients with severe apoenzyme defects excrete up to 5 to 6 g/day of methylmalonic acid with high blood concentrations up to 6 mmol/litre (Fig. 14). Propionate also accumulates in the blood and is excreted together with methylcitrate.

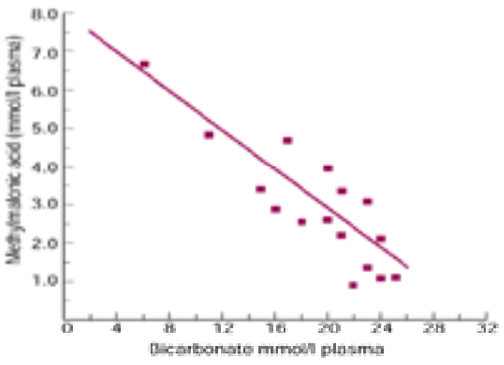

Fig. 14 Plasma concentrations of methylmalonate (a dicarboxylic acid) and bicarbonate in an affected teenage girl indicating that the acidosis is due almost entirely to the methymalonate. (From Brenton and Krywawych, unpublished data.)

Diagnosis

Diagnosis rests upon the detection of the relevant organic acids, their conjugates, or their carnitine esters in blood and urine.

Genetics

All three diseases are autosomal recessive. Isovaleryl coenzyme A dehydrogenase is a four-unit homopolymer with a single locus on the long arm of chromosome 15. Different enzyme variants cause phenotypic variation but severe neonatal and intermittent forms have been described in the same family.

Propionyl coenzyme A carboxylase has the subunit structure a_6/b_6. The a subunit gene is on chromosome 13 and the b subunit gene is on chromosome 3. Defects in the a chain (which binds the biotin) are associated with 50 per cent enzyme activity in heterozygotes and 1 to 5 per cent activity in homozygotes. Homozygous b-chain defects are similarly severe but heterozygotes have near normal activity. b chains are produced in half-normal amounts. b chains are normally produced in excess of a chains.

Methyl malonyl coenzyme A mutase is a dimer of subunit size 75 000 with adenosyl cobalamin bound to each subunit. The gene locus is on chromosome 6. The mutant mutase with no residual enzyme activity has no detectable enzyme protein, either because none is made or because it is highly unstable.

There is now considerable information on the causal mutations in all three diseases.

Prenatal diagnosis

Isovaleric acid in amniotic fluid is measured reliably by stable isotope dilution analysis, and isovaleryl coenzyme A dehydrogenase can be measured in cultured amniocytes. The measurement of methylcitrate in amniotic fluid and enzyme assay in cultured amniocytes has been used for diagnosis of propionic acidaemia. Similar approaches to prenatal diagnosis in isolated methylmalonic aciduria have used the measurement of methylmalonate acid in amniotic fluid and enzyme assays or studies of adenosyl vitamin B_{12} metabolism in cultured amniocytes. Molecular prenatal diagnosis will be increasingly used in families where the genotype is known.

Treatment

In the severe neonatal form of these diseases the initial treatment is concerned with removal of toxic organic acids by exchange transfusion (as urinary excretion of propionate is poor this may be followed in propionic acidaemia by peritoneal dialysis) and encouraging anabolism by the provision of calories as 10 to 20 per cent glucose and electrolyte solutions intravenously with or without insulin. Enteral feeding should be started by nasogastric tube as soon as possible (after 24 to 48 h); initially this should be with protein-free feeds, but soon changing to a low-protein feed (0.5 g/kg/day) and later increasing to tolerance and supplemented with amino acid mixtures that omit the amino acids whose metabolism is impaired. The requirements of these amino acids for growth are provided by the natural protein, whose intake must be adjusted accordingly. L-glycine supplements of 0.25 to 0.5 g/kg/day are helpful in isovaleric acidaemia because it increases the formation of the non-toxic isovalerylglycine. L-carnitine 100 mg/kg/day may help in all three diseases by replenishing carnitine and increasing the excretion of non-toxic carnitine acyl esters. Both insulin and growth hormone have been used to try and produce positive nitrogen balance and hasten recovery in catabolic states.

No true *in vivo* responsiveness to biotin has been demonstrated in isolated propionic acidaemia. However, *in vivo* response to hydroxocobalamin therapy in methylmalonic acidaemia occurs and should be tested in all such patients and continued long term if response occurs. Diet is needed long term in all three disorders; this is relatively easy in isovaleric acidaemia where a low-protein diet may suffice. A low-protein diet may also suffice in some patients with methylmalonic acidaemia, combined with regular oral sodium bicarbonate to control residual acidosis. Patients with propionic acidaemia are more difficult to manage and require a low-protein diet with more frequent supplements of amino acids. Chronic nasogastric feeding may be needed for anorexia. Oral metronizanole may reduce propionate production by gut bacteria in the intestine in propionic and methylmalonic acidaemia but therapeutic usefulness is not yet clear. Similarly, the use of L-carnitine on a chronic basis may help in all diseases but it is not proven. Patients with methyl malonic aciduria and renal failure may require renal transplantation. combined hepatic and renal transplantation carried out to cure the underlying metabolic defect has also had very variable outcome.

Disorders of g-aminobutyric acid metabolism

g-aminobutyric acid is formed from glutamate in the brain by the cytosolic enzyme glutamate decarboxylase, which requires pyridoxal phosphate. Pyridoxine-dependent seizures in neonates are postulated to be due to a deficiency of this enzyme, which is difficult to prove because other tissues have a genetically different mitochondrial glutamate decarboxylase. Glutamate can be regenerated from g-aminobutyric acid by transamination with ketoglutarate (g-aminobutyric acid transaminase), which is also pyridoxal phosphate dependent. The other product is succinic semialdehyde, which is dehydrogenated to succinate, which enters the citric acid cycle. Deficiency of succinic semialdehyde dehydrogenase leads to the excretion of 4-hydroxybutyric acid. Some more details of disordered g-aminobutyric acid metabolism are given in Table 11.

Defects of lysine metabolism

Lysine catabolism

The main pathway is via saccharopine to acetyl coenzyme A (Fig. 15); there are other less important pathways. Glutaryl coenzyme A dehydrogenase catalyses the conversion of glutaryl coenzyme A to crotonyl coenzyme A. Deficiency of this enzyme causes glutaric aciduria type I, a serious disorder. Other lysine degradation defects are of uncertain clinical consequence.

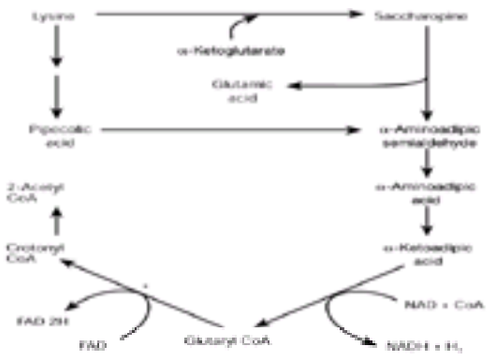

Fig. 15 The metabolism of lysine. The enzyme glutaryl coenzyme A dehydrogenase is asterisked.

Glutaric aciduria type I

Clinical

Abnormalities of development of the central nervous system begin before birth with macrocephaly and defective frontal and temporal lobe development, although early clinical development is often considered normal. Delayed motor development in the early years of life with hypotonia is followed by encephalopathic episodes precipitated by intercurrent illness with ataxia, athetosis, and other involuntary movements. Severe dystonia, pyramidal defects with extensor or flexor spasms, and severe dysarthria may follow. Intercurrent infection can also precipitate acidosis, seizures, coma, and paralysis, from which recovery is incomplete. The overall picture then resembles dystonic cerebral palsy. Computed tomography and magnetic resonance imaging have revealed progressive cerebral atrophy and hyperlucency of the caudate nucleus due to striatal necrosis. Even if the diagnosis is made in asymptomatic patients, acquired motor skills such as walking and writing may be slowly lost in the childhood years.

Biochemistry

Glutaryl coenzyme A dehydrogenase deficiency causes an accumulation of glutaryl coenzyme A (also derived from tryotophan degradation), increasing glutaric acid concentrations in plasma and urine, and increasing concentrations of 3-hydroxyglutarate and glutaconic acid. These are all inhibitors of glutamic acid decarboxylase, which may explain the low g-aminobutyric acid concentrations in the central nervous system. Neurodegeneration probably results from excessive stimulation of NMDA receptors by 3-hydroxyglutaric acid. Glutaryl carnitine is excreted in the urine even when free glutaric acid is absent. Systemic acidosis occurs in acute attacks with ketosis and hypoglycaemia.

Diagnosis

This cause of progressive dystonic cerebral palsy is usually indicated by the organic acids in plasma and urine. Sometimes the organic acids have not been detected, particularly between acute attacks. Enzyme assays on leucocytes or fibroblasts are then indicated.

Prenatal diagnosis

This has been carried out by finding glutaric acid in the amniotic fluid and enzyme assay on cultured amniocytes. Where the mutation in the family is known molecular prenatal diagnosis would be more accurate.

Genetics

The disease is an autosomal recessive. The glutaryl coenzyme A dehydrogenase gene is on chromosome 19p and over 70 mutations have been described.

Treatment

Low-protein diets reduce glutaric acid excretion. Carnitine supplementation corrects low plasma levels which are secondary to losses from glutaryl carnitine excretion. Riboflavin has been reported to diminish glutaric acid excretion in some patients, the treatment rationale being that increased flavine adenine dinucleotide might stabilize the enzyme. Baclofen has also been studied because it activates g-aminobutyric acid receptors. When treatment is started very early in life brain degeneration may be preventable. Delay results in irreversible damage to the caudate and putamen.

Defects in the final stages of carbon chain metabolism

Biotin-dependent carboxylation

Biotin is important in transferring a 1-carbon unit (carbon dioxide) to acceptor molecules. Defects in biotin metabolism disturb the function of four enzymes—pyruvate carboxylase, acetyl coenzyme A carboxylase, propionyl coenzyme A carboxylase, and 3-methylcrotonyl coenzyme A carboxylase (Fig. 16). These apoenzymes are converted to holoenzymes by the attachment of biotin, which needs the catalytic activity of an enzyme, holocarboxylase synthetase (Fig. 17). When the holoenzymes are themselves biologically degraded the biotin is initially released still attached to lysine peptides. The enzyme biotinidase frees biotin from these peptides. It also liberates dietary biotin from proteins in the gastrointestinal tract. In its absence biotin peptides are excreted, dietary biotin is not absorbed, and biotin deficiency occurs. Biotinidase is widely distributed throughout the body.

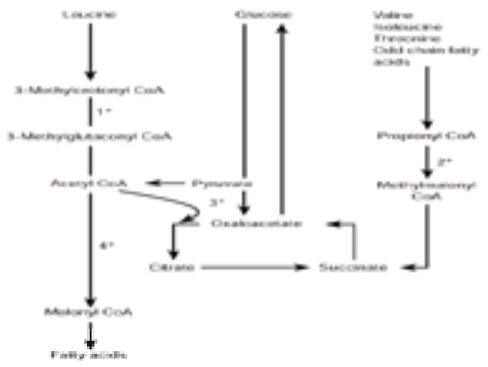

Fig. 16 Important carboxylases in amino acid metabolism. Asterisked enzymes are: 1, 3-methylcrotonyl coenzyme A carboxylase; 2, propionyl coenzyme A carboxylase; 3, pyruvate carboxylase; and 4, acetyl coenzyme A carboxylase.

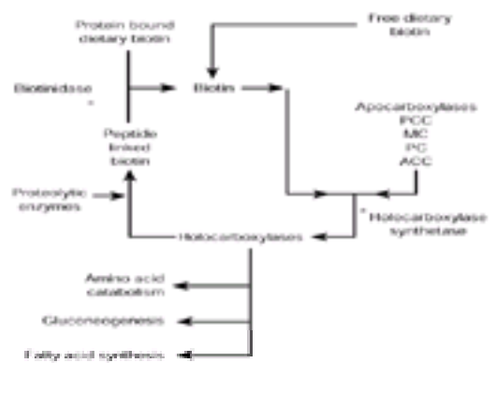

Fig. 17 The metabolism of biotin. MCC (3-methylcrotonyl coenzyme A) and PCC (propionyl coenzyme A carboxylase) are important in amino acid catabolism, PC (pyruvate carboxylase) is important in gluconeogenesis, and ACC (acetyl coenzyme A carboxylase) in fatty acid synthesis. Important enzymes are asterisked.

Electron transport and the acyl coenzyme A dehydrogenases

The electrons accumulating during oxidation in the citric acid cycle are carried by reduced nicotinamide adenine dinucleotide and reduced flavine adenine dinucleotide to be transferred along the electron transporting chain to molecular oxygen, with the generation of adenosine triphosphate and water. Transfer from reduced nicotinamide adenine dinucleotide takes place sequentially across four multienzyme complexes (I to IV), which are part of the structure of the inner mitochondrial membrane. The flavin-containing acyl coenzyme A dehydrogenases transfer electrons differently to an intermediate electron transferring flavoprotein and from there to ubiquinone catalysed by the enzyme electron transferring flavoprotein ubiquinone oxidoreductase (Fig. 18).

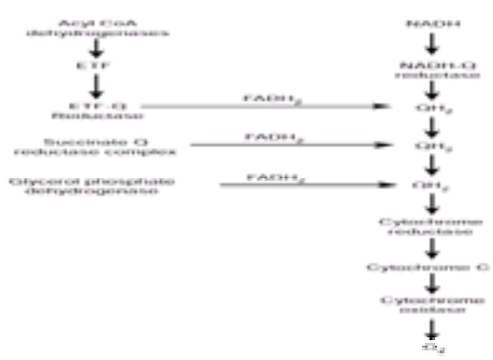

Fig. 18 The main electron transporting chain from reduced nicotinamide adenine dinucleotide (NADH) to oxygen is shown on the right, with other entry points for the flow of electrons coming from the left. ETF, electron transporting flavoprotein; $FADH_2$, reduced flavin adenine dinucleotide; QH_2, reduced ubiquinone.

Defects at this level affect not only amino acid catabolism but fatty acid oxidation, and the organic acid defects are complex. The affected acyl coenzyme A dehydrogenases include glutaryl coenzyme A dehydrogenase and defects in electron transport at this point in metabolism are labelled glutaric aciduria type II.

Multiple carboxylase deficiency

Clinical

Holocarboxylase synthetase deficiency causes neonatal acidosis with seizures, skin rash, and alopecia; it progresses to coma and death. Vomiting and ketosis are present. Biotinidase deficiency has a more variable clinical picture with progressive neurological deterioration including ataxia and seizures, developmental delay, and hypotonia. Other features of the neonatal form such as skin rash, alopecia, acidosis, and organic aciduria may not be prominent. Hearing loss and optic atrophy have

been described. Keratoconjunctivitis occurs.

Biochemistry

The carboxylase deficiencies cause a complex organic aciduria. Isovaleric acid (which imparts an unpleasant odour to the patient), 3-hydroxyisovaleric acid, methylcrotonic acid, and methylcrotonyl glycine result from the impaired activity of 3-methylcrotonyl coenzyme A carboxylase. Lactic acidosis, with more marked increases in cerebrospinal fluid levels of lactate, reflects defective pyruvate carboxylase activity. Impaired propionate metabolism also increases 3-hydroxypropionate and propionylglycine excretion. The accumulating acetyl coenzyme A results in ketosis.

Diagnosis

Apart from organic acid analyses biotinidase activity in plasma is reduced to 0 to 5 per cent of normal in genetic deficiency. Biotin itself can be measured in plasma and urine. The assay of holocarboxylase synthetase is difficult and possible in only a few places. The therapeutic response to biotin does not distinguish between the two defects.

Genetics

Both are recessive disorders and biotinidase deficiency seems to be more common than holocarboxylase synthetase deficiency. The gene for the latter has been assigned to chromosone 21 and several mutations described.

Prenatal diagnosis

This is only required in holocarboxylase synthetase deficiency and depends on amniotic fluid analysis for organic acids and enzyme assay in cultured amniotic cells.

Treatment

Biotinidase deficiency responds well, often dramatically, to 5 to 10 mg/day of oral biotin. Deficiency develops in a few days if biotin is stopped. Pre-existing neurological damage may not reverse. Most patients with holocarboxylase deficiency respond well to 10 mg daily, but larger doses may be needed and some have not fully responded to doses as high as 100 mg/day.

Glutaric acidaemia type II

Clinical

The most severe neonatal presentation, with associated congenital abnormalities, often leads to premature birth, metabolic acidosis, hypoglycaemia, hepatomegaly, and hypotonia. Severe cystic dysplasia of the kidneys is common; the kidneys may be palpable. Other defects include facial dysmorphism, 'rocker-bottom' feet, anterior abdominal wall defects, and defects of the external genitalia. Death usually occurs in the first week of life. Some affected neonates, without congenital defects, have the other clinical abnormalities of metabolic acidosis, hypoglycaemia, hypotonia, and hepatomegaly. The prognosis remains poor, with death in the early days or weeks of life, often with severe cardiomyopathy.

Milder forms presenting after the neonatal period, or survivors of early illness, may suffer recurrent encephalopathic episodes similar to Reye's syndrome. Cases with a predominantly late clinical presentation of lipid storage myopathy have been described. Adult presentation has been recorded. From their predominant organic acid pattern, some of these clinically milder patients are given the diagnosis of ethylmalonic-adipicaciduria.

Biochemistry

Glutaric aciduria type II is due to deficiency of electron transferring flavoprotein or electron transferring flavoprotein–ubiquinone oxidoreductase, the latter causing the severest neonatal form with congenital defects. The flavin-containing acyl coenzyme A dehydrogenases affected include glutaryl coenzyme A dehydrogenase, isovaleryl coenzyme A dehydrogenase, the long, medium, and short chain dehydrogenases used in fatty acid oxidation, and the dehydrogenases involved in sarcosine synthesis and breakdown. The organic acids found in urine as a consequence include short chain acids—isovaleric, 3-hydroxy isovaleric, glutaric, 2-hydroxyglutaric—the oxidation products of medium chain fatty acids—adipic, suberic and sebacic acids—ethylmalonic acid, 5-hydroxy hexanoic acid, and glycine conjugates of a variety of these. Carnitine concentrations in plasma are low and a range of acyl carnitines are found in urine and increased by carnitine therapy. Hypoglycaemia is very common.

Diagnosis

The florid organic acid pattern in severe patients is characteristic but in those more mildly affected it is less marked. Acyl carnitines are now well demonstrated by tandem mass spectrometry. Hepatomegaly and hypoglycaemia in older patients raise the diagnosis of glycogen storage diseases, but ketonaemia does not occur in glutaric aciduria type II. Electron transferring flavoprotein and electron transferring flavoprotein–ubiquinone oxidoreductase can be assayed in some centres using cultured fibroblasts.

Genetics

Both of the basic defects are autosomal recessive with assays of electron transferring flavoprotein and the electron transferring flavoprotein–ubiquinone oxidoreductase showing variable residual activity. The electron transferring flavoprotein protein has a and b subunits. The relevant genes have been localized to chromosomes 15 and 19 and mutations in both have been described.

Prenatal diagnosis

This has been carried out using amniotic fluid analysis and cultured amniocytes for electron transferring flavoprotein and oxidoreductase assays.

Treatment

Nothing has influenced severe early cases. Diets low in fats and protein reduce organic acid accumulation in milder cases and carnitine supplements increase the formation of the less toxic carnitine acyl esters. Oral riboflavin 100 to 300 mg/day has apparently been beneficial in some older patients, perhaps by stabilizing electron transferring flavoprotein or the oxidoreductase. Milder cases are helped by a high energy intake during intercurrent illness, which may need to to be intravenous.

Other defects of amino acid and organic acid metabolism

Many are not covered in the text because their rarity does not really justify it. Information is available in specialized texts.

Further reading

Adamson MD, Andersson HC, Gahl WA (1989). Cystinosis. *Seminars in Nephrology* **9**, 147–61.

Anikster Y, Shotelersuk V, Gahl WA (1999). CNS mutations in patients with cystinosis. *Human Mutations* **14**, 454–8.

Attree O *et al*. (1992). The Lowe's oculocerebrorenal gene encodes a protein highly homologous to inositol polyphosphate-5-phosphatase. *Nature* **358**, 239–42.

Batshaw ML, Bachmann C, Luckman M (1998). Advances in inherited urea cycle disorders. *Journal of Inherited Metabolic Disease* **21**, Supplement 1.

Blau N, Duran M, Blaskorvics ME (1996). *Physician's guide to the laboratory diagnosis of metabolic diseases.* Chapman and Hall, London.

Brenton DP *et al.* (1981). The adult presenting idiopathic Fanconi syndrome. *Journal of Inherited Metabolic Diseases* **4**, 211–15.

Brody LC *et al*. (1992). Ornithine delta amino transferase mutations in gyrate atrophy, allelic heterogeneity and functional consequences. *Journal of Biological Chemistry* **267**, 3302–7.

Burgard P, Link R, Schweltzer-Krantz S (2000). Phenylketonouria: Evidence-based clinical practice. *European Journal of Pediatrics* **159**, Supplement 2.

Camacho JA *et al.* (1999). Hyperornithinaemia–hyperammonaemia–homocitrillinuria syndrome is caused by mutations in a gene encoding a mitochondirial ornithine transporter. *Nature Genetics* **22**, 151–8.

Charnos LR *et al.* (1991). Clinical and laboratory findings in the oculo-cerebro-renal syndrome of Lowe with special reference to growth and function. *New England Journal of Medicine* **324**, 1318–25.

Chesney RW (1998). Mutational analysis of patients with cystinuria detected by a genetic screening network: Poweful tools in understanding the several forms of the disorder [editorial]. *Kidney International* **54**, 279–80.

Dent CE (1948). A study of the behaviour of some sixty amino acids and other ninhydrin-reacting substances on phenol-collidine filter paper chromatograms with notes as to the occurrence of some of them in biological fluids. *Biochemical Journal* **43**, 169–80.

Dhondt JL (1991). Strategy for the screening of tetrahydrobiopterin deficiency among hyperphenylalaninaemic patients: 15 years experience. *Journal of Inherited Metabolic Disease* **14**, 117–27.

Fernandes J, Saudubray J-M, van den Berghe G (1990). *Inborn metabolic diseases. Diagnosis and treatment*, 1st edn. Springer, Berlin.

Fowler B (1997). Disorders of homocysteine metabolism. *Journal of Inherited Metabolic Disease* **20**, 270–85.

Goodyer P *et al.* (1998). Cystinuria subtype and the risk of nephrolithiasis. *Kidney International* **54**, 56–61.

Haworth JC *et al*. (1991). Phenotypic variability in glutaric aciduria type I: report of 14 cases in five Canadian Indian kindreds. *Journal of Pediatrics* **118**, 52–8.

Holme E and Lindstedt S (1998). Tyrosinaemia Type I and NTBC. *Journal of Inherited Metabolic Disease* **21**, 507–17.

Kaplan P *et al*. (1991). Intellectual outcome in children with maple syrup urine disease. *Journal of Pediatrics* **119**, 46–50.

Lenke RL, Levy HL (1980). Maternal phenylketonuria and hyperphenylalaninemia. *New England Journal of Medicine* **303**, 1202–8.

Maestri NE *et al.* (1991). Prospective treatment of urea cycle disorders. *Journal of Pediatrics* **119**, 923–8.

Milliner DA (1990). Cystinuria. *Endocrinology and Metabolism Clinics of North America* **19**, 889–907.

Morton DH (1994). Through my window—remarks at the 125th year celebration of the Children's Hospital of Boston. *Pediatrics* **94**, 785–91.

Norden AG *et al.* (1991). Excretion of b_2 glycoprotein (apolipoprotein H) in renal tubular disease. *Clinical Chemistry* **37**, 74–7.

Paradis K *et al*. (1990). Liver transplantation for hereditary tyrosinaemia: the Quebec experience. *American Journal of Human Genetics* **47**, 338–42.

Rose WC *et al.* (1955). The amino acid requirements of man. XV The valine requirement. Summary and final observations. *Journal of Biological Chemistry* **217**, 987.

Rutchick SD, Resnick MI (1997). Cystine calculi: diagnosis and management. *The Urologic Clinics of North America* **24**, 163–72.

Santer R *et al.* (1998). Fanconi–Bickel syndrome—the original patient and his natural history; historical steps leading to the primary defect and a review of the literature. *European Journal of Pediatrics* **157**, 783–97.

Saudubray J-M *et al*. (1989). Clinical approach to inherited metabolic disease in the neonatal period: a 20-year survey. *Journal of Inherited Metabolic Disease* **12**, Supplement 1, 25–42.

Schneider JA *et al.* (1995). Recent advances in the treatment of cystinosis. *Journal of Inherited Metabolic Disease* **18**, 387–97.

Smith I (1993). Phenylketonuria due to phenyalanine hydroxylase deficiency: an unfolding story. Report of the MRC Working Party on P.K.U. *British Medical Journal* **306**, 115–19.

Smith I (1993). Recommendations on the dietary management of phenylketonuria. Report of the MRC Working Party on PKU. *Archives of Diseases in Childhood* **68**, 426–7.

Stephens AD (1989). Cystinuria and its treatment, 25 years experience at St Bartholomew's Hospital. *Journal of Inherited Metabolic Disease* **12**, 197–209.

Tada K, Kure S (1993). Non-ketotic hyperglycinaemia: molecular lesions, diagnosis and pathophysiology. *Journal of Inherited Metabolic Disease* **16**, 691–703.

Tuchman M, Holzknecht RA (1991). Heterogeneity of patients with late onset ornithine transcarbamylase deficiency. *Clinical and Investigative Medicine* **14**, 320–4.

Tuchman M, Knopman DS, Shih VE (1990). Episodic hyperammonaemia in adult siblings with hyperornithinaemia, hyperammonaemia and homocitrillinuria syndrome. *Archives of Neurology* **47**, 1134–7.

Van'T Hoff WG (2000). Molecular developments in renal tubulopathies. *Archives of Diseases in Childhood* **83**, 189–91.

Widhalm K *et al*. (1992). Long term follow up of 12 patients with the late onset variant of argininosuccinic acid lyase deficiency. *Pediatrics* **89**, 1182–4.

Wilcken DEL, Wilcken B (1997). The natural history of vascular disease in homocystinuria and the effects of treatment. *Journal of Inherited Metabolic Disease* **20**, 295–300.

Wolf B, Heard GS (1991). Biotinidase deficiency. *Advances in Pediatrics* **38**, 1–21.

11.3.1 Glycogen storage diseases

T. M. Cox

Glycogen, the main energy store in liver and muscle, is configured for the compact storage of glucose in a form that has a minimal osmotic effect but which is readily accessible and metabolically active. The molecule contains polymerized a- D-glucose units anchored covalently at their reducing termini to a small protein, glycogenin. The structure of glycogen is elaborate: its extensively arborized macromolecular arrangement is linked by a-1,4 glycosidic bonds with a-1,6 bonds at the branch points. These branch points are arranged in several tiers with increasingly long outer chains that terminate in non-reducing glucose residues. Thus the complex branched structure of glycogen also promotes its ready access to the enzymes of biosynthesis and degradation.

The liver and muscles contain between 200 and 300 g of glycogen and its polymerized structure can be seen with the electron microscope: liver glycogen consists mainly of a-aggregates or rosettes of smaller particles (b-particles) that are principally found in muscle cytoplasm. The molecular weight of glycogen in these tissues is several million daltons. Each b-particle contains up to 60 000 glucose residues, but despite its size the glycogen molecule undergoes remodelling as a result of constant breakdown and synthesis. Defects in the enzymatic steps for the synthesis, utilization, or degradation of glycogen lead to its pathological storage. Accumulation of glycogen may be generalized or involve certain tissues selectively; the stored glycogen may have a normal or aberrant structure.

Glycogen metabolism

The individual enzymatic steps for the formation and breakdown of glycogen are summarized in Fig. 1.

Fig. 1 The synthesis and degradation of glycogen.

Glycogen biosynthesis

The immediate precursor for glycogen synthesis is uridine diphosphoglucose (**UDPG**), which is formed from glucose 1-phosphate by UDPG pyrophosphorylase. This enzyme has a high affinity for its substrates and is abundant—no deficiencies have been recorded. In contrast, glycogen synthase is a highly regulated enzyme complex that exists in distinct isoforms in muscle and liver: the enzyme catalyses the transfer of UDP glucose units to glucose residues already covalently attached to a tyrosine residue of glycogenin, which acts as a primer. The tyrosine glucosyltransferase activity has not been identified but the glycogenin adduct possesses an intrinsic glucosyltransferase activity. Initially, one molecule each of glycogen synthase and glycogenin occur as a complex in each b-glycogen particle. After elongation, branching of the molecule is catalysed by amylo (1,4 ® 1,6) transglucosidase, 'branching enzyme'.

Glycogen synthase is subject to phosphorylation control that inhibits its activity: this inhibition is overcome by the allosteric activator, glucose 6-phosphate. The phosphorylation of at least nine serine residues is brought about by protein kinases. Glucagon and adrenaline, while stimulating phosphorylase via phosphorylase kinase, indirectly inhibit glycogen synthase by maintaining protein phosphatase I in its inactive configuration. Insulin stimulates glycogen synthase by promoting its dephosphorylation through the action of this same phosphatase: protein phosphatase I is activated by a cascade of protein kinases whose phosphorylation is initiated by the insulin receptor tyrosine kinase. Inherited deficiency of glycogen synthase activity is associated with reduced storage of liver glycogen and fasting hypoglycaemia. Branching-enzyme activity is essential for the formation of the compact spherical molecules of glycogen, especially in liver. It transfers a minimum of six a-1,4-linked glucose units from the distal ends of glycogen chains to a 1,6 position on the same or a neighbouring chain. Deficiency of branching enzyme leads to the accumulation of abnormal molecules that are partially resistant to degradation.

Glycogen breakdown

Glycogen is degraded by three enzymes: phosphorylase, debranching enzyme and acid a-glucosidase. Phosphorylase brings about the sequential release of glucose units from the a-1,4-linked chains of glycogen to liberate glucose 1-phosphate. After conversion to glucose 6-phosphate by phosphoglucomutase, free glucose is formed by the action of glucose 6-phosphatase. Debranching enzyme possesses transferase and a-1,6-glucosidase activities. When phosphorylase has degraded glycogen chains to within four a-1,4-glucosyl units of an a-1,6 linkage, three glucose residues are transferred to the end of another chain by the glycosyltransferase activity. Debranching enzyme then hydrolyses the remaining a-1,6 bond to release free glucose using its amylo-1,6-glucosidase activity. Debranching enzyme also cleaves the unique glucosyl–tyrosine linkage that anchors the terminal reducing glucose unit to glycogenin. Deficiency of debranching enzyme leads to the storage of glycogen that possesses short outer chains, 'limit dextrin'.

The main product of glycogen breakdown in muscle and liver is glucose 1-phosphate, which is produced by the sequential action of phosphorylase on a-1,4 glycosidic bonds. Glucose 1-phosphate is a key intermediate of glycolysis, gluconeogenesis, glycogenolysis, and the pentose-phosphate pathway, but, by virtue of phosphoglucomutase, the hepatic glucose 6-phosphatase system is the predominant metabolic source of blood glucose. Glucose 6-phosphatase exists as a multicomponent complex in the endoplasmic reticulum of hepatocytes and, to a lesser extent, in renal tubular cells—it is not found in muscle. The system contains glucose 6-phosphatase, several proteins that facilitate the transport of glucose, glucose 6-phosphate, and phosphate, as well as other stabilizing and regulatory moieties. Several genetic defects in this compartmentalized system are recognized to affect overall glucose 6-phosphatase activity: they are associated with severe

hypoglycaemia, metabolic acidosis, and hepatic disease.

Glucose 6-phosphate obtained from the breakdown of glycogen in skeletal muscle is used directly in glycolysis. Defects of muscle phosphorylase lead to a defective supply of adenosine triphosphate (**ATP**), especially during ischaemic exercise. There is a failure of conversion of glycogen to lactate, and exercise-induced muscle cramps reflect mild muscle necrosis with increased accumulation of glycogen. Phosphofructokinase-1 catalyses an irreversible step in the glycolytic pathway and is a key regulatory enzyme. Inherited defects that render it inactive or affect its positive allosteric regulation by the effectors adenosine monophosphate (**AMP**) and fructose 2,6-diphosphate resemble muscle phosphorylase deficiency. Because deficiency of phosphofructokinase affects the metabolism of endogenous glycogen as well as carbon units derived from extracellular glucose, the symptoms of phosphofructokinase-1 deficiency are more severe and of earlier onset than muscle phosphorylase deficiency. As expected, glucose 6-phosphate, fructose 6-phosphate, and glycogen, accumulate in the muscle cells.

Breakdown of glycogen in liver and skeletal muscle is brought about by the concerted activities of phosphorylase and debranching enzyme in the cytoplasm. Phosphorylase, which requires pyridoxal-5-phosphate, is activated by phosphorylation in response to hormonal or neural stimulation—a complex process that is mediated by phosphorylase kinases. Phosphorylase kinase is a multisubunit protein with regulatory, catalytic, and calcium-binding subunits that are encoded on separate genes. Separate isoforms are found in liver and muscle. The final common pathways for the regulation of phosphorylase kinase involve protein kinase A (cAMP-dependent protein kinase), calcium and kinase activation of calmodulin, and protein phosphatases 1 and 2A.

Another enzyme, acid a-1,4-glucosidase (otherwise known as acid maltase), has an important role in the metabolism of glycogen. This lysosomal hydrolase is present in all cells except erythrocytes and, although it has no relation to glycolysis, its deficiency causes a generalized disorder in which muscle disease, especially of the heart, is usually severe. Deficiency of acid a-glucosidase is associated with rapidly progressive cardiac hypertrophy with hepatic enlargement and generalized muscle weakness. Skeletal muscle symptoms may be prominent in patients with the infantile or late-onset forms of this condition but disease progression is usually rapid. Acid a-1,4-glucosidase deficiency was the first inborn lysosomal disease to be clearly recognized and represents a prototype for the other storage diseases: intracellular vesicles containing glycogen represent lysosomes distended by an undegradable substrate that accumulates as a result of autophagy. The accumulation of glycogen in lysosomes indicates that glycogen fragments are constantly being taken up for partial degradation and macromolecular remodelling.

Clinical features

The principal features of the different glycogen storage diseases are set out in Table 1, which also gives the primary enzymatic (or translocator) defect and chromosomal locus of the cognate human gene in each case.

Many of the manifestations of the glycogen storage disorders are common to several of these diseases and correlate with the main site of storage. However, in those disorders that affect the liver, the consequential effects of the primary metabolic lesions are often far-reaching and the function of many different tissues may be impaired as part of a pleiotropic disturbance of biochemical homeostasis. In several instances, for example glycogen storage diseases types III and IV, pathological storage affects both liver and muscle tissue (including cardiac muscle) (Table 2).

An additional set of clinical features is observed in the enzymatic defects that affect glycolysis: typically, these are associated with acute exercise-induced muscle symptoms and signs of rhabdomyolysis. These defects are usually restricted to those tissues with a high glycolytic capacity or dependence, such as muscle and red cells; mild haemolysis results from the impaired supply of ATP to the membrane sodium–potassium ATPase of the erythrocyte.

Several unusual features of the glycogen storage diseases have been reported that remain unexplained. These include the development of hepatic adenomas (which presage malignant transformation); leucocytes and macrophages in the translocator deficiencies (types 1b, c, and d) that predispose to microbial infections and granulomatous colitis, and vasoconstrictive pulmonary hypertension. Typically, the renal disease is preceded by a hyperfiltration syndrome and mild proteinuria. An unusual feature of late-onset glycogen storage disease type II due to acid maltase deficiency, has been the association with intracerebral arterial aneurysms; glycogen storage in arterial smooth muscle with prominent vacuolation has been documented.

Clinical genetics of the glycogenoses

The genes encoding the human enzymes that are defective in the individual glycogen storage diseases have been identified and mapped to their respective chromosomal loci, as indicated in Table 1. The individual disorders are inherited as autosomal recessive traits, with the exception of liver phosphorylase b kinase deficiency (type VIII) and Danon's disease (type IIb) which are X-linked diseases.

Diagnosis of glycogen storage diseases

Affecting the liver

The diagnosis may be suspected in infants and children with hepatomegaly, growth retardation, and hypoglycaemia, which is not invariable. Review of a previous biopsy may indicate glycogen deposition; glycogen deposits stain strongly within hepatocytes with the Periodic acid–Schiff reagent and the reaction characteristically is abolished by prior treatment with diastase. In many cases, a glucagon stimulation test (20 µg/kg intramuscularly) fails to induce the normal (>2 mmol/l) rise in blood glucose; however, definitive diagnosis by biopsy is warranted for prognosis, future antenatal diagnosis, and to direct treatment. Direct assay of liver tissue for glycogen and fat content as well as enzymatic analysis is desirable. Histochemical and electron microscopic study of glycogen structure provides useful additional information. Where possible, open wedge-biopsy of the liver should be carried out to obtain sufficient material for diagnosis and ensure haemostasis under direct vision; appropriate provision of platelets and blood coagulation factors should be made to correct the haemorrhagic diathesis before biopsy is carried out. However, the procedure is hazardous for young infants with acidosis or a bleeding tendency and close attention should be given to prevention of hypoglycaemia.

A particular difficulty arises in the diagnosis of certain variants of type I glycogen storage disease. The glucose 6-phosphatase system is uniquely incorporated into the endoplasmic reticulum: latency of its membrane-bound components renders diagnosis of specific lesions affecting the transport of substrates or products impossible when frozen tissue is thawed for analysis. Types 1B and 1C glycogen storage disease (in which glucose 6-phosphate translocation is defective) is an example where the study of fresh tissue is essential for establishing a diagnosis, since analysis of freeze–thawed material disrupts the integrity of the microsomal enzyme system and—by rendering it permeable to phosphate, pyrophosphate, and glucose 6-phosphate—overcomes the transport defect. Thus, where defects of glycogen storage are suspected, it is essential to seek the prior advice of a laboratory that is competent to carry out the appropriate investigations using fresh and deep-frozen biopsy material.

In muscle

Forearm exercise tests are useful for detecting defects in skeletal muscles that interfere with the supply of chemical energy in the form of ATP by the metabolic pathway that breaks down glucose and glycogen to lactate. In the absence of oxygen, glycolysis is the sole means by which ATP may be generated: the preferred energy source being glucosyl units derived from glycogen, rather than glucose obtained from the plasma. Defects in glycolysis (glycogenosis type VII and other enzyme deficiencies) cause similar symptoms. Exercise-induced cramps may occur in patients with the purine pathway disorder, myoadenylate deaminase deficiency, which may also be safely diagnosed by exercise testing. Unlike the earlier test devised by McArdle (1951), these provocative tests do not induce rhabdomyolysis accompanied by raised creatine kinase activity in the serum with acute myoglobinuric renal failure—features in the history that may indicate muscle glycogenosis.

After a 30-min rest, blood is taken from the antecubital vein of the non-exercising arm and a small sphygmomanometer cuff placed around the other wrist is inflated to 200 mmHg. A second standard cuff around the upper arm to be tested is inflated to mean arterial pressure and the patient squeezes as powerfully as possible 120 times over 2 min. Immediately afterwards, the second cuff is inflated to 200 mmHg. Blood is drawn through a needle placed in the antecubital vein of the exercising arm 2 min after completing the exercise and the upper cuff is released. To complete the test, five further samples are drawn at 1-min intervals. The samples are transported rapidly to the laboratory for analysis of lactate and ammonia. Reduced or absent generation of lactate is characteristic of glycogenolytic and glycolytic defects that affect muscle; in contrast, plasma levels of ammonia (as well as inosine and hypoxanthine) increase greatly in patients with glycogenosis types III, V, and VIII. These abnormalities reflect the excessive degradation of purines that occurs in the exercising muscles of patients in whom there is a disturbance of ATP generation. Measurement of ammonia release as well as lactate production also adds discriminatory value to the exercise test, as it controls for low levels of lactate release that result merely from inadequate exercise during performance of the test. The test may also identify myoadenylate kinase deficiency: in such patients lactate production is normal, but the failure to utilize the purine cycle to conserve intracellular nucleotides and provide alternative substrates for energy production is shown

by the failure of venous ammonia concentrations to rise.

Pompe's disease due to acid maltase deficiency is a generalized disorder that predominantly affects skeletal and cardiac muscle. Carbohydrate metabolism is otherwise normal, and phosphorylysis of cytosolic glycogen in the liver is sufficient to maintain euglycaemia. The diagnosis of infantile disease may be suspected on the basis of cardiac and liver enlargement in an infant with respiratory distress and hypotonia. Macroglossia is frequent and the electrocardiogram shows left axis deviation, a short P–R interval and broad QRS complexes. In the juvenile- and adult-onset forms of acid maltase deficiency the disease resembles limb-girdle and other myopathies as well as polymyositis; some patients have been reported with myotonic features. The activity of skeletal muscle creatine kinase (**CK**) in this variant (non-CK MB fraction) is elevated in the serum and may be the first sign of intrinsic muscle disease, especially in adult patients complaining of non-specific fatiguability and weakness. Myopathic changes—occasionally with pseudomyotonic discharges—are observed on electromyography and the diagnosis is revealed by biopsy, which shows vacuolar myopathy: massive deposits of glycogen in and between myofibrils. Under the electron microscope, free and lysosomal a-glycogen particles are observed. Enzymatic deficiency of acid a-1,4-glucosidase is readily confirmed in cultured amniocytes and all tissues except erythrocytes.

Recently, the molecular basis for an unusually perplexing vacuolar cardiomyopathy associated with glycogen storage has been identified (Danon's disease). This X-linked disorder has been principally reported in male infants, boys, and men with proximal muscle weakness and prominent hypertrophic cardiomyopathy including cardiac conduction defects. Although the ultrastructural studies revealed membrane-bound inclusions of glycogen resembling Pompe's disease, acid maltase (a-1,4-glucosidase) activity was normal. In those cases with normal phosphorylase kinase activity (an enzyme that also maps to the X-chromosome), no cause for the severe cardioskeletal myopathy was apparent until it was shown to be associated with mutations in the lysosomal membrane protein, LAMP2. Families with probable Danon's disease have been reported with mild mental intellectual impairment and systemic manifestations. Clinical expression has been reported in obligate carrier female subjects in affected pedigrees showing inheritance patterns typical of an X-linked trait; the severity of the storage disease appears to be highly variable in female heterozygotes, consistent with patterns of random X-inactivation. Danon's disease can be diagnosed by molecular analysis of the *LAMP-2* gene that maps to human chromosome Xq24, and thus represents the first example of a disease due to a structural protein of the lysosomal membrane.

Definitive diagnosis of muscle glycogenoses depends on biopsy with histochemical, ultrastructural, and biochemical analyses. Biopsy should be carried out after liaison with the laboratory so that, if necessary, tissue can be stored frozen for further study and enzymatic analysis. Biopsy and electromyography may be needed to differentiate suspected glycogen storage diseases from other myopathies, including Duchenne's dystrophy, Kugelberg–Welander disease, dystrophia myotonica, and mitochondrial and secondary disorders of muscle such as polymyositis.

Individual glycogen storage diseases

The main features of these disorders are surveyed and summarized in Table 1. Brief accounts of selected conditions are set out below.

Classical type I glycogen storage disease (von Gierke's disease)

In this disease, glucose formation from glycogen and gluconeogenesis is defective and affected infants develop hypoglycaemia on fasting or as a result of intercurrent infection or other stress. The liver is enlarged at birth. It contains excess glycogen and shows gross infiltration with fat but cirrhosis and portal hypertension are rare. In contrast, growth retardation, often combined with obesity, is common. The kidneys are enlarged by glycogen deposition. Progressive focal glomerulosclerosis and proximal tubular failure with a secondary Fanconi syndrome may also occur. Stress and starvation provoke acidotic attacks with marked lactic acidaemia. Poor metabolic control causes: growth arrest; hyperuricaemia and gout; marked hypertriglyceridaemia and hypercholesterolaemia with raised very low-density lipoprotein (**VLDL**) and normal low-density lipoprotein (**LDL**) cholesterol concentrations in the plasma (skin and retinal xanthomas accompany these findings); and prolonged bleeding time related to an acquired von Willebrand-like defect affecting the platelet. Patients with defects of the glucose 6-phosphate translocase system (type 1B) are prone to bacterial infection: there is neutropenia, and neutrophil migration and chemotaxis are impaired. These patients may develop episodes of severe diarrhoea in association with granulomatous infiltration of the colonic mucosa. Partial deficiencies of the glucose 6-phosphatase system lead to variable clinical expression, and subtypes of type I glycogen storage disease have been convincingly demonstrated in patients presenting with glucagon-unresponsive hypoglycaemia with or without liver enlargement in adult life. Adult patients or children with uncontrolled disease develop hepatic adenomas; frank hepatocellular carcinomas occur.

Metabolic disturbance

The metabolic disturbance in classical type I glycogen storage serves as a paradigm for the hepatic glycogenoses.

Hypoglycaemia in von Gierke's disease is often asymptomatic and tolerance of it improves with increasing age. The primary defect leads to a profound reduction in the supply of glucose from glucose 6-phosphate in the liver leading to marked substrate-cycling. Lactate delivered from extrahepatic sources is converted to glucose 6-phosphate, which is metabolized by the pentose-phosphate shunt or transferred back into glycogen. The pentose pathway supplies precursors for purine synthesis and reducing equivalents. Residual production of glucose probably occurs by lysosomal hydrolysis of glycogen and recycling through the glycogen synthase-debranching enzyme pathway, but metabolic adaptation of the brain, which can use lactate as an alternative substrate, is very important.

Failure to dephosphorylate glucose 6-phosphate stimulates substrate cycling and increases the activity of the pentose-phosphate pathway, with enhanced production of reduced NADP (**NADPH**, reduced form of nicotinamide-adenine dinucleotide phosphate), ribose 5-phosphate, and purines—this latter ultimately leads to the overproduction of uric acid through the action of xanthine oxidase. Increased delivery of fructose 6-phosphate from the pentose-phosphate pathway leads to the excess formation of lactate as a result of phosphohexosisomerase activity. Enhanced cycling of UDPG and the glycogen synthase reaction promotes glycogen accumulation. However, small quantities of free glucose can be liberated by the a-1,6-glucosidase activity of the secondary action of debranching enzyme but the co-ordinated action of glucosyltransferase and phosphorylase releases additional glucose 1-phosphate residues for recycling. An additional (fractional) degradation of the intracellular glycogen store is probably contributed by the a-1,4-glucosidase activity of lysosomal acid maltase.

Hypertriglyceridaemia is induced by the increased provision of reduced nicotinamide-adenine dinucleotide (**NADH**) and NADPH, glycerol, and acetyl-coenzyme A (**acetyl-CoA**) because of enhanced flux through glycolysis and underutilization of gluconeogenic precursors. Malonyl-coenzyme A, derived from acetyl-CoA, inhibits the carnitine acyltransferase system and blocks the oxidation of fatty acids; thus marked ketosis does not usually develop. Lactic acidaemia results from stimulation of glycolysis at the level of phosphofructokinase by high concentrations of glucose 6-phosphate (and hence fructose 6-phosphate); lactate cannot be recycled in the liver to form new glucose and lactic acidosis results. Lactate competes with urate for excretory pathways in the kidney and thus contributes to the hyperuricaemia. Uric acid is also overproduced in the liver: it arises from the degradation of purine nucleotides by AMP-deaminase and the co-ordinated action of xanthine oxidase on inosine phosphate and hypoxanthine. The deaminase is activated when the concentration of free phosphate falls as a result of sequestration in sugar phosphate esters.

Treatment

The main objective is to maintain euglycaemia: most of the other metabolic abnormalities are thereby corrected and the prognosis improves.

In infants, normoglycaemia is maintained throughout 24 h by intravenous alimentation at 0.25 to 0.5 g/kg per hour and, later by continuous nasogastric administration at night together with glucose supplements at intervals of 1 to 2 h during the day. These intensive regimens correct acidosis, hyperuricaemia, and hyperlipidaemia; they also promote normal development and allow catch-up growth to occur in stunted infants and children. After growth in later childhood and in adult patients, metabolic control can be maintained by the use of raw cornstarch, which serves as a source of glucose that is slowly released by hydrolysis: 1 to 2 g/kg is given orally every 4 to 6 h as a suspension in water.

In type Ib glycogen storage disease it is vital to avoid intercurrent infection, and prophylactic antimicrobial drugs may therefore be necessary. In several instances, infusions of granulocyte-colony-stimulating factor has been strikingly effective in reducing the rate of infection and controlling granulomatous colitis. Patients with type Ia disease may also require treatment for their bleeding tendency. The bleeding diathesis is associated with a qualitative defect of platelet function, prolonged bleeding time, and reduced factor VIIIc and von Willebrand factor activities. These abnormalities and the haemorrhagic tendency respond to the administration of 1-deamino-8-D-arginine vasopressin (**DDAVP**) at 0.3 µg/kg infused in 50 ml of saline over 30 min intravenously. Correction of the bleeding disorder lasts for several hours and is useful for the treatment of bleeding after trauma or surgery.

Failure of metabolic control in type I glycogen storage disease appears to be associated with tissue complications: hepatic adenomas or malignant transformation, renal disease due to hyperfiltration, focal glomerulosclerosis, and postinfective scarring. Lately, an inflammatory disorder of the colon, resembling granulomatous colitis, has been recognized in type Ib disease. Type Ic disease, characterized by the increased latency of hepatic microsomal inorganic pyrophosphatase activity has

now been reported. Phagocytic defects are not prominent in this disease subtype. Defective function of the microsomal glucose transporter has been reported and is designated type Id glycogen storage disease. Long-term, follow-up care with monitoring of biochemical parameters of kidney function and periodic ultrasonic examination of the liver is necessary. Continuing failure of growth, enlarging hepatic adenomas or progressive renal failure raise the question of organ transplantation. Transportal hepatocyte transplantation has been successfully achieved in this disease with correction of hypoglycaemia and lactic acidosis. Several successful renal, as well as hepatic, allografts have been carried out in patients with this condition using DDVAP infusions to control haemorrhagic manifestations. However, as regression of most complications, including hepatic adenomas, can be achieved by strict dietary measures, transplantation should be reserved for patients in whom nutritional treatment has failed. Survival into adult life (and parenthood) can be now expected.

Type II glycogen storage disease

Pompe's disease caused by acid maltase deficiency is usually a rapidly progressive disorder with effects on the heart, skeletal muscle, and nervous system. Affected children usually die within the first year or two of life, and until recently no measures other than supportive therapy and ventilatory assistance have been beneficial. Late-onset disease, usually without cardiomyopathy, occurs in juvenile and adult patients in whom it typically presents with skeletal myopathy affecting the proximal muscles. Ultimately, respiratory failure results from paralysis of the muscles of ventilation, including the diaphragm; voluntary muscles of deglutition may also be paralysed. Occasionally the disease resembles polymyositis or limb-girdle muscular dystrophy. Given that enzyme-replacement therapy is theoretically possible for lysosomal storage diseases, administration of purified acid a-1,4-glucosidase (acid maltase) has been attempted. Early trials of recombinant human acid maltase harbouring mannose 6-phosphate residues, to mediate targeting to cell-surface receptors for lysosomal uptake by skeletal myocytes, have been reported in infants with classical Pompe's disease. Two preparations (from the milk of lactating rabbits and from genetically engineered rodent cells) have been studied. Limited success was obtained in both trials, with improved muscle strength and transient mobility as well as delayed progression of myopathy. The long-term outcome is rendered uncertain by the development of neutralizing antibodies in many recipients and by the ability to manufacture sufficient enzyme. Bone marrow transplantation does not appear to be beneficial. In juvenile and adult acid maltase deficiency, muscle wasting may be arrested with improved or maintained function by institution of a high-protein, restricted-carbohydrate diet. Enzyme-replacement trials using recombinant human acid maltase have yet to be conducted in late-onset type II disease, although this treatment is likely to be more successful than in infantile disease where enzyme antigen is usually completely absent.

Type III glycogen storage disease

The clinical manifestations of Forbes–Cori's disease resemble those of type I glycogenosis, especially in infants, who present with hypoglycaemia, short stature, and hepatomegaly. Mild progressive myopathy, occasionally with signs of hypertrophic cardiomyopathy, may occur. The disorder is characterized by marked clinical variability. Generally the signs of liver disease regress during maturation and myopathy also improves with nutritional therapy as outlined for von Gierke's disease. Protein supplements, which may provide additional sources of energy, appear to benefit the muscle disorder.

Type IV glycogen storage disease

This disorder is one of the more severe glycogenoses because the deficiency of branching enzyme in Anderson's disease gives rise to the deposition of an abnormal glycogen in many tissues. Severe inflammation occurs in the liver, resulting in early cirrhosis, with splenomegaly due to portal hypertension. This fatal disorder is characterized by failure to thrive, hepatosplenomegaly, jaundice, and hypotonia. The myopathy is often prominent with a lordotic posture and waddling gait due to limb-girdle weakness. Cardiomyopathy leading to cardiac failure develops in severely affected infants and children. Diagnosis is based on the appearances of the liver biopsy and abnormal glycogen structure shown by histochemical and biochemical analysis. Deficiency of branching enzyme is demonstrable in leucocytes. No definitive therapy is available, but a few patients have survived hepatic transplantation without the development of neuromuscular or cardiac complications up to 7 years after the procedure. Generally the prognosis is poor: without transplantation most patients die before the age of 4 years with liver failure, variceal bleeding, and intercurrent infection. Prenatal diagnosis of branching enzyme can be conducted by enzymatic analysis of amniotic cells or chorionic villi; DNA analysis of the human branching enzyme on chromosome 3p12 may also be possible for at-risk families.

Type V glycogen storage disease (McArdle's disease)

This disorder is characterized by the late onset of muscle fatigue and cramps during adolescence or early adult life. Hepatomegaly is absent. Strenuous exercise may induce episodic myoglobinuria and biochemical evidence of rhabdomyolysis. A characteristic feature is the occurrence of the 'second wind' phenomenon: progressive weakness and fatigue develop during the first 10 to 15 min of exercise, with a rapid recovery that is complete on resting; after this adaptation phase, patients are often able to continue exercise without difficulty. The mechanisms involved in this adaptive phenomenon are not clear but include increased cardiac output, blood flow to the muscles, and metabolic changes, probably including different patterns of fibre recruitment and the use of oxidative pathways. Occasionally, acute myoglobinuric renal failure may result. Muscle biopsy may show abnormal muscle fibres with necrosis, atrophy, and hypertrophied fibres alongside. The course of this disease is benign; ingestion of glucose or pre-exercise administration of glucagon may partially ameliorate the symptoms but avoidance of strenuous exercise is advisable. The muscle phosphorylase gene maps to chromosome 11q13 and sequence analysis has identified common mutations in this glycogenosis; one mutation, involving formation of a stop codon within exon 1 at position 49 (arginine) is sufficiently common to be of diagnostic value.

Type VI glycogen storage disease and phosphorylase b kinase deficiency

These disorders cause hepatomegaly, intermittent hypoglycaemia, and markedly increased liver glycogen content. Although many polypeptides constitute the intact phosphorylase b kinase complex (encoded on autosomes and the X-chromosome), glycogen mobilization is usually only partially defective. X-linked phosphorylase b kinase deficiency is the most frequent variant and is associated with growth retardation, mild ketosis, and hyperlipidaemia in childhood. The symptoms improve with age and the disorder is compatible with a normal life expectancy. Cirrhosis of the liver is very rare, and the incompleteness of the defect is shown by almost normal hyperglycaemic responses to glucagon administration. Rare autosomal variants of phosphorylase kinase deficiency affecting liver and muscle or restricted to skeletal or cardiac muscle have been documented. These subtypes are associated with hypotonia or cardiac failure, respectively. Treatment of liver phosphorylase or kinase deficiency with frequent feeding to avoid hypoglycaemia may be needed, but intensive nutritional therapy is rarely indicated since the general prognosis is good. No specific treatment for the isolated cardiac form of kinase deficiency is known but cardiac transplantation could be considered if the diagnosis can be established.

Type VII glycogen storage disease (Tarui's disease)

This disorder, which is most frequent in patients of Japanese or Russian Ashkenazi ancestry, closely resembles type V muscle glycogenosis but severe symptoms usually come to light in childhood. There may be hyperuricaemia which is aggravated by exercise. Deficiency of red cell phosphofructokinase leads to chronic haemolysis; there is mild jaundice and a strong association with pigment-type gallstones. Decreased 2,3-diphosphoglycerate synthesis resulting from the metabolic block has been noted and probably contributes to exercise-induced symptoms by reducing oxygen delivery. Phosphofructokinase I catalyses an irreversible step in glycolysis and is an important regulatory enzyme, especially in muscle. Deficiency of phosphofructokinase I renders the pathway insensitive to positive allosteric regulation by the key effector molecules, fructose 2,6-diphosphate and AMP; hence myophosphorylase activity remains depressed. For this reason, Tarui's disease resembles a severe form of McCardle's disease. No specific therapy for this disorder is known—in contrast to McArdle's disease, neither glucagon nor glucose infusions improve exercise tolerance. Indeed, carbohydrate-rich meals aggravate the symptoms, presumably by diminishing the concentration of non-esterified fatty acids in the plasma, which serve as the alternative source of muscle energy production. Several very rare variants of phosphofructokinase deficiency are known: a severe infantile form with progressive and fatal myopathy and a late-onset form that causes fixed muscle weakness in middle-aged subjects are both clearly recognized. Approximately 15 mutations have been identified in the human muscle phosphofructokinase gene in patients with Tarui's disease; the three subunits encoding this isozyme originate from a locus on chromosome 12q13.3.

Glycogen synthase deficiency

Although glycogen synthase deficiency is very rare, it causes deficiency of glycogen formation in the liver. Most cases have been reported in infants and young children. It is, therefore, a disorder of storage rather than a true glycogenosis. The condition causes severe interprandial hypoglycaemia and marked ketosis; a notable feature is the rapid development of hyperglycaemia and lactic acidaemia on feeding. The disorder resembles fructose 1,6-bisphosphatase deficiency, but mutations in the liver glycogen synthetase gene II on chromosome 12p12.2 have been identified. Biopsy examination of the liver shows fatty infiltration and depletion of glycogen: uridine diphosphate-pyrophosphorylase, phosphorylase, glucose 6-phosphatase activities are normal but glycogen synthase is absent. Glucose polymers and uncooked cornstarch are effective therapy.

Further reading

Amalfitano A, *et al*. (2001). Recombinant human acid alpha-glucosidase enzyme therapy for infantile glycogen storage disease type II: results of a phase I/II clinical trial. *Genetic Medicine* **3**, 132–8.

Ambruso DR, *et al*. (1985). Infectious and bleeding complications in patients with glycogenosis Ib. *American Journal of Diseases of Children* **139**, 691–7.

Bao Y, *et al*. (1996). Hepatic and neuromuscular forms of glycogen storage disease type IV caused by mutations in the same glycogen-branching enzyme. *Journal of Clinical Investigation* **97**, 941–8.

Bianchi L (1993). Glycogen storage disease I and hepatocellular tumours. *European Journal of Pediatrics* **152**(Suppl 1), 563–70.

Braakhekke JP, *et al*. (1986). The second wind phenomenon in McCardle's disease. *Brain* **109**, 1087–101.

Burchell A (1992). The molecular basis of the type I glycogen storage diseases. *BioEssays*, **14,** 395–400.

Cabello A, *et al*. (1981). Glycogen storage disease in skeletal muscle. Morphological, ultrastructural and biochemical aspects in 10 cases. *Acta Neuropathologica (Basel)*, **Suppl VII**, 297–300.

Chen Y-T (2001). Glycogen storage disease. In: Scriver CR, *et al.*, eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 1521–51. McGraw-Hill, New York.

Chen Y-T, Cornblath M, Sidbury JB (1984). Cornstarch therapy in type I glycogen-storage disease. *New England Journal of Medicine*, **310**, 171–5.

Chen Y-T, *et al.* (1988). Renal disease in type I glycogen storage disease. *New England Journal of Medicine*, **318,** 7–11.

Chou JY and Mansfield BC (1999). Molecular genetics of type 1 glycogen storage diseases. *Trends in Endocrinology and Metabolism* **10**, 104–13.

Danon MJ, *et al.* (1981). Lysosomal glycogen storage disease with normal acid maltase. *Neurology* **31**, 51–7.

de Barsy T, Hers H-G (1990). Normal metabolism and disorders of carbohydrate metabolism. *Baillière's Clinical Endocrinology and Metabolism* **4**, 499–522.

Engel AG (1970). Acid maltase deficiency in adults: studies in four cases of a syndrome which may mimic muscular dystrophy or other myopathies *Brain* **93**, 599–616.

Faivre L, *et al.* (1999). Long-term outcome of liver transplantation in patients with glycogen storage disease type 1A. *Journal of Inherited Metabolic Disease* **22**, 723–32.

Fernandes J, *et al.* (1988). Glycogen storage disease: recommendations for treatment. *European Journal of Paediatrics* **147**, 226–8.

Furakawa N, *et al.* (1990). Type I glycogen storage disease with vasoconstrictive pulmonary hypertension. *Journal of Inherited Metabolic Disease* **13**, 102–7.

Gitzelmann R, *et al.* (1996). Liver glycogen synthase deficiency: a rarely diagnosed entity. *European Journal of Paediatrics* **155**, 561–7.

Haller RG and Lewis SF (1991). Glucose-induced exertional fatigue in muscle phosphofructokinase deficiency. *New England Journal of Medicine* **324**, 364–9.

Hendrickx J, *et al.* (1995). Mutations in the phosphorylase kinase gene PHKA2 are responsible for X-linked liver glycogen storage disease. *Human Molecular Genetics* **4**, 77–83.

Janecke AR, *et al.* (1999). Molecular diagnosis of type Ic glycogen storage disease *Human Genetics* **105**, 515–17.

Kroos MA, *et al.* (1995). Glycogen storage disease type II: frequency of three common mutant alleles and their associated clinical phenotypes studied in 121 patients. *Journal of Medical Genetics* **32**, 836–7.

Lee PJ, Dixon MA, Leonard JV (1996). Uncooked cornstarch—efficacy in type I glycogenosis. *Archives of Diseases of Children* **74**, 546–7.

Marti GE, *et al.* (1986). DDAVP infusion in five patients with type Ia glycogen storage disease and associated correction of prolonged bleeding times. *Blood* **68**, 180–4.

Muraca M, *et al* (2002). Hepatocyte transplantation as a treatment for glycogen storage disease type 1a. *Lancet* **359**, 317–18.

Nishino I, *et al* (2000). Primary LAMP-2 deficiency causes X-linked vacuolar cardiomyopathy and myopathy (Danon disease). *Nature* **406**, 906–10.

Pears JS, *et al.* (1992). Glycogen storage disease diagnosed in adults. *Quarterly Journal of Medicine* **82**, 207–2.

Raben N, Sherman JB (1995). Mutations in muscle phosphofructokinase gene. *Human Mutation* **6**, 1–6.

Roe TF, *et al.* (1992). Treatment of chronic inflammatory bowel disease in glycogen storage disease type Ib with colony-stimulating factors. *New England Journal of Medicine* **326**, 1666–9.

Shaiu W-L, *et al.* (2000). Genotype–phenotype correlation in two frequent mutations and mutation update in type III glycogen storage disease. *Molecular Genetics in Metabolism* **69**, 16–23.

Shin YS (1990). Diagnosis of glycogen storage disease. *Journal Inherited Metabolic Disease* **13**, 419–34.

Slonim AE, Goans PJ (1985). Myopathy in McArdle's syndrome: improvement with a high-protein diet. *New England Journal of Medicine* **312**, 355–9.

Slonim AE, *et al.* (1983). Improvement of muscle function in acid maltase deficiency by high-protein therapy. *Neurology* **33**, 34–8.

Talente, *et al* (1994). Glycogen storage disease in adults. *Annals of Internal Medicine* **120**, 218–26.

Van den Hout JM, *et al.* (2001). Enzyme therapy for Pompe disease with recombinant human a-glucosidase from rabbit milk. *Journal of Inherited Metabolic Disease* **24**, 266–74.

Vogerd M, *et al* (1998). Mutation analysis in myophosphorylase deficiency (McArdle's disease). *Annals of Neurology* **43**, 326–31.

Willems PJ, *et al.* (1990). The natural history of liver glycogenosis due to phosphorylase kinase deficiency: a longitudinal study of 41 patients. *European Journal of Pediatrics* **149**, 268–71.

Williams JC (1986). Nutritional goals in glycogen storage disease. *New England Journal of Medicine* **314**, 709–10.

Wolfsdorf JI, Rudlin CR, Cirgler JF (1990). Physical growth and development of children with type I glycogen-storage disease: comparison of the effects of long-term use of dextrose and uncooked cornstarch. *American Journal of Clinical Nutrition* **52**, 1051–7.

11.3.2 Inborn errors of fructose metabolism

T. M. Cox

There are three inborn errors of fructose metabolism recognized: (1) essential or benign fructosuria due to fructokinase deficiency; (2) fructose 1,6-diphosphatase deficiency; and (3) hereditary fructose intolerance (fructosaemia). There are discussed in relation to the overall metabolism of fructose, a major nutrient.

Metabolism of fructose

Phosphorylated forms of fructose are critical intermediates in the glycolytic and gluconeogenic metabolic pathways in all cells. Fructose is also an important component of the diet: it occurs as a free monosaccharide in fruit, nuts, honey, and some vegetables. Free fructose is released from sucrose in the gut lumen by sucrase–isomaltase in the brush-border membrane of the mucosal epithelium. Finally, the sugar alcohol, sorbitol (a constituent of medicines and tablets, as well as some foods for diabetics), is converted quantitatively to fructose in the liver and intestine. Most individuals in developed countries ingest 50 to 150 g fructose equivalents daily in the diet.

The pathways of fructose metabolism are summarized in Fig. 1. Fructose is absorbed rapidly by a carrier mechanism that facilitates transport across the intestinal epithelium; this process is mediated by the glucose transporter isoforms, GLUT5 and GLUT2, the latter probably contributing to efflux across the basolateral membrane of the enterocyte.

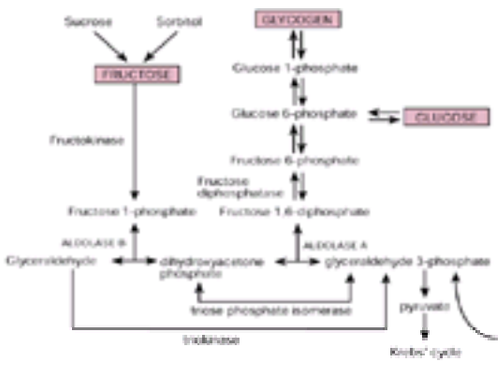

Fig. 1 Fructose metabolism.

It is then conveyed via the portal bloodstream to the liver, where it is assimilated. The jejunal mucosa and proximal tubule of the kidney are subsidiary sites of fructose metabolism. Assimilation of fructose depends on the concerted activities of the enzymes ketohexokinase (fructokinase), aldolase B, and triokinase, which are expressed specifically in these tissues. Uptake of fructose occurs independently of insulin and its incorporation into intermediary metabolism bypasses the regulation of glycolysis at the level of phosphofructokinase-1. For these reasons, solutions of fructose or sorbitol were advocated and, in the past, extensively used for parenteral nutrition. However, the occurrence of lactic acidosis, hyperuricaemia and other serious consequences have led to their withdrawal from hyperalimentation regimens in most, if not all, countries.

Fructokinase rapidly phosphorylates fructose at the 1-carbon position. This enzyme has a high affinity for its substrates and the intestinal mucosa and liver rapidly convert fructose to fructose 1-phosphate: in other tissues, the capacity of hexokinase to phosphorylate fructose at the 6-carbon position is limited. Similarly, the fate of fructose 1-phosphate in the fructose-metabolizing tissues is dependent on a specific isozyme of aldolase, aldolase B. This has greater activity towards fructose 1-phosphate than does its ubiquitous counterpart, aldolase A, the natural substrate of which is fructose 1,6-diphosphate. Cleavage of fructose 1-phosphate generates glyceraldehyde and dihydroxyacetone phosphate. These trioses enter the intermediary pools of carbohydrate metabolism, and, as a result of triokinase activity, glyceraldehyde is phosphorylated so that the two triose phosphates may be condensed by aldolase A to form the glycolytic and gluconeogenic intermediate, fructose 1,6-diphosphate.

Gluconeogenesis from triose phosphates, lactate, glycerol, amino acids, and Krebs cycle intermediates such as oxaloacetate, requires reversal of the committed reactions of glycolysis. It is the enzyme fructose 1,6-diphosphatase that releases the glucose precursor fructose 6-phosphate from fructose 1,6-diphosphate. Thus, when the remaining reactions of glycolysis are reversed, exogenous fructose provides a source of glucose or glycogen. Fructose 1,6-diphosphatase is active in the liver, kidney, and intestine; it is a key enzyme of gluconeogenesis.

Fructose malabsorption

Incomplete absorption of fructose with abdominal symptoms and diarrhoea reminiscent of intestinal disaccharidase deficiency is well recognized by gastroenterologists. The symptoms occur in adults and children after ingestion of fructose- or sorbitol-rich foods and drinks such as apple juice, and usually recede when these sugars are excluded from the diet. Many such individuals, as well as a high proportion of healthy control subjects, have suggestive evidence of fructose malabsorption based on breath-hydrogen tests. Unfortunately, the molecular basis of this syndrome and of the wide variation of tolerance to dietary fructose and its congeners is not known. Moreover, molecular analysis of the human *GLUT5* gene in several patients complaining of fructose-related intestinal symptoms, has hitherto failed to implicate this candidate sugar transporter. Preliminary studies suggest that lower intestines and colons of many patients who experience abdominal flatulence and diarrhoea after ingesting fructose-containing foods contain a bacterial population showing enhanced uptake and anaerobic metabolism of fructose. No conclusive evidence has yet been provided to support these observations and more fructose transport studies are needed on the mucosal epithelium of patients who complain of symptoms that indicate an intestinal malabsorption of this sugar.

Essential (benign) fructosuria (Mendelian inheritance in man (MIM) 229800)

This is a rare disorder (estimated frequency 1 in 130 000) of little clinical consequence. The abnormality is transmitted as an autosomal recessive condition and manifests itself by the presence of a reducing sugar in the blood and urine, especially after meals rich in fructose. The abnormality results from the deficiency of

fructokinase activity in the liver and intestine, significantly reducing the capacity to assimilate this sugar. Mutations in the human ketohexokinase gene on chromosome 2p23.3–p23.2 have been identified in patients with essential fructosuria, thus confirming the suspected molecular defect in this condition. Fructose metabolism occurs slowly in essential fructosuria as a result of conversion to fructose 6-phosphate by hexokinase in adipose tissue and muscle, but, while plasma concentrations remain high postprandially, large amounts of fructose appear in the urine. Essential fructosuria may be confused with diabetes mellitus if the nature of the mellituria is not defined with the use of glucose oxidase strips in preference to the older chemical methods for urinalysis, such confusion is now unlikely. No treatment beyond recognition and explanation appears to be necessary.

Fructose diphosphatase deficiency (MIM 229700)

Description

This very rare, recessively inherited disorder presents with hypoglycaemia, ketosis, and lactic acidosis in early infancy. Fewer than 30 cases have been reported since its original description in 1970. Severe, sometimes fatal, acidosis is associated with infection and starvation and most cases have presented within the first few days of life or in the neonatal period. Onset during the first year of life is the rule.

In newborn infants, the severe metabolic disturbance shows itself by acidotic hyperventilation, which may be accompanied by irritability, disturbed consciousness, seizures, or coma. The unusual combination of ketonaemia, lactic acidaemia, and hypoglycaemia is induced by fasting, the administration of fructose, sorbitol, and glycerol, and by ingestion of a diet rich in fat. Episodes in the neonatal period respond well to infusions of glucose and bicarbonate but, after an interval, further attacks occur, often provoked by intercurrent infection. Lethargy accompanied by hyperventilation is followed abruptly by prostration, coma, and seizures. Investigations reveal hypoglycaemia, ketosis, and profound lactic acidosis; there is hyperuricaemia, amino aciduria, and ketonuria. If the infant survives, hepatomegaly due to fatty infiltration may be detected but overt clinical disturbances of hepatic or renal tubular function are not seen. The untreated disease is associated with growth retardation.

The first infant to be affected by fructose diphosphatase deficiency in a given family may succumb before the diagnosis is established and in any case fares worse than siblings for whom the appropriate diet and prompt control of the condition are instituted. The response to treatment is favourable, however, and fructose diphosphatase deficiency is ultimately compatible with a benign course and with normal growth and development.

Metabolic defect

Deficiency of fructose 1,6-diphosphatase causes failure of gluconeogenesis in the liver—although the abnormality may be detected in intestinal mucosa, kidney, and in cultured mononuclear cells from peripheral blood. The muscle isozyme of fructose 1,6-diphosphatase is not affected.

Between meals, blood glucose is maintained by glucogenolysis and hence the onset of disturbed metabolism in fructose diphosphatase deficiency depends on the availability of hepatic glycogen. Since febrile illnesses accelerate the consumption of liver glycogen, the accompanying anorexia with or without vomiting may deplete glycogen stores critically. Acidosis results from the accumulation of gluconeogenic precursors including lactate, pyruvate, and alanine as well as ketone bodies, which cannot be utilized. Hypoglycaemia, which is unresponsive to glucagon and associated with exhaustion of glycogen stores, occurs: this does not respond to normal gluconeogenic substrates (for example, glycerol, amino acid solutions, dihydroxyacetone, sorbitol, or fructose), indeed administration of these aggravates the metabolic disturbance.

The pathogenesis of hypoglycaemia and accompanying disturbances in fructose diphosphatase deficiency is complex and not completely explained by exhaustion of hepatic glycogen stores. Well-fed patients have a normal response to glucagon but are intolerant of high-fat diets, as well as of fructose, sorbitol, alanine, glycerol, and dihydroxyacetone administration. Challenge with these nutrients induces hypoglycaemia, hyperuricaemia, and hypophosphataemia, accompanied by an exaggerated rise in blood lactate levels. The hypoglycaemia is then unresponsive to glucagon, indicating a secondary inhibition of phosphorylase activity in the liver, which results from the build-up of phosphorylated sugar intermediates that cannot be further metabolized in the context of reduced intracellular free inorganic phosphate. Adenosine deaminase is activated primarily because of reduced phosphate concentrations, so that purine nucleotides are broken down to uric acid. Failure to utilize glucogenic amino acids and metabolites such as dihydroxyacetone and glycerol appears to stimulate triglyceride formation in the liver, which induces steatosis. Unlike hereditary fructose intolerance (see below), high concentrations of hepatic fructose 1-phosphate do not occur, and profound disturbances of blood coagulation or hepatic or renal tubule function with progressive structural damage are absent in fructose diphosphatase deficiency. Similarly, aversion to foods that aggravate the disorder does not develop in affected infants and children; this may be explained by the absence of pain and abdominal symptoms in the condition.

Diagnosis

The importance of establishing the diagnosis of fructose diphosphatase deficiency cannot be overemphasized: proper dietary control and protocols for the institution of appropriate therapy depend upon recognizing the complex disturbance that underlies this disease.

Fructose diphosphatase deficiency should be considered in otherwise normal infants who develop unexplained severe acidosis or hypoglycaemia associated with episodes of infection. The combination of ketosis and lactic acidosis with hypoglycaemia is highly suggestive of a disorder affecting the gluconeogenic pathway, including deficiency of glucose 6-phosphatase, pyruvate carboxylase, pyruvate dehydrogenase, and phosphoenolpyruvate carboxykinase. The absence of abdominal distress, haemolysis, jaundice, coagulopathy, and disturbances of the proximal renal tubule differentiate the condition from hereditary fructose intolerance, tyrosinosis, and Wilson's disease. Confusion may arise with disorders associated with secondary defects in gluconeogenesis, especially the Reye-like syndrome caused by deficiencies of long-, medium- and short-chain acyl coenzyme A dehydrogenase activities, as well as defects of carnitine metabolism. Organic acidaemias are also readily distinguished by biochemical screening methods.

Provocative tests using food deprivation and the administration of infusions of fructose, sorbitol, or glycerol should be avoided in the acutely ill infant or child with suspected deficiency of fructose 1,6-diphosphatase (or fructose intolerance). The definitive diagnosis depends on the demonstration of selectively decreased fructose diphosphatase activity in tissue samples. Most frequently, the enzymatic defect will be identified by biochemical assay of a freshly obtained liver biopsy specimen, which allows other metabolic disorders and gluconeogenic defects to be confidently excluded. The defect may also be demonstrated in biopsy samples of jejunal mucosa and in cultured monocyte-derived macrophages obtained from peripheral blood. However, the presence of fructose 1,6-diphosphatase in these tissues is metabolically inconsequential and, although useful for confirmation of the diagnosis where it is strongly suspected, in practice decisive identification of this disorder normally depends on a systematic biochemical analysis of liver tissue in an experienced laboratory. The human fructose 1,6-diphosphatase (**FBP1**) gene maps to chromosome 9q22.2–q22.3 and inactivating mutations have been identified in the disease. Unlike fructose intolerance however, these mutations tend to be private and thus individually of less diagnostic significance for routine laboratory use in this disorder since mutational heterogeneity appears to be the rule. However, a minor exception to this occurs in the Japanese population, where one mutation (960–961 ins G) appears to account for almost one-half of mutant *FBPI* alleles.

Treatment

Dietary control and avoidance of starvation with rapid control of febrile illnesses is the mainstay of treatment. Minor infections and injuries require prompt attention, and intravenous glucose therapy should be instituted early in acute episodes to avoid hypoglycaemia and acidosis. Fasting should be avoided as far as possible, while night-time feeding may be needed in infants during recovery from injuries or infections and after strenuous exercise in older children. The habit of taking meals at regular 4-hourly intervals is best inculcated when the patient is young. The diet should exclude excess fat; sorbitol, sucrose, and fructose must be strictly avoided. Breast milk is rich in lactose, which is readily assimilated, but difficulties arise on transfer to artificial feeds during weaning. In addition, medications and syrups containing fructose, sucrose, or sorbitol present a special danger to patients with deficiency of fructose diphosphatase activity. A diet excluding these sugars but containing 56 per cent calories as carbohydrate with 32 per cent calories as fat and 12 per cent as protein has produced normal growth and development. Acute episodes of acidosis or hypoglycaemia are controlled rapidly by intravenous administration of glucose with or without bicarbonate as required.

Hereditary fructose intolerance (fructosaemia) (MIM 229600)

This disorder, first recognized in 1956, is the most common inherited defect of fructose metabolism with an estimated frequency of 1 in 20 000 births. The condition is transmitted as an autosomal recessive abnormality and, although it manifests itself first in early infancy, the effects of clinical disease may not be recognized until late childhood or adult life. Provided the diagnosis is made before visceral damage occurs, hereditary fructose intolerance responds completely to an exclusion diet.

The cardinal features of the illness are vomiting, diarrhoea, abdominal pain, and hypoglycaemia, which are induced by the consumption of foods, drinks, or medicines that contain fructose seizures, or the related sugars, sucrose or sorbitol. There is a generalized metabolic disturbance with lactic acidosis, hyperuricaemia, and hyperphosphataemia. Hypoglycaemia causes trembling, irritability, and cognitive impairment. Attacks are associated with pallor, sweating, and, when severe, loss of consciousness—sometimes accompanied by generalized seizures. These episodes usually occur within 30 min of feeds that contain large quantities of fructose or sucrose. Continued ingestion of noxious sugars is associated with renal tubular disease, liver damage with jaundice, and defective blood coagulation. There is failure to thrive and growth retardation. Persistent exposure to fructose in infants leads to structural liver injury with cirrhosis, amino aciduria, coagulopathy, and coma leading to death. The infant is first exposed to the offending sugars at weaning or upon transfer from breast milk to artificial feeds: survival is dependent on recognition of the effects of fruit and sugar by the mother or, especially in older infants, by vomiting or forcible rejection of food.

Infants who survive the stormy period of weaning, develop a strong aversion to sweet-tasting foods, vegetables, and fruits. This usually affords protection against the worst effects of fructose and sucrose, but abdominal symptoms with bouts of tremulousness, irritability, and altered consciousness due to hypoglycaemia usually continue. It has become clear that many cases escape diagnosis in infancy and childhood, but that the risk of illness, related to dietary indiscretion, remains throughout life. Characteristically, children and adults with hereditary fructose intolerance show a striking reduction in, or absence of dental caries.

Recently, a syndrome of chronic sugar intoxication has been recognized in older children and adolescents with hereditary fructose intolerance. General lack of vigour and developmental retardation are prominent features. Hypoglycaemia, though obvious after heavy fructose loading, may be insignificant after chronic low-level exposure in older children. Similarly, tests of hepatic and renal function may be only mildly abnormal. Persistent ingestion of fructose and sucrose is toxic to the kidney and liver, so that renal tubular acidosis (occasionally with calculi) as well as hepatosplenomegaly are frequently detectable in the younger patients. Severe growth retardation may be accompanied by rachitic bone disease that complicates the Fanconi-like syndrome of proximal renal tubular disturbance. Growth retardation responds to dietary treatment and is usually accompanied by regression of the other disease manifestations.

Provided that organ failure and serious tissue injury do not supervene, patients with hereditary fructose intolerance recover rapidly when the offending sugars are withdrawn. Children who survive by acquiring the protective eating-behaviour pattern avoid foods that they associate with abdominal symptoms. The aversion extends to most sweet-tasting items of food and drink as well as fruits and vegetables—it remains lifelong and consumption of fructose is usually reduced to less than 5 g daily. It has been shown that normal growth and development can be secured in children if less than 40 mg/kg fructose equivalents are ingested daily.

Metabolic defect

Hereditary fructose intolerance is caused by a deficiency of aldolase B in the liver, small intestine, and proximal renal tubule. These tissues suffer injury as a result of persistent exposure to fructose in patients affected by the disorder. In the absence of the fructose 1-phosphate splitting activity of aldolase B, the intracellular pool of inorganic phosphate is depleted. Studies *in vivo* by ^{31}P magnetic resonance spectroscopy show that 80 per cent of hepatic free phosphate is sequestrated as sugar phosphates after the infusion of small quantities of fructose (250 mg/kg body weight). The secondary metabolic disturbances are initiated by the accumulation of fructose 1-phosphate in a milieu where free inorganic phosphate is reduced: there is competitive inhibition of aldolase A and inhibition of phosphorylase activity so that glycogenolysis and gluconeogenesis are impaired. Thus challenge with fructose leads to hypophosphataemia and hypoglycaemia that is refractory to glucagon or the infusion of gluconeogenic metabolites such as glycerol or dihydroxyacetone. During challenge with fructose, high concentrations of fructose 1-phosphate cause feedback inhibition of fructokinase, thereby limiting the incorporation of fructose in the liver. As a result, fructosaemia occurs and when the concentration exceeds about 2 mmol/l in peripheral blood, fructosuria becomes apparent. Although the assimilation of fructose by the specialized pathway is blocked, only a small fraction of the fructose load is recovered in the urine. Studies show that 80 to 90 per cent of the fructose is taken up under these circumstances by adipose tissue and muscle, where it can be alternatively metabolized by phosphorylation to fructose 6-phosphate.

Electrolytic disturbances occur during challenge with fructose. Hypokalaemia results from acute renal impairment with defective urinary acidification. There is a defect of proximal tubule function with bicarbonate wasting and acidosis. Occasionally, acute flaccid weakness due to hypokalaemia accompanies the other effects of fructose exposure. In patients with hereditary fructose intolerance, the administration of fructose reproducibly increases serum magnesium concentrations. This is probably explained by the breakdown of magnesium–adenosine triphosphate (**ATP**) complexes, releasing intracellular magnesium ions as a result of nucleotide degradation by adenosine deaminase. Significant ingestion of fructose is thus also accompanied by marked hyperuricaemia in patients with hereditary fructose intolerance.

Pathology and molecular genetics

Chronic ingestion of fructose in hereditary fructose intolerance causes hepatic injury: there is diffuse fatty change and increased glycogen deposition. Hepatocyte necrosis with intralobular and periportal fibrosis occurs and fully developed cirrhosis results from continued exposure to fructose. After acute experimental challenge, electron microscopy has shown irregular electron-dense material surrounded by membranous structures, suggesting a florid lysosomal reaction to intracellular deposits of fructose 1-phosphate. Fatal administration of fructose or sorbitol parenterally is associated with the abrupt onset of hepatorenal failure associated with bleeding. Histological examination shows hepatic necrosis in these cases (Fig. 2). Loss of cellular functions, for example in the proximal renal tubule, is probably caused by depletion of ATP resulting from the arrested metabolism of fructose by the specialized pathway. The source of the severe abdominal pain that follows ingestion of fructose is unknown, but stimulation of visceral afferent nerves by the local release of purine nucleotides or lactate may be responsible.

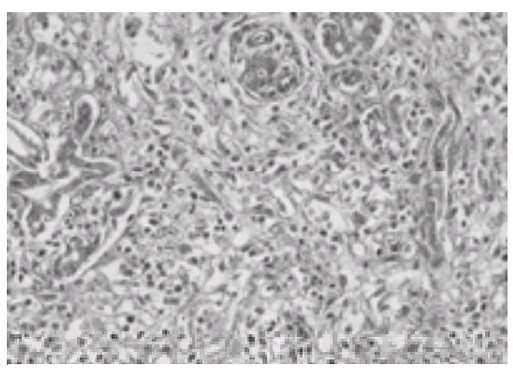

Fig. 2 The effects of fatal infusions of fructose in a young Italian girl.

The genetic basis of aldolase B deficiency has been studied intensively. The human aldolase B gene maps to chromosome 9q22.3. Several point mutations affecting the function of the enzyme are sufficiently widespread in patients of European origin to merit diagnostic investigation. One particular mutation, Ala ^{149}®Pro, which disrupts residues in a substrate-binding domain of aldolase B, is prevalent in Europe. This mutation accounts for most alleles responsible for fructose intolerance, but others, including Ala^{174}®Asp, Asn^{334}®Lys and a four-base deletion in exon 4, are sufficiently frequent and widespread to merit examination in the diagnostic laboratory (see below).

Diagnosis

In infancy and childhood, hereditary fructose intolerance most characteristically causes persistent vomiting, with failure to thrive, acidosis, hypoglycaemia, and jaundice. Clearly in very young infants there is a wide differential diagnosis, but fructose intolerance may be indicated by the nutritional history and feeding difficulties. The presence of reducing sugar in the urine may indicate that fructosuria and amino acids may also be present. Older children and adults report food aversion and may show a striking absence of dental caries. If fructose intolerance is considered, then sucrose, sorbitol, and fructose should be excluded completely before definitive tests can be carried out. Striking improvement, suggestive of hereditary fructose intolerance, may be seen within a few days. The differential diagnosis includes pyloric stenosis, galactosaemia, hepatitis, renal tubular disease, Wilson's disease, and tyrosinosis.

Since the prompt institution of strict dietary treatment has beneficial and, in infants and children, life-saving effects in those with fructose intolerance, every

reasonable effort should be undertaken to make a definitive diagnosis. This will have important consequences for relatives of the propositus and will provide information critical for the introduction of a rigorous and life-long exclusion diet.

The intravenous fructose tolerance test is often useful for diagnosis, particularly in adults: 0.25 g/kg (0.2 g/kg in infants) of D(+)-fructose is infused as a 20 per cent solution over a few minutes and blood samples for potassium ions, magnesium ions, phosphate ions, and glucose are taken at regular intervals over a 2-h period. Epigastric and loin pain accompany the infusion, and hypoglycaemic coma may occur. The hypoglycaemia does not respond to glucagon, therefore glucose for parenteral injection must be available. The test should be carried out under controlled conditions with medical personnel at hand: oral challenge with fructose or sucrose may produce severe pain and shock and is best avoided. Responses differ between individuals and hypoglycaemia is usually milder in adults, but typical responses in hereditary fructose intolerance and a control subject are depicted in Fig. 3. The tolerance test should not be carried out in patients with overt signs of liver disease where it may occasionally yield misleading results, particularly in infants and children.

Fig. 3 (a) Intravenous fructose tolerance tests in a 39-year-old woman with hereditary fructose intolerance proven by fructaldolase essay and DNA analysis. (b) An age- and sex-matched control subject with alcohol-related episodic hypoglycaemia.

Aldolase B deficiency is demonstrated definitively by enzymatic analysis of biopsy samples obtained from the liver or small intestinal mucosa. Biochemical assay of fructaldolases characteristically demonstrates reduced or absent fructose 1-phosphate cleavage activity with a partial deficiency of fructose 1,6-diphosphate aldolase. Since fructaldolase deficiency may accompany other parenchymal disease of the liver, these assays are of limited value in the acutely ill or jaundiced patient.

Recently, a direct diagnosis of hereditary fructose intolerance has been possible, particularly in patients of European ancestry: examination of aldolase B genes for the presence of common mutations responsible for the disease can be carried out by laboratories that specialize in the molecular analysis of genomic DNA. The ability to identify disease alleles by analysing tiny samples of blood or tissue may be beneficial for the investigation of infants with this disorder and, eventually, for postnatal screening before dietary exposure occurs. Tests for fructose intolerance based on the analysis of DNA may avoid the need for invasive or hazardous investigations using tissue biopsy procedures or challenge with parenteral sugar solutions.

Treatment

Dietary treatment of fructose intolerance alleviates the disorder but requires the almost complete exclusion of sucrose, fructose, and sorbitol. The daily consumption of sugar should be reduced to less than 40 mg of fructose equivalents per kilogram body weight (that is, 2–3 g for an adult) in order to reverse the disease manifestations and establish normal development in affected infants and children. The ubiquity of fructose and its congeners in the Western diet presents serious difficulties. Adult patients have usually restricted their consumption of fructose to less than 20 g daily and the source of the residual sugar may be difficult to establish. For this reason, the advice of an experienced dietitian should be sought (Table 1). Particular care needs to be taken with sugar-coated pills and, especially, liquid medications for paediatric use, as large amounts of fructose, sucrose, and sorbitol are frequently present. Children and adults with hereditary fructose intolerance may tolerate the taste of confectionery that contains large quantities of noxious sugars but in which the sweetness is masked by other flavours such as peppermint, which they enjoy. This behaviour may lead to unexplained hypoglycaemic symptoms and other signs of sugar toxicity. Occasionally, patients are unable to tolerate certain foods that are permitted on their diet sheets—in doubtful cases it is advisable to avoid the offending item or to have it analysed. Patients with hereditary fructose intolerance may lack folic acid and vitamin C. Supplements of these vitamins in particular are recommended, especially during pregnancy, but, as with other medicines, care has to be taken to avoid harmful sugars contained in the preparation: Ketovite®; (Paines and Byrne, Ltd, Surrey, England) is a satisfactory source of these vitamins.

Prognosis

Untreated hereditary fructose intolerance is a fatal disease in infants and young children in whom it generally causes irreversible liver disease and episodic, life-threatening, hypoglycaemia. Occasionally, adolescents and adult patients may succumb to the inadvertent use of parenteral fructose or sorbitol but, except in Germany, this practice is now obsolete. With the introduction of a strict exclusion diet, the disorder is compatible with a normal life expectancy.

Further reading

Ali M, Rellos P, and Cox TM (1998) Hereditary fructose intolerance. *Journal of Medical Genetics* **35**, 353–65.

Baker L, Wingrad AI (1970). Fasting hypoglycaemia and metabolic acidosis associated with deficiency of fructose-1,6-diphosphatase deficiency. *Lancet* **ii**, 13–16.

Boesinger P, *et al.* (1994). Changes of liver metabolite concentrations in adults with disorders of fructose metabolism after intravenous fructose by ^{31}P magnetic resonance spectroscopy. *Pediatric Research* **36**, 436–40.

Bell L and Sherwood WG (1987). Current practices and improved recommendations for treating hereditary fructose intolerance. *Journal of the American Dietetic Association* **87**, 721–8.

Chambers RA and Pratt RTC (1956). Idiosyncrasy to fructose. *Lancet* **ii**, 340.

Cox TM (1993). Iatrogenic deaths in hereditary fructose intolerance. *Archives of Diseases in Childhood* **69**, 413–15.

Cox TM (1994). Aldolase B and fructose intolerance. *Journal of the Federation of American Societies for Experimental Biology* **8**, 62–71.

Greenwood J (1989). Sugar content of liquid prescription medicines. *Pharmaceutical Journal* **243**, 553–7.

Kikawa Y, *et al*. (2002). Diagnosis of fructose 1,6-bisphosphatase deficiency using cultured lymphocyte fraction: a secure and noninvasive alternative to liver biopsy. *Journal of Inherited Metabolic Disease* **25**, 41–6.

Odièvre M, *et al.* (1978). Hereditary fructose intolerance in childhood. Diagnosis, management and course in 55 patients. *American Journal of Diseases of Childhood* **132**, 605–8.

Pagliara AS, *et al.* (1972). Hepatic fructose-1,6-diphosphatase deficiency. A cause of lactic acidosis and hypoglycaemia in infancy. *Journal of Clinical Investigation* **51**, 2115–23.

Sachs B, Sternfeld L, Kraus G (1942). Essential fructosuria: its pathophysiology. *American Journal of Diseases of Childhood* **63**, 252.

Steinmann B, Gitzelmann R, Van den Berghe G (2001). Disorders of fructose metabolism. In: Scriver CR, *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 8th edn, Vol II, pp 1489–520. McGraw-Hill, New York.

Wasserman D, *et al.* (1996). Molecular analysis of the fructose transporter gene (GLUT 5) in isolated fructose malabsorption. *Journal of Clinical Investigation* **98**, 2398–402.

11.3.3 Disorders of galactose, pentose, and pyruvate metabolism

T. M. Cox

Inborn errors of galactose metabolism

Galactose is derived principally from the milk sugar, lactose, in the diet by the action of mucosal lactase in the small intestine. The concentration of lactose in human breast milk is approximately 200 millimoles per litre. Newborn infants normally receive about one-fifth of their dietary energy supply in the form of galactose, which is derived from the breakdown of this lactose to galactose and glucose in equimolar amounts. After absorption, galactose serves as a source of glucose. Galactose is a component of many membrane glycoproteins and glycolipids; galactosylated lipids are abundant in nervous tissue.

The conversion of galactose to glucose involves reactions that lead to the formation of glucose 1-phosphate, which can enter the main pathways of carbohydrate metabolism, directly (Fig. 1). The first step involves phosphorylation to form galactose 1-phosphate, which is converted to glucose 1-phosphate and uridine diphosphate-galactose after reaction with the nucleoside diphosphate sugar, uridine diphosphoglucose. Uridine diphosphoglucose is regenerated by the action of uridine diphosphate-galactose-4-epimerase. The presence of this epimerase enables galactose to be produced from glucose for the synthesis of complex glycoconjugates and renders the individual potentially independent of exogenous galactose. Enzymatic defects in the interconversion of these metabolites lead to increased blood and tissue concentrations of galactose, especially after meals containing milk or dairy products. There are three inborn errors of galactose metabolism recognized: (1) galactokinase deficiency; (2) galactose 1-phosphate uridylyltransferase deficiency; and (3) uridine diphosphate-4-epimerase deficiency.

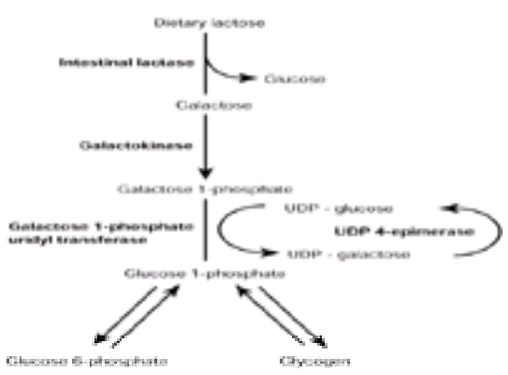

Fig. 1 Galactose metabolism.

Galactokinase deficiency: 'galactose diabetes'

Failure to phosphorylate galactose in the liver and other tissues impairs its clearance from the blood so that the free sugar and its metabolites, galactonic acid and galactitol, appear in the urine. Homozygous deficiency of galactokinase occurs with an approximate frequency of 1 in 100 000 live births.

Clinical features

Precocious formation of cataracts in infants and children is characteristic, with some heterozygotes developing cataracts before the age of 40 years. When blood concentrations are high, galactose is taken up by the lens and converted to the end-product galactitol by the action of aldose reductase: subsequent toxic or osmotic effects lead to swelling and irreversible damage to lens fibres. Patients with galactokinase deficiency persistently excrete reducing sugar in their urine but, apart from possible confusion with diabetes mellitus, this has no apparent significance.

Diagnosis and treatment

Galactokinase deficiency should be suspected in infants or children with cataracts and reducing sugar should be sought in the urine. This sugar will not react with glucose oxidase test strips. Definitive diagnosis by enzymatic assay of galactokinase in erythrocytes or cultured fibroblasts differentiates the disorder from classic galactosaemia and hypergalactosaemia due to vascular disease in the liver. Treatment with a strict lactose and galactose exclusion diet prevents cataract formation. The human gene for galactokinase maps to chromosome 17q24, with a putative second locus on chromosome 15. Several mutations responsible for galactokinase deficiency have been identified in the *GALK1* gene at the chromosome 17 locus.

Galactose 1-phosphate uridylyltransferase deficiency: galactosaemia

Unlike individuals in whom galactokinase is deficient, when patients who lack uridylyltransferase activity ingest lactose, there is a significant rise in intracellular galactose 1-phosphate as well as in the blood galactose concentration. The severe consequences of classical galactosaemia thus result from the toxic effects of galactose 1-phosphate principally in cells of the liver, proximal renal tubule, and brain. Although the exact mechanism of toxicity is unknown, as in hereditary fructose intolerance, the accumulated metabolite probably inhibits other enzymatic reactions involving phosphorylated intermediates and may lead to purine nucleotide depletion.

Recognition of galactosaemia in early infancy is of paramount importance since the acute effects of galactose poisoning may be reversed by institution of a lactose exclusion diet. However, the ability of dietary therapy to promote a completely healthy long-term outcome has now been questioned by follow-up studies in large cohorts of patients with classical galactosaemia and therefore more research is needed to improve our understanding about the pathogenesis of tissue injury in this nutritional disease.

Clinical and pathological features

The affected infant appears normal at birth, but vomiting or diarrhoea, jaundice, and hepatomegaly usually occur in the first few weeks. There is failure to gain weight, subcutaneous bruising, and progressive enlargement of the liver. Cataracts may be apparent at 1 month of age, by which time abdominal distension with ascites has developed. Mental retardation does not become manifest until later in the first year of life and varies greatly in severity. Many patients suffering from galactosaemia develop severe infections with *Escherichia coli* during the neonatal period, and Gram-negative bacterial sepsis may be the first indication of this disorder in young infants. A bactericidal defect in circulating leucocytes has been postulated. In adult patients after reversal of the acute galactose toxicity syndrome, the most obvious sequelae are growth failure, neurological deficit, and primary ovarian failure with infertility.

Occasional patients with galactosaemia remain asymptomatic while ingesting milk but gradually fail to gain weight. Such patients may come to light during childhood or even adult life, because of varying degrees of mental retardation and cataracts. Hepatomegaly and intermittent galactosuria are usually present, and often there is a history of feeding difficulties on institution of modified formula feeds during the neonatal period.

The neurological manifestations of classical galactosaemia are highly variable but, despite prompt institution of dietary therapy, a degree of mental retardation is common in affected children and adults. Characteristic learning difficulties in mathematics and spatial relationships with behavioural deficits have been observed. It appears that the galactose-free diet fails to confer benefit on mental development when instituted beyond the age of 2 years. In follow-up studies of galactosaemic children and adults, a range of neurological deficits, including seizures, apraxia, extrapyramidal disorders, and cerebellar signs, have been documented despite strict dietary measures.

Serum tests of liver function are non-specifically deranged: histological examination shows lobular fibrosis, fatty change, bile ductular proliferation and progression to frank cirrhosis. A haemorrhagic tendency is an early feature of galactosaemia and the diagnosis should be considered in jaundiced infants with signs of a bleeding diathesis. Involvement of the proximal renal tubule is shown by generalized aminoaciduria and occasionally a full-blown Fanconi syndrome with vacuolation of tubular epithelial cells. Histological examination of the brain shows non-specific signs of injury with gliosis and Purkinje cell loss in the cerebellum. Follow-up studies of female patients with galactosaemia has shown a high incidence of gonadal failure with ovarian atrophy: although this complication appears to be more common in patients in whom dietary therapy was delayed, no clear cause-and-effect relationship has been established. A toxic effect on the fetal ovary due to maternal hypergalactosaemia has been postulated to account for the hypergonadotrophic hypogonadism in affected women and girls. No evidence of gonadal failure has been found in male patients.

Genetic studies

Galactosaemia is transmitted as an autosomal recessive trait with an overall estimated frequency of 1 in 62 000. Classical galactosaemia is rare in Japan but frequent in some isolated groups, most notably in the modern Traveller population of Ireland. In this group, screening methods indicate a birth frequency of 1 in 480 compared with 1 in 30 000 in the non-Traveller Irish population. In Black patients from the United States a relatively mild disorder as been reported that is probably due to an unstable enzyme variant; uridylyltransferase activity is absent from their red cells but amounts to some 10 per cent of normal in samples of liver and small intestinal tissue. Individuals with the so-called 'Duarte variant' possess about half the normal enzyme activity in erythrocytes but remain asymptomatic.

The human galactosyl-1-phosphate uridylyltransferase gene maps to human chromosome 9p13 and encodes a protein of molecular weight 43 000 Da, which exists as a functional homodimer. Molecular analysis of the transferase gene indicates that most patients with classical galactosaemia harbour missense-type mutations and are compound heterozygotes. Several variant transferase enzymes have been described. Molecular analysis of the transferase gene has identified several widespread mutations; for example one mutant allele (Q188R) is in linkage disequilibrium with a restriction fragment-length polymorphism flanking exon 6 of the gene sequence in multiple populations worldwide, including the Irish Travellers – galactosaemic patients amongst whom, are all homozygous for Q188R. A less frequent mutation of diagnostic significance in White populations is designated R333W; the Duarte transferase mutation has been identified as N314D. Molecular analysis of the transferase gene now renders prenatal diagnosis of at-risk pregnancies possible.

Diagnosis

Galactosaemia may be suspected in an infant with growth failure, cataracts, liver disease, aminoaciduria, mental retardation, and especially where reducing sugar is present in the urine. The occurrence of unexplained bacterial sepsis, especially if due to *E. coli* infection in a newborn infant, may indicate galactosaemia. Cataracts may be detected by slit-lamp examination in the first few days of life.

The finding of hypergalactosaemia is not specific for those hereditary galactosaemias due to inherited deficiences of galactose-metabolizing enzymes. Recent studies show that persistent hypergalactosaemia may be commonly due to portosystemic venous shunts in infants that are often associated with patent ductus venosus or other congenital vascular abnormalities in the liver. Doppler ultrasonography is a convenient non-invasive investigation to search for such shunts in young infants.

Definitive diagnosis of hereditary galactosaemia is mandatory, and relies on the determination of galactose 1-phosphate uridylyltransferase activity and other galactose-metabolizing enzymes in red cells or leucocytes by means of a specific enzymatic assay. Reliable enzymatic or genetic testing for heterozygotes can be carried out in the parents of a child who died before the diagnosis was confirmed. In particular populations, neonatal screening for elevated blood galactose and galactose 1-phosphate concentrations is carried out routinely. Molecular analysis of the gene for galactose 1-phosphate uridylyltransferase deficiency in at-risk pregnancies has been requested by some affected families.

Treatment

Without strict dietary treatment, most patients with galactosaemia die in early infancy, although some may survive with liver disease and mental retardation beyond childhood. The course of galactosaemia is altered strikingly upon withdrawal of lactose (and galactose), although the outcome of neurological disease is often disappointing. However, lactose is present in many non-dairy foods and advice from an experienced dietician, as well as meticulous attention to detail, is required to eliminate it completely. In infants, soybean milks or commercial casein hydrolysates, 'Nutramigen', are used as milk substitutes and therapy is monitored by periodic assay of red cell galactose 1-phosphate concentrations. Despite reports that galactose may be reintroduced as the patient develops, lifelong strict adherence to the exclusion diet should be advocated. In subsequent pregnancies of heterozygous mothers who have had affected children, there is evidence that premature cataracts can be avoided in the fetus if the maternal intake of lactose is restricted. In late pregnancy, lactosaemia and lactosuria are common findings and result from the physiological induction of lactose biosynthesis in mammary tissue. In rare cases (see below) there is a risk of self-intoxication when women with homozygous deficiency of the transferase become pregnant and breast feed, so that additional dietary precautions are needed to maintain metabolic control during lactation.

Prognosis

The acute manifestations of galactosaemia and growth failure respond quickly to dietary therapy and cataract formation is prevented. Unfortunately, a proportion of patients have significant neurological deficits despite prompt and conscientious treatment. An international survey of the long-term outcome in 350 patients receiving dietary therapy has been published by Waggoner and colleagues. The presence of ovarian failure and elevated galactose 1-phosphate concentrations in patients apparently ingesting no lactose or galactose raises the possibility that an endogenous pathway of galactose 1-phosphate formation from the pyrophosphorylysis of uridine diphosphate-galactose may occur. This may also explain the late emergence of neurological disease in treated patients. Long-term follow-up and periodic neuropsychiatric, as well as physical, monitoring is recommended. Recently, several pregnancies have been reported in women suffering from classical galactosaemia, including subjects homozygous for the *Q188R* mutation. In such pregnancies, high concentrations of galactitol are found in amniotic fluid but cord blood values have been determined to be within the range found in galactosaemic patients receiving strict dietary therapy. Thus, although maternal galactitol traverses the placenta, it probably does not harm the heterozygous fetus.

Uridine diphosphate-4-epimerase deficiency

Epimerase deficiency is very rare but may be identified during screening for classic galactosaemia. In most cases no symptoms attributable to galactosaemia are apparent and follow-up studies have confirmed the usually benign nature of this anomaly. However, a few cases of marked deficiency of uridine diphosphate-4-epimerase have been discovered in patients otherwise manifesting the classic features of galactosaemia. In the absence of epimerase activity, the individual is dependent on exogenous sources of galactose, since this cannot be derived from glucose. The autosomal recessive nature of this inherited disorder has been confirmed by demonstrating a partial epimerase deficiency in the healthy parents of an affected infant. As a complete deficiency of the epimerase would lead to an absolute lack of uridine diphosphate (**UDP**)-galactose for glycosphingolipid synthesis, the ingestion of very small quantities of galactose has been recommended in this unusual disorder so that brain development and biosynthesis of essential galactosides can proceed. Because of the dual activity of the epimerase towards UDP-acetyl glucosamine as well as UDP-glucose, it has been suggested that small supplements of the aminoacetyl galactosamine should also be provided in the diets of patients with UDP galactose-4-epimerase deficiency. This condition may be contrasted with the transferase deficiency that allows the formation of small amounts of endogenous galactose in the presence of an intact epimerase. The gene for human UDP-galactose-4-epimerase has been mapped to chromosome 1p36–p35, and several mutant alleles has been identified.

Pentosuria

Pentosuria is caused by the excessive renal excretion of L-xylulose: this has no clinical significance, except that it may lead to the incorrect diagnosis of diabetes mellitus should tests for reducing sugar be carried out on the urine. Xylulose does not react with urinary test strips based on the glucose oxidase method.

Although pentosuria is a rare autosomal recessive trait, its frequency in Ashkenazi Jews may be as high as 0.05 per cent. It is caused by enzymatic deficiency of L-xylulose reductase in the oxidative pathway of glucuronate metabolism, which results in 1 to 4 g of xylulose and L-arabitol continuously appearing in the urine: output is greatly enhanced by the ingestion of glucuronic acid or drugs that are excreted as glucuronides.

Inborn errors of pyruvate metabolism

Pyruvate dehydrogenase

Deficiency of pyruvate dehydrogenase is the most common cause of lactic acidosis in newborn infants and children, but it is also associated with neurodegenerative syndromes in later life. Pyruvate dehydrogenase exists as a multienzyme complex representing the products of 10 distinct genes. However, defects in one subunit of pyruvate dehydrogenase itself (E1a) account for most patients so far investigated, although defects in dihydrolipoyl dehydrogenase (E3) are also described.

Biochemical defect

The pyruvate dehydrogenase (**PDH**) complex catalyses the conversion of pyruvate to acetyl-coenzyme A within mitochondria and operates at about 10, 40, and 70 per cent of capacity in the liver, heart, and brain, respectively. The PDH complex is critical for brain metabolism since this is normally entirely dependent on the oxidative breakdown of glucose. There are three main activities associated in the complex: (1) pyruvate dehydrogenase, a thiamine-dependent moiety (E1); (2) dihydrolipoyl transacetylase (E2); and (3) dihydrolipoyl dehydrogenase (E3). Also associated are a pyruvate dehydrogenase-specific kinase and phosphatase (both involved in overall metabolic regulation of the complex) as well as an essential lipoic acid moiety.

The accumulated pyruvate may either be reduced to lactate or transaminated to alanine, so that hyperalaninaemia and varying degrees of lactic acidaemia occur. Very rare defects in dihydrolipoyl dehydrogenase are associated with deficiency of branched-chain keto-acid dehydrogenase. Failure to carry out oxidative reactions in regions of the cortex and midbrain causes neuronal death, and deficiency of 4-carbon intermediates may critically impair neurotransmitter synthesis.

Clinical features and prognosis

Severe deficiency of pyruvate dehydrogenase affects intrauterine development and causes marked acidosis (blood lactate >10 mmol/l) at birth with early death. The clinical presentation of pyruvate dehydrogenase deficiency is strikingly heterogeneous. Many victims do not show clinically significant metabolic acidosis and come to light because of intrauterine growth failure, neonatal hypotonia asphyxia, and feeding difficulty. In some affected individuals the enzyme deficiency is responsible for a slowly progressive neurodegeneration associated with dysgenesis and other structural abnormalities in the olivopontocerebellar tract and periventricular grey matter. Cortical atrophy and agenesis of the corpus callosum have also been reported in association with spastic quadriplegia. In those with neurological manifestations, blood lactate concentrations do not exceed 10 mmol/l. Should feeding by gavage be instituted, there is a protracted course with failure of neurological development, microcephaly, quadriplegia, seizures, and blindness. Intermittent cerebellar ataxia or torsion dystonia has been recorded and choreoathetoid movements occur. Involuntary eye movements in children are associated with a progressively deteriorating course. In a few patients with intermittent cerebellar ataxia, hereditary spinocerebellar degeneration appearing in early adult life has been attributed to the deficiency of pyruvate dehydrogenase but there is no direct relationship to Friedreich's ataxia. In patients who present with severe acidosis at birth, subacute necrotizing encephalomyelopathy of the Leigh's type has been confirmed at necropsy and deficiency of pyruvate dehydrogenase activity has been demonstrated.

Genetics

The most common cause of pyruvate dehydrogenase deficiency is due to a defect in the E1a subunit—a protein encoded on the X chromosome. Although the disease is characteristically more severe in males, manifestations in the heterozygous female are unusually frequent for an X-linked disease and probably reflect the low functional reserve of the enzyme complex in the brain. Neonatal lactic acidosis is more frequent in males. An auxiliary gene for the E1a subunit is localized as a result of retroposition from the X-chromosome to the long arm of chromosome 4, but is expressed only during spermatogenesis; its presence, however, indicates the critical need for activity of the complex in nearly all tissues. Causal mutations in the *E1a* gene on the X chromosome have been described—most appear to be short deletions or duplications and, at present, are not generally applicable for diagnosis. However, analysis of X-chromosome inactivation patterns, by determination of methylation status, has proved useful for the evaluation of enzymatic assays of fibroblasts obtained from obligate carriers or female patients in whom the diagnosis is suspected.

Diagnosis and treatment

The diagnosis is suspected from the presence of severe acidosis at birth. It may also emerge during the investigation of neurological deficits, especially where they are associated with intrauterine growth failure. Routine screening of urine samples for organic acids may identify excessive pyruvate, lactate, and alanine excretion. In patients without clinically evident acidosis, cerebral disease is accompanied by striking elevations of lactate and pyruvate in the cerebrospinal fluid. Mutation analysis of the X-linked *PDH* gene and determination of the abundance of immunoreactive PDH protein now permits decisive diagnosis of this disease.

Neuroradiological procedures, including cerebral ultrasonography and computed tomography, reveal ventricular dilatation and cerebral atrophy. In several infant girls with PDH deficiency, magnetic resonance imaging showed hypoplasia of the corpus callosum as well as loss of normal white matter signal intensity. Proton magnetic resonance spectroscopy (**MRS**) revealed high-abundance signals for brain lactate with decreased intensity of *N*-acetylaspartate, while phosphorus MRS of skeletal muscle showed abnormally low muscle phosphorylation potentials, in keeping with the predicted biochemical disturbance. Pathological examination of previously affected siblings shows shrinkage of gyri, with involvement of the medulla shown by loss or hypoplasia of the pyramids. The pathological features of Wernicke's encephalopathy may be present. The corpus callosum may be absent. Definitive diagnosis, however, depends on genetic and enzymatic studies in skin fibroblasts or blood leucocyte samples.

Institution of a high-fat, low-carbohydrate, ketogenic diet may ameliorate the biochemical abnormalities, but, given the degree of neurological impairment that is normally present at diagnosis, little clinical improvement can be expected. Therapeutic responses to the administration of high-dose thiamine have been reported in patients with partial enzymatic deficiency, notably where ataxia and abnormal eye movements reminiscent of Wernicke's encephalopathy were conspicuous. In rare patients with the autosomally recessive condition due to dihydrolipoyl dehydrogenase deficiency, oral administration of lipoic acid has been reported to correct the organic acidaemia with clinical improvement.

Pyruvate carboxylase deficiency

Inborn defects in pyruvate carboxylase, a key gluconeogeneic enzyme, cause hypoglycaemia or profound metabolic acidosis with neurological disease. The manifestations of this latter syndrome closely resemble those caused by deficiencies of pyruvate dehydrogenase activity. A severe form associated with hyperammonaemia and citrullinaemia is also recognized, particularly in patients of French descent.

Metabolic defect

Pyruvate decarboxylase is a biotin-dependent enzyme that catalyses the first step in the formation of oxaloacetate from pyruvate and is activated allosterically by acetyl-coenzyme A. Thus, hypoglycaemia would be expected only after glycogen stores had been depleted. Krebs cycle intermediates may become depleted so that there is an insufficient synthesis of neurotransmitters. There may also be a reduced supply of aspartate for the arginosuccinate synthase reaction of the urea cycle.

Clinical features

Patients with severe deficiency of pyruvate carboxylase may present with the Leigh syndrome (necrotizing encephalomyopathy with lactate/pyruvate acidosis) or

hypotonia and neurological retardation. The presence of ataxia and abnormal ocular movements in life suggest the occurrence of midbrain disease resembling Wernicke's encephalopathy. Hypoglycaemia frequently occurs during intercurrent infection or during starvation and acidosis, requiring bicarbonate therapy. The most severe form, originally reported from France, progresses rapidly with evidence of liver damage, hyperammonaemia, and citrullinaemia.

Genetics

This disorder is transmitted as an autosomal recessive trait. In severely affected patients with hyperammonaemia, pyruvate carboxylase protein and its mRNA are absent in the liver. A partially inactive variant enzyme is detectable in other patients.

Diagnosis and treatment

The condition is suspected when acidosis and neurological disease occur in infants, especially in the presence of hypoglycaemia. Specific diagnosis requires enzymatic assay in fibroblasts, which can also be used for carrier detection. Disorders of pyruvate metabolism may be mimicked biochemically by mitochondrial diseases and acquired deficiencies of thiamine or biotin. Although biotin therapy has been disappointing in pyruvate carboxylase deficiency, occasional responses to high-dose lipoic acid and thiamine treatment, which may stimulate the pyruvate metabolism by the dehydrogenase complex, have been recorded.

Therapy

Episodes of acidosis are treated with intravenous sodium bicarbonate, and glucose may be required for hypoglycaemia. There is evidence that ketogenic diets containing 50 per cent fat and 20 per cent carbohydrate ameliorate the biochemical disturbance and delay the onset of neurological disease: the administration of glutamate and aspartate, which may act as a source of oxaloacetate, appear to have been beneficial in some patients.

Further reading

Inborn errors of galactose metabolism

Bowring FG, Brown ARD (1986). Development of a protocol for newborn screening for disorders of the galactose metabolic pathway. *Journal of Inherited Metabolic Disease* **9**, 99–104.

Cornblath M, Schwartz R (1991). Disorders of galactose: metabolism. In: Cornblath M, Schwartz R, eds. *Disorders of carbohydrate metabolism in infancy*, 3rd edn, pp 295–324. Blackwell Scientific, Boston.

Elsas LJ, Lai K (1998). The molecular biology of galactosemia. *Genetic Medicine* **1**, 40–8.

Gitzelmann R (1967). Hereditary galactokinase deficiency; a newly-recognized cause of juvenile cataracts. *Pediatric Research* **1**, 14–23.

Holton JB, Walter JH, Tyfield LA (2001). Galactosemia. In: Scriver CR, *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 8th edn, Vol 1, pp 1553–85. McGraw-Hill, New York.

Holton JB, *et al.* (1981). Galactosaemia. A new severe variant due to uridine diphosphate galactose-4-epimerase deficiency. *Archives of Diseases in Childhood* **56**, 885–7.

Kaufman FR, *et al.* (1986). Gonadal function in patients with galactosaemia. *Journal of Inherited Metabolic Disease* **9**, 140–6.

Mizoguchi N, *et al.* (2001). Congenital porto-left renal venous shunt as a cause of galactosaemia. *Journal of Inherited Metabolic Disease* **24**, 72–8.

Murphy M, *et al.* (1999). Genetic basis of transferase-deficient galactosaemia in Ireland and the population history of Irish Travellers. *European Journal of Human Genetics* **7**, 549–54.

Schweitzer S, *et al.* (1993). Long-term outcome in 134 patients with galactosaemia. *European Journal of Paediatrics* **152**, 36–43.

Waggoner DD, Buist NRM, Donnell GN (1990). Long-term prognosis in galactosaemia: results of a survey of 350 cases. *Journal of Inherited Metabolic Disease* **13**, 802–18.

Pentosuria

Hiatt HH (2001). Pentosuria. In: Scriver CR, *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 8th edn, Vol 1, pp 1590–9. McGraw-Hill, New York.

Inborn errors of pyruvate metabolism

Brown GK, *et al*. (1988). Cerebral lactic acidosis: defects in pyruvate metabolism with profound brain damage and minimal systemic acidosis. *European Journal of Pediatrics* **147**, 10–14.

Brown RM, Otero LJ, Brown GK (1997). Transfection screening for primary defects in the pyruvate E1-alpha subunit gene. *Human Molecular Genetics* **6**, 1361–7.

Brown GK, *et al.* (1994). Pyruvate dehydrogenase deficiency. *Journal of Medical Genetics* **31**, 875–9.

Dahl H-M, *et al.* (1992). X-linked pyruvate dehydrogenase E1-alpha subunit deficiency in heterozygous females: variable manifestation of the same. *Journal of Inherited Metabolic Disease* **15**, 835–47.

Hinman LM, *et al.* (1989). Deficiency of pyruvate dehydrogenase complex in Leigh's disease fibroblasts: an abnormality in lipoamide dehydrogenase affecting PDHC activation. *Neurology* **39**, 70–5.

Lissens W, *et al*. (2000). Mutations in the X-linked pyruvate dehydrogenase (E1) alpha subunit gene (PDHA1) in patients with a pyruvate dehydrogenase complex deficiency. *Human Mutation* **15**, 209–19.

Robinson BH, *et al*. (1987). The French and North American phenotypes of pyruvate carboxylase deficiency. *American Journal of Human Genetics* **40**, 50–9.

Shevell MI, *et al*. (1994). Cerebral dysgenesis and lactic acidemia: an MRI/MRS phenotype associated with pyruvate dehydrogenase deficiency. *Pediatric Neurology* **II**, 224–9.

11.4 Disorders of purine and pyrimidine metabolism

*Richard W. E. Watts**

These disorders are due to abnormalities in the biosynthesis, interconversion, and degradation of the purines, adenine and guanine and of the pyrimidines, cytosine, thymine, and uracil. All these compounds are heterocyclic bases which exist as tri-, di-, and monophosphorylated either deoxyribosylated or ribosylated derivatives (deoxyribose and ribose are pentose carbohydrates). The phosphorylated deoxyribosylated and ribosylated derivatives are termed 'nucleotides', while the purely ribosylated derivatives, which lack the phosphate group, are termed 'nucleosides'.

The polynucleotide deoxyribonucleic acid (**DNA**) contains equimolar amounts of adenylic acid (adenosine monophosphate, **AMP**), guanylic acid (guanosine monophosphate, **GMP**), thymidylic acid (thymidine monophosphate, **TMP**), and cytidylic acid (cytidine monophosphate, **CMP**). Uridylic acid (uridine monophosphate, **UMP**) replaces TMP in the polynucleotide ribonucleic acid (**RNA**).

DNA encodes, stores, and ultimately hands on genetic information. RNA transcribes and translates the genetic information so that the corresponding polypeptide strand emerges from its site of synthesis on the ribosome.

The purine nucleotides, their cyclic derivatives (CAMP and CGMP), and their more highly phosphorylated derivatives have functions in many aspects of intermediary metabolism. Purine compounds also function as signal transducers, neurotransmitters, vasodilators, and mediators of platelet aggregation.

Disorders of purine metabolism

The purine nucleotides are built up in a step-wise manner (*de novo* synthesis), undergo a series of interconversion and salvage reactions, and a final degradative process to yield uric acid, as shown in Fig. 1. Most human morbidity connected with diseases of purine metabolism is due to sodium urate monohydrate and to uric acid. The dietary intake of nucleoproteins is also an important factor in diseases due to sodium urate and uric acid. Ingested adenine and guanine nucleotides are degraded to free purine bases and, hence, to uric acid by enzymes in the intestinal juices and small-intestine mucosa so that the products of their metabolism do not mix with the corresponding endogenous metabolic pools except at the final uric acid stage.

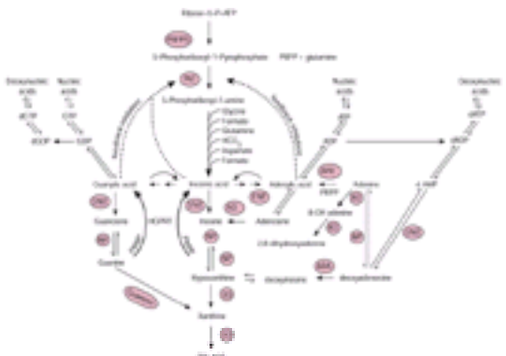

Fig. 1 Pathways of purine metabolism in humans. **ADA**, adenosine deaminase; **APRT**, adenine phosphoribosyltransferase; **HPRT**, hypoxanthine–guanine phosphoribosyltransferase; **5-NP**, nucleoside phosphorylase; **5'-NP**, 5'-nucleotidase; **PAT**, phosphoribosyl pyrophosphate amidotransferase; **PRPPS**, phosphoribosyl pyrophosphate synthetase; **XO**, xanthine oxidase.

De novo synthesis contributes about 300 to 600 mg (1.8–3.6 mmol/day) and dietary purines about 600 to 700 mg (3.6–4.2 mmol/day) to the dynamic urate metabolic pool of about 1200 mg (7.2 mmol) expressed as uric acid. Each day, about two-thirds of uric acid are excreted in the urine and about one-third is destroyed, mainly by bacterial uricolysis in the gut.

The urate anion is freely filterable at the renal glomerulus, only 5 to 10 per cent being very loosely bound to the plasma proteins (a $_{1–2}$-globulin fraction). The physiologically important p*K*a value of uric acid is 5.75, so that it exists mainly as the monovalent urate anion in plasma (pH 7.4) and assumes more of the free acid form when it passes into regions of the renal tubule, the contents of which are at lower pH values.

The kidney handles urate by:

1. glomerular filtration of virtually 100 per cent of the filtered load;

2. proximal tubular reabsorption by a urate/chloride exchanger in the endothelial brush border (99 per cent of the filtered load);
3. tubular secretion (equivalent to about 50 per cent of the filtered load);
4. postsecretory reabsorption (equivalent to about 40 per cent of the filtered load).

Thus, the net renal clearance of uric acid is approximately 10 per cent of the filtered load and is in the range of 6 to 11 ml/min per 1.73m ^{2} (1.73m^2 = average body surface area of an adult). The exact location of the reabsorptive, secretory, and postsecretory reabsorptive processes within the distal nephron is unclear.

Plasma urate levels

The currently quoted overall reference range for plasma urate (expressed as uric acid) in adults is 3.5 to 8.1 mg/dl (210–480 µmol/l) for men and 2.5 to 6.5 mg/dl (150–390 µmol/l) for women. The corresponding value for children is 1.0 to 4.0 mg/dl (60–240 µmol/l). It rises at puberty with female values being lower than those in men until the menopause, after which it gradually rises to the male value. Extrinsic factors, particularly diet, plumbism, the prevalence of a high ethanol intake in the community, and the prevalence of diseases such as malaria and thalassaemia, which lead indirectly to either increased purine biosynthesis or decreased excretion (Table 1), affect the plasma urate distribution in different populations.

The plasma urate concentration decreases during pregnancy, the reference range being 1.7 to 4.5 mg/dl (100–270 µmol/l). Hyperuricaemia is a characteristic and often an early feature of pre-eclampsia, preceding the proteinuria and hypertension, and is a diagnostically valuable parameter. It results from a reduced renal urate clearance and tends to be associated with hypocalciuria.

Epidemiological studies show significant variations in plasma urate concentrations between different ethnic groups: for example, Maoris and Polynesians have higher values than Western Europeans and Americans. This illustrates the genetic, presumably, polygenic aspects in the control of serum uric acid. Other epidemiological studies emphasize the importance of the environmental factors of purine, protein, and alcohol intake. For example, Gresser and Zöllner showed that the cumulated frequency of plasma urate, expressed as uric acid, rose from approximately 6.2 mg/dl (370 µmol/l) to about 9.0 mg/dl (536 µmol/l) between 1962 and 1971 in association with the improved nutritional state of the Bavarian population from the near-starvation conditions following the Second World War (Fig. 2).

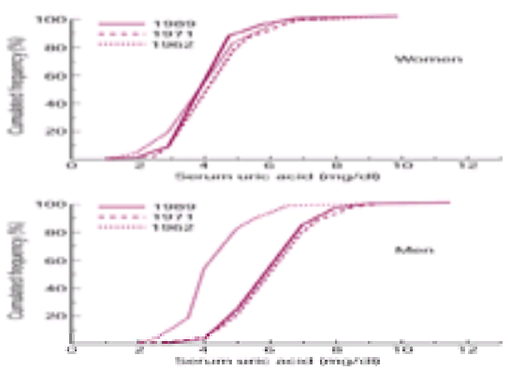

Fig. 2 Differences in the cumulated frequencies in urate levels in female and male blood donors in Bavaria between 1962 and 1989. (Reproduced with permission from Gresser and Zollner 1991.)

Similarly, the plasma urate levels in immigrant communities with low values in their home lands, move towards the values prevailing in the host country as they adopt the lifestyle and dietary habits of that country: for example, Filipinos migrating to the United States. Similarly, migrants with genetically determined high urate levels become even more hyperuricaemic.

The frequency distribution of plasma urate values based on asymptomatic populations is only approximately Gaussian, with an excess of higher values due to the inclusion of some asymptomatic hyperuricaemic subjects. Although plasma is saturated with monosodium urate at a concentration of 7.0 mg/dl (420 µmol/l), higher concentrations of urate can remain in a stable supersaturated solution in plasma without producing any symptoms. Ignoring the slight asymmetry of the frequency distribution and defining normality as the mean value ± 2 standard deviations about the mean, normal values of 7.0 mg/dl (420 µmol/l) for men and 6.0 mg/dl (360 µmol/l) for women have been widely adopted. This has led to considerable overtreatment of patients who have quite innocuous plasma urate concentrations.

Gout and hyperuricaemia

Gout is a classic example of a multifactorial disease in which there is an interplay of genetic and environmental factors. The overall effects of this interplay are wide, extending from cases where there is a clear-cut family history with autosomal dominant inheritance (Fig. 3) to those where environmental factors are the determinants, although often against a genetic background that may be either unifactorial or multifactorial. Gout *per se*, does not shorten life, although some of its complications may do so in the absence of treatment.

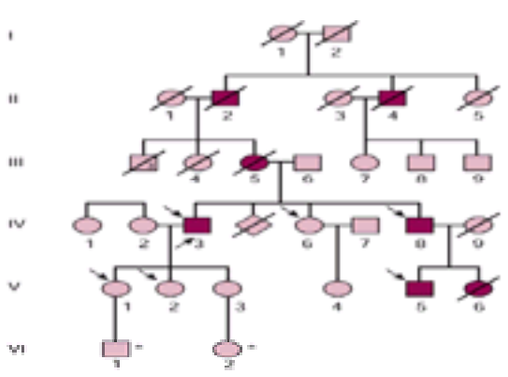

Fig. 3 Pedigree chart of a family showing autosomal dominant inheritance of gout complicated in some cases by renal failure (hyperuricaemia nephropathy). n • male and female subjects, respectively, with hyperuricaemia and renal failure; male and female subjects not known to be affected; Š; deceased male and female subjects, propositus; ¯ subjects whose rates of mononuclear cell *de novo* purine synthesis was measured and shown to be normal; * babies who were examined clinically but not further investigated. (Reproduced with permission from McDermott, *et al.* (1984). *Clinical Science* **67**, 249-58. ©Biochemical Society and Medical Research Society.)

Gout is defined as a syndrome brought about by the crystallization of monosodium urate monohydrate *in vivo* from body fluids supersaturated with this salt. The supersaturation results from either the overproduction or underexcretion of urate, or from a combination of these defects. The underlying causes of hyperuricaemia and gout are:

1. *Decreased net tubular urate secretion*: this is the major factor in the aetiology of the majority of those cases of gout previously described as being idiopathic (or primary), the hereditary predisposition to which is often compounded by environmental factors (e.g. high dietary purine intake and alcoholism).
2. *Identifiable enzymatic defects that accelerate urate* de novo *synthesis*: these are a hypoxanthine–guanine phosphoribosyltransferase (**HPRT**) deficiency which causes the Lesch–Nyhan syndrome and, in milder degrees of HPRT deficiency, some cases of the X-linked recessive hyperuricaemia, gout, and uric acid stone syndrome;
3. *Phosphoribosyl pyrophosphate (PRPPS) synthetase superactivity*: this also presents as X-linked recessive hyperuricaemia, gout, and uric acid stones and, in

some cases, neurological manifestations (e.g. deafness and autism).

Secondary causes of hyperuricaemia and gout are shown in Table 1.

The following abnormalities are commonly associated with, but not causally related to hyperuricaemia and gout:

1. obesity;
2. dyslipidaemia (usually type 4) with raised very low-density (**VLD**) lipoproteins and normal cholesterol levels, and sometimes hypercholesterolaemia with elevated low-density lipoprotein (**LDL**)-cholesterol and low high-density lipoprotein (**HDL**)-cholesterol levels;
3. hypertension;
4. insulin resistance with hyperinsulinaemia and impaired glucose tolerance;
5. ischaemic heart disease.

Thus, these patients may display the features of the 'metabolic syndrome X'.

There is no evidence that uric acid is toxic to the myocardium. Hyperuricaemia may be a marker of coincident cardiac disease, but **not** a causal risk factor. The elevated plasma uric acid concentrations observed in patients with ischaemic heart disease could arise from upregulated vascular adenosine synthesis associated with ischaemia and the subsequent degradation of adenosine to uric acid. However, the relationship of urate to endothelial function is complex. Plasma uric acid accounts for 60 per cent of the free-radical scavenging activity in human plasma: it interacts with peroxynitrile to form a stable nitric oxide donor, so promoting vasodilatation and reducing the potential for peroxynitrile-induced oxidative damage. Conversely, it could have an adverse effect on endothelial function by promoting leucocyte adhesion to the endothelium. However, there is no clear evidence that these actions are significant at the clinical level.

The fractional excretion of urate is the ratio of urate clearance to the glomerular filtration rate (**GFR**). In the presence of normal overall renal function, this can be measured on a random urine sample and a simultaneous plasma sample. The equation simplifies to:

$$\text{Fractional clearance of urate} = U_{\text{urate}} \times P_{\text{creatinine}} / P_{\text{urate}} \times U_{\text{creatinine}}$$

where '*U*' and '*P*' represent the urine and plasma concentrations. The fractional clearance can be used to assess the role of renal tubular dysfunction in the production of hyperuricaemia, provided that the overall renal function is normal.

Acute gouty arthritis

Acute gout is a sodium urate monohydrate-induced crystal inflammation of joints, bursas, and tendon sheaths. Clinically, the affected structures—classically, the first metatarsophalangeal joint is the first joint affected—become acutely inflamed, exquisitely tender, warm to the touch, and the overlying skin becomes red, shiny, and itchy and may desquamate as the inflammation subsides spontaneously over 5 to 15 days in the absence of treatment. Inflammation is usually maximal within 24 h of onset and is accompanied by pyrexia and malaise.

Joint aspiration yields a fluid containing polymorphonuclear leucocytes and negatively birefringent sodium urate monohydrate crystals. The attacks occur most frequently when the plasma urate level is rising or falling. Monosodium urate crystals may be found within monocytes in asymptomatic joints; it has recently been proposed that the inflammatory response to monosodium urate is influenced by the state of monocyte to macrophage differentiation, the balance being tipped towards acute inflammation by the recruitment of undifferentiated monocytes and neutrophils from the bloodstream by one of the precipitants for acute gout. At the beginning of an acute gouty attack, monocyte/macrophage activation leads to the production of inflammatory cytokines (interleukines, tumour necrosis factor (TNF-a)) and the activation of cyclo-oxygenase (**Cox**)-2. Apoptotic neutrophils and crystals are removed by activated macrophages as the inflammation subsides spontaneously.

The American College of Rheumatology criteria for the clinical diagnosis of acute gout are shown in Table 2. The presence of 6 of the 11 criteria has a 95 per cent specificity in differentiating gout from pseudogout (calcium pyrophosphate gout) and an overall sensitivity of 85 per cent. The final confirmation is the demonstration of negatively birefringent sodium urate monohydrate crystals as opposed to the positively birefringent crystals of calcium pyrophosphate.

Although acute gouty arthritis is typically a monoarthritis, some patients have short, recurrent, mild attacks of discomfort and swelling of affected joints. Some 10 per cent of attacks affect more than one joint and typical attacks may provoke migratory attacks in other joints. Multiple, simultaneous attacks are rare. Some attacks are triggered by trauma, intercurrent illness, surgery, alcohol, dietary excess, diuretics, and other medications (see Table 1). An acute septic arthritis is the most important differential diagnosis of acute gouty arthritis.

Chronic tophaceous gout

Large deposits (tophi) containing monosodium urate monohydrate crystals produce firm nodules over affected joints on the extensor surfaces of the fingers, hands, olecranon bursas (commonly bilateral), extensor surfaces of the forearm, Achilles tendon, the helix of the ear, and in the renal parenchyma. Tophi may discharge white chalky material, containing sodium urate monohydrate. They cause the bone erosions and joint destruction with secondary degenerative arthritis seen on radiographs. Tophus formation can be regarded as an attempted, but disordered, healing process in response to the presence of sodium urate monohydrate crystals in tissues.

Treatment of gout

The acute attack

Full doses of any of the non-steroidal anti-inflammatory drugs are effective in terminating attacks of acute gout. Indomethacin is particularly favoured by some clinicians. Colchicine remains a very effective remedy—an initial dose of 1.0 mg followed by 0.5 mg every 6 hours until either the attack subsides or a total dose of 6.0 mg has been achieved, or symptoms of toxicity (nausea, vomiting, and diarrhoea) occur. More frequent doses of colchicine, 0.5 mg every 2 to 3 h, deliberately inducing symptoms of toxicity was previously recommended. This is unnecessary now that the non-steroidal anti-inflammatory drugs are available. Heavy dosage with colchicine can also cause gastrointestinal haemorrhage and favour the development of other severe side-effects, including profuse diarrhoea, rashes, renal and hepatic damage, more rarely peripheral neuropathy, myopathy, and alopecia in the long-term. Intravenous colchicine is no longer recommended.

An attack of acute gout can be effectively terminated by the adrenocorticotropin analogue, tetracosactrin, or by a single intravenous dose of hydrocortisone. Rebound attacks of acute gout tend to occur unless the situation is covered by either colchicine or a non-steroidal anti-inflammatory drug.

Pharmacologically, colchicine disrupts the cellular microtubular architecture in the inflammatory cells. This mode of action gives it the potential to do more widespread damage and presumably underlies its inhibitory effects on mitosis, neutrophil migration, and phagocytosis. Short intensive courses of colchicine should not be repeated at less than 3-day intervals, although lower doses (0.5 mg-2 mg per day) can be used for longer periods, as in the treatment of familial Mediterranean fever.

Interval treatment

Asymptomatic hyperuricaemia should not be treated with urate-lowering drugs unless the patient experiences more than one acute attack of gout per year (Table 3). Allopurinol, a xanthine oxidate inhibitor, is effective in preventing acute gout; it acts by reducing the serum urate concentration to a value below the solubility of sodium urate monohydrate in plasma so that tophaceous deposits are mobilized and healing occurs. This applies to the tophi in bones as well as elsewhere. The drug should be introduced at a low level (e.g. 100–200 mg daily) and increased under cover of either colchicine or a non-steroidal anti-inflammatory drug, which should be continued until the serum urate concentration has stabilized at a normal level. Allopurinol is then continued indefinitely.

Initiating allopurinol without cover may cause attacks of acute gout as the serum urate concentration falls. Moderately severe gout may require as much as 300 to 600 mg of allopurinol daily, and occasionally as much as 700 to 900 mg per day given in divided doses. Between 10 and 20 mg/kg body weight per day is an appropriate

dose for children.

The incidence of adverse reactions to allopurinol is low but they can be severe and occasionally fatal. Reactions include erythema multiforme progressing to the Stevens–Johnson syndrome and toxic epidermal necrolysis, exfoliative dermatitis, vasculitis, interstitial nephritis, eosinophilia, hepatocellular damage, polyneuropathy, bone marrow depression, disturbances of vision and taste, as well as gastroenteropathy. Allopurinol potentiates the effect of coumarin anticoagulants (for example, warfarin), azathioprine, and 6-mercaptopurine, and predisposes to an ampicillin or amoxicillin rash. At high dosage and in the presence of greatly increased purine synthesis it may cause xanthine and oxipurinol urinary stones. There is also increased risk of toxicity with captopril (especially in the presence of renal failure) and with ciclosporin.

Much of the overall toxicity of allopurinol is due to the metabolite oxipurinol, which has a much longer half-life *in vivo* than the parent compound. Special care is necessary in the presence of renal failure, when a dose of 100 to 150 mg is usually sufficient. Patients with hyperuricaemia due to renal failure rarely develop gout, possibly due to immunoparesis.

Patients in whom allopurinol produces adverse reactions

Patients for whom the treatment of hyperuricaemia and gout is essential and who have developed severe adverse reactions to allopurinol present a special problem, especially if they have impaired overall renal function. The uricosuric drugs sulfinpyrazone, probenecid, and benzbromarone, together with a sufficiently high fluid intake to provide a measured urine output of at least 3 litres per 24 h and alkalization of the urine with sodium or potassium bicarbonate or sodium or potassium citrate, represent an approach to this problem, but may be inappropriate in the overall clinical context, for example in patients with cardiac or renal failure. Only sulfinpyrazone is readily available in the United Kingdom. Uricosuric drugs may be inefficient in the presence of renal failure and are contraindicated in the presence of uric acid urinary stones.

The use of recombinant uricase—either in its unmodified form or linked to polyethylene glycol (PEG) in order to reduce its immunogenicity—remains experimental, and is unlikely to be applied except in patients at risk of developing acute hyperuricaemic nephropathy and who cannot be given allopurinol.

The uricosuric agent benzbromarone is sometimes effective in patients with renal failure when other uricosuric agents have lost their efficacy. The use of oxipurinol (in low dosage) has also been proposed. Protocols are also available for the desensitization of patients who have experienced adverse reactions to allopurinol, and in whom the risk of uric acid stone formation with the potential for further reduction of renal function presents a problem.

Asymptomatic hyperuricaemia

Routine biochemical screening frequently identifies patients with hyperuricaemia. Guidance on their management is given in Table 3.

Acute uric acid nephropathy

This complicates the treatment of widespread malignant disease, particularly chemo- and/or radiotherapy of leukaemias and lymphomas. The nephropathy is of multifactorial origin and may form part of the acute tumour-lysis syndrome with accompanying tubular necrosis. These patients are usually underhydrated and acidotic and have high rates of uric acid production from nucleoprotein degradation in the apoptotic tumours. Acute uric acid nephropathy has occasionally been reported after extremely severe muscular exercise, after severe epileptic seizures, and in patients with gout and grossly increased rates of *de novo* purine synthesis.

The renal lesion is the intratubular precipitation of uric acid crystals. In addition, the renal pelvis and ureters may also be blocked by crystal aggregates and/or uric acid stones. Acute uric acid nephropathy can be avoided by giving allopurinol for several days before starting chemotherapy or radiotherapy. The condition presents as acute oliguric renal failure. Imaging techniques should be used to exclude the presence of bilateral ureteric obstruction by radiotranslucent uric acid stones. Treatment is by:

1. induction of an alkaline diuresis;
2. haemo- or peritoneal dialysis or haemofiltration;
3. percutaneous nephrostomy and/or ureteric catheterization may be needed if there is an element of postrenal obstruction due to impacted aggregates of sodium urate crystals or uric acid stones;
4. disruption or removal of impacted stones.

Chronic sodium urate nephropathy

Between 20 and 30 per cent of patients with untreated chronic tophaceous gout die from renal failure. These patients form an identifiable subgroup of the gouty population and an autosomal dominant inheritance is sometimes clearly apparent (Fig. 3). The term 'familial juvenile gouty nephropathy' is sometimes used for patients presenting in early life. Environmental factors exacerbate this hereditary predisposition. Such patients must be differentiated from another group of gout patients (20–30 per cent) with mild intermittent proteinuria and a good prognosis. Significant renal disease due to sodium urate deposition is very rare in asymptomatic hyperuricaemia. Patients with chronic sodium urate nephropathy have shrunken kidneys containing interstitial monosodium urate microtophi and show segmental destruction of the renal parenchyma due to tubular blockage by aggregates of uric acid crystals (microcalculi). These areas of segmental destruction have been referred to, inappropriately, as 'uric acid infarcts'.

Polycystic renal disease

Hyperuricaemia and gout may precede the onset of renal failure in patients with polycystic renal disease and about one-third of patients with polycystic renal disease develop gouty arthritis. This may be due to abnormal renal tubular handling of urate. A similar mechanism may operate in patients with medullary sponge kidney disease.

Ethanol and hyperuricaemia

Ethanol is oxidized to acetaldehyde by the liver. This raises the ratio of **NADH**:**NAD** (reduced nicotinamide–adenine dinucleotide:nicotinamide-adenine dinucleotide), which in turn promotes the reduction of pyruvate to lactate in the hepatocytes. Lactate competes with urate in the renal tubular excretory mechanisms and thereby promotes urate retention. There is also an element of starvation ketoacidosis in chronic alcoholics, with acetoacetate and betahydroxybutyrate also competing for the renal tubular excretory mechanisms which subserve urate tubular secretion. In addition, there is increased urate production associated with ethanol intake: first due to the high purine content of some alcoholic beverages (for example, beer) and second, the metabolism of alcohol involves increased dephosphorylation and degradation of adenine nucleotides in the liver. The free adenine produced is further metabolized to urate.

Uric acid urolithiasis

Pure uric acid stones account for 5 per cent of all urinary stones in patients in the United Kingdom. There is a much higher incidence elsewhere, for example in the Middle East. Uric acid urolithiasis occurs in 10 per cent of patients with gout. In Israel, about 40 per cent urinary calculi are composed of uric acid and 75 per cent of patients with primary gout develop renal calculus disease. Uric acid stones are more common in secondary than in primary gout and are sometimes associated with an impaired ability to alkalinize the urine. Ileostomy predisposes to uric acid urolithiasis because of (1) chronic bicarbonate loss, which leads to a persistent acidification of the urine and (2) a concentrated urine due to excessive water loss. Urinary uric acid concentrations close to, or greater than, those at which spontaneous crystallization begins are frequent in these circumstances. The genetic causes of uric acid urolithiasis are rare: (1) **HPRT** deficiency; (2) phosphoribosylpyrophosphate synthetase (**PRPPS**) superactivity; and (3) congenital renal hypouricaemia (congenital failure of the renal tubular reabsorption of urate). Renal hypouricaemia may be due to renal tubular damage by other genetic diseases or by toxic damage (Table 4) and this may be associated with other features of the Fanconi syndrome.

The urinary uric acid concentration is the main determinant of uric acid stone formation. The concentration depends on the state of hydration, the rate of purine *de novo* synthesis, the rate of metabolic turnover of purine compounds, the dietary intake of purines and alcohol, and the action of uricosuric drugs (for example,

sulfinpyrazone). Calcium oxalate stone formation is increased 30-fold in patients with gout and hyperuricosuria is common in non-gouty stone formers. Uric acid micro crystals may act as epitaxial nucleation sites for calcium oxalate crystallization. It is also possible that colloidal uric acid adsorbs urinary glycosaminoglycan inhibitors of crystallization and crystal growth.

Uric acid stone disease is treated by hydration to maintain a urine volume of at least 3 litres per 24 h, alkalization of the urine, and allopurinol if there is hyperuricosuria. The use of sodium and potassium salts for alkalization has to be carefully reviewed in the light of concurrent diseases, particularly impaired renal and cardiac function. The standard imaging techniques (particularly ultrasonography) are required for the diagnosis of these radiotranslucent stones. They can be fragmented or removed by standard procedures.

Congenital renal hypouricaemia and uric acid stones

Reduced net tubular reabsorption of urate occurs either as an isolated renal tubular reabsorption defect due to mutations in the gene directing the synthesis of the putative urate carrier protein, or in association with other inherited and acquired renal tubule transport defects (Table 4).

Isolated reduced net tubular reabsorption of urate (hereditary renal hypouricaemia) is inherited in an autosomal recessive manner. The hyperuricosuria may amount to 1000 mg (5.9 mmol) per 24 h in the homozygote. Lesser degrees of hyperuricosuria occur in heterozygotes. About 30 per cent of the homozygotes have an associated hypercalciuria. Uric acid urolithiasis occurs in about 25 per cent of the homozygotes, most commonly in patients with combined hyperuricosuria and hypercalciuria. Only two patients with hereditary renal hypouricaemia were found by searching the clinical biochemical data on 47 420 patients in a general hospital. The causes of hypouricaemia are summarized in Table 4.

The Lesch–Nyhan syndrome and variants

The Lesch–Nyhan syndrome results from mutations in the gene that directs the synthesis of hypoxanthine–guanine phosphoribosyltransferase (**HPRT**), an enzyme which normally catalyses the salvage of hypoxanthine and guanine to inosinic and guanylic acids (inosine monophosphate, **IMP**; and guanosine monophosphate, **GMP**), respectively, as shown in Fig. 1. The clinical spectrum extends from hyperuricaemia alone to hyperuricaemia with profound neurological and behavioural dysfunction. The biochemistry and molecular genetics of this disorder have been studied extensively. A recent survey of a database of 271 cases showed that mutation analysis does not provide precise information for predicting disease severity, but that it is a valuable tool for genetic counselling in terms of confirming diagnosis, the identification of carriers, and for prenatal diagnosis.

The clinical features of the most severely affected patients who are correctly referred to as having the 'Lesch–Nyhan syndrome' or as having 'complete or virtually complete HPRT deficiency', are summarized in Table 5.

Infants affected by HRPT deficiency have a lower than average birth weight, indicating some degree of intrauterine growth retardation (Fig. 4). The first clinical sign may be the presence of red grit (uric acid crystals with adsorbed urinary pigments) on the nappy. Affected infants are hypotonic from birth, although this is frequently not remarked upon before poor head control becomes apparent at the age of about 3 months.

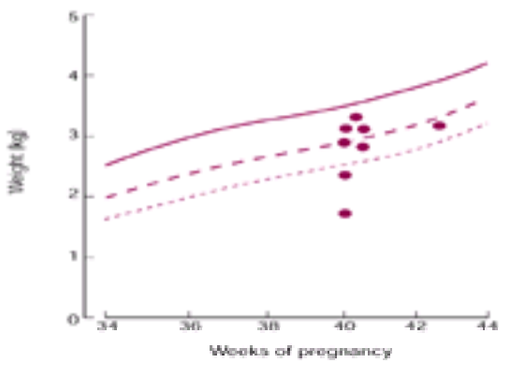

Fig. 4 Birth weight in eight boys who later developed the Lesch–Nyhan syndrome: the 50th (bold), 10th, and 3rd (interrupted) centiles are shown as lines. (Reproduced from Watts *et al.* 1987, with kind permission from Kluwer Academic Publishers.)

Postnatal growth, which becomes more marked after the second year of life, is also subnormal (Fig. 5) as indicated by sequential measurement of body weight, accurate assessment of body length being impossible due to the dystonic posturing. The overall pattern of weight growth follows centile lines for the first 2 years of life and thereafter slows to about 1 kg per year, or about half normal, a pubertal growth spurt is not observed. Head growth and bone development are less affected than weight. The poor weight gain cannot be attributed to either renal failure or malnutrition.

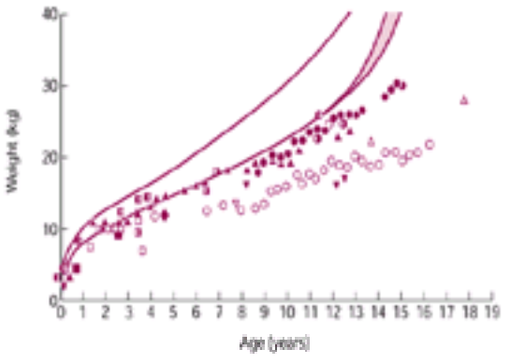

Fig. 5 Patterns of growth in weight of 13 boys with the Lesch–Nyhan syndrome: each patient is shown by a different symbol. The 50th and 3rd centiles are shown. (Reproduced from Watts, *et al.* 1987, with kind permission from Kluwer Academic Publishers.)

Torsion dystonia, with its two components of abnormal posturing and episodic rigidity, is superimposed on the basic hypotonia that is present between the dystonic episodes. Severe dysarthria is associated with dyskinesia of the face, mouth, pharynx, and the larynx, which greatly limits communication and even the ability to point accurately, thus leading to great frustration. The self-injurious behaviour and dyskinesia are eliminated or much reduced when the child is concentrating on a self-selected activity, such as watching an interesting television programme. Self-injury and dyskinesia are exacerbated by excitement such as the arrival of a visitor, fear, frustration, and unsuccessful attempts at volitional motor activity. The children also appear to be aware of the value of this behaviour as an attention-seeking manoeuvre, and sometimes appear to use it in a manipulative manner. This mixture of involuntary and volitional involuntary abnormal motor activity with an apparent interplay of unconscious and consciously meditated behaviour patterns should be common ground for behavioural scientists, neurochemists, and neuropharmacologists.

Although mental handicap has been stressed as a feature of the Lesch–Nyhan syndrome, it is of inconstant severity, and is neither marked nor specific. The apparent degree of mental handicap may be affected by the extensive disorder of expressive motor functions that exceeds the comprehension defect, by the lack of basic social and educational opportunities, and by the lack of intelligence tests for older children who have lacked these opportunities. However, for whatever combination of reasons, there does appear to be a decline of intellect from the age of 8 to 10 years.

Self-injurious behaviour usually begins at about 2 years of age. Its severity and the ingenuity with which the patients exploit new ways of self-injury exceeds that encountered in any other clinical situation. It is not a constant feature and some patients never show it; in the majority its severity waxes and wanes. Self-injury can produce very severe damage, such as complete destruction of the lower lip or traumatic amputation of a fingertip. The patients feel pain normally and are aware of their compulsion; they are afraid of it but are unable to control it. Nyhan and his colleagues consider it to be the hallmark of complete HPRT deficiency, as opposed to those patients with some residual enzyme activity (which may or may not be measurable in erythrocyte lysates).

There are no structural or ultrastructural changes in the brain as judged by light and electron microscopy. Computed tomography and electroencephalography also show no abnormality during life. MRI and PET scanning have not yet been applied to this problem.

Both the purine *de novo* synthesis and the HPRT-catalysed purine salvage pathways are present in all parts of the normal brain. HPRT activity is absent, but the *de novo* synthesis pathway remains active in patients with the Lesch–Nyhan syndrome. It has been suggested that the bone marrow and brain have particular requirements for the purine salvage pathway and that HPRT deficiency might constrain brain development. If this is so, it is not explained by particularly low activity of purine *de novo* synthesis activity in the brain. Indeed, demonstrable structural and ultrastructural changes in patients with the Lesch–Nyhan syndrome are also lacking, suggesting that the inability to salvage hypoxanthine and guanine in the Lesch–Nyhan syndrome causes a 'functional' aberration. Such a derangement could derive from either a postulated postsynaptic transmitter function for cyclic GMP (**cGMP**), or a related compound, or a consequential effect on the availability of synaptic neurotransmitters. So far, further studies have not supported the cGMP postsynaptic neurotransmitter hypothesis. The levels of HPRT activities are approximately uniform in a normal human brain. Purine salvage as well as purine *de novo* synthesis activity is also uniformly distributed in the different gross anatomical regions of the rat brain.

Evidence has been advanced for some aspects of the Lesch–Nyhan phenotype being related to dysfunction of the small central, but widely projecting, aminergic pathways involved in learning. Thus it has been suggested that self-injurious behaviour in the Lesch–Nyhan syndrome is due to an imbalance between the activities of catecholaminergic neurones and 5-hydroxytryptaminergic neurones. The catecholaminergic neurones are largely concerned with learning by reward, and the 5-hydroxy-tryptaminergic pathways with learning by punishment. Patients with the Lesch–Nyhan syndrome are insensitive to punishing stimuli and do not learn when such stimuli are used to reinforce the desired behaviour, which in this case is non-self injury. The ability to learn from rewarding stimuli is impaired. Psychotherapeutic techniques that are effective in eliminating self-injurious behaviour in other situations fail in patients with the Lesch–Nyhan syndrome. Thus, although the self-injurious behaviour in the Lesch–Nyhan syndrome could be modified by a programme of positive reinforcement of non-self injury and 'time out', this has proved difficult to achieve in the long term. The reinforcement strategy was found to be unsuitable for use at home because it involved apparently ignoring the self-injury and only paying attention to the child during periods of non-self injury. This was misinterpreted by friends and relations as unkindness or indifference.

The present view is that the neurological manifestations are brought about by a neurotransmitter imbalance (probably mainly in the basal ganglia). This imbalance is possibly due to a deficient supply of metabolic energy resulting from the non-salvage of hypoxanthine and guanine, thus causing a deficiency of adenine nucleotides that provide energy for short bursts of neurotransmitter synthesis.

Failure of pubertal development and testicular atrophy in HPRT deficiency are attributed to an inadequate supply of purine nucleotides to meet the increased metabolic energy requirement in the testis at this time. A similar inability to meet energy requirements may underlie the neurological manifestations. A partial defect in adrenocortical 11b-hydroxylation of steroids is demonstrable in patients with the Lesch–Nyhan syndrome after ACTH stimulation, and is thought to be linked with a failure to modulate mitochondrial function for this hydroxylation due to a deficiency of purine nucleotides.

Patients with Lesch–Nyhan syndrome whose hyperuricaemia has been controlled and who have not suffered renal damage, die in their teenage years, often with postmortem evidence of gastric aspiration during sleep. Less severe degrees of HPRT deficiency lead to the X-linked recessive hyperuricaemia gout and urolithiasis syndrome, which may also be associated with minor neurological abnormalities.

Treatment

Sufficient allopurinol should be administered to reduce the plasma urate and urine uric acid concentrations to normal in order to prevent gouty arthritis, urate nephropathy, and renal calculi. Relatively large doses of allopurinol are needed and the patient should be kept well hydrated to minimize the risk of xanthine and/or oxipurinol (the metabolic oxidation product of allopurinol) stones developing. Both types of stone are, like uric acid stones, radiotranslucent. Allopurinol treatment from birth does not prevent the behavioural phenotype. All therapeutic attempts at neuropharmacological manipulation have been unsuccessful.

Dental extraction, physical restraints with splints and bandages, and strapping the patient into a specially designed padded wheelchair fitted with a padded firm head support to prevent cervical spine injury during violent opisthotonic spasms, are usually needed to limit the effects of compulsive self-mutilation.

Children whose restraints have been temporarily released ask or indicate their wish for the bandages, straps, etc. to be replaced so that they are less able to damage themselves. Every effort should be made to exploit these patients' intellect and to keep them in a stimulating environment.

Clinical genetic aspects

The Lesch–Nyhan syndrome and its variants are inherited in a sex-linked recessive manner with no clinical manifestations in the female carriers. However, subtle alterations in purine metabolism, with small increases in the rates of *de novo* purine synthesis and increased uric acid excretion and occasionally mild asymptomatic hyperuricaemia, have been reported in females. Affected male hemizygotes are identified biochemically by HPRT assays on red cell haemolysates, the lack of HPRT being accompanied by an elevated level of phosphoribosylpyrophosphate. Genomic analysis is also possible. Carrier females are identified by the demonstration of mosaicism with respect to *HPRT*+ and *HPRT*– hair roots due to random inactivation of the X-chromosome, the hair roots being clonal in origin. Autoradiographic techniques can be used to demonstrate two cell populations (*HPRT*+ and $HPRT^{-}$) in fibroblast cultures.

Early prenatal diagnosis is possible using chorionic villus samples obtained during the ninth week of pregnancy, this permits elective abortion of an affected fetus before the end of the first trimester of pregnancy. *In vitro* fertilization with enzymatic assay on a cell removed at the four-cell stage to ensure that only unaffected embryos are implanted is theoretically possible.

Phosphoribosylpyrophosphate synthetase superactivity

This enzyme catalyses the production of phosphoribosylpyrophosphate, which is required for the first specific and rate-limiting reaction on the *de novo* pathway of purine synthesis. It is subject to feedback inhibition by purine nucleotides. The known mutations in the gene regulating the synthesis of phosphoribosylpyrophosphate synthetase diminish its sensitivity to this feedback inhibition, thereby leading to hyperuricaemia, hyperuricosuria, and gout. The condition is inherited in an X-linked recessive fashion.

Affected males develop uric acid lithiasis or gouty arthritis in childhood or early adult life. Hyperuricaemia is often severe and in the range 0.5 to 1 mmol/l, with uric acid excretion of 5 to 15 mmol/24 h. Heterozygotes remain asymptomatic, although some degree of increased purine synthesis *de novo* has been demonstrated.

In some families, the disorder presents in childhood with associated neurological features such as motor and mental retardation, ataxia, deafness, hypotonia, disturbed speech, and the development of polyneuropathy, intracerebral calcifications, and dysmorphic facial features. The constellation of associated disorders varies in different families.

Heterozygotes can be identified by studies in cultured skin fibroblasts. Amniocentesis, prenatal diagnosis, and preventive abortion are not justified in this condition unless one of the unusually severe phenotypes is known to be segregating in the family. The hyperuricaemia, primary purine overproduction, and uricosuria can be well controlled with allopurinol.

2,8-Dihydroxyadeninuria

These patients lack adenine phosphoribosyltransferase activity, adenine accumulates behind the metabolic block and is oxidized under the catalytic influence of

xanthine oxidase to the very insoluble compound, 2,8-dihydroxyadenine. This compound is excreted in the urine along with adenine itself, where it forms radiotranslucent stones that are white or pale fawn in colour. These rough and friable calculi have, in the past, been widely misdiagnosed as uric acid stones because 2,8-dihydroxyadenine reacts as if it were uric acid in colorimetric assays. The use of enzymatic uric acid assays has obviated this confusion.

Adenine phosphoribosyltransferase deficiency has an autosomal recessive pattern of inheritance and is clinically silent in heterozygotes. There are two subtypes (I and II). Type I patients have no detectable enzyme activity, being homozygotes or compound heterozygotes for null alleles. Type II patients have between 5 and 25 per cent residual enzyme activity. Whereas type I patients are encountered in many racial groups, the type II subtype has so far only been identified in the Japanese population. Heterozygotes for type I and type II can only be distinguished from one another by enzyme assays on extracts from cultured peripheral blood lymphocytes, both types show no activity in the red cell lysates that are generally used diagnostically.

Because of the extremely low solubility of 2,8-dihydroxyadenine in renal tubule fluid and urine, this condition often presents in early life. Severe obstructive uropathy and renal failure may occur in infancy.

Treatment is by hydration and xanthine oxidase inhibition with allopurinol, and with standard measures to disrupt or remove the stones and to manage urinary infections and renal failure.

Type I lycogenosis

Type I glycogenosis (hereditary glucose 6-phosphatase deficiency) is associated with hyperuricaemia. This is due to chronic hyperlacticacidaemia which leads to urate retention, and to increased urate production due to reduced serum phosphate concentrations. The phosphate ion inhibits AMP deaminase: the enzyme that catalyses the rate-limiting step in the metabolic pathway for the conversion of adenine nucleotides to uric acid. Thus, hypophosphataemia increases adenine nucleotide degradation to uric acid and adds to the accumulating urate burden; gouty arthritis may develop in childhood.

Treatment is by maintaining the blood glucose level in the normal range with frequent small meals and intragastric glucose infusion at night. Gout is treated in the standard manner with colchicine and/or non-steroidal anti-inflammatory drugs for the acute attacks, and with long-term allopurinol.

Xanthinuria

Xanthine stones occur in patients with xanthinuria (congenital xanthine oxidase/reductase deficiency) and occasionally in those who are being treated with the xanthine oxidase inhibitor, allopurinol. The latter is particularly likely in patients with accelerated purine *de novo* synthesis (for example, in patients with the Lesch–Nyhan syndrome). Xanthinuria is inherited in an autosomal recessive manner, and hypoxanthine and xanthine accumulate behind the metabolic block. The plasma urate concentration and urine uric acid excretion are less than about 0.06 mmol/l (1.0 mg/dl) and 0.30 mmol/24 h (50 mg/24 h), respectively, when the patient is taking an unrestricted diet. A search of general hospital clinical data on 47 420 unselected patients yielded no cases of xanthinuria. The plasma and urine 'oxypurines' (hypoxanthine plus xanthine) concentrations are characteristically elevated. Normal subjects have plasma levels between 0.00 and 0.15 mmol/l (0.00–0.25 mg/dl) and urine levels of 0.07 to 0.13 mmol/24h (11–22 mg/24 h); patients with xanthinuria have plasma levels between 0.03 and 0.05 mmol/l (0.05–0.9 mg/dl) and urine levels of 0.60 to 3.5 mmol/24 h (100–600 mg/24 h). Xanthine accounts for 60 to 90 per cent of the total xanthine plus hypoxanthine excreted, presumably reflecting the more active metabolic turnover of hypoxanthine and its efficient salvage by hypoxanthine phosphoribosyltransferase. Hypoxanthine and xanthine are mainly derived from adenine and guanine nucleotides, respectively (see Fig. 1). Hypoxanthine has a relatively high solubility and causes no problems.

At any age, about one-third of cases present with radiotranslucent xanthine stones. These stones are usually smooth, soft, and yellow–brown. Xanthinuric myopathy is a rare complication.

Xanthine stones also occur when there is a combined deficiency of the three molybdoflavoprotein enzymes—xanthine oxidase, sulphite oxidase, and aldehyde oxidase—because of defective molybdopterin cofactor synthesis. The clinical picture in these patients is overshadowed by the sulphite oxidase deficiency that produces severe brain damage and dislocation of the ocular lenses. Another subgoup of patients with xanthinuria only lack xanthine oxidase and aldehyde oxidase activity. These patients present with xanthine stones and are detected by their inability to convert allopurinol to oxipurinol, a reaction normally catalysed by aldehyde oxidase.

Adenylosuccinase deficiency

Adenylosuccinase (adenylate succinate lyase) catalyses the eighth step on the 10-step *de novo* purine synthesis pathway and the second step on one of the purine nucleotide interconversion pathways, the formation of ATP from IMP.

The patients present in infancy with severe psychomotor retardation, autism, and axial hypotonia with normal tendon reflexes. Self-mutilation has been recorded in some cases and cerebellar hypoplasia is present on computed tomographic (**CT**) scanning.

The presence of aspartic acid and glycine in body fluids suggests the diagnosis, and this is confirmed by finding succinyladenosine and succinylaminoimidazole carboxamide riboside in plasma, cerebrospinal fluid, and urine. There is gross purine overproduction with high levels of nucleosides in the urine. Urine and plasma uric acid levels are normal. Partial enzyme deficiencies have been demonstrated in liver, kidney, muscle, lymphocytes, and fibroblasts.

Adenylosuccinase deficiency is inherited as an autosomal recessive. The growth retardation has been improved by adenine (10 mg/day) and allopurinol. The latter promotes purine conservation by blocking hypoxanthine oxidation to xanthine and uric acid, and prevents the oxidation of administered adenine to 2,8-dihydroxyadenine.

Myoadenylate deaminase deficiency

Myoadenylate deaminase is the muscle-specific isoenzyme of adenylate deaminase which catalyses the deamination of adenylic acid to inosinic acid during muscle contraction. This reaction is necessary for normal muscle function. Myoadenylate deaminase deficiency may be congenital, due to a mutation in the gene directing the synthesis of the protein, or associated with a wide range of muscle diseases including the muscular dystrophies, polymyositis, and dermatomyositis.

Patients with congenital myoadenylate deaminase deficiency present at any age, including early childhood, with a syndrome of muscle weakness and muscle cramps during and after exertion. There is some decrease in muscle mass, some hypotonia, and a little muscle weakness. There may be a modest rise in plasma creatine phosphokinase levels and non-specific electromyographic changes. The lack of ammonia in the venous outflow from the affected muscles during exercise and the enzyme deficiency can be demonstrated histochemically. The pattern of inheritance is autosomal recessive, not all homozygotes have clinical symptoms and heterozygous carriers are clinically silent. A single mutant allele contains a non-sense mutation that leads to the production of a severely truncated enzyme. The acquired disorder may be due to the coincidental disease arising in a patient whose inherited myoadenylate deaminase deficiency would be otherwise silent. Genetic testing for the mutant allele can be utilized to determine whether congenital myoadenylate deaminase could be contributing to the patient's clinical presentation.

Oral ribose (2–60 g/day, or taking a dose before vigorous exercise) has been reported to produce symptomatic improvement. The risk of rhabdomyolysis has led some authors to recommend the avoidance of vigorous exercise, myoglobinuria following strenuous exercise having been reported in a few cases. Such advice is appropriate if exertion-related myoglobinuria has occurred or been suspected.

Inborn errors of purine metabolism and immunodeficiency

Adenosine deaminase (**ADA**) and purine nucleoside phosphorylase (**PNP**) catalyse sequential steps in the metabolism of purine ribonucleosides and deoxyribonucleosides (Fig. 6). These enzymes are highly expressed in lymphoid cells and their deficiency, which causes the lymphotoxic substrates 2'-deoxyadenosine (**dAdo**) and 2'-deoxyguanosine (**dGuo**) to accumulate (see Fig. 6), leads to lymphopenia and immunodeficiency.

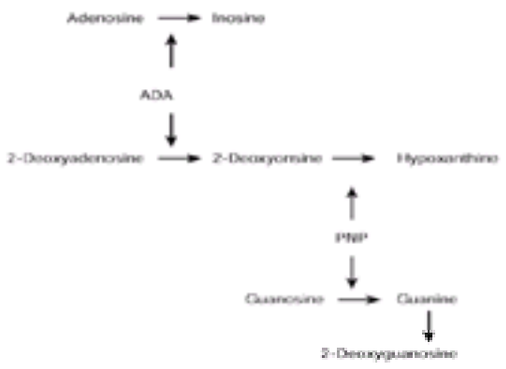

Fig. 6 Metabolic steps catalysed by adenosine deaminase (ADA) and by purine nucleoside phosphorylase (PNA).

Most patients with ADA deficiency lack both cell- (T cell) and humoral (B cell)-mediated immunity, resulting in severe combined immunodeficiency disease (**SCID**). Although PNP deficiency causes defective T-cell mediated immunity, these patients may possess either normal, hyperactive, or reduced humoral immunity. Most patients with these enzyme deficiencies present in infancy or early childhood with severe infections caused by pathogens or opportunistic organisms. About 50 per cent of patients with SCID have X-linked agammaglobulinaemia (Bruton's disease), a disease that is unrelated to ADA and PNP deficiencies and which displays an autosomal recessive pattern of inheritance.

Adenosine deaminase deficiency

About 85 per cent of patients with ADA deficiency are infants with SCID. In all patients with SCID, ADA deficiency accounts for a minority, possibly about 15 per cent. Although adenosine deaminase deficiency classically presents during infancy, a minority of patients have a clinically less severe variant and are diagnosed later. The prevalence of ADA deficiency has been estimated at between less than 1 in 10^6 and 1 in 2×10^5 live births.

ADA deficiency is inherited in an autosomal recessive fashion, the gene having been mapped to chromosome 20q13.11. The diagnosis is made by measuring ADA activity in erythrocytes. Heterozygote detection and prenatal diagnosis are best done using molecular probes for the ADA gene.

In addition to immunoparesis, about one-third of cases have multiple skeletal abnormalities, including fraying of the long bones, abnormally thick growth-arrest lines, and chondro-osseous dysphasia at the costochondral junctions. Other occasionally reported comorbidities are renal tubular acidosis, choreoathetosis, spasticity, and fine sparse hair.

The prognosis for patients with untreated adenosine deaminase-deficient SCID is very poor, with death due to multiple recurrent infections during the first year of life.

Adenosine and dAdo, derived from the breakdown of DNA due to cell death, accumulate proximal to the metabolic block—dAdo is the primary lymphotoxic precursor in adenosine deaminase deficiency, elevated levels of which are present in plasma and urine. Erythrocytes contain markedly raised levels of deoxyadenosine triphosphate (**dATP**) and reduced *S*-adenosylhomocysteine (**AdoHcy**) hydrolase activity due to inactivation by dAdo; erythrocyte ATP is reduced. The level of dATP in erythrocytes correlates with clinical expression and with the level of ADA activity expressed in *Escherichia coli* by mutant ADA alleles.

There are several mechanisms by which adenosine deaminase deficiency can impair immune function. Thus, accumulation of dATP can induce apoptosis in lymphoid cells, which may be related to dATP-induced inhibition of ribonucleotide reductase blocking DNA replication in dividing cells, and to dATP-induced DNA strand breaks in non-dividing lymphocytes. dATP also activates the protease (caspase 9) involved in apoptosis. AdoHcy also blocks *S*-adenosylmethionine (**AdoMet**)-mediated transmethylation reactions. The formation of dATP from dAdo activates IMP dephosphorylation, thereby leading to the depletion of cellular ATP. It has also been suggested that lymphocyte function may be impaired by aberrant signal transduction mediated by Ado acting through G-protein-associated receptors, or from an altered co-stimulatory function of T-cell associated ADA-complexing protein CD26/dipeptidyl peptidase IV.

Treatment

This is by bone marrow transplantation from a histocompatible donor. Repeated blood transfusions can provide temporary benefit, although repeated transfusion leads to iron overload. More sustained clinical improvement follows the weekly or twice-weekly administration of polyethylene glycol-modified bovine adenosine deaminase. The use of ADA-loaded erythrocytes membranes is also being explored.

Transplantation of T-cell-depleted marrow from an HLA-haploidentical donor has been tried, but it is associated with greater morbidity and is less effective than bone marrow transplantation in restoring immune function.

The *ex vivo* retrovirus-mediated transfer of ADA cDNA is the first attempt at somatic-cell gene therapy in humans. The efficacy of transducing stem cells has been low, but persistence of the vector myeloid cells and T lymphocytes has been demonstrated. The long-term evaluation of this approach is still awaited.

Purine nucleoside phosphorylase deficiency

PNP deficiency occurs less frequently than ADA deficiency. In addition to the clinical results of immunoparesis, more than 50 per cent of these patients have neurological abnormalities including disorders of muscle tone, delayed motor and intellectual development, ataxias, tremors, spastic tetraparesis, behavioural difficulties, and varying degrees of mental handicap. Autoimmune haemolytic anaemia and megaloblastic bone marrow have been occasional associations.

There appears to be a particular susceptibility to virus infections such as varicella, vaccinia, and cytomegalovirus. The tonsils and the thymus are small or absent and the lymph nodes are deficient in the thymus-dependent areas. Circulating lymphocyte counts are usually very low, with a low percentage of T lymphocytes and depressed or absent responsiveness to mitogen-induced transformation. Serum immunoglobulin levels and antibody responses to pneumococcal polysaccharide and keyhole limpet haemocyanin are typically increased in children with PNP deficiency, and the occasional finding of monoclonal IgG paraprotein strongly suggests that the changes in antibody production are the result of T-cell defects.

PNP deficiency is associated with the accumulation and excretion of dGuo and deoxyinosine, as well as guanosine and inosine. Paradoxically, there is massive purine overproduction and excretion, although all patients are severely hypouricaemic. Erythrocyte concentrations of dGTP are markedly raised in PNP-deficient cells. T cells, but not B cells, appear to be particularly susceptible to dGuo toxicity, probably as a result of the accumulation of dGTP, inhibition of ribonucleotide reductase, impairment of DNA synthesis, and, eventually, cell death.

The prognosis in children with PNP deficiency is often much better than that in adenosine deaminase deficiency. Since some children have remained healthy and free from viral infection until the age of 6 years, high-risk procedures such as bone marrow transplantation are currently not thought to be justified in all cases. Conservative treatment with gammaglobulin replacement and attempts at enzyme replacement with red cell transfusions in children with recurrent infections are the current approach to management.

Purine 5'-nucleotidase deficiency

Deficiency of the ecto enzyme 5'-nucleotidase is found in some patients with X-linked and 'acquired' adult-onset hypogammaglobulinaemia. There is no evidence that the enzyme deficiency causes the immunodeficiency in either case. It is currently thought much more likely to simply reflect an arrested stage of lymphocyte development in these patients.

Other disorders of purine metabolism

There are two unrelated conditions: (1) a regulatory mutation in liver adenylic deaminase as a cause of uric acid overproduction and gout in a single patient; and (2) erythrocyte adenylic acid deaminase deficiency in Japanese and Chinese peoples, which has no clinical phenotype.

Disorders of pyrimidine metabolism

The pathways of pyrimidine biosynthesis interconversion and degradation are shown in Fig. 7.

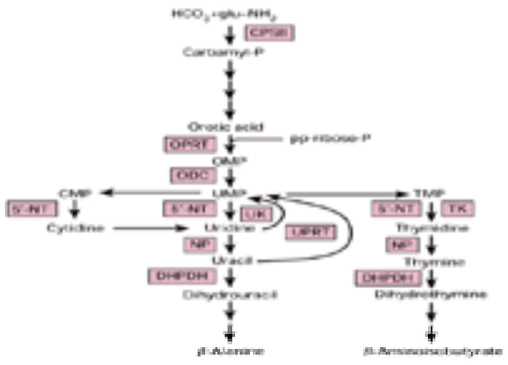

Fig. 7 Pathways of pyrimidine metabolism in humans. CPSH, carbamyl phosphate synthetase II; OPRT, orotate phosphoribosyltransferase; ODC, orotidine decarboxylase (OPRT + ODC from UMP synthase; 5'-NT, pyrimidine 5'-nucleotidase; NP, pyrimidine nucleoside phosphorylase; DHPD, dihydropyrimidine dehydrogenase; UK, uridine kinase; UPRT, uracil phosphoribosyltransferase; TK, thymidine kinase.

The *de novo* synthesis of pyrimidine nucleotides involves a series of six reactions beginning with the formation of carbamyl phosphate and concluding with orotidylic acid (**OMP**), which then undergoes a series of interconversion and salvage reactions as summarized in Fig. 7. The first three steps on the *de novo* synthesis pathway are catalysed by the multifunctional protein that encompasses carbamyl phosphate synthetase, aspartate transaminase, and dihydro-orotase. The fourth step is catalysed by dihydro-orotate dehydrogenase. The fifth and sixth steps are catalysed by the bifunctional protein encoding orotate phosphoribosyltransferase (**ORPT**) and orotidine-5'-monophosphate dehydrogenase (**OMPD**) activities. The pyrimidines are degraded to b-alanine and b-aminobutyrate (Fig. 7).

The inherited disorders of pyrimidine metabolism are much less common, or possibly much less easily recognized, than disorders of purine metabolism.

Orotic aciduria

Orotic aciduria is due to a deficiency of the bifunctional protein that encodes both OMP dehydrogenase and OPRT activities. There is massive overproduction of orotic acid due to loss of feed-back inhibition of carbamyl phosphate synthase, which is the first and rate-limiting step on the metabolic pathway.

Orotic aciduria presents during infancy with severe megaloblastic anaemia, orotic acid crystalluria, and, occasionally, radiotranslucent orotic acid urinary stones. Cardiac malformations, mild intellectual impairment, and strabismus have been reported. Orotic aciduria is inherited as an autosomal recessive.

Enzyme assays on erythrocyte lysates show either low levels of OPRT and OMPD (type 1 orotic aciduria) or a deficiency of ODC only (type 2 orotic aciduria). Administration of uridine (100–150 mg/kg per day), which is converted to UMP (Fig. 7), produces a prompt haematological response. Treatment needs to be started as soon as the diagnosis is made during infancy in order to minimize the possibility of persistent neurological deficits.

Some degree of orotic aciduria has been found in urea cycle defects, lysinuric protein intolerance, PNP-deficiency, normal pregnancy, and during allopurinol administration.

Pyrimidine 5'-nucleotidase deficiency

This autosomal recessive disorder leads to non-spherocytic haemolytic anaemia. Uridine triphosphate (**UTP**) and cytidine triphosphate (**CTP**) accumulate in the red cells, which show basophilic stippling. The enzyme is assayed in erythrocytes and activities between 0 and 30 per cent of normal have been reported. There is no effective treatment. Lead poisoning can also be associated with acquired erythrocyte pyrimidine 5_-nucleotidase deficiency.

Pyrimidine 5'-nucleotidase superactivity

Pyrimidine 5'-nucleotidase superactivity has been reported in four unrelated families with developmental delay and neurological abnormalities. Treatment with uridine is said to have been beneficial.

Deficiency of dihydropyrimidine dehydrogenase (DHPDH)

This autosomal recessive disorder presents with variable degrees of hypertonia, epilepsy, and autism. Some cases have only presented during adult life, when they have developed severe adverse side-effects following cancer chemotherapy with 5-fluorouracil. Uracil and thymine are elevated in the body fluids, including urine. Absent enzyme activities have been demonstrated in blood, cerebrospinal fluid, leucocytes, liver, and fibroblasts. There is no effective treatment for this condition and the prognosis for life is very variable.

N-Carbamyl-b-aminoaciduria

To date, just one patient has been detected with ureidopropionase deficiency causing *N*-carbamyl-b-aminoaciduria. This patient presented with choreoathetosis, hypotonia, and microcephaly.

---

*We are indebted to Professor George Nuki, who wrote on this subject in the third edition of the textbook, for permission to use Fig. 4 and Fig. 7 of that contribution in this chapter.

Further reading

Ahota AS, *et al.* (2001). Adenine phosphoribosyltransferase deficiency and 2,8-dihydroxyadenine lithiasis. In: Scriver CS, *et al.*, eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 2571–662. McGraw-Hill, New York.

Bax BE, *et al.* (2000). *In vitro* and *in vivo* studies with human carrier erythrocytes loaded with polyethylene glycol-conjugated and native adenosine deaminase. *British Journal of Haematology* **109**, 549–54.

Becker MA (2001). Hyperuricaemia and gout. In: Scriver CS, *et al.*, eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 2513–35. McGraw-Hill, New York.

De Ruiter CJ, *et al*. (2002). Muscle function during repetitive moderate-intensity muscle contractions in myoadenylate deaminase-deficient Dutch subjects. *Clinical Science* **102**, 531–39.

Desaulniers P, *et al.* (2001). Crystal induced neutrophil activation. VII: Involvement of Syk in the responses to monosodium urate crystals. *Journal of Leukocyte Biology* **70**, 659–68.

Fam AG (2001). Difficult gout and new approaches for control of hyperuricaemia in the allopurinol-allergic patient. *Current Rheumatology Reports* **3**, 29–35.

Gresser U, Zöllner N (1991). *Urate deposition in man and its clinical consequences*. Springer-Verlag, Berlin.

Harkness, *et al*. (1988). Lesch-Nyhan syndrome and its pathogenesis: purine concentrations in plasma and urine with metabolite profiles in CSF. *Journal of Inherited Metabolic Diseases* **11**, 239–52.

Hershfield MS, Mitchell BS (2001). Immunodeficiency diseases caused by adenosine deaminase deficiency and purine nucleoside phosphorylase deficiency. In: Scriver CS, *et al.*, eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 2585–625. McGraw-Hill, New York.

Hochberg MC (2001). Gout. In: Silman AJ and Hochberg MC, eds. *Epidemiology of the rheumatic diseases*, 2nd edn, pp 230–42. Oxford University Press, Oxford.

Jinnah HA, Friedmann T (2001). Lesch–Nyhan disease and its variants In: Scriver CS, *et al.*, eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 2537–70. McGraw-Hill, New York.

Jinnah HA, *et al.* (2000). The spectrum of inherited mutations causing HPRT deficiency. 75 new cases and a review of 196 previously reported cases. *Mutation Research* **46**, 309–26.

Landis RC, Haskard DO (2001). Pathogenesis of crystal induced inflammation. *Current Rheumatology Reports* **3**, 36–41.

Lipkowitz MS, *et al.* (2001). Functional reconstitution, membrane targeting, genomic structure and chromosomal localisation of a human urate transporter. *Journal of Clinical Investigation* **107**, 1103–15.

Liu R, *et al.* (2000). Extracellular signal-regulated kinase [1]/extracellular signal regulated kinase 2 mitogen-activated protein kinase signalling and activation of activator protein 1 and nuclear factor kappa b transcription factors play central roles in interleukin-8 expression stimulated by monosodium urate monohydrate and calcium pyrophosphate crystals in monocytic cells. *Arthritis and Rheumatism* **43**, 1145–55.

Liu R, *et al.* (2001). Src family protein tyrosine kinase signalling mediates monosodium urate crystal-induced IL-8 expression by monocyte THP-1 cells. *Journal of Leukocyte Biology* **70**, 961–8.

MacDermott K, Allsop J, Watts RWE (1984). The rate of purine synthesis *de novo* in blood mononuclear cells *in vitro* in patients with familial hyperuricaemic nephropathy. *Clinical Science* **67**, 249–58.

Schreiner O, *et al.* (2000). Reduced secretion of pro-inflammatory cytokines of monosodium urate crystal stimulated monocytes in chronic renal failure: an explanation for infrequent gout episodes in chronic renal failure patients? *Nephropathy, Dialysis and Transplantation* **15**, 644–9.

Stone TW, Simmonds HA (1991). *Purines: basic and clinical aspects*. Kluwer Academic, Dordrecht.

Van den Berghe G, Jacken J (2001). Adenylosuccinate lyase deficiency. In: Scriver CS, *et al.*, eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 2653–62. McGraw-Hill, New York.

Waring WS, Webb DJ, Maxwell SRJ (2000). Uric acid as a risk factor for cardiovascular disease. *Quarterly Journal of Medicine* **93**, 707–13.

Watts RWE (1985). Defects of tetrahydrobiopterin synthesis and their possible relationship to a disorder of purine metabolism (the Lesch–Nyhan syndrome). In: Weber G, ed. *Advances in enzyme regulation*, Vol 23, pp 25–58. Pergamon Press, Oxford.

Watts RWE, *et al.*(1987). Lesch–Nyhan syndrome; growth delay, testicular atrophy and a partial failure of 11b-hydroxylation of steroids. *Journal of Inherited Metabolic Diseases* **10**, 210–23.

Yakink DR, *et al.* (2000). Non-inflammatory phagocytosis of monosodium urate monohydrate crystals by mouse macrophages. *Arthritis and Rheumatism* **43**, 1779–89.

11.5 The porphyrias

T. M. Cox

The haem biosynthetic pathway holds great fascination for biochemists who marvel at the evolution of ancient enzymes which interact to bring about the formation of the pigments of life—haemoglobin, the cytochromes, chlorophyll, and the cobalamins (vitamin B_{12}). It is unfortunate that, because of complexities in their chemical structure and ambiguities in their technical nomenclature, the terminology of the porphyrin pigments and the diseases associated with their disturbed metabolism are perceived to be confusing and intimidating. These considerations apply particularly to the acute porphyrias which are rare but distressing syndromes that mimic other acute illnesses but for which recognition may be critical for the patient's survival; too often the diagnosis is not established until permanent disability or even death supervenes.

The porphyrias are caused by disturbances in the multistep pathway for the formation of haem—a pigment essential for oxygen transfer and the energy-yielding reactions of electron transport. The formation of haem is tightly regulated so that acquired or hereditary defects of any of its component reactions lead to the overproduction of haem precursors. Potentially photoactive macrocyclic compounds and toxic precursors of pyrroles thus accumulate. Most of the human porphyria syndromes result from uncommon genetically determined deficiencies of unitary enzymes of the haem biosynthetic pathway; but certain toxins including lead, iron, and hydrocarbons influence the pathway and cause porphyria in susceptible individuals. Similarly the metabolism of endogenous molecules, including steroid hormones, and xenobiotics such as alcohol and many therapeutic drugs, may disturb the delicate equilibrium that is achieved in asymptomatic patients with latent porphyria. Thus gene–environment interactions in previously fit individuals may precipitate sporadic attacks of acute porphyria.

Classification: types of porphyria

The porphyrias are disorders of metabolism characterized by overproduction of the precursors of haem synthesized principally in the liver and bone marrow. About 15 per cent of *de novo* haem biosynthesis occurs in the liver and about 80 per cent in the erythroid marrow. Hepatic synthesis of haem is subject to rapid and wide oscillations in flux; haem biosynthesis in the erythropoietic bone marrow is under most circumstances constitutive and stable. However, haem synthesis may be increased as the erythron expands and proliferates to meet the demands of blood loss or haemolysis, including ineffective erythropoiesis.

Hitherto the porphyrias have been classified into the hepatic and erythropoietic types depending on the principal location at which overproduction of haem precursors occurs. For clinical purposes, however, an operational definition of the porphyric syndromes is more usefully presented as the acute and the non-acute porphyrias. The acute porphyrias cause life-threatening neurovisceral manifestations typically precipitated by environmental factors that occur sporadically. The non-acute porphyrias are characterized by photosensitivity syndromes resulting from the overproduction of macrocyclic porphyrins which cause light-induced skin injury. Several of the acute porphyrias may be associated also with the overproduction of porphyrin intermediates and so may be accompanied at times by long-term photosensitivity which is often exacerbated during the acute attacks. In all instances it is the overproduction of haem precursors that characterizes the condition biochemically: this is the principal means by which a diagnosis can be made of the underlying enzymatic defect during the acute attack. Table 1, Table 2, and Table 3 set out the individual defects that characterize the clinical porphyrias and summarize the clinical features of these hereditary syndromes.

Formation of haem

Haem biosynthesis is catalysed by eight enzymes and is co-ordinated between mitochondrial and cytoplasmic compartments in the cell (Fig. 1). The first committed precursor, 5-aminolaevulinate, is formed in the mitochondria from glycine and the Krebs' cycle intermediate, succinyl-CoA, by one or other of the two isozymes of 5-aminolaevulinate synthetase. Precursor 5-aminolaevulinate is then exported to the cytoplasm where it undergoes condensation to form the monopyrrole, porphobilinogen, four molecules of which are then condensed to yield the macrocyclic tetrapyrrole, uroporphyrinogen III. This reaction is brought about by porphobilinogen deaminase and uroporphyrinogen III synthetase acting co-ordinately to reverse the orientation of one porphobilinogen molecule to yield the uroporphyrinogen III isoform that is the sole precursor of biological haem. Porphyrins of the I series do not serve as biological intermediates in the formation of protoporphyrin IX or haem.

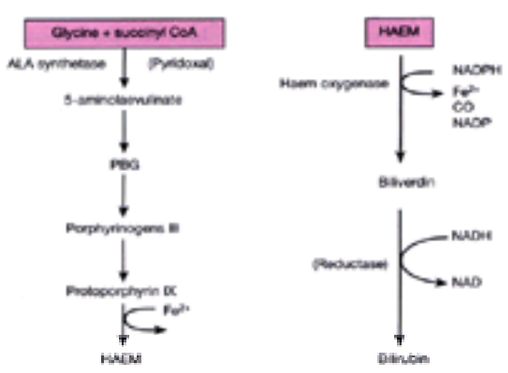

Fig. 1 Main pathways for haem biosynthesis and degradation in humans.

The cytoplasmic enzyme, uroporphyrinogen III decarboxylase decarboxylates the four acetate substituent side-chains to yield coproporphyrinogen III, which is then reimported into the mitochondrion for further oxidative decarboxylation. Coproporphyrinogen III oxidase modifies the two propionate side-chains to vinyl groups yielding protoporphyrinogen IX, the penultimate precursor of haem. Protoporphyrinogen oxidase removes six hydrogen atoms to yield protoporphyrin IX, which is the substrate for the final step in haem biosynthesis. The insertion of ferrous ions into the porphyrin macrocycle to form ferroprotohaem (haem) is catalysed by the mitochondrial enzyme ferrochelatase.

Haem serves as a key prosthetic group in haem proteins including cytochromes, myoglobin, and haemoglobin by which it fulfills its essential biological roles as a transporter of oxygen and electrons in the respiratory chain and in the metabolism of xenobiotics. The two isozymes (constitutive erythroid and the inducible hepatic isozyme) of 5-aminolaevulinate synthetase catalyse the rate-limiting step of haem biosynthesis. The hepatic isozyme maps to the autosome, chromosome 3, but the

erythroid isozyme of 5-aminolaevulinate synthetase (ALAS-2) maps to the X chromosome. These enzymes are subject to differential regulation principally involving transcriptional control in the liver and translational and post-translational control mechanisms in the erythroid cell by the end product, haem. These mechanisms regulate the activity of the whole biosynthetic pathway. Pyridoxal 5-phosphate (derived from vitamin B_6) is an essential cofactor for 5-aminolaevulinate synthetase isozymes. Deficiency of pyridoxine or interference with its metabolism leads to sideroblastic anaemia.

The second enzyme of the haem biosynthetic pathway, 5-aminolaevulinate dehydratase, is a multimeric enzyme with reactive sulphydryl groups that are particularly sensitive to the toxic effects of heavy metals, especially lead, so that 5-aminolaevulinate dehydratase activity is a sensitive measure of environmental and industrial toxicity. Moreover, 5-amino-laevulinate dehydratase is inhibited competitively by the metabolite succinyl acetone, concentrations of which rise to inhibitory levels in patients suffering from the defect of aromatic amino-acid degradation, tyrosinaemia type I. Patients with tyrosinaemia type I and lead poisoning suffer neurovisceral manifestations that resemble the acute porphyrias and it appears likely that overproduction of aminolaevulinate, as a result of arrest at the 5-aminolaevulinate dehydratase reaction, contributes to this effect.

In living cells most of the macrocyclic precursors of the haem biosynthetic pathway are present as their reduced porphyrinogen precursors which are not themselves photoreactive. However, when these tetrapyrroles (uroporphyrinogen, coproporphyrinogen, and protoporphyrinogen) are produced in excess, they diffuse into plasma and tissues where they react with ambient oxygen to form their parent porphyrins, which are spectacularly fluorescent. The double-bond resonance structure of these macrocyclic compounds promotes the formation of singlet oxygen by the transfer of absorbed energy to ground-state oxygen through light activation. It appears that generation of singlet oxygen brings about the photodermatoses associated with the porphyrias; these are characterized by photosensitization of the skin and tissues exposed to light in a broad region of the spectrum including the visible range (350 to 430 nm). Porphyrias associated with overproduction of formed macocyclic haem precursors are thus associated with photosensitivity; the particular skin reactions that develop differ between the particular enzyme defects. This may be explained principally by the degree of hydrophobicity of the overproduced porphyrins and their solubility in cellular membranes.

The first tetrapyrrole that serves as an immediate precursor to haem is uroporphyrinogen III, formation of which requires co-ordinated action of the two cytoplasmic enzymes, uroporphyrinogen I synthase (porphobilinogen deaminase) and uroporphyrinogen III cosynthase. In the absence of adequate cosynthase activity, there is a marked overproduction of porphyrins of the I series, which do not form biologically active ferroprotohaem. Deficiency of uroporphyrinogen III cosynthase leads to the very rare but disabling syndrome of Gunther's disease (congenital erythropoietic porphyria). This disorder is characterized by extreme photosensitivity, haemolysis, and the passage of pink urine containing abundant porphyrins of the I isoform. Persistently high concentrations of these toxic molecules in body fluids leads to staining of the teeth and bones and extreme photosensitive damage, often with cruel and painful skin disfigurement and hair loss.

Porphyria cutanea tarda is caused by deficiency of uroporphyrinogen decarboxylase, defects of which involve complex interactions between heredity and environmental factors. The enzyme activity is markedly decreased in the presence of excess tissue iron and, although rare familial cases of porphyria cutanea tarda occur, most patients have a sporadic disease which is provoked by exposure to environmental toxins such as alcohol, oestrogens, hydrocarbons, iron (often associated with mutations in the haemochromatosis gene, *HFE*), and hepatitis C. At the time of writing, the pathogenic relationship between these external factors and the manifestations of uroporphyrinogen decarboxylase deficiency is unclear.

The final step in the haem biosynthetic pathway involves insertion of ferrous iron into the protoporphyrin nucleus generated enzymatically from protoporphyrinogen IX by protoporphyrinogen IX oxidase. This latter step occurs in the mitochondrion. Ferrochelatase depends on the iron–transferrin cycle for the delivery of iron from plasma transferrin. In the bone marrow, when the iron supply is deficient, freely available zinc may be preferentially converted to zinc protoporphyrin rather than ferroprotohaem thus offering a convenient means to monitor iron-deficient erythropoiesis. Similarly, industrial lead exposure that inhibits both iron delivery and the activity of the sulphydryl enzyme ferrochelatase causes accumulation of zinc protoporphyrin and free protoporphyrin in erythroid precursors and reticulocytes. Deficiency of ferrochelatase leads to the accumulation of free protoporphyrin in liver tissue, plasma, and the skin, where it induces marked photosensitivity. The accumulation of excess protoporphyrin in red cell precursors leads to the characteristic fluorocytes (young red cells containing excess free protoporphyrin) that are the easily recognized hallmark of patients with burning photosensitivity caused by protoporphyria.

The highly regulated control mechanism of haem biosynthesis ensures that the free concentrations of the toxic intermediates involved in the pathway are kept low unless there is a metabolic arrest at one of the biosynthetic reactions; under these circumstances an overproduction of the intermediate compounds occurs which can be used for diagnosis. This overproduction predisposes to the development of the particular clinical porphyric syndrome. A knowledge of the enzymatic steps and of the differential solubility of the haem precursors facilitates appropriate diagnostic testing for the precise identification of suspected porphyria. In principle, overproduction of the early precursors such as aminolaevulinic acid is a common feature of those syndromes associated with neurovisceral manifestations or acute attacks of porphyria. Aminolaevulinate, in particular, represents a common biochemical marker of such attacks and those syndromes that mimic the porphyrias such as hereditary tyrosinaemia type I and lead poisoning. In patients with cutaneous photosensitivity, overproduction of the formed porphyrin macrocycles can be detected also in plasma, urine, and faeces in which they are distributed according to their aqueous solubility (Table 4).

The profile of molecules that are overproduced in a given syndrome may be predicted from the level at which the enzymatic arrest occurs as flux through the pathway is stimulated by diminished negative feedback. In those porphyrias where the principal site of production appears to be in the liver, including the acute porphyrias and porphyria cutanea tarda, oscillations in the flux through the biosynthetic pathway as a result of regulatory effects from environmental or endogenous factors can occur very rapidly; indeed minute-to-minute oscillations in biosynthetic haem fluxes have been recorded in the liver. Thus in starvation and on challenge with xenobiotic reagents (which place a demand for the production of haem to meet the needs for new cytochrome formation), as well as with endogenous hormonal changes, enhanced flux through the pathway leads to toxic overproduction of 5-aminolaevulinic acid. By the same token, rapid repression of the haem biosynthetic pathway in the liver can be induced by the administration of exogenous haem—a useful agent in the control of acute attacks and which rapidly corrects the disturbed metabolism (see below).

Haem formation in the erythron is more rapid than that in the liver but is not subject to sudden oscillations in synthetic rates. Nonetheless in patients with erythropoietic porphyrias, such as congenital porphyria, enhanced rates of red cell destruction when hypersplenism supervenes or in response to light exposure greatly exacerbate the overproduction of porphyrin intermediates and aggravate photosensitivity due to increased porphyrin release. Short-term experiments indicate that exogenous haem may partially repress the endogenous haem biosynthetic pathway in erythroid tissue but this has not proved to be useful for long-term relief in the erythropoietic porphyrias. Blood transfusion to suppress erythropoiesis or definitive replacement of bone marrow by transplantation has, however, proved to be successful in controlling the devastating manifestations of congenital erythropoietic porphyria.

Pathogenesis

The individual porphyria syndromes are described briefly below but the main manifestations (neurovisceral or phototoxic) remain the subject of further clinical research.

Acute neurovisceral attacks

These attacks occur in four of the porphyrias indicated in Table 1, Table 2 and Table 3. In all but one, Doss porphyria (aminolaevulinate dehydratase deficiency), the inheritance is as an autosomal dominant disease. 5-Aminolaevulinate dehydratase deficiency is inherited as an extremely rare recessive condition. Clinical expression is characterized by acute, life-threatening attacks of neuropathy that include abdominal pain, psychiatric symptoms, signs of sympathetic and hypothalamic autonomic overactivity, sometimes accompanied by convulsions and motor and sensory deficits. The syndrome is characteristically precipitated by drugs that induce hepatic haem formation and are metabolized by the hepatic cytochrome P-450 system. Neuropathological examination shows axonal degeneration and central chromatolysis in anterior horn cells and in the brain. Electromyography may reveal denervation compatible with a primary axonal neuropathy of peripheral nerves.

Although this acute porphyria is associated with lone overproduction of 5-aminolaevulinic acid, common to all those associated with acute manifestations, a toxic effect of this precursor is not the only potential mechanism of injury. The structure of aminolaevulinate is analagous to the inhibitory neurotransmitters g-aminobutyric acid and L-glutamate. It seems likely that 5-aminolaevulinate may interfere with the action of the GABA-ergic system—the best evidence for which appears to be its ability to inhibit melatonin production in the rat pineal gland *in vivo*, as has been described in patients with recurrent acute porphyric attacks. It has been further postulated that under the conditions of the acute attack there may be a deficiency of essential haem proteins, such as the cytochrome P-450 isozymes in the liver, with further disturbances in secondary metabolism; other possibilities include a decrease in the activity of hepatic tryptophan dioxygenase, leading to increased formation of 5-hydroxytryptamine (serotonin).

At present there is no clear resolution between combined or individual effects of acute porphyria on the production of neurotoxic pseudotransmitters

(aminolaevulinate) or secondary local deficiency of haem. However, early unpublished but apparently beneficial results of liver transplantation in patients with disabling recurrent attacks of acute intermittent porphyria indicate that the principal cause of the acute syndrome is the hepatic overproduction of toxic haem precursors. In any event, there is convincing evidence of abnormal neurotransmitter function and increased serotonin production—as well as direct interference of GABA receptors by toxic concentrations of 5-aminolaevulinate. Supplying exogenous haem during the acute attack, however, would be expected to correct both arms of this disturbed metabolism, which may account for the beneficial biochemical and clinical effects observed with its use. The recent development of a mouse model of porphyrinogen deaminase deficiency showing sensitivity to barbiturates serves as an authentic model of the biochemical and neuropathological manifestations of acute porphyria and may clarify much about the pathogenesis of this disturbing clinical syndrome. Detailed observations of the effects of hepatic transplantation in acute human porphyrias are also eagerly awaited.

Photosensitivity

Porphyrins absorb light maximally in the Soret region (400 to 420 nm) and in the visible wavelength region (between 500 and 600 nm); they re-emit this light energy at lower wavelengths to give pink, orange, or red fluorescence. This fluorescence is associated with the photodynamic effects and excitation to form triplet states; in the presence of oxygen in biological tissues, transfer of electronic energy leads to the generation of reactive oxygen species, including singlet oxygen, leading to complement activation and cutaneous toxicity. Careful studies examining the photoactive spectrum of skin from patients with various porphyrias has confirmed a cause-and-effect relationship between irradiance within the absorbing wavelength range of the given porphyrin and the development of weal-and-flare and other cutaneous phototoxic responses.

Distinct porphyric syndromes are associated with the accumulation of a particular formed macrocyclic porphyrin, each with its particular solubility properties in plasma and in cell membranes. In porphyria cutanea tarda, skin biopsies show subepidermal bullas and electron microscopy reveals vacuoles in the cells of the superficial dermal epithelium. In this disease, as in protoporphyria, the endothelium of the dermal capillary is thickened and the vessels are surrounded by complement and mucopolysaccharide deposits. In protoporphyria, an adequate oxygen supply has been shown to be critical for the development of experimental phototoxicity *in vivo.* Singlet oxygen and other radicals may lead to lipid peroxidation and cross-linking of membrane proteins with activation of late complement components. In the more severe disease, congenital erythropoietic porphyria, egress of uroporphyrin I from circulating erythrocytes, which may be destroyed within capillaries, leads to gross accumulation of porphyrin in dermal tissue and juxtaposed epithelium. Exposure to light is known to promote photohaemolysis indicating that light of the visible wavelength can penetrate the skin sufficiently to induce porphyrin photoactivation *in situ*.

Induction of acute porphyric attacks

Acute attacks of porphyria may be life-threatening illnesses which occur in genetically predisposed individuals who usually remain asymptomatic. The acute episodes develop on exposure to environmental or endogenous factors that place a demand for hepatic haem biosynthesis which leads to the overproduction of porphyrin intermediates and pyrrole precursors. The most frequent precipitating factors are changes in reproductive steroid hormones either due to natural hormone cycles or the administration of exogenous gonadal steroids. Starvation, including that associated with surgical procedures and anaesthesia, intercurrent infections, and many xenobiotics including alcohol as well as prescription drugs, over-the-counter agents, and chemicals present in health foods can precipitate acute porphyria.

Table 5 and Table 6 list drugs that have been classified as unsafe in the porphyrias either because they have been shown to be porphyrinogenic in animals or *in vitro* studies, or have been associated with acute attacks in patients with porphyria. The table is taken from the British National Formulary published by the British Medical Association and the Royal Pharmaceutical Society of Great Britain. It is pointed out in this publication that slight changes in the chemical structure can lead to marked differences in the ability of the drug to induce attacks of porphyria. A more complete list of drugs is provided in a review by Anderson *et al*. (2001) in the Further reading section.

Acute attacks of porphyria occur in the four conditions known as the hepatic porphyrias and particularly occur for the first time in latent carriers who are aged between 15 and 40 years. Attacks have been recorded in children before puberty but are extremely rare and usually occur during febrile illnesses precipitated by the use of porphyrinogenic cough medicines. Although the porphyrias occur in a latent state in men with a frequency that is equal to that in women, women suffering from acute porphyria outnumber men by at least 2:1.

Clinical features of acute porphyria

The clinical manifestations of an acute attack are very diverse and the condition may be indistinguishable from many other disorders. In Table 7 are listed common neurovisceral symptoms of acute porphyric attacks and of these abdominal pain is the most common, but not invariable, presenting symptom. The pain itself may be difficult to identify since it is usually constant but poorly localized and usually unassociated with tenderness. There may be an associated colicky component and later ileus with abdominal distension which may mimic a surgical emergency. Characteristically, constipation occurs but diarrhoea with increased borborygmi may develop. The patient often becomes very distressed and tachycardia is common.

A frequent feature is the development of pain in the limbs, particularly in the upper thighs, but also in any of the somatic muscles of the chest, lumbar region, shoulders, and neck. Ultimately, muscle weakness and respiratory paralysis may occur. The patient becomes restless or frankly disturbed and deluded as in a toxic confusional state. The inability of attending medical personnel to identify the cause of the pain and the distress associated with it often leads to alienation and an exaggeration of the patient's complaints which may be difficult to diagnose: often a suggestion of hysterical conversion syndrome or worse, malingering, is made by attending staff. Hypertension, sweating, and tremor together with tachycardia indicate marked sympathetic overactivity and cardiac arrythmias may ensue. In about 10 per cent of severe attacks, grand mal seizures develop; treatment of which may prolong the attack, since many anticonvulsants are porphyrinogenic. With sustained attacks there may be signs of a peripheral neuropathy which is related to axonal degeneration, principally affecting motor nerves. Peripheral neuropathy in its early stages may not affect the limb and tendon reflexes but with time these will be decreased or absent. Ultimately, progressive muscle weakness affecting the respiratory muscles, diaphragm, and swallowing may lead to paralysis and death in prolonged attacks in which the institution of lifesaving cardiorespiratory resuscitation measures and intensive care assessment is delayed.

In a full-blown attack, mental symptoms including anxiety, sleeplessness, and depression may be prominent. If the porphyric attack is sustained as a result of inadequate management or diagnosis, progressive alienation, visual and auditory hallucinations, and frank paranoia with progressive and homicidal outbursts may occur. These are extremely difficult to contain within the routine environment of the busy acute hospital. Although seizures may be a presenting sign of the acute attack, they are commonly attributable to marked hyponatraemia resulting from the inappropriate secretion of antidiuretic hormone that itself originates from hypothalamic sympathetic overactivity. Treatment of hyponatraemia due to this cause in the acute attack poses special difficulties (see below). The use of large volumes of hypotonic dextrose has in the past often aggravated the hyponatraemia and seizures—as well as cerebral odema.

Diagnosis of the acute attack is suspected on the basis of the past history including photosensitivity or the intermittent discoloration of urine. There may be a history of abdominal pain in first-degree family members, with or without photosensitivity. Confirmation of an acute attack of porphyria requires the demonstration of increased porphyrin precursors in the urine. Most commonly, increased excretion of the monopyrrole, porphobilinogen, is accompanied by increased excretion of urinary 5-aminolaevulinate. However, porphobilinogen excretion is not increased in the rare aminolaevulinate dehydratase deficiency nor in the pseudoporphyria of lead poisoning.

Acute attacks of porphyria appear to be more common in women as a result of changes in reproductive steroid hormones and many women who suffer from periodic attacks do so in the 1 or 2 days before onset of menstrual bleeding; usually as the menopause approaches, the pattern of these attacks may change or worsen, but with the onset of amenorrhoea, severe attacks of porphyria almost invariably cease. Sometimes, acute attacks lasting a day or two may have their onset in the mid-menstrual period around the time of ovulation. Many mild attacks of porphyria resolve spontaneously within a few days, either as a result of withdrawal of the precipitating factor or of natural hormonal rhythms. Prolonged attacks usually result from the interaction of adverse exogenous and endogenous cofactors and may last for many weeks or even months. The ensuing neurological injury, accompanied in severe attacks by bulbar and respiratory paralysis, may lead to prolonged or even permanent disability. Experience shows that in many such cases inappropriate drugs have been given to counter the early manifestations of the condition, for example analgesics, psychotropic drugs, and anticonvulsants. Thus the initiating medical interventions ultimately prove to be critical determinants of outcome where the diagnosis is not suspected or, if known, is ignored.

Outcome

An early series showed that during the first acute attack of porphyria half the patients died. However, perhaps as a result of better hospital facilities to deal with severe

or adverse outcomes, the mortality and effects of the disease in patients with acute attacks has improved. Reports from a single centre reported that about three-quarters of patients with acute intermittent porphyria or variegate porphyria were able to lead normal lives after an acute attack. Recurrent attacks of pain occurred only in a minority during a period of prolonged follow-up; these recurrent attacks were most likely to occur in the first 3 years.

The development of national centres for the treatment of porphyria, the early detection of genetic predisposition in at-risk first-degree relatives, and the dramatic reduction in prescriptions of porphyrinogenic drugs such as barbiturates and sulphonamides, together with better treatment of acute attack, are all responsible for the improved outcome. Nonetheless, acute porphyria remains life-threatening and deaths or marked disability due to prolonged, mismanaged, or undiagnosed attacks are all too frequent. Rapidly recurrent attacks of porphyria may be associated with severe motor neuropathy and sustained hypertension; postural hypotension may result from autonomic neuropathy. In severe cases, cranial nerve palsies, typically affecting the facial nerve and the vagus nerve, occur. Ischaemia of the occipital cortex during acute attacks has been associated in a number of instances with failed recognition of colours or of human faces (aprosopagnosis) and cortical blindness.

Although it appears that progestogens are principally responsible for cyclical or periodic attacks in women because they are more porphyrinogenic than oestrogens, pregnancy itself is not usually associated with adverse outcomes in women at risk from acute attacks. However, drugs such as metoclopramide that provoke attacks may be used mistakenly to control gastrointestinal symptoms in pregnancy and thus place the woman and her unborn infant at risk.

Individual porphyrias

Acute porphyrias

These are, in a descending order of frequency: acute intermittent porphyria, variegate porphyria, hereditary coproporphyria, and Doss porphyria (aminolaevulinate dehydratase deficiency). The first three of these disorders occur in at-risk heterozygotes for a single mutant allele in the cognate gene as autosomal dominant traits; 5-aminolaevulinate dehydratase deficiency is inherited as a very rare autosomal recessive trait.

The overall frequency of heterozygosity for acute porphyrias is estimated to be 1 in 10 000 of the population, of whom only 1 in 5 to 10 will develop an acute attack. In certain populations (South Africa and in the Lapps of Northern Sweden) the frequency rises to 1 in 1000 of the population. In South Africa, a high gene frequency results from the founder effects of the migration of a Dutch settler in the seventeenth century. Variegate porphyria has thus spread to all ethnic groups within the South African population, molecular analysis of which confirms the presence of a single dominant mutant allele of the protoporphyrinogen IX oxidase gene.

In the last decade or so there has been much interest in the identification of very rare homozygous forms of porphyria where the presence of two mutant alleles of the causative gene are generally responsible for severe clinical disease. In most instances, the condition is not truly homozygous since those individuals affected prove to be compound heterozygotes for two mutant alleles of the cognate gene rather than true homozygotes for the many discrete but rare mutations that occur in porphyria but which would only be expected to occur in consanguineous pedigrees.

Acute intermittent porphyria

This, the most frequent of the acute porphyrias, is caused by mutations in the porphobilinogen deaminase gene that maps to human chromosome 11q23 in which well over 100 mutations have been identified. Several widespread mutations have been identified in certain populations but the majority are reported in only one or two pedigrees.

Two isozymes of the human porphobilinogen deaminase enzyme occur in the tissues: an erythroid mRNA variant and a non-erythroid transcript that encodes 17 additional amino-acid residues in its N-terminus lead to synthesis of housekeeping ubiquitous isozyme and an erythroid-specific isozyme. Most mutations cause a decrease in the abundance as well as the activity of the porphobilinogen deaminase enzyme in all tissues. A small proportion of mutations associated with lack of the detectable protein product from the mutant allele are associated with reduction of the housekeeping isozyme but normal enzymatic activity of the erythroid-specific isozyme. Thus in such patients hepatic porphobilinogen deaminase activity may be reduced to approximately half normal values while the activity of the easily accessed red-cell enzyme is within the normal range.

A few mutations lead to the synthesis of a catalytically impaired but stable porphobilinogen deaminase protein from the cognate mutant allele but these appear to be in a minority. Molecular analysis of the porphobilinogen deaminase gene in patients with acute intermittent porphyria has been very valuable in establishing diagnosis of latent heterozygotes at risk in the affected family, for the provision of appropriate counselling and for the introduction of preventative strategies (see below).

Acute intermittent porphyria is characterized solely by acute porphyric attacks and cutaneous photosensitivity does not occur. In most instances the patients do not notice any change in their urine, although on standing, the increased excretion of pyrroles leads to the formation of coloured oxidation products of porphobilinogen (loosely called porphobilin) which may lead to obvious discoloration (Fig. 2 and Plate 1). During the increased excretion of porphyrin precursors, water-soluble porphyrins form as a result of non-enzymatic photochemical reactions induce a pink discoloration. During acute attacks, copious excretion of pyrrole precursors, including porphobilin, may occasionally give the urine a striking appearance resembling blackcurrant juice or strong solutions of potassium permanganate.

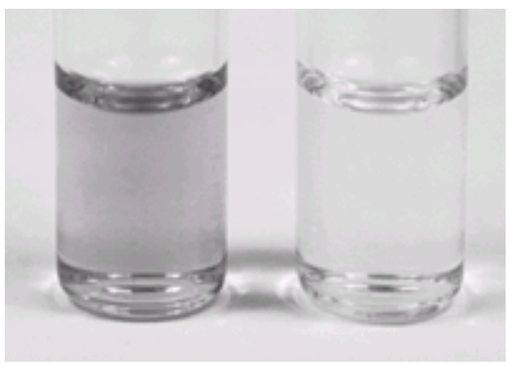

Fig. 2 Urine from a patient with acute intermittent porphyria around the time of an acute attack (left); control urine (right). A positive reaction with Ehrlich's diazo reagent is shown in the patient following the addition of 50 µl of urine to 1 ml of 2 per cent acidic dimethyl benzaldehyde. Subsequent tests showed that the pink diazo adduct was insoluble in chloroform and other organic solvents indicating the presence of excess porphobilinogen. (Urobilinogen in excess may give a positive reaction with the diazo reagent but the product is readily extracted into organic solvents.) (See also Plate 1.)

The incidence and severity of acute attacks in acute intermittent porphyria and variegate porphyria are generally greater than in hereditary coproporphyria. Various estimates indicate between 1 in 10 to 1 in 5 of heterozygotes experience acute attacks of porphyria during their lifetime. However, increasing use of molecular diagnostic methods for screening at-risk families, institution of appropriate avoidance, and the careful dissemination of information to family members and their medical advisers will further reduce the likelihood of disease in latent gene carriers. Latent carriers of acute intermittent porphyria have a high frequency of hypertension and although this should be treated, the potential for inducing attacks is increased by the uninformed prescription of antihypertensive drugs. A proportion of subjects appear to suffer depression and other chronic mental symptoms and at least one survey has reported an increased prevalence of acute intermittent porphyria in patients attending long-stay psychiatric facilities—again putting them at risk from the ill-considered use of porphyrinogenic neuroleptic and other psychoactive drugs.

Variegate porphyria

Variegate porphyria is particularly frequent amongst South African white people and other ethnic groups within that country. The condition is associated with typical acute attacks of porphyria as well as skin manifestations (the van Rooten skin). Acute attacks of porphyria occur very much as in acute intermittent porphyria. More than half the patients come to medical attention with skin lesions alone; in the same series only one-fifth of patients had acute neurovisceral disease and a similar

proportion had acute attacks as well as cutaneous disease.

Cutaneous photosensitivity resembles that seen in porphyria cutanea tarda and hereditary coproporphyria (see below) with fragility, milia, hyperpigmentation, and hairiness of light-exposed skin. During acute sunlight exposure, vesicles and even large bullas may form. Microscopic examination of the affected skin shows deposits of immunoglobulin and hyaline material that stains positively with the periodic acid–Schiff reagent in the dermal capillaries with proliferation of the basal lamina. As with porphyria cutanea tarda, ingestion of reproductive steroid, for example the oral contraceptive pill, may induce the cutaneous manifestations of variegate porphyria in otherwise latent heterozygotes.

A few severely affected patients with variegate porphyria have inherited mutations of the protoporphyrinogen oxidase gene (that maps to chromosome 1q22 to 1q23) from each parent, leading to homozygous 'dominant' variegate porphyria. These individuals present in childhood with a severe phenotype associated with marked photosensitivity, convulsions, and developmental delay; they have several skeletal abnormalities including medially deviated and shortened fifth digits. Developmental retardation is prominent, but surprisingly such patients appear to have few if any attacks of acute porphyria.

Hereditary coproporphyria

This condition is an infrequent and often mild form of acute porphyria which may be associated with cutaneous manifestations. It is due to mutations in the coproporphyrinogen III oxidase gene that maps to chromsome 3q12 and is transmitted as an autosomal dominant trait of low penetrance. The condition usually presents with acute attacks of abdominal pain as with the other acute porphyrias and about 30 per cent of patients develop cutaneous photosensitivity. As with several other porphyrias, several children presenting with marked photosensitivity in childhood have been shown to have inherited a mutant allele of the coproporphyrinogen III oxidase gene from each parent giving rise to so-called homozygous dominant hereditary coproporphyria. Particular mutations in the gene are usually restricted to individually infected pedigrees. As with the other acute porphyrias, molecular analysis of the coproporphyrinogen III oxidase gene may be of value in identifying at-risk heterozygotes for genetic counselling and provision of appropriate advice about the prevention and management of symptomatic disease.

5-Aminolaevulinate dehydratase deficiency (Doss porphyria)

Only a few affected homozygotes for this condition have been identified. Molecular analysis of the cognate gene has revealed the presence of compound heterozygosity and homozygosity for point mutations in the gene which maps to chromosome 9q34. As with the porphobilinogen deaminase gene, there are two promoter regions and alternative non-coding exons that allow for the synthesis of housekeeping and erythroid-specific transcripts. Less than 10 cases of this porphyria have been reported but it seems likely from the individual case histories of those identified that the disease will be underrecognized as the cause of acute abdominal crises usually presenting shortly after puberty and associated with neurological symptoms, including respiratory paralysis. The condition resembles acute lead poisoning. The urine contains an excess of 5-aminolaevulinate but excretion of porphobilinogen and tetrapyrrolic haem precursors is normal. Heterozygotes for aminolaevulinate dehydratase deficiency have been reported in at least one lead worker in whom peripheral neuropathy was ascribed to simple lead poisoning but it may have resulted from the susceptibility of the residual 5-aminolaevulinate dehydratase to inhibition by environmental lead.

Cutaneous porphyrias

Congenital erythropoietic porphyria is a classic but very rare syndrome now known to have an astonishing range of presentation from severe haemolytic anaemia *in utero*, severe photosensitivity presenting soon after birth (with excess porphyrins staining the teeth and urine), to mild late-onset forms presenting with skin lesions in adult life. Most patients have a mild to severe haemolysis with increased reticulocytosis, circulating normoblasts, decreased serum haptoglobin, and increased unconjugated bilirubin concentrations. Inclusion bodies are often seen in marrow, erythroid cells, and circulating normoblasts. Splenomegaly develops in childhood, thereby causing pancytopenia as a result of hypersplenism; this accelerates the haemolysis and leads to compensatory erythropoiesis in the bone marrow. Under these circumstances, splenectomy may help to control the condition.

The classic skin manifestations are of severe blistering lesions on sun-exposed skin, particularly of the hands and face, with the formation of vesicles and bullas that may become infected. There are pigmentary changes with greatly increased skin fragility. Healing of the lesions with or without consequential infection often leads to cutaneous deformities with loss of digits, scarring of the eyelids, nose, lips, scalp, and occasionally blindness due to corneal scarring. Examination of the teeth shows erythrodontia and deformities; exposure to ultraviolet light may reveal striking dental fluorescence. The condition is associated with osteoporosis and resorption of long bones as a result of gross expansion of the erythroid bone marrow.

Mutations in the uroporphyrinogen III synthase gene that maps to chromosome 10q25.3 to q26.3 have been shown to be responsible for this disease and thus may assist in the prenatal diagnosis of mothers harbouring an at-risk pregnancy and who have previously given birth to an affected infant. Constitutive activation of the haem biosynthetic pathway in erythroid cells leads to persistent overproduction of uroporphyrinogen I and coproporphyrinogen I as byproducts of the defective synthesis of uroporphyrinogen III, the sole precursor of protoporphyrin IX and haem. These reduced and colourless metabolites become oxidized to the fluorescent tissue and urinary porphyrins associated with the passage of pink urine that characterizes this often devastating disease.

Porphyria cutanea tarda

This disease is the most common of the cutaneous porphyrias and, unlike other hepatic porphyrias, is never associated with acute porphyric crises. The disease is characterized by skin blistering which is related to sunlight exposure. It occurs in several forms. Porphyria cutanea tarda may result from environmental exposure to dioxin or to hexachlorobenzene, particularly after industrial accidents such as that which occurred in Turkey in the 1960s. Occasional cases have been reported after exposure to other hallogenated phenols but under these circumstances it appears simply to be an environmental toxic syndrome. Toxic cutaneous porphyria appears to be separate from the sporadic porphyria cutanea tarda which is precipitated by other specific environmental factors: increased hepatic storage iron, excess ethanol consumption, administration of oestrogens, hepatitis C virus infection, human immunodeficiency virus infection and possibly, nutritional deficiencies including antioxidants such as vitamin C.

Most individuals who develop sporadic porphyria cutanea tarda prove to have increased iron stores in association with the presence of one or more mutant alleles for the *HFE* gene that also predispose to the development of hereditary adult haemochromatosis. In addition, many patients with sporadic porphyria cutanea tarda consume excess alcohol and smoke. There is a clear association between porphyria cutanea tarda and renal impairment in which the development of disease can be explained by the presence of iron overload (as a result of defective iron utilization with or without routine iron supplementation, particularly in patients on haemodialysis) and failure to excrete excess plasma porphyrins that do not readily diffuse through the peritoneal cavity or haemodialysis membranes. In sporadic porphyria cutanea tarda there is a partial deficiency of uroporphyrinogen III decarboxylase activity in the liver and no family history of the condition. The sequencing of the human uroporphyrinogen decarboxylase gene that maps to human chromosome 1p34 has not provided any evidence of mutations to account for the tissue-specific enzyme deficiency and no isoforms of the enzyme have yet been identified. At the time of writing the molecular pathogenesis of sporadic porphyria cutanea tarda is unknown, but it is also clear that iron and other environmental influences inactivate hepatic uroporphyrinogen decarboxylase. The relationship between regulators of iron homeostasis and the demand for haem biosynthesis in the hepatocytes of affected individuals is not understood but it appears likely from studies in experimental animals that genetic variation in the expression and activity of cytochrome isozymes such as P-450 IA2 may be critical for disease expression. Irreversible inhibition of hepatic uroporphyrinogen decarboxylase may also explain the occurrence of toxic porphyria cutanea tarda after exposure to halogenated hydrocarbons, metabolites of which cause experimental uroporphyria in animals.

Less than one-quarter of patients who suffer from porphyria cutanea tarda show a familial susceptibility to the condition. In these cases, mutations occur in one allele of the human uroporphyrinogen decarboxylase gene leading to catalytic deficiency of the enzyme in all cells, including erythrocytes. In most instances the genetic defect leads to partial reduction of the enzyme protein encoded by the mutant allele. Studies of pedigrees affected by familial porphyria cutanea tarda indicate that expressivity of the trait is very low: less than 10 per cent of heterozygotes develop clinical disease. Conversely, a very few patients present with a syndrome that closely resembles congenital erythropoietic porphyria with marked blistering skin lesions, excess hair growth, and cutaneous scarring in association with the excretion of pink or red urine. These individuals represent a homozygous form of uroporphyrinogen decarboxylase deficiency, termed hepato-erythropoietic porphyria, associated with a variety of mutations in the uroporphyrinogen III decarboxylase gene.

In hepato-erythropoietic porphyria, the activity of uroporphyrinogen decarboxylase is markedly deficient although residual activity remains to preserve essential haem biosynthesis in the erythron and liver. Most patients with hepato-erythropoietic porphyria ultimately develop splenomegaly with accelerated haemolysis closely resembling congenital erythropoietic porphyria. Molecular analysis of the human uroporphyrinogen decarboxylase gene may assist the prenatal diagnosis of at-risk

pregnancies in women who have already given birth to an affected infant.

The clinical features of porphyria cutanea tarda of whatever form are very characteristic and are confined to light-exposed skin (Fig. 3 and Plate 2). Most often, the only signs are of erosions resulting from minor trauma in skin with increased fragility as a result of light exposure, typically on the dorsum of the hands. Other changes include the development of large subepidermal bullae after exposure to light, which may burst leaving ulcerated lesions that are slow to heal. Increased, often accompanied by areas of decreased, pigmentation is a common feature combined with increased hair growth, particularly on the face.

Fig. 3 Porphyria cutanea tarda in a 60-year-old heterozygote for the *HFE* C282Y mutation. This man, a taxi driver, had noticed irritation after exposure of his hands to light transmitted through the windscreen. He had noticed fragility and blistering combined with pigmentary changes typical of this disorder. After treatment by controlled phlebotomy his skin complaint has regressed. (See also Plate 2.)

Patients with porphyria cutanea tarda do not always notice the photosensitivity and rarely experience marked pain unless exposed to brilliant sunlight. Occasionally there is evidence of dermal injury and loss of nails, damage to the conjunctivae, and hair loss. Careful examination of the affected areas shows small depigmented cutaneous scars and the formation of milia. If bacterial infection occurs and there is repeated exposure to sunlight, then severe and permanent scarring may result. Typically, porphyria cutanea occurs in middle-aged men with a history of alcohol use and in women after institution of oestrogen replacement therapy: in young persons, infection with hepatitis C or the immunodeficiency virus may precipitate the disease expression. Frank signs of hepatomegaly or iron overload are rare in porphyria cutanea tarda but have been noted; as with adult haemochromatosis, there is a significantly increased frequency of hepatocellular carcinoma.

Occasional patients with porphyria cutanea tarda may notice an increase in urine excretion of formed porphyrins which, especially after concentration overnight, may resemble the colour of tea or cola. The stool and urine contains large quantities of copro-and uroporphyrins that fluoresce intensely on exposure to long-wavelength ultraviolet light when placed in a suitable vessel for its transmission (namely silica rather than standard glass). Similarly, examination of liver biopsy specimens under ultraviolet light reveals bright red/orange fluorescence; microscopical examination may also show coincidental hepatitis with or without excess deposits of stainable tissue iron reflecting the increased iron storage of this disease. In sporadic porphyria cutanea tarda, increased storage iron is reflected by modest elevations of serum ferritin that often occur in association with the presence of one or more copies of the C282Y allele of the *HFE* gene that maps to human chromosome 6 and which is associated with adult haemochromatosis.

Treatment

Sunlight exposure should be avoided as much as possible until the porphyrin abnormality is corrected. Care is needed to protect fragile skin from mechanical injury and from infection; sunblock creams may also be useful until the metabolic disturbance is controlled.

Patients with porphyria cutanea tarda should moderate or stop their intake of alcohol and avoid the use of iron tonics and sex hormones, especially oestrogens. Screening should be undertaken for chronic infection with human immunodeficiency virus and hepatitis viruses, especially hepatitis C. Management should include imaging or biopsy of the liver if serum liver-related tests are abnormal as well as measurement of a-fetoprotein, since there is a risk of hepatocellular carcinoma in this disease.

Most patients with porphyria cutanea tarda respond to iron depletion by phlebotomy and initial iron status should be determined by measuring serum ferritin concentrations. Weekly or fortnightly removal of 500 ml of blood will usually correct the abnormal urine and plasma porphyrin profile within a few months but maintenance phlebotomy will be required, usually amounting to the removal of 2 to 4 units of blood at intervals each year. Successful therapy reduces the urinary excretion of porphyrins to normal. Patients with porphyria complicating renal failure should be treated with recombinant human erythropoietin and depleted of iron by gentle phlebotomy or parenteral desferrioxamine, if necessary.

The cutaneous manifestations of porphyria cutanea tarda respond rapidly to low-dose chloroquine treatment, which should be considered in patients with persistent symptoms or at the outset before iron storage has been fully corrected. This action of chloroquine was discovered empirically but the agent forms complexes with uroporphyrin deposits and promotes their external cellular disposal. Chloroquine promotes excretion of uroporphyrin from the liver and induces marked, but transient, porphyrinuria. Although chloroquine usually provides rapid relief from the cutaneous disease and photosensitivity, it does not correct the underlying metabolic defect in the liver; its long-term use is not recommended unless the other provocative factors in porphyria cutanea tarda have been removed. The usual effective dose of chloroquine is 100 to 200 mg once or twice weekly; larger doses are associated with marked hepatic toxicity in porphyria cutanea tarda. The drug is reported to have no therapeutic effect on other photosensitive porphyrias.

(Erythropoietic) protoporphyria

Protoporphyria is caused by the overproduction of the immediate precursor of haem, protoporphyrin IX, principally in the bone marrow. Protoporphyria causes an unusual cutaneous photosensitivity syndrome that presents in infancy. Protoporphyria is also a neglected cause of fatal hepatobiliary disease in about 5 per cent of those affected.

Recent studies indicate that protoporphyria is inherited as a recessive condition. Inheritance of mutations in the coding region of the ferrochelatase gene that partially inactivate the enzyme are coinherited in *trans* with a low-expression allele that occurs at polymorphic frequency in the population. Parent-to-offspring transmission of protoporphyria occurs in less than 10 per cent of cases but in all instances of the disease there is a marked deficiency of the enzyme ferrochelatase (less than 50 per cent of control values). The asymptomatic carrier parent shows only mild ferrochelatase deficiency. The gene for human ferrochelatase maps to chromosome 18q.

Protoporphyria characteristically presents with severe burning pain and cutaneous irritation on exposure to visible light and is usually obvious in infancy or early childhood. Erythema and diffuse oedema may follow marked light exposure but vesicles, blistering, and altered skin fragility are most unusual. After several years, increased pigmentation and thickening of the skin (lichenification) occur, especially over the knuckles. A typical feature is of shallow scarring in the malar regions of the cheeks and at the angle of the lips, where scarring is termed ragades. Overt scarring is unusual. There are no changes in urine colour. Protoporphyria is often the subject of delayed diagnosis because of the marked disparity between the severity of the symptoms and the development of physical signs in the skin.

The cutaneous pathology results from photoactivation of red-cell and plasma-derived protoporphyrin IX in skin capillaries (Fig. 4, Fig. 5 and Plate 3 and Plate 4). Protoporphyrin IX is a hydrophobic molecule that dissolves in cell membranes; it has a photoactivation spectrum in the Soret region with subsidiary activation by green and yellow light. Photoinjury is associated with complement activation and release of vasoactive factors; there is intracellular epidermal oedema accompanied by acute inflammatory changes and extravasated red cells. Deposits of hyaline material are found in superficial capillaries with thickening of the basement membranes. A supply of oxygenated blood appears to be essential for the development of photosensitive damage in protoporphyria.

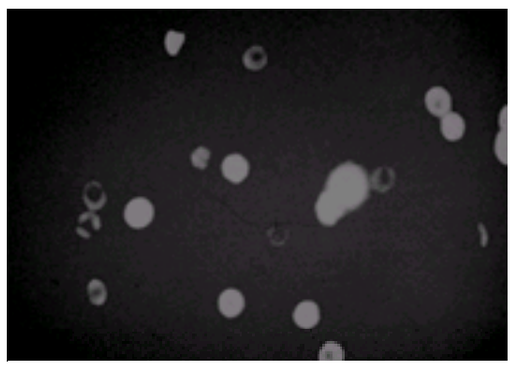

Fig. 4 Fluorescent microscopy of an unstained blood film from a patient with erythropoietic protoporphyria. Note the fluorescence of increased free protoporphyrin within individual young erythrocytes and reticulocytes. (See also Plate 3.)

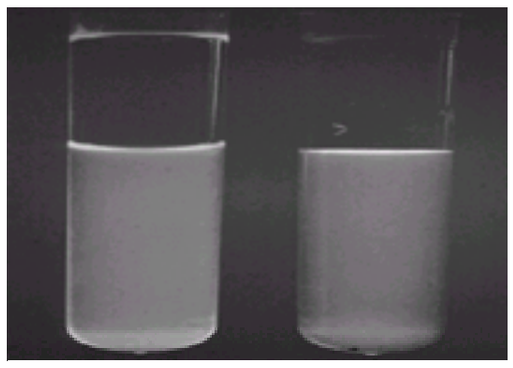

Fig. 5 Examination of human plasma under long-wave ultraviolet light. Plasma on the left was obtained from a patient with protoporphyrin hepatopathy and greatly increased photosensitivity and is compared with plasma obtained from a healthy subject on the right. Note the bright red fluorescence due to the presence of high concentrations of free protoporphyrin. Maximum fluorescence was obtained by exposure to visible light in the violet and green–yellow spectral regions corresponding to the absorbance bands of protoporphyrin.

Mild hypochromic microcytic changes with mild anaemia are usually the only manifestation of disturbed haem biosynthesis and iron metabolism in the bone marrow, although examination of the marrow may reveal occasional sideroblasts with intramitochondrial iron deposits. Haemolysis is usually clinically insignificant until severe cholestatic hepatic disease occurs when splenomegaly and hypersplenism aggravate haemolysis. The photosensitivity worsens under these circumstances and there is upper abdominal pain with splenic enlargement, jaundice, and extreme photosensitivity as concentrations of free protoporphyrin in the plasma rise (Fig. 5). A vicious cycle of decompensation is established with either fulminant hepatic failure associated with cholestasis due to protoporphyrin deposits within biliary radicals, or the development of cirrhosis. Without treatment, the prognosis is dismal and hepatic transplantation is required (see below).

Protoporphyric hepatic disease

Protoporphyria is normally associated with trivial abnormalities of serum liver-related tests but in a small proportion of patients micronodular cirrhosis with pigment deposition occurs. Examination of the liver under polarized light shows birefringent crystals with a characteristic Maltese cross appearance and examination under long-wave ultraviolet light reveals bright fluorescence. Gallstones containing precipitated protoporphyrin occur frequently in protoporphyria but cholestasis results principally from intracellular and cannalicular precipitation of protoporphyrin.

The principal source of protoporphyrin in protoporphyria is the erythron and although under emergency conditions hepatic transplantation may be effective, recurrence of protoporphyrin deposition with injury to the hepatic graft has been reported. The occurrence of this phenomenon, however, is not a contraindication to the use of hepatic transplantation when the illness requires it. Deteriorating hepatic disease is heralded by generalized abdominal pain, splenic enlargement, worsening jaundice, and haemolysis. Interruption of the enterohepatic circulation of protoporphyrin with charcoal or polymeric cationic resins such as cholestyramine may arrest the early downhill course by binding protoporphyrin or promoting hepatic bile acid secretion. However, once established, hepatic decompensation and accelerating photosensitivity is rapid.

Surgical management

Severe protoporphyrin hepatotoxicity is an indication for liver transplantation, preferably carried out by an experienced surgical team with the assistance of an informed anaesthetist and expert physicians in attendance. Consideration should be given to the simultaneous removal of the enlarged spleen at the time of the transplant procedure; there is evidence that splenectomy may reduced the haemolytic component of endstage protoporphyria.

In some patients with endstage liver disease due to protoporphyria, a bizarre neurological syndrome has been identified. In the perioperative period, axonal neuropathies requiring mechanical ventilation and cranial nerve palsies have been reported. Under these circumstances, coproporphyrin and uroporphyrins appear in the urine and may account for a blistering photosensitivity in endstage protoporphyric liver disease.

Operative treatment in patients with protoporphyria can be very dangerous as a result of phototoxic injury to visceral tissues and mucous membranes exposed to brilliant vertical lighting in the operating theatre. Surgical lights are best attenuated by the use of filters that reduce spectral power output below 530 nm; such precautions should be used throughout the perioperative period to reduce overall phototoxicity in the clinical environment. Theoretically, the definitive therapy of protoporphyria will require restoration of erythroid cell ferrocheletase activity in bone marrow. There is a single report of successful marrow transplantation in protoporphyria with coincidental myeloid leukaemia. This procedure cured the symptomatic protoporphyria. In future, either bone marrow transplantation or erythroid progenitor gene therapy will be used to correct this disease in patients suffering from life-threatening liver sequelae. Ancillary treatment by blood transfusion or red cell exchange transfusion will reduce the immediate source of plasma and red cell protoporphyrin, and in the immediate preoperative period plasmapheresis may also reduce phototoxicity. Neurological complications of fulminant protoporphyria may necessitate prolonged ventilatory support in the postoperative period.

Treatment of photosensitivity

Photosensitivity is managed by avoiding excessive light exposure, remembering that visible light of exciting green and violet wavelengths traverses ordinary window glass. Effective sunscreen preparations may assist management, especially in young children at risk. For many years b-carotene has been given to patients with protoporphyria. b-Carotene may absorb light energy at the appropriate wavelengths and also serve as a free-radical quenching agent. The preparation Lumitene (Hoffmann-LaRoche) at a dose of 120 to 180 mg/day is normally used. This causes orange staining of the skin due to carotenaemia but is otherwise well tolerated. It may improve tolerance to sunlight when plasma carotene concentrations between 10 and 15 µmol/l are achieved.

Treatment of an acute porphyric attack

It is essential to establish that the symptoms complained of are caused by an acute attack of porphyria. Of key importance is the careful laboratory analysis of urine and blood early in the course of the illness. This demonstrates elevated concentrations of porphyrins and haem precursors typified by elevated urinary 5-aminolaevulinate and porphobilinogen, which should be high in an attack of acute porphyria. The urine sample should be taken freshly from the patient and protected from light before analysis to avoid non-enzymatic conversion of the porphyrin precursors to porphyrins and hence misdiagnosis.

The immediate management of the acute attack of porphyria

This should involve scrupulous review of avoidable factors recently introduced that would have precipitated an attack. The precipitating factors are usually drugs, alcohol, exogenous or endogenous hormonal changes, fasting (including that due to dieting), or recent surgical procedures. More than 100 drugs may induce attacks of porphyria. Particular care should be taken to exclude agents that are obtained over the counter as tonics or herbal remedies, some of which may induce attacks. Tolerance of alcohol varies greatly in patients with porphyria, many of whom appear to tolerate modest amounts of alcohol. Alcohol is, however, best avoided. At the same time it is wise not to implicate alcohol in an acute attack, unless other causes have been excluded.

Abdominal pain and distress, together with anxiety, require prompt treatment: opiates which are safe in porphyria may be useful, although they often exacerbate constipation. Opiates may be combined with the phenothiazine tranquillizers, such as chlorpromazine, which may potentiate their action usefully.

Since starvation induces attacks of porphyria and haem biosynthesis may be suppressed by the ingestion of carbohydrate, it is advised that patients with minor attacks should eat regular meals containing carbohydrate in a complex form such as starch for its slow release. One-half to two-thirds of the energy intake should be derived from ingested carbohydrate. The management of an acute attack should involve repeated monitoring for the development of hyponatraemia, which may be very severe as a result of inappropriate secretion of antidiuretic hormone. In the past, intravenous glucose or fructose solutions have been advocated as a means to suppress haem biosynthesis in the liver. Great caution is needed in the use of these agents either as 5 or 20 per cent solutions since they exacerbate hyponatraemia and may cause fatal cerebral oedema. In the author's view, if the patient is sufficiently unwell not to be able to control the attack with oral carbohydrate-rich food, parenteral preparations of haem, such as haem arginate, rather than glucose or other sugar solutions, should be administered.

Haem therapy

Haem arginate is administered by a short intravenous infusion in porphyric crises of sufficient severity to merit hospital admission or those associated with limiting pain or metabolic disturbance. Haem arginate (Normosang) supplied by Orphan Europe (see below) is provided as a stable 25 mg/ml concentrate and should be administered at a dose of 3 mg/kg body weight once daily for up to 4 days to a maximum dose of 250 mg in 100 ml of physiological saline infused through a large antebrachial vein over at least 30 min. Haem arginate, like all preparations of haem, tends to polymerize and is unstable; thus the administration should be completed within 1 h after diluting the concentrate. The shelf-life of the concentrate is about 2 years. In the United States, haematin is supplied by Abbott Laboratories and appears to be a comparable preparation for suppressing hepatic haem synthesis and correcting the metabolic disturbance of the acute attack. Haem arginate and a preparation of haem albumin are apparently somewhat more stable than haematin, which tends to produce phlebitis or interfere with the action of coagulant proteins.

Recovery from an acute attack depends on the degree of damage to the nervous system and may occur within 1 or 2 days if haem therapy is introduced at the outset. Cast-iron proof of clinical benefit of haem treatment is lacking, but there is sufficient evidence for the beneficial use of therapy for it to be licensed in 19 countries, including the United Kingdom. Haem arginate therapy has a rapid effect on the excretion of aminolaevulinate and porphobilinogen in acute porphyria and retrospective studies suggest that the outcome of this treatment is better than that in patients previously documented before the use of the agent. Moreover, the results of a double-blind study comparing placebo and haem therapy showed a trend in favour of haem arginate in terms of duration of hospital stay and the requirement for pain relief but the differences did not quite reach statistical significance in the limited study of 12 patients. On the balance of probabilities, however, the evidence for a beneficial effect of haem arginate therapy, particularly at the onset of a porphyric attack, is very strong.

Haem therapy should be used in any patient with significant hyponatraemia, incipient neuropathy, seizures, or bulbar paralysis and in any patient with severe symptoms, particularly of abdominal pain. It must be recognized that patients with established neuropathy may take many months or even years to recover from an attack and, if it is to be effective, haem therapy should be introduced sufficiently early to halt its progress. Where haem therapy is not available, parenteral carbohydrate loading is the only alternative treatment for the acute attack: 2 litres of a 20 per cent w/v glucose solution is recommended over a 24-h period administered through a central venous catheter. There are risks from giving such therapy as outlined above and in the author's opinion the treatment has been superceded by the introduction of stable preparations of haem. Hypersensitivity reactions to haem arginate are rare and the drug has been used during attacks in pregnant women without injury to either the mother or child. Haem contains 10 per cent by weight of iron and the maximum daily dose of haem arginate would contain only 23 mg of elemental iron; the development of iron storage disease is therefore unlikely, except in very rare instances where the patient receives numerous infusions of haematin over prolonged periods.

Occasional patients, usually women, are seen in whom repeated acute attacks occur irrespective of the use of one or two courses of haem arginate. The reason for this is unknown but it is possible that haem arginate therapy induces tachyphylaxis as a result of exaggerated oscillation of haem catabolism by the induction of haem oxygenase in the liver. For this reason tin-protoporphyrin, an inhibitor of haem oxygenase, has been considered. This agent is only available in specialist centres and, because it contains toxic heavy metal and itself may induce photosensitivity, is currently not recommended for routine use. Recently, the combination of recurrent life-threatening porphyric attacks and poor venous access for administration of therapeutic haem preparations has led to the use of liver transplantation in a few young women stricken by this disease. Early (unpublished) reports indicate that this approach may, under exceptional circumstances, be successful.

Young women with cyclical porphyric attacks may require hormonal intervention by the use of gonadotrophin-releasing hormone analogues such as goserelin or buserelin for the release of gonadotrophins. These agents inhibit androgen, oestrogen, and progestogen production—as a result they induce menopausal-like symptoms and depression, as well as rapid decreases in trabecular bone density. Doses sufficient to suppress luteinizing and follicule-stimulating hormone concentrations in serum are required. Their prolonged use for more than a few months is not recommended but buserelin may be used intranasally and may be more convenient. To avoid the worse aspects of hypogonadism in women, low-dose oestrogen therapy under appropriate gynaecological supervision may be coadministered, once cyclic porphyric attacks have come under control. Hypertension is frequent in porphyric attacks and may be very severe as a result of sympathetic overactivity; during the attack, sinus tachycardia is frequent. b-Blockers are effective in the control of the hypertension and propanolol is safe; it also provides effective relief of sinus tachycardia.

Hyponatraemia may be very severe and in acute porphyria progresses on a daily basis during the course of the acute attack in most patients. Its management is critical and the rapid onset of severe hyponatraemia clearly contributes to the confusion and other mental symptoms associated with a porphyric attack. Prompt treatment by careful adjustment of fluid balance and fluid restriction is needed. Great care should be exercised in the presence of hyponatraemia with the use of intravenous solutions whose prescription should be reviewed frequently. The temptation to place a patient with abdominal pain on a surgical ward and administer a dilute solution of glucose is very great in current hospital practice: in the porphyric attack such management may contribute to death as a result of cerebral oedema or the complications of rapid-onset hyponatraemia. Where hyponatraemia progresses rapidly despite fluid restriction, once the diagnosis of inappropriate secretion of antidiuretic hormone is confirmed by determining urine and plasma osmolalities, hypertonic saline solutions or fluid restriction may be required for its correction.

Grand mal seizures in acute porphyric attacks pose a particular problem for management; they are often precipitated by hyponatraemia that frequently complicates the acute attack. Clearly appropriate management of the electrolytic abnormality (with the potential for life-threatening cerebral oedema) is an essential element of treatment. Status epilepticus poses special difficulties but has been treated successfully with parenteral diazepam or the related benzodiazapine, temazepam. Carbamazepine, lorazepam, and midazolam are probably (but not definitely) safe in acute porphyria. Clonazepam or valproate have been used for seizure prevention; the generally outmoded therapy of bromide may also have a role. Acetazolamide, which has been used as a minor agent in seizure prophylaxis, has been used safely in acute porphyria but many first-line drugs such as carbamazepine, sodium valproate, phenytoin, and chloral hydrate have been classified as unsafe or are frankly porphyrinogenic. Primidone and phenobarbitone are absolutely forbidden.

Further problems arise in the management of acutely disturbed patients who are not responsive to the safe phenothiazine, chlorpromazine. Thioridazine is categorized as unsafe but parenteral haloperidol has been used with good effect in occasional patients with uncontrollable or life-threatening manic aggression and paranoid disturbance. In all instances, prescription of any agent to a patient who has suffered from or is suffering from an acute porphyric crisis must involve consultation with a reliable pharmacopoea with individual drugs categorized for safety.

The ability of most drugs to initiate attacks of porphyria appears in many instances to be related to their effects on the induction of haem biosynthesis in the liver and specifically for the formation of the relevant P-450 xenobiotic metabolizing isoforms. One key isoform involved in the induction of porphyria is inhibited, at least *in vitro*, by the H_2-antagonist, cimetidine. It has been reported that cimetidine at 400 to 800 mg daily is sufficient to inhibit induction of this P-450 isozyme in adult humans. Cimetidine has been administered with occasional success as a means to inhibit or control spontaneous porphyric crises and as a last resort it might be considered in patients with life-threatening, otherwise uncontrollable, disease.

There is particular difficulty in young or middle-aged women with cyclical premenstrual attacks. Treatment with high-dose gonadotrophins continued for 1 to 2 years is likely to abort the attacks, but given alone will cause distressing symptoms of hypogonadism with depression and osteoporosis. The worst symptoms of hypogonadism can be overcome by the use of low-dose oestrogen replacement, for example with oestrogen patches, which have a significantly lower risk of provoking an attack of porphyria than progestagen-only hormone preparations. Clearly there is a risk of unopposed oestrogen therapy in patients with an intact endometrium and monitoring for the effects in those receiving oestrogen will be needed.

Acute perimenstrual attacks can be controlled by the prompt administration of haem arginate for 1 to 2 days at the predicted time of susceptibility. Although tachyphylaxis has not been recorded, there may be difficulties in withdrawing the haem arginate because of its effect on inducing haem oxygenase and hence amplifying the potential oscillations of haem biosynthesis in the liver once the haem arginate is withdrawn. The potential for iron overload developing as a result of haem arginate is most unlikely owing to its low content of iron at the doses recommended. Some authors have suggested the use of the haem oxygenase inhibitor, tin-protoporphyrin as an adjunct to the use of haem arginate. Although this may induce a more prolonged biochemical remission of the abnormalities of an acute porphyric attack, it does not induce a more rapid depression of the biochemical abnormality. Experience with tin-protoporphyrin where tachyphylaxis of haem arginate is suspected has been favourable in a few patients, but the drug itself induces photosensitivity. Tin is also potentially toxic as a heavy metal of which only limited excretion occurs. At present the use of tin protoporphyrin or its cogener, zinc deuteroporphyrin must remain speculative and more experience is necessary before these agents can be recommended for safe use in the long-term management of patients with current porphyric crises. The role of liver transplantation and, ultimately corrective gene therapy directed to the liver in acute porphyrias, awaits fuller evaluation in animal models of these diseases and in the few porphyric recipients of hepatic allografts so far recorded.

Sources of information and addresses

British National Formulary, British Medical Association, Tavistock Square, London WC1H 9JP.

United Kingdom and Royal Pharmaceutical Society of Great Britain, 1 Lambeth High Street, London SE1 7JN.

The United Kingdom Drug Information Pharmacists Group website: http://www.ukdipg.org.uk/

Haem arginate (Normosang) is manufactured by Leiras Medica, PO Box 415, SF 20101, Turku, Finland supplied in the United Kingdom by Orphan Europe (UK) Ltd, 32 Bell Street, Henley-on-Thames, Oxon RG9 2BH. Telephone: 44-(0)1491 414 333; Fax: 44-(0)1491 414 443; email: info.uk@orphan-europe.com

Patient associations

The British Porphyria Association, 14 Mollison Rise, Gravesend, Kent DA12 4QJ UK. Telephone: 44-(0)1474 350390.

The American Porphyria and Canadian Porphyria Foundations may also be accessed by the internet websites.

Additional information with emphasis on the molecular genetics of individual porphyrias may be found on the Online Mendelian Inheritance in Man (OMIM) website at ww.ncbi.nlm.gov/omim.

Warning jewellery: it is often valuable in patients with acute porphyrias for them to have a wrist bracelet or neck pendant that provides information about diagnosis in medical emergencies. Details in the United Kingdom can be obtained from The MedicAlert Foundation, 12 Bridge Wharf, 156 Caledonian Road, N1 9UU. Telephone: 44-(0)207 833 3034.

Further reading

Anderson KE *et al.* (1990). A gonadotrophin releasing hormone analogue prevents cyclical attacks of porphyria. *Archives of Internal Medicine* **150**, 1469–74.

Anderson KE *et al.* (2001). Disorders of heme biosynthesis: X-linked sideroblastic anemia and the porphyrias. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 8th edn, Vol II, pp 2991–3062. McGraw-Hill, New York. [This is a most comprehensive and up-to-date account of the human biosynthetic pathway in relation to the porphyrias, a large section within a four-volume treatise on inborn errors of metabolism.]

Elder GH, Smith SG, Smyth SJ (1990). Laboratory investigation of the porphyrias. *Annals of Clinical Biochemistry* **27**, 395–412.

Elder GH, Hift RJ, Meissner PN (1997). The acute porphyrias. *Lancet* **349**, 1613–17.

Gorchein A (1997). Drug treatment in acute porphyrias *British Journal of Clinical Pharmacology* **44**, 427–34.

Kauppinen, R, Mustajoki P (1992). Prognosis of acute porphyrias: occurrence of acute attacks, precipitating factors, and associated diseases. *Medicine (Baltimore)***71**, 1–13.

Mustajoki P, Nordmann Y (1993). Early administration of heme arginate for acute porphyric attacks. *Archives of Internal Medicine* **153**, 2004–8.

Poh-Fitzpatrick MB (1985). Porphyrin-sensitized cutaneous photosensitivity: pathogenesis and treatment. *Clinics in Dermatology* **3**, 41–82.

Schmid R, ed. (1998). The porphyrias. *Seminars in Liver Disease* **18**, 1–101. [An accessible and comprehensive review of the molecular genetics, biochemistry, clinical features, and treatment of human porphyria.]

11.6 Lipid and lipoprotein disorders

P. N. Durrington

Lipids are a heterogeneous group of substances, including oils and fats, that are distinguished by their low solubility in water and their high solubility in non-polar (organic) solvents. The difference between an oil and a fat lies in its melting point. Lipids are essential as energy stores and respiratory substrates, as structural components of cells, as vitamins, as hormones, for the protection of internal organs, for heat conservation, for digestion, and for lactation. Lipoproteins are macromolecular complexes of lipid and protein; their principal function is to transport lipids through the vascular and extravascular body fluids and they are also found as components of milk. Lipoproteins include apolipoproteins and enzymes. Increased concentration of a circulating lipoprotein is termed hyperlipoproteinaemia and a decreased concentration is termed hypolipoproteinaemia. Disturbed composition of circulating lipoproteins is termed dyslipoproteinaemia.

Atherosclerosis is the context in which lipid disorders most commonly present to clinicians and this will be the principal focus of this chapter; undoubtedly lipoproteins will ultimately be implicated in many pathological processes.

Lipid physiology

Triglycerides (triacylglycerols)

These are formed by the esterification of glycerol with fatty acids, which have a hydrocarbon group attached to a carboxyl group. Generally the hydrocarbon moiety is present in a long chain. Naturally occurring fatty acids usually have even numbers of carbon atoms, most of them linked by single bonds, but some contain double bonds. Those with double bonds are termed unsaturated, whereas those with only single bonds are the saturated fatty acids. Fatty acids with one double bond are termed monounsaturated and those with more, polyunsaturated. Each double bond creates the possibility of two stereoisomers according to whether the hydrogen atoms of the –CH=CH– are both on the same side of the double bond (*cis*) or on the opposite sides (*trans*). Naturally occurring fatty acids are mostly *cis* isomers. *Trans* isomers are, however, present in the milk of ruminants, such as the cow, and in margarines.

Triglycerides in adipose tissue provide our principal energy store. The body of a 70 kg man contains some 15 kg of stored triglycerides, representing 135 000 kcal (560 000 J) of energy, which would permit survival during total starvation for up to 3 months (compare this with the 225 g of stored glycogen, representing only 900 kcal (3800 J)). Obesity represents an excess of stored fat, and it is unfortunate for those wishing to slim that considerable and very prolonged dietary energy restriction is necessary to lose weight, given the large amount of energy stored in fat. Each gram of triglyceride produces 9 cal (38 J) of energy, whereas the same mass of carbohydrate or protein only produces 4 cal (17 J), and the latter are more difficult to store because they require an aqueous environment. Thus a muscle or liver cell can only store a minimal amount of glycogen. The adipocyte, on the other hand, contains a droplet of hydrophobic triglyceride surrounded by only a tiny rim of cytoplasm: about 85 per cent of the adipocyte is triglyceride. Thus each gram of adipose tissue yields almost 8 cal (33 J) of energy, whereas tissues containing cells packed to capacity with glycogen would not even approach a yield of 1 cal (4.2 J) for each gram.

For other organs to utilize the energy in adipose tissue the stored triglyceride must first be hydrolysed to its constituent glycerol and non-esterified fatty acids, a process known as lipolysis. This is accomplished by adipose tissue lipase, an intracellular enzyme which is inhibited by insulin (not to be confused with lipoprotein lipase, an extracellular enzyme located on the vascular endothelium of fat and muscle and which is activated by insulin (see below)).

The products of lipolysis are released into the circulation and non-esterified fatty acids bind to albumin. The normally circulating concentration of non-esterified fatty acids is 300 to 800 µmol/litre (8–23 mg/dl), but this falls when insulin is secreted following a meal and rises in starvation when insulin secretion is low. Their importance as a system for transporting lipid energy should not be underestimated, even at low concentrations, since their half-life in the circulation is only 2 to 3 min and their turnover is thus 100 to 200 g/day, and even more in starvation or uncontrolled diabetes.

Non-esterified fatty acids can be oxidized to acetyl-CoA by some tissues, such as muscle and liver, and then entered into the Krebs (carboxylic acid) cycle. Other tissues, which in the fed state rely on glucose as an oxidative substrate, cannot directly utilize non-esterified fatty acids. During starvation these tissues are supplied with water-soluble ketone bodies (acetone, acetoacetate, b-hydroxybutyrate), which the liver produces by partial oxidization (b-oxidation) of non-esterified fatty acids transported to it from adipose tissue. These ketone bodies, which can readily be entered into the Krebs cycle by tissues lacking the ability to oxidize fatty acids, constitute the second system for the transport of lipid energy. They are vital for survival when dietary energy is at a premium, but are also the cause of diabetic ketoacidosis when insulin production is insufficient to suppress the flux of non-esterified fatty acids from adipose tissue, so that the production of ketone bodies takes place at a faster rate than they can be respired. The amount of insulin required to decrease blood glucose increases in the presence of high levels of circulating non-esterified fatty acids. The higher flux of non-esterified fatty acids out of adipose tissue in diabetes thus contributes to insulin resistance. In the case of non-insulin-dependent diabetes a high rate of release of non-esterified fatty acids may have pre-dated the development of hyperglycaemia by many years because obesity, which is a common antecedent of this type of diabetes, is itself associated with an increased flux of non-esterified fatty acids through the circulation and with insulin resistance.

Phospholipids

These also have at least one fatty acyl group esterified to an alcohol and one phosphate group linked both to the alcohol and to another organic compound. The glycerolipids have glycerol as the alcohol. Examples of these are phosphatidylcholine (lecithin) and lysophosphatidylcholine (lysolecithin). Another abundant class of phospholipids are the sphingolipids, such as sphingyomyelin. Phospholipids are essential components of cell membranes and, because of the great diversity of physical properties permitted by their structure, are responsible for much of the variation in membrane structure.

Cholesterol

Cholesterol is also an essential component of cell membranes, where it allows the phospholipid molecules to pack more closely, increasing membrane rigidity. It is

also a precursor for the synthesis of steroid hormones, vitamin D, and bile acids. It is present in arterial fatty streaks and in atheromatous plaques (see below).

Cholesterol is an alcohol and may be unesterified as free cholesterol or esterified with a fatty acyl group (Fig. 1).

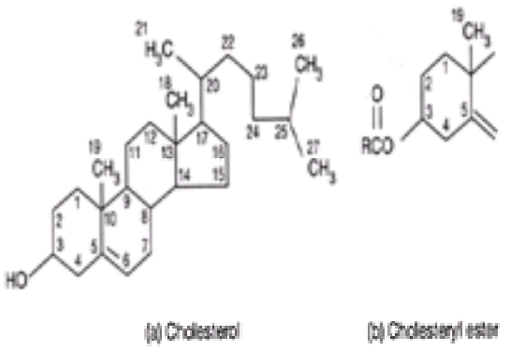

Fig. 1 The structure of free cholesterol and cholesteryl ester.

Lipoprotein physiology

Lipoprotein structure

The general structure of lipoprotein molecules is globular (Fig. 2). The physicochemical considerations, which govern the arrangement of their constituents, are similar to those involved in the formation of mixed micelles in the lumen of the intestine. Thus, within the outer part of the lipoprotein are found the more polar lipids, namely the phospholipids and free cholesterol, with their charged groups pointing out towards the water molecules. In physical terms, however, the role of bile salts, which are also in the outer layer in the mixed micelle, is assumed by proteins, so that the surface structure of a lipoprotein resembles the outer half of a cell membrane. Within the core of the lipoprotein particle are the more hydrophobic lipids, the esterified cholesterol and triglycerides. These form a central droplet to which are anchored, by their hydrophobic regions, the surface-coating molecules, phospholipids, free cholesterol, and proteins. The exception to this general structure is the newly formed or nascent high-density lipoprotein (**HDL**), which lacks the central lipid droplet and appears to exist as a disc-like bilayer consisting largely of phospholipids and proteins.

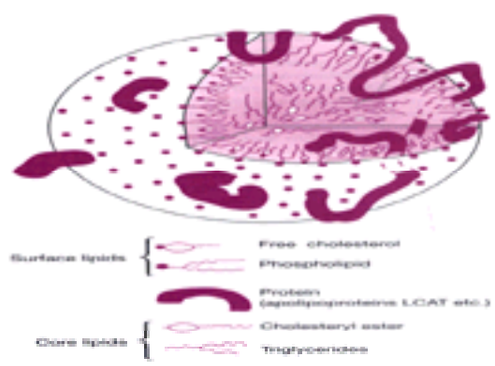

Fig. 2 Lipoprotein structure. The most hydrophobic lipids (triglycerides, cholesteryl esters) form a central droplet-like core, which is surrounded by more polar lipids (phospholipids, free cholesterol) at the water interface. Apolipoproteins are anchored by their more hydrophobic regions, with their more polar regions often exposed to the surface. (Reproduced from Durrington (2002) with permission.)

The protein components of lipoproteins are the apolipoproteins, a group of proteins of immense structural diversity, some of which have a largely structural role and others of which are important metabolic regulators. In addition, enzymes are found as components of lipoproteins. One example is lecithin:cholesterol acyltransferase which is located on HDLs, which are also its site of action.

Lipid transport from liver and gut to peripheral tissues

The products of fat digestion (fatty acids, monoglycerides, lysolecithin, and free cholesterol) enter the enterocytes from the mixed micelles. They are re-esterified in the smooth endoplasmic reticulum of these cells. Long-chain fatty acids (those with more than 14 carbon atoms) are esterified with monoglycerides to form triglycerides and with lysolecithin to form lecithin. Free cholesterol is esterified by the enzyme acyl-CoA:cholesterol *O*-acyltransferase.

The triglycerides, phospholipids, and cholesteryl esters are then combined with an apolipoprotein, known as apoB $_{48}$, in the enterocyte. The lipoproteins thus formed are secreted into the lymph (chyle) as chylomicrons. These are large (diameter > 75 nm; density < 950 g/litre) and are rich in triglycerides but contain only relatively small amounts of protein (Fig. 3). They travel through the lacteals to join lymph from other parts of the body and enter the blood circulation via the thoracic duct. In addition to cholesterol absorbed from the diet, the chylomicrons may also receive cholesterol that has been newly synthesized in the gut or transferred from other lipoproteins present in the lymph and plasma. The newly secreted, or nascent, chylomicrons receive C apolipoproteins from HDL, which in that respect appears to act as a circulating reservoir, since later in the course of the metabolism of the chylomicron, the C apolipoproteins are transferred back to the HDL pool. The chylomicrons also receive apolipoprotein E (**apoE**), although the manner in which they do so is unclear. Unlike other apolipoproteins, which are synthesized either in the liver or the gut or both, apoE is exceptional in that it is synthesized (and perhaps secreted) by a large number of tissues: liver, brain, spleen, kidney, lungs, and adrenal gland.

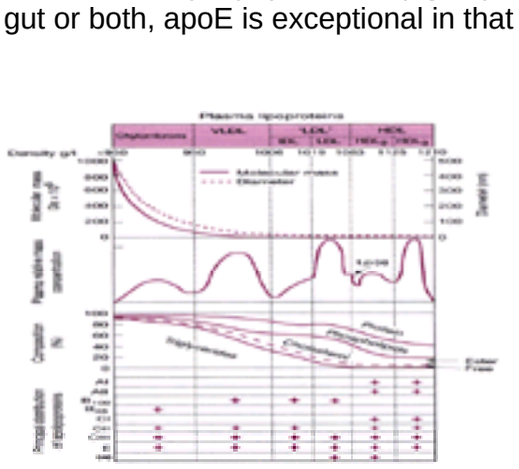

Fig. 3 The spectrum of plasma lipoprotein particles according to their hydrated density, molecular mass, molecular diameter, relative concentration, lipid composition, and apolipoprotein composition. (Reproduced from Durrington (2002) with permission.)

Once the chylomicron has acquired the apolipoprotein, apoC-II, it is capable of activating lipoprotein lipase (Fig. 4(a)). This enzyme is located on the vascular endothelium of tissues with a high requirement for triglycerides, such as skeletal and cardiac muscle (for energy), adipose tissue (for storage), and lactating mammary

gland (for milk). Lipoprotein lipase releases triglycerides from the core of the chylomicron by hydrolysing them to fatty acids and glycerol, which are taken up by the tissues locally. In this way the circulating chylomicron becomes progressively smaller. Its triglyceride content decreases and it becomes relatively richer in cholesterol and protein. As the core shrinks, its surface materials (phospholipids, free cholesterol, C apolipoproteins) become too crowded and they are transferred to HDL. The cholesteryl ester-enriched, triglyceride-depleted product of chylomicron metabolism is known as the chylomicron remnant. The apoB $_{48}$, present from the time of assembly, remains tightly anchored to the core throughout. The apoE also remains and regions of its structure are exposed, permitting chylomicron remnant catabolism via the 'remnant receptor' of the liver and also the low-density lipoprotein (**LDL**) receptors (also called apoB$_{100}$/E receptors), which can be expressed by virtually every cell in the body including the liver. Unlike the LDL receptor, which is discussed in more detail later, the 'remnant receptor' is not downregulated by intrahepatic cholesterol and involves the LDL receptor-related protein, LRP, which in addition to its binding site for apoB also has receptor sites for other proteins. ApoE is inhibited from binding to its receptors earlier in the metabolism of chylomicrons because its receptor-binding domain is blocked by the apolipoprotein, apoC-III. Remnants are largely removed from the circulation by the liver. Although the clearance of these particles by the LDL receptor is theoretically possible, this route is not likely to contribute greatly to remnant uptake in the adult, since the binding of remnant particles to the LRP is enhanced by a trapping mechanism in the space of Disse and elsewhere the remnant particles must compete for binding to the LDL receptor with LDL, the particle concentration of which is much higher than that of the chylomicron remnants (even more so in the tissue fluid than in the plasma). Also the LDL receptor is rapidly downregulated by the lysosomal release of free cholesterol into the cell, which follows the entry of lipoprotein–receptor complexes into the cell, whereas expression of the remnant clearance pathway is unaffected by entry of cholesterol into the liver.

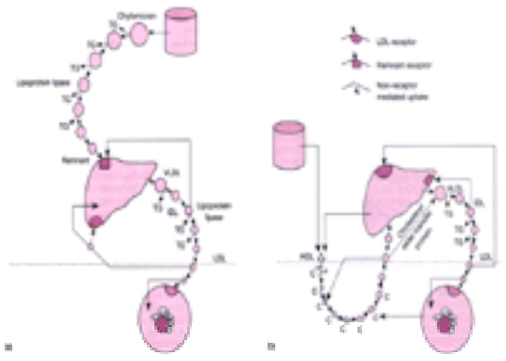

Fig. 4 Metabolism of (a) triglyceride-rich lipoproteins secreted by the gut and liver and (b) hepatic triglyceride-rich lipoproteins and lipoproteins transporting cholesterol to and from the tissues.

The liver itself secretes a triglyceride-rich lipoprotein known as very low-density lipoprotein (**VLDL**) which allows the supply of triglycerides to tissues in the fasting state as well as postprandially. Very low-density lipoprotein particles are somewhat smaller than the chylomicrons (diameter 30–75 nm, density < 1006 g/litre). Once secreted, they undergo exactly the same sequence of changes as chylomicrons; that is the acquisition of apolipoproteins and the progressive removal of triglycerides from their core by the enzyme, lipoprotein lipase. However, some additional transformations are involved in their metabolism in the human. In man, the liver, unlike the gut, does not esterify cholesterol before its secretion. This process differs in species such as the rat. In the human, most of the cholesterol released from the liver each day into the circulation is secreted in the VLDL as free cholesterol, and it undergoes esterification in the circulation. Free cholesterol is transferred to HDL along a concentration gradient. There it is esterified by the action of lecithin:cholesterol acyltransferase, which esterifies the hydroxyl group in the 3-position of cholesterol to a fatty acyl group. This it selectively removes from the 2-position of lecithin to give lysolecithin. The fatty acyl group in this position is generally unsaturated and the cholesteryl esters thus formed are frequently cholesteryl oleate or cholesteryl linoleate. Familial lecithin:cholesterol acyltransferase deficiency is a very rare disorder, in which HDL fails to mature and circulating free cholesterol levels increase. It leads to anaemia, corneal opacities, proteinuria, and renal failure.

Esterified cholesterol on HDL is transferred back to VLDL. This cannot take place by simple diffusion, because cholesteryl ester is intensely hydrophobic and because the concentration gradient is unfavourable. A special plasma protein, cholesteryl ester transfer protein or lipid transfer protein, transports cholesteryl ester from HDL to VLDL. It does this in exchange for triglycerides in VLDL and thus also contributes to the removal of core triglycerides from VLDL. The principal mechanism for the removal of triglycerides from VLDL is, however, lipolysis catalysed by lipoprotein lipase.

Another major difference between VLDL and chylomicrons is that the apolipoprotein B produced by the liver in man is not apoB $_{48}$ but is almost entirely apoB$_{100}$. As in the case of chylomicrons, the quantum of apoB packaged in the VLDL remains tightly associated with the particle until its final catabolism; its amount does not vary after secretion. It is probable that each molecule of VLDL contains one molecule of apoB $_{100}$. The apoB$_{100}$ produced in the liver contains the protein sequence necessary to bind to the LDL receptors, whereas that produced by the gut, although derived from the same gene, does not: a process of 'gene editing', which stops the ribosome translating the messenger RNA before the receptor-binding sequence, leads to an apoB with 48 per cent of the molecular mass of that from the liver. Microsomal triglyceride transfer protein is essential for the process by which both apoB $_{48}$ and apoB$_{100}$ are packaged with triglyceride in the enterocyte and hepatocyte to form chylomicrons or VLDL respectively; this is defective in abetalipoproteinaemia (see below).

The circulating VLDL particles become progressively smaller as their core is removed by lipolysis and surface materials are transferred to HDL. In normal man most of the VLDL is converted to smaller LDL particles through the intermediary of a lipoprotein known as intermediate density lipoprotein (**IDL**). This has a density of 1006 to 1019 g/litre and contains apoE. In this latter respect it is similar to chylomicron remnants. In some species, such as the rat, it is largely removed by the hepatic receptors, and LDL formation is thus bypassed. The enzyme hepatic lipase may be important in the conversion of IDL to LDL.

In man, LDL particles, which are relatively enriched in cholesterol but are small enough (diameter 18–25 nm, density 1019–1063 g/litre) to cross the vascular endothelium and enter the tissue fluid, serve to deliver cholesterol to the tissues. Their concentration in the extracellular fluid is probably about 10 per cent of that in the plasma. Cells require cholesterol for membrane repair and growth and, in the case of specialized tissues such as the adrenal gland, gonads, and skin, as a precursor for the syntheses of steroid hormones and vitamin D. Low-density lipoprotein is able to enter cells by two routes making a major contribution to its catabolism: one which is regulated according to the cholesterol requirement of each individual cell and one which appears to depend almost entirely on the extracellular concentrations of LDL.

The first of these two routes is by a cell-surface receptor, which specifically binds lipoproteins that contain apoB $_{100}$ or apoE. This is the LDL receptor. As mentioned previously, the receptor, although capable of binding apoE-containing lipoproteins, in practice binds mainly to the apoB $_{100}$-containing lipoproteins, of which LDL is the most widely distributed. After binding, the LDL–receptor complex is internalized and undergoes intracellular lysosomal degradation. The apoB moiety is hydrolysed to its constituent amino acids, and the cholesteryl ester is hydrolysed to free cholesterol. The release of this free cholesterol is the signal which regulates the cellular cholesterol content by three co-ordinated reactions. First the enzyme which is rate-limiting for cholesterol biosynthesis (3-hydroxy,3-methyl-glutaryl CoA reductase) is repressed, thus effectively centralizing cholesterol biosynthesis to organs such as the liver and gut. Secondly, the synthesis of the LDL receptor itself is suppressed. Thirdly, acyl-CoA:cholesterol *O*-acyltransferase is activated so that any cholesterol that is surplus to immediate requirements can be converted to cholesteryl ester, which, because of its hydrophobic nature, forms into droplets within the cytoplasm and is thus conveniently stored. The effect of the lysosomal release of free cholesterol on the expression of the LDL receptor contrasts with its effect on the hepatic remnant receptor, which is not subject to any similar downregulatory process. Free cholesterol released by lysosomal digestion of cholesteryl ester-rich, apoE-containing lipoproteins entering the hepatocyte via the 'remnant receptor' does not influence expression of this receptor mechanism; it will, nevertheless, downregulate the hepatic LDL receptors. Defective uptake of LDL by the LDL receptor is the basis of familial hypercholesterolaemia (see below).

The other quantitatively important mechanism by which LDL cholesterol may enter cells is by a non-receptor-mediated pathway: LDL binds to cell membranes at sites other than the LDL receptors and some of it passes through the membrane by pinocytosis. High-density lipoprotein is able to compete with LDL for this type of cell membrane association. The absence of a receptor means that the 'binding' is of low affinity and thus, at low concentrations, LDL entry by this route may have little significance. However, unlike receptor-mediated entry, non-receptor-mediated LDL uptake, is not saturable, but continues to increase with increasing extracellular LDL concentrations. When LDL levels are relatively high, entry of cholesterol into the cells by this route may thus assume greater quantitative importance than that via the LDL receptor, which will be both saturated and downregulated. This appears to be the situation in the typical adult consuming a high-fat diet whose LDL cholesterol is high compared with most animals and in whom only about one-third of LDL is catabolized by receptors and two-thirds by non-receptor-mediated pathways. In hypercholesterolaemia, an even greater proportion of LDL is catabolized via the non-receptor pathway (four-fifths in patients heterozygous for familial

hypercholesterolaemia, virtually all in homozygotes; see below).

Low-density lipoprotein may also be removed from the circulation by receptors other than the classical LDL receptor. These are probably responsible for the catabolism of only relatively minor amounts of LDL, but two groups of receptors present on the macrophage have excited considerable interest, because they are pertinent to atherogenesis. They are the b-VLDL receptor, a modified LDL receptor which allows the uptake of the b-VLDL from patients with type III hyperlipoproteinaemia (see below), and the scavenger receptors and oxidized LDL (**ox-LDL**) receptors, which permit the uptake of oxidized LDL by macrophages. Uptake by these receptors is so rapid that foam cells resembling those in arterial fatty streaks and atheromatous lesions are formed *in vitro*. On the other hand, uptake of unmodified LDL by the macrophage via the LDL receptor is too slow to allow foam cells to be formed. Oxidation of LDL may occur *in vivo* and is of potential relevance to atherogenesis (see below).

Lipoprotein (a)

Lipoprotein (a) (**Lp(a)**) is a lipoprotein first identified as a result of blood transfusion reactions occurring due to allotypic variation. Its exact location in the LDL and HDL_2 also varies from individual to individual, as does its serum concentration. It may be undetectable in some people or present at concentrations equalling those of LDL in others. The protein moiety of Lp(a), like that of LDL, contains $apoB_{100}$, but in addition apolipoprotein (a) (**apo(a)**) is also present. This contains homologous sequences of plasminogen, in which part of the plasminogen protein sequence (the kringle 4 domain) is repeated many times. The number of these repeats, which is determined at a genetic locus adjacent to the plasminogen gene, determines the molecular mass of apo(a), and individuals expressing polymorphisms with fewer kringle 4 repeats have the highest serum concentrations of Lp(a). Lp(a) is associated with the risk of coronary heart disease in people of European origin, particularly when serum cholesterol levels are also raised and when there is a family history of premature coronary heart disease. Lipoprotein (a) does not give rise to fibrinolytic activity because of a mutation in its activation site, and it may interfere with thrombolysis. Furthermore because Lp(a) binds to many different cells and connective tissue matrices, it is retained in the arterial wall longer than LDL and is thus more likely to be oxidized and taken up by macrophages, leading to atheroma (see below).

Transport of cholesterol from tissues back to liver

Cholesterol is exported by the gut and liver in quantities which greatly exceed its peripheral catabolism (largely conversion to steroid hormones and sebum). Therefore, except when the requirement for membrane synthesis is high, for example during growth or active tissue repair, the greater part of the cholesterol transported to the tissues (if it is not to accumulate there) must be returned to the liver for elimination in the bile, as bile acids and faecal sterols, or for reassembly into lipoproteins. The return of cholesterol from the tissues to the liver is termed 'reverse cholesterol transport'. It is less well understood than the pathways by which cholesterol reaches the tissues but it may well be critical to the development of atheroma. High-density lipoprotein has many features that make it very likely that it is directly involved in the reverse transport process.

The precursors of plasma HDL (nascent HDL) are disc-shaped bilayers composed largely of protein and phospholipid secreted mainly by the gut and liver (Fig. 4(b)). These are converted to the spherical, mature form of HDL by the action of lecithin:cholesterol acyltransferase. High-density lipoprotein components are also derived from surplus material (phospholipids, free cholesterol, and apoproteins) of triglyceride-rich lipoproteins released during lipolysis. ApoA-I and apoA-II, the major apolipoproteins of HDL, and apoE have been identified in nascent HDL. Other apolipoproteins and the bulk of its lipid are acquired as it circulates through the vascular and other extracellular fluids. In this respect the transformation of HDL from its lipid-depleted precursor to a relatively lipid-rich molecule is the inverse of that undergone by the other lipoproteins following their secretion.

High-density lipoprotein is a small particle compared with the other lipoproteins (diameter 5–12 nm, density 1063–1210 g/litre) and easily crosses the vascular endothelium, so that its concentration in the tissue fluids is much closer to its intravascular concentration than is the case for LDL. Because the serum HDL cholesterol concentration is only about one-quarter that of the LDL, it is often wrongly assumed that its particle concentration is lower. In fact, the particle concentrations of HDL and LDL in human plasma are often similar, and in the tissue fluids there are several times as many HDL molecules as those of other lipoproteins unless the capillary endothelium is fenestrated. Generally, therefore, cells are in contact with higher concentrations of HDL molecules than of any other lipoprotein. In man, unlike the rat, HDL serves no function in transporting cholesterol to cells.

Recently it has been suggested that cells express receptors for HDL, particularly HDL_3, which might permit the transfer of cholesterol out of the cell. Passage across the cell membrane depends on an ATP-binding cassette transporter, ABCA1, which has recently been identified as the cholesterol efflux regulatory protein. Homozygosity for mutations in the *ABCA1* gene cause analphalipoproteinaemia (Tangier disease, see below) in which nascent HDL disappears rapidly from the circulation without acquiring cholesterol. Heterozygotes for *ABCA1* mutations have low levels of HDL cholesterol and accelerated atherogenesis. Factors regulating the balance between intracellular cholesterol ester and free cholesterol (activities of acyl-CoA:cholesterol *O*-acyltransferase and cholesterol esterase) may also be important. Apolipoprotein E synthesized within certain cells may also mediate the egress of cholesterol to HDL. Once outside the cell, free cholesterol must be re-esterified in order that it can be transported in any quantity in the core of lipoproteins. Therefore, whether or not HDL is involved as the initial acceptor molecule, cholesterol must at some stage on its return journey to the liver reside on HDL, because it is the site of lecithin:cholesterol acyltransferase activity. However, once cholesterol has been esterified and packed into the core of HDL, simple clearance of the whole lipoprotein particle by the liver is not the route by which most cholesterol is returned to it. This is because LDL equivalent to 1500 mg of cholesterol is produced each day, whereas the rate of catabolism of the HDL apolipoproteins A-I and A-II would permit less than 200 mg of HDL cholesterol to be returned each day. Therefore:

1. the liver must be capable of selectively removing cholesterol from HDL and then returning the particle to the circulation with most of its apolipoproteins intact, or
2. the cholesterol in HDL must be transferred to another lipoprotein class which is capable of being cleared in quantity by the liver, or
3. a class of HDL, which contains little apoA-I or apoA-II, must be cleared by the liver at a much greater rate than the bulk of HDL.

In support of pathway (1), there is some evidence that hepatic lipase might act on the phospholipid envelope of HDL during its passage through hepatic sinusoids, and release the cholesteryl ester contained in its core, and that some hepatic trapping or even a receptor-mediated mechanism might enhance the process. On the other hand, in support of pathway (2) there is a well-established mechanism for the transfer of cholesteryl ester from HDL to VLDL through the agency of cholesteryl ester transfer protein. Once on VLDL, the conversion of this lipoprotein to IDL and then LDL means that the cholesteryl ester can then arrive at the liver via remnant receptors, LDL receptors, or by the non-receptor-mediated route for LDL uptake. Evidence for pathway (3), the return of cholesterol to the liver from HDL by a rapidly metabolized form of HDL present at low concentration in serum, is at present lacking in man, although it is possible that binding of the subclass of HDL containing apoE to hepatic remnant receptors permits the return of some cholesterol to the liver.

It is incorrect to regard HDL as a single homogeneous species, since it is known to be a mixture of particles differing in size, in lipid and apolipoprotein composition, and in function. Two main species can be resolved by ultracentrifugation, the less dense of which is designated HDL_2 (density 1063–1125 g/litre) and the more dense HDL_3 (density 1125–1210 g/litre). HDL_3 may be converted to HDL_2 by the acquisition of cholesterol, HDL_3 thus being a precursor of HDL_2. Whereas antisera to apoA-I precipitate virtually all of HDL, antisera to apoA-II do not, suggesting that some molecules of HDL contain apoA-I and apoA-II, whereas others contain only apoA-I. The apoA-I-only HDL molecules, which predominate in HDL_2, may arise from different metabolic channels than do the apoA-I/A-II particles. High-density lipoprotein containing apoE, as has previously been mentioned, may also have a different metabolic fate. Furthermore, HDL may contain other molecular species with overlapping density ranges, such as Lp(a); it is thus a highly diverse lipoprotein class.

Disorders produced by raised concentrations of lipoproteins

The incidence of coronary heart disease varies greatly in different parts of the world. Those countries with a Northern European culture (and in particular diet) have the highest rates, and places such as China, Japan, and rural Africa the lowest. Mediterranean countries are intermediate. There are, of course, many differences between these countries, but the variable that relates most closely to coronary heart disease is the median serum cholesterol of the middle-aged male population. It is of considerable interest that in a country such as Japan, where the average serum cholesterol is low, other coronary risk factors do not seem to operate. Thus in Japan coronary heart disease is comparatively uncommon, even in cigarette smokers and people with diabetes and hypertension.

Within populations there is an exponential relationship between serum cholesterol and the incidence of coronary heart disease (Fig. 5). This depends on the LDL cholesterol which comprises some 70 to 80 per cent of the total cholesterol in men and a little less in women. The greater part of the residual cholesterol in serum is on HDL, and the concentration of this HDL cholesterol is inversely related to the likelihood of developing coronary heart disease.

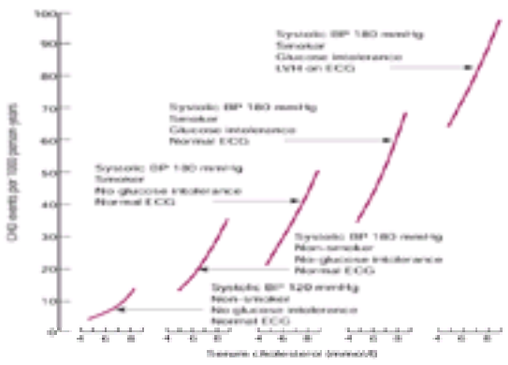

Fig. 5 The probability of 50-year-old men developing coronary heart disease each year as a function of serum cholesterol concentration, in the absence and in the presence of increasing numbers of risk factors. (Data from Kannel WB *et al.* (1973). The Framingham Study. An epidemiological investigation of cardiovascular disease. Section 28: the probability of developing certain cardiovascular diseases in eight years at specific values of some characteristis. Publication 74-618, US Department of Health Education and Welfare. Government Printing Office, Washington DC.)

In populations in which death from coronary heart disease is common, fatty streaks are evident in the arteries, such as the aorta, of men dying in their late teens of causes unrelated to cardiovascular disease. This was noted in American casualties of the Korean and Vietnam wars. The fatty streak is the precursor of atheroma (see Section 15). The epidemiological and histopatholgical evidence implicating LDL in atherogenesis seems overwhelming. Yet in tissue culture, LDL uptake by macrophages or smooth muscle cells proved disappointingly slow and foam cells were not formed. Subsequently, it was found that the macrophage has receptors which will allow the rapid uptake of LDL to form foam cells if the LDL has undergone some chemical modification. Initially these receptors were known as the acetyl-LDL receptors (after the first experimental chemical modification leading to their identification), but now scavenger receptors and another class of receptors, the ox-LDL receptors, are known to be responsible for the uptake of modified LDL by macrophages. It is likely that the chemical modification leading to LDL uptake in human atherogenesis is oxidation of the polyunsaturated fatty acyl groups of phospholipids of LDL which have crossed the arterial endothelium to enter the subintimal space. The lipid peroxides so formed break down to lysophospholipids and aldehydes, which directly damage the apoB of the LDL, which then binds to the scavenger and ox-LDL receptors. The same substances are directly cytotoxic and may further damage the overlying arterial endothelium, increasing its permeability. The oxidized LDL itself and the release of the cytokines it stimulates are chemotactic to blood monocytes (from which arterial wall macrophages are derived) and may thus recruit more inflammatory cells into the lesion. HDL may protect LDL against oxidative modification by a process which involves the enzyme paraoxonase, which is tightly bound to HDL. In addition to uptake of LDL through scavenger and ox-LDL receptors, macrophages phagocytose aggregated LDL to become foam cells and take up LDL–antibody complexes via Fc (immunoglobulin crystallizable fragment) receptors. Other lipoproteins can be taken up to form foam cells. In particular, the b-VLDL (a mixture of chylomicron remnants and IDL), which accumulates in the circulation in type III hyperlipoproteinaemias (see below), is rapidly taken up by a macrophage receptor.

Triglyceride-rich lipoproteins can also be taken up by macrophages by phagocytosis to form foam cells, although these large particles would not be expected to cross the vascular endothelium unless it is fenestrated. Thus in extreme hypertriglyceridaemia, lipid-engorged macrophages are present in the mononuclear phagocyte system. They may be observed on bone marrow biopsy, and are the cause of the hepatosplenomegaly associated with extreme hypertriglyceridaemia. When hypertriglyceridaemia occurs in association with elevated levels of LDL cholesterol, it increases the likelihood of atheroma developing still further, perhaps because this combination is associated with low serum HDL cholesterol, perhaps because of an increase in circulating IDL and delayed clearance of chylomicron remnants, perhaps because it is associated with smaller LDL particles, which are more readily oxidized, or perhaps because there are associated increases in the coagulability of blood due to increased plasma fibrinogen levels and factor VII activity. When, however, triglyceride-rich lipoproteins are increased without any increase in LDL, as in familial lipoprotein lipase deficiency (see below), there appears to be only a modest risk of atheroma. There is, however, an increased likelihood of acute pancreatitis in all types of severe hypertriglyceridaemia, both primary and secondary, particularly when serum triglyceride levels exceed 20 to 30 mmol/litre (2000–3000 mg/dl). The cause of this is not known for certain, but may be attributed to the release of fatty acids by lipolysis *in situ* due to pancreatic lipase.

Normal serum lipid concentrations

Whereas the average serum concentrations of most substances, for example sodium or fasting glucose, are much the same in all parts of the world, cholesterol displays considerable variation. In Britain the median serum cholesterol for a middle-aged man is 5.8 mmol/litre and deaths from coronary heart disease comprise around 40 per cent of total mortality at this age. In China the average for men of middle age is 2 mmol/litre less, and coronary heart disease accounts for less than 5 per cent of their deaths.

Conventionally, the normal range for a variable in a particular population is chosen to include values between the 2.5 and 97.5 percentiles, or sometimes the 1 and 99 percentiles, on the assumption that 19 out of 20 of the population, or 49 out of 50 respectively, are normal. To be rational, the implication in a medical context must also be that those people in the normal range are healthy. In the case of cholesterol, which is clearly linked to coronary heart disease, the healthy range would be more representative were it to include values from societies in which coronary heart disease is uncommon, such as China or Japan. This has led the National Institute of Health in the United States and the European Atherosclerosis Society to define healthy limits for serum cholesterol based on the risk of coronary heart disease. Thus an optimal serum cholesterol is 5.0 mmol/litre (200 mg/dl) or less. A level of 6.3 mmol/litre (250 mg/dl) (at the 75th percentile in the United States) is considered to indicate 'moderate risk' and 6.7 mmol/litre (270 mg/dl), which is the 90th percentile in the United States, 'high risk'. Some caution is required in using this concept. The risk of fatal coronary heart disease in an American middle-aged male population whose serum cholesterol is 5 mmol/litre (200 mg/dl) over the next 6 years is about 6 in 1000. At 6 mmol/litre (250 mg/dl) it is almost doubled, but that is only 10 in 1000, and at 7 mmol/litre (270 mg/dl) it is still less than 15 in 1000. Thus although these levels may be of great importance for public health initiatives aimed at reducing the cholesterol level in societies in which the risk of coronary heart disease is high, the clinician must be wary about overtreating men with cholesterol at these levels, if it is their only risk factor for coronary heart disease. The risk conferred by a particular level of cholesterol increases considerably when it is combined with another risk factor and this may considerably increase the benefits of treatment (Fig. 5). This is why there can be no single cholesterol level which demands a particular therapeutic response: the cholesterol value must always be viewed in the context of an individual's overall cardiovascular risk (see below).

An upper limit of normality for fasting serum triglycerides is often regarded as 2.2 mmol/litre (200 mg/dl). This is close to the 90th percentile for men and the 95th percentile for women. For serum HDL cholesterol a lower limit of normality of 0.9 mmol/litre (35 mg/dl) is frequently quoted, which is close to the 10th percentile for men and between the 5th and the 10th percentiles for women.

The Fredrickson/WHO classification

The concentration of four classes of serum lipoproteins when elevated can be regarded as pathological. These are chylomicrons, VLDL, LDL, and b-VLDL. The hyperlipoproteinaemias can be classified according to which of them is increased (Table 1).

The Fredrickson/WHO classification causes great confusion, largely because it is difficult to remember and is frequently wrongly regarded as a diagnostic classification when it is simply a way of reporting which of the serum lipoproteins is elevated. It is usually sufficient to remember that when cholesterol alone is elevated there is a type IIa hyperlipoproteinaemia. When both cholesterol and triglycerides are elevated the hyperlipoproteinaemia is generally type IIb, but occasionally it is type V (the serum will look milky if it is) and rarely type III. Type I is extraordinarily rare. An isolated increase in fasting serum triglycerides almost invariably signifies type IV hyperlipoproteinaemia.

All hospital laboratories, in addition to measuring cholesterol and triglyceride levels, should also measure HDL cholesterol in patients whose overall cardiovascular risk is being critically assessed when treatment of their hyperlipoproteinaemia with drugs is under consideration. Particularly in women, an elevated level of cholesterol may result from a relatively high HDL cholesterol concentration and thus not signify any increased risk of coronary heart disease. High serum HDL cholesterol does not have a Fredrickson/WHO class, but as evidence suggests it is associated with longevity, it cannot be regarded as hyperlipoproteinaemia in the pathological sense. It is low HDL cholesterol which is associated with an increased cardiovascular risk, particularly if total serum cholesterol and triglycerides are also elevated.

Primary hyperlipoproteinaemias

Primary hyperlipoproteinaemias in which there is hypercholesterolaemia (type IIa)

Serum cholesterol levels exceeding 5 mmol/litre are common in adults in Britain and much of Europe, the United States, Australia, and New Zealand. In Britain, for example, 80 per cent of middle-aged people have levels exceeding this, and the proportion in the United States is at least 50 per cent. Most of this hypercholesterolaemia does not represent the effect of any single cause, but is due to some combination of dietary fat, obesity, and individual susceptibility to develop hypercholesterolaemia. This susceptibility is partly genetic, probably involving more than one gene, and this common type of hypercholesterolaemia is usually referred to as polygenic hypercholesterolaemia. At the very top end of the cholesterol distribution are to be found individuals who have the less common monogenic condition, familial hypercholesterolaemia.

Familial hypercholesterolaemia

Heterozygous familial hypercholesterolaemia

Familial hypercholesterolaemia is dominantly inherited. The heterozygous form of the condition affects about 1 in 500 people in Britain and the United States, making it one of the most common genetic disorders in these countries. In some populations, such as the Lebanese Christians, the Afrikaner and Cape Coloured peoples of South Africa, and French Canadians, it is considerably more common. This is because such people have descended from a relatively small number of early settlers, a few of whom by chance had familial hypercholesterolaemia. This is known as a founder effect. In yet other populations, such as Africans who have not intermingled with Europeans, familial hypercholesterolaemia appears to be rare.

Typically, the serum cholesterol in adult heterozygotes is 9 to 11 mmol/litre (350–450 mg/dl). The condition is expressed regardless of diet or age, and elevated cholesterol levels are present throughout childhood. The lipoprotein phenotype is usually IIa, but occasionally there is a moderate increase in fasting serum triglycerides to produce a IIb pattern. There is a tendency for HDL cholesterol to be at the lower end of the range, particularly if triglycerides are elevated.

The clinical hallmark of familial hypercholesterolaemia is the presence of tendon xanthomas. These appear in heterozygotes from the age of 20 onwards. The most common sites for tendon xanthomas are in the tendons overlying the knuckles and in the Achilles tendons (Plate 1 and Plate 2). Less commonly, they may also be found in the extensor hallucis longus and triceps tendons, and occasionally other sites. It is also common to find subperiosteal xanthomas on the upper tibia where the patellar tendon inserts. The skin overlying tendon xanthomas is of normal colour and they do not appear yellow. The cholesteryl ester deposits are deep within the tendons. Tendon xanthomas feel hard because they are fibrotic. Indeed, it is not uncommon for those in the Achilles tendons to become inflamed from time to time, sometimes presenting as chronic Achilles tenosynovitis. More generalized tendinitis may follow rapid therapeutic reduction in serum cholesterol levels. Tendon xanthomas occur in only two disorders apart from familial hypercholesterolaemia, and these are so rare as not to pose any diagnostic difficulty. They are cerebrotendinous xanthomatosis, in which plasma cholestanol is elevated and deposited in tendons, and phytosterolaemia (b-sitosterolaemia), in which there is abnormal intestinal absorption of plant sterols, which are then deposited in tendons.

Corneal arcus is also a frequent occurrence in familial hypercholesterolaemia. When it occurs in adolescence or early adulthood it is more likely to be associated with familial hypercholesterolaemia than corneal arcus occurring in middle age or later. It is, however, not uncommon to encounter patients with familial hypercholesterolaemia who have florid tendon xanthomas but no arcus. It is thus not a very valuable physical sign. Xanthelasmata palpebrarum, although occurring with greater frequency and at a younger age in familial hypercholesterolaemia, affect only a minority of heterozygotes. Xanthelasmas are not specific for any particular type of hypercholesterolaemia and occur in polygenic hypercholesterolaemia, pregnancy, primary bilary cirrhosis, and hypothyroidism. They are also common in middle-aged women, often overweight, with no very marked increase in serum cholesterol, if any. They may run in families apparently independently of hypercholesterolaemia.

Identifying those heterozygous for familial hypercholesterolaemia as early as possible is important, because of their risk of coronary heart disease. Untreated, over half of affected men die before the age of 60 years. It is not uncommon for men to have their first myocardial infarction or develop angina in their thirties and occasionally even earlier. Some 15 per cent of women with familial hypercholesterolaemia die of coronary heart disease before the age of 60 years and the majority have symptomatic coronary disease by that age. Perhaps as many as 10 per cent of women have some evidence of cardiac ischaemia before their menopause. However, whereas it is exceptional for a man with familial hypercholesterolaemia to live to 70 without symptomatic coronary heart disease, almost a quarter of women do so. This largely explains why a family history of premature coronary heart disease is absent in as many as one-quarter of patients discovered to have familial hypercholesterolaemia on screening, or in men who are discovered to have familial hypercholesterolaemia when they present with a heart attack in early life: the condition has been inherited from their mother, who has herself not yet developed coronary symptoms. Most people with familial hypercholesterolaemia are not overweight and do not have risk factors for coronary heart disease other than hypercholesterolaemia and a family history of the premature disease. Those without a family history of premature coronary heart disease (approximately 25 per cent) will be missed in screening programmes for risk factors for coronary heart disease, in which cholesterol is only measured selectively.

Those patients with familial hypercholesterolaemia who develop coronary heart disease particularly early often come from families in which the affected members have all tended to develop coronary heart disease early. This may be because other genetic factors in the family predispose to coronary heart disease. Thus low serum HDL cholesterol and increased fasting triglycerides are associated with a worse prognosis. Serum lipoprotein (a) is increased in familial hypercholesterolaemia and any familial tendency to run a high level of Lp(a) is exacerbated in those members who also have familial hypercholesterolaemia. The apoE $_4$ genotype (see below) is also associated with more aggressive atheroma in familial hypercholesterolaemia. A knowledge of the average age at which affected members of a family developed coronary heart disease may be helpful in planning how actively to treat boys and young adult women.

There is an increased risk of atheroma in other parts of the arterial tree in heterozygous familial hypercholesterolaemia, but this is strikingly less so than in the coronary arteries. Some heterozygotes have aortic systolic cardiac murmurs due to deposits of atheroma in the aortic root, sometimes involving the aortic cusps.

Homozygous familial hypercholesterolaemia

Most cases of homozygous familial hypercholesterolaemia occur in societes in which consanguineous marriages and heterozygous familial hypercholesterolaemia are frequent. The chance of marriage between unrelated heterozygotes meeting by chance in countries such as the United Kingdom or United States is 1 in 500 ^{2}, and each of their children would stand a 1 in 4 chance of being homozygotes. Assuming no adverse effect on the survival of the conceptus, an incidence of homozygous familial hypercholesterolaemia of 1 in 10 ^{6} births would be predicted—a rare disorder.

Clinically, homozygous familial hypercholesterolaemia is characterized by the development of cutaneous xanthomas in childhood. These may be present in the first year of life or may not develop until late childhood. They are typically orange-yellow, subcutaneous, planar xanthomas, occurring on the buttocks, antecubital fossae, and the hands, frequently in the webs between the fingers. Tuberose subcutaneous xanthomas on the knees, elbows, and knuckles are also a feature. Serum cholesterol is typically greater than 15 mmol/litre (600 mg/dl). Myocardial infarction and angina frequently occur in childhood, sometimes even in infancy. Atheromatous deposits at the aortic root, invariably present by puberty, are so marked as to produce significant aortic stenosis, which contributes to the risk of sudden death. Death before the age of 30, and often considerably younger, was the rule before the advent of plasmapheresis and similar techniques for the extracorporeal removal of LDL (see below).

Polyarthritis, predominantly affecting the ankles, knees, wrists, and proximal interphalangeal joints, is common in homozygotes for familial hypercholesterolaemia.

The metabolic defect in familial hypercholesterolaemia

In familial hypercholesterolaemia there is decreased catabolism of LDL so that it remains for longer in the circulation. Normally the plasma half-life of LDL is 2.5 to 3 days, whereas in familial hypercholesterolaemia heterozygotes it is 4.5 to 5 days, and even longer in homozygotes. The molecular defect which causes this has been elucidated following the discovery of the LDL receptor (see above) by Goldstein and Brown, for which they received the Nobel prize for medicine in 1985. The gene encoding the LDL receptor protein is located on chromosome 19. Heterozygotes express only about half the LDL receptors of a normal person. Homozygotes have between none and 25 per cent of normal receptor activity. The mutations in the LDL receptor gene produce either receptors with no binding activity (receptor negative; because the receptor is not synthesized, is not transported to the cell surface, or, if it gets there, cannot be internalized after binding to LDL) or because, although the

mutation allows some LDL to be bound and to enter the cell, this occurs only slowly because the binding site is abnormal (receptor defective). Some 200 mutations have been described and undoubtedly more exist. In Afrikaners or French Canadians far fewer mutations are associated with familial hypercholesterolaemia. For example, three mutations account for 90 per cent of familial hypercholesterolaemia in Afrikaners. In societies such as Britain and the United States, however, the most frequent of these mutations is likely to occur in no more than 3 to 4 per cent of patients with familial hypercholesterolaemia. This means that the prospect of developing a DNA test for this condition in most countries is unrealistic. It also means that only patients in populations with a small number of mutations, or where intermarriage is common, are homozygous in the sense that both their LDL gene mutations are identical. Most will be compound heterozygotes. For clinical purposes it is reasonable to label as homozygotes patients who have the clinical syndrome. However, it is instructive to realize that some of the heterogeneity of the severity of the syndrome relates to the nature of the two LDL mutations present. Thus the worst prognosis is associated with inheritance of two receptor-negative mutations, and the best is with two receptor-defective mutations. The type of receptor mutation in heterozygotes is also probably of some importance, but here it is blurred against a background of other acquired or genetic factors, which can find expression over a much longer time than in homozygotes.

A small proportion (3 per cent) of patients who have the same clinical features as heterozygotes for familial hypercholesterolaemia do not have an LDL receptor defect but a mutation of apoB in which glutamine is substituted for arginine at amino acid residue 3500, which is part of the LDL receptor binding domain. This disorder has been termed familial defective apoB $_{100}$, and probably has a frequency of 1 in 500 to 600 in Britain and the United States. Only a minority of affected individuals have tendon xanthomas and typically the serum cholesterol associated with it is around 8.0 mmol/litre (310 mg/dl), which is less than in most heterozygotes for familial hypercholesterolaemia.

Common or polygenic hypercholesterolaemia

When a diagnosis of familial hypercholesterolaemia can be made, either because hypercholesterolaemia is present in childhood or an adult has the clinical features of the syndrome, a reasonably accurate estimate of clinical risk can be made and appropriate therapy given. In Britain, however, familial hypercholesterolaemia probably accounts for no more than 3 per cent of men dying of coronary heart disease before the age of 60. There is overlap between the range of LDL cholesterol levels encountered in familial hypercholesterolaemia and those due to the more common polygenic hypercholesterolaemia. Epidemiological studies have not included sufficient numbers of people with particularly high cholesterol levels to be certain, but it is probable that the risk in familial hypercholesterolaemia is greater than in polygenic hypercholesterolaemia. This may be because in the familial condition the hypercholesterolaemia has been present since birth, whereas polygenic hypercholesterolaemia is frequently not fully developed until the third or fourth decade. Furthermore familial hypercholesterolaemia, unlike many other types of hypercholesterolaemia, is associated with increased serum concentrations of Lp(a).

Estimates of how much different levels of cholesterol contribute to the overall cumulative male mortality from coronary heart disease by the age of 60 years are given in Table 2. The majority of such premature deaths come from the middle part of the cholesterol distribution, and therefore it has been argued that if a significant reduction in the incidence of coronary heart disease is to be achieved in countries such as the United Kingdom, efforts to lower cholesterol cannot simply be confined to those individuals whose plasma cholesterols lie at the upper end of the distribution. Nevertheless because the number of people in the middle range is so large (the vast majority of whom are not at increased risk of premature coronary heart disease), a different strategy must be applied to reducing their cholesterol from that applied to those in the upper part of the cholesterol distribution. This is the 'low-risk' or 'population' strategy, which aims to lower serum cholesterol by public health measures aimed at encouraging the adoption of a lower-fat diet and avoidance of obesity. Some patients from the middle range of serum cholesterol are, however, at much greater individual risk from their cholesterol level than the majority, because they have other risk factors for coronary heart disease which combine to increase their susceptibility. Probably the most potent of these is that the individual already has coronary heart disease.

In middle-aged survivors of myocardial infarction, serum cholesterol is an important indicator of cardiac prognosis (Fig. 6(a)), ranking after left ventricular function, but ahead of most of the other risk factors for coronary heart disease. Lipoproteins are also the most important risk factors for occlusion of coronary artery bypass grafts after the initial postoperative period. In people who have not yet developed coronary heart disease, the effect of risk factors such as cigarette smoking, hypertension, and diabetes synergizes with the risk from any given level of cholesterol (Fig. 5). A family history of coronary heart disease at an early age in a first-degree relative also increases the likelihood of coronary heart disease, and part of this effect is independent of other risk factors for coronary heart disease. The combination of all these factors with a relatively modestly increased serum cholesterol level can increase individual risk substantially to a level where clinical intervention is as justified as it is with more marked elevations in serum cholesterol. This is the 'high-risk' or clinical approach to prevention of coronary heart disease.

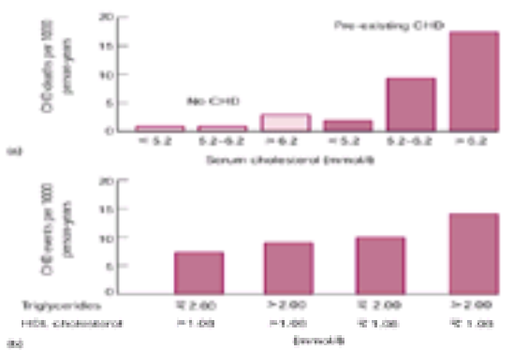

Fig. 6 (a) The risk of subsequent fatal myocardial infarction in survivors of myocardial infarction according to their serum cholesterol concentration. (Data from Pekkanan J *et al.* (1990) *New England Journal of Medicine* **322,** 1700–7.) (b) The likelihood of developing coronary heart disease in patients with moderately raised serum cholesterol concentrations (on average 6.9 mmol/litre) is increased when serum triglyceride levels are also raised and HDL cholesterol concentrations decreased. (Data from Manninen V *et al.* (1989).)

Metabolic defect in polygenic hypercholesterolaemia

In polygenic hypercholesterolaemia there is overproduction of VLDL by the liver. If this is rapidly converted to LDL there is no increase in serum triglyceride levels. The LDL receptor mechanism is probably overloaded in many individuals and in any case appears to catabolize only about one-third of LDL, so that the build-up of cholesterol in most patients is not due to any defect in the LDL receptor, but the inability of non-receptor-mediated catabolism to cope without a rise in the serum cholesterol concentration. Obesity and a high-fat diet (particularly saturated fat) are probably the major reasons for the enormous differences in the prevalence of polygenic hypercholesterolaemia in different parts of the world. Undoubtedly, however, individual responses to diet vary greatly and there is a complex interplay between dietetic and genetic factors in the genesis of polygenic hypercholesterolaemia. The rise in cholesterol with age, which occurs in both men and women until the climacteric, seems less evident in societies where the cholesterol level is, for nutritional reasons, lower. There is an impression that dietary modification aimed at lowering cholesterol in middle age in societies where serum cholesterol is high does not reduce it to the extent that might be anticipated from populations habitually consuming such a diet. Whether this is simply a matter of non-compliance with diet or represents some imprinted change in metabolism caused by a high-fat diet in early life is, at present, uncertain.

Primary hyperlipoproteinaemias in which there is hypercholesterolaemia combined with hypertriglyceridaemia

Type III hyperlipoproteinaemia

Type III hyperlipoproteinaemia has several synonyms: broad beta disease, floating beta disease, dysbetalipoproteinaemia, and remnant removal disease. It is rare, probably occurring in fewer than 1 in 5000 people. Type III hyperlipoproteinaemia has the distinction of being the first clinical syndrome associated with hyperlipoproteinaemia to be described (by Addison and Gull in 1851).

Type III hyperlipoproteinaemia is due to the presence in the circulation of increased amounts of chylomicron remnants and IDL, often collectively termed b-VLDL. This is the result of decreased clearance of these lipoproteins at the hepatic 'remnant' (or apoE) receptor. There is an increase in both the serum cholesterol and fasting triglyceride concentrations. Typical levels are 7 to 12 mmol/litre (270–470 mg/dl) for cholesterol and 5 to 20 mmol/l (450–1800 mg/dl) for triglycerides. Often the molar concentrations of cholesterol and triglycerides are similar, and this may be a clue that a patient has type III hyperlipoproteinaemia. Occasionally the condition is

associated with marked hypertriglyceridaemia due to overwhelming chylo-micronaemia.

Xanthomas are present in more than half of the patients who have the type III lipoprotein phenotype. Characteristic of the condition are striate palmar xanthomas and tuberoeruptive xanthomas. Striate palmar xanthomas may simply be an orange-yellow discoloration within the creases of the skin of the palms of the hands. They may, however, be more florid and appear as raised, seed-like lesions (sometimes even larger) in the skin creases of the palms, fingers, and flexor surfaces of the wrists. Tuberoeruptive xanthomas are raised yellow lesions, usually on the elbows and knees (Plate 3). They may be nodular or cauliflower like, often surrounded by smaller satellites. Sometimes they may be found over other tuberosities, such as the heels and dorsum of the interphalangeal joints of the fingers. They resolve entirely with successful treatment of the hyperlipidaemia.

Type III hyperlipoproteinaemia is rare in women before the menopause, perhaps because uptake of hepatic remnant particles is enhanced by oestrogen. It is also rare in childhood, but has a definite incidence in men by early adulthood. Type III hyperlipoproteinaemia is generally an autosomal recessive condition with variable penetrance. In all cases there appears to be a mutation or polymorphism of the apoE gene, which impairs the receptor binding of apoE. The most frequent is a polymorphism, called apoE_2, in which cysteine is substituted for arginine at position 158 of the amino acid sequence. At least 90 per cent of patients with type III hyperlipoproteinaemia are homozygous for apoE_2. More often than not, however, apoE_2 homozygosity, which is present in around 1 per cent of the population, does not itself impose such a severe strain on lipoprotein metabolism that hyperlipoproteinaemia develops: its combination with some other disorder, leading to overproduction of VLDL or some additional catabolic defect, is required. This explains the association of type III hyperlipoproteinaemia with diabetes and hypothyroidism. More often, however, the additional stimulus to hyperlipoproteinaemia is obesity or the coinheritance of a polygenic tendency to hypertriglyceridaemia. Rarer mutations of apoE have been described, which behave clinically similarly to apoE_2 homozygosity. More severe is a mutation leading to apoE deficiency, which in homozygotes does not require other factors for the expression of the type III phenotype. Heterozygous apoE deficiency finds little clinical expression, but, interestingly, mutations directly involving the receptor-binding domain of apoE (amino acids 124 to 150) produce the type III phenotype even in heterozygotes (dominant expression), implying that such mutations are a greater handicap to receptor clearance than mutations in which one gene is not producing apoE.

Type III hyperlipoproteinaemia undoubtedly causes accelerated atherosclerosis in the coronary, femoral, and tibial arteries. Intermittent claudication occurs at least as frequently as coronary heart disease and the incidence of the latter is about the same as that in familial hypercholesterolaemia. It is noteworthy that in familial hypercholesterolaemia peripheral arterial disease is uncommon relative to coronary heart disease, indicating that the leg arteries are much more susceptible to the larger lipoprotein particles in type III hyperlipoproteinaemia.

In the presence of typical xanthomas, the diagnosis of type III hyperlipoproteinaemia is not difficult. When these are absent the diagnosis must be made in the laboratory. Type IIb or V hyperlipoproteinaemia can give similar serum lipid levels. Lipoprotein electrophoresis is still available in some hospital laboratories and, when it clearly shows separate pre-beta (VLDL) and beta (LDL) bands, is useful in establishing type IIb rather than III hyperlipoproteinaemia. Frequently, however, the classical broad beta band associated with type III hyperlipoproteinaemia cannot be distinguished from a smear stretching from the origin into the pre-beta and sometimes beta region in the more severe IIb or type V phenotype. However, polyacrylamide isoelectric focusing or genotyping, available in many specialized centres, can identify apoE_2 homozygosity and this, in the presence of hyperlipidaemia, makes type III virtually certain. Rarely, the apoE mutation does not affect the electrical charge of apoE, or affects only one gene so that apoE_2 homozygosity is not found. The only way then to confirm the diagnosis is to send plasma to a centre that can provide ultracentrifugation to identify the cholesterol-rich VLDL (b-VLDL) typical of type III hyperlipoproteinaemia. It is also important in these circumstances to exclude paraproteinaemia, which can produce both hyper- and hypolipoproteinaemia and can mimic typical type III hyperlipoproteinaemia.

Type IIB hyperlipoproteinaemia

The common lipoprotein phenotype associated with a combined increase in serum cholesterol and triglycerides is IIb. In the majority of people with this, in whom it is primary, the cause is probably best regarded as a polygenic tendency exacerbated by acquired nutritional factors, such as obesity. A few patients will have tendon xanthomas, indicating familial hypercholesterolaemia (see above) but the great majority will not. Cardiovascular risk is greater for any given level of cholesterol when the serum triglyceride concentration is also elevated (Fig. 6(b)). Often the HDL is low, which further compounds the risk. In addition patients with hypertriglyceridaemia frequently have increased levels of a cholesterol-depleted small, dense LDL which contributes little to the total serum cholesterol concentration, but which is susceptible to oxidation and to which increasing attention is being paid because it may be highly atherogenic. Some authorities also believe that there is a specific syndrome in which there is a combined increase in serum cholesterol and triglycerides and a greatly increased coronary risk. They term this familial combined hyperlipidaemia. In this, multiple lipoprotein phenotypes occur in different family members: some IIa, some IIb, some IV, or occasionally even V. It is more than probable that what is being observed is the genetic tendency for hypercholesterolaemia and hypertriglyceridaemia running in the same family to combine in some members and not in others, and that when this occurs in a family susceptible to coronary disease, a particularly high premature mortality ensues. However, until the arguments about whether familial combined hyperlipidaemia is a distinct genetic entity are resolved, for practical purposes hypertriglyceridaemia (especially when HDL cholesterol is low) should be considered as an additional factor increasing the risk of hypercholesterolaemia. When these abnormalities are combined with a family history of premature coronary heart disease, there is a greatly increased risk of cardiovascular disease unless the condition is detected and treated.

Primary hyperlipidaemias in which hypertriglyceridaemia predominates

Severe hypertriglyceridaemia (types I and V)

Diagnosis and underlying mechanism

In any circumstance in which the serum triglycerides exceed 11 mmol/litre (1000 mg/dl) chylomicrons in addition to VLDL will be major contributors to the hyperlipidaemia, even when the patient is fasting. This is because in the circulation both chylomicrons and VLDL compete for the same clearance mechanism (lipoprotein lipase). The lipoprotein phenotype is usually type V. This severe hypertriglyceridaemia generally ensues when an increase in hepatic VLDL production, either familial or secondary to, for example, obesity, diabetes, alcohol, or oestrogen administration, is associated with decreased triglyceride clearance, which again may be genetic or acquired, for example hypothyroidism, beta blockade, or diabetes mellitus (diabetes can cause both an overproduction of VLDL and decreased lipoprotein lipase activity). With the clearance mechanism already overloaded with VLDL, the postprandial elevation in serum triglyceride concentrations when chylomicrons enter the circulation may be astronomical and they may spend days rather than hours in the circulation. The plasma takes on the appearance of milk and triglycerides may exceed 100 mmol/litre (9000 mg/dl) (Plate 4). Thus a patient, who might otherwise have a fasting serum triglyceride level of 5 mmol/l, can, with the injudicious use of alcohol or the development of intercurrent diabetes, achieve extraordinarily high serum triglyceride levels. Overall the frequency of severe hypertriglyceridaemia (> 11 mmol/litre (1000 mg/dl)) is probably no more than 1 in 1000 in adults and less in children.

Rarely, severe hypertriglyceridaemia is caused by familial lipoprotein lipase deficiency, a genetic deficiency in lipoprotein lipase activity. This is inherited as an autosomal recessive trait. Usually it is due to mutation in the lipoprotein lipase gene, leading to defective function or production, but occasionally it is due to a genetic deficiency of apoC-II, the activator of lipoprotein lipase. In familial lipoprotein lipase deficiency, severe hypertriglyceridaemia may be encountered in childhood. Occasionally, in children and young adults presenting for the first time, it produces type I hyperlipoproteinaemia, in which only serum chylomicron levels are elevated. It is not known why the VLDL is not also raised, but with advancing age the increase in both VLDL and chylomicrons, which might be expected if lipoprotein lipase is ineffective, becomes the rule.

Physical signs in severe hypertriglyceridaemia

Tuberoeruptive xanthomas are characteristic of extreme hypertriglyceridaemia. These appear as yellow papules on the extensor surfaces of the arms and legs, buttocks, and back. Often there is hepatosplenomegaly. Liver imaging shows the liver to be fatty, and bone marrow biopsy may reveal macrophages engorged with lipid droplets (foam cells). Because the triglyceride-rich lipoprotein may interfere with the determination of transaminases, giving spuriously high values, liver disease, in particular alcoholic liver disease, may be difficult to exclude, other than by the prompt resolution of the syndrome when a low-fat diet is instituted. Other features include lipaemia retinalis (pallor of the optic fundus, with both the retinal veins and arteries appearing white).

Complications of severe hypertriglyceridaemia

Atheroma is not a prominent complication of familial lipoprotein lipase deficiency, but it does complicate severe hypertriglyceridaemia in which there is residual lipoprotein lipase activity. It is difficult to make a precise estimate of the risk from the hyperlipidaemia *per se* because it is so frequently associated with insulin resistance or frank diabetes, which are themselves risk factors for atherosclerosis. If these are included as part of the syndrome, both coronary heart disease and

peripheral arterial disease are common. The explanation for the only modest risk of atheroma in patients lacking lipoprotein lipase is not understood but it may be because the incidence of diabetes is not increased in familial lipoprotein lipase deficiency. Also fibrinogen and factor VII activity are not increased; it is also notable that the conversion of VLDL and chylomicrons to the atherogenic IDL and remnant lipoproteins, respectively, is impaired in the absence of lipoprotein lipase.

Although atheroma may not be directly due to the high levels of triglyceride-rich lipoproteins, other complications are: acute pancreatitis may occur when serum triglyceride levels exceed 20 to 30 mmol/litre (2000–3000 mg/dl) (see above). The presentation of acute pancreatitis is similar to that from other causes (see Chapter 14.18.3.1). However, the diagnosis may not be confirmed by detecting increased serum amylase activity, because falsely low values may be encountered due to interference by triglyceride-rich lipoproteins in the laboratory method. All laboratories should inspect plasma or serum samples for milkiness before reporting normal or only moderately raised serum amylase activity in patients with severe abdominal pain (Plate 4). Clinicians may otherwise wrongly exclude the diagnosis of acute pancreatitis in favour, for example, of perforated peptic ulcer. Some patients do not develop acute pancreatitis, even when serum triglyceride levels exceed 100 mmol/litre (9000 mg/dl). Others, who are more susceptible, experience recurring acute episodes. Generally the pain subsides within a few hours or days of commencing nasogastric aspiration and intravenous fluids with nothing taken by mouth. Occasionally, if such treatment is delayed, pancreatic pseudocysts or abscesses may develop. Recurrent abdominal pain, not typical of pancreatitis, sometimes occurs in patients prone to marked hypertriglyceridaemia. It may mimic irritable bowel syndrome. Severe abdominal pain may also sometimes be the result of splenic infarction.

Pseudohyponatraemia is another consequence of extreme hypertriglyceridaemia, which may lead to serious misdiagnosis if the artefact is unrecognized. Spuriously low serum sodium values are reported, because much of the volume of the serum aliquot on which the sodium measurement is made is occupied by lipoproteins rather than water. When the serum triglycerides exceed 40 to 50 mmol/litre (3500–4500 mg/dl) the concentration of sodium in the aqueous phase (and thus the serum osmolality) may be normal while spurious serum sodium levels of 120 to 130 mmol/litre are being reported. The hazard is that these will be misinterpreted by the clinician and a patient already seriously ill with pancreatitis, or occasionally uncontrolled diabetes, will be made worse by restricting fluid intake or the infusion of hypertonic saline.

Focal neurological syndromes such as loss of vision, hemiparesis, memory loss, and loss of mental concentration may complicate extreme hypertriglyceridaemia, perhaps because of ischaemia due to sluggish microcirculation caused by the high concentrations of chylomicrons in the blood. Uniocular visual loss due to occlusion of the retinal microcirculation may likewise complicate hypertriglyceridaemia and is an indication for rapid institution of lipid-lowering therapy, and possibly antiplatelet agents. Paraesthesiae, especially in the feet, may also be an occasional feature, even in the absence of diabetes. Sicca syndrome and polyarthritis have also been described, but undoubtedly the commonest articular association is gout (see below).

Moderate hypertriglyceridaemia (type IV)

Raised fasting serum triglyceride levels in the range 2.2 to 10.0 mmol/litre (200–900 mg/dl) in the absence of a cholesterol level exceeding 5.0 mmol/litre (200 mg/dl) are occasionally discovered. Diabetes and excess ingestion of alcohol are important causes. Sometimes marked hypertriglyceridaemia is present in a fit, non-obese person with none of these factors. Family studies may then reveal similar increases in relatives, when the condition is called familial as opposed to sporadic hypertriglyceridaemia. Epidemiological studies show a univariate association between plasma triglyceride concentration and the risk of coronary heart disease, but there is little evidence that triglycerides are directly causal. Hypertriglyceridaemia is associated with low levels of HDL and glucose intolerance. When patients with established coronary disease and hypertriglyceridaemia whose serum cholesterol does not exceed 5.0 mmol/litre (200 mg/dl), are encountered they generally have low levels of HDL cholesterol and may have an increased level of cholesterol-depleted, small, dense LDL which is not evident from their cholesterol level. Such patients are likely to benefit from lipid-lowering therapy. They also have a greatly increased risk of developing diabetes mellitus over the next few years.

Hypertriglyceridaemia increases the risk of any associated increase in serum cholesterol (Fig. 6(b)), but present evidence would not favour its treatment in the absence of hypercholesterolaemia as a means of primary prevention of coronary heart disease. Occasionally, triglyceride concentrations of 5 mmol/litre (450 mg/dl) or less must be treated if they occur in patients prone to periodic exacerbations of more severe hypertriglyceridaemia associated with acute pancreatitis. Generally, levels exceeding 10 mmol/litre justify therapy, but for levels between 5 and 10 mmol/litre individual judgement should apply. In diabetes, evidence that serum triglycerides are an independent risk factor for coronary heart disease has been considered by some authorities to justify lipid-lowering therapy at lower levels than in non-diabetics when improvements in diet and glycaemic control have failed to decrease hypertriglyceridaemia.

Secondary hyperlipoproteinaemias

Secondary hyperlipoproteinaemias are those which are caused by another primary disorder (Table 3). When a disease that has hyperlipidaemia as a complication occurs in an individual who has a primary hyperlipoproteinaemia, the two frequently synergize to produce markedhyperlipoproteinaemia. This means that in societies in which poly-genic hyperlipoproteinaemia is prevalent, secondary hyperlipoproteinaemia will have most impact. The best-known example of this is diabetes mellitus, which in Japan is only rarely complicated by coronary heart disease, whereas in the United Kingdom and the United States, coronary heart disease is the most common cause of premature death in diabetics.

Diabetes mellitus

The dominant hyperlipidaemia in diabetes is hypertriglyceridaemia. This is more likely to be associated with hypercholesterolaemia and with decreased HDL cholesterol in type 2 diabetes. Despite this, the risks of coronary heart disease and peripheral arterial disease are increased in both types 1 and 2 diabetes. This may be because in both disorders the hypertriglyceridaemia results not simply in an increase in VLDL, but also from an increase in IDL and a small triglyceride-rich, cholesterol-depleted LDL particle. Since neither of these may contribute greatly to an increase in lipids, the term dyslipoproteinaemia is particularly aptly applied in diabetes. Also, plasma fibrinogen levels, which are increased in both types of diabetes, relate to serum triglyceride levels. Although lipoprotein abnormalities in type 1 diabetes may be less frequent than in type 2, the risk of coronary heart disease in type 1 is more often compounded by the presence of proteinuria. In diabetes uncomplicated by proteinuria, the risk of coronary heart disease is about two to three times (Fig. 5) that of non-diabetic people of a similar age. Proteinuria increases the risk by as much as 40 times. This may stem partly from hypertension and an exacerbation of the dyslipoproteinaemia, both of which may reflect the development of proteinuria. However, the increase in risk is greater than can be explained in this way (see Chapter 11.11) and may result because the proteinuria reflects a generalized increase in the permeability of arterial endothelium, enhancing the entry of macromolecules into the subintima and thus accelerating atherogenesis (see above).

The increased blood glucose in diabetes mellitus results from insulin resistance, insulin deficiency, or both. Insulin resistance may be present in non-diabetic, usually obese, people who are still able to secrete sufficient insulin to maintain control of blood sugar, but in such people there is often hypertriglyceridaemia with low HDL cholesterol and hypercholesterolaemia, hypertension, and increased risk of coronary heart disease. This syndrome is often referred to as the insulin resistance syndrome (syndrome X) or chronic cardiovascular risk syndrome. Clearly it has features in common with familial combined hyperlipidaemia and also with diabetes. Indeed, a proportion of people with syndrome X ultimately develop diabetes, sometimes not until after they have already developed coronary heart disease. This may explain in part why glycaemic control in diabetes seems to have little impact in preventing its atheromatous complications.

Diabetic women, particularly those with type 2 disease, tend to have a distribution of adipose tissue resembling that of obese men, being mostly around the abdomen and waist rather than the more female pattern which involves the buttocks and thighs, but leaves the waist relatively small. The relative protection from coronary heart disease which most women have, even those with familial hypercholesterolaemia, is largely lost by diabetic women, and it has been suggested that this may result from this androgenization. Many women with a similar body habitus, but who have not yet developed diabetes, are insulin resistant, hypertensive, have hyperlipidaemia, and have an associated increased risk of coronary heart disease.

Other secondary hyperlipoproteinaemias

Obesity

Obesity is a potent cause of hyperlipidaemia and has most impact in people with glucose intolerance. In its own right, obesity predominantly causes hypertriglyceridaemia (usually type IV), but, there is no form of primary hyperlipidaemia that it will not exacerbate. It therefore frequently accompanies hypercholesterolaemia as well as hypertriglyceridaemia. The exception appears to be familial hypercholesterolaemia, which is not associated with obesity. Alcoholic beverages, particularly wine and beer, are energy rich and may be a cause of obesity. Alcohol itself also induces hypertriglyceridaemia. Weight loss is generally associated with decreases in serum cholesterol and triglyceride levels. Anorexia nervosa is paradoxical in that it may be associated with quite marked elevations of

serum cholesterol.

Thyroid failure

In hypothyroidism, serum LDL cholesterol and, less frequently, serum triglycerides are raised. Levels of HDL tend to be increased. There is decreased receptor-mediated LDL catabolism and lipoprotein lipase activity may be decreased. Hypothyroidism should always be considered in the diagnosis of hyperlipidaemia, and it is particularly important to exclude it when marked hyperlipidaemia occurs in women and in diabetic patients.

Renal disease

Renal disease is becoming an important cause of secondary hyperlipidaemia in clinical practice, because improvements in long-term renal management are now exposing coronary heart disease as the major cause of premature death in many renal disorders. In nephrotic syndrome the major lipoprotein disorder is a rise in serum LDL cholesterol. In chronic renal failure hypertriglyceridaemia is produced by an increase in both VLDL and in LDL triglycerides. Haemodialysis, chronic ambulatory peritoneal dialysis, and high-energy diets exacerbate the hyperlipidaemia. Following renal transplantation, many of the lipoprotein abnormalities resolve if good renal function is established, but corticosteroid therapy, weight gain, antihypertensive therapy, and perhaps cyclosporin treatment mean that even then hyperlipidaemia persists in about one-quarter of patients. Lipoprotein (a) is markedly elevated in renal disease, even after transplantation.

Drugs

Drugs are a common cause of hyperlipidaemia. b-adrenergic antagonists without intrinsic sympathomimetic activity raise triglycerides and lower HDL cholesterol. Thiazide diuretics tend to increase both cholesterol and triglycerides. These effects may be relatively small in people whose serum lipids are not elevated at the outset, but in patients with hypertriglyceridaemia or with diabetes they may be substantial. Oestrogens tend to raise serum triglycerides, but will often lower LDL cholesterol after the menopause. They also raise serum HDL. Androgens have the opposite effect, decreasing triglycerides, raising LDL cholesterol, and lowering HDL. They may contribute to premature cardiac death in athletes unwise enough to use them in training. Glucocorticoids increase serum LDL cholesterol and triglycerides and often HDL cholesterol. Retinoic acid derivatives used in the management of skin disorders cause hypertriglyceridaemia. Phenytoin and phenobarbitone raise serum HDL cholesterol.

Liver disease

Cholestatic liver diseases, such as primary biliary cirrhosis, produce hypercholesterolaemia. This is not due to an increase in apolipoprotein B-containing LDL, but to an abnormal lipoprotein, designated lipoprotein X (**LpX**), produced largely as the result of reflux of biliary phospholipids into the circulation. Xanthelasmas are common in biliary obstruction and other xanthomas occasionally develop. In the later phase of chronic biliary obstruction, when secondary biliary cirrhosis and hepatocellular disease sets in, hepatic lipid biosynthesis plummets and the hyperlipidaemia of biliary obstruction resolves. Hepatocellular diseases may be associated with moderate hypertriglyceridaemia, probably because of impaired hepatic lipoprotein clearance. Concentrations of HDL are markedly decreased and lecithin:cholesterol acyltransferase activity is low. Some authorities believe that this defect in cholesterol esterification contributes to the complications of liver failure.

Hyperuricaemia

Hyperuricaemia is present in as many as half the men with hypertriglyceridaemia. It may lead to gout, particularly if such patients are receiving diuretic therapy. The association of hypertriglyceridaemia and hyperuricaemia appears to be more common than can be entirely explained by the coincidence of common aetiological factors, such as obesity and high alcohol consumption. Yet they are not causally related, because specifically lowering one does not usually decrease the other. They must therefore have some unknown antecedent in common.

Management of hyperlipoproteinaemia

Clinical trials have established beyond all question of doubt that reduction of serum cholesterol decreases both coronary morbidity and mortality and can prolong survival. The risk of coronary heart disease ascribable to a particular cholesterol level does, however, vary widely in different individuals depending on the presence of other risk factors for coronary heart disease. Thus a serum cholesterol of 6.0 mmol/litre (230 mg/dl) in a 50-year-old woman with an HDL cholesterol level of 1.9 mmol/litre (75 mg/dl) who does not smoke and is neither hypertensive nor diabetic will carry a risk of a coronary event of 1 in 40 over the next 10 years, whereas in a man of similar age the same serum cholesterol associated with an HDL cholesterol value of 0.9 mmol/litre (35 mg/dl) who is hypertensive, smokes and has diabetes the risk will be 1 in 3 over the same time interval. His likelihood of benefit from a given reduction in cholesterol will thus be much greater than hers, although both have the same concentration of serum cholesterol. The coronary risk attaching to the cholesterol level in individual patients could, were it known with reasonable accuracy, thus guide the clinician in deciding how rigorously treatment should be given. Below a certain level of risk, treatment may be more trouble than it is worth for the patient's lifestyle, presence of mind, or pocket. For a state healthcare system too there will also be a level of risk below which the cost-effectiveness of cholesterol lowering may mean that resources should be directed to some more cost-effective clinical practice.

Another consideration, as with any therapeutic intervention, is that there may be side-effects of treatment which should limit it to those patients whose risk of the disease it is intended to prevent (in this case primarily coronary heart disease) is substantially higher than the potential risk of serious side-effects. Dietary management is generally viewed as safe, and meta-analyses of dietary trials show no increase in non-cardiac mortality. Until recently, however, cholesterol-lowering drugs were often viewed with suspicion. Since 1994, however, results from six clinical trials of statin drugs have established that these drugs are safe with adequate medical supervision during the 5 to 6 years of the trials. In that time the risk of coronary heart disease was decreased by one-third, the decrease becoming greater with more prolonged therapy. The lowest average annual coronary risk of participants in these trials was 1 per cent (one event per 100 people per year) and the highest 4.5 per cent. The relative reduction in risk of one-third was similar regardless of the level of risk, so a greater number of coronary events was prevented when the risk was highest. Other cholesterol-lowering drug therapies such as fibrates and bile acid sequestrating agents decrease coronary risk, but their safety and the magnitude of their overall benefit is not as clear as with statins, partly because design and analysis of clinical trials were better in the more recent statin trials. In a recently reported trial of men with established coronary disease and relatively low levels of serum cholesterol and HDL cholesterol, gemfibrozil was found to be both safe and effective. Other fibrate trials are under way.

The essential point to grasp is that the decision to treat hyperlipidaemia is not based simply on any particular cholesterol value, but on an assessment of individual risk of coronary heart disease. It is sensible to select for treatment those patients with a high overall probability of dying prematurely of coronary disease. If the balance of risk suggests that they are not, they will be exposed to any possible ill-effects of such treatment with no likelihood of benefit. The identification of patients with established coronary heart disease, familial hypercholesterolaemia, or with more modest increases in serum cholesterol combined with multiple risk factors, including a bad family history (Fig. 5 and Fig. 6 ((a) and (b)) allows the targeting of cholesterol-lowering management to high-risk individuals, who can benefit most. Charts which can assist in the assessment of coronary risk are to be found in Section 15.16.1.

Dietary management

It is generally agreed that dietary advice should be given to people whose serum cholesterol exceeds a concentration of 5.0 mmol/litre (200 mg/dl). In a country such as the United Kingdom, however, two-thirds of men and women between the ages of 18 and 69 years have cholesterol concentrations exceeding this value. Thus, except in the case of the patients considered to be at high coronary risk, individual dietetic supervision beyond the provision of a diet sheet is not reasonable. It is particularly important to remember that cigarette smoking is a greater cause of ill-health than are minor elevations of serum cholesterol, and advice to stop smoking should be reiterated whenever a medical consultation occurs.

The principal aims of a cholesterol-lowering diet are to reduce obesity by a decrease in dietary energy intake and to decrease saturated fat consumption. Fat is a major source of dietary energy and the reduction in its intake should be the main objective of any weight-reducing diet. In the non-obese, dietary advice should focus on decreasing saturated fat to below 10 per cent of dietary energy intake and substituting it with a mixture of unrefined carbohydrate and monounsaturated and polyunsaturated fats (Table 4). Polyunsaturates in the form of linoleic acid (corn oil, sunflower oil) should not be the only fats to replace saturated fat, because it is not certain that in large amounts they do not have harmful long-term effects. In patients with established coronary disease there is increasing interest in the long-chain omega-3 fatty acids such as those found in fish oil (Table 4) which are more unsaturated and reduce sudden cardiac death, probably by suppressing ventricular arrhythmias. Eating fatty fish twice a week is thus recommended. Increasingly, too, oils rich in monounsaturated oleic acid such as olive oil, present in the diet of Mediterranean people since time immemorial, are being encouraged by nutritionists as substitutes for saturated fat. Rapeseed oil, which is much cheaper, contains

almost as much oleic acid as olive oil. Dietary cholesterol itself, although featuring prominently on food labels, usually has a smaller effect on serum cholesterol concentrations than saturated fat. Decreasing its absorption with foods enriched in plant sterol or stanol esters has a small hypocholesterolaemic effect, as also does mucilaginous fibre in fruit, vegetables, and oats. Avoiding coffee is probably pointless. Some authorities believe that the epidemiological evidence indicating that alcohol is protective against coronary heart disease is strong enough to justify encouraging moderate indulgence (red wine finds particular favour in view of the lower risk of coronary heart disease in southern as opposed to northern Europe). However, alcoholic beverages in excess can lead to obesity, hypertension, atrial fibrillation, and to exacerbation of hypertriglyceridaemia (see above), and a trial of abstinence should be considered in the patient with hyperlipidaemia suspected of overindulgence.

These dietary aims do not need to be modified for the treatment of moderate hypertriglyceridaemia and are also suitable for the management of diabetes. Carbohydrate-restricted diets are no longer in general use for either of these purposes. In patients with severe hypertriglyceridaemia it is necessary to limit the production of chylomicrons and so any fat in the diet must be avoided. Often a 25 to 30 g low-fat diet (in which, if the patient is not obese, carbohydrate is substituted to maintain dietary energy intake) can be employed, but occasionally even lower fat intakes must be achieved. Lipid-lowering drugs are frequently ineffective in patients with severe hypertriglyceridaemia, whereas dietary treatment can be particularly effective. Admission to a specialized centre with experienced dietetic services is often desirable.

Drug therapy of hyperlipidaemia

The indication for drug therapy is not the failure of serum cholesterol concentration to decrease below some arbitrary level despite dietary treatment in all patients. There are people with serum cholesterol concentrations as high as 8 mmol/litre (310 mg/dl) whose risk of coronary heart disease is not sufficiently high to justify the use of lipid-lowering drugs. However, when coronary risk is high there is no scientific basis for choosing different cholesterol (or LDL cholesterol) concentrations as thresholds for intervention or as therapeutic targets. Generally if the patient is in one of the following high-risk categories lipid-lowering therapy should be instituted if the serum cholesterol persists about 5.0 mmol/litre (200 mg/dl) (LDL cholesterol > 3.0 mmol/litre (> 120 mg/dl) and the aim of treatment should be to decrease serum cholesterol to less than 5.0 mmol/litre (200 mg/dl) (LDL cholesterol < 3.0 mmol/litre (< 120 mg/dl) or by 25 per cent (LDL by 30 per cent), whichever is the lowest. Whether dietary management should be instituted at the same time as lipid-lowering drug therapy or before in order to establish whether it alone will suffice is determined by the degree of risk and the degree to which the serum cholesterol is elevated. Dietary management does not typically decrease serum cholesterol by more than 0.5 mmol/litre (20 mg/dl). Certainly in patients with established CHD, statin treatment should be introduced without delay. The high risk categories are discussed in the subsections below.

Patients with established coronary heart disease or other significant atherosclerotic disease

Secondary prevention trials of cholesterol lowering using statin drugs provide strong evidence of prolonged survival due to a decrease in coronary events and strokes. Coronary angiography also provides evidence of regression of atheroma with lipid-lowering therapy. Lipid-lowering drugs are therefore indicated in patients with coronary heart disease (including those who have undergone coronary surgery or angioplasty). It is probably reasonable to extend this policy to patients with peripheral arterial, aortic, or significant carotid atherosclerosis, because there are angiographic studies to demonstrate favourable effects of lipid-lowering therapy on femoral and carotid atheroma and because this type of disease is closely associated with risk of coronary heart disease.

Familial hypercholesterolaemia and type III hyperlipoproteinaemia

The high risk of coronary heart disease and the known metabolic defects in these conditions justifies the use of lipid-lowering drug therapy. Familial hypercholesterolaemia should, if possible, be detected in childhood or early adulthood, and the age at which statin therapy should be commenced has therefore to be considered. Generally in boys a statin should be prescribed by the age of 20 years whereas in many women it can be left until the age of 30 years (discontinuing it temporarily if there is any possibility of conception). Some authorities advocate the earlier use of statins and this should certainly be considered if the family history of coronary disease is particularly adverse.

Patients with type III hyperlipoproteinaemia are generally encountered in adulthood and treatment should be initiated with a fibrate drug in all save the minority who respond to dietary management alone. It should not be assumed that dietary control is adequate if any degree of hypertriglyceridaemia persists, because this generally indicates that significant b-VLDL is still present in the circulation.

Multiple risk factors

The risk of coronary heart disease in some patients with additional adverse factors, whose serum cholesterol remains elevated despite diet, justifies the use of lipid-lowering drugs. Just how high the risk needs to be, and how it can be determined with any degree of exactitude, is a persisting problem for the clinician. Most national and international recommendations for primary prevention of coronary heart disease provide a means of assessing an individual patient's risk (usually based on the equation derived from the Framingham study by Anderson and colleagues) to assist in the clinical decision as to when to introduce lipid-lowering medication and increasingly when to treat mild hypertension. The National Cholesterol Education Program recommends an algorithm and recommends treatment at lower levels of risk than in Europe. The Joint European Guidelines recommend 20 per cent over 10 years as an appropriate threshold of coronary risk for statin therapy. In the Joint British Guidelines the minimum level of care is considered to be treatment when coronary risk is 30 per cent or more over 10 years, but that ideally a 10-year risk of 15 per cent or more should be targeted. A computer program is available from the British Hypertension Society website (www.hyp.ac.uk/bhs/management.htm) and the Family Heart Association website (http://www.familyheart.org/) for the prediction of both coronary and stroke risk. The author's own practice is to target 20 per cent and above. It is always important to seek evidence of existing coronary heart disease, since, if present, this clarifies the decision to start lipid-lowering therapy, and justifies investigation in its own right.

Diabetes mellitus

The relative reduction in coronary risk in patients with diabetes thus far included in statin trials is at least as great as in the non-diabetic participants. Because of the higher coronary risk in diabetes this means that even greater benefit, in terms of new events prevented, accrues from statin treatment than in non-diabetics. Results of statin and fibrate trials specifically conducted in diabetes are awaited. However, current evidence strongly supports the use of statin therapy in diabetic patients with coronary heart disease or any other atherosclerotic complication. What is uncertain is whether in the primary prevention of coronary disease diabetic patients should be stratified according to risk to determine when they receive lipid-lowering drug treatment as in non-diabetics or whether they should be a special category who are all treated in the same way as patients with established coronary heart disease. The latter approach, which has been adopted by the American Diabetic Association, would mean that most patients with serum cholesterol exceeding 5.0 mmol/litre (200 mg/dl) should receive statin therapy. In favour of this approach is the knowledge that coronary risk is substantially higher in diabetics for any given concentration of cholesterol than in non-diabetics, that prediction methods are likely to underestimate risk in non-insulin-dependent diabetes, and that there is no reliable method for predicting risk in insulin-dependent diabetes.

Markedly elevated cholesterol with no other risk factors and no clearly identifiable genetic syndrome

Many people fall into this category. Some will have familial hypercholesterolaemia but have not yet developed tendon xanthomas. Knowledge that a relative has these should weigh heavily in the decision to introduce therapy. The combination of high cholesterol with raised triglycerides and low HDL and an adverse family history also favours the introduction of lipid-lowering drug therapy even in the absence of other risk factors (see above). The risk will be underestimated from coronary risk prediction charts or computer programs when there is an adverse family history or hypertriglyceridaemia. If the risk cannot be read from the charts because the ratio of serum to HDL cholesterol is too high, treatment may in any case be justified.

In women with isolated hypercholesterolaemia who are peri- or postmenopausal, the possibility of prescribing hormone replacement therapy can be considered, particularly if the menopause is surgical or spontaneously premature, or if there are also menopausal symptoms. Hormone replacement therapy often decreases LDL cholesterol and may increase HDL cholesterol. However, at present there is no randomized, placebo-controlled clinical trial evidence that this therapy is beneficial in preventing coronary heart disease. Care should be exercised in patients with hypertension (because of the salt-retaining effects of progestogen), in patients with established coronary heart disease (because of its possible thrombogenic effects), and in patients with hypertriglyceridaemia (which sex steroids exacerbate). In women at high risk of coronary disease the quality of evidence in favour of statin therapy means that reliance should not be placed on hormone replacement therapy as a means of coronary prevention.

Lipid-modifying drugs

No major therapeutic decision, such as the introduction of a particularly restrictive diet or of lipid-modifying drug therapy, should be taken as the result of a single cholesterol determination, because this will be influenced both by biological and by laboratory variation. A laboratory result for cholesterol concentration is generally within ± 10 per cent of the true mean value, but may occasionally fluctuate more widely. Increasingly, portable or 'on-site' cholesterol analysers are being used in an attempt to make cholesterol measurement as immediate for the clinician as that of blood pressure. This has some advantages, but it must be remembered that such tests may be more expensive than those performed in the laboratory, they are generally less accurate unless performed by someone who is trained and regularly uses the instrument, and the calibrations usually differ from those employed in hospital laboratories.

Non-fasting cholesterol concentrations are satisfactory for the management of patients responding to simple dietary measures, but for those in whom drug therapy is under consideration, two fasting determinations of cholesterol, triglycerides, and HDL cholesterol are generally necessary (serum cholesterol and HDL cholesterol concentrations are not affected by meals, but serum triglyceride levels are). Knowledge of the HDL and triglyceride levels is essential at this stage because abnormal values for these would be an additional factor in favour of lipid-lowering drug therapy, and because their concentration may influence the choice of drug. Fasting blood glucose and serum creatinine and transaminases should also be measured, and urine should be tested for protein. Serum thyroxine should be measured if there is any suspicion of hypothyroidism, and some authorities advocate its measurement in all patients whose serum cholesterol exceeds 8 mmol/litre (310 mg/dl) even if hypothyroidism is not clinically evident.

The first-line therapy in all forms of hypercholesterolaemia, except that associated with triglyceride levels exceeding 5 mmol/litre (200 mg/dl), are the statin drugs (3-hydroxy-3-methylglutaryl CoA reductase inhibitors)—atorvastatin, fluvastatin, lovastatin, pravastatin, rosuvastatin, and simvastatin. These agents are often effective, even in marked hypercholesterolaemia, as monotherapy. They also have a triglyceride-lowering effect, which tends to be related to the extent to which they lower cholesterol but which is generally less than with a fibrate drug. Evidence that statins decrease coronary events is provided by three large secondary prevention trials using simvastatin or pravastatin, two primary prevention trials which employed pravastatin or lovastatin, and one trial which combined primary and secondary prevention patients and involved simvastation. Advantage may be taken of the synergism of statins with bile acid sequestrating agents (cholestyramine and colestipol), by prescribing two sachets in the morning with an evening dose of statin, in patients resistant to statins alone. Their use in combination with fibrate drugs requires strict clinical supervision, because there is an increased risk that myositis may ensue. There is a small incidence of this occurring spontaneously in patients on statins, and creatine kinase levels should be monitored. Erythromycin and cyclosporin also increase the risk of myositis, and care must be taken if statins are used after cardiac or renal transplantation. Statins may be particularly valuable in patients with renal disease in whom fibrates are contraindicated and in whom bile acid sequestrating agents may exacerbate hypertriglyceridaemia; these latter agents are particularly poorly tolerated in patients already receiving multiple drug regimes.

Bile acid sequestrating agents can be used in the treatment of hypercholesterolaemia in the absence of hypertriglyceridaemia, which they may exacerbate. A dose (two sachets) is best taken well soaked in fruit juice before breakfast. In larger more frequent doses these agents often cause nausea, heartburn, and constipation. Generally for this reason they have been increasingly relegated to the sidelines since the introduction of statins. In children and women of child-bearing potential, who have heterozygous familial hypercholesterolaemia and in whom drug therapy is justified because of a particularly adverse family history, the author remains reluctant to turn to other agents, although it is often better to wait until it is safe to commence statin treatment (which is better tolerated) than to alienate the patient from the clinic with unpalatable treatment earlier. If they are prescribed, folate and vitamin D supplementation should be considered, particularly in women who may become pregnant.

In patients whose hypercholesterolaemia is combined with more marked hypertriglyceridaemia, the fibrate drugs (bezafibrate, ciprofibrate, fenofibrate, gemfibrozil) are first-line therapy. They are also often highly effective in type III hyperlipoproteinaemia and useful in primary type V hyperlipoproteinaemia and in the dyslipoproteinaemia of diabetes mellitus. Fibrates are less effective in lowering LDL cholesterol than are statins. Most of their cholesterol-lowering effect is due to a decrease in VLDL cholesterol. They do, however, decrease small dense LDL levels. This is not readily evident from routine laboratory tests, because it is unaccompanied by any substantial reduction in cholesterol. In some particularly high-risk patients with combined hyperlipidaemia statin therapy may be added to fibrate therapy in order to satisfactorily lower LDL cholesterol. The fibrate drugs raise HDL cholesterol by more than statins. They must be avoided in patients with disturbed hepatic or renal function. They potentiate anticoagulants. The mode of action of fibrate drugs, which diminish serum triglyceride levels by stimulating lipoprotein lipase and decreasing circulating non-esterified fatty acids (NEFA), probably involves stimulation of the nuclear peroxisome proliferator-activated receptor a.

Nicotinic acid (niacin) can be used to lower serum cholesterol and triglyceride levels. The effective dose is usually associated with unpleasant flushing. This can be minimized if aspirin is taken before the nicotinic acid. There are also many other side-effects, and liver function must be monitored. Nicotinic acid has not found great therapeutic favour outside the United States. Unlike other lipid-lowering drugs, it is effective in lowering serum Lp(a). Acipimox, a niacin analogue, has a similar spectrum of action to the fibrate drugs and causes less flushing. Probucol is a cholesterol-lowering drug, which lowers HDL cholesterol relatively more than LDL. Despite its undoubted antioxidant properties, it requires further clinical evaluation before it can be regarded as beneficial in its overall action and is now of limited availability. Fish oil pharmacological preparations have triglyceride-lowering properties in daily doses of several millilitres, but do not lower LDL cholesterol and may even exacerbate diabetic hyperlipidaemia. Preparations which concentrate the omega-3 long-chain fatty acids, eicosapentaenoic and docosahexaenoic acid (Table 4), may have greater therapeutic potential. Fish oil improves survival after myocardial infarction in doses which have little effect on serum triglycerides and may relate to a decreased likelihood of ventricular arrhythmias.

Non-pharmacological lipid-lowering treatment

In addition to pharmacological agents and diet, extracorporeal removal of LDL is available in many centres for severe hypercholesterolaemia, usually homozygous familial hypercholesterolaemia in which it improves survival. Plasmapheresis or LDL apheresis, using systems that absorb LDL, are the two methods employed. Plasmapheresis and most methods of LDL apheresis also lower serum Lp(a). They must be repeated every 2 to 4 weeks. Occasionally, patients with homozygous familial hypercholesterolaemia have also been treated with liver transplantation to provide an organ with normally functioning LDL receptors. Partial ileal bypass surgery has been used to treat heterozygous familial hypercholesterolaemia (it is ineffective in homozygotes), but with the advent of more effective lipid-lowering drugs this is now very rarely necessary.

Hypolipoproteinaemia

Hypolipoproteinaemia is increasing as a clinical problem, because more cases are being discovered as a result of population screening for high cholesterol. People who have had a low serum cholesterol level all their lives do not seem to be at any disadvantage unless the decrease is profound, as in abetalipoproteinaemia. Indeed, their relative freedom from cardiovascular disease may lead to longevity. When the condition is discovered for the first time it is often difficult, however, to be sure that the low cholesterol is not due to an acquired disease, such as malignancy (for example colonic or prostatic neoplasms, leukaemia, reticulosis, or myeloma) or malabsorption (due, for example, to a short bowel, blind-loop syndrome, coeliac disease, pancreatic exocrine insufficiency, or giardiasis).

Some people with serum cholesterol levels around 1.0 to 3.5 mmol/litre (40 to 140 mg/dl) will have heterozygous familial hypobetalipoproteinaemia, an autosomal dominant condition in which truncated apoB mutations have been described. The condition is benign. However, homozygous hypoapobetalipoproteinaemia and another condition, abetalipoproteinaemia (inherited as an autosomal recessive), which produce more profound hypocholesterolaemia, are associated with retinitis pigmentosa, unusually shaped erythrocytes (acanthocytes), a syndrome resembling Friedreich's ataxia (preventable with administration fat-soluble vitamins), steatorrhoea (which can create diagnostic confusion with other causes of malabsorption leading to secondary hypocholesterolaemia), and fatty liver. Mutation of the *MTP* gene rather than of the apoB gene is associated with apobetalipoproteinaemia.

Analphalipoproteinaemia (Tangier disease) is a very rare autosomal recessive disorder with virtually absent HDL, reduced LDL cholesteryl ester, and cholesteryl ester deposition throughout the body, leading to enlarged orange-yellow tonsils and adenoids, lymph node enlargement, hepatosplenomegaly, bone marrow infiltration (thrombocytopenia), orange-brown spots on the rectal mucosa, neuropathy, and corneal cloudiness. Heterozygotes for this condition are at increased risk of premature coronary artery disease. A less severe form of this disorder (fish-eye disease) has been described. In another disorder, combined deficiency of apoA-I and apoC-III due to a rearrangement of DNA affecting the transcription of both their genes, which are clustered together on chromosome 11, leads to markedly decreased serum HDL levels, accelerated atherosclerosis, and corneal opacities. Some authorities believe that a much more common genetic HDL deficiency is the cause of HDL cholesterol levels in the lower 10 per cent of the frequency distribution. Evidence for this contention is incomplete.

Further reading

Anderson KM *et al.* (1990). Cardiovascular disease risk profiles. *American Heart Journal* **121**, 293–8.

Assmann G, van Eckardstein A, Brewer HB (1995). Familial high density lipoprotein deficiency: Tangier disease. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 2053–72. McGraw-Hill, New York.

Björkhem I, Boberg KM (1995). Inborn errors in bile acid biosynthesis and storage of sterols other than cholesterol. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 2073–99. McGraw-Hill, New York.

Breslow JL (1995). Familial disorders of high-density lipoprotein metabolism. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 2031–52 McGraw-Hill, New York.

Brooks-Wilson A *et al*. (1999). Mutations in *ABC 1* in Tangier disease and familiar high-density lipoprotein deficiency. *Nature Genetics* **22**, 336–45.

Brunzell JD (1995). Familial lipoprotein lipase deficiency and other causes of the chylomicronemia syndrome. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 1913–32 McGraw-Hill, New York.

Davies MJ, Woolf N (1993). Atherosclerosis: what is it and why does it occur? *British Heart Journal* **69** (Suppl.), S3–S11.

Downs GR *et al.* (1998). Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of the AFCAPS/TEXCAPS (Air Force/Texas Coronary Atherosclerosis Prevention Study). *Journal of the American Medical Association* **279**, 1615–22.

Durrington PN (1995). Lipoprotein (a). *Baillière's Clinical Endocrinology and Metabolism* **9**, 773–95.

Durrington PN (1998). Triglycerides are more important in atherosclerosis than epidemiology has suggested. *Atherosclerosis* **141** (Suppl. 1), S57–S62.

Durrington PN (2000). Diabetic dyslipidaemia. *Baillière's Clinical Endocrinology and Metabolism* **13**, 265–78.

Durrington PN (2002). *Hyperlipidaemia diagnosis and management*, 3rd edn. Butterworth Heinemann, Oxford.

Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults (2001). Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *Journal of the American Medical Association* **285**, 2486–97.

Gaw A, Shepherd J (1999). Fibric acid derivatives. In: Betteridge DJ, Illingworth DR, Shepherd J, eds. *Lipoproteins in health and disease*, pp 1145–60. Arnold, London.

Glomset JA *et al.* (1995). Lecithin:cholesterol acyltransferase deficiency and fish eye disease. In: In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 1933–51 McGraw-Hill, New York.

Goldstein JL, Hobbs HH, Brown MS (1995). Familial hypercholesterolaemia. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 1981–2030 McGraw-Hill, New York.

Gould AL *et al.* (1995). Cholesterol reduction yields clinical benefits. A new look at old data. *Circulation* **91**, 2274–82.

Grundy SM (1987). Dietary therapy of hyperlipidaemia. *Baillière's Clinical Endocrinology and Metabolism* **1**, 667–98.

Heart Protection Study, http://www.ctsu.ox.ac.uk/

Herz J (1999). Low-density lipoprotein receptor-related protein. In: Betteridge DJ, Illingworth DR, Shepherd J, eds. *Lipoproteins in health and disease*, pp 333–59. Arnold, London.

Illingworth DR (1999). 3-Hydroxy-3-methylglutaryl-coenzyme A reductase inhibitors. In: Betteridge DJ, Illingworth DR, Shepherd J., eds. *Lipoproteins in health and disease*, pp 1161–79: Arnold, London.

Kane JP, Havel RJ (1995). Disorders of the biogenesis and secretion of lipoproteins containing the B apolipoproteins. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 1853–85 McGraw-Hill, New York.

Karathanasis SK (1992). Lipoprotein metabolism: high-density lipoproteins. In: Lusis AJ, Rotter JI, Sparkes RS, eds. *Molecular genetics of coronary artery disease*, pp 140–71. Karger, Basel.

Law MR, Thompson SG, Wald NJ (1994). Assessing possible hazards of reducing serum cholesterol. *British Medical Journal* **308**, 373–9.

Law MR, Wald NJ, Thompson SG (1994). By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *British Medical Journal* **308**, 367–73.

Mahley RW, Rall SC (1995). Type III hyperlipoproteinemia (dysbetalipoproteinemia): the role of apolipoprotein E in normal and abnormal lipoporotein metabolism. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 1953–80. McGraw-Hill, New York.

Manninen V *et al.* (1989). High density lipoprotein cholesterol as a risk factor for coronary heart disease in the Helsinki Heart Study. In: Miller NE, ed. *High density lipoproteins and atherosclerosis II*, pp.35–42. Excerpta Medica, Amsterdam.

Pekkanan J *et al.* (1990). Ten-year mortality from cardiovascular disease in relation to cholesterol level among men with and without preexisting cardiovascular disease. *New England Journal of Medicine* **322**, 1700–7.

Rubins HB *et al.* for the Veterans Affairs High-Density Lipoprotein Cholesterol Intervention Trial Study Group (1999). Gemfibrozil for the secondary prevention of coronary heart disease in men with low levels of high-density lipoprotein cholesterol. *New England Journal of Medicine* **341**, 410–18.

Sacks FM *et al.* (1996). The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *New England Journal of Medicine* **335**, 1001–9.

Sampson MJ, Betteridge DJ (1999). Hyperlipidaemia and combination drug therapy. In: Betteridge, DJ, Illingworth DR, Shepherd J, eds. *Lipoproteins in health and disease*, pp 1213–29. Arnold, London

Scandinavian Simvastatin Survival Study Group (1994). Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *The Lancet* **344**, 1383–9.

Schumaker V, Lambertas A (1992). Lipoprotein metabolism: chylomicrons, very-low density lipoproteins and low density lipoproteins. In: Lusis AJ, Rotter JI, Sparkes RS, eds. *Molecular genetics of coronary artery disease*, pp 98–139. Karger, Basel.

West of Scotland Coronary Prevention Group (1996). West of Scotland Coronary Prevention Study: identification of high-risk groups and comparison with other cardiovascular intervention trials. *The Lancet* **348**, 1339–42.

Witztum JL, Steinberg D (1991). Role of oxidized low density lipoprotein in atherogenesis. *Journal of Clinical Investigation* **88**, 1785–92.

Wood D *et al.* (1998). Joint British recommendations on prevention of coronary heart disease in clinical practice. *Heart* **80** (Suppl. 2), S1–S29.

Wood D *et al.* with members of the Task Force (1998). Prevention of coronary heart disease in clinical practice: recommendations of the Second Joint Task Force of European and other societies on coronary prevention. *Atherosclerosis* **140**, 199–270.

11.7.1 Hereditary haemochromatosis

T. M. Cox

Definition

Haemochromatosis is an hereditary disorder generally caused by inappropriate absorption of iron by the small intestine that leads to iron deposition in the viscera, in endocrine organs, and at other sites. The toxic effects of iron impair the function of these organs and cause structural injury. Haemochromatosis cannot be diagnosed in the absence of excess tissue iron, and there is strong evidence for a cause-and-effect relationship between tissue iron storage and parenchymal injury. Several genetic syndromes associated with iron storage have been identified (Table 1); these may rarely involve specific tissues selectively, such as the lens of the eye or basal ganglia of the brain, or a characteristic range of tissues including the liver, heart, and endocrine system. By common agreement, the term 'hereditary haemochromatosis' refers to a group of inherited iron-storage diseases that affect diverse tissues and cause a multisystem disorder.

Pathological storage of iron

The body contains about 4 g of iron, 3 g of which is complexed with haem to form haemoglobin, myoglobin, and the cytochromes. The non-haem storage compartment, which consists of ferritin and its proteolytic degradation product, haemosiderin, represents up to 0.5 g of elemental iron in adult women and slightly more than 1 g in adult men. Excess storage of body iron (iron overload) is associated with an increase in hepatic iron concentrations and of the surrogate biomarker, serum ferritin. Minimal iron storage occurs when more than 1.5 g of total body iron is present; this is reflected in a hepatic iron concentration of approximately 30 µg atoms/g of tissue with a serum ferritin level of usually less than 250 µg/litre. Moderate iron-storage disease is reflected by a serum ferritin of approximately 500 µg/litre; under these circumstances the hepatic iron concentration rises to 100 µg atoms/g. Severe iron-storage disease (more than 5 g of storage iron) is shown by a hepatic iron concentration over 200 µg atoms/g liver tissue, with a serum ferritin level of at least 750 µg/litre—under these circumstances, tissue injury with impaired function is almost invariably present.

Clinical subtypes of haemochromatosis

Adult haemochromatosis

The familiar form of haemochromatosis is the classical adult type, which typically presents in middle age and is usually expressed in men. The disorder is inherited as a recessive trait and is due to mutations in a gene, *HFE*, that maps to the short arm of chromosome 6 in close apposition to the HLA class I loci of the human major histocompatibility complex (MHC). Expression of iron-storage disease in individuals carrying mutations in the *HFE* gene is very variable and is influenced by several environmental and sexual factors. Mutant alleles of the *HFE* gene that predispose to adult-type haemochromatosis are widespread and frequent in populations of North European origin. There is evidence from haplotype analysis that a single mutation arose on an ancestral chromosome 6 and spread throughout this population, probably as a result of the migration of the Vikings from Scandinavia. The disease occurs throughout the world as a result of intermarriage but is at its highest frequency in France, Germany, Great Britain, Scandinavia, Ireland, Northern Italy, Spain, and Eastern Europe as far as European Russia. Colonization has led to its appearance in all populations of the United States and in Australasia; for this reason hereditary adult-type haemochromatosis also occurs in South America.

Classical adult-type haemochromatosis is a slowly progressive disease affecting the liver, endocrine system, heart, and joints; it is often only diagnosed when irreversible tissue injury has occurred. The condition predisposes to the development of primary carcinomas of the liver. Other, rare, genetic forms of adult haemochromatosis occur in patients homozygous for mutations in the transferrin receptor gene type 2 (*TFR 2*) and in those heterozygous for mutant alleles of the human ferroportin gene.

Juvenile haemochromatosis

Since the identification of adult iron-storage disease by several European physicians during the nineteenth century, a similar disease has been recognized in children and young adults who may develop iron-storage disease of a more severe character, and which is now designated 'juvenile haemochromatosis'. Juvenile haemochromatosis is defined as iron-storage disease occurring before the age of 35 years; it evolves rapidly, typically affects the heart and endocrine system, and causes infantilism and hypogonadism as well as life-threatening cardiac arrhythmias. Juvenile haemochromatosis is inherited as a very rare recessive trait in which there is an increased frequency of consanguinity amongst the parents of affected subjects. Juvenile haemochromatosis resembles the severe iron-storage disease associated with the iron-loading anaemias, such as b-thalassaemia. Juvenile haemochromatosis affects males and females equally—an observation that reflects the overwhelming nature of the iron homeostatic defect: iron overload develops before the modifying effects of menstruation and dietary factors supervene. Recent studies have mapped at least one form of juvenile haemochromatosis to the long arm of chromosome 1. At the time of writing, the nature of the product of this gene is unknown.

Neonatal haemochromatosis

Neonatal haemochromatosis is a newly identified syndrome of uncertain cause, characterized by congenital cirrhosis or fulminant hepatitis associated with the

widespread deposition of iron in hepatic and extrahepatic tissues. Approximately 100 cases of neonatal haemochromatosis have been reported. Neonatal haemochromatosis occurs in the context of maternal disease—including viral infection, as a complication of metabolic disease in the fetus, and sporadically or recurrently, without overt cause, in siblings. In some families, although neonatal haemochromatosis appears to have an hereditary basis, no predictive genetic test is available to inform the outcome of at-risk pregnancies.

Prevalence and epidemiology

Juvenile and neonatal haemochromatosis are rare disorders that occur sporadically, but hereditary adult haemochromatosis is widely disseminated and of global importance. Removal of toxic iron by repeated venesection improves the outcome for adult haemochromatosis. If this treatment is instituted before irreversible tissue injury occurs, venesection may restore health and a normal life expectancy. For these reasons, there has been much discussion about the early recognition of iron-storage disease by the introduction of population-based screening programmes (using genetic testing or phenotypic biochemical screening methods) that can be applied to communities at risk.

In European populations, about 1 in 10 individuals carries one copy of an allele of the *HFE* gene that predisposes to iron-storage disease, and between 1 in 100 and 1 in 400 persons in these populations are homozygotes or compound heterozygotes with biochemical abnormalities of iron storage that may lead to full-blown clinical haemochromatosis. Thus the mutant allele, designated *C282Y* of *HFE*, which is the principal determinant of iron-storage disease, occurs at polymorphic frequency and is one of the most common genetic abnormalities leading to an autosomal recessive disease in populations of North-European origin. In European patients with iron-storage disease due to hereditary haemochromatosis, the frequency of homozygosity for the *C282Y HFE* allele ranges from about 35 per cent in Southern Italy to more than 90 per cent in the British Isles, including Ireland; in Australia, homozygosity for *C282Y* occurs in almost 100 per cent of patients with hereditary haemochromatosis. However, as discussed later, although useful for diagnosis, homozygosity for the *C282Y* mutation of *HFE* is **not** tantamount to a diagnosis of established iron-storage disease nor, therefore, of clinical haemochromatosis.

Clinical expression of haemochromatosis is highly dependent on age and it is very rare for there to be detectable disease in adults below the age of 20 years. As clinical disease is much more common in men than women, it is likely to reflect environmental factors and the modification of disease expression due to blood loss associated with menstruation and the investment in pregnancies, as well as the comparatively reduced dietary complement of iron in women. Other environmental factors, particularly the consumption of alcohol, appear to interact with predisposing genetic factors to induce the clinical expression of iron-storage disease in *C282Y* homozygotes. Most patients with the disease develop symptoms at, or above, the age of 40 years. However, studies of iron metabolism by biochemical measurements or tissue biopsy may reveal early evidence of iron storage in the long presymptomatic phase of this condition. With greater awareness of the diverse clinical manifestations of adult-type hereditary haemochromatosis, detection on the basis of early symptoms, for example arthritis or endocrine disease, may be possible. Thus, although the mutations that predispose to the development of haemochromatosis as a clinical entity are frequent in populations of European ancestry, there is a marked disparity in populations in which *C282Y* homozygosity is prevalent and the frequency with which symptomatic haemochromatosis is diagnosed.

Phenotypic expression of disease

For epidemiological purposes, since there is no internationally agreed case definition of haemochromatosis, caution is needed in interpreting claims that haemochromatosis is the most common inherited disorder affecting European peoples. Phenotypic expression of disease may range from the established clinical syndrome (cutaneous pigmentation, cardiomyopathy, endocrine failure—especially diabetes mellitus and hypogonadism, arthritis, and pigment cirrhosis) to a slight abnormality of blood parameters that reflect iron metabolism—elevated serum transferrin iron saturation and serum ferritin measurements. Such studies that are available to determine the penetrance and expressivity of the haemochromatosis gene have provided widely varying results in different populations: in Australia, where the mean intake of iron in the diet appears to be much greater than in the average European population today, most middle-aged male *C282Y* homozygotes appear to express at least one clinical manifestation of iron-storage disease. Similarly, a study of homozygous relatives (principally siblings) within pedigrees known to have haemochromatosis suggest that about half the men over 40 years of age, and about 1 in 6 of the women over 50 years of age, have at least one haemochromatosis-related clinical disorder. This latter survey, conducted in the United States, suggests that an important proportion of homozygous relatives of patients with established haemochromatosis, especially men, have conditions such as cirrhosis and arthropathy as well as abnormalities of serum liver-related tests that are not detected by spontaneous clinical referral.

Many reports of disease expression in haemochromatosis may, however, be questioned because of the prevalence of co-segregating genes within affected pedigrees, as well as early household environmental factors common to siblings that may predispose to disease expression. Studies in mice support this explanation, since it has been shown that several independent genetic determinants control the extent of iron-loading observed in mouse models of iron-storage disease generated by targeted disruption of the murine homologue of the *HFE* gene. In contrast, surveys conducted in outbred populations, for example in Jersey, show a great disparity between the predicted frequency of homozygosity for *C282Y* and the number of recorded cases with the disease attending local hospitals. These latter studies may reflect the widely suspected inability of clinicians to diagnose haemochromatosis, and an inability to bring together the unitary clinical manifestations of the disease into a unifying diagnostic category. However, widely differing degrees of disease penetrance almost certainly account for the apparent shortfall of diagnosed cases in populations at risk.

At present, no clear data in large unbiased population surveys are available to assess disease penetrance and the modifying effects of lifestyle factors (including alcohol, nutrition, diet) as well as pregnancy and menstruation that are likely to influence the effects and rate of iron storage in human *C282Y* homozygotes. Mortality figures show that death is rarely attributed to hereditary haemochromatosis in populations at risk. This fact contrasts starkly with the well-established known complications of the fully penetrant clinical syndrome, in which early death results from cirrhosis of the liver, hepatocellular carcinoma, endocrine failure or cardiac complications.

A contemporary study that examined the frequency of the most common symptoms of haemochromatosis in *C282Y* homozygotes, *C282Y/H63D* compound heterozygotes, and persons wild-type at these loci has been reported from California. In more than 41 000 individuals attending a health appraisal clinic, no evidence of an increased frequency of symptoms was identified in those genetically predisposed to iron-storage disease. The only significant clinical history identified in the at-risk group was that of 'hepatitis' or prior 'liver complaints'; only one of the 152 identified *C282Y* homozygotes had signs and symptoms of adult haemochromatosis. This provocative report, indicating a very low clinical penetrance (less than 1 per cent) of the haemochromatosis genotype in an unusual group of adults over the age of 26 years, raises important questions about the introduction of mass population screening for this potentially treatable iron-storage disease by genetic or even biochemical methods. However, the high prevalence of impotence, joint symptoms, chronic fatigue, and other complaints such as cardiac arrhythmias in the study group as a whole, raises disturbing questions about the valid application of this report to other populations. It is perhaps not surprising that in a group where, on average, more than 40 per cent complained of a general limitation of their health and/or joint symptoms, and in which more than 35 per cent of the male participants scored positively on symptom enquiry about impotence, a significant contribution from predisposing haemochromatosis alleles could not be identified. Nonetheless, this large study raises key questions about the utility of screening for adult haemochromatosis as a genetic disease. Before screening for haemochromatosis is introduced, there is clearly a need for other population surveys to be carried out in which the morbidity and mortality of individuals with the wild-type genotype as well as those harbouring disease alleles are investigated.

Pathophysiology and pathogenesis

Young patients with haemochromatosis absorb an increased amount of dietary iron in their upper intestine compared with normal control subjects. In established iron-storage disease, iron absorption continues at a rate that is inappropriate for the level of iron stores as reflected by serum ferritin and tissue iron determinations.

In the absence of an effective excretory pathway, the increased absorption of iron by the intestine leads to a progressive accumulation of the metal in the parenchymal cells of the liver, heart, endocrine glands, and specialized B-type synoviocytes. Excess iron accumulates in the pancreas where it is found in both acinar and endocrine cells of the islet, although there is a particular predisposition in the early phases of iron loading to the islet beta-cell. Iron also accumulates to toxic levels in the gonadotrophs of the anterior hypophysis, leading to hypogonadotrophic hypogonadism. Iron may accumulate in the adrenal gland, where it is concentrated particularly in those cells that secrete aldosterone, in the zona glomerulosa. Iron accumulates in the cardiac myocytes and conducting tissue of the heart, in the chief cells of the parathyroid, and in parenchymal cells throughout the body. The consequences of toxic iron storage include diabetes mellitus, cirrhosis of the liver, cardiomyopathy with or without conduction defects, hypogonadism, arthritis with chondrocalcinosis, adrenocortical deficiency, and, rarely, hypoparathyroidism. Evidence for the intrinsic toxicity of iron in haemochromatosis is provided by the regression of the pathological changes following measures taken to reduce iron, for example the use of iron chelators and removal of body iron by venesection. Venesection stimulates the mobilization and removal of iron from the storage compartment by increasing the demand for red cell production in the bone marrow.

Mechanism of iron toxicity

High concentrations of iron salts are toxic to cultured cells. The administration of iron chelates to experimental animals has induced diabetes with iron loading in the liver and pancreas as well as the generation of (renal) carcinomas. Injections of iron salts induce local sarcomas in experimental animals, with evidence of species susceptibility. In humans, sarcomas or carcinomas have arisen, albeit rarely, at sites of therapeutic injections of iron, and it is possible that the complications of silicosis and asbestos exposure result from the complement of iron associated with these particulates. A wealth of indirect but corroborative evidence indicates that the primary effect of excess free iron is to promote the formation of oxygen free radicals, which mediate the damage to cells and tissues that is observed in iron-storage disease. In established haemochromatosis, the iron-binding capacity of plasma transferrin may be exceeded so that a proportion of the iron present in the blood remains reactive as a low-molecular-weight species only loosely attached to plasma proteins. Non-transferrin iron in human plasma stimulates the peroxidation of unsaturated lipids and can form reactive complexes that react with DNA—thus suggesting a mechanism for genome toxicity and carcinogenesis related to iron overload. Iron is highly electroreactive, and coupling of the Fenton and Harber–Weiss reactions leads to the formation of hydroxyl radicals as a result of the catalytic interactions between superoxide and ferric ions. Tissues with significant iron storage show peroxidative injury in membrane lipid fractions.

The lysosomal compartment appears to be particularly susceptible to iron-mediated damage, since iron in the form of ferritin and its degradation product, haemosiderin, accumulates within lysosomes to form the particulate ferrugineous granules known as 'siderosomes'. In haemochromatosis, there is an increased activity of lysosomal enzymes with biochemical evidence of increased lysosomal fragility indicating disruption of the integrity of the lysosomal membrane by iron; these changes revert to normal when the tissue iron is removed by venesection or by the use of specific iron chelators. It seems likely that the electrochemical reactivity of iron and its particular propensity to accelerate the formation of oxygen free radicals mediate its injurious effects on cell membranes, as well as the nuclear genome leading to cancerous change. However, despite great advances in the understanding of free-radical chemistry, the cause-and-effect relationship between iron storage and tissue injury is difficult to prove unequivocally. Nonetheless, much experimental evidence points to the development of iron-mediated peroxidative injury of cellular membranes including the lysosome, as well as iron-mediated genotoxicity. Whatever their physiochemical basis might be, common mechanisms of iron toxicity clearly exist, since the pathological and clinical manifestations of all iron-storage syndromes, including secondary haemochromatosis associated with blood transfusion and the iron-loading anaemias, are almost identical.

Pathology of iron storage

Heavy deposits of iron in the tissues are associated with fibrosis and cell loss. Simple inspection reveals an overt rust-like discoloration of the liver, spleen, pancreas, heart, and lymph nodes. The liver is usually enlarged; haemosiderin is found in all cell types with the formation of fibrous septa and hyperplastic nodules. These nodules, which may be the forerunners of adenomas and hepatocellular carcinomas, contain little stainable iron—unlike the adjacent parenchyma.

The dominant site of iron deposition during the early phases is within hepatocytes but soon iron loading may be observed in all cell types, including the lining cells of biliary canaliculi, Kupffer cells, and stellate cells (Fig. 1 and Fig. 2 and Plate 1 and Plate 2).

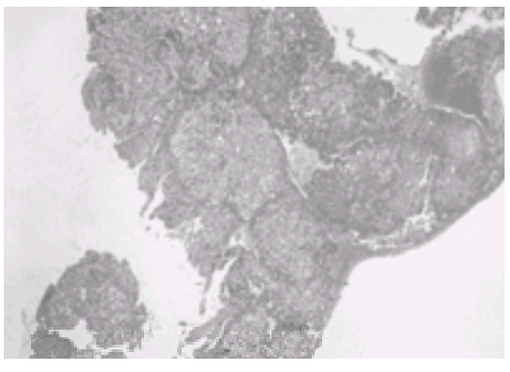

Fig. 1 Low-power, needle-biopsy appearance of liver specimen stained with haematoxylin and eosin from a 67-year-old man with adult haemochromatosis due to homozygosity for the *C282Y* mutation. Note the large hyperplastic nodules and fibrosis. (See also Plate 1.)

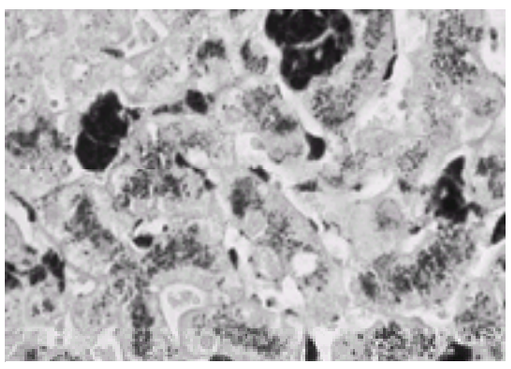

Fig. 2 High-power micrograph of the liver biopsy specimen shown in Fig. 1 stained with Perls' reagent. Note extensive deposits of ferric iron in all cell types including Kupffer cells, cells lining small biliary radicles, and in a punctate distribution within parenchymal hepatocytes. Liver cells are hyperplastic. (See also Plate 2.)

Similarly, in the pancreas there is fibrosis and iron deposition in the acini, ducts, and islets of Langerhans. Staining with Perls' reagent reveals arked haemosiderin deposition in the exocrine and endocrine glands, including many cell types in the testes. Haemosiderin is also markedly increased in the chief cells of the parathyroid, the adenohypophysis, the zona glomerulosa of the adrenal, and the thyroid.

In the joints, there is loss of the intra-articular space with chondrocalcinosis and deposits of haemosiderin in the synovium; electron microscopy shows selective deposits of ferritin and haemosiderin within type B synoviocytes. Radiological examination of the joints shows collapse of articular surfaces, subchondral cyst formation, and prominent formation of periarticular osteophytes. In the heart, pericardial constriction with fibrosis may occasionally be observed, but the principal abnormality is seen in the myocardium with degeneration and vacuolation of cardiac myocytes and intermyocyte fibrosis that involves conducting tissue in the septa. Surviving myocytes show eosinophilic degeneration and evidence of hypertrophy. Microscopical examination shows that in established cases of haemochromatosis, all tissues (except the choroid plexus) are affected by the iron-storage process. In the past, it was considered that transfusional and other types of secondary iron-storage disease predominantly affected the cells of the mononuclear macrophage system, such as the Kupffer cells of the liver, rather than the parenchymal cells. Iron deposits in the macrophage system may be less damaging than in other cell types, but it is difficult at present to relate evidence of iron-mediated injury to its cellular distribution. Progressive tissue injury follows the long-term cumulative toxicity of iron storage and its consequential effects on organ structure and cellular function. A striking, but unexplained, feature of iron-storage disease in the liver and other tissues is the absence of overt necrosis; careful study of the cellular effects of iron storage on apoptotic mechanisms in diseased tissues is clearly warranted.

Quantitative aspects of iron-storage disease

Chemical determination of tissue iron content yields useful information about the severity of iron loading in haemochromatosis, and may also provide a means to judge local responses to iron-depletion therapy, such as venesection. In normal individuals, the total concentrations of liver iron do not exceed 0.15 per cent by dry weight, but in established haemochromatosis the value is usually 1 per cent or more; in severely affected patients with untreated hereditary haemochromatosis or secondary haemochromatosis the amount of iron may exceed 5 per cent of the dry weight of tissue. The overall burden of body iron in patients with

haemochromatosis is usually in excess of 5 g in hereditary disease, a figure that rises with age. Estimates indicate that the total burden in patients with advanced haemochromatosis can be as much as 40 to 60 g—most of this accumulating in the liver—the pancreas and other organs such as the lymph, thyroid, pituitary, and salivary glands typically show an increase of more than 10 times the normal iron content.

Nature of the metabolic defect

In established haemochromatosis, where the burden of iron may increase body iron stores by at least tenfold, measurements usually show that iron absorption is within the normal range. Studies in young patients with rapidly progressive disease show a markedly increased absorption of iron, and all the evidence points to an increase in iron absorption to 2 to 3 mg daily throughout the lifetime of patients with haemochromatosis. After depletion therapy, the rate of recovery of iron stores is greatly enhanced for many years in patients with haemochromatosis, reflecting a persistent homeostatic abnormality in the retention of dietary iron. The daily absorption of between 2 and 4 mg of iron over a period of 30 to 40 years accounts for the degree of iron loading that occurs at presentation in patients with haemochromatosis, and compares with the normal absorption of 0.8 to 1.0 mg in men and in women, up to 2 mg daily. In effect, the abnormal absorption of iron represents a disturbed regulation of the final common pathway for the acquisition of iron from the environment by the small intestinal mucosa.

Iron absorption in hereditary haemochromatosis

A recent report, describing the transplantation of intestine and liver from an *HFE C282Y* homozygote into a recipient without haemochromatosis, has provided evidence that the small intestine is the principal site of expression of the hereditary defect in adult haemochromatosis. The transplantation was associated with early iron overloading in the recipient, together with raised serum transferrin iron saturations—a phenomenon not observed in recipients of hepatic allografts obtained from donors later found to be homozygous for the haemochromatosis gene. Studies *in vitro* and *in vivo* have suggested that there is a qualitative abnormality of the uptake and transfer of iron from the intestinal lumen in patients with hereditary haemochromatosis, although, until recently, the nature of this abnormality was unclear.

Latterly, genetic studies of mutant strains of mice with abnormalities of iron metabolism have shed light on the iron-absorption mechanism. The identification of a single gene encoding the divalent metal transporter protein, DMT 1, which is expressed in the upper small intestine and cells of the erythron, provides a molecular understanding of the iron deficiency—and the microcytic anaemia that occurs in the *mk/mk* mouse strain. A single point mutation in the *DMT1* gene interferes with the uptake of ferrous iron, since it disrupts the cognate transmembrane carrier protein mainly expressed in the mucosa of the proximal small intestine at the site of iron absorption and in the erythroid precursor cells. Since *in vitro* studies of the expressed protein DMT 1 show that it serves only as a carrier of divalent cations, and that interference with this pathway is sufficient to induce iron deficiency in a mammalian species, ferrous iron uptake is probably the main pathway by which inorganic iron is acquired by the intestine. Human *DMT1* maps to the long arm of chromosome 12 and encodes a 12 membrane-spanning protein that is expressed in the apical membrane of the upper intestine and in the apical membrane of differentiated human CaCo-2 cells of small intestinal phenotype. DMT 1 is also expressed in developing erythroid cells in which it is responsible for the intracellular delivery of iron derived from transferrin for haemoglobin synthesis.

The discovery of DMT 1 immediately indicated a possible role for this important protein in human haemochromatosis. Overexpression of *DMT1* mRNA has been identified in the intestinal mucosa of patients homozygous for the *C282Y* mutation with hereditary haemochromatosis, as well as in mice with iron-storage disease due to targeted disruption of the *HFE* gene. At the same time, studies in experimental animals have identified a cytochrome-containing ferrireductase that is also localized to the intestinal brush-border membrane; this reductase has been cloned from murine intestine and its human homologue has been identified. Expression of mucosal ferrireductase is specific to the apical microvillous membrane of mammalian intestinal mucosa and appears to be induced in response to nutritional iron deficiency. Mucosal ferrireductase reduces ferric irons derived from the diet in the lumen for delivery to the DMT 1 carrier protein, the final divalent pathway for inorganic iron uptake by intestinal mucosa. The mRNA species encoding murine *DMT1* exist in two isoforms, one of which contains an iron-response element (**IRE**) 3' region—which would allow for the post-transcriptional regulation of protein expression controlled by intracellular iron status. A similar translational control of transferrin receptor expression has been described with the 3' IRE in the mRNA encoding the human transferrin receptor. Since the IRE-containing isoform of DMT 1 is preferentially expressed in the duodenum, it seems likely that changes in intracellular iron status regulate the expression of this carrier protein in iron deficiency and haemochromatosis. Studies in *HFE* knockout mice indicate that the functional expression of the DMT1 protein is enhanced in the murine model of haemochromatosis, leading to increased iron uptake across the brush-border membrane of iron presented in the ferrous form; the action of non-rate-limiting ferrireductases at the brush-border membrane functionally coupled to DMT 1 activity appears to explain the enhanced isotopic uptake of ferrous iron in this model of haemochromatosis.

At present, our molecular understanding of transepithelial iron uptake in haemochromatosis and in health is somewhat rudimentary. A novel gene termed ' *ferroportin*', encoding a multitransmembrane domain protein, has been identified. The cognate protein may function as an exporter of iron across the brush-border membrane at the basal surface of the intestine as well as in placental syncytotrophoblasts. Ferrroportin appears also to be responsible for the export of iron retrieved by erythrophagocytosis by macrophages. After initial uptake, the enhanced transfer of iron across the mucosal epithelium in haemochromatosis and iron deficiency is mediated by, as yet, unknown iron-binding proteins. Delivery of the iron to the systemic circulation is mediated by the regulated downstream coexpression of the membrane protein, ferroportin. It seems likely that, in hereditary haemochromatosis and physiological iron deficiency, post-transcriptional control of carrier proteins responsible for the uptake and transfer of iron occurs in the absorptive epithelium on the tips of the intestinal villi. Thus homeostatic mechanisms in the proximal intestine operate to bring about the co-ordinated transfer of iron presented in the intestinal lumen specifically to meet body requirements. Proteins including hephaestin encoded on the X- chromosome, which is mutated in the sex-linked anaemic mouse *sla*, also mediates the transfer of iron across the intestinal mucosa.

The signal for regulating the absorptive activity of the ferrous ion transport pathway is not known. However, it seems likely that interactions between the wild-type HFE protein and transferrin receptors, including the newly described transferrin receptor 2 isoform that may be expressed in intestinal crypts, in some way instruct the developing epithelial cells within the intestinal crypt about body iron requirements. Although functional interactions of HFE molecules with the identified components of the absorptive pathway have yet to be clarified, the HFE protein probably influences iron status in intestinal stem cells within the crypt. By these means, the expression of key transport proteins such as DMT 1 and ferroportin may be imprinted, thus influencing their subsequent functional activity during ascent up the villus. At present, however, much more experimental work will be required to further our understanding of the signalling pathways by which the body iron status regulates the avidity of the proximal small intestine for nutritional iron presented within the lumen.

A variable but often substantial, component of dietary iron is present in the organic form as haem; a full molecular understanding of the uptake and transfer pathways for the absorption of iron complexes to the porphyrias is also needed. Whole-body studies show that the absorption of the radiolabelled iron moiety of haemoglobin is enhanced in patients with adult-type haemochromatosis. Early studies in dogs have shown that, in the presence of proteolytic digestion products of globin, the haem complex is taken up intact by mucosal epithelial cells; free iron is then released by the action of intracellular haem oxygenases. The contribution of haemoglobin, myoglobin, and cytochromes to the iron overload in patients with haemochromatosis has not been quantified but iron complexed to haem may well represent an important component of the total burden of body iron in symptomatic haemochromatosis.

Genetics and molecular biology of haemochromatosis

The principal determinant of adult haemochromatosis has long been known to be tightly linked to the human MHC loci on the short arm of chromosome 6. In 1996, mutations in the HLA class I-linked haemochromatosis gene, *HFE*, were shown to predispose to the adult form of the disease. The most common mutation in the non-classical MHC class I HFE protein affects a key cysteine residue, which contributes to the formation of the conserved a-3 helix that interacts co-translationally with the b_{2}-microglobulin protein. This association is required for the cell-surface expression of all class I MHC molecules. Most patients with haemochromatosis are thus homozygous for a cysteine-to-tyrosine mutation at codon 282 (*C282Y*) of the nascent HFE protein; an increased frequency of this mutation, in association with the more common *H63D* missense mutation, also occurs in adult haemochromatosis (Fig. 3). A minor variant, affecting the same region in the a-1 helix, *S65C*, is also occasionally associated with the *C282Y* allele in compound heterozygotes with adult iron-storage disease.

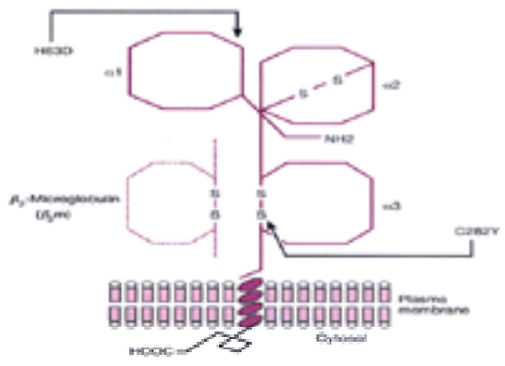

Fig. 3 Diagram of non-classical MHC class I-like HFE molecule shown in juxtaposition with the b $_2$-microglobulin. The location of the two frequent amino-acid substitutions (*C282Y* and H63D) that predispose to the development of adult haemochromatosis is indicated by the arrows.

These missense mutations in HFE occur at a much lower frequency in control populations without iron overload. Apart from reducing cell-surface expression of the mutant C282Y polypeptide, and thus the abundance of this protein within a population of cytoplasmic vesicles, a functional explanation for the qualitative abnormality of iron metabolism that characterizes haemochromatosis is not available. Recent unsubstantiated experiments have indicated the coexpression of HFE with transferrin receptor isoforms within a vesicular intracellular compartment. Structural studies have provided a molecular basis for this interaction based on the expression of truncated soluble HFE and transferrin receptor protein *in vitro*, combined with elegant structural studies using X-ray diffraction. Latterly, a novel isoform of the transferrin receptor, transferrin receptor type 2, has been identified in intestinal crypt cells where it may colocalize with the HFE protein within intracellular vesicles. Since rare cases of adult haemochromatosis have been reported with nonsense and inactivating (null) mutations in the *TfR2* (transferrin receptor 2) gene, it seems likely that the HFE protein participates in the regulation of iron delivery to cells through the transferrin ligand. HFE may affect the delivery of transferrin-bound iron by way of the transferrin receptor, or, also plausibly, by the transferrin isoform *TfR2* in intestinal crypt cells, thereby influencing their subsequent absorptive behaviour.

Mutations in *HFE* can be easily detected by the use of restriction enzymes and analysis of amplified *HFE* gene sequences obtained from genomic DNA. Point mutations in rare cases of non-HFE-associated haemochromatosis have been identified in the *ferroportin* and in the *TfR2* genes. Patients harbouring inactivating mutations in the gene encoding the iron-transfer ligand, serum transferrin, are also reported in association with severe iron-storage disease. This finding, together with the iron-storage disease associated with enhanced iron absorption in hypotransferrinaemic mice (*hpx*), implicates this transferrin-receptor subtype in the physiological regulation of iron homeostasis.

Clinical features

Adult haemochromatosis.

The clinical features of adult haemochromatosis include skin pigmentation. The pigment may be manifest as a generalized slate-grey coloration, due principally to melanin, or localized bronzed pigmentation particularly of the lower limbs, associated with iron deposits in adnexal dermal structures as well as melanin. Histological examination of the skin reveals increased melanocyte activity in conjunction with iron deposits, particularly in cutaneous sweat and apocrine glands. Increased skin pigmentation is a common but not invariable manifestation of haemochromatosis. It increases as the disease progresses and may be a late manifestation of the condition; absence of pigmentation should never, as a consequence, be regarded as a contraindication to the diagnosis of iron-storage disease.

Iron-storage disease invariably affects the liver. The liver is usually enlarged and may be cirrhotic, but portal hypertension and splenomegaly are rare endstage features of haemochromatosis. The enlarged liver, even in the absence of cirrhosis, may contain single or multifocal hepatocellular carcinomas. Hypogonadism is often present; it is typically preceded by a long history of fatigue, sexual asthenia, and impotence, as well as premature menopause and loss of libido in women. In men, there is gynaecomastia, circumoral vertical skin wrinkling, and loss of body hair; the genitalia show premature atrophy.

Many patients with haemochromatosis suffer from arthritis at an early phase in the illness and this may indeed be the sole manifestation of the condition for many years. The arthritis typically affects the second and third metacarpophalangeal joints of the hands and feet (Fig. 4 and Plate 3). These joints show painful swelling without obvious inflammatory changes. Distal interphalangeal joint disease is also recorded and is usually considered to be typical of osteoarthritis. Many joints, including the wrist, elbow, shoulder, and knee, may be affected and the changes in these joints are typically associated with chondrocalcinosis that is detected radiologically. The affected joints show loss of joint space, subchondral cysts, and, especially in the digits, prominent osteophyte formation (Fig. 5). Recent studies show that premature and disabling arthritis in the hip and other large joints is a characteristic feature of haemochromatosis.

Fig. 4 Arthropathy in a man with adult haemochromatosis forced to stop manual work because of painful arthritis especially in the second and third metacarpophalangeal joints; note increased skin pigmentation. (See also Plate 3.)

Fig. 5 Radiograph of hands in a 51-year-old woman with haemochromatotic arthropathy of the hands for many years. Note loss of joint space especially in metacarpophalangeal joints with subchondral cyst formation and osteophyte growth. Chondrocalcinosis is present in the ulnar fibro-cartilage at the wrist.

The symptoms of haemochromatosis are notoriously non-specific and slow in their progression. Fatigue is often reported and may be a manifestation of

hypogonadism and the onset of diabetes mellitus. Atrial fibrillation may be an early manifestation of cardiomyopathy. Later, paroxysmal arrhythmias and cardiac failure supervene, leading to shortness of breath and fatigue. Occasional patients with haemochromatosis present with isolated features, such as abnormal liver-related tests detected during routine examination for health insurance, or with arthralgia and signs of arthropathy in association with either diabetes, impaired libido, or sexual failure. Cardiomyopathy with heart failure or isolated arrhythmias is an unusual lone presentation of the disease.

The differential diagnosis of haemochromatosis is very wide, but the presence of diabetes with abnormal liver function or hepatomegaly, or an association with endocrine failure or arthropathy, should prompt consideration of iron-storage disease. Likewise, the presence of seronegative polyarthropathy with pigmentation, hepatomegaly, or any of the associated endocrinological changes should initiate immediate testing for evidence of iron-storage disease.

In young patients with hypogonadism or cardiomyopathy, iron-storage disease should be considered; juvenile haemochromatosis is often neglected by endocrinologists investigating young patients for infantilism or hypogonadotrophic hypogonadism. The condition may be responsible for cases of undiagnosed seronegative polyarthropathy. Iron-storage disease should be considered in any patient with signs and symptoms of chronic liver disease, including those with sustained mild elevation of serum transaminase activities, particularly since the liver is affected early in the course of the iron overload.

In fully established cases, skin pigmentation which may either be of a grey colour as a result of increased melanin, or, especially on the shins, a yellow-brown 'bronze' colour. Pigmentation in association with diabetes with or without arthropathy and hepatomegaly almost always signifies established iron-storage disease.

Diagnosis

It is critically important to make a diagnosis of haemochromatosis at the earliest opportunity. There is strong evidence that if treatment to remove iron before established structural injury occurs, then tissue function and symptoms improve. Several surveys indicate that removal of iron from patients diagnosed in the precirrhotic phase of adult haemochromatosis is associated with a normal or near-normal life expectancy.

Laboratory investigations

In adult haemochromatosis, the diagnosis can be usually established by demonstrating abnormalities of iron metabolism (fasting serum transferrin saturation with iron greater than 60 per cent), together with a measurement of serum ferritin concentration that provides evidence of increased iron stores. In most, but not all, untreated patients with pathological iron-storage disease due to haemochromatosis, the serum concentration of ferritin is elevated. Molecular analysis of the *HFE* gene for homozygosity for the common (*C282Y*) predisposing allele to the development of adult haemochromatosis may be very useful in patients of European ancestry. There is an increased frequency of compound heterozygotes for the *C282Y/H63D* or, more rarely, *C282Y/S65C* genotypes in patients with evidence of iron-storage disease.

Not all patients with adult haemochromatosis have mutations in the *HFE* gene; moreover, genetic tests are not at present available for the diagnosis of juvenile haemochromatosis or of neonatal haemochromatosis. In some patients with adult iron-storage disease, mutations have been identified in a newly identified gene that encodes a variant of the transferrin receptor—transferrin receptor type 2 (*TfR2*). Homozygosity for mutations in *TfR2* are found in some Southern European patients with adult haemochromatosis. Adult haemochromatosis due to mutations in the *HFE* gene is now known as HFE 1; juvenile haemochromatosis due to lesions in an, as yet, uncharacterized locus on chromosome 1q are now designated HFE 2, and mutations in the type 2 transferrin receptor gene cause HFE 3—another adult variant of hereditary haemochromatosis. As indicated above, HFE 2 (juvenile haemochromatosis) is a much more severe disease than either HFE 1 or HFE 3—although hypogonadism can be a presenting symptom in both HFE 1 and HFE 3 (see Table 1). A rare form of adult haemochromatosis, principally transmitted as a dominant trait, appears to be associated with point mutations in the newly described *ferroportin* gene (see above). It is believed that ferroportin is involved in the transport of iron from the intestinal epithelial cells to the body as well as to the liver. This type of haemochromatosis, now designated HFE 4, appears, if anything, to be slightly milder than haemochromatosis in the homozygous recessive forms of HFE 1, 2, and 3. HFE 4 responds poorly to iron-depletion therapy by venesection; histological examination reveals prominent iron storage within Kupffer cells.

Given the genetic variants that are now recognized as causes of haemochromatosis, it is clear that if any doubt exists as to the diagnosis, or molecular analysis of the *HFE* gene fails to identify known pathogenic mutations, then tissue diagnosis is indicated. This is usually carried out by liver biopsy with histochemical determination, and preferably chemical quantification, of tissue iron content. Although a liver biopsy is associated with small but definable risks, it does offer a key opportunity for the evaluation of liver structure and of the injury consequent upon iron deposition. The finding of cirrhotic change carries with it a poor prognosis; cirrhotic change is also a major predictor of the occurrence of hepatocellular carcinoma, which occurs rarely in non-cirrhotic subjects with iron-storage disease (Fig. 6 and Plate 4).

Fig. 6 Adult haemochromatosis. Section of liver lobe after surgical resection to remove a primary hepatocellular carcinoma arising in an iron-loaded but, unusually, non-cirrhotic liver in this disorder. The patient, aged 62 years, had been partially treated by venesection but recently noticed increasing lethargy: a raised serum a-fetoprotein concentration led to the diagnosis; moderate histochemical evidence of iron storage was found in the non-malignant tissue excised at surgery. (See also Plate 4.)

Serum iron-saturation determinations, and particularly serum ferritin concentrations, may signify conditions other than iron-storage disease. Serum ferritin is elevated in inflammatory states, certain malignancies such as Hodgkin's disease and in any condition associated with significant necrosis of parenchymal liver cells. Under these circumstances liver biopsy is recommended, since it is most likely to provide a definitive diagnosis of iron-storage disease. Sometimes, however, liver biopsy is not possible—either because the patient will not consent to it, or because of the presence of ascites and a bleeding disorder (especially thrombocytopenia). Under these circumstances, magnetic resonance imaging (**MRI**) may provide additional information but it is an insensitive test for the presence of iron-storage disease. If a liver biopsy is not possible and MRI of the liver does not reveal increased ferromagnetic signals indicative of iron storage, there are two further options: measurement of urinary iron excretion after parenteral administration of desferrioxamine, and, where the patient will tolerate it, quantitative phlebotomy. Injection of 500 mg of desferrioxamine intramuscularly in a patient with iron overload will usually induce the daily excretion of more than 2 mg of iron as the ferrioxamine complex in the urine. Ferrioxamine excretion may be increased in patients with haemolytic anaemia but when elevated is generally indicative of iron-storage disease. Weekly phlebotomy of 500 ml will remove approximately 225 mg of iron, and thus provides a means of estimating the amount of iron removed from the storage compartment when undertaken to induce a mild hypochromic anaemia of approximately 10.5 to 11.0 g of haemoglobin/dl or a serum ferritin concentration of less than 30 µg/l. Iron overload exists when the estimated iron removed by this method exceeds 1.5 g; unfortunately quantitative phlebotomy is cumbersome and may not be possible in patients with severe liver disease associated with hypoalbuminaemia.

Diagnosis in family members

The diagnosis of haemochromatosis, whether it be of adult or juvenile form, in an individual has immediate implications for first-degree relatives. All forms of haemochromatosis have a strong hereditary basis; and even some forms of neonatal haemochromatosis may, in some families, be inherited as an autosomal recessive trait. A dominant transmission pattern has been established in the case of HFE 4.

Although the penetrance and expressivity of homozygosity for the various alleles that predispose to haemochromatosis is not yet established, the risks of the disease in first-degree family members is sufficiently high to warrant systematic study. Clearly, the implications for asymptomatic or undiagnosed relatives of the propositus

are potentially very large. Hence, considerable care and sensitivity are needed in the means of informing them about the condition through the identified index case. In large families there may be formidable difficulties, so that the help of genetic counselling services as well as formal assistance from physicians practised in medical genetics may be needed. There can be little doubt, however, that at-risk relatives should be offered the opportunity for further diagnostic and clinical evaluation in relation to iron-storage disease. The condition is readily susceptible to iron-depletion therapy in its early stages; moreover, there may be additional considerations for patients who will wish to make reproductive choices and who will need to be reassured that appropriate testing can be carried out on their future offspring.

In HFE haemochromatosis, molecular analysis of the *HFE* gene (and, formerly, the tightly linked class I HLA class typing) may assist in assessing the risk of disease, particularly in asymptomatic siblings. Phenotypic screening, however, is useful at the level of clinical evaluation for evidence of liver disease, hypogonadism, arthritis, pigmentation, and diabetes. Determining the biochemical phenotype first involves assay of the serum parameters of disordered iron metabolism. Since the serum parameters may be abnormal before iron-mediated tissue injury has occurred, tissue biopsy should be offered to patients with serum ferritin concentrations in excess of 500 µg/l who are at risk from the liver disease.

In first-degree relatives, in whom HLA typing or molecular analysis of the *HFE* or *TfR2* genes indicates a genetic predisposition to the disease, periodic re-evaluation is needed by clinical and biochemical testing at intervals of not more than 5 years. In members of families affected by haemochromatosis due to mutations in the *HFE* or *TfR2* gene who were not found to carry the predisposing mutations and whose ferritin and iron parameters are normal, liver biopsy is not mandatory and the risk of the development of significant iron-storage disease in less than 5 or 10 years is extremely low. In patients with no known pregenetic disposition and normal tissue-biopsy findings, further follow-up screening is not indicated.

From the foregoing it can be seen that there is an urgent need to identify the gene responsible for juvenile haemochromatosis, accompanied by a complete genotype/phenotype evaluation for other forms of haemochromatosis. Unfortunately, no genetic locus has yet been identified for neonatal haemochromatosis, although this is a subject of continuing research. In at-risk pregnancies, neonatal haemochromatosis may be occasionally recognized by MRI during the third trimester, which may show increased iron signals in the fetal liver. After birth, biopsy of the oral mucosa on the gums or inner lip may reveal histological evidence of iron storage in minor salivary glands of affected infants.

Environmental cofactors and disease expression

Many patients with adult haemochromatosis give a history of excessive current or prior alcohol consumption. In the past, physicians have been tempted to attribute evidence of excess tissue iron in these individuals solely to the consumption of alcohol. In practice, however, it appears that those individuals who have biopsy-proven evidence of hepatic iron storage usually prove to carry two predisposing alleles of the *HFE* gene and therefore have true haemochromatosis. Although no clear predictors for the expression of disease in first-degree relatives at risk are available, disease expression is reduced in women of reproductive age; most practising clinicians consider that age and alcohol consumption are the main identifiable environmental factors that contribute to disease expression in predisposed homozygotes. Other comorbid factors, including heritable factors, that may influence the expression of *HFE* mutations in homozygous subjects, include the presence of adult coeliac disease. There are few data that define the relationship between haemochromatosis and coeliac disease but subclinical coeliac disease may ameliorate the long-standing effects of iron loading in *C282Y* homozygotes. Co-segregation of haemochromatosis and coeliac disease has not hitherto been reported.

Treatment

Since it is the toxicity of iron that is responsible for the manifestations of all forms of haemochromatosis, treatment is directed to the removal of iron at the earliest possible stage.

Venesection

In adult and juvenile haemochromatosis the preferred method of treatment is iron depletion by means of phlebotomy. This is best instituted by the removal of approximately 500 ml of venous blood each week by needle puncture of peripheral veins in the antecubital fossa. In young patients it may be possible to increase the frequency of venesection to twice per week after several once-weekly procedures. In elderly patients and those with hypoalbuminaemia as well as end-organ failure and heart disease, the frequency of venesection should be commuted to within the rate tolerated. Coincidental inflammatory disease may impede the erythropoietin-mediated drive to haemopoiesis, and, particularly in the early phases of treatment, mild haemorrhagic anaemia may ensure. Thus adjustments need to be made according to the early responses to venesection therapy, and regular monitoring of the haemoglobin concentration or haematocrit is advisable.

Difficulties may arise in delivering this deceptively simple treatment as a result of poor organization of health service provision and of the availability of suitable healthcare personnel to carry out the venesection procedure. Venesection should not be carried out by naïve or incompetent medical and nursing staff. Every practical effort should be made to ensure that the procedure is convenient for the patient—who is often a young or middle-aged person in full-time employment and who may find regular access to the treatment centre problematic. In cold weather, or in patients with poor circulation or inconspicuous superficial venous access, the use of local anaesthetic creams (such as EMLA cream™) or even local diffusable preparations of glyceryl trinitrate, applied 30 to 60 min before the venesection procedure may greatly improve venous access. Likewise, the simple technique of immersing the arm in warm water to improve peripheral blood flow may be critical for establishing confidence in treating staff. Since patients with haemochromatosis usually harbour a large burden of iron requiring repeated phlebotomy over a period of several years, every effort should be made to preserve the integrity of their peripheral veins. In the author's view, the use of a local anaesthetic is usually unwarranted since it involves further tissue invasion in the region of the antecubital fossa with needles; moreover, repeated injections of the irritant fluid often leads to sclerosis around the venous access site.

Duration of venesection therapy

One 500-ml unit of peripheral blood contains approximately 225 mg of elemental iron. Thus most patients with established haemochromatosis will require weekly phlebotomy for a period of 2 to 3 years. The objective of this treatment is to restore serum ferritin concentrations to within the low normal range and, if possible, to induce a mild iron-deficiency anaemia of approximately 11.5 g haemoglobin/dl. Having thus achieved a satisfactory depletion of body iron stores, interval maintenance phlebotomy, carried out according to ferritin measurements, four to six times per year is usually sufficient to maintain normal iron stores with a serum ferritin concentration less than 100 µg/l. Some authorities suggest that serum ferritin values below 30 µg/l should ideally be achieved. In patients with juvenile haemochromatosis, who have a higher than normal intestinal iron absorption, more frequent phlebotomy may be needed to maintain a healthy iron balance.

Iron chelation therapy

Alternative methods of iron removal are needed for patients with severe clinical manifestations of haemochromatosis, such as life-threatening cardiac arrhythmias and those with severe liver disease and hypoalbuminaemia, who are incapable of withstanding frequent phlebotomy. The preferred alternative involves chelation therapy with the parenteral agent, desferrioxamine. As indicated in Chapter 22.4.4, the subcutaneous administration of desferrioxamine brings about the removal of a maximum of 20 to 25 mg of iron daily and is thus generally less efficient than vigorous weekly phlebotomy. However, desferrioxamine may gain access to cellular pools of iron that are important in the pathogenesis of tissue injury in established iron-storage disease, and therefore may offer particular benefit in patients critically ill with arrhythmias due to haemochromatotic cardiomyopathy. Although the nature of this so-called 'chelatable iron pool' is unknown, there is strong circumstantial evidence that its depletion by means of intravenous desferrioxamine treatment may reverse the life-threatening consequences of terminal iron-storage disease in patients with haemochromatosis. Moreover, the removal of 140 mg of chelatable iron per week represents about two-thirds of the amount that can be removed by weekly phlebotomy. A biological advantage may also be gained by therapeutic access to a reactive low-molecular weight chelatable fraction responsible for the injurious effects of cellular iron overload.

Parenteral desferrioxamine may be given intravenously for life-threatening cardiac disease, as described in Chapter 22.4.4, or, in the non-emergent situation, by subcutaneous infusion using portable infusion pumps for 12 to 14 h five or six times per week. It must be stressed, however, that chelation therapy is not the preferred option for the treatment of established haemochromatosis and should be restricted to those patients unable to tolerate phlebotomy as a result of anaemia or hypoalbuminaemia, or in whom life-threatening cardiomyopathy or liver disease is present.

General measures

Attention should be given in patients with haemochromatosis to the diagnosis and treatment of end-organ failure. This particularly applies to the management of diabetes mellitus by diet and insulin where necessary, as well as hormone-replacement therapy for hypogonadism. (See Chapter 12.8.2.) In men, intramuscular depôt

injections of testosterone enantate (250 mg every 2–3 weeks) are recommended to improve libido and inhibit the development of premature osteoporosis; similarly, conventional sex hormone-replacement therapy should be used in women with premature gonadal failure as a result of haemochromatosis. Cardiac failure in patients with haemochromatosis due to cardiomyopathy and hepatic failure consequential upon pigmentary cirrhosis should be treated by standard methods; organ transplantation may be used successfully but correction of systemic iron overload should be undertaken as soon as practicable to restore normal function in all organ systems. Rarely, end-organ hormone deficiencies result from thyroid infiltration and parathyroid and adrenocortical disease. These deficiencies should be vigorously sought for in the clinical evaluation of the patient at presentation. The appearance of lethargy, faintness due to postural hypotension, or symptomatic hypocalcaemia demands immediate investigation and institution of appropriate replacement therapy.

Prognosis

The main causes of death in untreated patients with haemochromatosis are hepatocellular failure, primary carcinoma of the liver (including hepatocellular carcinoma), and, rarely, cholangiocarcinomas. Cardiac failure due to haemochromatotic cardiomyopathy and untreated diabetes also contribute to death. Although not categorically proven, evidence from retrospective surveys suggest that life expectancy is improved by removing iron from patients with haemochromatosis of whatever cause and the subsequent maintenance of normal iron homeostasis. Most patients experience an improvement in well being on iron-depletion therapy and, during its early phases, there is evidence that hypogonadotrophic hypogonadism may improve with this therapy. Similarly, the manifestations of cardiomyopathy with intractable cardiac failure or tachyarrhythmias can improve after the removal of iron.

The cirrhosis of haemochromatosis appears not to be reversible, although the earlier precirrhotic manifestations of hepatic disease improve greatly on the removal of iron with an apparent restoration of normal life-expectancy. In all patients there is at least a twofold increase in the survival rate at 5 years from the point of diagnosis with the introduction of phlebotomy. In patients studied during the 1950s and 1960s, the 5-year survival rate improved from 18 per cent to more than 65 per cent in all haemochromatosis subjects treated.

In patients diagnosed with haemochromatosis but without cirrhosis, iron-depletion therapy is associated with a near-normal or normal life expectancy compared with a sex- and age-matched control cohort derived from the same population. It is notable, however, that the indolent nature of this storage disorder and the long-term survival of patients who are affected by it has so far rendered long-term controlled studies of the effects of phlebotomy on eventual outcome almost impossible to achieve. However, a wealth of evidence, based on the understanding of the pathogenesis and documented responses to iron depletion in individual patient cohorts, indicates that early removal of iron is highly desirable: indeed it may be decisive in determining a good outcome from all forms of human iron-storage disease—including all subtypes of hereditary haemochromatosis so far established.

Hepatocellular carcinoma occurs mostly in patients with iron-storage disease who have established cirrhosis (which is irreversible). Although hepatocellular carcinoma and cholangiocarcinoma have been reported in non-cirrhotic patients with haemochromatosis, these are rare phenomena. Moreover, since all the evidence suggests that patients with haemochromatosis are more likely to have diabetes mellitus and other manifestations of the disease, every encouragement should be given to the prompt diagnosis of the condition and early institution of iron-depletion therapy.

Increasingly, it has been recognized that the arthropathy of haemochromatosis can be disabling whether or not it is associated simply with joint pain (arthralgia) or progressive and non-inflammatory joint destruction. The disease is associated with a loss of cartilage and, in many large joints, chondrocalcinosis. Although the response of the arthropathy to iron-depletion therapy is controversial, the weight of observation indicates that, once established, the arthropathy of haemochromatosis progresses independently of body iron status and of iron-depletion treatment. It seems intrinsically likely that effective removal of excess body iron stores before the development of joint symptoms will prevent their onset and progression. However, at present only cross-sectional data are available to support this contention.

In summary, observations in adult haemochromatosis suggest that once the disease is established in association with cirrhosis or diabetes mellitus, it diminishes life expectancy. The prognosis for cardiomyopathy in juvenile haemochromatosis is very poor but it may be improved by early diagnosis and the early institution of vigorous iron-depletion therapy. In several cases, the outcome has been improved by allogeneic cardiac transplantation. In adult patients with established pigment cirrhosis, hepatic transplantation has been undertaken and, provided the other systemic manifestations of haemochromatosis have been adequately treated, the procedure is associated with a good overall prognosis.

Prevention and control

The importance of early recognition and the institution of iron-depletion therapy in all forms of haemochromatosis cannot be overemphasized. Molecular analysis of the *HFE* gene or HLA class I haplotype screening, together with biochemical characterization using serum transferrin iron saturation estimations and serum ferritin concentrations, has the power greatly to assist in the detection of presymptomatic first-degree relatives of patients with haemochromatosis.

In relation to whole populations in which mutations in the *HFE* gene are frequent, the health implications based on mass screening remain contentious. Superficially, adult hereditary haemochromatosis due to mutations in the *HFE* gene appears to be an ideal condition for DNA-based mass-population screening. The condition is attributable to a single gene, and a single mutation of diagnostic significance is prevalent (gene frequency 5–10 per cent). Disease-related mutations in *HFE* (especially *C282Y*) are easily tested for by means of polymerase chain reaction-based techniques. At the same time, HFE-mediated haemochromatosis has a long incubation period without symptoms—and all the evidence suggests that the institution of treatment for presymptomatic disease is cheap, simple, and effective.

On the other hand, however, genetic identification of at-risk individuals is associated with problems of stigmatization, increased anxiety, and potential life-insurance weighting—all of which are familiar aspects in well-rehearsed debates about genetic testing in the general population. These aspects must be considered, together with the age-related penetrance of the homozygous state for *HFE C282Y* variants and, as yet, unknown combined genetic and environmental influences on disease expression. Uncertainty as to the significance of these factors has held back the introduction of mass-population screening by DNA-based methods. In light of the present state of knowledge, it is clear that homozygosity for the *C282Y* allele of *HFE* cannot be considered to be tantamount to a diagnosis of hereditary haemochromatosis.

More information is needed from outbred populations, rather than from homozygotes identified as a result of screening family members of index cases having full-blown clinical disease. Family studies provide a false measure of disease expressivity, presumably as a result of shared environments and of the co-segregation of potential disease-modifying genes within defined pedigrees. Finally, it must be emphasized that difficulties also occur for the evaluation of the burden of haemochromatosis in the population at large: although there are definitions of iron-storage disease that reflect the abnormal biochemical genotype, the manifestations of the clinical disease are variable and protean. Moreover, as pointed out earlier, no internationally agreed case definition of haemochromatosis exists, which creates additional difficulties for the introduction of public health measures and appropriate policy review of nationwide screening procedures.

Future directions

Although startling progress has been made in the discovery of many components that serve to regulate iron homeostasis in humans, more information is needed before a full molecular understanding of the mechanisms of iron homeostasis can be achieved. The genes responsible for severe juvenile neonatal and variant forms of adult haemochromatosis are yet to be characterized. At the same time, the functional interactions of the HFE molecule with the two transferrin receptor isoforms requires more study. An even more challenging task will be the identification of the environmental cofactors that determine the expression of iron-storage disease in genetically predisposed individuals; alcohol is a long-standing candidate but the mechanism by which it leads to increased delivery of toxic iron to the tissues is, at present, completely unknown.

Newly identified iron-storage diseases

By general agreement, the term 'haemochromatosis' is used to describe systemic syndromes of pathological iron storage that affects many tissue and disturbs the function of diverse organ systems. Conversely, several distinct clinical syndromes of local iron toxicity have been identified, especially in the eye and brain. Although these syndromes are individually rare, they are important because they are potentially accessible to measures that reduce cellular free iron (for example, metal chelation (see above)), and because they demonstrate the central importance of metabolic iron in selected tissues. A fuller understanding of these disorders and the cognate cell metabolic pathways they affect, may well shed light on ill-understood aspects of tissue iron physiology. Additional information is available by reference to the online Mendelian Inheritance in Man (OMIM) website at www.ncbi.nlm.nih.gov/omim.

Hereditary hyperferritinaemia cataract syndrome (OMIM 600886)

The sole clinical manifestation of this condition is of congenital bilateral ferrugineous nuclear cataracts due to the disposition of excess ferritin light-chain polypeptide in the ocular lenses. The serum ferritin concentrations are moderately elevated but no evidence of systemic iron storage is found. The disorder is caused by mutations in the non-coding 5' IRE of the ferritin L-chain gene that leads to unregulated translational overexpression of ferritin light chains. These polypeptides accumulate in the lenses and disturb their tissue organization and refractile properties. The hyperferritinaemia cataract syndrome is, as expected for an overexpression disease, inherited as a dominant trait; measurement of serum ferritin concentrations may identify at-risk family members. The gene encoding ferritin light-chain polypeptide maps to chromosome 19q3.3-qter.

Adult-onset basal ganglia disease (OMIM 606159)

A single pedigree has been identified with a dominantly inherited disorder showing features of late-onset extrapyramidal dysfunction resembling parkinsonism or Huntington's disease. Imaging and autopsy studies revealed cavitation of the basal ganglia with deposition of iron and ferritin protein in adjacent tissue, especially in the putamen and the globus pallidus; the macroscopic appearances showed widespread reddish discoloration of affected tissues. This disorder was mapped to chromosome 19q13.3 and a single mutation, a point insertion of a single adenine at nucleotide 461, was identified in exon 4 of the ferritin L-chain gene. The mutation is predicted to disrupt the carboxyterminal sequence of the ferritin light-chain molecule and disturb the iron-binding core of the hetero- or homomeric protein. Serum ferritin concentrations were found to be abnormally low in affected heterozygotes. Although this disorder has so far only been identified in a single large pedigree, it further illustrates the importance of ferritin in tissue iron metabolism and, especially, in selective regions of the brain. This disorder has been termed a 'neuroferritinopathy' and may be the first of several diseases affecting cellular iron pathways in iron-rich brain tissue.

Acaeruloplasminaemia with iron deposition (haemosiderosis) in basal ganglia (OMIM 277900)

This disorder is associated with mild systemic iron deposition and deficiency of the plasma copper-binding protein, caeruloplasmin. Caeruloplasmin has long been known to possess ferroxidase activity, and that it enhances the mobilization and delivery of iron to and from macrophages and hepatocytes: caeruloplasmin promotes iron loading of intact ferritin micelles. Acaeruloplasminaemia, due to mutations in the gene encoding caeruloplasmin on chromosome 3q21–24, is an autosomal recessive trait. The deficiency is associated with diabetes mellitus, dementia, and extrapyramidal features including parkinsonism, with choreoathetosis as well as cerebellar ataxia. MRI shows altered signals in the basal ganglia, and retinal degeneration may be apparent by fundoscopy. Excess systemic iron is demonstrable by examination of liver tissue and the serum ferritin concentration is moderately elevated; however, low serum iron transferrin saturations with hypochromic microcytic anaemia, reminiscent of copper deficiency, are usually present.

Infusions of plasma or purified caeruloplasmin may correct the systemic abnormalities of iron metabolism, but probably do not influence the dementia or the other neurological deficits—at least once these are established. The role of caeruloplasmin replacement or indeed parenteral chelation therapy with desferrioxamine or trientine, especially in the early evolution of the neurological syndrome, has not yet been established. The interplay between copper and iron metabolism is well illustrated by this severely disabling illness. Acaeruloplasminaemia illustrates the particular sensitivity of the basal ganglia to disturbances of iron metabolism. In this context, it is notable that caeruloplasmin expression is abundant in glia in the brain microvasculature juxtaposed to the pigment-containing dopaminergic neurones of the substantia nigra and inner layer of the retina.

Hallervorden–Spatz disease: pantothenate kinase-associated neurodegeneration—OMIM 234200

This disease has been familiar to neurologists and neuropathologists since its original description by two now discredited German neuroscientists of the Nazi period. The clinical features indicate basal ganglia disease and dementia with retinal degeneration leading to optic atrophy. The disorder often presents with equinovarus deformity in children and adolescents; extrapyramidal rigidity preceded by choreoathetosis usually follows rapidly. Dementia, optic atrophy, and generalized seizures occur in the latter stages, and death usually ensues by the age of 30 years. Although late-onset forms of the disease are known, a striking feature is the presence of iron pigment in the basal ganglia and substantia nigra, now easily recognized by MRI. The heredofamilial nature of this syndrome has been known since its first description. Hallervorden–Spatz disease is now known to be an autosomal recessive trait due to mutations in the pantothenate kinase 2 (**PANK2**) gene that maps to chromosome 20p13.

Pantothenate kinase-2 is abundant in the retina and target regions of the brain; it regulates the formation of coenzyme A. Deficiency of PANK-2 would deplete sensitive neural tissues with a high metabolic rate of coenzyme A; the defect may also lead to a consequential accumulation of cysteine, which normally condenses with the enzyme product, phosphopantothenate. In the presence of high concentrations of free iron, excess cysteine may accelerate the formation of cytotoxic oxygen free radicals. For some years, cysteine accumulation has been independently observed in the iron-rich nigrostriatal regions of the brain affected by this disorder. Identification of *PANK2* mutations offers the hope of improved diagnosis of this neurodegenerative disorder, and, more importantly, the prospect of specific therapy using supplementation to enhance local coenzyme A activity and phosphopantothenate concentrations in affected neural tissue.

Further practical information

Many patients' associations and societies exist to serve the needs of patients in their respective countries: International Association of Haemochromatosis Societies.

In the United Kingdom, useful information can be obtained from: The Haemochromatosis Society, Hollybush House, Hadley Green Road, Barnet, EN5 5PR. Fax: 44 (0) 208 449 1363; Email: info@ghsoc.org; Website: http://www.ghsoc.org/

Further reading

Adams PC, Speechley M, Kertesz, AE (1991). Long-term survival analysis in hereditary haemochromatosis. *Gastroenterology* **101**, 368–72.

Beutler E, *et al*. (2002). Penetrance of 845G->A (C282Y). *HFE* hereditary haemochromatosis mutation in the USA. *Lancet* **359**, 211–18.

Bomford A, Williams R (1976). Long-term results of venesection therapy in idiopathic haemochromatosis. *Quarterly Journal of Medicine* (New Series) **XLV(95)** 611–23.

Bulaj ZJ, *et al.* (2000). Disease-related conditions in relatives of patients with hemochromatosis. *New England Journal of Medicine* **343**, 1529–35.

Burke W, *et al*. (1998). Hereditary hemochromatosis. Gene discovery and its implications for population-based screening. *Journal of the American Medical Association* **280**, 172–8.

Camaschella C, *et al.* (2000). The gene TfR2 is mutated in a new type of haemochromatosis mapping to 7q22. *Nature Genetics* **25**, 14–15.

De Gobbi M, *et al.* (2002). Clinical expression of juvenile haemochromatosis compared with HFE C282Y and HFE3 haemochromatosis. *British Journal of Haematology* (In press)

Fargion S, *et al.* (1992). Survival and prognostic factors in 212 Italian patients with genetic haemochromatosis. *Hepatology* **15**, 655–9.

Feder JN, *et al.* (1996). A novel MHC class I-like gene is mutated in patients with haemochromatosis. *Nature Genetics*, **13**, 399–408.

Finch SC, Finch CA (1955). Idiopathic hemochromatosis, an iron storage disease. Iron metabolism in hemochromatosis. *Medicine (Baltimore)* **34**, 381–430.

Fleming ME, *et al.* (1999). Mechanism of increased iron absorption in murine model of hereditary haemochromatosis: increased duodenal expression of the iron transporter, DMT-1. *Proceedings of the National Academy of Sciences, USA*, **96**, 3143–8.

Griffiths W, Cox T (2000). Haemochromatosis: novel gene discovery and the molecular pathophysiology of iron metabolism. *Human Molecular Genetics* **9**, 2377–88.

Kelly AL, *et al.* (1998). Hereditary juvenile haemochromatosis: a genetically heterogenous life-threatening iron storage disease. *Quarterly Journal of Medicine* **91**, 607–18.

Kelly AL, *et al.* (2001). Classification and genetic features of neonatal haemochromatosis: a study of twenty-seven affected pedigrees and molecular analysis of genes implicated in iron metabolism. *Journal of Medical Genetics* **38**, 599–610.

McCance RA, Widdowson EM (1937). Absorption and excretion of iron. *Lancet* **233**, 680–4.

McKie AT, *et al.* (2000). A novel duodenal iron-regulated transporter, IREG1, implicated in baso-lateral transfer of iron to the circulation. *Molecular Cell* **5**, 299–309.

McKie AT, *et al.* (2001). An iron-regulated ferric reductase associated with the absorption of dietary iron. *Science* **291**, 1755–9.

Merryweather-Clarke AT, *et al.* (1998). The effect of HFE mutations on serum ferritin and transferrin saturation in the Jersey population. *British Journal of Haematology* **101**, 369–73.

Niederau C, *et al.* (1996). Long-term survival in patients with hereditary haemochromatosis. *Gastroenterology* **110**, 1107–19.

Roetto A, *et al*. (1999). Juvenile hemochromatosis locus maps to chromosome 1q. *American Journal of Human Genetics* **64**, 1388–93.

Sheldon JH (1935). *Haemochromatosis*. Oxford University Press, London.

Simon M, Bourel M, Genetet B (1977). Idiopathic hemochromatosis: demonstration of recessive transmission and early detection by family HLA typing. *New England Journal of Medicine* **297**, 1017–21.

11.7.2 Wilson's disease, Menke's disease: inherited disorders of copper metabolism

C. A. Seymour

Copper

Copper is ubiquitous and is present in relative excess in most diets. It is a prosthetic element of many metalloenzymes, playing a vital role in mitochondrial energy generation, melanin formation, and cross-linking of collagen and elastin. Since the liver efficiently maintains copper homeotstasis by regulating excretion of copper into the bile, acquired copper deficiency is rare. Excretion of copper is dependent on its incorporation into caeruloplasmin (the major copper-binding protein in plasma, normally 20–40 mg/dl); and defects in this homeostatic mechanism lead to toxic accumulation (Fig. 1).

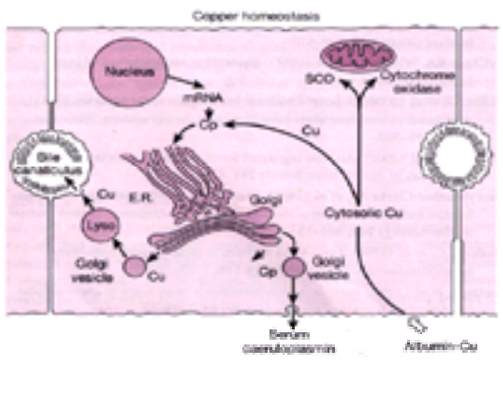

Fig. 1 Hepatic caeruloplasmin isoforms—copper transport in the liver.

Copper homeostasis

Total body copper is in the range of 50 to 150 mg, of which some 8 per cent is found in the liver. Relatively high concentrations are also found in the brain, kidney, heart, and bone. Neonatal and fetal liver tissue tolerate much higher quantities than does the adult liver, where copper is stored within lysosomes in association with metallothionein.

Copper (2–5 mg daily) absorbed from the diet is transported loosely bound to albumin. This accounts for about 10 per cent of circulating copper. Net uptake of copper (around 40–60 per cent) reflects its differential binding to low molecular weight ligands present in saliva, gastric and duodenal juice, and high molecular weight ligands present in bile. The regulatory mechanisms involved in intestinal copper transport are still unknown, although binding occurs to two cytosolic proteins, one similar to superoxide dismutase and metallothionein. Metallothionein and active transport of copper–amino acid complexes are involved in absorption but maintenance of copper homeostasis depends uniquely on its excretion in bile (1.5–1.7 mg/day); only about 0.7 µg/day is normally excreted in the urine. Any interruption in the secretion of bile leads to accumulation of copper in the liver.

Copper overload or toxicity

Copper toxicity occurs naturally in animals: the Bedlington terrier, Domenican toad (*Bufo marinus*), the mute swan (*Cygnus olor*), Long-Evans cinnamon rat, and in experimental animals given copper or in acquired copper toxicity in sheep. The Long-Evans cinnamon rat, an inbred mutant strain with autosomal recessive inheritance, most closely mimics Wilson's disease. In all these models, copper accumulates in the liver and not in the nervous system—as it does in man.

Chronic copper toxicity in man occurs in two major forms:

1. A primary (inherited) form, where copper accumulates in and damages the liver initially, and later the nervous system and other tissues, giving rise to hepatolenticular degeneration or Wilson's disease.
2. A secondary (acquired) form, where copper accumulates in similar amounts to Wilson's disease as a consequence of cholestasis, due either to biliary atresia or Indian childhood cirrhosis (congenital) or primary biliary cirrhosis (acquired). In chronic active hepatitis, lesser amounts of accumulated copper exacerbate pre-existing hepatocyte injury.

Wilson's disease

In 1912, Samuel Kinnier-Wilson, described a neurological condition with severe motor (movement) and mental disturbance due to a disorder of the basal ganglia; this was associated with cirrhosis of the liver. Perceptively, Wilson developed the hypothesis that abnormalities in the liver might be caused by 'a morbid agent' (toxin) generated within a cirrhotic liver.

Definition

Wilson's disease or hepatolenticular degeneration is an autosomal, recessively inherited disorder arising from an abnormal gene located on chromosome 13. Pathognomonic features of the disease are inadequate biliary excretion of copper (less than 1.5 mg/day), reduction in plasma caeruloplasmin (less than 200 mg/l), the major copper-carrying protein in plasma, with reduction in incorporation of copper, and consequent hepatic accumulation of copper (more than 25 µg/g dry weight). The genetic defect results in copper accumulation, initially within hepatocytes, with damage leading to cirrhosis, portal hypertension, and then as the liver is bypassed, to increased copper in the circulation which deposits in the central nervous system, cornea, kidneys, and other organs. Undiagnosed or untreated, the disorder has a fatal outcome within a few years of the onset of symptoms. Removal of copper (e.g. by chelation therapy) prevents progression of the disease, and may reverse some neurological and the corneal abnormalities. Regression of the neurological signs has also occurred after liver transplantation.

Incidence

The Wilson's disease gene is distributed world-wide and is present in all racial groups. Although generally considered a rare inborn error of metabolism, it has a prevalence of 1 in 30 000 live births, with an incidence of 15 to 30 per million live births, although this may be a low estimate as some patients are still undiagnosed when they die. Carrier frequency is about 1 per cent. There is no HLA association. A higher incidence has been noted in Jews of Eastern Europe, inhabitants of Southern Italy, Arabs, Japanese, Chinese, and Indians, and populations with high consanguinity, where the frequency may increase to 60 per million births. Because there is a high mortality associated with failure to diagnose the disease, point prevalence rates are lower than frequency at birth.

Genetics and molecular biology

Extended family studies confirmed an autosomal recessive mode of inheritance. Genetic studies of an Israeli-Arab kindred mapped the Wilson's disease gene to chromosome 13 by demonstrating a linkage between the Wilson's disease locus and the red cell enzyme, esterase D. Multipoint linkage techniques using highly variable markers in many families further localized the gene to 13q14–q21. By 1993, a candidate gene for Wilson's disease was identified by several groups, using different positional cloning strategies. The Wilson's disease gene is expressed predominantly in liver (in the transGolgi network), kidney, and placenta and less in brain, heart, lung, and pancreas. It shows functional homology with the Menkes' disease gene. Both genes are predicted to encode copper-transporting membrane p-type ATPases, with characteristic motifs and homology with the heavy-metal transporting ATPases found in bacteria and yeast. Sequence analysis of cDNA predicts that the Wilson's disease protein (ATP7B) has specific metal-binding domains, an ATP-binding domain, a cation channel, and a phosphorylation region which is involved in energy transduction from ATP hydrolysis to copper (cation) transport. Table 1 compares Wilson's with Menkes' disease.

About 70 mutations have been identified in patients with Wilson's disease. Gene deletions, nonsense and splice site mutations, likely to represent null alleles, are associated with a more severe form of the disease. A common mutation in European populations arises from substitution of histidine for glutamine (H1069Q) in the highly conserved ATP-binding region. This is associated with hepatic and neurological disease, and onset at around 20 years. Many other mutations have been reported in various ethnic groups, but none has yet clearly identified particular phenotypes. Detailed genetic and epidemiological studies have suggested that allelic heterogeneity may not be the sole cause of clinical variability of the disease; different ages of onset and disease course have been found in family members with identical mutant alleles. In the United Kingdom, a total of 37 different mutations, including H1069Q, have been reported in 52 British patients, which included 10 patients of mixed ethnic groups; 70 per cent of the mutations corresponded to those described in other Europeans. Thirty per cent of the mutations are not detectable by single mutation tests.

Clinical features (Table 1)

Wilson's disease may present in childhood, adolescence, or early adulthood. Symptoms and signs may be clinically undetectable under 5 years of age, and few present after the age of 35 years, although diagnosis over 55 years has been reported. In 90 per cent of patients, the disease presents with juvenile hepatic disease or with neurological/psychiatric manifestations. In large studies of Wilson's disease patients, initial manifestations were hepatic (40 per cent); neurological (30 per cent); psychiatric (10 per cent); haematological (12 per cent); and renal (1 per cent). 25 per cent of patients have two or more organs involved (usually liver and brain) at the initial assessment.

Clinical presentations

Haemolytic

During the early period (neonatal to under 5 years of age), copper may accumulate in the liver without clinical signs and excess copper in red cells may present as acute haemolysis or as chronic haemolytic anaemia.

Hepatic

Hepatic presentation of Wilson's disease usually occurs at 8 to 12 years. Acute hepatitis, chronic hepatitis/cirrhosis, and fulminant hepatic failure are the three principal patterns of liver disease. Before puberty, symptoms and signs of hepatic dysfunction are common and may mimic acute hepatitis. A diagnosis of Wilson's disease should be considered if these features coincide with abdominal pain and haemolysis, or in children with hepatomegaly, increased serum transaminases, and a fatty liver.

Five to thirty per cent of patients with Wilson's disease present with chronic liver damage which progresses to cirrhosis. In the early stages, patients are vaguely unwell but later develop more specific features of liver dysfunction such as nausea, easy bruising/bleeding, fluid retention, and jaundice. Portal hypertension develops with progressive hepatic insufficiency, splenomegaly, gastro-oesophageal varices, and ascites. In adolescents and older patients, splenomegaly should always raise the diagnosis of portal hypertension and liver disease. A minority of patients present with fulminant hepatitis, encephalopathy, and coagulopathy. Hepatocellular carcinoma is rare in the cirrhosis of Wilson's disease, unlike haemochromatosis.

Neurological

Neurological presentation usually occurs in older patients, between ages 14 and 40 years and are of two general patterns, movement disorders or rigid dystonia. Symptoms may be acute or chronic in onset and rapidly progressive. In all patients there is some degree of liver damage or cirrhosis.

A common presentation in adolescence is with the insidious onset of dysarthria, deteriorating physical performance at school, with clumsiness in using a knife and fork or chopsticks, deterioration in handwriting, and in physical performance at sport. Early physical signs include flexion–extension tremor of the hands, becoming a parkinsonian 'bat's wing' or intention type; abnormal movements become more obvious, with grimacing and choreiform movements. Orolaryngeal dysphagia and sialorrhoea are associated with hypokinesia. Later features include spasticity, rigidity of limbs and neck muscles, and convulsions, which may occur as a presenting sign or on commencement of treatment. Involuntary movement disorders respond to treatment, unlike the spastic-tonic features which mimic Parkinson's disease. Cognitive and sensory functions are usually preserved until a late stage.

Psychiatric

About 60 per cent of patients with neurological features also show evidence of behavioural or psychiatric disorders caused by excess cerebral copper. Adolescents may present with a fall off in intellectual ability at school and/or with truancy. About 20 per cent of patients present with early psychiatric symptoms. Presentations vary from depression, phobias, and compulsive disorders to aggressive and antisocial behaviour. In older patients, anxiety states, intellectual deterioration, and memory loss are more common. These are important to recognize since otherwise the patient may be placed solely in mental health care rather than in the joint care of a neurologist, psychiatrist, and physician who will offer specific therapy.

Ophthalmic

Kayser–Fleischer rings (Fig. 2 and Plate 1) are almost pathognomonic of Wilson's disease. They are due to the deposition of copper in the limbus of the cornea and

appear brown in a grey-blue iris, or grey in a brown eye; they are best seen by slit lamp examination. Similar appearances have been noticed in cryptogenic cirrhosis and with prolonged cholestasis. Rarely, the posterior membrane of the lens is involved, producing the appearance of a sunflower cataract.

Fig. 2 Kayser–Fleischer ring in Wilson's disease. (See also Plate 1.)

Renal

Renal tubular acidosis due to damage by copper in the proximal and/or distal tubules is not uncommon. Aminoaciduria and nephrolithiasis may also occur. Osteomalacia and vitamin D-resistant rickets may result from tubular loss of phosphate.

Joints

Skeletal abnormalities, particularly early osteoarthritis of the spine (Scheuermann's disease), polyarthritis, hypermobile joints, and chondromalacia patellae are recognized features. In the very disabled patient with neurological disease, hypokinesia may lead to flexion contractures.

Dermatological

Rarely, the skin may be hyperpigmented, appearing slightly grey with a bluish appearance of the lunulae of the nails. Long-standing copper excess may increase skin elasticity.

Cardiac/skeletal muscle

Cardiac abnormalities occur rarely, with cardiac hypertrophy associated with interstitial fibrosis, small vessel sclerosis and perivascular myocarditis, and rarely cardiomyopathy, which may lead to congestive cardiac failure. Copper-induced rhabdomyolysis has also been described.

Endocrine disturbances

Endocrine disturbances occur as a result of liver dysfunction (for example, gynaecomastia in men). Women with cirrhosis and copper toxicity have an increased frequency of abortion, stillbirth, premature delivery, and menstrual disturbance. Once chelation therapy has reduced the copper overload, successful pregnancies occur and chelating agents do not appear to harm the fetus or cause fetal copper deficiency. Copper may also injure other endocrine organs (e.g. causing hypoparathyroidism).

Pathology

In Wilson's disease the liver is almost invariably damaged and the histological changes vary with the amount of copper accumulated. The neonatal liver tolerates concentrations of copper that are six to eight times greater than those which injure the adult liver. The evolution of hepatocyte changes after the neonatal period is uncertain. There may be few changes in the liver lobular structure in the asymptomatic patient. Early changes evident on the liver biopsy are pericellular fatty droplet infiltration of the cytoplasm, 'glycogen' degeneration of the nuclei, and copper distributed diffusely in the cytoplasm. Other pathological changes are summarized in Table 2.

Diagnosis

General investigations

Most patients with Wilson's disease have Kayser–Fleischer rings (Fig. 2) and low plasma caeruloplasmin concentrations. Haematological investigations, such as a full blood count and haptoglobin, detect anaemia and haemolysis; mean corpuscular volume, prothrombin time, and clotting studies detect malfunction of the liver. Biochemical investigations will provide additional information. These include increased serum transaminases (altered liver cell turnover) and reduced albumin and urea (disturbed hepatocellular function). Measurement of autoantibody titres (e.g. antimitochondrial antibody) are important to exclude primary biliary cirrhosis.

Specific investigations

These are necessary to confirm the diagnosis of suspected Wilson's disease, and to monitor the efficacy and side-effects of treatment.

Serum caeruloplasmin

This can be measured enzymatically (copper oxidase), by radial immunodiffusion, or by reverse passive haemagglutination. The activity and concentration of this glycoprotein is reduced or absent (less than 200 mg/l) in 95 per cent of patients with Wilson's disease. Some patients have caeruloplasmin levels in the lower range of the normal distribution which overlap with obligate heterozygotes and normal subjects (Fig. 3). Hypocaeruloplasminaemia and acaeruloplasminaemia do not always indicate Wilson's disease. The normal neonatal liver mimics the Wilson's disease patient, with low or absent plasma caeruloplasmin and high hepatic copper concentration; synthesis and secretion of caeruloplasmin to the plasma start during the first 3 to 6 months of life and hepatic copper concentrations decrease within the first 2 years of life.

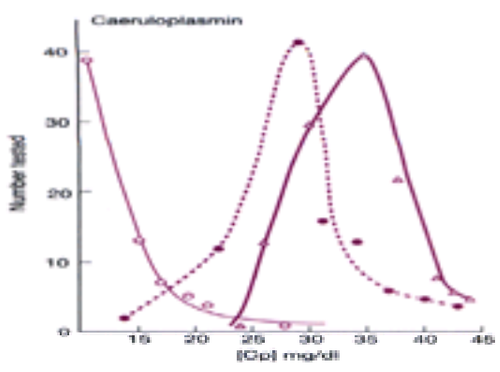

Fig. 3 Serum caeruloplasmin concentrations in controls, heterozygotes, and Wilson's disease.

Caeruloplasmin will also be reduced in protein malnourished states, reduced protein synthesis due to liver disease, protein loss due to the nephrotic syndrome, or a protein-losing enteropathy. Conversely hypercaeruloplasminaemia occurs as an acute phase protein reactant in acute inflammation, infection, pregnancy, and after oestrogen administration. This may increase previously low plasma caeruloplasmin concentrations to within the normal range.

Serum copper

Total serum copper will be reduced or low in Wilson's disease. However, non-caeruloplasmin-bound copper concentrations (loosely bound to albumin or amino acids; normally 50–100 µg/l) will be increased (more than 200 µg/l) in the untreated patient.

Urine copper

Urine excretion of copper is always increased in untreated Wilson's disease (more than 70–100 µg/24 h; normally less than 40 µg/24 h). Increased urinary copper may also occur in other liver diseases such as chronic active hepatitis and primary biliary cirrhosis. However, an increased urine copper in association with a low caeruloplasmin indicates Wilson's disease; the presymptomatic patient may still have a normal urinary copper excretion. Measurement of urine copper excretion is also important in monitoring the effects of chelating therapy where, early in treatment, urine levels may rise to 2000 µg/24 h and fall to less than 100 µg/24 h as the copper overload is reduced. Special care is needed when collecting urine samples to avoid contamination, and copper-free containers are required for the collection.

Provocation of urine copper excretion by giving penicillamine (500 mg orally) may be helpful in patients where the basal urinary copper is equivocal, and to assess the response of the Wilson's disease patient to treatment with chelating agents.

Liver copper concentration

Normal hepatic copper concentrations (15–55 µg/g dry weight of liver) can be measured by spectrophotometric assay, by atomic absorption spectrophotometry, or by neutron activation analysis. Liver biopsy allows measurement of liver copper concentration as well as histological assessment. In children or in the presence of significant liver disease (for example associated with coagulopathy), a transjugular liver biopsy may need to be considered.

In untreated Wilson's disease patients, the hepatic copper concentration is greater than 250 µg/g dry weight, and in heterozygotes, the concentration range is between 55 and 250 µg/g dry weight. Increased hepatic copper concentrations also occur in secondary copper overload conditions such as Indian childhood cirrhosis, primary biliary cirrhosis, sclerosing cholangitis, and chronic active hepatitis. These can be readily distinguished from Wilson's disease on clinical, biochemical, and histological grounds. Normal hepatic copper concentration excludes Wilson's disease in an untreated patient.

Radiocopper studies

Incorporation of orally administered radiocopper (^{64}Cu, ^{67}Cu, or ^{65}Cu) into caeruloplasmin at 1, 2, 3, and 48 h distinguishes clearly between normal patients and those with Wilson's disease, where little or no radiocopper is incorporated into the newly-synthesized caeruloplasmin.

Radiological imaging

Computer-assisted tomography and magnetic resonance imaging of the liver do not help in the specific diagnosis of Wilson's disease, although these imaging techniques will also detect non-specific associations of Wilson's disease such as splenomegaly and abnormalities in hepatic parenchyma. Central nervous system imaging with computer-assisted tomography or MRI demonstrate generalized cerebral atrophy and abnormalities in the basal ganglia.

Screening of family members

It is important that all first-degree relatives are screened for Wilson's disease once the diagnosis has been confirmed in the index patient. This is an essential part of the management of any patient with Wilson's disease. Screening of children should be after 3 years of age. It should include a history, clinical examination, slit-lamp examination of the eyes, liver enzyme and function tests, and serum caeruloplasmin concentration. If the results are suggestive of Wilson's disease, liver biopsy and a quantitative measurement of hepatic copper should follow (or radiocopper studies where liver biopsy is contraindicated).

Molecular tests

The diagnosis of Wilson's disease should be made on the basis of the clinical presentation, biochemical tests, and confirmed by DNA testing in patients with a high index of suspicion of the disease. Analysis of DNA from whole blood may be carried out to detect the common mutation H1069Q or, in Oriental populations, mutation H714Q. Most patients are compound heterozygotes carrying two mutations for the Wilson's disease gene. If the common mutations are not present, each of 21 exons in the gene must be screened by single strand conformational polymorphism (SSCP) or complete sequencing. Currently, this can only be undertaken in genetic units interested in Wilson's disease, and is not available for screening the general population. As methods for mutation detection improve, screening for this disease may become more practical.

Western blotting of caeruloplasmin isoforms obtained from whole blood or dried blood spots may be another, more simple, way of screening for Wilson's disease in the neonate and children over the age of 2 years.

Pathogenesis of the liver lesion

Copper, in free ionic form, is toxic to hepatocytes in man and in animals. Although the retained copper accumulates in lysosomes, there is no evidence that copper-filled lysosomes are more fragile, as is the case in haemochromatosis. Increased numbers of hepatic lysosomes and mitochondria have been described in untreated Wilson's disease, and lysosomes participate in excretion of copper into bile. Copper-containing mitochondria are more fragile, and it is likely that copper-induced hepatocyte damage results from impaired oxidative phosphorylation due to mitochondrial damage.

Several hypotheses have been advanced to explain the primary defect of failure in biliary excretion of copper. A link has been postulated between this gene defect and reduced plasma caeruloplasmin (Fig. 1). The liver in Wilson's disease does produce caeruloplasmin even when it is undetectable in plasma. Thus defects in this glycoprotein are likely to reflect abnormalities of processing or secretion. Evidence of reduced amounts of hepatic mRNA for caeruloplasmin has not yet been linked to the underlying metabolic defect. Two major molecular forms of caeruloplasmin are changed in Wilson's disease. The 132 kDa (plasma form) is reduced and the 125 kDa (biliary form) is absent, giving further support for a post-translational abnormality. It also suggests that caeruloplasmin may play more than a bystander role in the underlying metabolic defect. It is well established that reabsorption of biliary copper from the intestine in Wilson's disease is not increased. Characterization of the Wilson's disease gene suggests that the abnormality may be caused by an alteration in a specific membrane copper transporter (ATP7B) which may fail to bind copper and caeruloplasmin.

Management

The management of Wilson's disease involves the general care of any patient with liver disease and anaemia, and investigation of the family and siblings of the propositus, as well as specific therapy.

Diet

Strict dietary restriction of copper is not practical, but Wilson's disease patients should know which foods are high in copper. These include chocolate, liver, nuts, mushrooms, legumes, and shellfish. In addition, many Chinese dishes are high in copper and Oriental Wilson's disease patients and vegetarians will need special dietary advice. However, reducing copper in the diet is only an adjunct to the main pharmacological therapy.

Specific therapy

Walsh (1956) was the first to show how effective D-penicillamine was in removing copper from patients with Wilson's disease. The optimum time for treatment is in the early stages, and all patients with Wilson's disease should be treated, even if asymptomatic. Treatment is lifelong, unless the patient undergoes liver transplantation. The aim of treating a patient with Wilson's disease is to reduce toxic copper levels in the body tissues. This can be achieved by increasing the urinary excretion of copper. A negative copper balance should be monitored carefully since, with an increase in urine and faecal copper excretion which exceeds copper intake, increased urinary copper may reflect increased plasma non-caeruloplasmin copper (copper in its most damaging form). The effect of any therapy should be regularly monitored by clinical and radiological assessment, and by biochemical monitoring of abnormal liver enzymes and liver function. The agents commonly used in treatment of Wilson's disease, together with mechanisms of action and side-effects, are summarized in Table 3 and Fig. 4. (See also Plate 2).

Fig. 4 Penicillamine dermatopathy—elastosis perfringens serpiginosa. (See also Plate 2.)

Prognosis

Early diagnosis and chelation therapy is the only way to ensure a good outcome in Wilson's disease. Most symptomatic patients improve and have an almost complete resolution of their symptoms. The prognosis is best in the asymptomatic individuals who are detected early, often by meticulous screening of families of index cases. A poor prognosis is more likely in patients with severe liver damage, acute fulminant hepatic failure, acute neurological disease, and dystonia. In addition, in the presence of cirrhosis, even when copper-depleted, the risk of variceal bleeding and intercurrent infections remains.

Non-compliant patients who discontinue treatment may relapse rapidly and die; they are refractory to reversal of copper overload on restarting chelation therapy. In these patients, only liver transplantation would improve their prognosis.

Copper deficiency

Copper deficiency disorders in man are rare, occur in two forms, and have different clinical features:

1. Genetic: Menkes' disease ('steely' or 'kinky' hair syndrome) due to an X-linked recessive defect of intestinal absorption leading to defective synthesis of important copper-containing enzymes, and typical clinical features with damage to the brain and arteries. The occipital horn syndrome is a milder form of Menkes' disease.
2. Acquired: this may occur in malnourished children, adults with severe malabsorption syndromes, in patients taking regular total parenteral nutrition, or in patients regularly taking chelating agents (penicillamine) as treatment for rheumatoid arthritis. The diagnosis rests on low serum copper measurements, hypochromic microcytic anaemia, and evidence of bone demineralization.

Menkes' disease

Menkes first described the disease in 1962, noting the X-linked recessive inheritance in a family of English–Irish descent. Subsequently, a number of case reports defined the clinical features of this disorder, and pattern of inheritance. Danks (1972) described the defect in intestinal absorption of copper which caused serum copper deficiency and which explained the diverse clinical features.

In 1982, the gene for Menkes' disease was isolated by three different groups. The gene maps to human Xq13 and encodes a p-type copper transporting ATPase (ATP7A). In 1983, Peltonen described abnormalities of copper and collagen metabolism in cultured fibroblasts as a feature of Menkes' disease; an increased rate of copper incorporation and accumulation of metallothionein was shown. The occipital horn syndrome, with survival to adult life, was described in 1983. renatal diagnosis for Menkes' disease was commenced in 1974 by measuring ^{64}Cu accumulation in cultured amniotic cells. Treatment of Menkes' disease with copper histidine commenced in 1989, with some improvement in the neurological symptoms and survival (Table 1).

Definition

Menkes' disease is a multisystem, lethal disorder of intracellular copper metabolism. It presents with the symptoms of neurodegeneration and connective tissue manifestations in skin, hair, and bone. A pathognomonic feature is the sparse, coarse, depigmented hair termed 'kinky' or 'steely' hair syndrome (Fig. 5 and Plate 3). A defect in intestinal copper absorption is associated with defective synthesis of copper enzymes and tissue copper accumulation, and this explains the changes in hair, characteristic facies, skin, and other changes. Importantly, fibroblasts in this disease contain five times the copper of normal cells. The untreated condition results in death, usually under 2 years of age. A milder form of the disease, a possible allelic mutation, is the occipital horn syndrome.

Fig. 5 Appearance in Menkes' disease. (See also Plate 3.)

Incidence

The population frequency has been suggested by Danks to be around 1 in 40 000 per live births in Australia, but, as with Wilson's disease, this may be higher because some patients may remain undiagnosed even at death. More recent estimates suggest an incidence of 1 in about 300 000 live births in European countries, with an estimated mutation rate around 1.96×10^{-6} based on isolated Menkes' disease cases born during 1976 to 1987.

Clinical features

The first symptoms occur between 6 weeks and 6 months of age; major signs are poor growth, mental retardation, and hair abnormality. The disorder can present as a premature birth, and episodes of hypothermia may occur within the first few days or weeks of life. Some babies develop normally until 3 months of age, but many are ill from birth. The disease is progressive, culminating in death by the age of 3 years.

Classical characteristics are an abnormal facial appearance (Fig. 5) with flaccid skin, typical of cutis laxa, defective grey pigmentation of the skin, and lightly pigmented hair. Typically, the hair shaft is twisted to give pili torti, which is termed 'steely' or 'kinky' hair syndrome. Abnormalities of major arteries occur, giving thickened, tortuous vessels due to abnormal elastin fibres. Vascular complications, particularly subdural haematoma, can occur. Abnormalities in bone include Wormian bones in the skull and osteoporosis with widened metaphyses, which may fracture. Child abuse may be wrongly considered in these patients.

Progressive focal cerebral and cerebellar degeneration are also typical of the disorder, with associated convulsions and mental retardation. In the milder case of Menkes' disease, the occipital horn syndrome, ataxia and the complications of bladder or urinary diverticulae, hernias, and skin or joint laxity are prominent.

Genetics

Pedigree studies have confirmed Menkes as an X-linked recessive disease. The gene has now been mapped to Xq13.3 (Table 1).

Pathogenesis

Defective intestinal copper absorption leading to a copper deficiency and defective activity of copper-containing enzymes explain the clinical features of connective tissue and arterial abnormalities: tyrosinase (depigmentation and skin pallor); lysyl oxidase (abnormalities in arterial intima due to lack of formation of cross-links in elastin and collagen); monoamine oxidase ('kinky' hair); cytochrome c oxidase (hypothermia); and ascorbate oxidase (skeletal demineralization). The diagnostic hair abnormality probably arises from defective formation of disulphide bonds in keratin, another copper-dependent process. In Menkes' disease, most tissues including intestinal mucosa, kidney, placenta, and testis, with the exception of the liver, have an increased copper concentration. ^{64}Cu studies have demonstrated defective cellular efflux of copper.

Pathology

Major changes are in the arteries and central nervous system. In arteries, degenerative changes lead to aneurysmal dilatation, rupture or stenosis, and areas of intimal proliferation and thrombosis. In the larger arteries, fragmentation of the internal elastic lamina occurs, probably secondary to defective elastin formation. In the later stages of the disease, brain infarcts and haemorrhage occur.

Neuronal destruction and gliosis are found in the cerebral cortex. In the cerebellum, Purkinje cells are destroyed. Myelin is also defective. Changes in skeletal muscle, the iris, and retina have also been noted.

Diagnosis

Although hair changes may not be evident in the first few months of life, the initial diagnosis is based on the principal clinical features of Menkes' disease.

Biochemical tests

Reduced concentrations of serum copper and caeruloplasmin with reduction in gut mucosal and hepatic copper in association with the clinical features are diagnostic. Liver histology is normal but reduced concentrations of liver cuproenzymes, such as cytochrome oxidase and superoxide dismutase, are found in infants over 3 months of age. Accumulation of ^{64}Cu in cultured skin fibroblasts after a 20-h pulse and impaired efflux after a 24-h pulse confirms the diagnosis. Laboratories need to be skilled in radiocopper assays.

Molecular tests

Following identification of the Menkes' disease gene (*MNK*), fluorescence *in situ* hybridization, Southern blot hybridization, or PCR-based mutation analysis may detect various mutations. Chromosomal aberrations can be investigated by standard chromosome banding techniques and confirmed by *in situ* hybridization.

Prenatal diagnosis in the first trimester by measuring the copper content of chorionic villi using neutron activation analysis is unreliable. Contamination of samples gave false-positive results, as did intracellular copper analysis of cultured amniotic fluid cells. DNA analysis can now be carried out, but mutations must first be identified in the family and the mother's carrier status should be known. However, mutation analysis is a difficult undertaking because of the large size of *MNK*. Carrier detection may be possible by ^{64}Cu uptake in cultured fibroblasts, but, because of random activation of the X chromosome, may also be unreliable.

Management

Only administration of copper as copper histidinate has been shown to be successful in correcting skin, hair, and pigmentation abnormalities. The neurological damage appears to be irreversible unless treatment is started very early, as it is likely that copper damage to neurones is well established *in utero*. Thus, treatment should be commenced as early as possible before irreversible neuronal damage occurs. Late diagnosis treatment may prolong survival in patients up to 10 or 20 years but without significant improvement in the clinical features. Treatment with monoamine oxidase inhibitors, with the aim of correcting defective catecholamine synthesis due to deficiency of monoamine oxidase, has not been successful. What is now required is the means to deliver amounts of copper in an appropriate form, sufficient to replace tissue deficiency, but avoiding copper toxicity.

Prognosis

Prognosis varies with the mutations and clinical severity of the disease. Untreated, patients are unlikely to survive beyond 2 to 3 years of age. Use of copper histidine, particularly when commenced early, has prolonged survival to 20 years, with improvement in most of the clinical features except those due to neurological damage. In the occipital horn syndrome survival is longer.

Further reading

Wilson's disease

Brewer GJ *et al*. (1998). Treatment of Wilson's disease with zinc: XV long term follow-up studies. *Journal of Laboratory and Clinical Medicine* **132**, 264–78. [Review of zinc therapy.]

Gollan JL, Gollan TJ (1998). Wilson's disease in 1998: genetic diagnostic and therapeutic aspects. *Journal of Hepatology* **28**, 28–36. [Review of diagnostic aspects of Wilson's disease in relation to clinical presentation and monitoring of treatment.]

Gow PJ *et al*. (2000). Diagnosis of Wilson's disease: an experience over three decades. *Gut* **46**, 415–9. [Clinical experience of Wilson's disease.]

Hoogengraad T (1996). *Wilson's disease.* Saunders, Philadelphia. [Good general clinical and historical review of all aspects of Wilson's disease.]

Roberts EA, Cox DW (1998). Wilson's disease. *Baillère's Clinical Gastroenterology* **12**, 237–56. [Recent update on Wilson's disease.]

Menkes' disease

Christodoulou J *et al*. (1998). Early treatment of Menkes disease with parenteral copper (sic)-histidine: long term follow-up of four treated patients. *American Journal of Medical Genetics* **76**, 154–64.

Danks D (1995). Disorders of copper transport. In: Scriver CR, Beaudet AL, Sly WS, Valle D, *et al*. eds. *The metabolic basis of inherited disease*, Vol. 2, pp. 2212–35. McGraw-Hill, New York.

Menkes JH (1999). Menkes disease and Wilson's disease: two sides of the same coin. *European Journal of Paediatric Neurology* **3**, 243–53. [Comparison of the two diseases.]

Tanner MS (1999). Disorders of copper metabolism. In: Kelly DA, ed. *Disease of the liver and biliary system in children*, pp. 167–85. Blackwell Science, Oxford.

Tümer Z and Horn N (1997). Menkes disease: recent advances and new aspects. *Medical Genetics* **34**, 265–74.

11.8Lysosomal storage diseases

T. M. Cox

Definition

Lysosomal storage diseases represent the consequences of disordered lysosomal digestive function. The lysosome is an intracellular organelle that serves to degrade biological macromolecules derived either endogenously, from metabolism or cell structures, or from the breakdown of exogenous material incorporated by endocytosis.

Since lysosomes are found in most cells, disruption of their function leading to the storage of undegraded macromolecules usually affects many tissues. More than 40 lysosomal disorders are now recognized; they represent single or multiple defects in the organellar complement of specific acid hydrolases, their activator proteins, membrane proteins, or the intrinsic carrier proteins that transport the substrates or products of lysosomal digestion.

Lysosomal diseases, though often considered to be very rare, occur with a surprisingly high birth frequency. About one in 5000 live-born infants have a lysosomal storage disorder; many of these are disabling conditions: Gaucher's disease and Fabry's disease—which are glycosphingolipidoses—are possibly the most frequent. The recent addition of Batten's disease and other ceroid neuronal lipofuscinoses to this burgeoning family of disorders emphasizes the importance of the lysosomal diseases in biology and in medicine: they represent a large and disproportionate burden of illness in the population and for clinical services.

Biological importance of lysosomes

Since their discovery nearly 50 years ago by the medically qualified biologist Christian de Duve, lysosomes and their associated endosomal structures have been at the focus of an impressive body of research into molecular cell biology. De Duve recognized that, with their ready access to extracellular fluid, lysosomes could be used to deliver therapeutic proteins and other agents to the heart of the cell—a prophecy well rewarded with the development of targeted enzyme-replacement therapy for several lysosomal diseases today. The success of these treatments is predicated on detailed knowledge about the delivery of nascent lysosomal proteins to the organelle during its biogenesis, the definition of inherited defects in lysosomal function—and of the development of recombinant DNA technology for the manufacture and the post-translational modification of human proteins for therapeutic use. Greater understanding of lysosomal storage diseases has emanated from research into the metabolism of particular cellular macromolecules that accumulate in lysosomal storage diseases, and into the ebb-and-flow of substrates and products as they pass through lysosomal compartments.

Structure and function of lysosomes

Lysosomes are strikingly diverse in shape and size, but all belong to a single class of ubiquitous organelles containing numerous acid hydrolases (around 40 of which are known) with optimum activity in the pH range, 5 to 6. The matrix space of these organelles is surrounded by a single unit membrane containing transport proteins that facilitate the export of digestion products. Specialized carrier components use energy derived from ATP to maintain the optimal intraorganellar acid pH for enzymatic hydrolysis. The integrity of the lysosome membrane is maintained by highly glycosylated membrane proteins that offer protection from the activated luminal proteases with which they are in contact. Histochemical stains can identify pathological lysosomes in tissue preparations; together with electron microscopy, these methods can be used to detect abnormal lysosomes engorged with storage material in diseased tissues.

Delivery of macromolecules for lysosomal digestion

The lysosome is an intracellular digestive system but acquires complex macromolecules for breakdown by three main pathways: (1) receptor-mediated endocytosis; (2) engulfment and fusion; (3) autophagy. Receptor-mediated endocytosis occurs by means of clathrin-coated pits, a process in which molecules are delivered after internalization to a peripheral, and later to a perinuclear endosomal compartment, 'the endolysosome'. The endolysosome undergoes maturation to form a lysosome after the loss of certain membrane components and further acidification. Some molecules acquired by receptor-mediated endocytosis (for example, low-density lipoproteins) are specifically retrieved and ultimately returned to the cell surface having delivered their cargo of cholesterol within the lysosome. Other molecules that are not retrieved are ultimately degraded by fusion with mature lysosomes and enzymatic hydrolysis (for example, the epidermal growth factor (**EGF**) receptor system).

Lysosomes are also involved in a specialized process for the degradation of exogenous particulates and proteins, including microbes and effete cells such as erythrocytes and neutrophils. Although this engulfment and fusion process involving phagolysosomes is distributed throughout nature, it is particularly active in macrophages and dendritic cells that exhibit active phagocytosis. A specialized phagolysosome variant occurs in osteoclasts that are derived from myeloid cells of mononuclear phagocyte origin. The osteoclastic resorptive vacuole serves as a large exteriorized lysosomal compartment which is independently acidified for the process of bone resorption. In macrophages, cell-surface components that occur on bacteria and yeast, as well as exogenous cells, are recognized and bound by specific receptors on the plasma membrane. The phagocytes engulf foreign material to form large vesicles in which acidification and proteolysis, as well as the secretion of degradative molecules (including reactive oxygen and nitrogen species), is initiated. The phagolysosome fuses with lysosomes and further acidification occurs, so that the acid hydrolases are activated to bring about the breakdown of the ingested material.

Autophagy occurs within cells. It appears that, in a constant process of membrane fusion and flow, organelles (including the endoplasmic reticulum, mitochondria, peroxisomes, and other lysosomes) fuse with lysosomes, by which they are engulfed before breakdown. This process retrieves the basic building blocks of cellular components and proceeds hand-in-hand with *de novo* synthesis and the renewal of intracellular compartments throughout the life of all cells. When lysosomal function is impeded, the breakdown of endogenous organelle-derived macromolecules is impaired; this, together with a failure to breakdown exogenous macromolecules, results in the pathological storage of partially degraded and undegraded material.

Enzyme-replacement therapy for lysosomal storage diseases

Early experiments using fibroblasts obtained from patients suffering from mucopolysaccharidoses such as Hurler's disease and Hunter's syndrome, showed that the rate of degradation—rather than the rates of synthesis or secretion of radiolabelled glycosaminoglycans that accumulate in these diseases—is severely disrupted. In experiments in which fibroblasts obtained from genetically distinct storage disorders were co-cultured, the progressive accumulation of glycosaminoglycan that occurs

when the fibroblasts are cultured separately, was prevented. Biosynthetic labelling experiments showed that degradation of the previously accumulating substrates was restored to normal in these co-culture experiments.

An investigation of this phenomenon by Elizabeth Neufeld and colleagues later showed that each of the fibroblast cultures elaborated and delivered a specific corrective factor to the medium, which ultimately proved to be a high molecular weight form of the specific hydrolases that were lacking in each of the pathological fibroblasts. These corrective factors, when taken up from the medium, restore the normal intracellular degradation of glycosaminoglycans. This secretion–recapture process was shown to be mediated by an unusual carbohydrate component, the so-called 'recognition marker', mannose-6-phosphate. This sugar is found as a terminal moiety derived by post-translational modification and during the secretion of lysosomal enzymes. Neufeld's findings led to the identification of specific receptor pathways for the biosynthesis and uptake of nascent lysosomal proteins by the organelle during the course of organellar formation—and have subtended a great deal of productive biological research. From the therapeutic aspect, however, the experiments immediately raised the possibility of functional complementation of lysosomal storage disorders by supplying particular molecular isoforms of the enzymes that are deficient in the individual storage disorders.

Characterization of lysosomal recognition markers occurred at a time when other cell-surface glycoprotein recognition systems were being identified: for example, the asialoglycoprotein receptor which was implicated by Ashwell and colleagues in the uptake of plasma proteins by parenchymal liver cells *in vivo*. Thus the concept of enzyme replacement for lysosomal storage disorders was established but it was many years before an effective treatment based on glycoprotein targeting was brought into clinical practice. Indeed, the successful use of enzyme replacement was dependent on an understanding of glycoprotein chemistry, receptor-mediated endocytosis, and the molecular cell biology of lysosomal biogenesis—all subjects that themselves depended on a detailed molecular understanding of human lysosomal storage diseases. Identification of the secretion–recapture process in the lysosomal diseases provided the key theoretical underpinning to empirical complementation studies that preceded successful enzyme-replacement therapy. Cellular complementation, by providing a source of wild-type enzyme delivered from allogeneic bone marrow transplantation, has also had spectacular successes in several lysosomal disorders.

Alternative treatments for the glycolipid disorders

For many years it has been recognized that the accumulation of storage material within lysosomes is the precipitating factor for the development of tissue injury and the inflammatory response that accompanies the lysosomal storage disorders. Although it is principally a failure of degradation or export from the lysosome that leads to the excess storage, the concept of depleting the supply of macromolecular substrate to prevent the accumulation of storage material has been developed experimentally and brought to human clinical trials.

Of particular interest has been the discovery that certain iminosugars related to deoxynojirimycin selectively inhibit the glucosyltransferase step as the first committed reaction in the biosynthesis of glycosphingolipids. Experimental studies in cultured cells with pathological storage of glycolipids in lysosomes showed regression of the intralysosomal material after exposure to low concentrations of these natural product derivatives. Later studies demonstrated the reduced storage and delayed symptom onset in experimental animal models of debilitating human glycosphingolipidoses such as Tay–Sachs disease and Sandhoff disease (see Chapter 24.6.1). *N*-butyldeoxynojirimycin, a particular analogue of these iminosugars, has previously been used in clinical trials in an attempt to inhibit the replication of human immunodeficiency virus (**HIV**), as a result of its related inhibitory activity towards a-glucosidases; in these trials, such drugs appear to be relatively non-toxic in large doses. With these data in mind, studies were undertaken to investigate the concept of substrate depletion as a treatment for established glycosphingolipidoses. Evidence of disease regression was obtained in an open-labelled trial of *N*-butyldeoxynojirimycin in patients with Gaucher's disease, as shown by the reduction in visceral enlargement, enzymatic markers of Gaucher's disease activity (plasma chitotriosidase activity), and a slow improvement in haematological parameters. In Gaucher's disease, it was anticipated that the beneficial effects of substrate depletion would be indirect, since they would result from the decrease in the delivery to the macrophage system of glycolipid substrates on the membrane of blood cells. In any event, since the iminosugars are small molecules that penetrate the blood–brain barrier, the possibility of their use (either as a monotherapy or as a synergistic treatment with enzyme therapy) for neuronopathic Gaucher's disease has been raised, as well as for the treatment of the otherwise intractable glycosphingolipidoses that cause severe neurological disease. No effective treatment is currently available for Tay–Sachs disease, Sandhoff disease, and GM1 gangliosidosis, and the juvenile and late-onset variants are thus potential targets for substrate depletion with the iminosugars.

Substrate-depletion therapy depends upon the presence of residual enzymatic activity in the lysosomes. Disturbances of the dynamic equilibrium in the supply and handling of macromolecular lysosomal substrates occur, but most, if not all, patients with glycosphingolipid disorders express residual enzymatic function. At present, clinical trials are underway not only in Gaucher's disease but in patients suffering from severe neuronopathic glycolipidoses such as late-onset GM2 gangliosidosis as well as Niemann–Pick-disease type C. This latter disorder is itself associated with a secondary accumulation of toxic glycolipids within neuronal lysosomes (see below).

The iminosugars appear to be relatively well tolerated apart from causing diarrhoea, probably as a result of impaired biosynthesis of intestinal disaccharidases as a subsidiary effect on oligosaccharide processing. Indeed, experience shows that they appear to be well-tolerated once appropriate dietary restrictions are introduced. However, several cases of peripheral neuropathy have been reported in long-term studies of patients with Gaucher's disease. The sugars are absorbed after oral ingestion and offer the hope of a therapy to arrest the progression of several severe neurological sphingolipidoses that are otherwise beyond therapeutic correction. At present, the outcome of the clinical trials based on promising animal experiments with the iminosugars and other inhibitors of glycolipid biosynthesis derivatives is urgently awaited but few better general strategic options than substrate deletion appear to be available for exploration in patients who might otherwise be without hope.

Pharmaceutical development

Lysosomal storage diseases have been a focus for several prominent therapeutic discoveries. The cooperation of informed patient groups, applied medical research funded by government organizations, and the commercial interest of medium-sized pharmaceutical companies has been promoted by recently introduced Orphan Drug legislation. This legislation has facilitated the early exclusive licensing of products for rare diseases—and has greatly enhanced corporate pharmaceutical investment.

At present, at least six recombinant human enzyme preparations are in clinical use or late stages of development. Indeed, several companies continue to express interest in this area with gene therapy, recombinant proteins, and small-molecule products in active competitive development. There can be little doubt that industry has drawn encouragement from the commercial success of recombinant human products such as erythropoietin, human insulin, and interferon-b and -g. These are top-selling agents that have brought handsome rewards for their parent companies, whose manufacturing patents have several times been vigorously defended. At the time of writing, Genzyme Therapeutics, the United States-based company first involved in the development of targeted enzyme-replacement therapy for Gaucher's disease, is providing treatment to about 3000 patients worldwide. The company reports an annual operating profit of more than $500000000. As described here, many of the 40 or so known human lysosomal disorders are disabling and distressing conditions that cause pain and disability in infants, children and adults of all ages, and for which, until recently, no definitive treatments have been available. The unpredicted magnification of interest that has accompanied successful medical research into this area over a period of less than 50 years has been a model of utility and progress; it continues to provide for many patients the hope that, at last, definitive relief may be forthcoming.

Pathology

Although lysosomal defects occur in all tissues, the principal focus of each disease is observed in those tissues where turnover of the parent macromolecule with impaired degradation is greatest. For example, in Gaucher's disease, the turnover of parent glycolipids appears to be greatest in the mononuclear phagocytes. Here the accumulation of glycolipids derived from the breakdown of membranes present in the formed blood elements occurs; with mild or moderate impairment of the responsible enzyme, glucocerebrosidase, the pathology is restricted to the macrophage-containing tissues of the liver, spleen, bone marrow, and, occasionally, the lung. When inherited defects further impair the activity of glucocerebrosidase, additional pathology is seen in the nervous system where the accumulating glycolipid is derived from the turnover of endogenous neural sphingolipids.

Microscopic pathology shows storage within dilated vesicular spaces, which represent diseased lysosomes. Sphingolipids, being hydrophobic molecules, tend o accumulate in whorls known as 'membranous cytoplasmic bodies' where they assume a lamellar structure within lysosomal spaces. Paracrystalline and crystalline material in distended lysosomes may also be seen under electron microscopy, for example in the accumulation of the charged glycolipid sulphatide that occurs in metachromatic leucodystrophy (arylsulphatase A deficiency). With more water-soluble substrates, granular material accumulates within the vesicular spaces. These spaces represent distended and often fused lysosomes, filled for example with undegraded glycogen macromolecular complexes in acid maltase deficiency (Pompe's disease, glycogen storage disease type II, see Chapter 11.3.1). Pompe's disease was the first lysosomal storage disease to be identified. Its recognition by

Henri-Gery Hers (a colleague of de Duve) led to the concept of 'autophagy' and of the rapid turnover of normal macromolecular components of the cell—including glycogen—by the lysosomal compartment. Pompe's disease has also been the subject of intensive clinical research and early successes have been reported in trials of enzyme-replacement therapy (see the Further reading list).

Although the amount of storage material that accumulates within lysosomes in the lysosomal diseases is several hundred- or thousand-fold greater than normal, the absolute amount of material may amount to only a few grams or so. The quantity of storage material, however, bears no relationship to its pathophysiological effects. In some instances, the presence of a few grams of, for example, the sphingolipid sphingomyelin in Niemann–Pick disease, is associated with massive visceral enlargement with accompanying inflammatory ischaemic and other destructive changes due to the presence of storage cells. Similarly, marked pathological injury occurs: in the viscera and bone marrow spaces of patients with Gaucher's disease; in the heart and skeletal muscles of patients with a-glucosidase deficiency, due to modest glycogen accumulation in the sarcoplasm of striated and cardiac myocytes; and, in the form of neuronophagia, in the brain of patients with Tay–Sachs disease and GM_1 gangliosidosis. Microscopic examination of diseased tissues may also reveal pathognomic storage cells, for example the modified macrophage-derived microglia ('globoid cells') in Krabbe's disease, the striking pathological macrophages of Gaucher's disease, and the foam cells of Niemann–Pick disease types A, B, and C.

Pathogenesis

At present, there is only rudimentary knowledge about the pathological link between lysosomal storage, tissue injury, and the diverse clinical manifestations that accompany lysosomal storage diseases. In the case of the sphingolipidoses, sphingolipids participate in cell-recognition events and receptor biology; sphingolipid metabolic intermediates (lysosphingolipids) also function as signalling molecules in apoptotic and proliferative responses. However, the precise relationship between the storage material and the development of overt clinical disease is not fully understood. In one striking instance (Krabbe's disease due to b-galactosidase deficiency), however, it has been found that the unusual globoid cell can be induced *in vitro* by the pathological lysosphingolipid, psychosine, that accumulates in the diseased tissues. Psychosine and related glycolipids are thus implicated in the final pathological pathway that leads to the disease phenotype. Psychosine itself interacts with a G-protein-coupled receptor on human monocytic-lineage cells. This finding may have profound implications for other lysosomal disorders associated with cell loss and apoptotic as well as inflammatory fibrotic responses. Several indirect studies have indicated the release of inflammatory cytokines in at least one lysosomal storage disease (Gaucher's disease), which may explain the metabolic and plasma protein abnormalities associated with a sustained inflammatory response that characterizes the clinical syndrome.

In a scientific era in which the combined study of gene and protein expression offers a powerful means to understand complex functional networks that lead to tissue pathology, the lysosomal storage diseases represent a promising field for exploration using large-scale, high-throughput methods to investigate altered cell metabolism and signalling responses. An early potential application of this work has been the use of authentic experimental models of some of the more severe storage diseases generated by gene knockout technology; these models facilitate research on otherwise inaccessible tissues such as the brain during the development of the storage phenotype. Gene-expression profiling experiments conducted during periods of neuronal cell death have shown upregulation of genes related to the inflammatory process in the nervous system of mice that serve as a model of GM_2 gangliosidosis. The activation of local microglia is shown by the signature of upregulated macrophage expression markers and lymphocyte chemoattractants, as well as genes encoding antigen-presenting MHC class II molecules. Since this particular GM_2 gangliosidosis is partially ameliorated by bone marrow transplantation that supplies a population of genetically competent immune cells (and which is accompanied by the use of powerful immunosuppressant agents), it seems probable that the altered immunity accompanying bone marrow transplantation may itself modify the clinical expression of lysosomal storage diseases affecting the brain. This may occur without directly affecting the storage material.

Biochemical classification of the lysosomal storage diseases Table 1)

Lysosomal diseases result from inherited defects in lysosomal hydrolases and the mechanisms for delivering them to the organelle; lysosomal enzyme activators and cofactors; lysosomal membrane proteins; and carrier systems for the transport of the substrates and products of lysosomal digestion between the organelle and the cytoplasm. Most of the enzymatic defects are restricted to the activity of a single hydrolase but defects of activators and cofactors, as well as proteins involved in the processing of nascent lysosomal enzymes for organellar delivery, can lead to generalized defects of lysosome function.

As the clinical manifestations for the 40 or more diseases associated with disturbed lysosomal function are very diverse, the reader is referred to specialized literature including the paediatric literature for further information (see Further reading). Broadly, the lysosomal diseases may be associated with: slowly progressive, connective-tissue disease with skeletal abnormalities; neurological symptoms that include deafness and progressive mental deterioration; and visceral syndromes affecting the spleen, liver, lung, kidney, and heart. Ocular disease occurs particularly in the mucopolysaccharidoses and in the neurosphingolipidoses: this may cause corneal opacities or severe retinal changes, including a macular cherry-red spot. Neurological symptoms may have their onset from early infancy to adulthood, beyond even 60 years. In the past, patients with Tay–Sachs disease and other late-onset gangliosidoses were misdiagnosed as having demyelinating disease, Parkinson's disease, or even hereditary ataxia.

Diagnostic features

All lysosomal diseases disturb the catabolism of complex molecules in several tissues. Thus the symptoms are usually permanent and progressive; they show no relationship to food intake and are generally independent of intercurrent illness. Those disorders that affect metabolically active organs, such as the liver and kidney, often cause functional impairment, including the manifestations of liver failure, portal hypertension, and, in the case of the kidney, rickets, metabolic acidosis, and the effects of the renal Fanconi syndrome.

The occurrence of coarse facies, joint stiffness, vacuolation in circulating lymphocytes or neutrophils, with or without associated bone changes and hepatosplenomegaly, suggests the presence of one of the mucopolysaccharidoses or glycosphingolipidoses. Disorders associated with progressive neurological and mental deterioration with visceral and skeletal changes or other somatic disturbances strongly suggest the presence of a glycoproteinosis, such as mannosidosis or fucosidosis, or one of the mucopolysaccharidoses (**MPS**) including Hurler's, Hunter's, or Sanfilippo's disease (MPS I–III, respectively). Skin signs are unusual in lysosomal diseases, but the presence of small angiokeratomas, particularly in the region of the buttocks and genitals, strongly suggests the diagnosis of the sphingolipidosis, Fabry disease (see below), or fucosidosis or the rare condition, Schindler's disease (acetylgalactosaminidase deficiency).

Prominent bone necrotic lesions in association with the early onset of visceral disease, with or without supranuclear ophthalmoplegias, suggest either Niemann–Pick disease type A and C or neuronopathic Gaucher's disease (type III). Ataxia is a feature of GM_1 and GM_2 gangliosidoses, and a flaccid paraparesis in young children might suggest metachromatic leucodystrophy; widespread white-matter disease in association with a frontal dementia is a characteristic presentation of juvenile and adult forms of metachromatic leucodystrophy. Early-onset leucodystrophy is often caused by metachromatic leucodystrophy and Krabbe disease is a rare but important diagnostic entity in this group since the disease may be arrested by allogeneic marrow transplantation in the first 2 months of life. Lysosomal diseases with prominent neurological manifestations are often associated with progressive mental deterioration, with or without the onset of spasticity, myoclonic seizures, and optic atrophy. Extrapyramidal signs including parkinsonism, athetoid movements, and dystonia are frequent in this group of disorders.

Lysosomal diseases are a prominent cause of progressive neurological and mental deterioration in patients whose disease starts during adolescence up to mature adult life, and should always be considered in the diagnostic examination. Corneal opacities suggest cystinosis, I-cell disease, mucopolysaccharidoses, mannosidosis, Fabry disease, and galactosialidosis, as well as one form of Gaucher's disease with neuronopathic features (the D409H type IIIc variant). Specific syndromes are described below.

Diagnostic pathology

As described earlier, the pathological manifestations of the lysosomal diseases are diverse. They may range from enlargement of viscera with infiltration by abnormal macrophages containing storage material (foam cells of Niemann–Pick disease or Gaucher's cells) to bone infarction, neuronophagia, vacuolation of renal tubular cells, and diverse tissue infiltrates. Inclusion bodies may be observed in metachromatic-stained cells of the urine deposit or in circulating neutrophils and lymphocytes (Maroteaux–Lamy disease); staining with a periodic acid–Schiff reagent may reveal diastase-resistant glycolipid storage in the kidney and other organs in Fabry disease and other glycosphingolipidoses. The presence of metachromatic storage material in nervous tissue, including peripheral nerves, is characteristic of the sphingolipidosis, metachromatic leucodystrophy (Fig. 1 and Plate 1).

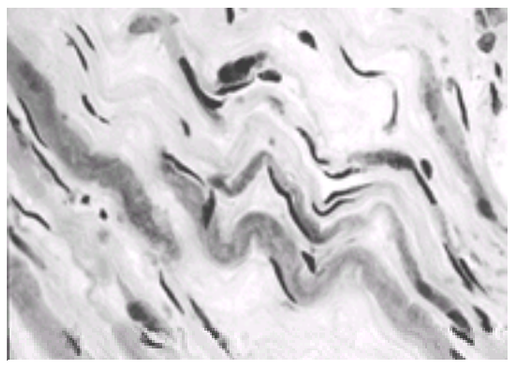

Fig. 1 Sural nerve biopsy stained with toluidine blue from the patient shown in Fig. 2 with metachromatic leucodystrophy. Note the brown-staining granular material within Schwann cells and perineurial macrophages typical of this disorder due to the deposition of the glycolipid sulphatide. (By courtesy of Dr. J. Xuereb, Addenbrooke's Hospital). (See also Plate 1.)

Ultrastructural examination is often diagnostic for lysosomal diseases: membrane-bound vesicles containing storage material that may show a crystalline or concentric appearance, or, in the case of glycogen in Pompe's disease, a granular appearance. The presence of concentric arrays of material strongly suggest a sphingolipidosis. Amorphous material accumulates within the lysosomal vacuoles in the mucopolysaccharidoses and glycoproteinoses. The secondary effects of lysosomal hypertrophy include increased staining for tartrate-resistant acid phosphatase and other lysosomal markers, including intrinsic lysosomal membrane proteins, for example LAMP-1.

Radiology

Ultrasonography, magnetic resonance imaging (**MR**I), and computed tomography (**CT**) may reveal visceral enlargement and infiltration, for example Niemann–Pick disease, mucopolysaccharidoses, Gaucher's disease. Skeletal radiographs may reveal bone expansion in vertebrae and in the phalangeal and long bones, sometimes associated with infarction and collapse, particularly in Niemann–Pick disease type B and Gaucher's disease. Echocardiography may reveal thickening and calcification of the cardiac valves (particularly of the aortic ring), infiltration of cardiac muscle causing ventricular hypertrophy in Pompe's disease, Fabry's disease, and, especially, in mucopolysaccharidoses I, IV, and VI.

Neuroradiology is of value—particularly in patients with mucopolysaccharidoses, and in Morquio's syndrome as well as MPS syndromes I, II, and VI where instability of the atlantoaxial joint may cause fatal subluxation in relation to increased soft tissue surrounding the dens. MRI of the cervical spine in MPS is critical in judging the need for joint stabilization by posterior fusion. Similarly, investigations of the lower spine may determine the cause of progressive spinal deformity due to lumbar kyphosis, and assist in the evaluation of the need for surgical intervention. MRI of the brain is invaluable in the assessment of dementing illnesses: cortical and/or white-matter disease may be delineated. Magnetic resonance imaging is often critical for diagnosing the striking white-matter changes that occur in patients with metachromatic leucodystrophy (Fig. 2).

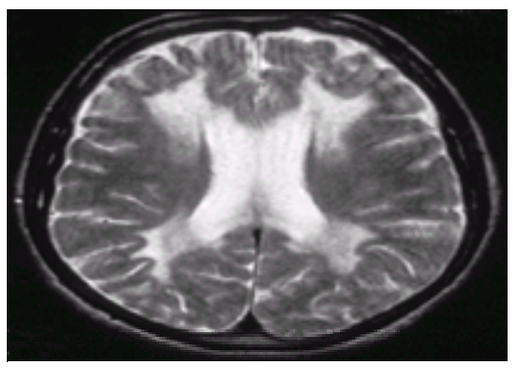

Fig. 2 Magnetic resonance imaging of the brain of a young woman with adult-onset metachromatic leucodystrophy. Notice the high signal intensity especially in the frontal white matter and periventricular regions. This patient presented with bizarre behaviour due to a frontal-type dementia; there are no neurological signs or symptoms. Short-term memory loss and lack of planning and higher executive functions are prominent features of her illness.

Diagnosis of lysosomal diseases

It is critical to establish a definitive diagnosis of a suspected lysosomal disease, even in critically ill patients, for two reasons: these disorders are inherited either as X-linked or as autosomal recessive traits, and the diagnosis may have important consequences for reproductive choice in other family members and for clarifying unexplained symptoms in at-risk relatives. Increasingly, enzyme-replacement therapy, bone marrow transplantation, or even substrate-reduction therapies may be available for their definitive treatment. Furthermore, several disorders respond well to palliative measures including renal transplantation and hepatic transplantation. Finally, a great deal of expertise is available from specialist groups with experience of treating these conditions.

Charitable associations now exist in many countries for members to share their experiences and provide advice and counselling. Above all, invaluable information about available medical services for specific conditions can be obtained through patient organizations. The Worldwide Web provides a useful entry into this, often untapped, resource where key information of importance to both patients and their doctors—and other relevant healthcare personnel—can be obtained.

The key to making the diagnosis of a lysosomal disease is enthusiastic and curious suspicion. In most circumstances, once suspected, the diagnosis can be made with relative ease by referral to a specialized regional reference laboratory for the diagnosis of metabolic disorders; senior laboratory staff will usually advise about the handling of appropriate tissue material for diagnostic studies. In the first instance, simple histochemical stains of existing biopsy material and examination of urine metabolites, including lipids and oligosaccharides, may narrow down the diagnosis. More commonly, specific enzymatic assays are used—generally carried out on leucocytes isolated from fresh heparinized blood samples, or on fibroblasts cultured from small skin biopsies.

Molecular analysis of genes encoding lysosomal enzymes may often support the enzymatic diagnosis, and may, on occasion, provide rough guidance as to the expected phenotype of the disease. DNA-based studies are of particular value for future prenatal diagnosis in a particular pedigree, and for the diagnosis of carrier status in at-risk females for heterozygosity in the X-linked diseases such as Hunter's and Fabry's disease. Lately, there has been a strong and justified trend in favour of specific enzymatic and genetic diagnoses, rather than for diagnoses based on the examination of biopsy material by light microscopy with or without the additional use of special histochemical stains. However, electron microscopy of biopsy material may be of particular value in recognizing the type of disorder but it is rarely essential for a specific diagnosis. Hitherto, histochemical and histopathological methods have led to diagnostic inaccuracies, but it must be admitted that many cases of lysosomal disease, particularly as they affect adults, have come to light as a result in the past of bone marrow examinations, liver biopsies, and other procedures carried out in an attempt to arrive at a diagnosis in an otherwise puzzling condition.

Fabry's disease, Niemann–Pick disease type B and C, as well as Gaucher's disease, have often been diagnosed as a result in young or adult patients who have presented with particular symptoms. General physicians, haematologists, nephrologists, neurologists, gastroenterologists and hepatologists, dermatologists, and even orthopaedic surgeons may be the first to evaluate the patient—all of whom identify the condition by following diagnostic pathways appropriate to their specialty. In any event, the diagnosis of lysosomal storage diseases is rarely difficult, provided the expertise of a trusted laboratory service is available for performing biochemical assays, diagnostic DNA studies, and wide-ranging histopathological examination. The value of good communications between laboratory staff and clinical

investigators to whom these patients are referred cannot be overestimated.

Specific lysosomal diseases

Gaucher's disease

This disorder may occur at any age and is the most frequent of the lysosomal storage diseases. The condition is usually due to a catalytic deficiency of glucocerebrosidase, although rare cases of deficiency of its cognate sphingolipid activator protein (**SAP-C**) may cause a severe disease intermediate between Gaucher's disease and metachromatic leucodystrophy. Numerous mutations responsible for the enzymatic deficiency have been identified in the human glucocerebrosidase gene and the reader is referred to the specialist literature for those genotype/phenotype correlations that broadly apply to this protean disorder.

Rarely, infants are born with an almost complete lack of glucocerebrosidase activity: they die within a few days of birth or are stillborn due to skeletal deformities and/or dehydration as a result of loss of skin integrity (collodion babies). Infantile Gaucher's disease is a rare condition that is associated with death in the first 2 years of life: there is neuronopathic disease with bulbar palsy, opisthotonus, and minor visceral enlargement. This disease is invariably fatal and does not respond to either systemic or intrathecal enzyme-replacement therapy. While neurological disease may occur in children, adolescents, and young adults with Gaucher's disease, it is less severe than in the infantile variant. This clinical variant is associated with supranuclear gaze palsies, myoclonus, ataxia, and, occasionally, seizures. The neurological condition usually deteriorates slowly but is exacerbated if splenectomy is performed for the accompanying splenomegaly and associated pancytopenia.

Where possible, and with vigorous enzyme therapy, splenectomy is best avoided—a partial splenectomy may be carried out to ameliorate pressure effects and life-threatening thrombocytopenia. Subacute neuronopathic disease is not always fatal and often improves with combined bone marrow transplantation and enzyme-replacement therapy (see below). Affected children may show striking visceromegaly, with the associated gaze palsies often playing a small part in the clinical presentation. Although juvenile ('neuronopathic') Gaucher's disease (type III) occurs sporadically in all populations, there is a small isolate in Northern Sweden where all individuals are homozygous for a single point mutation in the glucocerebrosidase gene (*L444P*) that has arisen by descent from a common ancestor.

The most frequent form of Gaucher's disease is the so-called 'adult non-neuronopathic form' (type I). This disease is found in all populations, but is over-represented in Jews of Ashkenazi origin. Although the condition does not affect the nervous system, visceral and skeletal manifestations are prominent. The pathognomic abnormality is the presence of large storage cells, which are activated macrophages (Gaucher's cells), typically found in the splenic sinusoids. The Gaucher's cells (Fig. 3, Fig. 4, and Plate 2) replace the Kupffer cells of the liver, alveolar macrophages of the lung and in the bone marrow.

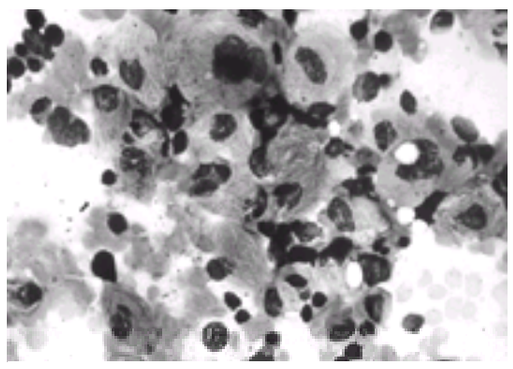

Fig. 3 Light micrograph of a Leishmann-stained bone marrow biopsy obtained from a 23-year-old man with type 1 Gaucher's disease. Note that the large, pale-blue staining Gaucher's cells with striated cytoplasm replace the Kupffer cells of the liver, alveolar macrophages of the lung and of the bone marrow. (See also Plate 2.)

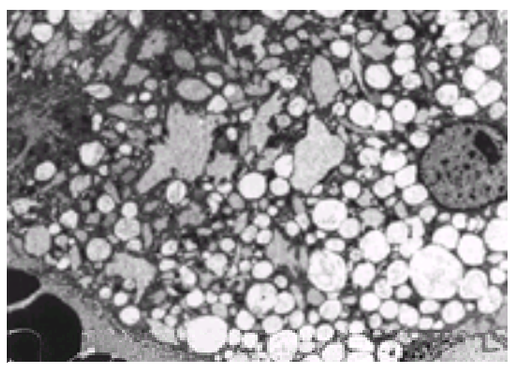

Fig. 4 Electron micrograph showing the cytoplasm of a Gaucher's cell in the spleen of a 56-year-old man removed because of life-threatening thrombocytopenia and pain due to a recent splenic infarct. Note the vesicular spaces filled with fibrillary glycolipid storage material.

Characteristically, Gaucher's disease presents with pancytopenia, with bleeding due to thrombocytopenia and splenic enlargement. Acutely painful episodes also occur in the bones, particularly during growth; these episodes are followed by avascular necrosis of the bone with consequential effects on the integrity of large joints, including the hip, knee, and shoulder (Fig. 5). In the era before enzyme-replacement therapy, splenectomy was often carried out during childhood to relieve the pressure effects of the enlarged organ and to ameliorate the effects of accompanying cytopenias. Although there appears to be an association between splenectomy and the development of severe bone disease, it is unclear as to whether this is directly due to the effects of the splenectomy or the consequential manifestations of disease severity. For this reason, splenectomy is avoided where at all possible. Splenectomy in Gaucher's disease carries a greatly enhanced risk of overwhelming infection; this includes infection with protozoa, such as babesia and malaria, as well as capsulated bacteria, for example the pneumococcus, *Haemophilus influenzae*, and *Neisseria meningitidis*.

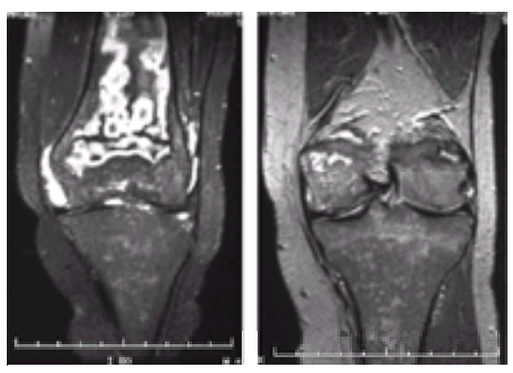

Fig. 5 T_1 (left)- and T_2 (right)-weighted magnetic resonance images obtained from the lower femur and upper tibia of a 30-year-old woman with non-neuronopathic Gaucher's disease experiencing pain due to acute avascular necrosis of bone. Note the geographical areas of increased signal intensity on the T_2-weighted image due to increased tissue water representing oedema surrounding the necrotic tissue. (By courtesy of Professor D. Lomas, Addenbrooke's Hospital.)

Painful episodes may occur in patients with Gaucher's disease either due to infarction of the liver and spleen or to the so-called 'bone crises'. These latter episodes represent acute bone necrosis due to infarction. The increased frequency of infarction events is an important aspect of Gaucher's disease that, as yet, has not been explained. Bone necrosis remains an aspect of the condition that often persists despite enzyme therapy and presents a significant challenge for clinical research.

Gaucher's disease is a truly multisystem disorder, which is accompanied by many ill-understood plasma and metabolic abnormalities. These include a polyclonal immunoglobulin response that may progress to monoclonal gammopathy, amyloidosis, or even frank myeloma. Low-density lipoprotein (LDL) and high-density lipoprotein (**HDL**) cholesterol fractions are abnormal in the plasma. Some lysosomal enzymes are elevated, including tartrate-resistant acid phosphatase, hexosaminidase, and a human chitinase, chitotriosidase. This latter enzyme has proved to be very useful for monitoring Gaucher's disease activity in response to treatment, and may reflect the severity of the disease. The enzyme is elevated sometimes several hundred-fold above normal in untreated Gaucher's disease.

Gaucher's disease may rarely be associated with pulmonary infiltrates, including reticulonodular opacities, restrictive lung defects, and various abnormalities of the pulmonary circulation causing pulmonary hypertension. The hepatopulmonary syndrome, accompanied by platypnoea and associated with severe scarring liver disease or cirrhosis and portal venous hypertension, has also been reported in severely affected adults. The osseous manifestations of Gaucher's disease are very diverse and include the presence of expanded bone lesions (Fig. 6) with surrounding cortical thinning related to Gaucher's cell infiltrates within the bone marrow ('Gauchomas'). Diffuse osteoporosis accompanied by pathological fractures may also compound the skeletal manifestations of Gaucher's disease. Kyphoscoliotic deformity due to crush fractures and vertebral avascular necrosis are common in untreated adults, particularly in postmenopausal women.

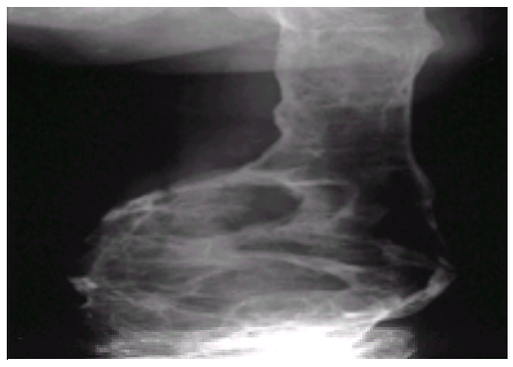

Fig. 6 Expanded lytic lesion at the distal end of the femur in a 44-year-old woman with severe Gaucher's disease complicated by osteoporosis, avascular necrosis, and, as shown, expansile lytic lesions in long bones leading to local infiltration of the marrow space by Gaucher's tissue.

In its untreated state, Gaucher's disease is a miserable condition leading to progressive skeletal deformity, pancytopenia, and visceral enlargement with failing organ function punctuated by painful visceral bone crises. The mean age of death in a single large series reported from Pittsburgh, Pennsylvania, was 60 years during the pretreatment era but this does not take into account the poor quality of life of most affected individuals. Some homozygotes for 'mild' missense mutations in the glucocerebrosidase gene (especially the widespread mutation, *N370S*) may escape detection and remain asymptomatic throughout a long adult life. Detailed investigation reveals only a mild thrombocytopenia and trivial splenomegaly in some cases. However, monoclonal gammopathy is frequently present after the age of 45 years. It is uncertain as to what extent the presence of such mutations in the population at large (homozygosity for *N370S* occurs in about 1 in 960 Ashkenazi Jews) contributes to the development of B-cell lymphoproliferative disorders, such as B-cell lymphoma and myeloma, in this at-risk group.

The diagnosis of Gaucher's disease is based on white-cell acid b-glucosidase activity, which may be accompanied by the elevation of one or more related marker enzymes such as chitotriosidase or tartrate-resistant acid phosphatase in the serum. Spleen tissue, liver biopsy material, or bone marrow aspirates may show the characteristic oligonucleate storage cells demonstrating striated cytoplasm on Leishmann staining (see Fig. 3), but which appear as pink sheets in tissue sections stained with haematoxylin and eosin. Molecular analysis of the glucocerebrosidase gene may identify widespread mutant glucocerebrosidase alleles that cause this disease and may assist in the diagnosis and investigation of family members at risk for this recessive disorder.

Treatment

Until recently, the treatment for Gaucher's disease was palliative. Bone marrow transplantation has been undertaken in a few infants and children with rapidly progressive disease, including those with the subacute neuronopathic form type III. When successful, this may correct most of the systemic manifestations of the condition and restore growth. It may arrest further neurological deterioration. However, bone marrow transplantation is no longer in routine use because of the accompanying severe risk resulting from the procedures and constraints in the supply of donors, especially MHC-matched, first-degree relatives.

Enzyme-replacement therapy was introduced during the early 1990s in the form of a natural product extracted from the human placenta, alglucerase, 'Ceredase'. A recombinant glycoform, imiglucerase, 'Cerezyme', is now available that, like alglucerase, is modified to reveal terminal mannose residues. The recombinant protein is supplied as a lyophilized powder which is reconstituted for intravenous infusion; the preparation is given at variable frequencies, from three times a week to once every 2 weeks, and infused over approximately 60 min. Modification of the sugar residues on this protein facilitates targeting to macrophage-containing organs, in which it complements the enzyme deficiency in the pathological storage cells. After a few weeks of enzyme administration, most patients show an improvement in the blood parameters of disease activity and a reduction of the chronic inflammatory response that accompanies Gaucher's disease: the platelet count rises; and there is a correction of the hypersplenic picture, with a reduction in hepatosplenomegaly and an improvement in the asthenia that complicates Gaucher's disease. Quality-of-life measures also show clear improvement.

Controversy remains as to the appropriate dose of imiglucerase; however, most authorities agree that the administration of the enzyme should be a lifelong therapy. There are two schools of thought as to whether enzyme therapy should be administered at a high dose to start with and then reduced as evidence of disease regression becomes clear, or whether a stepwise increase in this expensive agent can be undertaken in patients with long-standing disease until evidence of disease regression can be established. Disease activity is assessed by objective parameters, including visceral enlargement, and by determination of surrogate biomarkers such as chitotriosidase and blood counts. At present, there is no agreed protocol for therapy in adults with Gaucher's disease; but in patients with the subacute neuronopathic form of the condition, international guidelines suggest that a dose of at least 120 units of enzyme per kilogram bodyweight per month is necessary to secure disease regression.

Although it would not be expected that enzyme-replacement therapy would improve the neuronopathic aspects of Gaucher's disease, there is emerging evidence that clinical improvement may be induced by enzyme-replacement therapy given at this high dosage. Enzyme-replacement therapy for Gaucher's disease is very expensive and the doses recommended range from below 5 to at least 60 IU/kg per month. Thus for an adult, this may cost as much as £200 000 per year, so placing enormous demands on healthcare provision in the long-term. In response to these pressures, there have been initiatives to develop alternative methods to treat the condition—including the use of an oral agent that inhibits the formation of the substrate delivered to macrophages. When taken for several months, the trial agent, *N*-butyldeoxynojirimycin reduces glycolipid storage and clinical as well as laboratory parameters of Gaucher's disease activity.

Enzyme-replacement therapy, although very expensive, is a successful treatment for Gaucher's disease, and, since most patients do express the protein antigen endogenously, hypersensitivity and immune reactions are very rare. Apart from the inconvenience of periodic intravenous infusions, treatment is well tolerated and many patients in Europe and the United Kingdom choose to take their treatment as self-administered infusions at home. In relation to treatment with iminosugars, short-duration unwanted effects (including diarrhoea due to inhibition of intestinal disaccharidases) are frequent, although they usually respond well to dietary adjustments. However, at the time of writing it is unclear as to whether or not *N*-butyldeoxynojirimycin will be licensed as an Orphan drug, although clearly there is an international demand for a more readily available treatment for Gaucher's disease. The occurrence of peripheral neuropathy after long-term administration in a few patients with non-neuronopathic Gaucher's disease may indicate a neurotoxicity, and restrict the indications for use of *N*-butyl deoxynojirimycin. Substrate-reduction therapy may, however, have wider applications in the treatment of certain sphingolipidoses, including late-onset Tay–Sachs disease, GM1 gangliosidosis, and Niemann–Pick disease type C, which affect the brain and for which no other therapy is currently available. Moreover, since the iminosugars may penetrate the brain and are orally active, they may also be of value as an adjunctive therapy in type III Gaucher's disease. Several agents of this class have been found to be effective in

animal models of glycosphingolipidoses and successors to *N*-butyl deoxynojirimycin are actively undergoing preclinical study.

Treatment for Gaucher's disease should include appropriate immunization and antimicrobial prophylaxis in patients who have undergone splenectomy. The widespread osseous manifestations, including osteoporosis, may respond to therapy with bisphosphonate drugs, including pamidronate. Patients may require joint-replacement surgery to ameliorate the effects of bone infarction crises and, in rare instances, liver transplantation for endstage liver disease. Bone marrow transplantation probably does not have a role today, except in rare circumstances. Evidence of metabolic bone disease complicating the disorder should be always sought and osteoporosis should be treated promptly with hormone-replacement therapy, with the additional consideration of oral active or parenteral bisphosphonates. Where present, a deficiency of 25-hydroxyvitamin D should probably be treated with appropriate supplements. Because of the increased risk of infection due to intrinsic chemotactic and phagocytic defects as well as splenectomy, patients with Gaucher's disease undergoing surgery or with systemic infection should be promptly treated—preferably with parenteral antimicrobial agents.

Niemann–Pick disease

Niemann–Pick disease A and B are, respectively, neuronopathic and non-neuronopathic variants of acid sphingomyelinase deficiency, a sphingolipid disorder leading to the accumulation of sphingomyelin. Niemann–Pick disease resembles many of the manifestations of Gaucher's disease with a characteristic secondary storage cell which is also a macrophage. The Niemann–Pick cell has a foamy appearance rather than the characteristic striated cytoplasm of the Gaucher's cell. In Niemann–Pick disease, there is prominent infiltration of the lungs as well as the marrow cavity. At present, no specific treatments are available, apart from the prompt treatment of pulmonary infection and the management of the consequences of skeletal infiltrates and episodes of avascular necrosis. Some patients, including those previously misdiagnosed as having Gaucher's disease, may have undergone splenectomy to relieve pressure symptoms or the haematological effects of hypersplenism.

Niemann–Pick disease type A is associated with disabling neuronopathic features and dementia. At the present time no specific therapy for it exists. Niemann–Pick disease type B may occur in adults who have only trivial splenomegaly and minor pulmonary infiltrates that are only exacerbated at times of intercurrent pneumonic infection; they are at risk from osseous disease related to marrow infiltration, as with Gaucher's disease. Since this disease is primarily a disorder of macrophages, it should be susceptible to enzymatic complementation using the mannose receptor. At the time of writing, clinical research to develop macrophage-targeted, recombinant, human acid sphingomyelinase is well advanced. Unfortunately, no iminosugar derivatives are available for substrate-reduction therapy for the neuronopathic manifestations of Niemann–Pick disease type A, since the biosynthesis of sphingomyelin is not regulated by the uridine diphosphate-glucosylceramide synthase reaction.

Niemann–Pick disease type C (**NPC**) may present with jaundice in infants or children, but the initial hepatitic illness usually resolves. Later evidence of neuronopathic disease occurs, with ataxia, seizures, supranuclear palsy, and progressive diffuse cortical injury. NPC is not due to a primary defect of acid sphingomyelinase but to mutations in two distinct lysosomal membrane proteins. These are responsible for the NPC-1 and NPC-2 subtypes of disease. Although the function of the NPC-1 and NPC-2 proteins is not fully understood, they are implicated in the intracellular transport of cholesterol and cholesterol esters to and from the lysosomal compartment. NPC is also associated with the appearance of foam cells in the macrophages; the Kupffer cells of the liver may be enlarged and a cholesterol trafficking defect is apparent in most cells. Thus the defect, though not manifest in the skin, may be detected in skin and fibroblasts after culture and exposure to low-density lipoprotein-cholesterol: in NPC, cholesterol is taken up and accumulates in intracellular droplets that stain positively with the fluorescent dye filipin. Within the brain, NPC causes neuronophagia and the accumulation of gangliosides and other complex sphingolipid storage products that may induce neuronal injury. Clinical trials are under way with the use of the iminosugar *N*-butyldeoxynojirimycin. These trials follow on from the successful arrest of the disease in mice homozygous for a spontaneous mutation in the *NPC-1* gene that provide a convincing model of the human disease. The outcome of further investigations of these mice and of the trial of this therapy in patients with NPC is urgently awaited. Niemann–Pick disease type C is an intractable condition associated with progressive neurological disease in childhood and early adult life. Biological deterioration progresses inexorably and death usually occurs in the third or fourth decade. The use of statins and other agents that interfere with cholesterol metabolism has not been effective in arresting the course of this cruel illness.

Anderson–Fabry disease (usually shortened to Fabry's disease)

This disease is an X-linked disorder, unlike many of the lysosomal diseases, apart from Hunter's disease (MPS II). An unusual feature of Fabry's disease, is the presence of clinical signs and symptoms in the majority of heterozygous female carriers of the condition. Although these manifestations are usually less severe and of later onset than in affected hemizygous males, florid and life-shortening clinical disease has often been observed (and ignored) in affected women. Deficiency of a-galactosidase causes the accumulation of ceramidetrihexoside (otherwise known as globotriaosylceramide), which principally derives from the breakdown of lipids present in senescent red cells. Affected male hemizygotes have small, raised, red vascular skin lesions (angiokeratomas) particularly around the buttocks and genital region. These lesions are often detected in limited areas of affected heterozygous females and reflect X-chromosome inactivation patterns in the skin. With increasing age, impaired capillary circulation and progressive tubular, interstitial, and glomerular disease in the kidney leads to proteinuria and renal failure. Many patients require renal support, including haemodialysis, peritoneal dialysis, or kidney transplantation. Patients with Fabry's disease have disturbing gastrointestinal symptoms, characterized by diarrhoea shortly after eating and often abdominal pain associated with febrile attacks that are otherwise unexplained.

The most characteristic symptoms of the disease are the onset in early childhood of lancinating pain with background burning sensations in the extremities that are worse with exercise and with extremes of temperature. These attacks can be very disabling and represent neuropathic pain, which is notoriously difficult to control. The acroparasthesias are frequently attributed to Raynaud's phenomenon but this relationship is unclear, although many of the symptoms of Fabry's disease can be explained by neuropathy affecting autonomic nervous tone. Most men with established disease notice a striking absence of peripheral sweating, and impotence is very common. The abdominal symptoms may also result from autonomic neuropathy. High-tone loss of hearing is also a frequent feature of Fabry's disease, which appears to reflect a selective loss of functioning cochlear neurones. Cardiac hypertrophy occurs, especially of the left ventricle, with conduction defects leading to a shortened PR interval and a prolonged QRS complex—later accompanied by tachyarrhythmias and complete heart block. Left ventricular hypertrophy may or may not be associated with hypertension and cardiac embolic disease; disease of capillaries and medium-sized vessels in the brain associated with unusual microvascular changes, particularly in the posterior cerebral circulation, also causes stroke.

Stroke and renal failure are the most common causes of death in patients with Fabry's disease; in men, death occurs at a median age of 48 to 49 years, with a greatly reduced quality of life during the antecedent symptomatic period. Life expectancy in affected heterozygous women is also shortened. Sometimes the lancinating acroparasthesias are sufficient to cause severe depression and even suicide. Disease expression in many carrier females, who may develop renal failure, is accompanied by angiokeratomas restricted to certain dermatomes on careful examination and asymptomatic corneal opacification with whorl-like cataracts on slit-lamp examination.

Diagnosis is made by demonstrating the abnormal glycolipid in urine or plasma, as well as by assay of a-galactosidase in tears, plasma, white cells, or other tissue material. Molecular analysis of the a-galactosidase gene on the long arm of the X chromosome is worthwhile because it allows the unambiguous detection of female heterozygotes and may thus be useful during the reproductive period, particularly for antenatal diagnosis. Despite the presence of active disease, ceramide trihexoside concentrations and a-galactosidase assays are often within normal limits in affected female heterozygotes.

Hitherto, the treatment for Fabry's disease has been palliative, involving the use of anticonvulsants (including gabapentin) for the acroparasthesiae and neuropathic pain. Gastrointestinal symptoms sometimes respond to antimotility agents or to pancreatic enzyme supplements but these agents have not been subjected to control trials. Renal failure is managed by dialysis or by renal transplantation; occasionally, cardiac transplantation has been required for cardiomyopathy; pacemakers and antiarrhythmic drugs may also be needed. There is a rare cardiac variant in this disease, which appears to be predominantly manifested by restrictive cardiomyopathy in elderly patients who have a small amount of residual a-galactosidase activity in their tissues. In one remarkable instance, therapy with galactose infusions appears to have ameliorated this condition by stabilizing the nascent mutant enzyme, thereby enhancing residual a-galactosidase activity with slow clearance of cardiac glycolipid storage.

Recently, enzyme replacement using recombinant human a-galactosidase has been developed as a more definitive treatment. To date, two preparations—which may differ slightly in their post-translational glycosylation status for delivery to endothelial, epithelial, and other cells that represent the pathological focus of this disease—have been licensed: agalsidase-a (Replagal) and agalsidase-b (Fabrazyme). Administration of these preparations to male hemizygotes has improved lipid accumulation in the plasma and in renal biopsy samples from male hemizygotes with this disease. Agalsidase-a has also been shown in double-blind, crossover, placebo-controlled trials to improve clinical endpoints of the disease, including neuropathic pain, stabilization of renal function, and ventricular mass as well as conduction defects that represent infiltrative cardiomyopathy. Further clinical observations continue to be made but it is now clear that enzyme therapy is likely to be a safe and effective long-term treatment for Fabry's disease. Unlike Gaucher's disease, targeting to the affected cells and tissues in Fabry's disease probably results from uptake by the common lysosomal recognition marker, mannose-6 phosphate, as the principal ligand for protein delivery to this organelle.

The glycoproteinoses

These disorders fall into a group of lysosomal diseases associated with impaired degradation of glycoproteins. The most frequent of these disorders include fucosidosis and mannosidosis, which are inherited as autosomal recessive traits. Glycoproteinoses show some of the manifestations of the sphingolipidoses and the mucopolysaccharidoses: they are almost invariably associated with neurological disease and variable systemic manifestations. An unusual clinical feature of these conditions, which are associated with visceral enlargement (particularly hepatomegaly) and mental retardation occurring in childhood, is the appearance of angiokeratomas that are otherwise characteristic of Fabry's disease.

In a few patients, bone marrow transplantation has improved some of the systemic manifestations of mannosidosis and fucosidosis. However, trials of enzyme-replacement therapy have yet to be published beyond the study of experimental models, such as those that occur spontaneously in cattle, dogs, and other large animals. Diagnosis of the glycoproteinoses is usually prompted by the finding of increased excretion of oligosaccharides in the urine and by clinical suspicion of the diagnosis in patients with a multisystem disease that may have an obvious hereditary component. Definitive diagnosis is dependent on enzymatic studies in leucocytes or cultured skin fibroblasts. As with other lysosomal diseases, prenatal diagnosis using cultured amniocytes and chorionic villus cells may be offered by specialized laboratories.

Mucopolysaccharidoses (MPS)

These disorders are caused by a deficiency of lysosomal hydrolases that catalyse the cleavage of complex glycosaminoglycans—macromolecular components of connective tissues including joints, bones, heart, and the major arteries. Clinical manifestations of each of these disorders reflect an individual enzymatic deficiency and the resulting accumulation of mucopolysaccharide derivatives, of which dermatan, keratan, and chondroitin sulphate as well as heparan sulphate are the principal components. In general, the accumulation of the complex substrates that are normally linked to proteins to form proteoglycans is associated with visceral enlargement, as well as bony abnormalities, joint stiffness, corneal clouding, and short stature; the accumulation of heparan sulphate may particularly be associated with the development of brain disease, including thickening of the leptomeninges. Thus hydrocephalus is an often-neglected factor in cerebral impairment that may also be attributed to lysosomal storage affecting neurones of the brain and peripheral ganglia—as well as the retina.

Clinical features and pathology

Typically, these disorders are associated with coarse facial features, bone shortening, and skeletal abnormalities, as well as disturbances of dentition, the gums, and auditory canal. Abnormalities of the tracheobronchial cartilages and upper airways may be associated with respiratory infections. In the heart, the coronary arteries and valves may be infiltrated by glycosaminoglycans, leading to nodular thickening of aortic and mitral valves with clinical evidence of valvular disease malfunction. In some cases, accumulation of glycosaminoglycan occurs in the coronary arteries, which may be occluded. Similar changes may occur in peripheral arteries—particularly those supplying the viscera. In the eye, the basal layers of the cornea show swelling, cytoplasmic vacuolization, and storage granules leading to opacification.

Excess urinary excretion of glycosaminoglycan products, including dermatan sulphate and heparan sulphate, characteristically occur in the mucopolysaccharidoses. This abnormality should immediately prompt further investigations by enzymatic and genetic studies in blood leucocytes and/or fibroblasts obtained from cultured skin-biopsy samples. The inheritance pattern of the mucopolysaccharidoses is typical of autosomal recessive traits—with the exception of Hunter's disease (MPS II, which is due to iduronate sulphatase deficiency) that maps to the X chromosome and, unlike Fabry's disease, is expressed predominantly in boys and men. Female heterozygotes for Hunter's disease only very rarely show evidence of neurological impairment or connective tissue abnormalities.

Treatment of the mucopolysaccharidoses

Palliative treatment is a very important aspect of the management of these diseases, and should include the provision of multidisciplinary support for children and young adults with the accompanying developmental disabilities. Sustained provision for the long-term management of the condition in affected families is desirable.

Surgical procedures

Corneal transplantation may be required to improve vision where retinal degeneration is not dominant. Carpel tunnel syndrome with compression neuropathy of the median nerve is very common in the mucopolysaccharidoses and, when indicated, surgical treatment is often beneficial. Particular care is required in patients with mucopolysaccharidoses such as Hurler's syndrome when surgical procedures under general anaesthetic are required for relief of hydrocephalus, myringotomy, hernia repair, relief of airways obstruction due to laryngeal disease, and corrective spinal or joint surgery. Infiltration of the soft tissues of the upper and lower airways, as well as the heart and cervical spine (which may include subluxation of the atlanto-occipital joint) is associated with high perioperative mortality. Complications thus arise with the administration of a general anaesthetic beyond that of difficulties with endotracheal intubation. In particular, a tracheostomy may be required to avoid life-threatening complications of intubation. An extensive preoperative examination should be conducted when an anaesthetic is required for any procedure, particularly to assess the stability of the atlantoaxial joint, the airway, and the presence of coronary artery disease (that may predispose to perioperative myocardial infarction).

Specific treatment

Bone marrow transplantation using HLA-identical sibling and HLA-matched non-sibling donors has been extensively investigated in the mucopolysaccharidoses. Long-term clinical trials have confirmed the beneficial effects of successful transplantation with reversal of hepatosplenomegaly and obstructive airways disease. In some cases there is improved longevity, with a possible reduction also in the incidence of secondary hydrocephalus. However, at present, transplantation does not cure the condition and is unable to reverse established brain injury and most of the crippling skeletal manifestations of the mucopolysaccharidoses. If it is to be considered, bone marrow transplantation should thus be carried out early in the course of these diseases.

Enzyme-replacement therapy has long been under investigation in MPS I (Hurler's syndrome, Hurler–Scheie syndrome, and Scheie's syndrome), which was one of the first of such disorders to be subjected to intensive laboratory study. Recent studies with recombinant human a- L-iduronidase, given by weekly infusion intravenously, after 1 year clearly show a reduction in lysosomal storage: liver volume decreased; there was an improved rate of growth as well as improvement in the range of joint movements at sites characteristic of connective tissue infiltration in this condition. With a reduction in the storage material in the upper airways, there was also an improvement in episodes of hypoventilation during sleep. After a few weeks of enzyme treatment, urinary glycosaminoglycans abnormalities were corrected. Although a few patients developed serum antibodies, only transient immune reactions, including urticaria, occurred during the infusions. It appears that in patients with this disease, the first to be shown by experiments with fibroblasts *in vitro* to be corrected by enzymatic complementation, treatment with a recombinant human product reduces lysosomal storage and ameliorates several of the important clinical manifestations. Enzyme-replacement therapy is currently under clinical evaluation for patients suffering from MPS II (Hunter's syndrome with iduronate sulphatase deficiency) and MPS VI (Maroteaux–Lamy due to arylsulphatase B deficiency). Favourable responses to enzyme-replacement therapy have also been reported in animal models of related disorders, including the cone head mouse that represents a faithful model of MPS VII (Sly disease), due to deficiency of acid b-glucuronidase.

Although enzyme therapy is in an early clinical phase of application in the mucopolysaccharidoses, there has been a remarkable response from the pharmaceutical industry for the therapeutic development of enzyme-replacement therapy for lysosomal diseases. With the present state of knowledge, bone marrow transplantation seems to be ineffective for many patients with MPS; if it is to be restricted to use before the development of mental decline, the risks associated with the procedure and the need for matched donors to provide competent marrow cells limit its acceptability. Questions still arise of how clinical benefits and an improved quality of life can be best assessed. However, encouraging results showing an improved quality of life, mobility, nutrition, and educational achievements have already been documented in several MPS disorders in response to enzyme therapy—even where pre-existing developmental effects and mental retardation are established. The combined effects of marrow transplantation and enzyme replacement have yet to be systematically evaluated in clinical trials for evidence of therapeutic synergy. At the time of writing, the whole field of the lysosomal diseases such as MPS is attracting much attention as a promising area of clinical and pharmaceutical research. The earlier definition of Orphan diseases 'as those in which treatments offer little or no financial incentive for commercial development' has been abandoned: Orphan diseases are now conveniently defined 'as those which individually affect less than 1 per cent of the population'.

Recently characterized lysosomal diseases

The characterization of lysosomal defects in several ill-understood disorders with diverse clinical manifestations continues to reveal much about the role of the lysosome in cellular functions of significance in medicine and molecular physiology. Several recently studied lysosomal diseases in this category are briefly described here.

Neuronal ceroid lipofuscinoses

The neuronal ceroid lipofuscinoses (**CLN**) represent the most common group of progressive brain diseases that affect children and young adults. Childhood forms of these disorders are inherited as recessive traits and result in a progressive dementia combined with epilepsy, blindness, and an early death. The most familiar form of these diseases was previously known as Batten's disease. Pathological studies of affected patients show the characteristic accumulation of an autofluorescent storage material within neuronal lysosomes and lysosomes in other cells. The storage of this material occurs preferentially in the nervous system and is associated with progressive neuronal death leading to a marked atrophy of the brain; cerebral atrophy is particularly obvious in the early-onset forms of the neuronal lipofuscinoses. Although this complex of neurodegenerative conditions associated with lipofuscin pigments has long been recognized, the diseases were previously thought to represent disorders of other organelles. Mitochondria were implicated because degraded fragments of mitochondrial cytochrome C polypeptide were found to be one of the prominent storage molecules in neuronal tissue from patients with Batten's disease. Latterly, advances in molecular genetics have allowed the identification of defective genes and their protein products in several distinct clinical phenotypes of the neuronal ceroid lipofuscinoses.

Typical clinical forms of the neuronal ceroid lipofuscinoses include late infantile and juvenile neurodegenerative disorders; at least eight genetic loci, which encode proteins implicated in different aspects of lysosomal metabolism, have been assigned to distinct CLN phenotypes. Neuronal ceroid lipofuscinosis type I (CLN 1) is due to mutations in a gene encoding palmitoyl: protein thioesterase 1. CLN 2 is due to defects in the gene encoding tripeptidyl-peptidase, and CLN 8 is due to defects in the gene encoding human cathepsin D. To date, two lysosomal proteins of membrane location, CLN 4 and CLN 5, have also been identified. No specific therapy for Batten's disease yet exists, but the discovery of the basis of the condition and the genes involved allows for prenatal and postnatal diagnosis in affected pedigrees by molecular analysis of the implicated cognate genes. In most instances, neuronal ceroid lipofuscinoses represents defects in elements of intralysosomal protein catabolism, indicating that the turnover of the cognate proteins is very high in cortical neurones. Very recent *in vitro* studies have suggested that the use of the thiol agent, cysteamine, which is used with benefit in patients with cystinosis, may activate residual palmitoyl-protein phytoesterase activity in patients with ceroid lipofuscinosis type 1.

The realization that the neuronal ceroid lipofuscinoses in fact represent intrinsic disorders of lysosomal protein metabolism is very recent. The discovery clearly has important consequences for understanding the pathology of this family of diseases and for developing better diagnostic tools (especially for prenatal application) as well as new treatments.

Papillon–Lefèvre syndrome

This is an unusual syndrome resulting in periodontal disease with tooth loss and palmoplantar keratosis. Papillon–Lefèvre syndrome is associated with a selective deficiency of cathepsin C activity within the specific granules of neutrophilic polymorphonuclear leucocytes. It appears that the enzyme deficiency leads to the failure of bacterial clearance in the gums, thereby causing destructive periodontitis and tooth loss. The corresponding role of cathepsin C within the dermal epithelium is not known, but a failure of cathepsin C activity reproducibly leads to epithelial abnormalities and thickening of the skin, particularly on the soles of the feet. Papillon–Lefèvre syndrome is inherited as an autosomal recessive trait and several mutations have been identified within the gene encoding the cathepsin C polypeptide. Some patients with disabling skin manifestations have obtained benefit with the use of retinoids. These agents are, however, unlikely to improve early-onset destructive periodontal disease.

The importance of the Papillon–Lefèvre syndrome rests not only on the identification of lysosomal cathepsin C as an important component of immune defences against bacteria that specifically invade the privileged periodontal site, but also on the involvement of this enzyme in the normal turnover of keratinized skin. The molecular characterization of this disorder illustrates the protean manifestations of lysosomal defects and of the ubiquitous importance of lysosomes in the destruction and recycling of exogenous microbial, as well as endogenous cellular components.

Defects of organellar assembly: Chediak–Higashi (CHS) and Hermansky–Pudlak (HPS) syndromes

These rare disorders are inherited as autosomal recessive traits. Both cause oculocutaneous albinism in association with abnormal platelet granules and melanosomes in the skin and eyes. CHS predisposes to microbial infection and there are giant lysosomal granules in peripheral blood granulocytes; ceroid storage occurs in the nervous system and lungs. Although very rare, HPS occurs at a high frequency in the Swiss Alps and the Puerto-Rican population where it is the most common single gene defect. HPS causes a mild bleeding diathesis and platelet dense bodies are absent. Granulomatous colitis occurs and pulmonary changes lead to interstitial lung fibrosis; unexplained cardiomyopathy has been reported.

The Hermansky–Pudlak syndrome is caused by mutations in the adaptin, b-3A gene which is associated with altered trafficking of lysosomal proteins in melanosomes, lysosomes, and platelet-dense granules leading to storage pool deficiency. The gene maps to chromosome 10q. Chediak–Higashi syndrome has a clinical phenotype with a complex set of immune defects affecting natural killer cells and neutrophic leucocytes. Recurrent cutaneous and systemic pyogenic infections occur with defective neutrophil and monocyte migration; natural killer-cell cytotoxicity is absent. Neutrophils, melanocytes, neurones, muscle cells, and Schwann cells show giant inclusion bodies. Neurodegeneration is a prominent feature of the disease in young adults, but death often results from a rapidly progressive lymphoproliferative disorder. CHS is caused by mutations in the lysosomal trafficking regulator gene located on chromosome 1q44.

There are clear similarities between HPS and CHS, and further functional studies of their respective cognate proteins should reveal important information about the synthesis and assembly of lysosomes and related organelles.

Defects of lysosomal membrane function

In 1981 two cases of cardiomyopathy in male infants with skeletal myopathy and mental retardation were reported by Danon and colleagues. The skeletal pathology sugguested type II glycogenosis but no deficiency of acid maltase activity was present. Defects in LAMP–2, a major lysosomal membrane protein, have subsequently been identified in Danon's disease due to mutations in the gene encoding LAMP–2, located on the X-chromosome. The disease is responsible for rare forms of cardiomyopathy in adults and may also occur in heterozygous females.

Pycnodysostosis

Pycnodystostosis is a unique disorder of the skeleton caused by an inherited deficiency of another lysosomal-type acid hydrolase, cathepsin K. Cathepsin K is expressed prominently in osteoclasts, in which it is the most abundant secreted protein.

The clinical features of pycnodysostosis includes skull deformity, bone fragility with short stature, and mild osteopetrosis. Delayed fusion of the membrane bones of the skull leads to persistence of the anterior fontanelle; modelling deformities occur with micrognathia, as well as disordered eruption of the primary and secondary teeth. The dwarfed artist, Henri Toulouse-Lautrec—the product of a highly consanguineous marriage—exhibited many features of pycnodysostosis and was disabled by multiple pathological fractures that typically complicate the dense osteopetrotic bones.

A characteristic radiological feature of this disorder is the presence of short, resorbed terminal phalanges with shortened fingers; acro-osteolysis also occurs at the acromial end of the clavicle. Histological examination of affected bone reveals abnormal osteoclasts containing undigested type I collagen fragments within vacuolar spaces. Linkage studies mapped pycnodystostosis to the long arm of chromosome 1 in the same region on 1q21 as cathepsin K: multiple mutations in the cathepsin K gene have now been shown to segregate with the disease in affected pedigrees.

Thus cathepsin K, a cysteine-proteinase expressed in the lysosomes of the pathological macrophages in Gaucher's disease, is particularly abundant in the secretory products of the subosteoclastic resorptive vacuole and is critical for bone modelling and development. Defective digestion of the principal collagen in bone leads to a diffuse skeletal phenotype with clear applications for furthering our knowledge of bone physiology: at least one pharmaceutical company is currently exploring this system as a therapeutic target in osteoporosis.

Further reading

A key reference source for these and other inherited disorders is Victor McKusick's *Mendelian inheritance in man*, published by The Johns Hopkins University Press, Baltimore, MD. This catalogue of human autosomal, X-linked, and mitochondrial phenotypes contains invaluable information that is continually updated in the online version (OMIM) available through the worldwide web at: http://www.ncbi.nlm.nih.gov/htbit-post/omim

Detailed descriptions of all aspects of individual disorders are provided in the acclaimed reference book: *The metabolic and molecular bases of inherited disease*, 8th edition (2001), edited by CR Scriver, AL Beaudet, WS Sly, and D Valle, and published by McGraw-Hill, New York. The lysosomal diseases are principally described in Part 16, pp 3371–894 in the third volume of this large, four-volume work.

Additional references of relevance to particular disorders and new aspects of treatment include:

Amalfitano A, *et al*. (2001). Recombinant human acid alpha-glucosidase enzyme therapy for infantile glycogen storage disease type II: results of a phase I/II clinical trial. *Genetics in Medicine* **3**, 132–8.

Barton NW, *et al.* (1991). Replacement therapy for inherited enzyme deficiency: macrophage-targeted glucocerebrosidase for Gaucher's disease. *New England Journal of Medicine* **324**, 1464–70.

Cox T, *et al.* (2000). Novel oral treatment of Gaucher's disease with *N*-butyl deoxynojirimycin (OGT 918) to decrease substrate biosynthesis. *Lancet* **355**, 1481–5.

Cox TM (2001). Gaucher's disease: understanding the molecular pathogenesis of sphingolipidosis. *Journal of Inherited Metabolic Diseases* **24**(Suppl 2), 106–21.

Frustaci A, *et al*. (2001). Improvement in cardiac function in the cardiac variant of Fabry's disease with galactose-infusion therapy. *New England Journal of Medicine* **345**, 25–32.

Gahl WA, Thoene JG, Scheider JA (2001). Cystinosis: a disorder of lysosomal membrane transport. In: Scriver CR, *et al.*, eds. *Metabolic and molecular bases of inherited disease*, 8th edn, vol III, pp 5085–108. McGraw-Hill, New York.

Gelb BD, *et al.* (1996). Pyknodysostosis, a lysosomal disease caused by cathepsin K deficiency. *Science* **273**, 1236–8.

Kakkis ED, *et al.* (2001). Enzyme replacement therapy in mucopolysaccharidosis I. *New England Journal of Medicine* **344**, 182–8.

MacDermot KD, Holmes A, Miners AH (2001). Anderson–Fabry disease: clinical manifestations and impact of disease in a cohort of 98 hemizygous males. *Journal of Medical Genetics* **38**, 750–60

MacDermot KD, Holmes A, Miners AH (2001). Anderson–Fabry disease: chemical manifestations and impact of disease in a cohort of 60 obligate carrier females. *Journal of Medical Genetics* **38**, 769–807.

Peters C, *et al.* (1998). Hurler's syndrome II. Outcome of HLA-genotypically identical and HLA-haploidentical related donor bone marrow transplantation in fifty-four children. *Blood* **91**, 2601–8.

Schiffman R, *et al.* (2001). Enzyme-replacement therapy in Fabry disease: a randomized control trial. *Journal of the American Medical Association* **285**, 2743–9.

Spritz RA (1999) Multi-organellar disorders of pigmentation: tied up in traffic. *Clinical Genetics* **55**, 309–17.

Toomes C, *et al.* (1999). Loss-of-function mutations in the cathepsin C gene result in periodontal disease and palmoplantar keratosis. *Nature Genetics* **23**, 421–4.

Whitley CB (1993). The mucopolysaccharidoses. In: Beighton P, ed. *McKusick's heritable disorders of connective tissue*, 5th edn, pp 367–499. Mosby Year Book, Inc, St Louis, MO.

Zimran A, ed. (1997). Gaucher's disease. *Clinical haematology*, pp 621–846. Baillière Tindall, London.

11.9 Peroxisomal diseases

Ronald J. A. Wanders and Ruud B. H. Schutgens

Introduction

The peroxisomal disorders are relative newcomers to the field of inherited diseases in humans. The reason for this is that peroxisomes were the last true subcellular organelles to be discovered and were long thought to play only a minor role in cellular metabolism. Two key observations in patients suffering from a rare disease, the cerebrohepatorenal syndrome of Zellweger, changed this view completely.

A major hallmark in studies of Zellweger syndrome was the finding by Goldfischer and colleagues in 1973 that peroxisomes were completely absent in the hepatocytes and renal tubule cells of Zellweger patients. In 1982 the accumulation of certain saturated very long-chain fatty acids, notably hexacosanoic (cerotic) acid (C26:0), was reported in plasma from Zellweger patients whereas other fatty acids like palmitate, oleate, and linoleate were normal. At the same time our group reported that plasmalogens, a specific type of phospholipid, were deficient in tissues and erythrocytes from Zellweger patients. Subsequent studies soon resolved that these abnormalities were a direct consequence of the absence of peroxisomes in these patients.

Since then much has happened and it is now clear that Zellweger syndrome is the prototype of an expanding group of genetic diseases all caused by an impairment in one or more peroxisomal functions.

To provide the necessary background we will briefly describe the main functions of peroxisomes in humans as well as their biogenesis.

Functions of peroxisomes

Peroxisomes play an essential role in a range of cellular activities, which mainly have to do with lipid metabolism (see Table 1). The following functions are essential and are directly linked to certain peroxisomal disorders.

b-Oxidation of fatty acids and fatty acid derivatives

Just like mitochondria, peroxisomes are capable of oxidizing fatty acids by both a- and b-oxidation. The mechanism whereby fatty acids are b-oxidized in peroxisomes is the same as in mitochondria, involving four sequential steps of oxidation, hydration, dehydrogenation, and thiolytic cleavage.

It is well established now that there are major differences between the mitochondrial and peroxisomal b-oxidation systems, each catalysing the oxidation of a distinct set of substrates. Indeed, mitochondria take care of the oxidation of the vast majority of dietary fatty acids including oleate, palmitate, and linoleate whereas peroxisomes catalyse the b-oxidation of a range of fatty acids which are not so important for energy purposes but which do require breakdown.

These include (see Fig. 1):

1. Very long-chain fatty acids: fatty acids like tetracosanoic (lignoceric) acid (C24:0) and hexacosanoic (cerotic) acid (C26:0) cannot be b-oxidized by mitochondria but are good substrates for peroxisomal b-oxidation. Peroxisomes are not capable of fully degrading C26:0 and C24:0 to acetyl coenzyme A (**CoA**) units. Instead only a few cycles occur in peroxisomes after which the chain-shortened acyl CoA units move to the mitochondria for further oxidation.
2. Pristanic acid (2,6,10,14-tetramethylpentadecanoic acid): this branched-chain fatty acid is in part derived directly from dietary sources but is also formed from phytanic acid by a process called a-oxidation (see later). Pristanic acid undergoes three cycles of b-oxidation in the peroxisome to produce 4,8-dimethylnonanoyl-CoA which is then transported to the mitochondrion for full oxidation to CO_2 and H_2O (Fig. 1).
3. Di- and trihydroxycholestanoic acid: the CoA esters of di- and trihydroxycholestanoic acid undergo one cycle of b-oxidation in the peroxisome, giving rise to the CoA esters of chenodeoxycholic acid and cholic acid respectively, which are then converted into the corresponding taurine or glycine conjugates (tauro/glycochenodeoxycholate and tauro/glycocholate respectively). This is followed by transport into bile. This implies that peroxisomes play an indispensable role in bile acid formation.

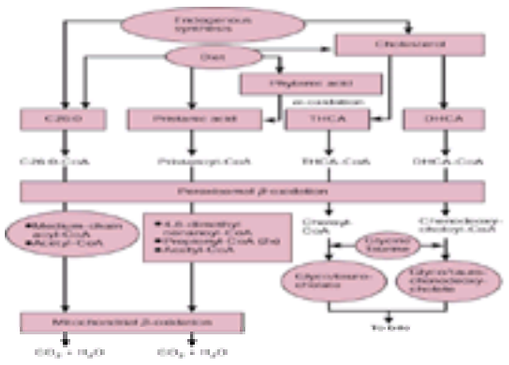

Fig. 1 Involvement of peroxisomes and mitochondria in the b-oxidation of hexacosanoic acid (C26:0), pristanic acid, and di- and trihydroxycholestanoic acid. See text for details.

It is now clear that peroxisomes contain multiple enzymes involved in b-oxidation. These include (see Fig. 2):

1. Two acyl-CoA oxidases: both these oxidases (acyl-CoA oxidase 1 and 2) react with a variety of straight-chain acyl-CoAs whereas only acyl-CoA oxidase 2 is reactive with branched-chain acyl-CoAs. This implies that acyl-CoA oxidase 2 is involved in the oxidation of pristanoyl CoA as well as dihydroxycholestanoic acid CoA and trihydroxycholestanoic acid CoA whereas acyl-CoA oxidase 1 is not. Acyl-CoA oxidase 1 is probably the major enzyme involved in C26:0 b-oxidation.
2. Two bifunctional proteins with both enoyl-CoA hydratase and 3-hydroxyacyl-CoA dehydrogenase activity: a major difference between the two proteins is that the first bifunctional protein forms and dehydrogenates L-3-hydroxyacyl-CoA in contrast to the second enzyme, which forms and dehydrogenates D-3-hydroxyacyl-CoA. Hence the names L-bifunctional protein and D-bifunctional protein.
3. Two peroxisomal thiolases: recent studies have led to the identification of a new thiolase called SCP_x or peroxisomal thiolase 2 which is the main if not exclusive enzyme involved in b-oxidation of pristanic acid as well as dihydroxycholestanoic acid and trihydroxycholestanoic acid, whereas the original thiolase (peroxisomal thiolase 1) is the main enzyme in C26:0 b-oxidation (see Fig. 2).

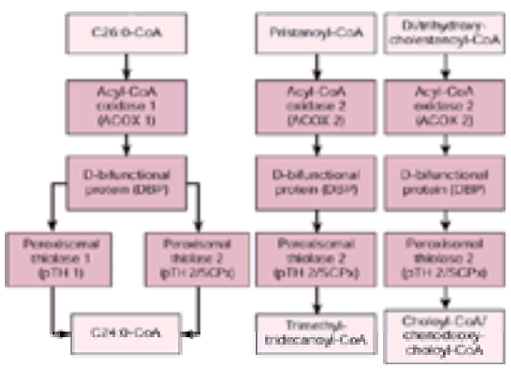

Fig. 2 Enzymology of the peroxisomal fatty acid b-oxidation machinery. See text for details.

Ether-phospholipid biosynthesis

A second major function of peroxisomes concerns their role in the synthesis of ether-phospholipids. Indeed, the two enzyme activities responsible for the introduction of the characteristic ether linkage in ether-linked phospholipids (i.e. dihydroxyacetonephosphate acyltransferase (**DHAPAT**) and alkyldihydroxyacetonephosphate synthase (**alkyl-DHAP synthase**)), are both localized in peroxisomes. The next enzyme, acyl/alkyl-DHAP:NAD(P) oxidoreductase, is localized in both peroxisomes and the endoplasmic reticulum so that the product of the alkyl-DHAP synthase reaction, (i.e. alkyl-DHAP), is converted into alkylglycerol-3-phosphate in the peroxisome or endoplasmic reticulum (Fig. 3). All subsequent reactions involved in ether-phospholipid synthesis take place in the endoplasmic reticulum.

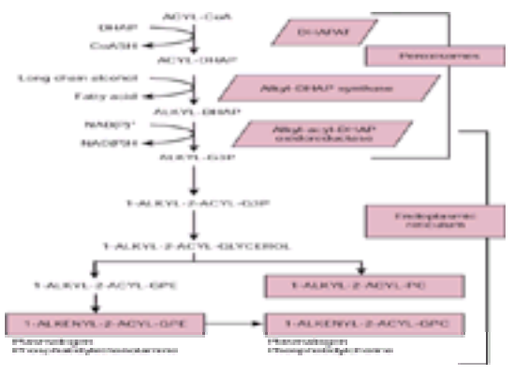

Fig. 3 Enzymology of ether–phospholipid biosynthesis. See text for details.

The functional role of plasmalogens has remained enigmatic until now. However, the identification of an isolated deficiency of either DHAPAT or alkyl-DHAP synthase in patients with severe clinical abnormalities comparable to rhizomelic chondrodysplasia punctata clearly shows that ether-phospholipids are of central importance in humans.

Phytanic acid a-oxidation

The pathway of phytanic acid a-oxidation has long remained an enigma but has recently been resolved. Phytanic acid first undergoes activation to phytanoyl-CoA after which a hydroxylation reaction takes place to generate 2-hydroxyphytanoyl-CoA as catalysed by the enzyme phytanoyl-CoA hydroxylase, which is deficient in Refsum disease (discussed later). Subsequently 2-hydroxyphytanoyl-CoA undergoes cleavage to produce formyl-CoA and pristanal, which is then dehydrogenated to pristanic acid. Activation of pristanic acid produces pristanoyl CoA, which can be degraded via b-oxidation in the peroxisome.

L-pipecolic acid oxidation

L-lysine is normally degraded by the saccharopine pathway involving the sequential action of L-lysine:2-oxoglutarate reductase and saccharopine dehydrogenase. However, L-lysine may also be degraded via the L-pipecolic acid pathway, which may especially be important in the brain. In the L-pipecolic acid pathway, L-pipecolic acid is produced from L-lysine via two enzymatic steps. L-pipecolic acid is then oxidized by L-pipecolate oxidase, a peroxisomal enzyme, at least in humans.

The function of the L-pipecolic acid pathway remains incompletely understood.

Biosynthesis of polyunsaturated fatty acids such as docosahexaenoic acid

In tissues and erythrocytes from patients with Zellweger syndrome there is a profound deficiency of docosahexaenoic acid (C22:6w3) suggesting the involvement of peroxisomes in the formation of docosahexaenoic acid. Subsequent studies have not only established the role of peroxisomes in docosahexaenoic acid formation but have also revealed that the last step involved in formation of docosahexaenoic acid, the desaturation of clupanodonic acid (C22:5w3) to docosahexaenoic acid, is not catalysed by a presumed ‡4-desaturase but involves a three-step pathway. In this pathway C22:5w3 is first elongated to C24:5w3, followed by ‡6-desaturation to C24:6w3. The latter compound is then chain-shortened to C22:6w3 (docosahexaenoic acid) via b-oxidation in peroxisomes.

This latter finding explains the deficiency of docosahexaenoic acid in patients with Zellweger syndrome.

Biogenesis of peroxisomes

In recent years much has been learned about peroxisome biogenesis, and many of the genes involved in peroxisome biogenesis, called *PEX* genes, are known. Although there is still discussion about the possible involvement of the endoplasmic reticulum, the model proposed by Lazarow and Fujiki is still generally accepted. The principal features of this model are:

1. Peroxisomal membrane and matrix proteins are synthesized on free polyribosomes.
2. The newly synthesized proteins are post-translationally imported from the cytosol into pre-existing peroxisomes.
3. Import of new polypeptides expands the peroxisomal compartment, making them grow until they reach a critical size which results either in division of peroxisomes into two daughter peroxisomes or in budding from the peroxisomal reticulum followed by subsequent growth.

The fact that proteins destined for peroxisomes are synthesized on free polyribosomes implies that these proteins must have specific signals which direct them to peroxisomes.

Studies, notably by Subramani and colleagues, have shown the existence of two such peroxisome targeting signals (**PTS**). Most of the peroxisomal matrix proteins are equipped with a PTS1 signal which involves a C-terminal tripeptide of the sequence serine–lysine–leucine or a conserved variant thereof. The second peroxisome targeting signal (PTS2) has been found in far fewer peroxisomal proteins and involves a stretch of nine amino acids of which amino acids numbers one, two, eight, and nine are essential with the following consensus: (arginine/lysine)–(leucine/valine/isoleucine)–XXXXX–(histidine/glutamine)–(leucine/alanine) in which X may be any amino acid. In mammals, the PTS2 signal has only been identified in peroxisomal thiolase 1, alkyl-DHAP synthase, and phytanoyl-CoA hydroxylase. Peroxisomal membrane proteins lack either a PTS1 a PTS2 signal which implies that there must be separate signals directing membrane proteins to peroxisomes.

The identification of the PTS1 and PTS2 targeting signals was soon followed by discovery of a growing number of so-called *PEX* genes required for peroxisome biogenesis. Most of these genes were first identified in the yeast *Saccharomyces cerevisiae*. Making use of databases of expressed sequence the human counterparts of these yeast genes have been identified in recent years, notably by Gould and colleagues. Most of these *PEX* genes code for peroxisomal membrane proteins with the exception of PEX5p and PEX7p. These two proteins are predominantly cytosolic and recognize the PTS1 and PTS2 signals respectively on peroxisomal proteins in the cytosol. In humans at least, the two loaded receptors functionally interact to form a complex which subsequently docks at the peroxisomal membrane. The proteins PEX13p and PEX14p play a central role in this docking process.

The following step involves dissociation of the complex followed by transport of the PTS1 and PTS2 proteins across the membrane and recycling of the receptors back to the cytosol for another round (Fig. 4). In this process multiple PEX proteins are involved, all of which essential since disruption of each single *PEX* gene is associated with a full block in peroxisome biogenesis.

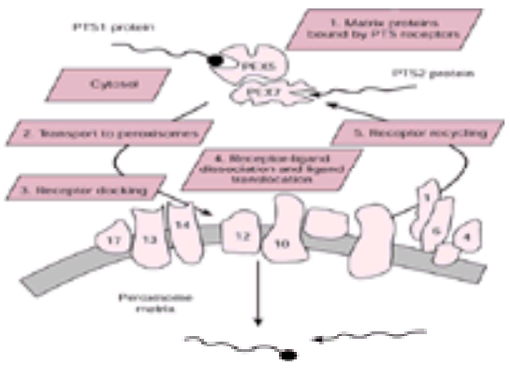

Fig. 4 Overview of the steps involved in peroxisome biogenesis including: (1) binding of ligands in the cytosol by PEX5p and PEX7p, the PTS1 and PTS2 receptors; (2) transport to and docking of PEX5p–ligand and PEX7p–ligand complexes with the peroxisomal membrane; (3) translocation of ligands into the peroxisome; and (4) recycling of the receptors back to the cytoplasm.

Figure 4 shows a current model for peroxisome biogenesis in humans. Our increasing knowledge about the human genes involved in peroxisome biogenesis has been of tremendous importance for studies on the genetic basis of Zellweger syndrome and the other disorders of peroxisome biogenesis. Indeed, at present, 11 of the 12 *PEX* genes which underlie the 12 complementation groups which have been identified so far are now known.

The peroxisomal disorders

Table 2 lists the peroxisomal disorders identified so far. Throughout the years several classifications have been proposed for the different peroxisomal disorders. The first proposed classifying the peroxisomal disorders into three groups reflecting the extent of peroxisomal dysfunction with a generalized (group A), multiple (group B), and single (group C) loss of peroxisomal function. A second classification divided the peroxisomal disorders into two groups with the disorders of peroxisome biogenesis in group 1 and the peroxisomal disorders in which a single peroxisomal function has been lost in group 2.

Below we will describe the characteristics of the individual peroxisomal disorders followed by the laboratory diagnosis of these disorders making use of a new classification based on the concept of diagnostic groups.

Cerebrohepatorenal (Zellweger) syndrome

The clinical presentation of Zellweger syndrome is dominated by craniofacial dysmorphia and profound neurological abnormalities. The craniofacial dysmorphia includes a high forehead, flat occiput, wide open sutures, large fontanelle, hypoplastic orbital ridges, epicanthus, high arched palate, external ear deformities, micrognathia, and redundant skin folds of the neck. Neurological abnormalities include severe hypo/areflexia, poor sucking, epileptic seizures, severe neonatal hypotonia, nystagmus, and sensorineural hearing loss. Furthermore, there is profound psychomotor retardation. There are also ocular abnormalities including cataracts, glaucoma, corneal clouding, Brushfield spots, pigmentary retinopathy, and optic nerve dysplasia. Because of the hypotonia and mongoloid appearance, Zellweger patients are sometimes wrongly suspected of suffering from Down's syndrome.

Pathological studies in Zellweger patients have shown a great number of abnormalities in various organs including the brain. Macroscopic abnormalities include deviant sulci and gyri with almost vertical parietal clefts and pachymicrogyria. The most striking and intriguing neuropathological abnormality is the impaired neuronal migration which leads to characteristic and unique cytoarchitectural abnormalities involving the cerebral hemispheres, the cerebellum, and the inferior olivary

complex.

Patients with Zellweger syndrome usually die early in life with an average lifespan of a few months.

Neonatal adrenoleucodystrophy

The first case of neonatal adrenoleucodystrophy was described in 1978 in a boy with hypotonia, convulsions, absent grasp reflexes, slight Moro response, and little spontaneous movement at birth. Characteristic signs of adrenoleucodystrophy were found in this patient which included demyelination of the central nervous system white matter, atrophy of the adrenal cortex, ballooned adrenocortical cells, and the presence of lamellar inclusions of electrondense leaflets separated by a clear space. However, there were also a number of additional central nervous abnormalities not described for X-linked adrenoleucodystrophy and it was noted that this new type of leucodystrophy resembled Zellweger syndrome in several respects, suggesting a common aetiology. This was soon found to be true when multiple peroxisomal abnormalities resulting from a defect in peroxisome biogenesis were described in patients with neonatal adrenoleucodystrophy.

Infantile Refsum disease

Infantile phytanic acid storage disease was first described in 1982 in three unrelated patients showing hepatomegaly, mental retardation, (minor) facial dysmorphia, retinopathy, neurosensory deafness, osteopenia, growth retardation, and elevated plasma phytanic acid levels. Because phytanic acid was known to accumulate in Refsum disease, the name infantile Refsum disease was coined. Since this initial report many additional patients have been described.

As with neonatal adrenoleucodystrophy it was soon realized that there were certain similarities between infantile Refsum disease and Zellweger syndrome, and we now know that infantile Refsum disease is also a disorder of peroxisome biogenesis which explains the multiple peroxisomal abnormalities observed in patients with infantile Refsum disease.

Hyperpipecolic acidaemia

Several cases have been reported in the literature of hyperpipecolic acidaemia. It is now clear that not all of these patients are true cases of isolated hyperpipecolic acidaemia. This implies that hyperpipecolic acidaemia has become obsolete in the classification of peroxisomal disorders. On the other hand, recent reports document the existence of apparently true cases of hyperpipecolic acidaemia although the underlying defect remains to be established.

Biochemistry of Zellweger syndrome, neonatal adrenoleucodystrophy, and infantile Refsum disease

In Zellweger syndrome, neonatal adrenoleucodystrophy, and infantile Refsum disease peroxisome biogenesis is defective resulting in the virtual absence of morphologically distinguishable peroxisomes in all the patient's cells. This has been demonstrated most convincingly in liver biopsy specimens from patients, notably by Roels and colleagues. The deficiency of peroxisomes can also be demonstrated in cultured fibroblasts from patients using immunofluorescence microscopy.

The absence of peroxisomes in patients with Zellweger syndrome, neonatal adrenoleucodystrophy, and infantile Refsum disease explains the multiplicity of biochemical abnormalities observed in these patients which are summarized in Table 3.

Molecular basis of the peroxisome deficiency disorders

The genetic basis of the different disorders of peroxisome biogenesis has been elucidated. Earlier studies had already shown that the genetic basis of Zellweger syndrome, neonatal adrenoleucodystrophy, and infantile Refsum disease is very heterogeneous. This was concluded from complementation studies, which involved fusion of fibroblasts of two different patients with the same abnormality such as a defect in peroxisome biogenesis so that hybrid cells (heterokaryons) are generated containing nuclei from each patient's fibroblasts. If the defective genes in the two patients' cell lines were different, one would expect restoration of peroxisome formation, whereas in the case when the mutant genes are identical, no complementation would occur. This is most easily done using catalase immunofluorescence. Large-scale complementation studies have now shown that there are at least 12 different complementation groups representing 12 different *PEX* genes. Interestingly, there is strong over-representation of one particular group, occurring in 60 to 70 per cent of all patients with a disorder of peroxisome biogenesis. The gene defective in this group is the *PEX1* gene which codes for a protein (PEX1p) belonging to the family of ATPases associated with diverse cellular activities which obviously plays an indispensable role in peroxisome biogenesis, although its precise mode of action is unknown. Eleven of the 12 genes which underlie the different complementation groups have now been identified. We have applied complementation analysis on a systematic basis to more than 200 patients with disorders of peroxisome biogenesis: most patients belong to the PEX1 group (63 per cent), followed by PEX6 (10 per cent), and PEX12 (5.4 per cent).

The identification of the mutant *PEX* genes is important for carrier detection and prenatal diagnosis in families at risk. On the other hand current prenatal diagnostic procedures which involve analyses at the enzyme (activity measurements) and protein (immunoblotting) level are very reliable.

Rhizomelic chondrodysplasia punctata type 1

Chondrodysplasia punctata represents a genetically heterogeneous group of bone dysplasias with stippling of the epiphyses as a common feature. The rhizomelic form is characterized by the presence of stippled foci of calcification within the hyaline cartilage with coronal clefts in the vertebral bodies associated with dwarfing, cataracts, multiple malformations due to contractures, and mental retardation in virtually all patients. Ichthyosis is frequent. Inheritance is autosomal recessive. Rhizomelic chondrodysplasia punctata must be distinguished from the milder autosomal dominant form of chondrodysplasia punctata (Conradi–Hunermann syndrome) with longer survival, absence of severe limb shortening, and usually intact intellect.

Apart from the classic presentation of rhizomelic chondrodysplasia punctata, a number of patients have now been described in the literature with a much milder presentation lacking the characteristic stigmata of rhizomelic chondrodysplasia punctata. Indeed, one such patient is still alive now at 16 years of age, has no rhizomelia, and only mild mental retardation. As we will describe below, there is not only clinical variability within rhizomelic chondrodysplasia punctata but also genetic diversity with two additional genetic types of rhizomelic chondrodysplasia punctata named types 2 and 3.

Biochemistry of rhizomelic chondrodysplasia punctata type 1 and molecular basis

Rhizomelic chondrodysplasia punctata is a true peroxisomal disorder, with the original description being of deficiency of plasmalogens in tissues and erythrocytes from patients and elevated phytanic acid levels in plasma. Subsequent studies have led to the remarkable finding of four distinct peroxisomal enzyme deficiencies at the level of:

1. dihydroxyacetonephosphate acyltransferase (DHAPAT)
2. alkyldihydroxyacetonephosphate synthase (alkyl-DHAP synthase)
3. phytanoyl-CoA hydroxylase, and
4. 41 kDa peroxisomal thiolase.

The underlying basis for this peculiar observation came when it was discovered that alkyl-DHAP synthase, phytanoyl-CoA hydroxylase, and peroxisomal thiolase all turned out to be PTS2 proteins, suggesting that the defect in rhizomelic chondrodysplasia punctata had to be in the PTS2 receptor. The gene coding for the PTS2 receptor was identified in 1997 and mutations were found in patients with rhizomelic chondrodysplasia punctata, thus establishing its molecular basis. Although many different mutations have been identified there is one frequent mutation, the Leu292Stop mutation, which occurs in more than 50 per cent of the mutant alleles.

Rhizomelic chondrodysplasia punctata type 2 (DHAPAT deficiency)

In 1992 the first patient was described with classic rhizomelic chondrodysplasia punctata but lacking the characteristic set of four peroxisomal abnormalities. Instead an isolated deficiency of only a single enzyme (DHAPAT) was found (Table 4). This enzyme catalyses the first committed step in plasmalogen synthesis and its deficiency is associated with the inability to synthesize plasmalogens.

At present several cases of rhizomelic chondrodysplasia punctata type 2 have been described. Most cases had a severe clinical presentation resembling classic rhizomelic chondrodysplasia punctata type 1. As could be expected, there is also clinical heterogeneity within rhizomelic chondrodysplasia punctata type 2. In all patients identified so far, the deficiency of DHAPAT is the single abnormality due to mutations in the structural gene coding for DHAPAT.

Rhizomelic chondrodysplasia punctata type 3 (alkyl-DHAP synthase deficiency)

The first case of rhizomelic chondrodysplasia punctata type 3 in a patient showing all the clinical stigmata of classic rhizomelic chondrodysplasia punctata was identified in 1994. In this patient plasmalogen biosynthesis was found to be blocked due to the deficient activity of the second enzyme involved in plasmalogen biosynthesis, i.e. alkyl-DHAP synthase. The least affected of those patients so far identified is alive at 7½ years with mild to moderate rhizomelia, generalized flexion contractures, inability to sit, roll, or crawl, cataract, continued seizures, and profound developmental delay. The deficient activity of alkyl-DHAP synthase in these patients has been found to result from different mutations in the structural gene coding for alkyl-DHAP synthase.

X-linked adrenoleucodystrophy

There is considerable clinical heterogeneity within X-linked adrenoleucodystrophy with at least six phenotypic variants. The classification of the different phenotypes of X-linked adrenoleucodystrophy is somewhat arbitrary and is based upon the age of onset and the organs principally affected. It should be noted that there are also patients not easily assignable to either phenotype, emphasizing the variable clinical expression of X-linked adrenoleucodystrophy. The two most frequently observed phenotypes, accounting for approximately 80 per cent of all cases, are childhood adrenoleucodystrophy and adrenomyeloneuropathy.

Childhood cerebral adrenoleucodystrophy

Childhood cerebral adrenoleucodystrophy is characterized by rapidly progressive cerebral demyelination. The onset is between 3 and 10 years of age. Frequent early neurological symptoms are behavioural disturbances, a decline in school performance, deterioration of vision, and impaired auditory discrimination. The course is relentlessly progressive, and seizures, spastic tetraplegia, and dementia develop within months. Most patients die within 2 to 3 years of the onset of neurological symptoms. Some patients survive longer, albeit in a persistent vegetative state.

In most patients with childhood cerebral adrenoleucodystrophy, cerebral magnetic resonance imaging typically reveals extensive demyelination in the occipital periventricular white matter, with sparing of the U fibres. Much less frequently, the frontal lobes are affected first. The early symptoms of childhood cerebral adrenoleucodystrophy are frequently attributed to an attention deficit disorder or hyperactivity. Before the advent of reliable diagnostic tests and magnetic resonance imaging, metachromatic leucodystrophy, ceroid lipofuscinosis, globoid cell leucodystrophy (Krabbe disease), and subacute sclerosing panencephalitis were sometimes diagnosed instead of childhood cerebral adrenoleucodystrophy.

Adrenomyeloneuropathy

Adrenomyeloneuropathy is the most frequent phenotype of X-linked adrenoleucodystrophy. The onset of neurological symptoms in this phenotype usually occurs in the third and fourth decade. Neurological deficits are primarily due to myelopathy and to a lesser extent to neuropathy. Patients gradually develop a spastic paraparesis, often in combination with a disturbed vibration sense in the legs and sphincter dysfunction. Approximately 50 per cent of men with adrenomyeloneuropathy show mild to moderate cerebral involvement on magnetic resonance imaging, and in some the abnormalities of white matter may resemble the demyelination seen in childhood cerebral adrenoleucodystrophy; the spinal cord is frequently atrophic. Nerve conduction studies and electromyography are compatible with a predominantly axonal, sensorimotor polyneuropathy. Life expectancy is probably normal, unless patients develop cerebral demyelination, or when adrenocortical insufficiency is not recognized and remains untreated. Many patients with adrenomyeloneuropathy were initially diagnosed with neurological diseases such as chronic progressive multiple sclerosis and hereditary spastic paraparesis.

Biochemistry and molecular basis of X-linked adrenoleucodystrophy

The biochemical hallmark of X-linked adrenoleucodystrophy is the accumulation of very long-chain fatty acids in plasma, fibroblasts, and other cell types. Analysis of plasma very long-chain fatty acids has turned out to be an extremely powerful diagnostic method with only a few if any false negatives, at least when analysed in experienced laboratories.

The accumulation of very long-chain fatty acids, notably C26:0, is due to their impaired oxidation in peroxisomes. It was initially thought that this was due to the deficient activity of a peroxisomal enzyme catalysing the activation of very long-chain fatty acids to their CoA esters which is the first obligatory step in fatty acid b-oxidation. The *X-ALD* gene was identified using a positional cloning strategy. Remarkably the deduced ALD protein did not appear to be an acyl-CoA synthetase but instead turned out to belong to the family of ABC proteins which also includes the CFTR protein involved in cystic fibrosis and the MDR protein involved in multidrug resistance. The ALD protein is a so-called halftransporter with six transmembrane spanning elements, and probably functions as a transporter of very long-chain fatty acid CoA esters across the peroxisomal membrane either as homo- or heterodimers localized in the peroxisomal membrane. The identification of the *X-ALD* gene has allowed molecular studies in X-linked adrenoleucodystrophy patients, which have shown diverse often unique mutations.

Acyl CoA oxidase deficiency (pseudoneonatal adrenoleucodystrophy)

In 1988 Poll-The and colleagues described patients with neonatal onset hypotonia together with delayed motor development, sensory deafness, and retinopathy with extinguished electroretinograms. There was no craniofacial dysmorphia in any of the patients. Psychomotor development was severely delayed, and after the first 2 years of life a progressive neurological regression set in. Computed tomography with contrast enhancement revealed bilateral enhancing lesions and a generally hypodense cerebral white matter which led to the diagnosis of neonatal adrenoleucodystrophy. The finding of elevated levels of very long-chain fatty acids in the plasma of these patients supported this conclusion. In contrast to the findings in other patients with neonatal adrenoleucodystrophy, however, these patients had peroxisomes normally present, albeit of enlarged size, and biochemical abnormalities were restricted to the accumulation of very long-chain fatty acids. The defect in these patients was at the level of acyl-CoA oxidase, the first obligatory enzyme involved in the peroxisomal b-oxidation of very long-chain fatty acids (see Fig. 2).

All patients identified so far have severe neurological abnormalities, mild to absent craniofacial dysmorphia, and failure to thrive. In patients with acyl-CoA oxidase deficiency, plasma abnormalities are restricted to the accumulation of very long-chain fatty acids which follows logically from the role of acyl-CoA oxidase in very long-chain fatty acid b-oxidation (see Fig. 2). A large deletion in the acyl-CoA oxidase gene in the two original patients has been reported.

D-Bifunctional protein deficiency

Bifunctional protein deficiency was first described in 1989 in a patient with severe hypotonia and seizures of neonatal onset. An electroencephalogram revealed multifocal spikes. No developmental progress was observed. There was no dysmorphia and no hepatomegaly. Fontanelles were open with open metopic sutures. Visual evoked responses and brainstem auditory evoked responses were grossly abnormal. A brain biopsy at 6 weeks revealed polymicrogyria. The patient died at 5 months of age after a clinical course marked by absent developmental progress and seizures refractory to treatment. Neuropathological studies revealed a polymicrogyric neocortex and focal areas of cortical heterotopia.

Laboratory analysis revealed elevated very long-chain fatty acids but liver morphology revealed the normal presence of peroxisomes. Subsequent immunoblot experiments revealed the absence of one of the peroxisomal b-oxidation enzyme proteins, i.e. bifunctional protein.

Many additional patients with bifunctional protein deficiency have since been described. Identification of such cases usually starts in a patient with a Zellweger-like phenotype in which subsequent laboratory studies reveal an isolated defect in peroxisomal b-oxidation with no abnormalities in other peroxisomal functions like plasmalogen biosyntheses and the normal presence of peroxisomes, although they are usually enlarged in both liver and in fibroblasts. Such data clearly point to a disorder of peroxisomal b-oxidation. Complementation studies have led to the identification of four groups with group 1 representing acyl-CoA oxidase deficiency, group 2 bifunctional protein deficiency, and groups 3 and 4 involving unknown defects.

Most patients turned out to belong to group 2. Remarkably, molecular studies in fibroblasts from patients belonging to group 2 failed to identify any mutations in the

gene for L-bifunctional protein. This puzzling situation was resolved when a new bifunctional protein called D-bifunctional protein or multifunctional enzyme 2 was discovered which turned out to be the enzyme deficient in group 2. Molecular studies clearly identified mutations in the gene for D-bifunctional protein, thus resolving the true molecular basis of bifunctional protein deficiency. Recent studies have shown that there are three subgroups within the bifunctional protein deficiency group, which finds its basis in the fact that bifunctional protein is a single protein with two catalytic activities.

The clinical presentation of patients affected by D-bifunctional protein deficiency is usually very severe and resembles that of Zellweger syndrome in many respects. Indeed, in a recent series neonatal hypotonia (94 per cent), dysmorphic features (80 per cent), neonatal seizures (92 per cent), hepatomegaly (43 per cent), developmental delay (100 per cent), poor feeding (86 per cent), and a disordered neuronal migration (88 per cent) were found.

Biochemistry of D-bifunctional protein deficiency

In patients suffering from D-bifunctional protein deficiency there is accumulation of very long-chain fatty acids, pristanic acid, and di- and trihydroxycholestanoic acid which follows logically from the scheme of Fig. 2. This is true for most cases, but in one form of D-bifunctional protein deficiency (type B) in which the enoyl-CoA hydratase component of D-bifunctional protein is functionally inactive there is no accumulation of bile acid intermediates although the underlying basis is unclear.

Peroxisomal thiolase-1 deficiency (pseudo-Zellweger syndrome)

In 1986 a girl from consanguineous parents showing marked facial dysmorphia, muscle weakness, and hypotonia at birth was described. The patient showed no psychomotor development during her 11-month life. At autopsy the patient had renal cysts, atrophic adrenals with striated cells, minimal liver fibrosis, hypomyelination in the cerebral white matter, foci of neuronal heterotopia, and a sudanophilic leucodystophy. Taken together, these clinical findings suggested that the patient was affected by Zellweger syndrome. Morphological analysis of the liver, however, revealed abundant peroxisomes. The subsequent finding that very long-chain fatty acids in plasma were clearly elevated did suggest a peroxisomal defect possibly restricted to the peroxisomal b-oxidation system. Further proof came when elevated levels of di- and trihydroxycholestanoic acid were found in a duodenal aspirate. Immunoblot studies revealed that both the 44 kDa precursor form of peroxisomal thiolase-1 as well as the mature 41 kDa form were completely missing, indicating peroxisomal thiolase deficiency. No additional patients have been described in the literature.

Peroxisomal 2-methylacyl-CoA racemase deficiency associated with a late onset motor neuropathy: a newly identified peroxisomal disorder

A new defect in the peroxisomal fatty acid b-oxidation pathway in a number of patients suffering from an adult onset sensory motor neuropathy has been described. Sensory motor neuropathy is associated with inherited disorders including Charcot–Marie–Tooth disease, X-linked adrenoleucodystrophy/adrenomyeloneuropathy, and Refsum disease. In the latter two, the neuropathy is thought to result from the accumulation of very long-chain fatty acids and phytanic acid respectively. The plasma of two patients with adult onset sensory motor neuropathy and additional signs suggesting Refsum disease (patient 1) and X-linked adrenoleucodystrophy (patient 2) had a similar profile: normal very long-chain fatty acids, marginally elevated phytanic acid, and definitely increased levels of the 2-methyl branched-chain fatty acids pristanic acid and di- and trihydroxycholestanoic acid. This suggested a specific defect in the peroxisomal b-oxidation of branched-chain fatty acids and not in the a-oxidation system, the first enzyme step of which is defective in Refsum disease. Studies in fibroblasts revealed normal values for all parameters measured except for pristanic acid b-oxidation, which was reduced to 20 to 30 per cent of control. The activities of the enzymes directly involved in the b-oxidation of branched-chain fatty acids, which includes branched-chain acyl-CoA oxidase, D-bifunctional protein, and peroxisomal thiolase 2 (pTH2/SCPx), were all normal.

Attention was focused on 2-methylacyl-CoA racemase. As described before, this enzyme is not directly involved in the b-oxidation process itself but is important in the b-oxidation of both pristanic acid and di- and trihydroxycholestanoic acid (Fig. 2) since the enzyme catalyses the interconversion of (2*R*) and (2*S*) stereoisomers of 2-methyl branched-chain fatty acyl-CoA esters. Measurement of racemase activity in fibroblasts using (25 *S*)-2,5,5-trihydroxycholestanoyl-CoA as substrate revealed a complete deficiency of the enzyme.

The finding of defined abnormalities in patients with late onset motor neuropathy resulting from a defect in the peroxisomal oxidation of 2-methyl branched-chain fatty acids at the level of 2-methylacyl-CoA racemase may have implications for the diagnosis of adult onset neuropathies of unknown aetiology.

Hyperoxaluria type 1 (alanine glyoxylate aminotransferase deficiency)

See Chapter 11.10 for further discussion.

Refsum disease

Refsum disease was first delineated as a distinct disease entity on a clinical basis by Sigvald Refsum in the 1940s under the name heredopathia atactica polyneuritiformis. Cardinal manifestations of the disease include retinitis pigmentosa, cerebellar ataxia, chronic polyneuropathy, and an elevated protein level in the cerebrospinal fluid with a normal cell count. Less constant features include sensorineural hearing loss, anosmia, ichthyosis, skeletal malformations, and cardiac abnormalities. The clinical picture of Refsum disease is often that of a slowly developing, progressive peripheral neuropathy manifested by severe motor weakness and muscular wasting, especially of the lower extremities.

Patients in whom Refsum disease is destined to develop appear to be perfectly normal as infants and do not show any obvious defects in growth and development. Onset has occasionally been detected in early childhood but not until the fifth decade in others. Most patients have clear-cut manifestations before the age of 20.

Biochemistry of Refsum disease

Studies in the early 1960s led to the identification of phytanic acid in tissues and plasma samples of Refsum patients. The enzyme defect in Refsum disease was identified at the level of phytanoyl-CoA hydroxylase.

Molecular basis of Refsum disease

The identification of the phytanoyl-CoA hydroxylase complementary DNA and gene structure has allowed molecular studies which show a variety of often unique mutations.

Glutaric aciduria type 3

In 1991 Bennett and colleagues described a patient who was investigated at 11 months of age because of failure to thrive and postprandial vomiting. Two abnormalities were found. First, she was shown to be homozygous for b-thalassemia and second, significant glutaric aciduria was found. Studies in fibroblasts revealed normal glutaryl-CoA dehydrogenase activity whereas glutaryl-CoA oxidase activity was not detectable. This study has not been followed up at the enzyme protein and/or DNA level. No additional cases have been identified so far.

Mevalonate kinase deficiency (see Chapter 11.12.3)

Mevalonate kinase deficiency is the only disorder involving the peroxisomal part of isoprenoid biosynthesis. Interestingly, apart from the classical form of mevalonate kinase deficiency which is associated with severe abnormalities early in life including profound developmental delay, facial dysmorphia, cataract, hepatosplenomegaly, lymphadenopathy, and early death, mevalonate kinase deficiency has also been observed in hyperimmunoglobulinaemia D and periodic fever syndrome. This syndrome is an autosomal recessive disorder characterized by recurrent episodes of fever associated with lymphadenopathy, arthralgia, gastrointestinal dismay, and skin rash. In patients with hyperimmunoglobulinaemia D/periodic fever syndrome mevalonate kinase was found to be strongly reduced (1.2 to 3.4 per cent of normal) but not fully deficient as in classical mevalonate kinase deficiency.

Acatalasaemia

Acatalasaemia is a rare disease which has mainly been described in Japan and Switzerland. In Japan, acatalasemia is associated with ulcerating, often gangrenous,

oral lesions whereas these abnormalities were not seen in Swiss patients.

Laboratory diagnosis of peroxisomal disorders: postnatal diagnosis

Although discussed in this chapter as a single group, the peroxisomal disorders comprise a heterogeneous group of disorders with a large spectrum of clinical signs and symptoms. Furthermore, different biochemical abnormalities are found in the various peroxisomal disorders reflecting the peroxisomal pathway(s) primarily affected. This immediately explains why there is not a single laboratory method allowing diagnosis of all peroxisomal disorders.

Earlier we introduced the concept of diagnostic groups in order to develop logical guidelines for the laboratory diagnosis of the various peroxisomal disorders. These diagnostic groups are:

I. The disorders of peroxisome biogenesis plus the disorders of peroxisomal b-oxidation with the exception of X-linked adrenoleucodystrophy which forms a distinct diagnostic group (group III). This group includes Zellweger syndrome, neonatal adrenoleucodystrophy, infantile Refsum disease, acyl-CoA oxidase deficiency, D-bifunctional protein deficiency, and peroxisomal thiolase deficiency.
II. Rhizomelic chondrodysplasia punctata complex.This includes: rhizomelic chondrodysplasia types 1, 2, and 3.
III. X-linked adrenoleucodystrophy complex.This includes all forms of X-linked adrenoleucodystrophy.
IV. The remaining peroxisomal disorders.This includes: racemase deficiency, hyperoxaluria type 1, Refsum disease, glutaryl-CoA oxidase deficiency, mevalonate kinase deficiency, and acatalasaemia.

Laboratory diagnosis of the disorders of diagnostic group I

As described earlier, the clinical presentation of the disorders of peroxisomal b-oxidation resembles that of the disorders of peroxisome biogenesis in many respects. This is immediately clear if is realized that the first cases of acyl-CoA oxidase deficiency were described under the name pseudoneonatal adrenoleucodystrophy and peroxisomal thiolase deficiency as pseudo-Zellweger syndrome. The similarity between the disorders of peroxisome biogenesis and peroxisomal b-oxidation is also clear in the case of D-bifunctional protein deficiency: patients with this defect show severe neurological abnormalities including hypotonia, seizures, and psychomotor retardation as well as craniofacial dysmorphia as discussed before.

The biochemical abnormalities of the different disorders belonging to diagnostic group I are listed in Table 3. Inspection of these data shows that very long-chain fatty acids are abnormal in all these disorders making very long-chain fatty acid analysis a reliable first-line diagnostic test to verify whether a patient with the clinical signs and symptoms of a disorder of peroxisome biogenesis or peroxisomal b-oxidation is truely affected by such a peroxisomal disorder. If abnormal, additional tests have to be performed to discriminate between either a disorder of peroxisome biogenesis or peroxisomal b-oxidation. This includes analyses in erythrocytes (plasmalogens) and plasma (bile acid intermediates, pristanic acid, and phytanic acid).

If plasmalogen levels are deficient, the diagnosis 'disorder of peroxisome biogenesis' is established. This should be followed by detailed studies in fibroblasts, to establish whether the defect is truely expressed in fibroblasts and to establish the gene defective in this patient. If plasmalogen levels are normal this usually, but not invariably, points to a disorder of peroxisomal b-oxidation thus emphasizing the value of examining fibroblasts.

Laboratory diagnosis of the disorders of diagnostic group II

Since the clinical characteristics of rhizomelic chondrodysplasia punctata type 1 (PTS2 receptor deficiency), type 2 (DHAPAT deficiency), and type 3 (alkyl-DHAP synthase deficiency) are very similar, it makes sense to include these three forms of chondrodysplasia punctata in a single diagnostic group.

The biochemical characteristics of these disorders are listed in Table 4. The fact that erythrocyte plasmalogens are deficient in all three types indicates that analysis of erythrocyte plasmalogens is a reliable initial laboratory test to establish whether one is dealing with rhizomelic chondrodysplasia punctata type 1, 2, or 3. Erythrocyte plasmalogens have always been found to be deficient, even in more mildly affected cases. This implies that if erythrocyte plasmalogens have been found to be normal, rhizomelic chondrodysplasia punctata type 1, 2, or 3 is excluded whereas the finding of deficient plasmalogen levels is diagnostic for all types of rhizomelic chondrodysplasia punctata. Detailed studies of fibroblasts are required to discriminate between types 1, 2, and 3.

Laboratory diagnosis of X-linked adrenoleucodystrophy complex (diagnostic group III)

As described above, there is great heterogeneity within X-linked adrenoleucodystrophy with childhood cerebral adrenoleucodystrophy and adrenomyeloneuropathy as the most frequent phenotypes. Studies in hundreds of patients have shown that analysis of plasma very long-chain fatty acids is a reliable initial test to verify whether a certain patient is affected by X-linked adrenoleucodystrophy. If abnormal, we usually proceed by doing a full study in fibroblasts followed by molecular analyses in blood cells of fibroblasts.

Fibroblast studies are not obligatory and it may be advisable to perform direct molecular studies in blood cells of the patient as soon as plasma very long-chain fatty acids have been found to be abnormal. Heterozygote detection is not so straightforward as for hemizygotes. Indeed, plasma very long-chain fatty acids levels have been found to be normal in about 15 per cent of obligate heterozygotes making such analysis unreliable for the detection of heterozygotes. For this reason we advocate molecular studies and omit very long-chain fatty acids analysis in families in which the molecular defect has been established. In case the family history is negative, we usually start by doing plasma very long-chain fatty acid analysis followed by detailed studies in fibroblasts including measurements of C26:0 b-oxidation activity, very long-chain fatty acid analysis, and analysis of the ALD protein by means of immunofluorescence microscopy. The latter method may be especially rewarding since it is a general feature of X-linked adrenoleucodystrophy that in many instances the product of the mutant *X-ALD* allele produces an unstable protein so that in cells from heterozygotes a mosaic pattern is observed with positive and negative cells upon immunofluorescence.

Laboratory diagnosis of the disorders of group IV

The disorders of group IV share no similarities and all require separate laboratory tests for diagnosis as described below.

Peroxisomal 2-methylacyl-CoA racemase deficiency

Patients with a deficiency of 2-methylacyl-CoA racemase are unable to degrade pristanic acid and the bile acid intermediates di- and trihydroxycholestanoic acid. For this reason pristanic acid and di- and trihydroxycholestanoic acid are elevated in patients whereas very long-chain fatty acids are normal. Although experience is limited, it is probably safe to say that postnatal diagnosis of such patients may either be based on analysis of pristanic acid by gas chromatography or mass spectrometry or bile acid intermediates preferably by tandem mass spectrometry. Definitive diagnosis of racemase deficiency requires detailed studies in fibroblasts including measurement of racemase activity making use of specifically synthesized substrates.

Primary hyperoxaluria type 1

In patients with hyperoxaluria type 1 alanine glyoxylate aminotransferase is deficient which leads to a block in glyoxylate detoxification. Glyoxylate may either be oxidized to oxalate or reduced to glycolate which explains why in most patients with primary hyperoxaluria type 1 there is increased urinary excretion of all three acids. Definitive diagnosis of primary hyperoxaluria type 1 requires a liver biopsy for assessment of alanine glyoxylate aminotransferase activity.

Refsum disease

In Refsum disease phytanoyl-CoA hydroxylase is deficient, leading to the impaired degradation of phytanic acid. Since phytanic acid is solely derived from exogenous sources, plasma phytanic acid levels may vary widely. Definitive diagnosis requires measurement of phytanoyl-CoA hydroxylase in fibroblasts followed by molecular analysis at the complementary DNA or preferably the genomic level.

Glutaryl-CoA oxidase deficiency (glutaric aciduria type 3).

Only a single patient with this defect has been described in the literature. In this patient there was increased urinary excretion of glutaric acid not due to glutaric aciduria type 1 (glutaryl-CoA dehydrogenase deficiency) or glutaric aciduria type 2 (electron transfer flavoprotein (ETF)/ETF dehydrogenase deficiency). Glutaryl CoA oxidase was deficient.

Mevalonate kinase deficiency

In the classic form of mevalonate kinase deficiency, urinary mevalonic acid is extremely high in all patients studied making urinary analysis of mevalonic acid a reliable method for identification. If abnormal, mevalonate kinase activity should be measured in various types of blood cell (except erythrocytes) and/or in fibroblasts followed by molecular studies.

In hyperimmunoglobulinaemia D/periodic fever syndrome mevalonate kinase is also deficient, although to a lesser extent compared with the classic type of mevalonate kinase deficiency. Importantly, urinary mevalonic acid may be completely normal in these patients, which implies that diagnosis should be based on direct enzyme analysis in white blood cells, platelets, and/or fibroblasts.

Acatalasaemia

The diagnosis of acatalasaemia is not discussed here.

Laboratory diagnosis of peroxisomal disorders: prenatal diagnosis

In recent years reliable methods for the prenatal diagnosis of virtually all peroxisomal disorders have become available, in most cases employing cultured chorionic villus samples. For this specialized area, the reader is referred to the literature listed below.

Further reading

Aubourg P *et al.* (1986). Neonatal adrenoleukodystrophy. *Journal of Neurology, Neurosurgery and Psychiatry* **49**, 77–86.

Elias ER *et al.* (1998). Developmental delay and growth failure caused by a peroxisomal disorder, dihydroxyacetonephosphate acyltransferase (DHAP-AT) deficiency. *American Journal of Medical Genetics* **80**, 223–6.

Ferdinandusse S *et al.* (2000). Mutations in the gene encoding peroxisomal alpha-methylacyl-CoA racemase cause adult-onset sensory motor neuropathy. *Nature Genetics* **24**, 188–91.

Goldfischer S *et al.* (1973). Peroxisomal and mitochondrial defects in the cerebro-hepato-renal syndrome. *Science* **182**, 62–4.

Gould SJ, Valle D (2000). Peroxisome biogenesis disorders: genetics and cell biology. *Trends in Genetics* **16**, 340–5.

Heymans HSA *et al.* (1985). Rhizomelic chondrodysplasia punctata: another peroxisomal disorder. *New England Journal of Medicine* **313**, 187–8.

Houten SM *et al.* (1999). Mutations in MVK, encoding mevalonate kinase, cause hyperimmunoglobulinaemia D and periodic fever syndrome. *Nature Genetics* **22**, 175–7.

Jansen GA *et al.* (2000). Human phytanoyl-CoA hydroxylase: resolution of the gene structure and the molecular basis of Refsum's disease. *Human Molecular Genetics* **9**, 1195–200.

Kelley RI *et al.* (1986). Neonatal adrenoleukodystrophy: new cases, biochemical studies, and differentiation from Zellweger and related peroxisomal polydystrophy syndromes. *American Journal of Medical Genetics* **23**, 869–901.

Lazarow PB, Moser HW (1995). Disorders of peroxisome biogenesis. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 2287–324. McGraw-Hill, New York.

Martinez, M *et al.* (1994). Blood polyunsaturated fatty acids in patients with peroxisomal disorders. A multicenter study. *Lipids* **29**, 273–80.

Moser HW, Smith KD, Moser AB (1995). X-linked adrenoleukodystrophy. In: Scriver CR *et al.*, eds. *The metabolic and molecular bases of inherited disease*, 7th edn, pp 2325–49. McGraw-Hill, New York.

Ofman R *et al.* (1998). Acyl-CoA-dihydroxyacetonephosphate acyltransferase—cloning of the human cDNA and resolution of the molecular basis in rhizomelic chondrodysplasia punctata type 2. *Human Molecular Genetics* **7**, 847–53.

Poll-The BT *et al.* (1987). Infantile Refsum disease: an inherited peroxisomal disorder. Comparison with Zellweger syndrome and neonatal adrenoleukodystrophy. *European Journal of Pediatrics* **146**, 477–83.

Poll-The BT *et al.* (1988). A new peroxisomal disorder with enlarged peroxisomes and a specific deficiency of acyl-CoA oxidase (pseudo-neonatal adrenoleukodystrophy). *American Journal of Human Genetics* **42**, 422–34.

Roels F, Espeel M, De Craemer D (1991). Liver pathology and immunocytochemistry in congenital peroxisomal diseases: a review. *Journal of Inherited Metabolic Diseases* **14**, 853–75.

Schram AW *et al.* (1987). Human peroxisomal 3-oxoacyl-coenzyme A thiolase deficiency. *Proceedings of the National Academy of Sciences of the USA* **84**, 2494–6.

Smith KD *et al.* (1999). X-linked adrenoleukodystrophy: genes, mutations, and phenotypes. *Neurochemical Research* **24**, 521–35.

Spranger JW, Opitz JM, Bidder U (1971). Heterogeneity of chondrodysplasia punctata. *Humangenetik* **11**, 190–212.

van Grunsven EG *et al.* (1999). Peroxisomal bifunctional protein deficiency revisited: resolution of its true enzymatic and molecular basis. *American Journal of Human Genetics* **64**, 99–107.

Wanders RJA, Schutgens RBH, Barth PG (1995). Peroxisomal disorders: a review. *Journal of Neuropathology and Experimental Neurology* **54**, 726–39.

Wanders RJA, Tager JM (1998). Lipid metabolism in peroxisomes in relation to human disease. *Molecular Aspects of Medicine* **19**, 69–154.

Wanders RJA, van Grunsven EG, Jansen GA (2000). Lipid metabolism in peroxisomes: enzymology, functions and dysfunctions of the fatty acid alpha- and beta-oxidation systems in humans. *Biochemical Society Transactions* **28**, 141–9.

Wanders RJA *et al.* (1988). Peroxisomal disorders in neurology. *Journal of Neurological Sciences* **88**, 1–39.

Wanders RJA *et al.* (1992). Human dihydroxyacetonephosphate acyltransferase deficiency: a new peroxisomal disorder. *Journal of Inherited Metabolic Diseases* **15**, 389–91.

Wanders RJA *et al.* (1994). Human alkyldihydroxyacetonephosphate synthase deficiency: a new peroxisomal disorder. *Journal of Inherited Metabolic Diseases* **17**, 315–18.

Watkins PA *et al.* Peroxisomal bifunctional enzyme deficiency. *Journal of Clinical Investigations* **83**, 771–7.

Wilson GN, Holmes RD, Custer J (1986). Zellweger syndrome: diagnostic assays, syndrome delineation and potential therapy. *American Journal of Medical Genetics* **24**, 69–82.

11.10 Disorders of oxalate metabolism

Richard W. E. Watts and C. J. Danpure

Introduction—oxalate, hyperoxaluria, oxalosis

The oxalate anion is metabolically inert in humans and its overall metabolism can be represented by a single-compartment model in which the oxalate metabolic pool is a little larger than the extracellular fluid volume and contains approximately 10 to 30 µmol of oxalate. The system is normally in equilibrium but expansion of the oxalate metabolic pool occurs in renal failure and when there is either overproduction or overabsorption of oxalate. Tissue deposition of calcium oxalate (oxalosis) only occurs when renal failure is combined with one of these other factors. The normal plasma oxalate concentration is 1 to 3 µmol/l and shows a circadian rhythm, being lowest in the morning and highest in the evening with superimposed postprandial rises. The urinary excretion of oxalate does not normally exceed 450 µmol/24 h in adults. The results in children are similar if they are adjusted to a standard body surface area (1.73 m^2), and adult levels are reached by about 14 years of age. The urinary excretion of oxalate also increases during the waking hours and shows seasonal variations related to dietary oxalate and calcium intakes; vitamin D supply also affects calcium absorption. The oxalate in the plasma and tissues is of dietary and biosynthetic origin. The following foods and beverages have particularly high oxalate contents: beets, beetroot, celery, chocolate, cocoa, nuts, rhubarb, strawberries, spinach, and tea. They provide about 0.8 to 1.7 mmol per day, of which, only about 5 per cent is absorbed depending upon the proportion that is in soluble form and the calcium content of the diet. Some dietary oxalate is degraded by gut commensal bacteria, for example *Oxalatobacter formigenes*. The main biosynthetic source of oxalate in humans is glycine; oxalate accounts for only about 1 per cent of the total glycine metabolic turnover, via the glyoxylate anion and the C_1–C_2 fragment of ascorbate.

Glycolate, hydroxyproline, serine and the sidechains of the aromatic amino acids have been shown to be minor metabolic precursors of oxalate in experimental animals. Negligible amounts of oxalate appear to be derived from carbohydrate and polyols (for example xylitol) under normal dietary conditions. The claim that the artificial sweetening agent diethylene glycol is converted to oxalate in humans has not been confirmed. The absorption of oxalate from the small intestine involves both an active carrier mediated transport system with oxalate–chloride exchange as well as passive diffusion.

The kidney handles the oxalate ion by 100 per cent filtration at the glomerulus, tubular secretion involving active tubular transport into the lumen, and passive backdiffusion into the peritubular capillaries. The ratio (oxalate clearance)/(glomerular filtration rate) is normally about 1.2, indicating net tubular secretion. Apart from acute oxalic acid poisoning, diseases attributable to oxalate present when calcium oxalate stones form in the urinary tract and when this salt crystallizes in either the renal parenchyma (calcium oxalate nephrocalcinosis) or in other tissues (oxalosis). The disorders of oxalate metabolism are due to either overproduction or excessive absorption of oxalate, although hyperoxaluria could, at least theoretically, arise from a renal tubular abnormality causing increased net renal tubular secretion of oxalate. Impaired renal function causes oxalate retention, and there is a linear relationship between the plasma oxalate and creatinine concentrations. However, the plasma oxalate concentration does not exceed the critical value (48.5 µmol/l) at which the solubility product of calcium oxalate in plasma is exceeded and oxalosis develops before endstage renal failure supervenes, unless there is also either oxalate overproduction, as in primary hyperoxalurias I and II, or oxalate hyperabsorption. Conversely, increased oxalate biosynthesis and hyperabsorption do not cause oxalosis unless recurrent oxalate urolithiasis and nephrocalcinosis have impaired renal function. The risk of oxalosis developing, as judged by rapidly rising plasma oxalate concentration and expansion of the oxalate metabolic pool, is greatly increased when the glomerular filtration rate decreases to about 25 ml/min/1.73 m^2. Hyperoxaluria is the hallmark of the disorders of oxalate metabolism and Table 1 lists its causes.

Primary hyperoxaluria type I

Epidemiology

The estimated prevalence of primary hyperoxaluria type I is about 10.5 per million inhabitants with an incidence of about 1 in 120 000 live births in European countries. The prevalence and incidence are greater in countries with a higher consanguinity rate.

Biochemistry and molecular biology

Primary hyperoxaluria type I is an autosomal recessive disorder of glyoxylate metabolism caused by deficiency of the liver-specific peroxisomal pyridoxal phosphate-dependent enzyme alanine:glyoxylate aminotransferase (EC 2.6.1.44). In some cases the deficient alanine:glyoxylate aminotransferase activity can be augmented by administration of pyridoxine. Alanine:glyoxylate aminotransferase normally catalyses the conversion of the intermediary metabolite glyoxylate to glycine, but its absence in primary hyperoxaluria type I allows glyoxylate to be oxidized to oxalate and reduced to glycolate instead (Fig. 1). The elevated synthesis of oxalate and glycolate leads to the hyperoxaluria and hyperglycolic aciduria characteristic of primary hyperoxaluria type I. In some families, a pseudodominant pattern of inheritance is apparent due to the segregation of three, rather than two, mutant alleles. Primary hyperoxaluria type I is phenotypically heterogeneous at both the enzymic and clinical levels. Three major enzymic categories are recognized, characterized by:

1. The absence of both alanine:glyoxylate aminotransferase catalytic activity and alanine:glyoxylate aminotransferase immunoreactive protein (ENZ–/CRM–).
2. The absence of alanine:glyoxylate aminotransferase catalytic activity but the presence of alanine:glyoxylate aminotransferase immunoreactive protein (ENZ–/CRM+).
3. The presence of both alanine:glyoxylate aminotransferase catalytic activity and alanine:glyoxylate aminotransferase immunoreactive protein (ENZ+/CRM+).

Many patients in the last category can have alanine:glyoxylate aminotransferase activities similar to those found in asymptomatic heterozygotes (Fig. 2). In most of these patients disease is caused by an unparalleled protein trafficking defect in which alanine:glyoxylate aminotransferase is mistargeted to the mitochondria, where it is unable to fulfil its metabolic function (i.e. glyoxylate transamination) properly.

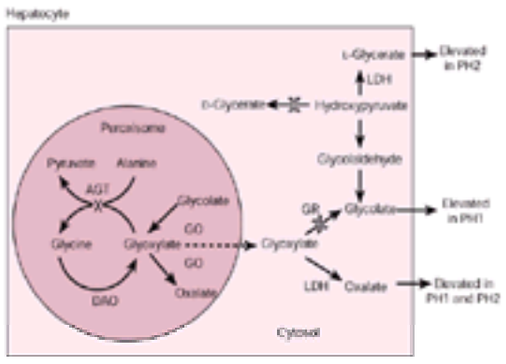

Fig. 1 Main pathways of glyoxylate metabolism in human liver cells. The black 'X' indicates the location of the defect in primary hyperoxaluria type I and the white 'X' indicates the same in primary hyperoxaluria type II: AGT, alanine:glyoxylate aminotransferase; GO, glycolate oxidase; DAO, D-amino acid oxidase; GR, glyoxylate reductase; LDH, lactate dehydrogenase. The peroxisomal membrane could be permeable to most or all of the metabolites shown. However, only the peroxisomal efflux of glyoxylate (dotted line) is shown to highlight the relationship between alanine:glyoxylate aminotransferase deficiency and the hyperoxaluria and hyperglycolic aciduria characteristic of primary hyperoxaluria type I.

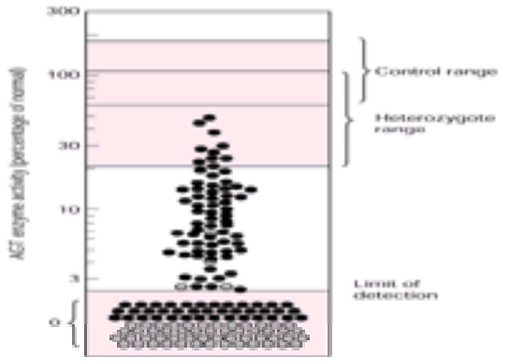

Fig. 2 Hepatic alanine:glyoxylate aminotransferase heterogeneity in patients with primary hyperoxaluria type I. Alanine:glyoxylate aminotransferase activity, expressed as a percentage of the mean normal control value, is shown for 162 patients with primary hyperoxaluria type I. Black circles, CRM+; grey circles, CRM–. Almost all patients with significant alanine:glyoxylate aminotransferase activity express the peroxisome-to-mitochondrion mistargeting phenotype.

Alanine:glyoxylate aminotransferase is encoded by the *AGXT* gene located on chromosome 2q37.3, where it spans about 10 kbases and consists of 11 exons and 10 introns. Numerous polymorphisms and mutations have been identified at the *AGXT* locus; more details can be found in the review by Danpure (2001) listed in Further reading. The most common mutation found in primary hyperoxaluria type I, with an allelic frequency of 30 per cent, leads to a substitution of arginine for glycine at residue 170 and, in combination with a proline®leucine polymorphism at residue 11, is responsible for the peroxisome-to-mitochondrion mistargeting of alanine:glyoxylate aminotransferase.

Pathology

In the early stages, the pathological findings are confined to the kidney, which shows a variable degree of hydrocalycosis and hydronephrosis with multiple calculi. Nephrocalcinosis forms later, causing severe renal fibrosis and shrinkage, and the kidney feels tough and gritty when incised. Changes due to renal hypertension and recurrent pyelonephritis are often present, and the renal tubules may be blocked by aggregates of calcium oxalate crystals, particularly if the terminal illness has been associated with a hypotensive oliguric episode. The characteristic rosette-like calcium oxalate monohydrate crystals are highly birefringent and easily recognized under a polarizing microscope. Their full extent will only be observed if either unfixed tissues are examined or if non-aqueous fixatives are used, but they are usually sufficiently insoluble for some to remain and be apparent after routine fixation in formal saline. They are found most extensively in the myocardium, the tunica media of muscular arteries and arterioles, the rete testes, and at sites of rapid bone turnover. Careful examination reveals a few crystals associated with the arterial supply of all organs and tissues. Similar deposits have been found intra-axonally in peripheral nerves.

Clinical aspects

Patients usually present with recurrent urolithiasis in childhood, and if untreated die from renal failure before they are 20 years old. The terminal phase of rapidly progressing oliguric renal failure usually lasts, at most, only a few months and is associated with dense calcium oxalate nephrocalcinosis and with the development of systemic oxalosis. Ischaemic lesions occur on the extremities, particularly in the pulps of the fingers and toes, and are attributable to the extensive crystallization of calcium oxalate in the walls of small muscular arteries and arterioles. Progressive peripheral neuropathy and mononeuritis multiplex are associated with calcium oxalate deposition within axons and in the walls of the vasa nervorum. These vascular and neurological manifestations, as well as a wider range of oxalotic features, occur particularly in patients in whom the terminal renal failure has been treated by standard haemodialysis, by peritoneal dialysis, or by an unsuccessful renal transplantation. Additional manifestations include livedo reticularis, subcutaneous calcinosis which may ulcerate, retinal changes (white flecks, exudates, infarcts, yellow crystalline deposits, especially along the courses of the ophthalmic arteries, black 'geographic' lesions at the macula), dilated cardiomyopathy, cardiac conduction defects, synovitis, a painful osteodystrophy with dense osteosclerosis and skeletal deformation, stress fractures (especially in the vertebrae), and the changes of coincidental secondary and tertiary hyperparathyroidism.

Hyperoxaluria may present during the first months of life with seizures, advanced renal failure, and dense nephrocalcinosis but few if any calculi (the infantile type). Another small group (the adult type) follows a benign course, presenting in adult life and surviving into the fourth and fifth decade with only occasional stone formation. This latter type shows less elevation of urinary oxalate excretion than is usual in the group (juvenile type) which follows the typical clinical pattern. However, the amount of oxalate excreted and the age at which renal failure develops are not very closely correlated.

Although the usual clinical pattern is one of recurrent stones, with or without nephrocalcinosis leading inexorably to renal failure, a few patients present with severe uraemia and may give no history of urolithiasis. A few present with symptoms arising from oxalosis involving principally the heart, arteries, bones, and peripheral nerves. Considering this clinical grouping in relation to the patient's age, pyridoxine responsiveness, and the plasma and urine oxalate concentrations gives a clinical guide to prognosis.

Although the 'metabolic lesion' of primary hyperoxaluria type I is located in the peroxisome, it does not share any of the features of the other peroxisomal diseases. Similarly, patients in whom alanine:glyoxylate aminotransferase has been misrouted into mitochondria do not show evidence of mitochondrial disease (impaired fatty acid oxidation and disorders of function of the respiratory chain).

Diagnosis

Primary hyperoxaluria type I should be considered in any child with urinary stones or nephrocalcinosis and in adults with recurrent calcium oxalate stones for which no alternative explanation has been found, especially if the clinical history extends back into childhood. The presence of calcium oxalate crystals in the urinary centrifuged deposit is not a specific diagnostic sign; the 24-h urinary oxalate excretion must be measured chemically and related to the creatinine excretion. The urinary oxalate excretion can be misleadingly low in patients in advanced renal failure. About 75 per cent of patients have an associated hyperglycolic aciduria. The definitive diagnosis is made by assaying alanine:glyoxylate aminotransferase activity on a percutaneous needle biopsy of the liver, which can also be examined by immunoelectron microscopy to establish whether the enzyme protein is present and its intracellular location. It is now possible to combine this with an assay for hydroxypyruvate/glyoxylate reductase, which is essential for the diagnosis of primary hyperoxaluria type II, on the same biopsy tissue. Figure 2 shows the results of

enzyme assays on patients, compared with the ranges obtained from carriers and other normals. The plasma oxalate concentration is determined when the patient is first evaluated and subsequently as a guide to prognosis and management. A progressively increasing plasma oxalate concentration indicates an increasing risk of oxalosis developing. A plasma oxalate value that is high relative to the corresponding creatinine is a valuable pointer to either oxalate overproduction or overabsorption. Early oxalosis is usually clinically silent. Some procedures that may be used to detect and evaluate it are listed in Table 2.

Prenatal diagnosis

Until a few years ago, prenatal diagnosis of primary hyperoxaluria type I was only possible by measuring alanine:glyoxylate aminotransferase activity in fetal liver biopsies in the second trimester. However, it is now possible to carry out the procedure in the first trimester by DNA (either mutation or polymorphism) analysis of material obtained from chorionic villi.

Treatment

Fluid and diet

Like all patients suffering from urinary stone disease those with primary hyperoxaluria type I should drink sufficient fluid to maintain a measured urine volume of 3 litres every 24 h with proportionately less in children. The diet should be low in oxalate and have minimum intakes of vitamins C and D, although this may need modification in children to meet the needs of growth and development.

Pyridoxine

The effect of pharmacological doses (150 to 1000 mg/day) of pyridoxine (pyridoxal phosphate is the prosthetic group in alanine:glyoxylate aminotransferase) on urinary oxalate excretion should be assessed over three 1-week periods, pretreatment, during treatment, and post-treatment, with assays of urinary oxalate and creatinine (a check for completeness of urine collection) on each 24-h urine collection. Smaller doses are occasionally effective. If the urinary oxalate decreases appreciably, pyridoxine should be continued indefinitely. A favourable response can probably be anticipated in between about 10 and 30 per cent of cases. The risks of pyridoxine-induced neuropathy have to be considered. Patients presenting in endstage renal disease may be given pyridoxine blindly.

Crystallization inhibitors

Citrates, either as sodium citrate (0.1–0.15 g/kg body weight/day) or equivalent doses of either sodium–potassium citrate (urolyte U Madaus®) or effervescent anhydrous sodium acid phosphate (phosphate Sandoz®) reduce the degree of urine calcium oxalate saturation. Neutral orthophosphates (equivalent to 2 g of elemental phosphorus per day) increase the excretion of pyrophosphate ions which inhibit heterogeneous calcium oxalate crystal nucleation, seeded growth, and aggregation. It also reduces calcium absorption. Magnesium supplements (for example 200 mg of magnesium oxide per day) also inhibit crystal growth and aggregation. The doses used should be sufficient to produce a material increase in either the urinary excretion of phosphate or magnesium. Phosphate and magnesium should be avoided if there is renal insufficiency. They may also produce diarrhoea.

Associated abnormalities

Coincidental urinary tract infections, hypercalciuria, and any urinary acidification defect should be treated vigorously on their merits. Excessive alkalization of the urine should be avoided because it may predispose to urinary tract infections and to the superimposition of phosphatic stones.

Treatment of urinary stones

Obstructive uropathy requires an immediate percutaneous nephrostomy to relieve the obstruction. Nephroscopic lithotomy, endoscopic lithotripsy with ultrasonic, electrohydraulic, and laser techniques, as well as extracorporeal shockwave lithotripsy, which produce relatively little damage to functioning kidney tissue, can be used to deal with asymptomatic stones. The kidneys should be kept as free from stones as possible and open lithotomy for large calculi should now rarely be needed. Stone debris may require either external drainage via a nephrostomy or internal drainage via a stent. However, stents and other foreign bodies in the urinary tract rapidly become encrusted with calcium oxalate deposits. Close follow-up is essential with regular radiological and/or ultrasonographic assessment. Patients who have previously passed stones often do so with little pain and unsuspected collections of stones may be found in the lower ureter on a routine abdominal radiograph. Acute hypotensive episodes are particularly dangerous causing intrarenal precipitation of calcium oxalate with acute and irreversible loss of renal function.

Transplantation

The oxalate ion has low dialysance and low filterability in the clinical context, and none of the currently available methods of replacing renal function can keep up indefinitely with the rate of oxalate overproduction once renal failure has occurred. Thus, almost all pyridoxine-resistant patients ultimately require orthotopic liver transplantation to correct the metabolic and genetic lesion together with a simultaneous renal transplant if they are approaching endstage renal disease. A small group of patients with relatively large amounts of residual catalytic enzyme activity may be suitable for an isolated renal graft. A renal transplant while the glomerular filtration rate is in the 15 to 20 ml/min/1.73 m^2 range may also be used to 'buy time' during which liver transplantation can be organized. The patient should be either haemodialysed or haemofiltered as vigorously as possible before, during, and after any grafting procedure in order to deplete the oxalate metabolic pool, minimize oxalosis, and reduce the risk to the grafted kidney.

Ideally, planning for liver and/or renal transplantation should begin when the glomerular filtration rate falls below 25 ml/min/1.73 m^2 (or 20 per cent of the mean predicted normal value) to minimize oxalosis and reduce the risk of oxalate deposits in the grafted kidney.

After a successful liver transplant the hyperglycolic aciduria returns to normal immediately. The plasma oxalate value normalizes over the course of a few weeks or months, and the urinary oxalate excretion returns to normal over the course of one or more years depending on the size of the oxalate deposits that are gradually mobilized from the tissues.

The plasma oxalate concentration should be followed sequentially before and after operation in these patients. Preoperatively, it is a guide to the extent to which the superimposition of oxalate retention on oxalate overproduction is occurring and hence to the rate at which the risk of oxalosis is increasing. Values approaching 35 µmol/l indicate that measures to reduce it are very urgently needed. Postoperatively, it is an indication of the risk of calcium oxalate damage to the grafted kidney.

Pre-emptive liver transplantation before renal failure has decreased to 25 ml/min/1.73 m^2 is an option if the disease is diagnosed early and is following an aggressive course. Partial orthotopic hepatic transplantation using tissue from a live related histocompatible donor has been proposed, but it is uncertain whether this would produce a clinically useful degree of metabolic correction. Heteroptic auxiliary liver transplantation is theoretically unsound. Although primary hyperoxaluria type I presents certain advantages as a candidate for gene therapy using retrovirus- or adenovirus-based constructs there has, as yet, been no definitive work in this area.

Primary hyperoxaluria type II

Primary hyperoxaluria type II is significantly rarer than primary hyperoxaluria type I and, as such, has been much less studied. Primary hyperoxaluria type II is caused by a deficiency of the widely-distributed enzyme glyoxylate reductase (EC 1.1.1.26/79, also known as hydroxypyruvate reductase and D-glycerate dehydrogenase, EC 1.1.1.29). Glyoxylate reductase catalyses a number of reactions, including the reduction of glyoxylate to glycolate and the reduction of hydroxypyruvate to D-glycerate (Fig. 1). Its absence in primary hyperoxaluria type II, however, allows glyoxylate to be oxidized to oxalate and hydroxypyruvate to be reduced to L-glycerate. The resulting increased synthesis of oxalate and L-glycerate, which leads to hyperoxaluria and hyper L-glyceric aciduria, is characteristic of primary hyperoxaluria type II.

Glyoxylate reductase is encoded by the *GRHPR* gene, which contains nine exons and eight introns, and spans about 9 kbases in the pericentromeric region of chromosome 9. A series of mutations in the *GRHPR* gene has been found in patients with primary hyperoxaluria type II. One case without hyper L-glyceric aciduria has been identified on the basis of immunoelectrophoresis.The clinical features, complications, and pathological findings are similar to those of primary hyperoxaluria type I. There have been no reports of possible biochemical and genetic heterogeniety of this disease. Enzyme replacement by organ transplantation has not been attempted in the type II disease, although if the liver proved to contain most of the complement of glyoxylate reductase activity, liver transplantation might be

beneficial. The expression of the enzyme in leucocytes suggests that bone marrow transplantation from a fully histocompatible sibling might also be a therapeutic option.

Primary hyperoxaluria type III

Patients with primary hyperoxaluria type III, which has been attributed to oxalate hyperabsorption, do not have an associated hyperglycolic or L-glyceric aciduria and have, therefore, to be distinguished from the approximately 25 per cent of patients with primary hyperoxaluria type I with isolated hyperoxaluria. The diagnosis of primary hyperoxaluria type III rests on: firm evidence of normal intestinal anatomy, histology, and absorptive function; the demonstration of excessive oxalate absorption (this may be difficult to establish by the available ^{14}C-labelled oxalate absorption and/or oxalate loading tests); normal hepatic alanine:glyoxylate aminotransferase and glyoxylate reductase levels. The urinary oxalate excretion (usually 1–2 mmol/24 h) is similar to that in some patients with the other types of primary hyperoxaluria, and type III patients are at risk of urinary stones, renal failure, and oxalosis. The metabolic lesion has not been identified but it might involve an oxalate–chloride exchanger in the small intestine. It has been reported that thiazides reduce urinary oxalate excretion in primary hyperoxaluria type III. As in secondary hyperoxaluria due to diffuse small intestinal disease, treatment is by a low-oxalate diet with oxalate binding agents such a cholestyramine and calcium ions (given as calcium carbonate). A marine hydrocolloid preparation (ox-absorb®) has recently been developed as an intestinal oxalate binding therapeutic agent. Patients with primary hyperoxaluria type III may not be a homogeneous group and some may represent an as yet unidentified disorder of either glyoxylate or glycolate metabolism. Others may be due to deficient colonization of the gut by *Oxalatobacter formigenes*. Disorders of renal tubular function are also a theoretical possibility.

Enteric hyperoxaluria

Enteric hyperoxaluria is an uncommon but potentially serious complication of the diseases listed under this heading in Table 1. It can cause extensive urolithiasis with nephrocalcinosis and renal failure. Expansion of the oxalate pool has been demonstrated and there is the same potential for oxalosis developing as in the primary hyperoxalurias. Treatment depends upon reducing dietary oxalate intake, the use of oxalate binding agents, and correcting the steatorrhoea or abnormal gut flora in the case of cystic fibrosis.

Further reading

Allan AR *et al*. (1996). Selective renal transplantation in primary hyperoxaluria type I. *American Journal of Kidney Diseases* **27**, 891–5.

Barratt TM, Danpure CJ (1996). Hyperoxaluria. In: Barratt TM, Avner ED, Harmon WE, eds. *Paediatric nephrology*, 4th edn, pp 609–24. Useful review with emphasis on paediatrics.

Cramer SD *et al.* (1999). The gene encoding hydroxypyruvate reductase (GRHPR) is mutated in patients with primary hyperoxaluria type II. *Human Molecular Genetics* **8**, 2063–9. Report of new findings.

Danpure CJ, Jennings PR (1986). Peroxisomal alanine: glyoxylate aminotransferase deficiency in primary hyperoxaluria type I. *FEBS Lett.* **201**, 20–4. First report of enzyme defect in primary hyperoxaluria type I.

Danpure CJ (2001). Primary hyperoxaluria. In: Scriver CR *et al*, eds. *The metabolic and molecular basis of inherited disease*, 8th edn, vol.II, ch. 133, pp 3323–67. McGraw-Hill, New York.

Danpure CJ, Rumsby G (1995). Enzymology and molecular genetics of primary hyperoxaluria type I. Consequences for clinical management. In: SR Khan, ed. *Calcium oxalate in biological systems*, pp 189–205 CRC, Boca Raton, FL.

Danpure CJ, Rumsby G (1996). Strategies for the prenatal diagnosis of primary hyperoxaluria type I. *Prenatal Diagnosis* **16**, 587–98.

Danpure CJ, Smith LH (1996). The primary hyperoxalurias. In: Coe FL *et al.*, eds. *Kidney stones: medical and surgical management*, pp 859–81. Lippincott-Raven, Philadelphia. Review.

Danpure CJ *et al*. (1989). An enzyme trafficking defect in two patients with primary hyperoxaluria type I: peroxisomal alanine/glyoxylate aminotransferase re-routed to mitochondria. *Journal of Cell Biology* **108**, 1345–52. Report of new findings.

Danpure CJ *et al*. (1989). Fetal liver alanine:glyoxylate aminotransferase and the prenatal diagnosis of primary hyperoxaluria type I. *Prenatal Diagnosis* **9**, 271–81.

Danpure CJ *et al*. (1994). Primary hyperoxaluria type I: genotypic and phenotypic herogeneity. *Journal of Inherited Metabolic Disease* **17**, 487–99. Review.

Danpure CJ *et al*. (1994). Molecular characterisation and clinical use of a polymorphic tandem repeat in an intron of the human alanine: glyoxylate aminotransferase gene. *Human Genetics 94* 55–64. Report of new findings.

Hoppe B *et al*. (1997). A vertical (pseudo-dominant) pattern of inheritance in the autosomal recessive disease primary hyperoxaluria type I. Lack of relationship between genotype, enzymic phenotype and disease severity. *American Journal of Kidney Disease* **29**, 36–44.

Latta K *et al*. (1995). Selection of transplantation procedures and perioperative management in primary hyperoxaluria type I. *Nephrology, Dialysis and Transplantation* 10: [Supplement 8]: 53–57

McKusick VA. OMIMTM. Online Mendelian Inheritance in Man. wep//http://www.ncbi.nlm.nih.gov/Onim/

Purdue PE *et al*. (1991). An intronic duplication in the alanine: glyoxylate aminotransferase gene facilitates identification of mutations in compound heterozygote patients with primary hyperoxaluria type I. *Human Genetics* **87**, 394–6.

Purdue PE *et al*. (1991). Characterization and chromosomal mapping of a genomic clone encoding human alanine: glyoxylate aminotransferase. *Genomics* **10**, 34–42.

Purdue PE, Takada Y, Danpure CJ. (1990). Identification of mutations associated with peroxisome-to-mitochondrion mistargeting of alanine/glyoxylate aminotransferase in primary hyperoxaluria type I. *Journal of Cell Biology* **111**, 2341–51.

Rumsby G, Creegeen D (1999). Identification and expression of a cDNA for human hydroxypyruvate/glyoxylate reductase. *Biochimica et Biophysica Acta* **1446**, 383–8.

von Schnakenburg C, Rumsby G (1997). Primary hyperoxaluria type I: a cluster of new mutations in exon7 of the *AGXT* gene. *Journal of Medical Genetics* **34**, 489–92.

Watts RWE (1994). Treatment of primary hyperoxaluria type I. *Journal of Nephrology* **7**, 208–14.

Watts RWE (1997). Primary hyperoxaluria: In: Sessa A *et al.*, eds. *Contributions to nephrology*, Vol. 122, pp 143–59. Karger, Basel.

Watts RWE, Veall N, Purkiss P (1983). Sequential studies of oxalate dynamics in primary hyperoxaluria. *Clinical Science* **65**, 627–33.

Williams HE, Smith LH Jr (1968). L-glyceric aciduria. A new genetic variant of primary hyperoxaluria. *New England Journal of Medicine* **278**, 233–8.

Yent ER, Cahanim M (1983). Absorptive hyperoxaluria: a new clinical entity-successful treatment with hydrochlorathiazide. *Clinical and Investigative Medicine* **9**, 44–50.

11.11 Disturbances of acid–base homeostasis

R. D. Cohen and H. F. Woods

In resting humans arterial blood, pH (pH_a) is normally maintained between 7.36 and 7.42, by control of the arterial partial pressure of CO_2 ($PaCO_2$) and plasma bicarbonate, between the limits 4.7 to 5.8 kPa and 24 to 30 mmol/l, respectively. Intracellular pH is also controlled, but varies substantially between organs within the range 6.3 to 7.4, depending on prevailing physiological or pathological circumstances. Some intracellular organelles are particularly acid, notably lysosomes. There is a substantial daily burden of hydrogen ions (protons) derived principally from metabolism (Table 1), and disordered neutralization or elimination of this burden shifts pH in the acid direction. Inappropriate loss of protons or proton-generating substrates, or excessive input of alkali, shifts pH in the alkaline direction.

Extra- and intracellular buffers, notably haemoglobin, other proteins, bicarbonate, and phosphate, play a transient role in countering acute pH changes but normally the acid burdens listed in Table 1 are ultimately eliminated quantitatively or neutralized. These burdens have been grouped into three classes according to their mode of disposal. Carbon dioxide derived from cellular respiration is much the largest potential generator of protons, the burden in the resting subject being an order of magnitude greater than that resulting from lactic and other organic acid production as well as from urea synthesis. Protons derived from the metabolism of sulphur- and phosphorus-containing compounds constitute the smallest source and represent a burden that is about one-hundredth of that derived from carbon dioxide. Disposal of carbon dioxide is dependent on adequate respiratory function. The metabolism of sulphur-containing amino acids in the diet eventually results in the production of so-called fixed acids, namely sulphuric acid and phosphoric acid, which may originate from many sources. Neither of these acids is volatile and they are thus excreted by the kidney.

The organic acids listed in Table 1 have pK values much below that of blood pH. They are therefore present in the blood as acid anions, rather than as the undissociated acids. The equivalent amount of hydrogen ions, generated at the site of production of these acids, titrate with local tissue and blood bicarbonate and other buffers. The organic acid anions (lactate, 3-hydroxybutyrate, acetoacetate, and fatty acids) are non-volatile but, unlike the fixed acids, may be eliminated by metabolism. Figure 1 shows an example of the important principle that when these organic acid anions are metabolized to electroneutral products (e.g. glucose, or carbon dioxide and water) protons are consumed and the bicarbonate consumed at their site of production is regenerated. Protons from organic acids can also be eliminated in the urine, but normally this is a much slower process than the metabolic route. Particularly in the case of the ketone bodies, for which the renal threshold is relatively low, substantial amounts can be lost in the urine when their plasma concentration is elevated. Although in maximally acidified urine (pH 4.5) about half of the urinary ketone bodies are in the form of the undissociated acid, the remaining free anion moiety represents loss of potential alkali, since it eludes metabolism to bicarbonate.

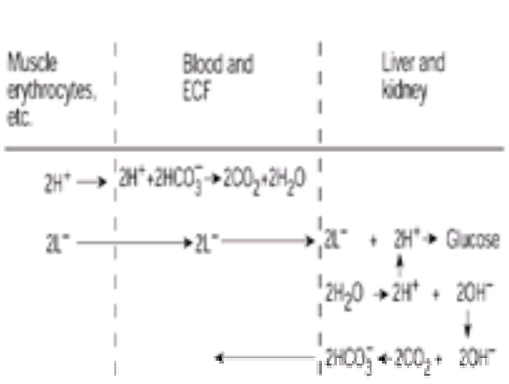

Fig. 1 A scheme, using lactate (L^-) conversion to glucose as an example, showing how conversion of the anion of an organic acid of low pK to an electroneutral substance consumes H^+ and regenerates HCO_3^-.

Despite the large quantitative differences in the burden due to the three classes of acid shown in Fig. 1, their correct elimination is, in a sense, equally important, for no class is able substantially to be disposed of by a route normally used to eliminate another class. Normally, the rates of production and elimination of each class of acid are matched in a long-term steady state. The homeostasis of arterial blood pH thus provided is given quantitative expression in the Henderson–Hasselbalch equation:

$$pH_a = 6.1 + \log_{10} \{[HCO_3^-]_a/(0.225 \times PaCO_2)\}$$

The constancy of arterial plasma bicarbonate concentration ($[HCO_3^-]_a$) is maintained by the removal of class I and II acids and by proton generation during ureagenesis (see below) and that of $PaCO_2$ by the lungs, thereby fixing pH_a within a narrow range.

The roles of the kidneys and liver in acid–base homeostasis

The interplay of these organs in acid–base homeostasis has been a matter of controversy. The authors' view is that the contribution of the kidneys is not entirely what it has appeared to be, and that the role of the liver requires emphasis. Classical descriptions are based on the kidneys as the principal controllers of $[HCO_3^-]$. Yet, as may be seen in Table 1, the liver is responsible for the major part of lactate disposal and consequent bicarbonate regeneration as well as the generation of ketoacids, with the opposite acid–base consequence. There is, however evidence that both these hepatic functions are normally controlled in an attempt to preserve acid–base homeostasis. Thus at normal concentrations of blood lactate, deviations in the acid direction enhance hepatic lactate disposal. As will be seen in the case of lactate, the homeostasis may be lost at higher concentrations of lactate. Ketogenesis is itself suppressed by increasing acidosis.

Urea production is another important feature of hepatic metabolism. The production of each molecule of urea (ultimately from ammonium and carbon dioxide) is accompanied by the generation of two protons. Ureagenesis is therefore a potential acidifying mechanism. Most of the protons produced in ureagenesis are neutralized by the bicarbonate generated during the oxidation of the carbon skeleton of amino acids, but normally there is a slight excess of protons produced, which have to be eliminated by the kidneys. Urea synthesis and accompanying proton production are negatively regulated by acidosis, which constitute another acid–base regulatory system intrinsic to the liver.

The renal tubules secrete protons and those not involved in the process of bicarbonate reabsorption are buffered by urinary phosphate. Under normal conditions, about 30 mmol/day of protons are excreted in this way. Classically, urinary ammonia (NH_3) is regarded as another buffer for hydrogen ions, which are therefore

removed as ammonium (NH_4^+); normally the excretion of hydrogen ions in the supposed NH_4^+ buffer amounts to about 70 mmol/day, but may increase to 500 mmol/day under severe acidifying stress in maximally acidified urine. It is, however, not possible to reconcile this view with the physicochemical fact that the ammonia/ammonium equilibrium is almost entirely in the form of ammonium at the time of generation from glutamine, and is therefore not available for buffering further protons. A more plausible explanation of the role of the kidneys is that the increase in ammonium excretion during acidosis serves to divert nitrogen from hepatic urea synthesis and consequent proton production, thus countering the acidosis.

These considerations provide a background to the interpretation of many acid–base syndromes described below.

Acid–base disorders

Definitions

The terminology of acid–base disturbances has always been confused. The terms acidaemia and alkalaemia simply indicate that pH_a is lower or higher than the normal range. Here we use the term acidosis to encompass both the situation where pH_a is low, and also that in which, although pH_a is normal, it would have been lowered if compensatory mechanisms had not occurred; an equivalent definition applies to alkalosis. When the primary disturbance is related to abnormal carbon dioxide elimination, the disturbance is referred to as 'respiratory'. All other primary disturbances, that is those related to disturbances of class II or III acid production or removal, are referred to as 'metabolic' or 'non-respiratory'. The term 'primary' is used to distinguish these processes from those which are compensatory in nature. Thus primary metabolic acidosis, lowering $[HCO_3^-]_a$, is compensated for by hyperventilation, which decreases $PaCO_2$; respiratory acidosis, driven by elevation of $PaCO_2$, is compensated for by metabolic events which result in an elevation of $[HCO_3^-]_a$.

The diagnosis of acid–base disturbances

Since the clinical manifestations of acid–base disturbances, described later, are frequently non-specific and may not be apparent until the disturbance is quite severe, laboratory investigation is indispensable. Measurement of pH_a, $PaCO_2$, and $[HCO_3^-]_a$ on arterial blood is the primary investigation. Estimation of plasma urea, creatinine, sodium, potassium, and chloride, and, when appropriate, lactate, ketoacids, and salicylate provides further important information.

Measurement of pH_a and $PaCO_2$—acid–base diagram

Blood gas analysers measure pH and PCO_2 and calculate plasma bicarbonate from the Henderson–Hasselbalch equation. Interpretation of results is best achieved by the use of an acid–base diagram which has pH_a and $PaCO_2$ as its axes. Diagrams which use $[HCO_3^-]_a$ on one of the axes are less suitable, since $[HCO_3^-]_a$ is calculated from pH_a and $PaCO_2$ and is not only subject to compounding errors in those measurements, but is affected by some poorly understood variations in pK_a in the Henderson–Hasselbalch equation in blood from severely ill patients.

The acid–base diagram in Fig. 2 has bands drawn in to show the ranges of data expected in uncomplicated acid–base disorders. It not only aids the diagnosis of acid–base disorders, but in addition the course of an individual patients disturbance and its response to treatment can be followed by serial plotting of data. The shaded square represents the approximate limits of pH_a and $PaCO_2$ in normal individuals. Thus a patient with uncomplicated metabolic acidosis will have values lying in the band marked 'metabolic' in the region above and to the left of the normal zone; the 'metabolic' band is the envelope of measurements of pH_a and $PaCO_2$ in patients with uncomplicated metabolic acidosis and alkalosis. The metabolic band is rather restricted on the alkalotic side, below and to the right of the normal zone. This is because compensation by hypoventilation for metabolic alkalosis is often poor; hypoxia may limit the degree of hypoventilation and metabolic alkaloses may be associated with intracellular acidosis, which could stimulate the respiratory centre. Marked hypocapnia is, however, occasionally seen in metabolic alkalosis.

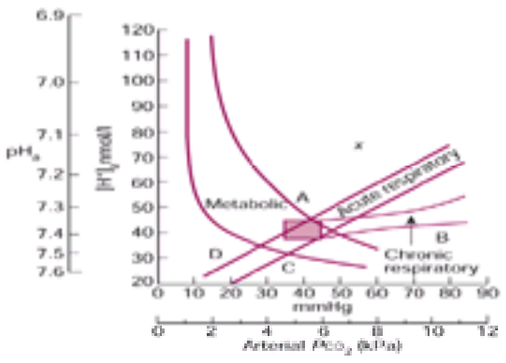

Fig. 2 A practical acid–base diagram. For explanation see text.

The band marked 'acute respiratory' is the 95 per cent confidence range of values obtained in normal individuals voluntarily hyperventilating or breathing air carbon dioxide mixtures for short periods. After a few days of carbon dioxide retention, an increase in plasma bicarbonate produces substantial or complete compensation for the respiratory acidosis; the band in chronic respiratory acidosis is therefore different from the acute response, the presence of the extra bicarbonate decreasing the fall in pH_a expected for a given rise in $PaCO_2$.

pH_a and $PaCO_2$ measurements in some patients will not fall within any of the defined bands in Fig. 2. Such patients have a mixture of acid–base disorders. Thus a patient whose pH_a and $PaCO_2$ are represented by the point '×' has mixed respiratory and metabolic acidosis, for example a patient with uraemic acidosis and an exacerbation of chronic bronchitis with respiratory failure. pH_a and $PaCO_2$ values lying in sectors A and C result from a combination of two primary acid–base conditions; in sectors B and D one of the two disturbances might be compensatory for the other.

Acid–base analytical equipment usually also provides at least two additional derived acid–base variables—the standard bicarbonate and base excess or deficit. The standard bicarbonate represents what the plasma bicarbonate would be if the blood had the normal $PaCO_2$ of 5.33 kPa (40 mmHg) rather than its actual value. Standard bicarbonate was introduced in an attempt to provide a measurement which was independent of respiratory disturbance and thus indicative of the underlying pure metabolic disturbance. Base deficit represents the amount of alkali in mmol needed to restore the pH of 1 litre of the patient's blood *in vitro* to normal (pH 7.4) at a PCO_2 of 5.33 kPa, and might at first sight be considered a quantitative measure of metabolic acidosis. Unfortunately the titration curve of blood *in vitro* is different from when it is circulating *in vivo*, since in the latter situation the interstitial and intracellular fluids also interact in the titration and may gain or lose bicarbonate from it; in addition, their buffering capacity differs from that of blood. These considerations detract from the usefulness of base excess/deficit as a guide either to diagnosis or therapy.

Further difficulties arise from ambiguities in the interpretation of base excess/deficit. Thus a patient with chronic respiratory acidosis will have a high standard bicarbonate and a base excess due to compensatory increase of plasma bicarbonate. It could be said, therefore, that this patient has simultaneously a respiratory acidosis and a metabolic alkalosis, as a base excess indicates the latter. This way of regarding the situation seems to us confusing and is incompatible with the definitions of acidosis and alkalosis we have given, which are intended to indicate the direction of the primary disturbance.

Use of the anion gap

In measurements of plasma electrolytes, the sum of the cations ($Na^+ + K^+$) normally exceeds that of the anions ($Cl^- + HCO_3^-$) by about 14 mmol/l (reference range 10–18 mmol/l). This difference is known as the anion gap and in health is attributable largely to the net negative charge on plasma proteins, but also to phosphate, sulphate, and several organic acids. Calculation of the anion gap is of great value in the differential diagnosis of metabolic acidosis, but the regrettably increasingly

common practice of omitting chloride estimation from sets of plasma electrolytes frequently deprives the clinician of this important diagnostic tool.

Metabolic acidoses may be divided broadly into those with high and those with normal anion gap. Metabolic acidoses with high anion gap are due to the ingestion or endogenous generation of acids, usually organic, whose anions are not measured in routine sets of plasma electrolytes. Plasma bicarbonate is titrated by these acids and therefore decreases; the anion gap is thus widened by the presence of these unmeasured anions. The most frequent organic acids concerned are lactic and ketoacids. In uraemic acidosis, the anion gap seldom exceeds 28 mmol/l. but considerably higher values may be found in severe lactic acidosis and ketoacidosis. It should be noted that there are causes of raised anion gap other than metabolic acidosis, for example therapy with sodium salts of relatively strong acids (e.g. lactate, acetate) and high-dose sodium carbenicillin treatment.

Metabolic acidoses with normal anion gap are due to the direct loss of bicarbonate from the body, either through the gut or fistulae, through the kidney, or, rarely, to the ingestion or infusion of acid or acidifying substances. When bicarbonate is lost, more chloride is retained by the renal tubules; thus low plasma bicarbonate is accompanied by hyperchloraemia and the anion gap remains unchanged.

Causes of acid–base disturbance

Table 2 classifies those conditions associated with high anion gap metabolic acidosis by the principal organic acid involved. Often a mixture of acids is involved but where possible the predominant acid has been shown in italics. Normal anion gap metabolic acidoses are shown in Table 3, classified according to whether they are due to gut or renal bicarbonate loss, or to ingestion or infusion of acidifying agents. Metabolic alkalosis (Table 4) is due either to ingestion or infusion of excessive alkali in circumstances when it cannot be excreted (e.g. poor renal function), or to the secretion of urine which is inappropriate both in its acidity and in its ammonium content. Most of the causes of the latter occurrence are related to the complex events in potassium and chloride deficiency and are dealt with later, as is the pathogenesis of the metabolic alkalosis of acute hepatic failure. Table 5 classifies respiratory acidosis according to the level of the problem, namely, the lungs and airways, the neuromuscular and mechanical aspects of respiration, and the central nervous system. Except in the case of deliberate or inadvertent external hyperventilation, respiratory alkalosis is always due to some form of stimulus to the respiratory centre, as classified in Table 6.

The effects of acid–base disturbances

These are widespread and we limit ourselves here to a brief description of those with known clinical consequences.

Respiratory effects

Both metabolic acidosis and acute respiratory acidosis induced by breathing high $P\mathrm{CO}_2$ gas mixtures result in hyperventilation. Deep sighing respiration (Kussmaul breathing) is a familiar sign of metabolic acidosis. pH control of ventilation is determined by the pH perceived by the carotid and aortic body chemoreceptors and by receptors in the medulla, which appear to monitor the pH of brain extracellular fluid. In the steady state, brain extracellular fluid pH is closely similar to that of cerebrospinal fluid. Sudden development of metabolic acidosis, resulting in low pH_a and arterial bicarbonate, induces hyperventilation by stimulating the carotid body and aortic chemoreceptors and $P\mathrm{aCO}_2$ is thus lowered. However, the first effect on brain extracellular fluid pH is to raise it. This is because brain extracellular fluid $P\mathrm{CO}_2$ is lowered since carbon dioxide is rapidly equilibrated across the blood–brain barrier. However, it takes many hours for the brain extracellular fluid bicarbonate to fall in response to the lowering of plasma bicarbonate, because movement of bicarbonate across the barrier is much slower than that of carbon dioxide. The temporary alkalinization of brain extracellular fluid somewhat offsets the extra ventilatory drive from the carotid and aortic chemoreceptors, so the hyperventilatory compensatory response takes some hours to reach its maximum. Though clinical circumstances usually prevent the observation of this sequence of events, the opposite, namely persistence of hyperventilation after restoration of normal pH during therapy of metabolic acidosis is commonly seen, and may last for over 24 h.

In chronic respiratory failure, with high $P\mathrm{aCO}_2$, direct depression of the respiratory centre occurs; the respiratory response to increments of $P\mathrm{aCO}_2$ is progressively lost and ventilation becomes increasingly dependent on hypoxic drive. Alkalosis also may depress respiration and increases the difficulties of weaning artificially ventilated patients from the respirator.

Cardiovascular effects

Acidosis decreases cardiac contractility (negative inotropism) and alkalosis has smaller but opposite effects. Acidosis and alkalosis both predispose to cardiac arrythmias. The negative inotropic effects are particularly related to changes in myocardial intracellular pH and are experimentally found to be rather greater in acute respiratory than in acute metabolic acidosis. In the rat, progressive metabolic acidosis reduces cardiac output as a result of bradycardia and negative inotropy; there is consequent hypotension and decreased renal and hepatic blood flow. This sequence of events may provide a model for the circulatory collapse which often occurs in patients after some hours of metabolic acidosis not originally attributable to shock. Mild to moderate metabolic acidosis has not often been associated with negative inotropic effects in the intact animal; this appears to be due to the protective effects of catecholamine release, which is increased in acidosis. In more severe acidosis, this protection breaks down. Patients receiving b-blockers may be more susceptible to the negative inotropic effects of acidosis.

Cerebral arterioles are very sensitive to the pH of brain extracellular fluid; they dilate when this falls and constrict when the pH rises. The cerebrovascular resistance is thus subject to the same type of phased responses to acid–base disturbances as described above for ventilation. Dilatation is also the response of most systemic arterioles to acidosis, although this response may be modified by catecholamine effects. The peripheral veins, however, constrict in acidosis, resulting in a shift of blood from the peripheral capacitance vessels to the central circulation. This effect has been shown to have important clinical consequences during treatment (see below).

Effects on intermediary carbohydrate metabolism

In all tissues in which observations have been made, glycolysis is inhibited by acidosis and stimulated by alkalosis, due to the effects of intracellular pH on phosphofructokinase, a rate-limiting enzyme of glycolysis. Respiratory alkalosis might therefore be expected to raise blood lactate but this effect is usually small, probably due to removal of lactate by the liver. However, in the presence of severe liver disease, gross elevation of blood lactate may be seen in association with respiratory alkalosis, and the increased production of lactate may partially compensate for the alkalosis.

Animal studies have shown that hepatic gluconeogenesis from lactate is inhibited by acidosis due to an effect on the metabolic step between pyruvate and oxaloacetate. This phenomenon may override the stimulatory effect on hepatic lactate disposal described earlier and may be responsible for perpetuating and worsening lactic acidosis.

Effects on nitrogen balance

Chronic acidosis produces negative nitrogen balance, mainly due to accelerated proteolysis in skeletal muscle. This effect is mediated through increased expression of the genes coding for ubiquitin and proteasome subunits.

Effects on blood oxygen uptake and delivery to the tissues

One of the factors determining blood uptake of oxygen during passage through the lungs, and subsequent delivery of oxygen to the tissues, is the position of the blood oxygen dissociation curve with respect to the abscissa ($P\mathrm{O}_2$). Right shifts of this curve improve unloading of oxygen in the tissues, but under some circumstances may impair oxygen uptake in the lungs. Left shifts have the opposite effect. The position of the curve is determined by three haemoglobin ligands, namely hydrogen ions, carbon dioxide, and 2,3-bisphosphoglycerate (2,3-BPG). Increases in all of these shift the curve to the right. Changes in intraerythrocytic pH and $P\mathrm{CO}_2$ are often rapid, but those of 2,3-BPG are much slower. In chronic acidosis, the synthesis of 2,3-BPG is inhibited and marked reductions in the erythrocyte content of this metabolite may occur, with opposite effects in alkalosis. These changes are, however, slow in comparison with the immediate effects of changes in pH and $P\mathrm{CO}_2$ (the Bohr effect).

The effect of these differences in time scale on oxygen delivery gives rise to a characteristic sequence of events during the development and treatment of acute metabolic acidosis. Initially, the acute acidosis causes a right shift of the curve, and thus improved oxygen release to the tissues. After several hours, erythrocyte

2,3-BPG falls, thus restoring the position of the curve towards normal. If the patient is now rapidly treated with alkali, the Bohr effect results in rapid shift to a position to the left of normal, because of the low level of 2,3-BPG. The resulting sudden deterioration of oxygen release may have adverse clinical effects unless the consequences of left shift are ameliorated by other factors, such as an increase in tissue blood flow. It may be many hours or days before erythrocyte 2,3-BPG concentrations are restored to normal.

Effects on the nervous system

Severe acidosis is frequently associated with impairment of consciousness, varying from mild drowsiness to coma. This effect is not closely related to systemic pH, and the mechanism is poorly understood. The effects on the respiratory and cardiovascular centres have been discussed above. The excitability of neural and muscular tissues is increased by alkalosis and diminished by acidosis. Tetany is a common feature of respiratory alkalosis, and may also be seen when chronic metabolic acidosis is corrected in patients with hypocalcaemia, a sequence of events which may occur in chronic renal failure. Epileptic attacks in susceptible individuals may be precipitated by alkalosis and suppressed by acidosis.

Effects on potassium homeostasis

Acute acidosis results in a shift of potassium out of the intracellular compartment into the extracellular fluid. Hyperkalaemia is thus often seen in the acidosis of renal failure, untreated diabetic ketoacidosis, and in acute respiratory failure; its mechanism is not entirely clear, factors other than extracellular pH being implicated. Alkali therapy in such patients causes a shift of potassium back into cells. As substantial amounts of potassium may be lost in the urine during the period of hyperkalaemia, overall depletion of body potassium occurs; thus alkali therapy may result in a rapid fall of plasma potassium to dangerous levels. This is a well-known hazard in the treatment of diabetic ketoacidosis and is even more dangerous in types 1 and 2 renal tubular acidosis in which plasma potassium is frequently low in the presence of acidosis (see also under Treatment below).

Chronic metabolic alkalosis is also frequently accompanied by potassium depletion, which results from distal tubular potassium secretion uninhibited by competition with hydrogen ions for secretion.

Effects on the kidney

The kidney is a major organ of acid–base regulation and many of its responses are therefore geared to acid–base homeostasis. Proton secretion is a principal function of tubular cells and in the proximal tubule is a crucial part of the mechanism for the apparent reabsorption of the large quantities of bicarbonate filtered at the glomerulus. In the cortical and medullary collecting tubules, where the main acidification of the urine takes place, the intercalated cell is equipped with the H^+-ATPase, residing in the apical (luminal) membrane, and the band 3 general anion exchanger, in the basolateral membrane. Under acid conditions H^+ and bicarbonate are generated by carbonic anhydrase within these intercalated cells. The protons are secreted into the lumen, where they titrate the phosphate buffer, or convert any bicarbonate present into carbon dioxide and water; the bicarbonate is transported by the anion exchanger in the opposite direction into the blood stream. There is now evidence of considerable plasticity of function of these cells so that in alkaline conditions the polarity of transporter expression is reversed, with the H^+-ATPase now predominating in the basolateral membrane and the anion exchanger in the luminal membrane, secreting bicarbonate into the urine. The maximum urinary pH thereby achieved is about 8.0.

As indicated earlier, there is a large increase in renal ammonium production and excretion in the urine in acidosis. The ammonium ions are derived from glutamine by the action of glutaminase in the proximal tubule; they are mainly secreted by pH-dependent non-ionic diffusion into the collecting tubule lumen, where the blood-lumen pH gradient is the greatest in acidosis. Chronic acidosis results in an increased expression of glutaminase and phosphoenolpyruvate carboxykinase. The latter enzyme is rate-limiting for gluconeogenesis, and an increase in renal gluconeogenesis is thought to play a crucial role in the high rate of ammonium production. A reinterpretation of the role of increased ammonium excretion in acidosis has been discussed above.

Effects on the distribution of metabolites and drugs

Many weak acids and bases are distributed between body compartments by the simple physicochemical process of pH-dependent non-ionic diffusion, which is based on movement of the non-dissociated hydrophobic moiety across the lipid membranes separating compartments, quite independently of any transporter. The pH differences between the compartments will determine the relative concentrations in the two spaces at equilibrium. Weak acids accumulate in the more alkaline compartment and weak alkalis in the more acid compartment. Examples of physiological metabolites distributed by this mechanism include ammonia/ammonium (weak base) and urobilinogen (weak acid). The distribution of ammonium and other amines present in advanced liver disease between blood and cerebrospinal fluid is partly determined in this way. Examples of drugs exhibiting this behaviour are salicylates and phenobarbitone (weak acids); use of their pH-dependent distribution is made in the treatment of poisoning with these drugs by forced alkaline diuresis.

Effects on bone

Bone acts as a buffer in chronic metabolic acidosis. Leaching out of bone calcium carbonate and exchange of extracellular phosphate for carbonate within the apatite crytal result in the neutralization of protons. The first of these mechanisms causes a negative calcium balance in chronic metabolic acidosis and in chronic uraemic acidotic subjects it has been shown that calcium balance can be restored by treatment with sodium bicarbonate. Although chronic metabolic acidosis in rats results in osteoporosis, renal tubular acidosis and the acidosis associated with ureterosigmoidostomy may lead to osteomalacia, which can be corrected by alkali therapy alone.

Effects on leucocytes

Severe acidosis is often associated with marked leucocytosis, unrelated to the presence of infection. Blood leucocyte counts of up to 60 000/mm^3 have been recorded in lactic acidosis and high values are also common in diabetic ketoacidosis. This phenomenon may be partly a specific reaction to acidosis and not merely a general manifestation of stress or dehydration.

Major acid–base syndromes

Lactic acidosis

In normal resting individuals, venous blood lactate concentration is in the range 0.6 to 1.0 mmol/l. In extreme exercise this may rise to 20 mmol/l or more. Lactate is the end-product of anaerobic glycolysis. Its production by many tissues, even at rest, is accompanied by equal amounts of hydrogen ions, since its pK is low (3.8) and the undissociated acid is therefore present only in minute amounts. These protons react with blood and tissue bicarbonate to form carbon dioxide and water, but the lost bicarbonate is quantitatively restored when the lactate is converted to glucose (see Fig. 1), mainly in the liver, or oxidized in many tissues to carbon dioxide and water. When lactate is produced at a rate which exceeds the disposal rate, the regeneration of bicarbonate is incomplete and acidosis results. The pathological mechanisms leading to lactic acidosis are therefore excess production, defective disposal, or, commonly, a mixture of both. As the acidosis develops, hepatic disposal of lactate by gluconeogenesis may become further inhibited (see above), leading to a vicious circle which provides a model for the often fulminating course of lactic acidosis.

Clinically, lactic acidosis falls into two main categories. In type A lactic acidosis, much the more common, there is clinical evidence of shock, poor tissue perfusion, or hypoxia. Though increased peripheral glycolysis is an important contributor, associated poor hepatic and renal perfusion limit the lactate disposal mechanisms. Indeed in circulatory failure, the liver and kidneys may produce lactate rather than dispose of it. In type B lactic acidosis, there is no evidence at the outset of circulatory insufficiency or hypoxia, though after many hours of increasing acidosis these may supervene. The original diagnosis and cause of acidosis may be obscured if the patient does not present until this late stage. Type A lactic acidosis is a frequent manifestation of haemorraghic, septic, cardiogenic, or traumatic shock and there is a direct relationship between the concentration of blood lactate and poor prognosis. The causes of type B lactic acidosis (Table 2) are varied and some of the mechanisms will be described below.

The initial clinical presentation in type B lactic acidosis is fairly uniform and consists of hyperventilation or dyspnoea, drowsiness or coma, vomiting, and abdominal pain, in approximately that order of frequency. The condition usually develops over a few hours, but may be more chronic, for example in the mitochondrial myopathies. Although by definition there is initially no clinical evidence of poor tissue perfusion or hypoxia, patients with severe type B lactic acidosis commonly

become shocked after a few hours.

The diagnosis of lactic acidosis is based on the clinical circumstances, including the presence of a known aetiological factor, the presence of a high anion gap acidosis, and the measurement of blood lactate, for which automated apparatus is now widely available.

Biguanide-induced lactic acidosis

This class of oral hypoglycaemic agent has widespread metabolic effects, including inhibition of gluconeogenesis and the monocarboxylate transporter responsible for movement of lactate ions across cell membranes, and stimulation of glycolysis. The lactic acidosis is of the type B variety though, as indicated above, circulatory insufficiency may eventually supervene. The principal culprits were phenformin and buformin and the mortality was about 50 per cent; these biguanides are no longer used. Metformin is, however, widely used; the incidence of lactic acidosis with metformin is less than one-tenth of that with phenformin. Since metformin is almost entirely excreted in the urine, lactic acidosis may be largely avoided by care not to prescribe this drug in patients with even mild degrees of renal insufficiency, or conditions such as uncontrolled heart failure which might be expected to diminish renal function. Attempts have been made to show that the risk of lactic acidosis in diabetics taking metformin is no greater than in diabetics not receiving the drug. There have, however, been serious criticisms of those studies, and since the use of metformin is likely to increase because of its value in the prevention of diabetic microvascular complications, we strongly advise that the above precautions for its use continue to be followed.

Postictal lactic acidosis

The severe muscular contractions during convulsions may produce severe lactic acidosis in the same way as vigorous exercise. The finding of lactic acidosis in these circumstances occasionally gives rise to confusion but may be distinguished from other causes of lactic acidosis by the rapid decline of blood lactate after the cessation of convulsions, with a half-life of approximately 20 min.

Lactic acidosis in liver disease

Though impaired disposal of an administered lactate load is readily demonstrable in chronic liver disease, clinical lactic acidosis is uncommon. However, in the later stages of fulminant hepatic necrosis it may be an important part of the clinical picture. Acid–base disturbances in the earlier stages are discussed below.

Lactic acidosis in severe falciparum malaria

Lactic acidosis is a common feature of severe malaria due to *P. falciparum*, particularly in children, where it is the strongest predictor of poor prognosis. Although shock may be a factor, the lactic acidosis is frequently of the type B variety, and is attributable to many factors, including production of lactate by the parasite itself, occlusion of the microcirculation by parasites, the direct effects of high circulating levels of certain cytokines, notably tumour necrosis factor, inhibition of gluconeogenesis from lactate because of decrease in hepatic blood flow, and overproduction of lactate during the convulsions which are a common feature of cerebral malaria. Hypoglycaemia in severe malaria may be linked with lactic acidosis; it may be a manifestation of decreased gluconeogenesis or of insulin release due to quinine therapy. Acidosis appears to increase the attachment of infected erythrocytes to capillary walls, perhaps thus worsening the capillary blockage seen in the cerebral circulation and other sites. It may also inhibit the uptake of antimalarial drugs into erythrocytes.

Lactic acidosis associated with treatment with nucleoside reverse transcriptase inhibitors

There have been reports of lactic acidosis, which may be severe, associated with AIDS therapy with nucleoside reverse transcriptase inhibitors. Two mechanisms have been described—riboflavine deficiency and a mitochondrial disorder, with myopathy associated with the characteristic ragged red fibres seen in inherited mitochondrial myopathies. In the former type, the lactic acidosis rapidly responds to the administration of riboflavine.

Ethanol and methanol-induced lactic acidosis

Ingestion of ethanol after a period of fasting is a well-known cause of hypoglycaemia, which may be severe. The phenomenon is due to inhibition of gluconeogenesis, which is the sole source of endogenous glucose output when glycogen stores have been depleted. The defect in gluconeogenesis may result in moderate lactic acidosis, because ethanol diverts some of the NAD^+ needed for the oxidation of lactate to pyruvate, the first step in lactate disposal, for its own oxidation, catalysed by alcohol dehydrogenase. With withdrawal of ethanol, administration of glucose, and refeeding, the condition is normally self-limiting.

In methanol poisoning, the main contributor to the acidosis is formic acid, but lactic acidosis also plays a part, due to inhibition of gluconeogenesis by similar mechanisms to those in ethanol-induced lactic acidosis.

Salicylate and ethylene glycol poisoning

See Chapter 8.1.

D(–) lactic acidosis

In all the lactic acidoses described above, the stereoisomer involved is L(+)lactate, the end product of mammalian glycolysis. However, a few cases have been described in which the acidosis has been due to D(–)lactate. There is a very minor pathway of D(–)lactate production in mammalian tissues, but in D(–)lactic acidosis the D(–)lactate arises as a product of glycolysis in bacteria in the gut, and all cases have been associated with short gut or jejunal–ileal bypass syndromes or the therapeutic ingestion of large quantities of *Lactobacillus acidophilus*. The lactic acidosis may be severe and is often associated with disturbances of consciousness; it is presumably due to absorption of large quantities of D(–)lactate from the gut, since it may be treated by appropriate oral antibiotics. In healthy individuals, infused D(–)lactate is cleared at approximately 70 per cent of the rate for L(+)lactate, but by the non-specific 2-hydroxybutyrate dehydrogenase rather than L(+)lactate dehydrogenase. The main problem in diagnosing D(–)lactate acidosis is that D(–)lactate is not detectable by the routine blood lactate assay which employs the enzyme L(+)lactate dehydrogenase. If the condition is suspected because of unexplained high anion gap metabolic acidosis in a patient with a predisposing condition, then D(–) lactate should be assayed either by gas chromatography or using a bacterial D(–)lactate dehydrogenase.

Diabetic ketoacidosis

The pathogenesis and clinical features of diabetic ketoacidosis are described elsewhere (Chapter 12.11). Only the acid–base disturbance will be discussed here. Though the acidosis is conventionally regarded as being due mainly to overproduction of ketoacids by the liver, recent evidence has suggested that the hydrogen ions are wholly or partly derived from other tissues, though the liver is, of course, the source of ketoacid anions. Hepatic gluconeogenesis, a major source of the hyperglycaemia of diabetic ketoacidosis, proceeds at increased rates, in spite of potential inhibition by systemic acidosis. This is because, unlike in acidoses of other origins, hepatic intracellular pH does not fall in diabetic ketoacidosis.

Diabetic ketoacidosis has usually been regarded as a typical high anion gap metabolic acidosis in which extracellular bicarbonate has simply been titrated by the ketoacids. If this were the case, the fall in plasma bicarbonate should roughly equal the rise in anion gap and the plasma concentration of ketoacids. However, Adrogué and colleagues have shown that, whilst in some cases this is true, the situation is frequently more complex. Patients who present in ketoacidosis with relatively well-preserved renal function tend to have a rise in anion gap which is much less than the fall in bicarbonate. This is due to the loss of large quantities of ketoacid anions in the urine, with concomitant tubular reabsorption of chloride to maintain electroneutrality. Hyperchloraemia develops, which, together with the urinary loss of ketoacids, results in a relatively small elevation of anion gap compared with the bicarbonate deficit. In contrast, patients who have relatively poor renal function on presentation, for example because of dehydration, have much smaller urinary losses of ketoacids and present with a more classical high anion gap metabolic acidosis.

The total blood ketone body concentration in the well-controlled diabetic is about 0.1 mmol/l. In diabetic ketoacidosis, the concentration is often above 10 mmol/l and can rise as high as 30 mmol/l. Some of the most severe acidoses seen in clinical practice occur in diabetic ketoacidosis, some with pH_a values as low as 6.8. Urinary pH reaches its minimum possible value (4.5–5.3). At the lower of these values, about half the urinary ketoacids are undissociated and some hydrogen ions are lost in

this way. Severe depletion of erythrocyte 2,3-BPG occurs, leading to left shift of the oxygen dissociation curve, especially during treatment, and with potentially adverse consequences (see below).

In 5 to 10 per cent of patients with diabetic ketoacidosis, there is an accompanying element of lactic acidosis, with blood lactate greater than 5 mmol/l. Lactic acidosis occurs particularly when the patient is shocked, but there are rare instances of lactic acidosis supervening when treatment of the initial ketoacidosis is well advanced. There are also occasional ketotic diabetics in whom the blood lactate is low. This effect is readily reproducible in experimental animals and is thought to be related to increased hepatic disposal of lactate and suppression of peripheral glycolysis by the acidosis.

The acidosis of renal failure

Metabolic acidosis of varying degree is a classical feature of acute and chronic renal failure. It has been traditionally attributed to failure of the kidneys to excrete protons derived from 'fixed acids'—the class III acids of Table 1. In chronic renal failure the remaining functional nephrons are usually able to lower the urinary pH to the normal minimum. However, failure of proximal tubular bicarbonate reabsorption may occasionally occur and leads to a bicarbonate leak, as in type 2 renal tubular acidosis (see Chapter 20.8); in this case urinary pH does not fall to its minimum until the filtered load of bicarbonate has been substantially reduced by the fall in plasma bicarbonate. In some conditions, for example chronic pyelonephritis and chronic obstructive uropathy, the renal medulla is particularly affected, and acidification of the urine may be impaired. Nevertheless, the usually normal acidification in chronic renal failure means that the phosphate buffers in the tubular lumen are titrated by protons to the same extent as is possible in normal kidneys. However, the excretion of ammonium ions is lower than normal in chronic renal failure because of the loss of glutaminase-containing proximal tubules and reduced renal blood flow decreases the supply of glutamine. The conventional explanation of the acidosis of renal failure has been that the diminished supply of ammonia from the glutaminase reaction lowers the ability of the luminal contents to buffer secreted protons, with the result that the minimum urinary pH is attained with less protons in the ammonia/ammonium buffer system, and thus disposing of fewer protons in the urine.

However, as indicated above, the ammonia/ammonium buffer system is already virtually entirely in the protonated form (i.e. ammonium) at the time of its generation in the glutaminase reaction, so there is no capacity remaining for this system to act as a urinary buffer, either in health or in renal failure. An alternative explanation is therefore required for the acidosis of renal failure. Atkinson and Camien have suggested that the nitrogen which would in health be excreted as ammonium ions in the urine is, in chronic renal failure, effectively diverted to the liver, where it is converted to urea with accompanying generation of protons (see above); the acidosis of renal failure is therefore due to relative overproduction of urea, rather than to failure of excretion of protons in the urine as ammonium ions.

The anion gap in uraemic acidosis seldom exceeds 28 to 30 mmol/l. The elevation is due to the accumulation of a relatively small quantities of several acid anions, including phosphate, sulphate, citrate, and other less well characterized contributions. When there is an element of proximal bicarbonate wastage, the anion gap may not be grossly raised; chloride may be reabsorbed instead of bicarbonate, leading to moderate hyperchloraemia.

The renal tubular acidoses, which are tubular disorders not, in the first instance, accompanied by glomerular failure, are discussed in Chapter 20.8.

Metabolic alkaloses associated with potassium and chloride deficiency

The most common cause of metabolic alkalosis is that associated with the use of potassium-losing diuretics; pyloric stenosis and Bartter's syndrome provide further examples of a complex aetiology. Chloride deficiency, indicated by low plasma chloride, may be due to a direct action of the diuretics, to loss from the gastrointestinal tract, as in pyloric stenosis, or to potassium deficiency itself, which has been shown experimentally to impair renal retention of chloride. Normally, most renal sodium reabsorption takes place in the proximal tubule, and it has to be accompanied by a readily reabsorbable anion to maintain electroneutrality. The most readily reabsorbable anion is chloride, and if the filtered load of chloride is low, due to hypochloraemia, some of the sodium which normally would have been reabsorbed proximally, passes to the distal segment of the nephron. Here sodium is reabsorbed by exchange with cations, principally potassium and protons, rather than accompanied by an anion. Since priority over acid–base regulation is accorded to the demands of extracellular volume control, the sodium reabsorption thus dictated causes further loss of potassium and protons into the urine, when the homeostatic response would have been to retain these latter ions. This accounts for the observation that the urine in these circumstances is acid when it should be alkaline ('paradoxical aciduria') and contains substantial quantities of potassium. Potassium loss in the urine is the principal cause of potassium depletion in pyloric stenosis, not loss in the vomit. The hypokalaemia is exacerbated by the fact that extracellular fluid volume is depleted in both diuretic therapy and in pyloric stenosis, leading to activation of the renin/angiotensin/aldosterone system, with further potassium loss. These considerations have important implications for therapy (see below).

An important cause of hypokalaemic hypochloraemic alkalosis, which is often marked, is deliberate overuse by patients of diuretics, notably frusemide, for reasons which may be related to psychological disturbances of the body image. Many of these patients are secretive about their use of diuretics; measurement of plasma frusemide is one way of diagnosing this dangerous condition.

Acid–base disturbances in fulminant hepatic failure

The most frequent acid–base disturbance in the earlier stages of fulminant hepatic failure is respiratory alkalosis, presumably due to the effects of ammonium and other amines on the respiratory centre. Metabolic alkalosis is also frequent, probably due to the failure of ureagenesis and its accompanying proton generation, but in some cases could be contributed to by potassium deficiency. Whatever the mechanism of the alkalosis, blood lactate concentration is frequently elevated, even in the absence of circulatory insufficiency. This phenomenon has been attributed to stimulation of peripheral glycolysis by alkalosis and to impairment of hepatic lactate disposal. Lactic acidosis may be a major feature in the later stages when the circulation is compromised, but is also occasionally seen in the very early stages. This early lactic acidosis is observed in paracetamol poisoning, a common cause of fulminant hepatic failure; it is associated with hypoglycaemia and may be largely related to a direct effect of paracetamol metabolites on hepatic gluconeogenesis from lactate; however, mild hypotension and dehydration also occurs.

Principles of treatment of acid–base disorders

The mainstay of treatment of acid–base disorders is to eliminate the cause of the disorder, the acid–base control mechanisms then restoring the normal situation in due course. However, it may be necessary to make a direct attempt to restore or partly restore normal acid–base status.

The treatment of respiratory acidosis is discussed in Chapter 17.11.

Acute metabolic acidosis

Major controversies still exist as to the advantages of treatment, especially with sodium bicarbonate. Randomized, controlled trials have not resolved these issues completely, largely because of the great variation of the physiological state of patients on presentation and the difficulty in establishing adequate sized, matching groups for trial purposes. Here we attempt to distinguish between conditions in which there is general consensus as to the best therapeutic approach and those in which there is less agreement. This is not an ideal approach, but is unavoidable at present.

The potential advantages of treating severe acidosis directly are improvement in cardiac performance, reduced risk of cardiac arrhythmia, redistribution of the blood volume away from the central circulation, correction of hyperkalaemia, and restoration of hepatic lactate disposal. Disadvantages lie in adverse effects on the oxygen dissociation curve, circulatory overload, especially if isotonic solutions have to be used, alkaline 'overshoot' when the acidosis is due to organic acids such as lactic acid and ketone bodies, and, allegedly, if bicarbonate is used, paradoxical intracellular acidification.

Paradoxical intracellular acidification is a concept arising from the observation that when sodium bicarbonate is infused, a significant rise in P_aCO_2 is observed, due to the titration of bicarbonate by hydrogen ions. Since there is an expectation that carbon dioxide will diffuse into cells much more rapidly than bicarbonate is translocated, it would be expected that intracellular pH would fall at the same time as pH_a rises; since many of the adverse effects of acidosis are directly related to effects on intracellular pH, this would be undesirable. A related observation is that in circulatory insufficiency, $P_{\bar{v}}CO_2$ in mixed venous blood may be much greater than in arterial blood, where it is often normal. On occasion, bicarbonate therapy may exaggerate this difference and it has been inferred that bicarbonate must be acidifying the intracellular compartment by the mechanisms outlined above. However, considerations of the mechanisms responsible for the mixed venous hypercapnia suggest that only if arterial P_aCO_2 is raised after passage of the blood through the lungs is bicarbonate therapy likely to cause intracellular acidification. Whether P_aCO_2 is indeed elevated by intravenous bicarbonate therapy is dependent on numerous factors, including cardiac output, ventilation, the pulmonary dead

space to total volume ratio, and, in particular, the rate of administration of bicarbonate.

Paradoxical intracellular acidification can indeed be demonstrated in closed systems such as platelets in which the carbon dioxide is not removed, and has been the reason for the development of alternative therapies such as an equimolar mixture of sodium bicarbonate and carbonate, for which the rise in P_aCO_2 during administration is substantially attenuated. However, it is difficult to demonstrate such an effect *in vivo*, when carbon dioxide is removed by the lungs, and in experimental animals either no change or actual elevation in intracellular pH in heart, liver, and skeletal muscle may be observed during bicarbonate administration, despite elevation of mixed venous PCO_2. In any case, if doubts remain concerning this issue, the problem may be avoided by the simple expedient of administering bicarbonate slowly. Hindman has made useful calculations of the rates of bicarbonate administration which avoid rises in P_aCO_2 under a range of circumstances.

The situations in which sodium bicarbonate therapy is generally agreed to be advantageous are as follows:

1. Metabolic acidosis in severe renal failure—in acute renal failure, sodium bicarbonate treatment may correct hyperkalaemia by shifting potassium into the intracellular compartment. It may also relieve distressing hyperventilation and make time for definitive renal support therapy to be introduced. If the patient is already fluid overloaded, the bicarbonate may be administered as a hypertonic solution (e.g. 8.4 per cent; 1 mmol/ml). If the patient is dehydrated then the isotonic solution (1.4 per cent; 0.163 mmol/ml) may be given. During haemofiltration for acute renal failure in the presence of lactic acidosis, there is little doubt that the use of bicarbonate rather than lactate-containing replacement fluid is preferable, because of failure of metabolism of lactate to bicarbonate in this situation (see below). In chronically uraemic patients, the use of oral bicarbonate to treat the acidosis may improve well-being, nitrogen balance, and the osteomalacic component of renal osteodystrophy.
2. In the acidosis of severe diarrhoea it has been shown, in the specific instance of cholera, where circulatory insufficiency and loss of alkali are prominent factors, that treatment with bicarbonate is superior to that with sodium chloride solutions. These patients have severe peripheral venoconstriction, displacing their blood volume towards the lungs. The administration of saline solutions may thus induce pulmonary oedema before the volume deficit has been replaced. Sodium bicarbonate appears to relieve the peripheral venoconstriction and full replacement is therefore less hazardous.
3. In renal tubular acidosis, both chronically and in exacerbations—this subject is discussed elsewhere (Chapter 20.8) but it is necessary to re-emphasize here that in types 1 and 2 renal tubular acidosis, where hypokalaemia is a prominent feature, it is mandatory to deal with the hypokalaemia either before or at least simultaneously with the acidosis. Treatment of the acidosis first will result in a further fall in plasma potassium, by driving potassium into the cells, with potentially fatal consequences.

It is in the treatment of lactic acidosis, particularly type A, and in diabetic ketoacidosis that the uncertainty of the value of bicarbonate therapy principally lies. In animal models of lactic acidosis, treatment with bicarbonate has been shown to produce less favourable or no better haemodynamic and metabolic results than with sodium chloride. Interestingly, in an acute model of haemorraghic shock there was little difference between bicarbonate and saline therapies; in these experiments, there had been no time for erythrocyte 2,3-BPG to fall. In contrast, in a model of diabetic ketoacidosis, developing over 48 h, in which 2,3-BPG was now virtually undetectable, treatment with bicarbonate produced a fall in blood pressure and evidence of tissue hypoxia, despite intracellular alkalinization. This suggests that acute acidoses of more than a few hour's duration may become increasingly susceptible to the adverse effects of bicarbonate. In a prospective, randomized trial of bicarbonate compared with saline in critically ill patients with lactic acidosis due to shock, there was no advantage of one therapy over the other, but this trial was not exempt from the general difficulties in mounting such trials.

The following practical guidance is therefore empirical, and largely based on current practice rather than formal trials:

1. Lactic acidosis—of paramount importance in type A lactic acidosis is the correction of hypovolaemic, cardiogenic, and other factors which are the primary cause of the condition. Such correction will promote aerobic metabolism and promote intracellular metabolism of lactate with consequent regeneration of bicarbonate (Fig. 1). Whether administration of exogenous bicarbonate can hasten this process or confer other benefit is uncertain. In short-duration lactic acidosis, there may be less concern about effects on oxygen dissociation related to 2,3-BPG levels, but it should be remembered that unless therapy directed at the primary cause has improved the circulation, alkalinization itself may produce an unfavourable effect on oxygen release from haemoglobin. Nevertheless, the possibility of bicarbonate helping in some circumstances cannot be ruled out. If it is given, this should be relatively slowly, as the isotonic solution, unless there is a circulatory overload problem, and only in amounts sufficient to raise pH_a to a relatively 'safe' level. In the special case of the acidosis of cardiac arrest, the previous priority given to bicarbonate administration has disappeared, because of lack of evidence of efficacy and the risk of alkaline overshoot when the high levels of lactate are metabolized on restoration of cardiac output. Hypertonic sodium bicarbonate is only now recommended as a secondary treatment after prolonged arrest. In type B lactic acidosis due to biguanides, it has been conventional to use bicarbonate therapy at least to bring pH_a to 7.2 to 7.4 over several hours and survival has been linked to the achievement of that goal. However, an alternative interpretation of the data is that those patients in whom acid–base status was restored to normal would have achieved this without the aid of bicarbonate, or with the use of saline instead. Other therapies for type B lactic acidoses of different causes are discussed below.
2. Diabetic ketoacidosis—it is generally accepted that provided pH_a is not below 7.0, bicarbonate treatment is not indicated. Rehydration and insulin therapy result in improved renal function, a fall in ketone body production and an increase in ketone body metabolism, all of which contribute to correction of the acidosis. When pH_a is below 7.0 many give just sufficient bicarbonate (isotonic) to bring pH_a just above 7.0, with careful attention to changes in plasma potassium. There is, however, no evidence that this regimen is better than rehydration (with saline), insulin and potassium replacement alone, and there are data showing that such therapy delays fall in ketone body levels and lactate (if raised). The delayed fall in lactate is consistent with tissue hypoxia related to low erythrocyte 2,3-BPG as discussed above. It is therefore common practice not to give bicarbonate even at low pH_a levels. If bicarbonate is given, then it must be isotonic (1.4 per cent,163 mmol/l); to give hypertonic bicarbonate would merely exacerbate the already present hyperosmolality. The amount required seldom exceeds 0.5 to 1 litre.

The amount of alkali therapy, if given, should be determined by an iterative process of administration of a relatively small amount (e.g. 80 mmol), followed by reassessment of the clinical state, pH_a and P_aCO_2, with the aid of serial plots on the acid–base diagram in Fig. 2 before repeating the cycle.

Alkalinizing agents other than bicarbonate have been considered. Sodium lactate has the disadvantage that it has to be metabolized to neutral products plus bicarbonate before it has an alkalinizing effect (Fig. 1) and if lactate metabolism is impaired, as in shock and hypoxia in particular, it has no effect. Lactate is not a buffer in its own right at pH values encountered in health or disease. The mixture of bicarbonate and carbonate referred to above ('Carbicarb') has not been shown to have clinical advantages and is less effective than saline in animal models of diabetic ketoacidosis. THAM (trishydroxyaminomethane buffer) has the theoretical advantage of producing intracellular as well as extracellular alkalinization, but is seldom used because of unwanted effects. Sodium dichloroacetate increases lactate disposal via oxidation by activation of pyruvate dehyrogenase, and markedly lowers blood lactate in critically ill patients with lactic acidosis. However, in a multicentre trial it did not improve survival over that in a saline control group, possibly because of the severity of the associated pathologies and the diverse clinical state of the patients. However, it is of proven value in the treatment of some congenital lactic acidoses. Thiamine produces dramatic resolution of the severe lactic acidosis sometimes seen in beri-beri. Riboflavine has produced similar results in nucleoside reverse transcriptase inhibitor-associated lactic acidosis.

Treatment of metabolic alkalosis

The first imperative is to identify the primary cause and if possible eliminate it. Where potassium-losing diuretics are responsible, they can be replaced by potassium-sparing preparations. In the most common form of chronic metabolic alkalosis, namely that associated with potassium and chloride deficiency, the primary therapies are potassium and chloride replacement. It has been shown that the potassium deficiency and alkalosis cannot be fully corrected unless there is also replenishment of chloride. Potassium supplements, whether oral or intravenous, should therefore be in the form of potassium chloride or contain other sources of chloride. It is also necessary to deal with any element of extracellular fluid volume contraction to switch off the drive to the renin–aldosterone system. It is seldom necessary to resort to administration of acid.

Further reading

Adrogué HJ, Wilson H, Boyd AE, Suki WN, Eknoyan G (1982). Plasma acid-base patterns in diabetic ketoacidosis. *New England Journal of Medicine* **307**, 1603–10.

Alberti KGMM, Darley JH, Emerson PM, Hockaday TDR (1972). 2,3-bisphosphoglycerate and tissue oxygenation in uncontrolled diabetes mellitus. *Lancet* **3**, 391–5.

Atkinson DE, Camien MN (1982). The role of urea synthesis in the removal of metabolic bicarbonate and the regulation of blood pH. *Current Topics in Cellular Regulation* **21**, 261–302.

Bellingham A, Detter JC, Lenfant C (1971). Regulatory mechanisms of hemoglobin oxygen affinity in acidosis and alkalosis. *Journal of Clinical Investigation* **50**, 700–6.

Chariot P, Drogou I, de Lacroix-Szmania I, *et al.* (1999). Zidovudine-induced mitochondrial disorder with massive steatosis, myopathy, lactic acidosis, and mitochondrial myopathy. *Journal of Hepatology* **30**, 156–60.

Cohen RD (1990). The metabolic background to acid-base homeostasis and some of its disorders. In: Cohen RD, Lewis B, Alberti KGMM, Denman AM, eds. *The Metabolic and molecular basis of acquired disease*, pp. 962–1001. Balliere Tindall, London.

Cohen RD (1991). Roles of the liver and kidney in acid-base regulation and its disorders. *British Journal of Anaesthesia* **67**, 154–64.

Cohen RD (1994). Lactic acidosis—new perspectives on origins and treatment. *Diabetes Reviews* **2**, 86–97.

Cohen RD, Woods HF (1976). *Clinical and Biochemical Aspects of Lactic Acidosis*. Blackwell, Oxford.

Cohen RD, Woods HF (1983). Lactic acidosis revisited. *Diabetes* **32**, 181–91.

Cohen RD, Woods HF (1999). Metformin and lactic acidosis. *Diabetes Care* **22**, 1010.

Cooper DJ, Walley KR, Wiggs BR, Russell JA (1990). Bicarbonate does not improve hemodynamics in critically ill patients who have lactic acidosis. *Annals of Internal Medicine* **112**, 492–8.

Emmett M, Narins RG (1977). Clinical use of the anion gap. *Medicine, Baltimore* **56**, 38–54.

Goldsmith DJA, Forni LG, Hilton PJ (1997). Bicarbonate therapy and intracellular acidosis. *Clinical Science* **93**, 593–8.

Hale PJ, Crase J, Nattrass M (1984). Metabolic effects of bicarbonate in the treatment of diabetic ketoacidosis. *British Medical Journal* **289**, 1035–8.

Hilton PJ, Taylor J, Forni LG, Treacher DF (1998). Bicarbonate-based haemofiltration in the management of acute renal failure with lactic acidosis. *Quarterly Journal of Medicine* **91**, 279–83.

Hindman BJ (1990). Sodium bicarbonate in the treatment of subtypes of acute lactic acidosis: physiologic considerations. *Anesthesiology* **72**, 1064–76.

Hood VL, Tannen RL (1998). Protection of acid-base balance by pH regulation of acid production. *New England Journal of Medicine* **339**, 819–26.

Kassirer JP, Berkman PM, Laurenz DR, Schwartz WB (1965). The critical role of chloride in the correction of hypokalaemic alkalosis in man. *American Journal of Medicine* **209**, 655–8.

Krishna S, Waller DW, ter Kuile F, *et al.* (1994). Lactic acidosis and hypoglycaemia in children with severe malaria: pathophysiological and prognostic significance. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **88**, 67–73.

Luzatti R, Del Bravo P, Di Perri G, Luzzani A, Concia E (1999). Riboflavine and severe lactic acidosis. *Lancet* **353**, 901–2.

Mitchell JH, Wildenthal K, Johnson RL (1972). The effects of acid-base disturbances on cardiovascular and pulmonary function. *Kidney International* **1**, 375–89.

Oh MS, Phelps KR, Traube M, *et al*. (1979). D-lactic acidosis in a man with the short bowel syndrome. *New England Journal of Medicine* **301**, 249–52.

Orringer CE, Eustace JC, Wunsch CD, Gardner LB (1977). Natural history of lactic acidosis after grand mal seizures. *New England Journal of Medicine* **297**, 796–9.

Record CO, Iles RA, Cohen RD, Williams R (1975). Acid-base and metabolic disturbances in fulminant hepatic failure. *Gut* **16**, 144–9.

Stacpoole PW, Barnes CL, Hurbanis MD, Cannon SL, Kerr DS (1999). Treatment of congenital lactic acidosis with dichloroacetate. *Archives of Diseases in Childhood* **77**, 535–41.

Weil MH, Rackow EC, Trevino R, Grundler W, Falk JL, Griffel MI (1986). Difference in acid-base state between venous and arterial blood during cardiopulmonary resuscitation. *New England Journal of Medicine* **315**, 153–6.

11.12.1 The acute phase response and c-reactive protein

M. B. Pepys

The acute phase response

Trauma, tissue necrosis, infection, inflammation, and malignant neoplasia induce a complex series of non-specific, systemic, physiological and metabolic responses including fever, leucocytosis, catabolism of muscle proteins, and the greatly increased *de novo* synthesis and secretion of a number of plasma proteins. The synthesis of albumin, transthyretin, and high and low density lipoproteins is correspondingly decreased, and the altered plasma protein concentration profile is called the acute phase response (Table 1). Most acute phase proteins are synthesized by hepatocytes, in which transcription is controlled by cytokines including, interleukin 1 (IL-1), interleukin 6 (IL-6), and tumour necrosis factor (TNFa). The circulating concentrations of complement proteins and clotting factors increase by up to 50 to 100 per cent whereas some of the proteinase inhibitors and a_1-acid glycoprotein can increase three to five fold. C-reactive protein (CRP) and serum amyloid A protein (SAA) are unique in that their concentrations can change by more than one-thousand fold. The response persists in individuals with chronic infections, chronic inflammation, or invasive or metastatic neoplasms, and is sustained, unless there is complete hepatocellular failure, until death. All endothermic animals mount a similar response, suggesting that it may have survival value, and increased availability of proteinase inhibitors, complement, clotting, and transport proteins presumably enhances host resistance, minimizes tissue injury, and promotes regeneration and repair. However, some acute phase proteins may be harmful. For example sustained, increased production of SAA can lead to the deposition of AA-type, reactive systemic amyloid, a serious and usually fatal condition that complicates chronic infective and inflammatory diseases. CRP, through its capacity to activate complement, can exacerbate ischaemic, and possibly also other forms, of tissue damage.

C-reactive protein

CRP was the first protein to be discovered which behaves as an acute phase reactant, and was named for its calcium-dependent interaction with the somatic C-polysaccharide of pneumococci, in which CRP recognizes phosphocholine residues. CRP also binds to other substances which contain phosphocholine, including phospholipids, some plasma lipoproteins, and the plasma membranes of damaged or apoptotic but not intact cells. In addition, CRP binds specifically to small nuclear ribonucleoprotein particles when these are exposed in dead or damaged cells.

Ligand-bound CRP activates the classical complement pathway via Cl, and can trigger the inflammatory, opsonizing, and complex-solubilizing activities of the complement system. A significant biological function of CRP may thus be to recognize and 'scavenge' cellular debris, promoting its safe clearance and helping to maintain tolerance to potential autoantigens. CRP may also protect against infection with pneumococci and *Haemophilus influenzae*, organisms that can express phosphocholine, and may thus contribute to innate immunity. On the other hand, CRP can also have tissue-damaging effects. Complement activation by CRP exacerbates ischaemic injury, the proinflammatory actions of CRP and its binding to phospholipids and lipoproteins may be proatherogenic, and the capacity of CRP to stimulate tissue factor production by macrophages may be prothrombotic. In contrast, CRP can suppress polymorph migration and infiltration *in vivo* and this may be anti-inflammatory.

The CRP molecule consists of five identical, non-glycosylated, non-covalently-associated polypeptide subunits, each of mass 23 027 Da, and containing 206 amino acid residues. The subunits have a flattened b-jellyroll fold with a single intrachain disulphide bond, and are arranged in an annular configuration with cyclic pentameric symmetry. There is a single calcium-dependent ligand binding site on the medial aspect of each subunit, all located on the same planar face of the molecule. A distinct but closely related plasma protein, serum amyloid P component (SAP), which is not an acute phase protein in man, has a very similar molecular structure with the same fold, characteristic of the 'lectin fold' superfamily. CRP and SAP belong to the pentraxin family that has been highly conserved in evolution and no structural polymorphism of CRP has been observed nor has any case of CRP deficiency been described.

Serum concentration of CRP

CRP is a trace protein in overtly normal, healthy individuals, the median value being 0.8 mg/l, with an interquartile range of 0.3 to 1.7 mg/l. Ninety per cent of apparently healthy subjects have levels of less than 3 mg/l and 99 per cent less than 10 mg/l. Serial studies of normal subjects and of monozygotic and dizygotic twins show that each individual's baseline CRP value is rather constant and is substantially genetically determined. Occasional higher values of CRP seen in ostensibly healthy people almost certainly reflect intercurrent subclinical pathology, and it is clear both that values greater than 3 mg/l are not normal and that, if they persist, they may have considerable clinical significance. In large surveys of the unscreened general population, there is a trend towards higher values with increasing age, with the median value rising to about 2 mg/l, and this probably reflects the higher incidence of pathological processes, such as atherosclerosis, osteoarthritis, and other diseases. Serum CRP concentrations are lower in healthy newborns but reach adult values within a few days.

Serum CRP concentration rises rapidly in the acute phase response and can exceed 300 mg/l by 48 h after a severe stimulus such as myocardial infarction, acute systemic bacterial infection, major trauma, or surgery. With uncomplicated resolution of injury or effective treatment of infection the circulating CRP concentration generally falls equally rapidly.

The speed of change and incremental range of CRP concentration are exceptional among all the acute phase proteins, apart from serum amyloid A protein which behaves in a similar fashion. The half-life of CRP in the circulation is 19 h and is constant in all conditions regardless of the presence of an acute phase response or its cause. In contrast to other acute phase proteins, such as clotting factors, complement proteins, transport proteins, and proteinase inhibitors, CRP does not undergo significant local sequestration or consumption, fragmentation, or complex formation. This means that, unlike most of the other acute phase reactants, the single major determinant of the circulating concentration of CRP is its rate of synthesis. Since this in turn is dependent on the intensity of the acute phase stimulus, the serum CRP level usually closely reflects the extent and activity of disease. These properties underlie the value in clinical practice of precise measurement of the serum CRP concentration. Drug or other treatments do not affect CRP production unless they also affect the disease process which is responsible for induction of CRP synthesis. The only exception is combined cyclosporin and steroid treatment given after renal transplantation. This suppresses the CRP response to renal allograft rejection, though not that provoked by infection. The only physical condition which seriously interferes with the capacity to interpret CRP levels is serious hepatocellular impairment, since CRP is made exclusively in the liver.

Clinical measurement of serum CRP concentration

Conditions associated with major elevation of serum CRP concentration

Most tissue-damaging processes, infections, inflammatory diseases of unknown aetiology, and malignant neoplasms are associated with a major acute phase response of CRP. CRP production is thus a non-specific response to disease and it can never, on its own, be used as a diagnostic test. However, if the CRP result is interpreted in the light of full clinical information about the patient it can provide exceptionally useful information for clinical management. Thus in nearly all the conditions listed in Table 2 the CRP level reflects quite precisely the extent and activity of disease. With deterioration the CRP value rises, whilst with spontaneous or

therapeutically induced remission the CRP level falls, and it thereby supplies an objective index of progress which is rarely available in any other way.

Infection

Most forms of systemic microbial infection are associated with high levels of serum CRP and, although the peak values attained in different patients cover a wide range, serial assays in individual subjects usually show an excellent correlation between the serum CRP concentration and the severity of disease and its response to treatment. Acute, systemic, Gram-positive and Gram-negative bacterial infections are among the most potent stimuli for CRP production. Systemic fungal infections occurring in immunodeficient hosts are also associated with high CRP values, whereas the levels in chronic bacterial infections such as tuberculosis and leprosy are usually rather lower, though nevertheless still markedly raised. Uncomplicated viral infections, particularly meningitis, may induce only a very modest response or none at all. Clinical rhinovirus infection (common cold) or influenza are associated with minor CRP elevation in a proportion of individuals, though this may reflect secondary bacterial infection. However, systemic cytomegalovirus or *Herpes simplex* infection of immunosuppressed patients does cause a major CRP response. Little is known about the CRP response to metazoan parasitic infestation in otherwise healthy subjects but malaria, especially *P. falciparum* infection, is associated with high CRP values as are *Pneumocystis* and *Toxoplasma* infections in immunodeficient patients.

Minor or localized, low-grade infection may not stimulate CRP production appreciably but the major CRP response in acute, serious bacterial infection is almost invariable and is present at all ages from premature neonates to the elderly. It also occurs in patients who are immunosuppressed or immunocomprised whether by a primary disease such as leukaemia, lymphoma or other malignancy, or AIDS, or by treatment with cytotoxic drugs, corticosteroids, or irradiation. This is of particular importance in the very young, in the old, in compromised hosts, and in any other patient in whom the usual clinical signs and symptoms of infection, including fever and neutrophil leucocytosis, may be masked or lacking (Fig. 1 and Fig. 2). Furthermore, at the onset of bacterial infection, especially in patients who are otherwise well following elective surgery or myocardial infarction, the CRP response frequently precedes clinical symptoms, including fever, by up to 24 to 48 h.

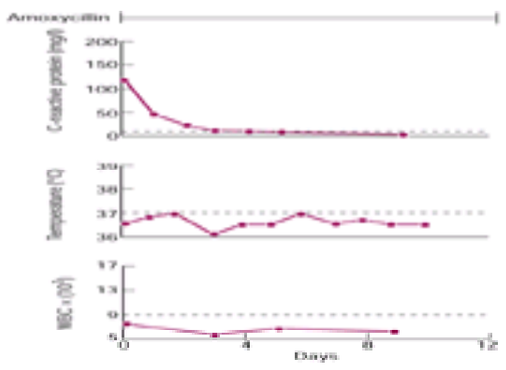

Fig. 1 A 69-year-old diabetic man was admitted with a 3-day history of confusion, cough, and incontinence of urine. There was clinical and radiological evidence of a left-sided pneumonia and although both the temperature and white cell count remained normal, the serum C-reactive protein was high (119 mg/l), confirming the suspicion of infection. Following treatment with amoxicillin, 250 mg thrice daily, the C-reactive protein level fell rapidly, in a characteristic exponential manner, and he made a speedy recovery with return of continence and improved mental state.

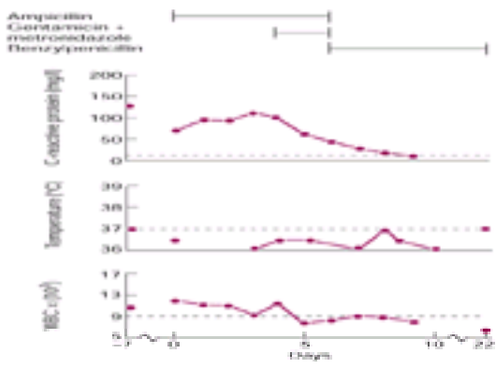

Fig. 2 An 86-year-old woman had been refusing food and drink for 6 weeks. She was dehydrated but rehydration in hospital failed to improve her mental state. She was paranoid and refused nursing and medical care. Paraphrenia was diagnosed and deterioration continued. A C-reactive protein of 130 mg/l and a white cell count of 13.5 × 10^9/l were then found. Chest radiograph, normal on admission, now showed a cavitating lesion from which 150 ml of pus was aspirated. Intravenous ampicillin reduced neither the C-reactive protein nor white cell count prompting a change of therapy to gentamicin and metronidazole. *Strep. equinus* was finally identified in the pus and treatment was changed to benzylpenicillin alone. The C-reactive protein then fell exponentially but rather slowly. The patient's clinical and mental state gradually improved and she was eventually discharged.

Once infection is diagnosed or suspected and antimicrobial treatment has been commenced, frequent monitoring of the serum CRP concentration provides an objective means of assessing the response which is often not available. Effective therapy is associated with a rapid, exponential fall in CRP level, with a half-time of about 24 h, and occurrence of this pattern is an encouraging prognostic sign (Fig. 2). Normalization of the CRP usually corresponds to clinical cure of the infection and may thus be used to determine the necessary duration of antimicrobial therapy. On the other hand, especially in neutropenic or immunodeficient patients, persistent elevation of CRP at the end of a course of antibiotics often presages relapse or recurrence of infection.

When bacterial infection is complicated by abscess formation or for any other reason is less readily eradicated by antimicrobial drugs, the serum CRP concentration may remain elevated or may fall linearly rather than exponentially during treatment. Such a pattern should raise questions regarding dosage of the drugs, sensitivity of the organism, and/or stimulate a diagnostic search both for localized pus and for other underlying, non-infective pathology such as malignancy. Indeed, in the absence of one of the chronic idiopathic inflammatory conditions which are known to be associated with high CRP levels (see below), the persistence of a raised serum CRP concentration is usually a grave prognostic sign, indicating the presence of either uncontrolled infection and/or other serious pathology likely to cause death. However, with alteration in antimicrobial drug regimen or the evacuation of pus or elimination of other pathology, the rapid fall in CRP which may then be observed is a most encouraging objective sign of clinical improvement.

These considerations apply at all ages and regardless of intercurrent pathology, with the exception of severe hepatocellular impairment. In view of the very small amount of serum required for the assay and the speed and precision of automated CRP immunoassays it is apparent that routine monitoring of serum CRP makes a valuable contribution to the recognition and management of infectious diseases. Situations in which these applications have been well documented are listed in Table 3.

Meningitis is of particular interest in view of its potential severity and the importance of rapid diagnosis and appropriate treatment. Bacterial meningitis is associated with much higher serum CRP levels at presentation than cases of aseptic or proven viral meningitis. The latter frequently have CRP concentrations within the normal range or which are only very slightly raised, unless they develop secondary bacterial infective complications, whilst patients with tuberculous have intermediate values. Appropriate therapy for either bacterial or tuberculous meningitis causes the CRP level to fall and this can be used to monitor objectively the response to treatment.

Baseline CRP values are much lower at birth and for the first few days than in older children or adults. Also, neonatal infections progress much more rapidly and can have a fatal outcome before the CRP response has produced concentrations detectable in routine assays. It is therefore essential to use high sensitivity methods capable of detecting and precisely measuring CRP in the range 0.05 to 5.0 mg/l, otherwise the critical initial acute phase response to infection will be missed.

Inflammatory disease

Most of the chronic inflammatory diseases of unknown aetiology (Table 2), with some notable exceptions described below, are associated with high CRP values when they are active. Serial measurements of CRP in individuals with any of these diseases generally reflect the extent and activity of their condition as determined by clinical examination and other laboratory tests. Rheumatoid arthritis is the most common and important disease in this group and the correlation between CRP values in individual patients and the extent and activity of arthritis is very well established. Importantly, there are appreciable differences between the CRP levels attained in different subjects with apparently similar severity of arthritis, but in each case the CRP value always reflects current disease activity. Furthermore, CRP values precisely predict future progression of bone erosion and joint damage. Left unchecked, high CRP levels are inevitably followed by progressive erosive disease, whilst treatments that lower CRP retard or arrest this process.

In some of the inflammatory disorders, for example systemic vasculitis or Crohn's disease (Fig. 3), unlike rheumatoid arthritis, the pathology is relatively inaccessible to direct examination and serum CRP measurement provides the best available, objective index of disease activity. Furthermore, the presence or absence of a CRP response can distinguish between symptoms or organ dysfunction which are due to currently active inflammation or which are the consequence of fibrosis and scarring from previous episodes. This can be very important when treatments include steroids and other powerful and potentially hazardous immunosuppressive, anti-inflammatory, and cytotoxic drugs. It permits precise titration of dosages and may help to avoid excessive or unnecessary use.

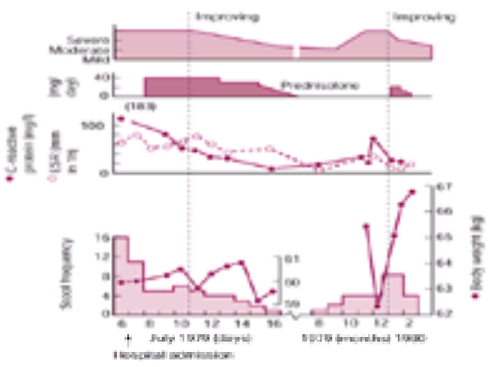

Fig. 3 A 26-year-old man with pancolonic Crohn's disease. He was admitted with severe exacerbation; temperature 38°C; pulse, 110 beats/min; 16 stools per day; haematocrit, 41.5 per cent, leucocytes 13.8 × 10^9/l. Rectal mucosa severely inflamed with histiocytic granulomata on biopsy. Rapid improvement occurred with oral and rectal prednisolone, ampicillin, and metronidazole, with complete clinical and histological remission on day 11. Relapse 5 months later responded promptly to a short course of oral and rectal prednisolone. C-reactive protein and ESR were both high during the initial exacerbation. The rapid response to treatment was paralleled by a prompt fall in C-reactive protein, whereas the ESR responded more slowly. Despite clinical remission and a normal ESR, the C-reactive protein remained slightly elevated, suggesting persistent low-grade inflammatory activity, and it rose further during a subsequent relapse when the ESR did not change. (Reproduced from Fagan A *et al*. (1982). Serum levels of C-reactive protein in Crohn's disease and ulcerative colitis. *European Journal of Clinical Investigation* **12**, 351–9, with permission.)

Induction of clinical remission and control of the underlying disease process is associated with prompt normalization of the CRP. However, CRP also becomes abnormal with intercurrent infection, a common complication of some of these disorders and their treatments, and this serves to focus diagnostic attention often before the infection has become too severe or even before it is clinically evident. Monitoring the CRP response to antimicrobial therapy can then help to confirm the diagnosis and the efficacy of therapy. Persistent elevation of the CRP after eradication of infection may indicate relapse of the underlying inflammatory disease, requiring additional anti-inflammatory treatment.

Necrosis

Myocardial infarction is invariably associated with a major CRP response, as is elective embolization leading to necrosis of tumours in the liver and elsewhere. The peak level of CRP occurs about 50 h after the onset of pain in myocardial infarction and correlates in magnitude, though not in timing, with the peak serum level of cardiac isoenzymes such as creatine kinase MB. In patients who recover uneventfully, the CRP falls rapidly towards normal in the usual exponential fashion. However, complications such as persistent cardiac dysfunction, further infarction, aneurysm formation, intercurrent infection, thromboembolism, or postinfarction syndrome are associated with either persistently raised CRP levels or secondary elevation after the initial decrease. Myocardial rupture is seen only in patients with high peak CRP values, greater than 200 mg/l, and the peak CRP concentration after acute myocardial infarction is inversely correlated with overall outcome, including survival, in the short, medium, and long term.

Stable angina and invasive investigation, such as coronary arteriography, do not stimulate CRP production, whereas some other causes of chest pain, such as pulmonary embolism, pleurisy, or pericarditis, are usually associated with raised CRP levels. Routine assays of CRP after infarction or in patients with chest pain may thus assist in diagnosis and in the recognition and management of complications, including iatrogenic infection associated with invasive cardiovascular monitoring. The role of high-sensitivity measurements of CRP in prediction of coronary heart disease is discussed below.

Serum CRP levels closely reflect the severity and progress of acute pancreatitis, providing a better guide to intra-abdominal events than other markers such as leucocyte counts, erythrocyte sedimentation rate (ESR), temperature, and the plasma concentrations of antiproteinase. A CRP concentration greater than 100 mg/1 at the end of the first week of illness is associated with a more prolonged subsequent course and a higher risk of the development of a pancreatic collection. Serial CRP measurements can therefore guide the use of appropriate imaging techniques and help to confirm resolution before discharge from hospital.

Trauma

The CRP concentration always rises after significant trauma, surgery, or burns, peaking after about 2 days and then falling towards normal with recovery and healing. Infections or other tissue-damaging complications alter this 'normal' pattern of CRP response and the failure of the CRP to continue falling or the appearance of a second peak may precede clinical evidence of intercurrent infection by 1 to 2 days.

Malignancy

Most malignant tumours, especially when they are extensive and metastatic, induce an acute phase response. This is particularly so with those neoplasms which cause systemic symptoms such as fever and weight loss, for example Hodgkin's disease (stage B), and renal carcinoma, but raised CRP levels are seen with many others. In some studies, notably of prostatic carcinoma and bladder carcinoma, the CRP level at presentation has been found to correlate with the overall tumour load and also with the prognosis, being higher for a given mass of tumour in those patients who subsequently fare worse. The CRP may also correlate better with progress and regression of tumour than other, more specific, tumour markers. However, given the non-specific nature of the acute phase response and the limited number of studies performed so far, a definite role for CRP measurements in the management of cancer patients, other than in cases of intercurrent infection, has not yet been established.

Allograft rejection

In the era before routine immunosuppression with combined cyclosporin and steroid treatment, rejection episodes following renal allografting were generally associated with increased production of CRP. However, such treatment almost completely suppresses the CRP response in this situation. In contrast, the acute phase response of SAA is unaffected, and importantly, intercurrent infection still stimulates high levels of both CRP and SAA.

Conditions associated with minor elevation of serum CRP concentrations

Despite unequivocal evidence of active inflammation and/or tissue damage, the conditions listed in Table 4 are usually associated with only minor elevations of the serum CRP concentration, and in many cases it may even remain normal in the face of severe disease. The contrasts between systemic lupus erythematosus (SLE) and rheumatoid and other arthritides shown in Table 2, and between ulcerative colitis and Crohn's disease, are very striking. However, intercurrent microbial infection does provoke a major CRP response in all the conditions shown in Table 4, and this is of great value in diagnosis and management, especially in SLE and leukaemia. The basis of the apparently selective failure of the acute phase response of CRP (which is also shown by SAA) is not known, but presumably involves defect(s) in the pathways which mediate the acute phase response to autologous inflammation and tissue damage. The inbred mouse strain NZB/W, which spontaneously develops antinuclear autoimmune disease, behaves just like human SLE patients with respect to its acute phase responses and the phenomenon may thus be genetically determined. SAP knockout mice also spontaneously develop antinuclear autoimmune disease and do not handle chromatin degradation normally, indicating that pentraxins have a key role in these processes.

Pyrexia is common in SLE and may be caused by microbial infection or by activity of the lupus itself. Both SLE and its treatment predispose to infection, and steroids and immunosuppressives can mask the usual symptoms and signs of infection. Furthermore, infection can trigger exacerbations of SLE. This is a serious clinical situation and infection remains one of the most common causes of death in patients with SLE. CRP values of 60 mg/1 or more are very rare in SLE in the absence of infection whilst levels below 60 mg/1 are seen in patients with documented infection only when it is rather mild and often localized, for example to the skin or lower urinary tract. Differential diagnosis and management of fever in SLE are thus considerably improved by the measurement of serum CRP concentration (Fig. 4).

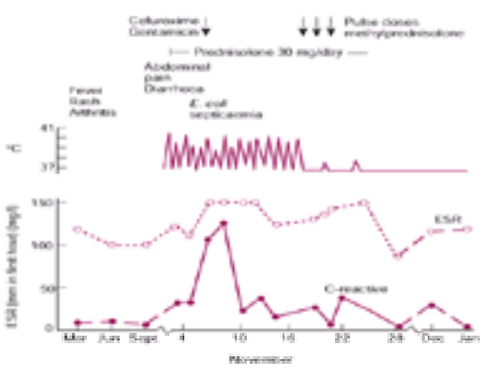

Fig. 4 A 12-year-old girl with a 3-year history of SLE; recurrent febrile episodes, polyarthritis, cutaneous vasculitis, and episodes of asymptomatic bacteriuria. Intermittent treatment was with prednisolone, azathioprine, and plasma exchange. Serum C-reactive protein was only marginally elevated throughout but ESR was persistently raised. Fever recurred with diarrhoea and abdominal pain. All microbial cultures were negative except for growth of *E. coli* from the urine. Despite oral cephalexin and prednisolone, her condition deteriorated with severe neutropenia, probably due to azathioprine. C-reactive protein rose from 36 to 101 and then 137 mg/l, and at this stage her blood culture grew *E. coli*. Intravenous antibiotics were given and the serum C-reactive protein level fell rapidly, but there was little clinical improvement. Active SLE appeared then to be the sole cause of the fever and this was confirmed by the development of a diffuse vasculitic rash and polyarthritis. Three pulse doses of methylprednisolone were given intravenously on successive days and produced a dramatic improvement in her clinical state with resolution of the fever. This case illustrates: (1) the differential response of the C-reactive protein to fever resulting from activity of SLE alone and fever due to bacterial infection; (2) the rapid response of the C-reactive protein both to the onset and to the effective treatment of serious bacterial infection; (3) the failure of ESR measurements to provide any useful information in this complex and rapidly evolving clinical situation. (Reproduced from Pepys MB, Langham JG, de Beer FC (1982). C-reactive protein in SLE. *Clinics in Rheumatic Diseases* **8**, 91–103, with permission.)

The reason why leukaemia patients fail to mount more than a modest CRP response, even during induction therapy when there is massive death of leukaemia cells, is not known, However, they do respond to infection. Since all febrile episodes in leukaemia must initially be treated as infective, the main value of CRP monitoring is to determine the response to therapy and assist in decisions about its duration. Acute or chronic graft-versus-host disease after bone marrow transplantion is usually associated with only a modest CRP response, if any. However, the immunosuppressive treatments used to prevent bone marrow rejection and to control graft-versus-host disease render the patients susceptible to intercurrent infections, often with unusual micro-organisms, and these are always associated with high levels of CRP. CRP monitoring therefore plays a valuable role in management in the post-transplant period.

Interpretation of clinical serum CRP measurements

Clinical measurements of serum CRP fall into two categories. First, routine measurements over the range 3 mg/l upwards in adult and general paediatric medicine. Secondly, high sensitivity measurements including the range up to 3 mg/l that are essential in neonatal medicine and for screening and prognostic investigation in adults with respect to atherothrombotic disease and osteoarthritis. The CRP response is not specific and CRP measurements on their own can never, therefore, be diagnostic of any particular condition. The CRP value can only be interpreted in the light of all other available clinical and laboratory information. Provided this is done it can make a most useful contribution to overall assessment of the patient and determination of the best management.

Routine CRP measurement

The applications fall into three main categories:

- Screening for organic disease
- Monitoring of extent and activity of disease:
  - infection
  - inflammation
  - malignancy
  - necrosis
- Detection and management of intercurrent infection.

Screening for organic disease

CRP production is a very sensitive response to organic disease. A normal CRP therefore eliminates many possible types of pathology and is a reassuring finding. Those serious conditions which only stimulate CRP production weakly if at all, for example SLE, ulcerative colitis, or leukaemia, are all readily recognized by clinical examination and other simple tests such as blood counts, rectal biopsy, or serology. The presence of a raised CRP is unequivocal evidence of active pathology though this may not necessarily be the cause of the complaint for which the patient presented. Such a finding, in the absence of other obvious abnormality, warrants a repeat CRP assay after a few days when a trivial cause such as upper respiratory tract infection will have resolved. Further investigation of a persistently raised CRP level will then depend on the severity of the complaint and other clinical findings.

Monitoring extent and activity of disease

Once the diagnosis is established, in those diseases which cause major elevation of the CRP, serial measurements reflect activity and response to treatment and can be used for monitoring. However, they can only be interpreted provided other possible intercurrent causes of an acute phase response, particularly infections, are excluded.

Detection and management of intercurrent infection

CRP production is a very sensitive response to most forms of infection and a raised level is thus a useful guide to the possible presence of infection in otherwise normal subjects or individuals with a primary condition which predisposes to infection. In disorders which themselves elevate the CRP concentration the decision as to whether infection is present or not must depend on clinical examination and other laboratory tests and the role of CRP testing is then to demonstrate rapidly and

objectively whether there is a response to whatever treatment is used. Effective antimicrobial therapy of infection is always associated with a prompt fall in the CRP whilst persistent CRP elevation indicates continuing infection and/or activity of the underlying disease. There is no other objective test which yields this sort of information so accurately. Changes in results of clinical examination and tests of organ function usually lag hours or days behind the CRP response.

CRP and body temperature

The acute phase response, which is best measured clinically by quantification of the serum CRP, is part of the systemic response of the body to disease. Monitoring of this same response by measurement of body temperature is an integral part of the physical examination and of patient management. CRP production is triggered by the same cytokines which cause fever, and the serum CRP concentration therefore may be considered in part to be a biochemical measurement of the body temperature. However, the CRP response is not susceptible to the many vagaries of thermoregulation itself and routine clinical measurement of body temperature. The precise numerical value of the CRP concentration and its changes over time reflect much more accurately than the temperature the intensity of the underlying stimulus. Furthermore there is often a CRP response in the absence of fever, especially in neonates and the elderly, though also at any age in many chronic inflammatory conditions, and a case therefore exists for the inclusion of a regular serum CRP chart together with the standard temperature chart in appropriate patients.

CRP or ESR?

The only other comparable, non-specific index of the presence of disease which is routinely measured is the erythrocyte sedimentation rate (ESR). The ESR reflects, in part, the intensity of the acute phase response, especially that of fibrinogen and the a-globulins, but is also largely determined by the concentration of immunoglobulins, which are not acute phase reactants. These proteins all have half-lives of days to weeks. The rate of change of the ESR is thus very much slower than that of the CRP level and it rarely reflects precisely the clinical status of the patient at the actual time of testing. ESR is also dependent on the number and morphology of the red cells, which bear no relation to the acute phase response. Finally, there is a significant diurnal variation in ESR, depending on food intake, which is not seen in the CRP. The ESR is therefore of limited use as an objective index of disease activity on which management decisions can be based. The dynamic range of the ESR is also much less than that of CRP and the precision and reproducibility of ESR measurements is poor compared to robust immunoassays for CRP. Thus, in all clinical situations which have been carefully evaluated, ranging from acute bacterial infections to the chronic remittent inflammatory diseases, such as Crohn's disease, rheumatoid arthritis and other inflammatory arthropathies, and systemic vasculitis in its various forms, frequent prospective measurements of CRP reflect disease activity very much more closely than measurements of the ESR. Finally, ESR does not provide the information given by the high sensitivity measurement of CRP, described below. However, the ESR remains a useful screening test for the detection of paraproteinaemias, especially multiple myeloma, which do not necessarily provoke an acute phase response.

High sensitivity CRP measurements

The limitations of conventional immunoassay technology for quantifying serum proteins imposed, until recently, a lower detection limit of about 5 mg/l in most routinely available methods. The advent of more sensitive assays has revealed important new indications for CRP measurement.

Neonatal medicine

Although newborns mount just as rapid and vigorous CRP responses as adults, their baseline values are lower; the median and range in normal cord blood are only 0.04 mg/l and 0.01 to 0.49 mg/l. Since infants can succumb to infection so rapidly, high sensitivity measurements are mandatory.

Atherosclerosis and coronary heart disease

In patients with severe, unstable angina admitted urgently to hospital, CRP values above 3 mg/l are significantly predictive of poor outcome, including death, acute myocardial infarction, or the need for urgent revascularization intervention. In patients undergoing coronary angioplasty, only those with raised CRP values mount an acute phase response to the procedure and the magnitude of this response predicts early reocclusion. Among outpatients with angina and also, remarkably, among healthy normal adult populations, those in the top quintile of the CRP distribution, that is with values above about 2.5 mg/l, have a two to five-fold increased risk of suffering a coronary event in future. Increased CRP production also predicts progression and atherothrombotic events in cerebrovascular and peripheral vascular disease, although the numbers of cases studied have been smaller.

The mechanisms underlying the relationship between even modestly increased CRP production and atherothrombotic events are not known. The CRP response may reflect the inflammation that is a major feature of atherosclerotic plaques, or it may be stimulated by low grade inflammation or infection elsewhere in the body, processes that are known to be associated with atherogenesis. Although CRP levels are associated with smoking, body mass index, hyperlipidaemia, and insulin resistance, all of which are risk factors for atherosclerosis and coronary heart disease, CRP values remain significantly prognostic of coronary events even after adjustment for these variables. It is therefore possible that, in addition to being a marker of inflammation, CRP itself may contribute to pathogenesis, perhaps through its interactions with lipids, lipoproteins, complement, and coagulation. Indeed CRP is detectable within atherosclerotic plaques. There is clearly much work to be done in this exciting new field but, regardless of the underlying mechanisms, the empirical results, from many large-scale independent studies in Europe and the United States, robustly show that high-sensitivity CRP measurements provide important prognostic information.

Osteoarthritis

Modest acute phase responses of CRP, within what was previously considered the normal range, are significantly associated with the presence and extent of osteoarthritis in middle-aged populations. Among affected subjects, the CRP values also predict future progression of the disease.

Serum amyloid A protein

SAA, an apolipoprotein of high-density lipoprotein particles, is a marked acute phase reactant, its concentration rising from normal levels of about 2 mg/l by as much as 1000 times. It is essential to monitor and control SAA levels in patients with reactive systemic, AA type amyloidosis (see Chapter 11.12.4). The other indication for routine SAA measurement is in renal allograft recipients, in whom the SAA response is the most sensitive marker of rejection episodes, despite suppression of the CRP response by immunosuppression with cyclosporin and steroids.

Further reading

Boralessa H *et al*. (1986). C-reactive protein in patients undergoing cardiac surgery. *Anaesthesia* **41**, 11–15.

Danesh J, *et al.* (2000). Low grade inflammation and coronary heart disease: prospective study and updated meta-analyses. *British Medical Journal* **321**, 199–204.

Fagan EA *et al*. (1982). Serum levels of C-reactive protein in Crohn's disease and ulcerative colitis. *European Journal of Clinical Investigation* **12**, 351–60.

Griselli M *et al*. (1999). C-reactive protein and complement are important mediators of tissue damage in acute myocardial infarction. *Journal of Experimental Medicine* **190**, 1733–9.

Hartmann A *et al*. (1997). Serum amyloid A protein is a clinically useful indicator of acute renal allograft rejection. *Nephrology Dialysis, Transplantation* **12**, 161–6.

Haverkate F *et al*. (1997). Production of C-reactive protein and risk of coronary events in stable and unstable angina. *Lancet* **349**, 462–6.

Koenig W *et al*. (1999). C-reactive protein, a sensitive marker of inflammation, predicts future risk of coronary heart disease in initially healthy middle-aged men. Results from the MONICA (Monitoring Trends and Determinants in Cardiovascular Disease) Augsburg Cohort Study 1984 to 1992. *Circulation* **99**, 237–42.

Liuzzo G *et al*. (1994). The prognostic value of C-reactive protein and serum amyloid A protein in severe unstable angina. *New England Journal of Medicine* **331**, 417–24.

Pepys MB, Lanham JG, de Beer FC (1982). C-reactive protein in systemic lupus erythematosus. *Clinics in Rheumatic Diseases* **8**, 91–103.

Ridker PM *et al*. (1997). Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *New England Journal of Medicine* **336**, 973–9.

Spector TD *et al*. (1997). Low level increases in serum C-reactive protein are present in early osteoarthritis of the knee and predict progressive disease. *Arthritis and Rheumatism* **40**, 723–7.

Starke ID *et al*. (1984). Serum C-reactive protein levels in the management of infection in acute leukaemia. *European Journal of Cancer* **20**, 319–25.

Wasunna A *et al*. (1990). C-reactive protein and bacterial infection in preterm infants. *European Journal of Pediatrics* **149**, 424–7.

van Leeuwen MA *et al*. (1997). Individual relationship between progression of radiological damage and the acute phase response in early rheumatoid arthritis. Towards development of a decision support system. *Journal of Rheumatology* **24**, 20–7.

11.12.2 Metabolic responses to accidental and surgical injury

Roderick A. Little

Introduction

Accidental injury remains the principal cause of death and disability in those aged less than 45 years of age in the developed world. Also, as the threats of infectious disease and starvation have been reduced in parts of the 'third world', they have been replaced by injury; at the same time, motor vehicle and interpersonal violence have increased. Since such injuries involve predominantly the young, it can be calculated that in the loss to society of productive years of life, injury presents more of a challenge than cancer and cardiovascular disease.

Humans are also subjected to the planned injury of surgery, recovery from which may be prolonged. This consumes health service resources within and outside hospital and also delays return to employment and/or independence.

Attempts to reduce the burden of injury should emphasize prevention but it is also important to lessen its biological effects. A full description and explanation of the biological responses to injury is needed so that treatment can be properly directed. For example now that the metabolic responses to major injury, surgery, and infection have been shown to be very similar, 'generic' scientifically based measures rather than empirical *ad hoc* interventions are used in treatment.

The most important advance in the description of the metabolic response to injury was the recognition, by Sir David Cuthbertson in 1932, that the response did not involve unconnected reactions but rather followed an ordered pattern. He divided the response into an initial transient 'ebb', followed by a prolonged 'flow' phase.

Ebb phase

This was first described as a period of depressed vitality or metabolism and anuria which occurred during the first 1 to 2 days after injury. Subsequent analysis of the data from Cuthbertson's four patients reveals little evidence for a reduction in energy metabolism. However, his interpretation may have been influenced by previous experiments which showed an acute depression of metabolism after hind-limb muscle injury in anaesthetized cats. In these animals, the depression in metabolism was caused by a failure of tissue oxygen delivery—an unlikely occurrence in Cuthbertson's patients with single, lower-limb fractures.

It now seems that the ebb phase can be redefined as the early stage after trauma during which tissue energy production is not limited by a failure of oxygen delivery (Table 1). It is a complex neuroendocrine response, the magnitude of which depends on the generation of somatic afferent nociceptive stimuli arising from damaged tissues and the loss of intravascular volume. As well as eliciting a neurohumoral response, the nociceptive stimuli also inhibit central thermoregulation and cardiovascular reflex activity. Many aspects of this early response are reminiscent of the alerting response of the defence reaction or preparation for fight or flight. Thus, the arterial baroreceptor reflex is inhibited allowing concomitant increases in arterial blood pressure and heart rate.

In this initial phase, there is also mobilization of energy stores to fuel the anticipated increase in activity. The increases in plasma glucose concentration, which are directly related to the severity of injury, arise from the breakdown of glycogen stores in the liver and skeletal muscle. This breakdown is mediated by rises in the counter-regulatory hormones (the catecholamines, especially adrenaline, cortisol, and glucagon) which occur at this time, together with increases in antidiuretic hormone. The catecholamines may also be sufficient in the severely injured to inhibit intracellular glucose uptake and oxidation by suppressing the release of insulin. The increased activity of the sympathetic nervous system also stimulates lipolysis but the relationship between plasma concentrations of non-esterified fatty acids and severity of injury is complex. In particular, there may be stimulation of fatty acid re-esterification within adipose tissue by the raised plasma lactate that is associated with severe injury or impaired perfusion of the fat depots.

Although the main changes in protein metabolism are associated with the 'flow' phase, the acute phase plasma protein response is initiated early after injury. Hence, 6 h or so after tissue damage, plasma concentrations of the acute phase reactants (e.g. C-reactive protein and fibrinogen) rise. These rises are due to an increase in their hepatic synthesis, probably induced by interleukin-6 (IL-6). Plasma albumin concentrations often fall rapidly after injury, as a result of increased microvascular permeability rather than to a reduction in synthesis.

The role of the cytokines in mediation of the acute responses to injury remains controversial. While cytokines have been shown to mimic some of these responses, levels of plasma cytokines are not universally elevated immediately after injury. This may, of course, reflect their autocrine/paracrine rather than endocrine function. However, elevated plasma concentrations of the proinflammatory cytokines tumour necrosis factor-a (TNF-a), IL-b, IL-2, IL-6, and IL-8 occur after accidental injury. Also IL-6 levels have been directly related both to the severity of multiple trauma and to the magnitude of surgical trauma. IL-6 may also be involved in the activation of the hypothalamo–pituitary–adrenal axis by peripheral tissue injury, probably by means of the induction of cyclo-oxygenase products at the blood–brain interface (e.g. the circumventricular organs).

If the injury is not overwhelming, and with appropriate control of the airway, breathing, and circulation, the 'ebb' phase gives way to the 'flow' phase.

Flow phase

The main metabolic features of the flow phase are an increase in metabolic rate (hypermetabolism), due to catabolism (especially of skeletal muscle) and resistance to the anabolic effects of insulin (Table 2).

Hypermetabolism

There is said to be an increase in metabolic rate which is directly related to the severity of injury. Although such a relationship has been described for burned patients treated by exposure, it is not so clear after other injuries. One of the problems is that to describe a patient as hypermetabolic it is necessary that we define a normal metabolic rate for that individual; but what is a normal rate for someone who might be paralysed and ventilated, bedridden for a considerable time, receiving inadequate nutrition, and has suffered an acute reduction in body mass? All of these factors can reduce metabolic rate and thereby counteract any hypermetabolic stimuli associated with injury. It is also possible that in the most critically ill, energy expenditure may be limited by a failure of oxygen delivery. This may be the explanation for the finding of 'normal' metabolic rates in critically ill, septic patients.

What are the hypermetabolic stimuli alluded to above? There is an upward resetting of metabolic rate, perhaps triggered by a cytokine/prostanoid cascade, and there are increases in efferent sympathetic activity and substrate cycling (a process in which ATP is consumed without concomitant change in the amount of substrate or its metabolic products). Also, there is an extra organ—the wound—which contributes to whole body energy metabolism in two main ways. First, it can be the site of origin of the cytokine/prostanoid cascade mentioned above and, second, the wound is a consumer of glucose which it converts to lactate by aerobic glycolysis (an inefficient producer of ATP). The lactate produced is reconverted to glucose in the liver in an energy consuming process, thereby increasing hepatic oxygen consumption and blood flow. The evaporation of water from a burn wound or from an area of granulation tissue will also increase energy expenditure. The catabolism of lean tissue protein can also increase energy expenditure, but only in those with severe head injuries or very extensive burns is it likely to contribute more than 20 per cent to the total expenditure.

Protein metabolism

Whole body protein turnover is increased after injury with the balance between synthesis and breakdown being modified by the severity of injury and the influence of nutritional intake on synthesis. Thus, the severity of the injury increases both protein synthesis and breakdown. However, after the most severe injuries, the increase in breakdown predominates and cannot be counteracted by even the most aggressive nutritional support. It is now recognized that the increase in proteolysis in such catabolic conditions involves the ubiquitin–proteasome pathway. The increase in muscle protein catabolism is reflected by concomitant increases in the urinary excretion of nitrogen, 3-methylhistidine and, creatinine.

The most obvious site of net protein catabolism is skeletal muscle—although it also occurs in the respiratory muscles, the wall of the gut, and possibly the heart. Thus, in addition to problems with mobility, ventilation and the maintenance/restoration of enteral nutrition can also be compromised.

The increase in proteolysis provides amino acids as precursors for hepatic gluconeogenesis. The persistence of this glucose production at a time when plasma glucose and insulin concentrations are normal or raised can be considered as another facet of insulin resistance (see below). Plasma concentrations of several amino acids can fall at this time although their hepatic uptake is maintained by the increase in hepatic blood flow (see above). One amino acid that has received a lot of attention is glutamine, the intracellular concentration of which falls in skeletal muscle in response to injury. Glutamine is an important fuel for cells of the immune system, it is a precursor for the synthesis of glutathione (a free radical scavenger), has a role in nitric oxide metabolism, and has also been implicated in the maintenance of the integrity of the gut mucosal barrier, which may be compromised after injury.

Insulin resistance

Resistance to the 'anabolic' effects of insulin after injury is manifest in a several ways. For example hepatic glucose production, lipolysis, and net efflux of amino acids from skeletal muscle persist at plasma glucose and insulin concentrations which are inhibitory in uninjured subjects. Also the uptake of glucose into skeletal muscle is reduced, an impairment which involves glucose storage rather than oxidation.

The cause of injury-induced insulin resistance is unclear, although a role for the counter-regulatory hormones (cortisol, adrenaline, and glucagon) has been suggested. Whilst infusion of these hormones can induce insulin resistance in healthy individuals, the plasma concentrations needed are higher than those in injured/septic patients who are known to be insulin resistant. In this respect, the proinflammatory cytokines may have a role in modulating insulin sensitivity.

Plasma IL-6 concentrations in patients with cancer are positively correlated with the degree of insulin resistance and TNF-a has been implicated in the insulin resistance of diabetes and obesity. A role for IL-1 in the insulin resistance associated with endotoxaemia has also been suggested. The nature of the intracellular defect underlying insulin resistance remains to be elucidated but it may involve serine phosphorylation of insulin receptor substrate-1.

Manipulation of the metabolic response

Most efforts have been focused on trying to attenuate or reverse the loss of muscle mass. The administration of glutamine and anabolic agents such as growth hormone, insulin like growth factor (IGF), and oxandrolone have all been investigated.

Since glutamine plays such a central role in host defence mechanisms, there have been several studies of glutamine supplementation in, for example, surgical patients. These have shown preservation of intestinal mucosal integrity, enhanced immune function, and attenuation of the decrease in muscle intracellular glutamine concentration and protein synthesis. There have been fewer studies in critically ill patients, although one trial has shown that glutamine-supplemented parenteral nutrition improved 6-month survival.

The use of recombinant human growth hormone (rhGH) has produced conflicting results. Improvements in nitrogen balance have been reported in surgical, burned, and injured patients, and in the surgical group the improvement was accompanied by preservation of handgrip strength. In burned children, rhGH reduced the length of hospital stay, probably reflecting improved wound healing. However, other studies have failed to demonstrate an improvement in nitrogen balance and attenuation of muscle breakdown. The administration of IGF-1 in studies in animals and human volunteers were encouraging, but several randomized, controlled studies in catabolic patients have failed to demonstrate any protein-sparing effect.

Oxandrolone is a testosterone analogue with anabolic activity that has been shown to be of benefit to patients with severe malnutrition and alcoholic hepatitis. A trial in patients with severe burns has shown that oxandrolone combined with an increased protein intake significantly increased the rate of weight gain and improved subjective measures of muscle strength.

Further reading

Aub JC (1920). Studies in experimental traumatic shock I. The basal metabolism. *American Journal of Physiology* **54**, 388–407.

Barton RN, Frayn KN, Little R A (1990). Trauma, burns and surgery. In: Cohen RD *et al*., eds. *The metabolic and molecular basis of acquired disease*, pp. 684–717. Bailliere Tindall, London.

Cuthbertson DP (1942). Post-shock metabolic response. *Lancet* **1**, 433–7.

Frayn KN *et al*. (1985). The relationship of plasma catecholamines to acute metabolic and hormonal responses to injury in man. *Circulatory Shock* **16**, 229–40.

Garlick PJ, Wernerman J (1997). Protein metabolism in injury. In: Cooper GJ *et al*., eds. *Scientific foundations of trauma*, pp. 690–728. Butterworth-Heinemann, Oxford.

Girolami A, Foex BA, Little RA (1999). Changes in the causes of trauma in the last 20 years. *Trauma* **1**, 3–11.

Griffiths RD, Hinds CJ, Little RA (1999). Manipulating the metabolic response to injury. *British Medical Bulletin* **55**, 181–95.

Little RA (1985). Heat production after injury. *British Medical Bulletin* **41**, 226–31.

Little RA, Kirkman E (1997). Cardiovascular control after injury. In: Cooper GJ *et al*., eds. *Scientific foundations of trauma*, pp. 551–63. Butterworth-Heinemann, Oxford.

Turnbull AV, Rivier CL (1999). Regulation of the hypothalamic-pituitary-adrenal axis by cytokines: Actions and mechanisms of action. *Physiological Reviews* **70**, 1–71.

Wilmore DW, Stoner HB (1997). The wound-organ. In: Cooper GJ *et al*., eds. *Scientific foundations of trauma*, pp. 524–9. Butterworth-Heinemann, Oxford.

11.12.3 Familial mediterranean fever and other inherited periodic fever syndromes

P. N. Hawkins and D. R. Booth

The hereditary periodic fever syndromes are a group of multisystem disorders characterized by recurrent episodes of fever in association with inflammation that variably affects serosal linings, joints, and skin. They include familial Mediterranean fever, which is by far the most common, familial Hibernian fever, the hyperimmunoglobulin D syndrome, the Muckle–Wells syndrome, and familial cold urticaria. Although some of the individual features of these diseases overlap, there are distinctions in their mode of inheritance, their clinical features, and the frequency of symptoms. There has been substantial progress in elucidating the molecular basis of the hereditary periodic fever syndromes, which has proved to be surprisingly diverse and has opened several new avenues in inflammation research. Molecular analysis has already had a major impact on clinical diagnosis and the genetic defect in familial Hibernian fever has suggested a specific form of treatment. All of these disorders are compatible with normal life expectancy but since each of them is associated with a prominent, acute phase plasma protein response, potentially fatal systemic AA amyloidosis may develop unless the underlying inflammatory condition can be suppressed.

Familial Mediterranean fever

Familial Mediterranean fever (FMF) is an inherited, inflammatory disease that occurs most commonly in Jewish, Armenian, Turkish, and Middle Eastern Arab populations but also rarely in individuals of any ancestry. It is characterized by recurrent, acute attacks of fever, sterile peritonitis and pleurisy, arthritis, and erysipelas-like rashes lasting from 12 h to about 3 days, which in most cases can be prevented by regular prophylactic therapy with colchicine. The disorder is also known as recurrent polyserositis, periodic disease, Armenian disease, and other names besides, but the term familial Mediterranean fever is most commonly used. The recently identified gene responsible has been called *MEFV*, (Mediterranean fever), is located on chromosome 16 and appears to be expressed only in neutrophils. It encodes a hitherto uncharacterized protein called pyrin or marenostrin, and its discovery should enable the molecular basis of the disease to be unravelled. Analysis of *MEFV* has already revolutionized the diagnosis of FMF. About 30 mutations in *MEFV* have now been associated with FMF, and pairs of *MEFV* mutations, presumed to involve both alleles, can be found in most FMF patients. These findings accord with the autosomal recessive mode of inheritance that is usually evident clinically and which has long been recognized in population studies.

Pathogenesis and genetics

Pyrin, the uncharacterized protein product of *MEFV*, consists of 781 amino acids with a molecular weight of about 90 kDa. Messenger RNA for pyrin is found exclusively in neutrophils and their precursor cells. The amino acid sequence of pyrin contains zinc finger motifs, potential phosphorylation sites, and a B30.2 domain, all typical of several other proteins that act within the nucleus. Pyrin could therefore be a transcription factor and, if so, it could suppress the production of a proinflammatory molecule, or upregulate the transcription of an anti-inflammatory one. However, the same features are also found in some proteins acting outside the nucleus and which are involved in protein–protein interactions. It is likely that the *MEFV* mutations which cause FMF disrupt the structure of pyrin sufficiently to reduce its function, leading to inappropriate neutrophil activation and migration. The recurrent, acute clinical attacks, along with massive influx of neutrophils into serosal and synovial linings, would be consistent with bursts of relatively uncontrolled neutrophil activity. Patients with FMF often have prolonged periods of apparently normal health between their attacks, suggesting that reduced pyrin function has clinical consequences only in certain circumstances. Physical and emotional stress, menstruation, and diet have been reported to increase susceptibility to FMF attacks and the irregular nature of symptoms supports the notion that clinical attacks of FMF may be triggered by exogenous factors. The characteristic pattern of inflammation in serosal and synovial membranes, and the paucity of evidence in patients with FMF for exacerbation of other inflammatory processes in which neutrophils are involved, suggest that pyrin may regulate a rather specific facet of neutrophil behaviour.

The 30 or so *MEFV* mutations that are associated with FMF encode either single amino acid substitutions or deletions in the pyrin molecule. The mutations that commonly cause FMF are in exon 10 of the pyrin gene, and in exon 5 in some populations. Although a polymorphism in exon 2 encoding the pyrin variant E148Q may contribute to FMF in some cases, homozygosity for E148Q does not seem to be sufficient to cause FMF. Whilst it is inherently likely that different mutations will impair the function of a protein to differing extents, several findings indicate that the methionine residue at position 694 may be critical for the normal function of pyrin. Three different pathogenic exon 10 mutations involving M694 have been identified (M694V, M694I, and deletion M694) and individuals who are homozygous for M694V are reported to have particularly severe and early-onset FMF disease, as well as a greater propensity to develop AA amyloidosis. It is also noteworthy that simple heterozygous deletion of residue M694, which is likely to produce more severe structural disruption than an amino acid substitution, has been reported, by itself, to cause typical FMF in several British families. Recognition that *MEFV* mutations affecting a single allele may give rise to FMF suggests that a 50 per cent complement of normally functioning pyrin is not sufficient to prevent susceptibility to the disease. Severe disruption of a single *MEFV* allele by one or more mutations may therefore account for the rare reports of autosomal dominant FMF. However, transmission of FMF is pseudodominant in most families in which more than one successive generation is affected by the disease and this usually reflects consanguinity or a high incidence of the FMF trait in the population at risk. Up to one individual in seven in some Mediterranean regions is an FMF carrier, and the apparently 'mild' pyrin variant E148Q occurs very widely indeed. Pyrin E148Q has been identified in European whites, black Africans, Punjabi Indians, and Chinese populations, with an allele frequency of about 20 per cent in the latter two groups. The remarkable global prevalence of pyrin E148Q supports the hypothesis that the FMF trait conferred a survival benefit during evolution. Possible mechanisms for this include enhanced resistance to microbial infection mediated by non-specific upregulation of the inflammatory response. It is worth considering that the same process may have the potential to exacerbate chronic inflammatory disorders.

Clinical features

Patients with FMF typically present with acute attacks of fever, peritonitic abdominal pain, pleuritic chest pain, arthralgia, and rash in approximately this order of frequency. However, the pattern of tissue involvement varies substantially between patients, and even in the same patient on different occasions. Other features occasionally include myalgia, headache, orchitis, and cutaneous and renal vasculitis. The symptoms of FMF usually start in the first decade of life, and only five per cent of patients develop symptoms after 30 years of age, although the frequency and severity of attacks can alter substantially during the course of a patient's life.

A brief prodrome may occur after which the temperature rises to 38 to 40°C in almost every case. Peritonitis occurs in more than 90 per cent of cases and evolves over a few hours. It is readily misdiagnosed as acute appendicitis or some other surgical crisis. Many patients become confined to bed with severe pain and vomiting but these features can be quite mild or absent in other cases. Diarrhoea is unusual. Pleurisy occurs in about 50 per cent of patients, is typically unilateral, and can be associated with atelectasis or a small effusion. Involvement of joints also occurs in about half of the patients, and can range from polyarthralgia without overt soft tissue swelling to a more classical monoarticular synovitis affecting the knees or hips. The rash of FMF is less common but comprises very characteristic 10 to 15-cm, erysipelas-like, warm swollen lesions on the lower leg. Pericarditis, perhaps curiously, is rare. The acute crisis of FMF usually resolves in less than 3 days, and attacks tend to recur at irregular intervals of weeks or months. Direct long-term sequelae of FMF, such as serosal fibrosis and adhesions, occur rarely although a small proportion of patients do develop a chronic inflammatory arthritis that can lead to severe joint destruction. The most important and life-threatening long term consequence of FMF is AA amyloidosis (see below).

Investigations

No laboratory test is specific for FMF. The identification of *MEFV* clearly represents a major advance both in elucidating the molecular basis of the disease and for its diagnosis but the results of genetic testing in any individual must be interpreted with care. This is essential because individuals with *MEFV* mutations seem to differ

greatly in their susceptibility to developing FMF and many have subclinical inflammatory disease. They may also have some other, unrelated inflammatory disease that show symptom overlaps with FMF. Another practical problem is that *MEFV* is a relatively large gene spanning 10 exons, and limited screening for common mutations will fail to identify abnormalities in many patients, particularly in those with atypical ethnic backgrounds in whom rare or new mutations are more likely. Certainly, in the absence of an exon 10 mutation, the diagnosis of FMF is unlikely in patients from ethnic groups in which FMF typically occurs. However, even the most comprehensive analysis of the *MEFV* coding region has identified only single allele mutations in some patients with classical FMF and no mutations at all in a few cases. Thus, mutations in regulatory regions of *MEFV*, or indeed mutations in other genes, might contribute to the pathogenesis of the disease in some patients; alternatively, certain heterozygotes may be unusually susceptible to FMF. The results of *MEFV* genotyping should therefore be interpreted in light of each clinical presentation and ethnic origin.

Diagnostic clinical criteria for the evaluation of patients with features suggestive of FMF are well developed. During the clinical attacks of FMF, the number of neutrophils in peripheral blood may rise from normal levels by several fold. Some patients have a mild normochromic anaemia and polyclonal hyperglobulinaemia but the most striking serological feature of the disease is the magnitude of the acute phase plasma protein response that is evoked. During a typical clinical attack, the concentration of C-reactive protein generally exceeds 100 mg/l and may be greater than 250 mg/l; the plasma concentration of serum amyloid A protein often attains even higher levels, frequently exceeding 500 mg/l (normal less than 3 mg/l). Both C-reactive protein and serum amyloid A protein have half-lives in the circulation of less than 24 h, and their concentrations fall to healthy baseline values within several days of resolution of an acute episode. Serial monitoring of these sensitive and dynamic markers have shown that bursts of inflammation occur frequently in patients with FMF in the absence of symptoms, and even, to a lesser extent, among those taking regular colchicine therapy. Our own practice in a patient with suspected FMF is to measure the concentration of C-reactive protein and serum amyloid A protein on four or five occasions at 2-weekly intervals. In many cases, this will objectively document self-limiting 'periodic' bursts of inflammation. It is often useful to repeat this type of analysis during a therapeutic trial of colchicine.

Amyloidosis

AA amyloidosis is the most significant complication of FMF since it is unpredictable, progressive and potentially fatal. AA amyloid fibrils are derived from a 76 amino acid N-terminal cleavage fragment of the 104 residue serum amyloid A protein. Before the beneficial effect of colchicine in FMF was known, the lifetime incidence of AA amyloidosis was reported to be up to 30 per cent or more. The risk of developing amyloidosis appears to be greater among Sephardic Jews and Turks than among non-Sephardic Jews and Armenians. Some patients with FMF present with AA amyloidosis before they have experienced symptoms of the inflammatory disease, a situation often referred to as (pheno)type II FMF. The lack of a clear association between the severity of symptoms in FMF and the likelihood of developing AA amyloidosis, and the different incidence of amyloidosis among various ethnic groups has led some investigators to question the nature of the relationship. However, the features of AA amyloid in patients with FMF are identical to those in patients with every other type of chronic inflammatory disease, ranging from the composition of the amyloid fibril protein to the distribution and clinical consequences of the amyloid deposits. Although there are evidently unknown genetic or environmental factors that influence susceptibility to AA amyloidosis generally, acute phase production of serum amyloid A protein is the only absolute prerequisite for AA amyloid deposition. Systematic serial monitoring of serum amyloid A protein has confirmed that a major acute phase response frequently occurs in patients with FMF in the absence of symptoms, and the magnitude and duration of this is probably a very substantial determinant of the risk of developing amyloidosis. Evaluation of this risk is likely to become more refined when the significance of different *MEFV* mutations, serum amyloid A protein isotypes, and other factors is better understood, but for the time being all patients with FMF who have had uncontrolled inflammation, with or without symptoms, must be regarded as susceptible. *MEFV* analysis in patients with FMF complicated by AA amyloidosis, who have been evaluated in our own clinic, has demonstrated a wide spectrum and combination of mutations suggesting that amyloidosis is not restricted to any particular genotype.

AA amyloid deposition may be extensive without causing symptoms and can develop at any time from early childhood to late adult life. AA amyloid initially accumulates in the spleen, and functional hyposplenism may eventually develop. However, the most common mode of presentation is with non-selective proteinuria, nephrotic syndrome, and/or renal insufficiency. Acute renal failure may be precipitated by minor insults, and end-stage renal failure and its complications are the most frequent cause of death. Patients sometimes present with hepatosplenomegaly, occasionally without overt renal dysfunction, but renal deposits are nevertheless always present in such cases. The adrenal glands are involved in at least one-third of cases and the liver in a quarter. Although the function of these latter organs is often well preserved, hepatic amyloidosis is a sign of advanced and extensive disease and has a poor prognosis. Histological involvement of the gut is common, but tends only to cause symptoms in patients whose disease is generally very advanced. AA amyloidosis rarely ever causes overt cardiac disease. The prognosis is chiefly determined by the extent of amyloid at diagnosis and the effectiveness with which production of serum amyloid A protein can be suppressed by colchicine therapy. Proteinuria should be sought routinely in patients with FMF, and the suspicion of amyloid followed-up by renal or rectal biopsy and Congo-red staining. Serum amyloid P component scintigraphy is a sensitive and specific non-invasive method for imaging amyloid in visceral organs and can be used serially to monitor the deposits in a quantitative manner. Scintigraphic follow-up studies in more than 100 patients with AA amyloidosis, some of whom had FMF, have confirmed the findings of numerous clinical and histological case reports by demonstrating that the amyloid deposits frequently regress when production of serum amyloid A protein is reduced to normal healthy baseline levels. Routine monitoring of serum amyloid A protein should be an integral part of the management of all patients with AA amyloid and automated immunoassay systems for serum amyloid A protein are available, standardized on a World Health Organization International Reference Standard.

Treatment

The only effective treatment for FMF is colchicine, a serendipitous discovery made by Goldfinger in 1972. Regular prophylactic treatment with colchicine at a dose of 1 to 2 mg daily prevents or substantially reduces the clinical manifestations of FMF in at least 95 per cent of cases. A proportion of the remainder are poorly compliant or intolerant of the drug. Studies in which serum amyloid A protein has been monitored in FMF patients who have been rendered free of symptoms by colchicine therapy show that the subclinical inflammation is also substantially, but often not completely, prevented. Colchicine modulates neutrophil function by binding to tubulin in microtubules which inhibits motility and exocytosis of the intracellular granules and diminishes neutrophil chemotaxis *in vitro* and *in vivo*. However, colchicine is not generally effective in other acute or chronic inflammatory diseases, even in those in which neutrophils are recognized to have an important role. The mechanism by which colchicine exerts its beneficial effect in FMF, and how, if at all, it influences the pyrin pathway are not yet known.

Regular, long-term use of colchicine is advisable in every patient with FMF and mandatory in those with AA amyloidosis. Not only does this prevent amyloidosis from developing in the first place but established amyloid deposition will usually be halted by an adequate dose of the drug. Existing amyloid deposits may gradually regress. The clinical effects of amyloid may also resolve, particularly proteinuria in patients whose glomerular filtration is well preserved. FMF patients with end-stage renal failure due to AA amyloidosis are often excellent candidates for renal transplantation, since graft amyloid can also be prevented by colchicine prophylaxis. Although colchicine is an extremely toxic agent in large quantities, the small regular dose required for the treatment of FMF is generally well tolerated. The most frequent adverse effect of low-dose colchicine is diarrhoea, but this seldom prevents its use and sometimes responds to a lactose-free diet. Despite concerns about the antimitotic potential of colchicine, the drug does not appear to cause infertility or lead to birth defects, even when used throughout pregnancy, or have any other long-term adverse effect. The concentration of colchicine in breast milk is sufficiently low to permit breast feeding.

Colchicine is largely ineffective in the acute management of FMF attacks. Analgesics and intravenous fluid replacement are required in some patients but a low threshold for intravenous fluids should be the policy for FMF patients with AA amyloidosis since hypovolaemia can precipitate irreversible renal failure in these patients.

Familial Hibernian fever (tumour necrosis factor receptor associated periodic syndrome—TRAPS)

The term familial Hibernian fever was coined by Williamson in 1982, who described a large Irish family with a periodic fever syndrome inherited in an autosomal dominant manner. Affected individuals had intermittent bouts of fever, abdominal pain, painful erythematous rashes, arthralgia, and myalgia. Flares of the disease are very much less distinct than in FMF and patients frequently remain symptomatic for weeks or sometimes months on end. The clinical manifestations of inflammation are associated with a very intense acute phase response which can be sustained even when symptoms abate. Several affected individuals have developed AA amyloidosis. Only a handful of other families, most of Northern European ancestry, have been reported with similar syndromes but its genetic basis has recently been discovered. This is not only of major diagnostic value in autosomal dominant periodic fever syndromes, but it has also identified a novel mechanism of inflammation that suggests a specific therapeutic approach.

The disease is caused by at least 16 different missense mutations in the gene on chromosome 12 which encodes tumour necrosis factor receptor 1 (TNFRSF1A). Tumour necrosis factor (TNF) is a potent proinflammatory cytokine that is implicated very widely in inflammation. TNFRSF1A is expressed on the outer membrane of most cells in the body where it acts closely with TNF receptor 2 to bind TNF and lymphotoxin cytokines and signal the activation of a complex inflammatory pathway. This signalling can be turned off by enzymatic cleavage and shedding of the soluble, extracellular portion of TNFRSF1A from the cell surface into the blood. This has a dual effect since it disrupts the signalling mechanism and the soluble TNFRSF1A that is released can bind and neutralize TNF in the circulation. Many of the

mutations that have been identified in autosomal dominant periodic fever syndromes are thought to disrupt conserved disulphide bonds in the extracellular TNFRSF1A domain and interfere with the physiological negative feedback control mechanism. The plasma concentration of soluble TNFRSF1A is abnormally low in many patients with these diseases, and measurement of soluble TNFRSF1A (also known as p55 protein) may be a useful way of screening for the disorder. It is now widely accepted that the dominantly inherited fever syndromes associated with TNF receptor mutations should be classified as tumour necrosis factor receptor associated periodic syndromes (TRAPS).

This disorder responds partially to colchicine in some cases but can usually be suppressed by high doses of corticosteroids and other immunosuppressive agents. The genetic basis of the disorder suggests that specific blockade of TNF might be beneficial, and early experience with this approach has been promising.

Hyperimmunoglobulinaemia D periodic fever syndrome

The hyperimmunoglobulinaemia D periodic fever syndrome (HIDS) was first described in the Netherlands in 1984. It is inherited in an autosomal recessive manner and is characterized by intermittent, irregular attacks of abdominal pain and fever in association with an acute phase response and persistent elevation of the plasma concentration of immunoglobulin D. Attacks characteristically occur every 1 or 2 months and last for 3 to 7 days, but with considerable individual variation. Unlike FMF, the abdominal pain does not usually have the overt features of peritonitis, and is frequently associated with diarrhoea as well as vomiting. Lymphadenopathy occurs in most patients and arthralgia affecting the knees and ankles is present in about two-thirds of cases. Most families are of western European origin. The acute phase response in HIDS tends to be less intense than in the other inherited periodic fever syndromes described here, and this is likely to be the reason for an apparently very low risk of developing AA amyloidosis.

HIDS has lately been attributed to several mutations affecting both alleles of the gene on chromosome 12 encoding mevalonate kinase, *MVK*. This enzyme is involved in isoprenoid synthesis and catalyses the phosphorylation of mevalonate to 5-phosphomevalonate. The identified mutations have been shown to reduce its enzymatic activity. Different mutations in *MVK* are the cause of mevalonic aciduria, another inherited and much more severe multisystem disease. Excretion of mevalonic acid in the urine is increased constitutively in mevalonic aciduria, but only during the febrile episodes, if at all, in HIDS. The pathogenesis of these disorders, including the dysregulation of immunoglobulin D production, is not understood and, as in FMF, a previously unknown pathway that can profoundly regulate an aspect of the inflammatory response may be involved.

Muckle–Wells syndrome and familial cold urticaria

Muckle–Wells syndrome and familial cold urticaria are very rare, dominantly inherited, periodic fever syndromes. Fever and rash are prominent features of both conditions, and both can be complicated by AA amyloidosis. The original family described by Muckle and Wells in 1962 had progressive sensorineural deafness, although this does not occur in other cases. Arthralgia and poorly characterized limb pain occur in both disorders, and both conditions tend to start in early childhood. The skin rash, fever, and other symptoms in familial cold urticaria are triggered by exposure to cold. The frequency and pattern of symptoms is extremely variable but they can be present for very prolonged periods in both disorders. The search for the genetic basis of these two disorders, which evidently have a number of overlapping clinical features, has lately linked both of them to chromosome 1q44. It is therefore possible that Muckle–Wells syndrome and familial cold urticaria represent different phenotypic manifestations of the same mutation, or that they might be due to different mutations in the same gene. Phenotypic expression of familial cold urticaria can presumably be influenced by climatic conditions. We have lately evaluated a family with this syndrome that resides in a warm part of India, perhaps accounting for the very mild symptoms in most of the affected members. Treatment for these disorders centres mainly on corticosteroids and immunosuppressive drugs. A therapeutic trial of colchicine is, however, worthwhile, since it seems to be quite effective in a proportion of cases. Most patients with these syndromes have a normal life expectancy, although in the absence of a generally effective treatment, the outlook is poor for individuals who develop AA amyloidosis.

Further reading

Drenth JPH *et al*. (1999). Mutations in the gene encoding mevalonate kinase cause hyper-IgD and periodic fever syndrome. *Nature Genetics* **22**, 178–81.

McDermott MF (1999). Autosomal dominant recurrent fevers. Clinical and genetic aspects. *Revue du Rheumatisme* **66**, 484–91.

Samuels J *et al*. (1998). Familial Mediterranean fever at the millennium. Clinical spectrum, ancient mutations, and a survey of 100 American referrals to the National Institutes of Health. *Medicine* **77**, 268–97.

11.12.4 Amyloidosis

M. B. Pepys and P. N. Hawkins

Introduction

Amyloidosis is a disorder of protein folding, characterized by extracellular deposition of abnormal protein fibrils. The underlying molecular abnormalities may be either acquired or hereditary and about 20 different proteins can form clinically or pathologically significant amyloid fibrils *in vivo* (Table 1 and Table 2). Amyloid deposits also contain glycosaminoglycans, some of which are tightly associated with the fibrils, and also a non-fibrillar plasma glycoprotein, amyloid P component. Small, focal, clinically silent amyloid deposits in the brain, heart, seminal vesicles, and joints are a universal accompaniment of ageing. However, clinically important amyloid deposits usually accumulate progressively, disrupting the structure and function of affected tissues and leading inexorably to organ failure and death. No treatment yet exists which can specifically cause resolution, but intervention which reduces the availability of the amyloid fibril precursor proteins may lead to regression.

Clinical amyloidosis

Introduction

Amyloidosis occurs in many clinical disorders. Amyloid deposits in the brain and cerebral blood vessels are a central part of the pathology of Alzheimer's disease, which is the fourth most common cause of death in the Western world, whilst amyloid is present in the islets of Langerhans of the pancreas in all patients with type 2, maturity onset, diabetes mellitus. Amyloid deposition in the bones, joints, and periarticular structures eventually affects most patients who are on long-term haemodialysis for endstage renal failure and is the most frequent cause of serious morbidity among the approximately 800 000 such individuals worldwide. Systemic amyloidosis complicating myeloma and other B-cell dyscrasias, or chronic infections and inflammatory diseases, is very important because diagnosis is often difficult and the prognosis is poor, but effective treatments are increasingly available. Hereditary amyloidosis is very rare, except in a few geographic foci, but its diversity is remarkable. It is important because of its poor prognosis, the complexity of clinical management, the difficult genetic issues involved, and its considerable value as a model for understanding the pathogenesis of amyloid deposition.

Although there are some correlations between fibril protein type and clinical manifestations, there are also many forms of acquired and hereditary amyloidosis in which there is little or no concordance between the fibril protein, or the genotype of its precursor, and the clinical phenotype. There are evidently genetic and/or environmental factors, which are distinct from the amyloid fibril protein itself, which determine whether, when, and where clinically significant amyloid deposits form. The nature of these important determinants of amyloidogenesis is obscure. Furthermore, the mechanisms by which amyloid deposition causes disease are poorly understood. Whilst a heavy amyloid load is invariably a bad sign, there is often a poor correlation between the local amount of amyloid and the level of organ dysfunction. Active deposition of new amyloid is often associated with enhanced deterioration compared with stable long-standing deposits. Nascent or newly formed amyloid fibrils, generated *in vitro*, are also cytotoxic to cultured cells, whilst aged or *ex vivo* fibrils are generally inert, although it is not known how this relates to effects *in vivo*.

Reactive systemic (AA) amyloidosis

Associated conditions

AA amyloidosis occurs in association with chronic inflammatory disorders, chronic local or systemic microbial infections, and occasionally malignant neoplasms. In Western Europe and the United States the most frequent predisposing conditions are idiopathic rheumatic diseases (Table 3). Amyloidosis complicates up to 10 per cent of cases of rheumatoid arthritis and juvenile inflammatory arthritis, although for reasons that are not clear the incidence is lower in the United States than in Europe. Amyloidosis is exceptionally rare in systemic lupus erythematosus and related connective tissue diseases and in ulcerative colitis, in contrast to Crohn's disease. Tuberculosis and leprosy are important causes of AA amyloidosis, particularly where these infections are endemic. Chronic osteomyelitis, bronchiectasis, chronically infected burns, and decubitus ulcers as well as the chronic pyelonephritis of paraplegic patients are other well-recognized associations (Table 3). Hodgkin's disease and renal carcinoma, which often cause fever, other systemic symptoms, and a major acute phase response, are the malignancies most commonly associated with systemic AA amyloidosis.

Clinical features

AA amyloid involves the viscera but may be widely distributed without causing clinical symptoms. More than 90 per cent of patients present with non-selective proteinuria due to glomerular deposition, and nephrotic syndrome may develop before progression to endstage renal failure. Haematuria, isolated tubular defects, nephrogenic diabetes insipidus, and diffuse renal calcification occur rarely. Kidney size is usually normal, but may be enlarged, or, in advanced cases, reduced. Endstage chronic renal failure is the cause of death in 40 to 60 per cent of cases but acute renal failure may be precipitated by hypotension and/or salt and water depletion following surgery, excessive use of diuretics, or intercurrent infection, and may be associated with renal vein thrombosis. The second most common

presentation is with organ enlargement, such as hepatosplenomegaly or thyroid goitre, with or without overt renal abnormality, but in any case amyloid deposits are almost always widespread at the time of presentation. Involvement of the heart and gastrointestinal tract is frequent, but rarely causes functional impairment.

AA amyloidosis may become clinically evident early in the course of associated disease, but the incidence increases with duration of the primary condition. The mean duration of chronic rheumatic diseases such as rheumatoid arthritis, ankylosing spondylitis, or juvenile rheumatoid arthritis before amyloid is diagnosed is 12 to 14 years, although it can present much sooner. For most patients the prognosis is closely related to the degree of renal involvement and the effectiveness of treatment of the underlying inflammatory condition. In the presence of persistent, uncontrolled inflammation, 50 per cent of patients with AA amyloid die within 5 years of the amyloid being diagnosed; however, if the acute phase response can be consistently suppressed proteinuria can cease, renal function is retained, and the prognosis is much better. Availability of chronic haemodialysis and transplantation prevents early death from uraemia *per se*, but amyloid deposition in extrarenal tissues is responsible for a less favourable prognosis than for other causes of endstage renal failure.

Amyloidosis associated with immunocyte dyscrasia: monoclonal immunoglobulin light chain (AL) amyloidosis

Associated conditions

Almost any dyscrasia of cells of the B-lymphocyte lineage, including multiple myeloma, malignant lymphomas, and macroglobulinaemia, may be complicated by AL amyloidosis but most cases are associated with otherwise 'benign' monoclonal gammopathy. Amyloid occurs in up to 15 per cent of cases of myeloma, in a lower proportion of other malignant B-cell disorders, and probably in fewer than 5 per cent of patients with a 'benign' monoclonal gammopathy. In some cases deposition of AL amyloid may be the only evidence of the dyscrasia. A monoclonal paraprotein or free light chains can be detected in the serum or urine of only about 90 per cent of patients with AL amyloid, but detection of immunoglobulin gene rearrangement in the bone marrow or peripheral blood sometimes confirms a monoclonal gammopathy in the remaining cases. The paraprotein may also appear after presentation and diagnosis of the amyloid, and subnormal levels of some or all serum immunoglobulins or increased numbers of marrow plasma cells may provide less direct clues to the underlying aetiology. Until recently it has been the practice to diagnose apparently 'primary' cases of amyloidosis, with no previous predisposing inflammatory condition or family history of amyloidosis, as AL type by exclusion. However, it has lately been recognized that autosomal dominant hereditary non-neuropathic amyloidosis, particularly that caused by variant fibrinogen a-chain, may be poorly penetrant and of late onset, so that there may be no family history. The coincident occurrence of a monoclonal gammopathy may then be gravely misleading and it is essential to exclude by genotyping all known amyloidogenic mutations, and to seek positive immunohistochemical or biochemical identification of the amyloid fibril protein in all cases.

Clinical features

AL amyloid occurs equally in men and women, usually over the age of 50 but occasionally in young adults. It has a lifetime incidence, and is the cause of death, of between 0.5 and 1 per thousand individuals in the United Kingdom. The clinical manifestations are protean, as virtually any tissue other than the brain may be directly involved. Uraemia, heart failure, or other effects of the amyloid usually cause death within a year of diagnosis, unless the underlying B-cell clone is effectively suppressed.

The heart is affected in 90 per cent of patients with AL amyloid, in 30 per cent of whom restrictive cardiomyopathy is the presenting feature and in up to 50 per cent of whom it is fatal. Other cardiac presentations include arrhythmias and angina. Renal AL amyloid has the same manifestations as renal AA amyloid, but the prognosis is worse. Gut involvement may cause disturbances of motility (often secondary to autonomic neuropathy), malabsorption, perforation, haemorrhage, or obstruction. Macroglossia occurs rarely but is almost pathognomonic. Hyposplenism sometimes occurs in both AA and AL amyloidosis. Painful sensory polyneuropathy with early loss of pain and temperature sensation followed later by motor deficits is seen in 10 to 20 per cent of cases and carpal tunnel syndrome in 20 per cent. Autonomic neuropathy leading to orthostatic hypotension, impotence, and gastrointestinal disturbances may occur alone or together with the peripheral neuropathy, and has a very poor prognosis. Skin involvement takes the form of papules, nodules, and plaques usually on the face and upper trunk, and involvement of dermal blood vessels results in purpura occurring either spontaneously or after minimal trauma and is very common. Articular amyloid is rare but may mimic acute polyarticular rheumatoid arthritis, or it may present as asymmetrical arthritis affecting the hip or shoulder. Infiltration of the glenohumeral joint and surrounding soft tissues occasionally produces the characteristic 'shoulder pad' sign. A rare but serious manifestation of AL amyloid is an acquired bleeding diathesis that may be associated with deficiency of factor X and sometimes also factor IX, or with increased fibrinolysis. It does not occur in AA amyloidosis, although in both AL and AA disease there may be serious bleeding in the absence of any identifiable factor deficiency.

Senile amyloidosis

Some amyloid is present in all autopsies on individuals over 80 years of age but it is not known whether this contributes to the ageing process or whether it is an epiphenomenon that becomes clinically important only when it is extensive.

Senile systemic amyloidosis

Up to 25 per cent of old people have microscopic, clinically silent systemic deposits of transthyretin amyloid involving the walls of the heart and blood vessels, smooth and striated muscle, fat tissue, renal papillae, and alveolar walls. In contrast to most other forms of systemic amyloidosis, including hereditary transthyretin amyloid caused by point mutations in the transthyretin gene, the spleen and renal glomeruli are rarely affected. The brain is not involved. Occasionally more extensive deposits in the heart, affecting the ventricles and atria and situated in the interstitium and vessel walls, cause significant impairment of cardiac function and may be fatal. The transthyretin involved is probably usually of the normal wild type but cases with transthyretin variants have been described which may be hereditary.

Senile focal amyloidosis

Microscopic and clinically silent amyloid deposits of different fibril types, localized to particular tissues, are very commonly present in old people. Deposits of b-protein (see below) as amyloid in cerebral blood vessels and intracerebral plaques seen in 'normal' elderly brains may or may not be the harbinger of Alzheimer's disease had the patient survived long enough. Amyloid deposits are present in most osteoarthritic joints at surgery or autopsy, usually in close association with calcium pyrophosphate deposits, affecting the articular cartilage and joint capsule. However, neither the clinical significance of this age-associated articular amyloid nor its biochemical nature are known. The corpora amylacea of the prostate are composed of b_2-microglobulin amyloid fibrils. Amyloid in the seminal vesicles is derived from an as yet unidentified exocrine secretory product of the vesicle cells. Isolated deposits of cardiac atrial amyloid consist of atrial natriuretic peptide. Focal amyloid deposits commonly present in atheromatous plaques of elderly subjects contain fibrils composed of the N-terminal fragment of apolipoprotein A-I.

Cerebral amyloidosis

Introduction

The brain is a very common and important site of amyloid deposition (Table 4), although possibly because of the blood–brain barrier there are never any deposits in the cerebral parenchyma itself in any form of acquired systemic visceral amyloidosis. However, cerebrovascular transthyretin amyloid may occur in familial amyloid polyneuropathy due to the most common transthyretin variant (methionine for valine at residue 30), and oculoleptomeningeal amyloidosis is caused by other very rare transthyretin variants. The common and major forms of brain amyloid are confined to the brain and cerebral blood vessels with the single exception of cystatin C amyloid in hereditary cerebral haemorrhage with amyloidosis, Icelandic type, in which there are major though clinically silent systemic deposits.

Alzheimer's disease

By far the most frequent and important type of amyloid in the brain is that related to Alzheimer's disease, which is the most common cause of dementia and affects over 3 million individuals in the United States and a corresponding proportion of other Western populations. It is generally a disease of the elderly and its prevalence is therefore increasing. The clinical differential diagnosis of senile dementia and the positive identification of Alzheimer's disease are difficult and often of limited precision in life. However, intracerebral and cerebrovascular amyloid deposits are hallmarks of the neuropathological diagnosis. The amyloid fibrils are composed of b-protein (Ab), a 39- to 43-residue cleavage product of the large amyloid precursor protein. The vast majority of cases of Alzheimer's disease are sporadic but there are also families with an autosomal dominant pattern of inheritance and usually early onset. In about 20 families there are causative mutations in the *APP* gene for amyloid precursor protein on chromosome 21, and most other kindreds have mutations in the genes for presenilin 1 (chromosome 14) and presenilin 2 (chromosome 1). All these mutations are associated with increased production from amyloid precursor protein of Ab1–42, the most amyloidogenic form of Ab. Since all individuals

with Down's syndrome, that is trisomy 21, develop Alzheimer's disease if they survive into their forties, there is evidently a close link between amyloid precursor protein, Ab overproduction, Ab amyloidosis, and the pathogenesis of Alzheimer's disease, although it remains unclear whether or how Ab *per se*, or the amyloid fibrils that it forms, contribute to the neuronal loss which underlies the dementia. Synthetic amyloid b fibrils formed *in vitro* are markedly cytotoxic and cause the death of cultured cells by apoptosis and necrosis, but it is not clear to what extent these findings reflect phenomena that may be responsible for neurodegeneration *in vivo*. There is controversy about the correlation between the severity of dementia in Alzheimer's disease and the extent of amyloid angiopathy and plaques. Nevertheless the fact that patients with Alzheimer's disease caused by amyloid precursor protein and presenilin mutations have exactly the same neuropathology as sporadic cases, including tangles, argues strongly that the amyloid precursor protein and b-protein pathway can be of primary pathogenetic significance.

In addition to the Ab amyloid deposits in the brains of patients with Alzheimer's disease and Down's syndrome, there are also extensive 'amorphous' deposits of amyloid b throughout the brain. These do not stain with Congo red, and are detectable only by immunohistochemical staining. Their significance is unknown. They apparently precede the appearance of histochemically identifiable amyloid but are not necessarily the precursor of it because they are present in areas such as the cerebellum in which Ab amyloid is never seen. The non-fibrillar, non-amyloid protein apolipoprotein E is demonstrable in many amyloid deposits, including those of Alzheimer's disease. The *ApoE4* gene (chromosome 19), encoding one of the three isoforms of this apolipoprotein, is strongly associated with predisposition to develop Alzheimer's disease and with increased amounts of amyloid in the brain, but the underlying mechanisms are unknown.

Another neuropathological feature of Alzheimer's disease, and some other neurodegenerative conditions, is the neurofibrillary tangle located intracellularly within neuronal cell bodies and processes. These tangles have a characteristic ultrastructural morphology of paired helical filaments, and although they bind Congo red and then give the pathognomonic green birefringence of amyloid when viewed in polarized light, they are completely different structurally from amyloid fibrils. They are composed of an abnormally phosphorylated form of the normal neurofilament protein, tau.

Senile cerebral amyloidosis and amyloid angiopathy

The cerebral blood vessels contain Ab amyloid in up to 60 per cent of aged brains of non-demented individuals and there may also be focal intracerebral Ab amyloid plaques. These deposits are usually clinically silent and may or may not be harbingers of Alzheimer's disease, had the patients survived long enough. Sometimes the amyloid angiopathy is more extensive and it is a rare but important cause of cerebral haemorrhage and stroke, to be distinguished from atherosclerotic cerebrovascular disease.

Hereditary cerebral haemorrhage with amyloidosis: hereditary cerebral amyloid angiopathy

Icelandic type

Cerebrovascular amyloid deposits composed of a fragment of a genetic variant of cystatin C are responsible for recurrent major cerebral haemorrhages starting in early adult life in members of families originating in western Iceland. There is autosomal dominant inheritance and appreciable but clinically silent amyloid deposits are present in the spleen, lymph nodes, and skin. There is no extravascular amyloid in the brain and the neurological deficits, often including dementia, of surviving patients are compatible with their cerebrovascular pathology.

Dutch type

In families originating in a small region on the Dutch coast the autosomal dominant inheritance of recurrent normotensive cerebral hemorrhages starting in middle age is due to deposition of a genetic variant of Ab as cerebrovascular amyloid. There are also 'amorphous' Ab deposits in the brain and early senile plaques, without congophilic amyloid cores. Multi-infarct dementia occurs in survivors but some patients become demented in the absence of stroke. Amyloid outside the brain has not been reported.

Cerebral amyloid associated with prion disease

The neuropathology of a group of progressive, invariably fatal spongiform encephalopathies which are transmissible and in some cases are hereditary, sometimes includes intracerebral amyloid plaques and amyloid cerebral angiopathy. These diseases, sporadic and familial Creutzfeldt–Jacob disease, the familial Gerstmann–Sträussler–Scheinker syndrome, and kuru are caused by prions (PrP^{Sc}), conformational isoforms of the normal physiological cellular prion protein (PrP^{C}). The human diseases are closely related to the animal diseases scrapie of sheep and goats, transmissible encephalopathy of mink, elk, and male deer, and bovine spongiform encephalopathy. Variant Creutzfeldt–Jacob disease is apparently the result of transmission of bovine spongiform encephalopathy to humans. The significance of amyloid *per se* in these disorders is not clear, because it is not always detectable histologically and is not seen, for example, in fatal familial insomnia or in bovine spongiform encephalopathy, which is apparently a result of the transmission of ovine scrapie to cattle. When scrapie or its human counterparts are transmitted to experimental animals by inoculation of affected brain tissue the development of intracerebral amyloid depends on the strain of infectious agent and the genetic background of the recipient. Even when amyloid is present in the brain it is not seen elsewhere, for example in the spleen, although the latter is a rich source of the infective agent. However, when the infective agent is exhaustively and highly purified from brain or spleen it forms typical congophilic amyloid fibrils, composed of the proteinase-resistant subunit which is the prion, PrP^{Sc}, and when amyloid deposits are present in affected brains they immunostain with antiprion antibodies. The amyloid fibril protein is thus directly related to the cause of the encephalopathy but gross amyloid deposition is evidently not necessary for expression of disease. Neuronal damage may perhaps be caused by cytotoxic prefibrillar PrP^{Sc} aggregates, or indeed by other mechanisms entirely. This is a different situation from the extracerebral amyloidoses and from cystatin C and non-hereditary cerebral amyloid angiopathies, in which amyloid deposition is invariably present when there is clinical disease.

Hereditary systemic amyloidosis

Familial amyloid polyneuropathy

Familial amyloid polyneuropathy is an autosomal dominant syndrome with onset at any time from the second decade onwards, characterized by progressive peripheral and autonomic neuropathy and varying degrees of visceral involvement affecting especially the vitreous of the eye, the heart, kidneys, thyroid, and adrenals. There are usually amyloid deposits throughout the body involving the walls of blood vessels as well as the connective tissue matrix, and the pathology is due to these deposits. Apart from major foci in Portugal, Japan, and Sweden, familial amyloid polyneuropathy has been reported in most ethnic groups throughout the world. There is considerable variation in the age of onset, rate of progression, and involvement of different systems, although within families the pattern is usually quite consistent. There is remorseless progression and the disorder is invariably fatal. Death results from the effects and complications of peripheral and/or autonomic neuropathy, or from cardiac or renal failure.

Familial amyloid polyneuropathy is caused by mutations in the gene for the plasma protein transthyretin, formerly known as prealbumin, the most frequent of which causes a methionine for valine substitution at position 30 in the mature protein, but over 60 amyloidogenic mutations have been described. There is often little correlation between the underlying mutation and the clinical phenotype, which is evidently determined by other genetic and possibly also environmental factors, although in a few cases certain mutations are uniquely associated with particularly aggressive or relatively organ-limited disease. The amyloidogenic transthyretin mutations are not always penetrant, and asymptomatic methionine 30 homozygotes over the age of 60 have been reported. Rare kindreds with the apolipoprotein A-I arginine 26 variant, which usually causes non-neuropathic amyloidosis, may present with prominent peripheral neuropathy resembling transthyretin familial amyloid polyneuropathy.

Familial amyloid polyneuropathy with predominant cranial neuropathy

Originally described in Finland but now reported in other ethnic groups, this autosomal dominant hereditary amyloidosis presents in adult life with cranial neuropathy, lattice corneal dystrophy, and distal peripheral neuropathy. There may be skin, renal, and cardiac manifestations and microscopic amyloid deposits are widely distributed in connective tissue and blood vessel walls, although life expectancy approaches normal. The amyloid fibrils are derived from variants of the actin-modulating protein gelsolin, encoded by point mutations. Individuals homozygous for these mutations have severe renal amyloidosis in addition to the usual neuropathy.

Non-neuropathic systemic amyloidosis

In this rare autosomal dominant syndrome of major systemic amyloidosis without clinical evidence of neuropathy, the patterns of organ involvement and overall clinical phenotype vary between families. The kidneys are often most severely affected leading to hypertension and renal failure, but the heart, spleen, liver, bowel, connective tissue, and exocrine glands may all be involved. Following clinical presentation, usually from the second decade onwards, there is inexorable progression to death or organ failure requiring transplantation. Clinical presentation is usually in early adulthood, although in a few kindreds it may be as late as the sixth decade. The amyloid proteins identified so far are genetic variants of apolipoprotein A-I, lysozyme, and the a-chain of fibrinogen.

Cardiac amyloidosis

Cardiac amyloidosis, without overt involvement of other viscera or neuropathy, progressing inexorably to death, is associated with certain transthyretin gene mutations and is inherited as an autosomal dominant with variable penetrance (see Table 2). By far the most common variant is isoleucine 122 transthyretin which occurs in 4 per cent of African-Americans and frequently causes cardiac amyloidosis from the sixth decade onwards.

Familial Mediterranean fever

Familial Mediterranean fever is an autosomal recessive disorder caused by mutations in the gene on chromosome 16 that encodes a neutrophil-specific protein of unknown function, called pyrin or marenostrin. The disease is characterized by recurrent episodes of fever, abdominal pain, pleurisy, or arthritis, and predominantly occurs in non-Ashkenazi Jews, Armenians, Anatolian Turks, and Levantine Arabs. In Sephardi Jews of North African origin, and in the other ethnic groups except Armenians and to a lesser extent Ashkenazi Jews, untreated familial Mediterranean fever is eventually complicated in a high proportion of cases by typical systemic AA amyloidosis. Furthermore, some patients with familial Mediterranean fever present with AA amyloidosis before they have experienced any symptoms, and this is consistent with the recent finding that a substantial acute phase plasma protein response is frequently present even in asymptomatic individuals. The different incidence of amyloid in patients with familial Mediterranean fever from different ethnic groups is not wholly explained by their specific pyrin gene mutations, and is another illustration of the unknown genetic determinants of clinical amyloidosis.

Haemodialysis-associated amyloidosis

Almost all patients with endstage renal failure who are maintained on haemodialysis for more than 5 years develop amyloid deposits composed of b $_2$-microglobulin. These deposits are predominantly osteoarticular and are associated with carpal tunnel syndrome, large joint pain and stiffness, soft tissue masses, bone cysts, and pathological fractures. Renal tubular amyloid concretions may also form. The serious clinical problems associated with b $_2$-microglobulin amyloidosis constitute the major cause of morbidity in patients on long-term dialysis. Furthermore, in some such patients more extensive deposition occurs, most commonly in the spleen but also in other organs, and a few cases of death associated with systemic b $_2$-microglobulin amyloid have been reported. The b $_2$-microglobulin is derived from the high plasma concentrations which develop in renal insufficiency and which are not cleared by dialysis. This type of amyloid also occurs in patients on continuous ambulatory peritoneal dialysis and has even been reported in a few patients with chronic renal failure who had never been dialysed.

Endocrine amyloidosis

Many tumours of APUD cells which produce peptide hormones have amyloid deposits in their stroma. These are probably composed of the hormone peptides, and in the case of medullary carcinoma of the thyroid the fibril subunits are derived from procalcitonin. In insulinomas the amyloid fibril protein is a novel peptide first identified in that site and subsequently shown to be the fibril protein in the amyloid of the islets of Langerhans in type II, maturity onset, diabetes. This peptide is called islet amyloid polypeptide (and also amylin) and shows appreciable homology with calcitonin gene-related peptide. Islet amyloid polypeptide amyloid is an almost universal feature of the pancreatic islets in type II diabetes and becomes more extensive with increasing duration and severity of the disease. Although the amyloid itself is probably not initially responsible for the metabolic defect in this form of diabetes, it is likely that progressive amyloid deposition leading to islet destruction subsequently does contribute to the pathogenesis. The possible hormonal or other role of islet amyloid polypeptide itself, which is produced by the islet B cells, is also not yet clear.

Rare localized amyloidosis syndromes

Amyloid deposits localized to the skin occur in both acquired and hereditary forms. Primary localized cutaneous amyloidosis presents in adult life as macular or papular lesions, the fibrils of which may be derived from keratin. Hereditary cutaneous amyloid lesions are rare, of unknown fibril type, and are sometimes associated with other, non-amyloid, multisystem disorders. Amyloid deposits in the eye cause local problems in the cornea (corneal lattice dystrophy) or conjunctiva, whilst orbital amyloid presents as mass lesions which can disrupt eye movement and the structure of the orbit. In one such case the fibril protein has been identified as a fragment of immunoglobulin G heavy chain.

Localized foci of AL amyloid can occur anywhere in the body in the absence of systemic AL amyloidosis, the most common sites being the skin, upper airways and respiratory tract, and the urogenital tract. They may be associated with a local plasmacytoma or B-cell lymphoma producing a monoclonal immunoglobulin, but often the cells, which must be present to produce the amyloidogenic protein, are scattered inconspicuously in the affected tissue. The clinical problems caused by these space-occupying amyloidomas are usually cured by surgical resection, but this is not always possible.

Amyloid fibrils

Regardless of their very diverse protein subunits, amyloid fibrils of different types are remarkably similar: straight, rigid, non-branching, of indeterminate length, and 10 to 15 nm in diameter. They are insoluble in physiological solutions, relatively resistant to proteolysis, and bind Congo red dye producing pathognomonic green birefringence when viewed in polarized light. Electron microscopy reveals that each fibril consists of two or more protofilaments, the precise number varying with the fibril type. The X-ray diffraction patterns of all the different *ex vivo* amyloid fibrils, and of synthetic fibrils formed *in vitro*, that have been studied demonstrate the presence of a common core structure within the filaments, in which the subunit proteins are arranged in a stack of twisted antiparallel b-pleated sheets lying with their long axes perpendicular to the long axis of the fibril. Recent observations show that many different proteins, including molecules totally unrelated to amyloidosis *in vivo*, can be refolded after denaturation *in vitro* to form typical, stable, congophilic cross b fibrils. Although it is not clear why only the 20 or so known amyloidogenic proteins adopt the amyloid fold and persist as fibrils *in vivo*, a major unifying theme that is currently emerging is that in all cases studied the precursors are relatively destabilized. Even under physiological or other conditions they may encounter *in vivo*, they populate partly unfolded states, involving loss of tertiary or higher-order structure, that readily aggregate with retention of b-sheet secondary structure into protofilaments and fibrils. Once the process has started, seeding may also play an important facilitating role, so that amyloid deposition may progress exponentially as expansion of the amyloid template 'captures' further precursor molecules.

Amyloid fibril proteins and their precursors

Immunoglobulin light chain

AL proteins are derived from the N-terminal region of monoclonal immunoglobulin light chains and consist of all or part of the variable (V $_L$) domain. Intact light chains may occasionally be found, and the molecular weight therefore varies between about 8000 and 30 000 Da. The light chain of the monoclonal paraprotein is either identical to, or clearly the precursor of, AL isolated from the amyloid deposits.

AL is more commonly derived from l chains than from _ chains, despite the fact that _ chains predominate among both normal immunoglobulins and the paraprotein products of immunocyte dyscrasias. A new l-chain subgroup, l$_{VI}$, was identified first as an AL protein in two cases of immunocyte dyscrasia-associated amyloidosis before it had been recognized in any other form, and it has subsequently been observed in many more cases of AL amyloidosis. Furthermore, there is increasing evidence from sequence analyses of Bence Jones proteins of both _ and l type from patients with AL amyloidosis, and of AL proteins themselves, that these polypeptides contain unique amino acid replacements or insertions compared with non-amyloid monoclonal light chains. In some cases these changes involve replacement of hydrophilic framework residues by hydrophobic residues, changes likely to promote aggregation and insolubilization, and in others the monoclonal light chains from amyloid patients have been demonstrated directly to have decreased solubility and a greater propensity for precipitation than control non-amyloid proteins. The inherent 'amyloidogenicity' of particular monoclonal light chains has been elegantly confirmed in an *in vivo* model in which isolated Bence Jones proteins were injected into mice. Animals receiving light chains from AL amyloid patients developed typical amyloid deposits composed of the human protein whereas

animals receiving light chains from myeloma patients without amyloid did not.

AA

The AA protein is a single non-glycosylated polypeptide chain usually of mass 8000 Da and containing 76 residues corresponding to the N-terminal portion of the 104-residue serum amyloid A protein (SAA). Smaller and larger AA fragments, even some whole SAA molecules, have also been reported in AA fibrils. Serum amyloid A is an apolipoprotein of high-density lipoprotein particles and is the polymorphic product of a set of genes located on the short arm of chromosome 11. Serum amyloid A is highly conserved in evolution and is a major acute phase reactant in all species in which it has been studied. Most of the SAA in plasma is produced by hepatocytes in which the synthesis is under transcriptional regulation by cytokines, especially interleukin 1, interleukin 6, and tumour necrosis factor, acting via nuclear factor _B-like and possibly other transcription factors. After secretion it is rapidly associated with high-density lipoproteins from which it displaces apolipoprotein A-I. The circulating concentration can rise from normal levels of up to 3 mg/l to over 1000 mg/l within 24 to 48 h of an acute stimulus, whilst with ongoing chronic inflammation the level may remain persistently high. Certain isoforms of SAA, the products of different genes, are predominantly synthesized elsewhere in the body by macrophages, adipocytes, and certain other cells. Although they also associate with high-density lipoproteins, their acute phase synthesis is stimulated differently and they presumably have different functions. There is also a closely related family of high-density lipoprotein trace apoproteins which are not acute phase reactants and which have been designated 'constitutive SAAs', although they do not form amyloid.

Circulating SAA is the precursor of amyloid fibril AA protein, from which it is derived by proteolytic cleavage. Such cleavage can be produced by macrophages and by a variety of proteinases but since further cleavage of AA is readily demonstrable *in vitro* it is not clear why the AA peptide persists in amyloid. Furthermore, it is not known whether in the process of AA fibrillogenesis, cleavage of SAA occurs before and/or after aggregation of monomers. Persistent overproduction of SAA causing sustained high circulating levels is a necessary condition for deposition of AA amyloid but it is not known why only some individuals in this state get amyloid. In mice, only SAA2, one of the three major isoforms of murine SAA, is the precursor of AA in amyloid fibrils. Human SAA isoforms are more complex but homozygosity for particular types seems to favour amyloidogenesis, although there may also be ethnic differences.

The normal functions of SAA are not known, although modulating effects on reverse cholesterol transport and on lipid functions in the microenvironment of inflammatory foci have been proposed. A protein, homologous with SAA, produced by rabbit fibroblasts has been reported to act as an autocrine stimulator of collagenase production *in vitro*. Other reports of potent cell regulatory functions of isolated denatured delipidated SAA have yet to be confirmed with physiological preparations of SAA-rich high-density lipoproteins. Regardless of its physiological role, the behaviour of SAA as an exquisitely sensitive acute phase protein with an enormous dynamic range makes it an extremely valuable empirical clinical marker. It can be used to monitor objectively the extent and activity of infective, inflammatory, necrotic, and neoplastic disease. Furthermore, routine monitoring of SAA should be an integral part of the management of all patients with AA amyloid or disorders predisposing to it, as control of the primary inflammatory process in order to reduce SAA production is essential if amyloidosis is to be halted, enabled to regress, or prevented. Automated immuno-assay systems for SAA are available standardized to a World Health Organization international reference standard.

Transthyretin

Transthyretin, formerly known as prealbumin, is a normal non-glycosylated plasma protein, with a relative molecular mass of 54 980. It is composed of four identical non-covalently associated subunits each of 127 amino acids. It is produced by hepatocytes and the choroid plexus and is a significant negative acute phase protein. Each tetrameric molecule is able to bind a single thyroxine or triodothyronine molecule and up to 15 per cent of circulating thyroid hormone is transported in this way. Transthyretin also forms a 1:1 molecular complex with retinol-binding protein, which transports vitamin A.

Transthyretin is encoded by a single copy gene but is appreciably polymorphic and about 70 different point mutations encoding single-residue substitutions have been identified so far. Normal wild type transthyretin is an inherently amyloidogenic protein which forms the fibrils in senile systemic amyloidosis, and *in vitro* exposure to reduced pH is sufficient to generate transthyretin amyloid fibrils from the pure protein. Most of the variant forms of transthyretin have been associated with hereditary amyloidosis, and show decreased stability *in vitro* compared with the wild type. Transgenic mice expressing variant human transthyretin with a methionine 30 substitution develop extensive systemic amyloidosis, but no amyloid deposits have yet been reported in the peripheral nerves even when the transgene is expressed in the choroid plexus and transthyretin amyloid is deposited in the meninges and choroid plexus. This is another example of the important, unknown, factors, other than the presence of an amyloidogenic protein itself, that determine where and when clinical amyloidosis develops.

Individuals heterozygous for transthyretin mutations have a mixture of wild type and variant transthyretin monomers in their circulating transthyretin, and if they develop amyloidosis both forms are often present although the variant may predominate in the amyloid fibrils. Although cleavage fragments of transthyretin are commonly present, intact transthyretin subunits are also found and fibrillogenesis does not depend on an initial proteolytic step.

Ab

The fibril protein in the intracerebral and cerebrovascular amyloid of Alzheimer's disease, Down's syndrome, and hereditary amyloid angiopathy of the Dutch type is a 39- to 43-residue sequence derived by proteolysis from a high molecular weight precursor protein, the so-called amyloid precursor protein, encoded on the long arm of chromosome 21. Several isoforms of amyloid precursor protein (APP) are generated by alternative splicing of transcripts from the 19 exon gene, and yielding major forms: APP695, APP751, and APP770. These are each single-chain, multidomain glycoproteins with the 47 residues of the carboxy terminal within the cytoplasm, a 25-residue membrane-spanning region, and the rest of the molecule lying extracellularly. APP751 and APP770 contain a 56-residue Kunitz type serine proteinase inhibitor domain encoded by exon 7. Following glycosylation and membrane insertion, APPs are cleaved extracellularly by so-called APP secretase activity, close to the transmembrane sequence, releasing, in the case of the isoforms containing the proteinase inhibitor domain, a molecule known as proteinase nexin II, which avidly binds factor XIa, trypsin, and chymotrypsin as well as epidermal growth factor-binding protein and the g subunit of nerve growth factor. Although mRNA encoding APP695, which lacks the proteinase inhibitor domain, is the predominant species found in brain, whereas mRNA for APP751 is the most abundant in other tissues, 85 per cent of secreted APP in the brain is proteinase nexin II. Interestingly, APP secreted by a glial cell line is substantially glycosylated with chondroitin sulphate glycosaminoglycan chains. APP also undergoes high-affinity interactions with heparan sulphate. These observations suggest that APP may have important functions in cell adhesion, cell migration, and modulation of growth factor activities. APP proteinase nexin II is present in and released by platelets and probably functions in the clotting cascade.

The amyloidogenic amyloid b, encoded by parts of exons 16 and 17, corresponds to the part of the APP sequence which extends from within the cell membrane into the extracellular space. Secretase cleavage of APP to release the soluble form cannot therefore generate intact amyloid b itself, or larger fragments containing it. However, there is an alternative processing pathway for APP, in which it is taken up whole by lysosomes and cleaved to yield fragments that do contain the whole amyloid b sequence. Furthermore, APP cleaved at the *N*-terminus of amyloid b, and also free soluble amyloid b itself, are normally produced by cell lines and by mixed brain cells in culture and are present in the cerebrospinal fluid. However, the source of the amyloid b in the intracerebral amorphous deposits and of that which aggregates as amyloid fibrils in the brain and cerebral blood vessels is still not known. The 42-residue form of amyloid b is markedly the most amyloidogenic, and all the mutations in the APP and presenilin genes that are associated with hereditary Alzheimer's disease result in increased production of this amyloid b1–42. Increased availability of the precursor is thus responsible for amyloidogenesis, but the pathogenesis of neuronal damage and dementia remain unclear.

Cystatin C

Cystatin C (formerly called g-trace) is an inhibitor of cysteine proteinases, including cathepsins B, H, and L. It is encoded by a gene on chromosome 20 and consists of a single non-glycosylated polypeptide chain of 120 residues. It is present in all major human biological fluids at concentrations compatible with a significant physiological role in proteinase inhibition. The normal concentration in cerebrospinal fluid is 6.5 mg/l (range 2.7 to 13.7, $n = 34$), but is much lower (2.7 mg/l, range 1.0 to 4.7, $n = 9$) in patients with the Icelandic type of hereditary cerebral amyloid angiopathy in whom fragments of the glutamine 68 genetic variant of cystatin C form the amyloid fibrils. This reduced concentration is useful diagnostically and is evident even in presymptomatic carriers of the cystatin C gene mutation. The point mutation that causes the disease encodes a glutamine for leucine substitution in the mature protein and the amyloid fibril protein consists of the C-terminal 110 residues of the variant. This amino-terminally truncated form is not detectable in the cerebrospinal fluid of affected patients, suggesting that cleavage takes place either in close proximity to fibril deposition or is a postfibrillogenic event. The variant cystatin C is less stable than wild type and readily forms fibrils *in vitro*. It is not known whether cerebral haemorrhage in cystatin C amyloidosis is caused simply by the damaging effects of vascular amyloid deposition or whether deficiency in inhibitory capacity for cysteinase proteinases also plays a part.

Gelsolin

Gelsolin (mass 90 000 Da) is a widely distributed cytoplasmic protein which binds actin monomers, nucleates actin filament growth, and severs actin filaments. Alternative transcriptional initiation and message processing from a single gene on chromosome 9 are responsible for synthesis of a secreted form of gelsolin (mass 93 000 Da), which circulates in the plasma at a concentration of about 200 mg/l. Its function in the blood is not known but may be related to clearance of actin filaments released by dying cells. In the Finnish type of hereditary amyloidosis the amyloid fibril protein is a 71-residue fragment of variant gelsolin with asparagine substituted for aspartic acid at position 15, corresponding to residue 187 of the mature molecule, and the same mutation has been discovered in affected kindreds from different ethnic backgrounds. In one Danish family with the same phenotype there is a different mutation at the same nucleotide, predicting a tyrosine for aspartic acid substitution at residue 187. Synthetic and recombinant peptides including the asparagine for aspartic acid substitution at residue 187 are less soluble than the wild type sequence and readily form amyloid fibrils *in vitro*.

Apolipoprotein A-I

Apolipoprotein A-I is the most abundant apolipoprotein amongst the high-density lipoprotein particles and participates in their central function of reverse cholesterol transport from the periphery to the liver. Apolipoprotein A-I variants are extremely rare and may be phenotypically silent or may affect lipid metabolism. However, nine different variants of apolipoprotein A-I, including single- and multiple-residue substitutions and deletions, have been associated with amyloidosis. Although inherited as an autosomal dominant, and usually highly penetrant, there are marked variations in age and manner of presentation even in the same family and in different kindreds with the same mutation. The amyloid fibril protein consists, in all cases studied, of the first 90 or so N-terminal residues even when the causative variant residue(s) are more distal. Wild type apolipoprotein A-I is also amyloidogenic, forming the deposits associated with atheromatous plaques in the elderly, and the various amyloidogenic mutations presumably encode sequence changes that render apolipoprotein A-I less stable and/or more liable to cleavage to yield the fibrillogenic N-terminal fragment.

Lysozyme

Lysozyme is the classic bacteriolytic enzyme of external secretions, discovered by Fleming in 1922. It is also present at high concentration within articular cartilage and in the granules of polymorphs, and is the major secreted product of macrophages. Lysozymes are present in most organisms in which they have been sought, although their physiological role is not always clear. The complete structures of hen egg white and human lysozymes are known to atomic resolution and their catalytic mechanism, epitopes, folding, and other aspects of their structure–function relationship have been analysed exhaustively. This contrasts with the absence of detailed three-dimensional structural information on all other amyloid fibril proteins and their precursors except transthyretin and b $_2$-microglobulin. Lysozyme, unlike transthyretin and b_2-microglobulin, is not inherently amyloidogenic, and is therefore a valuable model for investigation of amyloid fibrillogenesis. There is only one copy of the lysozyme gene in the human genome and no disease is associated with lysozyme other than amyloidosis. The mutations which cause amyloid produce substitution of threonine for isoleucine at residue 56 in one family and histidine for aspartic acid at residue 67 in others. These dramatic changes in residues which are extremely conserved throughout the lysozyme and related a-lactal-bumin protein families, destabilize the native fold so that the variants readily populate partly unfolded states even under physiological conditions and spontaneously aggregate *in vitro*, and evidently also *in vivo*, into amyloid fibrils.

Islet amyloid polypeptide

Islet amyloid polypeptide (amylin) is a 37-residue molecule encoded by a gene on chromosome 12 and with 46 per cent sequence homology to the neuropeptide calcitonin gene-related peptide. Islet amyloid polypeptide is produced in the b cells of the pancreatic islets of Langerhans and is stored in and released from their secretory granules together with insulin. It has been reported to modulate insulin release, and to induce peripheral insulin resistance, vasodilatation, and lowering of plasma calcium but neither its physiological role nor its contribution to diabetes are yet known.

Amyloidogenicity of islet amyloid polypeptide depends on the amino acid sequence between residues 20 and 29, as shown by *in vitro* fibrillogenesis with synthetic peptides. The synthetic decapeptide IAPP20–29 and even the hexapeptide IAPP25–29, glycine–alanine–isoleucine–leucine–serine–serine, form amyloid-like fibrils *in vitro*, whereas other islet amyloid polypeptide fragments do not. There is also a correlation between conservation of this sequence and deposition of islet amyloid polypeptide amyloid in the islets of diabetic animals of different species. However, the role of the amyloid in diabetogenesis remains to be established. In the degu, a South American rodent, spontaneous diabetes is associated with islet amyloid composed of insulin, and xenogeneic insulin can also form amyloid in humans at sites of repeated therapeutic insulin injections.

b_2-Microglobulin

b_2-Microglobulin is a non-glycosylated, non-polymorphic single-chain protein of 99 residues with a single intrachain disulphide bridge (relative molecular mass 11 815) encoded by a single gene on chromosome 15. It becomes non-covalently associated with the heavy chain of major histocompatibility class I antigens and is required for transport and expression of the complex at the cell surface. Amino acid sequence homology places b $_2$-microglobulin in the superfamily including immunoglobulins, T-cell receptor a- and b-chains, Thy 1, major histocompatibility class I and II molecules, secretory component, etc. Its three-dimensional structure is a typical b-barrel with two antiparallel pleated sheets comprising three and four strands respectively, and closely resembles an immunoglobulin domain.

b_2-microglobulin is produced by lymphoid and a variety of other cells in which it stabilizes the structure and function of class I antigens at the cell surface. When these complexes are shed by cleavage of the heavy chain at the cell surface, free b $_2$-microglobulin is released. The circulating concentration of b $_2$-microglobulin is 1 to 2 mg/l and the protein is rapidly cleared by glomerular filtration and then catabolized in the proximal renal tubule. Impairment of renal function is associated with retention of b_2-microglobulin and increased circulating levels because there is no other site for its catabolism. Daily production of b $_2$-microglobulin is about 200 mg and in patients in endstage renal failure on haemodialysis, plasma b $_2$-microglobulin levels rise to and remain at levels of about 40 to 70 mg/l. Isolated unaltered b_2-microglobulin can form amyloid-like fibrils itself *in vitro*, and most studies of *ex vivo* b_2-microglobulin fibrils show the whole intact molecule to be the major subunit, although fragments and altered forms of b $_2$-microglobulin have also been reported.

Glycosaminoglycans

Amyloidotic organs contain more glycosaminoglycans than normal tissues and at least some of this is a tightly bound integral part of the amyloid fibrils. These fibril-associated glycosaminoglycans are heparan sulphate and dermatan sulphate in all forms of amyloid which have been investigated. Fibrils isolated by water extraction and separated from other tissue components contain 1 to 2 per cent by weight of glycosaminoglycan, none of which is covalently associated with the fibril protein. Interestingly, in systemic AA and AL amyloidosis, the only forms in which this has been studied so far, there is marked restriction of the heterogeneity of the glycosaminoglycan chains, suggesting that particular subclasses of heparan and dermatan sulphates are involved. Immunohistochemical studies demonstrate the presence of proteoglycan core proteins in all amyloid deposits, and that these are closely related to fibrils at the ultrastructural level. However, in isolated fibril preparations much of the glycosaminoglycan material is free carbohydrate chains and it is not yet clear whether this represents aberrant glycosaminoglycan metabolism related to amyloidosis or is just an artefact of postmortem degradation of core protein.

The significance of glycosaminoglycans in amyloid remains unclear, but their universal presence, intimate relationship with the fibrils, and restricted heterogeneity all suggest that they may be important. Glycosaminoglycans are known to participate in the organization of some normal structural proteins into fibrils and they may have comparable fibrillogenic effects on certain amyloid fibril precursor proteins. Furthermore the glycosaminoglycans on amyloid fibrils may be ligands to which serum amyloid P component, another universal constituent of amyloid deposits, binds.

Amyloid P component and serum amyloid P component

Amyloid deposits in all different forms of the disease, both in humans and in animals, contain the non-fibrillar glycoprotein amyloid P component. Amyloid P component is identical to and derived from the normal circulating plasma protein, serum amyloid P component, a member of the pentraxin protein family which includes C-reactive protein. Human serum amyloid P component is secreted only by hepatocytes, is a trace constituent of plasma (women: mean 24 mg/l, SD 8, range 8 to 55, n = 274; men: mean 32 mg/l, SD 7, range 12 to 50, n = 226), and is not an acute phase reactant. Nevertheless, apart from the fibrils themselves, amyloid P component is always by far the most abundant protein in all amyloid deposits.

Serum amyloid P component consists of five identical non-covalently associated subunits, each with a molecular mass of 25 462 Da, which are non-covalently

associated in a pentameric disc-like ring. The tertiary fold of the subunit is dominated by antiparallel b-sheets, forming a flattened b-barrel with jellyroll topology and a core of hydrophobic sidechains. This is the so-called 'lectin fold', shared with a variety of other animal, plant, and bacterial carbohydrate-binding proteins (lectins). Serum amyloid P component is a calcium-dependent ligand-binding protein, the best defined specificity of which is for the 4,6-cyclic pyruvate acetal of b- D-galactose, but it also binds avidly and specifically to DNA, to chromatin, to glycosaminoglycans, particularly heparan and dermatan sulphates, and to all known types of amyloid fibrils. This last is the interaction responsible for the unique, specific accumulation of serum amyloid P component in amyloid deposits. Aggregated, but not native, serum amyloid P component also binds specifically to C4-binding protein and fibronectin from plasma, although serum amyloid P component is not complexed with any other protein in the circulation. In addition to being a plasma protein, serum amyloid P component is also a normal constituent of certain extracellular matrix structures. It is covalently associated with collagen and/or other matrix components in the lamina rara interna of the human glomerular basement membrane and is present on the microfibrillar mantle of elastin fibres throughout the body.

No deficiency of serum amyloid P component has been described and it has been stably conserved in evolution. There is a single copy of its gene on chromosome 1, no polymorphism of the amino acid sequence, and the single biantennary oligosaccharide chain attached to asparagine at residue 32 is the most invariant glycan of any known glycoprotein. These indications that serum amyloid P component is likely to have important physiological function(s) have lately been confirmed by the finding that mice with targeted deletion of the gene for serum amyloid P component spontaneously develop marked antinuclear autoimmunity and immune complex glomerulonephritis. Studies of these serum amyloid P component knockout mice also show that serum amyloid P component is involved in host resistance to some infections and contributes to pathogenesis of others.

The serum amyloid P component molecule is highly resistant to proteolysis and, although not itself a proteinase inhibitor, its binding to amyloid fibrils *in vitro* protects them against proteolysis. Once bound to amyloid fibrils *in vivo*, serum amyloid P component persists for very prolonged periods and is not catabolized at all, in contrast to its rapid clearance from the plasma (half-life 24 h) and prompt catabolism in the liver. These observations suggest that serum amyloid P component may contribute to the persistence of amyloid deposits *in vivo*, and indeed serum amyloid P component knockout mice show retarded and reduced induction of experimental AA amyloidosis, confirming that serum amyloid P component is significantly involved in pathogenesis of amyloidosis.

Other proteins in amyloid deposits

A number of plasma proteins, other than the fibril proteins themselves and serum amyloid P component, have been detected immunohistochemically in some amyloid deposits. These include a_1-antichymotrypsin, some complement components, apolipoprotein E, and various extracellular matrix or basement membrane proteins. None of these match the universality, quantitative, or selective importance of serum amyloid P component, and their role, if any, in pathogenesis of amyloid deposition or its effects is not known.

Diagnosis and monitoring of amyloidosis

Introduction

Until recently amyloidosis was an exclusively histological diagnosis, and green birefringence of deposits stained with Congo red and viewed in polarized light remains the gold standard. Furthermore, immunohistochemical staining of amyloid-containing tissue is the simplest method for identifying the type of amyloid fibril present. However, biopsies provide extremely small samples and therefore can never provide information on the extent, localization, progression, or regression of amyloid deposits. A major advance in clinical amyloidosis has been the development of radiolabelled serum amyloid P component as a specific tracer for amyloid. Combined scintigraphic imaging and metabolic analysis using labelled serum amyloid P component have provided a wealth of new information on the natural history of many different forms of amyloid and their response to treatment.

Histochemical diagnosis of amyloid

Biopsy

Amyloid may be an incidental finding on biopsy of the kidneys, liver, heart, bowel, peripheral nerve, lymph node, skin, thyroid, or bone marrow. When amyloidosis is suspected clinically, biopsy of the rectum or subcutaneous fat is the least invasive. Amyloid is present in these sites in more than 90 per cent of cases of systemic AA or AL amyloidosis. Alternatively, a clinically affected tissue may be biopsied directly.

Congo red and other histochemical stains

Many cotton dyes, fluorochromes, and metachromatic stains have been used, but Congo red staining, and its resultant green birefringence when viewed with high-intensity polarized light, is the pathognomonic histochemical test for amyloidosis. The stain is unstable and must be freshly prepared every 2 months or less. A section thickness of 5 to 10 µm and inclusion in every staining run of a positive control tissue containing modest amounts of amyloid are critical.

Immunohistochemistry

Although many amyloid fibril proteins can be identified immunohistochemically, the demonstration of amyloidogenic proteins in tissues does not, on its own, establish the presence of amyloid. Congo red staining and green birefringence are always required and immunostaining may then enable the amyloid to be classified. Antibodies to serum amyloid A protein are commercially available and always stain AA deposits, similarly with anti-b $_2$-microglobulin antisera and haemodialysis-associated amyloid. In AL amyloid the deposits are stainable with standard antisera to _ or l in only about half of all cases, probably because the light chain fragment in the fibrils is usually the N-terminal variable domain, which is largely unique for each monoclonal protein. Immunohistochemical staining of transthyretin, Ab, and prion protein amyloid may require pretreatment of sections with formic acid or alkaline guanidine or deglycosylation.

Electron microscopy

Amyloid fibrils cannot always be convincingly identified ultrastructurally, and electron microscopy alone is not sufficient to confirm the diagnosis of amyloidosis.

Problems of histological diagnosis

The tissue sample must be adequate (for example, the inclusion of submucosal vessels in a rectal biopsy specimen), and failure to find amyloid does not exclude the diagnosis. The unavoidable sampling problem means that biopsy cannot reveal the extent or distribution of amyloid. Experience with Congo red staining is required if clinically important false negative and false positive results are to be avoided. Immunohistochemical staining requires positive and negative controls, including demonstration of specificity of staining by absorption of positive antisera with isolated pure antigens.

Non-histological investigations

Two-dimensional echocardiography showing small, concentrically hypertrophied ventricles, generally impaired contraction, dilated atria, homogeneously echogenic valves, and 'sparkling' echodensity of ventricular walls is virtually diagnostic of cardiac amyloidosis. However, clinically significant restrictive diastolic impairment may be difficult to detect even by comprehensive Doppler and other functional studies. Imaging after injection of isotope-labelled calcium-seeking tracers has poor sensitivity and specificity and is of no clinical use.

In cases of known or suspected hereditary amyloidosis the gene defect must be characterized. If amyloidotic tissue is available the fibril protein may be known and the corresponding gene can then be studied, but if no tissue containing amyloid is available, screening of the genes for known amyloidogenic proteins must be undertaken.

Biochemical and immunochemical screening tests for the presence in the plasma of amyloidogenic variant protein products of mutant genes also exist, for example for transthyretin and apolipoprotein A-I variants, but molecular genetic analysis of DNA is easier to perform and is the most direct approach. However, regardless of the DNA results, it is desirable, if possible, to also directly identify the respective protein in the amyloid.

Serum amyloid P component as a specific tracer in amyloidosis

The universal presence in amyloid deposits of amyloid P component, derived from circulating serum amyloid P component, is the basis for use of radioisotope-labelled serum amyloid P component as a diagnostic tracer in amyloidosis. No localization or retention of labelled serum amyloid P component occurs in healthy subjects or in patients with diseases other than amyloidosis (Fig. 1(a)). Radio-iodinated serum amyloid P component has a short half-life (24) in the plasma and is rapidly catabolized with complete excretion of the iodinated breakdown products in the urine. However, in patients with systemic or localized extracerebral amyloidosis, the tracer rapidly and specifically localizes to the deposits, in proportion to the quantity of amyloid present, and persists there without breakdown or modification (Fig. 1(b), Fig. 1(c)). For clinical purposes, highly purified serum amyloid P component is isolated from the plasma of single accredited donors and is oxidatively iodinated under conditions that preserve its function intact. The medium-energy, short half-life, pure gamma emitter ^{123}I is used for scintigraphic imaging, and the long half-life isotope ^{125}I is used for metabolic studies. The dose of radioactivity administered (less than 4 mSv) is well within accepted safety limits and more than 3000 studies have been completed without any adverse effects. In addition to high-resolution scintigraphs, the uptake of tracer into various organs can be precisely quantified and, together with highly reproducible metabolic data on the plasma clearance and whole body retention of activity, the progression or regression of amyloid can be monitored serially and quantitatively.

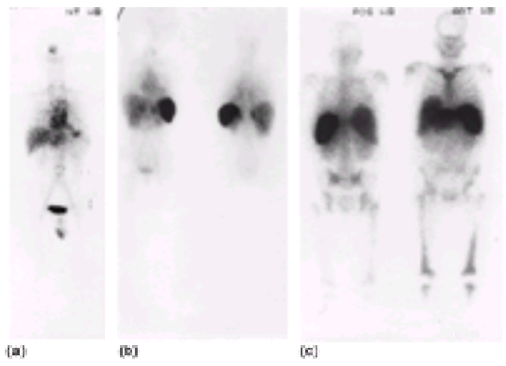

Fig. 1 Whole body scintigraphs 24 h after intravenous injection of ^{123}I-labelled human serum amyloid P component. (a) Anterior view of a normal control subject showing the distribution of residual tracer in the blood pool and radioactive breakdown products in urine in the bladder; note the absence of localization or retention of tracer anywhere in the body. (b) Posterior (left) and anterior (right) views of a patient with juvenile chronic arthritis complicated by AA amyloidosis. There is uptake of tracer in the spleen, kidneys, and adrenal glands, a typical distribution of AA amyloid in which the spleen is involved in 100 per cent of cases, kidneys in 75 per cent, and adrenals in 40 per cent. Note the reduced blood pool and bladder signal compared with (a). This patient, whose amyloid was diagnosed by renal biopsy 15 years ago when nephrotic syndrome developed, and who was then treated with chlorambucil, had been in complete remission for 10 years during which there had been no acute phase response. At the time of this scan there was no biochemical abnormality in blood or urine, despite the very appreciable amyloid deposits, illustrating the discordance between the presence of amyloid and clinical effects. (c) Posterior (left) and anterior (right) views of a patient with monoclonal gammopathy complicated by extensive AL amyloidosis. There is uptake and retention of tracer in the liver, spleen, kidneys, bone marrow, and soft tissues around the shoulder. This scintigraphic pattern of amyloid distribution is pathognomonic for AL amyloidosis; bone marrow uptake has not been seen in any other type. Note the complete absence of blood pool or bladder signal resulting from complete uptake of the tracer dose into the substantial amyloid deposits.

Important observations regarding amyloid (which have been made for the first time *in vivo*) include the following: the different distribution of amyloid in different forms of the disease; amyloid in anatomical sites not available for biopsy (adrenals, spleen); major systemic deposits in forms of amyloid previously thought to be organ-limited; a poor correlation between the quantity of amyloid present in a given organ and the level of organ dysfunction; a non-homogeneous distribution of amyloid within individual organs; and evidence for rapid progression and sometimes regression of amyloid deposits with different rates in different organs (Fig. 2). Examples of major regression of amyloidosis, when it has been possible to reduce or eliminate the supply of fibril precursor, are very encouraging. Studies with labelled serum amyloid P component thus make a valuable contribution to the diagnosis and management of patients with systemic amyloidosis, and these are available routinely for all known or suspected cases of amyloidosis in the National Health Service National Amyloidosis Centre at the Royal Free Hospital, London.

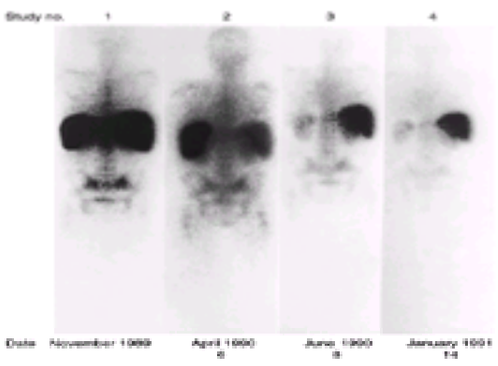

Fig. 2 Serial posterior whole body ^{123}I-serum amyloid P component scintigraphs of a man with AL amyloidosis complicating benign monoclonal gammopathy. At presentation, scan 1, there was uptake in the spleen, liver, and bone marrow, obscuring any possible renal signal. Chemotherapy was given before the second scan, which shows increased spleen uptake, reduced liver uptake, and some renal uptake, but no change in total amyloid load determined by measurements of the clearance and retention of the tracer (not shown). Subsequently he suffered from recurrent splenic infarction and splenectomy was performed. Thereafter, in scan 3, there was increased tracer uptake in liver, although a notably lower total amyloid load. Six months later, in scan 4, liver and kidney uptake, plasma clearance, and whole body retention of tracer were all reduced, indicating regression of amyloid. Clinically he was much improved and still remains well.

Management of amyloidosis

Although no treatments yet exist that specifically promote the mobilization of amyloid, there have been substantial recent advances in the management of systemic amyloidosis, in particular active measures to support failing organ function whilst attempts are made to reduce the supply of the amyloid fibril precursor protein. Serial serum amyloid P component scintigraphy in more than 1000 patients with various forms of amyloid has confirmed that control of the primary disease process, or removal of the source of the amyloidogenic precursor, usually results in regression of existing deposits and recovery or preservation of organ function. This strongly supports aggressive intervention, and relatively toxic drug regimes or other radical approaches can be justified by the poor prognosis. Such an approach, leading to reduced morbidity and improved survival, was the basis for the establishment of the National Health Service National Amyloid Centre. However, clinical improvement in amyloidosis is often delayed long after the underlying disorder has remitted, reflecting the very gradual regression of the deposits that is now recognized to occur in most patients. Continuing production of the amyloid precursor protein should be monitored as closely as possible long term, to determine the requirement for and intensity of treatment for the underlying primary condition. In AA amyloidosis this involves frequent estimation of the plasma SAA level, and in AL amyloidosis requires monitoring of proliferation of monoclonal plasma cells and immunoglobulin light chain production.

The treatment of AA amyloidosis ranges from potent anti-inflammatory and immunosuppressive drugs in patients with rheumatoid arthritis, to life-long prophylactic colchicine in familial Mediterranean fever, and surgery in conditions such as refractory osteomyelitis and the tumours of Castleman's disease. The alkylating agent chlorambucil can induce rapid and complete remission of inflammatory activity in many patients with rheumatoid and juvenile chronic arthritis, but its use must be considered very carefully since it is not licensed for this indication, it is potentially carcinogenic, and it causes infertility.

Treatment of AL amyloidosis is based on that for myeloma, although the plasma cell dyscrasias in AL amyloidosis are often very subtle. Prolonged low-intensity

cytotoxic regimes such as oral melphalan and prednisolone are beneficial in about 20 per cent of patients. Dose-intensive infusional chemotherapy regimes such as vincristine, doxorubicin (Adriamycin), and dexamethasone ('VAD'), and autologous peripheral blood stem cell transplantation are currently being evaluated with far more promising early results. However, very rigorous patient selection for transplantation is essential as the procedural mortality is high in individuals with multiple amyloidotic organ involvement, especially patients with autonomic neuropathy, severe cardiac amyloidosis, or a history of gastrointestinal bleeding, and in those aged over 55 years.

The disabling arthralgia of b_2-microglobulin amyloidosis may respond partially to non-steroidal anti-inflammatory drugs or corticosteroids, but even the most severe symptoms usually vanish rapidly following renal transplantation. The basis for this remarkable clinical response is unclear since although transplantation rapidly restores normal b_2-microglobulin metabolism, regression of b_2-microglobulin amyloid may not be evident for many years.

Hepatic transplantation is effective in familial amyloid polyneuropathy associated with transthyretin gene mutations since the variant amyloidogenic protein is produced mainly in the liver. Successful liver transplantation has now been reported in hundreds of patients with this condition and although the peripheral neuropathy usually only stabilizes, autonomic function can improve substantially and the associated visceral amyloid deposits have been shown to regress in most cases. Important questions remain about the timing of the procedure but, so far, early intervention seems advisable.

Supportive therapy remains critical in systemic amyloidosis, with the potential for delaying target organ failure, maintaining quality of life, and prolonging survival whilst the underlying process can be treated. Rigorous control of hypertension is vital in renal amyloidosis. Surgical resection of amyloidotic tissue is occasionally beneficial but, in general, a conservative approach to surgery, anaesthesia, and other invasive procedures is advisable. Should any such procedure be undertaken, meticulous attention to blood pressure and fluid balance is essential. Amyloidotic tissues may heal poorly and are liable to bleed. Diuretics and vasoactive drugs should be used cautiously in cardiac amyloidosis because they can reduce cardiac output substantially. Dysrhythmias may respond to conventional pharmacological therapy or to pacing. Replacement of vital organ function, notably dialysis, may be necessary and cardiac, renal, and liver transplant procedures have a role in selected cases.

Finally, a number of different therapies aimed specifically at inhibiting the formation of amyloid fibrils or promoting fibril regression are currently under development and will be evaluated clinically within the next few years. These approaches, directed at the generation of precursor proteins, the protein folding process, formed fibrils, glycosaminglycans, and serum amyloid P component, offer hope that in future amyloidosis may become a treatable condition.

Further reading

Booth DR *et al*. (1997). Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* **385**, 787–93.

Botto M *et al*. (1997). Amyloid deposition is delayed in mice with targeted deletion of the serum amyloid P component gene. *Nature Medicine* **3**, 855–9.

Drüeke TB (1998). Dialysis-related amyloidosis. *Nephrology Dialysis Transplantation* **13** (Suppl. 1), 58–64.

Gillmore J *et al*. (2001). Amyloid load and clinical outcome in AA amyloidosis in relation to circulating concentration of serum amyloid A protein. *Lancet* **358**, 24–9.

Hardy J (1997) Amyloid, the presenilins and Alzheimer's disease. *Trends in Neurosciences* **20**, 154–9.

Hawkins PN, Lavender JP, Pepys MB (1990). Evaluation of systemic amyloidosis by scintigraphy with ^{123}I-labeled serum amyloid P component. *New England Journal of Medicine* **323**, 508–13.

Hawkins PN *et al*. (1993). Serum amyloid P component scintigraphy and turnover studies for diagnosis and quantitative monitoring of AA amyloidosis in juvenile rheumatoid arthritis. *Arthritis and Rheumatism* **36**, 842–51.

Kyle RA, Gertz MA (1995). Primary systemic amyloidosis: clinical and laboratory features in 474 cases. *Seminars in Hematology* **32**, 45–9.

Kyle RA, Gertz MA, eds (1999). *Amyloid and amyloidosis 1998*. Parthenon Publishing, Pearl River, NY.

11.13 a_1-Antitrypsin deficiency and the serpinopathies

David A. Lomas

Introduction

People of European descent are susceptible to disease arising from a genetic deficiency of the plasma protein a $_1$-antitrypsin. This is a 394 amino-acid, 52-kDa, acute-phase glycoprotein synthesized by the liver and macrophages and present in the plasma at a concentration of between 1.5 and 3.5 g/l. It functions as an inhibitor of a range of proteolytic enzymes but its primary role is to inhibit the enzyme neutrophil elastase. Activated neutrophil leucocytes release elastase to break down connective tissue at sites of inflammation. This breakdown is limited by the antielastase activity of a $_1$-antitrypsin, but if the plasma concentration of this protein falls below 40 per cent then unimpeded tissue destruction may ensue.

Genetic deficiency

a_1-Antitrypsin is subject to genetic variation resulting from mutations in the 12.2-kilobase (**kb**), 7-exon gene on the long arm of chromosome 14. Over 90 allelic variants have been reported and classified using the **Pi** (protease inhibitor) nomenclature that assesses a $_1$-antitrypsin mobility in isoelectric focusing analysis. Normal a_1-antitrypsin migrates in the middle (M) and variants are designated *A* to *L* if they migrate faster than M, and *N* to *Z* if they migrate more slowly. Many of these variants have been sequenced at the DNA level and shown to result from point mutations in the a $_1$-antitrypsin gene. For example, the *Z* allele results from the substitution of a positively charged lysine for a negative glutamic acid at position 342. The *S* allele results from the substitution of a neutral valine for a glutamic acid at position 264. It is clear that such mutations alter the overall charge of the protein and explain the changes in mobility seen on isoelectric focusing. Point mutations are inherited by simple Mendelian trait; the normal genotype is designated *PiMM* or *PiM*, a heterozygote for the *Z* gene is *PiMZ*, and a homozygote is *PiZZ* or *PiZ*.

The most clinically relevant variants are the *S* and *Z* alleles and the uncommon *Null* (non-production) gene. Approximately 8 per cent of people of Northern European descent are heterozygotes for the *S* variant (*PiMS*), although this can be as high as 28 per cent in parts of Southern Europe. The *Z* variant is less common and is found in 4 per cent of Northern Europeans (*PiMZ*), with 1 in 1700 being homozygotes (*PiZ*). a_1-Antitrypsin alleles are co-dominantly expressed, with each allele contributing to the plasma level of protein. Moreover, each of the deficiency alleles results in a characteristic decrease in the plasma concentration of a $_1$-antitrypsin; the *S* variant forms 60 per cent of the normal *M* concentration and the *Z* variant 10 to 15 per cent. Thus combinations of alleles have predictable effects, the *MZ* heterozygote has an a_1-antitrypsin plasma level of 60 per cent (50 per cent from the normal *M* allele and 10 per cent from the *Z* allele), the *MS* heterozygote 80 per cent, and the *SZ* heterozygote 40 per cent.

Molecular basis of a_1-antitrypsin deficiency

a_1-Antitrypsin functions by presenting its reactive-centre methionine residue on an exposed loop of the molecule such that it forms an ideal substrate for the enzyme neutrophil elastase (Fig. 1). The exact fit between enzyme and inhibitor causes them to form a tightly bound 1:1 complex that inhibits the enzyme and allows it to be eliminated from sites of inflammation. The *Z* mutation (342glutamic acid®lysine) results in normal translation of the gene, but 85 per cent of the *Z* a_1-antitrypsin is retained within the endoplasmic reticulum with only 10 to 15 per cent entering the circulation. The *Z* mutation distorts the relationship between the loop and the b-pleated A sheet that forms the major feature of the molecule (Fig. 1(a)). The consequent perturbation in structure allows the reactive-centre loop of one a $_1$-molecule to lock into the A sheet of a second (Fig. 1(b)) to form a dimer which then extends to form chains of loop-sheet polymers (Fig. 1(c)). The formation of these polymers is temperature- and concentration-dependent and is localized in the endoplasmic reticulum of the hepatocyte. These chains of polymers become interwoven to form the insoluble aggregates that are the hallmark of a $_1$-antitrypsin liver disease (Fig. 2 and Plate 1). *S* a_1-antitrypsin (342glutamic acid®lysine) and the rare *I* variant (39arginine®cysteine) also result in the formation of polymers but at a much slower rate than *Z*. Therefore these variants do not accumulate in the liver and cause only mild plasma deficiency.

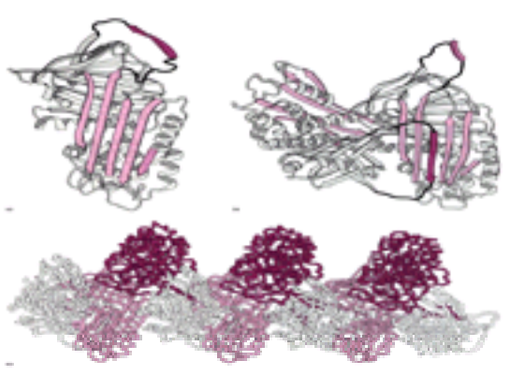

Fig. 1 The crystal structure of a $_1$-antitrypsin shows the reactive-centre loop (purple) held at the apex of the protein as a b-strand depicted as an arrow (a) The *Z* mutation opens the b-sheet A (pink) to allow the reactive loop of another molecule to insert to form a dimer (b) which can then extend to form long chains of polymers (c). In (c) white, pink, and purple represent different a $_1$-antitrypsin molecules linked together to form a polymer. (Figure prepared by Dr T. Dafforn, Cambridge Institute for Medical Research, and reproduced from *Nature Structure Biology* with permission.)

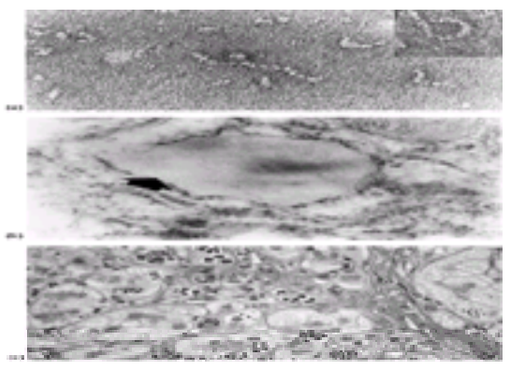

Fig. 2 (a) Electron microscopy of a chain of loop-sheet polymers isolated from a patient with a $_1$-antitrypsin deficiency. These polymers can form filaments or circlets

(inset) that tangle within the endoplasmic reticulum of the hepatocyte (b) to form the inclusions (arrowed) which are the hallmark of the disease. These intrahepatic inclusions are characteristically Periodic acid–Schiff (**PAS**)-positive and diastase-resistant (c) (see also Plate 1) and stain positive for a_1-antitrypsin on immunohistochemistry. ((a) and (b) reproduced from the *Journal of Biological Chemistry* and *Nature*, respectively, with permission.)

Clinical features

a_1-Antitrypsin deficiency and emphysema

The association between a_1-antitrypsin deficiency and the development of premature panlobular emphysema was first described by Laurell and Eriksson in 1963. Patients usually present with increasing dyspnoea and weight loss, with cor pulmonale and polycythaemia occurring late in the course of the disease. Chest radiographs typically show bilateral basal emphysema with paucity and pruning of the basal pulmonary vessels. Upper lobe vascularization is relatively normal. Ventilation–perfusion radioisotope scans and angiography also show abnormalities with a lower zone distribution. High-resolution computed tomography scans with 1 to 2 mm collimation are the most accurate method of assessing the distribution of panlobular emphysema and for monitoring the progress of the pulmonary disease, although this currently has little value outside clinical trials. Lung function tests are typical for emphysema with a reduced **FEV1/FVC** ratio (forced expiratory volume in 1 second/forced vital capacity), gas trapping (raised residual volume/total lung capacity ratio), and a low gas-transfer factor.

The association of a_1-antitrypsin with the development of premature emphysema has led to the wider conclusion that emphysema results from an imbalance between proteases and antiproteases within the lung. Undoubtedly the situation is more complex than a simple balance between elastase and a_1-antitrypsin, both in terms of the numbers of enzymes and inhibitors involved and the contribution of other mechanisms. Nevertheless, the elastase and a_1-antitrypsin balance clearly illustrates the processes involved in the development of emphysema and the interplay between the environmental and genetic factors that determine its onset.

Decline in lung function in individuals with a_1-antitrypsin deficiency

As with other tissues, there is a decline in the elasticity of the lungs with increasing age. Clinically, the most convenient measure of lung function is the FEV_1, which is approximately 3500 ml in young adults. After the age of 30 years in healthy non-smokers, the FEV_1 decreases by 35 ml/year, although there is considerable individual variation. By old age, most people will have an appreciable loss of lung function, but only occasionally in the non-smoker will this be clinically apparent. The assessment of symptomatic hospital patients has shown that the loss of FEV_1 may be accelerated to 80 ml/year in a *Z* a_1-antitrypsin homozygote. As a consequence there is a hastened but still variable onset of emphysema. In this study *PiZ* non-smokers were free from dyspnoea up to the age of 50 years, with the average age of death from respiratory disease being 67 years. Again there was considerable individual variation and, particularly in women, there was a good likelihood of a full lifespan without significant respiratory impairment. The outlook, however, was poor for the *PiZ* a_1-antitrypsin homozygote who was a heavy smoker, as the loss in FEV_1 increased to 300 ml/year. The onset of dyspnoea was approximately 30 years, with death from respiratory disease by the age of 50 years. More recently the Swedish, Danish, and NIH registries have reported a more favourable outcome. These registries include individuals identified by screening and family studies and are more representative of the disease process. They show a slower rate of decline in lung function in *PiZ* homozygotes who are non- or ex-smokers (approximately 50 ml/year). However, the studies reinforce the accelerated rate of decline in lung function in *PiZ* homozygotes who continue to smoke (70–132 ml/year).

a_1-Antitrypsin deficiency and liver disease

Z a_1-antitrypsin liver disease is characterized by the accumulation of diastase-resistant, Periodic acid–Schiff (**PAS**)-positive inclusions of a_1-antitrypsin in the periportal cells (Fig. 2(c)). This insoluble material accumulates within the endoplasmic reticulum of hepatocytes stimulating a massive increase in cellular degradative activity. The *PiMZ* individuals are able to degrade much of the abnormal a_1-antitrypsin, but not the *PiZ* homozygote in whom aggregation overwhelms the degradative process resulting in a_1-antitrypsin accumulation, hepatocellular damage, and cell death. The accumulation of a_1-antitrypsin within hepatocytes is also seen with two other rare mutations, S_{iiyama} (53phenylalanine®serine), which is the commonest cause of a_1-antitrypsin deficiency in Japan, and M_{malton} (52phenylalanine deletion), which is the commonest cause of a_1-antitrypsin deficiency in Sardinia. Both of these point mutations result in perturbations of a_1-antitrypsin and the ready formation of loop-sheet polymers. Cirrhosis has also been reported sporadically in *SZ* and *IZ* heterozygotes in whom the 'polymerogenic' *Z* and *S* or *I* a_1-antitrypsin can interlink to form chains of mixed SZ or IZ heteropolymers. The observation that polymer formation is temperature- and concentration-dependent may account for the variation in the number and density of liver inclusions between individuals. a_1-Antitrypsin is an acute-phase protein and, as such, undergoes a manifold increase in production in association with temperature increases of up to 41 °C. The increase in protein concentration and temperature during the inflammatory response favour polymerization, which in turn leads to inclusion formation and liver disease.

Neonatal jaundice and juvenile cirrhosis

Some 73 per cent of *Z* a_1-antitrypsin homozygote infants have a raised serum alanine aminotransferase level in the first year of life, but in only 15 per cent of people is it still abnormal by the age of 12 years. Similarly, the serum bilirubin level is raised in 11 per cent of *PiZ* infants in the first 2 to 4 months but falls to normal by 6 months of age. Cholestatic jaundice develops in 1 in 10 infants and 6 per cent develop clinical evidence of liver disease without jaundice. These symptoms usually resolve by the second year of life, but approximately 15 per cent of patients with cholestatic jaundice progress to juvenile cirrhosis. The reasons for this variable progression are unknown, but intercurrent illness and hormonal and genetic factors are likely to be involved. Indeed cholestatic jaundice in infancy is twice as common in boys than girls. The overall risk of death from liver disease in *PiZ* children during childhood is between 2 and 3 per cent.

Adult liver disease

All *PiZ* individuals have slowly progressive hepatic damage that is often subclinical and only evident as a minor degree of portal fibrosis. However, up to 50 per cent of *Z* a_1-antitrypsin homozygotes present with clinically evident cirrhosis and occasionally with hepatocellular carcinoma. The presence of *Z* a_1-antitrypsin deficiency, including the heterozygous *PiMZ* and *PiSZ* forms, should always be considered before making the diagnosis of cryptogenic cirrhosis.

Associated conditions

a_1-Antitrypsin deficiency is associated with an increased prevalence of asthma, panniculitis, Wegener's granulomatosis, pancreatitis, gallstones, and possibly bronchiectasis. There appears to be a reduced risk of cerebrovascular disease.

Diagnosis

The severe genetic deficiency of a_1-antitrypsin is readily diagnosed by low plasma levels and the virtual absence of the a_1-band on protein electrophoresis. As a_1-antitrypsin is an acute-phase protein, most laboratories will report levels with another acute-phase reactant, such as a_1-antitchymotrypsin, which allows the clinician to assess the likelihood of deficiency in the context of the inflammatory response. The acute-phase response raises the plasma level of a_1-antitrypsin, but never can the plasma level of the *PiZ* heterozygote reach the normal range. The deficiency variant is then assigned a Pi phenotype according to the migration of the protein on an isoelectric focusing gel.

Treatment

The treatment of a_1-antitrypsin deficiency depends largely on the avoidance of stimuli causing repeated pulmonary inflammation—primarily smoking. Patients with a_1-antitrypsin deficiency-related emphysema should receive conventional therapy with trials of bronchodilators and inhaled corticosteroids, pulmonary rehabilitation and, where appropriate, assessment for long-term oxygen therapy and lung transplantation. The role of lung volume-reduction surgery in this group is unclear as the

disease is basal rather than apical and resections of this region are technically more difficult. Mixed results have been reported in uncontrolled trials.

The lung disease results from a deficiency in the antielastase screen. This may be rectified biochemically by intravenous infusions of a $_1$-antitrypsin. Registry data suggest that individuals with a$_1$-antitrypsin deficiency and an FEV$_1$ of 35–49 per cent predicted may derive benefit from replacement therapy. The only controlled trial showed a non-significant trend towards reduced progression of emphysema in individuals receiving intravenous a $_1$-antitrypsin. a$_1$-Antitrypsin replacement therapy is currently unavailable in many European countries, including the United Kingdom. It is widely used in North America.

All *Z* homozygotes have some liver damage and, as such, would be wise to avoid alcohol abuse. The deduction that loop-sheet polymerization of a $_1$-antitrypsin complicates the acute-phase response highlights the importance of antipyretic agents in *PiZ* infants with antitrypsin deficiency. Although this has yet to be proven by clinical trials, there is anecdotal evidence that these intercurrent illnesses account for the variation in progression of liver disease in infants. Moreover, there is good reason to believe that conservative treatments to lessen pyrexia and the inflammatory response will be of value in reducing a $_1$-antitrypsin aggregation within hepatocytes and hence liver disease. *PiZ* homozygotes should be monitored for the persistence of hyperbilirubinaemia as this, along with deteriorating results of coagulation studies, indicates the need for liver transplantation. Parents with a child with severe *Z* a$_1$-antitrypsin liver disease may require genetic counselling. The likelihood of similar severe liver damage in a subsequent *Z* homozygote sibling is approximately 20 per cent.

The uncommon a$_1$-antitrypsin deficiency-associated panniculitis usually responds to dapsone, 100 to 150 mg daily, for 2 to 4 weeks, but occasionally it necessitates the administration of intravenous a$_1$-antitrypsin replacement therapy.

Other 'serpinopathies'

a$_1$-Antitrypsin is the archetypal member of a superfamily of proteins termed the *ser*ine *p*rotease *in*hibitors, or serpins, that have closely related structures and functions. These inhibitors control various inflammatory cascades, including coagulation (antithrombin), complement activation (C1-inhibitor), and fibrinolysis (a$_2$-antiplasmin). Pathological processes that underlie the deficiency of one member may account for deficiency of others. Indeed the process of polymer formation has also been reported in deficiency-mutants of antithrombin, C1-inhibitor, and a $_1$-antichymotrypsin. These polymers are inactive as proteinase inhibitors and so predispose the individual to thrombosis, angio-oedema, and chronic airflow obstruction disease, respectively. Moreover, polymerization also underlies a novel inclusion-body dementia that results from point mutations in a neurone-specific serpin, neuroserpin. The dementia, termed familial encephalopathy with neuroserpin inclusion bodies or FENIB, is inherited as an autosomal dominant trait with the inclusions of neuroserpin in the brain being PAS-positive and diastase-resistant, identical to those of *Z* a$_1$-antitrypsin in the liver. Heterozygotes with critical mutations develop early-onset dementia as the accumulated protein causes neuronal cell death. The recognition that serpin polymerization underlies all these disorders may allow the development of a common therapy.

Further reading

Davis RL *et al.* (1999). Familial dementia caused by polymerisation of mutant neuroserpin. *Nature* **401**, 376–9. [Description of two families with mutations in the serpin, neuroserpin, that form polymers *in vivo* and an inclusion-body dementia.]

Dirksen A, *et al*. (1999). A randomised clinical trial of a $_1$-antitrypsin augmentation therapy. *American Journal of Respiratory and Critical Care Medicine*, **160**, 1468–72. [The only randomised controlled trial of a$_1$-antitrypsin replacement therapy in patients with a$_1$-antitrypsin deficiency. The study showed a trend towards a reduced rate of progression of emphysema in patients receiving replacement therapy.]

Eriksson S, Carlson J, Velez R. (1986). Risk of cirrhosis and primary liver cancer in alpha $_1$-antitrypsin deficiency. *New England Journal of Medicine* **314**, 736–9. [Postmortem study demonstrating a high prevalence of liver disease in adults with PiZ a1-antitrypsin deficiency.]

Larsson C (1978). Natural history and life expectancy in severe a $_1$-antitrypsin PiZ. *Acta Medica Scandinavica* **204**, 345–52. [Report of 246 patients with PiZ a1-antitrypsin deficiency detailing age of onset of breathlessness. The subjects were largely ascertained from hospital populations.]

Mahadeva R, Lomas DA (1998). Alpha $_1$-antitrypsin deficiency, cirrhosis and emphysema. *Thorax* **53**, 501–5. [A review of the structural basis of a1-antitrypsin deficiency.]

Mahadeva R, *et al.* (1999). Heteropolymerisation of S, I and *Z* a$_1$-antitrypsin and liver cirrhosis. *Journal of Clinical Investigation* **103**, 999–1006. [Demonstration that different a1-antitrypsin variants can interlink to form polymers that are associated with cirrhosis.]

Piitulainen E, Eriksson S (1999). Decline in FEV$_1$ related to smoking status in individuals with severe alpha1-antitrypsin deficiency. *European Respiratory Journal* **13**, 247–51. [Report of rate of decline in lung function of 608 patients followed for 1–31 years. Current smokers have an accelerated rate of decline in lung function but ex-smokers have the same rate as non-smokers. The values are likely to be more representative than other reports as many subjects were ascertained from screening and family studies.]

Sveger T, Piitulainen E, Arborelius Jr M (1995). Clinical features and lung function in 18-year-old adolescents with a $_1$-antitrypsin deficiency. *Acta Pædiatrica* **84**, 815–16. [Report on the follow up of 127 subjects with PiZ a1-antitrypsin deficiency from birth to 18 years. This is the only long-term prospective study of patients with a1-antitrypsin deficiency and the only study that is free from selection bias.]

The alpha-1-antitrypsin deficiency registry study group (1998). Survival and FEV$_1$ decline in individuals with severe deficiency of a $_1$-antitrypsin. *American Journal of Respiratory and Critical Care Medicine* **158**, 49–59. [Report from the registry of 1129 patients with a1-antitrypsin deficiency who did or did not receive a1-antitrypsin replacement therapy. Replacement therapy may slow down the decline in lung function in patients with a predicted FEV1 of 35–49 per cent. This is not a randomized controlled trial and therefore the conclusions must be interpreted with caution.]

Useful web sites:

http://www-structmed.cimr.cam.ac.uk/serpins.html [An updated list of a1-antitrypsin mutants, their clinical effects, and their effect on the structure of the protein.]

http://www.alpha-1.priv.at/supportg.html [A list of international a1-antitrypsin support groups and other related web sites.]

12.1 Principles of hormone action

Mark Gurnell, Jacky Burrin, and V. Krishna K. Chatterjee

Definition

Endocrinology is the study of hormones secreted by glands or cells which, acting locally or at a distance, facilitate communication between cells and different organs thus co-ordinating their activities.

Classically, the production of hormones has been associated with specialized glands or tissues including the hypothalamus, pituitary, thyroid, parathyroids, gonads, pancreatic islet cells, adrenal glands, and placenta. It is now recognized that hormones are also produced by a range of other organs and tissues which are not considered to be classical endocrine glands. The heart is the primary source of atrial natriuretic peptide factor, which controls blood pressure and intravascular volume; endothelin and nitric oxide are derived from vascular endothelium and regulate vascular tone. Endocrine cells are distributed throughout the gastrointestinal tract and are a rich source of hormones such as cholecystokinin, gastrin, secretin, vasoactive intestinal peptide; many of these gastrointestinal hormones are also produced in the brain and central nervous system, where their role is less well understood. Erythropoietin, a circulating factor that stimulates erythropoiesis, is derived from the kidney. Adipose tissue has recently been shown to produce leptin, a circulating hormone which acts centrally to control appetite.

However, as understanding of intercellular communication has advanced, the lines of division that separate different physiological systems have become blurred. For example, neuroendocrinology represents intimate connections between the nervous and endocrine systems: peptide hormones produced in the brain exert effects via the hypothalamus to control hormone secretion from the pituitary gland; in the periphery, the sympathetic nervous system modulates hormone production by the adrenal medulla and pancreatic islets. Similarly, there are complex inter-relationships between the immune and endocrine systems: for example, glucocorticoid hormones exert powerful immunosuppressive effects; conversely, cytokines (for example, tumour necrosis factor-a and interleukin-6) produced by cells of the immune system markedly influence hormone secretion by glands such as the pituitary and adrenal.

Nature of hormones

In general, hormones can be classified into those that are based on proteins or peptides and those which are chemically derived. Small peptides include hypothalamic releasing factors produced by neuroendocrine cells, which act locally on the pituitary; larger polypeptides such as insulin or growth hormone are characteristically circulating hormones that act on more distant targets. Biogenic amines including catecholamines, dopamine, and serotonin (5-hydroxytryptamine) are derived from amino acids. The majority of protein and peptide hormones interact with membrane receptors located on the cell surface. Binding to membrane receptors activates downstream signalling pathways, leading to changes in cellular function which mediate responses to hormones.

A second class of hormones includes steroids and other lipophilic substances, which act by crossing the plasma membrane to interact with intracellular receptors. Steroid hormones are derived from cholesterol and include cortisol, progesterone, testosterone, and oestradiol. Vitamin D and retinoic acid, which are synthesized from dietary sources, and thyroid hormone produced by the modification of tyrosines in thyroglobulin, are structurally dissimilar to steroids but also act via nuclear receptors.

Development of endocrine glands

The hypothalamus develops from forebrain tissue adjacent to the third ventricle. Neurones secreting releasing factors send cellular processes which terminate in portal capillaries that perfuse the pituitary gland. The latter develops from ectoderm to form the adenohypophysis or anterior pituitary; the posterior pituitary or neurohypophysis is formed directly from axonal terminals of hypothalamic neurones that grow downward. The thyroid gland develops from endoderm in the floor of the oropharynx with the migration of cells caudally to its final position in the neck. During its descent, parafollicular C cells, derived from neural crest tissue within the ultimobranchial body and parathyroid glands from the third and fourth pharyngeal pouches, become incorporated into the thyroid gland. The adrenal glands comprise a steroid-secreting cortex developed from mesoderm, together with a catecholamine-producing medulla composed of chromaffin cells derived from neural crest. Germ cells within indifferent gonadal primordia differentiate to form the ovary, or in the presence of the Y chromosome-encoded sex determining gene (*SRY*), develop into testes. Endocrine cells of the pancreas are derived from endoderm and differentiate to form the islets of Langerhans. Various transcription factors that control the development of cells within endocrine glands and their differentiation to hormone biosynthesis are listed in Table 1.

Hormone synthesis, processing, and secretion

The organization of endocrine genes is homologous to those encoding many other proteins, although there are some characteristic features. Gene transcription is usually regulated by the promoter located in the upstream 5'-flanking region of the gene (Fig. 1). Typically, the promoter may contain three types of regulatory DNA sequence which are recognized by specific transcription factors: a hormone-response element (**HRE**) is recognized by nuclear receptors; a tissue-specific element (**TSE**) binds cell-specific transcription factors (see Table 1) which enhance the transcription of the hormone gene in a tissue-specific manner; a third class of response element mediates transcriptional activation in response to second messenger signalling pathways. A rise in intracellular cyclic AMP concentration leads to the phosphorylation of cyclic AMP response-element binding proteins (**CREBs**) which interact with CREs; cell-signalling pathways that activate protein kinase C induce the phosphorylation of the Fos–Jun (AP-1) transcription-factor complex which binds its cognate DNA regulatory sequence. Binding of transcription factors to regulatory DNA response elements, activates and stabilizes basal transcription factors (**BTFs**), promoting gene transcription and mRNA synthesis (Fig. 1).

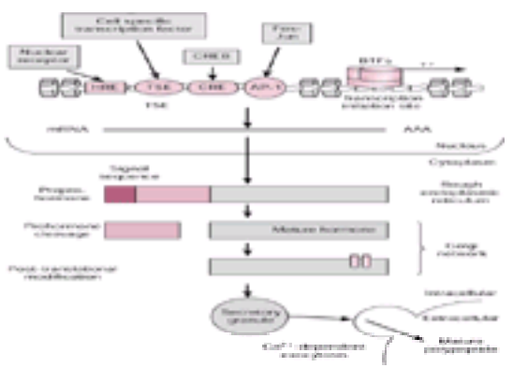

Fig. 1 Pathway of hormone synthesis, processing, and secretion.

Transcription of the gene generates mRNA, which undergoes translation in ribosomes leading to polypeptide synthesis. In some endocrine genes, alternate exon

splicing allows the substitution or removal of particular exons, such that peptides of differing sequence can be produced. For example, alternate splicing of the calcitonin gene in a tissue-specific manner directs the production of calcitonin in the C cells of the thyroid, whereas calcitonin gene-related peptide (**CGRP**) is produced preferentially in the brain.

Secreted polypeptide hormones incorporate a signal sequence at the amino terminus of the protein that directs its translocation across the endoplasmic reticulum where this sequence is cleaved (see Fig. 1). Many hormones are synthesized as larger polypeptides (prohormones) which undergo proteolytic cleavage to generate smaller functional peptides. Such proteolytic processing is mediated by specific proteases (prohormone convertase 1 and 2 (**PC1**, **PC2**)) which are highly expressed in cells of neuroendocrine lineage. Examples of hormone processing include the cleavage of proinsulin with removal of an internal C peptide to yield insulin, the active hormone. However, processing of the polypeptide precursor can also yield multiple functioning products. For example, pro-opiomelanocortin is cleaved by endopeptidases to yield adrenocorticotrophin (**ACTH**), melanocyte-stimulating hormone (**MSH**-a, -b, -g), b-endorphin, and lipocortin.

Hormones may also undergo post-translational modification such as amidation of neuropeptides or glycosylation. Modification of amino acids by the addition of carbohydrate side chains is a particular characteristic of the glycoprotein hormones (luteinizing hormone, follicle-stimulating hormone, thyroid-stimulating hormone, and human chorionic gonadotrophin) and such glycosylation affects both their biological activity as well as their half-life in the circulation (see Fig. 1).

Hormones such as growth factors and cytokines are not significantly concentrated within cells, but are released via small, clear, Golgi-derived transport vesicles that fuse with the plasma membrane, representing a 'constitutive' pathway of secretion. In contrast, many endocrine cells contain an additional 'regulated' secretory pathway, which allows the export of high concentrations of hormone stored in cytoplasmic dense-core vesicles. Chromogranin B, an acidic protein, and polypeptide proteases are additional constituents of secretory vesicles. Adrenal cells secreting catecholamine hormones contain chromaffin granules which include enzymes (for example, dopamine b-hydroxylase) that catalyse catecholamine biosynthesis. Dense-core vesicle exocytosis is mediated by a rise in intracellular calcium levels, which activates the cytoskeletal machinery, promoting vesicle translocation and docking with the plasma membrane (see Fig. 1). Cells secreting steroid hormones contain abundant mitochondrial and smooth endoplasmic reticulum which contain enzymes that mediate steroid biosynthesis. Mitochondrial side-chain cleavage enzyme converts cholesterol to pregnenolone. The latter is then converted to glucocorticoid, mineralocorticoid, or sex steroids dependent on the cell-specific expression of steroidogenic enzymes. Steroid hormones are not stored to any extent and are secreted constitutively.

Control of hormone production

The classical mechanism by which hormone-producing glands communicate is via **endocrine** pathways, whereby the products from one gland are secreted into the circulation (and exert effects on a different, distant target gland). Such endocrine pathways integrate the hypothalamus, pituitary, and various end-organs to control the production of major hormones (Fig. 2). Thus, peptide-releasing factors (for example, gonadotrophin-releasing hormone (**GnRH**), thyrotrophin-releasing hormone (**TRH**), growth-hormone releasing hormone (**GHRH**), and corticotrophin-releasing hormone (**CRH**)) from the hypothalamus, stimulate the production of trophic hormones from specific pituitary cell types; exceptions to this are somatostatin, which inhibits pituitary growth-hormone release, and dopamine, which is secreted continuously to inhibit prolactin secretion. The pituitary hormones act on end-organs to generate products, which, in turn, exert a negative feedback effect at both hypothalamic and pituitary levels to regulate their own synthesis. Tri-iodothyronine (T3) inhibits TRH and thyroid-stimulating hormone (**TSH**) production; gonadal steroids and inhibin negatively regulate hypothalamic GnRH and pituitary gonadotrophins; cortisol suppresses CRH and ACTH generation; circulating insulin-like growth factor-1 (**IGF-1**) inhibits GHRH and growth-hormone secretion (Fig. 2). Osmoreceptors in the hypothalamus sense changes in serum osmolality to control the release of vasopressin from the posterior pituitary.

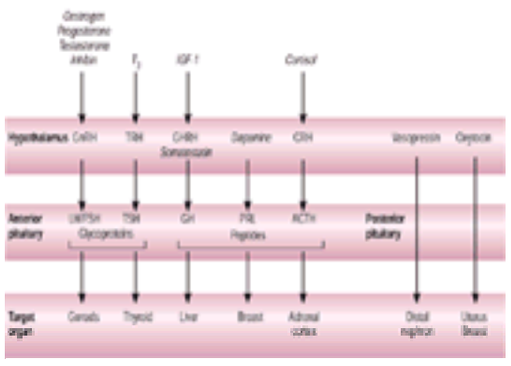

Fig. 2 Control of hormone production. Regulatory pathways integrating the hypothalamus, pituitary, and various end-organs. Hormones shown in italics exert inhibitory effects. Negative feedback regulation occurs at both hypothalamic and pituitary levels.

In addition to these endocrine control mechanisms, other types of local regulatory pathways are recognized. **Paracrine** regulation refers to factors that are released by one cell and act upon a nearby cell in the same tissue. For example, somatostatin produced by † cells in pancreatic islets inhibits the local production of insulin from b cells; in the testis, testosterone produced from Leydig cells exerts an effect on nearby Sertoli cells to enhance spermatogenesis. **Autocrine** control refers to a factor that acts upon the same cell in which it is produced. Examples include gonadotroph secretion of activin which stimulates the production of follicle-stimulating hormone (**FSH**) from the same cell; similarly, T cells produce interleukin-2 which acts to promote their own proliferation.

In addition to discrete hormonal responses, endocrine systems can respond to environmental stimuli by the integrated production of multiple hormones. For example, stress activates an array of pathways, with sympathetic activation mediating catecholamine release from the adrenals, and stimulation of the hypothalamus inducing multiple axes, resulting in the production of cortisol, growth hormone, prolactin, and vasopressin. The hormonal responses to starvation are also integrated by the hypothalamus. Here, diminished production of leptin rom adipose tissue inhibits hypothalamic GnRH and TRH secretion, with a consequent reduction in the production of both gonadal steroids and thyroid hormone to limit reproduction and energy expenditure.

In addition to the feedback regulatory mechanisms outlined above, many hormones are released in a rhythmic or pulsatile manner. Insulin is secreted in rapid (about every 10 min) pulses in response to changes in glucose concentration in the pancreatic b cell. Gonadotrophin-releasing hormone is secreted from the hypothalamus at a lower pulse frequency of every 1.5 to 3 h, stimulating similar pulses of pituitary luteinizing hormone (**LH**) and FSH release. This hormonal rhythm controls ovarian folliculogenesis and steroid production to establish the female reproductive and menstrual cycle. Pituitary growth-hormone secretion is regulated by pulses of stimulatory GHRH and inhibitory somatostatin from the hypothalamus that are out of phase with each other, corresponding to peaks and troughs of circulating growth hormone.

Many hormonal pathways are influenced by the light–dark cycle, with circadian variation in their circulating levels. For example, the hypothalamic–pituitary–adrenal axis exhibits most activity in the early morning with peak cortisol production, followed by a nadir in glucocorticoid levels in the evening. Sleep is another environmental regulator: puberty is associated with nocturnal surges of LH; growth-hormone secretion is also enhanced nocturnally and the release of vasopressin during sleep inhibits renal diuresis.

Hormone-binding proteins

Thyroid hormones and many steroids are transported in the circulation with serum-binding proteins. Thus, thyroxine (**T4**) and tri-iodothyronine (**T3**) are bound to thyroxine-binding globulin (**TBG**), albumin, and thyroxine-binding prealbumin. Cortisol and progesterone are bound to cortisol-binding globulin (**CBG**), while oestrogens and androgens are bound to sex-hormone-binding globulin (**SHBG**). The role of serum-binding proteins is to provide a reservoir of circulating hormone. The interaction of hormones with binding proteins is relatively weak compared to their affinity for receptors, enabling them to dissociate easily. Only free hormone interacts with its receptor to elicit a biological response. Hormone-binding proteins are produced by the liver, and their synthesis can be increased (for example, by oestrogens or in pregnancy) or decreased (for example, in liver disease) thereby affecting the circulating concentration of total hormones. Accordingly, wherever possible, the concentration of free hormones in the circulation (for instance, T4, T3) or urine (cortisol) should be measured. Some protein hormones also circulate associated with binding proteins, which may modulate their action. A range of insulin-like growth-factor binding proteins (**IGFBPs**) bind to IGF-1, with some inhibiting and others facilitating the action of this peptide on target-tissue receptors. Growth hormone circulates bound to the extracellular domain of its receptor derived by

cleavage from the membrane, with the complex prolonging the circulating half-life of the hormone.

Functions of hormones

The physiological roles of the major hormones can be broadly classified into three areas: control of growth and differentiation; maintenance of homeostasis; and regulation of reproduction. Some hormones have multiple functions and play a role in more than one area. In addition, some biological effects are mediated by the combined action of several different hormonal pathways. The principal actions of the major hormones are outlined in Table 2.

Linear growth is dependent on a complex interplay of many hormones and growth factors. Growth hormone plays a key role and exerts many of its effects by stimulating the hepatic production of IGF-1. Thyroid hormone also stimulates the epiphyseal growth plate in childhood, whereas production of sex steroids at puberty leads to epiphyseal closure. Other important actions of thyroid hormone include regulation of the basal metabolic rate, enhancement of myocardial contractility, and differentiation of the central nervous system.

The maintenance of homeostasis includes the control of metabolic pathways, fluid, electrolyte, and calcium balance, and regulation of blood pressure. Metabolic effects are mediated by several hormones: insulin lowers blood glucose by enhancing its cellular uptake and promotes glycogen synthesis; conversely, growth hormone, cortisol, glucagon, and adrenaline (epinephrine) act as counter-regulatory hormones to raise blood glucose. Glucagon and adrenaline stimulate glycogenolysis and, together with cortisol, promote gluconeogenesis. Other metabolic pathways are also influenced by these hormones: growth hormone and cortisol are lipolytic, whereas insulin mediates lipogenesis; insulin and growth hormone are also anabolic by promoting protein biosynthesis, whereas cortisol increases protein breakdown.

Circulating concentrations of ions and water balance are also under hormonal control. Vasopressin promotes water reabsorption via membrane channels (aquaporins) in the distal collecting ducts of the kidney; aldosterone acts at the renal distal convoluted tubule to stimulate sodium reabsorption and potassium excretion. Both parathyroid hormone (**PTH**) and vitamin D increase serum calcium levels; PTH mediates Ca^{2+} resorption from bone and kidney, whereas vitamin D acts on the gastrointestinal tract as well as these sites. Catecholamines and angiotensin-2 are potent vasoconstrictors and, together with cortisol, control blood pressure.

Hormones involved in reproduction exert effects from early in development. During embryogenesis, Müllerian-inhibiting substance (**MIS**) from the testis causes regression of female structures (uterus, fallopian tube), and testosterone promotes the development of the male structures (vas deferens, epididymis, seminal vesicles) derived from the Wolffian duct. Dihydrotestosterone promotes the development of male external genitalia. In both sexes, the gonadal axes are quiescent in childhood and become reactivated at puberty. Testosterone mediates virilization, secondary sexual characteristics, and spermatogenesis in the male; in females, ovarian production of oestrogen and progesterone induce secondary sexual features and control the menstrual cycle. In both sexes, gonadal steroids are required for the attainment of peak bone density at the end of puberty and its subsequent maintenance. During pregnancy, prolactin acts in concert with oestrogen to promote lactation; oxytocin stimulates uterine contraction at parturition and smooth muscle contraction in the mammary gland during suckling.

Hormone action

Hormones induce biological responses by interacting with receptors located either on the membrane or intracellularly in the cytoplasm or nucleus. Hormones bind to receptors with high affinity, such that low concentrations of free hormone associate and dissociate from receptors rapidly in a dynamic equilibrium. The interaction of hormones with receptors is usually highly specific, with individual receptors being highly selective for a single hormone even within a class of structurally related molecules (for example, steroid hormones). However, there are exceptions to this: parathyroid hormone (**PTH**) and parathyroid hormone-related peptide (**PTHRP**) share a common receptor and luteinizing hormone and chorionic gonadotrophin share another, generating similar biological responses; insulin and IGF-1 exhibit some degree of crossreactivity with their respective receptors; the mineralocorticoid receptor binds cortisol with equal or higher affinity than aldosterone.

Hormones that bind to membrane receptors act via effector proteins to activate second messenger signalling pathways. In turn, the second messengers stimulate a cascade of kinases, which then act upon target substrates in the cell membrane, the cytoplasm, or nucleus, to alter gene transcription or modulate a biochemical pathway, leading to a physiological response. Hormones that act through nuclear receptors are transported passively, or pumped actively, across the plasma membrane to interact with their targets. The hormone–receptor complex interacts with DNA sequences in target genes to either stimulate or repress their expression. The cellular actions of nuclear receptors are mediated by changes in target-gene transcription, altering mRNA synthesis, and in turn, the levels of protein product.

Signalling by membrane receptors

Membrane receptors can be divided into several groups (Table 3) depending on the signalling pathways that they utilize. The largest group consists of receptors with multiple transmembrane domains which are coupled to G-proteins; a second class of receptor contains an intracellular domain with tyrosine kinase activity; a number of hormones signal via membrane proteins that are homologous to cytokine receptors; a fourth class of hormone receptor contains an intracellular domain with serine or threonine kinase activity.

G-protein-coupled receptors (**GPCRs**) are characterized by seven separate hydrophobic domains that traverse the membrane phospholipid bilayer (Fig. 3(a)). They possess an extracellular domain of variable size, enabling further subclassification of these receptors: glycoprotein hormones or small molecule ligands (for example, calcium, g-aminobutyric acid (**GABA**)) interact with large amino-terminal extracellular domains; biogenic amines (for example, catecholamines, serotonin (5-hydroxytryptamine)) bind to residues that lie within the transmembrane domain; other polypeptide hormones interact with residues in both the extracellular and transmembrane domains. The intracellular domains of the receptor enable interaction with G-proteins.

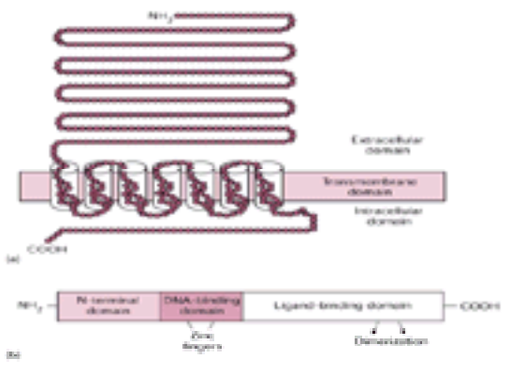

Fig. 3 Schematic representations of (a) G-protein-coupled receptor and (b) nuclear receptor illustrating their functional domains.

G-proteins typically form a heterotrimeric complex of a-, b-, and g-subunits that bind the guanine nucleotides GTP and GDP. The complex transduces signals from the receptor to downstream effectors such as adenylate cyclase, phospholipase C, or membrane voltage-dependent calcium channels. A family of different G-proteins (G_s, G_i, G_q, and others) exists with the ability to couple to different receptors and effectors, allowing a large array of potential receptor–G-protein–effector complexes, leading to diversity of cellular signalling.

A number of hormones signal via the cyclic AMP pathway (Table 4) and this mechanism is considered in further detail (Fig. 4). In the resting state, the G-protein complex is inactive and bound to GDP (Fig. 4(a)). Following hormone-binding to the receptor (Fig. 4(b)), the Ga-subunit binds GTP, becomes activated, and dissociates from the bg complex, to interact with adenylate cyclase (Fig. 4(c)). The latter converts ATP to the second messenger, cyclic AMP. This rise in the intracellular cyclic AMP level activates protein kinase A (**PKA**) which can phosphorylate a number of cellular targets: phosphorylation of a transcription factor, the cyclic AMP response-element binding protein (CREB), stimulates the transcription of genes containing CREs; other targets for PKA include enzymes in biochemical pathways or membrane ion channels.

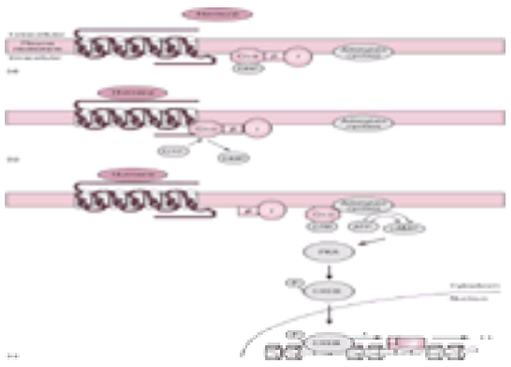

Fig. 4 G-protein-coupled receptor signalling via the cyclic AMP pathway.

At least two mechanisms serve to terminate signalling via a hormone–receptor complex. First, hydrolysis of GTP to GDP by the Ga-subunit promotes its reassociation with bg-subunits to reform an inactive complex; second, following hormone-binding, the G-protein-coupled receptors undergo phosphorylation of their intracellular domains by either PKA or other specific G-protein-coupled receptor kinases (**GRKs**). Such phosphorylation prevents further coupling to G-proteins and promotes receptor internalization. This receptor downregulation desensitizes the cell to hormone action, until further surface receptor is expressed.

Activation of their receptors by hormones such as somatostatin or dopamine is known to decrease intracellular cyclic AMP levels. Here, the hormone–receptor complex associates with a G-protein (Gi), whose a-subunit inhibits adenylate cyclase. Although many GPCRs signal via cyclic AMP, some receptors (for example, receptors for thyrotrophin-releasing hormone, gonadotrophin-releasing hormone, Table 4) are linked to different pathways. These receptors are coupled to G_q, whose a-subunit activates membrane phospholipase C (**PLC**) (Fig. 5). This enzyme catalyses the hydrolysis of phosphotidylinositol 4,5-bisphosphate (**PIP2**) to generate the second messengers inositol 1,4,5-triphosphate (**IP3**) and 1,2-diacylglycerol (**DAG**). IP_3 interacts with a specific receptor located on smooth endoplasmic reticulum, inducing the opening of intracellular channels and leading to a rise in cytoplasmic calcium levels (Fig. 5). Interaction of calcium with calmodulin (**CAM**), a cytoplasmic calcium-binding protein, activates a specific kinase (CAM kinase), which regulates a number of processes including hormone secretion, gene transcription, and metabolic enzymes. The rise in cellular calcium also facilitates DAG activation of protein kinase C (**PKC**), leading to phosphorylation of the Fos–Jun transcription-factor complex, inducing target-gene expression (Fig. 5). Hormones do not signal exclusively via a single pathway, with glycoprotein hormones and some peptides, for example, activating both cyclic AMP and phosphoinositide signalling (Table 4).

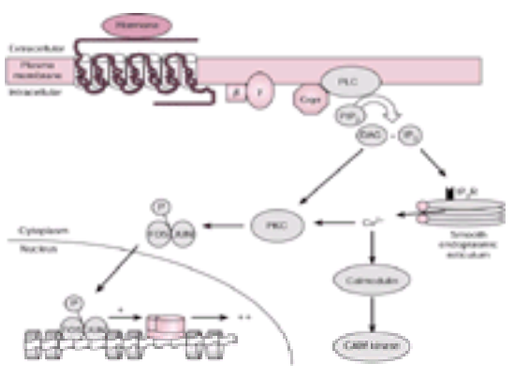

Fig. 5 G-protein-coupled receptor signalling via the phosphoinositide pathway.

The tyrosine kinase class of receptors is a diverse family that transduces signalling by insulin and IGF-1 as well as by epidermal, nerve, fibroblast, and platelet-derived growth factors. Growth-factor signalling differs from that of insulin, and the latter pathway will now be considered (Fig. 6). The interaction of insulin with its receptor promotes autophosphorylation of tyrosine residues in its cytoplasmic domains. In turn, this promotes the phosphorylation of substrates (for example, Shc and insulin-receptor substrate-1 (**IRS-1**)), followed by recruitment of adaptor proteins (Grb2/SOS). The Grb2/SOS complex recruits Ras, a GTP-binding protein. Ras activation induces signalling via a series of kinases (Raf, Mek, MAP kinase), culminating in the phosphorylation and activation of transcription factors that regulate target genes involved in mitogenesis or cellular differentiation. On the other hand, IRS-1 recruits phosphatidylinositol-3'-OH-kinase (PI3-kinase), which in turn activates protein kinase B. The latter mediates a number of the metabolic effects of insulin, enhancing translocation of a glucose transporter to the membrane to promote cellular glucose uptake, and activating enzymes involved in glycogen synthesis.

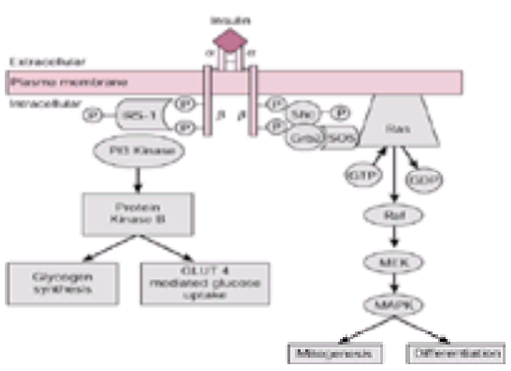

Fig. 6 Insulin action via its tyrosine kinase receptor and signalling cascade.

Hormones such as prolactin and GH interact uniquely with their receptors; a single polypeptide interacts simultaneously with two receptors promoting their dimerization (Fig. 7). The hormone–receptor complex recruits Janus kinases (**JAKs**) which phosphorylate **STATs** (signal transducers and activators of transcription). STATs translocate to the nucleus, interact with regulatory DNA elements, and promote target-gene transcription.

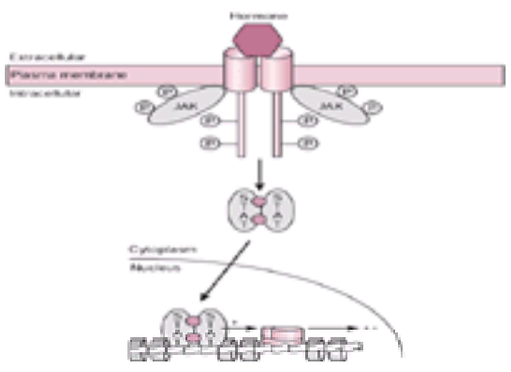

Fig. 7 Hormone signalling via the JAK–STAT pathway.

Activin and inhibin belong to the transforming growth-factor (**TGF**) class of peptides, which signal via a heterodimeric transmembrane-receptor complex with intrinsic protein serine/threonine kinase activity (Fig. 8). Here, hormone-binding promotes the association of two (type I, type II) surface receptors with differing properties. Subsequent transphosphorylation of the type I receptor by the intracellular kinase domain of the type II receptor leads to phosphorylation and dimerization of cytoplasmic Smad proteins. The Smad complex translocates to the nucleus to activate target-gene expression (Fig. 8).

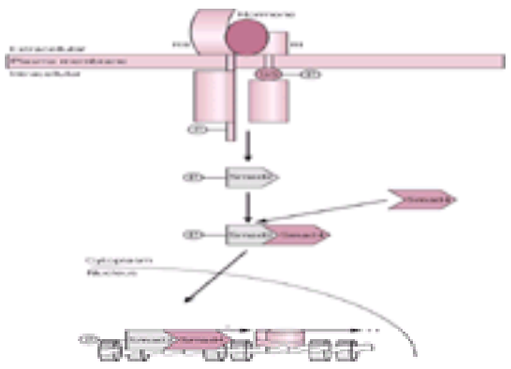

Fig. 8 Hormone signalling by the transforming growth-factor peptide family.

As described above, GPCR signalling is usually coupled to responses (for example, hormone secretion) by the Ga-subunit activation of cyclic AMP or phosphoinositide pathways. However, following receptor activation in some cellular contexts, the dissociated Gb/g-dimer subunit complex is also capable of stimulating effectors (for example, Ras, PI3-kinase), to enhance MAP kinase activity and elicit a mitogenic response.

Nuclear-receptor signalling

The nuclear receptors are a family of transcription factors that mediate the action of steroid and other lipophilic hormones. The human genome encodes approximately 60 to 70 different receptors, and it is clear that only a minority of these are targets for the action of major hormones (Table 5). The remainder comprise a large group classified as 'orphan receptors', reflecting the fact that either their ligands and/or physiological roles are still to be elucidated.

Based on homologies in their primary amino-acid sequence, nuclear receptors can be divided into distinct domains that mediate specific functions (Fig. 3(b)). A central DNA-binding domain contains cysteine-rich peptide motifs that chelate zinc to form two zinc fingers. The latter mediate receptor-binding to specific DNA sequences or hormone-response elements, usually located in target-gene promoters. The carboxy-terminal region of receptors encompasses their hormone-binding function as well as their ability to dimerize. Nuclear receptors can be divided into two major subclasses—the steroid receptors (homodimeric) and heterodimeric receptors—which differ in their mode of action.

Steroid receptors (for example, glucocorticoid, mineralocorticoid, [o]estrogen, progesterone, and androgen receptors (GR, MR, ER, PR, AR, respectively)) bind to hormone-response elements as homodimers (Fig. 9(b)). Some receptors (for example, GR, PR, AR) are bound to cytosolic heat-shock proteins. Hormone-binding to receptors promotes their dissociation from these, enabling translocation to the nucleus, dimerization, and interaction with DNA. In contrast, the thyroid, retinoic acid, and vitamin D receptors are constitutively nuclear and form heterodimers with a common partner (retinoid X receptor or **RXR**), to interact with DNA even in the absence of hormone or ligand (Fig. 9(a)). In some target-gene contexts, RXR can also form homodimers to mediate retinoid signalling.

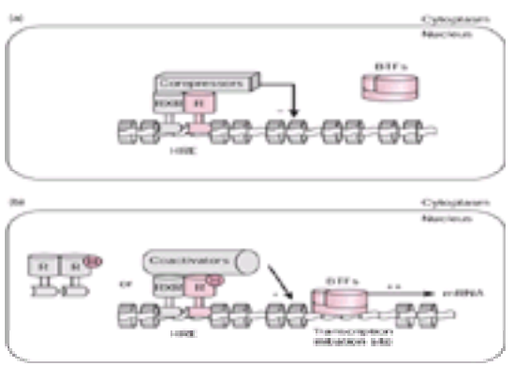

Fig. 9 Transcriptional regulation by nuclear receptors (a) In the absence of hormone, a subset of heterodimeric nuclear receptors (thyroid, retinoic acid) recruit co-repressors to inhibit gene transcription. (b) Hormone occupancy of homodimeric or heterodimeric receptors promotes their association with coactivators, leading to transcriptional activation.

In contrast to other transcription factors whose activity is controlled by post-translational modification (for example, phosphorylation), the hallmark of nuclear receptors is their ability to modulate gene expression in a hormone-dependent manner. Thus, in the absence of ligand, the thyroid and retinoic acid receptors actively silence target-gene transcription by recruiting a co-repressor complex of cofactors (Fig. 9(a)). For all nuclear receptors, hormone-binding induces a conformational change with dissociation of co-repressors and recruitment of coactivator proteins (Fig. 9(b)). This latter complex acts to relax the interaction between histone proteins and DNA in chromatin, thereby facilitating the access of basal transcription factors and RNA polymerase which induce gene transcription.

A further mechanism that controls signalling via nuclear receptors is regulation of the supply of their ligands to cells and tissues. Tri-iodothyronine, the ligand for the thyroid-hormone receptor (**TR**), is generated from circulating thyroxine by the action of type-1 or type-2 deiodinase enzymes expressed in the liver and central nervous system, respectively; the enzyme 5a-reductase converts testosterone to dihydrotestosterone in tissues of the male external genitalia. In contrast, the enzyme 11b-hydroxysteroid dehydrogenase type 2 catabolizes cortisol in renal cells, thereby enabling the mineralocorticoid receptor to respond selectively to aldosterone rather than to glucocorticoid, which it is also capable of binding to with high affinity.

Finally, in contrast to the classical effects of steroid hormones in modulating gene expression, recent evidence indicates that they can also modulate cellular functions such as hormone secretion or neuronal excitability within seconds or minutes. These rapid effects of steroid hormones occur independently of the genome, and are probably transduced by the same signalling pathways (for example, voltage-sensitive calcium channels) that mediate rapid responses to neurotransmitter hormones.

Genetic defects and endocrine disease

The majority of endocrine diseases can be divided into conditions of hormone excess, hormone deficiency, and hormone resistance. Defects in genes involved in hormone synthesis and action give rise to a spectrum of disorders (Table 6 and Table 7). Both germline gene defects causing inherited syndromes and somatic

mutations leading to acquired endocrine cellular dysfunction have been described.

Defects in developmental transcription factors are usually associated with endocrine gland hypoplasia: mutations in *HESX-1* cause optic and pituitary hypoplasia with agenesis of the corpus callosum; both *Pit-1* and *PROP-1* mutations disrupt the development of multiple pituitary cell types, resulting in a combination of hormone deficiencies; defects in *TTF-1*, *TTF-2*, and *PAX-8* result in thyroid dysgenesis manifesting as neonatal hypothyroidism; mutations in the *SRY* gene lead to failure of testis development and sex reversal in XY males.

Mutations in *DAX-1* or *SF-1*, orphan members of the nuclear-receptor family, disrupt both adrenal and gonadal development. Defects in other nuclear receptors (for example, VDR, TR, GR) are characterized by tissue resistance to their respective hormone ligands. Vitamin D resistance leads to rickets, together with abnormalities of skin differentiation, hair growth, and lymphocyte function, emphasizing its important extraskeletal actions. Point mutations in the androgen receptor are associated with a spectrum of phenotypes, ranging from complete feminization of XY individuals to mildly impaired virilization in men. In addition, expansion of a polyglutamine repeat sequence in the amino-terminal domain of AR is associated with adult-onset neuronal degeneration, leading to spinal and bulbar muscular atrophy. A homozygous defect in the oestrogen receptor in a male, led to failure of epiphyseal closure, resulting in tall stature together with severe osteoporosis. These manifestations suggest that testosterone effects on the male skeleton are, in part, mediated by its enzymatic conversion to oestrogens.

A growing number of disorders associated with defects in transmembrane receptors or their signalling intermediates have been described (Table 7). However, in addition to mutations that disrupt protein function, gain-of-function mutations causing constitutive activation of the receptor or signalling protein also occur. With G-protein-coupled receptors (GPCRs), diverse loss-of-function mutations, occurring most frequently in the extracellular domain, block hormone-binding or signalling, leading to insensitivity to hormone action. Such hormone resistance can lead to both hypofunction (for example, ACTH, TSH receptors) or hypoplasia (for example, LH, FSH receptors) of target glands expressing the receptor. Conversely, gain-of-function mutations in GPCRs typically occur in the third intracellular loop, causing constitutive activation of the receptor in the absence of hormonal ligand. Again, the functional consequence is either autonomous hyperfunction (for example, calcium, LH, FSH receptors) or excessive neoplastic proliferation (for example, TSH receptor, RET tyrosine-kinase receptor) of the target tissues in which the receptor is expressed (see Table 7). Constitutive activation of signal transduction may also result from G-protein mutations. Here, specific amino-acid substitutions in Gsa inhibit its intrinsic GTPase activity, and the GTP-bound protein constitutively activates adenylate cyclase leading to cyclic AMP accumulation. Somatic *Gsa* mutations occur in a proportion of pituitary growth-hormone secreting and thyroid adenomas; more widespread expression of a somatic *Gsa* mutation occurring early in development leads to polyostotic fibrous dysplasia, café-au-lait skin pigmentation, and hyperfunction of multiple endocrine glands, constituting the McCune–Albright syndrome. Similarly, germline, loss-of-function mutations which reduce cellular Gsa activity, are associated with resistance to multiple hormones, together with characteristic bone anomalies (Albright's hereditary osteodystrophy).

Further reading

Braverman LE, Utiger RD, eds. (2000). *Werner and Ingbar's the thyroid; a fundamental and clinical text*, 8th edn. Lippincott, Williams and Wilkins, Philadelphia.

DeGroot LJ Jameson JL, eds. (2001). *Endocrinology*, 4th edn. WB Saunders, Philadelphia.

Grossman A (1998). *Clinical endocrinology*, 2nd edn. Blackwell Science, Oxford.

Lodish H, *et al.*, eds. (1999). *Molecular cell biology*, 4th edn. WH Freeman, New York.

Yen SSC, Jaffe RB, Barbieri RL, eds. (1999). *Reproductive endocrinology*, 4th edn. WB Saunders, London.

12.2 Disorders of the anterior pituitary

Paul J. Jenkins and Michael Besser

Historical introduction

The anterior pituitary gland, often termed the 'conductor of the endocrine orchestra' secretes six hormones of known function. These control many of the peripheral endocrine organs, between them regulating growth and development, sexual behaviour and the menstrual cycle, lactation, as well as the thyroid and adrenal glands. Hippocrates in 400 BC gave the first description of a prolactinoma in stating that 'if a woman who is neither pregnant nor has given birth produces milk, and menstruation has stopped...'. There are several descriptions of giants in the Old Testament, both as individuals and also as families, (Joshua XIV: 'Arba was a great man amongst the Anakim'); it is possible that Goliath was slain by David because of his bitemporal hemianopia resulting from a suprasellar growth hormone-secreting pituitary adenoma. However, despite these early descriptions, detailed knowledge of the physiology and pathophysiology of the anterior pituitary gland has only become apparent over the last 60 or 70 years and is continuing to expand.

The first major advance was the ability to purify pituitary hormones, detect their actions by bioassay, and then synthesize them. The development of radioimmunoassay in the 1960s allowed for their rapid and more precise measurement and the elucidation of their physiological role. In Oxford in the 1950s Harris demonstrated the control of anterior pituitary function by hypothalamic factors; these were later extracted from several hundred tons of porcine pituitaries, principally by Schally and Guillemin in the 1970s, leading to their being awarded the Nobel prize. The availability of these hypothalamic factors facilitated studies of their physiological effects in both normal subjects and patients with endocrine disorders. The development of modern imaging techniques such as computed tomography (**CT**) and magnetic resonance imaging (**MRI**) has greatly enhanced our ability to visualize the anatomy of the hypothalamic–pituitary region and allow the precise localization of pituitary lesions. In the 1980s the advent of molecular biological techniques allowed greater understanding of the control mechanisms at a molecular level as well as fuelling interest in potential treatments. In experienced hands, trans-sphenoidal surgery has become a safe and effective firstline treatment for most pituitary tumours thus avoiding the morbidity associated with craniotomy.

Anatomy and development

The hypothalamus is the true conductor of the endocrine orchestra since it controls the pituitary gland. It is derived from forebrain tissue on either side of the lower parts of the third ventricle. The floor comprises the optic chiasm and tracts, the pituitary stalk, and mamilliary bodies; anteriorly it is limited by the lamina terminalis and the anterior commissure, whilst posteriorly it blends with the midbrain. The pituitary gland (sometimes called the hypophysis) lies within the pituitary fossa of the sphenoid bone above the sphenoid sinus. On either side are the cavernous sinuses containing the internal carotid artery, the third, fourth, fifth (first, and second divisions), and sixth cranial nerves. Superiorly, the gland is covered by a layer of dura through which the stalk passes. Occasionally congenital enlargement of this gap allows the usual pulsations of cerebrospinal fluid to be transmitted through into the fossa compressing the pituitary and ballooning the fossa and giving rise to the 'empty sella syndrome'. The normal pituitary is approximately the size of a large pea and weighs 100 mg; its dimensions are approximately 10 mm transversely, 9 mm anteroposteriorly, and 6 mm vertically. During pregnancy it undergoes enlargement up to almost twice its normal size as it becomes engorged with prolactin-secreting mammotroph cells under the influence of oestrogen. (Fig. 1).

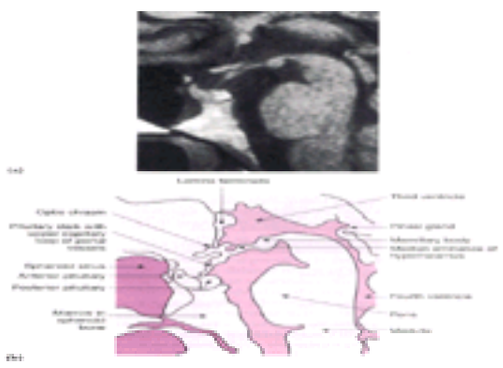

Fig. 1 Sagittal MRI scan of normal pituitary gland and an anatomical line drawing.

Embryologically and functionally the pituitary comprises two distinct parts. The epithelial portion which forms the anterior pituitary consists of the pars distalis and intermediate lobe; this originates from the stomodeal ectoderm of Rathke's pouch, which forms a vesicle that separates from the roof of the developing mouth. The neural portion which forms the posterior lobe, pituitary stalk, and infundibulum arises along with the rest of the hypothalamus from the diencephalic forebrain. Anterior pituitary cells also extend upwards to surround the neural tissue of the pituitary stalk forming the pars tuberalis.

The blood supply to the pituitary is in keeping with this dual origin. The anterior gland receives 80 to 90 per cent of its blood supply from the hypothalamo-hypophyseal portal veins which start as a plexus of capillaries in the median eminence/infundibulum of the hypothalamus, carrying blood and hypophyseotropic hormones down the stalk to bathe the cells of the anterior pituitary. The remaining blood supply is via the pituitary capsular vessels derived from the superior hypophyseal arteries. The neurohypophysis receives its blood supply from the inferior hypophyseal branches of the internal carotid artery. Venous drainage from the anterior pituitary is via the cavernous sinuses, principally into the petrosal sinuses and thence into the internal jugular veins.

Anterior pituitary differentiation and ontogeny is now known to be carefully regulated by tissue-specific transcription factors. The first marker of anterior pituitary differentiation is expression of the a subunit. Expression of prolactin, growth hormone, and thyroid-stimulating hormones in specific cells is controlled by the POU domain transcription factors Pit-I and Prop-I. Humans with mutations of Pit-I have a syndrome of short stature, congenital hypothyroidism, and prolactin deficiency.

General physiology and regulation

In common with all endocrine glands, secretion of anterior pituitary hormones is not autonomous. Each is subject to regulation by hypothalamic peptides and, with the possible exception of prolactin, subject to the fundamental endocrine regulatory mechanism of negative feedback by hormone from the target gland (Fig. 2; Table 1). This negative feedback control is at both the hypothalamic and pituitary levels and ensures precise homeostatic maintenance of physiologically appropriate hormonal secretion. Thus, failure of the primary gland results in reduced negative feedback and consequent increased hypothalamic and pituitary stimulation and secretion. Conversely, primary overactivity of the target gland results in increased negative feedback and diminished hypothalamic and/or pituitary stimulation. This central tenet

is fundamental to the laboratory interpretation of circulating hormonal levels and to the investigation of pituitary target gland disorders by means of either stimulatory or suppressive tests. In addition to this long-loop negative feedback, additional 'short-loop' feedback mechanisms exist between the pituitary and the hypothalamus.

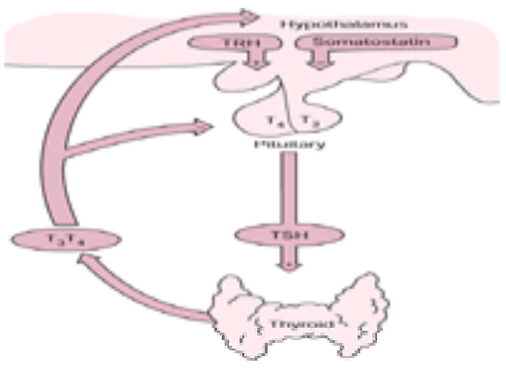

Fig. 2 Diagram of the hypothalamo-pituitary–thyroid axis showing negative feedback loops. TRH, thyrotrophin-releasing hormone; TSH, thyroid-stimulating hormone.

Pituitary hormones are synthesized as part of large precursor molecules; they are then cleaved into fragments which are secreted into the circulation. One fragment is the hormone concerned and the other cosecreted fragments have no known function as endocrine factors.

Causes of pituitary disease

Pituitary adenomas are the commonest cause of pituitary disease (Table 2); they can range from silent microadenomas to aggressive invasive tumours. Autopsy data suggest that the former may occur in up to 20 per cent of apparently normal people. Overall, clinically apparent pituitary tumours account for some 10 to 20 per cent of all intracranial tumours.

Pituitary tumours

With the advent of modern immunohistochemical techniques, the classification of pituitary adenomas has been simplified into that of the secreted hormone rather than the staining patterns of chromophobe, basophil, or acidophil adenomas resulting from periodic acid-Schiff staining (Table 2). Occasionally, despite these techniques, no tumour is visualized after trans-sphenoidal surgery even when the hormonal hypersecretion is cured postoperatively. Reasons for this may be that the tumour was so small it was missed on histological sectioning, the tumour was so small that it was not collected by the sucker during surgery, or the primary pathology was hypothalamic in origin.

Epidemiology of pituitary tumours

The annual incidence of clinically functioning pituitary tumours is estimated to be approximately 1 to 2 per 100 000 of the population. However, it is likely that this is an underestimate for two reasons: these are rare conditions which tend to be underdiagnosed and the incidence figures depend upon cancer registrations and not mortality data, which tend to be incomplete and are not universal. Despite these reservations, the following pituitary adenomas occur in decreasing order of frequency: non-functioning adenomas, prolactinomas, growth hormone-secreting, adrenocorticotrophic hormone (**ACTH**) secreting, thyroid-stimulating hormone secreting, and luteinizing hormone/follicle-stimulating hormone (**LH/FSH**) secreting tumours.

Clinical features of pituitary disease

The clinical features of pituitary dysfunction, usually resulting from a space occupying lesion, can be divided into local effects resulting from an expanding pituitary mass, anterior pituitary hormonal deficiency, and symptoms and signs of hormonal excess from hypersecretion of a pituitary hormone.

Local mass effects

An expanding mass within the pituitary fossa may give rise to headache, neuro-ophthalmological defects or facial pain according to the size and direction of expansion. Headaches usually result from dural stretching and are classically retro-orbital or bitemporal. They tend to be worse on waking and are relieved by analgesics; the somatostatin analogue octreotide may provide striking relief beyond any effect on hormone secretion, as it may have direct analgesic effects. Sudden catastrophic headaches may result from pituitary apoplexy. Very large pituitary masses may cause obstruction of the fourth ventricle or foramen of Munro resulting in hydrocephalus and expansion of the lateral ventricles. Rarely, inferior invasion and erosion of the sella floor may cause recurrent sinusitis or cerebrospinal fluid rhinorrhoea (confirmed by the presence of glucose in the fluid) and the risk of recurrent meningitis. Neuro-ophthalmological defects are common, particularly with macroadenomas, occurring in up to 60 per cent of such cases, although they are often asymptomatic. Although bitemporal hemianopia is the classic abnormality, any unilateral or bilateral visual field defect can occur depending on the site of impingement on the pathway of the optic nerve (Fig. 3). Lateral extension results in a squint from ocular nerve palsies. Extensive lateral invasion into the temporal lobe may result in temporal lobe epilepsy, whereas extensive superior extension may impinge upon the hypothalamus resulting in disorders of appetite, thirst, temperature regulation, and consciousness.

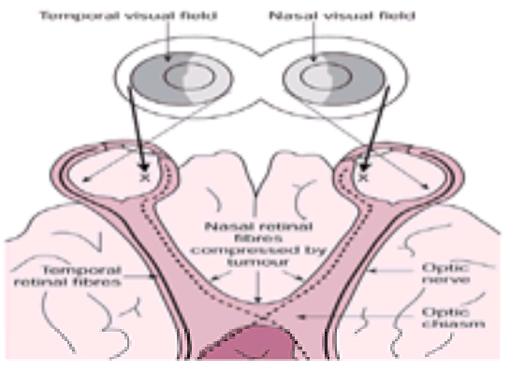

Fig. 3 Neuro-ophthalmological pathways and the classical bitemporal hemianopia that results from compression of the central optic chiasm by a pituitary tumour. However, any degree of unilateral or bilateral visual deficit can occur depending on the anatomical site of the lesion.

Hormonal deficiencies

Panhypopituitarism or varying degrees of loss of any of the six hormones may occur. Hypopituitarism resulting from a pituitary adenoma tends to occur in the sequential loss of LH, growth hormone, thyroid-stimulating hormone, and lastly ACTH and FSH. Thus, in adults the presenting clinical symptoms tend to be infertility, oligo/amenorrhoea, decreased libido, and erectile dysfunction. The clinical signs and phenotypic appearance reflect the loss of LH and growth hormone: there may be reduction of muscle bulk, decreased body hair, increased central adiposity, and small, soft testes. The facial appearance is almost pathognomonic—smooth skin with fine wrinkles, exaggerated by the loss of facial hair. Pallor may accompany ACTH deficiency. In children, hypopituitarism commonly presents with delayed puberty or impairment of growth. LH deficiency with preservation of growth hormone results in delayed fusion of the epiphyses of the long bones giving rise to a eunuchoid appearance, the span being greater than the height. Testicular size will depend on the stage of puberty prior to gonadotrophin failure and whether adrenarche,

influenced by ACTH-controlled adrenal androgen secretion, is maintained. Diabetes insipidus is almost never a presenting feature of primary pituitary tumours in either childhood or adults—it occurs most commonly in association with surgical treatment of pituitary adenoma.

Hormonal excess

If the pituitary adenoma is functioning, clinical symptoms and signs will also result from excess levels of either the secreted pituitary hormone itself, for example acromegaly (growth hormone), prolactin excess (prolactinoma), or from increased stimulation and secretion of the target gland hormone, for example Cushing's disease (ACTH), thyrotoxicosis (thyroid-stimulating hormone).

Evaluation of pituitary disease

The investigation of suspected anterior pituitary dysfunction requires the evaluation and integration in an appropriate clinical setting of: the presence of endocrine hyperfunction; the presence and degree of hypopituitarism; the radiological presence and extent of anatomical abnormalities; and assessment of neuro-ophthalmological function.

Hypothalamic pituitary function

Basal measurements

Measurements of basal levels of pituitary hormones with target gland hormone secretion are likely to be sufficient for the majority of cases of pituitary dysfunction, especially those involving thyroid-stimulating hormone, LH/FSH, and prolactin. It is only disorders of ACTH and growth hormone secretion that require dynamic investigation with stimulation or suppression tests according to the clinical scenario. Although the introduction of radioimmunoassays and then immunoradiometric assays revolutionized endocrine assessment, these have now largely been superseded by modern chemiluminescent assays which have the advantages of increased automation and sensitivity, avoidance of radioisotopes, and shorter assay time.

When interpreting any basal measurement of pituitary hormone, one needs to be aware of certain caveats:

1. Interpretation of all anterior pituitary hormonal levels, with the exception of prolactin, can only be made in the knowledge of the level of the primary target gland hormone.
2. The pulsatile nature of secretion of anterior pituitary hormones, especially ACTH, growth hormone, and LH/FSH, means that random isolated single levels may not be representative of overall secretion.
3. Specific factors such as time of day, stress, fed or fasting, asleep or awake, and stage of growth and pubertal development can all influence levels.

ACTH

Rapid proteolytic degradation in whole blood or frozen then thawed plasma, pulsatile secretion, and diurnal rhythm all need to be taken into account when measuring basal ACTH levels. In order to prevent degradation and thus uninterpretable results, plasma samples must be taken in an EDTA tube at 4 °C and immediately frozen. The time of day should be recorded (preferably 09.00) and samples should be taken from an in-dwelling cannula that has been *in situ* for at least 30 min in order to avoid the influence of stress. As ACTH is the prime regulator of cortisol secretion, a serum cortisol level (which does not require these logistical precautions) is a pragmatic measure of ACTH secretion; a 09.00 hour level of more than 200 nmol/l effectively excludes adrenal insufficiency under basal conditions, although it does not assess ACTH reserve (see later), whilst a 09.00 hour level of more than 550 nmol/l obviates the need for dynamic testing since the response to stress will be adequate.

Growth hormone

The markedly pulsatile secretion of growth hormone means that random levels of growth hormone are of very limited use. Furthermore, secretion is affected by nutritional status, being increased by amino acids. As it is a counter-regulatory hormone to insulin, levels are increased by hypoglycaemia and decreased by hyperglycaemia (the basis for its dynamic testing in situations of suspected oversecretion). There is also increased secretion during sleep (stages 3 and 4) and in response to stress. As a result of these factors, any meaningful assessment of growth hormone secretion requires dynamic testing. The action of growth hormone on tissues is classically mediated in an endocrine manner via hepatically derived circulating insulin-like growth factor I, although there is increasing evidence that most of its actions are effected through secretion of insulin-like growth factor I locally in the target tissues acting on cells adjacent to or the same as the cell of origin of insulin-like growth factor I (so-called paracrine or autocrine actions). Insulin-like growth factor II is not growth hormone dependent. A single serum level of insulin-like growth factor I has been claimed to represent an integrated index of growth hormone secretion, but it too is subject to separate influences and is discordant in up to 25 per cent of cases, particularly in adults. Thus, levels of insulin-like growth factor I in serum may be within the lower part of the normal range even though the patient has mild growth hormone deficiency. Growth hormone levels are markedly increased during puberty but decreased in pregnancy, due to the negative feedback on the pituitary by placental growth hormone, a variant of growth hormone produced by the placenta.

LH/FSH

More than for any other pituitary hormone, assays of LH/FSH must be interpreted with the simultaneous level of target gland hormone and the clinical scenario. In men a low testosterone in conjunction with low LH/FSH confirms the diagnosis of hypogonadotrophic hypogonadism rather than primary hypogonadism when LH/FSH levels would be high. In women the occurrence of the menstrual cycle is the definitive biological assay and excludes any deficiency. The occurrence of oligo/amenorrhoea requires measurement of gonadotrophins together with oestradiol, prolactin, and human chorionic gonadotrophin-b. Dynamic testing is very rarely required.

Prolactin

In the absence of pregnancy (during which levels rise markedly due to the action of increased oestrogen on the pituitary) a serum prolactin level of more than 4500 mU/l is almost always indicative of a prolactinoma. Mild elevations may be in response to stress and as such require repeated measurements. The large variety of causes of hyperprolactinaemia need to be considered (see later), particularly that due to coexisting drug therapy.

Thyroid-stimulating hormone

New ultrasensitive assays for thyroid-stimulating hormone have largely avoided the need for dynamic testing. A normal level of thyroid-stimulating hormone with a free thyroxine level in the normal range is indicative of euthyroidism. A subnormal level of thyroxine in conjunction with a normal or low thyroid-stimulating hormone (indicative of inadequate negative feedback) is strongly suggestive of secondary hypothyroidism. The commonly encountered low T_4 with an elevated thyroid-stimulating hormone is almost always indicative of primarily hypothyroidism, whilst conversely an elevated thyroxine or liothyronine level with a subnormal or undetectable thyroid-stimulating hormone confirms thyrotoxicosis. Secondary thyrotoxicosis as indicated by both an elevated thyroid-stimulating hormone and thyroxine level is very rare.

Dynamic endocrine testing

Where basal levels are inadequate or equivocal, dynamic endocrine tests will be required. The use of the combined anterior pituitary function test comprising LH-releasing hormone and thyrotrophin-releasing hormone is now no longer routinely used as it does not add to the clinical information obtained by basal hormone measurement. It is usually only disorders of the growth hormone and/or the ACTH axis that require either stimulatory or suppressive testing.

Insulin tolerance test

The insulin tolerance test remains the gold standard for assessing the AGTH/cortisol and growth hormone reserve. Its rationale is that insulin-induced hypoglycaemia is a marked stressor to hypothalamic neurones. A fall in blood glucose below 2.2 mmol/l invokes neuroglycopenic sympathetic stimulation, a stress reaction leading to

ACTH and cortisol release, which with catecholamines, glucagon, and growth hormone, act as counter-regulatory hormones liberating glucose from glycogen stores in the liver and returning blood glucose levels towards normal. Growth hormone is also released not only as a result of glucose levels falling below the neuroglycopenic threshold for stress induced release, but also in response to the subsequent stress.

Test procedure

The test must be performed in a properly equipped unit by well trained experienced personnel. In these circumstances it is an extremely safe and effective test. Contraindications are the presence of ischaemic heart disease, epilepsy or unexplained blackouts, a basal 09.00 cortisol level of less than 50 nmol/l (unless receiving exogenous glucocorticoids), and untreated hypothyroidism.

The patient should be fasted from midnight and the test performed at 09.00 hours. At least 30 min prior to this, an intravenous cannula is inserted. Blood is drawn for cortisol, growth hormone, and glucose analysis at 0, 30, 45, 60, 90 and 120 min. At 0 min 0.15 U/kg of soluble human insulin is injected intravenously (0.3 U/kg in insulin-resistant states such as acromegaly or Cushing's syndrome). Pulse rate and blood pressure along with the times of the characteristic features associated with hypoglycaemia are recorded with the blood samples. Symptoms and signs of hypoglycaemia usually occur between 30 and 45 min and comprise sweating, tachycardia, drowsiness, and hunger. Blood glucose must fall to less than 2.2 mmol/l and symptoms must be evident to be regarded as sufficient hypoglycaemic stress and to interpret the cortisol and growth hormone levels. A normal cortisol response is a rise to 580 nmol/l or more and growth hormone should increase to more than 20 mU/l. Failure to reach these levels indicates deficiency of ACTH and/or growth hormone. If the cortisol response is just subnormal, ACTH reserve may be adequate for day-to-day living but inadequate to cope with stresses such as illness or surgery and exogenous corticosteroid cover would be needed. A response to less than 450 nmol/l requires hydrocortisone replacement. Additionally all subnormal responses require a patient to carry a steroid card and MedicAlert bracelet.

An alternative to the insulin tolerance test to establish adequacy of ACTH/cortisol and growth hormone secretion in the presence of contraindications is the glucagon stimulation test (1 mg glucagon is given subcutaneously and sampling performed over 4 h) but the results are less consistent than with the insulin tolerance test.

The oral glucose tolerance test

As growth hormone secretion is inhibited by a rise in circulating glucose, the administration of exogenous oral glucose is used to confirm or exclude the diagnosis of acromegaly and remains the gold standard test for confirmation of this condition. The test is performed at 09.00 hours with the patient fasted from midnight. An intravenous cannula is inserted 30 min prior to the test and blood samples are drawn for analysis of glucose and serum growth hormone at –30, 0, 30, 60, 90 and 120 min. Glucose (75 g) dissolved in flavoured water (conveniently given as 370 ml 'Lucozade') is given to the patient immediately after the 0 min blood sample. In normal subjects, after oral glucose, serum growth hormone should be suppressed to undetectable levels (less than 0.5 mU/l). Failure to suppress is consistent with acromegaly, although it may also occur in diabetes mellitus, obesity, liver disease, and opiate dependence; in approximately 30 per cent of cases of acromegaly growth hormone levels paradoxically increase in response to glucose in this condition.

Pituitary imaging

The development of CT and subsequently MRI has revolutionized imaging of the pituitary gland, rendering the previous modalities of air encephalography and metrizamide cysternography obsolete. With their increasing availability, even the time-honoured skull radiograph has a limited role. These new techniques allow accurate visualization of the pituitary gland and, in the case of MRI, precisely delineates the extent of the surrounding invasion. If CT is used, imaging sections need to be thin (1.5 mm) and of high quality with multiplanar reconstructions; contrast should be given. Although CT is far more sensitive in demonstrating any calcification, as occurs commonly in craniopharyngiomas, MRI is nowadays the method of choice for delineating pituitary lesions. Its advantages are that it does not require ionizing radiation, it has the ability to image in any desired plane, and it shows the inherent contrast between tissues. Not only is it able to determine accurately the shape and dimensions of the anterior and posterior pituitary lobes (the latter has a high signal on T_1-weighted images in over 90 per cent of normal subjects) but it also delineates the hypothalamic region and optic chiasm. MRI allows accurate assessment of the size of pituitary adenomas, detecting lesions as small as 2 mm. It also determines the extent of any invasion superiorly, inferiorly, or laterally into the cavernous sinuses. On T_1-weighted images pituitary adenomas tend to be of lower signal intensity than the surrounding normal gland and enhance less briskly than the normal gland after injection with intravenous gadolinium contrast (Fig. 4).

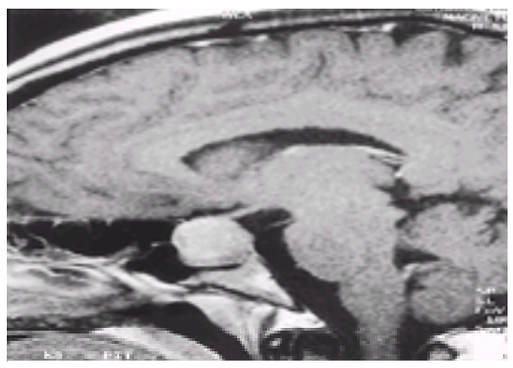

Fig. 4 MRI of a pituitary adenoma showing suprasellar extension and compression of the optic chiasm.

Neuro-opthalmological assessment

Neuro-ophthalmological assessment is mandatory in all cases of pituitary dysfunction. At the initial consultation visual acuity should be assessed with the use of Snellen charts and fundoscopy performed to exclude optic atrophy, retinal vein engorgement, or papilloedema from pressure on the visual pathways. Visual fields may be assessed by confrontation using a red pin. Patients with any clinical symptoms or evidence of compression of the optic chiasm from imaging studies require formal assessment of visual fields with Goldmann perimetery or visual evoked responses, stimulating each half-field in turn.

Although permanent loss of vision and/or visual field defects usually result from long-standing compression of the optic chiasm, the shorter the time of compression the easier and more complete is the reversal of any visual field deficit. Surgical decompression or shrinkage of prolactin-secreting tumours by medical therapy often results in rapid improvement in visual fields within hours or days, although the presence of optic atrophy reduces the likelihood of this occurring. Because onset is often insidious, patients may be unaware of any alteration in their vision, although once documented its presence requires them to inform the vehicle licensing authority as driving ability may be impaired. An exception to this usual gradual deterioration is pituitary haemorrhage when visual loss may be sudden with a loss of central vision and development of bitemporal field defects and possible ophthalmoplegia often accompanied by changes in mental function.

Pituitary surgery

With the exception of prolactinomas, for which medical therapy should be preferred as the primary therapy, trans-sphenoidal surgery is now regarded as the firstline treatment for pituitary adenomas. Originally performed by Harvey Cushing around 1910, the lack of adequate visualization prevented its reintroduction for routine use until the mid-1970s. In addition to its curative aim in pituitary adenomas, trans-sphenoidal surgery is also used where other treatments have failed, for example medical therapy in prolactinomas, and for debulking large tumours prior to irradiation. Pituitary apoplexy caused by significant haemorrhage into an adenoma is a neurosurgical emergency requiring prompt intervention. The most commonly used approach is with the patient in a semireclining position via a midline nasal route. Using a sublabial or direct nasal approach, the mucosa is cleaved off the nasal septum providing access to the sphenoid sinus with subsequent removal of the sellar floor. An endoscope is now in use. A less satisfactory alternative approach is via the ethmoidal sinus. Pituitary adenomas are usually soft and easily removed with curettes, although firmer and larger tumours may require piecemeal removal. The success of trans-sphenoidal surgery depends on a number of factors:

1. the size of the pituitary adenoma;
2. the degree of invasion into surrounding tissues, especially into the normal remaining pituitary gland, bone, meninges, and lateral extension into the cavernous sinuses;

3. the skill and experience of the surgeon; and
4. any previous therapy.

The aim is for selective adenectomy leaving sufficient functioning normal gland; in cases of microadenoma cure rates as high as 90 per cent can be achieved, although rates are usually much lower (in the region of 40 to 50 per cent) for macroadenomas. Visual field defects improve in approximately 80 per cent of patients.

Complications

In experienced hands, mortality is less than 1 per cent and is related to vascular complications, hypothalamic damage, or meningitis. Morbidity relates to local complications or endocrine dysfunction. Cerebrospinal fluid rhinorrhoea may occur and if persistent will require reoperation and sealing of the leak with autologous material such as fascia lata. The risk of meningitis can be minimized by the use of prophylactic antibiotics. Postoperative worsening of endocrine deficiency tends to increase with the size of the lesion and pre-existing hormonal deficiencies. Diabetes insipidus occurs in approximately 5 per cent of cases, though it is usually minor and transient; the exact prevalence depends on the stringency of the diagnostic criteria. The syndrome of inappropriate secretion of antidiuretic hormone is common, typically occurring transiently between the fifth and eighth postoperative days; it usually responds to fluid restriction.

While most macroadenomas, even those with a suprasellar extension of 2 or 3 cm, can often be substantially removed trans-sphenoidally, extremely large and more invasive pituitary tumours may require transfrontal removal. This approach is associated with additional risks of intracranial oedema/haemorrhage or damage to the optic nerve, frontal lobe, or hypothalamus.

Pituitary irradiation

With the advent of modern trans-sphenoidal techniques, pituitary irradiation tends to be reserved for patients who are either not fit enough to undergo surgery or in whom surgery is incompletely successful. In addition to so-called conventional megavoltage irradiation which has been in routine use since the mid-1960s, newer techniques of stereotactic or focused irradiation using one or a few large doses of irradiation have been introduced more recently.

Conventional irradiation uses a linear accelerator as the source; a cobalt source is less satisfactory. It is administered in a fractionated manner over 5 to 6 weeks with a total dose of 4500 cGy given in daily doses not exceeding 200 cGy. Irradiation should be via at least three portals (two temporal and one frontal) to prevent damage to the brain or optic chiasm. Modern CT techniques allow accurate planning and minimal variation in the daily dosage to surrounding structures. Widespread longstanding experience in this technique has revealed it to be both safe and effective. Hormone hypersecretion begins to decrease within 3 to 6 months with a rapid fall occurring within the first 2 years; thereafter there is a progressive exponential decline for up to 20 years. Thus most patients will be cured of their disease, although the time to achieve this varies. The use of this technique also dramatically reduces the risk of recurrence of both functioning and non-functioning adenomas. The major side-effect relates to the onset of hypopituitarism. In patients without preceding pituitary hypofunction, approximately 50 per cent will become deficient in growth hormone secretion after 5 years with slightly fewer becoming hypogonadal, followed much later (10 to 15 years) by deficiencies of ACTH and thyroid-stimulating hormone. These proportions increase with prior surgery. Thus any patient who has received pituitary irradiation requires lifelong careful follow-up. The other principal potential side-effect relates to damage to the optic chiasm, although this can be avoided by careful planning and keeping the daily fractionated dose to less than 200 cGy. Loss of cognitive function has been claimed to occur after irradiation, but does not occur if proper field planning and fractionated dosage is applied. The influence of prior surgery is also uncertain. Similarly there are reports of an increased incidence of secondary brain malignancies with a risk of perhaps 1 per cent at 20 years' post-irradiation. However, comparable control groups are difficult to obtain and it is likely that the predisposition to pituitary adenomas also predisposes to other cerebral tumours. The overall impression is that with careful radiotherapy the incidence of peripituitary second tumours is no greater in irradiated than non-irradiated patients with pituitary adenomas.

Stereotactic single high-dose pituitary irradiation using either the Gamma Knife (radiosurgery) or stereotactic multiple arc radiotherapy has received increasing attention in recent years, although long-term efficacy and safety data are not yet available. Care needs to be taken with tumours close to the optic chiasm. Initial impressions suggest that hypersecretory states fall to normal much earlier than after conventional radiotherapy but that hypopituitarism occurs just as often.

Individual pituitary hormones

ACTH

ACTH (adrenocorticotrophic hormone) is a 39-amino-acid peptide that is synthesized in the pituitary initially as part of a 231-residue peptide, pro-opiomelanocortin peptide. This then undergoes post-translational cleavage and modification before secretion. In the pituitary, pro-opiomelanocortin peptide is cleaved to b-lipotrophin, ACTH, a joining peptide, and an amino terminal fragment (Fig. 5). The biological activity of ACTH resides within its first 24 amino acids which are identical across species; synthetic $ACTH_{1–24}$ is used for clinical investigations. Pituitary secretion of melanocyte-stimulating hormone does not occur in humans; the increased pigmentation that is observed in conditions of ACTH excess is due rather to the melanocyte receptor stimulating properties of ACTH itself. ACTH G-protein coupled receptors are widely distributed in the adrenal cortex, predominantly in the zono fasiculata and reticularis. Upon binding of ACTH, cyclic AMP is generated which stimulates the synthesis and secretion of the glucocorticoid cortisol and the androgens androstenedione and dehydroepiandrostenedione. Mineralocorticoid secretion is largely controlled by the renin–angiotensin system. Long-term stimulation by ACTH acts to maintain adrenal size and growth, but there may be an additional, so far uncharacterized, hypothalamic–pituitary adrenocortical growth factors.

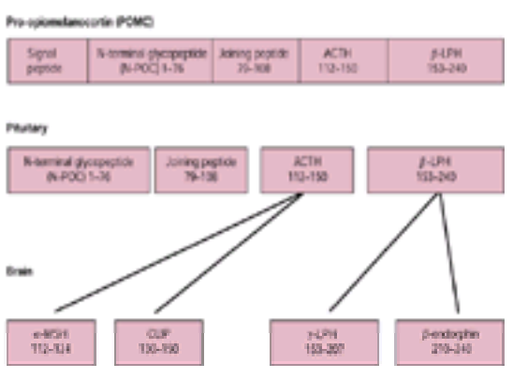

Fig. 5 Schematic representation of the pro-opiomelanocortin precursor protein and its subsequent tissue-specific cleavage products. LPH, lipotrophin; MSH, melanocyte-stimulating hormone; CLIP, coiticotrophinlike intermediate peptide.

Secretion of ACTH is tonically controlled by hypothalamic corticotrophin-releasing hormone, although arginine vasopressin also stimulates its release, particularly in response to stress. Cortisol feeds back negatively on the hypothalamus to inhibit the secretion of both of these factors, but also inhibits their actions at the pituitary directly; ACTH itself exerts negative feedback effects on the hypothalamus via a short loop. There are probably other physiologically important, as yet unidentified, ACTH-releasing hypothalamic hormones.

The secretion of ACTH is pulsatile but with a marked circadian rhythm. Secretory bursts start around 03.00 hours, are most frequent in the one to two hours before waking and the hour thereafter before declining progressively throughout the day to reach a nadir in the late evening, just before the normal time of sleeping. Secretion is predominantly controlled by the light–dark and sleep–wake cycles and is thus altered by time shifts—this accounts for at least some of the symptoms of jet lag. Secretion is also increased by major physical and psychological stress.

Conditions of ACTH deficiency

Isolated ACTH deficiency is rare; it more commonly occurs as a late component of the panhypopituitarism associated with pituitary adenomas, trans-sphenoidal surgery, or pituitary irradiation. Its symptoms are similar to those of Addison's disease but with two important differences: there is no pigmentation—indeed patients

are usually pale—and there is less risk of adrenal crises due to mineralocorticoid secretion being largely preserved. Patients typically complain of anorexia, malaise, loss of energy, and weight loss all of which contribute to them frequently being termed 'malingerers'. Severe cortisol deficiency may result in hypoglycaemia and predispose to circulatory collapse particularly during a supraimposed illness.

Conditions of ACTH excess

ACTH secretion is increased in both Addison's and Cushing's diseases (dealt with in other sections). It is also dramatically increased in the rare condition of Nelson's syndrome. This occurs in patients with Cushing's disease who have been treated by bilateral adrenalectomy. The loss of all negative feedback inhibition by cortisol can result in increasing growth of the pituitary adenoma with not only local invasion and mass effects, but also very high circulating ACTH levels which result in marked hyperpigmentation. Its incidence may be reduced by pituitary irradiation prior to adrenalectomy.

Treatment of ACTH deficiency

ACTH deficiency is best treated by glucocorticoid replacement. The standard regimen is 10 mg hydrocortisone on waking with a further 5 mg at lunchtime and 18.00 although this may need modifying according to the profile of measured serum cortisol throughout the day. Alternative regimes are prednisolone (5 and 2.5 mg) or dexamethasone (0.5 and 0.25 mg) although levels of either of these cannot be assayed in blood. It is essential that patients are educated to increase their dosage during illnesses or stress with at least a doubling of their oral medication. If vomiting occurs or during a perioperative period parenteral therapy will be required. Aldosterone production almost always remains adequate with ACTH deficiency unlike in primary adrenal insufficiency and fludrocortisone therapy is not needed.

Thyroid-stimulating hormone

The glycoprotein thyroid-stimulating hormone belongs in the same family as LH and FSH. Like other anterior pituitary hormones it is secreted in a pulsatile manner but with its effects mediated by circulating thyroxine and liothyronine from the thyroid gland. It is itself controlled by hypothalamic thyrotrophin-releasing hormone. Deficiency of thyroid-stimulating hormone almost never occurs in an isolated manner but occurs in conjunction with other pituitary hormone deficiencies.

Thyrotrophinomas

Thyrotrophin-secreting tumours (also known as TSHomas) are rare, comprising only 1 per cent of all pituitary tumours. Unlike primary thyroid gland diseases they have an equal sex incidence, with the majority presenting between the third and sixth decades of life, although cases at all ages have been recorded. Symptoms and signs result from the ensuing excess of thyroid hormone. As more than 90 per cent of the tumours are macroadenomas, visual field defects are common. Coexistant oversecretion of other pituitary hormones occurs in about 20 per cent of cases and is usually of growth hormone. Diagnosis is made on the basis of elevated circulating thyroxine with detectable or increased thyroid-stimulating hormone. The major differential diagnosis in such cases is the condition of resistance to thyroid hormone; distinguishing between the two requires both laboratory and radiological data. Generalized resistance to thyroid hormone is usually an autosomal dominant condition; however, when resistance to thyroid hormone is restricted to the pituitary it is often sporadic and patients are clinically hyperthyroid. In neither instance is a pituitary tumour present. Further differentiation is available with measurement of the a subunit of thyroid-stimulating hormone (and other glycoprotein pituitary hormones) and calculation of the a subunit/thyroid-stimulating hormone molar ratio; a ratio greater than 1.0, indicating oversecretion of the uncombined a subunit, is present in more than 90 per cent of thyrotrophin-secreting tumours, whilst a ratio of less than 1.0 is indicative of resistance to thyroid hormone.

Treatment

Surgery remains the initial treatment for the majority of cases of TSHomas, although the frequency of macroadenomas means that biochemical cure rates are only about 40 per cent. As almost all of these tumours express somatostatin receptors, the use of somatostatin analogues, for example octreotide, usually enables medical control of hypersecretion of thyroid-stimulating hormone to be achieved. This is often accompanied by a reduction in tumour size.

Gonadotrophins

The gonadotrophins LH (luteinizing hormone) and FSH (follicle stimulating hormone), together with thyroid-stimulating hormone and human chorionic gonadotrophin b comprise the family of glycoprotein anterior pituitary hormones. Whilst the a subunit is common to all these hormones, it is the b subunit that is unique to each and gives biological and immunological specificity. The actions of both LH and FSH are intimately involved with maintenance of libido and fertility; in women their secretion, subject to positive and negative feedback, controls ovarian gonadal secretion and the menstrual cycle. In men, LH controls testosterone production by the Leydig cells whilst FSH stimulates Sertoli cell spermatogenesis. Deficiency of gonadotrophin secretion is usually one of the earliest occurrences in pituitary tumours.

Gonadotrophinomas

Anterior pituitary tumours that secrete intact LH and/or FSH are termed gonadotrophinomas. Tumours secreting a subunit are more common, but as this is without biological effect and rarely measured in routine clinical settings these tumours are usually classified as non-functioning adenomas. Clinically both gonadotrophinomas and non-functioning adenomas usually tend to present in a similar manner with symptoms relating to a pituitary mass. In men, FSH-secreting tumours (known as FSHomas) cause testicular enlargement due to FSH-induced hypertrophy of the seminiferous tubules; as such the development of bilateral testicular enlargement should prompt investigation of a possible pituitary tumour.

The diagnosis of gonadotrophinomas is controversial; most clinicians require inappropriate elevation of serum LH and/or FSH. However, immunostaining of many clinically non-functioning adenomas will frequently demonstrate staining of either gonadotrophin or a subunit.

Treatment

Surgery is the main treatment for these tumours with adjunctive radiotherapy advocated by many clinicians to reduce the risk of recurrence. Medical therapy with long-acting gonadotrophin-releasing hormone agonists may reduce hormonal secretion, as may the use of somatostatin analogues, which may also result in tumour shrinkage.

Prolactin

Prolactin is a 199-amino-acid residue peptide that belongs to the same family as growth hormone with 16 per cent homology. Prolactin-secreting cells comprise approximately 10 per cent of the anterior pituitary cells in men but up to 30 per cent in multiparous women. Like other anterior pituitary hormones, prolactin is under the control of hypothalamic factors but is unique in that this comprises tonic inhibition by dopamine from the tuberoinfundibular neurones. Release also occurs after administration of thyrotrophin-releasing hormone and vasoactive intestinal peptide, which also occurs in the hypothalamus; although their physiological significance is clear in rodents it remains uncertain in humans. Prolactin synthesis is also strongly influenced by oestrogen which acts as a transcription factor for prolactin gene expression. Under this influence, concentrations are higher in premenopausal women than in men and increase during menarche and especially during pregnancy, when the gland may double in size and weight because of the increase in number of mammotroph cells. The causes of physiological and abnormal increased prolactin secretion are shown in Table 3. Following delivery, maternal levels decrease if breast feeding does not occur but remain elevated in response to suckling. Outside of lactation, nipple stimulation reflexly increases levels as does stress. Levels also increase in response to any disruption of the hypothalamic pituitary stalk or on administration of drugs that interfere with the synthesis or action of dopamine. Primary hypothyroidism may also cause levels to rise, falling again with thyroxine treatment (Table 3).

Action

Prolactin receptors, members of the cytokine superfamily, are widely distributed. Their predominant location is the mammary gland where their activation results in initiation and maintenance of lactation. During pregnancy, additional hormones are required for the development of the synthetic and secretory breast apparatus including oestrogen, insulin, cortisol, and placental mammotrophic hormones. However, the elevated oestrogen levels during pregnancy also inhibit lactation from the prepared prepuerperal breast, which occurs with their precipitous fall after delivery of the placenta. The actions of prolactin receptors at other sites is less clear. Direct effects on the hypothalamus inhibit pulsatile release of gonadotrophin-releasing hormone and thus gonadotrophin secretion, probably by increasing the inhibitory

opiate (endorphin) tone in the hypothalamus, resulting in impaired gonadal function. This is the basis for the contraceptive effect of lactation, which persists as long as prolactin levels are sufficiently high, maintained by the intensive reflex of breast feeding as the sole source of feeding the baby. It is also becoming increasingly clear that prolactin has widespread effects on immune functions as well as perhaps antiapoptotic actions in several tissues including the prostate.

Disorders of secretion

Absence of prolactin secretion rarely occurs on its own and is of unknown physiological significance other than precluding lactation. However, the usual coexistent loss of gonadotrophins and resultant infertility tends to precede this clinical effect.

Prolactinomas

Prolactinomas are the most common functioning anterior pituitary tumour encountered in clinical practice. They account for 25 to 30 per cent of pituitary adenomas and are much more common in women. The majority are sporadic, although occasionally they may be part of the multiple endocrine neoplasia type 1 syndrome. They may be either pure prolactin secreting or mixed somatomammotroph (growth hormone) tumours. They are of monoclonal origin. The clinical features can be divided into local mass effects or the effects of hyperprolactinaemia either on the breast or on the gonadotrophin-releasing hormone pulse generator. In premenopausal women the commonest symptoms are secondary oligo/amenorrhoea (in up to 95 per cent of patients), galactorrhoea (20 per cent), or infertility with regular cycles. Due to these symptoms, women tend to present earlier and thus the tumours are usually microadenomas. Although increased prolactin causes decreased libido and erectile dysfunction in men, these symptoms may be ignored and male patients, like postmenopausal women, present late with mass effects such as headaches, visual field defects, or involvement of other cranial nerves from cavernous sinus extension. Galactorrhoea is rare in men.

Treatment

In common with other pituitary tumours the aims of treatment are to reduce hormone secretion to normal, to reduce tumour size correcting visual field defects, and to restore normal anterior pituitary function.

Left untreated, the majority of microprolactinomas do not appear to undergo progressive enlargement. As such the need for treatment depends on the clinical situation and the patient's wishes. The introduction of the ergot-related synthetic dopamine agonist bromocriptine in 1971 revolutionized the treatment of prolactinomas. A variety of other dopamine agonists are now available, the only one with advantages being the long-acting cabergoline. All of these directly activate pituitary D_2 receptors, thereby mimicking endogenous hypothalamic dopaminergic action. They are highly effective in lowering prolactin levels, stopping galactorrhoea, and restoring normal gonadal function including fertility. They also have dramatic effects producing shrinkage of the tumour in 85 per cent of patients and improvement in visual fields. Disappearance of headaches is seen within days, and if imaging is repeated, shrinkage is confirmed within 4 to 6 weeks in most cases. Given these effects, it is mandatory that a serum prolactin level should be measured in all patients presenting with a visual field defect and a pituitary mass prior to surgery (normal levels are less than 400 mU/l or 20 ng/ml). A serum level of greater than 4000 mU/l (200 ng/ml) is indicative of a prolactinoma and introduction of dopamine agonist therapy will usually avoid the need for pituitary surgery. The usual starting dose for bromocriptine is 2.5 mg once daily taken during the main course of a bulk meal. This slows absorption and will minimize the side-effects of nausea and dizziness, although most patients experience tachyphylaxis and symptoms tend to settle with repeated dosages. An initial dose of cabergoline is 0.5 mg once weekly. Other side-effects include fatigue, headache, and nasal stuffiness. Most serious is psychosis which has been reported in approximately 1 to 2 per cent of patients, although more subtle psychiatric disturbances probably occur more often. A previous psychiatric history is therefore a contraindication to the use of this class of drugs. Most patients will require long-term therapy, although a significant proportion of microadenomas undergo spontaneous resolution/infarction and it is worthwhile having an intermittent trial off therapy from time to time. In the case of macroadenomas, pituitary irradiation is often advocated after initial shrinkage with dopamine agonists, in order to prevent recurrence. Occasional resistance to medical therapy may necessitate surgical intervention. While elevated levels lower than 4000 mU/l (200 ng/ml) may be found in patients with true prolactinomas, such levels may also be found in patients with pseudoprolactinomas. The term 'pseudoprolactinoma' is used for any peripituitary mass which interrupts delivery of dopamine to the normal pituitary via the pituitary stalk, i.e. a functional pituitary stalk section results and prolactin secretion from the normal gland is disinhibited. While prolactin levels as high as 6000 mU/l (300 ng/ml) have been described in association with prolactinoma, they are usually less than 2000 mU/l (100 ng/ml). While the clinical features in patients with pseudoprolactinomas may be identical, and are reversible with dopamine agonists, as in true prolactinomas, but pseudoprolactinomas do not shrink with medical therapy. Thus careful clinical supervision is needed during a trial of medical therapy for large tumours when prolactin levels are less than 6000 mU/l(300 ng/ml).

Hyperprolactinaema and pregnancy

As infertility is a common symptom of hyperprolactinaemia, it is likely that many women with prolactinomas will be trying to conceive whilst on dopamine agonists. Patients should be warned that they are likely to become fertile on treatment and the drug is usually stopped after confirmation of pregnancy. However, having been in use for nearly 30 years, it is clear that continued use of bromocriptine during pregnancy does not result in any adverse effects. Although there is less experience of cabergoline, there is no evidence to suggest that this should be any different.

During pregnancy, under the influence of increasing oestrogen levels, prolactinomas may enlarge like many peripituitary masses, especially meningiomas. Although the risk of this is small for microadenomas, there may be considerable expansion of macroadenomas. Such patients need very careful sequential monitoring for signs of chiasmal compression. In this event, reintroduction of medical therapy will shrink the mass and surgical intervention is rarely required. For patients with a macroadenoma, irradiation or trans-sphenoidal surgery has been advocated prior to attempting to conceive. Whilst this will reduce the risk of expansion during pregnancy, it may lead to growth hormone deficiency after some 5 years and gonadotrophin deficiency in about 50 per cent after 8 to 10 years.

Growth hormone

Human growth hormone is a 191-amino-acid single-chain protein containing two disulphide bonds. It is synthesized by cells located largely in the lateral part of the pituitary. Some 75 per cent circulates as a 22 kDa protein and 5 to 10 per cent as a smaller 20 kDa isoform with the remainder consisting of glycosylated and sulphated isoforms. Growth hormone is secreted in a marked pulsatile manner which is due to the integrated and co-ordinated effects of both stimulatory and inhibitory controlling hypothalamic hormones. Growth hormone-releasing hormone is a positive regulator of its synthesis whilst somatostatin (a 14-amino-acid peptide) is a potent inhibitor of both the frequency and amplitude of its secretory pulses. An additional positive regulator is the newly cloned growth hormone secretagogue 'ghrelin' which acts through a receptor which is distinct from that of growth hormone-releasing hormone. These regulatory peptides are themselves under numerous extrahypothalamic influences. Amino acids, sleep, stress, and a fall in blood glucose increase secretion of growth hormone; b antagonists are able to augment other stimulatory stimuli as do dopamine and cholinergic agonists, both of which are blocked by the cholinergic blockers. Oestrogen enhances the pulse amplitude of growth hormone. Thus secretion of growth hormone is subject to complex neuroregulatory control with levels increasing in response to several physiological stimuli (Table 4).

Overall growth hormone secretion is greater in females, with middle-aged women secreting approximately 45 µg in 24 h compared with 15 µg in men of equivalent age. Secretion increases from the time of puberty, peaking between 18 and 25 years, then inexorably falling to reach low levels after 60 years.

The growth hormone receptor

The growth hormone receptor is a 638-amino-acid receptor which is widely distributed throughout the body, being most abundant in the liver. It consists of extra- and intracellular domains; signal transduction requires dimerization of the receptor, which is facilitated by growth hormone binding. The identification of the specific high-affinity binding sites of the growth hormone molecule to the receptor has allowed their modification to provide specific modified growth hormone antagonists (see below). Signal transduction is via a number of pathways including activation of the JAK/STAT proteins, the insulin receptor substrate pathway and the MAP (mitogen-activated protein) kinase pathway. The complexity and variety of these pathways allows for the different anabolic/differentiative and proliferative effects of growth hormone in different tissues. Abnormalities of the growth hormone receptor occur in Laron's syndrome characterized by failure of growth and a distinct phenotypic appearance and high levels of growth hormone but low levels of insulin-like growth factor I.

Approximately 50 per cent of secreted growth hormone circulates bound to a growth hormone binding protein which consists of the cleaved extracellular domain of the transmembrane growth hormone receptor. Binding to this receptor fragment reduces the clearance rate of growth hormone thus prolonging its bioactivity. Growth hormone binding protein is absent, or abnormal, in Laron's syndrome and reduced or abnormal in other causes of insensitivity to growth hormone.

Actions of growth hormone

The actions of growth hormone can be divided into direct metabolic effects which act in the short term to increase glucose availability, regulate free fatty acids, and increase amino acid uptake, and powerful long-term effects on skeletal and soft tissue growth mediated via insulin-like growth factor I. Insulin-like growth factor I is a basic polypeptide of 70 amino acids (molecular weight 7.5 kDa) which was until recently thought to be synthesized almost solely by the liver. However, it is also synthesized by most tissues in response to growth hormone acting in a paracrine, autocrine, or juxtacrine manner. The circulating insulin-like growth factor I derived from the liver has little importance in determining the growth effects of growth hormone but is measured to reflect the overall functional status of growth hormone since tissue levels cannot be assayed. In the serum insulin-like growth factor I circulates bound to a group of binding proteins which also occur in tissue fluids; six have so far been identified. Insulin-like growth factor binding protein 3 is the predominant carrier of insulin-like growth factor I and its concentration is dependent on growth hormone itself; it thus plays a major role in regulating the bioactivity of insulin-like growth factor I. Insulin-like growth factor I acts via two specific receptors: type I, with a similar structure to that of the insulin receptor, has the greatest affinity for insulin-like growth factor I, whilst the structurally distinct type II receptor has the greatest affinity for insulin-like growth factor II. Insulin-like growth factor II is more important in the fetus and is not growth hormone dependent.

Disorders of growth hormone secretion

Acromegaly

The term acromegaly is derived from the Greek word summarizing its clinical manifestations: *akron*, extremity; *megas*, great. First described by Pierre Marie in 1886 it has an annual incidence of approximately 3 per million although this is almost certainly an underestimate. It is almost invariably due to a growth hormone-secreting pituitary tumour although very rare cases of ectopic secretion of growth hormone-releasing hormone from carcinoid tumours have been recorded. Approximately 5 per cent of cases are associated with multiple endocrine neoplasia type 1. Molecular analysis has revealed that approximately 50 per cent of cases are due to an activating mutation of the a subunit of the G-protein coupled receptor for growth hormone-releasing hormone.

Clinical features

Acromegaly affects both sexes equally, although its insidious onset means that the majority of patients are not diagnosed until the age of 40 to 60 years. Younger patients tend to have more aggressive disease and are detected earlier, although gigantism, occurring before fusion of the bony epiphyses, is a rare occurrence. In common with other pituitary tumours, the clinical manifestations can be related to the neighbouring effects of a local mass or to hypopituitarism, and to specific symptoms and signs relating to excessive secretion of growth hormone and insulin-like growth factor I (Table 5). The insidious onset means that the mean delay in diagnosis is about 7 years, although diagnosis may often take 10 to 20 years. In adults, the increase in soft tissue mass and skeletal effects are responsible for the clinical features including the classical coarse facial features, broad nose, and thick lips. Increase in the size of the mandible results in prognathism and interdental separation. Frontal bossing causes prominent supraorbital ridges. Sweating is one of the most prominent symptoms and the one most sensitive to growth hormone excess. Musculoskeletal symptoms are common, consisting of increased size and breadth of hands ('spade-like' hands) and feet; increasing ring size is common and is a sensitive index of excessive secretion of growth hormone. Degenerative arthropathy of the weight-bearing hips and knees is common, often necessitating joint replacement, and of the spine with pain and neurological pressure effects. Symptomatic carpal tunnel syndrome occurs in approximately half of patients and up to 80 per cent will have subclinical abnormalities. Thyroid enlargement is common resulting in a nodular goitre. Despite these signs the diagnosis is often overlooked.

Complications of acromegaly

Although originally regarded as a predominantly cosmetic disease, several epidemiological reviews have established that acromegaly is associated with significant morbidity and mortality. Early surveys revealed that over 50 per cent of patients died before the age of 60 years, principally due to cardiovascular and metabolic complications. With improved management of both these and the underlying disease, patients are now surviving longer. Reduction of mean serum levels of growth hormone to less than 5 mU/l (2 ng/ml) and/or a normal serum level of insulin-like growth factor I appear to be associated with a normal life expectancy.

Cardiovascular complications are a major source of morbidity and mortality. Although growth hormone in low doses may have a beneficial cardiac action, excessive amounts result in cardiomyopathy and increased left ventricular mass. Many patients develop arrythmias. Hypertension is a frequent complication requiring aggressive monitoring and management; it may persist despite lowering the levels of growth hormone. The increased cardiovascular risk factors also predispose to cerebral vascular ischaemic events. Due to the counter-regulatory properties of growth hormone on insulin, insulin resistance and impaired glucose tolerance is common often resulting in frank diabetes. This may be accompanied by a hypertriglyceridaemia. Increase in the soft tissues of the larynx and tongue can result in obstructive sleep apnoea which often requires domiciliary nocturnal continuous positive pressure airway ventilation. Anaesthetic intubation for any surgical intervention is more difficult (Table 6).

The question of increased risk of malignancy in acromegaly has been controversial for many years, but recent studies have confirmed that there is an increased risk of premalignant colonic tubulovillous adenomas and colorectal carcinoma. Although comparison with comparable control groups is difficult, this increased risk of colorectal cancer appears to be at least threefold. It appears to be an age-related complication with adenomas occurring after the age of 40 years and carcinomas occurring in or after the sixth decade. Although the precise mechanisms remain uncertain it is associated with elevated levels of insulin-like growth factor I which might both increase proliferation of epithelial cells and inhibit their apoptotic response to local environmental factors such as bile salts. These patients require regular colonoscopic screening, although their associated colonomegaly makes this technically challenging. There is a suggestion that acromegaly also predisposes to breast cancer and some circumstantial evidence indicates that elderly male patients should also be screened for prostatic neoplasia.

Aims of treatment

Although the aims of treatment are the same as those for any pituitary tumour, there has been controversy as to the level of reduction of growth hormone that should be considered as a satisfactory therapeutic aim. Epidemological surveys suggest that a normal life expectancy is associated with a mean serum growth hormone level of less than 5 mU/l (2 ng/ml) and/or a normal level of insulin-like growth factor I, although reference tends to be made to a 'safe' level of growth hormone rather than a cure, since physiological responses to growth hormone rarely return to normal. Larger epidemiological studies are required to confirm and refine these findings.

Treatment of acromegaly

Surgery: Trans-sphenoidal surgery remains the firstline therapy for the majority of patients. It is a safe operation with low morbidity. Its success rate depends on the size of the pituitary tumour, the presurgical levels of growth hormone, and the experience of the surgeon. Safe levels of growth hormone are achieved in 60 to 80 per cent of patients with microadenomas and 30 to 40 per cent of patients with macroadenomas.

Radiotherapy: Pituitary irradiation is a very effective means of reducing secretion of growth hormone. It tends to be reserved for patients in whom levels of growth hormone remain elevated after surgery. Its efficacy depends on the preirradiation level with approximately a 50 per cent fall occurring in the first 2 years but with an exponential decline thereafter which may continue for 15 to 20 years. The majority of patients therefore eventually achieve safe levels of growth hormone with a mean interval of 5 to 7 years being required. Normalization of insulin-like growth factor I occurs in over 50 per cent of patients by 10 years. Focused radiotherapy probably results in a more rapid fall. Medical therapy is required during the interim period.

Medical therapy: For many years dopamine agonists provided the only medical therapy for acromegaly. However, unlike in prolactinomas, the response rate is relatively poor with less than 10 per cent of patients achieving safe levels of growth hormone. Furthermore, the doses required are usually very much higher than in prolactinomas with a consequent increased frequency of side-effects. The introduction of the synthetic somatostatin analogue octreotide in the early 1980s provided an enormous advance in medical therapy. Natural somatostatin has a half-life of approximately 80 s and thus cannot be used therapeutically unless by continuous intravenous infusion. Octreotide by contrast has a prolonged effect lasting 6 to 8 h and thus thrice daily subcutaneous injections provide reasonably stable concentrations of drug in the plasma. Using this regimen, growth hormone is suppressed in more than 90 per cent of patients, but only approximately 60 per cent achieve safe levels and 50 per cent a normal level of insulin-like growth factor I. The usual total daily dose ranges from 150 to 600 µg, although occasional patients may require higher doses. There is evidence that octreotide may also shrink the adenoma in many patients, although this effect is not nearly as dramatic or complete as that seen in prolactinomas following therapy with a dopamine agonist. There is no consistent evidence that such shrinkage may aid subsequent surgery. Although generally well tolerated, the predominant side-effects are local pain at the site of injection and gastrointestinal—abdominal cramps, flatulence, and bulky or fluid stools are common although they tend to resolve with time. In the longer term, there is an increased frequency of gallstones which is due to both an inhibition of gall bladder

contractility and increased bacterial deconjugation of bile acids as a result of prolonged intestinal transit time. Atrophic gastritis is another long-term complication. More recently, subcutaneous octreotide has been superseded by the development of somatostatin analogue depot formulations, of which there are currently two, octreotide LAR and lanreotide SR, both of which are administered intramuscularly. The former is available at variable doses of 10, 20, or 30 mg and is administered every 4 or 6 weeks, whilst lanreotide SR is available as a 30 mg dose administered every 14 days or less. Their efficacy is similar to subcutaneous octreotide but with the obvious advantage of increased convenience for the patient. Pegvisomant is an entirely new medical therapy for acromegaly. It is a recombinant modified growth hormone molecule that has increased affinity to one receptor binding site but with a different modification of the other binding site that prevents receptor dimerization. Its conjugation to polyethylene glycol increases its molecular size and prolongs its half-life; it is given as a subcutaneous daily injection. Phase 3 clinical trials have shown it to be extremely effective, with over 90 per cent of patients achieving a normal level of insulin-like growth factor I. The fact that it is a modified growth hormone molecule means that it interferes with most growth hormone assays and thus this measurement cannot be used as a marker of efficacy. Indeed, growth hormone levels actually rise with this medication but to date there is no evidence of pituitary tumour growth. Whether pegvisomant will become a firstline treatment in acromegaly remains to be determined.

Growth hormone deficiency

Although long realized to be vitally important in children, the role and widespread influence of growth hormone in adult physiology has only become apparent over the last decade. It is now clear that not only does growth hormone have widespread functions, but its deficiency in the adult is associated with significant morbidity and mortality. Based on indirect evidence of the frequency of pituitary disease, the annual incidence of adult onset growth hormone deficiency is estimated to approximately 10 per million. In adults, pituitary adenomas remain the commonest cause, although iatrogenic causes, especially irradiation, are also common. In children, approximately 70 per cent of cases are idiopathic. As many of these children have a normal growth hormone axis on subsequent retesting in their teens and adulthood, the pathogenesis might be due to deficiency of hypothalamic growth hormone-releasing hormone or failure or delayed maturation of the hypothalamic–pituitary axis.

Clinical features

Growth hormone does not appear to be necessary for intrauterine growth and most children with congenital or childhood onset growth hormone deficiency tend to present with a falling off of growth velocity and failure to grow, which if unchecked leads to short stature. Bone age is delayed but with a normal weight for height. Thus, early detection requires accurate growth and weight charts.

Over the last 10 years, there have been a large number of studies detailing the effects of growth hormone deficiency, and in the case of adults it is now clear that these patients have a distinct phenotype (Table 7).

The impairment of psychological well-being and quality of life is probably the most important symptom of growth hormone deficiency. Its occurrence has been validated by a variety of questionnaires including the Nottingham Health Profile, the Psychological and General Well-Being Schedule, and the General Health Questionnaire, as well as a specific assessment of growth hormone deficiency. Consistent findings using these measures are decreased energy and mood, increased emotional lability and social isolation, increased anxiety, and reduced overall energy levels and vitality. These measures show significant improvement with administration of growth hormone in about 85 per cent of cases, which in some patients is dramatic. Lack of effect may reflect psychological dysfunction related to previous surgery or irradiation.

Growth hormone deficiency is associated with a reduction of lean body mass of approximately 7 to 8 per cent and an increase in fat mass of the same amount. This tends to be in a central distribution resulting in increased waist–hip ratio. There is decreased fluid volume comprising total body water. Administration of growth hormone will reverse all of these changes with an increase in lean body mass of up to 5 kg, predominantly skeletal muscle, and a corresponding reduction of fat mass with a consequent decreased waist–hip ratio. There is an increase in total body water. One of the most important effects of growth hormone appears to be a significant reduction in bone mineral density, with several studies showing an increase in osteoporotic fractures. Although recombinant growth hormone increases bone mineral density, it takes a minimum of 12 months for the effect to be measurable (using dual-energy X-ray absorptiometry scanning) and as yet there are no studies showing a consequent reduction in the rate of osteoporotic fractures. Impaired muscle strength and exercise performance is a frequent symptom and sign of growth hormone deficiency. Maximum oxygen uptake is reduced to approximately 80 per cent of that of age-, sex-, and height-matched controls. Whilst it improves with administration of growth hormone, not all measures of strength and muscle function are restored to normal and several years of treatment may be required.

Growth hormone deficiency has been associated with a two- to threefold increased mortality relating to vascular disease. Possible mechanisms might be related to the demonstrated increased thickness of the arterial intima and atherosclerotic plaques of carotid arteries, as well as the increased risk factors for ischaemic heart disease such as body mass, total cholesterol and low-density lipoprotein, and reduction in high-density lipoprotein. Metabolic effects of insulin resistance and hyperinsulinaemia with decreased carbohydrate metabolism might also be important. The abnormal lipid profile improves with growth hormone treatment, although an associated increase in lipoprotein (a), which is proposed to be an independent risk factor for atherosclerosis, remains of some concern.

Growth hormone deficiency is associated with impaired cardiac function comprising decreased left ventricular mass, impaired systolic function, and reduced ejection fraction. Whilst there is anecdotal evidence of dramatic increases in cardiac performance, the majority of studies have shown considerably milder effects on increased ventricular mass, stroke volume, and cardiac output.

Treatment of adult growth hormone deficiency

The aims of treatment are to improve and normalize the abnormalities associated with growth hormone deficiency. At present, most endocrine units only offer growth hormone replacement to patients who have a severe growth hormone deficiency (i.e. a peak growth hormone response of less than 9 mU/l after insulin-induced hypoglycaemia) and who are symptomatic or have abnormal metabolic/bone investigations. It is not yet routinely administered to all patients with growth hormone deficiency. In part this is due to its cost (in 2000 approximately £5000 per annum). There is, however, growing interest in the role of recombinant growth hormone therapy in the frail elderly with a view to improving their functional capacity, as it is suggested that the physiological decrease in growth hormone secretion in the elderly might be responsible for the associated changes in body composition. At present, however, trials of growth hormone administration in these patients are limited and there is obvious concern about the long-term safety issues (see below). Amongst patients in whom growth hormone treatment is indicated, the development and widespread availability of recombinant growth hormone has avoided the risk of transmission of Creutzfeldt–Jacob disease which was associated with the use of pituitary-derived natural growth hormone before 1985. The development of modern cartridge pens has also facilitated its administration, which is usually as a once daily subcutaneous injection in a manner similar to insulin. Currently, childhood doses are calculated according to body weight with the end point being growth velocity. Early studies in adults also used a weight-based regimen but the doses tended to be supraphysiological with an increased incidence of side-effects. It is now clear that individual dose titration should be performed with an initial daily dose of 0.8 U (0.3 mg). which is subsequently titrated according to the serum values of insulin-like growth factor I. The aim is to restore serum insulin-like growth factor I to the upper half of the age-matched normal range. This gradual titration ensures a more physiological replacement and minimizes the side-effects. The average maintenance dose in men is 0.8 U (0.3 mg) whilst in women it is higher at 1.2 U (0.4 mg). Regardless of the final dose, it has become clear that a minimum of 6 months' treatment is required to establish whether or not benefit has occurred.

Side-effects

Too large an initial dose, or too rapid an increase, results in sodium and water retention and thus oedema, weight gain, and carpal tunnel syndrome. There may be changes in glucose metabolism and alterations in cortisol metabolism such that hypopituitary patients taking hydrocortisone may need adjustment of their dose. Hypopituitarism seems to be associated with a reduced incidence of malignancy and there is a theoretical possibility that this might be increased back to normal, particularly as several epidemiological studies have demonstrated serum insulin-like growth factor I to be a risk factor for prostate, breast, and colorectal cancer. These safety issues will be clarified by continual analysis of the on-going long-term surveillance programmes that have been established, although there is no current evidence supporting this theoretical concern.

Craniopharyngiomas

Craniopharyngiomas are rare tumours with an incidence of approximately 1 to 2 per million population, but constitute the commonest intracranial tumour of childhood. They are derived from remnants of Rathke's pouch and as such arise within or above the pituitary fossa. They usually contain both cystic and solid components with the cyst containing oily fluid that is rich in human chorionic gonadotrophin b. Although histologically these tumours are benign, they are often locally invasive with a

strong tendency to recur. Their frequent involvement of the hypothalamus, pituitary, and optic pathways means that the clinical symptoms may be severe. Over half the cases occur in childhood when growth failure is the predominant complaint; presenting symptoms in adults include amenorrhoea and decreased libido. Hypothalamic symptoms, which are often most disabling, include obesity, hyperphagia, diabetes insipidus, and disturbances of sleep. In addition, as the tumours tend to be large, symptoms of a space occupying lesion and visual symptoms are common.

A unique feature is the frequent presence of calcification within the tumour, which aids in its diagnosis as it is visible on both skull radiographs and CT scans. MRI is optimal for visualizing the cystic components and for determining the extent of local invasion but does not show the calcification. There exists considerable controversy about the optimal treatment of these tumours. Early series tended to favour radical resection in order to reduce the chances of recurrence. However, such an approach, particularly in childhood, results in considerable damage to the optic apparatus and hypothalamus with an increased mortality and morbidity. Many units now favour the subtotal removal with postoperative irradiation. Recurrence of cysts is often dealt with by drainage and the insertion of yttrium-90 seeds into the cyst cavity.

Lymphocytic hypophysitis

Although not described until 1962, lymphocytic hypophysitis is increasingly recognized as a cause of pituitary dysfunction. It is an autoimmune disease with more than 90 per cent of cases occurring in women, often in relation to pregnancy when it occurs in the late third trimester or puerperium. However, men and patients of any age can be affected. Histologically, the disease is characterized by infiltration of the pituitary by lymphocytes and plasma cells with the formation of follicles containing germinal centres. There is subsequent destruction of the gland and fibrosis. Lymphocytic hypophysitis tends to present in either of two ways. In pregnancy, there is often a rapidly expanding pituitary mass presenting with headaches and visual failure. Alternatively, outside of pregnancy there may be an insidious onset of hypopituitarism. Diabetes insipidus occurs in approximately 10 per cent of cases. The major differential diagnosis is that of Sheehan's syndrome, although it is highly likely that many cases previously ascribed to Sheehan's syndrome were due to hypophysitis. Unlike pituitary adenomas, the destruction of pituitary cells in hypophysitis tends to affect corticotrophs and thyrotrophs predominantly, with deficiency of ACTH and/or thyroid-stimulating hormone occurring in approximately 80 per cent of patients. There are no specific radiological features, with both CT and MRI showing a varying appearance: the pituitary mass may appear hypo- or isointense with variable enhancement. More than 50 per cent show suprasellar extension.

The treatment for this condition is controversial. As there have been numerous reports of spontaneous resolution, particularly during pregnancy, the development of a pituitary mass in the third trimester or puerperium may warrant a conservative approach, although careful monitoring of visual fields is essential. Some endocrinologists familiar with this condition advocate a trial of steroids, although evidence of their efficacy remains anecdotal. Outside of pregnancy, the differential diagnosis of a pituitary mass is broad and trans-sphenoidal surgery is indicated both to remove the mass and to obtain a histological diagnosis. Although it is an autoimmune disease, there are no specific antibodies available for its diagnosis; however, the presence of antithyroid or other organ-specific antibodies is frequent.

Further reading

Carroll PV *et al.* (1998). Growth hormone deficiency in adulthood and the effects of growth hormone replacement: a review. Growth Hormone Research Society Scientific Committee. *Journal of Clinical Endocrinology and Metabolism* **83**, 382–95.

Ezzat S *et al.* (1994). Acromegaly. Clinical and biochemical features in 500 patients. *Medicine (Baltimore)* **73**, 233–40.

Ho KKY (2000). Growth hormone replacement therapy in adults. *Current Opinion in Endocrinology and Diabetes* **7**, 89–95.

Jenkins PJ *et al*. (1995). Lymphocytic hypophysitis: unusual features of a rare disorder. *Clinical Endocrinology* **42**, 529–34.

Jenkins PJ *et al.* (1997). Acromegaly, colonic polyps and carcinoma. *Clinical Endocrinology* **47**, 17–22.

Jenkins PJ *et al.* (2000). IGF-1 and the development of colorectal neoplasia in acromegaly. *Journal of Clinical Endocrinology and Metabolism* **85**, 3218–21.

Orme SM *et al.* (1998). Mortality and cancer incidence in acromegaly: a retrospective cohort study. *Journal of Clinical Endocrinology and Metabolism* **83**, 2730–4.

Trainer PJ, Besser M, eds (1995). *The Bart's endocrine protocols*. Churchill Livingstone, Edinburgh.

Trainer PJ *et al.* (2000). Treatment of acromegaly with the growth hormone-receptor antagonist pegvisomant. *New England Journal of Medicine* **342**, 171–7.

12.3 Disorders of the posterior pituitary

John Newell-Price and Michael Besser

Anatomy

The posterior pituitary lies immediately dorsal and caudal to the anterior pituitary gland and is an extension of the ventral hypothalamus. The major function of the posterior pituitary is the release of the nonapeptide peptide hormones arginine vasopressin (AVP) and oxytocin. These are synthesized in the magnocellular neurones of the supraoptic and paraventricular nuclei of the hypothalamus, and pass by axonal transport, to be stored in nerve endings in the posterior pituitary (Fig. 1). Efferent fibres from these nuclei also project to the median eminence, the brainstem, the floor of the third ventricle, and the spinal cord. Afferent fibres arise from osmoreceptors adjacent to, but separate from, the supraoptic and paraventricular nuclei, probably in the organum vasculosum of the lamina terminalis or subfornicial organ rostrally in the hypothalamus; signals are also received from the brainstem, and from the vagus and glossopharyngeal nerves that receive input from the pharynx and baroreceptors of the heart and great vessels. All these fibres carry the sensory signals modulating AVP and oxytocin release.

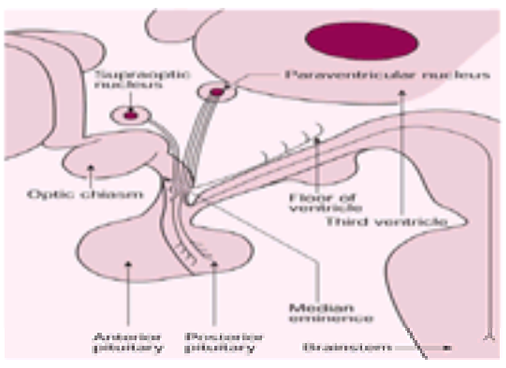

Fig. 1 Schematic representation of the neuronal pathways from the paraventricular and supraoptic nuclei. The nerves project to the posterior pituitary, the median eminence, the floor of the third ventricle, and the brainstem. Afferent fibres from the osmoreceptors and thirst centre are shown. (Reproduced from Besser GM and Thorner MO, 1994, with permission.)

Vasopressin and oxytocin

Vasopressin and oxytocin have molecular weights of 1087 Da and 1007 Da, respectively. The genes for both these hormones are located on chromosome 20q13 in close physical proximity. Each peptide is synthesized as a 145-amino acid precursor comprising a signal peptide, the peptide hormone AVP or oxytocin, and their specific neurophysin—I for oxytocin and II for AVP. The AVP precursor has, in addition, a glycoprotein at the C-terminus. AVP and oxytocin are composed of a six-member disulphide ring with a three-amino acid tail (Fig. 2). The differing amino acids at positions 3 and 8 have a profound effect on the biological action. In the synthesizing nuclei, granules are formed containing the precursor complex of neurophysin and oxytocin or vasopressin. During axonal transport processing cleaves off the active hormone from the neurophysin, and the products are stored in the nerve termini in the posterior pituitary. On firing of the nerves, AVP or oxytocin together with the relevant neurophysin are released into the systemic circulation. Most AVP circulates as free hormone and has a half-life of approximately 10 min. AVP is principally cleared by the liver and kidneys, whilst oxytocin is cleared in these sites and in the uterus.

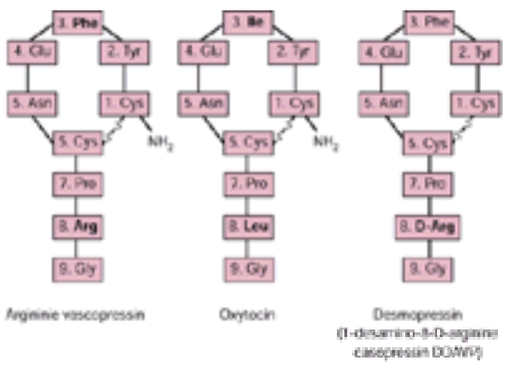

Fig. 2 The structure of vasopressin, oxytocin, and desmopressin. Amino acid differences are highlighted in bold.

Arginine vasopressin

Physiology

Actions of arginine vasopressin

At least three distinct receptors, V1a, V1b, and V2, mediate the actions of AVP. Each of these is a member of the superfamily of seven-transmembrane-domain, G-protein-coupled receptors. The first two signal by inositol-phosphate pathways, whilst the V2 receptor activates adenylate cyclase with an increase in intracellular cAMP. The V2 receptor is expressed almost exclusively in the collecting tubules of the kidney. When AVP binds to the receptor, an increase in intracellular cAMP results in the insertion of a water-conducting channel (aquaporin 2) into the apical membrane of the collecting duct. This allows water to pass from the lumen of the tubule along a concentration gradient into the cell, and then into the renal interstitium, via the AVP-independent aquaporin channels that are constitutionally active in the basolateral cell membrane, thus accounting for the antidiuretic action of AVP. Activation of the V1a receptor in vascular smooth muscle results in vasoconstriction and increases blood pressure but at much higher blood concentrations. Activation of the V1b receptor in the anterior pituitary, in synergy with corticotropin-releasing hormone (**CRH**), is involved in the release of ACTH, but the vasopressin involved in ACTH release derives predominantly from the paraventricular nuclei and represents a separate neuroanatomical system.

Control of arginine vasopressin secretion

There are three principal stimuli to AVP release—a rise in circulating osmolality, a drop in blood pressure, and a stressful event. An increase in plasma osmolality is sensed in the osmoreceptor cells, and this is the major physiological stimulus to the secretion of AVP. There is a tight linear positive correlation between the plasma osmolality and the release of AVP, and a similar relationship holds between osmolality and thirst, which is probably sensed in a centre distinct from, but adjacent to, the osmoreceptor. A loss of extracellular water will stimulate vasopressin secretion to conserve water, accompanied by thirst and a drive to drink. In combination, these processes maintain the plasma osmolality within the narrow range of 285 to 295 mOsmol/kg (Fig. 3). An increase in the plasma sodium concentration is a greater stimulus to AVP secretion than other solutes, and in humans maximum antidiuresis is achieved at plasma vasopressin concentrations between 2 and 4 pmol/l. After ingestion of fluid, there is a fall in the plasma vasopressin levels before a change in the osmolality, which is presumed to occur via a pharyngeal reflex that inhibits AVP release.

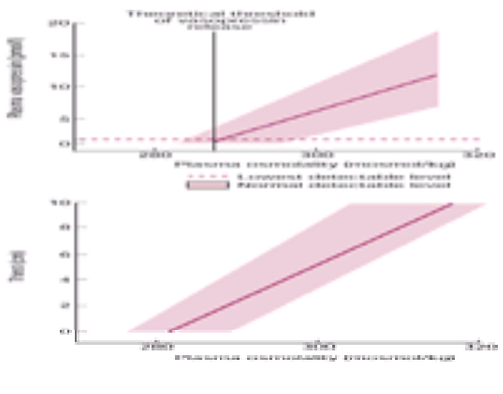

Fig. 3 The relationship between plasma osmolality and plasma arginine vasopressin concentration, and between plasma osmolality and thirst. AVP concentrations and thirst sensation rise in a linear fashion in relation to plasma osmolality. (Reproduced from Besser GM and Thorner MO, 1994, with permission.)

AVP release after a fall in blood pressure, as after haemorrhage, is sensed by baroreceptors in the heart, aorta, and the great vessels. The accompanying water retention helps, together with the sodium retention that follows aldosterone release, to restore the blood volume. AVP is also released under non-specific stress, particularly if associated with nausea or vomiting.

Disorders of arginine vasopressin secretion

Diabetes insipidus

Polyuria with dilute urine may result from vasopressin deficiency (cranial diabetes insipidus, **DI**), resistance to the actions of AVP (nephrogenic DI), and excessive fluid drinking (primary polydipsia). The causes of diabetes insipidus are shown in Table 1. Simple destruction or removal of the posterior pituitary gland, or damage to the distal part of the pituitary stalk, usually results in a temporary DI lasting for 6 weeks to 6 months, since the proximal nerve endings grow out to find systemic capillaries in any scar formed and begin direct secretion again. Upper stalk, median eminence, or more extreme hypothalamic damage results in permanent diabetes insipidus.

An individual deficient in AVP will pass approximately 40 ml/kg of urine in 24 h (between 3 and 20 litres), leading to the clinical features of polyuria, polydipsia, nocturia, and in children nocturnal enuresis. In the complete absence of AVP the maximally dilute urine has an osmolality of approximately 50 mOsmol/kg. As long as there is free access to water, normovolaemia and normonatraemia are maintained by an intact thirst centre. Glucocorticoids are necessary for adequate free-water clearance by the kidneys. Thus if there is a coexistent ACTH deficiency because of a hypothalamic or pituitary lesion, diabetes insipidus may not become manifest until the institution of corticosteroid replacement therapy. It the thirst centre is destroyed as part of the hypothalamic lesion, dangerous dehydration may ensue.

Familial causes are very rare and have been documented as being due to mutations in either the gene encoding the signal peptide or neurophysin molecule, but not of the gene for the AVP peptide itself. Interestingly, symptoms are not usually present at birth but gradually develop between 1 and 6 years of age. It is hypothesized that the changes resulting from such mutations interfere with correct protein folding, and cause retention of a mutant peptide within the neurones which then undergo degeneration, with diabetes insipidus becoming manifest when approximately 80 per cent have been destroyed. Diabetes insipidus is very rarely a presenting feature of an anterior pituitary tumour, although it is frequent after surgical treatment. Therefore, if DI is present at diagnosis of a pituitary mass lesion this should alert the clinician to possible primary hypothalamic lesions such as craniopharyngioma or germinoma.

Differential diagnosis of diabetes insipidus

In a normal individual the plasma osmolality will range from 285 to 295 mOsmol/kg, but urine can be concentrated to more than twice the concentration of plasma. Significant diabetes insipidus is excluded if the urine to plasma (U:P) osmolality ratio is more than 2 to 1, provided that the plasma osmolality is no greater than 295 mOsmol/kg. In diabetes insipidus, despite the raised plasma osmolality, the urine is inappropriately dilute with a U:P ratio below 2. Any cause of polyuria will give a relative resistance to the actions of AVP since the renal medullary concentrating gradient will be washed out. Furthermore, hypercalcaemia or hypokalaemia cause resistance to the actions of vasopressin, possibly by decreasing the level of cAMP within renal tubular cells. The commonest investigation used to discriminate normality from the various causes of diabetes insipidus is the water-deprivation test (Table 2).

There are no contraindications to this test in the fully hydrated patient. Before the test can be performed and adequately interpreted both the thyroid function and adrenal reserve must be normal, or the patient must be on replacement therapy. Impaired renal function, hypercalcaemia, and hypokalaemia need to be excluded. It is important that covert drinking is avoided and that the individual is continuously monitored to prevent severe dehydration. The patient is allowed free access to water overnight to avoid dehydration and the test is started in the morning. Drinking is not allowed for 8 h. Urine and plasma osmolality are followed and close adherence to a defined protocol is essential for safety (Table 2).

Interpretation of the water-deprivation test

A normal response is for the urine volume to gradually fall with fluid deprivation, with a increase in osmolality and a U:P osmolality ratio above 2 towards the end of the test; but the plasma osmolality should remain below 295 mOsmol/kg. In cranial diabetes insipidus there is little rise in urine osmolality, the U:P ratio is less than 2, and frequently the plasma osmolality rises above 295 mOsmol/kg. However, following desmopressin the urine concentrates normally, whereas in nephrogenic diabetes insipidus there is no rise in urine osmolality following desmopressin. Primary polydipsia often gives results that are difficult to interpret since the patient may be water-overloaded at the start of the test, and in addition has a decreased medullary concentrating gradient in the kidney. The combination of these two factors often results in U:P ratios of less than 2 at the end of the test and with a plasma osmolality of less than 285 mOsmol/kg, the threshold for AVP secretion.

In circumstances where the plasma osmolality does not rise above 295 mOsmol/kg, it is still possible that vasopressin secretion is abnormal despite a U:P ratio above 2. Unfortunately this may occur in mild cases of cranial and nephrogenic diabetes insipidus (DI), and in primary polydipsia causing diagnostic confusion. One excellent means to further investigate these patients is to give a 5 per cent hypertonic saline infusion and then measure plasma AVP levels; patients with cranial DI will demonstrate low levels of AVP, whilst the levels of AVP will be increased in those with nephrogenic DI. However, AVP is only measured in a few centres and an alternative is an extended water-deprivation test.

The extended water-deprivation test may be performed in mild cases and if the results from the standard water-deprivation test are equivocal. This is similar to the standard test except that the patient starts somewhat dehydrated, avoiding drinking from 18.00 h the day before the test, which is then started the next morning. Water deprivation is continued until the urine osmolality reaches a plateau (less than 30 mOsmol/kg increase between three consecutive samples), and then desmopressin 2 µg intramuscular is given and the patient allowed to drink. In normal subjects, endogenous AVP secretion under these circumstances is sufficient to maximally concentrate the urine so that no further increase in osmolality is seen after desmopressin. A urine osmolality rise of 9 per cent or more after desmopressin suggests partial cranial DI. Primary polydipsia is suggested when urine concentrates normally if the plasma osmolality rises above 290 mOsmol/kg before desmopressin, with

no further rise following desmopressin in the presence of polydipsia and polyuria.

Treatment of cranial diabetes insipidus

Desmospressin is a synthetic analogue of AVP with high selectivity for the V2 receptor and a prolonged half-life, giving it profound antidiuretic properties and, at antidiuretic doses, little pressor activity. In adults the normal daily dose by intranasal spray is between 10 and 40 µg in divided doses. It is essential to instruct the patient carefully in the administration of the spray so that it may be adequately absorbed from the olfactory mucosa. Oral preparations allow for easier administration, but variability in the bioavailability between patients means that the oral dose may range between 100 and 600 µg daily (and occasionally more) in divided doses. In the perioperative period, or in critically ill patients, it may be necessary to administer desmopressin parenterally at doses of between 1 and 4 µg daily. The main side-effect is hyponatraemia if excess fluid is ingested or administered. Patients should be encouraged to drink only in response to their thirst to avoid water intoxication. All other pharmaceutical treatments are inferior.

Individuals with cranial diabetes insipidus and an intact thirst centre usually do not present particular management problems. However, in the presence of destructive hypothalamic lesions the thirst centre may also be damaged and management of water balance is frequently difficult. In adipsic patients this is best achieved by setting a fixed dose of desmopressin and weighing them when they are normovolaemic. These individuals will then need to be instructed to maintain their daily weight by drinking, taking into account insensible losses, and in hotter climates this may involve twice-daily weighing. In some circumstances, an automated audible reminder to drink may be necessary if hypothalamic damage has affected the patient's memory.

Nephrogenic diabetes insipidus

Nephrogenic diabetes insipidus has diverse causes (Table 1). Congenital nephrogenic diabetes insipidus typically presents with profound polyuria and hypernatraemia from birth, in contrast to congenital cranial diabetes insipidus. The condition needs urgent recognition since repeated episodes of hypernatraemia with polyuria, vomiting, constipation, fever, irritability, and a failure to thrive may result in long-term cognitive impairment. The drive to drink may impair eating and lead to delayed growth. The X-linked condition is associated with mutations of the V2 receptor, whilst in autosomal recessive disease there is deficiency of aquaporin. Other than direct mutational analysis, infusion of desmopressin may allow discrimination between the two types: in the autosomal recessive condition there will be an increase in blood pressure, and in circulating Von Willebrand factor and factor 8 to two to threefold the basal level. As these effects are dependent on intact V2 receptor signalling they will not be seen in the X-linked form.

Treatment of nephrogenic diabetes insipidus

Drugs such as lithium should be withdrawn if possible. Thiazide diuretics reduce the urine output by enhancing sodium excretion to the expense of water and decreasing glomerular filtration rate. Amiloride may need to be coadministered to avoid hypokalaemia. Cyclo-oxgenase inhibitors such as indometacin may also be of benefit.

Syndrome of inappropriate antidiuresis (SIADH)

Excessive and inappropriate secretion of vasopressin either from the posterior pituitary or from ectopic sources, such as small-cell lung cancer, results in inappropriately concentrated urine, dilute plasma, and hyponatraemia with continuing renal sodium excretion. Since there are many causes of hyponatraemia the diagnosis of SIADH should only be entertained in the absence of oedema-forming states, hypovolaemia, or hypotension, and when renal and adrenal function are normal (Table 3). Typically, the condition is asymptomatic during the initial stages, especially if the fall in the serum sodium level is slow. Rapid onset of hyponatraemia is associated with confusion, drowsiness, convulsions, coma, and death. Symptoms are uncommon until the serum sodium falls to 120 mmol/l or less, or the plasma osmolality drops below 268 mOsmol/kg. In cases where vasopressin is secreted from an ectopic source, such as small-cell lung cancer, there is a complete disassociation of the concentration of plasma vasopressin from plasma osmolality. In contrast, in certain central nervous system conditions vasopressin is secreted in a regulated fashion but inappropriately at a low plasma osmolality. This is in keeping with an osmostat that is set at a lower level. Pragmatically, however, the management of the metabolic disturbance is the same whatever the cause. States associated with sodium depletion must be excluded. The urinary sodium concentration will be very low in patients with a low body sodium concentration, unless they are on diuretics, whereas it will be normal or high in those with SIADH.

Management of the syndrome of inappropriate antidiuresis

It is clear that the underlying cause will require treatment on its own merits. Fluid restriction is the cornerstone of treatment: Fluid restriction to between 500 and 750 ml/24h usually reverses any adverse clinical features and restores the circulating sodium level and osmolality to normal. The use of hypertonic saline infusions is very rarely required and only if severe drowsiness or convulsions are experienced, which are unresponsive to fluid restriction and if the serum sodium is around 100 mmol/l. Hypertonic saline should only be administered under close supervision to avoid rapid increases in the plasma sodium concentration and the risk of central pontine myelinolysis. The plasma sodium concentration should rise by no more than 0.5 mmol/l per hour. Drugs such as demeclocycline, which induce nephrogenic diabetes insipidus, may be helpful but the effects are often short lasting.

Oxytocin

Actions of oxytocin

Oxytocin binds to its specific cell-surface receptor expressed predominantly in the myometrial cells of the uterus and breast, and the ductal and epithelial cells of the breast. Expression of the receptor is increased by oestrogen, and receptor numbers increase during pregnancy. In the gravid uterus the myometrial cells are probably maintained in a tonic state of relaxation by the levels of cAMP induced by the actions of placental corticotropin-releasing hormone (CRH). The oxytocin receptor signals via phospholipase C; an increase in intracellular calcium causes phosphorylation of the CRH receptor, resulting in its desensitization and decoupling of CRH receptor signalling. The combined effect of these processes is to stimulate myometrial contraction. This is utilized in obstetric practice where infusions of synthetic oxytocin are used to initiate and sustain labour.

Control of oxytocin secretion

Cervical distension causes the release of oxytocin, which may be involved in the onset of parturition in the human as it is in lower species. In lactating women the release of oxytocin is stimulated by suckling, which then causes contraction of the myoepithelial cells within the breast alveoli and milk ejection.

Further reading

Baylis PH, Thompson CJ (1988). Osmoregulation of vasopressin secretion and thirst in health and disease. *Clinical Endocrinology (Oxford)* **29**, 549–76.

Baylis PH (1994). The posterior pituitary. In: Besser GM, Thorner MO, eds. *Clinical Endocrinology*, pp 5.1–5.14. Mosby-Wolfe, London.

Fujiwara TM, Morgan K, Bichet DG (1995). Molecular biology of diabetes insipidus. *Annual Review of Medicine* **46**, 331–43.

Moses AM, Notman DD (1982). Diabetes insipidus and syndrome of inappropriate antidiuretic hormone secretion (SIADH). *Advances in Internal Medicine* **27**, 73–100.

Oksche A, Rosenthal W (1998). The molecular basis of nephrogenic diabetes insipidus. *Journal of Molecular Medicine* **76**, 326–37.

Thompson CJ, *et al.* (1986). The osmotic thresholds for thirst and vasopressin release are similar in healthy man. *Clinical Science* **71**, 651–6.

Trainer PJ, Besser GM (1995). *The Barts endocrine protocols*. Churchill Livingstone, London.

12.4 The thyroid gland and disorders of thyroid function

Anthony P. Weetman

Structure of the thyroid gland

Development

The human thyroid develops as a diverticulum in the pharyngeal floor around 3 weeks of gestation. This median anlage moves caudally, remaining connected to the pharynx via the thyroglossal duct, which subsequently is obliterated when the thyroid begins to expand as two distinct lobes, around 2 months of gestation. The foramen caecum marks the point in the tongue where the thyroid develops and there is sometimes an upward extension of thyroid tissue from the isthmus, the pyramidal lobe, arising from the lower part of the thyroglossal duct. At the same time, the lateral anlage ultimobranchial bodies, derived from the fifth branchial pouches, fuse with the developing thyroid to which they contribute the parafollicular calcitonin-secreting clear (C) cells. Synthesis of thyroid hormone begins at week 11, at the same time as thyroid stimulating hormone (TSH) production by the pituitary. There is significant maternal-to-fetal thyroxine (T4) transfer so that babies with no endogenous thyroid hormone production are nonetheless largely protected from the adverse effects of fetal hypothyroidism on development of the brain, lung, and skeleton. Preterm infants of less than 27 weeks gestation have immature thyroid function and their neurological development may be improved by temporary thyroxine supplementation.

Anatomy and histology

The adult thyroid weighs 15 to 20 g and each lobe is around 4 cm in length and 2 cm in width, although the right lobe is often larger than the left. The isthmus connecting the two lobes lies just below the cricoid cartilage. The blood supply on each side is derived from the external carotid artery via the superior thyroid artery and from the subclavian artery via the inferior thyroid artery. There is adrenergic and cholinergic innervation which regulates blood flow. The thyroid is attached to the trachea by connective tissue and the recurrent laryngeal nerves lie between the trachea and the posterior aspect of the lobes.

The gland is made up of lobules, each comprising 20 to 40 spherical follicles. The follicles vary considerably in size, but average 200 μm in diameter, and are made up of a single layer of thyroid follicular epithelial cells (Fig. 1). The cells are cuboidal when quiescent and columnar when active, and have a microvillous apical membrane. The follicular lumen contains colloid, the principal constituent of which is the glycoprotein thyroglobulin, secreted by the thyroid cells. Each follicle is surrounded by a rich capillary network. C cells lie scattered between follicular epithelial cells or in the intersititium, and account for around 1 per cent of the epithelial mass.

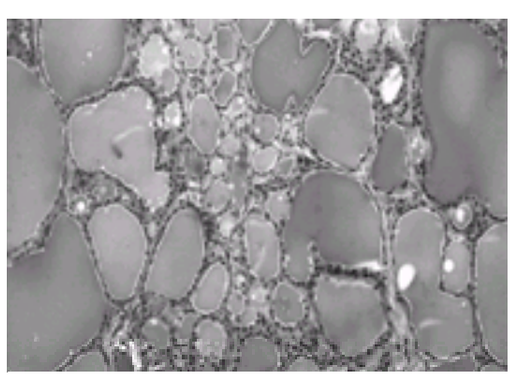

Fig. 1 Photomicrograph showing the histology of a normal thyroid. Thyroid epithelial cells are arranged in follicles containing colloid. (Original magnification ×200; photomicrograph by courtesy of Dr K. Suvarna.)

Thyroid hormone synthesis and metabolism

Synthesis and secretion

Thyroid hormone synthesis requires iodide uptake and oxidation, iodination of certain tyrosine molecules on thyroglobulin, and coupling of the iodotyrosines to form the thyroid hormones triiodothyronine (T3) and T4 (Fig. 2). Iodide is actively transported into the thyroid cell by the Na^+/I^-symporter, which is also expressed in breast tissue and the salivary glands. Perchlorate, thiocyanate, and pertechnetate are also transported by the same symporter and these anions can competitively inhibit iodide uptake. The recommended daily intake of iodine is 150 µg for adults (200 µg during pregnancy) but there is wide variation in actual intake, with many countries having borderline or frankly deficient intakes of less than 50 to 100 µg, while in Western Europe and North America intake is excessive (up to 750 µg/day).

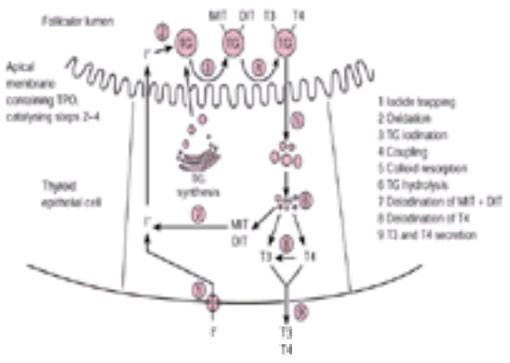

Fig. 2 Steps in the synthesis of thyroid hormones. TG = thyroglobulin, TPO = thyroid peroxidase, MIT = monoiodotyrosine, DIT = diiodotyrosine.

Iodide is oxidized by thyroid peroxidase, a haem-containing enzyme located at the apical border of the thyroid cell, and is rapidly incorporated into tyrosine residues to form monoiodotyrosine and diiodotyrosine. Thyroid peroxidase is also responsible for the coupling of these iodotyrosines, with different sites in the thyroglobulin molecule being responsible for the formation of T3 or T4. Normally, each thyroglobulin molecule contains 3 to 4 T4 molecules, but only 20 per cent of thyroglobulin molecules contain a T3 molecule. Thyroglobulin acts as slow turnover reservoir for thyroid hormone, thus ensuring maximum use is made of often scarce dietary iodine. Around a 7-week supply of T4 is contained in the normal thyroid. Thyroid hormone is released from the gland after endocytosis of colloid and lysosomal hydrolysis of the thyroglobulin to yield T4 and T3, which are secreted from the basal membrane into the capillaries in a ratio of 10:1. Released iodotyrosines are deiodinated for iodide recycling.

Thyroid hormone transport

Up to 90 per cent of the total T3 in the circulation is derived from peripheral conversion of T4 to T3 by deiodinase enzymes (see below) rather than thyroid secretion. Only 0.03 per cent of T4 and 0.3 per cent of T3 in the circulation exist as free hormone, able to diffuse into tissues. The remainder is protein bound. T4 binds predominantly to thyroxine-binding globulin, and to a lesser extent to transthyretin (or prealbumin); a little is bound to albumin. T3 binds to thyroxine-binding globulin and albumin, with little bound to transthyretin. Alteration in the concentration or binding capacity of thyroid hormone binding proteins can produce major changes in total but not free thyroid hormone levels (Table 1).

Metabolism of thyroid hormone

The half-life of T4 in the circulation is 7 days, contrasting with the much shorter half-life of T3 (24 h). The most important metabolic pathway for T4 is outer ring (5') deiodination to T3 (Fig. 3). This is catalysed by type 1 and 2 deiodinase, while type 3 deiodinase catalyses inner ring (5) deiodination leading to hormone inactivation. Type 1 deiodinase can also catalyse inner ring deiodination of T3 and T4. All three enzymes have a selenocysteine moiety as the active catalytic site. Type 1 deiodinase is expressed predominantly in the liver, kidney, thyroid, and brain, type 2 in the pituitary, brain, placenta, skeletal muscle, and heart (tissues critically dependent on thyroid hormone for development or function), and type 3 in the brain, placenta, and skin. The type 1 deiodinase is largely responsible for the generation of circulating T3 from T4, whereas T3 generated by the type 2 enzyme mainly provides intracellular T3 at specific sites.

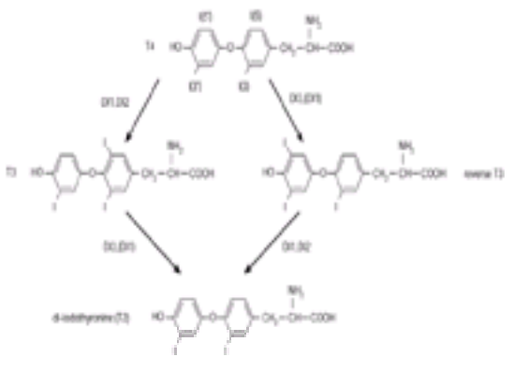

Fig. 3 Main deiodination pathway for thyroid hormones. DI = deiodinase enzyme; parentheses denote a minor contribution. Deiodination of T3 also yields 3,5-T2 and deiodination of reverse T3 also yields 3',5'-T2. T2 is further deiodinated to monoiodothyronine and thyronine.

Around 40 per cent of T4 is metabolized to T3 and 40 per cent is converted to reverse T3 by the type 3 deiodinase. This same enzyme is responsible for the main metabolic pathway for T3 which is converted to 3,3'-dioiodothyronine. Starvation, trauma, and drugs (propylthiouracil, amiodarone, glucocorticoids, propranolol) impair T4 to T3 conversion and must be borne in mind in interpreting tests of thyroid function (see below). In addition to deiodination, a small proportion of thyroid hormone is metabolized by conjugation of the phenolic hydroxyl group with sulphate or glucuronic acid, which increases water solubility and allows urinary and biliary excretion. Biliary iodothyronine glucuronides can be reabsorbed, constituting an enterohepatic cycle.

Thyroid hormone action

Thyroid hormone acts as a transcription regulatory factor, mediated by T3 binding to nuclear receptor isoforms which belong to the same superfamily as steroid and retinoic acid receptors. All such receptors possess a conserved DNA-binding domain, containing two zinc fingers which interact with specific DNA response elements, and a hormone-binding domain. Alternative splicing results in two pairs of thyroid hormone receptor (Fig. 4) whose tissue expression varies during development. Thyroid hormone receptors bind to DNA as homodimers or heterodimers (with the retinoid X receptor). Without ligand, basal gene transcription is inhibited by a corepressor. When T3 binds, homodimers dissociate, releasing corepressor and allowing gene transcription; the stable heterodimer binds coactivators in the presence of T3 with the same outcome. The a2 thyroid hormone receptor does not bind T3 and may act as a natural inhibitor of receptor activity. Knockout mice devoid of all known thyroid hormone receptors have an abnormal pituitary–thyroid axis and impaired growth and bone maturation, but not the severe manifestations of complete hypothyroidism, indicating the potential for other mechanisms of thyroid hormone action.

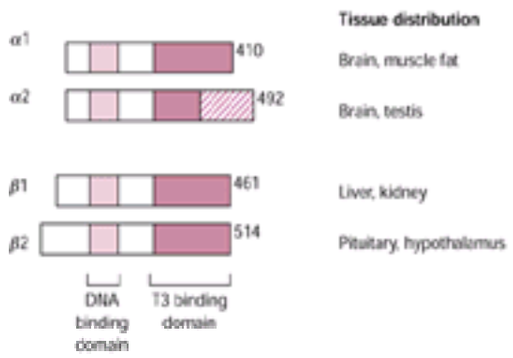

Fig. 4 Structure of the thyroid hormone receptors. The numbers indicate the amino acid content. Homologous areas are shaded; the lack of homology in the T3-binding domain of the a2 receptor (hatched area) prevents T3 binding and the function of this receptor is unknown.

Regulation of thyroid function

The main regulator of thyroid function is TSH (thyrotropin), secreted by thyrotrophs in the anterior pituitary in response to the tripeptide thyrotropin-releasing hormone (TRH), derived from the hypothalamic supraoptic and paraventricular nuclei. Thyroid hormones exert a classical negative feedback effect on thyrotrophs; the acute effect is mediated by T3 in the pituitary which is derived from T4 by type 2 deiodination. Thyroid hormones also inhibit hypothalamic TRH synthesis. TRH-stimulated TSH secretion is inhibited by dopamine and somatostatin, while a-adrenergic activation stimulates TSH release. Cytokines, particularly interleukin-1, interleukin-6, and tumour necrosis factor, inhibit TSH synthesis and may be responsible for the suppression of TSH seen in severe illness.

Within the thyroid, TSH binds to the G protein-coupled TSH receptor, leading to intracellular signalling predominantly via cyclic AMP. TSH increases iodide transport and organification, endocytosis of colloid and thyroid hormone secretion, as well as thyroid follicular epithelial cell division. Autoregulatory mechanisms can modulate thyroid function when TSH levels are constant. The most important is iodine intake. Increased iodide transport transiently decreases organification and reduces thyroid hormone synthesis (the Wolff–Chaikoff effect); after several weeks under normal conditions, the thyroid escapes and resumes hormone production. Sudden increases in iodine intake can also block thyroid hormone release acutely. In iodine deficiency, thyroid hormone production is switched to preferential T3 synthesis, but this effect is largely TSH-mediated rather than autoregulatory.

Laboratory investigation of thyroid function

Determining thyroid status

The introduction of sensitive immunoradiometric assays for circulating TSH, with a detection level of 0.1 mU/l or less, has transformed the evaluation of thyroid status. A normal TSH level rules out primary thyroid dysfunction. Low levels of thyroid hormones elevate TSH as a result of negative feedback, while excessive thyroid hormone suppresses TSH. The TRH test for detecting low TSH levels is now redundant. Besides primary thyroid disorders, other conditions may alter TSH levels and must be borne in mind when using TSH as a screening test for thyroid dysfunction (Table 2), as must the possibility of secondary (pituitary or hypothalamic) disturbances of thyroid function.

It is therefore essential to confirm thyroid status when TSH levels are abnormal, or when pituitary or hypothalamic abnormalities are possible, by measuring circulating thyroid hormone levels. Methods which measure total T3 or T4 are prone to artefacts caused by abnormal thyroid hormone binding (Table 1), although in the absence of such abnormalities these tests are reliable. When altered binding is suspected or found, compensation can be made by calculation of the free T3 or free T4 index. These indices are derived from the total hormone levels and measurement of the differential distribution of radiolabelled T3 between unoccupied protein binding sites in the sample and an absorbent resin (hence the term resin uptake test). Thyroxine-binding globulin levels can also be measured directly.

However, the ready availability of immunoassays for free T3 and free T4 has generally supplanted these methods. The immunoassays rely on the ability of a radiolabelled thyroid hormone analogue to bind to thyroid hormone antibody but not to plasma binding proteins. The analogue then competes for antibody binding with the free thyroid hormone in the sample. Despite initial concerns about the theoretical basis and performance of such assays, recent improvements allow generally reliable estimation of free thyroid hormones. In cases of doubt, free hormone levels can be measured by physical separation from bound hormone, using ultracentrifugation or equilibrium dialysis.

Several indirect methods can be used to determine thyroid status. The thyroidal uptake of radioiodine (^{123}I, ^{131}I) or ^{99m}Tc-pertechnetate is increased in hyperthyroidism and decreased in hypothyroidism, but can be affected by excessive dietary iodine and destructive processes in the thyroid, so that uptake is low when the patient is thyrotoxic (see Destructive thyroiditis below). Serum thyroglobulin is raised in hyperthyroidism of all types but is also raised in destructive thyroiditis and thyroid cancer. Its main role in investigation is follow-up of treated thyroid cancer (see Chapter 12.5). A number of non-specific tests have also been used to determine end organ responses to thyroid hormones, including basal metabolic rate, tendon relaxation time, and serum levels of cholesterol, ferritin, sex hormone-binding globulin, and liver enzymes.

Thyroid function in non-thyroidal illness and pregnancy

Assessing thyroid function in severely ill patients often reveals abnormalities termed the sick euthyroid syndrome. Many of the changes are due to cytokine release, but therapeutic agents such as dopamine and glucocorticoids also contribute, as do unknown factors. Any major, acute illness or starvation can result in a decrease in circulating T3 (total and free) with normal levels of T4 and TSH. Reverse T3 levels rise. The severity of the illness correlates with the magnitude of the fall in T3, and in very sick patients total T4 levels also fall. Analogue-based free T4 assays generally produce normal results but sometimes high or low values occur. In 10 to 15 per cent of sick individuals, TSH levels are abnormal (raised or lowered). Psychiatric illness can be associated with raised total and free T4 levels with normal T3.

There is no proven benefit from thyroid hormone administration in the sick euthyroid syndrome and the hormone changes may be protective, limiting catabolism (although this view is regularly challenged). The importance of these alterations lies in their potential to cause diagnostic confusion. Thyroid function tests should only be requested in ill patients when thyroid disease is genuinely suspected. Abnormal thyroid function tests due to the sick euthyroid syndrome return to normal after recovery and therefore repetition of testing is the simplest way to confirm the reason for unusual results.

Pregnancy also affects thyroid function testing. The most obvious change is the rise in thyroxine binding globulin, which elevates total but not free T3 and T4 levels. In addition, the reference ranges for free T3 and T4 are higher than normal in the first half of pregnancy, because placental human chorionic gonadotropin (hCG), at high levels, acts as a weak stimulator of the TSH receptor. There is a reciprocal fall in TSH levels during the first trimester, but TSH returns to normal in the second trimester as hCG levels decline. Occasionally, these changes are sufficient to cause transient 'gestational' hyperthyroidism, associated with hyperemesis gravidarum. Antithyroid drugs are usually unnecessary in this condition, and attention should be directed to controlling the vomiting and giving parenteral fluids. Renal clearance of iodine is increased in pregnancy, leading to maternal and neonatal goitre and mild hypothyroidism in areas where iodine intake is marginal (50 µg/day). These complications can be prevented by supplemental iodine, 150 to 200 µg/day.

Determining the cause of thyroid dysfunction

The most frequent cause of thyroid dysfunction in iodine-sufficient areas is autoimmunity and the simplest test for this is measurement of thyroid autoantibodies, particularly those directed against thyroid peroxidase (the 'microsomal' antigen). Antibodies against thyroglobulin are also easily measured but are almost always accompanied by thyroid peroxidase antibodies, so testing for the latter alone is usually adequate. Different methods, including haemagglutination, immunofluorescence, radioimmunoassay, and enzyme-linked immunosorbent assay, give different prevalence rates for thyroid autoantibodies. Almost all patients with autoimmune hypothyroidism, and around 75 per cent with Graves' disease, have thyroid peroxidase antibodies. Generally lower levels are found in 5 to 15 per cent of healthy women and 2 per cent of men, and in slightly higher proportions of patients with nodular goitre and thyroid cancer, and results therefore need to be interpreted carefully. Individuals with positive thyroid autoantibodies but normal thyroid function are at increased risk of developing autoimmune hypothyroidism (around 2 per

cent per year).

Thyroid imaging by scintiscanning is useful in determining the aetiology of thyroid disease when this is not obvious clinically, particularly in hyperthyroidism and ectopic thyroid tissue. Its role in the evaluation of a solitary thyroid nodule is considered in Chapter 12.5. ^{99m}Tc is usually used as it has a short half life (6 h) which allows safe administration of high activity and rapid scanning. ^{123}I is not as readily available but is also preferable to ^{131}I, especially in children, as it too has a short half life and does not emit beta radiation. ^{131}I imaging is particularly useful in planning treatment of thyroid cancer and localization of metastases. The place of ultrasound imaging is less clear. The technique allows accurate determination of thyroid size, which may be useful in follow-up of goitre, and can help to determine the nature of an atypical neck mass. Its role in evaluating nodular thyroid disease is considered in Chapter 12.5. CT scanning is particularly valuable in determining the extent of a retrosternal goitre and assessing tracheal compression (Fig. 5). In contrast, a standard chest radiograph can be misleading in evaluating tracheal compression, particularly in the anterior–posterior plane.

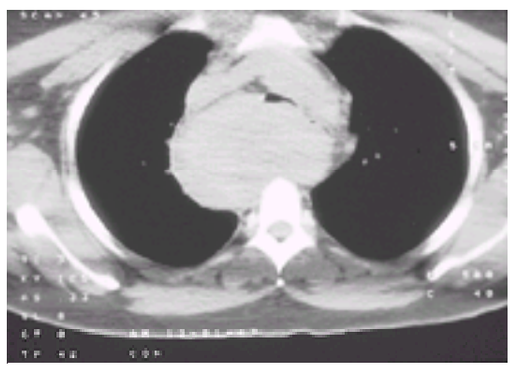

Fig. 5 CT scan of the chest in a patient with a large retrosternal goitre causing tracheal compression.

Goitre

The distribution of thyroid size in any population forms a continuous, positively skewed curve, whose shape depends on the age, sex, and country of residence of the individuals assessed. Hence a precise definition of goitre is impossible. Ultrasound is the most accurate method to assess thyroid size and estimates of goitre prevalence based on inspection and palpation underestimate the true frequency. However, simple schemes, such as that used by WHO/UNICEF, are useful in field studies of goitre prevalence:

- Grade 0 = no visible or palpable thyroid
- Grade 1 = thyroid enlargement that is palpable but not visible when the neck is in the neutral position
- Grade 2 = thyroid enlargement which is both visible and palpable with the neck is in the neutral position
- Grade 3 = goitre visible at a considerable distance

Of the many causes of goitre (Table 3), those associated with disturbances of thyroid function are considered later. The remainder can be classified broadly as endemic and sporadic non-toxic goitres.

Endemic goitre

Prevalence

Goitre is said to be endemic when the prevalence exceeds 10 per cent in children aged 6 to 12 years, although this figure is arbitrary and it has recently been suggested that a prevalence of more than 5 per cent should be used.. Over 200 million people are affected world-wide, especially in the Himalayas, Andes, and parts of Africa, although Eastern and Southern Europe are also involved.

Aetiology

The main cause is iodine deficiency, with goitre prevalence exceeding 30 per cent in areas with very low iodine intakes (<30 µg/day). However, endemic goitre is not exclusively related to iodine deficiency. Naturally occurring goitrogens, such as those in vegetables of the cabbage family and in cassava, exaggerate the effects of iodine deficiency by the action of thiocyanates and cyanoglucosides, respectively, on iodine transport. Where selenium and iodine deficiency coincide, thyroid cell destruction and gland fibrosis minimize goitre formation. In Japan, endemic goitre actually results from iodine excess, as well as goitrogens in seaweed, and in Kentucky, chemical pollution of water is goitrogenic.

Clinical presentation

Diffuse goitre is more frequent in girls, and gradually becomes nodular with age and increasing iodine deficiency. Endemic goitres can be massive but give few compressive symptoms. In areas of marginal iodine deficiency, such as Belgium, goitre only appears, and is then modest, when demands on thyroidal iodide metabolism are increased during puberty or in pregnancy.

The major impact of endemic goitre and iodine deficiency on health is the association with endemic cretinism. Two forms of cretinism can be delineated in separate geographical areas, but there is considerable overlap. Firstly, when maternal iodine intake is severely reduced, causing hypothyroidism, there is reduced placental transfer of T4 to the fetus, resulting in a profound neurological deficit in the infant, with mental deficiency, deafness, speech defects, and spastic gait. Secondly, hypothyroidism in the infant after birth produces the typical cretin features, in particular stunted growth. The thyroid in cretins may be enlarged or atrophic and it is clear from field studies that iodine deficiency alone cannot account for the multiple forms of endemic cretinism.

Management

Iodine supplementation is perhaps the simplest and cheapest of remedies and the condition it prevents has devastating consequences; it is sobering that iodine deficiency still persists. There are few complications from iodine supplementation, although thyrotoxicosis may result in a variable proportion of individuals (the Jod–Basedow phenomenon), some of whom have avoided this previously through lack of sufficient iodine. It is political, social, and economic inertia which are at the heart of continuing iodine deficiency. Effective programmes are best targeted at women intending pregnancy and children. Iodization of salt or bread is widely used in developed nations, but intramuscular or oral iodized oil as a single annual dose, or iodination of drinking water, is preferable in areas where distribution of iodized foodstuffs is a problem.

Sporadic goitre

Prevalence

Goitre occurs in around 5 per cent of the iodine-sufficient population and is four times more common in women. However, the prevalence varies with area and generally declines with age; over 60 per cent of goitres found in adolescents regress over the next 20 years. The character also changes over time, from a diffuse (sometimes called simple) goitre to a multinodular goitre. The presentation of single thyroid nodules is dealt with in Chapter 12.5, but it is worth mentioning here that solitary thyroid nodules increase in frequency with age.

Aetiology

The aetiology of sporadic goitre is largely unknown. Unidentified goitrogens may be responsible in a few patients, and in others, mild iodine deficiency in infancy may initiate goitrogenesis which persists despite a subsequently normal iodine intake. A large proportion are probably the result of mild defects in hormone synthesis; compensatory growth ensures normal thyroid function and current tests cannot identify the nature of the defect. Familial clustering of sporadic goitre supports this idea. Although TSH is the most obvious thyroid growth factor, TSH levels by definition are normal in sporadic goitre, which may therefore be the result of other autocrine and paracrine growth factors (e.g. insulin-like growth factor-1, epidermal growth factor, fibroblast growth factor). A role for growth stimulating autoantibodies has also been suggested, but remains controversial.

Progression to a multinodular goitre occurs when unencapsulated nodules form in a long-standing diffuse goitre. These nodules contain colloid-rich, polyclonal follicles, and are usually distinct from adenomas, which are encapsulated and derived from a single thyroid follicular cell with a somatic mutation conferring growth advantage. However, some goitres contain both nodules and adenomas, suggesting a spectrum of pathological changes. Because thyroid follicular cells are heterogeneous, nodules generally develop with varying degrees of function, giving rise to 'hot' and 'cold' areas on scintiscanning with radioiodine. Some nodules develop autonomy and may eventually cause hyperthyroidism, completing the evolution from non-toxic to toxic multinodular goitre (see below). Other nodules undergo degeneration with haemorrhage, fibrosis, and cyst formation.

Clinical presentation

Patients usually seek attention because of the appearance of the neck or a sensation of pressure or discomfort. Equally, they may be unaware of a long standing small goitre which is noticed on examination. Careful palpation is sufficient to distinguish true goitre, which moves on swallowing, from a prominent pad of fat over the front of the neck. Very large goitres can cause dysphagia or even stridor when the trachea is compressed, but these symptoms are uncommon. Venous compression at the thoracic inlet is even rarer; this sign is exacerbated by asking the patient to raise her arms (Pemberton's sign). Pain in the thyroid, which radiates to the jaw, is uncommon and suggests either destructive thyroiditis (see below) or haemorrhage into a cyst in a multinodular goitre. In the latter, the pain is usually unilateral, acute, and associated with a rapid change in thyroid size; symptoms resolve spontaneously in a few days.

Investigations

Thyroid function should be assessed by checking TSH levels, and then free T3 and T4 levels if the TSH is abnormal, to rule out goitre associated with thyroid dysfunction. The presence of thyroid peroxidase antibodies is also useful as a marker of an underlying autoimmune thyroiditis, which occurs in 10 to 20 per cent of multinodular goitres. Imaging, in my view, has only a limited place in the investigation of sporadic goitre, although in many centres, scintiscans and ultrasound are widely used. Ultrasound is useful in determining thyroid size accurately and may reassure an anxious patient that the thyroid is not enlarging. In most cases it is not necessary. Thyroidal uptake of radioisotopes (especially ^{99m}Tc) is indicated if destructive thyroiditis is suspected as a cause of goitre. Otherwise, the major role for imaging is to ensure there is no tracheal compression or intrathoracic/retrosternal component in a patient with suggestive symptoms, and a CT scan is then the preferred investigation (Fig. 5).

Treatment

Most patients with euthyroid sporadic goitre do not require treatment. Neck discomfort or cosmetic concerns may prompt intervention but it is necessary to take a careful history to ensure that discomfort or difficulty swallowing is indeed caused by the goitre. There is no cost–benefit analysis which shows the superiority of any single treatment. Thyroxine, given at doses to maintain slightly suppressed TSH levels (0.1 to 0.3 mU/l), leads to a reduction in goitre size in up to 60 per cent of patients but is unlikely to have any effect on a very nodular goitre or when the TSH level is already low (so-called subclinical hyperthyroidism, discussed below). There are concerns about the long-term effects of suppressive doses of thyroxine on the heart and skeleton, and treatment must be continued long-term to maintain any improvement.

Radioiodine has recently been used in some centres, with doses of ^{131}I ranging from 600 to 3400 MBq (hospitalization is required for doses greater than 800 MBq). Goitre size is usually reduced by more than 50 per cent at 2 years, and most of the improvement occurs within 2 to 3 months. Tracheal compression by a goitre can be treated with ^{131}I, despite theoretical concerns over acute worsening due to a radiation thyroiditis. However, there are as yet no long-term follow-up data on such patients, although hypothyroidism certainly occurs in 20 to 40 per cent by 5 years.

Surgery is used in other centres, and is particularly indicated for severe tracheal compression or retrosternal goitres, and if there is any suspicion of malignancy. Subtotal thyroidectomy is undoubtedly effective, but goitre may recur in around 20 per cent of patients within 10 years and does not seem avoidable by giving thyroxine replacement. Complications, including recurrent laryngeal nerve damage, hypoparathyroidism, and hypothyroidism, are more likely with the biggest goitres, near total thyroidectomy, and reoperation.

Hypothyroidism

Introduction

Impaired production of thyroid hormones is usually due to a primary abnormality of thyroid gland or iodine deficiency; occasionally it is secondary to pituitary or hypothalamic disorders, dealt with in Chapter 12.2 and Chapter 12.3. The onset of primary hypothyroidism is gradual and may be detected when the TSH is elevated (to compensate for impaired thyroid output) but the free thyroid hormone levels are normal. This state is subclinical hypothyroidism. As thyroid damage continues, TSH levels rise further but free T4 levels fall. The TSH at this stage is usually greater than 10 mU/l, symptoms become apparent, and the patient is said to have overt or clinical hypothyroidism.

Aetiology

The causes of hypothyroidism are listed in Table 4. The commonest cause world-wide is iodine deficiency, discussed in the preceding section. In iodine sufficient areas, autoimmune hypothyroidism and thyroid damage after radioiodine or surgical treatment for hyperthyroidism are the major causes.

Epidemiology

The prevalence of overt hypothyroidism in Caucasians is around 2 per cent in women and 0.2 per cent in men, with a mean age of 60 at diagnosis. Subclinical hypothyroidism is even more common (6 to 8 per cent of women and 3 per cent of men). Around 4 per cent of these individuals progress to overt hypothyroidism annually if thyroid peroxidase antibodies accompany the elevated TSH. Half this number progress in the absence of thyroid peroxidase antibodies. Focal lymphocytic infiltration of thyroid, associated with thyroid autoantibody positivity, occurs in up to 15 per cent of healthy women and 2 per cent of men without an elevated TSH, representing the earliest manifestation of thyroid autoimmunity; 2 per cent of these people progress to overt hypothyroidism annually. Congenital hypothyroidism occurs in about 1 in 4000 births and this high frequency has led to the widespread introduction of neonatal screening.

Pathogenesis

Autoimmune hypothyroidism is primarily the result of autoreactive T-cell-mediated cytotoxicity directed against thyroid follicular cells. Cytokines derived from the locally infiltrating T cells, macrophages, and dendritic cells impair thyroid cell function and enhance T-cell-mediated cytotoxicity. The role of thyroid autoantibodies in thyroid cell destruction is unclear, but thyroid peroxidase antibodies fix complement and may cause secondary damage. In 10 to 20 per cent of patients, antibodies which block the TSH receptor are partially or wholly responsible for hypothyroidism and transplacental passage of these antibodies (but not thyroid peroxidase antibodies) occasionally causes transient neonatal hypothyroidism. Genetic and environmental factors are involved in the aetiology but, as with most autoimmune disorders, the complex interaction of these factors has so far prevented a full understanding. Polymorphisms in the HLA-DR and CTLA-4 genes are associated with autoimmune hypothyroidism, and a high iodine intake may be an important environmental factor in some cases.

Congenital hypothyroidism is caused by thyroid aplasia or hypoplasia in 60 per cent of cases and in 30 per cent there is an ectopic gland. Mutations in thyroid-specific

transcription factors have been found in some of these cases. In the remaining 10 per cent, hypothyroidism is due to dyshormonogenesis (Table 4).

Clinical features

The cardinal features in adults with hypothyroidism are shown in Table 5. However, the ready availability of reliable screening tests for hypothyroidism, especially TSH assays, has led to the recognition of many patients in whom there are only vague or non-specific symptoms, such as tiredness, weight gain, and poor concentration. The differential diagnosis is accordingly vast but the high frequency of hypothyroidism should prompt its exclusion when any suggestive features are present, particular in middle aged women with chronic fatigue or depression.

Autoimmune hypothyroidism may present with a goitre (Hashimoto's thyroiditis) or without (atrophic thyroiditis or primary myxoedema). When present, the goitre is of variable size but is often hard and irregular, sometimes giving rise to suspicion of a malignancy, which then requires exclusion by fine needle aspiration biopsy. Primary lymphoma of the thyroid is a rare but important association (Chapter 12.5). Thyroid pain due to autoimmune thyroiditis is also a rare complication. Patients may notice a Hashimoto goitre before any thyroid dysfunction has developed and annual follow-up is then needed.

The most dramatic presentation of hypothyroidism is myxoedema coma, which is fortunately rare. In addition to the usual features, there is hypothermia (as low as 23°C) and coma, sometimes with seizures. Mortality is 50 per cent even with intensive treatment. Patients are typically elderly and either previously undiagnosed or poorly compliant with medication. There is generally an additional precipitant, such as respiratory depression due to drugs, chest infection, heart failure, stroke, blood loss, or exposure to cold.

Autoimmune hypothyroidism is frequently associated with other autoimmune conditions. In the type 2 autoimmune polyglandular syndrome, autoimmune thyroid disease (hypothyroidism or Graves' disease) is associated with type 1 diabetes mellitus and/or Addison's disease. This syndrome is autosomal dominant with variable penetrance. In the rare, autosomal recessive type 1 autoimmune polyglandular syndrome (chronic mucocutaneous candidiasis, Addison's disease, and hypoparathyroidism), autoimmune hypothyroidism is found in 5 to 10 per cent of patients. Other common associations include pernicious anaemia, vitiligo, and alopecia areata and there is a significant excess of autoimmune hypothyroidism in coeliac disease, dermatitis herpetiformis, chronic active hepatitis, rheumatoid arthritis, systemic lupus erythematosus, and Sjögren's syndrome. Breast cancer patients and individuals with Down's and Turner's syndromes have a higher than expected frequency of thyroid autoimmunity. Around 5 per cent of patients with thyroid-associated ophthalmopathy, discussed later in this chapter, have autoimmune hypothyroidism and 15 per cent of patients with Graves' disease successfully treated with antithyroid drugs develop hypothyroidism 10 to 20 years later. This relationship with Graves' disease is further emphasized by rare patients who oscillate between hyper- and hypothyroidism over a period of months. The likely explanation is fluctuation in the relative levels of TSH receptor stimulating and blocking antibodies, but the cause of these changes is unknown.

Juvenile hypothyroidism is uncommon. The features of adult hypothyroidism (Table 5) may be present, but the diagnosis is usually suggested by retarded growth and dentition, and an infantile face. Myopathy with muscle enlargement is common. Puberty is usually delayed, yet sometimes is precocious. Congenital hypothyroidism is typically unrecognizable at birth but, if not identified by screening, gives rise to prolonged jaundice, failure to thrive, impaired growth, feeding difficulties, constipation, and hypotonia. Left untreated, even for a few weeks after birth, there is permanent neurological damage, resulting in intellectual impairment.

Pathology

In Hashimoto's thyroiditis there is a prominent diffuse and focal lymphocytic infiltrate with germinal centre formation. The thyroid follicles show varying degrees of destruction and little or no colloid. The remaining thyroid follicular cells have an increased number of mitochondria, giving rise to oxyphil metaplasia (Askanazy or Hürthle cells). There is a variable degree of fibrosis. In atrophic thyroiditis, fibrosis is the most prominent feature, with a less obvious lymphocytic infiltrate than in Hashimoto's thyroiditis. Thyroid follicles are usually sparse, reflecting the later stage at which this form of autoimmune hypothyroidism is diagnosed. Whether there is a natural progression from Hashimoto's to atrophic thyroiditis is unclear, although the goitre usually decreases with thyroxine replacement.

Laboratory diagnosis

Measuring the serum TSH is the first step in diagnosing hypothyroidism, with the important caveat that this approach will miss most cases of secondary hypothyroidism in which the serum TSH measured by immunoassays may be low, normal, or even slightly raised, due to the secretion of bioinactive forms of the hormone. If secondary hypothyroidism is suspected, for instance in the follow-up of a patient with treated pituitary disease, it is essential to check the free T4 level. The TSH is elevated in other settings besides primary overt hypothyroidism (Table 2). It is therefore important to confirm the diagnosis by measuring the free T4 in all samples in which the TSH is elevated. Measurement of free T3 adds nothing to the diagnosis, especially as values may be within the reference range in a quarter of hypothyroid patients, due to extrathyroidal conversion of T4.

If myxoedema coma is expected, it is essential that treatment is initiated immediately without awaiting confirmation of the diagnosis. These patients often have dilutional hyponatraemia, hypoglycaemia, and electrocardiography changes (low voltage, prolonged QT interval, flat or inverted T waves, and heart block). Other non-specific features which may be found in any patient with hypothyroidism are elevation in serum liver and muscle enzymes (the raised creatine phosphokinase particularly may cause unnecessary concern), raised cholesterol, and anaemia. The anaemia is usually normocytic or macrocytic, but microcytosis occurs when hypothyroidism is accompanied by menorrhagia.

The aetiology is usually easily established. In the absence of a history of treated hyperthyroidism or iodine exposure, the majority of juvenile or adult onset primary hypothyroidism in iodine-sufficient countries is due to autoimmune hypothyroidism. Transient hypothyroidism due to destructive thyroiditis is considered later. The diagnosis of autoimmune hypothyroidism is confirmed by the presence of thyroid peroxidase antibodies, usually at high levels, although occasionally these antibodies are absent. Cytological diagnosis of Hashimoto's thyroiditis is possible using fine needle aspiration biopsy, but is only necessary if there is uncertainty over the cause of a nodular goitre.

Once congenital hypothyroidism is diagnosed by routine testing after birth, it is usual to initiate thyroxine immediately. Treatment can then be stopped without neurological consequences at age 3 to 4 years to establish that life-long thyroxine replacement is necessary. At this time, the aetiology can be established by scintiscanning and/or ultrasound. Dyshormonogenesis, suspected when there is detectable thyroid tissue and a family history, requires specialized investigation to establish the diagnosis and increasingly this is possible by direct analysis of gene mutations. The commonest of these defects is Pendred's syndrome in which there are mutations in the pendrin gene encoding a chloride/iodide transporter present in the thyroid and cochlea, leading to goitre, mild hypothyroidism, and deafness. The thyroid abnormalities usually appear in the second or third decade, rather than at birth. The diagnosis can be made easily by the perchlorate discharge test, which shows an excessive decline of radioactivity in the thyroid when potassium perchlorate is given 2 to 3 h after allowing the thyroid to take up a tracer dose of radioiodine.

Treatment

In adult patients without heart disease and below the age of 60, treatment can begin with the estimated replacement dose of thyroxine. If there is no remaining thyroid tissue (indicated by a very high TSH and very low or undetectable free T4), the full replacement dose is 1.6 µg thyroxine/kg body weight, which is around 100 to 150 µg/day. In practice, the typical starting dose is 50 to 100 µg thyroxine daily, the lower dose being reserved for patients with mild to moderate biochemical abnormalities. Dosage changes should be based on TSH levels measured 2 to 3 months after starting treatment, the main goal of treatment being to normalize the TSH. A similar period is required to assess the effect of any change to the dosage, made as 25 or 50 µg increments or decrements depending on how abnormal the TSH is. Treatment is usually straightforward, although if there is only partial thyroid failure when treatment is begun, the dose of thyroxine may require adjustment over many months.

Once on a full replacement dose, TSH levels should be checked at intervals of 1 to 3 years, depending on their stability. Fluctuating or elevated TSH levels in a previously stable patient, or thyroxine requirements in excess of 200 µg/day, usually indicate compliance problems. It is important to rule out malabsorption or abnormal thyroxine kinetics caused by drugs: cholestyramine, ferrous sulphate, lovastatin, aluminium hydroxide, rifampicin, amiodarone, carbamazepine, and phenytoin all alter the absorption or clearance of T4. A common cause for poor compliance is worsening angina. Optimization of antianginal treatment is then required, although some patients may simply prove intolerant of full thyroxine replacement if their coronary artery disease is extensive and irremediable. It is important to remind poorly compliant patients that, because of the long half-life of thyroxine, missed tablets should always be taken and that this is safe.

In the elderly, or in individuals with heart disease, the usual starting dose is 25 µg thyroxine daily (or on alternate days when there is severe angina). Dosage should

be increased slowly with increments of 12.5 to 25 µg thyroxine. Proportionately higher doses of thyroxine are needed during the first year of life than in adults, and the starting daily dose of thyroxine for congenital hypothyroidism is 10 µg/kg body weight. There is a continuing debate on the benefit of thyroxine in subclinical hypothyroidism. It is reasonable to commence thyroxine when subclinical hypothyroidism is coupled with the presence of thyroid peroxidase antibodies, as there is a high risk of progression to overt hypothyroidism. Modest improvements in mental function and lipid levels occur when thyroxine is given to some patients with subclinical hypothyroidism, but long-term studies on the benefits of treatment have not been conducted. At present, it seems reasonable to offer a 3-month trial of thyroxine to thyroid peroxidase antibody-negative patients with subclinical hypothyroidism. If the patient notices an improvement in the symptoms which prompted thyroid function testing, thyroxine is continued, but is stopped if there is no benefit. All patients with subclinical hypothyroidism or positive thyroid peroxidase antibodies should be offered annual testing for the development of overt hypothyroidism.

Another problem is posed by the occasional patient with overt hypothyroidism who continues to feel unwell or who fails to lose weight after the TSH is normalized with thyroxine replacement. It can take around 3 months from achieving full replacement for all symptoms to disappear, and weight gained during hypothyroidism will generally only be lost by following an appropriate diet. It is sensible to ensure that the TSH level is in the lower half of the reference range and sometimes a small increment of thyroxine can achieve this, improving symptoms but not suppressing the TSH. More controversial is the treatment of such patients with higher doses of thyroxine which suppress the TSH. This approach may resolve symptoms but at the risk of atrial fibrillation due to subclinical thyrotoxicosis. The other recognized adverse effect of excessive thyroxine is a decrease in bone mineral density, particularly in postmenopausal women who have previously had hyperthyroidism and therefore already have a low skeletal mass. However, the changes in bone mineral density are modest and no increase in fracture rate has been reported as a result of thyroxine given at supraphysiological doses. It should be emphasized that, in the absence of coronary artery disease, thyroxine has no adverse effects when given at doses which return TSH levels to normal.

Treatment of myxoedema coma consists of thyroid hormone replacement, treatment of any precipitating factor, and supportive therapy. If an intravenous preparation is available, 500 µg thyroxine is given as a single intravenous bolus; the same dose can be given by a nasogastric tube but absorption may be slow. Thereafter, 50 to 100 µg thyroxine is given daily. An alternative is to use triiodothyronine, which has the theoretical benefit of not needing conversion for its activity, but the potential disadvantage of excessive doses causing cardiac arrhythmias. The usual dose is 10 µg triiodothyronine every 4 to 6 h, and it can be given intravenously or by nasogastric tube. Some centres combine thyroxine (200 µg) with triiodothyronine (25 µg) as a single bolus. Patients usually require ventilation initially but external warming should be avoided as it may provoke cardiac failure through peripheral vasodilation and increased oxygen consumption. Instead, space blankets are used. Intravenous infusion of hypertonic saline or glucose may be required to correct metabolic problems. Parenteral hydrocortisone is given at doses of 50 mg every 6 h to treat the reversible decline in adrenal reserve which occurs in marked hypothyroidism. If infection is suspected as a precipitant, broad spectrum antibiotics should be used early.

Prognosis

Thyroxine treatment is usually life-long and, properly taken, restores normal health and lifespan. Occasional patients may discontinue thyroxine and remain euthyroid. Errors in initial diagnosis account for some of these; in others, a spontaneous decline in TSH receptor blocking antibody levels may be responsible. There is no easy means of ascertaining whether a patient continues to need thyroxine, short of stopping it and measuring the TSH 6 weeks later. As remission is uncommon, and of uncertain duration, few endocrinologists at present attempt withdrawal.

Special problems in pregnant women

Untreated hypothyroidism impairs fertility and increases the risk of miscarriage. Children born to such mothers have varying degrees of intellectual impairment. It is therefore essential that thyroxine replacement is monitored closely in women with hypothyroidism who intend to become or who are pregnant. Ideally the TSH and free T4 should be checked prior to conception, once pregnancy is confirmed, and at the beginning of the second and third trimesters. The requirement for thyroxine can increase by 50 to 100 per cent during pregnancy but reverts to normal after delivery. There are no implications for breast feeding.

Areas of uncertainty or needing further research

There has been a recent revival of the concept that thyroid hormone replacement should consist of both thyroxine and triiodothyronine, based on the observation that deiodinase activity varies between tissues, suggesting that in some organs the level of the active thyroid hormone, T3, is insufficient when only thyroxine is given. The short half-life of triiodothyronine makes it alone unsuitable for replacement. Improvements in mental function in a trial of 50 µg of the daily dose of thyroxine with 12.5 µg triiodothyronine have been modest and short-term, and further work is needed. Because hypothyroidism is frequent, routine screening of certain groups or even the entire population has been advocated (Table 6) but the cost–benefit of setting up new screening programmes is unclear. If widely adopted, screening will turn up many individuals with subclinical hypothyroidism, for whom the benefits of early treatment with thyroxine have not yet been fully established.

Thyrotoxicosis

Introduction

Thyrotoxicosis is defined as the state produced by excessive thyroid hormone. Hyperthyroidism exists when thyrotoxicosis is caused by thyroid overactivity but there are several types of thyrotoxicosis which are not due to hyperthyroidism, the most obvious being administration of excessive thyroxine.

Aetiology

The causes of thyrotoxicosis are shown in Table 7. Graves' disease is responsible for 60 to 80 per cent of cases and nodular thyroid disease (toxic multinodular goitre and toxic adenoma) accounts for most of the rest. Destructive thyrotoxicosis is dealt with in the next section.

Epidemiology

The prevalence of thyrotoxicosis is Caucasians is 2 to 3 per cent in women and 0.2 to 0.3 per cent in men. The peak age of onset for Graves' disease is between 20 to 50, whereas toxic multinodular goitre occurs more often in later life.

Pathogenesis

Graves' disease is caused by TSH receptor stimulating antibodies, clearly demonstrated by the occurrence of transient, neonatal thyrotoxicosis in babies born to mothers with Graves' disease whose antibody levels are high enough for transplacental transfer to affect the fetus. As with autoimmune hypothyroidism, genetic factors, including HLA-DR and CTLA-4 gene polymorphisms, are associated with the disease; the concordance rate in monozygotic twins is about 20 per cent and much less in dizygotic twins. A high iodine intake, smoking, and stress have all been identified as environmental factors, but in many patients the genetic and environmental triggers remain elusive. Smoking is a major risk factor for the development of thyroid-associated ophthalmopathy. These eye signs are due primarily to swelling of the extraocular muscles, the result of fibroblast activation by cytokines released by infiltrating T cells and macrophages, leading to glycosaminoglycan accumulation, oedema, and fibrosis. The close correlation between ophthalmopathy and thyroid disease is best explained by an unidentified, shared orbital and thyroid autoantigen.

Toxic multinodular goitre evolves from a non-toxic sporadic goitre (see above) and is particularly likely when iodine intake increases, either gradually as a result of changes in the diet, or acutely when iodine-containing agents (amiodarone, some contrast media) are given. More than 50 per cent of toxic adenomas are due to a somatic activating mutation in the genes encoding the TSH receptor or the associated Gsa protein, and a similar but unknown mechanism leading to constitutive activation of a clone of thyroid cells must underlie the remainder.

Clinical features

The typical features of thyrotoxicosis from any cause are shown in Table 8, but their presence and severity depend on the duration of disease and the age of the patient. Occasionally there are paradoxial manifestations, such as the weight gain which can occur in up to 10 per cent of patients when the increase in appetite exceeds the effects of increased metabolism, and apathetic or masked thyrotoxicosis in the elderly which mimics depression. The most dramatic but rare presentation

is thyrotoxic crisis or storm, with a mortality of 20 to 30 per cent even with treatment. Patients typically are previously undiagnosed or partially treated, and have an acute exacerbation of thyrotoxicosis precipitated by acute illness (infection, stroke, diabetic ketoacidosis) or trauma, especially directly to the thyroid (surgery or radioiodine). Exact diagnostic criteria for thyrotoxic crisis are not agreed and its frequency is sometimes exaggerated. There is marked fever (>38.5°C), delirium or coma, seizures, vomiting, diarrhoea, and jaundice, death being caused by arrhythmias, heart failure, or hyperthermia.

The differential diagnosis of thyrotoxicosis includes any cause of weight loss, anxiety, and phaeochromocytoma, but simple biochemical testing can readily distinguish thyrotoxicosis from these conditions. Once the diagnosis of thyrotoxicosis is made, it is essential to determine the cause (Table 7), as this determines treatment. Graves' disease is usually clinically distinctive; there is a small to moderate, diffuse, firm goitre and around a half of these patients have signs of thyroid-associated ophthalmopathy (Table 9, Fig. 6). There may be evidence of another autoimmune disorder, in the patient or her family, with the same associations as autoimmune hypothyroidism described above. Less than 5 per cent of patients have pretibial myxoedema, which is better called thyroid dermopathy as it can occur anywhere, especially after trauma (Fig. 7 and Plate 1). These patients almost always have moderate to severe ophthalmopathy and 10 to 20 per cent have clubbing (thyroid acropachy). Thyroid dermopathy most commonly occurs as non-pitting plaques with a pink or purple colour but no inflammatory signs. Nodular and generalized forms, the latter mimicking elephantiasis, also occur. Hyperplasia of lymphoid tissue, including splenomegaly and thymic enlargement, is sometimes found in Graves' disease.

Fig. 6 Thyroid-associated ophthalmopathy (a) upper lid retraction, periorbital oedema, and scleral injection; (b) chemosis (conjunctival oedema) and proptosis.

Fig. 7 Thyroid dermopathy (pretibial myxoedema) affecting the lateral aspect of the shin and the dorsum of the foot; the patient also had thyroid acropachy. (See also Plate 1.)

The absence of these features of Graves' disease and the presence of a multinodular goitre strongly suggest toxic multinodular goitre, although nodular thyroid disease is so common that occasional patients with Graves' disease may cause confusion when their thyrotoxicosis arises in a pre-existing multinodular gland. In toxic adenoma, the solitary thyroid nodule is usually readily palpable. Other, rare causes of thyrotoxicosis can usually be easily identified from the history and biochemical investigations.

Pathology

In Graves' disease, there is thyroid hypertrophy and hyperplasia. The follicles show considerable folding, contain little colloid, and are composed of tall columnar cells. Gland vascularity increases. There is a focal and diffuse lymphocytic infiltrate and lymphoid hyperplasia may occur in the lymph nodes, spleen, and thymus. These changes are all reversed by antithyroid drugs. Toxic multinodular goitre comprises a mixture of areas of follicular hyperplasia and nodules filled with colloid. There is a variable degree of fibrosis, haemorrhage, and calcification. Toxic adenomas are encapsulated and cellular, sometimes with little evidence of follicle formation, and occasionally containing unusual cell forms suggesting malignant change. However, capsular invasion is absent and this is the cardinal feature which distinguishes a follicular adenoma from carcinoma.

Laboratory diagnosis

Measuring the serum TSH is the simplest way to exclude primary thyrotoxicosis. A normal or slightly raised TSH level can rarely be associated with hyperthyroidism in the case of a TSH-secreting pituitary adenoma. A low TSH level is not always the result of thyrotoxicosis (Table 2). Therefore the diagnosis of thyrotoxicosis must be confirmed by measuring thyroid hormone levels. Free hormone assays are preferable to those for total hormone, to eliminate binding protein effects (Table 1). Measuring only free T4 alone is adequate in most cases of thyrotoxicosis, which can be confirmed by the presence of a suppressed TSH and elevated free T4 level. However, in up to 5 per cent of patients, only free T3 levels are elevated (T3 toxicosis), especially during the earliest phase of the disorder. Therefore, if both free T3 and T4 are not measured routinely by a laboratory, it is essential to request free T3 analysis in any sample showing a suppressed TSH but normal free T4 level. Rarely, the free T4 is elevated but the free T3 is normal. This arises when Graves' disease or nodular thyroid disease is precipitated by the administration of excess iodine (the Jod–Basedow phenomenon).

Although it is possible to measure TSH receptor stimulating antibodies and thus prove the existence of Graves' disease in a thyrotoxic patient, these assays are cumbersome or expensive, and at present therefore are not widely used. Almost as much information can be gained by measuring thyroid peroxidase antibodies which are present in around 75 per cent of patients with Graves' disease. In cases of diagnostic uncertainty, a thyroid scintiscan will demonstrate a diffuse goitre with high isotope intake in Graves' disease and reveal nodular thyroid disease as well as ectopic thyroid tissue in the extremely rare struma ovarii. In destructive and factitious thyrotoxicosis, the thyroid scan shows virtually no isotope uptake and the diagnosis of factitious thyrotoxicosis can be confirmed by measuring serum thyroglobulin levels, which are suppressed in contrast to the raised levels in all other causes of thyrotoxicosis. When a TSH-secreting pituitary adenoma is suggested biochemically, the diagnosis is made by demonstrating both an elevated level of the a-subunit common to glycoprotein hormones including TSH and a pituitary tumour on CT, or preferably MR, imaging. Prolonged thyrotoxicosis can cause a number of non-specific biochemical abnormalities, especially abnormal liver function tests, hypercalcuria, and elevated serum levels of ferritin. Less commonly, serum calcium and phosphate may be raised, glucose intolerance or diabetes may occur, and rarely there may be a microcytic anaemia or thrombocytopenia.

Treatment

Definitive diagnosis is the most important determinant of treatment selection for thyrotoxicosis. In particular, antithyroid drugs only achieve a cure in Graves' disease. When due to a subacute or silent thyroiditis, discussed below, spontaneous resolution of thyrotoxicosis is expected and symptomatic treatment with b-blockers such as propranolol, 20 to 80 mg three times daily, is indicated. Although b-blockers will rapidly alleviate symptoms in all types of hyperthyroidism, definitive treatment is

also necessary, and when euthyroidism is restored, b-blockers can be gradually withdrawn.

There are three types of treatment for Graves' disease: antithyroid drugs, radioiodine (^{131}I), and surgery. Local policy and patient age dictate the order of their use. For young or middle aged adults, antithyroid drugs are generally used initially in Europe and Japan, whereas radioiodine is preferred in North America. Surgery is particularly useful in patients with a large goitre, but is less frequently used in North America than elsewhere. The local availability of an experienced surgeon is crucial. There is more international agreement over the preferential use of radioiodine for a recurrence after antithyroid drugs and as first line treatment in the elderly with Graves' disease.

The main antithyroid drugs used in Europe are carbimazole and its active metabolite methimazole, whereas propylthiouracil is preferred in North America. There is little to choose between them in normal practice, as all exert their principal action by inhibiting iodide oxidation and organification by thyroid peroxidase. Propylthiouracil additionally inhibits the activity of type 1 deiodinase, reducing T3 formation in many tissues, but this activity is only of clinical importance in very severe hyperthyroidism, and more frequent dosing is necessary with this drug.

Two regimens are used to avoid antithyroid drug-induced hypothyroidism and achieve the best chance of remission, which occurs in 40 to 60 per cent of patients and is inversely proportional to dietary iodine intake. The first method is to titrate the dose of antithyroid drug, giving carbimazole (or methimazole) 20 mg two or three times daily, and then lowering the dose every 3 to 4 weeks or so, based on free T4 measurements, until a maintenance dose of 5 to 10 mg once daily is achieved. Equivalent starting and maintenance doses of propylthiouracil are 100 to 200 mg three times daily and 50 mg once or twice daily. Maximum remission rates occur after 18 to 24 months of treatment.

The second regimen is to start with the same dose of antithyroid drug but then to add thyroxine 100 µg daily after 3 to 4 weeks when free T4 levels are usually becoming normal, rather than lowering the dose of drug. Thereafter the patient is maintained on 40 mg carbimazole or methimazole once daily (alternatively, 100 to 150 mg propylthiouracil three times daily) and thyroxine, the latter being adjusted if necessary 4 weeks after starting to achieve normal free T4 levels. The block–replace regimen achieves the same remission rate as the titration regimen within 6 months; continuation beyond this time is not necessary but can be used if a patient wishes to ensure euthyroidism for a particular period of time. Patients with the biggest goitres almost always relapse after antithyroid drug treatment, but unfortunately there are no reliable predictors of which other patients will relapse, and therefore it is usual practice to follow patients closely (for example every 3 months) in the first year after stopping treatment. Thereafter, an annual check of thyroid function is warranted as recurrence occurs in 10 to 20 per cent 1 to 5 years after treatment, and autoimmune hypothyroidism may supervene in around 15 per cent.

The side-effects of antithyroid drugs are shown in Table 10; most occur in the first 3 months of treatment and there is a moderate dose dependency. Substituting propylthiouracil for carbimazole or vice versa usually reverses the common side-effects but further antithyroid drugs should be avoided if bone marrow disturbance develops. Lower doses of antithyroid drugs can be used in areas of low iodine intake. Lithium and potassium perchlorate have antithyroid actions and are alternatives when antithyroid drugs are not tolerated but these drugs are difficult to use, their side-effects are serious and they are given as a last resort. Anticoagulation with warfarin should be considered in all patients with atrial fibrillation; only 50 per cent of patients revert to sinus rhythm when euthyroidism is restored. In the remainder, attempts at cardioversion should be made, ideally when hyperthyroidism has been definitively treated with radioiodine. Digoxin is useful to control atrial fibrillation acutely but higher doses than normal are needed in the thyrotoxic state.

There are several dosage methods for radioiodine administration, which aim to achieve maximum cure rates with the minimum of subsequent hypothyroidism. Accurate dosimetry based on uptake tests has now largely fallen out of favour, as the results have been little or no better than more empirical methods of dose calculation. A simple formula is to give 200 MBq ^{131}I for a small goitre, 400 MBq for a large goitre, and 600 MBq for Graves' disease complicated by heart failure, but local policies vary, not least because less ^{131}I is needed when iodine intake is low. Around 5 to 10 per cent of patients treated this way require a second dose of ^{131}I, while hypothyroidism rates are 10 to 20 per cent after one year and 5 to 10 per cent annually thereafter. An alternative approach, based on the premise that hyperthyroidism is much more serious than predictable hypothyroidism, has been to attempt deliberate ablation of the thyroid with a fixed dose of 600 MBq ^{131}I. Even with this dose, some patients require a second treatment. Close follow-up is needed in the first year after treatment, and an annual test of thyroid function thereafter is recommended. Transient cytoplasmic, rather than nuclear, damage may cause hypothyroidism in the first 2 to 3 months after ^{131}I treatment, which then resolves. It is usual to delay a second dose of ^{131}I for at least 4 to 6 months after the first, as hyperthyroidism is controlled only slowly by radiation-induced nuclear damage. Antithyroid drugs or b-blockers are useful in the interim.

Radioiodine is contraindicated in pregnancy and breast feeding. There are no teratogenic risks if men or women attempt conception 4 months or more after treatment. Overall mortality rates from cancer are not increased by radioiodine, although there is a theoretical risk of an increase in the frequency and aggressiveness of thyroid cancer in children and adolescents, which makes many endocrinologists reluctant to use ^{131}I in this group, unless other treatments fail or are rejected. Another concern is the precipitation of thyrotoxic crisis by ^{131}I, but in practice this must be rare. To minimize the risk, antithyroid drugs can be given for up to 4 or more weeks prior to radioiodine, particularly in the elderly who are at special risk. Thyroid-associated ophthalmopathy may appear or worsen after radioiodine, especially if the patient smokes. A 3-month tapering course of prednisolone, starting with 40 mg daily at the time of ^{131}I administration, will prevent such worsening but an extended course of antithyroid drugs, with scrupulous maintenance of euthyroidism, may well be preferable until the orbital disease becomes inactive.

Surgery for Graves' disease consists of subtotal or near total thyroidectomy, and in the best centres achieves cure in more than 98 per cent of patients but with a hypothyroidism rate similar to radioiodine. Lower rates of hypothyroidism are inevitably associated with a higher recurrence rate. Patient preference is the main determinant of when surgical treatment is used to treat relapses after antithyroid drugs. Euthyroidism must be achieved with a further course of these drugs prior to surgery to avoid thyrotoxic crisis. Stable iodine (e.g. Lugol's iodine three drops three times daily) is often also given for 7 to 10 days prior to surgery, to block hormone synthesis acutely. Specific complications of surgery include haemorrhage leading to laryngeal oedema, damage to the recurrent laryngeal nerves, and hypoparathyroidism. These problems occur in less than 1 per cent of cases in experienced hands and the last two are often transient.

The management of thyroid-associated ophthalmopathy is summarized in Table 11. Symptoms and signs are usually mild to moderate, although still capable of creating considerable anxiety and disturbance of social function. Severe ophthalmopathy is fortunately rare (1 to 5 per cent of cases) and requires specialist ophthalmological management. Signs usually stabilize 12 to 18 months after onset, and may improve thereafter in 30 to 50 per cent of patients, although improvement is less likely for marked proptosis or diplopia. Corrective surgery for diplopia or cosmetic problems should only be considered in this stable phase. Thyroid dermopathy is left untreated and may resolve spontaneously. Surgical removal usually worsens the situation and, when troublesome, the best treatment is topical, high potency corticosteroids. Octreotide may also be beneficial.

Toxic multinodular goitre is usually managed by radioiodine treatment. Antithyroid drugs will control the hyperthyroidism but relapse is inevitable when the drugs are stopped. Long-term use of antithyroid drugs may be indicated in the very old or frail, or when incontinence poses an insuperable problem for the safe disposal of excreta after ^{131}I. The therapeutic dose of ^{131}I used for toxic multinodular goitre is generally higher than for Graves' disease (600 to 800 MBq) because there is uneven uptake of the isotope and usually a large goitre. Surgery is sometimes used as an alternative in patients with a retrosternal goitre or if there is any suspicion of a malignancy. Toxic adenoma is also usually treated with ^{131}I and the rate of subsequent hypothyroidism is low because the function of the normal thyroid issue is suppressed at the time the patient is hyperthyroid and therefore receives little irradiation. When there is a large (>5 cm) nodule or in young patients (<20 years) surgical excision is preferable and subsequent hypothyroidism is uncommon. Treatment of rare forms of primary hyperthyroidism is by surgical removal of the source of thyroid hormone or radioiodine. TSH-secreting pituitary adenomas causing secondary hyperthyroidism are usually treated by transphenoidal surgery, with radiotherapy for any residual tumour. Octreotide can also be used to lower TSH secretion.

Thyrotoxic crisis is a medical emergency whose management consists of antithyroid treatment, identification and treatment of any underlying precipitant, and supportive measures. Propylthiouracil is given as a loading dose of 600 mg and then 250 mg four times daily, either orally, by stomach tube, or per rectum. Carbimazole or methimazole are less effective alternatives, as they do not reduce T4 deiodination. Stable iodine blocks thyroid hormone synthesis and release but is safe to give only when iodine organification is arrested by propylthiouracil. Lugol's iodine, 5 drops four times daily, is given orally or by stomach tube 1 h after the first dose of propyl-thiouracil; ipodate 500 mg twice a day is an alternative. Control of the heart rate is central to managing the heart failure which frequently occurs, and in all but the most severe cases propranolol should be given (40 mg orally or 2 mg intravenously every 4 h), with careful monitoring of ventricular function. Diuretics and digoxin are used as needed. Plasmapheresis, dialysis, or cholestyramine (to interrupt the enterohepatic circulation of thyroid hormones) may be useful in extreme cases. Supportive measures include dexamethasone 2 mg four times daily, which also inhibits T4 to T3 conversion, oxygen, cooling, and intravenous fluids.

Prognosis

Although spontaneous remission occurs in Graves' disease, its exact frequency is unknown and is unlikely to be more than 10 per cent, with no guarantee of

persistence. Remission does not occur in other types of hyperthyroidism. Mortality rates in untreated hyperthyroidism are also uncertain but are probably around 30 per cent. Even after successful treatment, there is a three-fold increased risk of death from osteoporotic fracture and a 1.3-fold increased risk of death from cardiovascular disease and stroke. It is important that the patient with Graves' disease understands that the course of ophthalmopathy is independent of the thyroid disorder; eye signs appear one or more years before or after the onset of hyperthyroidism in a quarter of patients and progression of the orbital disease frequently occurs despite restoration of euthyroidism.

Special problems in pregnant women

Graves' disease during pregnancy is often treated with propylthiouracil, as carbimazole and methimazole have been associated with fetal aplasia cutis, but some dispute the significance of this association. The block–replace regimen is contraindicated in pregnancy, as preferential placental transfer of antithyroid drug will cause fetal hypothyroidism. Instead, the dose of antithyroid drug should be titrated to the lowest dose which results in maternal free T4 levels in the upper part of the reference range. TSH receptor stimulating antibodies decline during pregnancy and it is usually possible to stop treatment at the beginning of the third trimester. Subtotal thyroidectomy can be performed in the second trimester for women intolerant of antithyroid drugs.

Transplacental passage of TSH receptor antibodies causes fetal and neonatal thyrotoxicosis in 1 to 5 per cent of mothers with Graves' disease and can be predicted by demonstrating a high level of these antibodies in the maternal circulation at the beginning of the third trimester. Poor intrauterine growth and a high fetal heart rate also suggest this diagnosis. Fetal thyrotoxicosis is treated by giving the mother antithyroid drugs and the neonate requires treatment for 1 to 3 months after delivery. Failure to treat untrauterine and neonatal thyrotoxicosis causes low birth weight, premature closure of the sutures, and intellectual impairment. Breast feeding is safe with low doses of antithyroid drugs, but when high doses are needed (e.g. 20 mg or more carbimazole daily), thyroid function should be checked every 1 to 2 weeks in the baby. Patients with Graves' disease who have entered remission prior to or during pregnancy have an increased risk of relapse around 3 to 6 months after delivery, and should be offered thyroid function testing at this time.

Areas of uncertainty or needing further research

The pathogenesis of thyroid-associated ophthalmopathy is poorly understood, and remains an obstacle to developing better treatments. Outcome after antithyroid drug treatment in Graves' disease cannot yet be predicted but improved assays for TSH receptor antibodies may permit better assessment in the near future. Antithyroid drugs modulate the autoimmune response favourably in those patients whose Graves' disease remits, indicating the potential for more specific immunotherapy aimed at the cause of the disease, that would be preferential to present treatments which merely block or destroy the thyroid.

The evolution of hyperthyroidism is gradual and patients with multinodular goitre in particular are now recognized at the stage of subclinical hyperthyroidism, that is with a low or suppressed TSH but normal free T3 and T4 levels. Their optimum management is uncertain. There is a two to three-fold increased risk of atrial fibrillation over 10 years in subclinical thyrotoxicosis, as well as deleterious effects on bone mineral density, but no clinical trials have been performed to show a clear benefit from early intervention. Many endocrinologists simply follow such patients carefully, electing to treat when overt hyperthyroidism is shown by an abnormal free T3 level (T3 usually increases before T4). However, in the elderly with known cardiac disease, there is an increasing tendency to use radioiodine for sustained subclinical hyperthyroidism.

Destructive thyroiditis

Acute thyroiditis is rare and usually caused by bacterial infection of the thyroid via a pyriform sinus connecting the gland with the oropharynx. There is severe thyroid pain with fever and malaise, but thyroid function is rarely disturbed. Diagnosis is made by fine needle aspiration biopsy, with culture of the specimen, and treatment consists of antibiotics, surgical drainage of any abscess, and excision of the sinus which is identified by barium swallow.

Subacute (or de Quervain's) thyroiditis is due to thyroid infection by any of a number of viruses, especially mumps, Coxsackie, influenza, adenoviruses, and echoviruses. The most prominent symptom is pain in the thyroid, often radiating to the ears. A small, tender goitre can be palpated which is usually diffuse, but there can be asymmetrical involvement. Systemic upset with fever is variable but sometimes profound, and symptoms of a prodromal viral infection several weeks earlier may be recalled. There is a granulomatous thyroid inflammation with follicular destruction and the release of thyroid hormones often results in a transient thyrotoxicosis, lasting 1 to 4 weeks. Continuing thyroid destruction then leads to a phase of hypothyroidism once stored hormone is depleted. This lasts 4 to 12 weeks before euthyroidism is restored, but relapses occur in 10 to 20 per cent of cases. Sometimes only one phase of thyroid disturbance is seen. Confirmation of the clinical diagnosis is made by finding an elevated erythrocyte sedimentation rate (ESR) and low or absent radioiodine uptake by the thyroid. Thyroid function requires continuous monitoring as the disease evolves. Mild cases may resolve spontaneously with aspirin as symptomatic treatment, but most patients benefit from prednisolone 40 to 60 mg daily, which rapidly alleviates the pain. The dose is tapered over 6 to 8 weeks, depending largely on symptoms. Propranolol may be useful for thyrotoxic symptoms, and temporary thyroxine replacement is sometimes needed during the hypothyroid phase.

Silent thyroiditis is an autoimmune disorder in which there is a transient but painless thyroid destruction, giving rise to the same kind of thyroid function disturbances as subacute thyroiditis. As well as the absence of thyroid pain, there is no sign of a systemic inflammatory response (including a normal ESR) and the two conditions are therefore readily distinguished. The commonest setting for silent thyroiditis is in the postpartum period in a women with positive thyroid peroxidase antibodies and a mild underlying autoimmune thyroiditis, exacerbated for unknown reasons at this time. Such postpartum thyroiditis is common, being detectable in up to 5 per cent of women 3 to 6 months after delivery when repeated biochemical testing is done, although in many of these women the changes in thyroid function are mild and asymptomatic. Postpartum thyroiditis is three times more common in type 1 diabetes mellitus. Thyroid uptake tests are useful in the postpartum period to distinguish thyrotoxicosis due to postpartum thyroiditis from Graves' disease. ^{99m}Tc is used in preference to ^{131}I and only requires cessation of breast feeding for a day. Treatment is with propranolol for thyrotoxic symptoms and thyroxine for hypothyroidism. Thyroxine should be withdrawn 1 year after delivery and thyroid function tested 6 weeks later, as 90 per cent of women recover normal thyroid function. However, annual follow-up is needed as around 20 per cent of these women have permanent hypothyroidism 5 years later. The condition usually recurs in subsequent pregnancies.

Amiodarone inhibits T4 deiodination, and in all amiodarone-treated patients free T4 levels are in the upper half of the reference range or mildly elevated. Several months to years after starting amiodarone, however, effects on the thyroid may become manifest. In patients with mild thyroid dysfunction, especially autoimmune thyroiditis and positive thyroid peroxidase antibodies, the excessive iodine released from the drug causes hypothyroidism. This is treated as usual with thyroxine. Paradoxically, the high level of iodine causes hyperthyroidism in other subjects who are predisposed to this because of an underlying multinodular goitre or incipient Graves' disease (Jod–Basedow phenomenon). This is called type 1 amiodarone-induced thyrotoxicosis (AIT); type 2 AIT is due to thyroid destruction via drug-induced lysosomal activation. Colour-flow Doppler thyroid scanning shows an increase in vascularity in type 1 but not type 2 AIT, while serum IL-6 levels may be elevated in type 2 but not type 1 AIT. Mixed forms sometimes make an exact diagnosis impossible.

Treatment of AIT can be difficult and biochemical changes are often out of proportion to the symptoms. Amiodarone should be stopped if possible but often this cannot be done and in any case the drug has a very long half-life. Antithyroid drugs are often ineffective in type 1 AIT and potassium perchlorate may need to be added, 200 mg four or five times daily. There is a high frequency of agranulocytosis (up to 1 per cent) with this drug. Prednisolone is also used at doses of 40 to 60 mg daily, and is particularly helpful in type 2 AIT. Thyroidectomy is another alternative in severe cases.

Thyroid hormone resistance syndrome

Mutations in one allele of the b thyroid hormone receptor gene (Fig. 4) cause thyroid hormone resistance (homozygous mutation is lethal). The mutations affect the hormone binding domain and the mutant receptor inhibits the activity of normally encoded receptors, so called dominant negative inhibition, resulting in an autosomal dominant pattern of inheritance. The condition is usually discovered during screening for a goitre but children may sometimes present with short stature, hyperactivity, or mild learning difficulties. Thyrotoxic features in some patients were originally ascribed to selective pituitary resistance to thyroid hormone, leading to increased thyroid hormone secretion and therefore thyrotoxicosis in the peripheral tissues. However, the same receptor mutations occur in generalized and pituitary resistance syndromes, and although differential tissue expression of receptor subtypes presumably underlies the occasional expression of thyrotoxic signs and symptoms, the exact molecular basis is unknown.

The diagnosis is suggested by the presence of a normal or elevated TSH with elevated free T3 and T4 levels. Non-specific biochemical changes of thyrotoxicosis, such as elevated ferritin, sex hormone binding globulin, and liver enzymes, are absent. The main differential diagnosis is a TSH-secreting adenoma. Thyroid hormone resistance can be confirmed by direct mutational analysis. Treatment is usually not required as reducing thyroid hormone levels to normal causes hypothyroidism. If

thyrotoxic symptoms do occur, treatment is with b-blockers or thyroid hormone analogues (e.g. triac) aimed at lowering TSH secretion.

Further reading

Abramowicz MJ, Vassart G, Refetoff S (1997). Probing the cause of thyroid dysgenesis. *Thyroid* **7**, 325–6.

Amino N, *et al*. (1999). Screening for postpartum thyroiditis. *Journal of Clinical Endocrinology and Metabolism* **84**, 1813–21.

Arbelle JE, Porath A (1999). Practice guidelines for the detection and management of thyroid dysfunction. A comparative review of the recommendations. *Clinical Endocrinology* **51**, 11–18.

Bartalena L, Pinchera A, Marcocci C (2000). Management of Graves' opthalmopathy. *Endocrine Reviews*, **21**, 168–99.

Beck-Peccoz P, *et al*. (1996). Thyrotropin-secreting pituitary tumors. *Endocrine Reviews* **17**, 610–38.

Bodenner DL, Lash RW (1998). Thyroid disease mediated by molecular defects in cell surface and nuclear receptors. *American Journal of Medicine* **105**, 524–38.

Braverman LE, Utiger RD, eds (1996). *Werner and Ingbar's The Thyroid*, 7th edn. Lippincott-Raven, Philadelphia.

Brix TH, Kyvik KO, Hegedüs L (1998). What is the evidence of genetic factors in the etiology of Graves' disease? A brief review. *Thyroid* **8**, 727–32.

Burrow GN, *et al*. (1994). Maternal and fetal thyroid function. *New England Journal of Medicine* **20**, 1072–8.

Comtois R, Faucher L, Laflèche L (1995). Outcome of hypothyroidism caused by Hashimoto's thyroiditis. *Archives of Internal Medicine* **155**, 1404–8.

Davies TF, ed (1997). Newer aspects of clinical Graves' disease. *Baillière's Clinical Endocrinology and Metabolism* **11**, 431–601.

DeGroot LJ, *et al*. (1999). *Thyroid disease manager.* http://www.thyroidmanager.org/.

Delange F (1994). The disorders induced by iodine deficiency. *Thyroid* **4**, 107–28.

Dumont JE, *et al*. (1995). Large goitre as a maladaptation to iodine deficiency. *Clinical Endocrinology* **43**, 1–10.

Ferretti E, *et al*. (1999). Evaluation of the adequacy of levothyroxine replacement therapy in patients with central hypothyroidism. *Journal of Clinical Endocrinology and Metabolism* **84**, 924–9.

Franklyn JA, *et al*. (1998). Mortality after the treatment of hyperthyroidism with radioactive iodine. *New England Journal of Medicine* **338**, 712–8.

Gharib H and Mazzaferri EL (1998). Thyroxine suppressive therapy in patients with nodular thyroid disease. *Annals of Internal Medicine* **128**, 386–94.

Glinoer D (1997). The regulation of thyroid function in pregnancy: Pathways of endocrine adaption from physiology to pathology. *Endocrine Reviews* **18**, 404–33.

Göthe S, *et al*. (1999). Mice devoid of all known thyroid hormone receptors are viable but exhibit disorders of the pituitary-thyroid axis, growth and bone maturation. *Genes and Development* **15**, 1329–41.

Haddow JE, *et al*. (1999). Maternal thyroid deficiency during pregnancy and subsequent neurophysiological development of the child. *New England Journal of Medicine* **341**, 549–55.

Hermus AR, Huysmans DA (1998). Treatment of benign nodular thyroid disease. *New England Journal of Medicine* **338**, 1438–47.

Houghton DJ, Gray HW, MacKenzie K (1998). The tender neck: thyroiditis or thyroid abscess? *Clinical Endocrinology* **48**, 521–4.

Jarlev AE, *et al*. (1995). Is calculation of the dose in radioiodine therapy of hyperthyroidism worthwhile? *Clinical Endocrinology* **43**, 325–9.

Ko GTC, *et al*. (1996) Thyrotoxic periodic paralysis in a Chinese population. *Quarterly Journal of Medicine* **89**, 461–8.

Koutras DA (1999). Subclinical hyperthyroidism. *Thyroid* **9**, 311–5.

Laurberg P, *et al*. (1998) Guidelines for TSH receptor antibody measurements in pregnancy: Results of an evidence-based symposium organized by the European Thyroid Association. *European Journal of Endocrinology* **139**, 584–6.

Le Moli R, *et al*. (1999). Determinants of longterm outcome of radioiodine therapy of sporadic non-toxic goitre. *Clinical Endocrinology* **50**, 783–9.

Mandel SJ, Brent GA, Larsen PR (1993). Levothyroxine therapy in patients with thyroid disease. *Annals of Internal Medicine* **119**, 492–502.

Newman CM, *et al*. (1998). Amiodarone and the thyroid: A practical guide to the management of thyroid dysfunction induced by amiodarone therapy. *Heart* **79**, 121–7.

Nicoloff JT, LoPresti JS (1993). Myxedema coma: A form of decompensated hypothyroidism. *Endocrinology and Metabolism Clinics of North America* **22**, 279–90.

Radioiodine Audit Subcommittee of the Royal College of Physicians Committee on Diabetes and Endocrinology and The Research Unit of the Royal College of Physicians (1995). *Guidelines: The use of radioiodine in the management of hyperthyroidism*, 26 pp. Royal College of Physicians of London.

Rapoport B, *et al*. (1998). The thyrotropin (TSH)-releasing hormone receptor: Interaction with TSH and autoantibodies. *Endocrine Reviews* **19**, 673–716.

Rivkees SA, *et al*. (1998). The management of Graves' disease in children, with special emphasis on radioiodine treatment. *Journal of Clinical Endocrinology and Metabolism* **83**, 3767–76.

Schwartz AE, *et al*. (1998). Thyroid surgery—the choice. *Journal of Clinical Endocrinology and Metabolism* **83**, 1097–105.

Singer PA, *et al*. (1995). Treatment guidelines for patients with hyperthyroidism and hypothyroidism. *Journal of the American Medical Association* **273**, 806–12.

Toft AD (1999). Thyroid hormone replacement—one hormone or two? *New England Journal of Medicine* **340**, 469–70.

Vanderpump MPJ, *et al*. (1995). The incidence of thyroid disorders in the community: A twenty-year follow-up of the Whickham Survey. *Clinical Endocrinology* **43**, 55–68.

Van Sande J, *et al*. (1995). Somatic and germline mutations of the TSH receptor gene in thyroid diseases. *Journal of Clinical Endocrinology and Metabolism* **80**, 2577–85.

Van Wassenaer AG, *et al*. (1997). Effects of thyroxine supplementation on neurologic development in infants born at less than 30 weeks gestation. *New England Journal of Medicine* **336**, 21–6.

Volpé R (1993). The management of subacute (DeQuervain's) thyroiditis. *Thyroid* **3**, 253–5.

Weetman AP (1997). Hypothyroidism: Screening and subclinical disease. *British Medical Journal* **314**, 1175–8.

Weetman AP (2000). Medical progress: Graves' disease. *New England Journal of Medicine*, **343**, 1236–48.

12.5 Thyroid cancer

Anthony P. Weetman

Thyroid cancer is by far the most common endocrine malignancy but it constitutes less than 1 per cent of all cancers. Most thyroid cancers arise in the follicular epithelial cells of the thyroid and may be differentiated or undifferentiated (Table 1). Tumours of the parafollicular C cells (medullary carcinoma) and other malignancies such as lymphoma and sarcoma behave in a different fashion from thyroid follicular epithelial cancers and are dealt with at the end of this chapter.

Primary thyroid follicular epithelial tumours

Aetiology

Excessive stimulation of the thyroid by thyroid-stimulating hormone accounts for the higher proportion of follicular carcinomas compared with papillary carcinomas in iodine-deficient areas. The thyroid-stimulating antibodies of Graves' disease do not increase the risk of developing thyroid cancer, but incidental thyroid tumours that arise in this disorder may behave more aggressively because of activation of thyroid-stimulating hormone receptors. Low-dose external beam radiation (10–1500 cGy) to the head and neck increases the risk of papillary thyroid cancer over 10 to 30 years. Higher thyroid radiation doses, including those arising from radio-iodine given for treatment of hyperthyroidism, are not associated with an increased risk of malignancy because thyroid cells are destroyed rather than transformed. However, death from thyroid cancer, which is an unusual outcome, may be slightly increased by radio-iodine treatment, suggesting an effect of radiation on tumour dedifferentiation. In Belarus and Ukraine, the incidence of papillary carcinomas in children and young adults has increased 5 to 10 years after the disastrous release of radio-iodine and other radionuclides from the Chernobyl nuclear reactor. This is ascribed to the potent mutagenic effects of radio-iodine on the growing thyroid gland.

Familial forms of papillary and follicular carcinomas exist but are unusual (less than 5 per cent of cases). There are associations with familial adenomatosis polyposis, including the Gardner's syndrome variant, Cowden's disease (multiple hamartoma syndrome), Peutz–Jehgers syndrome, and ataxia-telangiectasia.

Papillary carcinomas do not arise from hyperplastic nodules or adenomas. In about one-third of these tumours one of five distinct rearrangements of the *ret* proto-oncogene, a member of the receptor tyrosine kinase family, occurs. The resulting chimaeric oncogenes are termed *ret/PTC* (for papillary thyroid carcinoma). *ret/PTC3* is particularly linked to radiation. In 10 to 20 per cent of papillary cancers there is oncogenic activation of the *trk* gene.

Follicular carcinomas probably arise, at least in some cases, from follicular adenomas. Activation of the *ras* oncogene occurs in both these tumours (and in some papillary thyroid cancers) and they share cytogenetic abnormalities, especially on chromosome 3. Rarely follicular carcinomas are associated with activating mutations of the thyroid-stimulating hormone receptor and Gsa protein (encoded by the *gsp* gene), similar to those found in toxic adenoma. Anaplastic carcinoma may arise in a papillary or follicular carcinoma and is associated with inactivating mutations of the p53 tumour suppressor gene.

Epidemiology

Papillary microcarcinomas are less than 1 cm in diameter and occur in 4 to 36 per cent of autopsy specimens; they are more frequent in areas of high iodine intake. Clearly most of these do not become malignant. Excluding tumours which are found coincidentally, the annual incidence of thyroid follicular epithelial cancer is around 4 per 100 000. In iodine-sufficient countries, papillary carcinoma accounts for more than 80 per cent of these; follicular carcinoma constitutes about 10 per cent, and anaplastic carcinoma 5 to 10 per cent. Women are two to four times more likely to develop thyroid cancer than men; the peak incidence is between 30 and 50 years of age.

Clinical features

Most patients present with an asymptomatic thyroid nodule: this may be noticed by themselves or relatives—sometimes the nodule is detected during physical examination for another complaint. The difficulty for diagnosis arises because thyroid nodules are frequent and only about 5 per cent of palpable thyroid nodules are malignant. Diffuse or multinodular thyroid enlargement occurs in around 10 per cent of the population, and is four times more common in women than men. Solitary thyroid nodules occur in up to 5 per cent of the population and are usually hyperplastic or colloid nodules. The remaining 5 to 20 per cent of nodules are neoplastic but this figure includes follicular adenomas as well as malignant tumours.

It can be seen that determining which thyroid nodules are malignant poses a great difficulty, which has been exacerbated by the widespread use of ultrasound examination of the neck. Up to 60 per cent of adult thyroids have nodules detectable by high-resolution ultrasound scanning. If these nodules are impalpable and there are no other associated features, such ultrasound findings can be ignored. However, a palpable thyroid nodule should always be properly evaluated. Another problem is determining which, if any, nodules warrant investigation in a multinodular goitre. It seems reasonable to investigate further if there have been any new symptoms (see below) or if one nodule is larger than the others.

There are usually no symptoms or signs to indicate that a solitary thyroid nodule is malignant because most tumours progress slowly and present before disease is advanced. Age and sex are important considerations, since a malignancy is more likely in a solitary nodule when the patient is a child or adolescent, is over 60 years old, or is a man between the ages of 20 and 60 years. Previous exposure to radiation and a family history of thyroid cancer should also arouse suspicion. A carcinoma is more likely if the nodule has grown recently or is hard, irregular, or fixed on palpation. Clinical assessment should include careful examination of the cervical, submental, and supraclavicular lymph nodes. Late-presenting features include hoarseness, dysphagia, or dyspnoea which may indicate local invasion, but these symptoms can occasionally occur with an enlarging benign goitre. Rarely the diagnosis only becomes apparent when metastatic disease is detected in bone or lung.

The relatively indolent presentation of papillary and follicular thyroid carcinoma contrasts with that of anaplastic carcinoma in which a rapidly enlarging and fixed thyroid mass occurs, sometimes with local pain. Extension to the oesophagus, trachea, and/or recurrent laryngeal nerves is frequent and the overlying skin may also be infiltrated.

Pathology

There are several variants of papillary thyroid carcinoma united by their characteristic cytological features. The nuclei are large, clear ('Orphan Annie', after the eyes of the cartoon character), and have longitudinal grooves and invaginations of cytoplasm (Fig. 1). Two-thirds of tumours are unencapsulated and display papillary and follicular structures, the remainder comprise the encapsulated, follicular, tall cell, sclerosing, and clear cell variants.

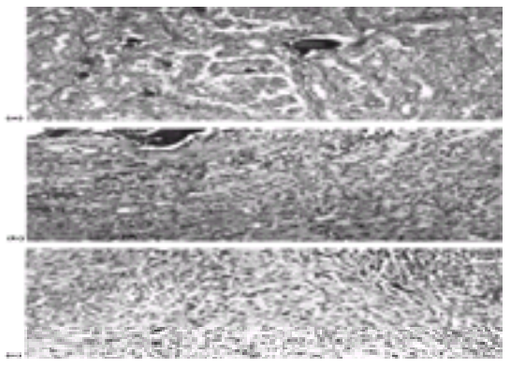

Fig. 1 Histopathological features of thyroid follicular epithelial carcinoma: (a) papillary carcinoma, with psammoma bodies and typical nuclear appearance; (b) metastatic follicular carcinoma, eroding vertebral bone; (c) anaplastic carcinoma showing pleomorphic spindle cells. (All sections, original magnification ×200; photomicrographs by courtesy of Dr K. Suvarna.)

The encapsulated variant has a better prognosis than average and the tall cell variant a worse prognosis. Half of papillary carcinomas contain degenerate calcified papillae, termed psammoma bodies. The tumour is multicentric in up to 80 per cent of cases if the resected thyroid is examined carefully. Metastasis is via the lymphatics and local lymph nodes are infiltrated in 40 to 50 per cent of cases (more in young patients). Distant metastases are found in less than 5 per cent of patients at presentation: the lung is the most common site.

Follicular carcinoma is characterized by follicular differentiation with a solid growth pattern and without the nuclear features of papillary carcinoma. The tumour is encapsulated but there is invasion of the capsule and vessels (Fig. 1). This invasion is the crucial feature which distinguishes follicular carcinoma from follicular adenoma, self-evidently a distinction only possible by histological examination. Minimally and widely invasive subtypes are recognized, the latter having a worse prognosis. When 75 per cent or more of the tumour cells exhibit oxyphilic staining due to mitochondrial accumulation, this is termed a Hürthle cell carcinoma, which probably also has a worse prognosis. Lymph node metastases are unusual, as is multicentricity in the thyroid. Metastasis occurs via the bloodstream, typically to bone and lungs.

When follicular differentiation is poor or absent, the tumour is classified as an insular carcinoma with a poor prognosis. In anaplastic carcinoma there is no capsule, the cells are atypical, including spindle, multinuclear, and squamoid forms, and mitoses are frequent (Fig. 1).

Diagnosis

Thyroid epithelial cancers generally fail to affect thyroid function. However, this should be evaluated in all patients presenting with a thyroid nodule: a low circulating level of thyroid-stimulating hormone strongly suggests an autonomous benign nodule. Anaplastic carcinoma may occasionally cause hypothyroidism, but the most frequent cause of an elevated level of thyroid-stimulating hormone with a hard, nodular thyroid is Hashimoto's thyroiditis. Some Hashimoto glands are so irregular that a malignancy may be suspected. There is no increased or decreased risk of thyroid epithelial carcinoma in Hashimoto's thyroiditis, but thyroid lymphoma almost always occurs in association with autoimmune thyroiditis. Therefore any dominant or atypical area in a Hashimoto goitre requires careful evaluation. Thyroid peroxidase and/or thyroglobulin antibodies occur in about one-quarter of patients with thyroid follicular epithelial carcinoma, coincident with the presence of a lymphocytic infiltrate which, in turn, is associated with a more favourable prognosis. Although the serum thyroglobulin concentration is extremely useful in follow-up, as discussed below, this investigation is useless in diagnosis: levels may not be elevated with some cancers and, even when elevated, cannot be causally distinguished from those which occur in benign adenoma, multinodular goitre, Graves' disease, or destructive thyroiditis.

In the past solitary thyroid nodules were investigated by radionuclide and/or ultrasound imaging but neither technique is able to diagnose malignancy accurately. Radionuclide scanning can be performed with $^{99}Tc^{m}$ pertechnetate or radio-iodine (^{123}I or ^{131}I), with similar information being obtained from either nuclide. Most thyroid cancers fail to take up radionuclide ('cold' nodules), but the more frequent benign lesions such as colloid nodules, cysts, adenomas, and thyroiditis behave similarly. About 20 per cent of nodules have normal or increased radionuclide uptake. Malignancy cannot be excluded with these appearances, however. The only exception is when the nodule is 'hot' and the surrounding thyroid tissue fails to take up radionuclide, indicating the presence of a toxic adenoma which is almost invariably benign. This type of nodule will cause suppression of thyroid-stimulating hormone and will be suspected from routine testing of thyroid function. In summary, radionuclide scanning adds little to diagnosis.

The use of ultrasound is more controversial. Cystic nodules have a lower risk of malignancy than those which are solid, but thyroid cancer cannot be reliably excluded by ultrasound. Predicting the presence of malignancy based on the echo pattern of the tumour, and more recently using colour-flow Doppler, may be successful in up to 80 per cent of cases but this depends on considerable experience. As well as the poor specificity of ultrasound, the technique is so sensitive that many small unsuspected nodules will be uncovered, complicating the evaluation. Ultrasound is useful for accurate measurement of thyroid and nodule size, which can be helpful in monitoring patients and in guiding biopsy, although this procedure is usually performed without imaging.

Fine needle aspiration biopsy is undoubtedly the current technique of choice for investigation of a thyroid nodule. Local anaesthetic is not needed because the procedure causes little discomfort. It is usual to take two to six biopsies to increase the sample yield. Essentially three diagnoses are possible: benign (65–75 per cent of specimens), malignant (5 per cent), and indeterminate (20–30 per cent), but an experienced cytopathologist is needed to obtain reliable results. Papillary carcinoma is readily diagnosed by fine needle aspiration biopsy, and medullary carcinoma and lymphoma can also be detected by use of immunohistochemical staining.

Follicular carcinomas cannot be distinguished cytologically from a follicular adenoma, and these tumours account for the bulk of needle aspiration specimens labelled indeterminate (or suspicious). Open biopsy and histological examination is the only secure diagnostic method in this setting. About 15 per cent of biopsies reported in experienced centres are considered unsuitable for diagnosis. It is relatively simple to repeat the biopsy but a persistently equivocal biopsy should be grounds for considering surgery, since malignant tumours will be found in about half of the cases. A cyst may be aspirated during biopsy. If this fails to reaccumulate and no lesion remains palpable, a malignancy is highly unlikely, but recurrence of a cyst may indicate malignant disease and require surgery for definitive diagnosis. Overall, the sensitivity and specificity of fine needle aspiration biopsy is greater than 90 per cent.

Treatment

Surgical excision

A total or near total thyroidectomy should be performed since papillary carcinomas are often bilateral and removal of thyroid tissue facilitates subsequent ablation by radio-iodine. There is controversy regarding surgery for low-risk unifocal papillary carcinomas (those less than 5 cm in diameter, especially in young patients) but bilateral resection is associated with a lower rate of local recurrence than unilateral total lobectomy. In papillary carcinoma, the ipsilateral central lymph nodes should be dissected, as should all palpable nodes.

After surgery, radio-iodine is usually administered to remove any remaining thyroid tissue, which then allows thyroglobulin or ^{131}I total body scanning to be used in follow-up to detect metastases. This treatment also destroys occult carcinoma and, by scanning after ablation, metastatic disease is revealed. Local policies vary, but in most centres an ^{131}I scan is performed 1 to 2 months after surgery and an ablation dose of 1100 to 3700 MBq ^{131}I is given, depending on the size of the remnant. In 15 to 30 per cent of patients a second treatment dose of ^{131}I is necessary to achieve ablation. Iodine exposure, including iodine-containing contrast media, may prevent accumulation of ^{131}I during treatment.

In patients whose tumour is less than 1.5 cm in diameter, excision alone without radio-iodine ablation is indicated. Whether all other patients require ablation is controversial. Clinical staging scores (see below) may help to identify other low-risk patients who do not require ablation excision.

Radio-iodine therapy

High levels of stimulation by thyroid-stimulating hormone are required to produce maximum uptake of ^{131}I; this is achieved by a period of 30 to 45 days without thyroxine replacement and can thus lead to the development of severe hypothyroid symptoms. The short action of tri-iodothyronine, 20 µg three times daily, as a replacement is therefore preferable in the weeks before scanning and ^{131}I treatment, because only 2 weeks are needed when this is stopped to increase endogenous thyroid-stimulating hormone. Even this short period without thyroid hormone may be troublesome for the patient and recombinant thyroid-stimulating hormone suitable for intramuscular administration is now available and can be given without cessation of thyroid hormone replacement.

Long-term thyroid replacement therapy

The third aspect of treatment is to maintain the patient for life on thyroxine at doses sufficient to suppress levels of thyroid-stimulating hormone to 0.1 mU/litre or less, because thyroid-stimulating hormone is a growth factor for thyroid carcinoma. The optimum level of thyroid-stimulating hormone is unknown, but higher levels can be accepted in those patients known to be disease-free for several years, compared with newly treated patients. In almost all patients, satisfactory suppression of thyroid-stimulating hormone can be achieved without inducing thyrotoxic symptoms. The effective thyroxine dosage is 2.2 to 2.8 µg/kg body weight.

Anaplastic carcinoma is rapidly fatal. The tumour does not take up radio-iodine. Surgery has a limited role in relieving obstructive symptoms and external beam radiotherapy is useful in palliation. The place of chemotherapy (usually doxorubicin combined with other drugs) is unclear.

Follow-up

Lifelong follow-up is necessary for papillary and follicular cancer because recurrence may occur many years after apparent cure. As well as monitoring the concentration of thyroid-stimulating hormone and performing careful neck palpation, serum thyroglobulin should be measured. Detectable levels of thyroglobulin after thyroid ablation indicate persistent or recurrent disease. In many centres, thyroglobulin levels are checked when the patient is not taking thyroxine replacement, as the rise in thyroid-stimulating hormone will stimulate thyroglobulin production and exaggerate any increase. This is particularly useful in high-risk patients: in those at low risk (see next section), it is reasonable to measure thyroglobulin without withdrawing thyroxine.

If thyroglobulin is detectable, the patient should have a total body ^{131}I scan and any recurrent disease can then be treated with a therapeutic dose of 3700 MBq ^{131}I. Many centres also perform a total body scan at 6 to 12 months after initial ablation, but repeated routine scans thereafter have now been superceded by measurement of thyroglobulin. The only exception is in the patient with thyroglobulin antibodies which interfere with many assays for thyroglobulin. If this is the case, repeated scans are the only way to ensure that the patient remains free of disease.

For metastatic disease, usually in the lung, treatment with radio-iodine can be repeated every 4 to 6 months, but there is little benefit above a cumulative dose of 18 500 MBq. Bone metastases may respond to ^{131}I or external beam radiotherapy. The best survival in metastatic thyroid cancer occurs in young patients with small metastases, indicating the overall value of early treatment for this disease.

Prognosis

At least seven scoring systems have been advocated to assess prognosis in papillary and follicular carcinoma. These take into account the age and sex of the patient, tumour characteristics (especially size, extension, and metastases), and completeness of excision. An example of the predictive power of such scoring is shown in Table 2. The risk of death increases with age, especially after 60, while tumour recurrence is commonest in those aged under 20 and over 60. Men have a worse prognosis than women, but the difference is small.

With proper treatment the rate of recurrence of papillary carcinoma is about 15 per cent, and the cause-specific death rate is approximately 5 per cent at 20 years. In other words, 85 per cent of these patients present with features of the group with the best prognosis, for instance achieving a score of less than 6 in the system described in Table 2. In follicular carcinoma, the cause-specific survival rate is 80 per cent at 20 years after treatment and 70 per cent at 30 years. However, in the subgroup with metastases at presentation the 10-year survival is only 20 per cent. The median survival time for anaplastic carcinoma is 4 to 12 months and those with distant metastases at presentation have a median survival time of only 3 months.

Prevention

In the event of a nuclear accident, prompt administration of stable iodine prevents the uptake of inhaled and ingested radioactive iodine isotopes. Emergency arrangements should be in place close to nuclear installations to provide for distribution of potassium iodate tablets, and arrangements for the United Kingdom are detailed in the Department of Health document PL/CMO (93) 1.

Special problems in pregnancy

A solitary nodule in a pregnant woman should be evaluated by fine needle aspiration biopsy. If the biopsy suggests malignancy, surgery can be undertaken in the second or third trimester, but if the nodule is discovered late in the third trimester, it is probably best to defer surgery until after delivery.

Medullary carcinoma of the thyroid

This accounts for 5 to 10 per cent of all thyroid cancers. About 80 per cent are sporadic with a peak incidence at 40 to 50 years of age. Hereditary autosomal dominant forms occur as part of multiple endocrine neoplasia type 2A or 2B or as isolated familial medullary carcinoma. These forms are associated with germline point mutations in the *ret* proto-oncogene and preneoplastic C cell hyperplasia (Table 3).

The pathological findings are of an encapsulated tumour with round, spindle shaped, or polyhedral cells arranged in a variety of patterns that have no prognostic significance. There is variable fibrosis and three-quarters of tumours show marked deposition of amyloid—a feature associated with a good prognosis. Heterogeneous staining for calcitonin, a hormone of C cells, is associated with a poorer outcome, reflecting dedifferentiation. Even the smallest medullary tumours may be associated with local lymph node metastases.

The presentation of sporadic medullary carcinoma is typically with a solitary thyroid nodule, accompanied by cervical lymphadenopathy in 50 per cent of cases. Lung, liver, or bone metastases are present at diagnosis in 10 per cent of cases. Symptoms due to local invasion or the paraneoplastic production of polypeptides and prostaglandins (flushing, diarrhoea, and Cushing's syndrome) are less common presenting features.

The diagnosis is often apparent from fine needle aspiration biopsy. Basal serum calcitonin concentrations are almost invariably elevated and confirm the diagnosis. There is controversy over the utility of routine serum calcitonin measurement in the work-up of all thyroid nodules: most centres perform aspiration biopsy initially. Newly diagnosed patients should be screened for other evidence of multiple endocrine neoplasia and a careful family history is also essential. In particular, phaeochromocytoma occurring as part of an inherited cancer syndrome must be excluded before surgery.

Testing genomic DNA for *ret* mutations in the germline is now widely available and should ideally be carried out on leucocyte DNA from all new patients. The absence of the most common mutations, coupled with a negative family history and the absence of C-cell hyperplasia or multicentric tumours in the resected thyroid, indicates that further family testing is not warranted. When a *ret* mutation is detected, there is a clear benefit from family testing, as prophylactic thyroidectomy in affected individuals improves outcome. However, there are some kindreds in whom familial medullary carcinoma occurs without a recognizable *ret* mutation and family screening must then be undertaken annually, up to the age of 35 to 40 years, using pentagastrin-stimulated serum calcitonin measurements as a guide to the presence of the inherited abnormality.

Medullary carcinoma should be treated by total thyroidectomy, with dissection of the central and other involved lymph nodes; this may require a second completion operation if the diagnosis is not made at the outset. Thyroxine replacement is needed at physiological doses rather than doses which suppress thyroid-stimulating hormone. After surgery the patient should be monitored by measurement of serum calcitonin concentration. Cure, defined as a persistently normal calcitonin level,

occurs in only about one-third of patients, but 80 to 90 per cent of patients in whom there is an elevated calcitonin level and only nodal disease survive for 10 years. The best management of persistent disease is unclear, but local recurrence with identifiable lymph node involvement should be dealt with surgically. Radiotherapy and chemotherapy have a variable and at best partial effect. Profuse (secretory) watery diarrhoea is frequently a troublesome feature of extensive disease; this may respond to treatment with loperamide, but somatostatin analogues, for example octreotide, may be needed.

Age, stage and size of tumour, and completeness of surgical removal are important prognostic features. Familial medullary carcinoma has the best outcome; in contrast, the tumour associated with multiple endocrine neoplasia type 2B is very aggressive. The overall 10-year survival is 70 to 80 per cent.

Primary thyroid lymphoma

Less than 5 per cent of thyroid malignancies are non-Hodgkin's B-cell lymphoma. The peak incidence is between 50 and 80 years of age, and women are affected three times more frequently than men. The typical presentation is a rapidly enlarging thyroid mass in a patient with Hashimoto's thyroiditis. The clinical features may suggest anaplastic carcinoma. The diagnosis can be made by fine needle aspiration biopsy and confirmed by large needle or open biopsy. Accurate staging is then necessary to plan treatment, which is with external beam radiotherapy and anthracycline-based chemotherapy. Intensive treatment has produced 8-year survival rates approaching 100 per cent.

Further reading

Ain KB (1998). Anaplastic thyroid carcinoma: behavior, biology, and therapeutic approaches. *Thyroid* **8**, 715–26.

Dulgeroff AJ and Hershman JM (1994). Medical therapy for differentiated thyroid carcinoma. *Endocrine Reviews* **15**, 500–15.

Fagin JA (1997). Editorial: familial nonmedullary thyroid carcinoma—the case for genetic susceptibility. *Journal of Clinical Endocrinology and Metabolism* **82**, 342–4.

Hay ID *et al.* (1993). Predicting outcome in papillary thyroid carcinoma: development of a reliable prognostic scoring system in a cohort of 1779 patients surgically treated at one institution during 1940 through 1989. *Surgery* **114**, 1050–8.

Hay ID *et al.* (1998). Unilateral total lobectomy: is it sufficient surgical treatment for patients with AMES low-risk papillary thyroid carcinoma? *Surgery* **124**, 958–66.

Hesmati HM *et al.* (1997). Advances and controversies in the diagnosis and management of medullary thyroid carcinoma. *American Journal of Medicine* **103**, 60–9.

Ladenson PW (1999). Strategies for thyrotropin use to monitor patients with treated thyroid carcinoma. *Thyroid* **9**, 429–33.

Matsuzuka F *et al.* (1993). Clinical aspects of primary thyroid lymphoma: diagnosis and treatment based on our experience of 119 cases. *Thyroid* **3**, 93–9.

Schlumberger MJ (1998). Papillary and follicular thyroid carcinoma. *New England Journal of Medicine* **338**, 297–306.

Schlumberger M, Baudin E (1998). Serum thyroglobulin determination in the follow-up of patients with differentiated thyroid carcinoma. *European Journal of Endocrinology* **138**, 249–52.

Suárez HG (1998). Genetic alterations in human epithelial thyroid tumours. *Clinical Endocrinology* **48**, 531–46.

Taylor T *et al.* (1998). Outcome after treatment of high-risk papillary and non-Hürthle-cell follicular thyroid carcinoma. *Annals of Internal Medicine* **129**, 622–7.

Wartofsky L *et al*. (1998). The use of radioactive iodine in patients with papillary and follicular thyroid cancer. *Journal of Clinical Endocrinology and Metabolism* **83**, 4195–200.

12.6 Parathyroid disorders and diseases altering calcium metabolism

R. V. Thakker

Calcium homeostasis

Most of the total of 1 kg of calcium in the healthy adult is present within the crystal structure of bone mineral and less than 1 per cent is in soluble form in the extracellular and intracellular fluid compartments. In the extracellular fluid compartment about half of the total calcium is ionized and the rest is principally bound to albumin or complexed with counter-ions. Ionized calcium concentrations range from 1.17 to 1.33 mmol/l, and the total serum calcium concentration ranges from 2.12 to 2.62 mmol/l. Measurements of free ionized calcium are not often undertaken because they are difficult; most laboratories report total serum calcium concentration for routine clinical use. However, the usual 2:1 ratio of total to ionized calcium may be disturbed by disorders such as metabolic acidosis, which reduces calcium binding by proteins, or by changes in protein concentration, caused by cirrhosis, dehydration, venous stasis, or multiple myeloma. In view of this, total serum concentrations are adjusted, or 'corrected', to a reference albumin concentration; thus, the corrected serum calcium may be related to a reference albumin concentration of 41 g/l and, for every 1 g/l of albumin above or below the reference value, the calcium is adjusted by 0.016 mmol/l, respectively. For example a total serum calcium of 2.70 mol/l with an albumin concentration of 47 g/l would be equivalent to a corrected serum calcium of 2.60 mmol/l, thereby correcting the initial apparent hypercalcaemic value to a normal value.

The extracellular concentration of calcium is closely regulated (Fig. 1) within the narrow physiological range that is optimal for those cellular functions that are affected by calcium. Indeed both hypercalcaemia and hypocalcaemia impair the function of many different organ systems. Regulation of extracellular calcium takes place through complex interactions (Fig. 2) at the target organs of the major calcium regulating hormone, parathyroid hormone (PTH), and vitamin D and its active metabolites, 1,25-dihydroxy (1,25$(OH)_2$) vitamin D. The parathyroid glands secrete PTH at a rate that is appropriate to, and depending upon the prevailing extracellular calcium ion concentration. Parathyroid gland disorders cause either hypercalcaemia or hypocalcaemia and these can be classified according to whether they arise from an excess of PTH, its deficiency, or insensitivity to its effects (Table 1, Fig. 2).

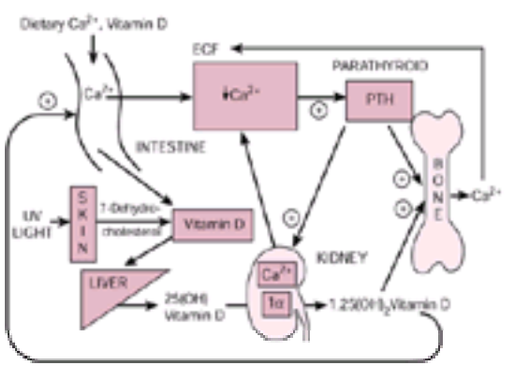

Fig. 1 Regulation of extracellular fluid (ECF) calcium (Ca^{2+}) by parathyroid hormone (PTH) action on kidney, bone, and intestine. A decrease in ECF Ca^{2+} is sensed by the calcium-sensing receptor (Fig. 2), and this leads to an increase in PTH secretion which predominantly acts directly on kidney and bone that possess the PTH-receptor (PTHR, Fig. 2). The skeletal effects of PTH are to increase (+) osteoclastic bone reabsorption but as osteoclasts do not have PTHRs, this action is mediated via the osteoblasts, which do have PTHRs and in response release cytokines and factors that activate osteoclasts. In the kidney, PTH stimulates (+) the 1a hydroxylase (1a) to increase the conversion of 25-hydroxy vitamin D (25(OH)D) to the active metabolite 1,25-dihydroxy vitamin D (1,25$(OH)_2$D). In addition, PTH, increases (+) the reabsorption of Ca^{2+} from the renal distal tubule and inhibits the reabsorption of phosphate from the proximal tubule, thereby leading to hypercalcaemia and hypophosphataemia. PTH also inhibits Na^+–H^+ antiporter activity and bicarbonate reabsorption, thereby causing a mild hyperchloraemic acidosis. The elevated 1,25$(OH)_2$D acts on the intestine to increase (+) absorption of dietary calcium and phosphate, and it is important to note that PTH does not appear to have a direct action on the gut. Thus, in response to hypocalcaemia and the increase in PTH secretion, all of these direct and indirect actions of PTH on the kidney, bone, and intestine will help to increase ECF Ca^{2+}, which in turn will act via the calcium-sensing receptor to decrease PTH secretion.

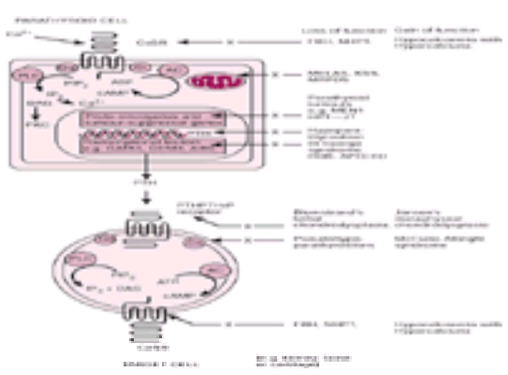

Fig. 2 Schematic representation of some of the components involved in calcium homeostasis. Alterations in extracellular calcium are detected by the calcium-sensing receptor (CaSR), which is a 1078 amino acid G-protein coupled receptor. The PTH/PTHrP-receptor, which mediates the actions of PTH and PTHrP, is also a G-protein coupled receptor. Thus, Ca^{2+} and PTH and PTHrP involve G protein-coupled signalling pathways and interaction with their specific receptors can lead to activation of Gs, Gi, and Gq, respectively. Gs stimulates adenylcyclase (AC) which catalyses the formation of cAMP from ATP. Gi inhibits AC activity. cAMP stimulates PKA which phosphorylates cell-specific substrates. Activation of Gq stimulates PLC, which catalyses the hydrolysis of the phosphoinositide (PIP_2) to inositol triphosphate (IP_3), which increases intracellular calcium, and diacylglycerol (DAG), which activates PKC. These proximal signals modulate downstream pathways, which result in specific physiological effects. Abnormalities in several genes, which lead to mutations in proteins in these pathways, have been identified in specific disorders of calcium homeostasis (Table 1). Adapted from Thakker RV, (2000). Parathyroid disorders. Molecular genetics and physiology. In: Morris PJ, Wood WC, eds. *Oxford Textbook of Surgery*, 2nd edn, pp. 1121–9. Oxford University Press, Oxford.

The *PTH* gene is located on chromosome 11p15 and consists of three exons (transcribed regions) which are separated by two introns. Exon 1 of the *PTH* gene is 85 bp in length and is untranslated whereas exons 2 and 3 code for the 115 amino acid pre-proPTH peptide. Exon 2 is 90 bp in length and encodes the initiation (ATG)

codon, the prehormone sequence and part of the prohormone sequence. Exon 3 is 612 bp and encodes the remainder of the prohormone sequence, the mature PTH peptide, and the 3' untranslated region. The 5' regulatory sequence of the human *PTH* gene contains a vitamin D response element 125 bp upstream of the transcription start site, which down-regulates *PTH* mRNA transcription in response to vitamin D receptor binding. *PTH* gene transcription (as well as PTH peptide secretion) is also dependent upon the extracellular calcium concentration, although the presence of a specific upstream 'calcium response element' has not yet been demonstrated.

The mature PTH peptide is secreted from the parathyroid chief cell as an 84 amino acid peptide; however, when the *PTH* mRNA is first translated it is as pre-proPTH peptide. The 'pre' sequence consists of a 25 amino acid signal peptide (leader sequence) which is responsible for directing the nascent peptide into the endoplasmic reticulum to be packaged for secretion from the cell. The 'pro' sequence is six amino acids in length and, although its function is less well defined than that of the 'pre' sequence, it is also essential for correct PTH processing and secretion. After the 84 amino acid mature PTH peptide is secreted from the parathyroid cell, it is cleared from the circulation with a short half-life of about 2 min, via non-saturable hepatic uptake and renal excretion.

PTH shares a receptor with PTH-related peptide (PTHrP); this PTH/PTHrP receptor (Fig. 1) is a member of a subgroup of the G protein-coupled receptor family. The PTH/PTHrP receptor gene is located on chromosome 3p21–p24 and is expressed in kidney and bone, where PTH is its predominant agonist. Expression of the PTH/PTHrP receptor also occurs in the brain, heart, skin, lung, liver, and testis where is mediates the actions of PTHrP. Mutations involving the genes that encode these proteins and receptors in this calcium regulating pathway (Fig. 2) are associated with hypercalcaemic and hypocalcaemic disorders (Table 1).

Hypercalcaemia

Clinical features and investigations

The clinical presentation of hypercalcaemia varies from a mild, asymptomatic, biochemical abnormality detected during routine screening to a life-threatening medical emergency. In general, the presence or absence of symptoms correlates with the severity and rapidity of onset of the hypercalcaemia. Thus, symptoms do not usually develop when serum calcium is below 3.00 mmol/l and are invariably present when the hypercalcaemia exceeds 3.50 mmol/l. However, there is a considerable variability and some patients may be symptomatic with mild hypercalcaemia (2.65–2.90 mmol/l). Although there are many causes of hypercalcaemia (Table 2), the signs and symptoms of hypercalcaemia are similar, regardless of aetiology. Indeed the clinical manifestations of hypercalcaemia involve several organ systems that include the renal, musculoskeletal, gastrointestinal, neurological, and cardiac systems (Table 3), and many of these have been referred to as 'moans, groans, pains, and stones'. Investigations should be directed at confirming the presence of hypercalcaemia and establishing the cause (Table 2).

The causes of hypercalcaemia can be classified according to whether serum PTH concentrations are elevated (i.e. primary hyperparathyroidism) or low (i.e. not due to a parathyroid tumour). Primary hyperparathyroidism and malignancy are the most common causes and account for more than 90 per cent of patients with hypercalcaemia. Detailed clinical history and examination will usually help to differentiate between these two diagnoses. In primary hyperparathyroidism, the hypercalcaemia is often less than 3.00 mmol/l, asymptomatic, and may have been present for months or years. If symptoms, for example nephrolithiasis, are present then they have usually been present for several months. However, in malignancy, the patients are usually acutely ill, often with neurological symptoms, the hypercalcaemia is more than 3.00 mmol/l, and the cancer (e.g. lung, breast, or myeloma) is often readily apparent. Hypercalcaemia from causes other than primary hyperparathyroidism or malignancy may also occur (Table 2) and a careful history (e.g. for vitamin D ingestion, drugs, renal disease) and examination (e.g. for thyrotoxicosis, adrenal disease, granulomatosis diseases), together with appropriate investigations (Table 4) are essential for establishing the diagnosis.

Management of hypercalaemia

The management of hypercalcaemia depends on the severity of the hypercalcaemia and the presence of symptoms. Thus, asymptomatic patients with mild hypercalcaemia, that is serum calcium below 3.00 mmol/l, do not usually need urgent treatment. However, a patient with severe hypercalcaemia, that is a serum calcium above 3.50 mmol/l, would require treatment regardless of symptoms, whilst a patient with moderate hypercalcaemia, that is a serum calcium in the range 3.00 to 3.50 mmol/l, would require urgent treatment if symptomatic. Before instituting treatment, it is always important to consider the underlying causes (Table 2) and to initiate investigations (Table 4).

The acute management of hypercalcaemia involves general measures to enhance hydration and diuresis, and specific measures using drugs to lower serum calcium. Dehydration due to hypercalcaemic symptoms, for example anorexia, nausea, vomiting and polyuria because of defective urinary concentration, is very common and patients may require 5 to 10 l of 0.9 per cent sodium chloride over a 24 to 48-h period. This vigorous hydration with normal saline may lower serum calcium by 0.25 to 0.75 mmol/l; it enhances urinary calcium excretion by increasing glomerular filtration and reducing proximal and distal renal tubular reabsorption of calcium and sodium. This saline diuresis may need adjuvant therapy with a loop diuretic, for example frusemide 20 to 100 mg to control complications due to volume overload, especially in the elderly and those with impaired cardiovascular and renal function. Saline diuresis may lead to hypokalaemia, hypomagnesaemia, and electrolyte imbalance, which will need correction.

If saline diuresis is not successful, and particularly if the hypercalcaemia is very severe, then more specific measures, for example dialysis and/or drugs, will be required. The drug of choice is pamidronate, which is a potent bisphosphonate that is administered parenterally. A recommended regimen is to administer 60 to 90 mg intravenously as a single infusion. Other bisphosphonates, for example etidronate and clodronate and other agents such as mithramycin, calcitonin, and gallium nitrate, have also been used in the past. Glucocorticoid therapy (e.g. hydrocortisone 120 mg/day in three divided doses) is particularly effective when the hypercalcaemia is mediated by the actions of 1,25-dihydroxy vitamin D, for example in granulomatosis disease or lymphoma (Table 2), or myeloma. Once the acute management of hypercalcaemia has been completed, then appropriate treatment for the underlying cause, for example parathyroidectomy for primary hyperparathyroidism, needs to be undertaken.

Hypercalcaemic diseases

Hypercalcaemia may arise through one more of three mechanisms: increased bone resorption, increased gastrointestinal absorption of calcium, and decreased renal calcium excretion. For example: lytic bone metastases cause increased bone resorption; thiazide diuretics lead to a decrease in calcium excretion; and excessive PTH will either directly or indirectly, by increasing 1,25-dihydroxy vitamin D production, stimulate bone resorption and calcium absorption from the gut and renal tubules. The hypercalcaemic diseases may be classified according to whether serum PTH concentrations are elevated or reduced (Table 2). In addition, hypercalcaemia may also be classified as being due to: an excess of PTH (e.g. primary or tertiary hyperparathyroidism) from parathyroid tumours; an excessive production of PTHrP; a defect in the PTH receptor (i.e. the PTH/PTHrP receptor); an excess production of down-stream mediators, for example 1,25-dihydroxy vitamin D; or an altered set point in the calcium-sensing receptor (Fig. 2).

Hyperparathyroidism

Hyperparathyroidism is characterized by high concentrations of serum immunoreactive PTH, and three types, referred to as primary, secondary, and tertiary, are recognized. Primary and tertiary hyperparathyroidism are associated with hypercalcaemia (Table 2), whereas secondary hyperparathyroidism is associated with hypocalcaemia (see below). Primary hyperparathyroidism may arise as an isolated endocrinopathy or as part of a multiple endocrine neoplasia (MEN) syndrome, and tertiary hyperparathyroidism usually arises in association with chronic renal failure.

Primary hyperparathyroidism

Primary hyperparathyroidism, which affects 1 in 1000 adults, is one of the two most common causes of hypercalcaemia and is due to an excessive secretion of PTH from one or more parathyroid tumours. In 80 per cent of patients this tumour is a solitary parathyroid adenoma, and in 15 per cent to 20 per cent of patients hyperplasia involving all four parathyroids is present. Parathyroid carcinoma accounts for less than 0.5 per cent of patients with primary hyperparathyroidism. Primary hyperparathyroidism usually occurs between the ages of 40 to 65 years, and is three times more common in females than males. The underlying causes of primary hyperparathyroidism are largely unknown, but abnormalities of several genes have been identified. Thus, abnormalities of the *cyclin D1* (*CCNDI*), *retinoblastoma*, *calcium-sensing receptor* (*CaSR*), *MEN type 1* (*MEN1*), and *type 2* (*MEN2*) genes together with other genes, yet to be identified, on chromosomes 1p and 1q (Table 1) are associated with the development of some parathyroid tumours.

Clinical features

Many patients with primary hyperparathyroidism will be asymptomatic and the hypercalcaemia, which is usually mild, will have been detected by chance at the time of biochemical screening for other reasons. However, it is important to note that nearly half the patients will have subtle neuromuscular symptoms such as fatigue and weakness and this becomes apparent only in retrospect after a successful parathyroidectomy.

Symptomatic hypercalcaemia (Table 3) predominantly affects the skeletal, renal, and gastrointestinal systems; peptic ulcers and pancreatitis may develop. The skeletal changes of osteitis fibrosa cystica due to subperiosteal resorption of the distal phalanges (Fig. 3), tapering of the distal clavicles, a 'salt and pepper' appearance of the skull, bone cysts, and brown tumours of the long bones are now identified in less than 5 per cent of patients. However, osteopenia, as assessed by bone mineral density, occurs in 25 per cent of patients. Renal stone disease (nephrolithiasis and nephrocalcinosis) occurs in 20 per cent of patients and hypercalciuria occurs in 30 per cent of patients; renal impairment may complicate this disease.

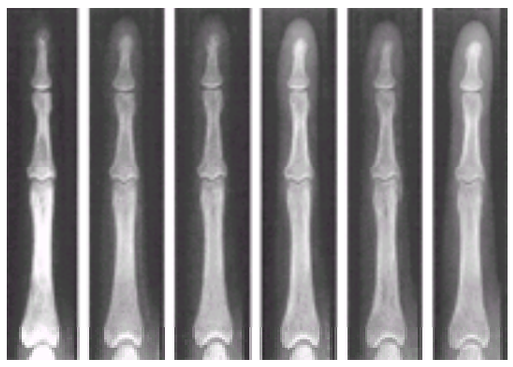

Fig. 3 Renal osteodystrophy over a 9-year period in a patient with chronic renal failure. Marked periosteal erosions were seen (a) despite treatment with 1a-hydroxy cholecaliferol, and a resolution was observed following dialysis (b). Note the vascular calcification. One year later a relapse was noted with periosteal erosions (c) and the use of calcitriol resolved these (d). Unfortunately, a relapse occurred 2 years later (e), and following renal transplantation a marked resolution was observed (f).

Investigations

In the presence of hypercalcaemia, the finding of elevated circulating PTH concentrations establishes the diagnosis, as the PTH will be elevated in approximately 90 per cent of patients with primary hyperparathyroidism, who will invariably have hypercalcaemia. However, it is important to make sure that the immunoradiometric (IRMA) and immunochemiluminometric (ICMA) assays for PTH are being used to measure the intact molecule, rather than the older radioimmunoassays which were not as reliable. The only other hypercalcaemic disorders in which PTH may occasionally be elevated are those related to familial benign hypocalciuric hypercalcaemia (FBH or FHH), immobilization, or lithium or thiazide use (Table 2), and a careful history and a cessation of drug use will help to exclude these possibilities. About one-third of patients with primary hyperparathyroidism will have a low serum phosphate and in the others it will be in the lower range of normal. In addition, some patients will have a small increase in serum chloride concentration and a concomitant decrease in bicarbonate concentration. Serum alkaline phosphatase activity may be elevated in some patients, and urinary calcium excretion is increased in 30 per cent of patients. The circulating 1,25-dihydroxy vitamin D concentration is elevated in some patients with primary hyperparathyroidism although it is not of diagnostic value as it is also elevated in other hypercalcaemic disorders such as sarcoidosis and lymphomas. The serum 25-hydroxy vitamin D concentration is within the normal range. Densitometric scanning is of use in detecting early skeletal changes. Patients with primary hyperparathyroidism develop reduced bone mineral densities (osteopenia) primarily of the cortical bone (e.g. distal third of forearm) rather than the cancellous bone (e.g. lumbar spine). The hip bones, which are an equal mixture of cortical and cancellous bone, show intermediate reductions in bone mineral density. Overall, the risk of bone fractures in patients with mild primary hyperparathyroidism is similar to those in matched, normal controls. However, successful parathyoidectomy does lead to an increase in bone mineral density over a 6 to 12-month period and this continues for up to 10 years. Indeed, bone mineral density measurements are used in the evaluation of patients with primary hyperparathyroidism and in deciding conservative as opposed to surgical management (Table 5).

Preoperative localization to define the site(s) of the parathyroid tumours may be undertaken. The non-invasive tests consist of ultrasonography, computed tomography (CT), magnetic resonance imaging (MRI), and scintigraphy with technetium-99m sestamibi. Sestamibi scintigraphy has now become established as the best and most convenient localization test; this can be performed with computed tomographic techniques (SPECT) to give a three-dimensional image with greater anatomical resolution. It is important to note that there is an appreciable incidence of false positive rates with all the non-invasive localization procedures and so a confirmation using two methods is preferable. Invasive localization tests consist of arteriography and selective venous sampling for PTH in the veins draining the thyroidal region. These tests are time-consuming, expensive, difficult, and dependent on the skill of the radiologist. It is generally accepted that these preoperative localization tests are indicated in those patients who have had previous neck surgery. However, their role in patients who have not had prior surgery remains to be established and at present the preferences and expertise of the local medical, radiology, and surgery teams usually determine the use of venous sampling procedures.

Management and treatment

Parathyroidectomy, which is the definitive cure, is a generally successful and safe procedure if undertaken by an experienced surgeon. Thus surgery is recommended for symptomatic patients and for those who have skeletal and renal complications. However, the decision to recommend surgery, which does have a small risk, may be difficult in asymptomatic patients, who may constitute over 50 per cent of patients with primary hyperparathyroidism. The natural history of primary hyperparathyroidism in most patients is to progress slowly or not at all. For example amongst asymptomatic patients, only 25 per cent will have progressive disease, which is usually manifested as a decrease in bone mineral density during a 10-year period. This has lead to a controversy regarding the indications for surgery, and guidelines have been provided by the Consensus Development Conference on the Management of Asymptomatic Primary Hyperparathyroidism (Table 5). However, these guidelines may not exclusively influence the decision for or against surgery and a careful evaluation and assessment of the risks and benefits is considered by most medical and surgical teams in conjunction with the patient. Clearly, some patients will not wish to continue living with a curable disease and will prefer surgery despite the guidelines (Table 5), whilst other patients will decline surgery, despite having guideline indications for surgery, because they may have coexisting medical conditions that make them feel that the risks of surgery are too great.

Patients who do not undergo parathyroidectomy should be evaluated clinically, and also monitored for serum calcium, creatinine, and PTH at 6 to 12-monthly intervals, and for bone mineral density and nephrolithiasis at 12-monthly intervals. In addition, the following medical guidelines are recommended. First, they should avoid dehydration and remain ambulant. Second, the dietary intake of calcium should be moderate, that is at or below 1000 mg/day, and thiazide diuretics should be avoided. Finally, they should avoid herbal and tonic remedies that may contain Vitamin D or A. These measures may help and at present an effective and safe drug for the treatment of primary hyperparathyroidism is not available. Drugs that have been used include oral phosphate, oestrogens in postmenopausal women, and the bisphosphonates—alendronate and clodronate. Phosphate is not used because of concerns related to soft tissue ectopic calcification, and although oestrogen does increase bone density in postmenopausal women, it has little effect on the serum calcium and PTH concentrations. The bisphosphonates inhibit bone resorption and do reduce serum calcium. However, these effects are not sustained. A more targeted approach using drugs that alter the function of the calcium sensing receptor, CaSR (see below), is being evaluated, and these calcimimetic agents may provide a medical therapy; for example use of such an agent has been shown to reduce serum calcium and PTH in postmenopausal women with asymptomatic primary hyperparathyroidism.

Uraemic hyperparathyroidism

Serum PTH levels rise in response to hypocalcaemia and this secondary hyperparathyroidism usually resolves with treatment of the underlying cause of hypocalcaemia (Table 6). However, in chronic renal failure the secondary hyperparathyroidism may persist for a longer time, and eventually the parathyroid cells gain an autonomous function, secreting excessive PTH despite hypercalcaemia; this state is referred to as tertiary hyperparathyroidism (Table 2). The cause of progression from the early, presumably polyclonal, secondary hyperplasia of the parathryoids to the later, presumably monoclonal, tumours is not understood and appears to involve genes other than those involved in the aetiologies of the sporadic and familial forms of primary hyperparathyroidism (Table 1).

Clinical features and treatment

In chronic renal failure, the ensuing phosphate retention and decreased production of 1,25-dihydroxy vitamin D result in hypocalcaemia and secondary hyperparathyroidism. This combination of biochemical abnormalities results in a severe bone disease that shows combined features of hyperparathyroidism and vitamin D deficiency (i.e. osteomalacia). Thus in renal osteodystrophy, bone erosions (Fig. 3) and osteomalacia are simultaneously observed. Treatment is based on correcting the hypocalcaemia, for example with oral administration of calcium salts, which also ameliorates the hyperphosphataemia by chelating phosphate in the intestines, and with calcitriol (1,25-dihydroxy vitamin D). The use of the most appropriate phosphate binder is not well established but it is clear that aluminium-containing compounds are to be avoided. Aluminium in these preparations and as a contaminant of dialysis solutions contributed in the recent past to the osteomalacic osseous disease and other aspects of metal toxicity in patients with renal failure (e.g. hypochromic anaemia and encephalopathy). Early treatment of the metabolic disturbance will prevent or delay the onset of severe secondary hyperparathyroidism and tertiary hyperparathyroidism, which requires parathyroidectomy.

Familial primary hyperparathyroidism

Primary hyperparathyroidism is most frequently encountered as a non-familial disorder. However, approximately 10 per cent of patients with primary hyperparathyroidism will have a hereditary form which may either be part of the multiple endocrine neoplasia type 1 (MEN1) and type 2 (MEN2) syndromes, or part of the hereditary hyperparathyroidism–jaw tumour (HPT-JT) syndrome. In addition, hereditary primary hyper-parathyroidism may develop as a solitary endocrinopathy and this has also been referred to as familial isolated hyperparathyroidism (FIHP). Investigations of these hereditary and sporadic forms of primary HPT have helped to identify some of the genes and chromosomal regions that are involved in the aetiology of parathyroid tumours (Table 1). FIHP has been reported in several kindreds, and some have been shown to harbour mutations of the MEN1 gene whilst in other families linkage to polymorphic loci from chromosome 1q21–q31, the region of the HPT-JT syndrome, has been shown. In addition, analysis of parathyroid tumours from FIHP patients has revealed loss of heterozygosity (LOH) involving chromosome 1q21–q31 loci. FIHP located on chromosome 1q21–q31 has been reported to be associated with a high incidence of early-onset parathyroid carcinomas. These familial syndromes associated with parathyroid tumours will be briefly reviewed.

Multiple endocrine neoplasia type 1 (MEN 1)

MEN1 is characterized by the combined occurrence of tumours of the parathyroids, pancreatic islet cells, and anterior pituitary. Parathyroid tumours occur in 95 per cent of MEN1 patients, and the resulting hypercalcaemia is the first manifestation of MEN1 in about 90 per cent of patients. Pancreatic islet cell tumours occur in 40 per cent of MEN1 patients and gastrinomas, leading to the Zollinger–Ellison syndrome, are the most common type and also the important cause of morbidity and mortality in MEN1 patients. Anterior pituitary tumours occur in 30 per cent of MEN1 patients, with prolactinomas representing the most common type. Associated tumours, which may also occur in MEN1, include adrenal cortical tumours, carcinoid tumours, lipomas, angiofibromas, and collagenomas. The gene causing MEN1, which is located on chromosome 11q13 and represents a putative tumour suppressor gene, consists of 10 exons that encode a 610 amino acid protein, referred to as 'MENIN'. The majority (>80 per cent) of the germ line *MEN1* mutations in families are inactivating. MENIN has been shown to be located in the nucleus, where it directly interacts with the N-terminus of the AP1 transcriptional factor JunD. MENIN suppresses JunD-activated transcription and thus MENIN acts via the transcriptional regulation pathway to control cell proliferation.

Multiple endocrine neoplasia type 2 (MEN2)

MEN2 describes the association of medullary thyroid carcinoma (MTC), phaeochromocytomas, and parathyroid tumours. Three clinical variants of MEN2 are recognized—MEN2a, MEN2b, and MTC-only. MEN2a is the most common variant, where the development of MTC is associated with phaeochromocytomas (50 per cent of patients), which may be bilateral, and parathyroid tumours (20 per cent of patients). MEN2b, which represents 5 per cent of all MEN2 cases, is characterized by the occurrence of MTC and phaeochromocytoma in association with a Marfanoid habitus, mucosal neuromas, medullated corneal fibres, and intestinal autonomic ganglion dysfunction leading to multiple diverticulae and megacolon. Parathyroid tumours do not usually occur in MEN2b. MTC-only is a variant in which medullary thyroid carcinoma is the sole manifestation of the syndrome. The gene causing all three MEN2 variants was mapped to chromosome 10cen–10q11.2, a region containing the c-*RET* proto-oncogene which encodes a tyrosine kinase receptor with cadherin-like and cysteine-rich extracellular domains and a tyrosine kinase intracellular domain. Specific mutations of c-*RET* have been identified for each of the three MEN2 variants. Thus in 95 per cent of patients, MEN2a is associated with mutations of the cysteine-rich extracellular domain and mutations in codon 634 (Cys®Arg) account for 85 per cent of MEN2a mutations. MTC-only is also associated with missense mutations in the cysteine-rich extracellular domain and most mutations are at codon 618. MEN2b is associated with mutations in codon 918 (Met®Thr) of the intracellular tyrosine kinase domain in 95 per cent of patients. Mutational analysis of c-*RET* to detect mutations in codons 609, 611, 618, 634, 768, and 804 in MEN2a and MTC-only, and codon 918 in MEN2b, has been used in the diagnosis and management of patients and families with these disorders.

Hyperparathyroidism–jaw tumour (HPT-JT) syndrome

The HPT-JT syndrome is an autosomal dominant disorder characterized by the development of parathyroid tumours and fibro-osseous jaw tumours. In addition, some patients may also develop Wilms' tumours, renal cysts, renal hamartomas, renal cortical adenomas, papillary renal cell carcinomas, pancreatic adenocarcinomas, testicular mixed germ cell tumours with a major seminoma component, and Hurthle cell thyroid adenomas. It is important to note that the parathyroid tumours may occur in isolation and without any evidence of jaw tumours, and this may cause confusion with other hereditary hypercalcaemic disorders such as MEN1, familial benign hypercalcaemia (FBH), which is also referred to as familial hypocalciuric hypercalcaemia (FHH), and FIHP. HPT-JT can be distinguished from FBH, as in FBH serum calcium concentrations are elevated from the early neonatal or infantile period whereas in HPT-JT such elevations are uncommon in the first decade. In addition, HPT-JT patients, unlike in FBH, will have associated hypercalciuria. The distinction between HPT-JT patients and MEN1 patients, who have only developed the usual first manifestation of hypercalcaemia (>90 per cent of patients), is more difficult and is likely to be influenced by the operative and histological findings and the occurrence of other characteristic lesions in each disorder. It is noteworthy that HPT-JT patients will usually have single adenomas or a carcinoma, whilst MEN1 patients will often have multiglandular parathyroid disease. The distinction between FIHP and HPT-JT in the absence of jaw tumours is difficult but important as HPT-JT patients may be at a higher risk of developing parathyroid carcinomas. These distinctions may be helped by the identification of additional features, and a search for jaw tumours, renal, pancreatic, thyroid, and testicular abnormalities may help to identify HPT-JT patients. The jaw tumours in HPT-JT are different from the brown tumours observed in some patients with primary hyperparathyroidism, and do not resolve after parathyroidectomy. Indeed ossifying fibromas of the jaw are an important distinguishing feature of HPT-JT from FIHP, and the occurrence of these may occasionally precede the development of hypercalcaemia in HPT-JT patients by several decades. The gene causing HPT-JT has been mapped to chromosome 1q25–q31.

Malignancy

The hypercalcaemia of malignancy is usually due to increased bone resorption, which may either be directly due to skeletal metastases or indirectly due to tumour-production of a humoral factor that stimulates osteoclastic bone resorption. The cancers that typically metastasize to produce lytic bone lesions are from the breast, lymphomas, or multiple myeloma (Table 2). The cancers that are typically associated with the humoral hypercalcaemia of malignancy (HHM) are squamous carcinomas of the lung, oesophagus, cervix, vulva, skin, head, or neck, but other types from the kidney, bladder, ovary, and breast may also occur. HHM accounts for up to 80 per cent of patients with malignancy-associated hypercalcaemia. The most common factor causing HHM is parathyroid hormone related peptide (PTHrP), which can be measured in the serum by immunoassay. However, these assays are relatively insensitive and the failure to detect serum PTHrP does not exclude the diagnosis of HHM. Patients with HHM generally have hypercalcaemia associated with lower or undetectable serum PTH levels, marked hypercalcaemia, and a reduced plasma 1,25-dihydroxy vitamin D level. Therapy of HHM is aimed at: (1) reducing the tumour load by surgery, radiotherapy, and/or chemotherapy; (2) reducing osteoclastic bone resorption by use of bisphosphonates or calcitonin; and (3) increasing renal calcium clearance by a saline diuresis.

Granulomatous disorders

Several granulomatous disorders are associated with hypercalcaemia (Table 2) and this is invariably associated with elevated circulating concentrations of 1,25-dihydroxy vitamin D, which is due to extrarenal synthesis. Sarcoidosis is the most frequently encountered granulomatous disorder associated with hypercalcaemia and 10 per cent of patients with sarcoidosis will have hypercalcaemia and about one-half will become hypercalciuric. The finding of raised serum angiotensin converting enzyme (ACE) activity may help in confirming the diagnosis. Glucocorticoids (e.g. 40 to 60 mg of prednisolone) decrease 1,25-dihydroxy vitamin D production and restore the calcium concentration to normal. Failure to achieve normal serum calcium concentrations within 10 days of glucorticoid therapy (e.g. hydrocortisone 40 mg, three times per day), which is referred to as the steroid suppression test, should suggest the coexistence of another cause for the hypercalcaemia, for example primary hyperparathyroidism or malignancy.

Endocrine causes of hypercalcaemia other than hyperparathyoidism

Several non-parathyroid disorders (Table 2) are associated with hypercalcaemia and these include thyrotoxicosis, phaeochromocytoma, Addison's disease, VIPomas, familial benign hypocalciuric hypercalcaemia, Jansen's disease, and William's syndrome.

Thyrotoxicosis

Mild hypercalcaemia (<3.00 mmol/l) frequently accompanies thyrotoxicosis, which leads to increased bone turnover and resorption. The hypercalcaemia may respond to treatment with b-adrenergic blockers.

Familial benign hypocalciuric hypercalcaemia and neonatal primary hyperparathyroidism

Familial benign hypercalcaemia (FBH), which is also referred to as familial hypocalciuric hypercalcaemia (FHH), is an autosomal dominant disorder with a high degree of penetrance, that is characterized by lifelong asymptomatic hypercalcaemia in association with an inappropriately low urinary calcium excretion (i.e. calcium clearance to creatinine clearance ratio <0.01). A normal circulating parathyroid hormone (PTH) concentration and mild hypermagnesaemia are also typically present. Although most patients with FBH are asymptomatic, chondrocalcinosis and acute pancreatitis have occasionally been observed. In addition, children of consanguineous marriages within FBH kindreds have been observed to have life-threatening hypercalcaemia due to neonatal primary hyperparathyroidism (NHPT). NHPT is defined as symptomatic hypercalcaemia with skeletal manifestations of hyperparathyroidism in the first 6 months of life. NHPT children often present in the first few days or weeks of life with failure to thrive, dehydration, hypotonia, constipation, rib cage deformities, and multiple fractures due to bony undermineralization. Children with NHPT often require urgent parathyroidectomy, which corrects the PTH-dependent hypercalcaemia and bone demineralization. FBH is due to heterozygous inactivating mutations of the calcium sensing receptor (CaSR) and NHPT is often associated with inactivating homozygous CaSR mutations (Fig. 2). However, NHPT has also been observed in children where only one parent had clinically apparent FBH, and many other NHPT patients appear to be sporadic, that is both parents have normal serum calcium concentrations. In such NHPT patients with heterozygous CaSR mutations, the mutant CaSR may exert a dominant negative action on the normal CaSR. The human CaSR is a 1078 amino acid cell surface protein which is expressed in parathyroids, thyroid cells, and kidney, and is a member of the family of G-protein coupled receptors. The *CaSR* gene is located on chromosome 3q21–q24.

Jansen's disease

Jansen's disease is an autosomal dominant disease that is characterized by short-limbed dwarfism, due to metaphyseal chondrodysplasia, and severe hypercalcaemia and hypophosphataemia, despite normal or undetectable serum levels of PTH. These abnormalities are associated with activating mutations of the PTH receptor (Fig. 2) and thus this represents a PTH-independent activation of the PTH receptor (PTHR). Two different mutations of the PTHR have been identified, and these involve codon 223 (His®Arg) and codon 410 (Thr®Pro). Expression of the mutant receptors in COS-7 cells resulted in constitutive, ligand-independent accumulation of cAMP, while the basal accumulation of inositol phosphates was not increased. These findings provide a likely explanation for the abnormalities observed in mineral homeostasis and growth plate development in this disorder.

William's syndrome

William's syndrome is an autosomal dominant disorder characterized by supravalvular aortic stenosis, elfin-like facies, psychomotor retardation, and infantile hypercalcaemia. The underlying abnormality of calcium metabolism remains unknown but abnormal 1,25-dihydroxy vitamin D_3 metabolism or decreased calcitonin production have been implicated, although no abnormality has been consistently demonstrated. Hemizygosity due to a microdeletion at the *ELASTIN* locus on chromosome 7q11.23 in over 90 per cent of patients with the classical William's phenotype has been demonstrated. This microdeletion has been reported to involve another gene, designated *LIM-KINASE*, that is expressed in the central nervous system. The calcitonin receptor gene has been localized to chromosome 7q21 and close to the region deleted in William's syndrome. However, the calcitonin receptor gene was not involved in the deletion found in four patients with William's syndrome, indicating that it is unlikely to be implicated in the hypercalcaemia of such children. While the involvement of the *ELASTIN* and *LIM-KINASE* genes in the deletions of William's syndrome patients can explain the respective cardiovascular and neurological features of this disorder, it seems possible that another, as yet uncharacterized, gene that is within this contiguously deleted region is likely to be involved to explain the abnormalities of calcium metabolism.

Drugs

Several drugs (Table 2) can cause hypercalcaemia by different mechanisms. Compounds containing vitamins D and A are common and frequently associated with hypercalcaemia. The use of thiazide diuretics is often associated with hypercalcaemia. The hypercalcaemia appears to be largely renal in origin, as thiazides enhance distal renal tubular calcium reabsorption. Hypercalcaemia reverses rapidly with discontinuation of the drug.

The milk-alkali syndrome was first described in the 1930s, generally in the context of ulcer treatment with large quantities of milk together with sodium bicarbonate. Today, the responsible agent is usually calcium carbonate, although consumption of large quantities of dairy products (milk, cheese, and yoghurt) may still contribute. Classical features include moderate to severe hypercalcaemia with alkalosis and renal impairment. The amount of calcium ingested by patients with this syndrome is usually 5 to 15 g/day. Treatment consists of: (1) discontinuing the ingestion of the calcium containing compounds(s) and antacids; (2) rehydration; and (3) saline diuresis.

Hypocalcaemia

Clinical features and investigations

The clinical presentation of hypocalcaemia (serum calcium <2.12 mmol/l) ranges from an asymptomatic biochemical abnormality to a severe, life-threatening condition. In mild hypocalcaemia (serum calcium 2.00–2.12 mmol/l), patients may be asymptomatic. Those with more severe (serum calcium <1.9 mmol/l) and long-term hypocalcaemia may develop: acute symptoms of neuromuscular irritability (Table 7); ectopic calcification (e.g. in the basal ganglia, which may be associated with extrapyramidal neurological symptoms); subcapsular cataract; papilloedema; and abnormal dentition. Investigations should be directed at confirming the presence of hypocalcaemia and establishing the cause. Hypocalcaemia (Table 6) can be classified by cause, according to whether serum parathyroid hormone (PTH) concentrations are low (i.e. hypoparathyroid disorders) or high (i.e. disorders associated with secondary hyperparathyroidism). Hypocalcaemia is most commonly caused by hypoparathyroidism, a deficiency or abnormal metabolism of vitamin D, acute or chronic renal failure, or hypomagnesaemia. In hypoparathyroidism, serum calcium is low, phosphate is high, and PTH is undetectable; renal function and concentrations of the 25-hydroxy and 1,25-dihydroxy metabolites of vitamin D are usually normal. The features of pseudohypoparathyroidism are similar to those of hypoparathyroidism except for PTH, which is markedly increased. In chronic renal failure, which is the most common cause of hypocalcaemia, phosphate is high and alkaline phosphatase, creatinine, and PTH are elevated; 25-hydroxy vitamin D_3 is normal and 1,25-dihydroxy vitamin D_3 is low. In vitamin D deficiency osteomalacia, serum calcium and phosphate are low, alkaline phosphatase and PTH are elevated, renal function is normal, and 25-hydroxy vitamin D_3 is low. The most frequent artifactual cause of hypocalcaemia is hypoalbuminaemia, such as occurs in liver disease or the nephrotic syndrome.

Management of acute hypocalcaemia

The management of acute hypocalcaemia depends on the severity of the hypocalcaemia, the rapidity with which it developed, and the degree of neuromuscular irritability (Table 7). Treatment should be given to symptomatic patients (e.g. with seizures or tetany) and asymptomatic patients with a serum calcium of less than 1.90 mmol/l who are at high risk of developing complications. The preferred treatment for acute symptomatic hypocalcaemia is calcium gluconate, 10 ml 10 per cent w/v (2.20 mmol of calcium) intravenous, diluted in 50 ml of 5 per cent dextrose or 0.9 per cent sodium chloride and given by slow injection (>5 min); this can be repeated as required to control symptoms. Serum calcium concentrations should be assessed regularly. Persistent hypocalcaemia may be managed acutely by administration of a calcium gluconate infusion; for example, dilute 10 ampoules of calcium gluconate, 10 ml 10 per cent w/v (22.0 mmol of calcium), in 1 litre of 5 per cent dextrose or 0.9 per cent sodium chloride, start infusion at 50 ml/h and titrate to maintain serum calcium concentrations in the normal range. Generally, 0.3 to 0.4 mmol/kg of elemental calcium infused over 4 to 6 h increases serum calcium by 0.5 to 0.75 mmol/l. If hypocalcaemia is likely to persist, oral vitamin D therapy (see below) should also be administered. In hypocalcaemic patients who are also hypomagnesaemic, the hypomagnesaemia must be corrected before the hypocalcaemia

will resolve. This may occur in the postparathyroidectomy period or in patients with severe malabsorption, for example those with established coeliac disease.

Management of persistent hypocalcaemia

The two main agents available for the treatment of hypocalcaemia are supplemental calcium, about 10 to 20 mmol calcium 6 to 12 hourly, and vitamin D preparations. Patients with hypoparathyroidism seldom require calcium supplements after the early stages of stabilization with vitamin D. A variety of vitamin D preparations have been used. These include: vitamin D_3 (cholecalciferol) or vitamin D_2 (ergocalciferol), 25 000 to 100 000 units (1.25–5 mg/day); dihydrotachysterol (now seldom used), 0.25 to 1.25 mg/day; alfacalcidol (1a-hydroxycholecalciferol), 0.25 to 1.0 µg/day; and calcitriol (1,25-dihydroxy cholecalciferol), 0.25 to 2.0 µg/day. In children, these preparations are prescribed in doses based on body weight. Cholecalciferol and ergocalciferol are the least expensive preparations but have the longest durations of action and may result in prolonged toxicity. The other preparations, which do not require renal 1a-hydroxylation, have the advantage of shorter half-lives and thereby minimize the risk of prolonged toxicity. Calcitriol is probably the drug of choice because it is the active metabolite and, unlike alfacalcidol, does not require hepatic 25-hydroxylation. Close monitoring (at about 1–2-week intervals) of the patient's serum and urine calcium concentrations are required initially, and at 3 to 6-monthly intervals once stabilization is achieved. The aim is to avoid hypercalcaemia, hypercalciuria, nephrolithiasis, and renal failure. It should be noted that hypercalciuria may occur in the absence of hypercalcaemia.

Hypocalcaemic diseases

Hypocalcaemic diseases (Table 6) may arise because of a destruction of the parathyroid glands, failure of parathyroid gland development, or reduced PTH secretion or PTH-mediated actions in target tissues. Thus, these diseases may be classified as being due to a deficiency of PTH, a defect in the PTH-receptor (i.e. the PTH/PTHrP receptor), or an insensitivity to PTH caused by defects down-stream of the PTH/PTHrP receptor (Fig. 2). The diseases may also be classified as being part of the hypoparathyroid disorders, of the calcium sensing receptor abnormalities, or of the pseudohypoparathyroid disorders.

Hypoparathyroidism

Hypoparathryoidism is characterized by hypocalcaemia and hyperphosphataemia, which are the result of a deficiency in parathyroid hormone (PTH) secretion or action. Serum concentrations of immunoreactive PTH are low or undetectable and the concentrations of 1,25-dihydroxy vitamin D_3 are usually in the low normal to low range but alkaline phosphatase activity is unchanged. The daily urinary excretion of calcium is reduced, although the fractional excretion of calcium is increased. Nephrogenous cyclic adenosine monophosphate (cAMP) excretion is low and renal tubular reabsorption of phosphate is elevated. Urinary cAMP, plasma cAMP, and urinary phosphate excretion increase markedly after administration of exogenous bioactive PTH (Chase–Aurbach and Ellsworth–Howard tests). Hypoparathyroidism may result from agenesis (e.g. the DiGeorge syndrome) or destruction of the parathyroid glands (e.g. following neck surgery, in autoimmune diseases), from reduced secretion of PTH (e.g. neonatal hypocalcaemia or hypomagnesaemia), or resistance to PTH (which may occur as a primary disorder (e.g. pseudohypoparathyroidism or secondary to hypomagnesaemia). In addition, hypoparathyroidism may occur as an inherited disorder (Table 1) that may either be part of a complex congenital defect (e.g. DiGeorge syndrome), or as part of a pluriglandular autoimmune disorder, or as a solitary endocrinopathy, which has been referred to as isolated or idiopathic hypoparathyroidism. Hypoparathyroidism may also complicate iron storage disease, especially secondary haemochromatosis in children and adolescents. In thalassaemic children, destruction of the parathyroids is associated with ill health and frank tetany, which may elude diagnosis and effective treatment unless hypoparathyroidism is suspected.

Isolated hypoparathyroidism

Isolated hypoparathyroidism may either be inherited or it may be acquired by damage to the parathyroids at surgery, by infiltrating metastases, or systemic disease (Table 6).

Inherited hypoparathyroidism

Patients with inherited forms of hypoparathyroidism may develop hypocalcaemic seizures in the neonatal or infantile periods and require life-long treatment with oral vitamin D preparations, for example calcitriol. Autosomal dominant, autosomal recessive, and X-linked recessive inheritances for hypoparathyroidism have been observed (Table 1). Some of the autosomal forms are due to mutations of the *PTH* gene, the calcium sensing receptor (see below), and the transcriptional factor *GCM2* (glial cells missing 2).

Acquired forms of hypoparathyroidism

Hypoparathyroidism may occur after neck surgery, irradiation, or because of infiltration by metastases or systemic disease, for example haemochromatosis, amyloidosis, sarcoidosis, Wilson's disease, or thalassaemia (Table 6). Surgical damage to the parathyroids occurs most commonly after a radical neck dissection, for example laryngeal or oesophageal carcinoma treatment, a total thyroid resection, or after repeated parathyroidectomies for multigland disease (e.g. in MEN1 or MEN2, see above). Hypocalcaemic symptoms begin 12 to 24 h postoperatively and may need treatment with oral or intravenous calcium. Parathyroid function often returns, but persistent hypocalcaemia requires treatment with vitamin D preparations.

Neonatal hypoparathyroidism resulting in hypocalcaemia may occur in the baby of a mother with hypercalcaemia caused by primary hyperparathyroidism. Maternal hypercalcaemia results in increased calcium delivery to the fetus, and this fetal hypercalcaemia suppresses fetal PTH secretion. Postpartum, the infant's suppressed parathyroids are unable to maintain normocalcaemia. The disorder is usually self-limiting, but occasionally therapy may be required. In addition, the feeding of cow's milk, which has a high phosphate content, to babies may also result in hypocalcaemia in some children.

Functional hypoparathyroidism may result from severe hypomagnesaemia (<0.40 mmol/l), which may be due to a severe intestinal malabsorption disorder (e.g. Crohn's disease) or a renal tubular disorder. It is associated with hypoparathyroidism because magnesium is required for the release of PTH from the parathyroid gland and also for PTH action via adenyl cyclase. Magnesium chloride, 35 to 50 mmol intravenously in 1 litre of 5 per cent glucose or other isotonic solution given over 12 to 24 h may be repeatedly required to restore normomagnesaemia.

Complex syndromes associated with hypoparathyroidism

Hypoparathyroidism may occur as part of a complex syndrome which may either be associated with a congenital developmental anomaly or with an autoimmune syndrome. The congenital developmental anomalies associated with hypoparathyroidism include the DiGeorge, the HDR (hypoparathyroidism, deafness, and renal anomalies), the Kenney–Caffey and Barakat syndromes, and also syndromes associated with either lymphoedema or dysmorphic features and growth failure (Table 1).

DiGeorge syndrome

Patients with the DiGeorge syndrome (DGS) suffer from neonatal hypoparathyroidism, T-cell immunodeficiency, congenital heart defects, and deformities of the ear, nose, and mouth (e.g. cleft lip and/or palate). Children with DGS often die from infections related to the immunodeficiency. The disorder arises from a congenital failure in the development of the derivatives of the third and fourth pharyngeal pouches with resulting absence or hypoplasia of the parathyroids and thymus. Most cases of DGS are sporadic but an autosomal dominant inheritance of DGS has been observed and an association between the syndrome and an unbalanced translocation and deletions involving chromosome 22q11.2 have also been reported. In some patients, deletions of another locus on chromosome 10p13–p14 have been observed in association with DGS and this is referred to as DGS type 2 (DGS2), whilst patients with the 22q11.2 deletions are referred to as DGS type 1 (DGS1). Studies of the DGS1 deleted region on chromosome 22q11.2 have revealed three genes (referred to as *RNEX40*, *NEX2.2-NEX3*, and *UDFIL* to be involved). However, although these are involved in the chromosomal deletions, the mechanisms by which they lead to the varied manifestations of DGS remain to be elucidated.

Hypoparathyroidism, deafness, and renal anomalies (HDR) syndrome

HDR is an autosomal dominant disorder in which patients often have asymptomatic hypocalcaemia with undetectable or inappropriately normal serum concentrations of PTH, and normal brisk increases in plasma cAMP in response to the infusion of PTH. Bilateral, symmetrical, sensorineural deafness involving all frequencies occurs, and the renal abnormalities consist mainly of bilateral cysts that compress the glomeruli and tubules and lead to renal impairment. Cytogenetic abnormalities

involving chromosome 10p14–10pter have been identified in HDR patients. HDR patients do not have immunodeficiency or heart defects, which are key features of DGS2, and indeed there are two non-overlapping regions; thus, the DGS2 region is located on 10p13–14 and HDR on 10p14–10pter. HDR patients have a haploinsufficiency of the zinc finger transcription factor GATA3.

Mitochondrial disorders associated with hypoparathyroidism

Hypoparathyroidism has been reported to occur in three disorders associated with mitochondrial dysfunction: the Kearns–Sayre syndrome (KSS), the MELAS syndrome, and a mitochondrial trifunctional protein deficiency syndrome (MTPOS). Kearns–Sayre syndrome is characterized by progressive external ophthalmoplegia and pigmentary retinopathy before the age of 20 years, and is often associated with heart block or cardiomyopathy. The MELAS syndrome consists of a childhood onset of mitochondrial encephalopathy, lactic acidosis, and stroke-like episodes. In addition, varying degrees of proximal myopathy can be seen in both conditions. Both the Kearns–Sayre and MELAS syndromes have been reported to occur with insulin dependent diabetes mellitus and hypoparathyroidism, and mitochondrial gene abnormalities have been identified in some patients. Mitochondrial trifunctional protein deficiency is a disorder of fatty acid oxidation that is associated with peripherial neuropathy, pigmentary retinopathy, and acute fatty liver degeneration in pregnant women who carry an affected fetus. Hypoparathyroidism has been observed in one patient with trifunctional protein deficiency.

Kenney–Caffey syndrome

Hypoparathyroidism has been reported to occur in over 50 per cent of patients with the Kenney–Caffey syndrome, which is associated with short stature, osteosclerosis, and cortical thickening of the long bones, delayed closure of the anterior fontanel, basal ganglia calcification, nanophthalmos, and hyperopia. Parathyroid tissue could not be found in a detailed post mortem examination of one patient and this suggests that hypoparathyroidism may be due to an embryological defect of parathyroid development. The molecular genetic defect has not been identified.

Additional familial syndromes

Single familial syndromes in which hypoparathyroidism is a component have been reported (Table 1). Thus, an association of hypoparathyroidism, renal insufficiency, and developmental delay has been reported in one Asian family in whom autosomal recessive inheritance of the disorder was established. The occurrence of hypoparathyroidism, nerve deafness, and a steroid-resistant nephrosis leading to renal failure, which has been referred to as the Barakat syndrome, has been reported in four brothers from one family, and an association of hypoparathyroidism with congenital lymphoedema, nephropathy, mitral valve prolapse, and brachytelephalangy has been observed in two brothers from another family. Molecular genetic studies have not been reported from these two families. A syndrome in which hypoparathyroidism was associated with severe growth failure and dysmorphic features has been reported in twelve patients from Saudi Arabia. Consanguinity was noted in 11 of the 12 patients' families, the majority of which originated from the Western province of Saudi Arabia. This syndrome, which is inherited as an autosomal recessive disorder, has also been identified in families of Bedouin origin and homozygosity and linkage disequilibrium studies have located this gene to chromosome 1q42–q43.

Blomstrand's disease

Blomstrand's chondrodysplasia is an autosomal recessive disorder characterized by early lethality, dramatically advanced bone maturation, and accelerated chondrocyte differentiation. Affected infants, who usually have consanguineous unaffected parents, develop pronounced hyperdensity of the entire skeleton with markedly advanced ossification, that results in extremely short and poorly modelled long bones. Mutations of the PTH/PTHrP receptor that impair its function are associated with Blomstrand's disease. Thus, it seems likely that affected infants will, in addition to the skeletal defects, have abnormalities in other organs, including secondary hyperplasia of the parathyroid glands, presumably due to hypocalcaemia.

Pluriglandular autoimmune hypoparathyroidism

This syndrome (Fig. 4) comprises of hypoparathyroidism, Addison's disease, candidasis, and two or three of the following: insulin-dependent diabetes mellitus, primary hypogonadism, autoimmune thyroid disease, pernicious anaemia, chronic active hepatitis, steatorrhoea (malabsorption), alopecia (totalis or areata), and vitiligo. The disorder has also been referred to as either the autoimmune polyendocrinopathy candidasis ectodermal dystrophy (APECED) syndrome or the polyglandular autoimmune type 1 syndrome. Antibodies directed against the adrenal, thyroid, and parathyroid glands are detected in the sera of some patients. The polyglandular autoimmune type 2 syndrome is characterized by adrenal insufficiency, insulin-dependent diabetes mellitus, and thyroid disease, and does not involve hypoparathyroidism. APECED, which has an autosomal recessive inheritance, has a high incidence in Finland and amongst Iranian Jews. The *APECED* gene, which has been located to chromosome 21q22.3, encodes a 545 amino acid protein that contains motifs suggestive of a transcriptional factor and includes a nuclear localization signal, two zinc-finger motifs, a proline-rich region, and three LXXLL motifs. The gene is referred to as *AIRE* (autoimmune regulator); six *AIRE* mutations have been reported in APECED families and a codon 257 (Arg®Stop) mutation was the predominant abnormality in 82 per cent of the Finnish families.

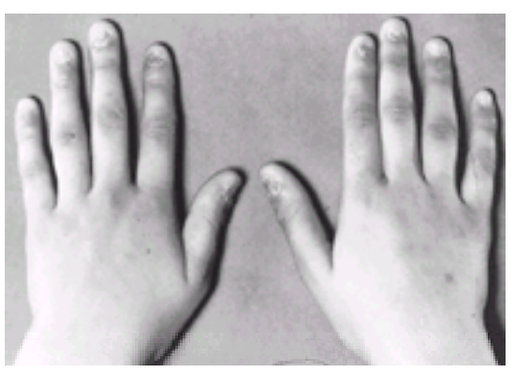

Fig. 4 Moniliasis and hyperpigmentation of the hands, particularly over the knuckles, is seen in this 8-year-old patient with hypoparathyroidism and Addison's disease. The patient also had vitiligo, and thus had some of the features of the polyglandular autoimmune syndrome type 1. Reproduced with permission from Thakker RV. Hypocalcaemic disorders. In: Thakker RV, Wass JAH, eds (1997). *Endocrine Disorders, Medicine*, vol. 25, pp. 68–70. The Medicine Group (Journals) Ltd, Abingdon, Oxon.

Calcium-sensing receptor (CaSR) abnormalities

The CaSR, which is located in the plasma membrane of the cell (Fig. 2), is at a critical site to enable the cell to recognize changes in extracellular calcium concentration. Thus, an increase in extracellular calcium leads to CaSR activation of the G-protein signalling pathway, which in turn increases the free intracellular calcium concentration and leads to a reduction in transcription of the *PTH* gene. CaSR mutations that result in a loss of function are associated with familial hypocalciuric hypercalcaemia (see above). However, CaSR missense mutations that result in a gain of function (or added sensitivity to extracellular calcium) lead to hypocalcaemia with hypercalciuria. These hypocalcaemic individuals are generally asymptomatic and have serum PTH concentrations that are in the low–normal range, and because of the insensitivities of previous PTH assays in this range such patients have often been diagnosed to be hypoparathyroid. In addition, such patients may have hypomagnesaemia. Treatment with vitamin D or its active metabolites to correct the hypocalcaemia in these patients results in marked hypercalciuria, nephrocalcinosis, nephrolithiasis, and renal impairment. Thus, these patients need to be distinguished from those with hypoparathyroidism.

Pseudohypoparathyroidism (PHP)

Patients with pseudohypoparathyroidism (PHP), which may be inherited as an autosomal dominant disorder, are characterized by hypocalcaemia and hyperphosphataemia due to PTH resistance rather than PTH deficiency. Five variants are recognized on the basis of biochemical and somatic features (Table 8) and three of these—PHP type Ia (PHPIa), PHP type 1b (PHPIb), and pseudopseudohypoparathyroidism (PPHP)—will be reviewed in further detail. Patients with PHPIa exhibit PTH resistance (hypocalcaemia, hyperphosphataemia, elevated serum PTH, and an absence of an increase in serum and urinary cyclic AMP and urinary

phosphate following intravenous human PTH infusion), together with the features of Albright's hereditary osteodystrophy (AHO), which includes short stature, obesity, subcutaneous calcification, mental retardation, round facies, dental hypoplasia, and brachydactyly (i.e. shortening of the metacarpals (Fig. 5), particularly the third, fourth, and fifth). In addition to brachydactyly, other skeletal abnormalities of the long bones and shortening of the metatarsals may also occur. Patients with PHPIb exhibit PTH resistance only and do not have the somatic features of AHO, whilst patients with PPHP exhibit the somatic features of AHO in the absence of PTH resistance. The absence of a normal rise in urinary excretion of cyclic AMP excretion after an infusion of PTH in PHPIa indicated a defect at some site of the PTH receptor–adenyl cyclase system (Fig. 2). This receptor system is regulated by at least two G proteins, one of which stimulates (Gsa) and another which inhibits (Gia) the activity of the membrane-bound enzyme that catalyses the formation of the intracellular second messenger cyclic AMP. Interestingly, patients with PHPIa may also show resistance to other hormones, for example thyroid-stimulating hormone (TSH), follicle-stimulating hormone (FSH), and luteinizing hormone (LH), that act via G-protein coupled receptors. Inactivating mutations of the Gsa gene (referred to as *GNAS1*), which is located on chromosome 20q13.2, have been identified in PHPIa and PPHP patients. However, *GNAS1* mutations do not fully explain the PHPIa or PPHP phenotypes, and studies of PHPIa and PPHP that occurred within the same kindred revealed that the hormonal resistance is parentally imprinted. Thus, PHPIa occurs in a child only when the mutation is inherited from a mother affected with either PHPIa or PPHP; and PPHP occurs in a child only when the mutation is inherited from a father affected with either PHPIa or PPHP. *GNAS1* mutations have not been detected in PHPIb, which has been considered to be due to a defect of the PTH/PTHrP receptor. However, studies of the PTH/PTHrP receptor gene and mRNA in PHPIb patients have not identified mutations, and linkage studies in four unrelated kindreds have mapped the PHPIb locus to chromosome 20q13.3, a location that also contains the *GNAS1* gene. In addition, parental imprinting of the genetic defect was observed and this is similar to the findings in kindreds with PHP-type Ia and/or PPHP. Two possible explanations for these observations have been proposed. First, PHPIb maybe due to a defect in a tissue- or cell-specific enhancer, or promoter, of the *GNAS1* gene and this may affect, directly or indirectly, the expression levels of the Gsa-specific transcripts. Or, second, PHPIb may be caused by a defect in a gene close to the *GNAS1* gene which is transcribed only from the maternal allele and affects PTH/PTHrP receptor or Gsa expression and/or function in some renal cells.

Fig. 5 Radiograph of both hands of a patient with pseudohypoparathyroidism type 1a. The patient has a normal right hand, but there is shortening of the left fourth metacarpal (brachydactyly). Metatarsals may be similarly shortened. Reproduced with permission from Thakker RV. Hypocalcaemic disorders. In: Thakker RV, Wass JAH, eds (1997). *Endocrine Disorders, Medicine*, vol. 25, pp. 68–70. The Medicine Group (Journals) Ltd, Abingdon, Oxon.

Further reading

Bilezikian JP (1999). Primary hyperparathyroidism. In: Favus MJ, ed. *Primer on the metabolic bone diseases and disorders of mineral metabolism*, 4th edn, pp. 187–92. Lippincott–Raven, Philadelphia.

Bilezikian J, Thakker RV (1998). Hypoparathyroidism. *Current Opinion in Endocrinology and Diabetes* **4**, 427–32.

Bouillon R (2001). Vitamin D: from photosynthesis, metabolism, and action to clinical applications. In: DeGroot LJ, Jameson JL, eds. *Endocrinology*, 4th edn, pp. 1009–28. WB Saunders, Philadelphia.

Bringhurst FR (2001). Regulation of calcium and phosphate homeostasis. In: DeGroot LJ, Jameson JL, eds. *Endocrinology*, 4th edn, pp. 1029–52. WB Saunders, Philadelphia.

Deftos LJ (1998). *Clinical essentials of calcium and skeletal disorders*, 1st edn. Professional Communications, Oklahoma.

Goltzman D, Cole EC (1999). Hypoparathyroidism. In: Favus MJ, ed. *Primer on the metabolic bone diseases and disorders of mineral metabolism,* 4th edn, pp. 226–30. Lippincott-Raven, Philadelphia.

Indridason OS, Quarles LD (1999). Tertiary hyperparathyroidism and refractory secondary hyperparathyroidism. In: Favus MJ, ed. *Primer on the metabolic bone diseases and disorders of mineral metabolism*, 4th edn, pp. 198–202. Lippincott-Raven, Philadelphia.

Levine MA (1999). Parathyroid hormone resistance syndromes. In: Favus MJ, ed. *Primer on the metabolic bone diseases and disorders of mineral metabolism*, 4th edn, pp. 230–5. Lippincott-Raven, Philadelphia.

Marx SJ (2000). Hyperparathyroid and hypoparathyroid disorders. *New England Journal of Medicine* **343**, 1803–75.

Roberts MM, Stewart AF (1999). Humoral hypercalcaemia of malignancy. In Favus MJ, ed. *Primer on metabolic bone diseases and disorders of mineral metabolism*, 4th edn, pp. 203–7. Lippincott-Raven, Philadelphia.

Shane E (1999). Hypercalcaemia: pathogenesis, clinical manifestations, differential diagnosis and management. In: Favus MJ, ed. *Primer on the metabolic diseases and disorders of mineral metabolism*, 4th edn, pp. 183–7. Lippincott-Raven, Philadelphia.

Shane E (1999). Hypocalcaemia: pathogenesis, differential diagnosis and management. In: Favus MJ, ed. *Primer on the metabolic bone diseases and disorders of mineral metabolism*, 4th edn, pp. 223–6. Lippincott-Raven, Philadelphia.

Stewart AF (1999). Miscellaneous causes of hypercalcaemia. In: Favus MJ, ed. *Primer on the metabolic bone diseases and disorders of mineral metabolism*, 4th edn, pp. 215–19. Lippincott-Raven, Philadelphia.

Thakker RV (1998). Disorders of the calcium sensing receptor. *Biochemica et Biophysica Acta* **1448**, 166–70.

Thakker RV (1998). Multiple endocrine neoplasia–syndromes of the twentieth century. *Journal of Clinical Endocrinology and Metabolism* **83**, 2617–20.

Thakker RV (2000). Parathyroid disorders. Molecular genetics and physiology. In: Morris PJ, Wood WC, eds. *Oxford textbook of surgery*, pp. 1121–9. Oxford University Press, Oxford.

Thakker RV (2001). Multiple endocrine neoplasia type 1. In: DeGroot LJ, Jameson JL, eds. *Endocrinology*, pp. 2503–17. WB Saunders, Philadelphia.

Thakker RV, Juppner H (2001). Genetic disorders of calcium homeostasis caused by abnormal regulation of parathyroid hormone secretion or responsiveness. In: DeGroot LJ, Jameson JL, eds. *Endocrinology*, pp. 1062–74. WB Saunders, Philadelphia.

12.7.1 Disorders of the adrenal cortex

P. M. Stewart

Introduction

Three main types of hormone are produced by the adrenal cortex—glucocorticoids (cortisol, corticosterone), mineralocorticoids (aldosterone, deoxycorticosterone), and sex steroids (mainly androgens). The biochemical pathways involved in their synthesis are shown in Fig. 1. Glucocorticoids are secreted in relatively high amounts (cortisol 10 to 20 mg/day) from the zona fasciculata under the control of ACTH, whilst mineralocorticoids are secreted in low amounts (aldosterone 100 to 150 µg/day) from the zona glomerulosa under the principal control of angiotensin II. Classic endocrine feedback loops are in place to control the secretion of both hormones—cortisol inhibits the secretion of both corticotrophin-releasing factor and ACTH from the hypothalamus and pituitary, respectively, and the aldosterone-induced sodium retention inhibits renal renin secretion.

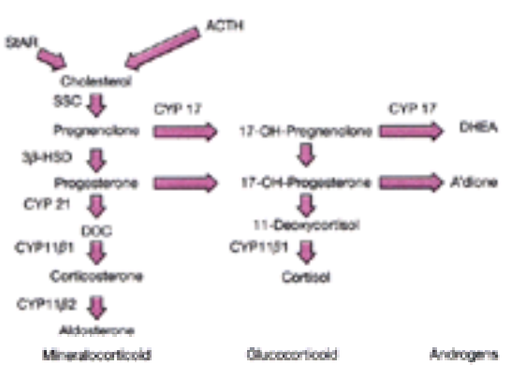

Fig. 1 Pathways of adrenocortical steroid biosynthesis.

Within the adrenal cortex, cholesterol is taken up from circulating cholesterol bound to low-density lipoprotein and an initial, rate-limiting step in adrenal steroidogenesis is the uptake of cholesterol by mitochondria dependent upon a recently characterized protein, steroidogenic acute regulatory protein (or StAR). Thereafter, the functional zonation of the adrenal cortex is achieved in part through the discrete expression and regulation of the final steroidogenic enzymes—aldosterone synthase, expressed in the glomerulosa, and 11b-hydroxylase in the fasciculata (Fig. 1).

Aldosterone acts physiologically to stimulate sodium transport across epithelial cells in the distal nephron, colon, and salivary gland. This involves the interaction of aldosterone with the mineralocorticoid receptor and the induction of the basolateral sodium–potassium ATPase pump and the apical sodium channel. This is mediated through the induction of a novel gene, serum- and glucocorticoid-induced kinase (sgk). The mineralocorticoid receptor, however, is non-selective *in vitro*; paradoxically cortisol and aldosterone have the same intrinsic affinity for this receptor, raising the question as to how aldosterone is the preferred mineralocorticoid *in vivo*. This selectivity is achieved at a prereceptor level through the expression of an enzyme, 11b-hydroxysteroid dehydrogenase type 2 (**11b-HSD2**), which efficiently inactivates cortisol to cortisone allowing aldosterone to occupy the mineralocorticoid receptor. Inhibition of 11b-HSD2 results in cortisol, conventionally regarded as a glucocorticoid, acting as a potent mineralocorticoid.

Glucocorticoids have more diverse and extensive roles than mineralocorticoids, regulating sodium and water homeostasis, glucose and carbohydrate metabolism, inflammation, and stress. These effects are mediated by the interaction of cortisol with the ubiquitous glucocorticoid receptors and the induction or repression of target gene transcription.

Adrenocortical diseases are relatively rare, but part of their importance lies in their relative ease of diagnosis and the availability of effective therapy. The diseases are most readily classified on the basis of whether there is hormone excess or deficiency (Table 1). In most instances this excess or deficiency arises from abnormal secretion of hormones. However, the defect may also relate to a change in corticosteroid metabolism (for example glucocorticoid-suppressible hyperaldosteronism, liquorice ingestion) or to defective receptors (for example glucocorticoid resistance).

Glucocorticoid excess–Cushing's syndrome

Harvey Cushing first described a case of the 'polyglandular syndrome' secondary to pituitary basophilia in 1912 and linked this to bilateral adrenal hyperplasia several years later. The first case of an adrenal adenoma was probably reported by H.G. Turney in 1913 (Fig. 2).

Fig. 2 H. G. Turney's case of Cushing's syndrome before and after developing the condition.

Definition

Cushing's syndrome comprises the symptoms and signs associated with prolonged exposure to inappropriately elevated levels of free plasma glucocorticoid (Fig. 2). This definition thus takes into account the elevated corticosteroid levels that may be found in severely depressed patients but which appear to be appropriate to the condition and also the increased total (but normal free) glucocorticoid levels found when there is an increase in circulating cortisol-binding globulin (for example in patients on oestrogen therapy). The use of the term glucocorticoid in the definition covers both endogenous (cortisol) and exogenous (such as prednisolone, dexamethasone) excess.

Classification of Cushing's syndrome

The condition is most readily classified into ACTH-dependent and ACTH-independent causes (Table 2). The term Cushing's syndrome is used to describe all causes whilst that of Cushing's disease is reserved for cases of pituitary-dependent Cushing's syndrome.

ACTH-dependent causes

Cushing's disease

When iatrogenic causes are excluded, the most frequent cause of Cushing's syndrome is Cushing's disease, which accounts for approximately 70 per cent of cases. The adrenal glands show bilateral adrenocortical hyperplasia with widening of the zona fasciculata and reticularis.

Cushing himself raised the question as to whether his disease was a primary pituitary condition or secondary to an abnormality in the hypothalamus. The release of ACTH from the pituitary is controlled by corticotrophin-releasing factor (**CRF**) acting synergistically with arginine vasopressin. If there was hypothalamic dysfunction in Cushing's disease, it might be expected that one or other of these would be produced in excess, yet measurement of CRF in both the circulation and cerebrospinal fluid has shown that the levels are low. This would suggest that CRF is not involved, but patients with Cushing's disease show an exaggerated ACTH response to CRF, suggesting that there may be an enhanced sensitivity of corticotrophs to this factor. However, *in vitro* experiments with microadenomas from patients with Cushing's disease have not confirmed this. By contrast there is some evidence for enhanced arginine vasopressin production in Cushing's disease and this might interact with CRF to promote tumour growth and ACTH release.

Whether or not the hypothalamus has an initiating role, there is abundant evidence that at presentation the condition is pituitary, rather than hypothalamus, dependent. In over 90 per cent of cases the disease is due to a pituitary adenoma of monoclonal origin; basophil hyperplasia is very uncommon. Selective surgical removal of the microadenoma usually results in cure with a very low recurrence rate.

Ectopic corticotrophin-releasing factor (CRF) production

This is a very rare cause of pituitary-dependent Cushing's disease. However, cases have been described in which a tumour (for example medullary thyroid, prostate carcinoma) has been shown to contain CRF but not ACTH. It has been suggested that ectopic CRF production may explain the metyrapone responsiveness and suppression with high-dose dexamethasone found in some patients with the ectopic ACTH syndrome.

Ectopic ACTH syndrome

Cushing's syndrome may be associated with non-pituitary tumours producing ACTH, most commonly a small-cell carcinoma of the bronchus (Table 3). These conditions are described further in Chapter 12.11.

Macroscopic nodular adrenal hyperplasia

In about 20 to 40 per cent of patients with Cushing's disease there is bilateral adrenocortical hyperplasia associated with one or more nodules. These may be up to several centimetres in diameter. Such nodules are a trap for the unwary (see below) as they may be mistaken for primary adrenal tumours.

ACTH-independent causes

Adrenal adenoma and carcinoma

With the exclusion of iatrogenic Cushing's syndrome, adrenal adenomas are responsible for about 10 per cent of cases and carcinomas for about the same. Carcinomas are the most common cause of Cushing's syndrome in children. The aetiology of these tumours is unknown.

Carney's syndrome

This is an autosomal dominant condition comprising mesenchymal tumours (especially atrial myxomas), spotty skin pigmentation, peripheral nerve tumours, and various endocrine tumours, one of which may lead to Cushing's syndrome. The adrenals then contain multiple, small, pigmented nodules. The condition has been described as pigmented multinodular adrenocortical dysplasia. It does not appear to be ACTH dependent and recent evidence suggests that the condition results because of mutations in the regulatory subunit R1A of protein kinase A.

McCune–Albright syndrome

In this condition fibrous dysplasia and cutaneous pigmentation may be associated with pituitary, thyroid, adrenal, and gonadal hyperfunction. The adrenal hypersecretion may produce Cushing's syndrome. The underlying abnormality is a somatic mutation in the a-subunit of the stimulatory G protein which is linked to adenyl cyclase. The mutation results in the G protein being constitutively activated (that is, in the adrenal mimics constant ACTH stimulation). Adrenal nodular formation may occur.

Aberrant receptor expression

Recently patients have been described with nodular hyperplasia, ACTH-independent Cushing's syndrome, and enhanced adrenal responsiveness to gastric inhibitory polypeptide (**GIP**). The biochemical clues were the presence of subnormal morning levels of plasma cortisol and a rise in cortisol after food. This food-dependent form of Cushing's syndrome resulted from the normal increase in GIP after eating. The adrenocortical tissue of these patients responded *in vitro* to low doses of GIP,

whereas there was no such effect in normal adrenal cortex, suggesting that in some unknown manner adrenal GIP receptors are linked to steroidogenesis in these patients. Not surprisingly, the clinical syndrome is related to food intake. Fasting can produce adrenal insufficiency. It remains to be seen whether abnormalities of adrenal sensitivity to GIP play a subtle role in other types of Cushing's syndrome. Similarly Cushing's syndrome due to a cortisol-secreting adrenal adenoma has recently been attributed to aberrant expression of receptors for interleukin 1.

Alcohol-associated pseudo-Cushing's syndrome

In the original description of this syndrome, urinary and plasma cortisol levels were elevated and were not suppressed with dexamethasone. Plasma ACTH has been found to be normal or suppressed. The frequency and pathogenesis of this condition remain unknown but a 'two-hit' hypothesis has been put forward to explain its aetiology. Chronic liver disease irrespective of the cause is associated with impaired cortisol metabolism, but in alcoholics this is associated with an increase in cortisol secretion rate, rather than concomitant suppression in the face of impaired metabolism. With abstinence from alcohol the biochemical abnormalities rapidly revert to normal.

Clinical features of Cushing's syndrome

The classic features of Cushing's syndrome with centripetal obesity, moon face, hirsutism, and plethora are well known following Cushing's initial description in 1912 (Fig. 2 and Fig. 3). However, this gross clinical picture is not always present. The signs and symptoms in patients with Cushing's syndrome are listed in Table 4 together with the most discriminatory features distinguishing Cushing's from simple obesity. Weight gain and obesity were the commonest symptom and sign, but the distribution of fat was not invariably centripetal; a 'buffalo hump' was present in about half the patients.

Fig. 3 Typical facies of a patient with Cushing's syndrome before and after treatment.

Gonadal dysfunction is very common, with menstrual irregularity in females and loss of libido in males. Hirsutism is frequently found in female patients, as is acne.

Psychiatric abnormalities have been reported in all series of patients with Cushing's syndrome regardless of cause. Depression and lethargy are among the commonest problems, but poor concentration, paranoia, and overt psychosis are also well recognized. Lowering of plasma cortisol by medical or surgical therapy usually results in a rapid improvement in the psychiatric state.

Most patients with long-standing Cushing's syndrome have lost height because of osteoporotic vertebral collapse. This can be assessed by measuring the patient's height and comparing it with their span; in normal subjects these measurements should be equal. Pathological fractures, either spontaneous or after minor trauma, are not uncommon. Rib fractures, in contrast to those of the vertebrae, are often painless. The radiograph appearances are typical, with exuberant callus formation at the site of the healing fracture.

The plethoric appearance of the patient with Cushing's syndrome is caused by thinning of the skin and is not due to true polycythaemia. In those with a high concentration of haemoglobin, the red cell mass is usually normal and the polycythaemia due to a reduced plasma volume.

The typical red-purple livid striae of the syndrome are found most frequently on the abdomen but may also be present on the upper thighs and arms. They are very common in younger patients and less so in those over 50.

Myopathy and bruising are two of the most discriminatory features of the syndrome. The myopathy involves the proximal muscles of lower limb and shoulder girdle. Complaints of weakness such as inability to climb stairs or get up from a deep chair are relatively uncommon, but observation of whether the patient can rise from a crouching position often reveals the problem. Bruising of the skin is often extensive and occurs with unknown or trivial trauma.

Hypertension is another prominent feature; even though epidemiological data show a strong association between blood pressure and obesity, hypertension is much more common in patients with Cushing's syndrome than in those with simple obesity.

Pigmentation is rare in Cushing's disease but common in the ectopic ACTH syndrome. However, in some pituitary tumours there is abnormal processing of the pro-opiomelanocortin (POMC) precursor molecule, with resulting pigmentation.

Infections are more common in patients with Cushing's syndrome. In many instances these are asymptomatic as the normal inflammatory response may be suppressed. Reactivation of tuberculosis has been reported. In the skin, fungal infection is frequently found. Glucose intolerance may be a predisposing factor, with overt diabetes being present in up to one-third of patients in some series.

Ocular effects may include raised intraocular pressure, chemosis, and exophthalmos (present in up to one-third of patients in Cushing's original series). Cataracts, a well-recognized complication of corticosteroid therapy, seem to be uncommon, except as a complication of diabetes.

Special features of Cushing's syndrome

Cyclical cushing's syndrome

Of particular clinical interest has been a group of patients with cyclical Cushing's syndrome, characterized by periods of excess cortisol production (for example, 40 days) followed by intervals of normal cortisol production (for example, 60 to 70 days). Some of these patients demonstrate a paradoxical rise in plasma ACTH and cortisol when treated with dexamethasone, and occasional patients show benefit with dopamine agonist (bromocriptine) or serotonin antagonist (cyproheptadine) therapy. Most patients have been thought to have pituitary-dependent disease and in many of these patients basophil adenomas have been removed, some with long-term cure. However, cortisol secretion may show some evidence of cyclicity in other causes of Cushing's syndrome, notably the ectopic ACTH syndrome.

Children

In children all the above features occur, but growth arrest is almost invariable. The dissociation between height and weight on the growth chart is obvious. It is important to try to obtain previous growth data so as to be able to calculate growth velocity. If the patient is growing along the same centile line, then the diagnosis of Cushing's syndrome is highly unlikely. In addition to glucocorticoid-induced growth arrest, androgen excess may result in precocious puberty.

Pregnancy

Pregnancy is rare in women with Cushing's syndrome because of associated amenorrhoea due to androgen excess or hypercortisolism. However, approximately 100

such cases have been reported, 50 per cent of which are due to adrenal adenomas. A few cases of true pregnancy-induced Cushing's syndrome have been reported with regression postpartum. In these cases the aetiology is unknown. Establishing a diagnosis and cause can be difficult; normal pregnancy is associated with a threefold increase in plasma cortisol due to increased production rates and increases in cortisol-binding globulin. Urinary free cortisol also rises and dexamethasone does not suppress plasma cortisol to the same degree as the non-pregnant state. Untreated the condition has a high maternal and fetal morbidity and mortality. Adrenal and/or pituitary adenomas should be excised. Metyrapone, which is not teratogenic, has been effective in many cases in controlling the hypercortisolism.

Adrenal carcinomas

In addition to the normal features resulting from glucocorticoid excess, the patient may present with other problems relating to (i) the tumour, for instance abdominal pain from the primary tumour or with secondary deposits, or (ii) the secretion of other steroids, such as androgens or mineralocorticoids. Thus, in females, in addition to hirsutism, there may be other features of virilization, with clitoromegaly, breast atrophy, deepening of the voice, temporal recession, and severe acne.

Ectopic ACTH syndrome

If this is due to a small-cell lung carcinoma, the clinical presentation more commonly resembles Addison's disease than Cushing's syndrome. The patients are very commonly pigmented and have lost weight, but the association of this with hypokalaemic alkalosis and glucose intolerance should alert the clinician. Patients with benign tumours, such as bronchial carcinoids which produce ACTH, present with the typical features of Cushing's syndrome.

Investigation of patients with suspected Cushing's syndrome

There are two stages in the investigation of a patient with suspected Cushing's syndrome: (1) Does the patient have Cushing's syndrome? (2) If the answer to (1) is yes, then what is the cause? Unfortunately many investigators fail to make this distinction and ill-advisedly use tests that are relevant to question (2) to try to answer question (1). In particular it is essential that radiological investigations are not undertaken until Cushing's syndrome has been confirmed biochemically. The principal diagnostic tests are listed in Table 5.

Diagnostic tests

Circadian rhythm of plasma cortisol

In normal subjects, plasma cortisol concentrations are at their highest first thing in the morning and reach a nadir at around midnight (less than 100 nmol/l). This circadian rhythm is lost in patients with Cushing's syndrome such that in the majority of patients the 09.00 h level of plasma cortisol is normal but nocturnal levels are raised. Random morning levels of plasma cortisol are therefore of little value in making the diagnosis. In addition, various factors such as stress of venepuncture, intercurrent illness, and admission to hospital may result in normal subjects losing their circadian rhythm. It is therefore good practice not to measure plasma cortisol until the patient has been in hospital for 48 h.

Very few laboratories have developed methods for the measurement of free levels of plasma cortisol. As more than 90 per cent of plasma cortisol is protein bound, the results of the conventional assay will be affected by drugs or conditions which alter levels of cortisol-binding globulin. Thus oestrogen therapy or pregnancy may elevate cortisol-binding globulin and hence total plasma cortisol. In practice, therefore, circadian rhythm is not a widely used screening test.

Urinary free cortisol excretion

For many years the diagnosis of Cushing's syndrome was based on the measurement of urinary metabolites of cortisol (24-h urinary 17-hydroxy-corticosteroid or 17-oxogenic steroid excretion, depending on the method used). However, the sensitivity and specificity of these methods is poor and most investigators have replaced these assays with the much more sensitive measurement of urinary free cortisol excretion. Urinary free cortisol is an integrated measure of plasma free cortisol. As cortisol secretion increases, the binding capacity of cortisol-binding globulin is exceeded and results in a disproportionate rise in urinary free cortisol. This is a useful screening test, but even so, it is accepted that urinary free cortisol may be normal in up to 8 to 15 per cent of patients with Cushing's syndrome.

Measurement of the cortisol–creatinine ratio on the first urine specimen passed on waking obviates the need for a timed collection and has been used by some as a sensitive screening test, particularly if cyclical Cushing's syndrome is suspected. Urine aliquots are stable when left at room temperature for up to 7 days and can then be sent by post to the local endocrine laboratory.

Low-dose/overnight dexamethasone suppression tests

In normal subjects, administration of a supraphysiological dose of glucocorticoid results in suppression of ACTH and hence of cortisol secretion. In Cushing's syndrome of whatever cause there is a failure of this suppression when low doses of the synthetic glucocorticoid dexamethasone are given.

The overnight test is often used as an outpatient screening test. Various doses of dexamethasone have been used, usually given at midnight. A normal response is a plasma cortisol of less than 50 nmol/l between 08.00 and 09.00 h the following morning. A dose of 1.5 or 2 mg gives a 30 per cent false-positive rate, whereas after 1 mg this is reduced to 12.5 per cent with a false-negative rate of less than 2 per cent. Thus, the outpatient overnight test has high sensitivity but low specificity, and further investigation is often required.

In the 48-h test, plasma cortisol is measured at 09.00 h on day 0 and 48 h later following dexamethasone given in a dose of 0.5 mg every 6 h for 48 h. This test is reported as having a 97 to 100 per cent true-positive rate and a false-positive of less than 1 per cent.

Certain drugs (phenytoin, rifampicin) may increase the metabolic clearance rate of dexamethasone thereby giving false-positive results.

Insulin tolerance test

Patients with severe depression may show many of the biochemical features of Cushing's syndrome (loss of circadian rhythm of plasma cortisol, increased urinary free cortisol, failure of cortisol suppression with low-dose dexamethasone). Patients with Cushing's syndrome are also frequently depressed. It is thus important in a depressed patient to take particular care in distinguishing the two conditions. In normal subjects and patients with severe endogenous depression, insulin-induced hypoglycaemia results in a rise in ACTH and cortisol levels, a response which does not usually occur in Cushing's syndrome.

Differential diagnostic tests

Once the biochemical diagnosis has been made, a series of investigations is required to determine the cause of the Cushing's syndrome.

Plasma ACTH at 09.00h

This will differentiate ACTH-dependent from ACTH-independent causes. ACTH is either within the normal reference range (50 per cent of cases) or elevated in patients with Cushing's disease. ACTH levels in the ectopic ACTH syndrome are high but overlap values seen in Cushing's disease in 30 per cent of cases and cannot therefore be used to differentiate these two conditions (Fig. 4). The measurement of ACTH precursors (pro-ACTH, POMC) is not routinely available but may be more useful in detecting an ectopic source of ACTH.

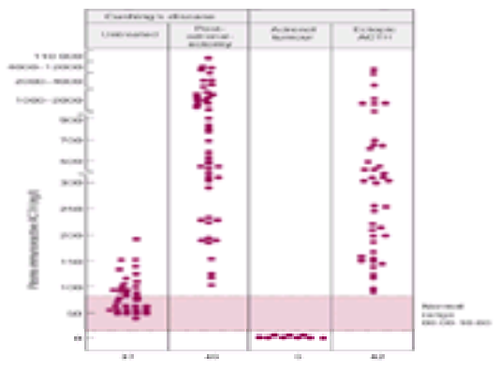

Fig. 4 Immunoreactive N-terminal ACTH levels in plasma samples taken between 08.00 and 10.00 h in normal subjects (hatched area), and patients with Cushing's disease either untreated or postadrenalectomy, patients with adrenal tumours, and in the ectopic ACTH syndrome (by courtesy of Professor Lesley Rees).

In patients with adrenal tumours, plasma ACTH is invariably undetectable. This can also occur with degradation of ACTH; consequently, non-haemolysed blood samples should be taken on ice and immediately separated.

Diagnosis is a problem in those patients whose plasma ACTH levels are low normal or intermittently detectable. This may occur in macronodular hyperplasia. The danger is that in some patients the asymmetry of the nodular hyperplasia may lead to a diagnosis of adrenal adenoma, the plasma ACTH is ignored, and an inappropriate adrenalectomy is performed. Conversely, in some patients with this syndrome an autonomous adrenal tumour develops and, despite detectable ACTH, unilateral adrenalectomy is required.

Plasma potassium (see also Chapter 20.2.2)

Hypokalaemic alkalosis is present in more than 95 per cent of patients with the ectopic ACTH syndrome, but is present in fewer than 10 per cent of patients with Cushing's disease. The aetiology of this is now becoming clearer. Patients with the ectopic syndrome usually have higher cortisol secretion rates that saturate the renal protective 11b-hydroxysteroid dehydrogenase type 2 enzyme resulting in cortisol-induced, mineralocorticoid hypertension (see apparent mineralocorticoid excess syndrome, below). In addition, these patients have higher levels of the ACTH-dependent mineralocorticoid, deoxycorticosterone.

High-dose dexamethasone suppression test

The rationale for this test is that in Cushing's disease there is negative feedback control of ACTH but it is set at a higher level than normal. Thus in this disease cortisol levels are not suppressed with a low dose of dexamethasone but are with a high dose. The original test introduced by Liddle was based on giving dexamethasone at a dose of 2 mg every 6 h for 48 h and measuring urinary 17-oxogenic steroids. Suppression was defined as a greater than 50 per cent fall in 24-h urinary 17-oxogenic steroids. In the modern test, plasma cortisol is measured at 0 and after 48 h or, less commonly, 8 mg of dexamethasone is given orally at 23.00 h and plasma cortisol taken at 08.00 h on the same day (basal sample) and at 08.00 h on the following morning. In both these tests greater than 50 per cent suppression of plasma cortisol in comparison with the basal sample has been used to define a positive response. In Cushing's disease about 90 per cent of patients have a positive 48-h test in comparison with 10 per cent with the ectopic ACTH syndrome. With overnight high-dose testing, 89 per cent sensitivity and 100 per cent specificity has been reported for Cushing's disease. A further variation on this test is the 5-h infusion of dexamethasone (1 mg/h).

Metyrapone test

Metyrapone is an 11b-hydroxylase inhibitor which blocks the conversion of 11-deoxycortisol to cortisol and deoxycorticosterone to corticosterone (Fig. 1). This lowers plasma cortisol and, via negative feedback control, increases plasma ACTH. This in turn stimulates an increase in the secretion of adrenal steroids proximal to the block. Given in doses of 750 mg every 4 h for 24 h, patients with Cushing's disease exhibit an exaggerated rise in plasma ACTH with 11-deoxycortisol levels at 24 h exceeding 1000 nmol/l. In most patients with the ectopic ACTH syndrome there is little or no response, but occasional patients (possibly those producing both ACTH and CRF) have an 11-deoxycortisol response which may be similar to that in Cushing's disease.

The metyrapone test was originally used to distinguish patients with Cushing's disease from those with a primary adrenal cause. However, these can be more reliably distinguished by measuring plasma ACTH and CT scanning of the adrenals. As indicated, the test does not reliably distinguish between Cushing's disease and the ectopic ACTH syndrome and the value of this test has been questioned. It should be reserved for patients when the results of other tests are equivocal.

Corticotrophin-releasing factor (CRF) test

CRF is a 41 amino acid peptide, identified by Vale in 1981 from ovine hypothalami. The ovine sequence differs by seven amino acid residues from that of the human, but despite this, stimulates the release of ACTH in humans. The test involves the intravenous injection of either ovine or human CRF in a dose of 1 µg/kg body weight or a single dose of 100 µg. The test can be performed in the morning or afternoon, and after basal sampling, blood samples for ACTH and cortisol are taken every 15 min for 1 to 2 h after administering CRF.

In normal subjects, CRF produces a rise in ACTH and cortisol, and this response is exaggerated in Cushing's disease. It is typically absent in the ectopic ACTH syndrome and in patients with adrenal tumours. In distinguishing pituitary-dependent Cushing's from the ectopic ACTH syndrome the response of ACTH to CRF has a specificity of 90 per cent, and with cortisol as the end-point, 95 per cent. Using an ACTH increase of 100 per cent over basal or a cortisol rise of 50 per cent as an end-point, this positive response eliminates a possible diagnosis of the ectopic ACTH syndrome.

As with the other tests, patients with macronodular hyperplasia may present a problem in diagnosis and show no response to CRF. The test is valuable in distinguishing patients who are obese and depressed from those with Cushing's disease; in the former the CRF response is either normal or reduced.

Inferior petrosal sinus sampling/selective venous catheterization

To distinguish Cushing's disease from the ectopic ACTH syndrome it may be necessary to identify the source of ACTH secretion. As blood from each half of the pituitary drains into the ipsilateral inferior petrosal sinus, catheterization of both sinuses with simultaneous sampling of venous blood can distinguish a pituitary from an ectopic source and aid in the lateralization of a pituitary microadenoma (Fig. 5). In patients with the ectopic ACTH syndrome, there is usually no ACTH gradient between the inferior petrosal sinus samples and simultaneously drawn peripheral venous levels. In Cushing's disease the ipsilateral:contralateral ACTH ratio is usually greater than 1.4. However, because of the problem of intermittent ACTH secretion, it is useful to make measurements before and at intervals (for example, 2, 5, and 15 min) after intravenous injection of 100 µg of synthetic ovine CRF. Using this approach, patients with Cushing's disease and bilateral inferior petrosal sinus ratios of less than 1.4 can be readily distinguished from those with the ectopic syndrome. The precise ratio that distinguishes Cushing's disease from the ectopic syndrome has been debated. Some authors use 2 rather than 1.4.

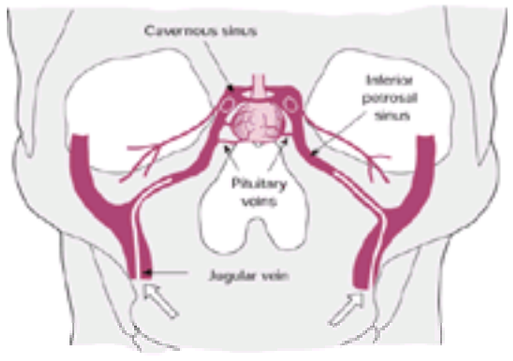

Fig. 5 Positions of bilateral catheters in inferior petrosal sinus sampling.

It is clear that inferior petrosal sinus sampling is a useful technique for making the differential diagnosis in ACTH-dependent Cushing's syndrome. However, it should be reserved for those cases where the differential diagnosis is still in doubt after high-dose dexamethasone and peripheral CRF testing. Inferior petrosal sinus catheterization may also be of value to the surgeon who is planning to explore the pituitary in a patient with Cushing's disease in whom imaging techniques have failed to demonstrate a microadenoma.

Rarely, selective catheterization of vascular beds may be required to identify the source of ectopic ACTH secretion, for example from a small pulmonary carcinoid or thymic tumour.

Tumour markers

Many tumours responsible for the ectopic ACTH syndrome also produce peptide hormones other than ACTH or its precursors.

Imaging

MRI/CT scanning of pituitary and adrenals

There is no doubt that high-resolution, thin-section, contrast-enhanced imaging using either CT or MRI has revolutionized the investigation of Cushing's syndrome. However, it is essential that the results of any imaging technique must always be interpreted in the light of the biochemical results if mistakes are to be avoided. In imaging the adrenals, asymmetrical nodular hyperplasia may lead to a false diagnosis of adrenal adenoma (Fig. 6). Owing to the presence of 'pituitary incidentalomas', pituitary MRI/CT scanning may produce false-positive results, particularly for lesions of less than 5 mm in diameter.

Fig. 6 CT scan of adrenals in patient with asymmetrical nodular hyperplasia. The macronodule on the left was initially thought to be an adrenal tumour. The biochemistry indicating ACTH-dependent Cushing's syndrome was ignored and a unilateral adrenalectomy performed without cure of the hypercortisolism. Further investigation confirmed Cushing's disease and a selective pituitary microadenomectomy resulted in cure.

Pituitary MRI is the investigation of choice if the biochemical tests suggest Cushing's disease, with a sensitivity of 70 per cent and specificity of 87 per cent (Fig. 7 and Fig. 8). About 90 per cent of ACTH-secreting pituitary tumours are microadenomas (that is, less than 10 mm in diameter). The classic features of a pituitary microadenoma are a hypodense lesion after contrast, associated with deviation of the pituitary stalk and a convex upper surface of the pituitary gland (Fig. 7). With such small tumours it is not surprising that the sensitivity of CT scanning is relatively low (20 to 60 per cent) with a similar specificity.

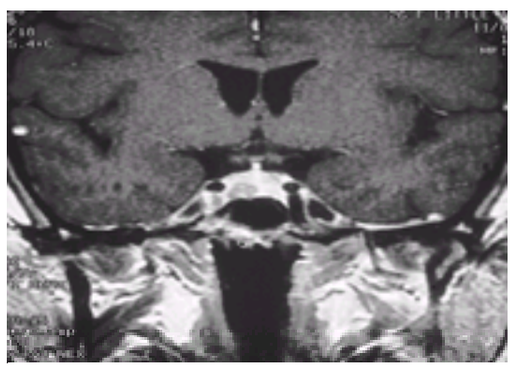

Fig. 7 MRI scan of pituitary demonstrating the typical appearance of a pituitary microadenoma. A hypodense lesion is seen in the left side of the gland with deviation of the pituitary stalk away from the lesion. Following a biochemical diagnosis of Cushing's disease, this patient was cured following trans-sphenoidal hypophysectomy.

Fig. 8 MRI scan of the pituitary gland demonstrating a large macroadenoma in a patient with Cushing's disease. In contrast to smaller tumours, these tumours are

invariably invasive and recur following surgery.

By contrast, for adrenal imaging, CT scanning rather than MRI is the investigation of choice offering better spatial resolution (Fig. 9). Once again, it is stressed that 'adrenal incidentalomas' are present in up to 5 per cent of normal subjects, and thus adrenal imaging should not be performed unless biochemical investigation suggests a primary adrenal cause. Adrenal carcinomas are large and often associated with metastatic spread at presentation (Fig. 10).

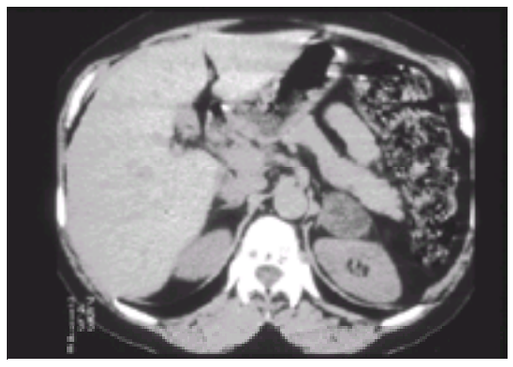

Fig. 9 Typical solitary left-sided adrenal adenoma on adrenal CT scanning.

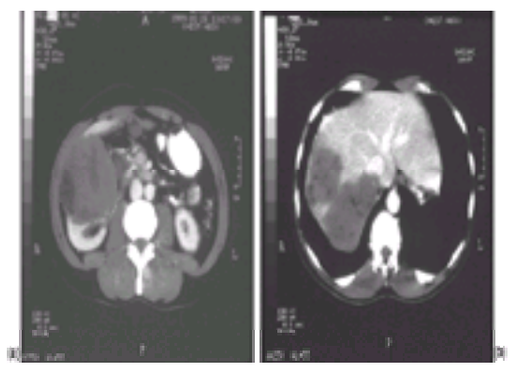

Fig. 10 CT scan of a patient with rapidly progressing Cushing's syndrome due to a right-sided adrenal carcinoma. An irregular right adrenal mass is shown (a) with a large liver metastasis (b).

In patients with 'occult' ectopic ACTH syndrome, high-definition MRI/CT scanning with images every 0.5 cm may be required to detect small ACTH-secreting carcinoid tumours.

Adrenal scintigraphy

This is of value in certain patients with primary adrenal pathology. The most commonly used agent is ^{131}I-6b-iodomethyl-19-norcholesterol. This is a marker of adrenocortical cholesterol uptake. In patients with adrenal adenomas the isotope is taken up by the adenoma but not by the contralateral suppressed adrenal. Adrenal scintigraphy is useful in patients with suspected adrenocortical macronodular hyperplasia, in which CT scanning may be misleading by suggesting unilateral pathology, whereas with isotope scanning the bilateral adrenal involvement is identified (Fig. 11).

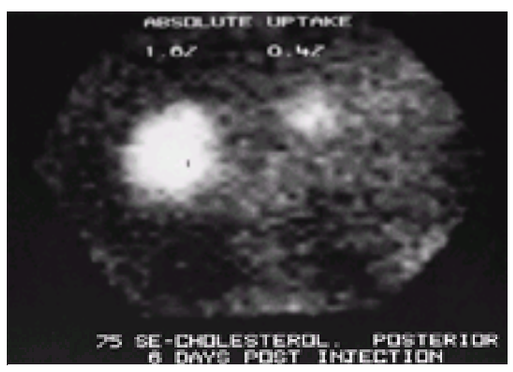

Fig. 11 Adrenal scintigraphy in a patient with Cushing's syndrome and macronodular hyperplasia. Note asymmetrical uptake in the adrenals, with 1.6 per cent uptake on the left and 0.4 per cent on the right.

Prognosis of untreated Cushing's syndrome

Studies carried out prior to the introduction of effective therapy suggested that 50 per cent of patients with untreated Cushing's syndrome died within 5 years. Even with modern management, an increased prevalence of cardiovascular risk factors persists for many years after an apparent 'cure'. Close follow-up of all patients is recommended.

Treatment of Cushing's syndrome

Adrenal causes

Adrenal adenomas should be removed by unilateral adrenalectomy, with 100 per cent cure rate. With the increasing experience of laparascopic adrenalectomy in most tertiary centres, this has now become the surgical treatment of choice for unilateral tumours, reducing surgical morbidity and postoperative hospital stay compared with traditional open approaches. After surgery it may take many months or even years for the suppressed adrenal to recover. It is wise therefore to give slightly suboptimal replacement therapy with dexamethasone at a dose of 0.5 mg in the morning, with intermittent measurement of the 08.00 h level of plasma cortisol prior to taking dexamethasone. When the morning plasma cortisol is above 180 nmol/l, dexamethasone can be stopped. A subsequent insulin tolerance test may then demonstrate whether the response to stress is normal.

Adrenal carcinomas have a very poor prognosis and most patients are dead within 2 years. It is usual practice to try to remove the primary tumour, even though metastases may be present, so as to enhance the response to the adrenolytic agent *o*,*p*'-DDD (Mitotane, see below). Radiotherapy to the tumour bed and to some

metastases, such as those in the spine, may be of limited value.

Pituitary-dependent Cushing's disease

The treatment of Cushing's disease has been improved by trans-sphenoidal surgery conducted by an experienced surgeon. Before the selective removal of a pituitary microadenoma the treatment of choice was bilateral adrenalectomy. This had an appreciable mortality even in the best centres (about 4 per cent) as well as morbidity. The main risk was the subsequent development of Nelson's syndrome (postadrenalectomy hyperpigmentation with locally aggressive pituitary tumour) (Fig. 12 and Plate 1). To avoid this, pituitary irradiation was often carried out following bilateral adrenalectomy. These patients required lifelong replacement therapy with hydrocortisone and fludrocortisone. Nowadays bilateral adrenalectomy is reserved for the occasional patient with Cushing's disease in whom no pituitary tumour can be found, or when pituitary surgery has failed, or where the condition has recurred.

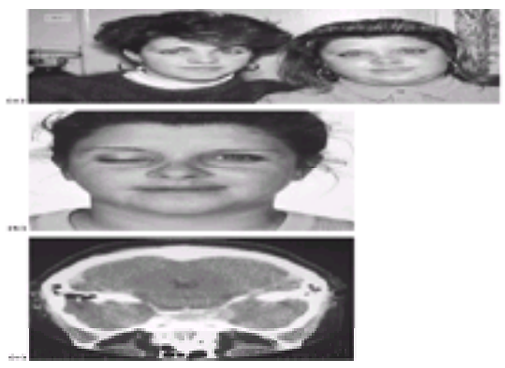

Fig. 12 A young woman with Cushing's disease, photographed initially alongside her identical twin sister (a). In this case treatment with bilateral adrenalectomy was undertaken and several years later the patient re-presents with Nelson's syndrome and a right III cranial nerve palsy due to cavernous sinus infiltration from a locally invasive corticotrophinoma. (See also Plate 1.)

After selective removal of a microadenoma, the surrounding corticotrophs are normally suppressed (Fig. 13). In these cases plasma cortisol concentrations are also suppressed postoperatively and glucocorticoid replacement therapy is required. Using the dexamethasone regime described above after removal of an adrenal adenoma, there is usually (but not invariably) gradual recovery of the hypothalamic–pituitary–adrenal axis (Fig. 14). A non-suppressed plasma cortisol postoperatively suggests that the patient is not 'cured', even though cortisol secretion may have fallen to normal or subnormal values. Close follow-up of such individuals is required.

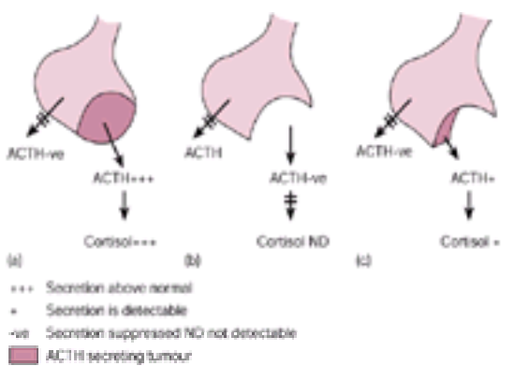

Fig. 13 Selective removal of a microadenoma and its effect on the hypothalamic–pituitary–adrenal axis. Because the surrounding normal pituitary corticotrophs are suppressed in a patient with an ACTH-secreting pituitary adenoma, successful removal of the tumour results in ACTH and hence adrenocortical deficiency with an undetectable (less than 50 nmol/l) level of plasma cortisol. A plasma cortisol of more than 50 nmol/l postoperatively implies that the patient is not cured. (Figure by courtesy of Dr Peter Trainer.)

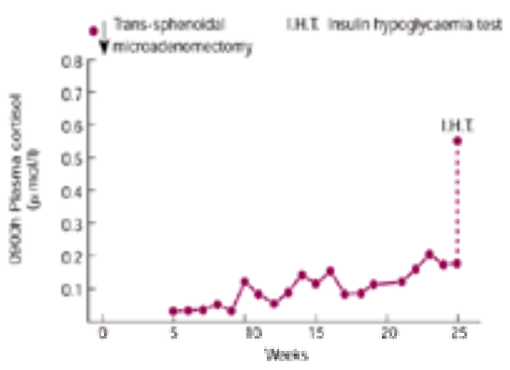

Fig. 14 Gradual recovery of function of the hypothalamic–pituitary–adrenal axis after removal of a pituitary ACTH-secreting microadenoma. The insulin hypoglycaemia test eventually demonstrated the return of a normal stress response.

In the past, pituitary irradiation was often used in the treatment of Cushing's disease. However, the improvements in pituitary surgery have resulted in far fewer patients being so treated. In children pituitary irradiation appears to be effective. Radiotherapy is not recommended as a primary treatment but is reserved for patients not responding to pituitary microsurgery or when bilateral adrenalectomy has been performed, or in those with established Nelson's syndrome.

Ectopic ACTH syndrome

Treatment of the ectopic ACTH syndrome depends on the cause. If the tumour can be found and has not spread, then its removal can lead to cure (for example bronchial carcinoid or thymoma). However, the prognosis for small-cell lung cancer associated with the ectopic ACTH syndrome is poor. The cortisol excess and associated hypokalaemic alkalosis and diabetes mellitus can be ameliorated by medical therapy (see below). Treatment of the small-cell tumour itself will also, at least initially, produce improvement (see Chapter 17.13). Sometimes, if the ectopic source of ACTH cannot be found, it may be necessary to perform bilateral adrenalectomy and then follow the patient carefully (sometimes for several years) to find the primary tumour.

Medical treatment of Cushing's syndrome

Several drugs have been used in the treatment of Cushing's syndrome. Their site of action is shown in Fig. 15. Most commonly, metyrapone has been given, often to lower cortisol concentrations prior to definitive therapy, or while awaiting benefit from pituitary irradiation. The daily dose has to be determined by measuring either plasma or urinary free cortisol. The aim should be to achieve a mean plasma cortisol of about 300 nmol/l during the day or a normal urinary free cortisol. The drug is usually given in doses ranging from 250 mg twice daily to 1.5 g every 6 h. Nausea may be produced and can be alleviated (if not due to adrenal insufficiency) by

giving the drug with milk.

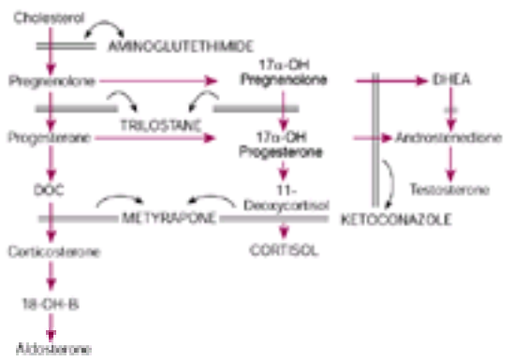

Fig. 15 Medical treatment of Cushing's syndrome: site of action of various drugs.

Aminoglutethimide is a more toxic drug which in high dose blocks initial steps in the biosynthetic pathway and thus affects the secretion of steroids other than cortisol. In doses of 1.5 to 3 g daily (start with 250 mg every 8 h) it commonly produces nausea, marked lethargy, and a skin rash.

Trilostane, a 3b-hydroxysteroid dehydrogenase inhibitor, is ineffective in Cushing's disease, since the block in steroidogenesis is overcome by the rise in ACTH. However, it can be effective in patients with adrenal adenomas.

Ketoconazole is an imidazole which has been widely used as an antifungal agent; it produces abnormal liver function tests signifying hepatitis in about 14 per cent of patients. Ketoconazole blocks a variety of steroidogenic cytochrome P450-dependent enzymes and thus lowers plasma cortisol levels. For effective control of Cushing's syndrome, 400 to 800 mg of ketoconazole daily have been required.

o,*p*'-DDD, or Mitotane, is an adrenolytic drug which is taken up by both normal and malignant adrenal tissue causing adrenal atrophy and necrosis. Because of its toxicity, Mitotane has been used mainly in the management of adrenal carcinoma. Doses of up to 8 g/day are required to control glucocorticoid excess, though evidence that it causes tumour shrinkage or improves long-term survival is scant. The drug will also produce mineralocorticoid deficiency and both glucocorticoid and mineralocorticoid replacement therapy may be required. Side-effects are common and include fatigue, skin rashes, and gastrointestinal disturbance.

Glucocorticoid deficiency—primary and secondary hypoadrenalism

Primary hypoadrenalism refers to glucocorticoid deficiency occurring in the setting of adrenal disease, whilst secondary hypoadrenalism arises because of deficiency of ACTH, the major trophic hormone controlling cortisol secretion. The principal distinction between these two conditions is that mineralocorticoid deficiency invariably accompanies primary hypoadrenalism but this does not occur in secondary hypoadrenalism because only ACTH is deficient; the renin–angiotensin–aldosterone axis is intact.

Primary hypoadrenalism

Congenital adrenal hyperplasia

Various inherited enzyme defects in the synthesis of adrenocortical hormones have been identified and this group of conditions is addressed in Chapter 12.7.2.

Addison's disease

Thomas Addison described this condition in his classic monograph published in 1855. Addison worked with Bateman, a dermatologist who produced one of the first classifications of skin disease. It seems likely that this stimulated Addison's interest in the skin pigmentation which is so characteristic of this disease.

Aetiology

This is a rare condition with an estimated incidence in the developed world of 0.8 cases per 100 000 population. The causes of Addison's disease are listed in Table 6.

Worldwide, infectious diseases are the most common cause of primary adrenal insufficiency. Leading causes include tuberculosis, fungal infections (histoplasmosis, crytococcus), and cytomegalovirus. Adrenal failure may occur in the acquired immunodeficiency syndrome. In tuberculous Addison's disease, the adrenals are initially enlarged with extensive epithelioid granulomas and caseation. Calcification eventually ensues in most cases (Fig. 16). Both the cortex and the medulla are affected.

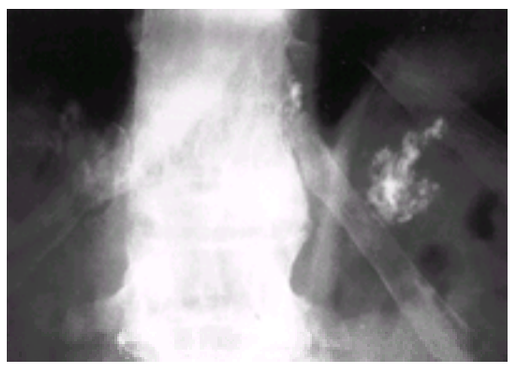

Fig. 16 Plain radiograph of the abdomen showing adrenal calcification in a patient with tuberculous Addison's disease.

In the Western world, autoimmune adrenalitis accounts for over 70 per cent of all cases of Addison's disease. Pathologically, the adrenal glands are atrophic with loss of most of the cortical cells, but the medulla is usually intact. Adrenal autoantibodies can be detected in up to 75 per cent of newly diagnosed cases and have recently been characterized. The major autoantigen is the adrenal enzyme, 21-hydroxylase, but antibodies directed against cholesterol side-chain cleavage and 17a-hydroxylase may rarely be detected. Fifty per cent of patients with Addison's disease have an associated autoimmune disease and these 'polyglandular autoimmune syndromes' have been classified into two distinct variants. Type I is inherited as an autosomal recessive condition and comprises Addison's disease, chronic mucocutaneous candidiasis, and hypoparathyroidism. The condition is rare and usually presents in childhood with either candidiasis or hypoparathyroidism. Other autoimmune conditions such as pernicious anaemia, thyroid disease, chronic active hepatitis, and gondal failure may occur but are rare. Autoantibodies to 21-hydroxylase are not normally present.

Type II polyglandular autoimmune syndrome is more common and may comprise Addison's disease, autoimmune thyroid disease, diabetes mellitus, and hypogonadism. The condition has an inherited basis with linkage to the HLA major histocompatibility complex, notably HLA DR3 and DR4. Autoantibodies to

21-hydroxylase are usually present and are predictive for the development of adrenal destruction.

With the exception of tuberculosis and autoimmune adrenal failure, other causes of Addison's disease are rare (Table 6). Adrenal metastases (commonest primary lung, breast) are often found at postmortem examinations but adrenal insufficiency from these is uncommon. Necrosis of the adrenals due to intra-adrenal haemorrhage should be considered in any severely sick patient and this may be due to infection, trauma, or hypercoagulability. Intra-adrenal bleeding may be found in any cause of severe septicaemia, particularly in children. When this is due to meningococcus, the association with adrenal insufficiency is known as the Waterhouse–Friderichsen syndrome. Adrenal replacement leading to glandular failure may also occur with amyloidosis and haemochromatosis. Adrenal hypoplasia congenita is an X-linked disorder comprising congenital adrenal insufficiency and hypogonadotrophic hypogonadism. The condition is caused by mutations in the DAX-1 gene, a known member of the nuclear receptor family which is expressed in the adrenal cortex, gonads, and hypothalamus.

X-linked adrenoleucodystrophy is a cause of adrenal insufficiency in association with demyelination within the nervous system due to a failure of b-oxidation of fatty acids within peroxisomes. Increased accumulation of very long-chain fatty acids occurs in many tissues and serum assays can be used diagnostically. Only males have the fully expressed condition and carrier females are usually normal. Two forms are recognized. Adrenoleucodystrophy presents at 5 to 10 years of age with progression eventually to a blind, mute, and severely spastic tetraplegic state. Adrenal insufficiency is usually present but does not appear to correlate with the neurological deficit. X-linked adrenoleucodystrophy accounts for about 10 per cent of cases of adrenocortical failure in boys and men. Adrenomyeloneuropathy, by contrast, presents later in life with the gradual development of spastic paresis and peripheral neuropathy. As both the childhood and the adult condition result from the same mutant gene (recently mapped to chromosome Xq28), it has been suggested that there is an additional autosomal modifier gene. Monounsaturated fatty acids which block the synthesis of the saturated very long-chain fatty acids have been used for treatment. A combination of erucic acid and oleic acid (Lorenzo's oil) has led to normal levels of very long-chain fatty acids, but this has not altered the rate of neurological deterioration; bone marrow transplantation appears to be more effective if undertaken in the early stages of the disease.

Familial glucocorticoid deficiency is a rare, autosomal recessive cause of hypoadrenalism which usually presents in childhood. The renin–angiotensin–aldosterone axis is intact and children usually present either with neonatal hypoglycaemia or later with increasing pigmentation, often with enhanced growth velocity. Patients have glucocorticoid deficiency with very high plasma ACTH levels—this occurs because of mutations in the melanocortin-2 or ACTH receptor. A variant syndrome is called the triple A or Allgrove's syndrome and refers to the triad of adrenal insufficiency due to ACTH resistance, achalasia, and alachrima. Mutations have not been found in the ACTH receptor and the molecular basis for this inherited syndrome is unknown.

Secondary hypoadrenalism (ACTH deficiency)

This is a common clinical problem and is most often due to a sudden cessation of exogenous glucocorticoid therapy or a failure to give glucocorticoid cover for intercurrent stress in a patient who has been on long-term glucocorticoid therapy. Such therapy suppresses the hypothalamic–pituitary–adrenal axis, with consequent adrenal atrophy and this may last for months after stopping glucocorticoid treatment. Adrenal atrophy and subsequent deficiency should be anticipated in any subject who has taken more than the equivalent of 30 mg of oral hydrocortisone per day (approximately 7.5 mg/day of prednisolone or 0.75 mg/day of dexamethasone) for longer than 1 month. In addition to the magnitude of the dose of glucocorticoid, the timing of administration of the dose may affect the degree of adrenal suppression. Thus prednisolone in a dose of 5 mg given last thing at night and 2.5 mg in the morning will produce more marked suppression of the hypothalamic–pituitary–adrenal axis compared with 2.5 mg at night and 5 mg in the morning because the larger evening dose blocks the early morning surge of ACTH.

Other causes of secondary adrenal insufficiency are rare (Table 6) and reflect inadequate ACTH production from the anterior pituitary gland. In many of these, other pituitary hormones are deficient in addition to ACTH, so that the patient presents with partial or complete hypopituitarism. The clinical features of hypopituitarism make this a relatively easy diagnosis to make (see Chapter 12.4). However, if there is isolated ACTH deficiency, this diagnosis may be readily missed.

Clinical features of adrenal insufficiency

The most obvious feature which differentiates primary from secondary hypoadrenalism is skin pigmentation (Fig. 17), which is nearly always present in primary adrenal insufficiency (unless of short duration) and absent in secondary. The pigmentation is seen in sun-exposed areas, recent rather than old scars, axillas, nipples, palmar creases, pressure points, and in mucous membranes (buccal, vaginal, vulval, anal). The cause of the pigmentation has long been debated but probably reflects increased melanocyte activity induced by POMC-related peptides including melanocyte stimulating hormone (MSH). In autoimmune Addison's disease there may be associated vitiligo (Fig. 17).

Fig. 17 Pigmentation in a patient with Addison's disease before and after treatment with hydrocortisone and fludrocortisone (by courtesy of Professor C.R.W. Edwards).

Patients with primary adrenal failure usually have both glucocorticoid and mineralocorticoid deficiency. In contrast, those with secondary adrenal insufficiency have an intact renin–angiotensin–aldosterone system. This accounts for differences in salt and water balance in the two groups of patients, which in turn result in different clinical presentations.

Primary adrenal failure may present with hypotension and acute circulatory failure (addisonian crisis). Anorexia may be an early feature, which progresses to nausea, vomiting, diarrhoea, and sometimes, abdominal pain. These crises may be precipitated by intercurrent infection or by stress, such as a surgical operation. Alternatively, the patient may present with vague features of chronic adrenal insufficiency—weakness, tiredness, weight loss, nausea, intermittent vomiting, abdominal pain, diarrhoea or constipation, general malaise, muscle cramps, and symptoms suggestive of postural hypotension. Salt-craving may be a feature and there may be a low-grade fever. The lying blood pressure is usually normal but almost invariably there is a fall in blood pressure on standing.

In secondary adrenal insufficiency due to hypopituitarism, the presentation may relate to deficiency of hormones other than ACTH, notably luteinizing hormone/follicle-stimulating hormone (infertility, oligo-/amenorrhoea, poor libido), thyroid-stimulating hormone (weight gain, cold intolerance), and growth hormone (hypoglycaemia). Patients with isolated ACTH deficiency present with malaise, weight loss, and other features of chronic adrenal insufficiency.

Laboratory investigation of hypoadrenalism

Routine biochemical profile

In established primary adrenal insufficiency, hyonatraemia is present in about 90 per cent of cases and hyperkalaemia in 65 per cent. The blood urea concentration is usually elevated. In secondary adrenal failure there may be a dilutional hyponatraemia with normal or low blood urea because glucocorticoids are required to maintain glomerular filtration rate and excrete a water load. Hypoglycaemia has been found in up to 50 per cent of patients with chronic adrenal insufficiency.

Plasma cortisol/ACTH

Clinical suspicion of the diagnosis should be confirmed with definitive diagnostic tests. Basal plasma cortisol and urinary free cortisol levels are often in the low normal range and cannot be used to exclude the diagnosis. In primary adrenal insufficiency the simultaneous measurement of plasma cortisol and plasma ACTH reveals an ACTH level which is disproportionately elevated in comparison with plasma cortisol (Fig. 18).

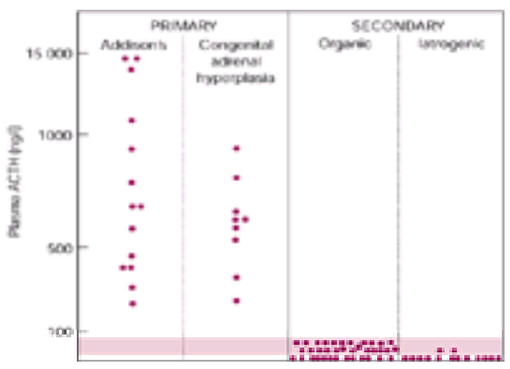

Fig. 18 Morning immunoreactive ACTH values in patients with hypoadrenalism. The reference range is indicated by the horizontal lines (by courtesy of Professor L. H. Rees.)

Mineralocorticoid status

Another difference between primary and secondary hypoadrenalism is in the renin–angiotensin–aldosterone axis. In primary hypoadrenalism there is normally mineralocorticoid deficiency with elevated plasma renin activity and either low or low normal plasma aldosterone. It is remarkable how frequently the investigation of zona glomerulosa activity is ignored in Addison's disease compared with the assessment of zona fasciculata function.

Stimulation tests

In practice all patients suspected of having adrenal insufficiency should have an ACTH stimulation test. This involves the intramuscular or intravenous administration of 250 µg of tetracosactrin (Synacthen ®), comprising the first 24 amino acids of normally secreted 1–39 ACTH. Plasma cortisol levels are measured at 0 and 30 min after ACTH administration and a normal response is defined by a peak plasma cortisol of more than 550 nmol/l. Levels of less than 550 nmol/l in response to Synacthen are found in both primary and secondary adrenal insufficiency, though rarely false-positive results have been reported, particularly in cases of secondary hypoadrenalism. A low-dose ACTH stimulation test giving only 1 µg ACTH has been proposed to screen for adequacy of function of the hypothalamo–pituitary–adrenal axis with the suggestion that it may be more sensitive than the conventional 250 µg test. At present there are insufficient data to support such a concept.

A prolonged ACTH stimulation test, involving the administration of depot tetracosactrin in a dose of 1 mg by intramuscular injection, with measurement of plasma cortisol at 0, 4, and 24 h will differentiate primary from secondary hypoadrenalism. In normal subjects the plasma cortisol at 4 h is more than 1000 nmol/l and the value at 24 h shows little further increase. Patients with secondary hypoadrenalism show a delayed response with usually a much higher value at 24 than at 4 h, but in primary hypo-adrenalism there is no response at either time. However, the test is now rarely required if plasma ACTH has been appropriately measured at baseline.

The insulin-induced hypoglycaemia or insulin tolerance test remains one of the most useful in assessing ACTH and growth hormone reserves. It should not be performed in patients with ischaemic heart disease (check ECG before test), epilepsy, or severe hypopituitarism (that is plasma cortisol at 09.00 h less than 180 nmol/l). The test involves the intravenous administration of soluble insulin in a dose of 0.1 to 0.15 U/kg body weight, with measurement of plasma cortisol at 0, 30, 45, 60, 90, and 120 min. Adequate hypoglycaemia (blood glucose less than 2.2 mmol/l with signs of neuroglycopenia—sweating and tachycardia) is essential. In normal subjects the peak plasma cortisol exceeds 500 nmol/l. However, the response to hypoglycaemia can be reliably predicted by the response to acute ACTH stimulation (see above); a safer, cheaper, and quicker test. If the ACTH test is normal, insulin-induced hypoglycaemia testing is not necessary in the vast majority of cases unless there is a need to document endogenous growth hormone reserve in a patient with pituitary disease. An insulin tolerance test is only required if, in a patient with suspected hypopituitarism, there is a subnormal response to ACTH. Some patients have an inadequate response to ACTH but then respond normally to hypoglycaemia. They do not require corticosteroid replacement therapy.

Other tests

Radioimmunoassays to detect autoantibodies such as those against the 21-hydroxylase antigen are now available and should be undertaken in patients with primary adrenal failure. In autoimmune Addison's disease it is also important to look for evidence of other organ-specific autoimmune disease. In long-standing tuberculous adrenal disease there may be adrenal atrophy with calcification on plain radiographs or CT scanning. Early morning urine samples should be cultured for mycobacteria if tuberculosis is suspected.

Treatment of acute adrenal insufficiency

This is an emergency, and treatment should not be delayed while waiting for definitive proof of diagnosis. However, in addition to measurement of plasma electrolytes and blood glucose, appropriate samples for ACTH and cortisol should be taken before giving corticosteroid therapy. If the patient is not critically ill, an acute ACTH stimulation test can be performed. However, if necessary, this can be delayed and carried out with the patient on corticosteroid therapy, provided the drug used does not interfere with the plasma cortisol assay (for example, change from hydrocortisone to dexamethasone).

Intravenous hydrocortisone should be given in a dose of 100 mg every 6 h. If this is not possible, then the intramuscular route should be used. In the shocked patient, 1 litre of normal saline should be given intravenously over the first hour. Because of possible hypoglycaemia, it is normal to give 5 per cent dextrose saline. The subsequent saline and dextrose therapy will depend on biochemical monitoring and the patient's condition. Clinical improvement, especially in the blood pressure, should be seen within 4 to 6 h if the diagnosis is correct. It is important to recognize and treat any associated condition, such as an infection, which may have precipitated the acute adrenal crisis.

After the first 24 h the dose of hydrocortisone can be reduced, usually to 50 mg intramuscularly every 6 h for the second 24 h and then, if the patient can take by mouth, to oral hydrocortisone, 40 mg in the morning and 20 mg at 18.00 h. This can then be rapidly reduced to the normal replacement dose of 20 mg on wakening and 10 mg at 18.00 h. Some patients will require more than 30 mg/day, but increasingly most patients can cope with less than this dose (usually 15 to 25 mg/day in divided doses). In primary adrenal failure, cortisol day curves with simultaneous ACTH measurements may provide some insight into adequacy of replacement therapy, but unfortunately there are no good objective tests in secondary adrenal failure. Nevertheless, crude objectives such as weight, well being, and blood pressure are important in this regard.

In primary adrenal failure, mineralocorticoid replacement is usually also required in the form of fludrocortisone (or 9a-fluorinated hydrocortisone) in a dose of 0.05 to 0.1 mg/day. The mineralocorticoid activity of this is about 125 times that of hydrocortisone. After the acute phase has passed, the adequacy of mineralocorticoid replacement can be assessed by measuring electrolytes, supine and erect blood pressure, and plasma renin activity; too little fludrocortisone may cause postural hypotension with elevated plasma renin activity, whilst too much causes the converse.

Patients receiving glucocorticoid replacement therapy should be advised to double the dose in the event of intercurrent febrile illness, accident, or mental stress such as an important examination. If the patient is vomiting and cannot take by mouth, parenteral hydrocortisone must be given urgently, as indicated above. For minor surgery, 50 to 100 mg of hydrocortisone hemisuccinate is given with the premedication. For major procedures this is then followed by the same regimen as for acute adrenal insufficiency.

Every patient on glucocorticoid therapy should be advised to register for a MedicAlert bracelet or necklace and must carry a 'Steroid card'.

Mineralocorticoid excess

Hypertension affects 10 to 25 per cent of the population. In most cases, no underlying cause for the patient's raised blood pressure is apparent, and they are labelled as having 'essential' hypertension. Mineralocorticoid-based hypertension may account for secondary causes of hypertension, and classically refers to hypertension caused by increased sodium and water retention by the kidney and expansion of the extracellular fluid compartment resulting in suppression of endogenous plasma renin activity. Unlike the majority of cases of secondary aldosteronism which arise either in the setting of reduced oncotic pressure (nephrosis, cirrhosis) or in patients with cardiac failure, oedema is not a feature of primary aldosteronism, probably because of the 'escape' phenomenon. Nevertheless, in the short term, intravascular volume is reset at a higher level and this leads to increased cardiac output and blood pressure. In the chronic state, hypervolaemia cannot be consistently demonstrated and other mechanisms may be equally important in raising blood pressure. Mineralocorticoid receptors have been characterized in the vasculature and heart, and depending upon the activity of local 11b-HSD, either glucocorticoids or mineralocorticoids may increase vascular tone, by potentiating catecholamine and angiotensin II-induced vasoconstriction, or by inhibiting endothelial relaxation. Mineralocorticoids can also modulate blood pressure centrally, independent of changes in renal electrolyte transport or vascular reactivity.

Mineralocorticoid hypertension: differential diagnosis

A comprehensive list of the causes of mineralocorticoid hypertension is given in Table 7.

Primary aldosteronism

First described by Conn in 1955, this is the commonest cause of mineralocorticoid hypertension. Prevalence rates of 0.5 to 2 per cent have been widely reported in the literature in unselected patients with 'essential' hypertension, but many of these studies relied on detecting hypokalaemia and, in the light of recent observations, will have underestimated true prevalence rates. By contrast, studies suggesting much higher prevalence rates of 5 to 12 per cent in hypertensive populations have been conducted in specialist centres and are therefore subject to selection bias.

Symptoms are often absent or non-specific but include tiredness, muscle weakness, thirst, polyuria, and nocturia due to hypokalaemia. Spontaneous hypokalaemia (less than 3.5 mmol/l) is rare in untreated hypertension; when found in a patient on diuretics these should be withdrawn, and potassium stores replenished and remeasured 2 weeks later. Despite this, it is now accepted that up to 40 per cent of patients with surgically confirmed primary aldosteronism will have normal serum potassium concentrations.

In approximately two-thirds of patients, primary aldosteronism is due to a small (0.5 to 2 cm), solitary aldosterone-producing adenoma of the adrenal which is commoner in women than men (male:female ratio 1:3). One-third of cases are caused by bilateral adrenal hyperplasia, and the remaining few (less than 2 per cent) by glucocorticoid-suppressible hyperaldosteronism or adrenal carcinomas. The aetiology of aldosterone-producing adenoma is unknown, although rarely it may have a genetic basis and can occur as a component of multiple endocrine neoplasia type I.

Primary aldosteronism is confirmed by demonstrating subnormal supine and erect plasma renin activity and an elevated plasma aldosterone concentration in a patient with no antihypertensive treatment for at least 3 weeks. However, primary aldosteronism may also occur with suppressed plasma renin activity and normal plasma aldosterone concentration, and some investigators advocate measures of aldosterone secretion over a 24-h period or salt suppression studies to further confirm the diagnosis. If severe hypertension prevents complete cessation of antihypertensive therapy during this diagnostic period, a-blockers such as prazosin or doxazosin interfere least with the renin–angiotensin–aldosterone axis. A single ratio of plasma renin activity/plasma aldosterone concentration may be a sensitive screening test, even in patients still taking antihypertensive medication, but depending upon the assays used and population salt intake, this requires validation in each centre.

The differential diagnosis of aldosterone-producing adenoma, bilateral adrenal hyperplasia, and glucocorticoid-suppressible hyperaldosteronism requires an understanding of the control of aldosterone secretion in each condition. In normal physiology, aldosterone secretion is under the control of angiotensin II through the renin–angiotensin system; ACTH and potassium are less important chronic secretogogues. Aldosterone-producing adenoma represents an autonomous source of aldosterone production which is not regulated by angiotensin II (ACTH assumes more importance in the control of aldosterone secretion in aldosterone-producing adenoma). By contrast, the zona glomerulosa in bilateral adrenal hyperplasia is more sensitive to angiotensin II; for a given angiotensin II infusion there is a much greater aldosterone response than normal. Finally, in glucocorticoid-suppressible hyperaldosteronism, aldosterone secretion and the secretion of intermediary metabolites (18-hydroxy and oxo-metabolites of cortisol and corticosterone) are under the control of ACTH. The optimal method(s) of establishing the differential diagnosis of primary aldosteronism are complicated and still controversial. In our clinical practice this is undertaken as a day case admission as illustrated in Fig. 19, measuring the response of aldosterone to erect posture (high in bilateral adrenal hyperplasia, absent in aldosterone-producing adenoma), to ACTH (absent in bilateral adrenal hyperplasia, increased in aldosterone-producing adenoma, exaggerated in glucocorticoid-suppressible hyperaldosteronism), and 18-hydroxy- or 18-oxo-cortisol/corticosterone in the plasma or urine. This study is only valid if plasma renin activity is seen to rise on adopting the erect posture and cortisol (reflecting underlying ACTH secretion) to fall between 08.00 h and 12.00 h.

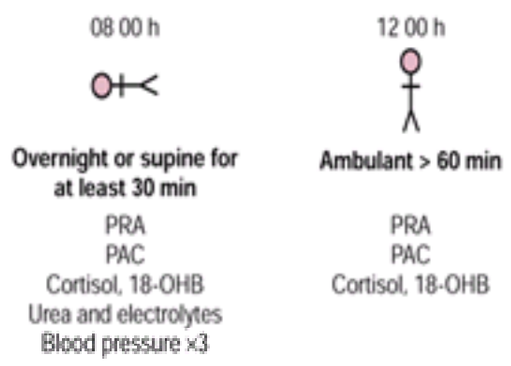

Fig. 19 Day case investigation of a patient suspected of having primary aldosteronism. Supine and erect plasma renin activity (PRA), plasma aldosterone concentration (PAC), cortisol (F), 18-hydroxycorticosterone (18-OHB) (or 18-OHF), electrolytes, and blood pressure are measured as shown. The posture study is only valid if there is a rise in PRA on adopting the erect posture and a fall in F concentrations between 08.00 h and 12.00 h (reflecting a circadian fall in ACTH levels). (Reproduced with permission of the author and *The Lancet.*)

Adrenal MRI/CT scanning should not be performed until a biochemical diagnosis has been made because of the high incidence of non-functioning adrenal incidentalomas. Thereafter, an MRI/CT scan should be the first localization procedure; CT has a better spatial resolution and may be more sensitive in detecting smaller aldosterone-producing adenomas (Fig. 20 and Plate 2). Adrenal scintigraphy studies (iodocholesterol scanning) have been found useful in some centres. Adopting the approach outlined in Fig. 19, few patients need invasive adrenal vein cannulation, although this may be required to make a diagnosis or to assist in the lateralization of a lesion if the posture and/or imaging studies are inconclusive. In particular, angiotensin II-responsive aldosterone-producing adenomas have been reported, and this may explain why the overall accuracy of posture studies in primary aldosteronism is only 70 to 80 per cent. Although technically difficult and not without risk, the demonstration of an aldosterone ratio of greater than 10:1 in one adrenal vein compared with the other remains the most sensitive diagnostic test. Simultaneous cortisol measurements ensure adrenal vein cannulation and, when expressed as an aldosterone/cortisol ratio, improve diagnostic accuracy.

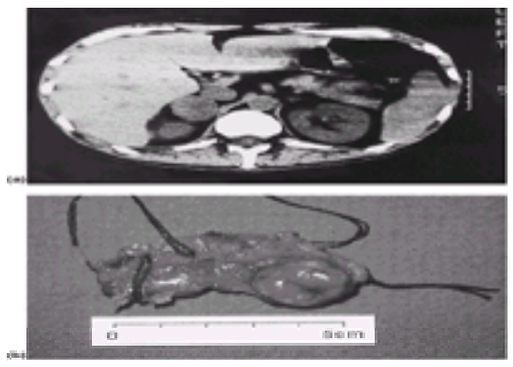

Fig. 20 (a) Adrenal CT scan demonstrating a solitary adrenal adenoma in a patients with Conn's syndrome. (b) The characteristic yellow appearance of the cut surface of the excised tumour reflects the high cholesterol content of these tumours. (See also Plate 2.)

One reason for establishing a definitive diagnosis is that treatment is surgical excision in the case of aldosterone-producing adenoma, but strictly medical for bilateral adrenal hyperplasia and glucocorticoid-suppressible hyperaldosteronism. The last responds well to dexamethasone at 0.25 to 0.5 mg/day. Patients with aldosterone-producing adenoma who are not suitable for surgery or decline operation and patients with bilateral adrenal hyperplasia should be treated with amiloride (starting dose 5 to 10 mg/day increasing to 30 mg/day depending upon blood pressure, urea, and electrolyte response). Spironolactone is as effective but in high doses frequently causes painful gynaecomastia and menstrual irregularity. In aldosterone-producing adenoma, normokalaemia is restored in 100 per cent of patients postoperatively and blood pressure falls to normal values in 70 per cent. With the ever increasing experience of laparascopic adrenalectomy, surgical morbidity can be kept to a minimum. Pre- and perioperative treatment should involve the co-ordinated management of surgeon and endocrinologist. Aldosterone secretion from the contralateral normal adrenal gland may be suppressed and hypoaldosteronism postoperatively should be anticipated and treated appropriately by increasing sodium intake and/or transient fludrocortisone therapy.

Monogenic hypertension

Hypertension is known to be a phenotype of some well-documented gene mutations; 17a-hydroxylase deficiency and 11b-hydroxylase deficiency are forms of congenital adrenal hyperplasia in which mineralocorticoid excess occurs because of ACTH-driven deoxycorticosterone excess. A similar process is thought to explain the hypertension seen in patients with glucocorticoid resistance due to mutations in the glucocorticoid receptor (GR) gene (Table 1). More recently, a significant advance in our understanding of the molecular basis of cardiovascular disease has been the elucidation of other single gene defects causing mineralocorticoid hypertension (Fig. 21).

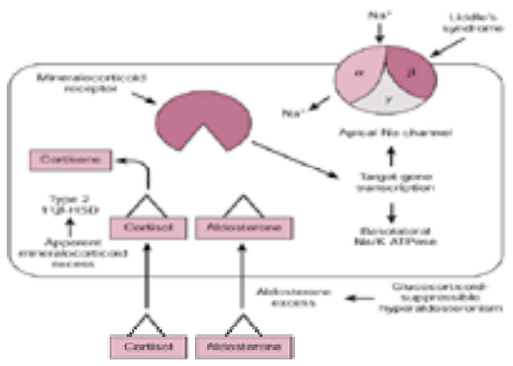

Fig. 21 Three causes of monogenic mineralocorticoid hypertension are detailed. A schematic diagram representing an epithelial cell in the distal colon or distal nephron is shown. In normal physiology, aldosterone interacts with the mineralocorticoid receptor (MR) to stimulate sodium reabsorption via induction of the apical sodium channel and serosal Na^+/K^+-ATPase pump. GSH (glucocorticoid-suppressible hyperaldosteronism) is a cause of aldosterone excess due to the production of a chimeric gene, 11b-hydroxylase/aldosterone synthase, within the adrenal cortex. Apparent mineralocorticoid excess results because cortisol cannot be inactivated to cortisone by the type 2 isoform of 11b-hydroxysteroid dehydrogenase (11b-HSD2); cortisol can then act as a potent mineralocorticoid. Liddle's syndrome occurs because of constitutively active mutations in the b- or g-subunits of the apical sodium channel.

Glucocorticoid-suppressible hyperaldosteronism

Glucocorticoid-suppressible hyperaldosteronism was first reported in 1966 and is an autosomal dominant form of low-renin hypertension characterized by aldosterone excess under the control of ACTH rather than the normal principal secretogogue, angiotensin II. There are two important consequences of this; first, there is dysregulation of aldosterone secretion because of loss of the negative feedback loop (aldosterone does not suppress ACTH secretion), and second, the exogenous administration of a glucocorticoid such as dexamethasone, by decreasing ACTH secretion, results in suppression of aldosterone secretion and can be used therapeutically. Long-term glucocorticoid therapy leads to reactivation and normal regulation of the renin–angiotensin–aldosterone axis. A further characteristic of glucocorticoid-suppressible hyperaldosteronism is the secretion of large quantities of 18-hydroxy- and 18-oxo-corticosterone/cortisol metabolites, again under the control of ACTH, and while there is some overlap with levels seen in aldosterone-producing adenoma, these provide a diagnostic marker for the condition.

The molecular basis for glucocorticoid-suppressible hyperaldosteronism was described by Lifton and colleagues following the cloning and characterization of the two final enzymes in cortisol and aldosterone synthesis, 11b-hydroxylase and aldosterone synthase, respectively. 11b-Hydroxylase converts 11-deoxycortisol to cortisol in the zona fasciculata, and aldosterone synthase converts corticosterone to aldosterone through an enzymatic step involving 11b-hydroxylation and 18-hydroxylation and oxidation. These enzymes are encoded by two genes, *CYP11B1* and *CYP11B2*, lying in tandem on chromosome 8. Despite the similarity in the coding sequences of 11b-hydroxylase and aldosterone synthase (more than 95 per cent), their 5' sequences differ, permitting regulation of 11b-hydroxylase by ACTH through cAMP and aldosterone synthase by angiotensin II through intracellular calcium ions, thereby establishing functional zonation of the adrenal cortex. In glucocorticoid-suppressible hyperaldosteronism a hybrid gene is formed at meiosis from unequal cross-over of the *CYP11B1* and *CYP11B2* genes and this contains proximal components of *CYP11B1* and distal components of *CYP11B2*. As long as the breakpoint of the hybrid gene is in or 5' to exon IV of the *CYP11B1* gene, the product of this gene can synthesize aldosterone, but is now under the control of ACTH (Fig. 22). The chimeric gene can be detected by Southern blotting or by long polymerase chain reaction, providing a screening test for glucocorticoid-suppressible hyperaldosteronism and the facility for prenatal diagnosis.

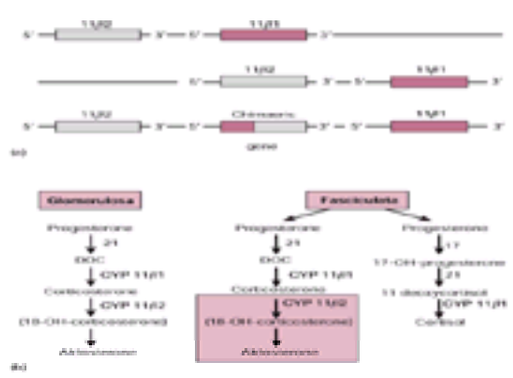

Fig. 22 (a) Chimeric gene responsible for glucocorticoid-remediable hyperaldosteronism and its impact upon adrenal steroid secretion. (b) The chimeric gene is

expressed in the zona fasiculata (boxed area) and can synthesize aldosterone but is under the regulatory control of ACTH.

In the wake of such advances, numerous kindreds have been reported with glucocorticoid-suppressible hyperaldosteronism and an international register for such cases has been established, which to date comprises 167 cases from 27 genetically proven pedigrees. Interesting observations to come from these larger cohorts are that potassium may be normal in up to 50 per cent of cases and there is poor correlation between genotype and phenotype (potassium, blood pressure) both between and within families. Severe mineralocorticoid excess has been reported in some individuals with this gene defect, but in other members of the same family, the gene defect has caused no abnormal phenotype. Patients with glucocorticoid-suppressible hyperaldosteronism are more susceptible to cerebrovascular haemorrhage.

Liddle's syndrome

In 1963, Grant Liddle described a family with several siblings affected by early-onset hypertension and hypokalaemia associated with low renin and low aldosterone levels. The condition responded well to inhibitors of epithelial sodium transport such as triamterene, but not to mineralocorticoid receptor antagonists such as spironolactone, and studies on erythrocytes suggested a generalized defect in sodium transport. Furthermore, in the proband of one of Liddle's original patients, renal transplantation resulted in blood pressure and potassium returning to normal levels, arguing against a circulating mineralocorticoid.

Mineralocorticoid-dependent epithelial sodium transport requires the activation of the apical sodium channel. Three subunits of this channel, a, b, and g, have been cloned and characterized. Full sodium conductance requires the concerted action of a/b or a/g subunits and cannot be sustained by any subunit in isolation. The b and g subunits lie in close proximity on chromosome 16 and mutations in these subunits have been described in kindreds affected with Liddle's syndrome. In each case these cause deletions of the C-terminus part of the protein (45 to 75 amino acids) producing a sodium channel which is constitutively active. Liddle's syndrome is inherited as an autosomal dominant trait and several other kindreds have been reported following the description of the genetic basis for the condition. As is the case with glucocorticoid-suppressible hyperaldosteronism, potassium has been reported to be normal in several patients.

Apparent mineralocorticoid excess and abnormalities of 11b-hydroxysteroid dehydrogenase type 2

Apparent mineralocorticoid excess was first described in detail by Ulick and New in the late 1970s. This is an autosomal recessive form of low renin, low aldosterone hypertension in which cortisol, conventionally regarded as a glucocorticoid, is able to act as a potent mineralocorticoid. The condition can be diagnosed from a 24-h urine collection analysed for cortisol metabolites using gas chromatography. Affected individuals have a characteristic increase in urinary cortisol compared with cortisone metabolites (tetrahydrocortisols/tetrahydrocortisone ratio or urinary free cortisol/urinary free cortisone ratio). Serum cortisol levels are unhelpful because although patients with apparent mineralocorticoid excess have a prolonged plasma cortisol half-life, a reduction in cortisol secretion rate mediated by the negative feedback mechanism ensures normal circulating concentrations. This defect in cortisol metabolism occurs because of loss of 11b-hydroxy-steroid dehydrogenase (11b-HSD) activity.

Two isozymes of 11b-HSD catalyse the interconversion of hormonally active cortisol (**F**) to inactive cortisone (**E**). 11b-HSD1 is predominantly found in the liver, adipose tissue, and gonad and acts principally as an oxo-reductase generating F from E, but it is the 11b-HSD2 isoform, acting as an efficient dehydrogenase inactivating F to E which is expressed in the mineralocorticoid target tissues, kidney, colon, and salivary gland, that is more important in modulating corticosteroid control of blood pressure. Aldosterone gains access to the mineralocorticoid receptor *in vivo* only when 11b-HSD2 activity is intact and F can be inactivated to E at a prere-ceptor level (Fig. 23). Homozygous inactivating mutations in the human 11b-HSD2 gene have been identified in over 20 patients with apparent mineralocorticoid excess and result in cortisol-mediated, mineralocorticoid hypertension. The condition is inherited as an autosomal recessive trait and the majority of heterozygotes, with a few notable exceptions, have a normal phenotype. Milder forms of apparent mineralocorticoid excess have been described and there appears to be a close correlation between genotype and phenotype. Spironolactone or amiloride (often in higher doses than those used to treat primary aldosteronism) can be used therapeutically, as can dexamethasone, which suppresses endogenous cortisol secretion, but itself is not a good substrate for 11b-HSD2.

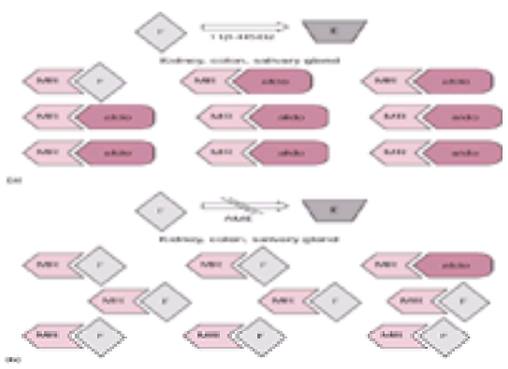

Fig. 23 (a) The role of 11b-hydroxysteroid dehydrogenase (11b-HSD2) in protecting the non-specific mineralocorticoid receptor from cortisol. (b) With congenital or acquired deficiency of the enzyme, F (cortisol) cannot be inactivated to E (cortisone) and acts as a potent mineralocorticoid.

Liquorice has been associated with a mineralocorticoid excess state since the late 1940s when Reevers, a Dutch Physician, used a liquorice preparation, 'succus liquoritiae', to treat patients with dyspepsia. This was the origin of the antiulcer drug, carbenoxolone, which also resulted in mineralocorticoid side-effects in up to 50 per cent of patients. The active 'mineralocorticoids' in both cases are glycyrrhizic acid and its hydrolytic product, glycyrrhetinic acid, which themselves have little inherent mineralocorticoid activity, but cause hypertension and hypokalaemia by inhibiting 11b-HSD2. Such patients will also have an increase in the urinary ratio of cortisol to cortisone metabolites (THF+allo-THF/THE), although not to the same degree as patients with apparent mineralocorticoid excess.

Cortisol is also the offending mineralocorticoid in patients with some forms of Cushing's syndrome. In ectopic ACTH syndrome, for example, the high cortisol secretion rate overwhelms renal 11b-HSD2, resulting in spill over on to the mineralocorticoid receptor. A high THF+allo-THF/THE ratio is also observed in some patients with pituitary-dependent Cushing's syndrome and this may explain the hypertension in these cases.

These unusual causes of mineralocorticoid hypertension have significantly enhanced our understanding of corticosteroid biosynthesis and hormone action. In addition they raise new questions as to the role of adrenal steroids in wider populations of patients with hypertension. Defects in the activity of 11b-HSD have been reported in patients with 'essential hypertension', but have not consistently been associated with mineralocorticoid excess. Endogenous circulating inhibitors of 11b-HSD2 have also been described, so-called 'glycyrrhetinic acid-like factors' or GALFs. Levels are higher in pregnancy, and some studies, but not others, report increased levels in patients with hypertension. At present, however, the identity of such GALFs is unknown.

Who should we suspect as having mineralocorticoid-based hypertension?

All patients with hypertension should have serum electrolytes measured and those with hypokalaemia must be investigated (Fig. 24). Because hypokalaemia may be absent in many cases of proven mineralocorticoid hypertension, patients with severe hypertension (for example those on triple antihypertensive therapy) and those with a family history of hypertension or cerebrovascular disease should also be screened. At present the incidence of primary aldosteronism, glucocorticoid-suppressible hyperaldosteronism, Liddle's syndrome, and apparent mineralocorticoid excess in unselected (that is, community rather than hospital-based) populations with 'essential' hypertension is unknown. Until this is defined, one cannot be more dogmatic in deciding who should and should not be screened for mineralocorticoid-based hypertension. In the interim these diagnoses will not be made unless they are considered. Wherever possible the elucidation of the underlying basis of a patient's hypertension should be sought so that appropriate therapy can be targeted to the patient.

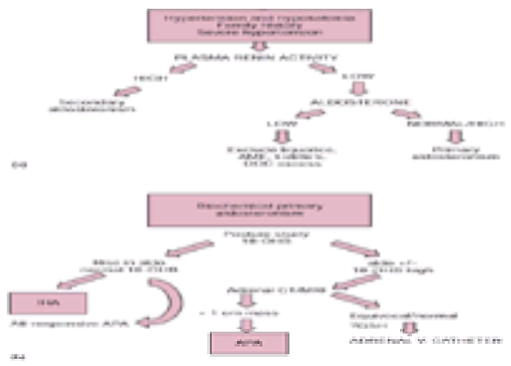

Fig. 24 Proposed algorithm for investigating mineralocorticoid-based hypertension. In practice the posture study (Fig. 19) can be included when the initial biochemical confirmation studies are performed.

Glucocorticoid resistance

A small number of patients have been described who have increased cortisol secretion but with none of the stigmas of Cushing's syndrome. These patients are resistant to suppression of cortisol with low-dose dexamethasone but respond to high doses. ACTH levels are elevated and lead to increased adrenal production of androgens and deoxycorticosterone. Thus the patients may present with the features of androgen and/or mineralocorticoid excess. Treatment with a dose of dexamethasone adequate to suppress ACTH (usually 3 mg/day) results in a fall in adrenal androgens and often return of plasma potassium and blood pressure to normal levels. Many of these patients have been found to have point mutations in the steroid-binding domain of the glucocorticoid receptor, with consequent reduction of glucocorticoid-binding affinity.

Mineralocorticoid deficiency

These syndromes are listed in Table 8. They can be divided into those that are congenital and others that are acquired.

Adrenal insufficiency

Mineralocorticoid deficiency may occur in some forms of congenital adrenal hyperplasia and these are discussed elsewhere (see Chapter 12.7.2). Similarly, other causes of adrenal insufficiency (for example Addison's disease and congenital adrenal hypoplasia) are discussed above.

Primary defects in aldosterone biosynthesis

Failure of conversion of corticosterone to 18-hydroxycorticosterone or of 18-hydroxycorticosterone to aldosterone usually presents as a salt-wasting crisis in neonatal life (Fig. 1). Hyperkalaemia, metabolic acidosis, dehydration, and hyponatraemia are found. The condition has been called corticosterone methyl oxidase (**CMO**) deficiency, but this was before the final enzyme(s) involved in the conversion of deoxycorticosterone to aldosterone were characterized and cloned. In fact a single enzyme, aldosterone synthase carries a multistep reaction involving 11-hydroxylation of deoxycorticosterone to corticosterone, 18-hydroxylation of corticosterone to 18-hydroxycorticosterone, followed by 18-dehydrogenation to aldosterone. Two variants of CMO deficiency are described; CMO I is characterized by low 18-hydroxycorticosterone and aldosterone levels, whereas patients with CMO II deficiency have hypoaldosteronism but high 18-hydroxy-corticosterone concentrations. In both cases, mutations in the gene encoding aldosterone synthase have been described and the discrepant 18-hydroxycorticosterone levels seem likely to be explained on the basis of variable 18-hydroxylase activity of the related CYP45011b-hydroxylase 1 enzyme. CMO II is much more common in Iranian Jews than in the Caucasian population.

Defects in aldosterone action: pseudohypoaldosteronism

Pseudohypoaldosteronism type I presents in infancy with severe salt-wasting and failure to thrive but with very high plasma aldosterone, and plasma renin activity levels with inappropriate urinary sodium loss. The mineralocorticoid receptor appears to be defective, as judged by studies looking at the binding of aldosterone to monocytes, but molecular studies have failed to show any abnormality in the mineralocorticoid receptor itself. Recently, inactivating mutations in the a, b, and g subunits of the epithelial sodium channel have been shown to explain the condition, that is, exactly the opposite phenotype of Liddle's syndrome. Acquired forms of pseudohypoaldosteronism can occur in patients after renal transplantation, following obstructive uropathy, and in premature infants.

Pseudohypoaldosteronism type II or Gordon's syndrome is an autosomal dominant disorder characterized by hyperkalaemia but not salt-wasting in contrast to the type I condition. Patients have resistance to the mineralocorticoid effects of aldosterone on tubular potassium transport, but not to those of sodium and chloride transport. As a result, affected individuals have hyperchloraemia, hypertension, and suppression of plasma renin activity. Recent intronic deletions in the gene encoding a serine-threonine kinase, WNK, have been described in affected cases.

Hyporeninaemic hypoaldosteronism

Angiotensin II is a key stimulus to aldosterone secretion, and thus damage or blockade of the renin–angiotensin system may result in mineralocorticoid deficiency. Various renal diseases have been associated with damage to the juxtaglomerular apparatus and hence renin deficiency. Of these the most common (more than 75 per cent of cases) is diabetic nephropathy.

The usual picture is of an elderly patient with hyperkalaemia, acidosis, and mild to moderate impairment of renal function. Plasma renin activity and aldosterone are low and fail to respond to sodium depletion, the erect posture, or frusemide administration. Unlike those with adrenal insufficiency, patients with hyporeninaemic hypoaldosteronism have normal or elevated blood pressure but no postural hypotension. Muscle weakness and cardiac arrythmias may also occur. Other factors may also contribute to the hyperkalaemia, including the use of potassium-sparing diuretics, potassium supplementation, insulin deficiency, and b-adrenoceptor blocking drugs, and prostaglandin synthetase inhibitors which inhibit renin release.

Treatment of primary renin deficiency is with fludrocortisone in the first instance together with dietary potassium restriction. However, these patients are not salt depleted and may become hypertensive with fludrocortisone. In such a setting, the addition of a loop-acting diuretic such as frusemide is appropriate. This will also increase the excretion of acid and thus improves the metabolic acidosis.

Adrenal 'incidentalomas'

With the more widespread use of high-resolution imaging procedures (CT, MRI), incidentally discovered adrenal masses have become a common problem. An adrenal mass will be uncovered in up to 4 per cent of patients imaged for non-adrenal pathology. Over 80 per cent of cases are non-functioning with phaeochromocytomas and cortisol- or aldosterone-secreting adenomas comprising the remainder. As a result, all patients with incidentally discovered adrenal masses should undergo appropriate endocrine screening tests (24-hour urinary catecholamines, urinary free cortisol and overnight dexamethasone suppression tests, plasma renin activity/aldosterone, adrenal androgens) to exclude a functional lesion. The possibility of malignancy should be considered in each case. In patients with a known extra-adrenal primary, the incidence of malignancy is obviously much higher (up to 20 per cent of patients with lung cancer, for example, have adrenal metastases on CT scanning). Primary adrenal carcinoma is rare—in one study only 26 of 630 incidentalomas were found to be adrenal carcinomas. In true incidentalomas, size appears to be predictive of malignancy—a lesion of less than 5 cm in diameter is most unlikely to be malignant. Non-functioning lesions of less than 5 cm can therefore be treated conservatively and patients followed with annual imaging. Functional lesions, or tumours larger than 5 cm in diameter should be removed by laparascopic adrenalectomy.

Further reading

Cushing's syndrome

Atkinson AB *et al.* (1985). Five cases of cyclical Cushing's syndrome. *British Medical Journal* **291**, 1453–7.

Kirschner LS *et al*. (2000). Mutations of the gene encoding the protein kinase A type I-alpha regulatory subunit in patients with the Carney complex. *Nature Genetics* **26**, 89–92.

Lacroix A *et al.* (1992). Gastric-inhibitory polypeptide-dependent cortisol hypersecretion—a new cause of Cushing's syndrome. *New England Journal of Medicine* **327**, 974–80.

Mampalam TJ, Tyrell B, Wilson CB (1988). Transsphenoidal microsurgery for Cushing's disease. A report of 216 cases. *Annals of Internal Medicine* **109**, 487–93.

Newell-Price J *et al.* (1998). The diagnosis and differential diagnosis of Cushing's syndrome and pseudo-Cushing's states. *Endocrine Reviews* **19**, 647–72.

Plotz CM, Knowlton AI, Ragan C (1952). The natural history of Cushing's syndrome. *American Journal of Medicine* **13**, 597–614.

Ross EJ, Linch DC (1982). Cushing's syndrome—killing disease: discriminatory value of signs and symptoms aiding early diagnosis. *Lancet* **ii**, 646–9.

Wallace C *et al.* (1996). Pregnancy-induced Cushing's syndrome in multiple pregnancies. *Journal of Clinical Endocrinology and Metabolism* **81**, 15–21.

Willenberg HS *et al.* (1998). Aberrant interleukin-1 receptors in a cortisol-secreting adrenal adenoma causing Cushing's syndrome. *New England Journal of Medicine* **339**, 27–31.

Zovickian J *et al.* (1988). Usefulness of inferior petrosal sinus venous endocrine markers in Cushing's disease. *Journal of Neurosurgery* **68**, 205–10.

Mineralocorticoids

Botero-Valez M, Curtis JJ, Warnock DG (1994). Brief report: Liddle's syndrome revisited—a disorder of sodium reabsorption in the distal tubule. *New England Journal of Medicine* **330**, 178–81.

Conn JW (1955). Primary aldosteronism: a new clinical syndrome. *Journal of Laboratory and Clinical Medicine* **45**, 6–17.

Edwards CRW *et al.* (1988). Tissue localisation of 11b-hydroxysteroid dehydrogenase-tissue specific protector of the mineralocorticoid receptor. *Lancet* **ii**, 986–9.

Fraser R, Davies DL, Connell JMC (1989). Hormones and hypertension. *Clinical Endocrinology* **31**, 701–46.

Gagner M *et al.* (1997). Laparoscopic adrenalectomy: lessons learned from 100 consecutive procedures. *Annals of Surgery* **226**, 238–46.

Gittler RD, Fajans SS (1995). Primary aldosteronism (Conn's syndrome). *Journal of Clinical Endocrinology and Metabolism* **80**, 3438–41.

Gordon RD *et al.* (1992). Primary aldosteronism: hypertension with a genetic basis. *Lancet* **340**, 159–61.

Hansson JH *et al*. (1995). Hypertension caused by a truncated epithelial sodium channel g subunit: genetic heterogeneity of Liddle syndrome. *Nature Genetics* **11**, 76–82.

Lamberts SWJ *et al*. (1992). Cortisol receptor resistance. The variability of its clinical presentation and response to treatment. *Journal of Clinical Endocrinology and Metabolism* **74**, 313–21.

Lifton RP *et al.* (1992). A chimaeric 11b-hydroxylase/aldosterone synthase gene causes glucocorticoid remediable aldosteronism and human hypertension. *Nature* **355**, 262–5.

Pascoe L *et al.* (1992). Glucocorticoid-suppressible hyperaldosteronism results from hybrid genes created by unequal crossovers between CYP11B1 and CYP11B2. *Proceedings of the National Academy of Sciences, USA* **89**, 8327–31.

Rich GM *et al.* (1992). Glucocorticoid-remediable aldosteronism in a large kindred: Clinical spectrum and diagnosis using a characteristic biochemical phenotype. *Annals of Internal Medicine* **116**, 813–20.

Shimkets RA *et al*. (1994). Liddle's syndrome: heritable human hypertension caused by mutations in the a-subunit of the epithelial sodium channel. *Cell* **79**, 407–14.

Stewart PM *et al.* (1987). Mineralocorticoid activity of liquorice: 11b-hydroxysteroid dehydrogenase deficiency comes of age. *Lancet* **ii**, 821–4.

Stewart PM *et al.* (1995). 11b-Hydroxysteroid dehydrogenase activity in Cushing's syndrome: Explaining the mineralocorticoid excess state of the ectopic ACTH syndrome. *Journal of Clinical Endocrinology and Metabolism* **80**, 3617–20.

White PC, Curnow KM, Pascoe L (1994). Disorders of steroid 11b-hydroxylase isozymes. *Endocrine Reviews* **15**, 421–38.

White PC, Mune T, Agarwal AK (1997). 11b-Hydroxysteroid dehydrogenase and the syndrome of apparent mineralocorticoid excess. *Endocrine Reviews* **18**, 135–56.

Wilson FH *et al.* (2001). Human hypertension caused by mutations in WNK kinases. *Science* **293**, 1030.

Young WF *et al.* (1996). Primary aldosteronism: adrenal venous sampling. *Surgery* **120**, 919–20.

Addison's disease

Addison T (1855). *On the constitutional and local effects of disease of the suprarenal capsules.* S. Highley, London.

Betterle C, Greggio NA, Volpato M (1998). Clinical review 93: Autoimmune polyglandular syndrome type 1. *Journal of Clinical Endocrinology and Metabolism* **83**, 1049–55.

Erturk E, Jaffe CA, Barkan AL (1998). Evaluation of the integrity of the hypothalamo-pituitary adrenal axis by insulin hypoglycaemia test. *Journal of Clinical Endocrinology and Metabolism* **83**, 2350–4.

Oelkers W (1996). Adrenal insufficiency. *New England Journal of Medicine* **335**, 1206–12.

Stewart PM *et al.* (1988). A rational approach for assessing the hypothalamo-pituitary adrenal axis. *Lancet* **i**, 1208–10.

Miscellaneous

Kloos RT *et al.* (1995). Incidentally discovered adrenal masses. *Endocrine Reviews* **16**, 460–84.

12.7.2 Congenital adrenal hyperplasia

I. A. Hughes

Congenital adrenal hyperplasia comprises a family of inherited disorders of adrenal steroidogenesis, characterized by deficiency of cortisol and an accumulation of substrate precursors. A pathophysiological consequence of inadequate cortisol and aldosterone production is ACTH hypersecretion and hyperplastic adrenals. Genital abnormalities are not a universal feature of all forms of congenital adrenal hyperplasia and the original adrenogenital syndrome nomenclature is now seldom used. Figure 1 shows the pathways of adrenal steroidogenesis. The rate-limiting step is the delivery of cholesterol from the outer to the inner mitochondrial membrane to act as substrate for the action of P450scc, a mixed-function oxidase side-chain cleavage enzyme. The intracellular transport of cholesterol is controlled by a number of proteins, including steroid acute regulatory (StAR) protein. The synthesis of cortisol is predominantly controlled by ACTH acting via a G protein-coupled receptor activation of cyclic AMP. Table 1 is a summary of the types of enzymes involved in adrenal steroidogenesis and the known localization of the genes that encode each enzyme. Deficiency of 21-hydroxylase activity is the cause for congenital adrenal hyperplasia in more than 90 per cent of cases.

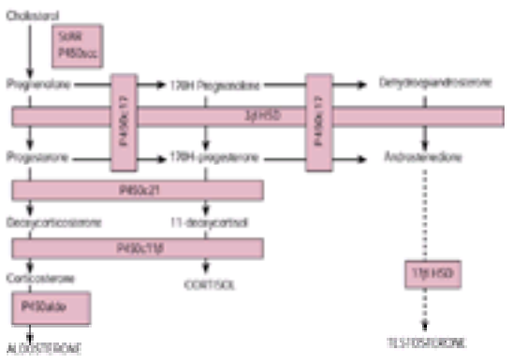

Fig. 1 Pathways of adrenal steroidogenesis. Enzymes involved as defined in the text are represented within the boxes. The dashed line denotes extra-adrenal synthesis of testosterone, catalysed by 17b-hydroxysteroid dehydrogenase. StAR, steroid acute regulatory protein.

21-Hydroxylase deficiency

Clinical presentation

Congenital adrenal hyperlasia due to 21-hydroxylase deficiency is a continuum of disorders which can manifest from birth to adult life (Table 2). The classical form of this enzyme deficiency presents in infancy in one of two forms. The most dramatic is with ambiguous genitalia of the newborn, whereby a female fetus becomes virilized *in utero* due to the effect of excess adrenal androgen converted peripherally to testosterone and masculinizing the external genital anlagen. Milder forms of virilization manifest either as isolated clitoromegaly or as isolated labial fusion. There is evidence that aldosterone biosynthesis is deficient in up to 75 per cent of cases; in affected males, salt loss is initially the sole manifestation as the onset of virilization in males is delayed beyond infancy.

Late-onset or non-classical forms of congenital adrenal hyperplasia are also recognized. There may be delay in the onset of virilization in the non-salt-losing male. The signs of precocious sexual development are also accompanied by tall stature, due to the growth-promoting effect of androgens. The testes remain prepubertal in size (less than 4 ml in volume), which is a useful distinguishing feature from other causes of precocious puberty associated with increased gonadotrophin secretion. The non-classical form of congenital adrenal hyperplasia in females may present with early onset of pubic hair growth or later after puberty with signs of hirsutism and symptoms of menstrual dysfunction. It is important to exclude an adrenal tumour as the cause of late-onset signs of virilization. In adult females, the symptoms and signs are similar to those associated with polycystic ovarian syndrome. Male infertility has also been ascribed to 21-hydroxylase deficiency. Tumours arising from testicular adrenal rests are a complication generally, but not exclusively, in inadequately treated males.

Tests to establish a diagnosis

The characteristic biochemical hallmark is an elevated plasma concentration of 17OH-progesterone, generally greater than 300 nmol/l (normal less than 10 nmol/l). Newborn screening for congenital adrenal hyperplasia is established in several countries in an effort to identify affected males before the onset of life-threatening salt loss. The world-wide incidence of 21-hydroxylase deficiency is approximately 1 in 14 000 births.

The diagnosis of late-onset 21-hydroxylase deficiency requires an ACTH stimulation test. Studies of women with signs of hyperandrogenism show only about 5 per cent of hormone profiles consistent with late-onset congenital adrenal hyperlasia. Polycystic ovarian syndrome is a well recognized and more frequent cause of hirsutism and infertility, although ultrasonographic evidence of polycystic ovaries is common in congenital adrenal hyperplasia. The disease frequency for the late-onset form is estimated as 0.1 per cent of the general population, but as high as 3 to 4 per cent of Ashkenazi Jews.

Treatment in early infancy, childhood, and adolescence

Glucocorticoid replacement in the form of hydrocortisone should be started; if the infant is in salt-losing crisis, parenteral hydrocortisone with intravenous saline and dextrose is needed. Hypoglycaemia may be a problem. The cortisol secretion rate is about 6 to 8 mg/m^2 per day so that an oral hydrocortisone dose no more than 15 to 20 mg/m^2 per day should be used initially. For salt-losers, aldosterone replacement is given in the form of 9a-fludrocortisone (florinef), 100 to 150 µg daily. Salt added to the feeds is sometimes needed during the first few months of life.

Surgical reconstruction of the external genitalia is required for the virilized female infant at 6 to 12 months of age. The procedure involves reducing the size of the clitoris and exposing the opening of the vagina on to the perineum. Further assessment of the vaginal opening is undertaken around puberty to determine whether treatment with dilators is required. Ongoing medical treatment is adjusted according to body surface area. The hydrocortisone dose should be trebled in times of stress, and parenteral steroids given to cover surgical procedures. Other glucocorticoid preparations used include cortisone acetate, prednisolone, and dexamethasone. Prednisolone and dexamethasone are both longer-acting glucocorticoids which appear to be useful for treatment in the postpubertal age group. Dexamethasone can be used as a single daily dose and appears to regulate menses more efficiently in the young adult female patient. Some girls, however, develop

side-effects, such as excessive weight gain and striae formation on quite small doses of dexamethasone. Dexamethasone is 80 to 100 times more potent than hydrocortisone in suppressing 170H-progesterone levels.

Monitoring treatment

The rate of linear growth is the main clinical yardstick of control before puberty (Table 3). Growth is normally rapid during infancy; over-treatment during this period may be a factor leading to reduced final height and to obesity in childhood. Growth monitoring is supplemented by bone age measurements. Androgens stimulate epiphyseal maturation through local conversion to oestrogens so that advanced skeletal maturation inevitably reduces the length of growing time and leads to short adult stature. The onset of the pubertal growth spurt is a milestone to monitor in both sexes, while delayed menarche in girls indicates inadequate control and increased plasma testosterone concentrations. In adults, regular menses and ovulation in the female and normal spermatogenesis in the male are reliable clinical indicators of adequate control.

Some of the tests listed in Table 3 are used as additional markers of control. There is a marked diurnal rhythm in 17OH-progesterone so that single random measurements are inadequate to monitor control. A daily profile is a useful measure of control, especially as capillary blood spot and saliva assays are available. Androstenedione can also be used as a marker of control. Random plasma testosterone measurements are useful to monitor control in infants, children, and adult females, but not in pubertal boys and adult males because of testosterone secretion by the testis. Serum electrolytes are an insensitive index of the adequacy of mineralocorticoid replacement, but renin measurement and blood pressure recordings should be undertaken routinely after infancy. An elevated plasma renin indicates the need for 9a-fludrocortisone treatment even if there has been no overt salt loss. Salt losers invariably like a salty diet even when adequately replaced with mineralocorticoid.

Outcome in treated 21-hydroxylase deficiency

The survival rate for the salt-wasting form of 21-hydroxylase deficiency has improved in recent years. However, an unequal sex incidence suggests that male infants still die in infancy from an unrecognized adrenal crisis. Achieving normal growth is a problem in management and patients rarely reach their predicted adult height. A recent clinical trial of lowering the glucocorticoid dose by blocking the action of androgen with an antiandrogen and the conversion of androgen to oestrogen with an aromatase inhibitor is showing promising results for improved growth.

Fertility is reduced in adult females. Contributory factors include inadequate vaginal introitus, and anovulatory cycles from increased progesterone concentrations acting as a 'mini-pill'. These problems are remediable and pregnancy rates are now improving. Nevertheless, there is evidence of lower maternalism and heterosexuality in adult females. Those who do become pregnant need careful monitoring to maintain steroid levels within the pregnancy-related range. Delivery is usually by caesarean section, and the offspring are normal even if testosterone levels increase slightly during pregnancy. Adult males who stop taking glucocorticoid replacement may develop oligospermia; this is usually reversible if treatment is restarted. Testicular tumours may also occur from ACTH hyperstimulation of testicular adrenal-rest cells. Laparoscopic adrenalectomy is a radical form of treatment in patients recalcitrant to medical management, particularly females with persistent signs of virilization.

Genetics of 21-hydroxylase deficiency

Congenital adrenal hyperplasia is an autosomal recessive condition; the *CYP21* gene is closely linked to the major histocompatibility complex on the short arm of chromosome 6. Studies of HLA haplotypes show an association between the uncommon A3, Bw47, DR7 haplotype in the classical salt-losing form, whereas in the non-classical late-onset form, there is a strong association with HLA-B14, DR1.

Two genes, *CYP21* and a pseudogene, *CYP21P*, were identified within the class III region of the HLA complex on chromosome 6p21.3 and are approximately 30 kb apart. They are adjacent to, and in tandem repeat with *C4A* and *C4B* genes which encode for the fourth component of serum complement (Fig. 2). *CYP21P* is functionally inactive because of a series of deleterious mutations, but is 97 per cent homologous with the active gene. Each gene comprises 10 exons. Misalignment and unequal crossing over between sister chromatids during meiosis leads to a major gene deletion. This is always associated with the severe, salt-losing form of congenital adrenal hyperplasia. The frequency of gene deletions as a cause of 21-hydroxylase deficiency is about 25 per cent and is highest in northern European populations. Another frequent genotype is associated with gene conversion events in which there is non-reciprocal transfer of multiple mutations from *CYP21P* to the active *CYP21* gene. Such large-scale conversions may account for a further 10 to 15 per cent of cases, all manifesting with the severe, salt-losing form. The majority of gene conversion events are small scale in nature. Several point mutations have now been identified while linked microsatellites are useful for prenatal diagnosis when the family genotype has previously been ascertained.

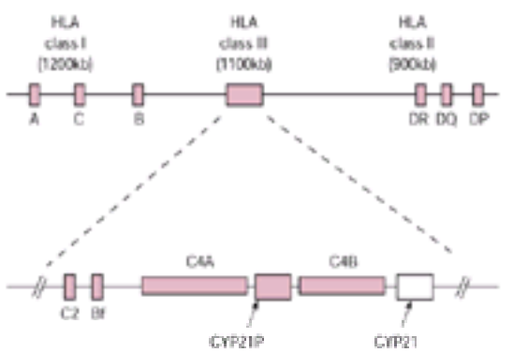

Fig. 2 Schema of the HLA gene region on chromosome 21p. The active 21-hydroxylase gene is *CYP21*, the inactive pseudogene, *CYP21P*. The two genes are in tandem repeat with complement C4A and C4B genes.

A sufficient number of alleles have now been studied to indicate general concordance between genotype and phenotype. The mutations which cause more than 90 per cent of cases of 21-hydroxylase deficiency are shown in Fig. 3 in relation to the expected phenotype. The most common mutation in classic 21-hydroxylase deficiency affects mRNA splicing due to a nucleotide base change (A/C to G) in the second intron. A stretch of nucleotides which is normally spliced out is retained, so that the translational reading frame is altered and an inactive protein synthesized. Most patients with this mutation have the salt-losing form of congenital adrenal hyperplasia, but some patients who are homozygous for this mutation are salt replete. Presumably, enough normally spliced mRNA is generated to produce some enzyme activity. Other examples leading to salt loss are shown in Fig. 3. *In vitro* functional assays of wild type and mutant CYP21 enzymes using progesterone and 170H-progesterone as substrate show total absence of enzyme activity for mutations leading to salt loss. A specific mutation associated with non-salt-wasting occurs in exon 4, changing isoleucine to an asparagine (Ile172Asn). This mutation results in an enzyme with about 1 per cent of normal activity, sufficient for adequate aldosterone production.

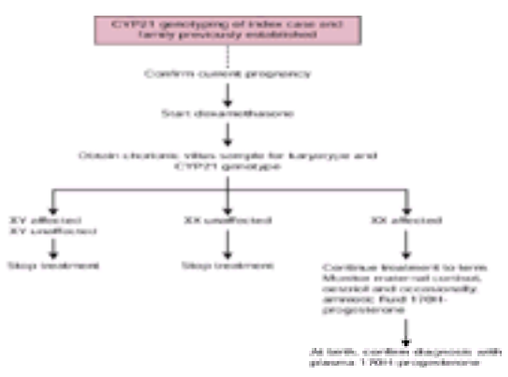

Fig. 3 Genotype: phenotype correlations for the 10 most frequent causes of 21-hydroxylase deficiency. E6 cluster refers to three mutations (Ile 236 Asp, Val 237 Glu, Met 239 Lys) in exon 6.

The non-classical or late-onset form of 21-hydroxylase deficiency is associated with a mutant enzyme which has 50 per cent and 20 per cent of normal activity *in vitro* with substrates 17OH-progesterone and progesterone, respectively. An example is Val251Leu in exon 7 which is associated with the haplotype HLA-B14, DR1. This single mutation accounts for the majority of non-classical cases of congenital adrenal hyperplasia, and is most frequently found in East European Jews. Genetic studies in 21-hydroxylase deficiency show that many patients are compound heterozygotes. In general, the phenotype reflects the less deleterious mutation.

Prenatal diagnosis and treatment

Chorionic villus sampling and molecular analysis of the *CYP21* gene has enabled an earlier and more reliable diagnosis to be made. Furthermore, there is the option of offering prenatal treatment to prevent virilization of an affected female fetus. Figure 4 outlines the current protocol used for the prenatal diagnosis and treatment of 21-hydroxylase deficiency.

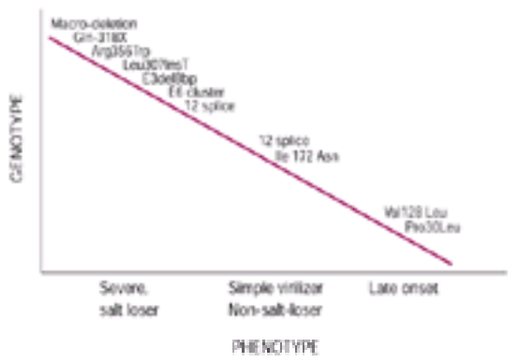

Fig. 4 Outline protocol for prenatal management of 21-hydroxylase deficiency.

Maternal dexamethasone treatment is started early as fetal adrenal steroidogenesis is established by the eighth week of gestation. *CYP21* genotyping of the index case, parents, and unaffected siblings should have been performed previously. DNA analysis is then more reliable, especially if the straightforward technique of linked microsatellite markers is used. Dexamethasone is chosen as the glucocorticoid since, unlike hydrocortisone and prednisolone, this steroid is not inactivated by placental 11b-hydroxy-steroid dehydrogenase. The conventional starting dose is 20 µg/kg per day based on prepregnancy bodyweight, administered in three divided doses. Once the diagnosis has been confirmed by molecular genetic analysis, treatment is only continued to term in the case of an affected female fetus. Thus, seven out of eight fetuses will be exposed unnecessarily to dexamethasone for about 6 weeks during early gestation. Fetal adrenal suppression is monitored by serial measurements of maternal plasma or urinary oestriol concentrations. This steroid metabolite is formed as a result of placental aromatization of weak androgen substrates produced uniquely by the fetal adrenal gland. More direct evidence of adrenal suppression can be obtained by collecting amniotic fluid for measurement of 170H-progesterone and testosterone.

The outcome of prenatal treatment is satisfactory in the majority of cases when treatment is started early and continues uninterrupted to term. Thus, the external genitalia in affected females are completely normal or so mildly affected that surgery is not required. There have been isolated reports of other abnormalities in treated infants but no specific pattern is emerging to suggest that prenatal dexamethasone treatment is detrimental to postnatal growth and development. This treatment to prevent a congenital malformation is still experimental and should be undertaken as part of a clinical trial which includes long-term outcome surveillance. Maternal side-effects of the treatment include excessive weight gain, striae formation, hypertension, and glucose intolerance.

Other forms of congenital adrenal hyperplasia

P450scc deficiency

This extremely rare form of the condition has also been variously called cholesterol desmolase deficiency and lipoid adrenal hyperplasia. The production of all classes of steroid hormones is deficient in gonadal as well as in adrenal tissue. Affected males have female external genitalia because of failure to synthesize testosterone. Affected 46XX females have developed puberty spontaneously suggesting the enzyme deficiency is less severe in the ovary compared with the testis.

The gene encoding P450scc is designated *CYP11A* and it was assumed that mutations in this gene caused lipoid adrenal hyperplasia. However, no mutations were found in several cases studied. Furthermore, adrenal tissue studied from an affected patient showed normal P450scc cDNA. When the human StAR gene was cloned, subsequent studies identified deleterious StAR mutations in patients with this form of congenital adrenal hyperplasia. The disorder is most common in the Japanese population; a premature stop codon at amino acid 258 substituting for glutamine is a frequently reported mutation. The StAR gene is expressed in the adrenals and gonads, but not the placenta, explaining why congenital lipoid adrenal hyperplasia is not a lethal disorder. Further analysis of the P450scc gene in a patient with lipoid adrenal hyperplasia who had a normal StAR gene, revealed a heterozygous mutation. The child had survived 4 years before developing an acute adrenal crisis.

3b-Hydroxysteroid dehydrogenase deficiency

3b-Hydroxysteroid dehydrogenase/isomerase (3bHSD) is a non-P450 membrane-bound enzyme which converts ‡^{5} to ‡^{4} steroids in the adrenals and gonads. Hence it is needed for the synthesis of glucocorticoids, mineralocorticoids, progesterone, androgens, and oestrogens. Two highly homologous genes, *HSD3B1* and *HSD382*, control the expression of two human isoenzymes. Type I enzyme is expressed predominantly in the placenta and peripheral tissues, whereas type II is expressed in the adrenals and gonads. Deficiency of 3bHSD activity also causes severe glucocorticoid and mineralocorticoid deficiency. Genital abnormalities occur, mainly in males because of the production of weak androgens by the testis. Diagnosis is confirmed by an elevated ratio of ‡ ^{5} (170H-pregnenolone) to ‡^{4} (170H-progesterone) steroids and analysis of urinary steroid metabolites.

Molecular studies show the majority of patients have missense mutations in the *HSD3B2* gene. Nevertheless, there is significant conversion of ‡ ^{5} to ‡^{4} steroids in peripheral tissues through the action of type I 3bHSD. Consequently, some patients have elevated levels of ‡ ^{4} steroids (17OH-progesterone, androstenedione), which has led to a mistaken diagnosis of 21-hydroxylase deficiency. Extraglandular steroid synthesis accounts for 40 per cent of androgen production in adult males and most oestrogen production in prepubertal and postmenopausal females. Spontaneous onset of puberty and menarche are reported in females with 3bHSD deficiency.

17a-Hydroxylase deficiency

A single P450c17 enzyme catalyses 17a-hydroxylase and 17,20-lyase reactions. Both are required for the synthesis of sex hormones (C19 steroids), whereas only 17a-hydroxylase activity is required to synthesize cortisol (C21, 17-hydroxysteroids). Mineralocorticoid biosynthesis is not dependent on the presence of the P45017a enzyme, so that ACTH-stimulated, low renin hypertension is a typical feature of P45017a hydroxylase deficiency due to excess production of C21, 17-deoxysteroids such as aldosterone. There is an accompanying hypokalaemic, metabolic alkalosis. Inadequate androgen in affected males causes a phenotype ranging from female genitalia to an ambiguous appearance or features of a hypospadic male. Females lack breast development and have primary amenorrhoea. The P450c17 enzyme operates as a qualitative regulator of steroidogenesis using pregnenolone as substrate.

Increased corticosterone, deoxycorticosterone, and progesterone and decreased levels of testosterone, oestradiol, and renin characterize this enzyme defect.

Measurements of steroid metabolites delineate patterns indicative of 17a-hydroxylase or 17,20-lyase deficiency alone, or combined. Most patients have a mutant P450c17 which leads to loss of both enzyme activities.

The 6.5-kb human *CYP17* gene comprises eight exons. A frequent mutation is a 4-bp duplication in exon 8 which, as a result of altering the reading frame, leads to a shortened carboxy-terminal sequence. Expression studies of the mutant protein show absence of both 17a-hydroxylase and 17,20-lyase activities. This is consistent with biochemical findings and a clinical female phenotype in affected males and females. The differential catalytic activity of this enzyme is manifest at adrenarche with the development of the zona reticularis and increased 17,20-lyase activity. This causes increased concentrations of dehydroepiandrosterone (DHEA) and its sulphate (DHEAS), independent of any change either in ACTH or cortisol levels. The increase in adrenal androgens may lead to early onset of pubic and axillary hair, body odour, and a moderate advance in skeletal maturation. Premature adrenarche must be differentiated from late-onset congenital adrenal hyperplasia or an adrenal tumour. Studies on the regulation of 17,20-lyase activity suggests a role for P450c17 phosphorylation in mediating the increased differential enzyme activity occurring at the time of adrenarche.

11b-Hydroxylase deficiency

Deficiency of 11b-hydroxylase activity accounts for about 5 per cent of cases of congenital adrenal hyperplasia. The enzyme is required for the terminal conversion of 11-deoxycortisol to cortisol and deoxycorticosterone to corticosterone. Consequences of increased ACTH stimulation are salt and water retention, low-renin hypertension, and virilization secondary to the increased production of deoxycorticosterone and adrenal androgens. Newborn females and infant males are more profoundly virilized than is the case with 21-hydroxylase deficiency. Prepubertal breast development is another specific and unexplained feature. Hypertension can develop in early childhood, the severity not necessarily correlating with plasma levels of deoxycorticosterone.

The diagnosis is confirmed by elevated concentrations of 11-deoxy-cortisol and deoxycorticosterone in plasma and their tetrahydro metabolites in urine. Plasma concentrations of androstenedione and testosterone are increased. Moderately elevated levels of 17OH-progesterone may lead to an erroneous diagnosis of 21-hydroxylase deficiency. Treatment requires only glucocorticoid replacement, although transient salt-wasting may follow an initial fall in levels of the potent mineralocorticoid, deoxycorticosterone. Antihypertensive treatment may be necessary if hypertension has been long-standing. Milder or late-onset deficiency occurs and manifests similar features to the late-onset form of 21-hydroxylase deficiency.

11b-Hydoxylase activity is a function of the *CYP11B1* gene which comprises nine exons and is located on chromosome 8q22. The gene is highly expressed in the adrenals; transcripts are controlled by ACTH and to a lesser extent, by angiotensin II. Located on the same chromosome, at a distance of about 40 kb, is the highly homologous *CYP11B2* gene which encodes aldosterone synthase (also referred to as corticosterone methyloxidase II) which catalyses the conversion of corticosterone via 18OH-corticosterone to aldosterone.

Mutations throughout the *CYP11B1* gene cause 11b-hydroxylase deficiency. The majority are missense mutations with some clustering occurring in exons 2, 6, 7, and 8. The highest incidence of this form of congenital adrenal hyperplasia occurs in an inbred population of Sephardic Jews in Morocco. A single amino acid substitution (Arg448His) is the cause. This alters the haeme-binding sequence which is a unique and conserved feature of all cytochrome P450 enzymes. Prenatal treatment with dexamethasone has also been used successfully in this form of congenital adrenal hyperplasia to prevent virilization of an affected female fetus. The late-onset or non-classic form of 11b-hydroxylase deficiency has been found in one study to be associated with compound heterozygote mutations of the *CYP11B1* gene. Studies of hirsute women with marginally elevated levels of 11-deoxycortisol have not revealed mutations in the *CYP11B1* gene.

Further reading

Acerini CL, Hughes IA (2000). 21-Hydroxylase deficiency defects and their phenotype. In: Hughes IA, Clark AJL, eds. *Adrenal disease in childhood. Clinical and molecular aspects*, pp. 93–111. Karger, Basel.

Alizai NK, Thomas DF, Lilford RJ, Batchelor AG, Johnson N (1999). Feminizing genitoplasty for congenital adrenal hyperplasia: what happens at puberty? *Journal of Urology* **161**, 1588–91.

Bose HS, Sugarawa T, Strauss JF III, Miller WL (1996). The pathophysiology and genetics of congenital lipoid adrenal hyperplasia. *New England Journal of Medicine* **335**, 1870–8.

Cerame BL, Newfield RS, Pascoe L, *et al*. (1999). Prenatal diagnosis and treatment of 11b-hydroxylase deficiency congenital adrenal hyperplasia resulting in normal female genitalia. *Journal of Clinical Endocrinology and Metabolism* **84**, 3129–34.

Forest MG, Morel Y, David M (1998). Prenatal treatment of congenital adrenal hyperplasia. *Trends in Endocrinology and Metabolism* **9**, 284–9.

Hughes IA (1998). Congenital adrenal hyperplasia—a continuum of disorders. *Lancet* **352**, 752–4.

Jaaskelainen J, Vouitilainen R (1997). Growth of patients with 21-hydroxylase deficiency: an analysis of the factors influencing adult height. *Pediatrc Research* **41**, 30–3.

Jaaskelainen J, Lavo A, Voutilainen R, Partanen J (1997). Population-wide evaluation of disease manifestation in relation to molecular phenotype in steroid 21-hydroxylase (CYP21) deficiency: good correlation in a well defined mutation. *Journal of Clinical Endocrinology and Metabolism* **8**, 3293–7.

Lajic S, Wedell A, Biu T-H, Ritzen EM, Holst M (1998). Long-term somatic follow-up of prenatally treated children with congenital adrenal hyperplasia. *Journal of Clinical Endocrinology and Metabolism* **83**, 3872–80.

Lo JC, Schwitzgebel VM, Tyrrell JB, *et al*. (1999). Normal female infants born of mothers with classic congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Journal of Clinical Endocrinology and Metabolism* **84**, 930–6.

Meyer-Bahlburg HFL (1999). What causes low rates of child-bearing in congenital adrenal hyperplasia? *Journal of Clinical Endocrinology and Metablism* **84**, 1844–7.

Miller WL (1998). Prenatal treatment of congenital adrenal hyperplasia: a promising experimental therapy of unproven safety. *Trends in Endocrinology and Metabolism* **9**, 290–2.

Miller WL, Auchus RJ (2000). Biochemistry and genetics of human P450c17. In, Hughes IA, Clark AJL, eds. *Adrenal disease in childhood. Clinical and molecular aspects*, pp. 63–92. Karger, Basel.

Moisan AM, Ricketts ML, Tardy V, *et al*. (1999). New insight into the molecular basis of 3b-hydroxysteroid dehydrogenase deficiency: identification of eight mutations in the HSD3B2 gene in eleven patients from seven new families and comparison of the functional properties of twenty-five mutant enzymes. *Journal of Clinical Endocrinology and Metabolism* **84**, 4410–25.

Pang S, Wallace AM, Hofman L, *et al*. (1998). Worldwide experience in newborn screening for classical congenital adrenal hyperplasia due to 21- hydroxylase deficiency. *Pediatrics* **81**, 866–74.

Premawardhana LDKE, Hughes IA, Read GF, Scanlon MF (1997). Longer term outcome in females with congenital adrenal hyperplasia (CAH): the Cardiff experience. *Clinical Endocrinology* **46**, 327–32.

Tajima T, Fujieda K, Kouda N, Nakae J, Miller WL (2001) Heterozygous mutation in the cholesterol side chain cleavage enzyme (p450scc) gene in a patient with 46,XY sex reversal and adrenal insufficiency. *Journal of Clinical Endocrinology and Metabolism* **86**, 2820-5..

Van Wyk JJ, Gunther DF, Ritzen EM, *et al*. (1997). The use of adrenalectomy as a treatment for congenital adrenal hyperplasia. *Journal of Clinical Endocrinology and Metabolism* **81**, 3180–9.

White PC, Speiser PW (2000). Congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Endocrine Reviews* **21**, 245–91.

White PC, Curnow KM, Pascoe I (1994). Disorders of steroid 11b hydroxylase enzymes. *Endocrine Reviews* **15**, 421–38.

Zucker KJ, Bradley SJ, Oliver G, Blake J, Fleming S, Hood J (1996). Psychosexual development of women with congenital adrenal hyperplasia. *Hormones and Behaviour* **30**, 300–18.

12.8.1 Ovarian disorders

H. S. Jacobs

Approach to the patient with ovarian disorders

Synchronization of the changes in the ovary and uterus and the hypothalamic–pituitary unit is complex and, not surprisingly, the ovulatory cycle is vulnerable to disturbances at any of the levels of endocrine organization. Ovarian cyclicity may also be disrupted by a deterioration in general health, a protective mechanism that avoids reproduction occurring in circumstances adverse to fetal development.

In obtaining the history, information about the regularity of the menstrual cycle is relevant in women concerned about fertility because the chance of conception is directly related to the rate of ovulation. If a woman ovulates only six rather than the usual 12 to 13 times a year, it will take her twice as long to conceive as a woman of the same age with a regular monthly cycle. Oligomenorrhea (interval between menstrual periods of more than 6 weeks but less than 6 months) is usually a consequence of the polycystic ovary syndrome (see later); amenorrhoea (no periods for more than 6 months) has a broader spectrum of causes (Table 1 and Table 2). The duration of a menstrual disturbance is relevant both to its cause and its consequences. For example, a history of delayed menarche implies that the disturbance was present from the age of 12, a common finding in women with polycystic ovary syndrome. A history of weight fluctuation is important because weight loss, often in the context of mild (and denied) anorexia nervosa, is a common cause of amenorrhoea and weight increase a common precipitant of the clinical expression of polycystic ovary syndrome. Bulimia may complicate either of the above conditions and, like anorexia nervosa, requires specialist management in its own right. It is always relevant to inquire about a family history of the patient's complaint since an increasing number of mono- and polygenic causes of reproductive disturbance are now recognized.

The association of amenorrhoea with galactorrhoea implies hyperprolactinaemia and an association with symptoms of oestrogen deficiency (flushing and sweating attacks and/or vaginal dryness and discomfort during intercourse) implies an increased risk of osteoporosis. Symptoms of hyperandrogenism (seborrhoea, acne, and excessive hair growth) imply increased production of androgens, which, in most cases, is caused by polycystic ovary syndrome. It is preferable to use the term 'unwanted hair' rather than 'hirsutism' to avoid provoking a debate with the patient about what is and what is not excessive. On the other hand, whether the unwanted hair is severe enough to be helped by present methods of treatment is a matter for the physician's judgement. Some women express a concern, which should not be trivialized, that the development of unwanted hair means they are turning into a man. They need a clear explanation of the underlying disorder and reassurance that treatment is available.

If the patient has previously used an oral contraceptive, one should determine whether it was primarily for contraception or had been prescribed to correct a menstrual disturbance. In the latter case, it is important to appreciate that treatment with an oral contraceptive does not cure such problems, which are therefore likely to recur when it is stopped. Oral contraceptives do not cause amenorrhoea after discontinuation. Amenorrhoea occurring after discontinuation of an oral contraceptive therefore needs the same investigation as amenorrhoea temporally unrelated to previous use of an oral contraceptive.

Most women with menstrual disturbances want reassurance about their present or future fertility. Their age is clearly of great importance in this evaluation which will also have to determine the extent to which failure to ovulate is an adequate explanation of the failure to conceive. Many older women need advice about the long-term effects of oestrogen deficiency and the wisdom of hormone replacement therapy.

Amenorrhoea

While the classification is usually into primary (the patient has never had a menstrual period) or secondary amenorrhoea (interval between periods exceeds 6 months) most of the common causes can present as either, as seen in Table 1 and Table 2. With the exception of structural abnormalities, such as an absent uterus, differences between primary and secondary amenorrhoea are outweighed by similarities so here they are considered primarily in terms of their aetiology.

In young women with primary amenorrhoea there may have been congenital abnormalities in the development of the ovaries, genital tract, or external genitalia or a perturbation of the normal process of puberty (Chapter 12.9.3). Investigation is appropriate when menstruation has not occurred by the age of 16 in the presence of normal secondary sexual development or by the age of 14 in its absence.

Developmental abnormalities

Developmental abnormalities of the mullerian duct, external genitalia, and the problems of intersexual abnormalities are dealt with elsewhere.

Gonadal dysgenesis

The commonest cause of gonadal dysgenesis is Turner's syndrome, in the severest form of which a 45XO karyotype is associated with a characteristic phenotypic appearance. Typically patients are of short stature, with cubitus valgus, webbed neck, low hairline, shield chest, and widely spaced nipples. The palate is often arched and the fourth metacarpal short. Lymphoedema, multiple pigmented naevei, and hearing loss are common. Coarctation of the aorta occurs in 10 to 20 per cent and hypertension independent of that abnormality occurs more commonly than in the normal population. The condition occurs in about 1 in 2500 female births. Spontaneous and, indeed, ovulatory cycles occasionally occur, particularly if there is chromosomal mosaicism, but in the long term premature ovarian failure is inevitable. It is important to determine the karyotype as the presence of a Y chromosome in an individual with gonadal dysgenesis means residual gonadal tissue must be removed because of an increased risk of malignancy.

Serum gonadotrophin concentrations are elevated compared with those of normal girls of the same age and may approach the menopausal range. Oestrogen levels are low, the uterus is small and bone densitometry shows skeletal decalcification; a history of spontaneous fracture is common. In addition to cardiovascular and renal assessment, autoimmune thyroiditis and diabetes mellitus should be excluded.

Management includes initiation of low-dose oestrogen therapy (starting with no more than 5 µg of ethinyl oestradiol per day) to promote breast development without prejudicing linear growth. The dose is gradually raised over 12 to 18 months. Maintenance therapy is with a cyclical oestrogen–progestogen preparation (such as an

oral contraceptive), as regular withdrawal bleeding is necessary to prevent endometrial hyperplasia and the risk of malignancy. While it is now possible to provide fertility to women with Turner's syndrome through ovum donation, shortage of oocytes remains an important and usually critically limiting factor. Given the prevalence of hypertension and cardiovascular disorder, careful medical assessment before referral is required.

Hypothalamic causes of amenorrhoea

Hypothalamic hypogonadotrophic hypogonadism may be functional or organic and can occur in an isolated form or in association with more widespread endocrine disorders, as in patients with infiltrating (sarcoidosis, tuberculosis), expanding (craniopharyngioma), or traumatic (head injury) lesions of the hypothalamus. Hypogonadotrophic hypogonadism that is potentially reversible occurs in primary and secondary iron overload syndromes and is a frequent presenting feature of juvenile haemochromatosis. Lesions of the pituitary stalk are increasingly recognized.

Idiopathic hypogonadotrophic hypogonadism is characterized by low serum levels of gonadotrophins and gonadal steroids in the absence of structural defects of the hypothalamic–pituitary axis. Idiopathic hypogonadotrophic hypogonadism usually results from a congenital defect of secretion of gonadotrophin-releasing hormone, although mutations of the gene encoding the gonadotrophin-releasing hormone receptor have also been described.

Hypogonadism with anosmia. In a number of idiopathic hypogonadotrophic hypogonadism pedigrees mis- and non-sense mutations and partial and complete deletions have been identified in a gene located in the distal part of the short arm of the X chromosome (Xp22.3). These mutations are associated with hypogonadotrophic hypogonadism and anosmia, as features of Kallman's syndrome in boys and men. This (*KAL*) gene lies next to the steroid sulphatase locus, so patients may exhibit a contiguous gene syndrome comprising complete X-linked Kallman's syndrome and ichthyosis. The *KAL* gene encodes a molecule with similarities to proteins involved in neural cell adhesion and axonal pathfinding. X-linked Kallman's syndrome arises from a defect of embryonic migration of olfactory and gonadotrophin-releasing hormone producing neurones from the anlage of the olfactory lobes into the hypothalamus. In affected females the disorder is usually inherited as an autosomal recessive or dominant trait, generally with incomplete penetrance. The molecular cause in these cases is not known.

In women Kallman's syndrome presents with delayed puberty and primary amenorrhoea. Characteristically the patient cannot smell curry or the difference between tea and coffee. There are signs of marked oestrogen deficiency, usually amounting to sexual infantilism. Adult stature is normal. Mirror movements, in which voluntary movements of one limb are associated with involuntary, non suppressible and homologous mirror movements of the contralateral limb, occur in 85 per cent and unilateral renal agenesis in 31 per cent of X-linked cases.

Investigations reveal subnormal serum gonadotrophin and oestradiol concentrations, on pelvic ultrasound the uterus and ovaries are small, and bone densitometry reveals skeletal demineralization (Fig. 1). Renal imaging may show an absent kidney; MRI of the brain shows absent or abnormal olfactory bulbs and sulci, with a normal pituitary gland.

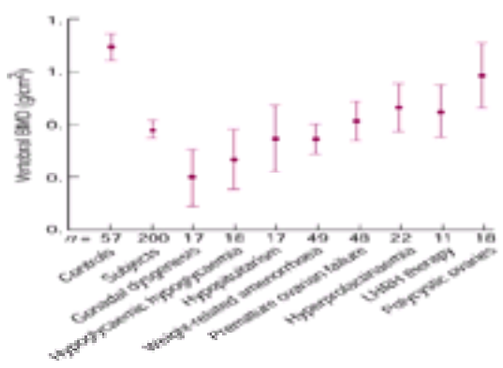

Fig. 1 Vertebral bone mineral density measurements in 200 young women with amenorrhoea, arranged by diagnosis. With the exception of the bone mineral density measurements in the women with polycystic ovary syndrome, the bone mineral density measurements of all other groups were significantly below those of the controls.

Management of idiopathic hypogonadotrophic hypogonadism involves induction of pubertal maturation, initially with small doses of oestrogen (not more than 5 µg/day of ethinyl oestradiol) to optimize breast development. Delay in recognition and treatment impairs full development and so has important psychosocial implications. Surgical referral for breast augmentation is appropriate if development remains inadequate after a year's treatment with gradually increasing doses of oestrogen. Maintenance hormone treatment is with an oral contraceptive preparation. Bone densitometry should be repeated to ensure that the dose of oestrogen is adequate. While an increase of 5 to 6 per cent per year for 2 to 3 years is expected, delay in diagnosis may result in failure to attain normal peak bone mass. Treatment with pulsatile gonadotrophin-releasing hormone or gonadotrophin injections allows an excellent prognosis for fertility.

Functional hypogonadotrophic hypogonadism

Weight-related amenorrhoea

Amenorrhoea is usual when the body fat content falls below 17 per cent or the body mass index falls below 19 kg/m ^{2}. The latter figure assumes an average amount of exercise; for competitive athletes, particularly those in track events, the figure would be higher. Depending on the age of onset, the patient may present with primary or secondary amenorrhoea.

The neuroendocrine mechanisms that signal this critical level of nutritional reserve are uncertain but centre on reduced secretion of leptin which results in impaired secretion of gonadotrophins, particularly luteinizing hormone. Serum insulin and insulin-like growth factor 1 levels are low so there is also reduced insulin drive to the ovaries. The consequent anovulation is protective because it avoids reproduction occurring in adverse circumstances. Quite apart from the obstetric evidence of premature delivery of immature babies to women who are underweight, there is also a link between deficient fetal nutrition and an increased prevalence in adults of cardiovascular and pulmonary disease.

Self-imposed starvation

The cause is often anorexia nervosa (Chapter 26.5.5). In patients presenting with amenorrhoea, loss of appetite is usually denied and the patient often draws attention to what she regards as a normal pattern of eating. These women characteristically maintain a good appetite which they have to suppress. In contrast, some women lose weight as a feature of the anorexia of depression, occasioned perhaps by the breakup of a relationship; in this condition the patient does lose her appetite. The prognosis for a return of normal weight and spontaneous menstrual cycles is much better than that in women with anorexia nervosa.

Exercise-related amenorrhoea

Amenorrhoea is common in ballet dancers and high-performance athletes, particularly during phases of intensive training and performance. The amenorrhoea is associated with reduced body weight and body fat content. Psychological stress is unlikely to be a major determinant because amenorrhoea is not common in Olympic swimmers, who, though as competitive as track athletes, have a normal body fat content.

Involuntary starvation

Worldwide, the most important causes of starvation are social disintegration, famine, and war. Increasingly we recognize that the deleterious physical effects of starvation are passed to the next generation, *inter alia*, through the adverse effects of fetal malnutrition on adult health.

Altered absorption

In patients with malabsorption, for instance in women with cystic fibrosis, amenorrhoea is associated with reduced body mass index and body fat content and resolves when nutrition improves.

Investigation of weight-related amenorrhoea

The diagnosis is made in part by identifying that the patient's weight is subnormal for her height, bearing in mind the effect of high-performance training on replacing body fat with more dense muscle. It is also made by exclusion, because self-imposed weight loss is common in young women and may coexist with other conditions.

Women with weight-related amenorrhoea have subnormal serum gonadotrophin (particularly luteinizing hormone) and oestradiol concentrations, with small ovaries and a small uterus on ultrasound scanning. Skeletal demineralization is the rule, except in those with coexisting polycystic ovaries (see later). Serum markers of bone resorbtion are low, indicating that, in contrast to the osteoporosis of oestrogen deficiency, this is a 'low-turnover' osteoporosis. The most likely cause is reduced osteotrophic stimulation by insulin-like growth factor 1 and leptin.

Management involves explanation, identification of psychiatric conditions for which specialist advice is needed, and correction of oestrogen deficiency. Optimally the latter is achieved through weight gain. Oestrogen deficiency may, however, be so severe that if the patient is unable to put on weight, it may be preferable for her to take oestrogen rather than have the clinician adopt a purist position. Bone mineral density, however, only improves when the patient gains weight: oestrogen treatment is ineffective. So far as fertility is concerned, medical induction of ovulation should be eschewed until the patient's weight has returned to normal so that nutritional risks to the unborn child (and adult) have been minimized. In the event, when the body mass index returns to normal, induction of ovulation is rarely required.

Pituitary causes of amenorrhoea

While a non-functioning pituitary tumour sometimes causes amenorrhoea, the commonest pituitary cause is hypersecretion of prolactin. The subject is discussed further in Chapter 12.2.

Ovarian causes of amenorrhoea

Primary ovarian failure

Primary ovarian failure occurs normally at the menopause ('age appropriate primary ovarian failure') because of the process of atresia (Fig. 2) that results in almost complete depletion of oocytes and follicles by about the age of 50. If the rate of atresia is faster than normal, the causes of which are discussed below, depletion of oocytes and follicles occurs prematurely and 'age inappropriate' or 'premature' ovarian failure develops.

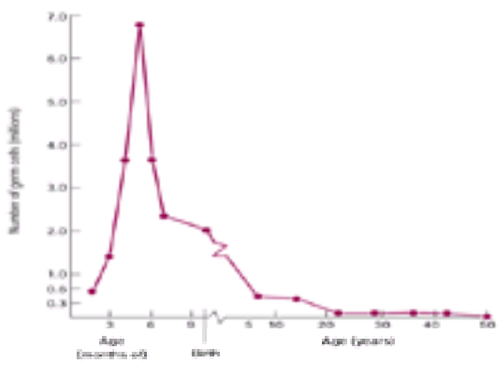

Fig. 2 The number of oocytes in the gonads in relation to age. Note that the maximum rate of atresia occurs before birth. (After Baker TG (1963). *Proceedings of the Royal Society (Biology)* **158**, 417)

Oocytes do not divide once they have been laid down in the ovary. Thus unlike the testis, the ovary has a finite complement of germ cells. Oocytes 'used' in the process of ovulation account for a minute proportion of those that are lost, and it can be seen from Fig. 2 that most atresia occurs during intrauterine life when endocrine influences are least important. The apoptotic process of atresia is controlled genetically. Loss of the second X chromosome, as in 45XO Turner's syndrome, accelerates the rate. This observation, together with the occurrence of familial cases of premature ovarian failure, has prompted the search for genes critical for normal ovarian development. Three loci on the X chromosome appear to be vital: the first is Xp22, which contains a series of genes that escape X inactivation, among which is the *ZFX* gene. Premature ovarian failure is one of the features of the transgenic mouse with this gene 'knocked out'. The *SOX3* gene (mapped to Xq26–27.2), the homologue of the *SRY* gene on the Y chromosome that determines testicular development, also escapes X inactivation and fulfils several of the criteria for a candidate gene. The third region of interest is Xq13–22 because several women with premature ovarian failure and break points in this region have been described. Familial premature ovarian failure also occurs without cytogenetic abnormality and autosomal dominant, autosomal recessive, and X-linked patterns of inheritance have been described. The condition may be present in families with galactosaemia, blepharophimosis, and fragile X syndrome. The presentation is particularly intriguing in the last case because the association is specifically with the premutation, until recently thought not to have a phenotype other than the risk of transmitting fragile X syndrome.

Destruction of genetic material by ionizing irradiation, anticancer chemotherapy, and viral infections, such as mumps oophoritis, can cause premature ovarian failure (Table 3). The association of premature ovarian failure with thyroiditis, adrenalitis, and diabetes mellitus has led to the hypothesis that in many cases the cause is autoimmune. While reliable tests are not widely available, autoantibodies to ovarian cells, oocytes, or gonadotrophin receptors have been reported in up to 80 per cent of cases, a result consistent with the author's finding of thyroid and adrenal autoantibodies in more than half of the patients remaining after those with chromosomal causes had been excluded. Finally, the presence in follicular fluid of toxic pollutants from tobacco smoke, such as cotinine, a congener of nicotine, may account for the earlier menopause that occurs in women who smoke cigarettes.

The symptoms of primary ovarian failure are those of oestrogen deficiency, together with infertility. Investigation reveals subnormal oestradiol and raised serum gonadotrophin concentrations. While the serum luteinizing hormone is often elevated in patients with polycystic ovary syndrome, a rise in the concentration of follicle-stimulating hormone always suggests primary ovarian failure. Autoantibodies should be sought because their presence alerts the clinician to the possible development of pluriglandular endocrine failure. Pelvic ultrasound shows undetectable or small ovaries and a small uterus. Bone densitometry usually reveals significant demineralization.

Patients with premature ovarian failure usually require hormone replacement therapy, the indications, precautions, etc. being the same as those for women in whom the menopause has occurred at a normal age. In some women there may be a spontaneous return of ovulatory menstrual cycles, albeit usually temporary, and therefore pregnancies do sometimes occur. Treatment with glucocorticoids or immunolytic drugs for women with autoimmune ovarian failure has occasionally proven successful but formal controlled trials are lacking and these treatments are not recommended. The ovaries do not respond to further stimulation with gonadotrophins so the only chance of childbearing for most women is through ovum donation. The problems of supplies of donor oocytes are formidable.

Resistant ovary syndrome

This ill-defined syndrome refers to women with amenorrhoea and elevated serum gonadotrophin concentrations in whom, paradoxically, oestrogen levels are well maintained. The persistent secretion of oestradiol suggests persistence of ovarian follicular activity, an implication occasionally confirmed histologically or by ultrasound assessment of the ovaries and by the occasional and unpredictable occurrence of pregnancy. The ovaries do not respond to further stimulation with

exogenous gonadotrophins. While both cause and prognosis remain obscure the condition is most easily understood as a transitional phase on the way to primary ovarian failure.

Polycystic ovary syndrome

A full description of polycystic ovary syndrome is given later. Its clinical expression, typically dating from the time of puberty, involves a menstrual disorder (usually oligomenorrhoea but amenorrhoea in 26 per cent of cases) hyperandrogenization, and weight increase.

The indications for treatment of amenorrhoea in women with polycystic ovary syndrome depend upon the patient's needs. For women who wish to conceive, induction of ovulation is required, combined with attempts to reduce insulin drive to the ovaries by diet, exercise, and insulin-sensitizing drugs such as metformin (see later). For women needing contraception, an oral contraceptive is appropriate, the choice of preparation depending on the degree of associated unwanted hair. An oral contraceptive is also appropriate for women troubled by the lack or unpredictability of menstruation. For the remainder, it is acceptable to remain amenorrhoeic, provided annual ultrasound scans of the endometrium shows that overstimulation has not occurred.

Investigation of amenorrhoea

The patient's stature, nutritional status, the presence of unwanted hair (male pattern or the lanugo of anorexia nervosa) or acanthosis nigricans, and the physical stigmata of oestrogen deficiency are important parts of the physical examination. Pelvic ultrasound reveals the ovarian dimensions and whether they have the characteristic internal echoes of polycystic ovaries and the size of the uterus and degree of endometrial thickening. Hormone assays reveal subnormal serum oestradiol levels associated with elevated (primary ovarian failure) or subnormal (hypogonadotrophic hypogonadism) gonadotrophin concentrations. Serum ferritin measurements are conducted to screen for excessive iron storage in patients with low gonadotrophin concentrations. Assessment of white blood cell karyotype and measurement of autoantibodies is undertaken in patients with primary ovarian failure. Bone densitometry may reveal demineralization of spine and hips. Radiological assessment of the pituitary is undertaken in patients with hyperprolactinaemia.

Treatment of amenorrhoea

In addition to correcting the cause whenever possible, management is directed at minimizing the long-term complications of oestrogen deficiency. When the cause cannot be corrected oestrogen replacement therapy, usually in the form of an oral contraceptive, is appropriate. Outcome should be monitored by ensuring that sufficient oestrogen is administered to cause withdrawal bleeds and that there is an improvement in bone density if skeletal demineralization has been demonstrated.

Hyperandrogenization

Normal hair growth and androgen production in women

At birth the fetus is covered in lanugo hair which rapidly disappears and is not seen again unless anorexia nervosa develops, although very rarely such hair appears as a non-metastatic complication of malignancy. The follicles in the skin of prepubertal children grow soft, short, and fair vellus hair, which, together with scalp, eyebrow, and eyelash hair, is known as non-sexual hair. In response to the secretion of adrenal androgens at puberty, both boys and girls develop terminal hair in the axillae and lower pubic triangle (ambosexual hair). Terminal hair is long, pigmented and coarse. In boys the further increase of (testicular) androgen secretion leads to the development of terminal hair in the upper pubic triangle and on the face, chest, abdomen, and arms and legs—that is, male pattern hair.

Perception of hairiness depends in part on its distribution and in part on its character and pigmentation. While there is racial variation in the density of hair follicles (Native Americans and Orientals having few and women from the Mediterranean littoral having many follicles per unit area of skin) it is the development of male pattern hair in women which constitutes hirsutism.

The sources of androgens in normal women are shown in Fig. 3 from which it can be seen that 50 per cent of directly secreted and 75 per cent of peripherally derived testosterone normally originates from the adrenal cortex. Testosterone circulates specifically bound to sex hormone binding globulin, from which it disassociates and diffuses into target tissues where it is either 5a reduced to a more powerful androgen, dihydrotestosterone, or aromatized to oestradiol. The dihydrotesterone–nuclear protein receptor complex associates with the specific DNA receptor to cause androgen-specific protein synthesis and the expression of androgen action.

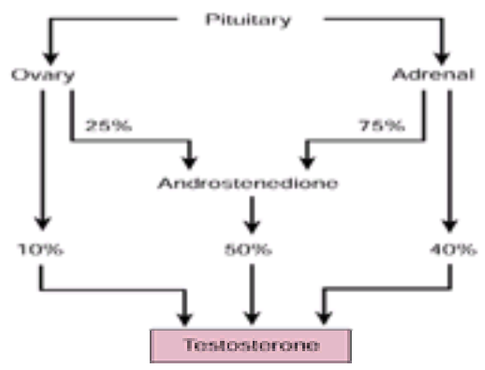

Fig. 3 Sources of androgens in normal women.

Synthesis of sex hormone binding globulin, whose concentration largely determines the total serum testosterone concentration, takes place in the liver, stimulated by thyroxine and inhibited by insulin and to a lesser extent by androgens. The rate of clearance of sex hormone binding globulin is reduced by oestrogen.

Clinical hyperandrogenization

Hyperandrogenization in women is manifest as seborrhoea, persistent acne, and the development of a male pattern of distribution and quality of hair. Male pattern hair loss may also occur, particularly in women with male family members who are bald. Clitoromegaly and increased muscle bulk are signs of severe and usually long-standing overexposure to androgens.

Although defects in adrenal steroid biosynthesis (congenital and late onset adrenal hyperplasia), Cushing's syndrome, and adrenal androgen secreting tumours may all cause oversecretion of androgens and therefore present with hirsutism, the commonest cause by far is polycystic ovary syndrome, in which condition there is an increase in the direct ovarian secretion of androgens.

Polycystic ovary syndrome

Polycystic ovaries are readily identified by pelvic ultrasound, because they are larger than normal (average volume three times that in normal women) and have a highly echodense central stroma in which cysts of 2 to 6 mm diameter are arranged around the circumference (Fig. 4). When ovaries with this appearance are detected in women complaining of specific symptoms, the term polycystic ovary syndrome is used (Table 4). Defined in this way, the polycystic ovary syndrome corresponds to the condition described over 50 years ago by Stein and Leventhal.

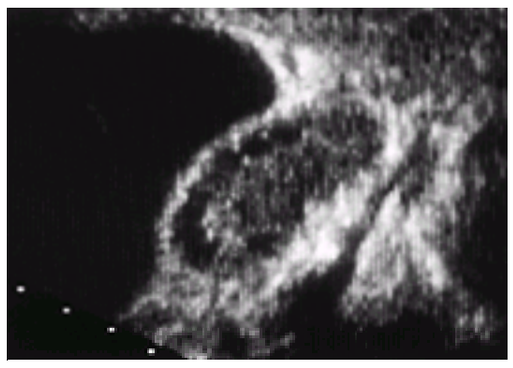

Fig. 4 Transabdominal ultrasound image of a polycystic ovary. Note the enlarged ovary with the echodense central stroma and the necklace of cysts around the circumference.

Patients with this condition commonly present in their late teens or early twenties, complaining of the consequences of hyperandrogenization or of a menstrual disturbance (Table 4). There is often a family history of similar complaints but even when that is absent an ultrasound scan usually reveals polycystic ovaries in first-degree relatives. Infertility is caused by failure of ovulation, although hypersecretion of luteinizing hormone is also important in this regard. Obesity, often associated with an increase in the ratio of waist to hip circumference, is the third classical feature.

Endocrine features

The classical profile is of hypersecretion of luteinizing hormone and androgens with normal circulating follicle-stimulating hormone, prolactin, and thyroxine concentrations. In fact a spectrum of endocrine findings occurs, reflecting the phenotypic heterogeneity of the polycystic ovary syndrome. In a study of more than 1500 cases, 44 per cent had an elevated serum luteinizing hormone and 22 per cent an elevated serum total testosterone concentration. Levels of luteinizing hormone were raised most commonly in the women complaining of infertility and of testosterone in those complaining of hirsutism.

The nature of the primary disturbance underlying these findings is uncertain. A central problem is failure of the polycystic ovary to convert androgens, made in excessive amounts by the abundant theca and interstitial cells of the hyperplastic ovarian stroma, into oestrogens. The androgens (predominantly androstenedione and testosterone) are released into the circulation and converted in the skin to dihydrotestosterone. In liver and fat tissue, they are converted into oestrogens at a rate which increases with the degree of obesity. The high levels of oestrogen (predominantly oestrone) inhibit secretion of follicle-stimulating hormone and may stimulate secretion of luteinizing hormone. The former effect contributes to persistent anovulation and the consequent lack of progesterone (which in the normal luteal phase limits the proliferative action of oestradiol) means that the action on the uterus of the normally weak oestrogen oestrone is unopposed. These patients are consequently at risk from endometrial hyperplasia and neoplasia. The raised levels of luteinizing hormone stimulate the excessive numbers of theca and interstitial cells to oversecrete androgens. Exposure of the ovaries to high levels of luteinizing hormone at inappropriate times of the cycle may also impair fertility through an action on the developing oocyte.

The above model does not explain the variable clinical presentation of polycystic ovary syndrome. It is likely that environmental factors lead to expression of the underlying, probably inherited, condition.

Many patients with polycystic ovary syndrome, particularly those who are anovulatory, are resistant to the action of insulin. While several types of insulin resistance are currently recognized (Chapter 12.11), in women with polycystic ovary syndrome the resistance is specifically to insulin-mediated extrasplanchnic disposal of glucose. As a consequence, euglycaemia can only be maintained through compensatory hypersecretion of insulin, the clinical clue to which is the development of acanthosis nigricans (see Plate 1). The insulin resistance spares the liver (the fasting glucose concentration is normal, serum sex hormone binding globulin and high-density lipoprotein concentrations are suppressed), the skin, and the ovary. Ovarian dysfunction results, in direct proportion to the intensity of compensatory hyperinsulinism.

A specific defect in transduction of the insulin signal has been described in women with polycystic ovary syndrome. In addition, as children enter puberty, insulin resistance develops in response to the increase of growth hormone secretion that underlies the acceleration in growth at this age. Obesity itself, present in some 40 per cent of women with polycystic ovary syndrome, worsens insulin resistance and so causes further deterioration of ovarian function. Should the patient come from a family with diabetes mellitus, there is the added risk of developing the insulin resistance of non-insulin-dependent diabetes mellitus.

Hypersecretion of insulin inhibits hepatic synthesis of sex hormone binding globulin which, particularly in obese patients, results in an apparent disparity between circulating testosterone concentrations and the degree of hirsutism. In these women the concentration of unbound testosterone, and by implication the rate of production of testosterone, is very high despite serum total testosterone concentrations which may be within the normal range. Hypersecretion of insulin also has non-reproductive adverse effects in patients with polycystic ovary syndrome. Thus an inverse relation of the cardioprotective high-density lipoprotein to cholesterol concentration and the fasting serum insulin concentration has been demonstrated, together with subnormal total high-density lipoprotein concentrations (Fig. 5). These data, together with reports of an increased incidence of coronary heart disease, hypertension, and diabetes in follow-up studies of patients with histologically verified polycystic ovaries, indicate the clinical importance of hypersecretion of insulin and its control in patients with polycystic ovary syndrome.

Fig. 5 Serum high-density lipoprotein concentrations in lean and obese women with polycystic ovary syndrome. Note that despite maintaining a normal body mass index, slim women with polycystic ovary syndrome have a statistically significant depression of their fasting high-density lipoprotein cholesterol concentration. The high-density lipoprotein falls even further in women with polycystic ovary syndrome who are obese.

The familial nature of polycystic ovary syndrome has led to extensive investigation of its genetics. Thus far linkage with genes involved in the steroid biosynthetic pathway (*CYP11A*) and the control of insulin secretion have been identified and confirmed. *CYP11A* encodes the side chain cleavage enzyme which converts cholesterol to pregnenolone, a rate limiting step in steroid biosynthesis. Presumably a mutation causing upregulation of this enzyme could result in an increase in androgen secretion, a cardinal feature of the syndrome. The linkage described with the large class III alleles of the insulin gene variable number tandem repeats (*INS VNTR*) also occurs with type 2 diabetes mellitus. Class III *INS VNTR* promotes insulin secretion, perhaps resulting in insulin resistance as a secondary event. The effects of these and other candidate genes may be expressed in fetal development or through obesity in later life.

Other ovarian causes of hirsutism

Hyperthecosis

Characterized pathologically by the presence of islands of luteinized theca cells within the ovarian stroma at a distance from follicles, the clinical features include very marked hypersecretion of androgens and of insulin. The condition is probably most easily regarded as a severe form of the polycystic ovary syndrome.

Ovarian tumours

Androgen secreting tumours of the ovary are derived from sex cord or stromal cells and include Sertoli–Leydig cell tumours (arrhenoblastomas), hilar cell tumours, lipoid cell tumours, and adrenal rest tumours. Other non-hormone secreting tumours (Brenner, cystadenoma, and cystadenocarcinoma) have been reported to stimulate androgen secretion by the surrounding ovarian stroma. These conditions are all very rare causes of hirsutism.

Diagnosis of hyperandrogenism

Adrenal causes of hyperandrogenism are discussed in Section 12.7 but hypersecretion of androgens by polycystic ovaries is much more common. While an ovarian tumour is suggested by a short history of rapidly advancing hirsutism, amenorrhoea, and a serum testosterone concentration in the male range, such lesions are in fact rare. A serum testosterone exceeding 10 nmol/litre is more commonly associated with polycystic ovary syndrome and severe insulin resistance, as suggested clinically by the presence of acanthosis nigricans, than with the development of an ovarian tumour.

Pelvic ultrasound will detect polycystic ovaries or an ovarian tumour. The serum total testosterone concentration reflects in part the rate of production of testosterone and in part the serum concentration of sex hormone binding globulin; it should be interpreted in the light of the patient's body weight and a concentration within the normal range should not therefore be dismissed in women who are overweight. Serum luteinizing hormone concentrations are often raised but with a normal level of follicle-stimulating hormone. Serum prolactin concentrations are modestly elevated (up to 2500 mu/litre) in 15 per cent of patients with polycystic ovary syndrome. About half of these cases have a microadenoma detected by MRI scan.

A small number of patients with hirsutism have no diagnosable cause for their cutaneous virilism. Labelled 'idiopathic hirsutism', these patients may have enhanced sensitivity of androgen-dependent tissues, perhaps caused by increased dermal activity of the 5a reductase enzyme.

Management of hirsutism

Medication reduces the rate of hair growth but cosmetic treatment is required to remove existing unwanted hair. Hair can be camouflaged by bleaching and removed by plucking, waxing, electrolysis, or laser. The latter two methods offer the possibility of long-term hair removal. Laser treatment works by selective thermolysis and, with present equipment, is only suitable for removal of dark hair from a fair skin.

In the United Kingdom the preferred drug for treatment of hyperandrogenization is cyproterone acetate. This steroid is a peripheral antiandrogen, a progestogen, and a mild glucocorticoid. In combination with oestrogen, it suppresses secretion of gonadotrophin and so reduces the secretion of ovarian androgens. It is also contraceptive. Its glucocorticoid activity may reduce secretion of adrenal androgens. Finally, it blocks uptake of the dihydrotestosterone–protein complex by the DNA acceptor protein in the nucleus of androgen sensitive cells, so acting as a peripheral antiandrogen.

Treatment is administered cyclically, together with oestrogen, given most conveniently in the form of Dianette (Fig. 6). Seborrhoea and acne usually clear up in about 6 weeks but it takes 12 to 18 months to realize the maximum improvement of unwanted hair. Cosmetic treatment is continued while on the medication, the impact of therapy being indexed by a reduction in the number of treatments required.

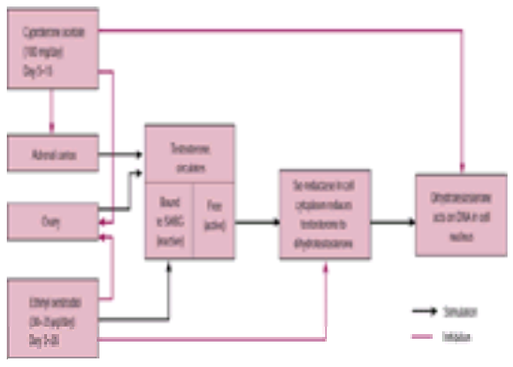

Fig. 6 The use of cyproterone acetate and ethinyloestradiol in the treatment of hirsutism. In the author's practice, ethinyloestradiol is usually replaced with Dianette®, a conveniently packaged formulation of ethinyloestradiol (35 µg) and cyproterone acetate (2 mg). (After *Medicine International* (1989))

Adverse reactions, contraindications, and surveillance are essentially those advised for treatment with the oral contraceptive pill. As symptoms remit, the dose of cyproterone acetate is reduced until the patient is taking the lowest dose compatible with symptomatic relief. Eventually treatment with Dianette alone is usually sufficient for maintenance. In patients not responding to this regimen, other antiandrogens, such as aldactone and flutamide, may be tried. Inhibition of the type II 5a reductase enzyme with finasteride has similar efficacy. Indeed the major difference between the available medications is in the pattern of adverse effects rather than in efficacy. Since these drugs are not contraceptive the patient must be warned of possible feminizing effects on a male fetus if they are taken inadvertently during pregnancy. They are therefore optimally prescribed with an oral contraceptive.

Infertility

Infertility can be defined as absence of conception after a year of unprotected intercourse, but it is most logically evaluated in relation to normal fertility. The maximum conception rate per ovulation is 25 to 30 per cent, so that the best cumulative conception rates are about 60 per cent after 6 months and 85 per cent after a year. Other than mechanical bars to conception (for example gynaecological problems such as occluded fallopian tubes) the important factors that reduce a woman's fertility are her age and any process that reduces the number of ovulations per unit time. The central strategy of medical management of female infertility is therefore the diagnosis and treatment of the causes of anovulation. An additional strategy is to ensure that ovulation, and thus conception, occurs in as favourable an environment as possible.

Ovulation is only proven by the occurrence of pregnancy, so indirect methods of detection are required. In practice, ovulation is usually inferred retrospectively by detection of a corpus luteum, indexed endocrinologically either by measurement of serum progesterone concentrations or indirectly by the effects of progesterone on basal body temperature or endometrial histology. Ovulation can be predicted ultrasonically by detecting the development of a preovulatory follicle of average diameter 20 to 22 mm, followed by its collapse and replacement by a solid structure, i.e. visualization of a corpus luteum. The preovulatory surge of luteinizing hormone can be detected by the patient herself using one of a number of commercially available immunological urine tests. These methods are used to determine whether anovulation can account for a couple's infertility and whether treatment has actually resulted in ovulation; prediction of ovulation is helpful for timing investigations and maximizing the chance of conception by ensuring intercourse around the time of ovulation.

Conception rates after the age of 35 are only half of those before the age of 25. Demographic changes in northern Europe (median maternal age at first birth in the United Kingdom is now 27 years) have resulted in a steady increase in the number of couples requesting consultations for infertility.

The causes of amenorrhoea are discussed above. All except primary ovarian failure are correctable (and that condition is treatable by oocyte donation) so the fertility prognosis for this group of patients is excellent. The commonest cause of oligomenorrhoea is polycystic ovary syndrome and while patients with this condition usually

ovulate readily in response to treatment, in about 40 per cent hypersecretion of luteinizing hormone impairs fertility, despite the occurrence (spontaneously or as a result of treatment) of otherwise normal ovulation. The mechanism is uncertain but may involve an adverse effect of the high levels of luteinizing hormone on completion of the final stages of oocyte maturation.

Failure to ovulate despite (more or less) regular menstrual cycles is an unusual but recognized cause of infertility.

Infertility not explained by ovulatory failure is usually treated by *in vitro* fertilization and embryo transfer.

Induction of ovulation

Table 5 shows the agents commonly used and the endocrine level at which they exert their actions. Anti-oestrogens enhance hypothalamic secretion of gonadotrophin releasing hormone by competing with oestrogen receptors, thus simulating oestrogen deficiency. The drug is taken for 5 days and provokes gonadotrophin secretion and thence follicular development. The most commonly used preparation is clomiphene, which is a racemic mixture, one isomer having oestrogenic and the other anti-oestrogenic activity. Most slim patients with polycystic ovary syndrome ovulate in response to clomiphene. Obese patients, who are usually hyperinsulinaemic, can be treated with metformin, which in a dose of 500 mg three times per day will often result in ovulation or enhance the response to treatment with clomiphene.

In patients not responding to the above treatments, pituitary secretion of gonadotrophins can be enhanced by injection of gonadotrophin releasing hormone. It is administered in a pulsatile fashion, usually by the subcutaneous route, the injections being given at 90 min intervals by a portable miniaturized pump.

For patients with structural lesions of the pituitary, gonadotrophin secretion can be replaced by injections of gonadotrophins. The preferred preparations are those synthesized by recombinant technology. Follicular development is induced by the injection of follicle-stimulating hormone (together with luteinizing hormone in those patients with hypogonadotrophic hypogonadism); ovulation is triggered and the corpus luteum maintained by a single injection of human chorionic gonadotrophin which has luteinizing hormone-like bioactivity and a very long half-life.

The objective of treatment is unifollicular ovulation with full-term delivery of a single infant. The response is monitored by ultrasound assessment of the ovaries and uterus and by measurement of plasma oestradiol concentrations. Human chorionic gonadotrophin is administered according to strict criteria (not more than three follicles of diameter equal to or greater than 16 mm, or six follicles equal to or greater than 14 mm diameter).

Complications of ovulation induction include multiple births (the perinatal mortality of twins is three times that of singletons) and the ovarian hyperstimulation syndrome. The latter condition occurs almost entirely in women with polycystic ovary syndrome who have received high-dose and inadequately monitored gonadotrophin therapy. It results from massive follicular luteinization and so only occurs after ovulation has been triggered by human chorionic gonadotrophin or, very rarely, by a spontaneous surge of luteinizing hormone. Symptoms usually appear 5 to 10 days after administration of human chorionic gonadotrophin. In its mildest form it consists of ovarian enlargement and discomfort but in the more severe forms abdominal distension, nausea, vomiting, and diarrhoea develop. As a result of increased vascular permeability, protein-rich fluid accumulates in the peritoneal and sometimes the thoracic cavity; hypovolaemia develops, associated with haemoconcentration, decreased central venous pressure, low blood pressure, and tachycardia. The patient develops a tense ascites, respiration is embarrassed, and urine formation is suppressed. A hypercoagulable state may develop with the risk of cerebral and peripheral venous thrombosis and embolism.

In managing this syndrome, its self-limiting nature should be borne in mind. Treatment is designed first to maintain blood volume while correcting fluid and electrolyte balance, second to avoid thromboembolic phenomena (by full heparinization if severe hypercoagulability is detected), and third to relieve abdominal and pulmonary symptoms (by paracentesis under ultrasound control).

Further reading

Berchuk A, *et al*. (1996). Role of BRCA1 mutation screening in the management of familial ovarian cancer. *American Journal of Obstetrics and Gynecology* **175**, 738–46.

Dunaif A (1997). Insulin resistance and the polycystic ovary syndrome: mechanism and implications for pathogenesis. *Endocrine Reviews* **18**, 774–800.

Kalantaridou SN, Davis SR, Nelson LM (1998). Premature ovarian failure. *Endocrinology and Metabolism Clinics of North America* **27**, 989–1006.

Sourander L, *et al*. (1998). Cardiovascular and cancer morbidity and mortality and sudden cardiac death in postmenopausal women on oestrogen replacement therapy (ERT). *Lancet* **352**, 1965–9.

12.8.2Disorders of male reproduction

F. C. W. Wu

Physiology of the hypothalamic–pituitary–testicular axis

The adult testis performs two functions—the production of androgens and spermatozoa (Fig. 1). These functions are dependent on trophic hormones from the hypothalamus and anterior pituitary which are responsive to the negative feedback action of testicular hormones, thus forming a closed-loop functional axis (Fig. 2).

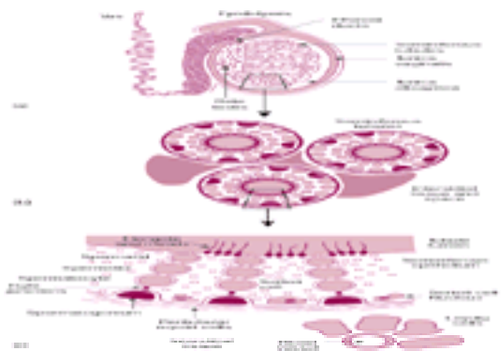

Fig. 1 (a) Human testis, epididymis, and vas deferens showing efferent ducts leading from the rete testis to the caput epididymis and the cauda epididymis continuing to become the vas deferens. (b) Cross-section through a seminiferous tubule showing central lumen, seminiferous epithelium, and interstitial space containing Leydig cells. (c) Anatomical relationships in the seminiferous epithelium between germ cells (spermatogonia, spermatocytes, and spermatids), Sertoli cells, peritubular myoid cells, and Leydig cells.

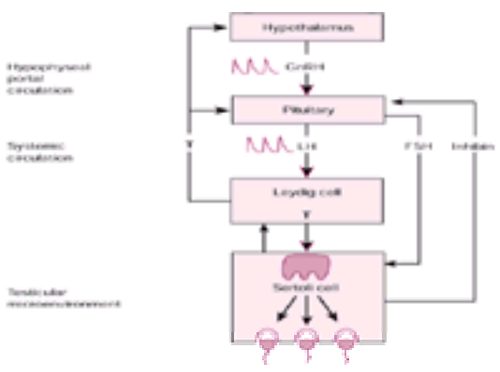

Fig. 2 Functional relationships in the hypothalamic–pituitary–testicular axis and testicular microenvironment. Gonadotrophin releasing hormone (GnRH) is secreted into the hypophysial circulation in an episodic manner which is reflected by an luteinizing hormone (LH) pulse in the systemic circulation. Open arrows represent positive stimulation and closed arrows negative feedback.

Gonadotrophin releasing hormone (GnRH) is synthesized in neurosecretory neurones in the hypothalamus and then released episodically into the pituitary portal circulation at a frequency of 1 to 2 hourly. GnRH stimulates synthesis and secretion of both luteinizing hormone (LH) and follicle stimulating hormone (FSH) in the gonadotrophs of the anterior pituitary gland. Each episode of GnRH secretion elicits an immediate release of gonadotrophins into the systemic circulation. The pulsatile pattern of LH secretion is more clearly defined, with its rapid release and shorter circulating half-life than FSH (Fig. 2). This intermittent mode of GnRH stimulation avoids desensitization of the pituitary gonadotrophs by continuous GnRH exposure, and is therefore obligatory for maintaining normal gonadotrophin secretion. Pituitary FSH secretion can also be modulated by locally produced activin, a peptide hormone of the inhibin family which stimulates FSH.

LH stimulates biosynthesis of androgenic steroids by binding to specific surface membrane receptors on the Leydig cells. This activates the cyclic AMP/ protein kinase and the steroidogenic acute regulatory protein which mobilizes cholesterol substrate, transfers cholesterol from the outer to inner mitochondrial membrane where it is converted to pregnenolone by splitting the side-chain at position C21. Figure 3 shows the principal steps in the steroidogenic pathway in which the carbon skeleton of the parent compound, cholesterol, is progressively hydrolysed to form various androgenic steroids. Testosterone is the main end product of the biosynthetic pathway in adult Leydig cells. The daily testicular production rate of testosterone is between 3 and 10 mg. As the principal circulating androgen secreted by the adult testes, testosterone exerts the major negative feedback action on gonadotrophin secretion by restricting the frequency of GnRH release from the hypothalamus and by reducing the amplitude of LH response to GnRH. Androgens are essential for the differentiation, growth, and function of the male genital ducts (epididymis) and accessory glands (seminal vesicles and prostate), male secondary sexual characteristics, and sexual potency (Table 1). Testosterone circulates in plasma bound to sex hormone binding globulin (SHBG) and albumin. In man, 60 per cent of circulating testosterone is bound to SHBG, 38 per cent to albumin, and 2 per cent is free. Free and albumin-bound testosterone constitute the bioavailable fractions of circulating testosterone. Androgen action is mediated through specific binding to intranuclear androgen receptors which increase transcription of specific androgen-responsive genes in target cells. In target organs, such as the fetal external genitalia, prostate, and facial hair follicles, full activation requires the local metabolism of testosterone by the enzyme 5a-reductase to 5a-dihydrotestosterone, an androgen which is several-fold more potent than testosterone (Fig. 3).

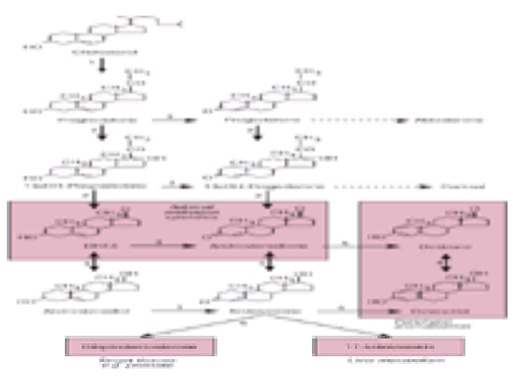

Fig. 3 Steroidogenic pathway from cholesterol to testosterone and further conversion of testosterone: (1) cholesterol side-chain cleavage; (2)

17a-hydroxylase/17,20-lyase; (3) 3b-hydroxysteroid dehydrogenase; (4) 17b-hydroxysteroid dehydrogenase; (5) aromatase; (6) 5a-reductase.

Recently described males with mutant oestrogen receptor (oestrogen resistance) and *CYP19* gene encoding aromatase (oestrogen deficiency), and the corresponding gene-knockout mouse models, have revolutionized our understanding of the role of oestrogens in men. An increasing variety of androgen-dependent functions in males are now known to be mediated by the oestrogen receptors a and b via conversion of testosterone to oestradiol by the widely distributed P450 aromatase in target tissues. These include pubertal growth spurt, skeletal maturation, fusion of epiphyses at the end of puberty, bone mass accrual and maintenance, some aspects of male-specific behaviour (in mouse models), fluid resorption from the testicular efferent ducts, and FSH feedback regulation. Much of the action of circulating testosterone is therefore regulated or refined locally at many different target tissues by 5a-reductase and aromatase expression. The relative contribution of circulating (endocrine action) compared with locally-produced (paracrine or intracrine action) hormones and the balance between androgen and oestrogen receptor activation are crucial to the physiological effects of androgens in man.

The endocrine (androgen synthesis) and gametogenic (spermatogenesis) functions of the testis are closely interlinked. Although testosterone is important as the principal circulating androgen, its local (paracrine) action within the testis is crucial, together with FSH, for the initiation and maintenance of normal spermatogenesis and hence fertility (Fig. 1 and Fig. 2). Since germ cells do not possess receptors for FSH or testosterone, these hormone signals are transduced through the Sertoli and peritubular cells. Sertoli cells create an insular microenvironment in the seminiferous tubules by providing the physical framework and elaborating an ever-changing chemical myriad of growth factors and cytokines for the developing germ cells enmeshed in their cytoplasm (Fig. 1). Sertoli cells also secrete inhibin B, a glycoprotein hormone, which inhibits FSH secretion by the pituitary (Fig. 2).

Spermatogenesis is a complex, repetitive series of cytodifferentiation processes in the seminiferous epithelium whereby cohorts of undifferentiated diploid germ cells (spermatogonia) proliferate and transform into greatly expanded populations of haploid spermatozoa (Fig. 1). The human testes produce around 200 million spermatozoa per day. Mitotic divisions of spermatogonial stem cells form subpopulations of spermatogonia which, at regular intervals of 16 days, differentiate into primary preleptotene spermatocytes to initiate meiosis. Meiotic reduction divisions of spermatocytes generate round spermatids which then transform (spermiogenesis) into compact, virtually cytoplasm-free, elongated spermatids. Condensed nuclear DNA forms the sperm head with an overlying Golgi-derived acrosome cap and a tail (containing nine pairs of microtubules arranged around a central pair) capable of propelling, flagellar movements. Mature spermatozoa are released from Sertoli cell cytoplasm into the tubular lumen some 74 days after their initial development from spermatogonia. The control systems regulating germ cell divisions and development remain poorly understood.

Male reproductive disorders

Male hypogonadism is a descriptive term for the clinical complex associated with androgen deficiency due to the failure of Leydig cell function. Concomitant impairment of spermatogenesis is likely since the seminiferous tubules will also be androgen deficient or directly involved by the same pathological process. However, infertility is usually an isolated abnormality of spermatogenesis where patients seldom show any clinical evidence of androgen deficiency. In the last 10 years, an increasing number of specific genetic defects have been identified, by genomic DNA mapping, to be associated with abnormal gonadal function and development. This has greatly improved our understanding of the pathogenesis of these conditions.

Male hypogonadism

Aetiology

A large number of pathological conditions can lead to destruction or malfunction of the hypothalamic–pituitary–testicular axis (Table 2). It is important to identify the underlying cause of hypogonadism and distinguish between pituitary-hypothalamic (secondary or hypogonadotrophic hypogonadism) and testicular (primary or hypergonadotrophic hypogonadism) disorders. The causal lesion may require specific treatment e.g. pituitary tumour, haemochromatosis. Hypogonadotrophic conditions are amenable to treatment aimed at inducing or restoring spermatogenesis while primary testicular failure, which is usually irreversible, is not.

Diagnosis

General clinical features of hypogonadism

The age of onset of androgen deficiency critically influences the manifestation of hypogonadism (Table 1). Prepubertal onset of testosterone deficiency gives rise to sexual infantilism and patients will present with delayed puberty. Eunuchoidal body proportions (arm span greater than height and heel–pubis exceeding crown–pubis lengths by at least 5 cm; Fig. 2) develop due to the continued growth of long bones (growth hormone-mediated) allowed by the delayed closure of their epiphyses and lack of the testosterone/oestradiol-induced spinal growth in late puberty.

Postpubertal onset of testosterone deficiency leads to regression of spermatogenesis, diminished sex drive and erection, loss of ejaculation, muscle atrophy, poor stamina, and decreased secondary sexual hair and shaving frequency. However, no change is observed in body and penile proportions or voice (Table 1). Symptoms and signs of hypogonadism usually develop and progress insidiously. It is therefore common for patients to present only after many years following the onset of hypogonadism. Furthermore, young patients who has never been adequately androgenized may not be aware, or even deny, that secondary sexual function is subnormal. In contrast, after surgical or traumatic/inflammatory castration, adults may experience hot flushes from acute withdrawal of androgens. Fetal onset of defective androgen action due to androgen receptor abnormalities or steroidogenic enzyme deficiency cause failure of masculinization of the genitalia resulting in intersexual states (Table 2).

Clinical findings associated with hypogonadism

Hypothalamic–pituitary tumours are suggested by headache, impairment of visual acuity or visual field loss, polyuria and polydipsia, or evidence of pituitary hormone excess such as Cushing's disease, acromegaly, and hyperprolactinaemia. Hyperprolactinaemia causes loss of sex drive even in the presence of normal testosterone. Primary testicular failure is suggested by a history of orchitis, testicular trauma, surgery, torsion, irradiation, or chemotherapy. An increasing number of chronic systemic diseases (Table 2) are associated compromised hypothalamic–pituitary–testicular function. With improved survival resulting from specific treatment, the role of gonadal dysfunction in the quality of life of these patients is increasingly important.

The use of recreational drugs and medications which interfere with pituitary–testicular function or androgen action should be sought (Table 2). Evidence of alcohol abuse should be noted. Ethanol causes a lowering of plasma testosterone through a direct toxic effect on Leydig cell steroidogenesis. Testicular atrophy and gynaecomastia, found in 50 per cent of men with hepatic cirrhosis, are due to altered androgen steroid metabolism, increased sex-hormone-binding globulin, and increased oestrogen production. These changes are usually irreversible.

Neurological diseases can be associated with hypogonadism. Postpubertal atrophy of the seminiferous tubules occurs in 80 per cent of patients with dystrophia myotonica, an autosomal dominant disorder characterized by myotonia, distal muscle atrophy, lens opacities, and premature frontal balding. Variable degrees of androgen deficiency also exist. Hypogonadotrophic hypogonadism is associated with familial cerebellar ataxia, Laurence–Moon, Bardet–Biedl, and Prader–Willi syndromes. Defective spermatogenesis is common in paraplegia or quadraplegia following spinal injury because of the inability to maintain a low scrotal temperature.

Specific condition

Klinefelter's syndrome is the commonest cause of male hypogonadism with an incidence of 2 per 1000 live births. It is a developmental disorder of the testis resulting from the presence of an extra X chromosome derived from the non-disjunction of parental (maternal origin in two-thirds of cases) germ cells during meiosis. The most common karyotype is 47 XXY (80–90 per cent) but rarer variants include 46 XY/47 XXY mosaic, multiple X + Y, and the so-called XX male syndrome. Accelerated atrophy of germ cells before puberty and hyalinization of the seminiferous tubules gives rise to sterility and small, firm testes. Leydig cells appear relatively hyperplastic but cell mass is in fact normal. The degree of Leydig cell steroidogenic defect (mechanism of which remains uncertain) is very variable ranging from the virilized, adult male presenting with infertility (see below) to the eunuchoidal youth who fails to complete sexual maturation. In midadulthood, 80 per cent of patients have reduced testosterone with elevated LH, FSH, and oestradiol. Other features include gynaecomastia, reduced body hair, long legs, tall stature, learning (verbal

and cognitive) difficulties, poor school performance, behavioural disturbances, and autoimmune endocrinopathies including diabetes mellitus. There is also an increased incidence of osteopenia, breast tumour, testicular and extratesticular (especially mediastinal and retroperitoneal) germ cell tumours, varicose veins, and leg ulcers. Mental retardation is associated with higher order X chromosome polysomy.

Kallmann's syndrome, with an incidence of 1 in 7500 males, is a sporadic or familial (X-linked or autosomal) form of congenital hypogonadotrophic hypogonadism associated with a number of somatic congenital abnormalities including anosmia or hyposmia (defective smell sense), red–green colour blindness, synkinesis, nerve deafness, cleft-lip or palate, and renal malformations. The X-linked variety is caused by deletion or mutation in the *KAL-1* gene in Xp22.3 encoding an cell adhesion protein. Faulty embryonic migration of GnRH-secreting neurones from their site of origin in the nose to the hypothalamus prevents normal axonal secretion into the pituitary portal circulation in the median eminence. GnRH is thus unable to target the gonadotrophs in the anterior pituitary. The same migratory defect affects the olfactory neurones in the nose, resulting in aplasia of the olfactory bulb and anosmia.

Total and free testosterone declines gradually but variably with age in men from the age of 40 years onwards. This is amplified by the age-related increase in SHBG and exacerbated by concomitant non-gonadal diseases and medications. In some elderly men, testosterone may fall below the young adult physiological range. Differentiation of non-specific symptoms of ageing, such as frailty, decreased muscle strength, lack of stamina, and decline in libido, from those of mild hypogonadism is difficult. Whether these functional changes, normally accepted as part of healthy ageing, are causally related to alterations in circulating testosterone is unclear. The existence and prevalence of a male climacteric remains controversial.

Investigation

Confirmation of hypogonadism

The clinical suspicion or diagnosis of hypogonadism must be confirmed by demonstration of low circulating testosterone before replacement therapy is commenced. Samples obtained between 8 and 9 a.m. avoid the physiological, diurnal trough levels of testosterone later in the day. In the presence of background changes in SHBG, such as in ageing, obesity, anticonvulsant medications, diabetes, thyrotoxicosis, and liver disease, free testosterone can be calculated from the total testosterone and SHBG concentrations. Free testosterone assays are technically demanding and result obtained by commercial kits can be misleading.

Assessment of the hypothalamic–pituitary–testicular axis and target tissue resistance

Measurement of LH, FSH, and prolactin is required to differentiate between primary and secondary hypogonadism. The physiological basis for differentiating between hypogonadotrophic and hypergonadotrophic hypogonadism is illustrated in Fig. 1. Pathologies in the hypothalamus and pituitary will give rise to low or low–normal gonadotrophins and low testosterone, that is a state of hypogonadotrophic hypogonadism or secondary testicular failure where the potential for stimulating testicular function by exogenous gonadotrophin or GnRH replacement is maintained. Conditions affecting the testes will interrupt normal testicular negative feedback. This results in elevated gonadotrophin levels with low testosterone, characteristic of hypergonadotrophic hypogonadism or primary testicular failure. Failure of spermatogenesis with reduced testicular size is commonly associated with a rise in FSH alone. The value of circulating inhibin B and mullerian inhibiting hormone for diagnostic purposes is currently being assessed. Patients with androgen insensitivity syndromes have elevated testosterone with high LH but normal to low FSH. Increased LH or FSH is associated with the very rare LH and FSH resistance syndromes.

Human chorionic gonadotrophin (hCG) stimulates Leydig cell steroidogenesis and plasma testosterone increases over 4 to 7 days. It is useful for detecting the presence of functional testicular tissue in patients with impalpable testes, assessing functional reserve of the testes prior to treatment with exogenous gonadotrophin or GnRH, and in differentiating hypergonadotrophic hypogonadism from rare cases who produce immunologically detectable, but biologically inactive, LH in excess.

Stimulation tests of gonadotrophin secretory reserve using clomiphene and GnRH seldom give additional information and has become largely obsolete, especially with the improved sensitivity and range of modern gonadotrophin immunoassays.

Assessment of the pituitary

Patients with hypogonadotrophic hypogonadism without the stigmata of Kallmann's syndrome should undergo full pituitary functional and anatomical assessment to exclude an underlying pituitary tumour. They require pharmacological tests of growth hormone and ACTH reserve, thyroid function tests, visual field charting, and MR or CT scanning of the pituitary and hypothalamus.

Other investigations

Suspected Klinefelter's syndrome should be confirmed by chromosome karyotyping on peripheral blood lymphocytes. Ultrasound and MR scan are useful in locating ectopic or intra-abdominal testes. DNA analysis can help confirm the diagnosis of androgen resistance syndromes and an increasing number of rare causes of hypogonadism (Table 2) such as haemochromatosis.

Treatment objectives

Treatment objectives are to:

1. relieve the symptoms of androgen deficiency;
2. prevent the long-term consequences of androgen deficiency such as osteopenia;
3. reproduce physiological, circulating, and tissue levels of plasma testosterone, dihydrotestoerone, and oestradiol;
4. induce fertility, if required, in hypogonadotrophic patients;
5. treat any specific underlying diseases.

The mainstay of treatment of the hypogonadal male is androgen replacement. Although hypogonadotrophic patients have the potential for fertility, gonadotrophin and pulsatile GnRH therapy should only be employed when there is a requirement for fertility because of the expense and complexity of these regimens. Previous testosterone treatment does not jeopardize subsequent response to gonadotrophin so that younger hypogonadotrophic subjects should be treated by testosterone in the same manner as hypergonadotrophic patients to initiate and maintain virilization and sexual function.

Androgen replacement

The circulating half-life of free testosterone is short (10 min) due to rapid degradation by the liver. To achieve sustained physiological circulating concentrations, testosterone has to be administered in a modified form or by a parenteral route so that its rate of metabolism or absorption is retarded.

Injectable testosterone esters are the commonest first-line androgen preparations. A mixture of four different testosterone esters (propionate, phenylpropionate, isocaproate, and decanoate) (Sustanon, Organon, Oss, the Netherlands), 250 mg 2 to 3-weekly, and testosterone enanthate (Primoteston depot, Schering, Berlin, Germany), 200 mg 2-weekly, are the most popular. Whilst undoubtedly effective, these preparations inevitably give rise to high supraphysiological peak testosterone levels within the first week which then fall sharply to lower limits of normal before the next dose. Some patients are disturbed by fluctuations in libido, mood, and stamina associated with the repeated rise and fall of testosterone levels as well as the painful, deep, intramuscular injections.

Crystalline testosterone compressed into cylindrical pellets, surgically implanted subcutaneously under local anaesthesia, provide a depot source of testosterone for several months. Peak testosterone levels are achieved after 2 to 4 weeks, followed by a gradual decline over subsequent months. A total dose of 800 mg (4 × 200 mg implants) can maintain physiological concentrations of testosterone over 6 months, which some patients find more convenient than more frequent injections. The implantation procedure can be complicated, though rarely, by haemorrhage and infection. Even in experienced hands, 10 per cent of implanted pellets are extruded. Implants should only be used as maintenance therapy for patients who have already shown satisfactory tolerance to androgen effects of shorter-acting preparations.

Testosterone undecanoate is administered orally and absorbed from the gut through intestinal lymphatics. Low bioavailability (<0.5 per cent), variable absorption, multiple daily dosing, and higher costs has restricted the use of testosterone undecanoate despite the obvious appeal of oral administration. To maintain testosterone

consistently within the physiological range, two to three times daily administration of 80 mg (2 × 40 mg capsules) of testosterone undecanoate is required. Intestinal 5a-reductase action gives rise to a disproportionate and unphysiological increase in dihydrotestosterone relative to testosterone. Oral testosterone undecanoate is useful in the induction of puberty in adolescents where lower doses are preferable, and as second line treatment in adults who are intolerant of injections or implants.

17a-alkylated androgens are relatively weak androgens but some may have more potent anabolic effects. 17a-alkylated compounds cause cholestatic jaundice in a reversible and dose-related manner while long-term treatment is associated with peliosis hepatis (haemorrhagic cysts in the liver) and liver tumours. Consequently, 17a-methyl testosterone, oxymetholone, and fluoxymesterone have now been withdrawn from the market in many countries. As a group, 17a-alkylated androgens are not recommended for clinical use but they are the most commonly abused anabolic steroids. Mesterolone, which is not hepatotoxic, is a weak androgen with low clinical efficacy but remains commercially available.

Transdermal testosterone preparations offer the advantages of stable physiological levels of testosterone without peaks and troughs, painless self-administration, minimal risk of overdosing and low potential for abuse. A 60 cm^2 translucent membrane (Testoderm™, ALZA, Palo Alto, United States) applied to the scrotum delivers testosterone at a rate of 4 or 6 mg per day. Daily renewal of Testoderm™ in the morning maintains plasma testosterone within the adult physiological range with a normal diurnal profile. 5a-Dihydrotestosterone levels are elevated because of abundant 5a-reductase activity in genital skin. This does not seem to have any adverse consequences even after several years of treatment. Local skin irritation is negligible. A non-scrotal transdermal system (Androderm™ or Andropatch™ SmithKline Beecham, Welwyn Garden City, United Kingdom) delivers testosterone at 2.5 mg (6.5 cm diameter) or 5 mg (13 cm diameter) daily. At a dose of 5 mg daily applied at bedtime, plasma testosterone, dihydrotestosterone, and oestradiol are maintained within the physiological range throughout 24 h with a small diurnal variation. While clinical efficacy is satisfactory, the major drawback of Andropatch™ is skin reaction at the application sites, which occurs in 60 to 70 per cent of cases. A new, non-scrotal testosterone (Testoderm TTS™, ALZA, Palo Alto, United States) has recently become available and is said to have a lower incidence of skin irritation.

Many novel androgen preparations are currently under clinical investigation or being developed. Most promising are testosterone gel, cyclodextrins, buccal preparations, and esters with long half-lives (testosterone undecanoate or decanoate in castor oil for intramuscular injection 2 to 3-monthly) and selective androgen receptor modulators such as 7a-methy-19-nortestosterone (MENT).

The choice of preparations depends on age of the patient, the patient's own preference, facilities for injections, and available experience for surgical implants. Many boys with constitutional delayed puberty will spontaneously enter or progress in puberty after short courses of testosterone, for example intramuscular testosterone enanthate 50 mg monthly or oral testosterone undecanoate 40 mg daily for 3 to 6 months. The low doses of testosterone will stimulate linear growth and promote virilization without premature epiphyseal fusion. In patients with no evidence of spontaneous progression, gradually increasing doses of testosterone over 3 to 4 years will ensure full virilization except for testicular growth. They can be maintained on adult replacement doses subsequently if hypogoandotrophic hypogonadism appears to be permanent. Testosterone treatment can be safely started after the age of 14. Indeed, delayed treatment can be associated with permanently impaired peak bone mass.

The invasive nature of the implantation procedure and the long duration of action make them less than ideal for the induction of puberty in adolescents and the initiation of treatment in androgen-naive young adults where a more gradual and flexible increase in dose is desirable. For these reasons, testosterone implant is usually reserved for maintenance treatment in young adults, replacement having been initiated with intramuscular or oral preparations. Almost all adult patients respond well to testosterone enanthate 200 mg 2-weekly, 300 mg 3-weekly or Sustanon 250 mg 2 to 3-weekly. In the absence of a satisfactory biological marker for androgen action, monitoring of treatment is best gauged by clinical response and documenting plasma testosterone within the low–normal range immediately before the next dose so that appropriate adjustments of dosing intervals can be made.

Hypogonadal patients over the age of 50 starting testosterone for the first time should be checked for pre-existing, occult prostatic cancer with digital rectal examination and prostate-specific antigen (PSA) measurement. These should also be repeated in the first 3 to 6 months after initiating treatment to ensure that there is no deterioration. Subsequent monitoring for prostatic disease should not differ from eugonadal men of comparable age since there is no increased relative risk in hypogonadal patients on long-term testosterone replacement.

Testosterone replacement therapy is safe and side-effects are rare. They may include acne, transient priapism, gynaecomastia, fluid retention, increase in haematocrit, obstructive sleep apnoea, and exacerbation of pre-existing behavioural disturbances. Testosterone is contraindicated in patients with known prostatic cancer and breast cancer. In older patients with benign prostatic hyperplasia, sleep apnoea, polycythaemia, dyslipidaemia, cardiac failure, liver disease, and renal failure, a cautious approach with reduced doses of testosterone, careful dose titration, and close supervision or specific management of the coexisting problems usually allow patients to benefit from androgen replacement.

Infertility

Infertility is defined as the inability of a couple to initiate a pregnancy after 12 months of unprotected intercourse. Some 8 to 15 per cent of married couples experience involuntary infertility. Of these, male factors alone are estimated to be responsible in 30 per cent and contributory in a further 20 per cent of subfertile couples. Thus, male infertility may affect 5 per cent of men of reproductive age. A secular trend of declining semen quality (sperm density) in men over the last 50 years has been reported in some but not other regions of Europe. This, together with a concurrent increase in incidence of testicular cancer, hypospadias, and cryptorchidism, has raised the question of possible environmental endocrine disruptors with oestrogenic or antiandrogenic actions influencing prenatal or neonatal testicular and genital tract development. The concern prompted the recent development of sensitive techniques for monitoring potential deleterious reproductive effects of environmental chemicals. However, there is currently no evidence that the incidence of male infertility is increasing.

Aetiologies

Male infertility, comprising a heterogeneous group of disorders (Table 2), represents the male partner's contribution to a couple's failure to conceive. This implied failure to fertilize normal ova is usually associated with defective spermatogenesis giving rise to absent (azoospermia) or low sperm output (oligozoospermia: <20 million/ml) and/or abnormal spermiogenesis giving rise to spermatozoa with poor motility (asthenozoospermia:<50 per cent of spermatozoa showing progressive motility) and abnormal morphology (teratozoospermia: <15 per cent normal forms). The pathogenic basis of defective spermaotgenesis or spermiogenesis remains poorly understood. Testicular histology may show quantitative reduction in all germ cell types (hypospermatogenesis), Sertoli cells only, or maturation arrest at the primary spermatocyte (premeiotic) or spermatid (postmeiotic) stage.

Idiopathic azoospermia/ oligozoospermia

By far the commonest form of male infertility (60 per cent) is idiopathic azoo/oligozoospermia, usually associated with asthenozoopermia and teratozoospermia. This probably represents the end result of a multitude of ill-defined pathologies which disrupt normal seminiferous tubular functions. However, recent molecular analyses have revealed that a substantial proportion of these cases hitherto classified as idiopathic have discrete gene defects associated with impaired spermatogenesis (see below).

Asthenozoospermia

Reduced velocity or vigour of sperm motility may be due to metabolic/ functional defects or ultrastructural malformations in the axonemal complex of the sperm tail usually associated with oligozoospermia or a high percentage of dead and abnormally-shaped sperm. The latter finding may indicate a recently-recognized condition, epididymal necro/asthenozoospermia. Testicular spermatozoa are normal, the defects occurring during epididymal transit. Rarely, complete asthenozoospermia (with normal sperm density) may result from absence of dynein arms (sites of Na/K ATPase activity) linking individual microtubules. This is associated with similar defects in respiratory cilia and a history of chronic respiratory infection, bronchiectasis, and sinusitis (immotile cilia syndrome). In addition, some of these patients have situs inversus (Kartargener's syndrome). Absence of the central pair of microtubules in the sperm tail is an even rarer cause of complete asthenozoospermia—the 9+0 syndrome.

Teratozoospermia

An extreme example of abnormal sperm morphology is the failure of acrosome cap development in the sperm head leading to formation of round-headed spermatozoa (globozoospermia) which are unable to bind to the zona pellucida of ova, a prerequisite for fertilization.

Chromosome disorders

Chromosome abnormalities identified by cytogenetic studies of blood lymphocyte are found in 15 per cent of azoospermic patients; 90 per cent of these have Klinefelter's syndrome. Other chromosomal abnormalities encountered include reciprocal X or Y autosomal translocations, XYY and XX males, reciprocal and robertsonian autosomal translocations, supernumerary autosomes, and inversion of autosomes.

Klinefelter's (47XXY) patients are azoospermic. Spontaneous pregnancies have been reported in Klinefelter's patients with 46XY/47XXY mosaicim. The mechanism by which an extra X chromosome gives rise to spermatogenic failure is not known. Inactivation of the X chromosome in primary spermatocytes is necessary for spermatogenesis to proceed normally through meiosis. Hyalinized seminiferous tubules devoid of germ cells are pervasive in the atrophic testes. Occasionally, isolated foci of tubules with preserved spermatogenesis can be identified in testicular biopsy of 47 XXY patients.

Y chromosome microdeletions

A major breakthrough in the understanding of the molecular genetics of male infertility is the recent characterization of three non-overlapping regions (designated azoospermic factors AZFa, ASFb, and AZFc) on the long arm of the Y chromosomes (Yq11) which contain multiple genes involved in spermatogenesis. Microdeletions in these AZF loci, identifiable only by PCR amplification of DNA but not routine karyotyping, have been found in 3 to 37.5 per cent of patients previously considered to have idiopathic azoospermia and severe oligozoospermia but not in fertile control populations. Larger deletions (involving more than one AZF locus) are associated with more severe testicular phenotypes and the incidence of microdeletions is highest amongst azoospermic patients with Sertoli cell-only histology. AZFc is by far the most frequently encountered deletion. Y chromosome microdeletions are emerging as the second most common specific aetiology of male infertility (after varicoceles).

Several cloned genes have been mapped to each of the AZF intervals. At least one strong candidate gene is associated with each deletion cluster— *DFFRY* in AZFa, *RBMY* in AZFb, and *DAZ* in AZFc. These are multicopy gene families scattered in both arms of the Y chromosome with the latter two being expressed only in the testis. The specific products of these candidate genes and their functional significance remain unclear. Male infertility associated with microdeletions of Y chromatin is probably attributable to reduced copy number of more than one of these gene families. Other, as yet unidentified, genes important in spermatogenesis within or outside the AZF loci of the Y chromosome are highly likely. A significant proportion of patients with microdeletions of the Y chromosome is oligozoospermic and not azoospermic. Transmission of specific Y chromosome microdeletions to male offspring by assisted conception techniques has been clearly documented.

Defects in target tissue

Mutations in the ligand binding or DNA binding domains of the androgen receptor cause defects in androgen action and varying degrees of failure of masculinization during primary sexual development (androgen insensitivity syndromes) despite raised levels of testosterone being produced by inguinal or intra-abdominal testes. These defects are, in descending order of severity:

- Complete testicular feminization—female phenotype and female external genitalia with absent uterus and Fallopian tubes. Presents with primary amenorrhoea.
- Incomplete testicular feminization—female phenotype and female external genitalia with minimal virilization such as clitoral hypertrophy and partial fusion.
- Reifenstein's syndrome—ambiguous genitalia with perineoscrotal hypospadias, poor penile development, bifid scrotum, and gynaecomastia at puberty.

In contrast to the above, expansion of CAG polyglutamine repeats to greater than 40 in the N-termial domain of the androgen receptor causes X-linked spinal bulbar muscular atrophy (Kennedy's disease) associated with gynaecomastia, poor virilization, and azoospermia due to 'late-onset' androgen resistance. Expansion of CAG glutamine repeats to between 25 to 40 is associated with a four-fold increased risk of oligozoospermia or azoospermia without clinical evidence of neuromuscular degeneration. This may represent a exclusively testicular form of androgen insensitivity.

5a-reductase-2 deficiency–deficient 5a-dihydrotestosterone action in the genital tract causes clitoral hypertrophy with perineoscrotal hypospadias and blind-ending pseudovagina, inguinal testes with epididymis and vas. Usually raised as girls, these patients dramatically virilize at puberty without gynaecomasita.

Males with oestrogen resistance and aromatase deficiency are normally virilized at birth and have normal pubertal development except for non-fusion of epiphyses resulting in extreme tall stature and osteoporosis in adulthood. Effects on spermatogenesis and fertility are currently unclear.

Cryptorchidism

Cryptorchidism has a prevalence of 2.5 to 5 per cent at birth which declines to 1 per cent by 1 year. Spontaneous descent rarely occurs after this age. Undescended testes can be a feature of many hypogonadotrophic conditions, and intersexual and dysgenetic states such as androgen insensitivity syndromes and Noonan's syndrome. The persistent mullerian duct syndrome is caused by defects in antimullerian hormone (AMH) production or action during fetal development. The presence of Fallopian tubes and uterus obstructs testicular descent. The lower temperature in the scrotum is a prerequisite for normal spermatogenesis. Undescended testes are therefore exposed to the harmful effects of the higher temperature in the abdomen and inguinal region. The testis which is not permanently in a low scrotal position by the age of 2 years will have sustained permanent damage to the seminiferous epithelium. Orchidopexy after 2 years of age for undescended testes does not improve fertility. For these reasons, treatment should ideally be undertaken between 1 and 2 years of age. hCG or intranasal GnRH are currently being increasingly used for early initial treatment of cryptorchidism. If hormonal treatment is unsuccessful, orchidopexy can be carried out by the age of 2 years. The risk of testicular tumour in a patient with a history of undescended testis, whether successfully treated by orchidopexy or not, is four to five-fold higher than the general population.

Testicular tumours

It is important to remember that infertility can be a presenting symptom of testicular tumours, the commonest malignancy in young adult men. With increasing use of testicular ultrasound, it has become clear that there is a significantly higher risk of testicular tumours in infertile men (in the absence of cryptorchidism) compared to the general population. Carcinoma *in situ*, an obligatory precancerous state, is occasionally encountered incidentally in diagnostic testicular biopsies. Without treatment, 50 per cent of carcinomata *in situ* progress to malignant seminoma or non-seminomatous germ cell tumours.

Varicocele

Varicocele is a dilatation of the scrotal portion of the pampiniform plexus due to reflux of blood in the internal spermatic veins, usually involving the left side from the renal vein. It usually gives rise to a reduction in ipsilateral testicular volume but varying degrees of hypospermatogenesis are often seen in both testes. Although a varicocele is clinically detectable in up to 40 per cent of male partners of infertile couples, its significance in male infertility remains controversial. Increased scrotal temperature, hypoxia, and exposure of the testes to adrenal metabolites have been postulated as possible mechanisms by which spermatic vein reflux can induce seminiferous tubular damage. Since varicoceles can be detected clinically in 15 per cent of fertile young men, it must not be assumed that this condition is invariably or solely responsible for infertility without actively excluding other possible aetiologies including those in the female partner.

Sperm autoimmunity

Immunological infertility is a specific disorder caused by sperm membrane-bound IgA antibodies found in around 5 per cent of men presenting with infertility. Conditions predisposing to sperm autoimmunity include vasectomy, testicular injury/inflammation, genital tract infection/obstruction, and family history of autoimmune disease. Male patients with significant antisperm antibody titres usually have severely suppressed fertility potential due to sperm agglutination, poor sperm transit through cervical mucus, and blocked sperm–oocyte fusion.

Genital tract infection

Infection in the lower genital tract is a major cause of male infertility in the global context. Chlamydia, gonococcus, Gram-negative enterococci, and tubercle bacillus are the usual pathogens. If not treated by appropriate antibiotics promptly, inflammation of the accessory glands and excurrent ducts may give rise to disturbed

function, formation of sperm antibody, and permanent structural damage with obstruction in the outflow tract. Asymptomatic prostatitis due to occult and usually focal infection is best diagnosed by transrectal ultrasound examination of the prostate.

Excurrent duct obstruction

Vasectomy and previous genitourinary infections, usually sexually transmitted or tuberculous, are the most common causes of obstructive azoospermia. Congenital bilateral agenesis of the wolffian duct-derived structures, corpus/cauda epididymis, vas deferens and seminal vesicles (CBAVD) characterized by impalpable scrotal vasa, distended caput epididymis, acidic non-coagulating semen of reduced volume (<2 ml) devoid of fructose and sperm is present in 95 per cent of males with cystic fibrosis. They carry homozygous or compound heterozygous mutations in the cystic fibrosis transmembrane regulator (*CFTR*) gene. More commonly (6 per cent of azoospermic men and 1–2 per cent of infertile males), patients present with CBAVD without frank respiratory tract disease or pancreatic insufficiency. They have milder heterozygous mutations and/or the 5T variant in intron 8 of the *CFTR* gene, giving rise to a predominantly genital phenotype of cystic fibrosis. Renal and urinary tract abnormalities are common in these patients. In Young's syndrome, progressive epididymal obstruction is due to progressive inspissation of amorphous secretion in the lumen. In these patients, the high incidence of chronic sinopulmonary infection from childhood and bronchiectasis is presumably the consequence of the same abnormality in the respiratory tract. Epidemiological data has recently raised the possibility of mercury poisoning in this condition.

Coital disorders

Inadequate coital frequency, technique (including the use of vaginal lubricants with spermicidal properties), and faulty timing of intercourse may contribute to continuing infertility but are rarely the only aetiological factor in the infertile couple.

Diagnosis

History

Particular attention should be paid to the following aspects. Previous surgery such as herniorrhaphy in childhood, trauma, or torsion suggests possible damage to the vas or testis. History of cryptorchidism and genitourinary infections are important aetiological factors. Delayed onset of puberty may suggest the possibility of gonadotrophin deficiency. A history of recurrent chest infection, sinusitis, or bronchiectasis may be obtained in patients with epididymal obstruction (Young's syndrome), immotile cilia syndrome, and CBAVD associated with cystic fibrosis. Chronic disorders such as renal failure, liver disease, malignancy, diabetes, and multiple sclerosis are associated with a variety of testicular and sexual dysfunctions. Each patient should be asked about episodes of pyrexia within the past 12 weeks because of transient suppression of spermatogenesis. Careful enquiry should also be made about occupational or environmental exposure to testicular toxins, radiation, current medications, previous treatment, or recreational drugs. Painful ejaculation, haemotospermia, and pain in the perineum are symptoms suggestive of chronic infection in the prostate and seminal vesicles. It is important to establish that vaginal intercourse takes place with appropriate frequency and timing without the use of vaginal lubricants.

Examination

Assessment of height, weight, body habitus, and secondary sexual development should be carried out in all patients. Measurement of testicular volumes by comparison with Prader's orchidometer provide a convenient clinical index of seminiferous tubular mass. Normal adult testicular volume is between 15 and 35 ml. Testicular volume is a key finding in differentiating between azoospermia due to seminiferous tubular failure (reduced volumes) and that arising from excurrent duct obstruction (normal volume). Testicular size is also a useful indicator of the degree of testicular development in hypogonadotrophic patients. If not in the scrotum, the lowest position of the testes should be defined with the patient upright. Irregular contour, induration, or abnormal consistency of the testis suggest previous orchitis, surgery, or malignancy. Special attention should also be paid to the palpation of the epididymis and scrotal vas. An enlarged and tense caput epididymis may be palpable in cases of obstructive azoospermia. Irregularity and induration of the epididymis and vas suggest previous infection. In congenital agenesis of wolffian duct-derived structures, the scrotal vasa are either impalpable or extremely thin. The patient should be examined standing so that varicoceles can become visible (grade 3) or palpable (grade 2), or detected as a venous impulse in the spermatic cord during valsalva manoeuvre (grade 1). Rectal examination may reveal irregular contour or abnormal consistency and tenderness in the prostate in the presence of chronic prostatitis and enlarged seminal vesicles due to ejaculatory duct obstruction.

Investigations

Conventional parameters of the semen analysis such as sperm density, percentage of motile sperm, quality of sperm movements, and sperm morphology provide a semiquantitative index of fertility potential. Although a variety of tests of sperm function, such as computer-aided sperm movement analyses, cervical mucus penetration, acrosome reaction, sperm-zona binding, and hamster oocyte penetration, have been devised, none are sufficiently reliable and accurate to be used routinely in clinical practice. Infertile men with oligozoospermia produce spermatozoa harbouring abnormal DNA with strand breaks and redundant cytoplasm which may produce excessive reactive oxygen species. Chromatin structure and cytoplasmic enzyme (LDH-X or CK-M) assays are being applied to assess functional integrity of spermatozoa. They may provide more reliable quantitative biochemical measures of male fertility to guide management in the future.

Measurement of plasma FSH is useful in distinguishing primary from secondary testicular failure and in identifying patients with obstructive azoospermia. In the presence of azoospermia or oligozoospermia, an elevated FSH, particularly with reduced testicular volume, is presumptive evidence of severe and usually irreversible seminiferous tubular damage. Low or undetectable FSH (usually associated with low LH and testosterone with clinical evidence of androgen deficiency) is suggestive of hypogonadotrophism. Conversely, azoospermia with normal FSH and normal testicular volume usually indicates the presence of bilateral genital tract obstruction. The potential role of inhibin B measurement as a circulating marker of Sertoli cell function in routine diagnostic workup of male infertility is currently being evaluated. Testosterone and LH measurements are only indicated in the assessment of the infertile male when there is clinical suspicion of androgen deficiency, Klinefelter's syndrome, or sex steroid abuse. A high LH and testosterone should raise the possibility of abnormalities in androgen receptors while low LH and testosterone suggest gonadotrophin deficiency. Hyperprolactinaemia is not a recognized cause of male infertility but prolactin measurement should be undertaken if there is clinical evidence of sexual dysfunction (particularly diminished libido) or pituitary disease leading to secondary testicular failure. Oestradiol measurement is rarely indicated except in the presence of gynaecomastia.

Chromosome analysis by karyotyping or fluorescent *in situ* hybridization (FISH) should be carried out in patients with azoospermia, testicular atrophy, and elevated FSH, primarily to confirm the diagnosis of Klinefelter's syndrome. Screening for Y chromosome microdeletions should be considered in all patients with sperm density less than 5 million/ml by an appropriate number of PCR-based DNA markers and confirmed by Southern blotting. The need for testicular biopsy has largely been superseded by the use of plasma FSH in recent years to differentiate between primary testicular failure and obstructive lesions. Undetectable or very low levels of seminal fructose is used to confirm the clinical diagnosis of vasal and seminal vesicle agenesis or blocked ejaculatory ducts in the presence of obstructive azoospermia. An increase in number (more than 1 million/ml) of peroxidase-positive or monoclonal antibody-detected leucocytes in the semen may indicate genital tract infection. Semen culture for pathogens are difficult because of the bactericidal properties of seminal plasma and urethral and skin commensals. Antisperm antibodies are detected by the mixed agglutination reaction where sheep red blood cells or polyacrylamide beads are coated with rabbit antibodies to specific classes of human Igs. These will attach to motile spermatozoa carrying specific IgA on the surface of the sperm head or tail. Ultrasound examination of the testis has become a routine investigation for infertile males with non-obstructive azoospermia or severe oligospermia to detect occult testicular tumours. In patients with persistent or treated cryptorchidism, testicular ultrasound should be carried out annually. Ultrasound of the urinary tract is indicated in patients with CBAVD. Transrectal ultrasound can aid the diagnosis of asymptomatic chronic prostatitis.

Management

Pregnancies can occur in subfertile couples without treatment albeit with a much reduced probability depending on the duration of infertility, age, and coexisting subtle abnormalities in the female partner in addition to the defects in sperm quality. Since the majority of patients with male infertility present no recognizable or reversible aetiologies, management remains largely empirical.

Subfertility due to idiopathic hypospermatogenesis

Although a wide variety of empirical medical treatments, including gonadotrophins, androgens, and antioestrogens, have been tried in attempts to improve fertility in subfertile men, none have been shown to be effective when assessed in randomized, controlled therapeutic trials and are therefore not recommended. Instead,

assisted conception techniques are increasingly applied to overcome idiopathic male infertility. This is based on the premise that placing a large number of 'prepared' motile spermatozoa in close proximity to ovulated or retrieved oocytes *in vivo* or *in vitro* can enhance the probability of fertilization. Intrauterine insemination (IUI) of more than 1 million washed, motile spermatozoa (freed of seminal plasma, leucocytes, and abnormal/ dead spermatozoa) is a relatively simple and inexpensive technique with few complications. Pregnancy rates of 5 to 10 per cent per cycle can be expected. This can be combined with controlled ovarian stimulation of the female using gonadotrophins but the risk of multiple pregnancies increases. *In vitro* fertilization (IVF) involves more intensive gonadotrophin stimulation of the female, suppression of spontaneous ovulation, and collection of multiple oocytes by laparoscopy or transvaginal-ultrasound-guided ovarian puncture which are then coincubated with prepared spermatozoa in culture medium. In patients with moderate oligozoospermia, average fertilization rates of 30 per cent and live birth rates of 5 to 12 per cent per treatment cycle can be expected. In those with severe and multiple defects in semen parameters, standard IVF is less effective. For these cases, microinjection of single live spermatozoon directly into harvested oocytes (intracytoplasmic sperm injection, ICSI) has become the treatment of choice. This bypasses the sperm–oocyte interactions normally required for fertilization in natural conception or IVF and can achieve a remarkably high fertilization and live birth rates ((55 and 26 per cent per cycle respectively) even with the most severely abnormal samples. Since only a few spermatozoa are required, ICSI has revolutionized management of extreme oligozoospermia and azoospermia irrespective of aetiology. Non-obstructive azoospermia is often intermittent and careful examination of centrifuged deposits of semen to detect and harvest occasional ejaculated spermatozoa for ICSI should be attempted repeatedly before resorting to alternatives. Even in patients with persistent azoospermia, isolated foci of spermatogenesis may be preserved so that testicular sperm extraction via multiple biopsies can often yield viable testicular spermatozoa (including several patients with Klinefelter's syndrome) for ICSI.

In obstructive azoospermia, epididymal spermatozoa can be aspirated by an open procedure or percutaneous needle puncture of the proximal epididymis. In these circumstances, cryostorage of harvested spermatozoa for subsequent ICSI is required; this does not appear to compromise efficacy. In children born after successful ICSI treatment, the incidence of major congenital abnormalities is not increased compared to natural pregnancies but there is a small increase in sex chromosome aneuploidy in some series. Concern regarding the developmental potential of children born after ICSI has been raised. Long-term follow-up of children from ICSI births is indicated.

Specific treatable conditions

Removal or withdrawal from antispermatogenic agent or drug exposure may lead to improvement in fertility. This is most commonly seen in patients with inflammatory bowel diseases changing treatment from sulfasalazine to 5-aminosalicylic acid which removes the offending moiety, sulfapyridine. Withdrawal from anabolic steroid abuse invariably leads to recovery of spermatogenesis although this may take many months because of the long half-lives of some preparations. Cryopreservation of semen should be offered to all male patients of reproductive ageing before commencing anticancer chemotherapy or testicular irradiation.

When patients with hypogonadotrophic hypogonadism desire fertility, they can discontinue exogenous androgen replacement and start on human chorionic gonadotrophin (hCG 2000 IU, subcutaneous, twice weekly) for 6 to 12 months. This should maintain normal testosterone levels. Patients with postpubertally-acquired gonadotrophin deficiency (e.g. from pituitary tumour) where spermatogenesis has previously been established, usually respond to hCG treatment alone to reinitiate germ cell development. If there is no spermatozoa in the ejaculate at the end of 12 months, human menopausal gonadotrophin (hMG), which contain both FSH and LH, or recombinant FSH should be added at 75 to 150 IU, subcutaneous, thrice weekly. Combined treatment may be required for a further 12 months. Most patients with congenital forms of hypogonadotrophic hypogonadism will require FSH to stimulate Sertoli cell division and initiate spermatogenesis. In general, around 70 per cent should show active spermatogenesis and 50 per cent could be expected to achieve spontaneous pregnancies even if sperm densities remain in the oligozoospermic range. Patients with hypothalamic GnRH deficiency, can be treated by pulsatile GnRH delivered 2-hourly by a battery-driven portable infusion minipump. Many find this form of chronic therapy impractical and too demanding. The outcome of treatment is similar to that obtained with exogenous gonadotrophin therapy.

Active infection in the genital tract should be treated by appropriate antibiotics (erythromycin, doxycycline, or norfloxacin) for 4 weeks for the patient and his partner.

Obstructive azoospermia due to epididymal obstruction can be treated by microsurgical epididymovasostomy. High pregnancy rates can only be achieved by a few experienced microsurgeon. A more feasible alternative is to obtain spermatozoa from the caput epididymis or efferent ducts proximal to the site of obstruction by direct needle aspiration (microepididymal sperm aspiration or percutaneous epididymal sperm aspiration) for use in assisted fertilization procedures (usually ICSI). In patients with CBAVD, *CFTR* mutation screening of the partner and genetic counselling should be undertaken beforehand because of the risk of cystic fibrosis in offspring.

Sperm antibody can be treated by immunosuppression with high-dose prednisolone 0.75 mg/kg per day or prednisolone 20 mg twice daily on days 1 to 10 and 5 mg on days 11 and 12 of the partner's cycle for three to six cycles. Side-effects are common, including irritability, sleeplessness, arthralgia, muscle weakness, peptic ulceration, glucose intolerance, and bilateral aseptic necrosis of femoral heads. Results of controlled trials of glucocorticoid treatment are conflicting. IVF and ICSI are increasingly being applied to manage immunological male infertility.

Varicocele can be treated either by open surgical ligation or transfemoral embolization of the internal spermatic veins. Results of treatment of varicocele from eight prospective, controlled therapeutic trials are confusing. Coexisting female factors contributing to infertility, insufficient samples size, high dropout rates, and lack of randomization/ blinding or sham procedures are some of the more important confounding variables which typify difficulties of treatment trials in male infertility. Nevertheless, the Royal College of Obstetricians and Gynaecologists recently concluded that treatment of varicocele in oligozoospermic, but not normospermic, subfertile men can significantly improve semen quality and pregnancy rate. The cost of varicocele treatment per live birth is less with surgical ligation (and embolization) than for assisted conception techniques.

Retrograde ejaculation can be treated medically with a-adrenergic, anticholinergic agents or imipramine. If unsuccessful, spermatozoa can be recovered from bladder catheterization and irrigation with culture medium for artificial insemination or IVF. Semen can be obtained by masturbation, vibrators, or electroejaculation from patients with various coital dysfunctions.

Untreatable sterility

Patients with persistent, non-obstructive azoospermia without retrievable postmeiotic germ cells, unable to undergo or failed to be helped by ICSI should be counselled regarding the options of continuing childlessness, adoption, and donor insemination.

Genetic screening and counselling

This has become important with the realization that genetic disorders could account for an increasing proportion of infertility previously believed to be idiopathic and that there is a high probability of transmitting infertility to male offspring if assisted reproductive treatment is successful. Furthermore, the long-term health of the ICSI offspring remains an unsettled question. Counselling should therefore be carried out in all couples considering microassisted fertilization techniques. It is also recommended that chromosome karyotyping and Y chromosome screening be performed in patients with azoospermia and severe oligozoospermia (less than 5 million/ml) regardless of the coexistence of other clinical abnormalities such as varicocele or cryptorchidism. This not only allows a firm diagnosis to be made but also encourages the clinician to forego empirical treatment and couples who conceived by assisted reproduction techniques may inform their son at a suitable age that he is likely to have fertility problems. Patients with obstructive azoopsermia due to CBAVD and their partners should undergo *CFTR* gene screening followed by genetic counselling if positive.

Erectile impotence

Erectile failure may be caused by neurological disorders such as autonomic neuropathy (usually complicating diabetes), multiple sclerosis and spinal injuries, vascular disease involving pelvic vessels, retroperitoneal and bladder neck surgery, medications (commonly a and b-adrenergic antagonists, psychotropic agents), alcohol abuse, severe systemic disease, psychological dysfunctions (including depression), relationship problems, androgen deficiency, and hyperprolactinaemia. Loss of libido characterizes androgen deficiency and hyperprolactinaemia, while normal spontaneous morning erection is suggestive of psychogenic impotence. Testosterone deficiency is uncommon (less than 5 per cent) in patients who present with erectile dysfunction without loss of libido. Management should aim to correct any reversible underlying disease (e.g. prolactinoma) or substitute offending medications. Androgen replacement is only indicated in patients with total or free plasma testosterone in the hypogonadal range. The use of phosphodiesterase inhibitor (PGE5 inhibitors), such as sildenafil, to enhance the neurovascular cGMP-mediated nitric oxide synthesis in penile vasculature has been remarkably successful in a treating a wide variety of erectile dysfunction. This has largely superseded the use of vacuum devices and intracavernosal injection of vasodilator agents such as papaverine or prostaglandin E1. These are reserved for cases with severe neurogenic

impotence unresponsive to PGE5 inhibitors.

Further reading

Bhasin S, ed. (1998). The therapeutic role of androgens. *Balliere's Clinical Endocrinology and Metabolism* **12**.

Griffen JE (1992). Androgen resistance—the clinical and molecular spectrum. *New England Journal of Medicine* **326**, 611–18.

Hargreave TB, ed. (1994). *Male Infertility*, 2nd edn. Springer-Verlag, Berlin.

Mooradian AD, Morley JE, Korenman SG (1987). Biological actions of androgens. *Endocrine Reviews* **8**, 1–27.

Nieschlag E, Behre HM, eds (1997). *Andrology Male Reproductive Health and Dysfunction*. Springer-Verlag, Berlin.

Nieschlag E and Behre HM, eds (1998). *Testosterone action. Deficiency. Substitution*, 2nd edn. Springer-Verlag, Berlin.

Royal College of Obstetricians and Gynaecologists (1998). *The Management of Infertility in Secondary Care—Evidence-based Clinical Guidelines no.3*. Royal College of Obstetricians and Gynaecologists Press, London.

Royal College of Obstetricians and Gynaecologists (1998). *The Management of Infertility in Tertiary Care—Evidence-based Clinical Guidelines no.6*. Royal College of Obstetricians and Gynaecologists Press, London.

Templeton A, Cooke I, O'Brien PMS, eds (1998). *Evidence-based Fertility Treatment*. Royal College of Obstetricians and Gynaecologists Press, London.

Wang C, ed. (1999). *Male Reproductive Function.* Endocrine Updates. Kluwer Academic Publishers, Boston.

World Health Organization (1992). *Guidelines for the Use of Androgens in Men*. WHO, Geneva.

Wu FCW, ed. (2000). Male fertility and infertility. *Balliere's Best Practice in Clinical Endocrinology*.

12.8.3The breast

H. S. Jacobs

Gynaecomastia

Gynaecomastia is defined as benign enlargement of the male breast caused by proliferation of the glandular components. Clinically the distinction from enlargement by fat tissue is made by examining the patient in the supine position, the breast being held between thumb and forefinger and the fingers gently moved towards the nipple. A firm or rubbery mobile disc-like mound of tissue arising concentrically from beneath the nipple and areola indicates the presence of gynaecomastia. The most important condition that needs to be excluded is carcinoma of the male breast. Cancer usually presents as a unilateral eccentric mass that is hard and fixed to underlying tissue; it may be associated with skin tethering, nipple discharge, or axillary lymphadenopathy. Mammography and fine needle aspiration are helpful in the differential diagnosis but if doubt remains biopsy is appropriate. While cancer of the male breast is rare, it has to be recognized that it is 16 times more common in Klinefelter's syndrome than in other men. Other causes of gynaecomastia are not, however, associated with an increased risk of breast cancer.

Pathogenesis

Microscopically breast tissue in both sexes appears identical at birth. It remains quiescent until puberty when, in boys, the ducts and surrounding mesenchymal tissue transiently proliferate, only to involute and ultimately to atrophy. Gynaecomastia is characterized by initial proliferation of the fibroblastic stroma and ductal system. Progressive fibrosis and hyalinization then occur in association with regression of the epithelial components. These regressive changes occur even if the stimulus (for example oestrogen treatment) continues. When gynaecomastia has been present for more than a year, clinical regression is rarely complete because the fibrosis persists even when the cause has been removed.

Since oestrogens stimulate and androgens inhibit development of breast tissue, gynaecomastia arises whenever there is an imbalance between these hormones. An alteration in the ratio of free androgen to free oestrogen, rather than a specific concentration of either, is thought to underlie most cases of gynaecomastia.

In men 98 per cent of testosterone is directly secreted by the testes, whereas the origin of oestrogen is more complex (Fig. 1): thus only about 15 per cent of oestradiol and less than 5 per cent of oestrone are directly secreted. In both cases the remainder is produced by extraglandular conversion (aromatization) of androgenic precursors in peripheral tissues such as adipose tissue, liver, and muscle. There is also substantial interconversion of oestrone and oestradiol. Treatment of normal men with human chorionic gonadotrophin results in an increase of directly secreted oestradiol in proportion to the increase of testosterone, so that while directly secreted oestradiol in normal men rarely amounts to more than 6 µg/day, when levels of luteinizing hormones are persistently high substantial amounts of oestradiol may be directly secreted by the testis.

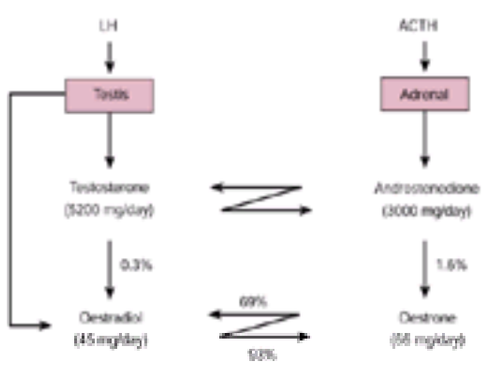

Fig. 1 Sources of oestrogen in men.

Causes

Gynaecomastia may be physiological or pathological. Physiologically it occurs at three times of life: the first is neonatally in response to transplacental passage of oestrogens, the second is during puberty for reasons that are not at all clear, and the third is in elderly men, probably because of the decline in Leydig cell function that occurs normally with age.

Pathological gynaecomastia is caused by a deficiency of testosterone formation or action, enhanced production of oestrogen, drugs, and unknown causes.

Testosterone deficiency

The commonest cause is Klinefelter's syndrome, about half the cases of which develop gynaecomastia at the time of puberty. The serum testosterone is usually about 50 per cent of normal, the gonadotrophin concentrations are raised, and the serum oestradiol is above normal. The diagnosis is suspected clinically and confirmed by white blood cell karyotype. Congenital causes of testosterone deficiency include defects in testosterone biosynthesis and congenital anorchia. Acquired causes include viral orchitis (usually mumps), trauma, neurological disease (myotonia dystrophica and spinal cord lesions), and renal failure. Androgen resistance syndromes are associated with high rates of secretion of testosterone and oestradiol; because of the large amounts of precursor (testosterone) there is also excessive extragonadal conversion of androgens to oestrogens.

Increased secretion of oestrogen

Testicular tumours, such as Leydig and Sertoli cell tumours, may secrete androgens and oestrogens autonomously; gonadotrophin secretion is therefore suppressed and azoospermia is common. They may be too small to be detected clinically but ultrasound can be very helpful. Some testicular tumours, for example choriocarcinomas, secrete human choionic gonadotrophin which then stimulates oestrogen secretion by the contralateral testis. Human choionic gonadotrophin may also be secreted by non-testicular tumours such as a bronchogenic carcinoma.

True hermaphroditism may be associated with gynaecomastia because of oestrogen secretion by the ovotestis.

Increased extragonadal production of oestrogens

Adrenal disease

Congenital adrenal hyperplasia caused by 2,1-hydroxylase or 3b or 17b steroid dehydrogenase deficiencies results in increased availability of adrenal androgen for peripheral aromatization. Adrenal carcinoma may be associated with massive oestrogen production, usually caused by extraglandular aromatization of the enormous amounts of androgen secreted by the tumour but occasionally directly secreted.

Liver disease

Cirrhosis, particularly alcoholic cirrhosis, is typically associated with gynaecomastia, testicular atrophy, and impotence. Plasma and urinary excretion of oestrogen is increased. The mechanism is in part decreased hepatic extraction of androstenedione and consequently an increase in its extrasplanchnic aromatization and partly reduced testosterone secretion by the testes. The gynaecomastia of starvation and refeeding may also be related to disturbed liver function.

Drugs

Oestrogens and oestrogen-like drugs

The most familiar use of oestrogen is in the treatment of advanced carcinoma of the prostate, indeed it is the development of gynaecomastia in this situation that has provided the model for most of our understanding of the evolution of the histological changes in the breast in gynaecomastia. Oestrogen residues in food (via injected animals) and cosmetic products have been reported as a cause of gynaecomastia.

Treatment with digitalis glycosides may cause gynaecomastia, the drug acting either as an oestrogen or as an oestrogen precursor.

Drugs that enhance oestrogen secretion

Treatment with human chorionic gonadotrophin and clomiphene can cause increased oestrogen secretion. The development of gynaecomastia in men treated with human chorionic gonadotrophin usually means that the dose used has been too high.

Drugs that inhibit testosterone secretion

Ketoconazole, an antifungal agent that is also used in the management of certain forms of Cushing's syndrome, blocks steroid synthesis in Leydig cells and, if a high dose is maintained, gynaecomastia may result.

Spironolactone causes gynaecomastia in as many as 50 per cent of men treated with 150 mg/day. The drug suppresses testosterone synthesis (by inhibiting 17,20-desmolase) but it also acts as a peripheral antiandrogen.

Drugs that block testosterone action

Cyproterone acetate and flutamide are two antiandrogens used in the management of advanced prostatic disease which usually produce gynaecomastia. Cimetidine, but not ranitidine, is antiandrogenic and is associated with a significant risk of gynaecomastia.

Whereas in most series between 50 and 75 per cent of cases of gynaecomastia are labelled idiopathic because no endocrinopathy can be identified, there are increasing reasons to suspect that environmental pollution, either with oestrogens or antiandrogens, is responsible for many of the cases.

Diagnosis

The history should include enquiry about drugs as well as possible environmental exposure to oestrogens and antiandrogens. Examination should include the testes. While small firm testes are characteristic of Klinefelter's syndrome, asymmetrical enlargement suggests a Leydig cell tumour (most readily diagnosed by ultrasound). Evaluation of alcohol intake and liver function is appropriate. Endocrine assessment should include measurement of testosterone, oestradiol, gonadotrophins, and dehydroepiandrosterone sulphate (high levels suggest adrenal disease) concentrations.

Treatment

Once gynaecomastia has been present for about a year medical treatment is unlikely to lead to a reduction in breast size because of the fibrosis which usually develops by this time. Consequently surgery is the mainstay of treatment. The psychological effects of persistent breast development in adolescent boys may be severe and surgery should be considered at an early stage: temporizing rarely produces resolution. Medical therapy with androgens or anti-oestrogens produces uncertain effects and can only be expected to have much benefit if administered early in the course of the disorder. Gynaecomastia caused by oestrogen treatment, as in the treatment of prostatic disease, can be prevented by pretreatment with low-dose irradiation of the breasts.

Galactorrhoea

Galactorrhoea is defined as a persistent discharge of milk or milk like secretion in the absence of parturition or beyond 6 months postpartum in a non-nursing mother. Galactorrhoea is not a sign of breast cancer and not a risk factor for it.

There are essentially two types of galactorrhoea—spontaneous galactorrhoea or galactorrhoea present on expression only. In the latter case the menstrual cycle is usually intact and an endocrine cause is rarely found. It is spontaneous galactorrhoea that is significant in endocrine terms and which is usually associated with amenorrhoea. It is in this situation that hyperprolactinaemia occurs.

In the puerperium there is a clear correlation between the amount of prolactin released in response to suckling and the volume of milk secreted. In contrast, in inappropriate lactation, that is, in galactorrhoea, no such relation exists, presumably because the breasts have not been prepared for lactation by the oestrogen- and progesterone-rich environment of pregnancy. When this observation is considered in relation to the ease and widespread availability of prolactin measurements, one can readily appreciate that the physical sign of galactorrhoea has ceased to have much diagnostic significance. In women with amenorrhoea caused by hyperprolactinaemia, for example, only about 20 per cent have galactorrhoea. Thus the evaluation of galactorrhoea nowadays is essentially the evaluation of hyperprolactinaemia.

Management

The essential investigation is the measurement of serum prolactin. The management of hyperprolactinaemia is outlined in Chapter 12.2. For women with non-hyperprolactinaemic galactorrhoea the important advice is first to stop expressing the milk, the second is the reassurance that it is not a sign of cancer, or a risk factor for it, and third is a trial of drug therapy with bromocriptine. In the author's experience, after they have received appropriate reassurance, patients with non-hyperprolactinaemic galactorrhoea rarely need drug therapy. Such patients are, however, very sensitive to the adverse effects of the drug and side-effects are common, even with low doses.

Further reading

Berchuck A, *et al.* (1998). Familial breast-ovarian cancer syndromes: BRCA1 and BRCA2. *Clinical Obstetrics and Gynecology* **41**, 157–66.

Haney AF (1997). Galactorrhea. *Current Therapy in Endocrinology and Metabolism* **6**, 393–6.

12.8.4 Sexual dysfunction

Raymond C. Rosen and Irwin Goldstein

Introduction

Sexual dysfunction is a common complaint in men and women and is associated with a broad range of medical, psychological, and interpersonal causes. Despite significant progress in basic research and clinical therapeutics in recent years, sexual problems remain among the most frequently overlooked and mismanaged patient complaints. Patient–physician communication difficulties, lack of knowledge, and inadequate reimbursement are frequently cited reasons for these shortcomings. Few physicians are adequately trained in the diagnosis and treatment of sexual dysfunction, and many patients seek assistance from inappropriate or unqualified providers. This trend is particularly unfortunate, since sexual problems frequently impact on patients' interpersonal functioning and quality of life. Moreover, new treatment options are available which offer significant therapeutic potential for many individuals.

Sexual problems are generally classified according to the four-phase model of sexual response originally proposed by Masters and Johnson. These include: (i) sexual desire disorders (hypoactive sexual desire disorder, sexual aversion disorder); (ii) sexual arousal disorders (erectile dysfunction, female arousal disorder); (iii) orgasmic disorders (female orgasmic disorder, premature or delayed ejaculation); and (iv) sexual pain disorders (dyspareunia, vaginismus). The specific dysfunctions and corresponding phases of the sexual response cycle are shown in Table 1.

Although public and professional attention has focused predominantly on erectile dysfunction, other sexual problems (e.g. hypoactive sexual desire, premature ejaculation) are more commonly reported in community-based surveys. Sexual dysfunction is more prevalent in women than men; recent British and American studies have found that about 40 per cent of women and 30 per cent of men have had one or more sexual problems in the past. Less than 10 per cent of these problems are typically brought to the attention of a physician or other health-care provider.

Medical management of sexual dysfunction should always begin with a thorough history and physical examination. In appropriate clinical contexts, at least one question regarding sexual function ought to be included in the initial examination of every patient, regardless of age or gender (such as 'Are you satisfied with your current sexual functioning or relationship?'). A physical examination and laboratory testing are important, since sexual problems are frequently associated with underlying medical illnesses or risk factors (such as diabetes, cardiovascular disease). The role of depression or other psychiatric disorders should also be considered, as well as possible iatrogenic effects of prescription or non-prescription drugs. It is also important to evaluate relationship and lifestyle factors in all cases. Finally, the patient's needs and expectations, as well as cultural or family issues should be taken into account.

Male sexual dysfunction

Male sexual dysfunction may be divided into erectile dysfunction, premature ejaculation, hypoactive sexual desire, and priapism.

Erectile dysfunction

Epidemiology

Erectile dysfunction is a significant and common medical problem, with recent epidemiological studies suggesting that approximately 10 per cent of men aged 40 to 70 have severe or complete erectile dysfunction, defined as the total inability to achieve or maintain erections sufficient for sexual performance. An additional 25 per cent of men in this age category have moderate or intermittent erectile difficulties. The disorder is highly age-dependent, as the combined prevalence of moderate to complete erectile dysfunction rises from approximately 22 per cent at age 40 to 49 per cent by age 70 (Fig. 1). Although less common in younger men, erectile dysfunction still affects 5 to 10 per cent of men below the age of 40. Findings from these studies show that erectile dysfunction impacts significantly on mood state, interpersonal functioning, and overall quality of life.

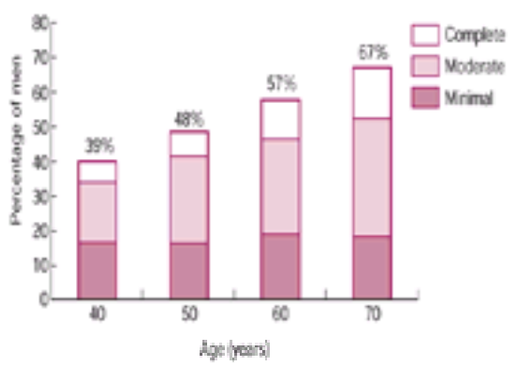

Fig. 1 Prevalence of erectile dysfunction (Feldman HA *et al.*, 1994).

Clinical context

Erectile dysfunction is strongly related to both physical and psychological health. Among the major risk factors are diabetes mellitus, heart disease, hypertension, and decreased high-density lipoprotein levels. Medications for diabetes, hypertension, cardiovascular disease, and depression may also cause erectile difficulties. In addition, there is a higher prevalence of erectile dysfunction among men who have undergone radiation or surgery for prostate cancer, or who have a lower spinal cord injury or other neurological diseases (such as Parkinson's disease, multiple sclerosis). Lifestyle factors, including smoking, alcohol consumption, and sedentary behaviour are additional risk factors. The psychological correlates of erectile dysfunction include anxiety, depression, and anger. Despite its increasing prevalence among older men, erectile dysfunction is not considered a normal or inevitable part of the ageing process. It is rarely (in fewer than 5 per cent of cases) due to ageing-related hypogonadism, although the relationship between erectile dysfunction and age-related declines in androgen remains controversial.

Pathophysiology

Basic research on neurovascular mechanisms has contributed greatly to our understanding of normal and pathological processes of erection. There is increasing evidence that the state of trabecular smooth muscle contractility is regulated by a delicate balance between neurotransmitter and vasoactive substances mediating erectile tissue contraction (consistent with flaccidity) and relaxation (consistent with erection). The neurotransmitter nitric oxide (NO) has been found to play a major role in inducing trabecular smooth muscle relaxation, as shown in Fig. 2: cGMP binds to cGMP-dependent protein kinases (PKG) and to cGMP-dependent ion channels, leading to lowering of intracellular calcium concentration and activation of myosin light-chain phosphatases, resulting in inhibition of smooth muscle contractility and enhancement of penile erection (Fig. 2).

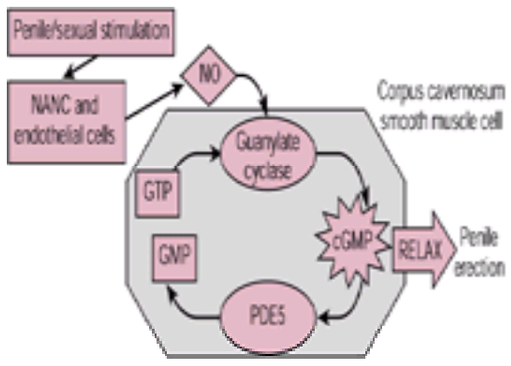

Fig. 2 Penile erection: the nitric oxide–cGMP mechanism. cGMP, cyclic guanosine monophosphate; GTP, guanosine triphosphate; NANC, non-adrenergic non-cholinergic neurones; NO, nitric oxide; PDE5, phosphodiesterase type 5.

Multiple vasoactive agents, including vasoactive intestinal polypeptide, prostaglandin E_1, forskolin, phosphodiesterase inhibitors, and a-adrenergic receptor antagonists affect smooth muscle contractility, each via a specific mechanism. Ultimately, all act by changes in intracellular calcium and modulation of specific smooth muscle myosin light-chain kinases and myosin light-chain phosphatases. These enzymes rapidly change the state of myosin phosphorylation and result in either smooth muscle contraction (flaccidity) or relaxation (erection). The role of central nervous system and spinal mechanisms has also been elucidated in recent studies. Understanding of these biochemical and neural mechanisms regulating erection has permitted the rational development of new pharmacological strategies for the management of erectile dysfunction.

Diagnosis and evaluation

Guidelines for diagnosis and evaluation were recently established by the 1st International Consultation on Erectile Dysfunction (Table 2). According to these guidelines, the first step in the process is the taking of a comprehensive sexual, medical, and psychosocial history. In obtaining a sexual history, special attention should be paid to personal or cultural sensitivities. History taking should be aimed at characterizing the severity, onset, and duration of the problem, and evaluating the need for specialized testing. A physical examination and selected laboratory testing should be performed on all patients with complaints of erectile dysfunction. Although not different from a routine physical examination, special emphasis is placed on review of genitourinary, endocrine, vascular, and neurological systems. The physical examination may corroborate aspects of the medical history (such as poor peripheral circulation), and may occasionally reveal unsuspected physical findings (such as Peyronie's plaques, small testes, prostate cancer). The physical examination also provides an opportunity for patient education and reassurance regarding normal genital anatomy.

Selective laboratory testing should be performed in all cases. This includes investigation of the hypothalamic–pituitary–gonadal axis via assessment of androgenic status, particularly if sexual desire is reduced. There is disagreement about the relative value of the various testosterone assays, including total, free, and bioavailable testosterone. However, strong consensus exists that at least one of these assays should be performed. A serum prolactin determination may be obtained in selected cases. Standard serum chemistries, full blood count, and lipid profiles may be of value and should be obtained, if not performed in the past year. Determination of serum thyroid-stimulating hormone may also be of value, as both hyper- and hypothyroidism are associated with erectile difficulties. Finally, measurement of serum prostatic specific antigen may be indicated based upon the patient's age and relative risk status.

Specialized diagnostic procedures, such as nocturnal penile tumescence and rigidity testing or other specialized vascular or neurological techniques, may play a role in selected cases. For example, CT and MR imaging, penile Doppler ultrasound, penile arteriography, cavernosal/pudendal/sphincter electromyography, or other tests may be of value in evaluating young patients with pelvic or penile trauma who may be candidates for reconstructive vascular surgery. Patients with complicated diabetes or other endocrinopathies may benefit from further endocrinological studies. Similarly, patients with complicated psychological or relationship problems may be suitably referred to a mental health or sex therapy specialist. Finally, patients with a history of cardiac disease or significant cardiovascular risk factors should be evaluated for potential cardiac risk associated with sexual activity: consensus guidelines have recently been established (Table 3).

Treatment

Results of the initial evaluation and specialized testing should be carefully reviewed with the patient and (if possible) their partner prior to initiating therapy. Potentially modifiable risk factors, such as cigarette smoking or alcohol abuse, should be addressed. Prescription drugs such as antihypertensives or antidepressants may be implicated in the patient's erectile difficulties, and should be altered when medically possible. Patients with specific endocrine deficiencies such as hypogonadism should be placed on hormone replacement therapy prior to initiation of direct therapies for erectile dysfunction. Sexual problems in the partner, such as a lack of lubrication, hypoactive sexual desire, or dyspareunia (painful intercourse) should also be addressed. Patients and partners should be fully informed about the range of treatment options available and the risks and benefits associated with each.

Direct therapies for erectile dysfunction can be stratified according to the mechanism of action, degree of invasiveness, ease of administration, reversibility, and relative costs associated with each (Table 4).

Sexual counselling and education

First-line options include sexual counselling or education, and oral agents (such as sildenafil). These options can be used alone or in combination. Brief sexual or couple's counselling is aimed at resolving specific psychological or interpersonal causes, such as relationship distress, sexual performance concerns, dysfunctional communication patterns, and comorbid sexual dysfunctions (such as hypoactive sexual desire). Advantages of sexual counselling include its non-invasiveness and relatively broad applicability. The major disadvantages are the lack of acceptability for many patients and uncertain efficacy.

Oral therapies

Sildenafil citrate (Viagra) is the first oral drug to be widely approved for the treatment of erectile dysfunction. It is a potent and selective inhibitor of type 5 phosphodiesterase, the primary form of the enzyme found in human penile erectile tissue (Fig. 2), thereby preventing the breakdown of cyclic guanosine monophosphate (cGMP), the intracellular second messenger of nitric oxide. Because nitric oxide is released following sexual stimulation, sildenafil only works when a man is sexually stimulated. Its mechanism of action is entirely peripheral and includes direct effects on smooth muscle relaxation and vasodilation of the penile arterioles. Sildenafil is administered 'on demand' in dosages of 25, 50, or 100 mg, and is effective in approximately 30 to 60 min, requiring ongoing sexual stimulation to be effective. Cardiac safety does not appear to be a major concern, except for patients receiving nitrates in any form, or those who have other cardiac risk factors associated with sexual activity itself (Princeton Consensus). Sildenafil is contraindicated for men receiving nitrate therapy, including short- or long-acting agents delivered by oral, sublingual, transnasal, or topical administration. Efficacy has been reported at 43 to 82 per cent, depending upon aetiology and severity of erectile dysfunction. The drug's side-effects include headaches, flushing, dyspepsia, and nasal congestion. A small percentage of men (2 to 3 per cent) may also experience mild alterations in colour vision (blue hue), visual brightness or sensitivity, or blurred vision.

Other oral agents are in development at the time of writing: some of these are based on a similar mechanism of action to sildenafil (i.e. PDE-5 inhibition), whereas others depend upon central dopaminergic activity, the dopamine–oxytocin pathway that originates from the hypothalamus and pituitary areas in the brain, or peripheral and/or central sympathetic inhibition. Some of these agents are in advanced clinical trials, but none has received regulatory approval yet in the United States or Europe. Combination oral agents are also being evaluated in some trials, and may play an important role in the future.

Local therapies

Injectable, intraurethral, or topical agents are classified as second-line therapies for erectile dysfunction according to recent guidelines. These should be selected based upon: (i) failure, insufficient response, or adverse side-effects associated with one or more of the first-line therapies, or (ii) patient/partner preferences. These interventions consist of intraurethral administration or intracavernosal injection of alprostadil. Vacuum pump devices can also be included in this category. Although widely utilized, these treatments are associated with variable efficacy, a high patient discontinuation rate, possible risk of side-effects, and moderately high costs.

Intracavernosal injection therapy

Prior to the approval of sildenafil, intracavernosal self-injection was the most common medical therapy for erectile dysfunction. Different forms of prostaglandin are used primarily for this purpose: alprostadil sterile powder and alprostadil alfadex are both synthetic formulations of prostaglandin E_1. Injection therapy is effective in most cases of erectile dysfunction, regardless of aetiology. It is contraindicated in men with a history of hypersensitivity to the drug employed, in those at risk for priapism (e.g. sickle-cell disease, hypercoagulable states), and in men receiving monoamine oxidase inhibitors. The effective therapeutic range is between 1 and 60 µg with the majority of responders (85 per cent) requiring less than 20 µg. In general, intracavernosal injection therapy with alprostadil is effective in 70 to 80 per cent of patients, although discontinuation rates are high in most studies. Side-effects include prolonged erections or priapism, penile pain, and fibrosis with chronic use. In addition to single-agent injection therapy, various combinations of alprostadil, phentolamine, and/or papaverine are widely employed in urological practice. Intracavernosal injections have been shown to be effective in about 70 to 80 per cent of men who fail first-line therapy with sildenafil.

Intraurethral alprostadil

Alprostadil may be administered intraurethrally in the form of a semi-solid pellet inserted by means of a special applicator. To obtain an effective concentration of alprostadil in the corpora cavernosa, 125 to 1000 µg of the drug are required. In a mixed group of patients with organic erectile dysfunction, 65 per cent of men receiving intraurethral alprostadil responded with a firm erection when tested in the office, and 50 per cent of administrations to that subset resulted in at least one episode of successful intercourse in the home setting. Side-effects associated with the intraurethral administration of alprostadil include penile pain and hypotension. Prolonged erections and penile fibrosis are rare, although the clinical success rate is low.

Vacuum constriction device therapy

The use of vacuum constriction device (**VCD**) therapy is a well-established, non-invasive treatment that has recently been approved by the United States Food and Drug Administration for over-the-counter distribution. It provides a useful treatment alternative for patients for whom pharmacological therapies are contraindicated, or who do not desire other interventions. Vacuum constriction devices apply a negative pressure to the flaccid penis, thus drawing venous blood into the penis, which is then retained by the application of an elastic constriction band at the base of the penis. Efficacy rates of 60 to 80 per cent have been reported in most studies. Like intracorporal injection therapy, VCD treatment is associated with a high rate of patient discontinuation. The adverse events occasionally associated with VCD therapy include penile pain, numbness, bruising, and delayed ejaculation.

Surgical treatments

Surgical implantation of a penile prosthesis, which was at one time the mainstay of treatment for erectile dysfunction, is now performed only in rare or special cases. For select cases of severe, treatment-refractory erectile dysfunction, for patients who fail pharmacological therapy or who prefer a permanent solution for the problem, surgical implantation of a semi-rigid or inflatable penile prosthesis is available. Various types of surgical prostheses have been described in the literature. The inflatable penile prosthesis provides a more aesthetic erection and better concealment than semi-rigid prostheses, although there is an increased rate of mechanical failure and complications (5 to 20 per cent) with the former. Despite the cost, invasiveness, and potential medical complications involved, penile implant surgery has been associated with high rates of patient satisfaction in previous studies. It should be noted, however, that these studies were conducted prior to the advent of newer forms of therapy (e.g. sildenafil).

Premature ejaculation

Definition and epidemiology

Premature ejaculation is difficult to define precisely. In part, it depends on the timing or speed of the partner's response, and whether satisfactory intercourse has been achieved. The preferred definition currently is that offered by DSM-IV: 'Persistent or recurrent ejaculation with minimal sexual stimulation or before, upon, or shortly after penetration and before the person wishes it'. In making the diagnosis, clinicians should consider the circumstances of the problem, including the degree of associated distress in the male and his partner. Some authors distinguish between primary or life-long premature ejaculation, and secondary or situational forms of the disorder. Other aspects of sexual functioning should be carefully evaluated, particularly since premature ejaculation can develop as a secondary reaction to erectile dysfunction. Otherwise, little is known about either the pathophysiology or aetiology of the disorder.

Premature ejaculation is a common disorder, affecting approximately 25 to 40 per cent of adult men at some time. In the National Health and Social Life Survey, approximately one-third of men in each age cohort reported problems with ejaculating too rapidly. Despite its greater prevalence, premature ejaculation has attracted less attention among health professionals and the lay public than erectile dysfunction. This may be due, in part, to the fact that men are able to function sexually (i.e. to perform intercourse) despite a mild or moderate degree of premature ejaculation. In the most severe cases, however, the male ejaculates before penetration is achieved. Far fewer men with premature ejaculation seek professional help for their problem, including those with the most severe forms of the disorder. This is unfortunate, since simple and effective therapies are widely available.

Treatment

Treatment of premature ejaculation consists of either behavioural/sex therapy training or pharmacological treatment. Masters and Johnson popularized the most widely used 'Stop–Start' technique in the 1970s. The method involves direct stimulation of the male until premonitory sensations are experienced just prior to orgasm. All stimulation is stopped at this point. After repeated practices over 4 to 8 weeks, the male becomes more aware of these anticipatory sensations and is able to delay or control his ejaculation to a much greater degree. This conditioning technique is often effective when practiced regularly by the male and his partner. However, it demands significant motivation on the part of the couple, and treatment efficacy is not always maintained over time.

The use of new pharmacological treatments, particularly the serotonin-uptake inhibiting drugs (**SSRIs**), has grown considerably in recent years. Among the specific drugs used for this purpose, sertraline and paroxetine have been most extensively studied. Clomipramine, which is not a true serotonin-uptake inhibitor, although

highly serotonergic, appears the most potent inhibitor of ejaculation to date, but can cause unpleasant side-effects, such as drowsiness or light-headedness, and may also cause decreased sexual desire or loss of erectile ability in some patients. Some authors recommend initial use of one of these agents on a daily basis, and then as needed during subsequent weeks. In one study, rapid ejaculation resumed shortly after withdrawal of clomipramine therapy that had been administered for the previous 8 weeks. To guard against this, patients should be encouraged to practice conditioning exercises along with use of the medication, and a combination of drug and non-drug therapies for premature ejaculation are currently recommended by most authors, although no controlled studies have been performed.

Hypoactive sexual desire

Hypoactive or low sexual desire is reported by about 10 to 15 per cent of men in population-based or community studies. It may be due to a variety of causes, including endocrine disorders (such as hypogonadism), other medical conditions (such as renal insufficiency), psychiatric illnesses (such as depression), or psychological factors. Various prescription (such as SSRIs) and non-prescription drugs (such as alcohol) can temporarily suppress sexual desire, as can partner conflicts or other sexual problems (such as erectile dysfunction).

Men with chronic low desire should receive a thorough medical and endocrine examination, including evaluation of the hypothalamic–pituitary–gonadal axis. Treatment should be based on identification of the relevant aetiological cause, wherever possible. Currently, there are no pharmacological agents which are safe and effective for treatment of hypoactive desire disorders in men.

Priapism

Priapism is a pathological condition defined by a persistent penile erection, greater than 4 to 6 h in duration, which persists after the cessation of sexual stimulation. Priapism is distinguished from prolonged erection, defined by a penile erection lasting up to 4 h, that spontaneously undergoes detumescence. There are two forms of priapism, veno-occlusive priapism, also referred to as low-flow or ischaemic priapism, and arterial priapism. It is essential to distinguish veno-occlusive from arterial priapism since the former is an emergency urological condition.

Veno-occlusive priapism

Introduction

Veno-occlusive priapism, the more common form of the disorder, typically presents as painful, tender, and persistent penile erection. Unless treated, it may cause irreversible damage to the erectile tissue, corporal fibrosis, and permanent erectile dysfunction. Veno-occlusive priapism is the consequence of a failure to regulate corporal veno-occlusion and is a 'closed-compartment syndrome' associated with absent cavernosal arterial inflow. The pathophysiology is based on unremitting corporal venous outflow obstruction, similar to other closed-compartment syndromes in the arms and legs.

Veno-occlusive priapism typically develops in stages. At first, in the presence of complete corporal veno-occlusion, cavernosal arterial inflow persists while intracavernosal pressures are lower than cavernosal artery systolic perfusion pressures. The presence of persistent cavernosal arterial inflow without corporal venous outflow results in intracavernosal pressure and volume increasing until maximal values of each are achieved.

In the next stage, intracavernosal pressure approximates cavernosal artery systolic pressure, thereby preventing further cavernosal arterial inflow. The consequences of absent cavernosal arterial inflow over time include: (i) an increase in corporal PCO_2, (ii) a decrease in corporal pH, and (iii) a decrease in corporal PO_2. The development of corporal tissue acidosis and hypoxia in this stage of veno-occlusive priapism results in adverse biochemical and cellular changes in the corporal tissue, which act to perpetuate the completeness of corporal veno-occlusion. A partial reversal of this process may occur during this phase with pharmacological or surgical intervention, when intracavernosal pressure lower than the cavernosal artery systolic pressure is transiently achieved. However, unless the primary aetiological cause is addressed, veno-occlusion and prolonged erection will probably recur. Over time, the intracavernosal blood will again become acidotic and hypoxic. Individuals in the intermediary second stage will either resolve the veno-occlusive priapism or progress to the third and final stage.

In the final stage of veno-occlusive priapism, irreversible changes to erectile tissue result from persistent exposure to acidotic and hypoxic corporal blood, which is clinically described as black, crankcase oil in appearance. It is not yet known when irreversible erectile tissue changes occur in an individual patient. It is likely that many factors contribute to the conversion from reversible stage II to irreversible stage III veno-occlusive priapism.

Aetiology

The main causes of venous priapism are listed in Table 5. Failure to regulate corporal veno-occlusion may be secondary to extraluminal or intraluminal obstructive pathophysiologies of the subtunical venules. Classic extraluminal obstructive mechanisms include pharmacological agents that induce persistent corporal smooth muscle relaxation, such as a-blockers and drugs with a-adrenergic effects (trazodone, clozapine), or intracavernosal/intraurethral agents (papaverine, prostaglandin E_1). Androgen administration may also cause priapism, perhaps related to facilitation of nitric oxide synthase (NOS) enzyme activity. Other extraluminal obstructive pathophysiologies are related to persistent neurological stimuli from central nervous system disorders such as spinal stenosis, as well as from erectile tissue infiltration with metastatic or local invasive tumours. Common intraluminal obstructive mechanisms include hyperviscosity, haematological disorders (sickle-cell disease, leukaemia), 20 per cent lipid administration in total parenteral nutrition, or fat embolism (pelvic fractures).

Management

A precise diagnosis should be established in all cases of priapism prior to therapeutic intervention. Veno-occlusive priapism can be diagnosed either by obtaining blood gas values (PO_2, pH, PCO_2) or determining cavernosal artery Doppler flow (absent in veno-occlusive priapism; present and bounding in arterial priapism). Aspects of the patient's history, particularly the presence or absence of penile pain, may also be of value in distinguishing veno-occlusive from arterial priapism. In patients with intact sensory nerves (i.e. excluding spinal cord injury, diabetes, etc.) veno-occlusive priapism is inevitably painful because of absent arterial inflow and resulting ischaemia. Arterial priapism is rarely associated with significant penile pain, although discomfort may be reported, particularly in the perineum—often the site of trauma. Since priapism may result in irreversible erectile dysfunction, it is important to establish the premorbid erectile function of the patient.

Other important aspects of the history include the occurrence and duration of previous episodes of prolonged erection or priapism, and response to previous therapies. Individuals with previous episodes of prolonged erection, often referred to clinically as 'stuttering' priapism, should be assessed for the presence of haematological conditions, such as sickle-cell disease or trait. Past or current use of intracorporal vasorelaxants (papaverine, prostaglandin E_1), psychotropic agents (trazodone, clozapine), a-blockers (prazosin), anticoagulants (heparin, coumarin), recreational drugs (cocaine), and over-the-counter adrenergic agonists (phenylephrine) should also be carefully assessed. Other aetiological factors include: penile or perineal trauma, such as falling on to a bicycle bar; haematological neoplasms such as leukaemia; metastatic or locally infiltrating neoplasms such as renal, bladder, or prostate carcinoma; neurological conditions such as spinal cord compression or acute lumbosacral disc prolapse; coagulopathy, such as disseminated intravascular coagulopathy; and elevated intravascular fat levels such as during total parenteral nutrition or following a fat embolism.

The goal of treatment for veno-occlusive priapism is to restore normal cavernosal arterial inflow. When appropriate, based on the duration and severity of the priapism, initial treatment may consist of non-specific medical management. This will involve aspiration of intracavernosal blood and/or interval saline lavage of the corpora cavernosa with large-bore butterfly needles (bilateral if indicated), with or without a penile block (1 per cent lidocaine without adrenaline). If cavernosal arterial blood flow is not re-established, intracavernosal adrenergic agonist therapy may be initiated to induce corporal smooth muscle contraction pharmacologically and thus initiate detumescence. Interval intracavernosal irrigation with large-bore butterfly needles may also be used. When administering intracavernosal adrenergic therapy, consideration for cardiovascular monitoring should be considered, especially if the agent selected has significant b_1-activity and the patient has a history of cardiovascular disease.

When veno-occlusive priapism is related to a specific aetiology, full resolution typically requires specific medical management. Such patients include those who have utilized pharmacological agents which have induced persistent corporal smooth muscle relaxation, who have a neurogenic pathophysiology, erectile tissue infiltration, or intraluminal obstructive mechanisms such as sickle-cell disease, leukaemia, or 20 per cent lipid administration in total parenteral nutrition. All remediable causes of priapism should be managed with disease-specific therapy. This may include discontinuation of the drug causing the persistent smooth muscle relaxation (cocaine, trazodone, etc.), irradiation (leukaemia), hydration and exchange transfusion (sickle-cell disease), appropriate therapy for any neurological condition, or substituting

10 per cent for 20 per cent lipids in total parenteral nutrition.

Surgical treatment should be considered if the erection recurs despite repeated adrenergic agonist administration, thereby resulting in loss of cavernosal arterial inflow. This may consist of a Winter's procedure, which involves the development of percutaneous shunts between the corpus spongiosum and the corpora cavernosa. Should a percutaneous shunt not restore arterial inflow in the cavernosal arteries, an Al Ghorab (distal) shunt is indicated. A transverse incision is fashioned in the glans penis 1 cm proximal to the corona. Blunt dissection allows exposure and excision of the distal portions of the corpora cavernosa. The glans is then closed, creating an open fistula between the corpora cavernosa and corpus spongiosum. Several proximal shunting procedures have been described for the management of recalcitrant veno-occlusive priapism, including cavernovenous (cavernosaphenous and cavernodorsal vein) and proximal cavernosal–spongiosal (Quackles). The role of these procedures and their indications are not clearly defined.

Arterial priapism

Arterial priapism is a rare condition involving the inability to regulate arterial inflow, the usual cause being trauma to the perineum or penis. Such trauma can result in a lacerated cavernosal artery that sends arterial blood directly to the lacunar spaces, bypassing physiological helicine arteriolar resistance mechanisms. An arterial–lacunar fistula can often be identified on duplex Doppler ultrasonography (usually located in the perineum) and selective internal pudendal arteriography. Clinical presentation is with a painless, non-tender, moderately (not fully) rigid penile erection.

Arterial priapism is not a medical emergency since the condition does not involve absent cavernosal arterial inflow with resultant erectile tissue ischaemia. Management involves 'watchful waiting'. For patients who wish to undergo treatment, surgical ligation of the cavernosal artery fistula or arterial embolization is typically performed.

Female sexual dysfunction

Introduction

Female sexual dysfunctions are broadly defined as alterations or disturbances in sexual functioning in women that are associated with subjective dissatisfaction or distress. These consist primarily of disorders of sexual desire, arousal, orgasm, and/or sexual pain (Table 6). While each of these disorders is uniquely defined, there is often significant overlap in affected patients. As noted above, about 40 per cent of adult women in community-based studies complain of one or more sexual dysfunctions in the past year. There has been limited investigation of the anatomy, physiology, and molecular biology of female sexual response, and the aetiology and pathophysiologies of specific sexual dysfunctions in women are not well understood. Stimulated in part by advances in male sexual dysfunction, there is increasing interest in basic science and clinical management aspects of female sexual dysfunction.

Anatomy and physiology

There are multiple anatomical structures which comprise the internal and external female genital tract, including the clitoris, labia minora, and corpus spongiosum (vestibular) erectile tissue, periurethral glans, urethra, anterior fornix, pubococcygeus muscle, and cervix. There are also multiple non-genital peripheral anatomical structures involved in female sexual response, such as salivary and sweat glands, cutaneous blood vessels, and nipples.

Internal and external genitalia

The vagina consists of a cyclindrical sheath of autonomically innervated smooth muscle (longitudinal outer, inner circular layer) lined by stratified squamous epithelium and a subdermal layer rich in capillaries. The vaginal wall consists of an inner glandular mucous-type stratified squamous cell epithelium supported by a thick lamina propia. This epithelium undergoes hormone-related cyclical changes, including slight keratinization of the superficial cells during the menstrual cycle. Deep to the epithelium lies the smooth muscle of the muscularis. There is a surrounding deeper fibrous layer above the muscularis, which provides structural support to the vagina and is rich is collagen and elastin, to allow for expansion of the vagina during sexual stimulation. Three sets of skeletal muscles surround the vagina, including the ischiocavernosum, bulbocavernosus, transverse perinei and levator ani, and pubococcygeus muscles.

The main arterial supply to the vagina flows to its superior aspect and arises from vaginal branches of the uterine artery. The middle portion of the vagina receives blood from the inferior vaginal artery, a branch of the hypogastric artery. The middle haemorrhoidal and the clitoral arteries send branches to the distal aspect of the vagina.

Autonomic efferent innervation of the vagina originates from the hypogastric plexus and the sacral plexus. These give rise to the uterovaginal nerves that contain both parasympathetic and sympathetic fibres, travelling within the uterosacral and cardinal ligaments to supply the proximal two-thirds of the vagina and the corporal bodies of the clitoris. Somatic afferent innervation is provided by the pudendal nerve, which reaches the perineum through Alcock's canal. There is an abundance of nerve fibres in the distal and anterior aspect of the vagina: these play a major role in sexual function and can easily be injured during pelvic surgery, or from blunt perineal trauma.

The vulva includes the labia minora, labia majora, the clitoris, the urinary meatus, the vaginal opening, and the corpus spongiosum erectile tissue (vestibular bulbs) of the labia minora. The labia majora are fatty folds covered by hair-bearing skin that fuses anteriorly with the mons veneris, or anterior prominence of the symphysis pubis, and posteriorly with the perineal body or posterior commissure. The labia minora are smaller folds covered by hairless skin laterally and by vaginal mucosa medially, and which fuse anteriorly to forms the prepuce of the clitoris, and posteriorly in the fossa navicularis. The perineal branch of the pudendal nerve innervates the labia. The main arterial supply arises from the inferior perineal branch of the internal pudendal, and by branches of the femoral artery.

The corpus spongiosum erectile tissues are paired structures located beneath the skin of the labia minora. The arterial supply is by the bulbar and posterior labial branches of the internal pudendal artery. The corpus spongiosum terminates in the glans clitoris. The corpora cavernosa of the clitoris measure up to 13 cm in length. The body of the clitoris consists of two paired erectile chambers composed of endothelial-lined lacunar spaces, trabecular smooth muscle, and trabecular connective tissue (collagen and elastin) surrounded by a fibrous sheath, the tunica albuginea. The arteries include the dorsal and clitoral cavernosal arteries, which arise from the iliohypogastric pudendal bed. The autonomic efferent motor innervation occurs via the cavernosal nerve of the clitoris, arising from the pelvic and hypogastric plexus.

As described above, the sexual response cycle in both men and women occurs in four stages: desire, arousal, orgasm, and resolution. The mechanisms underlying sexual response are less well understood in women than in men, particularly the role of central nervous system and hormonal processes. Sexual arousal responses in the genital and non-genital peripheral anatomical structures are largely the product of spinal cord reflex mechanisms. The spinal segments are under descending excitatory and inhibitory control from multiple supraspinal sites. The afferent reflex arm is primarily via the pudendal nerve; the efferent reflex arm consists of co-ordinated somatic and autonomic activity. One spinal sexual reflex is the bulbocavernosus reflex involving sacral cord segments S2, 3, and 4 in which pudendal nerve stimulation results in pelvic floor muscle contraction. Another spinal sexual reflex involves vaginal and clitoral cavernosal autonomic nerve stimulation and results in clitoral, labial, and vaginal engorgement.

In the unstimulated state, clitoral corporal and vaginal smooth muscles are under contractile tone. Following sexual stimulation, neurogenic- and endothelial-mediated release of nitric oxide (NO) plays a key role in clitoral cavernosal artery and helicine arteriolar smooth muscle relaxation. This leads to clitoral engorgement, a rise in clitoral cavernosal artery inflow, and an increase in clitoral intracavernosal pressure. The result is extrusion of the glans clitoris and enhanced sensitivity.

In the basal state the vaginal epithelium reabsorbs sodium from the submucosal capillary plasma transudate. Following sexual stimulation, neurotransmitters including NO and vasoactive intestinal peptide modulate vaginal and vaginal arteriolar smooth muscle relaxation. A dramatic increase in capillary inflow in the submucosa overwhelms sodium reabsorption leading to 3 to 5 ml of vaginal transudate, enhancing lubrication essential for pleasurable coitus. Vaginal smooth muscle relaxation results in increased vaginal length and luminal diameter, especially in the distal two-thirds. Vasoactive intestinal polypeptide is a non-adrenergic non-cholinergic neurotransmitter that plays a key role in enhancing vaginal blood flow, lubrication, and secretions.

Hormones

Steroid hormones play an important role in the regulation of female sexual function. Symptoms associated with diminished oestrogen include vaginal dryness, irritation, and pain, as well as causing complaints regarding sexual function such as decreased sexual arousal, decreased genital sensation, and difficulty achieving orgasm. Oestrogen replacement in postmenopausal or surgically menopausal women restores clitoral and vaginal pressure thresholds to premenopausal levels. Oestrogens have also vasoprotective and vasodilatory effects, which result in increased vaginal and clitoral blood flow, preventing atherosclerotic compromise of the iliohypogastric arterial bed. Thickness and rugae of the vaginal wall, as well as vaginal lubrication, are both oestrogen dependent. Oestrogen also improves integrity of vaginal mucosal tissue and has beneficial effects on vaginal sensation, vasocongestion, and secretions. By contrast, oestrogen deprivation leads to decreased pelvic blood flow resulting in clitoral fibrosis, thinned vaginal epithelial layers, and decreased vaginal submucosal vasculature.

In addition to conventional hormone replacement therapy and topical oestrogen creams, exogenous water- and oil-based lubricants have been used for alleviating symptoms of vaginal irritation and dryness. Although some of these agents have properties similar to vaginal fluid, they are limited by application time and duration of effectiveness. Oestrogen replacement therapy and oestrogen creams are effective for relieving symptoms of vaginal dryness in many women, but they are not a viable alternative for all women.

Low testosterone levels are often associated with a decline in sexual arousal, genital sensation, libido, and orgasm. Testosterone administration in women with decreased desire has been successful, although there are no currently approved testosterone preparations for this purpose. It has also been shown that menopausal women respond better to parental oestrogen–androgen than oestrogen alone in restoring sexual desire, energy, and sense of well being.

Specific sexual dysfunctions in women

Hypoactive sexual desire disorder

Hypoactive sexual desire disorder is defined as chronic or persistent deficiency of sexual interest or desire. It is distinguished from sexual aversion disorder, which is characterized by persistent or recurrent phobic aversion to, and avoidance of, sexual contact with a partner. Hypoactive sexual desire disorder is the most commonly reported sexual problem in women, affecting approximately 20 to 30 per cent of those aged 30 to 70 years. It is age-related and strongly associated with other medical and psychiatric disorders, particularly chronic illnesses and depression. It may be related to the use of prescription drugs, particularly antidepressants (e.g. SSRIs), alcohol abuse, or hormonal changes. Low sexual desire in women is also frequently associated with partner conflicts or the presence of other sexual dysfunctions (such as erectile dysfunction, anorgasmia). The degree of personal distress associated with the disorder varies widely and should be taken into account in making the diagnosis.

Female sexual arousal disorder

Female sexual arousal disorder refers to a chronic or persistent lack of subjective or physical arousal during sexual stimulation. This is characterized by an inability to achieve an adequate lubrication–swelling response of the vagina and labia for the completion of sexual activity, or a lack of subjective arousal during sexual activity. Although women with sexual arousal disorder may perform intercourse, the lack of adequate lubrication may result in pain or vaginal irritation. Findings from the National Health and Social Life Survey in the United States showed that approximately 20 per cent of women aged 18 to 59 reported difficulty in lubrication during sexual stimulation. Similar prevalence estimates were reported in a large-scale British study, which also found that sexual arousal difficulties became commoner with agieng, and that marital difficulties, anxiety, and depression were all significant risk factors.

Orgasmic disorder

Orgasmic disorder in women is a persistent or recurrent difficulty, delay in, or absence of orgasmic attainment despite adequate sexual stimulation and arousal. Primary anorgasmia refers to the woman who has never experienced orgasm through any means, whereas secondary orgasmic dysfunction occurs when a woman is unable to have orgasm with a particular partner or by means of specific stimulation (i.e. sexual intercourse). Both types of orgasmic dysfunction are common, affecting 10 to 20 per cent of women in population-based studies and associated with medical or psychiatric illnesses, use of prescription or non-prescription drugs, and specific neurological disorders (such as spinal cord injury). Relationship conflicts were also found to be significantly associated with the occurrence of orgasmic difficulties in women in a recent British study.

Dyspareunia

Dyspareunia, or pain associated with sexual intercourse, is common in women. The pain may occur before, during, or after intercourse. In most cases it is caused by a lack of lubrication or other physical cause. According to the National Health and Social Life Survey, about 15 per cent of women have experienced pain during sexual activity during the past year. In postmenopausal women, the prevalence of dyspareunia may be even higher. A wide variety of medical or organic conditions are associated, including hymenal scarring, pelvic inflammatory disease, and vulvar vestibulitis. However, dyspareunia is not reliably associated with any particular medical disorder. Moreover, anatomical or physiological factors that may have caused the original pain may not be the same factors responsible for maintaining it. Accordingly, some authors have recommended an interactive or multidimensional model of physical and psychological determinants of dyspareunia.

Vaginismus

Vaginismus, or involuntary spasms of the musculature of the outer third of the vagina, is a significant cause of penetration difficulties in women. The disorder is often seen in sex therapy clinics, occurring in approximately 15 to 17 per cent of women presenting for treatment who often complain of secondary dyspareunia in addition to other sexual dysfunctions. Women with primary or generalized vaginismus typically avoid gynaecological examinations and tampon use, in addition to sexual intercourse. Vaginismus can occur in association with vaginal pain due to various medical conditions, although it is more frequently related to psychological or interpersonal factors.

Management of female sexual dysfunction

Diagnostic assessment

There is no consensus on the routine diagnostic assessment of the woman with sexual dysfunction. Common features of the work-up typically include history (sexual, psychosocial, and medical), physical examination (external genitalia, internal genitalia), and laboratory testing (oestrogen, testosterone, glucose, full blood count, creatinine, liver function tests, cholesterol, urine analysis, vaginal cultures as appropriate). Specialized tests, such as vaginal photoplethysmography, have been reported in the literature, but are not used in routine clinical practice. Other specialized investigations may include: (i) neurological (somatosensory evoked potential, electromyography), (ii) hormonal, and (iii) psychological (depression, anxiety).

Treatment

There are limited data on safety and efficacy of the various primary psychological and hormonal treatments that have been used for female sexual dysfunction.

Psychological therapy has demonstrated efficacy in sexual desire, orgasmic disorders, vaginismus, and sexual pain disorders. Oestrogen and/or androgen replacement hormonal therapy may be useful in sexual desire, arousal, orgasmic, and dyspareunia sexual pain disorders. Oestrogen replacement is indicated in menopausal women to relieve hot flushes, prevent osteoporosis, lower the risk of cardiovascular disease, improve clitoral sensitivity, and decrease pain and burning during intercourse. In combination with oestrogen, methyl testosterone is used to treat symptoms of inhibited desire, dyspareunia, and lack of vaginal lubrication, as well as for its vasoprotective effects.

Other treatment options with scant efficacy and safety evaluations include vaginal lubricants, vasodilator agents such as sildenafil citrate, vaginal dilators, pelvic floor rehabilitation using biofeedback electromyography, and Kegel exercises.

Patient and partner education is an essential component in the management of female sexual dysfunction and should be a continuous element at each phase in the process of care. Modification of known risk factors (hypertension, hyperlipidaemia, prescription drugs, cigarette smoking, and alcohol abuse) and self-destructive

behaviours is part of good clinical practice.

Summary

Female sexual dysfunction is age-related and common, affecting 30 to 50 per cent of women. Improved understanding of both normal and dysfunctional sexual response in women is the keystone to enhanced management. A collaborative and comprehensive evaluation, patient and partner education, modification of reversible causes, and an individualized treatment plan should be the standard management of women with sexual dysfunction.

Basic science and clinical research in female sexual function and dysfunction are needed to overcome the gender gap, achieve a more balanced therapeutic perspective between psychological and physiological factors, and offer female patients enhanced opportunity for relief of a persistent or recurrent sexual condition that causes significant personal distress.

Further reading

Bastuba MD *et al.* (1994). Arterial priapism: Diagnosis, treatment and long term follow up. *Journal of Urology* **151**, 1231–7.

Cooper AJ, Cernovsky ZZ, Colussi K (1993). Clinical and psychometric characteristics of primary and secondary premature ejaculators. *Journal of Sex and Marital Therapy* **19**, 276–88.

DeBusk R *et al*. (2000). Management of sexual dysfunction in patients with cardiovascular disease: Recommendations of the Princeton Panel. *American Journal of Cardiology* **86**, 175–81.

Dunn KM, Croft PR, Hackett GI (1998). Sexual problems: a study of the prevalence and need for health care in the general population. *Family Practice* **15**, 519–24.

Feldman HA *et al.* (1994). Impotence and its medical and psychosocial correlates: Results of the Massachusetts Male Aging Study. *Journal of Urology* **151**, 54–61.

Goldstein I *et al.* (1998). Oral sildenafil in the treatment of erectile dysfunction. *New England Journal of Medicine* **338**, 1397–404.

Hakim LS *et al.* (1996). Evolving concepts in the diagnosis and management of high flow priapism. *Journal of Urology* **155**, 541–8.

Jardin A *et al.* eds (2000). *Erectile dysfunction: 1st International Consultation on Erectile Dysfunction. Plymouth, U.K*. Health Publications, United Kingdom.

Kulmala RV, Lehtonen TA, Tammela TLJ (1996). Preservation of potency after treatment of priapism. *Scandinavian Journal of Urology and Nephrology* **30**, 313–16.

Lauman EO, Paik A, Rosen RC (1999). Sexual dysfunction in the United States: Prevalence and predictors. *Journal of the American Medical Association* **281**, 537–44.

Lue TF (2000). Erectile dysfunction. *New England Journal of Medicine* **342**, 1802–13.

Masters WH, Johnson VE (1970). *Human sexual inadequacy*. Little, Brown, Boston.

Rosen RC, Lane RM, Menza M (1999). Effects of SSRIs on sexual function: A critical review. *Journal of Clinical Psychopharmacology* **19**, 67–85.

Spector IP, Carey MP (1990). Incidence and prevalence of the sexual dysfunctions: A critical review. *Archives of Sexual Behavior* **19**, 389–409.

Spycher MA, Hauri D (1986). The ultrastructure of the erectile tissue in priapism. *Journal of Urology* **135**, 142–7.

12.9.1 Normal and abnormal sexual differentiation

M. O. Savage

Introduction

Disorders of sexual differentiation are characterized by an abnormality in the formation of the internal or external genital structures. Most are genetically determined and are associated with an ambiguous appearance of the external genitalia. During the past three decades, the study of intersex disorders has changed in orientation. The emphasis has moved away from descriptive clinical syndromes towards the biochemical and molecular nature of the defects that cause them. If such an aetiological approach is to be used, a fundamental understanding of normal sexual differentiation is required. This provides the basis for the classification, investigation, and management of patients with abnormal sexual differentiation.

Physiology of fetal sexual differentiation

In the male, it is established that the castrated embryo develops as a female, indicating that the fetal testis is essential for male development. The chromosomal sex of the embryo, established at conception, directs the development of either ovaries or testes. In the male, specific genes on the short arm of the Y chromosome, known as the sex-determining region of the Y chromosome, code for testis determination, and hence contribute to testicular differentiation. Testicular Leydig cells synthesize and secrete testosterone from 8 weeks of gestation, aided by stimulation with placental human chorionic gonadotrophin (hCG). Testosterone diffuses locally to maintain and virilize the wolffian ducts which become the vas deferens, seminal vesicles, and epididymis. Antimüllerian hormone, or müllerian-inhibitory factor, is a glycoprotein which is secreted during the same time period by testicular Sertoli cells to inhibit the formation of the uterus, fallopian tubes, and upper vagina from the müllerian structures.

In androgen-dependent tissues, testosterone is converted to dihydrotestosterone which virilizes the external genitalia. Peripheral androgen action depends on the binding of the androgen in the target tissues to a receptor coded by an X chromosome.

In the female, ovarian development occurs in the presence of two X chromosomes and external gonadal development occurs spontaneously. Genital development in both sexes is completed by 20 weeks of fetal life. In the male, growth of the formed penis is dependent upon continued testicular testosterone secretion under stimulation by pituitary gonadotrophins.

Classification of intersex states

The classification that forms the basis of clinical assessment and management depends on gonadal morphology (Table 1). Female pseudohermaphroditism describes genital ambiguity resulting from abnormal virilization of a female with normal ovaries. The male counterpart—male pseudohermaphroditism—is the result of incomplete virilization of a male with differentiated testes. Thirdly, the true hermaphrodite possesses both ovarian and testicular tissue.

Female pseudohermaphroditism

Female pseudohermaphrodites have 46 XX karyotypes with normal ovaries and müllerian structures, but the external genitalia are virilized. The aetiology of female pseudohermaphroditism is given in Table 2. The degree of genital ambiguity can range from enlargement of the clitoris or fusion of the posterior labia to a completely male appearance, depending on the timing of androgen production and the concentration of androgens in the fetal circulation. Virilization may be caused by excessive production of either fetal or maternal androgens.

Virilization by fetal androgens

Congenital adrenal hyperplasia (see also Chapter 12.7.2)

The commonest cause of ambiguous genitalia in the newborn female is a recessively inherited enzyme defect of cortisol synthesis, with diversion of intermediates to androgen production. A reduction in steroid 21-hydroxylase or absence of 11b-hydroxylase or 3b-hydroxysteroid dehydrogenase can be the cause of this condition. These enzymes are part of the steroid biosynthetic pathways which link cholesterol with cortisol, aldosterone, and androgens. In the absence of, or lowered potential for, cortisol production, there are high adrenocorticotrophic hormone (ACTH) levels leading to adrenal hyperplasia and excess androgen production.

21-hydroxylase deficiency

This form of congenital adrenal hyperplasia accounts for 90 per cent of cases of female pseudohermaphroditism, and should be excluded before proceeding to assign other causes for ambiguous genitalia. The degree of virilization can be variable (Fig. 1). In Europe, 60 per cent of all cases will develop salt depletion due to decreased production of aldosterone in the first 2 weeks of life. Usually, there is enlargement of the clitoris associated with a degree of posterior labial fusion and the formation of a hypoplastic lower vagina which may open into the urethra.

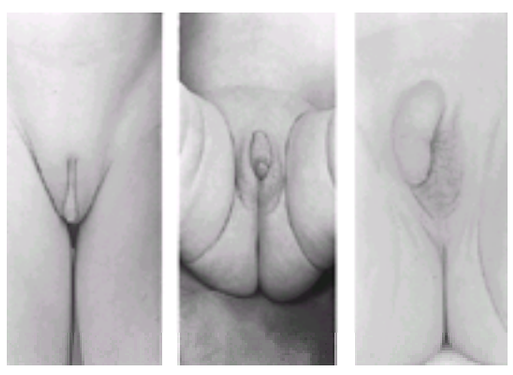

Fig. 1 Variation in degree of virilization in three female infants with 21-hydroxylase deficiency.

17-Hydroxyprogesterone is a biosynthetic precursor of cortisol, and plasma levels are elevated in 21-hydroxylase deficiency. After the third day of life, there is good discrimination between plasma levels of 17-hydroxyprogesterone in affected cases (100–800 nmol/l) and those of normal infants (less than 15 nmol/l). Measurement of plasma renin activity and aldosterone help to define the extent of mineralocorticoid deficiency. Analysis of urine steroid excretion using chromatography and mass spectrometry can provide a reliable diagnostic profile from day 3 after birth. A variant of 21-hydroxylase deficiency, due to a milder defect, is the non-classical form of the disease, characterized by absence of neonatal genital ambiguity but development of other signs of androgen excess such as hirsutism.

11b-hydroxylase deficiency

This defect probably accounts for about 5 per cent of all cases of congenital adrenal hyperplasia. The disorder is caused by mutations of the *CYP11B1* gene which abolish enzyme activity. The virilization may be severe, with affected females sometimes raised as males. The plasma concentration of 11-deoxycortisol (compound S) is elevated, and may exceed 1000 nmol/l. The urine steroid pattern will show high excretion of 6-hydroxytetrahydro-11-deoxycortisol as well as of tetrahydro-S.

3b-hydroxysteroid dehydrogenase deficiency

This rare defect was originally described in male infants with incomplete virilization. However, a paradoxical androgen affect may be seen in female infants due to a very high level of dehydroepiandrosterone, the steroid precursor immediately proximal to the enzyme block. A non-classical or attenuated form of 3b-hydroxysteroid dehydrogenase deficiency has been described which presents with virilization in postadrenarchal or peripubertal females.

Virilization by maternal androgens

Virilization of the external genitalia by a maternal ovarian or adrenal androgen-secreting tumour is a rare but well-recognized cause of female pseudohermaphroditism. The degree of virilization may be striking.

Other causes of fetal virilization

Female pseudohermaphroditism due to maternal administration of progestogen preparations became recognized about 30 years ago. A number of dysmorphic childhood syndromes may also be associated with virilized female genitalia.

Male pseudohermaphroditism

Male pseudohermaphroditism arises as a result of a disturbance of male genital development in patients with testes and a 46 XY karyotype. The genital anomaly can vary from apparently female to male external genitalia with a small penis or perineal hypospadias. The three main aetiological groups (Table 3) are impaired Leydig cell activity, peripheral androgen insensitivity, and deficient testosterone and antimüllerian hormone production by incompletely differentiated testes.

Impaired testicular secretion of testosterone

Inborn errors of testosterone biosynthesis

These rare disorders (Fig. 2) lead to defective testosterone synthesis during the critical period of fetal sexual differentiation. The result is inadequate testosterone secretion, either locally to virilize the wolffian ducts to form the vas deferens, seminal vesicles, or epididimis, or peripherally to virilize the external genitalia. Synthesis of antimüllerian hormone, being a glycoprotein rather than a steroid, is unaffected. When the enzyme deficiency, inherited as an autosomal recessive trait, is situated early in the biosynthetic pathway, adrenal steroid synthesis may also be affected.

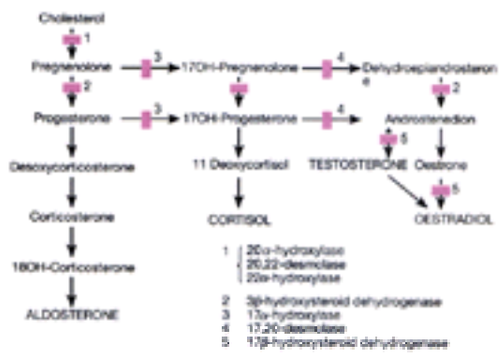

Fig. 2 Enzyme defects in testosterone biosynthesis.

Deficient formation of pregnenolone

Three closely related microsomal enzymes (20a-hydroxylase, 20,22-desmolase, and 22a-hydroxylase) are necessary for the conversion of cholesterol to pregnenolone. Deficiency of one of these enzymes leads to impaired synthesis of cortisol, aldosterone, and testosterone. Accumulation of cholesterol has been demonstrated in the hyperplastic adrenal leading to the term congenital lipoid adrenal hyperplasia. Recently, the gene responsible for the disorder has been cloned and validated by the demonstration of a non-sense mutation. The gene encodes for a protein named StAR (steroidogenic acute regulatory protein).

Deficiency of 3b-hydroxysteroid dehydrogenase

Male patients with this enzyme deficiency are poorly virilized and usually develop salt loss and adrenal failure in infancy. Some subjects have survived puberty, which has been characterized by virilization and gynaecomastia. Urinary pregnenetriol and plasma dehydroepiandrosterone are elevated, as is plasma renin activity. Several mutations of the type II 3b-hydroxysteroid dehydrogenase enzyme have been reported.

Deficiency of 17a-hydroxylase

This defect in cytochrome P450c17 results in decreased cortisol synthesis by the adrenal cortex and testosterone by the fetal testes, resulting usually in a omplete lack of virilization. Several different mutations of the *CYP17* gene have been reported. The disorder is identified biochemically by demonstrating high serum and urinary levels of progesterone and corticosterone. A compensatory increase in ACTH leads to excess mineralocorticoid secretion, causing hypertension, hypokalaemia, and low plasma renin activity.

Deficiency of 17,20-desmolase

This rare defect is related to 17-hydroxylase deficiency as both enzyme activities are coded by the same gene. Impaired virilization may be variable in degree. Biochemically, identification of the defect relies on elevation of plasma 17-hydroxyprogesterone, 17-hydroxypregnenolone, and urinary pregnanetriolone.

Deficiency of 17-ketosteroid reductase (17b-hydroxysteroid dehydrogenase)

Patients with this defect are born with female-looking external genitalia and a phallus closely resembling a normal clitoris. There are three types of 17b-hydroxysteroid dehydrogenase enzyme. Abnormalities of the type 3 enzyme, which has specific testicular expression, causes this disorder and several different mutations of the

respective gene have been described. There is a high prevalence of this disorder within the Arab population of the Gaza Strip. The enzyme defect interferes with conversion of androstenedione to testosterone, adrenal steroid synthesis being unaffected. There is elevation of the androstenedione to testosterone ratio, particularly after hCG stimulation. Subjects are usually raised as girls; however, gender conversion to male has been described, coinciding with the marked virilization which occurs at puberty.

Androgen insensitivity syndromes

Mechanisms of androgen action

Testosterone, the principal androgen secreted by the testis, circulates bound to two proteins, sex hormone binding globulin and albumin. The protein-bound steroid is in dynamic equilibrium with the free hormone. Free testosterone enters the target cell by a passive mechanism (Fig. 3). Inside the cell, testosterone can be reduced to dihydrotestosterone by the enzyme 5a-reductase. Testosterone or dihydrotestosterone binds with high affinity to specific receptor proteins to form an androgen–receptor complex. This complex enters the cell nucleus and, after transformation into a DNA binding state, binds to specific nucleotide sequences which promote transcription of messenger RNA, resulting in clinical virilization. The gene encoding the human androgen receptor has recently been cloned.

Fig. 3 Scheme of intracellular androgen action (adapted from Hughes and Pinsky, 1989). T, testosterone; DHT, dihydrotestosterone; R, receptor.

Although testosterone and dihydrotestosterone bind to the same receptor, the two hormones perform different roles in androgen physiology (Fig. 4). Testosterone regulates secretion of luteinizing hormone, virilizes the wolffian ducts during fetal life, and may be essential for spermatogenesis. Dihydrotestosterone is responsible for formation of the external genitalia and prostate, and for most secondary sexual effects, such as hair growth and enlargement of the genitalia.

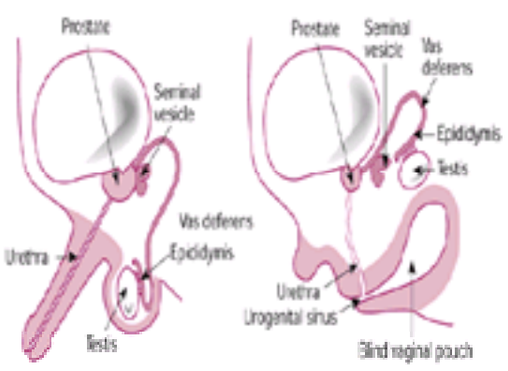

Fig. 4 Roles of testosterone and dihydrotestosterone in male sexual differentiation in the normal male (left) and the patient with 5a-reductase deficiency (right). Testosterone, heavy shading; dihydrotestosterone, light shading.

Clinical features of androgen insensitivity

Abnormalities of androgen action may have a severe effect on male sexual differentiation, resulting in incomplete virilization during fetal life and at puberty. The syndromes of androgen insensitivity probably account for the majority of cases of male pseudohermaphroditism. There are three main forms. The most important numerically is an X-linked cystolic androgen receptor defect. This may be manifested by a broad clinical spectrum from complete androgen insensitivity to a virtually normally formed male with infertility. The second form is known as receptor-positive resistance, where a similar spectrum of clinical defects is associated with apparently normal receptor function. Thirdly, 5a-reductase deficiency is an autosomally inherited enzyme defect resulting in impaired conversion of testosterone to dihydrotestosterone in the target cell.

Androgen receptor defects

These defects may be expressed clinically as a spectrum of deficient formation of the male internal and external genitalia.

Complete androgen insensitivity

The typical patient with presents after puberty with primary amenorrhoea, or before puberty with inguinal hernias and palpable testes. The phenotype and psychosexual orientation is female. Breasts develop as in a normal woman, but pubic and axillary hair is scanty and the vagina is blind-ending due to regression of müllerian structures, which are virtually always absent. Wolffian structures are usually absent and the gonads show Leydig cell hyperplasia with no spermatogenesis. There is a significant risk of gonadal malignancy occurring after puberty, when gonadectomy is recommended.

Incomplete androgen insensitivity

Here there is a range of impaired virilization from clitoral enlargement and labial fusion to small external genitalia, usually with hypospadias. At puberty, feminization may be dominant, with gynaecomastia a feature in some patients.

The endocrine features of androgen insensitivity are essentially similar in the range of clinical defects. Testicular androgen secretion is normal or increased. Plasma testosterone may be elevated in infancy but is normal during the remainder of the prepubertal period. At puberty, plasma testosterone, oestradiol, and sex hormone binding globulin levels are elevated. Plasma gonadotrophins are normal in childhood but both luteinizing hormone and follicle stimulating hormone levels are consistently elevated during and after puberty, due to insensitivity of the hypothalamic androgen receptor. Patients are usually infertile.

5a-Reductase deficiency

This autosomal recessive disorder is characterized by impaired conversion of testosterone to dihydrotestosterone in androgen-dependent target cells. It was first described in the Dominican Republic, and occurs principally in areas of high consanguinity. The clinical features can be summarized as showing male internal genital structures and female external genitalia (Fig. 4). Fetal dihydrotestosterone-dependent development is abnormal, resulting in a rudimentary phallus and absent prostate. The wolffian structures develop normally, and the testes differentiate with spermatogenesis capable of progressing to the spermatozoa stage. Most subjects

are raised as females but gender conversion to male occurs at puberty, coinciding with striking virilization of body habitus (Fig. 5) and male psychosexual orientation. Virilization during and after puberty, however, is incomplete, as the penis remains small and body and facial hair is sparse.

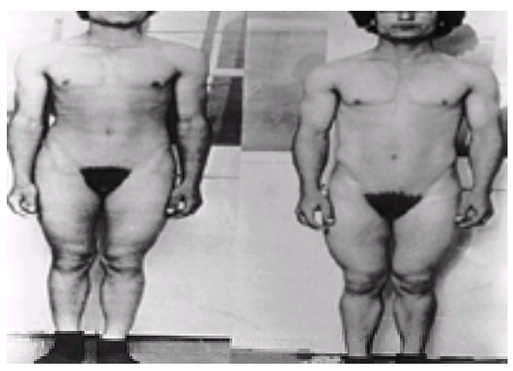

Fig. 5 Two Greek Cypriot brothers with 5a-reductase deficiency.

The endocrine features comprise low plasma dihydrotestosterone with normal testosterone and an elevated testosterone to dihydrotestosterone ratio. This abnormal ratio is the cardinal diagnostic feature. There is also elevation of the ratio of 5b : 5a androgen metabolites (i.e. aetiocholanolone to androsterone) in urine.

Male pseudohermaphroditism related to abnormal testicular differentiation

Incomplete differentiation of the fetal testes due to a defect of the Y-chromosomal genes responsible for testicular determination may cause genital ambiguity. Incompletely formed or dysgenetic testes secrete insufficient testosterone and antimüllerian hormone for normal male development. A number of clinical syndromes exist in this category.

Dysgenetic male pseudohermaphroditism

In this syndrome there are bilateral dysgenetic testes, persistent müllerian structures, cryptorchidism, and poorly virilized external genitalia.

Mixed gonadal dysgenesis

Here there is asymmetrical gonadal differentiation with a testis present on one side and a streak gonad on the other. The internal structures are also asymmetrical, reflecting the endocrine function of the ipsilateral gonad. Many patients have a mosaic XO/XY karyotype and features of Turner's syndrome.

Drash syndrome

This syndrome combines dysgenetic testes, genital ambiguity, glomerulonephritis, and Wilms' tumour.

True hermaphroditism

The diagnosis of true hermaphroditism is made when ovarian as well as testicular tissue is present in the same individual. Van Niekerk has published an extensive review of the literature, including a large personal series. The most common presenting symptoms are abnormal appearance of the external genitalia. Most patients have a 46 XX karyotype: about half are pure 46 XX and about a third are mosaics or chimeras with 46 XX cell lines, that is 46 XX/46 XY. A few patients with a pure 46 XY karyotype have been reported. Occasional familial cases of true hermaphroditism have been described in the literature.

Other 46 XX intersex states

Pure gonadal dysgenesis

This disorder, which may be familial, is usually associated with female external genitalia. Clitoromegaly is sometimes present. The gonads are streaks and the karyotype may be 46 XX or XY. 46 XY gonadal dysgenesis, inherited as an X-linked recessive or male-limited autosomal dominant condition, has also been described.

XX male

A number of XX males have been described. These are normal-appearing males with normal intelligence and male psychosexual orientation. Gynaecomastia, sparse facial hair, small genitalia, and hypospadias may occur in this syndrome. The testes are small and resemble Klinefelter testes histologically. There is absence of spermatogenesis, leading to sterility. Families have been reported containing both an XX male and a 46 XX true hermaphrodite.

Gonadal neoplasia and intersex states

It is now established that a number of intersex disorders carry an increased risk of gonadal tumours. Two important risk factors are the presence of dysgenetic gonadal tissue and a Y chromosome. Intra-abdominal gonads are more susceptible than scrotal glands. The commonest tumour is a gonadoblastoma which is a premalignant lesion but can progress to an invasive tumour.

Clinical and laboratory assessment of patients with intersex states

The assessment of patients with intersex states may be considered from the point of view of the paediatrician assessing an infant with ambiguous genitalia. The same principles apply to the older child or adult. It must be emphasized that the general appearance of the external genitalia, while important in deciding the appropriate gender for the child, is of very little help in defining the aetiology of the disorder.

Clinical assessment

The principles of clinical assessment are shown in Table 4. A history of a similar disorder in other family members may shed light on the likely diagnosis. Many of these conditions are genetically determined. Examination for other anomalies which could point to a dysmorphic syndrome known to be associated with abnormal genital development is also relevant. The most important aspect of the examination, however, is careful palpation of the gonads.

If no gonads are palpable, the most likely diagnosis is female pseudohermaphroditism due to congenital adrenal hyperplasia, and this is virtually certain if symptoms of salt loss develop. Other possible disorders are true hermaphroditism or male pseudohermaphroditism with intra-abdominal gonads. When both gonads are palpable in the scrotum or labial folds, the patient is likely to be a male pseudohermaphrodite, and measurement of plasma androgens will indicate whether the aetiology is a testicular or peripheral defect. A true hermaphrodite with bilateral ovotestes may also present in this way. The presence of only one palpable gonad or asymmetry of the perineum is suggestive of mixed gonadal dysgenesis; true hermaphroditism with asymmetrical gonads is the other differential diagnosis.

Laboratory assessment

A similar scheme may be devised as a guide to confirming the aetiology biochemically (Table 5). In all intersex patients a karyotype is indicated. If no gonads are palpable, determination of plasma 17-hydroxyprogesterone will confirm or exclude 21-hydroxylase deficiency. In 11b-hydroxylase deficiency the plasma 11-deoxycortisol concentration is elevated. The infant with two palpable gonads needs an hCG stimulation test to assess testicular androgen secretion. Numerous hCG regimens exist, of which two examples are 1000 IU daily for 3 days or a single injection of 1500 IU/m ^{2} body surface area. Basal and poststimulatory concentrations of testosterone, dihydrotestosterone, and androstenedione should distinguish a disorder of testosterone biosynthesis from a syndrome of androgen insensitivity.

If one gonad is palpable, gonadal biopsy may be helpful, particularly if ovarian tissue is suspected. Pelvic ultrasonography or exploratory laparotomy for identification of internal genital structures may also be indicated. In any patient with incomplete virilization, urethrography should be performed to identify a vaginal cavity communicating posteriorly with the urethra.

Medical management

Choice of gender

Parents are usually shocked to learn that there is doubt as to the sex of their child; they are often under the impression that the child may grow up to be neither male nor female. Temptation to give a provisional opinion should be avoided until the nature of the disorder is known and an informed answer can be given. The decision as to the appropriate sex-of-rearing is based mainly on the appearance of the external genitalia and on the likely pattern of secondary sexual development at puberty. This decision should be taken jointly by the endocrinologist, urologist, and the parents. The gender should be assigned as soon as possible; however, in some cases of severe ambiguity, there is a case for waiting to assess the effect of early treatment with depot testosterone (25–50 mg at monthly intervals) on phallic growth as a guide to androgen responsiveness.

The concept that, once established, gender identity and role are more or less fixed has now been questioned. Although change of gender may be extremely difficult, the possibility of gender conversion should be viewed with an open mind in the individual subject who, because of spontaneous virilization or feminization at puberty, finds existence in their original gender intolerable.

Sex hormone therapy

Long-term treatment with androgens to promote phallic growth in early childhood has rightly fallen into disrepute because of the acceleration of bone maturation, which leads to loss of ultimate growth potential. While standard testosterone treatment is effective for inducing pubertal development in males with androgen-responsive syndromes, it is of limited value in patients with androgen insensitivity. Induction of full masculization in these patients is still very unsatisfactory. It has, however, been demonstrated that some further virilization in adult patients may be effectively induced using supraphysiological doses of depot testosterone (500 mg weekly). Effects, albeit slow to appear, were seen specifically in penile length and facial and body hair growth.

Further reading

Eckstein B, Cohen S, Farkas A, Rosler A (1989). The nature of the defect in familial male pseudohermaphroditism in Arabs of Gaza. *Journal of Clinical Endocrinology and Metabolism* **68,** 477–85.

Forest MG (1981). Inborn errors of testosterone biosynthesis. *Pediatric and Adolescent Endocrinology* **8,** 133–55.

Hughes IA, Pinsky L (1989). Sexual differentiation. In: Collu R, Ducharne JR, Guyda HS, eds. *Paediatric endocrinology*, pp. 251–93. Raven Press, New York.

Imperato-McGinley J *et al.* (1982). Hormonal evaluation of a large kindred with complete androgen insensitivity: evidence for secondary 5-alpha-reductase deficiency. *Journal of Clinical Endocrinology and Metabolism* **54,** 931–41.

Kirk JMW, Perry LA, Shand WS, Kirby RS, Besser GM, Savage MO (1990). Female pseudohermaphroditism due to a maternal adreno-cortical tumour. *Journal of Clinical Endocrinology and Metabolism* **70,** 1280–4.

Pang S, Lerner AJ, Stoner LS (1985). Late onset adrenal steroid 3b-hydroxysteroid dehydrogenase deficiency: A cause of hirsutism in pubertal and post-pubertal women. *Journal of Clinical Endocrinology and Metabolism* **60,** 428–35.

Price P *et al.* (1984). High dose androgen therapy in male pseudohermaphroditism due to 5-alpha reductase deficiency and disorders of the androgen receptor. *Journal of Clinical Investigation* **74,** 1496–508.

Savage MO, Lowe DG (1990). Gonadal neoplasia and abnormal sexual differentiation. *Clinical Endocrinology* **32,** 519–33.

Savage MO, Sultan C (1999). Intersex. In: Mundy AR, Fitzpatrick JM, Neal DE, George NJR, eds. *Scientific basis of urology*, pp. 421–37. Isis Medical Media, Oxford.

van Niekerk WA (1981). True hermaphroditism. *Pediatric and Adolescent Endocrinology* **8,** 80–99.

Williams DM, Patterson MN, Hughes IA (1993). Androgen insensitivity syndrome. *Archives of Disease in Childhood* **68,** 343–4.

Wilson JD, Griffin JE, Leshin M, MacDonald PC (1983). The androgen resistance syndromes. In: Stanbury JB, Wyngaarden JB, Fredridison DS, Goldstein JL, Brown MS, eds. *The metabolic basis of inherited disease*, pp. 1001–26. McGraw-Hill, New York.

12.9.2Normal growth and its disorders

M. A. Preece

The normal curve of growth

The upper panel of Fig. 1 shows the curve of height attained (or distance) for a typical male from birth to 18 years of age. It contains all the relevant information about growth in height of an individual child. In the lower panel the growth data have been converted to height velocity in cm/year. This is calculated from the distance data by dividing the difference between two height measurements (as close to 1 year apart as possible) by the exact time elapsed between them. The calculated velocity is plotted at the midpoint of the time interval over which it is measured. This representation of growth is particularly useful as it emphasizes the dynamic nature of the growth process.

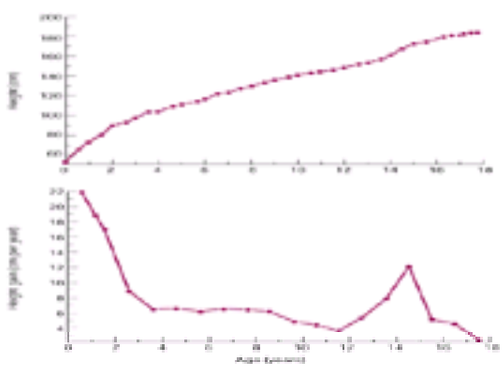

Fig. 1 The growth in height of an individual boy from birth to 18 years of age. The top panel shows the height attained or distance curve and the lower panel shows the data replotted as height gain or growth velocity. (Reproduced from Tanner JM (1962). *Growth at adolescence*, Blackwell, Oxford, with permission.)

The growth velocity in any one year is a more sensitive measure of events occurring in that year than is the coincident height distance datum, which is a measurement summating all previous growth. Thus, the more sensitive measures of the velocity curve may show rather dramatic change in growth during disease or during treatment, where the simpler distance curve would be less sensitive.

There are three epochs of growth: early, rather fast growth before the age of 2 years; relatively slow steady growth during the preschool and primary school years; and then the period around puberty when the adolescent growth spurt dominates the growth pattern.

Epochs of growth

Infancy

During the first year of life the infant has an average height velocity of about 25 cm/year. However, since this is a time when the velocity is changing dramatically, measurement over shorter time periods should be considered. The velocity during the first 3 months is equivalent to 3.3 cm/month in boys and 3.0 cm/month in girls, dropping to 1.2 cm/month and 1.3 cm/month respectively by the last 3 months of that year. During the next 3 years there is a further deceleration to a velocity of 0.5 cm/month, or 6 cm/year, which is the average through much of middle childhood. This continues with gentle slowing until puberty.

The first 3 years are also the time of increased channelling of the growth curve. At birth, the length of the baby is determined mostly by the fetal environment which, in turn, is much dependent on maternal size; the father's height is poorly correlated with the child's. During the next 2 to 3 years, the influence of the father's genetic make-up increases progressively until there is equal influence from both parents. This phenomenon can result in some rather bizarre growth patterns where mother and father have very different heights. For example, when the child is born to a tall mother but short father, the initially rather large baby will tend to grow unusually slowly until the genetically expected channel is achieved.

Middle childhood

Although in healthy children growth is a moderately steady process, there are fluctuations that occur in the short and middle term, the most striking of which are related to seasonal changes. Most children grow faster in spring and summer, with relatively slow periods in the autumn and winter; a few children show a reversed pattern, and others no regular pattern at all. There is a rather constant, but small, growth spurt at 6 to 7 years of age in both boys and girls, with no sexual dimorphism; it has been attributed to the onset of adrenal androgen secretion that occurs at about that age.

Puberty

There is little difference in growth rates between males and females before 10 years of age. Then the typical female starts her adolescent growth spurt and, for a few years, is taller than a male of the same age. About 2 years later the male starts his spurt and by the age of 14 is the taller. The major difference in adult height between males and females (14 cm in the United Kingdom) is established at puberty. About 11 cm comes from the extra 2 years of prepubertal growth of boys and 3 cm from their more intense growth spurt.

Sexual dimorphism in height during puberty is reflected in many other body dimensions. At this time there is a more dramatic growth spurt in males for most body dimensions, occurring about 2 years later than in females. A major exception is the more dramatic and sustained growth spurt of the female pelvis. There are also associated changes in soft tissues, leading to the greater muscularity of the male.

Metabolic and endocrine factors controlling growth

Each of the epochs of growth described above is associated with particular metabolic or hormonal factors which exert their effects in an almost sequential manner. These distinctions should not be overstated, however, and are only useful as a general guide to the most important factors at a given age. In essence, infant growth is dominated by nutritional considerations, the childhood period by growth hormone, and puberty by the sex hormones, but nutrition has an important role throughout and growth hormone is important in early infant growth, as shown by the reduced size at birth of children with congenital growth hormone deficiency. Thyroid hormone is also important throughout the growing period. Throughout growth the actions of growth hormone and possibly other hormonal and nutritional factors are in part

mediated by the insulin-like growth factors and their binding proteins.

Disorders of growth

Growth disorders are predominantly problems of childhood, and patients seen by the adult physician are inevitably left with a legacy of events which have occurred before maturity. In many cases where the condition was successfully treated during childhood there may be no residual problem with stature in adult life. In contrast, there are many situations when this happy outcome is not achieved and there is still a persistent problem requiring attention in the adult clinic. Thus the importance of the various disorders is different and what follows reflects this.

Short stature

Definition

What is considered as short (or for that matter tall) stature is essentially arbitrary. It is usually considered as a height that falls below the second centile for the relevant population; these values for adults are given in Table 1. This is easy when the discrepancy is severe but difficulties arise with patients whose height lies close to this limit. In this situation the perceptions of the patient may greatly colour the situation and its management. The most important decision is whether the apparent short stature is a symptom of a disorder that requires attention, and therefore most stress is placed upon diagnosis.

Classification of short stature

The major categories of short stature are shown in Table 2. The process of attributing such labels to an individual patient is largely one of clinical assessment, including critical appraisal of the pattern of growth combined with appropriate confirmatory investigations.

Familial short stature

This relatively common condition is among the most difficult to manage. It is sometimes referred to as normal or idiopathic short stature. Patients usually present in childhood with heights clustered around the second centile. The parents are usually of comparable size and the child's stature is simply reflecting the genetic inheritance. This apparently straightforward situation is made difficult because of the considerable social pressures that now exist; parents and their children simply find it hard to accept the situation and want it changed.

Diagnosis

The clinical picture is clear: the child will be at or below the second centile for height with a normal growth velocity measured over 1 year. The height will be appropriate for the family and it is critical that parental heights are properly taken into account. The simplest way to do this is to measure both parents and then determine their height centiles. The child's predicted adult height is calculated by an appropriate method that takes into account skeletal maturity; this predicted adult height should lie within a range of 10 cm above or below the midparental height centile, which represents the 2nd to 98th centiles for that family.

The general medical picture is of good health, apart from the usual childhood illnesses. Similarly, apart from short stature, no abnormalities are found on general examination. In this situation no investigations are necessary other than the assessment of skeletal maturity for the prediction of adult height.

Management

The main problem is to persuade the patient and the family that there is no medical problem. Even when this is achieved, there is often a wish to change the prognosis for height. In recent years there have been many clinical trials of human growth hormone in this situation. While definitive data are still scarce, there is increasing evidence that although a short-term acceleration of height velocity is almost always achieved, this is not maintained, and adult height is relatively unchanged. When such treatment is started after the age of 11 years this is moderately certain, but there remains some uncertainty if the human growth hormone is started at younger ages (6 to 8 years). Even if this is more successful, there remains a considerable debate about the ethics and cost-effectiveness of such treatment of normal healthy children.

Constitutional delay of growth and/or puberty

This problem is largely covered elsewhere (see Chapter 12.9.3). Here comment is restricted to noting that in many cases the delay in maturation is evident well before puberty when the presentation is purely of short stature associated with delayed skeletal maturation. There is often an associated element of familial short stature; the delay is amenable to treatment with anabolic steroids or androgens, whereas the familial element is unaffected.

Intrauterine growth retardation

This comprises a relatively large group of children who share the common feature of being born inappropriately small for gestational age, usually assessed in terms of birth weight. Many show a sustained period of accelerated or 'catch-up' growth and reach the normal centiles for height and weight. However, a significant number fail to do this and remain below the 2nd centile although growing at a normal velocity. In some cases this clinical picture is part of a more general dysmorphic or genetic syndrome, but it may occur alone.

Diagnosis

Traditionally intrauterine growth retardation is diagnosed when a baby is born with a birth weight more than two standard deviations below the mean for gestational age, maternal height, and parity. However, this criterion is changing to reflect intrauterine diagnosis by fetal ultrasound measurements.

In addition to the birth characteristics, the growth curve is typical and there are a number of other clinical features. The facies are small and triangular with a normal sized cranium which, compared with the small face, can give the impression of hydrocephalus. The general habitus is very lean and there may be a degree of limb asymmetry in about 50 per cent of cases. There are often other minor dysmorphic features.

Many children have severe feeding problems in the first few years of life, which may compound the lean appearance and create further anxieties for the families. There may be a degree of delayed skeletal maturation, but this is very variable. General and endocrine investigations are unhelpful.

Management

In all but the most severely affected children growth velocity is normal as long as the intrauterine growth retardation is not part of a more generalized dysmorphic syndrome. However, the long-term prognosis for height is usually rather worse than appears in childhood. This is because puberty tends to be prompt and the adolescent growth spurt attenuated, so that less growth is achieved in the teenage years than might be expected.

The only active treatment that is being pursued at present is the use of human growth hormone, but this is still under clinical trial. Early results suggest short-term benefit, but predicted mature height is little changed, as in the case of familial short stature.

Environmental short stature

Over the past 30 years it has become clear that children's growth may be adversely affected by their environment other than by malnutrition or infection. This effect may be due to emotional or physical abuse or neglect and may sometimes present in the most bizarre of ways, often imitating other organic disorders such as growth hormone insufficiency. It is most common in young children but may rarely present in young teenagers and, strikingly, may present in a single child in a family where

the other children thrive.

Diagnosis

This is a diagnosis which is notoriously easy to miss. In its most classical form it presents as an apparently straightforward diagnosis of growth hormone insufficiency although there will often be additional features of bizarre eating habits, behavioural abnormalities, or other evidence of abnormal family dynamics. In many cases where an initial diagnosis of growth hormone insufficiency is made, treatment with human growth hormone is started, often with early success. However, the accelerated growth soon falters and during the subsequent reappraisal the real diagnosis is uncovered.

As the above suggests, endocrine investigations are often misleading. Once the suspicion is raised the only way to confirm the diagnosis is a period of time away from the family environment, either by a hospital admission or by short-term fostering. A period of accelerated growth (initially weight gain followed by increase in height) makes the diagnosis. There is then the need for a detailed social and psychiatric appraisal of the family.

Management

Once the diagnosis is made the management is predominantly that of manipulation of the family circumstances, either by extensive social and psychiatric support for the family with the child *in situ*, or more commonly by removing the child from the family and placement elsewhere by fostering or adoption.

With proper management in childhood those who have suffered from environmental short stature should not present medical problems in adult life. However, it is highly likely that there may be continuing psychological problems and particularly the possibility of subsequent abuse of one or more of their own children.

Chronic paediatric disease

Any child with a longstanding chronic disease will show some degree of growth disorder if the primary activity of the underlying condition is less than optimally controlled. This may be due to the disease itself or to its treatment; the use of high doses of systemic corticosteroids is a particularly important example of the latter. However, it is very unusual for such a child to present with short stature as the primary symptom because the underlying disease is usually evident at an earlier stage. Notable exceptions to this rule are some gastrointestinal diseases, especially coeliac and inflammatory bowel disease. Irrespective of the primary disease, a degree of delayed skeletal maturation is prominent and may be the first feature of disordered growth to appear.

Management is targeted at the primary disease wherever possible. Adjunctive therapies, such as human growth hormone in chronic renal failure, are being studied within clinical trials. The long-term benefits of such treatments remain unclear at the present time.

Endocrine disease

The endocrine causes of short stature form the main group of readily treated growth disorders. Many of the details will have already been covered in the chapters on the relevant endocrine glands and only a brief summary is given here.

Hypothyroidism

Inadequately treated congenital or acquired juvenile hypothyroidism is always associated with marked growth failure. Congenital disease ought to be diagnosed by neonatal screening programmes, such that growth is never a problem, although some delay in initiating thyroid replacement may lead to the more serious problems of intellectual impairment. On the other hand, juvenile hypothyroidism, usually due to autoimmune thyroiditis, can be far more insidious and often presents with growth failure as the sole symptom.

Diagnosis

In any undiagnosed short child consideration should always be given to the possibility of juvenile hypothyroidism. The classical features of adult disease such as constipation and intellectual impairment are usually absent and poor growth may be the only abnormality. Most typically the child grows with a very low velocity, rapidly crossing centiles downwards. General medical examination may be quite normal unless the disease is long established and severe; such symptoms and signs as may be present are no different from those in adults. A goitre may be present, depending upon the stage of the disease.

A striking delay in skeletal maturation is often seen; to the unwary this may lead to an incorrect diagnosis of constitutional delay of growth and/or puberty, as the degree of maturational delay will usually be sufficient to apparently explain the height deficit. For this reason thyroid function tests should be undertaken early and will confirm or deny the diagnosis without ambiguity. Serum T_3 or T_4 will be low, with an elevated serum concentration of thyroid-stimulating hormone.

Management

Thyroid replacement with L-thyroxine is relatively straightforward, with an initial dose of 100 µg/m ^{2}/day; this is fine-tuned according to growth response, clinical examination, and maintenance of a suppressed thyroid-stimulating hormone concentration with either a normal serum T_3 concentration or serum T_4 in the upper normal range. The prognosis for height is generally very good, except in those diagnosed and treated well into puberty, when the outcome is less satisfactory.

Growth hormone insufficiency

Growth hormone insufficiency may occur as an isolated disorder of uncertain aetiology, as part of more complex diseases and developmental anomalies, or as part of more extensive hypothalamopituitary disease. A list of such conditions is given in Table 3. Recently, it has become clear that there are a number of genetic causes of growth hormone insufficiency and wider pituitary malfunction. Deletions or mutations of the structural growth hormone gene on chromosome 17 is now well recognized as are the cascade of homeobox genes responsible for pituitary development including *PIT1*, *PROP1* and *HESX1*.

The principal differences between the congenital and acquired forms are that the onset of the growth failure occurs differently. In the first case there is failure of normal growth from a very early age which can usually be detected within the first year of life. In contrast, acquired growth hormone insufficiency can lead to abnormally slow growth at any age before maturity. In the latter situation the assessment of height velocity assumes far greater importance as height may remain within the normal centiles for several years before the child becomes overtly short and below the third centile.

Particularly noteworthy is the pattern of growth in children receiving cranial or craniospinal irradiation as this may be rather different from other causes of growth hormone insufficiency. In both cases there may be associated early puberty, particularly with radiation doses below 2400 cGy, leading to a rather confusing picture as the early but rather attenuated adolescent growth spurt may mask the onset of growth hormone insufficiency. When the spine is involved in the radiation field there may be a combination of growth hormone insufficiency and direct damage to the spinal epiphyses, with subsequent failure of spinal growth which is not due to the endocrine abnormalities and is not responsive to endocrine replacement.

Diagnosis

Children with growth hormone insufficiency grow with a slow height velocity and, depending on whether the insufficiency is congenital or acquired later, fall progressively further below the second centile or cross height centiles downwards. The degree of short stature can range from relatively mild to very severe, depending on the degree of insufficiency, the age of onset, and parental heights. As in many short stature disorders, the parents' heights still modify the expression of the disorder such that children with growth hormone insufficiency born to tall parents will tend to be more normal in height for longer than those with equivalent disease born to shorter parents.

Other clinical features include a rather young appearance, with immature facies and sometimes a quite striking degree of frontal bossing. There is usually an excess of subcutaneous fat, giving a rather cherubic appearance, and in boys there may be hypoplastic genitalia not necessarily due to associated gonadotrophin deficiency

but caused by a severe growth hormone insufficiency in the intrauterine or neonatal perod.

The diagnosis is confirmed by a variety of endocrine measurements (see Chapter 12.2). Growth hormone is secreted in an episodic manner such that measurement in single serum samples is valueless. It is most commonly measured in several samples following pituitary stimulation by a provocative agent, of which the most important are listed in Table 4. The definition of a normal response is still rather arbitrary, but is usually set at a peak of more than 7 mg/l at some point following the administration of the provocative agent. It is important that the characteristics of different growth hormone assays are taken into account and centres carrying out these tests should establish their own cutoff values. The place of growth hormone releasing hormone is uncertain; it tests the integrity of the hypothalamopituitary axis but does not detect primary hypothalamic disease, which is probably the most important cause of growth hormone insufficiency.

Additional aids to diagnosis are the measurement of growth hormone concentrations in serial blood samples taken frequently (every 15 to 20 min) over 24 h, or the measurement of basal serum concentrations of insulin-like growth factor I together with insulin-like growth factor binding protein 3. The former approach is laborious and labour intensive and is probably unhelpful except in some rare situations where an assessment of physiological growth hormone secretion is required. Measurement of insulin-like growth factor I and its binding protein is rather too insensitive for routine clinical use.

Growth hormone insufficiency may be isolated or part of a wider constellation of pituitary hormone deficiencies, and it is important to check thyroid and adrenal function at an early stage. Gonadotrophin deficiency is relatively common but difficult to confirm prior to the age of puberty, and it may not be until puberty fails to occur spontaneously that suspicion is raised (see Chapter 12.9.3).

Having confirmed the diagnosis of growth hormone insufficiency, isolated or otherwise, it is important to determine the underlying aetiology and, in particular, seek an intracranial space-occupying lesion. Craniopharyngioma is the most important lesion and the use of modern imaging techniques, cranial computed tomography or magnetic resonance imaging, is mandatory. These lesions are most commonly associated with multiple pituitary hormone deficiencies, although these may take some years to become manifest (see Chapter 12.2).

Management

The treatment of growth hormone insufficiency is the same whatever the underlying aetiology: human growth hormone, 5 to 10 mg/m^2/week (0.2–0.4 mg/kg/week; 1 mg pure protein = 3 IU), by daily subcutaneous injection. The growth velocity is the best indicator of response and it should show a clear acceleration, usually to 10 cm/year or more. A poor response indicates the need to review the diagnosis. It is essential that any coexisting pituitary hormone deficiencies (such as deficiency of thyroid-stimulating hormone leading to secondary hypothyroidism) are adequately treated with appropriate replacement.

Treatment is continued until growth is complete and may therefore last for many years. For this reason it is particularly important to review dosage as growth occurs and ensure that injection sites are cared for adequately.

Growth hormone receptor deficiency

This is a very rare disorder whose importance lies in the ability to mimic the much more common severe growth hormone insufficiency. The clinical phenotypes may be indistinguishable but in the case of growth hormone receptor deficiency there is excessive secretion of growth hormone by the pituitary gland; the fault lies in the growth hormone receptor, which is either absent or non-functioning, leading to a deficiency of insulin-like growth factor I. The abnormality is due to one of several mutations in the receptor gene, which is inherited according to an autosomal recessive pattern. Until now it has been untreatable but clinical trials of recombinant insulin-like growth factor I look promising.

Adrenocortical excess (see also Chapter 12.7.1)

An excess of circulating corticosteroids is a potent cause of growth failure, whether due to endogenous overproduction or exogenous medication. In the former case, the aetiology may be hypothalamopituitary or adrenal in origin, as in adults. The diagnosis of Cushing's disease in childhood is rare and taxing to make, although once considered, the approach does not differ from that in adults. Iatrogenic glucocorticoid excess is much more common and is most often related to overuse of topical steroids for atopic disease; inhaled steroids for asthma and powerful dermatological preparations for eczema are particularly important. The management of the growth failure is entirely dependent on reducing the glucocorticoid load usually by the introduction of an alternative non-steroidal treatment for the underlying condition.

Genetic/chromosomal disorders, dysmorphic syndromes, and bone dysplasias

For the purposes of this book these last three categories can be considered together. For the main part they are individually rare, but are so many and varied that together they make a significant contribution to the causes of short stature. The approach to diagnosis depends heavily on clinical suspicion backed up by chromosomal and radiological investigation.

Turner syndrome

This is the only condition that will be discussed in any detail as it is relatively common (about 1 in 2500 to 3500 female births), surprisingly easy to miss, and amenable to useful treatment. Turner syndrome tends to present in two distinct age groups: birth or infancy and mid childhood. The young girls usually have a number of the classical features, including coarctation of the aorta leading to early clinical suspicion. However, a large number of affected girls only have subtle clinical signs, and in these patients short stature is virtually the only significant feature. Diagnosis is confirmed by chromosomal analysis, which may reveal a 45X karyotype with complete absence of one X chromosome, a more subtle structural abnormality of one X, such as an isochromosome, or a mosaic combination of cells with different chromosomal complements.

Untreated girls with Turner's syndrome reach adult heights of between 134 and 156 cm but this is dependent on their parents' heights; girls from tall families will be relatively tall for the diagnosis, even reaching into the lower part of the normal range. Puberty is usually, but not always, absent.

Treatment for both the short stature and the lack of puberty is possible. The latter requires the use of oestrogen and progestogens. A typical regimen would be the slow introduction of ethinyl oestradiol at about 12 to 13 years of age in a dose of 1 µg/day increasing to doses of 20 to 30 µg/day over 2 years. It should be given continuously at first but when adult doses are reached it should be omitted for one week in every four, when a withdrawal bleed will occur, mimicking menstruation. At the same time it is important that a progestogen, such as norethisterone 5 mg/day, is introduced for the last week of the cycle. More recently, clinical trials have shown that moderate growth benefit can be achieved by the combined use of human growth hormone (7–10 mg/m^2) and the mild anabolic steroid, oxandrolone (1.25–2.5 mg/day).

Tall stature

The causes of excessive growth are far fewer than those of short stature. Most common are variants of normal, usually with tall parents; pathological causes of tall stature are very rare.

Definition

This can be defined in a complementary way to short stature (see above); it is equally arbitrary. In practical terms, boys find difficulty in accepting heights above 200 cm, whereas most girls find 185 cm the limit.

Classification

The principal causes of tall stature are familial tall stature, pituitary gigantism, Sotos syndrome, Marfan syndrome, and homocystinuria; only the first will be discussed in any detail.

Familial tall stature

This is more or less the mirror image of familial short stature which is discussed above. There is often an element of advanced maturation with a skeletal age that exceeds chronological age by several years and early pubertal development. The diagnosis is made by clinical appraisal, including knowledge of parental heights and the demonstration that predicted adult height is appropriate for the family. Exclusion of other potential causes is often possible on clinical grounds, but the exclusion of excessive growth hormone secretion may be necessary.

Management

The calculation of a predicted adult height, which because of the advanced skeletal maturation is often less than the family fears, may be all that is necessary as the expected height is then acceptable. If this is not the case, then other pharmacological treatments may need discussion. At present these are unsatisfactory, although in girls it may be possible to curtail final height by induction of puberty early using physiological doses of ethinyl oestradiol, if they present early enough. In the past high-dose ethinyl oestradiol (100–300 µg/day) has been advocated in an attempt to accelerate skeletal maturation. However, the benefits are far from certain, there may be quite unpleasant side-effects, and the long-term safety is unclear.

The use of testosterone in boys, in an analogous manner to the use of oestrogen in girls, is of even less value and is probably contraindicated.

Other causes of tall stature

Pituitary gigantism with excessive growth hormone secretion is extremely rare but does occasionally require specific exclusion, usually by demonstration of normal suppression of growth hormone secretion to undetectable levels by oral glucose (1.75 g/kg). Serum concentrations of insulin-like growth factor I will usually be high but may overlap the normal range.

Marfan syndrome is characterized by disproportionately long limbs and digits (arachnodactyly), and is usually associated with a high-arched palate and pectus excavatum. It is an important diagnosis to make because of the risk of eye problems and dissection of the aortic root and arch. Ultrasound examination of the heart and aorta should be a regular routine.

Sotos syndrome and the other dysmorphic causes of tall stature are even rarer and can usually be diagnosed on other criteria. They are not considered further here.

Further reading

Dattani MT, Robinson IC (2000). The molecular basis for developmental disorders of the pituitary gland in man. *Clinical Genetics* **57**, 337–46.

Massoud AF, Hindmarsh PC, Brook CGD (1992) Disorders of stature. In: Grossman A, ed. *Clinical endocrinology*, 2nd edn, pp 855–84. Blackwell Scientific, Oxford.

Preece MA (1992). Principles of normal growth: auxology and endocrinology. In: Grossman A, ed. *Clinical endocrinology*, 2nd edn, pp 845–54. Blackwell Scientific, Oxford.

Preece MA (1999). Evaluation of growth and development. In: Barratt TM, Avner ED, Harmon WE, eds. *Pediatric nephrology*, 4th edn, pp 329–41. Lippincott, Williams and Wilkins, Baltimore.

Tanner JM (1962). *Growth at adolescence*. Blackwell, Oxford.

Woods KA *et al.* (1997). Phenotype–genotype relationships in growth hormone insensitivity syndrome. *Journal of Clinical Endocrinology and Metabolism* **82**, 3529–35.

12.9.3 Puberty

R. J. M. Ross and M. O. Savage

Introduction

Puberty, as defined by the *Concise Oxford dictionary*, is 'the state of being functionally capable of procreation' through the natural development of reproductive organs. The word is derived from '*puber*' meaning adult, and not 'pubic', which refers to the lower part of the abdomen, the pubes. There is a popular misconception that the onset of puberty is heralded by the development of pubic hair, but breast budding is usually the first sign of puberty in girls and an enlargement in testicular size in boys. A clear understanding of normal pubertal development is essential for the management of patients with disordered puberty, as in many cases counselling and reassurance is all that is required.

Sexual differentiation takes place at two stages of life: the first *in utero* extending to the perinatal period, and the second occurring at puberty. Between these two stages is the 'quiescent period'. The physiological changes that accompany sexual differentiation and the hormonal factors that control these changes are well defined, but what determines the duration of the 'quiescent period' and the onset of puberty remains to be established. What is known is that puberty is centrally driven, and this is well illustrated by the failure of pubertal development in children with Kallmann's syndrome and the changes that occur in anorexia nervosa. In Kallmann's syndrome there is a failure in the migration of gonadotrophin-releasing hormone (**GnRH**) neurones to the hypothalamus during fetal life. Affected patients present with hypogonadotrophic hypogonadism associated with anosmia. When puberty is delayed or arrested by anorexia nervosa it may be induced by the pulsatile administration of GnRH. The GnRH pulse generator is therefore essential for normal puberty, and the cues that switch it to pubertal mode include, most importantly, age and maturation of the central nervous system, environmental factors such as stress, social factors (probably the reason for an earlier onset of puberty in Western countries), and metabolic factors such as nutrition, body composition, and leptin.

The onset of puberty is characterized by an increase in basal luteinizing-hormone (**LH**) levels and in the amplitude and frequency of LH pulses independent of gonadal changes. Gonadal activation stimulated by the rise in gonadotrophin (LH and follicle-stimulating hormone (**FSH**)) secretion results in rising levels of the sex steroids. Apart from their action on sexual maturation, gonadal steroids have a direct effect in stimulating skeletal growth and also a central action in stimulating increased growth-hormone (**GH**) production. A consistent pattern of hormonal changes results in a relatively constant pattern of growth and pubertal development, characterized in girls by the development of breasts, pubic hair, and the onset of menstruation, and in boys by an increase in testicular volume, genitalia size, and the appearance of pubic hair. This is best appreciated by plotting a child's development on growth and development records. A loss of the normal pattern of development suggests pathology. For instance, a boy who at 8 years has a height above the 97th centile, stage 3 genitalia, and pubic hair, but testes less than 4 ml is likely to have an abnormal source of androgens, such as an androgen-secreting tumour or congenital adrenal hyperplasia.

Timing of puberty

Disorders of puberty can be classified by the timing of the onset of sexual characteristics into either precocious or delayed puberty. Precocious puberty is characterized by signs of sexual maturation appearing less than 2.5 standard deviations (**SD**) from the mean: before 8 years of age in a girl and before 9 years in a boy. In Western society puberty is considered delayed when there are no signs of pubertal maturation in a girl aged 13.4 years (2 SD) or a boy aged 13.8 years. As a simple working rule, investigation should be considered if there are no signs of puberty at 14 years of age. These ages are guidelines as they vary between populations and over time.

Precocious puberty

Precocious puberty can be classified into true (pituitary gonadotrophin-dependent) or pseudo (pituitary gonadotrophin-independent) precocious puberty (Table 1). In true precocious puberty, as in normal puberty, there is activation of the hypothalamopituitary axis and thus the normal pattern of puberty is preserved (complete precocious puberty). In pseudo precocious puberty, for example that caused by an adrenal adenoma, the normal pattern of puberty is lost (incomplete precocious puberty). Pseudo precocious puberty may be isosexual, with appropriate male or female puberty, or heterosexual, when there is virilization of a girl (as in congenital adrenal hyperplasia), or feminization of a boy (as in an oestrogen-producing Leydig cell tumour). Two conditions do not fit clearly into this classification: isolated thelarche and isolated pubarche.

Isolated thelarche and isolated pubarche

Breast enlargement in the absence of other signs of puberty is called premature thelarche. It is most common under the age of 2 years and may persist from neonatal breast enlargement. There is usually spontaneous regression and later a normally timed puberty. Isolated pubarche is the early appearance of pubic hair with or without axillary hair. It is more commonly seen in girls than boys and characteristically between 4 and 6 years of age. It is associated with adrenarche, an increase in adrenal androgen secretion seen in middle childhood. There can be a slight growth spurt and advance in bone age, but this is part of normal development. It can be differentiated from abnormal forms of virilization, including adrenal tumours and congenital adrenal hyperplasia, by measuring the sex steroid hormone profile, including dehydroepiandrosterone-sulphate (**DHEA-S**), and demonstrating normal suppression of adrenal androgens by dexamethasone. It has been suggested that premature adrenarche may be a precursor of the polycystic ovarian syndrome in girls, and some clinicians advocate long-term follow up for these patients.

Precocious puberty

Precocious puberty presents much more commonly in girls than boys, and in the majority of girls no organic cause is found and it is idiopathic and sporadic. In contrast, in boys idiopathic precocious puberty is rare and, although there are families with familial true precocious puberty, most commonly it is due to CNS tumours, either hypothalamic hamartomas or gliomas, or dysgerminomas (Table 1).

The clinical investigation of precocious puberty is first directed towards distinguishing between a true, pseudo, isosexual, or heterosexual condition. History and examination will help to establish whether there is a normal pattern of pubertal development, as in true precocious puberty, or an abnormal pattern as seen in pseudo precocious puberty. In girls, ultrasound of the pelvis will demonstrate the effect of oestrogens on uterine size and define the appearance of the ovaries. Measurement of the gonadotrophin response to GnRH should be made, as should basal measurements of b-human chorionic gonadotrophin, adrenocorticotrophic hormone, adrenal steroids (including cortisol, 17-hydroxyprogesterone, DHEA-S, and androstenedione), testosterone, oestrogen, and thyroid hormones. Steroid profiles may also be made on urine collections. Skeletal maturation should be determined by measuring bone age.

True precocious puberty

In true precocious puberty the gonadotrophins will show a pubertal response to GnRH with a greater rise of LH than FSH. In the normal prepubertal child there is only a small rise in the gonadotrophins and the response of FSH is greater than that of LH. In pseudo precocious puberty the gonadotrophins are usually suppressed unless true puberty has also been initiated, which may occur due to excessive sex-steroid secretion from any cause. Acquired hypothyroidism is associated with increased levels of FSH and may result in breast development and menstruation in girls and testicular enlargement in boys. These patients usually have stunted growth and the diagnosis is easily made by the measurement of thyroid stimulating hormone levels. Once a diagnosis of true precocious puberty is made, appropriate

scanning of the hypothalamopituitary axis should be performed using magnetic resonance imaging (**MRI**).

Pseudo precocious puberty

The further investigation of pseudo precocious puberty depends on the findings of the original screening tests. Adrenal tumours will be associated with an increased production of adrenal steroids which is not suppressed by a low-dose dexamethasone suppression test. Imaging will pick up adrenal carcinomas (usually greater than 6 cm in diameter) and most adenomas, although on occasion venous catheter sampling is required. Congenital adrenal hyperplasia in girls usually presents early with virilization and ambiguous genitalia; however, it may present later in boys with virilization and tall stature, but prepubertal testes. There is a typical urinary steroid profile, and in the commonest form there is 21-hydroxylase deficiency, with levels of ACTH and 17-hydroxyprogesterone raised, and low levels of cortisol. Ovarian tumours are best detected by ultrasound scanning, as are testicular tumours. The McCune–Albright syndrome (polyostotic fibrous dysplasia) is an unusual cause of precocious puberty due to postzygotic activating mutations in the gene encoding the G protein, G alpha s, resulting in activation of the signal-transduction pathway generating cyclic AMP. Girls usually present with autonomous ovarian activity, but this may be succeeded by true precocious puberty. Patients have patches of *café-au-lait* pigmentation with a ragged border and fibrous dysplasia of the bones. Testotoxicosis is an unusual inherited disorder due to activating mutations of the gene for the LH receptor. It is characterized by pubertal levels of testosterone, pubertal-sized testes, and a suppressed hypothalamogonadal axis. Tumours producing human chorionic gonadotrophin (**hCG**) can be detected by measuring hCG and scanning appropriate sites, including the gonads, liver, and pineal gland.

Treatment

There are four aims in the treatment of precocious puberty:

1. to remove the primary cause;
2. to treat the psychosocial consequences;
3. to allow a normal puberty; and
4. to promote normal growth.

Children with precocious puberty appear much older than they are, and this can result in considerable psychological difficulties and behavioural problems. Growth is stimulated both by a direct action of sex steroids on skeletal maturation and by the induction of GH secretion. This early maturation of the skeleton results in early fusion of the epiphysis, and although the child is initially tall, his or her ultimate height may be very short.

Girls with only slightly advanced pubertal development often require no treatment because puberty only advances slowly, and they do not have a significant loss in their height potential. GnRH analogues are now the treatment of choice in true precocious puberty. They act by downregulating the GnRH receptor and switching off the secretion of LH and FSH. GnRH analogues have been produced as nasal sprays, daily injections, and monthly depot injections. In our experience, depot injections have proved the most effective (goserelin 3.6 mg/month). GnRH may produce an initial period of stimulation, which can be prevented by giving concomitant treatment with cyproterone acetate (100 mg/m^2 once a day for the first 6 weeks). Occasionally the acute suppression of oestrogen production at the start of treatment will precipitate an oestrogen-withdrawal bleed.

In pseudo precocious puberty removal of the primary cause is the mainstay of treatment for patients with tumours, and treatment with glucocorticoids for patients with congenital adrenal hyperplasia. In patients with congenital activating mutations of receptors or residual disease after treatment of tumours, cyproterone acetate, a peripherally acting antiandrogen, at a dose of 50 to 100 mg daily is effective in halting the progress of the physical features of puberty, and is useful in suppressing menstruation. Cyproterone acetate is a weak glucocorticoid and may suppress ACTH and the adrenal glands. Testolactone, and a combination of this with spironolactone and GnRH analogues, has proven effective in improving height prediction.

Any effective treatment of precocious puberty will slow the growth rate through the consequent reduced secretion of sex steroids and GH. Patients treated to arrest puberty will have longer to grow, but their reduced growth rate and already reduced growth potential mean that, despite the use of GnRH analogues, they will not achieve the height of which they were originally capable. The addition of GH treatment for 2 to 3 years during conventional therapy with GnRH analogues may improve the height prognosis in children with a low growth velocity.

Delayed puberty

The causes are summarized in Table 2. The individual conditions which may cause it are discussed in other parts of this book. Here, discussion is limited to the management of constitutional delay of growth and adolescence.

Constitutional delay of growth and adolescence (CDGA)

CDGA occurs in otherwise normal adolescents who have relatively short stature, delayed puberty and bone age, and a height prognosis appropriate in relation to their parents. It presents far more commonly in boys than girls and is the commonest cause of delayed puberty in boys, with Turner's syndrome being the commonest in girls. CDGA needs to be distinguished from isolated gonadotrophin deficiency, but this is rarely easy as gonadotrophin levels are low, with a low or prepubertal response to GnRH in both conditions. A positive family history may indicate CDGA, and associated anosmia suggests Kallmann's syndrome. If in doubt and if treatment is indicated, then the patient should be reassessed after therapy (see below) to see if puberty then progresses without treatment.

Psychological problems are common in children with delayed puberty and short stature. Recent studies have suggested that delayed puberty may be associated with a reduced spinal bone density, putting adults at risk of bone fracture later in life. Thus, there are good reasons for treating this condition, which may be considered as a variant of normal growth.

Intervention with sex steroids or anabolic steroids is a safe treatment that brings forward the timing of the growth spurt without reducing the height potential. The object of treatment is to stimulate normal puberty and maximize linear growth. In boys a reasonable starting dose of testosterone esters is 25 to 50 mg monthly, increasing gradually to 250 mg every 4 weeks, although puberty may be induced more rapidly over a 6-month period and the course of treatment may be as short as 3 months. Oral Oxandralone (unlicensed, but available on a named-patient basis from Searle, UK), 2.5 mg daily, will similarly increase growth velocity.

In girls, ethinyloestradiol at an initial dose of 2 to 10 µg daily can later be increased to between 10 and 20 µg daily, with the addition of progesterone when the oestrogen dose has reached 20 µg (for example, medroxyprogesterone acetate 5 mg on days 1–14 of the calendar month).

Further reading

Klien KO (1999). Editorial: Precocious puberty: who has it? Who should be treated. *Journal of Clinical Endocrinology and Metabolism* **84**, 411–14.

Leschek EW, *et al*. (1999). Six-year results of spironolactone and testolactone treatment of familial male-limited precocious puberty with addition of deslorelin after central puberty onset. *Journal of Clinical Endocrinology and Metabolism* **84**, 175–8.

Pasquino AM, *et al*. (1999). Adult height in girls with central precocious puberty treated with gonadotropin-releasing hormone analogues and growth hormone. *Journal of Clinical Endocrinology and Metabolism* **84**, 449–52.

Saenger P (1992). Editorial: Premature adrenarche: a normal variant of puberty? *Journal of Clinical Endocrinology and Metabolism* **74**, 236–8.

Stanhope R, Albanese A, Shalet S. (1992). Delayed puberty. *British Medical Journal* **305**, 790.

12.10 Non-diabetic pancreatic endocrine disorders and multiple endocrine neoplasia

P. J. Hammond and S. R. Bloom

Pancreatic endocrine tumours

Pancreatic endocrine tumours (islet cell tumours, gastroenteropancreatic tumours) are rare, the most frequent, insulinomas and gastrinomas, occurring with an annual incidence of 1 per million, with others having an incidence of less than 1 per 10 million. Functioning tumours usually present with the symptoms of hormone excess. They may secrete the pancreatic hormones insulin, glucagon, and somatostatin, or ectopic hormones such as gastrin, vasoactive intestinal polypeptide (VIP), or parathyroid hormone related peptide (PTHrP) (see Chapter 12.6). Non-functioning tumours can reach a large size in an apparently well patient, as characteristically these tumours cause little non-endocrine systemic upset. They were once often mistakenly identified as adenocarcinomas, but are now increasingly diagnosed as a result of detection of their secretion of functionally inactive peptides, such as pancreatic polypeptide and neurotensin, or immunohistochemical staining for neuroendocrine markers, such as chromogranin and neurone-specific enolase. They probably account for 50 per cent of all pancreatic endocrine tumours. This section will initially consider aspects of tumour biology, diagnosis, and management common to all tumours, before describing each syndrome.

Natural history

These tumours were originally described as APUDomas because it was thought that they had a common origin from neural crest cells with the ability to perform amine precursor uptake and decarboxylation (APUD). However, this theory has since been disproved, and it has been proposed that the neuroendocrine and mucosal endocrine cells of the gastroenteropancreatic axis are derived from a common, bipotential endoplacal stem cell.

The genetic basis for the development of sporadic pancreatic endocrine tumours is largely unknown. However, about 25 per cent of them, particularly gastrinomas and insulinomas, occur as part of the familial autosomal dominant multiple endocrine neoplasia type 1 (MEN1) syndrome (see below), in which there is a mutation in the *menin* gene in the q13 region of chromosome 11, and loss of heterozygosity for this region has been demonstrated in some patients with sporadic gastrinoma and insulinoma.

Islet cells are pluripotential with respect to peptide production. Thus 70 per cent of tumours are associated with elevated pancreatic polypeptide levels, and in a small proportion of cases other hormones, particularly gastrin, may become elevated and cause secondary syndromes during the course of the disease. Altered processing of peptide precursor molecules may result in a variety of molecular weight forms of the same peptide being secreted, and not all the immunoreactive peptide is bioactive. This can have clinical implications; for example large molecular forms of glucagon (enteroglucagon) can cause villous hypertrophy and slowed intestinal transit, and large forms of somatostatin have been reported to cause hypoglycaemia, rather than the hyperglycaemia usually associated with the somatostatinoma syndrome.

Most pancreatic endocrine tumours are slow-growing and prolonged survival is often possible, even in the presence of metastatic spread, the median survival being about 5 years. However, some patients have aggressive, rapidly spreading disease, particularly those with non-functioning tumours, whose median survival is little over 2 years. Early in the disease, morbidity and mortality result from the effects of peptide hypersecretion rather than tumour bulk. The unpredictable nature of these tumours makes it difficult to give an accurate prognosis, occasional patients surviving for decades, and, combined with their rarity, this has made it difficult to assess the efficacy of different therapeutic strategies.

Diagnosis

Pancreatic endocrine tumours can usually be diagnosed by hormonal radioimmunoassay of a single fasting plasma sample, and for certain syndromes a small number of confirmatory tests. Several conditions other than tumour are associated with increased circulating gut hormone levels (Table 1), particularly renal failure, but the elevations are usually more modest than those associated with tumour syndromes. Gut hormone radioimmunoassays are not well standardized, and the use of different antibodies and assay techniques in different laboratories can give different values on the same sample. However, concentrations are usually of the same order of magnitude in all assays and show a similar percentage increase above normal.

Most glucagonomas, non-functioning tumours, and pancreatic VIPomas and somatostatinomas are large (greater than 2 cm) tumours, which may be calcified and have metastasized to the liver in the majority of cases. Such tumours are easily localized by abdominal computed tomography (CT) scanning and ultrasonography. However, localization of tumours producing more active hormones, which are therefore detected earlier in their lifecycle, may be very difficult; for example 40 per cent of gastrinomas and insulinomas are microadenomas, less than 1 cm in diameter. Insulinomas often occur in the distal two-thirds of the pancreas, and over 90 per cent of gastrinomas are found in the gastrinoma triangle, bounded by the third part of the duodenum, the neck of the pancreas, and the porta hepatis, about 20 per cent of these being in the duodenum. CT scanning and meticulous, highly selective angiography (Fig. 1) will localize 70 per cent of these microadenomas. Magnetic resonance imaging (MRI) may be more sensitive at detecting small pancreatic lesions than CT scanning, but this method requires further evaluation. Transhepatic percutaneous portal venous sampling is a sensitive method of detecting hormone gradients, but cannot give accurate enough resolution to assist the surgeon in most cases, and is an expensive procedure not without risk. In experienced hands, endoscopic ultrasonography may be more sensitive than conventional imaging, with a resolution of 2 mm and a detection rate of over 75 per cent for tumours in the pancreatic head (Fig. 2), but visualization is poorer for lesions of the pancreatic tail and duodenum. Intraoperative ultrasonography has a sensitivity of over 90 per cent for pancreatic tumours, and endoscopic transillumination of the duodenum may allow the surgeon to detect an occult gastrinoma. Functional radiological localization has been described for both insulinomas and gastrinomas. Injection of calcium or secretin into the artery supplying the tumour causes a marked rise in insulin or gastrin levels, respectively, in the hepatic vein, and allows equivocal lesions to be verified, or the site of unlocalized lesions to be more accurately predicted, and can be used to confirm tumour resection intraoperatively. Somatostatin receptor scintigraphy with radiolabelled somatostatin analogues has proved useful in demonstrating the extent of metastatic disease (Fig. 3), and may assist in the localization of extrapancreatic VIPomas, but is less effective in detecting microadenomas. There is controversy as to whether invasive imaging, usually angiography with provocative testing, or non-invasive imaging, principally endoscopic ultrasonography and somatostatin receptor scintigraphy, is the preferred option for localization of small tumours, but the choice will probably be determined by local expertise.

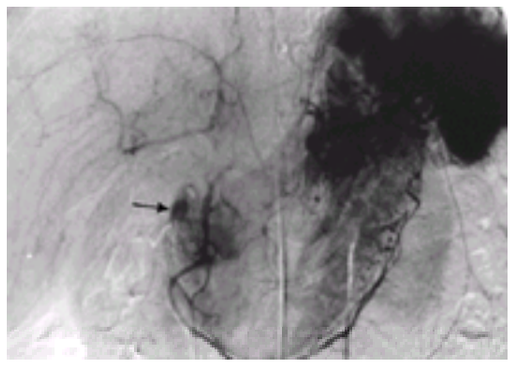

Fig. 1 Venous phase of coeliac axis angiogram demonstrating gastrinoma blush in duodenal wall (arrowed).

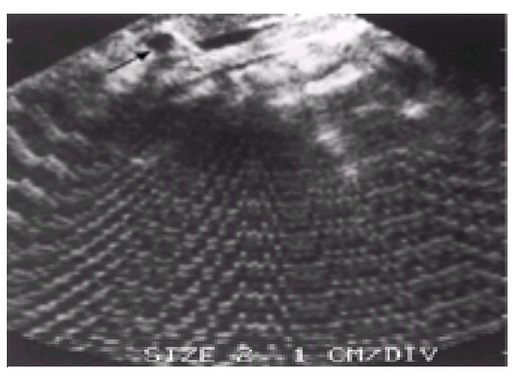

Fig. 2 Endoscopic ultrasound showing a 0.7 cm insulinoma in the head of the pancreas (arrowed).

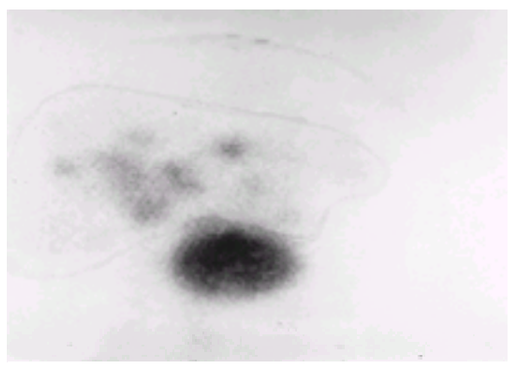

Fig. 3 ^{111}Indium-labelled somatostatin analogue scan showing the large primary tumour and diffuse hepatic metastases in a patient with a pancreatic glucagonoma.

Confirmation of the diagnosis can be made by immunocytochemical analysis of resection specimens or liver biopsies, in addition to conventional histology. Antisera against non-specific markers, such as chromogranins, provide evidence for a neuroendocrine origin, while antisera to different peptides identify the specific tumour type.

Treatment

Surgery offers the only hope of cure for pancreatic endocrine tumours, and all sporadic tumours without evidence of metastatic spread should be resected if possible. Surgical cures have been reported for a few patients with hepatic metastases amenable to enucleation, and, recently, liver transplantation has been successfully performed in patients with metastatic disease confined to the liver.

In the majority of patients, surgical cure is not possible and the aim of treatment in these cases is symptomatic palliation. Until the terminal stages of the disease, this is directed at reducing the symptoms of hormone excess in those with functional tumours. This can be achieved by reducing tumour bulk or inhibiting hormone secretion or action. Reduction of tumour bulk surgically is usually precluded by the operative morbidity. A variety of chemotherapy regimens have been reported as effective, although it is difficult to demonstrate that a particular regimen prolongs survival due to the small numbers of patients for analysis and the unpredictable nature of the tumours. The standard regimen consists of streptozotocin and 5-fluorouracil, with response rates of 80 per cent for functioning tumours, VIPomas responding particularly well, and about 50 per cent for non-functioning tumours. The combination of doxorubicin and streptozotocin has been reported to be more effective, preventing progression of disease for long periods and probably prolonging survival, and other studies have combined doxorubicin, streptozotocin, and 5-fluorouracil. Other agents advocated are dacarbazine for glucagonomas, and cisplatin and etoposide for anaplastic tumours. Another effective means of reducing tumour load in patients with extensive hepatic disease is embolization of the hepatic arterial supply to the metastases, a patent portal vein being needed to support the normal liver parenchyma. Response rates with this procedure are between 60 and 80 per cent. Additional benefit may be gained by chemoembolization, where chemotherapy is delivered via the catheter during the procedure.

Inhibition of hormone release and action is achieved by using the subcutaneous somatostatin analogue, octreotide. Native somatostatin inhibits multiple endocrine functions, acting particularly by blocking hormone effects on the target tissue, but has a half-life in circulation of only 3 min. Octreotide has a half-life of 2 h in the circulation and can be given in three doses daily. The clinical sequelae of peptide hypersecretion are often greatly diminished 24 h after the first injection, although patients become progressively resistant to its effects over many months or years, in part due to continued tumour growth. Recently, long-acting formulations of octreotide have become available, meaning that injections need only be given every 2 to 6 weeks, depending on the response.

Tumour syndromes

Gastrinoma

Gastrinomas are the commonest pancreatic endocrine tumour. Sixty per cent are malignant, 50 per cent of patients having metastases at the time of diagnosis, and up to 30 per cent of patients have the multiple endocrine neoplasia type 1 (MEN1) syndrome. The majority of tumours are pancreatic, but between 20 and 40 per cent are duodenal, and these are usually microadenomas, as little as 1 mm in diameter. Sporadic duodenal microgastrinomas are solitary, but in those patients with MEN1 they are usually multiple and associated with pancreatic microadenomas. Primary lymph node gastrinomas have been described but may represent metastases from duodenal microgastrinomas.

The gastrinoma syndrome was first described in 1955 by Zollinger and Ellison, who reported the triad of fulminating ulcer diathesis, recurrent ulceration with a poor response to therapy, and pancreatic non-b-cell islet tumours. The syndrome is the result of excess gastrin-stimulated gastric acid secretion. This causes severe, multiple peptic ulcers, which are usually duodenal, but may occur in the oesophagus and jejunum, and are often associated with complications such as haemorrhage, perforation, and stricture formation. Diarrhoea and steatorrhoea, due to acid inactivation of small bowel enzymes and mucosal damage, may be prominent features,

frequently preceding ulcer disease by 12 months or more.

The diagnosis of the gastrinoma syndrome requires the demonstration of a raised fasting gastrin concentration, associated with increased basal gastric acid secretion. The patient should, ideally, not take H_2 blockers for 3 days or omeprazole for 2 weeks before the test. Hypergastrinaemia and raised acid output may also arise from retained antrum following partial gastrectomy or the rare condition of G-cell hyperplasia. The intravenous secretin test distinguishes these conditions from gastrinoma and can aid diagnosis when other investigations are equivocal. In the presence of a gastrinoma, gastrin levels are elevated by at least 50 per cent following secretin, while there is no such increase in association with G-cell hyperplasia or retained antrum. Furthermore, gastrin levels are increased in response to a test meal in the latter conditions but not in association with a gastrinoma. Endoscopy may be valuable in demonstrating oesophageal and duodenal ulceration and hypertrophy of the gastric mucosa, while immunocytochemical analysis of antral biopsies may demonstrate G-cell hyperplasia. Localization of microgastrinomas may be aided preoperatively by endoscopic ultrasound or selective arterial secretin injection, or intraoperatively by ultrasonography or duodenotomy with transillumination and careful palpation. Small tumours often secrete gastrin rapidly and store little peptide so that histological diagnosis may only be possible by *in situ* hybridization to demonstrate synthesis of gastrin messenger RNA.

Since all gastrinomas may have the potential to metastasize, localized non-metastatic tumours should be resected, and regular attempts at localization should be made for occult tumours. In the past, the morbidity and mortality of the gastrinoma syndrome resulted from the severe peptic ulceration and associated complications. The best treatment for this was total gastrectomy to remove the source of acid hypersecretion. The H2-blockers provided relief of symptoms for many patients but often failed to suppress acid secretion adequately. The introduction of proton-pump inhibitors, which almost completely inhibits gastric acid production in all cases, has transformed the management of these patients, and offers the best palliation for those with metastatic disease. Most experience has been gained with omeprazole, but there is evidence that the newer agents are equally effective. Omeprazole is acid-labile and so initially should be administered with an H2-blocker. Since the introduction of proton-pump inhibitors, morbidity and mortality now occur much later and result from tumour bulk. Chemotherapy is effective in less than 50 per cent of cases, but hepatic embolization may be beneficial in the remainder.

VIPoma

VIPomas arise in the pancreas in 90 per cent of cases. The remaining tumours are mainly gangliomas or ganglioneuroblastomas originating in the sympathetic chain or adrenal medulla, and these tumours are especially common in children. Most extrapancreatic tumours are benign, but 50 per cent of pancreatic VIPomas have metastasized at the time of diagnosis, usually to local lymph nodes and the liver.

The features of the VIPoma (Verner–Morrison, pancreatic cholera) syndrome (Table 2) reflect the known biological actions of VIP. Large-volume diarrhoea without steatorrhoea is the cardinal symptom, most patients excreting more than 3 litres daily, with volumes of over 20 litres described. It is often intermittent at first, but in severe crises the volume loss coupled with the vasodilatory effects of VIP and the associated hypokalaemia may precipitate cardiovascular collapse.

Hypokalaemia results from loss in stools and activation of the renin–angiotensin system, and may be profound. The loss of bicarbonate in the stool leads to acidosis, which may mask the true potassium deficit. Achlorhydria or hypochlorhydria occurs in over 50 per cent of patients and distinguishes this diarrhoeal syndrome from that associated with gastrinoma, but its absence in a proportion of patients makes the acronym WDHA (watery diarrhoea, hypokalaemia, and achlorhydria) syndrome inappropriate. In up to 50 per cent of cases there is glucose intolerance as a result of the glucagon-like actions of VIP. Other features include: hypercalcaemia, probably due to PTHrP secretion and exacerbated by the dehydration; hypomagnesaemia due to loss in stools; and flushing of the head and neck, which can occur on tumour palpation and may be associated with a marked fall in systemic blood pressure. In advanced cases, extreme weight loss may occur.

VIPomas are usually associated with markedly raised plasma VIP concentrations, but because the half-life of VIP in circulation is only 2 min, the diagnosis is best confirmed by the finding of elevated circulating peptide histidine–methionine, which is produced from the prepro-VIP molecule, is more stable in plasma, and is cosecreted by VIPomas. Pancreatic polypeptide levels are elevated in 75 per cent of cases and neurotensin in 10 per cent. Ganglioneuroblastomas may secrete noradrenaline and adrenaline and so be associated with elevated urinary catecholamines and catecholamine metabolites.

VIPomas are usually large and so localization is rarely a problem. Occasionally, angiography may be necessary to detect small pancreatic lesions, or radiolabelled somatostatin or *m*-iodobenzylguanidine (MIBG) scanning to identify extrapancreatic tumours.

Resection specimens from pancreatic tumours show the structural and secretory patterns of epithelial endocrine tumours, while those from ganglioneuroblastomas show neurones and nerve fibres, together with Schwann cells. Immunocytochemistry detects VIP and peptide histidine–methionine, and electron microscopy shows poorly granulated tumours, with characteristic, small secretory granules.

Patients with non-metastatic disease should have surgical resections, and this is feasible in the majority of ganglioneuroblastomas. Chemotherapy provides very effective palliation for those with metastatic disease, and the excellent response to the comparatively non-toxic regimen of streptozotocin and 5-fluorouracil makes the use of other advocated agents, particularly a-interferon, unnecessary. Similarly, hepatic embolization is not usually indicated for metastatic VIPomas. Acute VIPoma crises should be managed with fluid and electrolyte support, and monitoring of central venous pressure is usually required. A number of drugs, including prednisolone, indomethacin, metoclopramide, lithium carbonate, and opiates, have been used with varying degrees of success to treat the diarrhoea. These have now been superseded by the somatostatin analogue, octreotide. Ninety per cent of patients respond to octreotide, with reduction of diarrhoea almost to normal and resolution of the electrolyte imbalance within 48 h, and it may be life-saving in an acute crisis. Unfortunately, the median duration of response to octreotide alone is less than 1 year, and so its use is probably best combined with chemotherapy.

Glucagonoma

Glucagonomas are a-cell tumours of the pancreas which secrete various forms of glucagon and other peptides derived from the preproglucagon molecule. They have an estimated annual incidence of 1 in 20 million, with a marginal female preponderance, and invariably present in adulthood. Over 70 per cent of patients have metastases at the time of diagnosis.

The characteristic feature of the glucagonoma syndrome is the rash of necrolytic migratory erythema, which occurs in almost all patients, although it often remains undiagnosed for many years (Plate 1). It usually starts in the groins and perineum, migrating to the distal extremities. The initial lesions are erythematous patches, which become raised and may be associated with bullae. These lesions break down and gradually heal, often leaving an area of hyperpigmentation, only to recur in another site. All mucous membranes may be involved, commonly leading to angular stomatitis, cheilitis, and glossitis. The cause of the rash is unknown. A direct effect of glucagon on the skin, glucagon-induced prostaglandin release, amino acid or free fatty acid deficiency, or zinc deficiency, due to the similarity with acrodermatitis enteropathica, have all been proposed as the underlying mechanism. The rash has been reported in a few patients without glucagonomas, who either had coeliac disease or cirrhosis, both of which may have led to elevation in glucagon and glucagon-like peptides. Other common features of glucagonomas include: impaired glucose tolerance, and occasionally mild diabetes requiring insulin therapy; progressive weight loss, which is occasionally severe enough to be fatal; venous thrombosis, which may be life-threatening; normochromic normocytic anaemia, probably as a result of direct bone marrow suppression by glucagon; bowel disturbance and nail dystrophy. Mental slowness, depression, and paraneoplastic neurological syndromes have also been described.

The diagnosis of glucagonoma is confirmed by demonstrating raised fasting plasma glucagon concentrations by radioimmunoassay; the elevation is usually 10- to 20-fold. Localization is almost never a problem, since tumours are invariably large and pancreatic, with metastases in the majority of cases. Barium studies often show thickened jejunal and ileal mucosa due to the trophic effects of large forms of glucagon on the small bowel. The tumour tissue contains large quantities of extractable glucagon which is localized to a-cells. Electron microscopy shows dense-core secretory granules and the core is often eccentric. Fifty per cent of tumours produce pancreatic polypeptide and coproduction of gastrin and insulin has been described. Skin biopsies show necrolysis of the stratum Malpighi of the epidermis in early lesions, but only a non-specific dermatitis at later stages.

Surgical cure of glucagonomas is rarely possible, although it has been claimed for patients with resectable metastatic disease; recently, successful liver transplantation for patients with metastatic glucagonomas has been reported. Surgery is complicated by the tendency to venous thrombosis, the catabolic effects of glucagon, and anaemia. A significant proportion of glucagonomas fail to respond to the combination of streptozotocin and 5-fluorouracil, and in these cases dacarbazine or hepatic embolization may be necessary. Octreotide is particularly effective in treating the rash, with resolution usually occurring within the first month of treatment and persisting for at least 6 months, but it has little impact on the other features of the syndrome. Other simple treatments for the rash which are worth using in all cases are topical or oral zinc and a high-protein diet. Amino acid infusions and blood transfusion may also be effective, but the tendency of the rash to spontaneous remission, often following hospitalization, throws doubt on the value of such procedures. The thrombotic tendency, which can result in fatal pulmonary

emboli, is refractory to conventional anticoagulation, but aspirin or dipyridamole may be of benefit.

Somatostatinoma

Somatostatinomas are extremely rare, with an estimated annual incidence of about 1 in 40 million. Fifty per cent of these tumours are pancreatic, the remainder arising in the duodenum. Approximately 50 per cent of duodenal somatostatinomas occur in association with neurofibromatosis type I (von Recklinghausen's disease) and these tumours are usually periampullary. Pancreatic tumours usually present late with hepatic metastases, but duodenal tumours are frequently identified earlier as a result of local effects.

The somatostatinoma syndrome is characterized by the triad of cholelithiasis, diabetes mellitus, and steatorrhoea, the latter occurring in almost all patients with pancreatic tumours. These features result from the inhibitory actions of somatostatin on gallbladder contraction and secretion, insulin secretion, and pancreatic exocrine secretions. Hypoglycaemia has occasionally been described, possibly due to larger molecular forms of somatostatin having a greater inhibitory effect on counter-regulatory hormones than on insulin. Other features of the syndrome include hypochlorhydria, anaemia, postprandial fullness, and weight loss. The full syndrome is rarely seen in association with duodenal somatostatinoma, gallbladder disease being the only common manifestation. These tumours usually present as a result of effects on local structures, such as obstruction of the ampulla of Vater causing jaundice or pancreatitis, or intestinal obstruction or haemorrhage.

Circulating levels of somatostatin are usually elevated greater than 10-fold in association with pancreatic tumours, but duodenal tumours are associated with much lower levels, probably because they are usually about one-tenth the size of the pancreatic lesions. Multiple molecular weight forms of somatostatin may be demonstrated by column chromatography of plasma or tumour extracts, and these may explain unusual clinical features. Localization is rarely a problem as barium examinations or endoscopy will identify duodenal lesions. Duodenal somatostatinomas are classified histologically as duodenal carcinoids. Duodenal carcinoids have the usual features of neuroendocrine tumours but often contain psammoma bodies. Those associated with neurofibromatosis type I are more likely to be pure somatostatinomas and to contain psammoma bodies.

Surgical resection of duodenal somatostatinomas is usually curative, although a Whipple's procedure may be needed to ensure clearance of periampullary tumours. Pancreatic tumours have almost always metastasized by the time of diagnosis and so palliation with chemotherapy or hepatic embolization are the only therapeutic options.

Pancreatic polypeptide, neurotensin, and other hormones

Pancreatic polypeptide (PP) can be extracted from almost all pancreatic endocrine tumours and is secreted by up to 75 per cent of them. The finding of elevated circulating pancreatic polypeptide in association with other tumour syndromes indicates a pancreatic tumour source. However, pancreatic polypeptide itself has no recognized physiological role and no associated tumour syndrome, and pure PPomas can be regarded effectively as non-functioning tumours. Similarly neurotensin, which is elevated in 10 per cent of VIPomas, does not cause a characteristic syndrome. Interestingly, neurotensin is produced by fibrolamellar hepatomas.

Hypercalcaemia is a feature of the VIPoma syndrome and may also occur in association with pancreatic endocrine tumours without other hormone syndromes. Secretion of parathyroid hormone related peptide (PTHrP) by pancreatic endocrine tumours has now been reported in a number of cases, and synthesis of PTHrP messenger RNA in normal and tumorous islets has been described. It is highly probable, therefore, that almost all cases of hypercalcaemia in association with pancreatic endocrine tumours are mediated by PTHrP. In these patients the hypercalcaemia responds to both octreotide and bisphosphonates.

The hypothalamic hormone, growth hormone releasing hormone (GHRH), was originally isolated from a pancreatic endocrine tumour, and there have been subsequent reports of patients with acromegaly and gigantism as a result of GHRH secretion by pancreatic endocrine tumours. Treatment options for these patients have included surgical resection, octreotide therapy, and liver transplantation.

Another hypothalamic releasing factor, corticotrophin releasing hormone, may be produced by pancreatic endocrine tumours, but this only causes Cushing's syndrome when the tumour also secretes corticotrophin. One patient with an enteroglucagon-secreting tumour of the right kidney, causing villous hypertrophy and slowed intestinal transit, steatorrhoea, and mild diabetes, has been reported, and there has been one case of acromegaly due to a growth hormone secreting pancreatic endocrine tumour. Other peptides produced by islet-cell tumours include neuropeptide Y, neuromedin B, calcitonin gene-related peptide, bombesin, and motilin, but these are not associated with recognized clinical syndromes.

Non-functioning tumours

Tumours not associated with a recognized hormonal syndrome may account for half of all pancreatic endocrine tumours. They usually present late with symptoms attributable to tumour bulk, such as anorexia and weight loss, or to effects on local structures, such as obstructive jaundice or intestinal obstruction or haemorrhage. They are often mistakenly diagnosed as adenocarcinomas, but the presence of elevated circulating gut hormones, such as pancreatic polypeptide or neurotensin, and the use of immunocytochemical analysis, can point to the correct diagnosis. Non-functioning tumours usually respond poorly to chemotherapy, but hepatic embolization may be beneficial. They have a poor prognosis as a result of their late presentation and lack of response to therapy.

Multiple endocrine neoplasia

The multiple endocrine neoplasia syndromes (MEN1 and MEN2) are familial conditions with an autosomal dominant pattern of inheritance and a high degree of penetrance. The genetic defect has been identified for both types of MEN—the MEN1 gene, *menin*, on chromosome region 11q13, and the MEN2 gene, *ret*, on chromosome region 10q11.2. In MEN2 there have been few mutations identified, so that rapid mutation screening is possible, and genotype–phenotype correlations have been identified. In MEN1 many different mutations have been identified, often occurring in only one kindred, and mutation detection can only be done by formal gene sequencing. Identification of specific gene defects in these syndromes may provide novel therapeutic options for tumour prevention in affected individuals.

Multiple endocrine neoplasia type 1 (MEN1)

MEN1 is characterized by the association of parathyroid hyperplasia, pancreatic endocrine tumours, and pituitary adenomas. This association was first described by Underdahl in 1953, and the autosomal dominant inheritance was first proposed in 1954 by Wermer, whose name provided the eponym for the syndrome. The prevalence of the condition has been estimated at about 1 in 10 000. The affected gene has been termed *menin* and is found on the 11q13 region of the long arm of chromosome 11. It encodes a nuclear protein which interacts with the JunD component of the transcription factor AP-1. The development of tumours fits the 'two-hit' model proposed by Knudson and demonstrated by familial retinoblastoma, whereby there is a germline mutation of the MEN1 gene on one chromosome 11, followed by a somatic deletion of the same region on the other chromosome, leading to loss of heterozygosity for that allele and subsequent tumour formation. A substantial proportion of MEN1 cases arise through sporadic mutations, and these patients present between the third and fifth decades, while familial cases can be identified earlier through screening, usually biochemical but increasingly genetic once the mutation affecting a kindred has been sequenced.

Parathyroid hyperplasia and adenomas

Hyperparathyroidism is the presenting feature of MEN1 in the majority of patients, and occurs in almost all cases. Patients present either with asymptomatic hypercalcaemia on biochemical screening or with the features of sporadic hyperparathyroidism. All four glands are diffusely hyperplastic and there may be nodule formation. Whether true adenomas develop remains controversial, but it is assumed that the presence of a capsule indicates adenomatous change. All patients should be operated on to prevent later morbidity from hypercalcaemia and there are two surgical approaches. Subtotal parathyroidectomy may be performed, but hyperparathyroidism will almost always recur, necessitating excision of the remaining parathyroid tissue. However, most surgeons would perform total parathyroidectomy, either with autotransplantation of one gland to the forearm, which can later be removed if hyperparathyroidism recurs, or with immediate replacement therapy with 1a-hydroxycholecalciferol.

Pancreatic endocrine tumours

Pancreatic endocrine tumours occur in about 70 per cent of patients with MEN1, and usually present between the ages of 15 and 50 if not identified by screening. They account for most of the morbidity and mortality of the MEN1 syndrome. Over 60 per cent of tumours are gastrinomas and about 30 per cent are insulinomas, the two coexisting in about 10 per cent of cases. VIPomas have rarely been described and there are only isolated reports of glucagonomas, but non-functioning tumours

may occur frequently. Diffuse hyperplasia of the pancreas is usually seen, similar to the parathyroid, and in the majority of cases there are multiple adenomas, most of which are less than 1 cm in diameter. Duodenal microgastrinomas are very common, probably accounting for almost half of all MEN1-associated gastrinomas, and are usually multiple, with up to 15 separate tumours described.

The surgical approach to pancreatic endocrine tumours in MEN1 is controversial. Surgical cure is best achieved by removing the pancreas and duodenum with adjacent lymph nodes, but such an aggressive approach is only justified in families in which the pancreatic disease has been extremely malignant, and in these kindreds it should be performed only when pancreatic disease is biochemically apparent. An alternative, potentially curative, approach is to perform a subtotal pancreatectomy with enucleation of palpable tumours in the head and careful exploration for duodenal lesions, which should also be resected. A more conservative strategy is to enucleate gross lesions to reduce the risk of developing metastatic disease, although size dose not necessarily correlate with metastatic potential, and then control hormonal syndromes with appropriate medical therapy. The latter approach may be appropriate for gastrinomas because proton-pump inhibitors are such an effective treatment, but for insulinomas, where medical therapy is often unsuccessful and symptoms usually recur after enucleation alone, more aggressive surgical management may be the best option. The treatment of metastatic disease is the same as in sporadic cases.

Pancreatic endocrine tumours associated with MEN1 are less malignant than sporadic tumours and carry a better prognosis, with a median survival of 15 years compared to 5 years for patients with sporadic tumours. This may reflect more indolent disease or earlier diagnosis.

Pituitary adenomas

The true incidence of pituitary adenomas in MEN1 is disputed. They are detected by screening in 30 per cent of patients, but are found at autopsy in over 50 per cent of patients. Unlike the pancreas and parathyroid, there does not appear to be diffuse pituitary hyperplasia, and loss of heterozygosity for the MEN1 locus is much less common in pituitary tumours than in parathyroid and pancreatic lesions.

Prolactinomas are the commonest tumours, occurring in about two-thirds of cases, with acromegaly accounting for about 30 per cent, and other functioning tumours being rare. Treatment is the same as for sporadic pituitary tumours (see Chapter 12.6).

Other lesions

Lesions in other tissues have been reported in association with MEN1, but their relationship to the syndrome remains controversial. Carcinoid tumours of the foregut, midgut, and thymuta occur in about 10 per cent of cases, and are often found in the pancreas, but are rarely symptomatic. Lipomas occur in a significant proportion of patients and act as a marker for affected individuals. Adrenal lesions are common autopsy findings in normal individuals, but do appear to occur more frequently in MEN1, with an incidence of up to 40 per cent. Furthermore, *menin* gene mutations have been demonstrated in individuals with atypical familial endocrine syndromes including phaeochromocytoma. Histology of adrenal lesions associated with MEN1 usually demonstrates nodular hyperplasia and there is no associated excess hormone secretion. Loss of heterozygosity of the MEN1 locus is not found in these lesions and it has been proposed that there may be a circulating adrenal growth factor, possibly secreted by the pancreas, since there is a strong correlation with pancreatic tumours, particularly insulinomas. MEN1-associated adrenal tumours showing loss of heterozygosity for 11q13 have been reported but are very rare. Thyroid disease has been reported in association with MEN1, but does not appear to occur more frequently than in the normal population.

Screening

The screening of first- and second-degree relatives of patients with MEN1 is aimed at early detection of parathyroid, pancreatic, or pituitary lesions in gene carriers, to reduce the associated morbidity. There is no evidence that screening reduces mortality, although the identification of affected individuals in 'malignant' kindreds with aggressive pancreatic disease may allow curative surgery which would be expected to prolong survival. Screening lowers the age of detection of the syndrome by about 20 years.

The most useful screening investigations are a serum calcium, fasting gastrin, and prolactin, although in practice a full gut hormone screen is usually performed. It has been suggested that the most sensitive markers of pancreatic disease are basal and test-meal-stimulated pancreatic polypeptide and gastrin, and basal insulin and proinsulin, identifying lesions at least 3 years before there are any radiological abnormalities. Since pancreatic tumours are the only life-threatening manifestation of the syndrome, such a screening protocol may be warranted. The MEN1 syndrome rarely develops before the age of 5 or after the age of 70, and so screening should be performed annually from 5 to 65, and at longer intervals thereafter. Eighty per cent of affected individuals will have been identified by the fifth decade. Screening of patients with apparently sporadic pancreatic endocrine tumours for evidence of MEN1 is probably justified, especially in those with gastrinomas or insulinomas. There is little evidence to support screening in those with sporadic pituitary tumours. MEN1 is present in 15 per cent of all patients with hyperparathyroidism, but hypercalcaemia may be associated with elevated fasting gastrin and pancreatic polypeptide, and, whereas in those at risk of MEN1 this finding would be highly significant, in those with sporadic hyperparathyroidism this very rarely indicates pancreatic disease, so screening of all patients is not warranted.

Genetic mutation analysis can be used to screen for affected individuals in kindreds where the gene mutation has been identified, but if this is not the case biochemical screening is still needed.

Multiple endocrine neoplasia type 2 (MEN2)

Multiple endocrine neoplasia type 2 is the association of medullary cell carcinoma of the thyroid (MTC) and phaeochromocytoma. The association was first recognized in 1932, but it was not until 1961 that it was noted that the risk of phaeochromocytoma in patients with MTC was increased 14-fold. MEN2 has since been subdivided: in MEN2A, or Sipple's syndrome, parathyroid hyperplasia may occur; MEN2B is associated with mucosal neuromas and marfanoid habitus. In addition, there is a familial form of MTC without other features. Germ line mutations of the *ret* proto-oncogene, a receptor tyrosine kinase, have been identified in all three syndromes. In MEN2A and familial MTC, mutations occur in the extracellular domain, whilst in MEN2B mutation in the tyrosine kinase domain has been demonstrated. The MEN2 phenotypes reflect the expression of *ret* in different tissues. Tumours in affected individuals are heterozygous for the *ret* mutation, and so it is the only known dominantly inherited proto-oncogene. It is likely that activation of *ret* leads to hyperplasia in affected tissues and that a somatic mutation in another oncogene is required for carcinogenesis. Thus loss of heterozygosity for the short arm of chromosome 1 has been described in phaeochromocytomas and MTC associated with MEN2. New mutations are uncommon in MEN2A, and probably account for less than 10 per cent of cases, whereas new mutations account for about 50 per cent of cases with MEN2B. Those patients with MEN2A not identified by screening usually present in the fourth and fifth decades, while those with MEN2B present much earlier due to their characteristic phenotype.

Medullary cell carcinoma of the thyroid (MTC)

MTC is a tumour of the C cells of the thyroid (see Chapter 12.4 and Chapter 12.5), which secrete calcitonin, and this acts as a tumour marker. Twenty-five per cent of cases are familial. The incidence of MTC in MEN2 is probably 100 per cent. Familial MTC alone is the most benign form of MTC, while MTC in association with MEN2B is the most malignant form of the disease. In MEN2, the initial thyroid lesion is C-cell hyperplasia, which has been found as early as the age of 3 years in MEN2A and may be present at birth in MEN2B. Over the subsequent 5 to 10 years, microscopic MTC develops and finally gross tumours become apparent. Metastases are invariably present when tumours are already palpable, but there is speculation that they may occur with clinically occult disease. All forms of hereditary MTC are bilateral, with multifocal tumours, usually occurring at the junction of the upper third and lower two-thirds of the thyroid.

In MEN2A, MEN2B, and familial MTC genotype screening has largely replaced biochemical screening for MTC using pentagastrin-stimulated calcitonin, with over 95 per cent of kindreds with MEN2 and over 80 per cent of those with familial MTC having germline mutations of the *ret* gene, 99 per cent of MEN2 kindreds having a mutation in codon 918. In families with a known *ret* mutation it has been recommended that a positive genotype should result in a total thyroidectomy with lymph node clearance by the age of 2 years in MEN2B kindreds and at age 3 in MEN2A and familial MTC kindreds. In MEN2B MTC has been reported as early as 15 months of age with metastases by the age of 3 years. Prior to thyroidectomy those with a positive genotype should be screened biochemically for phaeochromocytoma, which, if confirmed, should be resected first. In those patients with MTC not identified by screening, thyroidectomy should still be performed, unless distant metastases, usually to lung or liver, are present. It is probable that in all patients with palpable disease, metastases to local lymph nodes will be present, so a central lymph node dissection should also be performed, probably with lateral node sampling to look for further spread. The prognosis is poor in this group, with recurrent disease in about 20 per cent of patients with clinically occult but macroscopic MTC and in over 60 per cent of those with palpable MTC. It is particularly poor in individuals with

MEN2B who present with clinically apparent MTC. Their 10-year survival is about 50 per cent, and death from metastatic disease in the mid-twenties is common.

Phaeochromocytoma

Phaeochromocytoma is familial in 5 per cent of cases, 20 per cent of whom have MEN2. Fifty per cent of individuals with MEN2 develop phaeochromocytoma. About 70 per cent are bilateral, almost all are benign, and they are rarely extra-adrenal. The initial lesion, similar to that in the thyroid, is adrenal medullary hyperplasia, followed by nodule formation and, subsequently, development of multiple, multifocal phaeochromocytomas (see Chapter 15.16.2.4).

Symptoms and biochemical abnormalities are rare during the stage of medullary hyperplasia. MEN2-associated phaeochromocytomas are characterized by excessive adrenaline secretion, so that palpitation and other b-adrenergic symptoms predominate initially, with hypertension a late feature, although often present by the time of diagnosis. A urine adrenaline: noradrenaline ratio of greater than 0.15 in a patient with MEN2 indicates medullary hyperplasia or phaeochromocytoma. The treatment for adrenal medullary hyperplasia or phaeochromocytoma is bilateral adrenalectomy, since the incidence of bilateral disease is high, and the mortality from phaeochromocytoma in MEN2 about 15 per cent, usually due to sudden death. If an adrenal lesion is identified at the same time as MTC, the adrenalectomy should be performed first.

Other features of MEN2A

Parathyroid hyperplasia occurs in up to 80 per cent of patients with MEN2A, but less than 20 per cent have hypercalcaemia, the remainder being identified at the time of thyroidectomy. Parathyroidectomy should be performed in those with hypercalcaemia and in the remaining patients grossly enlarged glands should be removed at the time of thyroidectomy.

Cutaneous lichen amyloidosis, often preceded by intense pruritus, has been described in two kindreds with MEN2A and provides a phenotypic marker for the syndrome.

Other features of MEN2B

The characteristic phenotype of marfanoid habitus and mucosal neuromas (Fig. 4) identifies affected individuals with MEN2B and allows early intervention, since these features usually predate MTC and phaeochromocytoma. Neuromas are commonly ocular and oral, causing whitish-yellow or pink nodules on the anterior aspect of the tongue, lips, and eyelids, with thickening of the mucosa and often eversion of the lower lids. The nasal bridge may be broadened, pedunculated neuromas are found on cheek mucosa, and the corneal nerves are thickened and medullated. Involvement of peripheral motor and sensory nerves can cause a peroneal muscular atrophy type picture. Intestinal ganglioneuromatosis affects about 75 per cent of cases. Neuromas involve the autonomic nerves of both the myenteric and submucosal plexi and can cause poor suckling with failure to thrive, altered bowel habit, recurrent pseudo-obstruction, toxic megacolon, and occasionally dysphagia and vomiting, possibly due to achalasia. Almost all patients have a marfanoid habitus, usually associated with skeletal abnormalities, particularly slipped femoral epiphyses. Delayed puberty is another common feature of the syndrome.

Fig. 4 (a) Characteristic phenotype of MEN2B showing facial appearance. (b) Characteristic phenotype of MEN2B showing mucosal neuromas on tongue.

Screening

Screening to identify affected individuals by genotyping and for early biochemical detection of adrenal disease reduces both morbidity and mortality in MEN2. A positive genotype identifies individuals for thyroidectomy. If a kindred's genotype has not been identified an elevated pentagastrin-stimulated calcitonin but normal basal calcitonin identifies individuals at the stage of C-cell hyperplasia or microscopic MTC. Where it is indicated biochemical screening in MEN2A or familial MTC should commence at age 3 and be performed annually. In MEN2B individuals with the phenotype do not need screening, as they should have a thyroidectomy before the age of 2 years. Urinary metanephrines and catecholamines identify at least 95 per cent of phaeochromocytomas. Urinary vanillylmandelic acid levels are less useful, being associated with a high number of false positives and negatives. MIBG and/or CT scanning or measurement of plasma catecholamines may identify phaeochromocytomas missed by urinary assays. Serum calcium should be measured at the annual screening to identify overt hyperparathyroidism. In families where a mutation has been characterized, affected individuals can be identified by mutation screening.

Other syndromes associated with endocrine neoplasia

There are other syndromes which overlap with the multiple endocrine neoplasia syndromes and gene mutation analysis can identify those which are true variants of MEN1 or MEN2. Phaeochromocytomas may be associated with pancreatic islet-cell tumours alone, or in combination with other syndromes: von Hippel–Lindau syndrome is associated with a high incidence of phaeochromocytomas, islet-cell tumours, cerebellar haemangioblastomas, retinal angiomas, and renal cell carcinoma; neurofibromatosis type I (von Recklinghausen's syndrome) is often associated with phaeochromocytoma and, rarely, with duodenal somatostatinoma and medullary thyroid carcinoma; and phaeochromocytoma may be associated with prolactinoma as a mixed MEN syndrome.

Further reading

Ajani JA, Carrasco CH, Charnsangavej C, Samaan NA, Levin B, Wallace S (1988). Islet cell tumors metastatic to the liver: effective palliation by sequential hepatic artery embolization. *Annals of Internal Medicine* **108**, 340–4.

Arnold J, O'Grady J, Bird G, Calne R, Williams R (1989). Liver transplantation for primary and secondary hepatic apudomas. *British Journal of Surgery* **76**, 248–9.

Chandrasekharappa SC *et al*. (1997). Positional cloning of the gene for multiple endocrine neoplasia-type 1. *Science* **276**, 404–7.

Gorden P, Comi RJ, Maton PN, Go VL (1989). NIH conference. Somatostatin and somatostatin analogue (SMS 201–995) in treatment of hormone-secreting tumors of the pituitary and gastrointestinal tract and non-neoplastic diseases of the gut. *Annals of Internal Medicine* **110**, 35–50.

Grauer A, Raue F, Gagel RF (1990). Changing concepts in the management of hereditary and sporadic medullary thyroid carcinoma. *Endocrinology and Metabolism Clinics of North America* **19**, 613–35.

Hofstra RM *et al*. (1994). A mutation in the RET proto-oncogene associated with multiple endocrine neoplasia type 2B and sporadic medullary thyroid carcinoma. *Nature* **367**, 375–6.

Jensen RT, ed. (1989). Gastrointestinal endocrinology. *Gastroenterology Clinics of North America* **18**, 671–931.

Krejs G, ed. (1987). Gastrointestinal endocrine tumours. *American Journal of Medicine* **82** (Suppl. 5B), 1–3.

Moertel CG, Lefkopoulo M, Lipsitz S, Hahn RG, Klaassen D (1992). Streptozocin-doxorubicin, streptozocin-fluorouracil or chlorozotocin in the treatment of advanced islet-cell carcinoma. *New England*

Journal of Medicine **326**, 519–23.

Mulligan LM *et al*. (1993). Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A. *Nature* **363**, 458–60.

Oberg K, ed. (1991). Recent advances in diagnosis and treatment of neuroendocrine gut and pancreatic tumours. *Acta Oncologica* **28**, 301–449.

Rosch T *et al*. (1992). Localization of pancreatic endocrine tumors by endoscopic ultrasonography. *New England Journal of Medicine* **326**, 1721–6.

Rossi P *et al*. (1989). Endocrine tumors of the pancreas. *Radiologic Clinics of North America* **27**, 129–61.

Sheppard BC, Norton JA, Doppman JL, Maton PN, Gardner JD, Jensen RT (1989). Management of islet cell tumors in patients with multiple endocrine neoplasia: a prospective study. *Surgery* **106**, 1108–17.

Skogseid B *et al*. (1991). Multiple endocrine neoplasia type 1: a 10-year prospective screening study in four kindreds. *Journal of Clinical Endocrinology and Metabolism* **73**, 281–7.

Thakker RV and Ponder BA (1988). Multiple endocrine neoplasia. *Baillières Clinical Endocrinology and Metabolism* **2**, 1031–67.

Vasen HF *et al*. (1992). The natural course of multiple endocrine neoplasia type IIb. A study of 18 cases. *Archives of Internal Medicine* **152**, 1250–2.

Vinayek R, Frucht H, Chiang HC, Maton PN, Gardner JD, Jensen RT (1990). Zollinger-Ellison syndrome. Recent advances in the management of the gastrinoma. *Gastroenterology Clinics of North America* **19**, 197–217.

Wynick D and Bloom SR (1991). The use of the long-acting somatostatin analog octreotide in the treatment of gut neuroendocrine tumors. *Journal of Clinical Endocrinology and Metabolism* **73**, 1–3.

12.11.1

Diabetes

Gareth Williams

Diabetes mellitus can be defined as a state of chronic hyperglycaemia sufficient to cause long-term damage to specific tissues, notably the retina, kidney, nerves, and arteries.

This functional label gives little insight into the long and colourful history of this disease, its clinical and scientific importance, or its immense personal and socio-economic impact. Diabetes was recognized in antiquity, and its clinical features (with empirical treatment guidelines) were recorded over 3500 years ago in the Egyptian 'Ebers' papyrus. Our understanding of the disease has advanced greatly, especially during the last two decades, but many aspects of its management remain imperfect.

Diabetes is and will remain a threat to global health. In the United Kingdom, at least 2 per cent of the population (over 1 million people) are diabetic and the disease absorbs 5 to 10 per cent of the total health budget. Worldwide, diabetes probably affects 150 million people and its prevalence is predicted to double by 2015.

Diagnosis of diabetes

Blood glucose concentrations are is normally tightly regulated: fasting values lie between 3.5 and 5.5 mmol/l and even large carbohydrate loads do not raise the concentration above 8 mmol/l. It is logical to define diabetes by the blood glucose concentrations which cause the chronic complications of the disease but the choice of the diagnostic glucose levels has been contentious (and has stirred up much passion among epidemiologists). One difficulty is that some diabetic complications show a 'threshold' effect with the risk rising above a cutoff level (for example fasting plasma glucose of 6 to 7 mmol/l for retinopathy), whereas macrovascular disease (atheroma) does not (see later). Another problem is that even the current criteria are not self-consistent: for example, 30 per cent of newly diagnosed type 2 diabetic patients will have a 'normal' fasting plasma glucose, while another 30 per cent will have a 'normal' value 2 h after an oral glucose tolerance test.

The current diagnostic criteria for diabetes and other hyperglycaemic states (Fig. 1) have been approved by the World Health Organization and most national diabetes associations. All values refer to venous plasma glucose concentrations:

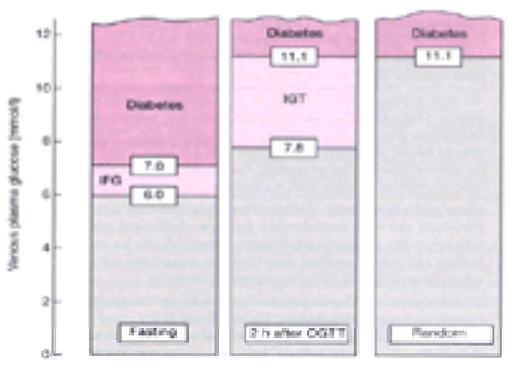

Fig. 1 Diagnostic thresholds for diabetes, impaired glucose tolerance (IGT), and impaired fasting glucose (IFG). From the Expert Committee on the Diagnosis and classification of diabetes mellitus (1997). *Diabetes Care* **20**, 1183–97. For conversion to mg/dl, multiply values in mmol/l by 18.

1. Diabetes mellitus: fasting glucose more than 7.0 mmol/l and/or a value exceeding 11.1 mmol/l, either at 2 h during an oral glucose tolerance test or in a random sample. The corresponding levels in non-SI units are 126 and 200 mg/dl respectively. The diagnostic fasting glucose level was lowered from the previous value of 7.8 mmol/l to reflect more accurately the risk of developing diabetic retinopathy.
2. Impaired glucose tolerance: fasting glucose less than 7.0 mmol/l and 2-h oral glucose tolerance test value between 7.8 and 11.1 mmol/l.
3. Impaired fasting glucose: fasting glucose 6.1 to 6.9 mmol/l (110 to 124 mg/dl).

Impaired glucose tolerance (**IGT**) and the recently distinguished impaired fasting glucose (**IFG**) are intermediate categories of hyperglycaemia that carry definite risks and so require follow-up and risk-factor management (see below). They are often transient stages and overlap to some extent: about one-third of subjects with

impaired fasting glucose also have impaired glucose tolerance, while one-quarter of those with impaired glucose tolerance also show impaired fasting glucose.

The new criteria put much emphasis on the fasting plasma glucose concentration. However, the time-consuming oral glucose tolerance test is still required in some cases with borderline fasting hyperglycaemia, because the 2-h oral glucose tolerance test value in such patients may be high enough to put them at risk of microvascular complications. Moreover, the oral glucose tolerance test remains the only way to define impaired glucose tolerance.

Practical screening and diagnostic procedures

Figure 2 shows an algorithmic approach to screening for and diagnosis of diabetes and its associated hyperglycaemic states. Certain high-risk groups need to be actively screened for type 2 diabetes, which may be present (and causing complications) for several years before it is noticed. These include subjects predisposed to develop type 2 diabetes through genotype and/or phenotype, those affected by diabetogenic conditions such as pregnancy, endocrine disorders or certain drugs, and those with other cardiovascular risk factors in whom hyperglycaemia must not be missed.

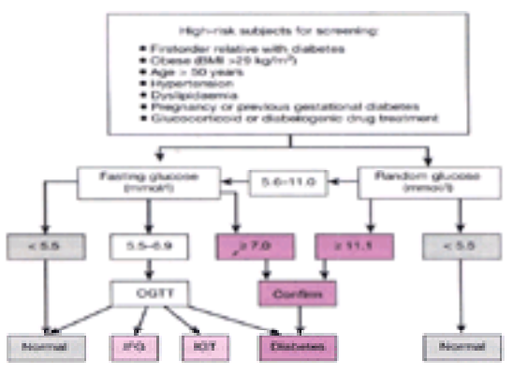

Fig. 2 Screening algorithm for diagnosing diabetes, impaired glucose tolerance, and impaired fasting glucose. All glucose values relate to venous plasma (mmol/l). Adapted from data in Shaw JE and Zimmet P (2000). Do we know how to diagnose diabetes and do we need to screen for the disease? In: Gill GV, Pickup JC, Williams G, eds. *Difficult diabetes*, pp 3–21. Blackwell Science, Oxford.

Diabetes is not a trivial diagnosis, and certain practical points must be carefully observed:

1. Glucose should be measured in venous plasma using a quality-controlled laboratory method. Capillary (finger-prick) samples contain higher glucose levels than venous blood, from which glucose has been extracted by the tissue bed; whole-blood glucose levels are lower than in plasma, because red cells actively metabolize glucose and so contain only low concentrations. These differences may reach 0.5 to 1.0 mmol/l. Portable glucose meters correlate well with laboratory glucose methods, but because of potential technical errors they should not be used to make or refute the diagnosis.
2. An oral glucose tolerance test is indicated for borderline hyperglycaemia (Fig. 2). After an overnight fast, the subject drinks 75 g of anhydrous glucose dissolved in 250 ml water (or 394 ml Lucozade); venous blood is sampled at baseline and 2 h later. Food intake should be normal during the preceding few days: poor nutrition can cause delayed hyperglycaemia with a raised 2-h value (the 'lag' curve).
3. Abnormal values need confirmation. Postchallenge glucose levels in particular can vary considerably. Because of this and possible laboratory error, the diagnosis of diabetes should ideally be verified using a further sample on another day. Obvious exceptions are grossly raised values in seriously ill patients, especially children.
4. Diabetes must not be diagnosed from indirect measures of hyperglycaemia such as raised glycated haemoglobin (HbA_{1c}) or fructosamine levels in blood, or glycosuria. HbA_{1c} and fructosamine reflect average blood glucose concentrations, but the measurements are not sufficiently sensitive or standardized (several different methods are in use) to be used diagnostically. Glycosuria depends on the renal threshold for glucose reabsorption and its presence does not necessarily indicate hyperglycaemia; conversely, glucose may be absent from the urine in diabetic subjects who also have a high renal threshold. However, abnormal results with any of these tests suggest diabetes and indicate the need for formal blood glucose screening.

Impaired glucose tolerance (IGT)

Impaired glucose tolerance is a metastable state: within 5 years, about 25 per cent of subjects with impaired glucose tolerance deteriorate into type 2 diabetes, while a further 25 per cent revert to normoglycaemia. The degree of hyperglycaemia in impaired glucose tolerance falls, by definition, below the threshold for microvascular complications but is enough to predispose to cardiovascular disease (see later).

Subjects found to have impaired glucose tolerance must be followed up because of the hazards of both diabetes and macrovascular disease. An oral glucose tolerance test should be repeated at least annually, and dietary and lifestyle advice given to decrease metabolic and cardiovascular risks; increased physical activity, a low-fat diet and weight loss convincingly reduce both the progression to type 2 diabetes and cardiovascular events. Risk factors such as smoking, hypertension, dyslipidaemia, and obesity should be managed actively. Specific antihyperglycaemic treatments (with metformin or thiazolidinediones, to improve insulin sensitivity) are currently being evaluated.

Impaired fasting glucose (IFG)

As with impaired glucose tolerance, the 5-year risk of progressing to type 2 diabetes appears to be about 25 per cent, and IFG predisposes to cardiovascular disease. Long-term monitoring and management should therefore be as for impaired glucose tolerance.

Metabolic basis of diabetes

Diabetes is due to inadequate production of insulin and/or 'resistance' to the glucose-lowering and other actions of insulin. To put this in context, key aspects of normal metabolism will be briefly reviewed.

The islets of Langerhans

There are about 1 million islets of Langerhans in the normal adult: insulin is produced by the b (B) cells, which make up the bulky core of each islet; b cells also synthesize the peptide known as amylin or islet-associated polypeptide. The other islet cell types, mostly surrounding the b-cell core, are the a (A) cells that produce glucagon, the † (D) cells that produce somatostatin, and the PP cells that synthesize pancreatic polypeptide. All islet cells are derived embryologically from the buds of gut endoderm which also give rise to the exocrine pancreatic tissue.

The various islet cell types communicate with each other through the hormones they secrete into the islet's rich capillary plexus and probably by paracrine effects on adjacent cells; these interactions presumably regulate hormone secretion. Insulin inhibits release of glucagon, while glucagon powerfully stimulates insulin secretion—an action exploited in the testing of b-cell reserve (see below). Somatostatin suppresses the secretion of insulin and glucagon. Amylin can inhibit insulin secretion under experimental conditions but its physiological role is uncertain. Amylin also polymerizes outside the b cell to produce fibrils of amyloid material, which have been implicated in the progressive b-cell damage of type 2 diabetes.

Insulin

Insulin has a molecular weight of 5800 Da; it is made up of an A chain (21 amino acid residues) and a B chain (30 residues), joined covalently by two disulphide bridges. The precursor molecule, proinsulin, consists of the A and B chains linked end-to-end through a connecting (C) peptide which is cleaved off during insulin processing. In the circulation, insulin is monomeric but in crystals and more concentrated solutions (for example in the insulin vial and the subcutaneous injection

site), six insulin molecules self-associate around a central Zn^{2+} ion. Self-association influences the pharmacokinetic properties of subcutaneously injected insulin: the rate-limiting dissociation of hexamers into monomers slows the absorption of even 'fast-acting' insulin.

Insulin regulates metabolism in birds, fish, and reptiles as well as mammals, and its structure is remarkably well conserved across the phyla. Three species of insulin are used therapeutically; the human sequence differs from porcine at a single residue (B30) and from bovine at two others. These differences affect the pharmacokinetic and immunogenic characteristics of the insulins (see below). The physicochemical behaviour of insulin has been successfully manipulated in synthetic 'designer' insulins that have improved absorption profiles: modification of the C terminus of the B chain, a region crucial for self-association, produces analogues that remain in the monomeric state and are therefore absorbed faster than the native soluble insulin (see below).

Insulin biosynthesis and processing

Insulin is a product of the preproinsulin (*INS*) gene, located on the short arm of chromosome 11, whose coding region contains three exons. Translation of *INS* mRNA in the rough endoplasmic reticulum produces preproinsulin, which is successively cleaved during its passage through the Golgi vesicles and secretory vesicles to yield first proinsulin and finally insulin and C peptide. Proinsulin is converted into insulin by the proteolytic excision of the C-peptide chain; the two intermediate cleavage products (with either end of the C peptide remaining attached to insulin) are called 'split products' of proinsulin. Normally, almost all proinsulin is processed through this 'regulated' pathway to yield equimolar amounts of insulin and C peptide. However, a 'constitutive' pathway may predominate in dysfunctional b cells (for example in type 2 diabetes and insulinoma), when processing is not complete and large quantities of proinsulin and split products may be released into the circulation.

C peptide is generally regarded as an inert byproduct of insulin production. However, its structure is also conserved across species and it may have vasoactive and other properties.

Insulinopathies are point mutations in the *INS* gene which either produce a mutant insulin (for example, insulin Chicago: a phenylalanine for leucine substitution at residue B25) or interfere with one of the cleavage sites of proinsulin so that the mutant split product cannot be further processed (for example, proinsulin Tokyo). These conditions are inherited as autosomal dominant traits; circulating insulin-like or proinsulin-like immunoreactivities may be extremely high but glucose intolerance is often surprisingly mild.

Insulin secretion

Glucose is the main insulin secretagogue; this action of glucose is modulated by other ingested nutrients, by hormones released by the islets and the gut, and by the autonomic innervation of the islet. The process gives insight into the mode of action of the sulphonylureas and related drugs, and the cause of maturity onset diabetes of the young (see below).

Glucose-stimulated insulin secretion

The amount of insulin released by the normal b cell is tightly coupled to blood glucose levels and begins to increase immediately when blood glucose rises. The ability of the b cell to sense ambient glucose levels accurately and rapidly depends on the glucose transporter isoform GLUT-2 and the glucose metabolizing enzyme glucokinase, while insulin release hinges on depolarization of the b-cell membrane which is controlled by a specific ion channel, the ATP-sensitive K^+ channel. The characteristics of GLUT-2 allow glucose at physiological concentrations to freely enter the b cell, where it is immediately converted by glucokinase into glucose-6-phosphate—the point of entry into the glycolytic pathway which ultimately yields ATP; ATP production within the b cell is therefore proportionate to extracellular glucose.

ATP binds to and closes the ATP-dependent K^+ channel; when open, this channel allows K^+ ions to leave the b cell along their concentration gradient and thus helps to maintain the negative charge inside the b-cell membrane. ATP-induced closure of the channel therefore causes K^+ ions to accumulate within the cell and the membrane to depolarize, which triggers the opening of specific (voltage-gated) Ca^{2+} channels in the membrane. Ca^{2+} ions then flood into the b cell from the outside and activate the contractile proteins which drag the secretory vesicles containing insulin and C peptide to the cell surface. Here, the vesicles fuse with the cell membrane and release their contents into the extracellular space (exocytosis), from where insulin and C peptide enter the islet capillaries.

Other factors affecting insulin secretion

Sulphonylureas induce insulin secretion by closing the same ATP-sensitive K^+ channel as glucose: they bind to a specific sulphonylurea receptor (SUR-1) linked to the K^+ channel protein (called Kir 6.2). Repaglinide also closes this K^+ channel, but binds to a different site from the sulphonylureas. By contrast, diazoxide locks the channel open, hyperpolarizing the b-cell membrane and inhibiting insulin secretion—hence its use in treating insulinoma.

Glucagon and glucagon-like peptide-1 7–36 amide (**GLP-1**; a gut peptide with insulin secretagogue (incretin) actions), both stimulate insulin secretion by raising cytosolic Ca^{2+} concentrations; binding to their receptors increases generation of cyclic AMP which blocks removal of Ca^{2+} into intracellular organelles. Conversely, somatostatin and possibly amylin act to decrease production of cyclic AMP and inhibit insulin secretion. Arginine stimulates insulin secretion, possibly by depolarizing the b-cell membrane as it enters the cell (it is cationic).

The autonomic nervous system is an important modulator of insulin secretion; it is stimulated by the parasympathetic (vagal) outflow and inhibited by the sympathetic. Vagal stimulation is mediated by acetylcholine acting via muscarinic receptors, while the inhibitory sympathetic neurotransmitter is noradrenaline, interacting with a_2 adrenoceptors.

Diseases due to defects in insulin secretion in perhaps 10 per cent of pedigrees is due to mutations affecting glucokinase (glucokinase-dependent maturity onset diabetes of the young). These impair ATP production from glucose, blunting the insulin response of the b cell to rising glucose and resulting in variable hyperglycaemia (see below). By contrast, familial neonatal hyperinsulinism is caused by mutations in *SUR-1* that close the ATP-sensitive K^+ channel, leading to sustained insulin secretion and severe hypoglycaemia soon after birth.

Normal pattern of insulin secretion

Insulin concentrations in peripheral blood show basal levels of about 10 mU/l that tend to fall overnight, on which are superimposed prandial peaks reaching 80 to 100 mU/l, roughly proportionate to the amount eaten. The prandial peaks are elicited by the insulin secretagogue effects of glucose and other nutrients, augmented by incretin gut peptides (such as GLP-1) and the vagal outflow (the early cephalic phase of insulin release).

Very frequent sampling (every minute) shows that 'basal' insulin secretion is in fact pulsatile, with clear but low-amplitude peaks every 9 to 13 min. This may help to keep the target tissues sensitive to insulin; loss of this pulsatility is an early sign of b-cell dysfunction in type 2 diabetes. An acute insulin secretagogue challenge (for example an intravenous glucose bolus) induces a sharp 'first-phase' insulin peak, loss of which is another early abnormality in type 2 diabetes.

The insulin response elicited by eating is larger than when an equivalent nutrient load is given intravenously. This is because glucose entering the gut stimulates neuroendocrine cells in the gut wall to release 'incretin' hormones which act on the b cell to enhance insulin secretion (the 'enteroinsular axis': see Chapter 14.8). An important incretin appears to be GLP-1, a product of alternative processing of the preproglucagon gene (glucagon itself is not produced, in contrast to the islet a cell). GLP-1 released from the small intestine augments insulin release in the presence of glucose, an effect being explored in the treatment of type 2 diabetes.

Peripheral insulin levels are lower than those in the portal vein, into which the islets drain, because up to 30 per cent of insulin is removed on its first pass through the liver—one of the main targets for insulin action. The kidney also actively clears and degrades insulin; the circulating half-life is only a few minutes.

C peptide provides a robust measure of residual b-cell function, because it is cleared more slowly than insulin and its plasma concentrations are therefore more stable. C peptide is generally measured after intense b-cell stimulation with the powerful insulin secretagogue glucagon; alternatives are a heavy oral load of carbohydrate, or simply the measurement of 24-h secretion of C peptide in urine (it is cleared largely intact through the kidneys). In normal subjects and most with type 2 diabetes, peak C-peptide concentrations at 6 min after 1 mg of intravenous glucagon are 1 to 4 nmol/l, whereas type 1 diabetic individuals are typically

'C-peptide negative', with peak levels less than 0.6 nmol/l.

The insulin receptor and signal transduction

The insulin receptor belongs to the family that also includes the insulin-like growth factor 1 receptor. Insulin receptors are found in the obvious insulin target tissues (fat, liver, and skeletal muscle) but also in unexpected sites, such as the brain and gonads, in which glucose uptake does not depend on insulin.

The insulin receptor is a 400-kDa heterotetramer composed of two a and two b glycoprotein subunits, interconnected by disulphide bridges (Fig. 3). Both a and b subunits are encoded within a complex gene (22 exons) on chromosome 19q. The a subunit (135 kDa) lies entirely extracellularly, while the b subunit (95 kDa) spans the cell membrane and extends into the cytoplasm. Part of the intracytoplasmic tail functions as a tyrosine kinase, attaching phosphate groups from ATP to tyrosine residues elsewhere on the receptor (autophosphorylation) and on other intracellular proteins. This tyrosine kinase activity is essential for insulin signalling and for insulin to exert its many effects on its target tissues. Insulin binds to a site on the extracellular a subunits, and binding triggers a conformational change in the receptor which activates the tyrosine kinase domain of the b subunits.

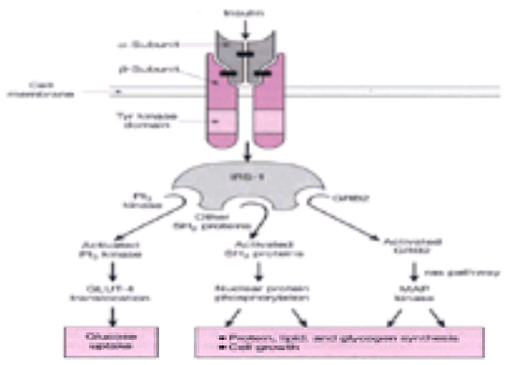

Fig. 3 The insulin receptor and signal transduction pathways within insulin's target cells. Binding of insulin to the extracellular a subunits of the receptor activates the tyrosine (Tyr) kinase domain of the intracellular b subunit. This phosphorylates insulin receptor substrate 1 (IRS-1) and the related IRS-2, which in turn phosphorylate other signalling proteins. These activated proteins then trigger other reactions that result in the biological actions of insulin, including enhanced glucose uptake, anabolic effects, and cell growth. PI_3 kinase, phosphatidylinositol 3 kinase; MAP kinase, mitogen-activated protein kinase.

Postreceptor mechanisms

The activated receptor phosphorylates tyrosine residues on specific intracellular proteins which initiate the signal transduction pathway within the target cell. One protein is the large (130 kDa) insulin receptor substrate 1, which has numerous phosphorylation sites that can accept other proteins possessing specific 'SH2' domains. 'Docking' and activation (phosphorylation) of these proteins by insulin receptor substrate 1 begins a cascade of intracellular reactions that lead ultimately to the effects of insulin on glucose, lipid, and protein metabolism and its many other actions. The details remain elusive, but the mitogen-activated protein kinase pathway is involved in glycogen synthesis, while the phosphatidylinositol 3 kinase pathway mediates glucose transporter translocation (see Fig. 3).

Receptor turnover

Receptors that bind insulin are 'internalized', i.e. taken up into the target cell by an invagination of the cell membrane that is coated with the protein clathrin. Bound insulin is degraded in the lysosomes, while most of the insulin receptors are carried back to the cell surface and reinserted into the membrane. The density of receptors on the cell surface is therefore a dynamic quantity, regulated partly by new receptor synthesis and partly by receptor recycling, which in turn is determined by insulin binding. Prolonged exposure to high insulin concentrations increases the proportion of internalized receptors and so decreases the density of receptors available on the cell surface. This 'downregulation' of receptors reduces the sensitivity of the target tissue to insulin.

Disorders due to insulin receptor defects

Many mutations have now been described in the insulin receptor, including point mutations that cause single-residue substitutions or truncation of the a or b subunits. Mutations affecting the tyrosine kinase domain interfere with insulin signalling and can lead to severe insulin resistance and glucose intolerance, sometimes with serious mental and physical abnormalities (for example in 'leprechaunism') which confirm the importance of insulin in fetal development.

Antibodies may develop against the insulin receptor and usually cause insulin resistance with variable hyperglycaemia (the 'type B' insulin resistance syndrome); rarely, hypoglycaemia results from antibodies that activate the receptor (analogous to thyrotoxicosis induced by antibodies to the thyroid-stimulating hormone receptor in Graves' disease).

Metabolic actions of insulin

Insulin functions as an anabolic hormone, favouring the uptake, utilization, and storage of glucose, the storage of lipids as triglyceride, and preventing the breakdown of protein.

Effects on carbohydrate metabolism

Insulin lowers blood glucose in two main ways (Fig. 4). At low basal concentrations (overnight and between meals) it shuts off the production of glucose by the liver, which is the main determinant of fasting glycaemia. Hepatic glucose output is fuelled by both glycogen breakdown (glycogenolysis) and gluconeogenesis (i.e. glucose synthesis from substrates including lactate, glycerol, and alanine and other amino acids); the rate-limiting enzymes for these processes are powerfully inhibited by insulin. Conversely, insulin stimulates glycogen synthesis.

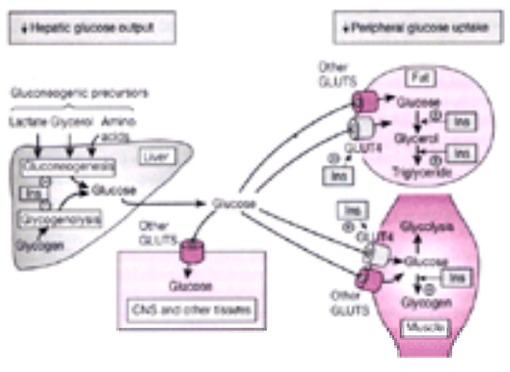

Fig. 4 Effects of insulin on glucose homeostasis. Insulin inhibits gluconeogenesis and glycogen breakdown in the liver, thus decreasing hepatic glucose output. Blood glucose is also lowered by increased glucose uptake into fat and skeletal muscle, mediated by the insulin-stimulated glucose transporter, GLUT-4. (Non-insulin mediated glucose uptake is effected by other GLUT proteins.)

At higher concentrations, such as after meals, insulin also stimulates glucose transport into skeletal muscle (where it is utilized to provide energy via glycolysis, or stored as glycogen) and into fat (where it is used to synthesize triglycerides). In both these tissues, insulin enhances glucose uptake through a specific glucose transporter protein, GLUT-4 (Fig. 4). Insulin causes GLUT-4 units to be translocated rapidly to the cell surface and inserted into the membrane: there, GLUT-4 units act as hydrophilic pores through which glucose can cross the otherwise impermeable membrane into the cell, following its concentration gradient. Insulin also stimulates GLUT-4 synthesis. Overall, insulin acting via GLUT-4 can increase glucose uptake into muscle and fat by up to 40-fold over the basal 'non-insulin mediated' glucose uptake. Non-insulin mediated glucose uptake occurs through other glucose transporter isoforms that operate in the absence of insulin, notably GLUT-1 in peripheral tissues and erythrocytes and GLUT-3 in brain.

Effects on lipid metabolism

Insulin inhibits triglyceride breakdown (lipolysis), while promoting its synthesis (lipogenesis). Lipolytic enzymes that split triglyceride into glycerol and free fatty acids are powerfully inhibited by insulin, even at low basal insulin concentrations. Profound insulin deficiency, such as in untreated type 1 diabetes, is therefore required before uncontrolled lipolysis occurs and generates enough free fatty acids to cause ketoacidosis (see below).

Effects on protein metabolism

Insulin inhibits protein catabolism and thus reduces the generation of amino acids which can act as gluconeogenic precursors to enhance glucose production by the liver and kidney. Insulin also promotes protein synthesis and cellular and tissue growth.

Other actions of insulin

These include vasodilatation, mediated by endothelial production of nitric oxide; growth and differentiation of the fetal nervous system; and enhanced tubular reabsorption of Na^+ ions by the kidneys.

Measurements of insulin action

Glucose lowering is the most easily tested biological action of insulin, and forms the basis for most measurements of 'insulin resistance'. Several methods are used in the research setting; theoretically, the simplest could be used in clinical diabetes care, to identify patients with marked insulin resistance who might benefit particularly from insulin-sensitizing drugs such as the thiazolidinediones.

1. *Homeostatic model assessment* (HOMA) is an index derived by mathematical modelling of the relationship between the fasting glucose and insulin concentrations: with decreasing insulin sensitivity, insulin secretion increases in an attempt to maintain euglycaemia, resulting in compensatory hyperinsulinaemia. Homeostatic model assessment can be performed on a single fasting blood sample and compares well with the insulin–glucose clamp.
2. *Insulin–glucose (hyperinsulinaemic–euglycaemic) clamp*. Insulin is infused intravenously to achieve constant high concentrations and a separate infusion of glucose is adjusted to maintain blood glucose 'clamped' at a normal value. The more glucose required, the greater is the insulin sensitivity. The clamp is generally regarded as the 'gold standard' method but demands blood glucose measurements every few minutes and takes some hours to perform.
3. *Intravenous glucose tolerance test*. An intravenous glucose bolus stimulates insulin release, and mathematical modelling of the relationship between the insulin peak and the decay in blood glucose levels can yield indices of both insulin secretion and insulin sensitivity.

Insulin resistance

Insulin resistance (or insensitivity) is a poorly defined term signifying decreased biological activity of insulin, and which is usually equated with impaired glucose-lowering.

There is no universal normal range for insulin sensitivity, because the ability of insulin to lower glucose varies considerably between and within individuals—it is influenced, for example, by levels of physical activity and fitness. Subjects with 'insulin-resistant' conditions such as type 2 diabetes or essential hypertension commonly show reductions of 40 to 60 per cent in glucose disposal (measured by the clamp technique), as compared with matched healthy controls, yet many apparently normal subjects also have comparable decreases in insulin sensitivity. There is no argument about extreme examples of insulin resistance: in some patients with so-called 'leprechaunism', over 20 000 U/day of insulin have failed to control hyperglycaemia and ketosis. A working definition of clinically relevant insulin resistance in insulin-treated diabetic patients is a daily requirement of more than 1.5 U/kg.

Causes of insulin resistance

Inherited causes

Inherited causes include the very rare mutations affecting the insulin receptor or postreceptor signalling pathways which can lead to extreme insulin resistance; milder polygenic defects contribute to the insulin resistance of type 2 diabetes (see below). Insulin receptor mutations cause clinically distinct syndromes, often with acanthosis nigricans and, in women, features of polycystic ovary disease and masculinization; hyperglycaemia is variable. Specific syndromes include the speculatively named 'leprechaunism' and various inherited lipodystrophies in which fat is lost from subcutaneous and other depots in defined but unexplained anatomical patterns (see Chapter 10.5). Recently, mutations affecting the *PPAR-g* gene (the target for the thiazolidinedione drugs; see below) have been shown to modify insulin sensitivity.

Obesity

Obesity induces insulin resistance, especially in skeletal muscle, while weight loss can improve insulin sensitivity in the obese. Insulin resistance is particularly associated with truncal (central) obesity, where fat is deposited in and around the abdomen; both the subcutaneous and intra-abdominal (visceral) fat depots have been implicated to various degrees that may reflect ethnic and other differences.

It is still not clear how an increased fat mass can decrease whole-body insulin sensitivity, but circulating fat-derived products are presumed to be responsible. Intra-abdominal fat depots would secrete potentially diabetogenic mediators into the portal circulation—where they would be delivered directly to the liver—and this may explain the association of visceral adiposity with insulin resistance. Possible candidates include free fatty acids and the cytokine tumour necrosis factor-a; both are secreted by adipocytes and, under experimental conditions at least, interfere with aspects of insulin action. Levels of free fatty acids are raised in obese subjects, apparently because lipolysis is enhanced, and free fatty acids may cause hyperglycaemia by competing with glucose metabolism in liver and muscle. In liver, free fatty acids enhance gluconeogenesis by stimulating the rate-limiting enzyme pyruvate carboxylase and so increase hepatic glucose production. In muscle, free fatty acids inhibit glycolysis at the level of phosphofructokinase and glucose oxidation via pyruvate dehydrogenase, causing a decrease in glucose utilization and a secondary reduction in glucose uptake (the 'glucose–fatty acid' or Randle cycle). *In vitro*, tumour necrosis factor-a inhibits the tyrosine kinase activity of the insulin receptor that is crucial for insulin signalling. Production of tumour necrosis factor-a by adipose tissue is increased in obesity but its role as a mediator of insulin resistance in human obesity is uncertain. Recently, a novel adipocyte product, adiponectin, has been shown to enhance insulin sensitivity in rodents; intriguingly, circulating adiponectin concentrations are decreased in human obesity.

Physical inactivity strongly predisposes to obesity and also promotes insulin resistance which can be reversed by regular exercise. The mechanism is unknown but physical training is known to stimulate translocation of GLUT-4 glucose transporters to the surface of muscle cells independently of insulin.

Other acquired causes

There are several other acquired causes of insulin resistance. Intrauterine growth retardation may contribute (see the 'Barker–Hales' hypothesis below). Physiological states of insulin resistance, due to the appropriate oversecretion of the counter-regulatory hormones whose metabolic actions oppose those of insulin, are puberty and pregnancy (see gestational diabetes, Chapter 13.10). Endocrine diseases that induce insulin resistance and can cause glucose intolerance and overt diabetes through excessive production of anti-insulin hormones include acromegaly (prevalence of diabetes and impaired glucose tolerance each around 25 per cent),

Cushing's disease (diabetes around 30 per cent), thyrotoxicosis, and the very rare glucagonoma (diabetes in more than 90 per cent of cases). In these disorders, diabetes is mostly non-ketotic, although insulin may be needed to control hyperglycaemia.

Intercurrent illnesses, for example myocardial infarction, stroke, or severe infections, induce the secretion of counter-regulatory stress hormones that can cause marked insulin resistance—insulin-treated diabetic patients may need twice their usual insulin dosages during such episodes. Many drugs decrease insulin sensitivity, including glucocorticoids, b_2 adrenoceptor agonists (ritodrine, salbutamol), and certain oral contraceptive pills containing high-dose oestrogen or levonorgestrel; glucocorticoid-induced hyperglycaemia commonly requires insulin treatment.

The 'type B' insulin resistance syndrome is due to the development of autoantibodies against the insulin receptor which interfere with insulin binding and/or signalling. Most patients are young women, usually with pre-existing autoimmune diseases such as lupus erythematosus, and masculinization often occurs. Acquired lipodystrophies are also associated with insulin resistance, sometimes severe. 'Immune insulin resistance' describes insulin-treated patients with very high insulin requirements (sometimes several thousand U/day) because of high titres of insulin-binding antibodies that bind and inactivate administered insulin. This has become very rare since the introduction of highly purified human-sequence insulin preparations with low immunogenicity (see below).

Metabolic and clinical features of insulin resistance

The metabolic disturbance due to insulin-resistant syndromes ranges from subclinical glucose intolerance to severely symptomatic hyperglycaemia, sometimes with ketosis. A crucial determinant is the capacity of the individual's b cells to secrete insulin in response to the rises in blood glucose that are due to impaired insulin action. The resulting hyperinsulinaemia is extremely variable, with plasma insulin levels ranging from twice normal in many obese subjects to 500 times normal in patients with defects of insulin receptors. Near-normoglycaemia can be maintained as long as hyperinsulinaemia can compensate for the underlying defect in insulin signalling; diabetes occurs when b-cell failure supervenes and insulin secretion falls below a critical level. In the total absence of functional insulin receptors (for example in 'leprechaunism'), massive endogenous hyperinsulinaemia or administration of industrial insulin dosages cannot prevent severe diabetes, although very high insulin concentrations may exert some metabolic actions through 'cross-talk' with the insulin-like growth factor-1 receptor.

Acanthosis nigricans, a characteristic skin manifestation of severe insulin resistance, may be due to high insulin concentrations activating growth factor receptors (perhaps the insulin-like growth factor-1 receptor) that drive the proliferation of keratinocytes and melanocytes. Hyperplasia of these cells leads to a velvety thickening and variable darkening of the skin, especially in the axillae, groin, and nape of the neck (see Chapter 23.1). Widespread acanthosis nigricans can also accompany gut tumours, which may also secrete dermal growth factors.

Increased androgen concentrations may lead to hirsutism and occasionally virilization in women with severe insulin resistance; high insulin concentrations may stimulate androgen production by the ovaries, which often show a polycystic appearance. Insulin resistance is a feature of polycystic ovary syndrome, especially in obese patients (Chapter 12.8.1). Enhancing insulin sensitivity through weight loss or treatment with metformin or the thiazolidinediones can decrease androgen levels and improve hirsutism and menstrual dysfunction.

The 'insulin resistance syndrome', or 'metabolic syndrome X'

This term identifies the co-occurrence of insulin resistance and glucose intolerance (ranging from mild to overt type 2 diabetes), with truncal obesity, dyslipidaemia (raised triglycerides and a high low density lipoprotein:high-density lipoprotein ratio), and hypertension (see Fig. 5).

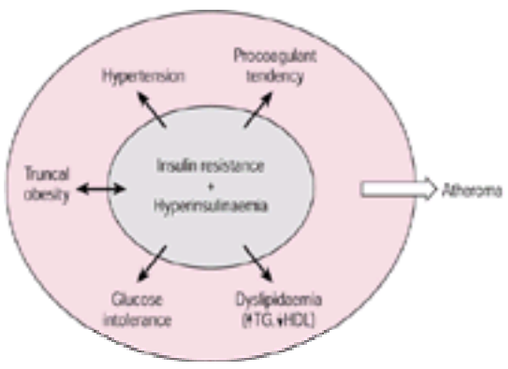

Fig. 5 Syndrome X, a constellation of atherogenic risk factors which may each be related to insulin resistance and/or the hyperinsulinaemia that accompanies insulin-resistant states. -TG, hypertriglyceridaemia; ¯HDL, reduced high-density lipoprotein cholesterol.

These abnormalities are all common in most westernized populations, and it is still not clear whether or not this constellation of cardiovascular risk factors represents a genuine syndrome with a common underlying cause. Reaven and others have argued that insulin resistance is the central abnormality, and that the key features can be explained either by loss of specific actions of insulin or by the effects of the compensatory hyperinsulinaemia on organs that remain relatively insulin sensitive. For example, raised insulin levels could contribute to hypertension by enhancing retention of Na^+ by the kidney; conversely, blood pressure could also be raised through loss of the direct vasodilator action of insulin. The pattern of abnormalities would therefore require 'insulin resistance' to affect certain tissues and specific actions of insulin but not others. Other proatherogenic defects identified in subjects with various features of syndrome X include increased coagulability of the blood (for example increased levels of plasminogen activator inhibitor 1) and impaired endothelial-mediated vasodilatation. The relationship of these abnormalities to insulin resistance is uncertain. Obesity, dyslipidaemia, hypertension, and glucose intolerance are all independent cardiovascular risk factors; any possible proatherogenic role of hyperinsulinaemia *per se* remains controversial.

The aetiology of syndrome X is unresolved. Adiposity, insulin sensitivity, and blood pressure show variable strengths of familial transmission that differ between populations and generally suggest polygenic inheritance of minor genes. On the other hand, Barker and Hales have suggested that fetal malnutrition programmes insulin resistance, hypertension, and dyslipidaemia in middle to late adult life. The underlying mechanisms remain elusive. Because obesity leads to insulin resistance and glucose intolerance, dyslipidaemia, hypertension, and atheroma, weight gain in middle age may be particularly hazardous in subjects who were underweight at birth.

Syndrome X is important clinically because it predisposes to atheroma formation and substantially increases the risk of dying prematurely from myocardial infarction or stroke. Treatment of syndrome X is currently based on correcting any factors (for example type 2 diabetes, hypertension, and dyslipidaemia) present in the individual patient. Lifestyle and dietary modification that achieves weight loss can improve most aspects of the syndrome, and specific insulin-sensitizing drugs such as the thiazolidinediones (see later) may prove beneficial.

Types and classification of diabetes mellitus

The current World Health Organization classification is based on aetiology (Table 1). Type 1 and type 2 diabetes together account for 90 to 95 per cent of cases and will be described in detail.

Type 1 diabetes

'Type 1' diabetes—now preferred to 'insulin-dependent' diabetes—is due to autoimmune killing of the b cells (the so-called 'type-1 process'). A similar clinical picture of insulin dependence can be caused by other forms of severe pancreatic damage.

Epidemiology and demographic features

Type 1 diabetes is considerably rarer than type 2, accounting for between 5 and 15 per cent of all diabetes and 30 to 50 per cent of insulin-treated cases in various populations. It appears predominantly in childhood, with a peak age at presentation of about 11 years in girls and 14 years in boys—hence the old description of 'juvenile-onset'. However, it can develop at any age, and about 5 per cent of newly diagnosed Caucasian diabetic patients over 65 years have type 1 diabetes.

The prevalence of type 1 diabetes varies considerably throughout the world. Incidence is highest in northern European countries (about 30 to 35 cases per 100 000 children per year in Scotland and Finland) and declines progressively towards the equator; there are some isolated 'hot spots' such as Sardinia, where the incidence is as high as in Finland. High susceptibility is found in European populations throughout the world, while African and Oriental populations are relatively spared (incidences of less than 1 per 100 000 per year). Superimposed on this geographical variation are time-related changes in incidence that hint at the importance of the environment in causing the disease. Type 1 diabetes presents more frequently during the winter months, particularly in children aged 10 to 14 years. In several countries (for example Sweden, Scotland, and Poland), there have been sharp 30 to 50 per cent increases in incidence over 10- to 20-year periods, although the explanation and significance of these secular trends are not clear.

Susceptibility to type 1 diabetes shows no gender bias.

Aetiology

Type 1 diabetes is an autoimmune, predominantly T-cell-mediated process that selectively destroys the b cells. Susceptibility is multifactorial, resulting from the impact of environmental agents in a genetically disadvantaged subject. Of these two components, the environment appears more important; genetic factors explain only 30 to 40 per cent of total susceptibility. Immunogenetic aspects are discussed in detail in Chapter 12.11.2.

Genetic factors

Over 20 genetic loci are associated with type 1 diabetes, although only two (*IDDM1* and *IDDM2*) account for most of the genetic predisposition.

IDDM1 lies within the major histocompatibility complex region on chromosome 6, that encodes several proteins intimately involved in immune responses. Of particular importance is the *DQB1* gene; this encodes the DQB1 peptide chain, which forms part of the cleft in the surface of the HLA class II molecule that is crucial in presenting peptide fragments of antigen to the T-helper lymphocyte. Changes in the structure of the DQB1 peptide could therefore influence the coupling between the class II molecule–peptide complex and the T-lymphocyte receptor, and thus modulate the immune response against the (auto)antigenic peptide. Specific *DQB1* polymorphisms have been shown to predispose to type 1 diabetes (for example DQB1*0302), whereas others (such as DQB1*0602) are protective—at least in certain racial groups. The relationships of these polymorphisms to the long-recognized influences of the DR3 and DR4 class II antigens (which increase several-fold the risk of type 1 diabetes) and of the protective DR2 are discussed further in Chapter 12.11.2.

IDDM2 corresponds to the insulin gene (*INS*), whose uniqueness as a b-cell product makes it an obvious candidate gene. As the insulin coding sequence is unchanged in type 1 diabetes, diabetogenic polymorphisms might affect the level of expression of the *INS* gene: variants which enhanced insulin expression could promote b-cell damage, because 'resting' the b cell (for example by giving exogenous insulin) in type 1 diabetic animals improves b-cell survival, in parallel with reduced expression of the putative autoantigen GAD65 (a specific isoform of the enzyme glutamic acid decarboxylase (**GAD**), expressed in the b-cell membrane).

Environmental factors

Viruses have long been popular candidates as a 'trigger' for diabetes. Some (for example mumps, coxsackie, cytomegalovirus, and rubella) infect the pancreas but normally damage the entire gland, particularly the exocrine tissue, rather than causing selective b-cell injury. Certain viruses target the b cell in animals (for example the Kilham rat virus) and can cause insulin-dependent diabetes, either though their direct cytolytic effects or by provoking a type 1-like autoimmune process. Important contenders in humans are coxsackie viruses (especially B4), rubella, and retroviruses.

Serological studies indicate that recent coxsackie B infections are relatively common among newly diagnosed patients with type 1 diabetes; these could represent the final insult in the disease's long natural history. Coxsackie viruses capable of damaging rodent b cells have also been isolated post-mortem from the islets of type 1 diabetic subjects. About 20 per cent of children who survive intrauterine rubella infection develop type 1 diabetes, with typical autoimmune markers. Retrovirus particles and RNA have recently been identified in b cells from type 1 diabetic patients, and two-thirds of newly diagnosed cases are reported to have antibodies against the protein encoded by the retroviral RNA.

Viruses could trigger or maintain autoimmune b-cell damage in various ways. Acute or persistent viral infection of b cells could release b-cell antigens that are normally sequestered beyond the reach of the immune cells. Certain viral proteins may elicit an immune response which crossreacts with specific b-cell antigens that happen to be similar ('molecular mimicry'): for example, peptide sequences of the P2-C capsid protein of coxsackie B viruses may crossreact with GAD65 in the b-cell membrane.

Other environmental factors are suggested to include bovine serum albumin from cows' milk and various toxins. Bovine serum albumin contains a peptide sequence that may crossreact with a b-cell surface protein (see below); this was suggested as an explaination for an apparent excess risk of type 1 diabetes among children fed with cows' milk in the neonatal period, although a protective effect for breast feeding remains controversial. Various toxins selectively damage b cells, including streptozotocin, a nitrosourea used to induce experimental diabetes in rodents. Related nitrosamine compounds have been blamed for the higher risk of type 1 diabetes in the children of women who eat fermented smoked mutton (a traditional delicacy in Iceland).

Autoimmune features

Type 1 diabetes has strong associations with endocrine and other autoimmune diseases, including Schmidt's syndrome (with hypothyroidism and adrenocortical failure) and the polyglandular autoimmune deficiency syndromes.

Most b-cell damage is probably inflicted by T lymphocytes. 'Insulitis'—infiltration of the islets with immune cells, mostly cytotoxic/suppressor (CD8+) T lymphocytes—is a pathognomic feature of the disease, and circulating T-helper lymphocytes can be identified that react against b-cell antigens including GAD65.

Various circulating autoantibodies also occur. Some target antigens unique to the b cell, while other autoantigens are shared by other islet cell types. Notable b-cell selective autoantibodies are those that recognize GAD65, a heat-shock protein (hsp60), and insulin itself. GAD catalyses the conversion of glutamic acid to g-aminobutyric acid (**GABA**), whose role in the b cell is uncertain. Studies in rodents with type 1 diabetes suggest that the level of GAD65 expression influences the intensity of the autoimmune attack on the b cells. The GAD67 isoform of the enzyme is also expressed in the central nervous system, and autoimmune damage of GABAergic neurones is presumed to explain the association of type 1 diabetes with the rare 'stiff man' syndrome (Section 24).

GAD65 antibodies are present in 70 to 90 per cent of newly diagnosed type 1 patients, and insulin autoantibodies in 40 to 70 per cent. Islet cell antibodies and islet cell surface antibodies are present in 80 to 90 per cent and 30 to 60 per cent respectively of newly diagnosed patients. Islet cell antibodies and islet cell surface antibodies recognize uncharacterized antigens that are common to b and non-b cells in the islets (which share the same embryological origin). These antibodies cannot explain the selective destruction of b cells, although some islet cell surface antibodies are complement-fixing and may target the b cell preferentially; islet cell antibodies and islet cell surface antibodies could be raised secondarily to b-cell damage.

High titres of each of these classes of antibodies have some value in predicting diabetes in high-risk individuals—a risk of about 50 per cent for GAD65 antibodies in first-order relatives of subjects with type 1 diabetes. However, they are clearly not the immediate cause of the disease: they are found in many subjects who do not go on to develop it, including a few per cent of the normal population. These antibodies are therefore general markers of autoimmunity against the b cell, rather than evidence of b-cell destruction, which is primarily cell mediated. Titres of all these antibodies tend to be high at presentation and (according to prospective studies of high-risk subjects) during the months leading up to this. Thereafter, antibody levels decline progressively and may even become undetectable, possibly through dwindling of the antigen load that perpetuates autoimmunity as any remaining b cells disappear.

Natural history of type 1 diabetes

b-cell damage might be initiated by direct viral attack, environmental toxins, and/or a primary immune attack against specific b-cell antigens such as GAD65, perhaps via molecular mimicry. T-helper lymphocytes (CD4+) are activated by b-cell antigens presented together with diabetogenic class II antigens by antigen-presenting cells (macrophages) and perhaps by b cells themselves ('aberrant' class II antigen expression may be induced in b cells by certain cytokines generated during the autoimmune process). Activated T-helper cells produce cytokines that attract T and B lymphocytes and encourage them to proliferate in the islet, leading to insulitis. B lymphocytes might then damage b cells by producing antibodies against released b-cell antigens, while cytotoxic (CD8+) T lymphocytes directly attack b cells carrying the target autoantigens. Insulitis is a patchy and unpredictable process that might flare up after encounters with new environmental triggers such as viral infections, but which can also fade and abort for unknown reasons.

Several years of progressive autoimmune damage usually precede the clinical onset of diabetes. This long prediabetic phase is asymptomatic, although careful testing (for example with the intravenous glucose tolerance test) reveals loss of the first phase, then increasingly obvious disturbances of insulin and C-peptide secretion, and eventually glucose intolerance. Finally, when the b-cell mass has been eroded to a critical level (probably 5 to 10 per cent of normal), falling insulin secretion can no longer restrain hyperglycaemia and clinical diabetes develops.

Residual b-cell mass is variable at presentation of type 1 diabetes: some newly diagnosed type 1 patients are C-peptide positive, and b-cell secretion may improve temporarily during the 'honeymoon period' that can follow the lowering of blood glucose when insulin treatment is started (see below). With continuing b-cell destruction, endogenous insulin production declines progressively, and more than 90 per cent of type 1 patients become permanently C-peptide negative within 5 years of presentation. Ultimately, insulitis burns itself out and the immune cells retreat, leaving islet remnants that are devoid of b cells but which still contain intact a, †, and PP cells.

The protracted prediabetic phase provides an opportunity to prevent subjects with active insulitis from developing clinical disease. A combination of autoantibody titres and genetic markers (HLA haplotypes) can be used to predict the chances of the disease developing in high-risk subjects, such as the siblings of children with type 1 diabetes; various immunosuppressive and immunomodulatory treatments are currently undergoing clinical trials.

Metabolic disturbances of type 1 diabetes

In untreated type 1 diabetes, insulin concentrations are generally 10 to 50 per cent of non-diabetic levels in the face of hyperglycaemia which would normally greatly increase insulin secretion. Such severe deficiency cannot sustain the normal anabolic effects of insulin and leads to runaway catabolism in carbohydrate, fat, and protein metabolism. Each of these processes accelerates hyperglycaemia, while the oxidation of excess free fatty acids generated by triglyceride breakdown can result in diabetic ketoacidosis.

Carbohydrate metabolism

Basal hyperglycaemia is due mainly to unrestrained production of glucose by the liver and is accentuated after eating by the failure of glucose to be cleared peripherally (see Fig. 4). Hepatic glucose output is boosted, especially by increased gluconeogenesis: the normal inhibition of the process by insulin is lost, while the supply of gluconeogenic precursors (glycerol from lipolysis, amino acids such as alanine from protein breakdown) is increased. Increased gluconeogenesis in the kidney may also contribute. Postprandial glucose uptake into muscle and fat, mediated by insulin and GLUT-4, is greatly decreased; this is partly offset by increased non-insulin dependent glucose uptake into peripheral tissues, via glucose transporters that do not require insulin.

The overall result is hyperglycaemia, commonly in the range of 15 to 25 mmol/l and higher after meals. Glucose concentrations of over 40 mmol/l are not uncommon during intercurrent illness and especially when insulin treatment is omitted or not increased sufficiently.

Fat metabolism

Lipolysis is stimulated by severe insulin deficiency, generating glycerol (a gluconeogenic precursor) and free fatty acids, the substrate for ketone formation. Ketogenesis is particularly enhanced by concomitant glucagon excess (see below). Mobilization of body fat contributes to weight loss in untreated type 1 diabetes.

Protein metabolism

Loss of the net anabolic effect of insulin encourages catabolism of proteins (primarily through the proteasome-mediated pathway), thus generating amino acids including gluconeogenic precursors such as alanine and glutamine. Muscle wasting may be prominent.

Role of counter-regulatory hormones

The effects of hypoinsulinaemia are compounded by the counter-regulatory hormones which are secreted in excess in response to stress (for example infections, myocardial infarction, trauma, surgery) and when circulating volume falls (for example in hyperglycaemic dehydrated patients). Insulin deficiency also leads to increased glucagon secretion because insulin normally inhibits the a cells.

Glucagon increases hepatic glucose production, both by driving glycogen breakdown and by increasing uptake of glucogenic amino acids by the liver and enhancing gluconeogenesis. It also stimulates ketogenesis by increasing entry of free fatty acids (as their fatty acyl-CoA derivatives) into liver mitochondria (see Fig. 8 below). Glucagon excess is an important factor that promotes diabetic ketoacidosis, acting synergistically with insulin deficiency (see below).

Cortisol and catecholamines enhance gluconeogenesis. Cortisol, catecholamines, and growth hormone oppose the lipogenic action of insulin and favour lipolysis, in the presence of hypoinsulinaemia. Cortisol is a powerful inducer of proteolysis, whereas growth hormone co-operates with insulin to stimulate protein synthesis.

Clinical features of type 1 diabetes

The classical presentation of untreated or poorly controlled type 1 diabetes reflects the consequences of catabolism and hyperglycaemia (Table 2). These features usually develop progressively and quite rapidly over a period of a few days to a few weeks.

Diuresis is due mainly to the osmotic effect of glucose remaining in the renal tubule, when its concentration exceeds the reabsorption threshold for glucose (corresponding generally to plasma glucose levels of about 10 mmol/l). The osmotic loads of urinary ketones and of electrolytes that are obligatorily lost with glucose also contribute. Urine output may reach several litres per day, causing polyuria, nocturia, and in children, enuresis.

Thirst generally parallels urine output and can be very intense; it is characteristically made worse by sugar-rich drinks. Taking water to bed at night is a useful sign of pathological thirst. A high fluid intake is an important homeostatic response to diuresis, and patients unable to drink (for example through nausea in ketoacidosis) can rapidly become dehydrated and hypovolaemic.

Weight loss, due to loss of fat and muscle and later to dehydration, can be dramatic and reach several kilograms over a few weeks. The energy deficit caused by catabolism and urinary losses of glucose can amount to several hundred calories per day. Appetite is often increased; the mechanism in humans is not known; falls in circulating leptin and insulin, both of which act on the central nervous system to inhibit feeding, are probably responsible for hyperphagia in diabetic rodents.

Systemic symptoms include tiredness, malaise, lack of energy, and muscular weakness.

Blurred vision is commonly due to changes in the shape of the lens due to osmotic shifts, typically causing longsightedness. Rarely, acute 'snowflake' cataracts develop because of reversible refractile changes, rather than the permanent denaturation of lens proteins in senile cataract.

Infections are often present because hyperglycaemia predisposes to infections and also because infections stimulate the secretion of stress hormones. Genital

candida infections, causing recurrent pruritus vulvae in women and balanitis in men, are frequent and should always prompt testing for diabetes. Pyogenic skin infections and urinary tract infections, sometimes complicated by severe renal damage, are also common, and certain rare infections have a particular predilection for diabetic people (see below).

Diabetic ketoacidosis presents with hyperglycaemic symptoms, which are usually severe, together with nausea and vomiting, acidotic (Kussmaul) breathing, the smell of acetone on the breath, and, especially in children, altered mood and clouding of consciousness that may progress to coma. Diabetic ketoacidosis is described in detail later.

Unlike type 2 diabetes, which is often present for several years before diagnosis, hyperglycaemia in newly presenting type 1 patients develops too acutely for chronic diabetic complications to appear. Because obvious symptoms appear quickly, very few cases are picked up fortuitously, although doctors who have forgotten to think of diabetes in their differential diagnosis of weight loss or hyperventilation may be surprised when hyperglycaemia is detected by routine screening.

Prognosis of type 1 diabetes

Before the introduction of insulin during the early 1920s, type 1 diabetes was invariably fatal, usually within months. With various semistarvation diets, hyperglycaemic symptoms could be improved somewhat and life extended by a few miserable months.

With modern insulin treatment, type 1 diabetic patients can be rescued from diabetic ketoacidosis, although one-third of deaths in diabetic children and young adults are still due to metabolic emergencies, notably ketoacidosis. The main threat to survival with type 1 diabetes is now chronic tissue damage, particularly renal failure from nephropathy, and vascular disease, notably myocardial infarction and stroke. Throughout adult life, the overall risk of dying within 10 years is about fourfold higher for patients with type 1 diabetes than for their non-diabetic peers.

There is encouraging evidence from Europe and the United States that the outlook for type 1 diabetes has improved over the last 10 to 20 years, with definite declines in the incidence of microvascular complications and extended survival—at least in countries able to afford effective diabetes care. This is partly attributable to tighter control of hyperglycaemia, which can reduce by 30 to 40 per cent the risks of nephropathy and retinopathy developing or progressing to a clinically significant degree (see below). Other measures have undoubtedly contributed, including better treatment of raised blood pressure and blood lipids.

Tragically, however, in many parts of the world patients with type 1 diabetes still die today as they did a century ago, simply because insulin is not available.

Type 2 diabetes

Type 2 diabetes is a heterogeneous condition, diagnosed empirically by the absence of features suggesting type 1 diabetes (Table 2) and of the many other conditions that cause hyperglycaemia (Table 1). Diagnostic accuracy may depend on the thoroughness of investigation: for example, up to 10 per cent of subjects with late onset diabetes show evidence of autoimmune b-cell damage and thus probably have slowly evolving type 1 diabetes.

The term 'type 2' replaces 'non-insulin dependent' which was both clumsy and confusing: many type 2 patients require insulin to control hyperglycaemia.

Epidemiology and demographic features

Type 2 diabetes accounts for 85 to 90 per cent of diabetes worldwide and is very common. It affects about 2 per cent of the Caucasian populations in most westernized countries, the prevalence rising with age to 10 per cent of those over 70 years. It is substantially commoner in certain immigrant populations in more affluent countries, for example 5 per cent or more of young and middle-aged adults in some Asian or Afro-Caribbean groups in the United Kingdom.

Type 2 diabetes is most commonly diagnosed in those over 40 years of age and the incidence rises to a peak at 60 to 65 years. However, much younger people are now presenting with type 2 diabetes, following the rapid rise in childhood obesity. Up to one-third of North Americans diagnosed as diabetic under 20 years of age have type 2 diabetes, Afro-Caribbeans and Mexicans being at particular risk. Maturity onset diabetes of the young, which commonly presents before 25 years of age, is now classified separately (see below for more details).

The prevalence of type 2 diabetes shows striking geographical variation—entirely different from that of type 1—and ranges from less than 1 per cent in rural China to 50 per cent in the Pima Indians of New Mexico. Prevalence is also rising rapidly and, worldwide, will double within 10 to 15 years. This pandemic is largely explicable by westernization (so-called 'Cocacolonization'), and is following in the wake of the obesity that is spreading throughout the world. The Pima Indians illustrate this process especially vividly, although most developed and developing countries are showing the same phenomenon albeit more slowly. Diabetes was rare while the Pimas led a frugal existence in desert conditions and were lean and physically active. Following urban resettlement and exposure to overnutrition and inactivity, there were rapid increases in the prevalence of obesity (currently 80 per cent of adult Pimas have a body mass index of over 30 kg/m^2) and later of type 2 diabetes. The Pimas' spectacular susceptibility to obesity and diabetes may be explained by the selection of 'thrifty' genes, i.e. those encouraging the storage of excess energy as fat, which would favour survival in their original harsh environment. In a setting of readily available food, cars, and television, the same thrifty genes would lead to obesity and ultimately diabetes (see below).

There is a 3:2 male preponderance among subjects with type 2 diabetes.

Aetiology

Type 2 diabetes is due to the combination of insulin resistance and b-cell failure, the latter preventing sufficient insulin secretion to overcome the insulin resistance. These two components vary in importance between different individuals, who may be clinically quite similar, and each has numerous possible causes. Susceptibility is determined by the interactions between genes and environment. The steeply rising prevalence of type 2 diabetes suggests that diabetogenic genes are common and are now enjoying an unparalleled opportunity to express themselves through the global spread of 'Cocacolonization' and obesity.

Genetic factors

Overall genetic susceptibility to type 2 diabetes is probably 60 to 90 per cent, rather less than was previously deduced from twin studies. Generally, transmission does not follow simple mendelian rules, and this polygenic pattern presumably reflects the inheritance of a critical mass of minor diabetogenic minor polymorphisms which interfere with insulin action and/or insulin secretion. Having a first-order relative with the disease increases by fivefold an individual's chances of developing it, representing a lifetime risk in Caucasians of about 40 per cent.

Genetic factors may partly determine both insulin resistance and b-cell failure, although to different degrees in different individuals. Insulin sensitivity appears to be largely genetically determined, at least in some populations. Genes leading to insulin resistance could encode regulators of energy balance, metabolic enzymes, or the proteins that signal insulin action, and presumably include 'thrifty' genes favouring fat deposition. Specific genes have not been firmly incriminated: particular polymorphisms associated with type 2 diabetes have been reported (for example the b_3 adrenoceptor implicated in lipolysis and energy expenditure, the insulin signalling protein insulin receptor substrate-1, and glycogen synthetase) but subsequently questioned. Mutations affecting the insulin receptor and glucose transporters do not appear to cause common type 2 diabetes, although mutations of insulin receptors can lead to severe insulin resistance (see above).

Diabetogenic genes leading to inadequate production of insulin could decrease insulin secretion in response to glucose, or impair b-cell viability. Mutations affecting glucokinase cause maturity onset diabetes of the young 2 (see below) and rare cases (less than 1 per cent) of apparently typical type 2 diabetes; otherwise, mutations in the known components of the b cell's glucose-sensing or insulin-secreting machinery are not responsible for common type 2 diabetes.

Environmental factors

These clearly play a critical part, because obesity and type 2 diabetes are spreading too rapidly to be explicable by changes in the genome; environmental factors are also important in practice because they may be modified to treat and prevent the disease. Known environmental diabetogenic factors mostly induce insulin resistance (for example obesity, pregnancy, intercurrent illness, certain drugs). Hyperglycaemia *per se* can both impair insulin sensitivity and inhibit insulin secretion (so-called

'glucotoxicity').

Specific risk factors for type 2 diabetes

Obesity, itself determined by both genes and environment, is one of the most important risk factors, apparently due to aggravation of insulin resistance (see above). The diabetogenic properties of excess fat depend not only on its bulk but also on its anatomical distribution and the time of life at which it is laid down. The risks of developing type 2 diabetes begin to increase steeply once the body mass index exceeds 28 kg/m ^{2}; some studies estimate the risk at a body mass index over 35 kg/m^2 to be 80-fold higher than for individuals with a body mass index of less than 22 kg/m ^{2}—a lifetime risk of about 50 per cent. Fat in the truncal (central) distribution is more diabetogenic than that deposited around the hips and thighs, and the visceral (intra-abdominal) depot is strongly associated with insulin resistance. Increasing adiposity after the early twenties, especially around the waist, aggravates the risk of high body mass index.

Physical inactivity, especially from the twenties onwards, is an independent predictor of diabetes in middle age, the risk increasing by about threefold for sedentary people as compared with regular athletes. This is due to worsening insulin resistance, which can be improved by physical training.

The still controversial Barker and Hales hypothesis suggests that poor fetal growth can 'programme' enduring metabolic and vascular abnormalities that are manifested in adult life, especially in people who were underweight at birth but then become obese. These abnormalities include key features of 'syndrome X' (hyperglycaemia, hypertension, dyslipidaemia), resulting in atheroma formation, myocardial infarction and stroke (see above). Evidence, mainly from animals, suggests that maternal and therefore fetal malnutrition during a critical early phase of fetal development can reduce b-cell mass and permanently impair insulin secretory reserve; deficiencies of sulphur-containing amino acids may be responsible in experimental animals but the relevance to humans is unknown. Other studies suggest that insulin sensitivity may also be reduced into adult life.

b-Cell failure in type 2 diabetes

b-Cell failure is an obligatory defect in the pathogenesis of type 2 diabetes: near normoglycaemia can be maintained even in severe insulin resistance (due for example to mutations in the insulin receptor), as long as the b cell can respond to the challenge and secrete enough insulin to overcome the resistance.

Subtle abnormalities of insulin secretion, including loss of the physiological pulses and of the first-phase response to intravenous glucose injection, are seen in normoglycaemic subjects who later develop the disease. These defects presumably indicate that the b cell is already stressed in trying to produce enough insulin to overcome insulin resistance. Normoglycaemic first-order relatives of type 2 diabetic subjects also show loss of pulsatility of insulin secretion which might indicate an inherited tendency to b-cell failure.

The mechanism of b-cell failure in human type 2 diabetes is not known. Histologically, the islets in type 2 diabetes show no features of type 1 autoimmune insulitis, and b-cell mass is not so dramatically reduced. Animal models of the disease suggest various causes, including synchronized b-cell apoptosis (possibly mediated by nitric oxide) in the Zucker diabetic fatty rat, and the deposition of amyloid fibrils (see above) in the rhesus monkey. Amyloid deposits are also prominent in the islets of some type 2 diabetic patients but may merely be due to dysfunctional b-cell hypersecretion rather than the cause of b-cell damage. Once hyperglycaemia is established, 'glucotoxicity' *per se* may further worsen both insulin secretion and insulin resistance.

In established type 2 diabetes, insulin secretion is unequivocally subnormal and tends to decline progressively with time, as illustrated by the long-term follow-up data from the United Kingdom Prospective Diabetes Study. Initially, plasma insulin levels may be higher than in non-diabetic subjects but are still inappropriately low, as the normal pancreas would produce much higher insulin concentrations in response to diabetic levels of blood glucose. Conventional radioimmunoassays may overestimate insulin levels in type 2 diabetic patients because of crossreaction with incompletely processed insulin precursors (proinsulin and its split products) released by the 'constitutive' pathway which operates in the malfunctioning b cell (see above). Many type 2 patients ultimately need insulin replacement; this indicates relatively severe insulin deficiency, although still not as profound as in type 1 diabetes. Some type 2 patients who require insulin early have autoimmune markers characteristic of type 1 diabetes, suggesting that they in fact have an indolent variant of type 1 diabetes.

Natural history

Longitudinal and cross-sectional studies indicate that insulin resistance develops first and that compensatory increases in insulin secretion can initially maintain near-normoglycaemia. Worsening insulin resistance is thought to drive the b cells towards maximal insulin output, a metastable stage that probably corresponds to impaired glucose tolerance (see above). Rescue is still possible if insulin resistance is decreased, for example through weight loss or insulin-sensitizing drugs: about 25 per cent of subjects with impaired glucose tolerance return to normoglycaemia within 5 years. However, if insulin resistance persists or worsens, the b cells fail and insulin production falls. At this point, the brake limiting hyperglycaemia is released and blood glucose rises into the diabetic range. The bell-shaped response of insulin secretion, initially increasing to compensate but ultimately failing, has been termed the 'Starling curve' of the b cells because it recalls the classical plot of cardiac output against preload in heart failure.

In common type 2 diabetes, these events usually take many years, and significant hyperglycaemia may have been present for several years at the time of diagnosis. The whole process can be greatly accelerated by acute increases in insulin resistance induced, for example, by steroid treatment or pregnancy.

Metabolic disturbances in type 2 diabetes

Hyperglycaemia is the most obvious abnormality, the extreme case being the hyperosmolar non-ketotic state. Lipid metabolism is also disturbed but true ketoacidosis occurs only exceptionally and is usually provoked by intercurrent events such as infections or myocardial infarction.

Blood glucose concentrations are raised both in the basal (fasting) state and after eating. This reflects the impairment of insulin action in both liver and skeletal muscle, where insulin respectively shuts off hepatic glucose production and stimulates glucose uptake after meals. Hepatic glucose output is increased, due mainly to unsuppressed gluconeogenesis, and this is largely responsible for hyperglycaemia overnight and before meals. In muscle, GLUT-4 activity and glycogen synthesis are especially decreased; this reduces insulin-stimulated glucose uptake into muscle after meals, although basal glucose uptake (non-insulin mediated glucose uptake; see above) is higher than in normal subjects because of the mass-action effect of hyperglycaemia. The degree of hyperglycaemia varies widely: many patients have fasting plasma glucose levels of 8 to 13 mmol/l with postprandial peaks of up to 20 mmol/l, while values exceeding 60 mmol/l are not uncommon in the hyperosmolar non-ketotic state.

Insulin deficiency is less profound than in type 1 diabetes, so mobilization of triglyceride (loss of body fat, ketoacidosis) and catabolism of protein (muscle breakdown) are not usually pronounced. Diabetic ketoacidosis may develop in patients with apparently typical type 2 diabetes who can subsequently be controlled by oral hypoglycaemic agents rather than insulin. Diabetic ketoacidosis is usually precipitated by severe intercurrent illness (for example myocardial infarction, stroke, or pneumonia) in which excessive secretion of counter-regulatory stress hormones exacerbates the metabolic disturbance caused by relative insulin deficiency.

Clinical features

Many cases present with classical symptoms of osmotic diuresis, blurred vision due to hyperglycaemia-related refractive changes in the lens, and genital candidiasis (Table 2).

Weight loss may occur but is generally less dramatic than with newly presenting type 1 diabetes, and may not be obvious because many type 2 patients—over two-thirds in the United Kingdom—are obese. Rapid or severe weight loss in patients who otherwise appear to have type 2 diabetes should be regarded with suspicion as it may point to an early need for insulin replacement (and possibly type 1 diabetes itself) or to coexisting illness: a well-recognized but unexplained association with recent onset type 2 diabetes is carcinoma of the pancreas.

The hyperosmolar non-ketotic state can present with confusion or coma (see below); as mentioned above, diabetic ketoacidosis is rare.

Chronic diabetic complications may be a presenting feature, because hyperglycaemia severe enough to cause tissue damage may already have been present for several years. Extrapolating the numbers of microaneurysms (which only develop at diabetic glucose concentrations) in type 2 patients at various intervals after

diagnosis suggests that significant hyperglycaemia is present for an average of 5 to 7 years before diagnosis. Common problems are arterial disease (myocardial infarction, stroke, and peripheral vascular disease), cataract—which are common in the older population—and retinopathy, especially maculopathy, which can damage central vision.

Increasing numbers of diabetics are detected by screening, either in high-risk groups such as the obese and those with cardiovascular disease, or at routine health checks. Many of these are nominally asymptomatic but will admit to symptoms such as nocturia or perineal irritation if asked directly.

Prognosis of type 2 diabetes

A long-held and prevalent misconception is that type 2 diabetes is 'mild'. Some patients do have relatively unexciting or asymptomatic hyperglycaemia but this can still be enough to cause complications which wreck the patient's life just as much as in type 1 diabetes. Moreover, hyperglycaemia can be as hard to control (even with insulin) as in type 1 patients.

Overall, life expectancy is shortened by up to a quarter in patients with type 2 diabetes presenting in their forties, with vascular disease (myocardial infarction and stroke) being the main cause of premature death. Renal failure from diabetic nephropathy is becoming more common in type 2 patients as their survival from vascular complications improves, and the disease is now the most frequent pathology among people waiting for renal replacement therapy in the United States and some European countries.

Type 2 diabetes is therefore an important threat to the patient's health and survival, and must be taken seriously by patients and their medical attendants, even if the blood glucose concentrations are not dramatically raised. Accordingly, treatment guidelines for the disease are rigorous (Table 3).

Maturity onset diabetes of the young (MODY)

In 1974, Tattersall described a rare familial form of non-insulin dependent diabetes that he distinguished from the generality of cases by its early age of onset, autosomal dominant inheritance, and apparently low risk of microvascular complications. MODY is now known to differ fundamentally from type 2 diabetes in its aetiology and is classified separately as type 3A. It probably accounts for about 1 per cent of non-insulin dependent diabetes and is diagnosed strictly by:

1. Early onset: diabetes is diagnosed before 25 years in the subject and in at least one other family member. Some cases with glucokinase mutations present before 5 years of age.
2. Absence of features of type 1 diabetes, with C-peptide positivity and no requirement for insulin within 5 years of diagnosis.
3. Autosomal dominant inheritance across at least three generations.

MODY is due to failure of the b cell to secrete enough insulin to maintain normoglycaemia: insulin is released in response to glucose but in amounts consistently lower than in non-diabetic subjects. This is explained by failure of the b-cell's glucose-sensing apparatus, which depends on the integrated operation of GLUT-2, glucokinase, and the downstream enzymes that generate ATP from glycolysis of glucose. In contrast to common type 2 diabetes, insulin sensitivity is normal.

Various molecular lesions have now been identified. The first were mutations in the glucokinase gene (on chromosome 7p), which interfere with the enzyme's ability to phosphorylate glucose—the first and rate-limiting step in glycolysis. Glucokinase-dependent MODY (now termed MODY 2) only accounts for about 10 per cent of cases. It is characterized by very early onset hyperglycaemia which is mild and worsens slowly, taking decades to reach truly diabetic levels; it follows that microvascular complications are late to develop.

Other forms of maturity onset diabetes of the young are MODY 1, due to mutations in hepatic nuclear factor 4a (on chromosome 20q) and accounting for only 5 per cent of cases, and the most frequent (65 per cent of cases), termed MODY 3 and caused by hepatic nuclear factor-1a mutations (chromosome 12q). The hepatic nuclear factors are a family of transcription factors that regulate the expression of various genes, but their pathological relevance here is not clear.

Management of MODY is as for type 2 diabetes and using the same treatment targets, because it is now clear that patients with MODY are not protected against chronic complications. Diet and oral hypoglycaemic agents are often effective, especially in the milder MODY 2, but insulin may ultimately be needed.

Other types of diabetes (see Table 1)

Diabetes in pancreatic disease

Chronic pancreatitis, most commonly due to alcohol abuse, causes diabetes that needs insulin in about one-third of cases. Widespread flecks of fine to medium calcification are often scattered through the pancreas, outlining it on a plain abdominal radiograph. Concomitant destruction of the islet a cells means that glucagon secretion is lost as well as insulin; diabetic ketoacidosis is therefore rare, while hypoglycaemia can be profound and prolonged—a particular hazard in those who continue to drink alcohol. Acute pancreatitis causes acute hyperglycaemia in 50 per cent of cases but few develop permanent diabetes.

Carcinoma of the pancreas is associated with newly presenting type 2 diabetes, and should be suspected in older patients with weight loss (especially when accompanied by abdominal or back pain and jaundice). The mechanism is unknown but appears to be due to tumour products that cause insulin resistance rather than to b-cell loss.

Genetic diseases that cause diabetes through pancreatic damage include haemochromatosis and cystic fibrosis. In one-half of cases of haemochromatosis, heavy deposition of haemosiderin in the islets causes diabetes, usually requiring insulin; associated features are slate-grey skin pigmentation due to deposition of iron in the dermis ('bronze diabetes'), cirrhosis, secondary gonadal failure, and pyrophosphate arthropathy. Magnetic resonance imaging shows abnormal signals in liver and pancreas, while serum ferritin concentrations are greatly elevated; diagnosis is usually possible by means of molecular analysis of the *HFE* gene but Perl's stain for iron deposition in a liver biopsy may be necessary (see Chapter 11.7.1). Cystic fibrosis causes pancreatic exocrine failure, with an increasing risk of diabetes (often requiring insulin) that approaches 25 per cent in subjects who survive beyond 20 years of age.

Gestational diabetes

This includes all degrees of hyperglycaemia (impaired glucose tolerance as well as overt diabetes) diagnosed during pregnancy in previously normoglycaemic women. It is covered in Chapter 13.10.

Malnutrition-related diabetes

This controversial diagnostic category was omitted from the most recent World Health Organization classification. It included 'fibrocalculous pancreatic diabetes' and 'protein-deficient diabetes mellitus'. Fibrocalculous pancreatic diabetes was identified by dense pancreatic fibrosis, the formation of discrete and often spectacularly large stones in the dilated pancreatic ducts, and recurrent abdominal pain; protein-deficient diabetes mellitus was a vaguer entity that lacked the pancreatic stones. Patients conforming to these 'syndromes' were rare even in the tropical zones where they were described (less than 5 per cent of all diabetes), and the current consensus is that they represent type 2 diabetes or chronic pancreatitis superimposed on malnutrition.

Management of diabetes

The treatment of diabetes has traditionally concentrated on correcting hyperglycaemia, the most obvious and easily monitored biochemical abnormality and the cause of troublesome symptoms as well as specific chronic diabetic complications. This approach has not been entirely successful, partly because it is difficult to normalize blood glucose but also because macrovascular disease—the principal cause of morbidity and premature death—is heavily dependent on other factors, notably hypertension and dyslipidaemia. The current treatment targets for both type 1 and type 2 diabetes (Table 3) are therefore more holistic, tackling cardiovascular risk factors and obesity in addition to hyperglycaemia.

This section describes the roles of lifestyle modification and antidiabetic drugs, followed by specific treatment strategies for type 1 and type 2 diabetes.

Diet and lifestyle modification and management of obesity

About 80 per cent of patients with type 2 diabetes are obese, as are at least 30 per cent of those with type 1 disease. Obesity is arguably one of the greatest obstacles to successful management of diabetes: it worsens insulin resistance, dyslipidaemia, and hypertension and is now recognized in its own right as a risk factor for coronary-heart disease. Proven benefits of 10 per cent weight loss in type 2 patients with a body mass index of 30 to 40 kg/m ^{2} include falls in fasting glucose of 2 to 4 mmol/l and a 1 per cent decrease in HbA_{1c}—comparable with sulphonylureas or metformin—and reduced dosages of antidiabetic drugs, including insulin. There may also be variable improvements in blood pressure and dyslipidaemia (decreased triglycerides and low-density lipoprotein cholesterol, increased high-density lipoprotein). The traditional focus on obesity has been in type 2 diabetes, but there is no reason to assume that the cardiovascular hazards of obesity do not also apply to type 1 diabetes.

Weight reduction is regarded as the 'cornerstone' for treating obese type 2 diabetics but is often undermined by a lack of determination. Accordingly, doctors have little confidence in its efficacy and tend to assume that most obese patients will be 'dietary failures'. However, with clear advice, better understanding of the causes of obesity, and the use of realistic targets, the currently poor track record of diet and lifestyle therapy can be greatly improved. All members of the diabetes team must understand the principles (but not the detail) of lifestyle management so that a strong and unified message can be given to the patient.

The notion of the 'diabetic diet' must now finally be laid to rest. Traditionally, carbohydrate intake was restricted because of the simplistic assumptions that sugar alone raised blood glucose and might even be diabetogenic; this strategy favoured a high fat intake that undoubtedly helped to sustain obesity and probably predisposed to atheroma. Current advice is close to the 'healthy eating' recommendations for the whole population and can therefore be suggested for all the patient's family, which will greatly increase the chances of compliance.

The following diet and activity recommendations apply to both type 1 and type 2 diabetes. The aims are to:

1. Correct obesity, which worsens insulin resistance, reduces the efficacy of glucose-lowering, antihypertensive, and lipid-modifying drugs, and is an independent risk factor for macrovascular disease. Management of obesity is discussed in detail in Chapter 10.5.
2. Reduce cardiovascular risk, by limiting fat, cholesterol, sodium, and alcohol intakes.
3. Avoid hypoglycaemia in patients receiving insulin or sulphonylureas by optimizing the timing and content of meals.

The steps in designing dietary advice for the individual patient are shown in Fig. 6.

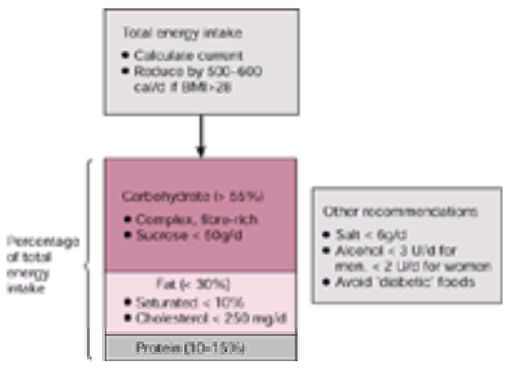

Fig. 6 Dietary recommendations for diabetic people. These guidelines now reflect 'healthy eating' for the general population, rather than a 'diabetic diet'.

Reducing total energy intake

This should be reduced by 500 to 600 cal/day (2100 to 2520 J/day) in patients who are overweight (body mass index over 28 kg/m ^{2}). This energy deficit mobilizes fat preferentially, whereas protein, glycogen, and water are also lost with more aggressive energy restriction; initially, the rate of weight loss will be 0.5 to 1.0 kg/week (adipose tissue contains around 7000 cal/kg (29 400 J/kg).

The desired energy intake should be calculated from standard formulae that employ the subject's age, sex, weight, and level of physical activity to estimate energy expenditure, which must equal energy intake under steady-state conditions. The standard dietary history is a waste of time for trying to assess energy intake, because overweight subjects consistently under-report how much they eat. Specific advice about how to cut energy intake is best left to the dietician, but hinges on reducing fat intake—a simple message that can be reinforced by the entire diabetes care team. Fat-rich foods not only have the highest energy density (9 cal/g (38 J/g) compared with 4 cal/g (17 J/g) for carbohydrate and protein), but also have poor satiating effects and so tend to encourage overeating.

The initial target should be 10 per cent loss of starting weight, not the 'ideal' body weight or body mass index, which is only rarely attained by obese diabetic patients. When energy intake is cut acutely, type 2 patients often show an immediate fall in blood glucose, due to a drop in hepatic glucose output, even before weight loss begins.

Weight loss during an energy deficit of 500 to 600 cal/day (2100 to 2520 J/day) is a slow process: for a 100 kg patient, 10 per cent weight loss may take several months. Frequent contact and encouragement are the best predictors of success, and the patient should be reassured that weight loss by a small but tolerable change in lifestyle is much more likely to be maintained than weight lost by a crash diet. As weight falls, resting energy expenditure also declines: it is proportional to lean body mass, which also decreases, although at a slower rate than fat. This means that greater reductions in energy intake (more than 600 cal/day (2520 J/day) will be needed to maintain the same rate of weight loss. If the 10 per cent target is met, further loss towards an 'ideal' body mass index of around 23 kg/m ^{2} may be feasible.

Weight loss is harder to achieve in diabetic patients than in their non-diabetic counterparts; possible reasons include fears about sugar rather than fat, and the adipogenic effects of insulin, sulphonylureas, and thiazolidinediones. In practice, weight loss of even 10 per cent is not commonly achieved by diet and lifestyle modification alone; only 15 to 30 per cent of newly diagnosed type 2 diabetic patients can normalize glycaemia initially by this means, and fewer than 10 per cent can sustain this for 5 years or more. The progressive b-cell dysfunction in type 2 diabetes (see above) makes it inevitable that the proportion of 'dietary failures' will increase steadily.

Improving dietary composition

Intakes of fat, salt, and refined sugar are generally too high in westernized populations. Current recommendations for healthy eating are based on evidence of beneficial effects on body weight, glycaemic control, lipids, and blood pressure (Fig. 6).

Fat should provide less than 30 per cent of total energy intake (in most industrialized countries, it accounts for 40 per cent). Polyunsaturated or monounsaturated fats (for example sunflower or olive oils respectively) are preferred to saturated animal fats, which should comprise less than 10 per cent of total energy intake. Patients may need to be reminded that 'good' unsaturated fats still contain 9 cal/g (38 J/g) and therefore sustain obesity just as effectively as the others. Cholesterol should be limited to less than 250 mg/day (less if dyslipidaemia is present).

Carbohydrates should account for more than 55 per cent of total energy intake, preferably in the form of foods rich in soluble fibre (such as pulses, root and leaf vegetables, and fruit); the current World Health Organization recommendation for the general population is for the consumption of at least four portions of fruit or

vegetables per day. Sugary drinks (especially fizzy glucose solutions that are supposed to give energy) should be avoided, except to treat hypoglycaemia. The present recommendation, which seems reasonable but is not based on evidence, is to limit added sucrose to less than 25 g/day and total sucrose intake to less than 50 g/day.

Protein should contribute 10 to 15 per cent of total energy—close to current levels in the general population. (For patients with renal impairment, see Chapter 20.10.1.)

Sodium intake should be less than 6 g/day, and less in patients with hypertension.

Alcohol contains 7 cal/g (29 J/g), and beers and wines in particular can be fattening. Intake should not exceed three units (30 g) per day in men and two units (20 g) per day in women, and should be further limited or avoided in those with hypertension or obesity. Alcohol can delay recovery from hypoglycaemia (see below); 'diabetic' beers (low sugar, but strong in alcohol) and spirits with sugar-free mixers are especially likely to provoke hypoglycaemia.

Moderate amounts of sucrose are acceptable (see above), while non-caloric sweeteners (such as aspartame) have no adverse metabolic effects. So-called 'diabetic' sweets and foods contain sorbitol or fructose instead of glucose, and are an expensive way to get diarrhoea; they should be avoided by patients, and withdrawn by the manufacturers.

Optimizing meal patterns

Judging the size and content of meals so as to limit glycaemic excursions remains an art rather than a science, and a skill which some patients develop with experience. Dosages of glucose-lowering drugs that act acutely to cover meals (short-acting insulin and sulphonylureas) can be tailored reasonably accurately to meals of similar composition but may not be matched to other meals, even when the total weights of carbohydrate, fat, and protein are similar.

There has been much interest in the ability of various foods to raise blood glucose, usually measured as the 'glycaemic index', i.e. the area under the curve of the rise in plasma glucose after eating a standardized load (50 g) of the food, expressed as a percentage of the area under the glucose curve after ingesting 50 g of glucose. Foods with a low glycaemic index include pulses and cereals, probably because of their high fibre and complex carbohydrate contents, while bread has a surprisingly high index. The glycaemic index of many foods such as potatoes and pasta varies widely according to the method of cooking (and even the shape of the pasta), and mixing different foods in a real-life meal has unpredictable effects on the overall postprandial glucose rise. It may be sensible to base meals around components with a low glycaemic index but it is clearly not feasible to use the index to adjust dosages of antidiabetic medication.

Increasing physical activity

Short-term exercise and improved physical fitness both increase insulin sensitivity, partly through increased translocation of GLUT-4 units to the surface of skeletal muscle cells; this effect is independent of insulin, and can enhance glucose uptake (under clamp conditions) better than metformin or the thiazolidinediones. Several studies, notably that conducted in Malmö, Sweden, have demonstrated that regular physical exercise reduces by 50 per cent the risk of impaired glucose tolerance progressing to type 2 diabetes, and also significantly decreases cardiovascular events. Exercise must therefore be encouraged in all diabetic patients, but the advice must be realistic, achievable, and safe. Brisk walking for 30 to 40 min every day is better physiologically than a hectic workout in the gym once or twice a week and is within almost everyone's reach.

Potential hazards of exercise are hypoglycaemia, which may be delayed by several hours (see below), and cardiac disease. Patients at risk should have an ECG, with consideration for an exercise tolerance test and echocardiography, and appropriate treatment for ischaemic heart disease or heart failure. Exercise remains beneficial and important in these cases but should be built up gradually.

Antiobesity drugs and bariatric surgery in diabetes

Antiobesity drugs may be indicated in selected obese diabetic patients with a body mass index over 28 kg/m ^{2} and who have demonstrated by losing weight beforehand through diet and exercise alone that they are prepared to make long-term changes in their lifestyle. Without this commitment, clinically useful weight loss is unlikely to be achieved or maintained beyond the period of drug prescription (currently 2 years for orlistat and 1 year for sibutramine); the medical and pharmacoeconomic benefits of modest weight loss for a couple of years in the obese patient's middle age are not known but are probably not dramatic.

Drugs currently available in many countries are orlistat, a gastrointestinal lipase inhibitor, and sibutramine, a combined serotonin/noradrenaline reuptake inhibitor. With each of these, up to 30 per cent of obese type 2 patients lose 10 per cent or more of body weight within 6 to 12 months, HbA_{1c} can fall by 1 per cent or more, and dosages of glucose-lowering drugs, including insulin, may be decreased.

Surgical treatment with gastric banding or gastric bypass operations is indicated in selected patients with a body mass index over 40 kg/m ^{2} (Chapter 10.5). This approach can achieve dramatic weight loss (up to 70 per cent of excess fat, maintained for several years), often with an impressive reversal of glucose intolerance: about 90 per cent of cases with type 2 diabetes or impaired glucose tolerance are returned to normoglycaemia.

Smoking

Smoking is at least as common among diabetic patients as in the general population. Smoking greatly amplifies macrovascular risk in diabetic subjects: 10-year mortality (mainly from myocardial infarction) is about 50 per cent higher than in diabetic non-smokers and twice as high as in non-diabetic non-smokers. Smoking may also accelerate the progression of nephropathy and possibly retinopathy.

Many diabetic people, especially young women, continue to smoke as a means of keeping thin, and because they fear gaining weight if they stop. Nicotine reduces fondness for sweet energy-dense foods and may also be mildly thermogenic. Weight gain after stopping smoking averages 3 kg but about 20 per cent of cases gain more than 6 kg; much of this weight is often lost within the following 1 to 2 years, and it can be limited or prevented by careful dietetic support beforehand and in the months after cessation. Moreover, the risks of continuing to smoke are much greater than this degree of weight gain, especially in diabetic people. Nicotine chewing gum may help patients to give up the habit.

Glucose-lowering drugs

Insulin

Insulin is the rational treatment for type 1 diabetes and the only drug that can normalize blood glucose in many type 2 diabetic patients. Unfortunately, subcutaneously injected insulin cannot match the physiological profile of normal insulin secretion (Fig. 7) and is a poor substitute for the finely tuned b cell with its nearly instantaneous capacity for 'in-flight' adjustment. Moreover, insulin given subcutaneously is absorbed into the systemic circulation rather than secreted into the portal system where an immediate effect on the liver, and first-pass clearance by that organ, are important in regulating the metabolic actions of insulin.

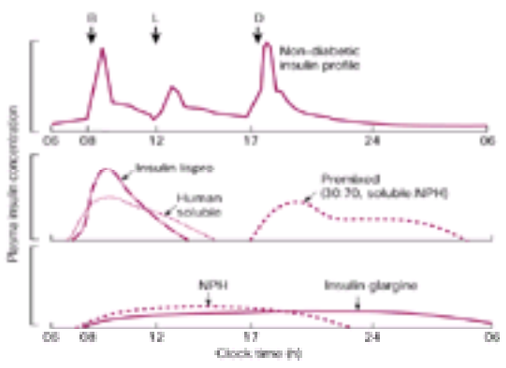

Fig. 7 Time course of insulin preparations, compared with the normal diurnal profile of plasma insulin concentrations in non-diabetic subjects (top). Breakfast (B), lunch (L), and dinner (D) were given as shown. Fast-acting analogues (such as lispro) act more rapidly than conventional soluble ones but are still sluggish compared with normal prandial insulin release. Premixed insulins injected in the early evening cover the evening meal adequately, but the long-acting component can cause hyperinsulinaemia and troublesome hypoglycaemia in the small hours. None of the conventional 'long-acting' insulins reliably lasts 24 h; new long-acting analogues such as insulin glargine may provide adequate background insulin levels with once-daily injections.

Insulin manufacture

Insulin was traditionally extracted from pork and beef pancreases in acid ethanol and purified by precipitation and recrystallization. Soluble (or 'crystalline') insulin prepared in this way was contaminated with other islet proteins, including glucagon and pancreatic polypeptide, which had an adjuvant-like effect and enhanced the immunogenicity of the injected insulin; immune reactions were relatively common with the 'dirty' animal insulins in use until the 1970s (see below). More sophisticated purification techniques including gel filtration yield 'highly purified' or 'monocomponent' insulins which only rarely provoke immune reactions.

Biosynthetic human-sequence insulin, produced by recombinant DNA technology, entered clinical practice in the early 1980s and was the first genetically engineered protein to be used therapeutically. The current approach is to introduce a synthetic gene for recombinant proinsulin or a novel insulin precursor into yeast; the secreted product is then cleaved enzymatically to yield insulin and C peptide.

There are some clinically relevant differences between the three species used therapeutically, although the shortcomings of insulin therapy relate mainly to the general pharmacokinetic misbehaviour of injected insulin. Human insulin is more lipophilic than porcine and bovine insulins and is slightly more rapidly absorbed: human soluble especially may lower glucose faster and patients being transferred from other species should be warned of this and prandial doses reduced initially by one-third. Human ultralente has a shorter and steeper action profile than its animal counterparts, particularly the bovine preparation; in real life, human ultralente behaves similarly to lente or isophane insulins and does not provide adequate basal levels for a full 24 h. Human insulin has been suggested to interfere with awareness of hypoglycaemia but the balance of evidence does not support this view (see below). Early beef insulins were especially prone to cause immune reactions (see below), although highly purified preparations do not appear to be particularly immunogenic.

Most insulin manufacturers are now turning to biosynthetic production of human-sequence insulin. Some patients prefer to continue using animal insulins—for reasons that may or may not appear scientifically sound—and these wishes should be respected by both clinicians and the pharmaceutical industry.

Insulin absorption

Absorption of insulin injected subcutaneously is slow and unpredictable. Individual day-to-day variability in the amount absorbed within a few hours can exceed 50 per cent. This means that small changes (less than 10 per cent) in insulin dosage are unlikely to influence glycaemic control, and that insulin treatment should generally not be adjusted on a daily basis.

Insulin absorption is influenced by the physical state of the insulin (soluble or delayed action), its speed of dissociation into monomers, the lipophilicity of the insulin species, and by blood flow and other characteristics of the injection site. Absorption is accelerated, and may lead to noticeably faster falls in blood glucose, by stimulating general or local blood flow through exercise, hot climate, saunas, and/or massaging the injection site. Conversely, absorption is slowed when subcutaneous blood flow is reduced, for example in cold conditions or hypovolaemic states. Lipohypertrophy, which may develop at frequently used injection sites, can significantly delay absorption—another reason for avoiding such areas.

The anatomical site of injection also influences the rate of subcutaneous absorption. It is fastest in the abdomen (also a good site to limit any effects of exercise) and arm, and slower in the leg. These differences are often eclipsed by the overall variability in absorption. Absorption from muscle is faster, presumably because of its higher blood flow, and this route is preferred for the emergency treatment of hyperglycaemia or ketoacidosis if the best option, controlled intravenous infusion, is not practicable.

Insulin preparations

Soluble (regular, or short-acting) insulin injected subcutaneously begins to lower glucose within 30 min, has a peak effect between 1 and 2 h and lasts 3 to 5 h (Fig. 7). This action profile is suitable for covering meals or hyperglycaemic emergencies and for use in insulin pumps or infusions. However, it would have to be injected several times per day to control hyperglycaemia around the clock, at the cost of frequent hypoglycaemia. Long-acting preparations are therefore used to cover basal insulin requirements.

Various approaches have been used to slow and prolong insulin absorption, especially the chemical combination of insulin into complexes that release it slowly. More recently, synthetic analogues have been designed whose structure promotes precipitation when injected subcutaneously (see Fig. 7).

Isophane insulins are also known as 'NPH' ('neutral protamine Hagedorn', from the director of the Danish laboratory where they were developed). They consist of a microcrystalline complex of insulin and the highly basic protein protamine (intriguingly isolated from fish sperm), together with trace amounts of Zn^{2+}. Isophanes were derived from protamine-zinc insulin which has a longer but highly unpredictable action profile. Isophanes produce peak plasma insulin levels at variable intervals between 4 and 8 h after injection, and their glucose-lowering action wears off rapidly after 10 to 12 h.

Insulin-zinc suspensions (lente insulins) employ higher Zn^{2+} concentrations which encourage insulin to form crystalline lattices. Varying the reaction pH can produce either larger crystals which are particularly slow to dissolve ('ultralente') or the amorphous 'semilente' which releases insulin faster; the familiar 'lente' is a 70:30 mixture of ultralente and semilente. Ultralente made with bovine insulin has a long, relatively flat action profile that can last 24 h or more, while human ultralente and the lente insulins of all three species have glucose-lowering profiles similar to that of isophane. These long-acting insulins have a cloudy appearance and need to be shaken before use to bring the insulin into suspension; visibly large particles or discoloration indicates that the insulin has become denatured and will have lost activity. Both lente and isophane insulins can be injected alone or mixed with soluble insulin.

Premixed insulins contain a short-acting soluble component together with a longer-acting lente or isophane. The aim is to provide prandial cover and then basal levels for several hours thereafter. Many preparations are available, with the proportion of short-acting insulin varying from 10 to 50 per cent. 30:70 mixtures are popular.

All these insulin types have been produced with porcine-, bovine- and human-sequence insulins, and are available in vials for pen injection devices.

Insulin analogues

The pharmacokinetic properties of native insulins of any species are poorly suited to subcutaneous injection: soluble insulins (despite their high-speed trade names) are too slow and prolonged in duration, while long-acting insulins do not provide reliable enough 24-h basal levels to be given once daily. Various synthetic insulin analogues, designed by molecular modelling, have improved physicochemical characteristics.

Fast-acting analogues are modified at the C-terminal end of the B chain, an area crucial in the self-association of insulin molecules, so as to resist dimerization and hexamerization. Insulin hexamers formed in the subcutaneous injection site, dissociate slowly into absorbable monomers, and this is a rate-limiting step in insulin absorption. Faster-acting analogues include insulin lispro (interchanging the B28 lysine and B29 proline residues of the normal human sequence) and insulin aspart, which carries aspartic acid at position B28 instead of the usual proline. They have an appreciably faster and shorter action profile (Fig. 7), and day-to-day variability in absorption and glycaemic responses may also be decreased. They can therefore reduce both prandial hyperglycaemia and the risk of postprandial hypoglycaemia. Overall, HbA_{1c} falls, with a reduced frequency of hypoglycaemia, when a fast-acting analogue is substituted for soluble insulin.

Long-acting insulin analogues are clear but precipitate when injected subcutaneously. Dissociation into monomers is at least as slow as with conventional long-acting

insulins and can provide basal levels for 24 h with a single daily injection, perhaps with less variability. An example currently entering clinical practice is insulin glargine (A21glycine, with two extra arginine residues extending the C terminus of the B chain).

Side-effects of insulin

Hypoglycaemia is the most common complication of insulin treatment and can be unpleasant, debilitating, and occasionally life-threatening.

Mild hypoglycaemia is common—many insulin-treated patients have at least one episode most weeks—but serious attacks causing unconsciousness or requiring the assistance of others are rare, about once every three patient years. Predictably, the frequency of both mild and severe attacks rises progressively when mean blood glucose levels are lowered by intensive insulin therapy; hypoglycaemia was three times more frequent in the tightly controlled group of the Diabetic Control and Complications Trial than in conventionally treated patients (see below).

The manifestations and treatment of hypoglycaemia are covered in detail later. As discussed there, there is no convincing evidence that the use of human as opposed to animal insulins specifically interferes with awareness of hypoglycaemic symptoms.

Weight gain is due to the anabolic effects of insulin, compounded by energy saved from glycosuria and sometimes by overeating after hypoglycaemia. Fear of weight gain discourages some patients, especially young women, from taking their full insulin dosages; deliberate omission or underdosing of insulin may be used surprisingly often to stay thin.

Lipohypertrophy is local thickening of subcutaneous tissue at frequently used injection sites, and is probably due to the lipogenic effects of high local insulin concentrations. Lipohypertrophy can be unsightly and can significantly delay insulin absorption. It can be prevented by rotating injections around several sites, and large lesions can be removed by liposuction.

Insulin allergy, now very rare with highly purified (especially human) insulins, can include local IgE-mediated erythematous reactions or even anaphylaxis. Lipoatrophy (localized pitting of the skin due to loss of subcutaneous fat) is apparently related to a chronic immune response generated around insulin crystals. 'Immune insulin resistance' was seen with impure animal and especially bovine insulins; high titres of insulin-binding antibodies mop up free insulin from the circulation, resulting in very high insulin requirements (occasionally more than 10 000 U/day), sometimes with unpredictable hypoglycaemia following release of antibody-bound insulin.

Insulin oedema is rare, and is usually seen in patients recovering from ketoacidosis who have been deprived of insulin for long periods. Fluid retention is probably due to the sodium-conserving effects of insulin on the renal tubule, and may cause ankle or generalized oedema. It usually resolves within a few days, although treatment with diuretics or ephedrine may be required.

Insulin regimens

Different individuals may need quite different insulin regimens, depending on their residual insulin reserve and severity of insulin resistance, as well as the desired tightness of control and the inconvenience that the patient will accept. Specific insulin schedules used in type 1 and type 2 diabetes are described later.

Insulin dosage

The healthy pancreas secretes about 40 to 60 U of insulin daily. Therapeutic insulin requirements range from less than this in thin type 1 patients (notably during the 'honeymoon period') to more than 200 U/day in very obese, insulin-resistant type 2 patients. High insulin requirements are often due to insulin resistance (see above), whereas low or falling dosages may be caused by weight loss (including anorexia nervosa), coeliac disease, or loss of counter-regulatory hormones in Addison's disease or hypothyroidism—all these conditions being associated with type 1 diabetes. Changing dosages, especially in previously stable subjects, should prompt investigation of these possibilities. Some patients with 'brittle' diabetes or psychological maladaptation to life with diabetes may pretend to take very high or very low dosages (see later).

Types of insulin

Formularies contain a bewildering assortment of insulins, many distinguished by imaginative claims about their action profile. Practically, prescribers should become familiar with regimens based on one or two preparations from the following broad classes:

1. Fast-acting insulin: either a soluble (regular) insulin such as Humulin S or Actrapid, injected 30 to 40 min before eating, or a faster-acting analogue (for example lispro or aspart) which can be given immediately before or even shortly after eating.
2. Long-acting insulin: either a lente insulin (for example Humulin Zn or Insulatard) or an isophane (for example Humulin I or Monotard). With either, circulating insulin falls to below useful levels after 10 to 14 h; they therefore need to be given twice daily in C-peptide negative patients, although those with residual insulin secretion (or who are given three premeal injections of soluble insulin) may be able to maintain good glycaemic control with a single bedtime injection. Bovine (but not human) ultralente can last a full 24 h, but its absorption is erratic. The long-acting analogues currently being introduced (such as insulin glargine) have flat, steady action profiles that may provide basal insulin levels with a single daily injection.
3. The timing of long-acting insulin injections does not have to be yoked to mealtimes as tightly as for soluble insulin but it is convenient to inject the morning dose at the same time as the prebreakfast soluble, either separately or mixed in the same syringe (glargine cannot be mixed with other insulins). When a second long-acting injection is needed, this is best given at bedtime, rather than together with the presupper soluble dose. This is because the action profile of long-acting insulin clashes with the physiological changes in insulin sensitivity that occur overnight. Growth hormone is normally secreted in large spikes on entering deep sleep, typically between 24.00 and 02.00 hours; this induces delayed insulin resistance which raises blood glucose during the hours leading up to breakfast. This 'dawn phenomenon' is accentuated if insulin levels are falling simultaneously—as happens if long-acting insulin is injected in the early evening. Another hazard with this timing is potentially dangerous nocturnal hypoglycaemia when insulin levels peak during the early morning (typically 02.00 to 04.00). Both problems can be reduced by delaying the long-acting injection to bedtime (22.00 to 23.00), when the risk of nocturnal hypoglycaemia is lower, and insulin levels generally persist long enough to counteract the insulin resistance of the dawn phenomenon.
4. Premixed insulins (for example 30 per cent short-acting with 70 per cent long-acting) are obviously more convenient than giving short- and long-acting insulins separately, but they lack flexibility. Premixed insulin injected 30 to 40 min before breakfast can achieve good glycaemic control through the morning and afternoon, but timing the evening dose is problematic: giving it before supper will tend to cause both early-morning hypoglycaemia and fasting hyperglycaemia because of the time course of the long-acting component, and simply increasing the evening dosage often makes nocturnal hypoglycaemia worse while failing to lower the prebreakfast glucose.

Insulin injections

Most insulin formulations are now available for both conventional syringes or pen injection devices. Pen injectors are compact, convenient, and easy to use: the required dose is 'dialled up' and injected by pressing the plunger; the ratchet mechanism of most pens gives an audible click that can help blind patients to count dosages.

Syringes and pens carry very fine (28 to 31 gauge) needles that allow insulin to be injected almost painlessly. The needle should be pushed in vertically and the insulin injected over a few seconds. 'Backtracking' of insulin to the skin surface, which can occasionally cause loss of several units of insulin, may be reduced by leaving the needle in place for a short while. A spot of bleeding may occur; very rarely, sudden hypoglycaemia may be due to direct injection of insulin into a subcutaneous vein.

Injections can be given into any site that is accessible and well upholstered with adipose tissue, especially the abdomen, thighs, buttocks, and upper arms. The abdomen has the advantage (theoretically at least) of relatively faster absorption that is less influenced by exercise, as compared with the limbs. Rotating injection sites, for example between the abdomen and leg, or around the quadrants of the abdomen, helps to avoid local reactions, especially lipohypertrophy which can make insulin absorption slow and erratic.

Jet injectors fire a metered dose of insulin as a high-pressure aerosol that penetrates the skin. These have is obvious appeal to patients with needle phobia, although there may be bruising and delayed discomfort at the injection site. Jet injectors are bulky and expensive and do not offer any pharmacokinetic advantages over

conventional injections.

Insulin pumps

Portable insulin pumps that administer continuous subcutaneous insulin infusion were developed by Pickup and colleagues in the late 1970s. Modern pumps are compact and light and worn in a belt or holster. Soluble insulin in a special cartridge is delivered through a fine-bore cannula and a butterfly-type cannula, which is inserted subcutaneously in the anterior abdominal wall and generally left in place for 1 to 2 days; the pump can be safely removed for 30 to 40 min for bathing or other activities. Different basal rates can be preprogrammed, and mealtime boluses are selected and given by pressing a button. Typical basal rates are 1 to 2 U/h during the day and 0.5 to 1 U/h overnight, with mealtime boluses (given 30 min before the main meals) amounting to about 50 per cent of the total daily dose. Insulin lispro has been used in pumps and may slightly improve control compared with conventional soluble insulin.

Continuous subcutaneous insulin infusion can achieve relatively steady insulin levels under laboratory conditions but cannot overcome the fundamental variability of subcutaneous insulin absorption. When used carefully by highly motivated patients who are supported by an experienced diabetes care team, continuous subcutaneous insulin infusion can achieve glycaemic control which is at least as good as that achieved with multiple injections; the two were used side by side in the Diabetic Control and Complications Trial. Insulin pumps are expensive (£1000 to £1500 (US$1500 to 2300)) as are consumables (another £1000 per year); medical backup can also be costly to provide. Continuous subcutaneous insulin infusion is only indicated for well-informed patients who are prepared to monitor their blood glucose frequently and take some responsibility for adjusting the pump. It is widely used in the United States and some European countries but less in the United Kingdom.

Infections at the infusion site with pyogenic skin commensals or unusual organisms (for example atypical mycobacteria) are uncommon but can be troublesome and cause rapid deterioration in glycaemic control. An increased rate of diabetic ketoacidosis was reported with earlier and less reliable pumps. With continuous subcutaneous insulin infusion, the subcutaneous insulin depot is only a few units, and any interruption of insulin delivery (such as with pump failure or cannula blockage) can lead to rapid rises in blood glucose and especially ketone levels. However, modern pumps carry no excess risk of diabetic ketoacidosis as compared with intensified injection therapy. Similarly, the risk of hypoglycaemia due to the pump over-running is now very low.

Continuous intraperitoneal infusion

The peritoneum is a good route for insulin administration: absorption is very rapid across its large surface area and insulin enters the portal circulation. Continuous intraperitoneal insulin infusion has been used in some cases, mostly employing a pump and reservoir implanted subcutaneously in the abdomen and delivering insulin through a flexible cannula sewn into the peritoneal cavity. The reservoir is filled with soluble insulin through an injection port lying just beneath the skin and is emptied by a liquid gas compression system at a rate that can be varied by an external electromagnetic control. Continuous intraperitoneal insulin infusion can provide basal insulin; meals need to be covered by additional insulin, usually injected subcutaneously.

Intraperitoneal pumps are expensive and convincing indications for their use are rare. They have been successful in some patients with apparently very high subcutaneous insulin dosages but surprisingly normal intravenous requirements. It is now clear that this situation is not due to a mysterious syndrome of 'subcutaneous insulin resistance', and that most of not all of these patients are interfering with their own treatment (see below). In this setting, continuous intraperitoneal insulin infusion is probably effective because these pumps are difficult to sabotage.

Oral hypoglycaemic agents

Sulphonylureas and repaglinide

The sulphonylureas were the first orally active glucose-lowering drugs to be used and were discovered in the 1930s when early sulphonamide antibiotics were found to cause hypoglycaemia. The 'first generation' (chlorpropamide, tolbutamide) have since been superseded by the 'second generation' (for example gliclazide and glibenclamide) and by newer agents such as glimepiride. Repaglinide acts in a similar way to the sulphonylureas.

Mode of action

Sulphonylureas are insulin secretagogues but insulin synthesis is not stimulated. Insulin levels peak within 1 to 2 h and decline within 4 to 6 h for the short-acting drugs (such as gliclazide) but may remain elevated for much longer with chlorpropamide and glibenclamide, which therefore carry a greater risk of hypoglycaemia. An 'extrapancreatic' action has also been attributed to sulphonylureas, i.e. improving insulin sensitivity. This effect is small and is probably explained by the non-specific decrease in insulin resistance ('glucotoxicity') when hyperglycaemia is corrected by any means.

Repaglinide acts in a similar way to the sulphonylureas but is structurally different. It is derived from the non-sulphonylurea part of the glibenclamide molecule (called meglitinide), which was found fortuitously to have glucose-lowering activity of its own.

Efficacy and potency

The ability of these agents to lower glycaemia depends on how much insulin is available for release from the b cells (which are already stimulated by hyperglycaemia) and by the severity of insulin resistance. In practice, all sulphonylureas lower basal and postprandial glucose levels by no more than 2 to 4 mmol/l and HbA_{1c} by 1 to 2 per cent; mild hyperglycaemia may therefore be corrected but patients with fasting glucose in excess of 13 mmol/l are very unlikely to achieve normoglycaemia (so-called 'primary failure'). Moreover, as b-cell function declines progressively in type 2 diabetes, many patients who initially respond well to sulphonylureas will subsequently need additional glucose-lowering drugs; this 'secondary failure' overtakes 5 to 10 per cent of patients per year, in a cumulative fashion. These limitations apply to all sulphonylureas and repaglinide: the more potent drugs have lower therapeutic dosages than the earlier agents but cannot lower glycaemia any further.

Pharmacokinetics

Most are taken twice daily with meals; glimepiride is taken once daily and repaglinide with each meal. Chlorpropamide has a very long action profile, while glibenclamide shows variable and sometimes prolonged hypoglycaemic activity. Sulphonylureas and repaglinide bind to circulating proteins and may be displaced by other strongly protein-bound drugs, causing hypoglycaemia (see below). All these drugs are cleared through the kidneys and can accumulate in renal failure, causing frequent hypoglycaemia and other side-effects. Gliquidone is metabolized mainly in the liver and may be slightly less hazardous in patients with renal impairment, although insulin is usually indicated in these cases.

Side-effects

Weight gain is due to the anabolic effects of hyperinsulinaemia, compounded by reduced losses of energy through glycosuria. Weight gain is typically 2 to 3 kg greater than with diet alone or metformin.

Hypoglycaemia is rarer than with insulin, but the risk is greater with longer-acting sulphonylureas (glibenclamide, chlopropamide), in renal failure, and in the elderly.

Sulphonylureas can cause allergic reactions including skin rashes (notably Stevens–Johnson syndrome) and marrow dyscrasias, and can precipitate acute intermittent porphyria. Side-effects exclusive to chlorpropamide include the syndrome of inappropriate secretion of antidiuretic hormone (see Chapter 20.2.1) and acetaldehyde-mediated facial flushing on drinking alcohol.

The cardiovascular safety of sulphonylureas has remained under a cloud since tolbutamide was associated with an excess of cardiovascular deaths during an essentially uninterpretable study (the UGDP) conducted in the 1970s; the presence of the SUR-2 receptor on cardiomyocytes has recently reinforced suspicions that these drugs may trigger ischaemia and arrhythmias (by preventing preconditioning). However, the recent United Kingdom Prospective Diabetes Study found no evidence that patients treated with sulphonylureas suffered cardiovascular events more often than those treated with insulin. Glimepiride is highly selective for SUR-1.

Indications and contraindications

These drugs are first-line therapy for type 2 patients in whom lifestyle and dietetic measures have failed to control hyperglycaemia. Because of their tendency to increase weight, they are best suited to non-obese patients. They can be usefully combined with metformin, which may partly offset weight gain, or insulin.

Insulin secretagogues are inappropriate for severely insulin deficient patients or during intercurrent illness, when insulin is needed, and are unlikely to be effective if fasting glucose exceeds 13 mmol/l. Sulphonylureas are contraindicated in renal failure: all should be stopped and insulin started if serum creatinine exceeds 250 µmol/l. Pregnancy has been viewed as a contraindication, because sulphonylureas cross the placenta and could cause fetal hyperinsulinaemia and perhaps teratogenesis; however, a recent study did not substantiate these concerns (see Chapter 13.10).

Many drugs interact with sulphonylureas, the most common interaction being hypoglycaemia due to displacement and/or decreased clearance of protein-bound sulphonylureas (for example by sulphonamides, fibrates, salicylates, and probenecid). Potential interactions must always be checked for any drug being contemplated in patients receiving sulphonylureas.

Choice of drug

There is little to choose between the newer agents; chlorpropamide is now obsolete. Glibenclamide should be avoided in the elderly because of its unpredictable tendency to cause hypoglycaemia.

Metformin

Metformin and phenformin are biguanides, the class of compounds responsible for the mild hypoglycaemic action of goat's rue (an otherwise undistinguished weed). Phenformin is no longer available in many countries because it carries a 10-fold greater risk of lactic acidosis, while metformin only recently entered clinical use in the United States.

Mode of action

Metformin acts primarily by inhibiting gluconeogenesis in the liver, thus reducing the raised hepatic glucose output which underpins basal and overnight hyperglycaemia; this effectively enhances the action of insulin on the liver. Peripheral glucose uptake may also be increased, while gastrointestinal side-effects may help to reduce fondness for food. Metformin does not stimulate insulin secretion.

Overall, metformin lowers blood glucose (especially postprandial) by 2 to 4 mmol/l and HbA_{1c} by 1 to 2 per cent, which is comparable to the effect of sulphonylureas. On its own, metformin does not cause hypoglycaemia, although this can obviously occur when it is combined with either a sulphonylurea or insulin. Weight does not usually increase with metformin, and may fall.

Metformin may have beneficial cardiovascular effects, as the United Kingdom Prospective Diabetes Study found a reduction in vascular events in the metformin-treated group only (see below). It is not clear whether this is related to specific metabolic effects of metformin (improved insulin sensitivity), to its antiobesity properties, or to other actions such as reported reductions in blood pressure and coagulability.

Pharmacokinetics

Metformin is given twice or thrice daily with meals. It is cleared mainly through the kidneys, and the increase in plasma levels in renal failure is a major risk factor for lactic acidosis.

Side-effects

Gastrointestinal symptoms (30 per cent of cases) include altered taste, loss of appetite, heartburn, abdominal discomfort and bloating, and diarrhoea (metformin is the commonest cause of this in the diabetic clinic). These problems are mostly mild, but may discourage the patient from taking the drug; they can be reduced by starting with a low dosage and increasing it slowly.

Lactic acidosis is very rare with metformin (about 3 cases per 100 000 patient-years) if it is carefully prescribed. This stems from the mode of action of metformin, namely the inhibition of hepatic gluconeogenesis—a process that constantly consumes the lactate produced by glycolysis. Blood lactate levels are modestly raised in patients receiving biguanides, and can escalate rapidly and cause life-threatening acidosis if lactate is overproduced (for example in respiratory or cardiac failure), or is not cleared by the liver (hepatic failure), or if metformin accumulates in renal failure. Lactic acidosis is described in detail later.

Megaloblastic anaemia can occur due to impaired absorption of vitamin B_{12}.

Indications and contraindications

Metformin is a first-line alternative to sulphonylureas in type 2 patients whose hyperglycaemia does not respond adequately to modification of diet and lifestyle; as it does not tend to cause weight gain, and may even reduce weight, it is often used in obese patients. Its addition can also be helpful in obese patients who are poorly controlled by sulphonylureas or insulin. Metformin is being evaluated in insulin-resistant conditions such as polycystic ovary syndrome and in impaired glucose tolerance.

Contraindications include all the major organ failures— renal, hepatic, cardiac, and respiratory. It should not be used when serum creatinine concentration exceeds 125 µmol/l.

Thiazolidinediones

Thiazolidinediones are a novel class of glucose-lowering drugs which improve insulin sensitivity. There are distinct differences between individual thiazolidinediones which influence their therapeutic spectrum and safety. Rosiglitazone and pioglitazone are currently available in many countries; troglitazone has been withdrawn because it caused rare but life-threatening hepatic damage.

Mode of action and pharmacokinetics

Thiazolidinediones bind to specific receptors in the nucleus which have the cumbersome title 'peroxisome proliferator activating receptor-g' (PPAR-g). PPAR-g and the related peroxisome proliferator activating receptor-a (the target for the fibrate class of lipid-lowering drugs) are ligand-activated transcription factors whose natural ligands appear to be fatty acid derivatives. PPAR-g that has bound a thiazolidinedione forms a heterodimeric complex with another nuclear receptor, RXR, bound to its own endogenous ligand, retinoic acid. The heterodimer then binds to specific recognition motifs found in the promoter sequences upstream of many genes, notably those involved in adipocyte and lipid metabolism.

The affinity of individual thiazolidinediones at PPAR-g parallels their glucose-lowering ability in animal models of type 2 diabetes, but their precise mode of action remains uncertain. Thiazolidinediones exert concerted effects that encourage the storage of triglyceride in mature adipocytes, including the differentiation of preadipocytes into adipocytes and enhanced expression of lipogenic enzymes; overall, circulating levels of free fatty acids fall and this may reduce hepatic glucose production and increase glucose uptake into muscle as described earlier. The net effect is to enhance the action of insulin—hence their description as 'insulin sensitizers'. Thiazolidinediones have negligible glucose-lowering action unless insulin resistance and hyperglycaemia are present. As with metformin, they do not cause hypoglycaemia when used alone, but can exaggerate the hypoglycaemic effects of insulin or sulphonylureas.

Efficacy and potency

Alone, all thiazolidinediones lower glucose by 2 to 3 mmol/l and HbA_{1c} by 1 per cent, somewhat less than the sulphonylureas. For unknown reasons, blood glucose declines slowly during thiazolidinedione treatment, and a maximal effect may not be reached for several weeks. Rosiglitazone is the most potent thiazolidinedione to date but, as with the more potent sulphonylureas, cannot lower blood glucose further than the other thiazolidinediones.

Pharmacokinetics

All are metabolized in the liver and cleared chiefly through the kidney. They are highly protein bound.

Side-effects

Weight gain, averaging 1 to 4 kg, is due mainly to fat deposition. This appears to spare the visceral depot associated with insulin resistance and does not negate the glucose-lowering action.

Fluid retention of unknown aetiology may cause a mild dilutional anaemia (haemoglobin typically falls by 1 to 2 g/dl) and ankle oedema (in 5 to 10 per cent of cases); rarely, heart failure may be precipitated in patients with pre-existing myocardial dysfunction, especially if they are also treated with insulin.

Hepatic damage, ranging from subclinical elevations of hepatic enzymes to fulminant and fatal hepatic necrosis (about one case per 1000 patient-years), has been reported with troglitazone but does not appear to be a risk with rosiglitazone or pioglitazone.

Indications and contraindications

Thiazolidinediones are currently re-garded as second-line drugs for treating type 2 diabetes when sulphonylureas or metformin are ineffective or insuitable. They can be combined with either a sulphonylurea or metformin, when HbA_{1c} may fall by more than 1 per cent; if HbA_{1c} has not fallen by more than 1 per cent within 6 months of adding a thiazolidinedione, it should be discontinued. When pioglitazone is used with insulin, insulin dosage can be reduced but weight gain may be problematic; rarely, heart failure may be precipitated (the combination of rosiglitazone with insulin is currently contraindicated). Subjects with impaired glucose tolerance treated with a thiazolidinedione have a lower risk of progressing to overt type 2 diabetes, and the drugs can improve hirsutism and menstrual dysfunction (sometimes inducing ovulation) in women with polycystic ovary syndrome.

Contraindications include hepatic dysfunction and congestive heart failure. Although these is no evidence of hepatotoxicity with thiazolidinediones other than troglitazone, it seems prudent to monitor liver enzymes periodically and to stop the drug if transaminases rise to more than 1.5 times the upper limit of normal, or if any other signs of hepatic dysfunction appear.

a-Glucosidase inhibitors

Acarbose (and the related miglitol and voglibose) are inhibitors of a-glucosidase, an enzyme of the brush border of the small intestine essential for the breakdown of dietary starch to disaccharides, which are then hydrolysed to the absorbable monosaccharides. Their rationale is to block digestion of complex carbohydrates and so damp post-prandial glycaemic rises. The therapeutic effect is small: post-prandial glucose may fall by 1 to 2 mmol/l, with predictably little impact on overnight glucose, and HbA_{1c} by 0.5 per cent or less. Side-effects due to carbohydrate malabsorption (flatus, abdominal bloating, gassy diarrhoea) are common and probably damage compliance. Despite its poor efficacy and low tolerability, acarbose is still widely prescribed and in some countries is regarded as a first-line drug.

Practical management of hyperglycaemia

Most newly diagnosed diabetic patients are easily allocated to either type 1 or 2 on clinical criteria (Table 2) and treatment is started accordingly. However, initial impressions may be misleading: a thin young patient may not need insulin because he has MODY, whereas a classical maturity-onset subject may lose weight rapidly and develop ketoacidosis because he has type 1 diabetes. Continuing monitoring and vigilance are therefore essential.

Type 1 diabetes

These patients must be given insulin immediately and for life. The insulin regimen will depend particularly on any remaining endogenous insulin, the patient's body weight, lifestyle, and motivation. Patients with residual insulin secretion, especially newly presenting and particularly during the 'honeymoon period'(see below), can often fill in gaps in insulin replacement and enjoy good glycaemic control with few injections and low insulin dosages. However, C-peptide negative patients will require exogenous insulin to cover both basal and prandial needs (Fig. 7) to achieve good control. Regimens include:

1. Twice daily long-acting insulin with preprandial short-acting insulin. Lente or isophane is injected before breakfast (and can be mixed with prebreakfast short-acting insulin) and at bedtime (see above). Soluble insulin is injected 30 min before breakfast and the evening meal, or a fast-acting analogue (such as lispro or aspart) given with food. Mid-day meals, unless large, are usually covered satisfactorily by the morning's long-acting dose and do not need separate short-acting insulin.
2. Once daily long-acting insulin with preprandial short-acting insulin is currently unsatisfactory because both lente and isophane run out too quickly, but longer-lasting analogues such as glargine may be effective when injected once daily at bedtime or breakfast. Short-acting insulin is given separately to cover meals, as above.
3. Premixed insulins injected before breakfast and before the evening meal suit some patients and many doctors, but often fail to control overnight and/or fasting glucose levels (see above).

Insulin dosages should be titrated according to blood glucose and HbA_{1c} monitoring (Table 3). Highly motivated patients may be suitable for continuous subcutaneous insulin infusion treatment as discussed above.

Starting insulin therapy

Patients at risk of ketoacidosis need hospital admission, while those who are clinically well can start insulin at home, supervised by a specialist diabetes nurse. Good control can often be achieved with long-acting insulin injected at breakfast and bedtime, starting with low dosages (for example 8–12 U in the morning and 4–6 U at night) to avoid potentially demoralizing hypoglycaemia. Short-acting insulin can then be added to cover excessive prandial hyperglycaemia. Wherever practicable, patients should be encouraged to give their own injections as soon as possible.

Newly diagnosed patients starting insulin need to be warned about a possible 'honeymoon period' of good glycaemic control, when the fall in glucose levels allows partial recovery of the remaining b cells. Blood glucose can often be easily controlled with low insulin dosages (and exceptionally, without exogenous insulin) but the honeymoon ultimately ends within a few months: blood sugar levels and insulin requirements then escalate, because of the progressive loss of remaining b cells.

Poor diabetic control and 'brittle' diabetes

In real life, relatively few type 1 patients approach the high-quality glycaemic control aspired to in Table 3. This largely reflects the pharmacokinetic shortcomings of current insulin preparations and the unpredictable nature of subcutaneous absorption. The patient's compliance is a crucial determinant of overall diabetic control; teenagers are notoriously resistant to advice abut diabetes, as with other matters, and many have markedly elevated HbA_{1c} concentrations. This clearly increases the risk of future diabetic complications.

A few patients have such poor metabolic control that they cannot live a normal life. Most have chronically high blood glucose and suffer recurrent hospital admissions with ketoacidosis; some suffer frequent hypoglycaemia, while others have an unstable or 'brittle' blood glucose profile that can swing rapidly between hyper- and

hypoglycaemia. Occasionally, endocrine or intercurrent illnesses are found to be responsible (Table 4), but most cases remain 'idiopathic' after even intensive investigation. It is now clear that poor compliance, often with deliberate interference with treatment, is responsible in many of these patients. Most are young women who tend to be obese and are generally hyperglycaemic despite apparently high insulin dosages; when tested under controlled conditions, however, their intravenous and subcutaneous insulin requirements are unremarkable. Many are probably omitting insulin or taking only small doses: common motives include escape from difficulties at school or home, or wanting to stay thin (disturbances of body image are common in this group). Initially, such patients may appear to lead charmed lives despite frequent hospital admissions but many die prematurely (especially from ketoacidosis or hypoglycaemia); significant diabetic complications frequently develop during their twenties or thirties.

Management can be extremely difficult. Patients with sustained poor control should be admitted selectively for intensive education, observation, and exclusion of other possible causes (Table 4). In some cases, it may be necessary to confirm that insulin is effective at conventional doses (for more information see the paper by Schade and Duckworth in Further reading). Even close supervision in hospital does not exclude ingenious interference with insulin treatment or glucose monitoring. Intensified insulin schedules or continuous subcutaneous insulin infusion may help in some cases but the key is more likely to be sympathetic counselling (perhaps with psychotherapy) of the patient and his or her family.

Experimental and future treatments for type 1 diabetes

Pancreatic transplantation, usually performed in conjunction with renal transplantation for patients with diabetic nephropathy, can achieve good results including long-term withdrawal of exogenous insulin in 10 per cent of cases. The whole gland or a segment is transplanted into the pelvis and anastomosed to the iliac vessels; to avoid damage from pancreatic exocrine secretions, the duct is either occluded or drained into the bladder (when urinary amylase excretion can indicate the health of the graft). Problems are the need for lifelong immunosuppression (required anyway for renal transplantation) and the global shortage of donor organs.

Pancreatic islet transplantation is gaining ground, especially transcutaneous injection into the portal vein of islets isolated from a donor pancreas; these colonize and function well in the liver, the first stop for insulin secreted physiologically. A novel immunosuppressive regimen without glucocorticoids can improve the outcome (most cases become insulin independent), but two or three donor pancreases are currently needed for each recipient.

Prevention of type 1 diabetes by aborting insulitis during the long prediabetic phase has been achieved by immunosuppression in some high-risk subjects, especially the siblings of type 1 patients who have diabetogenic HLA class II antigens and high titres of GAD and islet cell antibodies. Better immunosuppressive regimens may improve the results.

Antidiabetic drugs under development include aerosolized insulin that exploits the lung's large absorptive area and enters the bloodstream rapidly, and various low molecular weight insulin mimetics that either bind to the insulin receptor or enhance postreceptor signalling.

Management of type 2 diabetes

Dietary and lifestyle measures form an essential foundation for management of type 2 diabetes and must be maintained throughout, even though fewer than 10 per cent of patients can be controlled satisfactorily for more than a year by these means alone.

Patients who fail to meet the glycaemic targets in Table 3 should generally follow the steps outlined below, although compromises may be more appropriate in the elderly or those at risk of hypoglycaemia. Progress should be reviewed every 3 months or so if blood glucose is unacceptably high; the inexorable deterioration of b-cell function in type 2 diabetes means that there is no point in delaying decisions to increase drug doses or add insulin.

First-line oral hypoglycaemic agents for so-called 'dietary failure' are a sulphonylurea (or repaglinide) or, particularly for obese patients (those with a body mass index in excess of 30 kg/m^2), metformin. Dosages can be increased to a maximum over a few weeks. Should one class fail, the other can be tried but a dramatic improvement is unlikely.

Combination oral therapy: two of sulphonylurea (or repaglinide), metformin, or a thiazolidinedione can be used together. Some diabetologists would try adding acarbose at this stage. Triple therapy is generally not worth trying.

Long-acting insulin with a first-line oral agent: although seemingly illogical, a bedtime injection of isophane can control blood glucose overnight and before breakfast, and this apparently helps oral hypoglycaemic agents to act more effectively during the day. The combination of metformin (thrice daily with meals) with bedtime isophane often achieves good glycaemic control, while limiting the weight gain that commonly follows the introduction of insulin in type 2 patients. Isophane with a sulphonylurea or pioglitazone may increase weight; rosiglitazone is contraindicated in combination with insulin.

Insulin therapy can range from once- or twice-daily long-acting insulin in subjects with residual insulin, to the more intensified basal and prandial regimens used in type 1 diabetes. Large dosages (100 to 150 U/day) may be needed to achieve good glycaemic control in obese, highly insulin-resistant subjects. As in type 1 diabetes, rapidly acting and very long-acting analogues may be able to improve on the currently available native insulins.

Obesity (and therefore insulin resistance) may worsen when insulin treatment is started; possible reasons include reduced losses of energy through glycosuria, a tendency to relax dietary restriction when a more effective means of lowering glycaemia is introduced, and sometimes overeating during hypoglycaemic episodes. Increasing insulin resistance may lead to escalating insulin dosages. The possible hazards of insulin-induced obesity are not clear but could theoretically include vascular disease, which may be hinted at by the lower frequency of cardiovascular events among patients treated with metformin in the United Kingdom Prospective Diabetes Study trial. At present, however, the consensus is probably to aim for the glycaemic targets in Table 3 (which will reduce the risks of microvascular complications) and to accept an increase in weight, while actively treating other cardiovascular risk factors.

Experimental and future treatments for type 2 diabetes

GLP-1 is an incretin hormone that stimulates insulin secretion and may also induce satiety, particularly by delaying gastric emptying. Blood glucose can be lowered comparably to sulphonylureas with GLP-1 infused intravenously. Buccal and orally active preparations are under development, as are inhibitors of dipeptidyl peptidase IV, the enzyme which degrades GLP-1.

Combined peroxisome proliferator activating receptor a/g agonists have insulin-sensitizing and lipid-lowering actions in animals. Various novel insulin secretagogues and non-thiazolidinedione insulin sensitizers are in development; one agent (S15261) combines both these properties.

Antiobesity drugs could have an important impact in many type 2 patients. Agents currently in preclinical testing include neuropeptide Y receptor (Y5) antagonists, melanocortin-4 receptor agonists, and low molecular weight leptin mimetics (see Chapter 10.5).

Monitoring diabetic control

Treatment targets for blood glucose and type 1 and type 2 diabetes (Table 6) have been selected to reduce the risk of chronic diabetic complications. Avoiding acute episodes of hyper- and hypoglycaemia is also important.

Blood glucose monitoring

Blood glucose concentration can be easily and quickly measured in small drops of blood (a few microlitres or less), using various test strips; the ability to perform such measurements is an essential skill for all professionals delivering diabetes care and for most diabetic patients. Test strips contain glucose oxidase (which catalyses the oxidation of glucose to gluconic acid) together with a detection system to measure specific reaction products, either electrochemically or colorimetrically (using dyes sensitive to hydrogen peroxide). The signal is read by a reflectance meter or electrically, and converted into the glucose concentration in the sample. Colour-based test strips can also be read by eye against a printed standard scale, although this may be difficult for partially sighted or colour-blind patients.

A drop of blood is obtained by pricking the sides of the fingertip, avoiding the sensitive pads; various lancets and automatic finger-pricking devices are available.

Blood must cover the reaction area completely and be left in contact for exactly the period stipulated; some meters read out automatically at this point, whereas other strips must be wiped dry and left for the colour to develop. Failure to follow the manufacturer's instructions is the main cause of inaccurate readings, which are disturbingly frequent. With attention to detail, readings correspond closely to laboratory measurements of glucose (which also employ the glucose oxidase reaction) but are not reliable enough to be used for diagnosing diabetes.

Monitoring schedules

Type 2 diabetes treated with diet and oral agents can be monitored using fasting glucose and values in the midafternoon, both of which correlate with overall glucose level, measured once or twice per week.

Insulin-treated patients may need more frequent monitoring to adjust insulin dosages. Fasting glucose is determined by the previous evening's long-acting insulin, while values before the evening meal reflect mainly the morning's long-acting dose. Prandial short-acting insulin dosages can be titrated from the glucose rise 90 to 120 min after eating. Readings can be scattered across these time points on different days; most patients can be persuaded to check their glucose levels once or twice per day.

Written records help to bring out general patterns in glucose control. Patients must also be encouraged to check their glucose if they feel unwell and, crucially, frequently during intercurrent illness. Occasional tests during the night (especially between 02.00 and 04.00) are useful in patients at risk of nocturnal hypoglycaemia, including those injecting long-acting or premixed insulins in the early evening.

Checking the self-monitoring technique and the patient's action plan when glucose levels fall outside the target range is a core part of the patient's diabetic education.

HbA_{1c} and fructosamine

These tests measure the non-enzymatic reaction of glucose with circulating proteins (see below), and therefore reflect longer-term blood glucose levels. Glycated (glycosylated) haemoglobin (HbA_1) results from the combination of glucose with the N-terminal valine residue of the B chain of adult Hb (HbA), and can be separated from unaltered HbA by electrophoretic and other methods. HbA_1 includes the stable HbA_{1c} fraction, which is most closely related to average blood glucose levels over the preceding 6 to 8 weeks.

The various assay methods for HbA_{1c} are now being standardized to match the methodology used in the Diabetic Control and Complications Trial (DCCT), which defined the long-term risks of diabetic microvascular complications (see below). For assays conforming to DCCT standards, non-diabetic HbA_{1c} ranges from 3.5 to 5.5 per cent of total HbA, with 'good' control defined as less than 7 per cent and 'poor' control as more than 8 per cent; some poorly compliant patients have HbA_{1c} concentrations of 14 to 16 per cent. HbA_{1c} measurements are a useful index of medium-term glycaemic control, but may be invalidated by abnormal red cell turnover (values are spuriously low in haemolysis, bleeding, and pregnancy), in renal failure (carbamylated HbA coelutes with HbA_{1c}, falsely raising levels), and with abnormal haemoglobins (HbF also comigrates with HbA_1).

Serum albumin also undergoes glycation, which is measured by the fructosamine reaction. As albumin turns over faster than haemoglobin, the fructosamine concentration reflects mean blood glucose over the previous 1 to 2 weeks. Assays are cheap but not standardized between laboratories, and are generally less reliable and reproducible than measurements of HbA_{1c}.

Measurements of urinary glucose and ketones

Urinary glucose concentrations can be measured easily using glucose oxidase test strips, but are of limited use: urinary glucose concentration depends on the renal threshold (which can lie between 7 and 13 mmol/l), urine output, and the time since the bladder was last emptied. Crucially, hypoglycaemia cannot be detected. Urinary glucose measurements are acceptable in type 2 diabetic patients with a normal renal threshold who are not receiving hypoglycaemic medication (insulin or sulphonylureas) and in patients who decline to prick their fingers.

Urinary ketone measurements can be useful for predicting impending ketoacidosis, particularly during intercurrent illness when blood glucose is high. Moderate ketonuria can be caused by fasting or undereating, including during infections.

Structures for diabetes care

Diabetes is best managed by the combined efforts of a team of specialists with complementary and overlapping skills: physician, specialist diabetes nurse, dietician, and chiropodist. The specialist diabetes nurse has a crucial role in educating patients about diabetes and its practical management, and in starting and adjusting therapy. Many patients are more receptive and responsive to information given by specialist nurses than by doctors. There must be frequent contact with and easy access to other specialists (ophthalmologist, vascular surgeon, renal physican, obstetrician, and clinical psychologist), ideally in the setting of combined clinics. Each member of the team has a particular niche but all must agree common strategies (such as dietary advice for obesity) to avoid giving the patients conflicting or inconsistent information.

Diabetes care can be delivered effectively by hospital-based clinics, community 'miniclinics' run by well-informed general practitioners or practice nurses, or by 'shared care' schemes that involve both primary and secondary sectors. Because of the unpredictable course and potential complications of diabetes, all patients must be thoroughly reviewed each year and be rapidly referred for specialist help if the need arises.

A check list for the annual review is suggested in Table 5.

Diabetes education

Living and coping with diabetes is a considerable burden that is poorly appreciated by many doctors and nurses. Careful education about diabetes, its complications, and its practical management can provide great reassurance to patients and also reduce emergency hospital admissions and complications such as foot ulceration and amputation.

Diabetes education is most effectively provided by the specialist diabetes nurse, but all members of the diabetes care team should understand the key messages, and check and reinforce these whenever possible. These include:

- Causes of hyperglycaemia and diabetic symptoms.
- Own treatment: diet and lifestyle; drawing up and injecting insulin; oral agents; recognizing and treating hypoglycaemia.
- Self-monitoring technique; targets and danger levels; how to respond to poor control.
- 'Sick-day' rules: monitoring during intercurrent illness; how to adjust own treatment; when and how to call for help (see Table 6).

Employment, driving, and insurance

Because of the risk of hypoglycaemia, patients treated with insulin (type 1 or 2) are barred from active service in the police, fire service, or armed forces and from driving heavy-goods or public service vehicles. Specific diabetic complications, notably sight-threatening retinopathy, may preclude particular jobs or pastimes.

Patients must inform the driving licence authorities and their driving insurer that they are diabetic, and those receiving insulin or with clinically significant retinopathy may require periodic medical confirmation of fitness to drive. Frequent hypoglycaemia, especially with decreased awareness of symptoms, is a bar to driving.

Special life insurance policies are available from companies endorsed by patient-centred organizations such as Diabetes UK and the American Diabetes Association.

Many patients find it valuable to join these organizations.

Intercurrent events in diabetes and their management

Infections

Diabetic patients probably have increased susceptibility to pyogenic bacterial infections, especially when diabetes is poorly controlled. Hyperglycaemia can impair the killing of micro-organisms by neutrophils and macrophages and may also interfere with the function of T lymphocytes. Some infections particularly associated with diabetes include:

- Tuberculosis, often widespread and cavitating.
- Necrotizing fasciitis, rapidly spreading necrosis of subcutaneous tissues down to muscle, usually due to b-haemolytic streptocci with staphylococci and often anaerobes.
- Gas-forming infections with anaerobes and clostridia, including 'emphysematous' pyelonephritis, cholecystitis, cystitis, and foot infections. Plain radiography shows gas in the affected tissues.
- Diabetic foot ulcers (see below) are often infected, with the risk of osteomyelitis and deep soft-tissue spread.
- Urinary tract infections may be complicated by ascending infections with pyelonephritis and renal or perinephric abscess (sometimes with gas), and occasionally acute papillary necrosis. Severe loin pain and systemic symptoms, with deteriorating renal function, should suggest these possibilities and the need for urgent imaging.
- 'Malignant' or necrotizing otitis externa, due to pseudomonas infection, can invade the skull and facial nerve.
- Periodontal infections, sometimes causing tooth loss, are common.
- Rhinocerebral mucormycosis is a highly invasive fungal infection that originates in the sinuses but often spreads into the orbit and cranial cavity. Mortality is about 50 per cent, even with debridement and high-dose intravenous amphotericin B.

The bacterial infections often require aggressive intravenous antibiotic treatment with cover against anaerobes. Fastidious and rare organisms should be considered when standard antibiotic regimens are ineffective.

Diabetic control during infections

Minor viral infections rarely disturb diabetic control, but increased secretion of counter-regulatory stress hormones during severe infections, especially with fever, can rapidly worsen insulin resistance in both type 1 and 2 diabetes.

Type 1 patients may need twice as much insulin as usual, even if they are unable to eat. Failure to increase the insulin dosage will therefore allow glucose to rise, sometimes dramatically fast, and risk precipitating ketoacidosis. It is therefore essential to continue taking insulin, to monitor blood glucose frequently, and to increase insulin if sustained hyperglycaemia develops. A rise of 30 to 50 per cent in long-acting insulin is often enough, but requirements will be determined by blood glucose levels and should be decided in consultation with the diabetes care team. Avoidable deaths still occur every year because poorly educated patients (sometimes advised by ignorant doctors) reduce or even stop taking insulin because they feel ill, are not eating, and are worried about becoming hypoglycaemic. Clear 'sick-day' rules (Table 6) are a crucial part of diabetes education, which must be regularly checked and reinforced.

During severe infections, insulin needs may fluctuate rapidly and the safest way to give insulin is by continuous intravenous infusion, backed up by frequent (hourly) blood glucose measurements (see below).

Type 2 patients may similarly lose glycaemic control, and are often best transferred temporarily to subcutaneous or intravenous insulin.

Although not firmly evidence-based, it would seem best to maintain blood glucose between 5 and 10 mmol/l during intercurrent infections.

Myocardial infarction

This is discussed in detail later.

Surgery

Surgery can be hazardous to diabetic patients: the counter-regulatory stress response to surgical trauma can rapidly lead to hyperglycaemia and ketoacidosis, especially in insulin-deficient patients, while poorly controlled diabetes accelerates catabolism and delays wound healing. Moreover, insulin and the sulphonylureas can cause severe hypoglycaemia in fasted or anorexic patients, which can be particularly dangerous during general anaesthesia.

Glycaemic control must therefore be meticulous throughout the preoperative period. A routine management policy should be agreed between the diabetes care team, surgeons, anaesthetists, and ward staff, and this will greatly reduce the risks of operating on diabetic people. Fitness for surgery should be carefully assessed, in view of cardiovascular or other complications. Patients may need to be admitted some days before operation to optimize their treatment.

For type 2 patients who are well controlled by diet or oral agents and undergoing minor surgery only, long-acting sulphonylureas (glibenclamide) should be changed to short-acting ones (for example gliclazide) some days before surgery to reduce the risk of hypoglycaemia. Oral agents and breakfast should be omitted on the morning of operation and blood glucose should be monitored closely. Persistent hyperglycaemia should be treated with the intravenous glucose–potassium–insulin regimen described below.

For all other diabetic patients, subcutaneous insulin should be stopped on the morning of surgery, and a continuous intravenous infusion of balanced amounts of glucose, potassium, and insulin should be given (Table 7). If the patient is in steady state, the glucose–potassium–insulin (GKI) infusion will both maintain satisfactory glycaemic control (5–10 mmol/l) and prevent hypokalaemia. This regimen should be started on the morning of surgery and continued until the patient is able to eat and drink normally, when the usual treatment can be resumed. GKI bags must be changed if glucose levels are unsatisfactory. Alternatively, insulin may be given as a variable-rate intravenous infusion, which provides greater flexibility.

Acute metabolic complications of diabetes and their treatment

Diabetic ketoacidosis

This is uncontrolled hyperglycaemia with hyperketonaemia severe enough to cause metabolic acidosis. It remains a major cause of death in patients with type 1 diabetes under 20 years of age, and episodes still carry an overall mortality of 5 to 10 per cent (50 per cent in elderly patients with diabetic ketoacidosis precipitated by infection or myocardial infarction). Prompt diagnosis and careful management can prevent many deaths.

Causes

Diabetic ketoacidosis only develops when severe insulin deficiency, compounded by an excess of glucagon, stimulates lipolysis and a massive increase in ketogenesis (see above). It therefore almost always occurs in untreated or poorly treated type 1 diabetes and is generally regarded as the hallmark of that disease. However, diabetic ketoacidosis can occur in subjects with type 2 diabetes who are relatively insulin deficient, especially when the secretion of counter-regulatory hormones (especially glucagon) is increased by severe intercurrent illness. Precipitating factors include:

- Newly presenting type 1 diabetes.
- Omission or underdosing of insulin by established type 1 diabetic patients, which may be deliberate in patients with disturbances of body image.
- Intercurrent illness, such as infections, myocardial infarction, stroke, trauma, surgery, and burns. Many patients (and their doctors) fail to increase insulin

dosages or monitor blood glucose during such events.

About 30 to 40 per cent of episodes are unexplained; omitted or inadequate insulin treatment should always be suspected if no obvious infective or other cause is found.

Pathophysiology

Diabetic ketoacidosis is due to the accumulation of ketones, i.e. acetoacetate and its derivatives, 3-hydroxybutyrate (or b-hydroxybutyrate) and acetone (Fig. 8). They are generated by b oxidation of free fatty acids within the mitochondria of the liver. Free fatty acids enter the cytoplasm of hepatocytes and combine with coenzyme A (**CoA**) to form their fatty acyl-CoA derivatives. These are then transported into the mitochondria by the 'carnitine shuttle', a complex of two linked enzymes, carnitine palmitoyl transferase I (CPT-I) on the outer mitochondrial membrane and carnitine palmitoyl transferase II (CPT-II) on the inner. CPT-I, and the overall activity of the shuttle, is powerfully inhibited by insulin and stimulated by glucagon. Once inside the mitochondria, free fatty acids undergo b oxidation to yield ATP (oxidative phosphorylation) and acetyl-CoA. The latter is converted to acetoacetate, which may be oxidized to 3-hydroxybutyrate or undergo condensation to produce acetone.

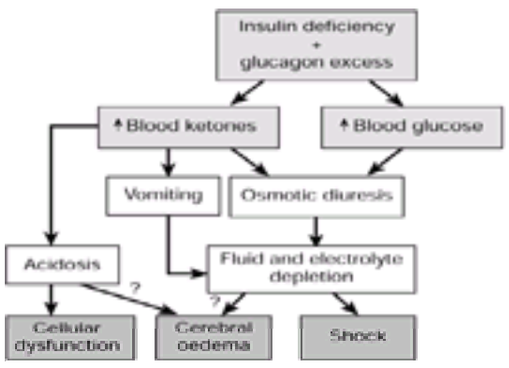

Fig. 8 Pathophysiological changes in diabetic ketoacidosis. Cellular dysfunction induced by intracellular acidosis, cerebral oedema, and shock are potentially life-threatening.

Ketones are transported out of the liver and are used as metabolic fuels by various tissues including the brain; they supply a few per cent of total energy needs after an overnight fast, but the proportion rises to over one-third during prolonged fasting. When produced in excess, they can accumulate rapidly, especially if plasma levels exceed 10 mmol/l (about 50 times normal), when tissue uptake mechanisms become saturated. Ketogenesis is greatly enhanced in uncontrolled type 1 diabetes because of the combination of low insulin with increased glucagon concentrations: lipolysis is unrestrained, and the uptake into liver mitochondria of the increased amounts of fatty acyl-CoA is stimulated by the synergistic effects on CPT-I of high glucagon and low insulin. The main consequences of raised circulating ketone levels are shown in Fig. 8 and listed below:

- Acidosis. Acetoacetate and 3-hydroxybutyrate are both moderately strong organic acids and lower the extracellular pH when the buffering capacity of plasma proteins is exceeded. Ion exchange across cell membranes leads to intracellular acidosis which compromises cellular metabolism because many crucial enzymes operate within a narrow pH range. Clinical measurements of acid–base status are confined to extracellular fluid and may underestimate the severity of intracellular acidosis.
- Diuresis. Ketones are filtered in the urine and are osmotically active. They therefore exacerbate the osmotic diuresis caused by glycosuria and the resulting polyuria, electrolyte losses, dehydration, and hypovolaemia.
- Nausea, through direct stimulation of the chemoreceptor trigger zone in the medulla.

Clinical features

Diabetic ketoacidosis usually presents with classical hyperglycaemic symptoms (Table 2), together with features of acidosis and hyperketonaemia:

- Acidotic (Kussmaul) breathing is deep, sighing hyperventilation which has been mistaken for panic attacks, pulmonary embolism, and left ventricular failure.
- Nausea and vomiting are ominous signs, because dehydration develops quickly in polyuric patients unable to drink.
- Drowsiness and coma occur late and may indicate early cerebral oedema.

The patient generally looks ill and may show postural hypotension and other signs of dehydration and hypovolaemia. Acetone (nail varnish remover) is volatile and may be smelled on the breath. Some patients are hypothermic due to heat loss from peripheral vasodilation, and this may mask the pyrexia of infection. Children with diabetic ketoacidosis often complain of abdominal pain, sometimes mimicking acute appendicitis or other surgical emergencies. A full examination is essential to identify any intercurrent illness.

Investigations and diagnosis

Once suspected, the diagnosis can be confirmed on the spot with a finger-prick blood glucose measurement and urinalysis for ketones. Treatment with intravenous saline and insulin should begin immediately, and baseline investigations carried out. Venous blood is taken for biochemical screening and arterial blood for pH and acid–base status. Additional tests to identify the cause of the episode should include a full blood count, urine and blood culture, chest radiograph, and, especially in older patients, ECG and cardiac enzymes or troponin levels.

Typical values and some diagnostic pitfalls in diabetic ketoacidosis are shown in Fig. 9. Plasma ketone levels are measured by some laboratories but are not usually needed for safe management. High ketone concentrations cause a large anion gap: i.e. plasma $[Na^+ + K^+]$ exceeds $[HCO_3^- + Cl^-]$ by more than 17 mmol/l.

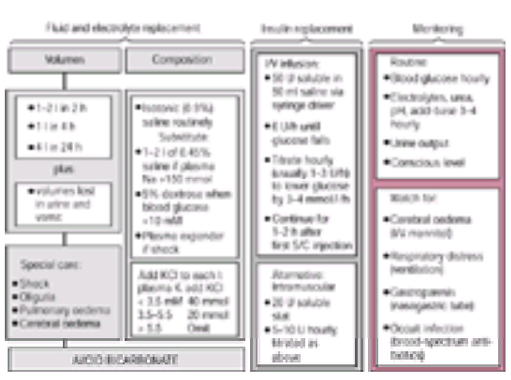

Fig. 9 Guidelines for the management of diabetic ketoacidosis.

Management

Diabetic ketoacidosis is a potentially life-threatening medical emergency that requires urgent treatment with scrupulous clinical and biochemical monitoring: many

avoidable disasters still happen because the patient is abandoned once treatment has been started. Severe diabetic ketoacidosis is best managed initially on a high-dependency or intensive care unit.

The highest priority is to correct hypovolaemia and dehydration, which will often improve acidosis and hyperglycaemia. Insulin replacement must also be started urgently. However, it now appears likely that the high mortality of diabetic ketoacidosis has been partly due to overenergetic replacement of intravenous fluids (especially bicarbonate) and perhaps insulin, which may predispose to the development of cerebral oedema. The treatment guidelines below (Fig. 9) are based on large studies that have reported very low mortality and morbidity.

Fluid replacement

Good intravenous access is crucial: a large peripheral vein may be used but a central venous cannula is safest for severely hypovolaemic patients and for the elderly or those at risk of heart failure, in whom monitoring of central venous pressure is essential.

Most patients recover rapidly with slower fluid replacement than was previously recommended. For those who are not shocked give:

- 1 to 2 litres in 2 h, then
- 1 litre over the next 4 h, then
- 4 litres over the next 24 h.

Fluid losses in urine or vomit should be added to these volumes. Shocked or oliguric patients may require faster fluid repletion, possibly with plasma expanders rather than saline, while slower replacement is safer in those with signs of fluid overload, heart failure, or any suspicion of cerebral oedema. Urine output must be monitored closely, as must blood pressure, central venous filling, and signs of pulmonary or peripheral oedema.

Saline containing potassium is the logical fluid to replace the losses of Na^+, K^+, and Cl^- induced by the osmotic diuresis of diabetic ketoacidosis. The use of intravenous bicarbonate to try to correct acidosis is contentious, both biochemically and in terms of clinical outcome (see below).

Isotonic (0.9 per cent) saline is used initially. Half-isotonic (0.45 per cent) saline has been suggested to replace 1 or 2 litres of isotonic saline, if severe hyperosmolarity (> 350 mosmol/kg) and/or hypernatraemia (> 150 mmol/l) are present. However, the rationale may be flawed: 0.9 per cent ('normal') saline is already hypotonic with respect to the patient's hypertonic plasma, and the use of even more hypotonic solutions would seem likely to exacerbate the intracellular movement of water which may lead to cerebral oedema. Five per cent dextrose is generally substituted when plasma glucose has fallen to 10 to 14 mmol/l to prevent hypoglycaemia (insulin is still required to prevent ketogenesis and promote glucose utilization in the tissues).

Intravenous sodium bicarbonate was previously recommended for severe acidosis. However, the hope that adding alkali will correct acidosis may be oversimplistic. HCO_3^- and H^+ ions (from 3-hydroxybutyric and acetoacetic acids) combine extracellularly to produce H_2CO_3, which dissociates to produce water and CO_2; this may reduce extracellular acidosis, but as cell membranes are impermeable to HCO_3^- ions, the all-important intracellular acidosis is not improved. Indeed, CO_2 can enter cells where it can combine with water to produce H_2CO_3, itself a weak organic acid that can dissociate into H^+ and HCO_3^- ions. Paradoxically, therefore, intravenous bicarbonate administration could worsen intracellular acidosis and there is evidence from animal models of acidosis that this occurs. Worryingly, a recent study identified bicarbonate administration as the most important independent predictor of cerebral oedema in children with moderately severe diabetic ketoacidosis. Another problem with high-strength (8.4 per cent) sodium bicarbonate solution is the intense thrombophlebitis it causes when given intravenously, which can obliterate even large central veins.

The current consensus is that bicarbonate is unlikely to do good but runs the risk of doing harm, and that it should not be used in the treatment of diabetic ketoacidosis

Potassium replacement

Diabetic ketoacidosis always depletes total body K^+ stores to a variable degree because of electrolyte losses through osmotic diuresis, but H^+/K^+ exchange across the plasma membrane encourages K^+ to leak out of cells in acidosis. Plasma K^+ levels can therefore be low, normal, or high, and dangerous hyperkalaemia can be present, especially if severe hypovolaemia causes prerenal failure. During insulin replacement, K^+ is carried intracellularly with glucose, and plasma K^+ levels can fall rapidly. Frequent monitoring of K^+ (every 3 to 4 h initially) is therefore essential in the safe management of diabetic ketoacidosis, and patients with marked K^+ disturbances should have continuous ECG monitoring.

Potassium replacement should be determined by current plasma K^+ levels:

- Add 20 mmol of KCl to each litre of intravenous fluid if K^+ is normal (3.5–5.0 mmol/l).
- Add 40 mmol/l of KCl to each litre if plasma K^+ is less than 3.5 mmol/l.
- Omit KCl if plasma K^+ is more than 5.0 mmol/l, because of the risk of precipitating arrhythmias.

Insulin replacement

Continuous intravenous infusion is the best way to give insulin in diabetic ketoacidosis; subcutaneous and intramuscular absorption are too erratic to be safe and the rate of fall of glucose (one of the factors implicated in cerebral oedema) cannot be easily controlled.

- 50 U soluble insulin (for example Actrapid or Humulin S) should be added to 50 ml isotonic saline (i.e. 1 U/ml) and delivered by a syringe driver pump, either into a separate vein or piggy-backed into the intravenous fluids line.

Because the half-life of insulin in the circulation is only a few minutes, blood glucose and ketone levels will rise rapidly if insulin delivery is interrupted; hourly monitoring of blood glucose is therefore mandatory during intravenous insulin. Failure of glucose to fall usually means that the pump has been turned off or that the infusion cannula is blocked. Initially, 6 U/h (i.e. 6 ml/h) should be given and once blood glucose has started to fall, the rate can then be titrated so that glucose falls by 3 to 4 mmol/l every hour. Faster rates of fall are unnecessary, commonly cause hypoglycaemia, and are thought to predispose to cerebral oedema. Most patients need 1 to 3 U of insulin per hour, and the requirement will become clear after 3 to 4 h of blood glucose onitoring.

A typical intravenous sliding scale (based on hourly glucose measurements) is:

- blood glucose less than 5 mmol/l: give 0.5U/h
- blood glucose 5 to 10 mmol/l: give 2 ml/h
- blood glucose more than 10 mmol: give 4 ml/h.

An alternative to the syringe driver is to dilute insulin into a larger volume (50 U soluble insulin into 500 ml of saline; 0.1 U/ml) and to regulate delivery (for example 20 ml/h (2 U/h)) using an electronic drip counter or a paediatric giving-set with a burette.

The GKI infustion used for perioperative management of diabetic patients is not appropriate because it assumes that the patient is in steady state (which is not the case) and because K^+ disturbances may be exacerbated; moreover, making up and changing GKI infusion bags is time-consuming and very rarely done as often as is needed to control the fall in glucose.

If it is impossible to give a controlled intravenous infusion, then intramuscular soluble insulin can be injected every 4 h or so, starting with 20 U and attempting to titrate subsequent dosages (for example 5–10 U hourly).

Other complications

Intercurrent illness must be treated energetically. Broad-spectrum antibiotics are often given prospectively. Myocardial infarction (see below) has a poor prognosis if it causes diabetic ketoacidosis.

Shock may lead to prerenal failure and sometimes acute tubular necrosis. Plasma expanders and inotropes may occasionally be required for severe hypotension, although rehydration as above is usually adequate.

Cerebral oedema still accounts for 50 per cent of fatalities in diabetic ketoacidosis, especially in children, although modern management protocols with slower fluid replacement and low-dose intravenous insulin infusion can markedly reduce its incidence. The cause is thought to be shifts of ions and water into the brain, particularly the movement of water into dehydrated, hypertonic cells when relatively hypotonic fluids reach the extracellular space. Such shifts would be predicted with isotonic and particularly with hypotonic fluids. Risk factors for cerebral oedema include over-rapid falls in blood glucose, excessive fluid replacement, and high insulin dosages. Insulin can affect various ion transport mechanisms in the brain, but its role remains mysterious and may simply reflect changes in extracellular osmolarity. Interestingly, CT scanning before fluid and insulin replacement has demonstrated subclinical cerebral oedema in children with diabetic ketoacidosis.

Swelling of the brain within the cranium causes coning, leading to cardiorespiratory arrest. It presents as a decline in consciousness, usually rapid and often when the patient's metabolic state has been stabilized. Papilloedema may be present, and CT or MR scanning will show characteristic swelling, with loss of cortical features and squashing of the ventricular system (Fig. 10). It is usually fatal (in more than 90 per cent of established cases), but intravenous mannitol (0.2 g/kg over 30 min, repeated hourly if there is no improvement) may help by raising the osmolality of extracellular fluid and drawing free water out of the brain; there is no firm evidence to support the use of dexamethasone.

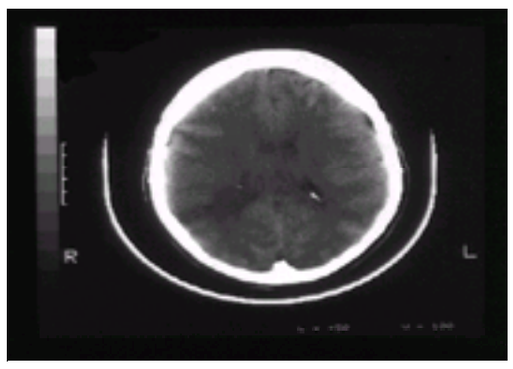

Fig. 10 Cerebral oedema in a patient recovering from diabetic ketoacidosis. The CT scan shows generalized swelling and loss of cortical detail with squashing of the cerebral ventricles.

Adult respiratory distress syndrome is due to accumulation of fluid in the alveoli, perhaps due to ionic and water shifts or to excessive leakiness of the pulmonary capillaries. Hypoxia is severe, and chest radiography shows an appearance like left ventricular failure but with a normal heart size. Risk factors include rapid fluid replacement. It carries a poor prognosis, but ventilation with high-concentration oxygen may be useful supportive treatment.

Acute gastric dilatation (gastroparesis) presents with vomiting and may show a succussion splash and a ground-glass appearance on abdominal radiograph. Nasogastric drainage may be needed to prevent aspiration, especially in the unconscious patient.

Hypothermia indicates a poor outcome. It may respond to rewarming with a space blanket.

Subsequent management

When the patient can eat and drink, intravenous fluids and insulin can be discontinued. There is no need for a GKI regimen; instead, the patient can be restarted on their usual insulin regimen (or on twice daily long-lasting insulin, if newly diagnosed). The intravenous insulin infusion should be maintained until the first injection has had time to act (3–4 h for long-acting insulin alone).

The causes of the episode must be determined if possible, and efforts made to prevent it from happening again. The patient's understanding of diabetes, including the 'sick-day' rules (Table 6), must be checked and reinforced if necessary. Recurrent diabetic ketoacidosis is a feature of brittle diabetes, and these patients need careful monitoring and counselling.

Hyperosmolar non-ketotic state (HONK)

Hyperosmolar non-ketotic state is distinguished from diabetic ketoacidosis by the absence of gross hyperketonaemia and metabolic acidosis. Hyperglycaemia can be greater than in diabetic ketoacidosis and, together with a rise in urea due to dehydration and prerenal failure, may elevate the plasma osmolality to well over 350 mosmol/kg.

Ketosis does not develop because circulating insulin levels are high enough to suppress lipolysis and ketogenesis; these patients are therefore C-peptide positive, with type 2 diabetes which is often previously undiagnosed. It is more common in people of Afro-Caribbean origin. Precipitating factors include myocardial infarction, stroke, infection, and diabetogenic drugs such as glucocorticoids and thiazide diuretics; fizzy glucose drinks may also contribute.

Presentation is typically with classical hyperglycaemic symptoms (polyuria, intense thirst, weight loss, blurred vision), without the features of ketoacidosis. Confusion, drowsiness, and coma are commoner than in diabetic ketoacidosis.

Complications include thrombotic events such as stroke and peripheral arterial occlusion, and deep venous thrombosis and pulmonary embolism, these being due to increased blood viscosity. Mortality exceeds 30 per cent because these patients are old and often have a serious precipitating illness.

Biochemical features of hyperosmolar non-ketotic state are:

- Hyperglycaemia: often over 50 mmol/l, sometimes over 90 mmol/l.
- Hypernatraemia: often over 155 mmol/l (may be artefactually depressed by high glucose levels).
- Uraemia due to dehydration, with or without renal failure.
- Hypersmolality: over 350 mosmol/kg.
- Blood and ketone levels are normal or only slightly raised (usually through anorexia).
- Arterial pH, venous bicarbonate, and anion gap show no features of severe acidosis.

Management is largely as for diabetic ketoacidosis:

- Saline replacement must be particularly cautious in older patients, in whom cardiac disease is common. Half-isotonic (0.45 per cent) solution is often given if plasma sodium exceeds 150 mmol/l or osmolality exceeds 350 mosmol/kg; the rationale for preferring this to isotonic saline is not proven, but the risks of cerebral oedema appear to be lower than in diabetic ketoacidosis.
- Potassium levels must be carefully monitored and replaced as above.

- Intravenous insulin infusion at low doses rapidly controls hyperglycaemia in most cases.
- Low-dose heparin (5000 U subcutaneously 8-hourly) should be given prophylactically, but full anticoagulation should be reserved for proven thromboembolism as the risks of fatal gastrointestinal bleeding are high. Intercurrent illness must be sought and treated appropriately.

After recovery, many of these patients can be successfully weaned off insulin. Drugs and other precipitating factors must be identified and avoided if possible.

Lactic acidosis

Lactate is generated by glycolysis and its levels rise rapidly during tissue anoxia (for example during shock, cardiac failure, or pneumonia) or when the liver is prevented from utilizing it as a gluconeogenic substrate (for example in hepatic impairment). Lactic acidosis is best known in diabetic patients as a rare but often fatal complication of the biguanides, phenformin and metformin, which act mainly by inhibiting hepatic gluconeogenesis. The risk is about 10 times higher with phenformin than with metformin, and it is very rare during metformin treatment as long as other predisposing factors (the major organ failures) are avoided.

Lactic acidosis presents as coma with metabolic acidosis (reduced arterial pH and venous bicarbonate) and a wide amino gap due to hyperlactataemia. Blood glucose levels are usually raised.

Treatment is still unsatisfactory. Intravenous sodium bicarbonate may paradoxically aggravate intracellular acidosis, although forced ventilation to blow off carbon dioxide may help (see above). Haemodialysis may both clear lactate and hydrogen ions, and correct any sodium overload following bicarbonate administration. Sodium dichloroacetate, which stimulates pyruvate dehydrogenase to metabolize lactate, is undergoing evaluation.

Mortality remains high (over 30 per cent), partly because of the organ failures that commonly coexist.

Hypoglycaemia

Hypoglycaemia is an inevitable side-effect of antidiabetic drugs that raise circulating insulin levels, namely insulin itself and sulphonylureas; it does not occur with metformin or thiazolidinediones alone, or with dietary restriction. Common contributory factors are:

- Accelerated insulin absorption, for example due to exercise or hot surroundings.
- Unfavourable timing of insulin injection: injecting too soon before eating can cause late postprandial hypoglycaemia, while long-acting insulins injected in the early evening often cause nocturnal hypoglycaemia.
- Too much insulin injected: dosage errors are quite common, particularly in the elderly.
- Inadequate food intake: missed, delayed, or small meals; vomiting, including gastroparesis.
- Exercise hastens insulin absorption while enhancing insulin action; delayed hypoglycaemia may occur many hours later because muscle continues to take up glucose to replenish glycogen.
- Alcohol inhibits hepatic gluconeogenesis, preventing the increase in hepatic glucose output that is crucial for restoring euglycaemia.
- Impaired awareness of early warning symptoms (see below).

Progressively more frequent or severe attacks may be caused by various conditions, which should always be sought:

- Weight loss, including anorexia nervosa and appetite disorders (relatively common in young women with type 1 diabetes).
- Loss of counter-regulatory hormones: Addison's disease, hypothyroidism, hypopituitarism, blunted glucagon secretion in long-standing type 1 diabetes.
- Intestinal malabsorption, notably coeliac disease (commoner in type 1 diabetes).
- Renal failure, which impairs the clearance of insulin.
- Deliberate inappropriate injection of insulin, often in the context of 'brittle' diabetes.

Manifestations

Clinical features of hypoglycaemia are due to an autonomic discharge, predominantly sympathetic, together with the cerebral effects of neuroglycopenia. Falling glucose levels are sensed by glucose-sensitive neurones, which are found in the periphery (vagal sensory endings in the portal vein) and medulla as well as the hypothalamus. This triggers a powerful sympathetic discharge that releases adrenaline from the adrenal medulla and noradrenaline from sympathetic nerve endings, causing the familiar 'flight or fight' response. Features include pallor (cutaneous vasoconstriction), sweating, tremor (a b_2-adrenergic effect on skeletal muscle) and tachycardia; systolic blood pressure rises due to increased cardiac output while pulse pressure widens—giving the typical bounding pulse—because b_2-mediated vasodilatation in skeletal muscle causes peripheral resistance to fall.

Hypoglycaemia also triggers the secretion of counter-regulatory hormones, namely glucagon and adrenaline (both crucial to restoring euglycaemia), growth hormone, and cortisol. Collectively, these inhibit insulin secretion and raise blood glucose by enhancing hepatic glycogenolysis and gluconeogenesis, causing glucose to pour out of the liver. Defects in glucagon or adrenaline release (which occur for example in long-standing type 1 diabetes), or in the ability of the liver to produce glucose (for example the presence of ethanol which inhibits gluconeogenesis, or a recent glucagon injection which depletes liver glycogen) will delay recovery of blood glucose.

The physiological and neurological features of hypoglycaemia usually develop in a fixed sequence when blood glucose is lowered in a controlled fashion in the laboratory. However, this hierarchy may not be apparent in real life, and some patients specifically lose their awareness of the early warning symptoms (see below). Key events as glucose falls are:

- ~3.8 mmol/l: increased glucagon and adrenaline secretion
- ~3.0 mmol/l: onset of hypoglycaemic symptoms
- ~2.8 mmol/l: neuroglycopenia and cognitive impairment
- at less than 1.0 mmol/l: coma.

Symptoms of hypoglycaemia

The symptom complex can be extremely variable, and hypoglycaemia should be suspected as the cause of any 'funny turn' in patients treated with insulin or sulphonylureas. Autonomic manifestations include sweating, tremor, tachycardia, and hunger, while neuroglycopenia can cause drowsiness, confusion, inco-ordination, dysarthria, and automatic or disinhibited behaviour; distinct neurological deficits include aphasia, diplopia, and hemiparesis. Non-specific malaise and headache afterwards are also common. Nocturnal episodes may pass completely unnoticed by the patient, or may cause sweating and restlessness (often obvious to the patient's partner), vivid nightmares, or a hung-over feeling the following morning.

Awareness of hypoglycaemic symptoms

Diabetic patients rely on the early autonomic symptoms (sweating, shaking, and hunger) to warn them of an impending hypoglycaemic attack, when corrective action can be taken. In some patients the early warning symptoms are attenuated or not noticed at all; this clumsily named 'hypoglycaemia unawareness' is potentially dangerous because severe neuroglycopenia (confusion, fitting, irrational behaviour, coma) may suddenly incapacitate the patient. Reduced awareness of hypoglycaemia occurs particularly in two settings, which may coexist:

- Long-standing type 1 diabetes. Some 30 to 50 per cent of patients with diabetes of more than 20 years' duration have decreased awareness of symptoms, and many also show a flat glucagon response to hypoglycaemia. Blunted recognition of hypoglycaemia by the central nervous system may be responsible.
- Excessively tight glycaemic control impairs awareness of hypoglycaemia; for unknown reasons, even a single episode can blunt perception of symptoms and counter-regulatory hormone responses for some days. Conversely, relaxing control and avoiding hypoglycaemia completely for several weeks can partially restore awareness of warning symptoms.

The use of human insulin has been suggested to impair awareness of hypoglycaemic symptoms. Human insulin is relatively lipophilic—hence its faster subcutaneous absorption—which could theoretically promote its entry into the brain. Insulin may act directly on the brain to affect various autonomic processes, but detailed comparisons of human and animal insulins, both in the laboratory and in real life, have not shown any species differences in counter-regulatory responses or the intensity of hypoglycaemic symptoms.

Sequelae of hypoglycaemia

Even the most dramatic neurological manifestations of acute hypoglycaemia—including aphasia, hemiparesis, fitting, and unconsciousness—usually resolve rapidly when blood glucose is normalized. Recovery from profound coma may take many hours or even days, and this is probably due to cerebral oedema. Patients who survive severe and prolonged hypoglycaemic coma may show permanent neurological damage, including memory loss, aphasia, and a vegetative state. There are concerns that repeated mild attacks, especially in children and perhaps particularly at night, can cause cumulative intellectual impairment, but this is not yet proven.

Severe hypoglycaemia has been implicated in precipitating myocardial infarction or stroke, particularly in the elderly; rises in blood pressure and increased coagulability of the blood following sympathetic stimulation may contribute. Like any convulsions, hypoglycaemic fits may cause injury, including limb and vertebral crush fractures.

Prolonged severe hypoglycaemia can be fatal and is one of the most common causes of death in young type 1 patients. Post-mortem studies show neuronal damage and necrosis in the hippocampus and cerebral cortex. Hypoglycaemia has been suspected as a cause of death in patients found unexpectedly dead in bed; arrhythmias may be responsible.

Diagnosis and detection of hypoglycaemia

Hypoglycaemia is easy to diagnose but is also easily missed; differential diagnoses include transient ischaemic attacks, psychosis, drunkeness, epilepsy, and migraine. Symptoms may be instantly recognizable to some patients, but may present atypically. If suspected, the blood glucose levels should be checked, taking care to avoid under-reading artefacts with reagent test strips. Urinalysis is obviously of no use—hence all patients receiving insulin or sulphonylureas must be able to check their blood glucose. The patient's close associates should also know how to diagnose and treat hypoglycaemia.

Various experimental hypoglycaemia detectors are undergoing development, including measurements of subcutaneous glucose by implanted sensors or transcutaneous near-infrared spectroscopy, but none is yet suitable for routine clinical use.

Prevention and treatment of hypoglycaemia

Insulin-treated patients fear hypoglycaemia as much as blindness or renal failure, and this may prevent them from tightening their diabetic control as much as their doctors would prefer. Many doctors underestimate the impact of hypoglycaemia; asking about it and trying actively to prevent it are an essential part of diabetes care.

Attention to the factors listed above should help to reduce the frequency and severity of attacks. Advice about exercise, moderating alcohol intake, and timing of insulin injections and meals are particularly important. Nocturnal hypoglycaemia can be reduced by checking the blood glucose at bedtime, and by taking long-acting carbohydrate (for example bread or cereal) if the level is less than 6 mmol/l.

Blood glucose levels of less than 3 mmol/l should be treated immediately (Table 8). Oral glucose or sucrose or other carbohydrate should be given if the patient can swallow safely. Give 20 to 30 g (e.g. six Dextrosol tablets or 150 ml Lucozade) initially; if possible, check blood glucose 15 min later and repeat if the glucose has not risen. Taking too much carbohydrate—which is understandable, given the unpleasantness of hypoglycaemia—can cause marked rebound hyperglycaemia.

If the patient is unconscious, give either:

- Glucagon 1 mg (0.5 mg in children), subcutaneously or intramuscularly; with either route glucose should rise within 10 to 15 min. Side-effects of glucagon include malaise, nausea, and abdominal discomfort and, because it acts primarily by breaking down hepatic glycogen (a limited resource), a second injection may be ineffective.
- Intravenous glucose: 15 to 20 g intravenously, as 50 per cent or 10 per cent solution (the former may cause painful thrombophlebitis, even if given into a large vein).

Glucose gels or jam can be smeared inside the mouth and cheeks in the unconscious patient, but these alone are unlikely to correct serious hypoglycaemia.

On recovery, blood glucose should be checked and oral glucose given as above. Slow recovery from coma may be due to cerebral oedema, which has a high mortality (around 10 per cent) but may respond to intravenous mannitol and forced ventilation with high inspired oxygen concentration.

Once the episode is treated, its cause must be identified if possible and corrective action taken to prevent it from happening again.

Chronic complications of diabetes

Long-term tissue damage is now the major burden of the disease, the greatest source of fear for diabetic people, and the most expensive item in the diabetes health-care budget. The list of complications is depressingly long but fortunately at least 40 per cent of diabetic patients escape clinically significant complications, and improved diabetes care should reduce the risks even further.

Microvascular complications—retinopathy, nephropathy, and neuropathy—are specific to diabetes and reflect damage inflicted on the microcirculation throughout the body. Retinopathy and nephropathy are obviously 'microvascular' disorders; the microcirculation of nerves (vasa nervorum) is also damaged in diabetic neuropathy, although other functional and structural abnormalities in the nerves themselves probably contribute. Macrovascular disease is simply atherosclerosis. This causes typical coronary heart disease, stroke, and peripheral arterial disease, but often behaves more aggressively than in non-diabetic people.

Other complications are due to irreversible biochemical and structural changes in tissues chronically exposed to hyperglycaemia. These include cataract, whose formation during normal ageing is accelerated by diabetes, and specific soft tissue disorders such as limited joint mobility (diabetic cheiroarthropathy).

Causes of chronic diabetic complications

Role of hyperglycaemia

Tissue lesions are identical in all types of diabetes, indicating that hyperglycaemia (or a closely related metabolic abnormality) is likely to be responsible. Microvascular disease in the retina, kidneys, and nerves is generally determined by the severity and duration of hyperglycaemia, although individual susceptibility varies considerably. By contrast, macrovascular disease does not display a clear dose–response relationship with hyperglycaemia: instead, the risk is increased above glucose values that lie below the 'diabetic' range (see above).

Recent intervention studies have confirmed that improving glycaemic control is rewarded by partial protection against microvascular complications but not atheroma. This principle is valid for both type 1 and type 2 diabetes, and is now embodied in their treatment targets (Table 3). Two 'landmark' studies are generally cited, although several smaller ones have also reached the same conclusion.

Type 1 diabetes

The Diabetic Control and Complications Trial (DCCT) was a 12-year North American study of over 1400 patients that compared 'intensive' insulin treatment (aiming

for an HbA_{1c} of 6 per cent) with 'conventional' (i.e. bad) regimens of once or twice daily injections (HbA_{1c} about 9 per cent). Intensive treatment consisted of at least three daily injections or the insulin pump (continuous subcutaneous insulin infusion), and achieved a mean HbA_{1c} of 7 per cent.

The DCCT concluded that improved glycaemic control reduced the risks of microvascular complications . In subjects who were initially free of complications, intensified treatment for 9 years decreased the prevalence of a defined degree of background retinopathy by 70 per cent (i.e. from 55 per cent with conventional treatment to 15 per cent), while the risks of developing microalbuminuria or clinical neuropathy fell by 33 per cent and 70 per cent respectively (Fig. 11). In subjects who already had background retinopathy at baseline, intensified treatment reduced the overall progression of retinopathy by 50 per cent; more importantly the risks of suffering sight-threatening retinopathy or requiring laser treatment were reduced by a similar degree. The development of clinical nephropathy (overt albuminuria) and neuropathy were each decreased by about 60 per cent. By contrast, intensified insulin treatment did not reduce the prevalence of macrovascular disease.

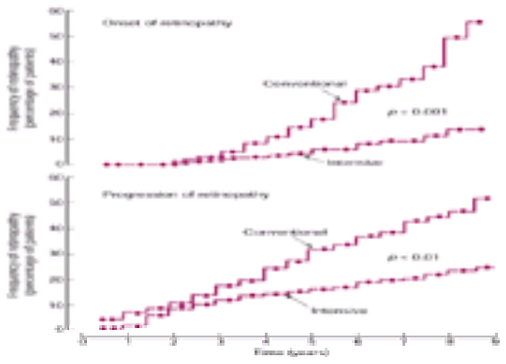

Fig. 11 Intensive insulin therapy and improved diabetic control reduces the risks of type 1 diabetic patients developing retinopathy (upper panel) and of established retinopathy progressing (lower panel). Data from the Diabetic Control and Complications Trial (DCCT).

Type 2 diabetes

The United Kingdom Prospective Diabetes Study (UKPDS) was guided through its 20-year course by the late Robert Turner, who died shortly after it was completed. This huge trial followed the outcome of over 5000 patients treated with diet and lifestyle alone (termed 'conventional' treatment), or together with sulphonylureas, metformin or insulin; confusingly, sulphonylureas and insulin treatments were both described as 'intensive' treatment. The trial confirmed the real-life difficulty of achieving good glycaemic control, especially against the progressive deterioration of type 2 diabetes: very few patients achieved and maintained the 'intensive' target fasting plasma glucose of 6 mmol/l. The trial has been criticised for its convoluted design (which diluted its statistical power) and both the lumping and splitting of data for outcome analysis. Nevertheless, it yielded useful messages about the importance of treating both hyperglycaemia and hypertension and about the natural history of the disease itself. Its conclusions were broadly similar to those of the DCCT: improved glycaemic control decreased the risk of microvascular complications. Lowering HbA_{1c} from 7.9 per cent (conventional) to 7.0 per cent (intensive) decreased the lumped rate of microvascular events by 25 per cent (Fig. 12), including sight-threatening retinopathy (20 per cent) and development of microalbuminuria (33 per cent). Across a reasonably wide range of HbA_{1c}, lowering HbA_{1c} by 1 per cent reduced the risk of microvascular disease by about one-third. Improved glycaemic control had no overall effect on macrovascular disease, although metformin treatment only significantly decreased cardiovascular events (see above).

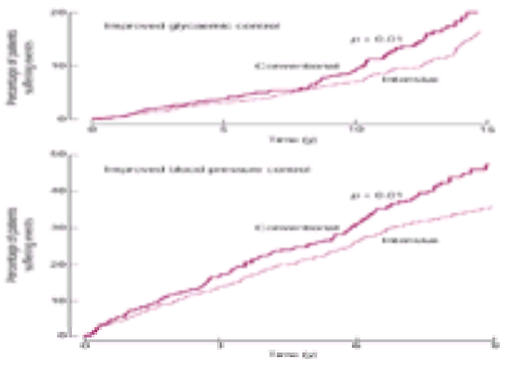

Fig. 12 Benefits of improving glycaemic and blood-pressure control in type 2 diabetic patients. Upper panel *:* 'intensive' treatment with glucose-lowering drugs reduces the risks of suffering any microvascular complication by about 25 per cent. Lower panel: tighter control of blood pressure reduces the risks of suffering any diabetes-related complication (including micro- and macrovascular disease) by about 24 per cent. Note the scale differences in both axes. Data from the United Kingdom Prospective Diabetes Study (UKPDS).

Possible mechanisms of hyperglycaemic tissue damage

High glucose levels can damage the function and structure of many tissues. The mechanisms currently thought most relevant to human diabetic complications probably operate to different degrees in different tissues.

Glycation of proteins and macromolecules

Glycation begins with the non-enzymatic combination of glucose and other reactive sugars with amino groups of proteins, and with acceptor groups of other long-lived macromolecules such as nucleic acids. Glycation is initially reversible, yielding a Schiff base which undergoes molecular rearrangement to form an Amadori product. Amadori products then undergo further reactions, including covalent crosslinking with the sugar groups in other glycated proteins. These irreversibly modified molecules, collectively termed 'advanced glycation endproducts', resist normal degradation mechanisms and thus accumulate.

Advanced glycation endproducts (AGE) can interfere with tissue structure and function in several ways. Stiffening of connective tissue in the limited joint mobility syndrome (see below) is related to crosslinking of collagen by AGE, while the same process in the proteins of the lens fibre (crystallins) causes cataract. AGE also damage blood vessels, and formation of AGE in the basement membrane increases vascular permeability. AGE in the arterial wall may bind low-density lipoprotein and promote atherogenesis. Curiously, endothelial cells carry specific receptors for AGE (so-called 'RAGE)', the binding of which generates oxygen free radicals that may induce oxidative damage and favour coagulation.

Overactivity of the polyol pathway

Polyols are sugar alcohols formed from their respective sugars (for example sorbitol from glucose) under the action of aldose reductase, the rate-limiting enzyme of the polyol pathway. This enzyme is expressed in various tissues susceptible to diabetic complications, notably the retina, glomerulus, lens epithelium, and Schwann cells of the nerves.

Glucose is preferentially shunted through the polyol pathway under hyperglycaemic conditions, generating sorbitol which is poorly diffusible and therefore accumulates intracellularly. This, together with reciprocal intracellular depletion of myoinositol (another polyol, involved in phosphatidylinositol metabolism) may lead to activation of protein kinase C (see below) and the production of highly reactive sugars that can glycate proteins. Increased glucose flux through the polyol pathway

also generates oxygen free radicals and can deplete antioxidants which normally mop up free radicals.

At present, the importance of the polyol pathway in humans remains uncertain, and aldose reductase inhibitors (for example sorbinil and ponaltrestat) have failed to show any convincing benefits in human microvascular complications.

Protein kinase C activation

This enzyme is stimulated by diacylglycerol, which is generated intracellularly in hyperglycaemia. Protein kinase C may mediate adverse effects such as increased vascular permeability and enhanced basement membrane synthesis, although the mechanisms remain obscure.

Abnormal microvascular blood flow

Diabetes interferes with blood flow through the microcirculation, potentially impairing the supply of nutrients and oxygen to the tissues. Resting blood flow is increased in the retina, glomerulus, and other tissues, apparently in response to hyperglycaemia; this may damage the endothelium, favouring thrombogenesis (diabetes also enhances the coagulability of the blood) and perhaps the release of vasoconstrictors such as the endothelins, which may cause microvascular occlusion.

Other factors

Individual susceptibility to microvascular and macrovascular complications varies widely, to a degree that is not entirely explicable by differences in hyperglycaemia. Other risk factors include hypertension, which predisposes to atheroma and is also crucial in determining the rate of deterioration of diabetic nephropathy (see Chapter 20.10.1) and perhaps retinopathy. Smoking is implicated in retinopathy and nephropathy as well as macrovascular disease. Familial clustering of markers for nephropathy has been reported (Chapter 20.10.1), but no convincing candidate genes have yet emerged.

Diabetic eye disease

Eye complications are greatly feared by diabetic patients, with good reason: in the United Kingdom and most westernized countries, diabetes (especially diabetic retinopathy) is the most common cause of blindness in people of working age. Annual screening is advisable with prompt referral for laser treatment if appropriate.

Diabetic retinopathy

This is an easily demonstrated example of the microvascular damage that diabetes inflicts throughout the body. The retina is particularly vulnerable because of its high metabolic and oxygen demands and its dependence on an intact blood–retinal barrier; moreover, small lesions that would pass unnoticed in other vascular beds can have a devastating impact on the patient and his or her quality of life. The Plate section for Section 25 shows the stages and lesions of retinopathy.

Epidemiology

Minor 'background' changes, especially the characteristic microaneurysms, are very common in type 1 patients. Microaneurysms begin to appear after 5 years, affecting about 50 per cent of cases at 10 years and virtually all after 20 years. By contrast, the formation of new vessels that defines proliferative retinopathy emerges after 10 years, reaching a plateau at about 40 per cent of all cases after 20 years. The incidence of maculopathy follows a similar curve, ultimately affecting 10 to 20 per cent of cases (more in older subjects). These different patterns suggest that distinct processes are responsible, and that susceptibility to neovascularization and maculopathy may be determined by factors additional to hyperglycaemia.

In type 2 patients, background changes and sometimes maculopathy and proliferative retinopathy may be present at diagnosis, consistent with the generally long duration of subclinical hyperglycaemia.

All grades of retinopathy can complicate any type of diabetes of sufficiently long duration, with some provisos. Retinopathy may be slow to appear in the mildly hyperglycaemic variants of MODY, while some racial groups (for example Native Americans and Afro-Caribbeans) appear more susceptible. The sexes are equally affected.

Aetiology and pathogenesis

Progression of retinopathy is generally related to the severity and duration of hyperglycaemia, while lowering blood glucose can slow or even prevent the process (see above). Hyperglycaemia damages the retinal vessels in various ways; glycation of key proteins and overactivity of protein kinase C appear to be more important than abnormalities of the polyol pathway.

When differences in glycaemic control are allowed for, there remains considerable individual variability in susceptibility. Genetic factors appear less important than in nephropathy (see above), while hypertension and possibly cigarette smoking may accelerate progression.

Increased vascular permeability, which leads to macular oedema and hard exudates, is an early abnormality, demonstrable by fluorescein angiography. Likely causes include glycation and other changes in the basement membrane of the microvessels, abolishing the negative charge which normally repels plasma proteins such as albumin, and endothelial cell damage, which opens up the tight intercellular junctions that constitute the blood–retinal barrier. Local production of vascular endothelial growth factor, which enhances permeability, may also contribute. Fallout of pericytes, the specialized contractile cells that enclose the capillaries, may weaken the capillary wall, increasing retinal blood flow and leading to the formation of microaneurysms. Increased retinal blood flow, perhaps following pericyte loss and hyperglycaemia *per se*, may cause endothelial damage and thrombogenesis and also enhance protein extravasation. Capillary occlusion is probably due to the formation of microthrombi following endothelial damage and diabetes-related changes in coagulability. Closure produces areas of capillary non-perfusion, which can be surprisingly widespread when shown by fluorescein angiography, and ultimately foci of ischaemia where the retina may infarct (causing cotton-wool spots) and angiogenesis may be stimulated.

New vessel formation is thought to be stimulated by growth factors released by ischaemic tissues, which cause endothelial cells to proliferate. A currently favoured angiogenic factor is vascular endothelial growth factor, a 46 kDa dimeric protein which is expressed, together with its receptors, by retinal endothelial cells; its expression is enhanced by hypoxia and its effects include both increased vascular permeability and endothelial cell proliferation. New vessels sprout initially as solid buds of endothelial cells that later canalize. They can grow within the retina, forward into the vitreous, or across the iris, and are indirectly responsible for the main vision-threatening complications of diabetic retinopathy. They are fragile and rupture easily, causing retinal, preretinal (subhyaloid), vitreous, or anterior chamber haemorrhages, while the fibrous tissue that proliferates around them can cause retinal traction and detachment and lead to glaucoma.

Lesions, clinical stages, and natural history

The stages of diabetic retinopathy are summarized in Fig. 13, and the lesions are illustrated in the Plates for Section 25.

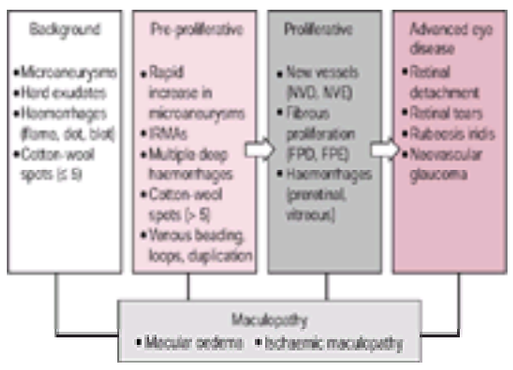

Fig. 13 Stages of diabetic retinopathy. Maculopathy may develop at any stage. IRMAs, intraretinal microvascular abnormalities; NVD/NVE, new vessels on the disc/elsewhere; FPD/FPE, fibrous proliferation on the disc/elsewhere.

Background retinopathy

Individual lesions may appear and regress but their total density tends to increase with lengthening duration of diabetes. Vision is not damaged unless maculopathy coexists; over 50 per cent of patients do not progress beyond this stage.

Microaneurysms are outpouchings of capillaries, perhaps representing ballooning of the weakened capillary wall or endothelial buds attempting to revascularize ischaemic retina. They appear as tiny red dots. Microaneurysms are not fixed features: about 50 per cent disappear within 3 years. The sudden appearance of numerous microaneurysms often indicates worsening retinal ischaemia and can herald preproliferative and proliferative changes.

Hard exudates are due to the precipitation in the retina of lipoproteins and other circulating proteins that escape from abnormally leaky retinal vessels. They are yellow-white spots or streaks with a waxy or shiny appearance, and often form clusters or arcs (a circinate pattern) around the macula and foci of capillary leakage. Like microaneurysms, their distribution and extent can vary markedly with time.

Haemorrhages are due to the rupture of weakened capillaries, their size and shape depending on their situation. Small 'dot' and larger 'blot' haemorrhages are spheroidal because they are contained within the densely packed deeper layers of the retina, whereas 'flame' haemorrhages track along nerve fibre bundles in the more superficial layers. Haemorrhages outside the retina (preretinal or vitreous) generally originate from new vessels and therefore indicate proliferative change.

Maculopathy

Disease of the macula, serious enough to affect central vision, can accompany any stage of diabetic retinopathy including background, and may be present in newly diagnosed type 2 patients.

Macular oedema is due to extravasation of plasma proteins across abnormally leaky capillaries. It may cause only retinal thickening which may be undetectable by routine fundoscopy, even when advanced enough to reduce visual acuity. Exudates, often circinate, and spotty 'cystoid' changes may occur.

Ischaemic maculopathy is the result of extensive capillary closure and can cause severe central visual loss. As with macular oedema, fundoscopy may appear deceptively normal; the macula may simply look featureless.

Maculopathy presents as progressive and painless loss of central vision. Testing of visual acuity is important in routine screening for maculopathy: poor acuity with no obvious explanation (for example cataract, vitreous haemorrhage) must always prompt further investigations for macular oedema or ischaemia. Retinal thickening can be identified easily by slit-lamp examination, while fluorescein angiography will demonstrate both ischaemic areas (hypofluorescent) and sites of vascular leakage (hyperfluorescent).

Preproliferative retinopathy

This stage indicates worsening retinal ischaemia which, if left untreated, often leads to the formation of new vessels. It is defined by one or more of the following, of which intraretinal microvascular abnormalities and venous beading are the most ominous:

- Multiple deep round haemorrhages, especially when appearing over a short period.
- Multiple (more than five) cotton-wool spots, which are due to the accumulation of axoplasm at the edges of retinal infarcts. They appear as dead-white patches with vague borders.
- Intraretinal microvascular abnormalities (IRMA) are flat clusters of abnormal capillaries which, unlike new vessels, are confined to the retina and do not leak fluorescein.
- Venous abnormalities include dilatation (due to general retinal hyperaemia and the shunting of blood around infarcted or non-perfused areas), beading, looping, and reduplication. Beading probably represents the terminations of occluded capillaries, while looping and reduplication may be due to local diversion of blood flow.
- Arterial abnormalities include occlusion, when the vessel is reduced to a thin white line.

Proliferative retinopathy

New vessels appear as fine fronds or arcades of abnormal structure, commonly arising on the optic nerve head ('new vessels disc' or 'NVD') or 'elsewhere' ('NVE'), especially at the bifurcation of veins. Greyish fibrous tissues and haemorrhages may be found in association.

Proliferative retinopathy threatens vision through the complications of the abnormal new vessels, namely haemorrhage, retinal detachment, and glaucoma. Overall, only 10 per cent of untreated patients retain useful vision after 10 years. New vessels on the disc (NVD) carry the worst prognosis: if left untreated, the chances of becoming blind within 5 years are over 50 per cent (compared with 30 per cent for new vessels elsewhere (NVE)). Vitreous haemorrhages tend to recur, and 30 per cent of eyes are blind within 1 year of the first bleed. Fortunately, laser photocoagulation has revolutionized the outlook for proliferative retinopathy.

Advanced diabetic eye disease

This represents endstage damage that commonly leads to blindness; vitreoretinal surgery has improved the prognosis somewhat:

- Vitreous and preretinal haemorrhages develop when new vessels grow forward from the retina, cross the potential preretinal (subhyaloid) space, and enter the vitreous. These vessels rupture easily: associated fibrous tissue contracts and tears them, as does the normal shrinkage of the vitreous with age. Vitreous haemorrhages appear as reddish or dark opacities that may completely fill the eye and block the view of the retina. Preretinal (subhyaloid) haemorrhages have a flat top (if the subject has been upright), because the red cells sediment within the haemorrhage cavity.
- Retinal detachment occurs: the retina is pulled off the underlying choroid by contracting strands of fibrous tissue associated with the formation of new vessels or previous vitreous haemorrhages. The retina may appear wrinkled (traction lines) or thrown into folds or bumps, sometimes with a visible tear.
- New vessels grow on to the iris (rubeosis iridis), usually in the context of widespread proliferative retinopathy. Vessels and diffuse reddening of the iris may be seen with the ophthalmoscope. The main complication is glaucoma, caused by proliferating fibrovascular tissue obstructing the filtration angle in the anterior chamber. Signs include circumcorneal injection, a fixed irregular pupil and corneal haze; the eye is often intensely painful.

Symptoms of diabetic retinopathy

There may be no visual symptoms, even with extensive proliferative changes, until sight-threatening complications occur:

- Vitreous haemorrhage and retinal detachment cause sudden loss of vision that is painless but often terrifying. Retinal detachment occurring behind a vitreous haemorrhage may be reported by the patient as a further deterioration in already poor vision.
- Maculopathy presents as a gradual painless decline in central vision, which may not be noticed by the patient but is picked up on routine eye screening.
- Rubeosis and particularly neovascular glaucoma cause worsening vision with pain and redness in the eye.

Other causes of visual loss need to be considered in diabetic patients, including cataract, stroke, retinal artery or vein occlusion, and hypoglycaemia and glaucoma.

Examination of the eyes in diabetic patients

This should be performed routinely on diagnosis, annually thereafter (or every 6 months if marked background or other changes are present), and immediately if the patient reports any change in vision. There should be close liaison with the ophthalmologist, and a low threshold for referral: indications for seeking expert advice are shown in Table 9.

Visual acuity must be checked with a Snellen chart, with the pupils undilated, and both uncorrected and corrected for refraction errors (with the patient's spectacles or a pinhole). Poor visual acuity (worse than 6/12) that is not correctable and has no other obvious cause (cataract, vitreous haemorrhage) is usually due to maculopathy, and this must be actively excluded (see below).

The iris and pupil are examined for evidence of rubeosis or glaucoma. Pupillary reflexes should be checked: an afferent pupillary defect (see Section 25) indicates severe retinal or optic nerve disease, such as retinal detachment.

Fundoscopy must be used to check both eyes through fully dilated pupils: peripheral new vessels may otherwise be invisible. Relative contraindications to mydriatics are intraocular lens implants; referral to the ophthalmologist is then advisable. The disc, entire retina, and macula must be carefully scanned. Binocular ophthalmoscopy provides good all-round and three-dimensional views, especially of vitreous haemorrhage and retinal detachment.

Additional specialist investigations include:

- Retinal photography: including 'non-mydriatic' cameras that allow photography of most of the retina, through partly dilated pupils (in a darkened room). These are widely used in community screening for retinopathy.
- Slit-lamp microscope: useful for examining the anterior chamber (for rubeosis and glaucoma) and assessing retinal thickness (for detecting macular oedema).
- Fluorescein angiography: fluorescein injected intravenously binds to albumin and so only escapes outside abnormally permeable vessels; sites of leakage are highlighted by persistent fluorescence when the retina is photographed under ultraviolet light. This is useful for showing the foci of leakage (for example for targeting laser photocoagulation) and for the diagnosis of macular oedema.
- B-scan ultrasound provides a cross-sectional image of the eye and can show retinal detachments that are invisible on fundoscopy because of a dense vitreous haemorrhage or cataract.

Management of diabetic retinopathy

Specific treatments for sight-threatening retinopathy have improved greatly, but general preventative measures are still crucial. These include:

- Tight glycaemic control, as highlighted by the DCCT and the UKPDS. Paradoxically, a rapid reduction in hyperglycaemia can provoke a transient deterioration in retinopathy, with worsening of the background condition or the development of preproliferative changes. This is probably due to an acute fall in retinal blood flow (which is elevated by hyperglycaemia), thus worsening ischaemia in already underperfused areas. Typically, the acute lesions resolve and the overall long-term outcome is improved if good glycaemic control can be maintained.
- Control of hypertension is likely to be important; the EUCLID study of enalapril showed beneficial effects which may reflect the inhibition of angiotensin-converting enzyme as well as lowering of blood pressure.
- Stopping smoking: smoking is thought to hasten the progression of retinopathy.
- Regular eye screening is essential, because even severe diabetic retinopathy may cause few or no symptoms.

Specific treatments

Laser photocoagulation can preserve useful vision in many cases of proliferative retinopathy and maculopathy. The blue-green light of the argon laser is maximally absorbed by vascular structures. It has a spot size of 50 to 500 µm and can be used to target discrete lesions such as clusters of leaking vessels identified by fluorescein angiography, but is usually employed to destroy larger areas of generally diseased retina. Panretinal photocoagulation ablates the peripheral retina with 1500 to 2000 burns that spare only a keyhole-shaped central area that includes the disc, the macula, and the maculopapillary nerve bundle running between them. Panretinal photocoagulation effectively concentrates the remaining retinal blood flow on to this crucial region which serves central, high-resolution colour vision, at the expense of the periphery. It is indicated for formation of new vessels (on the disc or elsewhere and rubeosis), and can be very effective: the chance of blindness within 5 years is reduced from 50 per cent to 25 per cent in patients at risk. It is increasingly used in the preproliferative phase to prevent the formation of new vessels. Photocoagulation is also used to treat macular oedema, in a 'grid' pattern around the central macula to destroy leaky capillaries; this reduces the 3-year risk of becoming blind from 30 per cent to 15 per cent.

Vitreoretinal surgery can now restore useful vision to some blind or severely impaired eyes. Techniques include vitrectomy (aspiration of vitreous haemorrhage and fibrovascular debris) and reattachment of detached or torn retina (using high-powered lasers to 'stitch' down the retina). Easy and rapid access to ophthalmologists and surgeons is often critical: for example, a detached retina must be repaired within a few weeks if it is to remain viable.

Practical aids for visual handicap range from pen injection devices and 'talking' blood glucose meters, to social support networks and national organizations for the visually impaired.

Cataract in diabetes

The normal lens transmits light because the fibre cells of the lens and the stacks of crystallin proteins which they contain are aligned in parallel. Normal ageing causes irreversible chemical modification ('browning') of the crystallins, with crosslinking and distortion that interrupt transmission of light and thus cause clouding of the lens. Diabetes accelerates the formation of these senile cataracts, probably through non-enzymatic glycation and crosslinking of AGE-modified crystallins.

Cataract is the most common cause of severe visual loss in diabetic patients over the age of 30, and is usually a typical 'senile' nuclear cataract with a characteristic radial spoke pattern. Much rarer is the 'snowflake' cataract with opacities scattered through the lens, which tends to occur in children presenting with severe hyperglycaemia. Here, the 'opacities' are reversible and are presumably due to local pockets of osmotic imbalance which distort the alignment of the lens crystallins.

Treatment of cataracts is conventional, usually removal and replacement by an intracapsular plastic lens. Long-term outcome is often not as good as in non-diabetic patients, because of coexisting maculopathy or proliferative retinopathy.

Gaucoma is more common in diabetics.

Other ocular problems

Cranial nerve palsies, especially of the third and sixth nerves, cause typical limitations of eye movement, often with acute onset of pain (see below and Chapter 23.13.15).

Eye infections include the rare but extremely destructive mucormycosis, which often spreads from the sinus to involve the orbit (see Chapter 7.12.1).

Diabetic neuropathies

Clinical syndromes

Diabetes damages nerves, both somatosensory and autonomic, in various ways that cause clinically distinct syndromes (Fig. 14). Subclinical nerve damage is common among diabetic patients, but significant neuropathic symptoms are fortunately unusual.

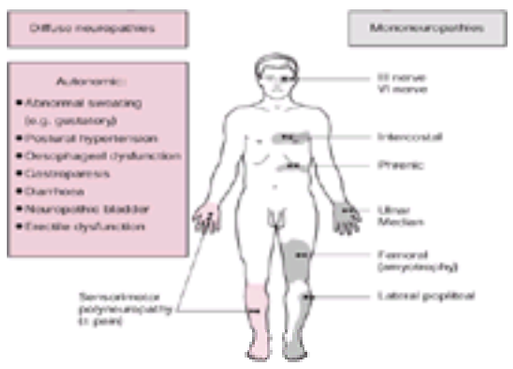

Fig. 14 Clinical manifestations of diabetic nerve damage.

Diffuse symmetrical polyneuropathy

This is classically a distal 'glove and stocking' peripheral polyneuropathy that affects all sizes of sensory and motor fibres; a variant selectively picks off small C fibres (see below). Both forms may be accompanied by neuropathy involving the sympathetic and parasympathetic divisions of the autonomic nervous system.

Aetiology

The diffuse nature of the nerve damage and its predilection for longer nerves is typical of toxic and metabolic neuropathies, and suggests cumulative damage that must reach a critical level before the function of a nerve is impaired. Both metabolic and vascular factors may contribute. The duration and severity of hyperglycaemia are generally related to the prevalence of neuropathy, while the DCCT confirmed that good glycaemic control decreased by the risks of developing neuropathy by about 70 per cent. Hyperglycaemia could damage nerves by glycation of key proteins. In the nerves of diabetic animals, polyol pathway overactivity has also been implicated (Schwann cells express aldose reductase), while myoinositol depletion may impair nerve conduction by inhibiting Na^{+},K^{+}-ATPase activity; in human diabetic neuropathy, however, the roles of polyols are less convincing and a role for aldose reductase inhibitors has not been demonstrated. There is some evidence of blockage of the vasa nervorum by microthrombi, which could follow the formation of AGE and the other mechanisms described earlier. Microvascular occlusion could contribute to intraneural hypoxia, demonstrated in the nerves of both animal and human diabetics.

Pathological changes affect both axons and Schwann cells. Axons fall out by dying back distally, sometimes accompanied by sprouting of regenerating nerve endings. Schwann cell damage leads to segmental demyelination. Nerve conduction velocity is slowed, in proportion to structural nerve damage.

Epidemiology and natural history

About 30 per cent of unselected diabetic patients have evidence of neuropathy on formal testing, but only 10 per cent suffer significant symptoms. Signs of neuropathy may be present in up to 10 per cent of newly diagnosed type 2 diabetic patients.

Nerve function generally worsens progressively over months or years, and areas of numbness may advance up the legs and occasionally involve the hands. Symptoms, including pain, may be variable; pain in particular may develop acutely, especially after weight loss or periods of poor diabetic control. Acute flareups tend to resolve after weeks or months, especially if glycaemic control is improved.

Symptoms and signs

Sensory symptoms are the commonest manifestation; muscle weakness occasionally predominates. Sensory symptoms may include loss of sensation, which can be profound, as well as 'positive' symptoms of pain, paraesthesiae, and allodynia (i.e. pain provoked by a normally innocuous stimulus, such as light touch or contact with bedclothes). Loss of the sense of touch and joint position may give the sensation of walking in thick socks, and Romberg's sign may be positive. Reduced pain sensation, which paradoxically may coexist with neurogenic pain, is potentially dangerous and an important cause of damage to neuropathic feet (see below). Horrifying accounts involving neuropathic diabetic feet include full-thickness burns to the soles after crossing a hot beach, pressure ulceration from a day's walking in tight new shoes, and transfixing the foot inside the shoe by stepping on a nail.

The mechanism of neuropathic pain is unknown; spontaneous firing of unstable regenerating nerves may be responsible. Pain is typically neurogenic, usually described as burning, shooting, or electric shock-like sensations, often with unpleasant pins and needles and allodynia. The feet and legs are usually affected, and the hands only rarely; neuropathic symptoms in the hands are usually due to damage to the ulnar and/or median nerves, which can be bilateral (Fig. 15 and Plate 1). Pain is characteristically worse at night, and can severely disturb sleep and cause depression (and suicide).

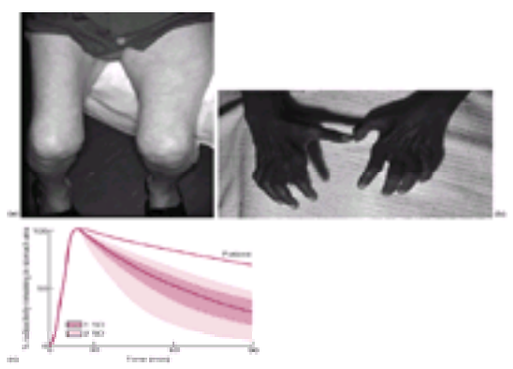

Fig. 15 (a) Diabetic amyotrophy: quadriceps right wasting due to femoral neuropathy (with thanks to Dr Geoff Gill, University Hospital Aintree, Liverpool). (b) Wasting of small muscles of the hands due to both ulnar and median nerve lesions. (See also Plate 1.) (c) Gastroparesis: grossly delayed gastric emptying. The normal range for clearance of the radiolabelled test meal from the stomach area is shown in in the darker shade (50 per cent confidence interval) and lighter shade (95 per cent confidence interval).

Examination may reveal symmetrical 'stocking' sensory loss affecting all modalities. Sensory deficits are frequently patchy and may be much less impressive than the symptoms suggest; clinical findings may even be normal in acute painful neuropathy. Tendon reflexes in the legs are often reduced or absent, and there may occasionally be marked muscle wasting. Neuropathic foot problems due to somatosensory and autonomic nerve damage may also be obvious, including ulceration, increased skin blood flow, and Charcot's arthropathy. The hands are less commonly involved.

Diffuse small-fibre neuropathy

This rare variant affects especially young women with type 1 diabetes, and may be autoimmune in origin. There is loss of temperature and pain sensation in a stocking distribution, other modalities remaining intact. Autonomic neuropathy is frequent, usually with postural hypertension, gustatory sweating, and diarrhoea. Ulceration and Charcot's arthropathy affecting the feet are common.

Autonomic neuropathy

Autonomic disturbances commonly accompany somatosensory neuropathy, and the autonomic nerves are presumably damaged by the same mechanisms. Up to 40 per cent of unselected diabetic people have abnormal tests of autonomic neuropathy, but only a few of these suffer major symptoms; however, these can be very debilitating and these patients have a significantly reduced life expectancy. The sympathetic and parasympathetic divisions are both affected.

Clinically apparent features are commonest in patients with long-standing diabetes and include:

- Abnormal sweating, mediated by cholinergic sympathetic nerves; this is one of the commonest autonomic symptoms. Profuse 'gustatory' sweating of the face and trunk (the area supplied by the superior cervical ganglion) may be provoked by eating, while sweating in the feet is often reduced.
- Postural hypotension, with a systolic fall exceeding 20 mmHg on standing, is due to failure of the normal sympathetically mediated increases in cardiac output and vasoconstrictor tone. This causes dizziness and blackouts, which may be mistaken for arrhythmias or myocardial ischaemia. Symptoms and the degree of postural drop may vary considerably with time. Postural hypotension may be exacerbated by the vasodilator effects of antihypertensives, nitrates, tricyclic antidepressants (used to treat neuropathic pain), and insulin.
- Disturbed gastrointestinal motility. Dysphagia may be due to oesophageal dysmotility. Gastric stasis, due to failure of the pylorus to relax when the antrum contracts, causes particular difficulties with emptying liquids and presents with recurrent vomiting. There may be obvious fullness in the epigastrium, sometimes with a succussion splash. Disturbances of motility in the colon most commonly lead to diarrhoea (characteristically but not always worse at night), which may be exacerbated by bacterial overgrowth of the relatively immotile small bowel. As with all these gastrointestinal symptoms, diarrhoea is often episodic and may alternate with constipation. Anorectal dysfunction is luckily rare, but can cause severe faecal incontinence.
- Neuropathic bladder, due to damage to the sacral nerves, prevents normal emptying and can lead to a permanently distended, sometimes palpable, bladder, with overflow incontinence. Hydroureter and hydronephrosis are other complications, and ascending urinary tract infections are common.
- Sexual difficulties include failure of erection (a parasympathetic response mediated by the sacral nerves) and sometimes failure of ejaculation (a sympathetic reflex transmitted by the lumbosacral outflow). Erectile failure is relatively common in diabetic men, affecting about 50 per cent of those over 55 years; as in the non-diabetic population, depression and anxiety (including fears about poor sexual performance) are common contributory factors. Arterial inflow to the corpora cavernosa may be compromised by atheromatous disease of the pudendal arteries or common iliac arteries, the latter causing the Leriche syndrome of impotence with claudication of the buttocks.
- Abnormal blood flow. Sympathetic denervation allows vasodilatation and relaxation of precapillary sphincters in the skin, hence the warm skin and distended veins characteristic of the neuropathic foot. Increased blood flow in bone may be an early abnormality in Charcot's arthropathy.
- Sudden unexplained death is more common in patients with severe autonomic symptoms. Possible causes include cardiorespiratory arrest and arrhythmias triggered by hypoglycaemia, awareness of which is blunted in many patients with long-standing diabetes.

Acute mononeuropathies

These syndromes are due to acute damage to isolated peripheral nerves, presumably due to a vascular event rather than metabolic damage. Limited histological studies show focal demyelination, probably consistent with this. Occasionally, two or more nerves can be affected more or less simultaneously (mononeuritis multiplex).

Diabetic amyotrophy

This is due to damage of one of the major nerve trunks or roots (radiculopathy) supplying the leg. The femoral nerve is most commonly involved, causing symptoms in the quadriceps muscle; other muscle groups are less often affected. Femoral neuropathy causes neurogenic pain of acute onset (burning or lancinating, usually severe), with weakness and often surprisingly rapid wasting in the quadriceps, and loss of the knee tendon reflex (Fig. 15). For unknown reasons, some patients have extensor plantar reflexes, in which case a spinal or cauda equina lesion must be excluded.

Amyotrophy most commonly presents in patients over 50 years of age, often following a period of poor diabetic control. Pain usually resolves spontaneously over several months, especially if diabetic control is improved, but muscle strength and tendon reflexes may take much longer to return.

Cranial and other nerve palsies

These are common, affecting the third and sixth nerves in particular. Third nerve palsy is often accompanied by pain behind the eye, and may need to be differentiated from an aneurysm of the posterior communicating artery; however, unlike in classical third nerve palsy, ptosis and pupillary dilatation are usually absent. Acute neuropathic damage may also affect the phrenic nerve (causing an elevated hemidiaphragm), and intercostal or truncal nerves, causing shingles-like pain and sometimes localized bulging of the abdominal wall. All these acute palsies tend to resolve spontaneously.

Pressure palsies

These include the median, ulnar, and occasionally lateral popliteal nerves, and are thought to be due to pressure damage superimposed on hypoxic or otherwise compromised nerves. These present in the classical way, but often recover slowly and incompletely, and respond poorly to surgical decompression (see Fig. 15).

'Insulin neuritis'

This is a transient deterioration in nerve function, often with pain and dysthaesiae affecting the legs symmetrically, which follows an acute improvement in glycaemic control, usually after starting insulin therapy. It may be due to an acute fall in nerve perfusion analogous to the decrease in retinal blood flow thought to explain a temporary deterioration in retinopathy under these circumstances (see above). Symptoms usually resolve within weeks or months.

Diagnosis of diabetic neuropathies

Peripheral sensorimotor neuropathies

A carefully taken history is usually diagnostic. The key qualities of neuropathic pain should distinguish it from claudication and night cramps.

Sensory deficits should be mapped on the legs and hands, for both large-fibre (vibration with a 128-Hz tuning fork, joint position sense, light touch, temperature, e.g. with a cold tuning fork) and small-fibre modalities (pin prick, light touch); objective losses may not match the patient's symptoms. A useful test uses the Semmes–Weinstein nylon monofilament, which is pressed against the skin until it buckles; the patient's inability to feel the 10-g filament indicates neuropathy severe enough to predict foot ulceration. Various bedside instruments can be used to assess specific sensory modalities, such as the biothesiometer (for vibration sense) and thermal threshold testers; age-related normal ranges are available for these methods but they are quite variable and add little to routine management. Muscle wasting and weakness should be sought, and the tendon reflexes checked.

Peripheral diabetic neuropathy must be differentiated from other metabolic neuropathies, including vitamin B_{12} deficiency, uraemia, and alcohol, all of which may affect diabetic patients.

Autonomic neuropathy

The most convenient tests are of cardiovascular autonomic function, which detect loss of the normal reflexes that modulate heart rate during respiration (mainly vagal) and that increase heart rate and blood pressure on standing (sympathetic). The simplest test is to measure heart rate (RR interval) from an ECG tracing during controlled deep breathing (5 s inspiration, then 5 s expiration, repeated for 1 min); the physiological bradycardia on expiration is lost, with a difference of less than 10 beats/min between inspiration and expiration. Reflex bradycardia during the Valsalva manoeuvre is similarly abolished. More sophisticated measures employing spectral analysis of variability of heart rate during normal breathing are more sensitive.

Postural hypotension is defined as a drop of more than 20 mmHg drop in systolic blood pressure, measured 30 s after standing. Postural drops often vary considerably through the day and from week to week.

Abnormalities of cardiovascular autonomic tests are common in patients with long-standing diabetes, especially type 1, and do not necessarily indicate that symptoms such as vomiting, diarrhoea, or erectile dysfunction are due to autonomic neuropathy. Other specific tests include:

- Gastroparesis: a plain abdominal radiograph may show a 'ground glass' appearance in the epigastrium, while endoscopy (always indicated to exclude pyloric obstruction) shows a dilated, poorly contracting stomach with a closed pylorus. Gastric emptying studies using radiolabelled test meals show delayed disappearance of radioactivity, particularly of a liquid test meal; however, abnormalities may not be consistent with the severity of symptoms (see Fig. 15).
- Neuropathic bladder and associated hydroureter and hydronephrosis can be confirmed by ultrasound or intravenous urography.
- Erectile failure is often multifactorial (see above). If it is due to autonomic neuropathy, other signs of autonomic dysfunction, especially neuropathic bladder, are usually prominent.

Treatment of diabetic neuropathies

General measures

Poor glycaemic control should be corrected. As well as helping to prevent the development of neuropathy, this may curtail pain in the acute syndromes; insulin neuritis is rare and usually self-limiting. Once established, chronic sensory motor neuropathy tends to progress, irrespective of glycaemic control. Specific treatments which aim to prevent or reverse diabetic nerve damage—aldose reductase inhibitors and aminoguanidine (which prevents the formation of AGE)—have so far been disappointing. Numb feet are at greatly increased risk of ulceration and require sensible shoes and good foot care (see below).

Pain may be difficult to treat; pain management programmes in specialized pain relief clinics may be helpful. The following should be tried in sequence:

- Simple analgesics (aspirin, paracetamol) are mostly unhelpful. Opiates are generally regarded as ineffective in neurogenic pain and are yet to be tested adequately in diabetic neuropathy.
- Tricyclic drugs suppress neurogenic pain, in addition to their antidepressant effects. Amitryptiline or imipramine can be started at 25 mg at bedtime (10 mg in the elderly), increasing weekly to a maximum of 75 to 150 mg. Side-effects, including postural hypotension, may limit the dosage. A phenothiazine such as fluphenazine (2.5–5 mg) is said to enhance the analgesic effect of tricyclics but its use is not evidence-based and it often exacerbates postural hypotension.
- Anticonvulsants, which stabilize the neurone membrane and may prevent spontaneous firing of C fibres, may be substituted for trycyclics or used in combination with them. Carbamazepine (initially 100 mg once or twice daily, up to 800 mg/day in divided doses) is often effective. Sodium valproate or phenytoin are alternatives. Gabapentin appears promising, but has yet to be compared adequately with tricyclics. Gabapentin has been given at doses of 900 to 3600 mg/day in divided doses—rather more than the standard antiepileptic regimen but apparently well tolerated.

Pain in the feet may respond to topical application of capsaicin ointment; capsaicin causes the burning sensaton of hot chillies, and depletes the pain-transmitting C fibres of the neurotransmitter substance P. Pain may be transiently worsened after application but relief can last for many hours. Contact hypersensitivity (allodynia) can be helped simply with a bed cradle to prevent contact with bed clothes, or by applying Opsite® adhesive plastic film to the skin.

Other drugs that have proved successful in some but not all trials include oral mexiletine (a class 1b antiarrhythmic agent that stabilizes nerve cell membranes; up to 450 mg/day in divided doses) and clonazepam (0.5–3 mg) in patients whose sleep is disturbed by 'restless legs'. Some patients unresponsive to drug therapy may benefit from implantation of a dorsal column stimulator, designed to exploit the 'gate' control of pain transmission.

Autonomic neuropathic symptoms may be treated as follows:

- Excessive sweating may be controlled by oral clonidine or topical 1 per cent glycopyrrholate ointment; systemic anticholinergics such as poldine have also been effective but have many side-effects, including urinary retention.
- Postural hypotension may be helped simply by raising the head of the bed at night. Fludrocortisone can be useful, but aggravates coexistent supine hypertension; very high doses (up to 1 mg/day) may be needed. The a_1-adrenergic agonist midodrine (2.5–10 mg daily) may also be helpful, but can also worsen hypertension.
- Vomiting due to gastroparesis often responds to metoclopramide or domperidone, and the prokinetic drug erythromycin (cisapride has been withdrawn because of arrhythmias). Some patients unresponsive to drug therapy may require intrajejunal feeding, most conveniently by percutaneous endoscopic gastrostomy; some patients benefit from surgical drainage procedures such as a roux-en-Y gastrojejunostomy.
- Diarrhoea is often improved or cured by erythromycin or tetracycline when bacterial overgrowth is a contributory factor.
- Neuropathic bladder may respond to regular bladder training, but intermittent self-catheterization may be needed.
- Erectile dysfunction can be treated with oral sildenafil 4 mg, which should be taken about 1 h before intercourse. It is effective in about 50 per cent of diabetic patients but is absolutely contraindicated in those taking nitrates in any form, because of the risk of profound hypotension and circulatory collapse. Alternatives, if sildenafil is contraindicated or ineffective, include the injection of vasodilators such as papaverine or prostaglandin E_1 into the corpus cavernosum, intraurethral alprostadil, or the use of vacuum tumescence devices. Coexistent contributory factors such as depression, alcohol, or drugs (including b-blockers and thiazides) should be sought and treated. Counselling of the couple is obviously important.

Diabetic nephropathy

This is covered in detail in Chapter 20.10.1.

Macrovascular disease

Diabetes of all types increases the background risk of atheroma, amplifying the hazards of additional cardiovascular risk factors such as hypercholesterolaemia, hypertension, and smoking. Atheroma appears earlier, spreads faster and more extensively, and carries greater morbidity and mortality than in non-diabetic people. Overall risks for myocardial infarction, stroke, and limb ischaemia are two to four times higher than in the general population, and diabetic women lose the premenopausal protection which their euglycaemic counterparts normally enjoy. This increased level of risk is comparable with that in non-diabetic subjects who have already suffered a myocardial infarct. Accordingly, it has been argued that primary cardiovascular prevention for diabetic patients should be as active as secondary prevention in the non-diabetic population. Macrovascular disease, especially coronary heart disease, is the main cause of premature death in type 2 diabetes.

As discussed earlier, cardiovascular risk is increased above glucose levels that lie below the 'diabetic' range: IGT also predisposes to coronary-heart disease, but not to retinopathy. This relationship probably explains why 'tight' glucose control has not yet been shown to prevent macrovascular disease, because none of the trials (for example the DCCT or the UKPDS) has managed to achieve even near normoglycaemia. There is a very strong relationship between microalbuminuria and premature death from cardiovascular disease (see Chapter 20.10.1). This presumably reflects widespread damage to the endothelium, which both predisposes to atheroma formation and enhances albumin leakage in the glomerulus.

Cardiovascular risk factors tend to cluster together with type 2 diabetes in the so-called 'metabolic syndrome X' (see above). These risk factors are also common in the type 1 diabetic and the non-diabetic populations. Because their impact is worsened by diabetes, they require active management. Treatment of obesity and smoking is discussed in Chapter 10.5 and Chapter 3.5.

Dyslipidaemia in diabetes

Type 1 and type 2 diabetes are both commonly accompanied by lipid disorders that are strongly atherogenic but may appear deceptively trivial on routine screening (Fig. 16).

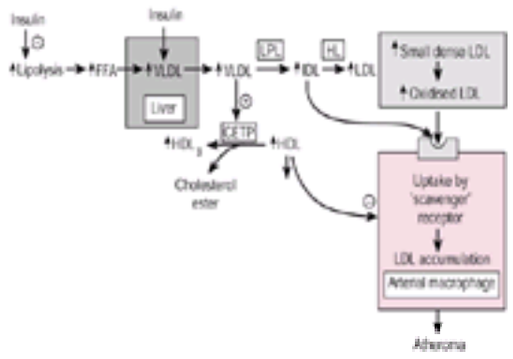

Fig. 16 Mechanisms of lipid abnormalities in diabetes. Insulin resistance or loss of insulin leads to increased lipolysis and hepatic secretion of very low-density lipoprotein (VLDL). High levels of VLDL stimulate cholesterol ester transfer protein which converts the antiatherogenic high-density lipoprotein (HDL) into HDL_3, which lacks this protective effect. VDLPs are stripped of triglyceride by lipoprotein lipase to yield atherogenic intermediate-density lipoprotein (IDL), and then by hepatic lipase to produce low-density lipoprotein (LDL), including the highly atherogenic small dense oxidized fractions. These, with IDL, are taken up by arterial wall macrophages, forming the 'foam cells' that initiate the atheromatous plaque.

In poorly controlled type 1 diabetes, the most obvious abnormality is hypertriglyceridaemia. This is due mainly to increased production of very low-density lipoprotein by the liver, driven by high levels of free fatty acids resulting from enhanced lipolysis in fat; triglyceride levels can be lowered by improved insulin therapy. Total cholesterol concentrations are often 'normal', while HDL may be increased, although this is predominantly the HDL_3 subclass which does not confer protection against atheroma (see Chapter 11.6). As stressed below, a 'normal' cholesterol is not necessarily reassuring because modified LDL particles in diabetes are particularly atherogenic and because cardiovascular risk in diabetic people is increased at all levels of cholesterol.

Dyslipidaemia in type 2 diabetes is also subtle, and similar to that of obesity and syndrome X. High-density lipoprotein cholesterol is reduced, increasing the ratio of LDL to HDL, while triglycerides are often modestly raised. The abnormalities are also due to excessive production of VLDL by the liver, due to insulin resistance rather than insulin deficiency: free fatty acids levels are raised by lipolysis (adipocytes are resistant to the normal antilipolytic action of insulin), while the inhibition by insulin of hepatic production of VLDL is lost. High levels of VLDL favour the production of two highly atherogenic cholesterol particles, namely IDL and small dense LDL, which is easily oxidized; both IDL and oxidized LDL are taken up, via specific receptors, by macrophages in the artery wall and are then transformed into foam cells (see Chapter 11.6). High triglyceride levels also accelerate the removal of cholesterol ester (via cholesterol ester transfer protein) from HDL, ultimately reducing levels of HDL and producing the non-protective HDL_3 fraction.

Risks of dyslipidaemia

Cardiovascular events including fatal myocardial infarction are three to four times commoner among diabetic patients than in the general population, across a wide range of cholesterol levels including 'normal' values. The risks of the slightly raised triglyceride concentrations are less obvious, and are currently the subject of large-scale trials with fibrates.

Management

This remains controversial, although the consensus is that diabetic people benefit at least as much from statin therapy as the non-diabetic population. The high background risk of diabetes is reflected in the lower treatment thresholds for hypercholesterolaemia in the various risk-factor tables (see Chapter 11.6), and lipid-lowering drugs are now used at ever lower cholesterol levels. Indeed, the American Diabetes Association recommends statin therapy for any diabetic patient with a total cholesterol of more than 5 mmol/l, aiming to maintain cholesterol below this level. It seems reasonable at present to consider adding a fibrate if triglycerides then remain in excess of 2.3 mmol/l.

Pharmacoeconomic arguments may intrude, and it is debatable whether dyslipidaemia should be treated exhaustively in patients who make no effort to stop smoking. Other factors that aggravate dyslipidaemia—obesity, poor glycaemic control, and drugs such as thiazides and b-blockers—should be tackled if possible.

Hypertension

Hypertension commonly accompanies diabetes: about 40 per cent of type 2 patients are hypertensive at diagnosis, and probably two-thirds of the diabetic population are inadequately treated with respect to current management guidelines (Table 3).

Causes

Essential hypertension is an integral feature of the metabolic syndrome X, associated with obesity and insulin resistance. Possible mechanisms include enhanced central sympathetic tone (possibly mediated in part by raised insulin levels), and increased total body sodium and extracellular fluid volume, to which the Na^+-retaining effects of hyperinsulinaemia could contribute; loss of insulin-induced vasodilatation due to insulin resistance could also play a role. Secondary hypertension may also develop because of specific diabetic complications, notably nephropathy (loss of the normal nocturnal dip in blood pressure can be an early feature: see Chapter 20.10.1), stenosis of the renal artery due to atheroma, and stiffening of the larger conduit arteries causing isolated systolic hypertension. Supine hypertension can coexist with postural hypotension—a particularly difficult combination to treat effectively.

Treatment targets

The blood pressure thresholds for active management have fallen progressively, with wider appreciation of the damage inflicted by even modest hypertension in diabetic patients and of the benefits of control of blood pressure. The American Diabetes Association now recommends 130/80 mmHg (recorded in the clinic) as the treatment target for blood pressure, and that consistently higher values require active treatment; levels exceeding 140/90 mmHg are increasingly regarded as unacceptably poor. Ambulatory blood pressure readings are lower than those recorded in the clinic, and mean daytime levels of less than 130/75 mmHg are the current target. Target blood pressure (clinic readings) for patients with microalbuminuria is 125/75 mmHg.

Impact of hypertension in diabetes

Like all cardiovascular risk factors, the atherogenic hazards of hypertension are amplified in the diabetic population. Hypertension also plays a crucial role in accelerating the progression of diabetic nephropathy and also of retinopathy. Several studies have shown that treating hypertension reduces the risks of myocardial infarction and stroke by 30 to 70 per cent; in the UKPDS, 'tight' blood pressure control (averaging 144/82 mmHg compared with 154/87 mmHg in less tightly controlled subjects) reduced stroke by 40 per cent and the need for photocoagulation by 30 per cent, although strangely, no effect on myocardial infarction was observed (Fig. 12).

Management

This should begin with lifestyle modification, including restricting energy intake and increasing exercise in the obese, and reducing alcohol and sodium intakes. If actually carried out, these general measures can lower blood pressure at least as effectively as many antihypertensive drugs. Most patients, however, will require drugs and many need combination therapy. Antihypertensive drugs can be used in the conventional stepped approach (Chapter 15.16.1.3) to achieve the target blood pressure, ideally 130/80 mmHg; less stringent targets may be appropriate for elderly patients or those who cannot tolerate tight blood pressure control because of other problems.

The choice of hypotensive drugs is less important than the level of blood pressure achieved, although some agents have properties that are better suited to diabetes.

All these drugs can worsen or precipitate postural hypotension in autonomic neuropathy.

Diuretics

High-dose thiazides (e.g. 5 mg bendrofluazide) can worsen hyperglycaemia in type 2 diabetes, apparently by impairing insulin secretion (a consequence of K^+ depletion) and possibly increasing insulin resistance. Low dosages (for example, 2.5 mg bendrofluazide) do not appear to aggravate glucose intolerance. Diuretics can precipitate hyperosmolar non-ketotic coma, while thiazides may also worsen dyslipidaemia.

b-Blockers

b-Blockers may also raise blood glucose in type 2 patients by increasing insulin resistance (possibly related to weight gain) and by interfering with insulin release (which is stimulated by b_2 adrenoceptors). b-blockers can also aggravate dyslipidaemia and impotence, while non-cardioselective agents can mask the sympathetically driven symptoms of hypoglycaemia. Low dosages of cardioselective b-blockers (for example, atenolol or metopropol) are safe, and are also indicated for treating angina and in the secondary prevention of myocardial infarction.

Angiotensin converting enzyme (ACE) inhibitors

These can effectively control blood pressure when combined with low-dose diuretics and are useful in the many diabetic patients who also have heart failure or left ventricular dysfunction, especially following a myocardial infarct. They reduce proteinuria, by relaxing the efferent arterioles in the glomerulus, and slow the development of both nephropathy and retinopathy; some evidence points to specific beneficial effects in nephropathy, in addition to the lowering of blood pressure (see Chapter 20.10.1). ACE inhibitors do not worsen blood glucose or lipids, and may even improve insulin sensitivity. They are contraindicated in renal artery stenosis, which is relatively common in arteriopathic diabetic patients, and can cause dangerous hyperkalaemia in those with hyporeninaemic hypoaldosteronism (type 4 renal tubular acidosis) which can be associated with diabetic nephropathy.

Calcium channel antagonists

These have no adverse metabolic effects and are useful in patients with angina or tachyarrhythmia.

Other drugs

Other drugs include a_1-adrenoreceptor antagonists (for example doxazosin), which may slightly improve insulin sensitivity; angiotensin II receptor antagonists (for example losartan, candesartan), which are useful alternatives when angiotensin converting enzyme inhibitors are poorly tolerated because of cough; and moxonidine, a centrally acting sympathetic drug acting on imidazoline I_1 receptors.

Coronary heart disease

Compared with their non-diabetic counterparts, clinically significant coronary heart disease is over twice as common in diabetic men and postmenopausal women and four times commoner in premenopausal women.

Angina

It is said that myocardial ischaemia can be painless in patients with long-standing diabetes and autonomic neuropathy, presumably because of sensory denervation of the heart; however, the overall prevalence of 'silent' myocardial ischaemia appears to be similar to that in the general population.

Angina should be treated first with conventional drugs, remembering the diabetogenic and other hazards of b-blockers and the potential of nitrates and calcium-channel antagonists to aggravate postural hypertension. Recent studies have confirmed that coronary artery stenting and bypass grafting markedly reduce fatal myocardial infarction in diabetic patients; it follows that exercise testing and coronary angiography should be performed sooner rather than later in those with worsening angina.

Myocardial infarction

The risk of fatal myocardial infarction is as high in diabetic people as in non-diabetics who have already suffered an infarct. Mortality from myocardial infarction is over twice as high as in matched non-diabetic controls, whether or not thrombolytic agents are used. About 75 per cent of diabetic patients are dead within 5 years of their first infarct, the main causes of death being heart failure and acute cardiogenic shock.

Primary prevention

Independent risk factors—hypertension, dyslipidaemia, smoking, and obesity—must be treated energetically in all diabetic patients, and poor glycaemic control should be tightened even though this alone is unlikely to protect against coronary heart disease. Specific prophylactic therapy, comparable to secondary prevention measures (i.e. measures given after an initial infarct) in the non-diabetic population, is increasingly recommended for diabetic people with any evidence of ischaemic heart disease, because their risk of infarction is so high. This includes a cardioselective b-blocker (for example, atenolol or metoprolol) and/or an angiotensin converting enzyme inhibitor if echocardiography shows left ventricular dysfunction with an ejection fraction of less than 40 per cent. b-Blockers improve survival at least as much as in the general population; the case for angiotensin converting enzyme inhibitors is not yet proven in diabetes. Some authorities recommend aspirin in all diabetic patients over 40 years of age, although the dosage (75–300 mg/day) remains undecided.

Acute myocardial infarction should be managed as follows (Table 10):

- Thrombolytic drugs should be given; survival is improved at least as much as in the non-diabetic population. Proliferative retinopathy is not a contraindication, and the risk of intraocular haemorrhage is very low.
- Glycaemic control should be optimized during the acute episode. There is evidence that long-term survival is improved if intensive insulin treatment is started on admission of a diabetic patient and continued for some months afterwards. The Diabetes Mellitus Insulin Glucose Infusion in Acute Myocardial Infarction (DIGAMI) Study showed that late (> 1 year) cardiovascular deaths were significantly reduced by one-quarter (from 44 per cent to 33 per cent after 3 years) in hyperglycaemic patients who received an insulin–glucose infusion in the coronary care unit, followed by multiple daily subcutaneous insulin injections for 3 months. Follow-up studies in progress should identify which elements conferred protection; theoretical mechanisms include lowering of free fatty acid levels, which may worsen ischaemic myocardial damage. In the meantime, it is reasonable to control hyperglycaemia (aiming for 5–10 mmol/l) during the first 48 h in all known diabetic patients, and in those diagnosed diabetic at admission (blood glucose over 11 mmol/l, with HBA_{1c} in the diabetic range). This can be done most easily with a simple sliding scale continuous intravenous insulin infusion as used to treat diabetic ketoacidosis (see above). Delivering insulin as a 1 U/ml solution with a syringe driver avoids unnecessary intravenous fluids—this is an important consideration, as heart failure is a common (and often fatal) complication of myocardial infarction in diabetic people. Alternatively, the (cumbersome) DIGAMI protocol can be used (see Malmberg K *et al*. in Further reading).
- Secondary prevention must include aspirin with a b-blocker and/or angiotensin converting enzyme inhibitors, as discussed above. Total cholesterol should be reduced to less than 4.8 mmol/l with a statin, with the option of adding a fibrate to control residual hypertriglyceridaemia. Early echocardiography and exercise testing or equivalent stress testing are needed, followed when appropriate by coronary angiography and ultimately referral for coronary angioplasty and stenting or bypass grafting.

Stress-induced hyperglycaemia and myocardial infarction

The intense sympathetic discharge triggered by an infarct can push blood glucose acutely into the diabetic range in previously normoglycaemic individuals. Stress-induced hyperglycaemia can be distinguished from previously undiagnosed diabetes because the HBA_{1c} will be normal. The acute management of these

individuals is uncertain; they were classified as 'diabetic' in the DIGAMI Study and also enjoyed improved survival with intensive insulin therapy. Pending further information, a pragmatic strategy would be to control glycaemia tightly during the admission with intravenous insulin for 48 h as described above, and then to monitor blood glucose closely and treat any persistent hyperglycaemia with insulin.

Heart failure

This is common in diabetic people with ischaemic heart disease and is a major cause of death in those who suffer an infarct. In addition to ischaemia from coronary artery disease, a specific diabetic cardiomyopathy may contribute to failure, as left ventricular function may be impaired in the absence of obvious atheroma on coronary angiography. Various defects in contractility and calcium flux within cardiomyocytes have been identified in diabetic animal models, but their relevance to humans is not clear. Echocardiography may show specific abnormalities early in diastole as well as areas of dyskinesia and a reduced ejection fraction. Moderate or severe left ventricular dysfunction, with an ejection fraction of less than 40 per cent, should be treated with an ACE inhibitor.

Stroke

Cerebrovascular accidents, mostly embolic, are two to five times commoner in diabetics than in the general population. They are managed conventionally. As with myocardial infarction, a stroke can acutely raise blood glucose in both diabetic and non-diabetic people. Trials are currently under way to determine whether tight glycaemic control can also improve the outcome of a stroke in hyperglycaemic individuals; for the moment it seems reasonable to keep blood glucose within the range 7 to 10 mmol/l, with the early use of insulin (probably subcutaneously) if this proves difficult.

Peripheral vascular disease

This is very common in the legs, often with diffuse disease distally as well as in the iliac and femoral arteries. Consequences include intermittent claudication, pain at rest, and gangrene which is usually dry and may lead to the loss of one or more toes. Severe atheroma in the iliac arteries can cause the Leriche syndrome, with buttock claudication and erectile failure; the latter may also be due to involvement of the pudenal arteries.

Investigation and management are conventional. Intermittent claudication may improve with the simple advice to stop smoking and keep walking, but a shortening claudication distance or pain at rest require urgent investigation. Angiography often shows widespread atheroma, and this may preclude reconstructive surgery. Otherwise, standard operations such as femoral–popliteal bypass and the use of the saphenous vein *in situ* for more distal disease can achieve good results and must not be withheld from people simply because they have diabetes. Nonetheless, diabetes still accounts for about one-half of all non-traumatic leg amputations.

Patients with diabetic renal failure or autonomic neuropathy may show calcification of the artery walls, often easily visible in the digital arteries on plain radiography of the feet (Fig. 17 and Plate 2). This medial sclerosis (Mönckeberg's) is not directly related to atheroma, although the two often coexist.

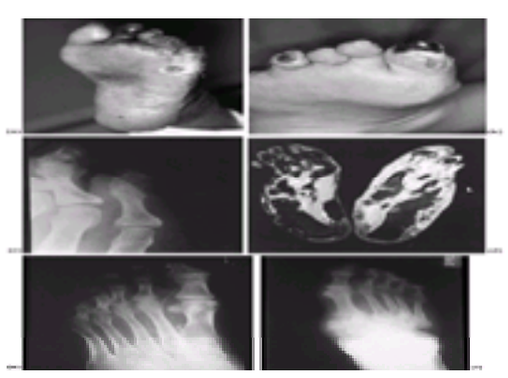

Fig. 17 The diabetic foot. (a) Typical punched-out neuropathic ulcer on the lateral aspect of the sole in an ischaemic foot with gangrene of the second, fourth, and fifth toes. (b) Ulceration and digital gangrene, caused by wearing tight shoes on a severely ischaemic foot. (c) Osteomyelitis in the diabetic foot. Early changes can be subtle: in this case, an erosion at the lateral edge of the distal end of the proximal phalanx of the fifth toe. ^{111}In-labelled white cell scanning showed an intense 'hot spot' at this site (with thanks to Dr Hans Laasch, Manchester Royal Infirmary). (d) Osteomyelitis affecting the proximal phalanx of the left big toe and the adjacent metatarsophalangeal joint, visible as an abnormally high signal on MR imaging. Associated oedema shows as a high signal in the surrounding soft tissues (with thanks to Dr King Sun Leong, Whiston Hospital). (e) Mönckeberg's medial sclerosis, outlining the digital arteries on this plain radiograph. (f) Charcot's arthropathy, showing massive destruction of the distal ankle joint. (See also Plate 2.)

Other manifestations of peripheral vascular disease include angina-like abdominal pain after eating, which is due to narrowing of the mesenteric artery, and renal artery stenosis which can contribute to hypertension and precipitate acute renal impairment soon after starting an angiotensin converting enzyme inhibitor.

Diabetic foot disease

The feet are at the mercy of various diabetic complications and problems such as ulceration and resistant deep infections often cause long and expensive hospital admissions. Ulceration and severe ischaemia leading to gangrene of the toes or forefoot are the commonest problems.

Diabetic foot disorders are best managed in a dedicated combined clinic. Prevention is extremely important. Many problems can be avoided by teaching the patients basic foot care, by regularly checking their feet and shoes, and by providing prophylactic chiropody and special footwear as appropriate. Charcot's arthropathy most commonly affects joints in the ankle or foot.

Neuropathy damages the foot through motor, sensory and autonomic involvement. Distal motor neuropathy alters the posture of the foot by weakening its small intrinsic muscles and allowing the unopposed action of the long extensors to claw the foot, concentrating pressure on the heel and the metatarsal heads. Shear forces generated by walking and shoes cause the skin over pressure points to thicken into callus; eventually, pressure damage leads to foci of liquefactive necrosis deep within the callus, and these break through to the surface to form an ulcer (Fig. 17).

Autonomic denervation opens up arteriovenous anastomoses, shunting blood to the skin—hence the warm skin and dilated veins of the neuropathic foot. Shunting may also deprive the tissue bed of oxygen and nutrients, thus worsening ischaemia from arterial disease. Increased blood flow may also be an initiating factor in Charcot's arthropathy (see below). Sensory denervation and loss of pain sensation can allow the foot to be damaged by agents such as over-tight new shoes, a drawing pin, or sharp stones in the shoe.

Ischaemia, due to peripheral vascular disease and possibly microvascular damage, can lead to ischaemic ulceration and to gangrene of the toes or forefoot (Fig. 17).

Foot ulceration

This is usually multifactorial, although one cause (for example, neuropathy) may initiate or dominate the process.

Trauma includes the normal wear and tear of walking in shoes and damage from foreign bodies, often undetected because of neuropathy. Inexpert do-it-yourself chiropody is another cause.

Infection commonly complicates diabetic foot ulcers, and often penetrates deep into the soft tissues and bone. Mixed organisms are usually responsible, including staphylococci, pseudomonas, and anaerobic organisms, sometimes gas-forming. Osteomyelitis is particularly ominous and requires urgent diagnosis and treatment

(Fig. 17).

Clinical features of diabetic foot ulcers

Primary neuropathy and ischaemic ulcers can generally be distinguished as below, but careful examination of the whole foot is essential so that all the possible contributory causes can be adequately treated.

- Primarily neuropathic ulcers occur at high-pressure sites (heel, metatarsal heads) and appear cleanly punched out of the surrounding callus (Fig. 17). The foot may be numb, with or without neuropathic pain, and the ulcer is often painless and may not have been noticed by the patient. Typical neuropathic features including clawed posture of the foot, warm skin, and sensory loss.
- Ischaemic ulcers tend to affect the edges of the foot and are often painful. There may be a history of intermittent claudication, absent foot pulses, and cold skin, sometimes with obviously ischaemic toes or previous amputation (Fig. 17).
- Traumatic ulceration may hint at its causes, for example symmetrical damage across the toes and margins of the feet from tight shoes (Fig. 17).
- Infection may cause local signs of inflammation, although the skin may appear deceptively normal even over extensive and serious deep infection, and pain may be absent. Anaerobes and pseudomonas characteristically produce a foul smell, while gas formation may occasionally cause crepitus in the subcutaneous tissues.

Investigation of diabetic foot ulcers

Effective treatment depends on identifying the cause(s) of ulceration. Neuropathy and ischaemia are assessed and managed as above. Swabs or curettings from deep in the ulcer should be cultured for both aerobic and anaerobic organisms. Plain radiography of the foot may show gas in the soft tissues or osteomyelitis, which can be difficult to distinguish from Charcot's arthropathy; MRI scans and ^{111}In-labelled white cell scans are highly specific for infection (Fig. 17).

Management of diabetic foot ulcers

Predominantly neuropathic ulcers are treated by the chiropodist to remove callus, and the foot protected with a lightweight plaster cast to unload pressure from the affected area, which accelerates healing while keeping the patient mobile; this may need to be worn for several weeks. Extra-depth or custom-built shoes, or pressure-absorbing socks, will reduce pressure loading and help to prevent recurrence. Ischaemia is treated as above, aiming to avoid or limit amputation.

Infection must be treated with appropriate antibiotics and repeated cultures may be needed to ensure that mixed infections, especially including fastidious organisms and anaerobes, are completely covered. Soft-tissue infections may respond to oral broad spectrum antibiotics such as amoxicillin and flucloxacillin, or coamoxiclav, combined with metronidazole if anaerobic infection is suspected; deep infections and osteomyelitis may require some weeks of intravenous treatment with drugs such as clindamycin, and surgical debridement. Amputation may be needed in refractory cases.

Charcot's arthropathy

This is fortunately rare, affecting fewer than 0.5 per cent of diabetic patients. It usually occurs in those with dense peripheral neuropathy and profound sensory loss, often with symptomatic autonomic damage. Reduced pain sensation is assumed to favour traumatic damage, and acute flareups are often preceded by injuries, which may be apparently trivial. Interestingly, blood flow to the affected area is increased early in the Charcot process, possibly because sympathetic denervation allows dilation of arterioles supplying the bone; this may stimulate osteoclast activity and bone resorption.

The ankle and joints in the mid- and forefoot are most commonly affected; non-weight-bearing joints are very rarely involved. The natural history is variable but can lead to massive destruction of the articular surfaces and resorption of adjacent bone, often with a large effusion that can become acutely inflamed and mimic septic or inflammatory arthritis. In advanced cases, the joint degenerates into a 'bag of bones'. The process is generally painless, but acute flareups can cause discomfort. The most important differential diagnosis is from septic arthritis and osteomyelitis. Radiographic appearances are characteristic in advanced cases (Fig. 17), but may be ambiguous early on; ^{99}Tc^{m} bone scans show increased uptake while the white cell scan is usually negative. Magnetic resonance imaging is probably best able to distinguish Charcot's arthropathy from osteomyelitis. An acutely inflamed joint may need to be aspirated to exclude infection, especially if systemic symptoms, neutrophilia, or raised erythrocyte sedimentation rate are present.

Treatment is often unsatisfactory. Non-steroidal anti-inflammatory drugs can provide symptomatic relief, while off-loading pressure with a plaster-cast boot may temporarily halt bone destruction, but neither appears to improve eventual outcome. Pamidronate may slow the disease process by inhibiting osteoclast activity. Surgery should be avoided if possible, because the Charcot process may then spread to neighbouring joints. Occasionally, amputation is the only option for a dangerously unstable or painful foot.

Other tissue complications of diabetes

Limited joint mobility, also known as the diabetic hand syndrome or cheiroarthropathy, is probably due to glycation of collagen and other connective tissue proteins. It is particularly common in type 1 diabetes, and may develop during childhood. It causes worsening flexion deformities of the fingers so that their palmar surfaces cannot be opposed when the hands are pushed together (the 'prayer sign'), often with Dupuytren's contracture. Median and ulnar nerve lesions, presumably compressive, are often associated. Rarely, thickening of skin over the metacarpophalangeal and interphalangeal joints causes Garrod's knuckle pads (Fig. 18 and Plate 3).

Fig. 18 The hands in long-standing diabetes. (a) Limited joint mobility (cheiroarthropathy), showing the 'prayer sign'. (b) Thickening of the skin over the knuckles and proximal interphalangeal joints (Garrod's pads). (See also Plate 3.)

Necrobiosis (lipoidica diabeticorum) is strongly associated with diabetes, although 25 per cent of affected patients are normoglycaemic. Necrobiosis is the hyaline degeneration of collagen. The lesions present as trophic, non-scaling yellowish areas, often with telangiectasias (Fig. 19 and Plate 4). They are commonest on the shins but may appear elsewhere, slowly enlarge, and may perforate. Their progression is unrelated to glycaemic control. Topical or locally injected steroids may be helpful. (The histologically similar granuloma annulare is not convincingly associated with diabetes.)

Fig. 19 Necrobiosis lipoidica diabeticorum (with thanks to Dr Geoff Gill, University Hospital, Aintree, Liverpool). (See also Plate 4.)

Diabetic dermopathy ('shin spots') is the commonest skin disorder in diabetic patients and is also seen in non-diabetic patients. These atrophic brownish or erythematous lesions, usually in the pretibial area, generally cause no problems and often resolve within a year or so.

Diabetic bullae (bullosis diabeticorum) are due to subepithelial splitting and present as tense and painful blisters which appear and heal within a few weeks. Differential diagnoses include pemphigoid.

Diabetic osteopenia: poorly controlled type 1 diabetes causes general loss of bone mineral, although this does not appear to increase fracture rate significantly. Plain radiographs of the feet may show low-density phalanges that taper and may be smoothly eroded— an appearance imaginatively described as resembling partly sucked candy.

Further reading

American Diabetes Association (2000). Clinical practice recommendations 2000. *Diabetes Care* **23** (suppl. 1).

DeFronzo RA (1999). Pharmacologic therapy for type 2 diabetes mellitus. *Annals of Internal Medicine* **131**, 281–303.

European Diabetes Policy Group (1999). A desktop guide to type 2 diabetes mellitus. *Diabetic Medicine* **16**, 716–30.

Glaser N *et al.* (2001). Risk factors for cerebral edema in children with diabetic ketoacidosis. *New England Journal of Medicine* **344**, 264–9.

King H, Aubert RE, Hennan WH (1998). Global burden of diabetes, 1995–2025. Prevalence, numerical estimates, and projections. *Diabetes Care* **21**, 1414–31.

Malmberg K for the DIGAMI (Diabetes Mellitus Insulin Glucose Infusion in Acute Myocardial Infarction) Study Group (1997). Prospective randomized study of intensive insulin treatment on long-term survival after acute myocardial infarction in patients with diabetes mellitus. *British Medical Journal* **314**, 1512–15.

Pickup JC, Williams G, eds (2002). *Textbook of diabetes*, 3rd edn. Blackwell Science, Oxford. [Comprehensive, well-illustrated, and up to date for most clinical aspects of diabetes and its management.]

Ramsey LE *et al.* (1999). British Hypertension Society guidelines for hypertension management 1999: a summary. *British Medical Journal* **319**, 630–5.

Schade DS, Duckworth WC (1986). In search of the subcutaneous insulin resistance syndrome. *New England Journal of Medicine* **315**, 147–53.

Serup P, Madsen OD, Mandrup-Poulson T (2001). Islet and stem cell transplantation for treating diabetes. *British Medical Journal* **322**, 29–32.

Shapiro A *et al.* (2000). Islet transplantation in seven patients with type 1 diabetes mellitus using a glucocorticoid-free immunosuppressive regimen. *New England Journal of Medicine* **343**, 230–8.

The DECODE Study Group (1999). Glucose tolerance and mortality: comparison of WHO and American Diabetes Association diagnostic criteria. *Lancet* **354**, 617–21.

The Diabetes Control and Complications Trial Research Group (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal of Medicine* **329**, 977–86.

The Diabetes Control and Complications Trial Research Group (1995). Adverse events and their associations with treatment regimens in the Diabetes Control and Complications Trial. *Diabetes Care* **122**, 561–8.

UK Prospective Diabetes Study Group (1998). Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes. *British Medical Journal* **317**, 703–13.

UK Prospective Diabetes Study Group (1998). Intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* **352**, 837–53.

Some useful websites

American Diabetes Association: http://www.diabetes.org/

International Diabetes Federation: http://www.idf.org/

Online Diabetes Resources: www.mendosa.com/faq.htm. Useful compendium of links covering clinical practice, research, and patient-based organizations.

Diabetes UK: http://www.diabetes.org.uk/

United Kingdom Prospective Diabetes Study: http://www.drl.ox.ac.uk/

12.11.2 The genetics of diabetes

J. A. Todd

Type 1 diabetes

Type 1 is the most severe form of diabetes, with acute onset, often in children. Within white European populations, incidence varies from three cases/100 000 individuals in the population per year in Romania to over 45/100 000 per year in Finland. Its main cause is a massive and irreversible destruction of the islet b cells by the body's own immune system, in an immune-mediated autoimmune response. Life-threatening insulin deficiency results. The median age at diagnosis is about 12 years. However, as many as 25 per cent of cases are now diagnosed under age 5 years and the disease can still be diagnosed over age 60 years

The immune attack of the islets probably begins, in most cases, before age 2 years, even though diagnosis may not occur until many years later. Most autopsy pancreas samples from children who died during acute onset of the disease show immune infiltration, especially cytotoxic T lymphocytes. Newly diagnosed patients and subjects in families with the disease are positive for autoantibodies against several proteins expressed in b cells. The first of these autoantibodies to appear are those against the insulin molecule itself, and current knowledge indicates that it is a primary, aetiological antigen in b-cell destruction. The immune-mediated nature of the disease is also supported by the observations that the immunosuppressant, cyclosporin, can alter the disease course and that it very strongly associated genetically with the T lymphocyte activation genes, the HLA class II genes. The disease is clustered in families, along with other immune-mediated diseases such as thyroid disease and rheumatoid arthritis. Spontaneous rodent models of type 1 diabetes also indicate that immune-mediated destruction of b cells underlies type 1 diabetes. Most importantly, the disease can be transferred from diabetic mice to healthy recipients using T lymphocytes only. It is unlikely that the b cells themselves are passive targets of a dysregulated immune system and resistance to type 1 diabetes may well involve genes expressed in b cells.

The factors that determine the initiation and duration of the, often long, prodromal phase before diagnosis are unknown. In many countries, the incidence of type 1 diabetes is increasing dramatically, two to three-fold over the last 20 to 30 years, particularly in children under age 5 years. This cannot be due a change in the gene pool and must reflect an increasingly permissive environment. Diet and infection have both been implicated, by analogy with conditions such as coeliac disease, in which gluten is the major triggering factor, but as yet no major factor has been identified. The role of the environment cannot be underestimated. Nevertheless, there can be very few cases with a purely environmental aetiology as at least 95 per cent of cases possess the known disease-associated alleles at the major locus, the HLA class II genes. The HLA genotype is a necessary factor but not a sufficient one to explain disease occurrence. Environmental factors (not best described as 'triggers') are probably not necessary and are definitely not sufficient, but because we live lifestyles with numerous diet and infection exposures they undoubtedly play a role in most cases, either in disease initiation or precipitation in individuals genetically predisposed to the disease. It is likely that no single type 1 diabetes non-HLA gene is necessary; it remains to be determined whether combinations of non-HLA alleles is necessary. Even if the genotype of individuals is identical (monozygotic twins or inbred strains of mice) and the environment shared, for example inbred diabetic mice in a animal facility with highly regulated and uniform, infection-free environment, there is a large discordance in disease occurrence owing to unknown stochastic or random developmental factors. Despite a fully disease-permissive environment and genotype, disease might still not develop.

Disease inheritance

Type 1 diabetes is classified as a multifactorial or complex disease and not a single-gene or simple mendelian disease because possession of the disease genotype does not guarantee development of the disease—the penetrance of the genotype is less than 100 per cent. Most evidently, only 25 to 70 per cent of genetically-identical twins both develop type 1 diabetes. The disease is strongly inherited but this inheritance is complex.

By analysis of the occurrence of disease in families, the mode of inheritance was modelled most parsimoniously by a single major locus in combination with many other loci but with lesser individual effect than the major locus. This is referred to as a 'polygenic background'. It has its roots in the classical theory of the genetics of continuous traits or phenotypes for which numerous alleles, each with small effect, contribute to the overall (disease) phenotype. The transmission of disease in families, measured as the risk of disease, follows a distinctive pattern: a rapid non-linear decline in risk from first-degree relatives (6 per cent) to third-degree relatives, at which point the risk is not much different from the risk in the general population (0.4 per cent). A simple, single-gene disease exhibits a linear decline in risk. The rapid decline in risk in type 1 diabetes families is explained by the need for susceptibility alleles at several loci to be present simultaneously. Hence, marriage of a susceptible family member to a carrier of a non-susceptibility allele at one of the disease-associated loci can substantially reduce risk. An analogy is the removal of one can in a pyramid of cans resulting in collapse of the entire pile. Multigenerational families with type 1 diabetes are very rare due to this polygenic inheritance.

Molecular genetics

The major locus, at the HLA region on chromosome 6p21, is not due to allelic variation in one gene, but in several genes that can be epistatic with each other. Such complexity can only be verified by finding the actual disease loci and identifying the relevant functions of their products. A necessary step in this process is to estimate how major is the major locus. These estimates, unfortunately, depend on which genetic model is assumed and include estimates of the contribution of HLA to the clustering of type 1 diabetes in families ranging from 40 to 70 per cent. In addition, HLA-identical siblings develop type 1 diabetes less frequently than identical twins, indicating the modifying effect on HLA disease genes by non-HLA genes across the genome.

Advances in technology, and painstaking collection of data from families with two or more affected siblings, the only type of multiplex families actually available, have allowed direct evaluation of the relative contributions of different regions of the genome to familial clustering. This is referred to as linkage analysis, in which the siblings share diabetes and they also share the chromosome regions that cause the diabetes more often than expected by chance (assuming random transmission of alleles under Mendel's laws). The first genome-wide searches for linkage showed conclusively that type 1 diabetes is a major locus disease because no convincing loci other than HLA were found, despite good statistical power. This is not the case for other, closely related autoimmune diseases such as rheumatoid arthritis, autoimmune thyroid disease, or multiple sclerosis in which, although the HLA region is linked to disease, its effect is not as dominating as in type 1 diabetes. In these three diseases, it appears that no single locus has a major effect. Yet, because adopted unrelated children brought up in families with multiple sclerosis have the same risk as a randomly selected member of the general population, it is certain that familial clustering is due to shared alleles. There are, however, many disease loci (with susceptibility or resistance alleles) scattered across the genome. This situation presents a practical problem in that genes of small effect require hundreds, if not thousands, of sib-pairs to achieve convincing statistical support for deviation from random allele sharing. Except in the context of international consortia combining data sets, or perhaps in certain populations such as Iceland where remarkable family records are available, this is not feasible.

Association studies

Genes for both rare, mendelian diseases and common, complex diseases can be detected and precisely mapped by analysing the sharing of alleles by unrelated cases compared to unaffected controls. This is referred to as association or linkage disequilibrium mapping. It assumes that a proportion of living cases have inherited the same mutation from the original person in whom the disease-predisposing mutation or polymorphism arose. Cases (and controls) that carry the allele of the mutation share not only this allele but also the flanking DNA. If the mutation arose very recently in the population then the amount of flanking DNA associated with the mutation will be much larger than if, for example, the mutation predated the colonization of Europe. This is because homologous recombination of chromosomes will not have had enough time to break up the physical linkage between the mutation allele and alleles of the flanking loci that were present at the time when the mutation occurred. When recombination events between nearby loci have recombined their alleles completely, such that combinations of alleles at neighbouring loci (haplotypes) occur at frequencies no different from that expected by chance (that is the product of the allele frequencies observed in the population), then the alleles

are said to be in linkage equilibrium. When alleles occur together non-randomly, they are said to be in linkage disequilibrium.

Whereas 2000 evenly spaced, biallelic or single nucleotide polymorphisms are sufficient to search the genome for chromosome regions that show linkage to disease, over 100 000 are required to probe the genome for chromosome regions or markers that are so close to a disease causal variant that the marker alleles and the disease locus alleles are in association or linkage disequilibrium. The density of polymorphic markers will depend on the demographic history of the population and the particular region of the genome. Single nucleotide polymorphisms occur as frequently as 1 per 1000 base pairs of DNA, so that as many as 3 million polymorphisms occur in the genome. Technologies are being developed that are capable of carrying out hundreds of thousands of assays to identify alleles at polymorphic loci across the genome, especially those that occur in the regions of genes that either encode proteins or determine their expression. By resequencing the human genome from tens of individuals, hundreds of thousands of polymorphisms are being identified and mapped onto the first draft of the sequence.

Until the ability to generate and type dense maps of polymorphic markers from large chromosome regions is a reality, both technically and economically, geneticists are restricted to focusing efforts on candidate genes—genes for which the expression or presumed function might be of relevance to the development of the disease. This more limited approach has historically been moderately successful. The candidate gene-association approach was used to identify the only two known genes for type 1 diabetes, the HLA class II genes and the repetitive DNA polymorphism in the 5' regulatory region of the insulin gene, on chromosomes 6p21 and 11p15, respectively (designated *IDDM1* and *IDDM2*). As marker typing technology improves, candidate genes will be analysed in much larger numbers.

Linkage studies in type 1 diabetes are continuing and meta-analyses of over 1000 affected sib-pair families are underway. This number of families provides the statistical power to detect disease loci with intermediate effects. Chromosome regions that show significant evidence of linkage to disease will be subjected to a more systematic association analysis in which very dense maps of polymorphisms will be developed, in the order of several single nucleotide polymorphisms per 10 kb of DNA, and typed in cases, controls, and families to find disease-associated regions that explain the linkage of the broader chromosome region to the disease. This is still a daunting task given current technology. Disease-linked chromosome regions usually encompass 15 cM to 15 Mb of DNA, indicating that at least 1500 evenly spaced polymorphic markers might have to be typed before the disease locus is located. This assumes that the linked region contains one disease locus, which seems unlikely given data from the mouse model of type 1 diabetes, and the fact that thousands of genes function in the immune system, for which a high polymorphism is beneficial in the defence against infection. Before embarking on such studies, it will be necessary to be confident that the evidence for linkage of the chromosome region to disease is convincing (P, probability, value less than 2×10^{-5}) and that the association study has sufficient statistical power to detect the association at P values in the range of 10^{-5} to 10^{-8}. These low P values take into account the very large number of tests carried out in studies of this kind. If a linked region contains several disease genes and each has several alleles, some or all which might be rare (1–5 per cent allele frequency in the population), then the prospects for sufficiently powerful studies are not favourable. If association studies are undertaken without prior evidence of linkage then, in the absence of convincing data implicating the function of the gene in disease, P values in the order of 10^{-8} are required to achieve results significant at the 5 per cent level (P = 0.05), taking into account the hundreds of thousands of tests that would be conducted across the whole genome. This assumes that hundreds of thousands of polymorphic markers, or at least all markers required to detect all haplotypes and constitute a continuous linkage disequilibrium map across the entire genome, have been typed in the study or are anticipated to be typed in future studies worldwide. Until technology is available that can tackle a 'whole genome assocation analysis' studies will be restricted to candidate genes or chromosome regions linked to disease.

Type 1 diabetes genes

We know the identity of two of the genes (termed insulin-dependent diabetes mellitus, IDDM, genes): *IDDM1* is comprised of a number of genes of related function, the human leucocyte antigen (HLA) class II genes, encoded in tandem in the major histocompatibility complex on chromosome 6p21. The *HLA-DQB1*, *-DQA1*, and *-DRB1* loci act as a 'superlocus' in disease with many alleles at the three different, but adjacent and homologous, loci interacting in epistatic ways. Their products allow T lymphocytes to recognize foreign and self-proteins in the form of peptide bound to the surface-expressed HLA class II molecules. They ensure that the body's immune system can recognize self-proteins and remain tolerant to them. Certain alleles, the most potent of which are referred to collectively as DR3 and DR4, are permissive for the anti-b-cell immune response, whereas other alleles, such as DR2, are dominantly protective against type 1 diabetes. The precise mechanisms underlying HLA class II mediated susceptibility and resistance are still uncertain. Moreover, this information, concerning a necessary and key step in disease pathogenesis, has not been translated into a way of modulating disease susceptibility. Although the translation of genetic–physiological research results such as these into therapy is the main goal, early attempts to block or modify HLA class II function directly in autoimmune disease have not been successful.

The second locus is referred to as *IDDM2*, now identified as a polymorphic DNA sequence composed of repeats of short oligonucleotides embedded in the 5' regulatory region of the insulin gene on chromosome 11p15. The two main variants of this polymorphism differ in size (that is number of repeats), sequence, and in their regulation of expression of the insulin gene. This differential regulation of insulin at the transcriptional level varies according to the tissue; very low levels of insulin are expressed in the thymus, the organ in which HLA class II molecules mediate the establishment and maintenance of immune tolerance. The type 1 diabetes protective allele of the insulin gene 5' repeat polymorphism is associated with increased expression of insulin (and its precursor preproinsulin) in the thymus. Since insulin is a key autoantigen in type 1 diabetes, a plausible model for the physiological action of this polymorphism in disease protection is increased tolerance to insulin and its precursors owing to their increased expression, and consequently HLA class II molecule-mediated immune tolerance. Peptides from insulin (and/or its precursors) bind to the HLA class II molecules in the generation of a healthy, b-cell tolerant T lymphocyte repertoire. This model fits much of the available data, particularly from rodent models of the disease. There are, however, missing links; for example the precise peptides involved, the affinity with which they bind to different HLA class II allotypes, and the consequences of these events for the immune system downstream of the HLA class II peptide T lymphocyte antigen receptor are uncertain. Again, it is not clear how this presumed pathway can be modulated to restore immune function to a level at which an individual is resistant to disease. This challenge is particularly acute since the very first events in disease and in immune tolerance may occur before age 2 years.

The risk of carrying disease-susceptible HLA class II and insulin gene alleles does not exceed 30 per cent. Informing a family member or an individual of this risk is not useful, and maybe harmful, in the absence of a safe way to modulate this risk and restore normal function. In a research context, the longitudinal follow-up of children from birth who carry the HLA susceptibility alleles or have type 1 diabetic parents could provide insights into the environmental factors that determine the development of the disease. The typing of protective alleles, such as HLA DR2, or more precisely the DQB1*0602 allele, already has some diagnostic benefit in clinical trials to prevent the disease. Positivity for the protective allele provides a criterion for exclusion from a trial, since DR2/DQB1*0602 positive subjects are at least 50 times less likely to develop type 1 diabetes than a randomly selected member of the population.

Current research

The search is proceeding for other type 1 diabetes genes using the genetic approaches of linkage and association. In concert with these efforts new tools are being applied, including the mass, parallel analysis of gene transcription using microarrays, the analysis of protein expression and modification using mass spectroscopy, the analysis of events *in vivo* using imaging techniques, and the investigation of cell metabolism using nuclear magnetic resonance spectroscopy. These approaches, in combination, will provide further details of immune system and islet cell physiology in the search for pathways and pathway–environmental factor interactions that could be manipulated to prevent the disease or rejection of transplanted islets or b cells.

Further reading

Bach JF (1994). Insulin-dependent diabetes mellitus as an autoimmune disease. *Endocrinology Reviews* **15**, 516–42.

Pipeleers D, Ling Z (1992). Pancreatic beta cells in insulin-dependent diabetes. *Diabetes and Metabolism Reviews* **8**, 209–27.

Risch NJ (2000). Searching for genetic determinants in the new millennium. *Nature* **405**, 847–56.

Rose NR, Mackay IR, eds (1998). *The Autoimmune Diseases*, 3rd edn. Academic Press, San Diego.

Strachan T, Read AP (1999). *Human molecular genetics 2,* 2nd edn. BIOS Scientific Publishers, Oxford.

Tisch R, McDevitt H (1996). Insulin-dependent diabetes mellitus. *Cell* **85**, 291–7.

Todd JA (1999). From genomics to aetiology in a multifactorial disease, type 1 diabetes. *Bioessays* **21**, 164–74.

Weiss KM, Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? *Nature Genetics* **26**, 151–7.

12.11.3 Hypoglycaemia

V. Marks

Introduction

Hypoglycaemia means low blood glucose concentration (<3.0 mmol/l). It is not a disease entity but a biochemical abnormality whose importance lies in its effects upon brain function. These are responsible, directly or indirectly, for the signs and symptoms produced by hypoglycaemia and often provide the first clue to the presence of curable or preventable disease.

The brain obtains glucose from the blood by means of facilitated transport and mainly utilizes the glucose-transporting protein GLUT 1. The activity of this protein is increased by hypoglycaemia and reduced by hyperglycaemia. It is not insulin-dependent. Although blood glucose concentration is the single most important factor determining glucose availability to the brain, it is not the only one. Indeed there is a poor correlation between blood glucose concentration and the severity and nature of cerebral symptoms—especially in diabetic patients. It is therefore important to distinguish hypoglycaemia, a description of blood glucose, from neuroglycopenia, which is responsible for the signs and symptoms to which hypoglycaemia gives rise.

Four distinct, but not mutually exclusive, neuroglycopenic syndromes are recognized:

1. Acute neuroglycopenia, the most common, is normally associated with iatrogenic and experimental hypoglycaemia and is characterized by profuse sweating, anxiety/nervousness, tremor, tachycardia, hunger, and paraesthesia—all of which can be attenuated by adrenergic and cholinergic blockade—and by speech and visual disturbances, unsteady gait, confusion, and a sense of fatigue, which are independent of the autonomic nervous system.
2. Subacute neuroglycopenia occurs in most varieties of spontaneous hypoglycaemia and, when it occurs in insulin-treated diabetic subjects, is called hypoglycaemia unawareness. It is characterized by a reduction in spontaneous movements and speech, somnolence, inefficient cerebration and work performance, personality change, and amnesia of varying severity. Other signs and symptoms common to acute and subacute neuroglycopenia include transient hemiplegia, hypo- or hyperthermia, convulsions, diploplia, and strabismus. The symptoms of both acute and subacute neuroglycopenia are ephemeral but, unless aborted by restoration of normoglycaemia, can progress to stupor, coma, or, in exceptional cases, death from cerebral oedema.
3. Chronic neuroglycopenia is rare and virtually confined to patients with hypoglycaemia due to insulinoma or diabetic patients over-zealously treated with insulin. It is characterized by insidious changes in personality, defective memory, psychosis—often with paranoid features—or mental deterioration resembling dementia.
4. Hyperinsulin neuronopathy, the clinical features of which may be mistaken for motor neurone disease, is a form of chronic neuroglycopenia. Temporary elevation of the blood glucose level has no discernible effect on cerebral or neuronal function but removal of the causative agent often does over the course of a year or two.

Experimental and iatrogenic hypoglycaemia

There is a hierarchy in the activation of brain centres as hypoglycaemia develops which is, however, not often observed in spontaneous hypoglycaemia. The tissue most sensitive to a falling blood glucose level is the normal pancreatic b-cell which virtually ceases secreting insulin as blood glucose concentrations fall to about 4.0 to 4.2 mmol/l. The sympathetic nervous system is activated and glucagon is secreted as the blood glucose concentration reaches about 3.7 mmol/l. Stimulation of growth hormone secretion occurs at glucose levels of about 3.5 mmol/l and that of ACTH and cortisol at about 3.3 mmol/l. The threshold for vasopressin release has not been determined. Most subjects experience symptoms only after their blood glucose concentration has fallen below 3 mmol/l but objective evidence of minor cognitive impairment, of which the subject is usually completely unaware, occurs at concentrations nearer 4 mmol/l.

Patients, whether diabetic or not, with recent experience of hypoglycaemia often tolerate lower blood glucose levels before symptoms develop and counter-regulatory hormones are secreted. They remain, however, just as sensitive to the deleterious effect of hypoglycaemia on cognitive function.

Some of the cerebral symptoms of neuroglycopenia and effects upon neuronal viability are due to liberation of the excitatory amino acids, glutamate and aspartate, rather than solely to decreased intracellular energy production.

Definition

Hypoglycaemia is defined arbitrarily by the blood glucose concentration; it is not determined by whether symptoms are present or not. For most purposes, an arterial (or capillary) blood glucose concentration below 3.0 mmol/l can be considered diagnostic of hypoglycaemia and one of 2.5 mmol/l or less pathological and demanding

of investigation as to cause. The possibility that a patient's symptoms are of neuroglycopenic origin should not be dismissed solely on the basis of blood glucose concentration. Normoglycaemic neuroglycopenia must be considered.

Classification

Iatrogenic hypoglycaemia as a consequence of insulin or sulphonylurea treatment for diabetes is common and accounts for most hypoglycaemia encountered in practice. It seldom presents diagnostic difficulties. There are, on the other hand some 100 or so causes of spontaneous hypoglycaemia, all of which are rare. Collectively, they are responsible for 0.1 per cent of all patients arriving in accident and emergency or medical investigation departments.

Table 1 lists the main causes of hypoglycaemia, which vary in frequency from country to country. Although all may occur in infants and children, the main causes of hypoglycaemia in this age group are not usually encountered in adults.

Presentation

Patients classically present either to accident and emergency in a stuporose or comatose state with concurrent hypoglycaemia or to outpatient departments with a normal blood glucose level but a history suggestive of recurrent neuroglycopenic episodes or progressive neurological/psychological dysfunction.

Management of the stuporose/comatose hypoglycaemic patient

Hypoglycaemia should be suspected in any case of altered consciousness, coma, hemiplegia, apparent alcoholic intoxication, or epilepsy, and eliminated or supported (though not established) by a point-of-care blood glucose determination. The diagnosis is confirmed by formal laboratory glucose analysis. Hypoglycaemia may also be caused by, and contribute to the symptomatology of, congestive cardiac failure, liver or kidney disease, malaria, and other severe infections. Management falls quite clearly into two separate phases.

Emergency treatment

Glucose 25 g (50 ml of 50 per cent w/vol) should be given intravenously to alleviate hypoglycaemia after sufficient venous blood (20 to 30 ml) has been withdrawn for subsequent laboratory analysis to determine its cause. Glucagon 1 mg may be given intramuscularly if venous access is not available, especially in cases of iatrogenic hypoglycaemia (in which it is usually effective).

Recovery of consciousness ordinarily occurs within 10 min. A further injection of 25 g glucose plus 100 mg hydrocortisone is indicated if recovery is delayed beyond 20 min. Specific measures to reduce brain swelling must be introduced if recovery does not occur within a further 20 min.

Prolonged, formerly called irreversible, hypoglycaemic coma is due to cerebral oedema and a consequence of profound hypoglycaemia generally lasting 5 h or more. Its treatment includes the use of intravenous mannitol and dexamethasone. Blood glucose must be monitored constantly and sufficient glucose infused to keep it within the range 5 to 10 mmol/l until consciousness is restored or permanent brain damage established. In cases of suicidal insulin or sulphonylurea overdose, glucose in doses up to 80 g/h given as a 25 to 50 per cent solution through a central line may be required.

Investigation

The second stage and third stages are similar to those employed in investigating patients suspected of suffering from a hypoglycaemic disorder but currently asymptomatic (Fig. 1).

Fig. 1 Investigation of a patient suspected of suffering from hypoglycaemia who is hypoglycaemic (but not unwell from some other cause, e.g. congestive cardiac failure, septicaemia, liver or renal failure) at the time of examination. It is customary to measure plasma total insulin immunoreactivity (IRI), C-peptide, proinsulin, b-hydroxybutyrate, growth hormone, IGF-1 and IGF-2, alcohol, and sulphonyureas simultaneously or sequentially on the initial hypoglycaemic blood sample. + to ++++: insulin >30 to 300 000 pmol/l; C-peptide >150 to 10 000 pmol/l; proinsulin >20 pmol/l; GH ³ 5 mU/l; BHB >600 µmol/l; alcohol ³2 to 100 mmol/l. –ve: insulin <25 pmol/l; C-peptide <100 pmol/l; GH <1 mU/l; B-OH <600 µmol/l; IGF-1 < 10nmol/l; IGF-2 <45 nmol/l. Abbreviations: IEM = inborn errors of metabolism, NICTH = non-islet cell hypoglycaemia, GH = growth hormone, B-OH = b-hydroxybutyrate, IR-AA = insulin receptor autoantibodies, AIS = autoimmune insulin syndrome.

Management of the asymptomatic patient suspected of having a hypoglycaemic disorder

Diagnosis takes place in three sequential stages:

1. suspicion of hypoglycaemia and its confirmation by measurement of the blood glucose concentration during a 'spontaneous' neuroglycopenic episode;
2. determination of its aetiology on the basis of specific investigative procedures;
3. localization of the lesion responsible if the hypoglycaemia has an anatomicopathological rather than a purely metabolic aetiology.

Confirmation of hypoglycaemia

Most patients, excepting those presenting in stupor or coma, are normoglycaemic and asymptomatic when first seen. Suspicion of hypoglycaemia is aroused by a history of subacute neuroglycopenia, that is episodes of altered behaviour or disturbed consciousness or, in a minority of cases, symptoms suggestive only of acute neuroglycopenia. Because of the nature of their illness, patients are often unable to supply a reliable history of their condition.

Exclusion or confirmation that a patient's symptoms are hypoglycaemic in origin can often be achieved by teaching them, or their relatives, to collect capillary blood during spontaneous symptomatic episodes occurring in the course of everyday life. Blood collected into specially prepared tubes or filter paper should be sent to the laboratory for glucose analysis since point-of-care monitoring systems are insufficiently reliable in the hypoglycaemic range to warrant initiation of detailed investigation and may cause confusion. A blood glucose concentration during a symptomatic episode greater than 3.5 mmol/l effectively eliminates hypoglycaemia as its cause. Glucose concentrations lower than this are unusual and require further investigation.

Hyperinsulinism

This is a misnomer for the syndrome to which insulinoma gives rise. It would better be called dysinsulinism since its hallmark is inappropriate rather than excessive secretion of insulin or proinsulin.

Insulinoma

Insulin-secreting tumours (insulinomas) are the most common type of neoplasm affecting the endocrine tissues of the pancreas. They have an incidence of one case or more per million of the population. Eighty per cent of insulinomas are benign and solitary, 7 to 10 per cent are multiple—often as part of the MEN 1 syndrome—and 8 to 10 per cent malignant. They occur at any age but are rare before the age of 10 and infrequently diagnosed after the age of 70. The lack of cases after age 70 may be due to their mode of presentation, which is often that of progressive dementia, rather than to their rarity. There is a 6:4 ratio in favour of women for benign but not for malignant tumours.

Insulinomas are composed mainly, or exclusively, of b-cells. Most are between 10 and 20 mm in diameter at diagnosis, though tumours as small as 5 mm in diameter have been associated with severe symptoms. Regardless of size, they occur at all sites in the pancreas with equal frequency.

Histological classifications, whilst valuable for the light they throw on insulin secretory mechanisms, contribute little to clinical management. Malignant insulin-secreting tumours are impossible to distinguish, clinically or histologically, from benign ones unless metastases are present. Some have the histological appearance of carcinoid tumours and both may contain and secrete other peptide hormones of which glucagon, somatostatin, ACTH, and GhRH are amongst the commonest. Only rarely, however, do these biochemical endocrinopathies manifest themselves clinically. There is no evidence that malignant tumours ever begin as benign tumours or that benign tumours ever become malignant.

The average time between the onset of symptoms and diagnosis of insulinoma is currently about a year but symptoms persisting over 30 years or more without evidence of permanent brain damage are not unknown. Diagnostic delays are usually due to reluctance by patients to seek help or failure by clinicians to suspect hypoglycaemia, rather than any difficulties in confirming the presence of an insulinoma once the possibility has been considered. Only very rarely is an insulinoma found at autopsy as the cause of unexplained death.

In a minority, probably not exceeding 1 to 2 per cent, functionally defective b-cells are distributed throughout the pancreas rather than in discrete tumours. Clinically and biochemically, such patients are indistinguishable from patients with insulinomas. Biologically, such patients resemble infants with persistent hyperinsulinaemic hypoglycaemia of infants (formerly nesidioblastosis).

Chemical pathology

Endogenous hyperinsulinism is characterized by failure of the abnormal b cells to stop secreting insulin in response to hypoglycaemia. This is ordinarily the most sensitive physiological response to a falling blood glucose concentration and becomes apparent at a level (4.2–4.0 mmol/l) well above the threshold for neuroglycopenic symptoms. A consequence of insulin secretion persisting during fasting is inhibition of hepatic glucose release and a gradual fall in blood glucose to below the level capable of sustaining normal brain function.

Paradoxically, the functionally abnormal b cells are often insensitive to hyperglycaemia *per se* and so produce glucose intolerance as well as fasting hypoglycaemia. They do, however, respond, often excessively, to other insulin secretagogues including glucagon, sulphonylureas, L-leucine, and the intestinal incretins GIP and GLP-1, and may therefore present with reactive rather than fasting hypoglycaemia.

Typically, plasma cortisol and growth hormone levels in patients with insulinomas are normal even in the presence of hypoglycaemia. This would ordinarily be considered evidence of hypothalamicopituitary insufficiency but responsiveness returns after restoration of permanent normoglycaemia. Plasma free fatty acid and b-hydroxybutyrate concentrations are typically suppressed (<600 µmol/l) but rise, though not to expected levels, during prolonged fasting.

Diagnosis

Diagnosis is made by demonstrating that the symptoms are caused by hypoglycaemia, provoked by fasting and/or rigorous exercise, relieved by intravenous glucose, and are caused by inappropriate insulin and/or proinsulin secretion. Plasma concentrations of total immunoreactive insulin, C-peptide, proinsulin, and proinsulin-like fragments are all high with regard to the prevailing blood glucose concentration but are not necessarily high in absolute (quantitative) terms.

Thus in the presence of concurrent hypoglycaemia (blood glucose less than 3 mmol/l), plasma total immunoreactive insulin concentrations of more than 30 pmol/l and C-peptide concentrations more than 100 pmol/l are inappropriately high. When both peptide levels are inappropriately high, a diagnosis of endogenous hyperinsulinism is virtually certain, providing sulphonylurea ingestion and various rare autoimmune diseases and infections, such as malaria, can be excluded. If hypoglycaemia is absent at the time of sampling plasma insulin, C-peptide and proinsulin assays become uninterpretable.

Fasting under close observation for up to 72 h produces symptomatic hypoglycaemia with inappropriate hyperinsulinaemia (proinsulinaemia and C-peptidaemia) in over 98 per cent of insulinoma patients but not in healthy men and women who, if they do become hypoglycaemic, invariably have appropriately suppressed plasma insulin levels. As an alternative to prolonged fasting, the overnight fasted patient can be exercised to exhaustion on a treadmill. In insulinoma patients, this fails to produce the normal suppression of plasma insulin and C-peptide secretion. It is, however, rarely necessary to subject a patients to these tests, especially if investigations are restricted to those who have blood glucose levels below 3 mmol/l during spontaneous episodes occurring in every-day life. Dynamic function tests including oral glucose, tolbutamide, glucagon, L-leucine, and insulin–hypoglycaemia/C-peptide suppression tests are unnecessary for the diagnosis of hyperinsulinism.

Some 5 to 10 per cent of insulinomas secrete only, or mainly, proinsulin and thus the diagnosis may be missed if an insulin-specific assay, rather than one capable of detecting total immunoreactive insulin, is used. Moreover, unusually efficient extraction of insulin by the liver can lead to low plasma total immunoreactive insulin concentrations in peripheral blood in the presence of genuinely inappropriate insulin secretion. This can occur in infants with nesidioblastosis as well as in adults with endogenous hyperinsulinism in whom inappropriately high plasma C-peptide levels will confirm the diagnosis. Hyperproinsulinaemia, that is a plasma proinsulin concentration of greater than 20 pmol/l, is found in some 95 per cent of patients with endogenous hyperinsulinism; its absence should raise doubts about the accuracy of the diagnosis.

Pre- and intraoperative localization

A diagnosis of endogenous hyperinsulinism, established on the basis of inappropriate hyperinsulinaemia, is almost synonymous with one of insulinoma. The treatment of choice is surgical ablation. Localization by an experienced surgeon at laparotomy is remarkably (96 per cent) successful but can be further improved by use of intraoperative ultrasound.

Though virtually every imaging technique has been advocated for preoperative localization of insulinoma, none is sufficiently reliable to justify dismissing a diagnosis made on sound clinical and biochemical grounds. Endoscopic ultrasound, with a 90 per cent prediction rate, and pancreatic intra-arterial calcium injection with hepatic venous sampling are currently the only imaging techniques that are useful for localization prior to operation. Venous sampling is especially indicated when surgery has failed to reveal a tumour and/or diffuse islet hyperplasia is suspected. It is the only way of establishing a diagnosis of non-insulinoma pancreatogenic hypoglycaemia.

Treatment

Surgical ablation ensures an excellent prognosis with no reduction in life expectancy except when the tumour is malignant. Even then, since these tumours grow slowly and rarely spread beyond the liver, removal of the primary tumour, and as many hepatic secondaries as possible, may add years of useful life. Operative mortality for adenomas is under 2 per cent, except in the elderly. Benign tumours recur in up to 5 per cent of patients.

In patients over 70 years of age, and others in whom surgery is impracticable, treatment with diazoxide (200–600 mg/day) combined with chlorothiazide (1 g/day) to increase its effectiveness, is well tolerated. It is the treatment of choice in hyperinsulinism due to diffuse islet hyperplasia, non-insulinoma pancreatogenic hypoglycaemia, and after surgical debulking in cases of metastatic insulinoma. Only when diazoxide/chlorothiazide treatment fails to relieve hypoglycaemia are other drugs, such as octreotide, b-blockers, or calcium channel blockers worth trying. In patients with malignant insulinomas, embolization or surgical debulking of hepatic

metastases may produce remissions lasting several years—as may treatment with cytotoxic agents such as streptozotocin and 5-fluorouracil.

Non-islet cell tumour hypoglycaemia (NICTH)

The symptoms of hypoglycaemia produced by non-islet cell tumours (NICTH) may be indistinguishable from that of insulinoma. The symptoms are almost invariably those of subacute neuroglycopenia and the features of autonomic nervous activation are absent. Biochemically, NICTH is characterized by fasting hypoglycaemia, hypoketonaemia, and low plasma total immunoreactive insulin, C-peptide, and proinsulin levels. Growth hormone, ACTH, and glucagon secretion are depressed during both hypo- and normoglycaemia and plasma IGF-1 levels are always low—unlike in insulinoma when they are normal or high.

NICTH can occur with almost every histological type of malignant tumour but is rare. Although sarcomas are disproportionately well represented, less than 1 per cent of them develop hypoglycaemia. It is, however, common in patients with haemangiopericytomas, which are themselves rare. Amongst the carcinomas, no histological type is exempt from NICTH but only in primary hepatomas is it at all common.

Chemical pathology

Regardless of histological type, hypoglycaemia due to non-islet cell tumours (NICTH) results from overproduction of an abnormally large form of IGF-2. This has many of the biological and immunological properties of IGF-2 itself but binds less avidly to plasma IGF binding proteins (IGF-BP) of which there are at least six normally present in plasma.

Big IGF-2 is generated by the removal of a 24 amino acid leader sequence from the N-terminus of prepro IGF-2. Normally, it then undergoes cleavage at its carboxy terminus to produce regular IGF-2. Failure to do so leaves the E-domain intact and the secretion of big IGF-2—the cause of NICTH.

There is characteristically a marked reduction in the most plentiful of the plasma binding proteins, IGF-BP3, and a partial compensatory increase in IGF-BP2 the net effect of which is, however, to reduce IGF protein binding capacity. Consequently plasma 'free (big) IGF-2' is increased without any corresponding increase in total immunoreactive IGF-2.

The exact mechanism by which big IGF-2 produces hypoglycaemia is unknown and may involve several steps, the most important of which is activation of insulin and IGF receptors on peripheral tissues and their increased uptake of glucose. The next most important is suppression of glucagon and growth hormone secretion.

Ectopic insulin secretion

Ectopic insulinomas are confined to the duodenum and are rare (<1 per cent). Ectopic insulin production by a non-islet cell tumour is extremely rare and has been established in only one case and suggested in five others. The coincidence of an insulinoma and another type of tumour has been described rather more often.

Diagnosis

The diagnosis of NICTH is seldom in doubt once thorough investigations into the cause of hypoglycaemia have been initiated:

1. Hypoglycaemia, once it has developed, seldom remits for more than very brief periods after meals.
2. The tumours are usually, though not invariably, sufficiently large to reveal themselves either on physical examination or as a result of comparatively straightforward imaging.

In the laboratory, findings of low plasma insulin, C-peptide, and proinsulin concentrations (<30, <100, and <20 pmol/l respectively) in the presence of hypoglycaemia and hypoketonaemia are highly suggestive of NICTH. Clinical laboratory assays typically measure both big and regular forms of IGF-2 and generally reported as normal (50–100 nmol/l) but IGF-I levels are invariably low (<10 nmol/l). Consequently, plasma IGF-2:IGF-1 ratios, expressed on a molar basis, are abnormally high (>10) and not seen in any other condition except gross undernutrition. Assays for the E-domain of proIGF-2 have been developed. Although useful for establishing recurrence, they provide less accurate initial diagnostic information than the IGF-2:IGF-1 ratio.

Treatment

Treatment of choice is surgical. In rare cases of benign tumour NICTH, the cure is permanent. In malignant cases, ablation or debulking of secondaries may produce prolonged remissions. Symptomatic relief may be obtained by the use of a combination of diazoxide and chlorothiazide but less predictably than with insulinoma. Prednisolone, in doses up to 60 mg/day, produces improvement in the biochemical profile and remissions from hypoglycaemia in many cases but has no effect upon tumour growth itself. Growth hormone and long-acting glucagon preparations also produce symptomatic relief given alone or with prednisolone.

The postprandial syndrome

The appearance of symptoms suggestive of acute neuroglycopenia in relation to the ingestion of food has been called the postprandial syndrome. It has many causes, one of the less common is hypoglycaemia.

Reactive hypoglycaemia

Following an initial rise, venous blood glucose concentrations may decrease in normal healthy volunteers as far as 2 mmol/l below fasting levels after ingestion of a liquid glucose load of 75 g or more on an empty stomach. A smaller fall in arterial blood glucose also occurs and may, in up to 50 per cent of normal healthy subjects, be accompanied by mild symptoms. This phenomenon, referred to as reactive hypoglycaemia, rarely occurs in every day life when normal mixed meals are eaten. When it does, diagnostic difficulties may arise since symptoms are usually vague, unspecific, and indistinguishable from those due to other illnesses, especially neurosis.

In the period 1950 to 80, the diagnosis of reactive hypoglycaemia, referred to by lay writers simply as 'hypoglycaemia', reached epidemic proportions in the United States. In most cases, the diagnosis was based on misattribution of the normal response to oral glucose. Whilst some patients with postprandial syndrome may have a lower threshold to neuroglycopenia, experiencing symptoms at (arterial) blood glucose levels of 3.5 to 4.0 mmol/l rather than the more customary level of 2.8 to 3.3 mmol/l, most do not. Nor do they manifest any abnormalities of glucose homeostasis.

Criteria for the recognition and diagnosis of reactive hypoglycaemia were laid down at the Third International Symposium on Hypoglycaemia, adherence to which has greatly reduced the number of persons misdiagnosed. The criteria include a history of food-stimulated autonomic symptoms appropriate to acute neuroglycopenia—a capillary blood glucose concentration measured during a spontaneous symptomatic episode below 3 mmol/l and rapid relief by oral glucose. Sometimes, when suspicion is high and blood collection during every day life proves difficult, it may be necessary to give the patient a standard meal and observe the glycaemic, symptomatic, and electroencephalographic responses over the ensuing 5 h. The oral glucose load test is not appropriate.

The term reactive hypoglycaemia is not a definitive diagnosis in its own right; it is only the first step towards determining causation. Almost every condition in which hypoglycaemia is induced by fasting, may present as reactive hypoglycaemia. Organic causes, including acquired and inherited metabolic derangements, must, therefore, be eliminated before making a diagnosis of idiopathic reactive hypoglycaemia—which is rare. Conditions in which patients experience only reactive, but not fasting, hypoglycaemia include partial gastrectomy and jejuno–oesophageal anastamosis (also referred to as alimentary hypoglycaemia), autoimmune insulin syndrome, and, recently identified, non-insulinoma pancreatogenous hypoglycaemia. Reactive hypoglycaemia, unaccompanied by fasting hypoglycaemia, occurs in up to 2 per cent of patients harbouring insulinomas.

Clinical features

Typically, patients present with a history of transient episodes of dizziness, anxiety, palpitations, sweating, hot flushes, and even convulsions or brief periods of altered consciousness extending over a period of 1 to 30 years. Between episodes they are well and asymptomatic. Patients rarely notice any relationship of

symptoms to food but may do so when prompted. Physical, including radiological, investigation is generally normal except in alimentary hypoglycaemia. In them, but few others, food-induced reactive hypoglycaemia may be of sufficient severity as to cause loss of consciousness.

Acute-neuroglycopenia-like symptoms experienced by some patients with the postprandial syndrome are rarely associated with any abnormality of glucose homeostasis or insulin secretion though exaggerated enteroglucagon, GIP, and *t*GLP-I response to food may occur. What role, if any, these hormones play in the aetiology of the syndrome is unknown. Some patients with these symptoms are sensitive to modest reductions in blood glucose concentration to which most healthy subjects would be oblivious and in them the possibility of non-hypoglycaemic neuroglycopenia may be entertained.

Alcohol-induced reactive hypoglycaemia

Symptomatic reactive hypoglycaemia may occur in healthy young subjects after ingesting a mixture of alcohol, sucrose, and quinine given as gin and tonic and, less commonly, with other mixtures of alcohol and carbohydrate on an empty stomach. Simultaneous ingestion of carbohydrate-rich snacks increases the severity of the hypoglycaemia; snacks rich in fat reduce it.

Diagnosis

Diagnosis of reactive hypoglycaemia is suggested by the clinical history and confirmed or refuted by glucose measurements made on capillary blood collected during spontaneous symptomatic episodes. Other laboratory tests, including measurement of plasma insulin, C-peptide, proinsulin, and b-hydroxybutyrate are used to exclude conditions such as non-insulinoma pancreatogenous hypoglycaemia, autoimmune insulin syndrome, and other conditions that do not produce hypoglycaemia during fasting but do require specific treatment.

Capillary blood glucose concentrations of less than 3.5 mmol/l, measured in an accredited laboratory on two or more occasions, establishes hypoglycaemia as a factor in the symptomatology. The oral glucose load test, formerly the lynch-pin for diagnosis, may be frankly misleading especially when conducted on individuals who have taken a self-prescribed low carbohydrate diet (<100 g/day) and should rarely be employed. A standard glucidic breakfast providing 100 g of readily assimilated starchy food has been advocated in its stead but is seldom indicated.

Treatment

Prevention of fluctuations in blood glucose is key to the management of reactive hypoglycaemia and is achieved by minimizing intake of rapidly absorbed carbohydrates such as sucrose, bread, and potato starch. Frequent small meals, rich in dietary fibre—and taken without alcohol—offer the best chance of symptomatic relief. Incorporation of soluble dietary fibre supplements, such as guar and glucomannan, in meals and taking a-glucosidase inhibitors, such as acabose and miglitol, with them reduce blood glucose excursions but their side-effects are often worse than discomfort from minimal hypoglycaemia.

Prognosis

Idiopathic postprandial syndrome is a self-limiting disorder but may be resistant to all physical treatments. Some patients respond well to psychotherapy and/or avoidance of alcohol.

Autoimmune hypoglycaemia

Autoimmune diseases are important causes of spontaneous hypoglycaemia. Three main types are recognized.

Autoimmune insulin syndrome

The autoimmune insulin syndrome occurs throughout the world but is rare outside Japan. It is due to polyclonal autoantibodies to insulin resembling those produced in response to exogenous insulin but more likely to bind proinsulin and its cleavage products including C-peptide.

Hypoglycaemia typically occurs as a late response to the ingestion of food. Insulin secreted early in response to a meal is sequestered by antibodies present in the plasma and rendered temporarily inactive. Dissociation of the insulin–antibody complex, after absorption is complete, produces an inappropriately high free plasma insulin level resulting in hypoglycaemia. This, though often profound, is of limited duration, rarely leading to coma and never to death.

There is often a history of autoimmune disease affecting other organs, especially the thyroid, and many patients have received treatment with methimazole, carbimazole, or other thiol-containing drugs.

Free plasma insulin concentrations are always inappropriately high and C-peptide usually depressed during hypoglycaemia. C-peptide concentrations may, however, be normal or high depending on the binding characteristics of the autoantibody.

Treatment is dietary and aimed at avoiding excessive insulin secretion in response to meals until spontaneous remission occurs, usually within a few years of onset. Surgery, in the mistaken belief that the patient has islet hyperplasia or insulinoma, must be avoided.

Insulin-receptor autoantibodies

Hypoglycaemia due to insulin-receptor autoantibodies is rare but may be the first indication of the causative disease. More often it develops in a patient already known to be suffering from an autoimmune disease or a neoplasm—especially lymphoma. Typically, hypoglycaemia is intractable but occasionally occurs only in response to food. Its immediate cause is binding of stimulatory autoantibodies to insulin receptors on hepatic and peripheral cell membranes, simulating the effects of insulin itself.

Clinically, the symptoms are indistinguishable from that of insulinoma though usually of shorter duration and greater severity. Plasma C-peptide and proinsulin concentrations are low (<20 pmol/l). Plasma insulin, though also often low, may be very high (>1000 pmol/l) due to its delayed clearance from the blood. Diagnosis can usually be inferred from the clinical associations and evidence suggestive of hyperinsulinism, that is coincident low blood glucose and b-hydroxybutyrate, but depressed plasma C-peptide, proinsulin (and usually insulin), concentrations rule it out. Definite diagnosis depends upon demonstrating antireceptor antibodies in the patient's plasma using *in vitro* bioassay techniques.

Treatment is that of the primary disease. Glucocorticoids and other immunosuppressants have been used with benefit in some cases but although remissions may occur, the prognosis is generally poor.

Islet-cell-stimulating antibodies

Antibodies capable of stimulating insulin release from isolated pancreatic b-cells *in vitro* have been held responsible for a form of hyperinsulinaemic hypoglycaemia analogous to Grave's disease of the thyroid. The evidence is, however, inconclusive.

Drug and toxin-induced hypoglycaemia

Medicines and toxins, such as alcohol, paracetamol, quinine, *Amanita* (toadstools), and *Blighia* (ackee) are collectively amongst the most frequent causes of non-iatrogenic hypoglycaemia. They produce their effects in various ways, mostly by interfering with hepatic glucose production, counter-regulatory hormone action, or by stimulating insulin secretion.

Alcohol hypoglycaemia

Alcohol-induced hypoglycaemia is the most common cause of non-iatrogenic hypoglycaemia. The patient is usually stuporose or comatose. Sometimes they are aggressively unco-operative and their symptoms attributed to alcoholic intoxication rather than to hypoglycaemia. Characteristically, hypoglycaemia develops within 6 to 36 h of the ingestion of moderate to large amount of alcohol (>30 g) by fasting or malnourished subjects who may be, but often are not, habituated to alcohol. Hypothermia is more common than with other causes of hypoglycaemia and may provide the first clue to diagnosis. Children, in whom there is a 25 per cent mortality, are particularly susceptible to this type of hypoglycaemia.

Blood glucose is less than 2.5 mmol/l and alcohol almost always present, generally at a concentration below 20 mmol/l (100 mg/100 ml). Plasma and urinary ketones are high but often overlooked because traditional tests for ketones detect only acetone and acetoacetate rather than b-hydroxybutyrate—the redox pair member normally present in alcoholic ketoacidosis.

Once considered, the diagnosis is seldom in doubt and is due to inhibition by alcohol of hepatic gluconeogenesis from lactate and glycerol. It can be confirmed by demonstrating hypoglycaemia, raised plasma b-hydroxybutyrate, and low plasma insulin, C-peptide, and proinsulin levels together, in most cases, with measurable amounts of alcohol.

Consciousness can be restored with intravenous glucose but not glucagon—which is ineffective. Long-term treatment is avoidance of the predisposing factors.

Accidental, factitious, and felonious hypoglycaemia

In these states, although hypoglycaemia is due to exogenous hypoglycaemic agents, this fact is not revealed by the history. The correct diagnosis emerges only from critical examination of laboratory test results and other non-clinical or forensic evidence. Typically the patient is hypoglycaemic and stuporose or comatose when first seen and—unless the possibility of drug-induced hypoglycaemia is suspected from the outset, and appropriate samples of blood and urine collected for insulin, C-peptide, proinsulin, and sulphonylurea assay—the correct diagnosis may never be made.

Sulphonylureas

Dispensing or prescription errors are an important cause of hypoglycaemia—a sulphonylurea being substituted for another drug with a similar name (e.g. diabinese™ for diamox™). In hospital, victims of accidental hypoglycaemia often have received medication intended for someone else. Because patients are often elderly and slip slowly into hypoglycaemic coma without warning, the diagnosis may be delayed or missed completely.

Deliberate sulphonylurea overdose with suicidal or felonious intent is uncommon. It may be difficult to distinguish from accidental overdose in a diabetic patient unless the plasma sulphonylurea level is abnormally high or a suicide note is found. Treatment with diazoxide and intravenous glucose may be required for many days to prevent recurrent hypoglycaemia.

Insulin

Factitious insulin-induced hypoglycaemia is as common in previously healthy subjects as in insulin-dependent diabetics and is due to deliberate, but concealed, injection of insulin. The history suggest insulinoma but is eliminated by the laboratory results which reveal high plasma insulin and low C-peptide (and proinsulin) concentrations during hypoglycaemia. In long-standing factitious hypoglycaemia, and in insulin-treated diabetics, insulin antibodies may be present in the plasma. Although once considered a strong pointer to factitious hypoglycaemia, the presence of insulin antibodies should nowadays suggest autoimmune insulin syndrome.

Suicidal overdosing with insulin is not confined to diabetic patients and is usually unsuccessful. Most patients are found within 12 h of injecting themselves and are restored to consciousness by appropriate treatment. Plasma C-peptide is unrecordably low and (free) insulin concentrations generally greater than 2000 pmol/l. In factitious hypoglycaemia, a form of Munchausen syndrome, plasma insulin concentrations are generally lower than this.

Murder or attempted murder with insulin is exceedingly rare and virtually confined to infants, critically ill patients, and the elderly. The victims are often dead when first seen; if suspected, the diagnosis can be made retrospectively by demonstrating inordinately high concentrations of insulin in blood drawn from a peripheral blood vessel or in tissue removed from the putative injection site. Blood, cerebrospinal fluid, and vitreous glucose measurements are uninterpretable after death.

Organ failure

Hypoglycaemia can occur, sometimes as a dominant feature, in almost any serious and life threatening illness. Most notably are: congestive cardiac failure; acute liver failure; chronic renal failure; bacterial, viral, and parasitic infections (especially malarial); and terminal malnutrition. The cause of the hypoglycaemia is seldom in doubt but its recognition and restoration of normoglycaemia sometimes dramatically alters the course of the illness.

Endocrine hypoglycaemia

Hypoglycaemia is a rare but important presenting sign of several endocrine disorders of which Addison's disease, pan-hypopituitarism, and isolated ACTH deficiency are the most common. The typical clinical features of endocrinopathy are inconspicuous and the diagnosis may be missed unless specifically sought through appropriate laboratory testing. Paradoxically, reactive hypoglycaemia is a rare manifestation of pheochromocytoma with which its symptomatology may be confused. Primary glucagon deficiency has only once been documented as a cause of hypoglycaemia.

Inborn errors of metabolism

Many inborn errors of carbohydrate metabolism—which usually present as hypoglycaemia in childhood—can first manifest themselves in adult life. Mild variants may be responsible for obscure cases of hypoglycaemia which occur only under very stressful conditions, such as prolonged fasting or exceptionally violent exercise, and for which no endocrine or organic cause can be found.

Further reading

Auer RN (1998). Insulin, blood glucose levels, and ischemic brain damage. *Neurology* **51**, S39–43.

Boles RG *et al*. (1999). Glucose transporter type 1 deficiency: a study of two cases with video-EEG. *European Journal of Pediatrics* **158**, 978–83.

Bolli GB, Fanelli CG (1999). Physiology of glucose counterregulation to hypoglycemia.. *Endocrinology and Metabolism Clinics of North America* **28,** 467–93.

Clark PM (1999). Assays for insulin, proinsulin(s) and C-peptide. *Annals of Clinical Biochemistry* **36**, 541–64.

Cryer PE (1999). Symptoms of hypoglycemia, thresholds for their occurrence, and hypoglycemia unawareness. *Endocrinology and Metabolism Clinics of North America* **28,** 495–500.

Grant CS (1999). Surgical aspects of hyperinsulinemic hypoglycemia. *Endocrinology and Metabolism Clinics of North America* **28,** 533–54.

Koch CA, Rother KI, Roth J (1999). Tumor hypoglycemia linked to IGF-II. In: Rosenfeld R, Roberts C Jr, eds. *Contemporary endocrinology: the IGF system*, pp. 675–98. Humana Press, Totowa, New Jersey.

Lteif AN, Schwenk WF (1999). Hypoglycemia in infants and children. *Endocrinology and Metabolism Clinics of North America* **28,** 619–46.

Marks V (1999). Murder by insulin. *Medico-Legal Journal* **67**, 147–63.

Marks V, Teale JD (1998). Tumours producing hypoglycaemia. *Endocrine-Related Cancer* **5**, 111–29.

Marks V, Teale JD (1999). Drug-induced hypoglycemia. *Endocrinology and Metabolism Clinics of North America* **28,** 555–77.

Marks V, Teale JD (1999). Hypoglycemia: factitious and felonious. *Endocrinology and Metabolism Clinics of North America* **28,** 579–601.

Redmon JB, Nuttall FQ (1999). Autoimmune hypoglycemia. *Endocrinology and Metabolism Clinics of North America,* **28,** 603–18.

Seckle MJ *et al*. (1999). Hypoglycemia due to an insulin-secreting small-cell carcinoma of the cervix. *New England Journal of Medicine* **341**, 733–6.

Service FJ (1999). Diagnostic approach to adults with hypoglycemic disorders. *Endocrinology and Metabolism Clinics of North America* **28,** 519–32.

Teale JD, Marks V (1998). Glucocorticoid therapy suppresses abnormal secretion of big IGF-II by non-islet cell tumours inducing hypoglycaemia (NICTH). *Clinical Endocrinology* **48**, 491–8.

Thomson GA *et al*. (1998). A comparative study of glucose meter accuracy during biochemical hypoglycaemia in humans. *Practical Diabetes International* **15**, 135–8.

12.12 Hormonal manifestations of non-endocrine disease

H. E. Turner and J. A. H. Wass

Introduction

Several endocrine syndromes may develop in association with diseases that are not primarily disorders of an endocrine gland. In most the cause is a tumour, usually but not invariably malignant, that develops in tissue not normally looked upon as the site of the particular hormone synthesized. Other 'non-endocrine conditions' may also be associated with either hormonal excess or deficiency, for example sarcoidosis and AIDS. Certain drugs may also modify hormonal biochemistry and cause hormonal imbalance syndromes.

Syndromes of ectopic hormone secretion

In 1941, Albright suggested that the hypercalcaemia sometimes associated with malignant disease without osteolytic metastases might be due to the secretion by the tumour of a parathyroid hormone-like peptide; we now know that this is true (parathyroid hormone related protein, PTHrP). Later it was shown that hypersecretion of adrenocorticotrophin (ACTH), not from the pituitary but from an ectopic site, was the cause in about one-fifth of patients with Cushing's syndrome.

Although 'ectopic' hormone secretion has classically been recognized in the context of neoplasia, and defined as the release of a hormone from a site different from the gland that normally produces the hormone, it is increasingly being recognized that many hormones are synthesized by 'non-endocrine' tissue. Thus the syndromes of neoplastic ectopic hormone secretion are actually due to the pathological over-secretion and/or inappropriate production of hormones. Increasing recognition of the importance of paracrine secretion of hormones such as insulin-like growth factors (IGF-1), their modulation by growth factors and binding proteins, for example IGFBPs1, 2, and 3, and their role in progression of neoplasia adds greatly to these complexities.

Many different hormones are ectopically secreted by neoplasms arising in diverse organs, notably the bronchus, breast, pancreas, kidney, and ovary as well as in mesenchymal tissue. Although a particular endocrinopathy may be associated with a specific type of tumour in a particular organ, the relationship is not invariable. An example is the lung, where squamous cell carcinomas are often associated with hypercalcaemia due to parathyroid hormone related peptide, while small cell lung cancer and bronchial carcinoid tumours are both associated with ectopic ACTH secretion, but with very different clinical manifestations. Indeed, many neoplasms elaborate more than one hormonal substance at the same or at different times and thus may produce a mixed endocrine picture (for example pancreatic endocrine tumours producing ACTH and insulin). Furthermore, the amount of ectopic hormone(s) produced may fluctuate from time to time (for example cyclical Cushing's syndrome in ectopic ACTH secretion). In some instances, the changes induced by the ectopic hormone may mimic very closely, and be clinically indistinguishable, from those found in the true endocrinopathy. In others, the picture is less characteristic and dominated more by abnormalities of biochemistry or hormone levels. Thus, in many cases of ectopic ACTH production by small cell lung cancer, the downhill course of the illness may be too rapid for the classical features of florid Cushing's syndrome to develop, and hypokalaemic alkalosis with diabetes predominates.

Criteria for diagnosis

The diagnosis of ectopic hormone production depends on a number of criteria, although it is seldom practicable or possible to confirm them all:

1. There is an association of the tumour with an endocrine syndrome.
2. Even though the endocrine syndrome may not be clinically florid, there is an elevated or inappropriately raised plasma level of the putative hormone.
3. Removal or suppression of the tumour induces a regression of the endocrinopathy and a fall in the hormone level.
4. The clinical picture and hormone levels are uninfluenced by removal of the gland that normally secretes the hormone.
5. The hormone level is higher in venous blood draining the tumour than in the arterial blood supplying it.
6. Extraction or immunohistochemical staining shows a higher concentration of the hormone in the tumour than in adjacent, non-involved tissue.
7. Demonstration can be made of tumour cell synthesis of identifiable hormones *in vitro* or of mRNA coding for the hormone.

Chemical structure

Most syndromes of ectopic hormone secretion are due to peptide hormones. It is rare for tumours to secrete steroid hormones because of the complexity of the enzyme cascade required for steroid biosynthesis. Tumours may, however, be associated with altered steroid metabolism—for example increased aromatase activity in hepatocellular carcinoma leads to feminization and gynaecomastia due to androgen conversion to oestrogens.

The precise amino acid sequences of hormones of ectopic origin are being increasingly defined. In general, they appear to resemble closely those of their normally occurring counterparts (except parathyroid hormone (PTH) and PTHrP). There is a tendency for a greater proportion of higher molecular weight precursors, prohormones, or subunits and fragments to be associated with an ectopic origin than with 'true' endocrinopathies but it is not always clear whether this is due to differences in biosynthesis or in intracellular or extracellular processing. Minor differences in molecular structure are sometimes reflected in disparities between

bioassay and immunoassay.

Prevalence

Clinically evident syndromes are less common than biochemical or hormonal abnormalities. The prevalence of ectopic production of ACTH, corticotrophin-releasing hormone (CRH), parathyroid hormone related protein (PTHrP), calcitonin, chorionic gonadotrophin (hCG), prolactin, or growth hormone, without clinical manifestations, is high when extensive biochemical and hormonal assays are applied to patients with cancer. These assays bring closer the prospect of finding a diagnostic 'marker' for tumours in general and, in particular, as is already the case with the monitoring of hCG or its subunits, to determine the response of tumours to treatment.

Hypercalcaemia in the absence of detectable bony metastases is the most common abnormality. It occurs in about 15 per cent of patients with squamous cell carcinoma, usually of the bronchus, carcinoma of the kidney, ovary, or breast. Next most common in neoplastic diseases is the syndrome of inappropriate antidiuresis (SIAD), usually associated with a small cell lung cancer and reported in 40 per cent of such cases. Cushing's syndrome due to ectopic ACTH or CRH secretion occurs in about 5 per cent of patients with small cell lung cancer, and in association with other neoplasms. Biochemical accompaniments of Cushing's syndrome in the absence of the clinical features are much more common, occurring in 50 per cent of patients with small cell lung cancer.

Pathogenesis

As techniques for molecular analyses have evolved, it has become clear that every somatic cell is capable of synthesizing every polypeptide hormone. However, only under pathological circumstances is that capability ever likely to be expressed. A variety of hypotheses for ectopic hormone synthesis and secretion have been made. None explains all of the observed facts. Fundamentally, all cells inherit an identical complement of DNA. They are therefore totipotential and have all the coded information required for the synthesis of all proteins and peptides, including protein hormones. The normal inability of non-endocrine tissue to synthesize hormones is ascribed to 'repressors' that mask specific segments of the DNA molecule. It seems possible that when a cell becomes malignant this normal repression becomes ineffective, allowing the unmasked DNA to synthesize proteins or peptides 'foreign' to the cell concerned. Such a 'de-repression' hypothesis does not explain why certain tumours are more prone to secrete certain ectopic hormones. Neuroendocrine cells, characterized by the presence of peptide hormone granules, are likely to be the origin of some tumours associated with hormone secretion, such as small cell lung cancer and bronchial carcinoids. Another hypothesis suggests that there are a small number of special proliferative cells in normal mature tissues that have fetal characteristics with the ability to produce peptide hormones—a process of 'dysdifferentiation' rather than 'de-repression'. There is currently no unifying mechanism with supportive experimental evidence to explain ectopic hormone production. Further information on the control of gene expression and hormone production, the role of oncogenes, and paracrine growth factors may provide further insight.

Treatment

Treatment of the clinical or biochemical abnormalities associated with endocrinopathies of non-endocrine origin is best directed at the primary disorder. In neoplastic disease, this may involve surgical excision, radiotherapy, or chemotherapy. Sometimes the tumour secreting the ectopic hormone is extremely difficult to locate even with the use of sophisticated imaging techniques such as magnetic resonance imaging (MRI), radiolabelled isotope scanning (e.g. ^{111}Indium-penetreotide imaging), or using selective venous catheterization.

More specific therapy may be necessary to contain the metabolic abnormality until such time as the fundamental disorder can be controlled. For example immediate measures may be required to reduce hypercalcaemia with fluids and bisphosphonates, or steps taken (administration of metyrapone) to diminish corticosteroid secretion from adrenal glands stimulated by ectopic ACTH secretion.

Ectopic secretion of calciotropic hormones

Malignancy is the most common cause of hypercalcaemia in the hospital inpatients and may be due to direct tumour spread to the bones or related to secreted calcium-releasing factors. Often several different mechanisms are involved in the same patient.

After its discovery in 1987, it was shown that PTHrP is responsible for hypercalcaemia in up to 70 per cent of patients with this tumour-associated phenomenon. PTHrP shares amino acid homology with PTH between positions 2 and 13 of the 84 residues of PTH and acts via the PTH receptor. The PTHrP gene is located on the short arm of chromosome 12; that of PTH is on chromosome 11. The PTHrP gene may be activated by transactivation, hypomethylation (renal carcinomas), or the effect of growth factors and cytokines, including IGF-1 and epidermal growth factor, while glucocorticoids and vitamin D_3 suppress PTHrP levels. Unlike PTH-mediated hypercalcaemia, dihydroxycholecalciferol is suppressed in PTHrP-mediated hypercalcaemia. PTHrP is made by squamous carcinomas as well as renal, bladder, ovary, skin, pancreas, and breast carcinomas, and lymphomas.

Other factors can be involved in hypercalcaemia unassociated with osseous metastases. 1,25-Dihydroxy vitamin D_3 is not uncommonly made by lymphoproliferative tumours, which are either high grade or widely disseminated. Transforming growth factor-a (TGFa) which stimulates osteoclastic bone resorption, is also made by squamous carcinoma, and renal and breast carcinomas. Some tumours cosecrete both TGFa and PTHrP. Interleukin-1, which is a very powerful stimulator of osteoclastic bone resorption, is also made by squamous carcinomas as well as some haematological malignancies. Tumour necrosis factor (TNF) and lymphotoxin also stimulate osteoclastic bone resorption. These related cytokines cause hypercalcaemia *in vivo*; lymphotoxin is produced by cultured myeloma cells *in vitro* and accounts for the hypercalcaemia seen in this condition. Prostaglandins of the E series may also cause hypercalcaemia.

It is important to remember that primary hyperparathyroidism itself is common, particularly in the elderly; two diseases may coexist. For this reason, primary hyperparathyroidism should always be considered when hypercalcaemia occurs, even if it is in a patient within the setting of malignant disease. It is now possible to differentiate between these two conditions by using the PTH two-site radioimmunoassay.

Paraneoplastic hypercalcaemia may be either asymptomatic or dominate the clinical picture and be life-threatening as a consequence of dehydration and renal failure. The features of hypercalcaemia and its general management are discussed elsewhere (see Chapter 12.6).

Oncogenic osteomalacia is a rare syndrome usually associated with benign mesenchymal tumours (e.g. haemangiopericytomas) where phosphaturia, hypophosphataemia, and normocalcaemia are associated with suppressed 1,25 dihydroxy vitamin D. It may be due to an unknown phosphaturic factor which also inhibits 1-a-hydroxylase.

Syndrome of inappropriate antidiuresis (SIAD)

This syndrome is usually, but not invariably, associated with high levels of circulating arginine vasopressin. Other, as yet unidentified, antidiuretic substances are sometimes involved. There is hyponatraemia and impaired water excretion in the absence of hypovolaemia, hypotension, or deficiency of cardiac, renal, thyroid, or adrenal function. Associated with hyponatraemia, there is a reduction in plasma osmolality and inappropriately normal/low urine concentration.

Bronchogenic carcinoma is the commonest malignancy associated with SIADH. Although SIADH in association with malignancy may be due to the tumour itself, it may also result from treatment (e.g. chemotherapy such as cyclophosphamide), an intercurrent illness such as pneumonia, or a complication such as hydrocephalus or cerebrovascular accident (Table 1).

Ectopic ACTH secretion

Pro-opiomelanocortin (POMC) is a 31 kDa precursor for both ACTH and b-lipotrophin as well as for other polypeptides derived from it, including g-lipotrophin and b-endorphin. A variety of non-pituitary tumours are capable of secreting POMC-derived peptides, accounting for about 20 per cent of patients with Cushing's syndrome. Approximately 50 per cent of these ectopic ACTH-producing tumours are in the lung and the rest are present in a variety of other tissues (Table 2). Some tumours, particularly pancreatic islet cell tumours which are seldom (< 5 per cent) associated with Cushing's syndrome, can, in addition to ACTH, also secrete a number of other hormones, including insulin, gastrin, and glucagon (see Chapter 12.10). This accounts for the usefulness, when screening for ectopic ACTH, of measuring other hormones (e.g. calcitonin, hCG) which may be cosecreted, the presence of which raises the suspicion of an ectopic hormone-secreting tumour. Very

rarely, corticotrophin releasing hormone (CRH) is secreted ectopically in association with ACTH.

While small cell lung cancer is the most common source, carcinoids anywhere, but particularly bronchial carcinoids, phaeochromocytoma, and medullary carcinoma of the thyroid may also secrete ACTH ectopically (Table 2).

The exact mechanism of synthesis of ectopic POMC-derived peptides is still debated. POMC mRNA can be found in the majority of tumours, but ACTH secretion is much less common, probably due to the lack of the signal sequence required for translocation. Changes in promoter usage and also in POMC processing may lead to ectopic secretion of ACTH. In addition, many tumours associated with ectopic ACTH secretion are of neuroendocrine morphology and may arise from progenitor cells associated with ACTH secretion.

Presentation

The clinical picture is variable. In patients with small cell lung cancer who have a rapidly progressive tumour, the physical features of Cushing's syndrome may not have time to develop. The major features are weight loss, proximal muscular weakness, polyuria, thirst, oedema, carbohydrate intolerance with glycosuria, and sometimes pigmentation due to ACTH. Hypokalaemic alkalosis is a characteristic finding; the plasma potassium is less than 3.2 mmol/l and the bicarbonate greater than 30 mmol/l, the urine potassium loss being the direct cause of most of the symptoms. This hypokalaemia is in part due to the very high cortisol levels, which have a mineralocorticoid action, and corticosterone and 11-deoxycorticosterone which may also be produced in excess. The 11b-hydroxysteroid dehydrogenase enzyme may also function abnormally, causing decreased inactivation of cortisol and corticosterone. The serum cortisol level is usually greatly elevated (>1000 nmol/l) and the plasma ACTH level is also raised (>200 µg/l). These high levels do not usually occur in pituitary-dependent Cushing's disease.

When the ectopic sources are other than a small cell lung cancer, the clinical manifestations may be quite indistinguishable from Cushing's disease and cushingoid features may antedate by months or years any evidence of a tumour causing ectopic ACTH secretion. The degree of elevation of ACTH is less marked than with small cell lung cancer and is proportional to tumour size. Some carcinoid tumours may be small and difficult to locate. The real problem is to differentiate ectopic ACTH secretion from pituitary-dependent disease (Table 3). The presence of a hypokalaemic alkalosis(< 3.2 mmol/l) is very useful test in the differential diagnosis. Lack of suppression on high-dose dexamethasone testing is found in 90 per cent patients with ectopic disease, but also up to 20 per cent with pituitary disease. However, the CRH test is very useful in differentiation as patients with ectopic ACTH secretion show an absent rise in cortisol whereas pituitary dependent disease is associated with an exaggerated response in 95 per cent patients. Because most of the tumours secreting POMC are in either the chest or abdomen, MRI or computed tomography (CT) will often reveal the source of ectopic hormone secretion. In patients in whom the lesion is not readily visible by imaging techniques, selective venous catheterization and sampling may help determine a source of ACTH by comparing levels at various sites within the venous system. Such sampling should include inferior petrosal sinuses in case of pituitary-dependent disease.

Treatment

Removal of the primary growth or its control with radiotherapy or chemotherapy will relieve the endocrine manifestations. A relapse may occur if metastases develop because these, too, usually secrete ACTH. When it proves impossible to control a primary tumour, adrenocortical hypersecretion may be reduced by 'medical adrenalectomy', giving the 11b-hydroxylase inhibitor of the conversion of 11-deoxycortisol to cortisol, metyrapone (500–4000 mg/day). Aminoglutethimide (1000–1500 mg/day) may also be used but frequently causes a skin rash. Ketoconazole (400–800 mg/day), which can cause fatal liver damage, and the adrenolytic drug mitotane are also useful. RU-486, a glucocorticoid antagonist at the receptor level, has been used as palliative therapy for some patients (10–30 mg/kg per day). Lastly, the long-acting somatostatin analogue, octreotide (0.3 mg/day, subcutaneously), has also been used in the treatment of ectopic ACTH syndrome.

Bilateral adrenalectomy is an alternative approach, but frequently it is not practical for patients with rapidly progressive metastatic disease. It may be possible to embolize the arterial supply of the adrenal gland if patients are not suitable surgical candidates for adrenalectomy. Medical treatment needs to be monitored carefully so that adrenal insufficiency is avoided.

Ectopic secretion of insulin-like growth factors

The insulin-like growth factors, IGF-I and II, share some sequence homology and actions of insulin. IGF-II is important in fetal growth, whereas IGF-I, synthesized in the liver, mediates most of the actions of growth hormone. IGFs circulate bound to one of six binding proteins (IGFBPs). Of these, the most important is IGFBP3, which itself is growth hormone-dependent and binds 75 per cent of IGF-I and IGF-II.

IGF-II secretion from tumours may be associated with hypoglycaemia. Usually the tumour is large and of mesenchymal origin, arising in the abdomen or thorax. Symptoms are those of neuroglycopenia—sweating, tachycardia, disorientation, drowsiness, fits, and coma. Histology shows a mesothelioma, a fibrosarcoma, or other sarcoma such as a leiomyosarcoma. Other neoplasms associated with hypoglycaemia are haemangiopericytoma, hepatoma, adrenal carcinoma, lung carcinoma, Wilms' tumour, and colonic carcinoma.

IGF-II secretion leads to suppression of growth hormone and insulin, and reduced production of IGFBP3, IGF-I, and acid labile subunit (ALS), leading to reduced formation of the IGF–IGFBP3–ALS complex which protects the IGFs from degradation. IGF-II circulates as a smaller complex which has enhanced tissue and receptor bioavailability, allowing access to the insulin receptor. There is also an increase in the large molecular weight molecules and the increased amounts of 'big' IGF-II not detected on radioimmunoassay. Growth hormone deficiency, decreased gluconeogenesis, and increased glucose metabolism by the tumour, which is usually large, may also contribute to hypoglycaemia. Treatment of these tumours is difficult. The hypoglycaemia is often not responsive to diazoxide, glucagon, octreotide, or corticosteroids. However, administration of growth hormone may be effective—increasing IGFBP3 and IGF-I and antagonizing the effect of excess IGF-II. The underlying tumour may be resistant to radiotherapy; surgery, although effective if possible, is not always feasible.

IGF-I and IGF-11 may also play an important role in tumour progression. Studies of breast cancer cells have suggested that IGF-I may have local mitogenic effects, and a role for IGF-II has recently been proposed in hepatocellular, colorectal, and adrenocortical tumours.

Ectopic human chorionic gonadotrophin secretion

Human chorionic gonadotrophin is a glycoprotein consisting of an a- and a b-subunit. The a-subunit is species specific and is the same for all glycoprotein hormones (luteinizing hormone (LH), follicle stimulating hormone (FSH), and thyroid stimulating hormone (TSH)). The b-subunit determines receptor interaction and specific hormone activity. The b-subunit of hCG is very similar to that of LH and this can cause problems with cross-reaction in assays. Clinically silent, ectopic secretion of hCG, with or without its free a- and b-subunits, occurs in many patients (Table 4).

In the first decade of life, ectopic hCG production may cause isosexual precocious puberty in boys with hepatoblastoma. hCG, through its LH-like action, causes Leydig cell stimulation in the testes. In turn, testosterone levels reach those of a normal adult, and secondary sexual characteristics develop together with premature skeletal maturity. The testes remain small because there is no seminiferous tubule growth as this is dependent on FSH. Precocious puberty is rare in girls.

Intracranial teratoma, choriocarcinoma, and pinealoma are associated with ectopic hCG secretion. In men this may be associated with gynaecomastia. In some this is due to cosecretion of oestrogen which may, in women, be associated with dysfunctional uterine bleeding. Other tumours associated with hCG secretion are testicular tumours, ovarian adenocarcinoma, and stomach, pancreatic, and liver tumours.

hCG is a useful tumour marker in gestational trophoblastic disease (choriocarcinoma) and in some men with testicular tumours, and provides an early warning of recurrent disease. However, it is important to measure other tumour markers, for example a-fetoprotein, which may also be secreted by non-seminomatous germ cell tumours. Discordance of marker levels and tumour progress may be seen. In central nervous system disease, cerebrospinal fluid/plasma ratios may help in the correct localization of tumours, as hCG does not cross the blood–brain barrier and levels in cerebrospinal fluid remain undetectable in pregnancy. Thus, cerebrospinal fluid concentrations higher than plasma suggest primary central nervous system disease.

In some patients, most commonly with choriocarcinoma and massive elevation of hCG, the latter, through its weak TSH activity, due to its biochemical similarity to TSH, may cause goitre and hyperthyroidism. This most frequently occurs in women, is not associated with eye signs, and is usually associated with modest biochemical abnormalities. Treatment of the tumour results in a resumption of a euthyroid state but, if this is not possible, carbimazole or propylthiouracil may be

required.

Ectopic human placental lactogen

Human placental lactogen (hPL), also called human chorionic somatomammotropin (hCS), is a trophoblastic hormone which may be secreted ectopically in association with lung tumours, testicular tumours, and trophoblastic disease. It is usually associated with gynaecomastia in men, and these tumours may also be associated with increased levels of oestradiol and hCG.

Ectopic growth hormone releasing hormone and growth hormone secretion

Most patients with acromegaly (98 per cent) have benign growth-hormone-producing pituitary adenomas. Less than 2 per cent of patients with acromegaly have ectopic growth hormone releasing hormone (GHRH) production. A patient with a carcinoid tumour of the pancreas producing GHRH enabled the final elucidation of the structure of this important hypothalamic peptide which stimulates anterior pituitary growth hormone secretion. Besides the pancreas, lung carcinoid tumours, small cell lung cancer, and phaeochromocytoma may produce GHRH ectopically and cause acromegaly by stimulation of the pituitary somatotrophs. Histologically, the two can be differentiated by the presence of somatotroph hyperplasia in ectopic GHRH secretion. These tumours are usually clinically apparent and GHRH levels in the circulation are elevated. GHRH can also be secreted by hypothalamic hamartomas, which also result in anterior pituitary somatotroph hyperplasia.

Ectopic growth hormone secretion has been reported in patients with bronchial, pancreatic, and gastrointestinal carcinoma, and cells cultured from an undifferentiated lung cancer have been shown to synthesize growth hormone *in vitro*. Breast carcinoma and ovarian tumours may also occasionally secrete growth hormone but no clinical syndrome has been clearly identified as caused by ectopic growth hormone.

Ectopic prolactin secretion

Prolactin may be secreted by bronchial carcinoma and renal cell carcinoma; the usual endocrine manifestation is galactorrhoea and there may be marked hyperprolactinaemia. These abnormalities are reversed if the tumour is controlled or removed. Difficulties in differential diagnosis may arise unless the underlying abnormality is clinically obvious or suspected, because in most instances the hyperprolactinaemia will be attributed to a prolactin-secreting adenoma. Suspicion of an ectopic source may only arise when the prolactin level is not lowered by bromocriptine treatment. An autocrine role for prolactin in breast and prostate cancer has recently been postulated.

Ectopic calcitonin secretion

Increased serum calcitonin levels are encountered in a variety of cancers apart from medullary carcinoma of the thyroid. The most common of these are small cell lung cancer, leukaemia, and neoplasms of the breast and pancreas. It is often produced as part of a multihormonal profile in conjunction with, for example, gastrin, ACTH, and somatostatin. Ectopic calcitonin may differ from the normal hormone in having more high molecular weight components; it does not cause any apparent symptoms and does not produce hypocalcaemia.

Ectopic renin secretion

Although hypertension associated with hyper-reninism and increased aldosterone production is usually due to a renal lesion, ectopic secretion of renin has also been described in association with cancer of the lung, pancreas, and ovary. The clinical picture is usually dominated by the underlying neoplasm but the patient has hypertension and the cause of this may be suspected from the associated hypokalaemia and its accompanying muscle weakness. Effective treatment of the primary lesion will reduce the increased renin and aldosterone levels and hence the raised blood pressure. When the underlying cause cannot be eradicated, the use of an angiotensin enzyme inhibitor will control the hypertension.

Ectopic aldosterone secretion

Hypertension and hypokalaemia related to ectopic secretion of aldosterone from a non-adrenal neoplasm have been described in patients with ovarian tumours. Its pathogenesis is different from the others described in this section. The aberrant production of a steroid, aldosterone, rather than a peptide, is presumably due to biochemical change in the ovarian steroidogenic cells. Attention is likely to be focused on a suspected lesion of the adrenal zona glomerulosa because the hyperaldosteronism is associated with low plasma renin activity. The ovarian lesion may initially be clinically silent and only revealed by pelvic imaging.

Endocrine manifestations of non-endocrine diseases

Systemic disease of non-endocrine glands may influence endocrine function due to a specific effect of the disease itself, due to a general response to either acute or chromic illness, or due to drug therapy used to treat the illness itself (Table 5). Often hormonal perturbations may be a complex mixture of all of these mechanisms, as may be seen for example in AIDS or critically ill patients on intensive therapy units. This section includes examples of systemic disease causing endocrine disorders.

A commonly encountered hormonal disturbance encountered in many hospital inpatients is the 'sick euthyroid syndrome'. Peripheral conversion of T_4 to T_3 is reduced, and typical thyroid function tests in this syndrome are a normal or reduced TSH in association with reduced T_3 and T_4 (and increased rT_3 if measured). Severe illness may also interfere with hypothalamopituitary function and lead to hypogonadotrophic hypogonadism. Possible mechanisms include increased cortisol levels, stress, cytokines, or opioids given as analgesia.

Disorders influencing hypothalamopituitary function

Anorexia nervosa is associated with complex changes in hypothalamopituitary function, with reduction in GnRH and gonadotrophin secretion leading to hypogonadotrophic hypogonadism but increased growth hormone secretion is associated with increased peripheral resistance to growth hormone.

Iron overload due to haematological conditions such as b thalassaemia major and to haemochromatosis may cause iron deposition in the anterior pituitary gland, and in particular in the gonadotrophs. This leads to hypogonadotrophic hypogonadism, which may be ameliorated to a degree by venesection and iron chelation therapy. Haemochromatosis may also lead to other hormonal changes due to pancreatic involvement causing diabetes mellitus, and cirrhosis associated with secondary hyperaldosteronism and hypogonadism.

Thyroid

Hyperemesis gravidarum in the first trimester of pregnancy may be associated with clinical and biochemical features of thyrotoxicosis, as the molecules hCG and TSH share very similar b subunits, allowing cross reactivity when high levels of hCG occur.

Opportunistic infections of the thyroid gland may occur, in conditions associated with immunosuppression such as AIDS. Infection with cytomegalovirus, *Cryptococcus*, and *Pneumocystis carinii* have been described. In addition, some patients with HIV infection have increased T_4 and T_3 due to increased thyroid binding globulin. As the disease progresses, T_4 and T_3 levels fall as patients develop biochemical features of sick euthyroidism.

Adrenal

Opportunistic infections (CMV, atypical mycobacteria, *Cryptococci*, *Toxoplasma*, and *Pneumocystis*), lymphoma and Kaposi's sarcoma may involve the adrenal glands in HIV and AIDS. The adrenal gland is the most commonly involved endocrine gland at autopsy. However, frank adrenal insufficiency is rare because this requires destruction of over 90 per cent of the adrenal cortex.

Gonads

Chemotherapy and irradiation may be associated with gonadal failure due to hypothalamopituitary gonadotrophin deficiency following, for example, cranial irradiation or due to testicular/ovarian damage following cytotoxic drug therapy such as cyclophosphamide, cisplatin, and busulfan.

Coeliac disease is associated with reversible male infertility due to androgen resistance, and improves on a gluten free diet. Alteration of gonadal steroid metabolism may occur in, for example, chromic liver disease, particularly if alcohol related. Elevated sex hormone binding globulin (SHBG) and oestradiol levels are associated with a reduction in bioavailable testosterone leading to testicular atrophy, gynaecomastia, and erectile impotence.

Gynaecomastia

Palpable breast glandular tissue is prevalent in population studies of men and boys. Subareolar glandular tissue more than 2 cm diameter is found in 35 to 60 per cent of men. Gynaecomastia may occur as a result of different conditions (Table 6) as well as drug therapy, and results from an alterations in the ratio of oestrogen to androgen. Gynaecomastia has been found in association with testicular and adrenal neoplasms, Klinefelter's syndrome, thyrotoxicosis, cirrhosis, primary hypogonadism, malnutrition, and ageing (Table 6). An increase in free oestrogen, a decrease in free endogenous androgens, androgen-receptor defects, and partially enhanced secretions of breast tissue may underlie these changes. Increased aromatization of oestrogen precursors occurs in patients with obesity, liver disease, and hyperthyroidism, and as a result of ageing.

Calcium

Hypercalcaemia in sarcoidosis is due to increased circulating 1,25 dihydroxy vitamin D. This is produced by alveolar macrophages in a dose-dependent fashion stimulated by g interferon, which is one factor responsible for the maintenance of the inflammatory process in sarcoidosis. Other granulomatous disorders (tuberculosis, histoplasmosis, coccidiomycosis, ruptured silicone breast implants) may rarely be associated with hypercalcaemia due to the same mechanism. Treatment with glucocorticoids or hydroxychloroquine are effective in lowering 1,25 dihydroxycholecalciferol and calcium.

HTLV-1 infection may be associated with hypercalcaemia, due to transactivation of the PTHrP gene on chromosome 12.

Drug-induced endocrine manifestations

Several pharmaceutical drugs may induce manifestations of endocrine disease. More commonly they may influence the results of hormonal assays and lead to mistaken diagnosis. It may not be a major problem when it is known that the patient is taking a particular compound and, from its molecular structure, it is appreciated that such a substance could influence the endocrine system or the results of hormonal assays. The problem is greater, however, when the drug in question has no clear relationship to a hormone and the mechanism by which it induces an endocrine manifestation, or interferes with an assay procedure, is not readily apparent.

Thyroid

Abnormalities of thyroid function test measurements

Drugs can interfere with thyroid function tests. Some act by inhibiting the conversion of thyroxine (T_4) to triiodothyronine (T_3), others by increasing thyroid-binding globulin. b-Blockers with membrane stabilizing properties, such as propranolol, inhibit peripheral conversion of T_4 to T_3. Oral cholecystographic agents and amiodarone, a heavily iodinated antiarrhythmic agent, are also potent inhibitors of T_4 to T_3 conversion and produce decreased serum T_3 concentrations and an increase in reverse T_3. Oestrogen increases thyroid-binding globulin, due to an increase in the sialic acid content of thyroxine-binding globulin, which prolongs its half-life in the circulation. Thus women on oestrogens, for example the contraceptive pill, have high total T_4 concentrations but are euthyroid. Such results may also be seen on tamoxifen. Heroin and methadone addicts also have raised levels of thyroxine-binding globulin, as do patients on the lipid-lowering agent, clofibrate.

A decreased serum T_4 does not necessarily indicate the presence of hypothyroidism. Many pharmacological agents lower the total T_4 concentration by interfering with the binding of T_4 to one or more of the thyroid-binding proteins. Therapeutic levels of phenytoin lower the level of serum T_4 and high concentrations are capable of inhibiting the binding of T_4 and T_3 to thyroid-binding globulin. High doses of salicylates have the same effect. Diclofenac, a non-steroidal anti-inflammatory drug structurally similar to thyroxine, also interferes with thyroid hormone binding. Phenylbutazone, anabolic steroids, and glucocorticoids may also be associated with a low total T_4 and normal thyroid function. Measurement of free thyroxine (fT_4) will obviate the problems of misleading results from the measurement of total T_4.

Drug-induced hyperthyroidism

Amiodarone may cause hyperthyroidism due to its high iodine content, or due to a destructive thyroiditis. Biochemically, there may be a marked elevation of total thyroxine, a relatively normal level of T_3, and a suppressed TSH. Often thyrotoxicosis is masked by the b-blocking effect of the drug. Because of the large iodine load, it may be very difficult to treat with antithyroid drugs, and steroids may also be necessary to suppress thyroid hormone levels into the normal range. Even if amiodarone is stopped, its effects continue for many weeks because it is predominantly stored in adipose tissue. Contrast media and iodine-containing cough medicines may similarly induce hyperthyroidism (Jod–Basedow phenomenon).

Drug-induced hypothyroidism

Increased iodide intake may also lead to decreased iodide trapping and a decrease in synthesis of thyroid hormones, hypothyroidism, and goitre. Iodine is contained in a number of 'tonics' and cough medicines. Amiodarone, besides producing thyrotoxicosis, may cause iodine-induced hypothyroidism in patients replete with iodine. Lithium blocks iodine uptake and the release of thyroid hormones. It also interferes with cAMP formation and thus inhibits the effects of TSH stimulation and may lead to goitre, although only 2 per cent of patients on lithium actually develop clinical features of hypothyroidism.

Adrenal cortex

Abnormalities of adrenal hormone measurements

Drugs may interfere with tests of adrenal function. Thus, for example, phenytoin accelerates metabolism of dexamethasone, and patients on phenytoin may not suppress cortisol normally during dexamethasone suppression tests. Furthermore, during the assessment of adrenal reserve, chronic topical application of steroids, as well as inhalation of steroids for asthma, may suppress adrenal function. Oestrogens, by enhancing hepatic production of cortisol-binding globulin, which binds between 90 and 97 per cent of circulating cortisol, increases cortisol-binding globulin two- to threefold. Thus, assessment of glucocorticoid replacement in patients on oestrogens is influenced by this effect and oestrogens should be stopped 6 weeks prior to the test.

Drug-induced Cushing's syndrome

Chronic, excessive intake of alcohol causes alcoholic pseudo-Cushing's syndrome. These patients behave biochemically as if they have Cushing's syndrome with absent dexamethasone suppression. This occurs through a centrally mediated mechanism with hypersecretion of pituitary ACTH and secondary secretion of cortisol by the adrenals.

Drug-induced primary aldosteronism

Primary aldosteronism can be mimicked by the mineralocorticoid effect of glycyrrhizic acid contained in both carbenoxolone and liquorice. Cortisol is normally inactivated by conversion to the inactive metabolite, cortisone, by the enzyme 11b-hydroxysteroid dehydrogenase but these compounds inhibit the enzyme, which is important in the kidney because it protects renal mineralocorticoid receptors from cortisol.

Drug-induced adrenal insufficiency

The antifungal agent, ketoconazole, and the short-acting anaesthetic, etomidate, are imidazole derivatives with significant inhibitory effects on 11b-hydroxylase. While they do not usually produce clinical insufficiency, they may do so in subjects with limited pituitary or adrenal reserve. Rifampicin and phenytoin, which both accelerate the metabolism of cortisol by inducing hepatic mixed-function oxygenase enzymes, can also provoke adrenal insufficiency in similar patients with limited pituitary or adrenal reserve. In such patients, increased doses of replacement therapy are necessary.

Gonads

Several drugs can affect testicular function, leading to hypogonadism and infertility. Mechanisms include the direct inhibition of testosterone synthesis or competitive inhibition of androgen action at receptor level. Spironolactone acts as a partial androgen receptor antagonist. Alcohol reduces testosterone levels acutely and chronically, by both a central and a gonadal effect on testosterone synthesis, secretion, and metabolism. Cimetidine has antiandrogen effects due to direct interaction with the androgen receptor and it may also exert antiandrogen effects at the pituitary and hypothalamus leading to gynaecomastia and impotence in males. Anticonvulsants, for example phenytoin, increase sex hormone-binding globulin and therefore decrease free testosterone levels. They also enhance testosterone to oestradiol conversion. Sulfasalazine causes reversible male infertility associated with oligospermia.

Infertility may occur as a result of cytotoxic therapy, caused in particular by the alkylating agents such as cyclophosphamide. These produce depletion of the germinal epithelium and lead to a raised FSH level, and oligo- or azoospermia, but normal LH and testosterone levels in males, and may lead to premature ovarian failure in women.

In women, hirsutism can be caused by a number of drugs, including danazol, phenytoin, diazoxide, and minoxidil.

Pharmacological doses of glucocorticoids may lead to hypogonadism because of inhibited gonadotrophin release. Drugs such as tricyclics, benzodiazepines, antihypertensives, and antipsychotics may also lead to hypogonadotrophic hypogonadism in both sexes.

Prolactin

Prolactin is controlled predominantly by a hypothalamic inhibitory mechanism through dopamine secretion. A number of drugs can cause hyperprolactinaemia and galactorrhoea usually acting through a dopaminergic mechanism. They may elevate prolactin to a sufficient extent to cause a clinical suspicion of prolactinoma, and in such patients a careful drug history is particularly important. Metoclopramide, pimozide, and sulpiride all act as dopamine antagonists and may considerably elevate prolactin, with all the attendant effects thereof. Fluoxetine may also lead to elevated serum prolactin, although tricyclic antidepressants are not usually associated with hyperprolactinaemia.

Phenothiazines, chlorpromazine, perphenazine, and trifluoperazine also act as dopamine antagonists, as do haloperidol and butyrophenone. Reserpine and methyldopa both decrease catecholamine stores and may cause hyperprolactinaemia. Oestrogens, in high doses, may slightly elevate prolactin but normal contraceptive pills do not. Verapamil, by decreasing dopaminergic tone, may also increase prolactin levels.

Gynaecomastia

Gynaecomastia may occur due to treatment with various drugs (Table 6). Drugs such as spironolactone and ketoconazole, which can displace steroids from sex-hormone binding globulin, displace oestrogens more easily than androgens. Activation of the oestrogen receptors in breast tissue may take place with drugs that have structural homology with oestrogen, such as digoxin; griseofulvin and cannabis may have the same effect. A decrease in androgen occurs in older men and with drugs such as spironolactone and ketoconazole that inhibit the biosynthesis of testosterone. The mechanism for the induction of gynaecomastia by captopril and calcium-channel blockers (nifedipine) is unclear. With cimetidine and omeprazole, this effect may be due to a direct antiandrogen effect or the inhibition of liver cytochrome P450.

Posterior pituitary

The syndrome of inappropriate antidiuresis is characterized by normovolaemic hyponatraemia with persistent secretion of vasopressin, despite a reduced plasma osmolality. A number of drugs can cause this syndrome, including thiazide diuretics, vincristine, vinblastine, cyclophosphamide, chlorpropamide, phenothiazines, carbamazepine, clofibrate, and tricyclic antidepressants (Table 1).

Nephrogenic diabetes insipidus can be induced by lithium in the therapeutic range, and up to 20 per cent of patients receiving long-term therapy may develop this complication. Demethylchlortetracycline produces dose-dependent nephrogenic diabetes insipidus, and both the concentrating defect and the unresponsiveness to vasopressin are reversible on cessation of the drug.

Parathyroid

Lithium therapy can cause an increase in parathyroid gland size, either with hyperplasia or adenoma. This hyperparathyroidism leads to mild hypercalcaemia and sometimes osteoporosis. Thiazide diuretics, by causing haemoconcentration and hypocalciuria, may also result in mild hypercalcaemia but this is usually transient (4–6 weeks); after this time, other causes of hypercalcaemia should be sought.

Vinblastine and colchicine inhibit parathyroid hormone secretion which may result in hypocalcaemia.

Further reading

Bell NH (1991). Endocrine complications of sarcoidosis. *Endocrinology and Metabolism Clinics of North America* **20**, 645–54.

Braunstein GD (1993). Current concepts: gynecomastia. *New England Journal of Medicine* **328,** 490–5.

Chopra IJ (1997). Clinical review 86: Euthyroid sick syndrome: is it a misnomer? *Journal of Clinical Endocrinology and Metabolism* **82**, 329–34.

Daughaday WH and Deuel TF (1991). Tumour secretion of growth factors. *Endocrinology and Metabolism Clinics of North America* **20,** 539–63.

Docter R, Krenning EP, DeJong M, Hennemann G (1993). The sick euthyroid syndrome: changes in thyroid hormone serum parameters and hormone metabolism. *Clinical Endocrinology* **39,** 499–510.

Grinspoon SK, Bilezikian JP (1992). HIV disease and the endocrine system. *New England Journal of Medicine* **327**, 1360–5.

Guise TA, Mundy GR (1998). Cancer and bone. *Endocrine Reviews* **19**, 18–54.

Howlett TA, Drury PL, Perry L, Doniach I, Rees LH, Besser GM (1986). Diagnosis and management of ACTH-dependent Cushing's syndrome: comparison of the features in ectopic and pituitary ACTH production. *Clinical Endocrinology* **24,** 699–713.

Hung W, Blizzard RM, Migeon CJ, Camacho AM, Nyhan WL (1963). Precocious puberty in a boy with hepatoma and circulating gonadotropin. *Journal of Pediatrics* **63,** 895–903.

Kovacs L, Robertson GL (1992). Syndrome of inappropriate antidiuresis. *Endocrinology and Metabolism Clinics of North America* **21,** 859–76.

Melmed S (1991). Extrapituitary acromegaly. *Endocrinology and Metabolism Clinics of North America* **20,** 507–18.

Penny E *et al.* (1984). Circulating growth hormone releasing factor concentrations in normal subjects and patients with acromegaly. *British Medical Journal* **289,** 453–5.

Turner HE, Wass JAH (1997). Gonadal function in men with chronic illness. *Clinical Endocrinology* **47**, 379–403.

Vaitukaitis JL (1991). Ectopic hormonal secretion and reproductive dysfunction. In: Yen SSC, Jaffe RB, eds. *Reproductive endocrinology*, 3rd edn, pp. 795–806. WB Saunders, Philadelphia.

Vanderpump MPJ and Tunbridge WMG (1993). The effects of drugs on endocrine function. *Clinical Endocrinology* **39,** 389–97.

Wass JAH, Jones AE, Rees LH, Besser GM (1982). HCGB producing pineal choriocarcinoma. *Clinical Endocrinology* **17,** 423–31.

White A, Clark AJL (1993). The cellular and molecular basis of the ectopic ACTH syndrome. *Clinical Endocrinology* **39,** 131–41.

12.13 The pineal gland and melatonin

T. M. Cox

The pineal gland is prominent in birds, reptiles, and other vertebrates in which it responds to sunlight by the secretion of hormones that affect reproduction and thermoregulatory behaviour. In humans the gland has been known since antiquity. Although an endocrine function was considered for many years, this was only given credibility in 1958 by the pioneering work of Lerner, who isolated a small molecule from bovine pineal glands that he named melatonin because it caused blanching of melanophores in vertebrate skin. Many questions remain unanswered about the function of the human pineal gland, but its secretion of the chronobiotic molecule, melatonin, has prompted enormous interest in the fields of travel medicine, neurophysiology, and endocrine research.

Structure

The pineal is less than 1 cm in its longest diameter and weighs less than 0.2 g; it lies above the posterior aspect of the third cerebral ventricle. Normal pineal tissue contains nests of large epithelial-like cells; it also contains neuroglial components, principally of astrocytic type, which occasionally become malignant. Human pineal tissue calcifies with age but this does not necessarily diminish its secretory activity. The pineal gland is considered to reside outside the functional blood–brain barrier.

Pathology

Pineal tumours are rare and are of three types: pinealomas, which represent a neoplastic expansion of the large epithelial cells that cluster within a fibrous stroma in the adult pineal; glial tumours, that resemble astrocytomas occurring elsewhere to cause gliomas in the central nervous system; and teratomas, which arise in the midline within residual pluripotential embryonic cells. It has been known for many years that pineal tumours may disturb sexual maturation and it had been long considered that the pineal secreted a substance that inhibits gonadal function, thus explaining the precocious puberty associated with pineal disease. Treatment of pineal tumours by surgical excision or radiation appears to suppress melatonin secretion leading to sleeping difficulties; melatonin replacement therapy has been reported to benefit such patients with defective melatonin release.

Melatonin

Melatonin, like serotonin, is an endogenous indoleamine derived from tryptophan. The first step in indoleamine synthesis is the 5-hydroxylation of tryptophan by tryptophan hydroxylase—an enzyme with requirements for dioxygen, iron, and tetrahydrobiopterin. The enzyme arylalkylamine, *N*-acetyl transferase, regulated by the sympathetic transmitter noradrenaline, appears to be the rate-limiting step in melatonin synthesis. The enzyme is localized principally in the pineal gland but also within specific cells in the upper gastrointestinal tract. Melatonin binds to specific receptors including the seven-transmembrane G-protein coupled receptors (Mel 1A and 1B) as well as nuclear receptors (including the transcription factor RXR/OR-a) that are associated with melatonin signalling. Membrane G-protein receptors for melatonin are principally expressed in the nervous system whereas the nuclear transcription factor is expressed in the periphery. Melatonin (formal chemical name, *N*-acetyl-5-methoxytryptamine) has been found within membrane-bound bodies in pinealocytes. In experimental animals these show light-dependent morphological changes associated with melatonin secretion under altered environmental light conditions. Melatonin appears to exert its main effects through MT-1 receptors in discrete regions of the hypothalamus that modulate the secretion of growth hormone-releasing hormone (GnRH). This occurs through a complex interneurone circuit which includes neurones containing serotonin, dopamine, and glutamate transmitters.

The role of melatonin and the pineal gland in photoperiodism

Exposure to light influences the secretion of melatonin, and melatonin release is suppressed particularly under illuminance with short-wavelength light. There is evidence that a photoreceptive system, which does not involve retinal rods or cones, mediates this effect. Rhythmic melatonin secretion leads to concentrations in the plasma or cerebrospinal fluid that are up to 10 times higher at night than in the daytime; maximum concentrations are observed in childhood and melatonin levels thereafter decline with age. It appears that melatonin serves as a chronobiotic molecule which acts to delay the sleep–wake cycle of the intrinsic body clock in the suprachiasmatic nucleus. There is evidence that the endogenous clock has a cycle of longer than 24 h, but becomes entrained to synchronize with daily environmental rhythms. Several syndromes associated with long-term insomnia in humans appear to result from slower or faster sleep–wake cycling. Melatonin may play a critical role in such entrainment, since synthesis of melatonin in the pineal induced at night is regulated by sympathetic outflow from the suprachiasmatic nucleus. Exposure to light at high illuminancy may improve disorders of the circadian rhythm that affect sleep. Experiments conducted in blind people have led to the use of synthetic melatonin, before the normal endogenous nocturnal peak, to re-entrain the biological clock. Melatonin-replacement therapy has thus been suggested as a means to improve sleep in night-shift workers, in elderly individuals with insomnia, in patients with pineal tumours, and in those suffering from the effects of jet lag, who have desynchronized rhythms. Careful balance studies have shown that jet lag is associated with asynchronous sleep–wake urine sodium secretion, disturbed fluid balance, and other biochemical abnormalities.

Pharmaceutical use of melatonin

Many studies have been carried out to investigate the efficacy of melatonin as a chronobiotic agent for the alleviation of symptoms associated with rapid eastward travel (jet lag). This application is based on the photoperiodism of humans and other primates and the demonstrated efficacy of melatonin in the regulation of circadian rhythm disturbances in experimental animals. The results of a recent comprehensive study by Herxheimer and Petrie to assess the effectiveness of oral melatonin, taken in different dosing regimens for alleviating jet lag after travel across several time zones, show that the agent is effective in preventing or reducing jet lag. Its short-term use appears to be safe on an occasional basis.

The review showed that in 9 out of 10 trials, melatonin taken close to the target bedtime at destination decreased jet lag in flights crossing five or more time zones. Daily doses of melatonin between 0.5 and 5.0 mg orally appeared to be similarly effective, although sleep was induced faster and better after 5 mg rather than 0.5 mg. Use of a slow-release preparation of 2 mg of melatonin was relatively ineffective in this study suggesting that a short-lived, higher peak was more effective. The drug appears to be relatively safe and side-effect reporting has been low—except in patients with epilepsy or those who are taking warfarin in whom convulsant effects or increased bleeding, respectively, have been reported. It appears possible that melatonin potentiates the action of warfarin. Given that melatonin acts physiologically to regulate the onset of puberty, excess use of melatonin may theoretically influence reproductive development in children and reduce sexual activity, if overused, in adults. No evidence of these effects has yet been reported.

In summary, the evidence is that oral ingestion of melatonin may be indicated for occasional use after transmeridian flights that would induce daytime fatigue and sleep disturbance associated with the gastrointestinal complaints, weakness, malaise, and loss of mental efficiency and other symptoms that typify with jet lag. Clearly, since the drug is not as yet licensed in all countries, routine pharmaceutical control quality must be established. Its use and safety in pregnancy has not yet been completely validated. At a time when prion-related diseases may result from the ingestion or injection of material derived from brain or other animal tissue, only pure biosynthetic melatonin should be considered for human use. Melatonin derived from bovine pineal or other biological sources should be avoided.

Melatonin is widely taken in certain communities, particularly in the United States, where it is claimed to provide indiscriminant protection against ageing, degenerative diseases, cancer, immune dysfunction, and reproductive and psychiatric illnesses. None the less it should be acknowledged that melatonin does have diverse physiological actions in humans, as in other vertebrates, which are incompletely understood. At present the principal indication for exogenous melatonin is for the control of sleep disorders and treatment of symptoms associated with jet lag, rather than the many conditions for which our scientific understanding of its proposed

benefits is as yet inchoate.

Further reading

Herxheimer A, Petrie KJ (2001). Melatonin for preventing and treating jet-lag. (*Cochrane Review*), *Cochrane Database Systems Review* **1**, CD 001520 issue (1).

Karasek M (1998). Melatonin in humans—where are we 40 years after its discovery. *Neuro-endocrinological Letters* **20**, 179–88.

Lewy AJ *et al.* (1992). Melatonin shifts human circadian rhythms according to a phase–response curve. *Chronobiology International* **9**, 380–92.

Zisapel N (2001). Circadian rhythm sleep disorders: pathophysiology and potential approaches to management. *CNS Drugs* **15**, 311–28.

13.1 Physiological changes of normal pregnancy

D. J. Williams

Since modern *Homo sapiens* emerged 100 000 years ago, it is estimated that the human population has increased from 50 000 to around 6000 million. Such reproductive success has defied gross reproductive inefficiency. Only 20 to 35 per cent of fertilized ova result in a successful pregnancy (10–25 per cent of *in vitro* fertilizations), most failing around the time of implantation with chromosomal abnormalities. The usual outcome of a successful pregnancy is a single offspring, produced after 9 months. Indeed, if the fetal brain was not programmed to outgrow the maternal pelvic outlet, anthropological comparisons with the great apes indicate that human pregnancy would last 16 months. The price of a relatively short pregnancy is neonatal immaturity and a prolonged period of nurture.

Preparing for pregnancy

The female body prepares for pregnancy during every menstrual cycle. It is not only the endometrium that anticipates implantation of a fertilized ovum, but the whole cardiovascular system. During the postovulatory or luteal phase of each menstrual cycle there is a decrease in systemic vascular resistance by approximately 20 per cent, leading to a 10 per cent fall in mean arterial pressure compared with the follicular phase. Cardiac output increases by almost 20 per cent, and renal vasodilatation increases both renal blood flow and glomerular filtration by approximately 10 per cent. All of these changes resolve with involution of the corpus luteum and onset of menses.

Haemodynamic changes in pregnancy

If fertilization is successful, there is progression of the haemodynamic changes established in the menstrual cycle. A progressive fall in systemic vascular resistance by up to 40 per cent creates a maximal decrease in mean arterial pressure by the end of the first trimester. Diastolic blood pressure falls between 5 and 15 mmHg, before rising to non-pregnancy levels at term, whilst systolic blood pressure remains unchanged throughout pregnancy. A gestational increase in heart rate from approximately 72 to 85 beats/min and of stroke volume by up to 30 per cent combine with the reduction in systemic vascular resistance to increase cardiac output. By 24 weeks, cardiac output reaches a maximum of 50 per cent above non-pregnant levels, which is sustained until term, except in the supine position during the third trimester when cardiac output falls as the gravid uterus compresses the inferior vena cava. Left ventricular wall thickness and left ventricular mass increase progressively throughout pregnancy, by up to 30 per cent and 50 per cent respectively. Cardiac output returns almost to prepregnancy levels within 2 weeks of delivery.

Distribution of increased cardiac output

Although it is technically difficult to measure blood flow to particular maternal viscera during pregnancy, it is clear that the timing and extent of changes to blood flow varies between organs. This is summarized in Fig. 1. Mammary artery blood flow increases early in pregnancy, breast tenderness and swelling being amongst the first symptoms.

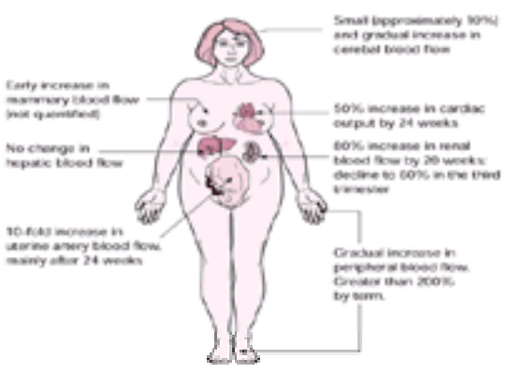

Fig. 1 Changes in maternal organ blood flow during healthy pregnancy.

Mechanism of haemodynamic change

The onset of physiological change during the menstrual cycle suggests that maternal rather than fetoplacental factors initiate gestational adaptation. Oestrogen, mainly in the form of 17B-oestradiol, is a potent vasodilator. It is produced by the corpus luteum during the luteal phase of each menstrual cycle and for the first 10 weeks of pregnancy. After 10 weeks, the placenta elaborates its own 17B-oestradiol, so that by term maternal oestradiol levels are approximately 250-fold higher than those found during the menstrual cycle. 17B-oestradiol relaxes vascular smooth muscle through both endothelium-dependent and independent mechanisms. All of the endothelium-derived vasodilators, nitric oxide, prostacyclin, and endothelial-derived hyperpolarizing factor have been implicated in the gestational fall of systemic vascular resistance. Much less is known about the vascular effects of progesterone. Circulating progesterone levels increase by a similar amount to 17B-oestradiol and may play a role in reducing pressor responsiveness to angiotensin II. Although the precise mechanism of maternal vasodilatation is likely to be different in different vascular beds, a healthy endothelium is essential for normal cardiovascular adaptation to pregnancy.

Fluid balance during pregnancy

Arterial dilatation creates a relatively 'under-filled' state, which stimulates the renin–angiotensin–aldosterone system. As a result, sodium and water retention throughout pregnancy leads to a 6 to 8 litre rise in total extracellular fluid volume. An increase in plasma volume is apparent by week 6 and continues until week 32 when it is 40 per cent (approximately 1.2 l) above non-pregnant levels. Furthermore, shortly after conception the osmotic threshold for thirst falls and plasma osmolality drops by 10 mosmol/kg. A concomitant fall in the threshold for secretion of antidiuretic hormone (AVP) prevents a water diuresis and sustains low plasma osmolality until term. During the second half of pregnancy, placental production of vasopressinase increases maternal antidiuretic hormone degradation, but plasma antidiuretic hormone levels remain stable as pituitary secretion of antidiuretic hormone increases four-fold. Plasma atrial natriuretic peptide levels are normal until the second trimester, when they rise by approximately 40 per cent.

Immunological changes during pregnancy

It is often presumed that pregnant women are immunosuppressed in order that the fetal 'semiallograft' can survive. This is not true. Certain aspects of maternal immunity are modulated, but it is the placenta that deserves most credit for eluding maternal immunity. Much harm is prevented by the physical separation of maternal and fetal blood. Fetal haemolytic disease is an example of the harm that can follow a breach in this barrier, a Rhesus-negative mother becoming isoimmunized against Rhesus-positive fetal blood.

In normal pregnancy, the placenta has to invade uterine tissue and become bathed in maternal blood. To avoid a hostile immune response, the surface layers of placenta express a unique non-polymorphic HLA G, rather than classical histocompatibility antigens. It is thought that HLA-G confers resistance to lysis by maternal T cells and natural killer cells. The placenta also expresses a plethora of complement control systems to protect itself from maternal complement (serum levels of complement factors C3 and C4 are elevated during pregnancy).

The number and activity of natural killer cells in maternal peripheral blood are diminished during pregnancy, and fetal survival is further enhanced by a shift away from maternal T-helper 1 responses that promote cell-mediated immunity, towards a stronger T-helper 2 response that promotes antibody production. In consequence, pregnant women are more prone to severe infections with intracellular pathogens such as malaria and leprosy and are more likely to suffer reactivation of viruses such as Epstein–Barr. Conversely, circulating levels of maternal immunoglobulin increase and once transferred to the fetus have a role in passive immunity. The mechanism by which these immune changes take place is unclear, but increased circulating levels of metabolically active cortisol may play a role.

Healthy pregnancy is also a proinflammatory state. Neutrophils increase in number and develop a proinflammatory phenotype. Mean total white cell count increases to 9.0×10^{9}/l and can rise as high as 40.0×10^{9}/l during labour, returning to normal within 6 days. Erythrocyte sedimentation rate (ESR) rises as a consequence of increased fibrinogen and globulin. An ESR over 30 mm/h is usual and up to 70 mm/h is within normal limits. Circulating levels of C reactive protein do not change during healthy pregnancy. Anatomical changes to the maternal immune system include involution of the thymus and enlargement of the spleen.

Ventilatory changes during pregnancy

The increased metabolic demands of pregnancy lead to a progressive increase in oxygen consumption, reaching almost 20 per cent by term. To compensate, pregnant women breathe more deeply, tidal volume increasing from approximately 500 to 700 ml, whilst respiratory rate remains unchanged. Effective alveolar ventilation actually surpasses the body's demand for oxygen, creating a respiratory alkalosis with PCO_2 falling from 5.0 to 4.0 kPa. Over-breathing is stimulated by direct effect of progesterone on the respiratory centre, particularly increasing sensitivity to CO_2.

Renal changes during pregnancy

An 80 per cent increase in renal blood flow and 55 per cent increase in glomerular filtration rate occur by 16 weeks gestation. The rise in renal blood flow causes the kidneys to swell so that they appear approximately 1 cm longer on ultrasonography. The renal pelvis and ureters dilate, sometimes appearing obstructed to those unaware of these changes.

Serum levels of creatinine and urea fall, so that levels considered normal outside pregnancy can suggest renal impairment during pregnancy. Proteinuria increases slightly during pregnancy, but levels over 260 mg/24 h should be considered abnormal. Gestational glycosuria reflects reduced tubular glucose reabsorption and does not necessarily indicate abnormal carbohydrate metabolism. Furthermore, reduced tubular absorption of bicarbonate creates a metabolic acidosis that compensates for the respiratory alkalosis, keeping maternal pH at 7.4.

The production of all three renal hormones, erythropoietin, active vitamin D, and renin, increases during healthy pregnancy, but their effects are masked by other physiological changes. In early pregnancy, peripheral vasodilatation exceeds the renin–aldosterone mediated plasma volume expansion, so blood pressure falls by 12 weeks. The 40 per cent expansion of plasma volume exceeds the effect of a two to four-fold increase in maternal serum erythropoietin levels, which stimulates only a 25 per cent rise in red cell mass. This creates a 'physiological anaemia', which should not normally cause haemoglobin concentration to fall to less than 9.5 g/dl (see Chapter 13.3). Similarly, active vitamin D circulates at twice non-gravid levels, but concomitant halving of parathyroid hormone levels, as well as hypercalciuria and increased fetal requirements, keeps plasma ionized calcium levels unchanged.

Liver metabolism during pregnancy

The size of the liver and its blood flow appear not to change during healthy pregnancy, hence liver blood flow accounts for proportionately less of the cardiac output as pregnancy progresses. There are, however, changes to hepatic synthetic function and metabolism. Circulating concentrations of fibrinogen, ceruloplasmin, transferrin, and binding-proteins, for example thyroid-binding globulin, increase, while serum albumin levels fall by approximately 25 per cent. Serum cholesterol increases by 50 per cent and triglycerides by up to 300 per cent. The normal ranges for aspartate transaminase, alanine transaminase, gamma glutamyl transferase, and bilirubin decrease by at least 20 per cent from the first trimester until term. After the fifth month, placental production of alkaline phosphatase increases maternal plasma levels by up to four-fold. Telangiectasia and palmar erythema are common signs of healthy pregnancy that resolve postpartum.

Gastrointestinal system

Nausea and vomiting affect about 60 per cent of women during the first trimester. The rise and fall of human chorionic gonadotrophin (hCG) levels correlate chronologically with the onset and improvement of these symptoms, but the role of hCG in gestational nausea is unproven and the cause remains unknown. Relaxation of intestinal smooth muscle by progesterone creates many of the other pregnancy-induced gastrointestinal changes. Gastric motility and small bowel transit are slowed, especially during labour. The gallbladder enlarges and empties slowly in response to meals. A decrease in lower oesophageal pressure makes gastro-oesophageal reflux more common.

Endocrine changes

Thyroid function

During pregnancy, the thyroid faces three challenges. Firstly, increased renal clearance of iodide and losses to the fetus create a state of relative iodine deficiency. In geographical areas where dietary iodine intake is low, pregnancy stimulates growth of thyroid goitres. Secondly, high oestrogen levels induce hepatic synthesis of thyroid binding globulin, but free thyroxine and tri-iodothyronine levels remain within the normal range throughout pregnancy. Thirdly, placental hCG shares structural similarities with thyroid-stimulating hormone and has weak thyroid-stimulating hormone-like activity. Although hCG rarely stimulates free T4 levels into the thyrotoxic range, trophoblastic disease and hyperemesis gravidarum are often associated with high hCG levels and can lead to hyperthyroxinaemia. In these circumstances, the mother remains clinically euthyroid.

Pituitary function

Once ovulation has occurred and the uterus is prepared for implantation, the maternal pituitary makes only a small contribution to a successful pregnancy. The only pituitary hormone to increase significantly during pregnancy (by approximately 10-fold) is prolactin, which is responsible for breast development and subsequent milk production.

Pituitary secretion of growth hormone is mildly suppressed during the second half of pregnancy by placental production of a growth hormone variant. The role of the latter is unclear, but may contribute to gestational insulin resistance.

Placental production of adrenocorticotrophic hormone leads to an increase in maternal adrenocorticotrophic hormone levels, but not beyond the normal range for non-pregnant subjects. Free cortisol levels double and in the second half of pregnancy may contribute to insulin resistance and striae gravidarum.

High oestrogen levels during pregnancy stimulate lactotroph hyperplasia, resulting in pituitary enlargement. These high levels, together with those of progesterone, suppress luteinizing hormone and follicular stimulating hormone. Plasma follicular stimulating hormone levels recover within 2 weeks of delivery, but pulsatile luteinizing hormone release is only resumed in women who do not breast feed. In suckling mothers, prolactin inhibits gonadotrophin-releasing hormone and hence luteinizing hormone.

Coagulation

In anticipation of haemorrhage at childbirth, normal pregnancy is characterized by low grade, chronic intravascular coagulation within both the maternal and uteroplacental circulation. There are increased levels of clotting factors (V, VIII, and X), decreased levels of the endogenous anticoagulant protein S and decreased fibrinolytic activity. These changes lead to an acquired protein C resistance in up to 38 per cent of pregnant women. However, postpartum contraction of the uterus by oxytocin is probably more effective at preventing haemorrhage than any changes to the coagulation system.

Carbohydrate metabolism

During the first trimester, women are more sensitive to insulin than when non-pregnant. From 20 weeks onwards, insulin resistance develops, hence women in the second half of pregnancy respond to a glucose load by producing more insulin, but with less effect. Obese women, who are already insulin resistant, are more likely to develop gestational diabetes. Hormones that might mediate this insulin resistance include cortisol, progesterone, oestrogen, and human placental lactogen. Placental production of human placental lactogen, a growth hormone-like protein, coincides temporally with insulin resistance.

Skin and hair during pregnancy

Hyperpigmentation affects up to 90 per cent of pregnant women. Areas that are normally hyperpigmented, such as the areolae and vulva, become darker. This may be mediated by oestrogen and progesterone, which are powerful melanogenic stimulants. Hair growth increases during pregnancy and hair loss is accelerated postpartum. The gestational rise in corticosteroids and ovarian androgens contributes to the number of hairs in the growing phase (anagen). Postpartum, the levels of these hormones fall and hairs move back into the resting phase (telogen).

Further reading

Chamberlain G, Broughton-Pipkin F, eds (1998). *Clinical physiology in obstetrics*, 3rd edn. Blackwell Science, Oxford. [Comprehensive text on physiological changes in healthy pregnancy.]

De Swiet M, ed (1995). *Medical disorders in obstetric practice*, 3rd edn. Blackwell Science, Oxford. [Each chapter describes the physiology of a different organ system before giving details of pathophysiology.]

References

Chapman AB *et al*. (1997). Systemic and renal hemodynamic changes in the luteal phase of the menstrual cycle mimic early pregnancy. *American Journal of Physiology* **273**, F777–82. [Carefully conducted study on physiological changes during the menstrual cycle.]

Chapman AB, *et al*. (1998). Temporal relationships between hormonal and hemodynamic changes in early human pregnancy. *Kidney International* **54**, 2056–63.[Serial study correlating haemodynamic with neurohumoral changes from preconception to 36 weeks gestation.]

Gill TJ (1997). Genetic factors in reproduction and their evolutionary significance. *American Journal of Reproductive Immunology* **37**, 7–16. [Review of evolution of reproduction and development of immunity.]

Lindheimer MD, Davison JM, eds (1994). Renal disease in pregnancy. *Baillieres Clinical Obstetrics and Gynaecology* **8**, 209–527. [Several chapters on renal function and fluid balance during healthy pregnancy.]

Poston L and Williams DJ (1999). The endothelium in human pregnancy. In: Vallance P, Webb D, eds. *Vascular endothelium in human physiology and pathophysiology*, pp. 247–81. Harwood Academic Publishers, Amsterdam. [Review of role of endothelium in vascular changes of pregnancy.]

Robson SC, *et al*. (1989). Serial study of factors influencing changes in cardiac output during human pregnancy. *American Journal of Physiology* **256**, H1060–5. [Comprehensive serial study on cardiovascular haemodynamics during healthy pregnancy.]

13.2 Nutrition in pregnancy

D. J. Williams

Introduction

The ability to adapt to different environmental and nutritional conditions is a key requirement for reproductive success. In the developed world, where food is generally plentiful, dietary recommendations are based on the average food intake amongst healthy pregnant women. In nations where food is scarce, dietary recommendations are based on minimal requirements for health and fall far below the average intake of a woman who eats to satisfy her appetite. Pregnant women adapt several metabolic pathways to minimize extra nutritional requirements and optimize fetal growth. However, despite these metabolic adaptations, millions of pregnant women are unable to provide enough nutrition for their fetus to thrive. Poor prenatal nutrition not only affects perinatal outcome, but also appears to dictate susceptibility to some adult diseases and possibly the health of the next generation.

Weight gain in pregnancy

Well-nourished mothers with free access to food gain up to 30 per cent of their prepregnancy weight, of which only 25 per cent is fetal. By contrast, mothers with limited access to food gain as little as 10 per cent of their prepregnancy weight, of which up to 60 per cent is fetal. Liberal weight gain increases birth weight, but also increases the rate of caesarean section and maternal complications such as gestational diabetes and pre-eclamspia. Limiting weight gain increases the incidence of low birth weight (defined as less than 2500 g at term). In 1990 the Institute of Medicine in the United States published guidelines for weight gain in pregnancy that were adjusted according to prepregnancy maternal weight (Table 1). These recommendations, based on large observational studies, have stood the test of many subsequent analyses and minimize the overall risk of both low and high (more than 4500 g at term) birth weights.

Caucasian women who kept within the Institute of Medicine recommendations retained 1 kg postpartum, while black women retained 3 kg. Opponents of the recommendations therefore believe that they are too generous and encourage postpartum weight retention. However, only 23 per cent of obese women could keep within the guidelines, and those unable to do so doubled their risk of a poor pregnancy outcome and increased the likelihood of postpartum weight retention. Unless the mother is under- or overweight (BMI less than 19.8 or more than 29 kg/m^2), measurement of weight gain during pregnancy is a poor predictor of pregnancy outcome.

Pregnancy weight gain in the developing world

In developing nations more than 20 per cent of babies are of low birth weight, of which only 25 per cent are premature; in comparison, in developed nations only 6 per cent of babies are of low birth weight, of which most (55 per cent) are premature. Nutritional supplements for malnourished women during pregnancy need to be administered with care. Chronic malnutrition limits maternal stature, including pelvic size. Hence protein and energy supplements in pregnancy may disproportionately increase fetal growth and lead to obstructed labour, a major cause of maternal and perinatal death in the developing world. A pragmatic recommendation is to be particularly aware of this possibility when including primiparous women of less than 1.5 m in height in supplementary feeding programmes aimed at accelerating fetal growth. Furthermore, improved obstetric care must accompany nutritional advice.

Energy requirements during pregnancy

The rate of human fetal growth is slow and the daily incremental energy stress of human pregnancy is relatively low compared with that in other species. This allows a mother time to adapt her metabolism and energy expenditure to diverse nutritional conditions. In well-nourished societies the total energy costs of pregnancy can be as high as 520 MJ (124 000 kcal), compared with –30 MJ (–7100 kcal) in countries where food is scarce (Fig. 1).

Fig. 1 Estimated total energy costs of pregnancy in different nutritional environments. The women from The Gambia were supplemented with a balanced protein-energy diet. (From Prentice and Goldberg (2000).)

The three major components of energy expenditure in an average well-nourished mother are growth of the fetus and reproductive tissues (about 18 per cent), new maternal fat stores (about 38 per cent), and increased maternal metabolism (about 44 per cent). Poorly nourished women try to maintain fetal growth by depressing their basal metabolic rate until late pregnancy and by laying down less fat. Although such adaptations usually result in successful reproduction, they are inevitably a compromise with regards to perinatal health. However, attempts to quantify minimal energy requirements for good perinatal health will always be confounded by huge individual variability and practical difficulties of attributing a single nutritional component to morbidity—a multifaceted problem.

In well-nourished women, the basal metabolic rate changes little until about 16 weeks' gestation, then increases rapidly until term (Fig. 2). During the middle trimester large amounts of maternal fat are laid down as energy stores. If food intake becomes limited during late pregnancy, maternal fat can be mobilized to support the period of most rapid fetal growth. This strategy of fat storage before anticipated energy demands is also used by birds before migration and hibernating mammals. Even poorly nourished women with low gestational weight gains lay down some extra fat. Conversely, well-nourished women with free access to food rarely need to

utilize all their fat stores to support late fetal growth. Excess fat remains difficult to lose postpartum.

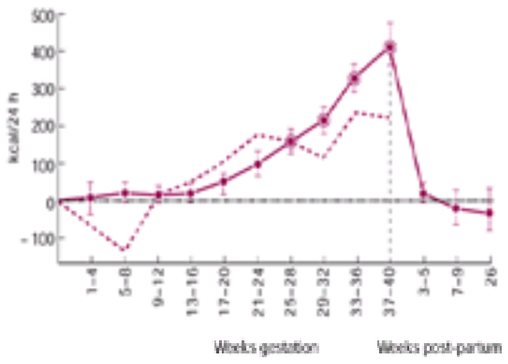

Fig. 2 The energy cost of gestational increase in basal metabolic rate in well nourished women in the United Kingdom (data derived from over 100 pregnancies). This amounts to 30 100 kcal (solid line) each pregnancy, when the total gestational energy requirement was 69 050 kcal. Total energy intake increased by only 22 000 kcal (dashed line). Maternal energy expenditure must reduce by 47 000 kcal to meet the extra demands of pregnancy. (Adapted from Durnin (1991), with permission.)

In the United Kingdom, the calculated total energy cost of pregnancy is about 70 000 kcal (Table 2). As a consequence, pregnant women have been recommended to increase energy intake by around 250 kcal/day during the last two trimesters. However, careful studies of well-nourished women have found that maternal energy intake actually increases by little more than 100 kcal/day after the first trimester (22 000 kcal in total) (Fig. 2). The shortfall of nearly 50 000 kcal is made up by an economy of energy expenditure, including reduced physical activity and diet-induced thermogenesis.

The mechanisms that control the diverse metabolic responses to pregnancy are not understood. Leptin, a protein produced by adipose tissue and the placenta, circulates in increasing amounts during pregnancy and controls peripheral energy status and body fat. Ob/ob mice (deficient in leptin) become insensitive to exogenous leptin when pregnant, consistent with the build-up of maternal fat stores until the last trimester. Furthermore, leptin is integrated with the hypothalamo–pituitary–gonadal axis and may explain why thin women with low leptin levels remain infertile until they have adequate fat stores. A precise role for leptin during pregnancy remains to be elucidated.

Carbohydrate metabolism

During the first half of pregnancy, women produce more insulin in response to a glucose load and are more sensitive to exogenous insulin than in the non-pregnant state. These changes affect carbohydrate and lipid metabolism to favour increased fat production and storage. During the second half of pregnancy a woman becomes increasingly resistant to insulin, so that at term the action of insulin is 50 to 70 per cent lower than in the non-gravid state. As a consequence the fat stores laid down in the first half of pregnancy are mobilized and postprandial blood glucose levels remain higher for longer. Circulating levels of fatty acids and glycerol increase and are used by the mother as an energy source in preference to glucose and amino acids, which are left for the fetus. As a consequence, fasting pregnant women oxidize fat and produce ketones far sooner than they do when they are not pregnant. Women with an exaggerated peripheral resistance to insulin develop gestational diabetes mellitus.

Protein metabolism

Pregnancy is an anabolic state. Protein and nitrogen metabolism adapt early and gradually throughout healthy pregnancy to provide for tissue growth. Well-nourished women are estimated to accumulate an extra 500 g to 1 kg of protein during pregnancy. Almost half of the protein accumulates as increased maternal lean body mass, while the rest lies within the fetus and reproductive tissues.

In the United Kingdom the daily increment of dietary protein has been calculated to increase gradually throughout pregnancy to 8.5 g at term, but this does not take into account reduced hepatic metabolism of branched chain amino acids and hence reduced urea synthesis. The rate of urea synthesis declines by 30 per cent during the first trimester and by 45 per cent during the third trimester, hence serum urea concentration falls, providing more nitrogen for protein synthesis.

Vitamins and micronutrients

In some parts of the developing world micronutrient deficiencies are endemic and have serious consequences for fetal, neonatal, and maternal well-being, for example hypothyroidism due to iodine deficiency and night blindness due to vitamin A deficiency. Such deficiencies are rare in developed countries.

Calculated increments in the recommended daily allowance of specific nutrients are derived from estimates of the cost of fetal growth and increased maternal metabolism. These calculations do not usually take account of maternal metabolic adaptations that make the need for extra nutrients unnecessary, for example increased intestinal absorption of calcium offsets the need for an increase in dietary calcium. Conversely, increased folic acid excretion leads to an underestimate of folic acid requirements. Furthermore, individual micronutrients interact with each other and changes to one may have a detrimental effect on the activity of another.

It is now widely accepted that supplemental folic acid (400 µg/day) during the first trimester reduces the risk of neural tube defects. With this exception, extra vitamins and micronutrients are not necessary for healthy pregnant women who eat a balanced diet. Indeed, excessive amounts of certain micronutrients can be harmful to the fetus.

Vitamin A

Vitamin A is a lipid-soluble vitamin essential for healthy embryogenesis and fetal growth. Preformed vitamin A is found in dairy products and liver: the recommended daily allowance during pregnancy is 2000 to 2700 iu/day (670–899 retinol equivalents; RE). Vitamin A deficiency is endemic in some parts of the world: some small studies have shown a minor increase in birth weight with maternal vitamin A supplements of 6000 to 8000 iu/day (2000–2670 RE/day). Breast milk is rich in vitamin A, and is important for neonatal immunity.

Excessive doses of vitamin A in the diet (more than 15 000 iu/day (5000 RE/day)), or supplements of vitamin A (more than 10 000 iu/day), are teratogenic. Pregnant women who take vitamin A supplements of more than 10 000 iu/day (3335 RE/day) have a 1 in 57 risk of a birth defect attributable to this. Drugs that are derived from vitamin A, such as the retinoids (for example isotretinoin), are associated with an estimated 25-fold increased risk of malformation. As a consequence, the American College of Obstetricians and Gynecologists has recommend that the daily dose of vitamin A should not exceed 5000 iu/day (1665 RE/day) during pregnancy. The carotenoids (b-carotene) that are precursors to vitamin A do not appear to be teratogenic and are now being substituted for preformed vitamin A in multivitamin preparations. In general, vitamin A supplements are unnecessary for well-nourished women and potentially harmful to the fetus.

Thiamine (vitamin B_1)

Thiamine deficiency is endemic in some developing countries, but is also a global problem in women with hyperemesis gravidarum. Severe and persistent vomiting during pregnancy leads to thiamine deficiency and can cause Wernicke's encephalopathy. Thiamine replacement is therefore an essential supplement for women with hyperemesis gravidarum.

Vitamins C and E

Serum vitamin C levels fall by about 50 per cent during pregnancy, hence it is recommended that this amount is supplemented, although benefits are unproven. The

antioxidant properties of vitamins C and E may reduce the risk of pre-eclampsia, but larger studies are needed to confirm this possibility.

Iodine

More than 800 million people live in iodine-deficient areas. Inadequate dietary iodine leads to maternal hypothyroidism and is detrimental to *in utero* growth and development. Supplemental iodine (usually added to salt) given to pregnant women can prevent these consequences.

Zinc

Zinc deficiency is associated with intrauterine growth restriction and teratogenesis. During pregnancy, maternal zinc levels remain stable through increased intestinal absorption. Excess iron supplements, smoking, alcohol abuse, or subsistence cereal diets high in phytate can all inhibit zinc absorption: under such conditions pregnant women may benefit from 25 mg zinc daily.

Iron

During pregnancy expansion in plasma volume exceeds the increase in red cell mass causing a fall in haemoglobin concentration. Healthy pregnant women not taking iron supplements drop their haemoglobin from 13.3 g/dl to 11.0 g/dl by 36 weeks' gestation. The minimum incidence of low birth weight (less than 2500 g at term) and preterm labour is associated with a haemoglobin of 9.5 to 10.5 g/dl. In the non-gravid state, a haemoglobin of 9.5 to 10.5 g/dl would indicate anaemia, but unless the mean corpuscular volume is less than 84 fl, supplemental iron is probably unnecessary. A meta-analysis of randomized controlled trials examining the benefit of supplemental iron found a significant reduction in women with a haemoglobin of less than 10 g/dl, but no effect, beneficial or harmful, on maternal or fetal outcome.

In the developing world, anaemia (of multiple causes) is endemic. The risk of maternal death is increased with severe anaemia (haemoglobin less than 7.0 g/dl), a condition where supplemental iron is unlikely to have much effect. Evidence that mild to moderate anaemia is associated with increased maternal and fetal risk is hard to find. Despite this, many developing countries advocate a policy of iron and folic acid supplementation for all pregnant women. More studies are necessary to monitor the effects of this policy on maternal and perinatal outcome.

Anaemia in pregnancy is discussed in more detail in Chapter 13.16.

Calcium

The growing fetus gains about 50 mg calcium per day by midpregnancy and about 300 mg/day at term. The breastfed infant receives about 250 mg of calcium in breast milk each day. The recommended daily allowance of calcium during pregnancy and lactation is 1.2 g/day, but women with much less dietary calcium undergo metabolic adaptations to meet the demands of pregnancy and lactation without any detriment to their health or that of the fetus.

During pregnancy, maternal calcium absorption increases twofold, stimulated by increased 1,25-dihydroxyvitamin D activity due to placental synthesis of 1,25-dihydroxyvitamin D and increased renal 1-a-hydroxylase activity. Although urinary calcium excretion doubles during pregnancy, fasting urinary calcium excretion, corrected for the increased creatinine clearance, is unchanged. The concentration of parathyroid hormone falls during pregnancy, suggesting that the pregnant woman receives enough calcium for her growing fetus. There are two caveats: one is the pregnant adolescent who needs to meet the demands of her own growth and that of the fetus; the other is the apparent benefit of supplemental calcium for women on a low-calcium diet, not a normal calcium diet, to prevent pre-eclampsia.

Following delivery, circulating 1,25-dihydroxyvitamin D concentrations return to non-pregnant levels. During the first 3 to 6 months of breastfeeding, mineralization of the maternal axial skeleton declines by approximately 3 to 5 per cent. After 6 months, bone demineralization recovers whether or not breastfeeding continues. Calcium supplements of 1 g/day given to lactating women do not prevent bone demineralization or improve the calcium concentration of breast milk, even if the woman is on a low-calcium diet. Furthermore, repeated long periods of breastfeeding in women with a low calcium intake do not contribute to osteoporosis in later life.

Fetal programming—the influence of fetal nutrition on adult disease

Epidemiological studies have found that low birth weight due to intrauterine growth restriction (rather than prematurity) is associated with an increased risk of cardiovascular disease in adulthood. It is hypothesized that a poorly growing fetus makes metabolic adaptations *in utero* to optimize growth and development. Despite these physiological adaptations birth weight remains low, and because of them the individual is indelibly programmed to insulin-resistant syndromes that are detrimental to long-term cardiovascular health. These issues are discussed in Chapter 15.4.1.1.

It has also been suggested that impaired insulin-mediated fetal growth and insulin resistance in later life is genetically determined. Fetal growth is partly genetically programmed, but the intrauterine environment appears to be more important: the relationship between birth weight and the body size of a surrogate mother who receives a donor egg is stronger than with the genetic mother.

Animal studies have shown that the composition of maternal diet can influence fetal growth and consequently blood pressure in her offspring. At present not enough is known about the mechanisms that control human fetal growth to give maternal nutritional advice that might eventually reduce the risk of cardiovascular disease in her children. Understanding these mechanisms may be fundamental to ameliorating the global epidemic of cardiovascular disease.

Foods to avoid during pregnancy

Food contaminated with *Listeria monocytogenes* can cause listeriosis. During pregnancy, this organism has a predilection to replicate at the uteroplacental site, leading to septic abortion in early pregnancy, or neonatal listeriosis in later pregnancy. To reduce the risk of infection with listeria, pregnant women should avoid eating soft ripened cheeses, all types of pâté, and undercooked meats. This is discussed in section 13.15.

Acute maternal infection with *Toxoplasma gondii* can cross the placenta to the fetus. Congenital infection is least likely during early pregnancy, but more severe when it occurs. The risk of congenital infection can be kept to a minimum by not eating undercooked meat, taking care while handling raw meat, and avoiding contact with cat faeces. This is discussed in Chapter 13.15.

Food cravings during pregnancy

Common food cravings during pregnancy are for dairy products and occasionally for non-organic material such as soil (pica). Common aversions are to alcohol, caffeine, and meats.

Further reading

Abrams B, Altman SL, Pickett KE (2000). Pregnancy weight gain: still controversial. *American Journal of Clinical Nutrition* **71** (suppl.), 1233S–1241S.

Atallah AN, Hofmeyr GJ, Duley L. (2000). Calcium supplementation during pregnancy for preventing hypertensive disorders and related problems. *Cochrane Database System Review*.

Butte NF (2000). Carbohydrate and lipid metabolism in pregnancy: normal compared with gestational diabetes mellitus. *American Journal of Clinical Nutrition* **71** (suppl.), 1256S–1261S.

Durnin JVGA (1991). Energy requirements of pregnancy. *Diabetes* **40** (suppl. 2), 152–6.

Campbell-Brown M, Hytten FE (1998). Nutrition. In: Chamberlain G, Broughton-Pipkin F, eds. *Clinical Physiology in Obstetrics*, 3rd edn, pp 165–91. Blackwell Science, Oxford. A thorough review of nutrition in pregnancy.

Hattersley AT, Tooke JE (1999). The fetal insulin hypothesis: an alternative explanation of the association of low birthweight with diabetes and vascular disease. *The Lancet* **353**, 1789–92.

Institute of Medicine (United States) (1990). *Nutrition during pregnancy. Report of the Committee on Nutritional Status during pregnancy and lactation, food and nutrition board*. National Academy

Press, Washington, DC.

Kalhan SC (2000). Protein metabolism in pregnancy. *American Journal of Clinical Nutrition* **71** (suppl.), 1249S–1255S.

Kalkwarf HJ *et al*. (1997). The effect of calcium supplementation on bone density during lactation and after weaning. *New England Journal of Medicine* **337**, 523–8.

Koop-Hoolihan LE *et al*. (1999). Longitudinal assessment of energy balance in well-nourished, pregnant women. *American Journal of Clinical Nutrition* **69** (suppl.), 697–704.

Mahomed K (2000). Iron supplementation in pregnancy (Cochrane Review). *The Cochrane Library*, Issue 3. Update Software, Oxford.

O'Brien SPM, Wheeler T, Barker DJP, eds (1999). *Fetal programming. Influences on development and disease in later life.* RCOG Press, London. A comprehensive series of reviews and research articles on fetal programming.

Prentice A. (2000). Calcium in pregnancy and lactation. *Annual Review of Nutrition* **20**, 249–72.

Prentice AM, Goldberg GR (2000). Energy adaptations in human pregnancy: limits and long term consequences. *American Journal of Clinical Nutrition* **71** (suppl.), 1226S–1232S.

Ramakrishnan U *et al*. (1999). Micronutrients and pregnancy outcome: A review of the literature. *Nutrition Research* **19**, 103–59.

Rothman KJ *et al*. (1995). Teratogenicity of high vitamin A intake. *New England Journal of Medicine* **333**, 1369–73.

Rush D (2000). Nutrition and maternal mortality in the developing world. *American Journal of Clinical Nutrition* **72** (suppl.), 212S–240S.

13.3 Medical management of normal pregnancy

D. J. Williams

Introduction

Until very recently *Homo sapiens* thrived with nothing but the most primitive antenatal care. The introduction of hospital-based childbirth in the United Kingdom was a disaster. In the mid 1800s it became clear to some that unhygienic medical practice was responsible for puerperal sepsis and a high maternal mortality rate. From the 1930s, maternal mortality in the United Kingdom fell from 1 in 100 deliveries in the worst maternity hospitals, to 1 in 10 000 deliveries today. Globally, however, there are still 585 000 pregnancy-related maternal deaths each year, meaning that one woman dies every minute of every day as a consequence of pregnancy and childbirth.

With the exception of a high death rate from AIDS in many developing nations, the causes of maternal mortality worldwide are similar to those found in developed countries before the implementation of modern obstetric practices (Fig. 1). In the developed world, the dramatic reduction in the number of maternal deaths from obstetric complications has not been matched by a similar fall in deaths associated with pre-existing maternal disease. This latter observation is partly due to the success of modern medicine in helping more women with congenital or chronic disease to survive until reproductive age, and partly due to the inability of physicians to manage otherwise familiar medical conditions during pregnancy.

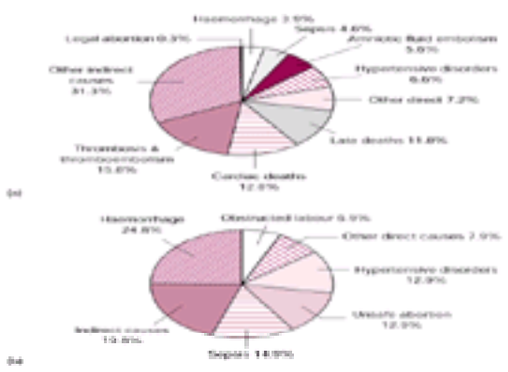

Fig. 1 (a) Causes of maternal mortality in the United Kingdom 1994–6 (adapted from Department of Health (1998)). (b) Causes of 585 000 maternal deaths worldwide in 1990. Direct deaths are a direct consequence of pregnancy; indirect deaths are due to the effects of pregnancy on pre-existing maternal disease. (Reproduced from Liljestrand (1999), with permission.)

Misplaced concern about fetal welfare often denies the mother life-saving investigations and treatment. Substandard care is therefore responsible for many maternal deaths in consequence. Clinical anxiety may be amplified when a doctor is presented with a healthy woman who has everything to lose by meddlesome intervention. The general physician should therefore be aware of the symptoms and signs of normal pregnancy and familiar with advice on how women should prepare for and maintain a healthy pregnancy.

Maternal factors that influence pregnancy outcome

Maternal age

More women are having babies later in life than ever before. In the United States, between 1969 and 1994, the median age at first birth increased from 21.3 to 24.4 years and the proportion of first time mothers aged 30 years or more increased from 4 per cent to 21 per cent. The prevalence of pregnancy-induced hypertension, gestational diabetes, and thrombosis is increased in women over 35 years.

Fetal aneuploidy, most notably trisomy 21 (Down's syndrome), also increases with maternal age. At 25 years of age, the risk of a pregnancy with trisomy 21 is 1:1250, at 35 years it is 1:385, and at 45 years it is 1:30. These risks can be refined during pregnancy (from 16 weeks) by information derived from measurement of maternal serum concentrations of a-fetoprotein, human chorionic gonadotrophin, and unconjugated oestriol (the triple test). Some centres derive a risk of chromosomal abnormality from an ultrasound measurement of skinfold thickness at the back of the fetal neck (nuchal translucency screening, around 12 weeks). Women found to be at high risk of a chromosomal abnormality can be offered diagnostic testing with amniocentesis (which carries a 0.5 to 1.0 per cent risk of miscarriage).

Maternal weight

Maternal health is threatened by a high prepregnancy weight. Pre-eclampsia and gestational diabetes mellitus are more common in overweight women (body mass index (**BMI**) more than 26), and the risk of late fetal death is also increased. However, maternal obesity protects against the delivery of an infant which is small for gestational age, whereas underweight women (BMI less than 19) are more prone to have babies with lower birth weights.

Weight gain during healthy pregnancy varies between 10 and 16 kg in Western societies, i.e. about 20 per cent of prepregnancy weight. Lean, nulliparous women who eat to appetite gain 0.65 to 1.1 kg during the first 10 weeks of pregnancy, about 0.45 kg per week during the second trimester, and about 0.36 kg per week during the last trimester. Maternal weight gain correlates poorly with fetal growth. Unless the mother is underweight before pregnancy (BMI less than 19), regular antenatal measurements of maternal weight are not helpful and fetal growth is most accurately assessed by serial ultrasound measurements.

Past medical history

Pregnancy is a medical stress test for the woman, which is particularly evident in those with chronic medical disorders. A diseased maternal organ may lose residual function attempting to accommodate the physiological demands of pregnancy. Furthermore, women with severe pre-existing disease are more likely to have an adverse fetal outcome.

Pregnancy can also uncover subclinical disease; for example inherited thrombophilias may lead to thrombosis only in combination with the hypercoagulable environment of healthy pregnancy. However, the physiological changes of pregnancy are not always damaging: some conditions improve, whilst others deteriorate (Table 1).

Family history

Gestational conditions tend to run in families. Pre-eclampsia, gestational diabetes mellitus, obstetric cholestasis, and probably hyperemesis gravidarum and postnatal depression have genetic components. Inherited thrombophilias may also have a direct impact on pregnancy outcome.

Infertility and multiple pregnancies

In 1978 the birth of the first baby by *in vitro* fertilization (**IVF**) gave hope to the 15 per cent of couples who are infertile. Since then over 200 000 children have been born throughout the world using IVF, and in the United Kingdom a healthy baby is born from 17.4 per cent of IVF cycles on average. The cause of infertility may itself lead to problems in pregnancy, for example women with polycystic ovary syndrome are at increased risk of pregnancy-induced hypertension and gestational diabetes.

In most cases of IVF, two embryos are returned to the woman. As a consequence, 28.9 per cent of deliveries from IVF conceptions lead to multiple births. Following natural conception, multiple births affect only 11 per 1000 pregnancies in the United States and Europe. Women with multiple pregnancies are more vulnerable to pre-eclampsia and premature delivery.

Ovarian hyperstimulation syndrome

To increase the yield of eggs, women receiving IVF undergo ovarian stimulation with gonadotrophins, following which up to 14 per cent develop an ovarian hyperstimulation syndrome, which is severe in 1 to 2 per cent of cases. In ovarian hyperstimulation syndrome, multiple follicles develop into corpus lutea that produce excessive amounts of progesterone, resulting in massive ovarian enlargement and increased vascular permeability. Protein-rich fluid shifts into serous cavities, causing ascites, and in more severe cases pleural and pericardial effusions. The fluid shift results in haemoconcentration and hypotension, increasing the risk of thrombosis and reducing renal perfusion. Most cases of ovarian hyperstimulation syndrome are mild, but death has followed acute respiratory distress, hepatorenal failure, thromboembolism, and rupture of grossly enlarged ovaries. Management is mainly supportive, including careful fluid balance, thromboprophylaxis, analgesia, and adjustment of luteal stimulation under the guidance of a specialist in assisted conception. In some cases, paracentesis relieves pressure symptoms.

Diagnosis of pregnancy

Pregnancy can be diagnosed within a day of missing a menstrual bleed by identifying a rise in concentration of urinary human chorionic gonadotrophin. At this time the embryo is 2 weeks old, but obstetric convention dictates that the gestation of pregnancy is calculated from the first day of the last menstrual period, i.e. 2 weeks earlier than embryonic age. Teratogenic drugs interfere with organ development in the 2 to 8 weeks postconception (embryonic period). After 9 weeks and until delivery, the conceptus is known as a fetus, but it is still vulnerable to the effects of drugs given to the mother.

Screening of maternal health during pregnancy

Pregnancy is an opportunity for women to be screened for occult disease. In the United Kingdom, healthy women are encouraged to register with an antenatal clinic at 12 to 14 weeks' gestation. However, by this gestation they will have missed the opportunity to take folic acid prophylaxis against neural tube defects, and may not recognize the need to adjust social behaviour (see below) or stop regular medications.

At the first antenatal visit, a medical and obstetric history is combined with cardiovascular examination, urinalysis, and laboratory tests. Identification of maternal infection with HIV, hepatitis B, or syphilis is crucial for appropriate management of the woman and her partner, and to minimize the risk of vertical transmission to the infant. Rhesus antibody screening allows prophylactic measures to prevent haemolytic disease of the fetus.

Further antenatal checks are usually performed at 20 weeks (combined with a detailed scan of the fetus), 26 to 28 weeks (combined with a glucose tolerance test and full blood count), 30, 32, 34 (full blood count), and 36 weeks, then weekly until delivery. At each visit, obstetric assessment is combined with a check of blood pressure and urinalysis.

Screening for asymptomatic bacteriuria during healthy pregnancy and subsequent treatment reduces the risk of maternal pyelonephritis and fetal morbidity. The cost-effectiveness of such screening depends on the prevalence of asymptomatic bacteriuria in the pregnant population. If the prevalence is less than 5 per cent, as in many developed nations, screening is probably not cost-effective. However, in the developing world asymptomatic bacteriuria is far more common and screening is worthwhile. As the recurrence rate of asymptomatic bacteriuria is about 30 per cent, women identified with an occult infection should be screened monthly throughout the remainder of their pregnancy. These issues are discussed in Chapter 13.5.

Symptoms and signs of healthy pregnancy

Fatigue

Fatigue is a common symptom that often begins early in healthy pregnancy. Towards term, changes in maternal size and shape as well as nocturia cause insomnia. If daily living is significantly compromised, anaemia or hypothyroidism should be excluded.

Cardiovascular system

The hyperdynamic circulation of pregnancy causes alterations to the cardiovascular system that can mimic heart disease (see Chapter 13.6). Furthermore, palpitations, dizziness, syncope, and dyspnoea are common symptoms of healthy pregnancy. Failure to distinguish between benign physiological change and significant pathology creates unnecessary anxiety and investigations.

Clinical examination

During healthy pregnancy, the peripheral pulses are full, bounding, and often collapsing, suggesting aortic regurgitation to the untutored. From mid-gestation onwards the jugular venous pressure becomes more obvious and may be raised due to increased intra-abdominal pressure. The apex beat is more forceful and because of the

increase in cardiac output may suggest cardiomegaly in normal patients. However, if the apex beat is more than 2 cm outside the midclavicular line, this should be considered abnormal. On auscultation, an ejection systolic flow murmur can be heard in up to 90 per cent of healthy pregnant women. During the last trimester, increased mammary blood flow can produce a bruit that varies with the pressure of the stethoscope.

In developed countries it is very rare for new heart lesions to be identified during pregnancy: most women with heart disease are diagnosed early in life. By contrast, women from developing nations are more likely to present with previously unrecognized cardiac abnormalities.

Palpitations

Transient sinus tachycardia, up to 130 beats/ min, and premature atrial and ventricular ectopic beats are common features of healthy pregnancy, especially in women who complain of palpitations. As pregnancy may expose previously asymptomatic abnormalities of cardiac conducting tissue, investigations should include a 12-lead ECG. During healthy pregnancy, the QRS axis moves to the left as the diaphragm becomes elevated and Q waves and inverted T waves are frequently seen in lead III and aVR. Pregnant women with syncope or presyncope coinciding with palpitations should have a 24-h Holter monitor. Thyrotoxicosis, anaemia, hypokalaemia, excess caffeine, or tobacco should be excluded.

Oedema

By the end of pregnancy, 80 per cent of healthy women will have some degree of oedema. This is due to a fall in plasma albumin concentration of 5 to 10 g/litre and reduced venous return. Unless peripheral oedema is very severe, or is associated with pulmonary oedema, diuretics should be avoided: they attenuate the plasma volume expansion of healthy pregnancy, which can lead to restriction of fetal growth. Severe and rapid onset of oedema, especially affecting hands and face, may herald pre-eclampsia and warrants further assessment.

Blood pressure

Peripheral vasodilatation leads to a slight fall in blood pressure by the end of the first trimester, which gradually returns to non-pregnant values during the third trimester. Systolic and diastolic readings are approximately 10 mmHg higher when measured sitting or standing as compared with the left lateral position, hence blood pressure should be measured with the mother in the same position at each antenatal visit. A blood pressure reading before 20 weeks' gestation is essential to allow later discrimination between pre-existing hypertension and pregnancy-induced hypertension.

Respiratory system

Dyspnoea

The physiological hyperventilation of pregnancy leads to a subjective feeling of breathlessness in about 70 per cent of women. The maximum incidence of breathlessness is between 28 and 31 weeks' gestation, but approximately 50 per cent of women will feel breathless before 20 weeks. The early onset of dyspnoea and improvement towards term suggests that the gravid uterus has little influence on this physiological symptom. Women with gestational dyspnoea are more sensitive to CO_2 and hypoxia than asymptomatic women and respond with excessive ventilation. However, physiological dyspnoea does not usually interfere with daily activities and further investigations are only necessary if symptoms or signs suggest cardiorespiratory disease, for example chest infection, pulmonary embolus, or heart failure.

Radiological imaging in pregnancy

In general, management of pregnant women should consider the health of the mother before that of the fetus. Nowhere is this consideration ignored more than with the use of X-rays. Although ionizing radiation is a known carcinogen, there is very little—if any—increased risk of childhood cancer following prenatal exposure to X-rays. Radiation from a chest radiograph is minimal (0.02 mSv), equivalent to 3 days of background radiation. During healthy pregnancy, chest radiographs show an increased cardiothoracic ratio and pulmonary vascular markings. Pregnant women suspected of a pulmonary embolus should not be denied a ventilation-perfusion scan (1.3 mSv).

Gastrointestinal system

Nausea and vomiting

During early pregnancy, approximately 75 per cent of all healthy women will feel nauseated and up to 50 per cent will vomit. Nausea usually begins around the fifth week; by the 14th week it will have resolved in 50 per cent of women, but 10 per cent of healthy pregnant women will still feel nauseated at 22 weeks. Contrary to popular belief, nausea is rarely confined to the mornings (less than 2 per cent), but affects 80 per cent of sufferers all day. Beneficial palliative measures include rest, eating carbohydrates, and drinking carbonated drinks.

Vomiting is severe and persistent in approximately 1.5 per cent of pregnant women. This progresses to hyperemesis gravidarum when there is dehydration, weight loss, and ketonuria (see section 13.09). Ptyalism is a frequent accompaniment, due to an inability to swallow saliva. Biochemical changes often include elevated liver transaminases, elevated free T_4, and depressed thyroid-stimulating hormone. Hyperthyroxinaemia associated with hyperemesis gravidarum coincides with the rise and fall of serum human chorionic gonadotrophin, which has thyroid-stimulating activity. Treatment of hyperemesis corrects the abnormal biochemistry.

Antiemetics have not been fully evaluated in early pregnancy. The clinician must therefore balance the potential risks of teratogenesis with the risks of leaving the mother malnourished, inadequately hydrated, and vulnerable to thrombosis. Most antiemetics, including antihistamines, phenothiazines, metoclopramide, pyridoxine (vitamin B_6), and ginger have been used to treat hyperemesis with some success and without fetal harm. More severe cases have responded to steroid treatment (prednisolone 30 mg daily) or serotonin antagonists. Intravenous rehydration and occasionally parenteral nutrition are necessary. Hyperemesis gravidarum can lead to Wernicke's encephalopathy, hence thiamine (vitamin B_1) supplementation is essential.

New onset of nausea and vomiting during the second half of pregnancy suggests pathology unrelated to hyperemesis and may herald pre-eclampsia. Gastro-oesophageal reflux is a common problem of late pregnancy that usually improves with antacids or a change in diet, but persistent symptoms during pregnancy have been safely treated with H_2 receptor antagonists or proton pump inhibitors. Increased circulating progesterone levels relax intestinal smooth muscle and commonly provoke constipation. Increased dietary fibre and avoidance of unnecessary iron supplements provide symptomatic relief.

Neurological system

Headaches are common in healthy pregnancy. Many pregnant women develop migrainous type headaches for the first time in early pregnancy: if these are recurrent or do not respond to occasional paracetamol, then regular aspirin 75 mg daily or propranolol 10 to 20 mg thrice daily are good prophylactic measures. Severe, persistent headache that presents for the first time in pregnancy, or is accompanied by focal neurological signs, requires investigation with magnetic resonance imaging (see section 13.12).

Introduction of an epidural catheter during labour can lead to accidental puncture of the dura and leak of cerebrospinal fluid, causing headache that improves when lying flat. If there is no improvement within 24 h, then an injection of 2 to 3 ml of autologous blood at the site of dural puncture (blood patch) usually resolves the headache.

Carpal tunnel syndrome affects approximately 20 per cent of healthy pregnancies. It begins during the second half of pregnancy and is associated with excessive weight gain and fluid retention. Pain and numbness of the first three fingers and wrist can be severe. Wrist splints alleviate symptoms, usually making surgical intervention inappropriate, as the majority of cases recover within a few weeks of delivery.

Musculoskeletal system

Low back and pelvic pain affect approximately 50 per cent of all pregnancies. A combination of mechanical stress on the spine and pelvis and the effects of relaxin, a hormone produced by the corpus luteum to relax ligaments in anticipation of childbirth, are believed to be responsible. Some women develop radicular symptoms as the uterus presses on nerve roots and the lumbar sacral plexus, but only 1 per cent develop true sciatica with a dermatomal distribution. Progressive neurological symptoms necessitate further investigations, often with magnetic resonance imaging. Most women will benefit from massage, exercises, or a maternity cushion. Others gain relief from transcutaneous electrical nerve stimulation or a trochanteric support belt. Non-steroidal anti-inflammatory drugs should be avoided in the third trimester, and used sparingly in early pregnancy because of fetal effects.

Skin

Pruritis is a common symptom of late pregnancy, thought to be related to increased cutaneous blood flow. If there is an associated rash, then gestational skin conditions need to be considered (see Chapter 13.13). If there is no rash, then liver function should be checked to exclude obstetric cholestasis.

Supplements for a healthy pregnancy

Folic acid and multivitamins

In the United Kingdom and United States spina bifida or anencephaly (neural tube defects) affect approximately 1 in 1000 pregnancies. The neural tube develops and then closes within 28 days of conception. Women who take 400 µg folic acid daily around the time of conception and for the first 2 months of pregnancy reduce their risk of a pregnancy complicated by neural tube defects by approximately 70 per cent. Foods fortified with folic acid provide only 100 µg folic acid and natural folate-rich foods even less: these lower doses of folic acid are of no proven benefit as prophylaxis against neural tube defects. Women who have had a baby affected by spina bifida, who are taking anticonvulsants, or who have coeliac disease, require higher doses of folic acid (5 mg daily).

Multivitamin preparations without folic acid do not reduce the risk of neural tube defects. Multivitamins taken periconceptually may reduce the risk of some congenital heart defects, but beyond the first trimester are of no proven benefit for healthy women on a balanced diet.

Iron

During healthy pregnancy the haemoglobin concentration falls as plasma volume expands. A gestational fall in haemoglobin of 3 g/dl to 9.5 to 10.5 g/dl is associated with the least incidence of small babies and premature delivery. Conversely, a haemoglobin of more than 12 g/dl at the end of the second trimester is associated with a three-fold increase in pre-eclampsia and intrauterine growth restriction (both are plasma contracted states). In the developed world, supplemental iron should be reserved for those who have a haemoglobin of less than 9.5 g/dl and a mean corpuscular volume of less than 84 fl in the third trimester. In the developing world, malnutrition and chronic infection diminish iron stores that are further exhausted during pregnancy. Under these conditions, routine supplemental iron and folate may have the potential to improve maternal and neonatal outcome (see section 13.02 for further discussion).

Prophylaxis against pre-eclampsia

Pre-eclampsia affects approximately 3 to 5 per cent of healthy nulliparous women. It is a heterogeneous, multisystem disorder to which predisposed women are vulnerable in different ways. Aetiology is uncertain, but likely to be multifactorial. It is no surprise, therefore, that none of the prophylactic measures given in an attempt to reduce the incidence of pre-eclampsia have proved to be beneficial in large randomized controlled trials. Such measures have included low-dose aspirin, dietary magnesium, zinc, and calcium, antihypertensive drugs, fish oil supplements, and antioxidant vitamins. A clearer understanding of the pathogenesis of pre-eclampsia is necessary before we can expect success from prophylactic measures, and when these are available they will probably need to be individualized.

Thyroxine and iodine

During the first trimester, neurodevelopment of the fetus depends on maternal thyroxine and subclinical hypothyroidism in the mother has been associated with impaired neurodevelopment of the infant. It has therefore been suggested that all women should be screened during early pregnancy, or before conception, for hypothyroidism. The difficulties of implementing this measure are probably outweighed by more easily applied public health measures to increase iodine intake. The benefit of thyroxine replacement in women with 'low normal' thyroxine levels has not been established.

Behavioural habits during pregnancy

Exercise

Pregnancy outcome is improved by regular exercise throughout a healthy pregnancy. The gestational increases in both cardiac output and respiratory work are enhanced further by exercise. In late pregnancy, non-weight-bearing exercises such as swimming are usually preferred. Exercise may be harmful to women with impaired cardiac or respiratory function who struggle to fulfil the physiological demands of pregnancy alone.

Alcohol

Heavy alcohol consumption during pregnancy leads to the 'fetal alcohol syndrome' in approximately one-third of offspring. The susceptibility of the fetus to alcohol depends on genetic vulnerability, nutritional status of the woman, and her abuse of other drugs. The developmental and neurological abnormalities that make up the fetal alcohol syndrome affect approximately 1 to 2 per 1000 live births. Drinking one to two units of alcohol each day has not been shown to be harmful to the fetus.

Tobacco

Women should stop smoking during pregnancy as it impairs fetal growth. Nicotine gum contains less nicotine than cigarettes and none of the other toxins, making them a preferable alternative during pregnancy. Nicotine patches provide a constant release of nicotine throughout the day that exceeds that of periodic nicotine gum. Smoking is a major source of oxidant stress, but paradoxically women who smoke before and during pregnancy suffer less pre-eclampsia than non-smokers.

Caffeine

Large quantities of caffeine (more than six cups of coffee a day) increase the risk of spontaneous abortion. Moderate caffeine consumption is unlikely to be harmful.

Travel

Aircraft are pressurized to an oxygen partial pressure equivalent to that found at 8000 ft (2440 m) above sea level. During a routine commercial flight, healthy pregnant women (32 to 38 weeks' gestation) increase their heart rate and blood pressure but drop their oxygen saturation. Despite these maternal responses, fetal heart rate remains unchanged. Airlines are reluctant to carry women after 36 weeks' gestation. Long flights also increase the risk of deep vein thrombosis.

Lactation

Breastfeeding is beneficial to the infant. However, the mother who breastfeeds for 6 months or longer transiently loses 4 to 5 per cent of bone density in her lumbar spine. Calcium supplementation does not prevent this transient loss of bone mineral density, which recovers spontaneously 6 months after delivery, whether or not the mother continues to breast feed.

Postnatal depression

Almost half of all women develop the 'maternity blues'. This is characterized by tearfulness, anxiety, and irritability, starting around the third to fifth postpartum days and usually resolving with nothing more than reassurance by the tenth day. Approximately 10 per cent of women develop non-psychotic postnatal depression 4 to 6 weeks postpartum, with a maximum incidence at 3 months postpartum. The depression is similar to that occurring at other times, but is often accompanied by thoughts of harming the baby. Although most women recover without treatment over 3 to 6 months, recovery can be hastened by counselling. Women who fail to respond to counselling or who have severe depression may benefit from antidepressant treatment. Small amounts of tricyclic antidepressants and selective serotonin reuptake inhibitors appear in breast milk, but not enough to recommend stopping breastfeeding. Nonetheless, the infant should be watched for possible unwanted effects. Women who have had postpartum depression are more likely to suffer depression in later life.

Future maternal health

'Gestational syndromes' must be monitored postpartum until they resolve or reveal occult disease. For example, proteinuria related to pre-eclampsia can take up to 12 months to disappear, but may expose the 2 to 5 per cent of women with pre-eclampsia who have occult renal impairment. Similarly, abnormal liver function that does not resolve postpartum suggests non-gestational liver disease.

Insulin resistance underlying gestational diabetes mellitus resolves immediately postpartum. However, women who have had gestational diabetes mellitus have a 20 to 60 per cent risk of developing type 2 diabetes mellitus within 5 to 16 years of pregnancy. Similarly, both pregnancy-induced hypertension and pre-eclampsia, but not eclampsia, increase the mother's risk of cardiovascular disease in later life (Fig. 2).

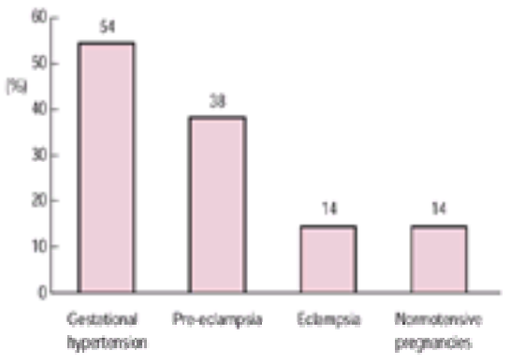

Fig. 2 Prevalence of hypertension 13.7 ± 5.1 years after pregnancy-induced hypertension (111 women), pre-eclampsia (80 women), eclampsia (14 women), and normotensive pregnancies (86 women). (Reproduced from Marin *et al*. (2000), with permission.)

Despite all of the above, most women complete an uncomplicated pregnancy. This bodes well for maternal health during subsequent pregnancies. Furthermore, those who were normotensive during pregnancy are less likely than the national average to have cardiovascular disease in later life.

Further reading

Botto LD *et al.* (1999). Neural-tube defects. *New England Journal of Medicine* **341**, 1485–90. Comprehensive review of neural tube defects.

Department of Health (1998). *Why mothers die. Report on confidential enquiries into maternal deaths in the United Kingdom 1994–1996*. Department of Health, London. Audit of maternal deaths in the United Kingdom with comments on management.

Garcia-Rio F *et al*. (1996). Regulation of breathing and perception of dyspnoea in healthy pregnant women. *Chest* **110**, 446–53. Thorough study of pattern and mechanism of dyspnoea during normal pregnancy.

Greer IA (1999). Thrombosis in pregnancy: maternal and fetal issues. *The Lancet* **353**, 1258–65. Review of management of thromboembolic disease in pregnancy.

Haddow JE *et al*. (1999). Maternal thyroid deficiency during pregnancy and subsequent neuropsychological development of the child. *New England Journal of Medicine* **341**, 549–55.

Human Fertilization and Embryology Authority (1999). *Eighth annual report*. HMSO, London.

James DK *et al.*, eds (1999). *High risk pregnancy*, 2nd edn. WB Saunders, London. Comprehensive review of management of normal and abnormal pregnancies.

Kalkwarf HJ *et al*. (1997). The effect of calcium supplementation on bone density during lactation and after weaning. *New England Journal of Medicine* **337**, 523–8.

Kjos SL, Buchanan TA (1999). Current concepts: gestational diabetes mellitus. *New England Journal of Medicine* **341**, 1749–56.

Lacroix R, Eason E, Melzack R (2000). Nausea and vomiting during pregnancy: a prospective study of its frequency, intensity and patterns of change. *American Journal of Obstetrics and Gynecology* **182**, 931–7.

Liljestrand J (1999). Reducing perinatal and maternal mortality in the world: the major challenges. *British Journal of Obstetrics and Gynaecology* **106**, 877–80. Commentary on global problem of perinatal and maternal mortality.

Marin R *et al*. (2000). Long-term prognosis of hypertension in pregnancy. *Hypertension in Pregnancy* **19**, 199–209. Thirteen year follow-up of women with hypertension in pregnancy.

Steer PJ (2000). Maternal hemoglobin concentration and birth weight. *American Journal of Clinical Nutrition* **71**, 1285S–1287S. Review of relationship between maternal haemoglobin and clinical outcome.

Whelan JG 3rd, Vlahos NF (2000). The ovarian hyperstimulation syndrome. *Fertility and Sterility* **73**, 883–96.

13.4 Hypertension in pregnancy

C. W. G. Redman

Cardiovascular changes in pregnancy

Cardiac output increases during the first trimester to about 1.5 l/min above the levels of non-pregnant women. No further increase occurs in the second and third trimesters. Towards full term, it declines in the supine but not lateral recumbent position, owing to the pressure of the gravid uterus on the inferior vena cava, which reduces venous return to the heart. In the third trimester, about two-thirds of the additional cardiac output is distributed to the placental circulation and to augment renal plasma flow. The increased output is the result of both a greater stroke volume and a higher pulse rate. Plasma volume increases progressively during the second and third trimesters and is significantly correlated with the birthweight of the conceptus, being higher in multiple pregnancies. Arterial pressure falls in the second half of the first trimester at about the same time as the cardiac output is increasing, meaning that peripheral resistance decreases relatively more than the cardiac output increases. The uteroplacental circulation is too small at this time to cause these changes, which must therefore result from a generalized arteriolar dilatation.

In the later weeks of pregnancy, there is a tendency for the diastolic pressure to rise slowly towards what it was before pregnancy began, the systolic pressure remaining more or less unchanged. However, in the supine position, with vena caval compression and reduced venous return, the arterial pressure may be atypically low with a narrowed pulse pressure and reflex vasoconstriction. The fall in systolic pressure may exceed 30 per cent in 10 per cent of cases and cause 'the supine hypotension syndrome' —evident as restlessness, faintness, hyperpnoea, and pallor.

Hypertension in pregnancy: definition, causes, and terminology

Definition

The average blood pressure during the first half of the second trimester is about 120/70; 140/85 and 160/95 correspond to two and three standard deviations above the mean, respectively. Hypertension in obstetric practice is conventionally recognized at or above an arbitrary threshold of 140/90. This is appropriate for the first half of pregnancy, but in the second half about one-quarter of all women will be hypertensive by this criterion, meaning that these limits are too low to define an unusual group that merits extra clinical attention. About 2.5 per cent have a maximum arterial pressure of 160/105 or more and about 1 per cent of 170/110 or more; these are more relevant limits for identifying third trimester hypertension.

Causes of hypertension in pregnancy

Hypertension in pregnancy has three possible aetiologies. Most important, it may be caused by the pregnancy as part of the syndrome of pre-eclampsia, a specific disorder of pregnancy that is common, dangerous, and poorly understood. Second, it may represent chronic hypertension, a long-term attribute of the woman. In some women chronic hypertension may be revealed for the first time during pregnancy—typically towards the end; but the condition is of the woman not of her pregnancy. Third, and much more rarely, it may be a new medical condition coinciding with pregnancy by chance.

Pre-eclampsia is a common syndrome that becomes evident in the second half of pregnancy (although its origins may lie in the first half). It is defined in terms of the transient development of new hypertension and proteinuria, these may be severe, but regress after delivery.

Terminology

Eclampsia is characterized by grand-mal convulsions. Pre-eclampsia (previously called pre-eclamptic toxaemia (PET)) is so called because it may precede eclampsia, as well as a number of other possible crises. These other crises (described below) are just as dangerous as eclampsia and occur as commonly or even more so. Not all cases of eclampsia are preceded by a prodromal illness of pre-eclampsia, so the terminology is altogether too simplistic to describe the events that can occur. Toxaemia is an obsolete expression, previously used to describe any hypertension or proteinuria in pregnancy, whether pregnancy-induced or not.

Pregnancy-induced hypertension (PIH), transient hypertension of pregnancy, or gestational hypertension are terms used to describe new hypertension, which appears after mid-term (20 weeks) and resolves after delivery. They therefore describe one of the components of the pre-eclampsia syndrome. For reasons that are historical rather than logical, and which are to some extent arbitrary, PIH is deemed to be a mandatory part of pre-eclampsia. However, PIH on its own (a common clinical presentation) is not pre-eclampsia; at least one more sign is required to make the diagnosis. The clusters of clinical features that comprise any syndrome are chosen for convenience; they describe outward appearances and embody no special truth about the underlying disease or diseases. When a syndrome such as pre-eclampsia is 'defined', rules are set that bring consistency to what is being discussed. The rules may be sensible or not, but their validity cannot be tested because there is no standard to which to refer. All the definitions of pre-eclampsia suffer from these limitations and none can be said to be the best. The conventional components of the cluster are PIH combined with new proteinuria that regresses after delivery.

Almost all hypertension presenting before mid-term (gestational age of 20 weeks) indicates pre-existing or chronic hypertension, the rare exceptions being women with atypical very early onset pre-eclampsia. However, normotension in the first half of pregnancy does not necessarily mean long-term normotension: the fall in blood pressure induced in early pregnancy may be exaggerated in some women and many with relatively severe hypertension may have normal blood pressures (without treatment) by 12 weeks; indeed, in one study as many as 60 per cent of women with chronic hypertension defined before pregnancy were normotensive by the end of the first trimester. In other words, some women enjoy the benefits of pregnancy-induced normotension just as others suffer the disadvantages of pregnancy-induced hypertension. Pregnancy-induced normotension tends to be lost in the third trimester. If the prepregnancy blood pressures are unknown then this may be misinterpreted as pregnancy-induced hypertension rather than recognized for what it is, namely re-establishment of the normal, long-term blood pressure.

Pregnancy-induced hypertension thus represents at least two clinical situations; early pre-eclampsia or occult, chronic hypertension. In many cases, the signs of pre-eclampsia are not confirmed, but nevertheless the blood pressure reverts to normal after delivery. It is possible that these cases represent the very early stages of pre-eclampsia, an alternative being that an innate tendency to hypertension has been revealed in pregnancy but will become overt only many years later. The studies

have not been done to confirm or refute this suggestion.

Pre-eclampsia

Pre-eclampsia becomes evident in the second half of pregnancy, during labour or even, for the first time and without apparent preceding problems, in the immediate puerperium, but it always resolves more remotely after delivery. It is common, can be dangerous to both mother and baby, and of unknown cause.

A typical definition of pre-eclampsia is given in Table 1. This and other current definitions require that hypertension and proteinuria, both pregnancy-induced, should be present before the syndrome is recognized. It is now acknowledged that the disorder originates in the placenta, meaning that pre-eclampsia is probably the most common form of secondary hypertension in clinical practice. It is increasingly evident that there are related syndromes, characteristic of the end of pregnancy, which share the same placental causes but do not necessarily provoke hypertension. Hence, although hypertension is conventionally required as a defining feature of pre-eclampsia, it is better considered to be one of several useful signs, but not a central part of the pathology which is more extensive and can involve the maternal liver, clotting, and nervous systems.

The incidence of pre-eclampsia depends on how it is defined and how assiduously the signs are sought. It is possible only to estimate the size of the problem; in the United Kingdom the incidence is of the order of one in 20 to 30 maternities.

Some of the factors that affect susceptibility are listed in Table 2; these include fetal-specific as well as maternal-specific components. Primigravidae are several times more prone to the condition. In parous women, pre-eclampsia particularly affects those who have had the problem before. The predisposition to pre-eclampsia is in part familial, probably genetic, but the pattern of inheritance is not clear. Other factors must also be relevant because pre-eclampsia does not affect identical twin sisters concordantly.

Certain medical problems predispose to pre-eclampsia, including some (chronic hypertension, renal disease) that can mimic the disorder. Superimposed pre-eclampsia refers to the mixed syndrome comprising pre-eclampsia in an individual with pre-existing hypertension or renal disease. In the absence of a specific diagnostic test for pre-eclampsia it can be difficult or impossible to disentangle what elements of proteinuric hypertension are caused by a chronic medical problem or which arise from superimposed pre-eclampsia, and the conventional definitions of pre-eclampsia cease to apply. If a woman is permanently proteinuric there are, for example, no accepted criteria for diagnosing 'proteinuric pre-eclampsia'. Nevertheless, it seems that chronically hypertensive women are three to seven times more likely to develop higher blood pressures and proteinuria ('superimposed pre-eclampsia') than normotensive women. Women with hypertension associated with chronic renal disease are particularly susceptible.

Aetiology and pathogenesis of pre-eclampsia

The primary pathology of pre-eclampsia is not known for certain, but the presence of a placenta is both necessary and sufficient to cause the disorder. A fetus is not required as pre-eclampsia can occur with hydatidiform mole. A uterus is probably not required because pre-eclampsia may develop with abdominal pregnancy (that is an ectopic pregnancy in the peritoneal cavity). Central to management is delivery, which removes the causative organ, namely the placenta. The primary involvement of the placenta explains why pre-eclampsia is associated with two syndromes not one; the fetal syndrome of nutritional and respiratory deprivation can be as important a part of the illness as the maternal syndrome, or even more so.

The placental problem appears to be a relative ischaemia secondary to deficiencies in the uteroplacental circulation or to an excessively large placenta (with multiple pregnancies for example). The uteroplacental circulation may be compromised by two lesions involving the spiral arteries, which are the end-arteries supplying the intervillous space. The first, poor placentation, is a partial lack of the structural modifications of the spiral arteries that occur between weeks 8 and 18 (before there is clinical evidence of pre-eclampsia), when the arteries become dilated in preparation for the hugely expanded uteroplacental blood flow of the second half of the pregnancy. The second is 'acute atherosis', when aggregates of fibrin, platelets, and lipid-loaded macrophages (lipophages) partially or completely block the ends of the arteries. Neither change is specific to pre-eclampsia; they can also occur with intrauterine growth retardation without a maternal syndrome. Hence the spiral artery changes may be only an associated, but not primary, feature of pre-eclampsia. The relationship of pre-eclampsia with processes that depend on placentation could mean that it originates much earlier than when the maternal syndrome becomes overt. Once pre-eclampsia is established, uteroplacental blood flow is reduced. There is no direct evidence that placental ischaemia can cause pre-eclampsia but in various animal models impeding the placental blood supply can induce a pre-eclamptic-like illness.

The secondary pathology of pre-eclampsia includes all the features of the maternal syndrome, short of decompensation. The maternal syndrome is typically variable in the time of onset, speed of progression, and the extent to which it involves different systems including arterial, coagulation, renal, central nervous, and hepatic. Until recently it was impossible to explain how a single pathological process might cause not only hypertension but also convulsions, disseminated intravascular coagulation, jaundice, hepatic dysfunction, or normotensive proteinuria (among others). However, the concept that the maternal endothelium is the target organ for the pre-eclampsia process has resolved this difficulty. In short, the maternal syndrome is not primarily a hypertensive problem, but the sum of the consequences of diffuse endothelial dysfunction, causing widespread circulatory disturbances in different organ systems as well as generalized arterial and coagulation abnormalities. Latterly, it has been shown that the endothelial dysfunction is one aspect of a more generalized, systemic maternal inflammatory response that also affects circulating leukocytes and other components of the inflammatory system (for example the clotting system). Moreover, this increased inflammatory response is well established in the third trimester of normal pregnancy, when it is not intrinsically different from that in pre-eclampsia except that it is milder. It is thought that pre-eclampsia develops when the pregnancy-induced, systemic inflammatory response causes one or other maternal system to decompensate. In other words, the disorder is not a separate condition but simply the extreme end of a range or continuum of maternal, systemic inflammatory responses engendered by pregnancy itself. This concept has profound implications for clinical practice. If true, it is unlikely that there ever will be a single cause, single diagnostic test, or single preventive measure for pre-eclampsia.

Under certain circumstances the secondary disturbances of pre-eclampsia can become so severe that they cause decompensation, and the tertiary pathology is what makes the condition so dangerous for the mother and baby. It leads to a number of crises that are listed in Table 3.

Clinical characteristics of pre-eclampsia

Usually, hypertension precedes proteinuria, although the converse can happen. Pre-eclampsia is more variable than is generally appreciated—in the time of onset for example. Thus, although pre-eclampsia is defined as presenting after 20 weeks, it may occur earlier or, at the other extreme, become evident only after delivery. The speed with which it progresses and how it involves different maternal systems is also variable.

The hypertension of pre-eclampsia appears to be caused by an increased peripheral resistance secondary to generalized, maternal endothelial dysfunction. There is no single haemodynamic pattern, both increased and decreased cardiac output having been reported. Some of the differences between studies may reflect drug use, for example treatment with vasodilators stimulates cardiac output by reducing afterload. The blood pressure is typically unstable at rest, possibly owing to reduced baroceptor sensitivity. Circadian variation is altered with, first, a loss of the normal fall in blood pressure at night then, in the worst cases, a reversed pattern with the highest readings during sleep.

Pre-eclampsia may cause arterial pressures that are well above the level (i.e. a mean pressure of about 140 mmHg) at which arterial and arteriolar damage would be expected. It is not, therefore, surprising that an important cause of maternal death from pre-eclampsia and eclampsia is cerebral haemorrhage, the pathology of which is similar to that seen in other hypertensive states. As far as it is known, cerebral haemorrhage is the only consequence of pre-eclampsia likely to be affected by antihypertensive treatment.

The involvement of the kidneys in pre-eclampsia has long been recognized. Proteinuria usually develops after the onset of hypertension, although in 10 per cent of cases it is detected first. The proteinuria is moderately selective, increases until delivery, and not uncommonly exceeds 10 g/24 h, pre-eclampsia being the commonest cause of nephrotic syndrome in pregnancy. It is associated with impaired glomerular perfusion and filtration, both reflected in a reduced creatinine clearance and increased plasma creatinine and urea concentrations. The typical renal glomerular lesion of pre-eclampsia is glomerular endotheliosis, when the endothelial cells of the glomeruli swell and block the capillary lumina so that the glomeruli appear enlarged and blood-less. The lesion, which represents direct histological confirmation of endothelial damage in pre-eclampsia, has been defined in research investigations; renal biopsy is never indicated for clinical management. Hyperuricaemia, resulting from a reduced renal urate clearance, is often an early feature of pre-eclampsia, preceding proteinuria and is useful for diagnosis at that stage. However, it is not consistently present so that its absence does not exclude the condition. It tends to be associated with hypocalciuria, another

early change in renal function. As the plasma urate rises, the plasma concentrations of urea and creatinine at first remain steady, tending to increase slowly after proteinuria has become established. The tertiary pathology of renal involvement in pre-eclampsia is acute renal failure arising from either tubular or cortical necrosis.

Generalized oedema is an inconsistent feature. It may develop suddenly and is associated with accelerated weight gain. Ascites is not uncommon with severe disease. Laryngeal oedema can cause respiratory obstruction and difficulties with intubation when general anaesthesia is required. Pulmonary oedema is a dangerous complication. In association with modern methods of ventilatory support it may progress to the adult respiratory distress syndrome, which is increasingly a cause of maternal death in this condition.

The clotting system is often, but not invariably, disturbed in pre-eclampsia, with accelerated intravascular generation of thrombin and parallel reductions in the platelet count ascribed to increased consumption. The time course is variable, but a fall in the platelet count may be a relatively early sign—antedating proteinuria for example. However, even when eclampsia supervenes, the majority of women have normal platelet counts at the time of presentation. The coagulation disturbances may decompensate to give overt disseminated intravascular coagulation (DIC). A further complication is microangiopathic haemolysis that may cause a sudden drop in haemoglobin associated with haemoglobinuria, fragmented or distorted red cells (schistocytes) on the peripheral blood film, and reduced serum haptoglobin concentrations.

The severe clotting abnormalities of pre-eclampsia, particularly DIC, are often associated with liver pathology, long recognized as an important and dangerous component of the disorder. When there is also the associated complication of haemolysis, the acronym HELLP syndrome has been used to label the concurrence of haemolysis, elevated liver enzymes, and low platelet counts. This is often not associated with marked hypertension or other conventional indices of severe pre-eclampsia. Indeed, liver damage and low platelet counts have been observed in primigravidae without hypertension or proteinuria but with the typical hepatic histology of pre-eclampsia, including fibrin deposition in the sinusoids. Epigastric pain and vomiting are the typical symptoms of the HELLP syndrome, which may present so suddenly as to be misinterpreted as biliary colic or other surgical emergencies. Hepatic tenderness and raised serum liver enzymes are the signs. Serum bilirubin is usually normal, but jaundice is possible and may be a presenting feature. In certain severe cases, typically of multiparae rather than primiparae, there may be bleeding under the liver capsule which may rupture to cause massive haemoperitoneum, shock, and (usually) maternal death. These issues are discussed in Chapter 13.9.

Eclampsia is the most dramatic evidence of involvement of the nervous system. It resembles other forms of hypertensive encephalopathy, having similar symptoms and cerebral pathology. One of the complications of hypertensive encephalopathy is cortical blindness, a feature of severe pre-eclampsia and eclampsia as well. Average blood pressures in eclampsia are high (170–195/110–120), but cases with much lower blood pressures are not as rare as with non-obstetric forms of hypertensive encephalopathy. Eclampsia is not associated with gross papilloedema or retinopathy. Ten per cent of cases of eclampsia are totally unheralded, that is without a warning prodrome of hypertension and proteinuria. The name hypertensive encephalopathy is misleading in that it suggests that the syndrome is caused by hypertension. The hypertension is an associated feature; there is no good evidence that hypertension causes eclampsia or other forms of hypertensive encephalopathy, and none that adequate medical control of the blood pressure prevents eclampsia. It is now generally agreed that eclampsia results from acute cerebral circulatory disturbances secondary to endothelial dysfunction. Cerebral vasospasm with focal ischaemia and oedema are major features that have been demonstrated by magnetic resonance imaging or CT scanning.

Diagnosis of pre-eclampsia

Pre-eclampsia is usually symptomless, hence its detection depends on signs or investigations. Nonetheless, one symptom is crucially important because it is so often misinterpreted. The epigastric pain, which reflects hepatic involvement and is typical of the HELLP syndrome, may easily be confused with heartburn, a very common problem of pregnancy. However, it is not burning in quality, does not spread upwards towards the throat, is associated with hepatic tenderness, may radiate through to the back, and is not relieved by giving antacids. It is often very severe, described by sufferers as the worst pain that they have ever experienced. Affected women are not uncommonly referred to general surgeons as suffering from an acute abdomen, for example acute cholecystitis.

In general, none of the signs of pre-eclampsia is specific; even convulsions in pregnancy are more likely to have causes other than eclampsia in modern practice. Diagnosis, therefore, depends on finding a coincidence of several pre-eclamptic features, the final proof being their regression after delivery. There are two ways to make the diagnosis. For research purposes the rules have to be followed that require the presence of both pregnancy-induced hypertension (PIH) and pregnancy-induced proteinuria. However, this is too restrictive for clinical practice, where there are presentations with the broad attributes of pre-eclampsia that do not fit these strict definitions; clinicians need to take a broader view and accept a wider range of combinations of the possible features of the syndrome, some of which are listed in Table 4. As with all syndromes, the more of the features that are clustered together the more certain is the diagnosis, but the absence of any one feature does not exclude the diagnosis. For example eclampsia can occur without proteinuria, and even hypertension seems not to be an essential component. To diagnose pre-eclampsia superimposed on long-term hypertension or renal disease there are, as stated already, no clear rules. In these circumstances, the diagnosis has to be made intuitively by judging the exacerbation of the long-term hypertension or proteinuria, in association with the appearance of other associated signs.

In practical terms, hypertension, proteinuria, and excessive weight gain have to be the signs of interest for screening in routine antenatal clinics. Different definitions have been proposed as to what constitutes hypertension, but the details are less important than the principle of an increment from a recording taken in the first half of pregnancy, which establishes the existence of pregnancy-induced hypertension (PIH). Between weeks 20 and 30 the blood pressure is normally steady so that even a small, consistent rise is clinically important. Between week 30 and term the diastolic will rise by about 10 mmHg on average. A sustained rise of at least 25 mmHg to a threshold of 90 mmHg or more is typical of pre-eclampsia. However, these are only guidelines: there is no clinical situation where rigid interpretation of the blood pressure is helpful.

The same applies to other measurements such as changes in the plasma urate. As a rough guide, abnormal levels are in excess of 0.30, 0.35, 0.40, and 0.45 mmol/l at 28, 32, 36, and 40 weeks respectively; or, if a baseline taken before 20 weeks is available, then increases of 0.10, 0.15, 0.20, and 0.25 mmol/l at 28, 32, 36, and 40 weeks respectively.

Proteinuria and evidence of reduced glomerular filtration rate are later signs. The changes in the measurements of renal function are usually within the normal range for non-pregnant individuals. In general, abnormal concentrations of plasma creatinine and urea are above 100 µmol/l and 6.0 mmol respectively. The proteinuria of pre-eclampsia ranges from 0.5 to 15 g/24 h depending on the individual and the stage of evolution of the disorder. In terms of stick testing, 0.5 g/24 h corresponds to at least + in every specimen of urine tested, and when this point is reached the disease can be said to have entered its proteinuric phase.

Thrombocytopenia ($<100 \times 10^9$/l) and increased plasma fibrin/fibrinogen degradation products (or specific fragments thereof such as the D-dimer) tend to be late developments, if they occur at all. The same is true for raised liver enzymes. In regard to the latter, it should be noted that plasma alkaline phosphatase is always elevated in late pregnancy because of the contribution from the placental isoenzyme, hence its measurement is not a useful guide to hepatic function. Serum bilirubin is rarely abnormal. Gamma-glutamyl transferase is increased only late in the evolution of the HELLP syndrome. The best simple tests are plasma aspartate amino transferase or lactate dehydrogenase.

New hypertension and the *de novo* occurrence of one other sign allows the diagnosis to be made with reasonable certainty, but PIH on its own is not pre-eclampsia, although the term is commonly, but wrongly, used to mean mild or early pre-eclampsia. It is true that PIH maybe the first indication of the onset of pre-eclampsia, but until other signs appear this remains unconfirmed. Often spontaneous or induced delivery prevents further developments so that a final certain diagnosis cannot be made.

Complications of pre-eclampsia

Complications of pre-eclampsia are listed in Table 3. Eclampsia complicates 1 in 2000 maternities in the United Kingdom and carries a maternal mortality of 2 per cent. The HELLP syndrome is commoner, probably about 1 in 500 maternities, but may be as dangerous as eclampsia itself. These two major maternal crises can present unheralded by prodromal signs of pre-eclampsia. In other words, our terminology is not exact in the sense that eclampsia can precede pre-eclampsia. Antepartum eclampsia is likely to occur earlier in gestation and is more dangerous than that presenting in labour or after delivery, and preterm disease is generally more dangerous than that at term (Table 5). Most postpartum crises develop in the first 12 h after delivery, but later is possible and eclampsia has been documented as late as 22 days after delivery.

Cerebral haemorrhage is a lesion that can kill women with pre-eclampsia or eclampsia. In that cerebral haemorrhage is a known complication of severe hypertension in other contexts, it must be assumed that this is a major predisposing factor in this situation, although this has not been proved. Adult respiratory distress syndrome

appears to have become more common, it is not known whether this is a consequence of modern methods of respiratory support rather than of the disease itself.

Prevention of pre-eclampsia

All the evidence is that, once it becomes overt, pre-eclampsia cannot be reversed except by delivery. Reliable methods of primary prevention are therefore needed, but none that is completely effective is known. If the concepts of pathogenesis described previously are correct, it is unlikely that any single measure will be effective for all susceptible women, but specific measures of reducing the susceptibilities among subgroups of women may be identified. Measures that may be effective or are definitely not effective are summarized in Table 6.

There is no evidence that blood pressure control attenuates the progression of early pre-eclampsia, nor that it prevents superimposition of pre-eclampsia in chronically hypertensive women who otherwise are more susceptible to the disorder. The only clear advantage of antihypertensive treatment is where the hypertension is so severe that delivery is essential to preserve maternal safety. At early gestational ages antihypertensive treatment can allow prolongation of pregnancy in this context, the benefit not being from prevention but palliation. However, the extent of the presumed benefit has not been measured because severe hypertension is a reason for exclusion from randomized trials of treatment, hence in all contexts antihypertensive treatment helps to protect the mother from the consequences of her problem, but not from the problem itself.

Trials of antiplatelet agents, in particular low doses of aspirin, have given mixed results. Those which have shown antiplatelet therapy to be ineffective have tended to be the largest with the least selective recruitment. There may be a modest effect in preventing or delaying the maternal syndrome, if low-dose aspirin is started early enough, well before the onset of signs. The benefits appear to be greatest in preventing early-onset pre-eclampsia (which is relatively rare) and least in preventing the disorder presenting at term (which is common). Antiplatelet therapy does not benefit women if started after the signs of pre-eclampsia have appeared. As yet there has been no clear demonstration that perinatal survival is improved. Low dose aspirin in pregnancy seems to be safe, but there may be a slight increase in maternal bleeding problems around the time of delivery. No adverse effect on the fetus has yet been identified.

At the time of writing there is one small trial suggesting benefit from the prophylactic use of antioxident vitamins C and E started in midgestation in carefully selected groups of at-risk women. It is too soon to know if this is likely to be confirmed as a safe and effective preventive regimen.

Management of pre-eclamptic hypertension

Control of the blood pressure is only a part of patient management. The definitive treatment is always delivery, which removes the cause of the problem, that is the placenta. If the affected woman can be delivered before irreversible damage has occurred (for example cerebral haemorrhage) a complete and rapid recovery is assured. Hence the purpose of medical management is to protect the mother from the dangers of her illness during the relatively brief interval after the disease is diagnosed and before elective delivery. The main objective is to prevent extreme hypertension. The threshold at which antihypertensive treatment should be started is a matter of opinion, a conservative criterion being to begin treatment if maximum readings (systolic or diastolic) repeatedly reach or exceed 170 or 110 mmHg, respectively.

Hydralazine has been the preferred antihypertensive agent for the treatment of acute severe pre-eclampsia, given intravenously by either continuous infusion (5–10 mg/h) or intermittent boluses (of 5 mg), or by intramuscular or subcutaneous injections (of 5–10 mg). After intravenous administration there is a delay of about 20 to 30 min in the onset of action and its effect is relatively short-lived, lasting 2 to 3 h. Side effects are common and include reflex tachycardia, anxiety, restlessness, hyper-reflexia, and severe headaches. These symptoms and signs may affect 50 per cent of women and simulate the features of impending eclampsia, when the symptoms of the disease cannot be disentangled from those caused by the treatment. Labetalol, a combined a and b-adrenergic blocking agent that can be given intravenously, lowers the blood pressure smoothly but rapidly without the tachycardia characteristic of treatment with hydralazine. A typical regimen starts with 20 mg/h, which is doubled every 30 min until control has been gained, but there are no adequate trials of its parenteral use in pregnancy to show how it might affect perinatal outcome. Sodium nitroprusside and nitroglycerine are rapidly acting vasodilators that have been used to manage hypertensive emergencies in pregnancy, but both should be reserved for use by specialists, usually in the context of intensive cardiovascular monitoring. The danger is of overdose with problems associated with extreme and sudden hypotension.

The calcium channel blocking agent, nifedipine, is an effective vasodilator that acts rapidly when given by mouth. Nifedipine capsules, which act too abruptly (within 10–15 min), should not be used. The slow-release tablets have a slower onset of action (about 60 min) but a more prolonged effect; whereas a long-acting preparation, formulated for once a day administration, is less convenient for acute control of pre-eclamptic hypertension. Nifedipine is at least as safe as hydralazine to use in pregnancy and is less likely to cause troublesome tachycardia. In theory, nifedipine could interact with parenteral magnesium sulphate given to prevent or treat eclampsia because the magnesium ion inhibits calcium channels; in practice there has been a report of two cases of profound hypotension in this context. Nimodipine, another calcium channel blocker but with a selective effect on the cerebral circulation, may have particular advantages for treating cerebral ischaemia in eclamptic women.

P>Diuretics are avoided because they exacerbate the hypovolaemia of pre-eclampsia, which is often severe, but they are indicated if complications such as pulmonary or laryngeal oedema occur.

Good blood pressure control in pre-eclampsia does not ameliorate its other features; the disease persists and remains relentlessly progressive until delivery. Escape from control is common. Adequate treatment does not prevent other complications such as eclampsia, the HELLP syndrome, abruption, or progressive fetal respiratory impairment. A persisting inability to control maternal arterial pressure is one of several indications for immediate delivery.

Longer-term control of pre-eclamptic hypertension

The control of pre-eclamptic hypertension must always be extended for a few days at the least, and frequently for longer. Therefore, once the blood pressure has been controlled acutely, the effects of treatment need to be prolonged. The requirements are for a drug that is safe in pregnancy, has an onset of action in 6 to 12 h, allows some titration of effect, and can safely be combined with a second drug if needed. The choice lies between methyldopa and various b adrenergic blocking agents (b-blockers).

In adequate doses, methyldopa can control the blood pressure within 6 to 12 h: a loading dose of 500 to 1000 mg is followed by 250 to 750 mg four times a day. Sedation is the rule for the first 48 h and thereafter tiredness is common. Postural hypotension is rarely a problem in the antenatal patient. Although b-blockers cause fewer subjective side-effects, their safety in pregnancy has not been so exhaustively investigated. A preparation such as atenolol, with its slow onset of action and flat dose–response curve, is not ideal for the day to day titration of blood pressure control, but its short-term safety for the fetus and neonate has been adequately demonstrated. Oxprenolol and labetalol are faster-acting alternatives; which agent is preferred depends on the clinician's familiarity with their use.

Recent evidence suggests that long-term lowering of the blood pressure has a modest but statistically significant effect in reducing the baby's birthweight. Whether this has any long-term implications is not known. However, it is a good reason for using antihypertensive agents parsimoniously and only where there is clear evidence of a maternal risk.

Prevention of eclamptic convulsions

Eclampsia is probably caused by focal cerebral vasoconstriction and ischaemia secondary to endothelial damage and therefore is neither the result of hypertension, nor prevented by antihypertensive treatment. The best mode of prevention is well-timed delivery. It is debated whether any prophylactic anticonvulsant medication needs to be offered routinely in all cases of advanced pre-eclampsia and, if so, what it should be. Intravenous diazepam (5 mg by slow intravenous injection) is preferred to stop eclamptic convulsions, although most are self limiting. Thereafter, it is reasonable to use medication to prevent recurrent convulsions. Agents that improve cerebral perfusion are more effective than those that suppress neuronal excitability. In the former category is parenteral magnesium sulphate, widely used in the United States to prevent or treat eclampsia; in the latter is phenytoin. There is now clear evidence from a large, double-blind, controlled trial that magnesium sulphate administration is superior.

Chronic hypertension complicating pregnancy

Pregnant women with chronic hypertension tend to be older, fatter, slightly taller, and frequently with clear family histories of hypertension. Owing to the physiological

changes of pregnancy, their hypertension may be ameliorated or masked by the beginning of the second trimester so that the diagnosis is missed unless prepregnancy blood pressure readings are available. As explained above, the blood pressure tends to climb back to the levels that characterize the non-pregnant state towards the end of the third trimester. If these are high this normal change can be misinterpreted as pre-eclampsia, the distinction being that the blood pressure fails to settle after delivery.

Pre-eclampsia superimposed on chronic hypertension tends to be more severe, to occur at earlier stages of pregnancy, to cause more fetal growth retardation, and to be recurrent in later pregnancies. Pre-eclampsia occurring in normotensive women tends not to recur. If a blood pressure of 140/90 in the first half pregnancy is taken as evidence of chronic hypertension, then affected individuals have an approximately five-fold increased risk of later pre-eclampsia compared to normotensive women. This close link between the two conditions led earlier clinicians to conclude that chronic hypertension is extremely dangerous when combined with pregnancy. It is now clear that the particular risks of chronic hypertension are entirely attributable to the increased chance of developing superimposed pre-eclampsia. The majority of chronically hypertensive women who do not get pre-eclampsia can expect a normal and uncomplicated perinatal outcome. In other words, the dangers of chronic hypertension in pregnancy have been over-emphasized.

Chronic hypertension can only be diagnosed with certainty during pregnancy on the basis of readings taken in the first half, preferably before 16 weeks of gestation. Without the benefit of such readings, hypertension in the second half of pregnancy cannot be interpretated because the possibility that it may represent pre-eclampsia cannot be excluded. The signs of pre-eclampsia in chronically hypertensive women are the same as in other women, except that the blood pressure increases from a higher baseline. There may be progressive hyperuricaemia, abnormal activation of the clotting system, or new proteinuria.

Treatment of chronic hypertension in pregnancy

If antihypertensive treatment has been started before conception, the patient may seek advice about the possible effects of her medication on the growth and development of her fetus. None of the commonly used antihypertensive drugs is known to be teratogenic, but this does not preclude the possibility of subtle problems that are, as yet, unknown. For this reason it is appropriate that women with no more than moderate hypertension stop treatment before conception. By the 12th week of pregnancy, the normal fall in blood pressure is such that treatment may no longer be needed, at least until the beginning of the third trimester. Although angiotensin converting enzyme (ACE) inhibitors are often considered to be teratogenic, this is not the case; there is clear evidence that they are fetotoxic (causing growth restriction, oligohydramnios, and intrauterine and postnatal renal failure in the second and third trimester) but none that they are teratogenic. After 3 months of pregnancy they are contraindicated but not before.

If chronic hypertension is diagnosed for the first time in pregnancy, it is necessary to treat those in whom it presents an immediate (as opposed to a long-term) hazard. The precise levels at which this is necessary is a matter of opinion not fact; we take a cut-off point at 170/110 mmHg. In general medical practice, the purpose of treating less severe chronic hypertension (that is 140–169/90–109 mmHg) is to prevent long-term complications such as heart failure, aortic dissection, or coronary and cerebral vascular disease. These problems are so rare in pregnant women that in themselves they cannot justify treatment for the brief period of pregnancy. Thus, moderate hypertension *per se* carries no intrinsic maternal risk over the brief period of 9 months, except insofar as it may be the precursor of more severe hypertension. However, the higher the arterial pressure the greater the eventual perinatal mortality. The risks evolve through simple progression if the mild hypertension indicates early pre-eclampsia. By contrast, if the mild hypertension indicates a pre-existing problem, then the risk is of later superimposition of pre-eclampsia which is, as stated above, several times more likely in chronically hypertensive women. Antihypertensive treatment can only be useful if either it halts the progression of mild pre-eclampsia or prevents the superimposition of pre-eclampsia in women with long-term hypertension. There is no evidence in support of either possibility, indeed there is good evidence that control of moderate, long-term hypertension does not prevent superimposed pre-eclampsia but does cause mild fetal growth retardation. Thus, there are neither clear fetal nor maternal indications for treating moderate hypertension in pregnancy.

Oral antihypertensive agents that are used in pregnancy

The choice of oral antihypertensive agents in pregnancy is dictated by considerations of fetal safety (Table 7). Methyl dopa is the preferred agent because its fetal effects have been defined more clearly than those of other agents. Its antihypertensive action and side-effects are the same as in non-pregnant individuals. The usual treatment schedule is 1.0 to 3.0 g/day in divided doses. It can be supplemented by nifedipine.

Labetalol is a popular alternative. However, long-term b-adrenergic blockade extending throughout the second and third trimesters has been associated with significant fetal growth retardation and for this reason should be avoided. ACE inhibitors are contraindicated in the second and third trimesters.

In the unlikely event that diuretics are essential for good blood pressure control, they can be continued throughout pregnancy, but their use carries certain disadvantages if pre-eclampsia supervenes, as already discussed.

Hypertension in the puerperium

In relation to both chronic hypertension and pre-eclampsia, the highest blood pressures are often recorded in the puerperium, typically peaking at about 5 to 7 days after delivery. Antihypertensive treatment therefore has to be continued, in some women for 3 to 6 weeks or even longer after delivery. There is no clear evidence that treatment interferes with breast feeding (Table 8).

Long-term sequelae of hypertension in pregnancy

Severe pre-eclampsia and eclampsia can cause irreversible maternal complications, particularly acute renal cortical necrosis or cerebral haemorrhage. In the absence of these problems, there is no evidence that long-term health is impaired. However, in terms of life expectancy, pre-eclamptic women fall into two groups. Those who become normotensive soon after delivery have a normal life expectancy. Those who remain hypertensive not only tend to suffer recurrent pregnancy-induced hypertension but have a higher incidence of later cardiovascular disorders and reduced life expectancy, compatible with the diagnosis of underlying arterial disease.

Conclusions

A raised blood pressure is one of many secondary effects of pre-eclampsia on the mother. In pre-eclampsia the main differential diagnosis is from chronic hypertension, which in its pure form does not share the renal, coagulation, hepatic, and placental abnormalities of pre-eclampsia. The perinatal risks of chronic hypertension in pregnancy result from superimposed pre-eclampsia.

Extreme hypertension (³170/110 mmHg) in pregnancy, whatever the underlying cause, is as dangerous as it is in any other medical situation and demands treatment. However, there is no clear reason for treating more moderate hypertension on either maternal or fetal grounds. As far as it is known, the progression of moderate pre-eclampsia is not delayed, nor is the later superimposition of pre-eclampsia on moderate chronic hypertension prevented.

Methyl dopa is the most thoroughly tested antihypertensive agent for use in pregnancy; no significant adverse reaction has been observed. Labetolol is a popular alternative. In general, b-adrenergic blocking agents are safe for short-term use but cause significant fetal growth retardation if administered over longer periods (from the second trimester), although the clinical trial data are less complete than for methyl dopa. Diuretics should primarily be reserved for the treatment of heart failure complicating pre-eclampsia. Angiotensin converting enzyme inhibitors are contraindicated for use in pregnancy because of adverse effects on fetal renal function.

Further reading

Davey DA, MacGillivray I (1988). The classification and definition of the hypertensive disorders of pregnancy. *American Journal of Obstetrics and Gynecology* **158**, 892–8.

Department of Health (1998). *Why mothers die. Report on confidential enquiries into maternal deaths in the United Kingdom 1994–1996*, pp. 36–46. HM Stationery Office, London.

National High Blood Pressure Education Program (1990). National High Blood Pressure Education Program Working Group Report on high blood pressure in pregnancy. *American Journal of Obstetrics and Gynecology* **163**, 1691–712.

Redman CWG (1991). Current topic: pre-eclampsia and the placenta. *Placenta* **12**, 301–8.

Redman CWG, Roberts JM (1993). Management of pre-eclampsia. *Lancet* **341**, 1451–4.

Redman CWG, Sacks GP, Sargent IL (1999). Preeclampsia: an excessive maternal inflammatory response to pregnancy. *American Journal of Obstetrics and Gynecology* **180**, 499–506.

Roberts JM, Redman CWG (1993). Pre-eclampsia: more than pregnancy-induced hypertension. *Lancet* **341**, 1447–51.

Sibai BM, Ramadan MK, Usta I, Salama M, Mercer BM, Friedman SA (1993). Maternal morbidity and mortality in 442 pregnancies with hemolysis, elevated liver enzymes, and low platelets (HELLP syndrome). *American Journal of Obstetrics and Gynecology* **169**, 1000–6.

The Eclampsia Trial Collaborative Group (1995). Which anticonvulsant for women with eclampsia? Evidence from the Collaborative Eclampsia Trial. *Lancet* **345**, 1455–63.

von Dadelzsen P, Ornstein MP, Bull SB, Logan AG, Koren G, Magee LA (2000). Fall in mean arterial pressure and fetal growth restriction in pregnancy hypertension: a meta-analysis. *Lancet* **355**, 87–92.

13.5 Renal disease in pregnancy

J. Firth

Changes in the kidneys and urinary tract during normal pregnancy

Anatomical

The most obvious anatomical change in the urinary tract during pregnancy is dilatation of the calyces, renal pelvis, and ureter. Contrary to popular belief, the ureters are not floppy and toneless, indeed tone is increased, but urinary stasis within the ureters may nevertheless contribute to the risk of asymptomatic bacteriuria developing into acute pyelonephritis. Ureteric dilatation can persist for 3 or 4 months after pregnancy, and long term in about 10 per cent of women who have had children. The kidney enlarges by about 1 cm in length during pregnancy.

Functional

Renal blood flow increases by 70 to 80 per cent between conception and midpregnancy, falling to a value 50 to 60 per cent above the non-pregnant level during the third trimester. Between conception and 16 weeks of pregnancy the glomerular filtration rate increases about 50 per cent above baseline and remains at this elevated level until delivery. Plasma creatinine decreases from a mean non-pregnant value of 73 µmol/l to 65, 51, and 47 µmol/l in successive trimesters.

Urinary excretion of glucose increases soon after conception and may rise 10-fold above non-pregnant values, hence glycosuria is common during pregnancy. This occurs because of decreased tubular reabsorption of glucose, but the reason for this is not known. Glucose excretion returns to normal non-pregnant levels within a week of delivery.

Plasma uric acid concentration decreases by about 25 per cent during normal pregnancy because of increased urinary excretion (the precise mechanism is unknown). In pregnancies complicated by pre-eclampsia or intrauterine growth retardation, the plasma uric acid concentration is higher than normal and serial measurements can be used to monitor progress. (See Chapter 13.4 for further discussion.)

The mean albumin excretion rate in pregnancy is 12 mg/day, with 29 mg/day the upper limit of normal (no different from that in the non-pregnant state). However, slightly increased urinary protein excretion is normal in pregnancy, such that proteinuria in pregnancy should not be considered abnormal until it exceeds 500 mg/day, which is over twice the upper limit of normal outside of pregnancy.

Pregnancy in women with known renal disease

Is pregnancy advisable?

Normal pregnancy is rare in women with serum creatinine more than 250 to 300 µmol/l, but many women who have renal disease will seek medical advice regarding pregnancy, the most typical questions being 'will pregnancy make my kidneys worse or do any long-term damage?' and 'will having kidney disease affect the pregnancy or baby?' The bottom line is that there is no strong medical contraindication to pregnancy from the 'renal point of view' if kidney function is only mildly compromised, proteinuria is not in the nephrotic range, and hypertension is nothing other than mild, although no woman can be given a 100 per cent guarantee. Table 1 gives a guide to the likely outcome of pregnancy in women with renal disease.

Women with a serum creatinine of less than 125 µmol/l usually have successful obstetric outcomes and there is no evidence that pregnancy adversely affects their renal prognosis. Some would suggest that more guarded advice should be given to patients with lupus nephropathy, mesangiocapillary glomerulonephritis, focal segmental glomerulosclerosis, and (perhaps) IgA nephropathy and reflux nephropathy.

When serum creatinine is in the range 125 to 250 µmol/l there is serious concern that pregnancy may cause immediate deterioration in renal function, severe hypertension, variable obstetric outcome, and an increased rate of decline of renal function postpartum.

Most women with serum creatinine more than 250 µmol/l are not fertile. The chances of conception, a normal pregnancy, and a healthy child are low; the risks to maternal health are high and pregnancy should be strongly discouraged. Women in this situation who are desperate to conceive should be told that their best chance of becoming pregnant and having a child is 1 year after successful renal transplantation.

Monitoring and management of pregnancy in women with renal disease

The chances of a successful outcome to pregnancy in a woman with renal disease are best if there is close co-operation between the patient, obstetrician, and nephrologist. At regular visits the following should be monitored: blood pressure, 24-h urine collection for proteinuria and creatinine clearance, urinary culture (for covert bacteriuria), and fetal development. Management of difficult cases requires achieving an acceptable balance between maternal and fetal interests, which is not always easy.

It is common for proteinuria to develop or to increase substantially during pregnancy in women with renal disease. Pregnancy can be allowed to continue as long as blood pressure is normal and renal function does not deteriorate. Reversible causes, such as volume depletion or urinary infection, should be sought if renal function does decline, but significant and otherwise unexplained deterioration in renal function is reason to recommend elective delivery.

The incidence of pre-eclampsia in patients with pre-existing renal disease is not known, hypertension and proteinuria being manifestations of both. The management of hypertension is the same, whether the cause is renal disease or pre-eclampsia, and is as discussed in Chapter 13.4, although whether or not mild hypertension should be treated in pregnant women with renal disease is a contentious issue. Blood pressure above 170/110 mmHg in the third trimester would be an indication for starting antihypertensive treatment in routine obstetric practice. By contrast, treatment of blood pressure above 140/90 mmHg in a non-pregnant woman of child-bearing age would be indicated in general medical practice to prevent long-term complications, and in those with renal disease (particularly if proteinuric) most nephrologists would recommend that arterial pressure be lowered even further than this. It is not known whether or not blood pressure in pregnancy should be treated more aggressively in women with renal disease than in those without.

Women with renal transplants

Fertility returns in women of child-bearing age after successful renal transplantation and over 90 per cent of pregnancies that survive the first trimester end successfully. It is generally recommended that women should not attempt to conceive for the first year after transplantation. Thereafter the risks are similar to those

suggested above for a woman with disease of her native kidneys; the risk of problems is low in the patient with good renal function, no proteinuria, and no hypertension, and there is no medical reason to dissuade such a woman from becoming pregnant. By contrast, the hypertensive woman with proteinuria and a poorly functioning graft is at high risk of adverse maternal and fetal outcome and should be strongly dissuaded from pregnancy. Prednisolone, azathioprine, and cyclosporin seem to be safe in pregnancy, there being no need for dosage adjustment in most cases. The presence of a renal transplant is not a contraindication to normal vaginal delivery and is not an indication for caesarean section. There is no increase in the incidence of congenital abnormalities in the infants of mothers with renal transplants.

Renal complications that can occur in pregnancy

Urinary tract infection

Asymptomatic bacteriuria

The incidence of asymptomatic bacteriuria is 2 to 10 per cent in pregnant and non-pregnant young women. If not treated, 40 per cent of women with asymptomatic bacteriuria will develop acute symptomatic infection during pregnancy, and treating pregnant women with asymptomatic bacteriuria should prevent approximately 70 per cent of all cases of symptomatic urinary tract infection in pregnancy. Untreated asymptomatic bacteriuria is associated with low birth weight and preterm delivery. Treatment is effective at clearing asymptomatic bacteriuria and reducing the incidence of pyelonephritis, preterm delivery, and babies with low birth weight.

E. coli is responsible for over 75 per cent of cases of bacteriuria in pregnancy. The choice of drug should be determined by the sensitivity of the organism isolated. Amoxicillin and cephalosporins are safe in pregnancy, as is nitrofurantoin, except in the last few weeks (it can produce neonatal haemolysis if used at term). Trimethoprim is contraindicated in the first trimester (folate antagonist) and quinolones should not be used at all in pregnancy (they cause arthropathy in animal studies). There is no good evidence to determine the first choice of antibiotic (pending culture results) or the duration of treatment. Use of single, high-dose therapy is controversial and some would recommend a 2-week course, with repeat urinary culture performed 1 week after treatment is stopped and monthly thereafter until delivery. About 25 per cent of patients will suffer recurrent infection and require a second course of treatment.

Symptomatic infection

Acute cystitis affects about 1 per cent of pregnant women. Treatment is as for asymptomatic bacteriuria, with the aim of abolishing symptoms and preventing acute pyelonephritis. A Cochrane review that included five studies of antibiotic treatment of symptomatic urinary infection in pregnancy concluded that there were no significant differences between any of the treatments studied and was unable to recommend any particular regimen.

Acute pyelonephritis presents with the same symptoms as it does in patients who are not pregnant (see Chapter 20.12), but the differential diagnosis in pregnancy includes uterine fibroid degeneration and abruptio placentae, and distinction from appendicitis can be particularly difficult. Preterm labour occurs in about 4 per cent of mild cases and 20 per cent of severe cases, where there is associated respiratory distress. Treatment is usually with intravenous antibiotics (typically amoxicillin or a cephalosporin) in the first instance, switching to oral therapy as the patient improves. Some advocate a 3-week course for acute pyelonephritis in pregnancy, but without firm evidence that this is necessary. However, since pyelonephritis recurs in up to 25 per cent of women during pregnancy it is important thereafter to monitor closely for evidence of recurrent urinary infection, with or without instigating prophylactic treatment with a bedtime dose of an appropriate antibiotic.

Acute renal failure specific to pregnancy

The priorities in dealing with acute renal failure in a pregnant woman are no different from those that apply to any other patient: treatment of life-threatening complications, fluid resuscitation (if necessary), establishing a precise diagnosis, treatment of any underlying condition (if possible), and timely provision of renal replacement therapy (if indicated). These issues are discussed in Section 20.5. There are, however, some causes of acute renal failure that are specific to pregnancy.

Obstetric acute renal failure

In the developed world, the incidence of obstetric acute renal failure has fallen dramatically over the last 40 years. In Leeds (United Kingdom), obstetric causes accounted for 26 per cent of cases of acute renal failure between 1956 and 1959, compared to 1.3 per cent between 1980 and 1988. A significant fall has also been seen in some parts of the developing world, where in Chandigarh (North India) obstetric causes were responsible for 25 per cent of cases of acute renal failure 30 years ago, compared with around 10 per cent today. By contrast, in a study from Turkey, the proportion of cases of acute renal failure due to obstetric causes changed only from 17 per cent to 14 per cent over the period 1980 to 1997. Where a decline in the incidence of obstetric acute renal failure has been seen, this is almost certainly attributable to improvements in perinatal care and a reduction in the numbers of septic abortions. Should acute renal failure develop in pregnancy or the puerperium then obstetric causes listed in Table 2 should be considered in addition to the conditions that can present in any patient (see Table 1, Table 2, and Table 3 of Chapter 20.4).

In most cases of acute renal failure in pregnancy, the clinical progress of the condition is typical of acute tubular necrosis, with full recovery of renal function if the patient survives. However, for reasons that are not known, pregnant women are particularly susceptible to acute cortical necrosis. The incidence of this has fallen in the developed world, from 1 in 10 000 in the decade from 1960 to less than 1 in 80 000 pregnancies in the decade from 1970 in one large study, but it still carries the risk of permanent renal failure, although delayed and partial recovery is not uncommon. (See Chapter 20.4 for further discussion.)

Acute fatty liver of pregnancy

Acute fatty liver of pregnancy is a disease of the third trimester or puerperium in which there is jaundice and severe hepatic dysfunction. In early series, the incidence of acute renal failure was over 50 per cent, but this is now much reduced, both as a consequence of the recognition of milder cases and through improved management of severe cases. The explanation for acute renal failure is unknown. Renal biopsy shows non-specific changes. (See Chapter 13.9 for further discussion.)

Haemolytic uraemic syndrome (idiopathic postpartum renal failure)

The haemolytic uraemic syndrome can present peripartum or at any time in the first few weeks after delivery, when it is often termed idiopathic postpartum renal failure. This can be a devastating condition (maternal mortality 5–20 per cent, fetal mortality 30 per cent), the cause of which is not known. Blood pressure can vary from low to very high. There may be severe heart failure and also neurological manifestations such as coma or epileptic fitting. The blood film shows a microangiopathic haemolytic anaemia, sometimes with a consumption coagulopathy in addition. Treatment is primarily supportive, but some authorities also recommend plasma exchange, although in this rare condition there are no randomized, controlled trials to justify its use. If the patient survives then renal function rarely, if ever, recovers completely and many remain dependent on dialysis. (See Chapter 20.10.6 for further discussion.)

Miscellaneous conditions

Acute hydroureter and hydronephrosis and the 'overdistension syndrome'

The anatomical changes associated with pregnancy can very occasionally be exaggerated, with massive distension of the ureters and renal pelvis. This is usually asymptomatic, but can rarely have clinical consequences. The 'overdistension syndrome' varies from transient, mild loin pain to recurrent attacks of severe loin or lower abdominal pain radiating to the groin. Variation in symptoms with posture and position are typical. Urine specimens are sterile, but can show microscopic haematuria. Diagnosis is confirmed by ultrasonography. Nursing in the knee–chest position often provides relief, but very uncommonly nephrostomy and/or ureteral stenting are required.

Rupture of the urinary tract

The development of severe, persistent pain or haematuria in a pregnant patient with acute pyelonephritis or the 'overdistension syndrome' suggests the very rare complication of rupture of the urinary tract. This can be retroperitoneal or intraperitoneal and is more likely to occur in those with pre-existing disease of the kidneys or

urinary tract, perhaps because mild weaknesses are exposed by the physiological changes of pregnancy. Distinction from other obstetric or abdominal catastrophes can be very difficult.

Further reading

Jones DC, Hayslett JP (1996). Outcome of pregnancy in women with moderate or severe renal insufficiency. *New England Journal of Medicine* **335**, 226–32.

Meyers SJ *et al*. (1985). Dilatation and nontraumatic rupture of the urinary tract during pregnancy: a review. *Obstetrics and Gynecology* **66**, 809–15.

Packham DK *et al*. (1989). Primary glomerulonephritis and pregnancy. *Quarterly Journal of Medicine* **71**, 537–53.

Pertuiset N, Grünfeld JP (1994). Acute renal failure in pregnancy. *Baillieres Clinical Obstetrics and Gynaecology* **8**, 333–51.

Romero R *et al*. (1989). Meta-analysis of the relationship between asymptomatic bacteriuria and preterm delivery/low birth weight. *Obstetrics and Gynecology* **73**, 576–82.

Selcuk NY *et al*. (1998). Changes in frequency and etiology of acute renal failure in pregnancy (1980–1997). *Renal Failure* **20**, 513–7.

Smaill F (2001). Antibiotics for asymptomatic bacteriuria in pregnancy (Cochrane review). *Cochrane Database Systematic Review* **2**, CD000490. http://www.update-software.com/abstracts/ab000490.htm

Vazquez JC, Villar J (2000). Treatments for symptomatic urinary tract infections during pregnancy. *Cochrane Database Systematic Review* CD002256. http://www.update-software.com/abstracts/ab002256.htm

Wolff JM *et al*. (1995). Non-traumatic rupture of the urinary tract during pregnancy. *British Journal of Urology* **76**, 645–8.

13.6 Heart disease in pregnancy

J. C. Forfar

Introduction

Up to 10 per cent of maternal deaths in the United Kingdom result from heart disease, and the proportion is higher in developing countries. Although rheumatic heart disease has declined world-wide, it remains an important and treatable condition in pregnancy. The improved survival of patients with treated complex congenital heart lesions presents difficult challenges for planning and management. Unfortunately, information on fetal and maternal risks in such situations is often inadequate, making it difficult to know what the best treatment is. Maternal involvement in the analysis of risks and benefits is of the greatest importance.

Cardiovascular changes in pregnancy

Significant circulatory changes occur during pregnancy (Table 1). Cardiac output increases by up to 50 per cent through increase in both heart rate and stroke volume secondary to a fall in systemic vascular resistance. The fall in blood pressure results from vasodilation mediated by gestational hormones, increased heat production, and the low resistance uteroplacental circulation. Pressure of the gravid uterus on the cava in the supine position can result in substantial falls in cardiac output and blood pressure in some women, causing weakness, light-headedness, dizziness, and even syncope. The supine hypotensive syndrome of pregnancy is relieved by turning to one side.

Blood volume increases early in pregnancy and more slowly from mid-pregnancy. These changes follow activation of the renin–angiotensin–aldosterone mechanism by oestrogens and secondary retention of salt and water. Increases in plasma volume (average 50 per cent) are faster and greater than red cell mass (average 25 per cent), thus accounting for the so-called 'physiological anaemia of pregnancy', when haemoglobin averages 11 g/dl.

Blood pressure, cardiac output, and oxygen consumption increase with the pain and anxiety of labour and from expansion of blood volume with uterine contraction. Despite blood loss on delivery, blood volume is effectively (although not actually) increased by the contracting empty uterus postpartum, and augmented preload acutely increases cardiac output. These changes return to normal over 24 h.

Cardiovascular evaluation in pregnancy

Assessment of heart disease in pregnancy is complicated by the functional changes described above, which may simulate or mask underlying heart disease. Routine investigative methods may be limited because of potential risks to the fetus.

Fatigue, reduced exercise capacity, breathlessness, and light-headedness are common symptoms during pregnancy. The increase in respiratory rate and reduced tidal volume may be perceived as breathlessness and this, combined with some peripheral oedema from increased blood volume and caval compression, with minor distension of the neck veins, may lead to an erroneous diagnosis of heart failure. The increased volume peripheral pulses, occasionally 'collapsing', and vasodilatation during later pregnancy can mimic aortic regurgitation or hyperthyroidism. The apical impulse is prominent and occasionally displaced, simulating volume loading. Heart sounds are commonly loud and may be palpable with exaggerated splitting. An apical S3 is not uncommon in pregnancy but S4 is less common and its presence should promote investigation for possible heart disease.

A soft systolic murmur at the base and lower left sternal edge occurs in most pregnancies and can radiate to the suprasternal notch and the neck. A continuous venous flow murmur may be audible in the neck along with a systolic or continuous murmur from increased flow to the breasts. Both murmurs decrease with stethoscope pressure and are less audible when the patient is upright. A diastolic murmur is rare in normal pregnancy and should prompt investigation. Murmurs associated with organic heart disease increase in intensity in pregnancy (from increased flow) and conclusions based on murmur intensity should therefore be made with caution.

Cardiovascular investigations in pregnancy

Electrocardiography (ECG)

Minor flattening of the T-waves and minor left or right QRS axis shift are common in normal pregnancy. Sinus tachycardia and atrial or ventricular premature beats are likewise frequent findings, particularly during labour and delivery. Exercise testing may confirm a clinical diagnosis of ischaemic heart disease or assess functional capacity. A low level protocol is recommended.

Chest radiography

Although the radiation dose is minimal, this test is best avoided unless there are clear indications. The uterus should be shielded. Straightening of the left heart border and prominence of the pulmonary conus are common findings, along with prominent pulmonary vascular markings. A small pleural effusion may be seen early postpartum and usually resolves quickly.

Doppler echocardiography

Cardiac ultrasound is safe and provides valuable anatomical and functional information throughout pregnancy. Late in normal pregnancy a small pericardial effusion can sometimes be seen. Trivial tricuspid and pulmonary regurgitation may be considered a variant of normal at this stage.

Cardiac catheterization

This should only be undertaken if an intervention is required for patient management. Balloon mitral valvuloplasty is an alternative to closed valvotomy for symptomatic severe mitral stenosis in pregnancy. Aortic and pulmonary balloon valvuloplasty have also been performed, as has coronary angioplasty. The brachial approach is preferable to minimize radiation exposure, and full shielding is appropriate. Radionuclide imaging techniques during pregnancy are best avoided because of uncertainty of the level of radiation exposure to the fetus.

Heart disease in pregnancy

Cardiac reserve (assessed by history and functional classification) is an important determinant of risk in pregnancy. Patients with heart disease and with no, or minimal, symptoms prior to pregnancy have a relatively small risk of complications. Patients with moderate or severe limitation prior to pregnancy are at much higher risk, and careful monitoring, and sometimes intervention, are required. Patients symptomatic at rest prior to pregnancy have a high maternal and even higher fetal mortality and pregnancy is contraindicated until their functional class can be improved.

It is therefore important that accurate diagnostic and functional evaluation of heart disease is undertaken in all patients prior to pregnancy, aiming to predict maternal and fetal risk as far as possible. The issues of maternal morbidity, long-term survival, and risks of fetal heart disease need to be discussed with the patient. It is preferable, in those with complex congenital heart disease, that such discussions take place well before pregnancy is contemplated so that appropriate action may be taken. Where predicted maternal mortality is above 30 per cent, sterilization or pregnancy termination may be most appropriate. The management of anticoagulant therapy before and during pregnancy requires special consideration (see below).

Congenital heart disease

This accounts for up to one-third of heart disease in pregnancy. Maternal and hence fetal outcome is determined by the nature of the disease, previous surgical repair, functional capacity, cyanosis, and the presence of pulmonary hypertension. Heart failure, arrhythmias, and systemic hypertension are more common with cyanotic congenital heart disease and when functional status is compromised. Infective endocarditis is a risk peripartum.

Fetal mortality averages 40 per cent in maternal cyanotic congenital heart disease, over twice that for the acyanotic mother with congenital heart disease. Low birth weight for gestational age and prematurity are more common in cyanotic mothers. The risk of important congenital defects in the fetus of mothers with congenital heart disease varies between 3 and 15 per cent.

Elective induction of labour at fetal maturity often allows better planning of monitoring and availability of key personnel. Caesarean section is usually performed for obstetric reasons rather than maternal status. Haemodynamic monitoring and maintenance of filling pressures minimize maternal risk. Antibiotic chemoprophylaxis is indicated during uncomplicated vaginal delivery in the presence of a valve prosthesis, a right to left shunt, or a surgical left to right shunt. In some centres, prophylactic antibiotics are used more widely.

Maternal risk can be classified according to structural diagnosis (Table 2).

Acyanotic shunt lesions

Atrial septal defect

This common lesion is usually well tolerated in pregnancy, even with a substantial left to right shunt. Antibiotic prophylaxis is not indicated for an isolated secundum atrial septal defect. Elective closure should be considered following delivery.

Ventricular septal defect

This lesion is well tolerated in pregnancy unless there is persistent pulmonary hypertension or impaired cardiac reserve.

Persistent ductus arteriosus

Percutaneous transcatheter closure of a persistent ductus arteriosus has occasionally been required during pregnancy because of the development of heart failure resistant to medical therapy, but the majority have a low risk provided pulmonary hypertension is not severe.

Acyanotic obstructive lesions

Aortic valve disease

Mild aortic stenosis can be missed in pregnancy because of the prevalence of a flow-related systolic murmur and confusion between a split first heart sound and the ejection click from a bicuspid valve. Asymptomatic patients with moderate or severe aortic stenosis may require careful haemodynamic monitoring during labour and delivery. Patients with symptomatic severe aortic stenosis should be advised against pregnancy, or should consider termination. Valvotomy and percutaneous balloon valvuloplasty have been performed successfully in the second and third trimesters for severe symptoms.

Coarctation of the aorta

Pregnancy is usually uneventful, although impaired fetal development, hypertension, heart failure, and aortic dissection have all been described. If possible, aortic coarctation should be corrected prior to pregnancy.

Pulmonary stenosis

The risks associated with pregnancy in this condition are low, although heart failure has been described. Balloon valvuloplasty or surgical valvotomy are intervention options, but clinical experience of these in pregnancy is small.

Cyanotic lesions

Tetralogy of Fallot

This is the commonest cyanotic congenital heart disease in adults. Pregnancy may cause major clinical deterioration because of augmented right ventricular pressure from increased blood volume and increased right to left shunting from reduced systemic vascular resistance. Reduced arterial oxygen saturation and a high haematocrit are worrying signs, and careful monitoring of labour and delivery is essential. Maternal and fetal risk are reduced where the defect has been surgically repaired, and this should be performed prior to pregnancy. Palliative aortopulmonary shunting leaves a significant risk during pregnancy.

Eisenmenger's syndrome

The high maternal mortality with this condition has been confirmed by several studies, although there are some isolated successes. One in four pregnancies reach term with a very high prevalence of growth retardation, prematurity, and perinatal death. Because of this, pregnancy is contraindicated and early termination should be considered. Thomboembolism is more common and anticoagulation should be considered in late pregnancy and postpartum. Intensive haemodynamic monitoring to avoid blood loss and blood pressure swings and to maintain filling pressures is appropriate.

Complex cyanotic congenital heart disease

The increasing success of palliative and corrective surgical procedures for complex cyanotic congenital heart disease has allowed pregnancy to be contemplated in those surviving to child-bearing age. Successful pregnancies have been reported with a single ventricle, corrected and uncorrected transposition of the great vessels, tricuspid and pulmonary atresia, but the risks are substantial and often inappropriate. Even in the absence of severe pulmonary hypertension, serious cardiovascular complications occur in one-third of pregnancies. The literature probably presents an underestimate of the risk because of a tendency to report success.

Rheumatic heart disease and other valve lesions

Rheumatic heart disease

Although rare in the United Kingdom, acute rheumatic fever may develop or recur during pregnancy and the associated carditis carries a substantial maternal risk. Antibiotic chemoprophylaxis has been advised throughout pregnancy in patients with a history of recurrent rheumatic fever.

Patients with rheumatic valvular disease should be managed according to disease site and severity. Careful haemodynamic monitoring may be necessary in late pregnancy, labour, and postpartum.

Mitral stenosis

This is the commonest cardiac lesion in pregnancy, with complications including heart failure, atrial arrhythmias (particularly fibrillation), and thromboembolism. Digoxin and b-blockade may be effective in controlling the ventricular rate and maintaining cardiac output when arrhythmias ensue. Diuretics reduce pulmonary congestion but hypovolaemia should be avoided. Asymptomatic or mildly symptomatic patients usually progress uneventfully. Severe symptomatic mitral stenosis can be successfully relieved by percutaneous balloon valvuloplasty or closed or open valvotomy. Both fetal and maternal risks appear to be modest with these procedures, although experience of balloon valvuloplasty is small. Systemic and pulmonary thromboembolism is a special complication in pregnancy and chronic anticoagulant therapy should be considered in those at higher risk based on standard criteria (see Chapter 13.7).

Mitral regurgitation

This lesion is usually well tolerated, presumably because of reduced systemic vascular resistance. Vasodilator therapy and digoxin may be indicated.

Mitral valve prolapse

This has been recognized in up to 15 per cent of women of child-bearing age and provided mitral regurgitation is not severe, there appears to be little risk to mother or fetus.

Aortic valve disease

Severe rheumatic aortic valve disease is uncommon in pregnancy. Aortic regurgitation is more frequent, but like mitral regurgitation is fairly well tolerated, partly because of the fall in systemic vascular resistance and tachycardia, both of which reduce myocardial work.

Cardiomyopathy in pregnancy

Hypertrophic cardiomyopathy

Pregnancy appears to be well tolerated, although symptoms of heart failure can develop or worsen at any stage. It is uncertain whether the risk of sudden death is increased by pregnancy. Careful haemodynamic monitoring may be indicated during labour and delivery, and avoidance of systemic vasodilatation, blood loss, and hypotension are essential. The role of b-blockade is uncertain, but these drugs can be useful in managing symptoms related to elevated left ventricular filling pressure and where there is an obstructive element to the condition.

Peripartum cardiomyopathy

This form of dilated cardiomyopathy causes left ventricular systolic dysfunction and heart failure in the last trimester of pregnancy or within the first 6 months postpartum. Other causes of dilated cardiomyopathy presenting in pregnancy should be excluded. The disease is rare in Europe but commoner in parts of Africa, with a prevalence of up to 1 per cent. Although the cause is unknown, the age at onset, geographical frequency variation, and substantial recovery of ventricular function in the majority of individuals suggest a unique and specific syndrome. Myocarditis, nutritional deficiency, small vessel coronary disease, and maternal immunological responses to a fetal antigen have all been proposed as mechanisms. Endomyocardial biopsy reveals inconsistent changes. The role of genetic factors is poorly defined.

Clinical examination reveals cardiomegaly, a third or fourth heart sound, and commonly functional mitral and tricuspid regurgitation. ECG abnormalities are widespread and usually non-specific, and a variety of arrhythmias have been described. The haemodynamic changes are similar to other forms of dilated cardiomyopathy.

Most patients respond to conventional management with digoxin, diuretics, and vasodilators. Angiotensin converting enzyme (ACE) inhibitors may affect fetal renal function and are not recommended. The role of immunosuppresive therapy is controversial and not established.

Fifty per cent of patients show substantial or complete recovery of ventricular function within 6 months postpartum. In the remainder, there is persistent left ventricular dysfunction and chronic heart failure. Progressive clinical deterioration and early death occur in a minority. There is a significant risk of thromboembolism and anticoagulant therapy is often appropriate.

The risk or relapse in a subsequent pregnancy is substantial (up to 40 per cent) and is greater in patients with persistently abnormal ventricular function.

Ischaemic heart disease

Symptomatic coronary artery disease in pregnancy has a prevalence of around 1 in 10 000. Conventional risk factors apply: particularly smoking and previous oral contraceptive use in combination, and hypercholesterolaemia. Peripartum myocardial infarction is occasionally associated with normal epicardial coronary arteries at subsequent angiography. Spasm, *in situ* thrombosis, dissection, and plaque rupture have been proposed as mechanisms.

Myocardial infarction is associated with a maternal mortality up to 25 per cent. Low-dose aspirin is relatively safe but use of thrombolytic drugs and percutaneous coronary intervention has not been adequately evaluated. Haemodynamic monitoring may be appropriate peripartum and elective caesarean section should be considered.

The Marfan syndrome

Pregnancy in patients with the Marfan syndrome is associated with an increased incidence of aortic dissection and death, particularly if existing cardiovascular disease is identified. An aortic root diameter of less than 4.5 cm and absence of progressive aortic dilatation and aortic regurgitation suggest a better outcome. However, selective literature reporting probably overestimates the true risk. In patients with established aortic root dilatation, pregnancy is undesirable and if root enlargement is progressive during the first trimester, termination must be considered. b-Blockade may reduce the rate of aortic dilatation during pregnancy, but clinical experience is limited. Regular monitoring of aortic root diameter is recommenced. Elective caesarean section may be preferable to vaginal delivery in those with significant aortic root dilatation.

Aortic dissection in the absence of the Marfan syndrome is rare, and occurs most frequently during the third trimester of pregnancy and peripartum. Precordial and transoesophageal echocardiography is the diagnostic investigation. Emergency measures to reduce blood pressure are required, although maternal and fetal mortality is high.

Primary pulmonary hypertension

A high maternal mortality rate (up to 40 per cent) has been reported and pregnancy is contraindicated. Haemodynamic deterioration during pregnancy is not predictable on the basis of preconception haemodynamics. Anticoagulation is recommended during pregnancy and early post partum.

Cardiac arrhythmias

Atrial and ventricular premature beats occur commonly in pregnancy and appear to have no adverse consequences. Patients with a substrate for supraventricular tachycardia, such as atrioventricular nodal or atrioventricular re-entry tachycardia, may experience symptomatic deterioration in pregnancy, or may develop symptoms for the first time. Bradycardia and heart block are rare and are usually congenital. Specific causes should be sought and, if appropriate, removed. Antiarrhythmic drug therapy should be initiated only for persistent arrhythmias threatening the mother or fetus (see below). Catheter ablation procedures should be undertaken postpartum to avoid unpredictable exposure to ionizing radiation.

Cardiac surgery in pregnancy

Although many cardiac operations have been described in pregnancy, including those on cardiopulmonary bypass, detailed evaluation of the risk–benefit balance is inadequate. There have been many reports of mitral valvotomy with low maternal and fetal mortality. In general, surgery considered essential for maternal health should be performed during the middle trimester or towards the end of the third trimester when elective or emergency delivery can be planned. Sustained uterine contraction is essential postpartum to minimize the risk from systemic heparinization during cardiopulmonary bypass. In appropriate circumstances, balloon valvuloplasty or coronary angioplasty seem attractive options.

Pregnancy and valve prosthesis

Increased cardiac output, a hypercoagulable state, deterioration of the bioprosthesis, and fetal risks from anticoagulation all potentially complicate pregnancy with a valve prosthesis. However, patients with no or minimal limitation before pregnancy usually have a favourable outcome. The risk of thromboembolic events with mechanical valves appears increased, partly related to anticoagulant control and the use of fixed-dose heparin regimens in the first trimester. Tissue valves are often recommended for women who wish to have children, but the long-term durability of heterograft valves in particular is questionable and premature degeneration from progressive leaflet calcification is more common in the young.

Cardiovascular drugs in pregnancy

All drugs should be avoided where possible during pregnancy and the risk/benefit balance carefully evaluated in terms of maternal health and fetal risk. For information on prescribing in pregnancy, see Chapter 13.18. Two specific issues will be discussed here: anticoagulation and the use of prophylactic antibiotics for those at risk of endocarditis.

Anticoagulation

There is no ideal anticoagulant in pregnancy. Use must depend on individual assessment of risks and benefits. Warfarin has been most widely used but has significant maternal and fetal side-effects. The risk of haemorrhage at delivery means that it should be discontinued approaching term. Fetal risks are derived from the transplacental passage of warfarin: there is a relatively high incidence of spontaneous abortion and stillbirth, and use during the first trimester has been associated with a 'coumarin embryopathy' in 5 to 15 per cent of infants so exposed. This syndrome includes hypoplasia of the nasal bone and epiphyseal stipling (chondrodysplasia punctata). Central nervous system disease, including optic atrophy and blindness, mental retardation, cerebral palsy, and intracranial bleeding have all been described and linked to warfarin.

Heparin, because of its molecular weight, does not cross the placenta. Its use in pregnancy is not without difficulty, although recent studies have shown satisfactory outcomes. Complications of long-term subcutaneous therapy include haematoma and abscess formation in the abdominal wall, thrombocytopenia, and osteoporosis. Self-injection of an adjusted dose of heparin subcutaneously every 12 h is a satisfactory approach. Fixed-dose regimens should be avoided. There is a consensus that intravenous heparin should be substituted close to term and discontinued at the onset of labour. Low molecular weight heparins have not been adequately studied during pregnancy. When the risk of thromboembolic complications is very high, for example with a mechanical mitral valve prosthesis continuous warfarin therapy (INR 2–3.5:1) may be a reasonable strategy.

A suggested strategy for the management of anticoagulation in pregnancy is shown in Fig. 1. In patients at moderate risk this involves early substitution of heparin for warfarin (preferably prior to conception) and return to warfarin from the end of the first trimester until 2 weeks before term. Continued heparin therapy throughout the pregnancy is an alternative widely practised in North America where litigation fears limit oral anticoagulation use. With the exception of premature infants, warfarin excretion in breast milk does not cause significant anticoagulation for the infant.

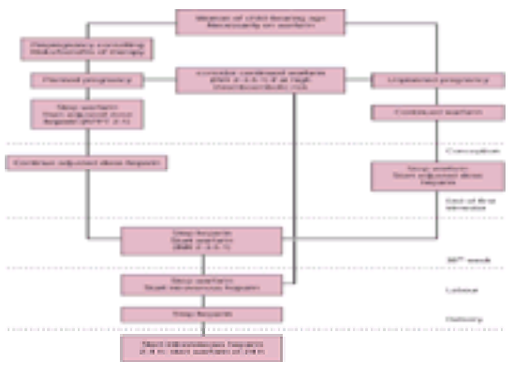

Fig. 1 A strategy for anticoagulation during pregnancy.

Prophylactic antibiotics

The incidence of bacteraemia associated with an uncomplicated vaginal delivery is low and thresholds for antibiotic chemoprophylaxis are uncertain. Routine prophylaxis is recommended for prosthetic heart valves, congenital heart disease with a right to left shunt, and when aortopulmonary anastomosis has been created with prosthetic material. Many physicians, however, also routinely administer prophylactic antibiotics to those at conventional risk from endocarditis, including valvular heart disease, ventricular septal defect, and hypertrophic obstructive cardiomyopathy. Treatment with intravenous ampicillin and gentamicin approximately 1 h prior to delivery is recommended with a second dose 8 h later. Slow intravenous infusion of vancomycin and gentamicin can be used in those who are allergic to penicillin. For patients at low risk, conventional chemoprophylaxis with oral amoxycillin is widely practised.

Further reading

Elkayam U, Gleicher N (eds) (1990). *Cardiac problems in pregnancy—diagnosis and management of maternal and fetal disease*, 2nd edn. John Wiley, Chicester.

Perloff JK (1992). Congenital heart disease. In: Gleicher N, ed. *Principles and practice of medical therapy in pregnancy*, 2nd edn. Appleton and Lange, Norwalk, Conn.

Sullivan JM, Ramanathan KB (1985). Management of medical problems in pregnancy—severe cardiac disease. *New England Journal of Medicine* **313**, 304–9.

Weiderhorn J, Rubin JM, Frishman WH, Elkayam U (1987). Cardiovascular drugs in pregnancy. *Cardiology Clinics* **5**, 651–745.

13.7 Thromboembolism in pregnancy

M. de Swiet

This section is concerned with thromboembolism in pregnancy, specifically deep vein thrombosis and pulmonary embolism. Arterial thromboembolism, which in pregnancy usually arises because of mitral valve disease and/or atrial fibrillation, cardiomyopathy, and the presence of artificial heart valves, is considered elsewhere. Cerebral vein thrombosis is considered in Section 15.

Significance and incidence

Pulmonary embolus is a leading cause of maternal mortality in the United Kingdom, responsible for 46 deaths between 1994 and 1996 (21 deaths per million maternities). In addition, deep vein thrombosis is a major cause of morbidity: about 80 per cent of women who have deep vein thrombosis in pregnancy have symptoms in the same leg at follow-up 11 years later. Because of the difficulties in diagnosis of non-fatal cases (see below) it is difficult to obtain precise data for the incidence of non-fatal pulmonary embolus or deep vein thrombosis. However, recent figures for the incidence of venous thromboembolism in the United Kingdom (Scotland) are 1.0/1000 maternities in women under 35 years of age and 2.4/1000 in those over 35 years. About twice as many episodes occur antenatally as in the 6-week postnatal period. Pulmonary embolism contributes 10 to 20 per cent of all cases; the remainder are deep vein thromboses.

In Africa and the Far East the condition has been almost unknown, in part due to the lower prevalence of inherited thrombophilias (see below); but those countries that have become more affluent have seen a corresponding increase in the incidence of thromboembolism.

Risk factors

It is generally believed that pregnancy itself is a risk factor for venous thromboembolism, presumably because of activation of the clotting system. Venous stasis in the lower limbs caused by obstruction of the venous return by the enlarging uterus is another factor. This effect is more marked in the left leg than the right and accounts for the fact that in pregnancy deep vein thrombosis is approximately ten times more common in the left leg than the right.

Analysis of fatal cases of pulmonary embolus shows that increasing age and parity are important risk factors: the risk of dying from pulmonary embolus in a woman aged over 40 years in her fifth pregnancy is nearly 100 times greater than in a primigravid woman aged 20 to 30 years. Caesarean section (and probably other forms of complicated instrumental delivery) increases the risk about three-fold. Oestrogen therapy to suppress lactation also increases the risk of thromboembolism and should no longer be used. Women who have had thromboembolism in the past have a 1 in 10 to 1 in 20 risk of thromboembolism in pregnancy, the risk being the same whether or not the previous episode occurred whilst taking oestrogen-containing oral contraceptives. It is not known whether previous thromboembolism in pregnancy increases the risk in subsequent pregnancies above the figures given. It is likely that prolonged periods of bed rest and obesity are also predisposing factors.

The most important specific haematological risk factors for thromboembolism are the thrombophilias. Acquired thrombophilia mainly comprises the antiphospholilpid syndromes, characterized by the presence of lupus anticoagulant and anticardiolipin antibodies, considered in Chapter 13.14. Knowledge of the inherited thrombophilias is expanding rapidly. Current estimates of the pregnancy risks of thromboembolism are as follow: factor V Leiden heterozygote: 1in 437 pregnancies; protein C deficiency: 1 in 113; type 1 antithrombin deficiency 1 in 3; type 2 antithrombin deficiency 1in 42. Increased risks that have not been accurately quantified also exist for those with protein S deficiency, the prothrombin gene mutation, and hyperhomocysteinaemia. Women with thrombophilia who have a family history of thromboembolism, and particularly those who have themselves already had an episode of thromboembolism, have a further increase in risk.

It must also be remembered that the fetus is at risk in maternal thrombophilia. The risks of miscarriage, second and third trimester loss, abruption, premature rupture of membranes, and pre-eclampsia are best characterized for acquired antiphospholipd syndrome, but they are increased in varying degree for the inherited thrombophilias as well.

Patients with any of the sickling conditions (Hb SS, Hb SC, Hb S thalassaemia) are at increased risk of developing the sickle lung syndrome in pregnancy, one component of which is probably thrombosis *in situ.* Management of haemoglobinopathies in pregnancy is considered in Chapter 13.16.

Diagnosis

The clinical features of pulmonary embolus and deep vein thrombosis are considered in Chapter 15.15.3.1. These do not differ in pregnancy, but the frequent occurrence in normal pregnancy of leg oedema, breathlessness, minor degrees of pleural effusion in the puerperium, and abnormalities of the electrocardiograph, make the clinical diagnosis even more difficult in those who are pregnant than in those who are not. Furthermore, the problems of anticoagulant therapy with its associated risks to the fetus are such that every effort must be made to make the diagnosis by objective criteria. The implications of this policy are as follows.

Deep vein thrombosis

All patients who have a history compatible with deep vein thrombosis and supporting physical signs in the legs should have a real-time ultrasound examination of the leg veins with additional Doppler blood flow studies if possible. This technique has been compared with venography in the non-pregnant state and shown to have high sensitivity and specificity for symptomatic femoral vein thrombosis. The direct comparison has not been made in pregnancy. Ultrasound is not yet accurate for calf vein thrombi, but these do not cause pulmonary emboli and the ultrasound test, being non-invasive, can be repeated to ensure that a calf vein thrombosis is not extending into the thigh. Ultrasound cannot be used above the inguinal ligament, but iliofemoral thrombosis is usually clinically obvious.

Pulmonary embolus

If a patient has a major pulmonary embolus, there is usually little doubt about the diagnosis. However, pulmonary embolus is often considered as a cause of collapse in pregnant women, particularly at the time of delivery, and the differential diagnosis of occult causes of collapse without obvious bleeding or inverted uterus is considered in Table 1. Perhaps the most important cause of confusion is intra-abdominal bleeding, with irritation of the diaphragm causing chest and shoulder-tip pain. Treatment of such a patient with anticoagulants would probably be fatal. The most important clinical features to note are the presence of abdominal signs and low jugular venous pressure in abdominal bleeding, and the raised jugular venous pressure and cardiac signs of pulmonary artery obstruction in pulmonary embolus. However, the problem of diagnosis usually arises in the patient who has pleuritic chest pain with, or without, physical signs, and either absent or non-specific signs on chest radiography. Arterial blood gas estimation may be helpful: blood samples should be taken with the patient sitting and, in cases of major pulmonary embolus, will show respiratory alkalosis with hypoxaemia. However, the definitive investigation is a lung scan, preferably a ventilation perfusion scan if there are any radiological signs in the lung parenchyma. The radiation exposure from the short-lived isotopes that are used ($^{81}Kr^{m}$ for ventilation and $^{99}Tc^{m}$ for perfusion) is trivial. If the patient

has normal blood gases and a negative lung scan, it is reasonable to exclude pulmonary embolus.

Treatment

Surgery

If the facilities and an expert team are available, the occasional patient who does not die from a massive pulmonary embolus and remains shocked and hypotensive (blood pressure less than 90 mmHg systolic, PaO_2 less than 60 mmHg, urine output less than 20 ml/h) one h after the onset of symptoms, should be considered for pulmonary embolectomy under cardiopulmonary bypass. If the patient reaches the operating theatre alive, the results are excellent. Alternatively, it may be possible to fragment the clot and improve pulmonary blood flow with a catheter introduced into the pulmonary artery. This requires only the simplest image-intensifying equipment, such as is available in most intensive care units (see Section 16).

The place of surgery (embolectomy) in massive iliofemoral thrombosis, where it might decrease the incidence of subsequent postphlebitic leg symptoms, is unclear.

Thrombolytic therapy

Streptokinase and/or urokinase treatment is probably underused for non-pregnant patients with pulmonary embolus. However, there are specific problems in pregnancy, such as bleeding, the initiation of premature labour, and the subsequent inco-ordinate uterine action associated with the release of fibrin degradation products. Therefore, in pregnancy, thrombolytic therapy should only be used in shocked patients with life threatening pulmonary embolism and as an alternative to embolectomy.

Filters

There is no substantive evidence of benefit of temporary or permanent caval interrupt devices in patients judged to be at high risk of pulmonary embolism. In pregnancy their use should be limited to those who have recurrent pulmonary embolism despite adequate anticoagulation. In practice, this happens very rarely; most recurrent pulmonary emboli occur in women who have not been adequately anticoagulated.

Anticoagulant therapy

For the reasons given above, the majority of patients will be treated with anticoagulants, the objective being to prevent further thromboembolism. As in the non-pregnant state, heparin should be used initially. Neither unfractionated heparin (UH) nor any of the low molecular weight heparins (LMWH) cross the placenta or are secreted into breast milk in significant quantities, and if heparin were to be secreted in breast milk it would not be absorbed from the infant's gastrointestinal tract, being denatured in the stomach. The only acute problem with heparin therapy is bleeding. Possible long-term problems are discussed below.

Traditionally, in the acute phase of treatment in pregnancy, heparin has been given by continuous intravenous infusion, starting at 40 000 units/24 h, adjusting the rate to double the activated partial thromboplastin time, and continuing for about 10 days. However, in the non-pregnant state clinical trials have shown that initial treatment of both deep vein thrombosis and pulmonary embolus with fixed high dose LMWH given by intermittent subcutaneous injection is as effective as intavenous infusion of UH. Typical regimens include enoxaparine 1 mg/kg 12 hourly. The advantages of LMWH regimen are considerable, in particular the lack of need for monitoring and dose adjustment, and the possibility of ambulatory treatment. It is very likely that the use of such regimen in the acute phase of thromboembolism treatment will also become standard in pregnancy, but there are concerns about this. The trials performed in the non-pregnant state have been of LMWH supported by very early additional warfarin therapy; this will not happen in pregnancy (see below), and LMWH will be used on its own for varying periods at varying intensities of therapy. Furthermore, the clotting system is activated in pregnancy, such that the doses of LMWH that have been successfully evaluated in the non-pregnant state may not be appropriate for pregnancy. Until more experience is obtained it is recommended that those using high-dose LMWH in acute phase treatment check the anti-Xa heparin assay, aiming for trough levels greater than 0.4 units/ml and peak levels less than 1.0 units/ml.

After the acute phase, the therapeutic options are oral anticoagulants, of which there is far more experience with warfarin than phenindione, and subcutaneous heparin of one sort or another. The problems with warfarin therapy in pregnancy are maternal bleeding, particularly in the puerperium, miscarriage, teratogenesis (chondrodysplasia punctata), fetal microcephaly, optic atrophy, and fetal bleeding, both retroplacental and intracerebral. The latter complications can occur with warfarin therapy at any gestational age. For these reasons warfarin should not be used in venous thromboembolism in pregnancy, although its use is necessary in those with artificial heart valves.

Once intravenous heparin has been discontinued, the patient should be given subcutaneous injections of heparin at prophylactic doses, either UH 10 000 twice daily or a LMWH such as enoxaparine 40 mg once daily. In practice, most patients receive LMWH because once daily injections are more convenient. Other possible advantages of LMWH, which remain unproven, are greater and more constant bioavailabilty, a superior risk ratio of antithrombotic activity to bleeding risk, less heparin-induced thrombocytopenia, and less bone demineralization (see below). Most patients can learn to inject themselves and administer the treatment at home. Bruising appears to be related more to the injection technique than to the type of heparin. It is not clear what degree of monitoring (if any) is necessary in this phase of heparin treatment. If the anti-Xa heparin level is less than 0.3 units/ml there should be no risk of bleeding, but the minimum heparin level for effective chronic phase or prophylactic therapy is not known. Common practice has been to check the anti-Xa heparin level at varying intervals and only adjust the dose downwards should the level be greater than 0.3 units/ml. In the absence of recurrence of thromboembolism the dose has not been increased to more than those levels recommended above. An alternative when using UH is to measure the thrombin time, which is very sensitive to excess UH (but not LMWH) activity: if it is not prolonged the patient should not be at any excess bleeding risk. In practice, when using injections of UH 10 000 twice daily or enoxaparine 40 mg once daily it is very uncommon for the anti-Xa level to be greater than 0.3 units/ml or for any of the standard clotting tests (activated partial thromboplastin time, thrombin time, INR) to be abnormal; hence the uncertainty about the need to perform these tests.

Assuming that the standard clotting tests are normal, there is no increased risk from bleeding in labour and subcutaneous heparin therapy should be continued through labour. Although there has been concern about the possibility of epidural haematoma formation in women taking subcutaneous heparin given epidural anaesthesia, this concern is not justified. Nevertheless, epidural block is often withheld until 2 h after an injection of UH or 4 h after an injection of LMWH. If an epidural catheter has been inserted similar constraints relate to its removal.

After delivery the dose of subcutaneous UH heparin is reduced to 7500 units twice daily; the dose of LMWH need not be changed. This treatment should be continued for the first week of the puerperium. After that time, the risk of secondary postpartum haemorrhage is small and patients may switch to oral warfarin therapy if that appears more desirable. Breast feeding is safe in patients taking warfarin: insignificant quantities of warfarin (though not phenindione) are secreted in the milk. Since blood-clotting parameters do not return to normal immediately after delivery, anticoagulation should be continued for some time after delivery in a patient who had thromboembolism in the antenatal period. Six weeks is often the time chosen, but the length of time is quite arbitrary.

Prophylaxis

Previous thromboembolism

Should any form of prophylactic therapy be given in pregnancy to women who have had an episode of thromboembolism in the past, granted the 5 to 10 per cent risk of recurrence? Trial data are not adequate to answer this question. Warfarin therapy is contraindicated because of the complications noted above. Some clinicians would use subcutaneous heparin throughout pregnancy, but the incidence of maternal side-effects—heparin induced thrombocytopenia (HIT) (see Chapter 13.16), alopecia, and bone demineralization—are a cause of considerable concern.

Heparin-induced thrombocytopenia is very uncommon in obstetric practice. The incidence may be less with LMWH than with UH, but patients who have had HIT with UH are certainly at risk if they take LMWH. A better alternative is the heparinoid orgaron. There are anecdotal reports of its use in pregnancy: it is also given parentally and does not cross the placenta. Clinical bone demineralization has been reported in patients taking as little as 10 000 units of heparin/day for 19 weeks, and there is evidence of subclinical bone demineralization in many patients taking heparin 20 000 units/day for more than 3 months.

If prophylaxis is to be given, the alternatives to 'standard dose' UH or LMWH prophylaxis are: the use of smaller quantities for shorter periods of time; low dose aspirin

(75 mg/day); or no prophylaxis in the antenatal period, starting heparin in labour, continuing in the puerperium for at least 1 week, and with the option of switching to warfarin for a further 5 weeks. None of these strategies has been fully evaluated.

Given the balance of risks, most clinicians would not advise prophylaxis for the patient who has simply had a single episode of thromboembolism in the past. They would, however, suggest that the patient accept the risk of prolonged subcutaneous heparin therapy if she has had more than one well-documented episode of thromboembolism, or if she has had a single episode of thromboembolism and has a known inherited or acquired thrombophilia. Patients who have had a single episode but also have a family history of thromboembolism in a first-degree relative should also receive prophylactic heparin throughout pregnancy: they may well have thrombophilia that has not yet been detected.

Thrombophilia

As indicated above, advances are being made very rapidly in this field. The risk of thromboembolism in pregnancy depends on the manner in which the condition was detected. Those found to have thrombophilia following an episode of thromboembolism are at greatest risk; those found because of a family history are at intermediate risk; and those found through population screening are at least risk.

Patients who have the antiphospholipid syndrome and have had previous thromboembolism should take prophylactic UH or LMWH throughout pregnancy and for at least 6 weeks after delivery. They are also likely to be taking low-dose aspirin for fetal reasons. If they have a particularly bad history of thromboembolism they should take LMWH in relatively high dose, granted their need to take warfarin aiming for INR 3 to 4 when not pregnant. If they have been taking heparin and aspirin in pregnancy for fetal reasons only, and have no history of previous thromboembolism, most clinicians would discontinue thromboprophylaxis at delivery unless there were other risk factors such as delivery by emergency Caesarean section.

Type 1 antithrombin deficiency has such a high risk of thromboembolism that women should take high-dose LMWH throughout pregnancy regardless of their history of thromboembolism. Antithrombin concentrate is now available and depending on the measured level of antithrombin may be given to cover labour or if the woman has an episode of thromboembolism.

It is difficult to be dogmatic about management of other forms of thrombophilia given the lack of trial data. However, patients with Type 2 antithrombin deficiency and those with homozygous factor V Leiden should probably have prophylactic low-dose heparin throughout pregnancy, whatever their past history, and if they have had a previous clot the level of anticoagulation could be increased by increasing the dose. In the remaining thrombophilias, protein C and S deficiencies, the prothrombin gene mutation, heterozygous factor V Leiden, and hyperhomocysteinaemia, heparin prophylaxis may be withheld during pregnancy but given peripartum if there is no past history of thromboembolism. Aspirin can be given to cover the antenatal period. If there is a past history of thromboembolism, prophylactic heparin should be given throughout pregnancy. Patients with hyperhomocysteinaemia should take folic acid supplements; the optimal dose is not clear, 5 mg daily should be sufficient.

Protein S deficiency cannot be diagnosed reliably in pregnancy since the levels of both free and bound protein S decrease in normal pregnancy. Genetic tests for factor V Leiden should be performed if the activated protein C resistance is low in pregnancy. However a low activated protein C resistance in pregnancy does not necessarily indicate abnormality since the activated protein C resistance is also decreased in normal pregnancy.

If a woman considering pregnancy is known to be protein C deficient, her partner's protein C status should be checked. Protein C deficiency is usually expressed in the heterozygous form. However, homozygous protein C deficiency can occur and the individual is then dependent on lifelong infusions of protein C concentrate to prevent recurrent episodes of thromboembolism.

Further reading

Dahlman T, Lindvall N, Hellgren M (1990). Osteopenia in pregnancy during long-term heparin treatment: A radiological study post partum. *British Journal of Obstetrics and Gynaecology* **97**, 221.

Department of Health (1998). *Why mothers die, report on confidential enquiries into maternal deaths in the United Kingdom 1994–96*. HMSO, London.

de Swiet M (1995). Thromboembolism. In: de Swiet M, ed. *Medical disorders in obstetrics practice*, 3rd edn, pp. 116–42. Blackwell Scientific Publications, Oxford.

de Swiet M, *et al.* (1983). Prolonged heparin therapy in pregnancy causes bone demineralization. *British Journal of Obstetrics and Gynaecology* **90**, 1129–34.

Greer IA (1999). Thrombosis in pregnancy: maternal and fetal issues. *Lancet* **353**, 1258–65.

Howell R, Fidler J, Letsky E, de Swiet M (1983). The risks of antenatal subcutaneous heparin prophylaxis: a controlled trial. *British Journal of Obstetrics and Gynaecology* **90**, 1124–8.

Maclon NS, Greer IA (1996). Venous thromboembolic disease in obstetrics and gynaecology: the Scottish experience. *Scottish Medical Journal* **41**, 83–6.

Nelson Piercy C, Letsky E, de Swiet M (1997). Low molecular weight heparin for obstetric thromboprophyllaxis: experience of 69 pregnancies in 61 women at high risk. *American Journal of Obstetrics and Gynecology* **176**, 1062–8.

13.8Chest diseases in pregnancy

M. de Swiet

Physiology

The pregnant woman at rest increases her minute volume by about 40 per cent within the first trimester of pregnancy. This is achieved by an increase in tidal volume rather than in respiratory rate, and is more than adequate to account for the increased metabolic rate of the mother and fetus, even in the later stages of pregnancy when maternal oxygen consumption increases by about 45 ml/min. The stimulus to increased ventilation is said to be increased progesterone secretion, but it is likely that this is not the only factor.

The majority view is that vital capacity and airways resistance do not change in pregnancy, although, surprisingly, transfer factor is reduced. The uterus enlarging within the abdomen partly accounts for a 20 per cent reduction in residual volume, but this reduction, together with change in the shape of the chest wall, occurs early in pregnancy, before there can be any mechanical effect due to the uterus. The reduction in residual volume is more marked when supine than sitting. $P\mathrm{aO}_2$ in the sitting or erect posture is unchanged in pregnancy, but falls by up to 2 kPa (15 mmHg) in patients who are supine in late pregnancy, probably due to unequal ventilation/perfusion ratios subsequent to airways closure during tidal breathing. Therefore, where possible, blood gases for diagnostic purposes should always be taken in pregnant patients when they are sitting. $P\mathrm{aCO}_2$ falls to about 4.0 kPa (30 mmHg) and plasma bicarbonate falls proportionately to about 20 mmol/l, hence there is no change in the arterial pH.

Breathlessness is a common symptom in pregnancy, presumably associated with the 40 per cent increase in ventilation that occurs in normal women. However, this cannot be the entire explanation because ventilation increases from before 4 weeks' gestation, whereas the maximum incidence of onset of breathlessness is at 28 to 31 weeks' gestation. Breathlessness is worrying for the doctor (and for the patient) because it may also be associated with cardiopulmonary disease, particularly pulmonary embolism. In the absence of other features of cardiopulmonary disease, and with normal findings on examination, useful investigations are chest radiography, arterial blood gas estimation, oximetry to determine oxygen saturation at rest and on exercise, and full lung function testing including measurement of transfer factor.

Lung disease in pregnancy—general considerations

Although ventilation does increase by 40 per cent in pregnancy, this increase is trivial in comparison to the marked increase (perhaps ten-fold) that is possible during exercise. This considerable reserve of ventilatory capacity is not greatly challenged by pregnancy and respiratory failure due to chronic respiratory disease is uncommon in pregnancy, the major problem for women with chronic conditions such as asthma or tuberculosis being the effect of therapy on pregnancy.

By contrast, acute respiratory failure is a major cause of maternal mortality: adult respiratory distress syndrome (ARDS) is the final common pathway for many obstetric disasters, carrying a mortality of about 70 per cent. Indeed, much of the practice of modern obstetrics is directed towards the avoidance of ARDS. For example the trend towards epidural rather than general anaesthesia reduces the risk of inhalation of stomach contents, and therefore of ARDS. Epidural, rather than general anaesthesia, is specifically indicated in all patients with significant chest disease.

Adult respiratory distress syndrome

The management of this condition is described in Chapter 16.5.2. Specific obstetric causes are indicated in Table 1: of these, pre-eclampsia is probably now the most common, followed by shock with or without disseminated intravascular coagulation. Some of these obstetric causes of ARDS warrant further consideration.

Inhalation of stomach contents

This only occurs in the absence of an effective gag reflex, which in obstetric practice is almost invariably associated with general anaesthesia. Most obstetric units starve their patients once labour is established, and it is a common though declining practice to give regular antacid therapy, since it is believed that the low pH of gastric contents makes them particularly harmful to the lungs, although ARDS has developed in patients given aluminium hydroxide in labour after inhalation of stomach contents at pH 6.4. Avoidance of inhalation is the single most important preventive measure. Since the maternal mortality from inhalation has fallen from 6 per cent of all maternal deaths in England and Wales in 1976 to 1978 to only one death in the United Kingdom in 1994 to 1996, current measures are probably having some effect.

Amniotic fluid embolism

This catastrophe occurs because amniotic fluid and other material of fetal origin enters the maternal circulation. Usually, though not invariably, it occurs at the end of a vigorous labour with intact membranes, generally with some obstetric intervention. Important elements in the pathogenesis are widespread deposition of platelet and fibrin thrombi, and disseminated intravascular coagulation (DIC) caused by the very high thromboplastin activity of amniotic fluid. The initial presentation is with profound hypotension and cyanosis. If the patient survives this, she is at risk of dying from haemorrhage due to DIC, and if she survives the DIC she is at risk from ARDS. In the anaesthetized patient, differentiation from inhalation is important. Bronchospasm is common in inhalation, but very rare in amniotic fluid embolus. DIC is an early presenting feature in amniotic fluid embolus, but it occurs late after inhalation. The differential diagnosis of other causes of collapse in pregnancy is considered in Table 1 of Chapter 13.7.

The diagnosis of amniotic fluid embolus can only be confirmed by finding fetal material, that is squames or hairs, in the maternal blood (from central venous pressure lines), in the sputum, or in the lungs at autopsy. However, even the finding of fetal material in maternal blood is not specific, since this has been reported in some normal women having Swann Ganz catheterization in labour. There is no specific treatment.

Asthma

Bronchial asthma is the commonest chest disease in pregnancy, with prevalence of 3 to 5 per cent. In keeping with the lack of change in airways resistance in normal pregnancy, there is little evidence that pregnancy consistently affects the clinical course of asthma. Those studies that have been performed suggest that airways responsiveness to a metacholine challenge improves during pregnancy, but this beneficial effect is trivial by comparison with the inappropriate reluctance of patients and their carers to continue normal therapy for asthma in pregnancy (see below).

There has always been concern about a possible effect of asthma on the outcome of pregnancy. Large epidemiological studies do suggest increased risks of preterm delivery, low birthweight, and congenital malformations, with relative risks of 1.3 to 2.2 compared to healthy controls. These risks almost certainly relate to poor treatment, though the cause of the birth defects is not clear. Other studies have shown no excess fetal risks when asthma specialists manage asthma in pregnancy

and achieve good control of symptoms.

It is unusual for patients to have acute attacks of asthma in labour. Perhaps the high circulating levels of endogenous catecholamines, corticosteroids, and prostaglandins are protective.

The treatment of patients with asthma does not require modification in pregnancy. Current management guidelines for asthma recommend a stepped care approach. Those with infrequent attacks should take inhaled betasympathomimetics as required to relieve symptoms, but if these are being used on a daily basis, additional regular inhaled glucocorticoids should be taken. The dose of inhaled glucocorticoid may be increased to for example beclomethasone 2000 µg per day. Alternatively a long acting b-agonist such as salmeterol or an anticholinergic drug (ipatropium) may be added. Oral prednisone should be used if these drugs do not achieve adequate control.

There is sufficient experience with all these classes of drug to recommend their use in pregnancy, but it seems sensible to use those with which there is most experience: namely oral salbutamol and inhaled beclomethasone, salmeterol, and ipatropium. Theophyllines can also be used in pregnancy, though their volume of distribution increases, making dosing difficult and emphasizing the need for monitoring of blood levels. Disodium cromoglycate has also been used extensively in pregnancy without problems, but little is used now because it is ineffective. There is insufficient experience to recommend the use of leukotriene antagonists in pregnancy.

The concern that oral steroids may be teratogenic and cause facial clefts is not supported by current studies. There is also no evidence of suppression of the fetal hypothalamo–pituitary–adrenal axis, at least with doses of up to 25 mg prednisone per day. Prednisone or hydrocortisone (by contrast with betamethasone or dexamethasone) are extensively metabolized by placental enzymes and little crosses the placenta. Women who have taken extended courses of oral steroids in the previous year should be given parenteral steroids to cover labour, as in any other situation where addisonian collapse is a possibility.

Of other drugs sometimes given to patients with asthma, aminoglycosides should only be given for more than 24 h in pregnancy if there is no alternative, and then with continuous monitoring of blood levels. This is because of the risk of damage to the fetal eighth nerve and kidney. Tetracycline should not be used because it causes permanent discoloration of the fetal teeth. Iodine-containing expectorants should not be used in pregnancy or in lactating women, since the odine freely crosses the placenta, is excreted in breast milk, and may cause hypothyroidism in the infant.

It has been suggested that ergometrine can cause severe bronchospasm in patients with asthma. Syntocinon ® should therefore be used for the management of the third stage of labour.

In summary, the management of patients with asthma requires little modification in pregnancy. In the unlikely event of an attack severe enough to require ventilation, maternal hypoxaemia should be avoided because of the associated severe fetal hypoxaemia; so also should hypocapnia (PCO_2 <17 mmHg, 2.3 kPa) and alkalosis (pH > 7.6) since these have been associated with fetal hypoxaemia, probably due to impaired placental transport.

Chronic bronchitis, bronchiectasis, and emphysema

These conditions are now very uncommon in pregnancy. Since pulmonary hypertension is poorly tolerated in pregnancy, cor pulmonale is likely to be the factor limiting maternal safety. The presence of arterial hypoxaemia puts the fetus at risk from intrauterine growth restriction.

Cystic fibrosis

Better management in childhood means that more patients with cystic fibrosis are surviving and wanting to have children. Recent analysis of 20 series concerning 217 pregnancies in 162 women indicated that pregnancy did not affect mortality, when pregnant women with cystic fibrosis were compared to non-pregnant women with cystic fibrosis. However, 24 per cent of all deliveries were preterm. Poor outcomes were associated with a pregnancy weight gain of less than 4.5 kg and a prepregnancy forced vital capacity of less than 50 per cent of the predicted value. This group is likely to produce a preterm infant and to suffer increased loss of pulmonary function and increased maternal mortality. In one small study, four out of seven women with prepregnancy FEV_1 below 60 per cent predicted died within 3.2 years of delivery, whereas all 15 women with prepregnancy FEV_1 above 60 per cent predicted survived.

Apart from high quality and intensity of obstetric and medical care, no specific measures are necessary in pregnant patients with cystic fibrosis. Most of the drugs used have been considered above. Inhaled aminoglycosides are probably safe in pregnancy.

Patients may have malabsorption due to pancreatic involvement in cystic fibrosis, and an increase in pancreatic supplements may be necessary. Diabetes mellitus can also become manifest for the first time in pregnancy, and all with cystic fibrosis should be screened for diabetes early in pregnancy, and at about 28 weeks' gestation. There is also an increased risk of pneumothorax in labour (see below).

There has been concern that women with cystic fibrosis should not breast feed their infants because of the possibility of very high sodium content of their milk (up to 280 mmol/l). This risk has probably been exaggerated, since the samples of breast milk initially analysed were taken from women who were not lactating freely, a situation in which all breast milk has high sodium content. More recent studies have indicated that once lactation has been established, the breast milk has normal sodium content. Lactation should not threaten the mother's weight providing this was reasonably maintained before pregnancy.

For couples who have had one child affected by cystic fibrosis, first-trimester prenatal diagnosis is possible on genetic material prepared from chorionic villi using linked DNA probes in at least two-thirds of cases. The prevalence of the gene in the community is said to be 1 in 20, although the overall prevalence of the disease is 1 in 2500. Therefore, women with cystic fibrosis should be counselled that there is a 1 in 20 to 1 in 44 chance that their child will have the condition if the father's status is unknown. If the father is a heterozygote, the risk is 1 in 2. All the children of affected mothers will be carriers.

Kyphoscoliosis

Mild degrees of kyphoscoliosis have no effect on pregnancy, and successful pregnancy is possible in patients with severe disease and a vital capacity of as little as 1000 ml. As in the other chest diseases, hypoxaemia and pulmonary hypertension are the limiting factors, and some with severe kyphoscoliosis become exhausted and then hypoxaemic in the last trimester. Any suggestion of excessive fatigue should be an indication for hospital admission for rest and nasal intermittent positive pressure ventilation if this is not available at home. Progressive hypoxaemia, with or without evidence of fetal compromise, is an indication for delivery. Labour and/or caesarean section are best managed with the assistance of epidural anaesthesia, which reduces the risk of atelectasis. It can be given to most patients, even those with very severe spinal abnormalities.

Pneumothorax and pneumomediastinum

There are rare complications of pregnancy (incidence less than 1 in 10 000) but they probably occur more commonly in those who are pregnant than those who are not. This is particularly so in labour, when it is supposed that the raised intrathoracic pressure due to straining is a contributing factor. However, there is often some predisposing condition such as asthma, cystic fibrosis, emphysema, or lymphangioleiomyomatosis. Cocaine use has also been implicated. Both pneumothorax and pneumomediastinum present with chest pain, and if there is a substantial leak of air the patient may be hypotensive and cyanosed (tension pneumothorax, malignant pneumomediastinum). The differential diagnosis of these and other occult causes of collapse in pregnancy are considered in Chapter 13.7. The physical signs and management of the pneumothorax are no different in pregnancy from those in the non-pregnant state, and are described in Chapter 17.12. If a patient has a past history of pneumothorax or pneumomediastinum, she should have an elective forceps delivery to prevent recurrence during straining in labour.

Tuberculosis

Before the advent of antituberculous therapy, tuberculosis was the cause of many maternal deaths, particularly in the puerperium. This is no longer so, and there should be no excess mortality from tuberculosis in pregnancy. However, a high index of suspicion is necessary to make the diagnosis in pregnancy: most centres in the United Kingdom have not screened for tuberculosis in pregnancy by routine Mantoux testing and/or chest radiographs for many years. In the future, and in areas

with high prevalence of tuberculosis in pregnancy, this policy may have to be changed.

The placenta is a very efficient filter, and intrauterine infection of the fetus almost never occurs, though neonatal infection from the mother can certainly be a problem.

None of the front line antituberculous drugs, rifampicin, pyrazinamide, isoniazid, or ethambutal, has been shown to be teratogenic despite extensive experience. However, isoniazid (and rifampicin) can cause serious hepatitis, which in the case of isoniazid seems to be a particular risk for pregnant women. Pregnant patients should therefore have regular liver function tests and receive supplementary pyridoxine when taking isonazid. A dose of 50 mg/day has been shown to give adequate blood levels in pregnancy, but the conventional dose of 10 mg/day may also be effective. Streptomycin (an aminoglycoside) should be avoided for the reasons given above. Ethionamide should not be used because of reports of multiple congenital abnormalities. Within these limitations, the pregnant patient with tuberculosis should be treated in the same manner as she would be in the non-pregnant state. In particular, patients with HIV disease or those with drug-resistant tuberculosis should be treated without regard to the pregnancy: the risk of inadequate maternal therapy is far greater than any potential risk of antituberculous treatment for the fetus.

Patients who have been adequately treated in the past for tuberculosis do not require prophylactic therapy in pregnancy. After birth, babies should only be isolated from their mothers if the mothers are still smear positive. Since modern antituberculosis regimes render the sputum sterile within 2 weeks and markedly reduce the number of organisms within 24 h, this should not occur frequently. The neonate should be treated with prophylactic isoniazid for 3 months. After this period BCG vaccination is given in the United Kingdom but not in the Unites States. It is not clear whether neonatal BCG vaccination adds any further protection to isoniazid prophylaxis. It is not without risks: skin ulceration and osteitis may occur, and occasionally disseminated disease, particularly if the mother has an immunodeficiency state. As isoniazid therapy does not affect the immunogenicity of BCG vaccine, there is no longer any rationale for the use of isoniazid-resistant BCG neonatal vaccination.

Sarcoid

This condition is rarely a problem in pregnancy. Patients who have had sarcoid in the past have no extra risk of relapse in pregnancy. Those who have active sarcoid during pregnancy tend to improve, possibly because of the increase in free as well as protein-bound cortisol levels. There is a tendency to deteriorate in the puerperium, but this should not be overemphasized. Since patients take many vitamins in pregnancy, those with sarcoid should be warned not to take vitamin D to which they may be very sensitive.

Pneumonia

Pneumonia is an uncommon complication of pregnancy, and unless the diagnosis is clear-cut, other conditions that produce chest symptoms and signs, such as pulmonary embolus, should always be considered. Bacterial pneumonia should be treated with antibiotics. For community-acquired pneumonia the antibiotics of choice in the United Kingdom are likely to be a penicillin or a macrolide such as erythromycin, both of which are suitable for use in pregnancy. Aggressive antipyretic therapy with tepid sponging, fans, and regular paracetamol should be used, because of the association between pyrexia and premature labour. During epidemics, such as the influenza epidemic in 1930, viral pneumonia has caused a high mortality in pregnancy, and clinicians should be aware that this condition should not be dismissed lightly.

Ten per cent of maternal varicella infections may be complicated by pneumonia, which has an appreciable mortality, particularly in pregnancy. All pregnant women who have close exposure to varicella-zoster virus and who have no demonstrable antibody should therefore receive zoster immune globulin. Those who develop varicella should receive acyclovir 10 to 30 mg/kg daily in three divided doses for 5 days.

Further reading

Brost BC, Newman RB (1997). The maternal and fetal effects of tuberculosis therapy. *Obstetrics and Gynecology Clinics of North America* **24**, 659–73.

Demissie K, Breckenridge MB, Rhoads GG (1998). Infant and maternal outcomes in the pregnancies of asthmatic women. *American Journal of Respiratory and Critical Care Medicine* **158**, 1091–5.

Department of Health (1998). *Why mother die, report on confidential enquiries into maternal deaths in the United Kingdom 1994–96*. HMSO, London.

de Swiet M (1995). Diseases of the respiratory system. In: de Swiet M, ed. *Medical disorders of obstetric practice*, 3rd edn, pp. 1–32. Blackwell Scientific Publications, Oxford.

de Swiet M (1998). The respiratory system. In: Chamberlain GVP, Broughton Pipkin F, eds. *Clinical physiology in obstetrics*, 3rd edn, pp. 111–28. Blackwell Science, Oxford.

Edenborough FP, Stableforth DE, Webb AK, Mackenzie WE, Smith DL (1995). Outcome of pregnancy in women with cystic fibrosis. *Thorax* **50**, 170–4.

Margono F, Mroueh J, Garely A, White D, Duerr A, Minkoff HL (1994). Resurgence of active tuberculosis among pregnant women. *Obstetrics and Gynecology* **83**, 911–14.

Milne JA, Howie AD, Pack AI (1978). Dyspnoea during normal pregnancy. *British Journal of Obstetrics and Gynaecology* **85**, 260–3.

Morgan M (1979). Amniotic fluid embolism. *Anaesthesia* **34**, 20–32.

White RJ, Coutts II, Gibbs CJ, MacIntyre C (1989). A prospective study of asthma during pregnancy and the puerperium. *Respiratory Medicine* **83**, 103–6.

13.9 Liver and gastrointestinal diseases during pregnancy

A. E. S. Gimson

Introduction

Gastrointestinal and liver disease in pregnancy includes those diseases specific to that condition, those occurring with increased frequency during pregnancy, and those already present at conception or arising coincidentally during the course of pregnancy (Table 1). Liver or gastrointestinal dysfunction is present in fewer than 5 per cent of pregnancies in Europe and the United States, but their recognition and management is important as increased maternal and fetal morbidity and mortality may result without prompt intervention.

There are significant physiological changes in hepatic function during pregnancy. Increased circulating blood volume and cardiac output are not associated with any changes in hepatic blood flow, but there is increased azygous flow, which results rarely in the formation of small oesophageal varices. Gallbladder motility is reduced and bile lithogenicity increased due to increased hepatic cholesterol synthesis and excretion into bile. Minor but important changes in laboratory blood tests occur due to haemodilution or alteration in hepatic synthesis (Table 2).

Increased gastric myoelectric activity may be manifest as nausea and vomiting, but there are few other significant changes in gastrointestinal function during normal pregnancy.

Liver disease specific to pregnancy

Hyperemesis gravidarum

Although nausea and vomiting may occur in up to 75 per cent of pregnancies, severe vomiting leading to dehydration, ketonuria, electrolyte disturbances, and nutritional deficiency is rare, developing in 2 to 16 of every 1000 pregnancies. Nutritional deficiency has been so severe as to progress to Wernicke's encephalopathy and changes in serum sodium may precipitate osmotic demyelination (central pontine myelinolysis). More common in younger women and in obesity, recent surveys do not suggest any relationship with parity or gravidity. Elevated transaminases, by two- to threefold, occur in 50 per cent of cases, with a minor rise in alkaline phosphatase and bilirubin in 10 per cent. Liver histology shows few abnormalities or hepatic steatosis only. The aetiology is unclear and may be multifactorial: positive *Helicobacter* serology has been reported in up to 90 per cent of cases, and changes in thyroid function are present in 50 per cent. An elevated free T_4 with suppressed thyroid-stimulating hormone correlates with elevated human chorionic gonadotrophin levels in these patients, raising the possibility that gestational thyrotoxicosis may also have role in pathogenesis. In some cases psychological factors are also important.

Management, which may include hospitalization, is symptomatic and includes rehydration and correction of nutritional deficiencies. Psychological support is crucial. Treatment of symptomatic gastro-oesophageal reflux is important and antiemetics are required, with metoclopramide and promethazine as effective as newer 5-HT_3 antagonists. There are uncontrolled reports of benefit with corticosteroids. In rare cases there may be recurrence in subsequent pregnancies.

Intrahepatic cholestasis of pregnancy

A cholestatic disorder of the second and third trimesters, this is the most common cause of jaundice during pregnancy, after acute viral hepatitis. Initially starting with pruritus, jaundice follows after 1 to 4 weeks in 20 to 60 per cent of cases, associated with pale stools and dark urine. Diagnosis is by history and the classic biochemical features of an elevated bilirubin (less than 100 µmol/litre) and increased aminotransferases (rarely more than 250 iu/litre), with no significant rise in alkaline phosphatase or g-glutamyl transpeptidase. Serum bile acids increase three- to 100-fold: those factors with the highest predictive diagnostic value being total bile acid concentration of more than 11.0 µmol/litre and a cholic/chenodeoxycholic acid ratio of more than 1.5 with a cholic acid percentage over 42. An ultrasound scan is necessary to exclude choledocholithiasis, and further imaging of the bile duct with magnetic resonance cholangiopancreatography will occasionally be needed. A liver biopsy is not required for diagnosis, but shows a canalicular cholestasis with no hepatocellular necrosis.

The epidemiology of intrahepatic cholestasis of pregnancy is interesting, with marked geographical variation. The highest incidence, over 10 per cent, has been recorded in Arauncanian Indians in Chile, with 2 to 3 per cent in Sweden and 0.1 per cent in Canada. It is reported to be rare in Afro-Caribbeans. Family studies have suggested a dominant mode of transmission in a few kindreds, and a non-sense mutation has been reported in the *MDR3* gene that encodes an ATP-dependent transporter of phosphatidylcholine across the canalicular membrane in one family. An association with *HLA-B8*, *HLA-B12*, and *DPB1* alleles is unconfirmed. It is more common in women with a history of contraceptive pill induced jaundice (50 per cent), those with benign recurrent intrahepatic cholestasis, where there are multiple gestations, and it may be more common in women who are positive for hepatitis C virus antibodies. In France it has been associated with use of progesterone in early pregnancy.

The aetiology is unknown, but one hypothesis suggests enhanced sensitivity of components of the bile salt excretion apparatus to oestrogen: pregnancy impairs sulphation of both monohydroxy bile salts and oestrogen, which may enhance the cholestatic potential of both compounds.

Intrahepatic cholestasis of pregnancy is associated with an increased incidence of fetal prematurity (fivefold increase), fetal distress, stillbirths, and meconium staining of amniotic fluid (1.5-fold increase), but perinatal mortality is normal with modern management. Reports of maternal morbidity from postpartum haemorrhage relate to vitamin K deficiency and are not a feature of recent series.

Treatment is symptomatic, with the bile salt ursodeoxycholic acid (10–15 mg/kg/day) the treatment of choice and better than *S*-adenosyl-methionine. Ursodeoxycholic acid relieves pruritus, reduces bile salt levels in maternal serum, and may reduce the frequency of fetal complications. Both bile salt sequestration with cholestyramine and dexamethasone have given variable results. Vitamin K should be given before delivery.

Most authors recommend early elective delivery at 38 weeks to prevent late fetal complications, and this policy has been shown to improve fetal outcome. More recently there have been suggestions that a policy of careful observation and induction of labour only for fetal distress may be used, but there is anxiety that classical

markers of fetal distress may not be adequate in this setting.

Intrahepatic cholestasis of pregnancy recurs in up to 60 to 80 per cent of subsequent pregnancies and is associated with a late increased incidence of gallstones. Oral contraceptives should be used with caution, although early reports of liver abnormalities were predominantly with high-dose pills and combined low-dose preparations may cause fewer problems.

Acute fatty liver of pregnancy

Acute fatty liver of pregnancy, a microvesicular steatosis during the last trimester of pregnancy, was first adequately described by Sheehan in 1940. Occurring in 1 in 14 000 pregnancies between the 34th and 36th weeks, it is more common in primigravidae, with male fetuses, and with twin pregnancies. Up to 40 per cent may have associated features of pre-eclampsia, with peripheral oedema, hypertension, and proteinuria, and occasionally the **HELLP** (Haemolysis, Elevated Liver enzymes and Low Platelet count) syndrome (see below). Acute fatty liver of pregnancy occurs only rarely before the third trimester, but postpartum presentations are well recorded. Initial symptoms are of headache, fatigue, nausea, and vomiting with abdominal discomfort. In severe cases jaundice develops within 14 days. This may progress to manifest all the features of acute liver failure, including coma, renal failure, and death. However, less severe cases are now described more commonly, with recent series reporting a 10 to 20 per cent fetal and maternal mortality.

The cause of many cases of acute fatty liver of pregnancy may be a fetal–maternal interaction resulting from abnormalities of mitochondrial fatty acid oxidation. Schoeman first reported a case of acute fatty liver of pregnancy associated with a defect of fatty acid oxidation in the fetus. More recently a mutation has been identified involving substitution of glutamine for glutamic acid at amino acid residue 474 of the alpha subunit of long-chain 3-hydroxyacyl-CoA dehydrogenase, an enzyme that forms one component of a trifunctional protein catalysing the last three steps in the b-oxidation of fatty acids within mitochondria. Heterozygote mothers carrying fetuses with long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency may develop either acute fatty liver of pregnancy or HELLP syndrome in up to 80 per cent of cases. Mitochondrial oxidation of fatty acids is already impaired during pregnancy, mediated by oestrogen and progesterone, and long chain 3-hydroxyacyl metabolites produced by the fetus can accumulate and be toxic to the liver. Because presentation of children with long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency may occur late after birth, with non-ketotic hypoglycaemia and sudden infant death, the early identification and treatment of such cases is important. For this reason diagnostic molecular testing is recommended on all mothers and offspring of acute fatty liver of pregnancy and HELLP pregnancies.

The important differential diagnosis is between acute viral hepatitis with liver failure, acute fatty liver of pregnancy, and hypertension-associated liver dysfunction of pregnancy. The presence of features of pre-eclampsia in up to 40 per cent may make this distinction difficult. In acute fatty liver of pregnancy transaminases are elevated to less than 500 iu/litre, in contrast to viral hepatitis where values are usually higher. Hypoglycaemia is more common in acute fatty liver of pregnancy than in pre-eclampsia, and a blood film commonly showing neutrophilia, normoblasts, thrombocytopenia, target cells, and giant platelets may also help to make the diagnosis of acute fatty liver of pregnancy. Prothrombin and partial thromboplastin times are prolonged with low antithrombin III levels. Although hepatic steatosis in acute fatty liver of pregnancy may be detected by ultrasonography, CT scanning is more sensitive with attenuation values less than half normal and less than spleen, the reverse of normal. Hyperuricaemia is present in 80 per cent, but is not pathognomonic as this may also be found in pre-eclampsia.

Despite all these biochemical and radiological tests, and because differentiation of acute fatty liver of pregnancy from acute viral hepatitis and hypertension-associated liver diseases is important in assessing the prognosis for future pregnancies, a liver biopsy may be necessary to distinguish between these three diagnoses. Histology demonstrates a microvesicular fat deposition with rare hepatocyte necrosis and minimal inflammation, a pattern similar to that seen in Reye's syndrome, tetracycline and sodium valproate hepatotoxicity, Jamaican vomiting sickness due to a toxin in unripe ackee fruit, some urea cycle enzyme deficiencies, and defects in mitochondrial fatty acid oxidation.

Other obstetric causes of renal failure that are associated with minor changes in liver blood tests rarely need to be considered. These include thrombotic thrombocytopenic purpura and haemolytic uraemic syndrome. The former is particularly associated with neurological signs, including epileptic fits, without evidence of disseminated intravascular coagulation, whereas haemolytic uraemic syndrome may occur postpartum and with microangiopathic anaemia.

After the diagnosis of acute fatty liver of pregnancy has been established, the most important component of management is early delivery of the fetus. A policy of careful monitoring has been proposed for the mildest cases, but extreme caution is required as deterioration can be sudden and unpredictable. Vaginal delivery can be tried first, but caesarean section will usually require general anaesthesia as a spinal anaesthetic in the presence of coagulation deficits may be dangerous. Hypoglycaemia is prevented by intravenous dextrose infusion; aggressive correction of coagulation abnormalities with fresh frozen plasma, cryoprecipitate, and antithrombin III has also been recommended. Liver transplantation has been used (very rarely) for cases failing to respond to early delivery and intensive care.

The risk of recurrence of acute fatty liver of pregnancy is very low, the few recorded cases most probably being due to associated recurrent metabolic defects in the fetus.

Hypertension-associated liver diseases of pregnancy

Hypertension occurs in up to 8 per cent of pregnancies and is the most common cause of maternal mortality in developed countries. Abnormalities of liver blood tests have been described in three clinical syndromes associated with hypertension in pregnancy: pre-eclampsia/eclampsia, spontaneous hepatic rupture, and the HELLP syndrome.

Pre-eclampsia/eclampsia

The development of peripheral oedema, proteinuria, and hypertension (blood pressure above 140/90 mmHg or an increase of 30 mmHg systolic and 15 mmHg diastolic pressure above pre-existing levels) occurs in up to 5 per cent of all deliveries, more commonly with primigravidae, extremes of age, multiple gestations, and family history. Pre-eclampsia is discussed in detail in section 13.04, but there are abnormalities of liver biochemistry in up to 25 per cent of mild cases and up to 80 per cent of those with severe disease, i.e. those with renal impairment, visual disturbance, headache, fits, and the onset of eclampsia. The usual abnormality is a rise in transaminases, with jaundice occurring only in the most severe cases. The alanine aminotransferase levels are usually less than 150 iu/litre, lower than in acute fatty liver of pregnancy, and bilirubin is less than 100 µmol/litre. Changes in coagulation parameters, with elevated D-dimers, reflect the intravascular activation and consumption of clotting factors. Antithrombin III levels are low. Liver histology in the early stages shows few changes except for deposition of fibrinogen within sinusoids and the space of Disse. Blockage of sinusoids by fibrin may progress to frank infarction of hepatic parenchyma, when the low levels of coagulation factors may then predispose to haemorrhage into these infarcted areas. This combination of infarcts and associated haemorrhage, often covert and without apparent clinical consequence, may be demonstrated on ultrasound or CT scan of the liver.

The management of liver dysfunction in this context is that for pre-eclampsia, with correction of hypertension and coagulation defects, early delivery being the most important aspect. Anticonvulsant prophylaxis must be considered. Very close fetal monitoring is important. Liver biochemistry improves after delivery, but a late cholestatic phase with rise in alkaline phosphatase and g-glutamyl transpeptidase is common.

Spontaneous hepatic rupture

Spontaneous rupture of the liver is fortunately rare, occurring in 1 in 100 000 deliveries. In 80 per cent of cases liver haematomas, segmental or larger infarcts, and rupture occur in patients with severe pre-eclampsia and eclampsia: the remainder occur in association with acute fatty liver of pregnancy, or underlying hepatic adenomata, hepatocellular carcinoma, haemangioma, choriocarcinoma, or liver abscess.

The classical presentation is with right upper quadrant pain, nausea, vomiting, and hypotension in an older woman with severe pre-eclampsia during the third trimester. Right upper quadrant tenderness may be associated with frank peritonism where rupture into the peritoneum has occurred. The diagnosis can be confirmed by ultrasonography or CT scanning. Current treatment is conservative if possible, starting with angiography and hepatic artery embolization, proceeding to laparotomy, use of collagen meshes, and hepatic artery ligation if necessary. Successful orthotopic liver transplantation has also been recorded, but associated coagulation abnormalities are difficult to manage. The baby should be delivered by caesarean section. Successful subsequent pregnancies are recorded, as is recurrent haemorrhage, hence careful monitoring of any future pregnancy is necessary.

Haemolysis, elevated liver enzymes, and low platelet count (HELLP syndrome)

Weinstein first described a syndrome of haemolysis, elevated liver enzymes, and a low platelet count in patients with severe pre-eclampsia or eclampsia. HELLP syndrome occurs in 10 per cent of pregnancies with severe pre-eclampsia. Strict criteria should be used for the diagnosis: haemolysis with a characteristic peripheral blood smear, serum lactate dehydrogenase greater than or equal to 600 U/litre, serum aspartate aminotransferase greater than or equal to 70 U/litre and platelet count less than 100×10^9/litre. An abnormal peripheral blood smear with fractured red blood cells (schistocytes, echinocytes, spherostomatocytes) is sensitive but not specific for HELLP, and the elevated serum lactate dehydrogenase is another indicator of haemolysis. Some cases display only one or two of the above criteria and have been labelled as having partial HELLP: these have been shown to follow a less severe clinical course. HELLP syndrome has also been classified (by Martin and colleagues) according to platelet count: class 1 platelet counts less than or equal to $50\,000 \times 10^9$/litre, class 2 more than 50 000 to less than or equal to $100\,000 \times 10^9$/litre, and class 3 more than $100\,000 \times 10^9$/litre, which is equivalent to partial HELLP.

The reason why some cases with severe pre-eclampsia progress to HELLP syndrome is not clear. Whilst the haemolysis is clearly related to intravascular deposition of thrombin and mechanical fracture of red cells, and there is fibrin deposition obstructing hepatic sinusoids, the factors causing particular damage to the hepatic microcirculation are unknown. Overt disseminated intravascular coagulation is not usually a major component; when defined as hypofibrinogenaemia (< 300 mg/dl) and elevated D-dimers (> 40 µg/ml) it occurred in 21 per cent of a series of 442 cases. Compared with other cases with severe pre-eclampsia, those with HELLP syndrome tend to be older, Caucasian, and multiparous.

Symptoms usually start in the second or third trimester, with 15 per cent starting prior to 26 weeks and 30 per cent only developing symptoms after delivery. HELLP syndrome has protean manifestations but universal early symptoms include malaise and fatigue, followed by nausea, vomiting, and headache shortly thereafter (Table 3). Epigastric and right upper quadrant pain are ominous signs, particularly when accompanied by right shoulder tip pain. Weight gain and peripheral oedema are found in 50 per cent, with diastolic blood pressure greater than or equal to 90 mmHg in all but a small minority.

The fall in platelet count and rise in transaminases usually reach their nadir in the first 2 days postpartum. Aspartate aminotransferase and lactate dehydrogenase are elevated in unison, along with other markers of hepatocellular or sinusoidal cell dysfunction, glutathione-*S*-transferase and hyaluronic acid.

Maternal complications associated with HELLP syndrome occur in up to 50 per cent of cases. Blood transfusion to correct hypovolaemia, anaemia, or coagulopathy is required in 50 per cent, with features of disseminated intravascular coagulation in 25 per cent, and pleural effusions or pulmonary oedema in 15 to 20 per cent. Renal failure due to acute tubular necrosis may occur in 3 to 8 per cent. Obstetric complications are also associated with the degree of fall in platelet counts, with placental abruption (16 per cent) and wound haematomas after caesarean section the most prevalent. Eclampsia is approximately two to three times more common in patients with class 1 HELLP than in those with milder varieties, and consistent with this the maternal mortality was 1.5 per cent in a large series, with a perinatal mortality in Martin *et al*.'s tertiary referral practice of 119/1000 infants. Overall perinatal mortality is strongly related to time of delivery, with rates as high as 30 per cent in some series, although this may be due to case selection. Mortality was 9.5 per cent in a large series assessing class 1, 2, and 3 cases. Preterm infants born before 32 weeks from mothers with HELLP syndrome had a higher frequency of severe intraventricular haemorrhage than other preterm infants. Birth weights tend to be lower in severe HELLP than in pre-eclampsia alone.

Preterm patients with HELLP syndrome should be treated at a referral centre with appropriate obstetric, anaesthetic, and haematological support. Management of the coexisting pre-eclampsia is crucial, with seizure prophylaxis using magnesium sulphate and blood pressure control with labetolol, ketanserin, or hydrallazine if blodd pressure is over 160/105 mmHg. Antenatal corticosteroids enhance fetal lung maturity if the pregnancy is of less than 32 weeks gestation. Careful fluid resuscitation is required to prevent volume overload, particularly in the presence of renal impairment, as is very close fetal monitoring.

Aside from signs of significant fetal distress, indications for urgent delivery include persistent severe right upper quadrant or shoulder tip pain, often associated with hypotension and thrombocytopenia, which indicate possible liver haematoma or impending rupture. Early delivery, at the safest time for mother and fetus, is strongly recommended in the absence of treatment that unequivocally improves the haematological abnormalities and both maternal and fetal outcome. Although arguments have been put forward for a more conservative approach in patients with mild disease—with careful monitoring of coagulation profiles, fetal growth and well being, and with the timing of delivery depending on clinical judgement—this management strategy is not without risk and has not been examined in a randomized controlled trial.

Attempts have been made to improve the outcome of HELLP with medical management alone in an effort to buy time to enhance fetal maturity and to improve the mother's clinical condition prior to delivery. Although plasma volume expansion, antithrombotic agents, plasma exchange, and corticosteroids have all been advocated, no therapies have been shown to allow safe deferral of delivery and improve outcome. In one study dexamethasone given predelivery resulted in a modest prolongation of pregnancy, whereas two other trials have shown significant improvements in haematological and biochemical parameters when given postpartum. In a trial of invasive haemodynamic monitoring, plasma volume expansion and afterload reduction, laboratory parameters improved with prolongation of gestation by 21 days, but no significant change in perinatal mortality.

Recurrence of HELLP syndrome in subsequent pregnancies is uncommon: it recurred in only 5 per cent of a series of 139 normotensive women after an index pregnancy with HELLP syndrome, despite 25 per cent developing pre-eclampsia. In hypertensive cases a further pregnancy was associated with pre-eclampsia in 70 per cent and HELLP syndrome in 8 per cent.

Differential dagnosis of jaundice during pregnancy

Many patients with acute fatty liver of pregnancy have signs of pre-eclampsia and such cases may be part of a clinical syndrome that includes hypertension-associated liver diseases. Evidence for this possibility includes the finding of microvesicular hepatic steatosis in cases with pre-eclampsia; indeed one study found fat deposition by special staining in all 41 cases studied. Histological evidence of both pre-eclampsia and acute fatty liver of pregnancy has also been demonstrated in some cases, and pregnancies associated with pre-eclampsia have been followed in the next by HELLP syndrome. Despite this there are usually features in the clinical history or laboratory findings (Table 4) that allow discrimination between these diagnoses. The size of the liver, degree of hyperbilirubinaemia, abnormalities on peripheral blood film, presence of hypoglycaemia, and disseminated intravascular coagulation are the most discriminatory tests. The differential diagnosis of jaundice and abnormal liver blood tests differs in the three trimesters of pregnancy (Table 5).

Liver diseases not specific to pregnancy

Budd–Chiari syndrome

Thrombosis in one or more hepatic veins has an increased prevalence during pregnancy and in those on the oral contraceptive pill. This relates to low antithrombin III levels and may be more common in those with an underlying procoagulant state or presence of antiphospholipid antibodies. Right upper quadrant pain, hepatomegaly, and maternal ascites should suggest the diagnosis, with confirmation by ultrasonography or hepatic venous angiography. Although hepatic venous balloon dilatation or insertion of a transjugular intrahepatic stent shunt has been recommended for Budd–Chiari syndrome, there are few data on their use during pregnancy. Maternal mortality remains very high.

Cholelithiasis

Gallbladder sludge and gallstones develop in 31 per cent and 9 per cent of pregnancies respectively, although most resolve thereafter. Prior use of oral contraceptives, increased cholesterol synthesis, reduced cholesterol carriage in bile, and impaired gallbladder motility all accounted for the increased lithogenicity of bile. Symptomatic gallstone disease should be managed in the usual way. Magnetic resonance cholangiopancreatography can accurately detect common bile duct stones without exposing the fetus to radiation, with endoscopic sphincterotomy and/or stent placement reserved for those in whom they are detected. In most cases surgery can be deferred until after delivery.

Viral hepatitis

Acute viral hepatitis is the most common cause of jaundice during pregnancy (Table 5), with no specific change to presentation, clinical course, or outcome for acute hepatitis A, B, cytomegalovirus, or Epstein–Barr virus infection.

Transmission of virus from a mother with acute hepatitis B to her offspring occurs in 50 per cent of cases, rising to 70 per cent when hepatitis starts in the third trimester. Transmission of virus from mothers with chronic hepatitis B carriage is less common, but depends on the level of viral replication. The rate is at least 90 per cent in those who are hepatitis B virus DNA positive, and who are usually hepatitis B e antigen positive, as is most common in Orientals. Following vertical transmission up to 80 per cent of offspring become chronic HBsAg carriers. Transmission of hepatitis B can be effectively interrupted by use of hepatitis B immunoglobulin at birth, with hepatitis B virus vaccination within 7 days and at 1, 2, and 12 months.

Transmission of hepatitis C from chronic carriers occurs in up to 8 per cent of cases, being higher in those with high maternal viral load. Antihepatitis C virus seroconversion of infants following transmission may take 6 to 12 months to appear, but detection of hepatitis C virus RNA by polymerase chaine reaction allows detection of transmission sooner.

Acute hepatitis E is due to an RNA virus and occurs, often in waterborne epidemics, predominantly in the Middle and Far East. In pregnancy it is associated with a mortality of up to 20 per cent due to development of acute liver failure during the third trimester, but transmission to offspring has not been recorded.

Variceal haemorrhage

Changes in splanchnic haemodynamics, increased cardiac output and azygous blood flow, and an increase in circulating blood volume have all been suggested as risk factors for variceal bleeding during pregnancy. Evidence for this remains controversial, although recent large series with non-cirrhotic portal hypertension report a haemorrhage rate of 13 per cent. Treatment of variceal bleeding during pregnancy should be with conventional endoscopic techniques, with use of transjugular intrahepatic stent shunts or surgical shunts reserved for rescue therapy.

Liver tumour during pregnancy

The first presentations of focal nodular hyperplasia, hepatic adenoma, hepatocellular carcinoma, and cholangiocarcinoma have all been reported during pregnancy. Adenomas, in some cases related to prior oral contraceptive use, may undergo vascular engorgement during pregnancy and rupture has been reported. Secondary tumours, including hepatic choriocarcinoma and ovarian teratomas, may also rupture.

Pregnancy following orthotopic liver transplantation

Fertility returns quickly following liver transplantation. Pregnancy does not alter the risks of cellular rejection, but immunosuppressive drug toxicity needs to be carefully monitored. Azathioprine may cause neonatal pancytopenia, and cyclosporin A is associated with a 40 per cent incidence of hypertension, which may be lower with Tacrolimus.

Pregnancy during chronic liver disease

Most patients with established cirrhosis are infertile, but a few remain fertile, although with a high rate of prematurity, low-birthweight babies, and stillbirths. There is little evidence that pregnancy results in deterioration in liver dysfunction in patients with cirrhosis, and improvement of inflammatory activity occurs in some cases of autoimmune chronic active liver disease. There is no increased rate of relapse after delivery. Patients with treated Wilson's disease are able to conceive and successful pregnancies whilst taking D-penicillamine or trientine have been reported.

Pregnancy during gastrointestinal disease

Only a few gastrointestinal diseases occur with altered frequency during pregnancy.

Gastro-oesophageal reflux

Symptomatic gastro-oesophageal reflux is present at some stage in up to 80 per cent of pregnancies. It is mainly due to a reduced lower oesophageal sphincter pressure rather than elevated intra-abdominal pressure from a gravid uterus. Treatment with antacid is recommended, with avoidance of H_2 antagonists or proton pump inhibitors unless symptoms and complications of gastro-oesophageal reflux outweigh potential drug toxicity. Acid-pepsin reflux combined with vomiting in early pregnancy may precipitate haematemesis, occasionally with a Mallory–Weiss tear, for which management should be as in the non-pregnant state. Upper gastrointestinal endoscopy is a safe procedure during pregnancy.

Inflammatory bowel disease

Stable inactive ulcerative colitis and Crohn's disease do not affect fertility, are not associated with increased fetal risk, and disease control is not impeded by pregnancy. There are few data on the effect of drug therapy on fertility, although sperm counts may be reduced in men on salazopyrine. The risk of relapse of inflammatory bowel disease during pregnancy has been assessed at between 30 and 50 per cent, but this is no higher than comparable non-pregnant control groups. Folate and iron supplementation are recommended, with regular monitoring of nutritional status.

Active inflammatory bowel disease is associated with involuntary infertility, and when very severe it is prudent to recommend deferring any attempt to conceive. Increased fetal loss may occur when active inflammatory bowel disease is first manifest during pregnancy, with recent reports suggesting that the site of disease activity (colonic or small bowel) does not affect outcome. Most studies have demonstrated that corticosteroids, sulphasalazine, and 5-aminosalicylic acid preparations are safe to use during pregnancy. Colonoscopy, in expert hands, can be performed during pregnancy without risk, although it is often possible to defer this procedure.

Acute appendicitis

The most common non-obstetric emergency requiring surgery, acute appendicitis occurs in 1 in 2500 to 1 in 3500 pregnancies. It is not clear if reports of a more aggressive clinical course reflect delays in diagnosis or reporting bias. Clinical management is similar to that of the non-pregnant case: surgery must not be deferred as the frequency of prematurity and perinatal mortality may be increased if perforation occurs.

Coeliac disease

Women with untreated coeliac disease have a markedly increased risk of abortion and low-birthweight babies, which can be reversed following institution of a gluten-free diet. Screening for coeliac disease should be considered in women with a previous history of abortion or unfavourable pregnancy outcomes.

Further reading

Audibert *et al*. (1996). Clinical utility of strict criteria for the HELLP syndrome. *American Journal of Obstetrics and Gynecology* **175**, 460–4.

Ibdah *et al*. (1999). A fetal fatty-acid oxidation disorder as a cause of liver disease in pregnant women. *New England Journal of Medicine* **340**, 1723–31.

Knox T, Olans L (1996). Liver disease in pregnancy. *New England Journal of Medicine* **335**, 569–76.

Kochhar R *et al*. (1999). Pregnancy and its outcome in patients with noncirrhotic portal hypertension. *Digestive Disease Science* **44**, 1356–61.

Korelitz (1989). Inflammatory bowel disease and pregnancy. *Gastroenterology Clinics of North America* **27**, 213–24.

Martin JN *et al*. (1999) The spectrum of severe preeclampsia: comparative analysis by HELLP (hemolysis, elevated liver enzyme levels, and low platelet count) syndrome classification. *American Journal of Obstetrics and Gynecology* **180**, 1373–84.

Martinelli P *et al*. (2000). Coeliac disease and unfavourable outcome of pregnancy. *Gut* **46**, 332–5.

Mayberry J, Weterman IT (1986). European survey of fertility and pregnancy in women with Crohn's disease; a case control study by the European Collaborative group. *Gut* **27**, 821–5.

Modigliani R (1997) Drug therapy for ulcerative colitis during pregnancy. *European Journal of Gastroenterology and Hepatology* **9**, 854–7.

Nicastri P *et al*. (1998). A randomised placebo-controlled trial of ursodeoxycholic acid and S-adenosylmethionine in the treatment of intrahepatic cholestasis of pregnancy. *British Journal of Obstetrics and Gynaecology* **105**, 1205–7.

Palma J *et al*. (1997). Ursodeoxycholic acid in the treatment of cholestasis of pregnancy: a randomized, double-blind study controlled with placebo. *Journal of Hepatology* **27**, 1022–8.

Sibai *et al*. (1993). Maternal morbidity and mortality in 442 pregnancies with hemolysis, elevated liver enzymes, and low platelets (HELLP syndrome). *American Journal of Obstetrics and Gynecology* **169**, 1000–6.

13.10 Diabetes in pregnancy

Michael D. G. Gillmer

Prior to the introduction of insulin treatment in 1921, diabetes was a rare complication of pregnancy with near 50 per cent maternal and fetal mortality. Within a decade of the introduction of insulin therapy, maternal mortality had fallen to between 2 and 3 per cent, but fetal mortality remained above 40 per cent until the 1950s despite recognition that 'rigid control' of diabetes was vital to achieve an optimal pregnancy outcome. This concept has remained central to the management of the disease in pregnancy, with hindsight suggesting that the early poor fetal outcome was due to incomplete understanding of the pathophysiology of the condition, and to a lack of suitable technology for assessing adequate diabetic control.

Metabolic changes in pregnancy

Pregnancy induces substantial alterations in carbohydrate, lipid, and amino acid metabolism, which have been described as a combination of 'facilitated anabolism' and 'accelerated starvation'. From a teleological standpoint these changes appear to ensure the optimal availability of nutrients for both fetus and mother.

Carbohydrate metabolism

Fasting plasma glucose concentrations gradually decline during pregnancy by approximately 0.5 mmol/l, reaching a nadir in the third trimester. Postprandial glucose concentrations increase, despite a rise in both basal and stimulated insulin secretion. This appears to be due to peripheral insulin resistance induced by placental hormones, and to the effects of oestrogen and progesterone on the maternal pancreas.

Although insulin sensitivity appears to increase transiently during the first trimester of pregnancy in some women, there is thereafter a progressive decline which is reflected by an increased insulin:glucose ratio. Human placental lactogen, a polypeptide hormone, is one of the main causes of the insulin resistance that characterizes pregnancy. Other possible factors include increased fat stores, raised prolactin and free cortisol concentrations, sequestration of insulin by the placenta, and changes in insulin receptor affinity and number.

Serial glucose tolerance tests indicate a progressive decline in tolerance with advancing gestation. After an oral glucose load in late pregnancy there are higher peak plasma glucose concentrations, a delay in the rise to the peak concentration and an increase in the total area under the glucose tolerance curve compared with the non-pregnant state. Despite these changes pregnant women maintain efficient glucose homeostasis, but with slightly lower preprandial and higher postprandial plasma glucose concentrations following mixed meals than in non-pregnant women.

Although insulin does not cross the placental barrier, glucose crosses freely by a process of facilitated diffusion. Fetal exposure to maternal hyperglycaemia causes premature stimulation of the fetal b-cells of the pancreatic islets of Langerhans and results in fetal hyperinsulinaemia. This stimulates excessive fetal growth, leading to the macrosomia that characterizes the infant of the diabetic mother.

Lipid metabolism

Plasma concentrations of triglycerides, cholesterol, phospholipids, and free fatty acids all increase during pregnancy. During early pregnancy increased food intake coupled with moderate postprandial hyperinsulinism create ideal conditions for lipogenesis, so-called 'facilitated anabolism'. During late pregnancy food intake declines, insulin resistance is established, and in the presence of high circulating levels of human placental lactogen, lipolysis is enhanced during the fasting state, when there is also a significant increase in ketones, so-called 'accelerated starvation'. The increase in circulating free fatty acids concentrations is thought to have an important influence on maternal metabolism, providing an alternate source of maternal fuel at a time in pregnancy when fetal and maternal glucose needs are maximal.

Plasma cholesterol increases by approximately 25 per cent during pregnancy, a change which probably reflects increased synthesis and decreased catabolism. Lipoprotein triglyceride and cholesterol do not cross the placenta, but free fatty acids cross freely by simple diffusion.

Protein and amino acid metabolism

Amino acids are crucial for fetal development and fetal protein accumulation occurs rapidly in late pregnancy. Despite this there is an increase in maternal amino acid excretion in the third trimester, consisting mainly of the non-essential amino acids glycine, histidine, serine, and alanine. In addition, most amino acid concentrations fall in pregnancy, in particular ornithine, glycine, taurine, and proline, while the postprandial peak concentrations of leucine, isoleucine, serine, and alanine following a mixed meal in late pregnancy are lower than those observed in non-pregnant subjects. Starvation in pregnancy causes a two to three-fold rise in valine, leucine, and isoleucine, but a fall in alanine concentrations.

The concentration of most free amino acids is higher in fetal than in maternal plasma, indicating placental amino acid transfer against a concentration gradient.

Gestational diabetes

Gestational diabetes may be defined as '...carbohydrate intolerance of variable severity with onset or first recognition during the present pregnancy...' This definition includes not only those women in whom diabetes occurs transiently during pregnancy and regresses after delivery, but also those in whom type 1 diabetes arises *de novo* during pregnancy and persists long term.

Screening for gestational diabetes has traditionally involved performing glucose tolerance tests on all women with 'risk factors' or 'potential diabetic features'. However, these are present in 30 per cent or more women in most communities, but not all of those who develop significant glucose intolerance during pregnancy, meaning that many women are subjected to unnecessary tests, whilst others who develop gestational diabetes are missed. 'Risk factors' are therefore of limited value

for screening purposes.

Universal screening programmes based on blood glucose measurement have become popular (although controversial) in recent years. The American Diabetes Association (ADA) and American College of Obstetricians (ACOG) have both endorsed the use of a 50-g oral glucose load at 24 to 28 weeks gestation. Venous plasma glucose is measured an hour later and a value equal to or greater than 7.8 mmol/l (140 mg/dl) is recommended as the threshold for a full diagnostic oral glucose tolerance test (GTT). This screening procedure has been shown to be the most sensitive (79 per cent) and specific (87 per cent) of the screening tests available, but is probably only appropriate in populations with a high prevalence of diabetes or for those at increased risk, such as older women and those who are grossly obese. A simpler and more cost effective protocol involves a so-called 'timed' random blood glucose measurement at antenatal booking and at 28 weeks gestation, repeated whenever glycosuria occurs. A full glucose tolerance test is indicated if the plasma glucose exceeds 6 mmol/l in the fasting state or 7 mmol/l within 2 h of a meal. Screening by means of glycosylated haemoglobin or plasma protein measurements, including fructosamine, have proved to be too insensitive for use in pregnancy.

The glucose tolerance criteria for the diagnosis of gestational diabetes are controversial and this, together with the poor reproducibility of the test, may explain some of the inconsistent results of screening programmes. The American College of Obstetricians and Gynaecologists has retained the 100-g oral glucose tolerance test, but in Europe a 75-g load is used, following the World Health Organization's recommendation. It has been suggested that the criteria for 'impaired glucose tolerance' after a 75-g glucose tolerance test should be used for the diagnosis of 'gestational diabetes', but this is not universally accepted (see below). The WHO criteria are shown in Table 1, together with the ACOG standards for a 100-g oral glucose tolerance test, Oxford data from a study of the 75-g oral glucose tolerance test in 491 women at 28 to 34 weeks gestation, and data from a multicentre study of the Diabetic Pregnancy Study Group (DPSG) of the European Association for the study of Diabetes (EASD) involving 354 women in the third trimester of pregnancy.

It is apparent from these data that if gestational diabetes is diagnosed using the WHO criteria for impaired glucose tolerance (IGT), this will lead to increased diagnosis of the condition, as a significant number of women will have a 2-h venous plasma glucose in excess of the WHO limit of 8.0 mmol/l. Widespread acceptance of the WHO criteria for IGT as diagnostic of gestational diabetes could therefore explain the continued uncertainty about the adverse clinical effects of gestational diabetes on pregnancy outcome. It is suggested that the modified WHO criteria shown in Table 2 are used in clinical practice, and that a clear distinction is made between women with impaired glucose tolerance and true diabetes during pregnancy.

Medical management

In the early 1970s it was recognized that the perinatal mortality in diabetic women is positively correlated with the mean maternal blood glucose concentration during pregnancy. This finding, together with the observation that blood glucose concentrations in normal pregnant women rarely exceed 6 mmol/l, except during the hour after a meal, focused attention on the need for 'rigid control' of the maternal diabetes. Ideally, blood glucose concentrations should be measured preprandially and postprandially, as shown in Table 3, on at least 2 days each week, or more frequently if indicated, and maintained between 4 and 6 mmol/l. Measurements of glycosylated haemoglobin or fructosamine are used in many units to provide an indication of medium to long-term glycaemic control and can prove helpful, especially in non-compliant patients.

Congenital malformations

Type 1 insulin dependent diabetes preceding pregnancy is associated with a significant increase in the risk of major congenital anomalies of between 7 and 14 per cent. The precise aetiology remains obscure in most cases, but the frequency is undoubtedly increased in women with poor diabetic control preceding pregnancy and during the first trimester. The incidence also varies depending on the definitions applied in diagnosing major malformations (Table 4).

There is a three to five-fold increase in the incidence of neural tube, cardiac, and renal anomalies which account for more than a half of current perinatal deaths and are also an important cause of avoidable long-term morbidity. As the organ systems commonly affected in diabetes are all fully formed by 9 weeks gestation, that is 7 weeks after the first missed menstrual period, it is vital that all women in the reproductive age group are advised that they must make serious efforts to achieve optimal diabetic control before planning a pregnancy and that this should be maintained throughout the period of embryogenesis. In addition, all diabetic women should be advised to take folate supplements for at least 4 weeks prior to conception to reduce the risk of delivering a child with a neural tube defect.

Team care

'Team care' is an essential part of the modern management of the pregnant diabetic woman. The most important member of the team is the woman herself: she obviously has responsibility for her diabetes on a day-to-day basis, and usually has the clearest understanding of how optimal glycaemic control can be achieved. She should ideally attend a joint diabetic–antenatal clinic where she can be seen by specialist diabetic nurses, midwives, dietitians, and medical staff including an obstetrician, a physician, and a neonatal paediatrician with a special interest in this condition.

It is important to see these patients as early as possible in pregnancy, after which the frequency of clinic visits will depend on several factors, including the blood glucose concentrations achieved and the occurrence of diabetic or obstetric complications. An average two or three-fold increase in insulin requirements occurs during pregnancy. It is therefore preferable to see all diabetic women at least every 2 weeks until 34 weeks gestation, and then weekly until delivery, as this facilitates the frequent alterations of insulin dose that need to be made as pregnancy progresses and also ensures adequate dietary advice. If control is poor and more frequent advice is required, this can usually be given by telephone contact.

Diet and insulin therapy

The management of women who are diagnosed as gestational diabetics depends on their preprandial and postprandial blood glucose concentrations. If these are 6 to 8 mmol/l, a high fibre isocaloric diet is advised initially, and the woman retested. If the preprandial plasma glucose concentrations remain above 6 mmol/l on this diet or if they initially exceed 8 mmol/l then insulin therapy is commenced, using a long-acting preparation at bedtime in the first instance. Preprandial short-acting insulin is added before meals if the postprandial values remain above 6 mmol/l. Oral hypoglycaemics are not used because they cross the placenta and stimulate the fetal pancreatic b-cells causing fetal hyperinsulinaemia, the pathological process that insulin treatment aims to avoid. Non-pregnant diabetics using oral hypoglycaemic agents should ideally be changed to insulin treatment when planning a pregnancy.

Human insulin is preferred, as this produces least antibodies and reduces the theoretical risks of fetal b-cell damage or macrosomia due to the transplacental passage of injected insulin bound to antibody.

The form of insulin therapy that is currently most widely used for the control of type 1 diabetes in pregnancy is a mixture of short-acting insulin three times daily before meals combined with an intermediate or long-acting injection at bedtime. However, twice daily injections of short and intermediate-acting insulin remain popular. Fixed ratio insulin mixtures have a limited role, but in practice the patient's prepregnancy insulin regimen should only be changed if it proves impossible to achieve the desired standard of control without doing so. The pregnant diabetic is particularly prone to overnight ketoacidosis and the continuous subcutaneous insulin infusion (CSII) pump is no longer considered appropriate in pregnancy as disruption of the infusion through pump failure, catheter blockage, or disconnection can rapidly lead to ketoacidotic coma, with the risk of fetal or even maternal death.

Pregnancy is characterized by a decline in fasting plasma glucose concentrations and a plentiful supply of alternate substrates for energy requirements, including ketones derived from the b-oxidation of free fatty acids. Hypoglycaemia is therefore rare, and unlike hyperglycaemia does not appear to have any demonstrable adverse effect on the fetus. However, because of endeavours to achieve very tight diabetic control, pregnant diabetics are at increased risk of hypoglycaemia. They should therefore be provided with glucagon that can be administered by a third party in the event of severe hypoglycaemia.

Management of diabetes during and after labour

It is important to maintain normoglycaemia during labour in order to reduce the risk of neonatal hypoglycaemia. This is most easily achieved using combined insulin and dextrose infusions. Dextrose 10 per cent solution is infused at 100 ml/h and blood glucose measurements are made every hour. Insulin (6 units in 60 ml normal saline) is administered simultaneously, at an initial rate of 1 unit (10 ml) /h using an infusion pump. The insulin infusion rate is doubled or halved as necessary to maintain the blood glucose concentration between 4 and 6 mmol/l. During labour the insulin requirement may fall dramatically, presumably because of the increased

glucose demand due to uterine work, and it is frequently necessary to switch the insulin infusion off towards the end of the first stage.

After delivery the insulin infusion rate must be halved to prevent hypoglycaemia as there is a rapid increase in insulin sensitivity following the delivery of the placenta. It is also essential to return to the prepregnancy insulin dose immediately the patient resumes her normal diet: profound hypoglycaemia can occur if the dose required prior to delivery is administered at this time.

Retinopathy

Rapid reduction of blood glucose concentrations has been shown to accelerate diabetic retinopathy in both pregnant and non-pregnant subjects. There is also evidence that pregnancy and hypertension complicating pregnancy may act as independent risk factors for the progression of diabetic retinopathy. Formal retinal assessment, with dilated pupils, should therefore be performed before pregnancy so that improved diabetic control can be achieved over 3 to 9 months before a planned conception. This should avoid the need for acute improvement of the blood glucose concentrations in early pregnancy and thus minimize the risk of exacerbating proliferative retinopathy. All women should have a full ophthalmic assessment in early pregnancy to assess their retinal state and determine the possible need for laser therapy.

Nephropathy

Overt nephropathy is associated with various pregnancy complications, including pre-eclampsia, growth retardation, and fetal distress, but there is little evidence to suggest that pregnancy will hasten the progression of overt nephropathy to endstage renal failure. Patients seeking advice about pregnancy should therefore be warned that although their renal disease may have an adverse effect on pregnancy, which could necessitate prolonged hospitalization and premature delivery, possibly by caesarean section, there is usually no need to avoid or terminate pregnancy.

Obstetric management

Antenatal assessment of the fetus

Accurate information about the duration of pregnancy, fetal growth, and fetal well being are vital in the management of the pregnant diabetic. Technological developments, particularly in the use of diagnostic ultrasound, have revolutionized fetal assessment and become central to the modern obstetric management of diabetes in pregnancy.

The fetal crown–rump length should be measured during the first trimester to confirm the duration of pregnancy. In some women this technique has identified 'early growth delay' in which the fetal crown–rump length measurement is smaller than expected from the gestational age. This condition is associated with an increased rate of congenital malformations and poor fetal growth, and is thought to be due to 'less-than-optimal' metabolic compensation in early pregnancy.

A biparietal diameter measurement is also performed in the mid-trimester, ideally at 16 weeks gestation, to provide additional information about gestational age. Blood for serum a-fetoprotein should also be taken at this gestation, both to screen for neural tube defects and as part of the 'triple test' used to screen for Down's syndrome. In assessing the result of these investigations it must, however, be borne in mind that the serum a-fetoprotein and unconjugated oestriol concentrations observed in diabetic pregnancy are lower than those in non-diabetic women and a specific algorithm is required for the interpretation of these screening tests in women with type 1 diabetes.

A detailed fetal ultrasound examination to exclude congenital anomalies is performed between 18 and 20 weeks, so that termination of the pregnancy can be considered if appropriate. Further cardiac anomaly scans may be performed at 28 and 34 weeks gestation. Serial studies of growth based on measurements of the fetal head and abdominal circumferences in the second trimester provide the best means of identifying those pregnancies in which the fetus is becoming macrosomic, when it is still possible to institute optimal metabolic control and reduce the likelihood of this complication. However, although an association between birthweight and maternal blood glucose concentrations has been demonstrated during the third trimester of pregnancy, the cause of fetal macrosomia in diabetes is still uncertain and there are many examples of women who deliver infants with birthweights above the 97th centile despite excellent metabolic control in late pregnancy. However, the situation is further complicated by the finding of a two-fold increase in small-for-dates infants (birthweight below the 10th centile) in diabetics with very tight control (mean blood glucose concentration of less than 4.8 mmol/l), indicating that excessively tight blood glucose control may have a deleterious effect on the growth of the diabetic fetus, and possibly on its development.

Obstetric complications of diabetes in pregnancy

Proteinuric hypertension occurs approximately twice as often in diabetics as in normal women. Serum urate and creatinine concentrations should therefore be measured at every antenatal visit and 24-h urine protein concentrations from 24 weeks gestation. These provide the earliest biochemical evidence of proteinuric pre-eclampsia and also serve to clarify those blood pressure changes in late pregnancy which are due to pre-existing essential hypertension. Although the reason for the increased incidence of pre-eclampsia in diabetics is unknown, a link with glycaemic control has been established and the incidence of this complication is reduced with optimal diabetic control.

Polyhydramnios is one of the hallmarks of diabetic pregnancy, and occasionally the presenting feature in gestational diabetes. The cause of this complication, which has an overall incidence of approximately 15 per cent, remains uncertain, but may be due to an osmotic diuresis induced in the fetus by feto–maternal hyperglycaemia. This would be in keeping with the fact that polyhydramnios generally lessens as diabetic control improves.

Premature labour is more frequent in diabetic pregnancy and may, in some instances, be due to underlying polyhydramnios. Conventional management with intravenous b-sympathomimetic agents causes hepatic glycogenolysis and insulin resistance, predisposing to hyperglycaemic ketoacidosis. This treatment is therefore potentially hazardous in diabetic women and should be avoided whenever possible or used with extreme caution, even in non-insulin dependent patients. Use of glucocorticoids in diabetic pregnant women may necessitate the administration of very high doses (up to 30 units/h) of intravenous insulin to maintain normoglycaemia.

Fetal well being and maturity

Unexplained intrauterine death during the last 3 to 4 weeks of pregnancy has been recognized as a considerable problem in the management of diabetic pregnancy since the preinsulin era. However, the so-called 'fetal biophysical profile', a real time ultrasound technique, has revolutionized the late pregnancy management of this condition and made it unnecessary to admit diabetic women routinely for daily monitoring in late pregnancy. These assessments, which should be performed at least weekly from 36 weeks gestation, have also made it possible to prolong diabetic pregnancies to near term or beyond.

Antenatal Doppler ultrasound assessments have also been used widely in diabetic pregnancy in recent years, the results providing the reassurance necessary to prolong uncomplicated diabetic pregnancies beyond term, but unlike the biophysical profile have not proved helpful in predicting fetal demise in diabetic women, unless the pregancy is complicated by fetal growth retardation.

Timing of delivery

Poorly controlled diabetes is associated with fetal pulmonary and hepatic immaturity, predisposing to the neonatal respiratory distress syndrome and jaundice. The optimal time for delivery in uncomplicated diabetic pregnancy appears to be in the 39th week (273 days). Despite this, some authors have advocated deferring delivery until 40 weeks or later as this allows a larger number of women to enter labour spontaneously. This policy is associated with a higher incidence of macrosomic and stillborn babies and has not been shown to have any significant benefit.

Management of labour

One of the main aims in management of the pregnant diabetic woman is to achieve a spontaneous vaginal delivery. Elective caesarean section may be indicated when there is significant fetal macrosomia, with abdominal circumference greatly in excess of the head circumference, fetal malpresentation, or a history of a previous

caesarean section. Continuous fetal heart rate and contraction monitoring is advised because of the increased incidence of fetal distress. Pain relief in labour is particularly important because painful uterine contractions cause catecholamine release, leading to glycogenolysis and hyperglycaemia. Epidural anaesthesia is ideal but not vital, especially in the multiparous patient who may have a rapid and uncomplicated labour. If intravenous fluids are required for 'preloading' prior to insertion of an epidural or for the administration of oxytocin, it is essential that 0.9 per cent saline or Hartmann's solutions and not dextrose are used to avoid fetal hyperglycaemia, which predisposes to neonatal hypoglycaemia.

Efforts to predict the risk of shoulder dystocia have been conspicuously unsuccessful to date. This possibility must always be considered when the abdominal circumference exceeds the head circumference, especially if the fetus is macrosomic.

The neonate

Insulin is present in the human pancreas from 11 weeks gestation, and although the pancreatic response to insulin secretogogues is sluggish in normal infants, fetal exposure to high concentrations and large fluctuations of glucose and amino acids during poorly controlled diabetic pregnancy appears to produce premature maturation of the fetal b-cells. This causes hyperinsulinaemia, which predisposes to neonatal hypoglycaemia that may occur during the first 24 h after delivery, when high circulating insulin concentrations inhibit both glycogenolysis and lipolysis, thus depriving the infant of alternative energy sources.

Other neonatal problems include, the respiratory distress syndrome, polycythaemia, jaundice, renal vein thrombosis, hypocalcaemia, hypomagnesaemia, and cardiomyopathy, all of which appear to be related directly or indirectly to fetal hyperinsulinaemia. The incidence and severity of neonatal complications is closely related to diabetic control during pregnancy, and the infant of the well-controlled diabetic mother does not usually require admission to a special care nursery, unless a problem arises after delivery.

The puerperium and contraception

Diabetics are at increased risk of wound infection following surgery and prophylactic antiobiotics are therefore advised following both elective and emergency Caesarean section or operative vaginal delivery.

Breast feeding is encouraged, but as this reduces the insulin requirement by approximately 25 per cent an appropriate reduction must be made once lactation is established. Women who choose not to breast feed or in whom breast feeding is unsuccessful should resume their prepregnancy insulin dose after delivery.

All diabetic women should be seen for a 6-week postnatal examination and should be offered contraceptive advice at this time. The nature of the advice will depend on age, parity, and future reproductive plans. The progesterone only (mini-pill) has virtually no effect on carbohydrate or lipid metabolism and is therefore suitable for the breast feeding diabetic woman. Provided she is prepared to accept the slightly higher failure rate of this method when ovulation resumes, then it may also be used long term. Modern, low-dose, combined oral contraceptive preparations have little effect on high or low density lipoprotein concentrations or carbohydrate metabolism and can be used safely, especially in younger, insulin-dependent and gestational diabetics. Early concerns about the apparently high failure rates of copper-containing intrauterine devices in diabetics have been refuted in recent studies. The woman who has completed her family should be encouraged to consider a laparosopic sterilization.

Further reading

Dornhorst A and Hadden DR, eds (1996). *Diabetes and pregnancy—an international approach to diagnosis and managemen*ı. John Wiley and Sons, Chichester. A comprehensive review of the subject.

Reece EA, ed. (1996). Diabetes in pregnancy. *Obstetrics and Gynecology Clinics of North America* **23**. A further review of this subject.

Reece EA, Coustan DR, eds (1995). *Diabetes mellitus in pregnancy*. Churchill Livingstone, New York. A comprehensive textbook.

13.11 Endocrine disease in pregnancy

John H. Lazarus

Introduction

During pregnancy the physiology of the mother and fetus changes constantly. For instance, maternal oestrogen concentrations rise and affect hepatic protein synthesis and plasma volume increases by as much as 50 per cent with consequent haemodilution. Endocrine function in the developing fetus is initially almost entirely dependent on maternal function as most endocrine glands do not produce hormones until the second trimester. Thereafter the fetus becomes less reliant on maternal hormones as the fetal glands develop and mature. This section will discuss the important therapeutic aspects of endocrine disease in pregnancy.

Pituitary disease

Prolactinoma

Pituitary adenomas are the most common pituitary disorder affecting pregnancy and prolactinomas are the most common of the hormone-secreting adenomas. Prolactinomas are a common cause of reproductive and sexual dysfunction and hyperprolactinaemia must be corrected to allow ovulation and fertility. The main concern during pregnancy is of symptomatic enlargement leading to visual impairment. There is less than a 2 per cent risk of this happening with a microprolactinoma, but a greater than 15 per cent risk with a macroprolactinoma. It is safe for patients to become pregnant following bromocriptine treatment; the prolactinomas may decrease in size, remain unchanged, or in some cases achieve complete resolution. Treatment with bromocriptine is safe during gestation, but a macroadenoma may require debulking prior to pregnancy. Cabergoline has also been used during pregnancy with no deleterious effects. The concern during pregnancy is the development of visual impairment which may occur with a macro-prolactinoma but not with a microprolactinoma.

Acromegaly

Fertility is impaired in acromegaly due to concomitant hyperprolactinaemia and decreased gonadotrophin reserve. The main aims of therapy in an acromegalic woman wishing to conceive are therefore to normalize prolactin and growth hormone levels to promote fertility. With the use of surgery as well as dopamine agonists and octreotide, many women with this condition are now able to achieve pregnancy.

The pituitary gland enlarges during normal gestation and may increase by 45 per cent during the first trimester. Pituitary adenomas can also enlarge, and pregnancy exacerbates acromegaly in about 17 per cent of cases. Patients with adenomas greater than 1.2 cm are at greater risk of visual loss during pregnancy. Management can be difficult, and risks need to be judged carefully. Those with microadenomas should discontinue medical therapy (bromocriptine or somatostatin analogues) during pregnancy and be assessed at each trimester (Fig. 1). In patients with macroadenomas removal before pregnancy leads to a greater risk of infertility, but if not resected there is a greater risk of pituitary enlargement and visual impairment during gestation. More detailed assessment during pregnancy is therefore recommended including regular visual field checks and MRI examinations in those patients with significantly large tumours at the beginning of pregnancy. However, although the metabolic and cardiovascular complications of acromegaly might be expected to result in increased risk to both mother and fetus, these potential hazards do not seem to be realized.

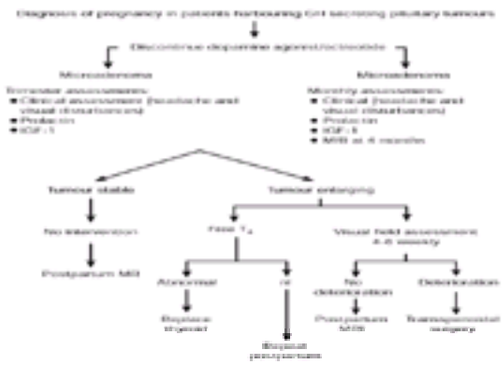

Fig. 1 Scheme for management of growth hormone secreting pituitary tumours in pregnancy (reproduced from Herman-Bonert *et al*. 1998, with permission).

Some women with apparent prolactinomas or non-functioning pituitary tumours will be found to have multiple endocrine neoplasia type 1 (MEN1). In this situation screening of the child should be offered.

Cushing's syndrome

Cushing's syndrome during pregnancy is associated with a high incidence of maternal and fetal complications, only about one-quarter of patients having an uncomplicated pregnancy. The diagnosis of the disease during gestation is difficult because many of the biochemical features, such as elevated cortisol levels and loss of the normal glucocorticoid feedback, are present during normal pregnancy. The use of corticotrophin-releasing hormone and dexamethasone testing can be helpful and MRI is valuable. Transphenoidal surgery has been successfully performed during pregnancy. This operation should be carefully considered if any pituitary tumour enlarges during pregnancy, especially when there is evidence of increasing visual field impairment.

Diabetes insipidus

Central diabetes insipidus may present during pregnancy. It is seen in women with Sheehan's syndrome, partial postpartum hypopituitarism, and associated with infiltrative disorders such as histiocytosis X. ADH deficiency is corrected using intranasal synthetic 1-deamino-8- D-arginine-vasopressin (DDAVP). During pregnancy this seems to be safe for both mother and baby. It does not affect delivery and has no adverse effects on the neonate.

Postpartum hypopituitarism

Sheehan's syndrome is caused by pituitary infarction following significant hypotension occurring at the time of delivery. Advances in obstetric care mean that it is now uncommon in the 'developed' world. However, lymphocytic hypophysitis is now increasingly recognized as a cause of hypopituitarism occurring late in pregnancy and in the postpartum period, and around 60 per cent of cases of women found to have adenohypophysitis are pregnant or have recently been delivered. They typically present with symptoms of an expanding pituitary tumour. Headaches, visual symptoms, inability to lactate, and amenorrhoea occur. Hyperprolactinaemia and elevated growth hormone levels are found. Computed tomography or magnetic resonance imaging reveals a pituitary mass mimicking an adenoma in about four-fifths of patients. Evaluation of pituitary function shows isolated or multiple anterior pituitary deficiency. Adenocorticotrophic hormone secretion is impaired, most frequently followed by that of thyroid-stimulating hormone (TSH), gonadotrophins, growth hormone, and prolactin. Histology shows lymphocytic infiltration and this may extend up to the pituitary stalk to the infundibulum. Antibodies to pituitary tissue may be present but often are not. Other autoimmune diseases, particularly postpartum thyroiditis, may be associated. In addition to the pituitary symptoms described, patients are at risk of adrenal failure and death has been reported.

Adrenal function should therefore be assessed in all patients. The diagnosis of lymphocytic hypophysitis may only be made at surgery, but if the condition is suspected beforehand, then surgery should be avoided and corticosteroids given, since these can reduce the size of the pituitary mass.

Thyroid disease

Maternal thyroid function during pregnancy

Thyroid physiology and function alter significantly during pregnancy. Thyroid volume increases in iodine deficient areas but not in areas of iodine sufficiency. There is increased synthesis of thyroxine-binding globulin, but free thyroxine decreases during the second and third trimester. During the first trimester the maternal thyroid is to some extent controlled by placental human chorionic gonadotrophin, a weak thyroid stimulator, and during this period TSH levels are suppressed.

Hyperthyroidism

Hyperthyroidism is associated with impaired fertility and the presence of thyroid antibodies is a marker for miscarriage and recurrent abortion even in those who are euthyroid. Nevertheless, hyperthyroidism occurs in 0.2 per cent of pregnancies. This is usually due to Graves' disease, but other causes include hydatidiform molar disease and gestational thyrotoxicosis due to high human chorionic gonadotrophin concentrations. Since the incidence of hyperthyroidism in pregnancy is low, and the symptoms of hyperthyroidism frequently overlap with those of the pregnancy itself, a high index of clinical suspicion is required to make the diagnosis. Further difficulties arise because thyroid function tests vary during normal pregnancy (see above) and each laboratory should ideally establish its own normal ranges for those who are pregnant. However, it is particularly important to consider whether a borderline test at this time really indicates thyroid dysfunction, and any result should be repeated before treatment is started. The diagnosis of hyperthyroidism is best made by noting an elevated serum free tri-iodothyronine in association with a suppressed TSH, the free thyroxine being elevated or at the upper limit of normal.

Untreated hyperthyroidism is associated with an increased risk of abortion and, if the pregnancy is completed, the baby may be of low birth weight and may show transient hyperthyroidism due to the transplacental passage of thyroid stimulating antibodies. Neonatal hyperthyroidism can also occur in babies born to euthyroid mothers who have been treated for Graves' hyperthyroidism in the past. It is therefore prudent to measure thyroid stimulating antibodies in the first trimester in any woman with active Graves' hyperthyroidism as well as those previously treated. If values are high, they should be measured again at 36 weeks and—if elevated—the obstetrician and paediatrician alerted to the possibility of a thyrotoxic infant.

Hyperthyroidism in pregnancy can be managed with either antithyroid drugs or surgery, the latter being optimally performed in the second trimester. Radioiodine therapy is completely contraindicated during pregnancy: non-pregnant women should undergo a pregnancy test 1 or 2 days before ^{131}I administration and be advised not to conceive for at least 4 months thereafter. If radioiodine has been administered after 12 weeks gestation (i.e. after the fetal thyroid is functional) there may be a case for therapeutic abortion. Propylthiouracil is the preferred antithyroid drug (Table) as fetal side-effects have been reported with carbimazole and methimazole. These are rare but significant, including aplasia cutis and a possible methimazole embryopathy, characterized by choanal atresia and other defects. If necessary, propylthiouracil should be continued in low dose right up to delivery and into the puerperium. An exacerbation of Graves' disease often occurs postpartum and this should be checked. Propylthiouracil does cross into breast milk but in lower concentrations than carbimazole or methimazole so that breast feeding can usually be permitted, but if breast feeding is prolonged, then thyroid function should be monitored in the neonate.

Hypothyroidism

Hypothyroidism is associated with relative infertility because of anovulation and menorrhagia. There is an increased incidence of stillbirths, congenital malformations, and maternal obstetric complications if the condition is untreated. In iodine-deficient areas the risk of neonatal brain damage is increased, resulting in cretinism in severe cases. This is due to the fact that the developing fetal nervous system is entirely dependent on T4 derived from the mother in the first trimester. Even in iodine-sufficient countries mild hypothyroidism during pregnancy is associated with impaired child development and a case for screening for maternal hypothyroidism in early pregnancy can be made. Pregnant patients with hypothyroidism of any degree should always be treated with thyroxine. In those already receiving T4 when found to be pregnant, the dose should be increased by at least 50 µg/day (and more if necessary) as it has been shown that dose requirements increase during gestation.

Fetal and neonatal thyroid dysfunction

The fetus of a mother with thyroid dysfunction during pregnancy should be regarded as a patient in its own right. Transplacental passage of maternal thyroid stimulating IgG immunoglobulins from a mother with Graves' disease may cause fetal hyperthyroidism as well as neonatal disease. Fetal hyperthyroidism is treated with propylthiouracil and the fetal heart rate monitored carefully to avoid hypothyroidism. If the mother is euthyroid, possibly with a history of previously treated Graves' hyperthyroidism, an hyperthyroid fetus may still occur due to thyroid stimulating IgGs. The fetal state is diagnosed by noting a high fetal heart rate and may be confirmed if necessary by the specialized procedure of fetal blood sampling. Treatment of the fetal hyperthyroidism in this situation is by maternal administration of high-dose propylthiouracil (300–450 mg/day). It is also necessary to give thyroxine during gestation to prevent maternal hypothyroidism.

In addition to transient neonatal hyperthyroidism, a hypothyroid state can also occur at birth due to maternal TSH-receptor-blocking antibodies. Both of these conditions are transient and the mother can be reassured. Other causes of transient neonatal hypothyroidism include maternal antithyroid drug administration, iodine deficiency, and ingestion of goitrogens by the mother.

Permanent neonatal hypothyroidism has an incidence of 1/4000 live births and is screened for routinely using whole blood TSH concentrations usually obtained by heel prick at 5 to 7 days. Recent advances in molecular biological diagnosis have revealed a number of specific hereditary gene defects (e.g. in the structure of thyroglobulin) which may require parental DNA analysis for appropriate genetic counselling.

Postpartum thyroid disease

Patients with hypothyroidism almost invariably have circulating antithyroid antibodies (usually antithyroid peroxidase (anti-TPO)). In euthyroid women, anti-TPO antibodies are found in 10 per cent at the routine antenatal booking clinic. These women are at risk of developing postpartum thyroid dysfunction (Fig. 2). Postpartum thyroiditis is a destructive process, not influenced by T4 or iodine treatment. It is characterized by transient hyperthyroidism followed by hypothyroidism, occurring in 5 to 9 per cent of anti-TPO positive women (Fig. 3). Although a minority of patients have been described with this condition who have no demonstrable immune abnormalities, the great majority have anti-TPO antibodies. Their hyperthyroidism is relatively asymptomatic but may require treatment with b-adrenoreceptor-blocking agents in some cases. Carbimazole is not effective. By contrast, the hypothyroidism is often symptomatic and thyroxine should be given. It persists at 1 year in 20 to 30 per cent of cases, and those that recover from transient postpartum thyroid dysfunction should have their thyroid function measured annually as nearly 50 per cent will develop hypothyroidism in the next 7 years. Mild depressive symptomatology is more common in postpartum women with anti-TPO antibodies than those without

these markers. Recurrent postpartum thyroid dysfunction will occur in up to 75 per cent of those women who have experienced a previous episode.

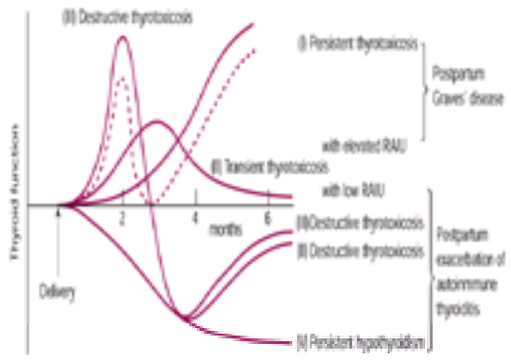

Fig. 2 Diagram to illustrate the progression to different types of thyroid dysfunction observed in the postpartum period (reproduced from Amino *et al*. 1999, with permission).

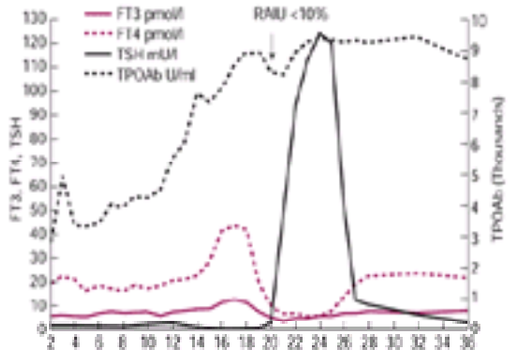

Fig. 3 Clinical chart illustrating the course of postpartum hyper- and hypothyroidism in a woman studied weekly for 36 weeks postpartum. Figures are weeks postpartum. Delivery at time 0 (reproduced from Lazarus *et al*. 1999, with permission). RAIU = radioiodine uptake.

At present, measurement of anti-TPO antibodies is not part of the routine antenatal screening performed at around 14 weeeks gestation. Such screening would predict the development of postpartum thyroiditis. Recent data also suggests that the development of children of anti-TPO positive mothers may be impaired. A case can therefore be made to support antenatal screening for anti-TPO antibodies. Similar arguments have been put forward for antenatal screening of TSH levels.

Thyroid nodules

Thyroid nodules may be clinically detectable in up to 10 per cent of pregnant women. Most of these lesions will be benign colloid nodules but between 5 and 20 per cent are true neoplasms, either benign follicular adenomas or carcinomas of follicular or para-follicular (C-) cell origin. It is often debatable whether investigation of a thyroid nodule should actually be performed during pregnancy. Many experts would pursue matters if a lump was found early on, but delay until the postpartum period if a nodule presented after 5 to 6 months gestation. Fine-needle aspiration biopsy is diagnostic in about 90 per cent of cases. In the remaining 10 per cent, ultrasound-guided biopsy may produce diagnostic tissue. If neck exploration is recommended, this should preferably be performed during the second trimester to avoid abortion in the first and premature labour in the third. Pregnancy does not adversely affect the course (generally favourable) of differentiated thyroid cancer: recurrence and distant metastases occur at the same rates in pregnant and non-pregnant women. In general, differentiated thyroid cancer should not be a contraindication to pregnancy or an indication for abortion. The position for medullary cell carcinoma is less clear cut.

Parathyroid disease

Although diseases of the parathyroid glands are uncommon in women of childbearing age, hyperparathyroidism during pregnancy can lead to acute pancreatitis, hypercalcaemic crisis, and toxaemia. There is an increased incidence of prematurity and neonatal hypercalcaemia if maternal calcium levels are high. If necessary, parathyroidectomy can be undertaken safely in the second trimester. Hypoparathyroidism is treated with vitamin D analogues, the dose of which may need to be increased to maintain normocalcaemia during gestation. Calcium levels should be monitored regularly throughout pregnancy, at least in each trimester.

Adrenal disease

The diagnosis of adrenal disease is often delayed in pregnancy. Adrenal tumours are very rare, but their pathophysiological consequences for mother and fetus are dire. In patients with phaeochromocytoma, hypertension may initially be mistaken for pregnancy-associated hypertension. This should be managed medically, as should primary aldosteronism, with surgical resection considered postpartum.

Occasionally phaeochromocytoma can be familial and screening of the child and other family members for measurable calcitonin levels, evidence of neurofibromatosis, von Hippel Lindau syndrome, or thyroid enlargement should be offered. It is also important to consider screening for multiple endocrine neoplasia type 2, associated RET mutations, and mutations in the von Hippel Lindau gene.

Addison's disease

Addison's disease is rarely diagnosed during pregnancy, but can present as a postpartum addisonian crisis. Symptoms may be attributed to pregnancy or its complications, leading to delayed diagnosis. The condition is readily confirmed by measurement of plasma cortisol, the short synacthen test, and adenocorticotrophic hormone levels. Antibodies to 21-hydroxylase should be measured to confirm the autoimmune nature of the disease, and other autoimmune conditions sought.

Congenital adrenal hyperplasia

Women with severe congenital adrenal hyperplasia have decreased fertility rates because of oligo-ovulation due to elevated androgen levels. Successful conception requires careful endocrine monitoring and possibly induction of ovulation. Problems occur during pregnancy in women with 21-hydroxylase deficiency (P450c21 deficiency),11-hydroxylase deficiency (P450c11), and 3 b-hydroxysteroid dehydrogenase deficiency. Gestational management must involve adequate adrenal steroid replacement and adrenal androgen suppression. Clinical status, serum electrolytes, and androgen levels should be measured regularly and glucocorticoid and mineralocorticoid therapy adjusted or increased as necessary. As the genetic basis of 21- and 11-hydroxylase deficiency is known, prenatal diagnosis is possible. Infants should also be evaluated clinically and, in most cases, biochemically.

Miscellaneous endocrine conditions

Gonadal dysgenesis (Turner's syndrome) is characterized by streak ovaries and infertility. Advances in *in vitro* fertilization and embryo transplantation have made pregnancy possible for some of these patients.

Further reading

Badawy SZ, Marziale JC, Rosenbaum AE, Chang JK, Joy SE (1997). The long-term effects of pregnancy and bromocriptine treatment on prolactinomas—the value of radiologic studies. *Early Pregnancy* **3**, 306–11. Useful clinical study.

Chittacharoen A, Phuapradit W (1997). Pheochromocytoma during pregnancy: case report. *Journal of Obstetric and Gynaecology Research* **23**, 209–12. Discussion of diagnosis and treatment.

Ciccarelli E, Grottoli S, Razzore P, *et al.* (1997). Long-term treatment with cabergoline, a new long-lasting ergoline derivate, in idiopathic or tumorous hyperprolactinaemia and outcome of drug-induced pregnancy. *Journal of Endocrinological Investigation* **20**, 547–51. Reviews of nearly 50 patient with hyperprolactinaemia.

Garner PR (1998). Congenital adrenal hyperplasia in pregnancy. *Seminars in Perinatology* **22**, 446–56. Discussion of biochemistry and management during pregnancy.

Glinoer D (1997). The regulation of thyroid function in pregnancy: Pathways of endocrine adaptation from physiology to pathology. *Endocrine Reviews* **18**, 404–33. Excellent review of this subject.

Hall R (1995). Pregnancy and autoimmune endocrine disease. *Bailliere's Clinical Endocrinology and Metabolism. Autoimmune Endocrine Disease* **9**, 137–55. Detailed discussion of this subject.

Harrington JL, Farley DR, van Heerden JA, Ramin KD (1999). Adrenal tumors and pregnancy. *World Journal of Surgery* **23**, 182–6. Clinical aspects of four adrenal tumours in pregnancy from the Mayo clinic.

Herman-Bonert V, Seliverstov M, Melmed S (1998). Pregnancy in acromegaly: successful therapeutic outcome. *Journal of Clinical Endocrinology and Metabolism* **83**, 727–31. Review by experienced endocrinologists.

Lo JC, Schwitzebel VM, Tyrrel JB, *et al.* (1999). Normal female infants born of mothers with classic congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Journal of Clinical Endocrinology and Metabolism* **84**, 930–6. Highlight key issues in management, particularly endocrine monitoring.

Maniker AH, Krieger AJ (1996). Rapid recurrence of craniopharyngioma during pregnancy with recovery of vision: a case report. *Surgical Neurology* **45**, 324–7. Unusual but important case of craniopharyngioma.

Mestman JH (1998). Parathyroid disorders of pregnancy. *Seminars in Perinatology* **22**, 485–96. Comprehensive review of subject.

Molitch ME (1998). Pituitary disease in pregnancy. *Seminars in Perinatology* **22**, 457–70. Review of all pituitary disease with emphasis on adenomas.

Molitch ME (1999). Medical treatment of prolactinomas. *Endocrinology and Metabolism Clinics of North America* **28**, 143–69. Detailed discussion of therapy of the commonest of pituitary tumours.

Ray JG (1998). DDAVP use during pregnancy: an analysis of its safety for mother and child. *Obstetrical and Gynecological Survey* **53**, 450–5. Large literature review of DDAVP use during pregnancy.

Ross RJ, Chew SL, Perry L, Erskine K, Medbak S, Afshar F (1995). Diagnosis and selective cure of Cushing's disease during pregnancy by transsphenoidal surgery. *European Journal of Endocrinology* **132**, 722–6. Complex case illustrates diagnosis and management of this condition.

13.12 Neurological disease in pregnancy

G. G. Lennox

Introduction

A wide range of neurological problems occasionally complicate pregnancy and the puerperium. Pre-existing neurological diseases, such as epilepsy and myasthenia gravis, sometimes become more troublesome or, like myotonic dystrophy, pose obstetric difficulties. New neurological disorders can occur, ranging from diseases of the peripheral nerves and muscles that are relatively common but generally benign, to diseases of the central nervous system that are rare but potentially life-threatening. In all these situations the presence of the fetus influences management.

Disorders of muscle and neuromuscular transmission

Muscle disorders

Muscle cramps, particularly on waking, are extremely common in the third trimester. They are almost never a symptom of serious neurological disease and often respond to calcium supplements. Restless legs syndrome, in which there is a feeling of discomfort in the legs that is relieved by movement, is also common, especially on retiring to bed. It may respond to correction of underlying anaemia (particularly if this is due to folate or iron deficiency); otherwise management in pregnancy is aimed at promoting the rapid onset of sleep, for example by reducing caffeine intake. Drug treatment (with levodopa, clonazepam, or codeine) is best avoided. Polymyositis, although rare in young women, can deteriorate during pregnancy. Treatment with corticosteroids is thought to be safe, as is azathioprine if another immunosuppressive agent is necessary.

Most of the congenital myopathies and muscular dystrophies, apart from myotonic dystrophy, cause no special problems in pregnancy unless they are of sufficient severity to compromise ventilation, either because of respiratory muscle weakness or associated scoliosis. Such cases should ideally be assessed in a specialist unit prior to pregnancy because the cardiorespiratory demands of pregnancy, combined with the splinting effect of the fetus on the diaphragm, can lead to ventilatory failure and a temporary need for mechanical ventilation.

Myotonic dystrophy

Myotonic dystrophy is an autosomal dominant disorder due to an expanded triplet repeat in the myotonin protein kinase gene. The expansion tends to increase during transmission from mother to child, so that a mildly affected or asymptomatic mother may have a severely affected fetus. This probably accounts for the excess of polyhydramnios and perinatal death. The myotonia affects the smooth muscle of the uterus, prolonging labour and increasing the risk of postpartum haemorrhage because of uterine inertia. Moderately affected mothers may also develop symptoms of cardiomyopathy during labour. (See Chapter 24.22 for further discussion.)

Myasthenia gravis

Myasthenia gravis deteriorates, improves, and remains stable during pregnancy in roughly equal proportions of patients, but the response is neither predictable nor reproducible in subsequent pregnancies. It deteriorates during the puerperium in about half of all patients, an effect that may also occur after abortion. The mechanism of these changes is not clear. Corticosteroids, oral anticholinesterases, and plasmapheresis can all be employed in the usual manner during pregnancy. It is reasonable to continue azathioprine where this has been prescribed before pregnancy for severe myasthenia, bearing in mind the risk of inducing neonatal leucopenia. Thymectomy can be performed during pregnancy (for example where a malignant thymoma is suspected) but may take up to a year to have a therapeutic effect and ideally should be performed well before any planned pregnancy. Myasthenia does not usually influence labour, although the second stage may be prolonged by fatigue and obstetric anaesthesia is complicated by the need to avoid drugs with adverse effects on neuromuscular transmission—regional anaesthesia being preferable when possible. Acetylcholine receptor antibodies can cross into the fetal circulation, giving rise to transient neonatal myasthenia in up to 20 per cent of the babies from affected mothers. Expert paediatric support must therefore be available at delivery. (See Chapter 24.17 for further discussion.)

Disorders of nerves and nerve roots

Facial palsy

The incidence of facial nerve palsy (Bell's palsy) is substantially increased during pregnancy and the puerperium (as Bell himself described). The reason for this is not known. There have been no studies of treatment in this specific context but it is reasonable to treat promptly with prednisolone, beginning at 40 mg daily and reducing over a 2-week course.

Mononeuropathies

Carpal tunnel syndrome, due to compression of the median nerve in the wrist, is very common in pregnancy, characteristically causing pain and tingling in the hands at night and after use. Most cases can be managed with nocturnal wrist splints, although steroid injections into the carpal tunnel may tide the patient over into the puerperium, when symptoms usually remit. Diuretics are of little value. Surgical decompression during pregnancy should be reserved for cases with severe pain, weakness, or wasting, when usually there have been symptoms either before pregnancy or early in the first trimester. In troublesome cases it is worth considering delayed surgery to prevent recurrence in subsequent pregnancies, which is common.

The lateral cutaneous nerve of the thigh can be compressed as it crosses the inguinal ligament. This is particularly common in the third trimester and causes tingling, hypersensitivity, or numbness in the midlateral thigh, which may be bilateral. Usually no treatment is required, but troublesome cases may respond to transcutaneous nerve stimulation or a local nerve block. Remission after delivery is the rule.

Lumbosacral root and plexus problems

Backache is very common in pregnancy, particularly in women with a past history of back pain or occupations that involve bending and lifting. It is traditionally blamed

on changes in posture and hormonally mediated relaxation of spinal and sacroiliac joints. The pain is usually confined to the lumbar region but may radiate into the buttock or thigh. There may be tenderness over one or other sacroiliac joints. Radiological investigations are not needed if there are no abnormal neurological signs. Management is conservative.

Abrupt onset of pain that radiates below the knee with focal weakness, numbness, or reflex loss is most likely to be due to a prolapsed intervertebral disc. Provided that the signs are unilateral with no sphincter impairment, conservative management is again appropriate, with analgesia and advice to keep mobile. If this fails then magnetic resonance imaging is thought to be a safe method of investigation prior to consideration of lumbar microdiscectomy.

Obstetric nerve palsies are becoming less common with improvements in obstetric care, but still occur in cases of prolonged or complicated labour (for example due to cephalopelvic disproportion, dystocia, and primiparity), in difficult forceps deliveries, and as a result of traction or haematoma formation in caesarean section. Damage to the common peroneal nerve from incorrectly positioned leg holders is now rare.

The baby may compress the lower parts of the lumbosacral plexus during labour. This will typically give rise to focal neurological deficits depending on which parts of the plexus have borne the brunt of the pressure. Most commonly there is a unilateral footdrop, which may only become apparent when the mother starts to mobilize. Examination also reveals sensory loss that characteristically involves the dorsolateral foot and leg, distinguishing plexus damage from a common peroneal palsy where the sensory loss is confined to the dorsum of the foot. Compression of the upper lumbosacral plexus leads to weakness of iliopsoas as well as the quadriceps muscles, which distinguishes it from more distal damage to the femoral nerve. Both may give rise to sensory loss in the anteromedial thigh and loss or depression of the knee jerk. In most cases the prognosis is good, with spontaneous recovery over a couple of months. Particular care must be taken in subsequent deliveries to avoid further damage to the same nerve, as recovery after repeated injury will tend to be less complete.

Long, complicated, or instrumental deliveries may also damage the obturator or pudendal nerves. Obturator neuropathy leads to weakness of hip adduction and rotation, together with some sensory loss in the upper medial thigh. Pudendal nerve damage may be initially asymptomatic but probably contributes to the subsequent development of perineal descent and stress incontinence.

Generalized neuropathies

Chronic inflammatory demyelinating polyradiculoneuropathy can present or relapse during pregnancy, and can be treated with corticosteroids or (if necessary) intravenous immunoglobulin. The incidence of acute Guillain Barre syndrome is increased in the puerperium and can be managed in the usual ways.

The combination of the nutritional demands of pregnancy and hyperemesis gravidarum can lead to thiamine deficiency. This most commonly causes a subacute sensory neuropathy, but cases of acute Wernicke's encephalopathy (with any combination of altered consciousness, ataxia, and ophthalmoplegia, leading if untreated to death) have been described. Both conditions respond promptly to parenteral thiamine 100 mg daily.

Pregnancy can precipitate relapse in acute intermittent porphyria: abdominal pain typically precedes autonomic and sensory neuropathy, sometimes with seizures and psychiatric disturbance. Finally, lepromatous neuropathies may present or deteriorate during pregnancy, making careful clinical supervision advisable.

Disorders of the central nervous system

Headache

The most common form of headache during pregnancy (as at other times) is chronic daily headache of the tension type. This may be a continuation of pre-existing headaches or a new phenomenon, compounded by anxiety, depression, or poor sleep. Neurological examination is usual in such cases and the treatment should concentrate on explanation and reassurance, coupled if necessary with advice about relaxation techniques. Occasional paracetamol may be helpful; aspirin should be avoided in the third trimester.

Migraine usually improves during pregnancy, but approximately 20 per cent of sufferers get worse and occasionally migraine actually begins in pregnancy. No single hormonal change has been convincingly linked to these divergent responses. Whilst migraine can generally be identified accurately on clinical grounds, diagnosis is more difficult when it presents for the first time in pregnancy, particularly if accompanied by the transient focal deficits of migraine aura such visual disturbances or hemiplegia. Other potential causes of headache in pregnancy, such as eclampsia (Chapter 13.4), subarachnoid haemorrhage, cerebral venous thrombosis, cerebral infarction, intracranial tumour, and intracranial infection must be considered and excluded by careful neurological and general examination, supplemented if necessary by brain imaging (see below). Some women present in pregnancy with migraine aura without headache, which can also give rise to diagnostic difficulty: it may be necessary to exclude causes of transient ischaemic attack.

Acute migraine attacks should be treated promptly with rest and paracetamol; prochlorperazine is probably a safe treatment for vomiting. Ergot derivatives must be avoided in pregnancy (and breastfeeding), and the triptan drugs have not yet been shown to be safe. If attacks are frequent then attention should be paid to relevant lifestyle factors such as irregular sleep or meals and worry. Prophylactic drug treatment is occasionally necessary, and the greatest experience lies with propranolol which, in doses of 20 to 80 mg three times a day, appears to be both effective and safe, despite its effect on placental blood flow. Women with migraine commonly develop a dull non-specific headache of variable severity in the first few days after delivery. This usually responds to simple analgesia, particularly if the woman has been warned about the phenomenon.

Tumours

Although the incidence of cerebral and spinal tumours is probably no greater than at other times, some tumours expand during pregnancy and may present unusually rapidly. This probably reflects a mixture of hormonal and vascular factors; most meningiomas and some neurofibromas and gliomas express oestrogen and progesterone receptors and placental growth factor.

Neurofibromatosis type 1 presents particular problems in pregnancy. Women with this condition experience an increased rate of spontaneous first trimester abortions, perhaps also intrauterine fetal growth retardation and stillbirths, and have a high rate of caesarean section. Most women notice that cutaneous neurofibromas grow or appear *de novo* during pregnancy.

Meningiomas are particularly liable to expand in the third trimester, causing local mass effects such as headache, cranial nerve palsies, hemiparesis, or paraparesis, which may remit after delivery. Corticosteroids can be given to reduce surrounding oedema, and surgery can often be delayed until after delivery. Gliomas tend to present earlier in pregnancy and have a reputation for following an aggressive course. They may require early surgical intervention, and it is sometimes also appropriate to consider termination of the pregnancy. Women with known intracranial mass lesions require careful assessment prior to delivery: prolonged Valsalva manoeuvres can increase intracranial pressure and elective caesarean section may be necessary.

Choriocarcinoma is a tumour peculiar to pregnancy and the most common form of malignancy associated with pregnancy. It usually presents after molar pregnancy or abortion, but 15 per cent of cases occur during or after normal pregnancy. Neurological manifestations due to brain or spine metastases are common. The brain metastases have a tendency to invade blood vessels, giving rise to strokes through infarction or haemorrhage. Spinal metastases cause cord or cauda equina compression that may be rapid in onset. There are usually multiple pulmonary metastases on chest radiography and the serum chorionic gonadotrophin is greatly elevated. Early diagnosis and treatment (with chemotherapy and radiotherapy) improves survival, but the mortality rate of cases with neurological manifestations remains high.

The normal pituitary gland and some pituitary tumours such as prolactinomas expand during pregnancy. Their management, the possibility of pituitary apoplexy, and the postpartum differential diagnosis of lymphocytic hypophysitis are discussed in Chapter 13.11.

Stroke

There is probably an increased incidence of stroke during pregnancy and the puerperium, although it remains rare. There are some causes of stroke which seem to be more common in pregnancy, but most cases are due to one of the usual causes of stroke in non-pregnant young women. Cerebral infarction due to large cerebral

artery occlusion may be slightly more common. Possible explanations for this include the mild hypercoagulable state that develops in the later stages of pregnancy and persists for a few weeks afterwards, also the phenomenon of paradoxical embolism from the leg or pelvic veins. Cerebral infarction may occur as a result of hypoxia–ischaemia or disseminated intravascular coagulation in the context of major obstetric emergencies such as amniotic fluid embolism. Unless the cause is obvious, ischaemic stroke in pregnancy should be investigated comprehensively in the same way that it would in a young non-pregnant woman.

There is a rare syndrome of segmental cerebral vasoconstriction in the puerperium, which usually presents with headaches, seizures, or focal deficits (especially visual field defects). It has a predilection for the posterior cerebral circulation and can give rise to multiple infarcts or haemorrhages. The condition, generally termed postpartum angiopathy, can occur spontaneously but has also been described in women taking bromocriptine. There are reports of successful treatment with corticosteroids and vasodilators, but there have been no prospective studies. The condition can recur in subsequent pregnancies.

Cerebral venous thrombosis, like deep vein thrombosis of the legs, is commoner in the puerperium, and the two conditions may coexist. Classically it gives rise to headache and neurological deficit that evolves over several hours and may become bilateral, with seizures and papilloedema; other cases present with the syndrome of benign intracranial hypertension. The diagnosis can usually be made with magnetic resonance imaging, including magnetic resonance venography. Although venous infarcts frequently undergo haemorrhagic transformation, the currently available evidence favours treatment with heparin. If the patient survives then recovery may be surprisingly complete.

The incidence of aneurysmal subarachnoid haemorrhage is not increased during pregnancy but its management is difficult. In general, neurosurgical considerations take precedence over obstetric ones and the aneurysm is treated in the usual way. If it is not technically possible to isolate the aneurysm then conventional wisdom is to deliver the baby (once it is mature) by caesarean section, although there is no definite evidence to suggest an increased risk of rebleeding during vaginal delivery. Intracranial and subarachnoid haemorrhage from arteriovenous malformations is much less common but the same principles of management apply. Women with untreatable vascular malformations (including cavernomas) should be counselled about the increased risk of bleeding (perhaps due to a mixture of hormonal and vascular factors) throughout pregnancy.

Epilepsy

Most women with pre-existing epilepsy have no change in the frequency of their seizures during pregnancy, but some 30 per cent have more frequent seizures. These are often women whose epilepsy has been hard to control at other times. Anticonvulsant plasma levels tend to fall in the later stages of pregnancy through increased volumes of distribution and rates of elimination: consensus guidelines recommend that plasma levels are monitored routinely and dosage adjusted to keep these steady. This approach is not entirely foolproof because most laboratories are unable to measure the changes in protein binding that also occur during pregnancy, increasing the availability of free drug. If the dosage is increased prophylactically during pregnancy then it is important to remember to reduce the dosage again in the postpartum period.

An equally important reason for deteriorating control of seizures is lack of adherence to anticonvulsant therapy because of fear of teratogenic effects. This is best addressed by counselling well before pregnancy, which should include a discussion of the risks of uncontrolled epilepsy to both mother and fetus, although it must be admitted that this is made difficult by the lack of quantitative data. As always, the aim of treatment is to control the epilepsy using a single anticonvulsant in the lowest effective dosage, and it is reasonable to try to reduce or withdraw anticonvulsants prior to pregnancy if there is a chance that the epilepsy may have remitted. It is unwise to attempt this during pregnancy itself.

All the anticonvulsant drugs are either known teratogens or of unknown safety in pregnancy. Epilepsy roughly doubles the risk of fetal malformation and most of this risk seems to be due to the treatment rather than the epilepsy or its cause. This translates into an absolute risk in the region of 5 to 10 per cent, although many of these malformations are minor. The risk is greatest in women taking two or more anticonvulsants. In the case of more serious abnormalities, phenytoin, barbiturates, carbamazepine, and sodium valproate all appear to be associated with facial clefts, cardiac septal defects, and a pattern of craniofacial and digital dysmorphism known as the fetal anticonvulsant syndrome. Sodium valproate and to a lesser extent carbamazepine appear to be associated with neural tube defects. Some of these defects may be secondary to drug-induced folate deficiency and it is good practice to offer folate supplements (5 mg daily) routinely to all potentially fertile women who are taking anticonvulsants, especially in the 3 months prior to and the first trimester of any planned pregnancy. Carbamazepine is traditionally regarded as the safest of the anticonvulsants for which we have reasonable data, but the teratogenic effects of sodium valproate appear to be dose dependent and dosages of less than 1 g daily may be equally safe. This is important when treating women with conditions such as juvenile myoclonic epilepsy which respond much better to valproate than carbamazepine. Several registers are currently collecting prospective data on the safety of the newer anticonvulsants (lamotrigine, gabapentin, vigabatrin, topiramate, tiagabine, etc.) but, at the time of writing, these must all be regarded as of unknown safety in pregnancy.

In addition to worries about teratogenic effects (which arise in the first 8 weeks of gestation), there are growing concerns about the potential for anticonvulsants to produce more subtle adverse effects on brain development and the subsequent behaviour and intelligence of the child. Such effects have been demonstrated in relation to anticonvulsant polytherapy including barbiturates, and it is possible that they may occur with other drugs. Prospective studies are in progress, but at the moment there is no clear guidance for women with epilepsy or their doctors in relation to the magnitude of these risks.

Carbamazepine, phenytoin, and the barbiturates accelerate vitamin K metabolism and increase haemorrhagic risks. Again it is considered to be good practice to offer the mother vitamin K supplements during the last month of pregnancy and to give the baby vitamin K at birth.

Epilepsy presenting for the first time in pregnancy requires investigation in the same way as adult-onset epilepsy in general. Idiopathic epilepsy that only occurs in pregnancy (so-called gestational epilepsy) is rare. Women presenting with serial seizures or status epilepticus are particularly likely to have an underlying secondary cause such as eclampsia, stroke, tumour, or encephalitis. Epilepsy during labour is usually either iatrogenic (for example, omission of normal anticonvulsant therapy) or again symptomatic of serious intracranial disease. Eclampsia (see Chapter 13.4) is clearly the first consideration, but other possibilities include amniotic fluid embolism and cerebral venous thrombosis.

All the anticonvulsants pass into breast milk to some extent, but this need not prevent breastfeeding. Only the barbiturates occasionally cause problems with excessive sedation, but this small risk must be balanced against the problems of effectively withdrawing barbiturates by not breastfeeding. This can lead to the baby becoming irritable and jittery; impaired suckling and withdrawal seizures have also been reported.

Multiple sclerosis

Pregnancy raises complex issues for women with multiple sclerosis. Preconceptual considerations include the small risk (approximately 3 per cent) of their child inheriting the disease and the practical burdens that child care imposes upon a mother with existing and potentially progressive disability. Several epidemiological studies have shown that the incidence of relapses of multiple sclerosis falls during pregnancy itself, with a compensatory rise in the puerperium (with between 20 and 40 per cent of women reporting an exacerbation of symptoms.) It has been suggested that this reflects the production of pregnancy-associated proteins with immunosuppressive properties, such as a-fetoprotein, and changes in T-lymphocyte subsets. There is no evidence of any long-term detrimental effect on disability, and no evidence of any adverse effect from epidural anaesthesia or breastfeeding.

Relapses in pregnancy are treated in the normal way, with rest supplemented by a short course of oral or intravenous steroid if there is serious new disability. High-dose steroids given late in pregnancy can cause neonatal adrenal suppression. The manufacturers of interferon-b advise women taking it to avoid pregnancy and discontinue it during pregnancy and breastfeeding unless there are compelling reasons to continue with therapy.

Many women with multiple sclerosis have impaired bladder emptying, which predisposes to urinary tract infection. Severe spinal cord disease is a particular risk because it may mask the usual symptoms of urinary infection; regular urine culture is a sensible precaution. Paraplegia (from any cause) otherwise has little effect on pregnancy, but can lead to premature and unheralded labour, hence regular monitoring is needed in the third trimester. High spinal cord lesions can cause autonomic instability during labour; this can be blocked by careful regional anaesthesia.

Movement disorders

Pregnancy aggravates any tendency to chorea, an effect termed chorea gravidarum. This should not be regarded as a specific diagnosis, and unless there is a definite history of previous Sydenham's chorea it should prompt a search for all the usual causes of the condition, including thyrotoxicosis and systemic lupus erythematosus. Chorea can be florid and exhausting so that treatment with a small dose of a neuroleptic such as haloperidol may be required. Recurrence in

subsequent pregnancies (or with the combined oral contraceptive) is common, perhaps because of the effects of oestrogens on the sensitivity of dopamine receptors.

Parkinsonism is rare in women of child-bearing age but tends to worsen slightly during pregnancy. Preconceptual counselling is difficult because there are no useful data in relation to the teratogenicity of the drugs used in young patients; levodopa has teratogenic effects in animals. Dystonic disorders also sometimes worsen in pregnancy, the effect being especially marked in dopa-responsive dystonia where an increase in levodopa therapy may be required. Wilson's disease is an exception and sometimes improves in pregnancy. Concerns about the potential teratogenic effects of therapy with penicillamine must be balanced against the risks of catastrophic neurological deterioration if therapy is abruptly withdrawn, although in the future treatments such as zinc may turn out to be a safe alternative.

Further reading

Aube M (1999). Migraine in pregnancy (review). *Neurology* **53** (4) (suppl. 1), 26–8.

Batocchi AP *et al*. (1999). Course and treatment of myasthenia gravis during pregnancy. *Neurology* **52**, 447–52.

Confraveux C *et al*. for the Pregnancy in Multiple Sclerosis Group (1998). Rate of pregnancy-related relapse in multiple sclerosis. *New England Journal of Medicine* **339**, 285–91.

Grosset DG *et al*. (1995). Stroke in pregnancy and the puerperium: what magnitude of risk? *Journal of Neurology, Neurosurgery and Psychiatry* **58**, 129–31.

Isla A *et al*. (1997). Brain tumour and pregnancy. *Obstetrics and Gynecology* **89**, 19–23.

Koch S *et al*. (1999). Long-term neuropsychological consequences of maternal epilepsy and anticonvulsant treatment during pregnancy for school-age children and adolescents. *Epilepsia* **40**, 1237–43.

Quality Standards Subcommittee of the American Academy of Neurology (1998). Practice parameter: management issues for women with epilepsy (summary statement). *Neurology* **51**, 944–8.

Rudnick-Schoneborn S *et al*. (1998). Different patterns of obstetric complications in myotonic dystrophy in relation to the disease status of the fetus. *American Journal of Medical Genetics* **80**, 314–21.

Shneerson JM (1994). Pregnancy in neuromuscular and skeletal disorders. *Archives of Chest Disease* **49**, 227–30.

13.13 The skin in pregnancy

F. Wojnarowska

The skin undergoes profound alterations during pregnancy as a result of endocrine, metabolic, and physiological changes. Some of these are trivial and chiefly cosmetic, producing no or minor symptoms, others can be distressing and/or of major medical importance. Pregnancy will profoundly modify expression of pre-existing skin disease and there are dermatoses which are specific to pregnancy: these are described in detail below.

Common skin changes in pregnancy

Vascular changes and lesions

There is increased skin blood flow during pregnancy and this makes the skin more prone to itch and to oedema, manifest as tightening of rings and shoes. Spider naevi and palmar erythema are common, as are haemangiomas. Pyogenic granuloma may develop: this is a benign tumour with a tendency ulcerate and to bleed, and is sometimes clinically confused with melanoma. It often recurs after local destruction.

Pigmentary changes and pigmented lesions

There is darkening of the nipples, genitalia, and linea alba. The unsightly and sometimes psychologically distressing facial pigmentation of melasma (chloasma) affects many women, is worse with sunlight, and can be reduced by the use of high protection factor (SPF 25) UVB and UVA sun screens.

Pigmented naevi can increase in number, size, and pigmentation. Melanoma may occur and is associated with a poor prognosis in pregnant women (see Section 23). Any rapidly changing, irregularly shaped, or irregularly pigmented mole should be biopsied to exclude a dysplastic naevus or melanoma.

Hair changes

There is diminished shedding of hair, due to prolongation of anagen. This is perceived as thickening of the hair, which increased sebum secretion makes appear more lustrous. The synchronized shedding after parturition gives rise to the distressing postpartum telogen effluvium. Hirsutism may begin or worsen in pregnancy as there is an associated increase in androgens.

Pilosebaceous changes

The increased oestrogens of pregnancy usually improve acne, but there may be worsening of acne in some unfortunate patients, and the entire skin is usually greasier.

Striae

Striae on the breasts and abdomen are very common in pregnancy, but do not necessarily relate to either the total weight gain or the rate of weight gain. There is much individual variation.

Cutaneous infections

Candida of the vulva as well as the vagina may occur. Cutaneous and genital warts thrive in pregnancy. Treatment of genital warts is by physical destruction as podophyllin must not be used in pregnancy. Genital herpes simplex infections can pose problems as regards delivery during active infections.

The pregnancy dermatoses

There are five major dermatoses which occur in pregnancy, and some that can be precipitated by pregnancy.

Pruritus of pregnancy

Itching occurs in about 20 per cent of pregnancies. Sometimes this is in association with an inflammatory dermatosis. Often there are no physical signs, other than scratch marks, and iron deficiency must be excluded. In about 2 per cent of women itching is related to cholestasis of pregnancy, when it is termed pruritus gravidarum. The itching begins in the third trimester and affects the abdomen, palms, and soles. Liver function tests are abnormal and bile salts are raised. It resolves postpartum, but will recur in subsequent pregnancies.

Management consists of emollients and sometimes antihistamines. Chlorpheniramine is the one usually recommended for pregnancy. The non-sedating antihistamines are probably ineffective.

Polymorphic eruption of pregnancy (pruritic urticated papules and plaques of pregnancy)

This dermatosis affects 1 in 240 singleton pregnancies but is more common in multiple births, and is thus being seen more often in the context of *in vitro* fertilization. It is more common with a male foetus. The dermatosis usually begins in the third trimester and occasionally postpartum. It is most common in first pregnancies or the first multiple pregnancy.

The lesions usually begin in the striae on the abdomen and thighs, and then spread to the whole trunk and limbs, including the hands and feet. The lesions are raised red papules (Plate) and plaques, occasionally polycyclic, and rarely may blister on the lower legs. The itching can be very severe, preventing sleep. The

histopathology shows oedema, perivascular lymphocytes, and eosinophils. Immunofluorescence does not demonstrate any circulating or bound immunoreactants.

The aetiology is unknown, but there is an association with a low serum cortisol. The increased frequency in multiple births may relate to the mechanical effect of the abdominal stretching or to an increased immune complex load.

Treatment is with reassurance and emollients, for example aqueous cream and 1 to 2 per cent menthol, is helpful, but is not always sufficient. Antihistamines and moderate to very potent topical steroids, which will have significant absorption (see Table 1), and occasionally systemic steroids or induction, may be required. The condition resolves over days to weeks after delivery. It does not usually recur. The outcome of the pregnancy is not adversely affected.

Prurigo of pregnancy

This is may affect 1 in 300 pregnancies. It commences at the end of the second or beginning of the third trimester. The eruption is scattered over the abdomen and limbs, and comprises excoriated papules. There is intense pruritus. It is essential to make sure that iron deficiency is not a contributing factor.

Histopathology shows a perivascular infiltrate with thickened epidermis. Direct and indirect immunofluorescence are negative. Treatment is with reassurance and emollients, and if this is not sufficient, with antihistamines and moderate to very potent topical steroids or occlusive coal tar or icthapaste bandages, which can be applied over topical steroids to the limbs. The condition resolves in days to weeks after delivery. It does not usually recur.

Pruritic folliculitis

This is rare and most commonly affects pregnancies with a male foetus. The lesions are pruritic papules and pustules that present in the third trimester, affect the trunk, and may spread to the limbs. Topical steroids may be helpful. Pruritic folliculitis is associated with a low birth weight.

Pemphigoid gestationis (herpes gestationis)

Pemphigoid gestationis is the most severe of the pregnancy dermatoses. It occurs in 1 in 50 000 pregnancies. The name herpes gestationis is best abandoned as the herpes refers to the herpetiform grouping of the blisters rather than herpes infection. Pemphigoid gestationis commences from the second trimester onwards and quite often in the first week postpartum (range 5 weeks gestation to 4 weeks postpartum). It usually occurs in the first and subsequent pregnancies, although 8 per cent of pregnancies are skipped.

The eruption begins around the umbilicus and spreads to the whole trunk, limbs, hands, and feet, including the palms and soles, and rarely the face. The mouth and vulva may be involved. The eruption usually commences as an annular red raised plaque around the umbilicus. The lesions comprise annular lesions, papules, and plaques. Vesicles and blisters are seen (Plate 2). The mucosal lesions may be blisters or erosions. Pruritus is severe and sleep often impossible. Transplacental transmission to the fetus occurs in about 3 per cent of affected pregnancies, and the neonate develops transient self-limiting blisters (Plate 3).

Histopathology demonstrates eosinophilia, subepidermal blisters, and tear-drop vesicles within the epidermis, continuous with the subepidermal blisters. Direct immunofluorescence demonstrates that C3 component of complement and IgG1 are bound at the basement membrane zone of the dermoepidermal junction. The patient's serum has circulating IgG1 basement membrane zone antibodies that bind C3. These immunoreactants are also found at the basement membrane zone of the amnion (Plate 4). The mothers have HLA DR 3, 4, and are C4 null, and there is an association with thyroid and less commonly other autoimmune disease.

The aetiology is only partially understood. The pathogenicity of the circulating basement membrane zone antibodies is demonstrated by transplacental transmission of the disease. The major target antigen is BP180/collagen XVII (chief epitope—the transmembrane NC16A domain) and BP230 is a further antigen. Both antigens are present in skin, mucosa, and amnion associated with the hemidesmosome and adhesion complex linking epithelium to dermis/mesenchyme, and are targets in other autoimmune blistering diseases. The placenta shows increased expression of antigen presenting cells, but it is unclear why breakdown of tolerance occurs, and why normal components of amnion and stratified squamous epithelium become antigenic.

Treatment with potent or very potent topical steroids and chlorpheniramine is sometimes successful, but usually systemic steroids, for example prednisolone 20 to 80 mg daily, are required, with the dose adjusted according to disease activity. There is usually a postpartum flare, necessitating increased steroids. The disease slowly resolves postpartum, but persists for several months. There is an increased incidence of premature births and small-for-dates babies. The classical teaching is that it recurs earlier and is more severe in subsequent pregnancies, but this has not always been our experience.

Dermatoses in response to pregnancy

Urticaria (hives) and dermographism (wealing in response to pressure, for example scratching) may be precipitated by pregnancy. Erythema multiforme due to pregnancy has been described.

Dermatoses and the effect of pregnancy

Atopic eczema

Atopic eczema is the commonest skin problem presenting in pregnancy. It can be severe and life ruining, and life threatening if secondary infection with herpes simplex (eczema herpeticum) or *Streptococcus* occurs. The effect of pregnancy on pre-existing atopic eczema is unpredictable: the immunosuppression can lead to improvement, but often there is deterioration of the eczema. The eczema becomes more widespread and may result in erythroderma in the most severe cases. Secondary infection with *Staphylococcus aureus* and *Streptococcus* is a frequent complication. The skin is red, dry, and scaly with areas of excoriation and thickening or lichenification.

Treatment is a major problem in pregnancy, as there is a dilemma in balancing the need for treatment with the wish to minimize the use of potent topical steroids which will be absorbed and may affect the foetus. The use of emollients may lessen the requirements for topical steroids, and steroids should be used in the minimum quantities and strengths necessary to control the disease (see Table 1). Many topical steroids contain antiseptics and antibiotics which will be absorbed and may be contraindicated in pregnancy. The sedating antihistamine, chlorpheniramine, may help with sleep. Secondary infection often requires systemic antibiotics such as erythromycin or flucloxacillin.

Psoriasis

Psoriasis may improve or deteriorate during pregnancy. Therapy poses special problems as all the systemic treatments are contraindicated: methotrexate is a folic acid antagonist, acitretin is teratogenic, ciclosporine results in intrauterine growth retardation, and psoralens with UVA are still not proven to be safe. Topical therapy with steroids should be avoided if possible. Coal tars and dithranol have been widely used in pregnancy but are not proven to be safe, and the new vitamin D analogues are not licensed for use in pregnancy. The ideal is minimum treatment, with emollients and if necessary UVB.

A severe form of pustular psoriasis, impetigo herpetiformis, may occur in pregnancy and is best managed with bedrest and emollients.

Autoimmune dermatoses in pregnancy

Cutaneous lupus erythematosus

Cutaneous lupus erythematosus does not seem to be adversely affected or improved by pregnancy. However such patients should be screened for anti-Ro and anticardiolipin antibodies etc. (see Chapter xxx), preferably prior to conception, to identify at-risk pregnancies.

Autoimmune bullous diseases

Linear IgA disease, an autoimmune blistering disease with IgA basement membrane zone antibodies, usually improves with pregnancy, such that some patients can discontinue their dapsone therapy. Despite the deposition of immunoreactants in the amnion basement membrane zone the fetus is not adversely affected. There is usually an exacerbation 3 months postpartum.

Pemphigus vulgaris, an autoimmune blistering disease with widespread mucosal and cutaneous erosions caused by antibodies to desmosomal components, can be transmitted across the placenta, with devastating results to the fetus. This does not occur in the related pemphigus foliaceus, which is endemic in Brazil.

Further reading

Collier P, Kelly SE, Wojnarowska F (1993). Linear IgA disease and pregnancy. *Journal of the American Academy of Dermatology* **30**, 407–12.

Holmes RC, Black MM, Dann J, *et al*. (1982). A comparative study of toxic erythema of pregnancy and herpes gestationis. *British Journal of Dermatology* **106**, 499–510.

Jenkins RE, Hern S, Black MM (1999). Clinical features and management of 87 patients with pemphigoid gestationis. *Clinical and Experimental Dermatology* **24**, 255–9.

Kelly SE, Wojnarowska F (1993). Pemphigoid gestationis. *European Journal of Dermatology* **4**, 16–20.

Muller S, Stanley JR (1990). Pemphigus: pemphigus vulgaris and pemphigus foliaceus. In: Wojnarowska F and Briggaman RA, eds. *Management of blistering disease*, pp 43–62.

Vaughan Jones SA, Hern S, Nelson-Piercy C, Seed PT, Black MM (1999). A prospective study of 200 women with dermatoses of pregnancy correlating clinical findings with hormonal and immunopathological profiles. *British Journal of Dermatology* **141**, 71–81.

13.14 Autoimmune disorders and vasculitis in pregnancy

Catherine Nelson-Piercy and Munther A. Khamashta

Autoimmune diseases affect 5 to 7 per cent of the population, are commoner in women of child-bearing age, and are frequently encountered in pregnancy. Pregnancy is associated with suppressed cell-mediated immunity and enhancement of humoral immunity, but these changes revert postpartum accompanied by sudden reductions of oestrogen, progesterone, and cortisol levels. The postpartum period is therefore a time of susceptibility to autoimmune disorders and women who already have an autoimmune disorder may suffer disease exacerbation following pregnancy. Conversely, autoimmune diseases may remit or improve during pregnancy, but this is not a universal rule and autoimmune rheumatic/connective tissue diseases can flare or present in pregnancy with disastrous consequences. This chapter considers the relationship between pregnancy and rheumatoid arthritis, systemic lupus erythematosus (SLE), antiphospholipid syndrome (APS), vasculitides, and scleroderma, and how pregnancy affects treatment of these disorders. The management of these conditions during pregnancy provides the obstetrician and physician with particular challenges and concerns related to not only the mother but also the fetus.

Rheumatoid arthritis

The adult form of the disease is more common in women (female to male ratio = 3:1), and approximately 1 in every 1000 to 2000 pregnancies are affected.

Effect of pregnancy on rheumatoid arthritis

Up to 75 per cent of women with rheumatoid arthritis experience improvement during pregnancy. Originally it was thought that this was the result of raised cortisol levels but these do not correlate with symptoms. Another hypothesis is a maternal immune response to fetal paternally inherited HLA class II gene products, supported by the finding that maternal–fetal disparities for HLA DQA were observed in 78 per cent of women whose rheumatoid arthritis went into remission during pregnancy, but in only 25 per cent in those whose disease remained active. Other theories attribute the improvement to high oestrogen levels, although ethinyl oestradiol given to postmenopausal women with rheumatoid arthritis is not effective. Pregnancy specific proteins such as a $_2$-glycoprotein (PAG) have also been implicated, and in experimental models a$_2$-glycoprotein improves arthritis. Removal of immune complexes by the placenta is another postulated mechanism.

Improvement usually begins during the first trimester, when rheumatoid nodules may also disappear, but 90 per cent of those who experience remission suffer postpartum exacerbations. The largest study of disease activity in pregnancy in 140 women with rheumatoid arthritis confirmed improvement in joint swelling and pain in two-thirds of subjects by the third trimester. However, only a minority had no joints with active disease, disability as assessed by the Health Assessment Questionnaire changed little compared to prepregnancy, and only 16 per cent went into complete remission. Disease response in a previous pregnancy was predictive of response in the index pregnancy. An increase in the mean number of inflamed joints was seen postpartum, but this could not be predicted from previous puerperal relapse. In women without the condition, there is an increased incidence of rheumatoid arthritis onset in the postpartum period.

Effect of rheumatoid arthritis on pregnancy

Unlike SLE, there seems to be no adverse effect of rheumatoid arthritis on pregnancy, and neither the fertility rate or spontaneous abortion rate is significantly altered. However, the infants of woman who have anti-Ro antibodies are at risk of neonatal lupus (see below). Atlantoaxial subluxation is a rare complication of a general anaesthetic for a caesarean section, and very rarely limitation of hip abduction is severe enough to impede vaginal delivery. The main concerns relate to the safety during pregnancy and lactation of the medications used to treat rheumatoid arthritis, although only 20 to 30 per cent of pregnant women with rheumatoid arthritis will require medications to control flares or systemic disease.

Treatment of rheumatoid arthritis (and other rheumatic disorders) in pregnancy

Paracetemol should be the first line analgesic, and there are no known adverse effects in pregnancy. Aspirin and non-steroidal anti-inflammatory drugs (NSAIDs) are not teratogenic, but salicylates (in analgesic doses) and NSAIDs may increase the risk of neonatal haemorrhage via inhibition of platelet function. NSAIDs may also lead to oligohydramnios via effects on the fetal kidney, and as they are prostaglandin synthetase inhibitors may cause premature closure of the ductus arteriosus (because prostaglandin E_2 relaxation of pulmonary vessels is inhibited) with neonatal primary pulmonary hypertension. They are usually avoided in pregnancy, especially in the last trimester, but the risk to the ductus arteriosus may have been exaggerated since premature closure has not been encountered when indomethacin is used for the treatment of premature labour. Impairment of ductal flow is rare before 27 weeks and resolves within 24 h of NSAID discontinuation. Oligohydramios is also reversible. In occasional circumstances, and especially prior to 28 weeks gestation, NSAIDs may be used for control of arthritic pain if there are relative contraindications to steroids, for example in patients on heparin. They should be discontinued at least 6 to 8 weeks prior to delivery. The recently introduced cyclo-oxygenase type-2-selective (COX-2) NSAIDs, although currently contraindicated in pregnancy, have been reported to show only minor renal and no ductal effects on the fetus when used to prevent premature labour.

Corticosteroids may be continued during pregnancy and are preferable to NSAIDs if paracetemol is insufficient to control symptoms in the third trimester. Prednisolone is metabolized by the placenta and very little (10 per cent) active drug reaches the fetus, unlike dexamethasone and betamethasone which cross the placenta more readily. Exceedingly large doses of cortisone are associated with an increased incidence of cleft palate in rodents, and one case–control study has suggested an increased risk of cleft lip following first-trimester exposure. There are many other studies supporting no increased risk of abortion, stillbirth, congenital malformations, adverse fetal effects, or neonatal death attributable to maternal steroid therapy. However, in women with antiphospholipid syndrome, treated with high doses of prednisolone throughout pregnancy, an increased frequency of premature rupture of the membranes has been reported. Although suppression of the fetal hypothalamic–pituitary–adrenal axis is a theoretical possibility with maternal systemic steroid therapy, there is no evidence that this occurs. Corticosteroid usage in pregnancy increases the risk of gestational diabetes, infection, and osteoporosis. If a woman is on long-term maintenance steroids, parenteral steroids should be administered to cover the stress of labour and delivery. Prednisolone is safe in breast-feeding mothers since <10 per cent of active drug is secreted into breast milk.

The alkylating agents cyclophosphamide and chlorambucil, and the folic acid antagonist methotrexate, are all teratogenic and fetotoxic and are contraindicated in pregnancy and lactation. Methotrexate should be discontinued at least 3 months prior to conception and folic acid supplementation given preconceptually. Azathioprine, the commonest cytotoxic used in rheumatoid arthritis and SLE, seems safe based on its successful use in large numbers of renal transplant mothers and women with SLE. However, neonatal immunosuppression has been noted and conflicting information exists regarding breast feeding while taking azathioprine. Cyclosporin has been associated with a higher rate of intrauterine growth restriction (IUGR) than azathioprine in renal transplant patients and should be avoided if possible.

D-penicillamine, a chelating agent used particularly in the management of the extra-articular features of rheumatoid arthritis, crosses the placenta and is teratogenic, with a 5 per cent risk of congenital collagen defect. It should therefore be stopped before conception in women with rheumatic diseases, but continued use is crucial for successful outcome of pregnancy in Wilson's disease, where there are about 90 reported cases. Gold salts are teratogenic in animals, but there is no conclusive evidence for such an effect in humans. They can therefore be continued during pregnancy if they are controlling disease, although most would avoid initiation of treatment during pregnancy.

Antimalarials, such as hydroxychloroquine, used in rheumatoid arthritis and particularly in subacute cutaneous lupus, are safe in doses used for malarial prophylaxis. There has been concern over larger doses, when concentration in the fetal uveal tract may result in retinopathy. Nevertheless, in 215 pregnancies in women exposed to chloroquine the congenital abnormality rate was no higher than background. There is also increasing experience of the use of hydroxychloroquine in pregnant women with SLE. It seems to have no adverse effect on the neonate, and should be continued through pregnancy for two reasons. Firstly, cessation may precipitate a flare of lupus. Secondly, it has a very long half-life, such that discontinuation will not prevent fetal exposure.

Sulphasalazine, another second-line agent, has been used extensively in the treatment of inflammatory bowel disease in pregnancy and appears to be safe. It may be continued throughout pregnancy, although concomitant folate supplementation is recommended.

Systemic lupus erythematosus (SLE)

SLE is much more common in women than men (ratio 9:1), particularly during the child-bearing years, (ratio 15:1). The prevalence is approximately 1 per 1000 women and may be increasing. The fundamental issues concerning the effect of SLE on pregnancy are the presence or absence of anti-Ro/La and antiphospholipid antibodies (see below), the activity of the disease, and the presence or absence of hypertension and renal involvement.

Effect of pregnancy on SLE

SLE flares may be difficult to diagnose during pregnancy since many features such as hair fall, oedema, facial erythema, fatigue, anaemia, raised ESR, and musculoskeletal pain also occur in normal pregnancy. Whether pregnancy exacerbates SLE and increases the likelihood of flare postpartum is controversial. Six controlled studies have addressed this issue. These differ in the ethnicity of the populations, the criteria for flare, and the SLE activity scales employed. Three found no increased risk of deterioration in pregnancy; three, including the most recent study from St Thomas', suggested that SLE was more likely to flare during pregnancy and the puerperium. The studies are consistent in showing that 58 to 70 per cent of women flare during pregnancy. Steroids do not prevent these flares, and it is not our practice to prescribe prophylactic steroids or increase steroid dosage prophylactically during pregnancy or postpartum.

In 242 pregnancies in 156 women with lupus nephritis, kidney function was unchanged in 59 per cent, transiently impaired in 30 per cent, and permanently deteriorated in 7.1 per cent. As in all types of renal disease, there is a greater risk of deterioration in patients with hypertension, heavy proteinuria, and high baseline serum creatinine.

Effect of SLE on pregnancy

SLE is associated with increased risks of spontaneous abortion, fetal death, pre-eclampsia, preterm delivery, and IUGR. A prospective study of 108 pregnancies in 90 lupus patients showed a live birth rate of 82 per cent, a 43 per cent incidence of prematurity, and a 30 per cent incidence of IUGR. Most of the fetal losses were in association with secondary antiphospholipid syndrome. Pregnancy outcome is particularly affected by renal disease. Even quiescent renal lupus is associated with increased risk of fetal loss, pre-eclampsia, and IUGR, particularly if there is hypertension or proteinuria. For women with SLE in remission, and without hypertension, renal involvement, or the antiphospholipid syndrome, the risk of problems in pregnancy is probably no higher than in the general population.

Management of SLE in pregnancy

When possible, this should begin with preconception counselling. Knowledge of the anti-Ro/La antiphospholipid antibody, and renal and blood pressure status allows prediction of the risks to the woman and her baby (see below). The outlook is better if conception occurs during remission. Pregnancy care is best undertaken in multidisciplinary, combined clinics where physicians and obstetricians can monitor disease activity and fetal growth and uterine and umbilical artery Doppler blood flow regularly.

Disease flares must be actively managed. Corticosteroids are the drug of choice (see above for discussion of anti-inflammatory agents and immunosuppression in pregnancy). Hydroxychloroquine should be continued since stopping may precipitate flare. Azathioprine is also usually continued, since this acts as a 'steroid-sparing' agent. Differentiation of active renal lupus from pre-eclampsia is notoriously difficult and the two conditions may be superimposed. Since hypertension, proteinuria, thrombocytopenia, and renal impairment are all features of pre-eclampsia, diagnosis of lupus flare requires other features, such as a rising anti-dsDNA antibody titre, the presence of red blood cells or cellular casts in the urinary sediment, or a fall in complement levels. Elevation of complement split products, particularly Ba and Bb, often accompanies flares so high ratios of CH50/Ba may differentiate pre-eclamptics from those with active lupus. The only definitive investigation to reliably differentiate a renal lupus flare from pre-eclampsia is renal biopsy, but this is rarely undertaken in pregnancy.

For control of hypertension, the drug of choice is methyl dopa, with nifedipine or hydrallazine as second-line agents.

Neonatal lupus syndromes

These conditions are models of passively acquired autoimmunity. Autoantibodies directed against cytoplasmic ribonucleoproteins Ro and La cross the placenta causing immune damage in the fetus. Several clinical syndromes have been described, of which cutaneous neonatal lupus is the most common, and congenital heart block is the most serious. They rarely coexist. More than 90 per cent of mothers of affected offspring have anti-Ro antibodies, and 50 to 70 per cent anti-La antibodies. About 30 per cent of patients with SLE are anti-Ro positive, and in such women the risk of transient cutaneous lupus is about 5 per cent and the risk of congenital heart block about 2 per cent. The risk of neonatal lupus is increased if a previous child has been affected, rising to 16 per cent with one affected child and 50 per cent if two children are affected. It is important to recognize that not all Ro-positive mothers of neonates with congenital heart block have SLE. A large proportion are asymptomatic, but 48 per cent of these developed symptoms of connective tissue disease in a mean of 2.6 years in one study. Mothers of babies with neonatal lupus have a higher frequency of HLA-DR3, often with A1 and B8. There is no correlation between the severity of maternal disease and the incidence of neonatal lupus.

The cutaneous form of neonatal lupus usually manifests in the first 2 weeks of life. The infant develops typical geographical skin lesions similar to those of adult subacute cutaneous lupus, usually of the face and scalp, which appear after sun or UV light exposure. The rash disappears spontaneously within six months, suggesting a direct antibody-mediated mechanism. Residual hypopigmentation or telangiectasia may persist for up to 2 years, but scarring is unusual. Sunlight and phototherapy should be avoided.

Congenital heart block appears *in utero*, usually around 18 to 20 weeks, and may be fatal. The mechanism of damage is not fully understood. There is no treatment that reverses congenital heart block, although salbutamol given to the mother may be beneficial if bradycardia is causing fetal heart failure. Dexamethasone and plasmapheresis have been used to treat non-immune hydrops resulting from myocarditis and pericarditis, but have no effect on the conduction defect. Perinatal mortality is increased with one-fifth of affected children dying in the early neonatal period, but most infants who survive this period do well, although two-thirds require pacemakers.

Antiphospholipid syndrome

Anticardiolipin antibodies and lupus anticoagulant (LA) are overlapping subsets of antiphospholipid antibodies. The combination of either of these with one or more of the characteristic clinical features (thrombosis, recurrent pregnancy loss, or adverse pregnancy outcome—premature birth before 34 weeks due to pre-eclampsia or IUGR) is known as the antiphospholipid syndrome (APS). Other associated features include thrombocytopenia and haemolytic anaemia, livedo reticularis, cerebral involvement (particularly epilepsy, cerebral infarction, chorea, and migraine), heart valve disease (particularly of the mitral valve), systemic and pulmonary hypertension, and leg ulcers. APS was first described in patients with SLE, but it is now recognized both that most patients with APS do not fulfil the diagnostic criteria

for SLE, and that those with primary APS do not usually progress to SLE. Although the clinical features of primary and SLE-associated APS are similar, and the antibody specificity is the same, the distinction is important, and patients with primary APS should not be labelled as 'lupus'.

Recurrent pregnancy loss, typically in the second trimester, is one of the most consistent features of APS. Fetal death is typically preceded by IUGR, oligohydramnios, and features of pre-eclampsia. Fetal loss represents one part of a spectrum of fetal compromise and a wide range of pregnancy morbidity has been reported in APS including recurrent first trimester miscarriage, second and third trimester loss, severe early-onset pre-eclampsia, IUGR, placental abruption, and prematurity. Since the classification criteria for APS have recently been amended, a patient with adverse pregnancy outcome may now be labelled as APS without a history of fetal loss. The risk of fetal loss is directly related to antibody titre, particularly the IgG anticardiolipin antibodies, although many women with a history of recurrent loss have only IgM antibodies. Quantifying the risk is difficult, and the presence of antiphospholipid antibodies does not preclude successful pregnancy. The antibodies should be regarded as markers for a high-risk pregnancy, but previous poor obstetric history remains the most important predictor of fetal loss in these women.

The prevalence of antiphospholipid antibodies in the general obstetric population is low (<2 per cent), so universal screening is not warranted. However, the prevalence of antiphospholipid antibodies is increased in women with pregnancy complications including severe early-onset pre-eclampsia, abruption, intrauterine fetal death, or IUGR without hypertension. In one study, 29 per cent of women with severe early-onset pre-eclampsia were found to have anticardiolipin antibodies compared with 2 per cent of controls.

The pathogenesis of fetal loss in these patients is not fully understood, but there is typically massive infarction and thrombosis of the placental and decidual vessels, probably secondary to spiral artery vasculopathy. One hypothesis is that anticardiolipin antibodies cause thrombosis by binding to co-factor b $_2$-glycoprotein, an endogenous coagulation inhibitor. Platelet deposition and prostanoid imbalance may be implicated, as they might be in pre-eclampsia.

Studies on pregnancy outcome in women known to have APS show differing rates of obstetric complications depending on their presentation. Those found to have APS as a result of recurrent miscarriage have lower rates of complications (see Table 1) than those discovered because of late losses, thrombosis, or other systemic manifestations.

Management of antiphospholipid syndrome in pregnancy

The management of pregnancy in women with APS is the subject of much debate. Aspirin inhibits thromboxane and may reduce the risk of vascular thrombosis, but its use as a single agent in APS pregnancy has only been subjected to randomized clinical trial in low-risk women. It was not found to be beneficial. There are several non-randomized studies in women with fetal loss suggesting that it is effective, and it can prevent pregnancy loss in experimental APS mice. Most centres now advocate low-dose aspirin (75 mg) for all women with APS, some even prior to conception, in the belief that the placental damage occurs early in gestation, and that aspirin prevents failure of placentation. We also advocate intra and postpartum (3–5 days) prophylaxis with heparin, especially in the event of caesarean section.

APS is the most frequent cause of acquired thrombophilia. However, unlike the congenital thrombophilias, thrombosis can be arterial or venous, and affects vessels of all sizes. The risk of recurrent thrombosis in patients with APS may reach 70 per cent, and women with APS and previous thromboembolism are at extremely high risk in pregnancy and the puerperium. Warfarin should be stopped and heparin started before 6 weeks gestation to avoid warfarin embryopathy. Subcutaneous unfractionated (10 000 u twice daily) or low-molecular weight heparin (enoxaparin 40 mg once daily; dalteparin 5000 once daily-twice daily) should be continued intrapartum and postpartum until warfarin has been reintroduced. It is occasionally necessary to use warfarin in women with previous arterial thromboses if heparin proves inadequate to prevent further transient ischaemic events.

Opinion is divided about the best antenatal therapy for those with recurrent pregnancy loss, but without a history of thromboembolism. Treatment with high-dose steroids (in the absence of active lupus) to suppress lupus anticoagulant and anticardiolipin antibodies, in combination with aspirin, was initially recommended because of improved (50 per cent) fetal survival compared to historical controls. However, high doses of prednisolone caused considerable maternal morbidity, and subsequent studies have failed to demonstrate better fetal outcome. Whether this can be improved by adding heparin to low-dose aspirin in women with recurrent miscarriage but without a history of thrombosis has been the subject of several studies, but remains controversial. Many centres have traditionally reserved the addition of heparin for those women with previous late losses or intrauterine deaths, and using such strategies live birth rates of 70 to 75 per cent can be achieved. However, two studies in different populations of women have suggested that aspirin and heparin can improve the live birth rate from about 40 per cent to about 70 to 80 per cent in woman with a history of recurrent miscarriage. In both, the excess losses in the aspirin alone group occurred prior to 13 weeks, after which outcomes were similar. This suggests that the beneficial effects of heparin occur before 13 weeks and raises the possibility that heparin may be stopped after this time if there is no history of thrombosis or a late loss. A more recent randomized controlled trial demonstrated no increase in live birth rate with aspirin and LMWH compared to aspirin alone.

Immunosuppression with azathioprine, intravenous immunoglobulin, and plasmapheresis have all been tried, but the numbers treated do not allow firm conclusions regarding efficacy. Trials of intravenous immunoglobulin in recurrent miscarriage have been stopped prematurely because of its cost and lack of obvious benefit.

Pregnancy complicated by APS requires expert care and a team approach by obstetricians, physicians, and haematologists. Close monitoring of both mother and fetus is essential. Ultrasound monitoring of fetal growth and uteroplacental blood flow is crucial, allowing for timely delivery. Uterine artery waveforms are assessed at 20 and 24 weeks gestation and those pregnancies with evidence of an early diastolic notch are monitored very closely with 2-weekly growth scans because of the high risk of IUGR. Where there are no notches, we recommend 4-weekly assessment of growth and amniotic fluid volume. Doppler flow studies of the umbilical artery may be used, as in other pregnancies at high risk of fetal compromise through uteroplacental insufficiency.

Vasculitis

Since the primary vasculitides occur principally in the post child-bearing years and are commoner in men, pregnancy is very uncommon in patients with Wegener's granulomatosis, polyarteritis nodosa (PAN), or Churg–Strauss vasculitis. Less than 50 pregnancies have been reported in total. In general, maternal and fetal outcome are dependent on disease activity. Reported cases would suggest that disease onset or flare is more likely during pregnancy or postpartum. Maternal death occurred in two of 20 pregnancies in women with Wegener's granulomatosis, none of seven pregnancies (in four women) in Churg–Strauss, and in all of seven women in whom PAN was diagnosed during pregnancy. Fetal outcome is usually successful in controlled or remitted disease, but active disease is associated with fetal demise. In the 20 pregnancies in Wegener's granulomatosis there were two intrauterine fetal deaths (both in women with active disease), four elective terminations, and the other infants survived. Despite the high maternal postpartum mortality in PAN, there were only two intrauterine deaths and nine surviving infants. Neonatal cutaneous vasculitis, resolving with treatment, has been reported in infants of mothers with cutaneous PAN.

In view of the significant maternal and fetal morbidity and mortality associated with active disease, it is important to adopt an aggressive approach to treatment with immunosuppression in the case of flare. Azathioprine is the drug of choice in pregnancy if immunosuppression with corticosteroids is insufficient, but life-threatening disease may necessitate the use of pulsed cyclophosphamide despite the risks.

Scleroderma

Although a rare disease, scleroderma is more common in women (female to male ratio = 3:1) and this is especially so during the reproductive years (10:1). Previously it was thought that fertility may be impaired in women with or destined to develop scleroderma, but more recent studies have demonstrated normal pregnancy rates in scleroderma. Indeed pregnancy may have an aetiological role: persistent fetal microchimerism being more common in women with scleroderma than controls, leading to the hypothesis that fetal antimaternal graft–versus–host reactions may be involved in pathogenesis (see Chapter 18.11.3).

Effect of pregnancy on scleroderma

Early case reports highlighted the risk from renal crises, even in women without a prior history of renal involvement, and renal disease was the commonest cause of death in these pregnancies. However, more recent case series and case–control studies have reported far fewer renal crises. A retrospective study of 86 pregnancies found no change in symptoms in 88 per cent, improvement in 5 per cent, and deterioration in 7 per cent. Ten-year survival of women with scleroderma, with or without pregnancy, was similar. A prospective series of 89 pregnancies in 58 women found stability in 61 per cent, improvement in 20 per cent, and worsening in 19 per cent. There were only five cases of renal crisis in these two series and all occurred in women with early diffuse scleroderma—the highest risk group for renal crises even outside pregnancy. Progressive cutaneous disease is unusual during or immediately after pregnancy. Raynaud's phenomenon usually improves in pregnancy as a

result of vasodilation and increased blood flow. Reflux and oesophagitis often worsen related to the decreased lower oesophageal tone of pregnancy. Arthralgias also worsen. There is no evidence that pregnancy worsens cardiac or respiratory disease, although those with severe pulmonary fibrosis and pulmonary hypertension are at extremely high risk of postpartum deterioration, as with pulmonary hypertension from any cause.

In general, women with limited scleroderma without organ involvement do better than those with diffuse disease. The extent of diffuse disease and systemic involvement (particularly lung, cardiac, and renal) influences prognosis, but there are no absolute rules. Those with early (less than 4 years) disease, diffuse disease, or antitopoisomerase (anti-ScL-70) antibodies are at greater risk of having more active aggressive disease than those with long-standing disease and anticentromere antibodies. Women with renal involvement often have associated hypertension and rapid deterioration is possible.

Effect of scleroderma on pregnancy

Although early literature suggested an increased risk of miscarriage in scleroderma, case–control studies found no increase when women with scleroderma were compared to those with rheumatoid arthritis or normal controls. A more recent, prospective study did find a 22 per cent incidence of miscarriage (compared with 13 per cent in controls) in women with diffuse disease. There is an increased risk of premature delivery (24 per cent in limited disease and 33 per cent in diffuse disease, versus 5 per cent in controls) and IUGR, but overall success is now reported to be 70 to 80 per cent. The risks of adverse outcome are highest for women with early diffuse disease. There is no increased risk of pre-eclampsia in the absence of hypertension or renal involvement.

Management

Women should be assessed prior to conception for the extent of organ involvement. Those with renal impairment, severe cardiomyopathy, severe restrictive lung disease or pulmonary hypertension should be advised against pregnancy. Those with early diffuse disease should delay pregnancy until the disease stabilizes. Disease-remitting drugs such as D-penicillamine and cyclosporin A should preferably be discontinued before conception if disease is stable, although inadvertent first trimester exposure should not cause undue concern.

High-level, joint obstetric and medical care is appropriate with frequent and regular multidisciplinary monitoring of disease activity and fetal growth. Particular attention to blood pressure monitoring is essential: hypertension may indicate a renal crisis. Coincident renal impairment or microangiopathic haemolytic anaemia should prompt the initiation of an ACE inhibitor. These drugs have revolutionized the management and survival in renal scleroderma and should not be withheld in pregnancy. They are usually contraindicated, but in scleroderma the benefit outweighs the risk of fetal renal toxicity.

Management of scleroderma during pregnancy is largely symptomatic. Calcium antagonists may be used for Raynaud's phenomenon and histamine blockers and proton-pump inhibitors for reflux. NSAIDs are best avoided, as previously discussed, and corticosteroids (more than 15 mg/day) must also be avoided in early diffuse scleroderma since they can precipitate a renal crisis. Caution is needed if b-adrenergic agonists are required for preterm labour since scleroderma patients may have silent myocardial damage making them more vulnerable to ischaemia and pulmonary oedema. Venepuncture, venous access, and blood pressure measurement may be difficult because of skin or blood vessel involvement. General anaesthesia may be complicated by difficult endotracheal intubation, and regional anaesthesia may also be difficult. Early assessment by an obstetric anaesthetist is advisable and epidural anaesthesia and analgesia is encouraged as vasodilation improves skin perfusion of the extremities. Other measures to reduce problems related to Raynaud's phenomenon include warming of the delivery room and any intravenous fluids as well as socks and gloves.

Close observation must continue in the immediate postnatal period, particularly in those with cardiac, pulmonary, or renal involvement. The most recent data, although derived from retrospective postal questionnaire, suggest that outcomes are not significantly worse than controls, provided pregnancy is well-timed and carefully monitored. Clinicians are therefore becoming more optimistic in counselling women with scleroderma considering pregnancy.

Further reading

Antirheumatic drugs and immunosuppressive agents in pregnancy

Bermas BL, Hill JA (1995). Effects of immunosuppressive drugs during pregnancy. *Arthritis and Rheumatism* **38**,1722–32. [Excellent in-depth review.]

Ostenson M (1998). Nonsteroidal anti-inflammatory drugs during pregnancy. Second international conference on rheumatic diseases in pregnancy. *Scandinavian Journal of Rheumatology* **27** (Suppl 107), 128–32. [Succinct review.]

Ostenson M, Ramsey-Goldman R (1998). Treatment of inflammatory rheumatic disorders in pregnancy. *Drug Safety* **19**, 389–410. [Comprehensive review.]

Ramsey-Goldman R (1998). The risk of cytotoxic drugs during pregnancy. *Scandinavian Journal of Rheumatology* **27** (Suppl 107), 133–5. [Succinct review.]

Rheumatoid arthritis

Barrett JH *et al*. (1999). Does rheumatoid arthriitis remit during pregnancy and relapse postpartum? Results from a nationwide study in the United Kingdom performed prospectively from late pregnancy. *Arthritis and Rheumatism* **42**, 1219–27. [Excellent report and review of the current knowledge.]

Nelson JL, Ostensen M (1997). Pregnancy and rheumatoid arthritis. *Rheumatic Disease Clinics of North America* **23**, 195–212. [Comprehensive review.]

Silman AJ (1998). Reproductive events and the risk of development of rheumatoid arthritis. Second international conference on rheumatic diseases in pregnancy. *Scandinavian Journal of Rheumatology* **27** (Suppl 107), 113–5. [Succinct review.]

Systemic lupus erythematosus

Buchanan NMM *et al*. (1996). Hydroxychloroquine and lupus pregnancy: a review of a series of 36 cases. *Annals of the Rheumatic Diseases* **55**, 486–8. [Largest reported series of hydroxychloroquine use in pregnancy.]

Khamashta MA, Hughes GRV (1997). Pregnancy in systemic lupus erythematosus. *Current Opinion in Rheumatology* **8**, 424–9. [Comprehensive review.]

Khamashta MA, Ruiz-Irastoza G, Hughes GRV (1997). Systemic lupus erythematosus flares during pregnancy. *Rheumatic Disease Clinics of North America* **23**,15–30. [Review of studies examining frequency of lupus flares in pregnancy.]

Lima F *et al*. (1995). Obstetric outcome in systemic lupus erythematosus. *Seminars in Arthritis and Rheumatism* **25**, 184–92. [Report of large series of lupus pregnancies.]

Oviasu E, Hicks J, Cameron JS (1991). The outcome of pregnancy in women with lupus nephritis. *Lupus* **1**, 19–25. [Important study suggesting that women with lupus nephritis should not be discouraged from becoming pregnant.]

Parke AL (1998). Antimalarial drugs, systemic lupus erythematosus and pregnancy. *Journal of Rheumatology* **15**, 607–10. [Comprehensive review.]

Neonatal lupus

Tseng CE, Buyon JP (1997). Neonatal lupus syndromes. *Rheumatic Disease Clinics of North America*, **23**, 31–54.

Waltuck J, Buyon JP (1994). Autoantibody-associated congenital heart block: outcome in mothers and children. *Annals of Internal Medicine* **120**, 544–51. [Largest recorded series of congenital heart block, long-term outcome.]

Antiphospholipid syndrome

Backos M *et al*. (1999). Pregnancy complications in women with recurrent miscarriage associated with antiphospholipid antibodies treated with low dose aspirin and heparin. *British Journal of Obstetrics and Gynaecology* **106**, 102–7. [A study describing obstetric outcome in different populations of treated APS pregnancies.]

Cowchock S, Reece A (1997). Do low-risk pregnant women with antiphospholipid antibodies need to be treated? *American Journal of Obstetrics and Gynecology* **176**, 1099–100. [Randomized controlled trial of aspirin for asymptomatic aPL positive.]

Cowchock FS, Reece EA, Balaban D, Branch DW, Plouffe L (1992). Repeated fetal losses associated with antiphospholipid antibodies: a collaborative randomized trial comparing prednisone with

low-dose heparin treatment. *American Journal of Obstetrics and Gynecology* **166**, 1318–23. [Important trial showing aspirin and heparin is better than aspirin and prednisolone.]

Dekker GA *et al*. (1995). Underlying disorders associated with severe early-onset pre-eclampsia. *American Journal of Obstetrics and Gynecology* **173**, 1042–8. [Excellent study showing that more than one-fifth of severe early-onset PET is aPL-related.]

Gordon C, Kilby MD (1998). Use of intravenous immunoglobulin therapy in pregnancy in systemic lupus erythematosus and antiphospholipid antibody syndrome. *Lupus* **7**, 429–33. [Review article.]

Granger KA, Farquharson RG (1997). Obstetric outcome in antiphospholipid syndrome. *Lupus* **6**, 509–13. [A study describing obstetric outcome in different populations of treated APS pregnancies.]

Harris EN, Spinnato JA (1991). Should anticardiolipin tests be performed in otherwise healthy pregnant women? *American Journal of Obstetrics and Gynecology* **165**, 1272–7. [Large study proving that aPL screening is not warranted in healthy pregnant women.]

Kerslake S, Morton KE, Versi E, *et al*. (1992). Early Doppler studies in lupus pregnancy. *American Journal of Reproductive Immunology* **28**, 172–5. [Succinct review.]

Khamashta MA *et al*. (1995). The management of thrombosis in the antiphospholipid-antibody syndrome. *New England Journal of Medicine* **332**, 993–7. [Seminal study highlighting need for long-term warfarin in APS.]

Kutteh WH (1996). Antiphospholipid antibody-associated recurrent pregnancy loss: treatment with heparin and low-dose aspirin is superior to low-dose aspirin alone. *American Journal of Obstetrics and Gynecology* **174**, 1574–89. [A study suggesting aspirin and heparin superior to aspirin alone for first trimester miscarriage in APS.]

Langford K, Nelson-Piercy C (1999). Antiphospholipid syndrome in pregnancy. *Contemporary Reviews in Obstetrics and Gynaecology*, 11, 93-8. [Comprehensive review.]

Lima F *et al*. (1996). A study of sixty pregnacies in patients with the antiphospholipid syndrome. *Clinical and Experimental Rheumatology* **14**, 131–6. [A study describing obstetric outcome in different populations of treated APS pregnancies.]

Lockshin MD (1993). Which patients with antiphospholipid antibody should be treated and how? *Rheumatic Disease Clinics of North America* **19**, 235–47. [Comprehensive review.]

Lubbe WF *et al*. (1983). Fetal survival after prednisone suppression of maternal lupus-anticoagulant. *Lancet* **i**, 1361–3. [Seminal study showing improved pregnancy outcome with steroids.]

Nelson-Piercy C (1997). Hazards of heparin: Bleeding, allergy, heparin-induced thrombocytopenia, osteoporosis. In: Greer I, ed. *Thromboembolic disease in obstetrics and gynaecology.* Bailliere's Clinical Obstetrics and Gynaecology **11**, pp 489–509. [Review of side-effects of heparin.]

Nelson-Piercy C, Letsky EA, de Swiet M (1997). Low molecular weight heparin for obstetric thromboprophylaxis: experience of 69 pregnancies in 61 high risk women. *American Journal of Obstetrics and Gynecology* **176**, 1062–8. [Largest reported series of low molecular weight heparin use in pregnancy.]

Rai R, Cohen H, Dave M, Regan L (1997). Randomized controlled trial of aspirin and aspirin plus heparin in pregnant women with recurrent miscarriage associated with phospholipid antibodies (or antiphospholipid antibodies). *British Medical Journal* **314**, 253–7. [A study suggesting aspirin and heparin superior to aspirin alone for first trimester miscarriage in APS.]

Vianna JL *et al*. (1994). Comparison of the primary and secondary antiphospholipid syndrome: a European multicenter study of 114 patients. *American Journal of Medicine* **96**, 3–9. [First study to compare primary and secondary APS.]

Ware-Branch D (1994). Thoughts on the mechanism of pregnancy loss associated with the antiphospholipid syndrome. *Lupus* **3**, 275–80. [Review of pathophysiology.]

Ware-Branch D *et al*. (1992). Outcome of treated pregnancies in women with antiphospholipid syndrome: an update of the Utah experience. *Obstetrics and Gynecology* **80**, 614–20. [A study describing obstetric outcome in different populations of treated APS pregnancies.]

Wilson WA *et al*. (1999). International consensus statement on preliminary classification criteria for definite antiphospholipid syndrome: Report of an international workshop. *Arthritis and Rheumatism* **42**, 1309–11. [Important paper discussing updated criteria for APS.]

Vasculitides

Lima F *et al*. (1995). Pregnancy in granulomatous vasculitis. *Annals of the Rheumatic Diseases* **54**, 604–6. [Case series and literature review.]

Ramsey-Goldman R (1998). The effect of pregnancy on the vasculitides. *Scandinavian Journal of Rheumatology* **27** (Suppl. 107), 116–7. [Succinct review.]

Scleroderma

Artlett CM, Smith B, Jimenez SA (1998). Identification of fetal DNA and cells in skin lesions from women with systemic sclerosis. *New England Journal of Medicine* **338**, 1186–91. [Microchimerism as a cause of scleroderma.]

Steen VD (1997). Scleroderma and pregnancy. *Rheumatic Disease Clinics of North America* **23**, 133–47. [Comprehensive review.]

Steen VD, Medsger TA (1998). Case-control study of corticosteroids and other drugs that either precipitate or protect from the development of scleroderma renal crisis. *Arthritis and Rheumatism* **41**, 1613–9. [Important study showing steroids may precipitate renal crisis.]

Steen VD, Medsger TA (1999). Fertility and pregnancy outcome in women with systemic sclerosis. *Arthritis and Rheumatism* **42**, 763–8. [Excellent and most recent study of this issue.]

13.15 Infections in pregnancy

Mark Herbert and Lawrence Impey

Introduction

Maternal immunity is suppressed in pregnancy and the fetal immune system is developmentally immature, thus infections in pregnancy can be devastating both for the mother, as is occasionally seen with varicella, and for the fetus, as is exemplified by congenital infections such as those caused by toxoplasmosis, syphilis, rubella, and cytomegalovirus (CMV).

Infections in pregnancy can be divided into four groups:

- Maternal illness made more severe by the pregnant state;
- Congenital infections acquired transplacentally, including *Toxoplasma gondii*, *Treponema pallidum*, rubella, CMV, *Listeria monocytogenes*, *Falciparum* spp., and *Trypanosoma* spp.;
- Fetal infection arising secondary to ascending maternal infection with preceding chorioamnionitis, caused by *Streptococcus agalactiae* and *Escherichia coli*;
- Neonatal sepsis acquired perinatally, such as HIV, HBV, and *Chlamydia trachomatis*.

Preterm delivery is an important cause of long-term handicap, and clinical evidence of infection in the placenta or neonate further increases the risk. Even with term delivery, maternal infection is associated with an increased rate of neonatal encephalopathy and cerebral palsy. This association is poorly understood. Chorioamnionitis is polymicrobial, and even *Candida* spp. have been implicated. It usually presents after preterm rupture of the membranes, but subclinical infection, usually ascending, may cause cervical changes and amniotic damage. Abdominal pain, fever, maternal and fetal tachycardia, and an offensive discharge are classic presentations, but chorioamnionitis is frequently asymptomatic in the early stages. Its management involves intravenous antibiotics and delivery, whatever the gestation.

In this chapter, we list the most important infective organisms in pregnancy, describe their maternal and fetal effects, and discuss their prevention, identification, and treatment. Detailed discussion of their pathology and features in adults is described elsewhere in the book.

Human immunodeficiency virus (HIV)

In the United Kingdom, the prevalence of HIV in pregnant women in inner London has risen from 0.04 per cent in 1989 to 0.4 per cent in 2000; in the rest of the United Kingdom, it is about 0.02 per cent. In parts of Africa, the seropositive rate among pregnant women exceeds 20 per cent, and an estimated 2.4 million HIV-infected women deliver every year.

Almost all HIV in infancy is acquired intrapartum or during breast feeding. Vertical transmission is highest in underdeveloped countries, if there are concomitant sexually transmitted diseases, in preterm delivery, and where the CD4 count is low and the viral load is high, as in early and late disease. Transmission rates are 15 to 20 per cent in Europe and North America and 25 to 35 per cent in Africa, India, and Thailand. Women with HIV are at greater risk of other sexually transmitted diseases and other coexisting disease, and consequently an increased vertical transmission rate and poorer obstetric outcomes. One-third of infected neonates die in infancy.

In the United Kingdom, viral antigen ELISA screening for HIV is offered to all pregnant women. Ideal management requires interdisciplinary co-operation, screening for other sexually transmitted diseases, and regular CD4 counts. Depending on the resources available, there are several drug regimens that are known to decrease vertical transmission, including:

- The PACTG 076 protocol: oral zidovudine 100 mg five times per day from 14 to 34 weeks gestation until labour; intravenous zidovudine in labour, 2 mg/kg intravenously over 1 h followed by 1 mg/kg per h until delivery, and oral zidovudine 2 mg/kg four times per day given to the infant for 6 weeks. Oral zidovudine given only to mother in late pregnancy, 300 mg two times per day from 36 weeks gestation, then every 3 h throughout labour.
- The HIVNET regimen of a single oral 200 mg dose of nevirapine given to mothers in labour and a single 2 mg/kg dose given to the infant at 48 to 72 h old. In sub-Saharan Africa, 110 000 HIV-positive births could be prevented over the next 5 years by this regimen.

Zidovudine and lamivudine together, from early second trimester are probably better than AZT alone. Occasional severe adverse effects in neonates have been reported. Antiretroviral therapy, elective caesarean section, and avoidance of breast feeding together reduce vertical transmission to less than 1 per cent. It is uncertain if elective caesarean section confers benefit in addition to retroviral therapy if the viral load is less than 1000 HIV copies/ml blood. Even in women on treatment there is a small but definite transmission rate with breast feeding, which should be avoided.

Rubella

A world-wide pandemic of rubella between 1962 and 1964 hastened the impetus for developing a vaccine, which has had a dramatic impact on the incidence of the congenital rubella syndrome. However, 10 to 20 per cent of the adult population in North America remain susceptible and small outbreaks still occur.

The incubation period is 16 to 18 days (range 14–21). There is a 1 to 5-day prodrome of low-grade fever, headache, malaise, anorexia, mild conjunctivitis, coryza, sore throat, cough, and lymphadenopathy (suboccipital, postauricular, and cervical) before the onset of a rash that lasts for 1 to 5 days. Arthritis and arthralgia occur in up to 70 per cent of adult women after the rash. Rare complications are thrombocytopenia, acute postinfectious encephalitis, myocarditis, Guillain–Barré syndrome, relapsing encephalitis, optic neuritis, bone marrow aplasia, and progressive panencephalitis.

The major sequelae in the fetus are cataracts, deafness, mental retardation, and heart disease, especially pulmonary arterial hypoplasia, patent ductus arteriosus, and coarctation of the aorta. The risk of the congenital rubella syndrome is greatest in early pregnancy: 90 per cent of infants will be affected before 11 weeks gestation, falling to 24 per cent at 15 to 16 weeks gestation. At 25 weeks, transmission to the fetus is approximately 25 per cent rising to 100 per cent at term, but the congenital rubella syndrome does not usually occur after the first trimester. Embryo resorption may occur in very early gestation, or abortion in later pregnancy.

Parvovirus B19

Parvovirus B19 binds to the P antigen present on erythrocytes, erythroblasts, and myocardium and can cause fetal anaemia (haemolytic and aplastic) and cardiac dysfunction. About 0.25 per cent of pregnant women are infected, but respiratory-borne epidemics, particularly in spring, may increase this fourfold. Half of adults are immune. Pregnancy does not alter the symptoms; 20 per cent are asymptomatic; the rest have a rubelliform or characteristic 'slapped cheek' rash, and 80 per cent with rash have arthralgia or arthritis.

Infection in pregnancy carries a 9 per cent fetal mortality rate, mainly arising from infection before 20 weeks gestation, and 3 per cent develop hydrops fetalis characterized by generalized fetal oedema, particularly ascites (Fig. 1). Two-thirds recover without intervention.

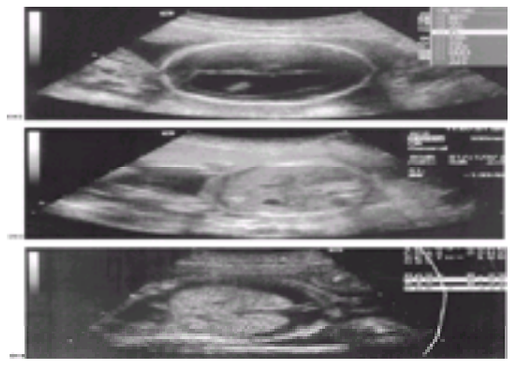

Fig. 1 Antenatal ultrasound scans: (a) fetal head showing ventriculomegaly secondary to congenital toxoplasmosis; (b) fetal abdomen showing intrahepatic calcification (marked with +) as seen in congenital varicella infection; (c) fetal abdomen showing ascites in parvovirus infection.

Maternal anti-B19 IgM antibody indicates recent infection. Ascertaining fetal involvement is more difficult because invasive sampling has a 1 per cent miscarriage risk and fetal IgM is not present before 22 weeks gestation. PCR for viral DNA can be performed on amniotic fluid, fetal blood, or other tissue. After diagnosis, regular ultrasound examination aids identification of fetal disease, and *in utero* transfusion of red cell concentrate via the umbilical vein is used to treat severe fetal hydrops. Adverse fetal effects are very rare more than 18 weeks after infection, thus regular ultrasound examinations are continued for this period.

Herpes simplex virus (HSV)

Genital herpes is predominantly caused by HSV-2, but up to 30 per cent of cases are due to HSV-1. Approximately 20 per cent of women in developed countries have been infected with HSV-2, but 90 per cent of these give no history. Primary infection in pregnancy occurs in 2 per cent of susceptible women. This is usually asymptomatic, but primary herpes may be characterized by genital pain and ulceration, discharge, dysuria, lymphoedema, and systemic symptoms. Acyclovir may be helpful in reducing the severity and frequency of recurrences. A rare but severe manifestation of HSV infection in pregnancy is isseminated disease, with necrotizing hepatitis, thrombocytopenia, leukopenia, disseminated intravascular coagulopathy, and more than 50 per cent mortality.

True congenital HSV is extremely rare in the West, but is important because of very high neonatal mortality. It occurs in an estimated 1 in 200 000 pregnancies, accounting for 3 per cent of all neonatal herpes. Neonatal infection following viral shedding in labour occurs in up to 30/100 000 in parts of North America, but is often much lower in other countries. Transmission is almost 50 per cent in active primary infection, but nearer 1 per cent with active recurrent herpes because of passive fetal immunity. Most neonatal herpes occurs in women without a history.

Caesarean section is advised for intrapartum primary herpes. If vaginal delivery is unavoidable or the membranes have been ruptured for more than 4 h, then the infant should be treated with acyclovir. With recurrent active herpes the risks of caesarean section probably outweigh the benefits. Caesarean section is not indicated if lesions are absent at the onset of labour.

Cytomegalovirus (CMV)

CMV is the commonest congenital infection in developed countries. Seventy five per cent of pregnant women are immune; less in higher socioeconomic classes or in developing countries. Around 1 to 4 per cent of women acquire primary infection in pregnancy; this may be asymptomatic, but commonly produces an infectious mononucleosis-like illness.

Transplacental transmission follows 40 per cent of primary infections and less than 1 per cent of secondary recurrences. After primary infection, 5 to 15 per cent of neonates are symptomatic, and of these more than 80 per cent develop severe neurological sequelae including mental impairment and sensorineural hearing loss. Even asymptomatic infants have a 5 to 15 per cent risk of hearing impairment. The outcomes of CMV infection in pregnancy are shown in Fig. 2.

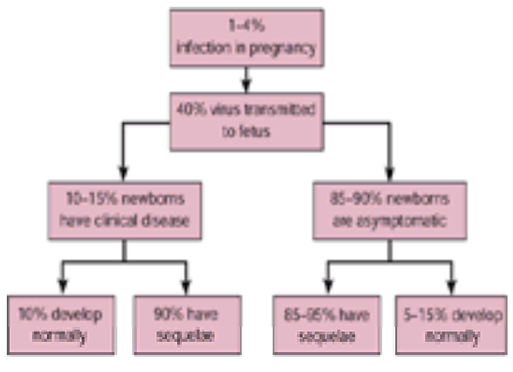

Fig. 2 The outcome from CMV infection in pregnancy.

In pregnant women, a diagnosis made by IgM or rising IgG titres can be confirmed by amniotic fluid PCR taken at least 6 weeks after maternal infection. High viral load and abnormal prenatal ultrasound findings such as intra-abdominal or cranial calcification, cerebral ventriculomegaly, and cardiomegaly, are associated with, but do not accurately predict, a poor long-term prognosis. Termination of pregnancy may be offered. There is no effective therapy. The chance of normal childhood development without evidence of fetal damage is approximately 75 per cent. Intravenous ganciclovir given to the infant reduces hearing loss in the most severely affected. CMV screening may be valuable in the future when orally available drugs are marketed.

Herpes zoster (VZV)

The varicella zoster virus incubation period is 10 to 21 days, infectivity being from 2 days before the rash appears to when all the vesicles are covered. Eighty five per cent of adults in the West are immune, infection only occurring in 1 in 2000 pregnancies. A much higher proportion of adults in developing countries are non-immune. The presence of maternal IgG indicates immunity, whilst IgM indicates primary infection. Invasive fetal testing is not often helpful.

Primary infection in pregnancy may be severe, with pneumonia in 10 per cent and occasional maternal death. Before 20 weeks, 2 per cent of fetuses develop the congenital varicella syndrome, characterized by neurological, optical, and limb anomalies. Ultrasound findings 5 weeks after infection include polyhydramnios and

echogenic foci in the fetal liver (Fig. 1).

Varicella zoster immunoglobulin (**VZIG**) should be given, preferably within 72 h of contact, to pregnant women who are non-immune as determined by serology. Women who develop varicella after 12 weeks gestation should be given acyclovir, intravenously if pneumonia develops. Specialist referral is advised with respiratory symptoms, dense lesions, or if new vesicles are still appearing after 6 days. Neonatal zoster may occur if maternal infection occurs 5 days before to 2 days after delivery; this is associated with up to a 30 per cent neonatal mortality. Babies born to mothers who develop perinatal chickenpox should be given intramuscular VZIG.

Hepatitis B

Ten per cent of hepatitis B infected adults become chronic carriers, compared to 90 per cent of infants following vertical transmission. Detectable hepatitis B surface antigen (HepBsAg positive) in blood indicates infectivity and is present from 4 to about 24 weeks after infection; thereafter its presence is associated with chronic infection. Individuals with hepatitis B surface antibodies (HepBsAb positive) are immunologically cured. Detectable hepatitis B e antigen (HepBeAg positive) indicates high infectivity.

In the West, less than1 per cent of pregnant women are HepBsAg positive, although the incidence is rising; in parts of Africa and Asia, the rate is 25 per cent. The risk of perinatal transmission relates to maternal viral antigen status. In HepBsAg positive/HepBeAg negative mothers the risk is 5 to 20 per cent, whereas in HepBsAg positive/HepBeAg positive it is 70 to 90 per cent. Targeted screening only identifies about half of chronic carriers, hence universal screening has been advocated in the West.

Vertical transmission can be reduced by more than 90 per cent by active immunization, with 0.5 ml hepatitis B vaccine, of all infants born to HepBsAg positive mothers, with additional passive immunisation (200 IU of hepatitis B immunoglobulin within 12 h of birth) for infants born to HepBeAg positive or HepBsAb negative mothers. Unfortunately, compliance with this regimen is often poor, and the WHO recommends universal vaccination in countries with high prevalence.

Hepatitis C

Approximately 3 per cent of pregnant women world-wide are infected with Hepatitis C. In inner London, the rate is 0.8 per cent, but it is 30 per cent in HIV positive women. Hepatitis C leads to chronic hepatitis in about 80 per cent. Progression is insidious and most pregnant women are asymptomatic. Liver transaminases may be normal, but if elevated, tend to reduce during pregnancy.

Vertical transmission of HCV occurs in approximately 6 per cent if HCV is detectable by PCR in the mother; otherwise the risk is negligible. Coexisting HIV infection increases the rate of vertical transmission to 23 per cent. Infected infants usually remain viraemic and prone to chronic hepatitis. Antibody levels are usually detectable within 3 months after infection. The ELISA has a sensitivity of 95 per cent, but should be supplemented by a recombinant immunoabsorbant assay (RIBA). PCR is used to confirm infection in infants.

Interferon-a reduces disease activity but is not recommended in pregnancy, although it may be used postpartum. Elective caesarean section, formula feeding, or administration of immune globulin do not reduce vertical transmission to the neonate.

Bacterial vaginosis

An overgrowth of anaerobic organisms such as *Gardnerella vaginalis* and *Mycoplasma hominis* causes this condition, which is characterized by excessive Gram-negative bacilli and cocco-bacillary organisms compared to lactobacilli on Gram staining. Bacterial vaginosis is not sexually transmitted, but is associated with sexually transmitted diseases and is rare before the onset of sexual activity. The prevalence varies from 5 to 20 per cent, depending much on the diligence with which the diagnosis is sought. Three of four Amsel's criteria are required for diagnosis: a thin white homogeneous discharge, clue cells, raised vaginal pH (>4.5), and a positive 'whiff test' (fishy odour when adding 10 per cent KOH to the discharge). At least 50 per cent of women with bacterial vaginosis have no symptoms, but an offensive, thin white discharge is often found. Its presence is strongly associated with a history of, and an increased risk for, late miscarriage and preterm birth. As prematurity is the major cause of neonatal mortality and morbidity in developed countries, the condition is extremely important. Bacterial vaginosis usually responds to metronidazole or clindamycin. Antibiotics reduce but do not eliminate the risk of prematurity. The role of screening for bacterial vaginosis in pregnancy is controversial.

Streptococci

Group A streptococci (*Streptococcus pyogenes*) remain an important cause of puerperal sepsis world-wide, but this is now rare in the West. Group B streptococci (*Streptococcus agalactiae*) usually cause less severe maternal disease. Overall, 12 to 26 per cent of pregnant women are colonized by Group B streptococci, particularly in the urine, and its presence is associated with preterm delivery. Ascending infection with chorioamnionitis and fetal infection may occur following rupture of the membranes. Intrauterine infection may cause stillbirth.

Early-onset neonatal streptococcal sepsis has a mortality of 6 per cent and occurs in 0.5 to 3.7 per 1000 live births. Risk factors include prematurity, prolonged rupture of the membranes, intrapartum maternal fever, heavy colonization, low maternal antibody levels, and a previously affected infant. Although 70 per cent of neonates born to carriers are colonized, only 1 to 2 per cent will develop disease.

Intrapartum penicillin greatly reduces early-onset neonatal disease. Preventative strategies are based on risk factors, either alone or in conjunction with screening. With the former, women are treated if they have a previous history, intrapartum fever, are in preterm labour, or where the membranes have been ruptured for more than18 h. By treating 18 per cent of women, 70 per cent of neonatal sepsis is prevented. In a combined screening and risk-based approach, third trimester vaginal and anal swabs are taken, and treatment of 27 per cent of all pregnant women can prevent 86 per cent of sepsis.

Listeria monocytogenes

Listeria monocytogenes is a Gram-positive bacillus that can cause serious disease in pregnant women, fetuses, and newborns. Infection is from salads contaminated with animal faeces, undercooked meats, unpasteurized milk, soft cheeses, and pates. In the United Kingdom, the incidence has been as high as 5 cases/100 000 live births, but world-wide the incidence has fallen as a result of public health campaigns about the likely source of infection.

Maternal disease manifests as bacteraemia, with fever, sore throat, headache, and chills. Diarrhoea, pyelitis, and back ache may occur. Symptoms generally recede without treatment, although a few mothers develop meningitis. Transplacental infection of the fetus is the most likely route of acquisition rather than ascending infection. Infection before 24 weeks gestation usually results in abortion; still birth or prematurity occurs after this gestation. The fetus may die, become macerated and then seed bacteraemia to the mother. Apparent meconium staining of liquor in preterm birth is highly suggestive of listeriosis. Two patterns of disease are seen, granulomatosis infantiseptica, in which the baby and the placenta are covered in miliary granulomata, and pneumonia without granulomata. Meningitis may coexist.

Listeria is resistant to all cephalosporins. The accepted treatment, based on animal model testing, is ampicillin with an aminoglycoside for synergy. There is little human data to support the view that ampicillin is better than penicillin G, but the usual practice is to change to ampicillin when listeriosis is confirmed.

Syphilis

The prevalence of *Treponema pallidum* infection in pregnancy is 0.2 per cent in London and 0.02 per cent in the rest of the United Kingdom, but in Africa, South East Asia, and Russia it is endemic. Pregnancy does not alter the clinical manifestations. Vertical transmission is predominantly transplacental, and occurs in up to 90 per cent of untreated women, particularly those with early disease. Most affected pregnancies result in congenital syphilis, miscarriage, preterm delivery, stillbirth, or neonatal death. Ultrasound appearances of the infected fetus may be normal or show hepatomegaly and other abnormalities. At birth, babies exhibit rhinitis, osteitis, and skin bullae. Hutchinson's triad of abnormal teeth, interstitial keratitis, and sensorineural deafness arise later in the untreated child.

Non-treponemal tests (e.g. VDRL) are usually employed for screening. Sensitivity is highest in secondary syphilis and lowest early in the infection, and false-positive

results occur with concomitant infections or autoimmune disease. The diagnosis should be confirmed with a specific treponemal test (e.g. FTA-ABS). Screening in pregnancy is cost effective, even where the disease is rare, and 121 women were identified by antenatal screening in the United Kingdom in 1994 to 97, meaning that 18 600 tests needed to be performed to detect one case.

Two intramuscular doses of benzyl penicillin, 2.4 MU 1 week apart, are given to treat syphilis in pregnancy. In true penicillin allergy, a 5 to 10-day regimen of high dose oral ceftriaxone is recommended. Treatment will prevent congenital infection in 98 per cent of cases. The rare Jarisch–Herxheimer reaction to treatment may precipitate preterm labour. VDRL titres should fall until undetectable or less than 1 in 4, otherwise retreatment is necessary.

Chlamydia trachomatis

The incubation period of *Chlamydia trachomatis* infection is 7 to 21 days. It is one of the commonest sexually transmitted diseases world-wide, and infects about 5 to 7 per cent of pregnant women. Endocervical swabs are best for screening, though urine testing may be easier. PCR has the highest sensitivity and specificity, followed by culture and then ELISA.

Chlamydia infection is mostly asymptomatic in pregnant women. Pelvic infection is very rare during pregnancy, but after delivery endometritis and salpingitis may lead to tubal damage and infertility, and 12 per cent of induced abortions are followed by pelvic infection. Maternal infection, particularly that which is recently acquired, is associated with prematurity, chorioamnionitis, and possibly stillbirth. Treatment reduces but does not eradicate these risks. Neonatal conjunctivitis occurs in up to 50 per cent exposed newborns and later-onset pneumonia in a smaller proportion.

Tetracyclines are contraindicated in pregnancy and erythromycin is not well tolerated. A single dose of azithromycin 1 g ensures compliance. Concerns about this drug in pregnancy remain unsubstantiated. Sexual contacts should be screened, and as reinfection rates are high, repeat testing is advised after at least 3 weeks to ensure a cure has been achieved. Screening or even prophylaxis of all mothers following abortion is cost effective.

Gonorrhoea

Neisseria gonorrhoea infection is endemic in many developing countries, but is declining in the West. Cervical culture will detect most infections, whereas PCR testing is expensive and does not enable antibiotic sensitivity testing. As with non-pregnant women, 80 per cent are asymptomatic in pregnancy. Pharyngeal and disseminated systemic infection with fever, rash, and septic arthritis are more common in pregnancy, but salpingitis is rare. Gonococcal cervicitis is associated with chorioamnionitis and a fourfold increase in prematurity. Ophthalmia neonatorum arises in 40 per cent if mothers are untreated. Gonococci have also been implicated in postpartum and postabortion endometritis and salpingitis. Penicillinase-producing strains are common, and treatment with 250 mg single dose intramuscular ceftriaxone, or 400 mg oral cefixime is recommended. Disseminated infection warrants intravenous therapy. The patient should be screened for other sexually transmitted diseases; indeed, antichlamydia therapy is often given at the same time. A test of cure should be taken at least 3 days after antibiotics.

Mycobacterium tuberculosis

In 1990, the WHO estimated that 90 million people developed tuberculosis and 30 million died from it. The highest rates of infection are in young adults, hence exposure during pregnancy must be common.

The pathogenesis and course are similar in pregnancy as in women who are not pregnant. The main additional concern is spread to the fetus, either through maternal disseminated disease and transplacental spread, or through direct extension of maternal genitourinary tuberculosis. Genital tuberculosis often has a long and indolent course with potential involvement of the fallopian tubes (90–100 per cent of women with genital infection), uterus (50–60 per cent), ovaries (20–30 per cent), and cervix (5–15 per cent). Genital infection is more likely to manifest as sterility and therefore is unusual as a cause of congenital infection.

Chemotherapy of tuberculosis during pregnancy provides the same excellent outlook as treatment in any other person. Providing proper treatment is received, there is no adverse effect of pregnancy, birth, the postpartum period, or lactation on the course of tuberculosis, and tuberculosis has little effect on the pregnancy outcome. Pregnancy is an opportunity to screen for tuberculosis. Women coinfected with HIV may have suppressed skin test reactivity and there is good argument for thoroughly investigating these women in pregnancy when they opportunistically present to health carers.

Congenital tuberculosis is very rare, but its incidence is increasing because of coinfection with HIV in pregnancy.

Toxoplasmosis

Toxoplasma gondii is a protozoan parasite acquired from eating uncooked meat or contaminated salad. The incubation period is 5 to 18 days. The prevalence of congenital toxoplasmosis is less than 0.1 per cent in most parts of the developed world.

Toxoplasmosis is frequently asymptomatic, but 10 to 20 per cent of mothers have lymphadenopathy or a flu-like episode. Microcephaly, hydrocephalus (Fig. 1 and Fig. 3), cerebral cystic lesions and calcification, and chorioretinitis are the severe sequelae in the fetus, leading to mental retardation and blindness. The risk of vertical transmission increases with gestation, but the severity of disease decreases, hence the highest risk for congenital toxoplasmosis with poor outcome is when maternal infection occurs between 10 to 24 weeks gestation. Congenital toxoplasmosis may rarely result from maternal infection up to 6 months prior to conception.

Fig. 3 MRI head of a 20-day-old baby with congenital toxoplasmosis, showing severe ventriculomegaly from hydrocephalus.

Diagnosis is difficult because serological tests have poor sensitivity, so much so that in proven cases no antibody may be detectable. Conversely, IgM may persist for months after infection and its identification should prompt a repeat test 3 weeks later. Screening is not advised in low prevalence countries. Confirmation of fetal infection is best achieved using PCR of amniotic fluid.

Treatment in pregnancy is with 1 g spiramycin three times per day until the diagnosis has been definitively made, but this is not of proven benefit in established congenital toxoplasmosis. Most authorities are now using pyrimethamine (50 mg/day), sulfadiazine (1 g three times/day), and folinic acid (50 mg/week) until delivery, but not starting this regimen until the second trimester onwards.

Cerebral ventriculomegaly or intracranial calcification on antenatal ultrasound examination may predict a worse outcome, and termination of pregnancy may be offered; yet normal development, including reversion of brain calcification and hydrocephalus, are possible following treatment of the neonate with pyrimethamine and sulfadiazine for 1 year.

Malaria

Millions of women world-wide are at risk of malaria during pregnancy each year. Parasitaemia is more frequent and the parasite load is higher in pregnancy, especially in primiparous women and during the second trimester; and cerebral malaria, pulmonary oedema, and renal failure occur more commonly. Severe malarial anaemia of pregnancy and placental infection cause spontaneous abortions, stillbirths, and intrauterine growth retardation. Malaria is one of the few preventable causes of low birth weight.

Congenital malaria from transplacental spread occurs in approximately 1 per cent of infected pregnancies. The newborns have fever, respiratory distress, pallor, anaemia, hepatomegaly, jaundice, and diarrhoea.

Controlled trials of antimalarial chemoprophylaxis during pregnancy demonstrate little risk and good benefit. Treatment of identified cases rather than chemoprophylaxis is not the best practice as most infected women are not diagnosed antenatally. In areas with sensitive plasmodium, chloroquine is recommended, but in many parts of the world mefloquine is preferable.

Trypanosomiasis

Trypanosoma cruzi (American trypanosomiasis; Chagas' disease) infection in pregnancy in Central and South America accounts for 1 to 10 per cent of all abortions. Infection occurs from insect bites or blood transfusions, with parasitaemia and fever occurring 2 to 3 weeks later. Congestive heart failure from myocarditis is more likely in pregnancy, and overall the disease has 10 to 20 per cent mortality; miscarriage, intrauterine growth retardation, or preterm delivery may also occur. Congenitally infected babies may have jaundice, anaemia, hepatosplenomegaly, encephalitis, pneumonitis, and hydrops fetalis from cardiac involvement. Diagnosis is through placental histology or blood smear examination for parasitaemia. There is no safe and reliable treatment in pregnancy.

T. brucei gambiense and *T. brucei rhodesiense* (African sleeping sickness) have similar adult mortality as American trypanosomiasis. Congenital trypanosomiasis is rarely described, but this is probably an under-reporting phenomenon.

A guide to the general safety of antimicrobial drugs in pregnancy is given in Table 1. The reader is advised to check current recommendations on every individual drug, for instance from the British National Formulary, before commencing any therapy in pregnancy. Table 2 gives brief notes on a few other infections in pregnancy.

Further reading

Brocklehurst P (2001). Interventions aimed at decreasing the risk of mother-to-child transmission of HIV infection (Cochrane Review). In: *The Cochrane Library*, Vol. 2. Update Software, Oxford.

Brown HL, Abernathy MP (1998). Cytomegalovirus infection. *Seminars in Perinatology*, **22**, 260–6.

Daffos F, Forestier F, Capella-Pavlovsky M, *et al*. (1988). Prenatal management of 746 pregnancies at risk for congenital toxoplasmosis. *New England Journal of Medicine* **31**, 271–5.

de March AP (1975). Tuberculosis and pregnancy: five to ten year review of 215 patients in their fertile age. *Chest* **68**, 800–5.

Grether JK, Nelson KB (1997). Maternal infection and cerebral palsy in infants of normal birth weight. *Journal of the American Medical Association* **278**, 207–11.

Hurtig A-K, Nicoll A, Carne C, *et al.* (1998). Syphilis in pregnant women and their children in the United Kingdom: results from national clinician reporting surveys 1994–7. *British Medical Journal* **317**, 1617–9.

Kenyon S, Boulvain M (2001). Antibiotics for preterm premature rupture of membranes (Cochrane Review). In: *The Cochrane Library*, Vol. 2. Update Software, Oxford.

King J, Flenady V (2001). Antibiotics for preterm labour with intact membranes (Cochrane Review). In: *The Cochrane Library*, Vol. 2. Update Software, Oxford.

Levy R, Weissman A, Blomberg G, Hagay Z (1997). Infection by parvovirus B19 during pregnancy: a review. *Obstetrics and Gynecology Survey* **55**, 254–9.

Liesnard C, Dooner C, Brancart F, Gosselin F, Delforge ML, Rodesch F (2000). Prenatal diagnosis of congenital cytomegalovirus infection: a prospective study of 237 pregnancies at risk. *Obstetrics and Gynecology* **95**, 881–8.

Marsden PD (1971). South American trypanosomiasis (Chagas' disease). *International Review of Tropical Medicine* **4**, 97–121.

McAuley J, Roizen N, Patel D, *et al*. (1994). Early and longitudinal evaluations of treated infants and children and untreated historical patients with congenital toxoplasmosis: the Chicago Collaborative Treatment Trial. *Clinical Infectious Diseases* **18**, 38–72.

Miller E, Fairley CK, Cohen BJ, Seng C (1998). Immediate and long term outcome of human parvovirus B19 infection in pregnancy. *British Journal of Obstetrics and Gynaecology* **105**, 174–8.

Mofenson LM and the Committee on Pediatric AIDS (2000). American Academy of Pediatrics. Technical report: perinatal human immunodeficiency virus testing and prevention of transmission. *Pediatrics* **106**, 1–12.

Murphy DJ, Sellers S, MacKenzie IZ, Yudkin PL, Johnson AM (1995). Case-control study of antenatal and intrapartum risk factors for cerebral palsy in very preterm singleton babies. *Lancet* **346**, 1449–54.

Ohto H, Terazawa S, Sasaki N, *et al*. (1994). Transmission of hepatitis C virus from mothers to infants. The vertical Transmission of Hepatitis C Collaborative Strudy Group. *New England Journal of Medicine* **330**, 744–50.

Remington JS, Klein JO, eds (2001). *Infectious diseases of the fetus and newborn infant*, 5th edn. WB Saunders, Philadelphia.

Schaefer G, Zervondakis IA, Fuchs FF, *et al*. (1975). Pregnancy and pulmonary tuberculosis. *Obstetrics and Gynecology* **46**, 706–15.

Smaill F (2001). Intrapartum antibiotics for Group B streptococcal colonisation (Cochrane Review). In: *The Cochrane Library*, Vol. 2. Update Software, Oxford.

Smith JR, Cowan FM, Munday P (1998). The management of herpes simplex infection in pregnancy. *British Journal of Obstetrics and Gynaecology* **105**, 255–60.

13.16 Blood disorders in pregnancy

E. A. Letsky

The physiological changes of pregnancy result in profound changes in the maternal haematological system. Plasma volume increases progressively, reaching a peak during the third trimester that is about 45 per cent or 1250 ml above non-pregnant values. Total red cell mass increases by about 20 to 30 per cent, resulting in haemodilution and hence a decline in haemoglobin concentration, packed cell volume, and red cell count. In the absence of iron deficiency the mean cell haemoglobin concentration remains at non-pregnant values and there is a slight increase (approximately 4 fl) in mean cell volume. As a result of these changes, anaemia cannot be diagnosed in pregnancy using criteria applied to non-pregnant individuals.

The changes in blood volume and haemodilution are so variable that the normal range of haemoglobin concentration can lie between 10.0 and 14.5 g/dl in healthy pregnancy at 30 weeks' gestation in women who have received parenteral iron. However, haemoglobin values of less than 10.5 g/dl in the second and third trimesters are probably abnormal and require further investigation. The World Health Organization (WHO) recommends that the haemoglobin concentration should not fall below 11.0 g/dl at any time during pregnancy.

Iron-deficiency anaemia

Effects of iron deficiency on the blood

The expansion of red cell mass represents the largest single demand for iron in pregnancy, amounting to a net gain of about 500 to 600 mg. In addition, 250 to 350 mg is needed for transfer to the fetus by active transport across the placenta, mainly in the last 4 weeks of pregnancy. Daily requirements for iron increase three- to four-fold and are met by an increased rate of absorption from the gut, together with mobilization of maternal iron stores. The mean serum iron concentration of healthy pregnant women is about two-thirds of the levels for non-pregnant individuals. Total iron-binding capacity is increased because transferrin levels more than double as pregnancy advances. In consequence, the saturation of iron-binding capacity is, in healthy pregnancy, lower (at about 25 per cent) than is normal for other situations. Serum ferritin (reflecting iron stores) declines during the first half of pregnancy to a nadir of about 15 to 20 µg/l, where it remains until delivery. Many women enter pregnancy with low or depleted iron stores even if the haemoglobin concentration is normal, and the majority of those who do not receive supplements have no stores at all at the end of pregnancy.

The anaemia of iron deficiency is the most common haematological problem in pregnancy. The earliest effect of iron deficiency on the erythrocyte outside pregnancy is a reduction in cell size, which appears to be the most sensitive index of underlying iron deficiency. This, however, is a very poor indicator of iron deficiency during pregnancy because of the established larger population of younger red cells. A fall in red cell haemoglobin concentration (mean corpuscular haemoglobin, mean corpuscular haemoglobin concentration) or the concentration of circulating haemoglobin is also a relatively late development in iron deficiency and is preceded by depletion of iron stores followed by a reduction in serum iron levels. A state of iron deficiency can be diagnosed before anaemia has developed by finding a serum ferritin of less than 30 µg/l and a serum iron of less than 10 µmol/l, with less than 15 per cent saturation of the total iron binding capacity. In recent years, serum ferritin, uninfluenced by recently ingested iron, has largely replaced serum iron and total iron-binding capacity as a non-invasive indicator of iron status.

Serum transferrin receptor (TfR) provides a new and allegedly reliable method for assessing cellular iron status. TfR is a transmembrane protein, present in all cells, that binds transferrin iron and transports it to the cell interior. Any reduction in iron supply results in an increase in TfR synthesis and, like serum ferritin, TfR has been shown to circulate in the plasma in small amounts which reflect the total body mass of TfR. As soon as cellular iron deficiency is established, the serum TfR rises in direct proportion to the degree of iron lack. This rise precedes the reduction in mean corpuscular volume and is therefore particularly valuable in identifying tissue iron deficiency in pregnancy. In combination with serum ferritin, serum TfR will give a complete picture of iron status, the serum ferritin reflecting iron stores (in the absence of chronic inflammatory disease) and the TfR reflecting tissue iron status. A bone marrow examination is now very rarely necessary to assess iron status during pregnancy.

Non-haematological effects of iron deficiency

Iron deficiency causes more than just anaemia and treatment with oral or parenteral iron results in improved well-being long before the haemoglobin rises significantly. The various effects of iron deficiency on cellular function may be responsible for the reported association between iron deficiency during pregnancy and preterm birth, and effects on neuromuscular transmission may underlie the anecdotal reports of increased blood loss at delivery in anaemic women.

Effects of iron deficiency on the fetus

The fetus derives its iron from maternal serum by active transport across the placenta in the last 4 weeks of pregnancy. The concentration of ferritin in cord blood is substantially higher than that in the mother's circulation at term and falls within the normal adult range, whether or not the mother is iron deficient. However, babies born to iron-deficient mothers have lower cord ferritin levels that those born to iron-replete mothers, and this has an important bearing on iron stores and development of anaemia in the first year of life when iron intake is very poor.

There have also been suggestions of behavioural abnormalities in children with iron deficiency, and iron deficiency in the absence of anaemia is associated with poor performance in the Bayley Mental Developmental Indices. Moreover, poor performance of 12- to 18-month-old iron-deficient, anaemic infants in mental and motor development can be improved to the level of iron-sufficient infants by treatment with ferrous sulphate.

Even more far-reaching effects of maternal iron deficiency during pregnancy have been suggested. A correlation has been shown between maternal iron-deficiency anaemia, high placental weight, and an increased ratio of placental weight to birthweight. This suggests that maternal iron deficiency results in poor fetal growth compared to that of the placenta. High blood pressure in adult life has been linked to lower birthweight and with those whose birthweight was lower than would be expected from the weight of the placenta. Prophylaxis of iron deficiency may therefore have important implications for the prevention of adult hypertension.

Giving iron in pregnancy

Because anaemia it is a late sign of iron deficiency, pregnant women are frequently given iron supplements prophylactically. The justification for doing this is still debated, although there is no question that iron deficiency is prevented. This is likely to be more important in poor than in developed countries, where controlled trials have demonstrated the effect of routine oral iron supplements on maternal well-being and perinatal outcome.

The main argument in favour of routine iron supplementation in pregnancy is that if iron deficiency is detected in the third trimester there may not be time to correct it by either oral or parenteral supplements. There is no haematological benefit in giving parenteral as opposed to oral iron and the maximal response to be expected in pregnancy is a rise in haemoglobin of approximately 0.8 g/l weekly. The main advantage of intravenous infusion is that it ensures adequate iron for response and avoids any undesirable side-effects associated with oral preparations, but parenteral iron is not without its risks. Otherwise, blood transfusions may be needed, with their associated hazards. On alance, a small daily supplement of oral iron given routinely, from the sixteenth week of pregnancy, seems the more sensible approach. It need not be considered as mandatory, but side-effects, which are most commonly bowel upsets, particularly constipation, are usually overcome by simple nutritional measures.

In iron-deficient women it may take more than a year after delivery for the haemoglobin to return to prepregnancy levels. By contrast, if iron supplements are given the haemoglobin is in the normal prepregnancy range 5 to 7 days after delivery, providing blood loss is not excessive.

Folic acid deficiency

Effects of folate deficiency on the blood and its treatment

When the diet is inadequate, pregnancy can lead to a state of negative folate balance. The pregnant woman needs approximately twice as much folic acid, 200 to 250 µg daily, as do non-pregnant individuals. The increase meets the needs of the growing uterus and conceptus and the expanded red cell mass. As pregnancy advances, serum folate falls to about half the non-pregnant value at term. The red cell folate content shows a slight decline over the same period.

Megaloblastic anaemia in pregnancy is usually the result of dietary folate deficiency. Its incidence is therefore variable, dependent on the socio-economic status of the population and whether or not folic acid supplements are given routinely as part of antenatal care. It occurs more frequently in multiple pregnancies, with about half of cases presenting in the third trimester and the remainder after delivery.

Deficiencies of iron and folate are often combined, with folate deficiency revealed by the failure of a patient to respond to iron supplements. The diagnosis can be difficult to make. The peripheral blood film may be unhelpful because the expected macrocytosis is masked by iron deficient microcytosis. Hypersegmentation of the neutrophils can also be seen in pure iron deficiency. Measurements of serum and red cell folate concentrations, as well as the excretion of formiminoglutamic acid after a histidine load (which is increased in normal pregnancy), can all be difficult to interpret because the results need to be related to the normal range that is expected for healthy pregnant women and not those derived from non-pregnant subjects. The diagnosis can only be confirmed by examination of a bone marrow aspirate.

There is a strong case for prophylaxis during pregnancy, particularly in countries where overt megaloblastic anaemia is frequent. Women with a poor diet should be given 300 µg of folate daily. The risk of adverse effects is very small because vitamin B_{12} deficiency in pregnancy is rare (see below): subacute combined degeneration of the cord has never been reported in these circumstances. Folic acid should be given with supplemental iron. If gastrointestinal megaloblastic changes are established, oral supplements will have no effect and absorption in general will be impaired. This situation can only be reversed by administration of intramuscular folic acid.

Non-haematological effects of folate deficiency

It is still argued to what extent folate deficiency may alter the outcome of pregnancy. Claims of an association between folate deficiency and placental abruption, abortion, and pre-eclampsia have not been substantiated. It has been shown that the incidence of prematurity and low birthweight can be reduced by giving supplements in poorly nourished populations, but has no effect in well-nourished women.

Effects of folate deficiency on the fetus

There is an increased risk of megaloblastic anaemia in the neonate of a folate-deficient mother, especially if delivery is preterm. There are also data suggesting an association between periconceptional folic acid deficiency, harelip, cleft palate, and, most important of all, neural tube defects. The latter has been confirmed by a multicentre, controlled trial of prepregnancy folate supplementation by the Medical Research Council.

In the United Kingdom, it is recommended that women contemplating pregnancy should take folate supplements of 400 µg daily. It has also been shown in Hungary, in a large, randomized trial, that periconceptional supplement of 800 µg of folic acid prevented the first occurrence of neural tube defects. The prevalence of harelip with or without cleft palate was not reduced by this supplementation.

Vitamin B_{12}

Maternal vitamin B_{12} stores, which are of the order of 3000 µg, are largely unaffected by pregnancy. The stores in the new-born infant are about 50 µg. The concentration of serum B_{12} falls during pregnancy from non-pregnant levels of 205 to 1025 µg/l to 20 to 512 µg/l at term, associated with preferential transfer of plasma B_{12} to the fetus. The recommended intake of vitamin B_{12} is 2.0 µg daily in the non-pregnant and 3.0 µg/day during pregnancy. This will be met by almost any diet that contains animal products, however deficient in other essential substances. Strict vegans, who will eat no animal produce whatsoever, can have a deficient B_{12} intake and should be given oral supplements during pregnancy. Recently, B_{12} deficiency associated with megaloblastic anaemia has been found in a proportion of women in Malawi whose main diet is maize and who eat little or no animal protein.

Vitamin B_{12} deficiency in pregnancy is very rare. Addisonian pernicious anaemia does not usually occur during the reproductive years, associated as it is with infertility. Pregnancy is only likely to occur after the deficiency has been corrected.

Haemoglobinopathies

It is important to recognize the genetic defects of haemoglobin structure and synthesis early in pregnancy, or preferably before conception, because:

- the clinical effects may complicate obstetric management and appropriate precautions can be taken
- it is now possible to offer prenatal diagnosis to those women carrying a fetus at risk of a serious defect of haemoglobin synthesis or structure at a time when termination of pregnancy is feasible (Fig. 1).

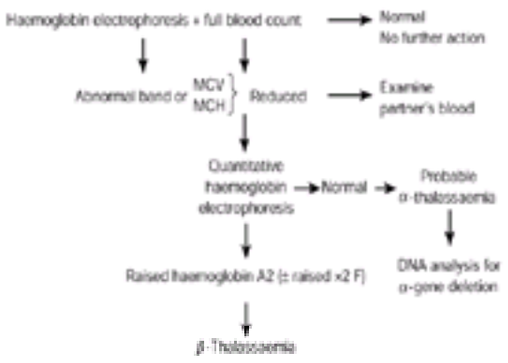

Fig. 1 Screening for haemoglobinopathies.

The two clinically significant groups of haemoglobinopathies, the sickle-cell syndromes and the thalassaemias, are described in detail in Section 22. Only special problems related to pregnancy will be discussed here.

Screening procedures vary from location to location and often only involve 'high-risk' populations. The most difficult situations arise in the United Kingdom, Europe, or the United States in antenatal obstetric units that care for a significant proportion of mothers from varying racial backgrounds. Figure 1 is a scheme that has been used with success in an obstetric unit serving a cosmopolitan population. This involves examination of red cell indices, haemoglobin electrophoresis, and, where indicated, quantitation of HbA_2 (and HbF) on every sample of blood taken at booking. If a haemoglobin variant or thalassaemic indices are found, the partner is requested to attend so that his blood can also be examined. By this means the chances of a serious haemoglobin defect can be assessed early in pregnancy, allowing counselling of the parents and the possibility of offering them prenatal diagnosis by fetal blood sampling or transabdominal chorion villus sampling, even though this will result in a late termination of pregnancy if indicated and desired. A recent audit of prenatal diagnosis, the first in 10 years in the United Kingdom, has shown that take-up is poor, particularly in Asian communities, and that at least six couples have sued their obstetrician for failure to inform them of the possibility of severe haemoglobinopathy in offspring.

Sickle-cell syndrome

It is essential to identify sickle-cell haemoglobin and the particular syndrome involved in the affected pregnant woman. The preferred procedure is to determine if there is an abnormal band on electrophoresis at booking (see above) and to perform a sickling test only on those cases where there is such a band, to confirm that this is sickle haemoglobin. The distinction between sickle-cell trait (HbA/S), sickle-cell anaemia (HbS/S), and HbS/C disease is then immediately apparent and nine out of ten unnecessary sickling tests are avoided in women of relevant racial groups.

Women with sickle-cell trait have no difficulties from overt sickling in pregnancy, but care is needed if a general anaesthetic is required during labour. There is an increased risk of pre-eclampsia, and tissue infarction can occur, even in the sickle-cell trait, if an adequate oxygen level is not maintained or if there is severe dehydration or shock.

Women with sickle-cell disease present special problems in pregnancy. Fetal loss is high, thought to be due to both impaired oxygen supply and sickling infarcts in the placental circulation. Abortion, severe pre-eclampsia, preterm labour, and other complications are more common than in women with normal haemoglobin. Although many women with sickle-cell disease have no complications, the outcome in any individual case is always in doubt.

Increasing numbers of obstetric units have adopted prophylactic transfusion regimes designed to maintain the proportion of HbA at 60 to 70 per cent of the total, but the benefit of such prophylaxis remains to be proven by a large multicentre trial with contemporary controls. A small, controlled trial in the United States suggested that the outcome is similar in women transfused prophylactically compared to those transfused only when indications arise. Prophylactic regimens are expensive and time consuming and may not be applicable in developing countries where the problem is widespread. In addition, there can be problems of alloimmunization to minor blood group red cell antigens: these have resulted in severe, delayed, and sometimes fatal haemolysis in the mother, and haemolytic disease of the new-born in some cases, not to mention the other hazards of blood transfusion.

Against this background of uncertainty, a reasonable compromise, particularly where blood transfusion facilities are limited, is to supervise pregnant women with sickle-cell anaemia with care throughout the pregnancy and to administer regular folate supplements. If they become severely anaemic or have crises during the second half of pregnancy, an exchange transfusion is appropriate. Alternatively, if they present with haemoglobin values of less than 7 g/dl it is also acceptable to transfuse them up to a level of 12 to 14 g/dl, since this will reduce the level of HbS to a safe value without the need for exchange transfusion. During labour, management is directed towards preventing dehydration and acidosis. There is some controversy concerning the safest form of anaesthetic during labour or to cover delivery, but if regional, rather than general, anaesthesia is used, precautions must be taken to prevent venous pooling in the lower limbs. In view of the non-functioning spleen, it is also sensible to give twice daily penicillin to protect against pneumoccocal infections even if pneumovax has been administered.

Special mention should be made of the problem of HbSC disease (see Section xx) in pregnancy. Many women with this disorder go through pregnancy with no complications, but occasionally there may be severe sickling episodes, either late in pregnancy or early in the puerperium, which may lead to maternal death due to massive infarctive crises, particularly in the lungs. Any woman with this disorder who develops a painful crisis late in pregnancy, or who develops symptoms and signs suggestive of a chest infection or small pulmonary embolus, requires urgent exchange transfusion. Examination of the stained blood film in the so-called chest syndrome will show large numbers of nucleated red cells. Cerebral sickling infarcts are not unusual and if not fatal may result in long-term morbidity.

As regards the fetus or neonate, most centres have developed programmes for antenatal diagnosis of sickle cell anaemia using DNA analysis of amniotic or chorionic cells, or of fetal blood. If these are not available, it is important to identify homozygous infants by agar gel electrophoresis or isoelectric focusing immediately after birth. The first 2 years of life are particularly hazardous for an infant with sickle-cell anaemia. Death due to infection and splenic sequestration is common, hence mothers must be advised to present the infants early with any unusual symptoms.

Thalassaemia syndromes

The clinical and haematological manifestations of the a- and b-thalassaemias are described in Section 22. Only certain points relevant to pregnancy will be discussed here.

a-Thalassaemia

The homozygous state for a-thalassaemia produces Bart's haemoglobin hydrops syndrome. Pregnancy with an a-thalassaemia hydrops is associated with severe, sometimes life-threatening, hypertension and proteinuria, the so-called mirror syndrome (of severe rhesus haemolytic disease). Vaginal deliveries are associated with obstetric complications resulting from the large fetus and bulky placenta and the small stature of the mother (usually of Far-Eastern origin).

If routine screening of the parents shows that the mother is at risk of carrying such a child, then she should be referred as early as possible for prenatal diagnosis so that termination of an affected fetus can be carried out before these severe obstetric problems occur. Although this is not yet a common problem in the United Kingdom, it may well become more frequent if there is an influx of immigrants from Hong Kong and the Far East, as has already occurred in the United States and Australia. Transfusion *in utero*, which has been successful in only a handful of cases, is not recommended. The fetus is usually not viable and may have a variety of associated physical defects.

b-Thalassaemia

Pregnancy is extremely rare in transfusion-dependent homozygous b-thalassaemics but is now being seen with increasing frequency in women with b-thalassaemia

intermedia. These patients may become profoundly anaemic and require regular transfusions during pregnancy.

Perhaps the most common problem associated with haemoglobinopathies and pregnancy is the anaemia developing in the antenatal period in women who have thalassaemia minor, heterozygous b-thalassaemia. They can be identified by examination of the blood sample taken at booking (see Fig. 1). The level of haemoglobin in early pregnancy may be normal or slightly below the normal range. Many women with thalassaemia minor enter pregnancy with depleted iron stores, hence they require the usual oral iron supplements in the antenatal period. Oral iron for a limited period will not result in significant iron loading, even in the presence of replete iron stores, but parenteral iron should never be given. A serum ferritin estimation can be carried out early in pregnancy, and if iron stores are found to be high then iron supplements can be withheld.

Folic acid, 5.0 mg daily, is recommended to cover the requirements of ineffective erythropoiesis. Blood transfusion may be indicated if the anaemia does not respond to oral iron and folate, the latter given parenterally as well as orally.

All women with mixed racial background should be screened for b-thalassaemia early in pregnancy. The partners of those who are found to be carriers should also be screened so that prenatal diagnosis can be offered where there is a risk of conceiving a homozygous child.

Miscellaneous anaemias specific to pregnancy

Many systemic medical conditions may further complicate the physiological haemodilution of pregnancy: these are discussed elsewhere.

A rare form of haemolytic anaemia appears to be specific to pregnancy. It remits after delivery but recurs in about half of affected women in later pregnancies. Although no autoantibody has been identified, the condition responds to corticosteroids or to infusion of human immunoglobulin (see below). The infant may be affected in about 20 per cent of cases.

There is a rare form of aplastic anaemia that occurs for the first time in pregnancy, remits after delivery, and then recurs again in subsequent pregnancies. The cause is unknown. It does not respond to any form of bone marrow stimulant or corticosteroid therapy and the management is symptomatic.

Disorders of haemostasis

Normal pregnancy is accompanied by major changes in the coagulation and fibrinolytic systems. There are significant increases in the procoagulant factors V, VIII, and X, and a very marked increase in plasma fibrinogen. In uncomplicated pregnancy, there is no change in antithrombin concentrations during the antenatal period, a fall during delivery, and then an increase 1 week postpartum. Protein C levels appear to remain constant or increase slightly, but protein S activity falls significantly during normal pregnancy.

The result of these physiological changes is to alter the usual balance between the procoagulants and anticoagulants in favour of the factors promoting blood clotting. In addition, fibrinolytic activity appears to be reduced during healthy pregnancy but returns to normal rapidly after separation of the placenta and completion of the third stage of delivery. This effect is mediated by placentally derived plasminogen activator inhibitor type II.

These changes in the haemostatic systems, together with the increase in blood volume, help to reduce the chances of abnormal haemorrhage at delivery, but they also convert pregnancy into a hypercoagulable state that may carry special hazards for both the mother and fetus. These hazards include a spectrum of haemostatic disorders, from thromboembolism through to the many conditions associated with disseminated intravascular coagulation.

Disseminated intravascular coagulation

This is associated with a wide variety of complications of pregnancy. It may be well compensated, with little change in tests of haemostatic function and no bleeding, as seen in prolonged retention of a dead fetus and mild pre-eclampsia, or it may result in intractable haemorrhage with gross consumption of coagulation factors and platelets and raised levels of fibrin degradation products, as seen classically in abruptio placentae (see Table 1).

Other complications of pregnancy in which disseminated intravascular coagulation may take a part include amniotic fluid embolism; septic abortion and intrauterine infection; hydatidiform mole; placenta accreta; pre-eclampsia and eclampsia, and prolonged shock from any cause (see Fig. 2).

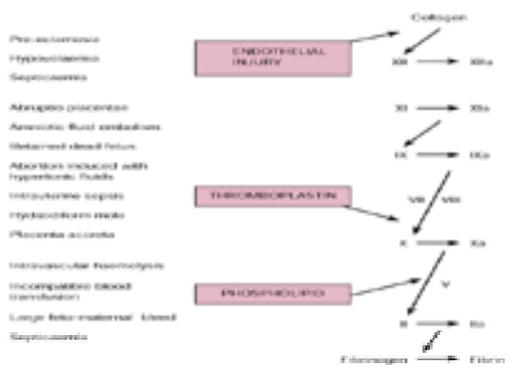

Fig. 2 Trigger mechanisms of disseminated intravascular coagulation during pregnancy. Interactions occur in many of these obstetric complications.

Disseminated intravascular coagulation in pregnancy is always a secondary process. Useful and rapid screening tests include the platelet count, partial thromboplastin time (intrinsic coagulation), prothrombin time (extrinsic coagulation), thrombin time, fibrinogen concentration, and levels of fibrin degradation products. Aside from standard supportive care (see Section 22) and removal of the triggering mechanism where this is known, management of the obstetric patient with disseminated intravascular coagulation should also seek to achieve an empty and contracted uterus.

Pre-eclampsia and haemostatic changes

Pre-eclampsia may be associated with marked changes in the normal physiological response of the haemostatic mechanisms during pregnancy (Table 2). The combination of a reduced platelet lifespan and a fall in the platelet count without platelet-associated antibodies indicates a low-grade coagulopathy. Once the disease process is established the most relevant haemostatic abnormalities appear to be the platelet count, factor VIII coagulation activity, and serum fibrin degradation products. Those women with the most marked abnormalities in these parameters suffer the greatest perinatal loss. It has been shown that if the platelet count is above 100×10^9/l there are no disturbances in haemostatic function and therefore no need for coagulation screening tests.

Rarely, in very severe pre-eclampsia, the patient develops microangiopathic haemolysis. This causes profound thrombocytopenia and leads to confusion in differential diagnosis between pre-eclampsia, HELLP syndrome, haemolytic uraemic syndrome (HUS), and thrombotic thromboctyopenia (TTP). The recent identification of deficient metalloproteinase in TTP, either due to an IgG antibody in the acute form or to a genetic defect in the chronic relapsing form, will help to distinguish this condition from severe pre-eclampsia, HELLP syndrome, and HUS, which is essential if appropriate management with fresh frozen plasma is to be applied. In relation to pregnancy, haemolytic uraemic syndrome usually presents in the postpartum period with renal failure.

Platelet disorders

Thrombocytopenia is a common haematological abnormality in pregnancy and can have important implications for both mother and fetus. It may occur as part of the

pathophysiology of pregnancy itself, or pregnancy may be superimposed on a background of haematological disease.

Incidental thrombocytopenia

As pregnancy advances there is a progressive small, but significant, fall in the platelet count in individual patients, probably due to haemodilution. Approximately 8 per cent of healthy pregnant women have thrombocytopenia at term, with platelet counts between 90 and 150×10^9/l. These women have no history of pre-eclampsia or immune thrombocytopenia and there is no increased incidence of thrombocytopenia in their offspring.

Thrombocytopenia and disseminated intravascular coagulation

Low-grade disseminated intravascular coagulation, as observed in pre-eclampsia, may be associated with further decrements but the platelet count rarely falls below 50×10^9/l, even in acute defibrination syndromes. Clearly, thrombocytopenia and platelet consumption represent only one aspect of this condition (see above) and will be corrected quickly when haemostatic mechanisms return to normal, usually without the use of, or need for, platelet transfusion.

Idiopathic thrombocytopenic purpura

Idiopathic thrombocytopenic purpura (ITP) is a rare condition, but relatively common in women of reproductive years. Patients in remission may still have elevated levels of platelet-associated IgG (PAIgG), especially following splenectomy. This is important in pregnancy because of the possibility of placental transfer of antibody resulting in fetal thrombocytopenia. Measurement of maternal platelet count, serum platelet antibody, and PAIgG are useful diagnostic tools but are not predictive of neonatal thrombocytopenia.

In the past, analysis of the literature gave an overall incidence of neonatal thrombocytopenia of 52 per cent with significant morbidity in 12 per cent, but we know now that this incidence was distorted because only symptomatic women were likely to have been investigated and reported. More recent analyses show an incidence of fetal thrombocytopenia of around 10 per cent, with severe thrombocytopenia (platelets fewer than 50×10^9/l) in less than 5 per cent overall. Fetal and neonatal morbidity and mortality is negligible even in the face of severe thrombocytopenia. Thrombocytopenia in the neonate tends to become more severe in the first few days of life and measures can be taken to correct this at birth before the nadir is reached, if indicated.

Pregnant women with ITP nearly always have the chronic form of the disease. The main clinical difficulty is that effects of treatment have to be considered in relation to the progress of pregnancy in both mother and fetus. The mild condition may require no treatment, but if the platelet count falls below 50×10^9/l, or there is clinical evidence of bleeding, then prednisolone is required (60 mg daily, reducing rapidly to the lowest possible dose that maintains the platelet count above 50×10^9/l). The prevalence of pre-eclampsia, gestational diabetes, postpartum psychosis, and osteoporosis are all increased by corticosteroids, hence the dose and duration of treatment should be the minimum needed to reduce the risk of bleeding or to raise the platelet count of any asymptomatic woman at term, allowing her to have epidural or spinal analgesia if desired or indicated.

The introduction of treatment by intravenous monomeric polyvalent human IgG has altered the management options in pregnancy dramatically. Used in the original recommended dose of 0.4 g/kg for 5 days by intravenous infusion, a persistent and predictable response is obtained in approximately 80 per cent of cases. There is no doubt about the value of this treatment in selected cases of severe symptomatic thrombocytopenia, but it cannot be advocated indiscriminately in view of its high cost and unproven benefit in milder cases. Analysis of recent reports indicates that the postulated beneficial transplacental effect on fetal platelets is unreliable and that exogenous IgG may not cross the placenta.

Splenectomy is now hardly ever indicated in the pregnant patient with ITP, but remains an option if all other attempts to increase the platelet count to safe levels fail. It is best carried out in the second trimester because surgery is best tolerated then and the size of the uterus does not make the operation technically difficult.

Mode of delivery

The most contentious problem in pregnancy associated with maternal ITP is the mode of delivery of the fetus. Even if the mother has to deliver in the face of a low platelet count, she is unlikely to bleed from the placental site once the uterus is empty, but she is at risk of bleeding from surgical incisions, soft-tissue injuries, or tears. Platelets should be available for transfusion but not given prophylactically.

The major risk at delivery is to the thrombocytopenic fetus, which as a result of birth trauma may suffer intracranial haemorrhage. Maternal platelet count, maternal platelet-associated IgG, splenectomy status, and history may give a crude indication of the likelihood of fetal thrombocytopenia but cannot be used in an individual case to predict the fetal platelet count. It is, however, unlikely for the fetus to have severe thrombocytopenia if the mother has no history of ITP before the index pregnancy and has no detectable IgG platelet antibody.

Platelet counts in blood obtained by transcervical fetal scalp sampling prior to, or early in, labour have been used to make a decision about the mode of delivery. However, this mode of sampling is not without risk of significant haemorrhage in the truly thromboctyopenic fetus, often gives false positive results, and demands urgent action to be taken on the results obtained. In addition, by the time that results are available the fetus may have already descended so far in the birth canal that caesarean section is technically difficult and traumatic for the fetus. The only way a reliable platelet count can be obtained is by a percutaneous transabdominal fetal cord blood sample, but this is not widely available and not without its difficulties, including a risk of approximately 1 per cent for the fetus. Decisions concerning the mode of delivery often have to be taken without knowledge of the fetal platelet count.

There is no good evidence that caesarean section is less traumatic than uncomplicated vaginal delivery, although this mode of delivery allows more overall control and there are usually no unpredictable complications. At the time of writing the emphasis of management is to return to a non-interventional policy of sensible monitoring, supportive therapy, and a mode delivery determined mainly by obstetric indications and not primarily by either the maternal or fetal platelet count.

Alloimmune thrombocytopenia

Fetal and neonatal alloimmune thrombocytopenia develops as a result of maternal sensitization to paternally derived fetal platelet antigen, the pathogenesis being analogous to that of Rh haemolytic disease of the new-born. The mother is not thromboctyopenic but the fetus can have a very low platelet count and is at risk of spontaneous intrauterine intracranial haemorrhage. This results from a specific antibody interfering with glycoprotein-binding sites and profoundly altering platelet function, particularly, aggregation. The most common platelet antigen involved is HPA-1, but a platelet incompatibility does not invariably result in alloimmunization. The maternal immune response appears to be restricted to those women with HLA-B8 and HLA-DR3 antigens. Thus, although 1 in 50 pregnancies are incompatible with respect to HPA-1 antigens (98 per cent prevalence of HPA-1a in the United Kingdom), only 1 in 5000 births are affected.

The children of first pregnancies (unlike rhesus disease) are often affected and the disease process can begin in early fetal life. Management is aimed at identifying the fetus at risk and correcting the thrombocytopenia *in utero*. Screening of women for HPA-1a status is not established. Investigation of neonatal intracranial haemorrhage or unexplained intrauterine death should include screening of parental blood for platelet antigens and maternal platelet antibodies. The approaches to the management of this problem are all controversial. One protocol involves fetal blood sampling at 20 to 22 weeks gestation and treating the mothers with thrombocytopenic fetuses with intravenous IgG 1 g/kg/week, with or without steroids, until delivery. This has been reported as successful in some units but not others. The overall results of multicentre trials of the efficacy of maternal intravenous IgG administration are variable, with more successful reports from North America and general scepticism from European centres. Another approach is to administer weekly compatible platelet transfusions to the fetus. This has been successful in a number of cases but involves frequent hazardous procedures. Whatever approach is used, immediate predelivery administration of compatible platelets to the fetus is recommended. The accepted management when the diagnosis is established shortly after birth is to transfuse specially prepared HPA-1a negative platelets from preselected blood-bank donors or, if facilities are available, washed platelets from the mother.

Inherited defects of haemostasis

Von Willebrand's disease

Von Willebrand's disease is the most frequent of all inherited haemostatic abnormalities and is therefore the most likely coagulopathy to affect women in pregnancy.

In normal pregnancy, a rise in both VIIIC and von Willebrand factor is observed. Patients with all but the severest forms of von Willebrand's disease show a similar but variable rise in both these factors during pregnancy, although there may not be a reduction in the bleeding time. After delivery, normal women maintain an elevated level of VIIIC for at least 5 days. In women with von Willebrand's disease the duration of this elevation seems to be related to the severity of the disorder. The general consensus is that the most important determinant for abnormal haemorrhage at delivery is a low factor VIIIC level. Appropriate factor VIII concentrates, containing von Willebrand factor activity as well as factor VIIIC, should be standing by to cover delivery but are rarely required to achieve haemostasis. Cryoprecipitate, with all the hazards of a fresh plasma product, is no longer recommended. There is virtually no place for desmopressin in obstetric practice, except perhaps in the puerperium. Any rise in factor VIII attributable to desmopressin will have been achieved under the influence of pregnancy itself. By contrast, desmopressin has a valuable place in women undergoing gynaecological or other surgery.

Haemophilia

The risks in pregnancy for the female carrier of haemophilia are two-fold:

1. She may, due to lyonization (random deletion of the X chromosome), have very low VIII or IX levels and be at risk of excessive bleeding, particularly following a surgical or traumatic delivery.
2. Fifty per cent of her male offspring will inherit haemophilia. This has important implications now that prenatal diagnosis of these conditions is possible.

It is important to identify carriers prior to pregnancy, not only to provide appropriate management for the rare case with pathologically low coagulation activity but also to provide genetic counselling. Changes in factor VIII complex may make the identification of carriers more difficult during pregnancy. Clinical problems occur more often in carriers for Christmas disease, since factor IX does not rise in response to healthy pregnancy in the same way as does factor VIII. Appropriate heat-treated concentrates are available to treat abnormal haemorrhage: cryoprecipitate and fresh frozen plasma should never be used unless heat-treated concentrates are not available.

It is possible to make an accurate prenatal diagnosis of haemostatic disorders. A fetal blood sample suitable for coagulation factor assays can be obtained from 18 weeks' gestation onwards and used to diagnose or exclude deficiencies of factors VIII and IX in male fetuses. In many cases at risk it is now possible to make rapid, early diagnosis of these conditions by DNA analysis of chorion villus samples or amniotic fluid cells obtained earlier in pregnancy. However, some few families remain where the recombinant DNA technology is not informative, so that fetal blood sample coagulation factor assays will continue to be of value.

Many women refuse prenatal diagnosis. If the fetus is male, scalp electrodes and sampling should be avoided during labour. Cord blood should be taken to establish the diagnosis to avoid haemostatic stress in the neonate.

Factor XI Deficiency (plasma thromboplastin antecedent deficiency)

This is a rare coagulation disorder, less common than the haemophilias. It has an autosomal recessive inheritance and therefore both men and women may be affected. Usually only homozygotes have clinical evidence of a coagulation disorder. Spontaneous haemorrhages and haemarthroses are rare but problems arise from the profuse bleeding which may follow major trauma or surgery if no prophylactic factor XI concentrate is given. The diagnosis is made by finding a prolonged partial thromboplastin time with a low factor XI level in an assay system in which all other coagulation tests are normal. Occasionally women with postpartum haemorrhage are found to have this abnormality. Fortunately the condition rarely causes problems either during pregnancy or delivery or in the newborn child. Prolonged bleeding at ritual circumcision is unusual. There is no justification in screening routinely for this condition either in the mother, fetus, or neonate. Women with factor XI deficiency should be given documentation of their defect so that appropriate measures can be taken to cover surgery or accidental trauma.

Genetic vascular disease

Ehlers Danlos Syndrome (EDS)

It is often forgotten that an essential part of the haemostatic system is healthy vasculature. The Ehlers Danlos syndrome may be associated with bleeding because of increased fragility of vessels due to defects in collagen synthesis. The disease has an autosomal dominant inheritance and has been subdivided into 10 subtypes of which type IV is the most severe and may have lethal complications, the most important of which is rupture of the long arteries. EDS IV is associated with an abnormality of collagen type III as a result of mutations in the corresponding gene COL3A1.

Surgical procedures should be avoided unless essential because the tissues are friable and massive bleeding may occur and healing of incisions may be delayed. Pregnancy and delivery will be involved with obvious potential hazards.

The diagnosis of this condition may be missed, especially in an obstetric gynaecological scenario, where many women complain of easy bruising, which is one of the main presenting symptoms, but platelet function and coagulation screening tests will yield normal results. Although there is no effective treatment or prophylaxis for this potentially lethal condition, appropriate management and precautions at least can be instituted, combined with sensitive genetic counselling regarding the autosomal dominant inheritance of this disorder.

Further reading

Anaemias and related disorders

Alberman E, Noble JM (1999). Commentary: Food should be fortified with folic acid. *British Medical Journal* **319**, 93.

Bothwell TH (2000). Iron requirements in pregnancy and strategies to meet them. *American Journal of Clinical Nutrition* **72**, 257S–64S.

Carriaga MT, Skikne BS, *et al*. (1991). Serum transferrin receptor for the detection of iron deficiency in pregnancy. *American Journal of Clinical Nutrition* **54**, 1077–81.

Czeizel AE and Dudas I (1992). Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. *New England Journal of Medicine* **327**, 1832–5.

Gill PS, Modell B (1998). Thalassaemia in Britain: a tale of two communities. *British Medical Journal* **317**, 761–2.

Goodall HB, Ho Yen DO, *et al*. (1979). Haemolytic anaemia of pregnancy. *Scandinavian Journal of Haematology* **22**, 185–91.

Howard RJ, Tuck SM, *et al*. (1995). Pregnancy in sickle cell disease in the UK: results of a multicentre survey of the effect of prophylactic blood transfusion on maternal and fetal outcome. *British Journal of Obstetrics and Gynaecology* **102**, 947–51.

Kadir RA, Sabin C, *et al*. (1999). Neural tube defects and periconceptional folic acid in England and Wales: retrospective study. *British Medical Journal* **319**, 92–3.

Koshy M and Burd L (1991). Management of pregnancy in sickle cell syndromes. *Hematology/Oncology Clinics of North America* **5**, 585–96.

Larrabee KD, Monga M (1997). Women with sickle cell trait are at increased risk for preeclampsia. *American Journal of Obstetrics and Gynecology* **177**, 425–8.

Letsky EA (1998). The haematological system. In: Broughton Pipkin F and Chamberlain GVP, eds. *Clinical physiology in obstetrics*, pp. 71–110. Blackwell Science, Oxford.

Letsky EA (2001). Maternal anaemia in pregnancy. Iron and pregnancy – a haematologist's view. *Fetal and Maternal Medicine Review* **12**, 159–75.

Medical Research Council (MRC) (1991). Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. MRC Vitamin Study Research Group. *Lancet* **338,** 131–7.

Modell B, Petrou M, *et al*. (1997). Audit of prenatal diagnosis for haemoglobin disorders in the United Kingdom: the first 20 years. *British Medical Journal* **315,** 779–84.

Perry KG Jr and Morrison JC (1990). The diagnosis and management of hemoglobinopathies during pregnancy. *Seminars in Perinatology* **14,** 90–102.

Rushton DH, Dover R, *et al*. (2001). Why should women have lower reference limits for haemoglobin and ferritin concentrations than men? *British Medical Journal* **322**, 1355–57.

Snyder TE, Lee LP, *et al*. (1991). Pregnancy-associated hypoplastic anemia: a review. *Obstetrical and Gynecological Survey* **46,** 264–9.

Van den Broek NR, Letsky EA, *et al*. (1998). Iron status in pregnant women: which measurements are valid? *British Journal of Haematology* **103,** 817–24.

Wald N J, Bower C (1995). Folic acid and the prevention of neural tube defects. *British Medical Journal* **310,** 1019–20.

Walter T (1994). Effect of iron-deficiency anaemia on cognitive skills in infancy and childhood. *Baillieres Clinical Haematology* **7,** 815–27.

Haemostasis

Burrows RF, Kelton JG (1995). Perinatal thrombocytopenia. *Clinics in Perinatology* **22,** 779–801.

David A, Letsky EA, *et al.* (In press). Factor XI deficiency presenting in pregnancy: diagnosis and management. *British Journal of Obstetrics and Gynaecology*.

Forbes CD, Greer IA (1992). Physiology of haemostasis and the effect of pregnancy. In: Greer IA, Turpie AGG, Forbes CD, eds. *Haemostasis and thrombosis in obstetrics and gynaecology*, pp. 1–25. Chapman and Hall, London.

Furlan M, Robles R, *et al*. (1998). Von Willebrand factor-cleaving protease in thrombotic thrombocytopenic purpura and the hemolytic-uremic syndrome. *New England Journal of Medicine* **339,** 1578–84.

Hamel BCJ, Pals G, *et al.* (1998). Ehlers-Danlos syndrome and type III collagen abnormalities: a variable clinical spectrum. *Clinical Genetics* **53**, 440–46.

Kadir RA. (1999). Women and inherited bleeding disorders: Pregnancy and delivery. *Seminars in Hematology* **36**, 28–35.

Leduc L, Wheeler JM, *et al*. (1992). Coagulation profile in severe preeclampsia. *Obstetrics and Gynecology* **79,** 14–8.

Letsky EA (in press). Coagulation defects. In: de Swiet M, ed. *Medical disorders in obstetric practice*, 4th edn. Blackwell Science, Oxford.

Letsky EA, Greaves M (1996). Guidelines on the investigation and management of thrombocytopenia in pregnancy and neonatal alloimmune thrombocytopenia. *British Journal of Haematology* **95,** 21–6.

Van den Broek N, Letsky EA. (In press). Pregnancy and the Erythrocyte Sedimentation Rate. *British Journal of Obstetrics and Gynaecology*.

Weatherall DJ, Letsky EA (2000). Genetic haematological disorders. In: Wald NJ, Leck I eds. *Antenatal and neonatal screening* 2nd edn, Oxford University Press, Oxford pp. 243–81.

13.17 Malignant disease in pregnancy

Robin A. F. Crawford

Introduction

Cancer is rare during pregnancy, quoted as occurring in only about 1 per 1000 live births. Most malignancies affecting this age group have been seen during pregnancy. Tumours of the uterine cervix, ovary, breast, or thyroid can metastasize to the placenta but not the fetus. Gestational trophoblastic disease arises from fetal chorion and is a malignant transformation of the placenta. Melanoma and haematological tumours, which also can invade the placenta, may cross into the fetal circulation. Pregnancy may cause enlargement of a pituitary tumour and a previously silent tumour may present with symptoms in pregnancy. Rare cases of colonic and neurological cancers developing in pregnancy have been reported in the literature. The diagnosis and treatment of the cancer may have been made prior to the pregnancy: discussion of teratogenesis and effects of treatment are also included in this chapter.

The placenta has several crucial functions. It is a fetal respiratory organ, a sophisticated endocrine unit, and a membrane that allows preferential and selective transfer of substrates from the mother to the fetus for fetal growth and development. In addition, the placenta is largely an effective barrier between the mother and the fetus. Transfer of fetal cells into the maternal circulation is common and probably occurs throughout gestation in all pregnancies. The transfer of maternal cells to the fetus, by contrast, is a relatively rare event.

Concurrence of pregnancy and cancer does raise complex therapeutic and ethical dilemmas, because the most appropriate and timely treatment for the mother may not be in the best interests of the fetus. Extra-abdominal surgery and anaesthesia during pregnancy rarely carry any risks to the fetus, and intra-abdominal surgery may be safely carried out in the second trimester. However, fetal cells divide and differentiate rapidly during the first trimester, and radiation and chemotherapy carry well-recognized risks to the fetus, including the risk of abortion, congenital abnormalities, or preterm birth. As a result, physicians may be reluctant to treat the mother aggressively at the time of initial diagnosis. Instead, in many cases they defer treatment for several weeks or months until the fetal lungs have matured. This delay, however, may substantially reduce the mother's chance of surviving the disease.

It is impossible to establish a threshold dose of ionizing radiation below which such treatment is safe for the fetus, inasmuch as exposure during the first trimester to a dose as low as 10 cGy appears to increase the risk of fetal abnormalities and exposure to 3 to 5 cGy increases the risk of childhood cancers. The risk is negligible if exposure to the fetus is less than 1 cGy. The dose of radiation, the gestational age of the fetus, and the practicability of shielding the fetus from radiation must be balanced against potential benefits to the mother.

Chemotherapy administered to the mother during the first trimester carries well-recognized risks including abortion or congenital abnormalities. Drugs that preferentially interfere with rapidly growing tissues, such as methotrexate, can harm the fetus. Use of antagonists of folate, purine, or pyrimidine synthesis during organogenesis results in congenital malformations in up to 25 per cent of fetuses, although this figure is much lower if the mother only receives therapy with a single agent. Treatment after the first trimester, when structural development is largely complete, is reasonably safe in many diseases and more appropriate than postponement of treatment. Chemotherapy after the first trimester has been associated with slight increases in the incidence of preterm birth and fetal growth retardation and, when administered shortly before delivery, with transient neonatal myelosuppression. Nevertheless, the long-term outcomes of the children of women who received chemotherapy during the second or third trimester are generally good.

In practical terms, acute leukaemia is virtually the only condition requiring immediate chemotherapy in a pregnant woman. When faced with cancer in the first trimester, the available information should be explained to the woman and her partner, who should give informed unhurried consent before treatment starts. In the majority of these situations, a consensus decision is reached between the woman, her partner, and the responsible physician to proceed with chemotherapy. However, the ethical issues are very complex and decisions have to be made on an individual basis.

As there is increasing success with the treatment of childhood cancers, more women will enter the reproductive age group having survived cancer treatment. There appears to be no overall increased risk of either congenital malformations or childhood cancers in the offspring of cancer survivors based on series of several thousand children. An increased incidence of spontaneous abortions, low-birth-weight babies, and neonatal deaths has been described for women with Wilms' tumour who had received at least 20 Gy abdominal radiation. Survivors of Hodgkin's disease who had received both radiation and chemotherapy (but not either alone) also appear to be at increased risk of spontaneous abortions.

Gestational trophoblastic disease

Gestational trophoblastic disease is a group of diseases which arise in the fetal chorion during various types of pregnancy. Histologically they are categorized as one of two types of hydatidiform mole (partial or complete), gestational choriocarcinoma or placental site trophoblastic tumour. Gestational trophoblastic disease is notable for several reasons. Firstly, the tumours are genetically different from the host, having antigens derived from the male partner. Secondly, apart from the placental site tumour, they secrete human chorionic gonadotrophin in amounts proportional to the viable tumour volume, allowing human chorionic gonadotrophin to be used as an ideal tumour marker. Thirdly, even metastatic disease can be cured with chemotherapy, the use of methotrexate in the early 1950s having shown reproducible results.

Complete and partial hydatidiform moles present as abnormal pregnancies ending in first or second trimester abortions. The complete mole is diploid (of paternal origin), is commonly diagnosed on ultrasound, and has no fetal elements present. The partial mole is triploid with paternal and maternal origin, has fetal elements present, and is usually diagnosed after the products of conception are examined pathologically. Gestational choriocarcinoma is a highly malignant tumour derived from syncytial and cytotrophoblastic cells. When villi are present in association with malignant trophoblasts, it is classified as a molar pregnancy. If there is diagnostic doubt about the possibility of combined molar pregnancy with a viable fetus, then the ultrasound scan should be repeated before intervention. If the twin pregnancy is associated with a partial mole, it should be allowed to proceed. If the twin pregnancy is associated with a complete mole, it may proceed after appropriate counselling. These pregnancies are associated with a reduced live birth rate of 25 per cent and are at risk of pre-eclampsia and haemorrhage. The subsequent need for chemotherapy in these rare cases is about 20 per cent, and is the same whether the pregnancy is terminated spontaneously or therapeutically, or allowed to proceed to term.

Gestational trophoblastic disease arises in various types of pregnancy, most of which are clinically recognized as abnormal. The incidence is 1.54 per 1000 live births. The most common are molar pregnancies, but gestational trophoblastic disease can also arise following abortions, ectopic pregnancies, or even after normal full-term pregnancies. Clinical surveillance of patients who have had a molar pregnancy is the only practical method of detecting and preventing gestational trophoblastic disease. In the United Kingdom, all patients with a histological diagnosis of a molar pregnancy are registered and followed up at one of three screening centres (Charing Cross Hospital in London, Sheffield, and Dundee). Only 7.5 per cent of women with hydatidiform mole require chemotherapy in the United Kingdom, and more than half of the patients who require chemotherapy for their gestational trophoblastic disease have a preceding molar pregnancy. Patients who develop gestational trophoblastic disease after an abortion or full-term pregnancy are more difficult to detect. They present with symptoms attributable to metastases. Sites of initial metastases (in order of frequency) are lung, vagina, brain, liver, gastrointestinal tract, and kidney. The interval between pregnancy and the development of metastatic gestational trophoblastic disease may be years. Because these tumours are rare, many clinicians do not consider gestational trophoblastic disease as part of any differential diagnosis.

Any woman of reproductive age who has an undiagnosed tumour or unexplained bleeding from any organ other than the uterus should have a human chorionic

gonadotrophin estimation to exclude the highly treatable gestational trophoblastic disease.

Patients with gestational trophoblastic disease are classified as having a low or high risk depending on a scoring system devised at Charing Cross Hospital and now modified by the World Health Organization. The score relies on factors such as age, the antecedent pregnancy, the interval between presentation and the previous pregnancy, the human chorionic gonadotrophin level, the blood group, the size of the largest tumour, site and number of metastases, and whether the patient had received prior chemotherapy. In the United Kingdom the low-risk group will be offered methotrexate with folinic acid rescue. The high-risk group and those low-risk patients who have resistant or persistent disease will be offered combination chemotherapy. The initial diagnosis may be made by surgical excision or biopsy of a suspicious lesion, but surgery otherwise has little role, excepting rarely to remove a cerebral metastasis to prevent a cerebral bleed.

The overall survival for patients with gestational trophoblastic disease is now about 94 per cent. Women should be advised not to conceive for 6 months after a negative human chorionic gonadotrophin reading. The risk of further molar pregnancy is low (1:74).

Cancer of the cervix

Carcinoma of the cervix may be diagnosed during pregnancy with an incidence of approximately 1 in 2200 pregnancies. In the United Kingdom, with the recent success of the cervical screening programme, the incidence is probably lower. Pregnant women with cervical cancer generally present with early stage disease and the prognosis is similar to that of non-pregnant patients.

The presenting symptom is usually vaginal bleeding. It is therefore important to check the cervix with a visual examination when pregnant women present with irregular vaginal bleeding. There is a tendency to assume that vaginal bleeding in early pregnancy is related to miscarriage, organize an ultrasound to check for fetal viability, and forget vaginal examination. In the case of an obvious cancer, a wedge biopsy under general anaesthetic is appropriate for diagnosis and staging. If there is any doubt, colposcopy can be used to assess the cervix. There is an increased risk of bleeding when taking a biopsy from the pregnant cervix, but there is no increased rate of fetal loss.

Patients with cervical intraepithelial neoplasia can be managed expectantly until after delivery. There is no contraindication for a vaginal delivery for women with cervical intraepithelial neoplasia. Indeed, there are several series which suggest that vaginal delivery is associated with a higher rate of regression of severe dysplasia than is usually seen. Standard practice would be to review with colposcopy at approximately 3 months after delivery. Management of women with microinvasion of the cervix is usually via cone biopsy under a general anaesthetic, allowing the pregnancy to continue.

When cervical cancer is diagnosed in early pregnancy, treatment options include immediate radical hysterectomy or to delay treatment until the fetus is viable, followed by classical caesarean section (scar in the upper segment of the uterus) and radical hysterectomy. This is appropriate for stage 1B cases, where the tumour is confined to the cervix and is less than 4 cm in diameter. In one series there was no difference in survival between the two modes of treatment. Typically, women diagnosed in the first trimester will be offered immediate surgery. Women diagnosed after 24 to 28 weeks' gestation are usually managed expectantly until after 32 weeks' gestation and then delivered by caesarean radical hysterectomy. Steroids are usually given to accelerate fetal lung maturity. The outlook may be worse for patients who deliver vaginally across a cervical cancer, but this has not been substantiated.

Cancer of the ovary

The stated incidence of adnexal masses occurring in pregnant women is from as rare as 1 in 2500 to as frequent as 1 in 81 live births. With the use of routine early ultrasound, the true incidence of adnexal masses is closer to the latter figure. Most of these (more than 95 per cent) are benign. Complications of a benign adnexal mass include pain due to torsion, rupture, and haemorrhage, obstruction of the pelvic outlet, and infection. Most cysts are managed conservatively, avoiding surgery. When necessary, surgery to remove cysts is usually performed in the second trimester. The advantage of waiting until the second trimester is that most cysts resolve spontaneously and that the rate of fetal loss is reduced.

The rationale for removing persistent adnexal masses is to exclude malignancy. Ovarian cancer in pregnancy is rare, with a reported incidence of 1 case per 17 000 to 38 000. This is because the usual age of childbirth is greater than the peak incidence of germ cell tumours and substantially less than the usual age of those with epithelial cancer. In addition, pregnancy protects against ovarian cancer. Two-thirds of the cancers detected are epithelial and the remaining are germ cell (usually dysgerminoma) and stromal cell types. Cysts which are simple on ultrasound and less than 5 cm in diameter have almost no malignant potential. Larger cysts with nodules, septa, or rapid growth are more likely to be malignant. Tumour markers are not helpful in pregnancy: CA 125 can be raised by pregnancy, as can a-fetoprotein and human chorionic gonadotrophin.

The management of the ovarian cancer is similar to that in the non-pregnant woman. Appropriate surgical staging is required: the author's preference being that removal of the cyst, taking of peritoneal washings for cytology, biopsy of the contralateral ovary, and biopsies of any abnormal areas are sufficient at the primary operation. It is also preferable to wait 48 h for a definitive diagnosis from paraffin sections, rather than expect the pathologist to give an immediate result from frozen section. This delay also allows the woman and her partner to consider the implications of the diagnosis. Most of the women seen with a malignant diagnosis in pregnancy will have early stage epithelial cancer: FIGO stage 1A or B, meaning well or moderately differentiated tumour confined to one or both ovaries, or to have borderline histology. No further therapy would then be necessary. Therapeutic termination is not required and pregnancy *per se* does not worsen outcome. Fuller staging may be considered 6 to 12 weeks after delivery. The decision to use chemotherapy postoperatively depends on the stage and differentiation of the tumour, the gestational age of the fetus, and the wishes of the mother. The treatment of malignant germ cell tumours can be carried out without affecting the pregnancy in the second two trimesters, especially if alkylating agents are avoided.

Cancer of the breast

Gestational breast cancer is defined as a breast cancer presenting either during pregnancy or up to 1 year postpartum. It was originally thought that pregnancy-related cancer carried a worse prognosis, but this has not been substantiated. Although breast cancer is regarded as a hormonal-dependent tumour, termination of pregnancy and oophorectomy do not provide a better outcome for the woman. Women becoming pregnant after treatment for breast cancer have a similar or better survival when controlled for age and stage.

Breast cancer is often diagnosed at a late stage as breast lumps may be difficult to detect against a background of pregnancy-related hypertrophy. Consequently, investigation of masses is often delayed. Mammography is not harmful to the fetus with appropriate shielding. When a breast mass is found, the most important step is to make a histological diagnosis. If the diagnosis of breast cancer is made, treatment is the same as for the non-pregnant woman. Obviously, chemotherapy in the first trimester is associated with risks for the developing fetus.

Suppression of lactation as a therapeutic manouevre is not necessary, with two exceptions. Firstly, if breast surgery is required during the puerperium, suppression of lactation can decrease the size and vascularity of the breast, allowing for a safer surgical procedure. Secondly, suppression of lactation is recommended in women receiving chemotherapy as some of the drugs can reach the breast milk and cause neonatal neutropenia.

Melanoma

The incidence of melanoma in pregnancy is between 0.14 to 2.8 cases per 1000 deliveries. Melanoma in pregnancy is unusual in that it can metastasize to the placenta and to the fetus. As this is a rare phenomenon, therapeutic abortion is not indicated, but careful examination and follow-up of the baby is warranted. Current evidence suggests that the clinical outcome for pregnant patients is similar to that of non-pregnant patients. Early detection and biopsy are performed as usual, and the surgical management is the same. Since most recurrences of melanoma occur in the first 3 years following initial diagnosis, it may be appropriate to delay further pregnancies until this time period has elapsed.

Thyroid cancer

It is not uncommon to find thyroid nodules which require further investigation during pregnancy. Most cancers are well differentiated with a very good prognosis. When a diagnosis is made, treatment proceeds as normal, with the exception that radio-iodine is contraindicated. Cancers discovered early in the pregnancy can be treated surgically in the second trimester. Tumours discovered in later pregnancy can have their investigation and treatment delayed until after delivery. Thyroxine is given to

reduce the level of thyroid-stimulating hormone. There is no evidence to suggest that pregnancy alters the outcome for thyroid cancer. Thyroid cancer is not an indication for termination of pregnancy.

Lymphoma

As Hodgkin's disease is a disease of young adults (mean age 32 years), it is not surprising that there are more cases diagnosed in pregnancy than there are of non-Hodgkin's lymphoma (mean age of diagnosis of 42 years). The reported incidence of Hodgkin's disease in pregnancy is between 1 in 1000 and 1 in 6000 deliveries.

Although historically believed to be exacerbated by pregnancy, there does not seem to be any influence of pregnancy on the outcome for Hodgkin's disease. If treatment is required, most patients can be managed without compromise to mother or fetus. Patients presenting with localized Hodgkin's disease relatively late in pregnancy may be observed with limited staging and not treated until after delivery.

By contrast, in non-Hodgkin's lymphoma, patients with Burkitt's lymphoma in pregnancy appear to have a highly aggressive disease involving the breast or ovary. The outlook for pregnant women with non-Hodgkin's lymphoma may be bleak, some dying prior to delivery.

Leukaemia

Leukaemia in pregnancy is rare, with an incidence of 1 per 100 000 pregnancies. This may be because the majority of cases of acute lymphoblastic leukaemia occur before reproductive age and the majority of cases of acute myeloid leukaemia occur afterwards. Chronic lymphocytic leukaemia is a disease of the elderly, hence chronic myeloid leukaemia constitutes 90 per cent of the cases of chronic leukaemia seen in pregnancy.

Since the introduction of intensive chemotherapy, the survival of pregnant women with leukaemia is similar to that of non-pregnant women. It does not appear that intrauterine exposure to antileukaemic chemotherapy produces detrimental late effects to the resulting children. Women treated with non-alkylating agents have no apparent decrease in fertility, although this is reduced by 33 per cent when alkylating agents are used.

Cancer of the colon

The reported incidence of colorectal cancer in pregnancy of 1 per 50 000 pregnancies may now be an underestimate as a reflection of the trend for women to delay pregnancy until later in life. A more recent study has reported an incidence of 1 per 13 000 live births. With increased awareness of inherited genetic traits and the availability of genetic testing, more and more patients at risk (for example those with familial adenomatous polyposis and hereditary non-polyposis coli) are undergoing screening. This may reduce the numbers of pregnant women diagnosed with colon cancer.

It appears that colorectal cancer in pregnancy is particularly common in the rectal region, below the peritoneal reflection. The importance of this is that 88 per cent of tumours are within reach of the flexible sigmoidoscope, allowing detection with a minimum of inconvenience to the patient and no risk to the fetus.

Presenting symptoms are similar to those in non-pregnant women. However, the combination of altered bowel habit, abdominal pain/swelling, and anaemia is common in pregnancy, such that these symptoms are frequently ascribed to the pregnancy itself. Assessment of the pregnant patient with colorectal cancer is similar to that of the non-pregnant patient. Radiological imaging is avoided in the first trimester. Carcinoembryonic antigen is not affected significantly by pregnancy and so can be used as a marker.

Patients younger than 40 years generally have a poorer prognosis due to delayed diagnosis and advanced stage at presentation. Pregnant women are no different in this respect. The overall fetal prognosis is relatively favourable as the diagnosis is usually made close to term and the fetus can be delivered coincident with the surgery for the colon cancer.

Further reading

Cappell MS (1998). Colon cancer during pregnancy. The gastroenterologist's perspective. *Gastroenterology Clinics of North America* **27**, 225–56. A good review of a rare condition.

Seminars in Oncology **16**, 335–436 (1989). This volume is dedicated to cancer in pregnancy and gives various overviews relating to gynaecological cancers, leukaemia and lymphoma, melanoma, and breast cancer. It is perhaps a little dated.

http://www.hmole-chorio.org.uk/ Choriocarcinoma UK Information website with up to date recommendations about management.

13.18 Prescribing in pregnancy

P. C. Rubin

Introduction

Prescribing in pregnancy is essentially about balancing risks. The damage that a drug may cause to the fetus must be weighed against the harm that may befall the mother and her unborn child if a disease goes unchecked.

While knowledge in most therapeutic areas has grown rapidly in recent decades, information on the use of drugs in pregnancy has developed sporadically, with case reports being more usual than large, prospective clinical trials. The reasons are not surprising and largely relate to concern about teratogenesis.

Thalidomide is a name inescapably associated with prescribing in pregnancy. Drug-induced fetal abnormality did not begin with thalidomide: there is an Old Testament exhortation to have 'no strong drink, neither eat any unclean thing' during pregnancy. However, the scale of the thalidomide tragedy brought to the general public for the first time the realization that drugs could harm the developing baby. Thalidomide was marketed in Germany in 1956 and subsequently in other countries as a sedative and hypnotic which had the particular attraction of being safe in overdose. Indeed, the drug was considered so safe that in some countries it was available without prescription. Then between 1960 and 1961 Germany experienced what amounted to an epidemic of phocomelia, a birth defect involving absence of the long bones with hands and feet being attached directly to the trunk. What had previously been an extremely rare condition (no cases had been reported in the 10 years to 1959) was being seen almost commonly. Various causes—viral, radioactivity, food preservatives—were considered as culprits, until one doctor retrospectively questioned his patients and found that 20 per cent had taken thalidomide in early pregnancy. On repeat questioning, asking specifically about the drug, 50 per cent admitted taking thalidomide, many having not mentioned it before since the drug was so obviously innocent. In fact, around 80 per cent of women who took thalidomide in the first trimester had a deformed baby. More than 10 000 such babies had been born before the drug was removed from the market.

The thalidomide experience had far-reaching ramifications. Drug regulation as we know it stems largely from the disaster. Doctors and their patients recognized that there is no such thing as a safe drug. In addition, the pharmaceutical industry has largely avoided obtaining systematic information on drug use in pregnancy. The reasons are obvious and understandable, but for the prescribing doctor the statement that 'the safety of this drug in pregnancy has not been established' is not helpful when faced with a woman who is, or may become, pregnant.

Identifying teratogenic drugs

Information on drug-induced fetal abnormality comes from case reports, case studies, and epidemiological studies. Case reports are a two-edged sword. Describing a single association between a drug and a fetal abnormality can be very useful in first identifying a real problem: warfarin was first linked to teratogenesis in this way. However, the problem with case reports is that they may be showing nothing more than a chance association, because fetal abnormalities occur in around 2 per cent of pregnancies, and caution must be exercised in their interpretation. This is well demonstrated by the Debendox® saga.

Most cases of morning sickness do not require treatment. However, some do, and the drug for which most information is available was withdrawn from the market in 1983 in view of mounting public concern about its safety. This drug was a mixture of doxylamine succinate and pyridoxine hydrochloride and was marketed as Debendox® or Bendectin®. Despite having been used by over 30 million pregnant women over a quarter of a century, and notwithstanding carefully designed clinical trials suggesting that the drug was not teratogenic, individual case reports linking the use of the drug to fetal abnormality were given considerable publicity and led to its withdrawal. In view of the extremely high number of exposures, many chance associations between drug use and fetal abnormality were inevitable. This episode illustrated that in an emotional area such as the use of drugs during pregnancy, well-chosen and carefully presented anecdotes can be more powerful than a substantial body of scientific data carefully accumulated over many years.

Case studies are more secure in that they describe several patients where the same drug and malformation were linked: phenytoin and the retinoids were found to be teratogenic in this way. Epidemiological studies are of two major types: cohort studies, which prospectively study exposed and unexposed groups, and case-control studies, which retrospectively compare the pregnancies of abnormal and normal offspring. So far as teratogenesis is concerned, case-control studies are the norm because of the size and expense of cohort studies. The relationship between diethylstilbestrol use in the first trimester and vaginal adenocarcinoma in teenage offspring was found in a case-control study.

The effect of drugs on the fetus

A drug can harm the fetus only if it crosses the placenta, but most drugs do. The placenta offers a lipid barrier to the transfer of drugs, and the rate at which a drug crosses from mother to baby will depend on its lipophilicity and polarity. However, with the exception of drugs administered acutely around the time of delivery, the rate of transfer is of little importance, and for any course of drug treatment it should be assumed that transfer will occur. The only notable exceptions are heparin—including low molecular weight preparations—and insulin.

Drugs can adversely affect the developing fetus in different ways depending on the gestation at which exposure occurs. For this reason it is appropriate to consider organogenesis, fetal growth and development, the breast-fed infant, and childhood growth and development separately.

Prescribing in the first trimester

Organogenesis occurs between 18 and 55 days of gestation and it is during this time that drugs can cause anatomical defects. A drug can cause a teratogenic effect only if it is present in the embryo during organogenesis, and even a definite teratogen will not cause a structural defect if it is given following this period. These seemingly obvious statements become relevant in prepregnancy counselling and in providing advice when exposure to a possible teratogen has occurred during pregnancy. Being present in the embryo during organogenesis is not necessarily synonymous with being prescribed during this period. The retinoids are stored in adipose tissue and released slowly, so a teratogenic effect can occur long after the course of treatment has been completed. It is important to recognize that teratogenic effects are not seen in all cases: on the contrary, most first trimester exposures to teratogenic drugs will not harm the baby. Clearly there is more to drug-induced fetal abnormality than simply the drug: the genetic make up of the baby is important too. Some drugs that are definitely teratogenic in the human, together with approximate risks, are listed in Table 1.

Preventing drug-induced teratogenesis—short of the obvious solution of not taking the drug—is difficult. Where there is a risk of neural tube defect, folic acid 5 mg daily should be prescribed from the time that pregnancy is planned. No direct evidence currently exists to support this approach, but the approach is logical: folic acid is known to be effective in the secondary prevention of naturally occurring neural tube defect, and anticonvulsants lower folate levels. Some reports have also claimed a direct relationship between dose and fetal abnormality. For example, one meta-analysis involving over 1000 babies exposed to sodium valproate found a higher risk of abnormality at doses above 1 g per day compared with less than 600 mg daily. In epileptic pregnancies, polypharmacy is accompanied by a greater risk of fetal abnormality, but since these women are likely to have the more severe forms of the disease, cause and effect is hard to establish. Many of the abnormalities caused

by these drugs can be detected by detailed ultrasound scanning at 18 to 20 weeks' gestation. However, the defects caused by warfarin involve mainly soft tissue and do not fall into this category.

Table 1 is not comprehensive and includes only those drugs commonly encountered in general medical practice. Some drugs used in specialist areas are teratogenic, for example several drugs used in cancer chemotherapy. Many more drugs may be teratogenic in a small percentage of exposures, but definitive information is not available because both prediction and detection of human teratogens is difficult. Predicting the effect of a drug in the human usually depends on studying its pharmacology in experimental animals. This is not fruitful in the area of teratogenesis because species variation is so great. For example, thalidomide causes phocomelia only in primates, while lithium causes cardiac abnormalities in humans at doses that produce no effect in the rat. Detecting teratogenic effects is complicated by the normal occurrence of fetal abnormalities, hence if a drug is teratogenic very occasionally it can be very difficult to distinguish its effects from those arising naturally.

Even if a drug is a teratogen, the balance of benefits and risks may still be in favour of its use. For example, chloroquine and proguanil are indicated for malarial prophylaxis in areas where *Plasmodium falciparum* remains sensitive. Currently available evidence suggests that chloroquine may cause a very small increase in birth defects: in one study 169 infants whose mothers took chloroquine-base 300 mg once weekly were compared with 454 children whose mothers took no drug. Abnormal babies were born to 1.2 per cent of the treated group, compared to 0.9 per cent of the controls: not a significant difference, but the study was too small to detect anything less than a fivefold increase in abnormality rate. By contrast to the possibility of this small increase in risk, malaria presents a major risk to the health and life of both mother and baby, particularly when an expatriate woman is travelling in an endemic area. The argument in favour of using prophylaxis is therefore overwhelming—but not so overwhelming as the advice for pregnant travellers to avoid malarial areas!

Similar arguments apply to corticosteroids, which have acquired a reputation for causing oral cleft defects. The evidence in support of this effect is at best conflicting and is easily outweighed by the benefits of steroids in conditions such as severe asthma, inflammatory bowel disease, systemic lupus erythematosus, or organ transplantation. The placenta inactivates around 90 per cent of prednisolone, whilst corticosteroids, such as betamethasone, that are used to accelerate fetal lung maturity have much greater penetration to the fetus.

Prescribing later in pregnancy

Beyond organogenesis, the fetus undergoes growth and development. The scope for producing anatomical defects has largely passed, exceptions being premature closure of the ductus arteriosus caused by indometacin and bleeding into the fetal brain produced by warfarin. Growth and function tend to be the targets of drug adverse effects for the remainder of the pregnancy.

The possible effects of some commonly used drugs later in pregnancy are shown in Table 2.

Drugs and breast feeding

Most women now elect to breast feed their babies, and the majority will take a drug during this time. Iron, mild analgesics, antibiotics, laxatives, and hypnotics are the most commonly used. Much work has been performed on the pharmacokinetic aspects of breast feeding, but systematic studies on the effect of drug ingestion by the mother on her breast-fed baby are lacking.

Milk consists of fat globules suspended in an aqueous solution of protein and nutrients. Drugs move from plasma to milk by passive diffusion of the unionized and non-protein-bound fraction. Since breast milk has a slightly lower pH than plasma, drugs that cross most extensively into breast milk are lipid-soluble, poorly protein-bound, weak bases. However, even for drugs that do cross readily into breast milk, considerable dilution has already occurred in the mother. Thus, when the concentration of a drug in breast milk and the volume of the milk consumed by the baby are translated into a dose, it is often the case that the baby receives too little drug to have any detectable pharmacological effect.

Some of the more commonly used drugs that, on the basis of experience, have a good safety record in breast-feeding mothers are listed in Table 3. It will be seen from this that many of the drugs that would be indicated for common medical problems in this context are safe to use. However, some qualification is needed about two of the drugs listed in Table 3. Oestrogen-containing oral contraceptives may suppress lactation if they are taken before the milk supply is well established, and in some women may do so even after this time: progestogen-only contraceptives do not influence lactation at any stage. Metronidazole is not harmful to the baby but is said to make the milk taste bitter and may therefore interfere with feeding.

Some drugs have been shown to affect the baby when ingested in breast milk: these are listed in Table 4. There are several other drugs for which theoretical risks exist, or for which isolated reports of serious adverse consequences have appeared. For example, aspirin is contraindicated in young children because of the possible association with Reye's syndrome, and some authorities consider that the drug should therefore be avoided in women who are breast feeding. No evidence is available to support this view, but unless the use of aspirin is considered essential in a breast-feeding woman (and such an eventuality must be rare), then it is probably best avoided. Similarly, indometacin has been associated with one case of neonatal convulsion when used during lactation: a decision with regard to its appropriateness in any given patient would depend on the likelihood of real benefit accruing from its use.

Behavioural teratology

The most obvious consequences of a drug-induced fetal abnormality occur at or shortly after birth in the form of anatomical defects, and studies in teratology have largely concentrated on immediate pregnancy outcome. However, drugs can, on occasion, cause problems that become manifest only after several years. The most striking example is diethylstilbestrol which, when given during early pregnancy, can lead to adenocarcinoma of the vagina in teenage offspring. In addition to late morphological effects, concern has been expressed that drugs given during pregnancy can influence behavioural development, although the available evidence is to the contrary.

Anticonvulsants

Several studies have claimed that the use of anticonvulsants during pregnancy is associated with impaired intellectual development of the children, but it is difficult to carry out studies in this area and the choice of control group is crucially important. When all children of treated epileptic mothers in a single hospital in Finland were studied prospectively, using the offspring of untreated epileptic women and age-matched children of the same social class as controls, no difference was found in intellectual development at the age of 5.5 years. At present it appears likely that, in the absence of any obvious morphological abnormality at birth, anticonvulsant use during pregnancy is not associated with impairment of intellectual development but studies of sufficient statistical power have not yet been performed.

Antihypertensive drugs

One of the earliest trials on the treatment of hypertension during pregnancy involved a comparison of methyldopa with no treatment. The children underwent physical and psychomotor assessment at 4 and 7.5 years. The 4-year-old children from the treatment group had a slightly smaller head circumference than their untreated controls, but there were no other physical or psychomotor differences. The evaluation at 7.5 years revealed no differences between the two groups. The reputation of methyldopa as a safe drug in pregnancy is largely based on this very well-conducted study.

The effects on childhood development of atenolol versus placebo have similarly shown no detrimental effects, a wide range of physical and psychomotor tests being performed on the children at the age of 1 year.

Effect of pregnancy on drugs

Influence of pregnancy on dose requirements

While the emphasis on what drugs can do to the pregnancy is both understandable and appropriate, the physiological changes of pregnancy can have a clinically important influence on drug disposition and effect. The plasma concentrations of some drugs fall to an extent that is clinically important during pregnancy.

Among the many physiological changes in pregnancy, the most important from the standpoint of drugs are those that influence clearance. By the third trimester renal blood flow has nearly doubled and the activity of some, but not all, liver metabolic pathways is increased during pregnancy. A further factor tending to reduce drug concentrations is an increase in body water, with around an additional 7 litres being retained by the end of pregnancy.

The importance of these changes is well illustrated by the influence of pregnancy on anticonvulsant dose requirements. The plasma concentrations of phenytoin and carbamazepine decrease as pregnancy progresses. An increase in systemic clearance is the main reason—for example, the clearance of phenytoin increases by over 100 per cent by the third trimester—with an increased volume of distribution making a further contribution. An example of the influence of pregnancy on the concentration of phenytoin is shown in Fig. 1. The reduction in anticonvulsant concentration can be substantial and, if the dose is not increased, then seizure control may be lost. The physiological changes of pregnancy resolve in the 6 weeks following delivery, and there is a progressive return to prepregnancy dose requirements during this time.

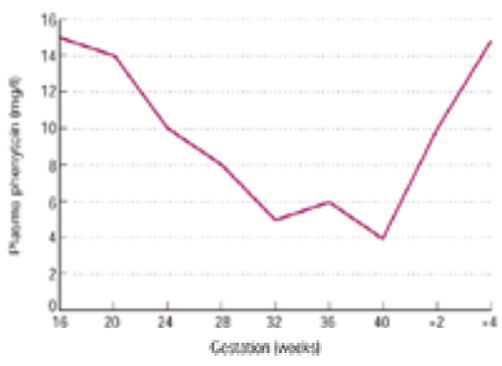

Fig. 1 Plasma phenytoin concentration during and following pregnancy in a woman who remained on a constant dose of 300 mg/day throughout. She had a seizure at 38 weeks' gestation and delivered at 40 weeks. The dose should have been increased when the phenytoin concentration began to fall.

Not all drugs metabolized in the liver show reductions in plasma concentration during pregnancy. For example, the clearance of propranolol is unchanged, presumably because this is determined by liver blood flow, which is not altered by pregnancy.

Since renal blood flow increases during pregnancy, the clearance of drugs eliminated by this route can also be expected to increase. Lithium clearance doubles during pregnancy, so that dose increases, guided by drug-level monitoring, are likely to be needed. Dose requirements fall rapidly following delivery and care must be taken to avoid the development of toxicity. The clearance of ampicillin nearly doubles during pregnancy. Formal pharmacokinetic studies have not been performed with cephalosporins, but plasma levels of around 50 per cent of those found in non-pregnant subjects have been reported. By contrast to drugs with a reasonably well-defined therapeutic range, the falling plasma levels of penicillin or cephalosporin antibiotics are of less obvious significance. However, it seems prudent to give doses at the higher end of the recommended range when using these agents to treat systemic infections during pregnancy.

Drug protein binding in pregnancy

The protein binding of drugs is also altered by pregnancy. The mechanism is not fully understood: although the concentration of albumin falls substantially in a normal pregnancy, there is no correlation between the concentration of albumin and the free fraction of the drug, at least not for all drugs. The free and pharmacologically active concentration of anticonvulsants is increased in pregnancy by 30 to 50 per cent, which has consequences for the interpretation of plasma drug levels.

Therapeutic drug monitoring during pregnancy

Epilepsy is the commonest condition for which therapeutic drug monitoring during pregnancy should be performed. This area is controversial because in non-obstetric practice therapeutic drug monitoring is considered of less value than seizure control as a guide to drug management. However, during pregnancy there is a high likelihood that drug levels will fall because of pharmacokinetic changes. In addition, measuring levels is a useful guide to poor compliance with treatment, which is a feature of the pregnant epileptic. Waiting for a seizure to occur is not without risk: women die from poorly controlled epilepsy in pregnancy. Since the free fraction of anticonvulsants increases during pregnancy, it is the unbound level that should preferably be recorded. Alternatively, saliva samples can be used to guide treatment, since these have been shown to correlate well with the plasma concentration of unbound drug.

Further reading

Briggs GG, Freeman RK, Yaffe SJ (1998). *Drugs in pregnancy and lactation*, 5th edn. Williams and Wilkins, Baltimore.

Rubin PC (2000). *Prescribing in pregnancy*, 3rd edn. British Medical Journal, London.

13.19 Benefits and risks of oral contraceptives

M. P. Vessey

Introduction

The basic physiological principles underlying a hormonal approach to contraception had already been elaborated by the mid 1930s, but the development of practical methods of hormonal birth control had to await the synthesis of potent orally active steroids some 20 years later. Much of the physiological and clinical development of the 'the pill' was done by Pincus and Rock in the United States in the 1950s; great credit must be given to these two for their contribution to one of the great medical breakthroughs of the twentieth century. Indeed, in 1999 it was estimated that about 110 million women were taking oral contraceptives, 62 million of them in developing countries.

There are several different types of oral contraceptive regimen, but the most important preparations include both an oestrogen and a progestogen. In the United Kingdom, only two oestrogens have been used, ethinylestradiol and mestranol, but seven progestogens are currently available (northisterone, norethisterone acetate, ethynodiol diacetate, levonorgestrel, desogestrel, norgestimate, gestodene) while others have been used in the past (e.g. norethynodrel, chlormadinone acetate, megestrol acetate). Since the dosage of the constituent steroids may be varied (the trend has generally been downwards over the years both for the oestrogen and progestogen components), it is not surprising that the number of different oestrogen–progestogen formulations marketed currently or in the past is very large—over 100 in the United Kingdom. This adds greatly to the difficulties confronting those trying to assess safety.

Oral contraceptives have many metabolic effects, although these are fewer with modern preparations than with earlier ones. None the less, it has been said that 'almost every metabolic parameter that is capable of laboratory investigation has been reported to be altered in one way or another by some contraceptive steroid'. This implies that the results of routine laboratory tests may be altered by oral contraceptives, a point of considerable practical importance. In this chapter, however, attention will be largely concentrated on effects of the pill on morbidity and mortality, as revealed by epidemiological studies.

Until the mid 1970s, most of the available data about the benefits and risks of the pill had been derived from uncontrolled clinical trials and from case-control studies. One large-scale randomized study, including 9757 women allocated either to an oral contraceptive or to a vaginal method of contraception, had been started in Puerto Rico in 1961 by Pincus, but it proved to have serious shortcomings and contributed little to knowledge. Since the mid 1970s, an enormous amount of epidemiological information has been obtained from two large British cohort studies, the Royal College of General Practitioners Oral Contraceptive Study and the Oxford–Family Planning Association (Oxford–FPA) Contraceptive Study. Between them, these investigations recruited 63 000 women of child-bearing age (about half of whom were users of the pill and half users of other methods or no method of contraception) who have now been carefully followed up for an average of around 25 years. Many of the findings described in this chapter are derived from these two cohort studies.

Information about the benefits and risks of oral contraception in the developing world is sparse. The reader is cautioned not to extrapolate the data summarized here to parts of the world to which they clearly do not apply.

Benefits of combined oral contraceptives

High efficacy

By far the most important beneficial effect of these preparations is their remarkable efficacy which, coupled with a high degree of acceptability (at least among the young), has given many women freedom from anxiety about pregnancy. If taken conscientiously, no more than about two to four women in every thousand using a combined preparation should become accidentally pregnant each year. In practice, pills are often missed and much less satisfactory results are then obtained.

Suppression of menstrual disorders

It has long been known that oral contraceptives suppress some menstrual disorders, notably menorrhagia and dysmenorrhoea, leading to a reduction in hospital referral for diagnosis and treatment of these conditions and to a lessened risk of iron-deficiency anaemia.

Suppression of benign breast disease

Epidemiological studies have consistently shown that use of older, higher-dose oral contraceptives seems to decrease the occurrence of benign lumps in the breast, reducing the need for hospital referral for diagnosis and treatment of such lesions by up to 50 per cent. The effect is most pronounced in long-term users, appears to wear off after discontinuation of use, and is probably attributable to the progestogen component of the pill. More recent studies considering lower-dose pills have tended to find less impressive effects on benign breast disease than did the earlier studies.

Pelvic inflammatory disease

Oestrogen–progestogen oral contraceptives reduce the risk of symptomatic pelvic inflammatory disease and probably reduce the severity of the disease as well. There is, however, evidence that the protective effect of the pill is less against pelvic inflammatory disease caused by chlamydial infection than against disease caused by other organisms.

Suppression of functional ovarian cysts

Since oral contraceptives act principally by inhibiting ovulation, it is not surprising that follicular cysts and particularly corpus luteum cysts are relatively uncommon in pill users, although this may apply less to modern, very low-dose pills than to older, high-dose ones.

Suppression of ovarian cancer and endometrial cancer

Epidemiological studies reported during the past 15 years have demonstrated that the risk of both epithelial ovarian cancer and endometrial cancer is reduced by up

to 50 per cent in women who have used combined oral contraceptives for at least 5 years. Longer durations of use offer additional protection while shorter durations of use still provide some beneficial effect. The protective effect appears to persist for many years after cessation of pill use; this is important from the public health point of view since ovarian and endometrial cancer are rare in young women amongst whom oral contraceptive use is most prevalent.

Other possible beneficial effects

While the beneficial effects already described may be considered established, a number of others have been reported in some studies and deserve mention. These include a lessened risk of thyroid disease, rheumatoid arthritis, fibroids, endometriosis, and colorectal cancer. An increased peak bone mass has also been reported in long-term users. Further work is necessary before the significance of these observations can be adequately assessed.

Risks of combined oral contraceptives

Oral contraceptives are well known to cause minor side-effects such as nausea, headache, and breast tenderness. Although such symptoms are common enough and unpleasant enough to lead to discontinuation of the pill by up to 25 per cent of women, they disappear when medication is stopped and do not represent a serious threat to health.

Cardiovascular effects

The best known substantial adverse effects of oral contraceptive use are cardiovascular, comprising venous thrombosis and embolism, thrombotic stroke, and acute myocardial infarction. The evidence concerning haemorrhagic stroke is rather less convincing. All these conditions are rare in young women and the absolute risk associated with pill use is very low. The risks, in most studies, are confined to current pill users and do not depend on duration of pill use. The risk of acute myocardial infarction (and to a lesser extent stroke) in pill users seems to be concentrated in women with other risk factors for cardiovascular disease, notably cigarette smoking and hypertension.

In 1995, findings were published from a multinational study by the World Health Organization which indicated that the risk of venous thromboembolism attributable to oral contraceptive use was twice as high in women using pills containing gestodene or desogestrel as in women using pills containing older progestogens. Some additional studies have confirmed this observation while others have not, and this remains a controversial topic.

The mechanisms underlying adverse cardiovascular reactions to the pill are uncertain. However, oral contraceptives have effects on the coagulation system, on serum lipids, on carbohydrate metabolism, on blood pressure, and on the structure of vessels. Any or all of these effects might be of significance. In recent years, there has been great interest in the effect of oral contraceptive use on venous thromboembolism in women with thrombophilic disorders. Thus the risks associated with factor V Leiden and current oral contraceptive use appear to multiply, such that a woman with both these characteristics may have a risk of venous thromboembolism which is about 30 times higher than that in a woman without either of them.

Breast cancer

Large numbers of studies have been conducted on oral contraception and breast cancer, but no consensus was reached on their interpretation until a collaborative re-analysis of 54 studies was published in 1996. It was found that the relative risk of having breast cancer diagnosed in current users of combined oral contraceptives in comparison with never-users was 1.24. There was also some increase in risk in ex-users, but this was no longer apparent 10 or more years after stopping. Cancers diagnosed in women who had used combined oral contraceptives were clinically less advanced than those diagnosed in women who had never done so. It is not known whether there is an increased risk of dying from breast cancer as well as an increased risk of having breast cancer diagnosed in women using the pill.

Hepatocellular adenoma and carcinoma

Hepatocellular adenoma and carcinoma are extremely rare (but serious) conditions in women of child-bearing age. In those without exposure to the pill, the incidence might be around one per million per annum. Oral contraceptive users suffer a higher incidence than this, but there is reason to believe that the increase in risk is much less in those taking modern, low-dose pills.

Impairment of fertility

Despite a vast literature, prior use of oral contraceptives has not been incriminated either as a cause of prolonged secondary amenorrhoea (say absence of periods for more than 6 months) or of prolactinoma of the pituitary, which is sometimes associated with this condition. Many women do, however, experience some temporary impairment of fertility after stopping the pill, especially those over the age of 30 trying to have a first baby. In the majority this lasts only a month or so, but in some recovery may be much slower. It seems unlikely that oral contraceptives are ever a cause of permanent infertility.

Other possible adverse effects

A large number of studies have indicated a positive association between long-term oral contraceptive use and cervical cancer. The interpretation of this association is uncertain, mainly because cancer of the cervix is so strongly associated with sexual activity that it is extremely difficult to isolate any independent effect of the method of contraception used.

Several studies have examined the possible relationship between oral contraceptive use and malignant melanoma; they have not given consistent results. The same is true for studies of hydatidiform mole and choriocarcinoma.

In the past, there was considerable anxiety about an increase in the risk of cholelithiasis in pill users. Further work has shown the effect, if present, to be minimal. The evidence concerning chronic inflammatory bowel disease is more convincing. Many other possible adverse effects of oral contraceptives have been suggested, including depression, urinary tract infection, and fetal malformation if taken inadvertently during pregnancy. In every case, the balance of evidence is not compelling. A recent concern centres around human immunodeficiency virus (HIV) infection being commoner in pill users, with worrying findings in prostitutes in Nairobi. These results have not, in general, been replicated in other studies and there is no clear evidence of an adverse effect of pill use on the risk of HIV infection.

Progestogen-only oral contraceptives

Low doses of progestogens taken every day by mouth have been extensively investigated as contraceptives. Such preparations do not consistently inhibit ovulation and their mode of action is uncertain. Their efficacy is lower than that of the oestrogen–progestogen pill, but they can give entirely adequate protection in older women who may prefer not to use conventional pills. A major disadvantage of progestogen-only oral contraceptives is their tendency to disrupt the menstrual cycle in many women, producing irregular bleeding, whilst those who become accidentally pregnant when using them have about a 5 per cent chance of an ectopic gestation. These drawbacks probably account for the fact that progestogen-only pills represent only a small fraction of all oral contraceptives consumed. Their main advantage is that they appear to be free from the undesirable metabolic effects of combined preparations. In addition, they can be taken safely by women who are breast feeding without risking the reduction in milk production that usually occurs if combined oral contraceptives are taken.

Balance of benefits and risks

A number of authors have provided analyses of varying degrees of complexity in which they have attempted to weigh up the benefits and risks of taking the pill. Three approaches are outlined here. In the first, which uses hospital inpatient morbidity data from the Oxford–FPA study, supplemented where necessary by other epidemiological data, a comparison is made over a 1-year period of women using either oral contraceptives or a condom for contraception. The approach is fully described elsewhere (see Further reading), but Table 1 summarizes the main findings. Despite the sizeable cardiovascular risks (older 50 µg pills were used in the Oxford–FPA study) the overall results are quite favourable as far as oestrogen–progestogen oral contraceptives are concerned. Similar analyses reported from the Royal College of General Practitioners Oral Contraceptive Study have produced comparable results.

The second approach involves constructing models that consider what is known about the effects of oral contraceptives and other methods of birth control on

mortality, and estimating the balance of benefit and risk over a period of many years. This approach is too complex to describe here because of space limitations, but details are available elsewhere (see Further reading). Not surprisingly, the results obtained in such analyses depend to an important extent on whether or not oral contraceptives increase the risk of death from breast cancer (as opposed to merely increasing the rate of diagnosis of the disease), the length of time after discontinuation of oral contraceptive use that the protective effect against epithelial ovarian cancer and endometrial cancer persists, and the ages at which oral contraceptives are used. Considering the current pattern of oral contraceptive use in developed countries, which is predominantly by women below the age of 35 years, the balance of benefits and risks seems entirely acceptable.

The final approach is a more direct one. It involves examination of the mortality rates observed in the major cohort studies, comparing oral contraceptive users with non-users. Data published in 1989 from the Oxford–FPA study relating to the first 238 deaths are shown in Table 2. Although the numbers are small, the pattern of mortality reflects what is known about the effects of oral contraceptive use from other studies. Overall, the mortality ratio comparing users with non-users is 0.93 (95 per cent confidence interval 0.71–1.22). In a corresponding analysis from the Royal College of General Practitioners study including 1599 deaths, the overall mortality ratio was 1.02 (95 per cent confidence interval 0.92–1.13), while the large Nurses' Health cohort study in the United States reported an overall mortality ratio of 0.93 (95 per cent confidence interval 0.85–1.01) based on 2879 deaths. These data clearly offer considerable reassurance about the effects of oral contraceptives.

The pill has been studied extremely intensively over the past four decades. On the whole, it has stood up well to close scrutiny. It remains an excellent method of contraception for younger women. There remains some doubt, however, about its suitability for those aged over 40. Women in this age group are, however, usually well served by progestogen-only pills if a reversible method of contraception is required.

Further reading

Beral V *et al.* (1999). Mortality associated with oral contraceptive use: 25 year follow up of cohort of 46,000 women from Royal College of General Practitioners' Oral Contraception Study. *British Medical Journal* **318**, 96–100.

Colditz GA for the Nurses' Health Study Research Group (1994). Oral contraceptive use and mortality during 12 years of follow-up: the Nurses' Health Study. *Annals of Internal Medicine* **120**, 821–6.

Collaborative Group on Hormonal Factors in Breast Cancer (1996). Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53,297 women with breast cancer and 100,239 women without breast cancer from 54 epidemiological studies. *Lancet* **347**, 1713–27.

Hannaford PC, Kay CR (1998). The risk of serious illness among oral contraceptive users: evidence from the RCGP's oral contraceptive study. *British Journal of General Practice* **48**, 1657–62.

International Agency for Research on Cancer (1999). Hormonal contraception and post-menopausal hormonal therapy. In *Monographs on the evaluation of carcinogenic risks to humans*, vol. 72. WHO, IARC, Lyon.

Oldfield K, Milne R, Vessey M (1998). The effects on mortality of the use of combined oral contraceptives. *British Journal of Family Planning* **24**, 2–6.

Vessey MP (1990). The Jephcott Lecture 1989. An overview of the benefits and risks of combined oral contraceptives. In: Mann RD, ed. *Oral contraceptives and breast cancer*, pp 121–32. Parthenon, Carnforth.

Vessey MP, Smith MA, Yeates D (1986). Return of fertility after discontinuation of oral contraceptives: influence of age and parity. *British Journal of Family Planning* **11**, 120–4.

Vessey MP *et al.* (1989). Mortality among oral contraceptive users: 20 year follow up of women in a cohort study. *British Medical Journal* **299**, 1487–91.

WHO Scientific Group (1998). Cardiovascular disease and steroid hormone contraception. *WHO Technical Report Series* no. 877. WHO, Geneva.

13.20 Benefits and risks of hormone replacement therapy

J. C. Stevenson

Introduction

The acute effects of female sex hormone deficiency such as vasomotor symptoms are well known, but the importance of the longer-term effects of ovarian failure have only been recognized recently. The menopause, the time of a woman's last menstrual period, is a useful marker for ovarian failure and occurs naturally at an average age of around 51 years, although it may occur at any time after puberty. A postmenopausal state can usually be inferred by the absence of menses for 12 months in a woman of appropriate age. The demonstration of elevated gonadotrophin levels may help to confirm the diagnosis in hysterectomized women.

Clinical features of menopause

A number of symptoms may arise soon after loss of ovarian function at the menopause. These include hot flushes and night sweats, and psychological symptoms such as mood swings, depression, anxiety and irritability, and difficulties with memory and concentration. Later there may be genitourinary problems such as vaginal dryness and dyspareunia, and increased urinary frequency and urge incontinence. However, it is the long-term consequences of hormone deficiency, particularly osteoporosis and cardiovascular disease, which pose a major health problem for women.

The menopause is recognized as a substantial risk factor for the development of osteoporosis, and perhaps one in every two women will have this disease by the end of their lives. The classical osteoporotic fractures are of the vertebrae, distal forearm, and proximal femur, but osteoporosis may also result in fractures of the ribs and pelvis, proximal humerus, ankle, and phalanges. Hip fracture is by far the most serious in terms of morbidity, mortality, and cost to the health service: one in five women die and at least 50 per cent of the remainder end up in institutionalized care, such that the health service costs of osteoporosis are now approaching a billion pounds annually in the United Kingdom.

Increased risk of cardiovascular disease is the most important consequence of ovarian failure. Coronary heart disease is the leading cause of death in women, and although it occurs at a later age than in men, overall more women than men die from the disease. The occurrence of coronary heart disease in women is frequently overlooked, and women are less likely than men to undergo both investigation and treatment for this disease.

It is now becoming apparent that oestrogen deficiency has profound neurological effects. Alzheimer's dementia is more common in elderly women than men, and the menopause has adverse effects on the central nervous system, including cognitive function.

Benefits of hormone replacement therapy

The main indications for the use of hormone replacement therapy are relief of menopausal symptoms and prevention of osteoporosis. Hormone replacement therapy will abolish vasomotor symptoms, often within days of starting treatment, whilst psychological and genitourinary symptoms may take weeks or even months to respond. It is therefore worthwhile persisting with therapy for several months in the absence of rapid symptomatic response, and treatment should be continued for at least several months after symptomatic relief has been obtained.

Hormone replacement therapy is well established for both the prevention and treatment of osteoporosis, and is as effective as any other agent currently available. It conserves, and to some extent increases, bone density and results in a reduction in the risk of fracture. Therapy should be offered to any woman considered at increased risk of osteoporosis, and particularly those with an early menopause. When the risk of osteoporosis is uncertain, bone density measurement can greatly aid clinical decision-making.

Hormone replacement needs to be given long-term when started in the early postmenopause, but few women are at immediate risk of osteoporotic fracture at this age. An alternative strategy is to commence hormone replacement in elderly women who are at a much greater risk for osteoporotic fracture. This approach results in a relatively rapid reduction in risk of fracture, and is thus more cost-effective. It also avoids the necessity of prolonged therapy. However, the elderly are less tolerant of the side-effects of treatment, particularly cyclical bleeding and mastalgia, and regimens that avoid bleeding are to be preferred for this age group. This includes those with lower doses of oestrogen than used in the early postmenopause, which are effective for bone conservation in the elderly. Cessation of hormone replacement therapy leads to a loss of bone density, but only at the usual postmenopausal rate, and the benefit gained by the skeleton from a suitable period of treatment persists into old age.

It is most likely that prevention of cardiovascular disease will become a major indication for hormone replacement therapy in the future. There are many mechanisms, both established and potential, whereby hormone replacement therapy might benefit the cardiovascular system, and these are summarized in Table 1. The effects vary depending on the type of oestrogen or progestogen and the route of administration. In general, hormone replacement produces a lowering of low-density lipoprotein cholesterol and an increase in high-density lipoprotein cholesterol, thus reversing the changes in lipids and lipoproteins brought about by the menopause. An improvement in glucose tolerance, due to enhancement of insulin secretion and elimination or a reduction in insulin resistance, may be seen. There are also direct effects of oestrogen on arteries, which improve their function by endothelium-dependent and non-endothelium-dependent mechanisms. Hormone replacement therapy should therefore be considered for use in women with increased cardiovascular risk, such as those with established coronary heart disease, diabetics, hypertensives, and cigarette smokers.

Population studies have shown that a reduction in the incidence of cardiovascular disease of around 50 per cent can be achieved with hormone replacement. However, a recent randomized prospective study of hormone replacement therapy for the secondary prevention of coronary heart disease failed to show any overall benefit in outcomes, although an eventual reduction in events by over one-third was observed by the end of the trial. Concerns have been raised about the reliability of these findings, due in part to certain aspects of the trial design, and further studies are awaited to clarify the position.

Therapeutic regimens

Hormone replacement therapy consists of oestrogen, which should be given continuously, with the addition of cyclical progestogen in women who have not had a hysterectomy. Progestogens are necessary to prevent endometrial hyperplasia and neoplasia, and to regulate any uterine bleeding that may occur. Oestrogen is given as oral estradiol 17b, estrone sulphate, or conjugated equine oestrogens. Alternatively, estradiol 17b can be administered transdermally through adhesive skin patches, or implanted subcutaneously as pellets. The synthetic alkylated oestrogens, such as ethinylestradiol, are not used in hormone replacement therapy because of their potency and unwanted side-effects. The progestogens used are either derivatives of 19 nortestosterone, such as norgestrel and norethisterone, or the less androgenic C-21 steroids, such as dydrogesterone and medroxyprogesterone acetate. Natural progesterone can be used but is often poorly tolerated because of drowsiness. Progestogens are usually given in the minimal dose necessary for endometrial protection for 12 or more days per month, and result in a regular uterine bleed. The usual doses of hormones are shown in Table 2.

The main drawback to current hormone replacement therapy regimens is the necessity of uterine withdrawal bleeding, although this is often fairly light, particularly in older patients, and tends to diminish with time. With a satisfactory and regular bleeding pattern, there is usually no need for endometrial screening. However, cyclical bleeding becomes less acceptable as women get older, and thus regimens that avoid such bleeding become preferable. Preparations giving continuous progestogen with continuous oestrogen are used to induce endometrial atrophy and hence abolish uterine bleeding, resulting in amenorrhoea in up to 70 to 80 per cent of women.

These therapies are less successful in younger women, where transient episodes of spontaneous ovarian activity may result in irregular bleeding.

Tibolone, a synthetic compound with oestrogenic, progestogenic, and androgenic properties, is an alternative that avoids cyclical bleeding. It relieves vasomotor symptoms and appears as effective as hormone replacement therapy for the prevention and treatment of osteoporosis, but it is not established whether it has other benefits associated with hormone replacement therapy such as desirable cardiovascular effects or effects on the central nervous system. Similarly, it is not known whether it has the same potential risks, such as for the breast.

Raloxifene is a so-called selective estradiol receptor modulator, a synthetic compound which binds to the oestrogen receptor but causes conformational changes that result in different tissue-specific actions. Thus it can act similarly to an oestrogen in the skeleton, preventing osteoporotic vertebral fractures, but like an anti-oestrogen in the breast, causing a reduction in incidence of breast cancer, at least in the short term. It does not cause uterine bleeding, but does not relieve vasomotor or genitourinary symptoms. Studies are awaited to determine what, if any, are its actions on the cardiovascular and central nervous systems.

Side-effects of hormone replacement therapy

Oestrogenic side-effects such as breast tenderness and nausea are sometimes experienced on commencing therapy, particularly by older patients who are many years postmenopause. These side-effects are transient and usually resolve after about 3 months of therapy. More commonly, side-effects are due to the progestogen and can include breast tenderness, abdominal and pelvic pain, backache, depression, irritability, and migraine.

Risks of hormone replacement therapy

The main concern about hormone replacement therapy, particularly with prolonged treatment, is the risk of breast cancer. Epidemiological evidence remains conflicting: whilst some studies show no overall increased risk of breast cancer, others show an increase with prolonged usage. However, in studies looking at mortality from breast cancer, women taking hormone replacement therapy who develop the disease appear to have a better survival than those not on treatment. It seems prudent to avoid hormone replacement therapy where possible in women with breast cancer, but the disease need not be considered a total contraindication in all cases.

Previous endometrial hyperplasia or neoplasia is not a contraindication, provided the disease has been eradicated. Similarly, endometriosis and uterine fibroids rarely cause a problem, although they may occasionally worsen. There is growing epidemiological evidence that hormone replacement therapy may result in a decrease of around 40 per cent in the incidence of colorectal cancer.

Despite previous beliefs, hormone replacement therapy does not cause hypertension, except as a rare idiosyncratic reaction. There is a small absolute increase in the risk of venous thromboembolism, although whether this is seen with non-oral low-dose therapy is not known. It is therefore prudent to exclude a pre-existing thrombophilia in patients with a relevant past or family history.

Many women gain weight after the menopause, most commonly due to excessive calorie intake, not as a result of taking hormone replacements. Weight gain may occasionally occur due to fluid retention, particularly associated with progestogen use, but increases in body fat are not caused by hormone replacement therapy, although there is a redistribution of body fat, with a reduction in the metabolically harmful central obesity.

Most of the other reputed adverse effects of hormone replacement therapy are unsubstantiated, and have largely arisen from an inappropriate extrapolation of data obtained with oral contraceptive use.

Hormone replacement therapy is a treatment with considerable benefits for many women. The choice of therapeutic agents should be tailored to suit the individual case. There are advantages and disadvantages of certain preparations and combinations, but overall the therapy used should be the one that the patient finds most acceptable. This will encourage compliance with long-term therapy, resulting in the greatest health benefits.

Further reading

Collaborative Group on Hormonal Factors in Breast Cancer (1997). Breast cancer and HRT: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *The Lancet* **350**, 1047–59.

Ginsburg J, Prelevic GM (2000). The place of tibolone in menopausal therapy. In: Studd JWW, ed. *The management of the menopause. The millennium review 2000*, pp 59–67. Parthenon Publishing Group, Carnforth.

Grodstein F, Newcomb PA, Stampfer MJ (1999). Postmenopausal hormone replacement therapy and the risk of colorectal cancer: a review and meta-analysis. *American Journal of Medicine* **106**, 574–82.

Henderson VW (1997). Estrogen, cognition, and a woman's risk of Alzheimer's disease. *American Journal of Medicine* **103**, 11S–18S.

Hulley S *et al*. (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Journal of the American Medical Association* **280**, 605–13.

Marsh MS, Stevenson JC (1997). Hormone replacement therapy and heart disease. In: Julian DG, Wenger NK, eds. *Women and heart disease*, pp 279–95. Martin Dunitz, London.

Oger E, Scarabin PY (1999). Assessment of the risk for venous thromboembolism among users of hormone replacement therapy. *Drugs Aging* **14**, 55–61.

Spencer CP, Stevenson JC (1997). Oestrogens and anti-oestrogens for the prevention and treatment of osteoporosis. In: Meunier P, ed. *Osteoporosis: diagnosis and management*, pp 111–22. Martin Dunitz, London.

Stevenson JC (1996). Metabolic effects of the menopause and oestrogen replacement. In: Barlow DH, ed. *Baillière's clinical obstetrics and gynaecology. The menopause: key issues*, pp 449–67. Ballière Tindall, London.

Stevenson JC. (1998). Various actions of oestrogens on the vascular system. *Maturitas* **30**, 5–9.

Willis DB *et al*. (1996). Estrogen replacement therapy and risk of fatal breast cancer in a prospective cohort of postmenopausal women in the United States. *Cancer Causes and Contro*l **7**, 449–57.

14.1 Introduction to gastroenterology

Graham Neale

Further reading

At the end of the nineteenth century clinical science emerged in French and German universities. The concepts generated there spread to the medical schools of North America and led to the development of academic medicine. In 1897 the American Gastroenterological Association was founded. However, clinicians were slow to apply the spirit of enquiry to their practice, and well into the twentieth century static electricity or bitter tonics were used to treat 'gastric neurasthenia'; sarsparilla and dandelion were used as 'biliary stimulants'; and colonic lavage was a popular treatment for 'autointoxication'. Few advances in the understanding of gastrointestinal physiology in the first half of the twentieth century had any direct impact on clinical practice. Before the Second World War clinicians with a special interest in abdominal disease remained general physicians, few of whom thought it worthwhile learning to use the semiflexible gastroscope. Oesophagoscopy was the province of otorhinological or thoracic surgeons; and sigmoidoscopy with rigid instruments was undertaken more often in operating theatres than in medical outpatient clinics.

After the Second World War the emergence of simple methods for obtaining biopsies of the liver and small intestine and the more rational management of inflammatory bowel disease led to the rapid development of gastroenterology as a specialty. The 1960s and 1970s brought major advances in the understanding of the pathophysiology of gastrointestinal disease which have laid the principal foundation for modern clinical practice (Table 1). This was the scientific growth period of clinical gastroenterology.

Then in the 1980s and 1990s the development of endoscopic techniques based on fibre optics revolutionized practice, taking us into the technological era of clinical gastroenterology. How long this will last is uncertain. It seems likely that gastroenterologists will become less involved in routine endoscopy. The simpler procedures will probably be undertaken by certified non-physician endoscopists (possibly with robotic instruments) and diagnostic endoscopic retrograde cholepancreatography and colonoscopy will be eliminated by improved scanning procedures. The speed of change may depend on the tenacity with which specialists in gastroenterology cling on to the simpler endoscopic techniques which often provide them with substantial earnings. However, ultimately just a few interventional endoscopists may remain, concentrating on 'high-tech' therapeutic techniques.

Moreover if straightforward curative treatments emerge for inflammatory bowel disease the role of the gastroenterologist will undoubtedly change. Already in the Western world more than half the referrals to specialists in gastroenterology are for advice on managing patients with an irritable bowel. Providers of funding for health care will get an increasing grip on cost efficiency. With better organization of medical knowledge and outcomes research, general practitioners should be able to provide the majority of routine care.

For the physician-gastroenterologist there will be a renewed emphasis on the need to develop advanced interpretative and diagnostic skills. These will underpin therapeutic decision-making and the counselling of patients—who are themselves now far better informed about health and disease. It is possible that some gastroenterologists will evolve into digestive health physicians working as members of teams involving surgeons, radiologists/scanners, pathologists, microbiologists, immunologists, and geneticists. These teams will deal with structural lesions by sophisticated minimally invasive procedures; they may control disease by new techniques of immune modification; and develop methods to correct genetic defects affecting the gastrointestinal tract and adjacent organs. They will depend on the material sciences, information technology, and new means of communication, and will be advising individuals about their health profiles. They will prescribe pharmaceuticals and work with dietitians to alter health risks and to produce health benefits. As a group, digestive health physicians will need to know a great deal about the role of nutrition in health and disease.

If these changes are to occur then educational policies will have to change. Most illnesses have multiple causes and management decisions will come to be made through sophisticated risk analysis. All clinicians should have a good background in clinical science in order to cope with the emerging information era. The sections on gastroenterology and clinical nutrition in the *Oxford Textbook of Medicine* provide such a background in an important segment of clinical practice.

Further reading

Booth CC (1985). What has technology done to gastroenterology? *Gut* **26**, 1088–94. An important warning that gastroenterologists should remain physicians and not become technologists.

Chen TS, Chen PS, eds (1995). *A history of gastroenterology*. Parthenon Publishing Group, New York. An interesting book of essays on key developments in gastroenterology.

14.1.1.1 Structure and function of the gut

D. G. Thompson

Introduction

This chapter provides a brief overview of the structure and function of the gastrointestinal tract (excluding the liver and pancreas). For more detail readers are referred to the Further reading list. Emphasis has been placed on those aspects of gastrointestinal anatomy and physiology which help an understanding of the nature of gastrointestinal symptoms and/or guide an approach to therapy.

Anatomy

Gross anatomy

The gastrointestinal tract is a hollow tube of approximately 5 to 6 m in length, stretching from the oral cavity to anal sphincter (Fig. 1). It is arbitrarily divided into a series of organs which serve different functions, and is joined to the liver and pancreas, the major organs of digestion.

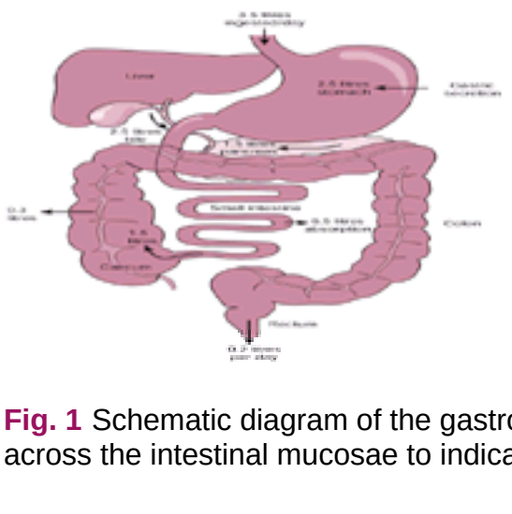

Fig. 1 Schematic diagram of the gastrointestinal tract showing the major organs of the tract and their connections. The figure also shows the average daily fluid flux across the intestinal mucosae to indicate sites and volumes of absorption and secretion in the various organs.

Anatomical structure

The gastrointestinal tract possess a broadly similar structure throughout its length (Fig. 2) with an innermost epithelium, a subepithelial lamina propria, and two muscle layers, an inner circular and an outer longitudinal layer, between which lies the myenteric plexus, the intrinsic neural control system of the musculature. While this description most accurately describes the small intestine, the other organs of the gastrointestinal tract differ only subtly from this stereotype.

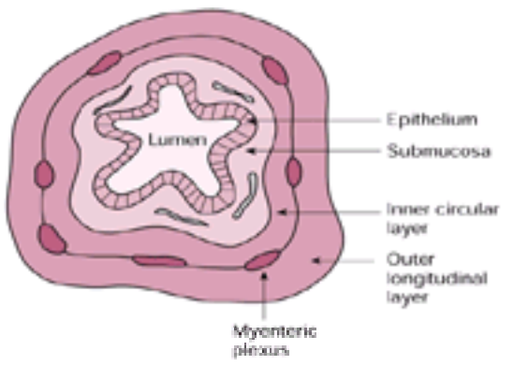

Fig. 2 A generalized structure of the intestine in cross section. A central lumen is bounded by an epithelial layer, which in turn is surrounded by a submucosal layer containing neural and vascular connections to the epithelium. Outside the submucosae lie circumferential and longitudinal muscular layers with the controlling neuronal myenteric plexus lying between.

Oesophagus

In the oesophagus, the innermost layer is a squamous rather than colomnar epithelium. The musculature in the upper third is striated and controlled directly via extrinsic neural pathways, unlike the lower two-thirds which has smooth muscle and a myenteric plexus.

Stomach

The anatomy of the stomach differs from the intestine, possessing an additional oblique muscular layer and at either end a sphincter—specialized musculature designed to act as a unidirectional valve to control the flow of luminal contents. The sphincter between the oesophagus and stomach (the lower oesophageal sphincter) lies at the level of the diaphragm. The sphincter between the stomach and small intestine is known as the pylorus.

Small intestine

The small intestine is arbitrarily divided into duodenum, jejunum, and ileum. The duodenum (so named because it is 12 fingers' breadth in length) is retroperitoneal, and possess on its medial aspect the ampulla of Vater which connects the pancreatic and common bile ducts to the duodenal lumen. The jejunum (Latin, empty, after death) is mobile and free on a mesentery. The ileum (Greek, twisted) begins indistinctly from the jejunum and ends at the caecum.

Colon and rectum

The colon differs from the small intestine in its muscular structure—the inner circular layer is similar but the outer longitudinal layer is condensed into three 'worm like' structures, the taeniae coli. At the proximal end of the colon, the caecum (Latin, blind ending) arises the vermiform appendix, named because of its worm like appearance. The ascending and descending colon are retroperitoneal whereas the transverse and sigmoid colon are freely mobile on a mesentery, extending to the pelvic floor following which it expands into the rectum.

Anal sphincter

The anal sphincter provides an important continence mechanism and has two parts, an internal sphincter of smooth muscle and an external sphincter of striated muscle.

Functional anatomy

The function of the gastrointestinal tract is closely associated with its structure.

The epithelial layer

The epithelium lies in contact with the luminal contents and ranges in permeability from being largely impermeable (oesophageal squamous epithelium) to highly permeable (intestinal epithelium). The absorptive function of the epithelial layer is modulated by a network of neurones, the submucous plexus, which receive input from the central nervous system. In addition, the neurones of the submucous plexus and the nerve terminals of extrinsic afferent nerves, particularly those running in the vagus trunk, are modulated by signals arising from the epithelium.

Neuromusculature of the gut

The striated muscle in the gastrointestinal tract (upper oesophagus and anus) is directly innervated by second order (lower motor) neurones (arising from the brainstem and spinal cord respectively) and therefore under direct central nervous system control whereas smooth muscle is largely autonomous, being controlled 'locally' by the enteric nervous system without direct innervation from the central nervous system. The central nervous system can, however, indirectly influence the muscular function of the gastrointestinal tract via its innervation of the myenteric plexus.

Immune system of the gastrointestinal tract

Throughout the gastrointestinal tract lie discreet clusters of immune cells which provide immunosurveillance and immune protection, that is Peyers patches in the small intestine and the appendix (see Chapter 14.4).

The function of the gastrointestinal tract

The function of the gastrointestinal tract is the transport, digestion, and elimination of ingested material to supply nutrients, vitamins, minerals, and electrolytes which are essential for life, together with the protection of the rest of the body from injurious or allergenic material.

Secretion/absorption

The gastrointestinal tract is responsible for movement of very large volumes across its lumen (Fig. 1). Overall, more than 8 litres enter the lumen per day. In contrast, only 200 to 300 ml is expelled per day as stool, the remainder being efficiently absorbed by the small intestine and proximal colon. The major digestive/absorptive organ of the gastrointestinal tract is the small intestine. Without the small intestine life is impossible whereas the possession of the small intestine without oesophagus, stomach, or colon is still compatible with reasonable nutrition. The various organs of the gastrointestinal tract subserve different functions to ensure that ingested nutrients are adequately digested or eliminated.

Oesophageal function

The oesophagus functions as a conduit to transport ingested food masticated by the mouth and salivary glands, through the thoracic cavity and into the proximal stomach.

Gastric function

The stomach acts as a storage, a sterilizing, and a digestive tank. Its receptive function enables large quantities of food to be eaten rapidly and stored and processed until adequately prepared for delivery to the small intestine. The presence of pathogens in food is reduced by the secretion of hydrochloric acid upon meal ingestion while the production of peptidases and lipase capable of operating in a low pH commence the process of digestion.

Small intestinal function

The small intestine is the major site of digestion and absorption. It regulates the speed of delivery of gastric contents via a sensing mechanism located in the epithelium, comprising endocrine cells sensitive to the pH, osmolarity, and chemical composition of the luminal contents, and signals both to intrinsic and to extrinsic neurones of the vagus to delay gastric emptying. This sensory signal also stimulates the delivery of bile and production of pancreatic secretion ensuring that these major digestive materials are delivered to the intestine only in the presence of nutrients.

The absorption of digesta is achieved through the intestinal mucosa. While some passes between the intestinal cells, most is actively transported through the epithelial cells via specific transporters (e.g. peptide, hexose transporters). The small intestinal is also a major absorptive organ, retrieving over 6 litres of fluid per day from the lumen (Fig. 1), the end result of which is the delivery of a small quantity of unabsorbed food (1.5 l) into the caecum.

Regional variation in intestinal absorption

The intestine shows regional differences in its absorptive function. The jejunum is responsible for the majority of nutrient and fluid absorption, whereas the ileum has additional, specific absorptive functions, in particular the absorption of vitamin B_{12} and the absorption of bile salts. Surgical resection of the ileum may thus be associated with development of B_{12} deficiency and of diarrhoea resulting from passage of bile salts into the colon where they induce secretion.

The colon

The colon's function is to salvage water and electrolyte from the small intestinal effluent, converting over a litre of material from the intestine into small pellets for elimination. In addition to its water and electrolyte absorptive function, the colon also salvages unabsorbed calories from the lumen, particularly undigested carbohydrate, for example starch polysaccharides. These are incompletely digested in the small intestine and thus pass to the colon where the anaerobic bacteria of the lumen ferment the carbohydrate to short chain fatty acids, which are absorbed to provide a secondary nutrient source.

The rectum

The rectum provides a storage function, enabling the elimination of colonic residue (defecation) to be restricted to times of personal convenience.

Neural control of gastrointestinal function

For the greater part of the time, the gastrointestinal tract is controlled by its own nervous system—the enteric nervous system. The enteric nervous system is not entirely autonomous however, and requires some local and central nervous system 'reflexes' for adequate co-ordination of functions along its length. For example the co-ordination of the passage of luminal contents into the small intestine from the stomach requires sophisticated control, which is provided by a vagally mediated reflex operating via the brainstem. This circuitry alters the gut from its fasted state to the fed state, that is gastric relaxation, the induction of gall bladder emptying, and pancreatic secretion, thus ensuring the provision of digestive enzymes at the appropriate time. An additional relay function is provided by prevertebral ganglia where visceral afferent neurones synapse with efferent relay neurones to integrate contractile patterns and contraction force.

The intrinsic nervous system

The intrinsic nervous system acts as a local control system with its own 'programmes', examples of which are the peristaltic reflex and the migrating motor complex.

Peristaltic reflex

This basic 'programme' responds to local luminal distension by an ascending muscular excitation pathway and descending inhibition (Fig. 3) which ensures aboral propulsion of luminal contents. This reflex is best seen in the oesophagus where it is known as secondary or non-swallow-related peristalsis. While the reflex can be induced in the small intestine or colon, it is not a major factor for luminal transit.

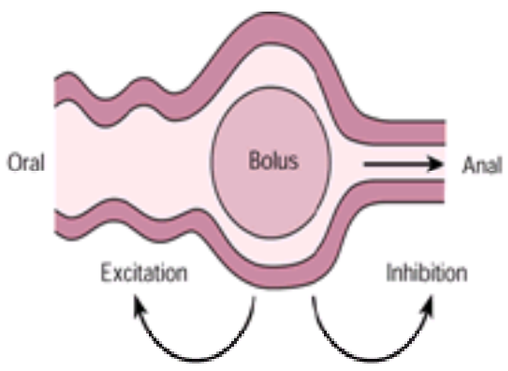

Fig. 3 Peristaltic reflex. The peristaltic reflex is the mechanical response of the intestine to the intraluminal distension. Note the presence of proximal motor excitation and distal inhibition which together propel the distending bolus from mouth to anus.

The migrating motor complex

This comprises a triphasic pattern of aborally propagating contractions in the distal stomach and small intestine during the fasted state which probably serve to maintain an empty lumen and reduce bacterial growth. Periods of quiescence are followed by irregular contractile activity which then terminates in a aboral migrating burst of regular contractions, which slowly migrate from the distal stomach down to the terminal ileum. This pattern, which characterizes the fasted state, is interrupted on food ingestion by a vagally mediated reflex which converts the pattern into a fed one.

The immune function of the gut

Being the major route of nutrient absorption, the gastrointestinal tract is also a potential portal for pathogen entry. The gastrointestinal tract therefore requires a sophisticated immune surveillance system together with a process for eliminating intestinal pathogens and the ability to either tolerate or eliminate ingested antigens. Details of this process are more fully dealt with in Chapter 14.4.

Disturbances of local physiological control mechanisms and origins of symptoms

Disturbances of local neuromuscular function are associated by the disturbances transit and elimination or secretion and absorption. An example of disturbed transit resulting from disturbed neuromuscular function is achalasia or slow transit constipation. Examples of disturbance of secretion and absorption are secretory diarrhoeas or the hyperacidity associated with *Helicobacter pylori* infection. Examples of symptoms which follow damage to extrinsic neural control are exemplified by the symptoms of truncal vagotomy, that is rapid transit and impaired nutrient–enzyme mixing result in poor digestion and an osmotic diarrhoea.

The relationships between gastrointestinal symptoms and the central nervous system are relevant to the understanding of functional gastrointestinal 'disorders'; it is well recognized that psychological disturbances, for example anxiety or depression combined with local disturbances of gastrointestinal physiology produce pain, nausea and vomiting, and altered bowel habit.

Further reading

Johnson LR, *et al*., eds (1981). *Physiology of the gastrointestinal tract*. Raven Press, New York.

Schultz SG, ed. (1991). The gastrointestinal system. In: *Handbook of physiology*, Section 6, Vols I–IV. American Physiological Society, Oxford University Press, New York.

14.1.1.2 Symptomatology of gastrointestinal disease

Graham Neale

The skilful analysis of symptoms indicating disorders of the digestive system is an integral part of the practice of internal medicine. Many patients with abdominal symptoms do not have easily defined organic conditions. The traditional skills of taking a careful history and examining the patient thoroughly are invaluable in managing patients who have functional disorders such as 'irritable bowel', non-ulcer dyspepsia, non-specific diarrhoea, recurrent abdominal pain, and somatization disorder. One might suppose that the enormous advances in endoscopic, scanning, and other techniques will have made clinical diagnosis less important, but this is not so: most gastrointestinal disorders are minor self-limited disorders of uncertain cause or are functional in nature, thereby often eluding definition by these procedures. Moreover the early suspicion of life-threatening disease and prompt referral of patients for investigation depends on clinical judgement.

Disorders of swallowing

Pharyngeal disease

Difficulty in swallowing is an important symptom that requires prompt resolution. Oropharyngeal disorders cause difficulty in initiating swallowing, regurgitation through the nasopharynx, a sensation of sticking in the throat, or the feeling of a lump in the throat on or after swallowing. Coughing and choking on swallowing is usually a symptom of pharyngeal disease and indicates failure to close the larynx. More rarely it is a sign of an obstructive lesion in the lower gullet which allows food and secretions to accumulate and spill into the larynx, especially at night. This symptom needs urgent attention to reduce the risk of aspiration pneumonia. Painful lesions of the oropharynx are usually demonstrated quite easily by simple inspection.

Patients with neurological disorders such as Parkinson's disease, motor neurone disease, myaesthenia gravis, and dermatomyositis only rarely present with disorders of swallowing and the clinician has only to be aware of the way in which known illnesses may affect the swallowing process. Oropharyngeal dyskinesia is common in the early phase of recovery after a stroke and may be persistently troublesome in patients with brainstem lesions.

Patients with the sensation of a lump in the throat on or after swallowing require an imaging technique to look for a pharyngeal pouch, a postcricoid web, or carcinoma and rarely pressure from a large osteophyte as a result of cervical spondylosis. It is unwise to submit patients with suspected pharyngeal lesions to conventional fibre-optic oesophagoscopy as a first examination because of the risks of perforation. Patients with a persistent feeling of 'a lump in the throat' without any demonstrable disease ('globus hystericus') usually respond well to the taking of a careful history, a single scanning procedure, and firm reassurance. They are more often women than men and they nearly always show signs of an anxiety state.

Oesophageal pathology

Dysphagia

Oesophageal dysphagia causes a sensation of food sticking in the gullet and is nearly always due to organic disease. The symptoms vary from discomfort to severe pain and the patient is rarely able to localize the site of the obstruction accurately. Associated symptoms such as regurgitation, vomiting, and coughing or choking are common. Oesophageal dysphagia is caused by obstructive lesions or more rarely by neuromuscular disorders. The common intrinsic lesions are inflammatory strictures secondary to reflux or tumours. Extrinsic compression may occur as a result of mediastinal lesions or vascular disorders (for example an aortic aneurysm). Neuromuscular disorders such as diffuse oesophageal spasm and dystrophia myotonica are much less common than obstructive pathology. The duration, progression, and frequency of symptoms help determine the likely nature of the pathology. Steady progression of dysphagia over a few weeks suggests malignant obstruction whereas association with a long history of heartburn suggests an inflammatory stricture. The dysphagia of neuromuscular disorders is usually confined to solids, and progresses, whereas that of achalasia is initially episodic and in the early stages often wrongly attributed to a functional disorder.

Heartburn

Heartburn is an extremely common symptom. It is an episodic lower retrosternal or epigastric burning that radiates upwards. It is caused by gastro-oesophageal reflux and commonly occurs an or hour or two after meals (especially if these are fatty or spicy); it may be precipitated by heavy physical work and bending. Symptoms often occur on lying down and are characteristically relieved by the ingestion of antacids. Most pregnant women suffer heartburn.

Oesophageal pain

Odynophagia is oesophageal pain felt within 15 s of swallowing and may be associated with the impaction of a lump of food at a site of mechanical blockage or a hold-up with oesophageal spasm. Odynophagia without hold-up occurs with intrinsic inflammatory disorders (such as reflux or candidal oesophagitis) and extrinsic disorders (such as mediastinitis). Hot liquids and alcohol may cause odynophagia in a normal gullet (the so-called 'tender oesophagus').

Oesophageal pain not clearly related to swallowing is characteristically retrosternal, often has a crushing quality, and may radiate to the jaw thereby mimicking cardiac pain. Patients with these symptoms are usually investigated for angina before being referred to a gastroenterologist: some will be shown to have reflux-associated chest pain or a primary disorder of motility (e.g. diffuse oesophageal spasm). High-amplitude contractions of the distal oesophagus are often discovered in patients with attacks of chest pain of uncertain cause and these are believed to be related to psychological stress.

Dyspepsia, nausea, and vomiting

Dyspepsia, nausea, and vomiting are extremely common symptoms which can be produced by a wide range of conditions from the most serious (such as end-stage neoplastic disease) to the most trivial (such as over-indulgence in food or alcohol). Patients may speak of indigestion (to describe any low-grade upper abdominal discomfort) and sickness (to describe either nausea or vomiting).

Dyspepsia

Dyspepsia is upper abdominal or lower chest discomfort or pain related to eating which may be described by the patient as a burning, a heaviness, or an aching and is often accompanied by other symptoms such as nausea, fullness in the upper abdomen, or belching. Although the symptoms of upper gastrointestinal disease are imprecise and non-specific, care in taking the clinical history will often facilitate making the correct diagnosis quickly and limit unnecessary investigation. All too often,

an over-stretched, relatively inexperienced clinician will spend a few minutes talking to the patient and then arrange a battery of blood tests, gastrointestinal endoscopy (and biopsy), ultrasonic examination of the abdomen, and some conventional radiology before telling the patient that he can find nothing wrong (and possibly implying that the patient is too ready to complain). Certain aspects of history-taking yield important clues:

- How clearcut is the patient's description of symptoms? Peptic ulcer often gives well-localized pain in the epigastrium. In about 40 per cent of patients with dyspepsia this comes on after a meal and wakes the patient at night. Attacks of central abdominal pain which cause the patient to double up may indicate gallstones although, if there is associated cholecystitis, the pain is more likely to be in the right upper quadrant (often radiating to the right shoulder).
- For how long has the patient had symptoms and how constant are they? A short history makes organic disease likely.
- Has the patient any associated diseases and what drugs are being taken? Many drugs cause upper abdominal symptoms especially aspirin and non-steroidal anti-inflammatory drugs.
- Has the patient lost significant weight? Are there associated symptoms? Vomiting suggests organic disease, alcoholism, or pregnancy.
- Has the patient any worries or anxieties that may be related to dyspepsia of recent onset (especially in women and in the young)? Sometimes patients will deliberately conceal their worries for fear that the clinician will too readily accept them as the cause of their symptoms.
- Details about dietary habits, smoking, and intake of alcohol should be obtained and it may be necessary not to take what the patient says at face value.

For the older patient (over 45 years) with dyspepsia of recent onset, gastroscopy is nearly always indicated in order to identify early gastric cancer. By the time the patient has the classical triad of symptoms of this disease—loss of appetite, loss of weight, and loss of strength—it is usually too late to achieve a surgical cure. Regrettably the early symptoms of gastric cancer are usually mild and non-specific.

The symptoms of patients known to have infection with *Heliobacter pylori* by serological testing are difficult to assess. Most infected patients are symptomless and most of those who have dyspeptic symptoms without ulceration are not cured by treatment with appropriate antibiotics.

Nausea

The term nausea should be restricted to the feeling of being about to vomit. Acute nausea is usually accompanied by hypersalivation. Nausea is caused by labyrinthine stimulation (as in motion sickness); distension of hollow viscera, or any severe somatic pain and by some drugs, especially opiates and those used in chemotherapy for malignant conditions.

Again the clinician has to define carefully what the patient means by nausea. It may be used to describe anorexia, an aversion to food, abdominal fullness, or a sinking feeling in the abdomen. In the absence of a recognizable cause persistent or frequent nausea without vomiting often proves to be psychologically determined.

Vomiting

Vomiting is the forceful ejection of gastric contents through the mouth by the co-ordinated contraction of abdominal and gastric muscles with relaxation of the lower oesophageal sphincter. Non-productive vomiting is called retching. Vomiting occurs with peptic ulceration, especially when there is delayed gastric emptying (pyloric stenosis), and with advanced gastric cancer. It occurs with disorders of the biliary tree (especially as a result of gallstones) and with acute pancreatitis (in which it is a prime symptom). It is an important symptom of intestinal obstruction, especially with lesions above the ileocaecal valve, and it may occur with any cause of peritoneal inflammation such as appendicitis. Metabolic causes of vomiting include diabetic ketoacidosis, hypoadrenalism, and uraemia. Drugs which cause vomiting include opiates, some antibiotics (for example erythromycin), and chemotherapeutic agents. Alcoholism, raised intracranial pressure, and pregnancy are important causes of early morning vomiting.

Effortless vomiting without a definable cause may be psychogenic. This is usually a disorder of young women many of whom have suffered psychological trauma (such as sexual abuse). It is not related to the vomiting of bulimia, a condition that is part of the anorexia nervosa syndrome (see Chapter 26.5.5). Rumination has to be distinguished from vomiting. Rumination is the repetitive regurgitation of gastric contents into the mouth after meals, the regurgitated material then being reswallowed. It is not associated with nausea, heartburn, or discomfort and often appears to be simply an acquired habit.

Abdominal pain

Pain in the 'acute abdomen'

Most patients with acute abdominal pain are promptly referred to a surgeon. However, this does not always happen and the end-result may be disastrous. Thus all clinicians should be able to assess a patient with acute abdominal pain. The site of the pain is usually helpful, but it is diffuse or atypical in at least 25 per cent of patients with acute gastrointestinal pathology. It is important to determine whether movement or coughing aggravate the pain as in appendicitis and generalized peritonitis (including perforated peptic ulcer and pancreatitis). Pain exacerbated by inspiration points to pathology in the upper abdomen (especially cholecystitis) or adjacent to the diaphragm. A detailed analysis of the type of pain is usually unhelpful but it is useful to know if it is intermittent, colicky, or constant. Some pain radiates in a characteristic manner—urological (loin-to-groin), gynaecological (to the back or thigh), and cholecystitis (to the shoulder tip). In contrast, the pain of appendicitis and diverticulitis does not radiate. The pain of intestinal obstruction is colicky and often associated with vomiting. Severe pain without physical signs and with normal routine tests (laboratory and simple radiological) raises the possibility of mesenteric vascular occlusion especially in patients over the age of 50 years or those known to be at risk of thromboembolic disease. Acute porphyria is the other condition to consider. The pain is diffuse and severe and most frequently occurs in the third decade with females being much more frequently affected than males. Tenderness is almost always absent but tachycardia and anxiety are prominent. During the attack the urine will contain excess †-aminolaevulinic acid and, usually, porphobilinogen.

Upper abdominal pain

Pain in the upper abdomen has been considered under the heading dyspepsia. Upper abdominal discomfort is so common that its presence alone is of no value in distinguishing between those patients with organic disease and those with a functional disorder. Moreover patients with an irritable bowel, diverticular disease, and occasionally with other colonic pathology may have discomfort in the centre of the abdomen which they may describe as indigestion. In such circumstances there is usually also a history of some change in bowel habit that will indicate the true nature of the condition. Disease in the small intestine is proportionately rather uncommon and is frequently misdiagnosed as an irritable bowel. Symptoms are rarely specific but should be reinterpreted in the light of screening investigations (such as the blood count, a straight radiograph of the abdomen, and assessment of serum markers of inflammatory disease).

Lower abdominal pain

With lower abdominal pain analysis of symptoms often does not help to determine the cause. Indeed inflammatory and neoplastic colonic disorders often may not give rise to pain and when pain does occur, it is usually diffuse and central. However, focal pain and tenderness in the left iliac fossa often indicates diverticulitis and, more rarely, colonic cancer. Focal pain in the right iliac fossa may be a marker of Crohn's ileocaecal disease. The passage of stool or flatus will often relieve the pain of colonic disease and an irritable bowel; in contrast it tends to exacerbate the pain of local rectal conditions. A history of recent-onset lower abdominal pain in patients over the age of 40 years is an indication for prompt investigation.

Proctalgia fugax is a very painful paroxysmal perineal pain occurring unexpectedly often at night and may last for up to half an hour. The pathogenesis is uncertain and it is unassociated with any signs of disease. The condition is often recurrent but is self-limiting.

Diarrhoea and constipation

The general public knows well what is meant by 'an attack of diarrhoea' or 'being constipated' but these conditions are not easy to define in medical terms. First one must recognize that to a lay person, constipation means passing a stool less often than normal and usually with difficulty. But 'normal' may be anything from two or three times a day to two or three times a week. Moreover the mass and consistency of a normal stool varies considerably depending on diet, gender, and individual factors. Subjects eating a Western diet pass 80 to 200 g of stool each day with more women at the lower end of this range. To add to the difficulty of assessing bowel function, the frequency of bowel habit may bear no relation to the volume of faecal material passed nor to the amount of stool in the colon. Thus care must be taken to define exactly what is happening and it is advisable to take the common-sense view that a change in normal bowel habit is significant. Two warnings are appropriate:

first, most people are loath to mention incontinence (they will usually say that they have diarrhoea and will have to be asked specifically about soiling); secondly in those with persistent unexplained diarrhoea, it is best to examine the stool because patients' descriptions of their faeces are often uninformative.

Diarrhoea

The clinician needs to understand some basic gastrointestinal physiology in order to understand the mechanisms of diarrhoea and to make sense of a patient's symptoms (Table 1). Most cases of diarrhoea can be diagnosed quite simply from the history, an examination of the stools, and, when appropriate, direct examination of the rectal mucosa. Acute infective diarrhoea is recognized by its recent onset, sometimes preceded by nausea or vomiting and a general systemic upset. Abdominal pain often occurs with *Campylobacter* infection, and although passage of blood and mucus can occur with any severe infection, it is more common with *Shigellosis* and infection with *E. coli* 0157.

In assessing chronic or recurrent diarrhoea, it is worth trying to distinguish diarrhoea of large bowel origin from that of the small intestine. Large bowel diarrhoea characteristically occurs on rising, may be associated with pain which is relieved by defaecation, and often contains mucus and sometimes red blood. The differential diagnosis usually rests between inflammation and neoplasm. Small bowel diarrhoea occurs at any time, and although often watery, the stool may also contain excess fat. Steatorrhoea occurs in coeliac disease, pancreatic insufficiency, stagnant loop syndromes, and massive intestinal resection. Drugs (such as beta-blockers, diuretics, antacids, and antibiotics) as well as excess intake of beer may also cause diarrhoea.

In assessing diarrhoea that otherwise remains unexplained, it may be worth distinguishing between osmotic and secretory mechanisms. This is done by measuring the concentration of the major cations (sodium and potassium) and stool osmolality. If the measured osmolality is little more than the sum of the cations multiplied by two (to allow for the unmeasured anions), then the patient has a secretory diarrhoea. In contrast, if there is a significant osmolar gap then there must be another solute in the stool. Unfortunately this test may be unreliable if the volume of stool is less than 500 to 600 ml per day.

Constipation

Each year about 1 per cent of the population consult their family doctor complaining of constipation. The mode of presentation ranges from the acute onset of colonic obstruction to a lifelong disability. The most common causes of constipation are shown in Table 2. In taking a clinical history it is important to determine exactly what the patient means by constipation. A relatively sudden change in bowel habit without any significant change in dietary habit or medication suggests an organic disorder. A desire to defaecate, especially if associated with colicky discomfort, suggests organic obstruction. Constipation from a young age increasing slowly with time indicates a disorder of normal colonic function, a condition that is much more common in women than men. Colon physiologists distinguish between 'slow transit' and 'outlet obstruction' constipation (Fig. 1) but this distinction is of limited value in management. 'Low fibre' diets (as ingested by many young people living away from the family home for the first time), some drugs (especially opiates and drugs with anticholinergic activity such as antidepressants), metabolic disorders, and neurological disease must be considered—although in most cases no cause can be established. In women it is also necessary to consider the obstetric history and to consider the possibility of pelvic floor dysfunction. Straining at stool over a prolonged period may lead to rectal prolapse and incontinence and it may be necessary to ask the patient specifically about such symptoms.

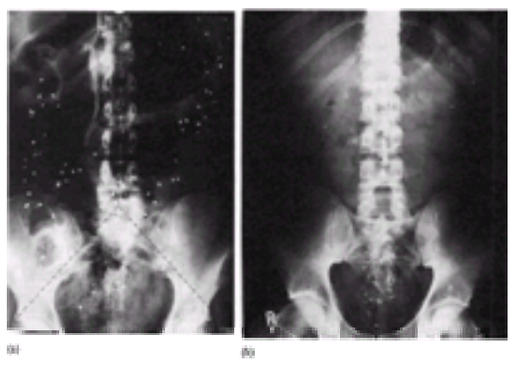

Fig. 1 Functional constipation illustrated by the use of markers (patient takes one marker capsule daily for 14 days, each capsule containing 10 radio-opaque pellets). (a) Abdominal radiograph of a 58-year-old woman with 'slow transit' constipation (colonic inertia). Seventy three pellets were retained after 14 days equally distributed between right and left colons (mean transit time 7.3 days, median normal for women 3.0 days). (b) Abdominal radiograph of a 30-year-old man who, a year previously, had an emergency laminectomy for an acute prolapsed disc. He complained of constipation since the operation and was shown to have 'outlet obstruction' presumably due to impaired rectoanal reflexes. Fifty seven pellets were retained with nine in the right colon, nine in the left colon, and 39 in the rectosigmoid segment (mean transit time 5.7 days, median normal for men 2.3 days). (By courtesy of Dr J. H. Cummings, MRC Microbiology and Gut Biology Group, University of Dundee. Figure 1(a) has been published previously in *Diseases of the gut and pancreas* ed. Misiewicz JJ, Pounder RE, and Venables CW, and is reprinted with the permission of Blackwell Scientific Publications.)

Further reading

Neale G (1998). Reducing risks in gastroenterological practice. *Gut* **42**, 139–42. Describes common errors in diagnosis and the importance of taking a careful history.

Snape WT Jr, ed (1996) *Consultations in gastroenterology*, section I, pp 1–155. WB Saunders, Philadelphia. A clinical approach to the diagnosis of nutritional and gastroenterological disorders.

Yamada T, ed (1999) *Textbook of gastroenterology*, pp 637–936. Lippincott, Williams and Wilkins, Philadelphia. An exceptionally comprehensive account of how to assess the symptoms of all patients with gastrointestinal disease.

14.2.1 Colonoscopy and flexible sigmoidoscopy

Christopher B. Williams and Brian P. Saunders

Development

Colonoscopes are similar to gastroscopes, but with greater flexibility to pass loops in the bowel. Early instruments had limited angulation and relatively clumsy characteristics. Colonoscopy was therefore initially looked on as a second-line procedure, time-consuming and potentially traumatic. Polypectomy was introduced in 1972 and this, coupled with recognition of the accuracy of the technique and (with biopsies) in diagnosing or excluding inflammatory disease or neoplasm, started an explosion of demand. Development of video-colonoscopes in the 1980s added high quality videotaping and prints which, with improved and more agile instruments, has contributed to making colonoscopy acceptable both for routine clinical use and for well-person surveillance and screening.

Colonoscopes range from 60 to 70 cm flexible sigmoidoscopes, thin very flexible paediatric instruments also used in adults with fixation or stricturing, up to 165 cm long colonoscopes with different flexibility characteristics and instrumentation channel sizes. Further technical improvements include higher resolution and zoom-magnification, better control ergonomics, adjustable shaft flexibility and electronic means of imaging shaft loops on-screen without fluoroscopy. Ultrasound colonoscopes, thin ultrasound probes for use with conventional instruments and magnetic resonance-compatible scopes are also being evaluated.

A large range of accessories can be introduced through the suction/instrumentation channel of a colonoscope, including biopsy or grasping forceps, washing, spraying or deflation tubes, cytology brushes, and injection needles. Therapeutic accessories include insulated forceps, polypectomy snares and retrieval devices, coagulating probes and argon plasma coagulating catheters, laser light guides, haemostatic clip and nylon loop applicators, dilating balloons, and metal stent introducers. It is also possible to preload onto the tip means of applying constricting loops or bands. Use of an external 'overtube' allows a looping instrument to be stiffened or exchanged for another without having to recommence the procedure. CO_2 insufflation apparatus is available which, since the gas is exhaled within 10 to 15 min, ensures that patients are not left distended after the procedure.

Cleaning and disinfection

As for all flexible endoscopes, skilled maintenance, regular checks, and meticulous cleaning are essential. All parts, including air, water, and instrumentation channels must be accessed during cleaning. It is not possible to sterilize a colonoscope but, after scrupulous mechanical cleaning and 'high level disinfection' (usually in an automated washing machine with 2 per cent glutaraldehyde solution and subsequent rinse cycles), all viral agents (including HIV and hepatitis B) and bacteria are inactivated. It takes at least 30 min to clean and disinfect an instrument so, allowing for repairs, three or more colonoscopes are needed to provide a routine service. Mycobacterial spores require even longer (60 min) exposure; prolonged disinfection is therefore recommended before and after examination of AIDS patients, not only because of patient susceptibility to infection but also their increased likelihood of harbouring mycobacteria. Accessories may be disposable or require autoclave sterilization or high-level disinfection, as directed by the manufacturers.

Patient preparation

Flexible sigmoidoscopy preparation is normally by disposable enema (hypertonic phosphate or similar) given or self-administered 10 to 15 min before the procedure. Some patients prefer to avoid the indignity of an enema by taking full oral preparation and this is also advisable in any patient with a colon narrowed by pronounced diverticulosis or stricturing.

Full bowel preparation is usually the most unpleasant part of colonoscopy. Oral preparation must be preceded by dietary restrictions, which include stopping iron or constipating agents in the preceding days, and taking low fibre diet (with no nuts, mushrooms, or iron-containing red wine) for 24 h.

Purgative preparation is cheap and generally best tolerated, but some individuals can suffer considerable cramping, explosive incontinence, or vasovagal reaction. A senna preparation on the afternoon or evening before exam is followed 2 to 3 h later by an osmotic purge (commonly magnesium citrate, sodium phosphate, or mannitol 10 per cent). The last dose should be taken only a few hours before the procedure to avoid the solidifying action of the proximal colon. Large quantities of clear fluids (including alcohol in moderation) are encouraged up to the time of examination to avoid adherent residue, which is difficult to flush or aspirate.

Fluid overload is the alternative approach, achieved by ingestion of 3 to 4 l of isotonic solution, usually polyethylene glycol (PEG)–electrolyte solution which avoids the electrolyte losses that occur if normal saline or 5 per cent mannitol are used. This approach is ideal for patients with active inflammatory bowel disease. Commercial PEG–electrolyte preparations can be prepacked for postage and are flavoured. Around 10 per cent of patients become nauseated, vomit, or become distended and stop drinking, any of which results in poor preparation, especially in the proximal colon.

If there is any possibility that the patient is obstructed, full oral preparation is contraindicated because of the possibility of perforation; a smaller volume should be given, supplemented by enemas. In the presence of massive bleeding it may be preferable to proceed direct to colonoscopy and rely on the purgative effect of blood, rather than waste time on preparation. In extreme cases of overwhelming blood loss peroperative 'caecostomy lavage' has been described, but nasal-tube lavage is a more usual compromise before emergency colonoscopy.

Psychological preparation of the patient should not be forgotten, since most of those scheduled for colonoscopy are apprehensive, whether through embarrassment, expected discomfort, or fear of colorectal cancer. Explanatory literature and a friendly telephone manner at the time of booking the exam can help a great deal. Equally, a warm and reassuring atmosphere on reception, whilst obtaining 'informed consent' and also during the procedure, can help transform colonoscopy from an ordeal into a reasonable and well-tolerated experience.

Medication

'Conscious sedation' is offered to most patients, unless they are likely to be easy to examine (flexible sigmoidoscopy, with a previously easy examination, or sigmoid resection or stoma) or are motivated to try without medication. The unpleasant gnawing quality of 'visceral pain' caused by the inevitable stretching of colon or attachments during insertion are tolerable for a few spells of 20 to 30 s. Prolonged examinations or the lowered pain threshold of patients presenting with features of irritable bowel syndrome or diverticulosis can be made considerably more pleasant with a minimal dose of a sedative–analgesic combination. Combing a low dose of benzodiazepine (typically, midazolam 2–3 mg, intravenous), with an opiate analgesic (pethidine 25–50 mg, intravenous) reduces discomfort and anxiety and gives the patient a well-deserved feeling of euphoria. Low-level sedation of this kind does not inhibit conversation, the ability to complain of pain or to change position when necessary. Smaller doses should be given in small, elderly, and sick patients but incremental larger doses (especially opiate) can be needed in apprehensive, younger ones. Pulse oximetry monitoring is routine and nasal oxygen, resuscitation equipment, and reversal agents (flumazenil, naloxone) should be immediately accessible. Sedated patients must be accompanied home. On-demand, self-administered nitrous oxide/oxygen inhalation has been shown by some to be equally effective for skilled endoscopists, giving rapid recovery and allowing return home without escort.

General anaesthesia is rarely needed and generally best avoided, since in an unconscious patient position change is made more difficult and removal of the warning given by pain may tend to more aggressive technique. For the few patients with particular reasons for general anaesthesia and an expert encoscopist, propofol anaesthesia gives excellent results and also more rapid recovery than after high doses of conventional sedatives. In certain countries with enough anaesthetists

available (France, Australia) this approach is commonly employed, whereas in others (Germany, northern Italy, Japan, China) unsedated colonoscopy is routine.

Antispasmodics are decried by many, on the basis that they are thought to elongate the colon and make insertion more difficult. In randomized trial, we have shown this to be untrue and routinely use them (hysocine-*N*-butyl bromide) both in order to speed insertion and to optimize the view for greater accuracy.

Antibiotics are only given to those with immunosuppression or immunodepression, previous endocarditis, heart valve prosthesis, septal defects, recent vascular prosthesis, or ascites.

Flexible sigmoidoscopy

Flexible sigmoidoscopy is the kindest and most logical means of examining proximal to the rectosigmoid junction (15 cm) whereas the rectum and the anal canal are better seen with the appropriate rigid instrument. Paradoxically the sigmoid colon, especially the sigmoid-descending colon junction, is the most difficult part of colonoscopy. In the presence of severe diverticular disease it may be impossible to reach even mid-sigmoid without expertise and a thin endoscope; after hysterectomy it may be cruel to examine without sedation. For this reason, although flexible sigmoidoscopy is both better tolerated and more accurate and effective than aggressive, rigid proctosigmoidoscopy, depth of insertion should be limited to what is tolerable by the individual patient. Some endoscopists mistakenly attempt to 'reach the splenic flexure' routinely. This is potentially unkind and also a mistaken concept. Without fluoroscopy or the use of newer means of imaging, even expert endoscopists can be completely mistaken between sigmoid-descending junction and splenic flexure. At 60 cm of insertion the 'scope tip can be anywhere between mid-sigmoid and caecum, and there are no positive localizing landmarks.

Insertion of the 'scope tip usually follows digital lubrication with jelly, the blunt tip of the instrument being gently inserted as the sphincters relax. Thereafter the 'scope is coaxed in as gently as possible, without haste or force, steering and 'cork-screwing' around bends with twisting movements. Blind insertion with 'red-out', guesswork, or 'push through' are all avoided as far as possible. Any small polyps (up to 5 mm) that may be adenomas are normally snared or destroyed at once, as they may be difficult to see on withdrawal or if left for subsequent colonoscopy. Because of the remote possibility of explosive gas concentrations after limited preparation, either repeated suction with air reinflation or use of CO_2 should precede electrosurgery; alternatively 'cold snaring' with physical removal of the polyp can be employed. If in doubt a biopsy can be taken, both to give some idea of the size of any lesion against the open forceps and to give partial histology.

Total colonoscopy

In expert hands, and in the absence of obstruction, a severely ulcerated colon, or other contraindication, total colonoscopy is possible in over 99 per cent of cases, with little sedation or suffering and virtually no complications. In less expert hands 'total colonoscopy' or 'completion' rates as low as 33 to 45 per cent have been reported and 75 per cent is common. The principle difference in technique between expert and inexpert is the ability, whilst keeping sufficient orientation for steering purposes, to pull-back and crumple the looped segment of colon already traversed, whilst simultaneously straightening the way or bend ahead. The ideal is to keep the colonoscope as straight as possible and to pleat or 'concertina' the colon over it, avoiding the unnecessary loops and pain by caused by pushing too hard or too long. The ideal is not always immediately achievable, so patience and determination—tempered by humanity—are essential qualities for the colonoscopist.

Paradoxically a freely mobile colon, without the conventionally fixed segments in the descending and ascending parts, can be as difficult to traverse as one with adhesions. This is principally because atypical loops may form, sometimes uncontrollable until the instrument tip eventually reaches a fixed point, giving a 'hold' and allowing the shaft to be straightened back. Happily, this type of long and mobile colon, whilst being a nightmare for the endoscopist, typically also has long enough attachments that the patient experiences little discomfort.

From the point of view of the patient, colonoscope stretch discomfort is frequently felt as 'wind pain', rapidly relieved as soon as the causative loop can be straightened. True over-distension is easily removed by aspiration. Further distress may be produced by the unpleasant illusion of incontinence given when the body-warmed and lubricated shaft is withdrawn through the sensitive anal canal; it is kind to prewarn the patient of this phenomenon and to preserve decorum by aspirating any fluid found during insertion through the rectosigmoid.

Once the colonoscope has successfully passed into the descending colon and has been straightened back to remove sigmoid colon looping, it is likely that the rest of the insertion phase will be considerably easier. When the 'scope shaft is straight it feels responsive and free, as do the angling controls—whereas the more looped the colonoscope is the more 'snarled up' everything becomes, and the more the patient suffers. Avoiding looping and responding to pain are the basis for successful, kind, and safe insertion. It also minimizes instrument repair bills and maximizes accuracy and ease of targeting lesions, since a straight instrument handles better.

This practical philosophy underlies the reason for simple but effective 'tricks of the trade' such as position change. It is obvious that in left lateral position there will be pooling of any fluid in the left colon; the transverse colon will also tend to sag down and so make the splenic flexure more acute. It therefore follows that, at the splenic and any flexure where there is a poor view or difficulty in insertion, position change may improve matters (to supine or right lateral at the splenic, but back to left lateral again for hepatic flexure). Adding only the simple principles of pulling back as often as possible to straighten each loop before tackling the next, avoiding over-distension to keep the bowel reasonably deflated and supple, and trying assistant hand-pressure whenever an unavoidable loop may be accessible (sigmoid and transverse colon) and the 'art of colonoscopy' is explained. Insensitivity, impatience, aggression result in the endoscopist being too muscular and tense to handle the shaft and controls sensitively, and so cause needless looping, pain, failure, and complications.

Contraindications, risks, and limitations

There are few contraindications to colonoscopy. It is, however, a relatively strong vasovagal assault with potential for arrythmias and so is contraindicated for 2 to 3 months after myocardial infarction. The tip, shaft, and air pressure involved in insertion have potential to exacerbate any existing risk of perforation. Colonoscopy is thus contraindicated in the acute phase and 2 weeks after an episode of diverticulitis, and also in severe acute or deeply ulcerated colitis of any variety (ulcerative, Crohn's, ischaemic, or infective). Patients with acute localized or rebound tenderness of the abdomen, free air, or dilated colon on radiograph should not be submitted to colonoscopy without special reason, due consultation, and by an expert endoscopist—who may decide to abandon the procedure.

The risks of diagnostic colonoscopy, as implied above, are to a great extent related to the training, personality, and handskills of the endoscopist. Regrettably large-scale audit shows figures of around 1 perforation in 1500 examinations, although this can be balanced against avoidance of the considerably worse complications of surgery. Therapy inevitably increases the likelihood of complications, principally bleeding but occasionally perforation after polypectomy or dilatation. Perforation may be actual (needing surgery) or threatened as the 'postpolypectomy syndrome' (managed conservatively with rest and antibiotics). Immediate bleeding can occur in around 1 per cent of polypectomies but is usually easily stopped by submucosal adrenaline injection (5–20 ml of 1/10 000 dilution, which is safe because of portal drainage) or by local electrocoagulation, clipping, or nylon loop application. Delayed haemorrhage can occur for up to 10 to 14 days after removal or local coagulation of even small polyps; it is usually self-limiting, but can be substantial and require admission to hospital and transfusion. Aspirin has been implicated as a risk factor in some series and should ideally be stopped for a week before and after polypectomy. Anticoagulants and other antiplatelet agents should similarly be withdrawn if possible or special measures instituted.

The greatest causes of colonoscopy-related mortality (1 in 10 000 exams) are patients referred (not always correctly) for surgery following endoscopic complications and deaths directly due to over-sedation, usually in the elderly. There have been unnecessary deaths when an internist has persisted in conservative management without involving a surgeon in management of suspected perforation. Surgical fatalities or major morbidity have resulted in others found at operation to have only a point perforation which would clearly have sealed spontaneously. In managing suspected perforation, due consultation between endoscopist and an endoscopically-aware, preferably laparoscopy-oriented, surgeon is essential. The need to modify sedation is described above and the endoscopist should also not be too proud to abandon an examination which is proving unreasonable, rather than to 'flatten' the patient. CT colography ('virtual colonoscopy') can image the proximal colon during the same visit.

The major limitations of colonoscopy relate to the fact that it is dependent on hand-skills and that tortuous, angulated, and haustrated colonic anatomy results in some 'blind spots' for the endoscopist. Areas that the endoscopist sees are extremely accurately evaluated, with a resolution of under 1 mm. The percentage of mucosa unseen is uncertain but is probably around 10 to 15 per cent overall. The likelihood of larger and 'significant' lesions being missed is much lower than this because colonic neoplasms are usually protuberant. Paradoxically, pathology can be missed in the capacious distal rectum or the anal canal, which can be avoided by retroverting the endoscope and/or examining with a rigid proctoscope as well.

Indications

The indications for colonoscopy are wide and constantly expanding, and are likely to continue to do so until alternative and less invasive techniques ('virtual colonoscopy' or genetic tests) are perfected. Where there is a shortage of endoscopic personnel, skill, or facilities it is possible to reduce the load of 'total colonoscopy' either by cross-referring for radiographs or related imaging techniques. It is also possible to combine these with prior flexible sigmoidoscopy on the same visit on the basis that limited colonoscopy covers the highest yield area, which is also the most prone to 'misses' or over-diagnosis.

High-yield indications include patients with bleeding, anaemia, or occult blood loss. Persistent bleeding, especially if dark or 'mixed-in' with the stool, is of sinister import, although it can be due only to local mucosal traumatization in diverticular disease. Good clinicians may select out for sigmoidoscopy alone patients with obviously fresh bleeding on defecation or with spotting 'on the paper'. However, the presence of blood in a patient of 50 years or more (so at-risk for colorectal neoplasia) is increasingly used as an excuse for the reassurance of a whole colon screening examination. Of all patients with blood loss referred for colonoscopy, around 10 per cent will have a 'significant' lesion, either a neoplastic polyp of 1 cm diameter or greater or malignancy. Colonoscopy is considered the investigation of choice for major bleeding, being readily available and offering immediate therapy and a high degree of diagnostic accuracy. Angiodysplasia, small ectatic vascular lesions in the proximal colon of elderly subjects with bleeding or anaemia, are relatively rare but are an example of a condition easily diagnosed and treated by colonoscopy—but by virtually no other method.

Chronic diarrhoea or known inflammatory disease is accurately and easily assessed by endoscopy and biopsy. The terminal ileum can be accessed or biopsied in over 80 per cent of cases by an experienced endoscopist. Endoscopic differential diagnosis between the focal or 'apthoid' (mouth-like) ulcers with intervening normal mucosa in Crohn's disease and the generally reddened surface of ulcerative colitis is easy and definitive in around 90 per cent of cases. A few remain as 'indeterminate colitis' and differential diagnosis can be more difficult in severe or chronic cases. The possibility of infective colitis, including tuberculous or amoebic, must be borne in mind and extra specimens taken for microscopy and culture if in doubt. Biopsies typically show somewhat greater extent of inflammation than is visible to the eye, and so must be taken at intervals around the colon in any patient with bowel frequency, to exclude 'microscopic colitis' or the related phenomenon of 'collagenous colitis'. Chronic inflammatory disease affecting more than half the colon carries an increased long-term risk of cancer or mucosal dysplastic (precancerous) change, and so indicates annual or 2-yearly surveillance from 8 to 10 years after onset of symptoms. Ischaemic colitis, typically affecting a short segment around the splenic flexure, can show changes ranging from mild reddening to marked ulceration or even near gangrene.

Polyps of almost any size can be removed endoscopically, leaving for transanal proctological management only very large, sessile rectal polyps up to 12 cm from the anal verge. Around 5 to 10 per cent of polyps will contain focal, high-grade dysplasia or invasive carcinoma. Even 'malignant polyps' or 'polypoid cancers' having no adenoma present can be managed conservatively by endoscopy alone if complete removal is confirmed histologically by a margin of 1 mm between the limit of invasion and the plane of excision, and the tumour is also well or moderately differentiated. Placing one or more India ink tattoos near the polypectomy site gives a permanent marker for endoscopic follow-up, or localizes it if surgery or laparoscopy is indicated. Lasers have been used for ablation of the postpolypectomy remnants of sessile polyps, but the newer alternative technique of argon plasma coagulation is cheaper, easier, and safer.

Cancer prevention or surveillance colonoscopy gives a strong guarantee to the patient even when negative, both because of the accuracy of colonoscopy and the generally slow time-course of development of colonic neoplasms. Follow-up at 3 to 5-year intervals after polypectomy does yield further adenomas in 30 to 50 per cent of patients, especially those with three or more or large polyps on the initial examination. A large series has suggested that the incidence of colorectal cancer is reduced by polypectomy and follow-up. Patients at genetic risk merit colonoscopic surveillance, especially those with a first-degree relative with colorectal cancer under 45 years of age, two or more affected first-degree relatives, or those assessed genetically as belonging to a 'hereditary non-polyposis colon cancer' family with autosomal dominant risk (see Chapter 14.15). Follow-up, ablating minute or 'flat' adenomas, is scheduled at 1 to 5-year intervals according to perceived individual risk.

Abnormalities on other diagnostic methods are ideally checked colonoscopically. Abnormalities seen on scanning frequently turn out to be spurious, presumably faecal, and the majority of positive occult blood tests prove to be 'false' for neoplasm. Supposed 'strictures' on barium enema often prove to be due to spasm or uncomplicated sigmoid diverticulosis. Others, such as anastomotic strictures, typically after Crohn's resection, are usually easily and effectively balloon-dilated. Even patients with typical malignant 'apple-core' strictures should ideally have preoperative total colonoscopy to exclude other synchronous neoplasms, if necessary using a small-diameter instrument (sometimes a paediatric gastroscope). If this proves impossible, endoscopy should be rescheduled within 6 months after resection. Colonoscopy has effectively supplanted 'diagnostic laparotomy'; it also avoids the numerous resections previously performed in diverticular disease when radiographs could 'not exclude the possibility of malignancy'—which the endoscopist can in a few minutes.

There are low-yield indications for colonoscopy, where alternative investigations such as flexible sigmoidoscopy, barium enema, 'virtual colonoscopy', or other approaches may be justified. These include patients with simple constipation of long standing, bloating, left iliac fossa discomfort, or combinations of these symptoms suggesting 'irritable bowel syndrome'. Where the patient is young, the extra accuracy of 'one-off' colonoscopy may be justified. In elderly patients, the greater likelihood of diverticular disease and difficult (and so more hazardous) colonoscopy may be a further disincentive, although these patients find the distension and manoeuvres of barium enema more unpleasant than sedated colonoscopy.

Cost effectiveness and relationship to other techniques

Flexible endoscopy seems superficially expensive and is demanding of professional time. However, modern colonoscopes are surprisingly robust and, properly handled and maintained, will perform thousands of examinations without expensive repairs.

Newer teaching methods, including use of computer simulation to improve handskills, should help to improve standards and speed training of doctors in colonoscopy and nurse practitioners to undertake flexible sigmoidoscopy.

Sigmoidoscopy provides rapid, unsedated examination, so allowing 'one-stop' patient management and avoiding the need to return for prepared exam–whether colonoscopy, radiography, or scanning. Flexible endoscopy has the general advantage that it leaves only gas (air or CO_2) in the colon, and so can be followed by any other modality if complete, whereas barium takes several days to clear.

Total colonoscopy, its accuracy increasing to near microscopic levels with newer and more agile instruments, is likely to remain the diagnostic 'gold standard' for the foreseeable future. Other, newer methods such as 'virtual' colonoscopy and genetic screening, may have an invaluable screening and selection role, but are unlikely to provide definitive tissue diagnosis or therapy. Endoscopy, providing that the bowel is prepared appropriately, can follow immediately to check or treat patients with a positive scan. It is therefore possible that colonoscopy will, in future, relinquish some of its present front-line role, but it is highly likely that requirements for it will increase if population screening for the prevention of colorectal cancer becomes routine.

14.2.2 Upper gastrointestinal endoscopy

Adrian R. W. Hatfield

Background

Subsequent to the pioneering work on the transmission of light down flexible optic fibres by Professor Hopkins in 1954, 'gastro-cameras' were replaced by early, flexible, fibre optic endoscopes in the mid-1960s, which led to the development of gastrointestinal endoscopy as we now know it. A major disadvantage of fibre optic endoscopes has been the deterioration of the fibre bundle with time leading to poor images. The recent availability of cheaper, miniaturized colour chips has led to the development of video endoscopes, providing an excellent, clear view which does not deteriorate with the age of the endoscope. With appropriate improvements in software, the endoscopic video image can be magnified and modern instruments will zoom up to 25 × magnification and the mucosal detail can also be enhanced electronically so that small lesions of a few mm can be seen quite clearly. The modern video endoscope image can be instantly printed out and archived digitally on a computer system.

The external specifications and handling of the new video endoscopes are similar to their previous, fibre optic counterparts and thus the techniques for disinfection and endoscopy are similar for both ranges of equipment. The disadvantage of the modern equipment is that the video endoscopy system needs considerable hardware. In most instances a video monitor, light source, and processor are located in an endoscopy unit and not easily moved to a different location such as an operating theatre for emergency endoscopy. In the acutely bleeding patient, the presence of blood in the lumen of the gastrointestinal tract diminishes the efficiency of the video chip and the image obtained is often poor. In this situation, a conventional fibre optic endoscope will often give a much better view and will be more easily taken into the operating theatre for such an emergency.

Endoscopy units and disinfection techniques

It is now well recognized that the care of the instruments and other equipment, together with the important aspects of patient safety, are greatly improved by having a purpose built endoscopy unit staffed by experienced endoscopic nursing staff who are trained in handling and disinfecting endoscopes and patient safety during and after intravenous sedation.

Most endoscopy units have a purpose built disinfecting machine which can take single or multiple instruments and, after suitable mechanical cleaning, a disinfecting agent will be automatically pumped through the channels of the instrument for a given period of time and flushed out afterwards. The choice of disinfecting agent varies between units but the trend has been away from hazardous agents such as glutaraldehyde (Cidex) to less harmful agents, such as Nu-Cidex or Tristel, that do not need sophisticated extraction and ventilation.

For routine, simple diagnostic upper gastrointestinal endoscopy many patients are now routinely endoscoped without sedation, after local anaesthesia to the throat only. 'No sedation endoscopy' is suitable for busy units with long lists of day cases. However, large numbers of endoscopies, particularly in apprehensive or sick inpatients and those needing more complicated procedures, are still performed under intravenous sedation.

There are now clear guidelines, drawn up by the British Society of Gastroenterology, as to the practice of administering intravenous sedation for endoscopic procedures. Patients are now monitored with pulse oximetry and oxygen is given routinely to the ill or elderly and to other patients if oxygen saturation falls during the procedure. The precise choice of sedation will vary between units and will depend on the patient and the type of procedure performed, however diazemuls and midazolam remain the two most common sedative agents used, often combined with pethidine for more lengthy or invasive procedures. It is not uncommon to reverse the effect of the benzodiazepine sedation with flumazenil (Anexate) and any opiate sedation with naloxone (Narcan). On rare occasions general anaesthesia will need to be used for endoscopy, usually for children or adults with ventilatory problems.

Specific risk of infection with endoscopy

Patients with heart murmurs were routinely given antibiotics to cover endoscopic procedures in the past. It is now recommended that only patients with prosthetic valves need be given routine prophylactic antibiotic cover, with a single parenteral dose of a broad-spectrum penicillin before the procedure.

Current disinfecting agents and schedules will cope with hepatitis B and C and HIV infection. All endoscopic staff wear disposable gloves and the nurse nearest the patient's mouth will usually wear a visor to cover eyes, nose, and mouth, particularly with a patient of known infective risk. As there is no effective way of sterilizing an endoscope against prions (at present thought to be the transmissible agent in Creutzfeldt–Jakob disease), the current Department of Health Guidelines make it clear that all equipment used on patients with suspected Creutzfeldt–Jakob disease should be destroyed afterwards. Patients with suspected Creutzfeldt–Jakob disease are not therefore endoscoped and alternative ways of diagnosis or treatment are usually sought.

Diagnostic endoscopy in the gastrointestinal tract

Endoscopy has now become the investigation of choice in patients with retrosternal or upper abdominal symptoms where, previously, barium radiology would have been employed. The advantages of detecting grades of inflammation and erosive change, rather than radiologically obvious ulceration, are obvious. Equally, the ability to take samples from the gastrointestinal tract with brush cytology or biopsy greatly enhances the diagnosis, not just in differentiating between benign and malignant ulcers and strictures, but also in assessing degrees of inflammatory change and in detecting dysplasia, for example in Barrett's oesophagus.

Endoscopy will detect oesophageal varices in patients with liver disease at an early stage. In the symptomatic patient with liver disease and gastrointestinal bleeding, endoscopy is particularly important as the site of bleeding may be variable and the management very different. Bleeding oesophageal varices can be managed endoscopically in a number of therapeutic ways.

In the last 10 years, it has become routine to take gastric biopsies in patients with peptic problems to detect the presence of *Helicobacter pylori*. The routine use of a simple CLO test, where mucosal biopsies are inserted into a gelatine well containing a colouring agent that turns yellow to red in the presence of *Helicobacter* urease, will be satisfactory. In some patients, gastric biopsies are necessary in this situation for culturing the bacteria to ascertain sensitivity in patients with infection resistant to multiple eradication therapies. In younger patients where malignant disease is less of a concern, serum, faecal, or breath test analysis is an acceptable alternative to establishing *Helicobacter* infection and thus such patients could be treated initially without endoscopy and gastric biopsy.

Most gastric cancers in the United Kingdom are diagnosed when the patient is symptomatic and thus the finding of a mucosal cancer is rare. Most lesions are straightforward to diagnose endoscopically and biopsies are usually confirmatory. Cancers that infiltrate the wall of the stomach below the mucosa are difficult to diagnose endoscopically as endoscopic biopsies are usually quite superficial. In this situation a 'double punch' type technique is useful, where a second biopsy is taken from the deeper submucosa through the small defect of the first biopsy. 'Linitis plastica' is difficult to assess endoscopically, particularly where Buscopan may have been used routinely to inhibit peristalsis at the start of the endoscopy. In such patients, a barium meal may help in the diagnosis by showing the lack of gastric motility.

Small bowel endoscopy (enteroscopy)

For many years, routine upper gastrointestinal endoscopes were not of sufficient length to pass beyond the duodenojejunal flexure into the small bowel. Enteroscopes are now made that can be advanced under direct vision down the upper small intestine or, alternatively, a thinner endoscope is allowed to pass down the small bowel spontaneously and then the bowel lumen is visualized on withdrawal. Such endoscopic procedures are lengthy and difficult and will not necessarily view the entire small bowel. A more comprehensive view is sometimes obtained, particularly in the hunt for obscure bleeding lesions, by passing a standard upper gastrointestinal endoscope up and down the small intestine through small enterotomies at the time of laparotomy, with a surgeon concertinering the small bowel over the shaft of the endoscope. Recently a video capsule has been developed which, when swallowed, transmits a good view of the entire small bowel during transit through the gut.

The rather lengthy, tedious, and unpredictable techniques of small bowel biopsy using a Crosby capsule have been largely superseded by routine upper gastrointestinal endoscopy with biopsies from the distal duodenum. Such biopsies have been shown to be very representative of the upper jejunal mucosa. This technique is now used routinely in the diagnosis of coeliac disease.

Therapeutic endoscopy in the upper gastrointestinal tract

Over the last 20 years, a wide range of therapeutic manoeuvres have been developed for use in various situations in the upper gastrointestinal tract.

Gastrointestinal bleeding

Oesophageal varices can be injected through the mucosa with ethanolamine oleate under direct vision. Paravasal injection is best avoided as it can lead to secondary bleeding from mucosal ulceration and sometimes later oesophageal stricture formation. Endoscopic sclerotherapy can be repeated at weekly or monthly intervals until the varices have been obliterated. Bleeding gastric varices can also be injected but these are more difficult to obliterate. More recently, endoscopic banding techniques have been employed, both in the acutely bleeding patient and the chronic situation. Single or multiple bands could be put on varices in the oesophagus or, sometimes, in the fundus of the stomach. The addition of thrombin into gastric varices after banding may enhance successful eradication and reduce the risk of bleeding if the bands slip off too early.

Bleeding erosions and ulcers can be injected with dilute adrenaline (1:10 000). This may be satisfactorily in reducing bleeding in the short term and can always be repeated if necessary. Endoscopic laser therapy can be employed around a visible vessel in the base of an ulcer. A similar effect can be obtained by the use of multicontact diathermy probes or heater probes. Bleeding vascular abnormalities, such as angiodysplasia, can be treated with thermal probes but more satisfactorily with non-contact laser which does not pull off a coagulum and has the extra benefit of destroying vessels just below the mucosa.

Benign oesophageal strictures

Commonly, a peptic stricture above a hiatus hernia secondary to reflux will produce dysphagia but benign strictures due to other causes, such the swallowing of corrosive substances and postsurgical anastomotic strictures, can be treated by the same endoscopic techniques. In the past, bougies of increasing size were passed over a previously endoscopically placed guidewire and the stricture slowly dilated. More recently, high pressure dilating balloon catheters, passed over the wire under radiological screening or directly through the scope under direct endoscopic vision, can be used more efficiently and safely with better patient tolerance.

Achalasia of the cardia can be treated with balloon dilatation using a larger balloon of 30 to 40 mm diameter, where the aim is to rupture muscle fibres to weaken the circular muscle sphincter. Alternatively, botulinum toxin can be injected through the mucosa into the muscle sphincter circumferentially at the time of endoscopy. The improvement in swallowing after this procedure is limited and may need to be repeated every 6 months.

Malignant gastro-oesophageal strictures

Most of patients with non-operable tumours of the stomach or oesophagus producing dysphagia are palliated by the insertion of some sort of oesophageal stent. By and large, silicone rubber prostheses have been replaced by self-expanding metal stents which can be very easily and safely placed through a malignant stricture without the need for prior dilatation, thus reducing the risk of perforation. Most of these stents now have a membrane to prevent tumour ingrowth through the mesh but this will sometimes occur at one or either end. Such tumour overgrowth can be treated with endoscopic laser therapy. Brachytherapy can be given via an endoscopically sited tube through the stricture prior to stenting.

Removal of foreign objects

Most solid objects such as marbles, rings, and coins should pass spontaneously and the need for removing foreign bodies is usually because they are sharp and may cause damage if left *in situ*. Most objects can be snared or trapped in a basket and removed intact. Sharp objects can be pulled into a endoscopic overtube to protect the oesophagus during withdrawal.

Polyps and small tumours

Most gastric polyps are entirely benign and do not need removing. Leiomyomas of the stomach or duodenum can be watched if small, but if over 5 cm in size should probably be removed. Often such patients will go directly to surgery but newer endoscopic techniques using submucosal resection can tackle lesions that do not infiltrate beyond the submucosa. Careful prior assessment with endoscopic ultrasound is usually needed to make sure that a small tumour can be technically removed in this way.

Assisted nutrition

There are now many types of enteral feeding tube that can be sited in the upper gastrointestinal tract. Although most fine-bore feeding tubes can be passed on the ward or under radiological control, the prior passage of an endoscopic guidewire into the stomach which is then rerouted through the nose, can allow feeding tubes to be positioned accurately, often through an oesophageal stricture or difficult anastomosis, or positioned in the duodenum in patients with gastric stasis. The endoscopic positioning of a nasojejunal feeding tube, beyond the duodenojejunal flexure, is now becoming a common alternative to intravenous feeding in patients with complicated pancreatitis where 'pancreatic rest' is needed.

Techniques for placing a gastrostomy tube endoscopically (PEG) are now simple and straightforward. After transabdominal puncture into a distended stomach under direct endoscopic vision, a PEG tube with diameters from 8FG to 24FG can be pulled back down the oesophagus through the stomach and a flange, balloon, or button will allow the tube to be anchored firmly up against the gastric mucosa. In patients where there is gastric stasis or in pancreatitis, a small jejunal extension tube can be inserted through the PEG and positioned endoscopically into the distal duodenum or beyond the DJ flexure (PEJ).

Endoscopic ultrasound

Special endoscopes are available with a dual capability of endoscopic and ultrasound imaging. Either a rotating or a fixed linear array transducer will provide an ultrasound image at a point where the endoscopist can accurately direct the probe in the lumen of the oesophagus, stomach, or duodenum. Although CT scanning will stage most larger tumours of the upper gastrointestinal tract, pancreas, and bile duct, endoscopic ultrasound is particularly useful in staging small tumours and particularly mucosal tumours. The linear array ultrasound endoscope can be used for needle biopsy of tumours in the wall of the gastrointestinal tract or head of

pancreas and sometimes adjacent lymph nodes. Although endoscopic ultrasound is still developing in the United Kingdom, it is used more commonly elsewhere in Europe where gastroenterologists are routinely trained in abdominal ultrasound.

Endoscopy and disorders of the pancreas and biliary tree

Diagnostic endoscopic retrograde cholangiopancreatography

The development of side-viewing duodenoscopes in the 1970s allowed endoscopic visualization of the papilla of Vater and cannulation of the pancreatic and biliary duct systems, endoscopic retrograde cholangiopancreatography (ERCP). For many years ERCP was the gold standard of investigating pancreatic and biliary disorders but, with the advent of CT and MR scanning, the need for diagnostic ERCP has diminished. ERCP is still extremely useful in the diagnosis of patients with gallstones, sclerosing cholangitis, and biliary tumours where scanning is normal or equivocal, in the absence of overt jaundice. A tissue diagnosis can be obtained with brush cytology and endoscopic biopsy within the bile duct, avoiding the need for percutaneous biopsy.

Diagnostic ERCP is still useful in the assessment of patients with pancreatitis, congenital abnormalities, such as pancreas divisum, and in some patients with a pancreatic mass on scanning where the diagnosis is not clear. Most patients with a carcinoma of the pancreas will present with obstructive jaundice and will need a therapeutic procedure, others without jaundice will usually be diagnosed on ultrasound or CT scanning.

In specialized centres, biliary and pancreatic manometry is performed to assess patients with pancreatobiliary pain with no apparent structural abnormalities. At ERCP, a perfused catheter can be inserted into the bile duct and into the pancreatic duct and pull-through manometry performed. This will show whether elevated basal and peak pressures indicate sphincter of Oddi dysfunction.

Therapeutic endoscopic retrograde cholangiopancreatography

Gallstones

The endoscopic removal of common bile duct stones at the time of ERCP is the treatment of choice for patients presenting with pain, abnormal liver function tests, jaundice, or cholangitis. Following previous cholecystectomy, about 10 per cent of patients will ultimately represent with bile duct stones and endoscopic management is far safer than further surgical exploration of the bile duct. Prior to laparoscopic cholecystectomy, it is particularly important to investigate and to endoscopically clear the bile duct of stones, if suspected. Failure to do so may increase the likelihood of postoperative bile duct leaks. At the time of ERCP if stones are located in the biliary tree, a small diathermy cut is made into the bile duct through the papilla and, through the sphincterotomy, stones can be extracted with a balloon or basket. If the stones are too numerous or too large to extract at the first procedure, small pigtail stents are inserted into the bile duct to guarantee good drainage without stone impaction and therefore reduce the incidence of postprocedure cholangitis.

Most large stones can ultimately be removed using a mechanical crushing basket (lithotripter) or sometimes with the help of extracorporeal shock wave lithotripsy, following which fragments can be removed from the bile duct at follow-up ERCP. In experienced hands, the technical failure rate is low and thus the need for surgical reintervention is uncommon. Only in patients with very large bile duct stones, intrahepatic stones, or stones above biliary strictures is there a need for further procedures, such as intraduct choledochoscopy, using small endoscopes with direct contact lithotripsy using a pulse dye laser or an electrohydraulic probe. Very elderly or frail patients with large bile duct stones can be managed long-term by simple placement of an endoscopic stent beside the stones for drainage to prevent jaundice and/or cholangitis. Such stents can be changed over the years as per necessary.

Benign strictures

Postoperative anastamotic strictures or those following bile duct damage at the time of cholecystectomy can initially be managed with intermittent biliary balloon dilatation at the time of ERCP or simple endoscopic stent placement. In the young patient after a trial of dilatation or stenting for a reasonable length of time, about 1 year, surgical reconstruction of the bile duct might be needed if it is clear that endoscopic treatment is not leading to resolution of the stricture. In patients with primary sclerosing cholangitis, there may be single or multiple strictures in the intrahepatic and extrahepatic biliary tree, often in association with pigment stones, which can be difficult to dilate or stent. A variable proportion of patients with primary sclerosing cholangitis develop a cholangiocarcinoma and this can be very difficult to prove even with good ERCP, biliary cytology, CT, and MR scanning.

Malignant bile duct obstruction

Pancreatic and bile duct cancer and carcinoma of the ampulla of Vater can all produce stricturing of the biliary tree at different levels. At ERCP the stricture can be dilated and then an endoscopic 10 or 12FG polyethylene stent placed to relieve jaundice. These stents are cheap and usually stay patent for 4 to 5 months. In pancreatic cancer, about one-third of patients will survive long enough to occlude their stent, in which case a further procedure is performed to remove the blocked stent and replace it with a new one. Self-expanding metal stents offer a way of palliating patients for longer as they have a lumen of 10 mm which gives excellent long-term drainage. At present, biliary metal stents have an open mesh and tumour infiltration may occur, causing recurrent jaundice and/or sepsis. In that situation, a plastic stent can be inserted through the blocked metal stent to achieve drainage. Membrane-covered metal stents are now becoming available which should avoid the problem of tumour ingrowth and hopefully remain patent for longer.

In some patients with cholangiocarcinoma at the hilum of the liver, separate obstruction to right and left main ducts or subsegments may be found. In such a situation, more than one stent may be necessary to relieve jaundice or sepsis. Brachytherapy for cholangiocarcinoma can be administered endoscopically using an iridium wire source inserted down an endoscopically placed catheter inside the stent within the cholangiocarcinoma. Photodynamic therapy (PDT) can also be administered using a diffuser laser fibre, endoscopically sited within the malignant biliary stricture(s).

Pancreatitis

In patients with acute, relapsing and chronic pancreatitis a variety of endoscopic therapies can be performed. After pancreatic sphincterotomy, stones can be removed from the pancreatic duct, strictures can be stented, and drainage of the dorsal duct in pancreas divisum can be achieved. Peripancreatic fluid collections and pseudocysts can also be managed by pancreatic duct drainage or direct endoscopic cyst puncture and stenting techniques. Pancreatic endotherapy is difficult and can be associated with complications. Nevertheless, in selected patients it may be very valuable and avoids difficult and complex pancreatic surgery.

Gastric outlet obstruction

About 10 per cent of patients with pancreatobiliary tumours will develop gastric outlet obstruction as a late complication as tumour infiltrates the duodenum. Conventionally, a surgical gastric bypass has been unavoidable and this has carried a substantial morbidity/mortality as these patients are often very frail in the latter stages of their malignant disease. A large-diameter, self-expanding metal 'enteral' stent can now be placed in the stomach and duodenum at the time of endoscopy. This rapidly relieves symptoms of gastric outlet obstruction and allows the patients to eat a reasonable diet, without vomiting, thus avoiding the need for bypass surgery.

Hazards and complications

Diagnostic endoscopy carries very few risks. With careful attention to nursing techniques and sedation protocol, cardiovascular problems during endoscopy and aspiration pneumonia after are extremely rare. Direct damage to the upper gastrointestinal tract during insertion and subsequent inspection down to the duodenum is extremely unusual but rarely the cricopharynx, lower oesophagus above the cardia, and duodenal cap are sites of direct perforation with the endoscope, more commonly with inexperienced endoscopists. An unrecognized pharyngeal pouch represents a real hazard during insertion of the endoscope and might lead to a perforation if undue force is applied.

Most complications of endoscopy occur during therapeutic procedures and are specific to the type of procedure being performed.

The perforation rate following oesophageal dilatation is extremely low now that techniques and equipment have improved. The development of self-expanding stents

in the oesophagus avoids the need for forceful dilatation of malignant strictures and this has radically lowered the postprocedure complication rate of perforation. Due to the size of the balloon used in dilating achalasia of the cardia, perforations can be seen. Anybody who develops pain or discomfort after oesophageal dilatation should be assumed to have developed perforation, a chest radiograph should be obtained and if there is evidence of mediastinal air or surgical emphysema, conservative management with nil-by-mouth, parenteral antibiotics, and intravenous feeding is advocated. Many patients will settle conservatively without the need for surgical intervention.

The complications of ERCP are well known and more frequent than those of other endoscopic manoeuvres in the upper gastrointestinal tract. Even with diagnostic ERCP, up to 2 per cent of patients may develop postprocedure pancreatitis after either manipulation at the papilla or the injection of contrast into the pancreas. Such pancreatitis is usually self-limiting and mild although the serum amylase may reach extremely high levels. Some patients have a very elevated amylase without any pain. After endoscopic sphincterotomy and any therapeutic manoeuvre in the pancreatic or biliary tree, pancreatitis and bleeding can occur. Between 2 and 5 per cent of patients may have some degree of bleeding, but only a small proportion of these will need a blood transfusion or, rarely, surgical intervention. With the use of periprocedure antibiotics and the routine use of biliary stents after incomplete gallstone clearance within the bile duct, the incidence of postprocedure cholangitis is minimal.

14.2.3Radiology of the gastrointestinal tract

Alan Freeman

The widespread introduction of endoscopic techniques has lessened the need for radiological examination of the intestinal tract, and has completely replaced it in the examination of the stomach. There is, however, still a major radiological role in the investigation of the small and large bowel; the small bowel will be considered first.

The small intestine

The small intestine may be examined by a number of radiological means which include: plain films, barium contrast studies, ultrasound, CT, nuclear medicine, and MR. Plain film radiography performed in cases of suspected small bowel obstruction typically include films taken with both vertical and horizontal beam, the object being to demonstrate dilated loops and air–fluid levels. This role, however, is increasingly replaced by CT and apart from this, plain films have no other function. Barium studies of the small bowel have been the mainstay of examination for almost a century and, correctly performed, provide good morphological detail of the bowel. There are two types of examination; the first is the so-called barium follow through wherein the patient drinks a quantity of barium sulphate and then sequential films are taken of its passage through the small bowel. The second is the small bowel enema or enteroclysis and in this technique a tube is passed into the third part of the duodenum and barium sulphate is continuously infused, thus outlining the small intestine. This latter technique can also be used to provide a double contrast effect by infusing methyl cellulose, which provides a negative contrast against the positive contrast of barium sulphate.

Both techniques have advantages and disadvantages. The follow through is simple to perform and of course is more comfortable for the patient. However, it is imperative that fluoroscopy is performed at regular intervals together with abdominal compression so that all loops of small bowel are outlined. Enteroclysis by its very nature means that the barium column is observed in a continuous fashion and any minor obstruction or abnormality is thus more likely to be observed. Proponents of enteroclysis therefore argue that it is a more accurate test, but it is difficult to make a direct comparison as it is obviously difficult to perform both techniques on a cohort of patients. Given the relatively low incidence of disease of the small intestine, it seems reasonable in most instances to perform a follow through study in the first instance and to reserve enteroclysis for unresolved problems, subacute obstruction, etc.

Both types of contrast study suffer from the fact they only demonstrate the mucosal surface of the bowel and disease in the wall of the bowel or outside of it may be easily overlooked. In this situation, some form of cross-sectional imaging, typically ultrasound or CT, may well give more information.

Ultrasound, to date, has not found a huge role because it is highly operator dependent and requires considerable time. Its advantage of course is that it is radiation free. Usually, a high frequency ultrasound probe, in the order of 5 mHz, is used and the whole abdomen is carefully covered quadrant by quadrant with graded compression. Dilated, fluid-filled loops of bowel are relatively easy to examine, but the presence of excess gas within bowel loops, deep pelvic loops, and patient obesity all provide major impediments to full examination by ultrasound.

These factors, however, provide no problem for CT which, despite its use of ionizing radiation, is finding an increasing role in the investigation of small bowel disease. This particularly applies to acute small bowel obstruction where it provides not only confirmation of the diagnosis but often demonstrates the site and the nature of the obstructing lesion. As the images are usually enhanced by means of intravenous contrast agents, they can provide useful information as to the viability or otherwise of the obstructed loops. Thickening and infiltration of the bowel wall is readily appreciated at CT, as is the demonstration of exophytic tumours such as leiomyomas and lymphoma.

Nuclear medicine studies, which involve the injection of a radio-labelled substance, have major roles in the demonstration of inflammatory conditions involving the small bowel and for the demonstration of potential bleeding sources from the small bowel. In the former, the patients white blood cells are extracted and then labelled with Tc 99m pertechnetate. These white cells are then reinjected into the patient. The cells concentrate at the site of inflammation and this activity is demonstrated on a gamma camera. The technique is known as labelled white cell scanning. Alternatively, indium may be used as the isotope when a longer half-life is required and this particularly applies in the diagnosis of possible intra-abdominal abscess. This does involve a higher radiation dose to the patient.

If bleeding from the small bowel is the problem, then the patient's red blood cells are extracted and again labelled with Tc 99m pertechnetate. The labelled cells are reinjected into the patient and the abdomen is scanned under a gamma camera. A bleeding source of more than approximately 0.5 ml/min, can be identified as a hotspot, indicating the probable site. A more specific bleeding source is a Meckel's diverticulum and if this is symptomatic it is likely to contain ectopic gastric cells which may be demonstrated by the simple injection by pertechnetate by itself.

Magnetic resonance imaging (MR) of the small bowel holds huge promise and is likely to become the most important technique in the future. It has major advantages in that it involves no radiation and image reconstruction can be performed in almost any plane. Indeed, image reconstruction is so fast that in effect, MR fluoroscopy of the bowel can produce images similar to those obtained at barium fluoroscopy. Because of the lack of ionizing radiation, areas of interest or difficulty can be 'revisited' time and again until fully evaluated. To obtain these images, the bowel does need to be distended which usually involves passing a nasoduodenal tube and utilizing water or methyl cellulose as the distention and contrast agents.

Anatomy of the small bowel

Strictly speaking, the small bowel commences at the pylorus and includes duodenum, jejunum, and ileum. However, the first part of duodenum, proximal to the superior duodenal flexure, is usually more closely aligned to the stomach in relation to pathology and investigation and for practical purposes, this chapter will discuss small bowel from mid-descending duodenum onwards. The third part of the duodenum commences at the inferior duodenum flexure and passes across the midline at roughly the level of L3 to ascend to the duodenal–jejunal junction. This is marked by the peritoneal ligament of Treitz. At this point, the small bowel emerges from the retroperitoneum to become an intraperitoneal structure. It is then divided roughly into equal lengths of jejunum and ileum, with jejunal loops occupying the left upper quadrant and ileal loops the right lower quadrant. The ileum terminates at the ileocaecal junction where it again becomes a retroperitoneal structure. Morphologically, jejunum can be distinguished from ileum both by its size (it is usually about 1 cm larger in diameter) and by the presence of valvulae connivente or plicae semilunaris. The latter gives it its characteristic fold pattern on contrast studies as opposed to the relatively featureless ileum.

Pathology of the small bowel

Coeliac disease and malabsorption states

Coeliac disease, caused by a sensitivity to gluten, is characterized by atrophy of the small intestinal villous pattern. The definitive diagnosis thus rests with duodenal or jejunal biopsy, but radiology retains a role in evaluating complications as well as excluding other causes or malabsorption.

Barium follow through studies have been the mainstay and demonstrate changes which include dilatation of jejunal loops (to more than 3.5 cm), flocculation of the barium suspension, thickening of the fold pattern (particularly if there is associated hypoproteinaemia), and delayed transit. Many of these features are very subjective and thus radiology cannot be used as a test to exclude coeliac disease. Marked mucosal effacement typically involving the duodenum, but sometimes the whole of the jejunum, may be a prominent feature leading to the so-called 'moulage' appearance which is specific for coeliac disease and for the rare graft-versus-host syndrome

(Fig. 1). Likewise, transient intussusceptions occurring during the examination are only seen with coeliac disease.

Fig. 1 Severe coeliac disease showing moulage phenomena. There has been complete obliteration of the small bowel fold pattern such that the barium follow through appearances resemble toothpaste squeezed from a tube, leading to the alternative term of 'tube of toothpaste' sign. This particular study was performed because the patient was thought to have a carcinoma of the stomach and thus illustrates one of the very varied ways in which coeliac disease can manifest.

Complications, such as the development of lymphoma, may be demonstrated by both barium studies and CT, whilst stricture formation following ulcerative jejunitis, often regarded as a prelymphomatous condition, is best assessed by barium studies.

Other causes of malabsorption include jejunal diverticulosis and short bowel syndrome, plus infiltrative causes such as intestinal lymphangectasia, Whipple's disease, eosinophilic gastroenteritis, amyloidosis, and mastocytosis:

- Jejunal diverticulosis causes malabsorption because of intestinal stasis and subsequent bacterial overgrowth. The diverticula usually protrudes through the mesenteric border of the jejunum and can be demonstrated on small bowel series as a number of sac-like structures containing barium, particularly if erect views are taken.
- Short bowel syndrome results if extensive lengths of small bowel have been resected, often either as a result of vascular accidents or following complications of Crohn's disease. The bowel will adapt by increasing its diameter, but an overall length of less than 50 cm is usually incompatible with life. This situation is particularly severe if the terminal ileum has been resected because of its specialized function in the absorption of vitamin B_{12} and in the enterohepatic circulation of bile salts.

The various small bowel infiltrations are rare:

- Intestinal lymphangestasia is characterized by the presence of dilated lymph vessels in the mucosa which cause protein and lymphocytic loss into the small bowel. It may be primary or secondary, with the secondary form being caused by conditions which obstruct the bowel lymphatics, such as lymphoma, carcinoma, postradiation changes, and occasionally in association with heart failure and constrictive pericarditis. High-quality barium studies may demonstrate the dilated lymphatics as very fine, mm-size nodules, mainly in the jejunum (Fig. 2).

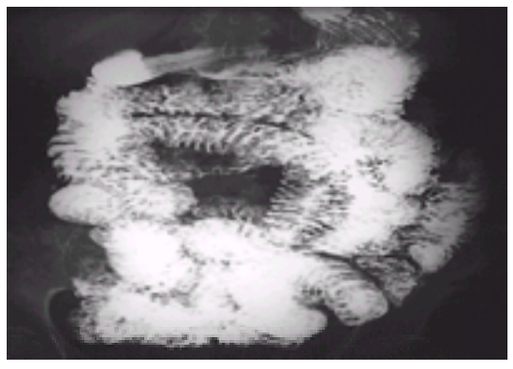

Fig. 2 Intestinal lymphangiectasia demonstrated by thickening and straightening of the fold pattern together with a faint granular background due to the dilated lymphatics. In this case lymphangiectasia was secondary to an underlying lymphoma.

- Whipple's disease is caused by a bacillus resulting in a syndrome of steatorrhoea, abdominal pain, and arthralgia. It has a male predominance. Barium studies show thickened folds together with slight dilatation of the bowel lumen. In addition, CT may demonstrate bowel wall thickening together with mesenteric lymphadenopathy, the lymph nodes usually being of low attenuation.
- Eosinophilic gastroenteritis—in this condition, the wall of the bowel is infiltrated with eosinophils and there is usually a peripheral eosinophilia as well. Often, there is a history of atopy and the triad of eosinophilic gastroenteritis, asthma, and mononeuritis multiplex comprise the Churg–Strauss syndrome. Symptoms resulting from small bowel involvement depend on the major site of location of eosinophilic infiltration, which may be either mucosal or serosal. With the former diarrhoea, malabsorption, and protein loss are dominant, whereas the latter is characterized by ascites. Barium studies in cases of mainly mucosal involvement show fold thickening together with nodular filling defects, whereas serosal involvement is better assessed by CT, which shows both the bowel wall thickening and ascites.
- Amyloidosis and mastocytosis are very rare causes of malabsorption with infiltration in both cases, causing thickening of the valvulae connivente. Amyloidosis is characterized by the deposition of amyloid fibres in the small bowel and may be primary or secondary (Fig. 3).

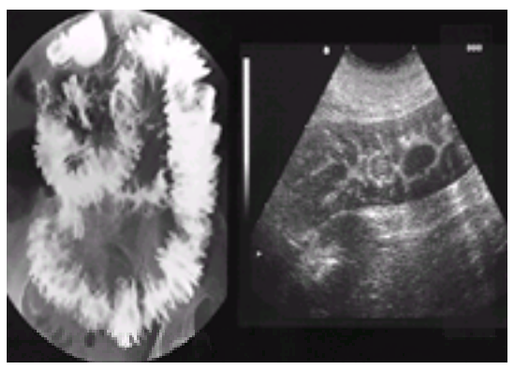

Fig. 3 Amyloidosis of the small bowel shown on a barium follow through (a) and ultrasound examination (b). There is complete disorganization of the normal small bowel fold pattern (a) and the degree of thickening of the small bowel wall is best appreciated on the ultrasound (b).

Tumours of the small intestine

These may be benign or malignant and are relatively uncommon:

- Benign tumours include leiomyomas, haemangiomas, and adenomas. Bleeding or intussusception are the commonest presentations, particularly from leiomyomas. These tumours may grow into the bowel lumen, but often have a significant exoenteric mass, in which case they can grow to a very large size without causing clinical symptoms. They may, therefore, be demonstrated on barium studies as a filling defect, sometimes with a central ulcer which results from necrosis, whereas CT is the best method for demonstrating the exocentric component. A number of syndromes include small bowel polyps as part of their features and these include Peutz–Jegher's with small bowel hamartomas, Gardener's with adenomas, and Cronkhite Canada with retention polyps. Many of these have an association with periampullary carcinoma.

Malignant tumours of the small bowel include carcinoid, carcinoma, lymphoma, and metastases:

- Carcinoid typically arises in the distal ileum and presents as a smooth, submucosal mass. At this stage, they behave like a benign tumour. Later, local infiltration occurs causing extensive tissue thickening and a desmoplastic reaction. This may be severe enough to cause bowel obstruction. Once the tumour has reached a size of 2 cm or more, it is likely to display true invasiveness, metastasizing via the portal vein to the liver. Radiological assessment is best by CT, which can show evidence of local infiltration into the mesentery typically with a central spiculated mass, and later very vascular liver metastases.
- Adenocarcinoma of the small bowel is rare with an approximate incidence of 1 in 100 000. These tumours are most commonly found in the jejunum. The usual macroscopic appearance is that of an annular, concentric tumour which gives rise to an apple-core appearance on barium studies or sometimes a large mass (Fig. 4). Prognosis is poor as the majority of tumours have metastasized at the time of surgery.

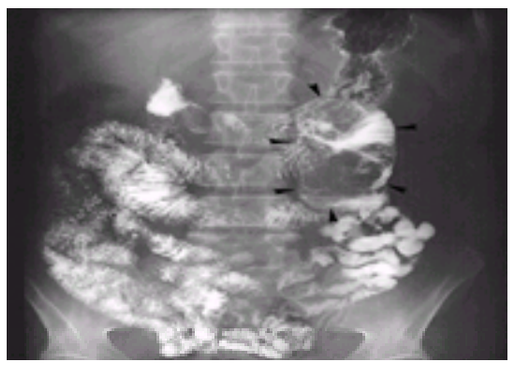

Fig. 4 Adenocarcinoma of the jejunum. This follow through examination in an anaemic patient demonstrates the large mass (arrowheads) arising from the proximal jejunum at a site just distal to the ligament of Treitz—a common position. Note also the normal small bowel pattern distal to this tumour when compared to the abnormal patterns shown on Fig. 1, Fig. 2, and Fig. 3.

- Lymphoma may involve the gastrointestinal tract and indeed is the most common location for extra nodal disease. In contradistinction to carcinoma, the ileum is more frequently involved than the jejunum. A number of macroscopic forms are seen including: (1) multiple nodular defects; (2) diffuse infiltration (Fig. 5); (3) the infiltration may destroy the muscularis propria leading to the so-called aneurysmal form with bowel dilatation; (4) large endoexoenteric masses with excavation; (5) large extraluminal masses. The latter is a particular feature of HIV-associated Burkett's lymphoma.

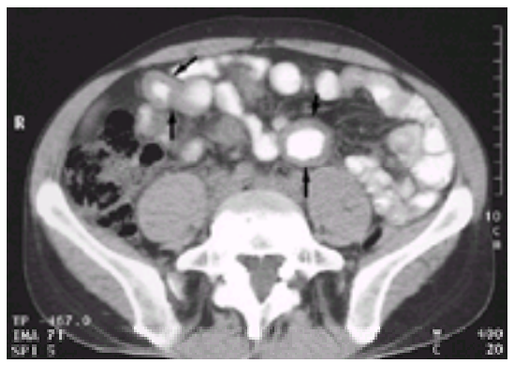

Fig. 5 Lymphoma of the small bowel as shown at CT examination. Several loops are involved demonstrating marked thickening of the bowel wall to a width of 5 to 6 mm (arrows).

- Metastatic disease to the small bowel may arrive by either transcoelomic spread or via the haematogenous route. Common tumours to seed across the peritoneal cavity are from the colon, pancreas, and stomach with the addition of the ovary in females. The commonest source for haematogenous metastases to the small bowel is malignant melanoma which produces characteristic submucosal masses with a central umbilical-type ulcer. Carcinoma of the bronchus and breast also are major causes of haematogenous metastases.

Crohn's disease and inflammatory small bowel conditions

Crohn's disease is the commonest inflammatory condition to affect the small bowel in Western populations and is characterized by a chronic course of progression with spontaneous remission. The cause remains unknown and despite the presence of non-caseating granulomas on biopsy, extensive investigation into infectious agents has so far proved inconclusive. It may involve any part of the gastrointestinal tract, but in the great majority of cases the small bowel demonstrates most macroscopic change, and when involved the terminal ileum is the usual site.

The disease starts with mucosal ulceration which then becomes transmural with fissure ulcers. Healing is accompanied by fibrosis and stricture formation and there is considerable thickening of the bowel wall. Skip lesions may occur affecting considerable lengths of small bowel. Fistulation is a characteristic of the disease and these may be enteric, enterocutaneous, or extend into the muscles of the abdominal wall and pelvis. Intra-abdominal abscesses also characterize severe disease. Radiological assessment is by means of Tc99m white blood cell scanning, barium follow through studies, CT, and MR. White cell scanning is used as the initial test for a patient with symptoms and signs of possible Crohn's disease (Fig. 6). It is also used to document disease activity and response to therapeutic measures. It is highly sensitive but may produce false positive results from other conditions which cause terminal ileal inflammation. Barium studies are then performed to demonstrate morphological detail. These may show the earliest change of Crohns disease which is an aphthoid ulcer resulting from ulceration on the tip of a lymphoid follicle. Next comes distortion and thickening of the fold pattern which in some instances becomes completely effaced. Finally, development of deep, interlacing, linear ulcers give rise to the characteristic cobblestone pattern due to oedema of mucosal islands between the ulcers. If stricture formation occurs, this is best appreciated by barium studies and if there is a long segment of stenosis, this is sometimes referred to as the string sign. The demonstrations of fistula is also shown by barium studies, but this is now increasingly in the realm of spiral CT which can also demonstrates intra-abdominal abscesses and abnormal intra-abdominal fat. The role of MR imaging in Crohn's disease is also evolving rapidly because this can demonstrate the thickened bowel wall, inflammatory changes in the mesentery, and, in particular, the presence and anatomy of fistula formation.

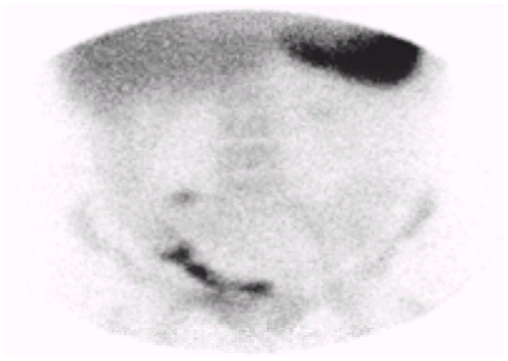

Fig. 6 Crohn's disease of the small bowel as demonstrated by a technetium 99m labelled white cell scan. There is grade III activity in the right iliac fossa indicative of highly active Crohn's disease in the terminal ileum. Note the normal accumulation of isotope in the spleen. (By courtesy of Dr Jane Dutton.)

Other inflammatory conditions of the small bowel include infection, such as *Yersinia enterocolitica* and tuberculosis. The former is associated with changes of terminal ileitis and is usually self limiting. The latter may affect any part of the gastrointestinal tract, but typically the ileocaecal region. Now that bovine tuberculosis has largely disappeared, the bacillus in most cases has arrived from swallowing infected sputum. Barium studies usually demonstrate changes in both the terminal ileum and caecum, a helpful point in distinguishing this from Crohn's disease, although the appearances can be very similar. There may be two type of appearance—the ulcerative form and the hypertrophic form. The latter is characterized by thickened and matted loops of bowel in the right iliac fossa. In addition, tuberculous enteritis may present as ascites from tuberculous peritonitis.

Postradiation enteritis and bowel ischaemia

Postradiation enteritis is considered here because the pathophysiology is one of endarteritis obliterans. This may follow any form of radiotherapy, either external beam or cavity therapy, and is most commonly encountered in female patients who have had therapy for carcinoma of the cervix. Thus, it typically involves small bowel loops which lie deep in the pelvis and these are made more susceptible if there is adhesive disease which fixes them in this position. Initial changes are of mucosal oedema but major problems from fibrosis and stricture formation occur later, in some instances after a latent period as long as 25 years. Sinuses and fistulae can be particularly problematic if there has been previous surgery to the radiation-damaged bowel. These changes are well appreciated at barium follow-through studies.

Bowel ischaemia and infarction may result from both arterial and venous occlusion. Arterial thrombosis of the superior mesenteric artery results in catastrophic infarction of most of the small bowel and carries a dismal prognosis. On the other hand, multiple, small emboli may cause episodes of non-occlusive ischaemia from which the bowel makes a good recovery. These typically are shown as areas of narrowing due to oedema with an abrupt transition to normal small bowel between the involved segments. Venous occlusion may result from volvulus of the bowel, blood dyscrasias, and malignant infiltration in the mesentery.

The colon

Colonoscopy has revolutionized imaging approaches to the colon because of its therapeutic as well as diagnostic role. However, it is not without risk, and barium enema examination still remains a much used alternative. CT is an increasingly used technique particularly since the advent of spiral CT allows three-dimensional reconstruction and thus so-called virtual colonoscopy.

For all three techniques a completely clean and empty colon is a prerequisite. This is achieved by colonic lavage, usually using an oral preparation the day before the examination, such as Picolax or Klean prep. For barium enema examination, the patient is placed on a fluoroscopic table and barium sulphate is run into the colon up to the level of the transverse colon. The excess is then drained out and air or carbon dioxide is insufflated into the colon to achieve a so-called double contrast effect, that is luminal distension with air and mucosal coating by barium sulphate. Up to 12 images are then taken of the colon with the patient in different positions so that the entire bowel is demonstrated in double contrast.

For CT examination, no positive contrast is used but the colon is simply distended with air, usually after the administration of a spasmolytic drug such as Buscopan. So-called volume rendering of the data following spiral CT examination permits a three-dimensional reconstruction of the lumen of the colon, thus allowing the operator to apparently 'fly' through the colon in a manner similar to colonoscopy, hence the term virtual colonoscopy.

CT without bowel preparation has also been used in the detection of gross colonic abnormalities in elderly patients, thus sparing them the discomfort of bowel preparation and a barium enema, both of which are likely to be less than optimal in this group.

Anatomy of the large bowel

The colon commences with the caecum and appendix in the right iliac fossa and these structures lead into the ascending colon which is a retroperitoneal structure. After the hepatic flexure the transverse colon is again intraperitoneal, suspended by the transverse mesocolon. The splenic flexure marks the transition to descending colon which is again retroperitoneal. Finally, the sigmoid and proximal rectum are intraperitoneal with the peritoneal reflection sited at the junction of the mid and lower third of rectum.

Pathology of the colon

Diverticular disease and irritable bowel syndrome

Diverticular disease and its complications are one of the commonest conditions to affect the colon. It results from raised intraluminal pressure causing a bleb of mucosa to herniate through the bowel wall at points of potential weakness, where the nutrient artery pierces the wall to supply the colon. This is accompanied by hypertrophy of the circular muscle fibres in the bowel wall, which is the first sign of this condition. Diverticula by themselves may not cause symptoms and indeed they are said to be present in one-third of the population over the age of 60. They may, however, become inflamed, resulting in a number of complications which include local and segmental abscess formation, perforation and fistula formation, stricture formation, and colonic bleeding. The usual location is the sigmoid colon which, with severe diverticular change, may become very distorted. A barium enema is the best technique for demonstrating the presence and extent of diverticulosis and these are shown as small barium-filled out-pouchings from the bowel wall. Most of the complications of diverticular disease, however, are best appreciated at CT. This demonstrates the marked thickening of the bowel wall as well as the presence of fistulae and abscesses. The latter are shown as soft tissue areas of fluid density either within or immediately adjacent to the wall of the bowel (Fig. 7). Stricture formation, however, is best assessed at barium enema and will usually require colonoscopy for biopsy purposes as the distinction between benign diverticular stricture and one secondary to malignancy is often difficult.

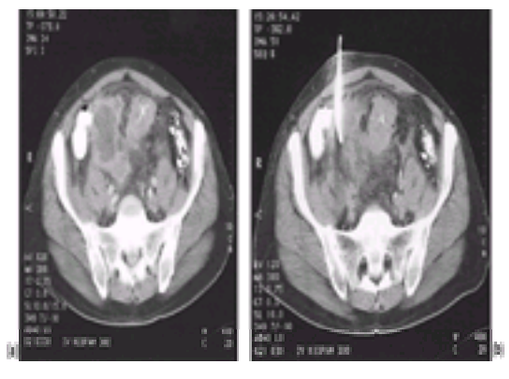

Fig. 7 Diverticular disease and its complication as shown at CT examination (a) demonstrating the marked thickening of the sigmoid wall and to the right of this is a fluid collection demonstrated by a low attenuation mass. This has been drained under CT control with the drainage catheter shown in position (b).

The diagnosis of irritable bowel syndrome is not radiological but clinical. However, in many instances the distinction from diverticular disease and colonic carcinoma is impossible and therefore a further examination, usually a barium enema, is indicated to exclude these diseases.

Colorectal cancer and polyp formation

Colorectal cancer is the second most common cause of death from cancer in the Western world and early detection can have a profound effect on prognosis. It is now clear that colorectal cancer arises as a result of a n umber of mutations resulting in a chromosome instability pathway. The first macroscopic change is the formation of an adenomatous polyp which over time, often two or three decades, will eventually become a frank carcinoma. The percentage of polyps, which, if left alone will become a carcinoma is unknown, but any attempt to reduce the mortality from colorectal cancer starts with their detection and subsequent removal. Polyp detection is achieved by both barium enema and CT colonography.

At barium enema, a polyp may be demonstrated as a filling defect in the barium column or as a ring of increased density on the air contrast views. They may be sessile or pedunculated.(Fig. 8). The size of the polyp is significant as lesions under 5 mm in size have little or no statistical association with an increased risk of cancer. Any polyp over the size of 10 mm should be removed and once they reach 20 mm they will almost certainly be malignant. For obvious reasons, their detection requires a colon that is completely clean as residual faecal material can readily be misinterpreted as polyp.

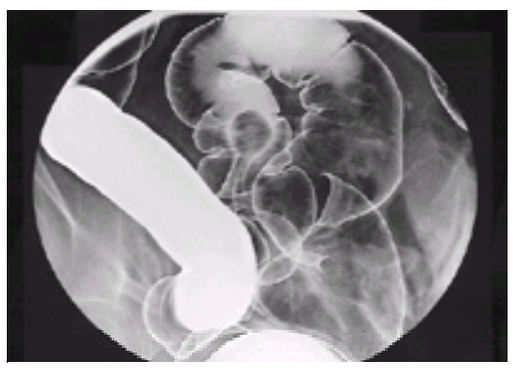

Fig. 8 Colorectal polyp in the mid sigmoid as shown on a barium enema. This is a pedunculated polyp and with the presence of a such a long stalk is likely to be benign. A sessile polyp of this size would have to be regarded as definitely suspicious and possibly malignant.

Carcinoma of the colon has several macroscopic appearances including annular stricture, a proliferative type, a large polypoid type, and schirrhous. Both barium enema and CT examinations detect of all of these.

CT colonography has two major advantages. First, if a tumour is demonstrated, then it can be formally staged at the same examination, that is by looking for involvement of the lymph nodes or the presence of liver metastases. Second, if a further clarification or a therapeutic manoeuvre is indicated, then colonoscopy may be performed immediately after the CT as only air has been insufflated into the colon.

Inflammatory bowel disease

The two main causes of idiopathic inflammatory bowel disease are ulcerative colitis and Crohn's disease with an approximate ratio of 3:1. Ulcerative colitis invariably starts in the rectum and spreads proximally, whereas Crohn's disease more commonly involves the right colon. Both conditions are best appreciated at barium enema, though the complications of Crohn's disease, that is fistulae and abscess formation, are best demonstrated by CT. The radiological appearances reflect the pathophysiology. Ulcerative colitis results in fine mucosal ulceration which always involves the rectum and then may spread proximally to involve the entire colon. There is loss of the haustral pattern and the mucosa demonstrates a fine granular appearance. Crohn's disease in contradistinction demonstrates more discreet and deep ulcers. The latter represent fissure ulcers which may extend throughout the whole thickness of the bowel wall. There is discontinuity of involvement as well as asymmetry. The picture is completed by terminal ileal involvement. An unusual feature of Crohn's disease is its earliest appearance, which is that of aphthous ulceration caused by ulcers appearing on the surface of hypertrophied lymphoid aggregates.

Other causes of colitis include pseudomembranous (caused by an overgrowth of *Clostridium difficile* often following antibiotic usage) as well as the various infective colitides, for example shigella, campylobacter, or CMV. The last named is usually only a problem in the immunocompromised patient. Postradiation colitis typically affects the sigmoid colon, for example after treatment for carcinoma of the cervix.

Further reading

Bartram CI (1999). Imaging in coloproctology. *Clinical Radiology* **54**, 413–14.

Freeman AH (2001). CT and bowel disease. *British Journal of Radiology* **74**, 4–14.

Herlinger H, Maglinte DT, Birnbaum BA, eds (1999). *Clinical imaging of the small intestine*, 2nd edn. Springer-Verlag, New York.

14.2.4 Investigation of gastrointestinal function

Julian R. F. Walters

Digestion and absorption of food by the gastrointestinal (**GI**) tract is achieved by the integration of multiple steps: the complex foods taken in through the mouth are digested into simpler molecules which can be transported across mucosal epithelial cells into the metabolic pool of the body. The contents of the gastrointestinal tract also need to be moved to regions where specialized digestive and absorptive functions can take place, and physical and immunological barriers must be maintained to prevent injury from toxic or immunologically active substances and bacteria.

These functions are assessed clinically in a variety of ways. Much can be learnt indirectly from techniques not principally aimed at defining GI function. Patients will describe appetite, dietary intake, weight changes, and the frequency and nature of their bowel movements. Clinical examination may reveal malnutrition—either generalized or specific. Many blood measurements of absorbed dietary components (such as iron, folate, vitamin B_{12}, cholesterol, and triglycerides), or their metabolic products (such as haemoglobin, albumin) can be abnormal when GI function is impaired, and their serial changes can be used to follow improvements with treatment. Radiological studies (see Chapter 14.2.3) can demonstrate functional changes as well as anatomy, and physiological measurements are central to studies of motility disorders (Chapter 14.12). Tests aimed at giving specific measurements of GI function are described below.

Intake and output

Nutritional assessment

The dietary history is a critical part of the investigation of GI function. For instance, a high intake of milk in a patient of African ethnicity will suggest that diarrhoea may be due to lactase non-persistence. Vitamin B_{12} deficiency in a vegan is most likely to be due to dietary deficiency rather than malabsorption.

Patients may not accurately recall what they eat. Assessment can be improved by keeping a detailed diary with formal recording of the diet over a week, this will enable the usual intakes of a full range of nutrients to be calculated. Total calories, fat, protein and nitrogen, water, electrolytes, individual vitamins, minerals, and trace elements can all be assessed in this way.

Assessment of nutritional status can indirectly provide evidence of gastrointestinal dysfunction. Calculation of the body mass index (kg/m ^{2}) gives a measure of obesity, and hence fat stores. More detailed estimates of body composition can be made by anthropometry, measuring skin-fold thickness, or body density. Dual-energy, X-ray absorptiometry will assess the percentage of fat, and bone mineral density as a measure of calcium stores.

These estimates of nutritional status can change over time. If the intake of any particular nutrient exceeds the losses, the body is in positive balance, as occurs during growth. If the losses are greater, the result is a negative balance. Losses from all sources must be included. Urinary loss is obvious for water, electrolytes and minerals, and nitrogenous compounds. Carbon dioxide and heat losses reflect metabolism and calorie consumption: research calorimetric techniques can estimate these accurately. Absorption by the GI tract is a major factor is determining overall balance—with unabsorbed nutrients and excreted matter being egested in faeces.

Faecal output

Stool weight and volume can vary in the healthy individual, but averages about 200 g/24 h. This volume increases in diarrhoea or other forms of malabsorption, and may be as high as several litres/24 h in patients with secretory diarrhoea, such as that due to cholera. Accurate measurements of faecal volume and electrolyte composition are then helpful in maintaining an accurate fluid balance.

Patients may complain of diarrhoea but mean urgent, frequent, or unformed stools, rather than an increase in volume. Stool charts, recording frequency and volume, help to define the change in the nature of the stools. Changes in frequency and volume with simple changes such as fasting help to differentiate osmotic diarrhoea from secretory or inflammatory causes. Stool electrolytes and osmolarity are also helpful here: a large osmotic gap suggests an unabsorbed ion from, for example, magnesium salts taken as laxatives, or the products of unabsorbed carbohydrate fermentation. Stool pH is low after carbohydrate malabsorption, as in lactase deficiency (non-persistence) or sucrase–isomaltase deficiency.

Faecal fat output is increased in most forms of generalized malabsorption and results in steatorrhoea. Patients frequently describe stools that float or are foul-smelling, but to be sure that this is due to fat, qualitative and quantitative estimations need to be performed. Fat droplets in the stool can be detected microscopically after staining with lipid-soluble dyes. Accurate estimation of the loss of fat in the faeces requires a 3-day collection of stools while the patient is on a defined fat-intake diet. An average output of more than 5 g/24 h for a 70 to 100 g daily fat intake is abnormal. Patients and clinical and laboratory staff dislike this test for obvious reasons.

Gas, wind, explosive stools, and foul odours are frequent complaints. Although volumes of flatus and the presence of various gases have been measured in research studies, these have not been adopted as routine clinical investigations.

Stool microscopy for faecal leucocytes can be helpful in diagnosing inflammatory diarrhoea. Detection of bacterial pathogens, parasites, ova, or toxins may show the cause.

Digestive secretions

In patients with possible malabsorption, who have an adequate dietary intake but large volume stools, an important clinical decision is whether digestion is at fault (such as pancreatic exocrine insufficiency) or whether absorption is the problem (as in coeliac disease). Often this is rapidly established by employing tests with high positive-predictive-values for common individual diseases, such as endomysial serology for coeliac disease or imaging studies for chronic pancreatitis (see Chapter

14.18.3.2).

Intubation of the lumen to collect the contents for measurements of secretory rates and composition is the definitive way to study digestive juices produced by the stomach, pancreas, liver, and intestine. Though necessary for basic physiological and pharmacological studies, tube tests are now rarely performed in a clinical setting. Endoscopes are generally now used, which enable direct vision and biopsy of anatomical lesions, and can, on occasions, also provide functional information.

A large number of tubeless tests have been developed to indirectly measure digestion and absorption, or to help differentiate between the two types of disorder. Many involve small doses of radionuclides. Other tests use markers of breath hydrogen or urinary excretion (Table 1). Selection of these tests in clinical practice depends on their predictive values, reliability, cost, and ease of use.

Gastric secretion

The gastric mucosa secretes acid, pepsin, and some other products such as intrinsic factor. Measurements of acid output have historically been important in diagnosing the cause and response to treatment of acid-related diseases such as duodenal ulceration. This has become less relevant with the discovery of *Helicobacter pylori* and potent acid suppression with histamine H2-receptor antagonists and proton-pump inhibitors.

Intubation of the stomach with a nasogastric tube allows the gastric contents to be sampled. Tube positioning is important. Swallowed salivary secretion will raise the pH and so will refluxing duodenal contents, common after retching. The yellow colour of bile will indicate duodenal reflux into the stomach. After an overnight fast, gastric aspiration allows the volume of resting secretions to be measured and the pH determined. Normal resting volumes are less than 50 ml. pH values above 4 suggest impaired acid secretion, as in the gastric atrophy and achlorhydria found in pernicious anaemia, or as the result of acid-suppressant drugs. Gastric pH can also easily be measured at endoscopy.

Estimation of basal and peak acid output requires continuous sampling and titration of the aspirate with sodium hydroxide. A marker can be infused to correct for loss of gastric contents into the duodenum. A basal acid output of about 5 mmol over 1 h is normal. Low values are found in achlorhydria. The detection of a high basal output is important in making the diagnosis of Zollinger–Ellison syndrome. However, this clinical decision is now usually made after finding a high serum gastrin level in a patient treated with a proton-pump inhibitor, who because of symptoms is unable to discontinue therapy.

Peak acid output following pentagastrin stimulation quantifies the ability of the stomach to maximally produce acid. There was considerable interest in the use of this test for research purposes before the aetiology of duodenal ulcer became clear. However, it is of little clinical use now that our understanding of pathophysiology and therapeutics has advanced.

Biliary secretions

Bile samples, mixed with pancreatic secretions and other duodenal contents, can be collected from the duodenum at upper endoscopy (or after intubation). Endoscopic retrograde cholangiopancreatography (**ERCP**) allows bile to be collected from the bile ducts. Microscopy of bile can detect cholesterol crystals. The proportions of bile acids, phospholipids, and cholesterol are relevant in the study of biliary cholesterol saturation and gallstone formation, but have little clinical use.

Bile secretion by the liver is not easily measured directly. Clearance from the circulation can be determined for a number of compounds normally secreted into the bile. Apart from measurements of bilirubin, such tests have found little clinical use. The nuclear medicine **HIDA** (hepatoiminodiacetic acid; lidofenin) scan gives a measure of biliary secretion as well as helping define functional anatomy. The gallbladder function of contraction in response to a meal, or cholecystokinin (**CCK**), can be measured by imaging techniques, of which ultrasound is the most convenient.

Pancreatic secretions

Pancreatic secretion of a large number of digestive enzymes and bicarbonate is crucial for the digestive process. Duodenal intubation allows their collection for assay of the activities of some of the key enzymes such as trypsin, lipase, and amylase; however, these duodenal contents will be mixed with bile and duodenal secretions, or with those from the stomach. Pure pancreatic juice can be collected at ERCP.

Although a number of stimulation tests have been used to diagnose pancreatic exocrine insufficiency, their clinical usage is very limited: virtually all patients can be managed without the need to resort to these function tests. The duodenum is intubated, usually under fluoroscopic control. Secretin, which stimulates bicarbonate secretion, or CCK, which stimulates enzyme secretion, are given intravenously, either alone or in combination. The contents of the duodenum are then aspirated for assay of bicarbonate and enzyme concentrations. The Lundh meal is an alternative standardized stimulus to pancreatic secretion.

Several indirect, tubeless pancreatic tests have been developed and are much simpler and more convenient to perform. The pancreolauryl test relies on the hydrolysis of fluorescein dilaurate by pancreatic esterases. The fluorescein is absorbed and excreted in the urine where it can be assayed easily. The test is administered on 2 days, each time with a standard breakfast and similar urine collections. On the first day, a test capsule containing fluorescein dilaurate is given, and on the second, a control capsule of non-esterified fluorescein. To diagnose pancreatic insufficiency, the ratio of fluorescein recovered in the urine with the dilaurate test substance will be less than 20 per cent of that with the control. Another test uses *N*-benzoyl-L-tyrosyl-*p*-aminobenzoic acid (**NBT-PABA**) as an alternative substrate. Following a similar principle, pancreatic chymotrypsin activity releases *p*-aminobenzoic acid (**PABA**), which is assayed in the urine. This test seems to be less reliable than that with pancreolauryl.

A different type of indirect test for pancreatic exocrine dysfunction involves the determination of proteolytic enzymes in the faeces—these enzymes are produced by the pancreas and are stable during passage through the intestine. Chymotrypsin activity has been used as a test for many years. An improved method using an immunological assay for human-specific elastase I has recently been developed.

All these indirect tests are generally reliable in patients with severe exocrine pancreatic insufficiency causing steatorrhoea, where they can have adequate specificity for pancreatic disease if borderline test results are repeated or ignored. Many patients with severe intestinal disease will have somewhat abnormal results. The sensitivity of the tests is such that patients with lesser degrees of pancreatic functional impairment will be missed and considerable pancreatic damage may have occurred before the tests become abnormal.

Intestine

The role of brush-border enzymes in digestion is clearly important. Individual enzyme activities (e.g. sucrase–isomaltase) can be measured directly in small-bowel biopsies, obtained at endoscopy or using a capsule or biopsy tube. Indirect tests of intestinal digestive enzyme function are closely linked with those of absorption and are described below. Intestinal secretions, such as those from the Brunner's glands in the duodenum or from the crypts in secretory diarrhoea, are not directly measured.

Absorption

Carbohydrates

The xylose absorption test has been used for many years as a simple measure of absorption and malabsorption. Xylose does not need to be digested, is absorbed by the jejunum, does not undergo significant metabolism, and is excreted efficiently by the kidney. Urine is collected for 5 h after oral administration of the sugar to a fasted subject. Usually 25 g is given and more than 17 per cent should be recovered in the urine. A smaller dose of 5 g causes less diarrhoea. The test measures the area of functioning mucosa and has reasonably good sensitivity for intestinal mucosal abnormalities such as coeliac disease. The measurement of blood levels is said to improve accuracy. False-positive low levels of xylose excretion can occur with delayed gastric emptying, ascites, and with insufficient urinary output. As described elsewhere, serological tests are now available that are simpler, more specific, and more sensitive in screening for coeliac disease.

Lactose tolerance testing can be performed to diagnose lactase deficiency. This is similar to the glucose tolerance test, except that 50 g of lactose is given instead of glucose. Blood glucose levels are measured, and should increase if brush-border lactase is present to split the lactose in to glucose and galactose. In subjects without

brush-border lactase, the lactose is not digested or absorbed and the blood glucose level does not rise. The lactose passes through the small bowel to the large intestine where it is then broken down by bacteria, thereby causing symptoms of gaseousness and diarrhoea.

The breath-hydrogen test for lactose intolerance is based on this principle (Fig. 1). Hydrogen is only produced in the body from the bacterial fermentation of carbohydrates in the gut. Hydrogen diffuses into the blood and is excreted in the breath. Breath hydrogen can be measured simply and relatively cheaply to an accuracy of a few parts per million (**p.p.m**). Colonic bacteria produce a background level of breath hydrogen from the fermentation of unabsorbed colonic contents. After taking oral lactose, breath hydrogen is measured every 30 min. An increase greater than 20 p.p.m. implies that, rather than being absorbed in the small intestine, the sugar has been broken down by bacteria in the colon. As described below, bacterial overgrowth in the small intestine elevates breath hydrogen after glucose administration, and can also do this when lactose is given. Similar breath-hydrogen tests can be performed with other sugars when fructose or sucrose malabsorption is suspected.

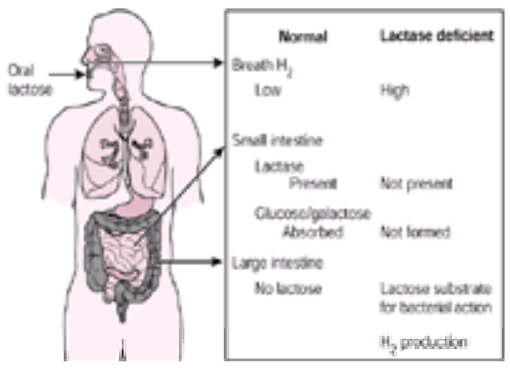

Fig. 1 Lactose-hydrogen breath test for lactase deficiency.

Fat

The estimation of non-absorbed fat, that is faecal fat, is discussed above. A tubeless test, which does not require faecal collection, is the [^{14}C]triolein test. This has been advocated as a simple method to determine lipid absorption. After digestion and absorption of the labelled triglycerides, metabolism in the liver produces ^{14}C-labelled CO_2 which s detected in the breath. This test is not widely used and is only sensitive to large changes in fat absorption.

Bile salts

Bile salts, critical for lipid absorption, undergo an enterohepatic circulation where the conjugated bile salts (such as glycocholate and taurocholate) are reabsorbed in the ileum. Failure to reabsorb these salts increases their concentrations in the colon where they produce a secretory diarrhoea. Tests have been developed to look for evidence of bile salt malabsorption.

The SeHCAT test uses radiolabelled selenohomocholic acid as the taurine conjugate. This is given orally and whole-body retention is measured with a gamma camera after 7 days. Low values result from an excessive loss of bile salts. Bile salt malabsorption, often producing SeHCAT retention values of less than 5 per cent, is found after ileal resection, with ileal disease such as Crohn's, and in idiopathic bile salt malabsorption due to transporter defects. Small intestinal bacterial overgrowth results in deconjugation of the bile salts, which will also impair absorption and retention.

Vitamin B_{12} and the Schilling test

Absorption of vitamin B_{12} (the cobalamins) is particularly complex. Vitamin B_{12} deficiency is common, and, if not nutritional, is due to one of several GI disorders. Intrinsic factor, produced by the stomach, binds cobalamins in the intestine. The intrinsic factor–vitamin B_{12} complex interacts with a receptor in the brush-border membrane of the terminal ileum and is taken up by the cell. Vitamin B_{12} is stored in the liver. Pancreatic enzyme activity is necessary to release dietary vitamin B_{12} from R-proteins in the diet so that it can bind to intrinsic factor. Bacterial overgrowth in the small intestine can split the intrinsic factor–vitamin B_{12} complex before it can be absorbed.

The Schilling test uses radioisotopes of cobalt to label cobalamin. A two-part test is usually performed, with two isotopes, ^{57}Co and ^{58}Co, simultaneously. One isotope is used to label free vitamin B_{12} and the other to label a complex of intrinsic factor and vitamin B_{12}. After an overnight fast, both are given together by mouth, and unlabelled cobalamin is given by intramuscular injection to ensure that binding sites are occupied. Absorbed radiolabelled vitamin B_{12} is excreted in the urine. This is collected for 24 h and should normally contain more than 10 per cent of the ingested dose, with an equal ratio of the two isotopes. In gastric disease, such as pernicious anaemia, or after gastrectomy, absorption is reduced when vitamin B_{12} is given alone but not when it is given with intrinsic factor. In terminal ileal disease, or after resection, absorption of both forms is low. Small-bowel bacterial overgrowth mimics ileal disease, but antibiotic treatment restores vitamin B_{12} absorption to normal.

Gastrointestinal transit

Physiological function tests have been developed to measure the motility of the gut in propelling the contents of the diet through the areas involved in digestion and absorption. These are described in full in Chapter 14.12.

Oesophageal function

Peristalsis in the oesophagus, with appropriately timed relaxation of the upper and lower oesophageal sphincters, is necessary for efficient and comfortable swallowing, without symptoms of dysphagia. Gastro-oesophageal reflux through the lower sphincter will produce heartburn. Oesophageal manometry, with multiple fine tubes passed through the nose and connected to sensors, will measure the pressures in the oesophagus and at the sphincters at rest and during swallowing. Disordered peristalsis, spasm, achalasia, nutcracker oesophagus, and conditions associated with reflux can be diagnosed (see Chapter 14.6).

Monitoring gastro-oesophageal acid reflux with pH-sensitive electrodes over 24 h is useful in diagnosing the severity of gastro-oesophageal reflux disease (**GORD**), and in relating atypical symptoms to episodes of reflux. A small, portable recording device allows the times of symptoms, meals, and sleeping to be recorded, so they can be analysed together with episodes of low pH. A composite score can be calculated, reflecting the severity of reflux. A bile-sensitive electrode is also available, and may be of use in patients with adequate acid suppression but who still suffer symptoms from duodenogastric and gastro-oesophageal reflux.

Gastric emptying

This can be measured by a range of imaging tests, of which radionuclide labelling of liquid and solid food is probably the most effective technique. Tracer quantities of technetium and/or indium are incorporated into simple foods. Gamma scanning over the stomach allows the time course of gastric emptying to be determined. These measurements are useful in conditions such as diabetic gastroparesis or after gastric surgery where bloating, nausea, or vomiting are problems. Radiological measurements of barium emptying, although widely available, are less reliable unless the contrast is incorporated into food. Ultrasound and magnetic resonance imaging (**MRI**) have also been used.

Intestinal transit

Radionuclide techniques, as used to measure gastric emptying, can also determine small-bowel transit by timing the appearance of counts over the caecum.

Estimates of transit times through the large intestine can be obtained with further imaging.

Mouth-to-caecum transit times can be estimated simply with breath-hydrogen testing. As described above, breath hydrogen is derived from the bacterial metabolism of unabsorbed carbohydrates. Lactulose, a non-absorbed sugar, is given by mouth and breath hydrogen sampled every 15 to 30 min. A rise in breath-hydrogen values indicates that the lactulose has reached the caecum. A rapid rise will occur if there is bacterial overgrowth in the small intestine.

Dye markers taken by mouth will give an estimate of the whole-gut transit time when they are detected in the stool.

Radiological markers, small differently shaped pieces of radio-opaque plastic, can be useful in determining transit through the large intestine. These are taken daily for several days. A plain abdominal radiograph shows the number remaining and their distribution in the parts of the colon.

Defecation and anorectal physiology can be measured with manometry in response to balloon inflation in the rectum.

Tests of gastrointestinal integrity and barrier functions

Several tests examine other aspects of GI physiology not discussed above. Evidence for infection, structural damage, or loss of barrier functions can be obtained.

Infection

Helicobacter pylori infection of the stomach is common and can be detected at endoscopy by microscopy, culture, or the biopsy urease test. The urea breath test, an indirect test, is now one of the commonest breath tests performed to look at gastric pathophysiology. Either radiolabelled [^{13}C]- or [^{14}C]urea is given by mouth to fasting subjects. The urease activity of *H. pylori* in the stomach metabolizes this to radiolabelled CO2, which is then exhaled. A standard amount of CO_2 is collected in a breath sample and the activity of the isotope determined—by mass spectrometry for ^{13}C or scintillation counting for ^{14}C. As there are no other sources of urease activity in the body, this is a very specific test. Low-level *H. pylori* infection, as can occur when acid production is suppressed with histamine H2-receptor antagonists or proton-pump inhibitors, can give negative results and reduce the sensitivity of this test.

Bacterial overgrowth in the small intestine can be detected by the glucose-hydrogen breath test (Fig. 2). Glucose given by mouth is normally fully absorbed, but some metabolism to hydrogen will occur if bacteria are present in the small intestine. This will be measurable in the exhaled breath, which is collected every 30 min for 2 h. A positive test produces a diagnostic rise of 20 p.p.m. above a low baseline. It may be necessary to give a diet low in non-absorbable polysaccharides to reduce baseline hydrogen production. This test has a sensitivity of about 80 per cent. Alternative tests, which are less popular, involve the administration of radiolabelled compounds such as the bile salt, cholyl-[^{14}C]glycine (glycocholate), and measuring ^{14}C-labelled CO_2 in the breath (Fig. 3).

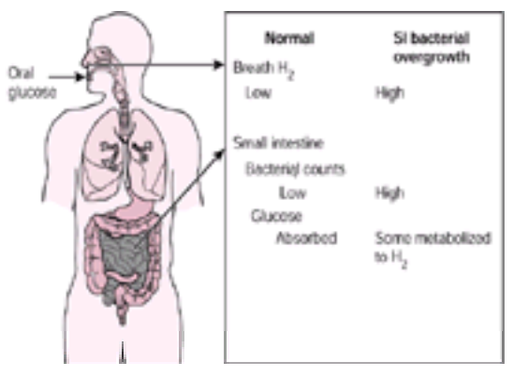

Fig. 2 Glucose-hydrogen breath test for small intestinal bacterial overgrowth.

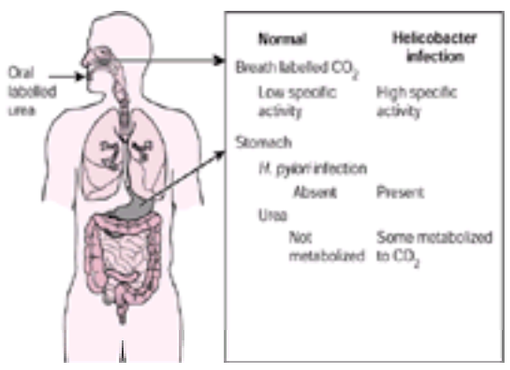

Fig. 3 Urea breath test (^{13}C or ^{14}C) for *H. pylori* infection.

Mucosal damage

Several tests will give abnormal results if the gastrointestinal mucosa is damaged. The faecal occult blood test becomes positive with a relatively small loss of blood per day. Dietary components containing blood or peroxidase can give positive tests. The presence of leucocytes in faeces suggest inflammatory causes of diarrhoea.

White-cell scanning has a role in diagnosing and assessing the activity of inflammatory bowel disease. Autologous white cells, labelled with indium or technetium, are reinjected intravenously and collect in areas of inflammation. Imaging with a gamma camera, at multiple time points, indicate white-cell accumulation at sites of diseased bowel and chemokine production. The complexity and the radiation dosage involved limits the use of this test.

Calprotectin is a calcium-binding protein found in leucocytes. It is stable in the gut lumen and so can be quantified in the stool. It can be detected in faeces in a variety of inflammatory conditions and its simplicity may make it suitable as a first-line screening test.

In patients with small intestinal lymphangiectasia, lymphocytes are lost into the lumen of the gut. The full blood count may then show a lymphopenia. Protein loss also occurs in this disorder. Many other inflammatory and ulcerative conditions can produce a protein-losing enteropathy, and several tests have been employed in attempts to quantify this loss. Since a1-antitrypsin is resistant to breakdown in the intestine, faecal measurements can provide an indication of the loss of serum proteins into the lumen. *In vivo*-labelled [^{51}Cr]albumin is also used to estimate GI protein loss.

The permeability of the intestine can be investigated using a number of different probes: usually sugars, which normally have little uptake in the gut but can enter the body in disease states. These probe substances will then be excreted and measured in the urine. Lactulose is one such sugar that is normally excluded. When administered together with another molecule which is taken up and excreted readily, such as rhamnose or mannitol, the ratio can correct for individual differences in gastric emptying or urine collection. Although the results of these tests are usually abnormal in coeliac disease, specific erological tests, such as endomysial IgA antibodies, are much more specific and sensitive for use in screening. Intestinal permeability changes have been described in other diseases but roles in their pathogenesis are uncertain.

Further reading

Bouchier IAD, *et al.*, eds (1993). *Gastroenterology: clinical science and practice*, 2nd edn. WB Saunders, London.

Yamada T, *et al.*, eds (1991). *Textbook of gastroenterology.* Lippincott, Philadelphia.

14.3.1 The acute abdomen

Julian Britton

Most patients with acute abdominal pain are treated in the community: a minority are admitted urgently to hospital and then, usually, put under the care of a surgeon. Two-thirds of such patients, wherever they live, are suffering from either abdominal pain for which no cause is found or acute appendicitis. In the remaining one-third many disorders account for the pain and, rarely, the cause lies outside the abdomen. Sometimes patients with a surgical diagnosis present to a physician because the abdominal symptoms and signs are overshadowed or suppressed by a medical condition. Occasionally patients in hospital for another reason will develop an acute abdomen.

The management of acute abdominal pain depends on clinical skill, and investigations play only a small part. Sometimes immediate treatment must take precedence over making a diagnosis. More commonly the clinician has time to take a history, to examine the patient, and to consider a differential diagnosis.

Diagnosis

The history

An accurate history is the essential foundation for the diagnosis of abdominal pain. This requires time, patience, and skill. The doctor should give the patient his or her undivided attention and take care not to interrupt the flow of words. The way patients tell their story is as important as the story itself. Any additional questions should be short, specific, and direct and must be couched in language the patient understands. Body language and sentence construction must not suggest any particular answer. Negative findings are always as useful as positive ones. It is important to obtain sufficient but no more than sufficient information. Unnecessary or irrelevant facts can be misleading and will always add to the difficulties of analysis.

Pain

The nature of the pain is the best guide to the diagnosis. The site and any radiation of the pain may suggest which intra-abdominal organ is the source of the symptom. How and when the pain started and whether it has been constant since then are useful points for consideration. Perforation of the bowel, rupture of an aneurysm, and acute ischaemia all occur suddenly and the pain is severe from the outset. Acute inflammation starts slowly and the pain is constant and increases in severity with time. Colic comes and goes and implies an origin from a viscus containing smooth muscle. Aggravation of the pain on sudden movement is equivalent to rebound tenderness on examination and suggests local peritonitis.

Other symptoms

Nausea and vomiting often accompany abdominal pain, but not necessarily together. Retching without vomiting suggests acute intestinal obstruction with impending strangulation. Anorexia is a non-specific symptom particularly in ill children. Disturbances of bowel function are always important and the complete absence of stool or flatus implies intestinal obstruction. The urinary and the genital systems are common sources of abdominal pain and patients must be asked about the associated symptoms. Complications of previous abdominal surgery cause pain and so, sometimes, do therapeutic drugs.

Examination

General observation of the patient is important. Demeanour is a good guide to the overall severity of the illness. Patients with peritonitis lie perfectly still. Patients with colic are restless. Abdominal distension can be often observed through the sheets and jaundice or the smell of melaena are also important signs.

Pulse, temperature, and blood pressure are essential measurements and changes over time are more valuable than isolated readings. Examination of the abdomen itself follows the pattern of inspection, palpation, percussion, and auscultation. It must include a rectal and vaginal examination and examination of the urine.

If palpation causes pain it is essential to decide if guarding, rigidity, and rebound tenderness are present as well. Rebound tenderness is a reliable guide to the presence of peritonitis. A silent abdomen carries the same implication, and in the context of pain is usually a clear indication that a patient requires an urgent laparotomy. Two surgical conditions which are commonly missed are a small strangulated femoral hernia in an obese patient and pelvic appendicitis. Careful examination of the groin should identify the former. In the latter case the abdominal signs are often subdued because somatic pelvic pain is not referred to the anterior abdominal wall. Tenderness on the right side of the pelvis may be the only positive physical sign and it is critically important to overcome any reluctance to perform a rectal examination.

Investigation

With one exception, simple blood tests do not help to diagnose acute abdominal pain but they are important in managing the surgical patient. The exception is the serum amylase which should be measured in every patient with abdominal pain unless the diagnosis is otherwise obvious. High levels of amylase activity are strongly suggestive of acute pancreatitis. Unfortunately serum amylase is also elevated in patients with biliary colic, perforated peptic ulcer, ischaemic bowel, or ruptured aortic aneurysm. A false negative result is found if the blood sample is taken too late after the onset of pancreatitis because amylase concentrations return to normal within about 48 h. A raised peripheral blood leucocyte count or the presence of white cells or red cells in the urine are useful pointers to intra-abdominal inflammation

or disease of the urinary tract, but neither test is specific.

Radiology

Plain abdominal radiographs are helpful in confirming intestinal obstruction and may show calculi in the gall bladder or the renal tract. Air under the diaphragm on an erect chest radiograph is definite evidence of perforation of the bowel or stomach but does not identify the site of the perforation (Fig. 1). About one in ten patients with a perforated peptic ulcer does not show free intraperitoneal air.

Fig. 1 An erect chest radiograph demonstrating free gas under the right diaphragm. (From Britton J (2000). The acute abdomen. In: Morris PJ, Wood WC, eds. *Oxford textbook of surgery*, 2nd edn. Oxford University Press, Oxford, 1823–42.)

An ultrasound examination may show a swollen tender retrocaecal appendix. Computed tomography will confirm acute pancreatitis and acute diverticulitis.

Making a diagnosis

Even experienced doctors only make a correct diagnosis in three-quarters of patients with abdominal pain, whilst junior doctors are right only about half the time. Patterns of disease provide a useful guide: all over the world acute appendicitis and non-specific abdominal pain are the most common diagnoses in a surgical setting (Table 1). Acute cholecystitis is the third most common cause of an acute abdomen in the developed world, whilst in Africa small bowel obstruction is frequent. In a medical ward it would be unusual to encounter non-specific abdominal pain: diverticulitis, a perforation, a ruptured abdominal aortic aneurysm, or mesenteric infarction from embolism secondary to atrial fibrillation or coronary thrombosis would be more likely.

Occasionally a patient presents with all the classical symptoms and signs of a condition. More commonly the clinician has to analyse the data in a number of different ways often based on anatomy, pathology, or, sometimes, by elimination. Pragmatic doctors simply decide if the abdominal symptoms and signs justify surgery. If they do then the diagnosis becomes apparent at operation. If they do not then the patient is re-examined after a few hours. If the signs worsen then surgery is required. If the patient improves then continued observation is appropriate. The only risk of active observation is that the patient deteriorates significantly in the time spent watching, with a consequent increase in the complication rate after surgery.

Surgical causes of abdominal pain

Acute appendicitis

The incidence of acute appendicitis is declining, but the disease remains frequent and often eludes diagnosis. Periumbilical pain moving to the right iliac fossa over a few hours and accompanied by nausea and fever are the classical symptoms. Tenderness, guarding, and rebound tenderness which is worse in the right iliac fossa are the important physical signs. When all these symptoms and signs are present there is little need for investigation and the treatment is a prompt appendicectomy. Prophylaxis against venous thrombosis is important and appropriate antibiotics given before operation are known to reduce the incidence of subsequent septic complications.

Laparoscopy is valuable when the diagnosis is not clear, particularly in young women in whom the rate of appendicectomy with a normal appendix is high. Laparoscopic removal of the appendix is also possible. Whichever method of appendicectomy is adopted, drinking followed by eating may start in a few hours. Most patients stay in hospital for 24 or 48 h.

Non-specific abdominal pain

Three out of every ten patients admitted to a surgical ward with acute abdominal pain are discharged without a diagnosis. This does not mean that there is no pain or that there is no cause. Acute appendicitis can sometimes resolve.

Acute cholecystitis

This is easy to diagnose, particularly with the assistance of an ultrasound examination. Bed rest, intravenous fluids, and, for most patients, antibiotics will allow the inflammation to resolve. If there is no improvement, the gall bladder should be drained either percutaneously or surgically (Fig. 2). In most instances a cholecystectomy should be performed on the next available operating list.

Fig. 2 This abdominal computed tomography scan shows acute cholecystitis. Fluid can be seen within the thickened wall of the gallbladder. The gall bladder was subsequently drained with a percutaneously placed pigtail catheter.

Intestinal obstruction

Abdominal pain, abdominal distension, vomiting, and constipation are the typical symptoms and signs of obstruction. Adhesions, hernias, and cancer of the large bowel are common causes. Obstruction due to adhesions will sometimes resolve itself during treatment with intravenous fluids and nasogastric suction. Hernias can

sometimes be reduced. In most other instances, surgery is required after appropriate resuscitation.

Perforated peptic ulcer

Surgery to close the perforation and to lavage the peritoneal cavity is the standard treatment. Patients benefit later from treatment to suppress acid secretion and to eliminate *Helicobacter pylori*.

Diverticulitis

This presents with pain and tenderness in the left iliac fossa accompanied by fever. The diagnosis is easily confirmed by computed tomography if necessary. Bed rest, intravenous fluids, and antibiotics are the standard treatment. When colonic perforation occurs, and this can be surprisingly silent, resection of the bowel and a temporary colostomy will be required.

Acute pancreatitis

This is a capricious disease and may complicate other medical or surgical conditions. Treatment is with intravenous fluids and analgesia. Computed tomography will confirm the suspected diagnosis and also helps in the management of complications such as a pancreatic abscess or a pseudocyst.

Medical causes of abdominal pain

Pneumonia

Pneumonia may present with referred abdominal pain in the young and the old especially when complicated by diaphragmatic pleurisy. Sometimes the pneumonia results from inflammation below the diaphragm and it is important to treat both conditions.

Hepatitis

Intrahepatic cholestasis due to drugs, viral hepatitis, or alcohol is easy to confuse with cholangitis caused by extrahepatic obstruction. If the bile ducts are not dilated on ultrasound examination 1 week after the onset of jaundice then large duct obstruction is unlikely to be the cause.

Herpes zoster

Persistent severe pain precedes the rash in herpes zoster (shingles). When one of the lower thoracic or upper lumbar dermatomes is involved patients will present with abdominal pain. The diagnosis is usually impossible before the rash has appeared and is obvious once it has.

Drugs

Digoxin, non-steroidal analgesics, laxatives, and drugs which cause constipation can all cause abdominal pain. The first is notorious for causing nausea, vomiting, and abdominal pain when given to excess.

Gastroenteritis

Severe colic is sometimes a presenting feature of gastroenteritis. Diarrhoea and vomiting are not far behind and the diagnosis is rarely difficult. Patients should be isolated as soon as the diagnosis is suspected and stool cultures should be requested promptly.

Diabetes mellitus

Children with acute type 1 diabetes may complain of abdominal pain. They are always ketotic and hyperglycaemic and the diagnosis can often be made by smelling the patient's breath. A surgeon should check that some septic focus is not an underlying cause and refer the patient urgently to a physician.

Rarities

Rare causes of abdominal pain such as sickle cell crises, acute porphyria, lead poisoning, or a spinal tumour are often only made after a considerable period of time and then usually by exclusion. The diagnosis is often suggested by appropriate review of the past history and family history of the patient. Once the possibility of these rarities is considered it is a simple matter to arrange the relevant tests: since these disorders greatly aggravate the risks of exploratory surgery under general anaesthesia, prompt diagnosis and rapid referral for definitive treatment is highly desirable.

Management of the medical patient with abdominal pain

Observation

Most patients who develop acute abdominal pain on a medical ward will require the opinion of an experienced surgeon. If this is deemed unnecessary or a surgeon is not immediately available then the patient should be placed under observation and reviewed at frequent intervals.

Pain relief

Analgesia suppresses abdominal signs and is best not given until a diagnosis has been made. However, a surgeon can allow for the diagnostically confounding effects of analgesia provided it is clear which drug has been given and when; patients in pain should be given the analgesia they need.

Antibiotics

Treating undiagnosed abdominal pain with antibiotics is not recommended. They are rarely curative.

Resuscitation

Patients who may need an abdominal operation should not be given any food or drink by mouth and should be resuscitated with appropriate intravenous fluids. There will then be the minimum of delay if surgery under general anaesthesia is required.

Further reading

Britton J (2000). The acute abdomen. In: Morris PJ, Wood WC, eds. *Oxford textbook of surgery*, 2nd edn, 1823–42. Oxford University Press, Oxford. [A more detailed account of the diagnosis and treatment of acute abdominal pain.]

de Dombal FT (1991). *Diagnosis of acute abdominal pain*, 2nd edn. Churchill Livingstone, Edinburgh. [A compilation of scientific studies of abdominal pain completed over 20 years.]

Martin RF, Rossi RL, eds (1997). Abdominal emergencies. *Surgical Clinics of North America*, **77**, 1226–470. [An up-to-date series of articles on the diagnosis and management of the acute abdomen with an American perspective.]

14.3.2Gastrointestinal bleeding

T. A. Rockall and T. Northfield

Introduction and definition

Acute gastrointestinal haemorrhage is classified by its origin, from either the upper or the lower gastrointestinal tract, anatomically demarcated by the ligament of Treitz. Only when gastrointestinal bleeding is acute does it constitute an emergency. Chronic, low-volume blood loss is usually subclinical until such time as it presents with iron deficient anaemia.

Acute upper gastrointestinal haemorrhage (AUGIH)

Epidemiology

The incidence of AUGIH in the United Kingdom is approximately 1 per 1000 adults/year, of which 15 per cent of cases occur in patients already in hospital. The male incidence is twice that of the female in all age groups except the elderly, where they are similar. The annual incidence increases dramatically with age, rising to nearly 5 per 1000/year in the over 75 age group. In the United Kingdom in 1993, about one-quarter of cases occurred over the age of 80 years.

Aetiology

Many gastrointestinal lesions may result in haemorrhage but peptic ulcer is the most frequent in the United Kingdom (Table 1). Each diagnostic group has its own aetiological factors. *Helicobacter pylori* infection and ingestion of aspirin and non-steroidal anti-inflammatory drugs (NSAID) are important for ulcer disease and erosions of the upper gastrointestinal tract. Ulceration may also occur at the site of surgical enterostomies (stomal ulcers) and in association with the Zollinger–Ellison syndrome. Peptic ulceration at specific sites is associated with major haemorrhage due to the anatomical relation of major arteries—the posterior wall of the first part of the duodenum (gastroduodenal artery), the lesser curve of the stomach (left gastric artery), and posterior wall of stomach (splenic artery).

Liver disease due to alcohol and hepatitis are the principal causes in Western countries of portal hypertension, which leads to oesophageal varices; variceal bleeding may rarely affect the stomach and the remaining gastrointestinal tract. Mallory–Weiss tears are mucosal lesions at the oesophagogastric junction associated with profuse vomiting; haematemesis occurs which is usually minor and always self-limiting. The most common malignant causes of upper gastrointestinal haemorrhage are adenocarcinoma of the stomach and gastric lymphoma, but acute bleeding is an unusual presentation. Other causes include benign tumours (or those with uncertain malignant potential) such as the stromal tumours (previously described as leiomyomas) which may bleed when the mucosal surface ulcerates, angiodysplasia (and other vascular lesions), aortoduodenal fistula, haemobilia, and trauma.

Clinical features

Upper gastrointestinal haemorrhage usually presents with haematemesis or 'coffee-ground' vomiting and melaena. Frank haematemesis indicates a severe bleed. It is not always a feature but melaena will always follow a significant bleed. In rapid bleeding, symptoms of hypovolaemia may precede haematemesis or melaena. These include postural hypotension, syncope, shock, and even death. In most cases, the causative lesion will not be known until diagnostic endoscopy is undertaken. The patient should be asked about ingestion of NSAIDs and whether blood was present in the first vomit (it is usually absent in Mallory–Weiss tear). Signs of chronic liver disease may be present in patients with oesophageal varices but this does not confirm the cause of blood loss since peptic ulcer is a common synchronous lesion. Melaena is a clinical diagnosis made on the observation of black, tarry, offensive stool on rectal examination or passed spontaneously. In patients with rapid haemorrhage, usually accompanied by shock, fresh blood may be passed per rectum (haemochezia) and may thus be difficult to distinguish from lower gastrointestinal haemorrhage. A mix of fresh blood and melaena may indicate a lesion in the lower small intestine (e.g. Meckel's diverticulum).

Laboratory diagnosis

AUGIH is a clinical diagnosis. The initial haemoglobin estimation is not a useful indicator of the volume of blood lost until time for haemodilution has passed. The haemoglobin may be normal in a patient with a large, acute haemorrhage. Equally, the haemoglobin may be low in a patient with iron deficiency anaemia resulting from chronic haemorrhage who presents with a small, acute bleed. The haemoglobin and haematocrit after volume resuscitation are more useful. Platelet count and coagulation studies are important to exclude a bleeding disorder and are of particular relevance in patients receiving therapeutic anticoagulants and in those with liver disease.

Treatment

The management of acute upper gastrointestinal haemorrhage falls into four principal stages.

Assessment, resuscitation, and monitoring

Important aspects of assessment are confirmation that a bleed has occurred and the degree of hypovolaemic shock that has resulted. Resuscitation is as for any hypovolaemic patient with the immediate aim of rapidly replenishing blood. Tachycardia, vasoconstriction, sweating, hypotension (including a postural drop), tachypnoea, and a low central venous pressure all indicate hypovolaemia. Adequate resuscitation is evidenced by a normal pulse rate, blood pressure and central venous pressure, production of urine, and an improving level of consciousness. Central venous access and placement of a urinary catheter will help in the resuscitation and monitoring of the more severe cases and those with major cardiovascular and respiratory comorbidity. Once circulating blood volume has been restored, management should be aimed at monitoring the patient for continued or recurrent bleeding, replacing blood, making a diagnosis, and instituting therapy. Regular pulse, blood pressure, central venous pressure, and urine output will give a good guide. Fresh haematemesis obviously indicates further acute haemorrhage. The passage of further fresh melaena has to be interpreted in the light of the cardiovascular signs and repeated estimations of blood haemoglobin concentration.

Diagnosis and haemostasis

In most instances, the diagnosis is obscure until upper gastrointestinal endoscopy is undertaken. This diagnostic, and potentially therapeutic, procedure should be undertaken as soon as possible after resuscitation is complete. The aim of endoscopy is fourfold:

1. to make a diagnosis;
2. to assess the risk of further haemorrhage based upon the site, size, and nature of the lesion (including stigmata of recent haemorrhage);

3. to apply haemostatic therapy where appropriate;
4. to inform the surgeon as to the site of the lesion in cases requiring urgent surgery due to rapid, ongoing blood loss and to exclude varices in these cases.

Haemostasis occurs spontaneously in most cases. When bleeding continues, haemostasis can be achieved by endoscopic, surgical, or radiological means. Endoscopic haemostatic therapy may be given in the form of injection of adrenaline or sclerosants (e.g. polidocanol) either alone or in combination, or other substances such as fibrin glue. The application of heat energy in the form of laser (Argon or NdYAG), heater probe, or diathermy are also effective for peptic ulcer with active bleeding or visible vessels. There is good trial evidence that endoscopic therapy reduces the rate of rebleeding in these subgroups and one trial has shown a reduction in mortality. There is no randomized, controlled trial evidence that planned, repeated endoscopic therapy further reduces rebleeding in peptic ulcer. Repeated injection or banding have been shown to reduce rebleeding and mortality from oesophageal varices.

Surgery is indicated in massive, acute bleeding not amenable to endoscopic therapy or where endoscopic therapy fails to control active bleeding. Many units would attempt a second endoscopic therapy especially in young patients before resorting to surgery. There is some evidence that early surgical intervention in those over 60 is appropriate. There is no place for a third attempt at endoscopic therapy and surgery is indicated. In cases where endoscopic therapy has failed and surgery is deemed to be exceptionally high risk, visceral angiography may allow for embolization (e.g. the gastroduodenal artery in duodenal ulcer).

Uncontrolled variceal haemorrhage may be controlled with a Sengstaken–Blakemore tube as a temporary measure before more definitive treatment. Where endoscopic therapies subsequently fail, transjugular intrahepatic portosystemic shunt (TIPSS) is a minimally invasive method of creating a portosystemic shunt. Finally, oesophageal transection is occasionally life saving where all other attempts at haemostasis have failed.

Treatment of causative lesion

Treatment of the causative lesion should be started as soon as possible after diagnosis. There is, however, no evidence that drug therapy alters the outcome of the bleeding episode in terms of rebleeding or death. Histamine (H_2) receptor antagonists and proton pump inhibitors both heal ulcers but neither have been shown to affect the outcome of the bleeding episode. Where the causative lesion is a tumour (benign or malignant) then elective surgery may be indicated. Angiodysplasia can be treated with laser or argon beam. Specific treatments may be required for rarer causes such as tuberculosis.

Prevention of recurrence

Recurrent episodes of bleeding from peptic ulcers can be achieved by eradicating *H. pylori* infection and through the avoidance of ulcerogenic drugs. Persistent ulceration despite these measures may require long-term acid suppressive therapy and Zollinger–Ellison syndrome should be excluded. Variceal haemorrhage can be prevented through programmes of variceal eradication by injection or banding and also by TIPSS, or ultimately liver transplantation where indicated.

Prognosis

Prognosis depends on many factors including the severity of the bleed, the age of the patient, the associated comorbidity of the patient, the diagnostic category, the endoscopic features (stigmata of recent haemorrhage), and whether continued or recurrent bleeding is a feature. Overall, the crude mortality for patients presenting to emergency departments with acute upper gastrointestinal haemorrhage is about 10 per cent. Most deaths occur in the elderly and those with severe comorbidity. Death in those under the age of 60 with no comorbidity is very low (0.1 per cent) regardless of the severity of the haemorrhage. The factors that contribute to mortality have been combined in a prognostic risk score. This is represented in Table 2. The mortality associated with each risk score is represented in Table 3.

Acute lower gastrointestinal haemorrhage

Epidemiology

The incidence of lower gastrointestinal haemorrhage has not been well defined but it is common. Most (90 per cent) of acute lower gastrointestinal bleeds will stop spontaneously although 35 per cent will require blood transfusion, and 5 per cent will require urgent surgical intervention.

Aetiology

As in upper gastrointestinal haemorrhage, several pathological causes are responsible. Most causative lesions are colonic or anorectal and only 3 per cent originate in the small bowel. In the Western world, diverticular disease represents the largest proportion of cases (40 per cent), followed by inflammatory bowel disease (20 per cent, including Crohn's, ulcerative colitis, infectious colitis, and ischaemic colitis), neoplasia (15 per cent), benign anorectal disease (10 per cent), and arteriovenous malformations (2 per cent). Other lesions are rare and include radiation injury, Meckel's diverticulum, other small bowel pathology, and varices. Bleeding is not uncommonly associated with coagulopathy but studies have shown the distribution of causative lesions in these cases to be the same. In severe cases, however, generalized mucosal bleeding may occur.

Iatrogenic causes of haemorrhage include postpolypectomy bleeding and anastomotic bleeding. The risk of haemorrhage after polypectomy is estimated to be between 0.2 per cent and 3 per cent. Haemorrhage is usually immediate but may be delayed. When identified, endoscopic haemostatic techniques are usually successful (injection of adrenaline, resnaring, recoagulating, placement of a ligature or clip).

Acute colonic diverticular bleeding is common. The estimated risk of bleeding with this disease is about 15 per cent. After a single bleed, the risk of recurrence is 25 per cent and after two bleeds it is 50 per cent. Eighty per cent of all bleeds stop spontaneously and no therapy is indicated. Operative intervention should be considered after two major bleeds because the risk of further recurrence is high. However, many of these patients are frail and elderly and continuation of conservative treatment for multiple, self-limiting episodes may be appropriate. Inflammatory bowel disease often manifests itself as bloody diarrhoea but more rarely may present with profuse haemorrhage. This is more common in Crohn's disease than in ulcerative colitis because the inflammation involves the whole thickness of the bowel wall. Up to 6 per cent of patients with this disease may sustain a major haemorrhage. About 50 per cent stop bleeding spontaneously but of these 35 per cent will rebleed. For this reason, urgent surgery is usually indicated for patients with a life-threatening haemorrhage as a result of inflammatory colitis. The operation usually required is a total colectomy. The rectum is usually preserved at this stage unless this is the site of major haemorrhage. Ischaemic colitis rarely causes severe haemorrhage. Bloody diarrhoea is more usual and may be accompanied by pain.

Benign and malignant colonic tumours may present as profuse bleeding although occult blood loss and minor fresh bleeding is more common. Rarely is urgent surgical intervention required. Vascular anomalies occur with increasing frequency with age. They may originate from chronic, partial venous obstruction of submucosal veins due to incompetence of the precapillary sphincters and arteriovenous malformations. These lesions are usually multiple and are most frequent in the caecum and ascending colon. Bleeding is usually slow, intermittent, and recurrent although once again it is occasionally massive (2–15 per cent). Most (90 per cent) stop spontaneously but 25 to 85 per cent will recur. The treatment of choice is endoscopic coagulation if the lesions can be identified. Colectomy is reserved for those with repeated major haemorrhage.

Benign anorectal disease does present as lower gastrointestinal haemorrhage and a careful examination of the anorectum is imperative before initiating more invasive examinations. However, anorectal lesions are common and complete colonic evaluation is usually required even after identifying an anorectal source such as haemorrhoids.

Diagnosis and treatment

A good history from the patient may give clues as to the cause of colorectal haemorrhage. Important points include a prior history of bleeding, the presence of liver disease, and drug usage (aspirin, non-steroidal anti-inflammatory drugs, and warfarin) as well as the exact nature of the bleeding—specifically the duration, the colour of the blood, the relationship to defaecation, whether the blood is mixed with or separate from the stool, an associated change in bowel habit, or mucus discharge. Bright red blood separate from the stool suggests an anorectal cause. Diarrhoea and mucous associated with darker blood mixed in with the stool suggests colitis or neoplasm. None of these clinical features, however, is absolutely diagnostic.

Resuscitation measures are as for bleeding from the upper gastrointestinal tract. However, since most lower gastrointestinal bleeds stop spontaneously, initial

management should be conservative with transfusion and correction of clotting abnormalities. Once haemorrhage has ceased, bowel preparation and colonoscopy can be undertaken in a stable patient and with a much higher chance of detecting the pathological lesion (85–90 per cent).

In the small proportion of patients in whom active colonic bleeding continues, investigation to localize the source of the haemorrhage is indicated so that directed treatment can be administered in the form of endoscopic therapy, interventional radiology, or surgery. Colonoscopy is favoured by many clinicians but the use of bowel preparation is still debated. Some favour the use of a purgative together with distal colonic washouts before endoscopy, whilst other authors have argued that this is unnecessary because of the cathartic effect of blood in the colon. In one study, the causative lesion was identified in three-quarters of patients without preparation. This compares favourably with studies which have used mechanical preparation methods. Colonoscopy should be abandoned if massive haemorrhage obscures the diagnosis or severe mucosal or ischaemic colitis is encountered, as the risk of perforation in these cases is high.

Nuclear scintigraphy can be used to detect active haemorrhage. It is very sensitive and can detect bleeding rates of less than 1 ml/min. Tc99m labelled sulphur colloid can be used, which has the advantage of no preparation but the half-life is very short and its rapid enhancement of the liver and spleen can obscure the diagnosis. A better method is the use of Tc99m labelled red cells. Unfortunately, although sensitive, it is also very non-specific and localizes the lesion very poorly. It may be useful immediately before angiography to confirm active haemorrhage before undertaking the more invasive procedure. Whenever there is massive, active haemorrhage, however, this is unnecessary and the patient should proceed directly to visceral angiography.

Selective mesenteric angiography can also detect a rate of bleeding of 0.5 to 1.0 ml/min. The sensitivity reported in various studies ranges from 40 to 86 per cent. Once the site of haemorrhage is identified, the patient can proceed directly to surgery or there is the therapeutic possibility of arterial infusion of vasopressin or selective embolization. Vasopressin infusion has considerable side-effects including mesenteric thrombosis, intestinal infarction, myocardial ischaemia, hypertension, arrythmias, and death. Nitroglycerin may be infused simultaneously to counteract the systemic effects of the drug. There is a wide range in the reported rate of initial control of the haemorrhage, and the rebleeding rate is high (22–71 per cent).

Selective embolization using coil springs or gelfoam into the most distal vessel results in high initial rates of haemostasis and the rate of intestinal infarction is low. It is a good technique for patients with a very high predicted operative mortality.

From a practical aspect, investigation should start with examination of the abdomen and rectum, followed by proctosigmoidoscopy. If there is any suspicion of an upper gastrointestinal cause, then upper gastrointestinal endoscopy is recommended. Colonoscopy is the investigation of choice to evaluate the colon and terminal ileum. Where this fails, visceral angiography should be undertaken, which may allow the site of haemorrhage to be identified and treatment instituted.

In about 5 per cent of cases, the source of bleeding remains obscure. Additional investigations may include small bowel enteroscopy or laparotomy with on-table enteroscopy.

Further reading

Northfield TC, Smith T (1970). Central venous pressure in the clinical management of acute gastrointestinal bleeding. *Lancet* **1**, 990–1.

Palmer KR, Church NI (1999). Therapeutic endoscopy for upper gastrointestinal bleeding. *Continued Medical Education Journal Gastroenterology, Hepatology and Nutrition* **2**, 75–8.

Peterson WL, Cook DJ (1998). Antisecretory therapy for bleeding peptic ulcer. *Journal of the American Medical Association* **280**, 877–9.

Rockall TA, Logan RFA, Devlin HB, Northfield TC (1995). Incidence of and mortality from acute upper gastrointestinal haemorrhage in the United Kingdom. *British Medical Journal* **311**, 222–6.

Rockall TA, Logan RFA, Devlin HB, Northfield TC (1996). Risk assessment following acute upper gastrointestinal haemorrhage. *Gut* **38**, 316–21.

Steele RJC (1989). Endoscopic haemostasis for non-variceal upper gastrointestinal haemorrhage. *British Journal of Surgery* **76**, 219–25.

Swain CP, Kirkham JS, Salmon PR, Bown SG, Northfield TC (1986). Controlled trial of Nd-YAG laser photocoagulation in bleeding peptic ulcers. *Lancet* **i**, 1113–6.

Vernava AM, Moore BA, Longo WE, Johnson FE (1997). Lower gastrointestinal bleeding. *Diseases of the Colon and Rectum* **40**, 846–58.

Williams SG, Westaby D (1994). Management of variceal haemorrhage. *British Medical Journal* **308**, 1213–7.

14.4 Immune disorders of the gastrointestinal tract

M. R. Haeney

Introduction

The gut is protected by several mechanisms. The intercellular tight junctions of the epithelial cells form an important physical barrier; these cells turn over every 24 to 96 h. Any injury to the epithelial barrier results in rapid migration of adjacent viable epithelial cells to cover the denuded area, a process called 'restitution', while lymphocytes and macrophages migrate through pores in the basement membrane to provide temporary protection. The acid pH of the stomach and the proteolytic enzyme content of the intestine are formidable chemical barriers to many organisms. A change in the normal microflora of the intestine or impaired gut motility may allow pathogenic bacteria to flourish. Microbial antigens that resist these defences and penetrate the epithelial surface encounter the mucosal immune system.

Functional morphology of the gut-associated lymphoid tissue

Lymphocytes are found at three sites within the mucosa (Fig. 1):

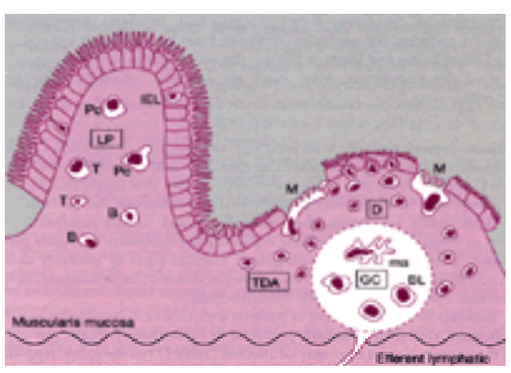

Fig. 1 Organization and structure of gut-associated lymphoid tissue. On the left, T and B lymphocytes and plasma cells (PC) can be seen in the lamina propria, with intraepithelial lymphocytes (IEL) between the columnar epithelial cells. On the right, there is a Peyer's patch covered by cuboidal epithelium with occasional 'M'cells. The Peyer's patch comprises three areas: the dome (D) of T and B lymphocytes; the thymus-dependent area (TDA); and the germinal centre (GC) containing macrophages (Ma) and B lymphoblasts (BL).

1. organized lymphoid aggregates (Peyer's patches) beneath the epithelium of the terminal small intestine;
2. lymphocytes within the epithelial cell layer (intraepithelial lymphocytes);and
3. lymphocytes scattered among other immunocompetent cells within the lamina propria.

Gut-associated lymphoid tissue is divided into two functional compartments: an afferent arm—Peyer's patches—where interaction occurs between luminal antigens and the immune system, and an effector arm—the lymphocytes of the intraepithelium and lamina propria.

Peyer's patches

These are covered by a specialized epithelium (follicle-associated epithelium) that has no microvilli but whose surface seems wrinkled or folded under the scanning electron microscope (Fig. 1). These microfold, or M, cells sample and transport particulate antigens from the lumen into the 'dome' area, where T and B cells mix freely with the microfolds of the M cells and priming of both types of lymphocyte occurs. Within Peyer's patches are specialized T cells that induce immature IgM-bearing B lymphocytes to switch isotype to IgA.

Lymphocytes are mobile: an array of cell surface receptors permits adhesion to endothelial cells and to components of the extracellular matrix. Primed B lymphoblasts, committed mainly to producing IgA antibody, migrate from Peyer's patches, via the lymphatics and mesenteric lymph nodes, to the thoracic duct and hence into the circulation. These cells return preferentially to the lamina propria, a process known as 'homing'. Once back in the gut, they mature into IgA plasma cells and are responsible for local and secretory antibody defences. The number of IgA-containing cells in the lamina propria far exceeds the numbers containing IgM, IgG, or IgE.

Intraepithelial lymphocytes

There is a similar migration pathway for T lymphocytes whereby T blasts from mesenteric nodes 'home' both to the epithelium and to the lamina propria. Intraepithelial lymphocytes are phenotypically and functionally distinct from peripheral blood lymphocytes. Peripheral T cells rarely express the human mucosal lymphocyte antigen (HML-1) but nearly all intraepithelial lymphocytes do. Human mucosal lymphocyte antigen CD103 is a novel aEb7 integrin which binds to its ligand, E-cadherin, expressed on gut epithelial cells. This interaction may direct homing of intraepithelial lymphocytes to the epithelium but is more likely to selectively retain intraepithelial lymphocytes within the epithelial compartment. Intraepithelial lymphocytes are not a homogeneous population: about 10 per cent do not express the CD3 antigen and therefore are not T cells. About 70 per cent are CD8+ and show increased expression of the g/† form of the T-cell receptor compared with peripheral blood lymphocytes (see Chapter 5.1). In experimental models, some intraepithelial lymphocytes are cytotoxic and some have natural killer activity, functions important in the control of enterovirus infection. Intraepithelial lymphocytes also seem to have a role in controlling the cell barrier function of epithelial cells, i.e. 'restitution'. However, the function of intraepithelial lymphocytes in humans is unclear.

Lamina propria lymphocytes

Large numbers of lymphocytes, natural killer cells, mast cells, macrophages, and plasma cells occur in the lamina propria. T and B lymphocytes are both found, but T cells predominate in a ratio of about 4:1. In contrast to intraepithelial lymphocytes, 80 per cent of these T cells are CD4+. They do not proliferate well after stimulation of the T-cell receptor, yet produce large amounts of cytokines interleukin 2, interleukin 4, interferon-g, and tumour necrosis factor-a. T-cell homing to the lamina propria is determined mainly by the integrin a4b7 on primed cells interacting preferentially with mucosal addressin cell adhesion molecule 1 expressed on the

microvascular endothelium in the lamina propria.

Secretory immunoglobulins

The plasma cells of the lamina propria secrete mainly IgA, which is specially adapted for its function. IgA is synthesized as a dimer with two IgA molecules linked by a smaller 'joining' peptide (the J chain), also produced by the plasma cells. The secretory component is a 70 kDa fragment of the polymeric immunoglobulin receptor synthesized by epithelial cells and is essential for transport of secretory IgA into the gut lumen. The polymeric Ig receptor binds the dimeric IgA; the complex is endocytosed and transported through the cytoplasm to the luminal surface of the cell where proteolysis of the receptor occurs. The IgA dimer is released into the gut attached to the proteolytic fragment of the receptor now called secretory component. Secretory component also protects the IgA molecule from degradation by proteolytic enzymes.

Secretory IgA predominates in the saliva and in gastric and intestinal secretions, where it tends to be concentrated in the mucous layer overlying epithelial cells. Secretory IgA neutralizes viruses, bacteria, and toxins and prevents the adherence of pathogenic micro-organisms to gut epithelium and so blocks the uptake of antigen into the systemic immune system.

Spectrum of intestinal immune responses

Ingestion of antigens can lead to local immunity, a systemic immune response, or a state of specific immune unresponsiveness (tolerance).

Local immune responses

These can occur independently of a systemic response. For example, immunization against poliomyelitis with oral Sabin vaccine gives better protection than the injected Salk vaccine, even though both induce serum antibodies. Local IgA antibody, produced in response to the oral vaccine, partly blocks uptake of pathogenic virus into the circulation.

Systemic immune responses

Macromolecules are absorbed by the intestine into the portal or systemic circulations, via either the glandular epithelium covering the villus or the M cells. Up to 2 per cent of a dietary protein load appears antigenically intact in the circulation. Sinusoidal phagocytes (Kupffer cells) of the liver destroy much of the antigen but enough passes through the liver to stimulate systemic antibody production, particularly in the spleen. Antibody formed in the spleen enters the portal circulation to complex with incoming antigen. Circulating immune complexes of IgA and dietary antigens are regularly found in normal people after meals.

Systemic tolerance

A unique feature of the mucosal immune system is its ability to downregulate immune responses to dietary antigens (oral tolerance). Native Americans knew that eating the leaves of poison ivy prevented contact dermatitis on subsequent exposure to the plant. This observation can be reproduced in animals by feeding them antigen; they become immunologically unresponsive (tolerant) to subsequent parenteral injections of that antigen. Oral tolerance can affect all aspects of the systemic immune response: a single feed of protein antigen suppresses systemic IgM, IgG, and IgE responses as well as T-cell mediated immunity. This has led to attempts to treat autoimmune diseases by feeding autoantigens to patients.

Immunological disorders of the gastrointestinal tract

Normally, the intestinal immune system steers a delicate course between the undesirable extremes of immunological incompetence, with resulting vulnerability to ingested pathogens (for instance, the gastrointestinal consequences of primary and secondary immunodeficiencies), and hypersensitivity to dietary antigens, with immunologically mediated reactions each time that antigen is eaten.

Primary immunodeficiency diseases

Immunocompromised patients are at risk from two sources of infection: common pathogens, which invade even the immunologically healthy, and 'opportunistic' agents that can invade and infect only those with weakened defences. In the compromised host, most infections are due to common pathogens that are readily identified and controlled. The difficult problems arise from opportunistic infections because these often elude isolation, may not respond to available drugs, and carry a high fatality. Indeed, the identification of certain opportunistic infections implies an underlying immunodeficiency that demands further investigation.

It is beyond the scope of this section to deal with all the gastrointestinal complications of every known form of primary and secondary immunodeficiency. Instead, attention will be focused on representative disorders.

Common variable immunodeficiency (CVI)

Common variable immunodeficiency is an example of one of the primary antibody deficiency syndromes described in Section 5.

Definition

Common variable immunodeficiency is a heterogeneous group of disorders characterized by low serum immunoglobulin levels, a normal or low proportion of circulating B lymphocytes and, in about one-third of patients, impaired cell-mediated immunity. It can present at any age and, in the United States and Western Europe, the prevalence is about 40 per million of the population. Most cases are sporadic, although some are inherited.

Clinical features

Patients present typically with recurrent sinopulmonary infections, most frequently caused by pneumococci, streptoccoci, and *Haemophilus influenzae.* Less commonly, they present with arthropathy, skin sepsis, meningitis, osteomyelitis, or other severe systemic bacterial infections (see Section 7).

With certain exceptions, these patients are not unduly susceptible to viral or fungal infections, because cell-mediated immunity is usually preserved. There are rarely any diagnostic physical signs of antibody deficiency, although examination often shows evidence of the consequences of previous infections, particularly bronchiectasis.

Between 30 and 50 per cent of patients with common variable immunodeficiency have gastrointestinal problems at some time. Virtually any part of the gastrointestinal tract may be affected (Table 1) but the most common symptoms are diarrhoea (intermittent or chronic) and weight loss. An approach to the diagnosis of these complications is shown in Fig. 2.

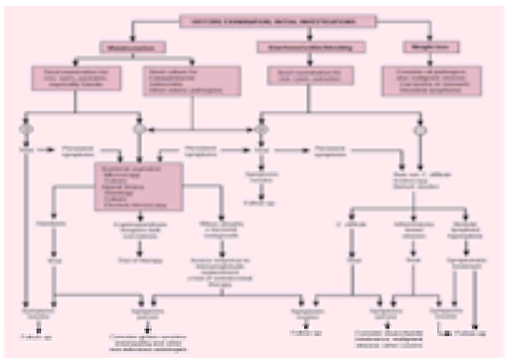

Fig. 2 A scheme for the investigation of gastrointestinal complications in patients with common variable immunodeficiency. (Reproduced from Haeney MR (1989). Gastrointestinal disease in the immunocompromised host. In: Turnberg LA, ed. *Clinical gastroenterology* by permission of Blackwell Science, Oxford.)

Stomach

Achlorhydria and pernicious anaemia

Achlorhydria is found in about 30 per cent of patients and the associated atrophic gastritis occasionally leads to a syndrome resembling pernicious anaemia except that the atrophic gastritis involves the whole stomach without antral sparing, the serum gastrin concentrations remain normal, and autoantibodies to gastric parietal cells and intrinsic factor are absent.

Gastric cancer

Patients with common variable immunodeficiency have a 47-fold increase in the incidence of carcinoma of the stomach. It is sufficiently common to warrant yearly gastroscopic examination in hypogammaglobulinaemic patients who have atrophic gastritis. The high concentrations of microbial enzymes and nitrites found in the gastric juices may lead to local production of carcinogenic *N*-nitroso compounds. Chronic active gastritis induced by *Heliobacter pylori* and overexpression of p53 may also play a role in gastric carcinogenesis.

Small intestine

Infective complications

Although infestation with *Giardia lamblia* is the most common identifiable cause of malabsorption, in many patients the cause is never found. Giardiasis is virtually confined to adults and is rarely seen in boys with X-linked hypogammaglobulinaemia. Giardiasis may also cause diarrhoea, villous abnormalities, vitamin B_{12} and folate malabsorption, steatorrhoea, disaccharidase deficiency, and protein-losing enteropathy but the pathogenetic mechanisms are poorly understood.

Examination of at least three consecutive fresh stool specimens is essential to detect the cysts of *G. lamblia* (Fig. 2). If this fails, duodenal aspiration and jejunal biopsy are needed to establish the diagnosis. In particularly difficult cases, a therapeutic trial of metronidazole can be useful, although infestation frequently recurs. Most patients show symptomatic improvement after treatment with either a 7-day course of metronidazole (2 g daily as a single dose) or mepacrin (100 mg, three times daily for 10 days). Other parasitic infestations occur. *Cryptosporidium* infection occasionally causes self-limiting diarrhoea but has a much more sinister outcome in boys with CD40 ligand deficiency (hyper IgM syndrome) and in patients with human immunodeficiency virus (**HIV**) infection.

Bacterial infections also cause diarrhoea in patients with common variable immunodeficiency and *Campylobacter jejuni* is frequently responsible. Rarely, campylobacter causes an ascending cholangitis and hepatitis. Treatment is a 2-week course of erythromycin (500 mg, four times daily) with follow-up stool culture to ensure that treatment has been effective.

Shigella or salmonella diarrhoea does not occur more commonly than normal. Similarly, while overgrowth of commensal bacteria is common, bacterial counts rarely exceed 10^5 organisms/ml, compared with counts of more than 10^6/ml in the blind-loop syndrome. Nevertheless, it is common practice to treat these patients empirically with tetracycline and metronidazole, often with symptomatic improvement.

Nodular lymphoid hyperplasia

Nodular lymphoid hyperplasia describes the presence of lymphoid nodules in the lamina propria of the gut. Although described in many disorders and occasionally in healthy individuals, nodular lymphoid hyperplasia should make the clinician suspect common variable immunodeficiency. It occurs in 20 to 50 per cent of patients but is not necessarily symptomatic. The nodules, which are 1 to 3 mm in diameter, appear as protrusions on fibreoptic endoscopy (Fig. 3 and Plate 1) and as multiple filling defects on barium studies (Fig. 4). Nodular lymphoid hyperplasia restricted to the rectum or colon can present with rectal bleeding, abdominal pain and features of intestinal obstruction, but rarely with diarrhoea.

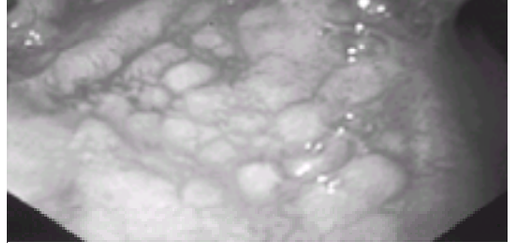

Fig. 3 The appearance of nodular lymphoid hyperplasia on upper gastrointestinal endoscopy. (See also Plate 1.)

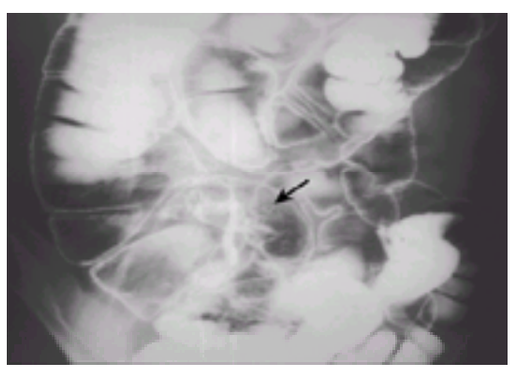

Fig. 4 A double-contrast barium enema showing nodular lymphoid hyperplasia in the terminal ileum (arrowed).

The ultrastructure of these nodules is similar to Peyer's patches, and lymphoblasts containing IgM are found in the centres of the follicles. The condition probably represents hypertrophy of the gut-associated lymphoid tissue in response to antigens in the gut lumen. In one series of nodular lymphoid hyperplasia in individuals with normal serum immunoglobulins, every patient had intestinal giardiasis, suggesting an aetiological link with persistent infestation. Although nodular lymphoid hyperplasia is not premalignant in patients with hypogammaglobulinaemia, intestinal lymphoma has been reported in apparently immunocompetent subjects with extensive small-bowel nodular lymphoid hyperplasia.

Hypogammaglobulinaemic sprue

In a few patients with unexplained diarrhoea, the mucosal lesion resembles coeliac disease or tropical sprue but with reduced or undetectable plasma cells within the lamina propria. In tropical regions, about 1 per cent of patients with 'sprue' may be suffering from a primary humoral immunodeficiency syndrome. Malabsorption in patients with hypogammaglobulinaemic sprue can improve rapidly after replacement immunoglobulin therapy.

Although extremely rare, patients with common variable immunodeficiency may have concomitant gluten-sensitive coeliac disease.

Inflammatory bowel disease

About 5 per cent of patients with common variable immunodeficiency have features of inflammatory bowel disease with radiological and histological findings of Crohn's disease. Others have proposed a specific common variable immunodeficiency enteropathy characterized by low-grade microscopic colitis, increased intraepithelial lymphocytes, and an intact crypt architecture with a good response to an elemental diet.

Non-granulomatous jejunoileitis

This is a rare feature of common variable immunodeficiency and has a poor prognosis.

Management

The cornerstone of treatment of antibody deficiency is immunoglobulin replacement; enough must be given to prevent further infections and reduce the incidence of complications. Intravenous or subcutaneous immunoglobulin therapy is the treatment of choice and is discussed more fully in Section 5.

Antibody-deficient patients respond as promptly as others to appropriate antibiotics but longer courses of treatment are usually needed to ensure complete eradication of the micro-organism.

Selective IgA deficiency (see also Section 5)

Definition

Selective IgA deficiency refers to a serum IgA concentration below the limit of detection (< 0.01 g/l). By definition, the serum IgG and IgM concentrations are normal.

Aetiology

Selective IgA deficiency is common and occurs in about 1 in 700 of healthy adults. Most cases are sporadic, but there is an association with inheritance of the HLA B8, DW3 haplotype, and with deficiencies of IgG2 and IgG4. It is sometimes linked with defects in chromosome 18, particularly in the autosomal recessive syndrome of ataxia telangiectasia. Selective IgA deficiency may also be due to drugs such as phenytoin or penicillamine.

Clinical features

Although selective IgA deficiency is associated with a range of disorders, most IgA-deficient individuals are asymptomatic, possibly because IgM-producing cells provide high local concentrations of IgM antibody or because symptomatic individuals are those who also have deficiency of IgG2 antibodies to polysaccharide antigens.

Gastrointestinal complications (Table 2)

Pernicious anaemia

Selective IgA deficiency is associated with pernicious anaemia. Unlike common variable immunodeficiency, the anaemia conforms to the classical Addisonian type in that atrophic gastritis and raised serum gastrin levels occur.

Malabsorption and steatorrhoea

IgA deficiency occurs in about 1 in 40 of patients with coeliac disease, over 15 times more frequently than in the general population. Patients with selective IgA deficiency and a flat jejunal mucosa respond to dietary gluten withdrawal in a way typical of classical coeliac disease.

Antibodies to dietary antigens

Secretory IgA helps prevent absorption of food antigens through the intestinal mucosa and there is a high prevalence of serum antibodies to food proteins in patients with selective IgA deficiency. For instance, about a third of IgA-deficient blood donors have serum antibodies to milk compared with 0.3 per cent of healthy controls. IgA-deficient subjects also tend to have autoantibodies to antigens such as collagen and IgA itself (see below).

Gastrointestinal infection

With the exception of *G. lamblia* infestation, other infections rarely persist. Even giardiasis is far less frequent than in common variable immunodeficiency. However, in the past, IgA-deficient patients were prone to develop chronic diarrhoea and malabsorption after truncal vagotomy and gastroenterostomy for duodenal ulceration. This was due to overgrowth of commensal bacteria in the upper intestinal tract, presumably because of the combined effects of deficiency of local antibody production, achlorhydria, and impaired gastrointestinal motility.

Inflammatory bowel disease

Crohn's disease and ulcerative colitis occur in patients with IgA deficiency but their frequency is difficult to judge from the widely varying published reports.

Malignant disease

Oesophageal, gastric, and colonic neoplasms have been reported but it is not certain whether the risk of malignancy is truly increased.

Management

Patients with selective IgA deficiency rarely warrant immunoglobulin replacement therapy, unless IgG2 deficiency is also present. Antibodies to IgA develop in about a third of patients with selective IgA deficiency: high titres of antibodies may cause severe reactions to plasma or blood transfusions or even the trace amounts of IgA present in intravenous immunoglobulin preparations.

Other types of primary immunodeficiency

Gastrointestinal problems occur in other types of immunodeficiency (Table 3) (see Section 5). These conditions are much rarer than primary antibody deficiency. Most defects involving cell-mediated immunity present within the first 6 months of life. Infants with severe combined immunodeficiency, for example, grow and develop normally for a few months but then fail to thrive, frequently with a clinical triad of pneumonia, mucocutaneous candidiasis, and intractable diarrhoea caused by one or more of a range of micro-organisms. Some disorders are associated with unusual gastrointestinal features (Table 3).

Secondary immunodeficiency

Secondary immunodeficiency describes conditions in which the immune defect results from underlying disease and is far more common than primary immunodeficiency. In many cases, the secondary immunodeficiency is of minor relevance to the clinical picture but occasionally its severity may mask the underlying condition. AIDS is a florid example of the gastrointestinal complications seen in patients with secondary defects predominantly involving cell-mediated immunity.

HIV and the gastrointestinal tract

The gastrointestinal tract is a major target organ in HIV infection and AIDS, irrespective of the route of acquisition of the infection. Breast feeding can transmit HIV in humans, implying that the intestine is also an important portal of entry for the virus. In a simian model, severe depletion of CD4+ lymphocytes in the lamina propria and of intraepithelial lymphocytes occurs during primary infection and persists throughout the course of the infection. These dynamic changes in intestinal T lymphocytes are more severe than those seen in blood or peripheral lymph nodes. About half of patients with HIV infection will have gastrointestinal involvement at some time, and any level of the tract, from mouth to anus, can be involved (see also Section 7).

There are three main mechanisms in the pathogenesis of gastrointestinal disease: direct infection of enterocytes by HIV; opportunistic and other infections; and opportunistic tumours (Fig. 5 and Fig. 6). A major change in the small intestine is a partial villous atrophy, detectable early in the natural history of HIV infection. Enteropathogens causing intestinal infections are of the same types as in immunocompetent subjects but the infections are much more aggressive and invasive, and elicit little host immune response, so familiar symptoms and signs may be absent. A systematic and thorough search for likely pathogens is essential (Fig. 6). Multiple infections and tumours may coexist, so the organism isolated is not necessarily the cause of the symptoms. Colonic complications increase in frequency as immunodeficiency worsens. Clinically, patients experience diarrhoea, intestinal bleeding, and abdominal pain. Toxic megacolon, intussusception, idiopathic colonic ulceration, and pneumatosis intestinalis have also been described.

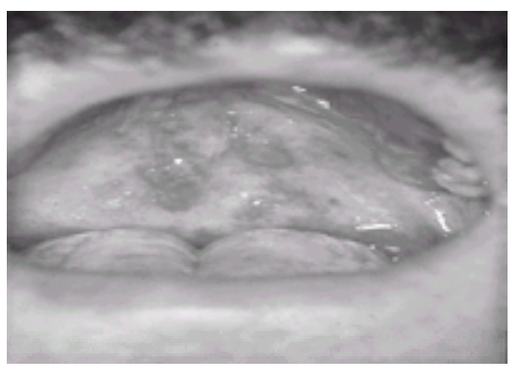

Fig. 5 Kaposi's sarcoma of the oral cavity in a patient with AIDS.

Fig. 6 Gastrointestinal consequences of HIV infection. (Redrawn from Chapel HM *et al.* (1999). *Essentials of clinical immunology*, 4th edn, by permission of the authors and Blackwell Science, Oxford.)

The clinical features of HIV infection and AIDS are discussed in detail in Chapter 7.10.21.

Immunodeficiency secondary to gastrointestinal disease

A low serum IgG concentration may be due to increased intestinal loss of immunoglobulin. A useful clue is a low serum albumin because there are no known conditions where immunoglobulin is selectively lost from the gut. The major causes of protein losing enteropathy are discussed in Chapter 14.22.

Intestinal lymphangiectasia (see also Chapter 14.15)

This immunodeficiency results from increased loss of lymphatic fluid containing immunoglobulins and lymphocytes. There is a selective loss of naive T lymphocytes expressing CD4/CD45RA. The basic defect is an abnormal dilatation of the lymphatic vessels in the intestine. There is a primary familial form in children, who present with diarrhoea, malabsorption, and growth retardation. Such children may have abnormal lymphatics elsewhere in the body causing chylous ascites, pleural effusions, and localized areas of oedema. The condition may also occur secondarily to lymphatic obstruction, for example due to intestinal lymphoma or constrictive pericarditis (see Section 15 and Section 22). The diagnosis should be suspected when there is T-cell lymphopenia, hypoalbuminaemia, and hypogammaglobulinaemia. The diagnosis is confirmed by finding dilated lymphatics in a jejunal biopsy (Fig. 7). The primary form of the disease responds well to a low-fat diet with additional medium-chain triglycerides. In secondary forms, correction of the underlying disease process is needed.

Fig. 7 A jejunal biopsy from a patient with intestinal lymphangiectasia showing dilated central lacteals.

Food allergy and intolerance

Introduction

Food allergy is one of the most controversial topics in medicine. It undoubtedly exists but extravagant claims that a staggering array of symptoms are due to food 'allergy' have confused the subject. Such claims are too rarely supported by objective, scientific observations and have provoked a sceptical response from many doctors. The major cause of confusion lies in the lack of agreement on definitions and diagnostic criteria.

Definition

Food allergy refers to a form of exaggerated reactivity (hypersensitivity) of the immune system to an ingested antigen. The term should be used only when the abnormal reaction is proved to be immunologically mediated, either by IgE or some other immune mechanism (Table 4). The term food intolerance should be used to describe all abnormal, reproducible reactions to food when the causative mechanism is unknown or is non-immunological. Food allergy and intolerance must be distinguished from food fads and psychological aversion to foods.

Aetiology

Food allergy

Although the gut provides a physical barrier to the antigen load in the lumen, up to 2 per cent of a protein meal can appear antigenically intact in the circulation. This was shown by injecting serum from a patient with known sensitivity to fish into the skin of a normal subject. A wheal and flare response at the skin test site (positive Prausnitz–Kustner reaction) was observed shortly after the normal subject ate the appropriate antigen, showing that this must have crossed the gut and triggered IgE-sensitized mast cells at the skin test site.

Atopic individuals have a higher prevalence of food allergy. The allergic components of foods are mainly glycoproteins with molecular weights between 10 and 70 kDa and most are heat stable and resistant to proteolysis, with the exception of those causing the oral allergy syndrome (see below).

In some forms of food allergy, damage involves immune mechanisms other than IgE. For instance, in coeliac disease there is strong evidence that exaggerated local T-cell mediated reactivity to dietary gluten causes the villous atrophy.

Food intolerance

Non-immunological mechanisms of reproducible, adverse reactions to food are much more common and include irritant, toxic, pharmacological, or metabolic effects of foods, enzyme deficiencies, or even the release of substances produced by fermentation of food residues in the bowel. Some foods contain pharmacologically active substances (such as tyramine or phenylethylamine) that act directly on blood vessels in sensitive subjects to produce migraine. Traces of drugs, food additives (for example monosodium glutamate), colouring agents (for example tartrazine), or preservatives (for example benzoic acid) can also cause symptoms in susceptible people by mechanisms which are ill understood, but are probably due to direct effects on mast cells.

Prevalence

The general public perceives food allergy to be a major health problem but epidemiological studies do not support this view. Food allergy affects 2 to 5 per cent of children under 5 years but only 1 per cent of adults. In one survey of 7500 households in the United Kingdom, about 20 per cent of the sample reported a food intolerance but this was confirmed by double-blind, placebo-controlled food challenge in only 1.4 per cent. However, the prevalence of peanut allergy appears to have increased significantly over the last 20 years.

Reactions to food additives, while they exist, are also not as common as most people believe. In a second population study in the United Kingdom, 7.4 per cent had symptoms suggestive of intolerance to food additives; further clinical assessment and additive challenge showed a true prevalence of 0.01 to 0.23 per cent.

Clinical features

Food reactions can be early or late, confined to the gastrointestinal tract or occur at sites remote from the gut (Fig. 8).

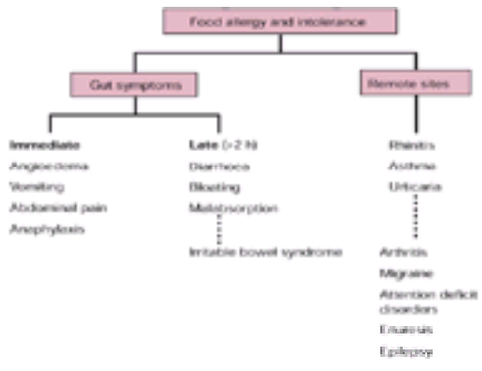

Fig. 8 Clinical spectrum of food allergy and intolerance. (Redrawn from Chapel HM *et al.* (1999). *Essentials of clinical immunology*, 4th edn, by permission of the authors and Blackwell Science, Oxford.)

Gut-related symptoms

About 75 per cent of young children, but only 10 per cent of adults, present with local gastrointestinal symptoms of food intolerance.

Early reactions

These are often 'immediate' in onset, occurring within minutes or up to 2 h after ingestion, recur on challenge testing, and include, apart from gastrointestinal disturbances, such features as perioral rash, angioedema of the lips or tongue, tingling of the throat, urticaria, asthma, or even anaphylaxis. Such acute and severe allergic reactions are mostly due to IgE antibodies to foods and are the least controversial form of food allergy. They are fairly easy to diagnose and the offending food is readily identified, usually by the parent: although any food may be responsible, 90 per cent of reactions are caused by cows' milk (in infants), peanuts, tree nuts, eggs, fish, and shellfish. Peanut is the most allergenic food known and peanut allergy is lifelong.

In some cases, anaphylaxis only occurs when the food is eaten 2 to 4 h before exercise, so-called food-dependent, exercise-induced anaphylaxis.

Allergy to latex rubber is increasingly common and several high-risk groups are recognized, notably patients with spina bifida or multiple urological procedures, and health care workers. Latex allergy may crossreact with plant defence proteins, called chitinases, in foods, typically melon, banana, avocado, chestnut, celery, passion fruit, or peach.

Oral allergy syndrome describes itching and lip swelling without involvement of other target organs. It occurs in patients with pollen allergy and is due to crossreactivity with epitopes in fresh (but not cooked) fruits and vegetables such as apples, carrots, hazelnuts, kiwi fruit, or raw potatoes.

Allergic eosinophilic gastroenteritis is characterized by intolerance to multiple foods, eosinophilic infiltration of the stomach and small intestine, peripheral eosinophilia, and positive skin prick tests and radioallergosorbent tests (see below) to foods.

Late reactions

Symptoms occurring over 2 h after food ingestion, such as diarrhoea, bloating, or a fatty stool are suggestive of food intolerance, if not allergy. Features of the irritable bowel syndrome (see below) may be accompanied by allergic symptoms elsewhere but usually occur in isolation and without any evidence of an immunological reaction.

Remote symptoms

Some patients with acute, IgE-mediated reactions to foods also experience rhinitis, asthma, urticaria, angioedema, or eczema. However, eating the implicated foods does not always cause these remote systems. Sneezing bouts, blocked nose, or asthma can also occur after taking wine or other alcoholic drinks because of the irritant effect of sulphite preservatives or other components. This is not an immunological reaction. Many patients with atopic eczema find that certain foods provoke a transient red and blotchy rash but it is mainly in children that food makes eczema worse. Elimination diets rarely improve atopic eczema in adults.

What is more debatable is whether food intolerance plays any part in remote symptoms such as hyperactivity/attention deficit disorder, enuresis, or arthritis.

Specific syndromes of food allergy

Food allergy contributes to a number of common intestinal disorders where the immunological mechanisms are not IgE mediated.

Coeliac disease

The characteristic histological lesion in untreated cases of coeliac disease (see Chapter 14.9.4) is loss of normal villi and a marked increase in the numbers of CD8+ intraepithelial lymphocytes, particularly those expressing the g/† T-cell receptor. It is believed that HLA class II molecules on antigen-presenting cells expose processed peptides from ingested wheat gliadin to immunocompetent T cells. Gliadin-specific, HLA DQ2-restricted T cells have been isolated from small intestinal biopsies of coeliac patients. T-cell infiltration of the small bowel epithelium is seen within hours of gluten exposure and resolves on treatment with a gluten-free diet, strongly suggesting that intestinal damage is due to a T-cell mediated reaction to gluten.

Cows' milk protein enteropathy

Milk proteins can cause a malabsorption syndrome similar to coeliac disease. Cows' milk protein enteropathy in babies causes failure to thrive, diarrhoea, malabsorption, and even intestinal bleeding and colitis. Jejunal biopsies show villous atrophy and lymphocytic infiltration.

Symptoms disappear when cows' milk is removed from the diet. Reintroduction of cows' milk causes a recurrence of symptoms. After a viral gastrointestinal infection, cows' milk may also be poorly tolerated for a while because of a temporary inability to digest lactose. Thus, in babies and small children with chronic gastrointestinal symptoms and failure to gain weight, trials (under medical supervision) of milk exclusion are justified and the diagnosis can usually be confirmed by food challenge. Recovery often occurs within a few months.

Recognized syndromes of food intolerance

In some conditions, a relationship to foods can be convincingly demonstrated in a proportion of patients. Sometimes, symptoms are provoked by the known irritant, pharmacological, or metabolic effects of food.

Irritable bowel syndrome

Irritable bowel syndrome is a descriptive term for several conditions that produce a similar range of abdominal symptoms (see Chapter 14.13). Irritable bowel syndrome is characterized by alternating constipation and diarrhoea, abdominal bloating, and colicky pain. In a variant of this disorder, however, constipation predominates and gastrointestinal transit times are greatly increased. Most cases are unrelated to food intolerance but in a minority of patients—usually those with predominant diarrhoea, with some bloating and pain—a relationship to specific foods can be demonstrated. Some patients who improve on a restricted diet are able to identify certain foods, notably cereals and dairy products, that provoke symptoms when reintroduced. However, not all gastroenterologists are convinced of a causal relationship between food and irritable bowel syndrome.

Lactose intolerance (see Chapter 11.3)

Many adults cannot digest lactose because of a deficiency of the enzyme lactase. In them, undigested sugar is fermented in the lower bowel, causing diarrhoea and wind. Lactose intolerance is not common in Europeans but affects up to 90 per cent of adult Africans and Orientals. It can also occur as a transient result of gastroenteritis and even as a secondary effect of cows' milk protein intolerance. This can cause confusion in diagnosis unless a lactose challenge is performed separately from a cows' milk protein challenge.

Fructose intolerance

This is discussed in Chapter 11.3.

Miscellaneous syndromes

Migraine and headache

Coffee and coffee withdrawal can provoke migraine in susceptible people. Certain cheeses cause headaches in many people, probably due to their tyramine or

phenylethylamine content. Red wines, especially port, cause headaches in susceptible people because of their content of congeners.

Asthma

Foods preserved by sulphites, particularly white wine, dried fruit, and fruit salads in supermarkets and restaurants, sometimes provoke asthma by the release of sulphur dioxide.

Urticaria

While IgE-mediated food allergy causes acute urticaria and angioedema, allergy rarely induces chronic urticaria. However, food dyes and preservatives often trigger chronic urticaria in sensitive subjects.

Chinese restaurant syndrome

Monosodium glutamate, used to enhance flavour in food and found in large amounts in Chinese food, may cause a syndrome of chest pain, sweating, nausea, dizziness, and fainting in susceptible individuals. However, double-blind studies have not convincingly demonstrated that monosodium glutamate is the culprit.

Controversial issues

Behavioural problems in children

The belief that foods and food additives can induce behavioural problems, particularly attention deficit disorder, is a controversial one. A diet free of preservatives, salicylates, and artificial flavours has been claimed to benefit up to 70 per cent of such children but most well-designed, double blind, placebo controlled challenges have failed to support a causal link. Children with behavioural disorders may improve temporarily for a few weeks when given a diet avoiding food additives but this appears to be a placebo effect. Parents who suspect food additive intolerance in their child may insist on maintaining the child on a restrictive diet, even when dietary challenges prove negative. Extreme examples have been termed 'Munchausen's syndrome by proxy'.

Psychological distress in adults

Some patients with a multiplicity of vague and variable symptoms, such as unexplained fatigue and malaise, and disturbances of sleep, appetite, or libido turn to the diagnosis of food allergy as an explanation. Only a few have clearcut psychiatric illness, others inadvertently cause symptoms by overbreathing or by somatizing their psychological distress. Having made their own diagnosis of food allergy, they have difficulty in accepting they are not allergic to foods, even though their food aversions may have resulted in a dangerously inadequate diet. They will frequently seek out practitioners who are prepared to endorse their views, whether valid or not. Early diagnosis and sympathetic management is essential if unnecessary consultations and inappropriate allergy tests are to be avoided.

Diagnosis

Food allergy should not be diagnosed without clear indications, as needless dietary restrictions can seriously disrupt not only the patient's life but also the whole family and may occasionally cause malnutrition. No test can replace a careful clinical history and thorough examination to exclude other, sometimes more likely, causes of the patient's symptoms.

Skin tests and radioallergosorbent tests

Skin prick tests and radioallergosorbent tests for detecting serum IgE antibodies are positive in about 75 per cent of patients who have IgE-mediated, acute, early reactions to foods such as nuts, egg, or fish. Usually, the offending antigen is obvious from the clinical history and confirmatory tests are needed only if there is clinical doubt. In patients with late symptoms at sites remote from the gut (Fig. 6), skin and blood tests are notoriously unreliable for many reasons:

1. Foods, as antigen sources, are poorly standardized and contain multiple, ill-defined antigens.
2. The antigen content of the food will depend on whether it is raw or cooked.
3. Some foods cause non-specific ('irritant'), positive skin reactions.
4. Patients may have IgE antibodies but no symptoms.
5. Food reactions can be mediated by mechanisms other than IgE antibodies.

For most patients with suspected food intolerance, laboratory tests are of little diagnostic value.

Elimination diets and challenge tests

In the absence of reliable laboratory tests, elimination diets and food challenge form the basis of diagnosis. To minimize bias and suggestion, the relationship between food and symptoms should be established by a placebo controlled, double blind challenge under medical supervision. In some cases, for example in chronic urticaria or when the symptoms are mild and largely subjective, it may be necessary to repeat the challenge before accepting that the association is not simply coincidental. In several series, only a quarter of reported 'adverse reactions' can be confirmed by double blind challenge. Although these rules are simple to state, they are difficult to carry out in practice. However, the alternative is that of prolonged, unsupervised, dietary manipulation, usually self-imposed or inflicted by parents on their children, with the attendant risks.

Food challenges are not without risk: there is a danger of precipitating an anaphylactic reaction. This is well-recognized in children with relatively mild symptoms of food intolerance, who develop anaphylaxis when the food, often cows' milk, is reintroduced after a period of avoidance.

Bogus or unproven laboratory tests

The absence of reliable laboratory tests has led to the promotion of controversial 'alternative' tests: these are at best misleading and at worst dangerous. New diagnostic procedures, like new drugs, require scientific validation: they must be reliable and reproducible. When presented with coded, duplicate samples, some 'alternative' laboratories in the United Kingdom were unable reliably to identify food allergies in patients known to have them; they gave inconsistent results for paired samples from the same patient; they reported many allergies in non-allergic subjects; and they often gave dubious and risky dietary advice.

Provocation-neutralization testing

This has been critically evaluated by the Royal College of Physicians of London, the American College of Physicians, and the California Medical Association: these bodies concluded that reported studies were seriously flawed and that the method lacked scientific validity. Under double blind conditions, the response of patients to active and control injections appeared to be due to suggestion and chance.

Leucocytotoxic testing

This involves incubating a patient's leucocytes with various food extracts and inspecting the cells for damage. The high number of false positive and false negative results led the American Academy of Allergy to conclude that there was no evidence that the test was effective in diagnosis of food allergy.

Other tests

Hair analysis, applied kinesiology, radionics, radiaesthesia, psionic medicine, and auriculocardiac reflex testing have never been objectively evaluated and are more a matter of gullibility and faith than science. Electrodermal (Vega) testing does not correlate with skin prick testing and cannot distinguish atopic from non-atopic

individuals.

Treatment

Dietary management

Recognition of the offending food and its elimination from the diet is the cornerstone of treatment. In patients with acute IgE-mediated reactions to a single food, such as shellfish, this is usually straightforward. Patients with anaphylactic reactions to foods need to be careful to avoid accidental exposure. A problem for such patients is the use of a food, most notably nuts, as an undeclared or 'hidden' ingredient in manufactured foods or restaurant meals. Where there remains a risk of accidental ingestion it may be appropriate for some patients to carry a preloaded syringe of adrenaline (epinephrine) for self-injection.

In less clearcut situations, certain foods or food additives are eliminated empirically because they are frequently implicated in that form of food intolerance: for example, a diet free of cereal grains and dairy products may be beneficial in certain patients with irritable bowel syndrome, while a diet free of azodyes, preservatives, and salicylates helps a proportion of patients with chronic intractable urticaria.

Patients who seem intolerant of a wide range of foods may need a very restricted diet, sometimes called a 'few-food' diet. If symptoms are improved, then foods can be reintroduced one at a time. This is both diagnostic and therapeutic, but care is essential as anaphylaxis can occur on reintroduction, especially in children. Expert advice from specially trained dietitians is essential to avoid nutritional deficiency.

Sodium cromoglycate

Oral sodium cromoglycate has been used as an adjunct to diet in selected patients with food allergy, especially those with accompanying allergic reactions in the eyes, nose, and skin. Its effectiveness is still unproven.

Immunotherapy

Although immunotherapy (hyposensitzation) is effective in wasp or bee venom anaphylaxis and in some forms of allergy to inhaled allergens, it has never been evaluated scientifically in food intolerance. There is considerable interest in future immunization with peptides containing T-cell epitopes.

'Alternative' therapies

Provocation-neutralization therapy and enzyme-potentiated desensitization are two treatments used by 'alternative' practitioners: neither is of proven value, although both induce significant placebo responses.

Food allergy seems particularly vulnerable in inducing unorthodox treatments which have not been scientifically validated by double blind, placebo controlled trials or confirmed by independent investigators. The hazard of the unconventional approach to therapy is that potentially serious problems can be misdiagnosed and mistreated.

Further reading

Ament ME, Ochs HD, Davis SD (1973). Structure and function of the gastrointestinal tract in primary immunodeficiency syndromes. A study of 39 patients. *Medicine (Baltimore)* **52**, 227–48.

Barrett S (2000). http://www.quackwatch.com/.

Blanshard C (1999). Gastrointestinal manifestations of HIV infection. *Hospital Medicine* **60**, 24–8.

Burks AW, Stanley JS (1998). Food allergy. *Current Opinion in Pediatrics* **10**, 588–93.

Corley DA, Cello JP, Koch J (1999). Evaluation of upper gastrointestinal tract symptoms in patients with HIV. *American Journal of Gastroenterology* **94**, 2890–6.

David TJ (1993). *Food and food additive intolerance in childhood*. Blackwell Scientific, Oxford.

Ernst PB *et al.* (1999). Regulation of the mucosal immune response. *American Journal of Tropical Medicine and Hygiene* **60**, 2–9.

Frieri M, Kettelhutt BV, eds (1999). *Food hypersensitivity and adverse reactions*. Marcel Dekker, New York.

Haeney MR (1994). Diagnostic tests in allergic disease. In: Spickett GP, Lewin I, eds. *Current themes in allergy and immunology*, pp 1–7. Royal College of Physicians, London.

Hein WR (1999). Organization of mucosal lymphoid tissue. *Current Topics in Microbiology and Immunology* **236**, 1–15.

Jewett DL, Fein G, Greenberg MH (1990). A double-blind study of symptom provocation to determine food sensitivity. *New England Journal of Medicine* **323**, 429–33.

Lewith GT *et al.* (2001). Is electrodermal testing as efficient as skin prick tests for diagnosing allergies? A double blind, randomised block design study. *British Medical Journal* **322**, 131–4.

Royal College of Physicians (1992). *Allergy. Conventional and alternative concepts*. Royal College of Physicians, London.

So ALP, Mayer L (1997). Gastrointestinal manifestations of primary immunodeficiency disorders. *Seminars in Gastrointestinal Disease* **8**, 22–32.

Strobel S, Mowat AM (1998). Immune responses to dietary antigens: oral tolerance. *Immunology Today* **19**, 173–81.

Teahon K *et al.* (1994). Studies on the enteropathy associated with primary hypogammaglobulinaemia. *Gut* **35**, 1244–9.

Viney JL, Fong S (1998). b7 integrins and their ligands in lymphocyte migration to the gut. *Chemical Immunology* **71**, 64–76.

Young E *et al.* (1994). A population study of food intolerance. *The Lancet* **343**, 1127–30.

Zullo A *et al.* (1999). Gastric pathology in patients with common variable immunodeficiency. *Gut* **45**, 77–81.

14.5 The mouth and salivary glands

T. Lehner

Dental caries and sequelae

Aetiology

Dental decay or caries is a very common chronic disease and causes much pain and discomfort. Caries is most frequent in children and young adults. It affects the pits and fissures of the occlusal surfaces, and the enamel of the approximal surfaces of teeth. Root caries (at the neck of the tooth) occurs later in life. Caries is an infection caused by the aggregation of bacteria on the surface of the tooth, referred to as dental plaque.

The development of dental caries requires the presence of cariogenic bacteria that produce acid below the critical pH (5.5) required for dissolving enamel and sugar in the diet that can be metabolized by bacteria. *Streptococcus mutans*, *Streptococcus sanguis*, *Lactobacillus acidophilus*, *Lactobacillus casei*, and *Actinomyces viscosus* are cariogenic. However, *S. mutans* appears to be the most efficient cariogenic organism. Germ-free studies have clearly shown that *S. mutans* induces caries rapidly in the absence of other organisms; it is a facultative anaerobic, non-haemolytic, acidogenic organism, producing extracellular and intracellular polysaccharides. The organism fulfils Koch's postulates as a cause of dental caries.

The most common carbohydrates in our diet are starch and sucrose, with smaller amounts of glucose, fructose, and lactose. However, the most important substrate in humans is sucrose. Sucrose gives rise to heavy plaque formation, with considerable amounts of extracellular polysaccharide. The most important polysaccharide is dextran (glucan), which is synthesized in large amounts by the constitutive enzyme glucosyltransferase (dextran-sucrase). Dextran allows plaque to stick to the surface of the enamel.

Streptococci do not possess a cytochrome system but contain glycolytic enzymes which will convert glucose to lactic and other organic acids. The pH inside the plaque may fall within 2 to 3 min of rinsing the mouth with glucose or sucrose from a level of about 6.5 to 5; the critical pH below which decalcification of enamel

occurs is thought to be about 5.5. Caries is the end result of a complex sequence of microbial and biochemical processes terminating in acid formation.

Pathology

Caries develops as a result of acid formed by the bacterial plaque acting on sucrose. The enamel becomes demineralized and plaque bacteria penetrate along the enamel prisms. This process progresses slowly through the enamel layer, but once the dentine is reached, destruction by decalcification and proteolysis of the dentine is rapid. The pulp reacts by an acute inflammatory response that results in necrosis, as the pulp is enclosed within the rigid walls of the tooth and the exudate cannot expand to adjacent tissues. Eventually, infection and toxic materials spread from the opening of the root canal to the tissues around the apex of the tooth and induce periapical inflammatory changes, which may terminate in an acute or chronic abscess or a chronic granuloma. If epithelial proliferation takes place within the granuloma or abscess, then a cyst may develop, which will increase in size over many years before it may be revealed clinically. A dental abscess represents a mixed infection with a variety of streptococci, staphylococci, and other organisms.

The immunological changes are complex, but serum IgG, IgA, and IgM antibodies, as well as cell-mediated immunity to *S. mutans*, can be correlated with the DMF (**D**ecayed, **M**issing, and **F**illed teeth) index of caries. Salivary IgA antibodies are also found. Although humans have the potential to mount humoral and cellular immune responses to *S. mutans* under natural conditions, the immunity achieved is commonly ineffective. Immunization experiments with *S. mutans* have been successfully carried out in rats and monkeys, with a significant reduction in caries. There are two principal immunological mechanisms of protection against caries. One involves salivary IgA antibodies, which can be induced by direct immunization of the minor salivary glands or by immunization of the gut-associated lymphoid tissue, from where sensitized B cells may home to the salivary glands. Salivary antibodies may prevent *S. mutans* from adhering to the tooth surface and thereby prevent caries. The alternative mechanism involves all the humoral and cellular components elicited by systemic immunization. Antibodies, complement, polymorphonuclear leucocytes, lymphocytes, and macrophages pass from the gingival blood vessels to the gingival domain of the tooth. Bacterial colonization of the tooth can therefore be influenced by systemic immunity and an important mechanism is probably that of IgG-induced opsonization, binding, phagocytosis, and killing of *S. mutans* by phagocytes.

Clinical features

The patient complains of toothache aggravated by hot or cold drinks or food. The throbbing pain becomes progressively worse, affects the patient especially at night, and may radiate to the face and ear. If relief is not sought the pain becomes excruciating in intensity, and the tooth becomes tender to bite on. This will be followed by death of the dental pulp and the development of an acute swelling due to an abscess or cellulitis. With an acute abscess the inflammatory exudate may penetrate through the bone to the soft tissues. Whilst the pain is reduced the oedematous swelling of the face increases, and if the upper canine is involved the swelling spreads to the eyelid and may present an alarming appearance. The regional lymph nodes are tender and enlarged and there may be fever and some malaise.

Much less commonly a cellulitis or infection by b-haemolytic streptococci may give rise to a spreading infection along the fascial planes, especially of the submaxillary and sublingual spaces. The inflammatory exudate may occasionally spread along the parapharyngeal spaces into the loose connective tissue of the glottis causing oedema of the glottis and respiratory obstruction. The attendant brawny swelling of the neck and floor and the mouth, difficulty in swallowing, trismus, fever, and malaise is referred to as Ludwig's angina. An alternative chronic course is the development of a chronic pulpitis, granuloma, abscess, and eventually cyst around the apex of the offending tooth, and these may proceed without symptoms or only slight discomfort.

Although the patient may point out the painful tooth this can be misleading because the pain often radiates to adjacent teeth. The offending tooth is located by finding the caries, most commonly in the pits and fissures of the occlusal surfaces or the approximal surfaces of adjacent teeth. The tooth responds with pain on application of a hot or cold stimulus, and later is tender to percussion and may be discoloured. Dental radiographs will confirm or localize the carious tooth and, at a later stage, periapical pathological changes.

Treatment

The principles of treatment are to remove the caries, apply a non-irritant material such as zinc oxide and eugenol dressing to protect the pulp, and then restore the tooth with a filling. The most common filling material is dental amalgam which contains mercury. Public anxiety has been aroused by anecdotal evidence that amalgam fillings are toxic and can cause poor memory and lassitude or even multiple sclerosis. There is, however, no scientific evidence to justify these claims, though as a precautionary measure, the United Kingdom Department of Health recommends that mercury-containing amalgam fillings should not be used in pregnant women. The alternatives to amalgam are a number of composite filling materials. If the pulp is damaged irreversibly it will have to be extirpated and root-canal therapy instituted. The alternative to conservative treatment is extraction of the offending tooth. A dental abscess is effectively dealt with by extraction of the diseased tooth, for this removes the source of infection and drains the pus.

If the tooth is to be saved, the pus is drained by an intraoral incision and/or establishing drainage through the root canal. Antibiotics are usually given for acute abscesses and oral penicillin such as phenoxymethylpenicillin, 250 mg four times a day for about 7 days, is adequate. Cellulitis should first be treated by intramuscular penicillin in the form of benzylpenicillin, 1 megaunit (MU) four times a day. The swelling should then be incised to relieve the pressure and provide drainage; extraction of the tooth under general anaesthesia should take place as soon as the patient's condition permits.

Prevention of dental caries is best practised by careful plaque removal by the individual, and by limiting the intake of sugar, especially the frequent consumption of sweets and sweetened drinks. The type of toothpaste used matters less than the method of tooth brushing, though fluoride in toothpaste decreases the incidence of caries in children by up to 40 per cent. Water fluoridation, however, is the most effective public health preventive measure. One part per million of fluoride in the drinking water will decrease the incidence of caries in children by up to 60 per cent. There is no evidence of toxicity from water fluoridation. The ethical and scientific issues of water fluoridation are complex and have been the subject of a report by the Royal College of Physicians of London.

Differential diagnosis

Toothache occasionally needs to be carefully differentiated from sinusitis and neuralgia. Throbbing pain that is exacerbated by thermal stimuli and is more severe at night is an important diagnostic feature. An abscess or cellulitis caused by dental caries has been confused with mumps, although mumps is confined predominantly to the parotid fascia, earache may be a prominent feature, and pain is elicited by pulling on the ear lobe. A chronic granuloma or a dental cyst are usually diagnosed radiologically, unless the cyst becomes large and a swelling becomes clinically evident.

Course and prognosis

The acute sequence of events from dental caries is acute pulpitis, periodontitis, resulting in an abscess or cellulitis. If treated promptly the sequelae can be prevented, but if not treated the patient will lose the tooth and may develop facial scarring due to a discharging sinus. With slow progression of caries or incomplete removal of decay chronic pulpitis may supervene followed by chronic periadenitis, which may result in a periapical granuloma, abscess, or cyst. Dental caries is in most instances a progressive condition and can be halted only by a dental surgeon.

Gingival and periodontal disease

Aetiology

A mild inflammation of the gingiva (gum) and slight destruction of the collagen fibres of the periodontal membrane are found in most adults. Advanced destruction of the periodontal membrane, including the supporting bone, is found in about half of the middle-aged or older population. A close association has been found between accumulation of bacterial plaque and gingivitis. During this process a change occurs from a predominantly Gram-positive coccal form of plaque to a complex population of filamentous organisms, spirochaetes, vibrios, and Gram-negative cocci. Of the Gram-positive organisms, *Actinomyces viscosus* appears to be involved in the development of gingivitis. Gram-negative organisms are thought to be essential in the development of periodontal disease. *Porphyromonas gingivalis*, *Actinobacillus actinomycetemcomitans*, *Fusobacteria* and *Capnocytophaga* spp. have been implicated in this disease. The cell walls of the Gram-negative organisms contain lipopolysaccharides and those of the Gram-positive organisms have lipoteichoic acids, dextrans or levans, which may be responsible for a variety of immunological functions.

The causative factors responsible for periodontal disease are not known, but bacterial plaque is thought to be involved. There are two views concerning the microbial

aetiology: that the non-specific mixed organisms in dental plaque are responsible for the development of periodontal disease or that specific organisms are responsible. The hypothesis of specific microbial aetiology has received support from observations that *P. gingivalis* is the predominant organism isolated from periodontal disease. Furthermore, a specific but rare type of juvenile and rapidly progressing adult periodontitis is associated with *A. actinomycetemcomitans*. Invasiveness of these micro-organisms probably plays an important part in their virulence, and some of the periodontopathic bacteria can be found in the gingiva of adult as well as juvenile periodontitis. However, innate and acquired host factors may be more important than bacteria in determining the development of periodontal disease.

Dental plaque may calcify, especially in adults and the elderly, to produce calculus. This is often found on the lingual surface of the lower incisors and the buccal surface of the upper molars, i.e. opposite the orifices of the major salivary glands. Chronic gingival inflammation may persist for many years and breakdown of the periodontal membrane, with loss of the supporting bone, may follow and increase in severity over the years. This is referred to as periodontitis, or 'pyorrhoea' as it used to be called, and is the most important cause of loss of teeth after the age of 40, when the incidence of dental caries has greatly diminished. An important feature of periodontitis is that it affects many teeth, and may result in complete loss of the dentition. As mentioned above, a very rare type of rapid destruction of the supporting dental tissues is found in children or young adults and is referred to as juvenile periodontitis; one or more teeth may become mobile and may be lost before 21 years of age.

Pathology

There are four immunopathological stages:

1. The initial lesion is found in the normal clinical state, with a localized inflammatory response of polymorphonuclear leucocytes; complement activation and chemotaxis generated by plaque antigens and possibly immune complexes may account for this stage.
2. The early lesion shows a localized infiltration of predominantly T with a few B lymphocytes. In the circulation, lymphocytes are sensitized at this stage to plaque antigens.
3. The established lesion is characterized by a localized plasma cell infiltration and peripheral blood lymphocytes can be stimulated to proliferate by plaque antigens. This stage can persist for years, with early pocket formation.
4. The advanced lesion marks the transition to a destructive immunopathological mechanism, with ulceration of the pocket epithelium and localized destruction of collagen and bone.

Periodontitis is a progressively destructive process leading to loss of teeth. The immunological processes are complex, and cytokines may play a significant role, especially raised levels of tumour necrosis factor-a, interleukin 1b, interferon-g, prostaglandins, and metalloproteinases. Destruction of the periodontal ligament and bone eventually leads to loss of support of the teeth.

Clinical features (Fig. 1 and Fig. 2)

Fig. 1 Chronic gingivitis, with erythema and oedema of the gingival margin of the lower teeth and especially the upper right lateral incisor.

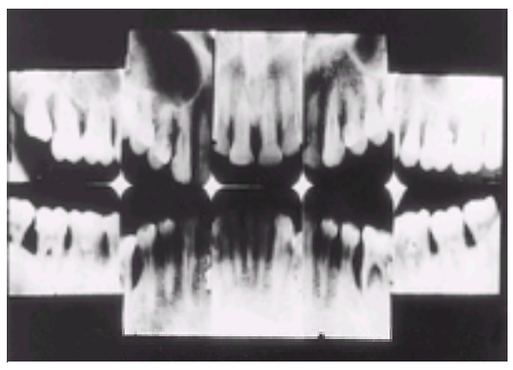

Fig. 2 Radiograph of teeth showing advanced periodontitis with loss of supporting bone of the teeth.

The symptoms of chronic gingivitis or periodontitis are usually so mild that they go unnoticed by the patient. They may, however, complain of discomfort from their teeth, bleeding of gums and associated halitosis, difficulty on eating, looseness of teeth, and occasionally abscess formation. A lack of severe symptoms allows the disease to progress to an irreversible stage before help is sought. Periodontal disease is commonly associated with diabetes. Smoking is a well-documented risk factor. Recently a significant link has been found between periodontal disease and cardiovascular thrombosis. Surprisingly, periodontal disease in pregnant women has been associated with an increased risk of low birth weight and premature birth. It is not clear whether transient bacteraemia in the last trimester might elicit a mechanism responsible for these findings.

Differential diagnosis

Chronic gingivitis can be differentiated from acute ulcerative gingivitis by the sudden onset, malaise, characteristic halitosis, pain, and ulceration of the gingiva in the latter. Herpetic gingivostomatitis occurs predominantly in children and again the onset is acute, with fever, malaise, pain, and ulceration of the gingiva and oral mucosa (see below). Desquamative gingivitis associated with bullous lesions and lichen planus may cause difficulties in differential diagnosis and the points to bear in mind are that the attached gingiva shows diffuse erosive areas and evidence of bullous lesions may be found in the oral mucosa. Periodontitis is clinically differentiated from gingivitis by the loss of connective tissue attachment to the teeth, leading to pocket formation. There is progressive bone loss, readily diagnosed by radiographic examination of the teeth, and eventually loosening of teeth.

Treatment

The aims in the management of gingivitis and mild periodontitis are to remove dental plaque and calculus by scaling the teeth, and this can be done only by a dentist or dental hygienist. Prevention is, however, much more effective by plaque control, which involves careful tooth brushing with the aid of plaque-disclosing solutions and regular use of dental floss and wood points. Chlorhexidine rinses (0.2 per cent) twice a day prevent the accumulation of plaque and decrease gingival inflammation. In some forms of periodontitis tetracycline or metronidazole can be helpful; preparations are available that can be applied locally, but these are generally used in addition to local surgical treatment. However, once deep periodontal pockets have been formed, these are treated by root planing, gingival

curettage, or surgically. It should be appreciated that the management of periodontal disease is dependent upon meticulous plaque control.

Course and prognosis

If the bacterial plaque is not removed, gingivitis may progress to periodontitis and after many years will result in increased mobility and loss of teeth. This process, however, is reversible by plaque control and, if necessary, eradication of pockets, as long as there is sufficient bone to support the teeth.

Herpes simplex and other viral infections

Herpes simplex virus type 1 is responsible for certain orofacial infections (see also Section 7).

Primary herpetic gingivostomatitis

Aetiology

Clinical or subclinical primary infections by herpes simplex virus type 1 are acquired in early childhood, probably in the second and third years of life. Primary herpetic infection in the first year is rare, because most mothers have neutralizing IgG antibodies to the virus that are transferred through the placenta to the fetus. Serum virus complement-fixing and neutralizing antibodies are found in about 50 per cent of children at 5 years of age. The disease is common in children, but is also seen, less frequently, in adults.

Pathology

Herpes simplex virus is a DNA virus and there are two types: type 1 is found predominantly in the orofacial region and type 2 in the genital region. There are three genes (a, b, g) and the b gene codes for viral glycoproteins gB, gC, gD, and gE. These viral glycoproteins have been well characterized: gB is involved in viral penetration of the cell membrane, gC constitutes the C3b receptor (binding activated C3b), and gE is the crystallizable fragment receptor for IgG. Antibodies against gD neutralize herpes simplex virus and block its penetration. Hence this viral infection generates many significant immunological molecules in the host cell, in addition to expressing a viral antigen on the cell surface.

Infection starts with the entry of herpesvirus into epithelial cells. Viral replication takes place inside the nucleus, and this is associated with the formation of intranuclear inclusion bodies and giant cells. As more epithelial cells become infected, degenerative and oedematous changes give rise to vesicle formation. The intraepithelial vesicles contain oedema fluid, with giant cells and degenerating cells with intranuclear inclusion bodies. The vesicles rupture early, resulting in ulcers that heal rapidly.

Clinical features

The disease is recognized by an acute onset of a sore mouth and often sore throat, fever, and extensive inflammation of the gums, followed by formation of vesicles and ulcers of the oral mucosa, and regional lymphadenitis. Infants display considerable fretfulness, sleeplessness, and refusal to eat. Initially there are crops of small ulcers but these coalesce to produce large, shallow, irregular ulcers with surrounding inflammation. Herpetic keratitis is not often associated with herpetic stomatitis, and herpetic encephalitis is extremely rare but may occasionally complicate herpetic stomatitis.

Diagnosis

The early phase of infection can be confused with a cold but the development of vesicles and ulcers makes that diagnosis unlikely. Recurrent aphthous ulcers may occasionally be misdiagnosed in the adult, though the important differentiating points are the acute onset, sore throat, fever, and lymphadenitis in herpetic infection. Laboratory tests can be useful in confirming the diagnosis. Direct examination of a smear from the lesion can be helpful if intranuclear inclusion bodies or giant cells are found. Culture of the virus may assist in the diagnosis, but the herpesvirus is also found in carriers. A rise in antibody titre to the virus during an infection can be a useful aid in diagnosis.

Treatment

Patients are advised to rest for 2 to 4 days; a soft diet is indicated, and an adequate fluid intake is emphasized. The mouth is cleansed by thorough rinsing with hot salt water six times daily and the teeth are cleaned with a wet flannel. In infants, special attention must be paid to the fluid intake and sleep. A useful sedative to use is promethazine elixir, given in doses of one teaspoonful (5 mg/5 ml) at night-time.

Acyclovir tablets (200 mg), two to four times daily, can be helpful if started at an early stage of infection. However, in late onset of primary herpetic infection, tetracycline mouthwash can speed up recovery; however, this should not be used in children.

Course and prognosis

The natural course of this infection is 7 to 14 days, during the initial days of which eating is usually difficult, but healing of the ulcers occurs spontaneously. Recurrence of herpetic lesions intraorally is rather rare in otherwise healthy subjects but occurs frequently in patients with cellular immunodeficiencies.

Recurrent herpetic infection

This is also called recurrent herpes labialis or cold sores.

Aetiology

The lesion is caused by herpes simplex virus type 1 and is commonly found from childhood to past middle age in both sexes. A variety of factors may precipitate the lesions: the common cold, fever, exposure to sunlight, local trauma, emotional stress, menstruation, dental treatment, and section of the sensory root of the trigeminal ganglion are among the best known. Severe herpetic infections, affecting the lips, perioral skin, and mouth, are seen in patients receiving immunosuppressive drugs.

Pathology

Primary herpes simplex infection is followed by viral latency in the trigeminal ganglion. The relation between primary infection, latency, and recurrent infection by herpes simplex virus has not been completely elucidated. However, there is evidence that primary infection induces immune responses to the virus; antibody and cell-mediated cytotoxic mechanisms kill most of the virus-infected cells. The virus is sequestrated to the nerves and will migrate along the axons to the trigeminal ganglion. Indeed, the entire herpes simplex virus genome can be found in the trigeminal ganglion, although the DNA is qualitatively different. A number of clinical precipitating factors may induce derepression of the viral genome and virus replication, which will then migrate along the axon to be shed at the nerve endings. In the presence of some defect in cell-mediated immunity acting at the neuroepithelial junction, recurrent herpetic lesions will be precipitated. Cytokine production, especially interferon-g, may be impaired, and a decrease in cytotoxic CD8 T cells is involved in recurrent herpetic infection.

Clinical features (Fig. 3)

Fig. 3 Recurrent herpes labialis vesicle on the vermilion border of the lower lip.

The lesions are usually limited to the vermilion border of the lips and adjacent skin. A single blister or a crop of blisters may develop a day after the prodromal phase. The duration of the lesion usually varies between 3 and 10 days, but secondary infection by *Staphylococcus pyogenes* commonly occurs. The lesion recurs at various intervals often at the same site for many years, and the rate of recurrence may be related to the type of precipitating factor involved. The significance of cellular immunity is highlighted by herpes simplex virus infections found in cell-mediated immunodeficiency states, such as AIDS, and in patients receiving immunosuppressive therapy.

Diagnosis

Localization to the vermilion border of the lips and the history of recurrences make this a readily recognizable condition. Laboratory assistance is rarely required but the findings are similar to those described for primary herpetic infection, except that there is an elevated initial antibody titre which does not usually increase during recurrent infection. Staphylococcal infection from the anterior nares should be excluded.

Treatment

Acyclovir (acycloguanosine) cream (5 per cent) can be effective if applied during the prodromal phase. Staphylococcal infection responds readily to mupirocin or fucidin ointment, applied three times daily. In the severe type of mucocutaneous herpetic infection in immunosuppressed patients, acyclovir tablets (200 mg) are administered two to four times daily.

Course and prognosis

The lesions heal usually within about 7 days but recurrences are difficult to prevent. If the precipitating factors are known, some preventive measures can be taken such as applying a barrier cream to the lips before exposure to the sun.

Herpes zoster infection

Herpes zoster infection of the skin of the face, innervated by the second or third branch of the trigeminal nerve, may be associated with unilateral oral vesicles. These break down early to produce ulcers along the oral distribution of the maxillary or mandibular branches.

Herpangina

This is a rare infection by the coxsackie group A viruses, usually affecting the soft palate and the oropharyngeal region. Children tend to be affected more often than adults and the mode of presentation of the disease is similar to that in primary herpetic stomatitis. The diagnosis can be firmly established only by isolating the virus from a lesion or by showing an increase in antibody titre. The disease appears to be self-limiting and specific treatment is not necessary.

Hand, foot, and mouth disease

This is another virus infection caused by coxsackie A5, 10, and 16 (see Section 7). The mouth is sore due to multiple small vesicles or ulcers, which most commonly affect the hard palate, tongue, and buccal mucosa. There are associated vesicular lesions on the hands and feet. The diagnosis is confirmed by isolating the virus from the lesion. The disease is self-limiting within about 2 weeks and no specific treatment is necessary.

Measles

This is an acute exanthematous virus infection of children (see Chapter 7.10.6). Whitish macules on the buccal mucosa, known as Koplik's spots, may precede the development of the red macular rash by 2 to 3 days.

AIDS (see Chapter 7.10.21)

Aetiology

This is an infection by the human immunodeficiency virus (**HIV-1**) affecting primarily CD4+ T cells, macrophages, Langerhans, and dendritic cells. In addition to the CD4 receptor all HIV-1 strains require the coreceptor CCR5 or CXCR4 for viral entry.

Pathology

Entry of HIV into the host is by the interaction between gp120, CD4, and CCR5 or CXCR4 on the cell membrane. The trimolecular complex enables the viral particle to enter the cell by fusion between the viral and cell membranes. Hence, the primary target of HIV is the CD4 subset of T cells which decrease in number as the cells become infected and killed, but the CD8 subset is not affected, resulting in a decrease in the CD4:CD8 cell ratio. It is not clear whether the virus kills CD4 cells directly or indirectly by an immune mechanism.

Clinical features

There are five main populations at risk:

1. Homosexual men: those with multiple sex partners and the anal-receptive partner in anogenital intercourse are at greatest risk.
2. Transmission of HIV during vaginal intercourse is common in parts of Africa and Asia; female prostitutes may carry the virus in their genital secretions.
3. Intravenous drug abusers can spread the virus via infected needles from one person to another, directly by the vascular route.
4. Those who have received a blood transfusion with HIV-infected blood, especially haemophiliacs treated with factor VIII.
5. Perinatal HIV infection of babies from infected mothers.

There is some epidemiological evidence that oral sex might lead to HIV infection, but it is unlikely that salivary transmission of HIV can occur. Comparative isolation studies of HIV from body fluids have been made and their results suggest that whilst HIV can be cultured from whole saliva, the frequency of isolation is low (up to 9 per cent) as compared with semen (21 per cent) or plasma (55 per cent). The quantity of HIV isolated from saliva is also low. The available evidence suggests that the HIV resides in the cellular fraction of oral fluid, presumably CD4+ T cells and macrophages, and not the fluid fraction originating mostly from the salivary glands.

The special significance to dentists of oral transmission of HIV is self-evident, as they work in a pool of saliva, often mixed with gingival blood. However, seropositive conversions were not found among about 1000 dental staff in the United States and the same number tested in Germany. Only 1 out of 1309 dentists in another study in the United States was seropositive and he did not wear protective gloves. This is an almost negligible prevalence of HIV seropositivity in a population of dentists, among whom more than 90 per cent admitted to needlestick injuries. The transmission of HIV during dental procedures remains a possibility, especially since a case has been documented of a Californian dentist passing HIV to his patients. However, the details of this case are most perplexing and the route of transmission has not been established.

This section will be confined to the oral manifestations of AIDS. It is of significance that oral candidiasis and hairy leucoplakia may predict the development of AIDS. A variety of opportunistic infections may develop in the mouth. Fungal infection with candida, especially *Candida albicans*, is common. All varieties of oral candidiasis have been recorded in AIDS, but it appears that the chronic hyperplastic and atrophic varieties are more frequent than the pseudomembranous variety. Other fungal lesions may occur but are rare (for example histoplasmosis and cryptococcosis).

Viral infections with herpes simplex virus give rise to recurrent oral herpetic lesions affecting the palate or gum and present as painful vesicles that ulcerate. It should be remembered that recurrent intraoral herpetic lesions are extremely uncommon in the rest of the population (unlike recurrent herpes labialis). Orofacial lesions due to herpes zoster have also been recorded but are rather rare. Epstein–Barr virus appears to cause hairy leucoplakia, which is a raised white plaque commonly affecting the tongue and is clinically similar to chronic hyperplastic candidiasis. Similar lesions have not been recorded in the general population. Papillomavirus may induce single or multiple warts in the mouth of AIDS patients.

Kaposi's sarcoma is a neoplasm of the vascular endothelial cells. Oral lesions present as red or purple macules or papules, often affecting the palate and tongue. Other neoplasias are less common but non-Hodgkin's lymphomas and carcinomas have been recorded.

Gingivitis and periodontitis may show changes similar to those of acute necrotizing ulcerative gingivitis (see below), except that these may be superimposed on rapidly progressing periodontitis. The condition can be painful and has been associated with rapid loss of soft tissue and bone support, leading to loss of teeth.

Recurrent oral ulcers are probably more common in patients with AIDS than in the general population. Enlargement of the salivary glands, especially of the parotid glands, might be caused by a viral infection.

Differential diagnosis

As oral manifestations of AIDS may occur early in the disease, oral candidiasis, herpetic infections, leucoplakia, oral or gingival ulcers, salivary gland swellings, and oral tumours should be suspected, especially in young men (and women) falling into the population groups most at risk of HIV infection.

Treatment

In addition to the general management of AIDS, the teeth and gums should receive a great deal of attention to maintain a high standard of oral hygiene. Otherwise the lesions should be treated as for any other oral condition. Routine dental treatment can be difficult to arrange in a dental practice, but most hospitals have made special arrangements for AIDS patients.

Fungal infections (see Section 7.12)

Candidiasis

This is also called moniliasis or thrush.

Aetiology

Candida is a commensal organism in the mouth found in 20 to 40 per cent of the normal population. Most normal subjects show serum-agglutinating antibodies and a cutaneous delayed hypersensitivity reaction to candida. It is not clear whether candida infection of the oral mucosa is endogenous or exogenous, but as the organism is ubiquitous, a suitable environment and impaired immune responses are the most important conditions conducive to infection by candida. Oral candidiasis can be an early manifestation of AIDS (see above). Although most species of candida can become pathogenic, *C. albicans* is most frequently found in oral infections.

Pathology

The different varieties of candidiasis have in common a superficial invasion of epithelium by fungal hyphae; it is unusual for the hyphae to penetrate the basement membrane. However, in immunodeficient patients candida may spread by the vascular route to the heart, kidneys, and brain. Raised titres of antibodies to candida are found in serum and secretory IgA antibodies in the saliva of patients with oral candidiasis. Antibodies and complement are necessary for optimal phagocytosis of candida by polymorphonuclear leucocytes or macrophages. There is evidence that serum antibodies to the 44 to 60 kDa candida antigen prevent systemic candidiasis. In contrast, cell-mediated immunity is involved in chronic mucocutaneous candidiasis, with a spectrum of cellular immunodeficiencies.

Clinical features

Oral candidiasis develops in a variety of conditions predisposing to candidal proliferation: diabetes mellitus, anaemias, cell-mediated immunodeficiencies (such as AIDS or thymic defects), broad-spectrum antibiotics, immunosuppressive drugs, and leukaemias. Local factors commonly predisposing to oral candidiasis are dry mouth due to Sjögren's or sicca syndrome, irradiation, dentures, or steroid sprays used for asthma.

Varieties of oral candidiasis

There are four main varieties of oral candidiasis.

Acute pseudomembranous candidiasis (thrush)

This disease is commonly seen in infants as well as in debilitated adults, particularly in diabetes mellitus and malignant diseases (especially leukaemia and lymphoma). Iatrogenic agents are also important predisposing factors; systemic antibiotics, corticosteroids, and immunosuppressive drugs seem to enhance candida infection. Local antibiotic and corticosteroid treatment can enhance oral candidiasis. Clinical manifestations of thrush are usually symptomless white papules or cotton-wool-like exudates that can be rubbed off leaving an erythematous mucosa.

Acute atrophic candidiasis (Fig. 4)

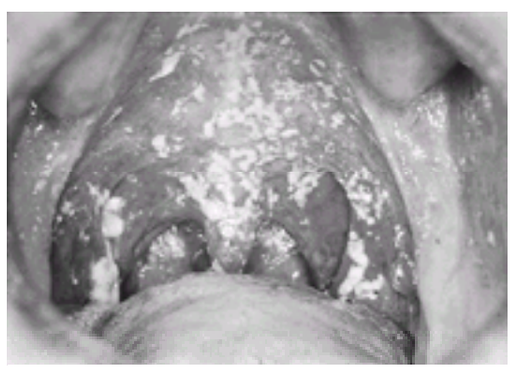

Fig. 4 Oropharyngeal thrush, following the application of a steroid spray in a patient with asthma.

This may follow acute pseudomembranous candidiasis and is usually associated with broad-spectrum antibiotic therapy; it is hence referred to as 'antibiotic sore tongue'. It is the only type of oral candidiasis that is consistently painful, showing a smooth erythematous tongue, with angular cheilitis and (less often) inflamed lips and cheeks.

Chronic atrophic candidiasis

This type of candida infection is better known as 'denture stomatitis', for it presents as a diffuse erythema of the palate limited to the denture-bearing mucosa. The denture covering the palatal mucosa predisposes to proliferation of candida. The lesion is usually symptomless but is often associated with angular cheilitis (Fig. 5).

Fig. 5 Angular cheilitis caused by candidal infection.

Chronic hyperplastic candidiasis

This lesion presents as a firm, diffuse white patch, or as numerous white papules with intervening erythema on the tongue, cheeks, or lips. The lesion may persist for many years or for life and should be distinguished from leucoplakia. This variety of candidiasis can be associated with skin lesions and there are three clinical types of mucocutaneous candidiasis:

1. Chronic localized mucocutaneous candidiasis. This starts in childhood as an intractable oral candida infection, with involvement of nails and sometimes the adjacent skin of the hands and feet. A number of other skin sites may show persistent candida infection.
2. Chronic localized mucocutaneous candidiasis with granuloma. This condition begins in infancy and the clinical manifestations are similar to those in the previous type of candidiasis, with the important additional feature of granulomatous masses affecting the face and scalp. Recurrent respiratory tract infection has been recorded in a quarter of affected children.
3. Chronic localized mucocutaneous candidiasis with endocrine disorder. This is found in children and young adults. A strong familial incidence is often found and candidiasis commonly precedes the endocrine abnormalities. The clinical features of candida infection are similar to those seen in the localized mucocutaneous variety. The association with hypoparathyroidism and Addison's disease, and less often pernicious anaemia and hypothyroidism, illustrates the relationship between cell-mediated immunodeficiencies and autoimmune endocrine disorders.

Differential diagnosis

Chronic hyperplastic candidiasis can cause some difficulties in differential diagnosis from leucoplakia and the laboratory tests are useful in this, as well as in the other types of candidiasis, in establishing the diagnosis. AIDS must be considered, particularly in homosexual males. A culture from the lesion yields candida, usually *C. albicans*, and direct examination of scrapings shows the Gram-positive hyphae and yeast cells of candida. Biopsy of the lesion in chronic mucocutaneous candidiasis is helpful, as in addition to the superficial invasion of epithelium by candida hyphae, there is usually extensive epithelial hyperplasia. The dermis shows an intense mononuclear cell infiltration with a large proportion of plasma cells.

A rise in convalescent serum antibody titre to candida may assist in the diagnosis of the acute types of candidiasis, but there may be an impaired antibody titre in the chronic type of candidiasis. Chronic mucocutaneous candidiasis usually shows some defects in cell-mediated immunity and this should be determined by investigating delayed hypersensitivity, lymphocyte proliferation, and generation of interleukins on stimulation with candida. It is essential that the endocrine function should be tested in children with chronic candidiasis of the mouth and nails.

Treatment

Oral candidiasis responds readily to topical oral treatment with antifungal drugs: sucking tablets of nystatin 500 000 IU four times a day or amphotericin B 100 mg four times a day for 1 to 2 weeks is very effective. Alternative antifungal agents, such as miconazole and fluconazole, are equally effective. Chronic mucocutaneous candidiasis, however, is often unresponsive to topical oral treatment and may necessitate intravenous administration of amphotericin B. Endocrine replacement therapy is essential if there is an associated endocrine disorder. Although almost complete eradication of the lesions can be accomplished, the disease tends to return after the drug is discontinued because of the underlying immunological defect which needs to be rectified.

Bacterial infections

Acute (necrotizing) ulcerative gingivitis

This is also called Vincent's gingivitis or acute fusospirochaetal gingivitis.

Aetiology

An infective cause of acute ulcerative gingivitis has been widely accepted, although the organisms thought to be responsible are disputed. *Fusobacterium fusiformis* and *Borrelia vincenti* have been favoured on account of their presence in large numbers in direct examination of smears from the lesions. *Bacteroides melaninogenicus* has also been implicated as the causative organism, but evidence is accumulating in favour of a mixed bacterial pathogenesis of Gram-negative organisms (fusobacteria, veillonella, bacteroides, leptotrichia), which may be responsible for the lesions due to their endotoxin activity.

Whatever role micro-organisms may play, a number of predisposing factors are recognized. Of the local factors poor oral hygiene with accumulation of dental bacterial plaque, defective restorations, and pericoronitis are most important. The prevalence of acute ulcerative gingivitis is rather high and it is seen more commonly in young adults and smokers. A lowered general resistance may also predispose to the disease.

Pathology

The gum undergoes an acute inflammatory reaction, with an intense polymorphonuclear response and fibrinous exudate. This leads soon to necrosis of the epithelium

and thrombosis of the small blood vessels.

Clinical features

Acute ulcerative gingivitis is readily recognized by the sudden onset of painful, bleeding gums and a characteristic foul breath. Except for primary herpetic stomatitis, this is the only other oral mucosal infection in which there is a rise in temperature, which may reach 39 °C, regional lymphadenitis, anorexia, and significant malaise. Oral examination reveals necrotic, punched-out ulcers, predominantly affecting the interdental gingiva. At times there are shallow necrotic ulcers affecting the oropharyngeal mucosa, which shows diffuse erythema; this has been referred to as Vincent's angina. In the presence of erupting wisdom teeth, the overlying gum can show ulceration and oedema causing partial trismus.

Diagnosis

This disease is often confused with primary herpetic stomatitis because of the acute onset. However, patients with primary herpetic stomatitis are usually younger and their breath is stale but lacks the distinct foul quality of that found in ulcerative gingivitis. First vesicles and then numerous well-defined ulcers are scattered over the oral mucosa, unlike the tendency for localization of necrotic sites to the gingiva in ulcerative gingivitis. Direct examination of a smear from the lesion reveals a large number of spirochaetal and fusiform organisms, with a decrease in the mixed bacterial flora.

Treatment

Metronidazole is very effective and should be taken 200 mg by mouth three times daily for 3 to 4 days. Phenoxymethyl penicillin, 250 mg taken four times daily for a week is equally effective in clearing the symptoms. Oxidizing agents, hydrogen peroxide mouthwash, and a variety of peroxyborate preparations are also useful. During the acute phase, patients are advised to use a soft toothbrush or a soft cloth to clean their teeth, and they are encouraged to rinse their mouths forcibly with warm saline every 3 h.

Although treatment by drugs is effective in clearing the acute phase, recurrences can be prevented only by careful attention to oral hygiene. The teeth have to be scaled and polished, and the patient is instructed as to the best method of tooth brushing and control of dental plaque. Frequent examinations by a dental surgeon are advisable.

Course and prognosis

In the absence of treatment the acute phase may gradually disappear leaving behind a partially necrosed gingiva and chronic inflammation. Inadequate treatment commonly leads to recurrent ulcerative gingivitis over many years, with halitosis, gingival bleeding, and recession.

Cancrum oris (noma)

This is a rapidly spreading gangrene of the lips and cheeks, mostly confined to children in parts of tropical Africa. It is thought to be an extension of acute ulcerative gingivitis when associated with other diseases, especially measles. Cancrum oris is very rare in the United Kingdom, but can be seen during the terminal stages in patients with leukaemia, especially when treated with a variety of cytotoxic, anti-inflammatory, and immunosuppressive drugs.

Tuberculosis

Oral tuberculosis is rare and usually associated with pulmonary tuberculosis. The presenting feature is usually a painful ulcer which may be single or multiple, often large, with a depressed, granulomatous floor and some induration of the base. The tongue, lips, and cheeks may be affected. Diagnosis is based on microscopical and cultural demonstration of *Mycobacterium tuberculosis* and a biopsy of the lesion, which will show a tuberculous granuloma. With the rise in the prevalence of tuberculosis, especially due to AIDS, oral tuberculous lesions may also reappear. Oral tuberculosis responds readily to specific chemotherapy.

Syphilis

Treponema pallidum may affect the mouth in all stages of syphilis but is uncommon (see also Chapter 7.11.33).

Primary stage

A chancre appears within 2 to 4 weeks of infection. The lesion presents on the lip or tongue as a painless, small, firm nodule that breaks down and forms an ulcer with raised indurated edges. The regional lymph nodes show discrete, rubbery enlargement. The diagnosis depends on direct observation of *T. pallidum* by darkground illumination. This stage is highly infective, but serological tests are usually negative during the initial 3 to 4 weeks.

Secondary stage

This develops 1 to 4 months after infection and presents as a generalized maculopapular rash and lymphadenitis. Shallow, snail-track ulcers affect the tonsils, tongue, or lips, and the saliva is highly infective. The serological tests for syphilis are positive.

Tertiary stage

This is delayed by 3 to 15 years after infection. Gumma and leucoplakia are the typical oral manifestations at this stage. A gumma starts as a swelling of the palate, tongue, or tonsils; it undergoes necrosis and results in a painless, punched-out, deep ulcer, with a 'wash-leather' floor. The lesion may heal by scarring, or give rise to perforation. Leucoplakia usually affects the dorsum of the tongue as an irregular, diffuse white patch that cannot be rubbed off.

The treatment of oral syphilis is the same as that used in other sites, but the response in the tertiary stage is rather poor.

Oral ulceration

In view of the great variety of oral ulcers a classification will be given first (Table 1). Only recurrent oral ulcers will be dealt with fully and the other types of ulcer will be considered predominantly under differential diagnosis.

Recurrent oral ulcers

Three types of ulcer will be described: minor aphthous ulcers, also known as aphthae; major aphthous ulcers, also referred to in the literature as periadenitis mucosa necrotica recurrens; and herpetiform ulcers. Aphthous stomatitis is another term used to describe these ulcers.

Aetiology

These are the most common lesions affecting the oral mucosa and the prevalence varies between 10 and 34 per cent in the population. Although a number of causes has been suggested, the aetiology of recurrent aphthous ulcers has not been fully established. Trauma is unlikely to play an essential role, though it might precipitate ulceration, as is seen following dental treatment. There is no evidence that vitamin deficiency or food allergy is involved. Infection by the herpes simplex virus has been excluded as a cause of this type of ulceration. Whilst emotional stress may often influence the pattern of the disease, it is unlikely to be the direct cause. A family history of recurrent aphthous ulcers is often present and the highest incidence of ulcers is recorded in siblings in whom both parents have recurrent aphthous ulcers. A hormonal disturbance may play a part, as in some female patients there is a relationship between the ulcers and the menstrual period; the onset of ulceration may coincide with puberty, or the ulcers may develop only after the menopause and the ulcers often disappear during pregnancy. The part that autoimmunity may play in the pathogenesis of this disease has not been fully elucidated. However, oral epithelial cells share common antigens with the 65 kDa heat

shock protein that is found in Gram-positive organisms. A specific peptide of 15 amino acid residues (91–105), derived from the sequence of the 65 kDa heat shock protein has recently been found to stimulate lymphocytes from patients with recurrent oral ulcers. The role of this peptide in the pathogenesis of oral ulceration is under investigation.

Pathology

An early intense lymphomonocytic infiltration, especially with a perivascular distribution, is a constant histological finding suggesting a delayed hypersensitivity reaction. This is followed by a polymorphonuclear infiltration. Immunohistological investigations suggest an enhanced immune response, with a significant increase in the number of CD4 and CD8 subsets of T cells, Langerhans cells, and macrophages and the expression of HLA DR in the epithelial cells.

Clinical features (Fig. 6)

Minor aphthous ulcers

About 80 per cent of recurrent oral ulcers are of this type; they are very common, especially in the 10 to 40 year age group, and they are found more frequently in females than males.

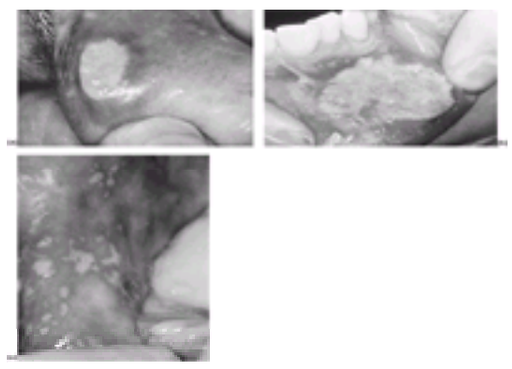

Fig. 6 The three types of recurrent oral ulcer: (a) minor aphthous ulcer; (b) major aphthous ulcer; (c) herpetiform ulcer.

A prodromal phase is recognized by most patients 1 to 2 days before the onset of ulceration, as a soreness or burning sensation. With the breakdown of epithelium and associated inflammatory reaction the pain increases in severity, particularly on eating. The ulcers are round or oval, up to five in number, and enlarge in size, although they remain well under 1 cm. They have a yellow floor with a slightly raised margin and often marked surrounding erythema and oedema. The most common sites of involvement are the mucosa of the lips and cheeks and margin of the tongue, and the ulcers last for 4 to 14 days. The rate of recurrence varies from 1 to 4 months and is usually irregular, though in some females ulcers may precede the menstrual period.

Major aphthous ulcers

These are severe variants of minor aphthous ulcer and fewer than 10 per cent of patients with recurrent oral ulcers have this type of ulcer. The pain that develops after the prodromal symptoms can be severe and persistent, so that patients find it difficult to eat and swallow food and often lose weight. Examination may reveal one to ten ulcers at a time and some of these may enlarge to about 3 cm. The ulcers are necrotic with a raised margin and inflammation of the adjacent tissue, so they occasionally mimic a carcinomatous ulcer. In addition to the lips, cheeks, and tongue, the soft palate and tonsillar region are commonly involved. There may be some regional lymph node enlargement. Healing of an ulcer may take 10 to 40 days and recurrences are so frequent that the patient suffers from continuous ulceration. Multiple small scars may result from large ulcers and these may assist in the diagnosis of major aphthous ulcers. The prevalence of major aphthous ulcers is raised in ulcerative colitis. A striking association has been found in smokers who give up the habit and develop recurrent aphthous ulcers.

Herpetiform ulcers

These are recurrent crops of up to a hundred minute ulcers, affecting any part of the mouth including the gum, palate, and dorsum of the tongue. They account for fewer than 10 per cent of recurrent oral ulcers and are much more common in females than males. Patients present with pain on eating and talking, and often with dysphagia; malaise and loss of weight can be prominent features. The lesions persist for 7 to 14 days and new ulcers commonly appear before the previous crop has healed, so that ulceration becomes continuous.

Diagnosis

The differential diagnosis of the three types of recurrent oral ulcer is given in Table 2. It is important to differentiate these ulcers from those found in patients with iron, folate, or vitamin B_{12} deficiency, who constitute fewer than 5 per cent of patients with recurrent oral ulcers. About 2 per cent may suffer from coeliac disease due to gluten enteropathy, and these ulcers respond readily to a gluten-free diet.

Agranulocytosis or neutropenia may manifest as shallow necrotic ulcers, predominantly affecting the oropharyngeal region. The ulcers tend to persist, unlike major aphthous ulcers which recur at different sites. However, cyclical neutropenia can mimic minor aphthous ulcers and the diagnosis depends on serial weekly white blood cell counts.

One of the most common diagnostic errors is to confuse the effects of denture trauma with aphthous ulcers, although the former are usually localized to the mucosa covering the mandibular and maxillary alveolus and the buccal and lingual sulci. The relation between denture trauma and ulceration is usually simple to find and requires the attention of a dentist.

The differential diagnosis from pemphigus, benign mucous membrane pemphigoid, and erythema multiforme is important and will be described below.

Not infrequently, patients with major aphthous ulcers are suspected of having a carcinoma, though a careful history will make it evident that these ulcers have recurred at different sites in the mouth. Although major aphthous ulcers may have a raised margin this is due to inflammation and not invasion, so that palpation fails to elicit the induration usually detected in carcinomatous ulcers. If in doubt biopsy of the ulcer is warranted.

Treatment

Topical corticosteroids are the best means to relieve aphthous ulcers. They are most effective if application is started during the prodromal phase when the mucosa has not yet ulcerated. If steroids are applied early ulceration may be prevented, but application at a later stage may still reduce the severity and duration of ulceration. The most useful preparations are triamcinolone in orabase, containing 0.1 mg triamcinolone per 100 g of an adhesive base, betamethasone, containing 0.5 mg steroid per tablet, or beclomethasone spray, up to six puffs daily. The tablets are kept in the mouth, or the ointment is applied to the ulcers, three to four times daily until the ulcer disappears. Systemic prednisolone has to be resorted to occasionally in patients with major aphthous ulcers when topical corticosteroids fail to control the ulcers.

Topical tetracycline is the drug of choice in suppressing herpetiform ulcers but is also useful in controlling some major aphthous ulcers, particularly when there is severe inflammation. Its mode of action is not clear and an effective preparation is to use capsules containing 250 mg tetracycline; the powder from a capsule is dissolved in 10 ml of water and kept in the mouth four times daily. Chlorhexidine solution (0.2 per cent) can be used as a mouthwash, which keeps the teeth free of

dental plaque, and may facilitate remission of ulceration.

Course and prognosis

Minor aphthous ulcers may recur from early childhood for many years, and these ulcers may often cause only transient discomfort to which the patient becomes accustomed. However, major aphthous and herpetiform ulcers usually cause a great deal of discomfort, difficulty in eating, and loss of weight. In children, major aphthous ulcers are particularly troublesome and need careful management. In the majority of patients with recurrent oral ulceration the disease burns itself out, but this may take many years. In a very small proportion of patients extraoral sites may become involved, of which the vulvovaginal region is most common, to form part of Behçet's syndrome. There is no way of predicting the development of Behçet's syndrome in patients with recurrent oral ulcers (see Chapter 18.10.5).

Bullous lesions

These are diseases that often affect the skin and mouth, but sometimes involve only the oral mucosa. Three conditions will be discussed in this section: pemphigus vulgaris, benign mucous membrane pemphigoid, and erythema multiforme (see Chapter 23.1).

Pemphigus vulgaris

Aetiology

This is a rare disease which in many instances presents in the mouth, although oral lesions are found at some stage of the disease in all patients. Autoimmunity plays a part in the pathogenesis of pemphigus vulgaris, with IgG antibodies targeted to normal epithelial membrane glycoproteins (66, 150, and 210 kDa). Autoantibodies may bind to keratinocytes and cause a loss of interepithelial adhesion. A significant association has been established with HLA DR4 and DRW6 in patients with pemphigus and either of these gene products may confer disease susceptibility.

Pathology

This shows loss of interepithelial adhesion, intraepithelial bullae, and acantholytic cells, with a diffuse leucocytic infiltration of the lamina propria.

Clinical features (Fig. 7)

The disease affects females two to three times as often as males, usually those over the age of 30 years. Painful, fluid-filled blisters or bullae may appear in any part of the mouth and burst within a few hours, resulting in shallow ulcers. These persist for weeks or months, but new lesions appear throughout the disease process. Oral manifestations of the disease may persist for many months, without overt ill health, but skin lesions, malaise, and loss of weight may occur at a later stage.

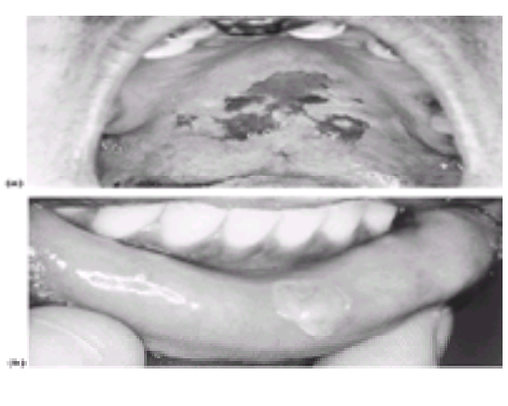

Fig. 7 (a) Bullae and erosions of pemphigus vulgaris affecting the palate. (b) Bullae and erosions of pemphigus vulgaris affecting the lower lip.

Differential diagnosis

Clinically the lesions are differentiated from recurrent aphthous ulcers by the presence of bullae, and when these ulcerate the edges lack the well-defined character of aphthous ulcers. Only occasionally is the Nikolsky sign helpful—rubbing the mucosa to induce a bulla. The most important diagnostic test is the presence of acantholytic cells on microscopic examination of direct scrapings from the lesion and a biopsy must always be taken. Antibodies to interepithelial antigens assist in the diagnosis. Pemphigus must be differentiated from pemphigoid and dermatitis herpetiformis (see below).

A less severe and rather rare variant of pemphigus vulgaris is pemphigus vegetans. Vegetation may be found on the oral mucosa and lips, and histological examination shows intraepithelial abscesses containing numerous eosinophils.

Treatment

Systemic corticosteroids such as prednisolone are given initially in doses of 40 to 60 mg/day and this is gradually reduced to the minimal dose that will prevent formation of new lesions. In order to keep the steroid dose to a minimum azathioprine can also be used, with a dose of 200 mg/day.

Course and prognosis

Treatment with corticosteroids must be maintained for life and has completely changed the prognosis of the disease. Patients rarely die now from the disease but they may develop the side-effects of steroid therapy.

Benign mucous membrane pemphigoid

Aetiology

This is a rare disease, affecting women twice as often as men, usually those over the age of 40 years. The aetiology is ill understood but there is some evidence that autoantibodies to the epithelial basement membrane may play a part in this disease.

Pathology

This shows subepithelial bullae, and the epithelium tends to detach itself from the underlying lamina propria. IgG, IgA, or IgM, with or without complement, are found in the basement membrane.

Clinical features

Bullous lesions involve the oral mucosa, conjunctiva, and the skin around the genitals, but in some patients only the mouth is involved. The bullae rupture within a day or two leaving erosions and ulcers. The gingiva is commonly involved, giving rise to persistent pain, bleeding, and a diffuse, raw, fiery red lesion. Other mucous membranes can be involved, such as the nose, larynx, pharynx, oesophagus, vulva, vagina, penis, and anus. The oral lesions usually heal without scarring unlike

those of the conjunctiva.

Differential diagnosis

Benign mucous membrane pemphigoid can be differentiated from pemphigus vulgaris on clinical grounds but only a biopsy examination will establish the diagnosis. There are no acantholytic cells and the bullae are subepithelial and not suprabasilar. Furthermore, autoantibodies can be detected, probably in fewer than half of patients, binding to the basement membrane of epithelium and not to the interepithelial substance. The disease should be differentiated from linear IgA disease, which shows linear deposition of IgA in the basement membrane, and dermatitis herpetiformis, in which IgA deposits are found in the papillae.

Treatment

If the disease is confined to the mouth topical corticosteroids are often adequate to control the lesions. However, when other sites are involved systemic corticosteroids are indicated, as in pemphigus.

Course and prognosis

This is a chronic disease which persists, often with exacerbations and remissions, over many years. The conjunctivitis may result in adhesions, corneal opacity, and blindness.

Erythema multiforme

Aetiology

Erythema multiforme may develop at any age but often occurs in young males. Many agents have been associated with this disease—drugs, such as sulphonamides and barbiturates, microbial infections, especially with herpes simplex virus—but a large proportion appears to be idiopathic.

Pathology

There is intracellular oedema with a zone of liquefaction degeneration of the upper layers of epithelium. Subepithelial bullae are often present and the lamina propria is infiltrated with lymphocytes, monocytes, neutrophils, and eosinophils.

Clinical features (Fig. 8)

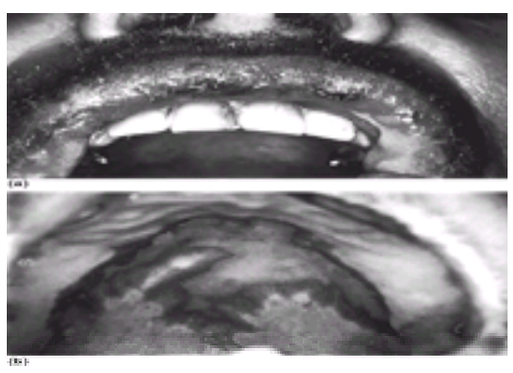

Fig. 8 Erythema multiforme: (a) haemorrhagic crusted upper lip; (b) diffuse erosion of the palate.

Oral manifestations may not be a significant feature. However, the mouth can be affected without skin involvement and the diagnosis is then more difficult. The patient develops painful, extensive erosions and ulcers with a predilection for the palate, tongue, and cheeks. The gum may show extensive erosions, which tend to bleed. Haemorrhagic crusting of the lips is often seen. A severe variant of erythema multiforme, which affects the eyes and genitalia in addition to the skin and mouth, is referred to as Stevens–Johnson syndrome.

Differential diagnosis

The diagnosis of oral lesions without the typical skin manifestations can be difficult. The clinical features to note are the extensive erosions affecting the palate, tongue, cheeks, and gingiva, and the haemorrhagic crusting of the lips. These features should avoid confusion with aphthous ulcers. An association with drugs or microbial infection is helpful in making the diagnosis. The age and sex prevalence differs from that in benign mucous membrane pemphigoid. A biopsy examination can formally exclude pemphigus and erosive lichen planus. The differential diagnosis of Stevens–Johnson syndrome from Behçet's syndrome has been discussed and the points noted about Reiter's syndrome also apply here (see Chapter 18.6).

Treatment

Whenever possible the offending drug or infection should be eliminated. The oral lesions often respond to topical tetracycline. Treatment with systemic corticosteroids may be indicated for extraoral manifestations.

Course and prognosis

If the causative agent is not found the lesions may recur over many years and cause a great deal of discomfort. In Stevens–Johnson syndrome, blindness may result from intercurrent bacterial infection.

Lichen planus

This is a disease that may affect the skin, the mouth, or both mucocutaneous surfaces (see Chapter 23.1).

Aetiology

Although the prevalence of oral lichen planus is not known it is surprisingly common in adults. Very little is known about its aetiology, but the condition can develop in graft versus host reaction after bone marrow transplantation. Several drugs are capable of inducing lichenoid changes in the mouth (for example penicillinase, colloidal gold). It seems to be associated with emotional or psychiatric stress. However, in most patients no cause can be determined.

Pathology

The pathological changes are hyperkeratosis, hyperplasia, and a characteristic liquefaction degeneration of the basal cell layers of the epithelium. The lamina propria shows a well-defined lymphomonocytic infiltration.

Clinical features (Fig. 9)

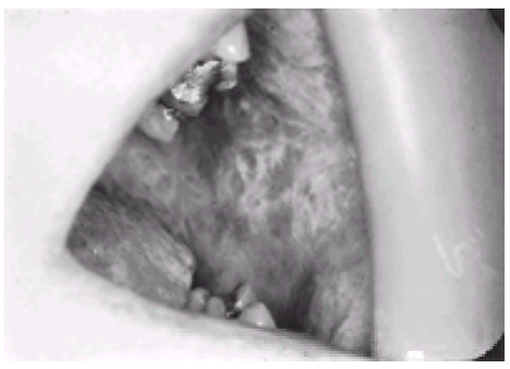

Fig. 9 Striae of lichen planus affecting the buccal mucosa and tongue.

In the mouth the lesions may remain symptomless for years and not infrequently they are first noticed by a dentist during routine examination. Some patients complain of a furry thickening of the mucosa and others of pain or bleeding from the gums on eating. There are three types of oral lichen planus: hypertrophic, erosive, and bullous. The hypertrophic variety is most common and is usually seen in all three types. There are white striae and minute papules, most commonly affecting the posterior part of the buccal mucosa, lips, and dorsum of tongue, though the palate, gum, and floor of the mouth are also involved. The striae crisscross giving rise to a fine lacy or fern-like pattern, and less commonly a honeycomb or annular patter. At times the striae may fuse together and result in a diffuse, somewhat smooth, shiny white plaque which may be difficult to differentiate from leucoplakia. Indeed, the dorsum of the tongue usually manifests diffuse white patches instead of the striated pattern.

In bullous lichen planus a bulla is rarely seen, presumably because it bursts to produce ulcers. Erosive lichen planus, however, is common, and patients complain of pain and discomfort on eating. There may be large shallow ulcers up to 3 cm in size surrounded by white striae and papules. The favoured sites are the same as in the hypertrophic variety, and whilst the latter may break down to result in erosive lichen planus, it is remarkable how often the hypertrophic variety remains unchanged. Except for discomfort, difficulties with eating, and occasionally loss of weight, there are no general manifestations and the regional lymph nodes are not enlarged, except with secondary infection. Not infrequently lichen planus may affect only the gum, inducing a diffuse, fiery red gingivitis and scattered erosions. This is a particularly troublesome type of lichen planus, referred to as desquamative gingivitis, with pain and bleeding, and tends to be resistant to treatment. It should be stressed that many patients with oral lichen planus do not have skin lesions.

A great deal of attention has been paid to the potential for carcinomatous transformation of lichen planus. Applying critical criteria, 1 to 2 per cent of oral lesions of lichen planus may transform to squamous cell carcinoma.

Differential diagnosis

The striae and papules of lichen planus are sufficiently distinctive features in the mouth to differentiate lichen planus from other lesions without the necessity for a biopsy examination. However, the diffuse hypertrophic variety can be confused with leucoplakia and then a biopsy is helpful. Erosive lichen planus may very occasionally lack the distinctive striae, and then erythema multiforme and benign mucous membrane pemphigoid should be excluded. Both systemic and discoid lupus erythematosus can present in the mouth as central erosions surrounded by a keratinized margin.

Treatment

In the absence of symptoms, hypertrophic lichen planus does not require any treatment. The patient, however, needs to be appraised as to the nature of the disease. Topical corticosteroids are usually effective in the treatment of erosive lichen planus but also suppress the striae and papules of the hypertrophic variety. Triamcinolone in orabase ointment applied three to four times a day is useful in localized lesions, but betamethasone (as sodium phosphate) is more effective and is usually used in the form of 0.5 mg tablets, kept in the mouth three times daily. An alternative is to use an aerosol inhaler containing beclomethasone, and applying four to six puffs daily. For these drugs to be helpful, they must be applied for one to several months. The lesions almost invariably recur, although the length of remissions varies greatly and corticosteroids may have to be applied with every remission.

Cleaning the teeth tends to be painful and the accumulation of a large amount of dental plaque aggravates the gingivitis. The patient should use a very soft toothbrush and needs to have the teeth scaled every 3 to 6 months. Chlorhexidine mouthwash can be helpful in controlling dental plaque.

Course and prognosis

The disease is chronic and tends to persist for years, with natural remissions and exacerbations. Topical corticosteroids prolong the remissions, and the erosions and discomfort are kept under control. Since carcinomatous transformation, especially of the erosive type of lichen planus, can take place in a small proportion of patients, they should be followed up regularly at a stomatological clinic.

Leucoplakia

White patches of the oral mucosa that cannot be removed by scraping are referred to as leucoplakia. By convention, lichen planus and lupus erythematosus are excluded from this group.

Aetiology

The prevalence of leucoplakia is not known, but it seems that during the past two decades it has become less frequent. There are many causes of leucoplakia and as these may have distinctive features they will be classified below. It should be noted, however, that in about half the leucoplakias a cause cannot be found. Syphilitic, candidal, and AIDS leucoplakias have been discussed, elsewhere. Causes include:

1. Physical and chemical agents: frictional keratosis, smoker's keratosis.
2. Microbial infection: chronic hyperplastic candidiasis, tertiary syphilis, and AIDS.
3. Congenital and hereditary leucokeratosis.
4. Idiopathic causes.

Pathology

The microscopic features of leucoplakia show a spectrum of changes; at the benign end is epithelial keratosis alone, followed by hyperplasia, and then epithelial atypia at the premalignant end. The lamina propria shows a parallel increase in mononuclear cells, especially plasma cells. Carcinoma *in situ* is the least common histological finding.

Clinical features

The white patches vary from a soft, slightly thickened mucosa, involving a small or large mucosal surface, to hard, irregular white plaques with intervening normal, erosive, or ulcerated sites. The latter is often referred to as speckled leucoplakia and must be recognized clinically because of its greater propensity to carcinomatous

transformation. Any part of the oral mucosa or gum may be involved but the cheeks and tongue are most often affected.

Frictional keratosis is usually found along the occlusal line of the buccal mucosa and presents as a linear white patch of even consistency.

Smoker's keratosis (Fig. 10) shows a characteristic distribution of the soft and adjacent hard palate, as keratinized papules with central red dots. The distribution is due to involvement of the palatal mucous glands and the red dots are the openings of the ducts. It is usually caused by pipe smoking, but cigarette smoking may also lead to keratosis of a diffuse type, most commonly affecting the cheeks.

Fig. 10 Smoker's keratosis of the palate.

Congenital and hereditary leucokeratosis can be distinguished by the presence of diffuse, soft, white plaques, often with a folded surface. The lesions tend to be symmetrical and affect the floor of the mouth. Other members of the family may have similar lesions.

Differential diagnosis

All leucoplakias should be biopsied, except smoker's keratosis of the palate, as even small white patches have at times proved to be early carcinomas (Fig. 11). It is also essential to find out the degree, if any, of epithelial atypia as this affects the prognosis of leucoplakia. Direct examination of scrapings can be helpful in the presence of candida hyphae; cultures should also be set up for candida. Serological tests can further aid in the diagnosis of candidiasis but are essential in the diagnosis of syphilitic leucoplakia.

Fig. 11 Leucoplakia of the tongue, which on biopsy examination showed a well differentiated squamous cell carcinoma.

Treatment

Smoker's keratosis is reversible in many instances if the patient gives up smoking. Frictional keratosis can also be cleared if some local cause of irritation is removed. Candida leucoplakia should be treated with topical antifungal drugs, though this rarely results in permanent clearance of the lesion. Syphilis should be managed by a course of penicillin and stringent follow-up, so as to detect any carcinomatous transformation early. Leucoplakia showing evidence of epithelial atypia should be excised and if the lesion is large a skin graft may be required. However, in many cases the lesion recurs, even after repeated excision. There is no satisfactory treatment for leucoplakia and the most important point is long-term follow-up so as to detect in time the development of an incipient carcinoma.

Course and prognosis

Leucoplakia may persist for life, without any discomfort or change. However, about 5 per cent of all leucoplakias may undergo malignant changes and this figure increases to about 30 per cent in leucoplakias showing histological evidence of epithelial atypia. It seems that epithelial atypia is more commonly associated with speckled leucoplakia, and the latter as well as syphilitic leucoplakia has a worse prognosis. In contrast, smoker's keratosis and frictional keratosis have a very good prognosis if the offending cause is removed. Congenital or hereditary leucokeratosis is thought to be free of malignant changes, although recently a few cases with carcinomatous transformation have been reported.

Benign neoplasms, cysts, and developmental and inflammatory lesions of the soft tissues

There are numerous benign neoplasms and soft tissue lesions of the mouth. This section will be restricted to some essential features of the following lesions: papilloma, fibroma, lipoma, neurofibroma, hamartoma, pigmented naevus, lymphangioma, denture granuloma, giant cell reparative granuloma, fibrous polyp, pregnancy tumour, mucous retention, and extravasation cysts.

Aetiology

The cause of benign neoplasms is unknown and the parts that physical or chemical irritation and microbial infection may play are ill understood. Mucous retention or extravasation cysts are caused by trauma or obstruction of the orifices of the ducts of the minor salivary glands. Whereas true benign neoplasms are rare, inflammatory lesions and cysts are commonly found in the mouth.

Clinical features

The soft tissue tumours present as painless, slow-growing swellings affecting any part of the mouth, but if they originate from the gum they are referred to as epulides. Fibrous polyps are the most common inflammatory lesions of the oral mucosa and result from trauma or irritation from rough edges of carious teeth. Most of the tumours are sessile, some are pedunculated as with some fibromas, and others are flat and pigmented as with the naevi. They are usually symptomless except for bleeding from hamartomas and giant cell reparative granulomas.

Differential diagnosis

There are some distinguishing clinical features, but the definitive diagnosis will depend on the histological examination of the excised specimen. A papilloma can be

recognized by its firm, small, keratinized, finger-like processes. Lymphangiomas are soft swellings which may cause considerable enlargement of the lip or tongue. Hamartomas are flat or nodular red lesions that may blanch when compressed; they are occasionally confused with pregnancy tumours, which are rather vascular granulomatous swellings of the gingiva found during pregnancy. Giant cell reparative granulomas are also very vascular, maroon-coloured lesions originating from the gingiva. Denture granulomas can be readily recognized from their relation to the flange of a denture; the lesion is often elongated, and can be indented or ulcerated by the denture. Mucous retention or extravasation cysts are small, often bluish, swellings affecting the lips or cheeks.

Treatment

Surgical excision, with a margin of normal tissue at the base of the lesion, is usually indicated. Pregnancy tumours, however, commonly regress spontaneously.

Course and prognosis

The soft tissue neoplasms will enlarge over the years and interfere with the normal functions of the mouth. Bleeding from any of the lesions is rarely profuse. Only the giant cell reparative granuloma has a tendency to recur after excision.

Oral carcinoma

Aetiology

Carcinoma of the mouth accounts for about 2 per cent of all cancers in Britain and the United States. The prevalence increases significantly after the age of 40 years and more than twice as many men as women are affected. The incidence of oral cancer, however, in India and Sri Lanka may account for about 40 per cent of all cancers. As in other carcinomas the cause is unknown, but smoking and alcohol have been implicated. There is some epidemiological evidence to support this, but unlike lung cancer it is pipe or cigar rather than cigarette smoking that have been associated with oral cancer. The association with chronic oral sepsis and irritation has not been critically examined. There is some evidence that microbial agents, particularly *Treponema pallidum*, *Candida albicans*, human papillomavirus, and HIV, may directly or indirectly influence the development of carcinoma.

Among the predisposing lesions, leucoplakia is the best-known; in 5 per cent of all patients and in about 30 per cent of those showing evidence of epithelial atypia the leucoplakia may undergo carcinomatous transformation (Fig. 12). Submucous fibrosis is another precancerous condition and is found predominantly in India and Sri Lanka. It seems to be related to eating chillis and possibly chewing betel nuts; it affects the palate, buccal mucosa, and tongue.

Fig. 12 Leucoplakia of the buccal mucosa, the lower edge of which is raised and on biopsy proved to be a well-differentiated squamous cell carcinoma.

Pathology

Squamous cell carcinoma in the mouth is usually a well-differentiated keratinizing neoplasm invading the surrounding tissue. Poorly differentiated, anaplastic oral carcinomas are much less frequent and are especially rare with carcinoma of the lip. Spread occurs by local invasion: lymph node metastasis is less common than is generally thought, and occurs at a late stage.

Clinical features

The presenting features of carcinoma vary with the site of involvement but there are two types, a lump or an ulcer. The patient complains of a swelling or ulcer that is resistant to healing and gradually enlarging in size. There may be little pain initially, but at a later stage discomfort and occasional bleeding may occur. Cancer of the tongue may give rise to local pain and earache. Whereas some patients complain of an excess of saliva, especially with the larger tumours, a dry mouth may be found during the early stages of malignant change and should be noted as another feature favouring malignancy. A small lump may enlarge to a hard swelling before the covering mucosa breaks down. A malignant ulcer shows a raised and often everted edge, and the most important feature is induration at the base of the lesion. Any part of the mouth can be involved but the lips (usually the lower lip) and tongue are most common, each accounting for about 25 per cent of oral carcinomas. The floor of the mouth, gingiva, cheek, hard and soft palate, and oropharynx may account for about 10 per cent of the carcinomas. In most patients there is only one lesion but some patients may have two or even multiple carcinomas. Metastasis may occur at a late stage to the submandibular or upper cervical lymph nodes, and occasionally to the submental nodes.

Differential diagnosis

Any long-standing or indurated lesion in the mouth, especially of elderly or middle-aged patients, should be queried for malignancy and biopsy examination is essential. A traumatic ulcer caused by a denture can be confused with a malignant ulcer, but it may lack induration, the offending part of the denture may fit into the ulcer, and removing the denture for about a week may bring about healing of the lesion. Major aphthous ulcers have been mentioned elsewhere (see above), but the salient differentiating features are a history of recurrent ulcers at different sites of the mouth over many years.

Adenocarcinoma of the small salivary glands may present as a lump of the soft palate, lips, or cheeks and only a biopsy will establish the diagnosis firmly. Carcinoma *in situ* is rare in the mouth, but it may present as a diffuse, erythematous, somewhat velvety lesion, affecting a part of the soft palate or cheek. Again a biopsy examination must be carried out for diagnosis.

Treatment

The principles of treatment of oral carcinomas are those applied to other carcinomas of the body. Surgical excision of the lesion and a margin of adjacent healthy tissue is the most common practice, and this may be extended if necessary to block dissection of the regional lymph nodes. Radiotherapy is an alternative approach and is commonly used in primary treatment of cancer of the lip, in inoperable cases, or with recurrent carcinoma following surgery. Chemotherapy is used less often in the management of cancer of the mouth and the results are variable. Management of oral cancer is a complex subject outside the scope of this section. It should be emphasized that oral hygiene is particularly important with any treatment so as to avoid ascending parotitis. A dry mouth usually follows radiotherapy and again meticulous oral hygiene should be advised, so as to prevent rampant caries and candida infection.

Course and prognosis

The 5-year survival rates differ considerably with the anatomical site of the cancer. Carcinoma of the lip has by far the best prognosis, irrespective of whether treatment is by surgery or radiotherapy, and the 5-year survival rate is about 80 per cent. In contrast the figures for carcinoma of the tongue range from 25 to 35 per cent, floor of the mouth 20 to 40 per cent, cheek 30 to 50 per cent, and oropharynx, palate, and gingiva at about 25 per cent. The prognosis is significantly better in

the absence of lymph node involvement.

Diseases of the salivary glands

Xerostomia

Xerostomia is a term describing dryness of the mouth and can be due to a variety of conditions.

Aetiology

Dry mouth is a common manifestation, especially in middle-aged women, and can be caused by anxiety and emotional and mental stress. Iatrogenic xerostomia is secondary to a number of drugs, the most common of which are antihistamines, tricyclic and other antidepressants, phenothiazine, hypotensive agents, diuretics, and preparations containing atropine. Another common cause is secondary to radiotherapy, but the salivary flow tends to recover although it may take many months. Some diseases affect the salivary glands directly and cause dryness of the mouth, for example Sjögren's syndrome and sialadenitis. Another large group of agents cause xerostomia by inducing changes in fluid balance; diabetes, anaemia, dehydration, and oedema are common examples.

Pathology

Diseases affecting the salivary glands cause a destruction of the secretory components by mononuclear cell infiltration and fibrosis of the salivary acini.

Clinical features

The patient complains of dryness of the mouth and sometimes the eyes, soreness of the mouth, especially the tongue and throat, and discomfort on swallowing of solids and at times difficulty in speaking. The most convincing clinical evidence of xerostomia is an atrophic, dry oral mucosa, often fiery red, due to infection by candida. Inspection of the duct orifices of the major salivary glands will fail to reveal salivary flow. Whole salivary flow rates are readily determined by collecting unstimulated (resting) or lemon juice stimulated saliva. Arbitrary levels of resting saliva of less than 0.1 ml/min and stimulated saliva of less than 0.5 ml/min are indicative of impaired salivary function. The patient may develop rampant caries or if he or she wears dentures there may be difficulties with retention.

Differential diagnosis

A thorough history may establish psychogenic or iatrogenic causes and diseases affecting fluid balance. Sialography and labial gland biopsy may be necessary in the diagnosis of Sjögren's syndrome, though a raised erythrocyte sedimentation rate, rheumatoid factor, antinuclear factor, autoantibodies, and HLA typing may assist in the diagnosis. Nevertheless there will be a large proportion of patients in whom a specific cause cannot be found.

Treatment

Management of the patient is clearly directed to elimination of the cause of xerostomia but this may be difficult or at times impossible to achieve. In such cases, palliative measures are helpful and these include frequent sips of water, meticulous oral hygiene, and early treatment or preferably prevention of candidiasis by topical nystatin or amphotericin B. Each patient responds differently; some prefer glycerin as a lubricant, others carboxymethylcellulose, and the latter can be taken as a solution or spray (Glandosane). A mucin preparation can also be helpful as a spray or as a lozenge (Saliva Orthana).

Sialadenitis

Bacterial or viral infections and rarely allergic reactions may cause inflammation of the salivary glands. These agents may give rise to acute, chronic, or allergic sialadenitis, and recurrent parotitis.

Aetiology

Ascending infection of the parotid gland used to be a common complication in elderly, postoperative patients who were predisposed by dehydration, reduced salivary flow, and lack of oral hygiene. Acute parotitis may also follow the use of drugs causing xerostomia. The most common micro-organisms involved are *Staphylococcus aureus*, *Streptococcus viridans*, and pneumococcus. The most common acute parotitis is mumps (see Section 7). The salivary glands are sometimes affected by HIV infection, with an enlargement of the parotid glands. Chronic sialadenitis is usually associated with duct obstruction and therefore affects the submandibular gland. Recurrent sialadenitis is a disease of unknown aetiology and may be associated with a decreased salivary flow causing retrograde infection. The disease may affect both adults and children.

Pathology

Acute sialadenitis shows an acute inflammatory reaction of the salivary tissue with a predominantly neutrophil infiltration, except in mumps in which there is an infiltration by mononuclear cells. In both chronic and recurrent sialadenitis there is a marked periductal and acinar infiltration by mononuclear cells, with some epithelial hyperplasia of the duct accompanied by acinar atrophy and fibrosis.

Clinical features

The presenting symptoms of acute sialadenitis are a painful swelling in one of the parotid glands of an elderly patient. Commonly the patient has a low-grade fever, oedema of the cheek, some trismus, and a purulent discharge may be expressed from the duct opening. In contrast, mumps affects healthy children and young adults.

In chronic sialadenitis there are usually clinical features of obstruction of the duct of one of the submandibular glands. There is pain and swelling in the submandibular or retromandibular region, with a reddened duct orifice discharging pus. Recurrent parotitis presents as an acute pain and swelling of one or both parotid glands, with erythema of the duct orifices and pus discharging from them. There may be an associated fever and malaise. Recurrences vary from weeks to months and after repeated attacks the affected gland may remain enlarged.

Differential diagnosis

There is little clinical difficulty in the differential diagnosis between acute sialadenitis of the parotid gland in the elderly patient due to ascending infection and mumps in the healthy young subject. Any discharging pus should be cultured for organisms and its antibiotic sensitivity should be determined. Recurrent parotitis, however, can cause difficulties; in addition to a history of recurrent painful swelling and discharging pus, sialography may show sialectasis and duct dilatation. In chronic sialadenitis there is usually clinical or radiological evidence of calculus and sialography may show duct dilatation.

Several granulomatous diseases may very occasionally affect the salivary glands, such as sarcoidosis, tuberculosis, syphilis, and actinomycosis. When there is bilateral salivary and lacrimal enlargement this is often referred to as Mikulicz's syndrome. Allergic sialadenitis is also rare and to determine the allergic agent can be difficult as drugs, foods, pollen, and other agents have been implicated.

Treatment

In acute, chronic, or recurrent sialadenitis the relevant antibiotics should be used to control the infection, but occasionally surgical drainage may also be necessary. Careful oral hygiene measures are important in all types of sialadenitis. In chronic sialadenitis the cause of obstruction, such as a calculus, should be removed. The treatment of recurrent parotitis is more difficult and if antibiotics do not control the disease, surgical intervention should be considered.

Course and prognosis

Acute sialadenitis will resolve with the aid of antibiotics and general management of the patient. Chronic sialadenitis may persist for many years and may lead to destruction of the gland unless the cause of duct obstruction is removed early. Recurrent parotitis in childhood may show spontaneous recovery after puberty.

Salivary duct obstruction due to calculus

Aetiology

The submandibular salivary ducts and, to a lesser extent, glands are the most common sites for the development of stones. Calcium phosphates and carbonates are deposited from the saliva round a nidus of desquamated cells or micro-organisms.

Clinical features

Salivary calculus is usually found in adults and the presenting symptoms are a sudden unilateral swelling and pain of the gland related to eating. The swelling may take minutes to appear and hours to subside. Examination reveals a soft swelling of the affected gland and careful digital palpation along the course of the salivary duct will localize the calculus. This may vary in size from a small grain to a concretion 10 to 20 mm in length. The presence and localization of a stone in a duct needs to be confirmed by radiographs, but the presence of calculi in the gland can be diagnosed only by radiography. Over 80 per cent of calculi are found in the submandibular duct or gland.

Differential diagnosis

Recurrent unilateral swelling associated with eating is characteristic of salivary gland obstruction but occasionally this may be caused by external agents. Trauma from a denture or sharp tooth may cause obstruction of the orifice of the parotid duct.

Treatment

If the calculus is near the orifice of the duct it can occasionally be teased out, otherwise surgical removal is indicated.

Course and prognosis

Single calculi do not tend to recur, but if treatment has been delayed numerous calculi may have formed inside the gland which may occasionally have to be excised.

Salivary gland tumours

A variety of epithelial tumours affect the major and minor salivary glands, of which the commonest is pleomorphic adenoma, or mixed salivary tumour (74 per cent), followed by adenocarcinoma (12 per cent), adenoma (8 per cent), mucoepidermoid tumour (3 per cent), and acinic cell tumour (2 per cent); the percentages give the prevalence in the parotid glands. Only pleomorphic adenoma will be considered in any detail and further reading should be consulted for other tumours.

Pleomorphic adenoma

Aetiology

The cause of this tumour is unknown, although salivary gland tumours can be produced in animals by carcinogenic hydrocarbons, polyomavirus, and other agents. The tumour originates from epithelial cells of the ducts, acini, or myoepithelial cells and these are thought to be capable of producing the stromal mucins of this tumour.

Pathology

The epithelial cells proliferate to form duct-like structures, sheets, and cords within a connective tissue stroma, which may show mucous, cartilaginous, or hyaline appearance. The tumour is encapsulated, although satellite tumours are often found outside the capsule.

Clinical features

The tumour is usually found in adults and the parotid salivary gland is most commonly affected, followed by the submandibular gland and rarely the sublingual gland. The minor salivary glands, however, are also affected, and the most frequent sites are the glands of the palate, lips, and cheeks. The tumour presents as a small, painless swelling, which may take years to enlarge and is not attached to the overlying skin or mucosa.

Differential diagnosis

As the tumour is slow growing it needs to be differentiated only from other tumours. Adenocarcinoma, mucoepidermoid carcinoma, and adenoid cystic carcinoma may mimic pleomorphic adenoma in its slow growth, but some may grow more rapidly, invade the adjacent skin or mucosa, and metastasize. These tumours can often be differentiated only on histopathological examination, and wherever possible an excision biopsy should be done.

Treatment

Surgical excision with a margin of normal tissue is the treatment of choice, as the tumour is radioresistant.

Course and prognosis

If left untreated the tumour may enlarge to a grotesque size. A small proportion of pleomorphic adenomas may undergo carcinomatous transformation. The tumour has a bad record for recurrences after excision and this is thought to be due to leaving behind satellite tumours outside the capsule.

Neoplasms, cysts, developmental lesions, and dystrophies of the bones and teeth

This section covers a very large number of lesions found in the jaws. Only essential features, especially of differential diagnosis, will be covered in the following disorders:

1. benign neoplasms: osteoma, chondroma, fibroma, ossifying fibroma, and giant cell tumour;
2. malignant neoplasms: osteosarcoma and chondrosarcoma;
3. cysts and tumours of dental origin: periodontal and dentigenous cysts, keratocysts, and ameloblastoma;
4. dental malformations or odontomes;
5. osteodystrophies: giant cell reparative granuloma, brown tumour of hyperparathyroidism, fibrous dysplasia, and Paget's disease.

Aetiology

The cause of the neoplasms and osteodystrophies is not known. Periodontal cysts, which are the most common lesions in this group, develop as a consequence of

chronic periapical infection.

Clinical features

The bony tumours and cysts are commonly symptomless unless they have reached a large size and the patient notices a swelling, or a denture ceases to fit. Pathological changes are often noticed by the dentist through movement of teeth or on routine radiographic examination of the teeth. Hyperparathyroidism should be excluded in cases when a giant cell granuloma is suspected. Cysts can be found at any age, but giant cell reparative granulomas, ossifying fibroma, and fibrous dysplasia are often seen in young people, unlike Paget's disease of bone which is seen only in the elderly. There is a predilection for the mandible to be involved more commonly with ossifying fibroma and giant cell reparative granuloma. Odontomes are developmental malformations of dental tissues that become calcified. This is a diverse group of disorders and varies from a simple enamel pearl, consisting of a nodule of ectopic enamel attached to a tooth, to a complex composite odontome, which is an irregular mass of calcified dental tissues. Ameloblastoma is a rare but important epithelial neoplasm of the jaws. Young adults are most often affected, the tumour is slow-growing, and affects the mandible more often than the maxilla. The neoplasm is locally invasive but does not metastasize. Osteosarcoma and chondrosarcoma are found in children or young adults but may develop in the elderly with Paget's disease. They present as fast-growing, painful, and firm swellings and they may metastasize to the lungs early.

Differential diagnosis

The diagnosis of bony lesions of the jaws is made on the basis of radiological appearances and the histological features of the biopsy. Periodontal cysts are very frequent and show a radiolucent rounded area with a sharply defined outline. If the crown of a tooth is enclosed within the cyst, it is referred to as a dentigerous cyst. The latter and keratocysts are usually found in the young, but with some keratocysts a tooth may be missing. Dental cysts must be differentiated from ameloblastomas, which tend to show multilocular and sometimes a honeycomb pattern on radiographs. These radiolucent lesions should also be differentiated from secondary carcinoma and myelomatosis. Giant cell reparative granuloma and tumour (osteoclastoma) show a radiolucent area, sometimes loculated, and the outline is not as well defined as a dental cyst. Hyperparathyroidism can be excluded by the radiographic appearance of other bones and by the calcium and phosphate levels in the blood. Ossifying fibromas are more common than fibromas and radiographs show a well-defined radiolucent area with speckled calcification. This can usually be distinguished from the 'ground glass' appearance, without a distinct border, found in fibrous dysplasia. In Paget's disease there is a distinctive 'cotton wool' appearance on radiographic examination and the alkaline phosphatase levels are high. Odontomes can be readily recognized on clinical examination, but those that are unerupted, particularly the compound and complex composite odontomes, show on radiographs a mass of overlapping denticles and an irregular radio-opaque mass respectively. Osteosarcoma and chondrosarcoma show patchy areas of bone resorption and deposition.

Treatment

The treatment of dental cysts is by enucleation of the cyst lining and usually extracting the involved tooth. The tumours and malformations are excised but some, such as giant cell reparative granuloma, can be curettaged. Brown tumours will recur unless the underlying hyperparathyroidism has been treated. Fibrous dysplasia may require removal of excessive tissue for cosmetic or functional reasons, but this should be delayed until normal bone growth has ceased. Bony changes in Paget's disease are best not interfered with, except when there are functional reasons such as inability to fit a denture. Composite odontomes should be removed surgically. The treatment of ameloblastoma is by local excision, with a generous margin of normal bone, or by hemimandibulectomy. Sarcoma of the jaw must be dealt with by early radical excision.

Course and prognosis

If the cysts or benign tumours are removed surgically they do not recur, except with keratocysts and the reparative granulomas. Ameloblastomas may recur after several excisions, without metastases, and this is why some surgeons prefer to do a hemimandibulectomy. The prognosis of the jaw sarcomas is very poor and the 5-year survival rate is between 25 and 40 per cent. Fibrous dysplasia tends to be self-limiting, but in Paget's disease there may be progressive enlargement, especially of the maxilla.

Miscellaneous disorders

In this section a brief discussion will be given on the following three topics: oral manifestations of blood disorders, halitosis, and disorders of the temporomandibular joint.

Oral manifestations of blood disorders

Mild anaemias or deficiencies of iron, folate, or vitamin B_{12} may manifest themselves as glossitis (Fig. 13) with a sore tongue or mouth, angular cheilitis, or recurrent ulceration (see Section 22). The tongue is commonly depapillated, the corners of the mouth may be inflamed and fissured, and occasionally there may be small shallow ulcers affecting the lips, tongue, and cheeks. The cause of any haematological deficiency should be investigated and, especially with folate deficiency, coeliac disease should be excluded. Replacement therapy usually deals with the clinical features effectively. It should, however, be emphasized that the complaint of a sore tongue can be associated with many other causes, such as erythema migrans, candidiasis, lichen planus, recurrent aphthous ulceration, and black hairy tongue.

Fig. 13 Smooth, depapillated, erythematous tongue in a patient with iron deficiency anaemia.

Erythema migrans (geographical tongue) is particularly common and is characterized by oval, depapillated areas with a well-defined edge affecting the dorsum of the tongue (Fig. 14). The lesions move from one site to another. The aetiology of erythema migrans is unknown and treatment is rather unsatisfactory. It is noteworthy that a sore tongue is a frequent complaint in middle-aged women, often without any demonstrable aetiological factor.

Fig. 14 Round or oval, depapillated lesions with a raised margin in a patient with erythema migrans of the tongue.

Acute leukaemia, particularly the myelomonocytic form, may occasionally present in the young in the form of sore, bleeding gums. This may vary from slight inflammation to that showing bulbous enlargement of the gingiva. There are usually inadequate local causes for such a gingivitis and anaemia may be evident; blood tests should be requested to exclude leukaemia.

Leucopenia and agranulocytosis, especially those due to drugs, may become clinically evident by ulceration of the throat or the mouth. Purpura may be associated with a deficiency of platelets, so that bleeding from the gum may also be a feature.

Many haemorrhagic disorders may become evident after extraction of a tooth, because bleeding does not stop. Less commonly, gingival bleeding may direct attention to the blood disorder.

Halitosis

Bad breath is usually a trivial complaint, though it is heightened by social pressures. There are four possible sources of halitosis: the mouth, nasopharynx, lungs, and the gastrointestinal tract. Altered blood round the gum may be the most important oral cause, and this may be associated with debris or pus from gingivitis and periodontal pockets. A characteristic halitosis is found in acute ulcerative gingivitis. It should be noted that bad taste and bad breath are subjective sensations which are often confused. Excessive bacterial plaque on the teeth is not a principal cause of halitosis; nevertheless, meticulous oral care should be advised.

Chronic tonsillitis may be responsible for halitosis but atrophic rhinitis causing ozaena is probably the most important cause to be excluded. Occasionally, respiratory tract infections may cause halitosis and a variety of gastrointestinal disorders have been associated with bad breath but there is little evidence to substantiate this. Frequently all these sources of halitosis may be excluded without finding a cause and these patients may have a fixation about bad breath related to emotional or sexual problems.

Temporomandibular joint disorders

The patient complains of pain, clicking, or limitation of movement. It is found in young women more often than men. Examination may reveal limitations in jaw movement, tenderness of the joint, and crepitus on movement, discovered by palpating the head of the condyle through the overlying skin. The cause is difficult to establish but malocclusion might be one of several factors. The condition may clear spontaneously but in some patients the occlusion should be checked and a bite-raising appliance is often helpful. Rheumatoid arthritis and osteoarthritis of this joint are occasionally seen clinically. Dislocation of the joint, which becomes fixed in the open position, may be caused by a blow on the jaw or during dental extractions under general anaesthesia. Ankylosis of the joint is nowadays extremely rare but in the past was caused by osteomyelitis.

Further reading

Atkinson JC *et al.* (1990). Major salivary gland function in primary Sjögren's syndrome and its relationship to clinical features. *Journal of Rheumatology* **17**, 318–22.

Bouquot JE, Weiland LH, Kurland LT (1988). Leukoplakia and carcinoma *in situ* synchronously associated with invasive oral/oropharyngeal carcinoma in Rochester Minn., 1935–1984. *Oral Surgery, Oral Medicine, Oral Pathology* **65**, 199–207.

Carlsson J (1989). Microbial aspects of frequent intake of products with high sugar concentrations. *Scandinavian Journal of Dental Research* **97**, 110–14.

Chau MN, Radden BG (1989). A clinical-pathological study of 53 intra-oral pleomorphic adenomas. *International Journal of Oral and Maxillofacial Surgery* **18**, 158–62.

Dummer PMH *et al.* (1990). Factors influencing the caries experience of a group of children at the ages of 11–12 and 15–16 years: results from an ongoing study. *Journal of Dentistry* **18**, 37–48.

Eley BM (1997). The future of dental amalgam: a review of the literature. Part 3. Mercury exposure from amalgam restorations in dental patients. *British Dental Journal* **182**, 331–8.

Fox PC, Busch KA, Baum BJ (1987). Subjective reports of xerostomia and objective measures of salivary gland performance. *Journal of the American Dental Association* **115**, 581–4.

Gibbons RJ (1989). Bacterial adhesion to oral tissues: a model for infectious diseases. *Journal of Dental Research* **68**, 750–60.

Goldberg HI *et al.* (1994). Trends and differentials from cancers of the oral cavity and pharynx in the United States, 1973–1987. *Cancer* **74**, 565–72.

Greenspan D, Greenspan JS (1996). HIV-related oral disease. *Lancet* **348**, 729–33.

Hasan A *et al.* (1995). Recognition of a unique peptide epitope of the mycobacterial and human heat shock protein 65–60 antigen by T cells of patients with recurrent oral ulcers. *Clinical Experimental Immunology* **99**, 392–7.

Helander SD, Rogers RD (1994). The sensitivity and specificity of direct immunofluorescence testing in disorders of mucous membranes. *Journal of the American Academy of Dermatology* **30**, 65–75.

Herrod HG (1990). Chronic mucocutaneous candidiasis in childhood and complications of non-Candida infection. *Journal of Pediatrics* **116**, 377–82.

Hogewind WF *et al.* (1989). The association of white lesions with oral squamous cell carcinoma: A retrospective study of 212 patients. *International Journal of Oral and Maxillofacial Surgery* **18**, 163–4.

Holmstrup P *et al.* (1988). Malignant development of lichen planus-affected oral mucosa. *Journal of Oral Pathology* **17**, 219–25.

Huilgol SC, Bhogal BS, Black MM (1995). Immunofluorescence of the immunobullous disorders. *European Journal of Dermatology* **5**, 186–95.

Kashima HK *et al.* (1990). Human papilloma virus in squamous cell carcinoma, leukoplakia, lichen planus, and clinically normal epithelium of the oral cavity. *Annals of Ontology, Rhinology and Laryngology* **99**, 55–61.

Larsson KS (1995). The dissemination of false data through inadequate citation. *Journal of Internal Medicine* **238**, 445–50.

Lehner T (1992). *Immunology of oral diseases*. Blackwell, Oxford.

Lloyd RE, Ho KH (1988). Combined CT scanning and sialography in the management of parotid tumors. *Oral Surgery, Oral Medicine, Oral Pathology* **65**, 142–4.

Lozada-Nur F, Gorsky M, Silverman S (1989). Oral erythema multiforme: clinical observations and treatment of 95 patients. *Oral Surgery, Oral Medicine, Oral Surgery* **67**, 36–40.

Newbrun E (1989). Frequent sugar intake—then and now: interpretation of main results. *Scandinavian Journal of Dental Research* **97**, 103–9.

Page RC *et al.* (1997). Advances in the pathogenesis of periodontitis: summary of developments, clinical implications and future directions. *Periodontology* **14**, 216–48.

Seaman S, Thomas FD, Walker WA (1989). Differences between caries levels in 5 year old children from fluoridated Anglesey and non-fluoridated mainland Gwynedd in 1987. *Community Dental Health* **6**, 215–21.

Van der Meij EH *et al.* (1999). A review of the recent literature regarding malignant transformation of oral lichen planus. *Oral Surgery, Oral Medicine, Oral Pathology* **88**, 307–10.

Van der Waal I *et al.* (1997). Oral leukoplakia: a clinicopathologivcal review. *Oral Oncology* **33**, 291–301.

Williams DM (1989). Vesiculobullous mucocutaneous disease: pemphigus vulgaris. *Journal of Oral Pathology and Medicine* **18**, 544–53.

Williams RC (1990). Periodontal disease. *New England Journal of Medicine* **322**, 373–82.

14.6 Diseases of the oesophagus

John Dent and Richard H. Holloway

Introduction

Oesophageal function testing

Techniques are now available for the precise measurement of oesophageal function. The benefits of oesophageal function testing are most apparent when it is the only means for the accurate diagnosis of a treatable disorder. However, it is also valuable for the recognition of disorders for which there is no definitive therapy, since symptoms can be explained. Oesophageal function testing is relatively expensive and time-consuming and so should not be requested for the assessment of trivial symptoms or when the information gained will not aid management.

Barium radiology

The barium swallow is often undervalued for the diagnosis of structural abnormalities. Useful information can also be obtained about the motor function of the pharynx and oesophagus by videotaping the images and analysing, in slow-motion replay, repeated tests of standardized stimuli. The sensitivity of barium radiology is enhanced by the use of provocative solid boluses, such as bread, or barium tablets. Videofluroscopy is essential for the proper analysis of pharyngeal motor function, and complements other tests of oesophageal motor function such as manometry. Optimal results from radiology are achieved when there is a partnership between a clinician and a radiologist who have a special interest in oesophageal motor disorders; if the particular question(s) that are being pursued by the clinician are well communicated to the radiologist the examination technique can then be tailored accordingly.

Oesophageal manometry

Oesophageal manometry provides the most direct indication of patterns of oesophageal motor function. It is most helpful in the diagnosis of dysphagia, after exclusion of fixed, structural defects. Manometry plays only a minor role in the management of reflux disease; it assists if there is no oesophagitis and mild dysphagia is present, or if surgery is planned. Prolonged, ambulatory 24-h manometry, which requires sophisticated miniaturized equipment, may be useful in patients with dysphagia or non-cardiac chest pain.

24-h ambulatory oesophageal pH monitoring

The principal value of this procedure is to determine the association between symptoms and episodes of acid reflux. Investigation of this relationship is only of importance in a minority of patients in whom the origin of troublesome symptoms is unclear (Fig. 1).

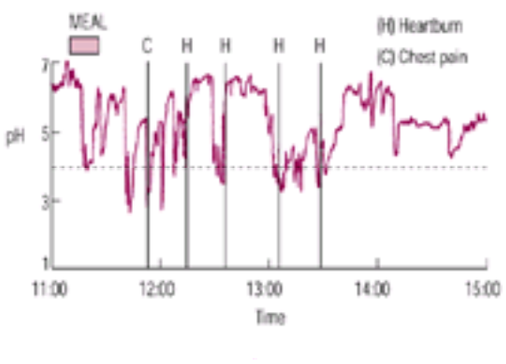

Fig. 1 Section of a 24-h oesophageal pH monitoring study showing the association of heartburn with episodes of acid reflux after a meal.

Radionuclide measurement of oesophageal transit

Computerized scintigraphic analysis of the movement of swallowed radiolabelled boluses can give quantitative information about the patterns of movement of material down the oesophagus. However, its poor spatial resolution makes it an inadequate method for the display of oesophageal anatomy. If structural abnormalities have been adequately excluded, slow or interrupted transit suggests abnormal motility, although patterns of transit are usually non-specific. If good barium radiology is available, radionuclide testing of oesophageal transit becomes redundant.

General management of oesophageal dysphagia

Symptomatic treatment of dysphagia is frequently necessary because of the limited options and efficacy of treatments for oesophageal disorders. Although these measures may appear obvious, this aspect of management is commonly neglected by both patient and physician.

Optimization of bolus consistency

Large particles of solid food may impact on strictures. Large boluses require greater propulsive force even in the absence of stricture, and may trigger oesophageal spasm. Boluses should therefore be small and, in some circumstances, reduced to semiliquid or liquid form. This can be achieved by the use of fluids in bolus preparation and avoidance of hard fibrous foods. Poor dentition should be treated. In some patients, defects of oesophageal function may be so severe that the diet should be pureed. Consultation with a dietitian will assist patients in identifying and preparing suitable food and in maintaining nutrition.

Assistance with oesophageal transit

Liquids assist transit by reducing the viscosity of food and providing a pressure head in the oesophagus. Gas generated within the oesophageal body from effervescent drinks can act as a piston which displaces oesophageal contents into the stomach in the erect position and may be sufficient to overcome an achalasic sphincter. The value of gravity in assisting transit should never be forgotten. Patients with severely impaired oesophageal transit should be advised to swallow medications in the upright position and with plenty of water so as to avoid injurious contact of the oesophagus with potentially corrosive tablets.

Alternative/supplementary approaches to feeding

Rarely, the above measures fail to maintain nutrition. Percutaneous endoscopic gastrostomy should then be used.

Gastro-oesophageal reflux disease

This is by far the most common oesophageal disorder. Reflux symptoms occurring at least once in 6 months are experienced in about one-third of people. Management should be tailored to the wide range of severity.

Definition

Gastro-oesophageal reflux occurs to some degree in everybody. It should only be considered a disease when it gives rise to significant symptoms or complications sufficient to impair the quality of life. The terms reflux or peptic oesophagitis should be reserved for circumstances when endoscopy shows that the oesophageal mucosa is clearly breached by the action of the refluxed gastric contents. Minor changes such as erythema, oedema, or friability have been shown to be very unreliable indicators of the presence of oesophagitis.

Aetiology

In most patients reflux disease arises from the excessive exposure of the distal oesophagus to refluxed acid, usually because of an abnormal frequency of reflux episodes. In a few patients, however, symptoms arise with relatively normal levels of acid exposure, presumably because of sensitization of the oesophageal mucosa.

In most patients reflux occurs as a result of defective neural control of the lower oesophageal sphincter. Severe reflux disease can also result from damage to the oesophagus as in scleroderma. Hiatus hernia is common in patients with reflux disease and causes displacement of the sphincter from the hiatus formed by the diaphragmatic crura. The hiatus provides important extrinsic support to the sphincter and helps to maintain gastro-oesophageal competence, particularly during straining. Hiatus hernia therefore impairs sphincter function and also acid clearance.

Most reflux occurs during the day, usually after food, but nocturnal reflux is also very important. Refluxed acid is cleared from the oesophagus by peristalsis and swallowed saliva. Slow clearance of oesophageal acidification contributes significantly to prolonged acid exposure in about 50 per cent of patients.

Consequences of excessive reflux

Symptoms

These are an important source of disability. Heartburn is most important and when it occurs on more than 2 days a week causes significant impairment of quality of life. However, presentation may be with the less specific pattern of dyspepsia, or with regurgitation, haematemesis, and dysphagia due to either stricture or motor dysfunction of the oesophageal body. Reflux may cause respiratory symptoms such as hoarseness, persistent cough, and asthma, which may predominate in some patients.

Oesophagitis

The chemical insult from excessive exposure of the mucosa to acid and pepsin leads to distal oesophageal erosion or ulceration in between 40 and 60 per cent of patients with troublesome reflux symptoms. The extent of ulceration varies greatly, from tiny patches of erosion to extensive circumferential ulceration in a small minority.

The risks of oesophagitis are not well defined. Peptic stricture and/or oesophageal columnar metaplasia (Barrett's oesophagus) are typically only associated with severe oesophagitis.

Stricturing causes dysphagia which may be debilitating and lead to malnutrition. Treatment with dilatation (bougienage) is a burden and is associated with a risk of perforation.

Columnar metaplasia (see below) is associated with the risk of developing oesophageal adenocarcinoma. It may also be associated with deep benign oesophageal ulceration within the columnar-lined segment. Occasionally such ulcers erode into mediastinal structures or the pleural space, sometimes with fatal consequences.

Bleeding from oesophagitis is relatively common, but is rarely life-threatening except when it occurs from a deep ulcer associated with columnar metaplasia..

Diagnosis and assessment of severity

History

The history is pivotal for diagnosis because of the extremely high prevalence of reflux-induced symptoms and the lack of a definitive, inexpensive diagnostic test for reflux disease. Fortunately, the specificity of patterns of symptoms of reflux disease is arguably the highest of any of the more common gastrointestinal diseases and

the majority of patients can be diagnosed confidently on the basis of their history. The strategic use of history and initial empirical therapy as a means to assess diagnosis and severity is summarized in Fig. 2, and discussed in detail in the section on treatment below.

Fig. 2 Principal decision paths for management of reflux disease.

Endoscopy

When investigation is needed, endoscopy is the first choice as it is the only test that can give sensitive recognition and grading of oesophagitis and reliable diagnosis of oesophageal columnar metaplasia. Endoscopy also allows for the effective identification of significant peptic strictures, other types of oesophagitis, and other upper gastrointestinal disorders such as peptic ulcer disease and oesophageal and gastric carcinoma. As discussed above, however, most patients with reflux disease do not have endoscopically visible mucosal damage, so a negative endoscopy does not exclude the diagnosis of reflux disease. The value of endoscopy as the initial investigation is greatly enhanced by the accurate diagnosis of endoscopic biopsy and, where indicated, cytology brushings.

Oesophageal function tests

The place of these is summarized in Fig. 2. Oesophageal manometry and ambulatory pH monitoring have a limited but important role in the diagnosis of reflux disease. Oesophageal pH monitoring is most useful in patients with troublesome symptoms but without endoscopic signs of oesophagitis in whom a trial of therapy has failed, and patients with atypical symptoms that cannot be clearly related to reflux. Patients with suspected reflux symptoms but with no endoscopic evidence of oesophagitis who are being considered for antireflux surgery should also undergo oesophageal pH monitoring.

Barium swallow and meal

This is an inappropriate primary diagnostic test; it is of no value for the detection of abnormal reflux, and is insensitive for the diagnosis of oesophagitis and cannot grade it. Other pathologies such as gastric ulcer and oesophageal stricture are demonstrated with reasonable sensitivity, but adequate evaluation of these findings requires endoscopic biopsy. In contrast, barium swallow has an important secondary role in the investigation of the mechanisms of troublesome dysphagia. Barium swallow is the best method for recognizing extrinsic oesophageal compression which may be producing symptoms that could be interpreted as being due to reflux, and in the assessment of anatomically complex hiatus hernia. The mere demonstration of hiatus hernia, however, does not necessarily indicate the presence of reflux disease.

Principles of management

New treatments have transformed management in recent years. The major aims of treatment are to provide adequate symptomatic relief and control of oesophagitis. Reduction of oesophagitis to minor patchy erosions is probably sufficient to prevent the complications of oesophagitis, although adequate symptomatic relief is usually achieved only when oesophagitis is completely healed. Management steps are easily confused, as endoscopy, assessment of symptoms, and treatment are used not only for diagnosis, but also assessment of severity and titration of therapy.

Cost-efficient, secure diagnosis

The steps needed for diagnosis are shown in Fig. 2. When the history is classical, the symptoms are only mild to moderate, and there are no alert symptoms, endoscopy is effectively redundant. Given the high proportion of patients who have no endoscopic abnormality, a negative endoscopy should not detract from the soundness of a diagnosis based on symptoms. When symptoms are indeterminate, or alert symptoms of dysphagia, haematemesis, or weight loss are present, endoscopy is the primary diagnostic approach. The place of endoscopy in patients with troublesome or severe but classical symptoms is becoming less relevant because of the advent of highly effective initial high-level medical therapy with proton pump inhibitors.

Assessment of severity

Morbidity arising from symptoms and the presence and severity of oesophagitis are the two most important measures of severity of reflux disease. However, they show only poor correlation. The response of symptoms to low- and medium-level therapy (see Table 1 and below) gives an indirect approximation of the severity of oesophagitis and helps to determine further action.

Tailoring and titration of therapy

Classification of therapies by level of efficacy regardless of their mechanism of action (Table 1) provides a framework for the tailoring and titration of long-term management. The lowest effective dose of any agent should be used in long-term therapy. Figure 2 and Table 2 summarize the logic of using endoscopic findings to choose an appropriate level of initial therapy. As the responsiveness of individual patients varies, therapy needs to be titrated upwards or downwards through levels of efficacy in order to find the lowest level that is effective, on the basis that this will minimize drug costs.

Most patients with reflux disease will have an initial trial of empirical therapy. There are two main options for this. Initial low- to medium-level therapy is the traditional model outlined in Fig. 2. This has the disadvantage of giving less crisp diagnostic information, and often gives only slow relief of symptoms, but will usually identify patients with severe oesophagitis since these will usually not respond adequately. Initial high-level medical therapy is now available. This is most likely to give crisp confirmation of the diagnosis and prompt relief of symptoms. For optimum cost-effectiveness, it should be followed by a step-down approach to long-term therapy as outlined in Fig. 2.

Options for treatment of oesophagitis and symptoms arising directly from reflux

Non-drug measures and antacids

The efficacy of these traditional approaches is often over-rated. The most useful measures are avoidance of large meals and provocative foods, drinks, and physical activities. The benefits to reflux disease of stopping smoking, weight loss, and elevation of the bedhead are uncertain. Antacids will not usually prevent symptoms, but may be effective in aborting episodes of heartburn. These low-cost measures are worth a trial in patients with mild intermittent symptoms (Table 1 and Table 2), and should be used as maintenance therapy if they prove effective, provided that their impact on lifestyle is acceptable to the patient. They are unlikely to succeed in people with troublesome reflux symptoms.

Acid suppression

Inhibition of secretion of gastric acid makes gastric juice less injurious but does not stop reflux. This has deservedly become the most widely used drug therapy because of its high efficacy and adjustability. The more severe the oesophagitis, the higher the level of acid suppression that is needed (Table 2). Proton pump inhibitors have a special role because of their effectiveness in reduction of food-stimulated acid secretion and their greater overall efficacy in control of acid secretion compared with histamine-2 receptor antagonists.

Long-term treatment with acid suppressants maintains patients free of symptoms and oesophagitis indefinitely but withdrawal is usually associated with prompt relapse. The maintenance dose appears to be the same as the lowest effective healing dose. There have been concerns about the safety of long-term acid suppression ever since the introduction of histamine-2 receptor antagonists. To date, follow-up of patients treated continuously for 10 or more years with acid suppression has shown no evidence of any effects of significance, but in the context of patients who may require treatment with these agents for decades, more extensive follow-up is still needed. Given these theoretical safety considerations and also drug cost, long-term treatment of reflux disease with these agents should use the lowest effective dose.

Motility stimulants

Only cisapride has been adequately researched. It has medium efficacy for both short- and long-term management. It appears unlikely that much is gained from an increase in dosage. The principal effect of cisapride appears to be on oesophageal acid clearance. Unfortunately, it has been recognized recently that cisapride has effects on cardiac conductance that may rarely lead to sudden death at peak serum levels encountered during therapy in some patients, especially when various drugs are coadministered. Because of the risk of cardiac effects, cisapride must be regarded as a second- or third-line therapy that needs to be used with special precautions.

Combination medical therapy

Use of cisapride and histamine-2 receptor antagonists in combination gives moderately improved results but is less effective than monotherapy with proton pump inhibitors. Given the safety concerns described above, and its relatively high cost, such combination therapy should be reserved for special cases.

Antireflux surgery

In skilled hands, antireflux surgery is a very effective long-term therapy. Negative factors are the dependence of the results on the expertise of the surgeon and the morbidity and small (approximately 0.5 per cent) mortality associated with the surgery itself. Laparoscopic antireflux surgery is a major advance, as it achieves good control of reflux with a major reduction in the morbidity inherent in the more traditional approach. More information is needed about long-term results.

Choice between therapies

Selection of a medical or surgical therapy should take account of the severity of disease and the risks of antireflux surgery specific to the patient. It should also take account of the patient's age, both from the point of view of operative risk and the time over which the patient will need treatment for reflux disease, the cost of effective medical therapy, and, naturally, the preferences of the patient. In the United States, good open antireflux surgery becomes cost-effective compared with medical therapy after about 10 years, although this assessment does not take into account the cost of mortality. The breakeven point is likely to be shorter with laparoscopic surgery and in countries where the costs of surgery are lower than in the United States. Cost comparisons also need to take into account the decreasing price of acid suppressing agents. The choice between medical therapies should be largely governed by the local cost of the alternatives that give the necessary level of treatment, as all of the first-line options are safe and well tolerated.

Management of complications of reflux disease

Peptic stricture

Dysphagia secondary to stricture (Fig. 3) needs to be distinguished from the more common dysphagia seen in patients with reflux disease which is due to defective triggering and control of oesophageal body peristalsis (see the section on non-specific oesophageal motor disorders below). Peptic stricture is managed by a combination of peroral dilatation and healing of oesophagitis by either medical or surgical means. Provided oesophagitis is healed, stricture is usually not an ongoing problem.

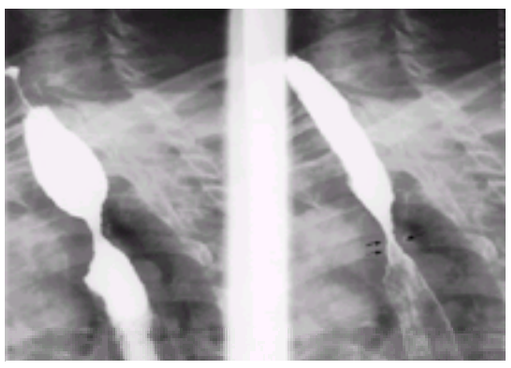

Fig. 3 Peptic stricture: asymmetrical circumferential narrowing with associated intramural diverticulosis suggests benign aetiology; the site suggests Barrett's mucosa, subsequently proven by endoscopy. (By courtesy of Dr H. Harley.)

Oesophageal columnar metaplasia (Barrett's oesophagus)

This increasingly recognized consequence of oesophagitis is dealt with in the section on oesophageal neoplasia. The association of columnar metaplasia with oesophageal adenocarcinoma contributes to the logic of vigorous treatment of severe oesophagitis.

Respiratory complications

Respiratory disease may occur as a result of either direct aspiration of refluxed gastric contents or from the reflex effects of gastro-oesophageal reflux. It is difficult to prove that reflux disease which coexists with respiratory disease is actually the cause of the respiratory problem. The best investigative approach is probably a trial of high-level acid inhibition with at least a double-dose of proton pump inhibitor for at least 2 months. Management of respiratory disease by antireflux surgery is a gamble that can only be supported primarily by clinical evaluation.

Regurgitation

Voluminous regurgitation is the main symptom in a small subgroup of patients with reflux disease. They may present complaining of vomiting, but a detailed history reveals that there is no prior nausea, and no effort involved in the appearance of the gastric content in the mouth. The determinants of high-volume reflux and regurgitation have not been defined. Treatment with proton pump inhibitors can have substantial benefits, but in more severe cases antireflux surgery is usually the only effective management.

Non-cardiac chest pain

Reflux is an important cause of non-cardiac chest pain (see below).

Primary oesophageal motor disorders

Idiopathic achalasia and achalasia-like states

Definition

These disorders are characterized by absent or incomplete relaxation of the lower oesophageal sphincter and impairment of peristalsis of the oesophageal body. Idiopathic achalasia, which was first described over 300 years ago, accounts for most cases and has an annual incidence of approximately 1 to 2 per 100 000. It affects all ages, but is diagnosed most often in early to mid adult life. The syndrome is also seen in Chagas disease and can sometimes accompany the intestinal pseudo-obstructive syndrome. Achalasia may be a manifestation of paraneoplastic neural dysfunction and may also be secondary to oesophageal amyloidosis.

Secondary or pseudoachalasia can arise from neoplastic infiltration of the gastro-oesophageal junction, and has been reported with carcinoma of the stomach, oesophagus, lung, pancreas, prostate, and with lymphoma.

Aetiology

Impairment of inhibitory neural control of the distal oesophagus is the universal abnormality. The syndrome can probably be produced by neural damage at several sites. The clearest evidence is degeneration of myenteric inhibitory neurones which, in the early stages, is associated with an inflammatory response.

Symptoms

Dysphagia with solids is almost universal. Regurgitation is also prominent. The regurgitated material tastes bland because it never enters the stomach. Cramping chest pain occurs in some patients during an early hypercontracting phase of the disorder. Weight loss is seen in patients with disabling dysphagia. The course of symptoms over time is variable. In some patients, symptoms remain static for many years but in others there is a progression with increasing problems with regurgitation over several years, as a result of development and increase of oesophageal dilatation. When this occurs, respiratory problems secondary to aspiration can become a major feature.

Diagnosis

Idiopathic achalasia is diagnosed on average 2 years after its first presentation, Delay is especially likely if oesophageal dilatation is absent. Dilatation varies in degree from a minor increase in oesophageal calibre to a grossly enlarged, colon-like oesophagus. Barium swallow shows oesophageal retention with a gastro-oesophageal junction that tapers smoothly to a closed sphincter, with occasional spurts of flow into the stomach (Fig. 4). In the absence of dilatation, a barium swallow is often reported as normal.

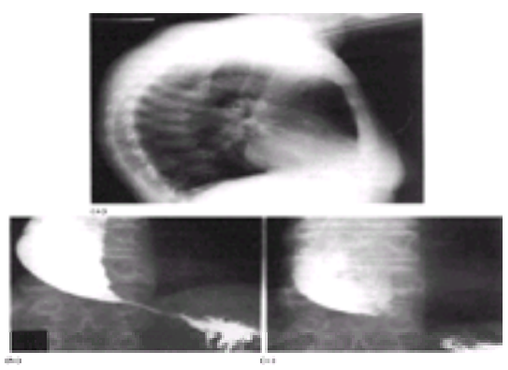

Fig. 4 Achalasia. (a) A lateral chest film shows the air–fluid level in the mid oesophagus just behind the air column of the trachea. (b) On a posteroanterior view this level is almost invisible. Barium-filled dilated oesophagus; intact mucosa in distal achalasia segment.

Oesophageal manometry is the only sensitive method for demonstration of the characteristic motor dysfunction. It is not unusual for manometry to be diagnostic of achalasia even though barium studies have been judged to be normal.

Idiopathic achalasia and achalasia-like states should be distinguished from constriction of the gastro-oesophageal junction by an infiltrating or encasing malignancy at the cardia. This diagnosis is often difficult to make. As a minimum, patients should be evaluated clinically for any symptoms or signs suggestive of malignancy, and upper gastrointestinal endoscopy should be done with mucosal biopsies. Computed tomography scanning is also of value.

Treatment

There are four potential approaches to treatment: drug therapy with agents that relax the lower oesophageal sphincter, mechanical disruption of the sphincter by either pneumatic dilatation or surgical myotomy, and pharmacological poisoning of the remaining excitatory nerves to the sphincter with botulinum toxin. The results of reduction of lower oesophageal sphincter pressure with drugs such as calcium antagonists and b-adrenergic agonists compare poorly with mechanical disruption of the sphincter, and by inference botulinum toxin.

Oesophagomyotomy, which is now being done increasingly as a laparoscopic or thoracoscopic procedure, is highly effective but is associated with a 5 to 10 per cent risk of troublesome gastro-oesophageal reflux. This risk can be minimized by the incorporation of an antireflux procedure.

Balloon dilatation is an attractive approach because of its simplicity and low cost, but it often needs to be repeated and in some hands fails in up to 40 per cent of patients, especially those who are young. It also carries a risk of perforation of about 5 per cent. With the development of minimally invasive surgery for oesophagomyotomy, balloon dilatation will probably be used most in older patients who have other medical problems that increase the risks of surgery.

Endoscopic injection of the sphincter with botulinum toxin is the most recent innovation. This toxin acts on residual excitatory nerves thereby lowering sphincter pressure. Short-term results are comparable to those of pneumatic dilatation but the procedure usually has to be repeated within 1 to 2 years. The toxin is also relatively expensive. It is a simple, low-risk procedure and most applicable to patients with significant coexisting morbidity which renders them unfit for dilatation or myotomy.

When oesophageal dilatation is present, prompt treatment is indicated to prevent its worsening, because of the morbidity and poor therapeutic outcome associated with gross oesophageal dilatation. Oesophageal emptying can be assisted by effervescent drinks (see above).

Prognosis

If effective treatment is applied before the development of major dilatation results are excellent, despite the persistence of major physiological abnormalities. Achalasia carries a very significantly increased risk for oesophageal carcinoma up to many years later which ranges from 2 to 7 per cent in authoritative reports. There is no apparent reduction of this risk with treatment. The average interval from diagnosis of achalasia to development of carcinoma has been estimated as 28

years. It is not usual practice to undertake surveillance for this risk, but some clinicians recommend periodic screening endoscopy.

Diffuse oesophageal spasm

Definition

Episodic chest pain and/or dysphagia resulting from spastic contractions of the distal half of the oesophageal body in the absence of any precipitating structural stenosis. There are no generally agreed criteria for diagnosis.

Aetiology

It is widely assumed that this is a dysfunction of neural control but there is a lack information that addresses this or other possibilities. What is known of the epidemiology is unhelpful with regard to aetiology. Stress is an unlikely primary precipitant but may exacerbate the problem. Good prevalence data are lacking. Diffuse spasm affects all ages and is much less common than achalasia.

Symptoms

Virtually all patients have episodic, crushing central retrosternal pain which can be excruciating; cardiac ischaemia is often the first diagnosis. Intermittent dysphagia occurs in about two-thirds of patients and leads to temporary abandonment of eating until symptoms abate—usually over about 30 min, but episodes of oesophageal obstruction can last for several hours. In most patients, symptomatic episodes occur less than once a month but in severe cases these may occur several times a week, or each time food intake is attempted.

Diagnosis

As with any intermittent fault, full-blown dysfunction is usually absent during investigation, but in a minority of patients there is asymptomatic motor dysfunction. Barium swallow then shows trapping of beads of contrast in the distal oesophagus—'the corkscrew oesophagus'—or sustained obliteration of the distal oesophageal lumen. Oesophageal manometry may show intermittent, simultaneous, prolonged, and vigorous oesophageal contractions interspersed with normal swallow-induced peristalsis. Relaxation of the lower oesophageal sphincter is normal. 24-h ambulatory manometry improves diagnostic accuracy by increasing the likelihood of capturing symptomatic episodes but is not necessary in all patients.

Most frequently the diagnosis is made on the basis of the history and the exclusion of other problems that may mimic diffuse oesophageal spasm. Most important amongst these is Schatski ring (see below). Achalasia is readily excluded by manometry. When appropriate, myocardial ischaemia should be excluded.

Treatment

There is no specific therapy. Smooth muscle relaxants such as nitrites, nitrates, and calcium antagonists may reduce symptoms but their use is often limited by side-effects. In many patients, reassurance is the most important management since the intensity and nature of symptoms gives rise to great concern. Opiate therapy is sometimes necessary. In the rare case of frequent, disabling spasm, long oesophagomyotomy can give good relief.

Prognosis

The major significance is impairment of quality of life and concern about life-threatening cardiac disease. There is no consistent progression over time. There are several reports of progression of diffuse oesophageal spasm to achalasia but in most of these it seems likely that early, spastic achalasia was initially misdiagnosed as diffuse oesophageal spasm.

Hypertensive peristalsis or nutcracker oesophagus

Definition

This is defined purely by the manometric criterion of primary peristaltic pressure waves in the oesophageal body that have peaks in excess of 250 mmHg (Fig. 5). There is preservation of the normal peristaltic pattern of a broad progression of the time of onset of the contraction wave in the oesophageal body.

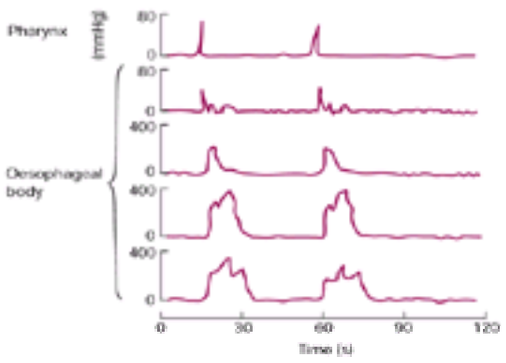

Fig. 5 Oesophageal manometric tracing made from points along the oesophagus in a patient with hypertensive peristalsis. The time of onset of the pressure waves shows a normal peristaltic gradient, but the contractions in the distal half of the oesophagus are very vigorous so that the peak pressures are abnormally high. The high-amplitude pressure waves are also somewhat prolonged and multipeaked.

Aetiology

This is not understood. It is not clear if it is a true motor disorder, and it may just represent the upper end of a continuum of peristaltic wave amplitudes. It has been shown to vary over time within individuals. There are indications that psychological factors can influence peristaltic amplitude. A minority of patients with hypertensive peristalsis also experience episodes of diffuse oesophageal spasm suggesting that the underlying dysfunction may be related to diffuse oesophageal spasm and is likely to involve neural control mechanisms.

Symptoms

The only clinical significance of hypertensive peristalsis is its relationship to non-cardiac chest pain. Hypertensive peristalsis alone does not produce dysphagia or derangement of oesophageal transit, because, by definition, peristalsis is preserved.

Treatment and prognosis

These are discussed in the section on non-cardiac chest pain.

Non-specific oesophageal motor disorders

Definition

This is a ragbag of manometrically defined oesophageal motor abnormalities which occur in isolation from other more clearly defined syndromes of oesophageal dysfunction, or in association with diseases such as gastro-oesophageal reflux disease, diffuse oesophageal spasm, and diabetic and other autonomic neuropathies. The pragmatic definition of these dysfunctions is that they are departures from normal patterns of oesophageal motor function which do not actually define specific diseases, but which are of clinical significance. Non-specific oesophageal motor disorder is the commonest single functional diagnosis made in most oesophageal manometric laboratories.

Aetiology

This is unknown, but it should not be assumed that only one mechanism is involved. Intermittent occurrence of dysfunctions suggests that they are due to defective neural control.

Symptoms

Multipeaked, swallow-induced distal oesophageal body contraction waves stand out from the other patterns not only functionally but also symptomatically. This pattern is loosely associated with the hypercontraction disorders of diffuse oesophageal spasm and hypertensive peristalsis, but sometimes does not appear to have any clinical significance.

Hypocontraction dysfunctions, recently termed 'ineffective' peristalsis, are associated with defective triggering and progression of both primary and secondary peristalsis. Failure to develop a propagated pressure wave of sufficient strength to maintain closure of the oesophageal lumen leads to deranged oesophageal transit. This probably explains the association of these disorders with mild intermittent dysphagia which occurs characteristically with solids. The non-obstructive dysphagia and slow oesophageal acid clearance seen in gastro-oesophageal reflux disease are due to such dysfunction. Secondary oesophageal body peristalsis has not yet been widely evaluated, but defects of this are probably an important cause of intermittent dysphagia, since, at least in patients with non-obstructive dysphagia and reflux disease, dysfunction of secondary peristalsis is substantially more common than primary peristaltic dysfunction. Oesophageal manometry with an adequate number of recording points in the oesophageal body is the only sensitive means for diagnosis.

Treatment

In most cases patients with symptoms found to be due to non-specific oesophageal motor disorders are in search of an explanation of the origin of their symptoms and reassurance rather than relief of symptoms. Cisapride may improve triggering and amplitude of peristaltic contractions and so theoretically transit. Usually, however, the symptoms arising from primary peristaltic dysfunction are not sufficiently severe to warrant continuous therapy. Secondary peristaltic dysfunction may be more troublesome, but there is no good information on the effect of prokinetic or other drugs on this.

Prognosis

These dysfunctions do not remit spontaneously. Patients are often helped by the measures outlined in the section on general management of oesophageal dysphagia, which minimize the demands on oesophageal transport mechanisms and provide propulsive forces that substitute for oesophageal contractions.

Non-cardiac chest pain

Definition

Implicit in this rather circuitous and negative label is the view that this pain has a cardiac-like quality, but there is no evidence for a cardiac origin. The oesophagus is the next most likely origin, but it is unlikely that all such pain arises from the oesophagus.

Aetiology

Long-term monitoring of oesophageal pH and motility has given us mixed messages. Evidence for triggering of pain by reflux or oesophageal motor dysfunction has been found in between one-fifth and one-half of patients evaluated. Oesophageal mucosal pain due to gastro-oesophageal reflux is the most common and rewarding positive diagnosis. Frank oesophageal spasm associated with achalasia and diffuse oesophageal spasm is an unusual but convincing cause of non-cardiac chest pain. In the majority of patients, most episodes of pain occur independently of reflux and any motor abnormality, although many of these patients have non-specific oesophageal motor disorders or hypertensive peristalsis (see above). Recently, sustained contraction of the longitudinal muscle has been identified by prolonged intraluminal ultrasonography in association with a high proportion of episodes of pain. Nevertheless, in many patients non-cardiac chest pain appears to be a primary oesophageal pain disorder and any motor disorder may be an epiphenomenon.

Symptoms

By definition, the pain resembles cardiac pain in its sensation and distribution. It can be very intense and distressing, can disturb sleep, and may be worse during periods of emotional stress. Postprandial occurrence, in association with heartburn, suggests that it may be caused by reflux. When pain is associated with dysphagia, vigorous achalasia or oesophageal spasm are very possible.

Diagnosis

Investigation is demanding and relatively unrewarding. Firstly, myocardial ischaemia should be assumed to be the cause until proven otherwise. Endoscopy should then follow. In patients who are having recurrent problems, monitoring of oesophageal pH and oesophageal motility studies are both indicated.

Treatment

If the pain is triggered by gastro-oesophageal reflux, high-level therapy should be tried (see section on gastro-oesophageal reflux disease). Achalasia and diffuse oesophageal spasm should be treated on their own merits. Half or more of patients will still have no clearcut diagnosis. In these, treatment with anxiolytics and antidepressants has been found to be moderately effective. Agents that reduce the strength of oesophageal contraction, such as calcium antagonists, appear ineffective in hypertensive peristalsis.

Prognosis

When the pain is clearly due to reflux disease, diffuse oesophageal spasm, and achalasia, the prognosis is as for these conditions. In patients in whom there is no such clear relationship, continuing anxiety about the origin of the symptoms and the fear that this might be cardiac tends to persist, with repeated admissions to hospital because of attacks of pain.

Oesophageal motor disorders secondary to systemic disease

Oesophageal motility may be affected by a number of systemic diseases (Table 3). These diseases may affect the striated or smooth muscle itself or the neural control mechanisms.

The division of the oesophageal musculature into striated and smooth muscle components is revealed clearly by the myopathic diseases that affect the oesophageal musculature. In patients with peripheral myopathy this would normally have been diagnosed. Weak or absent oesophageal contraction in the affected segment has the expected adverse impact on oesophageal transit, with a pattern of symptoms similar to the hypocontraction states of non-specific oesophageal motor disorders

(see above). The management of these dysfunctions is along general lines (see section on general management of oesophageal dysphagia).

Diseases of oesophageal smooth muscle

Scleroderma

Definition and aetiology

This eventually involves the smooth muscle of the oesophagus in at least three-quarters of patients. It may be part of the CREST (**C**alcinosis, **R**aynaud's syndrome. (o)**E**sophageal dysphagia, **S**clerodactyly, **T**elangiectasia) syndrome The time of onset of symptoms from oesophageal involvement is very variable in relation to other manifestations but is sometimes the presenting symptom. Muscle atrophy and fibrosis are the cardinal features, but neuropathic abnormalities may also contribute to dysfunction. Smooth muscle peristalsis is feeble or totally absent and the tone of the lower oesophageal sphincter subnormal or absent.

Symptoms

Troublesome reflux symptoms are the most common consequence of loss of function. The pattern of dysphagia resembles that seen in non-specific oesophageal motor disorder (see above). If dysphagia is severe, peptic stricture should be excluded, as complete loss of oesophageal smooth muscle peristalsis rarely leads to disabling dysphagia.

Treatment

Reflux disease is frequently severe and should be managed by high-level medical therapy in order to prevent complications such as stricture (see above). Antireflux surgery is relatively contraindicated because of the poor propulsive function of the oesophageal body.

Other disorders

A scleroderma-like picture of oesophageal dysfunction is sometimes seen in other connective tissue disorders such as mixed connective tissue disease. The smooth muscle segment is also involved in systemic myopathies including polymyositis–dermatomyositis and myotonic dystrophy.

Abnormal oesophageal motility is common in diabetes mellitus, and may be a feature of amyloidosis, chronic alcoholism, and the pseudo-obstructive syndrome. In these disorders, the disturbance is believed to be primarily due to dysfunction of neural control mechanisms.

Disorders of striated muscle

Involvement of the striated muscle segment of the oesophagus is rare and usually present with high dysphagia, often in association with oropharyngeal dysfunction (see Chapter 14.5). The inflammatory myopathies (dermatomyositis, polymyositis, and inclusion body myositis), the muscular dystrophies (myotonia dystrophica and oculopharyngeal dystrophy), and myasthenia gravis are the most common causes.

Abnormalities of oesophageal anatomy

Non-neoplastic abnormalities which distort oesophageal anatomy may interfere with normal function or may merely pose difficulties in the interpretation of findings.

Sliding hiatus hernia

Definition

Around 90 per cent of hiatus hernias are of this type in which the gastro-oesophageal junction is displaced upwards into the thorax, giving a simple shaped pouch of intrathoracic stomach.

Aetiology

The phreno-oesophageal ligament is effaced in sliding hiatus hernia, but it is not clear whether this is a primary defect of gastric anchorage.

Symptoms

Many patients with hiatus hernia are asymptomatic. Despite this, physiological studies indicate that herniation of the gastro-oesophageal junction impairs its function as an antireflux barrier by removing the normal diaphragmatic crural compression from the lower oesophageal sphincter. Thus, hiatus hernia can be taken as a risk factor for reflux disease, but not an abnormality that makes the diagnosis.

Treatment

Symptoms of gastro-oesophageal reflux are the only ones of major significance. These should be treated along conventional lines (see gastro-oesophageal reflux disease).

Prognosis

This is essentially that of any associated reflux disease.

Rolling or para-oesophageal hiatus hernia

Definition

A variable part of the stomach herniates through the hiatus alongside a normally situated gastro-oesophageal junction. This pattern of herniation may produce a gross disturbance of gastric anatomy, usually with a narrow exit from the herniated pouch into the main stomach cavity. Some rolling hernias are also associated with displacement of the gastro-oesophageal junction above the hiatus in which case these are known as mixed hernias.

Symptoms

Obstruction and distension of the pouch causes upper abdominal discomfort and can progress to strangulation. Gastric volvulus can occur because of the laxity of the gastric anchorage and may obstruct the gastro-oesophageal junction. Both of these problems have a very high mortality and demand urgent surgery. Elective surgery is normally recommended to reduce and anchor rolling hiatus hernias in order to remove these risks.

Prognosis

Unfortunately, there are no adequate data on the degree of risk associated with rolling hiatus hernia or on any anatomical factors that are especially hazardous.

Schatzki ring (B ring)

Definition

This is a characteristic short luminal stenosis which occurs at the gastro-oesophageal junction (Fig. 6). It is made up only of mucosa and submucosa, and may narrow the lumen to a few millimetres or cause a clinically insignificant minor indentation.

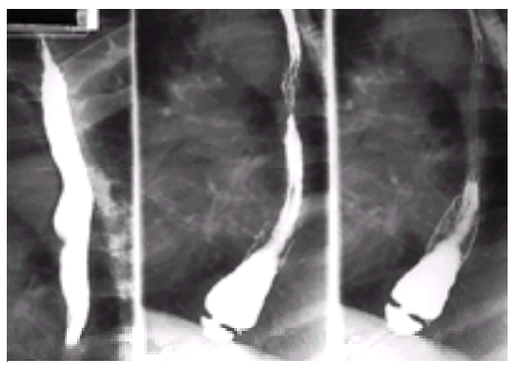

Fig. 6 Schatzki ring: a thin (2 to 4 mm in height), annular constriction at the gastro-oesophageal junction, best shown on prone-oblique views.

Aetiology

This is unknown. There is no firm evidence that reflux oesophagitis is the cause although this is sometimes assumed to be the case.

Symptoms

With mechanically significant rings, intermittent dysphagia occurs on eating solids. Meat is often the culprit, leading to the common term of 'steakhouse syndrome'. Episodes of bolus obstruction are not unusual, with associated chest pain caused by powerful oesophageal contractions. Failure to recognize a Schatzki ring frequently leads to the incorrect diagnosis of primary diffuse oesophageal spasm. Sensitive diagnosis is only achieved by an expert radiologist who has been asked to look for this abnormality. Adequate distal oesophageal distension during the barium swallow is essential for detection and this is best achieved by prone-oblique views. A less well-tailored barium examination and endoscopy frequently fail to show a symptomatic ring.

Treatment

Disruption of the ring by simple peroral dilatation or endoscopic diathermy or laser is very rewarding, as the dysphagia and chest pain are cured, sometimes after many years of symptoms. However, there is a significant incidence of recurrence and repeated dilatations at intervals are often needed.

Other rings and webs

Other short oesophageal stenoses may develop because of peptic stricture, muscular rings, and cervical webs with (Plummer–Vinson syndrome) or without iron-deficiency anaemia.

Oesophageal diverticula and pseudodiverticula

Wide-mouthed multiple diverticula are characteristic of scleroderma oesophagus. In the non-sclerodermatous oesophagus, diverticula occur in the mid and distal oesophagus, both types probably being 'blow-outs' secondary to hypercontraction motor disorders. These can become very large. It is rare for them to cause symptoms, but they may be associated with dysphagia and regurgitation of retained contents. Unless symptoms are disabling, they are best left undisturbed because leakage is common following surgical removal.

Multiple intramural outpouchings of barium are characteristic of intramural pseudodiverticulosis which appears to be due to dilatation of the ducts of submucosal glands by an unknown process.

Extrinsic oesophageal compression

This is a relatively common cause of dysphagia, and is most often a result of malignant mediastinal lymphadenopathy. Barium swallow or endoscopy usually show a relatively long constriction of the oesophageal lumen of variable calibre, associated with a normal mucosal appearance. Dilatation of such a compression is usually unrewarding because of its elastic recoil.

Mechanically significant extrinsic compression may also result from an enlarged heart, a dilated or unfolded aorta, or an aortic aneurysm. Kyphosis may accentuate the mechanical impact of these abnormalities. Congenital vascular abnormalities can also compress the oesophagus in adults, an aberrant right subclavian artery being by far the most common.

Mechanical, chemical, and radiation trauma

Mallory–Weiss tear

These mucosal tears extend across the gastro-oesophageal junction and are normally induced by vigorous straining associated with vomiting. Bleeding is the only consequence of significance. In 10 per cent of cases bleeding is severe enough to cause hypovolaemia. The history is usually quite characteristic, but definitive diagnosis requires endoscopy. Continued bleeding usually responds to endoscopic injection or electrocoagulation, vascular embolization, or vasopressin infusion. Very rarely, surgery is needed to underrun a persistently bleeding artery at the base of the tear.

Barogenic oesophageal rupture (Boerhaave's syndrome)

In this uncommon condition, straining and vomiting cause oesophageal rupture, most often in the left lower third of the oesophagus. High-volume spillage of the gastric contents into the pleural space causes shock and pain in the chest and upper abdomen with radiation to the back, left chest, or shoulder. The chest radiograph becomes abnormal only some hours after rupture. Surgical repair and drainage are usually necessary, and if this is delayed beyond 24 h, the mortality is very high. Unfortunately, diagnostic delay is not unusual.

Iatrogenic oesophageal perforation

Physicians encounter this problem most often as a result of their involvement in dilatation of oesophageal strictures, pneumatic bag dilatation for achalasia, or through problems with the management of oesophageal varices by balloon tamponade. Even with meticulous technique and appropriate equipment, oesophageal perforation can occur. Perforation is strongly suggested by development of chest or epigastric pain directly after instrumentation, sometimes with dyspnoea. Pneumothorax and surgical emphysema are diagnostic. Any suspicion of perforation should be acted upon by taking a chest radiograph which should be repeated in several hours if it is negative. Broad-spectrum antibiotics should be given on suspicion, as they are most effective in minimizing the risks of mediastinitis when given from the outset. Surgical consultation should occur promptly; the choice between conservative and surgical management needs to be individualized. Increasingly, instrumental perforation is being managed non-surgically with nasogastric suction, antibiotics, and intravenous nutrition with good results, primarily because instrumental injury

usually occurs when the stomach is empty.

Caustic ingestion

Definition and aetiology

Strong acids and alkalis are both very damaging to the oesophagus and are found in high concentrations in many agents commonly used in the household for cleaning and maintenance. Laryngeal and gastric injuries may overshadow oesophageal injury. Because of their relative lack of taste, alkaline solutions are more likely to be swallowed accidentally in large amounts. Alkaline injury is especially deep; acid tends to form a superficial coagulant, which limits penetration.

Symptoms

The severity and extent of injury are immensely variable and cannot be predicted accurately from estimates of the volume ingested. Around half of patients with a history of caustic ingestion have no significant injury. Oropharyngeal and laryngeal injury confirm caustic ingestion and can be a major threat to the airway, but do not predict the existence and severity of oesophageal injury which causes odynophagia, dysphagia, or haematemesis. Prompt fibreoptic panendoscopy appears to be safe. This may be normal or show only patchy mucosal oedema, erythema, and small haemorrhagic ulcers, indicative of superficial damage with a good prognosis. Extensive and circumferential ulceration, and grey or brown/black ulceration suggest transmural injury.

Treatment

Patients with severe injury must be observed closely for signs of perforation. Nasogastric suction should be used with the administration of broad-spectrum antibiotics as these appear to reduce the severity of infective complications. The use of steroids is controversial, the balance of evidence tending to oppose their use. Oesophageal stricture is to be expected with severe injury and appears not to be prevented by routine dilatation in the first 2 weeks after injury. A barium study should be done at 2 to 3 weeks to screen for stricturing, and then subsequently at about 3-monthly intervals thereafter for a year, so that the development of stricturing is recognized at a stage when dilatation may have some impact.

Prognosis

The main short- to medium-term risk is the development of stricture. Caustic strictures are difficult and hazardous to treat by peroral dilatation so that about half of patients require oesophageal resection. In the long term (average onset 40 years after injury) carcinoma of the oesophagus is a major hazard, the risk being 1000 to 3000 times the expected risk.

Medication-induced oesophagitis

Definition and aetiology

This entity was only recognized in 1970. The chemical properties of medications pose hazards to the oesophageal mucosa because of the relative susceptibility of this to injury through pH-dependent and other mechanisms. This susceptibility arises in part from the high local concentrations of medications that occur in the oesophageal lumen when a tablet gets 'hung up'. Pills move surprisingly slowly through the normal oesophagus. Defective oesophageal transport, poor pill design, increased mucosal susceptibility to injury, and poor pill-taking technique contribute to the problem. Medications known to have an especially high risk for oesophageal damage are listed in Table 4.

Symptoms

Symptoms are those for any form of oesophagitis with stricturing which can be very difficult to manage. Probably, much pill-induced injury goes unrecognized. Pill-induced injury is by far the most likely cause of oesophagitis and/or benign stricture at the level of the aortic arch, where pills can lodge for prolonged periods. Injury at the distal oesophagus, the other common site of hold-up, is probably usually misdiagnosed as being due to reflux disease.

Treatment and prognosis

Medications and formulations with a high risk of injury should be identified and avoided if possible, especially in elderly patients with reflux disease or abnormal oesophageal transit. Pill transit is facilitated if pills are taken in the erect position with plenty of water. Pharmaceutical companies need to pay more attention to the use of shapes, sizes, and coatings that can assist transit of pills through the oesophagus. Stricturing may require surgery.

Chemotherapy-induced oesophageal problems

Chemotherapy causes oesophageal problems in several ways. Therapy may impair mucosal defences by affecting cell turnover leading to 'mucositis'. This in turn may reduce resistance of the mucosa to damage from other agents, and increase susceptibility to infective oesophagitis from immune suppression. Oesophageal transit and acid clearance may be impaired through the neurotoxic effects of some agents. Fistulation or perforation may occur through cytotoxic effects on a malignancy in the oesophageal wall. The striking recent observation that combination chemotherapy is associated with the development of oesophageal columnar metaplasia in women being treated for breast cancer demands further investigation.

Oesophageal neoplasms

Cure is only possible in the small minority of patients whose disease presents early. Several approaches are possible for palliation and data are somewhat conflicting about the relative merits of each. Some approaches require considerable technical skill.

Squamous cell carcinoma

Definition

This is simply defined as a squamous carcinoma arising from the squamous oesophageal mucosa. It is by far the most common oesophageal neoplasm. In some parts of the world it is the most common of all cancers, but in the Western world it accounts for approximately 4 per cent of cancer deaths and has an annual incidence in the United States of 5 per 100 000 in Whites and 17 per 100 000 in Blacks.

Aetiology

The striking geographical variation in incidence suggests a major contribution from environmental factors. There are multiple proven or putative risk factors which include heavy alcohol use and intake of carcinogens from smoking, from soil and water, and from high rates of consumption of nitrosamines and aflatoxins. Other factors implicated are vitamin A deficiency, chronic candida infection, injury to the oesophageal mucosa due to ingestion of a corrosive substance years previously, and chronic irritation from oesophageal retention in achalasia. Some hereditary conditions such as tylosis predispose to squamous carcinoma. Invasive carcinoma is preceded by mucosal dysplasia and carcinoma *in situ*.

Symptoms

Dysplasia and carcinoma *in situ* are asymptomatic, and are only recognized by screening programmes set up in very high-risk areas, usually using blind cytological sampling methods. Inexorable progression of dysphagia over a few weeks is the almost universal presentation. Dysphagia usually only occurs when the tumour has become circumferential. Rarely, malignant mucosal ulceration presents with pain of the oesophageal mucosa due to malignant oesophageal ulceration. Substantial

weight loss has often occurred by the time of presentation.

Barium swallow typically reveals a stricture with an irregular, lobulated mucosal outline (Fig. 7), but occasionally the appearance mimics benign peptic stricture. The diagnosis is best proven by fibreoptic endoscopy, with mucosal biopsy and brush cytology. Occasionally, an asymptomatic oesophageal carcinoma is diagnosed when endoscopy is done for some other reason. Early lesions are often unimpressive, so that any minor mucosal irregularity should be sampled thoroughly by biopsy and cytology.

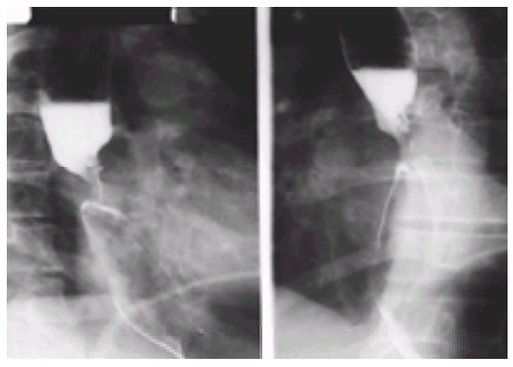

Fig. 7 Squamous carcinoma: a circumferential luminal narrowing which is asymmetrical and irregularly shouldered; the mucosal aspect is ulcerated.

Treatment

In very early, usually asymptomatic, carcinoma, resective surgery is the treatment of choice as it achieves high rates of cure. Curative resective surgery should only be attempted after careful staging of the tumour by clinical examination, chest radiograph, thoracic and abdominal computed tomography scanning, bronchoscopy, and liver function tests. Endoscopic oesophageal ultrasound is a much needed advance in the sensitivity of definition of the local extent of the tumour, but is not yet widely available. Palliation poses many challenges. There is a sorry lack of critical comparison of options. Resective surgery is unattractive, especially in elderly patients, because of its morbidity and mortality. Radiotherapy, with or without chemotherapy, is usually the best option for management of malignant obstruction. Repeated peroral dilatation, peroral placement of stenting tubes, laser photocoagulation, and injection of sclerosants are all options for the management of recurrent malignant strictures which have potential for improving the quality of life. Oesophagopulmonary fistula is a distressing development which usually causes pneumonia and persistent cough and which can sometimes be controlled by stenting.

Prognosis

This remains dismal except for regions where screening programmes identify early, asymptomatic cases. Only about one-quarter of patients are deemed to be potentially curable by surgery, and of these, about one-quarter will be alive and free of disease after 5 years. Thus, the overall 5-year survival rate is approximately 6 per cent. Such figures must be interpreted cautiously, because of differences between studies in the scope of presurgical staging, definitions of resectability, and criteria for exclusion of patients from consideration for surgery on the grounds of debility, old age, and other medical problems.

Adenocarcinoma and oesophageal columnar metaplasia (Barrett's oesophagus)

Definition

Between 80 and 90 per cent of adenocarcinomas arising in the oesophagus occur in association with oesophageal columnar metaplasia, or Barrett's oesophagus. In the minority of adenocarcinomas occurring in a squamous-lined oesophagus, the oesophageal mucus glands appear to be the source of malignant change.

Aetiology

Oesophageal columnar metaplasia (Barrett's oesophagus) develops as a result of the healing of severe reflux oesophagitis with metaplastic epithelium. This occurs from the gastro-oesophageal junction upwards over a distance that varies from 2 or 3 cm to the full length of the oesophagus. Oesophageal columnar metaplasia carries a 40-fold risk for development of oesophageal adenocarcinoma. Surveillance programmes in patients with oesophageal columnar metaplasia have shown a rate of development of adenocarcinoma that varies from 1 in 50 to 1 in 175 patient years. Occurrence of adenocarcinoma is very strongly associated with prior development of high-grade dysplasia in the metaplastic segment. The reasons for an apparently real increase in oesophageal adenocarcinoma are unknown. An increase in the prevalence of reflux oesophagitis, related to the reduced prevalence of *Helicobacter pylori* infection and a consequent increase in gastric acid secretion, is a plausible explanation.

Symptoms

The presentation of adenocarcinoma resembles that of squamous carcinoma (see above). Adenocarcinoma tends to be more fleshy and intraluminal but still presents at a very late stage. Metastatic disease is more common on presentation of adenocarcinoma than with squamous carcinoma.

Initial diagnosis and staging are along the same lines as for oesophageal squamous carcinoma.

Treatment of established adenocarcinoma

Careful staging is the cornerstone of appropriate management. Because of the usually distal site of occurrence of oesophageal adenocarcinoma, resection with oesophagogastrostomy is often best. Adenocarcinoma appears to respond less frequently to radiotherapy and chemotherapy than squamous carcinoma.

Management of the risk for adenocarcinoma in oesophageal columnar metaplasia

This is a very active field of research. Development of high-grade dysplasia in the columnar metaplastic segment precedes development of adenocarcinoma. This dysplasia can be recognized with sensitivity if at least four radially spaced biopsies are taken at every 2 cm of columnar-lined oesophagus. It is controversial whether such expensive surveillance methods are justified by the relatively low rate of recognition of early adenocarcinoma. There is also controversy about how a diagnosis of high-grade dysplasia should be acted upon. Some authorities recommend oesophageal resection on confirmation of this diagnosis, whilst others favour close surveillance with endoscopic ultrasound and repeated biopsies, with oesophageal resection being reserved for when there is clear evidence of disruption of the structure of the oesophageal wall, indicative of early invasive carcinoma. Others advocate perendoscopic ablation by laser or argon beam coagulation, or mucosal resection in the case of true intramucosal carcinoma. In large part, the approach to management of high-grade dysplasia is substantially moulded by the morbidity and risks of resective surgery and the availability of endoscopic ultrasound. In younger, fit patients, the balance is more strongly in favour of early resection than it is in older, less fit patients. Lack of detailed knowledge about the natural history of high-grade dysplasia makes the decision-making process especially difficult.

Failure to discuss the risk for adenocarcinoma and the option of endoscopic surveillance with a patient who has oesophageal columnar metaplasia could well be viewed as an indefensible lapse of practice, despite the uncertainties about cost-effectiveness.

Prognosis

For established symptomatic adenocarcinoma, prognosis is every bit as dismal as for squamous carcinoma. Post-mortem studies have shown that only a small proportion of patients with oesophageal columnar metaplasia are diagnosed as having this condition during life. Consequently, the impact of endoscopic surveillance

can at best only be limited. For those patients in whom screening is undertaken, it has been clearly established that high-quality screening leads to diagnosis of oesophageal adenocarcinoma at a stage when it may be cured by resection.

Other oesophageal tumours

Primary malignant tumours

Primary malignant tumours other than squamous carcinoma and adenocarcinoma are rare and all have a poor prognosis. These include malignant melanoma, lymphoma, carcinoid, leiomyosarcoma, neuroendocrine carcinoma (small cell carcinoma), adenoid cystic carcinoma, and pseudosarcoma. These tumours show a mixture of polypoid and infiltrating features and are usually only clearly distinguished from the more common malignancies by histology.

Benign oesophageal tumours

Leiomyoma is a relatively common oesophageal tumour which rarely causes symptoms. It is usually intramural but can become pedunculated. Around two-thirds of benign oesophageal tumours are leiomyomas. They usually only cause symptoms if they are very large, or on a long pedicle. Other benign intramural tumours of the oesophagus include lipomas and granular cell tumours. The main risk of these is that they are mistaken for malignant tumours and operated on inappropriately.

Squamous cell papillomas of the mucosa can mimic a polypoid squamous carcinoma and so should be removed endoscopically for histological diagnosis.

Infective oesophagitis and other non-neoplastic mucosal diseases

Most of these cause symptoms because of mucosal hypersensitivity. When the course is prolonged, interference with food intake may become a dominant problem in patient management. Viral oesophagitis can sometimes cause major haemorrhage. Some disorders damage the full thickness of the oesophageal wall and so lead to stricturing. Infective oesophagitis is by far the most important of these disorders, and has become more prevalent with the increased number of people who are immunosuppressed through HIV infection or chemotherapy.

Diagnosis is often aided by the setting in which the oesophageal problem occurs. Cutaneous or oral disease can suggest what is happening in the oesophagus, but barium swallow adds relatively little to the assessment of mucosal hypersensitivity. Endoscopy is the diagnostic method of choice. Mucosal appearance and the distribution of oesophageal lesions can be virtually diagnostic. In addition, biopsies and brushings allow for histological diagnosis and identification of infectious agents. Endoscopy has most to offer in patients with chronic symptoms, or those who are immunosuppressed.

Infective oesophagitis

The more important causes of infective oesophagitis are summarized in Table 5. Immune status is a major determinant of the pattern of infection. Though infective oesophagitis may be severe in immunocompetent patients it is characteristically self-limited and topical therapy is normally all that is needed (Table 5).

Immunocompromised patients usually need aggressive, systemic therapy, otherwise the infection does not resolve. The infection can be difficult to eradicate, tends to recur, and can cause major disability. Two or more infections are not unusual (Table 5).

Helicobacter pylori does not appear to be of any primary significance in the pathogenesis of oesophageal mucosal disease.

Other non-neoplastic mucosal diseases

Skin and systemic diseases associated with lesions of the oropharynx may also involve the oesophagus. These include epidermolysis bullosa, Behçet's disease, lichen planus, pemphigus vulgaris, bullous pemphigoid, benign mucous membrane (cicatrial) pemphigoid, and drug-induced disease (Steven's Johnson syndrome and toxic epidermal necrolysis).

Chronic, and less frequently acute, graft versus host disease may cause severe oesophageal problems through mucosal desquamation or mural damage. Resultant stricturing shows considerable variation in appearance.

Rarely, Crohn's disease can cause indolent, craggy ulceration and/or stricturing. Oesophageal sarcoidosis can mimic Crohn's disease.

Further reading

Anand BS *et al*. (1998). A randomized comparison of dilatation alone versus dilatation plus laser in patients receiving chemotherapy and external beam radiation for esophageal carcinoma. *Digestive Diseases and Sciences* **43**, 2255–60.

Balaban DH *et al*. (1999). Sustained esophageal contraction: a marker of esophageal chest pain identified by intraluminal ultrasonography. *Gastroenterology* **116**, 29–37.

Cook IJ, Kahrilas PJ (1999). AGA technical review on management of oropharyngeal dysphagia. *Gastroenterology* **116**, 455–78.

Dent J *et al*. (1999). An evidenced-based appraisal of reflux disease management—the Genval Workshop Report. *Gut* **44** (Suppl. 2), S1–S16.

Dent J, Holloway RH (1996). Esophageal motility and reflux testing. State-of-the-art and clinical role in the twenty-first century. *Gastroenterology Clinics of North America* **25**, 51–73.

DeVault KR (1996). Lower esophageal (Schatzki's) ring: pathogenesis, diagnosis and therapy. *Digestive Diseases* **14**, 323–9.

Ellis FH Jr. (1998). Long esophagomyotomy for diffuse esophageal spasm and related disorders: an historical overview. *Diseases of the Esophagus* **11**, 210–14.

Falk GW (1999). Endoscopic surveillance of Barrett's esophagus: risk stratification and cancer risk. *Gastrointestinal Endoscopy* **49**, S29–S34.

Fennerty MB (1999). Perspectives on endoscopic eradication of Barrett's esophagus: who are appropriate candidates and what is the best method? *Gastrointestinal Endoscopy* **49**, S24–S28.

Kahrilas PJ (1997). Anatomy and physiology of the gastroesophageal junction. *Gastroenterology Clinics of North America* **26**, 467–86.

Lagergren J *et al*. (1999). Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma. *New England Journal of Medicine* **340**, 825–31.

Lundell LR *et al*. (1999). Endoscopic assessment of oesophagitis: clinical and functional correlates and further validation of the Los Angeles classification. *Gut* **45**, 172–80.

McCord GS, Staino A, Clouse RE (1991). Achalasia, diffuse spasm and non-specific motor disorders. *Bailliere's Clinical Gastroenterology* **5**, 307–35.

Roth JA and Putnam JBJ (1994). Surgery for cancer of the esophagus. *Seminars in Oncology* **21**, 453–61.

Spechler SJ (1999). AGA technical review on treatment of patients with dysphagia caused by benign disorders of the distal esophagus. *Gastroenterology* **117**, 233–54.

Tobin RW (1998). Esophageal rings, webs, and diverticula. *Journal of Clinical Gastroenterology* **27**, 285–95.

Vaezi MF (1999). Achalasia: diagnosis and management. *Seminars in Gastrointestinal Disease* **10**, 103–12.

Weston S *et al*. (1998). Clinical and upper gastrointestinal motility features in systemic sclerosis and related disorders. *American Journal of Gastroenterology* **93**, 1085–9.

Young MA, Rose S, Reynolds JC (1996). Gastrointestinal manifestations of scleroderma. *Rheumatic Diseases Clinics of North America* **22**, 797–823.

14.7 Peptic ulcer diseases

*John Calam**

Introduction

Peptic ulcers are common: individuals in Western populations have a lifetime risk of 1 in 10 of developing an ulcer. They are a major cause of illness and death and are economically important. Research in the field of peptic ulcers initially focused on the elevated secretion of gastric acid in duodenal ulcer disease. Treatments to combat acid progressed from alkaline antacid preparations through histamine H_2-receptor antagonists, to proton pump inhibitors; maintenance therapy was required because ulcers recurred when the medication was stopped. Our understanding of peptic ulcer disease was radically changed in 1983 by the discovery of *Helicobacter pylori* and thereafter its role in gastric and duodenal ulcers. Most patients with ulcers are infected with *H. pylori*, and eradicating the infection permanently cures the ulcers (Fig. 1). Inhabitants of developing countries are also commonly infected with *H. pylori*, but the prevalence is lower and is falling in regions where better sanitation and hygiene limit transmission.

Fig. 1 Demonstration in 1988 by Marshall and colleagues that eradication *of H. pylori* greatly prolongs remission of duodenal ulcer disease. CLO+ and – refer to the diagnosis of infection after attempted eradication by the biopsy urease test, the CLO test, which Marshall also invented. (Reproduced from Marshall BJ *et al*. (1988). *Lancet* **ii**, 1437–42, with permission.)

As *Helicobacter* is controlled, another ulcerogenic factor has become important. Non-steroidal anti-inflammatory drugs (**NSAIDs**), often prescribed for musculoskeletal conditions which particularly affect the elderly, lead to ulcers in a population that is less well able to withstand the major complications of haemorrhage and perforation. The chance of dying as a result of a peptic ulcer is currently about a thousand times greater in the elderly than in the young, and peptic ulcers thus remain a challenge to modern medicine.

Definition

A gastrointestinal ulcer is defined as a breach in the epithelium that penetrates the muscularis mucosae. If the muscularis is not breached it is called an erosion. Duodenal ulcers and gastric ulcers are often considered together as peptic ulcers but differ considerably with regard to epidemiology, pathogenesis, presentation, and

management: they are discussed separately. This chapter will also address acute erosive gastritis and non-ulcer dyspepsia. Oesophageal ulcers are discussed with erosive oesophagitis in Chapter 14.6. The Zollinger–Ellison syndrome is discussed in Chapter 12.10 and Chapter 14.8.

The major causes of peptic ulcers are *H. pylori* infection and NSAIDs.

Helicobacter pylori

Spiral bacteria were observed by European investigators in the stomachs of animals in the nineteenth century and in humans at the beginning of the twentieth century; but interest waned when a prominent American investigator reported no such bacteria in gastric biopsies from 1000 patients. In 1981 a British study showed that duodenal ulcers remain healed for longer after administration of the antibacterial bismuth than after the H_2-antagonist drug cimetidine. In 1983, the Australians Warren and Marshall succeeded in culturing *H. pylori* (originally known as *Campylobacter pyloridis*) from human gastric biopsies. They were assisted by new selective culture media, and serendipity—growth was achieved when the culture plates were incubated for longer than planned during a holiday. Warren and Marshall soon noticed the association between *H. pylori* infection and gastritis, that almost all patients with duodenal ulcers are infected, and that eradication of infection largely prevents recurrence of ulcers.

The epidemiology of *H. pylori*

Population surveys reveal two patterns of *H. pylori* infection. In developing counties about 80 per cent of the population are infected by the time they are adults and remain infected. In the West, the prevalence rises gradually throughout life to about 60 per cent in old age (Fig. 2). But this is due to a cohort effect rather than gradual acquisition of infection throughout life. The infection is usually acquired before the age of 5 years. Older Westerners were children when infection was much more prevalent than it is now. An exception is that adults frequently acquire infection during wartime: the prevalence of *H. pylori* in the West is distinctly lower in those born after the Second World War. It remains high in the poor and in immigrants from developing countries.

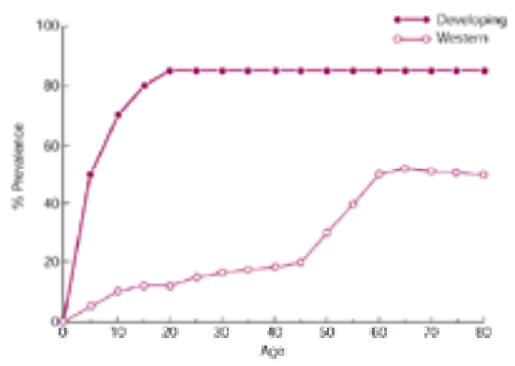

Fig. 2 *H. pylori* prevalence in Westernized versus developing countries. (Adapted from Marshall BJ (1994). *American Journal of Gastroenterology* **89**, S116–118, with permission.)

Bacteriology and transmission of *H. pylori*

Helicobacter pylori is a Gram-negative microaerophilic bacillus (Fig. 3). It uses its spiral shape and four to six flagellae, located at one end, and to move through the gastric mucus layer. It is well adapted to the gastric environment; for example its abundant urease generates alkali by splitting urea into two NH_4^+ ions and one HCO_3^- ion. Indeed *H. pylori* can only survive on gastric-type mucosa, perhaps because its adhesins bind specifically to certain motifs on gastric cells. It is readily cultured from gastric biopsies on selective media, but transport to the laboratory must be rapid. Transmission is from person to person, often within families. It is unclear whether spread is usually oral–oral or faecal–oral. Infection via drinking water occurs in developing countries. There is no significant animal reservoir. The genomes of several *H. pylori* strains have been sequenced and vary considerably. Different strains contain different lengths of 'cag pathogenicity island'. This is a cassette of genes involved in pathogenicity, plus the gene *cagA*. The presence of antibodies to cagA protein in a patient reflects the presence of the island, and is associated with ulcers and cancer. Variations in the gene *vacA* that encodes the *H. pylori* vacuolating toxin also appear to influence the likelihood of these consequences of infection.

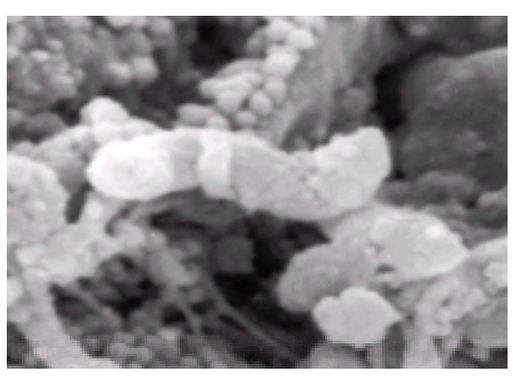

Fig. 3 Electron micrograph of *H. pylori*.

Diagnosis of *H. pylori* infection

Three methods can be used to detect *H. pylori* in gastric biopsies:

1. The biopsy urease test depends on the the ability of the bacterium to generate alkali.
2. *H. pylori* bacteria are readily detected histiologically using special stains.
3. Bacterial culture allows the antibiotic sensitivity of the patient's strain to be determined.

Two tests allow *H. pylori* to be diagnosed without endoscopy:

1. Serology is accurate and convenient but remains positive for several months after successful eradication, and is not useful for determining whether eradication has been successful.
2. The urea breath test. The patient drinks a solution of urea containing carbon atoms labelled with ^{13}C or ^{14}C. Labelled CO_2, generated by bacterial urease, can be detected in the breath by mass spectroscopy or radioactive counting if the infection is present (Fig. 4). This test is ideal for testing the success of eradication, if this is required (see below).

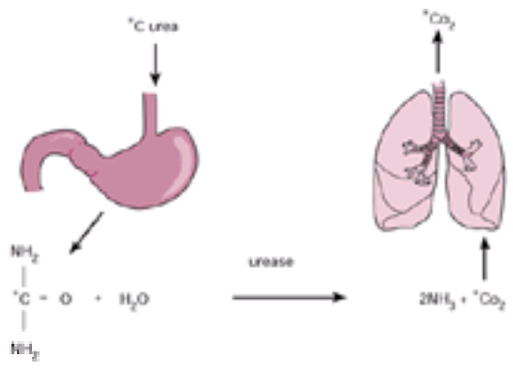

Fig. 4 Urea breath test. The patient drinks a solution of urea containing ^{13}C or ^{14}C carbon. If *H. pylori* is present then the carbon isotope appears in exhaled carbon dioxide.

Factors that determine the accuracy of the diagnostic tests

These tests are all quite accurate but can give incorrect results under certain circumstances: all except serology can give false negative results if the patient has received antibiotics during the past month. Also, the bacterium tends to move from the antrum to the proximal stomach when proton pump inhibitors are given. Biopsies should then be taken from the proximal gastric mucosa as well as the antrum. Proton pump inhibitors can inhibit *H. pylori*'s urease, which can affect results based on this enzyme. Serological tests remain positive for up to a year after the infection has been eradicated.

Disease associations of *H. pylori* infection

Gastritis

This was noted to occur in subjects who were accidentally or experimentally infected with the bacterium. A first infection is followed by a period of low or absent acid secretion, which lasts for weeks or months and which was called 'epidemic achlorhydria' before the cause was discovered. Before *H. pylori* was identified, it was well known that patients with duodenal ulcers have antral gastritis, whilst those with gastric ulcers and gastric cancer have pangastritis. *H. pylori* clearly contributes to the gastritis associated with these diseases because the inflammation resolves when the infection is eradicated. *H. pylori* gastritis is typically 'chronic superficial', meaning that lymphocytes and macrophages are located superficially. Infiltration with neutrophils ('activity') is variable. Prolonged infection predisposes to mucosal atrophy in which the number of chief and parietal cells in the gastric mucosa falls. Atrophy may then proceed to intestinal metaplasia in which the gastric epithelium is replaced with one resembling small or large intestine. The inflammatory response may diminish or even eradicate the infection by reducing acid secretion and the number of gastric cells to which *H. pylori* can adhere. Inflammatory mechanisms in *H. pylori* gastritis have been studied in detail because of their role in the pathogenesis of ulcers and cancer and their relevance to vaccine development. This topic is beyond the scope of this chapter. Briefly, *H. pylori* bacteria remain extracellular but induce gastric epithelial cells to express class molecules and chemokines such as IL-8. These recruit and activate inflammatory cells that express diverse inflammatory mediators including tumour necrosis factor-a, and products such as reactive oxygen species. *H. pylori* gastritis cannot be diagnosed on the basis of endoscopic appearance alone because the mucosa usually either looks normal or mildly reddened.

Duodenitis

Some individuals develop patches of metaplastic gastric mucosa in their proximal duodenum. This change depends on the rate of gastric acid secretion, being abundant in the Zollinger–Ellison syndrome and absent in pernicious anaemia. Gastric metaplasia is important because it allows *H. pylori* to colonize the duodenum leading to duodenitis. This can be erosive and increases the risk of chronic duodenal ulceration.

Gastric cancer

The World Health Organization has classified *H. pylori* as a gastric carcinogen. Infection is associated with an approximately eightfold increased risk of gastric cancer. Eradication of *H. pylori* from Japanese patents with early gastric cancer greatly diminished the risk of recurrent cancer after endoscopic resection. Gastric cancer is beyond the scope of this chapter but bears on the question of whether to eradicate *H. pylori* from individuals without ulcers—an issue which is currently unresolved.

Gastric MALT lymphoma (see Chapter 14.9.4)

Tumours of gastric mucosa-associated lymphoid tissue (**MALT**) are much more prevalent in *H. pylori* infection. This is as expected because uninfected gastric mucosa contains hardly any lymphocytes. Lymphomas restricted to the gastric mucosa usually disappear when *H. pylori* is eradicated because *H. pylori*-specific T cells provide contact-dependent help for growth of malignant B cells. The abnormal B-cell clone often remains detectable by gene rearrangement studies, and tumours reappear if the patient is reinfected. These lesions are less likely to respond to *H. pylori* eradication alone if they extend beyond the gastric mucosa. Chemotherapy or surgical excision may then be indicated, and therefore these tumours are generally best managed in a specialist centre.

Likelihood of disease following *H. Pylori* infection

Most individuals who are infected with *H. pylori* do not develop clinical disease. The likelihood of disease depends on the nature of the infecting strain, which is highly variable. The genome of more aggressive strains contains a pathogenicity island with genes encoding proteins that stimulate cytokine production. Infection with such strains can be identified by the presence of antibodies to cagA protein which is highly antigenic. Variations in *vacA* also affect the likelihood of serious clinical outcomes of the infection. However, people from parts of the world where heavy infection is frequent often harbour infection with more than one strain of *H. pylori,* which complicates this relationship.

Treatment of *H. pylori* infection

At the time of writing the following 1-week triple therapies are recommended in the British National Formulary because they provide the best combination of efficacy and convenience:

- Proton pump inhibitor twice daily plus amoxicillin 1000 mg twice daily plus clarithromycin 500 mg twice daily.
- Proton pump inhibitor twice daily plus metronidazole 400 mg thrice daily plus clarithromycin 500 mg twice daily.

The proton pump inhibitor can be either omeprazole 20 mg twice daily, or lansoprazole 30 mg twice daily. An H_2 antagonist, such as cimetidine, or an antibacterial agent such as bismuth citrate 400 mg twice daily can be used instead of the proton pump inhibitor.

The choice between these regimens depends on the local rate of metronidazole resistance. This is increased in inner-city areas because of immigration from developing countries where metronidazole is widely used, but is lower elsewhere in the United Kingdom. Resistance to clarithromycin is currently about 5 per cent in the United Kingdom but higher in the rest of Europe. Resistance to amoxicillin is exceedingly rare. It is important to encourage each patient to comply accurately, because about 60 per cent of strains will become resistant to metronidazole or clarithromycin if they are exposed to these without being eradicated. This lessens the chance that further attempts at eradication will succeed. An effective regimen for such cases is a proton pump inhibitor given twice daily, tripotassium dicitratobismuthate (TDB, DeNol) 120 mg four times daily 30 min before meals and at night, tetracycline 500 mg four times daily, and metronidazole 400 mg thrice daily all given for 2 weeks. *H. pylori* eradication regimens are subject to continued modification and it is thus recommended that current British National Formulary or local guidelines be consulted.

Whether eradication therapy has been successful or not can be determined by the urea breath test, but this must be performed at least 4 weeks after the end of the

eradication regimen to avoid false negative results. Patients must be retested if persistent infection would be hazardous, for example after haemorrhage of an ulcer, but this is unnecessary if symptoms from an uncomplicated ulcer have disappeared.

The role of non-steroidal anti-inflammatory drugs in peptic ulceration

Ingestion of NSAIDs (including aspirin) is the main cause of gastric and duodenal ulcers that are not associated with *H. pylori* infection. The ratio of *H. pylori*- to NSAID-associated ulcers is about 95:5 in the duodenum and 80:20 in the stomach because NSAIDs affect the stomach more than the duodenum. The ratio also depends on local *H. pylori* prevalence and NSAID use. In prosperous areas with low *H. pylori* prevalence and many elderly persons taking NSAIDs, the latter is now the main culprit. This is important because the elderly are less able to withstand the ulcer-related complications of haemorrhage and perforation. Use of NSAIDs greatly increases the risk of admission to hospital with an ulcer, and the chance of ulcer haemorrhage during ingestion of NSAIDs increases with the dose and duration of therapy. It also varies from drug to drug. NSAIDs act by inhibiting the enzyme cyclo-oxygenase which converts arachidonic acid to prostaglandins which normally protect gastrointestinal epithelia by increasing blood flow and secretion of mucus and bicarbonate. The isoenzymes cyclo-oxygenase-1 and cyclo-oxygenase-2 are largely responsible for the mucosal protective and anti-inflammatory effects respectively of NSAIDs (Fig. 5). Therefore newer NSAIDs, such as celecoxib and rofecoxib, that are highly selective for cyclo-oxygenase-2, control inflammation with less risk of causing ulcers.

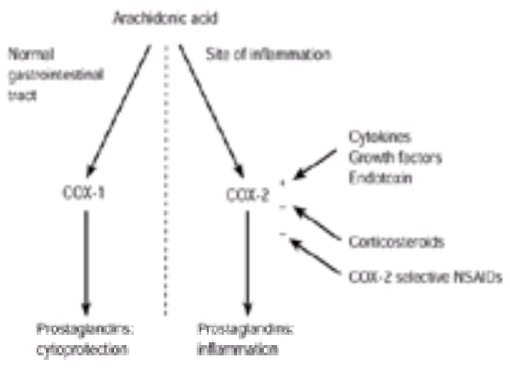

Fig. 5 The different roles of cyclo-oxygenase-1 and -2. Most existing NSAIDs inhibit both cyclo-oxygenase-1 and cyclo-oxygenase-2.

Ingestion of NSAIDs leads to a 'chemical' gastritis that is a milder form of the gastritis produced by reflux of bile into the stomach. The mucosa shows foveolar hyperplasia, oedema, vasodilatation and congestion, and few inflammatory cells. Ingestion of NSAIDs very often causes acute erosive gastritis that typically goes unnoticed unless it causes frank bleeding. Erosions tend to heal by the process of mucosal adaptation when NSAIDs are taken for more than about a week. Chronic ulcers develop with more prolonged ingestion of NSAIDs.

Prevention and control

The use of alternative medications such as paracetamol is the best way to prevent NSAID-associated ulcers. NSAIDs are now generally avoided in non-inflammatory conditions such as osteoarthritis. When NSAIDs are required a cyclo-oxygenase-2-selective drug is preferred. If an ulcerogenic NSAID has to be continued in a patient at risk then ulcer prophylaxis is indicated. Proton pump inhibitors and misoprostol are both effective in healing and preventing NSAID-associated ulcers. The former are generally preferred because they have fewer side-effects: misoprostol in particular tends to cause diarrhoea (see below).

Other factors which increase the risk of peptic ulceration

Cigarette smoking strongly predisposes to peptic ulceration. This may be because nicotine stimulates acid secretion and reduces mucosal blood flow, but eradication of *H. pylori* prevents the ulcerogenic effect. A moderate alcohol intake is not harmful and might even decrease the risk of duodenal ulceration. Duodenal ulceration is considerably more common in patients with cirrhosis or chronic pancreatitis, presumably because bile and pancreatic juice normally neutralize gastric acid when it enters the duodenum. It is usual to recommend a 'bland' diet in ulcer disease but no particular food is known to increase the risk of duodenal ulcer disease. Indeed the traditional Japanese diet, high in salt, pickles, and raw fish, may contribute to the low prevalence in Japan of duodenal ulcer disease and the high prevalence of gastric ulcers and cancer by causing corpus gastritis, which diminishes acid secretion. Hyperparathyroidism increases the prevalence of duodenal ulcer disease. Elevated circulating calcium concentrations stimulate gastrin release, and parathyroid tumours are associated with gastrinomas in the multiple endocrine neoplasia-1 syndrome (see Chapter 12.10).

The epidemiology of peptic ulcers

Ulcers of the stomach and duodenum led to 4111 deaths in England and Wales in 1996. This remains a cause for concern, but most peptic ulcers do not produce life-threatening illness.

Duodenal ulcer epidemiology

Duodenal ulcers are often asymptomatic so their prevalence can only be established by investigating apparently healthy people. Such a study in Finland showed that 1.4 per cent of the entire population had a duodenal ulcer at any time. The lifetime prevalence was 10 per cent in males and 4 per cent in females. In the United Kingdom duodenal ulcers are more common in the north of the country and in urban rather than rural regions. The incidence of duodenal ulcers gradually increases with age but peaks at about 60 years of age. In most parts of the world duodenal ulcers are about three times as common as gastric ulcers, but gastric ulcers are more common in some places including Japan, Sri Lanka, and the Andean region. The epidemiology of peptic ulcers is changing quite rapidly: in late nineteenth-century England, gastric ulcers occurred in young women and were much more common than duodenal ulcers. From the turn of that century until about 1960 the incidence of duodenal ulceration rose several times to become more common than gastric ulceration, which is now uncommon under the age of 40 years. Factors that might have contributed include cigarette smoking and the spread of *H. pylori* during the Depression and wartime. Also the integrity of the acid-secreting mucosa might have improved as the increasing use of refrigeration diminished the salt content of the diet and fresh fruit and vegetables provided antioxidants. Since 1960, the incidence of duodenal ulceration has stopped rising and may even have declined. The declining prevalence of *H. pylori* and early treatment of it has further changed the situation so that it is now becoming unusual for a gastroenterologist to see active duodenal ulcer disease in some prosperous regions in the west.

Gastric ulcer epidemiology

In most parts of the world chronic gastric ulcers are less common than duodenal ulcers, but gastric ulcers are still quite common. The most accurate epidemiological studies are from Finland. At any time about 0.3 per cent of that population has a gastric ulcer. The lifetime prevalence of gastric ulceration is 4 per cent in males and 3 per cent in females and the prevalence of *H. pylori* infection is similar in males and females. Gastric ulcers are rare in people under the age of 40 years and tend to occur in the lower socio-economic groups.

Duodenal ulcers

Duodenal ulcer is a common illness with serious complications. Over 90 per cent of these ulcers are due to *H. pylori* infection and can be permanently cured by its eradication (Fig. 1) but it is not possible to achieve this worldwide. Duodenal ulceration in the absence of *H. pylori* infection is the exception and is usually due to NSAIDs, Crohn's disease, or the Zollinger–Ellison syndrome.

Pathogenesis

Duodenal ulcer disease is associated with antrum-predominant *H. pylori* gastritis: the proximal stomach that secretes acid is relatively spared. *H. pylori* antritis

suppresses expression of the inhibitory peptide somatostatin and increases release of the acid-stimulating hormone gastrin. Gastrin elevates acid secretion both immediately and through the trophic effects of gastrin on the oxyntic mucosa. Eradication of *H. pylori* reverses these effects on gastric physiology. The increase in acid secretion directly damages the duodenal mucosa. Acid hypersecretion also produces gastric metaplasia in the proximal duodenum (Fig. 6). This allows *H. pylori* bacteria to colonize the duodenum and produce duodenitis that further impairs mucosal integrity and diminishes duodenal bicarbonate secretion. These changes result in duodenal ulcers.

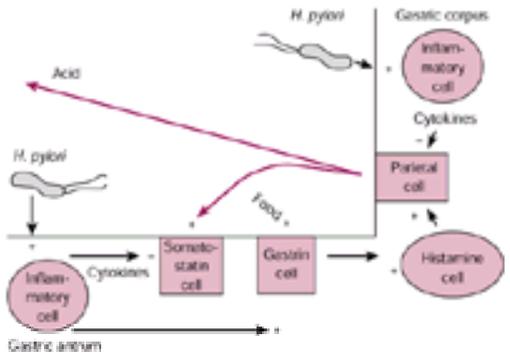

Fig. 6 Pathophysiology of *H. pylori* infection. Food stimulates gastrin cells in the gastric antrum to release gastrin. Circulating gastrin stimulates enterochromaffin-like cells in the gastric corpus to release histamine that stimulates adjacent parietal cells to secrete acid. Normally acid stimulates secretion of somatostatin that inhibits further gastrin release. Cytokines released in *H. pylori* antral gastritis suppress antral somatostatin-cells, leading to hypersecretion of gastrin and acid. But cytokines released in *H. pylori* gastritis of the gastric corpus tend to suppress parietal cells leading to diminished acid secretion.

Clinical presentation

Duodenal ulcers typically present with pain that is dull and located in the epigastrium or to the right of it over the duodenum itself. It is characteristically relieved by eating, then gets worse when the stomach empties. The pain usually wakes the patient from sleep in the middle of the night and is relieved by eating food, drinking milk, or taking an alkali preparation (antacid). Night pain is due to high nocturnal acid secretion without the buffering effect of food. The pain is also episodic with exacerbations lasting a few weeks separated by pain-free periods. These last for several weeks or months and probably reflect spontaneous healing of the ulcer. Pain radiating to the back suggests a posterior penetrating ulcer. Note that many ulcers do not cause pain and present with bleeding. This is particularly likely in the elderly, or if the patient is taking NSAIDs. Patients with duodenal ulcer often have other symptoms such as retrosternal burning and acid regurgitation. Nausea and vomiting are unusual and appetite is preserved. Persistent vomiting suggests pyloric stenosis. Symptoms resulting from other complications are described below.

Differential diagnosis

Gastro-oesophageal reflux disease also causes pain when the patient is in bed but patients usually notice an acid taste in the mouth or burning in the chest. Pancreatic pain typically radiates to the back, is exacerbated by eating, and is relieved by leaning forwards. Gallstone pain tends to be colicky and is exacerbated by eating fat. Pain due to the irritable bowel syndrome can occur in the epigastrium and be affected by eating but it usually extends to the lower abdomen and is typically affected by defaecation. Duodenal ulceration seen at endoscopy is usually due to *H. pylori* but NSAIDs, Crohn's disease, and the Zollinger–Ellison syndrome need to be considered, particularly if *H. pylori* is absent or if ulceration persists after it is eradicated.

Pathology

In duodenal ulcer disease the duodenum contains patches of gastric metaplasia colonized by *H. pylori* bacteria leading to infiltration with lymphocytes and neutrophil polymorphs. The ulcer itself consists of a breach in the epithelium with ulcer slough, inflammatory cells, and varying amounts of collagen scar in the base.

Diagnosis of duodenal ulceration

Duodenal ulcers are diagnosed most accurately by endoscopy. Antral biopsies can be tested for *H. pylori* and local treatment can be applied if the ulcer is bleeding (see below). Between episodes of ulceration the duodenum may show scarring or deformity. The presence of duodenitis also provides a clue to the diagnosis. High-quality double-contrast barium radiology is only slightly inferior. If an ulcer is seen, the likelihood of *H. pylori* infection is about 90 per cent—which is sufficient to justify treatment.

Treatment

Healing of duodenal ulcers can be accelerated by a number of acid-suppressing or mucosal protective drugs, but *H. pylori* eradication is now the mainstay of treatment (see above). Of the acid suppressors, proton pump inhibitors are more effective that histamine H_2-receptor antagonists, but the difference is not great in duodenal ulcer disease. *H. pylori* eradication should be commenced immediately. It is usually unnecessary to continue acid suppression after eradication therapy because eradication heals duodenal ulcers rapidly. However, it is sensible to continue acid suppression for about 8 weeks if the ulcer was complicated by bleeding or pyloric stenosis. *H. pylori* occasionally proves impossible to eradicate: healing may then be achieved and maintained by acid suppression. Alternatively prolonged remissions can be induced by courses of tripotassium dicitratobismuthate (DeNol).

Prognosis

Recurrence of *H. pylori*-related duodenal ulceration is uncommon after successful eradication. The rate of reinfection with *H. pylori* is about 0.7 per cent per annum in Western adults. Higher rates of reinfection have been reported in developing countries but this is not universal and the reinfection rate is 1 per cent per annum in China. Apparent reinfection in the West is often actually persistence of the initial infection. Recurrent ulceration in the absence of *H. pylori* may be due to NSAIDs, Crohn's disease, or the Zollinger–Ellison syndrome. In some patients recurrent ulceration after eradication of *H. pylori* is associated with persistent high acid output in the absence of a gastrinoma.

Gastric ulcers and erosions

Breaches in the gastric epithelium take two forms. Chronic benign gastric ulcers are relatively large and usually single. Acute erosive gastritis produces many small ulcers or erosions.

Chronic benign gastric ulcer

Gastric ulcers are important because they cause ill health, bleed and occasionally perforate, and because they are sometimes difficult to distinguish from gastric cancers. Most are associated with *H. pylori* infection and eradicating the bacterium greatly diminishes recurrence of ulcers, which demonstrates the causative role of the bacterium. Most other chronic gastric ulcers are associated with ingestion of NSAIDs. At the beginning of the twentieth century gastric ulcers were more common in the United Kingdom than duodenal ulcers and often occurred in young women. Now they are less prevalent than duodenal ulcers and typically occur in older people with low incomes. In some countries, including Japan, gastric ulcers are still more common than duodenal ulcers.

Pathogenesis

Ulceration occurs when luminal aggressive factors overcome mucosal defence. Suppression of acid accelerates healing of gastric ulcers so acid clearly contributes, but rates of acid secretion are normal or slightly below normal in these patients so acid cannot be regarded as the main cause. The duodenogastric reflux is increased

in patients with gastric ulcers, and the refluxed fluid weakens the mucosal barrier. *H. pylori* might cause ulceration directly by releasing a toxin such as its vacuolating toxin. Alternatively products of the inflammatory cells that it attracts, such as oxygen radicals or proteolytic enzymes, might compromise mucosal integrity (Fig. 7). The hydrophobic barrier of the mucosa is impaired in *H. pylori* infection. Non-steroidal anti-inflammatory drugs cause ulceration largely through inhibiting the production by cyclo-oxygenase-1 of protective prostaglandins in the gastrointestinal tract. In future this can be avoided by using highly selective inhibitors of cyclo-oxygenase-2, such as celecoxib and refecoxib. This is the isoenzyme that is involved in the inflammatory diseases for which these drugs are used (see above).

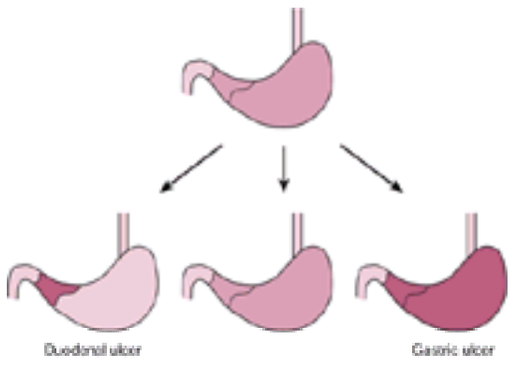

Fig. 7 Relationship between the distribution of *H. pylori* gastritis and the clinical outcome of the infection. Initially the infection affects the entire stomach. Most patients then develop a mild pangastritis and no clinical disease. Patients with duodenal ulcer have antrum-predominant gastritis, a healthy corpus and high acid secretion. Marked pangastritis is associated with low acid secretion and an increased risk of gastric cancer.

Clinical features

Patients with chronic gastric ulcers tend to be over 40 years old and from lower socio-economic groups. Epigastric pain is the most frequent symptom. It occasionally radiates to the back if the ulcer is located posteriorly. Food or antacids usually relieve it. It typically occurs in exacerbations lasting for several weeks with symptom-free periods in between. Night pain occurs in a minority of patients with gastric ulcer compared with most of those with duodenal ulcers. Gastric ulcers quite often produce nausea, anorexia, or weight loss suggestive of gastric cancer. Some patients vomit, but many ulcers have no symptoms until the patient presents with haemorrhage, and in practice one cannot reliably distinguish between these diseases on the basis of symptoms. Most patients with epigastric pain have non-ulcer dyspepsia or gastro-oesophageal reflux disease, rather than gastric or duodenal ulcers: gallstones can also cause epigastric pain but this tends to be colicky, at the right costal margin, and is exacerbated by fatty food. Pancreatic pain tends to be more constant, radiates to the back, and may be relieved by leaning forward. The differential diagnosis of gastric ulceration includes gastric cancer, lymphoma, Crohn's disease, syphilis, tuberculosis, and sarcoidosis. Importantly, gastric cancers occasionally produce ulcers that resemble simple peptic ulcers and these can heal when acid secretion is suppressed; further confusion may arise because biopsies from the ulcer edge do not always contain malignant cells.

Pathology

Histological examination of a chronic gastric ulcer shows a breach in the gastric mucosa that penetrates the muscularis mucosae. The base is composed of chronic inflammatory cells, and slough and fibrous tissue in varying proportions. The edge of the ulcer shows evidence of increased proliferation. The surrounding gastric epithelium may show *H. pylori* gastritis or chemical gastritis caused by NSAIDs or bile reflux.

Laboratory diagnosis

Cancer is excluded by examining multiple biopsies from the edge of the ulcer. *H. pylori* is sought as described above, but biopsy-based methods are prone to give false negative results because the stomachs of these patients typically contain areas of atrophy or intestinal metaplasia which are not colonized by the bacterium.

Treatment

Treatment of chronic gastric ulcers is by removing the cause if possible and giving ulcer healing agents if necessary. *H. pylori* is eradicated as described above. NSAIDs are discontinued if possible. In addition, administration of a proton pump inhibitor such as omeprazole accelerates healing of the ulcer. Until the ulcer heals, the patient should have an endoscopic examination every 4 to 6 weeks and multiple biopsies should be taken from the ulcer edge to exclude cancer. If NSAIDs have to be continued, the stomach can be protected against further ulceration by a proton pump inhibitor or misoprostol.

Prognosis

The immediate dangers are from complications and undiagnosed malignancy. Gastric ulcers are very likely to recur if they are healed by acid suppression alone. However, they unlikely to reappear if the cause is removed or adequate prophylaxis is given.

Acute erosive or haemorrhagic gastritis

Introduction

Acute gastric erosions are breaches in the gastric epithelium that are often multiple and small ({lt} 3 mm), reflecting an acute diffuse response to the gastric insult. If they do breach the muscularis mucosae they are called acute gastric ulcers. Cushing and Curling ulcers are historical terms used to describe ulcers in patients with head injury and burns respectively, but most acute gastric ulcers are now associated with severe illness or ingestion of NSAIDs or alcohol.

Acute gastric erosions in severe illnesses

These are common—80 per cent of patients receiving respiratory support from mechanical ventilators studied after 3 days in a British intensive care unit had erosions:

a. Acute gastric ulcers in patients with injury to the central nervous system were described by Cushing. Ulcers that frequently bleed occur in 50 to 70 per cent of patients with head injuries. They also occur after spinal injuries and after surgery to the central nervous system. The injury might cause the ulcers by increasing the release of gastrin because these patients have elevated acid secretion and plasma gastrin levels.
b. Curling described acute duodenal ulcers in patients with severe burns; however, acute gastric ulcers are even more common, and were present in 86 per cent of such patients in one recent study.
c. Acute gastric ulcers occur in patients suffering from many severe illnesses, including severe trauma, sepsis, shock, and major organ failure, presumably because these diminish gastric perfusion and compromise the metabolic integrity of the gastric mucosa.

Acute gastric ulcers due to ingested substances

Non-steroidal anti-inflammatory drugs including aspirin produce acute gastric ulcers by inhibiting synthesis of protective prostaglandins by cyclo-oxygenase-1. These impair mucosal protection by diminishing mucosal blood flow and the secretion of mucus and bicarbonate. Ulceration is most extensive shortly after the start of therapy: the stomach then adapts and ulceration subsides.

Alcohol induces acute gastric ulceration by a direct toxic effect. Studies in experimental animals suggest that vascular thrombosis also contributes to alcohol-induced ulceration.

Clinical features

The patient may complain of epigastric pain, nausea, anorexia, or vomiting but symptoms are usually absent until haemorrhage occurs The bleeding may be occult, show itself as small amounts of blood or 'coffee grounds' in gastric aspirates, or occur suddenly as a substantial haemorrhage.

Diagnosis

Acute gastric ulcers and erosions are readily seen with an endoscope, but the appearance varies from multiple erosions to submucosal haemorrhages to one or more points of active bleeding. During healing the ulcer may be elevated by surrounding oedema. Endoscopy is the principal diagnostic procedure and angiography is rarely needed.

Treatment and prevention

This disease is both treated and prevented by minimizing injurious factors. This involves correction of shock, sepsis, and respiratory and renal failure and avoidance of NSAIDs and alcohol. It is also important to correct coagulopathy that is frequently present in sick patients. Suppression of acid secretion is protective. This may be achieved with a histamine H_2-receptor antagonist or a proton pump inhibitor. However, it is recognized that acid suppression allows overgrowth of bacteria in the stomach lumen. This increases the risk of sepsis, particularly through aspiration in patients with impaired consciousness or cough reflex; hence prophylaxis with the antipepsin agent sucralfate, is generally preferred in patients undergoing intensive care.

The management of patients with established acute gastric ulceration involves blood transfusion, correction of coagulopathy, and suppression of acid secretion. The benefit from acid suppression is slight once haemorrhage has occurred. Endoscopic intervention and surgery are not usually appropriate because the condition is diffuse. If an operation is required it may need to be radical to control the haemorrhage (for example total gastrectomy).

Complications of ulcer disease

Haemorrhage

Haemorrhage remains a challenging problem and is the main cause of death from peptic ulcers. Blood loss may be slow and present as unexplained anaemia but more typically presents acutely with haematemesis or melaena or both with varying degrees of hypovolaemic shock. Such cases should be managed jointly by physicians and surgeons from the outset. Older patients need particular attention because they are much more vulnerable to the effects of hypovolaemia. Immediate action should be directed to the correction of circulatory shock by blood transfusion. Further transfusion may be required to keep the patient's haemoglobin level above 10 g/dl. Disorders of coagulation, pre-existing or due to transfusion, should also be corrected. Endoscopy is performed, preferably after the patient's condition has stabilized, to define the source of bleeding and to apply endoscopic treatments. Ulcers which are actively bleeding or show stigmata, such as adherent clot or a visible vessel, which make further bleeding likely can be treated with lasers, heater probes, or local injection of adrenaline. Rebleeding is an indication for surgery. After the acute episode, it is important to attend to the cause of the ulcer. Eradication of *H. pylori*, if it is present, greatly diminishes the frequency of further episodes of bleeding in future, but this is a measure that is frequently overlooked. Non-invasive tests can be used to diagnose the infection if biopsies were not taken at the time of endoscopy during the acute bleeding. NSAIDs are contraindicated after haemorrhage from NSAID ulcers. It is generally unnecessary to give long-term prophylaxis if the cause of ulceration has been removed, but this may be advisable if the patient is elderly, frail, or does not have rapid access to hospital.

Perforation

Ulcer perforation typically presents with a sudden onset or worsening of pain with considerable abdominal tenderness followed by the onset of peritonitis with board-like rigidity, rebound tenderness, and loss of bowel sounds. Gas in the peritoneum may lead to loss of liver dullness to percussion, and is usually visible beneath the diaphragm on erect chest radiograph. The patient is unwell with tachycardia, leucocytosis, and sometimes fever. The differential diagnosis includes acute pancreatitis, acute cholecystitis, and other causes of an acute abdomen such as gut infarction or perforation of other organs (see Chapter 14.3.1). The patient is resuscitated with intravenous fluids and antibiotics before transfer to theatre for repair of the perforation. Note that perforation, like other abdominal emergencies, may have a less specific presentation in the elderly, who may collapse or suffer confusion rather than the characteristic pain. Again, once the acute event has been dealt with, it is important to identify and treat the cause of the ulcer.

Pyloric stenosis

Repeated duodenal ulceration sometimes leads to stenosis of the pyloric canal or proximal duodenum. The narrowing is due to oedema as well as fibrosis so it can resolve without the need for surgery. The main symptom is vomiting which may contain food eaten the previous day. Typical symptoms of duodenal ulceration may or may not precede the onset of vomiting. Patients rapidly become dehydrated and develop hypokalaemia with a metabolic acidosis, they may also be malnourished. A succussion splash, which is normally present up to 4 h after a meal, is present at other times. Barium radiology or upper endoscopy show a distended stomach containing retained food and secretions and with a narrowed pyloric canal. The differential diagnosis includes cancer of the distal stomach. Most of these patients settle without the need for further intervention if treated by gastric aspiration, acid suppression, and intravenous fluids for a few days. If not, the stenotic region can be dilated using a balloon passed via an endoscope. A few patients require an operation, but again the cause of ulceration must be identified and appropriately treated.

Gastrinoma

The Zollinger–Ellison syndrome comprises an association of pancreatic tumour, gastric hypersecretion, and intractable ulceration. The tumour causes the syndrome by releasing gastrin so it is called a gastrinoma. This disease is discussed in Chapter 12.10.

Non-ulcer dyspepsia

Dyspepsia is upper abdominal or lower chest discomfort or pain, related to eating, which may be accompanied by other gastrointestinal symptoms such as nausea, vomiting, anorexia, or distension. Non-ulcer dyspepsia refers to cases of dyspepsia where no ulcer or other cause such as gallstones or gastro-oesophageal reflux disease is found. Historically, it was the introduction of modern investigative techniques, particularly endoscopy, which allowed these patients to be identified.

Advances in clinical investigation allow some patients considered to have non-ulcer dyspepsia to be diagnosed and treated appropriately. For instance oesophageal pH studies might change the diagnosis to endoscopy-negative gastro-oesophageal reflux disease (GORD) (see Chapter 14.6). Some patients may therefore elude diagnosis because in effect they are not fully investigated. The diagnosis of non-ulcer dyspepsia is not intellectually satisfying but it is important to make. First the patient benefits from the reassurance of knowing that he or she has a recognized medical condition; secondly insight will be gained ultimately into the causes of non-ulcer dyspepsia and how it behaves in response to different treatments.

Epidemiology

Non-ulcer dyspepsia is exceedingly common. In one survey in southwest England, 38 per cent of the entire adult population had had dyspepsia during a 6-month period and a further 25 per cent gave a past history of dyspeptic symptoms. Many or even most of these individuals have non-ulcer dyspepsia.

Symptoms and their pathogenesis

Patients report a variety of symptoms. Research reveals a series of abnormalities of physiological function. Symptoms and the corresponding pathophysiology are described together below; in practice the relationship between symptoms and disorders of function remains ill defined and many patients have more than one symptom. A common theme is that irritation, mild injury, or anxiety diminish thresholds for perception of pain and discomfort in the gastrointestinal tract.

Burning

Many patients with dyspepsia report burning in the epigastrium or chest, or other symptoms suggesting gastro-oesophageal reflux. Irritation of the oesophagus, for example by alcohol or previous reflux episodes, diminishes the threshold for oesophageal pain, so that very mild reflux then causes symptoms. If studies of oesophgeal pH and manometry show definite reflux, the patient can be reclassified as having endsocopy-negative gastro-oesophgeal reflux disease. In busy clinical practice this distinction may be considered unnecessary because the treatment of the two conditions is similar and usually empirical.

Distension

A frequent complaint is that the abdomen feels distended, a symptom that also occurs in irritable bowel syndrome, but if the distension is largely felt in the upper abdomen and after meals, it is considered to be typical of non-ulcer dyspepsia. In addition, there is frequently a sensation of early statiety after meals and a proportion of non-ulcer dyspeptics show delayed gastric emptying on scintigraphy, the cause of which is unknown, although anxiety may contribute. Reflux of duodenal contents into the stomach also occurs more frequently in non-ulcer dyspeptics. The differential diagnosis includes ascites and obesity. Some patients create a bizarre appearance of distension by contracting their diaphragms and increasing their lumbar lordosis.

Pain

Pain is a frequent symptom in non-ulcer dyspepsia, although its origin is usually difficult to establish. It may be due to muscular spasm. The differential diagnosis includes diffuse oesophageal spasm and achalasia as well as other painful conditions including pancreatitis and gallstone disease.

Nausea, vomiting, and satiety

If nausea and vomiting persist in the absence of an organic lesion it is important to exclude other causes such as drugs, metabolic disease, and disorders of the inner ear or central nervous system. Otherwise it is important to consider bulimia, anorexia nervosa, and psychogenic vomiting.

Diagnosis

The physical examination is unremarkable apart from the upper abdomen which is often tender. Investigation aims to exclude other diseases. This is partly to direct treatment and partly so that the patient can be reassured. The condition is so prevalent that it is unnecessary and impractical to investigate all patients fully. Clinical judgement is required. Upper endoscopy is indicated if the picture suggests organic disease or if symptoms persist. Upper abdominal ultrasound scanning is indicated if gallstones are suspected. Whether to test all dyspeptics for *H. pylori* remains controversial (see below). The infection has proved to be no more prevalent in patients with non-ulcer dyspepsia than in the general population.

Treatment

Explanation of the diagnosis with reassurance may relieve anxiety and thus diminish symptoms. Remaining symptoms can then usually be managed with antacids without the need for further medical attention. If treatment is prescribed proton pump inhibitors are most effective but also most expensive (Fig. 8). Alkalis, histamine H_2-receptor antagonists, and prokinetics are often helpful and less expensive. Most large randomized double-blind studies show that eradication of *H. pylori* does not improve symptoms in non-ulcer dyspepsia.

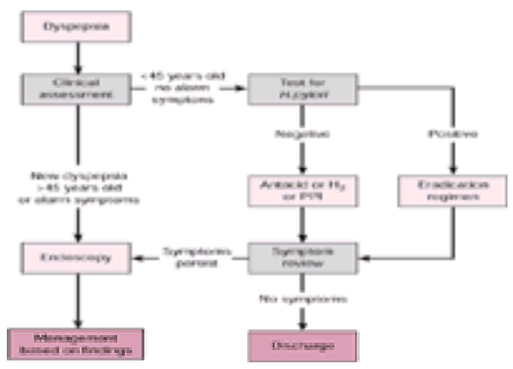

Fig. 8 Flow chart for the management of a patient presenting with dyspepsia of unknown cause that has failed to respond to lifestyle measures and antacids: H_2 = histamine H_2-receptor antagonist; PPI = proton pump inhibitor.

Prognosis and areas of uncertainty

Non-ulcer dyspepsia is a benign condition with no complications and a good prognosis. Only a minority of patients require long-term maintenance therapy. The problem is that some of the many dyspeptics in the community will develop serious organic disease. In a recent study, Swedes who reported symptoms of reflux were eight times more likely than healthy control subjects to develop adenocarcinoma of the gastro-oesophageal junction. Refluxed acid causes Barrett's oesophagus where these cancers originate (see Chapter 14.6). Therefore we need to learn how to manage this risk in the population of patients with dyspepsia. The second area of controversy is whether to eradicate *H. pylori* if it is found in a patient with non-ulcer dyspepsia.

Strategies for management of peptic ulcers and dyspepsia: the role of tests for *H. pylori*

The investigation and treatment of these conditions consume considerable healthcare resources (Fig. 9). Appreciation of the pathogenic role of *H. pylori* has prompted debate over how dyspepsia can be managed more efficiently. All agree that patients with recent onset of dyspepsia over the age of 45 years should be endoscoped promptly to exclude early gastric cancer. Three strategies have emerged for the management of younger patients:

1. 'Treat symptomatically and endoscope those with persisting symptoms'. This is the traditional approach. Initial treatment is with alkalis and less expensive acid-suppressing drugs. Those with persisting or troublesome symptoms are referred for endoscopy. Diagnoses such as ulcer, oesophagitis, and non-ulcer dyspepsia and the *H. pylori* status can then be made precisely, and management organized accordingly. The argument against this policy is that it does not make intelligent use of testing for *H. pylori* to rationalize demand for endoscopy.
2. 'Test and investigate'. The rationale is as follows: Research shows that patients under 45 years of age who are serologically negative for *H. pylori* and not taking NSAIDs are very unlikely to have peptic ulcers or gastric cancer. Therefore if these individuals are not normally endoscoped, endoscopy waiting lists and the number of negative examinations can be reduced. Negative *Helicobacter* serology allows the patient to be reassured that they have no serious disease and can be treated symptomatically with alkalis and acid-suppressing drugs. In practice, the main issue is whether the absence of *H. pylori* is sufficiently reassuring to relieve anxiety and the pressure for further investigations.
3. 'Test and treat'. Eradication of *H. pylori* heals and prevents gastric and duodenal ulcers. The World Health Organization has declared that *H. pylori* is a class 1 carcinogen. Therefore it is proposed to test all patients under 45 years of age with dyspeptic symptoms for *H. pylori*, and treat those who are infected. Endoscopy is reserved for patients whose symptoms persist after *H. pylori* eradication and those with new dyspepsia over the age of 45 years. The issue here is whether it is desirable to eradicate *H. pylori* from patients without ulcers. Most dyspeptics under 45 years of age do not have ulcers, even if they are infected with *H. pylori*. Overall this policy appears seems sensible. The main arguments against it are that antibiotic use will increase and that there is some controversial evidence that *H. pylori* protects against oesophageal disease. Overall, policies based on testing for *H. pylori* are less efficient in populations where the prevalence of the infection is very low, such as prosperous parts of North America, or very high, as in developing countries or in recent immigrants from them. Furthermore, in parts of the world with inadequate sanitation, reinfection rates can be as high as 30 per cent per annum, which diminishes the value of eradication.

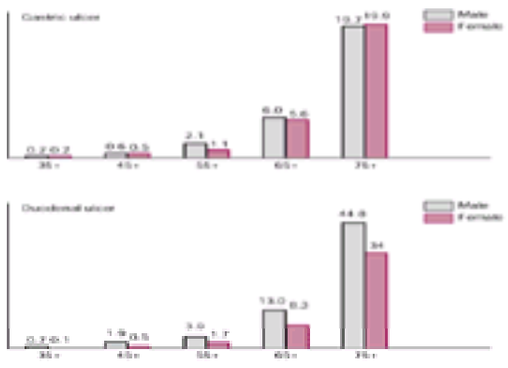

Fig. 9 (a) and (b) Age-specific death rates from peptic ulcer for England and Wales as rates per 100 000 per annum in 1998.

Treatments for peptic ulcer disease and non-ulcer dyspepsia

Antacids

Alkalis provide temporary symptomatic relief of pain in peptic ulcer disease but do not accelerate healing compared with placebo unless very large quantities are given. They are more useful in GORD and non-ulcer dyspepsia than in peptic ulcer disease. They usually contain the alkaline salts or hydroxides of magnesium or aluminium. Those based on magnesium are laxative and those based on aluminium tend to cause constipation. Some preparations contain mixtures of magnesium and aluminium to overcome this. Some preparations contain alginate which forms a flocculent raft on the gastric contents to reduce the effects of bile and acid reflux and protect the oesophageal mucosa.

Bismuth preparations

Tripotassium dicitratobismuthate (bismuth chelate, DeNol)

This bismuth complex acts mainly by suppressing *H. pylori* infection. It also has a protective effect on the mucosa, probably by stimulation of the synthesis of mucosal prostaglandins. Tripotassium dicitratobismuthate heals gastric and duodenal ulcers about as effectively as histamine H_2-receptor antagonists. It produces longer remissions of duodenal ulcer disease than do H_2 antagonists. This is probably because of its effect against *H. pylori* and because it remains in the gastric mucosa for several weeks after treatment. Currently tripotassium dicitratobismuthate is chiefly used as a component of second-line *Helicobacter* eradication regimens (see above). Bismuth preparations cause black stools—not to be confused with melaena. The bismuth content of tripotassium dicitratobismuthate is quite low. In the past 'epidemics' of bismuth encephalopathy occurred in Europe after patients ingested large amounts of other preparations containing unchelated bismuth. Tripotassium dicitratobismuthate only causes this complication if patients with impaired renal function take excessive doses for long periods.

Ranitidine bismuth citrate

This complex contains bismuth and the histamine H_2-receptor antagonist ranitidine. It can be used to treat duodenal ulceration associated with *H. pylori* infection. However, its main role is as a component of some highly effective *H. pylori*-eradication regimens (see above). It is contraindicated in moderate to severe renal failure and is not used for maintenance therapy since it may cause bismuth encephalopathy. It darkens the stools and can cause a black tongue.

Sucralfate

Sucralfate is a complex of aluminium hydroxide and sulphated sucrose, the properties of which include protection of the mucosa and ulcer healing. It is a weak antacid and may act by stimulating the synthesis of mucosal prostaglandins. It also binds bile salts. It is about as effective as histamine H_2-receptor antagonists at healing gastric and duodenal ulcers, but this role has now largely been taken over by proton pump inhibitors and *H. pylori* eradication. Currently sucralfate is mainly used in the prophylaxis of acute erosive gastritis in severely ill patients (see above). Side-effects are few but include mild constipation. It is to be avoided in patients with renal failure who occasionally develop aluminium toxicity.

Misoprostol

Misoprostol is a synthetic prostaglandin that increases the resistance of the stomach and duodenum to damage. Mechanisms include increased blood flow and secretion of mucus and bicarbonate. Misoprostol also mildly inhibits acid secretion. It accelerates healing of gastric and duodenal ulcers but this role has largely been taken over by proton pump inhibitors which are more effective, and by *H. pylori* eradication. The main use of misoprostol is in the prevention of gastroduodenal damage by NSAIDs, when these have to be given. The idea of replacing prostaglandins, the production of which is blocked by inhibitors of cyclo-oxygenase, is elegant, but this role has largely been taken over by proton pump inhibitors that are more effective and are about equally priced. The main side-effect of misoprostol is diarrhoea. This can be severe but is less likely if single doses do not exceed 200 µg. This is taken two to four times daily for prophylaxis. Note that misoprostol can induce abortion, so it should be avoided in women of childbearing age.

Histamine H_2-receptor antagonists

Secretion of gastric acid is normally stimulated by histamine, released from enterchromaffin-like cells in the gastric mucosa and acting on histamine H_2 receptors on parietal cells. Histamine H_2-receptor antagonists inhibit acid secretion by blocking this receptor. Sir James Black was awarded the Nobel prize for discovering this class of drug (as well as b-adrenergic antagonists). During the 1980s these drugs provided optimal treatment of peptic ulcer disease before being superceded by proton pump inhibitors and *H. pylori* eradication in the 1990s. However, they remain useful and are currently much less expensive. They are of use in non-ulcer dyspepsia and milder cases of gastro-oesophageal reflux disease (see Chapter 14.6). They protect the duodenum, but not the stomach, from ulceration due to NSAIDs. These drugs are not effective in haematemesis and melaena. Protection against acute erosive gastritis has not been shown consistently in trials. The four H_2 antagonists are cimetidine, ranitidine, nizatidine, and famotidine; but only the first two of these are widely used, largely due to the effects of marketing rather than to any differences in efficacy. Compared with ranitidine, cimetidine is less expensive, but it interacts with the metabolism of warfarin and anticonvulsants; it may cause mental confusion in the elderly.

Proton pump inhibitors

These drugs act by inhibiting the hydrogen/potassium adenosine triphosphatase enzyme system (H+/K+ATPase or 'proton pump') which pumps acid into the gastric lumen. They are prodrugs which are weak bases that are taken up into the acidic spaces in parietal cells, where the low pH creates a sulphenamide that is hydrophobic and therefore cannot diffuse away. The drug then forms disulphide bonds with cystine residues in the alpha chain of the proton pump, leading to its irreversible inactivation. Acid secretion is only restored when the cell synthesizes new proton pump protein. So these drugs are longer acting than H_2 antagonists even though they are cleared rapidly from the circulation. Once-daily dosing is satisfactory in most cases. Proton pump inhibitors are considerably more effective at preventing and healing gastro-oesophageal reflux disease than any other class of drug. In peptic ulcer disease their therapeutic advantage is smaller but significant. Their role in therapy is:

a. As a component of *H. pylori* eradication regimens. In this role they act by elevating the intragastric pH so that bacterial multiplication is encouraged and bacteriocidal antibiotics can act.
b. They accelerate healing of ulcers during and after *H. pylori* eradication. This is particularly relevant to chronic gastric ulcers that tend to be slow to heal because they are large.

c. They prevent NSAID-related ulcers if these agents have to be given. In this respect, proton pump inhibitors are more effective at protecting the stomach and duodenum than other classes of drug.
d. Omeprazole is the treatment of choice for the Zollinger–Ellison syndrome (see Chapter 14.8). Proton pump inhibitors are also highly effective in the treatment of non-ulcer dyspepsia, but most cases can be managed successfully with less expensive remedies.

They are currently the class of drug that is responsible for the greatest drug-expenditure in the United Kingdom and there is pressure to restrict their use. The main side-effects are diarrhoea and headache. The former may be less frequent with omeprazole than lansoprazole. Proton pump inhibitors can interact with the metabolism of other drugs including anticoagulants and anticonvulsants. Antacids and sucralfate impair absorption of lansoprazole.

Motility stimulants

Prokinetic drugs are not used to treat peptic ulcers, but are of use in non-ulcer dyspepsia as well as gastro-oesophageal reflux disease (see Chapter 14.6). Metoclopramide and domperidone are dopamine agonists that increase gastric emptying, small bowel transit, and the tone of the lower oesophageal sphincter. They can cause hyperprolactinaemia, leading to gynaecomastia, galactorrhoea, and diminished libido. Dystonias can occur, particularly in younger patients, and more often with metoclopramide than domperidone. Both are inexpensive. Cisapride is a motility stimulant that is believed to act by promoting release of acetylcholine in the gut wall. It has several side-effects and important drug interactions. In particular cisapride can cause serious arrhythmias and is contraindicated if the QT interval is prolonged. Concomitant medication with macrolide antibiotics increases the circulating concentrations of cisapride and the risk of arrhythmia.

Prevention and control

H. pylori infection will diminish worldwide if its transmission can be diminished by public health measures. Immunization has been shown to be effective in animal models but a human version is still some way off. Mass eradication is not a serious option due to the cost and the risk of generating antibiotic resistance. NSAID-associated ulceration is diminished by a policy of not using these drugs in non-inflammatory arthritis, and by using cyclo-oxygenase-2-selective drugs.

Special problems in pregnant women

Pregnancy decreases the frequency of peptic ulcers but gastro-oesophageal reflux is frequent. Treatment is largely with dietary and lifestyle changes, together with antacids or sucralfate. Persistent symptoms are treated with H_2-receptor antagonists. If symptoms continue and are severe despite these interventions, upper endoscopy and/or therapy with proton pump inhibitors may be considered during the second or third trimester. Misoprostol should be avoided because it can induce abortion.

Occupational, quality of life, and psychological aspects

Chronic peptic ulcer disease certainly impairs quality of life, but happily this can be improved either with acid-suppressing drugs or *H. pylori* eradication. There is no objective evidence to support the widespread belief that psychological stress predisposes to peptic ulcers. The myth is perpetuated by using the obsolete term 'stress ulcer' to describe acute erosive gastritis.

The need for further research

The main area of uncertainty is over how to apply our knowledge about *H. pylori* to the general population. It could be argued that whole populations should be tested and treated if found to be infected because the bacterium clearly plays a major causative role in important diseases including peptic ulcers and gastric cancer. Against this there is the cost, which would be considerable, and a significant risk of generating antibiotic-resistant strains of *H. pylori* and other bacteria. Some studies raise the possibility that *H. pylori* infection might protect against oesophageal diseases, but this is controversial and several studies do not support the idea. Long-term studies of the effects of eradication compared with no eradication in healthy infected patients would help to resolve this, but would be a major undertaking. At present no country is evaluating mass eradication, with the exception of Japan where *H. pylori* and gastric cancer are both unusually prevalent. Developing countries lack the resources to mount such a programme. General improvements in public health and hygiene in the west are diminishing the prevalence of *H. pylori* infection without the need for specific action against this bacterium. On the other hand it is quite usual to eradicate *H. pylori* from infected patients without ulcers, even though the overall conclusion from trials is that this does not improve symptoms. This may be irrational but it points to the difference between the science of medicine and its art. Informing the patient that a serious infection has been found and eradicated can have a very useful and reassuring effect and ultimately might even prevent late complications of persistent infection.

---

*It is with great regret that we report the death of John Calam during the preparation of this edition.

Further reading

Calam J (1996). *Clinician's guide to Helicobacter pylori*. Chapman and Hall Medical, London.

Farthing MG, Patchett SE, eds (1998). *Helicobacter* infection. *British Medical Bulletin* **54**, 1–263.

Hawkey CJ (199). COX-2 inhibitors. *The Lancet* **353**, 307–14.

Talley NJ, ed (1998). Dyspepsia. *Baillière's clinical gastroenterology*, vol. 12, pp 417–630. Baillière Tindall, London.

14.8 Hormones and the gastrointestinal tract

P. J. Hammond, S. R. Bloom, A. E. Bishop and J. M. Polak

Introduction

The discovery of secretin, the first recognized hormone, by Bayliss and Starling in 1902 marked the birth not only of gastrointestinal endocrinology, but of endocrinology itself. This was followed in 1905 by the identification of gastrin, but the technique of identifying hormones was, thereafter, more successfully applied to the study of secretions from the ductless glands, and gastrointestinal endocrinology languished for the next six decades. The determination of the amino acid structure of gastrin following its extraction from a solid tumour in 1964 marked a renewed interest in the field, and the introduction of techniques for large-scale chemical extraction and purification of gut peptides resulted in the discovery of further gut peptides. Most of the gut peptides, such as cholecystokinin and substance P, have been identified within the central and peripheral nervous systems, playing a neuromodulatory role in many organs. These neurocrine peptides are synthesized in nerve cells rather than endocrine cells in the gut and act locally as peptide neurotransmitters or neuromodulators.

The endocrine cells of the gastrointestinal tract are not grouped into anatomically distinct glands, like most endocrine cells, but are scattered through the length of the gastrointestinal tract. The principal role of gut peptides is in the integration of gastrointestinal function, and they regulate the actions of the epithelium, muscles, and nerves throughout the gastrointestinal tract. This local effect of peptides may be either autocrine, regulating the function of the cell secreting them, or paracrine, influencing the behaviour of neighbouring cells of different type. Thus somatostatin, originally identified as a hypothalamic inhibitor of growth hormone release, has been shown to have inhibitory effects in many different organ systems. It is locally released and its main mechanism of action is a direct one on neighbouring cells, for example to inhibit gastric acid and insulin secretion. In addition to altering gastrointestinal function many peptides, such as gastrin, secretin, and enteroglucagon, probably play an important paracrine role in controlling the growth and development of the gastrointestinal tract. In contrast, for most gut peptides there is little evidence that they act as true hormones in an endocrine fashion.

Two techniques have contributed to the increased understanding of gastrointestinal endocrinology. Molecular biology has helped identify members of peptide families by molecular cloning techniques, and has provided information about peptide processing, which has shown that different peptides may originate from a single common precursor. Sensitive peptide radioimmunoassay has allowed detection of gut peptides, which have very low concentrations in plasma and tissues. Furthermore, the specific peptide antibodies can be used for immunocytochemistry to demonstrate the cellular and neuronal localization of gut peptides (Fig. 1), and for immunoneutralization studies to elucidate the pathophysiological functions of gut peptides. Peptide localization is further defined by electron microscopy, which demonstrates specific peptide storage granules (Fig. 2), and *in situ* hybridization, which allows the sites of peptide synthesis to be identified. The most recent advance in gastrointestinal endocrinology has been the molecular characterization of hormone receptors by cloning techniques. This has demonstrated different receptors for the same ligand and provides an explanation for the diverse biological actions of many gut peptides in the same tissues. The development of agonists and antagonists to specific receptors will allow the physiological roles of the gut peptides to be fully characterized, and may be of therapeutic benefit in restoring normal gastrointestinal function in a disease.

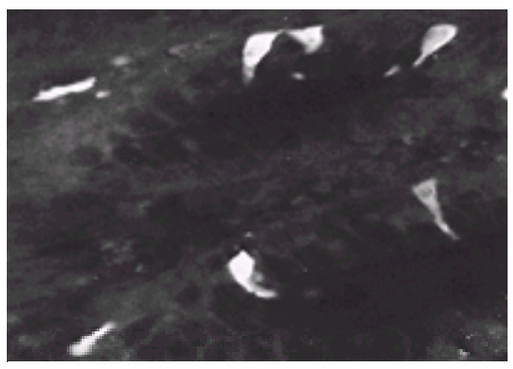

Fig. 1 Somatostatin cells, immunostained using the technique of indirect immunofluorescence, in the mucosa of human colon (×300).

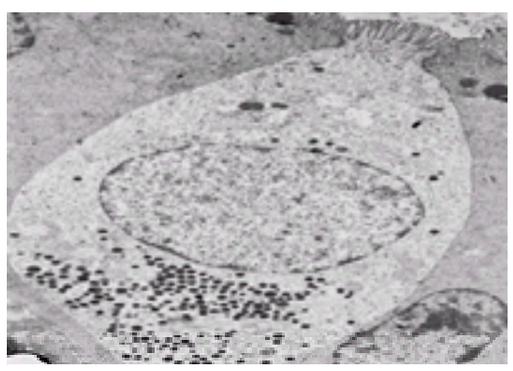

Fig. 2 Electron micrograph of a mucosal endocrine cell showing well-developed microvilli and secretory granules grouped at the basal membrane (×5500).

This section describes the gut peptide hormones and neurotransmitters, classifying them by common structure or precursor peptides, and then outlines abnormalities in gastrointestinal disease. The roles of gut peptides in the syndromes associated with gastroenteropancreatic tumours are considered in detail in Chapter 14.18.3.3, while the carcinoid syndrome is described at the end of this section.

Hormones and paracrine peptides

Gastrin–cholecystokinin family

Gastrin

Gastrin occurs in a variety of molecular forms but all the biological activity resides in the four carboxy-terminal amino acids. The major molecular forms contain 17 (G17; 2098 Da), 14 (G14; pentagastrin), and 34 (G34; big gastrin) amino acids. Larger molecular forms have been described but may be artefacts. In humans, gastrin is particularly localized to the gastric antrum, where G17 is the predominant form, but is also found in the upper small intestine, mainly as G34. These two are the predominant circulating forms. Gastrin is synthesized in G cells and electron microscopy shows gastrin granules to be large and electron lucent.

Gastrin release is particularly stimulated by protein ingestion and gastric distension. Its main physiological action is the stimulation of gastric acid secretion. Gastrin's other important physiological role appears to be its trophic effect on the gastric mucosa. Infusion of gastrin stimulates gastric motor activity and contraction of the lower oesophageal sphincter, but the physiological significance of this action is unclear.

Cholecystokinin

Cholecystokinin has an identical, five amino-acid, carboxy-terminal sequence to gastrin, but its specificity is conferred by the adjacent three amino acids, and this octapeptide confers its biological activity. It is found in the gut in 33, 39, or 58 amino-acid molecular forms predominantly, and is produced by the I cells of the duodenal and jejunal mucosa. The octapeptide cholecystokinin is a neurotransmitter in the central nervous system and a small amount is found in specific enteric neurones of the upper gastrointestinal tract.

Cholecystokinin secretion is stimulated by long-chain fatty acids and certain amino acids. The development of cholecystokinin antagonists specific for the two cholecystokinin receptor subtypes (cholecystokinin A, which is cholecystokinin specific, and cholecystokinin B, which appears to be also the only gastrin receptor) has allowed the important physiological roles of cholecystokinin to be characterized. The cholecystokinin A receptor appears to be involved in stimulation of gallbladder contraction and trophic effects on the duodenum and pancreas. The ability of cholecystokinin A receptor antagonists potently to inhibit meal-stimulated gallbladder contraction may be of therapeutic value in biliary colic.

The secretin family

The secretin family comprises a number of peptides with significant sequence homology. These include, in addition to secretin, glucose-dependent insulinotropic peptide, glucagon, enteroglucagon (see below), vasoactive intestinal peptide, peptide histidine methionine, and growth hormone-releasing factor. The last is released from the hypothalamus, mainly as a 44 amino-acid peptide, to stimulate release of growth hormone, but is also found in significant concentrations, mainly in a 40 amino-acid form, in the small intestinal mucosa, where its function is unknown.

Secretin

Secretin is a 27 amino-acid peptide (3056 Da), which appears to occur in only one molecular form, the whole molecule being needed for full biological activity. Circulating concentrations of secretin are lower than those of most other gut hormones. It is produced by S cells sparsely scattered throughout the duodenal and jejunal mucosa and is stored in characteristic secretory granules.

The main stimulus to secretin secretion is a duodenal pH of less than 4.5, although this occurs rarely. It is probably also secreted late after a meal, but the timing and quantities of this secretion are uncertain. The main physiological role of secretin is stimulating production of watery, alkaline pancreatic juices in response to acid in the duodenum. It may play an important part in the developing gastrointestinal tract, concentrations of secretin being particularly high in the early postnatal period.

Glucose-dependent insulinotropic peptide

Glucose-dependent insulinotropic peptide (**GIP**) is a 42 amino-acid peptide (5105 Da) with considerable sequence homology at the amino-terminal to secretin, glucagon, and vasoactive intestinal peptide. It is produced by K cells, predominantly in the upper small intestinal mucosa, but also in the gastric antrum and ileum, and is stored in large granules.

At pharmacological doses, GIP inhibits gastric secretions, and was originally named gastric inhibitory peptide. However, its physiological role appears to be as a component of the enteroinsular axis, being released in response to a mixed meal, particularly carbohydrates and long-chain fatty acids, and stimulating insulin release. This potentiation of insulin release in response to oral as opposed to intravenous glucose is the incretin effect. GIP has recently been implicated in the stimulation of cortisol release in two patients with ACTH-independent Cushing's syndrome whose serum cortisol rose postprandially.

Vasoactive intestinal peptide

Vasoactive intestinal peptide (**VIP**) is a 28 amino-acid peptide neurotransmitter (3326 Da) widely distributed through the central and peripheral nervous systems. The highest concentrations of VIP occur in the submucosa of the intestinal tract, where it is found in postganglionic intrinsic nerves (Fig. 3). VIP is a potent stimulator of small intestinal and colonic enterocyte secretion of water and electrolytes, acting via an elevation in cAMP. Other important actions include: smooth-muscle relaxation, both in the alimentary tract and in the systemic vasculature; stimulation of insulin release, counteracted by a direct glucagon-like effect of VIP in stimulating hepatic gluconeogenesis and glycogenolysis; stimulation of pancreatic bicarbonate secretion; and relaxation of the gallbladder, pyloric sphincter, and circular muscle of the small intestine with contraction of the longitudinal muscle. VIP inhibits release of gastric acid but not at physiological concentrations in humans.

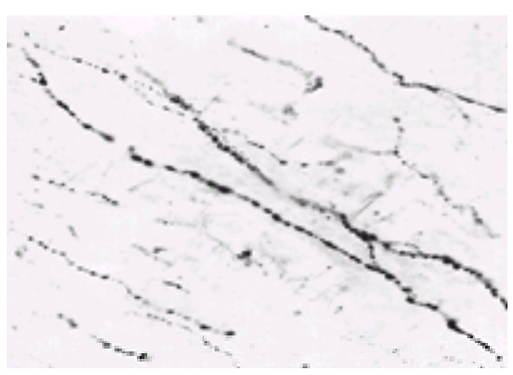

Fig. 3 Vasoactive intestinal polypeptide fibres, immunostained using the unlabelled antibody enzyme (PAP) method, in the submucosa of human colon (×500).

Peptide histidine methionine is a 27 amino-acid neuropeptide with considerable sequence homology to VIP and derived from the adjacent exon of the prepro-VIP gene. It mimics the actions of VIP, probably acting via the same receptor, but is less potent.

Pituitary adenylate cyclase-activating peptide is a recently identified peptide occurring in 27 and 38 amino-acid forms and with considerable sequence homology to VIP. It has a similar tissue distribution to VIP and shares the same receptor outside the central nervous system and pituitary gland. It has similar actions to VIP on intestinal secretion and motility.

Peptide products of preproglucagon

In the pancreas the major product of the preproglucagon molecule is pancreatic glucagon, but in the intestinal L cells preproglucagon is cleaved into enteroglucagon, a 69 amino-acid peptide containing the entire sequence of pancreatic glucagon, and the two glucagon-like peptides (GLP-1 $_{7–36}$ NH_2 and GLP-2; see below).

Enteroglucagon

Enteroglucagon (also termed glicentin) is found in high concentrations in the mucosa of the ileum, colon, and rectum. It is released after a mixed meal, particularly of carbohydrate and long-chain fatty acids. Pure enteroglucagon has not become available for infusion studies and so evidence for its physiological role remains circumstantial. The amount of enteroglucagon secreted is proportional to the amount of unabsorbed food entering the colon, and high enteroglucagon concentrations are found in conditions associated with loss of the small intestinal absorptive capacity. Thus it has been postulated that enteroglucagon has a trophic effect on the small intestinal mucosa and may be important in gut adaptation. Enteroglucagon is further cleaved by the L cells to produce oxyntomodulin, a 37 amino-acid peptide released into the circulation, which is a potent inhibitor of pentagastrin-stimulated gastric acid secretion.

Glucagon-like peptide 1

Glucagon-like peptide 1 (**GLP-1**) is a 36 amino-acid peptide, which is secreted in a cleaved form containing the 30 carboxy-terminal amino acids (GLP-1 $_{7–36}$ NH_2). It is a more potent stimulus to insulin secretion than GIP, and appears to be the most important incretin in humans. It also inhibits secretion of glucagon and potentiates release of somatostatin. Infusion of GLP-1$_{7–36}$ NH_2 greatly reduces insulin requirements following a meal in patients with type 1 or 2 diabetes, and this effect may have therapeutic potential.

Pancreatic polypeptide, neuropeptide Y, and peptide tyrosine tyrosine

Pancreatic polypeptide, neuropeptide Y, and peptide tyrosine tyrosine are peptides with structurally similar genes and propeptide molecules probably derived from a common ancestral gene.

Pancreatic polypeptide

Pancreatic polypeptide is a 36 amino-acid peptide (4226 Da) first isolated as a contaminant during the purification of insulin. It is produced by specific cells found at the periphery of the pancreatic islets, particularly those in the head of the pancreas, and scattered through the exocrine pancreas. Pancreatic polypeptide granules are small and electron dense.

Concentrations of pancreatic polypeptide rise dramatically after a meal, particularly one high in protein, and this is at least in part due to activation of cholinergic fibres from the vagus. At physiological plasma concentrations, this polypeptide inhibits pancreatic exocrine and biliary secretion, and these may represent its biological actions, although there are no obvious consequences of its deficiency or excess.

Neuropeptide Y

Neuropeptide Y is a 36 amino-acid peptide neurotransmitter, which is often colocalized with noradrenaline. It is found in both extrinsic adrenergic nerves to the myenteric plexus and in intrinsic nerves in the myenteric and submucosal plexi, and highest concentrations occur in the upper intestine and distal colon. It is a potent vasoconstrictor, inhibits intestinal secretion, and depresses colonic motility.

Peptide tyrosine tyrosine

Peptide tyrosine tyrosine (**PYY**) is a 36 amino-acid peptide found in endocrine cells of the ileum, colon, and rectum. It has a similar distribution to enteroglucagon, with which it is often colocalized. It is released after a meal, particularly one containing carbohydrates or long-chain fatty acids, and its main function appears to be to slow intestinal transit, allowing more time for absorption. Other actions include delaying gastric emptying, decreasing intestinal motility, and inhibiting gastric acid secretion.

Bombesin and the gastrin-releasing peptides

Bombesin is a 14 amino-acid peptide (1620 Da) initially isolated from amphibian skin. It was found to be a potent stimulator of gastrin, and hence of gastric acid secretion. Its mammalian counterparts have similar properties and so were named gastrin-releasing peptides. In humans, gastrin-releasing peptide is a 27 amino-acid peptide found in the gut in the intrinsic neurones of the myenteric and submucosal plexi, particularly in the stomach and pancreas. In addition to its effect on gastrin, it stimulates release of motilin and cholecystokinin, and pancreatic enzyme secretion. Gastrin-releasing peptide has been shown to be an autocrine growth factor for small-cell lung carcinomas, and probably has trophic effects on the developing gut.

Opioids

The opioid peptides leu- and met-enkephalin and dynorphin are widespread through the nerves of the myenteric and submucosal plexi of the gastrointestinal tract. Their principal actions appear to be inhibition of gastrointestinal secretion and increased smooth muscle contractility.

Tachykinins

Substance P is an 11 amino-acid peptide (1345 Da) whose existence was demonstrated in 1931 through its ability to cause smooth-muscle contraction and vasodilatation. A number of homologous peptides have now been characterized, and are collectively known as tachykinins, because of their rapid action. In humans there are two tachykinin genes, preprotachykinin A encoding substance P and neurokinin-a, and preprotachykinin B encoding neurokinin-b. These three tachykinins are localized to neurones in the myenteric and submucosal plexi throughout the gastrointestinal tract, with high concentrations in the duodenum and jejunum. Their principal effects are smooth-muscle contraction, vasodilatation, and inhibition of intestinal absorption.

Other gut peptides

Motilin

Motilin is a 22 amino-acid peptide (2700 Da) secreted by small intestinal M cells, whose density decreases from duodenum to ileum. The biological activity resides in the 9 amino-terminal amino acids. Peaks in motilin secretion coincide with initiation of the duodenal myoelectric complex, and so motilin appears to control the reflex motor activity of the small intestine, which occurs at approximately 2-hourly intervals in the fasted state, keeping the small intestine free of debris. Circulating amounts of motilin rise after a meal or drinking water and it may have a physiological role in accelerating gastric emptying and colonic transit. The macrolide antibiotics, such as erythromycin, are motilin-receptor agonists, hence their side-effects of diarrhoea and abdominal cramps.

Neurotensin

Neurotensin is a 13 amino-acid peptide (1673 Da) present throughout the central nervous system, and in enteric neurones and N cells of the ileal mucosa. It was originally isolated from bovine hypothalamus.

Plasma neurotensin concentrations rise after a meal, particularly those with a high fat content, and the rise is proportional to the size of the meal. At physiological doses, neurotensin inhibits gastric acid secretion and gastric emptying, and stimulates pancreatic exocrine and intestinal secretion. However, as with pancreatic polypeptide, there are no obvious consequences of neurotensin excess.

Somatostatin

Somatostatin was initially isolated from the hypothalamus as a 14 amino-acid peptide (1640 Da) that inhibited the release of growth hormone. It is widely distributed throughout the central and peripheral nervous system, and is found in a variety of endocrine tissues. In the gastrointestinal tract it occurs in 14 and 28 amino-acid forms. Somatostatin is secreted by specialized (D) cells distributed throughout the gut mucosa and on the inner rim of the pancreatic islets. D cells have all the characteristics of endocrine cells, but also possess axon-like basal elongations along which the peptide can be transported and secreted directly on to local cells. Five human somatostatin receptors have now been identified and cloned, the type 1 receptor predominating in the gastrointestinal tract. As gastrointestinal and other neuroendocrine tumours often possess high-density somatostatin receptors, scintigraphy with radiolabelled somatostatin analogues has been used for tumour localization (see Carcinoid syndrome below).

Somatostatin inhibits hormone release and blocks the response of the effector tissue, and inhibits a wide range of gastrointestinal functions (Table 1). Its acts principally as a paracrine factor or neurotransmitter, although small amounts of somatostatin are released into the plasma in response to physiological stimuli, including food ingestion, and so it may have an endocrine role.

Chromogranin-derived peptides

These structurally related, acidic proteins are present in the secretory granule matrix of neuroendocrine cells and are useful markers of normal and neoplastic neuroendocrine cells. To date, this family of proteins has been shown to consist of three molecules: chromogranin A, the first to be identified and also known as (parathyroid) secretory protein I; chromogranin B, or secretogranin I; and chromogranin C, or secretogranin II. It appears that the chromogranins have dual physiological roles: they may act in the processing of some regulatory peptides and prohormones. Their latter property was suspected when the primary structures of the three proteins were determined. All were found to contain multiple pairs of basic amino acids, forming sites for potential proteolytic cleavage. Chromogranin A gives rise to several peptides, including catestatin, chromostatin, vasostatin, and parastatin. Another derived peptide, pancreastatin was first characterized as a potent inhibitor of insulin release and later found in mucosal cells throughout the gut, where it is often co-stored with other peptides. It is released by gastrin from enterochromaffin-like cells of the gastric fundus, fitting with its action of enhancing meal-stimulated gastric acid secretion. The chromogranin B molecule yields, amongst other peptides, GAWK (from its first four amino acids: glycine, alanine, tryptophan, and lysine), a peptide distributed abundantly throughout the gut in both mucosal endocrine cells and intramural nerves. Chromogranins are proving to be of relevance to clinical medicine as plasma concentrations of chromogranins A and B can be used to determine the presence of a neuroendocrine tumour and as a means to monitor the efficacy of treatment.

Other peptide neurotransmitters

Calcitonin gene-related peptide is a 37 amino-acid peptide produced by alternative splicing of the calcitonin gene transcript. It is a widespread neurotransmitter and in the gut occurs in both extrinsic sensory nerves and intrinsic neurones. It inhibits gastric acid and pancreatic secretion, and causes relaxation of vascular smooth muscle.

Galanin is a 29 amino-acid peptide neurotransmitter isolated from porcine intestine. It is widely distributed in enteric nerve terminals and in nerves supplying the liver and pancreatic islets. Its main actions are inhibition of intestinal smooth-muscle contraction and inhibition of postprandial insulin release.

The potent vasoconstricting peptide endothelin has been demonstrated in the plexi of the gastrointestinal tract and in mucosal epithelial cells. However, its role in the regulation of gastrointestinal function is unknown.

Gut peptides in gastrointestinal disease

Gastric pathology

The most common cause of an elevated level of gastrin is achlorhydria, which may be the result of atrophic gastritis, pernicious anaemia or uraemia, or from iatrogenic causes such as the use of H_2-receptor antagonists or the proton-pump inhibitor omeprazole, or following vagotomy. The elevation in gastrin is a consequence of the loss of negative feedback on gastrin secretion by the low stomach pH. If the antrum is mistakenly retained after gastric surgery, this similarly removes the antral G cells from exposure to gastric acid and is associated with high gastrin concentrations. Achlorhydria-related hypergastrinaemia results in hyperplasia of the gastric histamine-producing enterochromaffin (**ECL**) cells. Atrophic gastritis in humans, and prolonged achlorhydria as a result of antisecretory therapy (for example, omeprazole) in rats, are associated with gastric carcinoid tumours, and these are thought to develop as a result of the direct trophic effect of gastrin on the ECL cells. Antisecretory therapy in humans has not been associated with the development of these tumours, but recommended therapeutic doses should not be exceeded and in patients on long-term therapy, hypergastrinaemia should be avoided.

Peptic ulcer disease is not usually associated with abnormalities in gut peptide secretion, although a decrease in somatostatin release in patients infected with *Helicobacter pylori* may influence the paracrine regulation of gastric function.

After gastrectomy or truncal vagotomy, patients may develop the dumping syndrome due to accelerated gastric emptying. In these individuals there is a marked increase in the postprandial rise of VIP, neurotensin, PYY, and enteroglucagon, and a decrease in the release of motilin. VIP and neurotensin may both contribute to the postprandial hypotension associated with dumping, but neurotensin may have a beneficial effect in slowing gastric transit. The long-acting somatostatin analogue octreotide, which inhibits release of these peptides and inhibits gastric emptying, is often a very effective treatment for this condition.

Malabsorption

Malabsorptive conditions are associated with a decrease in the amount of peptides produced in the affected region, and a compensatory elevation of other peptides, particularly those trophic peptides implicated in the bowel's adaptation to loss of absorptive surface, such as enteroglucagon.

Coeliac disease is an autoimmune disease resulting from dietary gluten sensitivity and it is associated with villous atrophy of the upper small intestine (see Chapter 14.9.3). The postprandial peptide response in patients with untreated coeliac disease shows greatly reduced secretion of GIP and secretin, which originate from the affected region of bowel. In contrast, there is marked elevation of enteroglucagon, neurotensin, and PYY (Fig. 4). The decrease in secretin and increase in PYY may be responsible for the reduced pancreatic exocrine and biliary secretion found in this condition. Enteroglucagon stimulates enterocyte turnover in the affected segment, despite the villous atrophy. It may have a trophic effect on the remaining small intestinal mucosa and delay gut transit time, and neurotensin may help to improve absorption by delaying gastric emptying. In tropical sprue, a postinfective malabsorptive state usually seen in travellers to Asia and Central and South America, a different profile of postprandial peptide release is seen. There is marked elevation in enteroglucagon and PYY, as in coeliac disease, but also in motilin secretion, whilst other peptides behave normally. After successful treatment of coeliac disease or tropical sprue, peptide responses return to normal.

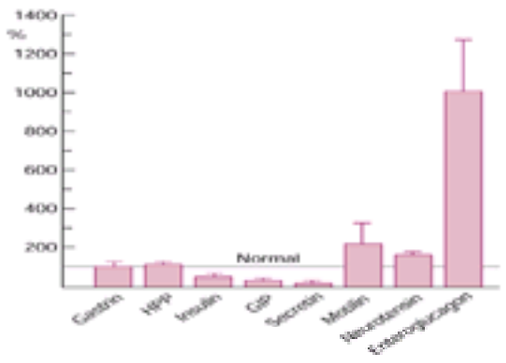

Fig. 4 The percentage incremental rise in gut hormones following a standard test breakfast in patients with coeliac disease compared with normal controls.

Chronic pancreatitis (see Chapter 14.18.3.2) results in varying degrees of pancreatic endocrine dysfunction in addition to the exocrine insufficiency. Thus patients often have insulin-dependent diabetes, and basal and arginine-stimulated glucagon concentrations may be reduced, although they are elevated in some individuals. Basal, and meal- and secretin-stimulated concentrations of pancreatic polypeptide are reduced if steatorrhoea is associated with chronic pancreatitis. The early loss of pancreatic polypeptide secretion in most individuals probably reflects the location of its secretory cells throughout the exocrine pancreas and on the periphery of the islets. However, secretion of pancreatic polypeptide is occasionally preserved and its concentration is not of diagnostic value in chronic pancreatitis.

Cystic fibrosis is often associated with diabetes mellitus and pancreatic exocrine insufficiency. Fasting and milk-stimulated concentrations of pancreatic polypeptide are usually suppressed. GIP concentrations fail to rise after a milk stimulus, implying a failure of the enteroinsular axis, and this may contribute to the associated glucose intolerance.

The malabsorption associated with pancreatic exocrine insufficiency of any cause leads to an excess of nutriments in the colon, and as a result the concentrations of enteroglucagon, PYY, and neurotensin are raised. The gut adaptation resulting from the effects of these peptides may contribute to the improvement in absorptive function with age in patients with cystic fibrosis.

Intestinal resection

Intestinal resection has profound effects on gut peptide concentrations. A jejunoileal bypass used to be constructed in patients with gross obesity. Peptide concentrations were normal preoperatively, but patients were hyperinsulinaemic and glucose intolerant. After the procedure there was an almost complete absence of the prandial GIP response and consequently a much reduced first-phase insulin response. The initial beneficial effects of the operation were ultimately negated by massive hypertrophy of the remaining bowel. The appearance of large volumes of undigested nutrients in the distal ileum is associated with a 16-fold increase in enteroglucagon responses and an eightfold increase in neurotensin secretion, and this may provide an explanation for the hypertrophy. After partial ileal resection, the concentrations of gastrin, enteroglucagon, pancreatic polypeptide, motilin, and PYY are elevated, but after colonic resection only gastrin and pancreatic polypeptide are raised, as there is a decrease in production of the other predominantly colonic peptides.

Diarrhoea

In acute infective diarrhoea, the concentrations of enteroglucagon, PYY, and motilin are increased, probably contributing to the altered gut motility and aiding mucosal repair. Patients with Crohn's disease have an elevated pancreatic polypeptide, GIP, motilin, and enteroglucagon, while in ulcerative colitis there is a modest elevation in pancreatic polypeptide, GIP, motilin, and gastrin, the last in response to the hypochlorhydria associated with the disease. Elevated levels of endothelin have been reported in ulcerative colitis and Crohn's disease and oral administration of an endothelin receptor antagonist in a model of colitis was found to ameliorate diarrhoea and tissue damage. No demonstrable abnormalities in gut peptides account for disordered motility in the irritable bowel syndrome.

Intestinal tumours

The trophic effects of gut peptides may contribute to proliferation of malignant gut tumours. In particular, colon carcinoma cells have receptors for a number of potentially mitogenic peptides, including gastrin, gastrin-releasing peptides, and VIP.

Neuropathic disease

In conditions associated with destruction of intrinsic enteric nerves there is loss of the neurocrine peptides found in the affected region. Chagas' disease (see Section 7) results from chronic infection with *Trypanosoma cruzi* and in the gastrointestinal tract can result in mega-oesophagus and megacolon. Concentrations of VIP and substance P and of their nerve fibres are greatly reduced in biopsies from affected segments. Similar changes are seen in the affected bowel from children with Hirschsprung's disease, which results from an aganglionic colonic segment. In contrast, neuropeptide Y-containing, mostly adrenergic, nerves are not reduced. Also, patients with the Shy–Drager syndrome, who have chronic autonomic failure with loss of preganglionic extrinsic nerves, have no abnormalities in neurocrine peptides or peptidergic nerve fibres on rectal biopsies (Fig. 5). Acquired immune deficiency disease is frequently accompanied by diarrhoea without evidence of secondary infection, and reduced immunostaining for substance P, VIP, and somatostatin in biopsies suggests a neuropathic process may be responsible. Alterations in neuroactive peptides have been observed in a number of inflammatory diseases. Increased density of VIP innervation has been reported in several gut diseases including reflux oesophagitis, radiation colitis, ulcerative colitis, and Crohn's disease. Calcitonin gene-related peptide has been shown to mediate the protective effect of sensory nerves in experimental colitis. It was recently reported that upregulation of the galanin-1 receptor is a mechanism for the increased colonic fluid secretion in infectious diarrhoea resulting from various pathogens.

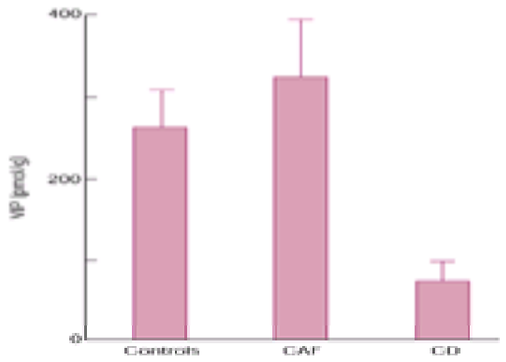

Fig. 5 Rectal vasoactive intestinal polypeptide concentrations (pmol/g wet tissue) in controls and patients with chronic autonomic failure (CAF) and Chagas' disease (CD) with gastrointestinal involvement. Reduced concentrations were seen in Chagas' specimens (reproduced from *Lancet*, 1980, **i**, 559, with permission).

Carcinoid syndrome

Introduction

The term *Karzinoide* was originally used by Obendorfer in 1907 to describe a carcinoma-like lesion without malignant qualities. It has now come to refer to tumours capable of producing serotonin (5-hydroxytryptamine; **5-HT**). However, several different cell types either synthesize or take up 5-HT and so the term carcinoid is

applied to a variety of malignant tumours with different biological behaviour grouped by their similar histological appearances. This section will focus primarily on those tumours associated with the classic carcinoid syndrome.

Primary gastrointestinal carcinoid tumours are derived from the embryonic foregut (thyroid, bronchus, stomach, common bile duct, and pancreas), midgut, or hindgut. The most common sites for carcinoid tumours are the appendix and rectum, but these tumours are often found incidentally on histological examinations of appendicectomy and rectal biopsy specimens. These tumours are almost always benign. Rectal neuroendocrine tumours generally produce glucagon-like peptides and PYY rather than 5-HT and are not usually associated with a clinical syndrome, even when they metastasize.

The carcinoid syndrome occurs in about 10 per cent of patients with carcinoid tumours. It does not develop when the tumour drains through a normal liver, and so midgut tumours have almost always metastasized, usually to the liver, before symptoms develop. The carcinoid syndrome is most commonly due to a metastatic midgut tumour, about 50 per cent of which metastasize to the liver. Primary carcinoid tumours are bronchial in origin in about 10 per cent of cases, and rarely occur in the ovary and testis. Tumours in these sites may be associated with the syndrome in the absence of metastases. The annual incidence of the carcinoid syndrome is about 1 in 500 000.

Clinical manifestations

The cardinal feature of the classic carcinoid syndrome is the flush. The carcinoid flush predominantly involves the head and upper thorax, and is usually associated with a tachycardia, hypotension, and increased skin temperature. Patients may have a sensation of intense heat and wheezing may occur. Rarely, flushing extends to the trunk and limbs, and may be associated with lacrimation, facial oedema, and great distress. Attacks are paroxysmal, and usually unprovoked, although precipitating factors include alcohol or food ingestion, stress, emotion, or exertion. Flushing initially lasts for only a few minutes, but as the disease progresses may become almost continuous, and such patients often develop a chronically reddened and cyanotic facial hue, with widespread telangiectasia, the leonine facies. This fixed flush is more commonly seen with bronchial carcinoids, which are often metabolically inactive, but when associated with flushing can cause severe attacks lasting for hours or days, occasionally with profound hypotension and even anuria. Gastric carcinoids are often associated with raised, localized, wheal-like areas of flushing, which are usually pruritic and may migrate.

The other characteristic feature of the syndrome is secretory diarrhoea, which may be profuse, with passage of several litres a day occasionally accompanied by electrolyte disturbance. It may be associated with cramping abdominal pain, nausea, and vomiting. Rarely these symptoms may result from small bowel obstruction from a large ileal carcinoid tumour, but the majority of primary tumours are small, usually being less than 1 per cent of total body tumour weight. Hepatic metastases may cause right hypochondrial pain, particularly if the liver capsule is involved or stretched, and acute exacerbations may occur if metastases become ischaemic and undergo autonecrosis. Weight loss and, in the later stages, cachexia are common as a result of poor dietary intake, malabsorption, and increased catabolism. Pellagra with dermatitis of sun-exposed areas may occur, the increased conversion of 5-hydroxytryptophan into 5-HT causing nicotinamide deficiency.

Cardiac valve abnormalities affect about 50 per cent of patients. They occur as a result of endocardial fibrosis, with plaques of smooth muscle in a collagenous stroma deposited on the valves. Lesions are almost always on the right-hand side, left-sided valve damage only occurring in association with bronchial carcinoids, which drain into the left atrium, or atrioseptal defects with right to left shunting. The most common lesions are tricuspid incompetence and pulmonary stenosis, and the usual clinical outcome is oedema and breathlessness due to right ventricular failure, which can be fatal. The other causes of breathlessness in association with the carcinoid syndrome are bronchospasm, which affects a small number of patients, often occurring with flushing attacks, and metastatic involvement of the lung and pleura. Arthritis occurs in a small number of patients, and sclerotic bone metastases may be seen, usually in association with foregut tumours.

Carcinoid tumours, in common with other gastroenteropancreatic tumours, have the potential to produce a variety of peptide products and may be associated with other syndromes, with or without the carcinoid syndrome. The most common of these associated syndromes is Cushing's, due to an ectopic ACTH-secreting, bronchial or pancreatic carcinoid. Carcinoid tumours may also be a feature of multiple endocrine neoplasia type 1 (see Chapter 12.10).

Biochemistry

The biologically active metabolite characteristically produced by metastatic carcinoid tumours is 5-HT, synthesized from the amino acid tryptophan (Fig. 6). 5-HT probably plays a part in the pathogenesis of some of the symptoms of the carcinoid syndrome, particularly the diarrhoea and bronchoconstriction. It is metabolized to 5-hydroxyindole acetic acid (**5-HIAA**), which accounts for 95 per cent of the urinary excretion of 5-HT.

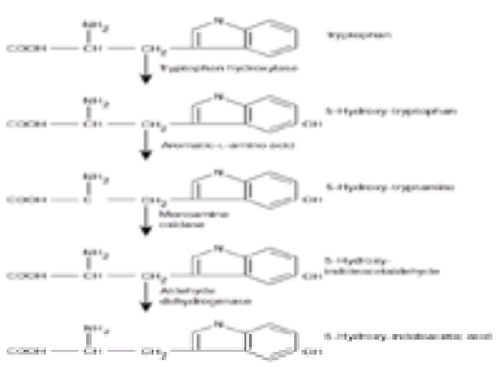

Fig. 6 Biochemical pathway for the synthesis and degradation of 5-hydroxytryptamine.

A variety of vasoactive substances may be secreted by carcinoid tumours and have been implicated in the pathogenesis of the flush. Flushing can be provoked by intravenous noradrenaline, which has been shown to activate kallikrein in the tumour, leading to synthesis and release of bradykinin. Other possible mediators of the flush include histamine, the tachykinins substance P and neurokinin A, and prostaglandins, although the flush is rarely affected by inhibitors of prostaglandin synthesis, such as indomethacin. Gastric carcinoids are derived from histamine-producing enterochromaffin cells and histamine is probably the cause of the characteristic wheal-like flush seen with gastric tumours.

Investigations

The diagnosis of carcinoid syndrome is made on the basis of elevated concentrations of 5-HIAA in a 24-h urine collection, and urinary 5-HIAA acts as a marker of disease progression. Various foods, including avocados, bananas, aubergines, pineapples, plums, and walnuts, should be avoided while collecting specimens to prevent false-positive results. A number of drugs and other substances interfere with the spectrophotometric assay: paracetamol, fluorouracil, methysergide, and caffeine give false-positive results, and ACTH, phenothiazines, methyldopa, monoamine oxidase inhibitors, and tricyclic antidepressants false-negatives. The other products of carcinoid tumours are not routinely assayed. Circulating markers of neuroendocrine tumours, such as pancreatic polypeptide and chromogranin, may corroborate the diagnosis, and other gut hormones are occasionally elevated in association with carcinoid tumours, most frequently gastrin.

Localization of carcinoid tumours is rarely a problem, as most have gross hepatic metastases, visible on computed tomographic (**CT**) scanning or abdominal ultrasonography, at the time of diagnosis. In those rare cases where the syndrome occurs in the absence of metastases, tumour localization may offer the prospect of cure. These tumours are unlikely to be in the gastrointestinal tract and so chest radiographs and CT scans of the chest and pelvis should be taken. The recently developed, indium-labelled, somatostatin analogue pentetreotide may prove valuable in localizing these tumours, although the resolution is only about 1 cm and bronchial carcinoids are frequently atypical and do not bear somatostatin receptors. An alternative method of isotopic localization is using 123- *m*-iodobenzylguanidine (MIBG) scanning, which may be equally effective. These scanning techniques can be useful in patients with metastatic disease to demonstrate the extent of spread (Fig. 7), particularly in those who are being considered for liver transplantation, which would be precluded by the presence of extrahepatic metastases. Angiography may be of value in assessing suitability for hepatic embolization. Carcinoid tumours have characteristic histological features being composed of regular polygonal cells arranged in nests. The capacity of 5-HT to reduce silver salts (so-called argentaffinity) led to the development of a diagnostic histochemical stain, but more

recently, immunostaining for serotonin has been used to identify the tumours.

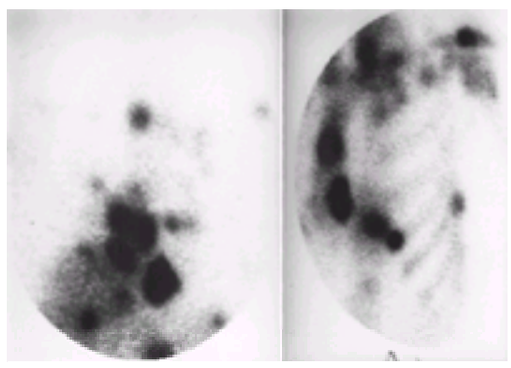

Fig. 7 Indium-111-labelled somatostatin analogue scan (left) in a patient with carcinoid syndrome showing metastases to liver, bone, and intrathoracic lymph nodes compared with a conventional base scan (right).

Treatment

The realistic aim of therapy in patients with the carcinoid syndrome is to relieve the symptoms. Simple treatments such as codeine phosphate, diphenoxylate, and loperamide may help to control the diarrhoea. Many of the symptoms can be controlled with the peripheral 5-HT antagonists: cyproheptadine, a 5-HT type 2 receptor blocker, often helps the diarrhoea; ketanserin may be effective in reducing flushing; and the 5-HT type 3 receptor antagonist ondansetron can alleviate nausea and anorexia. Parachlorphenylalanine, an inhibitor of tryptophan hydroxylase, and chlorpromazine block synthesis of 5-HT, but are rarely used. Histamine may mediate some of the features of the syndrome, especially in patients with gastric carcinoids, and in these cases H_1- and H_2-receptor blockade may be useful. With the exception of simple antidiarrhoeal agents, these treatments have been largely superseded by the long-acting, subcutaneously administered, somatostatin analogue octreotide. This inhibits the release of the mediators of the syndrome by the tumour and antagonizes their peripheral effects. Octreotide is effective in alleviating symptoms in over 90 per cent of patients. It is rarely associated with significant side-effects: the acidic solution can cause pain at the injection site; gallstones often develop, but are rarely of clinical significance; and a few patients develop steatorrhoea, which can be prevented by giving pancreatic enzyme supplements. Octreotide is now the first-line treatment for most patients and may be lifesaving in the carcinoid crisis, when symptoms become severe and continuous. In addition to octreotide, during crises patients usually need close monitoring of fluid and electrolytes, often by measurement of central venous pressure, and appropriate replacement therapy. Minor injury including overenthusiastic clinical examination of large carcinoid masses in the liver may induce a life-threatening state akin to the tumour lysis syndrome characterized by pain, fever, shock, and renal failure compounded by hyperuricaemia and hyperphosphataemia. Allopurinol, and the judicious use of infusions of sodium bicarbonate to make the urine more alkaline, as well as antimicrobials, may reduce the threat of renal failure in this condition. Corticosteroids may improve shock.

The principal disadvantage of octreotide is that patients develop resistance with time, and most become refractory to any form of treatment after about 4 years. Vitamin supplements containing nicotinamide are necessary when patients have pellagra, and these can be given prophylactically. The treatment of cardiac manifestations is the same as for valve disease and cardiac failure of other causes. Patients with painful bony metastases may benefit from palliative radiotherapy.

In patients who fail to respond or are intolerant to octreotide, tumour debulking may provide palliative relief. Surgery is rarely indicated, although enucleation of large metastases may give some benefit. Carcinoid tumours rarely respond to chemotherapy, either with streptozotocin and 5-fluorouracil, or with a variety of other agents, including cyclophosphamide and doxorubicin, although recent evidence suggests interferon-a may be of more benefit. The most effective means of debulking is hepatic embolization, which devascularizes the tumour while the blood supply to the normal liver is maintained by the portal vein. Octreotide should be given in high dose during this intervention, as the necrotic metastases release large quantities of vasoactive mediators that can cause a severe carcinoid crisis with profound hypotension, leading to acute renal failure (see tumour lysis syndrome above).

Prognosis

Carcinoid tumours behave like other gastroenteropancreatic tumours, with the majority following an indolent course. The median survival from the time of diagnosis is about 5 years, with a range of up to 20 years. Thus palliation is very worthwhile in these patients, allowing them to lead a normal life until the terminal stages of the disease.

Multiple endocrine neoplasia and non-diabetic pancreatic endocrine disorders are described in Chapter 12.10.

Further reading

Besterman HS *et al.* (1978). Gut hormone profile in coeliac disease. *Lancet* **i**, 785–8.

Bloom SR, Long RG, eds. (1982). *Radioimmunoassay of gut regulatory peptides*. Saunders, London.

Cook GC *et al.* (1979). Gut hormone responses in tropical malabsorption. *British Medical Journal* **i**, 1252–5.

Gorden P *et al.* (1989). NIH conference. Somatostatin and somatostatin analogue (SMS 201–995) in treatment of hormone-secreting tumors of the pituitary and gastrointestinal tract and non-neoplastic diseases of the gut. *Annals of Internal Medicine* **110**, 35–50.

Gronbech JE, Soreide O, Bergan ATI (1992). The role of resective surgery in the treatment of the carcinoid syndrome. *Scandinavian Journal of Gastroenterology* **27**, 433–7.

Gutniak M *et al.* (1992). Antidiabetogenic effect of glucagon-like peptide-1 (7–36) amide in normal subjects and patients with diabetes mellitus. *New England Journal of Medicine* **326**, 1316–22.

Hodgson HJ, Maton PN (1987). Carcinoid and neuroendocrine tumours of the liver. *Baillière's Clinical Gastroenterology* **1**, 35–61.

Jensen RT, ed. (1989). Gastrointestinal endocrinology. *Gastroenterology Clinics of North America* **18**, 671–931.

Kvols LK (1989). Therapy of the malignant carcinoid syndrome. *Endocrinology and Metabolism Clinics of North America* **18**, 557–68.

Lacroix A *et al.* (1992). Gastric inhibitory polypeptide-dependent cortisol hypersecretion—a new cause of Cushing's syndrome. *New England Journal of Medicine* **327**, 974–80.

Long RG *et al.* (1980). Neural and hormonal peptides in rectal biopsy specimens from patients with Chagas' disease and chronic autonomic failure. *Lancet* **ii**, 559–62.

Long RG, Adrian TE, Bloom SR (1981). Gastrointestinal hormones in pancreatic disease. In: Mitchell CJ, Kelleher J, eds. *Pancreatic diseases in clinical medicine*, pp 223–39. Pitman Medical, Tunbridge Wells.

Maton PN, Jensen RT (1992). Use of gut peptide receptor agonists and antagonists in gastrointestinal diseases. *Gastroenterology Clinics of North America* **21**, 551–664.

Moss SF *et al.* (1992). Effect of *Helicobacter pylori* on gastric somatostatin in duodenal ulcer disease. *Lancet* **ii**, 930–2.

Oberg K, ed. (1989). Neuroendocrine gut and pancreatic tumours. *Acta Oncologica* **28**, 301–449.

Oberg K, ed. (1991). Recent advances in diagnosis and treatment of neuroendocrine gut and pancreatic tumours. *Acta Oncologica* **28**, 301–449.

Reznik Y *et al.* (1992). Food-dependent Cushing's syndrome mediated by aberrant adrenal sensitivity to gastric inhibitory polypeptide. *New England Journal of Medicine* **327**, 981–6.

Thompson JC (1991). Humoral control of gut function. *American Journal of Surgery* **161**, 6–18.

Winkler H, Fischer-Colbrie R (1992). The chromogranins A and B: the first 25 years and future perspectives. *Neuroscience* **49**, 497–528.

Wynick D, Bloom SR (1991). The use of the long-acting somatostatin analog octreotide in the treatment of gut neuroendocrine tumors. *Journal of Clinical Endocrinology and Metabolism* **73**, 1–3.

14.9.1 Differential diagnosis and investigation of malabsorption

Julian R. F. Walters

Malabsorption by the gastrointestinal tract results in excess loss of dietary nutrients in the faeces and, if dietary intake does not increase to compensate, nutritional deficiency in the body. The normal physiology of nutrient absorption is complex—specific molecular mechanisms have evolved for each of the various types of nutrient. Understanding the principles involved in normal absorption enables different causes of malabsorption to be appreciated, appropriate differential diagnoses to be made, and investigations planned accordingly.

Principles of normal absorption

Absorptive capacity

For each of the classes of nutrients, the overall efficiency of intestinal absorption varies. Some compounds, such as components of dietary fibre, are not absorbed even in health. Others are normally almost completely absorbed, but in disease, absorption is insufficient to cope with the load, giving symptoms of diarrhoea from excess faecal water, or steatorrhoea from excess faecal fat.

The principal determinants of the maximum absorptive capacity are the area of the intestinal mucosa, increased by surface folding, villi, and microvilli to about 200 m^2, and the function of the individual cellular transporting mechanisms. As part of the total absorptive process, the intestine also has to reabsorb endogenous secretions produced to aid digestion. Approximately 7 litres of digestive fluids from salivary, gastric, biliary, pancreatic, and intestinal sources add significantly to the absorptive requirements for water, electrolytes, protein, and fat. Secretory diarrhoea and protein-losing enteropathy are conditions where endogenous output exceeds the absorptive capacity of the bowel.

Sites of absorption

Gastrointestinal motility mixes food with digestive secretions and propels them from the mouth to the anus. During this passage, nutrients are exposed to specialized areas of the gut with specific digestive or absorptive functions. The duodenum and proximal jejunum are mostly involved with digestion and fluid secretion. However, the more acidic pH in this area means the solubility and hence the absorption of polyvalent cations such as iron and calcium is high. The bulk of nutrient absorption takes place in the more distal jejunum and ileum. The terminal ileum is specialized for cobalamin (vitamin B_{12}) and bile salt absorption. The colon salvages fluid and electrolytes not absorbed by the small intestine and absorbs short-chain fatty acids produced by colonic bacteria from poorly digested carbohydrates. Loss of specialized areas by surgical resection or disease activity can produce specific patterns of malabsorption.

The intestinal epithelial cells differentiate as they move from crypt to villus tip. The older villus-tip enterocytes perform most of the absorptive functions, though some digestive enzymes are found in less mature cells. Fluid secretion probably occurs from the crypts. Goblet cells secrete mucus, trapping an unstirred water layer which is a relative barrier to the diffusion of large molecules but allows the smaller products of digestion to reach the surface of the epithelium. Other epithelial cells secrete various hormones or have immunological functions.

Mechanisms of absorption

Absorption occurs by transcellular and paracellular pathways. The paracellular pathway is through the tight junctions which link the epithelial cells. By this pathway, passive absorption of small molecules occurs by diffusion down electrical and concentration gradients. Solvent drag is the term used to describe movement down concentration gradients, which are themselves created by the movement of water. Active transport takes place through the epithelial cell against these gradients and necessitates the expenditure of energy generated within the cell.

Three steps are involved in transcellular absorption: entry to the cell at the apical (brush-border) membrane, passage through the cytoplasm, and exit from the cell at the basolateral membrane. Polarization of the enterocyte produces differences in structure and function of the apical and basolateral membranes. Specific carrier molecules are present in one of these membranes but not the other; this asymmetry generates vectorial flow in a single direction through the cell. The molecular basis for absorption of most types of nutrients has now been defined.

Diagnosis of malabsorption

The diagnosis of malabsorption is often missed until it is obvious. Diseases of the small intestine, colon, pancreas, liver, and stomach can all produce malabsorption; these may be obvious (such as the result of previous surgery) or may not be suspected until malabsorption is diagnosed.

History

Delayed growth and development, loss of weight, lassitude, and weakness may be described, but can be due to many other conditions. Changes in the nature of the stool or frequency of bowel habit suggest gastrointestinal disease but are not invariable, as apparently normal stools or constipation can also be found. In describing their faeces, patients may indicate the features of steatorrhoea, rather than watery diarrhoea. Careful questioning is needed to differentiate descriptions of changes in stool frequency or volume, and the passage of gas, liquid, oil, or grease. Bloating, borborygmi, and abdominal discomfort are often reported and seepage of oil from the anus may be described.

A previous history of abdominal surgery, radiation, and alcohol or drug usage may immediately make obvious the likely cause of malabsorption. A family history of coeliac disease or dermatitis herpetiformis makes gluten sensitivity more likely. Malabsorption of nutrients such as iron, folate and vitamin B_{12}, calcium, or vitamin K can give specific histories of anaemia, bone disease, and fractures, or bleeding and bruising.

Examination

General nutritional status can be assessed by height, weight, body mass index, and by anthropomorphic measurements such as skin fold thickness. Anaemia, bruising, petechias, ascites, oedema, glossitis, mucosal changes, and neuromuscular irritability (including positive Trousseau's or Chvosteck's signs) may be found and indicate specific deficiencies. Pigmentation and clubbing can occur. Abdominal distension, scars from previous surgery, masses, or fistulas can suggest specific diagnoses. The nature of the stools must be examined.

Evidence from routine investigations

Commonly obtained haematological and biochemical investigations can show a reduced haemoglobin concentration, microcytosis or macrocytosis, raised red cell distribution width, thrombocytopenia, and low serum iron, transferrin saturation, B_{12}, and folate. The prothrombin time or international normalized ratio (INR) may be raised. Albumin, calcium, phosphate, 25-hydroxyvitamin D, zinc, and other nutrients may be reduced. Elevated alkaline phosphatase and parathyroid hormone can suggest metabolic bone disease secondary to malabsorption.

Differential diagnosis

When malabsorption is suspected, two parallel diagnostic pathways need to be followed: first, to define the extent of the nutritional deficiency, and second, to define the cause of the malabsorption.

Generalized or isolated nutrient malabsorption?

When an isolated nutritional deficiency, such as that of B_{12} or iron, is identified, it must be remembered that this may be due to a more generalized process and evidence for malabsorption of other nutrients may be found if looked for. However, abnormalities in specific transport pathways, either genetic or acquired, account for some common types of malabsorption (Table 1).

Malnutrition, maldigestion, or malabsorption?

Evidence of deficiency of a nutrient does not necessarily imply malabsorption. In many cases, nutritional intake is impaired, or insufficient to meet increased demands. Pregnancy places additional demands on iron, calcium, and many other nutrients. Menorrhagia requires extra iron intake. Excessive loss of protein, electrolytes, and water may require increased intake, and additional calories are required in catabolic states such as infection, surgery, and critical care. Assessment of intake and requirements may determine that poor nutrition is the principal factor and dietary supplementation is required.

Impaired digestion, usually from pancreatic insufficiency, will produce a clinical state similar to malabsorption resulting from intestinal disease. Absorption of simple nutrients, as in an elemental diet, will be normal, but complex foods, particularly fats, will not be hydrolysed to forms that can be absorbed. Evidence of pancreatic disease should be looked for with imaging and function tests.

Conditions that cause malabsorption are listed in Table 2. Some such as coeliac disease are common, others such as the short bowel syndrome in patients with extensive surgery may be obvious, but other diagnoses may not be made unless specifically sought.

Investigation of malabsorption

Function tests

Tests to investigate absorptive functions are described in Chapter 14.2.4. These may be necessary to define the extent of nutrient malabsorption, but in most cases are not needed in routine clinical practice, where the emphasis is usually on defining the precise pathological cause of the malabsorption. The Schilling test for B_{12} malabsorption is particularly useful in diagnosing functional ileal disease.

Faecal fat measurements are often necessary to confirm that malabsorption (or maldigestion) needs to be investigated. The absorption of fat depends on a large number of different steps and is a sensitive indicator of malabsorption. Faecal collections are made over several days on a defined fat intake. Despite the unpleasantness of these collections and assays, this may be the only way of confirming that fat malabsorption is in fact occurring. A number of other tests have been developed in attempts to circumvent these problems.

The pancreolauryl test (see Chapter 14.2.4) is a simple way to look for impaired lipid digestion, and is sufficiently sensitive to diagnose pancreatic insufficiency when it is severe enough to produce steatorrhoea. False positives can occur in a range of intestinal conditions.

Xylose absorption tests have been used traditionally to help differentiate pancreatic maldigestion (when they are usually normal) from intestinal malabsorption such as coeliac disease (when they are usually abnormal). This test may give many false-positive results and has largely been superseded by tests aimed at detecting specific pathologies. Testing stools or urine for laxative abuse may be necessary, and endocrine causes of diarrhoea and malabsorption, although rare, should not be forgotten.

Serological tests for coeliac disease

IgA-class endomysial antibody serology is a highly sensitive and specific test which has simplified screening for coeliac disease in patients with any suggestion of possible malabsorption. These antibodies are detected by immunofluorescence on sections of monkey oesophagus or umbilical cord, and are a subclass of the previously used reticulin antibodies. False-negative results in coeliac disease occur in the presence of selective IgA deficiency (approximately 1 in 50 of the population), but IgG-class endomysial antibodies are detected. With effective treatment of coeliac disease with a gluten-free diet, endomysial antibodies become negative. Gliadin antibodies have much lower specificity and are not valuable in screening.

The antigen recognized by endomysial antibodies is now known to be tissue transglutaminase. More easily performed tests with this antigen, such as enzyme-linked immunoassays (ELISAs), will replace immunofluorescent testing for endomysial antibodies when they are confirmed to be as reliable.

The availability of simple blood tests, with high positive and negative predictive values for coeliac disease, means that this condition can now be strongly suspected before intestinal biopsy is performed.

Endoscopy and small bowel histology

Endoscopy is widely available and enables tissue to be taken from the small intestine to make a histological diagnosis of the pathology causing malabsorption. Oesophagogastroduodenoscopy allows biopsies from the upper duodenum. At colonoscopy, biopsies can be taken from the terminal ileum. Enteroscopy allows much more of the small bowel to be inspected and biopsies taken. Fortunately, common intestinal diseases such as coeliac disease are diffuse and diagnosis can usually be made from duodenal biopsies taken at routine upper endoscopy. Multiple biopsies reduce the likelihood of sampling error.

Biopsy of the more distal parts of the jejunum and ileum may be performed using the Crosby capsule or other similar designs. The capsule attached to a fine tube is swallowed allowing biopsies to be taken under fluoroscopic control, after which the capsule is recovered. Systems for multiple biopsies have been developed. These capsules are generally safe, but are used infrequently now that coeliac disease is usually diagnosed by endoscopic duodenal biopsies. Unfortunately, none of these methods allows reliable targeting to specific areas of small intestine seen on radiology. Laparotomy (or laparoscopy) may be needed to take full-thickness biopsies from jejunal and ileal lesions.

Mucosal appearances at endoscopy may be abnormal, suggesting histological diagnoses. With modern endoscopes, the resolution and magnification is such that villi can be detected. A smooth mucosa, with reduced folds of Kerckring, scalloped valvulae conniventes, pallor, and a mosaic appearance suggest villous atrophy, as in coeliac disease. Small white spots can indicate areas of intestinal lymphangiectasia. In the terminal ileum, ulceration can indicate Crohn's disease. Biopsies can be

directed to abnormal areas.

Histology will be definitive in most small bowel diseases. Villous atrophy with crypt hyperplasia is most frequently caused by coeliac disease. Tropical sprue, allergy to cows' milk protein in children, and a range of other conditions occasionally cause a similar picture. Intestinal lymphangiectasia, lymphoma, eosinophilic enteritis, Whipple's disease, amyloid, and abetalipoproteinaemia have characteristic appearances. Parasites including *Giardia* sp. may be seen.

Radiology

Plain abdominal films may show calcification in chronic pancreatitis, faecal loading, or abnormal gas-filled loops of bowel.

Contrast studies of the small intestine will define abnormal anatomy. Small bowel enema (enteroclysis) is preferred for showing mucosal detail, although barium follow-through studies are more easily tolerated and can give better images of proximal duodenum and terminal ileum. Mass lesions or strictures from Crohn's disease, tuberculosis, lymphoma, other tumours, fibrosis, radiation, ischaemia, or drug-induced injury will be demonstrated. Enteric fistulas and diverticula are diagnosed. Post-surgical anatomy can be defined and the length of remaining intestine in the short bowel syndrome estimated.

Endoscopic retrograde cholangiopancreatography (ERCP), and increasingly magnetic resonance cholangiopancreatography (MRCP), will show pancreatic abnormalities that can cause pancreatic exocrine insufficiency. CT, ultrasound, and angiography have roles in further defining conditions associated with malabsorption.

Microbiology

Small bowel bacterial overgrowth is a common and frequently undiagnosed cause of malabsorption. Absolute bacterial counts in proximal small-intestinal fluid are hard to obtain without contamination, so diagnosis tends to be on clinical suspicion confirmed indirectly via glucose hydrogen breath testing (see Chapter 14.2.4).

Infestations causing malabsorption (including *Giardia* and *Strongyloides* spp.) will be diagnosed by demonstrating parasites in fresh stool, or duodenal aspirates.

Response to treatment

Confirmation of the diagnosis of malabsorption is often made by assessing the response to treatment. Symptomatic patients with typical villous atrophy and positive antibody tests, who respond clinically and serologically to a gluten-free diet, can be confirmed to have coeliac disease and do not routinely need further biopsies. In small intestinal bacterial overgrowth, the diagnosis is often only finally made by the response to broad-spectrum antibiotics, which will also correct the abnormal Schilling test in this condition. Pancreatic exocrine insufficiency can be confirmed with a satisfactory response to enzyme replacements.

Further reading

American Gastroenterological Association (1999). American Gastroenterological Association medical position statement: guidelines for the evaluation and management of chronic diarrhea. *Gastroenterology* **116**, 1461–3.

Booth CC, Neale G, eds (1985). *Disorders of the small intestine*. Blackwell, Oxford.

British Society of Gastroenterology. Guidelines in gastroenterology: tests for malabsorption. http://www.bsg.org.uk/guidelines/27727.html

Walker-Smith JA *et al.* (1990). Revised criteria for diagnosis of coeliac disease. *Archives of Disease in Childhood* **65**, 909–11.

14.9.2Small bowel bacterial overgrowth

P. P. Toskes

Introduction

The occurrence of malabsorption in a patient with overgrowth of bacteria in the small intestine is known as small bowel bacterial overgrowth (**SBBO**).

Until recently, SBBO was typically associated with patients in whom disordered motility or structural abnormalities of the gastrointestinal tract were identified (Table 1). It is now recognized that SBBO occurs in a range of conditions associated with abnormal motility including gastroparesis, irritable bowel syndrome, and chronic pancreatitis—three conditions that now account for most patients in whom bacterial overgrowth is documented. Thus SBBO should be suspected in patients with these conditions whose symptoms prove resistant to conventional treatment or in whom steatorrhoea, weight loss, or flatulence develop unexpectedly. SBBO is now also recognized as the most important cause of malabsorption in elderly subjects, in whom a structurally intact small intestine becomes inhabited by colonic flora. Establishing a diagnosis of suspected SBBO by rigorous investigation is of key importance to its proper treatment.

Indigenous bacterial populations of the normal gastrointestinal tract

An understanding of SBBO is based upon a thorough knowledge of the indigenous bacterial populations of the normal gastrointestinal tract (Table 2). The proximal small intestine is normally inhabited by a few bacteria. Qualitative and quantitative changes appear at the ileum and become quite striking in the colon. Table 3 indicates the endogenous factors that prevent SBBO in humans. Of the factors listed, by far the two most important are the normal intestinal motility and an appropriate amount of gastric acid secretion. When either or both of these mechanisms are inhibited, SBBO may ensue.

Thus the stomach and proximal small intestine normally contain relatively few bacteria, which are usually lactobacilli, entercocci, Gram-positive aerobes, or facultative anaerobes present in concentrations of up to 10^4 viable organisms per gram of jejunal contents. Coliforms are rarely found in the healthy proximal small intestine. Anaerobic *Bacteroides* are not found in the proximal small intestine of a healthy gastrointestinal tract. The ileum represents a zone of transition from the sparse populations of the proximal small intestine and the very dense bacterial populations of the colon. In the colon the bacterial population increases up to one million times and reaches 10^9 to 10^{12} bacteria per gram of colonic content. The quality of the bacteria also change remarkably in the colon. Here there are fastidious anaerobic bacteria such as *Bacteroides*, anaerobic lactobacilli, and clostridia. These anaerobes outnumber the aerobic bacteria by as much as 10 000 to 1. The complexity of the colonic flora is such that more than 400 different species may be present in the colon of a single individual.

Bacteria normally metabolize bile acids, androgens and oestrogens, exogenous and endogenous cholesterol, unabsorbed dietary lipids, proteins, and carbohydrates as well as fibre, protein and urea, and other substances. The by-products of this metabolism may be of benefit or harm to the normal host. An exaggeration of this metabolism occurs in the presence of SBBO. It is noteworthy that the normal bacterial flora also are important in the metabolism of some drugs and other xenobiotics. Drugs metabolized by intestinal bacteria are listed in Table 4. The importance of the excessive metabolism of medications that may occur in SBBO is yet to be defined. Perhaps the relatively frequent occurrence of SBBO in the elderly may lead to ineffective medication of this age group.

Clinical conditions associated with SBBO

Table 5 lists the recognized clinical conditions associated with SBBO. In the past when much more gastrointestinal surgery was performed, the common causes of clinically significant bacterial overgrowth were structural abnormalities (for example Billroth II, anastomosis, surgery for Crohn's disease). Stagnant loops of intestine resulting from fistulas or surgical enterostomies and leading to SBBO were also common. Duodenal and jejunal diverticula can lead to SBBO, particularly if there is an associated hypo- or achlorhydria. Figure 1 depicts multiple duodenal and jejunal diverticula in a patient with cobalamin (vitamin B_{12}) malabsorption and steatorrhoea secondary to SBBO. Obstruction of the small intestine caused by Crohn's disease, adhesions, radiation damage, lymphoma, or tuberculosis may cause SBBO. Devastating malabsorption may occur secondary to SBBO associated with a gastrocolic or gastrojejunocolic fistula with colonic contents passing into the stomach or upper small intestine. SBBO may result from the ileal anal pouch procedure used to treat ulcerative colitis and adenomatous polyposis. The dysmotility syndrome, especially if combined with hypo- or achlorhydria, may lead to SBBO. Such motility disturbances include scleroderma, intestinal pseudo-obstruction, and diabetic autonomic neuropathy. Figure 2 demonstrates a diffusely dilated small intestine in a patient with diarrhoea, steatorrhoea, and scleroderma associated with SBBO. Antibiotic therapy completely reversed the abnormalities. Subjects with an absent or disordered migrating motor complex may develop SBBO. Such patients have no radiographic abnormalities and present with unexplained malabsorption. Elderly patients may develop malabsorption secondary to SBBO, and indeed it has been suggested that bacterial overgrowth may be the most common cause of clinically important malabsorption in the elderly. The elderly often have motor disorders (often induced by previous gastrointestinal tract surgery) and decreased acid secretion. The importance of both normal intestinal motility and normal gastric acid secretion in the prevention of clinically significant SBBO cannot be overemphasized. For example, patients with scleroderma and reflux oesophagitis who are well while receiving H_2 receptor antagonists may develop marked malabsorption manifested by diarrhoea and steatorrhoea after introduction of a proton pump inhibitor.

Fig. 1 Multiple duodenal and jejunal diverticula in a patient with cobalamin malabsorption and deficiency and steatorrhoea associated with SBBO.

Fig. 2 An upper gastrointestinal and small bowel series in a woman with scleroderma who presented with severe weight loss, cobalamin (vitamin B_{12}) deficiency and marked steatorrhoea. These abnormalities were corrected by broad spectrum antibiotics. Note the marked dilatation of intestinal segment was seen throughout the entire small bowel.

Up to 40 per cent of patients with chronic pancreatitis may have concomitant SBBO. The management of such patients may be quite problematic unless the clinician recognizes the need to treat both the pancreatic insufficiency and the SBBO. These patients with chronic pancreatitis may develop SBBO because of a decrease in intestinal motility resulting from pain, use of narcotics, inflammatory changes or obstruction from the large inflamed pancreas, or previous pancreatic surgery.

Several other clinical entities are associated with SBBO, but the pathogenesis is ill understood. These include endstage renal disease, cirrhosis, myotonic muscular dystrophy, fibromyalgia, chronic fatigue syndrome, and various immunodeficiency syndromes such as chronic lymphocytic leukaemia, immunoglobulin deficiencies, and selected T-cell deficiency.

Clinical manifestations of SBBO

No matter what the clinical condition leading to SBBO may be, there are several typical features as listed in Table 1. In addition, non-specific symptoms of nausea, bloating, abdominal distention, and abdominal pain may be the presenting symptoms of SBBO.

In suspected cases, a thorough evaluation is warranted; many of these patients may have had small bowel diverticula for years before suddenly developing marked symptoms as a result of SBBO. It may be that such patients needed to have a significant reduction in their gastric acid secretion before the structural abnormality could contribute to the SBBO. It is important to realize that SBBO may be superimposed on several clinical conditions whose initial symptoms are exactly those of SBBO. Patients with Crohn's disease, radiation enteritis, short bowel syndrome, or lymphoma may have superimposed SBBO. To what extent the malabsorption is the result of the primary intestinal disease or the consequence of SBBO is often difficult to determine. Weight loss associated with clinically apparent steatorrhoea has been observed in about one-third of patients with SBBO severe enough to cause cobalamin deficiency. Osteomalacia, vitamin K deficiency, night blindness, and even hypocalcaemic tetany as well as the vitamin E deficiency syndromes (neuropathy, retinopathy, T-cell abnormalities) may result.

Mechanisms of the metabolic abnormalities associated with SBBO

The malabsorption of nutrients associated with SBBO can largely be attributed to the abnormal intraluminal effects of the overgrowth flora combined with enterocyte injury induced by the overgrowth flora. A patchy small-intestinal mucosal lesion has been identified in experimental animals with SBBO and in human subjects with the condition. Steatorrhoea associated with SBBO results from bacterial alteration or bile salts, which leads to an impairment of micelle formation. In addition, accumulation of toxic concentrations of free bile acids may also contribute to the steatorrhoea by inducing a patchy intestinal mucosal lesion, thereby impairing the transport of fat. The predominant cause of the anaemia associated with SBBO is cobalamin deficiency. The anaemia is megaloblastic and serum cobalamin levels are low. Neurological changes indistinguishable from those of pernicious anaemia may ensue. The anaemia can be corrected by physiological doses of cobalamin. Cobalamin malabsorption that cannot be corrected by exogenous intrinsic factor is a characteristic of clinically significant SBBO. Competitive uptake of cobalamin, particularly by Gram-negative anaerobes, appears to be the mechanism responsible for the cobalamin malabsorption in SBBO. Iron deficiency may also occur in association with SBBO due to blood loss through the gastrointestinal tract, perhaps resulting from patchy ulceration. These patients may have blood detected on examination of their stools together with a microcytic and hypochromic anaemia. In some patients there may be two populations of red blood cells, microcytic and macrocytic. Folate deficiency is not a common occurrence in SBBO because the overgrowth flora synthesize folate and it is available for the host to utilize; serum folate levels may be elevated.

Hypoproteinaemia is frequent is SBBO and is occasionally severe enough to lead to oedema. Its causes are multifactorial but include decreased uptake of amino acids by a damaged small intestine, intraluminal breakdown of protein and protein precursors by bacteria, and protein-losing enteropathy.

A decrease in urinary xylose excretion is frequently seen in patients with SBBO. The primary reason for the decreased urinary xylose excretion is intraluminal degradation of the sugar by the overgrowth flora. Diarrhoea has many potential causes: the overgrowth flora may produce organic acids that increase osmolarity of the small intestine and decrease intraluminal pH. Furthermore, bacterial metabolites such as free bile acids, hydroxy fatty acids, and organic acids stimulate secretion of water and electrolytes into the lumen.

General diagnostic approach to patients suspected of having SBBO

SBBO should be considered in the differential diagnosis of any patient who presents with diarrhoea, steatorrhoea, weight loss, or macrocytic anaemia, particularly if the patient is elderly and has had previous abdominal surgery. A history of previous surgery for small intestinal obstruction should raise the question of whether the obstruction was bypassed by an end-to-side anastomosis, leaving a blind pouch, or side-to-side anastomosis, resulting in recirculation of the contents of the small intestine. The presence of dysphagia in a patient with malabsorption should suggest the diagnosis of scleroderma, and repeated bouts of intestinal obstruction without obvious organic cause should suggest intestinal pseudo-obstruction. Figure 3 suggests an algorithm for the evaluation of patients with malasorption including those with SBBO. This algorithm emphasizes the use of non-invasive, inexpensive tests. It also focuses on the fact that, in most medical centres, clinically important malabsorption is more likely to be due to pancreatic insufficiency or SBBO then coeliac disease or tropical sprue. When the history suggests that SBBO may be contributing to the malabsorption, further evaluation is necessary for optimal management. The presence of steatorrhoea should be documented. If the patient has clinically significant bacterial overgrowth, cobalamin absorption is frequently impaired, even though the patient may not yet have developed low levels of serum cobalamin. Intrinsic factor will not improve cobalamin absorption in these patients. The urinary excretion of xylose may be decreased and the serum folate level may be increased in some, but not all, patients with SBBO.

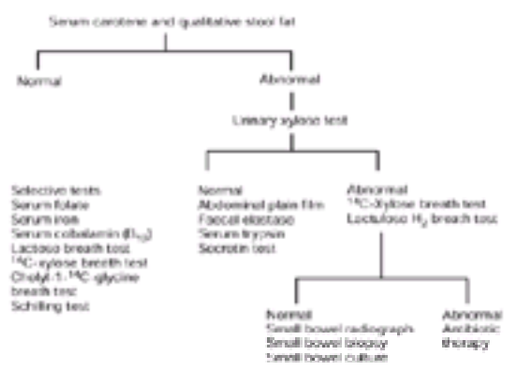

Fig. 3 Algorithm for evaluation of malabsorption. (From Toskes, 1992. Malabsorption. In: Wyngaarden JB, Smith LH, Bennett JC, eds. *Cecil's textbook of medicine*,

19th edn, pp 687–99. WB Saunders, Philadelphia, with permission.)

Specific diagnosis of SBBO

The definitive diagnosis of SBBO requires a properly collected and appropriately cultured aspirate from the proximal small intestine. The specimen should be obtained under anaerobic conditions, serially diluted, and cultured on several selected media. In patients with SBBO, the total concentration of bacteria generally exceeds 10^5 organisms per millilitre of jejunal secretions. *Bacteroides*, anaerobic lactobacilli, coliforms, and enterococci are all likely to be present in varying numbers. Although in most patients the intraluminal microbial proliferation can be documented in the proximal jejunum, it is important to recognize that pockets of overgrowth may be missed by a single culture and that bacterial overgrowth may occur in the more distal parts of the small intestine.

A properly collected and analysed intestinal culture requires intubation of the small intestine and it is both time-consuming and expensive. In many clinical practices today, small intestinal intubation for quantitative cultures to demonstrate SBBO is simply not performed. Consequently, a variety of surrogate tests for detecting SBBO have been devised based on the varied metabolic actions of the bacteria within the overgrowth flora. Table 6 lists these various tests and compares them in respect to sensitivity, specificity, and simplicity. Measurement of urinary excretion of indican, phenols, drug metabolites, and deconjugated para-amino benzoic acid suffer from a lack of sensitivity and specificity in distinguishing SBBO from other causes of malabsorption. The quantification of deconjugated bile acids and short-chain fatty acids in jejunal secretions requires an intubation of the intestine and thus is resisted by clinicians and patients for the same reasons that cultures of the intestine are not popular.

Another approach to diagnosing SBBO is the timed analysis of breath excretion of volitle metabolites produced by intraluminal bacteria. Both the measurement of expired, labelled carbon dioxide after oral administration of ^{14}C- or ^{13}C-labelled substrates, and breath hydrogen after administration of non-labelled fermentable substrate have been utilized.

The first breath test to be utilized clinically to detect SBBO was the bile acid or ^{14}C-cholyglycine breath test. This test, unfortunately, suffered from signfiicant false-negative and false-positive results. The bile acid breath test does not distinguish SBBO from ileal damage or resection with excessive breath $^{14}CO_2$ production resulting from bacterial deconjugation within the colon of the unabsorbed labelled bile salt. This was particularly problematic because SBBO may be superimposed on ileal damage in conditions such as Crohn's disease, lymphoma, and radiation enteritis. False-negative results have also been described with this test in 30 to 40 per cent of patients with culture-proven SBBO.

Studies in experimentally induced SBBO demonstrated that the overgrowth flora had the capacity to metabolize significant quantities of xylose and produce carbon dioxide. A 1-g ^{14}C-xylose breath test was developed and found to be sensitive and specific for detecting the presence of SBBO in patients with culture-proven SBBO. Xylose was chosen as a substrate because: (i) it is catabolized by Gram-negative aerobes which are always part of the overgrowth flora; (ii) it is predominantly absorbed in the proximal small intestine in contrast to the predominant ileal absorption of bile salts, leading to virtually 'no dumping' of xylose into the colon; and (iii) it is metabolized substantially less than other proximally absorbed substrates such as glucose. Elevated $^{14}CO_2$ levels appear in the breath of 85 per cent of patients with culture-proven SBBO within the first 60 min of the test, with the 30-min sample being the most reliable. Laboratories throughout the world have demonstrated the reliability of the ^{14}C-xylose breath test when compared with intestinal culture. In those studies that utilized intestinal culture as the gold standard and evaluated shorter sampling intervals, particularly the 30-min time point, the sensitivity and specificity approximated 90 per cent. Some recent studies have raised doubts as to the reliability of the ^{14}C-xylose breath test, but those studies evaluated patients with severe disorders of motility and it is quite possible that the xylose never left the stomach appropriately to come in contact with the overgrowth flora in the proximal small intestine. In addition, there must be an overgrowth of Gram-negative coliforms for the xylose test to be positive. In at least one of the recent studies failing to confirm the reliability of the ^{14}C-xylose breath test, the cultures also lacked Gram-negative coliforms. Others have suggested refinement of the ^{14}C-xylose breath test to include a transit marker for intestinal motility, which may enhance its specificity.

It is not recommended that ^{14}C-labelled xylose be used as a substrate in the diagnosis of SBBO in children or fertile women. Consequently, ^{13}C-labelled xylose has been developed and demonstrated to be an effective test in detecting SBBO. These ^{13}C-labelled substrates have been utilized in selected centres but are not yet in general use in clinical practice.

Breath hydrogen analysis allows a distinct separation of metabolic activities of the overgrowth flora from that of the human host because hydrogen is not produced to any significant extent in mammalian tissue. Excessive breath hydrogen production has been noted in patients with bacterial overgrowth after the administration of 50 to 80 g of glucose or 10 to 12 g of lactulose. A fasting elevation of breath hydrogen is an excellent test for detecting SBBO, but only about one-third of subjects with culture-proven SBBO will have elevated fasting levels of breath hydrogen.

There must be rigorous attention paid to methodological details when tilizing a hydrogen breath test. Certain foods that cause prolonged excretion of hydrogen must be avoided the night before the test, and 2 h must elapse after cigarette smoking or physical exercise sufficient to produce hyperventilation before taking the test. It is also recommended that a mouthwash be performed before testing to eliminate the possibility of an early hydrogen peak resulting from oral bacteria. Finally, strict interpretation criteria must be adopted. Even with careful attention to these details, the sensitivity and specificity of the hydrogen breath test is disappointing when used to detect SBBO. Recent studies have demonstrated that up to 27 per cent of normal subjects failed to show any rise in breath hydrogen following the administration of lactulose. The non-radioactive nature and the ease of performance of hydrogen breath tests make them quite attractive. However, many studies indicate that these tests have an unacceptable lack of sensitivity and specificity for clinical use and clinical laboratories do not pay attention to the critical details required for their proper conduct.

Therapeutic approach to the management of SBBO

The aim of therapy for SBBO is to correct when feasible the cause of the stasis, but surgery is often impractical (scleroderma, multiple diverticula, diabetes, intestinal pseudo-obstruction) and unacceptable to the patient. Thus, management of patients with SBBO is lifelong. Antibiotic therapy is the cornerstone of treatment and remarkable improvement in symptoms can be achieved in most patients. It is important to emphasize once again that SBBO may be a treatable component of the malabsorption seen in patients with conditions such as Crohn's disease, intestinal lymphoma, or radiation enteritis. The deterioration in absorption in such patients may not be caused by their primary disease process but by the associated overgrowth. Clinicians also must be aware that bacterial overgrowth may be present without causing any disease. Not all patients who have a pathological flora in the proximal small intestine develop clinically important symptoms. An abnormal breath test or a pathological culture must be put into perspective before therapeutic decisions are made.

It would seem attractive to select the appropriate antibiotic by evaluation of the sensitivity of the bacteria present in the small bowel lumen. However, this approach is very problematic because there are many different bacterial species present, often with very different antimicrobial sensitivities. Under such conditions, it may be extremely difficult to select the most appropriate agent on the basis of the sensitivity results. It is important to select an antibiotic that is effective against both aerobic and anaerobic enteric bacteria. Although most patients with clinically significant malabsorption secondary to SBBO have a flora that is largely overgrown with anaerobes, malabsorption associated predominantly with the overgrowth of Gram-negative aerobes also occurs.

Table 7 lists antimicrobial agents that have been effective in treating SBBO whether in controlled clinical trials or extensive clinical practice. Antibiotics whose activities are largely limited to anaerobes, such as metronidazole or clindamycin, are not usually effective as monotherapy. Antibiotics that are known to have poor activity against anaerobes should not be used in treating SBBO; such antibiotics include penicillin, ampicillin, the oral aminoglycosides, kanamycin, and neomycin. Historically, the treatment of first choice has been tetracycline, but recent experience in the United States suggests that up to 60 per cent of patients with SBBO do not respond to tetracycline largely because of *Bacteroides* resistance to this drug.

In most patients, a single course of therapy (7 to 10 days) markedly improves symptoms and the patient may remain symptom-free for months; in others, symptoms recur quickly and acceptable results can only be obtained with cyclic therapy (1 week out of every 4); and in still others, continuous therapy may be needed for 1 to 2 months. If the antimicrobial agent is effective there will be a resolution or marked diminution of symptoms within 1 week. Diarrhoea and steatorrhoea will decrease and cobalamin malabsorption will be corrected.

Prolonged antibiotic therapy poses potential clinical problems including diarrhoea, enterocolitis, patient intolerance, and bacterial resistance. A prokinetic agent that could help clear the small intestine of the overgrowth flora would be an attractive therapy. Experimental animal studies suggest that SBBO might be favorably influenced by prokinetic agents. There have been two small studies of these agents in patients with SBBO, one utilizing cisapride and one using octreotide. Both agents led to positive results in respect to SBBO following the prokinetic treatment. Another study utilizing octreotide and erythromycin in patients with scleroderma and SBBO attained positive responses following prokinetic therapy. Large controlled trials of prokinetic therapy in patients with SBBO have yet to be completed. Since the days of Metchkinoff, it has been thought that one could manipulate the intestinal flora by giving live 'probiotic' microbial supplements that would change the balance in the intestinal flora. Studies to date with probiotic therapy in subjects with SBBO have been disappointing. A recent placebo-controlled, randomized cross-over trial compared norfloxacin, amoxicillin–clavulanic acid, and *Saccharomyces boulardii* in 10 symptomatic patients with SBBO. Both antibiotic treatments led to significant decreases in symptoms and a substantial improvement in the results of hydrogen breath testing. The probiotic treatment with *S. boulardii* did not result in any improvement in these parameters.

Nutritional support is an important part of treatment of SBBO and may be needed despite attempts to control the bacterial overgrowth by antimicrobial agents because there may be irreversible damage to the enterocytes. Therefore, a lactose-free diet and substitution of a large proportion of dietary fat by medium-chain triglycerides may be necessary. Patients with cobalamin malabsorption should receive monthly injections of cobalamin (1000 µg). Deficiencies of other nutrients such as calcium and vitamin K should also be corrected.

Clinicians should have a low threshold for suspecting SBBO as a cause of malabsorption. The algorithm presented in Fig. 3 emphasizes the performance of simple, outpatient testing to pinpoint the cause of the malabsorption. Therefore, an appropriate attempt should be made to document whether SBBO is present or not. The consequences of SBBO can lead to serious malabsorption that results in clinically important deficiencies of several nutrients and, moreover, can be easily diagnosed and treated.

Further reading

Attar A *et al*. (1999). Antibiotic efficacy in small intestinal bacterial overgrowth-related chronic diarrhea: a cross-over, randomized trial. *Gastroenterology* **117**, 794–7.

Bishop WP (1997). Breath hydrogen testing for small bowel bacterial overgrowth—a lot of hot air? *Journal of Pediatric Gastroenterology and Nutrition* **25**, 245–9.

Bouhnik Y *et al*. (1999). Bacterial populations contaminating the upper gut in patients with small intestinal bacterial overgrowth syndrome. *American Journal of Gastroenterology* **94**, 1327–9.

Corazza GR *et al*. (1990). The diagnosis of small bowel bacterial overgrowth. *Gastroenterology* **98**, 302–5.

DeBoissieu D *et al*. (1996). Small-bowel bacterial overgrowth in children with chronic diarrhea, abdominal pain, or both. *Journal of Pediatrics* **128**, 203.

Fried M *et al*. (1996). Duodenal bacterial overgrowth during treatment with omeprazole in outpatients. *Gut* **35**, 23–7.

King CE, Toskes PP (1986). Comparison of the 1-gram [^{14}C]xylose, 10-gram lactulose-H_2, and 80-gram glucose-H_2 breath tests in patients with small intestine bacterial overgrowth. *Gastroenterology* **91**, 1447–51.

Saltsman J *et al*. (1994). Bacterial overgrowth without clinical malabsorption in elderly hypochlorhydric subjects. *Gastroenterology* **106**, 615–18.

Soudah H, Hasler W, Owyang C (1991). Effect of octreotide on intestinal motility and bacterial overgrowth in scleroderma. *New England Journal of Medicne* **325**, 1461–7.

Toskes P, Kumar A (1998). Enteric bacterial flora and bacterial overgrowth syndrome. In: Feldman M, Scharschmidt B, Sleisenger M, eds. *Sleisenger and Fordtran's gastrointestinal and liver disease*, 1523–35. WB Saunders, Philadelphia.

14.9.3Coeliac disease

D. P. Jewell

Definition

Coeliac disease is an inflammatory disorder of the small intestine induced by the prolamins of certain cereals, namely the gliadins of wheat, hordeins of barley, and secalins of rye. The inflammation is associated with loss of villous height and crypt hypertrophy and leads to malabsorption. The functional and histological abnormalities are reversed towards normal following exclusion of those cereals from the diet; they reappear on luminal challenge with the noxious prolamins.

History

Coeliac disease may well have been recognized in ancient times, as Aretaeus, the Cappadocian, wrote of the 'coeliac affection'. This was clearly a malabsorptive illness with steatorrhoea, affecting children and adults, but whether it represents a gluten-sensitive enteropathy is impossible to know. Samuel Gee of St Bartholomew's Hospital, London gave an excellent account of the disease in 1888 and concluded that 'if the patient is to be cured at all, it will be by means of a diet'. It was the Dutch paediatrician, W.K. Dicke, who finally recognized the role of wheat and his initial observations made in the 1930s were confirmed during the 'winter of starvation' in Holland in 1944. He noted that children with coeliac disease paradoxically improved as bread became virtually unobtainable. Using dietary challenge and faecal fat output as an indicator, Dicke together with J. H. van de Kamer and H. A. Weijers showed that it was gliadin, the alcohol-soluble component of gluten, that was the damaging substance. The demonstration of a flat intestinal mucosa by J. Paulley in 1954 and the development of a technique to take biopsies from the small intestine by Margot Shiner in 1956, who studied the histological findings with I. Doniach, as well as Rubin and colleagues in 1960, characterized the disease histologically and allowed easy and accurate diagnosis. The ability to follow the response of the mucosa to dietary manipulation by serial biopsy also allowed clinicians to demonstrate that coeliac disease in children was the same disease as idiopathic steatorrhoea in adults. Therefore, the disease is now referred to as coeliac disease or as a gluten-sensitive enteropathy.

Pathology

Coeliac disease affects the small intestine but the mucosal inflammation can vary in severity and in extent. Many patients have very mild proximal disease and the mucosal damage can be patchy. The characteristic, but not specific, feature is loss of villous height so that, under a dissecting microscope, the mucosa appears completely flat (Fig. 1). This is confirmed on histological examination (Fig. 2). The mucosa may be completely flat or there may be very short, broad villi—this appearance is often called subtotal villous atrophy. However, the total mucosal thickness (surface epithelium to muscularis mucosae) is usually normal or only slightly reduced because the crypts become elongated—usually referred to as crypt hypertrophy. The surface epithelial cells become flattened, the basal polarity of their nuclei is lost, and the microvilli of the brush border become short and irregular. This last change is revealed by electron microscopy.

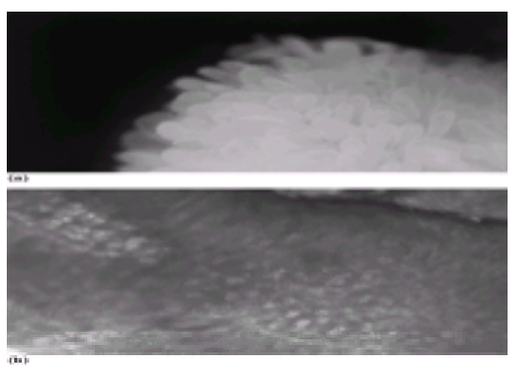

Fig. 1 (a) Dissecting microscopic appearance of a normal jejunal biopsy. (b) Dissecting microscopic appearance of coeliac disease.

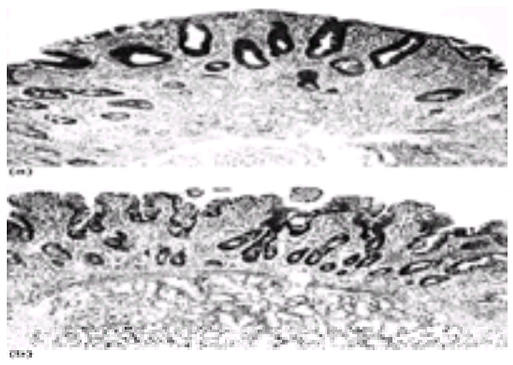

Fig. 2 Histological appearances of a distal duodenal biopsy in a patient with coeliac disease before (a) and after (b) a 3-month period on a gluten-free diet. Following treatment, there is much less inflammation and the villous pattern has begun to reappear.

Within the lamina propria, there is a marked infiltration of chronic inflammatory cells—plasma cells and lymphocytes. In addition, there is an increase in neutrophils, eosinophils, and mast cells. The proportion of intraepithelial lymphocytes is also increased in comparison with the number of enterocytes, although the absolute number is probably not increased.

Within the plasma cell population, there is an increase in the cells producing IgA, G, and M isotypes, although IgA cells still predominate. There is a marked increase in the proliferation rate of crypt cells, which, in a histological section, is shown by numerous mitoses at the base of the crypts. This increase in crypt cell proliferation is

thought to be mediated by cytokines released by the underlying lymphocytes and macrophages and leads to elongation of the crypts and loss of villous height.

It has been proposed that three stages in the development of coeliac disease are defined—infiltrative, hyperplastic, and destructive. These stages have been identified on the basis of challenge studies as well as biopsies of asymptomatic members of coeliac families. In the infiltrative stage, the epithelium becomes infiltrated with increased numbers of lymphocytes and this is the lesion frequently seen in patients with dermatitis herpetiformis. This stage leads on to the point where there is some inflammation of the lamina propria, with elongation of crypts. Both these stages are asymptomatic and can be regarded as latent coeliac disease. The destructive stage is the full lesion with loss of villi and a marked inflammatory infiltrate. Whether this classification is helpful, or indeed representative of what actually happens, is not yet clear.

As well as these characteristic changes in the small intestine, there may be a diffuse infiltration of other mucosal surfaces with lymphocytes and plasma cells. In particular, a proctitis has been recognized recently, but this is only detected if a rectal biopsy is taken and is virtually never severe enough to cause symptoms. This generalized mucosal infiltration presumably represents homing patterns of lymphocytes sensitized within the small-intestinal lamina propria.

Following a gluten-free diet, these histological changes return towards normal. For the majority of patients, a repeat biopsy after 3 months will show much less inflammation, the villous height will have increased, and the crypt elongation will have diminished. The mucosa usually returns to normal in children, but in adults, minor changes may persist with a crypt:villous ratio of 1:2 rather than the normal ratio of 1:4.

When patients in remission on a gluten-free diet are challenged with gluten, histological changes may be seen within a few days. In fact, electron microscopic changes may occur within a few hours of challenge and a fall in brush-border disaccharidase activity occurs in 24 h. However, some patients may take much longer to relapse and there have been some individuals who have virtually no histological change for up to a year.

Epidemiology

Coeliac disease is primarily a disease of Caucasians and, as it is closely associated with the extended haplotype HLA B8-DR3-DQ2, it is rare in those parts of the world where this haplotype is uncommon. Over 90 per cent of individuals with coeliac disease possess HLA DQ2, most of the remainder having HLA DQ8. The prevalence of coeliac disease in Europe and North America is about 1 in 300—much higher than was previously thought. This apparent increase has been due to more frequent diagnosis resulting from the accuracy of screening programmes that measure antiendomysial antibodies (see below).

There is little sex difference, although some studies have shown a preponderance of women. There is a familial incidence, with about 10 per cent of first-degree relatives being affected, and studies of monozygotic twins have shown a concordance of 70 per cent. These data indicate a strong genetic susceptibility.

Pathogenesis

There are two clear facts about the aetiopathogenesis of coeliac disease. The first is that fractions of gliadin, the alcohol-soluble component of gluten, are the toxic dietary constituent, together with similar fractions of rye and barley prolamins. The second is that there is a genetic susceptibility to gluten intolerance because of the close association with the HLA haplotype B8-DR3-DQ2 in northern Europeans and with B8-DR5/7-DQ2 in southern Europeans. This difference is due to the fact that the same DQ a–b heterodimer can be encoded on the same chromosome (*cis* position) in DR3 individuals or on opposite chromosomes (*trans* position) in DR5/DR7 individuals (see Section 5). This suggests that DQ molecules confer most of the susceptibility. What is not clear is how this particular haplotype interacts with gluten to produce mucosal inflammation in one person whereas the majority of people with this haplotype are able to ingest gluten with impunity. The other unsolved question is the exact nature of the toxic peptide.

An attractive hypothesis to explain why certain individuals lose oral tolerance to gluten is that it occurs as a result of an infection with adenovirus 12. This hypothesis was suggested by the observation that there was a dodecapeptide on the surface of a-gliadin which was similar to a peptide contained within an E1b protein of the virus. Many patients with coeliac disease demonstrate cellular and humoral immune responses to the virus, and so it is conceivable that the immune response to the virus cross-reacts with the gliadin peptide and thus induces intestinal inflammation. The gliadin dodecapeptide is now known to bind avidly to DQ2.

Some progress has been made in determining the toxic peptide, but in most studies toxicity has been assessed by *in vitro* methods rather than by direct feeding studies in patients. The amino acid sequences that seem to be shared by these toxic peptides are pro–ser–glu–glu or ser–pro–glu–glu. Two recent studies have identified an epitope (PQPQLPY) which appears to be dominant for the induction of a T-cell response. Furthermore, it requires the deamidation of a specific glutamine (at position 65) by tissue transglutaminase to elicit the response. B-cell responses also appear to depend on deamidation in this sequence. This is particularly interesting as tissue transglutaminase is now known to be the antigen to which the antiendomysial antibody is directed (see below).

The inflammatory lesion in the small intestine seems to result from an immunological reaction to the gluten peptides, with both cellular and humoral responses being involved. The mucosal abnormality is usually a proximal one and is presumably a reflection of luminal concentration of the relevant peptides. The release of cytokines and inflammatory mediators is thought to amplify the immune response and to influence epithelial stem-cell kinetics, with subsequent crypt elongation and loss of villous height. Malabsorption occurs because of loss of absorptive area and the presence of a population of immature surface epithelial cells whose absorptive and secretory function may be additionally impaired by cytokines and inflammatory mediators.

Clinical features

Coeliac disease in infants classically presents soon after weaning at the point that cereals are introduced. The babies usually fail to thrive, are miserable, refuse to eat, and lose weight. The abdomen becomes distended, there is muscle wasting, and they may have diarrhoea, which usually has the features of steatorrhoea. Abdominal pain and vomiting may be prominent symptoms and can mislead the clinician. Rectal prolapse may occur.

In older children, growth retardation is a common presentation and if the gastrointestinal symptoms are minimal, the diagnosis can be overlooked. Nutritional deficiencies can occur and may again be the reason for presentation, anaemia being the most common deficiency. Delayed puberty is another mode of presentation.

In adults, the most common presentations are anaemia and variable abdominal symptoms of discomfort, bloating, excess wind, and an altered bowel habit. Mouth ulcers are also frequent and can be the presenting symptom. The anaemia is most commonly due to iron deficiency and frequently occurs in the absence of intestinal symptoms; the macrocytic anaemias that sometimes occur in coeliac disease are described in Section 22. Many patients presenting with diarrhoea, wind, and abdominal pain are wrongly diagnosed as having an irritable bowel syndrome and there may then be a considerable delay before the true diagnosis is made. Patients suspected of having an irritable bowel syndrome should be specifically questioned about mouth ulcers and weight loss, either of which can be a pointer to organic disease. They should also be asked about feeding difficulties as a child, about growth milestones, and the age of achieving puberty. Less commonly, patients will present with a more typical history, with features of steatorrhoea, weight loss, bruising, and other symptoms of nutritional deficiencies resulting from malabsorption.

The classic findings in infants are those of an irritable child with stunted growth, muscle wasting, and a 'pot belly'. The infant usually has feeding difficulties and may show evidence of colic, and has a marked diarrhoea. However, in toddlers and older children who present with growth failure or anaemia, there may be few signs. Similarly, in adults, there may be no physical signs in those presenting with symptoms suggestive of an irritable bowel syndrome. Signs of iron deficiency may be present and there may be aphthous ulcers in the mouth, mild finger clubbing, and evidence of recent weight loss. It is very unusual, nowadays, for patients to show evidence of bleeding and osteomalacia (or rickets in children). Even less common are patients who are so malnourished that they have signs of ascites and hypoproteinaemic oedema.

Diagnosis

The crucial test to establish the diagnosis is a small-intestinal biopsy. This has traditionally been taken from the duodenal–jejunal junction (the ligament of Treitz) using a Crosby capsule. However, a distal duodenal biopsy taken at endoscopy is being used increasingly to make the diagnosis and comparative studies with a true jejunal biopsy have justified its use.

Several serological tests have been developed as screening tests which include antibodies to gliadin (IgA or IgG isotype), IgA antibodies to reticulin, and IgA antibodies to endomysium. The endomysial antibody has proved to be the most useful with a specificity and sensitivity of 90 to 95 per cent. The antibody is detected

by immunofluorescence using monkey oesophagus or human umbilical vein. The antibody is directed towards tissue transglutaminase and is increasingly being detected using enzyme-linked immunosorbent assay (ELISA) with transglutaminase-coated wells. The titre falls as the disease goes into remission on a gluten-free diet, but can take 3 to 12 months to become negative. Failure of the serum to become negative for the antibody, or reappearance of the antibody suggests non-compliance with the diet. Since it is an IgA antibody that is measured, a false-negative result may occur in the 5 per cent or so of patients with coeliac disease who have a coexistent IgA deficiency. The antiendomysial antibody is useful for screening high-risk populations, such as family members of patients with coeliac disease, diabetes, or osteoporosis.

Assessment of malabsorption

Careful documentation of nutritional deficiency as a result of malabsorption must be made and should include the following.

Full blood count

The haemoglobin level may be low, but the mean corpuscular volume may be low (iron deficiency), high (vitamin B_{12} or folate deficiency), or within the normal range. This can occur either because there is no significant deficiency of a haematinic or because there is a mixed deficiency, usually a combination of iron and folate. The red cell folate is a more reliable indicator of folate deficiency than the serum concentration and if it is low, there may be a pancytopenia. Vitamin B_{12} concentrations are only low in patients with extensive involvement of the small intestine and so are usually normal. Serum iron, total binding capacity, and ferritin concentrations should be measured to record the patient's iron status.

Biochemistry

Quantitative estimations of faecal fat excretion are becoming progressively more difficult to obtain despite their obvious value in assessing small-intestinal function. Qualitative assessment of excess fat by staining faecal smears with Sudan black or oil red O can be a reasonable alternative but merely records the presence or absence of a steatorrhoea. Fat malabsorption is inevitably accompanied by malabsorption of the fat-soluble vitamins A, D, E, and K. Serum concentrations of b-carotene, calcium, alkaline phosphatase, vitamin D, and the prothrombin time (INR—international normalized ratio) should therefore be assayed. Patients with diarrhoea may become hypokalaemic. Serum magnesium concentrations may also be low in severe coeliac disease and, with hypocalcaemia, can lead to tetany. Serum albumin is often low, as is the concentration of zinc.

Immunological tests

In addition to the titres of diagnostic antibodies discussed above, serum immunoglobulin concentrations should be measured. The most common pattern of abnormality is a raised IgA with a low IgM, but virtually any pattern may be seen. However, 5 per cent of patients with coeliac disease have an associated IgA deficiency, and loss of villous height and crypt hypertrophy frequently accompany common-variable acquired immunodeficiency.

Radiology

Barium radiology cannot give a positive diagnosis of coeliac disease and is not usually necessary unless it is required to exclude other small-intestinal diseases. In patients with mild disease the appearances may be normal, but if abnormalities are present, the appearances vary according to the radiological technique used. If a barium meal and follow-through is done, the small intestine may appear dilated and the barium often segments and flocculates. The proximal loops may appear smooth with a corresponding accentuation of the valvulae conniventes in the ileum—the so-called jejunization of the ileum. If a small-bowel enema is used (enteroclysis), the features are those of dilation and oedema of the valvulae conniventes.

Differential diagnosis

Few patients present with overt malabsorption, so the diagnosis of coeliac disease requires a high index of suspicion. However, once a biopsy is obtained that shows appearances compatible with coeliac disease, the differential diagnosis is limited.

For infants, the most common differential diagnosis is cow's milk allergy. An eosinophilia in the lamina propria as well as in peripheral blood is common, but this can also occur in coeliac disease. A soya milk allergy can also cause a flat small-intestinal mucosa. The precise diagnosis is usually dependent on a dietary history and the results of dietary exclusion.

In adults, infection with giardia, common-variable hypogammaglobulinaemia, lymphoma, Crohn's disease, and other small-intestinal diseases such as radiation enteritis, amyloid, and Whipple's disease may all show villous flattening and mucosal inflammation. Tropical sprue is usually associated with less marked changes—so-called partial villous atrophy—but has to be considered in patients who have spent time in endemic areas. Rarely, patients may be seen with a flat biopsy but with crypt hypoplasia—these do not respond to a gluten-free diet. Some patients with crypt hypoplasia also have a thickened band of subepithelial collagen, so-called collagenous sprue. Systemic diseases such as the vasculitides and systemic sclerosis may also be associated with an abnormal mucosal biopsy. Bacterial overgrowth of the small intestine may be associated with some mucosal inflammation and minor villous changes but they are rarely sufficiently severe to be confused with coeliac disease.

Dermatitis herpetiformis is commonly associated with an abnormal mucosal biopsy. The mucosal inflammation can be as severe as coeliac disease and responds to gluten withdrawal. The skin lesions also respond to a gluten-free diet, albeit slowly.

Non-specific infections of the small intestine may lead to a degree of malabsorption and mucosal inflammation. In only a small proportion can giardia be detected, and the precise aetiology of the majority is never determined. The illness is usually sudden in onset and gets better spontaneously over several weeks. It has been named 'temperate sprue'. Although this entity is uncommon, it can mislead clinicians. Most patients presenting in this way are started on a gluten-free diet as soon as the result of the biopsy is known and their improvement is regarded as a dietary response. Thus a patient presenting with a very short history of symptoms and whose mucosal biopsy suggests coeliac disease must be considered carefully. HLA typing and the presence of serum endomysial antibodies may be very helpful in making the correct diagnosis. If there is still doubt, the patient should be given a gluten-free diet but when the mucosa has recovered, a gluten challenge with subsequent biopsy should be undertaken.

Associated diseases

There is an increased prevalence of autoimmune diseases in patients with coeliac disease, especially those that are associated with the HLA B8-DR3 phenotype. These include diabetes, thyroid disease, and Addison's disease. Fibrosing alveolitis, systemic lupus erythematosus, and polyarteritis have also been reported. There may be an increase in epilepsy, especially temporal-lobe epilepsy, in patients with coeliac disease.

About 5 per cent of individuals with coeliac disease have an isolated IgA deficiency but the reason for and significance of this are not clear.

Treatment

Once the diagnosis is confirmed by a small-intestinal biopsy, patients should be started on a gluten-free diet. This diet should also exclude barley and rye, but with oats the need is less clear. As the evidence for oat toxicity is confused, many clinicians allow oats and only exclude them if the repeat biopsy does not show a good histological response. However, others exclude oats as the diet is begun and then consider reintroducing them once the disease has gone into histological remission.

All patients should be advised of the diet by a dietician and should be told to keep to it strictly. For children, the parents must be well briefed, including being told of the dangers of many sweets and 'fast foods'. For all patients, the diet is a lifelong necessity. Many children inevitably ingest small amounts of gluten during adolescence and many of them remain asymptomatic and develop normally. This has given rise to the highly erroneous view that children can 'grow out' of the disease. If the diagnosis was correct in the first place, then the patient has coeliac disease for life and if the diet is not strict, it may predispose them to complications in the future.

Patients should also be given details of the national coeliac society, if there is one. Most countries in which coeliac disease occurs have such a society. They provide a considerable amount of information about the disease and update patients about the gluten contents of new products appearing in the supermarkets. They also give invaluable advice about diet and foreign travel, as well as providing a social forum for patients and opportunities for fund raising to support research.

Nutritional supplements may be necessary at the start of treatment. If there are low serum concentrations of iron and folate, or biochemical evidence of osteomalacia, appropriate supplements are clearly required. However, once a gluten-free diet has begun, mucosal recovery occurs rapidly so that long-term supplementation is rarely necessary.

Patients with extensive mucosal damage are unable to digest lactose because of lactase deficiency. These patients may need a lactose-free as well as a gluten-free diet until there is histological recovery.

Once patients have been on the diet for 3 to 4 months, a further small-intestinal biopsy must be obtained to check for histological recovery. If the mucosa is still inflamed and villous height has not returned towards normal, a thorough review of the patient's diet is needed. Hidden sources of gluten (Communion wafers being a classic example) can often be found by a skilled dietician. If oats have not been excluded, then this is worth doing. In children, additional exclusion of soya products is occasionally needed.

Long-term follow-up, preferably in specialized clinics, is desirable but need only be on an annual basis once patients are stabilized on their diet. This allows patients to be seen by a dietician as well as their physician and reinforces the need to comply with a strict diet. Compliance should also be checked by measuring antiendomysial antibody—it should be negative if the disease is in remission.

Complications

The two major complications of coeliac disease are an ulcerative jejunoileitis and a T-cell lymphoma, and some investigators consider the jejunoileitis to be a manifestation of a lymphoma. They usually occur in middle age and usually present with weight loss, anaemia, abdominal pain, and diarrhoea. Thus any coeliac presenting with these symptoms having been previously well on a gluten-free diet must be carefully screened for these complications. Biopsies should be snap-frozen in liquid nitrogen to allow immunohistochemical analysis of T-cell markers and the detection of T-cell receptor rearrangements. The prognosis of a lymphoma complicating coeliac disease is poor.

In addition, there is a slight increase in the frequency of small-bowel carcinoma, although this is still very rare. There is also an increased incidence of other gastrointestinal cancers, especially oesophageal tumours, although the reasons for this increase are obscure.

There is some evidence suggesting that the patients who develop malignant disease, especially lymphoma, are those who have been poor compliers with the diet. Although this association is not absolutely proven, it provides the basis for continuing to advise a strict diet and to monitor compliance on a regular outpatient review.

Osteoporosis

Bone density scans have shown that most patients with coeliac disease have reduced bone density at diagnosis, but that this improves after a year or so of a strict gluten-free diet. The recommended intake of calcium for individuals with coeliac disease is 1500 mg daily, which is high, and requires skilled dietetic advice. For those patients with frank osteoporosis, treatment with calcium and vitamin D is required. For those with osteopenia, a follow-up scan should be done following a year on a gluten-free diet and calcium supplementation.

Prognosis

Provided that patients adhere to a strict diet, the prognosis is excellent and there are no data which suggest that there is an excess mortality in this group. Children develop normally and proceed into adolescence without delay. However, as mentioned above, compliance with the diet is often poor in adolescents and some clinicians still believe that many of them can have a more liberal diet if they are asymptomatic. It is this group that often present some years later with anaemia, mouth ulcers, or more serious evidence of malabsorption.

Unresponsive disease

A rare group of patients who are found to have a flat small-intestinal biopsy fail to respond to a gluten-free diet despite meticulous attention to avoid even minute amounts of gluten over many months. By definition these patients do not have coeliac disease, although they are often referred to as 'non-responsive coeliacs', a phrase that is misleading and should be avoided. Treatment of this group is difficult. Corticosteroids with or without azathioprine may help some, and more recent case reports suggest oral cyclosporin may also be of benefit. Excluding other dietary items such as soya can be tried, or an elemental diet that removes all dietary antigens. Some of these patients have a variety of central and peripheral neurological signs that do not fit classic vitamin-deficiency syndromes. The aetiology and pathology of these neurological lesions is unknown and the initial suggestions that they represented vitamin E deficiency have not been substantiated. The prognosis for these patients is poor because the neurological damage is slowly progressive and patients gradually lose weight despite full nutritional support.

It is always worth reviewing the intestinal biopsies in patients who do not respond to gluten withdrawal. Small-intestinal lymphoma, loss of villous height with crypt hypoplasia, and collagenous sprue are alternative diagnoses that may not have been recognized during the initial assessment.

Further reading

Anderson RP, Jewell DP (2001). Coeliac disease. In: Hunt R, Irvine J, eds. *Evidence based gastroenterology*, pp.307–22. B. C. Decker Inc., Ontario.

Marsh MN (1992). Gluten, major histocompatibility complex, and the small intestine. *Gastroenterology* **102**, 330–54.

Sategna-Guidetti C, Grosso S (1994). Changing pattern in adult coeliac disease: a 24-year survey. *European Journal of Gastroenterology and Hepatology* **6**, 15–19.

van Berge-Henegouwen GP, Mulder CJJ (1993). Pioneer in the gluten-free diet: Willem-Karel Dicke 1905–1962, over 50 years of gluten free diet. *Gut* **34**, 1473–5.

Van De Wal Y *et al.* (2000). Coeliac disease: it takes three to tango! Gut **46**(5), 734–7.

14.9.4Gastrointestinal lymphoma

P. G. Isaacson

Introduction

Gastrointestinal lymphomas, the most common extranodal lymphomas, are almost exclusively of non-Hodgkin's type, primary gastrointestinal Hodgkin's disease being extremely rare. A primary gastrointestinal lymphoma is defined as a lymphoma that has presented with the main bulk of disease in the gastrointestinal tract, with or without involvement of contiguous lymph nodes, necessitating direction of treatment to that site. The stomach is the commonest site of primary gastrointestinal lymphoma followed by the small intestine. Oesophageal and colorectal lymphomas are rare. The lymphomas that may arise in the gastrointestinal tract are listed in Table 1. Two of these, namely B-cell lymphoma of mucosa-associated lymphoid tissue (**MALT**) and enteropathy-associated T-cell lymphoma (**EATL**), do not arise in peripheral lymph nodes and will be discussed in more detail in this section. Any of the lymphomas that normally arise in lymph nodes may present as a primary gastrointestinal tumour, the most frequent being diffuse large B-cell lymphoma and mantle-cell lymphoma, which usually manifests in the gut as lymphomatous polyposis. Burkitt's lymphoma, which is the commonest childhood gastrointestinal lymphoma, is an especially common primary small intestinal lymphoma in the Middle East. The increasingly important group of B-cell lymphoproliferative conditions associated with immunodeficiency commonly present in the gastrointestinal tract, but they are more properly considered in the context of immunodeficiency-related lymphoproliferative conditions as a whole.

MALT lymphomas

The term MALT-lymphoma is used to designate a group of low-grade B-cell lymphomas whose histology recapitulates the features of mucosa-associated lymphoid tissue (MALT) as exemplified by the Peyer's patch. Paradoxically, there is usually no lymphoid tissue in the sites where MALT lymphomas occur, but lymphoid tissue of MALT-type accumulates prior to the development of lymphoma. In the stomach this is usually the result of chronic inflammation in response to *Helicobacter pylori* infection and its associated autoimmune phenomena. Intestinal MALT lymphomas are less frequent but include the entity known as immunoproliferative small intestine disease (**IPSID**) that has interesting parallels with gastric MALT lymphoma.

Gastric MALT lymphoma

Clinical presentation

Gastric MALT lymphoma typically occurs in patients over 40 years of age but can occur at any age. The sex incidence is equal. The presenting symptoms are usually those of non-specific dyspepsia and more suggestive of gastritis or peptic ulcer than a neoplastic lesion. Likewise, endoscopy more often shows inflamed, sometimes eroded mucosa than a tumour mass.

Pathology

Most MALT lymphomas of the stomach arise in the antrum, and macroscopically are characterized by an ill-defined thickened inflamed and ulcerated mucosa. The histological features closely simulate those of MALT (Fig. 1). Reactive non-neoplastic follicles are surrounded by the lymphomatous infiltrate in the region corresponding to the Peyer's patch marginal zone. The infiltrate extends into the surrounding tissue and invades individual gastric glands to form characteristic lymphoepithelial lesions (Fig. 2). Although the term centrocyte-like is most commonly used to describe the cells of MALT lymphoma, their cytological characteristics are more variable and they may more closely resemble small lymphocytes or show the features of so-called monocytoid B-cells. Scattered transformed blasts are usually present and plasma-cell differentiation, characteristically maximal beneath the surface epithelium, is present in one-third of cases. The lymphoma cells may specifically colonize the reactive follicle centres in a way that may lead to an appearance closely resembling follicular lymphoma. Gastric MALT lymphoma is characteristically multifocal.

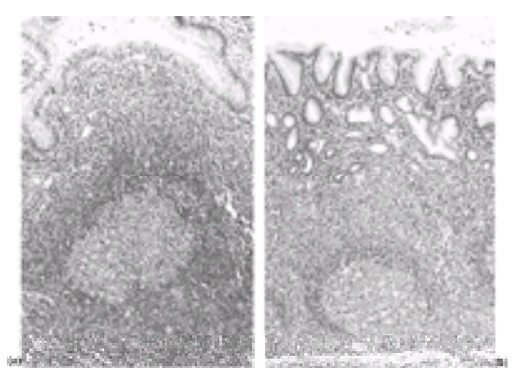

Fig. 1 (a) Peyer's patch comprising a B-cell follicle surrounded by a mantle zone external to which is the marginal zone. There are collections of small B lymphocytes within the dome epithelium. (b) Gastric MALT lymphoma. The tumour cells surround the reactive B-cell follicle in the marginal zone and invade gastric glands to form lymphoepithelial lesions. The overall structure is similar to the Peyer's patch.

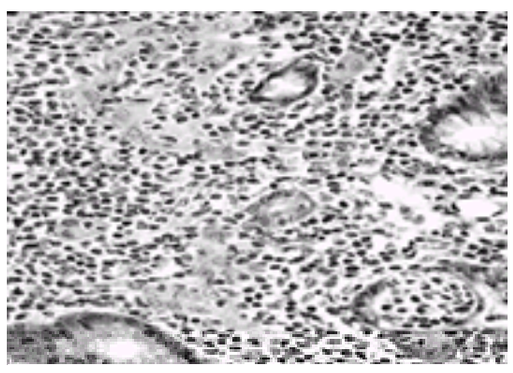

Fig. 2 Detail of the neoplastic infiltrate in a gastric MALT lymphoma showing 'centrocyte-like' cells invading gastric glands to form lymphoepithelial lesions (top left

and bottom right).

Biopsy appearances of gastric MALT lymphoma

There are many pitfalls in making the diagnosis of MALT lymphoma in small endoscopic biopsies. Amongst these are the retrieval of inadequate tissue, the presence of predominantly submucosal lymphoma, and the presence of cryptic foci of high-grade lymphoma or an associated adenocarcinoma. The differential diagnosis between MALT lymphoma and florid *H. pylori*-associated chronic gastritis (follicular gastritis) can be especially difficult. Molecular evidence of B-cell monoclonality is also helpful, but a diagnosis of lymphoma should never be made unless the histological criteria are fulfilled.

Dissemination to lymph nodes and other sites

Most gastric MALT lymphomas are at clinical stage 1_E at the time of diagnosis, but approximately 20 per ent have spread to the gastric lymph nodes or beyond. The more common distal sites include the small intestine, spleen, and bone marrow. In both lymph nodes and spleen the lymphomatous infiltrate tends to concentrate in the marginal zone.

The phenotype and genotype of gastric MALT lymphoma

The B-cells of MALT lymphoma express surface and, to a lesser extent, cytoplasmic immunoglobulin (usually IgM), which shows light-chain restriction. The cells express mature B-cell antigens including CD21 and CD35. They are CD5- and CD10-negative. This phenotype is homologous with that of marginal zone B cells, which are now acknowledged as the normal-cell counterpart. Various cytogenetic abnormalities have been described in MALT lymphomas, including trisomy 3, t(1;14) and t(11;18). The immunoglobulin (**Ig**) genes are mutated with ongoing mutations. Detection of monoclonal Ig gene rearrangement by Southern blotting or, more usually by the polymerase chain reaction (**PCR**), can assist in making the diagnosis of lymphoma in gastric biopsies, but caution is required since PCR evidence of monoclonality has been reported in biopsies from cases of florid *H. pylori*-associated gastritis.

High-grade transformation of gastric MALT lymphoma

Transformation of low- to high-grade MALT lymphoma is heralded by the emergence of increased numbers of transformed blast cells, which eventually form sheets or clusters (Fig. 3) and finally grow to confluence effacing any trace of the preceding low-grade tumour. This gives rise to difficulty both in grading some MALT lymphomas and in classifying large B-cell lymphomas of the stomach and other parts of the gastrointestinal tract. Those large-cell lymphomas in which no MALT component is evident are best classified as diffuse large B-cell lymphoma without reference to MALT.

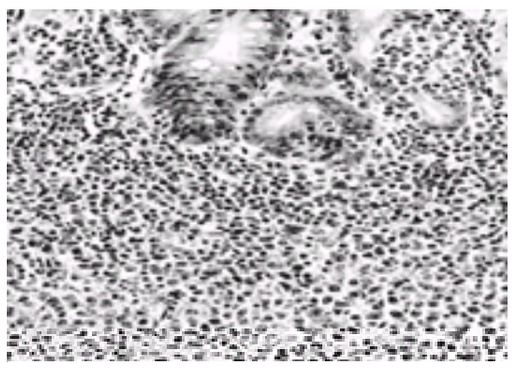

Fig. 3 MALT lymphoma showing transformation from low-grade (small cell) histology (upper half of figure) to high-grade (large cell) lymphoma (bottom half of figure).

The clinical behaviour of gastric MALT lymphoma

In comparison with nodal low-grade B-cell lymphomas, such as follicular lymphoma, which, at the time of diagnosis are characteristically at an advanced stage, MALT lymphoma is usually at stage I_E or II_E when diagnosed and is slow to disseminate. Hence low-grade MALT lymphomas respond favourably to therapy and there is an excellent overall survival approximating 90 per cent at 10 years. The survival for cases in which there is evidence of high-grade transformation is significantly worse, 45 per cent at 10 years.

Helicobacter pylori and gastric MALT lymphoma

There are several lines of evidence that implicate *H. pylori* in the pathogenesis of gastric MALT lymphoma. These include the fact that normal gastric mucosa is devoid of organized lymphoid tissue which, however, accumulates as a consequence of *H. pylori* infection, and the observation that the organism can be detected in most cases. The epidemiological study of Parsonnet *et al.*, which showed that there was a significantly higher frequency of preceding *H. pylori* infection in patients with gastric lymphoma compared to matched controls with non-gastric lymphoma, added further support to this association. The evidence became even more compelling following *in vitro* studies, which showed that the cells of low-grade gastric MALT lymphoma respond to *H. pylori* antigens via a T-cell mediated mechanism. The clinical significance of these findings was first shown by Wotherspoon *et al.* who described the regression of gastric MALT lymphoma in patients following eradication of *H. pylori* using appropriate antibiotics. Subsequent studies have shown that eradication of *H. pylori* may result in striking regression of the lymphoma in approximately 75 per cent of cases. Deeply invasive lymphomas, those in which there are foci of high-grade transformation, and cases with t(11;18) are unlikely to respond.

Immunoproliferative small intestinal disease (IPSID)

This condition is a subtype of MALT lymphoma, which occurs most commonly in the Middle East, although small numbers of cases have been reported from elsewhere. It is a disease of young adults and usually presents with severe malabsorption. The histology of IPSID is similar to that of gastric MALT lymphoma, except that plasma-cell differentiation is much more prominent both in the intestine and mesenteric lymph nodes. These plasma cells synthesize large amounts of alpha heavy chain without light chain, which can be detected in the serum. Hence the term 'alpha chain disease' which was first used for this condition. IPSID remains localized to the small intestine for prolonged periods and patients usually die from the severe malabsorption. High-grade transformation may also occur.

In its early stages, IPSID may be responsive to broad-spectrum antibiotics, which presumably eradicate bacterial spp. from the intestinal lumen. Thus, there is a remarkable parallel with the relationship between *H. pylori* and gastric lymphoma, although no specific organism has yet been implicated in IPSID.

Enteropathy-associated T-cell lymphoma (EATL)

An association between malabsorption and intestinal lymphoma has long been recognized, and at first it was thought that the lymphoma was in some way responsible for the malabsorption. It subsequently became clear that the reverse is true, and that intestinal lymphoma, in common with a variety of other tumours, was a complication of the malabsorption, which was most likely due to coeliac disease (gluten-sensitive enteropathy). In 1978, Isaacson and Wright characterized the lymphoma associated with malabsorption as a single entity, namely a variant of malignant histiocytosis. Later, Isaacson *et al*. showed that both the phenotype and genotype of this lymphoma were those of T cells rather than histiocytes, hence the term 'enteropathy-associated T-cell lymphoma' (EATL) was coined to describe the disease.

Clinical features

EATL is characteristically a disease with an equal sex incidence that occurs in the sixth and seventh decades of life, although sporadic cases have been described in younger patients. The commonest presentation is the sudden onset of abdominal symptoms, usually with the reappearance of steatorrhoea, after a short (months to years) history of successfully treated adult coeliac disease. Some of the patients may have first presented with dermatitis herpetiformis. In a minority of cases there is a well-documented history of childhood coeliac disease. The lymphoma may also present as an abdominal emergency with no history of malabsorption, the features of coeliac disease are found in the uninvolved portion of the resected small intestine. Abdominal pain, weight loss, fever, finger clubbing, and an ichthyotic rash are all common presenting symptoms and signs. The lymphoma usually results in intestinal perforation or haemorrhage rather than obstruction.

Jejunal biopsy in patients with EATL usually shows villous atrophy with crypt hyperplasia, but it may show only minor changes that, in some cases, may be limited to an increase in intraepithelial lymphocytes. It is unusual to obtain lymphoma tissue in the biopsy, although evidence of active or healed ulcers is sometimes present.

Most patients with EATL are subjected to a laparotomy. The lymphoma may involve any segment of the small intestine but is more common in the jejunum where is occurs as multiple nodules, ulcers, and strictures or, less frequently, as a large mass. The small intestine may appear normal, although there is usually considerable enlargement of mesenteric lymph nodes.

The clinical course of EATL is very unfavourable, except in a minority of cases where resection of a localized tumour has been followed by long remission. In most cases the lymphoma involves multiple segments of intestine rendering resection impossible, or has already disseminated beyond the mesenteric lymph nodes and out of the abdomen.

Pathology

EALT may involve any part of the small intestine and, rarely, other parts of the gastrointestinal tract including the colon and stomach, but most cases arise in the jejunum. The tumour is usually, but not always, multifocal and forms ulcerating nodules or large masses that may be accompanied by benign-appearing ulcers and strictures. The mesentery is often infiltrated and mesenteric lymph nodes are commonly involved. There is sometimes remarkably little macroscopic evidence of disease in the intestine in contrast to mesenteric lymphadenopathy.

The histological features of EATL show great variation both between cases and within any single case. The tumour cells may be only slightly larger than normal small lymphocytes or resemble immunoblasts but, more usually, are strikingly pleomorphic (Fig. 4). Intraepithelial tumour cells may be prominent. Interpretation of the histology is further complicated by the heavy inflammatory component, often containing many eosinophils, and extensive necrosis, which, together, may mask the neoplastic infiltrate (Fig. 5). Granulomas may be present and cause confusion with Crohn's disease. Non-specific 'benign' ulcers are frequently present in EATL and microscopically these show only chronic inflammation (see below).

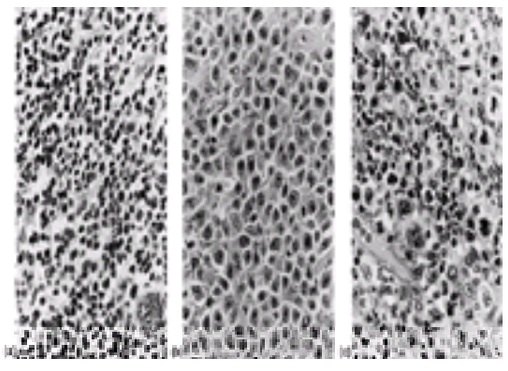

Fig. 4 Histological appearances of three different cases of EATL showing the cytological variability. In (a) the tumour is composed of small- to medium-sized lymphocytes; in (b) the tumour is composed of monomorphic, large immunoblasts; in (c) the tumour shows striking pleomorphism.

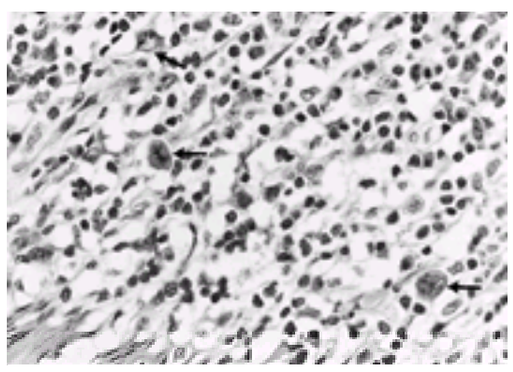

Fig. 5 A higher magnification of the ulcer base in which isolated malignant cells are evident (arrows).

The histology of the small intestine remote from the site of the tumour is an important consideration in the diagnosis of EATL. In most cases the changes are identical with those of coeliac disease with villous atrophy with crypt hyperplasia, plasmacytosis of the lamina propria, and an increase in intraepithelial lymphocytes. The degree of intraepithelial lymphocytosis may be spectacular and so extreme as to virtually obscure the epithelial cells (Fig. 6). The lymphocytes are small, without neoplastic features, and in these extreme cases spill into the lamina propria where they may merge with the lymphomatous infiltrate.

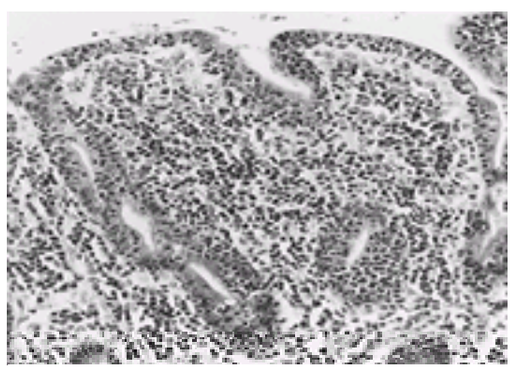

Fig. 6 Uninvolved mucosa from a case of EATL showing an extreme degree of intraepithelial lymphocytosis with spilling of lymphocytes into the lamina propria.

Episodes of ulceration followed by remission with healing may occur before the manifestation of EATL (so-called ulcerative jejunitis; see below). This can lead to a

confusing appearance of the mucosa with scarring and distortion of mucosal architecture, and the appearance of cells of the ulcer-associated cell lineage.

Lymph node involvement

The mesenteric lymph nodes are commonly involved, and almost always show accompanying hyperplasia that may mask the malignant cells which may be present in remarkably small numbers. Selective necrosis of lymph nodes, often involving entire nodes, remote from the main lesion is a feature of some cases. The cause of this necrosis is obscure.

Immunophenotype and genotype

In most cases the cells are CD3+, CD7+, CD4–, CD8-, CD56-, CD103+ and contain cytotoxic granules. Although the cells are CD4/8-negative they do not express the g/† T-cell receptor. Occasionally the lymphoma cells fail to express CD3 and may be CD8+; more rarely they are CD56+. These properties suggest that EATL arises from intraepithelial T cells. No characteristic genotypic features have been described.

EATL and coeliac disease

There is strong evidence that the enteropathy in EATL is a consequence of coeliac disease. Its histology and distribution are those of coeliac disease, and the HLA type of patients with EATL and coeliac disease are identical. Moreover, gluten sensitivity has been demonstrated in numerous EATL patients and a gluten-free diet has been shown to protect against the development of lymphoma. There remains the dilemma posed by those cases with minimal or even absent enteropathy. These patients are thought to suffer from the so-called latent coeliac disease, which can be confirmed by the finding of a positive endomysial antibody test diagnostic of coeliac disease.

Refractory coeliac disease, chronic ulcerative jejunitis, and EATL

Patients with established coeliac disease who become resistant to a gluten-free diet are said to have developed refractory coeliac disease (refractory sprue). Patients with this disorder may present with gluten-resistant malabsorption *de novo*, and in these cases the diagnosis can be substantiated by the finding of endomysial antibodies. A subgroup of patients with refractory coeliac disease develops multiple intestinal ulcers, and this syndrome has been termed chronic ulcerative jejunitis. Patients with refractory coeliac disease, particularly those with ulcerative jejunitis, may progress to develop EATL. It has recently been shown that the intraepithelial lymphocytes in patients with refractory coeliac disease and ulcerative jejunitis comprise a monoclonal population with an abnormal immunophenotype (cCD3+, CD4–, CD8–) similar to that of EATL. Moreover, when these patients develop EATL there is clonal identity between the intraepithelial lymphocytes and the subsequent lymphoma. Interestingly, the intraepithelial lymphocytes in the 'non-lymphomatous' small intestinal mucosa of EATL likewise share the abnormal phenotype and clonal identity of the lymphoma. Thus, patients with refractory coeliac disease can be said to be suffering from a clonal T-cell disorder that is directly linked to EATL. The optimum treatment for this condition remains to be clarified.

Management

The treatment of EATL is most satisfactory in those cases with a localized tumour, when surgical excision may be followed by long remission or even cure. Most cases are multifocal or have already disseminated at diagnosis and require treatment appropriate for a high-grade, non-Hodgkin's lymphoma, which may include bone marrow autografting. This form of therapy is particularly hazardous in EATL because of the danger of intestinal perforation. Some cases of ulcerative jejunitis, even when small foci of lymphoma are present, may respond to steroids.

Further reading

MALT lymphoma

Akbulut H, *et al*. (1997). Five-year results of the treatment of 23 patients with immunoproliferative small intestinal disease: a Turkish experience. *Cancer* **80,** 8–14.

Isaacson PG, Norton AJ (1994). *Extranodal lymphomas.* Churchill Livingstone, Edinburgh.

Isaacson PG, Spencer J. (1987). Malignant lymphoma of mucosa associated lymphoid tissue. *Histopathology* **11**, 445–62.

Parsonnet J, *et al*. (1994). *Helicobacter pylori* infection and gastric lymphoma. *New England Journal of Medicine* **330**, 1267–71.

EATL

Bagdi E, *et al*. (1999). Mucosal intra-epithelial lymphocytes in enteropathy-associated T-cell lymphoma, ulcerative jejunitis, and refractory celiac disease constitute a neoplastic population. *Blood* **94**, 260–4.

Wright DH (1997). Enteropathy-associated T-cell lymphoma. In: Wotherspoon AC, ed. *Lymphoma, cancer surveys*, Vol. 30, pp. 249–61. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

14.9.5 Disaccharidase deficiency

T. M. Cox

Disaccharidases are specific glycosidases that are required for the complete assimilation of nearly all dietary carbohydrate apart from free glucose and fructose. The enzymes are found on the luminal surface of the small gut; their activity may be reduced by genetically determined deficiencies or acquired by generalized disease of the intestinal mucosa. Disaccharidase deficiency causes a characteristic syndrome of carbohydrate intolerance.

Physiology of carbohydrate digestion Fig. 1)

Free disaccharides occur in the diet or are derived from the luminal hydrolysis of starch and glycogen by salivary and pancreatic a-amylase. Because amylase cannot hydrolyse the a-1,6 branching linkages and has little specificity for a-1,4 bonds adjacent to these points, the initial products of starch digestion are branched oligosaccharides containing at least one a-1,6 bond. Maltase-glucoamylase is a mucosal a-glucosidase that removes glucose moieties sequentially from the non-reducing terminus of linear oligosaccharides. a-Dextrinase (isomaltase) continues the hydrolysis of branched carbohydrate polymers by cleaving the a-1,6 glycosidic bonds of the limit dextrins that remain. a-Dextrinase is a component of the bifunctional enzyme complex, sucrase–isomaltase, the sucrase moiety of which hydrolyses sucrose into fructose and glucose. The disaccharides sucrose, lactose, and trehalose, like the a-dextrins, are poorly absorbed: to be assimilated, they are also split into monosaccharides by glycosidases located on the brush-border membrane (sucrase, lactase, and trehalase). Mucosal disaccharidases are optimally active at pH 6.0 and are present principally in the duodenum and jejunum—activity also persists in the ileum but is absent in the colon.

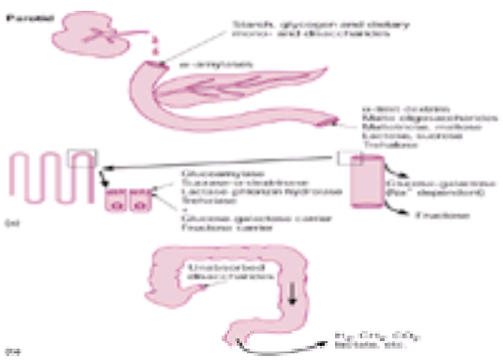

Fig. 1 Carbohydrate digestion and absorption.

Specific carriers in the microvilli for the transport of glucose and galactose, as well as fructose, mediate the uptake of monosaccharides released by the mucosal glycosidases—and absorption occurs rapidly. Active transport by the sodium-dependent glucose–galactose carrier is accompanied by the passive flux of water from the lumen. Maldigestion of osmotically active sugars thus leads to retention of fluid in the gut. For most carbohydrates, hydrolysis in the lumen and at the mucosal surface is sufficiently rapid to saturate the pathways for glucose and fructose transport. For lactose, however, the rate of hydrolysis, rather than glucose and galactose uptake by the mucosa, may become limiting. Hence the functional reserve of lactase in the human intestine is restricted and assimilation of lactose is often impaired in the early stages of mucosal disease.

Although the biosynthesis of surface disaccharidases continues throughout the life of the epithelium, the enzymes are only active in mature cells on the upper reaches of small-intestinal villi. Complete turnover of the enzyme molecules occurs several times during the lifespan of the mature enterocyte. Brush-border disaccharidases are complex glycoproteins that undergo proteolytic processing; extensive glycan modification in the Golgi apparatus occurs before insertion into the membrane. The mature enzymes are derived from large, single-chain polypeptides. The genetically determined mechanism by which lactase expression is normally reduced after infancy is not fully understood but in most individuals it appears to regulate transcriptional activity of the lactase gene. Unabsorbed carbohydrate resulting from the maldigestion of disaccharides is fermented by bacteria in the colon to short-chain organic acids, hydrogen, and methane. In these circumstances, ingestion of carbohydrate may cause pain by distension of the bowel with fluid and gas, accompanied by an irritant and watery diarrhoea.

Carbohydrate intolerance syndrome Table 1)

Abdominal symptoms are usually noticed within an hour of the ingestion of foods containing the offending sugars. There is nausea, bloating, and distension of the abdomen accompanied by borborygmi and flatulence. Colicky pain precedes a watery diarrhoea, usually associated with flatus, and it may be explosive. Diarrhoea due to the maldigestion of carbohydrate can occur several hours after ingestion of the noxious food or drink. These symptoms may result from consumption of only a few grams of the offending sugar. Intestinal hurry aggravates fat malabsorption in disaccharidase deficiency and may obscure the underlying cause of the diarrhoea. Deficiency of particular disaccharidases is responsible for the dietary intolerance of specific foods and drinks: milk-containing products in the case of lactase deficiency; table sugar and starch in asucrasia; mushrooms (and probably shellfish) in the rare trehalase deficiency. Identification of a cause-and-effect relationship between particular items and the intolerance syndrome is often impossible, given the ubiquity of sucrose and lactose in commercial foods.

Lactose intolerance

Most patients suffering from intolerance of lactose in the diet suffer either from lactase deficiency acquired as a result of intestinal disease, especially postinfective gastroenteritis in children, or as a result of genetically determined restriction of lactase expression.

Congenital lactase deficiency

A few infants have been reported in whom diarrhoea occurred after the first feed with breast milk and who responded completely to a lactose-free formula feed. This disorder is distinct from congenital glucose–galactose malabsorption, in which lactose exclusion alone is ineffective. Congenital lactose intolerance is associated with a severe inherited deficiency of mucosal lactase activity and, unlike the intolerance of lactose associated with prematurity or secondary to diffuse intestinal disease, remains lifelong. This syndrome leads to lactosuria due to the abnormal absorption of intact lactose, principally in the stomach; renal tubular acidosis and aminoaciduria have been recorded in this autosomal recessive disease that leads to vomiting, failure to thrive, and dehydration.

Lactase deficiency of prematurity

Unlike the other mucosal glycosidases, which appear early during fetal development, intestinal lactase activity is not fully expressed until after the 28th week of gestation and transient intolerance of milk feeds is common before this age. Abdominal distress due to gaseous distension and diarrhoea requires careful attention to the diet and fluid balance in premature infants.

Lactase restriction in children and adults

The capacity of the infant's intestine to digest lactose is retained into adult life by only a minority of individuals. Persistence of high intestinal lactase activity is an unusual state in adult mammals, and in humans is believed to have become prevalent in populations in which dairy culture was introduced about 10 000 years ago. Thus, tolerance of lactose in milk, dairy products, and many processed and ready-to-eat foods (Table 2) is found mainly in peoples of Northern European descent and those with a tradition of dairy farming. In about 5 per cent of Northern European adults, compared with more than 90 per cent in parts of Africa and Asia, there is a genetically determined and physiological decline in mucosal lactase activity after weaning. In most instances, reduction of mucosal lactose activity is associated with reduced synthesis of the precursor protein in the epithelial cells with apparently normal processing to the mature enzyme. The physiological decline in activity occurs between 3 and 5 years of age.

The element that determines lactase activity in adults acts in *cis* with the human lactase gene on chromosome 2q21. Recent studies of multiple polymorphic sites at this locus have identified a C/T polymorphism approximately 14 kb upstream of the gene that is tightly linked to the deficient/persistent lactase phenotype in several distinct populations. Homozygosity for the *C* allele at nucleotide –13910 is associated with adult lactase 'deficiency'—a finding consistent with the ancestral origin of lactase persistence. Although only low levels of lactase activity remain, this need be of no consequence when the consumption of dairy products is insignificant. Symptoms develop on exposure to excessive milk-and lactose-containing foods or medicines in late childhood or early adult life. The selective pressures that maintain this physiological reduction in mucosal lactase deficiency in childhood are unknown but the concept of 'lactase deficiency' in adults is difficult to justify, since lactase persistence is the least frequent variant. Nonetheless, with the increasing migration of peoples and their adoption of Western-style diets, this physiological loss of lactase activity is a prevalent cause of abdominal distress. A significant proportion of patients considered to have spastic colon, irritable bowel disease, or other 'functional' disturbances may prove to have lactase deficiency.

The speculative possibility arises that lactase-deficient subjects are at risk from osteoporosis in countries of the Northern hemisphere because of a dietary deficiency of calcium or vitamin D. A modest positive selection for the lactase persistence allele in Northern Europe would explain its present frequency if it arose at the time dairy farming was introduced. The relative lack of functional reserve of mucosal lactase activity also explains the frequency with which lactose malabsorption becomes manifest after partial gastrectomy and related procedures that enhance delivery of carbohydrate to the jejunum.

Diagnosis of lactose malabsorption

Intolerance of dietary carbohydrate caused by the maldigestion of lactose may be suspected from the dietary history of a patient typically complaining of abdominal pain, flatulence, and diarrhoea. Symptoms are often related to changes in social circumstances; they are frequently reported by Oriental immigrants to Western countries. The stool has an acidic pH (<6) and the osmolality of stool water is generally greater than 350 mosmol/kg due to the presence of lactate and other organic anions. Breath-hydrogen analysis is a useful confirmatory test. Hydrogen excretion determined by rebreathing 2 h after the ingestion of 50 g of lactose identifies patients with lactase deficiency diagnosed by enzymatic assay of jejunal mucosa. Other investigations, such as the lactose barium-meal examination and determination of blood glucose profile after oral challenge with lactose, are cumbersome and, because they give false-positive results, are now obsolete.

Secondary lactase deficiency

Lactase activity may be depressed by mucosal disease of the small intestine. This may occur transiently after infective gastroenteritis. It is particularly frequent in infants suffering from viral gastroenteritis, and continuing symptoms provoked by milk feeds can persist for days or some weeks. In infants, dehydration may develop rapidly, accompanied by prominent bloating; disaccharidu ria is found and acid, sour-smelling stools may be obvious. The symptoms resolve rapidly when dairy products are excluded from the diet. Decreased lactase activity also accompanies extensive and long-standing mucosal disease—the milk intolerance syndrome due to maldigestion may complicate coeliac disease, intestinal giardiasis, and Crohn's disease.

In secondary deficiencies of disaccharidases, because of the critical relationship between lactase activity and the rate of hydrolysis, intolerance of lactose predominates. However, the use of high-calorie supplements containing disaccharides other than lactose (especially maltose and sucrose) in patients with nutritional disturbances caused by intestinal disease may also cause the syndrome of carbohydrate maldigestion.

Sucrase–isomaltase (a-dextrinase) deficiency

This recessively inherited enzyme deficiency of the mucosal brush border is rare in all populations except the Inuit of Greenland, in whom the frequency of homozygotes is up to 10 per cent. Cetacean mammals also lack sucrase–isomaltase. Several defects of the human gene on chromosome 3q appear to be responsible; in some, there is aberrant glycosylation and the enzyme is inefficiently transported to the brush border. Substantial degradation of the abnormal polypeptide occurs within the epithelial cell.

Intolerance of sucrose is responsible for most of the symptoms, which develop as table sugar and sugar-containing foods are introduced during weaning. Intolerance of starch is less prominent because the osmotic contribution of the larger a-dextrin molecules that remain unsplit in the gut lumen is less. However, ingestion of large, starchy meals may induce cramping discomfort, flatulence, and diarrhoea. Whilst taking a normal diet, patients with deficiency of sucrase–isomaltase have persistent diarrhoea with the passage of acid and frothy stools containing increased concentrations of lactate.

The diagnosis may be suspected on the basis of the history of diarrhoea at weaning and on the character of the stools. Differentiation from coeliac disease, cow's milk allergy, infective or postinfective gastroenteritis, pancreatic failure, and disaccharide intolerance syndromes in relation to other inflammatory disease of the bowel is important, and biopsy of the jejunal mucosa for enzymatic assay and histological examination should be considered. In inherited sucrase–isomaltase deficiency, these activities are selectively reduced to less than 10 per cent of control values in histologically normal mucosa. Hydrogen breath tests after ingestion of sucrose and isomaltose may also prove to be useful in diagnosis, but experience is limited.

Trehalase deficiency

A few patients have been reported with mushroom intolerance due to the absence of mucosal trehalase. Trehalase is a brush-border a-glycosidase that cleaves the unusual 1a–1a bond of trehalase into its component glucose moieties. Trehalose is found in the haemolymph of arthropods and in fungi, so that intolerance of crustacean shellfish as well as mushrooms in the diet might be expected. Given that intolerance of edible fungi is not uncommon, trehalase deficiency may prove to be more frequent than previously supposed. Trehalase deficiency has also been reported to occur in 10 to 15 per cent of Greenland Inuit but the functional significance of this is unknown.

Treatment

Dietary exclusion of the offending sugar is the best method of preventing symptoms in individuals with primary or acquired disaccharidase deficiency. Symptoms recur as soon as excessive lactose or sucrose is reintroduced and advice from a professional dietitian may be needed to avoid indiscretions. In hypolactasia, complete elimination is not usually required, as lactase deficiency is rarely absolute; nevertheless, if symptoms persist there are many potential sources of lactose that warrant investigation (see Table 2).

An early, alternative method for preventing symptoms in lactose malabsorbers was the use of b-galactosidases obtained from yeast or other micro-organisms. These enzymes were added to dairy products before consumption and often changed the taste. In the United States, b-galactosidase has been produced commercially from yeast ('LactAid') and has been shown to reduce symptoms as well as breath-hydrogen excretion in subjects with maldigestion of lactose. Similar studies have demonstrated the efficacy of b-galactosidase derived from *Aspergillus oryzae* ('Lactrase'), in children with late-onset intolerance of lactose. The enzymes are taken in

tablet form immediately before challenge with lactose, but their cost, compared with dietary exclusion, may not be justified. In the future, microbial b-galactosidases might be used for food supplementation programmes in countries where lactose intolerance and nutritional deprivation in the adult population are widespread.

Complete absence of sucrase–isomaltase activity in most patients with sucrose intolerance together with the ubiquity of sucrose in modern diets complicates symptom management. Modest reduction of amylopectin-rich foods usually suffices to improve symptoms of starch intolerance but complete avoidance of sucrose-containing foods can be difficult especially in infants and young children. It has been reported that ingestion of dried brewer's yeast (containing invertase or sucrase but little lactase activity) after food is effective in patients with sucrase–isomaltase deficiency. However, dried yeast is rather unpalatable and not usually accepted by children. Recently, a high-potency liquid preparation of invertase used for the industrial manufacture of fructose from unrefined sugar cane juice ('Sacrosidase'; Universal Foods Corporation), which is approved by the Food and Drug Administration of the United States, has been used in a double-blind, randomized, controlled trial in patients with sucrase–isomaltase deficiency. The agent was found to be safe, acceptable, and effective for the treatment of all the associated symptoms and signs of this disease in patients receiving a low-starch diet.

Further reading

Anonymous (1992). Lactose intolerance. *Lancet* **338**, 663–4.

Clare H, Ruth M (1997). Phylogenetic analysis of the evolution of lactose digestion in adults. *Human Biology* **69**, 605–28.

Ennatah NS, *et al.* (2002). Identification of a variant associated with adult-type hypolactasia. *Nature Genetics* **30**, 233–7.

Gray GM (1975). Carbohydrate digestion and absorption. Rôle of the small intestine. *New England Journal of Medicine* **292**, 1225–30. [An informative and accessible review]

Hoskova A, *et al.* (1980). Severe lactose intolerance with lactosuria and vomiting. *Archives of Diseases in Childhood* **55**, 304–16.

King CE, Toskes PP (1983). The use of breath tests in the study of malabsorption. *Clinics in Gastroenterology* **12**, 591–610.

Madzarovova-Nohejlova J (1973). Trehalase deficiency in a family. *Gastroenterology* **65**, 130–3.

Medow MS, *et al.* (1990) b-Galactosidase tablets in the treatment of lactose intolerance in pediatrics. *American Journal of Diseases of Children* **144**, 1261–4. [Promising results with enzyme replacement therapy]

Simoons FJ (1978). The geographic hypothesis and lactose malabsorption. *American Journal of Digestive Diseases* **23**, 963–80.

Treem WR (1995). Congenital sucrase–isomaltase deficiency. *Journal of Pediatric Gastroenterology and Nutrition* **21**, 1–14.

Treem WR, *et al.* (1999). Sacrosidase therapy for congenital sucrase–isomaltase deficiency. *Journal of Pediatric Gastroenterology and Nutrition* **28**, 137–42.

Wang Y, *et al.* (1998). The genetically programmed down-regulation of lactase in children. *Gastroenterology* **114**, 1230–6.

14.9.6 Whipple's disease

H. J. F. Hodgson

Whipple's disease is an uncommon infection caused by the recently characterized actinomycete *Tropheryma whippelii.* The organism is widely distributed in tissues in affected individuals, and may cause disease in many systems. The condition is most commonly diagnosed when overt small intestinal disease occurs, leading to malabsorption, but systemic features such as arthralgia and fever may have been present for many years. The disease may also present with involvement of many other systems, including the brain and heart. Molecular techniques for identifying the organism are increasing the frequency with which individuals with minimal or absent gastrointestinal disease are diagnosed, thus expanding the spectrum of manifestations of Whipple's disease. The condition responds well to antibiotics but may relapse. The organism is probably acquired as an enteric infection, because it has been identified in effluent samples from sewage plants.

Pathology and aetiology

Advanced or fatal cases of Whipple's disease predominantly show severe intestinal and intra-abdominal pathology. The presence of fatty deposits in the small intestine and mesenteric lymph nodes prompted Whipple to call the disease intestinal lipodystrophy. The small intestine is thick and oedematous, with stubby or absent villi and dilated lacteals (secondary lymphangiectasia reflecting obstructed lymph flow). The absorptive enterocyte layer is virtually normal, but the lamina propria is stuffed with macrophages containing foamy material which stains brilliant magenta with periodic acid-Schiff reagent (diastase resistant) (Fig. 1). There is little inflammation otherwise. There are fatty deposits, and occasionally granulomas, in the mesenteric nodes as well as the characteristic macrophages. Other organs are involved to a varying degree, with foamy macrophages in spleen, lymph nodes, central nervous system, liver, lung, heart, and joints. Valvular endocarditis and localized brain deposits account for two of the most severe forms of the disease.

Fig. 1 Jejunal biopsy specimen from a 50-year-old man with Whipple's disease showing stunted villi and infiltration of the lamina propria with densely staining macrophages (periodic acid-Schiff stained, ×150).

Rod-shaped micro-organisms, the source of the periodic acid-Schiff-positive material, are identifiable at light and electron microscope level in affected tissues. After many decades of failure to identify these by conventional microbial culture methods, molecular characterization of the bacterial 16S ribosomal RNA gene in the tissues assigned the organism to a previously unrecognized class of actinomycetes. More recently human macrophages deactivated with anti-inflammatory cytokines and dexamethasone have allowed culture and passage of *Tropheryma whippelii*. The use of the polymerase chain reaction technique to identify sequences encoding the specific bacterial ribosomal RNA provides an invaluable diagnostic tool. The condition is an example of the value of molecular techniques in demonstrating that a disease process is infective in origin. The presence of the organism in sewage-exposed water suggests that infection occurs by invasion of the alimentary tract, and may explain the preponderance of small intestinal disease, with subsequent haematogenous or lymphatic dissemination. It remains unclear whether those who become infected have an underlying immunodeficiency, but host factors are probably relevant as there is evidence of diminished monocyte function persisting after successful treatment. A weak HLA B27 association has also been reported.

Demonstration of bacterial RNA-encoding sequences by means of the polymerase chain reaction now offers an alternative to the classical diagnostic techniques of biopsy and histology. However, some reports on saliva suggest the organism may reside as a commensal, emphasizing the need for clinical interpretation. Most patients continue to be diagnosed by the examination of the histology of the small intestine, reflecting the ease with which tissues can be obtained at routine upper gastrointestinal endoscopy. However, polymerase chain reaction positivity has been reported on peripheral blood, lymph nodes, synovial tissue, bone marrow, and even in faeces, and may yield positive results on small intestinal and other tissues in which characteristic histology cannot be identified. Application of diagnostic techniques based on the polymerase chain reaction to patients with arthritis, pyrexia of unknown origin, and other chronic undiagnosed conditions can identify cases in which small intestinal disease is not apparent, and this technique is likely to become widely used as it become more available. The use of the polymerase chain reaction has suggested that intractable idiopathic thrombocytopenia, quadriparesis, isolated muscle weakness, and juvenile chronic arthritis may form part of the clinical spectrum of Whipple's disease.

Differential diagnosis

Histological appearances suggestive of Whipple's disease have been reported in AIDS patients affected with other organisms—atypical mycobacteria and rhodococci.

Clinical features and diagnosis

The condition is most frequently diagnosed in middle aged or elderly men but women and children may be affected. The typical patient is diagnosed with relatively advanced disease with malaise, weight loss, diarrhoea, and arthralgias, and on examination may show marked pigmentation, lymphadenopathy, anaemia, finger clubbing, hypotension, and oedema. Rarely gastrointestinal bleeding may also occur. In such cases, investigation of an obvious gastrointestinal complaint should quickly establish the diagnosis. Recognition is far more difficult if symptoms are limited to fever or arthritis, or another systemic manifestation, which may be present transiently or intermittently for many years before the disease is diagnosed. The arthritis is migratory, non-deforming, and seronegative, predominantly affecting peripheral joints and in some series affects up to 90 per cent of patients. Other early features include respiratory symptoms with pleurisy and pulmonary infiltrates, and pericarditis. Chylous or serous ascites, endocarditis, cardiac conduction defects, coronary arteritis, and neurological abnormalities may occur with progression of the condition. In a recent survey over 80 per cent of patients had gut disease at diagnosis but 15 per cent had no gastrointestinal disorder at any time. Joint symptoms were present at some time in 83 per cent of patients, 20 per cent had neurological disease, 17 per cent cardiovascular disease, and 15 per cent mucocutaneous manifestations (pigmentation or sarcoid-like plaques). The central nervous system manifestations are diverse and include depression, apathy, fits, and mycoclonus, and a variety of ocular manifestations including ophthalmoplegia, papilloedema, scotomata, pseudotumour, and uveitis. Meningitis and a hypothalamic syndrome with insomnia, hyperphagia, and polydipsia also occur. Oculomastatory myorhythmia is said to be diagnostic.

Supplementary investigations are of value in confirming the involvement of different organs, but are not diagnostic of the disease. Radiographs of the small intestine characteristically show dilatation. Ultrasonography and computed tomography of the abdomen may show lymphatic masses, and computed tomography or magnetic resonance imaging of the brain may show multiple lesions in the white matter and grey–white junction with characteristic appearances. The sedimentation rate is generally but not inevitably elevated and an anaemia due to folate or iron deficiency may be present. Eosinophilia and thrombocytosis may be apparent on blood

films. Steatorrhoea, hypocalcaemia, vitamin deficiencies, and an elevated alkaline phosphatase occur with advanced gut disease, which may also give rise to hypoproteinaemia and a protein-losing enteropathy.

Treatment and prognosis

Whipple's disease progresses slowly, but unrecognized disease is eventually fatal. Antibiotic therapy is effective, although short-term corticosteroid therapy may occasionally be required in malnourished individuals to correct the metabolic and nutritional state. Administration of many different oral and parenteral antibiotics has been successful, including penicillin alone, penicillin plus streptomycin, tetracycline, and cotrimoxazole. A comparison of tetracycline versus cotrimoxazole (trimethoprine-sulphamethiazole) indicated superiority of the latter. Clinical improvement occurs within a few weeks, but prolonged treatment for at least a year is recommended. In particular it appears that the risk of a relapse with central nervous system manifestations is reduced if the initial regime involves drugs that pass the blood–brain barrier. The third-generation cephalosporin cefiximine has been reported to be effective in relapsing central nervous system Whipple's disease. A Herxheimer-like syndrome with fever and vasculitic manifestations has been reported at the start of treatment. The histological appearance of the gut mucosa returns to normal within a few months, although scattered periodic acid-Schiff-positive macrophages may persist for longer. Serial studies show that analysis of affected tissues by the polymerase chain reaction becomes negative in advance of histological improvement, and patients with clearance of tissues, documented by the polymerase chain reaction, appear to have a low risk of subsequent relapse. However, it is important to be aware of the possibility of relapse even after many years, especially when progressive central nervous system disease occurs in the absence of other systemic manifestations.

Further reading

Durand DV *et al.* (1997). Whipple disease. Clinical review of 52 cases. *Medicine Baltimore* **76**, 170–84.

Dutly F *et al*. (2000). *Tropheryma whippelii* DNA in saliva of patients without Whipple's disease. *Infection* **28**, 219–22.

Ectors NL *et al.* (1994). Whipple's disease: a histological, immunocytochemical, and electron microscopic study of the small intestinal epithelium. *Journal of Pathology* **172**, 73–9.

Gubler J *et al.* (1999). Whipple endocarditis without overt gastrointestinal disease *Annals of Internal Medicine* **131**, 112–16.

Louis ED *et al.* (1996). Diagnostic guidelines in central nervous system Whipple's disease *Annals of Neurology* **40**, 561–8.

Maizel H, Ruffin J, Dobbins W (1993). Whipple's disease: a review of 19 patients from one hospital and a review of the literature since 1950. *Medicine Baltimore* **72**, 343–55.

Marth T *et al.* (1997). Defects of monocyte interleukin 12 production and humoral immunity in Whipple's disease. *Gastroenterology* **113**, 442–8.

Misbah SA *et al.* (1997). Whipple's disease without malabsorption: new atypical features. *Quarterly Journal of Medicine* **90**, 765–72.

O'Duffy JD *et al.* (1999). Whipple's arthritis: direct detection of *Tropheryma whippelii* in synovial fluid and tissue. *Arthritis and Rheumatism* **42**, 812–17.

Playford RJ *et al.* (1992). Whipple's disease complicated by a retinal Jarisch–Herxheimer reaction. *Gut* **33**, 132–4.

Pron B *et al.* (1999). Diagnosis and follow-up of Whipple's disease by amplification of the 16S rRNA gene of *Tropheryma whippelii*. *European Journal of Clinical Microbiology and Infectious Diseases* **18**, 62–5.

Ramzan NN *et al.* (1997). Diagnosis and monitoring of Whipple disease by polymerase chain reaction. *Annals of Internals Medicine* **126**, 520–7.

Schoedon G *et al.* (1997). Deactivation of macrophages with IL4 is the key to the isolation of *Tropheryma whippelii*. *Journal of Infectious Diseases* **176**, 672–7.

Wilson K *et al.* (1991). Phylogeny of the Whipple's-disease-associated bacterium. *The Lancet* **338**, 474–5.

14.9.7 Effects of massive small bowel resection

R. J. Playford

Introduction

Large resections of the small bowel may cause multiple nutritional and other medical abnormalities, now commonly termed the 'short bowel syndrome'. More patients with this condition are surviving, due in part to improvements in anaesthetic and surgical techniques which have enabled patients previously considered to be inoperable to undergo radical procedures. The survivors present formidable medical problems, but scrupulous attention to their nutrition and general care, with the application of physiological principles, has improved their quality of life and independence.

Aetiology and prevention

The two main reasons why adults require massive intestinal resection are major vascular events involving the superior mesenteric artery, usually thrombosis or embolus, or multiple surgical resections of the small bowel in patients with Crohn's disease (regional ileitis). To reduce the number of those patients with Crohn's disease who will develop life-threatening consequences of recurrent disease in the residual intestine, it is imperative that the minimum amount of bowel is resected. Stricturoplasty, rather than resection, may be possible and multiple small segments of relatively normal intestine should be retained *in situ* and joined in series, rather than removed. Preservation of only a few additional centimetres of gut may be enough to allow the patient to be maintained on oral rather than parenteral nutrition.

The principal conditions requiring massive resection in children include segmental volvulus in the prenatal period and necrotizing enterocolitis postnatally. Rarer causes affecting all ages include trauma, retroperitoneal tumours, radiation enteritis, and strangulation, mainly resulting from adhesions.

Physiology

Although digestion and absorption of water, electrolytes, and nutrients occurs throughout the small intestine, there are regional differences. Regional functions of the jejunum include iron and folate absorption and disaccharide digestion and, in combination with the duodenum, the production of cholecystokinin and secretin.

The ileum is the principal site for absorption of vitamin B_{12} and bile salts and, in contrast to the jejunum, is capable of absorbing sodium against a steep gradient. It also plays a key role, in combination with the proximal colon, in mediating the 'ileal brake', in which intestinal transit and secretions are reduced when nutrients reach the terminal small bowel. Hormones, particularly peptide YY, probably mediate this phenomenon.

Factors, including adaptation, that influence the metabolic consequences of massive resection

The ability of the residual bowel to adapt after resection varies greatly between patients; it influences the development of symptoms and may determine the long-term requirement for parenteral nutrition. Four main factors influence the patient's ability to absorb nutrients:

1. Extent and site resected. The length of the small intestine varies between individuals. As a general rule, patients who have an intact duodenum but less than 50 cm of additional small bowel if the colon is *in situ*, or less than 100 cm if the colon has been removed, will require long-term total parenteral nutrition. Conversely, a requirement for parenteral nutrition is unlikely if more than 25 per cent of the small bowel remains.
2. Condition of the remaining intestine. The capacity of the residual bowel to adapt postoperatively is influenced by any underlying condition. Patients in which the residual bowel is damaged or abnormal due to conditions such as Crohn's disease or radiation enterocolitis are more likely to have metabolic disturbances.
3. Presence of the ileocaecal valve. It is not unusual for part or all of the colon to be removed along with segments of small intestine. Removal of the ileocaecal valve has a major impact on subsequent clinical progress and troublesome watery diarrhoea that compounds malabsorption is frequent. Factors contributing to this include faster intestinal transit, possibly related to loss of the 'ileal brake' mechanism, and a much higher likelihood of bacterial overgrowth.
4. Function of other digestive organs. Pancreatic hypofunction, resulting from malnutrition and reduced hormonal stimulation, may exacerbate fat malabsorption; this is sometimes compounded by gastric hypersecretion that inactivates pancreatic enzymes in the lumen.

Pathophysiology

Because regional differences in the function of the small intestine exist, the clinical sequelae of resection vary according to the site removed. Resection of most of the jejunum can usually be compensated for by the distal bowel, and the consequences of proximal resections are usually slight. Patients may experience iron and folic acid deficiency as well as lactose intolerance, resulting in abdominal bloating and watery diarrhoea.

Clinical problems are more likely to occur following large resections that include most of the ileum. Intractable (cholerheic) diarrhoea, often with steatorrhoea, and consequential metabolic abnormalities including vitamin B_{12} deficiency occur.

Diarrhoea

This is probably the most troubling symptom. Multiple factors are involved in its aetiology (Table 1):

- Transit time is decreased due to the reduced length of bowel and alteration in the control of its motility.
- Luminal osmolality is increased, partly due to reduced absorption of lactose and other carbohydrates, which are then metabolized by colonic bacteria. Severe metabolic (lactic) acidosis may develop—the increased 'anion gap' being due to the microbial generation of D-lactate.
- Disruption of the enterohepatic circulation of bile salts reduces the total body pool of bile salts. This is initially compensated for by a homeostatic upregulation of bile salt production by the liver. Increased delivery of bile salts into the colon, however, stimulates colonic adenylate cyclase activity, increasing colonic secretion of water and electrolytes, resulting in watery diarrhoea sometimes termed cholerheic diarrhoea.
- If most of the ileum has been removed, the compensatory upregulation of bile salt production may be insufficient to balance losses. This leads to decreased micelle formation in the lumen of the small bowel with a resultant reduction in absorption of water-insoluble fatty acids, causing the patient to have steatorrheic diarrhoea. Resection of the terminal 100 cm of ileum is typically associated with clinically significant malabsorption of bile salts. The presence of excess a-hydroxy fatty acids derived from bacterial metabolism in the colonic lumen stimulates adenylate cyclase, further increasing secretion of fluids and electrolytes.
- In massive intestinal resections, the reduced micellar solubilization of fat and consequential impairment of lipolysis is compounded by the loss of absorptive

mucosa, thus aggravating the effects of maldigestion and fluid loss.

Stones

Gallstone formation is two to three times more common after ileal resection and may be of the cholesterol rich or pigment type. Reduced concentrations of bile salts within the bile due to depletion of the body pool of bile salts, in combination with gall bladder hypomotility, facilitate the formation of cholesterol crystals.

Renal stones (usually calcium oxalate) commonly result from increased absorption of oxalate and hyperoxaluria. The availability of free oxalate within the colon is increased by excessive complexation of calcium by fatty acids which normally promote formation of insoluble (non-absorbable) calcium oxalate. Although concentrations of bile salts in the small intestine may be reduced, the failure to reabsorb bile salts in the ileum increases luminal bile salts in the colon; this increases colonic permeability and further promotes oxalate absorption.

Gastric hypersecretion

This phenomenon occurs in some patients, although its severity tends to lessen over time. Hyperacidity may inactivate pancreatic enzymes by precipitating bile salts and lowering intraduodenal pH as in Zollinger–Ellison syndrome.

Nutritional status

Many patients undergoing resections will be malnourished preoperatively and energy consumption increases in the immediate postoperative period. If not appropriately managed, long-term protein-energy malnutrition, as well as life-threatening mineral and vitamin deficiencies develop.

Adaptation

Morphological and functional adaptive changes follow resection of the small intestine. The residual bowel undergoes mucosal hyperplasia and its capacity to absorb fluids and nutrients increases over a period of weeks or months. The molecular events that underly these changes are unclear but may include circulating trophic factors and growth factors present in pancreatic juice or secreted into the intestinal lumen. Early intervention is required to achieve maximal adaptation, and maintenance of a supply of luminal nutrients is a prerequisite for the adaptive changes. It is therefore important that luminal feeding is started as early as possible after surgery even if the patient also requires parenteral nutrition.

Management

Initial therapy

In the initial postoperative period, vigorous intravenous fluid and electrolyte replacement is required to prevent dehydration and to compensate for intestinal losses. Many patients will also require parenteral nutritional supplements while the residual bowel adapts. Ingestion of water may exacerbate diarrhoea and be counterproductive. The use of an oral iso-osmolar saline–glucose solution containing bicarbonate, similar to that used for the treatment of cholera, may often assist in reducing intravenous requirements without increasing intestinal fluid loss.

Nutrition

Oral nutrition, initially consisting of elemental or polymeric diets administered by nasogastric or enteral tube feeding, should ideally be started within the first few days of surgery. The introduction of luminal nutrition tends, however, to exacerbate the diarrhoea. Many high-calorie enteral supplements for use in malnourished patients who have little or no impairment of small intestinal function have a very high osmolality, thereby inducing catastrophic egress of luminal fluid and diarrhoea in patients with large resections. These preparations are to be used with great caution or avoided altogether in patients suffering the effects of massive bowel resections.

Subsequently, small-volume, frequent, solid or semisolid meals with low fat and oxalate content should be introduced. Low-fat meals and supplements containing large quantities of medium-chain fatty acids tend to be unpalatable. Compliance of patients with dietary advice is therefore best if symptoms are used as a guide to the amount of fat that is included in the diet. Since much of the energy content of the ingested diet may well be lost in the stool, the daily intake of calories often has to be greater than expected. This is best provided in a complex form including glucose polymers and starch which have little osmotic effect in the lumen and are hydrolysed rapidly by brush-border hydrolases at the site of absorption. Lactose intolerance, seen particularly in patients following significant jejunal resections, may induce bloating and exacerbation of diarrhoea but usually responds to reduction in lactose-containing dairy products. Low-fibre diets are helpful in some patients although they may aggravate symptoms in others; treatment must be tailored to the individual. Patients should be encouraged to take multivitamin and mineral supplementation at levels two to five times the normal recommended daily requirements; vitamin B_{12} injections are required following terminal ileum resection. In all patients, regular long-term monitoring of fat-soluble vitamins (A, K, and D), vitamin B_{12}, folate, magnesium, zinc, and bone status is required.

In some patients, adequate fluid and nutritional balance cannot be maintained by the oral route alone and long-term total parenteral nutrition is needed. These patients should be encouraged to continue oral nutrition, for social and psychological reasons as well as to minimize the amount of parental nutrition required.

Drugs

Most patients will require antiperistaltic drugs to increase the time of contact between luminal contents and residual bowel. A step-wise approach should be used, starting with agents such as loperamide or codeine phosphate. Long-term administration of the more potent constipating, but potentially addictive, opiates should only be used in intractable cases. Since diarrhoea may be particularly troublesome in the initial postoperative period, liquid or occasionally intravenous formulations may be needed.

Administration of H_2-receptor antagonists or proton pump inhibitors may reduce diarrhoea and promote digestion, as well as prevent peptic ulceration, by decreasing gastric secretions and preventing inactivation of pancreatic enzymes. Cholerheic diarrhoeas may respond well to bile acid sequestrants such as cholestyramine but its use can worsen the fatty component of diarrhoea by exacerbating the deficiency of bile salts. Similarly, the use of long-acting somatostatin analogues can reduce gastrointestinal secretions and fluid loss but may exacerbate steatorrhea and formation of gallstones. In patients with marked steatorrhea who do not respond to restriction of fat intake, the addition of oral pancreatic enzyme supplements to food may assist lipolysis and improve digestion.

Bacterial overgrowth

Colonization of the small bowel by colonic or pathogenic bacteria results in exacerbation of diarrhoea, malabsorption, and nutritional deficiencies. Culture and analysis of small bowel aspirates is required for definitive diagnosis but is a moderately invasive procedure. Because results from many of the usual non-invasive tests, for example glucose or lactulose hydrogen breath tests, are abnormal in all patients after significant resection, empirical trials of antibiotics may be justified.

Surgical options

In patients with severe intractable diarrhoea further surgery should be considered, although it is usually of limited benefit. Although not in general clinical use, reversal of a small segment of small bowel can delay gut transit; however, if too long a segment is used, obstruction may occur. Longitudinal lengthening may also be of value, particularly in paediatric patients.

Small bowel transplantation is now available in a limited number of centres. Because of the high morbidity and mortality associated with transplantation, it is usually only offered to those patients who cannot be maintained on total parenteral nutrition. Patients who have undergone intestinal transplantation are particularly prone to infections and lymphoma. Problems with acute and chronic rejection are also common. Patients therefore require detailed counselling about the risks of any such procedure.

Future directions

Administration of gut trophic factors, such as glucagon-like peptide 2, hepatocyte growth factor, epidermal growth factor, and growth hormone (possibly in combination with glutamine supplements), during the early postoperative period may increase the rate and extent of mucosal adaptation that occurs in the intestine. In patients who prove not to adapt adequately, continuing advances in techniques for small bowel transplantation and in antirejection therapy offers future hope. In the longer term, advances in tissue engineering technology may allow intestinal mucosa to be obtained from humanized animal gut or to be reconstituted in culture from the patient's own residual bowel, thereby removing the problems of rejection and immunosuppression.

Further reading

Grand D (1999). Intestinal transplantation: 1997 report of the international registry. Intestinal Transplant Registry. *Transplantation* **67**, 1061–4. Review of the survival figures provided by 33 intestinal transplant programme centres.

Kim SS, Vacanti JP (1999). The current status of tissue engineering as potential therapy. *Seminars in Pediatric Surgery* **8**, 119–23.

Robinson MK, Ziegler TR, Wilmore DW (1999). Overview of intestinal adaptation and its stimulation. *European Journal of Pediatric Surgery* **9**, 200–6. Discusses the use of growth factors and dietary constituents to stimulate adaption.

14.9.8 Malabsorption syndromes in the tropics

V. I. Mathan

Introduction

Patients, in whom the 'digestive fire is weakened and food is expelled from the body without contributing to growth' were described in *Charaka-Samhita*, an ancient Indian treatise on medicine, compiled some time between the sixth and twelfth centuries BC. The clinical description in the section on '*Grahani Vyadhi*' or diseases of the organ of assimilation, clearly describes patients gradually wasting with chronic diarrhoea and loud borborygmi. Malabsorption of nutrients with its sequelae has therefore long been recognized as a clinical entity in some tropical regions.

Malabsorption is the result of the failure of intestinal function. It is multifactorial and may be due to failure in the luminal and mucosal phases of digestion and absorption, as well as of transport from the intestine. Detailed investigation of the individual patient is essential for diagnosing the causes of malabsorption.

In the context of the 'global village' of the third millennium, are there defined 'tropical' malabsorption syndromes? All the causes of nutrient malabsorption prevalent in temperate climates also occur in the tropics, but there are certain conditions: the majority being chronic enteric infections or infestations that are geographically limited to the tropics. Expatriates from other parts of the world to the tropics may have a higher susceptibility to some of these conditions.

Causes of malabsorption primarily prevalent in the tropics

The small intestinal and colonic mucosa of apparently healthy residents of many tropical countries, in comparison to residents of temperate-zone industrialized countries, show minor morphological and functional abnormalities. These changes, designated 'tropical enteropathy' and 'tropical colonopathy', are the normal background on which clinically significant malabsorption occurs.

Malabsorption in the tropics, as elsewhere, may have an identifiable underlying aetiology, when it is classified as secondary malabsorption. When no primary cause has yet been identified it is considered primary or idiopathic malabsorption (Table 1).

The importance of a variety of protozoal infections, especially intracellular protozoans, was recognized in temperate-zone countries at the beginning of the AIDS epidemic, these organisms having been identified as opportunistic infections. In tropical countries these protozoa have been identified in symptomatic and asymptomatic immunocompetent subjects.

Capillaria philippinensis infestation has been reported in epidemics from The Philippines and as sporadic cases from other tropical countries including India. Hyperinfection with *Strongyloides stercoralis* can occur rarely. Both these helminths burrow into the mucosa and form tunnels.

Abdominal tuberculosis occurs much less frequently than pulmonary tuberculosis and is often secondary. Malabsorption in abdominal tuberculosis is the result of bacterial colonization of the small intestinal lumen secondary to strictures and extensive ulceration, or due to obstruction of the lymphatic outflow (*Tabes mesenterica*).

Progressive wasting in African people infected with the human immune deficiency virus (**HIV**) is known as 'slim disease', and is a consequence of malabsorption secondary to enteric opportunistic infections. There is some evidence to suggest that a primary HIV enteropathy can also contribute to malabsorption.

Calcific pancreatitis affecting young adults, particularly in economically disadvantaged sections of society, is another cause of malabsorption unique to the tropics.

Immunoproliferative small intestinal disease (**IPSID**) is due to the clonal expansion of immunocytes producing altered alpha heavy-chain immunoglobulin. This is also known as Mediterranean lymphoma and has been reported from several tropical countries. The characteristic histology in the premalignant stage is diagnostic and can be reversed by prolonged antibiotic therapy at this stage. Once malignant transformation has occurred the treatment is as for other lymphomas, but the prognosis is guarded.

In many tropical regions, particularly south and south-east Asia, the high lactase activity in the intestinal epithelium in the neonates declines rapidly after weaning. Since most adults in such countries do not regularly consume milk or lactose in their diets, this abnormality is of relatively small significance. The use of fermented milk and milk products (for example, yoghurt) can ensure that milk-based nutritional supplementation is still possible in such populations.

Malabsorption, or increased secretion, is an invariable part of all acute diarrhoeal infections. These episodes, most frequent in children, are of short duration, usually a few days. A small proportion of infants and young children have diarrhoea that persists for longer than 2 weeks following an acute episode. This persistent diarrhoea syndrome, seldom if ever seen in adolescents and adults, is also not considered as one of the malabsorption syndromes.

The concept of a 'postinfective malabsorption state' associated with the presence of a mixed bacterial flora in the small intestine has been postulated to explain, in particular, the persistent diarrhoea and malabsorption reported in many European travellers to the Indian subcontinent. By extension, it has been suggested that the syndrome of primary malabsorption in the tropics, tropical sprue, is only another form of postinfective malabsorption. Several factors contradict this assumption. Significant bacterial colonization of the small intestine has been found in apparently healthy asymptomatic adults resident in the tropics. Detailed investigation of many overland travellers from Europe to the Indian subcontinent identified several of the infections described earlier as the cause of persistent malabsorption, along with an altered luminal bacterial flora. The epidemiology of tropical sprue is distinctly different from that of acute infectious diarrhoea. Detailed clinical and laboratory investigations of adults and children in over 20 epidemics of acute diarrhoea, studied in south India, identified no single case of persistent malabsorption, other than the background prevalence of tropical enteropathy. There are also well-documented instances of expatriates from Europe who developed a primary malabsorption syndrome many years after their return to temperate climes. A careful analysis of the available literature therefore suggests that significant persistent and symptomatic malabsorption following acute enteric infectious diarrhoea is a rare event in the tropics.

When patients with conditions that can give rise to secondary malabsorption unique to the tropics or elsewhere are excluded, a group remain who have chronic diarrhoea, malabsorption, and its nutritional sequelae. Such patients are relatively rare in temperate climates, but they are frequently encountered in areas such as southern India and the Caribbean islands. This primary or idiopathic malabsorption syndrome has been called 'tropical sprue'. Tropical sprue occurs on the background of tropical enteropathy in the indigenous population of these regions.

Tropical enteropathy

The intestinal mucosal morphology of germfree and conventionally reared, litter-mate, rats is different; the latter having shorter villi, higher crypts, and increased mononuclear cells infiltrating the lamina propria and epithelium. These differences are attributed to the modulating effect of the microbial flora in the intestinal lumen of the conventionally reared litter-mates.

Similar morphological differences are found between the jejunal mucosa of apparently healthy asymptomatic individuals living in temperate-zone industrialized countries and those in tropical preindustrialized countries. The morphological features of this tropical enteropathy are characterized by the replacement of finger- and tongue-shaped villi by broader structures in the upper small intestine, reduction in the height of villi with an increase in crypt thickness, and an increased infiltration by mononuclear cells in the lamina propria and the epithelium. Similar mucosal morphological changes have also been shown to occur in the large intestine.

The morphology of fetal intestinal mucosa is identical in both geographical regions, the earliest differences appearing shortly after birth. The morphological changes are not apparent in biopsies from residents of Singapore—although this is a tropical country, it has standards of environmental hygiene and nutrition that equal those in temperate-zone industrialized countries. People expatriated from temperate countries to tropical countries develop these mucosal morphological changes over time. Even if expatriates from tropical countries living in a temperate zone continue to ingest a diet similar to that in their original home, they eventually revert to having a temperate-zone morphology. The evidence therefore suggests that the morphological alteration in the small intestine of residents of tropical countries is not a result of climatic differences, but is probably a reflection of an adaptation to environmental factors. *In vitro* organ culture studies have shown a slightly accelerated cell turnover in the jejunal mucosa of people living in the tropics, further supporting an adaptive response as the basis for the change.

Extensive bacterial colonization of the upper small intestinal lumen and mucosa, in apparently healthy adults, by aerobic and anaerobic bacteria, has been documented in studies from southern India. There is also circulation of enteric pathogens in asymptomatic individuals. It is also known that the first dose of oral immunization agents to enteric pathogens usually results in a secondary response, even in children as young as 2-years old living in the tropics. There is no evidence that the macronutrient deficiency widely prevalent in many tropical countries influences intestinal structure or function. No such information is available regarding micronutrient deficiencies. Conceptually, it may be useful to categorize 'specific pathogen-free' populations (temperate zone) and 'conventional' populations (tropical)!

Minor abnormalities in absorption can also be demonstrated in these healthy subjects, with xylose malabsorption in 40 per cent, mild steatorrhoea in 10 per cent, and vitamin B_{12} malabsorption in 3 per cent. The overall absorption of calories is reduced by about 5 per cent, while the colonic bacterial mass is increased. Colonic salvage of unabsorbed calories is thereby reduced. There is no evidence that these changes in the lining epithelium of the intestinal tract, a primary barrier between the internal and external environment of the body, significantly affects the health of these 'conventional' populations. However, the reduction in overall caloric absorption can raise the question as to whether the absence of tropical enteropathy can increase the effective availability of food without an increase in supply.

Tropical sprue

Definition

This primary (idiopathic) malabsorption syndrome affecting residents of, or visitors to, certain tropical regions, with characteristic enterocyte damage, is usually associated with chronic diarrhoea and the nutritional sequelae of persisting malabsorption. The aetiology(ies) underlying this syndrome is not yet understood. There are differences in the presentation, epidemiology, and clinical course in different geographical regions and between expatriates to endemic regions and indigenous residents affected by the syndrome.

History

William Hillary described a chronic wasting diarrhoea in European expatriates in Barbados in 1759, probably the first description of the syndrome in the English literature. The disease apparently attained epidemic proportions 3 years after he arrived in Barbados. The syndrome in expatriates was well recognized by British and Dutch physicians in south and south-east Asia with expanding colonization. However, no cases were described from tropical Africa. Tropical sprue assumed epidemic proportions during the Second World War and was a major factor for repatriation from the Assam and Burma theatres of war. Indian troops were also affected. It was only in the postcolonial era, with the work of Baker and colleagues in southern India and of Klipstein in Puerto Rico and Haiti, that the extent of the problem of tropical sprue was defined in indigenous populations.

Epidemiology

Endemic cases in indigenous and expatriate residents and epidemics in troops and indigenous populations have both been described. Endemic tropical sprue is apparently geographically restricted to south and south-east Asia and the Caribbean islands other than Jamaica, with a few case reports from Central and South America and Sub-Saharan Africa. In fact, much of the literature up to the 1960s is limited to expatriate populations. In India, only two large medical institutions, with well-developed laboratory facilities, have reported detailed studies, suggesting that in marginally nourished indigenous populations many cases may be missed due to the poor availability of diagnostic facilities.

Apart from the reports during the Second World War, large epidemics have only been described from southern India. The first such reported epidemic in 1960–61 affected approximately 100 000 patients, with a 40 per cent case fatality. This was reflected in the unusual death rates in the North and South Arcot Districts of Madras state in the 1961 census of India. The last epidemic was detected in 1978. In all the epidemics, patients initially developed an apparent episode of acute diarrhoea accompanied by vomiting in about 30 per cent and fever in 25 per cent of cases. Significant malabsorption of fat, carbohydrate, and vitamin B_{12} was present even during the first week of illness and 50 per cent of those affected had diarrhoea for longer than 1 month. The epidemics evolved over a period of months to years: adults had a significantly higher attack rate and were affected earlier during the course of the epidemic. The epidemiological data suggested an infective aetiology, but no causal viral, bacterial, or parasitic agent could be found.

Clinical features

The patient with tropical sprue is usually an adult with a history of loose or watery stools lasting for several weeks or months and with symptoms and signs of nutritional deficiency. There is usually anorexia, a feeling of abdominal distension, and loud abnormal borborygmi. The signs of nutritional deficiency include pallor due to anaemia, angular stomatitis, glossitis, oedema, and the skin and hair changes of severe hypoproteinaemia. The prevalence of nutritional deficiency, measured by clinical or laboratory parameters, is higher in those patients with a longer duration of symptoms. In the epidemic situation the prevalence of nutritional deficiency in patients during the first month of illness was no different from that in the unaffected people in the same village. However, in patients affected in the epidemics persistent malabsorption begins during the first few days of illness. The diarrhoea can be severe enough to produce life-threatening dehydration: in the epidemics the early deaths were mainly due to fluid and electrolyte imbalance. These could be prevented by maintenance of hydration. As the disease progresses the sequelae of severe malnutrition and consequential acute infections, especially of the respiratory tract, contribute to mortality. The natural history of the illness shows periods of remission, relapses, and spontaneous recovery, which make an evaluation of specific therapy difficult. Although patients have been followed for up to 25 years in southern India, intestinal neoplasms have not developed.

Investigation

Investigation of these patients should confirm the presence of intestinal malabsorption, exclude conditions that can give rise to secondary malabsorption, and evaluate the nutritional sequelae of malabsorption (Table 2).

A simple faecal smear stained with a fat stain such as Sudan 3 can often detect fat globules and fatty acid crystals associated with steatorrhoea. The extent of tests for confirming the presence of malabsorption is determined by the availability of facilities, which in many tropical areas is still limited. Tests of xylose absorption should be interpreted in the light of xylose malabsorption as a part of tropical enteropathy in the particular community.

The exclusion of conditions that can give rise to secondary malabsorption is of importance since many of these conditions are amenable to therapy. The diagnosis of the syndrome of primary malabsorption is one of exclusion.

Evaluation of the nutritional sequelae of malabsorption, especially the presence of megaloblastic anaemia, provides useful benchmarks for appropriate nutritional

rehabilitation.

Pathology and pathogenesis

The wide availability of peroral mucosal biopsies confirmed the report, as early as 1924, that the primary lesson in tropical sprue was in the small intestinal mucosa. Electron microscopic examination of jejunal mucosal biopsies confirmed the presence of damage to enterocytes in the crypt (regenerative) and villous (functional) compartments. This damage can be demonstrated in the first weeks of illness in patients affected during epidemics. Accelerated cell turnover in the regenerative compartment and increased loss of enterocytes from the functional compartment was demonstrated by *in vitro* culture of jejunal mucosal biopsies labelled with tritiated thymidine. In fact these changes in the enterocyte lifecycle explain the observed mucosal architecture, which has often been called partial villous atrophy. In contrast to the situation in coeliac disease, where the initial damage to enterocytes occurs in the functional compartment and the crypts are hypertrophied with normal enterocytes, in tropical sprue the primary lesion appears to affect the regenerative compartment. These findings have only been confirmed in patients studied in southern India.

The mucosal lesion in tropical sprue is not confined to the small intestine, since functional and structural abnormalities have also been demonstrated in the stomach and the colon. Significant water malabsorption in the colon may contribute to the severity of diarrhoea.

An appreciation of regional differences in the patient profile is essential for understanding the pathology and pathogenesis. All patients have malabsorption, but vitamin B_{12} malabsorption is only found in about 70 per cent of patients in southern Indian, while it is almost invariable in expatriates from the temperate zones and in the Caribbean. In Haiti and Puerto Rico the observed seasonal incidence is ascribed to small intestinal colonization by toxin-producing coliforms, probably secondary to the consumption of rancid pork fat. In southern India, the extent and severity of small bowel colonization by enterotoxin-producing coliforms is no higher than in matched controls. Identification of one or more 'agents' that can damage the mucosal epithelial cells will enable a clearer understanding of these differences.

Treatment

Provision of symptomatic relief from diarrhoea, correction of fluid and electrolyte abnormalities and nutritional deficiencies, and attempts at specific curative measures are the cornerstone of treatment. Diarrhoea and abdominal distension can be helped by the judicious use of loperamide and dimethyl polysiloxane. Increasing the nutrient intake and providing therapeutic supplements such as vitamin B_{12} and folic acid, as indicated, is beneficial. However, specific therapy to cure the condition awaits understanding of its aetiology.

Empirical evidence from patients in the Caribbean and European expatriate community, indicates that folic acid can alleviate symptoms in cases of less than 2 months' duration. In patients with a longer duration of symptoms, the addition of oral tetracycline for up to 6 months leads to the restoration of normal intestinal absorption. In southern India, the results of therapy with vitamin B_{12}, folic acid, and tetracyclines were not so clear-cut and a few patients were resistant to all therapy. Nevertheless, the recommended management includes the use of all three of these therapeutic agents for up to 6 months.

Areas needing further research

Malabsorption syndromes occur in tropical regions and there are conditions that are unique to this geographical region. While the advent of HIV infection has created an added dimension, it is important to continue the search for agents that can directly damage the endoderm-derived epithelial cells. The public health significance of tropical sprue is probably declining, with the last reported epidemic occurring over 20 years ago and the improving nutritional status of many tropical populations enhancing their ability to compensate for mild to moderate degrees of intestinal malabsorption.

Nevertheless, understanding tropical sprue, a primary or idiopathic malabsorption syndrome of the tropics, continues to be an intriguing challenge.

Further reading

Baker SJ (1973). Geographical variation in the morphology of the small intestinal mucosa in apparently healthy individuals. *Pathology and Microbiology* **294**, 222–37.

Baker SJ, Mathan VI (1972). Tropical enteropathy and tropical sprue. *American Journal of Clinical Nutrition* **25**, 1047–55.

Manson-Bahr PH (1924). The morbid anatomy and pathology of sprue and their bearing upon aetiology. *Lancet* **1**, 1148–51.

Mathan M, Mathan VI, Baker SJ (1975). An electron-microscopic study of jejunal mucosal morphology in control subjects and patients with tropical sprue in southern India. *Gastroenterology* **68**, 17–32.

Mathan M, Ponniah J, Mathan VI (1986). Epithelial cell renewal and turnover and its relationship to the morphological abnormalities in the jejunal mucosa in tropical sprue. *Digestive Diseases and Sciences* **31**, 586–93.

Mathan M, *et al.* (1990). Ultrastructure of the jejunal mucosa in human immuno deficiency virus infection. *Journal of Pathology* **16**, 119–27.

Mathan VI (1988). Tropical sprue in southern India. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **82**, 10–14.

Ramakrishna BS, Mathan VI (1982). Water and electrolyte absorption by the colon in tropical sprue. *Gut* **23**, 843–6.

The Wellcome Trust (1971). *Tropical sprue and megaloblastic anaemia*. Churchill Livingstone, London.

14.10 Crohn's disease

D. P. Jewell

Crohn's disease is a chronic inflammatory disease of the gastrointestinal tract, the cause of which remains unknown. It is characterized by a granulomatous inflammation affecting any part of the tract, frequently in discontinuity, and by the tendency to form fistulas.

History

The first clear description of the disease affecting the terminal ileum (regional ileitis) was given by Crohn, Ginzburg, and Oppenheimer in 1932. However, the disease certainly existed long before then and many of the early descriptions of ulcerative colitis would now be regarded as Crohn's disease. Dalziel, in 1913, described an inflammatory process of the ileum and colon consisting of ulceration, submucosal oedema, fibrosis, and mesenteric lymphadenopathy. He reported the presence of granulomas on microscopy but could find no evidence of tuberculosis. Similar cases were described in the 1920s by oschowitz and Willensky.

After the description by Crohn and his colleagues, it was clearly recognized that the colon could also be involved and, on occasions, it could be the sole site of the disease. The disease therefore became known as regional enteritis or, preferably, Crohn's disease. Colonic disease is often referred to as Crohn's disease of the colon, Crohn's colitis, or granulomatous colitis.

Epidemiology

Crohn's disease is well recognized in Europe, Scandinavia, North America, and Australia but is rarely seen in India, tropical Africa, and South America. This may be largely due to the difficulty of diagnosing Crohn's disease in areas where intestinal tuberculosis is common and to the problems of long-term follow-up. However, it is now being recognized in India in specialist gastrointestinal units. The disease is about 10 times less common in Japan than in the West, but its prevalence in Japan appears to be increasing.

There has been a striking increase in the incidence and prevalence of Crohn's disease in Europe and Scandinavia since 1950 (Table 1). This is also shown by examining the annual discharge rates in England and Wales. For Crohn's disease, the rate rose from 2.8 per 100 000 in 1958 to 7.2 per 100 000 of the population in 1971, whereas the rate for ulcerative colitis during the same period was unchanged at 10 to 12 per 100 000. Recent studies in Scotland and in Stockholm have suggested that the incidence has begun to decline in adults, but there is a widespread belief that Crohn's disease is increasing in children, supported by studies in Scotland.

The reasons for the changing patterns of incidence are not clear. Much of the increased incidence is due to an increased frequency of colonic disease and it might be argued that this represents diagnostic transfer from ulcerative colitis to Crohn's disease. The annual discharge rates for England and Wales, quoted above, make this explanation unlikely. Whether similar changes in frequency have occurred in North America is uncertain, although the data available suggest that the incidence has probably not altered. It is possible that the changing incidence may result from an infective or environmental factor.

Crohn's disease occurs in all age groups but it is rare in early childhood and most commonly affects young adults. There is no marked sex difference, and no association with social class or occupation. There may be an increased incidence amongst Ashkenazi Jews, especially in the United States.

Genetics

Evidence for a genetic susceptibility to developing Crohn's disease is suggested by a much higher concordance in monozygotic twins (approximately 45 per cent) compared with dizygotic twins (approximately 15 per cent), and by the finding that 10 to 15 per cent of patients will have at least one other family member affected. The highest risk is between siblings (ls 20 to 25—this is a measure of relative risk), but there is still an increased risk of affected offspring (ls 15) if a parent is affected. However, given the low incidence of Crohn's disease, the absolute risk to family members if one individual develops the disease is still low. Within a multiply-affected family, there is a remarkable concordance with respect to disease type (most will also have Crohn's disease, although Crohn's disease and ulcerative colitis can occur in the same family) and disease behaviour. However, there is no clear mode of inheritance and recent linkage studies utilizing microsatellite markers to link disease to areas of the genome have suggested that multiple genes are involved. It seems likely that there are a variety of genes which render individuals susceptible to 'inflammatory bowel disease' but that other genes determine the type and behaviour of disease. Crohn's disease has been linked to a pericentromeric region of chromosome 16 in many different populations, thus providing strong evidence that there is a relevant gene in this region. In 2001 this gene was identified as *NOD2*, encoding a protein which is an intracellular receptor for bacterial lipopolysaccharide. A number of point mutations and a frameshift mutation are found in up to 40 per cent of patients with ileal disease but are not associated with colonic Crohn's disease. Linkages to chromosomes 6 and 14 have also been replicated.

Aetiology

The cause of Crohn's disease is unknown but clearly involves an interplay between genetic and environmental factors. The latter include smoking and intestinal luminal factors. There is a relative risk of 4 to 6 for Crohn's disease in smokers compared with non-smokers, which is in striking contrast to the reverse association seen in ulcerative colitis. Whether smoking predisposes to Crohn's disease by altering mucosal blood flow, synthesis of mucus, or by an effect on endothelial cells is unknown. The role of luminal factors is suggested by the tendency of Crohn's colitis to heal if the colon is rendered non-functional by an ileostomy and by the effectiveness of an elemental diet for the treatment of active disease. Other mechanisms may contribute to the pathogenesis and include diet, infective agents, ischaemia, and immune mechanisms. Recent claims that the measles virus or the **MMR** (measles, mumps, rubella) vaccine might predispose to Crohn's disease later in life have not been confirmed by subsequent data. Likewise there are no confirmatory data that *Mycobacterium paratuberculosis* is involved in disease pathogenesis.

Diet

Several investigators have reported that patients with Crohn's disease have a higher intake of refined sugar than a control population or a matched group of patients with ulcerative colitis. In addition, patients with Crohn's disease may also have a reduced intake of fibre, especially that derived from fruit and vegetables. However, the significance of these changes is unclear, especially as a controlled trial was unable to show that a low-sugar, high-fibre diet had any effect on the cause of the disease over a 2-year period. Claims have been made that many patients will benefit from dietary exclusion determined by a period on an elimination diet followed by challenge with individual foods. This might suggest a role for dietary factors, but the long-term benefit of dietary exclusion is by no means proven. Elemental diets consisting of glucose and amino acids have been shown to have equal efficacy to prednisolone for treating active Crohn's disease, but whether this effect is mediated by influencing bacterial populations, by removing dietary antigens, or by some other mechanism is unknown.

Infective agents

Viruses, cell-wall deficient bacteria, and atypical mycobacteria have been claimed to be the cause of Crohn's disease. Most of these claims have been discredited but there is current interest in the role of *M. paratuberculosis* and measles virus.

M. paratuberculosis is the organism that causes Johne's disease in cattle and other farm animals. This resembles Crohn's disease in so far as it is a granulomatous inflammatory disorder of the intestine. Over the last 20 years, an atypical mycobacterium has been isolated from intestinal tissue of a few patients with Crohn's disease and most of the isolated organisms have been shown to be identical to *M. paratuberculosis* using DNA analysis. The organism is very slow growing and it has taken 1 to 2 years of culture in order to demonstrate it. The possibility that such an organism might be involved in the aetiology of the disease was strengthened by the detection of *M. paratuberculosis* DNA in intestinal tissue in about two-thirds of patients using a specific probe and amplification by the polymerase chain reaction. However, specific DNA for the organism was also detected in 10 per cent of control tissue and other investigators have been unable to detect specific DNA in any patient with Crohn's disease. Antituberculous therapy has not proved effective so far. Thus, it is still very uncertain whether *M. paratuberculosis* is an aetiological agent, whether it is responsible for causing disease in all or just a subgroup of patients, whether some of the findings are laboratory artefacts, or whether it is a secondary invader of inflamed tissue.

The initial reports linking measles virus or the MMR vaccine to the risk of developing Crohn's disease later in life have not been substantiated by other investigators. At best, the case for measles virus remains unproven.

Ischaemia

Marked abnormalities of mucosal arterioles have been detected in resected specimens of intestine affected by Crohn's disease by making resin casts of the arterial tree. Many of the small vessels have been shown to be thrombosed, but in a rather patchy distribution, suggesting that much of the inflammation may arise from multifocal infarction. However, as many cytokines and inflammatory mediators released during an immunological or inflammatory response damage endothelium, it is not possible to be sure whether these vascular changes represent a primary abnormality or are merely secondary to the inflammation. It seems more likely that they occur as a consequence of the inflammation but, nevertheless, they may still contribute to the pathogenesis of chronic inflammation.

Immune mechanisms

Patients with Crohn's disease usually have normal serum concentrations of immunoglobulins and complement components, although raised concentrations may occur in association with active disease. Neutrophil and monocyte functions, *in vitro*, show no defect, although inhibitors of cell motility are often present in the serum of patients with active disease. The absolute number of peripheral blood T lymphocytes may be reduced but the proportion of phenotypic subsets (CD4, CD8) remains unchanged.

Within the inflamed tissue, there is a marked increase of plasma cells, lymphocytes, macrophages, and neutrophils. As with ulcerative colitis, the increased immunoglobulin production is predominantly of the IgG isotype, but in Crohn's disease, there is a greater proportional increase in the IgG2 subclass compared with the IgG1 or IgG3 subclasses. Antibodies to bacterial antigens and autoantibodies to epithelial and neutrophil antigens are much less common than they are in patients with ulcerative colitis, but antibodies to *Saccharomyces cerevisiae* (baker's yeast) have been frequently reported, especially with small intestinal Crohn's disease.

T cells and macrophages in the lamina propria are activated, as shown by the increased expression of activation markers and also, in the case of macrophages, by functional assays. The possibility that chronic inflammation results because of a defect in immunoregulation has been explored in a variety of assays but no consistent defect has been described. However, some evidence suggests that there may be an impaired facility to induce antigen-specific suppressor cells; if this is confirmed, it might explain some of the immunological overactivity that is characteristic of the disease. Cellular activation results in the release of cytokines and inflammatory mediators, which will influence the nature of the inflammatory response. Crohn's disease may reflect a predominant T_{H1} response. The possibility that the course of the disease, whether fibrosing or fistulating, may be determined by the cytokine profile is an attractive but unproven hypothesis.

Pathology

Crohn's disease may occur anywhere in the gastrointestinal tract, although the most common pattern is an ileocolitis. The disease is often discontinuous, giving rise to the so-called skip lesions. Isolated involvement of the mouth, oesophagus, stomach, and anus is recognized but such cases are extremely rare. Macroscopically the bowel is thickened and frequently stenosed. The serosal surface may be inflamed and the mesentery becomes oedematous. The regional mesenteric nodes are usually enlarged. The earliest macroscopic lesion on the mucosal surface is an aphthoid ulcer—a small, superficial lesion often surrounded by hyperaemia. In areas of more severe disease, deep, fissuring ulcers occur in the oedematous and inflamed mucosa, giving rise to a cobblestone pattern. Long, serpiginous ulcers are a further characteristic feature. Strictures occur as a result of submucosal fibrosis and, because of serosal inflammation, the affected intestine may become adherent to adjacent loops of intestine or other structures (such as the bladder or vaginal vault) with the subsequent formation of fistulas.

Histologically, the inflammation is transmural and consists principally of lymphocytes, histiocytes (tissue macrophages), and plasma cells. Granulomas are found in only 65 per cent of patients and they occur more commonly the more distal the disease; that is, they are present in most cases with rectal disease but are much less common in ileal disease. The granulomas appear to be in the walls of either blood vessels or lymphatics. The mucosal architecture is well preserved despite heavy inflammation and, in the colon, goblet cells are usually present even though the glands are being infiltrated with inflammatory cells. Fissures, penetrating into the submucosa and lined with histiocytic cells, are frequently present.

Quantitative histological and enzyme studies have suggested that the whole of the gastrointestinal tract is abnormal in patients with Crohn's disease even though only one segment may be overtly involved at any one time.

Immunofluorescent and immunoperoxidase studies have shown a large increase in IgG- and IgM-containing cells with a smaller rise in IgA-containing cells. Even in quiescent disease, the IgG- and IgM-containing cells appear to be increased compared with the normal intestine.

Clinical features

The manifestations of Crohn's disease are protean and are partly determined by the anatomical location of the disease. The majority of patients complain of diarrhoea (70 to 90 per cent), abdominal pain (45 to 66 per cent), and weight loss (65 to 75 per cent). Fever is also common (30 to 49 per cent). Obstructive symptoms (colic, vomiting) are much more commonly associated with ileal disease than colonic Crohn's disease. Colonic disease causes rectal bleeding more commonly than ileal disease, but even so, it is present in only about 50 per cent of patients with Crohn's colitis. Colonic disease is also associated with perianal disease (in about one-third of patients) and with extraintestinal manifestations, which are not commonly seen when the disease is confined to the ileum. Symptoms of anaemia are common and usually occur as a result of iron deficiency from intestinal blood loss or, less frequently, from vitamin B_{12} or folate deficiency. Other features of malabsorption are infrequent, but in patients with extensive small-bowel disease, symptoms and signs of osteomalacia may occur and there may be a bleeding tendency secondary to vitamin K malabsorption. Nutritional deficiencies may also be present, for example deficiencies of magnesium, zinc, selenium, ascorbic acid,

and the B vitamins, but these are uncommon and are usually only seen in patients with diffuse small-intestinal disease.

A few patients present with the clinical features of acute appendicitis but at operation they are found to have an acute terminal ileitis. Only a minority of these prove to be due to Crohn's disease. Diagnostic difficulties may also occur when the disease presents without gastrointestinal symptoms. These include patients presenting with fever, weight loss, and anaemia without diarrhoea or abdominal pain, and those with ileocaecal disease presenting with urinary frequency and dysuria due to ureteric involvement.

Physical examination may be normal but many patients will show evidence of anaemia. Glossitis and aphthous ulcers in the mouth, beaking or frank clubbing of the nails, evidence of weight loss, and a tachycardia are common features. Abdominal examination usually reveals tenderness over the affected bowel, which can often be felt to be thickened. An abdominal mass is frequently palpable when small- intestinal disease is present. Anal examination often shows the presence of fleshy skin tags, which have a characteristic violaceous hue. Anal fissures, perianal fistulas, and abscesses are particularly associated with colonic disease.

The extraintestinal manifestations of Crohn's disease are similar to those of ulcerative colitis. Table 2 lists those that are most frequently seen.

Complications

Patients with Crohn's disease can develop an acute dilatation of the bowel (defined as a colonic diameter of 5.5 cm or more on a plain radiograph), perforation, or massive haemorrhage, especially when the disease involves the colon. However, these complications occur less frequently than they do in ulcerative colitis. The more usual complications are intestinal obstruction due to strictures in the small or large intestine and fistulas. The latter may occur between other parts of the gastrointestinal tract (such as gastrocolic, enterocolic) or between the affected loop of intestine and the bladder or vagina. Pneumaturia, the passage of faeces in the urine, or a faecal vaginal discharge are cardinal features of the latter forms of fistula formation. The gross malabsorption that occurs with a gastrocolic or ileocolic fistula is largely due to bacterial overgrowth of the small intestine. External fistulas to the skin also occur, but this is usually secondary to surgical intervention. Crohn's disease affecting the terminal ileum or the right side of the colon may involve the right ureter, giving rise to frequency with a sterile pyuria, a frank urinary tract infection, or a ureteric stricture with subsequent hydronephrosis. Left-sided disease may occasionally involve the left ureter, but this is very uncommon. Hyperoxaluria and oxalate stones may be complications of ileal disease associated with steatorrhoea. The mechanism is currently thought to be due to binding of calcium to unabsorbed fat, leaving the oxalate free to be absorbed from the colon.

Carcinoma of the colon may complicate Crohn's colitis. The incidence is about 3 to 5 per cent, a frequency similar to that of colonic carcinoma associated with ulcerative colitis. The risk factors are not yet established, however, although histological dysplasia has been noted in some cases of Crohn's disease. Small-bowel carcinomas have been reported in association with ileal Crohn's disease.

Amyloid is another complication of Crohn's disease; it may occur within the bowel or systemically, for example in liver, spleen, and kidney. It usually occurs in patients with poorly controlled Crohn's disease complicated by complex fistulas and abscesses. If renal function is deteriorating, the affected bowel should be resected as the amyloid may then regress with concomitant improvement in renal function.

Radiological appearances

A plain radiograph of the abdomen should always be obtained in patients with severe disease, together with decubitus films. These are often normal but may show evidence of intestinal obstruction or suggest an inflammatory mass in the right iliac fossa. In acute Crohn's colitis, evidence of mucosal oedema and ulceration may be clearly seen on the plain films. This appearance could obviate the need for barium studies, which should, if possible, be avoided in the presence of severe, active disease. The plain film can also provide evidence of sacroiliitis or ankylosing spondylitis.

Examination of the oesophagus, stomach, and duodenum is best done endoscopically because the radiological appearances are often non-specific and biopsies are required for histological confirmation. The small intestine may be examined with a standard barium meal and follow through, but more information is obtained with the barium infusion technique (small-bowel enema, enteroclysis). After colonic preparation, a tube is passed until the tip lies just beyond the ligament of Treitz and a dilute barium suspension is infused (800 to 1200 ml). The earliest lesions are thickening of valvulae conniventes and small, discrete aphthoid ulcers. In more severe disease, cobblestoning, fissure ulcers, and thickening of the wall occur (Fig. 1). Longitudinal ulcers may also occur but these are uncommon. Areas of stenosis and dilatation may be present, and sinus tracts and fistulas may be demonstrated. Asymmetry of the bowel is often present, although this may be an unreliable sign. The abnormal segment of the intestine is usually well demarcated from the normal bowel.

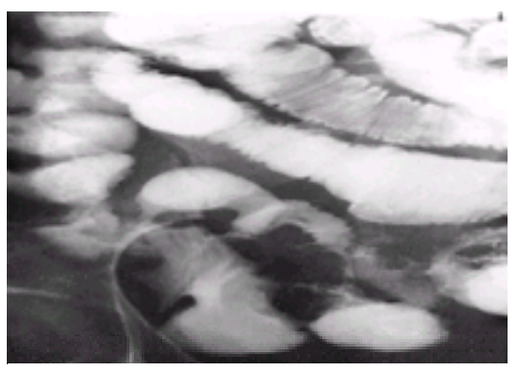

Fig. 1 Small-bowel enema demonstrating Crohn's disease of the terminal ileum with fissure ulcers, ileocaecal fistulas, and partial obstruction. (By courtesy of Dr D.J. Nolan.)

Radiological examination of the colon is made with a double-contrast barium enema after a thorough but gentle preparation. Characteristically there is rectal sparing but the appearances of Crohn's colitis are otherwise similar to those described for the small intestine (Fig. 2). Table 3 lists the main features that differentiate the radiological appearances of Crohn's colitis from ulcerative colitis. The barium enema is a good means of showing internal fistulas and fistulas to other organs.

Fig. 2 Barium enema showing Crohn's disease of the colon and terminal ileum. Distal sigmoid, rectum, and a segment of ascending colon are normal. The diseased segments show loss of haustration, shortening, and fissure ulcers. (By courtesy of Dr D.J. Nolan.)

If fistulas to the surface are present, sinograms should be taken to delineate the anatomy.

CT scans may be helpful in detecting intra-abdominal abscesses and also thickened loops of intestine. Magnetic resonance scans are particularly useful for visualizing peranal and perirectal fistulas, which often migrate between and through the sphincters and arbonize through the levator muscles.

Endoscopy

Sigmoidoscopy and rectal biopsy should be done in all patients. The rectal mucosa is frequently normal but may show a granular proctitis and occasionally the typical appearances of Crohn's disease. Nevertheless, histological examination of a rectal biopsy specimen from a macroscopically normal rectum often shows an inflammatory infiltrate, which is often focal and may contain granulomas. The indications for colonoscopy are: (i) to examine the colon and obtain biopsies in suspected cases where the barium enema is normal or equivocal; (ii) to obtain biopsies from strictures; (iii) to obtain biopsies when the differential diagnosis is in doubt; and (iv) to assess activity and extent of disease in symptomatic patients when there is little clinical evidence of activity. A further advantage of colonoscopy is that biopsies can often be obtained from the terminal ileum.

Endoscopically the earliest lesion of Crohn's disease is a small aphthoid ulcer surrounded by normal mucosa with a normal vascular pattern. This contrasts with the erythema and loss of vascular pattern seen in ulcerative colitis. In more severe disease the mucosa becomes oedematous and is penetrated by fissuring ulcers to give a cobblestone appearance. The ulcers are often linear and may eventually become confluent. A diffusely inflamed, granular, friable, and dark-red mucosa is more typical of ulcerative colitis, although discrete ulceration may occur in severe cases. Pseudopolyps and mucosal bridges occur in both diseases.

Multiple biopsies should be taken, even from apparently normal areas of mucosa, because granulomas may be present, which allows a precise diagnosis to be made.

Upper gastrointestinal endoscopy is not routinely required in these patients and is only indicated in the presence of appropriate symptoms or if abnormalities are noted on a barium meal. Although Crohn's disease of the stomach or duodenum may occur as an isolated phenomenon, most cases are associated with disease elsewhere in the gastrointestinal tract. Deep, longitudinal ulcers may occur in the stomach together with rugal hypertrophy and a cobblestone appearance. In the duodenum the major differential diagnosis is duodenal ulcer, but there is usually a 'cobblestone' mucosa surrounding the frank ulceration. Biopsies are usually helpful, although granulomas are found infrequently.

Examination of the small intestine is now possible using a push-type or sonde-type enteroscope (see Chapter 14.3.1). This is an expensive, lengthy procedure and, for these reasons, it is unlikely to become widely available. Nevertheless, it may be helpful for patients in whom the diagnosis is suspected but in whom the small-bowel enema is not diagnostic.

Laboratory data

Anaemia is common and is often due to mixed deficiencies. Iron deficiency from intestinal blood loss is the most common cause but serum folate and vitamin B_{12} concentrations may also be low. The blood film and mean corpuscular volume may therefore show microcytosis or macrocytosis. Serum ferritin is the best indicator of iron stores in those patients with chronic disease. A neutrophil leucocytosis is usually, but not invariably, associated with active disease and there may also be a thrombocytosis. The total lymphocyte count and the absolute number of circulating T lymphocytes may be reduced.

Hypokalaemia is associated with severe diarrhoea and the plasma urea concentration is often low, reflecting a poor dietary intake of nitrogen. Serum albumin is reduced in the presence of active disease, largely due to down-regulation of albumin synthesis by cytokines such as interleukin 1 (**IL-1**), tumour necrosis factor, and IL-6, but studies using albumin labelled with chromium-51 often demonstrate a protein-losing enteropathy. Serum immunoglobulins are normal or mildly elevated and there may be a rise in the a_2-globulins. A low serum calcium, when corrected for albumin, is unusual unless there is extensive small-bowel disease, and a low urinary calcium is more likely to reflect a poor diet rather than osteomalacia. Liver function tests are frequently abnormal, usually consisting of mild elevations of the aspartate transaminase and alkaline phosphatase. Persistence of abnormal liver tests suggests associated liver disease and should be investigated by liver biopsy and visualization of the biliary tree. Patients with extensive ileal disease or with ileal stricture may have increased faecal fat excretion. This is usually secondary to bacterial overgrowth rather than loss of absorptive surface, and is compounded by the low circulating pool and increased excretion of bile salts, which is often present in patients with long-standing ileal disease. It is important not to miss magnesium, zinc, and selenium deficiencies, which are occasionally present.

Diagnosis

This may be delayed for several years. Intermittent abdominal symptoms and diarrhoea without systemic symptoms are often labelled as an irritable bowel syndrome. Weight loss, fever, and anaemia without gastrointestinal symptoms are another source of misdiagnosis. The diagnosis of Crohn's disease in children may be considerably delayed when it presents as failure to thrive or delayed puberty but without gastrointestinal symptoms.

Even when the clinical diagnosis seems sound, all patients must have: (i) stool examination to exclude pathogens; (ii) sigmoidoscopy and rectal biopsy—characteristic features (such as granuloma) may often be present in the biopsy specimen even when the mucosa is macroscopically normal; (iii) radiographs of the small and large intestine to confirm the diagnosis and establish the extent of the disease; and (iv) colonoscopy with multiple biopsies is indicated where the above investigations are equivocal or normal and there are strong clinical reasons for suspecting Crohn's disease. Colonoscopy should also be done if the differential diagnosis is in doubt or if strictures are present.

Differential diagnosis

Few patients with an acute ileitis and a clinical picture of acute appendicitis subsequently develop Crohn's disease. Serological examination helps to diagnose those cases caused by yersinia; the aetiology of the remainder is unknown. The main differential diagnosis of ileal Crohn's disease is tuberculosis, especially when the disease occurs in patients from areas where intestinal tuberculosis is common. Laparoscopy may be helpful if serosal tubercles are present, as biopsies can be taken from them and cultured. Stool culture and circulating antibodies to mycobacteria are unhelpful. Colonoscopy with multiple biopsy specimens may be helpful. If genuine doubt exists, corticosteroid therapy for Crohn's disease must be covered with antituberculous therapy. Other differential diagnoses include abdominal lymphoma, a-chain disease, actinomycosis, amyloid, Behçet's disease, and carcinoma of the small bowel.

The major differential diagnosis of Crohn's colitis is ulcerative colitis (Table 3). Crohn's disease should also be considered in patients presenting with proctitis, as 30 per cent of patients with ileal Crohn's disease may have a proctitis and may present in this way. When a segmental colitis occurs, ischaemia, tuberculosis, and lymphoma have to be excluded. Young adults may present with an acute segmental colitis, which is self-limiting. The cause is unknown, although in women, oral contraceptives have been implicated. Crohn's disease can be overlooked on the barium enema when it occurs in association with severe diverticular disease.

As indicated above, Crohn's disease may have to be considered in the differential diagnosis of a fever with weight loss, malabsorption, and delayed development.

Assessment of activity

There is no satisfactory method of assessing activity of the disease and this poses a major clinical problem. Symptoms such as fever or continuing weight loss are obvious indicators, but severe disease can be present in the absence of any major symptom. Laboratory evidence of activity includes a reduced serum albumin, and a rise in acute-phase reactants (C-reactive protein, orosomucoid) and in the erythrocyte sedimentation rate. Recently, a number of activity indices have been developed in order to standardize assessment for the purpose of multicentre studies, two examples being the American Crohn's disease activity index and the Dutch activity index. However, they are mostly too complex for normal clinical use. Furthermore, they tend to measure different aspects of disease activity. At present it is worth remembering that disease activity can be assessed by the clinical picture (symptoms and signs), morphologically (for example by radiography or endoscopy), and by laboratory indices. Another technique for the assessment of activity is the use of indium-labelled neutrophils. The labelled cells preferentially migrate to inflamed mucosa and the increased uptake of isotope can be detected using a g-camera (Fig. 3). Faecal excretion of the labelled cells can also be quantified and this has shown good correlation with the Crohn's disease activity index and albumin loss.

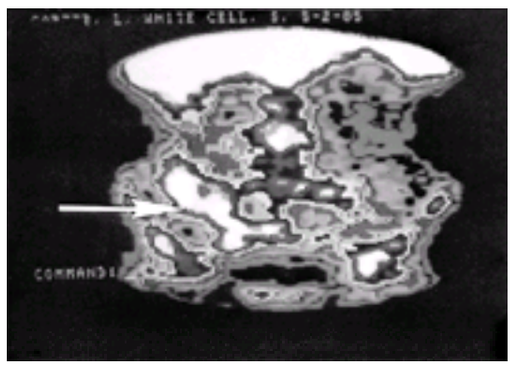

Fig. 3 A neutrophil scan labelled with indium-111 in a patient with active Crohn's disease. Active inflammation in distal ileal loops is well shown (arrow).

Labelling neutrophils with technetium-99 using hexamethyl propylene amine oxime (HMPAO) as a chelator is gradually replacing indium because it is easier, quicker, and less expensive. It appears to provide similar sensitivity and specificity but it cannot be used to assess faecal excretion of neutrophils.

Management

The management of Crohn's disease involves a team approach between physician and surgeon and includes nutritional support, medical therapy, and surgical treatment.

Nutritional support

For most patients, a well-balanced diet should be advised. A low-residue diet should be used for patients with strictures and a low-fat diet may be helpful for those with a steatorrhoea. Some centres claim good long-term results with elimination diets, but the value of this approach for all patients is far from proven. A lactose-free diet is obviously indicated for those with hypolactasia.

Iron and vitamin B_{12} deficiencies are common and need to be excluded. In patients with chronic active disease, the serum iron, binding capacity, and ferritin may all be low, making it difficult to know whether the anaemia is due to iron deficiency or to chronic inflammation. However, if the serum iron is less than 10 per cent of the iron-binding capacity it is reasonable to diagnose iron deficiency and treat accordingly. Many patients are intolerant of oral iron and for these a total-dose intravenous infusion is the best form of treatment. Patients with extensive small-intestinal disease may develop deficiencies of the fat-soluble vitamins (A, D, E, K). Deficiencies of folic acid and B vitamins may also occur because of poor dietary intake. Parenteral nutrition is often indicated for seriously ill patients who are being prepared for surgery or who have a short bowel syndrome.

Drug therapy

Active disease

In general, Crohn's disease is only treated if it is causing symptoms; there is no indication for treatment in asymptomatic patients. Active disease can usually be controlled with corticosteroids, the dose and route depending on the severity of the disease. Severe disease requires admission to hospital and treatment with intravenous prednisolone (60 to 80 mg daily) or hydrocortisone (400 mg daily), together with fluids and electrolytes. Most patients settle within 5 to 7 days and can then be given prednisolone orally (for example 40 mg daily). Patients with less severe disease can be treated with 20 to 40 mg of prednisolone daily. There is no defined duration of corticosteroid therapy but most patients will have made a good symptomatic response by 4 to 6 weeks; the dose can then be reduced over the next 3 to 6 weeks and finally stopped.

Active disease can also be treated with liquid diets. Elemental diets have been repeatedly shown to be as effective as prednisolone in controlling active Crohn's disease, but the major problem is one of compliance. Some patients intubate themselves at night with a fine-bore nasogastric tube and feed themselves during sleep. Polymeric diets may be as effective as elemental diets and are certainly more palatable. If an individual patient responds well and is willing to carry on with the diet, it should be continued until symptoms have settled and the laboratory indicators of inflammation have returned to normal. Experience then differs as to the course of the disease once normal food is introduced. Some data suggest that relapse occurs more rapidly than if remission has been achieved by the use of corticosteroids.

There is some evidence that antibiotics can be useful in treating active disease (for example metronidazole, ciprofloxacin, or clarithromycin), but the controlled trials are small and many are only published in abstract. However, they are often indicated to treat bacterial overgrowth, perianal sepsis, or abscesses associated with fistulas. Randomized controlled trials of antimycobacterial regimens have failed to confirm the optimism of anecdotal series.

The role of 5-aminosalicylic acid drugs for treating active disease is unclear. Using high doses of mesalazine (4 g daily), the initial trial showed benefit after 12 to 16 weeks but this has not been confirmed in subsequent trials. Meta-analysis of all trials shows no overall benefit compared with placebo.

Chronic active disease

There is good evidence from several trials, supported by meta-analysis, that azathoprine (2.0 to 2.5 mg/kg) or 6-mercaptopurine are effective treatment for steroid-dependent or steroid-resistant disease, although they act slowly and it may take up to 16 weeks before an effect is seen. Furthermore, both drugs are effective maintenance therapy in those patients who respond well. Retrospective studies suggest they maintain their effect over 4 to 5 years without inducing long-term side-effects. About 15 per cent of patients are unable to tolerate the drug (mostly because of gastrointestinal symptoms, myalgia, and occasionally hepatitis). Acute pancreatitis is rare. Bone marrow suppression is also rare and is confined to those who are homozygous for a deletion in the gene encoding the thiopurine methyl transferase enzyme (approximately 1 in 300 of the population). Interestingly, patients who develop nausea, diarrhoea, or myalgias on azathioprine are frequently able to tolerate 6-mercaptopurine. For patients failing to respond, or who are intolerant of these drugs, methotrexate may induce a response in two-thirds as shown by a clinical trial and by open series. In the trial, 25 mg of methotrexate by weekly intramuscular injection was used, but open series suggest a similar response when it is given orally. Recently, monoclonal antibodies to tumour necrosis factor have been developed. Infliximab is now licensed as a single infusion for chronic active disease unresponsive to steroids and immunosuppressants. It is an IgG1 chimeric antibody containing 75 per cent human immunoglobulin. A dose of 5 mg/kg by slow intravenous infusion over 2 h appears to give maximal response and this occurs in about 70 per cent of cases. Endoscopic healing can be shown but the response is usually temporary and, after a few months, patients begin to relapse. Side-effects are few (nausea, headaches, and rarely, infusion reactions can occur) but about 12 per cent of patients develop antibodies to double-stranded DNA, probably secondary to apoptosis induced by the antibody.

Maintenance of remission

There is no good maintenance therapy for Crohn's disease. The 5-aminosalicylic acid drugs have very little effect if remission has been induced by medical therapy, although high doses (3 g daily) can delay recurrence following surgical resection. However, eight such patients need to be treated to prevent one recurrence over a 5-year period. As described above, azathioprine and 6-mercaptopurine are effective long-term in steroid-dependent patients who make a good initial response to immunosuppression. Undoubtedly the best maintenance therapy is to stop patients smoking—that is associated with a more benign course and lower recurrence rates following surgical resection.

Osteoporosis

Many patients at the time of diagnosis have reduced bone density (chronic inflammation, poor nutrition) and this can be exacerbated subsequently if frequent courses of corticosteroids are required to control active disease. Dual energy X-ray absorptiometry (DEXA) scans should be obtained in patients with long-standing

troublesome disease, and appropriate treatment begun (see Chapter 19.4).

Surgery

The majority of patients (70 to 80 per cent) will require at least one operation during the course of their disease. Indications for surgery include failure to respond to medical therapy, strictures causing mechanical obstruction, fistulas, and other complications such as abscess and perforation.

If surgery is required, the following principles apply. Resection should be limited to removing the most severely affected segment and an end-to-end anastomosis should be made, even if there is some inflammation in the tissue being anastomosed. Wide resections have not been shown to diminish the subsequent recurrence rate. Bypass procedures (such as ileotransverse colostomy) should not be done. If surgery is for internal fistulas, nutrition must be corrected, infection controlled, and active disease controlled with steroids before the operation. The anatomy of the fistulas must also be determined by sinograms or by CT scans or MRI. The fistula is excised together with the segment of affected intestine and the subsequent anastomosis is usually best protected with a temporary ileostomy. For colonic disease, the choice is a conservative operation of a split ileostomy or a proctocolectomy with terminal ileostomy. Colectomy with ileorectal anastomosis is associated with a high recurrence rate. Defunctioning the colonic disease with an ileostomy often allows the disease to settle and the patient's nutritional state to be restored. In Oxford, our practice is to reconnect after 12 to 18 months and the majority of patients then remain well. However, other clinicians claim a rapid relapse following restoration of continuity and will only use a split ileostomy as a means of getting patients fit for more radical surgery.

Perianal disease is common, both in ileal and colonic Crohn's disease. It may consist of simple fissures but, more frequently, consists of complex fistulas with abscess formation. Management has to include demonstration of the anatomy (e.g. a pelvic MRI scan), drainage of abscesses, control of infection and suppression of disease activity. The use of setons has been a major advance in allowing fistulas to heal by providing effective drainage and they can be kept in for months. Immunosuppression with thiopurines may allow some degree of healing but, recently, infliximab has been highly effective. In a randomized clinical trial, two-thirds of patients assigned to infliximab closed their fistulas, significantly more frequently than a placebo infusion. However, MRI scans have shown persistent fistula tracks even though the cutaneous orifice heals. Thus, it is not surprising that the apparent 'healing' is usually only temporary. Nevertheless, infliximab (usually an infusion of 5 mg/kg) at 0, 2, and 6 weeks) has been a major advance in the overall management of fistulas.

Some patients with small-intestinal disease present with multiple short strictures. Once active disease is controlled with corticosteroids, the strictures should be dealt with by stricturoplasty rather than by multiple resections. This gives good symptomatic relief and it is unusual for further stricturing to occur at sites of previous stricturoplasties. This conservative approach has greatly reduced the need for repeated resections and has therefore minimized the chances of a short bowel syndrome.

Management during pregnancy

Crohn's disease should be treated in the pregnant woman along the lines outlined above. Overall, the outcome of the pregnancy is not influenced by the disease except in very severe cases where there may be an increased risk of abortion. Corticosteroids and sulphasalazine are safe to use and have not been associated with fetal abnormalities. Likewise, azathioprine has not been clearly demonstrated to be teratogenic and can be used if there is sufficient clinical indication. Methotrexate is known to be teratogenic and is therefore completely contraindicated for women who are trying to conceive or who are pregnant.

Management in children

There is no essential difference in the principles of management from those described for adults, although dosages may need to be reduced. Alternate-day steroids should be employed, especially if long-term treatment seems likely. Excellent but uncontrolled results have been reported in adolescents using maintenance corticosteroids, as an alternate-day regimen, which allowed puberty and growth to develop normally. One of the major effects of the disease in children is growth retardation. Corticosteroid therapy often promotes a growth spurt but great emphasis should be paid to the child's nutrition. Dietary intake should be assessed and supplemented to provide a high-calorie, high-nitrogen intake. Many paediatric gastroenterologists use elemental or polymeric diets as first-line therapy for active disease, often administered via a percutaneous gastrostomy.

Course and prognosis

Patients are never cured of Crohn's disease and they are subject to relapses of their disease and to recurrence following surgical resection. Most patients (70 to 80 per cent) will receive surgical treatment at some point during the course of their illness. After a resection, the disease recurs in about 30 per cent of patients during the subsequent 5 years and in 50 per cent of patients during the subsequent 10 years; of these, half will require further surgery. Although there is still some controversy, the balance of evidence suggests that the risk of requiring second or third operations is no greater than the risk of requiring the initial operation. Patients with Crohn's colitis who have a proctocolectomy appear to have a lower risk of recurrence than those who have an ileal or ileocolic resection.

Recent endoscopic visualization of the neoterminal ileum has shown that the recurrence rates, when assessed by endoscopic appearance, are even higher than the rates quoted above, which are based on symptoms. For patients who have had an ileal or ileocolic resection, 70 to 80 per cent of them will show endoscopic lesions just proximal to the anastomosis within the first postoperative year. The more severe lesions, such as aphthoid ulcers, predict a high chance of symptomatic recurrence. Mesalazine has been shown to have no influence on the endoscopic recurrence rate and it remains to be seen whether steroid compounds with low systemic bioavailability (such as budesonide) will be effective in this context without inducing systemic side-effects.

The overall mortality of Crohn's disease varies from 10 to 15 per cent in different studies. Some of these reports have suggested a worse prognosis for women than for men, and for patients over the age of 50 years, although this was mainly associated with higher operative mortality. Overall, age and gender probably have little influence on the outcome of the disease. The Oxford experience has suggested that mortality is not appreciably increased during the first 15 years of the disease but then becomes progressively greater during subsequent follow-up. In contrast, however, data from Birmingham suggest that the highest mortality occurs in young people during the early stages of the disease.

In general, most patients with Crohn's disease will have a good prognosis with a mortality of only about twice that expected. Considerable morbidity can be expected but this will be intermittent and the overall quality of life should be good.

Further reading

Allan RN *et al.* (1997). *Inflammatory bowel diseases*, 3rd edn. Churchill Livingstone, Edinburgh.

Jewell DP, Warren BF, Mortensen NJ (2001). *Inflammatory bowel disease.* Blackwell Science, Oxford.

Kirsner JB (2000). *Inflammatory bowel disease*, 5th edn. WB Saunders Co., Philadelphia.

14.11 Ulcerative colitis

D. P. Jewell

Ulcerative colitis is a chronic inflammatory disease of the colon of unknown cause. It always affects the rectum and extends proximally to involve a variable extent of the colon. It is characterized by a relapsing and remitting course.

The disease was first described in 1859 by Samual Wilks, a physician at Guy's Hospital, who recognized that 'simple, idiopathic colitis' could be distinguished from other forms of colitis, mainly bacterial dysentery. It took many years for the concept to be accepted, but finally, in 1931, Sir Arthur Hurst was able to give a complete description of the disease including the sigmoidoscopic appearances. Nevertheless, he still considered the disease to be primarily infective, even though its chronic nature might be induced secondarily by other factors.

Epidemiology

Ulcerative colitis is a worldwide disease, although it may be difficult to diagnose in areas where infective colitis is prevalent. Accurate figures for incidence and prevalence are not universally available but the disease is now recognized in most countries. Table 1 lists data for the high-incidence areas and also shows that there have been no trends to suggest the disease is becoming more common, which is in contrast to Crohn's disease. The low-incidence areas include Eastern Europe, Asia, Japan, and South America where the incidence rates are at least tenfold less.

The age of onset peaks between 20 and 40 years but the disease may present at all ages from the first few months of life to the 80s. Some series show a secondary peak of onset in the 60- to 70-year-old age group, but this has not been a universal finding. Earlier series suggested a predominance of the disease in women, but more recently, there has been little difference between the sexes.

Both in the United States and Cape Town, Jews are more prone to ulcerative colitis than non-Jews by a factor of 3 or 4. Within Israel, Ashkenazi Jews have a higher incidence than Sephardim but it is still less than the incidence in Jews in the United States or, indeed, than the European incidence. This suggests that environmental factors may be involved in addition to genetic factors. However, the differences in incidence between urban as opposed to rural communities or between different socio-economic groups have been slight and inconstant.

Genetics

The familial incidence of ulcerative colitis has long been recognized, with 10 to 20 per cent of patients likely to have at least one other family member affected either with ulcerative colitis or with Crohn's disease. Most of the familial association is within first-degree relatives, but there is controversy about the precise relation, with a preponderance of parent–sibling combinations being found in the United States, whereas in the United Kingdom the disease is more commonly shared by siblings. Within a multiply-affected family there is a high degree of concordance for disease characteristics (for example extent, severity, presence of extraintestinal manifestations).

A study of twin pairs in Sweden showed that of 16 pairs of monozygotic twins in whom one member had ulcerative colitis, only one pair was concordant for the disease whereas all 20 dizygotic twins were discordant. This gave a proband concordance rate of 6.3 per cent, which is very much lower than 45 per cent for Crohn's disease.

This low concordance rate might suggest that familial clustering reflects environmental influence rather than inherited genetic susceptibility. However, the incidence of ulcerative colitis in spouses of probands is extremely low, although, of course, that does not exclude environmental factors operating early in life.

The mode of inheritance is unknown, but as with Crohn's disease, multiple genes are probably involved in determining disease susceptibility and its behaviour. Studies of multiply-affected families, using microsatellite technology, have demonstrated linkage to chromosome 12 and, less strongly, to chromosomes 3 and 7. These susceptibility loci are shared by patients with Crohn's disease although some evidence suggests that the locus on chromosome 12 contains two separate genes, one for ulcerative colitis and a second for Crohn's disease. In most studies, ulcerative colitis is more strongly linked to chromosome 6p (the HLA region) than Crohn's disease. Study of individual HLA alleles has shown that possession of HLA DR103 is likely to be associated with severe disease. In Japan and in the Jewish population of California, the disease is associated with HLA DR1502 (an allele of DR2 common in these populations but very uncommon in Europeans). The occurrence of extraintestinal manifestations also appears to be related to genetic make-up. For example, patients who develop a reactive, large joint arthropathy in association with active disease are likely to possess the HLA DR103 allele (35 per cent) compared with patients who do not (8 per cent) or healthy controls (3 per cent). In contrast, the small joint, seronegative arthropathy is associated with HLA B44 (77 per cent) and MICA-8 (98 per cent). Nevertheless, it is not possible to be

sure on the basis of present knowledge whether these associations are biologically meaningful or whether they represent linkage disequilibrium with a nearby gene.

Aetiology

The cause of the disease remains unknown. The main hypotheses that have been proposed include infection, allergy to dietary components, immune responses to bacterial or self-antigens, an abnormality in epithelial cell integrity, and the psychosomatic theory. There are virtually no data to support a primary role for psychosomatic factors in the aetiology of the disease, although they may play a secondary role in determining the pattern of symptoms and must always be considered when managing individual patients.

Infection

No specific infective organism has been consistently isolated from patients with ulcerative colitis. However, the recognition that the strains of *Escherichia coli* in the normal colon are continually changing has led to the concept that patients may carry strains which, by releasing enzymes or other toxic products, might damage the mucosa. The demonstration that, even in remission, patients with ulcerative colitis are more likely to harbour *E. coli* expressing adhesins than control subjects is a particularly interesting observation, as these may allow the bacteria to adhere readily to the epithelium. The role of sulphate-reducing bacteria is also of interest as these organisms are found more commonly in those with colitis. They reduce sulphate to sulphide which, in turn, inhibits butyrate oxidation in epithelial cells. Several investigators have demonstrated reduced activity of butyrate dehydrogenase within colitic epithelium, even in remission, raising the possibility that luminal bacteria may have a deleterious effect on epithelial cell metabolism and, hence, integrity.

Food allergy

The early suggestions of allergic responses to milk proteins, eggs, and other dietary proteins have not been substantiated as an aetiological factor. Milk-free diets may be beneficial in a minority of patients but it is not clear whether this results from an associated hypolactasia, an immunological response, or some other mechanism. The failure of ulcerative colitis to respond either to intravenous nutrition avoiding oral food or to colonic isolation by means of a split ileostomy are further pointers that dietary factors play little part.

Environmental factors

As well as infection and diet, smoking and the use of oral contraceptives may influence disease. Many studies have now shown that ulcerative colitis is more common in non-smokers than smokers, with a relative risk of 2 to 6. Ex-smokers have a particularly high incidence and this is highest for former heavy compared with light smokers. Women taking oral contraceptives may have a slightly increased risk of the disease but this association is weak and loses significance when the data are corrected for smoking habits and social class.

Immunopathogenesis

The intense infiltration of the inflamed mucosa with plasma cells, B and T lymphocytes, and macrophages suggests immunological activity. Whether activation of both humoral and cellular immune mechanisms merely reflects increased antigenic absorption through an abnormal epithelium, a response to a specific aetiological agent, or an underlying defect in mucosal immunoregulation is unknown.

There is an increase in plasma cells synthesizing all three of the major immunoglobulin isotypes—IgA, IgG, and IgM. However, the largest proportional increase is in IgG-producing cells and this is predominantly of the IgG1 and IgG3 subclasses, which is in contrast to Crohn's disease where an IgG2 response is predominant. IgG1 and IgG3 are synthesized in response to protein antigen and are effective in fixing complement. Complement activation is known to occur in active colitis, probably as a result of the formation of antigen–antibody complexes, and is likely to be one of the principal effector mechanisms in establishing the inflammatory lesion. Some of the increased mucosal IgG synthesized is known to have antibody specificity for bacterial and epithelial antigens. As antibody to epithelial antigens, especially a 40-kDa protein, is a feature of ulcerative colitis, rather than Crohn's disease, it is possible that autoimmunity plays a part in ulcerative colitis. This concept is strengthened by the association with other autoimmune disorders and with circulating antibodies to neutrophils (pANCA), neither of which is associated with Crohn's disease. Nevertheless, whether anticolon antibodies or pANCA have a pathogenetic role is still very uncertain.

The main subsets of T cells (CD4+, CD8+) are present in increased numbers in the inflamed mucosa but their proportions do not change significantly. Several lines of evidence suggest that the T cells are activated and release a variety of cytokines. Whether there is a failure of T cells to either upregulate or downregulate the mucosal immune response has not been clearly shown. However, data suggest that there may indeed be a failure to induce suppression to specific antigens, which could lead to some of the immunological overactivity that is observed in this disease. Intraepithelial T lymphocytes isolated from colons resected for severe ulcerative colitis also fail to suppress T-cell proliferative responses to specific antigens, a property that is not due to the increased numbers of intraepithelial T lymphocytes using g† T-cell receptors.

As well as T-cell activation, there is also a marked increase in the population of activated macrophages, which not only release inflammatory mediators (reactive oxygen metabolites, leukotrienes, platelet-activating factor) but serine proteases, metalloproteinases, and cytokines. The release of interleukin 1 (**IL-1**), IL-6, and tumour necrosis factor will not only lead to tissue damage but will initiate an acute-phase response, downregulate albumin synthesis, and induce fever. Release of interferon-g from activated T cells induces HLA class II molecules on colonic epithelial cells, which, in turn, are then able to present antigen to the adjacent CD4+ lymphocytes and to activate the CD8+ intraepithelial T lymphocytes. Changes in epithelial permeability induced by interferon-g and inflammatory mediators, endothelial damage by a wide variety of cytokines and mediators leading to local ischaemia, and stimulation of collagen synthesis by transforming growth factor-b, IL-1, and IL-6 may all contribute to the inflammatory process.

Pathology

Macroscopic

Ulcerative colitis always involves the rectum but in about 40 per cent of patients the disease is limited to the rectum and sigmoid. In adults, only about 20 per cent will have the whole colon involved, although this proportion rises to about 50 per cent in children. In mild disease, the mucosa is hyperaemic and granular, but as the disease becomes more severe, small punctate ulcers appear, which may then enlarge and extend deeply into the lamina propria. The ulceration may be linear along the line of the taeniae coli. The mucosa can become intensely haemorrhagic. In patients with long-standing disease, inflammatory polyps (pseudopolyps) may develop. They are usually found in the colon and rarely in the rectum. Inflammatory polyps are of no significance and have no malignant potential. In occasional patients, they may regress.

When the disease goes into remission, the colonic appearances may return to normal, but, especially in patients who have had recurrent attacks, the mucosa becomes atrophic and featureless. There is often narrowing and shortening of the bowel. Fibrous strictures complicating long-standing chronic disease are extremely rare.

If an acute dilatation occurs in a patient with severe disease, the bowel becomes thin and congested. There is usually severe ulceration, with only small islands of mucosa remaining. An acute dilatation may be accompanied by a perforation.

Microscopic

The inflammation of ulcerative colitis is largely confined to the mucosa. The lamina propria becomes oedematous, with dilated and congested capillaries, and extravasation of red blood cells. There is a cellular infiltrate of acute and chronic inflammatory cells: neutrophils, lymphocytes, plasma cells, macrophages, mast cells, and eosinophils.

The neutrophils invade the epithelium, usually in the crypts, giving rise to a cryptitis and eventually to a crypt abscess. The triggers for this migration of neutrophils are unknown, but chemotactic peptides of colonic bacteria (for example formyl methionyl leucyl phenylalanine) as well as IL-8, leukotriene B4, platelet-activating factor, and activated complement are potential candidates. Damage to the crypts leads to increased epithelial cell turnover and a discharge of mucus from goblet

cells. With increasing inflammation, the surface epithelial cells become flattened, irregular, and eventually ulcerate. Deep ulcers may extend into the lamina propria, leading to inflammatory changes in the submucosa—this may be accompanied by an acute dilatation or perforation.

Many of the acute changes of ulcerative colitis are non-specific and may also be seen in infective colitides. However, the diagnosis of ulcerative colitis can be made with some accuracy (more than 80 per cent probability) if features of a chronic inflammatory process are present. These include distorted crypt architecture, crypt atrophy, basal lymphoid aggregates, and a chronic inflammatory infiltrate.

Once the disease has gone into remission, the histological appearances may return to normal. However, there is frequent evidence of bifid or shortened crypts, hyperplasia of the muscularis mucosae, neuronal hypertrophy, and Paneth-cell metaplasia at the base of the crypts.

Clinical features

Patients usually present with a gradual onset of symptoms, often intermittent, but which become progressively more severe. Occasionally, ulcerative colitis can present much more rapidly and may mimic an infective colitis. Indeed, some patients begin with a documented infection (such as a campylobacter or salmonella colitis) but continue to have symptoms that ultimately lead to the correct diagnosis.

The principal symptoms include diarrhoea, rectal bleeding, the passage of mucus, and less frequently, abdominal pain. When the inflammation is confined to the rectum (proctitis), patients often pass fresh blood, which is usually mixed with the stool but can be streaked on the surface. These patients often complain of constipation rather than diarrhoea and, on clinical symptoms alone, may be mistakenly diagnosed as suffering from haemorrhoids. When the inflammation extends beyond the rectum, there is usually diarrhoea with the passage of partly altered blood. The diarrhoea is often accompanied by urgency and tenesmus, and patients can be incontinent. Nocturnal diarrhoea is a common symptom in the presence of severe inflammation. With a severe ulcerative colitis affecting most or all of the colon, patients are usually anorexic, nauseated, and have lost weight. They usually have severe diarrhoea (in excess of six motions daily) that becomes a slurry of faecal material, pus, and blood—it may resemble anchovy sauce and, indeed, some patients may fail to recognize that they are passing blood.

Patients may also complain of malaise, lassitude, and symptoms referable to chronic iron deficiency or to some of the extraintestinal manifestations, especially recurrent aphthous ulcers of the mouth.

On examination, patients with mild or moderate attacks usually look well and exhibit few abnormal physical signs. Weight should always be recorded and, for children and adolescents, both height and weight should be recorded on growth charts. Abdominal examination may reveal a tender colon but is often normal. Bowel sounds are normal and rectal examination is also normal apart from blood.

Patients with a severe attack may also look deceptively well and a tachycardia or a tender colon may be the only abnormal signs. However, many of these patients are obviously ill, with fever, salt and water depletion, anaemia, and evidence of weight loss. There may be oral candidiasis, aphthous ulceration, signs of iron deficiency, and finger clubbing. The skin changes of hypoalbuminaemia and dependent oedema may occur. The abdomen is often distended and tympanitic, with reduced bowel sounds and marked colonic tenderness.

Minor perianal disease, such as a fissure, may occur in patients with an active ulcerative colitis but it is never as severe as is seen in patients with Crohn's disease.

Assessment of disease severity

This can be done clinically, by grading the degree of inflammation seen endoscopically or histologically, and by using laboratory tests of inflammatory activity.

Clinical grading

1. Mild—there are less than four stools daily, with or without blood, with no systemic disturbance and a normal erythrocyte sedimentation rate.
2. Moderate—this is between mild and severe.
3. Severe—there are at least six stools daily, with bleeding, and evidence of systemic illness as shown by fever, tachycardia, a falling haemoglobin, hypoalbuminaemia, and raised erythrocyte sedimentation rate and C-reactive protein.

Laboratory markers of inflammation

Active disease is often accompanied by a neutrophil leucocytosis, thrombocytosis, and a rise in acute-phase proteins (C-reactive protein, oroso-mucoid) and in erythrocyte sedimentation rate. There may also be a fall in haemoglobin and albumin levels. These inflammatory markers are useful when measured serially during the course of treatment as an indicator of disease activity. However, if corticosteroids are used, the white cell count can no longer be used as a marker of disease activity because it will often rise in response to the steroids. Patients with a proctitis rarely have a rise in C-reactive protein unless the inflammation is particularly severe.

Diagnosis

The diagnosis is made on the basis of the history, the absence of faecal pathogens, and the endoscopic and histological appearances of the colon.

Stool cultures should be set up for all patients presenting for the first time and, ideally, for all those presenting with a relapse of established disease. Special culture conditions are required for campylobacter, yersinia, gonococci, and *Clostridium difficile*. The possibility of an infection with *E. coli* 0157 must also be considered, especially in patients in whom bleeding and abdominal pain are predominant symptoms. An infective colitis with opportunistic organisms in patients with immunodeficiency syndromes has become much more common and has to be remembered in differential diagnosis.

Sigmoidoscopy is safe, even in patients with a severe attack, and not only confirms rectal inflammation but also allows a biopsy specimen to be taken and an assessment of severity to be obtained. Although some centres use colonoscopy in severe attacks, this is rarely necessary for diagnosis, for assessment of severity, or for determining management. It is best avoided in the acute stage. The earliest signs of colitis on sigmoidoscopy are blurring of the vascular pattern associated with hyperaemia and oedema, leading to blunting of the valves of Houston. With increasing severity, the mucosa becomes granular and then friable. With severe inflammation, the mucosa shows spontaneous bleeding and ulceration. These changes begin in the rectum, they are diffuse, and extend proximally to affect a variable length of the colon. Pseudopolyps (inflammatory polyps) often occur in patients with long-standing disease but tend to be in the colon rather than the rectum.

Colonoscopy with multiple biopsies is useful for assessing the extent of disease and is mandatory for patients with a colonic stricture. It is also required for cancer surveillance (see later). Preparation of the colon should follow the normal methods and osmotic purgation is the most satisfactory. However, a more gentle approach is needed if colonoscopy is done in the presence of severe inflammation, but this is rarely indicated.

All patients with a severe attack must have a plain abdominal radiograph. Not only does this exclude a dilated colon but it may provide prognostic information (mucosal islands, distended small bowel loops) and demonstrate the extent of the disease. An abnormal haustral pattern, thickening of the bowel wall, and mucosal oedema can be detected on a plain film (Fig. 1). As an inflamed colon does not hold faecal material, the presence of faecal matter in the ascending or transverse colon will indicate that the inflammation is distal. In a severe attack, barium radiography is virtually never indicated, but if it is done, a single-contrast study in an unprepared colon with barium entering the colon at low pressure should be used. In less severe disease, a double-contrast barium enema can be safely given (Fig. 2), but the colon must not be overdistended and the procedure must be stopped if the patient complains of pain.

Fig. 1 Plain abdominal radiograph of a 24-year-old man with severe ulcerative colitis. The ascending and transverse colon are grossly oedematous and diseased with loss of the normal haustral pattern. In addition, there are multiple loops of distended small intestine.

Fig. 2 A double-contrast barium enema in a patient with active ulcerative colitis. The figure is a close-up view of the splenic flexure to show extensive mucosal ulceration, loss of haustration, and narrowing of the colon. The patient also has diverticula in the descending colon.

Biopsy specimens must be taken at sigmoidoscopy or colonoscopy, preferably with small, cupped forceps. Histological assessment contributes to grading severity as well as the differential diagnosis.

Laboratory data

These are required for assessing severity, as discussed above, and to document haematological or biochemical complications.

Iron deficiency is common as a result of chronic iron loss; this can be exacerbated by a severe attack, in which 0.5 g of elemental iron can be lost. Thus, a hypochromic, microcytic anaemia is frequently present. A neutrophil leucocytosis, thrombocytosis, eosinophilia, or monocytosis may also be present and are indicators of active inflammation.

Biochemical abnormalities are rare in mild or moderate attacks, but hypokalaemia, hypoalbuminaemia, and a rise in a $_2$-globulin frequently accompany a severe attack. Minor elevations of the aspartate transaminase or alkaline phosphatase are also frequently seen in patients with a severe attack, but they return to normal when the disease goes into remission. They probably reflect a fatty liver, together with the effects of toxaemia or poor nutrition. Persistent elevation, especially of alkaline phosphatase, may indicate underlying chronic liver disease and needs further investigation (see below).

Serum immunoglobulins rarely exceed the upper limit of normal during a relapse, but usually fall as remission occurs.

Differential diagnosis

If the patient has a history of slow onset of symptoms, including blood and mucus, and has diffuse inflammation on sigmoidoscopy, the diagnosis of ulcerative colitis is highly probable. The major differential diagnosis is Crohn's disease (see Chapter 14.10). If clinical, radiological, endoscopic, and histological information is considered together, less than 10 per cent of patients fall into the category of indeterminate colitis. The recently recognized collagenous colitis usually has only a mild inflammation on colonoscopy and is diagnosed on the basis of a thickened subepithelial collagen band (wider than 15 µm) seen in a rectal biopsy specimen. Microscopic or lymphocytic colitis has a normal endoscopic appearance but shows a diffuse infiltration of the lamina propria with lymphocytes and eosinophils on histological examination. Although ischaemic colitis classically occurs around the splenic flexure, it may occur in the rectum, especially in the elderly, and can be diagnosed histologically. Radiation damage to the rectum may occur, especially in men who have had radiotherapy to the prostate.

Rarely, a drug-induced colitis may occur. The drugs that have been implicated include non-steroidal anti-inflammatory drugs, gold, penicillamine, and 5-aminosalicylic acid. The last drug may cause considerable diagnostic confusion in patients who already have ulcerative colitis. An antibiotic history must be taken but a pseudomembranous colitis secondary to *Cl. difficile* can occur in the absence of antibiotic usage, especially in the elderly.

For those patients presenting with a much more acute history, infective forms of colitis must be excluded by stool culture. A sudden onset of symptoms, the predominance of abdominal pain, the ingestion of potentially infected food (chicken, shellfish), and evidence of diarrhoeal disease in contacts are obvious pointers to an infection. Sigmoidoscopic appearances are usually very similar to ulcerative colitis but a rectal biopsy can be very useful in distinguishing an infective from a more chronic ulcerative colitis. The presence of a chronic inflammatory infiltrate, architectural disturbances of the glands, and basal lymphoid aggregates favour ulcerative colitis. The common organisms causing an infective colitis are salmonella, shigella, and campylobacter. Yersinial infections may also cause a colitis and can pursue a chronic course over many months before resolving. Special culture conditions may isolate the organism from stool, but a rising titre of serum antibody is often the more reliable method of identifying the infection. *E. coli* 0157 is a recognized cause of an acute colitis, especially in institutions, and massive bleeding is often a characteristic feature. Children may develop a haemolytic uraemic syndrome. Diagnosis is difficult because most laboratories are not equipped either to detect this strain of *E. coli* or to measure specific antibody. For patients who have travelled in endemic areas, amoebic and schistosomal colitis must be considered—stool examination and histological demonstration of amoebas or schistosomal ova in rectal biopsy specimens make the diagnosis.

Other causes of infective colitis can occur in immunosuppressed patients and include cytomegalovirus, herpes simplex, and *Mycobacterium avium intracellulare*. Although these organisms are usually associated with fairly characteristic sigmoidoscopic appearances, they can be associated with a more diffuse pattern of inflammation. Other sexually transmitted causes of proctitis (gonorrhoea, chlamydia, lymphogranuloma) do not usually cause diarrhoea and, especially with gonorrhoea, are associated with the passage of watery pus.

Ulcerative colitis also has to be differentiated from irritable bowel syndrome, colonic polyps or carcinoma, diverticular disease, solitary rectal ulcer syndrome, and factitious diarrhoea. Sigmoidoscopy usually clarifies the diagnosis, but if the ulceration of the solitary rectal ulcer syndrome becomes circumferential, this can be mistaken for ulcerative colitis. A biopsy specimen showing strands of smooth muscle radiating up into the lamina propria between the glands is characteristic of the solitary ulcer syndrome.

Extraintestinal manifestations

Table 2 lists the extraintestinal manifestations.

Skin

The most common skin rash seen in patients with ulcerative colitis is a hypersensitivity rash to sulphasalazine (related to the sulphapyridine moiety), which may be photosensitive. Erythema nodosum occurs in about 2 per cent of patients and is mostly associated with active disease. The lesions occur most commonly on the anterior aspect of the lower legs. Pyoderma gangrenosum is rare (1 to 2 per cent) and is usually seen in patients with active disease, but occasionally persists despite inactive colitis. The lesions usually begin as sterile pustules, usually on the limbs, which break down as they enlarge and finally coalesce. Ulceration leads to necrosis and the lesions become surrounded by black, necrotic tissue. Treatment of the colitis is usually followed by regression of the skin lesions.

Mouth

Crops of aphthous ulcers are common in patients with active disease. A sore tongue and angular stomatitis often accompany chronic iron deficiency.

Eyes

Episcleritis or an anterior uveitis occur in 5 to 8 per cent of patients. Local corticosteroids and treatment of active colitis usually lead to resolution.

Joints

An acute arthropathy occurs in 10 to 15 per cent of patients with active disease. It affects the larger joints (knees, hips, ankles, wrists, elbows) and is usually asymmetrical. It is a non-erosive condition and settles as the colitis goes into remission. A less common joint complication is a symmetrical small-joint polyarthropathy which is seronegative and is unrelated to the activity of the colitis.

Low back pain is a common symptom and is usually due to a sacroiliitis, which can be seen radiologically in 12 to 15 per cent of patients. It is unrelated to disease activity, is not strongly associated with HLA B27, and rarely progresses to ankylosing spondylitis. The latter disease occurs in only 1 to 2 per cent of patients and 60 per cent of these have the HLA B27 phenotype. There is a 2:1 ratio in favour of males with this complication. The spondylitis may present before the colitis becomes apparent or may follow the intestinal symptoms. Its natural history is independent to that of the colitis and should be treated with physiotherapy, hydrotherapy, and if necessary, non-steroidal anti-inflammatory drugs. However, these drugs can occasionally worsen the colitis and should therefore be used cautiously.

Liver disease

Patients with severe attacks of ulcerative colitis often have minor elevations of alkaline phosphatase or transaminases. The cause of these enzyme rises is probably multifactorial, including malnutrition, sepsis, and a fatty liver, which occurs in up to 60 per cent of patients coming to urgent colectomy. The liver enzymes return to normal activities when remission is achieved.

However, there may be persistent abnormalities in liver enzymes in about 3 per cent of patients, usually a rise in alkaline phosphatase. The overwhelming majority of these patients will have primary sclerosing cholangitis when the bile duct is visualized by endoscopic cholangiography. Histologically, liver biopsy specimens show evidence of chronic liver disease, but the spectrum of appearances ranges from those of an autoimmune hepatitis to the classic picture of concentric periductular fibrosis with obliteration of bile ducts.

Many patients with ulcerative colitis and sclerosing cholangitis remain well for many years. The colitis is often very mild, though frequently affecting the whole colon, but the liver disease is progressive and ultimately leads to portal hypertension and liver failure. Sclerosing cholangitis is a premalignant condition and explains the well-recognized association between ulcerative colitis and cholangiocarcinoma. Pathogenesis and treatment of the liver disease are discussed in Chapter 14.20.2.3.

Rare associations

Pericarditis with or without an effusion has been described in association with an acute attack of colitis but a true association is not yet proven. Autoimmune haemolytic anaemia has been reported in ulcerative colitis and may recur when the colonic disease becomes active. Amyloid rarely occurs in ulcerative colitis—it is much more likely to be associated with Crohn's disease. A rapidly progressing bronchiectasis has also been described in some patients with ulcerative colitis.

Medical management

The main principles of therapy for the treatment of ulcerative colitis are: to control active disease rapidly, to maintain remission, to select patients for whom surgery is appropriate, and to ensure as good a quality of life as possible.

Treatment of active disease

The most effective drugs for controlling active disease are the corticosteroids, which may be given systemically, topically, or in combination. Drugs containing 5-aminosalicylic acid (sulphasalazine, olsalazine, balsalaizide, mesalazine) are often used to treat a mild colitis but prednisolone has been shown to be more effective and to control symptoms more rapidly, which make it the drug of choice. The dosage and route of administration are largely governed by disease severity. Once active inflammation has been controlled and remission obtained, the corticosteroids should be tailed off because they are ineffective as maintenance therapy and prolonged use puts the patient at risk of long-term side-effects such as osteoporosis.

Proctitis

Proctitis refers to disease limited to the rectum—in practice, it refers to inflammation that does not extend beyond the limits of a rigid sigmoidoscope. It can be remarkably difficult to treat. Initial therapy is usually a 5-aminosalicylic acid drug by mouth in combination with topical therapy. The latter can be a corticosteroid or 5-aminosalicylic acid in the form of a suppository. For patients who do not respond, oral prednisolone may be given. Some have sufficiently severe proctitis to warrant intravenous steroids, and occasionally colectomy may be necessary. Many patients with a refractory proctitis develop a severe proximal constipation, which can cause considerable abdominal discomfort, bloating, and nausea. Relief of the constipation, usually by gentle osmotic purgation, will often give considerable symptomatic benefit but may also be associated with a marked improvement in the inflammation. Some patients appear to be refractory because foam or enema preparations are used: changing to a suppository allows a much higher concentration of drug in the rectum and is frequently associated with improvement in symptoms.

Mildly active disease

Patients who have no more than four motions daily on average, with inflammation extending beyond the limits of the rigid sigmoidoscope, should be given 20 mg of oral prednisolone daily, together with topical steroids or 5-aminosalicylic acid. Treatment should be given for at least 4 weeks before being tailed off over the subsequent 3 to 4 weeks. Clinical trials have shown that 5-aminosalicylic acid formulations are more effective than placebo in treating active disease, especially in high doses. However, corticosteroids achieve remission more quickly and in a higher proportion than 5-aminosalicylic acid drugs.

Moderately active disease

Patients who have, on average, more than four bowel motions daily but who are not systemically ill should be given 40 mg of prednisolone by mouth daily. Giving larger doses (such as 60 mg daily) provides only a marginally better effect but increases the frequency of side-effects quite considerably. The dose is reduced to 20 mg daily over 2 to 3 weeks and the regimen then follows that described for mild disease.

Severe disease

This is defined as an attack in which the patient has more than six bowel motions daily, with blood, and who is systemically ill as shown by tachycardia, fever, and anaemia. The colon is usually tender on palpation. These patients should be admitted to hospital and assessed by both physician and surgeon. Fluid and electrolyte losses are replaced intravenously; a blood transfusion should be given if the haemoglobin is less than 10 g/dl. Patients are given intravenous corticosteroids (such as 100 mg of hydrocortisone, 6-hourly) together with a twice daily rectal drip of hydrocortisone (100 mg in 100 ml water). Parenteral nutrition is indicated for patients who are malnourished, but for the majority, intravenous saline and dextrose–saline are sufficient, together with potassium supplements. Most patients with a severe attack prefer to have only clear fluids by mouth during the first 24 h. Thereafter, there is no evidence that a light diet has any adverse effect on the disease, but many clinicians will leave the patient on only clear fluids for the first few days.

Provided the patient is improving, treatment is continued for 5 to 7 days. At this time, a good response is one in which the patient feels well, there is no fever or tachycardia, the colon is not tender on abdominal palpation, and the diarrhoea has largely settled, usually to less than four motions daily. At this stage, the stools are rarely formed but macroscopic bleeding has stopped. These patients can then go on to oral prednisolone (for instance 40 mg daily), a retention enema, an oral 5-aminosalicylic acid drug, and a light diet. Patients who deteriorate during the first few days of intravenous treatment or those who have not made a substantial improvement by the end of the first week should be advised to have urgent surgery. The more difficult decision is when patients have made some improvement but are still not well—they may still be anorexic, have an intermittent low-grade fever, tachycardia, and continuing diarrhoea. Continuing intravenous therapy for more than 7 to 10 days is rarely beneficial and surgery is usually required. It is in this group of patients that the introduction of a light diet towards the end of the first week of treatment often provides a guide to future management. If the pulse rises or a fever develops in response to feeding, urgent colectomy is required. For the group of patients who do not make a rapid response to intravenous steroids, the addition of cyclosporin will induce remission in 60 to 80 per cent. The drug is usually given by continuous intravenous infusion (4 mg/kg), although the new oral formulation (Ne-oral) may be as effective. Intravenous cyclosporin should not be given in patients with a low cholesterol level (less than 3.00 mmol/l) as it can be complicated by fits. In these patients oral cyclosporin should be used as it is the cremaphor that is used to suspend the intravenous form that causes the fits. If cyclosporin is to be used, it should be added after 3 to 5 days of steroid therapy. Most patients respond very quickly (3 to 4 days) and thus, for those in whom surgery becomes necessary, a decision about colectomy can be made after a week or so of treatment. Those who do respond to drug therapy can be converted to oral treatment with decreasing doses of prednisolone. Practice varies between continuing oral cyclosporin for some months, changing to azathioprine, or using the two in combination. No controlled trial data are yet available to provide evidence-based guidelines.

So far, the use of cyclosporin has not been associated with major side-effects when used in the way described above, although prolonged use of high-dose steroids and cyclosporin can be associated with *Pneumocystis carinii* pneumonia. Minor, reversible side-effects are frequent with long-term use of oral cyclosporin.

Approximately 25 per cent of patients with a severe attack will require an urgent colectomy. These patients can often be identified early on using clinical and radiological features, which have been shown to have prognostic significance. These are the passage of more than nine stools daily, a pulse rate greater than 100/min, or a temperature greater than 38°C during the first 24 h of treatment. A serum albumin level of less than 30 g/l during the first few days or the failure of acute-phase proteins such as the serum C-reactive protein to fall are also poor prognostic signs. Seventy-five per cent of patients showing mucosal islands in the colon or having more than three loops of distended small bowel on a plain abdominal radiograph will come to urgent surgery. These findings, based on retrospective studies, have been confirmed by a prospective series which showed, in addition, that if, on day 3 of steroid therapy, patients were still having more than eight motions daily or four to six daily with a C-reactive protein greater than 45 mg/l there was an 85 per cent chance that urgent colectomy would be required. Thus, for patients falling into this category, cyclosporin should be added at this stage unless there are other reasons for proceeding directly to surgery.

Chronic active disease

Some patients repeatedly relapse when they come off corticosteroids or receive a daily dose of less than 10 to 15 mg of prednisolone. Immunosuppression therapy with azathioprine or 6-mercaptopurine is often beneficial in this group. In the United Kingdom, azathioprine is the drug that is most used, in doses of 2.0 or 2.5 mg/kg, and may allow the prednisolone to be withdrawn. It usually takes 4 to 6 weeks before an effect is seen and the drug is then continued for several months. Although few long-term sequelae have been encountered, most clinicians do not usually continue therapy for more than 18 to 24 months. Oral cyclosporin (5 mg/kg) has also been used for chronic active disease but no formal clinical trials have been made. High-dose prednisolone (40 mg) given on alternate days is another approach that may be useful. However, if the patient's lifestyle is impaired by chronic disabling symptoms or by the side-effects of treatment, surgical management should be considered.

Maintenance of remission

Sulphasalazine and its active moiety, 5-aminosalicylic acid, are able to maintain the disease in remission when given over many years and reduce the relapse rate by about fourfold. Thus, provided they are well tolerated, they should be given indefinitely.

For sulphasalazine, the optimal dose to obtain good therapeutic efficacy with the least side-effects is 2 g daily. Common side-effects are nausea, anorexia, and headache, which are dose related and are caused by the sulphapyridine component. Other side-effects, which are also usually due to the sulphonamide but are not dose related, include hypersensitivity skin rashes, male infertility, agranulocytosis, and Heinz-body haemolytic anaemia. Overall, 10 to 15 per cent of patients are unable to take the drug, although the nausea and headache can often be overcome by starting at a low dose and gradually increasing it.

Sulphasalazine is an unusual drug in that it is poorly absorbed in the stomach and small intestine. When it reaches the colon, the azo-bond linking the 5-aminosalicylic acid and sulphapyridine moieties is split by bacterial azoreductases. The sulphapyridine is absorbed, metabolized in the liver, and excreted in the urine. The majority of the 5-aminosalicylic acid (about 70 per cent) is poorly absorbed and excreted in the faeces. As it is the 5-aminosalicylic acid that is the active compound, several drugs are now available that present 5-aminosalicylic acid to the colon without the sulphapyridine which causes the majority of the side-effects of sulphasalazine. The 5-aminosalicylic acid cannot simply be given by mouth as it is rapidly absorbed. Thus, it is either given as a delayed-release formulation (the mesalazine group) or as a prodrug (olsalazine, balsalazide). Table 3 lists these and details their characteristics.

Which drug containing 5-aminosalicylic acid should be prescribed as maintenance therapy for ulcerative colitis? Sulphasalazine is well tolerated by 85 per cent or so of patients, it is cheap, and serious side-effects (such as Stevens–Johnson syndrome, agranulocytosis, and pancreatitis) are very rare. The newer drugs are much more expensive than sulphasalazine but they have equal therapeutic efficacy. In general they are associated with fewer side-effects. However, occasional patients develop typical salicylate reactions (rhinitis, urticaria, and a colitis). About 10 to 12 per cent of patients will develop loose stools when given olsalazine. This gradually settles if treatment is continued, but about 5 per cent will develop a severe watery diarrhoea, which usually necessitates stopping the drug. The risk of diarrhoea can be minimized by taking the drug with food. There have been reports of renal failure, mainly due to an intestinal nephritis, and have been mostly associated with the delayed-release forms of mesalazine. It is a rare complication and the mechanism is unknown although 5-aminosalicylic acid has structural similarity to phenacetin. Both Asacol and, especially, Salofalk give higher plasma concentrations of 5-aminosalicylic acid than either Pentasa or the prodrugs (olsalazine, balsalazide, and sulphasalazine).

Diet

Patients with recurrent, severe disease have a slightly higher prevalence of hypolactasia and a lactose-free diet may be beneficial. Individual patients may be intolerant of dairy products, wheat, eggs, and other dietary constituents but the majority of patients should have a normal, well-balanced diet.

Local complications

Perianal lesions

Minor lesions such as fissures, perianal abscesses, or haemorrhoids may occur in patients with ulcerative colitis, but extensive lesions such as fistulas are exceptional and, if they occur, suggest Crohn's disease. Treatment of fissures involves treatment of active inflammation. Surgical treatment should be avoided wherever possible and, if necessary, should be conservative.

Massive haemorrhage

This occurs in association with severe attacks but is rarely seen. Intravenous corticosteroids and blood transfusion usually allow the bleeding to stop. However, if patients have already received six or more units of blood and are still bleeding, urgent colectomy must be considered.

Perforation

This is the most dangerous of the local complications and carries appreciable mortality. In patients receiving corticosteroids, the physical signs of peritonitis may not be obvious, and malaise, tachycardia, and reduced or absent bowel sounds may be the only clinical features. Plain abdominal films usually show free intra-abdominal gas. It may complicate an acute dilatation but can occur in its absence. Management consists of immediate intravenous fluid, electrolytes, antibiotics, and hydrocortisone. As soon as the patient's condition improves, urgent colectomy is performed immediately. The mortality of a perforation is as high as 16 per cent, even in specialist centres.

Acute dilatation

This is defined as a transverse colon with a diameter of greater than 5.0 to 6.0 cm with loss of haustration seen on a plain radiograph in a patient with a severe attack of ulcerative colitis. It occurs in about 5 per cent of patients with a severe attack and can be precipitated by hypokalaemia or the administration of opiates. Physical signs are often minimal but the patient is usually obtunded, the bowel sounds are reduced, and the abdomen may become distended. If the colon is already dilated on presentation of the severe attack, medical therapy with intravenous steroids should be given. Approximately 50 per cent of patients will settle on medical therapy alone, but urgent surgery is required for those who continue to deteriorate or do not improve within 24 h. If the colon dilates during the course of treating a severe attack, colectomy should be performed.

Strictures

These occur very rarely in patients with long-standing ulcerative colitis with a shortened, narrow colon. Colonoscopy with multiple biopsies must be carried out as there should be a high index of suspicion for carcinoma.

Pseudopolyps

These are common and may be filiform, sessile, or may form bridges. They can occur throughout the colon but often spare the rectum. They are not premalignant and may occasionally regress.

Colonic carcinoma

The risk of cancer is mainly in patients who have had extensive disease for more than 10 years, especially if they have had recurrent attacks. The most recent series studying primary cohorts suggest that the cumulative risk for patients with extensive disease is about 7 to 15 per cent at 20 years, with very little risk up to 15 years of disease.

Carcinoma is usually, but not always, preceded by dysplasia. This can be detected histologically and has led to the use of colonoscopic surveillance programmes for patients with long-standing ulcerative colitis affecting most or all of the colon. Provided no dysplasia is found, the examination is repeated every 1 to 3 years. If high-grade dysplasia is present, prophylactic colectomy is usually considered. For low-grade dysplasia, repeat colonoscopy within a few months is usually advised, but there are increasing reports of cancers occurring in association with low-grade dysplasia. Thus colectomy is increasingly being recommended whenever dysplasia is recognized regardless of grade. As large numbers of colonoscopies are involved in a surveillance programme the question of cost–benefit has been raised. However, two recent studies have shown that patients have a worse outcome with respect to cancer if they are not in a surveillance programme. The possibility of using flow cytometry to detect DNA aneuploidy in biopsy specimens as a means of increasing the sensitivity of surveillance has been explored, but is probably no better than a histological assessment given by an experienced intestinal pathologist.

Surgery

The indications for surgery have already been mentioned and are:

1. severe inflammation unresponsive to medical therapy;
2. acute complications—perforation, dilatation;
3. for chronic active disease; and
4. to prevent cancer.

The choice of operation is partly determined by the expertise available and the activity of the disease. When surgery is done for a severe attack, a one-stage proctocolectomy with a Brooke ileostomy has been shown to be a safe and effective procedure. The major problems with the operation are poor healing of the perineal wound, adhesion obstruction, and ileostomy dysfunction. Sexual dysfunction in males rarely occurs if a perimuscular excision of the rectum is made. However, with the advent of restorative proctocolectomy with the formation of an ileoanal reservoir or pouch, many surgeons will do only a colectomy in the acute stage. The rectal stump is either oversewn (which is not recommended as it often leaks with abscess formation), or brought out as a mucous fistula either in the lower end of the wound or in the left iliac fossa. This allows histological examination of the whole colon to exclude Crohn's disease. The rectum is excised and the pouch formed some months later when nutrition has been restored and patients are not taking corticosteroids or immunosuppressive drugs.

Restorative proctocolectomy has become the procedure of choice for specialist centres provided the anal sphincter is intact. For this reason, this operation is not advised over the age of 65 years. Either the two-limb (J) or four-limb (W) pouch is now the favoured design and most surgeons preserve the anal transitional zone by anastomosing the pouch 1 to 2 cm above the dentate line. This allows for a stapled anastomosis which shortens the operation and is associated with better continence than a mucosectomy which inevitably requires anal dilatation by retractors. If the operation is being carried out for high-grade dysplasia or cancer, most surgeons will perform a mucosectomy and hand-sew the pouch to the dentate line.

The majority of patients who undergo a pouch operation have excellent function, with less than 10 per cent having any leakage, which is usually limited to night-time soiling. Nevertheless, all patients should be advised to wear a pad at first after a pouch procedure. The pouch usually requires emptying 6 to 12 times daily within the first few weeks of functioning and loperamide is usually needed. Adaptation occurs during the first few months, and by the end of a year the emptying frequency is around four to six times daily but without urgency. Complications of the pouch, once the immediate surgery is over, include anal stenosis, adhesion obstruction, and pouchitis. Pouchitis occurs in 10 to 20 per cent of patients and consists of diarrhoea with blood and evidence of inflammation on endoscopy. It usually responds to antibiotics such as metronidazole or ciprofloxacin but occasionally requires topical treatment with corticosteroids or 5-aminosalicylic acid.

The causes of pouchitis are heterogeneous and include ischaemia, infection with a recognized pathogen (such as campylobacter), and poor emptying, but most pouchitis attacks are unexplained.

Poor emptying can be recognized by isotopic scanning using a radiolabelled artificial stool and usually responds to regular catheterization of the pouch. The idiopathic pouchitis is particularly interesting in so far as it is only seen in patients who have previously had ulcerative colitis and is rarely, if ever, seen in patients who have a pouch for other reasons. After the formation of a pouch, for whatever indication, the ileal mucosa undergoes colonic metaplasia. The triggers for this are unknown but almost certainly involve luminal stasis. Thus, whatever factors first render an individual susceptible to developing ulcerative colitis also seem to render him/her susceptible to developing acute inflammation in ileal mucosa that has undergone colonic metaplasia.

Course and prognosis

Most patients with ulcerative colitis have intermittent attacks of the disease, but the duration of remission between attacks can vary from a few weeks to many years. About 10 to 15 per cent of patients will have a chronic continuous course and rarely achieve a full remission for any appreciable time. A few (5 to 10 per cent) will

have a severe first attack requiring urgent surgery, but fewer, if any, have one attack only and never relapse.

Patients with extensive or total disease are much more likely to have a severe attack within 1 year of diagnosis than patients with distal disease and are therefore at greater risk of colectomy. However, a year from diagnosis the risk of colectomy is similar in all groups with a cumulative rate of about 1 per cent per year. Patients with disease limited to the rectum are a special group in so far as most of them continue to have very limited involvement. Only about 30 per cent will develop more extensive disease in the 20 years after diagnosis.

Despite having a chronic relapsing disease, 90 per cent or so of patients are able to work with very few days of sick leave each year. Nevertheless, quality of life can be impaired in many patients. During active inflammation, lassitude, discomfort, and urgency of defaecation are the major symptoms that limit everyday activities. Sexual and marital problems are not uncommon but may be no more frequent than that seen in other populations of patients with acute-on-chronic illnesses. Most of these problems disappear during remission, although fear of relapse and the need for continuing treatment and medical supervision can cause considerable anxiety. Many patients will alter their lifestyle with respect to daily activity, travel, and diet, but with prompt treatment of active disease and supportive medical care, most are able to have a normal life for most of the time. The development of patient self-help groups (such as The National Association for Colitis and Crohn's Disease in the United Kingdom) has been of tremendous value in providing education and an environment in which patients can regain their confidence and overcome the problem of isolation, an important and common factor in patients with an uncommon and socially unpleasant disease.

There has been a dramatic fall in the mortality rates for ulcerative colitis since the introduction of corticosteroids in the 1950s and the improvement in the management of severe attacks. The mortality rate for a severe attack, including urgent surgery, should now be less than 2 per cent. In the longer term, mortality differs hardly at all from that expected in a matched healthy population, a fact which the majority of life assurance companies fail to recognize.

Ulcerative colitis in pregnancy

Women with ulcerative colitis have normal fertility, are not at increased risk of having a spontaneous abortion, and there is no evidence that pregnancy is a risk factor for relapse. If they do become pregnant, the chance of having a normal baby is the same as for healthy women. Furthermore, there is no good evidence that corticosteroids, drugs containing 5-aminosalicylic acid, or even azathioprine are harmful. Therefore, maintenance treatment should be continued throughout the pregnancy and, if a relapse does occur, it should be treated aggressively with corticosteroids to obtain a rapid remission.

Ulcerative colitis in childhood

Ulcerative colitis is less common in children than in adults and, for the United Kingdom, the prevalence is about 6 to 7 per 100 000. Nevertheless it can present within the first few weeks of life, although the mean age of presentation is about 10 years. The symptoms are those of diarrhoea, rectal bleeding, abdominal pain, and failure to thrive. There may be evidence of delayed growth but this is more commonly a feature of childhood Crohn's disease. The proportion of children with a total colitis is about 50 per cent, which is higher than in adults, and probably accounts for the higher rate of colectomy reported in most series.

Treatment follows the same principles as for adults, although dosages are adjusted for the child's weight. In addition, great attention must be made to nutrition to allow for adequate growth. For children requiring repeated courses of corticosteroids, an alternate-day regimen usually controls the disease activity but prevents growth retardation. If colectomy becomes necessary, a restorative proctocolectomy should be done.

Further reading

Allan RN *et al.* (1997). *Inflammatory bowel diseases*, 3rd edn. Churchill Livingstone, Edinburgh.

Jewell DP, Warren BF, Mortensen NJ (2001). *Inflammatory bowel disease.* Blackwell Science, Oxford.

Kirsner JB (2000). *Inflammatory bowel disease*, 5th edn. WB Saunders Co., Philadelphia.

14.12 Functional bowel disorders and irritable bowel syndrome

D. G. Thompson

Introduction

Functional bowel disorders

Symptoms suggestive of disturbed lower gastrointestinal function without adequate explanation are very common in the adult population of the Western world. Surveys from the United Kingdom and United States indicate that up to 15 per cent of the adult population experience such symptoms at any one time, although most do not seek medical advice. The chief questions that remain largely unresolved are whether the symptoms of those individuals who do seek medical help have a different pathophysiological basis from those who do not, and whether the seeking of medical advice is an indication of a worried individual rather than of disturbed gut function.

Given these difficulties, it has to be accepted that most of the currently used terms are best viewed as an attempt by clinicians to provide some clinically useful categorization of such patients and their symptoms. Since knowledge of the physiology and the psychology of the problem remains incomplete, therapy remains largely empirical. Current observations about functional bowel disorders should therefore be regarded as the latest (but by no means the last) attempt at rationalizing a complex interrelationship between brain and gut function.

Definition of terms used

Over the last century many attempts were made to categorize functional bowel disorders. It is not surprising that most have failed to stand the test of time, as the symptoms suffered (whilst being genuine and troublesome to the patient) are often difficult to define, variable in their expression, and defy pathophysiological explanation. The latest and perhaps most comprehensive attempt has been made by a working group whose recommendations, known as the 'Rome criteria', are now accepted as useful research tools. Whether these criteria will stand the test of time as clinical diagnostic tools will, however, depend upon whether they turn out to provide a better understanding of the pathogenic mechanisms of the disease or to aid its therapy. The Rome Working Group has suggested the division of functional bowel disease into a number of symptom-based categories (Table 1). Because they do seem to have some practical value in guiding approaches to management, this chapter is based on some of the Rome categories, with emphasis on those previously encompassed by the term 'irritable bowel syndrome'. However, it must be recognized that the Rome criteria do not necessarily include all symptoms presented by patients with abnormal bowel function. Failure to allocate a patient into one or other category should therefore not be taken to mean that the patient does not have a functional bowel disorder.

Irritable bowel syndrome

Definition

This syndrome is characterized by the presence of abdominal pain associated with defaecation, or a change in bowel habit, together with disordered defecation and the sensation of abdominal distension. For practical purposes, its recognition relies upon the presence of abdominal pain that is relieved by defaecation and of an associated change in frequency in defaecation and/or stool consistency, together with two or more of the following symptoms:

1. altered stool frequency;
2. altered stool consistency;
3. altered ease of defaecation;
4. passage of mucus;
5. sensation of abdominal distension.

These criteria are themselves based on the studies of Manning and of Kruis, which identified that the above features were reported most frequently in patients with functional bowel problems but were very unusual in patients with structural disease of the colon.

Diagnosis

The diagnosis of irritable bowel syndrome is clinically based, and relies on a carefully taken history and examination, there being no specific endoscopic, radiological, or laboratory investigation that is yet capable of providing a positive diagnosis. Despite the absence of a specific pathological indicator, the identification of irritable bowel syndrome is usually not difficult, and in most cases it is unnecessary to investigate the patient extensively in an attempt to exclude other, more serious disease.

Clinical features

The history

In addition to the careful elicitation of the above specific symptoms, other features may be found that serve to increase clinical confidence. For example, many patients have upper-gut symptoms, for example food-related abdominal distension. Women may also complain of menstrual and bladder symptoms, and there is also an increased prevalence of psychosexual problems.

Examination

Clinical examination is important. Whilst there is no physical abnormality that is diagnostic of irritable bowel syndrome, a number of features occur commonly. Palpation over the site of the lower colon, particularly in the left iliac fossa, may produce discomfort, and a sigmoid colon containing faeces is often palpable. Similar tenderness may be present under the rib margins and in the right iliac fossa.

Rectal examination and sigmoidoscopy should be conducted as part of the initial clinical assessment. Characteristic findings are the presence of pellety stools in the rectum and a mucosa of normal appearance, evidence of mucosal inflammation is incompatible with the diagnosis. A further helpful pointer is the response to air insufflation during the sigmoidoscopy; abdominal discomfort is often reproduced and relieved by air expulsion. Evidence of a pigmented rectal mucosa (melanosis coli) may be found in patients who have been taking stimulant laxatives and is a useful indicator of the chronicity of the problem.

Further laboratory investigations remain at the discretion of the clinician, depending upon the confidence with which a clinical diagnosis is made. Routine haematological and biochemical screening is usually done on the assumption that they will be normal, and thus to provide reassurance both to the patient and the doctor. Radiological and endoscopic examination of the colon is not mandatory unless a clinical suspicion of a structural colonic disorder, particularly neoplasia, remains after the history and examination have been completed.

Features that raise the suspicion of organic disease and indicate a need for further investigation include the onset of symptoms in the middle-aged or elderly, weight loss, or blood in the stool. The development of new colonic symptoms in a patient with a long history of irritable bowel syndrome should also be taken seriously, as there is no evidence that the syndrome protects against the development of other disease and the incidence of colonic neoplasia increases with age.

Pathophysiology

Despite much interest and many painstaking clinical studies, our understanding of the pathophysiology of irritable bowel syndrome remains limited and the following hypotheses are discussed solely as a guide to current thinking.

Neuromuscular dysfunction

The most popular hypothesis is that these patients have a disorder of neuromuscular function of the gastrointestinal tract. However, whilst this seems eminently plausible, evidence is lacking. Manometric studies of the colon do show an increased contractile activity in patients with this syndrome, particularly after food, but the neurophysiological basis of this finding and its relationship to symptoms remains to be determined.

Visceral hypersensitivity

Another currently popular hypothesis is that visceral sensation from the gastrointestinal tract is somehow enhanced in these patients. This idea is based on the observation that distension of the rectum and colon produces greater discomfort than in people with normal bowel function. This increased sensory awareness appears to be viscerally specific, as cutaneous responsiveness is normal. However, it remains to be determined whether the mechanism for this hypersensitivity is peripheral (abnormal mechanoreceptor responsiveness in the gut) or central (abnormal sensory processing by the brain and spinal cord).

Psychiatric disease

There is convincing evidence that psychiatric disease and abnormal illness behaviour are more prevalent in patients with irritable bowel syndrome. The relationship between the psychological problem and any neuromuscular abnormality remains uncertain, although it is recognized that a heightened awareness of visceral sensation is a feature of affective disorders, particularly depression.

Diet

It is customary to regard diet as being a pathogenic factor and to attribute constipation symptoms to fibre deficiency, on the basis that irritable bowel syndrome is uncommon in those parts of the world where a high-fibre diet is consumed. While it is true that faecal bulk can be increased by ingesting more fibre and that constipation is improved, careful studies of fibre intake and symptom development do not show a clear causal relationship. Food 'allergy' or sensitivity is occasionally confused with irritable bowel syndrome because abdominal pain and diarrhoea can accompany both problems. Classic food allergy with measurable immunological alterations in response to a particular food (for example, eggs, shellfish) is readily distinguishable from irritable bowel syndrome by a clear relationship between ingestion of the implicated food and symptom development. More subtle forms of food intolerance (for example, lactose intolerance, fructose intolerance) that produce gut symptoms without an accompanying immune response are much more difficult to recognize because the nutrient in question is often present throughout the diet. Recognition requires a painstaking dietary history and the clear demonstration of a relationship between symptoms and food intake. In most patients such a relationship is not found.

Functional abdominal bloating

Definition

This is characterized by symptoms of abdominal fullness or distension, awareness of audible bowel sounds, and excessive flatus with no evidence of either maldigestion and malabsorption or excessive consumption of poorly absorbed fermentable carbohydrate.

Pathophysiology

Distension of the colon at sigmoidoscopy characteristically produces greater discomfort than normal, suggesting increased gut sensitivity. However, there is no evidence that intestinal gas production is increased. As in irritable bowel syndrome, the prevalence of psychological disorders is high.

Clinical features

The clinical assessment of such patients is identical to that for irritable bowel syndrome and features will be identical.

Functional constipation

Definition

This is arbitrarily defined as either persistently difficult, infrequent defaecation or the sensation of incomplete defaecation. Usually, two or more of the following are present: straining at defaecation; lumpy or hard stools; the sensation of incomplete evacuation; and two or fewer bowel movements per week.

Clinical evaluation

As with the other categories of functional bowel disorders, the diagnosis is based on a carefully conducted history and examination designed to exclude the possibility of more serious colonic disease, particularly cancer. When considering the diagnosis, it is important to enquire about immobility, concomitant drug therapy (particularly opiate analgesia), and a low roughage diet, which are well recognized as contributing to constipation, particularly in the infirm.

An abnormality of pelvic-floor function on attempted defaecation is an unusual cause of constipation that should be suspected in individuals who feel the need to defecate but cannot expel faeces despite severe straining. Such a problem should be considered when symptoms develop following pelvic trauma or difficult childbirth. Clinical evidence of diabetes, hypothyroidism, and hypercalcaemia must also be sought, as these may also lead to altered colonic function and

constipation.

Physical examination should include a rectal and vaginal examination. The absence of perineal descent on straining is a simple indicator of impaired pelvic-floor relaxation, while descent below the level of the ischial tuberosities indicates pelvic-floor weakness. Sigmoidoscopy is required to identify the presence of formed faeces, and to exclude faecal impaction and organic obstruction of the lower colon and rectum.

Laboratory examination

Extensive laboratory investigation is usually unnecessary in the absence of clinical indicators of systemic disease and in the presence of the above criteria. A plain abdominal radiograph is often helpful to confirm the presence of faecal material throughout the colon and to allow estimation of the diameter of the small intestine and colon, which helps to exclude the rare cases of intestinal pseudo-obstruction and megacolon caused by intestinal myopathies and neuropathies.

Transit studies using radio-opaque markers are commonly performed as part of the investigation of constipated patients to determine the severity of transit delay, and to distinguish those with a pancolonic abnormality from a more localized problem of pelvic relaxation. However, measurement of whole-gut transit should not be regarded as necessary for the diagnosis—documented infrequent defaecation is usually sufficient. The electrophysiological and radiological assessment of anorectal function is only indicated if there is evidence of abnormal perineal descent or rectal prolapse, as the accurate recognition of pelvic-floor dysfunction can influence the choice of therapy. Such investigations are indicated when Hirschsprung's disease is suspected.

Pathophysiology

The cause of functional constipation is uncertain. Factors likely to be of relevance are similar to those proposed for irritable bowel syndrome, in particular, enteric neural dysfunction.

In the mildest cases, dietary-fibre deficiency may be relevant; however, in the more severely affected patients, fibre supplementation does not abolish the problem and may even worsen symptoms, making, in such individuals, a causal role for fibre untenable. By contrast, a histological abnormality of the enteric nerves of the colon or muscle of the colon may be found in the most severe cases; for the great majority of constipated patients, however, no structural abnormality has been identified.

In a proportion of patients, almost invariably female, defecatory dysfunction appears to be the major factor. A failure of the pelvic-floor muscles to relax on attempted stool expulsion is identifiable in these patients; this appears to be a 'learned' phenomenon with a psychophysiological aetiology rather than peripheral nerve dysfunction. In other patients, low tone in the pelvic floor and rectal prolapse appear to be the result of damage to the pudendal nerve from straining at stool or parturition and thus may be a consequence of the constipation rather than its cause.

In some severely affected women, there is a relationship between symptom severity and the luteal phase of the menstrual cycle, which has led to the suggestion of a sex-hormonal aetiology. In support of this hypothesis is the fact that colonic muscle tone is reduced by progesterone and that constipation is a frequent accompaniment of normal pregnancies. Against the hypothesis, however, is the failure to demonstrate abnormal colonic sensitivity to progesterone in constipated women, leaving the possibility that the menstrual cycle-related events are merely the expression of a normal cyclical progesterone effect on a malfunctioning colon.

Functional diarrhoea

Definition

This is defined as the frequent passage of unformed stool without the presence of other features of irritable bowel syndrome. Neither abdominal pain nor the frequent passage of formed stools are included in the symptoms.

The diagnosis of functional diarrhoea depends on the presence of two or more of the following: unformed stool; three or more bowel movements per day; and increased stool weight, greater than 200 g/day.

Clinical features

This disorder is recognized only after excluding other, more medically serious conditions, in particular, inflammatory bowel disease and secretory diarrhoeas. The possibility of surreptitious laxative use should always be borne in mind. Some patients identify the time of onset of the problem to a specific life-event, particularly a bout of severe gastroenteritis. The possibility of a chronic intestinal infection needs to be considered carefully in such patients, although evidence of an infective agent will be lacking in most and the label 'postinfective diarrhoea' is usually applied.

Physical examination should determine the extent of nutritional deficiency, exclude metabolic disorders such as hyperthyroidism, and rule out intra-abdominal structural abnormalities. Careful examination of stool samples for pathogens and laxatives is required.

Laboratory investigations

Unlike the other functional diseases, it is important to make a careful search for a structural mucosal disease in such patients. Key diagnoses that must be excluded are: chronic malabsorption due to pancreatic insufficiency or gluten sensitivity, inflammatory bowel disease, infections, and infestations of the gastrointestinal tract.

Pathophysiology

In the absence of any definable structural abnormality, functional diarrhoea is generally assumed to be a disorder of neuroenteric control of intestinal epithelial transport.

In some patients, there is a clear relationship between psychological state and symptoms, with diarrhoea worsening whenever anxiety occurs. However, whether the relationship is truly causal is unknown.

Functional abdominal pain

Definition

While this symptom category is commonly included amongst the functional bowel disorders, the relationship between the abdominal pain and a disturbance of gastrointestinal-tract function is difficult to ascertain. Abdominal pain is frequent, recurrent, or continuous, and characteristically persists for many months. The relationship between pain and recognizable physiological events such as eating, defaecation, or menstruation is lacking, and evidence of organic disease in the abdomen is absent. Most of these patients show a major loss of daily functioning capacity and exhibit chronic illness behaviour.

Management of functional bowel diseases

The management of patients with functional bowel disorders remains empirical. Perhaps it should not be surprising that in an area of human suffering with such symptom diversity and in which the pathophysiological mechanisms remain obscure, no single pharmacological agent or group of agents have ever been found to be consistently effective.

A review of randomized, double-blind, placebo-controlled trials for the treatment of irritable bowel syndrome examined 43 trials and concluded that none offered convincing evidence that any therapy was effective, a conclusion which is perhaps as much an indictment of trial design as the efficacy of the drug therapy. Furthermore, in a condition in which the patient's mental state plays such an important part in defining symptom severity, it is not surprising that in most clinical trials placebo responses have been very high, usually up to 50 per cent. Also, short-term trials of therapeutic agents in diseases where symptoms are intermittent may be

unable to distinguish a true drug effect from a placebo response.

So what can the clinician do to help patients with functional bowel disease? As in all chronic problems without a cure, a principal task is to give an explanation and reassurance. Therapy must be patient-centred and designed to provide a solution for the patient's personal needs and expectations. The clinician should give a full explanation of the likely nature of the problem and firm reassurance that organic disease is not likely to be present. Attention to the patient's psychological state is very important, as it is clear that mood is a powerful modulator of symptoms.

In more severe cases of irritable bowel syndrome, psychological treatment using a variety of techniques has been found to provide greater improvement in a patient's sense of well being than drug therapy alone. Good prognostic factors for improvement seem to be overt psychiatric symptoms, particularly anxiety or depression, together with intermittent pain exacerbated by stress. In contrast, patients in whom the abdominal pain is constant, and who exhibit evidence of chronic illness behaviour, do not seem to be helped by a psychotherapeutic approach but they may respond to antidepressants.

In mild cases, attention to the individual and his/her symptoms is usually the approach taken. In patients with predominant constipation, supplementary dietary fibre and poorly absorbed fermentable carbohydrates increase faecal bulk, soften the stool, and may ease defaecation. On occasion, however, this approach can exacerbate symptoms of abdominal distension, probably as a result of increased colonic gas produced by the fermentation of the unabsorbed carbohydrate. Wherever possible, long-term use of stimulant laxatives is best avoided because of the concern that such drugs may themselves damage the colonic enteric-neural function and eventually make the problem worse. Osmotic laxatives and enemas are the mainstay of therapy of the severely constipated patient with slow transit.

For the patient who appears unable to relax the pelvic floor musculature on attempted defaecation, a variety of biofeedback techniques are now available that help the individual to 'relearn' the process. Success is high in those able to engage closely with the therapist in the process.

For patients with diarrhoea-predominant symptoms, attention to diet is also often helpful, as the size of and timing of meals is likely to influence the frequency and social inconvenience of the diarrhoea. Fermentable carbohydrates are best taken in moderation because they can exacerbate symptoms. In the more persistent cases of diarrhoea, symptoms can be improved by simple antidiarrhoeal agents, the dose being adjusted according to the symptoms and administered before a meal.

In the management of patients with unexplained abdominal pain it is tempting to prescribe opiate-derivative analgesics. These are unlikely to be of benefit in the long term, however, and may even exacerbate symptoms because of their constipating effect. Antidepressants are often prescribed empirically in low doses, with evidence of benefit at least in mood elevation. Antispasmodics (for example, hyoscine butyl bromide) are frequently employed. Whilst there are undoubtedly a number of patients and doctors who are convinced of their value, a beneficial effect has yet to be proven beyond doubt by clinical trial.

Relaxation therapy, in particular hypnosis, seems to benefit those individuals who are prepared to participate. In the right conditions, programmes of self-delivered 'autohypnosis' may offer a satisfactory approach for some sufferers.

Surgical intervention for symptoms of functional bowel disorders is usually best avoided, as benefit is unlikely. On occasions, however, subtotal colectomy and ileorectal anastomosis will provide symptomatic benefit in carefully selected patients with severe constipation.

The management of patients with functional bowel disease therefore remains a major challenge that cannot be shirked by clinicians, with whom responsibility exists to provide a careful, individually oriented explanation, and support. The careful conduct of clinical trials of current and new therapies remains a priority for these disorders.

Further reading

Afzalpurkar RG, *et al.* (1992). The self-limited nature of chronic idiopathic diarrhoea. *New England Journal of Medicine*, **327**, 1849–52.

Anuras S, ed. (1992). *Motility disorders of the gastrointestinal tract*. Raven Press, New York.

Christensen J (1992). Pathophysiology of the irritable bowel syndrome. *Lancet*, **ii**, 1444–7.

Creed FH, Craig T, Farmer RG (1988). Functional abdominal pain, psychiatric illness and life events. *Gut*, **29**, 235–42.

Guthrie E, *et al.* (1991). A controlled trial of psychological treatment for the irritable bowel syndrome. *Gastroenterology*, **100**, 450–7.

Klein KB (1988). Controlled treatment trials in the irritable bowel syndrome: a critique. *Gastroenterology*, **95**, 232–41.

Kruis W, *et al.* (1984). A diagnostic score for the irritable bowel syndrome: its value in the exclusion of organic disease. *Gastroenterology*, **87**, 1–7.

Manning AP, *et al*. (1978). Towards a positive diagnosis of the irritable bowel. *British Medical Journal*, **2**, 653–4.

Read NW, Timms JM, Barfield LJ (1986). Impairment of defaecation in young women with severe constipation. *Gastroenterology*, **90**, 53–61.

Wexner SD, *et al*. (1992). Prospective assessment of biofeedback for the treatment of paradoxical puborectalis contraction. *Diseases of the Colon and Rectum*, **35**, 145–50.

Whorwell PJ, Prior A, Faragher EB (1984). Controlled trial of hypnotherapy in the treatment of severe refractory irritable bowel syndrome. *Lancet*, **ii**, 1232–4.

14.13 Colonic diverticular disease

N. J. McC. Mortensen and M. G. W. Kettlewell

Diverticula can be found throughout the gastrointestinal tract, but are seen most commonly in the sigmoid and descending colon.

Epidemiology

Asymptomatic diverticular disease is much more common than clinical diverticulitis. Autopsy studies in the United Kingdom and Australia have shown that the prevalence of colonic diverticula increases with age. It is rare in those under 30 years of age but occurs in more than 50 per cent of those over 70 years. On the other hand, colonic diverticulosis is very rare in African and Asian countries and right-sided disease predominates in Japan. This geographical distribution is not due to race, as West Indians and Asians living in Britain, American Blacks, and Japanese who have moved to Hawaii or the mainland United States are just as prone to the disease as Caucasians. Patients presenting with complicated diverticular disease have a low intake of dietary fibre, whilst vegetarians have a low incidence of the disease.

In Edinburgh, 23 per cent of all barium enemas demonstrated diverticula. The annual incidence increased from 0.17/1000 in those under 45 years to 5.7/1000 in those over 75 years of age. Women were affected more than men. In spite of the introduction of high-fibre diets, there is no evidence that the incidence of acute diverticulitis is declining.

Aetiology

Diverticular disease is said to be a disease of the twentieth century. It was rarely described in the nineteenth century literature, and in Britain there is a correlation between the rising incidence at the beginning of the twentieth century and an increased consumption of refined flour and sugar. Sugar consumption has trebled since 1860, and in the late 1870s the stone grinding of flour was replaced by roller milling, which removes more fibre. Modern white and some brown breads contain little fibre compared with the amount in wholemeal bread, which was previously a staple part of the diet.

The development of diverticula therefore can be ascribed to a lifelong diet deficient in dietary fibre. An unrefined, high-fibre diet produces swiftly passed, soft stools that subject the colon to little strain. Modern, fibre-deficient diets on the other hand give rise to stiff, viscous stools that need high intracolonic pressures to propel them. High luminal pressures cause a protrusion of the mucosa through vulnerable points in the sigmoid and descending colon. They usually occur at the site where colonic blood vessels penetrate the wall. This hypothesis is supported by the observation that, although basal intracolonic pressures are similar in health and diverticular disease, when the diseased colon is activated by emotion, eating, mechanical stimuli, or drugs such as morphine or prostigmine, high pressures are generated in those segments that have diverticula. This is due to hypersegmentation by the colonic smooth muscle, and the difference has been recorded in the earliest stage of disease and may explain its progressive nature. In symptomatic patients an increase in dietary fibre causes a relief of symptoms in many cases.

Changes in the colon wall also play a part. With age, and following episodes of diverticulitis, the colonic wall becomes stiff and less distensible, aggravating the effects of raised intracolonic pressure. An increase in elastin and changes in collagen have been reported. Diabetic patients are prone to diverticular disease at an earlier stage, suggesting a defect in glycolysation of colonic collagen with advancing age. In those with connective tissue disorders such as Ehlers–Danlos syndrome or Marfan's disease, diverticula are also seen at an unusually early age.

The distinction between symptomatic and asymptomatic diverticular disease is important, for whilst something is known about the formation of diverticula it is not known why some diverticula become symptomatic.

Pathology

A diverticulum consists of a herniation of mucosa through the colonic musculature, and as it enlarges its muscle covering atrophies, so that the fully developed diverticulum consists of mucosa, connective tissue, and peritoneum. The striking abnormality is in the thickening of the circular and longitudinal muscle, which both narrows the colonic lumen and shortens the sigmoid like a concertina to give a saw-tooth appearance on barium enema. The diverticula occur as slit-like apertures between the muscle clefts.

Inflammation in diverticular disease is the result of infection around diverticula, which spreads within the pericolic fat to form a dissecting abscess. Usually a single diverticulum is the cause of a pericolic abscess, perhaps initiated by the presence of a faecolith. Involvement of the peritoneum results in local peritonitis, which may become generalized in the event of a perforation. This may also give rise to intra-abdominal abscesses or fistulae to the bladder, small bowel, vagina, or uterus. Repeated episodes of diverticulitis lead to a contracted, narrowed sigmoid colon surrounded by fibrous tissue. Bleeding in diverticular disease can often be traced to an infected diverticulum. This may cause either the erosion of a vessel in its wall or the formation of granulation tissue inside the diverticulum, which then bleeds.

Clinical features

As diverticulosis is so common, most diverticula are asymptomatic. They are usually discovered incidentally and only some 10 per cent produce symptoms, and around 1 per cent require surgery. The symptoms usually result from disordered motility rather than secondary complications of the disease.

Uncomplicated diverticular disease

Pain can be felt along the course of the colon, particularly over the sigmoid, and is often accompanied by a change in bowel habit with the passage of broken, pellety stools after considerable straining. These symptoms may be indistinguishable from those of the irritable bowel syndrome. The passage of blood with an unformed stool is unusual and should alert one to the possibility of other pathology.

Management

All patients should have a rigid or flexible sigmoidoscopy in addition to a barium enema to exclude a rectal or sigmoid carcinoma (Fig. 1). They should be reassured that there is no serious underlying disease and a high-fibre diet should be recommended. This must include wholemeal bread, wholewheat breakfast cereals, rough porridge or muesli, and fresh fruit and vegetables daily. Fibre increases stool bulk in three ways—by holding water, by proliferation of bacteria, and from the byproducts of bacterial fermentation. The coarser the fibre the greater is the faecal bulk, and unpalatability, and although cooking bran improves its taste, it reduces its water-holding capacity. A good clinical response is usually achieved by including two tablespoons of bran with the morning cereal, but about half the patients will experience gaseous distension or cramps on starting the high-fibre diet. It is worth warning them that this is likely to happen and that it will resolve within a month or so if they persist with the diet.

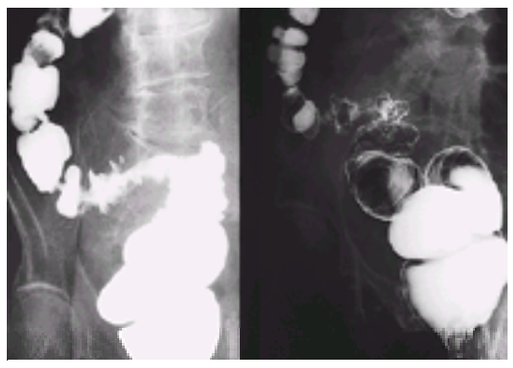

Fig. 1 Barium enema showing a narrowed sigmoid colon with a few diverticula. This appearance can be confused with those of a carcinoma and colonoscopy would be indicated to clarify the diagnosis.

In patients with pain, antispasmodics such as mebeverine may be useful, and in a minority with repeated severe attacks an elective resection is then indicated (Table 1). This is probably more effective than sigmoid myotomy, an operation popularized in the mid-1960s. In this procedure the circular muscle is divided with a longitudinal incision to widen the colonic lumen. The incision is made through the taenia so as to avoid opening diverticula, and is deepened until the mucosa is just seen. The operation lowers the sigmoid intraluminal pressures and improves symptoms but, after 3 years, pressures return to their former levels. The need for myotomy has declined but it may still be useful in some elderly or obese patients.

Complicated diverticular disease

It is important to distinguish the minority of patients who suffer from a febrile attack with left iliac-fossa peritonism, sometimes called left-sided appendicitis, from those with chronic pain and diarrhoea. The inflammation may settle with minimal symptoms or develop into a pericolic abscess or peritonitis.

Acute diverticulitis

Pain is felt over the left lower abdomen, and the patient may have pyrexia, malaise, anorexia, and nausea. The white blood count is raised.

Treatment is with rest, antibiotics, usually cefuroxime 750 mg and metronidazole 500 mg 8-hourly, and analgesia. Most cases settle and the diagnosis can be confirmed after 2 to 3 weeks by barium enema. A narrow segment can sometimes be difficult to distinguish from a carcinoma and any doubtful cases can be clarified by subsequent colonoscopy (Fig. 2).

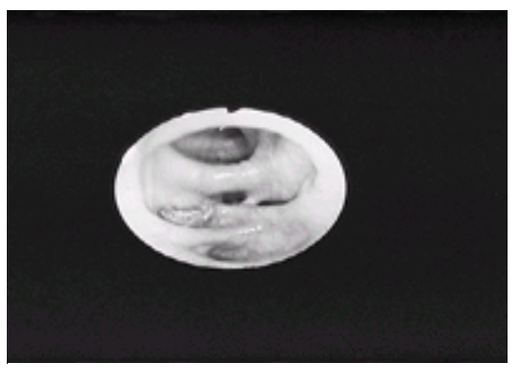

Fig. 2 The typical appearance of diverticula seen at colonoscopy. Note the muscular haustra and the mouths of diverticula—one with a faecolith. (Reproduced from the *Slide atlas of gastroenterology*, Gower Medical Publishing, London, with permission.)

If symptoms fail to resolve, or recur, resection of the sigmoid colon may be necessary. When it is necessary to resect an acutely inflamed and unprepared colon, a Hartmann's operation may be safer than a primary anastomosis.

For recurrent diverticulitis operated electively, a primary anastomosis would be ideal.

Diverticular abscess

Acute diverticulitis can lead to a local peritonitis with abscess formation, either in the paracolic or pelvic area. There may be a palpable mass and a swinging fever. When in doubt the diagnosis can be confirmed by ultrasound or computed tomography (**CT**) with rectal contrast (Fig. 3).

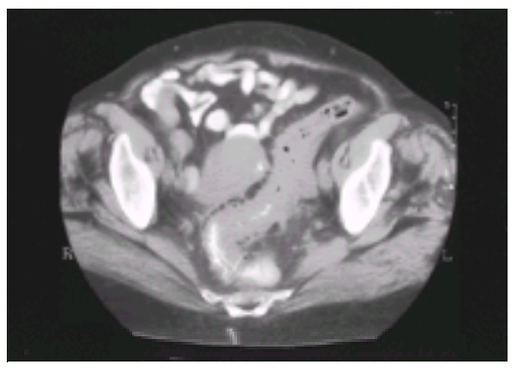

Fig. 3 Computed tomography of the pelvis in a patient with acute diverticulitis. The sigmoid colon is grossly thickened, the lumen narrowed, and pockets of air are seen in the diverticular disease.

It is wise to let an abscess localize whilst treating the patient with rest, antibiotics, and analgesia. Some abscesses will be amenable to drainage by direct incision, over them or via the rectum or vagina. More complicated collections are best drained by CT-guided aspiration or drain placement. There is rarely any need to do a proximal transverse colostomy. If drainage persists, an elective sigmoid colectomy with primary anastomosis can be done at a later time. Even when an abscess is localized, however, the condition remains potentially dangerous as it may rupture into the peritoneal cavity giving rise to peritonitis.

Perforated diverticulitis

Acute diverticulitis can be complicated by generalized purulent peritonitis, either by direct spread from the inflamed colon or by rupture of a peridiverticular abscess.

The clinical picture is of severe intraperitoneal sepsis with toxaemia, ileus, and abdominal pain, and septicaemia will often follow. Emergency laparotomy is almost always required, although time must be allowed for adequate rehydration, correction of electrolytes, and starting antibiotic therapy—again cefuroxime and metronidazole.

Other causes of the acute abdomen that may not require surgery should be excluded, including pelvic inflammatory disease, ureteric calculus, and even pulmonary embolus. In these circumstances a CT scan is invaluable.

There has been a shift away from the more conservative procedures in this situation. At one time, peritoneal toilet, pelvic drainage, and a defunctioning transverse colostomy was the favoured procedure, but this has the disadvantage that the 'septic colon' is left in place and that there is a column of faecal material below the stoma and above the perforation. There is the further problem of the unsuspected carcinoma within the inflammatory mass.

For these reasons more radical measures are favoured by experienced surgeons. A Hartmann's procedure—removing the diseased sigmoid, oversewing the distal rectum, and bringing out an end colostomy—is the most frequently used procedure (Fig. 4). In favourable cases it may be possible to do an on-table colonic lavage via the appendix stump and make an immediate anastomosis but this carries the risk of leakage.

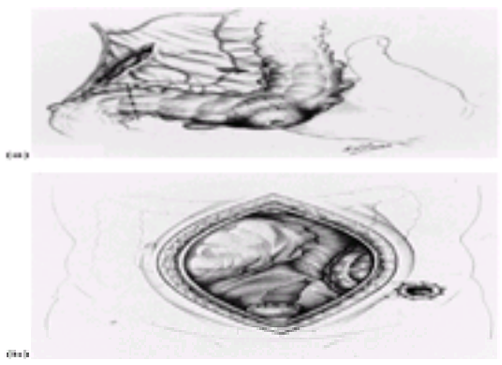

Fig. 4 (a) The area of sigmoid colon resected for perforated diverticular disease. (b) Hartmann's operation—the sigmoid colon has been resected, the rectum oversewn, and a left iliac fossa colostomy fashioned.

Hartmann's procedure is safe and effective, although subsequent reconnection may involve a major operation in elderly patients. Purulent peritonitis carries a mortality of around 15 per cent.

Faecal peritonitis

This is a catastrophic complication with a mortality of around 50 per cent particularly in the elderly. A diverticulum ruptures, often with little or no inflammation, liberating quantities of faeces into the peritoneal cavity. Rapid and severe shock with septicaemia ensues. Energetic resuscitation is necessary, followed promptly by surgery and a Hartmann's operation. These patients often need to be stabilized in an intensive care unit postoperatively.

Intestinal obstruction

Recurrent inflammation with fibrosis and muscular hypertrophy can lead to progressive stenosis and colonic obstruction, which is usually chronic but may present acutely. Conservative treatment is worth trying at first, provided a carcinoma has been excluded. With the aid of a stool softener the symptoms may resolve and the stricture gradually dilate. If these measures fail, the bowel should be prepared for a resection, with care taken not to aggravate the obstruction.

Small-bowel obstruction is sometimes a complication of acute diverticulitis, as the bowel may adhere to the inflammatory mass. It usually resolves as the inflammation subsides but on occasion a laparotomy and division of adhesions or even a small-bowel resection may be necessary.

Colonic fistulae

A colovesical fistula usually presents with recurrent urinary tract infections together with pneumaturia or faecuria. The fistula arises in the sigmoid, which has often folded over into the pouch of Douglas, and adheres to the apex of the bladder. This is the most frequent cause of colovesical fistula but carcinoma and Crohn's disease should be excluded.

Fistulae may also occur between the sigmoid and vagina, uterus, ureter, and ileum. They seldom heal spontaneously but do not always give rise to disabling symptoms and so represent a relative indication for surgery. Sigmoid colectomy as a one-stage procedure is the best option, and colostomy is rarely required. A fistula into the bladder is simply closed and urethral catheter drainage continued for a week.

Haemorrhage

Major haemorrhage is an uncommon but well-recognized complication. It is usually self-limiting, only requiring transfusion and supportive measures. The precise reason for the bleeding is not known but angiographic and colonoscopic studies suggest that many bleeds attributed to diverticula are caused by other lesions such as polyps and angiodysplasia.

Repeated or minor haemorrhage is seldom caused by diverticula and is more likely to be due to carcinoma or polyps. It is therefore vital to exclude other sources of bleeding by barium enema or colonoscopy. The source of a persistent major bleed must be sought urgently, and selective angiography whilst the patient is bleeding is essential. As the haemorrhage can be from any part of the colon, good localization is an essential prelude to any operation. Blind colonic resections have a particularly poor record and if the site of bleeding has still not been located, on-table colonic lavage via the appendix stump and intraoperative colonoscopy will usually target the bleeding segment.

A recent study reported the results of urgent colonoscopy in bleeding diverticular disease. Instead of the tradititional conservative measures patients were given a bowel prep and colonoscoped within 12 h. Bleeding sites thus identified were treated by colonoscopic diathermy and the number of major bleeds, blood transfusions, and operations was reduced together with length of hospital stay. It remains to be seen whether this will result in a major shift of emphasis in management.

Further reading

Boulos BP *et al.* (1984). Is colonoscopy necessary in diverticular diverticula disease? *The Lancet* **i**, 95–6.

Eastwood MA *et al.* (1977). Variation in the incidence of diverticular disease within the city of Edinburgh. *Gut* **18**, 571–4.

Eastwood MA *et al.* (1978). Comparison of bran, ispaghula and lactulose on colon function in diverticular disease. *Gut* **19**, 1144–7.

Gear JSS *et al.* (1979). Symptomless diverticular disease and intake of dietary fibre. *The Lancet* **i**, 511–14.

Gianfranco JA, Abcarian H (1982). Pitfalls in the treatment of gastrointestinal bleeding with blind subtotal colectomy. *Diseases of the Colon and Rectum* **25**, 441–5.

Grief JM, Fried DO, McSherry CK (1980). Surgical treatment of perforated diverticulitis of the sigmoid colon. *Diseases of the Colon and Rectum* **23**, 483–7.

Heaton KW (1985). Diet and diverticulosis—new leads. *Gut* **26**, 541–3.

Hughes LE (1969). Postmortem survey of diverticular disease of the colon. *Gut* **10**, 336–51.

Hyland JMP, Taylor I (1980). Does a high fibre diet prevent the complications of diverticular disease? *British Journal of Surgery* **67**, 77–9.

Jensen DM *et al.* (2000). Urgent colonoscopy for the diagnosis and treatment of severe diverticular haemorrhage. *New England Journal of Medicine* **342**, 78–82.

Kettlewell MGW, Moloney GE (1977). Combined horizontal and longitudinal colomyotomy for diverticular disease: preliminary report. *Diseases of the Colon and Rectum* **20**, 24–8.

Krukowski ZH, Mattheson NA (1985). Emergency surgery for diverticular disease complicated by generalised and faecal peritonitis: a review. *British Journal of Surgery* **71**, 921–7.

Krukowski ZH, Koruth NM, Mattheson NA (1985). Evolving practice in acute diverticulitis. *British Journal of Surgery* **72**, 684–6.

Painter NS (1975). *Diverticular disease of the colon*. Heinemann Medical, London.

Reilly M (1966). Sigmoid myotomy. *British Journal of Surgery* **53**, 859–63.

Smith AN, Attisha RP, Balfour T (1969). Clinical and manometric results one year after sigmoid myotomy for diverticular disease. *British Journal of Surgery* **56**, 895–9.

Whiteway J, Morson BC (1985). Elastosis in diverticular disease of the sigmoid colon. *Gut* **26**, 258–66.

14.14 Congenital abnormalities of the gastrointestinal tract

V. M. Wright and J. A. Walker-Smith

Although present at birth, congenital abnormalities of the gastrointestinal tract usually manifest shortly after birth, but on occasion symptoms may be delayed for months or even years. For example duodenal atresia presents in the first few days of life whereas duodenal stenosis may not present until adult life.

The widespread use of ultrasound to assess the fetus allows many of these abnormalities to be recognized prenatally; in particular, abdominal wall defects, fluid-filled stomach in duodenal atresia, dilated bowel in the more distal atresias, and cystic masses. In addition, associated anomalies may indicate a major chromosome abnormality or cardiac lesion. This allows parental choice to continue with or terminate the pregnancy.

Embryology of the congenital abnormalities of the gastrointestinal tract

The primitive gut is initially a simple tube of endoderm, the muscle and connective tissue developing from the splanchnopleuric mesoderm. Cranially, the gut terminates at the buccopharyngeal membrane and caudally at the cloacal membrane. Both membranes disappear; failure of the cloacal membrane to do so results in one of the rarer forms of imperforate anus. The primitive foregut diverticulum gives rise to the respiratory system, oesophagus, stomach, duodenum to the level of the ampulla of Vater, liver, and pancreas. The primitive oesophagus lengthens rapidly, becomes narrow, and frequently the lumen is transiently obliterated. A longitudinal, ventral diverticulum of the foregut forms the trachea with ridges on either side that fuse, initially caudally with progression cranially, until the primitive respiratory system is separated from the oesophagus. Failure of this complex process results in the various forms of oesophageal atresia and tracheo-oesophageal fistula. Dilatation of the foregut distal to the oesophagus produces the stomach, initially slung from the dorsal body wall by the dorsal mesentery and from the septum transversum by the ventral mesentery. Rapid differential growth results in the stomach rotating through 90° on its long axis, the dorsal border becoming the greater curvature and the ventral border the lesser curvature. The dorsal mesentery forms the greater omentum. The ventral mesentery, into which the liver bud grows, forms the falciform ligament and coronary ligaments attaching the liver to the diaphragm, and the lesser omentum. Congenital abnormalities of the stomach are excessively rare. The liver arises as a shallow groove on the ventral aspect of the duodenum. The groove becomes tubular and invades the septum transversum and the ventral mesentery. Bile is secreted from the fifth month, and gives meconium its characteristic dark-green appearance. The mesoderm of the septum transversum forms the fibrous tissue of the liver.

The pancreas develops as two outgrowths of the duodenum. One comes from the dorsal aspect, the other from the ventral. The dorsal bud grows into the dorsal mesentery and the ventral bud is swept around dorsally into the mesentery when the duodenum rotates to the right. These two primordia fuse, the ducts fuse, and the main pancreatic duct joins the bile duct to enter the duodenum at the ampulla of Vater. If the ducts do not fuse, an accessory pancreatic duct persists. Annular pancreas is a congenital anomaly where the pancreas surrounds the duodenum, which may be atretic or intrinsically stenosed. Annular pancreas is not the primary cause of the duodenal obstruction in these cases.

The duodenum is derived partly from foregut and partly from the midgut. The loop of primitive duodenum is fixed at the pyloric end, and by the ligament of Treitz at the duodenojejunal flexure to the left of the first lumbar vertebra. By rotating to the right, the entire duodenum comes to lie retroperitoneally in a curve around the head of the pancreas. Failure of the duodenum to fix in this position is a fundamental reason for the gut failing to rotate correctly. During rapid growth the duodenal lumen is obliterated and partial or total failure of recanalization will result in the anomalies of duodenal atresia or stenosis. The small intestine and colon, suspended on the dorsal mesentery, rapidly lengthen and outgrow the primitive peritoneal cavity, and herniation occurs into the umbilical sac during the fifth week of development. Growth in length continues, the loop of bowel rotating through 90° anticlockwise, the cranial limb lengthening more than the caudal limb. About the tenth week the loops of bowel return to the peritoneal cavity, undergoing a further 180° anticlockwise rotation. The small intestine goes first, the large intestine subsequently. Thus the large intestine lies in front of the small. The caecum is initially subhepatic, the large liver occupying the right side of the abdomen, eventually retreating to the right upper quadrant and allowing growth in the length of the ascending colon. The caecum, ascending colon, and descending colon become fixed to the posterior abdominal wall; thus the small bowel is suspended from a mesentery that runs from the left side of the first lumbar vertebra to the right iliac fossa. Failure of the duodenum to rotate and fix, coupled with a failure of normal rotation of the bowel with consequent lack of normal fixation, gives rise to malrotation of the intestine. Abnormal bands run from the caecum, which lies to the left of the midline, to the region of the gallbladder and may compress the duodenum. The narrow mesentery of the small intestine predisposes to a volvulus of the entire midgut.

At the apex of the midgut loop, the primitive gut is in continuity with the extraembryonic yolk sac via the vitellointestinal duct, which runs in the umbilical cord. Obliteration and disappearance of this duct occurs, allowing the bowel to return from the umbilical sac to the enlarged peritoneal cavity. Failure of the duct to disappear may result in a Meckel's diverticulum, a band connecting the ileum to the umbilicus, a communication between the lumen of the ileum and the umbilicus, or failure of the gut to return completely to the peritoneal cavity, resulting in a small umbilical hernia.

Persistence of the umbilical sac will result in an exomphalos, with the sac containing a variable amount of gut and much of the liver. The embryology of gastroschisis is disputed. It may be due to early rupture of the umbilical sac allowing the primitive gut to extrude into the extraembryonic coelom, or failure of fusion of the lateral body folds producing a defect in the anterior abdominal wall adjacent to the umbilicus.

The midgut comprises the duodenum distal to the ampulla of Vater, jejunum, ileum, caecum, and colon as far as the left transverse colon. Atresia affecting the midgut may occur at single or multiple sites. The cause is probably intrauterine interference with the blood supply to that part of the gut which is affected, with consequent resorption of the ischaemic bowel.

The hindgut gives origin to the left third of the transverse colon, the descending colon, sigmoid, rectum, and upper part of the anal canal, and a considerable part of the urogenital system. The hindgut terminates in the primitive cloaca, which is separated from the proctodaeum (a shallow ectodermal depression) by the cloacal membrane. The primitive cloaca communicates with the hindgut and the allantois. Early in development the cloaca is joined by the pronephric ducts. A coronal septum (the urorectal septum) arises in the angle between the allantois and hindgut, grows caudally, fuses with the cloacal membrane, and divides the cloaca into a dorsal primitive rectum and a ventral primitive urogenital sinus. The cloacal membrane breaks down, establishing continuity between the endodermal hindgut and the ectodermal part of the anal canal. There are many varieties of imperforate anus. Absence of a variable length of rectum and anal canal, known as the 'high' anomaly, is frequently associated with the bowel terminating via a rectourethral or rectovaginal fistula. Ten per cent of babies with an imperforate anus will have oesophageal atresia, with or without a fistula, suggesting that the division of trachea and oesophagus and urogenital system and rectum must be occurring at a similar time in gestation, with possibly a similar mechanism producing the division. Anomalies of the urogenital system occur in a very high proportion of affected infants. Abnormalities of the ectodermal component of the anal canal result in 'low' imperforate anus.

The ganglion cells of the gut lie in the submucosa and intermyenteric plane. Ectodermal in origin, they migrate caudally along the length of the gut. Failure of migration down to the internal sphincter of the anal canal results in an aganglionic segment extending for a variable distance proximally, and is the underlying abnormality in Hirschsprung's disease.

Mucosal differentiation occurs in the early months. The inner circular muscle differentiates earlier than the outer longitudinal. Thus the fetal intestinal tract is prepared for digestion, absorption, and propulsion at a comparatively early stage in development.

Oesophageal atresia and tracheo-oesophageal fistula

The incidence of this condition is approximately 1 in 3500 live births.

The upper oesophagus ends in a blind pouch. In the majority of cases, the lower oesophagus communicates at its upper end with the trachea, that is there is a tracheo-oesophageal fistula. Although much less common, there are a number of well-recognized anatomical variations illustrated in Fig. 1.

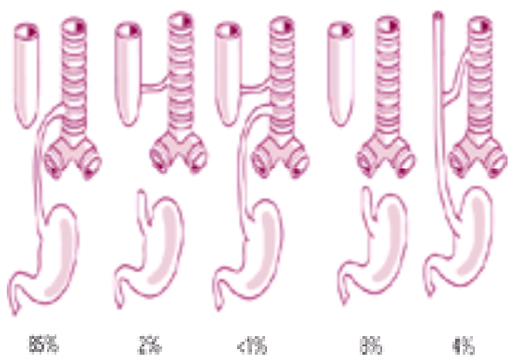

Fig. 1 Anatomical variations of oesophageal atresia and tracheo-oesophageal fistula, indicating the relative frequency.

Clinical features

Frequently the infant with oesophageal atresia is premature or small for gestational age. In 50 per cent there is a history of polyhydramnios. Shortly after birth, because swallowing is impossible, copious amounts of frothy saliva dribble from the mouth, associated with choking, dyspnoea, and cyanotic episodes. Frequent suction is required to keep the airway clear. The infant with a tracheo-oesophageal fistula without associated oesophageal atresia coughs, chokes, and becomes cyanosed during feeds. Because air escapes through the fistula into the oesophagus, gaseous distension of the abdomen is frequently present. Aspiration of feed into the airway results in pulmonary collapse/consolidation.

Over 50 per cent of infants with oesophageal atresia will have significant associated anomalies. Of particular importance are cardiac, anorectal, urogenital, and skeletal anomalies. The premature infant or the infant who is small for gestational age are more likely to have multiple anomalies than is the full-term infant.

Survival of infants with oesophageal atresia depends on birth weight and associated abnormalities. All infants with a birth weight greater than 1.8 kg and no associated abnormalities or pneumonia should survive; this is also true of the larger infant with a moderately severe associated abnormality or pneumonia. The mortality for the infant less than 1.5 kg, or one with multiple severe congenital abnormalities, remains in the region of 20 to 30 per cent.

Diagnosis

When oesophageal atresia is suspected a size 10 or 12 FG catheter is passed through the mouth and into the oesophagus. If the oesophagus is obstructed, the catheter meets a resistance 9 to 11 cm from the gum margin. A smaller catheter may curl up in the obstructed oesophagus. Contrast studies of the oesophagus are rarely necessary. A chest and abdominal radiograph will show the position of a radio-opaque tube in the upper oesophagus, and the presence of gas in the bowel if a tracheo-oesophageal fistula is present. Complete absence of gas in the abdomen is diagnostic of an oesophageal atresia without a distal tracheo-oesophageal fistula. The radiograph will also reveal any abnormalities of ribs or vertebrae, signs of pneumonia, and may provide evidence of an associated cardiac abnormality.

In isolated tracheo-oesophageal fistula, very careful contrast studies of the oesophagus are required to demonstrate the fistula. Endoscopic examination of trachea and oesophagus is usually diagnostic.

Management

Early division of the tracheo-oesophageal fistula and anastomosis of the oesophagus are possible in the majority of cases. Postoperatively, mechanical ventilation may be necessary, but usually the full-term infant with no preoperative complications only needs careful suction of the nasopharynx to maintain a clear airway. A gastrostomy or a transanastomotic nasogastric tube is usually used to enable the infant to be fed within 48 h of operation. A primary anastomosis may not be feasible in pure oesophageal atresia, extreme prematurity, or where the infant's general condition is poor. In such cases a tracheo-oesophageal fistula, if present, would be divided and a feeding gastrostomy established. Subsequently, an oesophageal anastomosis, after a delay of 4 to 6 weeks, having left the upper oesophageal pouch intact and kept empty of saliva by continuous suction, may be feasible. Alternatively, a cervical oesophagostomy is done with the intention, when the infant's condition

permits, of establishing continuity between mouth and stomach, using a length of colon, a tube of stomach, or the whole stomach. The choice depends on the surgeon's preference.

Anterior abdominal wall defects

The incidence of exomphalos and gastroschisis is approximately 1 in 3000 births. An exomphalos occurs because the intra-abdominal contents herniate through the umbilical ring into the base of the umbilical cord and are covered by a translucent membrane composed of peritoneum and amnion. Exomphalos major indicates that the diameter of the defect is greater than 5 cm, exomphalos minor that the defect is less than 5 cm. The contents of the exomphalos almost always include liver and a variable amount of bowel. On occasion, a very small amount of bowel alone herniates into the base of the cord. The diagnosis is frequently made on a prenatal ultrasonographic scan and prompts a search for associated major abnormalities, particularly anencephaly, chromosomal trisomies, major cardiac anomalies, and the Beckwith–Wiedemann syndrome. Associated abnormalities occur in 40 per cent.

The Beckwith–Wiedemann syndrome also termed the exomphalos macroglossia gigantism (EMG) syndrome, usually presents as a large-for-dates infant with a small exomphalos. The tongue is strikingly large, there are frequently ridges in the ear lobes, and a prominent naevus flammeus on the forehead. Hypoglycaemia as a result of hyperinsulinism produced by islet-cell hyperplasia is a common early problem, which may require steroids, glucagon, and rarely subtotal pancreatectomy to effect control. In the long term, children with this syndrome have an increased incidence of solid tumours, particularly nephroblastoma and hepatoblastoma.

In gastroschisis there is a full-thickness defect in the anterior abdominal wall, usually to the right of the umbilical cord. The defect is small but most of the gastrointestinal tract may be extruded through it. In contrast to exomphalos, other intra-abdominal organs are rarely eviscerated and abnormalities outside the gastrointestinal are unusual. Again, prenatal diagnosis on ultrasonographic scan is common.

Exomphalos

Clinical features

The lesion will be obvious at birth. Occasionally the membrane will rupture during, or shortly after, delivery. Careful examination for associated defects is essential.

Management

A nasogastric tube is passed to decompress the bowel. The sac can be very satisfactorily covered and supported by wrapping clingfilm around the exomphalos and the baby's trunk. Plain radiographs of chest and abdomen are taken preoperatively in order to study the cardiac contour, the intestinal gas pattern, and to look for evidence of an associated diaphragmatic hernia. If the contents of the sac can be reduced into the peritoneal cavity, the abdominal wall can be closed in layers. If closure of all layers of the abdominal wall is impossible, skin closure alone may be used, or a synthetic material such as Silastic sheeting or Prolene mesh is used to enclose the sac after suturing it to the margins of the defect. Gradual reduction of the contents into the peritoneal cavity is then possible, with delayed closure of the abdominal wall. An alternative is to paint the sac with an antiseptic solution such as 70 per cent alcohol or one of the iodine-based preparations. This results in the formation of a dry eschar that separates after some weeks, leaving a granulating surface, which gradually epithelializes. Any method that does not achieve muscle closure will leave a ventral hernia, which requires surgery at a later date.

Postoperatively, ventilatory support may be necessary. Antibiotics commenced preoperatively are continued postoperatively, particularly if an artificial material is used. Parenteral nutrition will be necessary if oral feeds cannot be given. Survival is related to the size of the lesion and the severity of any associated abnormalities.

Gastroschisis

Clinical features

Babies with this abnormality are frequently small for gestational age. After delivery, heat loss from the exposed bowel rapidly causes hypothermia. Hypoproteinaemia is very common. The small size of the defect in the anterior abdominal wall and the often narrow pedicle from which the bowel is suspended may impair the blood supply and result in infarction of much of the extruded intestine. Atresia may have occurred because of intrauterine impairment of the blood supply.

Management

A nasogastric tube is passed and the bowel decompressed. The bowel can be enclosed in clingfilm wrapped around the baby's trunk, or the baby can be placed in a large polythene bag taped around the chest. This keeps the bowel moist and prevents excessive heat loss. Antibiotics are commenced preoperatively and colloid is given to counteract the existing hypoproteinaemia and hypovolaemia. At operation the anterior abdominal wall is stretched and any meconium washed out *per rectum* to reduce bulk. Reduction of the extruded bowel is attempted and abdominal wall closure achieved where possible.

In about 10 per cent of cases, primary closure is not possible and a Silastic sheet or Prolene mesh is used to form an artificial sac to enclose the intestine. The material is sutured to the margins of the defect and the size of the sac gradually reduced over some days, squeezing the bowel back into the peritoneal cavity until closure of the abdominal wall becomes feasible—usually after 10 to 14 days. Ventilatory support postoperatively is often necessary. Parenteral nutrition is essential and may need to continue for many weeks until gastrointestinal motility and absorption are adequate. Sepsis is a considerable hazard. The mortality is now 5 to 10 per cent compared with 80 per cent 10 years ago. Improved postoperative management is largely responsible for this.

Congenital pyloric stenosis

Congenital hypertrophic pyloric stenosis is a disorder characterized by hypertrophy of the circular muscle of the pylorus and so obstruction to the gastric outlet. The incidence is 2 per 1000 live births. The aetiology is unknown. Theories include primary muscle hypertrophy, abnormalities of the maturation of ganglion cells, absence of a certain type of ganglion cell, or a response to abnormally high concentrations of circulating gastrin. Genetic and environmental factors play an important part. There is an increased incidence of pyloric stenosis in siblings of an affected child and in the offspring of a woman who has had the condition. Environmental factors include social class, type of feeding, and a seasonal variation with an increase in the winter months. In any large series the male:female ratio is 3 or 4:1 and half the cases will be first-born children.

Clinical features

The onset of symptoms is usually between 3 and 6 weeks of age, but may present shortly after birth. Vomiting of increasing severity is the cardinal symptom, eventually occurring after most feeds and becoming projectile. The vomitus is milk and mucus, and may contain altered blood suggesting an oesophagitis or gastritis; bile is never present. The baby stops gaining weight and becomes constipated. Characteristically the baby is alert, anxious, and hungry. If diagnosis is delayed, severe malnutrition may develop.

Examination reveals evidence of weight loss and in advanced cases signs of dehydration will be evident. When the stomach is full, waves of peristalsis travelling from left to right in the epigastrium will be seen (visible peristalsis). The thickened pylorus is felt as an olive-sized tumour lying deep to the edge of the right rectus and is often most easily felt when the stomach is empty. The diagnosis of pyloric stenosis is made on clinical grounds in the majority of cases. A plain radiograph of the abdomen may be very helpful in revealing a large stomach with a paucity of distal gas. A barium meal is diagnostic when the 'string' sign of the elongated pylorus is demonstrated. The barium study may also reveal gastro-oesophageal reflux, which is commonly associated with pyloric stenosis. Ultrasound is now widely used—pyloric length more than 1.2 cm and wall width more than 3 mm supporting the diagnosis.

Management

In the child presenting early, electrolyte disturbance and dehydration are minimal. In the later case, dehydration with hypochloraemic alkalosis and marked potassium depletion occurs. Preoperative correction of water and electrolyte deficits is essential. The operation of pyloromyotomy, described by Ramstedt in 1912, splits the hypertrophied muscle longitudinally allowing the mucosa to bulge through the defect, thus enlarging the pyloric canal. Postoperatively, various feeding regimens are

advocated; all aim to have the baby on a normal feeds by 48 to 72 h postoperatively. The prognosis is excellent.

Atresia and stenosis of the small intestine

An intrinsic obstruction may produce either complete or partial obliteration of the bowel lumen. Complete obliteration may be due to a gap between the two ends of the small intestine, with or without a connecting band between these ends, or a complete mucosal diaphragm. Such complete obstruction is known as atresia. When obstruction is incomplete it may be due to a narrowing of the lumen—a stenosis—or a mucosal diaphragm with a hole. Small-intestinal atresia is a more common finding than is stenosis. The duodenum is most often affected, followed by jejunum, and least often ileum.

Associated abnormalities of the gastrointestinal tract, including malrotation, oesophageal atresia, imperforate anus, biliary atresia, and annular pancreas are a feature of duodenal atresia/stenosis. Localized volvulus and meconium ileus are associated with jejunoileal atresias.

Intrinsic obstruction of the small intestine of congenital origin presents most often in the neonatal period but when the obstruction is partial it may first present much later, in infancy and childhood.

Congenital intrinsic duodenal obstruction

When duodenal obstruction is complete, vomiting usually occurs within a few hours of birth and is bile stained unless the obstruction is proximal to the ampulla of Vater, when the vomiting is persistent and copious but not bile stained. Meconium may be passed normally and there may be obvious epigastric distension. In view of the association with other abnormalities, these should be sought carefully. In particular the infant should be examined for evidence of Down's syndrome. Duodenal lesions are an association of this syndrome and occur in 10 per cent of cases.

When obstruction is incomplete the symptoms may be intermittent and the diagnosis delayed.

Congenital intrinsic duodenal obstruction may be accompanied by an annular pancreas; this is a sign of failure of duodenal development rather than an obstructive lesion *per se*. In infants with duodenal atresia, at operation, it often looks as if there is an annular pancreas because there is interposition of the pancreas between the two ends of the duodenal atresia.

Congenital intrinsic duodenal obstruction is not, in general, associated with multiple atresias in the remainder of the small intestine, but there may be obstruction at two levels in the duodenum.

Jejunoileal obstruction

Symptoms, typically bile-stained vomiting and abdominal distension, usually occur within the first 2 days of life. Meconium may or may not be passed. When obstruction is incomplete the diagnosis may again be long delayed and the child may present with intermittent vomiting, abdominal distension, and even with features of malabsorption—a clinical picture that may resemble coeliac disease.

Diagnosis

Plain radiographs of the abdomen are usually diagnostic in infants who present with a complete obstruction. In duodenal atresia there is the characteristic 'double bubble' (Fig. 2). When duodenal obstruction is incomplete there may be small amounts of air in the lower bowel. A barium meal may be necessary to demonstrate the obstruction and may suggest an associated malrotation. When there is complete jejunoileal obstruction there are usually multiple dilated loops of intestine. A barium enema may reveal an unused microcolon. When obstruction is incomplete a barium follow-through may be needed to establish the diagnosis. Rarely, laparotomy may be the final court of appeal.

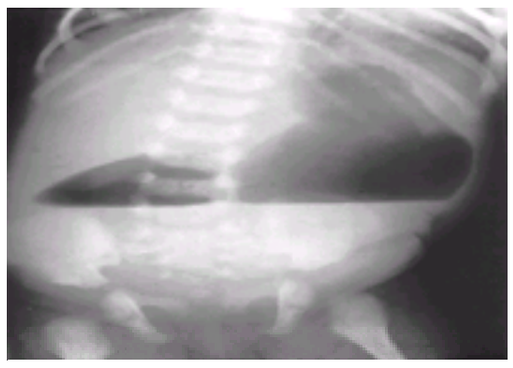

Fig. 2 Plain radiograph of the abdomen of an infant with duodenal atresia showing characteristic 'double bubble'.

Management

A nasogastric tube is passed to empty the stomach and allow accurate measurement of gastric losses. Correction of fluid and electrolyte disturbances, if present, should precede surgery, provided that gangrenous or ischaemic bowel is not suspected. At laparotomy, care should be taken to exclude any other gastrointestinal abnormality. In duodenal obstruction, the operation of choice is duodenoduodenostomy. In jejunoileal lesions, adequate resection of the proximal dilated gut reduces the great discrepancy in size between the two blind ends and so facilitates end-to-end anastomosis, although an oblique-to-end anastomosis is sometimes necessary. Leaving the dilated gut immediately proximal to the anastomosis results in ineffective peristalsis and delay in establishing enteral feeds.

Considerable loss of intestinal length may occur as a result of the intrauterine process producing the atresia; surgical correction, particularly of multiple atresias, will result in further loss. Every effort is made to preserve some ileum and the ileocaecal valve. Loss of considerable lengths of jejunum is well tolerated. Loss of ileum, particularly if the ileocaecal valve is also lost, presents management problems throughout childhood because malabsorption of a variety of important nutrients occurs. The enterohepatic circulation may be impaired. Early liver damage is a consequence of prolonged parenteral nutrition and episodes of sepsis.

Duplication of gastrointestinal tract

Definition

Duplications are cystic or tubular structures whose lumen is lined by a mucous membrane, usually supported by smooth muscle. They occur most often within the dorsal mesentery of the gut. They are also sometimes described as enteric cysts, neurenteric cysts, and reduplications. Duplications may occur anywhere along the alimentary tract but they are found most often in relation to the small intestine, particularly the ileum. They may not communicate with the lumen of the gastrointestinal tract. Duplications may be found in association with intestinal atresias. Sometimes those associated with the small intestine are lined by gastric mucosa and peptic ulceration of the adjacent small-intestinal mucosa, with bleeding, may occur. Those associated with the colon never contain ectopic gastric mucosa.

Clinical features

These are congenital malformations that present most often in early infancy. Later presentation, even into adult life, is well recognized. Duplications may present in infancy as a small-bowel obstruction, or a small cystic duplication may form the lead point of an intussusception. A palpable abdominal mass in infancy, as well as rectal bleeding and volvulus, may also be modes of presentation of this disorder. The clinical diagnosis is often difficult and the diagnosis may sometimes be made

only at laparotomy. A technetium scan may be helpful by demonstrating ectopic gastric mucosa. Initial presentation may be a posterior mediastinal cystic mass, possibly associated with cervical or upper thoracic vertebral abnormalities. The mass is likely to communicate through the diaphragm with an intestinal duplication.

Management

Excision of a cystic duplication with or without the adjacent intestine is usually straightforward. Any associated thoracic cyst will also need excision. Short tubular duplications can be excised with the adjacent intestine; very extensive tubular duplications can be opened longitudinally and the mucosa stripped out, leaving the common muscle wall.

Small-intestinal malrotation with or without volvulus

Malrotation of the small intestine is due to disordered movement of the intestine around the superior mesenteric artery during the course of development of the embryo.

Two main abnormalities that produce symptoms may occur. First, there is a gross narrowing of the base of the mesentery, which may allow the midgut to twist around and cause a volvulus. This may occur acutely, causing complete obstruction, or it may occur intermittently, producing bouts of partial or complete obstruction that release themselves spontaneously. Secondly, there may be partial duodenal obstruction from extrinsic compression of the small intestine by peritoneal bands (Ladd's bands) that extend from the caecum to the subhepatic region.

Malrotation may be associated with duodenal atresia or stenosis. It is also found in association with diaphragmatic hernia, omphalocele, and gastroschisis. However, malrotation may be asymptomatic and is sometimes discovered only as an incidental finding on a barium study. The majority of children who develop symptoms related to malrotation do so within the neonatal period, presenting with features of intestinal obstruction, complete or incomplete. When there is a volvulus there may also be obstruction to the blood supply to the bowel, which if complete will lead to extensive gangrene of the small bowel. The passage of bloody stools may be an early sign of this complication.

Those children with malrotation who present later in childhood may do so with features of intermittent obstruction such as episodes of vomiting, often bile stained, and abdominal pain, but sometimes they may manifest with features of malabsorption and many clinical features suggestive of coeliac disease. This is due to intestinal stasis with bacterial overgrowth in the lumen of the small intestine. Steatorrhoea may be accompanied at times by protein-losing enteropathy from obstruction of the mesenteric lymphatics, and chylous ascites may also occur.

Diagnosis

The diagnosis needs to be considered in the differential diagnosis of small-intestinal obstruction in infancy.

Plain radiographs of the abdomen may be very useful, typically revealing an air-filled stomach with some gas scattered through the lower part of the abdomen. However, a malrotation may not be accompanied by any abnormality on the plain radiograph of the abdomen and a barium meal will then be necessary to reveal the presence of malrotation by outlining the failure of the duodenum to cross to the left of the vertebral bodies with the fourth part lying adjacent to the first lumbar vertebra. A barium enema may be useful if it demonstrates the abnormal position of the caecum, but a barium meal is more reliable.

Management

Surgical intervention is indicated when a firm diagnosis is established. Ladd's operation is usually the procedure of choice. This involves, in general, the placement of the colon on the left and the small intestine on the right, having divided any bands and adhesions between the duodenum and large bowel, and, by dissection, broadened the base of the mesentery as much as possible. After a volvulus, total bowel necrosis is untreatable, but severe bowel ischaemia can be reversible and a 'second look' laparotomy may be necessary.

Small-intestinal lymphangiectasia

Small-intestinal lymphangiectasia has been described as a primary, that is a congenital, abnormality or as a secondary manifestation of some other disease process such as constrictive pericarditis. The primary abnormality may be accompanied by generalized lymphatic abnormalities including lymphoedema, chylous ascites, and hypoplasia of the peripheral lymphatic system, but the lymphatic abnormality may be confined to the small bowel and its mesentery. It is usually, but not invariably, accompanied by hypoproteinaemic oedema. Radioisotope studies have demonstrated that the hypoproteinaemia is due to abnormal protein loss into the gut. The pathogenesis of the hypoproteinaemia has been attributed to the rupture of dilated lymphatic channels or to protein exudation from intestinal capillaries via an intact epithelium, where there is obstruction of lymphatic flow.

Clinical features

It is a rare condition, which may present throughout life but most often in the first 2 years with diarrhoea and failure to thrive and, later, generalized oedema with hypoproteinaemia. The clinical picture may resemble coeliac disease. There is lymphopenia in the presence of a normal bone marrow and reduction of serum albumin, serum IgG, and carrier proteins such as protein-bound iodine. The severe protein loss may be accompanied by enteric calcium loss, leading to hypocalcaemia. Steatorrhoea is often found in this disorder.

Diagnosis

Diagnosis is made by showing the characteristic lymphatic abnormality on small intestinal biopsy, that is dilated lacteals, but the lesion is patchy. One negative biopsy does not exclude the diagnosis. Radioisotope demonstration of abnormal enteric protein loss using a technique such as intravenous $CrCl_3$ is helpful in diagnosis but is not specific. Barium studies in most cases show coarse mucosal folds.

Pathology

Autopsy studies reveal a considerable variation in the distribution of the lymphatic abnormality along the length of the small intestine. Dilated lacteals may occur irregularly along the small bowel and there may be gross dilatation of lymphatics projecting into the lumen. Lymphatic proliferation and dilation may also occur within the mesentery, as well as the serosal, muscular, and submucosal layers of the small-intestinal wall, and extend into the lymph nodes and occupy part of the nodal tissue.

Treatment

This is usually dietetic, as the lymphangiectasia is rarely localized enough to allow surgical excision to effect a permanent cure. The amount of long-chain fat in the diet, which is normally absorbed via the intestinal lymphatics, should be limited. This leads to a reduction in the volume of intestinal lymph and in the pressure in the dilated lymphatics. It is best done by placing the child on a low-fat diet (5–10 g/day) and adding medium-chain triglycerides, instead of the usual long-chain dietary fats, in unrestricted amounts. A milk containing medium-chain triglyceride such as Pregestimil ^{R} may be used with medium-chain triglyceride oil for cooking. Some children may be resistant to this therapy when the abnormality is very extensive and, on occasion, death may result despite therapy. Albumin infusions are of little value in management as their benefit is so transitory. Steroids have been advocated but there is little evidence to justify their use. In a follow-up study of children, although there was a continuing chyle leak, as shown by persistent lymphopenia and hypoalbuminaemia, there was a rapid and sustained improvement in dependent oedema following the use of the diet recommended above, although asymmetrical oedema from peripheral lymphatic abnormalities was unaffected. Their growth rate improved on the diet. Clinical relapse occurred quickly when the diet was relaxed. Continued adherence to a strict diet, at least through puberty, is therefore recommended. Indeed it seems probable that this is a life-long disorder and that some dietetic management may usually need to be permanent.

Meckel's diverticulum

This diverticulum is the vestigial remnant of the vitellointestinal duct. Although most people who have such a diverticulum are asymptomatic, complications may arise, which may present in a variety of ways. In children, these complications chiefly arise in association with the presence of ectopic gastric mucosa in the diverticulum. Other ectopic tissue, for example pancreatic tissue and colonic mucosa, may be found in some cases.

The diverticulum is located in the distal ileum within 100 cm of the ileocaecal valve. It is always antemesenteric.

Clinical features

Rectal bleeding is the main symptom. This is usually the passage of bright blood rather than tarry melaena stools. Typically the stool is at first dark in colour but later bright red. Bleeding may be acute, with shock requiring urgent blood transfusion, or it may be chronic. From a practical viewpoint any child who has a massive, painless, rectal bleed should be regarded as having a Meckel's diverticulum until proved otherwise. Most often bleeding from a Meckel's diverticulum is associated with ulceration of the small bowel adjacent to ectopic gastric or pancreatic mucosa but this is not always the case as bleeding may occur in the absence of ectopic mucosa.

Small-intestinal obstruction may also be a mode of presentation. This may be as a volvulus associated with a band, or an intussusception with the diverticulum as the lead point. Acute diverticulitis occurs and may produce a picture indistinguishable from acute appendicitis.

Diagnosis and management

This depends upon the mode of presentation. When rectal bleeding occurs, other causes need excluding. Investigation may include colonoscopy to exclude colonic causes and upper endoscopy to exclude peptic ulceration or oesophagitis.

Barium follow-through is usually an unrewarding investigation. a technetium scan is usually the most important investigation. The radionuclide technetium-99m concentrates in the gastric mucosa. When it is given intravenously, ectopic gastric mucosa appears as an abnormal localization on abdominal imaging with a gamma-camera. In this way a Meckel's diverticulum with ectopic gastric mucosa or indeed a duplication with such ectopic tissue may be diagnosed. A negative scan may prompt angiography. However, negative investigations in a child with severe bleeding should not deter a surgeon from proceeding with a diagnostic laparotomy, or laparoscopy if appropriately skilled. Indeed, when considering the other modes of presentation of Meckel's diverticulum it is often only at laparotomy that the role of a Meckel's diverticulum in the child's intestinal pathology is appreciated.

Meconium ileus

This is a manifestation of cystic fibrosis, the disorder sometimes known as fibrocystic disease of the pancreas. Meconium ileus is the earliest mode of presentation of this disorder during the neonatal period. A similar syndrome in older children and young adults who have cystic fibrosis may occur—the meconium ileus equivalent. The abnormally viscid consistency of the meconium produces an intraluminal obstruction. It may result from several factors including the lack of pancreatic enzymes during fetal life, which may account for the high protein content of the meconium. There is also evidence of reduced secretion of water and electrolytes in such infants, which may further render the meconium more viscid. The meconium, because of its high viscosity and tendency to adhere to the mucosa, cannot be propelled along the bowel and so small-intestinal obstruction results. This occurs most often in the distal ileum.

Clinical features

The neonate with this disorder usually develops signs of intestinal obstruction within the first 24 to 48 h of life, with the classical signs of bile-stained vomiting, progressive abdominal distension, and failure to pass meconium. In simple meconium ileus, the meconium is the sole source of the obstruction, but meconium ileus may be complicated by perforation of the gut and, when this occurs *in utero*, intraperitoneal calcification may be observed on a plain radiograph of the abdomen, providing evidence of meconium peritonitis. Perforation may also occur in the neonatal period. Volvulus and atresia may also complicate meconium ileus.

In simple meconium ileus, the plain radiograph of the abdomen may show dilated bowel but few fluid levels. Sometimes there is the appearance of bubbly meconium in the right lower quadrant. Bowel loops may be palpable. If a contrast enema is performed a microcolon, a consequence of disuse, will be demonstrated. Atresia associated with meconium ileus is frequently indistinguishable radiologically from an atresia of ischaemic origin.

Management

When meconium ileus is complicated by atresia or perforation, gangrene, peritonitis, or associated volvulus, surgical intervention is essential. Surgical options include the formation of a double-barrelled stoma with subsequent irrigation of the meconium from the distal bowel over a week or so, or intraoperative irrigation of the bowel with an immediate end-to-end anastomosis. In both options, an associated atresia or necrotic bowel are resected. The treatment of uncomplicated meconium ileus using enemas containing pancreatic enzymes, mucolytic agents such as acetylcysteine, and the detergent Tween 80 had been advocated for some time. Noblett in Melbourne, in 1969, used a Gastrografin enema to relieve intraluminal obstruction. Gastrografin is a radio-opaque, hyperosmolar solution that is effective because of its hypertonicity. This technique should not be used until a plain radiograph of the abdomen has excluded the possibility of complicated meconium ileus. An initial barium enema should exclude Hirschsprung's disease and demonstrate a microcolon extending to the proximal colon. The retrograde passage of contrast medium through the ileocaecal valve should demonstrate intraluminal meconium with passage into proximal dilated ileum, thus excluding an ileal atresia. After a successful Gastrografin enema, large amounts of meconium will be passed.

Although there may be no signs clinically or radiologically of pulmonary complications in the neonatal period, physiotherapy should be started and any chest infections treated with antibiotics when they occur (as for older children with cystic fibrosis). A pancreatic enzyme preparation should also be started, at first in small dosage when milk feedings have begun. The diagnosis should be confirmed by sweat electrolyte estimations; concentrations of sweat sodium above 60 mmol/l are abnormal. In the majority of infants with cystic fibrosis, the finding of the abnormal gene † *F508* or one of the other recognized mutations confirms the diagnosis. In a minority the abnormal gene is not identifiable.

Congenital short intestine

There is a syndrome of congenital short intestine in association with malrotation with clinical features similar to those that follow massive intestinal resection. There is also another syndrome of congenital short intestine in association with pyloric hypertrophy and malrotation. This latter syndrome is due to an absence or diminution of argyrophil ganglion cells in the small-intestinal wall. These cells normally organize peristalsis and ensure that the bolus moves forward at the correct speed. In the absence of such innervation, smooth muscle of the small-intestinal wall contracts spontaneously and rhythmically, but segmentation is not co-ordinated and the food bolus does not move forward, and there is work hypertrophy of smooth muscle. Both syndromes are rare and often only diagnosed at laparotomy.

Colonic atresia

Atresia of the large intestine is rare. In any series of cases of intestinal atresias, fewer than 10 per cent will have isolated colonic atresia.

Clinical features

The baby presents in the first 24 to 48 h with marked abdominal distension, vomiting, and failure to pass meconium.

Diagnosis

Abdominal radiographs reveal multiple dilated loops of bowel with fluid levels; the position of the loops may suggest a large bowel obstruction. Confirmation of the level of the atresia is obtained by barium enema.

Management

Nasogastric suction and intravenous fluids are commenced preoperatively. At laparotomy the lesion may be an isolated atresia or associated with multiple atresias of small and large bowel. If the atresia is solitary, it may be possible to perform an anastomosis after resection of the atresia and a length of the grossly dilated proximal bowel. Frequently a colostomy is fashioned to allow the dilated proximal bowel to contract before an end-to-end anastomosis some weeks later.

Hirschsprung's disease

In this condition, ganglion cells are absent in the bowel wall. The distal rectum is always aganglionic and the aganglionosis extends proximally for a variable distance. In 70 per cent the rectosigmoid is involved, in 20 per cent the aganglionosis extends proximal to the sigmoid for a variable distance up the colon, and in 10 per cent the aganglionosis extends into the small intestine. The aganglionic bowel is incapable of co-ordinated peristalsis and passively constricts, resulting in a mechanical obstruction. The incidence is approximately 1 in 5000 births.

Clinical features

Hirschsprung's disease is not associated with a high incidence of prematurity, and most of the babies have a birth weight appropriate for gestational age. This contrasts sharply with most of the other congenital obstructions of the alimentary tract. Associated abnormalities are rare. The most important association is with Down's syndrome.

Symptoms of Hirschsprung's disease are present in the first few days of life in almost all cases. Exceptionally, a baby will have no symptoms during the early neonatal period. The major symptoms are failure to pass meconium within 36 h of birth, abdominal distension, vomiting, and poor feeding. These may occur singly or in combination. Frequently, a rectal examination will relieve the obstruction by passively dilating the aganglionic segment. Twenty to 50 per cent of patients with Hirschsprung's disease are not diagnosed in the early weeks of life. Later presentation is with constipation that dates back to the neonatal period. It is not accompanied by soiling and is frequently associated with failure to thrive. Presentation may be delayed for months or years.

Hirschsprung's enterocolitis may be the mode of presentation in the infant of a few weeks of age. This condition, the precise cause of which is unknown, presents with abdominal distension, profuse diarrhoea, and circulatory collapse. The infant is gravely ill and the mortality is 20 per cent. The child with this complication, successfully treated initially, may have absorptive problems for some time, suffer recurrent episodes of enterocolitis despite successful surgery, and the surgery is attended by a higher rate of complications. The incidence of enterocolitis can be greatly reduced if the diagnosis of Hirschsprung's disease is made in the first week of life.

Diagnosis

In the neonatal period a plain abdominal radiograph will reveal distension of small and large bowel. A barium enema may show the narrow aganglionic bowel with dilated proximal bowel (Fig. 3) but a normal barium enema does not exclude Hirschsprung's disease. A 24-h film showing retained barium in the colon is often more helpful than the actual enema in confirming the clinical suspicion of Hirschsprung's disease. The definitive diagnostic procedure is a rectal biopsy. Suction biopsy enables the pathologist to look for ganglion cells in the submucosal plexus; full-thickness biopsy provides the intermyenteric plexus as well but this is usually unnecessary. In Hirschsprung's disease, ganglion cells are absent, hypertrophic nerve trunks are present, and if a histochemical stain for acetylcholinesterase is used, this reveals excessive amounts of this enzyme in the bowel wall. Anorectal manometry in Hirschsprung's disease typically shows failure of relaxation of the internal sphincter in response to rectal distension but this reflex is frequently absent in normal term babies until after the second week of life. This method of diagnosis is therefore unreliable in the neonatal period, requires considerable expertise to obtain reliable results, and cannot be regarded as suitable for the routine diagnosis of Hirschsprung's disease.

Fig. 3 Barium enema in Hirschsprung's disease illustrating a narrow aganglionic rectum with dilation proximally.

Management

Following diagnosis, either definitive surgery is carried out or a colostomy is fashioned in ganglionic bowel and definitive surgery deferred for a period of time. Definitive surgery consists of excision of aganglionic bowel with a 'pull through' procedure, enabling an anastomosis to be made between the anus and ganglionic colon. The three operations most often performed are those described by Swenson, Duhamel, and Soave. Provided that the surgery is uncomplicated, the long-term complications, which include faecal and urinary incontinence, and impotence, should be minimal. Bowel control is likely to be imperfect for a number of years, with soiling as a major problem, but good bowel control will be achieved in the majority of patients treated by experienced surgeons.

Imperforate anus

The exact incidence of this abnormality is not known but the usual incidence quoted is 1 in 5000 births. The basic classification differentiates between the high anomalies, where the bowel terminates above the pelvic floor, the bowel narrowing down to communicate with the urethra in the male (a rectourethral fistula) and the vagina or vestibule in the female (a rectovaginal/vestibular fistula) in the majority of cases. In the low anomalies, the bowel passes through the pelvic floor and either opens on to the perineum in an ectopic position, or lies just beneath the skin-covered anus. The high anomaly is more likely to occur in boys, the low in girls. Overall, more boys than girls present with an imperforate anus. Associated anomalies of the urogenital tract, oesophagus, heart, and skeletal system are common.

Clinical features

Early examination of the perineum will establish the presence of an anorectal anomaly. In the male, the presence of meconium on the perineum usually indicates a low anomaly. In the female, careful inspection is necessary to differentiate meconium being passed *per vaginum*, indicating a high anomaly, from meconium emerging from a perineal site, suggesting a low anomaly. Careful probing of any opening will enable the direction in which the bowel is running to be established. In the female, doubt about the precise anatomy of the anomaly may be resolved by contrast studies. In the male, differentiating a completely covered anus from a high anomaly may be difficult in the early hours after birth. Examination of the urine microscopically may reveal the presence of squamous cells or debris, suggesting a fistula between bowel and urethra. Occasionally, meconium is passed *per urethra*.

A lateral film of the pelvis taken after the infant has lain 'bottom up' over a foam wedge for some minutes will often reveal the level at which the rectum terminates, but this film cannot be reliably interpreted in the first few hours after birth because air may not have reached the distal bowel. In boys, a micturating cystourethrogram will demonstrate a rectourethral fistula in a high proportion of cases, but is rarely necessary as an initial diagnostic procedure. Having defined the nature of the anorectal

anomaly, evidence of any associated abnormality should be sought by careful clinical examination and radiographs of chest, abdomen, and the vertebral column.

Management

A low anomaly usually requires a perineal procedure to enlarge the opening. Dilatation alone may suffice, but in the majority of cases a simple anoplasty produces a more satisfactory result. In the long term, the functional results for the low anomalies should be very good. A high anomaly necessitates a defunctioning colostomy in the neonatal period. Definitive surgery involves division of any fistula and positioning the bowel accurately within the pelvic floor and sphincter muscles. Delay in achieving bowel control is common and a number of secondary operations designed to improve control have been advocated. However, if the initial surgery is meticulous, acceptable continence should be achieved in over 80 per cent of children within the first 10 years. A permanent colostomy should rarely be necessary. The high incidence of associated genitourinary abnormalities makes it mandatory to investigate carefully the urinary tract at an early stage. The mortality for anorectal anomalies is largely dictated by the presence of other serious abnormalities.

Further reading

Brown RL, Azizkhan RG (1999). Gastrointestinal bleeding in infants and children: Meckel's diverticulum and intestinal duplication. *Seminars in Pediatric Surgery* **34**, 202–9.

Dalla Vecchia LK, Grosfeld JL, Cono J, Khoury MJ, Weatherly MR, Moore CA (1998). Intestinal atresia and stenosis: a 25 year experience with 277 cases. *Archives of Surgery* **133**, 490–6.

De Backer AI, Parizel PM, De Schepper A, Vaneerdeweg W (1997). A patient with congenital short small bowel associated with malrotation. *Journal Belge Radiologie* **80**, 71–2.

Freeman NV, Burge DM, Griffiths DM, Malone PSJ, eds (1994). *Surgery of the newborn.* Churchill Livingstone, London.

Langer JC (1996). Gastroschisis and omphalocele. *Seminars in Pediatric Surgery* **5**, 124–8.

Larsen WJ (1997). *Human embryology.* Churchill Livingstone, New York.

Pierro A, Fasoli L, Kiely EM, Drake D, Spitz L (1997). Staged pull-through for rectosigmoid Hirschsprung's disease is not safer than primary pull-through. *Journal of Pediatric Surgery* **32**, 505–9.

Roberts HE, Cragan JD, Cono J, Khoury MJ, Weatherly MR, Moore CA (1998). Increased frequency of cystic fibrosis among infants with jejunoileal atresia. *American Journal of Medical Genetics* **78**, 446–9.

Shaul DB, Harrison EA (1997). Classification of anorectal malformations—initial approach, diagnostic tests, and colostomy. *Seminars in Pediatric Surgery* **6**, 187–95.

Swaniker F, Soldes O, Hirschl RB (1999). The utility of technetium 99m pertechnetate scintigraphy in the evaluation of patients Meckel's diverticulum. *Journal of Pediatric Surgery* **34**, 760–4.

Veereman-Wauters G (1996). Normal gut development and postnatal adaptation. *European Journal of Pediatrics* **155**, 627–32.

14.15 Tumours of the gastrointestinal tract

A. F. Markham, I. C. Talbot, and C. B. Williams

Introduction

Breakthroughs in our understanding of the basic molecular pathology of tumours of the gastrointestinal tract have provided a paradigm for explaining the development of malignancy. Insights into the mechanisms of carcinogenesis in specific tumours of the gut now inform aspects of their clinical management, but carcinomas of the digestive tract remain difficult to treat successfully. Indeed, only minor improvements in 5-year survival rates have been achieved over several decades. The realization that certain gene products are overexpressed early in the development of specific malignancies has led to the evaluation of known drugs in chemoprevention studies. The best example is the use of non-steroidal anti-inflammatory drugs (**NSAIDs**) to inhibit colorectal adenoma and cancer development in predisposed individuals. NSAIDs inhibit cyclooxygenase-2 (**COX-2**), which is upregulated in early colonic adenomas and may drive carcinogenesis. The ability to discriminate between genetic variants (phenocopies) of colorectal and other gastrointestinal cancers will allow more definitive clinical trials of specific treatments for particular forms of these conditions.

Screening strategies to detect surgically resectable gastrointestinal cancers at ever earlier stages are proving successful, particularly for gastric cancer in Japan and for colorectal malignancy in many countries. An appreciation of the natural history of colorectal cancer has informed the design of screening programmes, and also enables the diagnostic yields expected from faecal occult-blood testing, flexible sigmoidoscopy, or colonoscopic examination to be predicted. The complex health economic issues that such screening programmes raise are being subjected to examination.

As well as the gene encoding the adenomatous polyposis coli (**APC**) tumour suppressor protein, which is mutated in familial adenomatous polyposis (**FAP**) (OMIM175100; this refers to the number designation of the condition in McKusick's catalogue, *Online Mendelian inheritance in man*, available online at http://www.ncbi.nlm.nih.gov/, which provides detailed clinical information on inherited conditions), seven different variants of hereditary non-polyposis colon cancer (OMIM114500) (**HNPCC** 1–7) are now recognized. Most of these involve mutation in one of the genes encoding components of the protein complex that repairs DNA mismatches introduced erroneously at replication. These Mendelian diseases are responsible for 5 per cent of colorectal cancers.

Genetic loci responsible for at least a proportion of the disease burden in juvenile intestinal polyposis (OMIM174900), mixed hereditary polyposis (OMIM601228), hyperplastic polyposis, and Peutz–Jeghers syndrome (OMIM175200) have been mapped and in some cases the mutant genes identified. Mutations in the E-cadherin gene have been shown to lead to familial gastric carcinoma (OMIM192090). The gene mutated in 'tylosis with oesophageal cancer' (OMIM148500) has been mapped to chromosome 17q24. Thus, there have been significant recent developments in all these disease areas.

Vogelstein and colleagues have added to their familiar model of the adenoma-to-carcinoma sequence of genetic alterations in colorectal cancer, with the concept of 'gate-keeper', 'caretaker', and 'landscaper' tumour suppressor genes in gastrointestinal cancer. *APC* is the classic example of a gate-keeper gene. The protein plays a key role in preventing carcinogenesis, and loss of such a tumour suppressor gene (**TSG**) leads inevitably to malignancy, with very high penetrance. The mismatch repair genes represent examples of genetic 'caretakers'. Mutations do not lead inevitably to cancer, but loss of their function leads to widespread genetic damage (characterized in the cell by 'microsatellite instability') including secondary mutation in gate-keeper genes, which can eventually generate a malignant phenotype. Thus, for example, colorectal cancer will develop in some 80 per cent of male carriers of an HNPCC mutation by the age of 85 years.

The concept of genetic 'landscaping' emerged from study of the paradox that colorectal cancers with microsatellite instability (**MSI**, sometimes referred to as 'replication error-positive' or **RER**+ tumours) did not contain the MLH1 mismatch repair protein, even though at least one allele of the gene encoding this protein appeared normal on DNA sequencing. This was the result of an epigenetic cause of tumorigenesis, which has proved to be a widespread factor in malignancy. DNA hypermethylation of CpG residues in the promoter regions of this or other genes starts a complex process involving histone deacetylation and permanent promoter silencing. This leads to the loss of function at a TSG allele, equivalent to that caused by a point mutation in the gene, or gene deletion.

In many cases, the roles of mutations in the genes discussed above in human cancers have been confirmed by introducing the corresponding mutations into transgenic mice. Although the phenotypes are by no means always identical, these animal models of human malignancy do act as model systems in which to develop the next generation of therapeutic agents.

The consensus emerging from these progressively more comprehensive, evidence-based surveys of the optimum approaches to the clinical management of gastrointestinal tumours has been incorporated herein. The reader is encouraged to consult these accessible sources directly.

Oesophageal tumours

Benign

Submucosal leiomyomas are the least rare of benign oesophageal lesions. Rarely, fibrovascular polyps, lipomas, haemangiomas, or other mesenchymal tumours are detected on barium-swallow examination. Endoscopic ultrasound can confirm the intramural localization of these lesions to exclude tracheal compression. Tumours large enough to cause dysphagia are removed by enucleation. Mucosal papillomas, foci of leucoplakia, or acanthosis nigricans also occur.

Malignant: oesophageal carcinomas

These constitute only 1 per cent of all malignancies and 6 per cent of gastrointestinal cancer, but cause some 10 per cent of all cancer deaths given their 5-year survival of less than 10 per cent. Historically, the majority of malignant tumours of the oesophagus were squamous carcinomas. However, there appears to be a current epidemic of adenocarcinoma in the lower third of the oesophagus, so that over 50 per cent of the disease in the Western world is now of this histological type.

Other malignant tumours are rare. They include metastases (primarily from the stomach), malignant melanoma, plasmacytoma, stromal tumours, and spindle-cell

carcinoma. Some 50 per cent of patients dying with AIDS have gastrointestinal Kaposi's sarcoma. Although the incidence in the oesophagus is less than in the mouth or hypopharynx, oesophageal extension is associated with a poorer prognosis.

Epidemiology and aetiology

Squamous oesophageal carcinoma

Only half the 3000 cases of oesophageal cancer diagnosed annually in the United Kingdom are now squamous carcinoma. A threefold greater incidence in men may reflect an association with tobacco smoking and alcohol intake. The incidence is much higher in an area extending from the Caspian Sea to China. The high incidence in this 'oesophageal cancer belt' suggested that carcinogenic *N*-nitroso compounds in pickled foods in the diet might be responsible. The possibility that these may be population isolates with predisposing genotypes or even hereditary forms of the disease has only recently begun to be explored.

The incidence of squamous carcinoma is increased in patients with caustic oesophageal strictures due to the ingestion of corrosives, or after radiotherapy. Persistent achalasia predisposes to disease, possibly related to stasis above the narrowed segment. There is also an association with coeliac disease. The Patterson–Kelly (Plummer–Vinson) syndrome, characterized by iron deficiency anaemia, glossitis, postcricoid webs, and dysphagia, particularly in Scandinavian populations, has been associated with squamous carcinoma in up to 20 per cent of cases.

The autosomal dominant disease 'tylosis with oesophageal cancer' (OMIM148500) eventually causes carcinoma in 95 per cent of patients. Pedigrees segregating the characteristic hyperkeratosis of the palms and soles, have allowed the gene for this disease to be mapped to chromosome 17q24. Its identification can be anticipated. This may allow consideration of prophylactic oesophagectomy in carriers.

A number of loss-of-heterozygosity (**LOH**) studies have been performed with squamous carcinomas and several loci show consistent allelic losses compared with matched normal tissues, as would be expected at a tumour-suppressor gene locus. Pedigrees in northern China with consanguinity, and possibly segregating autosomal recessive forms of the disease, have recently been described.

Adenocarcinoma of the oesophagus

Barrett's oesophagus is attributed to chronic gastro-oesophageal reflux, resulting in replacement of the squamous epithelium by an abnormal columnar, metaplastic epithelium. This specialized intestinal metaplasia, with mucus-secreting goblet cells, is thought to increase the risk of developing adenocarcinoma 40-fold. Increasing epithelial atypia defines low to high grades of dysplasia. The frequent association of high-grade dysplasia with adenocarcinomas in surgical-resection specimens suggests that the former is a true precursor of the latter. The claimed annual incidence of adenocarcinoma in an established Barrett's oesophagus is approximately 1 per cent.

However, optimal surveillance of individuals with Barrett's oesophagus remains ill-defined. There is no correlation between the severity of symptoms nd the extent of Barrett's metaplasia. The cancer risk in Barrett's oesophagus may have been overestimated because of publication bias towards reporting positive studies. The incidence of adenocarcinomas appears to be higher in long-segment than in short-segment metaplasia. Aggressive antireflux therapy or endoscopic ablation of the metaplastic epithelium have not been proved to decrease the risk of oesophageal cancer. There is no difference between proton-pump inhibitors and H_2-antagonists in preventing the progression of Barrett's oesophagus. Our lack of understanding of the pathophysiology of Barrett's metaplasia means it is not clear what endpoint in treating reflux disease (symptom relief, epithelial healing, or elimination of reflux) is relevant in preventing progression to oesophageal cancer.

High-grade dysplasia may produce a much higher risk of disease progression. Given the cost implications of endoscopic surveillance programmes, which may not even measure progression effectively, there is an urgent need for reliable molecular markers of high risk. A few pedigrees have been described that apparently segregate an hereditary predisposition to Barrett's oesophagus and associated adenocarcinomas in an autosomal dominant, Mendelian fashion. These may provide the opportunity to identify predisposing mutant genes. Molecular markers associated with other malignancies in the gastrointestinal tract (TP53, RB1, E-cadherin, and cyclin D1) may be informative in that, for example, high levels of cyclin D1 expression in biopsy specimens appear to correlate with an increased risk of carcinoma.

Clinical features

Patients generally present with dysphagia, initially for solids but eventually also for liquids and even saliva. Unfortunately, dysphagia usually reflects circumferential disease. Impact pain from food, or pain in the front or back of the chest, are bad prognostic features. Loss of weight and appetite more often reflect the cachexia of advanced disease, rather than simply the difficulty in swallowing. Regional lymph node involvement is present in approximately 50 per cent of cases at diagnosis. Direct spread may involve the bronchi and aorta. Perforation may result in tracheo-oesophageal fistulas or mediastinitis. Hoarseness may be apparent on involvement of the recurrent laryngeal nerve. Obstruction with difficulty swallowing saliva may result in aspiration pneumonia. However, the physical signs may be limited to weight loss, anaemia, and/or cervical lymphadenopathy.

Diagnosis and staging

Barium swallow characteristically shows abrupt, irregular lumenal narrowing, in contrast to the smooth narrowing of a benign stricture. Endoscopy allows biopsy of the lesion, producing a diagnostic accuracy of over 95 per cent. Palliative dilatation or intubation may then be performed. **TNM** (primary tumour, regional nodes, and metastasis) staging is achieved using endoscopic ultrasound to image the tumour in the oesophageal wall and to assess regional lymph nodes. Computed tomography (**CT**) or magnetic resonance imaging (**MRI**) will highlight invasion of adjacent structures, more distal lymph node involvement, and metastatic disease.

Treatment

Patients with superficial carcinomas at stage T_1N0M0 may achieve 5-year survival rates of over 50 per cent. However, overall the results of treatment are dire, with a 5-year survival rate of less than 10 per cent. Unfortunately, perioperative mortality is rarely less than 10 per cent. Adjuvant chemo/radiotherapy has been advocated for otherwise inoperable disease that has spread beyond the oesophageal wall. Squamous-cell carcinomas are more sensitive to both these treatment modalities than adenocarcinomas. Chemotherapy and radiotherapy are the subjects of Cochrane Reviews, but there is no clear evidence of improved survival in patients with potentially resectable lesions. Palliative therapy is based on radiotherapy and therapeutic endoscopy for endoscopic dilatation or insertion of a stent to overcome dysphagia and bridge any fistulas which may have formed. Ablation of tumour growing into the lumen of the oesophagus may be attempted using laser photocoagulation or electrocautery. Ethanol injection induces tumour necrosis and is effective, particularly when combined with intubation.

Stomach tumours

Benign

Benign gastric tumours may be derived from the mesenchyme or epithelium. Gastrointestinal stromal tumours (GISTs) arising from the gastric wall are most frequent, though not common. These project into the gastric lumen and occasionally become superficially ulcerated leading to haemorrhage. When large, these tumours have low grade malignant potential (see section on small bowel tumours). Even rarer submucosal lesions include lipomas, haemangiomas, hamartomas, gastric carcinoids, and lymphoid hyperplasia. The latter may be difficult to discriminate from a mucosa-associated lymphoid tissue (**MALT**) lymphoma.

Gastric epithelial polyps are mainly hyperplastic or inflammatory and usually asymptomatic. The historical view that they are not susceptible to malignant change, may be questioned. Only 15 per cent of gastric polyps are adenomatous. Although the risk of gastric cancers in these adenomas is not as clear-cut as in the large bowel, their association with FAP in particular, the eventual development of gastric cancer in FAP, and their occurrence at the margins of gastric carcinomas indicate that they should be removed endoscopically. Villous adenomas and polyps larger than 2 cm have a particularly high malignant potential. Rarely, large pedunculated polyps may ulcerate and bleed, or obstruct the pyloric outflow. Gastric hamartomas occur in Peutz–Jeghers syndrome, Cronkhite–Canada syndrome, Cowden's disease, and juvenile intestinal polyposis and these polyps are associated with an increased incidence of malignancy, highest in Peutz–Jeghers syndrome. This is obscured by the higher risk of carcinomas elsewhere in the gastrointestinal tract in these conditions. Paradoxically the most numerous gastric polyps in FAP are cystic gland (mucus retention) polyps which have no malignant potential or clinical significance whatever.

Malignant: carcinoma of the stomach

Pathology

The antrum is the most frequent site of gastric adenocarcinoma. The incidence of proximal cancers (which are smoking-related) and gastro-oesophageal junction tumours is rising for reasons that are not fully understood. Pathological classification is not entirely clear-cut. The histological approach of Lauren defines two subtypes. The intestinal type (60 per cent) is glandular and polypoid, associated with metaplasia, chronic gastritis, and the presence of *Helicobacter pylori*. The diffuse type (30 per cent) consists of scattered clusters of cells and spreads submucosally to ultimately result in linitis plastica. It has a worse prognosis and occurs in younger individuals. Tumours of mixed type (10 per cent) are also recognized.

Morphological classifications have been attempted, including that of Borrmann for advanced gastric cancer: type 1, polypoid without ulceration; type 2, fungating with surface ulceration; type 3, ulcerated with a wall-like edge and surrounding infiltration; and type 4, diffusely infiltrative. This bears some resemblance to the Japanese classification of early gastric adenocarcinomas: type 1, protruded, polypoid; type 2, superficial (elevated, flat, or depressed); and type 3, excavated. An alternative histological classification defines early cancer as not penetrating the submucosa (T1 disease), with advanced disease involving extension through the submucosa to the muscularis propria or serosa (T2), breaching the peritoneum (T3), and into adjacent tissues (T4). The TNM classification defines lymph node status as N_1 when one to 6 nodes are involved, N_2 when 7 to 15 nodes are posistive and N3 when more than 15 nodes are involved. Differences in classification between Japanese and Western pathologists have resulted in attempts at International consensus statements: the Vienna and the Padova classifications—which, unfortunately, are not identical.

An autosomal dominant, familial form of gastric cancer of diffuse type occurs, with presentation at early ages (OMIM192090). This disease is caused by germline mutations in the E-cadherin gene. The finding is of significance because the intracellular domain of this protein interacts at the epithelial adherens junction with b-catenin, an important participant in colorectal carcinogenesis. Gastric cancers frequently demonstrate *P53* mutations and consistent LOH in a number of specific chromosome regions. HNPCC provides a fourfold increased risk of gastric cancer, presenting in 10 per cent of carriers. *APC* mutations also occur in early sporadic gastric cancers.

Epidemiology and aetiology

There are approximately 10 000 new cases in the United Kingdom annually, making gastric cancer the sixth commonest malignancy, and, like oesophageal cancer, accounting for about 10 per cent of cancer deaths. There has been a sustained slow decline in incidence in the United Kingdom, but an increase in the number of cases involving the cardia. There are marked geographic differences in incidence, the rate in Japan being eight times higher than amongst Whites in the United States. Gastric cancer rates in Japanese immigrants to the United States halve in one generation, suggesting an environmental factor in its aetiology. Mass radiographic screening for gastric cancer has been undertaken in Japan since 1960. Some 5 million barium studies conducted annually detect 6000 cases. Approximately half of these are early stage, offering the potential for curative resection. Screening in the United Kingdom cannot be justified given the lower incidence. The incidence in men still approaches twice that in women, worldwide.

H. pylori has emerged as a causative agent for distal gastric cancer and is now classified as a class 1 carcinogen. It causes gastritis, gastric and duodenal ulcers, and both adenocarcinoma and lymphoma in the stomach. Some 60 per cent of the world's population are infected. In the developing world most of the population is infected from early childhood. In contrast, the prevalence of infection is falling in the developed world. This probably reflects decreased childhood infection rates, rather than eradication of infection in adults. Bacterial isolates with the *Cag-A* genotype are particularly implicated in carcinoma of the middle and distal thirds of the stomach, of both intestinal and diffuse types.

The organism can be detected by urease breath-testing or immunologically by a serum enzyme-linked immunosorbent assay (**ELISA**). Initial infection induces pangastritis with achlorhydria. There is a brisk serum immune response and the appearance of IgA in gastric secretions. However, *H. pylori* is not eradicated and gastritis is maintained by a combination of *H. pylori*-induced cytokine release from gastric epithelium and a TH1-type T-cell response. A sustained period of antral gastritis ensues, associated with duodenal ulceration. Persistent reduction in acid secretion may lead to the reduced effectiveness of antioxidant mechanisms in the stomach. Eventually, multifocal atrophic gastritis develops with metaplasia, followed by dysplasia, leading on to malignancy. This pathway remains consistent with the model originally proposed by Correa in 1988 (Fig. 1).

Fig. 1 Flow chart of the pathogenesis of gastric cancer.

The increased frequency of gastric carcinoma in lower socioeconomic groups may reflect their increased prevalence of *H. pylori* infection. A decreased incidence of *H. pylori* infection in childhood may explain the reduced rate of gastric carcinoma in Japanese immigrants to the United States. Eradication of the organism leads to resolution of the gastritis and normalization of basal acid secretion over several months. Reinfection rates in adults appear to be low in the developed world. This offers the possibility of achieving a long-term reduction in cancer incidence.

There is no convincing evidence that the prolonged use of H_2-blockers or proton-pump inhibitors is associated with an increased incidence of gastric cancer, secondary to prolonged suppression of acid secretion. Dietary factors such as nitrites or a lack of free-radical scavengers in the diet (for example, vitamin C, selenium) have long been suspected of causation in gastric carcinoma. Conversion of nitrates to nitrosamines by bacteria at neutral pH has been suggested to be carcinogenic. Smoking and increased alcohol use are associated with gastric cancer. Whether these dietary agents exert their effects indirectly through their influence on *H. pylori* infection rates remains to be established. An increased incidence in blood group-A individuals has been recognized since 1955. There is no direct connection between peptic ulcers and gastric cancers. Malignant ulcers may heal on initial treatment for dyspepsia, which can confuse the diagnosis. Furthermore, because they are both associated with *H. pylori* infection, the two conditions do tend to coincide.

Pernicious anaemia also leads to gastric atrophy in the body and fundus of the stomach and, presumably as a consequence, is associated with a threefold increase in the incidence of gastric cancer. The incidence may approach 10 per cent in patients followed up for 20 years. Intestinal metaplasia is recognized as a predisposing feature secondary to gastritis as it is found associated with carcinoma in resected specimens. However, this is again likely to be secondary to *H. pylori* infection in the main. Intestinal metaplasia and chronic active gastritis occur in gastric stumps postgastrectomy. These show an increased incidence of cancer, particularly after gastrojejunostomy. A prolonged bile reflux may be a cause of this gastritis. As in gastric atrophy, 10 per cent of patients may develop cancer during a 20-year follow-up for chronic gastritis. This is consistent with the need for multiple genetic hits to generate a malignant phenotype.

Clinical features

Although early gastric cancer may be asymptomatic, clinical suspicion of dyspepsia in older patients may lead to its early diagnosis, with the possibility of successful surgical intervention in early-stage disease. Common presenting symptoms include epigastric pain, which may be relieved or made worse by food. This pain can be constant and severe. The clinical features of anaemia, anorexia, early satiety, and weight loss suggest advanced disease at presentation. Large proximal tumours of

the cardia may cause dysphagia. Tumours of the fundus are more likely to result in anaemia and nausea. Distal tumours of the antrum may cause outlet obstruction and result in vomiting.

Gross haematemesis from gastric cancers is unusual. Metastasis occurs to the peritoneum (with ascites and possibly ovarian involvement—Krukenberg's tumour), to the liver, and latterly to the lung and other sites. An epigastric mass may be palpable in 30 per cent of patients, and suggests advanced disease. Palpable lymphadenopathy (Virchow's node) in the left superclavicular fossa (Troisier's sign) may be present. Carcinoma of the stomach is the malignancy most frequently associated with dermatomyositis or acanthosis nigricans.

Diagnosis

Although double-contrast plain imaging of the stomach can provide up to 90 per cent diagnostic accuracy, flexible endoscopy and biopsy has become the definitive diagnostic procedure (Fig. 2). Sensitivity increases with the number of biopsy samples taken from the base of an ulcer and its margins—10 biopsies of adequate depth will give a reliability approaching 100 per cent and eliminate false-negative results.

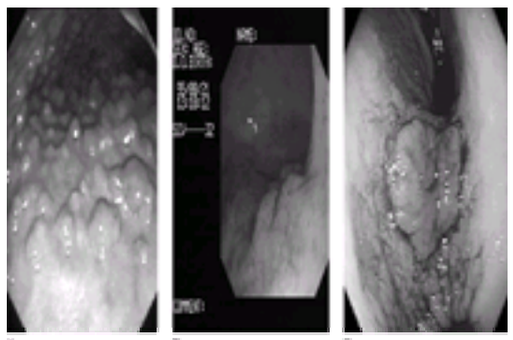

Fig. 2 Early gastric cancer – endoscopic appearance. (a) Initial view. Better view in retroversion (b) and (c) after dye-enhancement.

Staging of disease prior to attempted surgery relies on a number of imaging modalities. CT scanning defines the spread of the primary tumour and gross lymphatic and metastatic disease. CT is approximately 75 per cent accurate in detecting involved lymph nodes larger than 5 mm. However, it often fails to detect small involved lymph nodes and peritoneal deposits. Resolution may be improved with the use of intravenous contrast media. It is particularly important to detect distant involved lymph nodes because these would not normally be removed during surgery, and would therefore result in treatment failure. MRI scanning may help in this context.

Endoscopic ultrasound is useful in the local staging of gastric carcinoma. The normal stomach wall comprises five hyperechoic layers. Carcinomas present as hypoechoic lesions that disrupt the normal pattern. The depth of tumour invasion of the gastric wall can be assessed accurately, as can the presence of involved perigastric lymph nodes. T staging with 90 per cent accuracy and N staging with 70 per cent accuracy are achievable by correlation with pathological examination of specimen obtained during surgery. However, the technique is limited in its ability to detect distant enlarged lymph nodes. Classification of a tumour as stage T4 by endoscopic ultrasound renders surgical cure highly unlikely, so that palliative bypass or endoscopic techniques may be the preferred clinical option. Laparoscopy may be necessary to confirm the presence of peritoneal metastases. MRI or ultrasonography will identify hepatic metastases.

Treatment

Surgical resection of a tumour and involved lymph nodes is undertaken in the absence of metastases. The site of the tumour dictates the surgical approach. The goal is a tumour-free margin of 5 cm with lymphadenectomy to an uninvolved point along the arterial tree (defined as resection levels R1, R2, or R3, reflecting increasingly distal nodal involvement). This may require total gastrectomy, or some variant of a proximal or distal gastrectomy for tumours in these sites, ideally avoiding postoperative biliary reflux, which may itself predispose to disease recurrence. Endoscopic surgery for early gastric cancers may be possible, provided the tumours are small, confined to the mucosa, and non-ulcerated. This approach requires a high degree of certainty that local lymph nodes are not involved. Endoscopic ultrasound-guided needle biopsy may be useful in this context.

Palliative procedures in symptomatic patients with disseminated disease may also include gastrectomy, which can double survival time, or less radical interventions including stenting, feeding gastrostomies, or laser photoablation. Adjuvant chemotherapy regimens including 5-fluorouracil (combined with various other agents) have produced small survival advantages in randomized clinical trials. Radiotherapy alone has not been shown to be beneficial.

Predicting the prognosis for an individual patient is complex. The stage of the cancer at operation is crucial and 5-year survival rates approaching 90 per cent have been reported for early gastric cancer detected by screening in Japan. However, apparently superior stage-specific survival rates in Japan may reflect a lower incidence of diffuse or proximal tumours. This dependence of 5-year survival rates on stage at presentation is mirrored in the United Kingdom, where new cases present with roughly a third each at stage 1/2, stage 3, and stage 4. Whilst patients in the first category may achieve a 30 to 50 per cent 5-year survival, this falls to less than 15 per cent in those with stage 3 disease. Those at stage 4 have less than a 3 per cent 5-year survival rate.

The 5-year survival rate is marginally lower for patients with proximal tumours than distal tumours, with those with tumours in the middle third showing a slightly better prognosis. Unfortunately, the latter category of tumours (20 per cent) is less common than the others (one-third of cases each), with disseminated diffuse cancer making up the remainder. Cancers with a diffuse histology appear to have a worse 5-year survival rate than those of intestinal type.

In summary, the combination of tumour stage, lymph node involvement, position in the stomach, and histological subtype all exert subtle influences on outcome in the absence of metastases. These factors and more sensitive detection of metastatic disease leading to less operations with intent to cure, make interpretation of possible improvements in 5-year survival rates after supposedly curative resections difficult to interpret. The need for *H. pylori* prophylaxis and for improved chemotherapy approaches are pressing. The requirement for vitamin B_{12} supplementation after gastrectomy should not be overlooked.

Malignant: gastric lymphoma

Mucosa-associated lymphoid tissue (MALT) lymphomas are B-cell, non-Hodgkin's lymphomas, associated with *H. pylori* infection in over 90 per cent of cases. These represent 5 per cent of all gastric malignancies. The disease arises as a low-grade MALT lymphoma in areas of chronic gastritis, characterized by large centroblast-like cells. High-grade lesions have immunoblastic features. The clinical presentations of these patients are not readily distinguishable from gastric carcinoma. *H. pylori* eradication can result in lymphoma regression or complete remission, especially for low-grade disease, with an overall response rate of over 60 per cent. Surgery to debulk the disease, chemotherapy (using CHOP (cyclophosphamide, hydroxydaunomycin, Oncovin, prednisone) regimens), and radiotherapy are effective in high-grade disease. The disease stage and histological grade determine prognosis. In the absence of distal lymph node involvement and bulky disease, the 5-year survival rate is over 70 per cent.

Cochrane Reviews have been published on the eradication of *H. pylori* (in the context of non-ulcer dyspepsia) and Protocols have appeared for Cochrane Reviews of extended-versus-limited lymph node dissection techniques for adenocarcinoma of the stomach.

Small-bowel tumours

Given the length of the small intestine, the incidence of neoplasia (5 per cent of all gastrointestinal tumours) is remarkably low. Some two-thirds of small-bowel tumours are malignant. Attempts to explain the rarity of these tumours on the basis of the reduced carcinogenic effects of small-bowel contents, their bacterial flora, or levels of protective secretory IgA are not convincing. There are, however, clear-cut associations between specific genetic traits, including FAP and Peutz–Jeghers

syndrome, and small-bowel neoplasia.

The stem-cell compartment in the small bowel expresses low levels of the antiapoptotic protein, BCL-2, compared to the higher levels in the colonic crypt stem-cell compartment. It is possible that epithelial stem cells suffering genetic damage in the small intestine undergo apoptosis and are eliminated without the development of a neoplasm. In contrast, mutant colonic stem cells are resistant to apoptosis and persist to undergo further genetic hits, and a 50-fold increased rate of carcinoma compared to the small bowel. Paradoxically, the small bowel is the most common site in the gastrointestinal tract for metastatic melanoma.

Benign tumours

Adenomas and lipomas are the least rare, solitary benign tumours (with the exception of tumours arising in a variety of inherited forms of polyposis, discussed below). Adenomas predominate in the duodenum and are premalignant; haemangiomas and neurofibromas are more common in the jejunum; and fibromas and lipomas are more frequent in the ileum.

These tumours are frequently asymptomatic, incidental findings at operation. Symptoms where they do occur include obstruction, intussusception, pain or chronic haemorrhage with anaemia. Endoscopic examination and polypectomy in the duodenum, proximal jejunum and terminal ileum may be achievable. Barium follow-through is the appropriate diagnostic tool supplemented by angiography, for example where bleeding from an angioma or ectasia is suspected. Symptomatic benign small-bowel tumours are treated by surgical resection.

Gastrointestinal stromal tumours (GISTs) (formerly known as smooth muscle tumours) arise in the bowel wall, can grow to a large size and have low grade malignant potential, sometimes metastasizing to the liver many years after initial presentation and surgical resection. Evidence of malignancy is large size and more than 5 mitoses in 50 high power fields.

Malignant: carcinoma of the small bowel

Many adenocarcinomas of the small intestine occur in specific conditions. Management of FAP by total colectomy means that patients present with small-bowel adenomas, which have malignant potential. These occur commonly (50 per cent) in the duodenum (and rarely the stomach), particularly around the ampulla of Vater. Hamartomatous polyps are common in the jejunum in Peutz–Jeghers syndrome. These again undergo malignant change, imparting a 500-fold relative risk of small-bowel carcinoma. Coeliac disease is associated with an increased incidence of adenocarcinoma, as well as lymphoma, in the small bowel and an increased incidence of malignancies throughout the gastrointestinal tract (for example, the oesophagus) and elsewhere (for example, testes). Gluten-free diets do protect against the development of malignancy. Crohn's disease of prolonged duration slightly increases the incidence of small-bowel adenocarcinoma.

Clinical presentation usually involves abdominal pain, anorexia, cachexia, and/or anaemia. Patients may present with diarrhoea or, rarely, a palpable abdominal mass. Diagnosis involves radiological follow-through examination (Fig. 3) and CT or MRI scanning. Surgical treatment involves resection to uninvolved margins with any locally involved lymph nodes. The 5-year survival rates are dependent on tumour grade, lymph node involvement, and distant spread: with disease limited to the mucosa they approach 100 per cent, while the rate with lymph node involvement and distal metastases is essentially zero. The overall 5-year survival rate is around 25 per cent. There are very limited data on the value of adjuvant chemotherapy or radiotherapy. Combined preoperative radiotherapy and 5-fluorouracil have been claimed to be beneficial in duodenal adenocarcinoma. Where ampullary carcinomas are unresectable, palliation involves bile-duct stenting to avoid jaundice.

Fig. 3 Small bowel follow-through: compression view of mid-jejunum shows annular 'apple-core' appearance with pre-stenotic dilatation typical of an adenocarcinoma.

Carcinoid tumours are usually found in the ileum. Patients with these tumours have 5-year survival rates in excess of 75 per cent, in the absence of liver metastases. Multifocal small-bowel disease is present in approximately 30 per cent of cases. Carcinoids that are more than 2 cm in diameter have frequently metastasized when detected. Carcinoid syndrome due to excessive amounts of serotonin occurs with liver metastases. Treatment involves resection of the primary tumour and hepatic metastases, with the use of somatostatin analogues (octreotide) to control the effects of hormone excess (see Chapter 14.8).

Malignant: lymphoma of the small bowel

Primary small-bowel lymphoma occurs not infrequently in the ileum (Fig. 4) and represents 5 per cent of all lymphomas. It is also a feature of disseminated lymphoma. Lymphomas constitute one-third of primary small-bowel malignancies: distinct groups of B- and T-cell lymphomas are recognized.

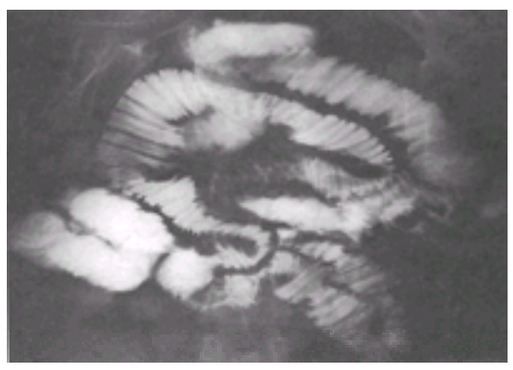

Fig. 4 Small bowel follow-through showing diffuse thickening of folds in ileum indicating submucosal infiltration (proved to be a lymphoma).

Immunoproliferative small-intestinal disease (IPSID)

These B-cell, non-Hodgkin's lymphomas occur mainly in the Middle East, Africa, and South America. This 'Mediterranean' lymphoma demonstrates an unusually strong association with low socioeconomic status. IPSID occurs in young adults as well as in the elderly. Initially the mucosa becomes infiltrated with plasma cells. It is sometimes associated with intestinal giardiasis, reminiscent of the association of gastric lymphoma with *H. pylori*. The clinical features resemble adenocarcinoma but with progressive diarrhoea, cachexia, and finger clubbing. Treatment of the infection, lymphoid infiltration, and secondary malabsorption is usually with tetracycline, or metronidazole and ampicillin. The progression to a large B-cell immunoblastic lymphoma histologically requires cytotoxic treatment with CHOP (cyclophosphamide, hydroxydaunomycin, Oncovin, prednisone) or one of its variants. IPSID frequently involves extensive areas of the small bowel so that surgical resection is often

impossible. It tends to occur proximally and is associated with the production of monoclonal immunoglobulin heavy chains, which may be detectable in the serum.

Enteropathy-associated T-cell lymphoma (EATCL)

EATCL is less common than B-cell small-bowel lymphomas. The condition develops in patients with long-standing coeliac disease or dermatitis herpetiformis. Consequently, their relative risk of a T-cell lymphoma is 50-fold higher than the general population. Clinical suspicion should be raised in patients with coeliac disease who experience increased malabsorption despite maintaining their gluten-free diet. The disease is usually high-grade and consequently of poor prognosis. The lesions tend to ulcerate with a risk of intestinal perforation. They are notoriously unresponsive to standard lymphoma chemotherapy. Again surgical resection is precluded by the large areas usually involved, particularly in the jejunum.

Other small-bowel lymphomas

Non-IPSID, non-EATCL primary small-bowel lymphomas are usually distal, and unifocal, so that surgical resection to debulk them prior to chemotherapy with CHOP, may be appropriate. Small intestinal lymphoma is frequent in patients with the acquired immunodeficiency syndrome (**AIDS**), and a form of Burkitt's lymphoma has been described in the terminal ileum in susceptible populations.

Gastrointestinal polyps and the polyposis syndromes

A tumour projecting into the lumen of the gut is termed a polyp, from the Greek *polypous* for squid, which pedunculated polyps resemble. The nomenclature used to describe the wide range of gastrointestinal tumours encompassed by this general definition can be confusing (Fig. 5). However, recent advances in elucidating the genetic basis of several inherited polyposis syndromes have clarified some of these classification difficulties. These are autosomal dominant diseases in which multiple polyps (from five or more, to many thousands in particular conditions) of a given histological type occur. All these familial forms of polyposis predispose, to some extent, to the development of carcinoma in individual polyps.

Fig. 5 Diagrammatic representation of the histology of adenomatous polyps.

In the subgroup of polyps which are adenomas, the progression from adenoma to carcinoma is well established. The molecular basis for this has been elucidated in the autosomal dominant disease familial adenomatous polyposis. The same mechanism has subsequently been shown to drive carcinogenesis in solitary adenomas occurring sporadically in individuals without the germline mutations present in the inherited condition.

Solitary hyperplastic (metaplastic), hamartomatous, or juvenile (retention) polyps have not been regarded as premalignant. However, the strong predisposition to malignancy in familial forms of polyposis of these histological types calls this into question. Identification of the different tumour-suppressor genes (TSGs) mutated in these familial diseases means that it is now possible to look for somatic mutations of the same TSGs in the corresponding histological types of sporadic polyps occurring in the general population. By analogy with adenomatous polyps, this may highlight other types of isolated polyp having a genotype suggestive of malignant potential.

Polyps can be classified simplistically on the basis of their gross shape and appearance as either sessile (villous), pedunculated, or flat. The latter type is of particular current interest as the significance of flat adenomas comes under greater scrutiny. However, these macroscopic descriptions offer little insight into the underlying histology. Polyps can range in size from almost invisible at less than 1 mm, to larger lesions several centimetres in diameter. They may occur singly or literally thousands of polyps may carpet the bowel in the polyposis syndromes. The system of classification of polyps as 'non-neoplastic' or 'neoplastic' requires revision. Many conditions historically included in the former category clearly predispose to the development of carcinoma. With this proviso, the traditional histological classification into non-neoplastic and neoplastic polyps is retained herein (Table 1). The majority of colorectal polyps are adenomas (neoplastic) in which a carcinoma may develop. Thus the detection and removal of these lesions are crucial. This can also be the case with many of the so-called non-neoplastic types of polyp.

Non-neoplastic polyps

Hyperplastic (metaplastic) polyps

These metaplastic lesions are a common finding on proctosigmoidoscopy, usually presenting as 2- to 4-mm pale shiny nodules. Metaplastic polyps are increasingly recognized to occur in association with so-called 'serrated adenomas' and mixed hyperplastic–adenomatous polyps, these latter demonstrating epithelial dysplasia. Large hyperplastic polyps are frequently found on the right side of the colon and are associated with the development of carcinoma. Multiple hyperplastic polyps (hyperplastic polyposis) are strongly associated with a family history of colorectal cancer. Hyperplastic polyposis appears to involve an early loss of chromosome 1p in about 25 per cent of patients. It seems increasingly likely that a metaplastic polyp–mixed hyperplastic/adenomatous polyp–carcinoma sequence occurs, reminiscent of the more familiar adenoma to carcinoma sequence.

Hamartomatous polyps

Polyps of this general histological type have usually been considered together as non-premalignant lesions. However, the natural history of familial diseases in which hamartomatous polyps are a feature casts doubt on this conclusion.

Peutz–Jeghers syndrome (OMIM175200) is an autosomal dominant disease characterized by hamartomatous polyps throughout the gastrointestinal tract, particularly in the jejunum, with mucocutaneous pigmentation of the buccal mucosa, perioral region (Fig. 6), and digits. Peutz–Jeghers hamartomatous polyps have a characteristic histopathological appearance distinct from other gastrointestinal polyps. They contain frond-like epithelium with cystic dilatation of glands overlying a network of characteristic fibromuscular bundles. Hypermucinous goblet cells are often prominent. Clinically there is a risk of intussusception or infarction of Peutz–Jeghers polyps that are more than 1 cm in diameter.

Fig. 6 Peutz-Jeghers patient showing characteristic lip pigmentation.

The disease has been mapped to chromosome 19p13.3 with possible genetic heterogeneity and a second locus on 19q. A putative tumour-suppressor gene, *STK11,* which encodes a serine/threonine kinase, harbours germline mutations in patients. Consistent with its function as a tumour suppressor, the normal allele of *STK11* is lost in hamartomas and in associated adenocarcinomas, suggesting that the former may be precursors of the latter.

Recent meta-analysis of the risk of a wide variety of malignancies in Peutz–Jeghers syndrome has confirmed the very high relative risks of developing cancer of the oesophagus (57 times higher than the general population), stomach (relative risk (**RR**) 213), small bowel (RR 520), colon (RR 84), and pancreas (RR 132), with increased risks also for lung, breast, uterine, and ovarian cancer. The cumulative risk for all cancer is over 90 per cent by the age of 65 years. Thus patients with the Peutz–Jeghers syndrome are at a very high relative and absolute risk for both gastrointestinal and non-gastrointestinal cancers. Endoscopic follow-up should be at 2-yearly intervals, barium follow-through series is needed at similar intervals. Periodic laparotomies may be needed to remove large (2 cm+) polyps and avoid the danger of intussusception and gangrene

Juvenile intestinal polyposis

Juvenile intestinal polyposis (OMIM174900) is another autosomal dominantly inherited hamartomatous polyposis syndrome that is genetically heterogeneous, with loci mapped to chromosomes 18q21.1 and 10q23.3. A further variant is the atypical juvenile polyposis syndrome called hereditary mixed polyposis syndrome (OMIM601228). These patients also have an increased incidence of inflammatory and metaplastic polyps. This autosomal dominant disease has been mapped to chromosome 6q, although identification of the actual gene involved has not yet been achieved.

Juvenile polyps are pedunculated hamartomas, prone to surface ulceration, which can cause bleeding and stalk torsion that can autoamputate the polyp with bleeding. The main distinction from Peutz–Jeghers syndrome is a lack of muscle fibres. Juvenile polyps are characteristically overlaid with a layer of normal epithelium in contrast to the dysplastic changes in adenomas. Single juvenile polyps are usually located in the distal colon and have a macroscopic cherry-red appearance. Up to 1 per cent of children may have a juvenile polyp, and in 70 per cent only a single polyp is present. Juvenile intestinal polyposis, defined by the presence of more than three to five polyps, is therefore relatively rare.

About one-quarter of juvenile intestinal polyposis pedigrees segregate mutations in the *SMAD4/DPC4* gene on chromosome 18q21.1. This gene was originally identified as a tumour suppressor deleted in pancreatic cancers. The protein functions in the signal transduction pathway of the inhibitory growth factor **TGF-b** (transforming growth factor-beta). Absence of this protein would be expected to lead to a loss of response to the growth inhibitory effects of TGF-b. Other pedigrees show genetic linkage to chromosome 10q23.3. The *PTEN* tumour-suppressor gene maps in this region. It encodes a phosphatase, which works through the PI3-kinase pathway to modulate the cell-cycle control protein p27. *PTEN* is mutated in pedigrees with the allelic conditions Cowden disease (OMIM158350) or Bannayan–Zonana syndrome (OMIM153480), in both of which hamartomas in multiple tissues and macrocephaly occur. Juvenile polyps are features of both these autosomal dominant diseases, but the number of polyps is not as high as in juvenile intestinal polyposis. The occurrence of *PTEN* mutations in juvenile intestinal polyposis remains unconfirmed and another gene at 10q23 may be involved.

Juvenile polyps also occur in patients with the Gorlin syndrome (OMIM 123456), an autosomal dominant disease primarily characterized by basal-cell carcinomas of the skin and odontogenic keratocysts in the mouth. The genes mutated in this condition encode the receptor for the sonic hedgehog (**SHH**) protein. Again, however, no mutations have been detected in pedigrees segregating juvenile intestinal polyposis. Interestingly, SHH signalling indirectly controls the expression of the *SMAD4/DPC4* gene, which is causative in some juvenile intestinal polyposis pedigrees. SHH signalling is also indirectly involved in wingless (**WNT**) expression. Loss of signalling control through the WNT pathway leads to FAP.

Juvenile intestinal polyposis predisposes to adenomatous transformation and an increased risk of colorectal cancer and cancers elsewhere in the gastrointestinal tract. The accepted view of solitary juvenile polyps has been that they do not predispose to malignancy in children, who do not therefore require follow-up. It will now be possible to confirm this if it can be shown that *DPC4* mutations (or mutations in the putative gene at 10q23) are not found in solitary retention hamartomas. The occurrence of mutations would suggest these tumours have malignant potential.

Cronkhite-Canada syndrome is a rare, non-Mendelian condition characterized by juvenile-type polyps throughout the gastrointestinal tract (with the exception of the oesophagus). Ectodermal abnormalities include nail dystrophy, alopecia, and brown hyperpigmentation of the skin. Patients present with severe diarrhoea, malabsorption, and weight loss. This condition is rapidly progressive with inflammation of the entire intestinal mucosa. In some cases the condition resolves spontaneously with corticosteroids, but up to one-third of patients die within months. Survivors have an increased incidence of adenomatous change and carcinoma warranting regular colonoscopic surveillance. Surgical resection is necessary for otherwise uncontrollable bleeding. The pathogenesisof this disease remains unknown.

Inflammatory polyps

These pseudopolyps arise during the healing phase after attacks of severe colitis (ulcerative colitis, Crohn's disease, and schistosomiasis).have led to extensive mucosal ulceration. Histologically the colonic epithelial-cell layer is often normal. Larger inflammatory polyps may be composed of granulation tissue in the healing mucosa. These polyps may be friable and bleed, and can be multiple (Cap polyposis). Biopsy may be necessary to distinguish between inflammatory polyps and adenocarcinoma but those over 1 cm diameter are usually removed endoscopically for surety.

Other types of non-neoplastic polyp

Lipomas are usually located in the right colon. With the exception of intussusception, they are asymptomatic and do not require removal. Other rare benign tumours arising from the colonic submucosa include leiomyomas, neurofibromas, and polypoid haemangiomas. Pneumocystic disease (also called pneumatosis cystoides or pneumatosis coli) leads to multiple submucosal cysts, which can be punctured and collapsed in patients when they cause abdominal pain. The cysts may resolve when treated with oxygen therapy.

Neoplastic (adenomatous) polyps and the adenoma to carcinoma progression

Adenomas are dysplastic epithelial growths with malignant potential. Most frequently they are tubular (Fig. 7), having glandular organization, with a peduncle of normal tissue. Adenomatous polyps with a tubulovillous mixed character occur as well as the less common villous adenomas, which are sessile lesions, frond-like histologically, often quite large, and with the highest tendency to malignant change. Our understanding of the molecular events during first adenoma then carcinoma development has mainly been gained from genetic studies of familial adenomatous polyposis.

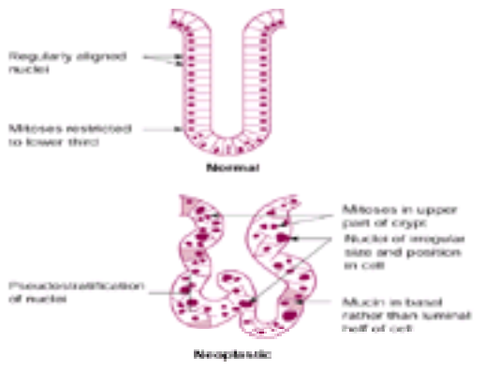

Fig. 7 Diagrammatic representation of the difference between the histology of a normal colonic crypt and the disorder of a neoplastic (dysplastic) crypt.

Aetiology, epidemiology, and pathology (Table 2)

'Sporadic' colonic adenomas are extremely common and are usually asymptomatic. Adenomas are found in around 20 per cent of individuals in their sixth decade and the incidence increases steadily with age. The distribution of adenomas in the colon reflects the site-specific incidence of colorectal carcinomas, with only minor differences. The majority of adenomas are found on the left side of the colon and in the rectum, although more right-sided lesions are found with increasing age. The role of environmental factors in adenoma development is suggested by their increased incidence in the developed countries. Furthermore, migrant communities such as the Japanese in the United States acquire an adenoma (and carcinoma) incidence comparable to the local population.

Around 60 per cent of adenomas can be accessed by flexible sigmoidoscopy. A distal adenoma is associated with the presence of a second more proximal adenoma in around 30 per cent of individuals. Like FAP adenomas, sporadic colorectal adenomas are thought to progress slowly over time. This means that only 5 per cent of 50-year-olds will eventually develop colorectal cancer, even though their incidence of adenomas is much higher than this. Because they are slow growing, only some 10 per cent of adenomas are greater than 1 cm in diameter when found. The incidence of malignant progression increases with tumour size.

Adenomas may arise in aberrant crypt foci, which are small areas of epithelium with irregular glandular organization, but without apparent dysplasia. These develop into single dysplastic crypts (unicryptal adenomas) so that the earliest recognizable adenomas have dysplastic epithelium. Dysplasia progresses through mild, moderate, to severe, which is reflected in the increasing malignant potential of the polyp. Adenomas may also be classified as early, intermediate, and late in histological appearance, and these histological changes have been correlated with the acquisition of a series of genetic alterations by Vogelstein and colleagues. Endoscopic excision is usually facilitated by the macroscopic difference in appearance between the normal mucosa and neoplastic epithelium, which appears hyperaemic and raised above the surrounding surface. In around 5 per cent of adenomas, carcinoma is detected (when neoplastic cells have breached the muscularis mucosae) (Fig. 8). Only at this stage is there access to the local lymphatics and blood vessels with the possibility of metastasis.

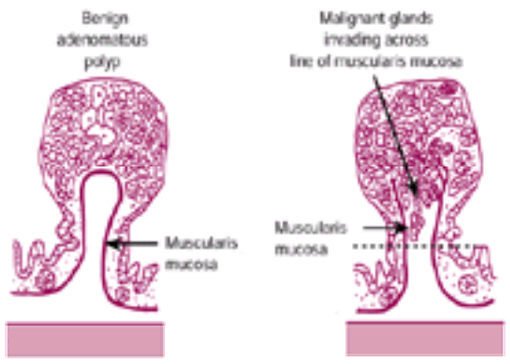

Fig. 8 Diagrammatic representation of the histology of a benign adenoma and a 'malignant polyp' showing adenoma and invasive adenocarcinoma (dotted line indicates transaction line to achieve complete removal).

There is increasing recognition of the existence of flat adenomas. These polyps are only very slightly raised and indeed may appear as depressed lesions. Flat adenomas are difficult to see endoscopically, but pioneering studies by Japanese gastroenterologists using non-absorbable mucosal dyes to provide contrast has led to a more widespread recognition of these lesions. Their clinical significance is that they can be more severely dysplastic than polypoid adenomas of comparable size, and therefore may represent a much-increased risk of carcinoma. The difficulties of studying the natural history of flat adenomas and their progression to so-called 'flat carcinomas' is complicated by the fact that they are more common on the right side of the colon. Some individuals with germline *APC* mutations appear to develop a flat adenoma phenotype.

The location of colonic carcinoma within about 5 per cent of adenomas suggests that the adenoma is a precursor of carcinoma. The time for progression to carcinoma is slow, usually 10 years or more. Patients with large or villous adenomas have an increased risk of coexisting synchronous carcinomas, or for the subsequent development of another carcinoma and so merit colonoscopy and follow-up. Later stage cancers contain less adenomatous material, presumably because this is overgrown by the malignant cells. Epidemiological data also support the adenoma to carcinoma sequence.

Additional mutations accumulate as the adenoma moves through intermediate to late stages (Fig. 9). These include activating mutations in the K-*ras* proto-oncogene and loss of tumour-suppressor function at chromosome 18q21. It appears unlikely that this is the so-called 'deleted in colorectal cancer' (*DCC* gene) and is more likely to involve the *DPC4/SMAD4* pancreatic TSG. It is the accumulation of these different genetic changes rather than their stepwise occurrence that drives carcinogenesis. K-*ras* mutations occur in 50 per cent of sporadic colorectal tumours. Loss of function of the *P53* tumour-suppressor gene occurs in 70 per cent of sporadic colorectal cancers. The resulting lack of apoptosis correlates with a metastatic phenotype, in that the incidence of *P53* mutations in carcinomas is much higher than in associated adenomas. Additional mutations occur in distant metastases.

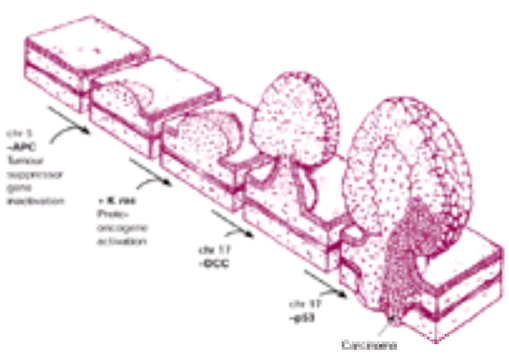

Fig. 9 The 'adenoma-carcinoma sequence' – a possible genetic model. Accumulated genetic changes to the stepwise progression from normal epithelium through increasing polyp size to eventual malignant change. Genetic alterations, which do not occur in a rigid sequence, include inactivation of tumour suppressor genes (adenomatous polyposis coli (APC), deleted in colorectal cancer (DCC) and p53, as well as the activation of proto-oncogenes such as K-ras. (Adapted from Fearon and Vogelstein (1990). *Cell* **61**, 759.)

Clinical features, management, and surveillance

Colonic adenomas do not usually cause symptoms. Abdominal pain, altered bowel habit, or intermittent bleeding are unusual. Large villous adenomas of the rectum may rarely present with mucoid diarrhoea and hypokalaemia. Polyps are usually discovered accidentally on barium enema or scanning examination or during endoscopic screening. Histological discrimination from other types of polyp is mandatory. The detection of a distal adenoma by proctoscopy or flexible sigmoidoscopy justifies colonoscopy to exclude the presence of proximal adenomas. The health economic implications of this are considerable given the high incidence of adenomas in the adult population.

Polyps are removed by endoscopic snare polypectomy or electrocoagulation. This is painless except for polyps within 2-5 cm of the anal margin. Around 90 per cent of patients with polyps have only two lesions, the majority of which are under 1 cm, stalked, and easy to snare. The tendency to bleed after polypectomy is greater with larger tumours. Very large, sessile polyps may need to be removed in multiple portions, but few are so large as to need surgical resection. Endoscopic removal of an adenomatous polyp that subsequently proves to contain a focus of carcinoma is a satisfactory outcome not requiring subsequent operation if there are clear histological margins.

Post-polypectomy surveillance is necessary in patients who have had an adenoma detected and removed. They are at higher risk of second adenomas, particularly when multiple (five or more) polyps are present at first examination. Villous adenomas are particularly liable to recur locally. Repeat colonoscopy is performed at 3 to 5 yearly intervals, depending on number and size of polyps found. The decision to cease surveillance at 75 years of age is justified after a negative examination because of the slow natural history of the disease, despite the increased incidence of both polyps and carcinoma with age.

Familial adenomatous polyposis (FAP)

The principal feature of FAP (OMIM175100) is the presence of hundreds to thousands of adenomatous polyps throughout the colon (Fig. 10). The condition is inherited as an autosomal dominant, although about one-quarter of cases arise without a family history, thus implying a high rate of new mutations. As well as the colonic polyps, adenomas are found in the duodenum in the majority of patients, and also in the stomach and elsewhere in the small intestine. Adenomas usually begin to develop during the second decade so that screening is performed between 12 and 14 years of age. The risk of colorectal cancer developing in one of the adenomas is essentially 100 per cent by the age of 40 years, so that this condition is fully penetrant. Screening by flexible sigmoidoscopy with biopsy of polyps for histological diagnosis confirms the condition and allows surgery before the age of 20. Screening family members by flexible sigmoidoscopy confirms or eliminates the diagnosis in relatives. The creation of Polyposis registers makes FAP an example of practical colorectal cancer prevention.

Fig. 10 Familial adenomatous polyposis (FAP) colectomy specimen, showing carpeting of the colon by hundreds of adenomas.

There is a 10- to 20-year period between the appearance of an adenoma and its progression to colorectal carcinoma. Clinically, therefore, colectomy with ileorectal or ileoanal anastomosis is usually performed on FAP subjects in their late teens, as there is low risk of carcinoma before this age. Careful long-term follow up is required. There is a 10 per cent risk of carcinoma in any rectal stump over a 30-year period, although rectal polyps may undergo regression, suggesting that adenoma progression to carcinoma is not inevitable. Total proctocolectomy with ileostomy is also an option with large numbers of rectal polyps if the increased operative risks of ileoanal anastomosis with formation of a pelvic pouch are not regarded as acceptable.

Historically, FAP caused 1 per cent of all colorectal cancers. Surveillance programmes have reduced this incidence, but the prolonged survival of patients postcolectomy has highlighted the other features of FAP, which increasingly create clinical problems. Carcinomatous change occurs in adenomas in the duodenum, particularly around the ampulla of Vater and in the bile duct. It may also occur in the stomach. Follow-up is required to avoid these eventualities. Mesenchymal desmoid tumours, particularly those occurring in the retroperitoneum, can grow to an enormous size. Their position and tendency to re-growth makes them difficult to remove and they may cause devastating consequences due to pressure effects.

The wide range of additional features of FAP led to the description of Gardner's and Turcot's syndromes that have proved to be allelic with FAP. Additional features include osteomas of the mandible and skull, dental abnormalities, sebaceous cysts, and brain tumours. About one-half of patients with Turcot's syndrome have germline *APC* (adenomatous polyposis coli) mutations and develop cerebellar medulloblastomas. It is notable that the other 50 per cent have germline mutations in their DNA mismatch repair genes (that is, *HNPCC*). These individuals undergo somatic *APC* mutations but develop glioblastoma multiforme. This link between HNPCC and mutations in *APC* is intriguing.

FAP is caused by mutations in the *APC* gene on chromosome 5q21. This large 2843 amino-acid protein performs multiple functions in the cell, but in the context of colonic neoplasm development its key role is in the WNT pathway. In the absence of WNT signalling, APC forms a complex with two other proteins, axin and glycogen synthase kinase-3b (GSK-3b). This complex phosphorylates the b-catenin protein, which can then be tagged with ubiquitin and degraded by the proteasome system. In this manner, levels of cellular b-catenin are controlled. b-Catenin plays roles in intracellular signalling from E-cadherin in the epithelial adherens junction (which is also mutated in gastric cancer), but when present in the cell at high levels it can bind a second protein (TCF-4), enter the nucleus, and switch on the expression of a number of genes, including the c-*myc* oncogene and the cyclin D1 gene, and others. In the presence of a WNT signal, GSK-3b is inactivated and b-catenin levels rise to drive gene expression and cellular proliferation. With only mutant APC, the cell is also incapable of degrading b-catenin and again proliferation occurs.

APC expression increases as the colonocyte migrates up the crypt to the lumenal surface. This may reflect a role in controlling cell movement. In FAP, the vast majority of germline mutations are in the *N*-terminal half of the protein and are nonsense mutations leading to truncated proteins. The key region of APC for b-catenin binding lies between amino acids 1000 and 2000. Germline mutations occur almost exclusively before amino acid 1650. Some 85 per cent of sporadic colorectal adenomas and carcinomas also have mutations in the *APC* tumour-suppressor gene. The majority of these occur in the so-called 'mutation cluster region' between amino acids 1280 and 1500. Again, this creates truncated proteins that cannot properly degrade b-catenin. Consistent with Knudsen's hypothesis, the second normal *APC* allele is consistently lost in adenomas, so that the cell has no functional 'gate-keeper' tumour-suppressor protein.

Identical germline *APC* mutations can cause a variety of phenotypes, from polyposis alone to the full spectrum of Gardner's syndrome. This suggests that there are genetic modifier loci for FAP. Mutations at the extreme 5'-end of the gene may be associated with a milder 'attenuated' phenotype. Mutations towards the 3'-end of the gene (around amino acid 2600) cause the familial desmoid syndrome. Congenital hypertrophy of the retinal pigment epithelium (**CHRPE**) is another recognized clinical sign in FAP. Only germline mutations downstream of codon 422 cause CHRPE, for reasons that remain unclear. This means, however, that opthalmoscopic detection of CHRPE cannot be regarded as a reliable screening method.

Determination of the mutant sequence in FAP pedigrees now allows precise genetic counselling. However, the large size of this gene and protein make sequencing it difficult in individual patients. The protein truncation test (**PTT**) is used to detect *APC* mutations because so many generate truncated proteins. These are detected by

electrophoresis after *in vitro* transcription and translation from patients' *APC* mRNA.

The critical role of APC in colorectal adenomas (and carcinomas) has been further emphasized by analysis of sporadic tumours that do not harbour *APC* mutations. These 15 per cent of colorectal neoplasms prove to have mutations in the b-catenin gene itself. Moreover, these mutations alter those amino acids which undergo phosphorylation by the APC–GSK-3b–axin complex. Thus, these mutant forms of b-catenin are constitutively active. This same mechanism seems to apply in melanoma. Mutations in the axin gene can also lead to colorectal neoplasia.

NSAIDs were serendipitously found to reduce the number and size of colorectal adenomas in FAP. The effect has been demonstrated with sulindac in particular. Aspirin also reduces the risk of colorectal cancer. Colorectal cancers overexpress cyclo-oxygenase-2 (COX-2). In *Apc* knockout mice, deletion of the murine *Cox-2* gene causes a significant reduction in the number of polyps formed. Aspirin or selective COX-2 inhibitors may delay the onset of FAP adenomas and hence their progression to carcinoma. The attraction of this possibility is the advantage of postponing elective colectomy into the third decade.

Polymorphic *APC* variants may exist in the population, which, although associated with an increased risk of adenoma development, are of such low penetrance that they do not seem Mendelian in character. This would be consistent with the marked familial clustering seen in colorectal cancer.

Colorectal cancer (Fig. 11)

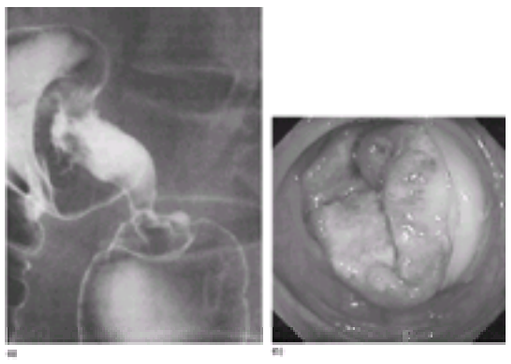

Fig. 11 Colonic carcinoma appearances. (a) Air-contrast barium study showing 'apple-core' stricture. (b) Colonoscopy appearance of encircling 'annular' tumour.

The natural history of colorectal tumorigenesis now provides a rational basis for the introduction of screening programmes in an attempt to control this common malignancy, which is curable if detected early or at the precancerous (polyp) stage. Several screening modalities have been shown to decrease colorectal cancer (**CRC**) mortality, and screening guidelines have been published in several countries. The incidence of the disease is falling in the United Kingdom and United States as a result of increased detection and removal of adenomas.

Epidemiology and aetiology

Colorectal cancer is the most common malignancy in the United Kingdom after lung cancer: 19 000 deaths each year reflect a 5-year survival rate for CRC at all stages of approximately 50 per cent. The lifetime risk of colorectal cancer in the United States approaches 6 per cent. There is a higher incidence of rectal cancer in men. The estimate for the United States in 2000 was 130 200 new cases of colorectal cancer and 56 300 deaths, a lower population death rate than in the United Kingdom. The incidence of cancers of the colon and rectum increases with age, the mean age at diagnosis being 65 years. The 5-year survival rate is dependent upon the tumour stage at diagnosis: Dukes' stage A disease, over 90 per cent; Dukes' B disease, 65 per cent; stage C1 disease, 30 per cent; stage C2 disease, 20 per cent; and stage D disease, a 5-year survival of 2 per cent or less.

A small proportion of patients with Dukes' stage A colorectal cancer succumb to their disease. Although only 15 per cent of cases are diagnosed at this stage, it would be invaluable to identify the small proportion of them that might benefit from adjuvant therapies. This issue becomes more important in patients with Dukes' stage B carcinomas, which represent some 40 per cent of tumours diagnosed. It would be valuable to be able to discriminate those 35 per cent of patients at diagnosis who will go on to develop progressive disease and target adjuvant chemo/radiotherapy on them. Conversely, it would be desirable to avoid exposing patients who have had what proves to be curative surgery to the toxicity of adjuvant chemotherapy.

There are clearly environmental factors that influence the incidence of colorectal cancer. These are presumably mainly dietary. The incidence of colorectal cancer is higher in Northern Europe and North America than in the developing world. The risk of adenoma formation and colon cancer in Japanese peoples rises when they are exposed to a westernized diet. Although carcinogens or co-carcinogens may be produced by bacterial metabolism of bile salts, or from animal fat, or as a result of prolonged transit times of low-fibre diets, the evidence for this is not clear-cut.

It is increasingly recognized that genetic factors other than those causing Mendelian cancer syndromes are important in colorectal carcinoma with, for example, a threefold increased frequency of a family history of bowel cancer in individuals diagnosed as having an adenomatous polyp compared with normal controls. Thus the underlying aetiology of colorectal carcinogenesis is complex. Implantation of the ureters into the sigmoid colon (ureterosigmoidostomy) results in a high incidence of adenomas with subsequent carcinoma development 10 years and more after creation of the ureteric stoma. Long-term inflammation, as in extensive ulcerative colitis of over 20 years' duration (and probably also Crohn's colitis), provides a risk of ulcerative colitis-associated colorectal carcinoma (**UCACRC**). This may affect 5 per cent of these individuals, even where their disease is reasonably well controlled or healed. It may be multifocal. These cancers often have a mucinous appearance, with or without 'signet ring' cells. Ileorectal anastomosis to treat refractory inflammatory bowel disease (**IBD**) also leaves a risk of local malignant change in the rectum. Some protection is afforded by annual or biennial colonoscopy, with serial biopsies to detect any warning epithelial dysplasia.

Hereditary non-polyposis colon cancer

A second, autosomal dominant disease accounts for up to 5 per cent of cases of colorectal cancer. Hereditary non-polyposis colorectal cancer (HNPCC) was originally divided into Lynch syndromes type 1 and 2, with a range of other malignancies associated with the latter. A total of seven different mutant genes have now been characterized to define HNPCC types 1 to 7, superseding the Lynch definition. HNPCC kindreds were initially defined by the so-called 'Amsterdam' criteria, which are:

1. at least three family members in two or more successive generations must have colorectal cancer;
2. one of these must be a first-degree relative of the other two;
3. cancer must be diagnosed before the age of 50 in at least one family member;
4. FAP must be excluded.

HNPCC is associated with other malignancies in mutation carriers: endometrial cancer is the most common, ovarian cancer is frequent, and tumours in other organ systems are less common (breast, stomach, possibly prostate). The risk of uterine cancer may exceed that for colorectal cancer in females in pedigrees, emphasizing the necessity for uterine screening. HNPCC has an approximate penetrance of 74 per cent for colorectal cancer in men by the age of 70 compared with only 30 per cent in females. The risk of uterine cancer is 42 per cent in these women. Hysterectomy is usually accompanied by bilateral oophorectomy because of the risk of subsequent ovarian cancer.

The syndrome arises as a result of the germline transmission of mutations in one of six genes encoding components of the DNA mismatch repair enzyme system. Germline mutations in the TGF-b receptor II gene are now also classified as HNPCC. The encoded proteins repair DNA mismatches arising as a result of replication errors. These tumours were therefore described as RER+. Replication of DNA is particularly prone to error in those regions of DNA known as microsatellites where multiple short nucleotide repeat sequences occur. Loss of DNA repair function leads to the so-called 'microsatellite instability' of these regions. This means that the allele sizes of microsatellites are different in tumours than in normal cells from the same individual. PCR-based methods are readily able to detect this to establish that

a given tumour has an RER+ phenotype. The extent of microsatellite instability varies between tumours and there is some disagreement as to how many loci need to be modified before a tumour can be classified as RER+.

In addition to patients with HNPCC, between 10 and 15 per cent of sporadic colorectal cancers are also RER+ to some extent. The inability to repair errors introduced during DNA synthesis renders the genome in HNPCC cells liable to mutation. Interestingly, the *APC* gene itself appears prone to mutation in mismatch repair deficiency, leading to a frame-shift mutation and premature termination of translation. Thus it is possible that colorectal adenomas arise in patients with HNPCC through a similar final common pathway to that in FAP.

The commonest mutations occur in the mismatch repair genes *MSH2* and *MLH1*. There is some evidence that HNPCC polyps and tumours occur predominantly on the right side of the colon. It has also been suggested that these have a better prognosis, although this may be confounded because right-sided lesions tend to be larger (given the reduced solidity of the stool in the ascending colon) at presentation and therefore likely to be at more advanced stages. It is clear that a variety of mutations can predispose to colorectal cancer, but intriguingly all of these seem to feed into a final pathway of tumorigenesis similar to that originally elucidated for FAP. Patients with HNPCC should be included in genetic counselling and surveillance programmes usually by colonoscopy performed at 2 to 3 yearly intervals.

Colorectal cancer screening and surveillance

A number of screening modalities have been evaluated. These include testing for faecal occult blood, flexible sigmoidoscopy, double-contrast barium enema examination, and colonoscopy. The health economic implications of implementing such screening programmes are considerable. However, there is clear benefit in screening and surveillance of high-risk individuals such as those with previously detected adenomas or with a strong family history. In these cases colonoscopic examination is justifiable, as in the follow-up of individuals who have had carcinomas resected.

In the average-risk population, there is increasing consensus that screening should begin at 50 years of age. A number of protocols are undergoing trial to establish their diagnostic yield in terms of cost per year of life saved. Despite its lack of sensitivity and specificity, annual faecal occult-blood testing (**FOBT**) can be considered. Others advocate flexible sigmoidoscopy every 5 years, with colonoscopy if any adenomas are detected. Whether repeat flexible sigmoidoscopy every 5 years in individuals in whom no polyps are found is worthwhile needs to be established, as its avoidance would generate considerable cost savings. Specific guidelines have been published in the United States and a Cochrane Review of FOBT screening has been published. Cochrane Reviews are in preparation on the benefits of colorectal adenoma surveillance programmes and on detecting colorectal cancer in patients with IBD. Population screening using flexible sigmoidoscopy is the subject of an ongoing, national multicentre study in the United Kingdom. There is the real possibility of achieving significant reductions in the incidence of this common, but avoidable, malignancy by screening. Colonoscopic surveillance of those with adenomas or significant genetic risk is at 2 to 5 year intervals thereafter.

Pathology

Most colorectal cancers occur distally, in the rectum and sigmoid colon. An approximate estimate is that one-third of cancers occur in the rectum, one-third in the sigmoid colon, and one-third more proximally. Carcinomas are typically polypoid masses with central ulceration, eroded edges that bleed easily, and which may result in strictures. Infiltrating scirrhous carcinomas are rare. Colorectal carcinoma initially spreads by local invasion, which accounts for its good prognosis when confined to the bowel wall. Invasion through the mucosal and muscle layers of the bowel may lead to involvement of adjacent organs including the bladder and prostate. Secondary complications may therefore include fistula formation. Venous and lymphatic spread results in the predominance of liver metastases. It is increasingly recognized that even multiple liver metastases may be amenable to surgical resection. Lymphatic spread is to local nodes (Dukes' stage C1 disease) then to more proximal nodes in the mesentery (Dukes' stage C2 disease).

The Dukes' classification of colorectal carcinomas (Fig. 12) remains the mainstay of tumour staging. Stage A disease is confined to the bowel wall with no extension to the serosal fat. Stage B disease involves the full thickness of the bowel wall with extension through to the serosa. Stages C1 and C2 involve the spread of tumour to draining lymph nodes. Stage D disease involves distant spread, primarily to the liver. These different stages very clearly correlate with 5-year survival. The alternative TNM staging system has the advantage over the Dukes' system in distinguishing early tumours involving only the submucosa (T1) and amenable to endoscopic resection, from slightly more advanced tumours (T2) which extend into the muscularis propria. T3 tumours extend into the serosal fat and T4 tumours are those which have perforated or involve the peritoneal surface or other viscera. The extent of histological differentiation is also an important prognostic indicator with poorly differentiated tumours particularly aggressive. As well as liver metastases, secondary deposits appear in the lung (50 per cent), peritoneum (15 per cent), and bones (15 per cent). Unfortunately, approximately 50 per cent of colorectal cancer is stage C or D at diagnosis because it remains asymptomatic, so that the overall 5-year survival rate is reduced. Skewing the stage of detection towards Dukes' A lesions would have a major impact on overall survival.

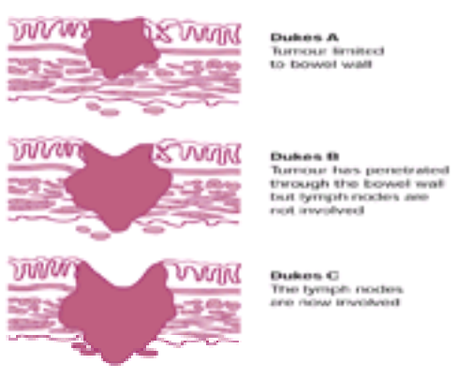

Fig. 12 Dukes' classification of colorectal carcinoma.

Clinical features

The site of a colorectal cancer influences the resultant symptoms. Most colorectal neoplasms grow into the lumen of the bowel where they may cause obstruction or bleed because of their friable epithelial surfaces. Obstruction and bleeding are increasingly likely with more distal lesions as the stool becomes more solid. Blood is typically dark and mixed in with the stool, in contrast to the fresh or dripping bleeding observed from haemorrhoids. Alteration of bowel habit may involve frequency or increased constipation where solid stool flow is obstructed, particularly with left-sided lesions. Ascending colon tumours can grow large without affecting faecal flow. In these cases presentation is often with iron-deficiency anaemia as a consequence of long-standing but invisible blood loss. Diagnosis of right-sided lesions is further complicated by the difficulty of achieving complete evacuation, the challenge of barium enema or scanning studies in this region, and the difficulty of colonoscopic access. Pain in a patient with colorectal cancer suggests obstructive disease or invasion, and weight loss is a late symptom suggestive of advanced disease with poor prognosis. A palpable mass may reflect faecal retention, but the tumour itself may still be resectable. Thus the clinical picture in colorectal cancer is relatively non-specific and must be differentiated from the anorectal bleeding of haemorrhoids or fissures, and the pain and altered bowel habit of irritable bowel disease.

The occurrence of so many tumours in the rectum makes digital rectal examination essential. Flexible sigmoidoscopy has become a routine procedure requiring a simple enema preparation and usually no sedation. It will detect 60 per cent of adenomas and carcinomas. Colonoscopic examination has the advantage that other synchronous polyps or tumours are detected. Other imaging modalities are becoming increasingly important, for example CT scanning as 'virtual colography' or to stage disease before operation in order to detect involvement of adjacent structures such as the ureters. MRI scanning is particularly useful to assess rectosigmoid disease. Transmural ultrasound is used in the rectum for assessing local invasion. Serological tests such as carcinoembryonic antigen levels are of some value in postoperative management and in the detection of tumour recurrence, but are secondary to accurate histological staging.

As a result of these investigations the differential diagnoses can usually be discounted. These include benign lesions such as areas of ischaemia, solitary rectal ulcer syndrome, masses of granulation tissue, endometriosis, or amoebomas, all of which may mimic carcinoma on radiography or endoscopy. An appendix abscess may mimic a caecal carcinoma, as may tuberculosis, actinomycosis, or Crohn's disease. Sigmoid diverticular disease with pericolic abscess formation may mimic distal carcinoma. Other malignancies of the colon, including lymphoma, sarcoma, and carcinoid tumours are rare. Kaposi's sarcoma presents as violaceous nodules in patients with AIDS. A variety of anal tumours occur that reflect the different structures and embryological derivation of this area. They include adenocarcinomas,

condylomas, and squamous carcinomas (which are common in immunosuppressed patients due to papillomavirus infections), and malignant melanomas. Rarely, the anal margin may be the site of premalignant skin conditions such as leucoplakia, Paget's disease, or Bowen's disease.

Treatment

Surgical resection is the primary treatment and is potentially curative. The aim is to remove the neoplasm with an adequate uninvolved resection margin of normal tissue, as well as the entire draining lymphatic field. The need for careful mesorectal dissection to prevent local recurrence is established in rectal cancer surgery. In the presence of metastases, palliative resection of the primary tumour is usually desirable in order to prevent obstruction, so that 90 per cent of patients undergo surgery. The operative mortality for colorectal surgery is less than 1 per cent. Bowel preparation is essential, with oral purgatives and/or enemas, to avoid anastomotic complications. Prophylactic antibiotics minimize postoperative sepsis. Thorough preoperative preparation may be impossible in patients presenting in emergency with perforation or obstruction. Some 25 per cent of patients with colorectal cancer present in this manner, when the hazards of surgery are much increased.

The site of the tumour dictates the operation performed: caecal carcinomas are treated by right hemicolectomy; ascending and transverse colon tumours by extended right hemicolectomy; left hemicolectomy for descending colon cancer; and sigmoid colectomy for sigmoid tumours. All these operations involve resection back to the respective tributaries of the superior and inferior mesenteric arteries.

The operation performed for rectal cancer is also dependent on the tumour's position. An anterior resection is performed where it is possible to leave a disease-free margin above the anal canal. When the cancer is too low for this, an abdominoperineal resection is performed with resection of the distal sigmoid colon, rectum, and anus, and creation of a permanent sigmoid colostomy. Rectal surgery has been transformed by the availability of a variety of stapling instruments to reanastomose the bowel. Low anterior resections are now commonly performed with good results, so that anal excision is avoided. Abdominoperineal resection is unavoidable when the tumour involves the anal sphincter. A temporary colostomy is sometimes advisable, particularly acutely, when there is a high risk of postoperative anastomotic leakage.

There is increasing interest in the use of laparoscopic techniques to remove colorectal tumours. The great advantage of this approach is that surgery is less traumatic and postoperative recovery is accelerated. A variety of techniques have been developed to avoid the problem of disease recurrence at the sites of laparoscopic entry ports. Multicentre, randomized controlled trials are currently comparing outcomes, especially 5-year survival rates after laparoscopic versus open surgery for colorectal cancer. Given the different 5-year survival rates for tumours at different Dukes' stages, the outcomes of colorectal surgery essentially reflect the proportions of patients presenting at these different stages (Table 3).

Dukes' staging is also central to the appropriate use of chemotherapy and radiotherapy. 5-fluorouracil (with folinic acid) prolongs the survival of patients with stage D disease. Randomized trials have also shown that 5-fluorouracil (**5-FU**) improves survival in stage C disease when used as adjuvant therapy to surgery. It is not clear that portal vein infusion of 5-FU provides any additional advantage. The evidence of survival advantage from adjuvant chemotherapy in stage B disease is less clear-cut. The complicating factor is that 60 per cent of these patients will survive for 5 years after surgery alone, and any additional survival advantage from chemotherapy must be balanced against the risk of life-threatening toxicity (stomatitis, leucopenia, etc.) in up to 5 per cent of patients. Intuitively, it has been assumed that agents beneficial in stage C disease should provide a therapeutic benefit in the 40 per cent of stage B patients (and 10 per cent of stage A patients) who do not survive for 5 years after surgery alone. Some newer drugs show promise in the treatment of colorectal cancer: these include other thymidylate synthase inhibitors and drugs that inhibit topoisomerase (the 'tecan' class). However, this begs the question of how patients with tumours with atypically poor prognosis can be identified.

Randomized clinical trials of preoperative radiotherapy versus surgery alone show a reduction in local recurrence rates and survival advantages, particularly in rectal cancer. However, this must be balanced against the disadvantages of radiotherapy: impaired bowel function in up to one-third of patients and minor increases in early postoperative morbidity. Based on the available evidence, preoperative radiotherapy is better than postoperative treatment because the incidence of surgical complications is lower. Thus, adjuvant chemotherapy combined with preoperative radiotherapy is established practice for Dukes' stage C colorectal cancer. Recurrent disease is incurable so complete tumour excision is vital, particularly total mesorectal excision of rectal carcinomas. Therefore, follow-up after potentially curative surgery is primarily aimed at detecting further adenomas or cancers in these patients, who are at increased risk. Colonoscopy at 5 years seems to be as effective as more intensive surveillance programmes.

Further reading

Cochrane Reviews have appeared on the clinical management of gastrointestinal tumours (available at http://www.update-software.com/). Their topics include:

preoperative chemotherapy for resectable thoracic oesophageal cancer

preoperative radiotherapy for oesophageal carcinoma

palliative chemotherapy for advanced or metastatic colorectal cancer

screening for colorectal cancer using the faecal occult-blood test

Protocols have also been published for Cochrane Reviews on:

colorectal adenoma surveillance and colorectal cancer

surgery for obstructing left-sided colorectal cancer: primary or staged resection?

follow-up strategies for non-metastatic colorectal cancer

intravenous 5-FU-based regimes as adjuvant to surgery for colon cancer with local lymph node spread

local surgical treatment in early rectal cancer with or without adjuvant therapy

effect of preoperative radiotherapy and surgery for localized rectal carcinoma

strategies for detecting colon cancer or dysplasia in inflammatory bowel disease.

In addition, relevant material appears in publications from the NHS Centre for Reviews and Dissemination (e.g. Management of upper gastrointestinal cancers, *Reviews and Dissemination*, Vol. 6, 2000). Further authoritative reviews of United Kingdom best practice are published in *Clinical Evidence* (see Issue 4, December 2000). These include meta-analyses of randomized controlled trial evidence for or against:

treatment of gastro-oesophageal reflux disease in preventing progression of Barrett's oesophagus

adjuvant chemotherapy for stomach cancer

radical versus conservative surgical resection for stomach cancer

total mesorectal excision for rectal cancer

particular follow-up strategies after colorectal cancer

preoperative radiotherapy in colorectal cancer

adjuvant chemotherapy in colorectal cancer.

The consensus emerging from these progressively more comprehensive, evidence-based surveys of the optimum approaches to the clinical management of gastrointestinal tumours has been incorporated therein. The reader is encouraged to consult these accessible sources directly.

14.16 Vascular and collagen disorders

Graham Neale

Clinical disorders of the gastrointestinal tract caused by vascular and collagen disorders are rare in general medical practice. Vascular insufficiency (Table 1) may cause dramatic disease with infarction, perforation, bleeding, ulceration, or strictures, and in the very sick it may damage mucosal function and lead to ischaemia-reperfusion injury. In contrast, patients with collagen disorders often have abdominal symptoms due to disorder-related vasculitis, but such symptoms rarely dominate the clinical picture.

Causes of segmental mesenteric ischaemia

The blood supply to the gut may be impaired by compression, by intraluminal occlusion of vessels, or by intrinsic vascular pathology, including vasospasm.

Compression of mesenteric vessels

The mesenteric vessels may be compressed by torsion or strangulation of the mesentery, retroperitoneal haematomas, neoplastic infiltration, and rarely by proliferating fibrous tissue such as occurs in retroperitoneal fibrosis or occasionally around carcinoid or desmoid tumours. Venous occlusion with thrombosis is a rare complication of excessive gaseous pressure during laparoscopy.

Intraluminal occlusion of mesenteric vessels

Thrombosis

Thrombosis may occur in arteries or veins. In arteries thrombosis usually occurs on an ulcerated atheromatous plaque but it may occur spontaneously in polycythaemia, sickle cell disease, cryoglobulinaemia, and amyloidosis. Thromboangiitis obliterans (Buerger's disease) is a rare condition which does not usually affect mesenteric vessels but intestinal infarction has been described as a first manifestation of the disorder in women who smoke. Mesenteric arterial thrombosis has also been described in cocaine addicts.

Mesenteric venous thrombosis is less common than arterial occlusion. An inherited thombophilia (such as deficiency of protein C, protein S, or antithrombin III) is the most likely cause. It may also occur in the primary antiphospholipid antibody syndrome and with dysfibrinogenaemia and has been described in women taking oral contraceptives (especially those who smoke) after splenectomy and in association with pancreatitis.

Arterial embolism

Emboli cause up to one-third of cases of mesenteric vascular occlusion. They arise from the heart, especially in patients with mitral stenosis and atrial fibrillation, with myocardial infarction and endomyocardial thrombosis, and with bacterial endocarditis. Paradoxical embolism through a patent foramen ovale and embolism from aortic mural thrombi are uncommon causes.

Diffuse microthrombosis

Diffuse thrombosis of small vessels may occur as a result of disseminated intravascular coagulation. The haemolytic–uraemic syndrome and thrombotic thrombocytopenic purpura are related conditions in which platelet aggregation occurs in small vessels (see Chapter 20.10.6 and Section 22). Renal failure usually dominates the clinical picture, but in most patients there is evidence of widespread involvement of the intestine and related organs. The haemostatic abnormalities and their relationship to microvascular injury are complex and give rise to diverse and often confusing clinical and laboratory findings. Infusions of platelets and fresh plasma are usually beneficial but therapy must be individualized depending on the nature of the underlying pathology and the site and severity of haemorrhage or thrombosis.

Intrinsic vascular pathology

Atheromatous occlusion

Atheroma is the most common cause of mesenteric vascular insufficiency, which often remains undiagnosed in life, probably because the slowness of the pathological process allows for the development of a compensatory collateral circulation. Intestinal infarction secondary to atheromatous occlusion of one major mesenteric vessel is uncommon. Indeed, all three vessels may be occluded without visceral damage.

Inflammation of blood vessels

Vasculitis is primarily a disorder of small vessels and affects many organs (Table 2). Involvement of the mesenteric circulation rarely causes infarction of long segments of the gut but may cause a wide spectrum of gastrointestinal disorders that are described later in this chapter.

Damage to the arterial wall in the splanchnic circulation sometimes occurs a few days after surgical correction of coarctation of the aorta. There is necrotizing arteritis with fibrinoid necrosis which is most marked at arterial bifurcations and appears to be related to the sudden sustained increase in blood pressure. It usually resolves spontaneously but may occasionally lead to intestinal infarction requiring operative intervention.

The walls of the mesenteric arteries may be involved in fibroelastic hyperplasia and in Takayasu's disease. In malignant hypertension, intimal hyperplasia and fibrinoid necrosis of arteriolar walls may lead to patchy ischaemia of the intestines. Irradiation of the abdomen leads to vascular necrosis and thrombosis, which in turn may cause ischaemic ulceration that on healing leaves a fibrosed intima and a poorly perfused segment of intestine.

Recently mesenteric veno-occlusive disease has been described as a discrete entity that may affect adults of any age. Previous cases were regarded as idiopathic

venous thrombosis.

Non-occlusive mesenteric ischaemia (Table 3)

Sporadic and epidemic cases of necrotizing enteritis but without evidence of vascular occlusion have been described from many parts of the world and in all age groups. Neonatal necrotizing enterocolitis occurs in the first week of life of premature and low-birth-weight infants. Artificial hyperosmolar feeds may promote damage, whereas breast milk appears to be protective, possibly by providing passive enteric immunity. Lesions occur most frequently in the stomach, the distal ileum, and the colon; as mucosal integrity disintegrates bacterial invasion enhances the damage. The condition may also occur in infants of normal birth weight who have a hyperviscosity syndrome or who have been exposed to cocaine *in utero*.

In adults, infarction of the gastrointestinal tract without vascular occlusion is seen mainly in the elderly with severe low-output cardiac failure but it has been described in women in middle-life without obvious risk factors. In tropical and subtropical areas the condition is more common and may be related to environmental factors including diet, infection, and infestation.

Non-occlusive focal ischaemia appears to underlie the pathogenesis of uraemic colitis, radiation enteritis, potassium-induced ulcers, and multiple stress ulcers of the upper gastrointestinal tract. The verotoxin of *Escherichia coli* 0157:H7, which causes haemorrhagic colitis, is another aetiological agent. This is a particularly potent cause of disseminated intravascular coagulation, which may occur as a result of the absorption of bacterial endotoxin and thromboplastins from damaged tissues. It is associated with the development of the haemolytic–uraemic syndrome in children and more rarely with thrombotic thrombocytopenic purpura in adults (Table 2).

Intestinal ischaemia: the clinical syndromes

The clinical effects of mesenteric vascular insufficiency will be considered under four headings: acute mesenteric ischaemia, chronic mesenteric ischaemia, ischaemic colitis, and ischaemia-reperfusion injury. Focal ischaemia of the intestine will be considered in the section on collagen disorders.

Acute mesenteric ischaemia

Clinical features

Necrosis, incipient or complete, of that part of the gut supplied by the superior mesenteric artery is life-threatening. The onset is usually abrupt. Abdominal pain is the key symptom and at the onset it is usually colicky in nature. As the condition progresses the pain becomes constant and unremitting. Initially it is felt in the right iliac fossa and then spreads over the entire abdomen. Diarrhoea is usual and the motions may contain blood. Vomiting occurs in some cases but haematemesis is rare. There may be slight tenderness in the right iliac fossa and some exaggeration of bowel sounds. Over the course of hours (or at the most a day or two) the abdomen becomes distended and silent with increasing tenderness and a positive rebound sign. At this stage there are usually signs of peripheral circulatory failure. The patient is pale, anxious, sweating, and tachypnoeic. Later the blood pressure falls and the patient becomes cyanosed and anuric; by now damage to the intestine is irrecoverable.

Diagnosis is often delayed because there are no clinical signs apart from overwhelming patient distress which itself indicates the need for urgent specialist attention. Duplex Doppler ultrasonography and contrast-enhanced magnetic resonance imaging are probably the best means of confirmation but are rarely available. For practical purposes, the diagnosis still depends on the efficiency with which other causes of an apparent abdominal catastrophe can be excluded. Plain radiographs of the abdomen may show non-specific dilatation of loops of intestine with multiple fluid levels. The presence of gas bubbles in the portal vein is diagnostic of intestinal necrosis at a stage when the patient is beyond recovery. Needle aspiration of the peritoneal cavity may be a helpful procedure because intestinal infarction usually produces blood-stained fluid.

Management

Early laparotomy is essential. First the clinician has to combat the effects of loss of water, electrolytes, and protein leading to hypovolaemia and impaired tissue perfusion, bacterial invasion, and disseminated intravascular coagulation. The value of pharmacological agents (such as phenoxybenzamine, glucagon, or dopamine) for improving the mesenteric circulation remains uncertain.

As soon as the patient is sufficiently fit, the abdomen must be opened. If a large vessel is occluded, the surgeon may be able to undertake embolectomy or reconstruct an occluded artery. In both occlusive and non-occlusive vascular disease it is necessary to decide how much intestine to resect. If there is doubt about the viability of the residual intestine the abdomen should be closed and re-explored 24 h later. Infiltrating the coeliac and mesenteric plexuses with local anaesthetic may help relieve vascular spasm and should be used if there is no evidence of vascular occlusion. Treatment with anticoagulants may also be indicated.

Chronic intestinal ischaemia

Chronic intestinal ischaemia is usually due to atheroma. The coeliac axis and superior mesenteric artery are commonly affected and the inferior mesenteric artery to a much lesser extent. Stenotic lesions occur at the aortic origins of the vessels. Diffuse, severe atheroma throughout the intestinal arterial tree is uncommon and therefore arterial reconstruction may be very rewarding.

Clinical features

Patients with chronic ischaemia of the gut suffer poorly localized severe cramping abdominal pain. This occurs every day and is usually worse 20 to 60 min after eating. The pain may be relieved by simple analgesics or by vasodilator drugs. As the condition progresses the patient becomes afraid to eat and loses weight. There are no diagnostic physical signs and in particular it is not helpful to find a vascular bruit. Thus clinicians must request vascular imaging whenever there is reasonable suspicion and other tests (including computed tomographic scanning of the abdomen) have failed to give a diagnosis.

Magnetic resonance imaging is best able to provide functional information about the degree of arterial stenosis and quantitative information about blood flow and blood oxygenation. Unfortunately only a few specialized centres are likely to develop sufficient experience with this sophisticated technique.

Compression of the coeliac-axis

In occasional patients with chronic abdominal pain and an abdominal bruit (which may be exacerbated by inspiration), aortography shows apparent constriction of the coeliac axis by the median arcuate ligament. It has been claimed that the arteries in the territory of the coeliac axis 'steal' blood from that of the superior mesenteric artery, thereby causing an intestinal angina that may be relieved by dividing the median arcuate ligament and possibly by reconstructing the coeliac axis. The validity of this syndrome is uncertain.

Ischaemic colitis

The colon is more prone to ischaemic damage than the small intestine. The transverse and descending segments of the colon are supplied by marginal branches of the middle colic (superior mesenteric territory) and left colic (inferior mesenteric territory) arteries. An arterial and lymphatic watershed exists close to the splenic flexure, which is supported to a variable extent by an additional vascular arcade. This segment of the colon is at risk when the mesenteric circulation is compromised. In addition, distension of the colon may impair blood flow. Thus ischaemic colitis may occur in the segment of intestine immediately proximal to an obstructing lesion (stercoral ulceration) or with colonic pseudo-obstruction. Venous occlusion may also cause ischaemic colitis.

Clinical features

In the acute phase of ischaemic colitis the clinician has to differentiate between mild injury, which responds quickly and effectively to supportive measures and treatment with appropriate antibiotics, and severe injury, in which gangrene may develop. Typically the affected person complains of pain in the left iliac fossa,

nausea, and vomiting followed by the passage of a loose motion containing dark blood.

Marked tenderness in the left iliac fossa is the most constant physical sign. At colonoscopy the mucosa may be blue and swollen without contact bleeding. The rectum is invariably spared. Plain radiographs of the abdomen may show an abnormal segment of large intestine outlined with gas.

Contrast enema examination of the colon is a most useful way of demonstrating ischaemic damage. In the early phase 'thumb printing' is the characteristic sign. This may persist for several days (Fig. 1). Subsequently the mucosal appearance may return to normal or progress to mucosal ulceration, giving an appearance that may be indistinguishable from segmental ulcerative colitis or Crohn's disease. These changes may resolve spontaneously or progress to tubular narrowing of the intestine with or without sacculation on the antimesenteric border.

Fig. 1 Ischaemic colitis: barium enema showing thumb printing at the splenic flexure. (By courtesy of Dr A. Freeman, Addenbrooke's Hospital.)

Ischaemic colitis may be confused with dysenteric conditions, acute diverticular disease of the colon, acute inflammatory bowel disease, perforation of a hollow viscus, or left-sided peritonitis caused by pancreatitis. The most important distinguishing features are the association with degenerative cardiovascular disease and the distinctive, although not pathognomonic, radiographic and colonoscopic appearances.

Management

On establishing the diagnosis of ischaemic colitis the treatment is initially expectant. The patient should be given intravenous fluid as necessary, together with systemic broad-spectrum antibiotics. Well over 90 per cent of recognized cases resolve spontaneously. A stricture may develop in up to a third of patients but this is usually asymptomatic and only rarely needs to be resected. Surgery is indicated if there is evidence of peritonitis, persistent bleeding, or of an underlying colonic disorder (such as carcinoma).

Intestinal reperfusion injury

The gut mucosa may be transiently but seriously damaged after a period of ischaemia which is followed by apparently adequate reperfusion. The damage appears to be caused by the generation of reactive oxygen metabolites (including superoxide, hydrogen peroxide, and hydroxyl radicals). These alter the vascular permeability of endothelial cells and damage epithelial cells by peroxidation of cell membranes. The injured tissues are strongly chemotactic for neutrophils and this leads to an acute inflammatory response. Such damage diminishes the barrier function of the gut, increasing intestinal permeability and allowing the translocation of bacteria. Injury of the intestinal mucosa is now recognized as an important factor in the prognosis of critically ill patients in intensive care. Serial measurements of serum intestinal fatty acid protein (iFABP) have been shown to provide a sensitive marker of damage to intestinal mucosa.

Vasculitis nd the collagen disorders: effects on the gut

The gut may be involved in any of the systemic collagen–vascular disorders. Vasculitis may cause focal ischaemic damage of the intestine (Table 2) but this is a feature of many conditions other than the collagen disorders, for example drug-induced ulceration (for example by potassium salts), the after-effects of blunt trauma to the abdomen, irradiation, and rarely infective disease (for example typhoid or leprosy) (Table 3).

In the collagen disorders the visceral muscle may be damaged and the resulting dysmotility may cause dysphagia, delayed gastric emptying, small intestinal stasis with bacterial overgrowth, or colonic inertia. Gas may infiltrate the tissues, giving rise to pneumatosis intestinalis.

The specific pathological diagnosis is usually based on the systemic features of the illness and the laboratory findings rather than on the mostly non-specific abdominal complications. But the inquisitive physician may also recognize curious associations in patients with multisystem disorders. Thus, intestinal malabsorption and protein-losing enteropathy have been described in association with systemic lupus erythematosus and rheumatoid arthritis, pancreatic insufficiency with systemic sclerosis, acute pancreatitis during the course of Behçet's disease, and apparently classical inflammatory bowel disease with systemic lupus erythematosus.

Systemic sclerosis

In primary systemic sclerosis (see Chapter 18.11.3), fibrous connective tissue proliferates. In the gastrointestinal tract it may replace smooth muscle, especially in the oesophagus (which is involved in 80 per cent of cases), to a lesser extent in the small intestine (although duodenal involvement is quite common), and rather rarely in the colon.

Overt vasculitis is a less common feature but occasionally causes intestinal infarction. Pneumatosis cystoides intestinalis is also described, especially in association with intestinal pseudo-obstruction or a pneumoperitoneum.

Clinical features

In systemic sclerosis progressive dysphagia is the most frequent gastrointestinal symptom. Initially there is a decrease in the incidence and amplitude of contractions of the lower oesophagus and incomplete relaxation of the lower oesophageal sphincter. In addition, the resting tone of the sphincter is reduced, allowing reflux of gastric juices, oesophagitis, shortening of the oesophagus, and occasionally stricture formation. Associated hiatal herniation is common.

More rarely the stomach is involved, causing delayed emptying that on occasion is exacerbated by associated stenosis of the pyloric canal. Changes lower down the gastrointestinal tract also occur. Characteristically the duodenum is dilated, the valvulae of the small intestine are thickened, and pseudodiverticula may form. These changes are associated with abdominal discomfort, distension, and borborygmi, especially after the taking of meals. The impaired motility of the small intestine leads to stasis of its contents and bacterial overgrowth causing malabsorption, especially of fat and vitamin B_{12} (see Chapter 14.9.2) (Fig. 2). Progressive constipation due to impaired colonic motility is uncommon.

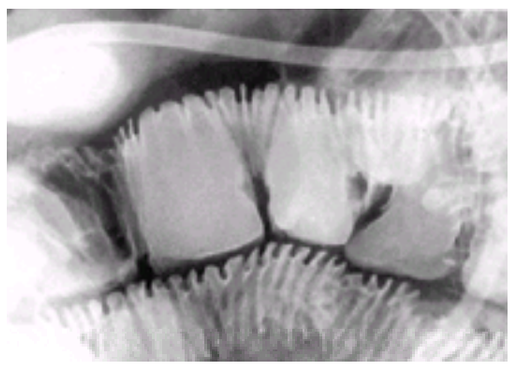

Fig. 2 Systemic sclerosis. Typical 'sacculation' appearance of the bowel.

Management

There is no specific treatment of primary systemic sclerosis. Lesions in the gastrointestinal tract need management on their merits. It is important to recognize early the patient with gastro-oesophageal reflux. A proton pump inhibitor should be given to prevent the ravages of acid-peptic digestion of the oesophageal mucosa. When strictures occur these should be dilated by bouginage. Surgical intervention is occasionally necessary.

A breath test (see Chapter 14.2.3 and Chapter 14.9.2) is a useful screening test for delayed passage of contents and bacterial proliferation in the small intestine. Patients with a positive breath test should be assessed for evidence of malabsorption, and if this is found intermittent therapy with antibiotics for an indefinite period may be of clinical value.

Systemic lupus erythematosus

Systemic lupus erythematosus (see Chapter 18.11.2) may cause abdominal symptoms arising from any part of the gastrointestinal tract. Anorexia, weight loss, nausea, vomiting, and diarrhoea are relatively common. Dysphagia, abdominal pain, distension due to ascites, and gastrointestinal bleeding are less frequent symptoms. Occasionally a patient with systemic lupus erythematosus develops an acute abdomen, which may be due to localized or widespread lupus vasculitis causing ischaemic damage to the gut or its related organs, including the gallbladder and pancreas. Arteriography may be helpful in diagnosis by revealing diffuse irregularities in the branches of mesenteric vessels.

Treatment with oral corticosteroids usually relieves minor abdominal symptoms and will lead to rapid resolution of simple ascites. In the acute stage of the disease, however, surgery may be necessary to deal with infarcted intestine, serious bleeding, or intestinal obstruction.

Other systemic disorders

In rheumatoid arthritis, vasculitis is associated with long-standing disease, seropositivity, and florid subcutaneous nodule formation (see Section 18.11). Occasionally, a severe diffuse and necrotizing angiitis causes infarction in the gallbladder, pancreas, or intestine. Symptoms vary from vague abdominal pain, with or without diarrhoea, to the development of an acute abdomen.

Dermatomyositis rarely causes damage to the viscera although thrombosis of small vessels occasionally causes gastrointestinal ulceration.

In Behçet's syndrome the triad of relapsing iritis, painful ulcers of the mouth, and genital ulceration is only part of the syndrome (see Chapter 18.11.5). Again, vasculitis appears to be the underlying histopathological lesion. In the gastrointestinal tract this may lead to ulceration of the colon, malabsorption (sometimes with lymphangiectasia), and pancreatitis.

Primary vasculitis

Henoch–Schönlein purpura (anaphylactic purpura)

This is a self-limiting disorder of unknown cause characterized by small-vessel vasculitis (see Section 18.11). Gastrointestinal disease occurs in at least two-thirds of cases and is manifest as abdominal pain and gastrointestinal bleeding. Intramural haematomas are common and rarely may be complicated by intussusception, perforation, or an infarcted segment of gut.

Polyarteritis nodosa

Abdominal pain and other gastrointestinal symptoms are common in patients with polyarteritis nodosa. The underlying cause is usually recognized by evidence of systemic disease such as skin lesions, renal involvement, hypertension, and eosinophilia. Mesenteric angiography is useful as a diagnostic tool because up to two-thirds of cases have recognizable aneurysms of mesenteric and renal vessels. A small proportion of patients with polyarteritis have acute abdominal episodes including ulceration, haemorrhage, perforation and segmental necrosis of intestine, cholecystitis, pancreatitis, and hepatic infarction. Kawasaki disease (infantile acute febrile mucocutaneous lymph node syndrome) (see Chapter 18.11.8) proceeds to a disorder indistinguishable histopathologically from infantile periarteritis nodosa. Cardiac involvement is most common, but the gastrointestinal tract is affected in up to a third of cases.

Antineutrophilic cytoplasmic antibody-positive vasculitides

Wegener's granulomatosis, Churg–Strauss syndrome, and microscopic polyarteritis are conditions frequently associated with the finding of antineutrophilic cytoplasmic antibodies. Gastrointestinal symptoms are common in these conditions, although the intra-abdominal pathology has not been well characterized except when angiitis has led to a life-threatening condition such as visceral perforation or infarction.

Giant cell arteritis

This characteristically affects the larger cranial arteries including the ciliary and central retinal arteries (see Chapter 18.11.4) and rarely limb arteries. Very occasionally a similar pathology affects mesenteric arteries and causes bowel infarction.

Localized arteritis

Arteritis has been described causing pathology solely in the appendix, the gallbladder, and the pancreas. The relationship of a localized arteritis to systemic polyarteritis is uncertain. Similarly, localized leucocytoclastic (hypersensitivity) vasculitis has been described in the abdominal cavities.

Other vascular disorders that may affect the gut

Aneurysms of the aorta and its major branches

Rarely, aneurysms fistulate into the stomach or duodenum. This usually causes catastrophic bleeding and rapid death. Even more rarely there is intermittent bleeding (for example from the splenic artery into the stomach), which may be difficult to diagnose.

Superior mesenteric artery syndrome

A syndrome of postprandial epigastric pain, distension, and vomiting may occur in asthenic young people, especially those who have lost weight or who are fixed in a position of hyperextension after spinal injury. Barium studies show a distended proximal duodenum with a sharp cut-off at the line where the superior mesenteric artery crosses the duodenum. Symptoms may be relieved if the patient adopts the prone position after meals and usually disappear as the patient gains weight. Surgery is occasionally necessary. The condition must be distinguished from duodenal ileus caused by mesenteric bands, a condition that is associated with partial malrotation of the midgut.

Haemangioma

Haemangiomas are uncommon but they may cause painless bleeding especially in the jejunum.

Intestinal telangiectasia

These lesions occur most commonly with Osler–Weber–Rendu disease (see Section 22.04.04) and may lead to microscopic bleeding with anaemia, especially in adult life.

Vascular dysplasia

This is a more recently recognized and not uncommon disorder causing occult bleeding from the gut in older subjects. The lesions occur as small arteriovenous malformations or as foci of ectatic capillaries or veins with little supporting stroma. They are found predominantly in the caecum and ascending colon. There may be an association with aortic stenosis but none with cutaneous telangiectases and no familial aggregations have yet been described.

Patients give a history of recurrent anaemia or episodes of bleeding from the gut, have usually been investigated repeatedly without getting a firm diagnosis, and sometimes have had one or more operations (including resection of a segment of the gastrointestinal tract) without relief of symptoms. The diagnosis of vascular dysplasia should be considered in all cases of obscure gastrointestinal haemorrhage and may be made by direct visualization of the intestinal mucosa (Fig. 3) or by selective mesenteric arteriography (Fig. 4). The lesions may be multiple, in which case resection of the affected segment of gut may be necessary. Many patients, however, can be treated successfully by fulguration of the lesion through an endoscope. If the lesion(s) cannot be obliterated, in women a trial of treatment with an oestrogen–progesterone preparation is often effective.

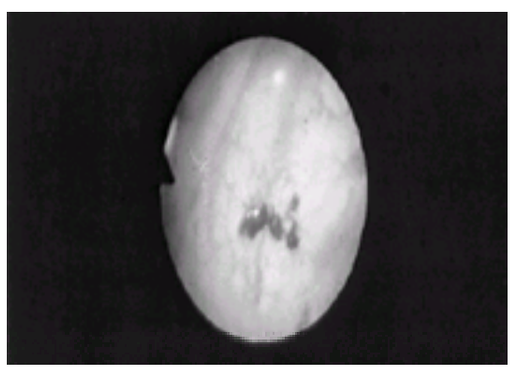

Fig. 3 Angiodysplastic lesion in the caecum, photographed through a colonoscope. (By courtesy of Dr R. Hunt, RN Hospital, Haslar.)

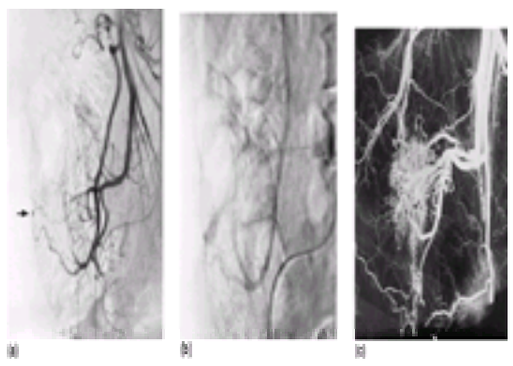

Fig. 4 (a) Angiodysplastic lesion in the caecum: superior mesenteric angiogram in a 53-year-old man with anaemia for 20 years (no lesion found at previous operations). Vascular lake in caecum (arrowed). (b) Angiodysplastic lesion in the caecum: superior mesenteric angiogram in a 53-year-old man with anaemia for 20 years (no lesion found at previous operations). Capillary phase, showing early filling vein arising from lesion. (c) Angiodysplastic lesion in the caecum: superior mesenteric angiogram in a 53-year-old man with anaemia for 20 years (no lesion found at previous operations). Injected specimen magnified ×30. (By courtesy of Dr D. J. Allison, Royal Postgraduate Medical School and previously published in *British Journal of Hospital Medicine*, 1980; **23**: 358.)

Intramural bleeding

Bleeding into the wall of the bowel may occur as a result of treatment with anticoagulants or from the inflammation of small vessels (as occurs classically in Henoch–Schönlein purpura). The usual presentation is with colicky abdominal pain, with bleeding into the lumen of the gut. Appropriate barium examination may show the classical sign of 'thumb printing'. The condition usually resolves spontaneously providing that the underlying disorder can be treated. A blood transfusion may be needed.

Further reading

Bryant DS, Pellicane JV, Davies RS (1997). Non-occlusive intestinal ischemia: improved outcome with early diagnosis and therapy. *American Surgeon* **63**, 334–9. Good article on diagnosis and management of a poorly recognized condition.

Cappell MS (1998). Intestinal (mesenteric) vasculopathy (Parts I and II). *Gastro-enterological Clinics of North America* **27**, 783–860. Comprehensive review of the mesenteric vasculopathies.

Heiss SG, Li KC (1998) Magnetic resonance angiography of mesenteric arteries. A review. *Investigative Radiology* **33**, 670–81. Important non-invasive method of assessing mesenteric vessels and their blood flow.

Hunder GG, ed (1992). Vasculitic syndromes. *Current Opinion in Rheumatology* **44**, 1–55. Good review of vasculitic syndromes.

Jamieson CW (1986). Coeliac axis compression syndrome. *British Medical Journal* **293**, 159. Classical description of coeliac compression syndrome.

Lie JT (1997). Mesenteric inflammatory veno-occlusive disease (MIVOD): an emerging and unsuspected cause of digestive tract ischemia. *Vasa* **26**, 91–6. An newly described condition.

Marston A (1986). *Vascular disease of the gut*. Arnold, London. Classical book on vascular disease of the gut.

Pastores SM, Katz DP, Kvetan V (1996). Splanchnic ischemia and gut mucosal injury in sepsis and the multi-organ dysfunction syndrome. *American Journal of Gastroenterology* **91**, 1697–710. Description of the ischaemia-reperfusion syndrome.

14.17 Gastrointestinal infections

Davidson H. Hamer and Sherwood L. Gorbach

Introduction

Diarrhoea, the most common manifestation of intestinal tract infections, is a leading cause of death in most developing countries where its greatest impact is seen in infants and children. Infectious diarrhoea may be accompanied by numerous complications (Table 1). The financial burden associated with medical care and lost productivity due to infectious diarrhoea amounts to more than 20 billion dollars a year in the United States alone.

The aetiology and severity of gastrointestinal infections are determined by several epidemiological factors. Young children and the elderly are at greatest risk for more severe disease and complications. The presence of underlying medical conditions, especially those that compromise immunity, greatly enhances the risk of acquiring an infection and its ultimate severity. Poor sanitation, inadequate water supplies, and increasing globalization of food transport systems all predispose to the development of large epidemics of food- and water-borne outbreaks of gastrointestinal disease. Seasonal or cyclic weather variations also influence the epidemiology of diarrhoeal disease and food poisoning.

A wide array of bacterial, protozoal, and viral pathogens is responsible for gastrointestinal tract infections. The characteristics of specific organisms are described in detail in Section 7 of this book: here are presented the pathophysiology, common clinical syndromes, diagnosis, management, and prevention of gastrointestinal diseases.

Pathophysiology

Host factors

Normal intestinal flora

The proximal small bowel, including the stomach, duodenum, jejunum, and upper ileum, has a relatively sparse microflora, with most organisms being derived from the oropharynx. Colonization of the upper intestine by Gram-negative bacilli is an abnormal event, one that is characteristic of illness due to pathogens such as *Vibrio cholerae* and *Escherichia coli*. The large bowel has an abundant microflora, with total concentrations of 10^{11} bacteria per gram of content. Anaerobes including *Bacteroides* spp., *Clostridium* spp., and anaerobic streptococci outnumber aerobic bacteria, such as coliforms, by 1000-fold. During an episode of acute diarrhoea, regardless of the aetiology, the colonic flora becomes less anaerobic because of the rapid transit of intestinal contents. As a consequence, strictly anaerobic bacteria decrease in number while there is an increase in coliforms, which are often aberrant types such as *Enterobacter*, *Klebsiella*, and *Proteus* spp. The pathogen itself assumes a dominant position in the flora, so that the major faecal isolate may be *Salmonella* spp. or *V. cholerae.*

In addition to the longitudinal distribution of bacteria in the gastrointestinal tract, the bowel microflora is found both within the lumen and adherent to the mucous layer overlying epithelial cells. Invasive pathogens such as *Campylobacter*, *Shigella*, *Salmonella*, and *Yersinia* spp. can penetrate the mucosal surface and infect epithelial cells, or translocate into the mesenteric lymph nodes and bloodstream.

Control mechanisms

At the portal of entry, gastric acid suppresses most organisms that are ingested. In the setting of reduced or absent gastric acid, there is a higher incidence of bacterial colonization of the upper small intestine. Consequently, persons with hypochlorhydria, achlorhydria, or those using drugs such as proton-pump inhibitors that inhibit gastric acid secretion are susceptible to diarrhoeal diseases. A critical element in maintaining the sparse flora of the upper bowel is propulsive motility. The antibacterial properties of biliary fluid may control the intestinal flora. The glycocalyx and intestinal mucins secreted by epithelial cells provide a mechanical barrier to invasion by gut pathogens. Finally, antibacterial substances produced by the normal intestinal microflora help to maintain the stability of normal populations of organisms and to prevent the implantation of pathogens.

Intestinal immunity (see Chapter 14.4)

The intestinal immune system plays a major role in the host's response to enteric pathogens. The human gut contains a large amount of lymphoid tissue in the form of intraepithelial lymphocytes, lamina propria lymphoid cells, and Peyer's patches. The latter are lymphoid aggregates in the mucosa and submucosa of the distal small intestine which serve as sites for the presentation of antigens to B and T lymphocytes. After activation by antigens, bacteria, or viruses in the Peyer's patches, the lymphocytes migrate to the lamina propria and the intraepithelial portion of the intestinal lining where, along with macrophages and other types of white blood cells, they protect the host from specific pathogens. Plasma cells in the lamina propria produce secretory immunoglobulin A, which is released into the intestinal lumen. When the mechanical barrier of the gut fails, then the intraepithelial and lamina propria lymphocytes provide the next level of protection against pathogenic enteric organisms.

Microbial factors

The number of organisms that need to be ingested to establish a gastrointestinal tract infection varies from as few as 10 to 100 in the case of *Shigella* spp. to as many as 10^8 for *V. cholerae*. In the presence of reduced gastric acidity or underlying immunosuppression, the inoculum needed to establish infection is reduced.

Enteric pathogens can cause intestinal disease by means of enterotoxins, adherence to gut mucosa, or invasion of enterocytes.

Toxins

Bacterial enteric pathogens can elaborate enterotoxins that act directly on intestinal epithelial cells (for example, cholera toxin) or preformed toxins that are ingested in contaminated food (for example, *Bacillus cereus* toxin). While invasive bacteria penetrate the mucosal surface of the gut as the primary event, they may also secrete enterotoxins. Production of enterotoxin can be demonstrated in the laboratory by *in vivo* tests—such as the rabbit ileal loop model and the suckling mouse model, or by *in vitro* tests involving a tissue culture line, such as Y-1 adrenal cells or Chinese hamster ovary cells.

Many organisms elaborate enterotoxins that cause fluid and electrolyte secretion in the gut. Diarrhoeal toxins can be grouped into two categories: cytotonic, which produce fluid secretion by activation of intracellular enzymes such as adenylate cyclase, without causing any damage to the epithelial surface; and cytotoxic, which

cause injury to the mucosal cell while also inducing fluid secretion, but not primarily by activation of cyclic nucleotides. *V. cholerae* and enterotoxigenic *Escherichia coli* (**ETEC**) are examples of pathogens that cause dehydrating diarrhoea by producing enterotoxins of the cytotonic type (see Chapter 7.11.7).

Intestinal fluid loss is the primary manifestation of cholera—this results from the action of enterotoxin on the small bowel epithelial cells. These organisms colonize the small intestine, adhering to epithelial cells and then elaborating enterotoxin. There is no invasion of the mucosal surface so there is no evidence of damage to the mucosal architecture and bacteraemia is not a complication. The faecal effluent is watery, often voluminous, and produces the clinical features of dehydration. The most sensitive areas are the upper bowel, particularly the duodenum and upper jejunum; the ileum is less affected, and the colon is usually in a state of absorption since it is relatively insensitive to the toxin. This is a form of 'overflow' diarrhoea, with a large volume of fluid produced in the upper intestine that overwhelms the capacity of the lower bowel to absorb.

ETEC produces two types of enterotoxins: a heat-labile (**LT**) and a heat-stable toxin (**ST**). LT is a protein that is destroyed by heat and acid; like cholera toxin, it activates adenylate cyclase, causing secretion of fluid and electrolytes into the lumen. In contrast, ST can withstand heating to 100 °C and acts by activating guanylate cyclase. Despite the differences between the two toxins, the ultimate effect of both enterotoxins is a non-inflammatory secretory diarrhoea.

Invasion

Whereas toxigenic organisms usually involve the upper intestine, invasive pathogens target the lower intestine, particularly the distal ileum and colon. Histological findings include evidence of mucosal ulceration with acute inflammation in the lamina propria. Principal pathogens in this group are *Salmonella* spp., *Shigella* spp., enterohaemorrhagic *E. coli* (**EHEC**), enteroinvasive *E. coli* (**EIEC**), *Campylobacter spp.,* and *Yersinia* spp. Although there are important differences among these organisms, they all have in common the property of mucosal invasion as the initiating event. To date, three theories have been invoked to explain the mechanism of fluid production in invasive diarrhoea. First, fluid production may result from an enterotoxin, at least in the initial phase of the illness. Most *Shigella* strains elaborate an enterotoxin that differs substantially from cholera toxin, but which does result in fluid and electrolyte secretion by the intestine. A similar toxin has been proposed for *Salmonella spp.,* and there is suggestive evidence that *Campylobacter* and *Yersinia* spp. elaborate enterotoxins. Second, invasive organisms lead to an increased local synthesis of prostaglandins at the site of the intense inflammatory reaction that may be responsible for fluid secretion and diarrhoea. Third, damage to the epithelial surface may prevent reabsorption of fluids from the lumen and thereby result in a net accumulation of fluid in the bowel lumen, resulting in diarrhoea.

A series of pathogenic factors, each controlled by plasmids or chromosomal loci, are used by pathogenic strains of *Salmonella*. Specific plasmids encode for bacterial spread from Peyer's patches to other sites in the body, for the ability of certain strains to survive within macrophages following phagocytosis, and for the ability of salmonellae to elicit transepithelial signalling to neutrophils (see Chapter 7.11.7). Invasion by *Shigella* spp. is also associated with diverse virulence factors related to various stages of invasion. The end result is the death of the intestinal epithelial cell, focal ulcers, and inflammation of the lamina propia. The shigella virulence factors are encoded by chromosomal and plasmid genes, all of which are needed for the full expression of virulence. Various genetic loci encode for an invasion plasmid antigen (*ipa*), which seems to determine recognition of the epithelial cell, *inv* invasion factors, and a series of *vir* loci that are involved in regulation within the infected cell. After penetrating the mucosal surface of the gut, *Shigella* spp. multiply within epithelial cells and extend the infected area by direct cell-to-cell migration of bacilli. *Shigella* species rarely penetrate beyond the intestinal mucosa and therefore do not usually invade the bloodstream.

Adherence

Specific fimbriae or adhesins mediate the attachment of pathogenic bacteria to gut mucosal cells. For example, the attachment of *V. cholerae* is mediated by a fimbrial colonization factor, known as the toxin-coregulated pilus. Some enteric pathogens such as enteropathogenic *E. coli* (**EPEC**) attach to the intestinal mucosa in a characteristic manner, producing ultrastructural changes known as attachment-effacement lesions; this leads to the elongation and destruction of microvilli. Protozoal parasites such as *Giardia lamblia* use a ventral adhesive disc to attach to the mucosal surface of the small intestine. Thus, enteropathogens have devised a number of different ways to adhere to the surface of the gut.

Clinical syndromes of gastrointestinal infections

Gastrointestinal infections usually result in three principal syndromes: non-inflammatory diarrhoea, inflammatory diarrhoea, and systemic disease. Non-inflammatory diarrhoea primarily involves the small intestine, whereas inflammatory diarrhoea predominantly affects the colon. The location of infection influences the clinical characteristics and certain diagnostic features of the diarrhoeal disease (Table 2). Thus, the organisms that target the small intestine tend to produce watery, potentially dehydrating diarrhoea, while those infecting the large intestine cause bloody mucoid diarrhoea associated with tenesmus.

Non-inflammatory diarrhoea

Bacteria

Cholera, the prototypic non-inflammatory diarrhoea, can cause dehydration and death within 3 to 4 h of onset. Like many other infectious diseases, there is a spectrum of clinical manifestations—from an asymptomatic carrier state to severe dehydration with shock. Initial symptoms of vomiting and abdominal distention are rapidly followed by diarrhoea, which accelerates over the next few hours to frequent purging of large volumes of 'rice-water' stools. The acutely ill patient has marked dehydration manifested by poor skin turgor, 'washerwoman's hands', feeble to absent pulses, reduced renal function, and hypovolaemic shock.

Non-01 cholera vibrios have also been associated with severe, dehydrating diarrhoea as well as wound infections and septicaemia. *V. vulnificus* is one of the most important non-cholera vibrios, based on the severity of illness that it causes, especially in patients with underlying liver disease and especially iron-storage disease. This infection can be acquired by direct consumption of seafood, usually raw oysters, or as a wound infection in people who have direct contact with salt water. Since this infection can be lethal in susceptible people, such persons should be warned about eating raw seafood, especially oysters.

ETEC infections are one of the most common causes of diarrhoea in travellers to less developed countries and children living in these regions. The incubation period of this infection is usually between 24 and 48 h, after which the disease often begins with upper intestinal distress, followed soon thereafter by watery diarrhoea. The infection can be extremely mild, with only a few loose movements, or it can be quite severe, mimicking cholera with profuse watery diarrhoea leading to severe dehydration. Other strains of *E. coli* such as enteroaggregative, diffusely adhering, and enteropathogenic *E. coli*, may also be associated with watery diarrhoea.

Viruses (see also Chapter 7.10.8)

Numerous viruses are responsible for as many as 30 to 40 per cent of self-limited episodes of non-inflammatory diarrhoea, especially in children. Rotavirus causes a range of clinical manifestations from asymptomatic carriage to severe, potentially fatal dehydration. The disease occurs primarily in children aged between 3 and 15 months; infections continue into the second year of life, but after this age are less common. Adults can develop mild infections with group A rotaviruses, especially if there is a sick child in the household. The disease process often begins with vomiting, followed shortly thereafter by watery diarrhoea. The incubation period is between 1 and 3 days, with an average duration of illness of 5 to 7 days, although some instances of chronic diarrhoea have been described.

Caliciviruses are single-stranded RNA viruses that are responsible for human and animal infections. Recent molecular studies have shown that the Norwalk and Norwalk-like viruses have a genetic composition that places them in the taxonomic family of Caliciviridae. This family of viruses typically causes disease mainly in infants and young children, especially in day-care centres. The illness is generally mild and indistinguishable from that due to rotavirus or even epidemic Norwalk disease. The Norwalk virus causes explosive epidemics of diarrhoea that sweep through communities with a high attack rate.It shows no respect for age, as it can affect virtually all age groups except infants. Infections caused by the Norwalk agent tend to be relatively mild and short-lived, with common symptoms including diarrhoea, nausea, abdominal pain, vomiting, and myalgias. Generally, the clinical illness lasts no longer than 24 to 48 h.

Astroviruses are responsible for outbreaks of diarrhoea in day-care centres and in communities with infants.The disease is characterized by watery or mucoid stools, nausea, vomiting, and, occasionally, fever, but it tends to be milder than rotavirus diarrhoea as there is less dehydration. Adenovirus serotypes 40 and 41 are responsible for day-care centre and nosocomial outbreaks of gastroenteritis in children under two years of age. As opposed to rotavirus or Norwalk virus, infection with enteric adenovirus has a long incubation period lasting approximately 8 to 10 days, and the illness can be prolonged for as long as 2 weeks.

Other infestations (also see Section 7)

Giardia lamblia is responsible for clinical syndromes ranging from asymptomatic cyst passage, to self-limited diarrhoea, to chronic diarrhoea with malabsorption and weight loss. After an incubation period between 1 and 2 weeks, patients experience the onset of frequent, loose to watery bowel movements associated with abdominal cramps, bloating, belching, nausea, anorexia, and flatulence.

Patients with cryptosporidiosis present with watery diarrhoea associated with abdominal pain, nausea, vomiting, low-grade fever, malaise, and anorexia. Faecal output may be voluminous and dehydrating in immunocompromised patients, particularly those with underlying human immunodeficiency virus (**HIV**) infection. Symptoms usually resolve by 5 to 10 days. Infection with *Cyclospora cayetanensis* is manifested by anorexia, intermittent diarrhoea, and nausea. Diarrhoea is usually self-limiting, but it can last for several weeks in immunocompetent patients and result in significant weight loss. *Isospora belli* also causes a self-limited illness characterized by watery, non-bloody diarrhoea, abdominal cramping, anorexia, weight loss, and, less commonly, fever. Any of the parasitic infections is more severe and longer lasting in immunocompromised patients, such as those with HIV or organ transplants.

Food poisoning

Food poisoning is most commonly caused by the consumption of food contaminated with bacteria or bacterial toxins. Food poisoning can also be due to parasites (for example, trichinosis), viruses (e.g., hepatitis A), and other toxins (e.g., mushrooms see Section 8.3). The most well-recognized causes of bacterial food poisoning are the following: *Clostridium perfringens, Staphylococcus aureus, Vibrio* spp. (including *V. cholerae* and *V. parahaemolyticus*), *Bacillus cereus, Salmonella* spp., *C. botulinum, Shigella* spp., toxigenic *E. coli* (ETEC and EHEC), and certain species of *Campylobacter, Yersinia, Listeria,* and *Aeromonas.*

An enterotoxin elaborated by type A strains of *C. perfringens* is responsible for food-borne outbreaks with high attack rates but which are of short duration. *C. perfringens* food poisoning is characterized by severe, crampy abdominal pain and watery diarrhoea, usually without vomiting, beginning 8 to 24 h after the incriminating meal. Fever, chills, headache, or other signs of infection are usually absent. Strains of *C. perfringens* type C elaborate a similar enterotoxin that has been implicated in outbreaks of enteritis necroticans secondary to the consumption of rancid meat in Europe, also known as 'pigbel' in Papua New Guinea. This is a much more severe, necrotizing disease of the small intestine and carries a high mortality rate.

Staphylococcal food poisoning presents with severe vomiting, nausea, and abdominal cramps, often followed by diarrhoea. *B. cereus* is an aerobic, spore-forming, Gram-positive rod that has been associated with two clinical types of food poisoning—a diarrhoea syndrome and a vomiting syndrome. The latter has a short incubation period of about 2 h, after which nearly all affected persons experience vomiting and abdominal cramps. In contrast, the diarrhoea syndrome has a median incubation period of 9 h; clinical illness is characterized by diarrhoea, abdominal cramps, and vomiting. *B. cereus* is particularly associated with the ingestion of contaminated rice that has been kept for a long time in a warm or partially cooked state in take-away food outlets. Fevers are uncommon with all three of these bacterial toxin-mediated syndromes. Episodes of staphylococcal and *B. cereus* food poisoning are short-lived, usually resolving within 24 h. Often the staphylococcus has been introduced by contamination from a small abscess, whitlow, or other discharging lesion present during preparation of food, which is allowed to remain warm and not fully cooked before serving.

Travellers' diarrhoea

People who travel from industrialized countries to less developed areas of the world are at risk of contracting traveller's diarrhoea, with as many as 25 to 50 per cent or more suffering from one or more episodes of diarrhoea. The greatest frequency of diarrhoea occurs in students or low-budget tourists. Business travellers are at intermediate risk, while travellers who are visiting relatives have the lowest risk. Young travellers—particularly those 20 to 29 years old—have the highest risk, whereas the lowest rates of travellers' diarrhoea are noted in those over 55 years of age. The disease does not begin immediately but generally starts 2 to 3 days after the traveller's arrival. While most people have three to five watery, loose stools daily, about 20 per cent can have as many as 6 to 15. A minority of patients, approximately 2 to 10 per cent, has fever, bloody stools, or both—these people are more likely to have shigellosis. Diarrhoea is frequently associated with gas, cramps, fatigue, nausea, abdominal pain, fever, and anorexia. The illness usually resolves without specific therapy within 3 to 5 days, although a few unfortunate travellers will have persistent diarrhoea.

Infectious micro-organisms in contaminated food and drink are the main source of travellers' diarrhoea. Especially risky foods include uncooked vegetables, meat, and seafood. Tap water, ice, unpasteurized milk and dairy products, salads, and unpeeled fruits are also associated with an increased risk. Although an array of pathogens has been found, the leading culprits are various forms of *E. coli,* particularly ETEC. *C. jejuni* is encountered in a significant proportion of cases, particularly during cooler seasons. Viruses, *Shigella*, *Salmonella, Giardia, Cryptosporidium,* and *Cyclospora* spp. are responsible for a minority of travellers' diarrhoea cases.

Prudent selection of beverages and foods can help reduce the risk of developing travellers' diarrhoea. Bottled carbonated beverages, hot coffee or tea, beer, and boiled water are generally safe choices for fluids. Avoiding salads, unpeeled fruit, ice, and undercooked or raw meat, poultry, and seafood can help lower the risk. Because the venue of food consumption determines the risk of contracting travellers' diarrhoea, travellers should be advised to avoid eating food from street vendors. While studies have shown high protection rates when prophylactic antimicrobial agents such as ciprofloxacin or cotrimoxazole are taken, this approach is generally not advised because of the risk of side-effects and emergence of antibiotic-resistant enteric flora.

Chronic non-inflammatory diarrhoea

Certain pathogens cause chronic diarrhoea of small intestinal origin. Some patients with giardiasis develop chronic diarrhoea associated with fatigue, steatorrhoea, weight loss, and intermittent constipation, along with malabsorption of fat, vitamins A and B_{12}, protein, and D-xylose. Acquired lactose intolerance is common, but a lactose-free diet should be recommended in such cases. Cryptosporidiosis can become a chronic, dehydrating diarrhoea in immunocompromised patients, especially in those with the acquired immunodeficiency syndrome (**AIDS**). Complications of chronic cryptosporidiosis include malabsorption, wasting, and biliary tract disease. Patients with AIDS are also at risk for chronic, non-inflammatory diarrhoea due to diffusely adherent *E. coli*, microsporidia, and *Isospora* and *Cyclospora* spp.

About 1 to 3 per cent of travellers returning from a developing country will have persistent diarrhoea that may last for 1 month or more. *Giardia*, *Cyclospora*, and, rarely, *Shigella*, *Salmonella* spp., or *C. jejuni* may be responsible for persistent diarrhoea in travellers. A causative agent is not identified in many travellers suffering from prolonged diarrhoea. Some of these unfortunate individuals will respond to empirical therapy with broad-spectrum antibiotics since they have 'tropical jejunitis' or a mild form of tropical sprue.

Bacterial overgrowth in the small intestine can result in chronic diarrhoea, steatorrhoea, bloating, abdominal pain, and wasting. Factors contributing to the development of this problem include achlorhydria, decreased motility (as may be seen in diabetes mellitus or scleroderma), and stasis due to diverticula or blind loops of bowel. Treatment with amoxicillin/clavulanic acid, erythromycin, or tetracycline in conjunction with a lactose-free diet will often lead to resolution of the diarrhoea.

Inflammatory diarrhoea

Acute inflammatory diarrhoea is the result of infection with bacterial enteropathogens such as *Shigella*, *Campylobacter*, *Salmonella* spp., EHEC, *V. parahaemolyticus*, and *C. difficile*. Among the parasites, *Entamoeba histolytica* is the most common cause of dysenteric illness although *Balantidium coli*, *Schistosoma mansoni*, *S. japonicum*, *Trichuris trichiura*, hookworms, and *Trichinella spiralis* can all cause bloody, mucoid diarrhoea (see Section 7).

Dysentery is an oft-used term that refers to a diarrhoeal stool that contains an inflammatory exudate composed of blood and polymorphonuclear leucocytes. Patients with bacillary dysentery classically present with crampy abdominal pain, rectal burning ('tenesmus'), and fever, associated with multiple small-volume, bloody mucoid, bowel movements. The most constant findings are lower abdominal pain and diarrhoea. Fever is present in less than half of patients and the typical dysentery stool, consisting of blood and mucus, in only one-third. Sigmoidoscopy reveals acute mucosal inflammation with ulcerations and focal haemorrhage.

Bacteria

The *Shiga* bacillus, *S. dysenteriae* type 1, produces the most severe form of dysentery, while *S. sonnei* produces the mild disease. *S. flexneri* is the most commonly encountered serogroup in tropical countries, whereas *S. sonnei* is the most common in industrialized nations. Many patients with shigellosis manifest a biphasic

illness. The initial symptoms of fever, abdominal pain, and watery, non-bloody diarrhoea result from the action of enterotoxin. The second phase, starting 3 to 5 days after the onset of symptoms, is notable for tenesmus and small-volume bloody stools. This period corresponds to invasion of the colonic epithelium and acute colitis. Infection with *S. dysenteriae* type 1 and malnutrition, especially in young children, are factors associated with a more severe course. Complications of shigellosis include intestinal perforation, protein-losing enteropathy, hypoglycaemia, seizures, thrombocytopenia, and haemolytic–uraemic syndrome—the latter three being particularly common in children.

Campylobacter species, especially *C. jejuni*, have gained in prominence as invasive diarrhoeal pathogens. Clinically, disease manifestations range from frank dysentery, to watery diarrhoea, to asymptomatic excretion. Most patients have diarrhoea, fever, and abdominal pain—about 50 per cent will note bloody stools. Constitutional symptoms such as headache, myalgias, backache, malaise, anorexia, and vomiting are often present. The illness usually resolves in less than 1 week, although symptoms can persist for 2 weeks or more, and relapses occur in as many as one-quarter of patients. Rare complications include gastrointestinal haemorrhage, toxic megacolon, pancreatitis, cholecystitis, haemolytic uraemic syndrome, bacteraemia, meningitis, and reactive arthritis, and Guillain-Barré syndrome.

Recent years have seen an increasing frequency of outbreaks of *Salmonella enteritidis* associated with the consumption of uncooked or raw eggs. Salmonella gastroenteritis is characterized by initial symptoms of nausea and vomiting, followed by abdominal cramps and diarrhoea which is accompanied by fever in about 50 per cent of persons. The diarrhoea varies from a few loose stools, to dysentery with grossly bloody, purulent faeces, to a cholera-like syndrome.

Yersinia enterocolitica can cause illness ranging from acute non-bloody diarrhoea to invasive colitis and ileitis. Fever, abdominal cramps, and haem-positive diarrhoea that may persist for several weeks characterize yersinia enterocolitis. *V. parahaemolyticus* outbreaks have been associated with the consumption of raw fish or shellfish. Illness is generally characterized by explosive, watery diarrhoea, abdominal cramps, nausea, vomiting, and headaches. In some cases a bloody dysenteric syndrome is observed.

EIEC strains are capable of invading epithelial cells and producing a shiga-like toxin. Patients with EIEC present with diarrhoea, tenesmus, fever, and abdominal cramps. EHEC strains possess at least two virulence factors that produce intestinal damage: an adherence mechanism causing attachment-effacement lesions similar to those seen with EPEC; and the production of two shiga-like cytotoxins (SLT I and II). Some EHEC strains produce only SLT I or II, whereas others produce both toxins. After a mean incubation of 3 to 4 days, illness begins with watery, non-bloody diarrhoea associated with severe abdominal cramping, nausea, vomiting, chills, and low-grade fever. The diarrhoea then often progresses to visibly bloody stools. Leucocytosis with a shift to the left is usually present, but anaemia is uncommon unless infection is complicated by the development of the haemolytic–uraemic syndrome (**HUS**) or thrombotic thrombocytopenic purpura (**TTP**). The median duration of diarrhoea is 3 to 8 days—longer durations have been described in children and persons with bloody diarrhoea.

Parasites

While infection with a number of different intestinal nematodes and trematodes can be associated with an inflammatory diarrhoea, *E. histolytica* is by far the most common parasitic cause of dysenteric illness (see Chapter 7.13.1). Approximately 50 million cases of invasive colitis due to *E. histolytica* occur worldwide each year, primarily in developing countries. In industrialized countries, populations at high risk of infection include institutionalized persons, especially the mentally impaired, recent immigrants, returning travellers, and sexually active male homosexuals. Malnutrition, malignancy, glucocorticoid use, pregnancy, and young age are risk factors for greater severity of infection.

There are two distinct species of *Entamoeba* that can be differentiated on the basis of antigenic structure, isoenzyme analysis, host specificity, *in vitro* growth characteristics, *in vivo* virulence, and DNA characterization. The two species, *E. histolytica* and *E. dispar*, have the same lifecycle and are morphologically identical. However, *E. dispar* is associated with an asymptomatic carrier state, while *E. histolytica* is capable of invading tissue and causing symptomatic infection.

A spectrum of clinical illness occurs with *E. histolytica* infections including asymptomatic carriage, non-bloody diarrhoea, acute dysenteric colitis, fulminant colitis with perforation, chronic non-dysenteric colitis, and the formation of an amoeboma, an annular lesion of the colon that can be confused with colon cancer. Patients with acute amoebic dysentery usually present with a 1- to 3-week history of bloody diarrhoea, tenesmus, and abdominal pain. Fever and dehydration are present in a minority of patients. Complications of amoebic colitis include intestinal perforation and toxic megacolon. Although nearly all patients have blood in the stool, faecal leucocytes are usually absent, probably as a result of the lysis of inflammatory cells by trophozoites. Amoebic liver abscess can occur with or independent of acute colitis. The fulminant variant of amoebic colitis is characterized by the rapid onset of fever, bloody mucoid diarrhoea, diffuse abdominal pain with peritoneal signs, and leucocytosis. Chronic non-dysenteric amoebiasis is a syndrome usually lasting more than 1 year with intermittent diarrhoea, mucus, abdominal pain, flatulence, and weight loss.

Antibiotic-associated colitis

Although the mechanism has not been fully elucidated, it appears that the normal bowel flora inhibits overgrowth by *C. difficile* in the large intestine. Factors such as antibiotic use or chemotherapy disrupt the suppressive effects of the microflora and allow *C. difficile* to propagate and to secrete its toxins. This organism produces two cytotoxins, one of which, cytotoxin A, appears to be responsible for damaging the colonic mucosa, while the other, cytotoxin B, is used for diagnosis based on its cytotoxic effects in tissue culture.

Antibiotic-associated diarrhoea and colitis due to toxin-producing strains of *C. difficile* can be community-acquired or acquired in hospitals and chronic-care facilities. Recent treatment with antibiotics, especially cephalosporins and clindamycin, or chemotherapeutic agents such as methotrexate precedes the development of illness. Clinical findings range from asymptomatic carriage to fulminant colitis with perforation. Symptomatic patients have frequent, malodorous bowel movements that are not grossly bloody. Associated signs and symptoms include crampy abdominal pain, fever, and abdominal tenderness. Leucocytosis with an increase of immature neutrophil forms is often present. Complications of *C. difficile* colitis include toxic megacolon, perforation, electrolyte disturbances, and hypoalbuminaemia.

Invasive infections

There are many infections of the gastrointestinal tract that do not present with diarrhoea, but instead are manifested by a systemic illness in which constitutional symptoms and signs predominate. Enteric fever, particularly that caused by *S. typhi*, may be the most common invasive bacterial infection worldwide.

Typhoid fever

After ingestion in contaminated food or water, *S. typhi* penetrates the small bowel mucosa and makes its way rapidly to the lymphatics, the mesenteric nodes, and finally the bloodstream. Following an initial bacteraemia, the organism is sequestered in cells of the reticuloendothelial system where it multiplies and re-emerges several days later in recurrent waves of bacteraemia, an event that initiates the symptomatic phase of infection.

Typhoid fever is a febrile illness of prolonged duration, characterized by hectic fever, delirium, persistent bacteraemia, splenomegaly, abdominal pain, and a variety of systemic manifestations. Pulse–temperature dissociation is present in some patients. In approximately 50 per cent of patients, there is no change in bowel habits; in fact, constipation is more common than diarrhoea in children with typhoid fever. As a result of recurrent waves of bacteraemia, patients with typhoid fever can develop pneumonia, pyelonephritis, osteomyelitis, septic arthritis, and meningitis. Intestinal haemorrhage and perforation, the most common complications, often occur in the third week of infection or during convalescence.

While *S. typhi* is the main cause of typhoid fever, other serotypes of *Salmonella* occasionally produce a similar clinical picture, known as enteric or paratyphoid fever. These serotypes include *S. paratyphi, S. schottmülleri* (formerly *S. paratyphi B*), and *S. hirschfeldii* (formerly *S. paratyphi C*), as well as others such as *S. typhimurium.*

Parasitic infestations

Certain gastrointestinal parasites are associated with systemic signs and symptoms during the extraintestinal stages of their lifecycles. Gut infections with *Strongyloides stercoralis* manifest with vague symptoms such as abdominal pain, bloating, and diarrhoea, frequently associated with eosinophilia. During the migration of this parasite through the skin and lung, specific symptoms attributable to the local inflammatory response in these tissues may occur. Hyperinfection or disseminated strongyloidiasis develops in immunocompromised patients, especially those with HIV infection, haematological malignancies, or those treated with systemic steroids or other immunosuppressive agents. Individuals with the hyperinfection syndrome have heavy worm burdens that can lead to intestinal obstruction,

meningitis, respiratory failure, or Gram-negative bacteraemia.

Other intestinal parasites such as hookworm, *T. trichiura*, and *Schistosoma* species can cause gradual blood loss from the intestine that, in prolonged infections, can lead to clubbing, severe malnutrition, pica, stunting of growth, and congestive heart failure secondary to severe anaemia. Chronic infections with all *Schistosoma* species with the exception of *S. haematobium* can cause significant morbidity and mortality as a result of granuloma formation in the intestine and liver. The resulting hepatic fibrosis leads to portal hypertension that can eventually be complicated by splenomegaly, oesophageal varices, haematemesis, and death.

Intestinal tuberculosis

Mycobacterium tuberculosis is responsible for most cases of intestinal tuberculosis. In some developing countries, however, cases caused by *M. bovis,* an organism found in unpasteurized dairy products, still occur. The most frequent sites of intestinal involvement are the distal ileum and caecum, although any region of the gastrointestinal tract can be involved. Most patients with intestinal tuberculosis are asymptomatic. The most common complaint is chronic, non-specific abdominal pain. Weight loss, fever, diarrhoea or constipation, and blood in the stool may be present. An abdominal mass, commonly located in the right lower quadrant of the abdomen, is appreciated in about two-thirds of patients. Complications include haemorrhage, obstruction, perforation, fistula formation, and malabsorption.

Peritoneal tuberculosis results from the haematogenous spread of *M. tuberculosis* to mesenteric lymph nodes. Ascites is the most common presenting feature and is often associated with fever, lethargy, and weight loss. The ascitic fluid is notable for an elevated white blood cell count with a lymphocytic predominance, and a high albumin concentration.

Diagnosis and management of gastrointestinal infections

Diagnosis

Although there is considerable overlap in presenting signs and symptoms, nevertheless a pathophysiological approach can be used to make a presumptive aetiological diagnosis in patients with infectious diarrhoea (Table 2). By separating micro-organisms that target the upper small intestine from those that attack the large bowel, the clinician can categorize the general type of pathogen based on the initial symptoms and the type of diarrhoea. In the case of the non-inflammatory bowel pathogens, microscopy of the stool reveals no leucocytes or erythrocytes, whereas these are often abundant in the faeces of patients with invasive diarrhoeal pathogens. Several organisms including *Salmonella*, *Yersinia* spp., *V. parahaemolyticus,* and *C. difficile* produce variable findings on microscopic examination of stools. Depending on the invasiveness of the strain and the extent of colonic involvement , there can be few to many red blood cells and/or polymorphonuclear leucocytes in the stool.

A diagnostic algorithm can be used to help decide which patients should be treated symptomatically and which require further diagnostic studies and treatment (Fig. 1). Approximately 90 per cent of cases of acute diarrhoea fall into the 'no studies–no treatment' category. Because of the significant morbidity and cost associated with infectious diarrhoea, making a specific laboratory diagnosis can be useful epidemiologically, diagnostically, and therapeutically. A definitive diagnosis is achieved mainly through study of faecal specimens, using bacteriological culture, viral culture, or direct electron microscopy for viral particles, and identification of microbial antigens (viruses, bacteria, parasites, or toxins). DNA probes, polymerase chain reaction, and immunodiagnostic tests can now be used to identify several pathogens in stool specimens. Although some diseases can be diagnosed by elevations of serum antibody titres, this method is usually retrospective and often inaccurate.

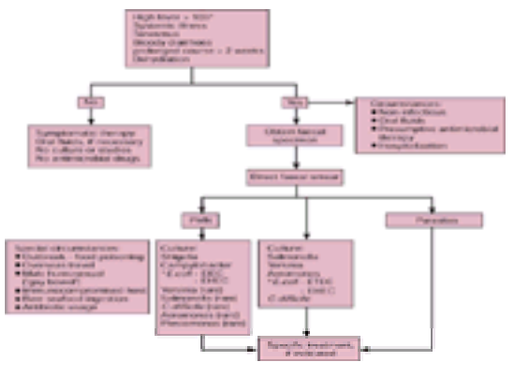

Fig. 1 Algorithm for the diagnosis and treatment of diarrhoea.

Invasive procedures such as sigmoidoscopy or upper endoscopy generally play a limited role in the diagnosis of bacterial infections of the gastrointestinal tract. If performed, proctoscopy or sigmoidoscopy of patients with colitis due to *Shigella* spp., *Salmonella* spp., and other invasive pathogens will show a diffusely ulcerated, haemorrhagic, and friable colonic mucosa. Large bowel involvement with *C. difficile* manifests as an acute inflammatory colitis with or without pseudomembranes. While the demonstration of pseudomembranes by colonoscopy can provide a rapid diagnosis, this method is relatively insensitive. Sigmoidoscopy with biopsy of the rectal mucosa is often helpful in identifying parasitic infections such as *E. histolytica* or *S. mansoni*. Endoscopy with duodenal aspirates or biopsies may help to establish the diagnosis of giardiasis, cryptosporidiosis, microsporidiosis, or strongyloidiasis. These procedures should be carried out during the evaluation of patients with chronic diarrhoea if stool cultures and examinations for ova and parasites have failed to elucidate the aetiology.

Differential diagnosis

Non-inflammatory diarrhoea

A large number of non-infectious causes of food poisoning such as heavy metals (for example, arsenic or cadmium), mushrooms (for example, *Amanita phalloides*), and other chemical substances can result in acute diarrhoea, nausea, and vomiting. Various toxin-mediated forms of shellfish or seafood poisoning, including ciguatera, scombroid, and toxic encephalopathic shellfish poisoning, can all present with nausea, vomiting, and diarrhoea as part of a constellation of symptoms. A history of recent seafood consumption and the presence of other characteristic symptoms or signs should alert the clinician to the cause.

Endocrine disorders associated with diarrhoea include thyrotoxicosis and Addison's disease. Some secretory tumours such as carcinoid, medullary tumour of the thyroid, and vasoactive intestinal peptide-secreting adenomas have watery diarrhoea as a prominent symptom. Chronic, non-bloody diarrhoea is seen in patients with coeliac disease, laxative abuse, Whipple's disease, short-gut syndrome, and pancreatic insufficiency.

Inflammatory diarrhoea

Bloody diarrhoea due to invasive enteropathogens is difficult to distinguish from that caused by inflammatory bowel disease. Two features help to distinguish dysentery from an acute attack of idiopathic ulcerative colitis: a positive culture for a pathogen and a self-limited course without relapse. However, positive cultures are encountered in only 40 to 60 per cent of reported dysentery cases. Biopsy of colonic mucosa from patients with both bacterial dysentery and ulcerative colitis show oedema, neutrophils in the lamina propria, and superficial cryptitis with preservation of the normal crypt pattern. Yet, biopsy from idiopathic ulcerative colitis also reveals signs of chronicity such as crypt distortion and plasmacytosis in the lamina propria. In clinical practice, the main diagnostic quandary is the patient with severe, acute colitis who has failed to respond to antimicrobial therapy. Presumptive treatment should include a fluoroquinolone for bacterial pathogens and metronidazole for protozoa. The decision to use other treatments, such as corticosteroids and antimetabolites, rests on the distinction between these diseases, although it may be difficult to make this decision based on culture or histopathological findings. In addition to inflammatory bowel disease, often non-infectious causes such as ischaemic colitis, acute diverticulitis, and, rarely, colon cancer can present with bloody diarrhoea.

Enteric fever

Because the initial presentation of typhoid and paratyphoid fever is pyrexia, there is a large differential diagnosis during the early stages of enteric fever. Depending on epidemiological and clinical factors, a range of infectious (for example, malaria, Gram-positive sepsis, brucellosis, occult abscess) and non-infectious (for example, rheumatological diseases and malignancy) aetiologies need to be considered. Blood cultures are an essential part of the diagnostic evaluation for enteric fever.

Management

Rehydration

Since the most devastating consequences of acute infectious diarrhoea result from fluid losses, the major goal of treatment is the replacement of fluid and electrolytes. While the intravenous route of administration has been traditionally used, oral rehydration solutions (**ORS**) have been shown to be equally effective physiologically and logistically more practical and less costly to administer, especially in developing countries. ORS is the treatment of choice for mild-to-moderate diarrhoea in both children and adults, providing vomiting is not a major feature of the gastrointestinal infection. ORS can also be used in severely dehydrated patients after initial parenteral rehydration.

Although there is no doubt about the value of ORS in treating dehydrating diarrhoea, the optimal concentration of sodium that should be used remains in dispute, particularly in regard to the treatment of mild-to-moderate diarrhoea in well-nourished children in industrialized countries. The high concentration of sodium (90 mmol) in the standard World Health Organization ORS formulation may cause hypernatraemia and even seizures in children with non-cholera watery diarrhoea. Consequently, lower concentrations of sodium and a reduced osmolarity solution have been found to be effective for rehydration and not to be associated with any serious adverse clinical events. The substitution of starch derived from rice or cereals for glucose in ORS has been another approach. Rice-based salt solutions produce lower stool losses, a shorter duration of diarrhoea, and greater fluid and electrolyte absorption than do glucose-based solutions in treating childhood and adult diarrhoea.

Diet

The traditional approach to an acute diarrhoeal illness, dietary abstinence, restricts the intake of necessary calories, fluids, and electrolytes. During an acute attack, the patient often finds it more comfortable to avoid spicy, high-fat, and high-fibre foods, all of which can increase stool volume and intestinal motility. Although giving the bowel a rest provides symptomatic relief, continued oral intake of fluids and foods is critical for both rehydration and the prevention of malnutrition. In children, it is particularly important to restart feeding as soon as the child is willing to accept oral intake.

Because certain foods and fluids can increase intestinal motility, it is wise to avoid fluids such as coffee, tea, cocoa, and alcoholic beverages. Ingestion of milk and dairy products can potentiate fluid secretion and increase stool volume. Besides the oral rehydration therapy outlined above, acceptable beverages for mildly dehydrated adults include fruit juices and various bottled soft drinks. Carbonated drinks should be allowed to 'de-fizz' by letting them stand in a glass before ingestion. Soft, easily digestible foods are generally acceptable to the patient with acute diarrhoea.

Antimicrobial therapy

Since most patients with infectious diarrhoea, even those with a recognized pathogen, have a mild, self-limited course, neither a stool culture nor specific treatment is required for such cases (Fig. 1). For more severe cases, however, empirical antimicrobial therapy should be instituted, pending the results of stool and blood cultures. Gastrointestinal infections likely to respond to antibiotic treatment include cholera, giardiasis, cyclosporiasis, shigellosis, *E. coli* diarrhoea in infants, symptomatic travellers' diarrhoea, *C. difficile* diarrhoea, and typhoid fever. The choice of antimicrobial drug should be based on *in vitro* sensitivity patterns, which vary from region to region. A fluoroquinolone antibiotic is a good choice for empirical therapy, since these agents have broad-spectrum activity against virtually all bacterial pathogens responsible for acute infectious diarrhoea (except *C. difficile*) and resistance to this drug remains limited in most parts of the world. In patients with severe community-acquired diarrhoea—characterized by more than four stools per day lasting for at least 3 days or more with at least one associated symptom such as fever, abdominal pain, or vomiting—there is a high likelihood of isolating a bacterial pathogen. In this setting, a short course of a fluoroquinolone, namely 1 to 3 days' duration, will generally provide prompt relief with a low risk of adverse effects. Fluoroquinolones will not be effective for parasitic infections—specific antiparasitic drugs should be prescribed after identification of the offending pathogen in stool smears.

There are conflicting reports regarding the efficacy of antimicrobial drugs in several important infections, such as those caused by *Campylobacter* spp., and insufficient data for infections caused by *Yersinia and Aeromonas* spp., vibrios, and several forms of *E. coli.* In cases of EHEC, there is evidence that antibiotics are not helpful and may even be harmful.

The duration of antimicrobial therapy has not been clearly defined. While courses of anywhere from 3 to 10 days of treatment have been recommended, there are several studies that included severe forms of diarrhoea which suggested that a single dose is as effective as more prolonged therapy. For example, single-dose fluoroquinolone therapy is highly effective for infections due to *V. cholerae, V. parahaemolyticus*, and most *Shigella* species. On the other hand, short-course treatment of salmonella gastroenteritis with fleroxacin has not been found to be clinically beneficial. When treatment is indicated, a number of studies have shown that the combination of an antimicrobial drug and an antimotility drug provides the most rapid relief of diarrhoea.

Antidiarrhoeal agents (Table 3)

Antimotility drugs are particularly useful in controlling moderate-to-severe diarrhoea. These agents disrupt propulsive motility by decreasing jejunal motor activity. Opiates may decrease fluid secretion, enhance mucosal absorption, and increase rectal sphincter tone. The overall effect is to normalize fluid transport, slow transit time, reduce fluid losses, and ameliorate abdominal cramping.

Loperamide is the best agent because it does not carry a risk of habituation or depression of the respiratory centre. Treatment with loperamide produces rapid improvement, often within the first day of therapy. Although there has been a long-standing concern that antimotility agents might exacerbate cases of dysentery, this has largely been dispelled by clinical experience. Patients with shigellosis, even *S. dysenteriae* type 1, have been treated with loperamide alone and have had a normal resolution of symptoms without evidence of prolonging the illness or delaying excretion of the pathogen. However, as a general rule, antimotility drugs should not be used in patients with acute severe colitis, whether infectious or non-infectious in origin.

Bismuth subsalicylate (**BSS**), an insoluble complex of trivalent bismuth and salicylate, is effective in treating mild-to-moderate forms of diarrhoea. Bismuth possesses antimicrobial properties, while the salicylate moiety has antisecretory properties. In trials of diarrhoea among travellers in Mexico and West Africa, BSS reduced the frequency of diarrhoea significantly relative to placebo, but results were generally better when a high dose (for example, 4.2 g per day) was used. A number of studies have shown that the combination of an antimicrobial drug and an antimotility drug provides the most rapid relief of diarrhoea.

Prevention

Strict adherence to food and water precautions as outlined above will help travellers to less developed areas of the world to decrease their risk of acquiring gastrointestinal infections. Parasitic infections, such as strongyloidiasis and hookworms, can be avoided by the use of footwear. Avoiding contact with fresh water such as rivers and lakes in endemic areas serves to prevent schistosomiasis.

Immunization represents an ideal way to prevent certain bacterial and viral diseases, but has not yet proved successful for combating many gastrointestinal pathogens. The cholera vaccine that has been available for decades suffers from low efficacy, a moderate risk of side-effects, and a short duration of action. Newer oral cholera vaccines, including inactivated and live-attenuated forms, appear to be more promising. Immunization has been partially effective for the prevention of typhoid fever, especially in endemic areas. Although the efficacy of the currently available typhoid vaccines has not been determined in persons from industrialized regions, these vaccines are widely used for the prevention of typhoid fever in travellers to developing countries.

Further reading

Avery ME, Snyder JD (1990). Oral therapy for acute diarrhea: the underused simple solution. *New England Journal of Medicine* **323**, 891–4.

Acheson DWK, Keusch GT (1995). *Shigella* and enteroinvasive *Escherichia coli*. In: Blaser MJ, *et al*. eds. *Infections of the gastrointestinal tract*, pp 763–84. Raven Press, New York.

Blacklow NR, Greenberg HB (1991). Viral gastroenteritis. *New England Journal of Medicine* **325**, 252–64.

DuPont HL, Capsuto EG (1996). Persistent diarrhea in travellers. *Clinical Infectious Diseases* **22**, 124–8.

Echeverria P, Sethabutr O, Serichantalergs O (1993). Modern diagnosis (with molecular tests) of acute infectious diarrhea. *Gastroenterology Clinics of North America* **22**, 661–82.

Gerding DN, *et al*. (1995). *Clostridium difficile*-associated diarrhea and colitis. *Infection Control and Hospital Epidemiology* **16**, 459–77.

Gorbach SL (1997). Treating diarrhoea. *British Medical Journal* **314**, 1776–7.

Gorbach SL, Edelman R, eds. (1986). Travellers' diarrhea: National Institutes of Health Consensus Development Conference. *Reviews of Infectious Diseases* **8**(Suppl. 2), S109–S233.

Hamer DH, Cash RA (1999). Cholera and enterotoxigenic *Escherichia coli.* In: Armstrong D, Cohen J, eds. *Infectious diseases*, pp 22.1–22.4. Harcourt Brace, London.

Hamer DH, Gorbach SL (1998). Use of the quinolones for the treatment and prophylaxis of bacterial infections. In: Andriole VT, ed. *The quinolones*, 2nd edn, pp 267–85. Academic Press, San Diego.

Mishu Allos B, Blaser MJ (1995). *Campylobacter jejuni* and the expanding spectrum of related infections. *Clinical Infectious Diseases* **20**,1092–101.

Simon GL, Gorbach SL (1995). Normal alimentary tract microflora. In: Blaser MJ, *et al*. eds. *Infections of the gastrointestinal tract*, pp 53–69. Raven Press, New York.

Su C, Brandt LJ (1995). *Escherichia coli* 0157:H7 infection in humans. *Annals of Internal Medicine* **123**, 698–714.

14.18.1 The structure and function of the liver, biliary tract, and pancreas

A. E. S. Gimson

The liver and biliary tract

The liver weighs 1.2 to 1.5 kg and has a highly vascular architecture. The classic descriptions of liver anatomy demonstrating the complexity of different parenchymal and non-parenchymal elements have a long history, but only recently have they been united with an increasing understanding of the intricate functional organization and physiological compartmentalization of liver structure. This has had a profound effect on our understanding of the control of physiological processes and the development of liver surgery. A grasp of the hepatic anatomy is key to an appreciation of these complex functional arrangements.

Morphological anatomy

This describes the classic structure of the liver into two lobes, right and left, and the accompanying vascular structures, lymphatics, and biliary tract.

The liver, situated in the right upper quadrant of the abdomen, is covered by Glisson's capsule, a visceral continuation of the peritoneum. Three ligaments attach to surrounding structures—the falciform ligament anterior and superiorly, and the two posterior triangular ligaments which enclose the retrohepatic vena cava and the small bare area of the liver. Inferiorly Glisson's capsule attaches to the lesser curve of the stomach and at the hepatic hilus encases the hepatic pedicle consisting of hepatic artery, portal vein, and common hepatic bile duct.

Hepatic lobes

The two major lobes, right and left, and two accessory lobes, quadrate and caudate, are defined by points of surface anatomy (Fig. 1(a)). The larger right lobe comprises the dome of the liver under the diaphragm and is limited anteriorly and medially by the falciform ligament and posteriorly by the right border of the inferior vena cava. The quadrate lobe inferiorly abuts on to the antrum of the stomach and first part of the duodenum and is bordered by the posterior transverse hilar fissure, the gallbladder fossa laterally, and the umbilical fissure medially. The caudate lobe lies posterior and superior to the quadrate lobe limited by the vena cava and the ligamentum venosum. Finally, the left lobe has the umbilical fissure medially and the falciform ligament anteriorly.

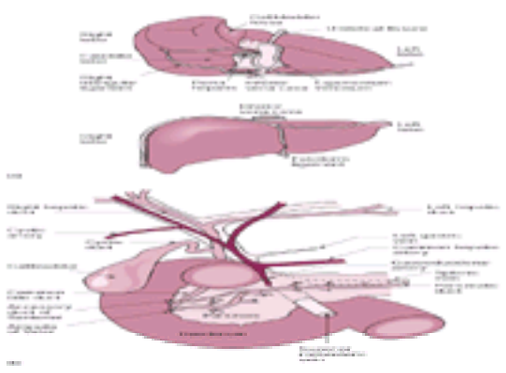

Fig. 1 (a) Lobar anatomy and relations. (b) Hilar, portal biliary tract, and pancreatic anatomy.

Vascular anatomy

The portal vein, hepatic duct, and hepatic artery form the hepatic pedicle with the bile duct anterior in the free edge of the lesser omentum and the portal vein posteriorly (Fig. 1(b)). The latter is formed by the confluence of the superior mesenteric vein and the splenic veins running posteriorly in the pedicle, dividing into left and right branches to supply each lobe. The left gastric vein also drains into the portal vein and may, in the presence of portal hypertension, be a major feeding vessel for gastro-oesophageal varices.

The hepatic artery arises from the coeliac axis as the common hepatic artery before dividing into a gastroduodenal and the main hepatic artery. There are several common anatomical variants of the arterial supply of the liver, which are of no functional significance but which are of importance in liver transplantation and during surgical resection. The standard division into single left and right hepatic arteries is present in approximately 70 per cent of cases (Fig. 1(b)), but common variants include: a separate second right hepatic artery (10 per cent), a separate right and left hepatic arteries (8 per cent), and origin of the main hepatic artery off the superior mesenteric artery (2.5 per cent). Variants of the left hepatic arterial supply also occur with a separate left hepatic artery arising from the left gastric artery in 10 per cent of cases.

Venous drainage of the liver is through the three main hepatic veins, right, left, and middle, the latter two coalescing before joining the inferior vena cava. The caudate lobe drains separately through an array of small spigelian veins directly into the inferior vena cava. The functional anatomy of the liver (see below) describes the relationship between the main divisions of the portal vein and their draining hepatic veins running in the right, left, and main scissures (Fig. 2).

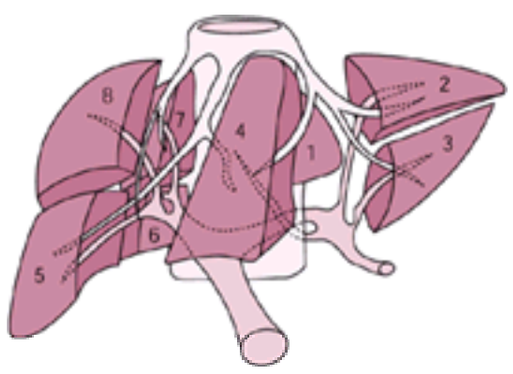

Fig. 2 Functional anatomy of the liver with Couinaud's segments.

Biliary anatomy

Biliary canaliculi drain into left and right hepatic bile ducts forming the common hepatic duct until entry of the cystic duct after which it is designated the common bile duct and has a diameter of less than 8 mm. The left hepatic duct follows a nearly horizontal course, partially extrahepatic. Anatomical variants are again quite frequent, and are surgically important, the most common being drainage of the cystic duct directly into the right hepatic duct. The common bile duct passes behind the first part of the duodenum, through pancreatic tissue to the ampulla of Vater joining drainage of the pancreatic duct (Fig. 1(b)). The gallbladder lies in a shallow depression in the underside of the liver, may contain up to 50 ml of bile, and is connected to the cystic duct with a spiral valve.

Lymphatics

The liver has a high blood flow and a highly permeable microcirculation. The consequent production of interstitial fluid, intrahepatic lymph, is formed in the perisinusoidal space of Disse between the hepatocytes and sinusoidal lining endothelium. Lymphatic vessels drain via the portal tracts, closely applied to the hepatic arterial branches, to the hilum and thence to the thoracic duct. A smaller proportion drains with the hepatic veins and some interstitial fluid drains through Glisson's capsule into the peritoneum. Lymph flow acts to drain from the liver that interstitial fluid and protein that forms inevitably through microvascular filtration. The lymph flow rate in mammalian liver is approximately 0.5 ml/kg of liver per minute making up 25 to 50 per cent of thoracic duct lymph flow and may be increased either by elevated microvascular pressure (hydrostatic pressure) through increased hepatic venous pressure or increased inflow pressure, or by reduced transcapillary oncotic pressure.

Nervous system

Both sympathetic and parasympathetic efferent innervation of the liver are described, an anterior plexus around the hepatic artery and posterior around the portal vein. Sympathetic stimulation increases glucose release and glycogenolysis, and reduces oxygen consumption, ammonia uptake, and bile formation. Hepatic vascular resistance also rises as does portal pressure and there is rapid expulsion of blood out of the liver into the systemic circulation. An intrinsic nervous system with a wide variety of neurotransmitters including noradrenaline, prostanoids, neuropeptide Y, substance P, and vasoactive intestinal peptide is closely located to smooth muscle cells, fibroblasts, endothelial lining cells, and biliary epithelium within the liver and may be involved in chemoreception and osmoreception.

Extrinsic nervous regulation of hepatic physiological processes seems to be of minor importance as there is no apparent impairment of liver metabolism or bile formation following orthotopic liver transplantation. It may be more relevant during pathophysiological stress: the existence of a hepatorenal reflex is patients with cirrhosis has been postulated whereby an increase in sinusoidal pressure is associated with increased efferent renal sympathetic activity and reduced renal blood flow. In animal models of chronic liver disease, the metabolic consequences of sympathetic nerve stimulation are impaired but the haemodynamic responses exaggerated.

Functional anatomy

Following the initial descriptions by Cantlie in 1898, there has been an increasing appreciation of the importance of the functional anatomy of the liver, the culmination of which was the description by Couinaud of the present eight liver segments that underpins all modern hepatic surgery. Each segment is a complete functional unit with a single portal pedicle and a hepatic vein (Fig. 2). There are four portal pedicles, two for each lobe, each supplying a sector of the liver, divided from each other by the three hepatic veins lying in a right, middle, and left scissure. This separates the liver into a right and left liver, different from lobes, with independent vascular supply and biliary drainage. Within each sector of the liver there are further subdivisions into segments. The caudate lobe (segment 1) has its own venous drainage, manifest during the Budd–Chiari syndrome with thrombosis of hepatic veins when all venous drainage attempts to pass through this segment with consequent lobar hypertrophy.

The left liver consists of the left posterior sector of segment 2 alone, and a left anterior sector of segment 3 medially and segment 4 laterally separated by the umbilical fissure. The right liver comprises a posterior sector of segment 7 superiorly and segment 6 inferiorly and an anterior sector of segment 5 inferiorly and segment 8, being most of the dome of the liver, superiorly (Fig. 2).

Structural organization

Within the functional segments of the liver the structural unit is the hepatic lobule, a polyhedron (2 mm by 0.7 mm) surrounded by four to six portal tracts containing hepatic arterial and portal venous branches from which blood perfuses through sinusoids, surrounded by walls of hepatocytes that are a single cell thick and lined by specialized endothelial cells with 'windows' (fenestrae), to the centrilobular region and the central hepatic veins (Fig. 3).

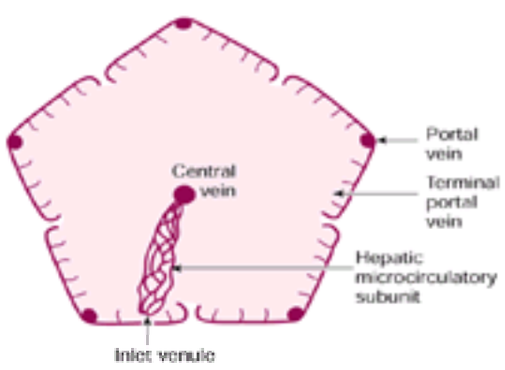

Fig. 3 Hexagonal lobule with portal venous branches and hepatic microcirculatory subunit—sinusoids.

The portal vein branches give off numerous terminal portal venules that run around the lobules in the interlobular septa accompanied by arterioles and bile ductules, and subsequently branch into inlet venules which each supply a hepatic microcirculatory subunit consisting at the base of numerous interconnected sinusoids and, at the apex, the central vein (Fig. 3).

Sinusoids

Sinusoids are specialized capillaries without a basement membrane and lined with endothelial lining cells through which proteins of low molecular weight may percolate into the space of Disse. The sinusoidal membrane of the surrounding hepatocytes is covered by microvilli that increase the surface area sixfold (Fig. 4). Within the sinusoids, Kupffer cells and liver-associated lymphocytes may be found, and within the space of Disse, the hepatic stellate cells (also called Ito, fat storage, or perisinusoidal cells), which respectively make up 2, 0.2, and 1.4 per cent of the lobular parenchyma (Table 1).

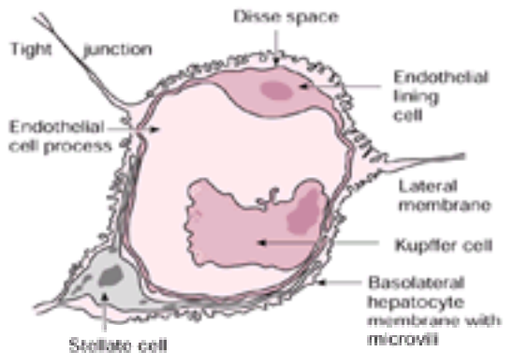

Fig. 4 Hepatic sinusoid, sinusoidal cells, and functional spaces.

Biliary canaliculi

Bile secreted through the canalicular membrane of the hepatocyte collects in biliary canaliculi, which pass around hepatocytes until draining through the short canal of Hering into the bile ductule. Cholangioles are lined by three or four cells that eventually become cuboidal epithelium.

The volume and flow rate of bile are low; secretion into the duodenum is controlled by gallbladder contraction and sphincter of Oddi tone. Agents that cause gallbladder contraction, including cholecystokinin, secretin, and motilin, also relax the sphincter of Oddi (Table 2). Factors modulating biliary motility have received increased attention recently with the realization that the syndrome of biliary dysmotility may be the cause of biliary-type pain in some cases. Changes in gallbladder motility may also be important in gallstone pathogenesis.

Cellular elements

Hepatocytes are arranged in unicellular plates (Remak's plates) that branch and divide around sinusoids, and are covered by specific membranes at each surface: sinusoidal (70 per cent of surface area) for exchange of material between the Disse space and intracellular compartment (endo- and exocytosis); canalicular membrane (15 per cent) for exchange with the smallest of biliary canaliculi or hemicanals; and lateral membrane (15 per cent) separated from the former by tight junctions and involved in intercellular transport between hepatocytes. There is abundant smooth and rough endoplasmic reticulum, numerous mitochondria, and glycogen. There is an extensive cytoskeleton. Other cells making up 6 per cent of all parenchyma include sinusoidal-lining endothelial cells, Kupffer cells, hepatic stellate cells (Ito cells, fat-storing cells), and pit cells (intrahepatic lymphocytes) (Table 1). These cells each differ in morphology, patterns of function, reactions to stimuli and disease, and expression of surface molecules and receptors. Interplay between these cells is critical, with communication via tight junctions allowing complex modulation of hepatocyte growth and function by sinusoidal lining cells. Parenchymal cells may clear mediators, including cytokines, secreted by endothelial lining and Kupffer cells. Waves of cellular activity may pass down the length of sinusoids. Importantly some cells show heterogeneity of function relative to their zonal location. Periportal hepatocytes differ from perivenous cells in both the direction of carbohydrate metabolism and ammonia/glutamine synthesis. Ito cells show zonal differences in desmin and cytokeratin staining, vitamin A storage, and a-smooth muscle actin.

Endothelial lining cells

These cells are central to the processes that control entry and exit trafficking of molecules from the sinusoidal flow into the Disse space. Fenestrae with a diameter of 100 nm, occupying up to 8 per cent of the sinusoidal surface, act as a physical barrier to access of parenchymal cells by large molecules including lipids, cholesterol, vitamin A, and possibly some viruses. Endothelial cells also possess numerous specialized endocytotic mechanisms, some linked to specific receptors including mannose, transferrin, caeruloplasmin, modified high-density lipoprotein (**HDL**), low-density lipoprotein (**LDL**), glucosaminoglycans, and hyaluronic acid. Non-specific endocytosis of molecules and small particles up to 0.1 µm also occurs. Endothelial cells are also capable of expressing a range of surface adhesion molecules including E- and P-selectins, intercellular adhesion molecule 1 (**ICAM-1**), and lymphocyte function-associated antigen-4 (**LFA-4**) that enhance polymorphonuclear leucocyte and lymphocyte adherence, activation, and migration towards sites of inflammation.

Kupffer cells

These cells represent part of the mononuclear phagocyte system and are adherent to the sinusoidal surface of endothelial lining cells, predominantly in a periportal distribution. Covered with numerous microvilli and with a number of intracytoplasmic vesicles, their main function is to phagocytose a range of particulate material including cellular debris, senescent red blood cells, parasites, bacteria, endotoxin, and tumour cells. Phagocytosis is via a range of mechanisms including coated pits, macropinocytotic vesicles, and phagosomes aided by opsonization of particles by fibronectin or opsonin. Kupffer cells may be activated by molecules including *Escherichia coli* endotoxin, interferon-g, tumour necrosis factor-a (**TNF-a**), and arachidonic acid as well as zymosan and phorbol myristate to release a range of inflammatory mediators that include oxygen radical species, nitric oxide, proteases, TNF-a, interleukins 1, 6, and 10 (**IL-1**, **-6**, **-10**), transforming growth factor-b (**TGF-b**), prostanoids, and interferon-a and -g. Some of these may act in an autocrine or paracrine loop to further activate other Kupffer cells. These inflammatory products have a range of effects including significant modulation of parenchymal cell function (downregulation of albumin synthesis and upregulation of acute-phase protein gene expression), and induction of adherence of polymorphonuclear leucocytes and lymphocytes to endothelial lining cells due to enhanced expression of endothelial adhesion molecules.

Hepatic stellate cells

Stellate cells (Ito cells, fat-storing cells) have a similar morphology to fibroblasts with the addition of fat droplets, and are located within the Disse space. A fine branching array of cytoplasmic processes circle sinusoids under the endothelial cells. Stellate cells contain most of the body's stores of vitamin A. Retinoids are taken up from chylomicrons by specific receptors on hepatocytes and stellate cells and stored within the latter. These cells are central to the process of hepatic fibrogenesis, responding to mediators released by parenchymal and Kupffer cells, causing transformation into myofibroblasts. TGF-b initiates this process, stimulating production by the transformed stellate cell of extracellular matrix products (collagen type I, III, and IV, fibronectin, laminin, chondroitin sulphate, and hyaluronic acid) in addition to products for matrix degradation (collagenase, metalloproteinase, and its inhibitor TIMP-1). Activation of stellate cells is also an important mechanism for control of sinusoidal perfusion, through cytoskeletal actin within branching cellular processes beneath the endothelium.

Pit cells

Similar to large granular lymphocytes and located in clefts within endothelial lining cells, pit cells have natural killer cell properties with spontaneous activity against tumour cells in the absence of prior activation. They may also play a role in hepatic regeneration

Physiological processes

Hepatic blood flow

The liver receives approximately 25 per cent of cardiac output, one-third from the hepatic artery and two-thirds from the portal vein with a plasma flow at rest of 1600 ml/min in women and 1800 ml/min in men. Hepatic blood flow increases after feeding and with expiration and decreases with standing, inspiration, and sleep. In contrast to other organs, metabolic autoregulation of blood flow is not observed. Changes in hepatic oxygen consumption do not seem to control hepatic blood flow. Vascular autoregulation of hepatic arterial blood flow mediated by adenosine is present, but may not be of great physiological importance. Hepatic arterial resistance increases with increasing hepatic venous pressure due to a stepwise myogenic response in the hepatic artery to increased pressure. There is an important reciprocity between portal venous and hepatic arterial flow with a reduction in portal venous input being associated with significant compensatory decrease in hepatic arterial resistance and rise in arterial flow. The mechanism for this relationship is unproven but may be due to adenosine-mediated arterial vasodilatation.

The portal venous system is passive, without pressure-dependent autoregulation, and the major physiological factors controlling flow are those modulating supply to

the intestines and spleen. The sites of portal venous resistance are not fully defined in humans but may be at sinusoidal or post-sinusoidal levels. The significant capacitance of the hepatic circulation, with blood comprising up to 20 per cent of liver volume, is reflected in the important role of the liver and splanchnic circulation in acting as a blood reservoir. Sympathetic nerve stimulation may reduce hepatic blood volume by up to 50 per cent.

Sinusoidal perfusion

Blood pressure in sinusoids ranges from 4.8 to 1.7 mmHg, with flows of 270 to 410 ml/s. There is likely to be considerable heterogeneity of the unidirectional sinusoidal flow, control for which can be considered as either passive (haemodynamic) or active. Passive control mechanisms include: (i) the arterial input pressure and flow at the level of the arteriosinous twig at the origin of the sinusoid; and (ii) changes in right atrial pressure, central venous pressure, and hepatic venous pressure that are transmitted to the sinusoid from the centrilobular veins. Active control mechanisms include: (i) the presence of 'functional' sphincters at the inlet and outlet of the sinusoid due to indentations by the cell bodies of sinusoidal lining cells, which under different physiological stimuli may change dimension and alter sinusoidal perfusion; (ii) plugging by leucocytes, which are less compressible than erythrocytes and may under physiological stimuli adhere to endothelial lining cells; (iii) activation of Kupffer cells within sinusoids and release of other vasoactive mediators including nitric oxide, cytokines, and prostanoids; and (iv) transformation of hepatic stellate cells into activated contractile myofibroblasts that constrict the sinusoidal lumen. Sinusoidal flow will also affect the transendothelial traffic into and out of the Disse space by the processes of forced sieving and endothelial massage that may affect, respectively, the passage of lipoprotein particles and the appropriate mixing of the interstitial fluid. Therefore, sinusoidal flow is likely to have a profound effect on numerous hepatic metabolic functions and clearance of xenobiotics.

Bile formation

The formation of bile by hepatocytes and its modification by bile ductular epithelium serves many functions (Table 3). In humans the daily production of 600 ml of bile is made up of 75 per cent of canalicular origin and 25 per cent from ductules. Bile is formed by osmotic filtration, with the secretion of the two primary bile salt anions, taurine and glycine conjugates of cholic acid and chenodeoxycholic acid, across the canalicular membrane by an active transport mechanism against a concentration gradient of 5000:1 (Fig. 5). Negatively charged intercellular tight junctions prevent back diffusion of these anions, allowing the selective passage of cations, predominantly sodium, and to a smaller extent potassium, calcium, and magnesium, followed by the passive transit of water, transcellularly or between cells. The resulting bile salt-dependent bile flow makes up 50 per cent of canalicular bile flow, with the remaining bile salt-independent flow resulting from the active secretion of bicarbonate and glutathione.

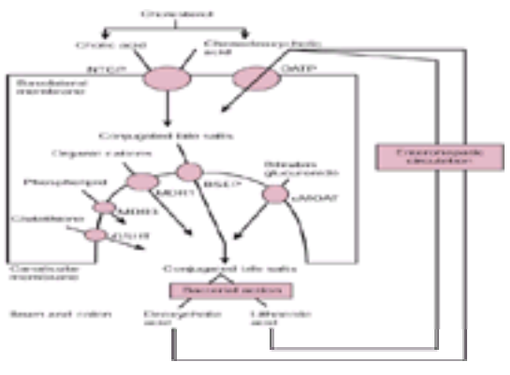

Fig. 5 Bile salt metabolism and enterohepatic pathway. NTCP, Na-taurocholate cotransporters; conjugated bile salt uptake from portal blood. OATP, organic anion transporter; bile salt, organic anion, and amphipathic solutes uptake. BSEP, bile salt export pump; ATP-dependent bile export—bile salt-dependent bile flow. MDR1, multidrug resistance-1 P glycoprotein; organic cation, xenobiotic export. MDR3, multidrug resistance-3 P glycoprotein; translocation of phosphatidylcholine. cMOAT, multispecific organic anion transporters; bilirubin glucuronide export; bile salt-independent bile flow. GSHT, glutathione transporter; gluthathione transport independent of bile flow.

Bile in biliary ductules is further modified by reabsorption of glucose, amino acids, and bile salts, as well as active secretion. Reabsorption of bile salts, the cholehepatic shunt pathway, occurs after their protonation in bile with the generation of further bicarbonate into bile stimulating bile flow. Active secretion of bicarbonate and chloride within ductules is mediated by the secretin receptor and the cystic fibrosis transmembrane receptor. Gallbladder epithelium further modifies and concentrates bile by an active anion transport process.

Bile salt conjugates secreted from hepatocytes into bile are deconjugated in the jejeunum and ileum with reabsorption and reuptake by the liver—this enterohepatic circulation conserves bile acids and maintains their high concentration within bile. The 5 per cent of bile acids passing through the ileocaecal valve are fully deconjugated by colonic bacteria and reabsorbed as the secondary bile acids deoxycholic acid and lithocolic acid, which are in turn secreted as taurine and glycine conjugates.

Metabolic processes

Hepatic metabolic processes have a central role in protein, carbohydrate, and lipid metabolism and fuel economy, orchestrating a diverse interplay between central splanchnic and peripheral organs. Interruption to these processes results in the major metabolic consequences of acute and chronic liver disease. Modulation of these metabolic processes can occur at a number of levels. Transport of molecules across membranes and through cells is an important control mechanism as are rate-limiting enzyme levels, controlled at a number of transcriptional and translational points. There is important zonal heterogeneity of hepatocyte function, with periportal zone 1 cells with a higher oxidative capacity and larger mitochondria involved in gluconeogenesis, b-oxidation of fatty acids, amino acid catabolism, ureagenesis, cholesterol synthesis, and bile secretion, whereas perivenular cells are more involved with glycolysis, lipogenesis, ammonia clearance with glutamine synthesis, detoxification, and biotransformation.

Bilirubin metabolism (Fig. 6) (see Chapter 14.19.3)

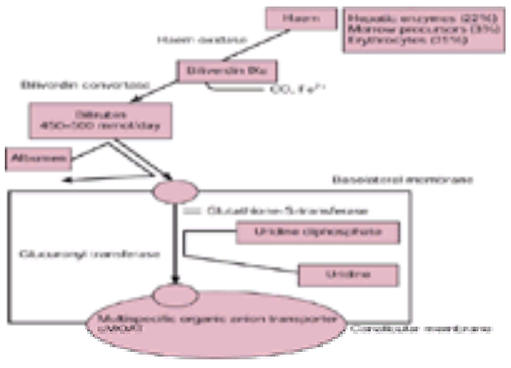

Fig. 6 Metabolism of haem and bilirubin with clearance through canalicular membrane to bile.

The first step in the production of bilirubin is the formation of biliverdin IXa by the action of haem oxidase on haem-containing proteins including catalases, cytochromes as well as haemoglobin in senescent red cells, with the release of carbon monoxide and Fe^{2+}. Biliverdin convertase within the cytosol reduces biliverdin to unconjugated bilirubin. Both biliverdin convertase and haem oxidase are predominantly found within reticuloendothelial cells.

Bilirubin is transported within plasma bound with high affinity to albumin. A few substances may displace bilirubin from albumin including sulphonamides and fatty acids. Unbound bilirubin, which is insoluble in water, is only present in nanogram quantities but may cause significant cellular toxicity in neonates and in the Crigler–Najjar syndrome.

Bilirubin uptake by hepatocytes occurs via an organic anion-binding protein receptor. Within the hepatocyte the unbound bilirubin is transported by organelles and a number of transport proteins including glutathione- *S*-transferase (ligandin) to the endoplasmic reticulum. This reduces back diffusion into sinusoids of the lipid-soluble unbound bilirubin. Glucuronidation to the mono- and diglucuronides renders bilirubin water soluble. Secretion across the canalicular membrane occurs at the canalicular multispecific membrane organic anion transporter.

Bile salt metabolism (Fig. 5)

In addition to their role in digestion, bile acids are the principal mechanism for clearance and metabolism of cholesterol, which acts as a substrate for their synthesis and in turn promotes biliary cholesterol secretion as lamellar vesicles. The first step in bile acid synthesis is rate limiting and involves cholesterol 7a-hydroxylase. Transcriptional control of the cholesterol 7a-hydroxylase gene has been demonstrated with thyroxine and glucocorticoids increasing, and glucagon decreasing, gene expression. Preformed (non-dietary) cholesterol and bile acids may also control this enzyme. The close association between bile acid and cholesterol metabolism is reflected in the often parallel activation of 7a-hydroxylase and HMG-CoA reductase, which is of critical importance in bile acid synthesis. The two major bile acids, cholic acid (60 per cent of bile acid pool) and chenodeoxycholic acid are secreted into bile as taurine and glycine conjugates.

Carbohydrate metabolism (Fig. 7)

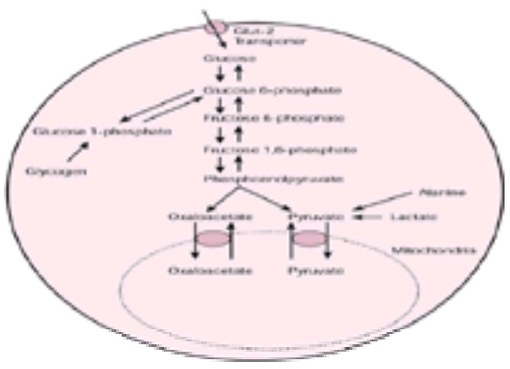

Fig. 7 Carbohydrate metabolism and pathways for glycolysis and glycogenesis.

The liver has a central role in maintaining blood glucose within a narrow margin. During fasting, hepatic glucose release is contributed to by both glycogenolysis (33 per cent) and gluconeogenesis (67 per cent) from lactate, pyruvate, glycerol, and the glucogenic amino acids alanine and glutamine. This process is regulated by at least four levels: (i) hormonal control, with glucagon accounting for up to two-thirds of basal fasted glucose output, and cortisol, growth hormone, and catecholamines also contributing; (ii) the supply of substrates, fatty acids, lactate, pyruvate, and amino acids for hepatic gluconeogenesis; (iii) metabolic regulation of hepatic enzyme activity; and (iv) the degree of hepatocellular hydration. The direction of gluconeogenesis or glycogenolysis is controlled at the level of three paired enzyme cycles—glucose/glucose-6-phosphate, fructose-6-phosphate/fructose-1,6-bisphosphate, and pyruvate/phosphoenolpyruvate. In contrast, after a glucose load, insulin suppresses hepatic glucose release and activates glucose synthetase, whilst autoregulation of hepatic glucose extraction by glucose itself within the portal venous circulation is an important factor in controlling the distribution of the load between liver and peripheral tissues.

Amino acid and ammonia metabolism

The liver is the most important organ in controlling the plasma concentration of amino acids. During prolonged starvation, hepatic proteolysis stimulated by glucagon increases splanchnic export of amino acids, whereas during the post-prandial absorptive state, amino acid uptake is significantly increased. The gluconeogenic amino acids are preferentially extracted and metabolized, whereas the branch-chain amino acids valine, leucine, and isoleucine are only cleared in the liver for protein synthesis and are catabolized in the muscle. During sepsis and under the influence of cytokines IL-1, IL-6, and TNF-a, the liver may significantly enhance gluconeogenesis and protein synthesis of acute-phase reactants (C-reactive protein, serum amyloid A).

The liver has a critical role in clearing portal venous ammonia generated within the gut lumen, by both formation of carbamoyl phosphate and entry into the urea cycle in periportal hepatocytes, and glutamine synthetase-driven glutamine synthesis in perivenous hepatocytes.

Protein synthesis

Most circulating plasma proteins with the exception of immunoglobulins and von Willebrand factor are produced by hepatocytes. The major controlling factors for this constitutive protein secretion are substrate delivery and the degree of hydration of hepatocytes. Acute-phase protein secretion is also specifically controlled by cytokines with a reciprocal relationship to albumin and other carrier protein synthesis.

Lipid and lipoprotein metabolism (Fig. 8) (see Chapter 11.6)

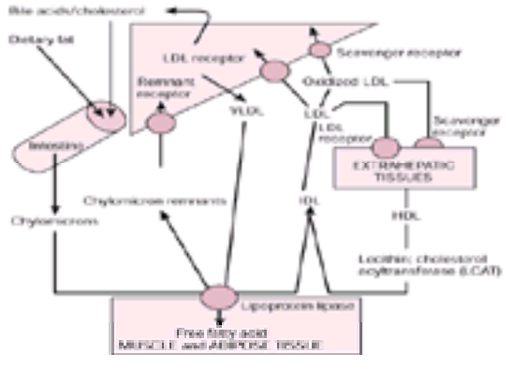

Fig. 8 Lipoprotein metabolism.

Plasma lipoproteins are particles with an outer layer of cholesterol, phospholipids, and apoproteins and an inner core of cholesterol esters and triglycerides. The various lipoproteins differ in the relative proportions of these elements. Dietary derived chylomicrons, consisting of more than 90 per cent triglyceride, are processed within muscle and adipose tissue by lipoprotein lipase, extracting free fatty acids and the remnant, enriched in cholesterol, are extracted by the liver—an exogenous lipid pathway. During carbohydrate feeding, free fatty acids formed within the liver are exported as very-low-density lipoprotein (**VLDL**) and taken up by muscle and adipose tissue with extraction of free fatty acids, leaving intermediate-density lipoprotein and subsequently low-density lipoprotein (LDL). Specific LDL receptors on hepatocytes or scavenger receptors on Kupffer cells remove LDL where cholesterol may be utilized for bile salt metabolism or excreted into bile. Peripheral LDL receptors in extrahepatic tissues also extract cholesterol. Export of cholesterol from peripheral tissues in high-density lipoprotein is modified in plasma by lecithin;

cholesterol acyltransferase (LCAT) and LDL is formed for further recirculation.

Pancreas structure and function

A retroperitoneal organ receiving arterial supply from splenic, superior mesenteric, and gastroduodenal arteries, the pancreas is composed of an exocrine portion centred on acini producing digestive enzymes draining through a ductal system into the duodenum, and the islets of Langerhans which make up 1 to 2 per cent of the whole volume and are predominantly located along arterioles.

Pancreas development and congenital anomalies

The pancreas develops from ventral and dorsal buds of the primitive duodenum. With rotation around the duodenum the two portions fuse together and the duct originating from the dorsal portion (duct of Santorini) forms the accessory duct whilst the main drainage of the gland is through the duct of Wirsung to the ampulla of Vater. Failure of ductal fusion, pancreas divisum, in which most of the gland drains through the duct of Santorini to the minor papilla, occurs in approximately 8 per cent of the population, and in a small proportion may lead to recurrent acute pancreatitis. Annular pancreas results from pancreatic tissue remaining wrapped around the duodenum during rotation of the ventral portion. Ectopic pancreatic tissue may occur in a submucosal location within the stomach and duodenum.

Exocrine pancreas

The pancreas secretes up to 2 litres of fluid per day although resting secretion rates are very low (0.3 ml/min). Acini are located in lobules draining into extralobular ducts. Cells lining the ducts secrete bicarbonate, the major anion within pancreatic juice. The acinar cells are pyramidal with the nucleus and endoplasmic reticulum towards the base and zymogen storage granules towards the apex and draining duct. Two classes of proteolytic enzymes are secreted—the serine proteases and the exopeptidases. Serine proteases all require activation either by intestinal endopeptidase in the case of trypsinogen or by trypsin itself in the case of chymotrypsin, elastase, and protease E. Serine protease act at various cleavage points whereas the carboxypeptidases A and B (exopeptidases) cleave C-terminal amino acids. The lipolytic enzymes include phospholipase A_2, lipase, and carboxylesterase. Other proteins found in pancreatic secretions include lysosomal proteins, ribonucleases, and amylase.

Control of the secretory process involves hormones as well as sympathetic and parasympathetic nerve fibres. Secretin is the main stimulus to ductal bicarbonate secretion, whereas cholecystokinin, acetylcholine, and to a lesser extent gastrin and neurotensin stimulate zymogen release of digestive enzymes at the apical membrane. Although often described as having cephalic, gastric, and intestinal phases to indicate the origin of the pancreatic stimulus, this distinction is physiologically artificial since the phases run concurrently. Somatostatin and glucagon inhibit pancreatic pro-enzyme secretion.

Endocrine pancreas

Islets of Langerhans represent an endocrine organ consisting of four cell types: A cells secreting glucagon, B cells secreting insulin, D cells secreting somatostatin, and PP cells secreting pancreatic polypeptide. B cells constitute 80 per cent of islet volume and form the central core around which the others cells form a mantle. The principal physiological function of these cells is to maintain stable glucose concentration irrespective of substrate delivery.

B cells act as a sensor of glucose concentration over a wide range, with rapid equilibration of glucose levels across the cell membrane by the GLUT-2 transporter. The molecular basis for this sensor is considered to be glucokinase, the activity of which closely follows glucose levels. Enhanced glucose metabolism increases adenosine triphosphate/adenosine diphosphate ratios, which in turn blocks potassium ionchannels, andthe subsequent change in membrane potential allows an influx of calcium that promotes exocytosis of insulin-containing granules. Many other hormones, neuropeptides, and neurotransmitters also modulate glucose-dependent insulin secretion (Table 4).

Further reading

Balabaud C *et al*. (1988). Light and transmission electron microscopy of sinusoids in human liver. In: Bioulac Sage P, Balabaud C, eds. *Sinusoids in human liver; health and disease*, pp 87–110. Kupffer Cell Foundation, Rijswik.

Erlinger S (1993). Intracellular events in bile acid transport by the liver. In: Tavoloni N, Berk PD, eds. *Hepatic transport and bile secretion. Physiology and pathophysiology*, pp 467–75. Raven Press, New York.

Gumucio JJ (1999). Functional organisation of the liver. In: Bircher J *et al. Oxford textbook of clinical hepatology*, pp 437–46. Oxford University Press.

Kang S, Davis RA (2000). Cholesterol and hepatic lipoprotein assembly and secretion. *Biochimica et Biophysica Acta* **1529**(1–3), 223–30.

Knook DI, Wisse E, eds (1982). *Sinusoidal liver cells.* Elsevier Biomedical Press, Amsterdam.

Tukey RH, Strassburg CP (2000). Human UDP-glucuronosyltransferases: metabolism, expression, and disease. *Annual Review of Pharmacology and Toxicology* **40**, 581–616.

14.18.2Ct and mri of the liver and pancreas

C. S. Ng, D. J. Lomas, and A. K. Dixon

Introduction

Computed tomography (**CT**) and magnetic resonance imaging (**MRI**) are cross-sectional imaging techniques which allow excellent, non-invasive evaluation of the anatomical and parenchymal detail of both these organs. Diffuse and focal lesions can be demonstrated, and information on the associated vascular and biliary structures can also be obtained.

Both imaging techniques rely on the detection of a variety of physical properties of tissues, in the case of CT, on X-ray attenuation (essentially physical density); and in the case of MRI, on multiple factors including the radiofrequency response, proton density, and T_1 (longitudinal spin–lattice) relaxation and T_2 (transverse spin–spin) relaxation times of tissues. The successful detection of focal lesions within tissues depends on the ability of the imaging technique to identify sufficient differences in the relevant physical properties between the lesions and background parenchyma of the organ in question.

Major technological advances have been made in recent years in the speed of image acquisition in both CT and MRI. The main contributions have been, in the case of CT, 'spiral' (or 'helical') and multidetector technology; and in the case of MRI, improvements in the speed at which magnetic field gradients can be switched. These advances now permit breath-hold imaging in both CT and MRI. This has reduced breathing-related imaging artefacts, and has permitted the introduction of multiphasic or dynamic enhancement techniques following a bolus intravenous injection of contrast medium. This in turn has improved lesion detection and characterization. Multiplanar or three-dimensional reformation of data can, on occasions, improve the spatial appreciation of the anatomy.

In general, body CT is more widely available than MRI. Combined with its relative flexibility in imaging more than one body region, CT can be regarded as the current optimal investigation for these organs, even though ultrasound is often used as a preliminary and sometimes definitive investigation. The lesion-detection sensitivities for CT and MRI are comparable. Unfortunately, some lesions can be overlooked by both, or one or other investigation. In general, MRI has an advantage over CT in characterizing lesions, and is often used as a problem-solving examination. Although lesions as small as 2 or 3 mm can be demonstrated, lesions smaller than 10 mm are very difficult to characterize. If necessary, and technically feasible, tissue specimens for cytopathology or histopathology can be obtained by imaging-guided, fine-needle aspiration (**FNA**) or histological core biopsy (Fig. 1). New developments such as CT fluoroscopy and open MRI machines are likely to improve the interventional scope of these techniques, which can include therapy (for example, radiofrequency or laser ablation of neoplastic lesions). However, such techniques are still under evaluation and only available in a very limited number of centres. Some of the relative merits of CT and MRI are presented in Table 1. In many instances, other imaging techniques, such as ultrasound, angiography, and nuclear medicine, provide complementary information.

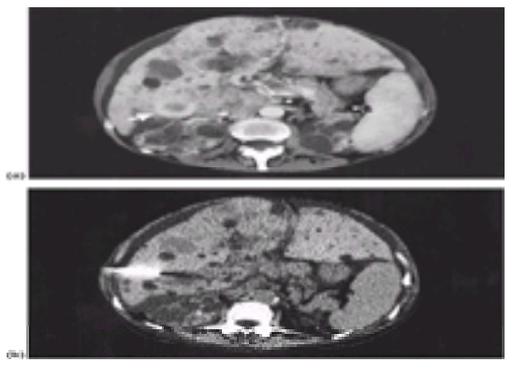

Fig. 1 (a) and (b). CT-guided biopsy of ring-enhancing liver lesion (arrow) in a complicated patient with polycystic liver and renal disease. This yielded adenocarcinoma. Note ascites.

CT and MRI of the liver

Modern spiral CT machines are able to obtain images of the liver in a single breath-hold (in the order of 15 to 25 s). Lesion detection is improved with the use of iodinated intravenous contrast media. In certain circumstances, the sensitivity and specificity of lesion detection is further improved by 'biphasic' imaging during the infusion of intravenous contrast media (that is, during arterial-dominant and portal-dominant phases). This is particularly true for hypervascular lesions, such as hepatoma, which may only be detected in the arterial-dominant phase of intravenous contrast medium infusion (Fig. 2). For the clinician, this means that the likely diagnosis(es) and clinical questions must be conveyed to the radiologists in order to adapt the CT protocol appropriately.

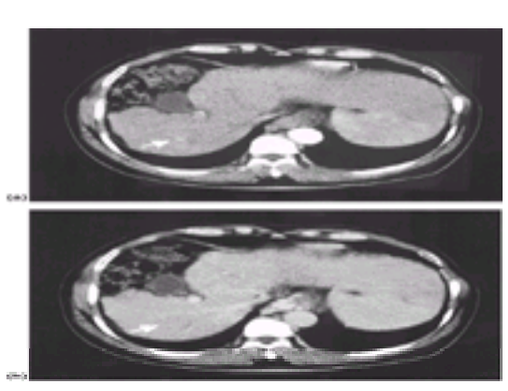

Fig. 2 CT showing hepatoma within a cirrhotic liver. Note the small shrunken right lobe. The small hepatoma (arrow) shows well in the arterial phase (a) due to the increased arterial supply. It is very much more difficult to see in the corresponding portal phase image (b), the conventional method of examining the liver.

MR imaging of the liver centres on T_1-weighted and T_2-weighted images (the latter employing 'conventional' or 'fast' spin-echo sequences, often with 'fat suppression') (Fig. 3). As with CT, breath-hold imaging is now possible, and 'dynamic' images during infusion of intravenous contrast media (for example,

gadolinium–diethylene-triamine-pentaacetic acid; **Gd-DTPA**) can add sensitivity and specificity. The addition of 'liver-specific' agents may also contribute to lesion detection (for example, paramagnetic iron particles which are taken up by Kupffer cells) or manganese-based agents (which are taken up by hepatocytes). Flow-sensitive sequences can provide information on the patency or involvement of the major hepatic vessels.

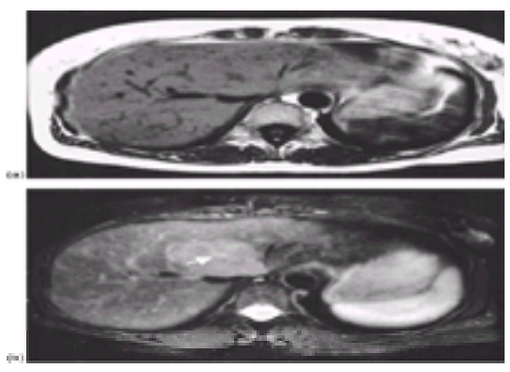

Fig. 3 Liver MRI. (a) T_1-weighted image. At first glance this appears normal. The intrahepatic veins are seen as signal voids because of the flowing blood. (b) T_2-weighted image. A large high-signal intensity lesion in segment IV is now apparent. This was focal nodular hyperplasia with a small central scar (arrow), which can just be identified on the T_1-weighted image in retrospect.

CT arterioportography and CT following intra-arterial introduction of Lipiodol are further techniques that attempt to improve lesion detection; but these are invasive, requiring selective mesenteric intra-arterial catheterization. Both are subject to important artefacts, mainly due to minor anatomical vascular variants, which can cause false-positive results.

Focal liver lesions

The detection of focal liver lesions plays an important role in the management of patients with known or suspected malignancy. However, the differentiation between benign and malignant lesions can be extremely challenging. The former include cysts, focal nodular hyperplasia (Fig. 3), haemangiomas (Fig. 4), adenomas, focal fatty infiltration (and sparing), regenerating nodules in cirrhosis, and small abscesses/granulomas. Difficulties arise when lesions do not display their characteristic imaging appearances, and particularly when they are small (<1 cm). In addition, there is considerable overlap in the appearances of benign and malignant lesions. The overall accuracy of CT in the diagnosis of hepatic metastases is in the region of 60 to 80 per cent; high-quality MRI yields slightly better results. Intraoperative ultrasound is generally considered to provide the best sensitivity. A key problem is the interpretation of a solitary small lesion in a patient with a known malignancy; such lesions are now commonly identified due to the advances in imaging technology; most will prove to be benign, even in patients with known malignancy.

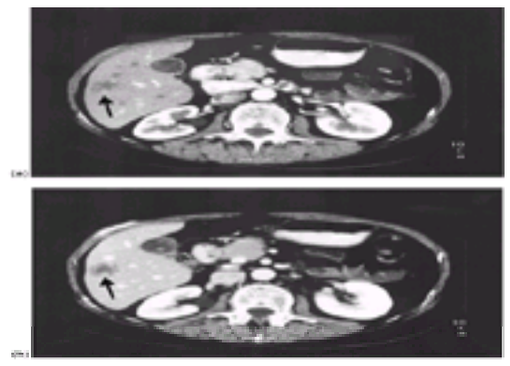

Fig. 4 CT showing liver haemangioma. On unenhanced images, the lesion had the same CT attenuation as blood in the aorta. It still shows as a low attenuation in the arterial phase (a). In the portal phase (b), it is filling in from venous lakes in the periphery (arrow).

Surgical resection for hepatic lesions, particularly metastases from colorectal carcinoma (Fig. 5), is a therapeutic option in many centres. CT and MRI can be used to determine the anatomical relationships of the tumour(s). For these purposes, the ability to delineate the functional divisions of the liver (Couinaud's nomenclature) is particularly useful from a surgical point of view (Fig. 6).

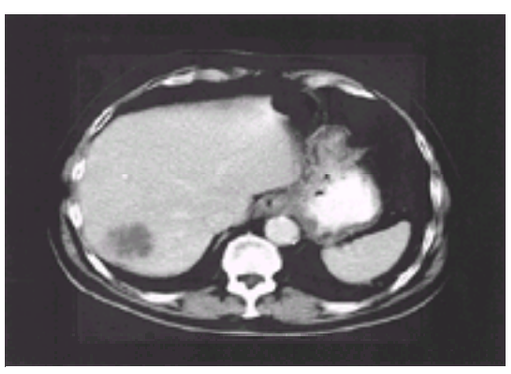

Fig. 5 CT showing liver metastasis from colorectal carcinoma.

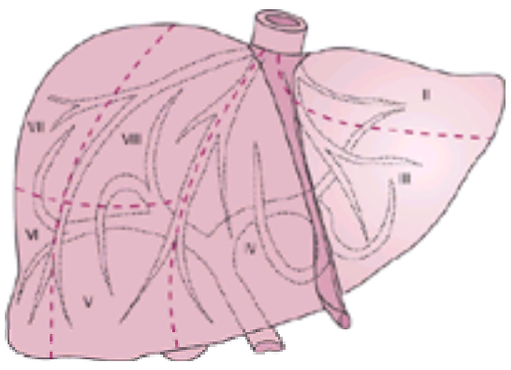

Fig. 6 Functional division of the liver and its segments according to Couinaud's nomenclature. (Reproduced from Merran S, Hureau J, Dixon AK, eds. (1989). *CT and MRI radiological anatomy*, with permission from Butterworth Heinemann.)

Diffuse liver lesions

Cirrhosis may be suggested in the presence of a distorted liver, an irregular liver surface, multiple hepatic nodules, or signs of portal hypertension (splenomegaly, ascites, or varices). However, in general, diffuse liver processes, which can include tumours and inflammatory conditions, are not reliably detected or characterized. Notable exceptions are the striking low CT attenuation caused by fatty infiltration (Fig. 7), and the abnormally high CT attenuation and low T_2-weighted MR signal intensity caused by heavy metal excess (usually iron; occasionally iodine following prolonged amiodarone therapy) (Table 2). In this regard, it is worth noting that the excess of copper found in Wilson's disease does not result in high CT attenuation; the excess is only in the microgram range; and the accompanying cirrhosis tends to reduce the attenuation.

Fig. 7 CT showing a fatty liver and chronic pancreatitis. Dilated pancreatic duct (arrow), and calcification in pancreatic body (arrow). Note the markedly low attenuation of the liver (the normal density relationship of unenhanced hepatic vessels and liver is reversed), due to diffuse fatty infiltration (steatosis), which is turn is most commonly from alcoholic excess.

Abnormalities involving the vascular tree can, on occasion, be detected, for example invasion or occlusion of the portal or hepatic veins by a tumour or thrombus. MRI, with its flow-sensitive capability, is superior to CT in this regard. Budd–Chiari syndrome produces a recognizable, but not entirely specific, parenchymal enhancement pattern. However, the key feature of an occlusion or thrombus within hepatic veins is not always reliably demonstrated by CT or MRI (nor indeed Doppler ultrasonography).

CT and MRI of the pancreas

The pancreas lies in a retroperitoneal location and its orientation in an oblique transverse plane is ideally suited to the axial planes of CT and MRI. Surrounding bowel and retroperitoneal fat are frequently helpful in delineating the pancreas in CT and MRI; conversely, these factors typically hinder ultrasound visualization of the organ. Although breath-hold, thin-section CT images of the pancreas are still anatomically superior to those obtained from MRI, the opportunities of magnetic resonance cholangiopancreatography (**MRCP**) images (heavily T_2-weighted, fluid-sensitive images) on modern MRI machines offer unique supplementary information about the biliary tree. The complementary role of endoscopic ultrasound, which continues to expand, should not be forgotten.

Acute pancreatitis

CT is generally considered the imaging investigation of choice, and can be valuable when the diagnosis is in doubt (Fig. 8(a)). On occasions, the causative distal common bile duct calculus can be identified (calculi in this location can be difficult to identify on ultrasound). Complications of acute pancreatitis may also be identified, these include necrotizing pancreatitis, abscess and/or pseudocyst formation, splenic vein thrombosis, and pseudoaneurysm formation. An assessment of the amount of viable pancreatic tissue can be made from the proportion of pancreas that enhances following administration of intravenous contrast media (Fig. 9). Differentiation between a simple pseudocyst and an infected collection can be difficult (Fig. 8(b)), but CT-guided diagnostic aspiration and/or drainage of collections can be undertaken.

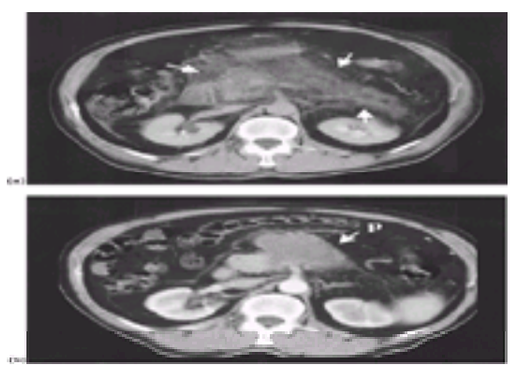

Fig. 8 CT showing acute pancreatitis progressing to pseudocyst formation. (a) In the acute phase note the poorly defined tissue planes (arrows), especially around the pancreatic head and Gerota's fascia. (b) One year later, the tissue planes around the pancreatic head are well defined, despite the distortion caused by the pseudocyst (p).

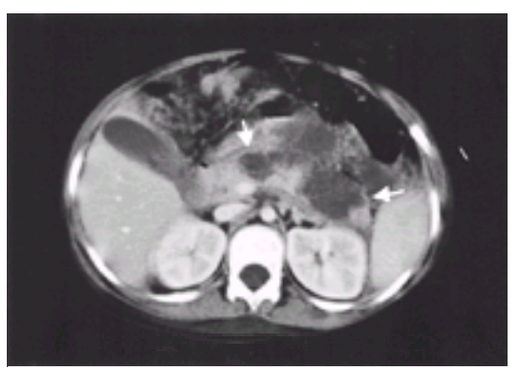

Fig. 9 CT of an acute necrotizing pancreatitis. Note ill-defined tissue planes due to retroperitoneal inflammation and non-enhancing foci following intravenous contrast medium (arrows), indicating areas of pancreatic necrosis.

Chronic pancreatitis

A diagnosis of chronic pancreatitis can be inferred on CT if there are the typical appearances of an atrophied gland with associated coarse parenchymal calcification and a dilated pancreatic duct (Fig. 7). However, the dilated and beaded pancreatic duct pattern that is the hallmark of early chronic pancreatitis on endoscopic retrograde cholangiography (**ERCP**) cannot be demonstrated by CT. At present, MRCP can only identify quite severe changes reliably; ERCP remains the 'gold standard' for subtle changes.

Focal pancreatic lesions

These include adenomas, adenocarcinomas (the most common), islet-cell tumours, cysts (which may be congenital, benign, or malignant), and focal pancreatitis. The role of imaging is in identifying the lesions, assessing treatment response, and evaluating surgical resectability.

Characterization of focal pancreatic lesions by CT or MRI is not well refined. Although, hypervascular lesions are typical of islet-cell tumours (Fig. 10), the majority of pancreatic lesions are hypovascular and there is considerable overlap in appearances. In particular, differentiation between pancreatic adenocarcinoma (Fig. 11) and focal pancreatitis can be difficult. So too may be the distinction between pancreatic pseudocyst, cystadenoma, and cystadenocarcinoma. In such circumstances, it may be necessary to perform to image-guided, fine-needle aspiration or biopsy. However, it should be emphasized that obtaining diagnostic material (for example, from pancreatic adenocarcinomas) is notoriously difficult, even for the surgeon who may have the pancreas exposed.

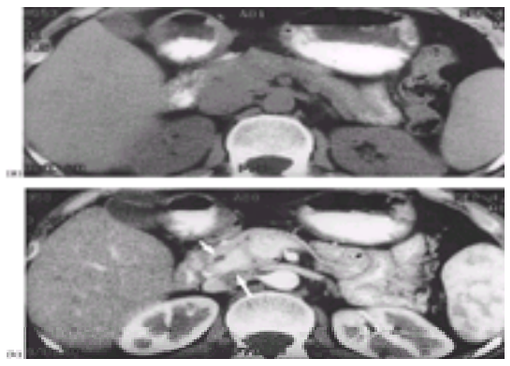

Fig. 10 (a) and (b). CT showing an islet-cell tumour of the pancreas. High-density lesion (arrow) adjacent to a normal common bile duct (small arrow) visualized on the arterial-dominant image (b) of intravenous contrast-medium enhancement (but not on the unenhanced image (a)).

Fig. 11 CT showing a small pancreatic head carcinoma. A low-attenuation tumour (straight arrow), medial to a biliary stent, occupies the uncinate process and extends towards the normal pancreatic head. The superior mesenteric vein (curved arrow) and smaller accompanying artery are opacified well; they have normal adjacent fat planes.

The accuracy of CT and MRI in assessing resectability in pancreatic adenocarcinoma is comparable.

Diffuse pancreatic changes

CT can show fatty changes within the pancreas. In type II diabetes, the outline of the pancreas can become irregular with increased fat within. In cystic fibrosis, the pancreas becomes fatty and will eventually atrophy. Similar appearances can be demonstrated by MRI.

Further reading

Bearcroft PW, Gimson A, Lomas DJ (1997). Non-invasive cholangio-pancreatography by breath-hold magnetic resonance imaging: preliminary results. *Clinical Radiology* **52**, 345–50.

Dixon AK, Walshe JM (1984). Computed tomography of the liver in Wilson disease. *Journal of Computed Assisted Tomography* **8**, 46–9.

Grainger RG, Allison DJ, eds. (1997). *Diagnostic radiology*, 3rd edn, pp 49–81, 1155–99, 1269–93. Churchill Livingstone, Edinburgh.

Husband JES, Reznek RH, eds. (1998). *Imaging in oncology*, pp 129–90. ISIS Medical Media, Oxford.

Megibow AJ, *et al*. (1995). Pancreatic adenocarcinoma: CT versus MR imaging in the evaluation of resectability—report of the radiology diagnostic oncology group. *Radiology* **195**, 327–32.

Miller FH, *et al*. (1998). Using triphasic helical CT to detect focal hepatic lesions in patients with neoplasms. *American Journal of Roentgenology* **171**, 643–9.

Miller WJ, *et al*. (1994). Malignancies in patients with cirrhosis: CT sensitivity and specificity in 200 consecutive transplant patients. *Radiology* **193**, 645–50.

Ros PR, Payley MR. (1997). MR imaging of the liver – a practical approach. *Magnetic Resonance Imaging Clinics of North America* **5,** 415–29.

Ros PR, Taylor HM. (1998). Hepatic imaging: an overview. *Radiological Clinics of North America* **36**, 237–45.

Schwartz LH, *et al*. (1999). Prevalence and importance of small hepatic lesions found at CT in patients with cancer. *Radiology* **210**, 71–4.

14.18.3.1Acute pancreatitis

C. W. Imrie

Incidence and epidemiology

In different countries incidence figures vary from 40 to 500 new cases of acute pancreatitis per million of the population arise each year. Biliary disease and alcohol abuse are the main associated factors, gallstones accounting for 40 to 70 per cent of all cases. Older men and most women have biliary pancreatitis, while younger men develop pancreatitis due to alcohol abuse, as shown by studies of urban populations. This is typified by a study from Gothenburg, where, from the late 1950s to the mid 1970s, there was a change from a 68 per cent association with biliary disease to exactly same proportion with pancreatitis due to alcohol abuse.

In recent large population studies from Scotland and Finland the incidence of the disease has risen steadily to the current 400 patients per million population per year (higher figures from the United States are less reliable). From 1960 onwards, the mortality has reduced: it was 7.5 per cent overall in 1985 to 1994 with half the deaths occurring in the first week of illness. Death is usually due to multiple organ failure in which respiratory failure predominates.

Aetiological factors (Table 1)

Major factors

Biliary disease and alcohol abuse together account for over 80 per cent of patients in most prospective studies. With recent diagnostic advances, particularly the early use of endoscopic retrograde cholangiopancreatography (**ERCP**) and endoscopic ultrasound (**EUS**), it is clear that almost all the remaining patients have very small stones. Biliary sludge or bile crystals may also be indicative of a tendency to stone formation. Studies in the United Kingdom reveal that 40 to 65 per cent of patients have small gallstones. In Ohio in the United States a similar incidence has been detected, while in Argentina over 80 per cent of patients have been shown, by faecal sieving, to have small gallstones. The means by which the transient migration of small stones causes acute pancreatitis is not understood, but small stones which more easily exit the gall bladder to impact (usually transiently) in the ampullary area of the bile duct are a major factor.

Pancreatitis due to alcohol abuse occurs in over 80 per cent of patients from New York, and around 70 per cent in Helsinki. This association is usually found in young males who drink in excess of 80 g alcohol per day. Up to 10 per cent of patients have both biliary disease and abuse alcohol. Alcohol may provoke acute pancreatitis by acinar stimulation with simultaneous ampullary spasm.

Minor factors (Table 1)

Iatrogenic causes

Surgical or endoscopic procedures involving the ampulla of Vater can induce pancreatitis. The frequency of post-ERCP pancreatitis is increasing, even though the procedure only carries a risk for acute pancreatitis of about 1 per cent. Following a therapeutic endoscopic sphincterotomy the risk of acute pancreatitis is approximately 3 per cent. It is possible that failure to clear the duct of stones and inadequate sphincterotomies are the two most common predisposing features because they often allow stones to impact at the ampullary area. Manometric studies by patients themselves in a small group of patients with sphincter of Oddi dysfunction (SOD) are associated with acute pancreatitis in around 30 per cent of cases.

Viral infection

Viral infection, particularly mumps, coxsackie B, and viral hepatitis, can cause acute pancreatitis which is often missed. One clinical feature that may prove useful is prodromal diarrhoea, which is rare in all other types of acute pancreatitis.

Drug-induced acute pancreatitis

The drugs most commonly implicated in acute pancreatitis are valproic acid, azathioprine, L-asparaginase, and corticosteroids. However, unless viral titres have been determined, together with adequate biliary investigations, it is unwise to ascribe acute pancreatitis to a particular drug. Repeat exposure to the same drug again causing acute pancreatitis is the strongest evidence of a direct association.

Hyperparathyroidism

This is now recognized to be an uncommon accompaniment of acute pancreatitis. Indeed, many of the reported patients have also had gallstones. The association is calculated at 0.1 per cent. Removal of a parathyroid adenoma usually prevents further acute pancreatitis since persistent hypercalcaemia appears to be the provoking

factor.

Hyperlipoproteinaemia

Patients with type I and V hyperlipoproteinaemia (see Chapter 11.6) may develop acute pancreatitis. The significance of this association can be difficult to validate. It has been found that acute pancreatitis associated with hyperlipoproteinaemia is rare in patients who do not, at the same time, have a high alcohol intake. Nevertheless, several experimental studies point to the importance of this association; patients with primary hyperlipoproteinaemia with chylomicronaemia and hypertriglyceridaemia are prone to attacks of acute pancreatitis in the absence of alcohol ingestion.

Hypothermia

This is a particularly important association in the elderly when pancreatitis may be associated with myxoedema coma. In younger patients, alcohol abuse may be linked, particularly if patients fall asleep out of doors or in a cold, unheated house.

Blunt trauma (Fig. 1)

This is a notable cause of acute pancreatitis, particularly in young children. Sports injuries from rugby, football, ice hockey, martial arts, and similar activities may result in acute pancreatitis, usually from a crush injury to the body of the pancreas against the vertebral column. Of greater importance numerically are victims of road traffic accidents when the diagonal section of seat belts is sometimes incriminated.

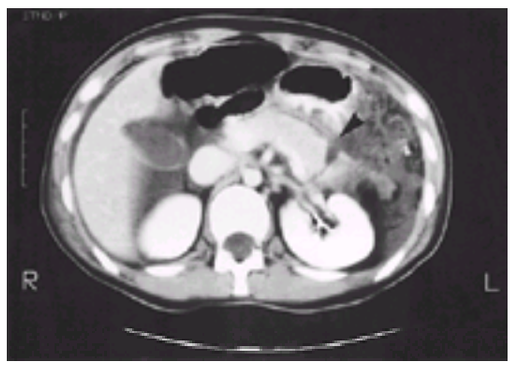

Fig. 1 Blunt trauma causing pancreatitis by transection at the arrowpoint on the CT scan.

Periampullary adenoma or cancer

This is an important association, best diagnosed by ERCP. With the increase in this approach to diagnosis, tumours at or close to the ampulla have been shown to have a greater association with acute pancreatitis then hyperparathyroidism (0.4 per cent). Effective treatment of the tumour abolishes recurrent attacks. This usually involves surgical resection, but in older less fit patients endoscopic laser therapy can be effective.

Hereditary

This form is increasingly being studied since the discovery of two genetic familial defects of trypsinogen (*N21I* and *RII7H*) have been identified with clinical acute pancreatitis usually presenting in children and young adults. Chronic pancreatitis follows from 20 to 50 years and an appreciable incidence of pancreatic carcinoma by 65 to 70 years. These cationic trypsinogen defects are autosomal dominant and shed light on the mechanism of acute pancreatitis.

Least common causes

The unusual causes of acute pancreatitis are listed in Table 2. The link with pancreatic cancer and metastases to the pancreas is well documented.

Microscopic pathology

All patients with acute pancreatitis have microscopic evidence of necrosis, while macroscopic changes, particularly black discoloration, are confined to the most severe cases. It is more frequent for this gross degree of necrosis to occur in the peripancreatic fatty tissue than in the pancreas itself. When present in the pancreas there is usually a panlobular necrosis and it is impossible to delineate where the disease initiated.

In a classical paper, Foulis claimed that the most common microscopic abnormality seen in humans, periductal necrosis, is typical of biliary and alcohol causation. Less commonly a perilobular necrosis is found, usually in patients with hypothermia or gross hypotension.

In experimental acute pancreatitis the initial lesion is now considered to be intracellular, featuring coalescence of lysozymes and zymogen granules. Acinar cell disruption is found with many of the hyperstimulation models such as caerulin-induced acute pancreatitis. It is now believed that this initial event may be associated with oxidative stress. Although this is very difficult to prove, it has formed the basis of some putative approaches to treatment.

Clinical presentation and laboratory abnormalities

Clinical features

Sudden onset of upper abdominal pain with vomiting is the most common manner of presentation.

The pain may focus in the epigastrium or right or left upper quadrant with penetration through to the back. Occasionally it encircles the upper abdomen. Patients who have experienced both a myocardial infarct and acute pancreatitis usually describe the latter pain as being much more severe. However, it tends to lessen in severity progressively over the first 72 h, and it is not usually a significant factor beyond this time. The pain on presentation is very similar to that of a perforated duodenal ulcer, but vomiting is less common in those with perforated ulcers.

Up to 90 per cent of patients with acute pancreatitis have troublesome vomiting in the first 12 h of illness, and this contributes to hypovolaemia and hypotension.

Patients with stones in the common bile duct may well be jaundiced and cholangitis can supervene in a minority. Much milder degrees of jaundice may occur from external compression of the lower bile duct in patients with alcohol-induced disease.

Hypotensive shock only occurs in very severe cases, but loss of circulating volume due to the extravasation of albumin, coupled with vomiting, leads to dehydration; the patient is thirsty, but fears drinking because of vomiting.

Bowel sounds are rarely present in the early phase of the disease and paralytic ileus may occasionally extend beyond 4 days. Despite these observations early nasojejunal feeding is possible even in clinically severe acute pancreatitis, and this is now a new therapeutic approach often begun within 48 h of the onset of disease.

The course of mild acute pancreatitis

Those patients who fail to meet objective criteria of severe acute pancreatitis tend to have a low mortality (maximum 2 per cent) and rarely need to be in hospital beyond 7 to 10 days. Simple therapeutic measures normally suffice through the first 24 to 48 h, at which time nasogastric suction and urinary catheterization can usually be discontinued. It is safer to assume that a patient may move into the more severe group and to provide early monitoring of the volume of nasogastric aspirate and hourly urinary output to maintain adequate fluid replacement. Even in this category of patients it may occasionally be necessary to provide 4 to 5 litres of intravenous fluid in the first 24 h of the illness. It is important to lower the risk of further attacks by clearing gallstones in the same admission.

The course of severe acute pancreatitis

Patients who meet objective criteria of severe acute pancreatitis may be pyrexial and are hypotensive, markedly tachypnoeic, and suffer from abdominal ascites, pleural effusions, and prolonged paralytic ileus. Body wall staining at the umbilical area (Cullen's sign) or in the flanks (Grey Turner's sign) can occur, usually appearing around the fourth day of illness.

Respiratory insufficiency or failure

Hypoxaemia is the hallmark of acute pancreatitis and reflects its severity. The basic mechanism of the hypoxaemia is unknown but high levels of various cytokines, as well as leucocyte elastase, are implicated, together with factors that contribute to localized intravascular coagulation. Shunting of blood occurs in the pulmonary vascular bed and accounts for up to 30 per cent of the cardiac output.

The initial clinical sign, which may easily be overlooked, is a fast respiratory rate; the patient may be cyanosed but there is no substitute for the measurement of arterial oxygen pressure. Almost all of the systems for objective monitoring of severity of disease include an arterial oxygen pressure of less than 60 mm Hg (8 kPa) as an index of severity. Hypoxaemia can usually be reversed by the provision of humidified oxygen, and in severe cases the pattern is similar to adult respiratory distress syndrome of other causes. Pleural effusions may be large enough to warrant aspiration, and when humidified oxygen is insufficient to reverse the hypoxaemia, assisted ventilation is necessary. Hyaline membrane formation has been found in severe cases, and even in milder cases complete reversal of the respiratory insufficiency takes many weeks.

Basal atelectasis and respiratory compromise are very common and must be expected. Urine output and arterial oxygen saturation must be monitored. Pulse oximetry is useful in monitoring, but arterial gas analysis may be needed three or four times in the first 24 h in order to make sensible decisions about humidified oxygen therapy and possible ventilator support. Single organ insufficiency or failure necessitates high-dependency care and often full intensive care management, as this may presage multiorgan failure.

The cardiovascular system

The initial hypovolaemia is of great importance in cardiac and renal function. Where cardiac output is more compromised and simple fluid replacement does not restore circulating volume, support drugs such as catecholamines may well be necessary. In the most severe acute pancreatitis the cardiovascular changes are very similar to those encountered in septic shock, with a high cardiac output and low peripheral vascular resistance. Stress on the heart may cause arrhythmias and ischaemic changes.

Renal impairment

Patients with acute pancreatitis are at risk from the development of acute renal failure which is related to hypovolaemic shock and the acute inflammatory changes associated with the illness.

The single most important corrective measure is to provide adequate fluid replacement. Low-dose dopamine may be a useful drug, and diuretics such as frusemide and mannitol may still have a place.

Additional measures are required in any patient failing to produce 30 ml of urine per hour. Most respond to increasing the rate of intravenous fluid replacement but more vigorous measures are necessary in the sickest patients, including haemoperfusion peritoneal dialysis or haemodialysis.

Biochemical abnormalities

A multitude of biochemical phenomena are found in acute pancreatitis—various pancreatic enzymes are released that are useful as diagnostic markers. With acinar cell disruption, high serum activities of amylase, lipase, trypsin, chymotrypsin, phospholipase, elastase, as well as breakdown products such as trypsinogen activation peptide and phospholipase activation peptide, are all found. The cheapest and most durable of these measurements as a diagnostic marker has been the total activity of serum amylase. Levels over four times the upper limit of normal in blood are usually taken as diagnostic of acute pancreatitis, provided the clinical course corresponds. The serum lipase activity is a more specific measure and is now almost as cheap to measure; levels of twice the upper limit of normal are significant. Other body fluids contain elevated activities of these enzymes.

Measurements of serum trypsin and chymotrypsin activity or antigens are expensive and the antiprotease defence mechanisms are efficient at releasing a_2-macroglobulin and a_1-antiprotease (also known as a_1-anti-trypsin) which rapidly counteract free trypsin and chymotrypsin within body fluids. Thus measurement of both trypsin and chymotrypsin can be unrewarding while measurement of urinary trypsinogen activation peptide shows good potential. This small peptide molecule is excreted in urine very early in the disease course in patients with severe acute pancreatitis. It is therefore a good measure of the degree of disruption of acinar cells and a commercial assay is now available. This may be a valuable investigation for both diagnosis and for gauging severity of acute pancreatitis.

For many years it was believed that the antiprotease defence mechanisms required supplementation in patients with acute pancreatitis. Thus aprotinin (Trasylol), gabexate mesilate (FOY), and purified plasma derivatives have been administered intravenously in the hope of improving the clinical course. It is now clear, however, that the antiprotease defence system is intact and that there is no need to boost it—a conclusion that is supported by many clinical trials. Aprotinin was also given into the peritoneal cavity in the hope that it would improve survival and reduce complications, but was unsuccessful.

Very high concentrations of circulating cytokines are found in the blood at an early stage in the disease. The proinflammatory tumour necrosis factor-a, platelet activating factor, and interleukin 6 are present in greatest concentration in those with severe pancreatitis. The stimulus to the release of cytokines is thought to be endotoxin, probably from the gut, and there is great interest in the possibility of administering agents that inhibit endotoxin or the cytokines as a potential therapy. Alternatively exogenous interleukin 10 (an anti-inflammatory cytokine) may move from experimental to clinical use, but early clinical studies of post-ERCP pancreatitis have produced conflicting results.

Calcium-albumin

It has long been known that hypocalcaemia is a feature of acute pancreatitis; however, a significant proportion of patients have only a drop in protein-bound calcium and the primary pathology is the loss of albumin from the intravascular space rather than decreased ionized calcium. However, even after correction factors are applied there is an undoubted tendency for serum ionized calcium to fall; this is usually counteracted by a marked elevation in parathyroid hormone.

Haematological abnormalities

There is marked haemoconcentration associated with hypovolaemia so that initial haemoglobin levels of over 16 g/100 ml may be found. After rehydration, the haemoglobin level falls but it is unusual for blood transfusion to be required as the degree of haemorrhage in and around the pancreas is usually not of great moment. Later in the course of the disease, bleeding from gastric erosions, peptic ulcer, or haemorrhage into a pseudocyst may require blood products and endoscopic, angiographic, or surgical intervention.

In addition the acute-phase response results in high levels of liver-derived C-reactive protein, a $_1$-antiprotease, factor V, factor VIII, and fibrinogen. Platelet and a_2-macroglobulin levels fall in the first week, usually returning to normal by days 8 to 10. The common finding in patients with severe acute pancreatitis is hypercoagulation, and disseminated intravascular coagulation, while it does occur, is very uncommon; its management is considered in Section 22. The rapidity of mediator response makes leucocyte elastase, tumour necrosis factor-a, and interleukin 6 potentially excellent markers of severity; C-reactive protein levels in excess of 150 mg/litre, are another useful marker of severity.

Pyrexia

This reflects cell damage and necrosis as with any condition associated with tissue destruction. A low-grade fever is typical of the first 3 to 4 days of illness. Especially in those who have the clinical signs of obstructive jaundice, ascending cholangitis should be suspected and appropriate antimicrobials given. In the most severely ill patients, transudation of bacteria from the transverse colon into necrotic tissue around the pancreas, and occasionally in the gland itself, has been detected within 48 to 96 h of onset, although it is more typical to find such sepsis at a later stage.

Making an accurate diagnosis

The diagnosis is usually made from the clinical presentation, particularly the rapid onset of upper abdominal pain and vomiting. Gross elevations of amylase and lipase in blood usually support the diagnosis, while urinary amylase levels of greater than four times the upper limit of normal can be helpful in less typical cases.

Peritoneal aspiration (after catheterization of the urinary bladder and nasogastric intubation) can be used where diagnostic doubt still exists. The aspiration of more than 20 ml of free fluid without bacterial contamination (evident by smell and Gram stain) is indicative of a severe form of the disease. The darker the colour of the fluid, the more severe the disease. This procedure is especially effective in patients with alcohol-induced acute pancreatitis. The presence of bacterial contamination indicates an alternative diagnosis and the need for an immediate laparotomy, as visceral perforation of the duodenum or small bowel is more likely.

Computed tomography (**CT**) will reveal pancreatic swelling, fluid collection, and change in density of the gland. Contrast enhanced CT scanning is mandatory to identify areas of pancreatic ischaemia and infarction. Magnetic resonance (**MR**) scanning may ultimately replace contrast enhanced CT scanning in this area. Either CT or MR can help in the difficult diagnosis.

Differential diagnosis (Table 3)

A dissecting aortic aneurysm usually presents with an initial history of chest pain in a known hypertensive; abdominal pain and loss of arterial pulses may occur later. A minor degree of pancreatitis due to ischaemia of the pancreas should not obscure the main diagnosis. Elevated amylase activities are frequently present in patients with renal failure, while a lifetime of high amylase occurs in those with macroamylasaemia; failure to filter the amylase complex results in very low urine levels.

High amylase activity in ectopic pregnancy derives from the fallopian tubes but the clinical presentation should not be mistaken for acute pancreatitis. Patients with diabetic ketoacidosis may occasionally have very high levels of amylase but this should not distract from the diagnosis.

Small bowel obstruction is associated with multiple gas–fluid interfaces on erect abdominal radiographs. The differentiation from an early perforated duodenal ulcer (less than 5 h) will rest on the combination of the finding of free gas under the diaphragm on radiographs and the lack of a significant rise in blood amylase to greater than twice the upper limits at the acute stage of illness. A more difficult diagnostic problem arises in the patient who has had a perforated duodenal ulcer for some hours because a marked elevation of serum amylase can occur. Similarly, mesenteric ischaemia or infarction can be associated with biochemical changes akin to those of acute pancreatitis, but in both these situations bacterial contamination of the peritoneal cavity will be detected by peritoneal aspiration.

Grading disease severity

The importance of objective grading of disease severity is that less experienced clinicians can direct the more serious cases to high-dependency or intensive care facilities at an early stage of their illness, or instigate contrast enhanced CT scanning and early ERCP for patients who will derive most benefit. Grading is also useful for trials of different forms of therapy.

The original Ranson grading system of 11 prognostic factors was developed for patients with acute pancreatitis due to alcohol abuse but later a system was introduced for those with gallstones. An alternative single system, validated for both the common causes, is the Glasgow scoring system of eight prognostic factors (Table 4). Validation came from a multicentre randomized British study that assessed the place of peritoneal lavage in the management of severe acute pancreatitis.

Atlanta criteria

In 1992 in Atlanta an international conference on disease nomenclature decided that severe acute pancreatitis was the presence of failure of one or more organs or the development of a major later complication—infected necrosis, abscess, or pseudocyst.

CT scanning

This can be very useful in confirming the diagnosis and also to grade severity of disease (Table 5). This is not a more accurate system of grading than the Glasgow score but it is very helpful in assessing an individual patient (Fig. 2). It is expensive and is not usually necessary in the initial few days of illness. The area of non-perfused pancreas corresponds to the extent of necrosis. Greater than 50 per cent necrosis (especially if the head of the pancreas is involved) is associated with the most severe disease.

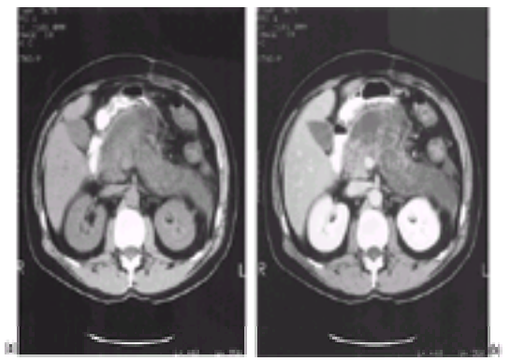

Fig. 2 (a) Severe acute pancreatitis with diffuse pancreatic swelling (CT scan). (b) Same scale level as in (a) with angiogram enhancement, revealing hypodense areas of poor perfusion.

The APACHE II system (Fig. 3)

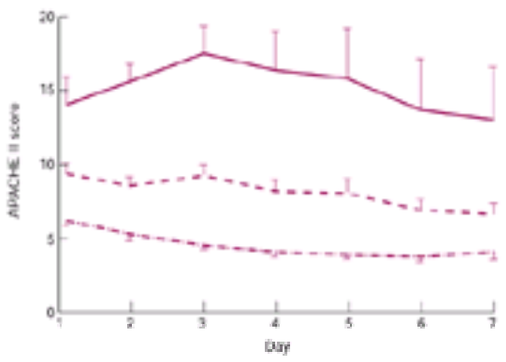

Fig. 3 Mean daily APACHE II scores by outcome in 119 patients with an uncomplicated course (–• – •–), 26 patients with a complicated course (– – – –), and 12 patients with a fatal outcome (—). The differences between fatal and uncomplicated and between complicated and uncomplicated were highly significant ($p < 0.001$) for each day (Mann–Whitney U test). (Published with permission from the *British Journal of Surgery*.)

This can be used to grade the severity of many diseases and has been shown to be useful in acute pancreatitis. It takes time for an individual clinician to learn to use the system, but it has the advantage that it can be applied throughout the first week of illness. The higher the score the worse the prognosis. Patients with the most severe acute pancreatitis have scores in excess of 10. A recent large (more than 1500 patients) clinical study assessing the potential role of a platelet activating factor antagonist in the management of higher-risk patients revealed major concerns about an entry criterion of an admission APACHE score of greater 6 or greater to select patients as the mortality rate in each of three groups of approximately 500 patients was less than 10 per cent.

Organ failure scoring

Clinical assessment by an expert is probably better than any of the other systems described at identifying the most ill patients. Quantification of the clinical assessment by scoring 0 to 4 points for each of several organ systems has revealed that 44 per cent of patients with an APACHE II score of 6 or more will show an organ failure score of 2 or more at admission. Most of these improve with supportive intravenous fluids and oxygen, but roughly a third continue with this degree of organ failure score or deteriorate.

It is this dynamic aspect of organ failure that is not recognized by the Atlanta criteria and probably caused most problems in previous grading systems. In the group who do not improve after admission with an organ failure score of 2 or more, and those who later develop this feature, we have found (in prospective data collected from 121 patients in Glasgow) a mortality in excess of 50 per cent. The application of such a scoring system for organ compromise and failure has the promise of greater accurary than Ranson, Glasgow, or APACHE II methods.

Obesity

Morbid obesity is usually described as a body mass index of over 40 kg/m^2. Acute pancreatitis carries a significantly higher mortality and morbidity in patients with a body mass index greater than 30 kg/m^2 (obesity) mainly because of increased risk of hypoxaemia but also from other associated factors.

C-reactive protein

Being an acute-phase reactant the main value of this marker is around 36 to 48 h of illness when the baseline levels of less than 10 mg/litre are far greater in those with severe acute pancreatitis. A cut-off at 150 mg/litre is a useful guide in groups of patients being assessed, but rogue results occur (Fig. 4). Of all the simple markers of severity, most experience has been gained with C-reactive protein. Values between 200 and 600 mg/litre are found in the most severely ill patients. The test is cheap and easily performed.

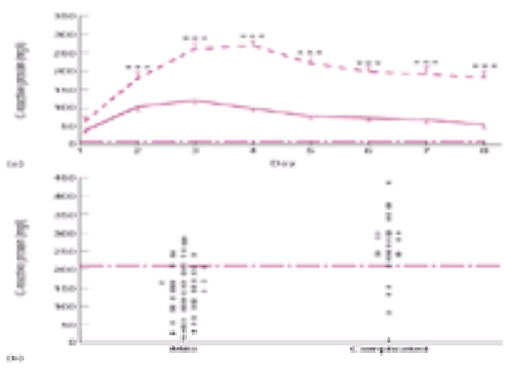

Fig. 4 (a) Sequential C-reactive protein concentrations in 47 patients with mild pancreatitis (—) and 25 with complicated attacks (– – –). Results are expressed as mean ± standard error of the mean: *$p < 0.05$; **$p < 0.01$; ***$p < 0.01$ (mild versus complicated); (— —), upper limit of normal for C-reactive protein. (b) Scattergram showing discrimination between mild and complicated attacks of pancreatitis based on the peak C-reactive protein concentration recorded on days 2 to 4 (— —). The peak concentration providing the best discrimination was greater than or equal to 210 mg/litre. (Published with permission from the *British Journal of Surgery*.)

Serum amyloid A

Prelimary studies measuring this substance hold considerable promise that it may be the most useful single marker of disease severity.

Trypsinogen activation peptide

The activation of trypsinogen releases trypsin and a small peptide (trypsinogen activation peptide) that passes unchanged in the urine, where its level can be used as a marker of severity. It has now been employed in two clinical studies with the promise of this being a valuable step forward in assessment of severity.

Clinical management

Pain is usually treated with intramuscular pethidine or buphrenorphine; intravenous benzodiazepines may also be required in severe cases. The effect of the combination of intravenous midazolam and pethidine can be particularly difficult to predict and the combination must be used with great care. Morphine has a strong spastic effect on the sphincter of Oddi and is contraindicated. Haloperidol is useful in managing the agitation of alcoholics with acute pancreatitis.

Correction of hypovolaemia requires the rapid infusion of high-volume electrolyte solutions. There is a tendency to underestimate fluid requirements in the initial 12 h of treatment and monitoring the central venous pressure is essential.

Catheterization of the bladder to monitor urine output should be done immediately and a minimum flow of 30 ml/h obtained. Nasogastric aspiration is beneficial and both this and urinary catheterization can be discontinued at an early stage if the disease proves to be mild. In such patients there is little justification for the routine use of antibiotics or H_2-receptor antagonists as nearly all of them improve within a few days. If gallstones have been identified by ultrasound scanning, laparoscopic or open cholecystectomy should be done in the same admission to minimize the risk of recurrent attack. In older and infirm patients, ERCP sphincterotomy alone is

considered a satisfactory alternative.

The severely ill patient

Those who are graded as having severe acute pancreatitis are usually particularly ill at the time of admission or within 24 h and warrant high-dependency or intensive care therapy. Monitoring for system failures and biochemical or haematological abnormalities is now routine. An algorithm for suggested steps in the management of severe acute pancreatitis (Fig. 5) is based on the United Kingdom National Guidelines published in 1998.

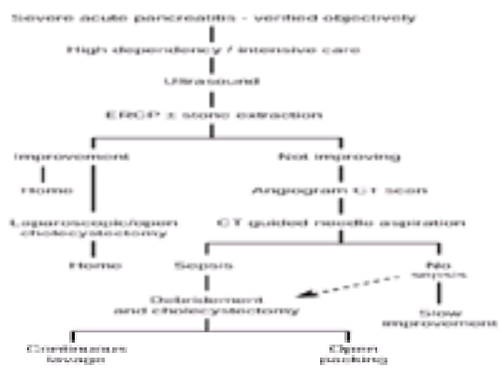

Fig. 5 Summary of management of acute pancreatitis.

In addition to monitoring individual systems and providing support, early ultrasound scan and liver function test data may indicate consideration for diagnostic and therapeutic ERCP. Two controlled studies and a large body of clinical data indicate that endoscopic sphincterotomy in those with severe gallstone acute pancreatitis provides a significant therapeutic advantage. It has also been the author's experience that rapid improvement can occur in some of the most severely ill patients soon after early endoscopic duct clearance. The greatest advantage is in jaundiced patients with or without cholangitis. Intravenous antimicrobial therapy (as a single IV bolus) is recommended as a prophylactic measure during sphincterotomy and the drugs most widely used include the second-generation cephalosporins.

The prophylactic administration of parenteral imipenem to patients with objectively graded severe acute pancreatitis throughout the first 2 weeks of illness has been found to be of value in several studies, possibly because of its high tissue penetration. We advocate a more selective use of antibiotics depending on specific indications such as proven cholangitis, the use of ERCP, and evidence of infection by culture of organisms at percutaneous fine needle aspirate in patients with infected necrosis. Fungal infection can be a problem in long-term management of those with severe acute pancreatitis when broad-spectrum antibiotics are given for more than 10 days. The mortality of candida septicaemia was about 50 per cent in two recently published experiences.

Another approach has been to strive to reduce the concentration of intestinal organisms by selective gut decontamination combined with intravenous antimicrobial therapy, but this has only been done in one study of over 100 patients in The Netherlands; Gram-negative and fungal sepsis was reduced.

The reason for this attention to the lowering of the risk from sepsis is the belief that transudation of organisms from the transverse colon into the peripancreatic necrotic tissue is a life-threatening complication of this disease. While this is usually a late complication after the first week of illness, it has been claimed that it may occur much earlier. The majority expert view currently recommends parenteral use of broad spectrum antibiotics for the initial 10 to 14 days of illness in clinically severe disease but the randomized studies each contain fewer than 80 patients and the case has yet to be proved that parenteral antibiotics confer benefit other than lowering the incidence of pancreatic infection. Neither mortality nor morbidity have been significantly reduced.

Early nasoenteral feeding

The biggest single change in therapy of severe acute pancreatitis has been the revolutionary advocacy of early nasojejunal feeding previously considered impossible due to the supposed stimulatory effect on the inflamed pancreas. Clinical studies began with the experimental evidence (in acute pancreatitis) of very early loss of small bowel mucosal integrity and the possible protective effect of a peptide and carbohydrate feed. Encouragement was also provided by data from comparative studies of patients fed by intravenous and percutaneous jejunostomy in United States trauma units, which showed clinical advantage in terms of shorter intensive care and total hospital stay for enteral feeding.

A randomized study of 38 objectively graded patients with severe acute pancreatitis found that it was both safer and cheaper (fewer infective complications and £30 versus £100 per day) to immediately use nasojejunal feeding than parenteral feeding. Although there was no mortality difference, other trials demonstrated that early nasojejunal feeding was better than parenteral feeding in terms of enhanced antioxidant capacity, decreased levels of endotoxin antibodies, and a faster resolution of markers of systemic inflammatory response.

Clinical studies from Brussels and Glasgow involving almost 100 patients with severe acute pancreatitis have verified the practicality of early nasoenteral feeding within 48 h of onset, and this approach has now become standard practice in many hospitals. This is a rare instance of a clinical advance proving cheaper than its predecessor.

The role of surgery

Patients who develop infection in the necrotic tissue around the pancreas and in the pancreas itself require open surgical debridement of the necrotic tissue by a combination of gentle finger and forceps dissection; the necrotic material will not readily drain along percutaneously introduced tubes. After removal of the infected tissue there are two options. One is to establish a postoperative lavage system, which may necessitate up to three inflow and three outflow drains because of the tendency for retroperitoneal extension of the infected necrosis down the paracolic gutters and upwards towards the diaphragm. Alternatively, if venous ooze of blood is a particular problem, packing of the abdominal cavity with large cotton packs wrapped in non-adhesive paraffin gauze, together with limited or non-closure of the abdominal wall, has been advocated. Such patients are invariably in intensive care on ventilator therapy. The packs should be changed at intervals of 48 to 72 h, with removal of any extension of infection or necrosis. Abdominal wall closure and postoperative lavage may be established after the second or subsequent operations.

Debate continues as to whether only patients with infected necrosis warrant such surgical intervention Proof of infection depends on either the presence of retroperitoneal gas radiologically or the results of fine needle aspiration guided by ultrasound or CT scanning. Most experts advocate that uninfected necrosis can be managed successfully without surgical intervention, while others are concerned about the limitations of methods for detecting infection, and would therefore widen indications for surgical intervention. Irrespective of approach, the proportions of patients coming to this surgery vary between 4 and 15 per cent of those with objectively graded severe acute pancreatitis. Recently a method of minimally invasive decompression of pus and subsequent removal of necrotic tissue has been pioneered utilizing either a left flank retroperitoneal approach (80 per cent) or right anterior drainage (20 per cent); this is a promising development since it is less traumatic than open operation. A randomized comparison of the standard open approach with this minimally invasive one is now necessary to strive to identify which patients derive most benefit from each approach. Retroperitoneal pancreatic necrosectomy involves following the route of a radiologically placed guidewire at the time of fine needle aspiration (FNA). The track is dilated to 10 mm diameter using the Amplatz dilator system.

Gallstone eradication

Cholecystectomy and common duct clearance are indicated within the same admission in patients with stones or biliary sludge. The later complication of pancreatic pseudocyst has a higher morbidity and mortality in those with biliary than alcohol-abuse pancreatitis. This is largely attributable to postoperative complications of sepsis and haemorrhage, which are much more common in the gallstone group, particularly if they have not had the gallbladder removed at the primary operation. Thus all patients with severe acute pancreatitis coming to surgery should have a cholecystectomy, especially as it is often impossible to identify small stones/sand by either ultrasonography or palpation at operation.

Pancreatic pseudocyst

This condition is probably overdiagnosed. The 1992 Atlanta conference on nomenclature agreed that the fibrous wall around a pseudocyst took approximately 4 weeks to develop from the onset of acute pancreatitis. It recommended that the term 'acute fluid collection' be used at an earlier stage in the disease process because these frequently disappear spontaneously. Even established pseudocysts can spontaneously resolve in around 50 per cent of patients. For those not resolving, synthetic somatostatin therapy (octreotide) given subcutaneously three times a day may be helpful in speeding resolution although there is no evidence base for this view as no controlled study has yet been carried out.

Percutaneous aspiration alone invariably results in recollection of the fluid quite rapidly, while infection is potentially associated with long-term percutaneous drainage. In younger patients pseudocysts are best dealt with by internal surgical drainage to the stomach or a defunctioned Roux loop of jejunum. Cystogastrostomy has been done laparoscopically, an approach that can also be used combined with a cholecystectomy. Pancreatic pseudocyst most commonly occurs in the lesser sac and often represents a closed pancreatic fistula, as a breach in the main or major pancreatic duct can be demonstrated at ERCP. This investigation is potentially hazardous as it may lead to the introduction of infection and should not be done without an appropriate antimicrobial in the injection fluid or planned endoscopic or surgical drainage. In recent years endoscopic transgastric/duodenal decompression of pancreatic pseudocysts has been increasingly employed but the evidence that this is superior to surgery has yet to be provided. Likewise the alternative of the placement of a plastic stent along the main pancreatic duct across the breach in the duct has yet to be critically assessed against other approaches, but it can be very useful and is sometimes combined with transgastric/duodenal stent decompression. Endoscopic ultrasound increases safety by accurately identifying the optimum site of pseudocyst drainage.

Pancreatic ascites

This condition occurs either when a pancreatic pseudocyst spontaneously decompresses into the gut or peritoneal cavity or a major pancreatic duct disrupts after trauma (Fig. 1) or pancreatitis, with escape of pancreatic juice into the peritoneal cavity. Treatment comprises either a combination of intravenous nutrition and octreotide therapy or surgical excision of the disconnected segment of pancreas. Success has also been obtained with intrapancreatic main-duct stents placed endoscopically.

Rare complications

Rarer complications of severe acute pancreatitis include splenic vein thrombosis and subcutaneous fat necrosis. The latter condition can mimic erythema nodosum.

Further reading

Acosta JM, Ledesma CL (1974). Gallstone migration as a cause for acute pancreatitis. *New England Journal of Medicine* **290**, 480–7.

Balthazar EJ *et al*. (1990). Acute pancreatitis: value of CT in establishing diagnosis. *Radiology* **156**, 767–72.

Beger HG (1989). Surgical management of necrotising pancreatitis. *Surgical Clinics of North America* **69**, 529–69.

Bradley EL (1987). Management of infected pancreatic necrosis by open drainage. *Annals of Surgery* **206**, 542–50.

Bradley EL (1993). A clinically based classification system for AP: Atlanta International Symposium summary. *Archives of Surgery* **128**, 586–90.

Carter CR *et al*. (2000). Percutaneous necrosectomy and sinus tract endoscopy in the management of infected pancreatic necrosis: an initial experience. *Annals of Surgery* **232**, 175–80.

Corfield AP *et al*. (1985). Prediction of severity in acute pancreatitis: prospective comparison of three prognostic indices. *The Lancet* **ii**, 403–7.

Foulis AK (1982). Morphological study of the relation between accidental hypothermia and acute pancreatitis. *Journal of Clinical Pathology* **35**, 1244–8.

Gudgeon AM *et al*. (1990). Trypsinogen activation peptide assay in the early prediction of severe AP. *The Lancet* **i**, 4–8.

Heath DI *et al*. (1993). Role of interleukin-6 in mediating the acute phase protein response and potential as an early means of severity assessment in acute pancreatitis. *Gut* **34**, 41–5.

Imrie CW *et al.* (1977). Arterial hypoxia in acute pancreatitis. *British Journal of Surgery* **64**, 185–8.

Imrie CW *et al*. (1978). Parathyroid hormone and homeostasis in acute pancreatitis. *British Journal of Surgery* **65**, 717–20.

Imrie CW *et al*. (1978). A single centre double blind trial of Trasylol therapy in primary acute pancreatitis. *British Journal of Surgery* **65**, 337–41.

Kelly TR (1976). Gallstone pancreatitis: pathophysiology. *Surgery* **80**, 488–92.

Kivisaari L *et al*. (1984). A new method for diagnosis of acute haemorrhagic necrotising pancreatitis using contrast enhanced CT. *Gastrointestinal Radiology* **9**, 27–30.

Larvin M, McMahon MJ (1989). APACHE II score for assessment and monitoring of AP. *The Lancet* **ii**, 201–4.

Leese T *et al*. (1987). Multicentre clinical trial of low volume fresh frozen plasma therapy in acute pancreatitis. *British Journal of Surgery* **74**, 907–11.

London NJM *et al*. (1989). Contrast enhanced abdominal computed tomography scanning and prediction of severity of acute pancreatitis: a prospective study. *British Journal of Surgery* **76**, 268–72.

Lucarotti ME, Virjee J, Alderson D (1993). Patient selection and timing of dynamic computed tomography in acute pancreatitis. *British Journal of Surgery* **80**, 1393–5.

Luiten EJ (1995). Controlled clinical trial of selective decontamination for the treatment of severe acute pancreatitis. *Annals of Surgery* **222**, 57–65.

McKay CJ *et al*. (1999). High early mortality rate from acute pancreatitis in Scotland. 1984–1995. *British Journal of Surgery* **86**, 1302–6.

Murphy D, Pack A, Imrie CW (1980). The mechanisms of arterial hypoxia occurring in AP. *Quarterly Journal of Medicine* **49**, 151–63.

Neoptolemos JP *et al*. (1987). Acute cholangitis in association with acute pancreatitis: incidence, clinical features, outcome and the role of ERCP and endoscopic sphincterotomy. *British Journal of Surgery* **74**, 1103–6.

Pederzoli P *et al*. (1993). A randomised multicenter clinical trial of antibiotic prophylaxis of septic complications in acute necrotizing pancreatitis with imipenem. *Surgery, Gynaecology and Obstretics* **176**, 480–5.

Pickford IR, Blackett RL, McMahon MJ (1977). Early assessment of severity of acute pancreatitis using peritoneal lavage. *British Medical Journal* **2**, 1377–9.

Poulakkainen P *et al*. (1987). C-reactive protein (CRP) and serum phospholipase A2 in the assessment of severity of AP. *Gut* **28**, 764–71.

Ranson JHC *et al*. (1974). Prognostic signs and the role of operative management in acute pancreatitis. *Surgery, Gynecology and Obstetrics* **139**, 69–81.

Viedma JA *et al.* (1992). Role of interleukin-6 in acute pancreatitis. Comparison with C-reactive protein and phospholipase A. *Gut* **33**, 1264–7.

Wilson C *et al.* (1989). C-reactive protein, antiproteases and complement factors as objective markers of severity of AP. *British Journal of Surgery* **76**, 177–81.

Wilson C *et al*. (1990). Prediction of outcome in acute pancreatitis: a comparative study of APACHE II, clinical assessment, and multiple scoring systems. *British Journal of Surgery* **77**, 1260–4.

Windsor AC *et al*. (1998). Compared with parenteral nutrition, enteral feeding attenuates the acute phase response and improves disease severity in acute pancreatitis. *Gut* **42**, 431–5.

14.18.3.2Chronic pancreatitis

P. P. Toskes

Introduction

Patients with chronic pancreatitis usually come to medical attention with abdominal pain or maldigestion (diarrhoea, steatorrhoea, weight loss), but the frequency of chronic pancreatitis has been underestimated because of inadequate investigation of these symptoms. The realization that impaired pancreatic exocrine function can occur without obvious dilatation of the main duct, that is 'small-duct disease', has greatly influenced management. Symptomatic variability and the many causes of this disease have made its classification difficult.

There are three forms of chronic pancreatitis now recognized: (1) chronic calcifying, (2) chronic obstructive, and (3) chronic inflammatory. Alcohol abuse and/or malnutrition are the most common causes of the calcifying type. Obstruction of the main pancreatic duct with secondary fibrosis in that part of the pancreas proximal to the obstruction leads to the obstructive type. Chronic inflammatory pancreatitis is not well characterized and many patients with chronic pancreatitis of unknown cause fall into this group. Often irreversible changes occur in the gland, making a cure improbable. Nevertheless, the chief complaints of pain and/or maldigestion can be effectively treated.

Histologically, in advanced stages of chronic pancreatitis, the gland may be fibrotic and calcified and the main duct may be dilated. Inflammation and sclerosis with progressive damage to the acini and ducts are the histological hallmarks of chronic pancreatitis. Islet cells are usually lost more slowly than the exocrine part, so that diabetes is a late feature.

Aetiology

Table 1 classifies chronic pancreatitis into a number of different conditions associated with this disease. Chronic alcoholism and cystic fibrosis are the most frequent causes in adults and children, respectively. Gallstones rarely cause chronic pancreatitis because a cholecystectomy is almost always performed after the first or second attack of acute pancreatitis related to gallstones, after which the pancreas recovers. Hypertriglyceridaemia may cause chronic as well as acute pancreatitis. Some patients with chronic pancreatitis may have suffered autoimmune pancreatitis: they have had enlargement of the pancreas, strictures of the pancreatic duct, autoantibodies in the serum, elevated plasma immunoglobulins, and histology showing a dense lymphocytic infiltrate. A few have responded to steroid therapy. Tropical pancreatitis (Africa and Asia) is characterized by calcific disease, glucose intolerance, and infrequent pain. Pancreatic exocrine impairment occurs in patients with haemochromatosis and a_1-antitrypsin deficiency, but the pancreatic disease is usually asymptomatic. Secondary pancreatic exocrine insufficiency may occur after gastric surgery, leading to postcibal (postprandial) asynchrony; usually the maldigestion is not very severe. Similarly, the acid hypersecretion associated with gastrinoma may irreversibly inactivate lipase, causing steatorrhoea. Hereditary pancreatitis and developmental anomalies leading to pancreatitis are discussed later.

Idiopathic chronic pancreatitis remains controversial and may account for up to 20 per cent of cases of chronic pancreatitis, depending on the population. Many patients with idiopathic pancreatitis present solely with unexplained abdominal pain and no evidence of maldigestion. These patients have small-duct disease, often without overt radiographic abnormalities. Direct intubation (hormone stimulation) tests are essential to identify this condition, although they are not universally carried out. Endoscopic retrograde cholangiopancreatography (**ERCP**), which is often used to diagnose chronic pancreatitis, may miss up to 30 per cent of patients with chronic pancreatitis who have abnormal hormone-stimulation tests. One questions how many patients with unexplained abdominal pain may indeed suffer from small-duct chronic pancreatitis! Some will be thought to have non-ulcer dyspepsia; others with idiopathic pancreatitis may present at an older age with painless diarrhoea, steatorrhoea, and secondary diabetes mellitus and often have pancreatic calcification.

Several investigators have documented mutations of the cystic-fibrosis, transmembrane-conductance regulator (**CFTR**) gene which functions as a cyclic AMP-regulated chloride channel in idiopathic chronic pancreatitis. More than 900 mutant alleles of the *CFTR* gene have been identified. Attempts to elucidate the relationship between the genotype and pancreatic manifestations have been hampered by the number of mutations. Two reports have shown *CF* (cystic fibrosis) gene mutations in 13 to 40 per cent of patients with idiopathic chronic pancreatitis (the observed frequency of a single *CFTR* mutation was 11 times greater than expected); moreover, the frequency of two mutant alleles was increased 80-fold. Most of these patients were adults with chronic pancreatitis, none of whom had any clinical evidence of pulmonary disease; the results of sweat chloride measurements were not diagnostic of cystic fibrosis. A further study examining all known *CFTR* mutations noted abnormalities in 55 per cent of 16 patients, 14 of whom had idiopathic chronic pancreatitis. Some of these patients with either one or two mutations had evidence of defective CFTR-mediated ion transport in nasal epithelium. It is not yet clear whether these CFTR abnormalities are primarily responsible for chronic pancreatitis or whether they are, at least in some cases, unrelated.

Pathophysiology

Although alcohol-induced chronic pancreatitis has been studied extensively, it remains uncertain as to whether the biochemical and histological lesions are caused by a reduced secretion of pancreatic-stone protein or alcohol toxicity. Ingestion of alcohol may decrease the stone protein secretion below a critical level, allowing calcium and other secretory components to precipitate and obstruct pancreatic ductules. On the other hand, alcohol may cause abnormalities in acinar cells, leading to an imbalance of proteases and their cognate inhibitors, resulting in the initiation of a necroinflammatory process. In tropical pancreatitis, a combination of protein deficiency and a dietary toxin that occurs in cassava or sorghum may be responsible. A primary defect in the permeability of the ductal epithelium to electrolytes in patients with cystic fibrosis reduces secretory fluxes, so that the hyperconcentrated proteinaceous fluid precipitates and obstructs the pancreatic ducts. The pathophysiology of the other causes of chronic pancreatitis is not understood.

Although it is widely believed that when a patient develops their first attack of acute alcoholic pancreatitis they have already sustained chronic damage to the pancreas, contemporary investigations indicate that some individuals who do not abuse alcohol regularly develop acute pancreatitis after ingesting uncommonly large quantities of alcohol (binge-drinking).

Incidence

The exact prevalence and incidence of chronic pancreatitis is unknown. Most opinions are based on clinical experiences, which vary greatly. The prevalence in autopsy studies varies from 0.04 to 5.0 per cent. The only prospective study (Copenhagen Pancreatic Study) found a prevalence of 26.4 cases per 100 000 population and an incidence of 8.2 new cases per 100 000 per year. However, this study mainly reflects alcohol-induced pancreatitis.

Clinical features

Abdominal pain is the cardinal symptom of chronic pancreatitis; its pattern, severity, and frequency vary considerably. Whereas the pain of acute pancreatitis is often located in the epigastrium and bores through to the back, the pain of chronic pancreatitis has no characteristic features and may be constant or intermittent. Eating

often increases the severity of the pain, resulting in the avoidance of food and a subsequent weight loss. The pain may be mild, requiring no therapy, or severe, leading to the frequent use of analgesics and narcotic addiction.

Patients with abdominal pain may develop steatorrhoea and/or diarrhoea, or the abdominal pain may remain their primary symptom. Approximately 15 per cent of patients never suffer with abdominal pain but present with steatorrhoea, diarrhoea, and weight loss. In those who only have abdominal pain, there are few physical findings except for abdominal tenderness and mild pyrexia. There is a marked disparity between the severity of the abdominal pain and the physical findings.

Signs and symptoms of liver disease may be present in patients with maldigestion and weight loss due to alcohol-induced pancreatitis. Clinically apparent deficiencies of fat-soluble vitamins or vitamin B_{12} are un-common.

Diagnosis

Computed tomographic (**CT**) scans may reveal diffuse enlargement of the pancreas and, occasionally, a pseudocyst (Fig. 1). Ultrasonography may reveal calcification and dilatation of the pancreatic duct (Fig. 2); calcification may also be seen on plain abdominal radiographs (Fig. 3). Blood tests rarely contribute to a diagnosis of chronic pancreatitis. The plasma activities of pancreatic enzymes (amylase, lipase, trypsin) are usually normal except in patients who have a pseudocyst of the pancreas. There may be evidence of cholestasis (elevated alkaline phosphatase, elevated bilirubin) caused by inflammatory reactions around the common bile duct. Some patients with severe disease may have raised fasting blood-glucose levels.

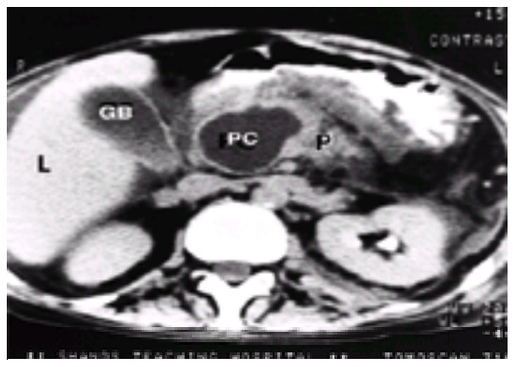

Fig. 1 CT scan demonstrating a pseudocyst (PC) in the head of the pancreas, diffuse enlargement of the pancreas (P), a normal liver (L), and normal gallbladder (GB).

Fig. 2 Ultrasonogram of chronic pancreatitis. The large closed arrow points to pancreatic calcification; the small closed arrow shows a dilated pancreatic duct; and the open arrow identifies the splenic vein.

Fig. 3 Plain film of the abdomen showing diffuse pancreatic calcification; the arrow points to one of the calcified areas.

Table 2 lists selected tests of pancreatic function and structure, but abnormalities of function generally precede abnormalities in structure. The most sensitive tests are at the top of the table, the least sensitive at the bottom. Currently, the most accurate means of detecting chronic pancreatitis is a combination of a hormone-stimulation test and ERCP. As many as 30 per cent of patients with chronic pancreatitis may have a normal ERCP but an abnormal hormone-stimulation test. Occasionally the converse will be true. The two significant causes of a false-positive (abnormal) ERCP are normal ageing and recent acute pancreatitis. Ageing does not appear to affect the hormone-stimulation test. Simple, non-invasive tests (bentiromide, pancreolauryl, trypsin) are not sensitive and are used to confirm the clinical impression. The same can be said for radiography other than ERCP.

Almost any test listed in Table 2 will identify patients with severe disease, but a hormone-stimulation test is often needed to diagnose those with abdominal pain only. Although it is generally accepted that a hormone-stimulation test is the most sensitive way to detect mild to moderate impairment of exocrine function, comparisons of the true 'gold standard' (histological examination of the pancreas) with any pancreatic test have been lacking until recently. The hormone-stimulation test correlates with the pancreatic histological findings obtained at surgery. A recent study found that the most discriminatory function was the maximum bicarbonate concentration, followed by volume and amylase output. A significant correlation was found between pancreatic function and histology. In 29 of the patients with histologically confirmed pancreatitis, the cholecystokinin–secretin test had a sensitivity of 79 per cent and ERCP 66 per cent. A simple, inexpensive test that has a sensitivity and specificity approaching that of hormone stimulation is needed. A cost-effective approach to the evaluation of patients suspected of having chronic pancreatitis would first be to use a simple non-invasive test like serum trypsin (or bentiromide) and to initiate pancreatic enzyme therapy if the result was abnormal. However, if this first-order test was normal, the next step would be to perform a hormone-stimulation test, and finally, if needed, an ERCP.

Relatively recent techniques for imaging the pancreas include magnetic resonance cholangiopancreatography (**MRCP**) and endoscopic ultrasonography (**EUS**). MRCP provides a satisfactory morphological assessment of the main pancreatic duct. However, this technique does not provide detailed imaging of the secondary ducts or even the main pancreatic duct if it is small. MRCP has been utilized successfully in elderly patients where the risk of pancreatitis from ERCP has influenced the clinician to not perform an ERCP. Endoscopic ultrasonography provides a detailed assessment of both the pancreatic duct and parenchyma. A total of nine abnormal features have been defined, and more than three criteria have been required in most studies for a diagnosis of chronic pancreatitis. EUS has replaced

ERCP as a diagnostic modality in selected patients with suspected chronic pancreatitis. What is not yet clear is how sensitive and specific EUS is in defining abnormalities in patients considered to have early or mild chronic pancreatitis. In patients in whom only the hormone-stimulation test has been found to be abnormal will EUS be helpful in diagnosing chronic pancreatitis, but its sensitivity and specificity require further study.

Management

The cornerstone of the medical management of chronic pancreatitis is the use of pancreatic enzyme formulations. The principles of therapy are similar for treating pain or steatorrhoea. A potent enzyme formulation must be used to ensure that the relevant enzymes (protease for pain, lipase for steatorrhoea) escape destruction by gastric acid and reach the duodenum.

Abstinence from alcohol is recommended. The diet should be moderate in fat (30 per cent), high in protein (24 per cent), and low in carbohydrate (40 per cent). Non-narcotic analgesics are the pain-relieving medications of choice.

To date, three controlled trails have shown that pancreatic enzymes decrease abdominal pain in some patients with chronic pancreatitis. Pain relief was obtained in 75 per cent of the patients evaluated. Those most likely to respond have small-duct disease, that is to say a minimal to moderate impairment of exocrine function (abnormal hormone-stimulation test, minimal abnormalities on ERCP, normal fat absorption) (Fig. 4). Patients with severe (large-duct) disease (abnormal hormone-stimulation test, marked abnormalities on ERCP, steatorrhoea) (Fig. 5) do not respond well to enzyme therapy for pain. These clinical observations fit well with findings in experimental animals and humans, which demonstrate the negative-feedback regulation of pancreatic secretion controlled by the amount of proteases within the proximal small intestine. Treatment comprising eight tablets or capsules of a potent, non-enteric-coated enzyme preparation should be given at mealtimes and at bedtime, with appropriate adjuvant therapy (Table 3). Enteric-coated preparations are not the prepar-ations of choice because they often release their proteases in the jejunum or ileum rather than the duodenum, thus failing to deliver to the feedback-sensitive segment of the intestine; these preparations may also cause acute colonic disease and rupture.

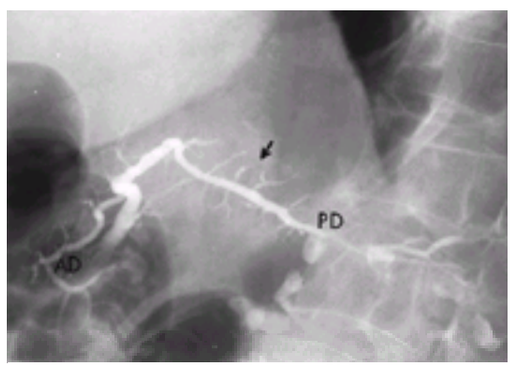

Fig. 4 Early ERCP changes of chronic pancreatitis. A non-dilated main pancreatic duct (PD) and accessory duct (AD) with mild dilatation and clubbing of the side branches (arrow) are shown.

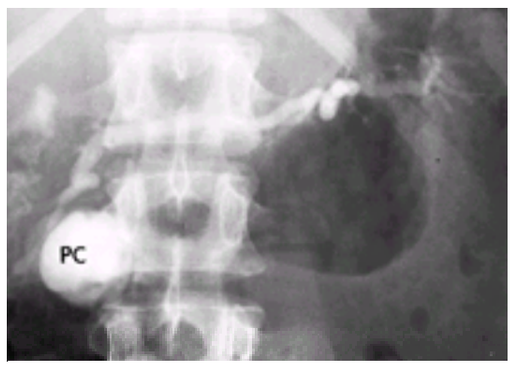

Fig. 5 ERCP showing a dilated main pancreatic duct with a communicating pseudocyst (PC).

Figure 6 outlines an approach to patients with abdominal pain thought to be caused by chronic pancreatitis. After other causes of abdominal pain have been excluded, an ultrasonography should be obtained. If no pseudocyst or mass is found, a hormone-stimulation test should be performed; this will invariably be abnormal in patients with abdominal pain secondary to chronic pancreatitis. A 4-week trial of pancreatic enzymes (with adjuvant) is indicated, as described above. If pain is not relieved, ERCP is appropriate to characterize the pancreatitis as small- or large-duct disease, and possibly to define the surgical approach.

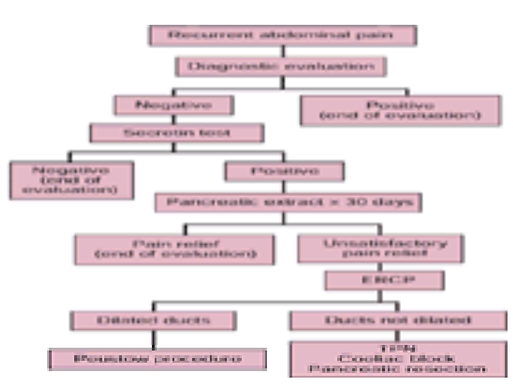

Fig. 6 Approach to the management of chronic pancreatitis and abdominal pain.

If there is large-duct disease (diameter of the main pancreatic duct greater than 7 mm), a lateral pancreaticojejunostomy (Peustow procedure) should be performed. Immediate pain relief occurs in 80 per cent of patients, with satisfactory pain relief sustained in about 50 per cent at 1 to 3 years' follow-up. If the ducts are not significantly dilated, most patients can eventually have their pain controlled by adjusting the enzyme and adjuvant therapy; for example, substitution of a proton-pump inhibitor for an H_2-receptor antagonist, total parenteral nutrition with no food orally for several weeks, or by performing a nerve block. It is now rare for a major pancreatic resection to be undertaken to control pain. In some preliminary controlled trials, octreotide in doses up to 200 µg three times daily given subcutaneously has been effective in reducing pain in patients with severe chronic pancreatitis. Whether ductal decompression or major resection are performed, enzyme therapy for enhancing digestion should be given.

Endoscopic therapy for the pain of chronic pancreatitis has been disappointing. This therapy has included dilatation or stenting of duct strictures, removal of calculi, and treatment of biliary obstruction. With the exception of acute biliary decompression, none of these therapies has been shown to be effective. Complications such as bleeding, sepsis, pancreatitis, and perforation have occurred after stent placement; moreover, stents can induce progressive ductal changes similar to the

abnormalities seen in chronic pancreatitis.

Steatorrhoea in chronic pancreatitis is a late finding and does not occur until lipase secretion is reduced by 90 per cent. With eight conventional or three enteric-coated enzyme tablets or capsules (Table 3), control of steatorrhoea and diarrhoea, and weight gain, can be readily achieved, even though some steatorrhoea persists. Formulations containing 25 000 units or more of lipase have recently been associated with the occurrence of colonic strictures in patients with cystic fibrosis who were taking large doses of these high-potency preparations. In the United States, all pancreatic enzyme preparations with more than 20 000 units of lipase per capsule have been taken out of clinical use.

Decreasing the amount of long-chain triglycerides in the diet and/or adding medium-chain triglycerides (which do not require pancreatic lipase for absorption) should decrease the steatorrhoea and enhance weight gain and energy.

Complications

Table 4 lists the structural and metabolic complications of chronic pancreatitis. Inflammatory masses are common. Ultrasonography and computed tomography greatly assist in discriminating phlegmon from a pseudocyst and from an abscess. The management of pseudocysts is currently being re-evaluated. Most clinicians have used drainage if the pseudocyst persists for longer than 7 weeks. However, the ability of pseudocysts to undergo late resolution may have been underestimated and the incidence of serious complications exaggerated. In patients with minimal symptoms who have pseudocysts and do not actively abuse alcohol, a mature pseudocyst that has benign radiological appearances should be observed: nine out of ten such pseudocysts will resolve without complications.

Pancreatic ascites occurs when there is a rent in the pancreatic duct or a leaking pseudocyst. The amylase content in the ascitic fluid is extraordinarily high, averaging 20 000 IU/l. True pancreatic ascites should be distinguished from 'reactive ascites' in patients with pancreatitis. In reactive ascites the amylase content of the fluid while increased, is not nearly as high as in the pancreatic ascites. Pancreatic stimulation should be avoided in patients with pancreatic ascites, they should receive total parenteral nutrition and no food by mouth; proton-pump inhibitors or H_2-antagonists will reduce pancreatic stimulation resulting from gastric acid release into the duodenum. Surgery may be needed if the ascites persists after several weeks of conservative therapy. An ERCP may be needed to determine the site of duct leakage.

Although obstruction of the common bile duct is common, it may be temporary, due to the resolution of inflammation. Biliary obstruction due to fibrosis of the pancreas rarely leads to cholangitis. Conservative management is justified unless the alkaline phosphatase level remains very high or cholangitis develops. In a few patients, the obstruction may require surgical relief by anastomosing the dilated common bile duct to the duodenum or jejunum.

Gastrointestinal bleeding may arise from portal hypertension associated with splenic vein thrombosis caused by inflammation of the tail of the pancreas. Bleeding may also occur if a pseudocyst erodes into the duodenum or from a pseudoaneurysm within the wall of a pseudocyst. However, the most common cause of bleeding in chronic pancreatitis is a related duodenal ulcer or alcohol-induced gastritis.

Up to 30 per cent of patients with chronic pancreatitis have impaired glucose tolerance. Although pancreatic diabetes is usually manageable, destruction of glucagon-containing cells may render hypoglycaemia more likely. Diabetes retinopathy occurs as often in pancreatic diabetes as in other types of diabetes mellitus. Retinopathy due to a zinc or vitamin A deficiency may occur.

Cobalamin (vitamin B_{12}) malabsorption is common in chronic pancreatitis, but clinical vitamin B_{12} deficiency is rare. It is caused by the failure to release free cobalamin by proteolysis of transcobalamin complexes. The exogenous administration of pancreatic enzymes corrects the maldigestion.

Hereditary and familial diseases

Hereditary pancreatitis

Familial or hereditary pancreatitis is an autosomal dominant disorder with approximately 80 per cent penetrance. About 1 per cent of all cases of chronic pancreatitis is due to hereditary pancreatitis. Often these patients present in childhood with recurrent acute pancreatitis which causes chronic pancreatitis and often pancreatic insufficiency. The lifetime risk of developing pancreatic cancer is quite high. The genetic abnormality is a defect in the cationic trypsinogen gene that maps to chromosome 7, which appears to interfere with the inactivation of trypsin after it is cleaved. These findings should prove to be revealing about the pathogenesis of idiopathic chronic pancreatitis.

Schwachmann's syndrome (pancreatic insufficiency and bone marrow disease)

This familial disorder affects the pancreas, bone marrow, and skeletal system. It is second only to cystic fibrosis as a cause of pancreatic insufficiency in infants. Unlike cystic fibrosis the sweat chloride test is normal. Neonates with this condition present with severe steatorrhoea. The associated neutropenia leads to frequent infections. The steatorrhoea is well treated by pancreatic enzymes, but severe skeletal defects result in dwarfism; there is a high lifetime risk of transformation into acute myeloid leukaemia for patients with Schwachmann's syndrome.

Isolated pancreatic enzyme deficiencies

Protease deficiencies result from a lack of enterokinase (proximal small-intestine mucosal enzyme) or trypsinogen. The addition of exogenous enterokinase to duodenal secretions will differentiate these two deficiencies; it will not activate duodenal secretions lacking trypsinogen. Both conditions respond to pancreatic enzyme therapy. Lipase and colipase deficiencies are also rare isolated deficiencies that cause steatorrhoea. Patients with these pancreatic lipase deficiencies retain a residual fat-absorbing capacity, presumably from the action of other lipases such as gastric lipase.

Developmental anomalies

Annular pancreas

A failure of the ventral and dorsal anlage of the pancreas to unite produces a ring of pancreatic tissue encircling the duodenum. This may lead to intestinal obstruction in the neonate or the adult. Non-specific symptoms of postprandial fullness, nausea, abdominal pain, and vomiting may be present for years before the diagnosis is made. The radiographs show fixed symmetrical dilatation of the proximal duodenum, with bulging of the recesses on either side of the annular band, effacement of the duodenal mucosa without obstruction of the mucosa, and accentuation of the findings in the right anterior oblique position. The differential diagnosis should include duodenal webs, tumours of the pancreas or duodenum, postbulbar peptic ulcer, Crohn's disease of the proximal intestine, and adhesions. Patients with an annular pancreas have an increased incidence of pancreatitis and peptic ulcer. Because of these and other intestinal complications, surgery may be necessary, even though the condition has been present for years. Retrocolic duodenojejunostomy is the procedure of choice, although some surgeons prefer a Billroth II gastrectomy with gastroenterostomy and vagotomy.

Pancreas divisum

Pancreas divisum is the most common, congenital, anatomical abnormality of the human pancreas. It occurs when the ventral and dorsal parts of the pancreas fail to fuse, so that pancreatic drainage is accomplished mainly through the accessory papilla. Current evidence indicates that this anomaly predisposes, albeit infrequently, to the development of pancreatitis. The combination of a pancreas divisum and a small accessory orifice could result in dorsal-duct obstruction. The challenge is to identify this subset of patients. Cannulation of the dorsal duct by ERCP is not as easy as cannulation of the ventral duct. Patients with pancreatitis and pancreas divisum demonstrated by ERCP should be treated conservatively, including pancreatic enzyme therapy. Many of them have idiopathic pancreatitis unrelated to the pancreas divisum and will respond well to pancreatic enzymes. Endoscopic or surgical intervention is indicated only when these methods fail. Surgical ductal decompression is indicated if marked dilatation of the dorsal duct can be demonstrated. However, the appropriate therapy for those patients without dilatation of the dorsal duct is not yet defined. It should be emphasized that the ERCP appearance of pancreas divisum (that is, a small-calibre ventral duct with an arborizing pattern) may be confused with an obstructed main pancreatic duct caused by a pancreatic tumour.

Further reading

Amann ST, Toskes PP (1998). Hyperlipidemia and pancreatitis. In: Berger HG, *et al*., eds. *The pancreas*, pp 311–16. Blackwell Science, Oxford.

Forsmark CE (2000). The diagnosis of chronic pancreatitis. *Gastrointestinal Endoscopy* **52**, 293–8.

Josephson S, Toskes PP (1996). Chronic pancreatitis: medical management. *Practical Gastroenterology* **20**, 6–22.

Lowenfels AB, *et al*. (1994). Prognosis of chronic pancreatitis: an international multi-center study. International Pancreatitis Study Group. *American Journal of Gastroenterology* **89**, 1467–72.

Saforkas GH, *et al*. (2000). Long-term results after surgery for chronic pancreatitis. *International Journal of Pancreatology* **27**, 131–42.

Somogyi L, *et al*. (2000). Synthetic porcine secretin is highly accurate in pancreatic function testing in individuals with chronic pancreatitis. *Pancreas* **21**, 262–5.

Vitab GJ, Sarr MG (1992). Selected management of pancreatic pseudocysts: operative versus expectant management. *Surgery* **III**, 124–30.

Walsh TN, *et al*. (1992). Minimal change chronic pancreatitis. *Gut* **33**, 1566–71.

Whitcomb DC, *et al*. (1996). Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. *Nature Genetics* **14**, 141–5.

14.18.3.3 Tumours of the pancreas

Julian Britton

Every year an average of 10 individuals from a Western population of 100 000 people will develop a tumour of the pancreas. Nine of the 10 with ductal adenocarcinoma will be dead a year later. Very few will survive 5 years. Two-thirds of patients are over the age of 65 years and there are slightly more men than women. The incidence of pancreatic cancer is increasing and we have little understanding of the cause. Smoking is a definite risk factor and about one in seven patients either have or develop diabetes mellitus before the cancer becomes evident. Hereditary chronic pancreatitis, which is rare and for which the causative gene has recently been identified, is associated with an increased risk of developing pancreatic cancer. It is less certain that chronic pancreatitis due to other causes is a risk factor.

Pathology

Malignant tumours of the exocrine pancreas are the commonest histological type. Endocrine tumours and benign tumours are rare (Table 1). In clinical practice, two-thirds of ductal adenocarcinomas occur in the head and uncinate process of the gland. All the adenocarcinomas grow locally and then spread to the regional lymph nodes and the liver. Perineural and vascular invasion are common features on histology. Occasional patients develop secondaries in the lung. Cystic adenocarcinoma is a separate histological type with a better prognosis. Cystic/papillary tumours occur exclusively in young women. They grow to a large size and rarely metastasize. Adenocarcinomas of the ampulla, the bile duct, and the duodenum are separate and distinct tumours with a better prognosis.

Insulinomas are the most common endocrine tumour of the pancreas. Most are small, benign, and hard to find, but 10 per cent are multifocal or malignant. Gastrinomas have usually spread to the liver by the time they are discovered. About one in five islet cell tumours do not secrete sufficient hormones to produce clinical symptoms.

Breast cancer occasionally metastasizes to the pancreas and lymphoma can cause obstructive jaundice.

Diagnosis

A short history of epigastric abdominal pain which radiates through to the back, jaundice, and weight loss are the cardinal symptoms of a cancer in the head of the pancreas. About half the patients also complain of itching. Examination reveals a wasted and jaundiced patient with an enlarged palpable gall bladder (Courvoisier's sign) (Fig. 1).

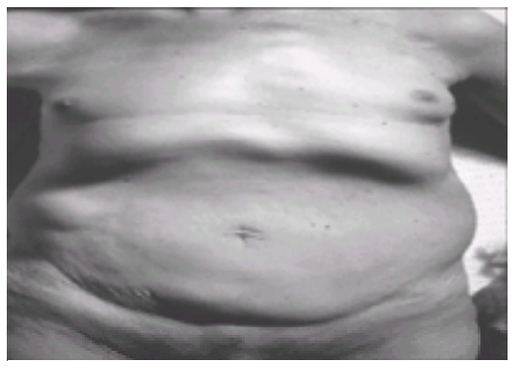

Fig. 1 This patient has cancer of the head of the pancreas and demonstrates Courvoisier's sign. The jaundice, the distended gall bladder, and the loss of weight are easy to see.

Not all patients have a palpable gall bladder and not all tumours in the head and uncinate process of the gland cause jaundice. These tumours and tumours in the body and tail of the gland present with pain and disturbance of digestion. The symptoms are often initially diagnosed as gastritis or peptic ulceration. Occasional patients present with acute pancreatitis or thrombophlebitis migrans. Pancreatic tumours are rarely palpable, and they present late in the course of the disease.

Insulinomas release insulin and present with intermittent hypoglycaemia. Excess gastrin leads to severe peptic ulceration and steatorrhoea and vasoactive intestinal polypeptide causes diarrhoea. Often, however, the clinical picture is subtle and the key to diagnosis is appropriate biochemical analysis once the diagnosis is entertained. Non-functioning endocrine tumours behave like ductal adenocarcinoma and are discovered on investigation.

Investigation

Ultrasonography

Ultrasound examination of the bile ducts and the pancreas is the first investigation when a tumour in the pancreas is suspected. Dilatation of the bile ducts is easy to see. Experienced ultrasonographers may be able to identify the level of the obstruction or to see a mass in the pancreas. More commonly part or all of the pancreas is obscured by gas in the stomach, duodenum, or transverse colon. When ultrasound is unhelpful and there is a reasonable suspicion of pancreatic pathology then computed tomography is mandatory (Fig. 2). This is usually the only way to identify cancer of the body and tail of the gland. Magnetic resonance imaging is an alternative and has the advantage of imaging the pancreas and also obtaining a pancreatogram and a cholangiogram.

Fig. 2 The body and tail of the pancreas are swollen and enlarged and there is streaking in the peripancreatic fat. This patient had a biopsy proven carcinoma of the pancreas, but it can be very difficult to tell pancreatic cancer from chronic pancreatitis on radiological grounds.

Endoscopic cholangio- and pancreatography, and percutaneous transhepatic cholangiography

Endoscopic cholangio- and pancreatography is the next preferred investigation. In cancer of the head of the gland, the bile duct and the pancreatic duct are usually narrow and irregular (the double-duct sign). In about three-quarters of these patients it is then possible to place a stent across the obstruction in the bile duct and so relieve the jaundice and pruritus. Percutaneous transhepatic cholangiography will also outline a biliary stricture but there is always a small risk of peritonitis because the needle, catheter and guide wire all cross the peritoneal cavity. However, it is slightly easier to place a biliary stent by this route (Fig. 3).

Fig. 3 Percutaneous transhepatic insertion of a biliary stent to relieve obstructive jaundice due to cancer of the pancreas. The catheter system to insert the self-expanding metal stent can be seen in the top left of the picture.

Biopsy

Biopsy of the duodenal wall is possible at endoscopic retrograde cholangiopancreatography and biliary strictures can be brushed over a guide wire for cytology. Alternatively a percutaneous biopsy can be obtained using computed tomography or ultrasonic guidance.

Computed tomography, magnetic resonance imaging, and angiography

Irregularity or narrowing of the arteries or veins around the pancreas on angiography indicate that a tumour is not resectable. Modern methods of imaging using three phase computed tomography or magnetic resonance imaging with suitable contrast agents can give the same information without a direct arterial injection. They are also more reliable at identifying liver metastases and a pancreatic resection is not justified if they are found.

Tumour markers

No serum marker has yet been found to be useful in the diagnosis of pancreatic cancer. During treatment an elevated CA 19-9 level will usually fall and this can be helpful in management.

Endocrine tumours

A high concentration of serum insulin in association with a low serum glucose is the key diagnostic feature of an insulinoma. Raised levels of other hormones will identify patients with other types of islet cell tumour. Although the tumours always develop in the pancreas or the duodenal wall most of them are less than a centimetre in size and they are hard to find. Ultrasound or computed tomography scanning and angiography are not worthwhile. Magnetic resonance imaging of the pancreas and endoscopic ultrasound are still developing and may be helpful. In most patients, once the diagnosis is established, a laparotomy by an experienced surgeon is the next best investigation. Most tumours can be seen or felt (Fig. 4), but when this is not the case intraoperative ultrasound with the probe applied directly to the pancreas will find the tumour. Transhepatic portal venous sampling, which will locate a tumour to a particular area within the pancreas, is best used after an initial unavailing laparotomy.

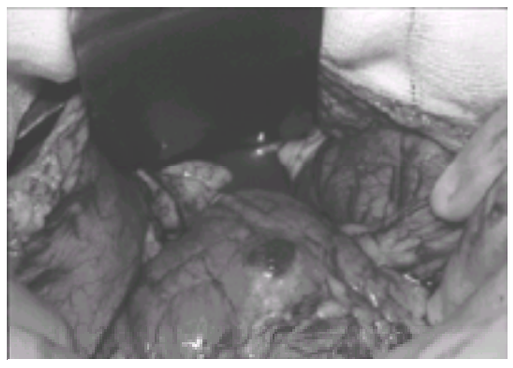

Fig. 4 The insulinoma is easily seen as a red swelling on the anterior surface of the body of the pancreas. It was a simple matter to enucleate the tumour. Leakage of pancreatic juice is rarely a problem afterwards.

Strategies for investigation

Cancer of the head of the pancreas commonly presents with jaundice. These patients need an urgent ultrasound examination which will confirm dilatation of the bile ducts. Either endoscopic retrograde cholangio-pancreatography or percutaneous transhepatic cholangiography should be performed next and the choice depends on

the available skills and local preference. In most patients, the jaundice can be relieved at the same time by inserting a stent across the biliary stricture. It is then necessary to identify those patients who would most likely benefit from surgery. Pancreatic resection is still a major procedure, despite a mortality rate now below 5 per cent. It is therefore preferable to offer radical surgery to patients under the age of 65 years, without significant other disease and whose tumour is less than 5 cm in diameter and apparently confined to the pancreas. Before surgery, this small group of patients should have portal venography carried out under computed tomography examination. In a few more patients this will show that the tumour is not resectable. The remaining patients should be offered an operation. Cancer of the body and tail of the pancreas is diagnosed on cross-sectional imaging and in most instances the computed tomography scan will also confirm that the tumour is not resectable.

Most patients with adenocarcinoma of the pancreas require palliative care. Biopsy of the pancreas or a liver metastasis is essential to confirm the diagnosis if chemotherapy is contemplated and is important for most other patients. Biopsy is not generally recommended in patients who are offered surgery because of the risk of seeding the tumour, despite the fact that a few are subsequently discovered to have a benign stricture.

Treatment

Palliation

Itching and intolerable pain are the two symptoms that require treatment. Itching rapidly improves once the bile duct is drained and the jaundice fades. Plastic stents last for an average of 4 months and can then be changed. Metal stents last rather longer. Most patients only survive a few months and never require a change of stent.

Pain from cancer of the pancreas is difficult to control. Early assistance should be sought from a specialist in palliative care. Regular opiate analgesia is usually required but a coeliac plexus block provides worthwhile pain relief in many cases. At operation 20 ml of absolute alcohol can be injected around the coeliac plexus to ablate sensory pathways. In other patients, the injection can be given by inserting the needle from the back alongside the lumbar spine and under the guidance of computed tomography.

A few patients will develop obstruction of the gastric outlet which will require a gastrojejunostomy. Surgical bypass of the biliary tract is now rarely required.

Radiotherapy has no part to play in palliation but many patients will wish to discuss chemotherapy with an oncologist. New drugs are being developed, and ideally most patients should be entered into a trial. About one in five patients will show some response to chemotherapy.

Surgery

Surgical resection of a localized pancreatic cancer is the only treatment that currently offers the possibility of long-term survival. Patients with apparently resectable tumours after investigation should therefore be offered an operation. A preliminary laparoscopy will identify patients with liver metastases which are too small to identify on scanning. Laparoscopic ultrasound in expert hands can identify involvement of the superior mesenteric or portal vein which also precludes resection. If the patient passes these tests the tumour should be assessed at a laparotomy. A trial dissection must show that the portal vein can be dissected off the tumour and that the superior mesenteric artery can be preserved. If this is the case then a pancreaticoduodenectomy (Whipple's operation) should be done (Fig. 5). The head and uncinate process of the pancreas along with the attached duodenum is removed. The antrum of the stomach is excised and the common bile duct is divided just above the duodenum. Some surgeons also remove the gall bladder. There are many techniques for restoring continuity. The pancreatic anastomosis is the most difficult and a pancreatic fistula is the cause of most postoperative complications. Occasionally a total pancreatectomy is required because the whole gland is diseased. In those rare cases in whom resection of a cancer in the body and tail of the gland is possible a splenectomy is always required as well. Even then tumour is often left around the origin of the splenic artery and postoperative radiotherapy to this area may be justified.

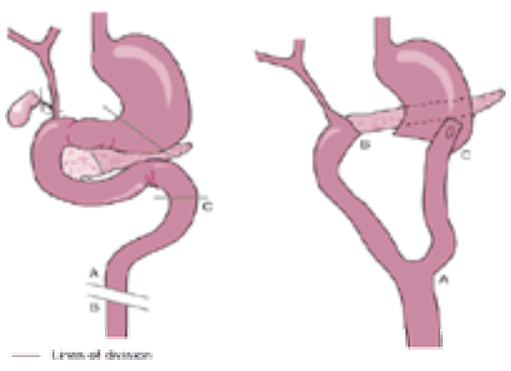

Fig. 5 The drawing on the left shows the structures that are removed in a classical pancreaticoduodenectomy (Whipple's operation). One method of restoring continuity is shown on the right.

Insulinomas are well circumscribed and can be enucleated from the pancreas by careful blunt dissection. Multifocal tumours require a pancreatic resection or even total pancreatectomy. Metastases in the liver can be enucleated, although medical management is often more appropriate.

Prognosis

Most insulinomas are benign and patients are cured when the tumour is removed. Malignant endocrine tumours metastasize and kill patients either because of their mass effect or as a consequence of intractable hypoglycaemia, diarrhoea, and hypokalaemia or peptic ulceration unless the hormonal effects can be effectively blocked.

Overall the prognosis of adenocarcinoma of the pancreas is very poor. Ninety per cent of patients are dead within 12 months of diagnosis. In the small proportion who undergo successful surgery, about one in five will survive 5 years. This will include some patients with lymph node metastases at the time of the original surgery. A significant proportion of the long-term survivors of resection will develop both endocrine and exocrine pancreatic insufficiency. This will require replacement with insulin and enzyme supplements but surprisingly, digestion is often remarkably normal: the diabetes is not usually difficult to control and most survivors can lead a near normal life.

Overall most patients with pancreatic adenocarcinoma die rapidly once the diagnosis is made but a few will survive for a long time after successful surgery. Improving outcome in the future will depend on understanding the cause of this disease.

Further reading

Carter DC and Trede M (1993). Tumours of the exocrine pancreas and periampullary region. In : Trede M, Carter DC, eds. *Surgery of the pancreas*, pp 383–544. Churchill Livingstone, Edinburgh. [A comprehensive review of cancer of the pancreas from a European perspective.]

Yeo CJ, Cameron JL (2000). Pancreatic cancer. In: Morris PJ, Wood WC, eds. *Oxford textbook of surgery*, 2nd edn, pp. 1785–1808. Oxford University Press, Oxford. [A similar review from America.]

14.19.1 Congenital disorders of the liver, biliary tract, and pancreas

J. A. Summerfield

Pathogenesis of congenital disorders of the biliary tract

During the fourth week of gestation the liver arises as a bud of cells (the hepatic diverticulum) from the ventral wall of the foregut. At about the eighth week of gestation a layer of liver precursor cells around the portal vein branches differentiate to form a sleeve, termed the ductal plate. This sleeve duplicates to form a double layer of cells which by twelve weeks is remodelled by dilatation of segments of the double-layered ductal plate to form tubules hat become the intrahepatic bile ducts. Non-tubular parts of the plate disappear and the bile ducts form part of the portal tracts.

Congenital disorders of the biliary tract are classified into two main groups: diseases characterized by inflammatory destruction of the bile ducts (the biliary atresias) and diseases marked by ectasia of the bile ducts with varying degrees of fibrosis (the fibropolycystic diseases). Both of these groups of disorders are related to the persistence or lack of remodelling of the embryonic ductal plate. They are termed 'ductal plate malformations'.

Ductal plate malformations can be seen on ultrasound or CT scans as a circular lumen containing a fibrovascular cord. Figure 1 shows ductal plate malformations in a CT scan of a patient with Caroli's disease (a fibropolycystic disease).

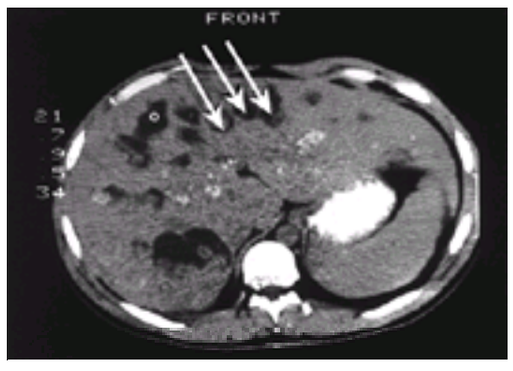

Fig. 1 Caroli's disease. Intravenous contrast enhanced CT scan of the liver shows dilated intrahepatic bile ducts containing filling defects which are portal vein branches (arrowed). This is an example of a ductal plate malformation. (From Sherlock S, Summerfield JA (1991), with permission.)

Biliary atresia

Classification

Biliary atresias are classified into extrahepatic biliary atresia and intrahepatic biliary atresia (paucity of intrahepatic bile ducts). Biliary atresia does not represent agenesis of the bile ducts but is the result of progressive bile duct destruction from an inflammatory disease of unknown cause. In extrahepatic biliary atresia the destructive cholangitis affects not only part or the whole of the extrahepatic bile duct but also intrahepatic bile ducts and leads to paucity of intrahepatic bile ducts. In intrahepatic biliary atresia the destructive cholangitis is restricted to the intrahepatic bile ducts. Intrahepatic biliary atresia can be classified further into a non-syndromatic or a syndromatic type (Alagille's syndrome or arteriohepatic dysplasia). About a quarter of patients with extrahepatic biliary atresia have evidence of ductal plate malformation indicating that the destructive cholangitis started early in fetal life.

Symptoms and signs

Biliary atresia presents as cholestatic jaundice starting after the first two weeks of life. The infant develops jaundice with pale stools, dark urine, and hepatomegaly. Itching is often prominent. Bile pigments may stain the growing teeth greenish. The jaundice steadily deepens and xanthomas of the palm and knees, rickets, a bleeding tendency, and growth failure may develop. Biliary atresia may eventually causes biliary cirrhosis with pigmentation (due to melanin), portal hypertension, ascites, and liver failure.

The progress of biliary atresia depends on the type. Infants with extrahepatic biliary atresia (usually girls) have a steadily deepening jaundice and biliary cirrhosis soon develops. Untreated, these children usually die by six months of age. The fate of infants with intrahepatic biliary atresia depends on whether they have a syndromatic or non-syndromatic atresia. Children with non-syndromatic intrahepatic biliary atresia survive longer than those with extrahepatic biliary atresia, but biliary cirrhosis eventually develops in later childhood. In contrast, patients with syndromatic intrahepatic biliary atresia (Alagille's syndrome) tend to recover normal liver function as they become adolescent. Infants with Alagille's syndrome can be recognized by the associated features, which include a characteristic facies (a flattened and triangular-shaped face), pulmonary stenosis, vertebral abnormalities, and a change in the eyes (embryotoxon). Some patients have growth and mental retardation. Alagille's syndrome is associated with mutations in the Jagged 1 gene which encodes a Notch ligand.

Differential diagnosis

Jaundice is common in early infancy. In the early neonatal period jaundice is usually due to haemolysis and impaired bilirubin conjugation. After two weeks, jaundice is usually cholestatic. There are many causes of cholestasis in infancy and childhood. The most common are extrahepatic and intrahepatic biliary atresias, neonatal

hepatitis (such as hepatitis A, B, and C, rubella, and cytomegalovirus infection), metabolic causes (such as galactosaemia, a $_1$-antitrypsin deficiency, and tyrosinaemia), and the 'inspissated bile syndrome' (congenital spherocytosis).

Laboratory investigations

Liver function tests show a cholestatic (biliary obstructive) pattern. Serum bilirubin and alkaline phosphatase levels are markedly raised with only modest elevations of serum transaminases. Later, very high levels of serum cholesterol may develop.

Histological examination of the liver cannot distinguish between intrahepatic and extrahepatic biliary atresia. Liver biopsy shows severe centrizonal cholestasis and a prominent giant-cell reaction. In the portal tracts bile ducts are reduced. Later in the course of the disease the portal tracts are devoid of bile ducts and biliary cirrhosis is present.

Imaging

The initial step in the management of infants with cholestasis is to differentiate between intrahepatic and extrahepatic biliary atresia. Since the clinical and laboratory findings are similar, this distinction requires imaging techniques. In extrahepatic biliary atresia, scintiscanning with $^{99}Tc^{m}$-labelled HIDA (dimethyl acetanilide iminodiacetic acid) shows accumulation of the label in the liver but none enters the biliary tree. Percutaneous and endoscopic cholangiography provide more precise anatomical detail.

Treatment and prognosis

General supportive measures include parenteral administration of fat-soluble vitamins A, D, K, and E. Medium-chain triglycerides as a source of fat, cholestyramine to relieve itching, and ursodeoxycholic acid as a choleretic help some patients.

Extrahepatic biliary atresia

Hepatic portoenterostomy (Kasai's operation) has been the treatment of choice for extrahepatic biliary atresia and is still widely performed. Approximately 25 to 35 per cent of patients who undergo a Kasai portoenterostomy will survive more than 10 years without liver transplantation. A third of the patients drain bile but develop complications of cirrhosis and require liver transplantation before the age of 10. For the remaining third of patients, bile flow is inadequate following portoenterostomy and the children develop progressive fibrosis and cirrhosis. The portoenterostomy should be done before there is irreversible sclerosis of the intrahepatic bile ducts. Consequently, a prompt evaluation for conjugated hyperbilirubinaemia is indicated for any infant older than 14 days with jaundice.

Intrahepatic biliary atresia

All infants should receive general supportive measures. Definitive treatment depends on the type of intrahepatic biliary atresia. Non-syndromatic intrahepatic biliary atresia eventually progresses to biliary cirrhosis and liver failure. Liver transplantation should be performed before the onset of liver failure. Syndromatic intrahepatic biliary atresia (Alagille's syndrome) has a good prognosis in most children and few develop biliary cirrhosis. General supportive measures are usually sufficient until the cholestasis disappears.

Fibropolycystic disease

Fibropolycystic disease encompasses a family of rare congenital hepatobiliary diseases that arise due to malformations of the embryonic ductal plate. These diseases include fibropolycystic disease (polycystic liver), congenital hepatic fibrosis, congenital intrahepatic biliary dilatation (Caroli's disease, Fig. 1), choledochal cysts, and microhamartomas (von Meyenberg complexes). Many patients will have more than one disease. The combination of congenital hepatic fibrosis and Caroli's disease is characteristic as these patients develop first variceal haemorrhage (due to congenital hepatic fibrosis) and later recurrent cholangitis (due to Caroli's disease). Associated kidney defects are common. Malignant change may complicate congenital hepatic fibrosis, Caroli's disease, choledochal cysts, and microhamartomas. These diseases are of widely differing severity and the prognosis in an individual patient is determined by the fibropolycystic diseases present.

Polycystic liver disease

The infantile type is inherited as an autosomal recessive disease and is usually rapidly fatal due to the associated renal disease. Adult polycystic liver disease is more common and has a dominant inheritance. The patient is usually a woman presenting in the fourth or fifth decade. The liver contains many thin-walled cysts filled with a clear or brownish liquid (due to altered blood). The cysts vary in size from a pinhead to about 10 cm in diameter. The remainder of the liver is normal. Patients present with right upper quadrant pain and increasing girth. Examination reveals an enlarged liver as the cause of the upper abdominal swelling. Liver function tests are normal. Provided no other fibropolycystic diseases are present, polycystic liver disease is benign. Some patients with polycystic liver disease also have polycystic kidneys or nephrocalcinosis. The associated renal disease may cause serious complications including renal failure. The diagnosis can be confirmed by ultrasound or CT scanning, which show numerous thin-walled cysts of low density (Fig. 2). The enlarged polycystic liver causes some patients considerable discomfort. It is best treated by percutaneous aspiration of the larger cysts using ultrasound guidance in order to reduce liver size. Percutaneous aspiration treatment can be performed repeatedly.

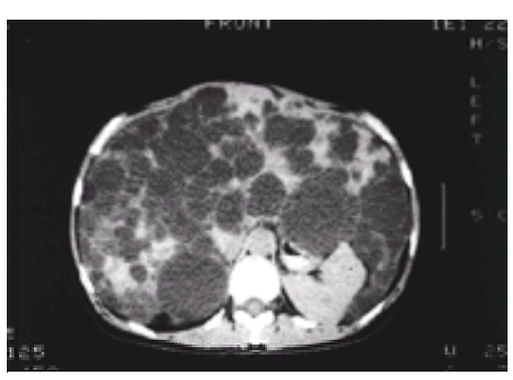

Fig. 2 Polycystic liver disease. CT scanning shows the liver contains many cysts of low density indicating that they are fluid filled. (From Sherlock S, Summerfield JA (1991), with permission.)

Congenital hepatic fibrosis

This is a rare autosomal recessive condition which is usually diagnosed before 10 years of age. The main complication is portal hypertension. Children present with a large, very hard liver and splenomegaly or bleeding from oesophageal varices. Congenital hepatic fibrosis may be misdiagnosed as cirrhosis. Liver function tests are normal or only slightly deranged. Ultrasound scans show the liver contains many bright areas due to the dense bands of fibrous tissue. The diagnosis is made by liver biopsy which shows normal liver parenchyma surrounded by fibrous septa containing structures resembling bile ducts.

Patients with congenital hepatic fibrosis bleed repeatedly from oesophageal varices, but because liver function is well preserved they do not develop portosystemic encephalopathy. Portocaval shunts will stop the variceal bleeding and are well tolerated. Liver transplantation has also been used successfully.

The long-term prognosis in congenital hepatic fibrosis is usually determined by the associated renal disease. Renal lesions include renal dysplasia, medullary cystic disease, and infantile or adult-type polycystic kidneys. The kidneys are rarely normal and renal failure eventually develops in many patients. However, renal

transplants have been successful.

Congenital intrahepatic biliary dilatation (Caroli's syndrome)

In Caroli's syndrome the common bile duct is normal but the intrahepatic ducts have bulbous dilatations with normal ducts between (Fig. 3). The mode of inheritance is unknown. While the cystic dilatations of the bile ducts remain uninfected the patient is symptom free. Eventually, ascending infection leads to cholangitis, which can be intractable with the formation of gallstones and liver abscesses. Caroli's syndrome usually presents in early adulthood as cholangitis. Most patients are male. Liver function tests show cholestasis with elevations of serum bilirubin and alkaline phosphatase and modest elevations of the transaminases. The diagnosis is made by endoscopic cholangiography. CT scans can also demonstrate the syndrome (Fig. 1). The natural history of Caroli's disease is of recurrent cholangitis which is very resistant to antibiotics. Biliary cirrhosis eventually develops. Bile duct cancer develops in about 10 per cent of cases. Treatment is difficult, antibiotics are usually only partially effective and liver transplantation is compromised by the extensive sepsis.

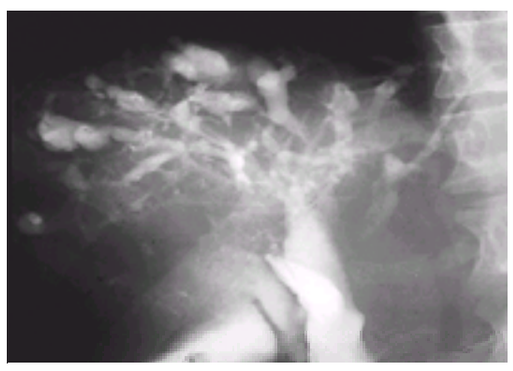

Fig. 3. Caroli's disease. An endoscopic cholangiogram shows bulbous dilatations of the intrahepatic bile ducts. The rest of the biliary tree is normal. (From Sherlock S, Summerfield JA (1991), with permission.)

About half the patients with congenital hepatic fibrosis or Caroli's disease will also have the other disease. The clinical presentation in these patients is distinctive. As in Caroli's disease, males predominate. The first complication is variceal haemorrhage followed about 10 years later by recurrent cholangitis.

Choledochal cyst

Choledochal cyst is a congenital dilatation of part or the whole of the common bile duct (Fig. 4). It is more common in girls and usually presents in childhood but may appear in early adulthood. Choledochal cysts classically cause a triad of intermittent pain, jaundice, and a right hypochondrial mass. Choledochal cysts are particularly common in Japanese and Chinese individuals. Liver function tests show cholestasis, similar to Caroli's disease. Ultrasound and CT scans show cystic dilatation of the bile duct. The diagnosis is made by endoscopic or percutaneous cholangiography. Choledochal cysts should be treated by surgical excision because of the risk of bile duct malignancy. Caroli's disease is a common associated disease.

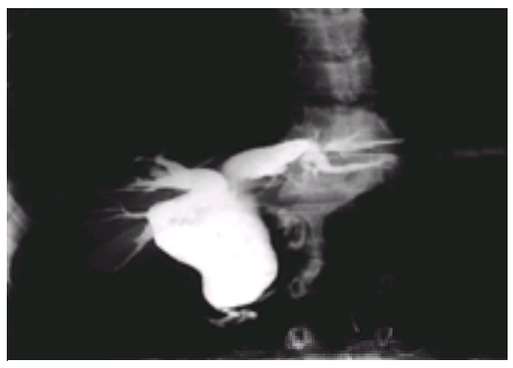

Fig. 4 Choledochal cyst in a 20-year-old woman. The endoscopic cholangiogram shows a massively dilated common bile duct. The gallbladder was normal but obscured by the dilated bile duct. (From Sherlock S, Summerfield JA (1991), with permission.)

Microhamartomas (von Meyenberg complexes)

Microhamartomas are groups of rounded biliary channels embedded in a collagen stroma located around portal tracts. The appearances are of localized islands of congenital hepatic fibrosis. Microhamartomas are usually asymptomatic and discovered incidentally on liver biopsy. They may be associated with other fibropolycystic diseases and are a rare cause of portal hypertension. Bile duct and pancreatic cancers are commoner in these patients.

Congenital disorders of the pancreas

Agenesis of the pancreas

Pancreatic agenesis is rare and may occur as an isolated anomaly or be associated with other defects. These children usually die soon after birth. Agenesis of either the dorsal or ventral pancreas may occur, although agenesis usually involves the dorsal segment.

Annular pancreas

This is a rare condition where pancreatic tissue encircles the descending duodenum. It results from persistence of part of the ventral pancreas during embryonic development. Annular pancreas is the most common cause of duodenal obstruction in infancy and often involves growth of pancreatic tissue into the duodenal wall. However, the clinical presentation is variable and annular pancreas may first present as an incidental finding at surgery or autopsy.

Pancreas divisum

Pancreas divisum results from failure of fusion of the ducts of the dorsal and ventral portions of the pancreas. The body and tail of the pancreas drain through the narrow duct of Santorini into the accessory papilla. Only the head of the pancreas drains into the ampulla of Vater (Fig. 5). This is the commonest congenital abnormality of the pancreas occurring in about five per cent of patients. Pancreas divisum appears to be associated with an increased incidence of pancreatitis affecting the body and tail of the pancreas which drains into the accessory papilla. Endoscopic sphincterotomy of the accessory papilla is reported to lead to clinical improvement in this type of pancreatitis.

Fig. 5 Pancreas divisum. An endoscopic pancreatogram following injection of contrast medium into the ampulla of Vater shows only the ducts of the head of the pancreas, characterized by a trefoil pattern. The body and tail of the pancreas drain via the accessory ampulla.

Hereditary pancreatitis

This rare form of pancreatitis is inherited as an autosomal dominant disorder. Recurrent attacks of abdominal pain start in childhood or the second decade. Hereditary pancreatitis tends to be troublesome rather than life-threatening and attacks become less severe as the patient gets older. They often disappear by middle age. Hereditary pancreatitis is associated with mutations in the cationic trypsinogen gene, which probably render the protease more resistant to autocatalytic trypsinogen breakdown.

Other rare abnormalities causing pancreatic disease

Congenital abnormalities adjacent to the pancreas are rare causes of pancreatitis. These include duodenal diverticulum, duplication of the duodenum, stenosis of the sphincter of Oddi, and choledochal cyst. These abnormalities seem to cause pancreatitis by obstructing the pancreatic duct.

Further reading

Chardot C *et al.* (1999). Prognosis of biliary atresia in the era of liver transplantation: French national study from 1986 to 1996. *Hepatology* **30**, 606–11.

Desmet VJ (1992) Congenital diseases of intrahepatic bile ducts: variations on the theme 'ductal plate malformation'. *Hepatology* **16**, 1069–83.

Sherlock S, Dooley JS (1997). *Diseases of the liver and biliary system*, 10th edn. Blackwell Scientific Publications, Oxford.

Sherlock S, Summerfield JA (1991). *A colour atlas of liver disease*, 2nd edn. Wolfe Medical Publications, London.

Summerfield JA *et al.* (1986). Hepatobiliary fibropolycystic diseases; a clinical and histological review of 51 patients. *Journal of Hepatology* **2**, 141–56.

14.19.2 Diseases of the gallbladder and biliary tree

J. A. Summerfield

Anatomy

The biliary system comprises the collection of ducts extending from the biliary canaliculus of each hepatocyte to the ampulla of Vater opening into the duodenum. The biliary canaliculi drain into interlobular and then septal bile ducts. These further ramify to form the intrahepatic bile ducts which are visible on cholangiography (Fig. 1). They eventually form the right and left hepatic ducts draining bile from the right and left lobes of the liver, respectively. The junction of the hepatic ducts at the porta hepatis forms the common hepatic duct. The cystic duct, linking the gallbladder to the bile duct, arises from the lower end of the common hepatic duct. The gallbladder rests in a fossa under the right lobe of the liver. Anatomical variations in the size and position of the gallbladder and the insertion of the cystic duct into the bile duct are of major surgical importance. The common hepatic duct becomes the common bile duct below the insertion of the cystic duct. The common bile duct passes through the head of the pancreas and the sphincter of Oddi to drain into the duodenum via the ampulla of Vater. The bile duct usually exits through a common channel with the pancreatic duct in the ampulla of Vater, although anatomical variations are frequent.

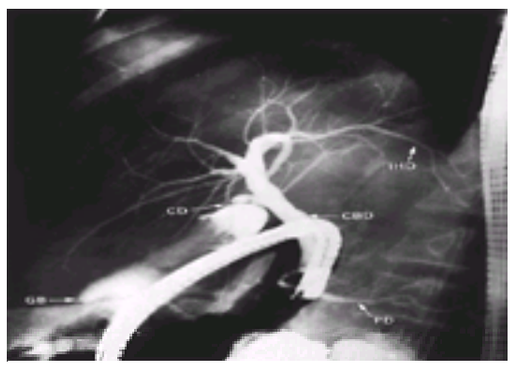

Fig. 1 The normal biliary tree. The intrahepatic bile ducts (IHD) taper smoothly and extend deep into the liver. The gallbladder (GB) drains via the cystic duct (CD) into the common bile duct (CBD). The pancreatic duct (PD) has also been opacificed in this endoscopic retrograde cholangiogram.

The investigation of biliary disease

Objectives

The clinical and laboratory features of biliary disease may also be caused by hepatic disorders. Consequently, the primary objective of investigations is to establish that the cause is due to biliary and not hepatic disease. The secondary objective is to define the anatomy of the lesion to permit a rational choice of the many surgical and non-surgical therapeutic options which are now available. To achieve these objectives requires not only a careful history and physical examination, but also the use of various imaging techniques and sometimes aspiration liver biopsy.

Symptoms and signs

Disorders of the biliary system usually give rise to the symptoms and signs of biliary obstruction (cholestasis). The repertoire is rather limited: pain, jaundice, itching, nausea and vomiting, fevers, and rigors. The pain can range from abdominal discomfort described as 'dyspepsia' to severe right hypochondrial colic caused by a sudden rise in biliary pressure. Jaundice, dark urine, and pale stools indicate obstruction of the bile duct. Itching is an important sign of biliary obstruction. Nausea and vomiting may be prominent in sudden obstruction of the bile duct, usually by a gallstone. The milder symptoms of flatulence and intolerance of fatty food are more common. Fever and rigors indicate bacterial infection of the biliary tract, which frequently accompanies partial obstruction. In jaundiced patients weight loss is usual and results from fat malabsorption due to the lack of bile acids reaching the gut; it may also indicate a malignant tumour. Prolonged biliary obstruction leads to skin changes: increased pigmentation (due to melanin) and cholesterol deposits (xanthelasma and xanthoma). Finally, biliary cirrhosis may develop causing the signs of portal venous hypertension and liver cell failure.

Laboratory investigations

In general, disorders of the biliary system give rise to the biochemical picture of biliary obstruction (cholestasis). A notable exception is gallstones in the gallbladder (cholelithiasis) where the liver function tests are usually normal. In cholestasis, the serum bilirubin concentration may be normal or raised and most of the bilirubin is esterified (conjugated). Bilirubinuria is present. The disappearance of urobilinogen from the urine indicates complete biliary obstruction. Elevation of the serum alkaline phosphatase is an important but not invariable sign of biliary obstruction; the rise is usually greater than three times normal. Other biliary canalicular enzymes accumulate in the blood, including g-glutamyl transpeptidase. This enzyme is only found in the liver and is estimated if there is doubt as to whether the alkaline phosphatase is of bony or hepatic origin. This may be required in children and patients with malignancy. Serum transaminases, such as aspartate aminotransferase, show only modest elevation in contrast to the rises which ccur in hepatitis. The serum cholesterol concentration rises and may cause abnormalities of red cell shape (target cells) (see Section 22). A raised concentration of serum bile acids is a sensitive index of biliary disease. A prolonged prothrombin time reflects intestinal malabsorption of fat-soluble vitamin K owing to a lack of bile acids. Vitamin A and D deficiency may also develop. The serum albumin and gammaglobulin levels are normal until biliary cirrhosis develops. A polymorphonuclear leucocytosis accompanies bacterial infections of the biliary system.

Imaging techniques

A plain radiograph of the abdomen may reveal an enlarged liver, calcified gallstones, or air in the biliary tree. Plain radiographs of the abdomen are now rarely performed. The preferred first investigation is ultrasonography (Fig. 2). Computed tomography (**CT** scan) and magnetic resonance imaging (**MRI**) are used in complicated diagnostic problems. These tests reveal dilated bile ducts and may also indicate the position of the obstruction in the biliary tree and dense structures such as gallstones. Hepatic scintiscanning with $^{99}Tc^{m}$-labelled HIDA (dimethyl acetanilide iminodiacetic acid) is an alternative and is of value in the diagnosis of acute cholecystitis. Oral cholecystograms are rarely performed nowadays but are useful to determine whether the gallbladder is functioning in patients with gallstones being assessed for oral bile acid dissolution therapy (see below). Intravenous cholangiography is obsolete. However, these non-invasive investigations usually provide insufficient anatomical detail for diagnosis or planning of treatment. An invasive cholangiographic technique such as percutaneous transhepatic cholangiography (**PTC**) or endoscopic retrograde cholangiopancreatography (**ERCP**) is necessary. ERCP is the preferred investigation in the first instance. PTC is reserved for patients in whom ERCP fails. Both these techniques carry small risks including haemorrhage, biliary peritonitis, and cholangitis (with PTC), and bowel perforation, cholangitis, and pancreatitis (with ERCP). Should cholangiography reveal a normal biliary system in a jaundiced patient, a liver biopsy is indicated.

Fig. 2 Ultrasound scan of the gallbladder shows gallstones (arrowed) as bright round objects which cast acoustic shadows.

This diagnostic approach is ideal but expensive both in terms of human and material resources. The apparatus required is costly and procedures such as ERCP require considerable expertise. Obviously local factors will determine the diagnostic pathway that is adopted. Nevertheless, these techniques have revolutionized the management of patients with biliary disease. It is now a routine matter to achieve a precise diagnosis rapidly. In addition, a series of non-operative therapeutic options ranging from the introduction of endoprostheses for the management of benign and malignant biliary structures to endoscopic sphincterotomy for the removal of the biliary calculi are direct consequences of these diagnostic approaches. Developments in MRI indicate that soon ERCP may be superseded by the non-invasive technique of magnetic resonance cholangiography (MRC).

Bile composition and gallstone formation

Bile composition

Bile is secreted by the hepatocytes and its water and electrolyte composition altered during its passage down the biliary system. Between meals much of the bile is diverted to the gallbladder where it is concentrated by the removal of sodium, chloride, bicarbonate, and water. In response to food, the gallbladder contracts, emptying bile into the duodenum. Apart from water (97 per cent) the major components of bile are bile acids, phospholipids, and cholesterol. Bile is also the major excretory route of other compounds including bilirubin and certain drugs and their metabolites. Cholesterol is insoluble in water but is held in solution by the detergent action of bile acids with the aid of phospholipids.

Cholesterol is synthesized primarily in the liver and small intestine. The rate-limiting enzyme for cholesterol production is hydroxymethylglutaryl-CoA reductase, which catalyses the first step, the conversion of acetate to mevalonate. Subsequently, non-esterified (free) cholesterol is secreted into bile. Dietary cholesterol also contributes to biliary cholesterol secretion. The control of cholesterol metabolism is complex. It is not yet clear what proportion of biliary cholesterol is derived from circulating lipoproteins and what proportion is newly synthesized by the liver.

The primary bile acids, cholic and chenodeoxycholic acid, are synthesized in the liver from cholesterol. The economy of the bile acid pool is preserved by efficient reabsorption, principally in the terminal ileum. About 95 per cent of the bile acids are reabsorbed and pass back to the liver in the portal venous system (enterohepatic circulation). The remainder enters the colon where bacteria form the secondary bile acids, deoxycholic and lithocholic acid, from cholic and chenodeoxycholic acid, respectively. Some of the secondary bile acids are absorbed from the colon but most are excreted in the faeces. The normal bile acid pool is about 3 to 5 g and circulates 6 to 10 times each day. Synthesis is controlled by the negative feedback of bile acids returning in the portal venous blood, which act on the rate-limiting hepatic enzyme, cholesterol-7a-hydroxylase. The principal phospholipid in bile is lecithin. It is produced in the liver and secreted into the bile. In the intestine lecithin is hydrolysed to lysolecithin by pancreatic phospholipase and is subsequently reabsorbed.

Above a certain level (the critical micellar concentration) bile acids coalesce to form micelles that have a hydrophilic external surface and hydrophobic internal surface. Cholesterol is incorporated into the hydrophobic interior. Phospholipids are inserted into the micellar wall so that the micelles are enlarged; these 'mixed micelles' are thus able to hold more cholesterol.

Consequently, the solubility of cholesterol in bile depends on the concentrations of bile acid and phospholipid. In the presence of a relative excess of bile acids and phospholipid (on a molar basis) the cholesterol-holding capacity of bile is increased and it is said to be unsaturated. However, if there are insufficient micelles of bile

acid and phospholipid to hold the cholesterol, the solution is referred to as saturated and the excess cholesterol tends to precipitate. With a knowledge of the molar concentration of cholesterol, phospholipid, and bile acids, the cholesterol saturation of bile can be predicted using triangular co-ordinate diagrams.

Gallstone formation

Gallstone disease is common and afflicts between 10 and 20 per cent of the world's population. Gallstones are classified according to their composition into two main groups: cholesterol stones and bile pigment stones. Cholesterol stones are composed mainly of cholesterol (more than 70 per cent) and can be subdivided into pure cholesterol stones (usually solitary) and mixed stones which contain cholesterol in a matrix of calcium bilirubinate, calcium phosphate, and protein (Fig. 3 and Fig. 4). Mixed stones are usually multiple and faceted. Bile pigment stones can also be divided into two main groups. Brown pigment stones are soft and friable and consist of calcium bilirubinate, cholesterol, and calcium soaps. Pure pigment stones ('black stones') are black, hard, and brittle and contain an insoluble black pigment, calcium bilirubinate, calcium carbonate and phosphate, calcium salts of fatty acids, and bile acids. All pigment stones contain a large amount of mucoprotein matrix (up to 70 per cent). Gallstones are rare before the age of 10 years. The incidence increases progressively with age. Cholesterol gallstones account for about 75 per cent of the gallstones in Europe and the United States.

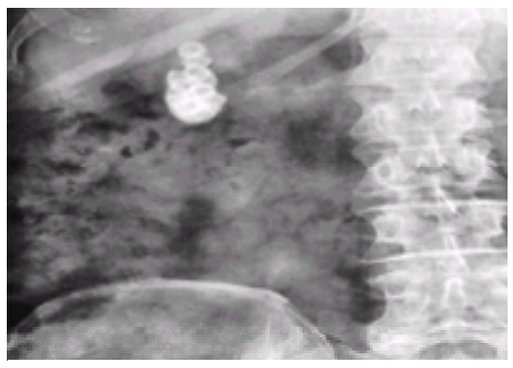

Fig. 3 Calcified gallstones. Gallstones contain sufficient calcium to be visible on a plain abdominal radiograph in about 10 per cent of patients. The gallbladder stones are surrounded by a ring of calcium slats. (Reproduced from Sherlock S, Summerfield JA, 1979, *A colour atlas of liver disease*, Wolfe Medical Publications, London, with permission.)

Fig. 4 Cholesterol gallstones. An intravenous cholangiogram has opacified the gallbladder showing multiple faceted radiolucent gallstones. These are typical features of cholesterol stones.

Cholesterol gallstones

Cholesterol gallstones result from the secretion of cholesterol-saturated bile by the liver. The cause of the saturation is unclear. Patients with gallstones usually have a smaller bile acid pool than controls and it circulates more frequently. The rapid recycling of bile acids may be responsible for the smaller bile acid pool by excessive inhibition of the enzyme which controls bile acid synthesis, cholesterol-7a-hydroxylase. However, diminished bile acid synthesis is probably not the most important factor in the production of saturated bile. This appears to be an elevated biliary cholesterol secretion rate, due either to increased hepatic cholesterol synthesis or increased transfer of plasma lipoprotein cholesterol into bile. Nevertheless, saturated bile may be encountered in normal subjects, especially during fasting. It is therefore likely that other factors such as the condition of the gallbladder, the mechanism of seeding (nucleation) of gallstones, and the control of gallstone growth are important. Furthermore, racial differences, advancing age, female sex, obesity, diet, drugs (such as the contraceptive pill and clofibrate), and gastrointestinal disease (such as Crohn's disease) are known to have a significant influence on the development of gallstones.

Bile pigment gallstones

In contrast to cholesterol stones, little is known of the aetiology of bile pigment stones. The soft, friable brown-pigment stones are especially common in the Far East and are associated with *Escherichia coli*, bacteroides, and clostridium infection of the biliary tract. It is probable that these bacteria contribute to stone formation by producing b-glucuronidase that deconjugates bilirubin diglucuronide to form free unconjugated bilirubin. This combines with calcium to form sparingly soluble calcium bilirubinate that precipitates.

The black, hard, and brittle pure-pigment stones are the type commonly encountered in the West. The incidence of pure pigment stones increases with age and they are found in patients with cirrhosis, chronic bile duct obstruction (such as biliary strictures), chronic haemolytic anaemias including haemolysis induced by prosthetic heart valves, and malaria. Pure pigment stones affect both sexes equally. The mechanism of stones production is unclear, but does not appear to be due to cholesterol saturation of hepatic or gallbladder bile. About 50 per cent of all pigment stones are radio-opaque and they account for about 70 per cent of all opaque stones.

Natural history of gallstones

The majority of gallstones remain in the gallbladder (cholelithiasis) and may give rise to no symptoms ('silent' gallstones), being discovered incidentally during investigation or at autopsy. Impaction of a gallstone in the neck of the gallbladder results in gallbladder inflammation and the symptoms and signs of acute or chronic cholecystitis. Acute cholecystitis will subside if the stone spontaneously disempacts, or may progress to gangrene and perforation of the gallbladder or empyema of the gallbladder. Gallstones may pass through the cystic duct into the bile duct (choledocholithiasis) resulting in biliary obstruction and jaundice. Bacterial infection (cholangitis) commonly accompanies choledocholithiasis and can lead to a liver abscess. Gallstones may perforate through the inflamed gallbladder wall to form an internal fistula, usually to the small intestine or colon. A large gallstone passing into the small intestine may impact in the ileum resulting in intestinal obstruction (gallstone ileus). Finally, surgical treatment for gallstones, while usually curative, may result in a postcholecystectomy syndrome or a benign stricture of the bile duct.

Treatment

The usual treatment for gallstones remains cholecystectomy although medical treatments may be employed in selected patients (see below). The advent of laparoscopic cholecystectomy has swung the balance in favour of surgery since this technique carries so little morbidity and a very short hospital stay. Treatment is obviously indicated for symptomatic gallstones and for their complications. However, in patients in whom 'silent' gallstones are discovered incidentally and in patients with minimal symptoms it is by no means clear that treatment is always the best solution. The problem revolves around the probability of serious complications in the future. It is appropriate to offer treatment to young patients (who, with many years ahead of them, will have a greater likelihood of developing the complications of

gallstones) and to advise against treatment in the elderly with other major medical problems. However, in fit middle-aged patients with no or minimal symptoms it is reasonable to tell the patient of the finding and to withhold surgery until it is warranted by symptoms or complications.

Gallstone dissolution and disruption

Cholesterol gallstones can be removed from the gallbladder and bile ducts in a proportion of patients by medical treatments. These techniques avoid the discomfort, disability, and risks of general anaesthesia and surgical exploration of the abdomen and bile ducts. However, with the widespread availability of laparoscopic cholecystectomy these techniques are now used rarely. There are two types of medical method: chemical agents that dissolve gallstones, and physical methods such as endoscopic sphincterotomy and extracorporeal shock-wave lithotripsy (**ESWL**). Judicious combinations of chemical and physical methods yield the best results.

Chemical methods

Oral bile acid therapy

Oral treatment with chenodeoxycholic acid or ursodeoxycholic acid can dissolve cholesterol gallstones. These bile acids, normal constituents of bile, reduce the cholesterol saturation of bile and result in the leaching of cholesterol from gallstones. They act by reducing the hepatic synthesis and biliary excretion of cholesterol. Ursodeoxycholic acid has advantages over chenodeoxycholic acid in that it does not cause diarrhoea or elevations of serum transaminases. These bile acids differ in the way that they remove cholesterol from gallstones and have been shown to dissolve gallstones better in combination than singly. Combination therapy is the preferred treatment.

Contact dissolution of allstones

Cholesterol stones in the gallbladder can be dissolved by the direct instillation of methyl tertbutyl ether (**MTBE**) into the gallbladder via a percutaneous catheter. MTBE is a foul-smelling, volatile, inflammable colourless liquid that remains liquid at body temperature. The gallbladder is catheterized by the transhepatic route, entering it through the area of attachment of the gallbladder to the liver and MTBE is continually infused and aspirated with vigour until the stones have disappeared (which typically takes 5 to 7 h).

Physical methods

Extracorporeal shock-wave lithotripsy

ESWL is a non-invasive and safe but expensive way of rapidly shattering gallstones into a coarse powder. The gallbladder must contain no more than three stones to allow accurate focusing of the shock waves.

Endoscopic sphincterotomy

Endoscopic sphincterotomy can remove gallstones from the bile duct. The bile duct is entered by a cannula passed via a duodenoscope and the bile duct is opened by diathermy cutting of the ampulla of Vater. Stones are removed by balloon or wire catheters.

Patient selection and results

Medical treatment with oral bile acid therapy, ESWL, or contact dissolution are suitable for patients with cholesterol gallstones in a functioning gallbladder (as judged by an oral cholecystogram). Calcified gallstones do not dissolve. Radiolucent gallstones are usually, but not always, composed of cholesterol. CT scans are useful for detecting low levels of gallstone calcification. These treatments should be reserved for patients with mild or no symptoms in whom the risk of cholecystectomy is high, including those with pre-existing disease, the elderly, and the very obese. They are also of value in patients who refuse surgery. Drugs which increase the cholesterol saturation of bile should be avoided; these include oestrogens, the oral contraceptive pill, and clofibrate.

Oral bile acid therapy is protracted but safe. It dissolves gallstones in about 25 per cent of patients fulfilling the selection criteria by 6 months. It should not be taken during pregnancy. The preferred treatment is combination therapy with chenodeoxycholic acid (7 mg/kg) and ursodeoxycholic acid (7 mg/kg). Proprietary combination tablets are available. Gallstone dissolution usually requires 6 to 24 months of therapy depending on stone size. Oral cholecystograms are performed every 6 months to assess progress. Combining oral bile acid therapy with ESWL speeds up the process greatly: gallstones will be cleared in more than 90 per cent of patients within 18 months. Furthermore, slightly calcified gallstones can be treated in this way. MTBE therapy is invasive and the ether is unpleasant to use, but dissolution is rapid. Endoscopic sphincterotomy removes gallstones from the common bile duct. Any type of stone can be removed up to about 20 mm in diameter.

Side-effects and toxicity

The most frequent side-effect of oral bile acid therapy is diarrhoea. It is dose related and usually mild and transient. It can be minimized by slowly increasing the dose to the required level. Transient elevations of serum transaminase activity are also common; liver function tests should be monitored. Ursodeoxycholic acid may cause calcification of gallstones. Gallstone recurrence remains a major problem with oral bile acid therapy. One year after gallstone dissolution about 30 per cent of patients will have had a recurrence. Unwanted effects of ESWL include biliary colic, skin petechias, and haematuria. The principal unwanted side-effects of MTBE are sedation, burning upper abdominal pain, nausea, and vomiting. Endoscopic sphincterotomy can cause gastrointestinal haemorrhage and acute pancreatitis.

Acute cholecystitis

Aetiology

Acute cholecystitis is associated with gallstones in over 90 per cent of patients. It follows the impaction of a gallstone in the cystic duct. Continued secretion by the gallbladder leads to a rise in pressure. Inflammation of the gallbladder wall results from the toxic effects of the retained bile and bacterial infection. The gallbladder bile is usually turbid but may become frank pus (empyema of the gallbladder). Intestinal organisms, especially anaerobes, are commonly cultured from the gallbladder. Ischaemia in the distended gallbladder wall may lead to infarction and perforation. Generalized peritonitis may follow, but the leak is usually localized to form a chronic abscess cavity. Some patients have repeated attacks of acute cholecystitis which are probably exacerbations of chronic cholecystitis. Acute cholecystitis in the abscence of gallstones (acalculous cholecystitis) is usually very rare. However, acalculous cholecystitis is a particular problem in patients with the acquired immunodeficiency syndrome (**AIDS**). Cytomegalovirus and cryptosporidium are the most commonly associated organisms in acalculous cholecystitis in AIDS.

Symptoms and signs

The typical patient is an obese, middle-aged female, and the acute attack is often precipitated by a large or fatty meal. However, there are many exceptions to this pattern. The principal symptom is pain, of fairly sudden onset, which is severe, continuous or minimally fluctuating, and localized to the epigastrium or right hypochondrium. The pain often radiates to the back. The constancy of the pain is in contrast to the repeated short bouts of biliary colic. In uncomplicated cases the pain gradually subsides over 12 to 18 h. Flatulence and nausea are common but persistent vomiting suggests the presence of a stone in the common bile duct. Examination reveals an ill, sweating patient with shallow, jerky respiration. Fever indicates a complicating bacterial cholangitis. Jaundice may accompany acute cholecystitis but is usually a sign of a stone in the bile duct. The abdomen moves poorly with respiration. Right hypochondrial tenderness is present and is exacerbated by inspiration (Murphy's sign). Muscle guarding and rebound tenderness are common. The gallbladder is usually impalpable but occasionally a tender mass of omentum and gallbladder may be felt under the liver.

Laboratory investigations

The white cell count is usually moderately elevated (12 to 15 × 10^{9}/l) due to a polymorphonuclear leucocytosis. Serum bilirubin concentrations between 17 and 68 µmol/l (1 and 4 mg/dl) may be seen in uncomplicated acute cholecystitis, but should raise the suspicion of a stone in the bile duct. Modest rises in the serum alkaline phosphatase, aspartate transaminase, and amylase may also be seen. An abdominal radiograph will show gallstones in about 10 per cent of patients. Ultrasound

scanning of the gallbladder is the preferred first investigation. Scintiscanning with $^{99}Tc^{m}$-labelled HIDA provides similar information. It is important to establish the correct diagnosis before surgery is performed.

Differential diagnosis

Acute cholecystitis may be confused with other abdominal emergencies including perforated peptic ulcer, acute pancreatitis, retrocaecal appendicitis, perforated carcinoma or diverticulum of the hepatic flexure of the colon, and liver abscess. Cardiac infarction and pneumonia with right-sided pleurisy should also be considered.

Complications

Gangrene of the gallbladder

Pain, tenderness, and fever progressively increasing or persisting for longer than 24 to 48 h are indications of gangrene of the gallbladder. The prognosis is poor if necrosis and perforation occur. In patients who are elderly and obese, perforation of the gallbladder can occur without definite signs. Perforation into an adjacent viscus may produce a cholecystenteric fistula and may lead to gallstone ileus.

Cholangitis

Intermittent high temperatures often accompanied by rigors indicate bacterial infection of the bile duct and usually follow the passage of a stone into the bile duct.

Treatment

In most patients acute cholecystitis subsides in a few days with conservative treatment. Cholecystectomy is performed either a few days after the symptoms have settled or 2 to 3 months later. In the latter event, if the symptoms recur during the interval, cholecystectomy is performed without delay. Immediate surgery is mandatory if signs of gangrene or perforation develop.

Conservative treatment

Oral feeding is stopped. Intravenous fluids, and analgesia with nalbuphine or pethidine (demerol) and atropine are administered. Antibiotics are given to all but the most mild cases; tetracycline, amoxicillin, or a cephalosporin are satisfactory for general use. The patient should be observed frequently with abdominal examination and sequential leucocyte counts to detect signs of gangrene of the gallbladder or cholangitis.

Surgical treatment

Cholecystectomy is the operation of choice. Laparoscopic cholecystectomy is the preferred approach. About 10 per cent of patients with acute cholecystitis will have stones in the common bile duct. The bile ducts should be assessed by ERCP and bile duct stones removed by endoscopic sphincterotomy. If an open cholecystectomy is performed, intraoperative cholangiography may be performed to determine whether bile duct stones are present. In high-risk patients and when technical difficulties are encountered a cholecystotomy may be performed.

Chronic cholecystitis

This is the most common form of gallbladder disease that results from gallstones. Pathologically it is characterized by chronic inflammation and thickening of the gallbladder wall. In addition to stones the gallbladder may contain a brown sediment ('biliary mud'). A proportion of these patients have cholesterolosis of the gallbladder ('strawberry gallbladder'). This describes the deposition of yellow specks of cholesterol in the pink gallbladder wall and is a consequence of cholesterol-saturated bile. Cholesterolosis of the gallbladder is asymptomatic but about half the patients develop gallstones. Chronic cholecystitis usually develops insidiously but may follow an attack of acute cholecystitis.

Symptoms and signs

Some patients complain of bouts of constant right hypochondrial or epigastric pain. If it is intermittent, that is, biliary colic, the height of the pain is separated by 15- to 60-min intervals. The pain may last several hours or be as brief as 15 to 20 min. It may radiate to the right shoulder or the back. More commonly the symptoms are vague and ill-defined and include abdominal discomfort and distension, nausea, flatulence, and intolerance of fatty foods. Unfortunately, many patients who do not have chronic cholecystitis complain of these symptoms. Examination of the abdomen may reveal tenderness over the gallbladder and a positive Murphy's sign. Laboratory investigations are usually unhelpful.

Imaging techniques

An ultrasound scan is used to detect gallstones. A plain radiograph of the abdomen may reveal calcified stones or opacification of the gallbladder caused by high concentrations of calcium carbonate ('limey bile') but is not often used now. If these investigations fail to show stones, but stones are still suspected on clinical grounds, an ERCP should be performed before surgery is undertaken.

Differential diagnosis

Dyspepsia and fat intolerance are common symptoms that may be caused by many conditions including peptic ulcers, hiatus hernia, irritable bowel syndrome, chronic relapsing pancreatitis, and tumours of the stomach, pancreas, colon, or gallbladder. Other functional disorders may also mimic chronic cholecystitis.

Complications

The complications of chronic cholecystitis include acute exacerbations (acute cholecystitis), passage of stones into the bile duct (choledocholithiasis or Mirizzi's syndrome), pancreatitis, cholecystenteric fistula formation and gallstone ileus, and rarely carcinoma of the gallbladder. Occasionally the accumulation of mucus and gallstones produces hydrops of the gallbladder, which is characterized by a tender mass without the symptoms of acute cholecystitis.

Treatment

In established cases of chronic cholecystitis the treatment of choice is cholecystectomy. When the diagnosis is in doubt, especially when vague symptoms are associated with a well-functioning gallbladder containing stones, a conservative approach is worth trying. This includes weight reduction and a low-fat diet, especially if fatty food is associated with the symptoms. Oral bile acid therapy may also be considered (see above).

Prognosis

Chronic cholecystitis carries a good prognosis. Cholecystectomy is curative and should have a mortality below 1 per cent. However, if cholecystectomy is performed indiscriminately on patients with 'dyspeptic' symptoms who happen to have incidental gallstones, the results will be unpredictable and often unsatisfactory.

Choledocholithiasis

Most stones in the common bile duct originate in the gallbladder. About 15 per cent of patients with cholelithiasis have common duct stones. This proportion rises with age so that in the elderly nearly 50 per cent of patients with cholelithiasis may have common duct stones. Stones may develop in the bile duct in diseases causing chronic biliary obstruction such as benign bile duct strictures and sclerosing cholangitis.

Clinical features

The classic triad of symptoms is right upper abdominal pain, jaundice, and fever. The abdominal pain is typically colicky, severe, and persists for hours. It is often associated with vomiting. Fever and rigors indicate cholangitis, which commonly accompanies bile duct stones. Jaundice is variable; it may be mild or deep and is often intermittent. The urine is dark due to conjugated bilirubin and the faeces are pale. Frequently, the amount of pigment in the faeces varies. Itching may be prominent. However, common bile duct stones may also be silent, especially in the elderly. Alternatively, only one of the triad of symptoms may be present; the patient presenting with jaundice, abdominal pain, or cholangitis. The liver is moderately enlarged and there may be tenderness in the right upper quadrant. Prolonged biliary obstruction lasting months or years eventually leads to biliary cirrhosis with portal venous hypertension and liver cell failure.

Laboratory investigations

Liver function tests show a cholestatic (biliary obstructive) pattern. The prothrombin time may be prolonged due to inadequate absorption of vitamin K. A polymorphonuclear leucocytosis is common and indicates biliary infection. Blood cultures should be performed repeatedly during the fevers to isolate the organism and determine sensitivities.

Imaging techniques

A plain radiograph of the abdomen will show calcified gallstones in 10 per cent of patients, but is rarely performed now. Ultrasonography is useful for demonstrating the dilated biliary tree that results from obstruction and may reveal biliary gallstones. Unfortunately, ultrasound frequently fails to detect common duct stones obstructing the lower end of the bile duct. Cholangiography by ERCP is required in these patients (Fig. 5). Common bile duct stones should be removed by endoscopic sphincterotomy before the patient is submitted to cholecystectomy.

Fig. 5 Choledocholithiasis. An endoscopic retrograde cholangiogram shows multiple faceted radiolucent stones in a dilated bile duct. The gallbladder has not been opacified.

Differential diagnosis

Common duct stones are the most common cause of cholestatic (biliary obstructive) jaundice. Next in frequency are carcinomas of the head of the pancreas, bile duct, and ampulla of Vater (Table 1). Intrahepatic diseases may also cause a cholestatic jaundice; the causes include viral and alcoholic hepatitis, drugs, and pregnancy.

Treatment

Common bile duct stones must be removed. The optimal treatment is endoscopic sphincterotomy to remove bile duct stones followed by laparoscopic cholecystectomy. This approach avoids the hazards of open exploration of the common bile duct. Endoscopic removal of common duct gallstones without cholecystectomy is appropriate in patients unfit for surgery. Few patients will have further problems from the gallbladder that remains. Stones overlooked at surgery (residual calculi) are best treated by endoscopic sphincterotomy or, if a T-tube is in place, removed by a steerable basket-catheter manipulated down the T-tube track. Open exploration of the common bile duct is required if gallstones are too large to be removed endoscopically (more than 2 cm). Preoperative preparation includes appropriate antibiotics for cholangitis, the correction of fluid and electrolyte balance, nutrition, and anaemia, and if the prothrombin time is prolonged, parenteral vitamin K.

Postcholecystectomy syndromes

After cholecystectomy a proportion of patients continue to complain of symptoms such as right upper quadrant pain, flatulence, and fatty food intolerance. However, the vast majority of patients with gallstones are improved by surgery. The persistence of symptoms in many is probably a consequence of the wrong diagnosis being made before surgery and other diseases such as oesophagitis, pancreatitis, or functional bowel disease should be sought. In others, technical problems during surgery may have resulted in a benign post-traumatic biliary stricture or residual calculi. However, there remains a group of patients where the cause appears to be due to less common biliary disorders such as long, dilated cystic duct remnants, amputation neuromas of the cystic duct, and spasm or stenosis of the sphincter of Oddi. The biliary tract must be carefully investigated in these patients, especially if colicky pain, fever, jaundice, or cholestatic liver function tests persist. Biliary tract manometry is of value when spasm or stenosis of the sphincter of Oddi is suspected.

Biliary infections

Bacterial cholangitis (suppurative cholangitis)

This is usually associated with common bile duct calculi and benign biliary structures. Malignant structures produce complete obstruction and the bile remains sterile. Other conditions associated with cholangitis are biliary enteric fistulas—both spontaneous and surgical—sclerosing cholangitis, and congenital intrahepatic biliary dilation (Caroli's disease). Organisms of the gut flora are usually cultured in these infections, including aerobes such as *E. coli*, *Streptococcus faecalis*, *Proteus vulgaris* and staphylococci, and anaerobes such as bacteroides, aerobacter, and anaerobic steptrococci.

Clinical features and treatment

The onset of malaise, fever, and rigors is followed by pain, vomiting, jaundice, and itching. The urine turns dark and the faeces pale. The biliary obstructive features are probably due to oedema of the bile duct wall. Recurrent attacks are common. Hepatic abscesses may result. Repeated blood cultures are performed during the fever to isolate the organisms. Culture of a liver biopsy fragment may also yield the organism. The main element of treatment is drainage of the biliary tract, which is best achieved by emergency endoscopic sphincterotomy. Additionally, appropriate antibiotics such as cefuroxime and metronidazole are given. For recurrent attacks of cholangitis, tetracycline, amoxicillin, or cephalexin are usually effective.

Infestations

Infestations (see Section 7) with the roundworm *Ascaris lumbricoides* and the liver fluke *Clonorchis sinensis* are particular problems of the Far East. Both lead to cholangitis. *C. sinensis* infestation predisposes to bile duct carcinoma and primary liver cancer. The common sheep fluke *Fasciola hepatica* may be encountered as a cause of cholangitis in Europe during wet summers.

Benign biliary strictures

In about 95 per cent of patients these are a consequence of biliary tract surgery. The remainder are caused by gallstones eroding the bile duct and, rarely, blunt injury to the abdomen. Signs of biliary stricture may be detected in the immediate postoperative period but are often delayed. Disasters such as ligation or section of the bile duct present early with jaundice and drainage of bile from the wound drains. With lesser damage to the duct the patient presents after an interval with cholangitis and jaundice. Liver function tests reveal a cholestatic pattern and blood cultures may yield an organism. The precise delineation of the stricture requires ERCP or PTC. Biliary stricture is not a benign condition; untreated it will usually progress to biliary cirrhosis with portal venous hypertension and liver failure. Treatment is surgical and should be performed by a surgeon skilled in this difficult repair.

Malignant biliary stricture

This is most commonly due to adenocarcinoma of the head of the pancreas but may also be caused by adenocarcinomas of the bile ducts, of the ampulla of Vater, and rarely of the gallbladder. Occasionally the cause is lymph node enlargement at the porta hepatis due to malignant metastases or lymphoma.

Symptoms and signs

Cancers of the pancreas and biliary tree (Fig. 6 and Fig. 7) usually affect middle-aged and elderly individuals. The onset is insidious with deepening jaundice, itching, and weight loss. A dull nagging upper abdominal pain which radiates to the back is common. In contrast to choledocholithiasis and benign strictures, cholangitis is unusual. Examination reveals a deeply jaundiced patient often excoriated from scratching. The liver is enlarged but not tender. If the malignant obstruction is below the level of the cystic duct, the gallbladder is distended and may be palpable (Courvoisier's law). The urine is dark and the stools pale. In cancer of the ampulla of Vater, a film of blood on the pale stool may give it a silvery colour ('silver stools').

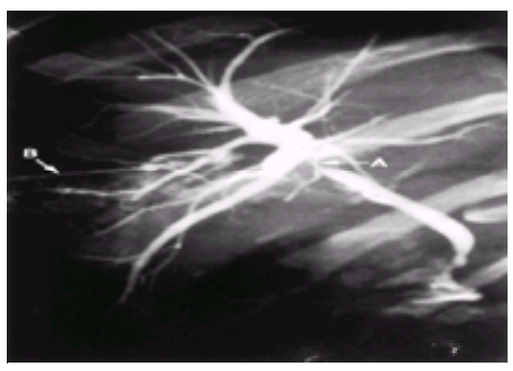

Fig. 6 Carcinoma of the bile duct. A percutaneous transhepatic cholangiogram (PTC) shows a stricture (a) high in the bile duct at the porta hepatis intrahepatic bile ducts are moderately dilated. The transhepatic track of the 'skinny' needle used for the PTC is also visible (b).

Fig. 7 Carcinoma of the pancreas. The percutaneous transhepatic cholangiogram shows a very dilated biliary tree which terminates in a blunt 'nipple-like' obstruction (arrow) at the lower end of the common bile duct. This is the usual finding in the cancers of the head of the pancreas which obstruct the bilary system.

Laboratory investigation

Liver function tests reveal a cholestatic pattern. The serum bilirubin may be very high (600 µmol/l; 35 mg/dl). A microcytic hypochromic anaemia indicates blood loss from the tumour.

Imaging techniques

An ultrasound or CT scan examination will reveal dilatation of the biliary tree and may demonstrate the level of the obstruction. Ultrasound-guided percutaneous needle biopsy may be employed to provide a histological diagnosis. Bile duct carcinoma frequently causes obstruction at the porta hepatis and, consequently, at laparotomy the extrahepatic biliary tract appears non-dilated. Even if operative cholangiography is performed, the constrast medium frequently fails to pass the obstruction and fill the dilated intrahepatic biliary tree. Therefore it is important to establish the diagnosis precisely before surgery is contemplated by performing an ERCP or PTC. This is particularly important because most of these patients are best treated by endoscopic or percutaneous biliary stents rather than surgery (see below).

Treatment

Occasionally small tumours confined to the head of the pancreas and ampulla of Vater may be treated curatively by a Whipple's operation. Unfortunately the great majority of pancreatic and bile duct cancers can only be treated palliatively with a bypass procedure such as a cholecystojejunostomy. The prognosis for these patients is poor. An alternative treatment is endoscopic or percutaneous transhepatic introduction of prostheses (stents) through the biliary stricture. Patients with endoscopic prostheses have the same median survival as those with surgical bypass procedures, but the operative mortality and morbidity rate is much lower for endoscopic prostheses. Endoscopic prostheses are the preferred treatment for unresectable biliary and pancreatic cancers. The prostheses may block after about 3 months and need to be replaced.

Other causes of bile duct obstruction

Pancreatitis may obstruct the common bile duct during its passage through the head of the pancreas. Transient jaundice is common in acute pancreatitis due to compression by pancreatic oedema. In chronic pancreatitis, especially alcoholic, persistent jaundice can develop requiring a surgical bypass procedure such as a cholecystojejunostomy. This biliary obstruction is probably a consequence of pancreatic fibrosis. Pancreatic cysts may rarely cause extrinsic compression of the bile duct. Haemobilia or haemorrhage into the biliary tract is uncommon but may follow trauma, liver biopsy, biliary tumours, and gallstones. In addition to jaundice, the blood clots cause biliary pain. Massive gastrointestinal haemorrhage may occur. The diagnosis of these conditions relies on accurate cholangiography (usually ERCP).

Sclerosing cholangitis

Sclerosing cholangitis is the description applied to multiple strictures and bead-like dilatations of the intrahepatic and extrahepatic biliary tree.

Primary sclerosing cholangitis (Fig. 8)

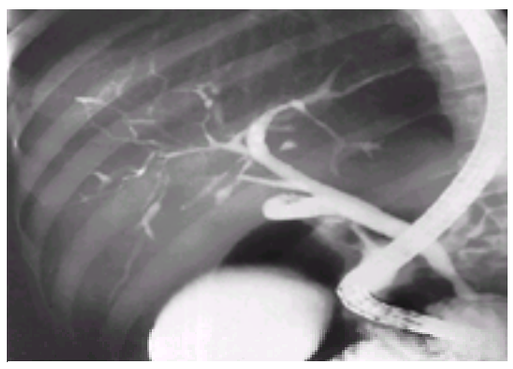

Fig. 8 Primary sclerosing cholangitis. The intrahepatic bile ducts show alternate strictures and dilatations ('beading'). The common bile duct, cystic duct, and gallbladder appear normal in this study but may also be involved.

This should only be diagnosed if the following criteria are satisfied: (i) absence of gallstones; (ii) absence of previous biliary surgery; and (iii) sufficiently long follow-up to exclude carcinoma of the bile duct. Primary sclerosing cholangitis affects males more than females (2:1) and about 70 per cent of patients have ulcerative colitis. The usual clinical presentation is cholestatic jaundice and cholangitis. However, a significant proportion of patients are asymptomatic or present with cirrhosis and portal venous hypertension. There is associated retroperitoneal fibrosis or Riedel's thyroiditis in some cases. Serum biochemistry shows cholestatic liver function tests. A raised serum alkaline phosphatase is almost invariable. Consequently the diagnosis should be considered in patients with cirrhosis whose liver function tests show cholestatic features. The IgM concentration is commonly elevated. Liver biopsy may be helpful and usually indicates large bile duct obstruction. The diagnosis is established by cholangiography with ERCP or PTC. Laparotomy should not be performed. Lone tight strictures and stones can be treated by endoscopic techniques. Primary sclerosing cholangitis is being recognized more frequently as a result of the widespread use of ERCP and PTC. It may be confused with primary biliary cirrhosis, but the serum mitochondrial antibody is always negative in primary sclerosing cholangitis. Treatment is unsatisfactory, neither corticosteriods nor azathioprine are of proven value. Ursodeoxycholic acid improves liver function tests but has not been shown to prolong survival. Pruritus may be helped by cholestyramine. The prognosis is variable but most patients eventually develop cirrhosis and liver failure. Liver transplantation yields excellent results in these patients. Bile duct adenocarcinoma is a late complication.

Secondary sclerosing cholangitis

Several causes of secondary sclerosing cholangitis are now recognized. These include recurrent bacterial cholangitis due to gallstones or benign biliary strictures. Children with primary immunodeficiency syndromes and patients with AIDS also develop sclerosing cholangitis. Cytomegalovirus and cryptosporidium are the organisms most commonly associated with AIDS-related sclerosing cholangitis. Sclerosing cholangitis may also develop in patients treated by hepatic arterial infusion of cytotoxic drugs and after the introduction of caustics into hydatid cysts.

Congenital disorders of the gallbladder and biliary tract

This subject is discussed in Chapter 14.19.1.

Further reading

Angulo P, Lindor KD (1999). Primary sclerosing cholangitis. *Hepatology* **30**, 325–32.

Donovan JM (1999). Physical and metabolic factors in gallstone pathogenesis. *Gastroenterology Clinics of North America* **28**, 75–97.

Ko CW, Lee SP (1999). Gallstone formation. Local factors. *Gastroenterology Clinics of North America* **28**, 99–115.

Schiff L, Schiff ER (1993). *Diseases of the liver*, 7th edn. Lippincott, Philadelphia.

Sherlock S, Summerfield JA (1991). *A colour atlas of liver disease*, 2nd edn. Wolfe Medical Publications, London.

Sherlock S, Dooley JS (1997). *Diseases of the liver and biliary system*, 10th edn. Blackwell Scientific Publications, Oxford.

14.19.3 Jaundice

R. P. H. Thompson

Jaundice is the clinical sign of hyperbilirubinaemia, and hence usually indicates disease of the liver or biliary tree. The pigment in the tissues in best seen as yellowing of the sclera; eventually the skin and soft palate become tinted, but not saliva or sputum. The urine usually becomes dark. Rarely, carotenaemia, from eating excessive carrots or vitamin A, can mimic jaundice, but its colour is more prominent in the palms than the sclera.

Physiology of bilirubin (Fig. 1)

Fig. 1 The porphyrin–bilirubin pathway.

All haem molecules in haemoglobin or cytochrome enzymes are stoichiometrically (1:1) degraded to bilirubin, especially in the spleen and liver, but also in macrophages in other tissues, including skin, and in renal tubular cells. Haem oxygenase enzymes break open the asymmetric tetrapyrrole haem molecule specifically at the a-methene bridge, releasing carbon monoxide and iron. One principal isomer of biliverdin, namely IXa, is formed, although small amounts of the other three possible isomers (b, g, and †) can be detected in bile. The excretion of carbon monoxide in breath can be used quantitatively to determine the breakdown of haem to bilirubin, of which 200 to 350 mg (340 to 600 µmol) is produced daily. About 85 per cent of biliverdin and bilirubin is derived from the delayed breakdown of the haemoglobin in effete red blood cells, while the remainder is either from the breakdown of haem proteins, chiefly in the liver, or from ineffective erythropoiesis in the bone marrow; these constitute the so-called 'early labelled' bilirubin, defined by isotopic studies *in vivo*.

Biliverdin is green and is directly excreted in bile by birds, amphibians, and reptiles but not by mammals. Biliverdin is reduced by the cytosolic enzyme biliverdin reductase in liver and spleen to the yellow bilirubin IXa, which has then to be excreted. The reason for this difference was obscure, for bilirubin is lipid soluble, and potentially toxic, and has to be conjugated before it is excreted in bile, while biliverdin is water soluble and can be readily excreted in urine and bile by mammals. However, bilirubin is an antioxidant or free radical scavenger in plasma and bile, particularly when bound to copper, and this may be especially important in the neonate, especially when levels of the antioxidant ascorbate are low. Hence bilirubin probably has a function and is not just a waste product. Bilirubin is surprisingly lipid soluble, this being due to internal hydrogen bonding in the molecule so that it forms a tight, non-polar, non-linear, three-dimensional structure. After its release from macrophages, it is firmly bound to plasma albumin, so that none enters the urine. At high concentrations in the blood it slowly diffuses into tissues, where it can be toxic, particularly in the neonatal brain (kernicterus) or kidney. Jaundice is usually less obvious in unconjugated hyperbilirubinaemia since its diffusion into the tissues is more limited.

The circulating pool of bilirubin in the plasma (about 100 µmol) is almost all unconjugated. Routine measurements still rely on the Van den Bergh diazo reaction, which yields either an indirect (unconjugated bilirubin) or direct reaction (conjugated) and, although this overestimates the true level of conjugated bilirubin, the results indicate whether or not circulating bilirubin is wholly unconjugated. The direct and indirect reactions depend on the slow reaction of the unconjugated bilirubin with the reagent, which is accelerated when solvents such as methanol, which break the internal hydrogen bonding, are added. The normal range of plasma bilirubin is wide (about 5 to 19 µmol/l), reflecting wide variation in the rate of conjugation in the liver, and is higher than in most other mammals in whom clearance and excretion are more efficient. The distribution of values is Gaussian, so that the true upper limit of normal is arbitrary (see Gilbert's syndrome). Hepatic enzyme-inducing drugs reduce the plasma level by increasing conjugation and hence clearance.

Bilirubin is selectively removed by hepatocytes from sinusoidal blood, although its plasma clearance (about 50 ml/min) is low compared, for instance, with that of bile acids, and hence its extraction (1.5 per cent of plasma pool/min) is dependent more upon hepatocyte distribution and function than on blood flow. It is initially surprising that bilirubin can be displaced from its plasma binding sites and enter hepatocytes, and specific hepatic cytoplasmic binding proteins have been described. Nevertheless, binding to the active site of the microsomal conjugating enzyme uridyl diphosphate (**UDP**)-glucuronyl transferase should be sufficient to maintain a low level of free bilirubin in the cytoplasm, which, without the need for specific transfer proteins, should alone produce a gradient sufficient to allow bilirubin slowly to enter the hepatocyte. Uptake of bilirubin is facilitated by the direct contact of plasma with the hepatocyte in the interstitial space of Disse through fenestrations in the endothelium of hepatic blood capillaries. Although uptake into the cell predominates, dynamic studies show that there is considerable reflux of bilirubin back out of the cell to the plasma.

Within the hepatocyte bilirubin is principally conjugated by one of the two specific isoforms of the microsomal enzyme UDP-glucuronyl (glucuronate-glucuronosyl) transferase, chiefly with two glucuronic acid moieties. Conjugated bilirubin is excreted out of the endoplasm reticulum and then across the microvillous intercellular canalicular membrane by the anionic conjugate transporter protein (**mrp2**; or cMOAT). Hepatocytes have a separate canalicular bile salt export pump protein (bsep). mrp2 also transports other multivalent anions, such as conjugated bromsulphthalein. Minor quantities of bilirubin are conjugated with one glucuronic acid molecule (monoglucuronide) or with combinations of related sugars (xylose, glucose); a small amount of unconjugated bilirubin also appears in bile. The chemical properties of the conjugated molecules are quite different from those of unconjugated bilirubin, for there is no internal hydrogen bonding of bilirubin—they now become more linear, fully water-soluble molecules and are efficiently excreted in bile. In many liver diseases conjugated bilirubin readily refluxes back into blood and, since it is water soluble and less firmly bound to albumin than unconjugated bilirubin, about 1 per cent is filtered across the glomerular membrane and enters the urine (choluria).

Excretion of conjugated bilirubin is increased by the bile acids that also accumulate in cholestasis. If renal function is normal, renal excretion of bilirubin matches production when conjugated bilirubin levels in the plasma reach about 600 µmol/l. With renal failure, or haemolysis, plasma levels rise higher. Little bilirubin, even conjugated bilirubin, diffuses through renal dialysis membranes.

Recently it has been shown that deconjugated bilirubin can undergo a substantial enterohepatic circulation; it is absorbed from the colon, particularly when there is bile acid malabsorption and hence the concentration of bile acids is increased in the colon, for example as a result of ileal disease or resection. This reabsorption increases the concentration of bilirubin re-excreted in bile, and may in part explain the increased incidence of pigment gallstones in patients with ileal disease. Similarly, fasting increases unconjugated bilirubin levels in the plasma by increasing the reabsorption of bilirubin.

In the distal intestine, conjugated bilirubin is deconjugated and reduced to a series of sterco- and urobilinogens that give the normal colour to faeces. Some colourless urobilinogen is normally absorbed from the colon and undergoes an enterohepatic circulation, with a small amount being excreted in urine. If this circulation and biliary excretion is impaired in liver disease, or increased in haemolysis, then excess urobilinogen is excreted in urine, where it can oxidize on standing to brown urobilins. Urobilinogen is easily detected by routine clinical 'stix'. Ehrlich's aldehyde reagent was at one time used, and when added to urine containing excess urobilinogen it turns it red; the urobilinogen pigment can then be extracted into an organic solvent such as chloroform, unlike the pigment formed from the more polar porphobilinogen adduct in acute porphyria, which remains in the aqueous phase.

Management of jaundice

Complex algorithms of management of the patient with hyperbilirubinaemia or jaundice have been published, but a simple pragmatic approach is proposed here (Fig. 2).

Fig. 2 Investigation of jaundice.

Raised plasma bilirubin levels, and eventually frank jaundice, are due primarily to excessive unconjugated or conjugated bilirubin in blood, depending on whether the abnormality in bilirubin metabolism is in its production and/or conjugation, or in the subsequent hepatic excretion of conjugated bilirubin, respectively. Impaired excretion is almost always combined with impaired bile flow and is best termed cholestasis. When other liver-related blood tests are abnormal, especially the biliary enzymes alkaline phosphatase and g-glutamyl transpeptidase, serum bile acids are also raised, there is often itching, and the microvilli lining the biliary canaliculi are injured. Examination of a liver biopsy specimen taken from a patient with cholestasis under light microscopy may show bile plugs. This finding is, however, often termed 'obstructive jaundice', an unfortunate term, since this implies extrahepatic obstruction of the biliary tree. The molecular events underlying some forms of cholestasis are now being unravelled (see below).

History

Dark urine and, less commonly, pale stools indicate cholestasis. Many drugs, including alcohol, can be a cause of unconjugated and conjugated hyperbilirubinaemia and should be rigorously enquired about. Fever (hepatitis, cholangitis, abscesses), travel (hepatitis, amoebiasis), sexual history (hepatitis A, B, or C), surgery and anaesthesia (postoperative jaundice, see below; biliary tract disease), herbal medicines (e.g. West Indian teas, Chinese herbs), and transfusions and blood products (hepatitis B or C) can be important clues.

Clinical examination

Stigmata of chronic liver disease (e.g. spider naevi, facial telangietases, parotid enlargement, Dupuytren's contractures, muscle wasting, hepatosplenomegaly, and ascites) are important, but do not define the cause of jaundice.

Testing for unconjugated hyperbilirubinaemia

If serum bilirubin alone is abnormal among the liver-related blood tests, unconjugated hyperbilirubinaemia should first be excluded by testing whether the bilirubin in blood is predominantly conjugated or unconjugated. A normal reticulocyte count will usually exclude haemolysis severe enough to cause raised bilirubin levels and blood film examination may be additionally informative. Suspected haemolysis is investigated as described elsewhere. If no cause of unconjugated hyperbilirubinaemia is identified, then benign constitutional unconjugated hyperbilirubinaemia (Gilbert's syndrome) is diagnosed (see below).

Conjugated hyperbilirubinaemia

The familial syndromes without cholestasis (Dubin–Johnson and Rotor) are rare (see below).

Routine liver-related blood tests cannot differentiate between intra- or extrahepatic causes of jaundice, unless the transferases are very high (e.g. more than 1000 IU/l), in which case hepatitis (e.g. viral, alcoholic) is certain. A greatly raised alkaline phosphatase level does not necessarily imply an extrahepatic lesion; intrahepatic causes are common (Table 1). Research methods for assessing liver function (e.g. galactose tolerance test, aminopyrine breath test) are of no value in the management of the patient with jaundice.

Cholestasis should be investigated first with abdominal ultrasonography, which should accurately detect a dilated intra- and/or extrahepatic biliary tree and often also reveal its cause (e.g. gallstones, tumour). Cholecystography or intravenous cholangiography will fail in the presence of jaundice. If biliary disease is thus suspected, then an endoscopic retrograde cholangiogram (ECRP) or, failing that, a fine-needle, percutaneous, transhepatic cholangiogram (PTC), will define the anatomy more accurately and often provide definitive therapy (removal of biliary stones, stenting), thus often avoiding surgery. Magnetic resonance cholangiography (MRC) is increasing in sensitivity and availability and can now produce high-quality non-invasive images of the biliary tree and pancreas; it is useful even for intrahepatic biliary disease. It cannot, of course, be therapeutic. Endoscopic ultrasonography (EUS) can show accurately the presence of stones, biliary or pancreatic tumours, and sclerosing cholangitis, while g-camera scans with technetium-labelled **HIDA** (hepato-iminodiacetic acid) can be used to indicate biliary obstruction, particularly in the neonate, if ultrasonography is normal.

If intrahepatic cholestasis is suspected because of a non-dilated biliary tree on ultrasound, the following tests should be considered: hepatitis A, B, or C serology, autoantibodies (antimitochondrial for primary biliary cirrhosis, smooth muscle and liver–kidney microsomal for autoimmune chronic hepatitis), serum caeruloplasmin and copper for Wilson's disease if less than 40 years of age, or plasma a $_1$-antitrypsin concentrations for homozygous deficiency. Intrahepatic masses on ultrasound will prompt measurement of a-fetoprotein for primary hepatoma and other tumour markers. A percutaneous needle liver biopsy (or aspiration of an abscess) may then be indicated, provided that blood coagulation parameters and the platelet count are normal. Guidelines for liver biopsy have recently been published. A transjugular venous approach for the biopsy is appropriate if the risks of bleeding are increased.

Unconjugated hyperbilirubinaemia

Plasma bilirubin levels are exponentially and positively related to the half-life of circulating red blood cells (i.e. bilirubin load), and negatively to the hepatic clearance rate of bilirubin. This relationship is analogous to that of muscle breakdown, plasma creatinine, and glomerular filtration rate. Hence, as the rate of haemolysis rises or clearance falls, bilirubin levels may rise rapidly in response to small changes of the input load or removal rate from plasma, or both.

Haemolytic jaundice is most commonly encountered in the haemoglobinopathies of sickle-cell anaemia (homozygous SS or heterozygous SC disease) or homozygous thalassaemia major, although dark skin may render it difficult to detect. The 'acholuric jaundice' of hereditary spherocytosis is rare. Mildly elevated bilirubin levels are described in ineffective erythropoiesis of the bone marrow, in vitamin B_{12} deficiency (pernicious anaemia), or in thalassaemia minor.

Drugs may cause haemolysis (e.g. methyl DOPA, sulphasalazine), or impair hepatic bilirubin clearance (e.g. rifampicin). Infections (e.g. malaria) or mismatched blood transfusions can produce massive haemolysis, but this overshadows the raised bilirubin levels. Autoimmune haemolytic anaemia, glucose-6-phosphate dehydrogenase deficiency, and haemolysis due to leaking prosthetic cardiac valves may cause obvious clinical jaundice that may escape diagnosis before the true nature of the hyperbilirubinaemia is recognized.

Familial unconjugated hyperbilirubinaemia

A series of defects of the hepatic conjugating enzyme UDP-glucuronyl transferase produce various degrees of unconjugated hyperbilirubinaemia due to impaired bilirubin clearance; they have long fascinated physiologists and more recently molecular biologists.

At least 3 per cent of the normal adult population have mildly raised unconjugated bilirubin levels in blood that rise excessively on fasting. This 'phenomenon' is commonly termed Gilbert's syndrome, although it is unclear whether the eponym is justified. The raised concentrations of bilirubin develop in early adult life and are often associated with mild degrees of haemolysis. Various associated defects of hepatic drug metabolism have also been described and these are probably linked genetic abnormalities. Any combination of an increased bilirubin load from the haemolysis and a mildly impaired clearance will increase plasma bilirubin concentrations more than would either alone, and hence together they bring the condition to notice. It is probably not a discrete entity but rather different defects of conjugation that elevate bilirubin levels above an arbitrary upper limit of normal. Determination of the bilirubin-conjugating capacity of liver biopsy tissue has shown that the activity of glucuronyl tranferase is reduced by 60 to 70 per cent, and this impairs bilirubin clearance.

Gilbert's syndrome is recognized by a fluctuating, raised serum bilirubin concentration with other routine liver-related blood tests being normal, and a normal reticulocyte count to exclude overt haemolysis. It can be confirmed by measuring the unconjugated fraction of the bilirubin, which should be greater than 90 per cent. Measuring the pronounced increase of plasma bilirubin that occurs after a 48-h fast on 400 kcal/day or provocation with intravenous nicotinic acid are now generally considered to be research procedures of little clinical value. A liver biopsy is rarely needed. Reassurance that the results do not indicate liver disease and will not affect life insurance is important. Plasma bilirubin concentrations rise in patients with Gilbert's syndrome during intercurrent illness when frank jaundice may be observed.

It is said that Gilbert's syndrome can follow an attack of viral hepatitis, although this may be simply due to ascertainment bias in a population with a high underlying prevalence of the biochemical anomaly.

Crigler–Najjar syndromes

Two syndromes of more severe unconjugated hyperbilirubinaemia have been described, namely the rare type I (100 cases reported), which often causes neonatal death, and the more common, and benign type II. Both are due to severe deficiency in the UDP-glucuronyl transferase enzymes.

In type I, first reported in 1952, with an autosomal recessive inheritance, neonates rapidly become progressively jaundiced in the first days of life (bilirubin levels reach 350 to 950 µmol/l) and, if untreated, develop kernicterus or brain damage. Death usually occurs within a year but delayed kernicterus has been reported. There is no conjugated bilirubin in bile, but small quantities of unconjugated bilirubin can be found in bile and cross the intestinal wall.

The inheritance of type II is complex, and is reported both to be dominant with incomplete penetrance or autosomal recessive. Bilirubin levels are lower (less than 350 µmol/l), and persistent mild jaundice is only noticed in childhood. Brain damage does not occur, and the only problem is cosmetic. One-third of the conjugated bilirubin in bile is present as the monogluronide (normally less than 10 per cent).

In the Gunn strain of rat, severe unconjugated hyperbilirubinaemia occurs and glucuronyl transferase activity is absent in the liver, as it is in Crigler–Najjar type I. In type II, enzyme activity is less than 10 per cent of normal, but measurable.

It has long been known that there is a spectrum of bilirubin levels in type II Crigler–Najjar and Gilbert's syndromes and indeed both conditions have been observed within families, suggesting different degrees of enzyme activity. Phenobarbitone or other hepatic microsomal enzyme-inducing agents markedly reduce bilirubin levels in Gilbert's and Crigler–Najjar type II syndromes, although unfortunately not in Crigler–Najjar type I, and increase the activity of glucuronyl transferase. Such treatment, however, is not needed, except possibly for cosmetic purposes.

Molecular analysis of the genes encoding human UDP-glucuronyl transferases has clarified the genetic basis and inheritance of these disorders. The two complementary DNAs for the two human bilirubin UDP-glucuronyl transferase isoforms have been sequenced; they differ from those that encode the other glucuronyl transferases that conjugate, for instance, steroids. The UDP-glucuronyl transferases map to human chromosome 2, where at least seven exons encode the specific mRNAs of the isozymes, together with a common region for the glucuronyl transferases. Analysis of DNA from five patients with type I Crigler–Najjar syndrome has identified homozygous or heterozygous defects in the common structural exons encoding the two bilirubin glucuronyl transferase isoforms—similar defects occur in the Gunn rat.

In Crigler–Najjar type II syndrome, initial studies have found mutations in the upstream regulating region of the gene encoding the active bilirubin glucuronyl transferase isoform. Phenobarbitone induces the expression of the abnormal enzyme, explaining its efficacy in this condition. Probably a heterozygous combination of an abnormality of the promoter region (Gilbert's defect; see below) and this Crigler–Najjar defect is responsible for the phenotype of type II Crigler–Najjar syndrome. This explains the presence of patients with Gilbert's syndrome within families with type II Crigler–Najjar.

The genetic basis of Gilbert's syndrome remains controversial but it appears that it is an autosomal recessive condition in which there is a homozygous abnormality in the promoter region affecting expression of the specific glucuronyl transferase gene. Heterozygotes have normal bilirubin levels. There must be another factor responsible for the increased bilirubin concentrations as the heterozygote abnormality occurs in 40 per cent of healthy individuals without hyperbilirubinaemia. Moreover, some individuals who are homozygous for the defect have normal plasma concentrations of conjugated bilirubin. The bilirubin load from red cell breakdown will influence plasma concentrations.

Crigler–Najjar type I syndrome can now be successfully treated by whole-body blue-light phototherapy for 16 h daily (see below), or by plasmapheresis until orthotopic liver transplantation can be carried out as a definitive cure. Severe kernicterus is a contraindication to transplantation as it is not reversible. About 20 patients have so far received transplants, with two deaths. Hepatocyte transplantation, in which donor hepatocytes are infused into the portal vein, has been partially successful, and clearly has potential for development. Drugs that displace unconjugated bilirubin from albumin (sulphonamides, salicylates, penicillin) will increase brain damage in Crigler–Najjar syndrome and must be avoided.

Neonatal jaundice

Unconjugated hyperbilirubinaemia, often with mild clinical jaundice, occurs in all full-term newborn infants. Bilirubin concentrations are maximal at 2 to 5 days after birth but the plasma bilirubin rarely exceeds 90 µmol/l; neonatal jaundice is more severe in premature infants. It is attributed to a combination of immaturity of hepatic glucuronyl transferase and the added load of bilirubin from rapid haemolysis of surplus fetal red blood cells in the neonatal period. Before birth, fetal bilirubin is

excreted by the mother, and meconium as well as stools are pale because of the reduced excretion of bilirubin.

If haemolysis is increased, as in rhesus or other fetomaternal incompatibility of red cell antigens, causing haemolysis of fetal red blood cells by maternal antibodies, then severe jaundice and kernicterus can occur. Acidosis and some drugs (sulphonamides, salicylates, penicillin) may increase kernicterus by displacing unconjugated bilirubin from albumin. Glucose-6-phosphate deficiency can also cause jaundice and anaemia in the neonatal period and is usually observed in infants of Mediterranean, African, or Chinese ancestry.

Treatment with phenobarbitone induces hepatic glucuronyl transferase but its effect is slow unless given to the mother before birth. Exchange transfusion or plasmapheresis are more effective. Phototherapy, namely exposure of the near-naked infant to blue light in an incubator, is also very effective. Being yellow, bilirubin absorbs light at approximately 450 nm and is sensitive to light, which oxidizes it to water-soluble, non-toxic products. Hence, exposure of the bilirubin in skin capillaries to light reduces plasma concentrations; the breakdown products are excreted safely in urine and bile. Reabsorption of bilirubin from the intestine can also be reduced by giving agar by mouth, thus interrupting the enterohepatic circulation.

Breast feeding slightly increases bilirubin levels and about 1 in 40 breast-fed infants develop jaundice, which remits on transfer to cow's milk for 24 h; this jaundice does not always recur when breast milk is reintroduced. Breast feeding increases the enterohepatic cycling of bilirubin from the intestine, since stool weights and frequency are less than with formula feeds. A further effect of steroid molecules that inhibit glucuronyl transferase activity in the neonatal liver and that are present in breast milk has also been postulated.

Hypothyroidism increases jaundice and should be sought in patients with unexplained hyperbilirubinaemia since it may not be associated with obvious neonatal cretinism. The rare Crigler–Najjar type I syndrome presents with florid jaundice in the first few days of life.

Sickle-cell anaemia and b-thalassaemia

Jaundice is common in homozygous sickle-cell anaemia due to the unconjugated hyperbilirubinaemia from persistent haemolysis. During crises jaundice often deepens in association with increasing anaemia, suggesting accelerated haemolysis although transient bone marrow failure may also occur. Occasionally, conjugated hyperbilirubinaemia with dark urine occurs during these episodes, and hepatic histology may show areas of necrosis due to thrombosis and bile thrombi. Patients with sickle-cell anaemia are also prone to pigment gallstones, due to the excessive bilirubin excreted, and these can cause extrahepatic biliary obstruction. Unconjugated and conjugated hyperbilirubinaemia both occur in homozygous thalassaemia as a result of increased red cell destruction and intramedullary haemolysis associated with ineffective erythropoiesis.

Cholestasis

There are many causes of intrahepatic cholestasis (Table 1).

Neonatal cholestasis

Conjugated hyperbilirubinaemia or cholestasis in the neonate, with dark urine and pale stools, is always pathological and if it continues beyond 2 weeks of age requires urgent investigation. There are many causes.

In many instances the cause is never established and although once called neonatal hepatitis, it is better termed the hepatitis syndrome; hepatic histology shows hepatitis, sometimes with giant cells. Some babies recover, while perhaps half progress to hypoplasia of the intrahepatic bile ducts, which then overlaps with extra- and intrahepatic biliary atresia.

Infections, particularly urinary, can cause transient cholestasis. Syphilis is now rare, as is toxoplasmosis. Various viral infections (rubella, cytomegalovirus) can cause neonatal jaundice. The hepatotrophic hepatitus B virus contracted from an HBe antigen-positive mother rarely causes jaundice. Metabolic diseases that may be causes of neonatal jaundice include galactosaemia, hereditary fructose intolerance (fructosaemia), and tyrosinosis—all of which need to be diagnosed quickly so as to start dietary treatment early—as well as homozygous a_1-antitrypsin deficiency, and intravenous feeding *per se*. Other genetic diseases include trisomy 13 and trisomy 18 (one-quarter of babies developing the hepatitis syndrome) and cystic fibrosis.

Several familial syndromes presenting with neonatal cholestasis have been described, some with other congenital abnormalities, such as arteriohepatic dysplasia (Alagille's syndrome), and others solely with progressive familial cholestasis, with persistent jaundice, raised serum bile acids, hepatosplenomegaly, steatorrhoea, and failure to thrive, such as Byler's syndrome in Amish families. Bile duct hypoplasia, cirrhosis, and liver failure often follow, unless liver transplantation is carried out. There are several different mutations in the various syndromes of progressive familial cholestasis that affect the function of the FIC-1 gene, so that canalicular biliary bile acid and conjugated bilirubin excretion are severely impaired, and hence cholestasis develops.

Extrahepatic cholestasis in the neonate is most commonly due to biliary atresia, but choledochal cyst or bile duct perforation can also cause jaundice at this age. Biliary atresia appears to represent a form of sclerosing cholangitis with progressive loss of intra- and extrahepatic ducts. HIDA scans, percutaneous liver biopsy, and retrograde cholangiography can establish the diagnosis without laparotomy.

Benign recurrent intrahepatic cholestasis

In this rare syndrome, recurrent reversible episodes of cholestasis start in childhood or adult life. Each attack is characterized by jaundice, anorexia, and itching for several months, which then subsides with no residual effects. Hepatic histology only shows cholestasis. Phenobarbitone or ursodeoxycholic acid may shorten and attenuate attacks. A locus for benign recurrent intrahepatic cholestasis has been mapped to chromosome 18, and the abnormality is similar to some cases of progressive familial cholestasis.

Postoperative jaundice

Jaundice due to halothane hepatitis, post-transfusion viral hepatitis, incompatible blood transfusion, drugs, and bile duct damage is described elsewhere.

Prolonged intrahepatic cholestasis used to be common after cardiac surgery, and for no clear reason is much less commonly seen in intensive care units. It is related to the length of surgery and intraoperative cardiac function, and may be due to reduced hepatic blood flow during surgery. Improvement in intra- and postoperative care seems to have improved hepatic function and rendered the syndrome uncommon. Transfused red blood cells are prone to rapid haemolysis and this increases the bilirubin load, while impaired renal function reduces the urinary excretion of conjugated bilirubin. Drug-induced liver injury should be considered.

Cholestasis of pregnancy

Slight impairment of the hepatic excretion of bilirubin can be demonstrated during normal pregnancy or after the administration of oestrogens, but rarely bilirubin and alkaline phosphatase levels rise during the third trimester and intolerable itching and frank jaundice develop, all of which rapidly remit after delivery. The severity of the syndrome increases in successive pregnancies. The fetus is probably not affected, but premature induction of labour may be needed for the mother's sake. The contraceptive pill frequently causes a milder syndrome in the same susceptible women. Ursodeoxycholic acid is reported to ameliorate the condition and is safe, at least during late pregnancy. Phenobarbitone may help the itching, although there may be a small risk of impairing neonatal respiration. Cholestyramine has also been used.

Other causes of jaundice in late pregnancy should be remembered, including acute fatty liver, extrahepatic biliary obstruction, such as from gallstones, and toxaemia.

Pregnancy, by affecting bilirubin excretion, may bring the jaundice of primary biliary cirrhosis or the Dubin–Johnson/Rotor syndromes to notice.

Sepsis

Abnormal liver-related blood tests, and occasionally cholestatic jaundice, often develop during bacterial/viral infections, unrelated to the administration of drugs. In animals this has been shown to be due to endotoxins and cytokines that rapidly down-regulate and translocate the canalicular transport protein mrp2, which excretes conjugated bilirubin into the canaliculus. At the same time other pump proteins are up-regulated, a complex rearrangement that may protect the hepatocyte against oxidative damage. The degradation and impaired synthesis of mrp2 may explain the strange, slow time courses of some of the remitting cholestatic syndromes. Jaundice is especially common in patients with glucose-6-phosphatase deficiency when they develop sepsis, such as pneumonia, since the haemolysis exacerbates the jaundice. The combination of high bilirubin levels and sepsis is particularly damaging to the kidney.

Dubin–Johnson and Rotor syndromes

There are two rare, familial forms of non-haemolytic, conjugated hyperbilirubinaemia without cholestasis.

The Dubin–Johnson syndrome, first described in 1954, is a chronic, relapsing jaundice, without itching or raised serum bile acids. Other liver-related blood tests are normal, but there are associated defects in the excretion of other anions, such as bromsulphthalein, radiographic dyes, and urobilinogen. Hence cholecystography fails, there is excess urobilinogen in the urine, and a delayed rise of the plasma levels of bromsulphthalein after an injection of the dye due to reflux of the conjugated anion from hepatocytes. Jaundice increases during pregnancy or when taking the contraceptive pill because oestrogens impair bilirubin excretion further. A black pigment accumulates in the liver so that at laparoscopy the liver appears strikingly black, as do needle biopsy specimens. Urinary coproporphyrin excretion is abnormal. The inheritance seems to differ between families, and a similar condition occurs in a mutant strain of Corriedale sheep, although in this instance photosensitivity also occurs. Other families have been described in which there are similar findings but no hepatic pigment, the so-called Rotor syndrome. No treatment of either syndrome is required apart from reassurance, and support when seeking life assurance.

Further reading

Chowdury JR, Chowdury NR (1993). Unveiling the mysteries of inherited disorders of bilirubin glucuronidation. *Gastroenterology* **105**, 288–92.

Elferink RPJO, van Berge Henegouwen GP (1998). Cracking the genetic code for benign recurrent and progressive familial intrahepatic cholestasis. *Journal of Hepatology* **29**, 317–20.

Grant A, Neuberger J (1999). Guidelines on the use of liver biopsy in clinical practice. *Gut* **45**(Suppl IV), 1–11.

Jansen PLM (1996). Genetic diseases of bilirubin metabolism: the inherited unconjugated hyperbilirubinemias. *Journal of Hepatology* **25**, 398–404.

Jansen PLM, Müller M (1998). Early events in sepsis-associated cholestasis. *Gastroenterology* **116**, 486–8.

Soloway RD (1996). The increasingly complex molecular life cycle of bilirubin. *Gastroenterology* **110**, 2013–14.

14.20.1 Viral hepatitis—clinical aspects

H. J. F. Hodgson

Many viruses can infect the liver (Table 1). In some patients, hepatic involvement is merely one facet of a systemic infection, and the liver involvement is generally trivial, although occasionally it can be dominant. The liver bears the brunt of infection with the major hepatitis viruses. Thus far, five such viruses have been clearly delineated—A, B, C, D, and E. Other hepatitis viruses are being described and their clinical relevance is under investigation. Viral hepatitis is a major clinical problem worldwide, particularly in developing countries, but no society is exempt.

The clinical effects of infection with a hepatitis virus depend on the severity of the inflammation induced in the liver, and on whether the virus is rapidly cleared from the liver or persists long-term. These in turn reflect the characteristics both of the virus and of the host's immune response. The clinical picture of viral hepatitis is very variable, including fatal fulminant acute hepatitis, acute hepatitis with complete recovery, or chronic infection leading to cirrhosis and predisposing to hepatocellular carcinoma. This chapter will describe the clinical and pathological consequences of viral hepatitis in general, identify virus-specific clinical patterns, and discuss the investigation, management, and prophylaxis of viral hepatitis.

Clinical outcome of hepatitis virus infection

The commonest clinically recognized manifestation of viral hepatitis is an episode of acute icteric hepatitis, generally a self-limited condition with a low mortality and complete recovery. In a typical attack, after an initial prodrome lasting from several days to a couple of weeks—comprising malaise, anorexia, mild fever, and upper abdominal discomfort—jaundice appears. The icteric period typically lasts for a few days to a few weeks, after which the jaundice slowly subsides. Pruritis may occur, generally after the onset of jaundice. Development of ascites or oedema is uncommon but may occur in more severe cases. Return to normality after an attack of hepatitis may take several weeks to a few months and residual fatigue is common.

There are a number of variations on the clinical course of acute hepatitis. Often jaundice does not occur (anicteric hepatitis), and the episode is asymptomatic or dismissed as 'flu-like'. This may occur more frequently than clinically recognized attacks. In cholestatic hepatitis, jaundice with pruritus, pale stools, and dark urine persists for up to 2 or 3 months. Before recovery eventually occurs, a small number of patients have relapsing hepatitis with a transient worsening of jaundice after an initial improvement. Acute hepatitis is only rarely fatal. If it is, patients usually rapidly develop hepatic encephalopathy, and the timing of onset of this has been used to define a variety of syndromes of fulminant hepatitis. One definition is that in 'fulminant hepatitis' encephalopathy develops within 2 weeks of jaundice. Encephalopathy occurs later than this in patients with 'subfulminant hepatitis'. Hepatitis A, B, C, and E can all initiate an acute self-limited hepatitis, although hepatitis C is particularly unlikely to give rise to the fulminant form. Only hepatitis B and C have the propensity to cause chronic viral hepatitis: generally an indolent disease, in which viral carriage in the liver persists over years or decades, with inflammation that varies in intensity. Hepatitis D, which coinfects patients infected with hepatitis B, can contribute to either acute or chronic inflammation.

Features of acute hepatitis caused by different viruses

Hepatitis A virus (HAV)

This causes acute self-limited hepatitis, but not chronic viral carriage or chronic liver disease. The RNA virus is acquired orally. The incubation period is between 2 and 6 weeks. Transmission generally follows the ingestion of food or water contaminated with faeces from an HAV-infected individual. Viral shedding in the faeces ceases at approximately the onset of clinical symptoms. Transmission may occur in epidemics, following floods, or after sewage contamination of shellfish beds. The disease is also endemic in all parts of the world. In developing countries, infection is frequent; there is serological evidence of past infection in up to 100 per cent of 10-year-olds in some countries. In Western countries, evidence of prior infection varies, typically ranging from 5 to 40 per cent dependent on age, social class, and other factors. Promiscuous homosexual males have a high incidence of infection. Very rarely, pooled blood products have transmitted the disease parenterally. Clinically the disease is often anicteric or mild, particularly in young children. About 10 per cent of patients have a relapse before recovery. The mortality rate is low, about 0.3 per cent. Deaths occur predominantly in the elderly amongst whom mortality rates may exceed 2 per cent, and pre-existing chronic viral hepatitis B or C may predispose to a fatal outcome. A rare sequel is aplastic anaemia some months after recovery from hepatitis.

Hepatitis B virus (HBV)

Hepatitis B viral infection was recognized by its parenteral transmission route, classically as serum- and then transfusion-associated hepatitis. The incubation period of this DNA virus varies from 4 to 24 weeks. In between 90 and 95 per cent of adult cases the infection is self-limited and the HBV is cleared. In infants, clearance rates are as low as 5 to 10 per cent. The incidence of acute hepatitis B varies widely, and is very high in the Far East and Africa. Transmission may be vertical—that is, infection of a newborn or infant child usually by a chronically infected mother, either at the time of birth or during close family contact. Horizontal transmission routes include blood transfusion and blood products, the use of contaminated needles medically or by drug addicts, exposure in dialysis units, tattooing, and sexual contact. Promiscuous homosexuals and heterosexuals are at risk.

Anicteric attacks of acute HBV are common. If the acute infection is recognized clinically, in addition to the typical clinical features of any acute hepatitis, the preicteric prodrome may include prominent arthritis, fever, and an urticarial rash, due to immune complex deposition. Hepatitis B is fulminant in about 0.3 per cent of cases, and both the strain of HBV and a very active host immune response may contribute. In the great majority of cases, in which HBV is cleared after acute infection, **HBsAg** (hepatitis B surface antigen) disappears from the blood within weeks to a few months. Failure to clear within 6 months defines 'chronic carriage'.

Hepatitis C virus (HCV)

This was recognized as a cause of transfusion-associated hepatitis. The incubation period ranges from 2 to 26 weeks, but usually between 5 and 12. Apart from blood transfusion and blood product administration, drug addiction and renal dialysis are strong epidemiological associations. Sexual transmission and horizontal transmission are uncommon, but not unknown. Fulminant hepatitis due to HCV is rare. Indeed, the initial acute episode is most often subclinical, but after acquiring the infection 85 per cent of individuals fail to clear the virus. HCV infection is therefore usually not recognized until the chronic phase.

Hepatitis D virus (HDV)

The unique position of hepatitis D virus, an RNA virus 'parasitic' on HBV, has been discussed (Chapter 7.10.19). If HBV and HDV coinfect simultaneously, either unremarkable acute hepatitis, or on occasion fulminant disease, result. If HBV is cleared, HDV must be so also. Acute HDV infection can superinfect a chronic HBV carrier, and result in worsening of liver function, particularly if the hepatitis B virus has previously caused significant liver disease. In such a carrier, the superinfection with HDV may be transient, or chronic hepatitis D carriage may persist. Both coinfection and superinfection are recognized initiators of fulminant hepatitis. In the West, intravenous drug abuse is a prominent epidemiological association, but all parenteral modes demonstrated by HBV occur, including sexual transmission. The southern Mediterranean, the Far East, and South America are areas of high or moderate incidence.

Hepatitis E virus (HEV)

This enterally acquired RNA virus, as is HAV, causes acute hepatitis without chronic carriage. Most major epidemics of acute hepatitis in the Indian subcontinent and the Far East are due to the HEV. Such epidemics affect adults as well as children, indicating that immunity in those areas is not regularly acquired in childhood. Flooding and sewage contamination often precede epidemics. The incubation period is about 6 weeks, and faecal excretion of the virus may persist for nearly 2 months after the onset of hepatitis. A striking feature of HEV infection, and the main clinical difference from HAV, is the propensity to induce fulminant hepatitis if acquired during mid-trimester pregnancy, and mortality rates of 10 to 40 per cent are recorded amongst pregnant women.

Other hepatotrophic viruses

Other hepatotrophic viruses remain to be described, in particular to explain non-A/B/C/D/E fulminant and transfusion hepatitis. Although the hepatitis G virus and transfusion-transmitted virus (**TTV**) have been well characterized, they do not appear to give rise to significant disease.

Clinical examination

Jaundice and right upper-quadrant abdominal tenderness characterize acute hepatitis. Skin manifestations include spider naevi (which often disappear after recovery), scratch marks in the pruritic phase, and rarely a vasculitic or urticarial rash. Mild hepatomegaly is common, but a rapid shrinkage in hepatic size may occur in severe or fulminant hepatitis. Splenomegaly is uncommon, and suggests alternative viral causes such as Epstein–Barr virus or cytomegalovirus, or pre-existing liver disease. Marked nausea and persistent vomiting indicate a severe hepatitis and increase the chance of developing hypoglycaemia. Stools become pale and urine darkens as jaundice is established. Ascites and peripheral oedema may occur in prolonged or severe episodes. The most significant clinical indicator of deterioration is the development of hepatic encephalopathy, indicating the onset of hepatic failure.

Laboratory investigations

Virological investigations depend on serological testing as outlined in Table 2. Figure 1 and Figure 2 indicate the typical serological evolution of self-limited episodes of A and B viral hepatitis.

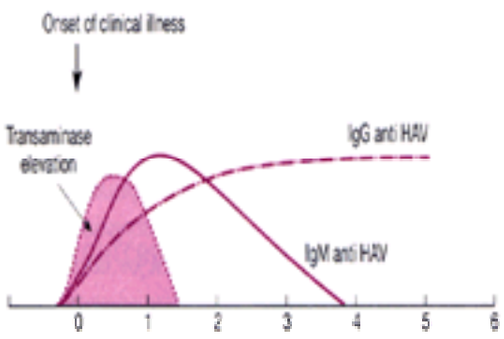

Fig. 1 Typical serology of hepatitis A infection.

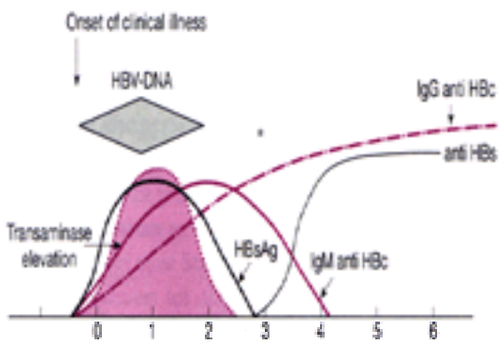

Fig. 2 Serological changes during acute hepatitis B with viral clearance. *, 'Window phase' after elimination of HBsAg and before the emergence of anti-HBs, during which anti-HBc may be the sole indicator of infection.

Typically, hepatocellular enzyme levels in blood (AST, aspartate aminotransferase; ALT alanine aminotransferase) are prominently raised at the time of the onset of symptoms, often more than 10-fold above normal, whilst the serum alkaline phosphatase level is only slightly increased, less than 2.5-fold (Fig. 3). As an episode evolves, transaminase levels fall and alkaline phosphatase may rise, notably if there is prolonged intrahepatic cholestasis. Urinary analysis shows excess urobilinogen in early and late phases of an episode, with excess bilirubin at the height of jaundice. The severity of the attack is best reflected in the synthetic parameters of albumin and clotting factors: in particular, progressive prolongation of the prothrombin time mirrors the onset of liver failure. A low factor V (below about 30 per cent of normal) level has been used as an indicator of irreversible failure.

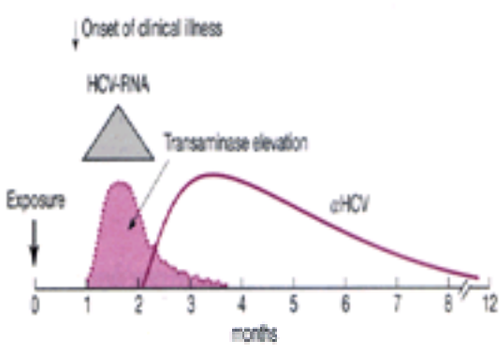

Fig. 3 Serological changes during acute hepatitis C with viral clearance.

Hepatic imaging techniques such as ultrasound contribute to diagnosis primarily by excluding other causes. Patients with uncomplicated hepatitis do not require a liver biopsy, but hepatic histology is very helpful if there is diagnostic uncertainty or an unusual course in severity or duration (Fig. 4 and Plate 1). In such cases, biopsy may require correction of clotting factors and use of the transjugular route or 'plugged' biopsy techniques.

Fig. 4 Haematoxylin and eosin staining of liver biopsies from (a) mild hepatitis with inflammation restricted to the portal tracts and minimal fibrosis, and (b) severely active but still precirrhotic liver with bridging necrosis (by courtesy of Professor P. Dhillon). (See also Plate 1.)

Differential diagnosis

Drug-induced jaundice is the most common differential diagnosis, and its course may be very similar. Drug history and drug screening should particularly enquire about the use of acetaminophen and non-steroidal anti-inflammatory drugs. Other potential drugs and toxins include halothane, antituberculous drugs, carbon tetrachloride, and mushroom poisoning. Alcoholic hepatitis often presents with less marked elevations of serum transaminases and a high circulating leucocyte count. About one-third of patients with autoimmune hepatitis present with a clinical picture of acute hepatitis. Autoantibody testing is generally helpful: the majority of patients with autoimmune disease having high levels of circulating autoantibodies. However, there are often low- or moderate-titre antinuclear and anti-smooth muscle antibodies in uncomplicated viral hepatitis. Similarly, there may be some increase of immunoglobulin levels in acute hepatitis, though not the doubling characteristic of autoimmune hepatitis. Although uncommon, acute Wilson's disease is an important diagnosis to make, because of the high incidence of acute liver failure and the rapid necessity for transplantation. 'Surgical' obstructive jaundice tends not to raise transaminase levels markedly, but serum alkaline phosphatase levels are high. Obstruction is generally confirmed by imaging techniques, notably ultrasound. Individual causes may be suspected from features in the history such as nausea and biliary colic in cholelithiasis, and painless jaundice without systemic upset in an elderly patient with pancreatic cancer. Pregnancy-associated syndromes—acute fatty liver and **HELLP** (haemolysis, elevated liver function tests, low platelets)—are in fact less common in pregnancy than acute hepatitis. Occasionally ischaemia, generally after profound hypotension over many hours, and rapidly progressive malignant infiltration may mimic acute viral hepatitis.

Management

Uncomplicated cases of hepatitis recover spontaneously. Classical studies in military personnel demonstrated no benefit from bed rest, though whether the same applies to the elderly is unknown. In any case malaise and nausea often enforce rest. Clinicians must be alert to signs of impending liver failure and ensure that hypoglycaemia is avoided, if necessary by parenteral administration of glucose. No diets are of established benefit, but dietary fat is often poorly tolerated. There is no rationale for protein restriction unless evidence of hepatic encephalopathy has emerged. Alcohol and potentially hepatotoxic drugs should be withdrawn. Troublesome pruritus can be treated with colestyramine, which is preferable to antihistamines because of potential hepatotoxicity. There is no proven therapy to enhance recovery. Corticosteroids do not speed recovery or improve survival, although they do lower serum bilirubin levels. In hepatitis B, and particularly hepatitis C, the use of interferon has been advocated to enhance the chance of elimination of the virus. There is little evidence of its efficacy with HBV, but some reports in HCV infection are very encouraging.

In fulminant hepatic failure, which in the setting of viral hepatitis carries a mortality risk of 80 per cent, patients should, if possible, undergo orthotopic liver transplantation. Criteria for listing differ in different centres, but include a marked abnormality of clotting parameters (e.g. prolongation of the prothrombin time to >50 s or a factor V level <20–30 per cent) and the development of significant encephalopathy. Patients awaiting transplantation require glucose supplementation, full intensive-care monitoring, and prophylaxis of infection. Some patients require renal support, such as haemofiltration, and ventilation. Some units invasively monitor intracerebral pressure, which may rise dangerously, so that cerebral oedema may be treated with intravenous mannitol.

Prevention of viral hepatitis

Sanitation and hygiene reduce the frequency of the enteric-borne infections HAV and HEV. Passive protection against hepatitis A (to close family contacts) and hepatitis B (after exposure to risk factors such as sexual contact with an individual incubating acute hepatitis B, or a needlestick injury) are available using gammaglobulin preparations (standard preparations for protection against HAV, specific high-titre preparations for HBV). Active immunization to HAV, using formalin-inactivated viral preparations, provides a high level of protective immunity within a few days—suitable, for example, for use prior to travel from the West to highly endemic areas, and also advisable in patients with established chronic liver disease, particularly chronic viral hepatitis. Active immunization to HBV is discussed below, which also protects against HDV. Vaccines are not yet available for HCV or HEV.

Chronic viral hepatitis—hepatitis B

Up to 10 per cent of adults and more than 90 per cent of infants become chronic B carriers after infection, defined by the persistence of HBsAg in the blood for more than 6 months. Subsequently, a low proportion of patients will clear the virus spontaneously each year, but most are infected long-term. Failure to clear the virus is more common in neonates or those infected as infants, in males, and those with natural or iatrogenic immunosuppression. Carriage rates in the population vary widely geographically, and are notably high in the Far East and Southern Africa (10–20 per cent), and low in Northern Europe and North America (<1 per cent).

The consequences of long-term carriage are varied, reflecting the strength of the immune response mounted by the host, the duration of infection, and alteration in the mechanisms of viral replication with time (Fig. 5). Viral mutation may also contribute to modulation of the host response and viral replication. During the early years, the 'replicative' phase, HBeAg (hepatitis B e antigen)-expressing virus replicates independently of the host chromosomes, resulting in the production of fully infectious viral particles in the blood with high levels of HBV-DNA (Table 3). The early replicative phase is associated with a state of relative immune tolerance by the host, and may be very prolonged if infection is acquired as an infant. In the later replicative phase there is expression of immune responses associated with inflammation. Thereafter, HBeAg expression is often lost, HBV-DNA in the blood may be at low or undetectable levels, and HBsAg production may be driven by viral sequences integrated into the genome (integrative phase).

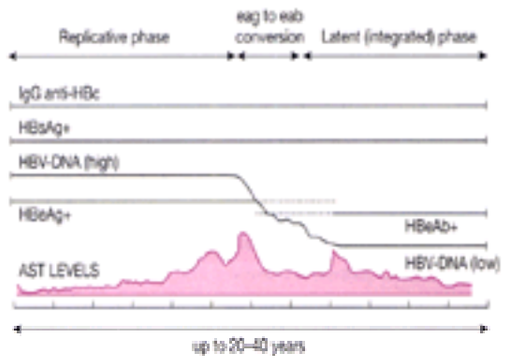

Fig. 5 Serological markers during chronic hepatitis B. Note that HBV mutants not expressing HBeAg may emerge and that HBV-DNA quantitation is required to assess their replicative phase.

Chronic hepatitis B infection is associated with a spectrum of histological damage and clinical manifestations. The inflammatory response to the virus may sometimes be so slight that the histological appearances of the liver are virtually normal, with the exception of evidence of virally infected hepatocytes seen as 'ground-glass' cells on routine eosin staining or by histochemistry. More commonly, the immune response is adequate to inflame the liver but inadequate to clear the virus. The resulting chronic inflammation may be confined to the portal tracts, with a chronic lymphocytic infiltration, associated to varying extent with periportal and/or lobular inflammation, a tendency to develop fibrosis spreading from the portal tracts, and in some cases eventually cirrhosis. These appearances can be categorized in terms of inflammatory activity and fibrosis (Table 4). In general, the replicative phase of HBV infection with HBeAg-positivity, particularly in its later phase, is associated with more marked inflammation than the subsequent HBeAb-positive stage.

Chronic hepatitis B infection may be clinically silent for years, or give rise only to non-specific symptoms of fatigue. The condition may be recognized on screening (for example, during pregnancy) or the investigation of coincidentally detected abnormal liver function tests. Some patients present with non-specific indications of chronic liver disease (malaise or hepatomegaly), or at a late stage with a complication of established cirrhosis. Episodes of enhanced inflammation ('flares') may give rise to transient worsening of liver function tests, particularly transaminase elevations and jaundice at any stage of the disease; precipitating events may include a reduction of prior immunosuppression, or the time of conversion from HBeAg- to HBeAb-positivity. Usually the progression to cirrhosis takes many years, but the rate varies. The incidence of hepatocellular cancer in chronic hepatitis B is high, probably increased 100-fold over non-infected controls. Most, but not all, patients with hepatocellular cancer will have cirrhosis.

Chronic HBV infection may also give rise to a number of extrahepatic manifestations. These include membranous glomerulonephritis, polyarteritis nodosa, and cryoglobulinaemia.

After establishing the diagnosis of chronic HBV infection, it is necessary to define the virological status of the patient with respect to infectivity and viral replication (see Table 3), and the hepatic status with respect to the presence of inflammation and liver damage. Interpretation of viral status may be complicated by the emergence of viral mutants, particularly the 'pre-core' mutant that results in absent HBeAg expression, but which, none the less, is associated with active inflammation and circulating HBV-DNA levels.

Treatment

The general measures relevant to chronic liver disease of any aetiology are discussed elsewhere.

With respect to HBV infection, the prospects for inducing viral clearance in an individual patient are relatively low, but the viral load, infectivity, and intensity of hepatic inflammation can often be reduced. The two current approaches are the use of a-interferons, which have both immunomodulatory and antiviral properties, and the use of inhibitors of viral replication, currently nucleoside analogues.

Patient selection is important. Those with active inflammation, viral replication independent of host DNA, but low levels of HBV-DNA have the greatest potential to benefit. Most patients treated will have circulating HBeAg present. In the absence of elevated transaminases the response to treatment is very poor. Whilst some patients will clear HBsAg from the blood in response to treatment, loss of HbeAg is a more common event.

a-Interferons (**IFN-a**) act predominantly by enhancing T-cell-mediated viral clearance, by processes including the enhancement of hepatocyte class I-HLA expression. Treatment involves parenteral IFN-a (5 MU daily or 10 MU three times weekly, for 4 months). If viral clearance or HBeAg to HBeAb conversion occurs, there is generally an inflammatory flare during the second or third month. Side-effects include malaise, fever (particularly in the first weeks of treatment), anaemia, alopecia, and depression. HBeAg to HBeAb conversion or loss of HBV-DNA occurs in 30 to 40 per cent of cases, HBsAg clearance in about 10 per cent. Women, those with a shorter duration of carriage, occidentals, and those without an additional immunosuppressed background (for example, human immunodeficiency virus (**HIV**) infection) respond more favourably. Relapse after the clearance or sustained loss of HBV replication is rare (5–10 per cent). Successful treatment slows histological progression and reduces liver-related mortality (including hepatocellular cancer).

Nucleoside analogues can inhibit viral reverse transcriptase and inhibit replication. Drugs such as lamivudine are orally bioavailable and now licensed for use. Lamivudine for 12 months markedly reduces HBV-DNA for the duration of treatment, leads to HBeAg to HBeAb conversion in one-third of patients, and reduces inflammatory activity on histological and liver-function test criteria in 50 per cent of cases. However, the loss of HBsAg is infrequent. One difficulty with this drug is that cessation of therapy can occasionally be followed by a disease flare. Moreover, in a significant proportion of cases lamivudine-resistant strains emerge due to mutations in DNA polymerase (*YMDD* mutants). The inflammatory response to this mutant form of the virus may however be diminished. The role of lamivudine and other nucleoside analogues such as famciclovir is under intensive investigation.

Prevention of hepatitis B

Active immunization for the prevention of HBV infection initially involved the use of a vaccine derived from viral proteins in infected blood, but it now uses recombinant HBsAg proteins. Vaccination strategies range from universal vaccination in infancy to the vaccination of only high-risk individuals. In areas of high carriage in the Far East, universal vaccine programmes have already reduced the national incidence of infection, carriage, and hepatocellular cancer. Conventional three-dose immunization in adults leads to protective immunity, as judged by anti-HBsAg, in 90 per cent of individuals.

Passive immunization with anti-HBsAg hyperimmune globulin provides rapid protection after exposure (e.g. after needlestick injury). A combination of passive and active immunization is recommended for children born to infected mothers. In some infants, chronic infection with a mutant 'escape' virus has subsequently occurred.

Chronic hepatitis C

Around 85 per cent of patients who become infected with HCV fail to clear the virus and become chronic carriers. In the majority of cases the initial presentation will have been asymptomatic, and HCV infection is generally recognized in the chronic phase. Following the availability of tests to diagnose HCV infection, it is clear that an asymptomatic indolent necroinflammatory response to the virus in the liver may persist long term, and often, but not inevitably, lead to cirrhosis after 15 to 25 years and predispose to hepatocellular cancer thereafter.

Mechanisms of transmission and prevalence rates vary geographically. In the West, blood transfusion and treatment of clotting disorders with plasma concentrates prior to the early 1990s, and intravenous drug abuse constitute the main routes of transmission. Medical use of unsterilized needles, including in vaccination programmes, tattooing, dentistry, and communal shaving practices, may all contribute worldwide. Vertical transmission is rare. Sexual transmission is low (<5 per cent in stable heterosexual relationships).

As with HBV, the severity of liver damage reflects host–virus responses. Severe inflammation is less common if the virus is acquired in childhood and progression to

cirrhosis less frequent and probably slower. HBV coinfection and alcoholism worsen disease and increase the likelihood and rate of developing cirrhosis. The histological response of the liver shows a similar variety of response to that seen in HBV infection, from minimal to severe portal inflammation, periportal hepatocyte necrosis, and progressive fibrosis leading to cirrhosis. The presence of lymphoid follicles in portal tracts and parenchymal steatosis are characteristic of the response to HCV.

Patients may be diagnosed coincidentally during the investigation of fatigue or abnormal liver function tests, or with manifestations of chronic liver disease. In addition, there are a variety of extrahepatic manifestations thought to reflect either antigen–antibody complex formation or the induction of crossreacting autoimmunity. These include a vasculitic rash associated with cryoglobulinaemia (type 2, polyclonal immunoglobulin plus rheumatoid factor), glomerulonephritis, abnormal thyroid function, thrombocytopenia, and porphyria cutanea tarda. As in HBV, assessment of a patient with HCV involves both the virological and the hepatic status.

The initial screening test for HCV is detection of circulating anti-HCV antibody. If present, confirmation is required either by more specific antibody testing using immunoblots, or more often by using **PCR** (polymerase chain reaction) for viral RNA (Fig. 6). Quantitative PCR and viral genotyping (types 1–4) are becoming more relevant as treatment strategies evolve (see below). Changing sensitivities of antibody tests, and varying techniques for PCR testing, render specialist interpretation mandatory.

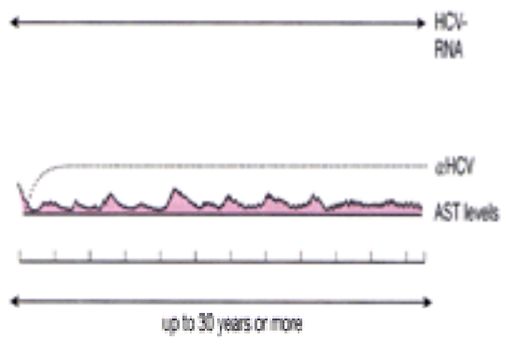

Fig. 6 Serological changes during chronic hepatitis C.

The severity of inflammation in the liver is poorly judged from routine liver function tests such as aminotransferase level measurements. Histological assessment may reveal both significant inflammation and progressive fibrosis despite normal serum enzyme levels, and are necessary if treatment is to be considered. There is consensus that moderate or severe precirrhotic disease as judged by activity grading and well-compensated cirrhosis should be considered for treatment. The position in minimal or mild disease is uncertain, in part because patients with mild disease histologically, and normal aminotransferase levels, do not respond well to current treatment.

Treatment of chronic HCV

The main aim of treatment in chronic HCV infection is to clear the virus, which has been established to be associated with a reduction in necroinflammation and slowing in the rate of accumulation of fibrosis in the liver.

Initial studies using IFN-a monotherapy demonstrated that between 25 and 50 per cent of patients responded temporarily, with normalization of aminotransferase levels and loss of PCR-positivity for viral RNA recorded at the end of 6 months' therapy. However when assessed 6 to 12 months later this response was sustained in far fewer patients—overall sustained virological response rates were only between 10 and 12 per cent. Rates were higher with more prolonged treatment, but did not exceed 20 per cent. Clearance rates were lower with genotype 1 infection and when initial aminotransferase levels were normal.

Subsequently, the combination of interferon with ribavirin has been shown strikingly to enhance sustained response rates, with an approximate doubling of clearance rates after both 24 and 48 weeks. For genotypes 2 and 3 however, 6 months' treatment appears adequate. Currently, for these genotypes, the combination of 3 MU of IFN-alpha three times weekly plus ribavirin 1000 to 1200 mg orally is recommended. For genotype 1, treatment for 12 months is recommended. With this regime, in addition to the side-effects of IFN-a listed discussed above for HBV treatment, ribavirin induces cough, rash dyspnoea, and insomnia in about 25 per cent of patients, and there is predictably a dose-dependent haemolytic anaemia. Both pregnancy and fathering children need to be avoided whilst taking ribavirin. Strategies for the early identification of non-responders, by assessing whether there is a reduction in viral load in the blood, are evolving. The full role of therapy in established cirrhosis, and the long-term benefits of reducing inflammation with these therapies in patients in whom viral clearance has not been achieved, are currently under investigation. Improved interferon formulations (such as interferon conjugated to polyethylene glycol, PEG) and other antiviral drugs are being introduced.

Prevention of hepatitis C

There is no vaccine available for HCV. Passive immunization with HCV-antibody-containing gammaglobulin in not protective.

Hepatitis D

Chronic HDV generally follows superinfection of a chronic HBV carrier in whom ample HbsAg to permit HDV encapsulation is already present. The spectrum of chronic liver disease associated with the double chronic infection is as variable as with HBV alone. Overall, however, the liver tends to be more severely affected, and in 10 to 15 per cent of chronic carriers there may be a rapid (1 to 2 year) evolution to cirrhosis. HDV acts to suppress HBV infection, so that markers of HBV activity, such as HBV-DNA, in the serum may become suppressed. Many patients with HDV may therefore be HBeAb positive.

The treatment of HDV mirrors that for HBV. Prolonged courses of IFN-a (6 or more months) may transiently clear HDV in some patients in whom HbsAg persists. In general, HbsAg clearance is required to cause sustained HDV clearance.

Liver transplantation

Liver transplantation is indicated both in fulminant hepatic failure due to acute hepatitis and in advanced chronic hepatitis with cirrhosis. Recurrence of viral hepatitis after transplantation is a major concern. The use of hyperimmune globulin, interferon, and nucleoside analogues allows control in most cases of HBV. Severe recurrence remains a significant problem after transplantation for HCV.

Further reading

Alter MJ (1997). Epidemiology of hepatitis C. *Hepatology* **26**(Suppl 1), 62S–65S.

Bell BP, *et al.* (1998). The diverse patterns of hepatitis A epidemiology in the United States—implications for vaccination strategies. *Journal of Infectious Diseases* **178**, 1579–84.

Chang MH, *et al.* (1997). Shau Universal hepatitis B vaccination in Taiwan and the incidence of hepatocellular carcinoma in children. *New England Journal of Medicine* **336**, 1855–9.

Farci P, *et al*. (1994). Treatment of chronic hepatitis D with interferon alfa-2a. *New England Journal of Medicine* **330**, 88–94.

Feitelson MA (1994). Biology of hepatitis B virus variants. *Laboratory Investigation* **71**, 324–49.

Guillevin L, *et al.* (1995). Polyarteritis nodosa related to hepatitis B virus. A prospective study with long-term observation of 41 patients. *Medicine* **74**, 238–53.

Hoofnagle JH (1997). Hepatitis C: the clinical spectrum of disease. *Hepatology* **26**(Suppl 1), 15S–20S.

Irshad M (1997). Hepatitis E virus: a global view of its seroepidemiology and transmission pattern. *Tropical Gastroenterology* **18**, 45–9.

Lai CL, *et al.* (1998). A one-year trial of lamivudine for chronic hepatitis B. *New England Journal of Medicine* **339**, 61–8.

Lee WM (1997). Medical progress: hepatitis B virus infection. *New England Journal of Medicine* **337**, 1733–45.

Lemon SM, Thomas DL (1997). Vaccines to prevent viral hepatitis. *New England Journal of Medicine* **336**, 196–204.

Poynard T, *et al.* (1998). Randomised trial of interferon alpha2b plus ribavirin for 48 weeks or for 24 weeks versus interferon alpha2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus. *Lancet* **352**, 1426–32.

Scheuer PJ, Davies SE, Dhillon AP (1996). Histopathological aspects of viral hepatitis. *Journal of Viral hepatitis* **3**, 277–83.

Sjogren MH (1996). Serologic diagnosis of viral hepatitis. *Medical Clinics of North America* **80**, 929–56.

Thursz MR, Thomas HC (1997). Host factors in chronic viral hepatitis. *Seminars in Liver Disease* **17**, 345–50.

Vento S, *et al.* (1998). Fulminant hepatitis associated with hepatitis A virus superinfection in patients with chronic hepatitis C. *New England Journal of Medicine* **338**, 286–90.

14.20.2.1 Autoimmune hepatitis

H. J. F. Hodgson

Autoimmune hepatitis describes chronic inflammation in the liver attributed to immune responses against self-antigens in the liver. Patients generally have circulating autoantibodies, and 60 per cent have other autoimmune diseases in addition. In severe cases autoimmune hepatitis can lead to acute liver failure, and untreated there is often progression to cirrhosis. There is generally a good response to corticosteroid therapy.

The term 'autoimmune hepatitis' should be reserved for patients with clinically significant liver disease, and not used to describe the very mild chronic inflammation often seen in the livers of patients with systemic autoimmune conditions. Previous terms for autoimmune hepatitis include 'autoimmune chronic active hepatitis' and 'lupoid hepatitis'.

Aetiology

Autoimmune hepatitis is often familial, and a constellation of autoimmune diseases (for example, thyroiditis, type-1 diabetes) may occur in affected families. Some of the predisposing genetic factors have been characterized. There are strong HLA associations. In the United Kingdom and United States the strongest association is with HLA DR haplotype B1*030 (50 per cent of patients compared with 20 per cent of controls), and a secondary association with *0401, but in other geographical areas there are different associations. The DR4 association is strongest in Japan. Deficiency of the C4 component of complement also predisposes to the disease. In the West, DR3 is associated with more severe disease and a younger onset, and DR4 with an older onset and better treatment response. In most cases there is no clear initiating event for the development of autoimmune hepatitis, but occasionally drugs (a-methyldopa, oxyphenisatine, nitrofurantoin, isoniazid, minocycline, pemoline, dihydralazine, tienilic acid) can precipitate the condition. As discussed below, in some patients with chronic hepatitis C, autoimmune manifestations develop and may contribute to the inflammatory processes in that chronic viral condition.

Histological appearances and immunopathogenesis

The histological appearances used to be referred to as 'chronic active hepatitis', but pathologists now grade chronic hepatitis in respect of aetiology, disease activity (grading), and progression (staging). There is a marked portal tract infiltrate containing both plasma cells and T cells. Cytotoxic T cells spread out across the limiting plate of the portal tract, associated with piecemeal necrosis of periportal hepatocytes, and lymphocytes are often present diffusely within the parenchyma (Fig. 1). Compared with the inflammation seen in chronic hepatitis B and C, the plasma-cell component of the portal tract infiltrate is more prominent, as is regenerative 'rosette' formation by the hepatocytes. The portal tract lymphoid aggregates and steatosis common in hepatitis C are less frequent in autoimmune hepatitis. With progression, periportal fibrosis, bridging necrosis linking portal tracts to central veins, and hyperplasia of regenerating hepatocytes all occur leading to cirrhosis.

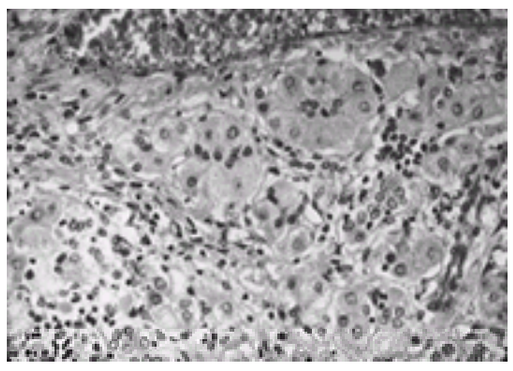

Fig. 1 Haematoxylin and eosin stained liver histology showing 'rosettes' of regenerated hepatocytes, surrounded by lymphocytes that have spread into the hepatic parenchyma.

Antibody and T-cell reactivity to a panel of hepatic antigens, both intracellular and on the cell surface, are described. Some of these may be the primary mechanism of liver damage, and others a secondary response to tissue injury. Cell-surface targets such as the hepatocyte-specific asialoglycoprotein receptor may be particularly relevant to the development of tissue damage.

Epidemiology

Most reported cases are of European ancestry, although cases occur worldwide. Most patients are female—with female:male ratios of up to 8:1 reported in some series. Northern European prevalence figures are about 100 to 200 per 100 000, similar to those for primary biliary cirrhosis and primary sclerosing cholangitis. The age of onset varies widely, but young adults and children are most commonly affected. Perhaps 20 per cent of cases occur after the age of 65.

Clinical manifestations

In about 30 per cent of cases the onset is indistinguishable clinically from acute viral hepatitis, with anorexia, nausea, hepatic discomfort, and the development of jaundice. Other patients present just with malaise, or with extrahepatic manifestations such as arthralgia, arthritis, or fever, and subsequent investigation demonstrates high circulating transaminase levels. Despite an acute onset, many patients have cirrhosis at the time of presentation.

Clinical signs may therefore vary greatly. Cutaneous manifestations of palmar erythema and prominent spider naevi, a maculopapular or acneiform rash, and occasionally abdominal striae (without corticosteroid therapy) may be found both in acute and insidious presentations. In an acute presentation, jaundice and tender hepatomegaly may be prominent, with severe cases demonstrating ascites and most sinisterly the development of encephalopathy. Acute severe autoimmune hepatitis may progress to fulminant hepatic failure. Splenomegaly, which is often marked, indicates that cirrhosis has already developed. There may also be evidence of other autoimmune conditions (Table 1).

Investigations

High transaminase levels, marked hyperglobulinaemia (particularly IgG), and circulating autoantibodies characterize the condition. Transaminase levels are often five or ten times the upper normal limit. Hyperglobulinaemia may include IgG levels of more than 30 g/l when the condition is active. The major serological diagnostic markers are the presence of autoantibodies demonstrated by tissue immunofluorescence, prominently antinuclear antibodies (**ANA**), titre 1:40 or greater, and an anti-smooth muscle (**SMA**) titre 1:80 or greater. In addition, other autoantibodies may be present, in particular to liver–kidney microsomes (liver and proximal renal tubule, LKM-1). However, the antibody profiles in the disease are variable and may alter over time: some cases identified on clinical and histological grounds may lack circulating antibodies on routine testing. Other autoimmune associations of the disease may be manifest with antithyroid, parietal cell, and intrinsic factor antibodies, or antibodies leading to immune thrombocytopenia or haemolysis. Other positive immune tests include a moderate frequency of antibodies to double-stranded and single-stranded DNA. The positive lupus erythematosus cells described in early reports of the condition led to the designation of 'lupoid hepatitis'.

Diagnostic criteria

There are no critical signs, symptoms, or liver test abnormalities that are sufficiently specific to provide diagnostic criteria. The diagnosis is therefore made on a constellation of features, each of which may also occur in other conditions. Diagnosis during the active phase of the disease generally reflects the presence of elevated serum transaminase levels without marked elevation of alkaline phosphatase, elevated globulins, positive antibodies to ANA, SMA, or LKM-1, seronegativity for hepatitis viruses (although the overlap with hepatitis C is discussed below), and characteristic or compatible liver histology. A 'scoring system' to identify definite or probable autoimmune hepatitis from a combination of clinical and histological features has been proposed.

Differential diagnosis

The differential diagnosis during investigation includes not only viral and drug-induced hepatitis, but the metabolic conditions of Wilson's disease and a1-antitrypsin deficiency. Some alcoholic patients may manifest histological appearances overlapping with autoimmune hepatitis. The immune biliary diseases primary biliary cirrhosis (**PBC**) and primary sclerosing cholangitis (**PSC**) are generally easy to differentiate by the predominant elevation of serum alkaline phosphatase and the antimitochondrial antibody in PBC, and antineutrophil cytoplasmic antibodies (**ANCA**) and cholangiography in PSC. However, as already mentioned, each of the suggestive indicators for autoimmune hepatitis may be absent in particular cases, and in some series up to 20 per cent of cases lack any ANA, SMA, and anti-LKM-1 antibodies. Furthermore, there is overlap in the autoimmune profiles of various immunological liver diseases, with, for example, low titres of ANCA often found in autoimmune hepatitis, and SMA in both PBS and PSC. Autoimmune hepatitis is not part of the spectrum of systemic lupus erythematosus.

Autoimmune hepatitis subtypes

A subclassification of autoimmune hepatitis into four groups according to a combination of autoantibody profiles and clinical features has also been suggested (Table 2):

- **Type I**—Classical autoimmune hepatitis with high-titre ANA and SMA. Typically, patients are young adult females. The SMA antibody has anti-F-actin specificity, as can be demonstrated using cell lines as immunofluorescent substrates.
- **Type II**—Anti-LKM-1 antibodies are present, with specificity for a cytochrome P-450 antigen, CYP4502D. The disease may be particularly severe, and tends to affect children. However anti-LKM-1 may also be found in sera from patients with chronic hepatitis C infection, which can itself also predispose to a variety of autoimmune manifestations. It has therefore been suggested that Type-II autoimmune hepatitis should be further subdivided to separate those with a primary autoimmune condition (IIa) from those in whom hepatitis C has triggered an autoimmune response (IIb). This, however, introduces further confusion as some patients with hepatitis C have ANA and SMA but not anti-LKM.
- **Type III**—Antibodies to a soluble liver antigen (**SLA**), often without ANA, SMA, or LKM antibodies.
- **Type IV**—No antibodies detectable.
- The value of this classification has been questioned, and the issue is complex. Many other autoantibodies are being described in these patients, to both intracellular and surface antigens of liver cells (for example: antibodies to the hepatocyte-specific asialoglycoprotein receptor, to other hepatic membrane antigens, and to cytosolic enzymes; and to a 'liver–pancreas' antigen). Furthermore, antibodies identified by immunofluorescence are often relatively non-specific, and reactions to different epitopes may give similar staining patterns. Many of the newly described antibodies are found in more than one type of autoimmune hepatitis and in other immune conditions affecting the liver. Also, as described below, there are well-described 'overlap' cases, with patients with autoimmune hepatitis also showing manifestations of another immune liver disease. Subtyping on autoantibody criteria alone is therefore unlikely to be definitive. The delineation of Type IV, with no detectable autoantibodies, serves to emphasize that if other clinical and laboratory features are present, a trial of corticosteroids may be worthwhile even if autoantibodies have not been identified.

Associated liver diseases and overlap conditions

Up to 10 per cent of cases may show mixed autoimmune hepatitis and immune biliary disease, generally PBC and less frequently PSC. The disease may also coexist with autoimmune cholangitis (which resembles PBC but is antimitochondrial antibody-negative).

Natural history

Untreated, autoimmune hepatitis may occasionally spontaneously remit, but there is a marked tendency for progression with the development of hepatic fibrosis and cirrhosis, and subsequently the complications thereof. Childhood-onset and Type-II autoimmune hepatitis have a bad prognosis. Over half of the patients with severe disease die within 5 years if untreated. Epidemiological features suggest that inapparent autoimmune hepatitis is the cause of a significant number of cases of cryptogenic cirrhosis. Cirrhosis arising as a consequence of autoimmune hepatitis only rarely leads to hepatocellular carcinoma.

Treatment regimes

Specific therapy for autoimmune hepatitis is aimed at reducing or abolishing inflammation. Patients, particularly those with cirrhosis, may require treatment for the complications of portal hypertension and liver failure discussed in Chapter 14.21.2.

There is consensus that severe cases of autoimmune hepatitis should be treated with an immunosuppressive regime for 1 to 2 years. Evidence of benefit is firm in patients with transaminase levels more than five times normal, and those with histological evidence of bridging necrosis on biopsy. Whether patients with mild disease (minor elevations of transaminase, minor inflammation on biopsy without developing fibrosis) benefit from immunosuppressive therapy is unclear, and in such patients a period of observation followed by repeat biopsy to gauge progression may be helpful. The finding of established cirrhosis should not be taken as a reason to prevent treatment, provided there is active inflammation.

Corticosteroid treatment of patients with severe disease reduces inflammation in 80 to 90 per cent of cases, reduces the chance of progression to cirrhosis if it has not already occurred, and prolongs survival. In assessing the short-term effects of corticosteroid or other therapy, responses are characterized as 'complete' if transaminase levels normalize and remain so for a year or more, or if repeat histology shows only minimal activity. A 'partial response' describes improvement but the persistence of transaminase abnormalities at more than twice the upper limit of normal, or persistent histological activity despite the normalization of transaminase levels. In the most successful cases, complete responses with regression of fibrosis and normalization of architecture have been reported.

A variety of corticosteroid-based regimes are in use. Comparative studies demonstrate equivalent anti-inflammatory efficacy of a solely corticosteroid regime of prednisolone 20 mg daily, and a steroid-sparing combination of 75 mg azathioprine plus 10 mg prednisolone, but a lesser incidence of side-effects over 2 years with the latter. Many physicians use higher doses of prednisolone initially, 30 to 45 mg, until there is a definite improvement in liver function tests. Amongst those who respond, corticosteroid treatment will significantly reduce or normalize transaminase levels within a few weeks to a few months. Symptoms tend to resolve over a similar period, but corticosteroids may not abolish inflammation, as judged histologically, for a year or more. Therefore, if a complete clinical and biochemical remission is established, generally within a few months, most physicians advocate treatment for between 1 and 2 years and a repeat liver biopsy after that to see if there is still histological evidence of inflammation. A biopsy at this interval may also demonstrate that cirrhosis has supervened, despite clinical and biochemical control of the disease. If there is no active inflammation, or only mild inflammation, on the follow-up biopsy, cautious withdrawal of corticosteroids—for example, reducing the dose at a rate of 1 mg per month—may allow their discontinuation. However, the chances of relapse are high (perhaps 70 per cent) either during withdrawal or later, in which case immunosuppressive therapy will again be required and should then be continued long-term. When this is required, azathioprine

alone may be all that is required. For patients who have only partially responded to corticosteroids—judged by the failure to reduce transaminase levels significantly or by persisting inflammation on follow-up biopsy—long-term immunosuppression is advocated, and increased corticosteroid or azathioprine dosage may be required to achieve remission.

Minimizing the incidence and severity of corticosteroid side-effects is very important. Glucose intolerance and hypertension should be screened for, and the enhanced risk of infection warrants increased vigilance. Strategies to reduce bone loss include oral supplementation with calcium and vitamin D, or the use of bisphosphonates to restore lost bone mass. The orally administered halogenated corticosteroid budesonide is rapidly metabolized by the liver, and reports indicate that it may be effective in treating autoimmune liver disease with less effect on the pituitary–adrenal axis than prednisolone. Theoretically, lesser effects on bone would also be anticipated.

For patients who fail to respond to corticosteroid-based immunosuppression there is anecdotal evidence of a reasonable response rate to ciclosporin. A number of alternative immunosuppressive agents—methotrexate, mycophenylate, and tacrolimus—have also been anecdotally reported to be useful. Failure to respond should also prompt reconsideration of the diagnosis, in particular exclusion of previously unrecognized Wilson's disease.

In the overlap syndrome of primary biliary cirrhosis-autoimmune hepatitis, the periportal inflammatory element of the condition and the serum transaminase elevations generally respond to corticosteroid therapy, but the biliary component is not improved. Conversely, bile acid therapy with ursodeoxycholic acid can improve the biliary manifestations whilst leaving the transaminase levels and periportal inflammation unaffected. Combination of the two therapeutic approaches may be necessary to restore normal biochemical markers of liver disease. Patients with the autoimmune hepatitis-PSC overlap syndrome respond poorly to treatment.

Prognosis after treatment

Amongst patients presenting without cirrhosis, long-term survival is excellent (over 95 per cent at 10 years), whilst the 10-year survival rate is about 65 per cent if cirrhosis is present initially. Corticosteroid therapy increases the survival of both cirrhotic and non-cirrhotic patients.

Transplantation

Endstage cirrhosis due to autoimmune hepatitis, and acute non-responsive autoimmune hepatitis leading to acute or subacute liver failure, provide firm indications for orthotopic liver transplantation. Failure to achieve an early response in acute disease, with a shrinking liver volume, should prompt consideration of transplantation. Overall, the prognosis after transplantation is good, with 5-year survival rates in excess of 80 per cent, and a similar incidence of acute rejection episodes (50–60 per cent) to that seen in other immunological liver diseases. Autoantibodies persist after transplantation, though at lower titre. There is a tendency for the disease to recur in the transplanted liver, with a frequency of about 20 per cent at 5 years, sometimes necessitating retransplantation. Whether more aggressive immunosuppressive antirejection regimes will prevent this remains to be established.

Further reading

Alvarez F *et al.* (1999). International Autoimmune Hepatitis Group Report: review of criteria for diagnosis of autoimmune hepatitis. *Journal of Hepatology* **31**, 929–38.

Bansi D, Chapman R, Fleming K (1996). Antineutrophil cytoplasmic antibodies in chronic liver disease: prevalence, titre, specificity and IgG subclass. *Journal of Hepatology* **24**, 581–6.

Cassani F, *et al.* (1997). Serum autoantibodies in chronic hepatitis C: comparison with autoimmune hepatitis and impact on the disease profile. *Hepatology* **26**, 561–6.

Chazouilleres O, *et al.* (1998). Primary biliary cirrhosis-autoimmune hepatitis overlap syndrome: clinical features and response to therapy. *Hepatology* **28**, 296–301.

Czaja AJ (1998). Frequency and nature of the variant syndromes of autoimmune liver disease. *Hepatology* **28**, 360–5.

Czaja AJ, Carpenter HA (1993). Sensitivity, specificity, and predictability of biopsy interpretations in chronic hepatitis. *Gastroenterology* **105**, 1824–32.

Dufour JF, DeLellis R, Kaplan MM (1997). Reversibility of hepatic fibrosis in autoimmune hepatitis. *Annals of Internal Medicine* **127**, 981–5.

Fernandes NF, *et al.* (1999). Cyclosporine therapy in patients with steroid resistant autoimmune hepatitis. *American Journal of Gastroenterology* **94**, 241–8.

Johnson PJ, McFarlane IG, Williams R (1995). Azathioprine for long term maintenance of remission in autoimmune hepatitis. *New England Journal of Medicine* **333**, 958–63.

Meyer zum Buschenfelde KH, Dienes HP (1996). Autoimmune hepatitis. *Virchows Archiv* **429**, 1–12.

Newton JL, *et al.* (1997). Autoimmune hepatitis in older patients. *Age and Ageing* **26**, 441–4.

Ratziu V, *et al.* (1999). Long term follow up after liver transplantation for autoimmune hepatitis, evidence of recurrence of primary disease. *Journal of Hepatology* **30**, 131–41.

Strettell MD, *et al.* (1997). Allelic basic for HLA encoded susceptibility to type 1 autoimmune hepatitis. *Gastroenterology* **112**, 2028–35.

14.20.2.2Primary biliary cirrhosis

M. F. Bassendine

Primary biliary cirrhosis is a chronic, cholestatic liver disease in which the biliary epithelial cells lining the small intrahepatic bile ducts are the target for immune-mediated damage leading to progressive ductopenia. The cause is unknown but evidence points to an autoimmune aetiology, in particular the strong association with disease-specific autoantibodies. It affects women in over 90 per cent of cases and usually has an insidious onset in middle age. Patients with early disease are often recognized following the incidental discovery of antimitochondrial antibodies or elevated levels of serum alkaline phosphatase. Progression may be slow but eventually many patients develop cirrhosis and, ultimately, death may occur from liver failure or complications of cirrhosis such as bleeding oesophageal varices. It is a common indication for liver transplantation.

Epidemiology

There is a marked geographical variation in the prevalence of the disease; it is commonest in Northern Europe but rare in the Indian subcontinent and Africa. It was rare and accounted for fewer than 5 per cent of patients dying of cirrhosis in Western communities, but its incidence appears to be increasing. In the North-East of England the prevalence rose from 202 per million adults and 541 per million women over 40 in 1987 to 335 per million adults and 940 per million women over 40 in 1994. It remains unclear whether this increase represents better diagnosis or a true increase in prevalence. Death rates from all causes are nearly three times greater in patients with biliary cirrhosis compared with the general population after adjusting for age and sex.

Aetiology and pathogenesis

In common with most autoimmune disorders, genetic factors partially determine susceptibility to primary biliary cirrhosis, but the pattern of inheritance is very complex. Familial clustering is well documented, and the sibling relative risk is 10.5, similar to values seen in other autoimmune disorders where it is thought that genetic factors may contribute up to 50 per cent of the total risk. There is no association of the disorder with major histocompatibility complex (**MHC**) class I antigens but several associations with class II antigens have been reported, in particular HLA DR8 with a two- to sixfold increase in patients compared with controls. Information on MHC class III associations is conflicting. The genetic predisposition conferred by HLA is however neither sufficient nor necessary for disease development and other genes are likely to be involved. The gene encoding cytotoxic T lymphocyte-associated antigen-4 has recently been examined as a candidate gene and this locus is important for conferring susceptibility not only to primary biliary cirrhosis but also to autoimmunity in general.

Over 95 per cent of patients have antibodies to mitochondria, with the dominant autoantibody response being directed against two components (dihydrolipoamide acetyltransferase (E2) and E3-binding protein) of pyruvate dehydrogenase complex (**PDC**) (Table 1). The loss of tolerance to these autoantigens is an early event in this progressive disease with antimitochondrial antibodies (**AMA**) being detectable in serum before abnormalities in liver function and long before the onset of symptoms. One hypothesis is that the development of these AMA marks the exposure of a genetically susceptible individual to an initiating environmental factor. Autoreactive T cells play a central role in the development of various autoimmune diseases and an immunodominant T-cell epitope within pyruvate dehydrogenase complex E2 (peptide 163 to 176) has been identified in patients with primary biliary cirrhosis. T-cell clones reactive to this peptide can also be activated by mimicry peptides derived from several microbial proteins, supporting the hypothesis that autoreactive T cells present in the peripheral blood can be activated and clonally expanded by antigenic stimulation by mimicry peptides derived from environmental non-self antigen.

Aberrant expression of pyruvate dehydrogenase complex occurs on the apical surface of biliary and salivary epithelial cells in patients with primary biliary cirrhosis and secretory IgA autoantibodies have been found in bile and saliva. IgA autoantibodies to pyruvate dehydrogenase complex may thus interact with components of pyruvate dehydrogenase complex within the epithelial cells leading to metabolic consequences and subsequent cell damage.

Antinuclear antibodies occur in a minority of patients with primary biliary cirrhosis (Table 1) and display unique immunofluorescence patterns such as nuclear dots or a nuclear ring-like pattern. Disease-specific nuclear antigens include a 210-kDa glycoprotein of the nuclear pore membrane (gp 210), nucleoporin p62, and Sp100, an interferon-inducible nucleoprotein with a molecular mass of 100 kDa.

Despite progress in characterizing the reactivity of the disease-specific autoantibodies, little is understood of the way in which the autoimmune response is induced or the effector mechanisms that cause tissue damage. The concept of the T_{H1}/T_{H2} paradigm has gained importance in autoimmune reactions; in primary biliary cirrhosis there is dominance of T_{H1} cells. Interferon-g is the main cytokine in the liver and CD8+ cytotoxic T cells infiltrate the portal tract as part of the chronic inflammation that characterizes primary biliary cirrhosis. The cytotoxic lymphocyte must interact closely with its putative target in order to recognize peptide in association with MHC class 1 and this is facilitated by adhesion to intercellular adhesion molecule 1 (ICAM-1) and CD58, both of which are increased on inflamed biliary cells. Biliary epithelial cells undergo apoptosis but the mechanisms involved and the role of the other effector systems are less clear.

Clinical features

Patients with early disease may complain only of fatigue or symptoms of coexisting autoimmune disease. Those with more advanced disease have evidence of cholestasis, with jaundice, pruritus, light stools, easy bruising, and weight loss. The pruritus may first be noticed during pregnancy or when the patient is on the contraceptive pill. Occasionally, patients present with gastrointestinal bleeding from oesophageal varices or associated peptic ulcer.

Findings on examination vary widely. At one extreme, there may be no abnormality, whereas at the other the patient is jaundiced, with scratch marks and signs of long-standing cholestasis. The planus form of xanthoma occurs characteristically as xanthelasmas around the eyes and in the palmar creases. Tuberous lesions develop late on the extensor surfaces around the knees, elbows, wrists, ankle, and on pressure points such as buttocks. Occasionally they affect tendon sheaths and nerves, producing xanthomatous peripheral neuropathy.

The liver is often enlarged and firm, and splenomegaly may be present, with or without portal hypertension. Spider naevi and palmar erythema are less frequent than in patients with alcoholic cirrhosis. Fluid retention with ascites and oedema is usually a late complication, as is bleeding from oesophageal varices. Steatorrhoea occurs primarily in patients who have advanced cholestasis, leading to malabsorption of fat-soluble vitamins, especially vitamin D. Bone pain due to osteomalacia, with tenderness and fractures involving vertebrae, can occur, as can liver failure with encephalopathy. Such late manifestations of disease are now rarely seen in Western countries as liver transplant is performed in most patients before their development. Osteoporosis is also well recognized but may largely reflect the gender and age of patients with primary biliary cirrhosis. Deficiency of vitamin K sometimes results in easy bruising or other haemorrhagic phenomena. Clubbing of the fingers and leuconychia are rare findings. There is an increased incidence of gallstones and peptic ulceration, and features of these conditions may form part of the clinical picture.

Primary biliary cirrhosis is associated with past smoking and a number of other autoimmune diseases. These include Sjögren's syndrome, seropositive and seronegative arthropathy, thyroiditis, scleroderma, and renal tubular acidosis. The CREST syndrome (calcinosis, Raynaud's phenomenon, sclerodactyly, and telangiectasia), pulmonary fibrosis, psoriasis, and coeliac disease have also been reported.

Pathology

The characteristic early lesion of primary biliary cirrhosis is inflammatory duct destruction. Later there is fibrosis, often patchy, and eventually a frank cirrhotic picture. Histologically this disease appears to evolve from a florid duct lesion to cirrhosis. This has led to a morphological classification into four stages. It must be recognized, however, that overlap between stages is common in different parts of the liver. In stage 1, the duct lesion is florid (Fig. 1) with the epithelium irregular, hyperplastic, or ulcerated. There is a heavy infiltrate of lymphocytes, plasma cells, and neutrophils, with occasional eosinophils. Aggregates of histiocytes with granulomas ranging from foci of epithelioid cells to rounded lesions with multinucleated giant cells are present. In stage 2 there is established duct destruction and the bile ducts may be replaced by lymphoid aggregates with fibrosis. In stage 3 there is relatively little inflammation, though lymphoid aggregates may be present and fibrous septa extend from the portal tract. In stage 4 there is an established cirrhosis, paucity of bile ducts, and lymphoid infiltration (Fig. 2). Mallory bodies similar to those seen in alcoholic liver disease may be present adjacent to the areas of inflammation and there is excess stainable copper-binding protein, a reflection of the cholestasis.

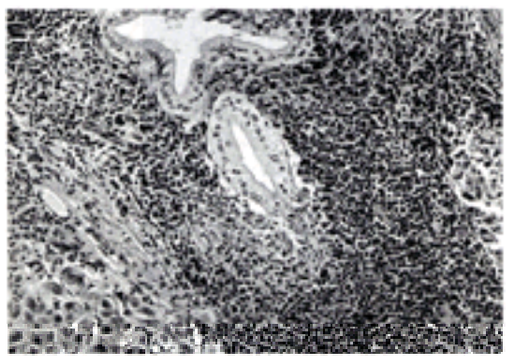

Fig. 1 Bile duct lesion in primary biliary cirrhosis. There is granulomatous destruction of a medium-sized bile-duct radicle in which the epithelium appears hyperplastic. Epithelioid macrophages are surrounded by a chronic inflammatory cell infiltrate. Haematoxylin and eosin stain. (By courtesy of A.D. Burt.)

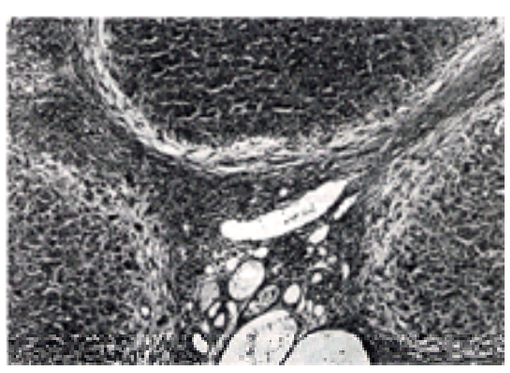

Fig. 2 Stage 4 primary biliary cirrhosis: an established micronodular cirrhosis; the halo effect seen around the nodules is a characteristic feature of biliary cirrhosis. Haematoxylin and eosin stain. (By courtesy of A.D. Burt.)

Malignancy

There is a dramatically increased risk for the development of hepatobiliary malignancies in patients with primary biliary cirrhosis with a reported relative risk of 46 for women and 55 for men. Hepatocellular carcinoma is a recognized complication of cirrhosis from any cause. Men are afflicted at least twice as often as women, and primary biliary cirrhosis is no exception to this rule. Hepatocellular carcinoma is a relatively common cause of death in male patients with concomitant primary biliary cirrhosis and liver cirrhosis and screening with regular liver ultrasound is recommended.

Diagnosis

The diagnosis is based on the serological and biochemical changes, together with the clinical findings and liver histology. A positive AMA may antedate all other abnormalities. Liver function tests reflect cholestasis, with increases in serum alkaline phosphatase and g-glutamyl transferase, but only modest changes in transaminases. At presentation, total serum bilirubin is usually normal or only modestly increased. The serum globulins are usually raised, particularly the IgM, but the serum albumin is usually maintained until late in the disease. Other tests such as erythrocyte sedimentation rate, cholesterol, and autoantibodies other than AMA are less specific. In a patient with a strongly positive AMA and the typical symptoms and biochemical abnormalities, a liver biopsy may not be essential, but is helpful. The features are very specific, and although several different histological stages may be found in one biopsy, the presence of fibrosis or cirrhosis indicates a worse prognosis.

The main differential diagnosis is from other causes of cholestasis. Good ultrasound examination of the liver and biliary tree is mandatory to exclude extrahepatic biliary obstruction or gallstones. Computed tomography, magnetic resonance cholangiopancreatography (**MRCP**), or endoscopic retrograde cholangiopancreatography (**ERCP**) may be necessary for patients without detectable AMA, many of whom have a positive antinuclear antibody and may be thought to have 'autoimmune cholangitis'. There is an overlap with autoimmune hepatitis, which can be diagnosed on liver histology, whilst primary sclerosing cholangitis will be evident on MRCP or ERCP.

Treatment

This consists of therapy aimed at modifying the disease process and progression to cirrhosis, and treatment of symptoms and late complications.

Numerous trials of specific therapy have been undertaken in the last 30 years; the agents that have been assessed are shown in Table 2. Over recent years the naturally occurring bile acid, ursodeoxycholic acid, has become an established treatment for primary biliary cirrhosis. Many randomized controlled trials comparing ursodeoxycholic acid with placebo have been published. All the trials reported an improvement in standard liver enzyme tests and serum bilirubin levels. Results from three large studies have been included in a combined analysis, which indicated that treatment with ursodeoxycholic acid was associated with a significant delay in the time to death or transplantation. However, this was not confirmed in a long-term study (median follow-up of 3.4 years), nor in a meta-analysis of published randomized controlled trials. Data suggest that ursodeoxycholic acid does not prevent ongoing bile-duct destruction but that it exerts its beneficial effect by protecting against the consequences of bile duct destruction. Ursodeoxycholic acid therapy does not benefit the symptom of fatigue and has a variable effect on pruritus. At best ursodeoxycholic acid, in a dose of 12 to 15 mg/kg per day, may retard progression to cirrhosis (histological stage 4), but it does not halt the disease and cannot be seen as a cure. However, it is safe and well tolerated.

Trials of combination therapy using ursodeoxycholic acid with methotrexate, colchicine, predisolone, oral budesonide, and mycophenolate mofetil have been reported; most are too small to evaluate the efficacy adequately, but one pilot study suggests ursodeoxycholic acid plus budesonide is superior in patients with early disease.

Itching can be an intolerable symptom in primary biliary cirrhosis and the first line of treatment is with cholestyramine. Improvement in itching has also been reported with rifampicin and opioid antagonists (nalmifene and naloxone). There is no indication for a fat-free diet unless the patient has symptoms related to steatorrhoea or xanthelasma with high serum cholesterol levels. Supplementation with medium-chain triglycerides may be necessary if adequate weight or nutrition cannot be sustained. A prolonged prothrombin time is treated with intramuscular vitamin K at a dose of 10 mg monthly. Injections of vitamin A (100 000 iu) and vitamin D (100 000 iu) are usually given every 2 months in jaundiced patients and vitamin E supplements may also be required. Osteomalacia is now rare, given such treatment. The principles developed for monitoring and treating postmenopausal osteoporosis can be followed for patients with primary biliary cirrhosis. The complications of portal

hypertension and of liver failure are treated appropriately.

Liver transplantation is now the accepted treatment for endstage primary biliary cirrhosis. Referral to a transplant centre should be considered as the bilirubin approaches 100 µmol/l, although patients with disabling symptoms such as intractable itching may need to be considered individually. Recurrence of primary biliary cirrhosis occurs in about 10 per cent of patients in the first few years after liver transplantation; the cumulative risk increases with time and may be affected by the immunosuppression used.

Prognosis

The progression of primary biliary cirrhosis is extremely variable. Asymptomatic patients have a reduced survival compared with an age-and gender-matched general population and about 40 per cent develop symptoms within 5 to 7 years. The most reliable determinant of prognosis is the serum bilirubin concentration; other factors associated with poor prognosis include weight loss, hepatomegaly, splenomegaly, histological stage, patient age, and impaired liver synthetic function. Several prognostic models have been validated in clinical studies; the most widely used is the Mayo risk score.

Quality of life aspects

Fatigue is present in more than 80 per cent of patients and is one of the worst symptoms. It is not related to disease severity and may be associated with depression. The Fisk Fatigue Severity Score is a reproducible measure of fatigue severity and can be used in the clinical assessment of patients and in therapeutic trials.

Further reading

Adams DH, Shields PL (2000). Lymphocyte recruitment and activation in primary biliary cirrhosis. *Immunological Reviews* **174**, 15–26.

Agarwal K, Jones DEJ, Bassendine MF (1999). Genetic predisposition to primary biliary cirrhosis. *European Journal of Gastroenterology and Hepatology* **11**, 603–6. [Part of 'Review in depth' of primary biliary cirrhosis.]

Gershwin ME *et al*. (2000). Primary biliary cirrhosis: an orchestrated immune response against epithelial cells. *Immunological Reviews* **174**, 210–25.

Goulis J, Leandro G, Burroughs AK (1999). Randomised controlled trials of ursodeoxycholic-acid therapy for primary biliary cirrhosis: a meta-analysis. *Lancet* **354**, 1053–60.

Heathcote JE (1999). Evidence based therapy for primary biliary cirrhosis. *European Journal of Gastroenterology and Hepatology* **11**, 607–15. [Part of 'Review in depth' of primary biliary cirrhosis.]

Heathcote EJ (2000). Management of primary biliary cirrhosis. *Hepatology* **31**, 1005–13. [Guidelines developed under the auspices of, and approved by, the Practice Guidelines Committee of the American Association for the Study of Liver Diseases.]

James OFW *et al*. (1999). Primary biliary cirrhosis once rare, now common in the United Kingdom? *Hepatology* **30**, 390–4.

Jones DEJ, James OFW, Bassendine MF (1998). Primary biliary cirrhosis: clinical and associated autoimmune features and natural history. In: Lindor KD, Dickson ER, eds. *Clinics in liver disease: primary biliary cirrhosis, primary sclerosing cholangitis, and adult cholangiopathies*, Vol 2, pp 265–82. WB Saunders, Philadelphia. [Part of an in depth review of adult cholangiopathies, including primary biliary cirrhosis.]

Jones EA, Bergasa NV (1999). The pathogenesis and treatment of pruritus and fatigue in patients with PBC. *European Journal of Gastroenterology and Hepatology* **11**, 623–31. [Part of 'Review in depth' of primary biliary cirrhosis.]

Kim WR, Dickson ER (1998). Predictive models of natural history in primary biliary cirrhosis. In: Lindor KD, Dickson ER, eds. *Clinics in liver disease: primary biliary cirrhosis, primary sclerosing cholangitis, and adult cholangiopathies*, Vol 2, pp 313–31. WB Saunders, Philadelphia. [Part of an in depth review of adult cholangiopathies, including primary biliary cirrhosis.]

Metcalf JV, James OFW (1997). The geoepidemiology of primary biliary cirrhosis. *Seminars in Liver Disease* **17**, 13–22.

Neuberger J (1997). Primary biliary cirrhosis. *Lancet* **350**, 875–9.

Nijhawan PK *et al*. (1999). Incidence of cancer in primary biliary cirrhosis: The Mayo experience. *Hepatology* **29**, 1396–8.

Shimoda S *et al*. (2000). Mimicry peptides of human PDC-E2 163–176 peptide, the immunodominant T-cell epitope of primary biliary cirrhosis. *Hepatology* **31**, 1212–16.

Yeaman SJ, Kirby JA, Jones DEJ (2000). Autoreactive responses to pyruvate dehydrogenase complex in the pathogenesis of primary biliary cirrhosis. *Immunological Reviews* **174**, 238–49.

14.20.2.3Primary sclerosing cholangitis

R. W. Chapman

Primary sclerosing cholangitis is a chronic cholestatic liver disease characterized by an obliterative inflammatory fibrosis of the biliary tract. It may lead to bile-duct obstruction, biliary cirrhosis, hepatic failure, and in some patients, cholangiocarcinoma. Primary sclerosing cholangitis was initially considered to be a rare disease; however, the advent of endoscopic retrograde cholangiopancreatography (**ERCP**) in the early 1970s established the diagnosis in a progressively larger number of patients. This led to the realization that primary sclerosing cholangitis has a much wider clinical and pathological spectrum than was previously recognized.

The generally accepted diagnostic criteria of primary sclerosing cholangitis are: (i) generalized bleeding and stenosis of the biliary system on cholangiography (Fig. 1); (ii) absence of choledocholithiasis or a history of bile-duct surgery; and (iii) exclusion of bile-duct cancer, usually by prolonged follow-up.

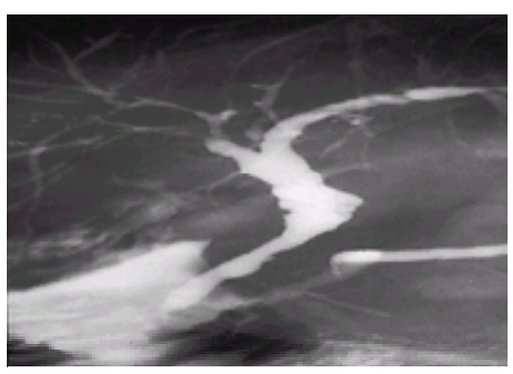

Fig. 1 Endoscopic retrograde cholangiogram showing the typical features of primary sclerosing cholangitis with stricturing and dilatation of the intra- and extrahepatic biliary tree.

The term secondary sclerosing cholangitis is used to describe the typical bile-duct changes when a clear predisposing factor to duct fibrosis, such as previous bile-duct surgery, can be identified. The causes of secondary sclerosing cholangitis are shown in Table 1.

Aetiology

The cause of primary sclerosing cholangitis remains unknown. There is a very close association, however, between primary sclerosing cholangitis and inflammatory bowel disease, particularly ulcerative colitis. Approximately two-thirds of Northern European patients with primary sclerosing cholangitis have coexisting ulcerative colitis, and primary sclerosing cholangitis is the most common form of chronic liver disease found in ulcerative colitis. In Southern Europeans about one-half of patients with primary sclerosing cholangitis will have ulcerative colitis. This difference in populations may be real or may represent differences in case finding as not all the patients studied had had colonoscopy and colonic biopsies performed. Three to 10 per cent of patients with ulcerative colitis will develop primary sclerosing cholangitis, and the prevalence is greater in patients with substantial or total colitis than in those with distal colitis only. In a Swedish study, the prevalence of ulcerative colitis was 171 per 100 000 population and primary sclerosing cholangitis 6.3 per 100 000 population. It is clear that any proposed factors in the aetiopathogenesis of primary sclerosing cholangitis must explain this close association with inflammatory bowel disease. Current studies have suggested that genetic and immunological factors are important in the pathogenesis of primary sclerosing cholangitis.

Immunogenetic factors

Case reports of families in whom members developed ulcerative colitis and primary sclerosing cholangitis led to the search for an HLA association. A close link with the *HLA A1-B8-DR3* haplotype has been found, in common with other organ-specific autoimmune diseases such as autoimmune chronic active hepatitis. *HLA DRw52a*, which is in linkage disequilibrium with *DR3*, is also closely linked to the development of primary sclerosing cholangitis. In British patients who are *DR3* and *Drw52a* negative, an increased prevalence of *HLA DR2* is found. *HLA A1-B8-DR3* and *DR2* are equally distributed in patients with primary sclerosing cholangitis, with or without ulcerative colitis. An independent association with *HLA DR6* has also been documented. It has been suggested that *DRw52a*, *DR6*, and *DR2* encode for amino acids in the HLA b-chain that may enhance antigen presentation by the HLA molecule to the T-cell receptor. Further evidence of an autoimmune basis for this condition has been provided by many studies that have shown humoral and cellular immune abnormalities.

Humoral immune abnormalities

Like primary biliary cirrhosis, a disease with which it shares many features (see Chapter 14.27.2), symptomatic primary sclerosing cholangitis is characterized by hypergammaglobulinaemia, often with a disproportionate elevation of serum IgM concentrations in adult patients. In contrast, high concentrations of serum IgG are found in all children with primary sclerosing cholangitis. Smooth-muscle antibody and antinuclear factor are also found in approximately one-third of patients with primary sclerosing cholangitis, usually in low titres.

Recently, a cytoplasmic antineutrophil antibody was found in the serum of 80 per cent of patients with primary sclerosing cholangitis and approximately 30 per cent of patients with ulcerative colitis. However, it is not specific for primary sclerosing cholangitis and is found in 50 per cent of patients with autoimmune chronic active hepatitis (type 1). It is not found in primary biliary cirrhosis. The antigens in primary sclerosing cholangitis are distinct from those found in Wegener's granulomatosis, which have been shown to be proteinase 3 and myeloperoxidase. Current evidence suggests that the antigen may be a nuclear envelope protein. The pathogenetic significance of the circulating antibody is not clear, but it may prove to be useful in a diagnostic test. Titres of the antibody do not change after hepatic transplantation.

Cellular immune abnormalities

Elevated circulating immune complexes associated with activation of complement via the classic pathway have been found in the serum and bile of patients with primary sclerosing cholangitis. In common with other autoimmune diseases, there are reduced levels of T-suppressor cells circulating in the serum of these patients, leading to an increased ratio of T-helper to T-suppressor cells. Infiltration of portal tracts by increased numbers of mononuclear cells is seen in liver biopsies from patients with primary sclerosing cholangitis. The majority of these cells are activated T lymphocytes.

Current evidence suggests that primary sclerosing cholangitis is an immunologically mediated disease, perhaps triggered in genetically susceptible subjects by acquired toxic or infectious agents, which are presented through antigen-presenting cells to activated T lymphocytes. Unlike normal biliary cells, the biliary epithelial cells in primary sclerosing cholangitis express HLA class II molecules and also intercellular adhesion molecules (**ICAM**) such as ICAM-I. It has not yet been confirmed, however, that bile-duct cells can act as antigen-presenting cells, as expression of other costimulatory molecules such as B7, which are needed for antigen presentation, are not consistently found on biliary epithelial cells (cholangiocytes).

Alternative hypothesis—exposure to bacterial components

An alternative hypothesis has been proposed in which the initial event is the reaction of an immunologically susceptible host to bacterial cell wall products. This reaction would result in hepatic macrophages producing tumour necrosis factor-a and endotoxin. The exposure to bacterial components and increased gut permeability would be increased by the presence of inflammatory bowel disease, but could also, in theory, occur during episodes of gut infection. The resulting increase in peribiliary cytokine and chemokine secretion would attract activated neutrophils, monocyte/macrophages, T cells, and fibroblasts. The deposition of concentric fibrosis could result in atrophy of the biliary epithelial cells secondary to ischaemia. The resulting bile duct loss would lead to progressive cholestasis, fibrosis, and secondary biliary cirrhosis. This hypothesis does not explain the relative scarcity of patients with Crohn's colitis and does not take into account the strong circumstantial evidence of immune mediation and autoimmunity, previously described.

Clinical features

There is a clear male predominance, with a male:female ratio of 2:1. The majority of patients present between the ages of 25 and 40 years, although primary sclerosing cholangitis may be diagnosed at any age. Indeed, it has become recognized recently as an important cause of chronic liver disease in children.

The clinical presentation is variable: some patients may present with fatigue, intermittent jaundice, weight loss, right upper quadrant pain, and pruritus. Attacks of acute cholangitis are surprisingly rare and usually follow instrumental biliary intervention, such as ERCP. Physical examination is abnormal in approximately half of symptomatic patients; the most common findings are jaundice and hepatosplenomegaly. Many patients with primary sclerosing cholangitis are asymptomatic at diagnosis, which is made incidentally when a persistently raised serum alkaline phosphatase is discovered, usually in the setting of ulcerative colitis.

Serum biochemical tests usually indicate cholestasis, but primary sclerosing cholangitis may cause no abnormalities of serum biochemistry. The serum alkaline phosphatase is often raised to greater than three times normal, and mild elevations in liver transaminases are seen in the majority of patients. Serum bilirubin is not usually elevated until later stages of the disease. Levels of bilirubin and alkaline phosphatase may fluctuate widely in an individual patient during the course of the disease. Hypoalbuminaemia is unusual until the disease becomes advanced. As mentioned above, increased serum IgM concentrations are seen in about half of the symptomatic adult patients, but high concentrations of IgG are always found in children with primary sclerosing cholangitis.

In addition to the serum antineutrophil antibodies, low levels of antinuclear antibody and smooth-muscle antibody may be found in approximately one-third of patients, but serum mitochondrial antibodies are absent.

Diagnosis

Radiological features

The cholangiographic appearances on ERCP are usually diagnostic and consist of multiple, irregular stricturing and dilatation (beading of the intrahepatic and extrahepatic biliary ducts) (Fig. 1). Occasionally, involvement is localized to the intrahepatic system, and even more rarely, only the extrahepatic bile ducts may be involved. Small diverticula are found along the common bile duct in about 20 per cent of patients and are pathognomonic (Fig. 2). Magnetic resonance cholangiopancreatography provides a non-invasive method of imaging the biliary tree, and will become the standard technique for the diagnosis of primary sclerosing cholangitis. (Fig. 3) Approximately 20 per cent of patients have stricturing of the main pancreatic duct, although exocrine pancreatic insufficiency is rare.

Fig. 2 Endoscopic retrograde cholangiogram from a patient with primary sclerosing cholangitis showing a diverticular appearance of the common bile duct.

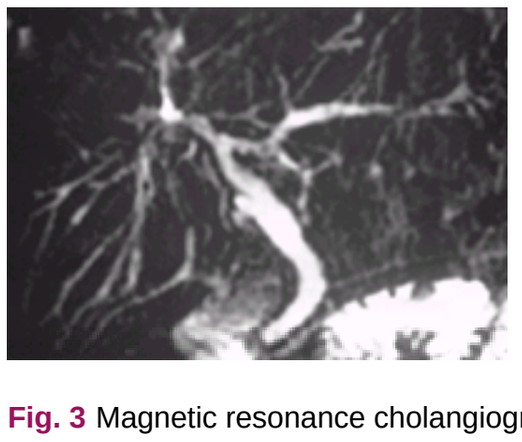

Fig. 3 Magnetic resonance cholangiogram demonstrating hilar stricture and intrahepatic involvement of the biliary tree.

Pathological features

The histological appearances of liver are not usually diagnostic for primary sclerosing cholangitis, although some form of biliary disease can usually be identified. The characteristic early features of primary sclerosing cholangitis are periductal 'onion skin' fibrosis and inflammation, portal oedema, and bile ductular proliferation resulting in the expansion of the portal tracts (Fig. 4). Later, fibrosis spreads into the liver parenchyma to form fibrous septa, leading inevitably to biliary cirrhosis. As in primary biliary cirrhosis, with disease progression an obliterative cholangitis occurs, leading to complete replacement of the intralobular bile ducts by connective

tissue—the so-called vanishing bile-duct syndrome. In addition, piecemeal necrosis, copper-binding protein, cholestasis, and occasional portal phlebitis may be present.

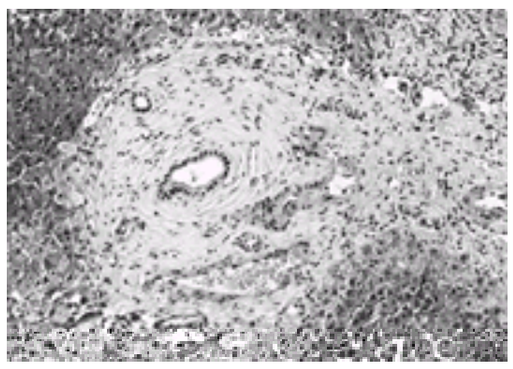

Fig. 4 The hepatic histological changes of early primary sclerosing cholangitis showing a concentric (onion skin) fibrosis around the bile ducts.

Association with other diseases

A large number of diseases have been associated with primary sclerosing cholangitis (Table 2). The most important association, as discussed above, is with inflammatory bowel disease, particularly ulcerative colitis. The extent of the colitis is usually total but symptomatically, and paradoxically, mild, often with no rectal bleeding and characterized by prolonged remission. Rectal sparing is found in 20 per cent of patients with ulcerative colitis and primary sclerosing cholangitis compared with 5 per cent of patients with ulcerative colitis alone. Although the symptoms of ulcerative colitis usually develop before those of primary sclerosing cholangitis, the onset of the latter may precede the symptoms of colitis by some years. The outcome of primary sclerosing cholangitis is completely unrelated to the activity, severity, or clinical course of the colitis, and colectomy has no effect on the progression of the cholangitis. Primary sclerosing cholangitis is less common in Crohn's disease, occurring in less than 1 per cent of patients and only in those with Crohn's colitis. Patients with primary sclerosing cholangitis and ulcerative colitis are at greater risk of developing colorectal dysplasia and colonic cancer than those with ulcerative colitis alone. In a Swedish study, the absolute accumulative risk of developing colorectal dysplasia/cancer in the primary sclerosing cholangitis/ulcerative colitis group was 9, 31, and 50 per cent, respectively, after 10, 20, and 25 years of disease duration. In the group with ulcerative colitis alone, the corresponding risk was 2, 5, and 10 per cent, respectively.

Natural history and prognosis

The course of primary sclerosing cholangitis is highly variable. The median survival from presentation to death or liver transplantation in symptomatic patients is approximately 10 to 12 years, whilst approximately 75 per cent of asymptomatic patients survive 15 years or more. The majority of patients die in hepatic failure following deepening cholestatic jaundice. However, approximately 10 to 30 per cent of patients with long-standing primary sclerosing cholangitis die from the development of bile duct carcinoma, which often follows a very aggressive course. The mean survival after the diagnosis of cholangiocarcinoma is only 9 months. Unfortunately, there are no factors that will predict which patients will develop this cancer. Tumour markers such as CEA and CA 19–9 have been investigated as potential serum markers of the development of bile duct cancer in primary sclerosing cholangitis. Although some centres have found elevations in serum CA 19–9 a useful predictor, these results have not been confirmed in other units. Attempts to model factors that will predict the risk of progression to liver failure and death have yielded conflicting data from different centres. It is probable that the majority of asymptomatic patients will progress insidiously to symptomatic liver disease, liver failure, and death.

Treatment

Symptomatic measures

There is no curative treatment for primary sclerosing cholangitis. This is indicated by the plethora of medical, endoscopic, and surgical approaches that has been advocated.

Management of cholestasis

Symptomatic patients are frequently troubled by pruritus. This is best managed initially by cholestyramine and the dose should be increased until relief is obtained. Second line treatments include rifampicin and the opioid antagonist naltrexone. In addition, replacement of fat-soluble vitamins is necessary when patients become jaundiced. Metabolic bone disease (usually osteoporosis) is a common complication of advanced primary sclerosing cholangitis. Calcium supplementation with vitamin D_3 should be given prophylactically in jaundiced patients and bisphosphonates considered in patients with osteoporosis.

Management of complications

Broad-spectrum antibiotics such as ciprofloxacin should be given for acute attacks of cholangitis, but they have no proven prophylactic value and should not be used in the long term routinely. If cholangiography shows a well-defined obstruction to the main extrahepatic bile ducts, then mechanical relief must be considered. In many patients the best approach is to introduce a prosthesis (stent) through the obstruction. This may be placed non-operatively by the percutaneous transhepatic route or at ERCP. Balloon dilatation of the strictures before stenting may prove useful in a minority of patients with well-defined localized strictures and can lead to a striking improvement in symptoms and serum biochemistry.

Another common complication is the development of small biliary stones (brown pigment) and biliary sludge, which can lead to a rapid clinical or biochemical deterioration. In these patients, endoscopic sphincterotomy with extraction of the biliary debris can be beneficial.

Small duct disease

A few patients with ulcerative colitis will have persistently abnormal cholestatic liver function tests with typical histological appearances such as concentric fibrosis but with normal bile ducts at cholangiography. The term 'small-duct primary sclerosing cholangitis' has been proposed to replace the term 'pericholangitis' in this group of patients as the evidence suggests that these conditions are all part of the same disease spectrum. Only a minority of patients with 'small-duct disease' will progress to develop extrahepatic biliary involvement and they are not predisposed to develop cholangiocarcinoma.

Specific treatment

Medical

The medical treatment of primary sclerosing cholangitis has included trials of corticosteroids, immunosuppressive drugs, cholecystogogues, and antibiotics, either alone or in combination. The results have been universally disappointing, although assessment of treatment of this uncommon disease is difficult because the clinical course fluctuates, survival is variable, and some patients may remain asymptomatic for long periods of time. The role of corticosteroid therapy is unclear. There have been no large controlled trials, but corticosteroids have been used topically and systemically in small and generally uncontrolled studies. However, there is evidence that, even in male patients, metabolic bone disease may be accelerated by corticosteroids and in general they should not be used in this condition.

Ursodeoxycholic acid is a non-hepatotoxic hydrophilic bile acid which has been used widely for the treatment of cholestasis—it reduces levels of cholestatic liver enzymes. Controlled trials in concentrated doses (10 to 15 mg/kg body weight) have shown no effect on symptoms, histology, or survival. Recent trials suggest that

larger doses are needed to produce a beneficial effect, with improvement in histology at 2 years.

A number of immunosuppressant agents have been tried, either alone or in combination, including azathioprine, methotrexate, and cyclosporin. Overall, the results have been disappointing.

Surgical

The role of hepatobiliary surgery in the treatment of primary sclerosing cholangitis remains controversial. Good results have been claimed for the resection of the extrahepatic biliary tree followed by biliary reconstruction with silastic transhepatic stents. However, controlled trials are needed to confirm the efficacy of these and other surgical techniques, as previous biliary surgery will increase perioperative mortality from hepatic transplantation.

Transplantation

Orthotopic liver transplantation is the only option available in young patients with primary sclerosing cholangitis and advanced liver disease. Primary sclerosing cholangitis is now the second most common indication for liver transplantation in the United Kingdom. Recent results have been very encouraging, with 5-year survival rates of 80 to 90 per cent being obtained in most centres. These rates compare favourably with those for other forms of chronic liver disease. It has become clear that primary sclerosing cholangitis recurs in the transplanted liver in 20 per cent of patients at 1 year post-transplant, but only rarely has recurrence led to problems with liver decompensation requiring retransplantation. Proven cholangiocarcinoma is a contradiction to transplantation because the tumour recurs rapidly after transplantation with immunosuppression. As patients suffering from primary sclerosing cholangitis in combination with ulcerative colitis have an increased risk for the development of colon cancer after transplantation, yearly colonoscopy has been recommended in this group. Several centres have noted a worsening in the symptoms of ulcerative colitis after transplantation; the explanation for this phenomenon remains unclear.

Further reading

Broome U *et al.* (1995). Primary sclerosing cholangitis and ulcerative colitis: evidence for increased neoplastic potential. *Hepatology* **22**, 1404–8.

Broome U, Olsson RK, Loof L (1996). Natural history and prognostic factors in 305 Swedish patients with primary sclerosing cholangitis. *Gut* **38**, 610–15.

Chalasami N *et al.* (2000). Cholangiocarcinoma in patients with primary sclerosing cholangitis: a multicentre case–control study. *Hepatology* **31**, 7–11.

Chapman RW *et al.* (1980). Primary sclerosing cholangitis—a review of its clinical features, cholangiography and hepatic histology. *Gut* **21**, 870–7.

Donaldson PT *et al.* (1991). Dual association of HLA DR2 and DR3 with primary sclerosing cholangitis. *Hepatology* **13**, 129–33.

Graziadei IW *et al.* (1999). Long term results of patients undergoing liver transplantation for primary sclerosing cholangitis. *Hepatology* **30**, 1121–7.

Graziadei IW *et al.* (1999). Recurrence of primary sclerosing cholangitis following liver transplantation. *Hepatology* **29**, 1050–6.

Johnson GK *et al.* (1991). Endoscopic treatment of biliary tract strictures in sclerosing cholangitis: a large series and recommendations of treatment. *Gastrointestinal Endoscopy* **37**, 38–43.

Lindor KD (1997). Ursodiol for primary sclerosing cholangitis. Mayo Primary Sclerosing Cholangitis–ursodeoxycholic Audit Study Group. *New England Journal of Medicine* **336**, 691–5.

Lo SK, Fleming KA, Chapman RW (1992). Prevalence of antineutrophil antibody in primary sclerosing cholangitis and ulcerative colitis using an alkaline phosphatase method. *Gut* **33**, 1370–5.

Ludwig J *et al.* (1981). Morphological features of chronic hepatitis associated with primary sclerosing cholangitis and ulcerative colitis. *Hepatology* **1**, 632–40.

Manns MP *et al.* (1998). *Primary sclerosing cholangitis*. Kluwer Academic Publishers, London.

Mitchell SA *et al*. (2001). A prelimary trial of high dose ursodeoxycholic acid in primary sclerosing cholangitis. *Gastroenterology* **122**, 900–7.

Terjung B *et al.* (1998). Atypical antinuclear cytoplasmic antibodies with perinuclear fluorescence in chronic inflammatory bowel diseases and hepatobiliary disorders colocalise with nuclear lamina proteins. *Hepatology* **28**, 332–40.

14.21.1 Alcoholic liver disease and non-alcoholic steatosis hepatitis

O. F. W. James

Alcoholic liver disease

Only 10 to 30 per cent of heavy, persistent, alcohol drinkers develop cirrhosis, although well over 50 per cent have fatty livers. Individual susceptibility depends on many factors. The contribution of nutrition remains controversial, but is seems possible that both undernutrition and obesity act synergistically with direct alcohol toxicity to increase the likelihood of liver damage. High alcohol consumption in patients infected with hepatitis C virus also increases the possibility of severe liver disease.

In a large group of males with alcoholic cirrhosis, average alcohol consumption was 160 g/day (equivalent to over two bottles of wine, 4.5 litres of normal strength lager or two-thirds of a bottle of spirits) over 8 years. In females the corresponding figure was 110 g/day. It seems likely that almost no risk of significant alcohol-related liver damage exists below about 40 g/day equivalent to 30 units per week in men, rather less in women, assuming no other associated risk factors. Current 'sensible' limits recommended in the United Kingdom are 21 units (200 g) for men and 14 units (130 g) for women, but these figures are arbitrary.

Susceptibility to alcoholism and to alcoholic liver damage each have genetic components; probably one in three alcoholics will have at least one parent who is alcoholic. Analysis of twin studies suggests hereditability of excess drinking of alcohol in the range of 0.3–0.6 (where 0 = no hereditability, 1.0 = complete hereditability).

Pathology

Fatty liver is the first histological lesion; it occurs in most heavy drinkers at one time or another, but is completely reversible on alcohol withdrawal. More serious is alcoholic hepatitis, which may occur in up to 40 per cent chronic ethanol abusers. The most severe changes are seen in the perivenular area; including ballooning and necrosis of liver cells, in some of which Mallory bodies may be seen; pericellular fibrosis around hepatic venules ('chicken wire' fibrosis); and a patchy inflammatory-cell infiltrate, mainly polymorphs, often only seen around a few hepatocytes. Ultimately, fibrous septa link hepatic veins to portal veins and regeneration occurs, disturbing normal liver architecture with the formation of nodules and cirrhosis.

Clinical features

Symptoms and signs of alcoholic liver disease (Table 1) correlate only very broadly with underlying histology or abnormal tests of liver function.

Patients with fatty liver may have no symptoms or complain only of nausea and malaise. Liver function tests may be mildly deranged but in severe cases, there may be cholestasis or even, very rarely, liver failure and portal hypertension.

Mild alcoholic hepatitis is often indistinguishable clinically from fatty liver, with which it usually coexists, but in more severe cases anorexia, nausea, abdominal pain, and weight loss may develop. Severe alcoholic hepatitis, with or without cirrhosis, is a medical emergency; patients are at risk of ascites, bleeding, and encephalopathy. They may also become infected—urinary tract infection, pneumonia, spontaneous bacterial peritonitis, or septicaemia. Detection of such infection is complicated by the fever and leucocytosis caused by the liver disease itself. Those most severely affected may develop profound cholestasis and hypoglycaemia.

The picture in established cirrhosis is variable. In some who have stopped drinking there may be no symptoms and liver function tests may be near normal. More often, patients suffer malaise and will have lost weight and show classical physical signs (Table 1), may be jaundiced, and may bleed from varices. Zieve's syndrome of marked jaundice from a combination of cholestasis and haemolysis with hyperlipidaemia is rare.

History and investigation

Many patients are reluctant to admit their alcoholism, even when liver disease is gross (see Chapter 26.7.2). Liver function tests are often unhelpful in establishing severity, except at the late stage. Early abnormalities may include raised serum g-glutamyl transferase and macrocytosis. Measurement of blood or urinary ethanol is helpful if high levels are found. There may be a disproportionate elevation of serum aspartate transaminase compared to alanine transaminase. Serum ferritin may be very elevated in active heavy drinkers. Alcoholism is a common cause of combined hyperlipidaemia, indeed serum may be very hyperlipaemic. Liver biopsy allows confirmation of histological severity and exclusion of other pathologies. White-out on colloid liver scan implies severe alcoholic hepatitis.

Prognosis

Prognosis is above all related to whether the patient continues to drink or stops. In patients with fatty change alone the outlook is excellent, provided patients stop or substantially cut down drinking, although some may progress to more advanced liver disease. Mild alcoholic hepatitis has a similar prognosis to fatty change. In severe, acute alcoholic hepatitis, whether or not superimposed upon cirrhosis, there is a 12 to 50 per cent mortality within 6 months of presentation. Particularly adverse features are raised bilirubin level and abnormal blood clotting. This had led to the use of a discriminant function to help assess prognosis and decide upon treatment (Table 2). In alcoholic cirrhosis overall survival at 5 years is about 50 per cent, among abstainers 70 per cent, and in those who continue to drink 35 per cent. A second important prognostic feature is age at presentation. In a recent United Kingdom study, 3-year survival was 77 per cent in patients under the age of 60, and 46 per cent in those presenting over that age. Nutrition (possibly reflecting socio-economic status) also influences survival. Unfortunately, hepatocellular cancer can arise in patients with long-standing, often inactive, alcoholic cirrhosis, particularly men.

Treatment (see also Section 26.7)

The best treatment remains total withdrawal of alcohol and subsequent long-term abstinence in all patients with liver disease worse than moderate fatty change alone, in which case, very moderate drinking after a period of abstinence may be an option. Good prognostic features include recognition of the problem by the patient, a supportive family, steady employment, and willingness to accept treatment. No other treatment is required for patients with fatty liver or mild alcoholic hepatitis.

Severe alcoholic hepatitis

Meta-analysis of more than 10 trials of high-dose corticosteroid treatment added to conventional therapy has led to the following recommendations. In patients with discriminant function over 32 (Table 2) but who have no overt sepsis or active bleeding, 40 mg prednisolone for 21 days probably provides a 20 per cent improvement

in mortality. Insulin and glucagon, anabolic steroids, colchicine, enteral and parenteral nutrition, and a variety of other so-called 'hepato-protective drugs' have all been used in trials but without real evidence of benefit with respect to mortality.

Cirrhosis

Treatment is directed against its complications, particularly portal hypertension, ascites, spontaneous bacterial peritonitis, and encephalopathy.

Transplantation

The indications for transplantation in alcoholic cirrhosis are similar to those for other endstage liver diseases (see later), with the caveat that even in advanced alcoholic cirrhosis, abstention can lead to enormous clinical improvement and long-term survival. Many transplant units consider patients only after a 6-month period of abstention, both to detect patients in whom transplantation is no longer necessary and to exclude individuals who continue to drink heavily. Estimated 5-year survival following transplantation is now over 70 per cent.

Non-alcoholic steatosis hepatitis

This is increasingly recognized as an important distinct clinical entity. Originally described in, usually very, obese females and in diabetics, or patients with hyperlipidaemia, also in patients following jejunoilial bypass surgery for obesity. This condition is now recognized in individuals of both sexes who are only marginally obese but who may have changed weight recently. Patients often present because of detection of abnormal liver function tests. It is important to make this diagnosis both for prognostic reasons and to clearly state the non-alcoholic nature of this disease in an individual for purposes of employment and insurance.

Pathology

This is identical to alcoholic liver disease, most patients having simple steatosis, others develop an alcoholic hepatitis-like appearance with or without fibrosis, a small proportion (perhaps 5–10 per cent) develop cirrhosis.

Clinical features

Most are asymptomatic but are 'accused' of having alcoholic liver disease. Up to 40 per cent have persistent right upper abdominal pain and may complain of lethargy and malaise. About 5 to 10 per cent, usually very obese and diabetic, develop cirrhosis with its complications.

Investigations

Patients must have a history of high alcohol consumption exhaustively excluded as a cause for their liver disease. Liver function tests show raised serum transaminases, unlike alcoholic liver disease serum alanine transaminase is raised compared with aspartate transaminase. Serum g-glutamyl transferase is also raised. Liver ultrasound shows a fatty appearance but cannot reliably distinguish the extent of fibrosis.

Prognosis

This is excellent except in the small proportion of those who develop cirrhosis where complications and clinical course are as for other causes of cirrhosis.

Treatment

No treatment has yet been proven to be effective. Probably slow weight reduction, a reduced fat diet, and a period of abstinence from any alcohol consumption are most effective.

Further reading

Day CP, Bassendine MF (1992). Genetic predisposition to alcoholic liver disease. *Gut* **33**, 1344–7.

Hislop WS, *et al*. (1983). Alcoholic liver disease in Scotland and north-eastern England; presenting features in 510 patients. *Quarterly Journal of Medicine* **206**, 232–3.

Sherlock S, Dooley J (2001). *Diseases of the liver and biliary system*, 2nd edn. Blackwell Science, Oxford.

14.21.2Cirrhosis, portal hypertension, and ascites

Kevin Moore

Introduction

Ascites is the accumulation of fluid in the peritoneal cavity. It has fascinated doctors for many years, and studies on its pathogenesis were initiated as long ago as the seventeenth century. Richard Lower (1631–1691), a physician based in Oxford, demonstrated that ascites developed in dogs following ligation of the inferior vena cava. Ernest Henry Starling (1866–1927), a physiologist based at University College London, made the greatest contribution to the study of oedema formation, with the demonstration that both hydrostatic forces and oncotic forces were involved. He also showed that the increase in thoracic lymph flow following obstruction of the inferior vena cava is mainly derived from the liver.

Aetiology

Ascites is a common complication of cirrhosis and indicates the presence of portal hypertension and hepatic decompensation. It occurs in at least 50 per cent of patients within 10 years of the diagnosis of cirrhosis, which accounts for over 75 per cent of cases presenting with ascites. Ascites may be due to malignancy, pancreatitis, tuberculosis, cardiac failure, myxoedema, or other rarer causes, each of which may also occur in patients with cirrhosis (Table 1). Ascites does not occur in patients with portal vein thrombosis or other forms of non-cirrhotic portal hypertension such as congenital hepatic fibrosis, except as a transient finding following a gastrointestinal haemorrhage. It frequently occurs in patients with the Budd–Chiari syndrome or late-onset hepatic failure (subfulminant hepatic failure), and, to a lesser extent, where small amounts of peritoneal fluid accumulate in cases of acute liver failure.

Other (rare) causes of ascites include constrictive pericarditis, malnutrition, stromal tumours and Meigs' syndrome, hypothyroidism, Budd–Chiari syndrome, veno-occlusive disease, or lymphatic leak (chylous ascites). Rare infections include candidiasis and filariasis. Granulomatous liver disease such as sarcoidosis may cause severe portal hypertension, and occasional ascites. Although ascites commonly occurs in patients with cardiac failure, it is not usually a presenting feature. Ascites may also occur in the ovarian hyperstimulation syndrome in women undergoing fertility treatment.

Epidemiology

Cirrhosis is the eleventh leading cause of death in the United States. It heralds the beginning of a usually rapid decline of liver function so that about half the patients die within 2 years of the onset of ascites.

Pathogenesis of ascites due to cirrhosis

The presence of portal hypertension is essential for the development of ascites: fluid accumulation does not occur at a portal pressure below 8 mmHg. However, factors other than portal pressure are important, since ascites does not develop spontaneously in patients with portal vein thrombosis. Ascites develops as a consequence of sodium and water retention, which in the presence of portal hypertension causes transudation of fluid into the peritoneal cavity, and together with an increased production of hepatic lymph may cause a massive accumulation of fluid—a moderate to marked ascites will comprise about 5 to 25 litres of fluid. Portal hypertension is an essential prerequisite for the development of ascites, since it does not develop when salt and water retention is due to mineralocorticoid excess (for example, Conn's syndrome or a secreting adrenal carcinoma), in which the cardinal manifestation is hypertension. Whilst it is well recognized that abnormal sodium handling occurs in patients with advanced cirrhosis, it less well known that sodium handling is abnormal even in the preascitic stage. Experiments by Dudley and colleagues have shown that the proximal tubular reabsorption of sodium is enhanced in early cirrhosis, and that some patients exhibit glomerular hyperfiltration. Sodium retention does not occur as a result of hyperaldosteronism, for approximately 60 per cent of patients with ascites have normal aldosterone levels at initial presentation. There is no doubt, however, that aldosterone plays a role in sodium balance, as it increases sodium retention in the distal tubules. The administration of high doses of spironolactone increases natriuresis in patients with cirrhosis and ascites, despite apparently normal plasma aldosterone levels. This has given rise to the concept that renal sensitivity to aldosterone may be enhanced in cirrhosis. Other factors known to be involved in sodium homeostasis include the sympathetic nervous system, which is activated in decompensated liver disease and which enhances sodium reabsorption along the proximal tubules. A third important factor is the rate of sodium delivery to the tubules. Renal blood flow is decreased in cirrhosis, and decreases further with decompensation. During decreased sodium delivery, sodium reabsorption is enhanced. Other factors such as atrial natriuretic peptide, endothelin-1, or urodilatin may be involved, but their role is as yet undefined.

The underlying cause of the activation of sodium-retaining pathways is still disputed, but data are available to support both the underfill and overfill hypotheses. There is no doubt that at different stages of disease development the same patient may exhibit signs of both an expanded and a contracted central blood volume. A unifying hypothesis, known as the 'vasodilatation hypothesis', was put forward to explain the observations regarding the development of salt-retaining states (Fig. 1). In this, the central stimulus for activation of neurohumoral pathways is a decrease in the central blood volume, which differs in severity depending on the intensity of liver disease. Whilst superficially attractive, this hypothesis fails to explain why many patients develop salt retention in the presence of normal aldosterone concentrations, and why systemic vasodilatation is only observed in the supine state.

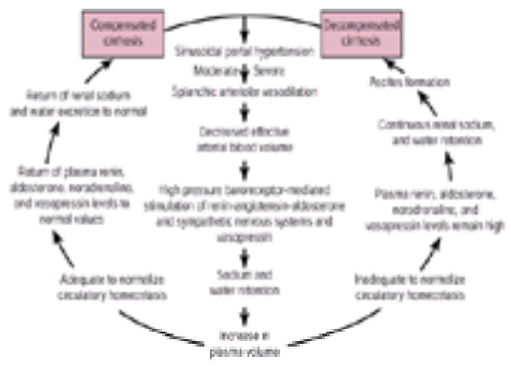

Fig. 1 Outline of peripheral vasodilation hypothesis (Schrier *et al*. (1988). Peripheral arterial vasodilation hypothesis: a proposal for the initiation of renal sodium and water retention in cirrhosis. *Hepatology* **8**, 1151–7).

Clinical features

Ascites is graded 1 to 3 depending on its severity. Grade 1 ascites is mild, and only detectable by ultrasound examination. Grade 2 ascites is moderate, and is manifest by moderate symmetrical distension of the abdomen; whereas grade 3 ascites is large or gross, with marked abdominal distension. A grade 2 ascites is most easily detected by a shifting dullness. Grade 3 ascites is usually tense and easily detected by the presence of a fluid thrill on palpation. There is often divarification of the rectus abdomini muscles, and prominent veins may be evident on the abdominal wall (see Fig. 2). Paraumbilical hernias develop in about 20 per cent of patients with ascites, an incidence that increases to up to 70 per cent in those with long-standing recurrent tense ascites. The main risks are rupture and strangulation. Pleural effusions (hepatic hydrothorax) develop in about 5 per cent of patients with cirrhosis. Hepatic hydrothorax may develop in patients with no discernible ascites. The pleural effusions are right-sided in 85 per cent of cases, left-sided in 13 per cent and bilateral in 2 per cent of cases.

Fig. 2 Alcoholic cirrhosis with ascites is often associated with marked anorexia.

Laboratory diagnosis

The cause of ascites or its precipitation is often self-evident. Where there are no obvious clues to its aetiology, tests must be directed at diagnosing both the presumed cause of liver disease and/or at excluding other causes of ascites such as malignancy or tuberculosis, etc. Other causes of abdominal distension such as huge masses, Meigs' syndrome, or pregnancy should be considered. The essential investigations on admission of a patient to hospital include:

- *Serum urea and electrolyte concentrations*—patients with ascites due to cirrhosis are prone to hyponatraemia or renal impairment, either spontaneously or following diuretic therapy.
- *Ascitic aspiration* with microscopy, determination of albumin or protein content, culture, cytology, and amylase measurement should be performed to confirm or exclude spontaneous bacterial peritonitis, tuberculosis, malignancy, or pancreatic disease. A Gram stain is usually uninformative.
- *Ultrasound scans* are needed to evaluate liver appearance (nodular and cirrhotic) or congested (for example, congestive cardiac failure (CCF)), as well as blood flow in the portal vein (to exclude the portal vein thrombosis that occurs in 8 per cent of patients with cirrhosis, and which may precipitate hepatic decompensation), a semi-quantitation of the amount of ascites, and the presence of tumour in the liver or other masses.

Paracentesis

An ascitic tap is used for either diagnostic purposes or for the therapeutic removal of large volumes of ascites. The most common site for aspirating ascites is about 15 cm lateral to the umbilicus, with care being taken to avoid an enlarged liver or spleen. The epigastric arteries run just lateral to the umbilicus towards the mid-inguinal point and should also be avoided.

For diagnostic purposes, between 20 and 50 ml of ascitic fluid should be withdrawn, and 3 to 5 ml placed under aseptic conditions (i.e. the needle changed) into each of two blood culture bottles (tuberculin cultures have to be requested specifically). Fluid for culture should **not** be sent in plastic containers for culture—when a sample of infected ascites is placed into such containers the culture positivity is only 40 per cent, whereas it is more than 90 per cent positive when inoculated into blood culture bottles. A 5-ml aliquot of fluid should be sent in a plastic container to the microbiology department for a polymorphonuclear neutrophil (PMN) or lymphocyte count. The microbiologist should report a neutrophil count (e.g. < 250 PMNs/mm ^{3}) together with a limited differential (i.e. PMN or lymphocytes). Ascitic fluid occasionally clots, in which case a sample should be placed in a tube containing EDTA (ethylenediaminetetraacetic acid). Coulter-counter estimations of ascitic neutrophil numbers are probably unreliable at the lower end of a pathologically increased white blood cell count. A 5-ml aliquot should be sent for measurement of ascitic protein content or, ideally, ascitic albumin. Ascitic cytology should involve liaison with the cytopathologist. Cytology requires 20 to 50 ml of ascitic fluid. There is little diagnostic value in an analysis of ascitic fluid pH or of lactate or glucose concentrations.

Ascitic fluid investigations

Ascitic protein

The use of ascitic protein in the differential diagnosis of the causes of ascites is much over-rated, and misinterpreted. Conventionally, the type of ascites is divided into exudates and transudates: ascitic protein concentration over 25 g/l or under 25 g/l, respectively. The purpose of this subdivision is to narrow the differential diagnosis of the causes of ascites. However, many physicians assume that cardiac ascites will have a low level of ascitic protein, when this is rarely the case, and that patients with tuberculous peritonitis have a high ascitic protein content, when in fact it is low in 30 per cent of patients. Moreover, about 15 per cent of cases of cirrhotic ascites have an ascitic protein level of more than 25 g/l, while 20 per cent of patients with a malignancy have a low ascitic protein level. The causes of transudative and exudative ascites are given in Table 2. For those patients with cirrhosis, a very low ascitic protein level (< 10 g/l) is associated with an increased risk of spontaneous bacterial peritonitis (**SBP**) at the time of hospital admission. SBP is present in about 15 per cent of all patients admitted with cirrhotic ascites. The use of ascitic protein estimations to subdivide patients into exudative or transudative causes of ascites is considerably enhanced by measuring the difference between ascitic and serum albumin levels (see below).

Serum-ascites albumin gradient

Several studies have compared the value of ascitic protein with serum-ascitic albumin gradient measurements in patients with ascites resulting from cirrhosis and

other causes. In one study comprising 44 patients, it was reported that 5/29 (17 per cent) patients with cirrhotic ascites had an ascitic protein level above 25 g/l, whereas 3/15 (20 per cent) with malignant ascites had an ascitic protein level below 25 g/l. In contrast, the overlap in each group was reduced to 1/29 and 1/15 respectively when a serum-ascitic albumin gradient above 11 g/l was used. Based on the many inaccurate predictions of aetiology based on the measurement of ascitic protein, it is clearly preferable to measure the serum-ascitic albumin gradient in patients presenting with ascites. This method, which involves subtraction of the ascitic albumin concentration from that observed in ascites, very accurately divides patients into two groups: those with a high gradient (>11 g/l) and those with a low gradient (< 11 g/l). The overall accuracy of this method is 97 per cent. (Table 3)

Ascitic amylase

The ascitic fluid amylase level should always be measured in patients with an exudative or unexplained ascites. A very high value is obtained when ascitic fluid results from a pancreatic pseudocyst or mass.

Ascitic fluid microscopy

An ascitic neutrophil count of more than 250 PMNs/mm^3 is diagnostic of spontaneous bacterial peritonitis. An elevated lymphocyte count should raise the possibility of tuberculous peritonitis. Excess red blood cells are most commonly due to a traumatic tap, but should raise the possibility of malignancy.

Ascitic fluid culture

Classically, aspirated fluid is placed in a sterile container and sent to the microbiology department for microscopy and culture. For infected fluid handled in this way, a positive culture will be obtained in only about 40 per cent of samples. However, if ascitic fluid is treated in the same way as blood cultures, and fluid inoculated directly into blood culture bottles at the bedside, the positive culture rate increases to 92 per cent (see above). A single study has evaluated the effectiveness of cytospin with cell lysis to improve the efficacy of culturing ascitic fluid, and again found that direct inoculation of ascitic fluid into blood culture bottles was far superior (79 per cent positivity) compared with the cell-lysis method (46 per cent culture-positive).

Ascitic fluid cytology

Ascitic cytology should involve liaison with the cytopathologist so that the index of suspicion and type of potential tumour are discussed. A 20- to 50-ml sample of ascitic fluid is required to produce a cell concentrate for cytology—obtained by centrifuging the ascites fluid, removing supernatant, and resuspending the cells. A sample of the concentrate then undergoes a cytospin to deposit cells on to microscope slides, following which the cells are stained. Typical stains include the Papanicolaou and May-Grunwald–Giemsa stain.

Ascitic volume

Ascitic volume is not usually determined in clinical practice. It can, however, be quantified radiologically or by indicator-dilution. As a rough guide, patients with barely detectable ascites usually harbour between 1 and 4 litres, those with moderate ascites 4 to 8 litres, and those with marked ascites more than 8 litres of fluid. Ultrasonographic determination of ascitic volume involves measurement of the abdominal circumference and the deepest vertical depth of the fluid, and modelled as a segment of a sphere. Isotopic determination of ascitic volume involves the injection of radiolabelled ^{99}Tc- macroalbumin.

Treatment

Patients with ascites can be divided into those that are easy to treat and those that are difficult. In general, patients with their first presentation of ascites and normal renal function, who have a spot urine sodium concentration of more than 20 mmol/l, or an identifiable source of dietary sodium excess, respond well to simple measures. Likewise, when ascites has developed as a consequence of bleeding or infection, it usually resolves more readily. The treatment of ascites is summarized in Table 4.

Bed rest

Bed rest is probably of no benefit in patients with a preserved renal function, as indicated by a serum creatinine concentration of less than 125 µmol/l and a good initial response to diuretics. However, there is data to suggest that it may be beneficial in those with a poor response to diuretics, but further studies are required.

Dietary salt restriction

There is a general consensus that dietary salt restriction is important in the management of patients with cirrhosis and ascites. However, it is important to maintain an adequate level of nutrition in patients with cirrhosis. Therefore it is generally agreed that sodium intake should be restricted to less than 90 mmol/day, which in effect amounts to a 'no added salt' diet with avoidance of preprepared meals. It is also generally agreed that salt restriction should be an adjunct to diuretic therapy, and that it is rarely effective alone.

Role of water restriction

There are no studies evaluating the role of water restriction on the resolution of ascites. In many studies from the United States and Europe, it has been customary to restrict water intake to between 1 and 1.5 litres/day, a recommendation that has appeared in many major texts for the last 20 to 30 years. It would appear that this treatment has simply crept into current dogma, with no clinical or scientific basis to support this treatment. It is well known that water follows salt, and thus fluid loss will occur with salt restriction or adequate diuresis. In a study of 55 patients with ascites, none of whom had taken any diuretics for 2 weeks, 21 had spontaneous hyponatraemia and 34 were normonatraemic. In all patients with normonatraemia, the free-water clearance was normal. Of those with hyponatraemia, 13 had a marked reduction in free-water clearance and glomerular filtration rate. The remaining 8 hyponatraemic patients had a relatively normal free-water clearance. The patients with hyponatraemia and poor renal function did not respond to diuretic therapy, and had a poor prognosis: 60 per cent inpatient mortality compared with 15 per cent in the remaining patients. Hyponatraemia is caused by excessive water retention, primarily as a result of increased circulating vasopressin levels. Since the hyponatraemia is due to water excess, patients with significant hyponatraemia are usually subject to a water restriction of less than 500 ml/day. The response, however, is slow, and many regard this approach as being relatively ineffective. It is sometimes more prudent to try and improve renal function with volume expansion, in the first instance with colloid.

Thus, patients with ascites who are normonatraemic or mild/moderately hyponatraemic (serum sodium concentration above 125 mmol/l) should not be water-restricted. Water restriction should be reserved for those with severe hyponatraemia in whom the free-water clearance is decreased, after ensuring that their intravascular volume is adequate.

Diuretics

Since they first became available, diuretics have been the mainstay of the treatment of ascites. Diuretic dosage should be increased stepwise if there is an insufficient diuretic response, as defined by a weight loss of less than 1 kg in the first 7 days and/or 2 kg every 7 days thereafter, until the ascites is adequately controlled. The safe upper limit of the rate of weight loss is contentious. However, most experts agree that, in clinical practice, the diuretic dose should be adjusted to achieve a rate of weight loss below an average of 500 g per day in patients without peripheral oedema or 1 kg per day in those with peripheral oedema. Best practice is to add a loop diuretic (furosemide (frusemide) 40 mg/day) once a patient fails to respond to the equivalent of 200 mg spironolactone per day. Many diuretic agents have been evaluated over the years, but in the United Kingdom and Europe this has been mainly confined to spironolactone, amiloride, furosemide (frusemide), and bumetanide.

Diuretic agents

Spironolactone

Spironolactone is an aldosterone antagonist, acting mainly on the distal tubules to increase natriuresis and conserve potassium. In a controlled study comparing the

efficacy of spironolactone in 40 non-azotaemic patients with ascites who were excreting less than 12 mmol of sodium/day, 18 of 19 patients responded to spironolactone alone, whereas only 11 of 21 patients responded to furosemide (frusemide) alone. Most patients responding to spironolactone required 150 mg/day, and a few required 300 mg/day. In all cases diuresis occurred by the third day. For those given furosemide, most responders required 80 mg/day, but a few required 160 mg/day. Furosemide was associated with a decrease in the serum potassium concentration, which necessitated potassium supplementation. In those given spironolactone, serum potassium concentrations increased appreciably. The side-effects of spironolactone include gynaecomastia, hyponatraemia, hyperkalaemia, impotence, menstrual disturbance (although most ascitic patients are amenorrhoeic), and osteomalacia.

Amiloride

Amiloride was first used in 1968, and, combined with either furosemide (frusemide) or ethacrynic acid, resulted in a satisfactory diuresis in most cirrhotic subjects with ascites. In a larger study, Yamada and Reynolds evaluated the efficacy of amiloride in patients with cirrhosis and ascites resistant to bed rest and a very low salt diet (<20 mmol Na/day). When used alone it induced a satisfactory response in 80 per cent of patients at doses of 15 to 30 mg/day.

Furosemide (frusemide)

Furosemide is a loop diuretic that causes a marked natriuresis and diuresis in normal subjects. Its efficacy compared with spironolactone is discussed above. In normal subjects it has a half-life of about 75 min, increasing to about 130 min in patients with cirrhosis. It is generally used as an adjunct to spironolactone treatment and has poor efficacy when used alone in cirrhosis. This is probably because there is salt retention in the distal tubules in the subset of patients with high plasma aldosterone concentrations. Furosemide should be used in a dose not exceeding 160 mg/day; its use is associated with severe electrolyte disturbance, and therefore should be prescribed cautiously.

Complications and benefits of diuretic therapy

Diuretic therapy generally improves morbidity and well being, since it causes resolution of ascites, allows a more liberal diet, decreases portal pressure, and increases the opsonic activity of ascitic fluid thereby decreasing the risk of spontaneous bacterial peritonitis. However, diuretic use is associated with complications in between 10 and 70 per cent of patients. The high incidence of side-effects from diuretics seen earlier is now becoming less common. The main complications of diuretic therapy are shown in Table 5.

Therapeutic paracentesis

Paracentesis has been in use for at least 2000 years, and was widely used in the earlier part of the last century. When diuretics first became available in the 1940s the practice declined but it was still used as an adjunct to therapy until the early 1960s. It gradually fell into disrepute with the recognition that repeated paracentesis resulted in salt depletion and oliguria, and became virtually banned as a treatment. The use of paracentesis re-emerged in the mid-1980s when several controlled clinical studies demonstrated that paracentesis with colloid replacement was safe and associated with fewer complications than diuretic therapy. In a large controlled study, patients with tense ascites were randomized to receive either paracentesis with intravenous albumin (40 g after each paracentesis) or diuretics (spironolactone 200–400 mg/day plus furosemide (frusemide) (40–240 mg/day). Patients with significant renal impairment (serum creatinine concentration > 250 µmol/l) were excluded. Paracentesis (4–6 litres/day) was effective in all patients, and no significant change in electrolytes or renal function was observed. Diuretics were effective in 28 out of 34 patients. There was, however, a significant increase and decrease in serum creatinine and sodium levels, respectively, in the diuretic-treated group, and the duration of inpatient treatment was considerably longer. Total paracentesis does, however, ensure a decrease in blood pressure (Fig. 3) by 7 to 10 mmHg. This research was followed by many other studies evaluating the speed of paracentesis, the haemodynamic changes following paracentesis, and the need for colloid replacement therapy.

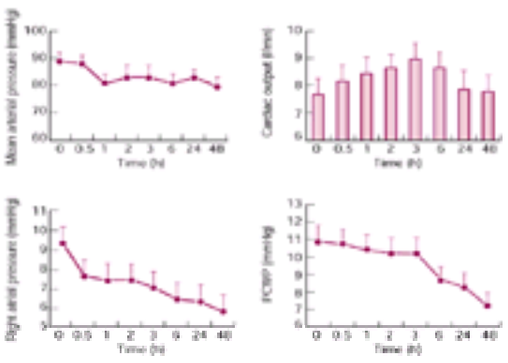

Fig. 3 Haemodynamic changes following acute total paracentesis of approximately 10 litres of ascites over 1 h. Paracentesis was commenced at time 0 h and sequential changes were monitored by a Swan–Ganz catheter, without albumin replacement (modified from Panos *et al*. 1990).

Practical aspects of therapeutic paracentesis

The most important feature of a paracentesis cannula is that it should have multiple side perforations to avoid obstruction by omentum. All ascitic fluid should be drained in a single session as rapidly as possible over 1 to 4 h. If 25 litres of ascites can be drained within 2 to 4 h it is quite safe to do so. The old dogma that rapid paracentesis causes marked hypotension is false, and the haemodynamic changes that occur do so after the removal of as little as 1 litre of ascites. There is a rapid increase in cardiac output, and a corresponding decrease in systemic vascular resistance that peaks at 3 h. There is an immediate fall in right atrial pressure (within 30 min), due to a decrease in intra-abdominal pressure and a decrease in compression of the right atrium. Pulmonary capillary wedge pressure remains constant for 6 h (in the absence of colloid), and decreases after this interval in the absence of colloid replacement. Mean arterial pressure decreases by about 8 mmHg. These changes are shown in Fig. 3. The drainage system should never be left in place overnight since this carries a risk of infection.

Colloid replacement

It is very important that colloid replacement is given following paracentesis to prevent circulatory disturbances. After total paracentesis, synthetic plasma substitutes may be used if the volume of ascites removed is less than 5 litres. However, albumin should be used when more than 5 litres is removed. All or most trials have used albumin at a dose of 8 g/l of ascites removed. There are no data on whether smaller or larger amounts of albumin have differing degrees of efficacy. Based on studies of the haemodynamic changes that follow paracentesis, it is clear that colloid should be given after paracentesis has been completed.

Contraindications to paracentesis

It is generally agreed that there are no contraindications to paracentesis, although studies to date have excluded several subsets of patients, primarily because of inadequate data. In practice, some clinicians have concerns about carrying out paracentesis in patients who have a severe coagulopathy or marked thrombocytopenia in case localized bleeding complications arise, but there are no data to support this view.

Benefits of paracentesis

Paracentesis provides immediate relief from ascites and a tense abdomen. There is a suspicion that paracentesis makes patients more responsive to diuretic therapy. Whilst there is no direct evidence to support this, the observation that plasma arginine vasopressin (**AVP**) levels are proportional to intra-abdominal pressure lends credibility to this idea. Despite the assumption that during paracentesis the observed exacerbation of vasodilatation involves the splanchnic bed, studies have shown that paracentesis causes a significant reduction of portal pressure. A further benefit includes the relief on respiratory muscles. Tense ascites clearly restricts breathing, and increases both the workload of respiration and energy expenditure. Paracentesis provides immediate relief. Likewise, patients who present with fluid overload and ascites can be rapidly relieved by paracentesis. Ascites increases the resting energy expenditure, and this may be improved following paracentesis.

One potential beneficial effect of paracentesis may be to enhance salt and water excretion, due to the acute reductions of renal venous pressure and the increase in renal perfusion that follow. A second and unexpected benefit may relate to water metabolism. Studies by Solis-Herruzo have shown that paracentesis is followed by an acute fall in plasma arginine vasopressin levels—this hormone is implicated in water homeostasis, since it directly affects the water permeability of the collecting tubules and ducts. This acute decrease in plasma concentrations is directly related to intrathoracic or intra-abdominal pressure, since inflation of the abdomen with air to form a pneumoperitoneum, which increases intrathoracic and intra-abdominal pressure, had a similar effect.

Role of albumin infusion

There is a persistent belief that the infusion of albumin is beneficial to patients with cirrhosis. The role of albumin infusion has already been mentioned in the section relating to paracentesis, and it also has a role in the treatment of SBP. The identification of new hepatitis viruses, HIV, and the advent of new-variant Creutzfeldt–Jakob disease should make all clinicians cautious in the administration of human products. During the 1940s, several studies evaluated the effect of fractionated human albumin solution in patients with cirrhosis. These studies demonstrated that the infusion of albumin could correct the low plasma levels observed and result in a modest diuresis in some patients, but overall the results were disappointing.

Role of angiotensin-converting enzyme (ACE) inhibitors

The therapeutic effect of ACE inhibitors is attractive since they directly target the system involved most intimately with salt and water retention. However, the acute administration of either captopril or enalapril may cause an acute fall in blood pressure. Some studies have suggested that in patients with ascites, the chronic administration of enalapril suppressed plasma aldosterone levels and increased urinary sodium excretion and glomerular filtration rate (**GFR**), as well as increasing urinary prostaglandin (**PG**) E_2 and 6-oxo-PGF_{1a}. More studies on the efficacy of these drugs and of angiotensin antagonists on portal pressure and the treatment of ascites are expected. Recent studies have suggested that low doses of ACE inhibitors may enhance salt excretion, but they should be used very carefully.

Treatment of patients with severe liver dysfunction and ascites

Patients with endstage liver disease often have subclinical renal impairment, with a typical GFR of around 60 ml/min. For patients with alcoholic liver disease with or without alcoholic hepatitis, the presence of ascites does not generally affect the clinical outcome, unless the ascites becomes a focus for infection. Up to 90 per cent of patients dying from alcoholic hepatitis develop renal failure. For those with alcoholic hepatitis that resolves spontaneously or with treatment, the ascites usually improves as salt and water excretion increase with improvement of the liver disease. Most patients with severe liver failure have been excluded from clinical studies assessing the efficacy of diuretics or paracentesis. It is recommended that extreme caution is exerted when trying to diurese patients with endstage liver disease. It is probably safer to paracentese such patients rather than give potentially nephrotoxic drugs, although few studies have been conducted in such patients.

Treatment of refractory ascites

Diuretics may be ineffective for a variety of reasons. Approximately 5 to 10 per cent of patients do not respond adequately to diuretics. This may be because the diuretics induce an electrolyte disturbance or encephalopathy, necessitating a temporary and recurrent withdrawal of medication. Alternatively, the patient may be genuinely resistant to the diuretics given. In both these groups, there is invariably significant renal dysfunction when assessed by creatinine clearance or other techniques measuring GFR. Because of confusion over the random use of the term 'refractory ascites' in the literature, in 1994 the International Ascites Club agreed the following definition: 'refractory ascites is defined as ascites that cannot be mobilized or the early recurrence of which cannot be satisfactorily prevented by medical therapy'. It is subdivided into diuretic-resistant ascites and diuretic-intractable ascites (see below); however, it should be noted that the amount of sodium restriction used in the following definitions is now below the recommended sodium restriction:

- *Diuretic-resistant ascites* is defined as an ascites that cannot be mobilized or prevented postparacentesis because of the lack of a response to dietary sodium restriction (<50 mmol/day) and maximal diuretic therapy. Maximal diuretic therapy is defined as spironolactone 400 mg/day together with furosemide (frusemide) 160 mg/day.
- *Diuretic-intractable ascites* is defined as an ascites that cannot be mobilized or prevented postparacentesis because of the development of diuretic-induced complications that preclude the use of an effective or maximal dose of diuretics.

The mainstay of treatment for these patients is repeated paracentesis.

Shunts

Transjugular intrahepatic portosystemic shunts (TIPS) for refractory ascites

Several studies have suggested that TIPS may improve natriuresis in patients with diuretic-resistant ascites. In one study, 50 patients with refractory ascites were treated by TIPS, sufficient to decrease the portal pressure gradient by over 60 per cent. Some 75 per cent of all patients showed complete resolution of their ascites by 3 months, and 20 per cent achieved a partial response. A new onset of hepatic encephalopathy occurred in a further 10 per cent of patients. Other studies, however, have reported a 45 per cent incidence of hepatic encephalopathy post-TIPS, which was severe and disabling in 15 per cent of all treated patients, and this has been confirmed. TIPS was also associated with a decrease in the mean serum creatinine concentration from 133 µmol/l to 80 µmol/l at 6 months. The mean 1-year survival post-TIPS is less than 50 per cent. The effect of TIPS on sodium excretion and renal function is, for some reason, delayed, not being apparent until about 1-month postinsertion. Insertion of a TIPS is associated with a deterioration of liver function, and some patients develop severe haemolysis.

Peritoneovenous shunts (Le Veen shunts or Denver shunts)

Peritoneovenous shunting became very popular in the 1970s, with numerous publications on its benefits to renal function and resolution of ascites. However, it soon became apparent that many shunts became blocked or infected and caused scarring of the peritoneum, which can make liver transplantation difficult. There have been three clinical trials evaluating the efficacy of the peritoneovenous shunt. This technique offers no survival advantage over medical therapy or repeated paracenteses. With respect to its effects on renal function, it has been shown that shunting had no overall effect on mortality in patients with the hepatorenal syndrome.

Prognosis

The occurrence of ascites in patients with cirrhosis is associated with a poor prognosis. Survival rates vary between 50 per cent at 1 to 2 years, but a somewhat better survival in alcoholic patients with ascites who stop drinking. The development of bacterial peritonitis in patients with ascites is associated with a mortality of 75 per cent at 1 year. Thus, the development of this complication is associated with an overall poor prognosis, and, unless contraindicated, all patients should be considered for orthotopic liver transplantation.

Important complications

The complications of ascites are shown in Table 6 and discussed below.

Pleural effusion

Pleural effusions (hepatic hydrothorax) develop in about 5 per cent of patients with cirrhosis. Fluid tracks up into the pleural cavity via defects in the diaphragm (for example, holes or blebs), which occasionally close spontaneously. Hepatic hydrothorax may develop in patients with no discernible ascites. The pleural effusions are right-sided in 85 per cent of cases and bilateral in 2 per cent of cases. To confirm the diagnosis, if doubt exists, a radiotracer should be injected under aseptic conditions into the abdomen and its appearance followed in the pleural fluid. A pleural effusion should be managed as for conventional ascites unless it is unresponsive and causing severe dyspnoea, in which case a TIPS (transjugular intrahepatic portosystemic shunt) should be inserted. TIPS is a highly effective treatment for hepatic hydrothorax.

Paraumbilical hernia

Paraumbilical hernias develop in about 20 per cent of patients with ascites, an incidence that increases up to 70 per cent in those with long-standing recurrent tense ascites. The main risks are rupture and strangulation.

Hepatorenal syndrome

Hepatorenal syndrome is the development of renal failure in patients with advanced liver disease (acute or chronic) in the absence of any pathological cause of renal failure. It is due to a reduction of renal blood flow, an increased renal sympathetic drive, and increased circulating or increased renal production of various vasoactive mediators such as endothelin-1, cysteinyl-leukotrienes, thromboxane A_2, or F(2)-isoprostanes. This syndrome is discussed in detail elsewhere.

Hypercatabolic state

Many patients with ascites present in a hypercatabolic state. This may be secondary to the low-grade endotoxaemia present in many patients, together with their general state of malnutrition. This can be reversed, particularly in alcoholic patients who stop drinking and improve their nutritional lifestyle. However, successful TIPS can also improve the nutritional status of these patients, although this may of course be secondary to their cessation of alcohol intake.

Respiratory difficulties

Increasing abdominal distension due to the accumulation of peritoneal fluid increases the effort required for breathing. Occasionally, this may precipitate extreme difficulty in breathing and should be treated by rapid paracentesis.

Spontaneous bacterial peritonitis (SBP)

The spectrum of bacterial peritonitis includes spontaneous bacterial peritonitis, monomicrobial bacterascites, culture-negative neutroascites, and secondary bacterial peritonitis. Spontaneous bacterial peritonitis is now defined as the combination of a positive ascitic fluid culture, an ascitic fluid neutrophil count of more than 250 cells/mm^3, and no evident intra-abdominal source of infection. Secondary bacterial peritonitis is identical, except that an intra-abdominal source is apparent and the organisms are frequently polymicrobial.

The risk of SBP has been evaluated. Of patients presenting with ascites, about 11 per cent will develop SBP within 1 year and 15 per cent within 3 years. For those with an ascitic protein level below 10 g/l the risk is 24 per cent within 3 years. For patients admitted to hospital with ascites with or without other complications (for example, bleeding) the incidence of SBP on admission, based on a review of several reports, is about 10 per cent.

The symptoms of SBP are shown in Table 7.

The pathogenesis of bacterial peritonitis is shown below in Fig. 4. It is apparent that a source for bacteraemia gives rise to organisms in the hepatic lymph and thence the ascitic fluid. Before an inflammatory reaction occurs, an ascitic tap will yield a positive culture, but a low neutrophil count. This is termed 'monomicrobial bacterascites'. If there is a polymicrobial growth with a low ascitic neutrophil count, then the tap is likely to have been traumatic. This occurs in 0.6 per cent of all ascitic taps. In the absence of any intervention it is estimated that two-thirds of these cases will resolve as a consequence of complement-mediated bacterial lysis: that is to say, it will not develop into SBP but will be resolved by the normal antimicrobial defences of the body. If the organisms multiply and neutrophils are mobilized, the ascitic neutrophil count increases. An ascitic tap at this stage yields a positive culture and an elevated white blood cell count. If the infection resolves at this stage (that is, the organisms are lysed) and an ascitic tap is then performed, the ascitic neutrophil count will be increased but the ascitic fluid will be sterile. This is termed 'culture-negative neutrocytic ascites' (**CNNA**). The most common cause is, however, a poor culture technique. As shown, the opsonic activity is important in determining whether the monomicrobial bacterascites resolves or develops into SBP. The organisms isolated are shown in Table 8, and the risk factors for the development of SBP are summarized in Table 9 and discussed below.

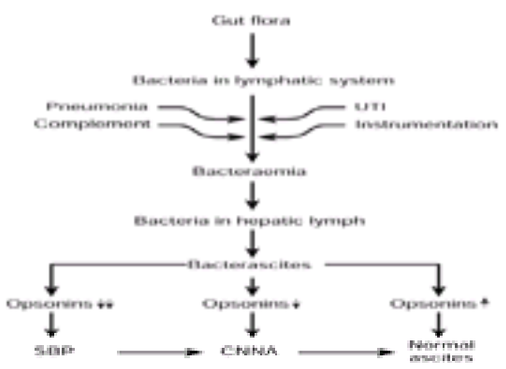

Fig. 4 Pathogenesis of spontaneous bacterial peritonitis. Bacteria can enter ascites through the lymphatic system. In many cases this is resolved through complement mediated bacterial lysis. When opsonins are decreased (e.g. low ascitic protein) or host defence is poor, bacteria multiply and cause spontaneous bacterial peritonitis (SBP). CNNA = culture negative neutrocytic ascites.

Risk factors

Decreased opsonic activity

The protein concentration of ascitic fluid does not change with the advent of SBP. Patients with a low protein content have an increased risk of developing SBP, compared to those with a high ascitic protein content. Patients with cirrhotic and nephrotic ascites are prone to infection, whereas those with malignant ascites or cardiac ascites are not. The risk of SBP is increased sixfold for those with an ascitic protein of less than 10 g/l. When bacteria enter ascitic fluid they may be lysed by the activity of complement if they are serum-sensitive, or they may be coated with opsonins such as IgG or the third component of the complement pathway (C3). Complement deficiency also predisposes to infection. The opsonic activity of ascitic fluid correlates with the total protein, as well as that of CH100 (total haemolytic complement), C3, and C4 concentration. These concentrations may be increased by diuresis, thus decreasing the risk of SBP. Paradoxically, diuretics may also increase the neutrophil count in ascitic fluid.

Recent instrumentation

Recent instrumentation such as endoscopy or sigmoidoscopy increases the risk of SBP. However, the risk of developing SBP is not increased by paracentesis. A study compared the risk of SBP in two groups of patients and found that it was slightly less, if anything, in those treated with paracentesis than in those treated with diuretics.

Gastrointestinal haemorrhage

There is a 21 per cent incidence of SBP in patients admitted to hospital with a gastrointestinal haemorrhage. Whilst it is assumed that bleeding predisposes to infection, there is data to suggest that infection may predispose to bleeding. Since the incidence of SBP immediately following a gastrointestinal bleed is high, many clinicians now advocate the prophylactic use of broad-spectrum antibiotics in such situations.

Previous spontaneous bacterial peritonitis

SBP is a recurrent condition. The recurrence rate for SBP is 47 per cent at 6 months and 69 per cent at 1 year.

Treatment of SBP

The mortality associated with SBP is approximately 30 per cent in most series. Therefore, SBP must be treated as soon as a presumptive diagnosis is made following microscopy of ascitic fluid. For patients with bacterascites but no rise in the neutrophil count, the ascitic tap should be repeated, while those with an increased neutrophil count should be treated. The following therapeutic regimen is used in the United Kingdom, which is based on the types of organism most frequently encountered during SBP: treatment is continued until complete resolution of all signs of infection and the ascitic neutrophil count decreases to within the normal range. This is generally achieved within 1 week of treatment. Appropriate antibiotics include cefotaxime, ciprofloxacin with amoxicillin, and piperacillin with tazlocillin (PIP/TAZ). Recent studies from Barcelona have also shown that the administration of albumin at a dose of 1.5 g/kg at the time of diagnosis and 1 g/kg at 48 h decreases the incidence of renal dysfunction, and improves survival. The treatment of SBP is summarized in Table 10.

Prevention and control

The prevention and control of ascites have already been covered under the treatment of ascites. Currently there is an ongoing National Institutes of Health (**NIH**)-funded multicentre study evaluating the efficacy of b-blockade on the development of portal hypertension, but the results of this are not yet available. However, there have been many studies investigating prophylaxis against SBP.

Prophylaxis against SBP

Many clinicians in the United Kingdom now use ciprofloxacin. A French study has evaluated the efficacy of this antibiotic in patients with cirrhosis who have an ascitic protein level below 15 g/l. This study observed that the administration of 750 mg ciprofloxacin given in a single dose per week decreased the incidence of SBP from 22 per cent to 4 per cent at 6 months, with a corresponding decrease in hospital admission over this period (18 days to 9 days). These authors concluded that prophylaxis with ciprofloxacin was effective, and cost-analysis studies have shown such antimicrobial prophylaxis to be cost-effective.

Special problems in pregnant women

Women with cirrhosis and ascites rarely, if ever, become pregnant, since ovulation has usually ceased before the onset of ascites.

Occupational, quality of life, and psychological aspects

Ascites is significant because it indicates the development of hepatic decompensation, and a corresponding poor prognosis. It also carries a significant morbidity—increasing the workload on the patient, causing backache, etc.—and may be associated with the development of electrolyte disturbance and hepatic encephalopathy. It has enormous cost implications to the National Health Service in terms of repeated hospital admissions, long-term treatment of salt and water retention, and iatrogenic complications. It may also be complicated by the development of spontaneous bacterial peritonitis, which is itself associated with gastrointestinal haemorrhage, and a high mortality.

Areas of controversy

Current areas of controversy include the use of albumin or plasma substitutes following paracentesis. Although albumin has been shown to be superior to plasma substitutes in preventing the activation of hormonal systems that indicate hypovolaemia, there is no hard data to show that albumin is more effective than colloid substitutes in terms of overall patient survival or duration of hospital stay. There is also controversy over the issue of central blood volume, with two groups reporting diametrically opposite findings. This is a crucial argument since the current peripheral vasodilatation hypothesis is based on the premise of a decreased central blood volume. Studies by Mauro Bernardi's group have shown that vasodilatation disappears during upright posture, and it appears that patients may exhibit features of both underfill and overfill depending on their posture and severity of liver disease.

Areas needing further research

Ideally, longitudinal studies conducted over many years need to be performed in newly diagnosed patients with cirrhosis, including baseline clinical, hormonal, and sodium-balance studies, as well as long-term (10 years or more) investigations of systemic haemodynamics and portal pressure. Why do patients develop vasodilatation? Although current ideas favour a role for nitric oxide, the evidence to support this is unclear. Why do patients develop ascites after liver transplantation? This complication is unusual but can be very striking and disruptive to management. Finally, controlled studies are needed to determine whether it is beneficial to paracentese patients with spontaneous bacterial peritonitis.

Further reading

Arroyo V, Ginés P (1992). Arteriolar vasodilatation and the pathogenesis of the hyperdynamic circulation and renal sodium and water retention in cirrhosis. *Gastroenterology* **102**, 1077–8. [This paper reviews the vasodilatation hypothesis of cirrhosis.]

Arroyo V, *et al.* (1976). Prognostic value in spontaneous hyponatremia in cirrhosis with ascites. *Digestive Diseases* **21**, 249–56. [This paper highlights the incidence and prognostic value of serum sodium or free-water clearance in cirrhosis with ascites.]

Bernard B, *et al.* (1995). Prognostic significance of bacterial infection in bleeding cirrhotic patients a prospective study. *Gastroenterology* **108**, 1828–34. [Many patients with variceal haemorrhage have an underlying bacterial infection.]

Bernardi M, *et al.* (1995). Hyperdynamic circulation of advanced cirrhosis: a re-appraisal based on posture-induced changes in haemodynamics. *Journal of Hepatology* **22**, 309–18. [Several papers from Bernardi's group have shown that the vasodilatation is only evident in patients when they are supine, and this affects the renal handling of sodium and glomerular filtration.]

Campra JL, Reynolds TB (1978). Effectiveness of high-dose spironolactone therapy in patients with chronic liver disease and relatively refractory ascites. *Digestive Diseases* **23**, 1025–30. [This paper highlights the use of spironolactone in the management of ascites.]

Dolz C, *et al.* (1991). Ascites increases the resting energy expenditure in liver cirrhosis. *Gastroenterology* **100**, 738–44. [This study demonstrates that patients with ascites have an increased energy expenditure rate which decreases with treatment of ascites.]

Henriksen JH, *et al.* (1989). Reduced central blood volume in cirrhosis. *Gastroenterology* **97**, 1506–13. [This paper demonstrates the reduction of central blood volume in cirrhosis.]

Inadomi J, Sonnenberg A (1997). Cost-analysis of prophylactic antibiotics in spontaneous bacterial peritonitis. *Gastroenterology* **113**, 1289–94. [It is cost-effective to give prophylactic antibiotics to prevent SBP in at-risk individuals.]

Luca A, *et al.* (1994). Favorable effects of total paracentesis on splanchnic haemodynamics in cirrhotic patients with tense ascites. *Hepatology* **20** (Part 1), 30–3. [This study shows that paracentesis is followed by a reduction of portal pressure.]

Nevens F, *et al.* (1996). The effect of long-term treatment with spironolactone on variceal pressure in patients with portal hypertension without ascites. *Hepatology* **23**, 1047–52. [This paper shows that treatment with spironolactone also lowers the portal pressure.]

Panos MZ, *et al.* (1990). Single, total paracentesis for tense ascites: sequential haemodynamic changes and right atrial size. *Hepatology* **11**, 662–7. [This study demonstrated the haemodynamic changes over 48 h following a single, total, large-volume paracentesis.]

Pare P, Talbot J, Hoefs JC (1983). Serum-ascites albumin concentration gradient: a physiologic approach to the differential diagnosis of ascites. *Gastroenterology* **85**, 245–53. [Several papers have shown that measurement of serum-ascites albumin gradient is far superior to measurement of ascitic protein in helping with the differential diagnosis of causes of ascites.]

Perez-Ayuso RM, *et al*. (1983). Randomized comparative study of efficacy of furosemide versus spironolactone in non-azotemic cirrhosis with ascites. *Gastroenterology* **84**, 961–8. [One of the few

controlled trials on the use of diuretics in the management of ascites.]

Rolachon A, *et al.* (1995). Ciprofloxacin and long-term prevention of spontaneous bacterial peritonitis: results of a prospective controlled trial. *Hepatology* **22**, 1171–4. [One of several studies of antibiotic prophylaxis for SBP.]

Runyon BA (1986). Low-protein-concentration ascitic fluid is predisposed to spontaneous bacterial peritonitis. *Gastroenterology* **91**, 1343–6. [Low ascitic protein concentration is a risk factor for the development of SBP.]

Runyon BA (1997). Treatment of patients with cirrhosis and ascites. *Seminars in Liver Disease* **17**, 249–60. [This provides a view from the US on the management of ascites, and extensively reviews the historical and recent literature.]

Runyon BA, Hoefs JC (1984). Culture-negative neutrocytic ascites a variant of spontaneous bacterial peritonitis. *Hepatology* **4**, 1209–11. [This was one of several landmark papers by Runyon that helped us to understand the pathogenesis of SBP.]

Runyon BA, *et al.* (1990). Bedside inoculation of blood culture bottles with ascitic fluid is superior to delayed inoculation in the detection of spontaneous bacterial peritonitis. *Journal of Clinical Microbiology* **28**, 2811–12. [Inoculation of blood culture tubes with ascitic fluid is far superior to conventional methods.]

Salerno F, *et al.* (1991). Randomized comparative study of hemaccel vs. albumin infusion after total paracentesis in cirrhotic patients with refractory ascites. *Hepatology* **13**, 707–13. [This study suggests that plasma expanders are equally effective in the prevention of postparacentesis complications.]

Simón M-A, Díez J, Prieto J (1991). Abnormal sympathetic and renal response to sodium restriction in compensated cirrhosis. *Gastroenterology* **101**, 1354–60. [This study suggests that sodium restriction causes activation of the sympathetic nervous system in non-ascitic patients with cirrhosis.]

Solà R, *et al.* (1995). Spontaneous bacterial peritonitis in cirrhotic patients treated using paracentesis or diuretics results of a randomized study. *Hepatology* **21**, 340–4.

Solis-Herruzo J, *et al.* (1991). Effect of intra-thoracic pressure on plasma arginine vasopressin levels. *Gastroenterology* **101**, 607–17. [This shows that plasma AVP levels decrease following paracentesis.]

Stanley MM, Ochi S, Lee KK, and the Veterans Administration Co-operative Study on Treatment of Alcoholic Cirrhosis with Ascites (1989). Peritoneovenous shunting as compared with medical treatment in patients with alcoholic cirrhosis and massive ascites. *New England Journal of Medicine* **321**, 1632–8. [A controlled trial comparing two therapies.]

Strauss RM, Boyer TD (1997). Hepatic hydrothorax. *Seminars in Liver Disease* **17**, 227–32. [This is a good review of the subject.]

Tító L, *et al.* (1988). Recurrence of spontaneous bacterial peritonitis in cirrhosis frequency and predictive factors. *Hepatology* **8**, 27–31. [SBP is a recurrent disease.]

14.21.3 Hepatocellular failure

E. Anthony Jones

Introduction

Hepatocellular failure is the syndrome that occurs when loss of hepatocytes and/or hypofunction of hepatocytes exceeds the capacity of hepatocytes to regenerate and/or repair hepatocellular injury. Its clinical manifestations include hepatic encephalopathy, a haemorrhagic diathesis, ascites, and hepatocellular jaundice. The syndrome may complicate any disease in which the pathophysiology includes hepatocellular necrosis or apoptosis, or hypofunction of hepatocellular organelles. The duration of evidence of hepatic dysfunction before the onset of hepatocellular failure is variable, ranging from a few days to many years. The term hepatocellular failure does not necessarily imply impaired function of hepatic cells other than hepatocytes. Although many biochemical lesions induced by specific chemical, immunological, or cytopathic hepatotoxic factors have been documented, with the notable exception of hypoxia, the precise mechanisms by which such factors induce hepatocellular failure are poorly understood. Factors that may contribute to hepatocellular injury include immunological damage mediated by cytotoxic T lymphocytes, macrophage activation, direct cytopathic effects of viruses, cytokine-induced activation of cellular interactions, and oxidative stress. An influx of calcium ions into hepatocytes appears to be a late phenomenon in the sequence of biochemical events culminating in hepatocellular necrosis.

Definitions

Acute hepatocellular failure

The syndrome of hepatocellular jaundice, hypertransaminasaemia, and prolongation of the prothrombin time associated with an acute liver disease.

Fulminant hepatic failure

Classically defined as the syndrome of acute hepatocellular failure complicated by hepatic encephalopathy occurring within 8 weeks of the onset of clinical evidence of liver disease. The King's College Hospital (London) group have introduced the terms hyperacute liver failure for the occurrence of encephalopathy within 7 days of the onset of jaundice and late-onset liver failure for the syndrome in which hepatic encephalopathy occurs 8 to 24 weeks after the onset of clinical evidence of liver disease. In addition, the Beaujon Hospital (Paris) group has proposed that the term fulminant hepatic failure be applied to acute liver failure with a plasma factor V level less than 50 per cent of normal and hepatic encephalopathy occurring less than 2 weeks after the onset of jaundice, and that the term subfulminant hepatic failure be used for acute liver failure with a plasma factor V level less than 50 per cent of normal and hepatic encephalopathy occurring 2 weeks to 3 months after the onset of jaundice.

Chronic hepatocellular failure

This is the syndrome of decompensated chronic liver disease, which is chronic hepatocellular disease complicated by hepatic encephalopathy, coagulopathy, ascites, and/or hepatocellular jaundice.

Hepatic encephalopathy (portosystemic encephalopathy)

This is the complex neuropsychiatric syndrome attributable to impaired hepatocellular function and increased portosystemic shunting. The terms hepatic encephalopathy and portosystemic encephalopathy are usually used interchangeably. However, whereas the term portosystemic encephalopathy may be appropriate for encephalopathy complicating increased portosystemic shunting in the absence of overt hepatocellular failure, it may be inappropriate to use the term hepatic encephalopathy in this context.

Aetiology

Acute hepatocellular failure

The most common causes of fulminant hepatic failure are acute viral hepatitis and drugs. About one-third of cases appear to be due to non-A, non-B, non-C hepatitis of undetermined aetiology. Markers of acute infection with specific hepatitis viruses (such as IgM anti-HAV, IgM anti-HBc, IgM anti-HDV) may be useful in suggesting the aetiology. A syndrome similar to acute liver failure with encephalopathy associated with infection by other viruses (such as herpes, varicella) may occur, particularly in immunocompromised patients. Only drugs that can cause acute hepatocellular injury (rather than cholestasis) have the potential of inducing fulminant hepatic failure. Examples are paracetamol, halothane, and antiretoviral drugs. Fulminant hepatic failure caused by poisoning may be due to *Amanita* mushrooms or industrial solvents, particularly chlorinated hydrocarbons. Hypoxic hepatocellular injury may be attributable to reduced hepatic perfusion, but rarely leads to fulminant hepatic failure (for example following cardiac arrest). Important vascular causes of fulminant hepatic failure include the Budd–Chiari syndrome and veno-occlusive disease. The latter may be induced by pyrrolizidine alkaloids, chemotherapy, or irradiation. A rare cause of fulminant hepatic failure is heat stroke. Intravascular haemolysis suggests Wilson's disease. Autoimmune chronic active hepatitis may present with a syndrome similar to subfulminant hepatic failure with type 1 antibodies to liver and kidney microsomes. Fulminant hepatic failure may be precipitated by partial hepatectomy (removal of more than 80 per cent of a normal liver). Fulminant hepatic failure soon after orthotopic liver transplantation may be due to hyperacute allograft rejection or hepatic arterial thrombosis. In carriers of the hepatitis B or C viruses, fulminant hepatic failure may be precipitated by modulation of the host's immune response to the virus as a consequence of

immunosuppressive chemotherapy or its withdrawal. In Reye's syndrome a fulminant hepatic failure-like syndrome may occur, but there are, in addition, mitochondrial changes in the brain which are not specific for liver failure.

Chronic hepatocellular failure

Chronic hepatocellular failure may complicate any progressive chronic hepatocellular disease or any lesion causing chronic hepatic central venous congestion.

Manifestations

Cardinal features

Hepatic encephalopathy

Impaired mental function in liver failure may lead to a wide spectrum of psychiatric and neurological changes. Impaired psychometric test results and/or abnormal brain electrophysiological function in a patient with chronic liver disease, in whom a routine neurological examination is normal, may imply subclinical hepatic encephalopathy. The earliest clinical signs are psychiatric and behavioural changes. These changes are primarily due to subtle impairment of intellectual function that reflects predominantly bilateral forebrain dysfunction. Conventionally, four clinical stages of hepatic encephalopathy are recognized (Table 1). Increased muscle tone with cogwheel and neck rigidity, and myoclonic twitching may occur. Asterixis ('liver flap') can often be elicited (Table 1, Fig. 1). The mouth may be difficult to open. With progression, deep tendon reflexes may be increased and subsequently decreased. One or both plantar responses may be extensor. With progression, the frequency of the electroencephalogram decreases and its amplitude increases. With further progression, the amplitude decreases and triphasic waves may occur. Both the clinical and electrophysiological manifestations of hepatic encephalopathy are non-specific. Hepatic encephalopathy complicating chronic liver disease may be acute or chronic. When acute it is usually associated with one or more recognized precipitating factors (Table 2). With the notable exception of sedative hypnotic drugs, the mechanisms by which common precipitating factors exacerbate encephalopathy are poorly understood. Failure to identify a precipitating factor may imply deterioration of hepatocellular function. The term chronic portosystemic encephalopathy is often preferred when hepatic encephalopathy complicating chronic liver disease is persistent or episodic.

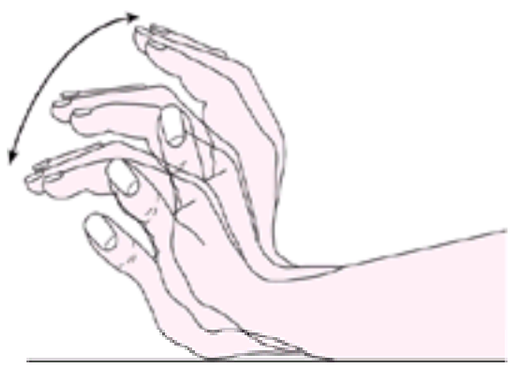

Fig. 1 The 'liver flap' is a slow, flapping tremor (one flap every 1 to 2 s), which can be elicited by asking the patient to dorsiflex the hands with the arms outstretched and the fingers extended and parted. It is due to neuromuscular incoordination between flexor and extensor muscles (negative myoclonus). The hands tend to fall forward, but this involuntary movement is rapidly corrected by readoption of the dorsiflexed position, thereby creating a 'flap'. The same phenomenon may be elicited by asking the patient to squeeze the physician's extended finger. Neuromuscular incoordination is indicated by repeated intensification and relaxation of the intensity of the squeeze ('milkmaid's grip').

Hepatic encephalopathy is considered to be a reversible metabolic encephalopathy with a multifactorial pathogenesis. Traditionally, gut factors have been considered to play important roles (Table 2 and Table 3). In liver failure there is decreased hepatic extraction and metabolism of constituents of portal venous plasma and decreased exposure of hepatocytes to these constituents as a consequence of their passage through intrahepatic and extrahepatic portosystemic venous collateral channels. Consequently, constituents of portal venous plasma tend to accumulate in the systemic circulation. If some of these compounds are neuroactive and can cross the blood–brain barrier, modulation of brain function may occur. The blood–brain barrier is normally highly permeable to non-polar substances, such as non-ionic ammonia and benzodiazepines, but has a low permeability to polar compounds. However, in liver failure the permeability of the barrier to polar compounds, such as the inhibitory neurotransmitter GABA, may increase. Most of the manifestations of hepatic encephalopathy appear to be attributable to a global suppression of central nervous system function, due predominantly to a net increase in inhibitory neurotransmission, as a consequence of increased neurotransmission mediated by inhibitory neurotransmitters (such as GABA) and/or decreased neurotransmission mediated by excitatory neurotransmitters (such as glutamate). Currently, the two factors considered to be most important in pathogenesis are raised brain concentrations of ammonia and increased GABA-mediated neurotransmission.

Increased GABAergic neurotransmission is associated with impaired motor function and decreased consciousness, two of the cardinal manifestations of hepatic encephalopathy. Potential mechanisms for increased GABAergic tone in hepatic encephalopathy include: (i) increased availability of GABA at $GABA_A$ receptors in synaptic clefts; (ii) increased astrocytic synthesis and release of neurosteroids that are agonists of the $GABA_A$ receptor; and (iii) increased brain concentrations of natural benzodiazepine receptor agonist ligands.

Ammonia was originally implicated in pathogenesis because it was recognized to be neurotoxic, plasma concentrations tend to be raised in liver failure, and lasma ammonia readily enters the brain. Plasma ammonia concentrations higher than those usually found in liver failure (more than 1 mmol/l) are associated with increased neuronal excitation and seizures. In contrast, plasma ammonia concentrations typically found in patients with precoma stages of hepatic encephalopathy (stages I to III) (100 to 400 µmol/l) may enhance neuronal inhibition by: (i) directly facilitating GABA-gated chloride conductance; (ii) selectively increasing the binding of agonist ligands to the $GABA_A$/benzodiazepine receptor complex; and (iii) stimulating astrocytic synthesis and release of neurosteroids that are potent $GABA_A$ receptor agonists.

Possible roles for neurotransmitter systems, other than the GABA system, have been postulated. Some of the symptomatology of hepatic encephalopathy can be explained by disturbances in functional loops of basal ganglia, which could arise as a consequence of an imbalance between glutamatergic and GABAergic neurotransmission.

Haemorrhagic diathesis

The basis of the haemorrhagic diathesis is multifactorial. Of major importance is impaired synthesis of hepatocyte-derived blood clotting factors; this leads to prolongation of the prothrombin time (see Chapter 22.5.1), which is not corrected by parenteral vitamin K. Thrombocytopenia is often present; it may be secondary to the hypersplenism of portal hypertension (see Chapter 14.21.2). However, in fulminant hepatic failure, platelet structure and function are abnormal and the capillary bleeding time is greater than that predicted from the platelet count. Mild disseminated intravascular coagulation is often detectable, but is rarely of clinical significance. Upper gastrointestinal haemorrhage (for example from gastritis, gastro-oesophageal varices, or ulcers) frequently occurs. A common clinical manifestation of the bleeding tendency is bruising around venepuncture sites.

Ascites

Ascites due to hepatocellular failure complicates lesions that cause sinusoidal portal hypertension (such as cirrhosis) or impaired hepatic venous drainage. However, hepatocellular failure is not invariable when ascites is associated with hepatic venous congestion (see Chapter 14.21.2).

Hepatocellular jaundice

The jaundice of hepatocellular failure has an orange tint and is attributable to conjugated hyperbilirubinaemia due to impaired secretion of conjugated bilirubin into the bile canaliculus; the transport maximum for conjugated bilirubin across the bile canaliculus is reduced relative to bilirubin production and conjugation (see Chapter 14.19.3). In acute hepatitis the degree of conjugated hyperbilirubinaemia reflects the extent of hepatocellular necrosis, but even when jaundice is deep, other features of hepatocellular failure (such as a prolonged prothrombin time) are often absent, reflecting the large normal hepatic reserve. In contrast, in chronic non-cholestatic liver disease, hepatocellular jaundice usually reflects severe hepatocellular failure.

Other features

Increased susceptibility to infection

About 80 per cent of infections are bacterial, but about one-third are complicated by tissue invasion by fungi (such as aspergillosis, candidiasis). There may be no fever or leucocytosis. The mortality is high. Fungal infection is suggested by antibiotic-resistant fever. The increased frequency of infections may be related to reduced levels of complement components and opsonins, reduced phagocytic and bactericidal properties of polymorphonuclear leucocytes, and reduced clearance function of Kupffer cells. Spontaneous bacterial peritonitis is a common complication of ascites (see Chapter 14.21.2).

Fetor hepaticus

Fetor hepaticus is the term applied to a particular smell of the breath that commonly occurs in patients with cirrhosis and extensive portosystemic shunts or fulminant hepatic failure. Descriptions vary and include a sweetish, slightly pungent, or faecal smell, similar to that of a rotten apple, mice, or a freshly opened corpse. Being subjective there is considerable variation in its recognition. It has been attributed to gut-derived, sulphur-containing products of methionine metabolism.

Acid–base and electrolyte changes

A wide range of abnormalities occur, particularly in fulminant hepatic failure, and may contribute to altered neurological and cardiac function. Hyponatraemia may be due to impaired free-water clearance, failure of the sodium pump, or diuretics. Hypernatraemia is usually iatrogenic. Respiratory alkalosis, secondary to hyperventilation of central origin, is common in fulminant hepatic failure. Loop diuretics often precipitate a hypokalaemic metabolic alkalosis. Metabolic acidosis may be associated with extensive tissue damage, hypoxia, and lactic acidaemia. Respiratory acidosis may be associated with hypercapnia and respiratory infection.

Cerebral oedema and raised intracranial pressure

Cerebral oedema and raised intracranial pressure frequently complicate fulminant hepatic failure, occurring in about 80 per cent of patients with stage IV encephalopathy, but are uncommon in patients with chronic hepatocellular failure. These complications may be classified separately from hepatic encephalopathy. Herniation of the cingulate, uncus, or cerebellar tonsil secondary to raised intracranial pressure is a frequent cause of death in fulminant hepatic failure. Antemortem diagnosis of cerebral oedema and raised intracranial pressure is suggested by sudden deterioration of consciousness, increased muscle tone, unequal pupils, abnormally reacting pupils, myoclonus, focal seizures, decerebrate posturing, fixed pupils with spontaneous respiration, and/or absent ciliospinal reflexes. Sudden changes in pulse and blood pressure unrelated to haemorrhage, rapid deterioration of the electroencephalogram, sweating, tachycardia, arrhythmias, intermittent systemic hypertension, sudden severe hypotension, bursts of hyperventilation, and fever may all be manifestations of raised intracranial pressure. Papilloedema is rare. Signs of raised intracranial pressure become apparent when intracranial pressure exceeds 30 mmHg. A failure of cellular osmoregulation, with intracellular accumulation of osmolytes, such as glutamine, appears to be a pathogenic mechamism (cytotoxic). Compensatory loss of other intracellular osmolytes, such as inositol, may be more effective in chronic liver disease than in fulminant hepatic failure. Increased blood-to-brain transfer of fluid across the blood–brain barrier (vasogenic), and expansion of the extravascular space (interstitial or hydrocephalic) may also contribute to pathogenesis.

Hypoglycaemia

Severe hypoglycaemia (blood glucose less than 40 mg/dl) occurs in about 40 per cent of patients with fulminant hepatic failure (particularly children) and may exacerbate encephalopathy. The clinical and electroencephalographic features of hepatic and hypoglycaemic encephalopathies are similar. In acute liver failure, hypoglycaemia may occur in the absence of hepatic encephalopathy. Hypoglycaemia may develop rapidly and may recur with sepsis. It is due primarily to impaired hepatic glucose release secondary to glycogen depletion. In contrast to hepatic encephalopathy, hypoglycaemic coma may cause irreversible brain damage.

Cardiovascular changes

Hepatocellular failure is associated with systemic vasodilation and a hyperdynamic circulation. Cardiac output is increased, peripheral vascular resistance decreased, blood pressure reduced, and splanchnic and capillary flow increased, but perfusion of the renal cortex is decreased. Features of a hyperdynamic circulation include a bounding pulse, capillary pulsation, vasodilated extremities, a precordial heave, and an ejection systolic murmur. The increased cardiac output has been attributed to an increased vascular capacitance and hence relative hypovolaemia with low jugular venous pressure.

Recently, endogenous cannabinoids acting at vascular CB_1 receptors have been implicated in this state of vasodilation. Arrhythmias, other than sinus tachycardia, frequently occur with hypoxia and stage IV encephalopathy due to fulminant hepatic failure. Cardiac arrest (unrelated to respiratory arrest) may occur.

Hepatorenal syndrome

Renal failure is common and may be rapidly progressive. In only a minority of cases is it attributable to hypovolaemia or a lesion of the urinary tract. It is typically functional and characterized by reduced glomerular filtration rate and oliguria. Acute tubular necrosis may supervene. Absorption of large quantities of nitrogenous substances from the gut after a gastrointestinal haemorrhage may contribute to azotaemia. Plasma urea and creatinine are not reliable indices of renal function in fulminant hepatic failure; hepatic synthesis of urea is reduced and tubular secretion of creatinine is increased. Functional renal failure is associated with intense renal arterial vasoconstriction. The kidneys in this syndrome function normally when transplanted into subjects without liver disease. Several humoral systems have been implicated in pathogenesis (see Chapter 14.20.2).

Hepatopulmonary syndrome

This syndrome is defined as the triad of liver disease, intrapulmonary peripheral vascular dilatation with decreased pulmonary vascular resistance (right to left shunt), and an increased alveolar–arterial oxygen gradient. Hypoxaemia (PaO_2 less than 70 mmHg) is common and may be associated with cyanosis. The hypoxaemia is usually reversed by 100 per cent oxygen and is attributable to abnormal ventilation–perfusion ratios and impaired diffusion capacity, but uncommonly in cirrhosis hypoxaemia, not reversible by 100 per cent oxygen, may be due to large pulmonary arteriovenous shunts. Portopulmonary shunting and pulmonary hypertension may develop. Chest radiographs may show a high diaphragm, basal pulmonary infiltrates, or pulmonary oedema. Pulmonary oedema, not attributable to left ventricular failure, occurs in fulminant hepatic failure. Respiratory arrest of central origin may occur. The mechanism of the pulmonary vasodilation is unknown. Oxygen does not diffuse readily into the centre of dilated vessels and increased cardiac output limits the time for gas exchange.

Skin changes

Recognition of certain skin changes in a patient with chronic liver disease alerts the clinician to the possibility of incipient or overt chronic hepatocellular failure. However, no skin changes are specific for hepatocellular failure. Spider naevi are often present in patients with cirrhosis. They consist of a central protuberant arteriole from which small vessels radiate in a manner that has been likened to the appearance of a spider's legs (Fig. 2 and Plate 1). Their diameter is usually less than 0.5 cm. They occur in the area of drainage of the superior vena cava and should be distinguished from telangiectasia, corkscrew scleral vessels, and purpura. Development of new 'spiders' suggests progressive hepatocellular disease. Palmar erythema occurs less frequently than spider naevi. It is characterized by an exaggeration of the normal mottling of palmar surfaces of the hands, resulting in well-demarcated redness of the thenar and hypothenar eminences, and of the pulps of the fingers (Fig. 3 and Plate 2). Dilated, thread-like blood vessels in the skin, having an apparently random distribution, may occur, and may resemble a United States dollar note ('paper money' skin). White nails with loss of demarcation of the lunulae (leuconychia, Terry's nails) (Fig. 4 and Plate 3) and finger clubbing

(Lovibond's angle greater than 180°) (Fig. 5 and Plate 4) may also occur.

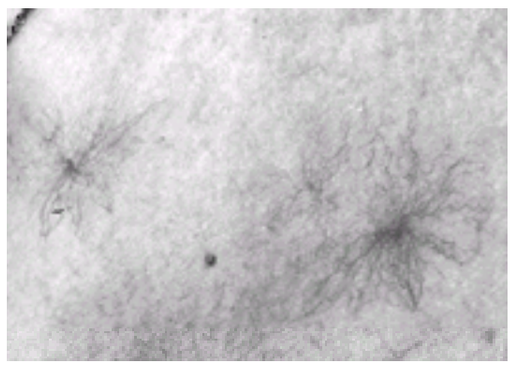

Fig. 2 Spider naevi in a patient with cirrhosis. (See also Plate 1.)

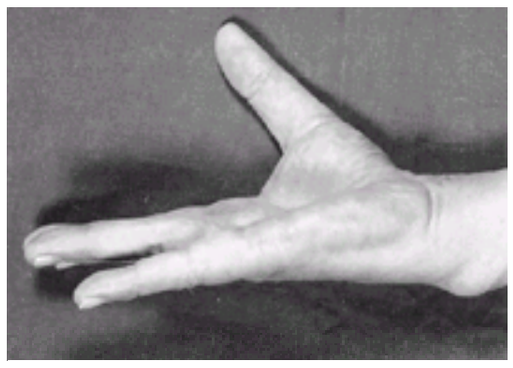

Fig. 3 Palmer erythema in a patient with cirrhosis. (See also Plate 2.)

Fig. 4 White nails in a patient with cirrhosis. (See also Plate 3.)

Fig. 5 Finger clubbing in a patient with cirrhosis. (See also Plate 4.)

Endocrine changes

Chronic liver disease may be associated with reduced concentrations of testosterone. Some male patients with cirrhosis develop hypogonadism and feminization. The former is characterized by testicular atrophy, decreased potency and libido, and a reduced need to shave; the latter is characterized by gynaecomastia and female hair distribution and body habitus. Some female patients with cirrhosis develop infertility, scanty irregular menstruation, and an asexual appearance due to loss of female characteristics. Unilateral or bilateral (tender) gynaecomastia may be a complication of cirrhosis (Fig. 6) or spironolactone therapy.

Fig. 6 Gynaecomastia in a male patient with cirrhosis.

Fatigue

Severe disabling fatigue that seems to be out of proportion to a patient's general condition may occur in chronic liver disease, before the onset of overt hepatocellular failure.

Abnormal protein metabolism

In cirrhosis the degree of hypoalbuminaemia reflects both decreased hepatic synthesis and an increase in plasma volume. Because of albumin's long plasma half-life, hypoalbuminaemia may not be present early in the course of fulminant hepatic failure. Chronic hepatocellular failure is associated with increased protein catabolism and a loss of skeletal muscle mass.

Fever

A low-grade fever may occur with severe active hepatocellular disease (such as acute alcoholic hepatitis) in the absence of infection.

Anaemia

A normochromic normocytic anaemia is a feature of chronic hepatocellular failure.

Osteopenia

Osteopenia is common in patients with decompensated cirrhosis.

Diagnosis

The syndrome of hepatocellular failure constitutes a clinical spectrum from acute liver failure at one extreme to decompensated chronic hepatocellular disease at the other. A patient dying of hepatocellular failure usually exhibits all four of the cardinal manifestations with or without complicating sepsis.

Hepatic encephalopathy

This is a clinical diagnosis that is usually made by recognizing the presence of encephalopathy and excluding non-hepatic causes. No individual clinical or laboratory abnormality is specific for hepatic encephalopathy. Special attention is paid to changes in personality, hypersomnia, and deterioration of performance at work or school. Asterixis (liver flap) is not pathognomonic. Signs of portal hypertension, non-specific cutaneous stigmata of liver disease, and/or fetor hepaticus may be present. It is necessary to recognize disorders with neurological manifestations that may mimic hepatic encephalopathy, such as Wernicke's encephalopathy, alcohol intoxication, or subdural haematoma. More than one type of encephalopathy may coexist. Psychometric tests are useful in detecting and monitoring subtle mental dysfunction in patients with subclinical or prestupor stages of hepatic encephalopathy. The quantitiative number connection test is frequently applied, but allowance must be made for the effects of learning and age on test scores. Electroencephalographic abnormalities are non-specific. There is usualy a fairly good correlation between the clinical stage of encephalopathy and the degree of abnormality of the electroencephalogram. The electroencephalogram is of value in differential diagnosis; it can reveal focal lesions in the brain, seizure activity, and other findings that might suggest an alternative diagnosis. Visual event-related potentials that depend on cognitive function, such as P300 potentials, are sensitive in the detection of electrophysiological changes in the brain in patients with cirrhosis who do not have overt encephalopathy. Routine laboratory tests aid in the differential diagnosis of encephalopathies, and in the detection of factors that may precipitate hepatic encephalopathy (Table 2). Plasma ammonia concentrations are modestly increased in the majority of patients with hepatic encephalopathy, but correlate poorly with the clinical stage and are not useful in management. An elevated plasma ammonia may be helpful in suggesting a hepatic origin for an undiagnosed encephalopathy.

Haemorrhagic diathesis

The most important readily obtainable laboratory markers of this diathesis are the prothrombin time and the platelet count. Plasma activities of individual clotting factors that are synthesized in the liver are reduced (see Chapter 22.5.1).

Ascites

When the presence of ascites on physical examination is in doubt, the issue may be resolved by ultrasonography of the abdomen, which can detect as little as 100 to 200 ml of intraperitoneal fluid. Careful examination of the jugular veins is necessary in the exclusion of cardiac causes. On ultrasonography diffuse inhomogeneity of the liver suggests cirrhosis, and difficulty in visualizing major hepatic veins suggests the Budd–Chiari syndrome. A small diagnostic ascitic fluid tap is done routinely; analysis of the fluid includes determination of concentrations of leucocytes and protein, examination for malignant cells, and culture.

Hepatocellular jaundice

The conjugated hyperbilirubinaemia of hepatocellular failure has to be distinguished from acquired intrahepatic cholestatic disease, cholestasis due to large duct biliary obstruction, and rare congenital hyperbilirubinaemias in which other routine serum biochemical liver tests are normal. Recognition that conjugated hyperbilirubinaemia is attributable to hepatocellular failure is usually possible from clinical and routine haematological and serum biochemical data and an ultrasound showing no evidence of dilated bile ducts. Unconjugated hyperbilirubinaemia is not a feature of hepatocellular failure.

Acute hepatocellular failure

Acute hepatocellular disease associated with a conjugated hyperbilirubinaemia may be classified as acute liver failure when prolongation of the prothrombin time occurs. If acute liver failure is due to hypoxia, a cause is usually obvious, such as a hypotensive episode during surgery.

The diagnosis of fulminant hepatic failure requires the presence of encephalopathy, elevated serum alanine aminotransferase levels early in the course, and marked prolongation of the prothrombin time. Serum alanine aminotransferase concentrations exceeding 50 times the upper limit of normal are common in massive hepatocellular necrosis, but may be less than three times the upper limit of normal with minimal hyperbilirubinaemia when fulminant hepatic failure is associated with microvesicular hepatic steatosis (see Pathology). Abdominal pain may occur with poisoning. Rarely, hepatic encephalopathy precedes jaundice and abnormal behaviour may have to be distinguished from non-hepatogenous acute psychiatric disease. However, patients with fulminant hepatic failure due to massive hepatocellular necrosis, who survive more than a few days, develop deep jaundice. Lumbar puncture should usually be avoided, because of the coagulopathy and possible raised intracranial pressure. However, a baseline CT scan of the brain may be useful. Evidence for the presence of other types of encephalopathy is routinely sought. The syndrome of fulminant hepatic failure may occasionally be mimicked by severe sepsis or falciparum malaria. In subfulminant hepatic failure, ultrasonography may reveal inhomogeneity of the liver due to nodular transformation.

Chronic hepatocellular failure

The diagnosis of chronic hepatocellular failure requires the demonstration of an appropriate chronic liver disease and evidence of hepatocellular failure. Mild conjugated hyperbilirubinaemia and a modest prolongation of the prothrombin time tend to occur before the development of overt hepatic encephalopathy or ascites. In contrast to diseases that lead to sinusoidal portal hypertension (such as cirrhosis), those that cause presinusoidal portal hypertension (such as schistosomiasis) do not usually progress to hepatocellular failure.

Pathology

There is no single hepatic histological change that is pathognomonic of hepatocellular failure. Fulminant hepatic failure is usually associated with massive or

confluent hepatocellular necrosis. However, occasionally, when due, for example, to acute fatty liver of pregnancy, or hepatotoxicity caused by intravenous tetracycline, valproic acid, or antiretroviral drugs, liver histology reveals microvesicular hepatocellular steatosis, in which the nucleus retains its central location within hepatocytes. In an appreciable proportion of autopsies on patients who succumb to fulminant hepatic failure, there is evidence of cerebral oedema and raised intracranial pressure, such as increased brain weight, tense dura, flattened cortical gyri, dilated ventricles, and cingulate, uncal, or cerebellar herniation. The histological appearances of the brain in fulminant hepatic failure are essentially normal. In contrast, histology of the brain of patients who died from chronic liver failure typically shows an increase in the number and size of Alzheimer type 2 astrocytes. Functional renal failure is associated with no gross pathological changes in the kidney.

Course and prognosis

Acute hepatocellular failure

In patients with acute liver failure, who do not develop encephalopathy, such as the typical case of acute icteric viral hepatitis, complete recovery is the rule.

The course of fulminant hepatic failure is variable. There are no reliable criteria that enable prediction of whether an individual patient will die or regain consciousness and ultimately survive. Overall survival appears to have improved with advances in intensive supportive care, and may currently be about 40 per cent without liver transplantation. Mortality tends to be greater when the encephalopathy is severe and prolonged, and when coagulopathy is profound (prothrombin time greater than 100 s). Mortality also tends to be greater if the patient's age is below 5 or over 40 years, or if encephalopathy occurs more than 8 days after the onset of jaundice. The mortality is particularly high (more than 80 per cent) in cases caused by halothane, or drugs other than paracetamol, but is about 50 per cent when paracetamol is implicated. Major complications increase mortality. Small or decreasing liver size, convulsions, cardiac arrhythmias (other than sinus tachycardia), and marked fetor hepaticus are ominous signs. Serum concentrations of aminotransferases may decrease abruptly, but have no prognostic value. The course of fulminant hepatic failure can be divided into five phases.

Pre-encephalopathy

In acute liver failure a progressive increase in prothrombin time is ominous and often precedes the onset of hepatic encephalopathy. After paracetamol overdosage the onset of encephalopathy may be predicted from plasma concentrations of the drug.

Encephalopathy

About one-third of patients die within 2 days of the onset of stage IV encephalopathy. In about 20 per cent of cases, death appears to be due to acute liver failure with progressive encephalopathy. In other cases, death can be attributed to one or more complications of the syndrome, such as upper gastrointestinal haemorrhage, cerebral oedema and raised intracranial pressure, sepsis, and renal failure. In subfulminant hepatic failure, death due to raised intracranial pressure and/or sepsis is more common than in fulminant hepatic failure.

Hepatic regeneration

The key factor in determining the outcome of fulminant hepatic failure, in the absence of liver transplantation, is the ability of the liver to regenerate. Nodules of hyperplastic regenerating liver tissue may be found at autopsy in patients who survive more than 10 days after the onset of encephalopathy. In general, such patients have usually died of a complication of fulminant hepatic failure at a time when indices of hepatocellular function were improving. Serum concentrations of a-fetoprotein, which are regarded as an index of hepatic regeneration, do not usually become elevated until at least 10 days after the onset of encephalopathy. The concentrations tend to correlate fairly well with the amount of hepatic regeneration found at autopsy. Recovery is usually heralded by clinical improvement in encephalopathy, which may be preceded by a decreasing prothrombin time. The electroencephalogram may remain abnormal for several days after consciousness is regained.

Cholestasis

A phase of profound cholestasis often develops 2 to 3 weeks after patients regain consciousness. When death has occurred during this phase, large regenerative nodules and intense cholestasis in hepatocytes have been found at autopsy.

Long-term sequelae

Complete restoration of normal hepatic function and structure usually occurs in survivors of fulminant hepatic failure, even after cerebral oedema, decerebrate rigidity, and episodes of flattening of the electroencephalogram. Serum biochemical liver tests and hepatic histology typically return to normal 45 to 75 days after the onset of hepatic encephalopathy. Permanent neurological sequelae have been reported when recovery has occurred after respiratory arrest.

Chronic hepatocellular failure

In patients with chronic hepatocellular disease, hepatic encephalopathy is often reversible, particularly if a precipitating factor is identified. An MRI of the brain typically reveals symmetric pallidal hyperintensities, possibly due to increased deposition of manganese in the basal ganglia. These hyperintensities appear to correlate with the degree of impairment of hepatocellular function, but not with hepatic encephalopathy. Elevated serum conjugated bilirubin in cirrhosis or precirrhotic alcoholic liver disease is associated with a poor prognosis. An increasing serum conjugated bilirubin in a patient with a chronic cholestatic liver disease may reflect progression of the disease and/or the development of hepatocellular failure. The serum bilirubin is regarded as a good index of prognosis in primary biliary cirrhosis. When ascites first develops in a patient with cirrhosis, 1-year survival is about 50 per cent and 5-year survival about 20 per cent. Survival after the onset of the hepatorenal syndrome is usually only a few weeks or months.

Management

The first issue is whether there is any effective therapy for the underlying liver disease. A treatment that suppresses the pathological process responsible for impairing hepatocellular function may decrease or reverse manifestations of hepatocellular failure. For acute liver failure, corticosteroids are ineffective and may be harmful, except in uncommon patients in whom the underlying lesion is an autoimmune hepatitis, and in carriers of the hepatitis B or C virus in whom acute liver failure has been precipitated by the withdrawal of immunosuppressive chemotherapy. Antiviral therapy has not been shown to be efficacious for acute viral hepatitis. Acetylcysteine has been shown to improve survival in patients who have taken an overdose of paracetamol, and this may apply even when the antidote is given after hepatic encephalopathy has developed. When viral infections, other than viral hepatitis, are diagnosed, antiviral treatment is instituted. Interruption of pregnancy has been advocated to improve survival in patients with fulminant hepatic failure due to acute fatty liver of pregnancy. It is useful to consider management of acute and chronic hepatocellular failure separately. The chronic syndrome accounts for the great majority of cases of hepatocellular failure, whereas the acute syndrome, when severe, is one of the most challenging in clinical medicine, presenting the physician with a unique constellation of difficult problems. In addition to discontinuing drugs that might have contributed to the clinical condition, especially neuroactive, hepatotoxic, and nephrotoxic drugs, it is necessary to take into account hepatocellular disease-associated alterations in drug pharmacokinetics and pharmacodynamics when prescribing for the patient in hepatocellular failure. Drugs may modify the clinical manifestations of hepatocellular failure. For example, patients with cirrhosis exibit increased sensitivity to the central neuroinhibitory and muscle relaxant effects of benzodiazepines. Whether liver transplantation is an appropriate therapeutic option must be considered in all patients with hepatocellular failure (see Chapter 14.21.4).

Chronic hepatocellular failure

As the patient has irreversible architectural changes in the liver, there is no potential for complete recovery with medical treatment. In such cases, management consists of trying to reduce the manifestations of hepatocellular failure that are amenable to treatment, especially encephalopathy and ascites; optimizing nutritional status with a high protein diet, if tolerated; treating complicating infections; and assessing suitability for liver transplantation. Non-specific clinical deterioration raises the possibility of bacteraemia, spontaneous bacterial peritonitis, or hepatocellular carcinoma.

Hepatic encephalopathy

The following general principles are relevant in the management of hepatic encephalopathy: (i) removal or correction of any precipitating factor (Table 2); (ii) reduction of absorption of nitrogenous substances from the gut; (iii) reduction of increased portosystemic shunting; and (iv) reversal of contributing neuropathophysiological mechanisms with drugs that act directly on the brain (Table 3).

Acute hepatic encephalopathy

All drugs that might contribute to encephalopathy, including diuretics, are stopped, and consideration is given to administering an appropriate antidote, such an naloxone or flumazenil. Meticulous attention is paid to maintaining fluid and electrolyte balance, and an adequate urine flow. Dietary protein intake is restricted, and enemas, such as magnesium sulphate or phosphate, are given. Lactulose, or another disaccharide with similar properties, such as lactitol, is given routinely. There is no disaccharidase on the microvillus membrane of enterocytes that hydrolyses lactulose. Its metabolism by colonic bacteria leads to production of lactic acid and other organic acids, a fall in colonic pH, and increased ionization of nitrogenous compounds. These changes may lead to a decrease in the absorption of nitrogenous compounds, including ammonia. Lactulose is a cathartic and is widely believed to be efficacious in the management of hepatic encephalopathy. It may induce hypernatraemia due to increased faecal fluid loss. In addition, an enterically administered, broad-spectrum, poorly absorbed antibiotic may be given to reduce the enteric bacterial flora. Neomycin (up to 6 g daily) has been most extensively used; potent alternatives include kanamycin and paramomycin. Metronidazole, which is effective against anaerobes, may also be given. If improvement in consciousness occurs, dietary protein is increased incrementally.

Chronic portosystemic encephalopathy

In the absence of protein intolerance a nutritious diet that includes a high protein content (80 to 100 g/day) is encouraged to maintain a positive nitrogen balance and optimize liver function. Vitamins are given empirically and thiamine replacement may be indicated in malnourished patients who are alcoholic. Vegetable protein diets seem to be well tolerated and tend to be cathartic due to their fibre content. Oral branched-chain amino acids may decrease protein catabolism and facilitate maintenance of a positive nitrogen balance. When protein intolerance develops, management consists of reducing dietary protein intake to as low as 40 g/day. Lactulose or lactitol is given in doses sufficient to produce two or three semiformed bowel actions daily, and precipitating factors are carefully avoided. If disaccharide intolerance develops, a broad-spectrum antibiotic may be tried. However, long-term neomycin should be avoided because of the risk of ototoxicity and nephrotoxicity. Metronidazole may induce a peripheral neuropathy.

If intractable chronic portosystemic encephalopathy occurs in a patient with a large spontaneous or surgically-induced portosystemic shunt, the invasive technique of balloon occlusion, coupled with embolization of a collateral vein, may reverse portal blood flow from hepatofugal to hepatopetal, improve hepatocellular function, and ameliorate the encephalopathy. Similarly, in a patient with chronic portosystemic encephalopathy and a patent transjugular intrahepatic portosystemic stent (TIPSS) the shunt can be narrowed or closed.

A new approach

A new therapeutic approach is to give a drug that acts on the target organ of hepatic encephalopathy, the brain, by reversing contributory neuropathophysiological mechanisms. The benzodiazepine antagonist, flumazenil, is the first promising drug of this type. It competes with high specificity with other benzodiazepine receptor ligands for binding to central benzodiazepine receptors, and rapidly and completely reverses the sedative and other neurological effects of benzodiazepine agonists, such as diazepam. When given as a bolus intravenously, flumazenil induces transient, incomplete clinical and electrophysiological ameliorations of overt encephalopathy in an appreciable proportion of patients with encephalopathy complicating acute liver failure or cirrhosis.

Ascites

Treatment is discussed in Chapter 14.21.2.

Acute hepatocellular failure

Prevention

The incidence of fulminant hepatitis B should be substantially reduced by widespread vaccination. Fulminant hepatic failure can be prevented by avoiding re-exposure to an agent that has induced an idiosyncratic acute hepatitis (such as halothane), and may be prevented by giving *N*-acetylcysteine after paracetamol overdose.

Acute liver failure

Treatment of acute liver failure in the absence of encephalopathy is expectant, but frequent monitoring is necessary when the prothrombin time is prolonged and prompt admission to hospital is indicated at the first sign of encephalopathy. Referral to a specialized liver unit has been recommended before encephalopathy develops if levels of clotting factors fall to less than 50 per cent of normal.

Fulminant and subfulminant hepatic failure

Routine management for acute hepatic encephalopathy is instituted (by extrapolation from the management of hepatic encephalopathy complicating chronic liver disease). With the onset of stage II hepatic encephalopathy, intensive supportive care should be instituted and transfer of the patient to a unit with the potential of undertaking orthotopic liver transplantation is recommended. All patients are considered to have potentially reversible disease. Treatment is designed to buy time for hepatic regeneration to take place and to avoid iatrogenic deterioration. No factor reported to stimulate hepatic regeneration experimentally is of proven clinical benefit. Conventional intensive care for the unconscious patient is instituted. A fluid intake of 1 to 2 litres daily is usually adequate. A nasogastric tube is used to decompress the stomach and detect upper gastrointestinal haemorrhage. Despite the coagulopathy, an arterial catheter is useful for continuous blood-pressure monitoring, frequent blood sampling, and measurement of blood gases. Caloric intake is maintained by infusing hypertonic dextrose (10 to 50 per cent), usually 200 to 300 g/day into a central vein. Intravenous lipids and amino acids may also be given. Vitamins may be given empirically. Unless the aetiology is known to be non-infectious, blood and all secretions are considered to be infectious. In this circumstance, attending personnel should wear gowns, gloves, and masks. Enteric isolation procedures are enforced and all specimens from the patient are labelled as infectious.

Blood is withdrawn at the outset for serological markers (such as for hepatitis viruses, cytomegalovirus), screening for common drugs, and estimation of serum copper. Blood glucose is monitored as frequently as every 1 to 2 h. The following investigations are carried out every 12 h: haemoglobin, total and differential leucocyte count, platelet count, urea, creatinine, potassium, sodium, chloride, and bicarbonate. Daily investigations include prothrombin time, total and direct bilirubin, alkaline phosphatase, alanine and aspartate aminotransferases, albumin, amylase, calcium, phosphate, magnesium, fibrinogen, and fibrinogen split products. Chest radiographs are obtained daily. Serial ultrasonic determinations of liver size may be useful in following the course. Needle biopsy of the liver is contraindicated. Frequent semiquantitative assessment of neurological status (such as the Glasgow coma score) and continuous monitoring of the electrocardiogram and electoencephalogram should be instituted. Patients should be monitored frequently for complications, which must be treated promptly and vigorously. A major goal is the prevention of brain damage. If agitation, piercing cries, delirium, or seizures occur, the patient should be restrained in a dark quiet room and the temptation to administer sedatives should be resisted.

Specific problems

Susceptibility to infections

Intensive microbiological monitoring is necessary. In fulminant hepatic failure, daily cultures of blood, urine, sputum, and swabs of intravenous cannulas are recommended. For patients admitted to a liver failure unit, prophylactic antimicrobial therapy has been advocated, such as intravenous, broad-spectrum antibiotics, oral or nasogastric amphotericin B suspension, and in females, vaginal clotrimazole cream. Nephrotoxic aminoglycosides are avoided. Antibiotics are recommended

to cover invasive procedures. Potential sources of infection, such as intrauterine devices, are removed.

Acid–base and electrolyte disturbances

Alkalosis does not require treatment. Acidosis should be managed by specific treatment of the cause; intravenous sodium bicarbonate increases body sodium. Hypokalaemia (potassium less than 3.5 mmol/l) is corrected by adding potassium chloride to intravenous fluids. The serum potassium is not increased above 4.0 mmol/l if liver transplantation is an option, as graft reperfusion may precipitate hyperkalaemia. Addition of sodium chloride to intravenous fluids is not indicated in the presence of hyponatraemia unless there is clear evidence of excessive loss of sodium. Sudden changes in sodium concentration should be avoided; they have been causally related to central pontine myelinolysis.

Cerebral oedema and raised intracranial pressure

Hepatic encephalopathy in fulminant hepatic failure may be compounded by cerebral oedema and raised intracranial pressure, hypoglycaemia, hypoxia, renal failure, and acid–base/electrolyte changes. Cerebral oedema precedes raised intracranial pressure and is not always demonstrable on CT scan; it may be indicated by a loss of demarcation between grey and white matter. To avoid precipitating an increase in intracranial pressure, patients are nursed in a quiet room with the trunk and head elevated 40°; jugular venous compression is avoided. To measure intracranial pressure, direct monitoring is necessary. A parietal or temporal bur hole is required to place an extradural or subdural pressure transducer. This procedure is controversial—it is potentially hazardous due to the coagulopathy and should be undertaken by a neurosurgeon in an operating theatre. Epidural monitoring is safer, but less accurate, than subdural monitoring. The cerebral perfusion pressure (mean arterial pressure minus intracranial pressure) should be maintained at a minimum of 60 mmHg. The best time to introduce a transducer is uncertain, but may be when progression to stage III encephalopathy occurs or when the patient becomes a candidate for liver transplantation. Although monitoring intracranial pressure has not been associated with increased survival, it may facilitate optimal management before liver transplantation. Mannitol has been shown to reduce elevated pressures that are not greater than 60 mmHg. It is given as an intravenous bolus of a 20 per cent solution (1 g/kg), which can be repeated every 4 h (0.5 g/kg) if the previous infusion induced a diuresis, plasma osmolarity does not exceed 315 mosmol/l, and azotaemia is not present. Mannitol has variable and potentially deleterious effects on intracranial pressure when the initial pressure is over 60 mmHg, and should probably not be given without prior measurement of intracranial pressure. Thiopentone (185 to 500 mg intravenously over 15 min), indomethacin, or induced hypothermia may reduce intracranial pressure if mannitol is ineffective. Corticosteroids and controlled hyperventilation do not appear to be effective treatments.

Haemorrhagic diathesis

The haemorrhagic diathesis requires no treatment in the absence of overt bleeding. Vitamin K (10 mg) is usually given intravenously, in spite of the low risk of inducing anaphylaxis. Fresh frozen plasma is not given routinely, so that plasma levels of clotting factors can be used as indices of prognosis. Skin puncture sites may require protracted pressure to achieve haemostasis. Administration of an H_2-antagonist to maintain gastric pH above 5.0 decreases transfusion requirements in fulminant hepatic failure; dose reduction may be necessary in the presence of renal failure. Infusion of platelets and fresh frozen plasma may be indicated to cover invasive procedures. Clotting factor concentrates, which exacerbate disseminated intravascular coagulation, are contraindicated. Heparin is not indicated for mild disseminated intravascular coagulation. Standard regimens of endoscopic diagnosis and therapy are instituted when haemorrhage occurs from the gastrointestinal tract. A haematocrit of at least 30 to 35 per cent should be maintained. A substantial increase in intracranial pressure may be due to an intracranial haemorrhage and is an indication for a CT scan of the head.

Hypoglycaemia

Hypoglycaemia must be prevented. Dextrose is administered intravenously to maintain a plasma glucose of 60 to 200 mg/dl. Occasionally, massive amounts of dextrose are required (for example more than 2 kg).

Cardiovascular changes

Maintenance of a normal blood pressure may reduce the risk of cerebral oedema or lessen its severity. Ionotropes may increase tissue hypoxia and have not been shown to be beneficial. However, if sepsis is suspected as a cause of hypotension, an ionotrope infusion may be warranted. If hypertension occurs, hypotensive or vasodilator drugs, which might adversely affect intracranial pressure, are not given. Arrhythmias may subside with correction of hypoxia, or acid–base or electrolyte disturbances.

Hepatorenal syndrome

Optimization of blood volume by infusing 20 per cent albumin may transiently improve renal function by correcting haemodynamic disturbances. Care is taken not to overload the circulation to avoid an adverse effect on intracranial pressure. Any therapeutic agents that may contribute to impaired renal function, including diuretics, are avoided. Severe acid–base/electrolyte disturbances or fluid overload, and rarely azotaemia, may be an indication for ultrafiltration or renal dialysis. Such procedures, which must be undertaken carefully because of cardiovascular instability and coagulopathy, may be necessary to optimize a patient's condition before liver transplantation, but would not be expected to alter the course of the hepatic or renal dysfunction. In the presence of raised intracranial pressure, ultrafiltration is preferred. Continuous venous access may be obtained using a double-lumen tube. The left femoral vein may not be used to facilitate venovenous bypass during liver transplantation. Use of vasopressin analogues is experimental.

Hepatopulmonary syndrome

No attempt should be made to correct hyperventilation. Hypoxaemia is an indication for 100 per cent oxygen by face mask. Endotracheal intubation is recommended at the onset of stage III encephalopathy. A tube large enough to permit bronchoscopy should be used. The procedure may be facilitated by curarization. Assisted mechanical ventilation is indicated for patients with stage IV encephalopathy, respiratory failure (increasing PCO_2), or pulmonary oedema. Positive end-expiratory pressure, which may reduce hepatic blood flow and increase intracranial pressure, should be avoided.

Convalescence

Abstinence from alcohol for a period of 6 months is recommended after recovery from an episode of acute liver failure not precipitated by alcohol. If alcohol was implicated in such an episode, lifelong abstinence is advocated. Other identified precipitants, such as halothane, must be rigorously avoided.

Temporary hepatic support

The original rationale for providing temporary hepatic support was based on the assumption that the hepatic lesion in fulminant hepatic failure is potentially reversible, provided that the patient can be kept alive sufficiently long for hepatic regeneration to take place. Theoretically, the patient selected for treatment with temporary hepatic support would die if treated by conventional intensive supportive care alone and would survive if the functions of the liver could be provided artificially over a finite period (Fig. 7). Temporary hepatic support should not only maintain the general condition of the patient, but also prevent life-threatening complications of fulminant hepatic failure. As there is a lack of detailed understanding of the biochemical disturbances that need to be corrected in fulminant hepatic failure, the design of artificial liver support systems has been largely empirical. Attempts have been made to (i) remove substances that have accumulated in the body using non-biological systems, such as charcoal haemoperfusion; (ii) provide deficient factors normally synthesized by the liver as well as clear accumulated substances using biological systems, for example haemoperfusion using devices containing hepatocyte preparations; and (iii) combine both approaches (hybrid systems). It has not yet been established that the risk/benefit ratio associated with application of any liver support device favours the patient. The provision of temporary hepatic support in the management of chronic irreversible liver disease (such as cirrhosis) may be limited to preparing a patient for liver transplantation. When a patient with fulminant hepatic failure or cirrhosis is a candidate for liver transplantation, effective temporary hepatic support may not only reduce operative or perioperative mortality but may also beneficially increase the waiting time for a donor liver.

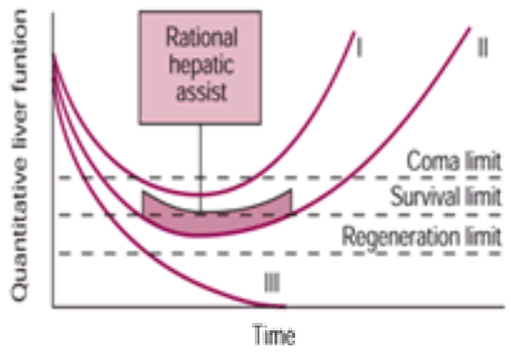

Fig. 7 Three hypothetical courses of fulminant hepatic failure (as envisaged by N. Tygstrup). In group I patients, liver function deteriorates below the coma limit, but does not fall below the survival limit; these patients should survive with intensive conventional medical supportive care alone. In group II patients, liver function deteriorates below the survival limit, but, if it can be maintained above this limit for a sufficient time by providing effective temporary hepatic support (rational hepatic assist), liver function would not fall below the regeneration limit and these patients should also survive. In group III patients, liver function deteriorates below the survival and regeneration limits irrespective of whether temporary hepatic support is provided. Liver transplantation offers the only hope of survival for group III patients. To facilitate optimal selection of patients for temporary hepatic support and liver transplantation, it is necessary to develop reliable criteria that indicate which course an individual patient will follow.

Further reading

Basile AS, Jones EA, Skolnick P (1991). The pathogenesis and treatment of hepatic encephalopathy: evidence for the involvement of benzodiazepine receptor ligands. *Pharmacological Reviews* **43**, 27–71.

Batkai S, *et al*. (2001). Endocannabinoids acting at vascular CB_1 recptors mediate the vasodilated state in advanced liver cirrhosis. *Nature Medicine* **7**, 827–32.

Blei AT, Butterworth RF, eds (1997). Hepatic encephalopathy. *Seminars in Liver Disease* **16**, 233–338.

Chang S-W, Ohara N, eds (1996). The lung in liver disease. *Clinics in Chest Medicine* **17**, 1–169.

Gines P, Arroyo V, Rodes J (1998). Ascites and hepatorenal syndrome: pathogenesis and treatment strategies. *Advances in Internal Medicine* **43**, 99–142.

Jones EA, Weissenborn K (1997). Neurology and the liver. *Journal of Neurology, Neurosurgery and Psychiatry* **63**, 279–93.

Pappas SC, Jones EA (1983). Methods for assessing hepatic encephalopathy. *Seminars in Liver Disease* **3**, 298–307.

Williams R, ed. (1996). Fulminant hepatic failure. *Seminars in Liver Disease* **16**, 341–444.

14.21.4 Liver transplantation

Graeme J. M. Alexander and M. Allison

Introduction

Pioneer work in the dog in the 1950s laid the basis for human liver transplantation, which was first undertaken in 1968 but liver transplantation was only established as a treatment for human therapy for liver failure in the mid 1980s. The number of centres and countries providing liver transplant services and the number of recipients continue to rise, accompanied by a combination of increased surgical and anaesthetic competence, more effective immune suppression, and improved organ preservation.

Liver transplantation services are now part of a sophisticated, multidisciplinary specialty. One-year survival of over 90 per cent should be achieved for low-risk elective cases. Successful liver transplantation enables most patients with severe liver disease to return to normal life.

The donor organ

Size match between donor and recipient is critical. Transplantation can be performed across the ABO barrier in exceptional circumstances but at the expense of long-term graft function. Matching for blood group is routine. Donor organ quality has a major impact on the immediate postoperative outcome and is influenced by many factors including experience of the surgical retrieving team, ensuring adequate, healthy artery, vein, and bile duct for the anastomoses. The introduction of cold University of Wisconsin solution *in situ* to the donor organ has prolonged the acceptable cold ischaemia time and transplantation can now be undertaken during working hours.

Fatty changes in the donor liver (the most common cause of primary graft non-function and the most frequent reason that a retrieval team refuse a liver) is important since such livers are less tolerant of cold and warm ischaemia. Relative donor shortage, increasing waiting lists, and prolonged waiting time compel the use of donors of marginal viability. Early postoperative graft function is likely to become more prevalent. Learning to use donor livers of impaired quality is a critical area for future research.

Surgical aspects

After recipient hepatectomy, the donor liver is implanted into the abdomen and anastomoses of the upper inferior vena cava, portal vein, lower inferior vena cava, and the hepatic artery are carried out. Surgeons may have to improvise the hepatic artery anastomosis, since about 20 per cent of donor hepatic arteries have anomalous anatomy. Venovenous bypass with extracorporeal circulation of blood via the portal vein back into the systemic circulation is associated with improved outcome in difficult cases and reduced rates of renal failure and sepsis.

Biliary anastomosis is carried out end-to-end, duct-to-duct unless there are doubts about the vascular supply to the biliary tree, with a second or subsequent transplant, or with biliary tract disease when a Roux loop is used. Whether this is appropriate in primary sclerosing cholangitis is contentious, since residual biliary tissue may retain a risk of future cholangiocarcinoma. In the past, the gall bladder was used as a conduit between recipient and donor ducts and was associated with a high rate of biliary strictures and gallstones.

Living-related donation has the advantages of a planned procedure, avoidance of cold ischaemia, minimal warm ischaemia time, and donor/recipient relatedness. However, the risk to such donors is not negligible. Auxiliary transplantation is one option for metabolic disease and acute liver failure but is not established because of difficulties in establishing an additional supply of blood. In an era of donor shortage, split livers are attractive, whereby the liver is divided and offered to two recipients, is an attractive option. Multiple organ transplantation is associated with reduced rates of rejection of organs transplanted with the liver, at the expense of prolonged delay waiting for a suitable donor.

Complication rates are higher in paediatric recipients because of size, the increased likelihood of in congenital anomalies and the use of 'cut-down' livers. Previous surgery is associated with an increased risk of haemorrhage. Established portal vein thrombosis is also a relative contraindication to transplantation.

Indications

There are no fixed rules for liver transplantation, which should be considered for patients with progressive disease where death is a likely end point (Table 1). It should be considered for patients with poor quality of life (a subjective assessment) who are able to withstand the procedure. For certain disorders guidelines are available to aid decisions.

Complications of liver disease

Liver transplantation may be used to treat the complications of liver disease: hepatic encephalopathy, ascites, subacute bacterial peritonitis, variceal haemorrhage (particularly gastric varices), jaundice, malnutrition, hepatic osteodystrophy, hepatopulmonary syndrome, hepatorenal syndrome, reversed portal vein flow, and superimposed hepatocellular carcinoma.

Early referral

Early referral allows time for the introduction of the patient to the concept of liver transplantation and improves outcome. Patients with a small body habitus and blood

group O have a prolonged wait and should be assessed early. Increasing age, renal dysfunction, poor nutritional status, high Child–Pugh score, and jaundice are associated with a poor outcome.

Assessment

As well as assessing the severity of their underlying liver disease, patients should be assessed for fitness for surgery. The presence of hepatopulmonary syndrome and pulmonary hypertension might complicate anaesthesia, but are not absolute contraindications. Some patients require psychiatric evaluation for alcohol or drug addiction. Early nutritional assessment is advised, to address weight loss in advance.

Accepted contraindications

Patients with AIDS and those with viraemia would not be considered without introduction of antiviral therapy. Extrahepatic malignancy is considered a contraindication, with the exceptions of neuroendocrine tumours and some cases of haemangioendothelioma. Metastatic disease involving liver is a contraindication.

Relative contraindications

The presence of HIV antibody may be regarded as a contraindication, despite an adequate CD4 count and absence of viraemia. Age alone is not a contraindication. Severe psychiatric disease requires psychiatric assessment. Continued alcohol or substance abuse, or a history of recidivism, are relative contraindications. Portal vein thrombosis in the absence of an adequate alternative vessel is a surgical contraindication. The likelihood of non-compliance with hospital attendances and immunosuppressive medication is difficult to assess but those deemed incapable of complying with these demands should not be considered.

Waiting list

Whilst waiting for transplantation, patients may develop new complications of their liver disease. Sepsis is common and requires suspension from the waiting list for treatment. Hyponatraemia (<125 mmol/l) is a risk factor and should be corrected in advance to avoid the risk of central pontine myelinolysis. Non-specific deterioration should prompt imaging of the portal vein and liver for new thrombosis or tumour respectively, and a search for sepsis. A proportion of patients develop the hepatorenal syndrome whilst under surveillance.

Specific disorders

Acute liver failure

Compared with chronic liver disease, the outcome for acute liver failure is less favourable (Table 2). However, the patients are often younger and may be cured by transplantation. For patients with paracetamol (acetoaminophen) poisoning, the decision to proceed to transplantation is made on the basis of prothrombin time, bilirubin, hepatic encephalopathy, renal failure, and acidosis; late rises in prothrombin time denote a grave prognosis and often indicate the presence of sepsis. Such patients require psychiatric assessment at the earliest opportunity to identify treatable conditions and to assess future compliance with medication and follow-up.

For patients with acute liver failure of other cause (drugs, hepatitis A, hepatitis B) the decision to proceed to transplantation is based on considerations of age, aetiology, jaundice to encephalopathy time, bilirubin, and prothrombin time.

Subacute hepatic failure

The prolonged period between onset of jaundice and encephalopathy allows patients with subacute hepatic failure to be identified. These are likely to survive long enough to be offered a liver and have an excellent outcome. Without transplantation, most patients with subacute hepatic failure die.

Chronic liver disease

Primary biliary cirrhosis

Several models have justified the introduction of transplantation for patients with primary biliary cirrhosis based on knowledge about life expectancy without liver transplantation. Such analyses are useful in decision making since definitive, controlled trials have never been done. A bilirubin in excess of 100 µmol/l predicts high mortality without transplantation. Some patients require transplantation because of extreme fatigue and pruritus prior to the onset of significant jaundice.

Primary sclerosing cholangitis

Patients suffering from primary sclerosing cholangitis are amongst the hardest to assess for liver transplantation. The presence of frequent episodes of jaundice, increasing jaundice, or cholangitis should prompt consideration of liver transplantation. There is a significant lifetime risk of cholangiocarcinoma (which would be a contraindication) and an ever-present concern that it may have evolved. Patients with primary sclerosing cholangitis should be considered for a transplant early in the course of their illness rather than late. Measurement of serum CA19.9 identifies patients with an evolving cholangiocarcinoma, but this value can rise very significantly in the presence of severe biliary disease and sepsis and should not be relied upon alone. Inflammatory bowel disease should be sought and treated before transplantation, even when asymptomatic.

Autoimmune hepatitis

Transplantation is indicated for complications of autoimmune hepatitis, poor control despite adequate therapy, or those patients who are slow to achieve control or escape control with immunosuppressant drugs.

Chronic viral hepatitis

Patients with chronic hepatitis B virus infection (HBV), chronic hepatitis C virus infection (HCV), and chronic hepatitis D virus infection (HDV) should be considered on merit according to the presence or absence of the complications of chronic liver disease and symptoms.

Alcohol

Patients with alcoholic liver disease undoubtedly constitute the most contentious group: liver transplantation is restricted largely to those with progressive disease despite abstinence. Most centres insist on complete abstinence from alcohol, first because the liver disease improves with prolonged abstinence and second because it is thought that patients who are able to maintain abstinence for six months have a lower risk of recidivism after transplantation. Assessment of abstinence is difficult and the ethics of random testing for alcohol abuse without consent are complex. It is important that details of any contract between patient and doctor are recorded accurately. In practice, those patients able to maintain a stable home life, job, and partner are more likely to be offered a liver transplant than those who are homeless, jobless, and isolated.

Tumours

For the treatment of hepatocellular carcinoma evolving on a background of cirrhosis, there are two options—resection and liver transplantation. Resection is unrealistic for most with cirrhosis and restricted to those with accessible tumours and Child–Pugh class A. Transplantation is indicated for those with a low risk of tumour recurrence. Accurate assessment of the extent of disease is therefore critical. Patients with tumours greater than 5 cm, or more than three nodules, or evidence of extrahepatic disease are likely to have recurrent disease after the transplant procedure. Radiological assessment of the extent of disease usually includes multiple modalities (hepatic angiogram, CT, MR, and ultrasound scans). Despite this, capsular or vascular invasion is often discovered at the time of transplantation.

The fibrolammellar variant of hepatocellular carcinoma usually occurs in young patients without cirrhosis and is probably more slow growing. However, resection is more appropriate for this group but although the 5-year survival figures for fibrolammellar disease are good, a significant risk of tumour recurrence remains.

Cholangiocarcinoma

The risk of postoperative recurrence is so high that this condition is regarded as a contraindication to hepatic transplantation.

Neuroendocrine tumour

P>Where disease is limited to the liver with a long lead-time, transplant offers reasonable 5-year survival. Recurrence after transplantation is frequent, although recurrent disease may be amenable to other strategies—chemotherapy, hormone modulation, radiotherapy, and direct tumour ablation.

Haemangioendothelioma

The reported prognosis after transplantation varies from early recurrence to none. A firm recommendation is impossible and the decision to go ahead with the procedure is often based on factors other than histology.

Budd–Chiari syndrome

Patients with acute liver failure in relation to the Budd-Chiari Syndrome should be managed as any other patient with acute liver failure. For chronic disease, there are alternative approaches, including surgical shunts and transplantation. One expressed view is that liver transplantation is the most reliable and effective shunt procedure. Most patients with Budd–Chiari syndrome have an underlying disorder predisposing to thrombosis, which should be thoroughly investigated and treated before the transplantation procedure is carried out.

Metabolic/genetic disease

Transplantation is successful in a number of metabolic/genetic disorders associated with liver disease, including Wilson's disease, a-1-antitrypsin deficiency (with maintained respiratory function), Gaucher's disease, glycogen storage disease, Crigler–Najar syndrome, and Bylers syndrome. Patients with haemochromatosis have increased morbidity, perhaps as a consequence of associated disorders (diabetes and cardiomyopathy), as well as an increased incidence of sepsis in the early postoperative stages that may be related to systemic iron overload.

Polycystic liver disease rarely requires liver transplantation but where associated with renal polycystic disease, combined liver/kidney transplantation is considered. The Rendu–Osler–Weber syndrome may be associated with hepatic haemangiomas causing portal hypertension or shunting that result in biliary ischaemia, and thus is best treated by liver transplantation.

Patients with cystic fibrosis represent a difficult group, presenting with portal hypertension and cardiorespiratory disease requiring heart/lung transplantation. Waiting for a triple organ donation for a recipient with a small body habitus may be futile, while proceeding with single organ transplantation in a patient with several diseased organs that will ultimately require transplantation procedures in their own right, may also be mistaken.

Metabolic disorder with a structurally normal liver

Patients with a structurally normal liver have undergone liver transplantation for hypercholesterolaemia and hyperoxaluria. In the former, it is essential to address the cardiovascular complications before transplantation. For the latter, it is important to recognize that the kidneys transplanted simultaneously remain at risk of oxalate-induced damage for several years after otherwise successful hepatic transplantation has been carried out.

Complications

Immediate

The most immediate, usual complication of transplantation is perioperative haemorrhage.

The presence of one or more of: coma following withdrawal of sedation, a rapidly rising prothrombin time, acidosis, high insulin requirement, thrombocytopenia, and hyperkalaemia, prompts consideration of three diagnoses—primary non-function of the graft, non-thrombotic graft infarction, or vascular thrombosis. Imaging of the hepatic artery and portal vein with ultrasound and angiography may greatly aid distinction between these possibilities. The patient may need a further hepatic graft as an emergency and delay in this situation can be catastrophic for survival.

Hyperacute rejection of the liver is rare (in contrast to the kidney) and liver transplantation can be undertaken successfully in the presence of a positive cross-match.

Early

Hepatic artery thrombosis may be identified by worsening liver function tests (confirmed by ultrasound and/or angiography), by ischaemia seen on liver biopsy, or biliary leak due to ischaemia of the bile duct. Portal vein thrombosis may present with either an ischaemic graft (less acute than that seen in the immediate postoperative phase) or the presence of ascites. Caval stricture might present with rapid accumulation of ascites and peripheral oedema. With an abdominal drain *in situ*, many litres of ascites might need to be removed daily. Cholangitis can be recognized by fever, neutrophilia, pain, and jaundice. A stricture should be sought by ultrasound.

Acute rejection

Acute rejection of the grafted organ may present with pain, fever and jaundice. Most often however, acute rejection is identified by noting a deterioration in liver biochemistry (particularly bilirubin) associated with peripheral blood eosinophilia. A liver biopsy is essential to determine severity and may confirm the presence of tissue eosinophils in association with lymphoblasts. The target tissues are bile duct and endothelial cells (hepatic artery, portal vein, and, less frequently, central vein). Not all acute rejection requires therapy but patients with acute rejection are prescribed supplemental high-dose corticosteroids. Severe rejection, steroid resistant rejection, and multiple episodes of acute rejection carry a poor long-term prognosis for the graft.

Graft-versus-host disease

Graft-versus-host disease is characterized by a skin eruption, gastrointestinal disturbance, and malnutrition. Often, patients with graft-versus-host disease are malnourished, have alcohol-related liver disease, are lymphopenic, and do not develop acute rejection. The diagnosis is confirmed by identification of host and recipient lymphocytes in the circulation. The best form of therapy is likely to be monoclonal antibody directed against activated T cells. Overall, the prognosis is poor.

Hepatic artery and portal vein strictures

These are identified usually as a consequence of abnormal liver biochemistry or a slow recovery following transplantation; occasionally, however, liver biochemistry may be normal. Ultrasound is unreliable for identification of vascular strictures and CT perfusion scanning or angiography are recommended.

Biliary strictures

Persistent elevation of the alkaline phosphatase and an ultrasound revealing intrahepatic duct dilatation reveal strictures of two types. The most common is

anastomotic and usually amenable to dilatation at ERCP. With recurrence, reconstruction should be considered. A hilar stricture should prompt a search for ischaemia. In these circumstances reconstruction and/or retransplantation are considered.

Bacterial infection

Bacterial sepsis is common in chest, urine, blood, abdomen, and intravascular cannulae. There is no universal recommendation for antibiotics although most physicians recommend the use of antimicrobial prophylaxis for up to 48 h. Thereafter, antimicrobial therapy is based on careful observation, clinical assessment, culture, and local knowledge of likely pathogens.

Viral infections

Herpes simplex virus infections are often clinically apparent and can be managed with topical acyclovir, unless there is a suspicion of systemic disease, requiring parenteral acyclovir. It seems probable that genital herpes will also be reactivated.

In the past, cytomegalovirus (CMV) infection has been the principal viral cause of infection in this period but there have been significant advances in the past decade. Prophylaxis with oral ganciclovir to prevent clinical expression of disease has proved effective but it may occur with reduced severity once ganciclovir is withdrawn at 3 months. For patients who have completed prophylaxis and recipients with antibody at transplantation, recrudescence or infection with a donor strain can be investigated by means of PCR for CMV DNA in serum. CMV viraemia is associated closely with clinical disease, affecting the gastrointestinal tract, central nervous system, respiratory system, the bone marrow, retina, and liver. Patients who become CMV PCR positive should receive systemic ganciclovir. Newer, quantitative assays might guide therapy more accurately. Ganciclovir resistance has been reported. Shingles is a common complication of the early postoperative period and should be treated with systemic antiviral therapy with acyclovir. Amitriptyline and carbamazepine reduce the risk of postneuralgic syndrome.

Fungal infections

Candidal species are isolated commonly and prophylaxis is in routine use. Fluconazole is effective, but drug interactions, particularly with immunosuppressive agents, may render its use problematic. Oral therapy with nystatin or amphotericin is recommended at a later stage, up to 3 months. Systemic candidal or *Aspergillus* infections are more likely in those with severe liver disease, ischaemic grafts, a poor postoperative course, and renal impairment.

Protozoal infections

Co-trimazole prophylaxis for lymphopenia has almost eradicated *Pneumocystis carinii* infection in many centres. Toxoplasmosis is rare since the introduction of prophylaxis with pyrimethamine or co-trimazole for 6 weeks for patients with prior serological evidence of infection.

Late

Multiple biliary strictures may present with recurrent cholangitis and may be a late expression of ischaemia. A proportion comes ultimately to retransplantation, usually after a period of years.

Lymphoproliferative disease

This affects 2 to 4 per cent of patients and is less common with liver than with other solid organ grafts. Usually the disorder is a B-cell lymphoma and most are Epstein–Barr virus related. The most frequent location for lymphoma is the liver. Management is by reducing immunosuppression to a minimum and chemotherapy according to conventional guidelines.

Carcinoma

There is a substantial increase in the incidence of almost all carcinomas, including squamous and other skin cancers.

Osteodystrophy

A large proportion of patients has severe bone disease at presentation, which then worsens so that there is an increased fracture rate in the postoperative period. The severity of bone disease has improved considerably over the past decade because of the reduced dose of corticosteroids used. Bone mass improves significantly over the first 2 to 3 years after transplantation but it is wise to use bisphosphonates (and sex hormones) for patients who show clear evidence of reduced bone density.

Chronic rejection

This is a devastating consequence of liver transplantation and is rarely reversible. It can be predicted on the basis of severe acute rejection, steroid-resistant acute rejection, or multiple episodes of acute rejection. It leads to graft loss and the requirement for a further transplant. It is uncertain what the main target for the process is and the precise immunological nature of the process remains to be determined. However, the main branches of the hepatic artery become obliterated with foamy macrophages and the histological pattern resembles chronic rejection of other solid organs. Bile ducts are lost (the vanishing bile duct syndrome), probably as a consequence of chronic ischaemia.

Cardiovascular disease

Patients who have undergone liver transplantation have a significant increase in their cardiovascular risk profile in the early postoperative phase, which extends long term. It is probable that this represents a consequence of immunosuppression, since hyperlipidaemia, hypertension, renal impairment, and weight gain are common features.

Recurrent disease

It is recognized that immunosuppressive regimes should be modified according to the primary indication for transplant, especially for prevention or treatment of recurrent graft disease.

Hepatitis B virus

Serum HBV DNA at the time of transplantation or the presence of HBcAg in host liver predicts accurately the likelihood that the graft will become infected with HBV. Graft infection without treatment is associated with a significant morbidity and mortality such that in the 1990s many patients with HBV were not considered.

The situation has been revolutionized. Hepatitis B immunoglobulin, although expensive, reduces the rate of graft infection, particularly in those HBV DNA negative in serum at the time of transplantation. Lamivudine prevents graft infection if given prior to transplantation and is effective therapy for graft infection. Regrettably, lamivudine resistance evolves rapidly and mutated virus can cause liver damage. Adefovir for patients with lamivudine resistant virus post-transplantation has been useful and resistance to the combination has not yet been reported. The role of adefovir in preventing graft infection has not been assessed. Corticosteroids should be withdrawn by week 6 and maintenance on single therapy with tacrolimus rather than cyclosporin is recommended.

Hepatitis C virus

HCV infects the graft inevitably, and is associated with a rapidly progressive fibrosis which is not usual in HCV-positive patients not treated by hepatic transplantation. A significant proportion of patients have cirrhosis by 5 to 10 years and a small proportion of patients lose the graft to HCV infection within 1 to 2 years of

transplantation. Trials of interferon-a and ribavirin in the transplant setting are underway but there is no evidence yet that this combination is able to prevent graft infection or to prevent liver damage once graft infection occurs. Most established transplant units recommend minimal immunosuppression for this group of patients.

Autoimmune hepatitis

Undoubtedly this can affect the graft and cause graft loss. It may be predicted in advance of significant deterioration by monitoring immunoglobulins. A small proportion of patients develops 'autoimmune hepatitis' *de novo*. Triple therapy with long-term azathioprine and long-term low dose prednisolone, tacrolimus, or cyclosporin is advised.

Primary sclerosing cholangitis

This may affect the graft and is associated with an increased incidence of biliary complications. Immunosuppression does not appear to prevent recurrence in the graft, which may be lost.

Primary biliary cirrhosis

This can affect the graft and cause graft loss. It appears that immunosuppression does not prevent recurrence. Indeed, cirrhosis with portal hypertension has occurred within 2 years of grafting for this indication.

Alcohol

Recidivism for alcoholism and the development of alcohol-related liver disease after transplantation for alcohol-related disease are well recognized but uncommon. Liver damage can lead to graft loss within a period of 12 months if the consumption of alcohol is resumed.

Immune regulation

Tolerance

Tolerance is probably rare and there are no adequate tests to identify tolerance in human grafts when it does occur. Studies of liver tissue in well patients with normal liver biochemistry have invariably shown many with abnormal liver histology; only a small minority are normal (and probably tolerant). It is recognized that calcineurin inhibitors prevent the development of tolerance, which is an active process. Nevertheless, these drugs are currently the best available for transplantation. Murine and other animal models indicate that a range of antibodies to T-cell markers can induce tolerance experimentally (including CD4, CD2, CD3, CTLA-4, CD45RB and CD40L).

Immunosuppression

Over the past decade the introduction of additional immunosuppressive agents has been a major therapeutic advance: individual disorders and individual patients can receive tailored therapy (Table 3). Maintenance in the early stages is based on triple therapy—cyclosporin or tacrolimus, which are prescribed indefinitely, azathioprine for 1 year, and prednisolone for a variable period. Our current practice is to stop prednisolone at 3months, adrenal function permitting. The daily dose of prednisolone used nowadays rarely exceeds 20 mg and the daily dose of azathioprine rarely exceeds 1 mg/kg. Large, parenteral doses of prednisolone are given for acute rejection.

Cyclosporin causes significant cardiovascular morbidity with hypertension, hyperlipidaemia, and weight gain. Neuropsychiatric illness is also reported. Peculiar to cyclosporin is the development of hirsutism and gum hypertrophy, making this particularly unsuitable for females. Renal impairment is a common problem which is probably under-reported and under-recognized. Renal grafting has been required for cyclosporin-induced renal failure.

Tacrolimus

Tacrolimus shares most of the side effect profile of cyclosporin, in particular the cardiovascular complications (hypertension, hyperlipidaemia, and weight gain) as well as neuropsychiatric disease and renal toxicity. The question of whether diabetes mellitus is induced by tacrolimus remains contentious.

Other immunosuppressive agents

Other agents used are unproven and at present lack clear clinical indications. These include rapamycin, which is associated with poor wound healing, bone pain, gastrointestinal upset, bone marrow suppression, and hypertriglyceridaemia. The absence of hypertension, weight gain, or renal toxicity may prove invaluable for those complications arising with calcineurin inhibitors. Mycophenolate mofetil is another unproven agent for liver transplant patients. It reduces cell proliferation in similar fashion to azathioprine but appears to be more lymphocyte-specific—gastrointestinal disturbances are common with this agent.

Antilymphocyte globulin and antithymocyte globulin are T-cell antibodies utilized in the treatment of steroid-resistant acute rejection. Some utilize one or other in the initial induction regime. They carry an increased risk of infection, long-term lymphopenia, and lymphoproliferative disease; the use of these compounds in routine immunosuppression is in decline.

A number of newer, monoclonal antibodies are in development or subject to evaluation in clinical trials. These are directed largely at activated T cells or adhesion molecules. Their clinical role remains to be defined.

Future

Three main areas can be identified for the future development of hepatic transplantation. First, increasing the number and quality of donor livers; at the same time research to improve rescue of damaged donor organs is critically important. Second, donor-specific tolerance remains a goal—in this respect, considerable encouragement has been gained by the successful induction of tolerance in experimental animals as a result of continued world-wide investment in basic transplantation research by immunologists. Finally, improved management of the long-term complications of transplantation, in particular cardiovascular and malignant disease, would greatly extend the duration and quality of life of the many patients who have received donor livers for otherwise fatal hepatic diseases.

Further reading

Balen V, Marsh JW, Rekele J (1999). Liver transplantation. In: Bircher J, Benhamou J-P, McIntyre N, Rizzetto M, Rodes J, eds. *Oxford textbook of clinical hepatology*, 2nd edn, pp. 2039–63. Oxford University Press, Oxford.

Carrithers RL Jr (2000). Liver transplantation: American Association of the Study of Liver Diseases practice guidelines. *Liver Transplantation* **6**, 122–35.

Devlin J, O'Grady J (1999). Indications for referral and assessment in adult liver transplantation: a clinical guideline. *Gut* **45** (Suppl. 6).

Morris RE (1996). Mechanism of action of new immunosuppressive drugs. *Kidney International* (Suppl.), S26–S38.

14.21.5 Primary and secondary liver tumours

Iain M. Murray-Lyon

Benign and malignant tumours may arise in the liver from the hepatocytes, bile-duct epithelium, or supporting mesenchymal tissue. With the exception of hepatocellular carcinoma all the primary malignant tumours are rare, but the liver is frequently the site of secondary (metastatic) deposits of malignant tumours elsewhere in the body.

Hepatocellular carcinoma

This occurs either as a single mass or as scattered nodules of tumour, and in around 80 per cent of patients there is pre-existing cirrhosis. The tumour tends to invade the portal and hepatic veins, and spreads to the abdominal lymph nodes and bone. Histologically the tumour is typically composed of cells resembling hepatocytes, which are arranged in cords. A number of other distinct histological subtypes are now recognized, including the fibrolamellar variant in which clumps of eosinophilic carcinoma cells are surrounded by a characteristic fibrous stroma. This tumour occurs in young adults in a non-cirrhotic liver.

Epidemiology

Although this is a comparatively rare tumour in Western Europe and North America where the annual incidence is around 1 to 2 per 100 000 of the population, there is recent evidence that it is becoming more common, and in Africa and South-East Asia the incidence is 20 to 30 times higher. In patients with underlying cirrhosis, males greatly outnumber females but in non-cirrhotic cases this sex difference is less striking. In areas of high incidence the peak age is in the third and fourth decades of life but in Europe and North America most cases occur in the fifth and sixth decades.

Aetiology

In all countries in the world, cirrhosis, particularly the macronodular form, is present in about 80 per cent of cases. In Western Europe and the United States this is usually due to chronic alcoholism or chronic hepatitis B or C, and at least 10 per cent to 15 per cent of such patients will develop a hepatocellular carcinoma. In Africa and Asia, chronic liver disease is usually associated with hepatitis B or C virus infection. Rare cases may complicate cirrhosis due to other causes, and may follow prolonged use of the oral contraceptive pill or prior investigations using the radioactive contrast agent Thorotrast.

In parts of Africa and the Far East there is increasing evidence implicating aflatoxin. This is a potent carcinogen derived from the mould *Aspergillus flavus*, which often contaminates food.

The hepatitis B virus (**HBV**) is now recognized to have an important role in the development of hepatocellular carcinoma, particularly in areas of high incidence. Long-term, follow-up studies of large numbers of HBV carriers have confirmed that the risk of developing hepatocellular carcinoma is at least 100 times higher than in matched uninfected controls. The HBV can be identified in the tumour as well as the surrounding liver, and integration of viral DNA in the genome of hepatocellular carcinoma has been shown. This HBV DNA integration may result in major structural rearrangements in adjacent cellular DNA, and a range of deletions, duplications, and translocations between chromosomes has been reported. In geographical areas of high endemicity of HBV as well as exposure to aflatoxins, one of a variety of mutations of the *p53* gene on chromosome 17 is a frequent finding. How these molecular events are initiated and progress is still unclear.

The hepatitis C virus (**HCV**) is also closely linked with the development of hepatocellular carcinoma, especially in geographical areas such as North America, Western Europe, and Japan where HBV is not hyperendemic. Almost all cases are associated with cirrhosis. Prospective studies of patients with chronic post-transfusion HCV infection indicate the latent period before tumour development may be as long as 25 to 30 years.

Hepatocellular carcinoma is a largely preventable disease, and the extensive use of hepatitis B vaccination has already led to a reduction in the incidence of this tumour in Taiwan. Until such time as a vaccine against HCV is available, introduction of all possible measures to reduce transmission of this virus is important. Furthermore, there are indications that the successful treatment of chronic hepatitis B and C will reduce the cancer risk.

Clinical features

In Africa and other high-incidence areas, patients usually present with a short history of right upper abdominal pain, often associated with fever and weight loss. There may be considerable abdominal swelling due to liver enlargement, with or without ascites. Catastrophic intraperitoneal bleeding sometimes occurs due to tumour rupture. In low-incidence areas the disease is often more insidious and presents as a general deterioration in the health of a patient already known to have cirrhosis. There is usually hepatomegaly and a bruit may be heard over the liver. A number of non-metastatic systemic manifestations may also rarely occur, such as hypoglycaemia, hypercalcaemia, and porphyria cutanea tarda.

Because of the use of screening in high-risk groups, more small (less than 3-cm diameter), asymptomatic tumours are now being detected.

Investigations

Haematological and biochemical indices

The haematological and biochemical changes, apart from alpha-fetoprotein, are non-specific and reflect the space-occupying lesion as well as the underlying cirrhosis present in about 80 per cent of cases.

Alpha-fetoprotein is a glycoprotein synthesized by the fetal liver and its plasma concentrations reach their maximum at the end of the first trimester (3 to 4 mg/ml) and then decline. After birth, concentrations fall rapidly to adult levels (1 to 10 ng/ml). Raised levels are found in about 80 per cent of patients with hepatocellular carcinoma and tend to be higher in African and Far-Eastern populations than in those in low-incidence areas and in those patients with small tumours. Concentrations above 500 ng/ml in a patient with liver disease are highly suggestive of hepatocellular carcinoma. However, in interpreting alpha-fetoprotein levels it should be remembered that high plasma levels are found in some patients with germinal-cell tumours of the testis and ovary as well as occasional patients with carcinoma of the stomach or pancreas, usually with hepatic metastases. Below 500 ng/ml there is a diagnostic 'grey zone', for such levels may be found in patients with severe viral hepatitis and active cirrhosis. But subsequent readings tend to fall towards normal in patients with these conditions, whereas in patients with hepatocellular carcinoma the levels rise progressively. Sequential readings are therefore of great diagnostic value, and the measurement of hepatoma specific isoforms may improve diagnostic specificity and sensitivity.

Other tumour markers for hepatocellular carcinoma have been described in the serum, including an abnormal vitamin B_{12} binding protein which is usually present with the fibrolamellar histological variant.

Liver imaging

Real-time ultrasound

This is a sensitive and specific test and picks up hepatocellular carcinoma in 85 to 90 per cent of cases. False-negative results usually occur in patients with tumours of less than 2 cm in diameter.

Abdominal computed tomographic (CT) scanning

This technique is probably no more accurate in detecting hepatocellular carcinoma than ultrasound and should be reserved for cases in which doubt persists. Sensitivity can be increased by contrast enhancement. Dynamic spiral contrast-enhanced CT scanning is even more sensitive.

Magnetic resonance imaging (MRI)

This technique, particularly with the addition of a contrast agent, is proving useful in identifying and characterizing focal liver masses. Manganese dipyridoxyl diphosphate (**Mn-DPDP**) targets hepatocytes and super-paramagnetic iron oxide (**SPIO**) targets reticuloendothelial cells, resulting in increased prominence of the lesion with respect to normal liver tissue (Fig. 1). Lesions that do not contain reticuloendothelial cells or hepatocytes (haemangiomas and metastases) do not have their signal intensities altered.

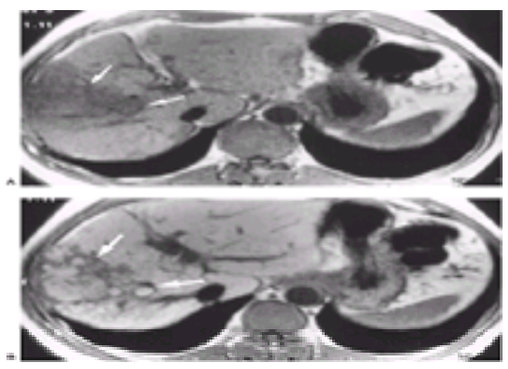

Fig. 1 (a) Axial T_1-weighted image through the liver in a patient with hepatitis C. An ill-defined area of reduced signal in the right lobe of the liver at the junction of segments 7 and 8 is suspicious for a liver tumour (arrowed). (b) Axial T_1-weighted image through the liver in the same patient following an infusion of Mn-DPDP, a hepatocyte specific contrast agent. There is a general increased signal in the normal liver and a heterogeneous increased signal in the liver tumour, characteristic of uptake by a hepatocellular carcinoma (arrowed).

Hepatic arteriography

Excellent visualization of the hepatic artery can usually be obtained by selective catheterization using the Seldinger technique. As the major vascular supply to a hepatocellular carcinoma is usually arterial, diagnostic changes are seen in a high proportion of cases. Information gained on the anatomical distribution of the tumour and the vascular anatomy is essential if surgical resection or transplantation is being contemplated, and consideration can also be given at the time of arteriography to intra-arterial chemotherapy and hepatic artery embolization. The sensitivity of arteriography can be increased by combining it with dynamic spiral CT scanning together with late films to show the portal venous system, as well as injection of the iodine-containing contrast medium Lipiodol. CT scanning 10 to 14 days later can visualize Lipiodol selectively retained in tumours as small as 2 to 3 mm in diameter.

Liver biopsy

For definitive diagnosis, liver biopsy is essential, although this is not always possible because of the prolongation of the prothrombin time. The diagnosis can be considered highly likely without liver biopsy proof if the alpha-fetoprotein level is greater that 500 ng/ml and the hepatic arteriogram shows a tumour circulation. Biopsy may be conveniently done at the time of laparoscopy or ultrasonography and suspicious areas can be sampled under direct vision. Because of the risk of tumour spread, biopsy is often avoided if curative resection or transplantation is planned.

Screening

Patients with an increased risk of developing hepatocellular carcinoma, such as those with cirrhosis or chronic HBV or HCV infection, should be considered for regular screening by alpha-fetoprotein assay and abdominal ultrasonography. While such a strategy has been shown to pick up early tumours, and there is evidence of improved survival figures for these patients in the Far East, the benefits of screening have so far proved disappointing in Europe.

Prognosis

This is a highly malignant tumour and the mean survival in most series is around 4 to 6 months. In Africa the disease tends to run a more malignant course. Patients with cirrhosis have a poorer prognosis than those without. Encapsulated tumours and the fibrolamellar histological variant, as well as small tumours picked-up at screening, have a better prognosis.

Treatment

Curative

Only complete resection or orthotopic transplantation hold out any chance of cure and these procedures should be considered in every case.

Resection

Resection is only possible in about 10 per cent of cases because of underlying cirrhosis or tumour in both lobes. Often a major resection is needed and the anatomical possibilities are illustrated in Fig. 2. In the presence of cirrhosis only a limited resection is possible as liver regeneration is defective, but this procedure may be curative if the tumour is small. In China, screening programmes to detect early hepatocellular carcinoma have led to higher rates of tumour resection in cirrhotic patients and improved long-term survival figures.

Fig. 2 Diagram to illustrate the main types of hepatic resection.

The best results are achieved in patients with well-compensated cirrhosis and small tumours (less than 3 cm), but 5-year survival rates of only 20 to 30 per cent are usually quoted due to progression of the underlying liver disease, tumour recurrence, and the development of new tumours.

Transplantation

This is an attractive option for patients with cirrhosis and a hepatocellular carcinoma and the procedure can cure both the tumour and the underlying cirrhosis. Long-term results are best with tumours less than 5 cm in diameter when survival figures up to 70 per cent at 5 years are recorded. The best results are obtained when a hepatocellular carcinoma is discovered incidentally in the resected liver when transplantation is performed for liver failure, and with the fibrolamellar histological variant.

Palliation

Radiotherapy

External-beam X-irradiation does not produce consistent improvement. Intra-arterial injection of Lipiodol mixed with iodine-131 has given some encouraging preliminary results.

Cytotoxic drugs

Doxorubicin (Adriamycin) is one of the few drugs that may produce worthwhile regression, but only 20 to 30 per cent of cases respond and no survival benefit has been established. Mitoxantrone (mitozantrone), which is structurally similar to doxorubicin, gives a similar response rate and has fewer toxic side-effects.

The presence of hormone receptors in hepatocytes has prompted attempts to modify tumour growth. Results are poor with tamoxifen but octreotide has been reported to be of value.

Targeted therapies

A wide variety of local targeted therapies have been developed and assessed in recent years. Very few, however, have yet been submitted to a prospective, randomized controlled trial.

Percutaneous ethanol injection

Sterile alcohol is injected directly into the tumour under ultrasound guidance causing tumour necrosis. Repeated injections may be given into more than one tumour mass. The best results are obtained with tumours less than 3 cm in diameter and the survival figures are comparable with those of limited surgical resection.

Lipiodol-targeted chemotherapy

Cytotoxic drugs may be emulsified with Lipiodol (see above) and delivered directly into the liver at selective hepatic arteriography. There are some data to show that the duration of action of the drugs is prolonged because of the retention of Lipiodol in the tumour. There is, however, no convincing evidence of benefit.

Transcatheter arterial embolization (TAE)

Embolization with foreign materials, such as gel foam, can be achieved at the time of hepatic arteriography and may result in substantial tumour necrosis, particularly in highly vascular tumours, which derive the bulk of their blood supply from the hepatic artery. In patients with decompensated cirrhosis and those with portal vein occlusion the procedure is contraindicated. Broad-spectrum antibiotics are given for some days because of the risk of anaerobic infection in the ischaemic liver. As tumour necrosis is never complete, embolization of the tumour should be combined with targeted chemotherapy (chemoembolization). The gel foam particles may be soaked in doxorubicin or cisplatin or the TAE may be immediately preceded by Lipiodol targeted chemotherapy. While such treatment may result in tumour necrosis, shrinkage, and symptomatic improvement, three controlled trials have not established survival benefit.

Cryoablation and thermal ablation

These techniques are intended to destroy tumour cells by physical means. No controlled trials have been reported. Cryotherapy probes are inserted into the tumour either at laparotomy or laparoscopy and liquid nitrogen is circulated through them. Thermal energy can be applied via probes placed percutaneously in the tumour using either laser, radiofrequency, or microwaves. All three techniques have proved safe, have few side-effects and can be repeated. Current data are insufficient to allow comparison of the efficacy of these ablative techniques.

Cholangiocarcinoma

Carcinoma may arise in any part of the biliary tree from the small intrahepatic bile ducts down to the lower end of the common bile duct. Two clinical varieties occur in the liver—a peripheral form, which consists of one single or multiple nodules; and the much commoner hilar form usually situated at the confluence of the right and left

hepatic duct. This invades locally and causes obstruction of the biliary tree. The histological appearances are identical whatever the site of origin. It is an adenocarcinoma with a simple ductular arrangement of columnar or cuboidal cells, usually with a prominent fibrous stroma.

Epidemiology

This tumour is much less common than hepatocellular carcinoma and accounts for about 7 to 10 per cent of primary malignant tumours, except in the Far East where it makes up about 20 per cent. The peak age is in the sixth and seventh decades and the sex incidence shows only a slight male predominance.

Aetiology

Thorium dioxide (Thorotrast) is a well-recognized but rare cause of the intrahepatic variety of the tumour. In the Far East, infestation of one of a variety of distomes (*Clonorchis sinensis*, *Opisthorchis viverrin*ı) is probably commonly related.

Patients with long-standing ulcerative colitis occasionally develop carcinoma in the biliary tree, and the risk is about 10 times greater than for the general population. Primary sclerosing cholangitis— whether or not associated with inflammatory bowel disease and various types of cystic disease of the biliary tree such an congenital hepatic fibrosis, polycystic disease of the liver, and Caroli's disease—may all be complicated by the development of malignant change. Unlike hepatocellular carcinoma neither long-standing HBV or HCV infection nor cirrhosis seem to predispose to cholangiocarcinoma.

Signs and symptoms

In the peripheral intrahepatic type, patients present with upper abdominal pain, anorexia, malaise, and weight loss. With hilar tumours, jaundice is an early feature. Hepatomegaly is usual and splenomegaly may be found if secondary biliary cirrhosis develops owing to prolonged biliary obstruction.

Diagnosis

The liver function tests show cholestatic features with elevation of bilirubin and alkaline phosphatase levels. Alpha-fetoprotein concentrations are usually normal or only slightly raised. Levels of Ca 19.9 and carcinoembryonic antigen (**CEA**) may also be raised, although sensitivity and specificity do not approach 100 per cent as raised levels are found in obstructive jaundice due to other causes.

Ultrasonography and CT scanning may demonstrate the tumour mass and with hilar tumours show dilatation of the intrahepatic biliary tree. On hepatic angiography the tumour tends to be avascular but encasement and occlusion of vessels occurs. Biliary tree obstruction in the hilum may be demonstrated on MRI cholangiography or by endoscopic retrograde cholangiography (**ERCP**) prior to insertion of a stent (see below).

Prognosis

Most patients deteriorate progressively, with an average survival from diagnosis around 12 to 18 months. If biliary drainage can be achieved in patients with hilar tumours, the prognosis may be better, for these tumours are often slow growing.

Treatment

For the peripheral tumours the principles of treatment are the same as for hepatocellular carcinoma (see above). The response to therapy is disappointing and no controlled trials have been reported. Hilar tumours may sometimes be suitable for curative resection with anastomosis of a Roux loop of jejunum to the biliary tree in the hilum. More usually curative excision is not possible, and the aim must be to establish biliary drainage. A stent can be placed through the growth at laparotomy, or at ERCP, or via the percutaneous transhepatic route thus avoiding surgery. The use of self-expanding metal stents is a recent advance.

Conventional radiotherapy and high-dose local irradiation within the biliary tree by means of iridium-192 wire may sometimes produce useful symptomatic relief. If biliary drainage can be achieved by these procedures, survival for 1 to 2 years is not unusual. Because of the high risk of tumour recurrence, liver transplantation is seldom indicated.

Angiosarcoma (Kupffer-cell sarcoma)

This is a rare tumour consisting of spindle-shaped malignant endothelial cells. It is often multifocal and may arise in a cirrhotic liver.

Considerable progress has been made in identifying aetiological agents. Like hepatocellular carcinoma and cholangiocarcinoma, it occurs in patients who were exposed to Thorotrast 15 to 25 years earlier, and chronic exposure to arsenic has also been implicated. More recently the tumour has been found in workers in the vinyl chloride industry, particularly those exposed to high concentrations of vinyl chloride monomer while cleaning the autoclaves. Since this discovery strict safety regulations have been introduced but because of the long latent period new cases continue to present. A few cases have occurred in long-term androgen takers, but no aetiological factor has yet been identified in the majority of cases. As with other liver tumours, patients present with abdominal pain and hepatic enlargement and blood-stained ascites is common.

This is a highly malignant tumour and curative resection is rarely possible. No form of palliative treatment has so far proved effective.

Epithelioid haemangioendothelioma

This is a rare malignant tumour of vascular endothelial origin with a characteristic histological appearance. It is usually multifocal. It occurs in younger patients than angiosarcoma and does not have the same aetiological associations. The tumour is usually slow growing and prolonged survival has been reported after resection and liver transplantation.

Other primary malignant tumours

These are extremely rare and include fibrosarcoma, leiomyosarcoma, and lymphoma. Children develop both hepatoblastoma and hepatocellular carcinoma.

Hepatic metastases

The liver is a favoured site for metastatic spread and about 50 per cent of malignant tumours in the portal venous drainage area eventually gives rise to hepatic metastases.

Diagnosis

The diagnosis is easy when physical examination reveals a large nodular liver but detection of small or solitary deposits is difficult. Liver function tests may be normal, but the alkaline phosphatase usually rises as the tumour mass enlarges. Ultrasound scanning should pick up tumours greater than 1 cm in diameter but accuracy is greatest when the metastases are large or numerous. The diagnosis can be confirmed by targeted liver biopsy at the time of laparoscopy or ultrasonography.

Prognosis

The prognosis is obviously worse when there is extensive liver replacement by tumour with severe disturbance in liver function tests or ascites. The site of the primary growth is also relevant and deposits from colorectal cancer have a better prognosis (untreated mean survival 9 to 12 months) than most other tumours.

Treatment

The range of possible treatments is the same as has been discussed for hepatocellular carcinoma. Partial hepatectomy to remove deposits may occasionally lead to prolonged survival or cure and, as mentioned above, the results are best in patients with colorectal cancer. A special situation exists with respect to hepatic metastases from the carcinoid tumour. This is often a slow-growing neoplasm and the main problem is the distress caused by flushing and diarrhoea. Resection of tumour bulk with no attempt at total removal often gives symptomatic relief for some years, as does embolization. Transplantation should be considered for slow-growing tumours.

The choice of chemotherapy will be determined by the origin of the primary tumour and this will not be discussed here. As with hepatocellular carcinoma the poor results with systemic chemotherapy led to trials with intra-arterial perfusion. Such treatment has been simplified by the development of implantable pumps, but while objective tumour regression may occur with improved quality of life, as yet there is little convincing evidence that survival is prolonged.

Benign tumours

Haemangioma

This is the most common benign tumour and is usually asymptomatic, being found incidentally either during ultrasonography or CT scanning. Occasionally when large it may cause abdominal pain or shock due to rupture leading to surgical excision.

Although the appearances with ultrasonography are usually diagnostic, it may be necessary to proceed to CT scanning with contrast, angiography, or MRI for diagnostic certainty.

Hepatic adenoma

The incidence of this tumour seems to have increased markedly since the introduction of the oral contraceptive pill, with most reported cases having occurred in females who have been on the pill for 5 years or more. It should be emphasized, however, that the risks for the individual woman is infinitesimal. Patients are often asymptomatic and a mass is discovered on physical examination or incidentally on ultrasound examination. Some patients complain of upper abdominal pain and others present acutely with shock due to intraperitoneal bleeding. The tumour is usually solitary but may be multiple. It consists of cords or acini of hepatocytes without bile ducts or portal tracts, and fibrous tissue septa are sparse. It may be encapsulated. There is little or no disturbance in liver function and alpha-fetoprotein concentrations are normal.

Ultrasonography shows a focal lesion of variable echogenicity, CT scanning shows marked arterial enhancement and sometimes areas of haemorrhage within the tumour which can be confirmed on MRI scanning.

In some cases the tumour has regressed after withdrawal of the contraceptive pill, but surgical resection is usually recommended because of the risk of intraperitoneal bleeding and the occasional development of malignant change.

Focal nodular hyperplasia

This is a benign condition of uncertain pathogenesis that is frequently confused with hepatic adenoma. The lesion is composed mainly of hepatocytes and Kupffer cells. Typically it has a central stellate scar with radiating septa containing arterial and venous channels and bile ductules. It is much more frequent in women than men, but no relationship to the oral contraceptive pill has been established. The mass is usually solitary and asymptomatic but rupture with intraperitoneal bleeding occasionally occurs. The findings on imaging are often different from those of hepatic adenoma and may allow definitive diagnosis. In particular Doppler ultrasound may show an arterial signal within the tumour, CT scanning or MRI scanning may demonstrate the central stellate scar and biliary scintiscanning may show a late hotspot in the tumour.

The prognosis is excellent and malignant change is not recorded. If the imaging techniques do not provide a definite diagnosis, however, surgical excision is often recommended.

Other benign tumours

These are very much rarer and include fibroma, lipoma, leiomyoma, and cystadenoma.

Further reading

Cherqui D, *et al*. (1995). Management of focal nodular hyperplasia and hepatocellular adenoma in young women: a series of 41 patients with clinical, radiological, and pathological correlations. *Hepatology* **22**, 1674–81.

Clavien P-A, ed. (1999). *Malignant liver tumours. Current and emerging therapies*. Blackwell Science, Malden, MA.

De Groen PC, *et al*. (1999). Biliary tract cancers. *New England Journal of Medicine* **341**, 1368–78.

Mathurin P, *et al*. (1998). Review article: overview of medical treatments in unresectable hepatocellular carcinoma—an impossible meta-analysis? *Alimentary Pharmacology and Therapeutics* **12**, 111–26.

Schafer DF, Sorrell MF (1999). Hepatocellular carcinoma. *Lancet* **353**, 1253–7.

Vauthey J-N, *et al*. (1996). Arterial therapy of hepatic colorectal metastases. *British Journal of Surgery* **83**, 447–55.

Williams R, Rizzi P (1996). Treating small hepatocellular carcinomas. *New England Journal of Medicine* **334**, 728–9.

14.21.6 Hepatic granulomas

C. W. N. Spearman, P. de la Motte Hall, and S. J. Saunders

Introduction

Granulomas are localized collections of modified macrophages, known as 'epithelioid' cells, that have become transformed from a predominantly phagocytic cell to a more secretory cell in response to ingested antigens. The epithelioid cells, which are derived from blood monocytes, have abundant amounts of eosinophilic cytoplasm. Langhans' or foreign body-type giant cells, which form by fusion of the epithelioid cells, are often seen in granulomas. Granulomas are usually surrounded by a rim of mononuclear cells, predominantly lymphocytes. Granulomas may be progressively replaced by collagen.

The aetiology of hepatic granulomas varies with the patient population and geographical origin. In the developed world, sarcoidosis, primary biliary cirrhosis, and drug-induced hepatic granulomas probably account for most, whilst infectious causes such as mycobacterial infections, schistosomiasis, and AIDS-related infections predominate in the developing world.

The diagnosis of granulomatous hepatitis is usually made during the investigation of a systemic illness, frequently presenting as a pyrexia of unknown origin (**PUO**). However, granulomas are found in 10 to 15 per cent of all liver biopsies and may be an unexpected finding. The histomorphology of the granuloma, their distribution in the liver, and special stains, for example Ziehl–Neelsen for mycobacteria and a methenamine silver for fungi, may yield a definite diagnosis.

Pathogenesis

Granuloma formation represents a specialized cellular immune-mediated response involving the presentation of antigen, either endogenous or exogenous, by activated macrophages to CD4 lymphocytes, which are in turn activated by the secretion of macrophage-derived interleukin-1 (IL-2). The activated CD4 lymphocytes secrete interferon, resulting in the upregulation of MHC class II molecules on the surface of the activated macrophages. Upregulation of the HLA DR-positive macrophage and the resulting increased interaction with stimulated CD4 lymphocytes, is accompanied by the consequent increase in IL-2 receptor expression and IL-2 secretion. This results in a clonal increase in the CD4 lymphocytes, leading to the recruitment of B cells which are activated and produce immunoglobulins, antibodies, and autoantibodies. Persistent antigenaemia or poorly degradable antigens, such as chemicals or toxins, provide an ongoing stimulus for the cytokine cascade which results in the focal accumulation of activated lymphocytes and macrophages, with the macrophages undergoing epithelioid transformation.

In infections, the micro-organisms, together with their by-products, are the sensitizing exogenous antigens, whereas in malignancy, or immune complex disease, sensitizing endogenous antigens may trigger an interaction between the activated macrophages and lymphocytes.

Depending on the cause, differences are seen in the above-described structural/functional arrangement of the granuloma. The well-studied sarcoid granuloma has a central core of HLA DR-positive macrophages, epithelioid cells, and giant cells with a peripheral rim of CD4 lymphocytes. The macrophages surrounding the central core are distinguishable from those in the centre by their reactivity with the macrophage monoclonal antibody RFD-1 as opposed to RFD-2. CD8 suppresser cells, and some CD4 cells, may be found at the periphery but not in the centre of the granuloma. The epithelioid cells in sarcoid granuloma secrete a number of compounds, including angiotensin-converting enzyme, lysozyme, glucuronidase, collagenase, elastase, and calcitriol. In AIDS patients with *Mycobacterium avium intracellulare* (**MAI**) infection, there is a paucity of CD4 lymphocytes in the granulomas. While in granulomas infected with *Schistosoma mansoni*, the CD4 cells show increased Th2 co-operation. The granulomas in tuberculoid leprosy contain very few bacilli, while bacilli are profuse in lepromatous leprosy.

Aetiology

The many causes of hepatic granulomas are shown in Table 1. Granulomas are frequently non-specific in appearance and a clinicopathological correlation is essential for diagnosis. In 10 to 30 per cent of hepatic granulomas, the aetiology remains unknown. However, caseating granulomas (Fig. 1 and Plate 1) are characteristic of mycobacterium tuberculosis; the presence of ova and non-caseating granulomas permit the diagnosis of schistosomiasis; fat droplets are seen in the granuloma (lipogranuloma) that accompanies mineral oil ingestion; and fibrin-ring granulomas are highly suggestive of Q fever.

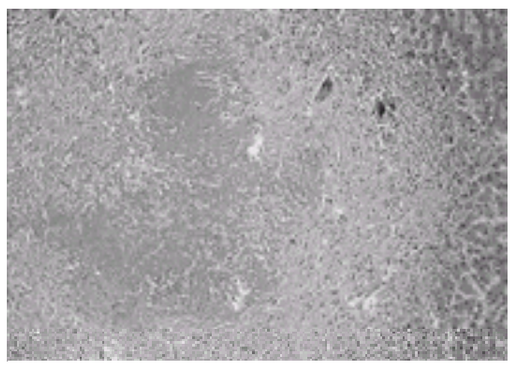

Fig. 1 Liver showing a portion of a large caseating granuloma from a patient with miliary mycobacterium tuberculosis. Several Langhans' giant cells are also seen. (Haematoxylin and eosin, total magnification 25.) (See also Plate 1.)

Clinical presentation

Fever (PUO) is the most common presenting symptom. The diverse clinical features, which include weight loss, anorexia, fatigue, hepatosplenomegaly, and abdominal pain, depend on the underlying aetiologies. Serum transaminase activities are frequently normal, but alkaline phosphatase activity may be elevated. Jaundice is uncommon unless there is bile-duct injury, for example primary biliary cirrhosis and sarcoidosis.

Infectious causes

Mycobacterium tuberculosis infection

Tuberculosis is a common cause of hepatic granuloma in the developing world. There may be evidence of pulmonary tuberculosis or of tuberculosis elsewhere, or the patient may present with a pyrexia of unknown origin. Hepatic granulomas are common in miliary tuberculosis. Caseation occurs in about one-third of cases but, while characteristic of tuberculosis, is not unique to it, occurring also in candidiasis, histoplasmosis, and cryptococcosis. Acid-fast bacilli are seen in 10 to 15 per cent of cases, and in 31 per cent of autopsy specimens, and therefore the biopsy must always be cultured. More recently, use of the polymerase chain reaction (**PCR**), based on amplification of IS6110 insertion sequences, was shown to have a sensitivity of 58 per cent in the diagnosis of hepatic granulomas of definitive tuberculosis origin and a specificity of 96 per cent, and is a useful test as it can be performed on paraffin-embedded tissue. BCG vaccination may also cause hepatic granulomas.

HIV/AIDS

Currently more than 30 million people worldwide test positive for the human immunodeficiency virus (**HIV**). Opportunistic infections, especially *Mycobacterium avium intracellulare* (**MAI**), are an important cause of hepatic granulomas in patients with AIDS. The MAI granulomas are composed of epitheliod cells with striated bluish cytoplasm, due to the presence of large numbers of micro-organisms (Fig. 2 and Plate 2).

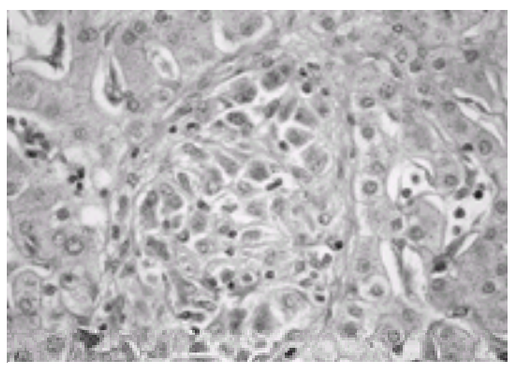

Fig. 2 Liver from an HIV/AIDS patient infected with *Mycobacterium avium intercellulare*, showing a granuloma composed of epithelioid cells which contain large numbers of micro-organisms. (Diastase/periodic acid–Schiff stain, total magnification 100.) (See also Plate 2.)

Histoplasmosis, cryptococcosis, toxoplasmosis, and cytomegalovirus may also cause granulomas in AIDS patients. However, an infectious cause is not always found, presumably there are as yet unidentifiable infections.

Drugs, Hodgkin's disease, and non-Hodgkin's lymphoma may also cause hepatic granulomas in AIDS patients.

Leprosy

Millions of people living in the Indian subcontinent have leprosy and this disease is also common in Africa. Hepatic granulomas are more common in lepromatous leprosy. The diagnosis is usually made from the characteristic skin and peripheral nerve lesions, and only occasionally, the physician is alerted to the diagnosis by the finding of an otherwise unexplained hepatic granuloma.

Histoplasmosis

Histoplasmosis is an important cause of hepatic granuloma in the United States. The fungus may be seen in the granulomas and may be cultured from liver biopsies, blood, or bone marrow. The chest radiograph is usually abnormal and the diagnosis is confirmed serologically.

Q fever

The patients usually present with a PUO or an illness resembling viral hepatitis. The typical granuloma contains inflammatory cells and fat droplets and has a fibrin ring within or at its margin. However, fibrin-ring granulomas may be seen in patients with cytomegalovirus infection, Hodgkin's disease, and leishmaniasis, and in drug reactions to, for example, allopurinol.

Schistosomiasis

Schistosoma japonicum and *Schistosoma mansoni* infestation occurs commonly in Africa, South America, and in the Far East. Ova are usually deposited in the portal tracts within the portal vein radicles where a granulomatous reaction occurs around the ova. Presinusoidal portal hypertension may occur with a 'pipe-stem' cirrhosis. Common presenting features are hepatomegaly and portal hypertension. Eosinophilia also occurs.

Hepatitis C virus (HCV)

Sparse non-caseating hepatic granulomas have been described in patients with chronic hepatitis C infection and in liver-transplanted patients who have subsequent recurrent HCV infections. There is controversy as to whether the presence of hepatic granulomas is predictive of a favourable response to interferon therapy.

Non-infective causes

Sarcoidosis

This is the most common non-infectious cause of hepatic granuloma formation. Although thought to be immunologically mediated, the triggering factor remains unknown. Hepatic granulomas are found in 60 per cent of liver biopsies performed on patients with sarcoidosis.

Well-formed, non-caseating granulomas occur in clusters in the parenchyma and the portal tracts. There may be associated portal fibrosis, which in some patients, progresses though bridging fibrosis to cirrhosis. Other features that are sometimes seen include bile duct damage and loss, which needs to be differentiated from primary biliary cirrhosis, other causes of ductopenia, and acute cholangitis associated with bile-duct obstruction. A lobular hepatitis, portal inflammation with interface hepatitis may also occur. The development of nodular regenerative hyperplasia as well as fibrosis and cirrhosis may be associated with granulomatous phlebitis of the portal and hepatic veins.

Hepatic sarcoidosis is frequently asymptomatic. While significant hepatomegaly is uncommon, splenomegaly is often present. The liver enzymes may be normal or

there may be elevated levels of alkaline phosphatase and aminotransaminases. Portal hypertension and cholestasis are rare complications occurring in only 1 in 300 European patients, but appear to be more common in the United States, and particularly in Black males. Patients may present with intrahepatic cholestasis and later cirrhosis. Differentiation from primary biliary cirrhosis is often difficult. The portal hypertension is presinusoidal and is thought to be due to granulomatous involvement of the portal venous radicals. These patients may present with variceal bleeds in the absence of cirrhosis.

Treatment with steroids often results in the reduction of the size of the hepatosplenomegaly and in improvement in liver enzymes. Steroids have little effect on portal hypertension but may benefit patients with intrahepatic cholestasis.

Primary biliary cirrhosis

Hepatic granulomas are found in approximately 25 per cent of patients with primary biliary cirrhosis (**PBC**). The granulomas are found in portal tracts and tend to surround damaged bile ducts. Occasionally, granulomas are seen in the parenchyma. The granulomas are seen in the early stages of PBC. PBC has a female to male ratio of 8:1, tends to present in the fifth decade of life, and usually has no extrahepatic systemic manifestations. Pruritis tends to be marked, clubbing and hepatomegaly is frequent. These clinical features help to differentiate PBC from sarcoidosis. The latter has an equal female to male ratio, occurs in the third and fourth decades of life, and frequently has extrahepatic manifestations such as erythema nodosa, uveitis, pulmonary involvement, hilar adenopathy, and abnormalities of calcium metabolism with hypercalcaemia and hypercalciurea. Antimitochondrial antibodies are positive in over 95 per cent of patients with PBC. The serum angiotensin-converting enzyme activity test may be elevated in both conditions.

Neoplasia

Non-caseating hepatic granulomas occur with a variety of neoplasms, including lymphoma and carcinoma. These granulomas may occur in the absence of tumour deposits and may represent an immune response to tumour antigens.

Lymphoma

Hepatic granulomas may occur in both Hodgkin's and non-Hodgkin's lymphoma. Granulomas may mask the infiltrates of malignant lymphoma and in Hodgkin's disease. Immunohistochemistry may demonstrate a clonal lymphocytic population in or around the granulomas associated with lymphoma.

Chronic granulomatous disease of childhood

This is a classical X-linked disorder, usually presenting at about 5 years of age with hepatosplenomegaly, generalized lymphadenopathy, granulomatous skin lesions, and diffuse miliary lung infiltration. The neutrophils of children with chronic granulomatous disease are unable to kill ingested bacteria, as they are deficient in those enzymes required for the superoxide respiratory burst. The diagnostic test for this condition is the inability of neutrophils to reduce nitroblue tetrazolium from colourless to blue-black formazan granules in their cytoplasm in the presence of a bacterial infection.

Crohn's disease

Non-caseating granulomas are found in the intestine, perineum, lip, and, occasionally, in the liver. The clinical manifestations are typical of Crohn's disease and diagnosis is not usually problematic.

Hepatic granulomatous disease

Hepatic granulomatous disease is a diagnosis of exclusion. The granulomas are seen in the portal tracts and lobules but there is no evidence of hepatitis. There appear to be two variants: one an acute febrile illness characterized by respiratory symptoms, a high white-cell count, and splenomegaly; and the other a more chronic condition presenting more frequently in middle-aged men with recurrent fevers, rigor, sweating, general malaise, and loss of weight. Neuralgia and arthralgia are common, there may be mild, tender hepatomegaly, and liver enzyme abnormalities are non-specific, including bilirubinaemia, mild elevation of transaminases, as well as elevated alkaline phosphatase levels. There is usually a good response to steroid therapy, with pyrexia resolving and liver enzymes improving. The cause is unknown.

Drugs and chemicals

Many drugs cause granulomatous hepatitis (Table 2). Drug-induced liver injury is frequently due to a hypersensitivity reaction. Although drug reactions may result in eosinophil-rich granulomas, in most cases the granulomas are non-specific.

Drug-induced hepatic granulomas may be completely asymptomatic or present with features suggestive of a drug allergy with a swinging fever, skin rash, eosinophilia, and abnormal liver enzymes. Although the diagnosis of drug-induced granulomatous hepatitis is one of exclusion, drugs should always be considered when granulomas are found in the liver.

Exposure to various chemicals such as beryllium, silicone, starch, talc, and suture material has also been associated with the development of hepatic granulomas.

Investigation and management of hepatic granulomas

To obtain the correct diagnosis and determine the appropriate therapy, it is important always to consider the epidemiological background of the patient. In the Western world, the common causes of granulomas in the liver are sarcoidosis, PBC, drugs, and neoplasms, whilst in patients in developing countries infectious causes must be considered first and excluded before considering others. However, the global problem of HIV/AIDS may well place mycobacterial and other infections as the most frequent causes of granulomatous hepatitis.

Further reading

Barceno R, *et al*. (1998). Post-transplant liver granulomatosis associated with hepatitis C. *Transplantation* **65**, 1494–5.

Denk H, *et al*. (1994). Guidelines for the diagnosis and interpretation of hepatic granulomas. *Histopathology* **25**, 209–18.

Devaney K, *et al*. (1993). Hepatic sarcoidosis. Clinicopathologic features in 100 patients. *American Journal of Surgical Pathology* **17**, 1272–80.

Diaz ML, *et al*. (1996). Polymerase chain reaction for the detection of *Mycobacterium tuberculosis* DNA in tissue and assessment of its utility in the diagnosis of hepatic granulomas. *Journal of Laboratory and Clinical Medicine* **127**(4), 359–63.

Emile JF, *et al*. (1993). The presence of epithelioid granulomas in hepatitis C virus-related cirrhosis. *Human Pathology* **24**, 1095–7.

Farrell GC (1995). Drug-induced granulomatous hepatitis. *Drug-induced liver disease*, pp 301–17. Churchill Livingstone, Edinburgh.

Goldin RD, *et al*. (1996). Granulomas and hepatitis C. *Histopathology* **28**, 265–7.

Ishak KG (1998). Sarcoidosis of the liver and bile ducts. *Mayo Clinic Proceedings* **73**, 467–72.

Lee RG, *et al*. (1981). Granulomas in primary biliary cirrhosis: a prognostic feature. *Gastroenterology* **81**, 983–6.

Lefkowitch JH (1999). Hepatic granulomas. *Journal of Hepatology* **30**, 40–5.

O'Connell MJ, *et al*. (1975). Epithelioid granulomas in Hodgkin's disease: a favourable prognostic sign. *Journal of the American Medical Association* **233**, 886–9.

Simon HB, Wolff SM (1973). Granulomatous hepatitis and prolonged fever of unknown origin: a study of 13 patients. *Medicine* (Baltimore) **52**, 1–20.

14.21.7 Drugs and liver damage

J. Neuberger

Introduction

Drug-induced liver injury is relatively uncommon, but unless it is recognized early and the drug discontinued it may cause death. Adverse drug reactions are responsible for between 0.1 and 3 per cent of all hospital admissions. Reliable data are difficult to come by: a relatively recent study has shown that in 1986 to 1987 about 1600 cases per year of adverse drug reactions were reported in England. Hepatic reactions accounted for 3.5 per cent of which 7 per cent were fatal. Similar figures were found by Pillans in New Zealand. A total of 205 drugs were associated with 943 reports of adverse liver injury between 1974 and 1994: 20 drugs accounted for nearly 60 per cent of reports. Most reactions are of jaundice and hepatitis; the more common are due to antibiotics and non-steroidal anti-inflammatory drugs. Halothane, perhexiline, and erythromycin were common causes of death and diclofenac, augmentin, and flucloxacillin were the most important causes of liver damage. In Denmark, Dossing and colleagues estimated that drug-induced liver injury accounts for between 1 in 600 and 1 in 3500 hospital admissions, amounting to 2 to 3 per cent of all hospital admissions due to adverse reactions and about 3 per cent of all jaundiced patients. In general practice, the spectrum of liver damage is slightly different: drugs associated with a high incidence of acute liver injury (greater than 100 per 100 000 users) were chlorpromazine and isoniazid; drugs with intermediate incidence of acute liver damage (more than 10 per 100 000 users) were amoxicillin/clavulanic acid and cimitedine.

Drug-induced liver damage may be caused by agents not considered as conventional drugs, such as herbal remedies, and by 'recreational drugs' such as ecstasy (methylenedioxymethampfetamine) and cocaine (see Box 1). The wide regional and individual variation in reporting rates and failure to report reactions after deliberate overdose combine to underestimate the frequency and severity of adverse drug reactions. Most adverse drug reactions are not fatal and withdrawal of the drug will usually lead to resolution of the liver damage. A study reporting the outcome of 110 cases of presumed drug-induced liver damage found a significant proportion of continuing liver damage, sometimes associated with continuing use of the hepatotoxic drug.

Box 1 The 'rules' of drug-induced liver disease

- Assume that all drugs may cause liver damage
- All patterns of liver damage have been associated with drug toxicity
- Some drugs may cause more than one pattern of liver damage
- Always take a full drug history
- Ask about other drugs—including herbal remedies, recreational drugs, vitamins
- The diagnosis of an adverse drug reaction is one of exclusion and temporal relationship
- Drug withdrawal is not always associated with improvement in liver function
- Reports of drug-associated liver damage do not necessarily mean causality
- Clinical challenge is rarely justified, may be fatal, and may be misleading

Almost all patterns of liver disease can be induced by drugs (Table 1) and some drugs may be associated with more than one type reaction. For example, oral contraceptives are associated not only with the development of cholestasis but also with adenoma, hepatocellular carcinoma, peliosis hepatis, and Budd–Chiari syndrome. It is important therefore to consider the possible contribution of drugs in a patient with any type of hepatic abnormality.

The diagnosis of drug-induced liver damage is largely circumstantial and by exclusion of other causes of liver disease. It must be remembered that the reporting of an associated drug reaction does not prove causality. The temporal association between the onset of damage and timing of drug exposure, and the response to drug withdrawal (Table 2) and the known patterns of drug reaction all help in establishing a drug as the cause of liver damage. Rarely, the presence of specific serological markers may help confirm the association between the drug and liver damage. For example, an antibody to tifluoroaceylated proteins is found in halothane-associated hepatitis, and antiliver–kidney microsomal antibodies occur in tienilic acid-associated hepatitis. Use of a clinical challenge is rarely justified, may be rnisleading, and may prove fatal.

Acute hepatitis

The severity of liver cell necrosis associated with drugs varies from a mild elevation of serum transaminases without symptoms to fulminant hepatic failure. Many drugs have been associated with acute liver failure (Table 3). Clinically, the picture may be indistinguishable from that of viral hepatitis. Occasionally, right upper quadrant pain may he so severe as to lead to the mistaken diagnosis of acute cholecystitis. The serological changes are those of acute hepatitis with initial elevations of serum aminotransferases. Prolongation of the prothrombin time and jaundice may occur in more severe cases. Histologically, the appearances vary from a mild focal necrosis to massive liver cell damage. In some cases, paracetamol for example, the damage is predominantly centrilobular, whereas in others, such as a-methyldopa, the whole lobule is affected. Steatosis, granulomas, and eosinophilia are variable features. The most common causes of drug-associated fulminant hepatic failure are paracetamol overdose and halothane hepatitis. Liver failure may also be associated with 'recreational' drugs.

The development of abnormalities of liver tests during prolonged drug use poses particular problems, as for example with antituberculous therapy. Derangement of serum aminotransferases occurs in approximately 10 per cent of patients and, if the noxious drug is continued, up to l0 per cent of these develop severe hepatic necrosis. Identification of those patients who will develop severe hepatic failure is difficult and the clinician has to decide whether the risks of continuing therapy outweigh the potential benefits. Drugs such as heparin are commonly associated with abnormal liver enzymes but very rarely with liver disease. The reason is not known but it may be due to loss of a few sensitive hepatocytes or to adaptation.

Conventionally, hepatic drug reactions are classified into predictable and idiosyncratic (Table 4). Predictable reactions are dose dependent; that is, the greater the amount of drug ingested, the greater the probability of developing liver damage. Because animal models can usually be developed, screening will detect many of these drug reactions and the drug withdrawn before reaching the market. Hence this type of drug reaction is relatively uncommon, except in overdose. The classic example is paracetamol toxicity, which is described in detail elsewhere.(Irene: cross reference?) None the less, between individuals there may exist great variability in the probability of developing predictable drug reactions.

With very few exceptions, drugs require metabolism before cytotoxicity develops. Variations in susceptibility may, therefore, be a consequence of genetic variations in drug metabolism. Well-recognized genetic polymorphisms include variations in the cytochrome P450 isoenzymes, drug oxidation, acetylation, and hydroxylation. Age, too, is associated with differences in susceptibility to toxicity. In general, younger children metabolize drugs differently from adults. Those taking enzyme inducers

such as alcohol, rifampicin, or phenobarbital are at a greater risk of increased metabolism of the drug and hence of forming toxic metabolites. Those with reduced glutathione stores, due to fasting, malnutrition, or associated disease for example, may be at greater risk of developing paracetamol toxicity because detoxification mechanisms are impaired. Other factors determining susceptibility include smoking and coexisting diseases, so that, for example, methotrexate toxicity is more common in those with diabetes. Finally, liver disease itself may alter susceptibilities to drug toxicity. However, because of potential alterations in absorption, volume of distribution, protein binding, detoxification, and excretion it is difficult to predict the effect of disease on susceptibility to drug toxicity. Many drugs induce hepatitis by apoptosis which may be accompanied by simultaneous or secondary necrosis.

In contrast, idiosyncratic drug reactions are dose independent and may be due either to metabolic idiosyncrasy or the involvement of immune mechanisms. Immune involvement rather than metabolic idiosyncrasy is suggested by a rapid onset after subsequent exposure and the appearance of markers such as peripheral and intrahepatic eosinophilia, granulomas, circulating immune complexes, autoantibodies, and other autoimmune phenomena, for example haemolytic anaemia. Two drugs in particular have been well studied with respect to immune-mediated hepatitis—halothane and tienilic acid. Halothane hepatitis occurs rarely and after multiple exposures. Risk factors include female sex, obesity, and repeated or subsequent exposure within 3 months. Immune involvement is suggested by an increased incidence of organ non-specific autoantibodies, peripheral eosinophilia, and circulating immune complexes, and the presence of antibodies reacting with a variety of halothane-associated liver cell macromolecules. In other examples. antibodies to drug-metabolizing enzymes are present in serum. Tienilic acid-associated hepatitis is associated with a circulating liver–kidney microsomal antibody that reacts with the cytochrome P450, CYP 2C9, associated with metabolism of the drug; antibodies to CYP 1A2 are associated with hydralazine and disulfiram hepatitis; alcohol and halothane hepatitis are associated with antibodies to CYP 2E1. Iproniazid hepatitis is associated with antibodies to MAO-B. Whether these antibodies are involved in the pathogenesis of the disease remains uncertain.

Cross-reaction between two drugs may occur. Thus, halothane sensitization may predispose to toxicity from other halogenated hydrocarbon anaesthetic agents such as isoflurane. This may be due to the two drugs inducing similar antigenic determinants, leading to cross-sensitization, or to a different mechanism of toxicity, as suggested for captopril and enalapril hepatotoxicity, where a similar metabolic pathway of toxicity has been postulated.

Acute cholestatic hepatitis

Acute cholestatic hepatitis is characterized by jaundice, pruritus, pale stools, and dark urine. There are usually few clinical findings, although the liver may be enlarged. Serologically, in the early stages there is elevation of the serum alkaline phosphatase and g-glutamyl transpeptidase; as the disease progresses, hepatocellular enzymes start to rise. Histologically, the liver shows dilated sinusoids with cholestasis often predominating in the centrilobular region. There may be an associated portal inflammation and liver cell necrosis. In the majority of cases there is rapid resolution following withdrawal of the drug, although with chlorpromazine and other phenothiazines the cholestasis may take up to 1 to 2 years to resolve. Many drugs cause a mixed hepatitis, where there are features both of cholestasis and liver cell damage (Table 5).

Bland cholestasis

Bland cholestasis is characterized by cholestasis in the absence of hepatitis and is due to specific interference with bile secretion. The two main groups of drugs associated with this condition are oral contraceptives and oestrogens and anabolic steroids. Cholestasis occurs in women taking oral contraceptives and in pregnancy. Prevalence varies, being low in southern Europe and North America (1 in 10 000) and high (1 in 4000) in parts of Chile and Scandinavia. Cholestasis associated with anabolic and contraceptive steroids is well recognized and may occur in association with virtually all the anabolic steroids with a C17 group; these drugs include norethandrolone, oxymethalone, danazol, stanozalol, and methyltestosterone. Other drugs are listed in Table 6. In some cases, drug induced cholestasis leads to a progressive, vanishing bile duct syndrome. Treatment is symptomatic: the itching may be intense and sometimes responds to cholestyramine or colestipol; other therapies include antihistamines, ursodeoxycholic acid, rifampicin, androgenic anabolic steroids, and opiate receptor antagonists.

Steatosis

Steatosis may be micro- or macrovesicular. Differentiation is important because the clinical features and outcomes are different. (Table 7).

Microvesicular steatosis

In microvesicular steatosis, the fat is distributed in small lipid droplets and the hepatocellular nucleus is not displaced. There may be an associated hepatitis. Extensive microvesicular steatosis, even in the absence of liver cell necrosis, may lead to a serious clinical syndrome with haemorrhage, syncope, hypotension, lethargy, coma, and hypoglycaemia. In some cases, renal failure and pancreatic inflammation may occur. Biochemically, serum aminotransferases and bilirubin are not greatly increased, although the prothrombin time may be greatly prolonged. Microvesicular steatosis is thought to be related to drug inhibition of mitochondrial oxidation of fatty acids.

Macrovesicular steatosis

In contrast, macrovesicular steatosis is usually far less serious. The hepatocyte contains a large droplet of fat, which displaces the nucleus to the periphery. Liver tests are usually only minimally deranged. Damage is thought to be related to impaired release of lipids from liver cells.

Granulomatous hepatitis

The spectrum of granulomatous hepatitis varies from an asymptomatic finding to a systemic illness characterized by generalized aches and pains, pruritus, jaundice, and hepatomegaly. Serologically. the main abnormality is an increase in serum alkaline phosphatase. Histologically the liver is infiltrated by granulomas—small, rounded foci of epithelioid cells with multinucleated giant cells. Drugs associated with granulomatous hepatitis are listed in Table 8.

Phospholipidosis

Phospholipidosis is characterized by the accumulation of phospholipids in liver cell lysosomes. The major drugs associated with this form of liver damage, perhexiline and amiodarone, are cationic, amphiphilic compounds that accumulate within the liver cell lysosomes where they form complexes with phospholipids. Accumulation can be detected by immunohistochemistry or electron microscopy. The compounds are stored in these complexes and may be released very slowly, even after ingestion has stopped. The extent to which these complexes accumulate in patients without toxicity remains uncertain.

Non-alcohol steatotic hepatitis

Long-term treatment with perhexiline and amiodarone may be associated with a syndrome that is clinically and histologically identical to alcoholic hepatitis. The disease develops insidiously and is characterized by hepatomegaly, jaundice, ascites, and encephalopathy. Other drugs implicated in this syndrome include diltiazem and nifedipine.

Fibrotic and vascular disease Table 9)

Perisinusoidal fibrosis

Perisinusoidal fibrosis is characterized by accumulation of collagen within the space of Disse. This may be asymptomatic or lead to hepatomegaly and portal hypertension. The most common causes of perisinusoidal fibrosis due to drugs are large doses of vitamin A given for prolonged periods, or methotrexate. Liver damage may be associated with alopecia. Characteristically the liver shows hyperplasia of the Ito cell as a consequence of vitamin A accumulation. Serum concentrations of vitamin A may he normal, even in the presence of marked liver damage. Patients with a high intake of alcohol are at greater risk of fibrosis.

Peliosis hepatis

Peliosis hepatis is a histological diagnosis and is characterized by blood-filled cavities, bordered by hepatocytes, which may be distributed throughout the liver.

Originally described in association with tuberculosis, it is now appreciated that peliosis hepatis may be drug induced and is often asymptomatic. The major drugs involved are the anabolic steroids, androgenic steroids, azathioprine, vinyl chloride, and pyrizolide derivatives.

Hepatic venous damage

Obstruction of the large hepatic veins results in the Budd–Chiari syndrome, characterized by the onset of abdominal pain and ascites, often with diarrhoea. In the acute form the patient may develop liver failure. Most cases of Budd–Chiari syndrome are due to myeloproliferative disorders, either clinically apparent or latent, but it may be associated with the use of oral contraceptives and some antineoplastic drugs such as dacarbazine, doxorubicin, and cyclophosphamide.

Obstruction of the small veins leads to hepatic veno-occlusive disease, characterized by non-thrombotic, concentric narrowing of the small centrilobular veins. Clinical presentation is often chronic but rarely may be acute. Veno-occlusive disease was initially described in association with ingestion of the pyrrolizidine alkaloids present in senecin plants but may be seen in patients treated with immunosuppressives, especially with organ transplantation.

Hepatic tumours

Hepatic tumours may be benign or malignant (Table 10). Hepatocellular adenoma has been associated with the use of oral contraceptives and anabolic steroids. These tumours have a potential for malignant transformation. Usually withdrawal of the steroid results in a reduction in the size of the tumour.

In contrast, hepatocellular carcinoma is also associated with the anabolic and androgenic steroids, oral contraceptives, and thorium dioxide. Although the risk of malignancy increases with the prolonged use of oral contraceptives, up to eightfold after 8 years, it must be emphasized that the overall risk of developing hepatocellular carcinoma with oral contraceptives is extremely small, and must be balanced against their beneficial, therapeutic effects. Angiosarcomas and cholangiosarcomas may also be related to drugs, although the association is less clear-cut.

Chronic disease

Cirrhosis and chronic hepatitis

Some drugs are associated with chronic liver disease. It may be that the initial lesions develop subclinically and that only prolonged use of the drug will result in cirrhosis. Rarely, a short-term exposure to a drug results in chronic liver disease. In some instances, there is a syndrome resembling autoimmune hepatitis: although corticosteroids may be given, withdrawal of the drug usually leads to resolution of the hepatic inflammation. Some of the drugs associated with the development of cirrhosis and chronic hepatitis are listed in Table 11.

Intrahepatic chronic cholestasis

In some instances of drug-related cholestasis, jaundice or cholestatic liver tests persist for 6 months or more (Table 12). In these cases it is important to exclude other causes of cholestatic disease, such as primary biliary cirrhosis or primary sclerosing cholangitis, which may have been brought to light by drug-induced disorders. However, some drugs may be associated with a chronic vanishing bile duct syndrome , which may be indistinguishable from primary biliary cirrhosis. A syndrome virtually identical to primary sclerosing cholangitis can be induced by infusion into the hepatic artery of floxuridine for the treatment of intrahepatic malignancy. Sclerosing cholangitis may develop several months after starting chemotherapy. The outcome is variable. A vanishing bile duct syndrome has been associated with carbamazepine, thiobendazole, flucloxacillin, haloperidol, ajmaline, cyproheptidine, and chlorpromazine. There has been a suggestion that primary biliary cirrhosis is associated with the use of benoxaprofen. The cause of the chronic cholestasis is uncertain; both immune mechanisms and the recirculation of toxic metabolites have been implicated.

Further reading

Aithal PG, Day CP (1999). The natural history of histologically proved drug induced liver disease. *Gut* **44**,731–5.

Bem JL, Msann R, Rawlins MO (1988). Review of yellow cards. *British Medical Journal* **296**, 1319.

Danan O (1990). Consensus meeting. Criteria of drug induced liver disorders. *Journal of Hepatotogy* **11**, 272–6.

Dossing M, Sonne J (1993). Drug-induced hepatic disorders. *Drug Safety* **9**, 441–9.

Friis H, Andreason P (1991). Drug induced hepatic injury: an analysis of 1100 cases reported to the Danish Committee on Adverse Drug Reactions between 1978 and 1987. *Internal Medicine* **232**, 133–42.

Garcia Rodriguez LA, Ruigomez A, Jick H (1997). A review of epidemiologic research on drug-induced acute liver injury using the general practice research data base in the United Kingdom. *Pharmacotherapy* **17**, 721–8.

Kaplowitz N, ed. (1990). Recent advances in drug metabolism and hepatotoxicity. *Seminars in Liver Disease* **10**, 235–338.

Lewis J, Zimmerman H (1989). Drug induced liver disease. *Medical Clinics of North America* **73**, 77–96.

Neuberger J (1989). Drug induced jaundice. *Clinical Gastroenterology*, **3**, 447–66.

Pessayre D *et al.* (1999). Withdrawal of life support, altruistic suicide, fratricidal killing and euthanasia by lymphocytes: different forms of drug-induced hepatic apoptosis. *Journal of Hepatology* **31**, 760–70.

Pillans PI (1996). Drug associated hepatic reactions in New Zealand: 21 years experience *New Zealand Medical Journal* **109**, 315–19.

Shaw D *et al.* (1997). Traditional remedies and food supplements. A 5 year toxicological study (1991–1995). *Drug Safety* **17**, 342–56.

Stricker NH, Spoelstra P (1985). *Drug induced hepatic injury.* Elsevier, Amsterdam.

Zimmer HJ (1990). Update of hepatoxicity due to class of drugs in common clinical use. *Seminars in Liver Disease* **10**, 322–33.

14.21.8 The liver in systemic disease

J. Neuberger

The liver is affected in many systemic diseases. It most instances, disturbance of liver structure and/or function is a minor component of the illness, but in some cases abnormalities of liver function may be the presenting symptom. This chapter describes abnormalities of liver function that occur in systemic diseases.

Cardiovascular disease

Congestive cardiac fallure

Most patients with congestive cardiac failure have few symptoms related to hepatic congestion, although nausea, vomiting, and right upper quadrant pain may occasionally occur. Hepatomegaly is frequent in moderately severe heart failure. Rarely, cardiac cirrhosis develops and may be associated with splenomegaly and ascites. Jaundice occurs in about one-quarter of patients with persistent hepatic venous congestion.

The standard serum liver-related tests may show a rise in bilirubin, which rarely exceeds 50 µmol/litre. The level of unconjugated bilirubin usually exceeds that of conjugated bilirubin. The serum aminotransferases may also be elevated but rarely exceed twice the upper limit of normal. However, in severe, acute heart failure, concentrations in excess of 1000 IU/litre may be found. Serum alkaline phosphatase is rarely elevated. The prothrombin time is often prolonged by a few seconds. The liver is usually enlarged and a cut section shows the classical nutmeg appearance, with the pale periportal zones alternating with darker centrilobular zones. Microscopically there is congestion, with dilatation of the terminal hepatic venules and adjacent sinusoids, and areas of centrilobular necrosis due to hypoperfusion injury. With chronic heart failure centrilobular necrosis may be associated with fibrosis.

Constrictive pericarditis

Hepatic complications of constrictive pericarditis occur late in the course of the illness. Cardiovascular features of constrictive pericarditis are described elsewhere (see Section 15). The liver is enlarged and there may be associated splenomegaly. Jaundice and ascites may develop. Ultrasonography will show enlargement of the liver with dilated hepatic veins.

Tricuspid incompetence

Tricuspid incompetence most commonly occurs as a result of failure of right heart dilatation but may also result from congenital or acquired disease of the tricuspid valve. The liver is enlarged and pulsatile.

Tumours of the heart

Tumours of the right atrium, including myxoma and myosarcoma, may infiltrate he hepatic veins resulting in Budd–Chiari syndrome (a syndrome of hepatic venous thrombosis, characterized by abdominal pain, progressive ascites, and diarrhoea). Cardiac myxoma may be associated with abnormalities of liver function tests, including increased serum bilirubin and alkaline phosphatase and a reduction in serum albumin and total protein.

Drug reactions

As described elsewhere (see Section 15), many drugs used for the treatment of heart disease may be associated with adverse reactions that involve the liver. Thus, chronic active hepatitis may be associated with methyldopa, and granulomatous hepatitis with hydralazine. Quinidine, amiodarone, and perhexiline may cause phospholipidosis and a syndrome resembling alcoholic hepatitis.

Hypoxia

Hypoxic episodes, especially during surgery, may lead to an acute liver injury resulting from an ischaemic hepatitis. The clinical severity ranges from an asymptomatic elevation of serum aminotransferases to fulminant hepatic failure. The syndrome may be followed by a period of cholestasis. The aminotransferases may become greatly elevated (in excess of 10 000 IU/litre); histologically there is hepatocellular necrosis in the absence of inflammation, most marked in acinar zone 3 (the perivenular area). Similar changes occur in patients with heat stroke.

Syndromes affecting both the heart and the liver

Several conditions affect both the heart and the liver; these include Alagille syndrome (a multisystemic disorder, associated with a paucity of intrahepatic bile ducts, leading to a biliary cirrhosis with pulmonary stenosis and other cardiac abnormalities), biliary atresia (where up to 10 per cent of patients may have congenital heart disease), and cardiomyopathy which may be associated with a variety of inherited and acquired conditions affecting both organs, these include alcohol toxicity, haemochromatosis, tyrosinaemia, and mitochondrial cytopathy. Drugs, such as the immunosuppressive agent tacrolimus, may also cause cardiomyopathy.

Pulmonary disease

Cirrhosis

Lung disease may complicate cirrhosis; the hepatopulmonary syndrome, discussed elsewhere, may resolve after treatment of the underlying liver disease or after liver transplantation. In contrast, abnormalities of liver function in patients with pulmonary disease arise either as a consequence of that disease or of diseases affecting both lung and liver. In most patients with chronic lung disease, abnormalities of liver function are mild and may be manifest only by abnormalities of bromosulphophthalein clearance. In more advanced disease, associated with hypoxia, there may be more widespread disturbances of liver function, with elevation of serum aminotransferase, bilirubin, alkaline phosphatase, and g-glutamyltransferase. However, abnormality of liver function in patients with pulmonary disease is associated mainly with pulmonary hypertension rather than lung disease or hypoxia *per se.*

Pneumonia

Some patients with pneumococcal pneumonia may have jaundice. It usually occurs on the fourth or fifth day of the illness and is seen particularly in patients with consolidation of the right lower lobe. The serum bilirubin rarely exceeds 100 µmol/litre, and abnormalities of other liver tests are unusual. The cause of the jaundice is not known: factors that have been implicated are glucose-6-phosphatase deficiency, associated acute haemolysis, hypoxia, fever, and direct toxicity. The increased amounts of inflammatory cytokines seen in such patients may also contribute to the jaundice.

Abnormal liver function tests are also seen in patients with Legionnaire's disease and are characterized by elevation of aspartate aminotransferase and alkaline phosphatase. Jaundice is less common and tends to occur only in patients who are severely ill.

Diseases that involve lung and liver

a_1-Antitrypsin deficiency (see Chapter 11.13)

a_1-Antitrypsin deficiency was initially described in relation to pulmonary emphysema but it is now known that the liver, kidney, and pancreas can also be diseased. In children, a_1-antitrypsin deficiency often presents as neonatal hepatitis: in one-third it resolves, one-third develop fibrosis, and the remainder develop progressive cirrhosis, often requiring transplantation. In adults the disease often presents with cirrhosis or its complications. The liver shows the characteristic histological features of periodic acid-Schiff-positive, diastase-resistant globules in the liver. These globules are not diagnostic of the disease. Patients usually express the Pi zz phenotypic a_1-antitrypsin variant.

The course is unpredictable, but many patients develop progressive disease that requires liver transplantation. There is no other proven effective treatment. The onset of cholestasis often heralds liver failure. In cases where lung and liver disease coexist, the only effective therapy is transplantation of the heart, lung, and liver.

Cystic fibrosis

The increasing success in treating respiratory complications in children with cystic fibrosis has resulted in a greater number surviving to develop liver disease. Abnormal liver tests are found in up to half the children, and in adults up to a quarter of patients with cystic fibrosis develop a biliary cirrhosis. These patients present with cholestasis and jaundice. The pathogenesis and aetiology of this cholestasis are poorly understood. In most cases, liver disease is characterized by the development of a focal biliary cirrhosis that increases with time. Early involvement of the liver is characterized by the presence of eosinophilic granular material in the portal ducts. There is proliferation of bile ducts and portal fibrosis. This progresses to a focal biliary cirrhosis, which then develops into a multilobular cirrhosis with the onset of symptoms of cholestasis and jaundice. However, many patients have evidence of biliary obstruction shown by imaging the biliary tree by magnetic resonance cholepancreatography or endoscopic retrograde cholepancreatography. There is some evidence that infusion of *N*-acetyl cysteine into the biliary tree may relieve the obstruction in the extrahepatic biliary tree. The onset of jaundice and ascites is associated with a poor prognosis. Standard liver tests may underestimate the severity of the liver disease. Treatment with ursodeoxycholic acid will improve the liver tests and may improve liver function.

Other causes of cholestasis in patients with cystic fibrosis include gallstones and pancreatic insufficiency associated with increased loss of faecal bile salts, a consequent decrease in the size of the bile salt pool, and the development of lithogenic bile.

Treatment is uncertain. Open studies have suggested that ursodeoxycholic acid, 10 mg/kg/day, may result in biochemical improvement, weight gain, and improved nutrition. However, whether this agent has any long-term effect remains to be established.

Sarcoidosis

Sarcoidosis is a systemic granulomatous disease of unknown aetiology. The liver is commonly affected with evidence of infiltration in up to 70 per cent of cases. However, symptoms and signs of hepatic disease are uncommon. Hepatomegaly and splenomegaly occur in about one-quarter of patients. While jaundice is rare, elevation of the serum alkaline phosphatase is frequently observed. Complications of granulomatous infiltration of the liver are unusual. Liver failure may supervene, but portal hypertension occurs more frequently and causes bleeding varices or ascites.

As with sarcoid elsewhere, the diagnosis is usually made by the finding of non-caseating granulomas around the portal tracts. These granulomas are usually large and consist of multinuclear giant cells with T lymphocytes. Most patients respond to corticosteroids, although the portal hypertension may persist, possibly due to established presinusoidal fibrosis.

Overlap syndromes with primary biliary cirrhosis are well recognized: in patients with typical sarcoid infiltration in lungs and liver in the presence of bile duct damage consistent with primary biliary cirrhosis (PBC) and with positive antimitochondrial antibody tests, treatment should be directed against the sarcoid because, as yet, there is no -treatment that improves survival in primary biliary cirrhosis. Other causes of granulomatous hepatitis are discussed elsewhere.

Drugs

As indicated elsewhere, many drugs used for the treatment of lung diseases may be associated with abnormal liver function. Inhaled disodium chromoglycate reportedly causes a syndrome that transiently resembles primary biliary cirrhosis. However, inhaled medications otherwise rarely cause significant abnormalities of liver-related tests.

Disorders of the gastrointestinal tract

Inflammatory bowel disease

The range of liver abnormality associated with inflammatory bowel disease includes fatty change, pericholangitis, sclerosing cholangitis, chronic active hepatitis, cirrhosis, and amyloidosis. The reported incidence of serum liver test abnormalities in inflammatory bowel disease varies from 3 to 10 per cent. In general, abnormalities of liver tests correlate poorly with the severity of liver disease determined histologically. Ulcerative colitis is more commonly associated with abnormal serum liver-related tests than is Crohn's disease.

There is no clear relation between the onset of symptoms of inflammatory bowel disease and liver abnormalities. In general, symptoms of ulcerative colitis precede changes in liver function tests by about 8 years, but liver disease may precede by many years the onset of clinically apparent inflammatory bowel disease. Conversely, liver disease may become manifest several years after colectomy. Furthermore, there is little correlation between the severity of inflammatory bowel disease and the incidence or severity of liver disease. Indeed, in many patients with sclerosing cholangitis the colitis tends to be a pancolitis but is often quiescent (Table 1). Fatty change is relatively common on histological examination of the liver in patients with inflammatory bowel disease and is probably multifactorial in origin, relating to the degree of ill health, poor nutrition, and use of corticosteroids. As a patient's condition improves, the fatty infiltration resolves. Primary sclerosing cholangitis is associated with inflammatory bowel disease in about 10 per cent of cases, whereas nearly 90 per cent patients with sclerosing cholangitis have inflammatory bowel disease. Primary sclerosing cholangitis is a premalignant condition, associated with bile duct carcinoma in 5 to 20 per cent of cases. In patients with primary sclerosing cholangitis and ulcerative colitis, there is an increased risk of colon cancer.

Cirrhosis occurs in up to 10 per cent of patients who die with ulcerative colitis. The cause of the cirrhosis is not known, but in some cases it may be caused by chronic hepatitis C infection from drug transfusions or by drug toxicity, rather than primary sclerosing cholangitis.

Although chronic active hepatitis in association with inflammatory bowel disease is rare, its recognition is important because it resembles autoimmune chronic active hepatitis and may respond well to corticosteroid treatment. Other hepatic complications of inflammatory bowel disease include granulomatous hepatitis, amyloid infiltration of the liver, bile duct carcinoma, gallbladder cancer, and gallstones.

Coeliac disease

In patients with coeliac disease there may be minor abnormalities of liver function tests, characterized by elevation of serum aminotransferases; they usually resolve with institution of a gluten-free diet. Coeliac disease may also be associated with autoimmune diseases affecting the liver, including primary biliary cirrhosis, cryptogenic cirrhosis, sclerosing cholangitis, and autoimmune hepatitis. Up to 4 per cent of patients with primary biliary cirrhosis may have coeliac disease and up to 3 per cent of patients with coeliac disease may have PBC.

Gastrointestinal bypass surgery

Jejunoileal bypass surgery may be associated with liver disease; the changes in the liver range from simple fatty infiltration to cirrhosis. In a few cases there may be features identical to those of alcoholic hepatitis. In those in whom liver function tests are deranged, progressive injury is likely; although treatment with metronidazole has been advocated, restoration of normal anatomy appears to be the only effective measure.

Total parenteral nutrition

The association of hepatobiliary disorders with total parenteral nutrition has been recognized over the last two decades. Although the pathogenesis remains obscure, most studies suggest that the incidence is now less than 5 per cent. Total parenteral nutrition-associated hepatobiliary disease varies from a mild, asymptomatic disease with acalculous cholecystitis, biliary sludge, or hepatomegaly to jaundice, cirrhosis, and liver failure. Biochemically, the severity of abnormalities will reflect the severity of the disease but elevations of serum liver enzymes such as aspartate and alanine transferases, lactate dehydrogenase, and alkaline phosphatase, and serum bilirubin, are common. The histological features vary from a mild fatty infiltrate or cholestasis to a more severe picture resembling alcoholic fatty liver. Cirrhosis will develop in chronic cases. The mechanism is uncertain but hypoxic enterocytes, nutritional depletion, sepsis, toxicity of certain unidentified amino acids, and even carnitine deficiency have been implicated (Table 2). Once a patient develops abnormal liver function, and provided that other causes have been excluded, there is little alternative other than to reduce or stop parenteral nutrition and find other ways of providing adequate nutrition.

Obesity

Obesity is occasionally associated with abnormalities of serum liver tests, especially of the serum aminotransferases. Cutaneous manifestations of chronic liver disease may develop. The liver ultrasound will show a fatty liver and liver histology a macrovesicular fatty infiltration. Rarely, a non-alcoholic steatohepatitis syndrome will develop.

The liver in endocrine disease

Autoimmune hepatitis may be associated with autoimmune endocrine disorders, such as thyroid disease, vitiligo, and diabetes mellitus.

In haemochromatosis, where both liver and pancreas are affected, diabetes itself is associated with liver abnormalities. The liver may be enlarged due to excess stores of fat and glycogen. In severe cases there may be a non-alcoholic steatohepatitis syndrome which can lead to hepatic fibrosis and cirrhosis. Liver abnormalities are seen much more commonly in non-insulin-dependent diabetes (20–75 per cent of patients) compared with well-controlled type I diabetes (< 1 per cent of patients).

Hypothyroidism may be associated with a mild hyperbilirubinaemia and elevated serum transaminases (which may be of muscular origin). Ascites occurs very rarely. Hyperthyroidism is also associated with mild abnormalities of liver function which resolve on treatment of the thyroid disorder; severe uncontrolled thyrotoxicosis is associated with cardiac failure, atrial fibrillation, and jaundice (Habershon's jaundice) in which hypoperfusion and the toxic effects of thyroid hormone lead to hepatocellular dysfunction —this syndrome carries a poor prognosis and requires urgent treatment for thyroid stores.

The liver in haematological diseases

In general, diseases of the blood do not affect the liver. However, diseases associated with abnormal blood clotting, such as protein C or S deficiency and paroxysmal nocturnal haemoglobulinuria, may lead to Budd–Chiari syndrome (hepatic vein thrombosis).

Haemolysis

Jaundice may accompany haemolysis and is principally associated with an increase in unconjugated bilirubin. In patients with underlying liver disease both conjugated and unconjugated bilirubin are elevated out of proportion to the degree of haemolysis. Patients with chronic haemolytic anaemia are at risk of developing haemosiderosis. Iron is deposited initially in the Kupffer cells, but spread to the parenchyma will subsequently occur. The identification of the gene for hereditary haemochromatosis has assisted in the distinction between primary and secondary iron overload. The haemolytic anaemias are associated with an increased risk of pigment gallstones, which may lead to liver and biliary tract disease.

Sickle cell disease

Most of the abnormalities of liver function in sickle cell disease are due to haemolysis or infections transmitted by blood transfusion. Kupffer cell hyperplasia, haemosiderosis, fibrosis, or cirrhosis may be due to iron overload following multiple transfusions. Sometimes, patients present with severe pain in the right upper quadrant and rapid enlargement of the liver is part of the hepatic sequestration syndrome. Liver histology may show clumps of sickled red cells in the sunusoids, erythrophagocytosis, and sinusoidal dilatation. There is an increased risk of gallstones.

Thalassaemia

As with sickle cell disease, there is an increase in haemolysis and the complications of blood transfusions. Gallstones are common.

Multiple transfusions

Patients who receive regular blood transfusions or blood products, those with thalassaemia or haemophilia for example, are at risk of developing viral hepatitis B or C. It is now clear that such patients are at an increased risk of hepatitis C which may lead to cirrhosis and liver cell cancer. With screening, the incidence of hepatitus C virus is likely to fall. Blood transfusion is also associated with secondary haemochromatosis since each unit of blood provides the equivalent of 225 mg of iron which ultimately causes iron storage disease in the liver and in other organs.

Lymphomatous disease

In patients with Hodgkin's disease, liver function tests are of limited value in predicting liver involvement, although jaundice is a recognized feature and may be due to a number of different causes. For example, haemolysis may complicate Hodgkin's disease, and occasionally there is a bland cholestasis in the absence of infiltration, which resolves when the disease is treated. The clinical manifestations of liver involvement in Hodgkin's disease relate to the degree of infiltration. In rare cases, patients with Hodgkin's disease develop fulminant hepatic failure: the clue to infiltration is a large liver, as most cases of viral or drug-related fulminant hepatic failure are associated with small livers. Liver biopsy may be diagnostic. Primary lymphoma of the liver has been described, but is rare. The liver may be involved in both non-Hodgkin's lymphoma and leukaemia and the diagnosis is usually made by biopsy. Some patients with non-Hodgkin's lymphoma have chronic hepatitis preceding diagnosis or treatment. A causal effect cannot be excluded. Hodgkin's disease and non-Hodgkin's lymphoma may be associated with obstructive jaundice due to invasion of the bile duct by invasive nodal disease in the hilum.

The liver and infections

Abnormal liver function may occur during systemic infections but it is rare for patients with sepsis to present primarily with liver symptoms (Table 3). However, jaundice, abnormal liver function tests, or even, occasionally, fulminant hepatic failure may be the principal presenting feature.

Bacterial infections

Pneumococcal infections are discussed above. Meningococcal infections are occasionally associated with features suggestive of viral hepatitis. Jaundice may be associated with the toxic shock syndrome associated with *Staphyloccus aureus*. Gonococcal infection is a well-recognized cause of liver disease. The classical Fitzhugh–Curtis syndrome, perihepatitis, is characterized by sudden onset of severe pain in the right upper quadrant, occurring classically in a woman with a previous history of pelvic inflammatory disease. On examination there may be little to find, although tender hepatomegaly and a hepatic rub may be present. Where laparotomy has been performed in the mistaken diagnosis of cholecystitis, perihepatitis with adhesions and pus around the liver may give the clue to the diagnosis. In chronic infection, adhesions develop between the surface of the liver and the anterior abdominal wall. The condition usually resolves without treatment, although the use of penicillin promotes more rapid resolution. Abnormalities of liver function may occur in gonococcal bacteraemia, peritonitis, and endocarditis. Perihepatitis is also reported in association with syphilis and chlamydial infections. Chlamydial infection is today the most frequent cause of chronic perihepatic adhesions.

In childhood, some infections with *Escherichia coli* may be associated with hepatitis and jaundice. Jaundice is rare in older patients, although pregnant women seem more susceptible. Abnormalities of liver function occur in systemic streptococcal and staphylococcal infection and in enteric fevers, paratyphoid, and typhoid. Hepatomegaly is common in typhoid infection and jaundice occurs in about 10 per cent of patients, although up to a third have abnormal liver function tests with increased levels of aminotransferase and normal values for alkaline phosphatase. The hepatomegaly rapidly responds with treatment. In gas gangrene, deep jaundice may occur may occur in up to a fifth of patients. The liver may be infected and a plain radiograph of the abdomen may show gas within the liver. Liver damage and jaundice are associated with *Listeria monocytogenes* and *Legionella pneumophila* infections

Brucellosis may also be associated with jaundice and abnormal liver function tests. All three species of *Brucella* have been associated with abnormal liver function. Characteristically, the liver biopsy shows a marked inflammatory infiltrate and fibrosis with multiple large or small granulomas scattered throughout the parenchyma. Some reports have suggested that granulomatous hepatitis due to *Brucella* may cause cirrhosis, but this is questionable. The common causes of liver granulomas, including infections, are listed in Table 4. The liver damage asociated with leptospirosis (Weil's disease) is described elsewhere (Section 7).

Actinomyces spp. are commensal organisms that rarely cause disease. Actinomycotic infection of the liver may occur, the patient presenting with abdominal pain, anorexia, and fever. In one case report the liver was found to have small, multilocular abscesses.

Tuberculosis may present with granulomatous hepatitis, biliary tuberculosis, a solitary tuberculoma, or tuberculosis of the biliary tract. The liver is involved in up to 85 per cent of patients with tuberculosis, especially in those with miliary disease. The presence of multiple granulomas in the liver should raise the possibility of tuberculosis, although the differential diagnosis of granulomatous hepatitis is long (see Table 4). With the increasing incidence of atypical mycobacterial infections, lesions similar to tuberculosis can be found. In those infected with *Mycobacterium avium intracellulare* there are numerous acid-fast bacilli, often in the absence of granulomas.

Leptospirosis

Leptospiral infections are described in Section 7. Acute leptospirosis is frequently accompanied by jaundice, although frank liver failure is uncommon. The jaundice is mainly cholestatic, although there may be liver cell injury.

Rickettsial infection

Liver injury in Q fever (*Coxiella burnetti*) is recognized, although symptoms of liver disease are uncommon. Hepatomegaly is frequent and liver function tests may show an elevation of serum alkaline phosphatase and, rarely, a picture resembling viral hepatitis. Histologically, the liver has areas of focal necrosis, Kupffer cell proliferation, lipogranuloma formation, and mononuclear cell infiltration in the portal tracts. The characteristic histological feature of Q fever is eosinophiliic fibrinoid necrosis but this is not specific. Treatment is with chloramphenicol or tetracycline. Liver disease is much more frequent in Rocky Mountain spotted fever (*Rickettsia rickettsia*).

Fungal infections

The liver may be involved in fungal infection, often in patients with immunodeficiency such as with AIDS, following chemotherapy, and after organ transplantation. Histoplasmosis, cryptococcosis, aspergillosis, blastomycosis, and candidiasis are all causes of liver damage. The liver is usually involved in disseminated fungal infections. Cryptococcal infection has also been associated with a primary biliary cirrhosis-like condition.

Protozoal infections

Protozoal infections are described in detail in Section 7.13 many of them involve the liver. In toxoplasmosis, while most patients are asymptomatic and liver involvement is mild, hepatitis may occur and *Toxoplasma gondii* may be found in liver biopsy samples. In malaria, due to either *Plasmodium falcipartun* or *P. vivax*,

abnormalities of liver tests may be observed. Hepatomegaly is common and is often associated with jaundice. The jaundice is in part due to haemolysis but liver tests may provide a picture suggestive of viral hepatitis. Histological examination may show characteristic features of Kupffer cell proliferation with black malarial pigment and mononuclear cell infiltrate. Frank hepatic failure is extremely rare.

Schistosomiasis is one of the most common causes of liver disease worldwide. A heavy infection of fertile schistosomes in the portal system results in deposition of eggs that induce an immune response, leading to portal fibrosis and granuloma formation, portal hypertension with consequent splenomegaly, ascites, and variceal haemorrbage. Hepatocyte function is well preserved. There is a complex interaction between schistosomal eggs and the immune system; the degree of fibrosis is directly related to the number of eggs and the duration of infection. The diagnosis is made on stool examination or finding schistosomes in the liver. Serological tests are unreliable at present. Treatment is described in Chapter 7.16.1. Successful treatment is associated with a significant but variable improvement in the degree of portal hypertension. Treatment of the portal hypertension is dependent on the medical facilities available. As parenchymal function is well preserved, these patients usually tolerate a portosystemic shunt.

Coinfection of patients with schistosomiasis and hepatitis B or C virus is associated with an aggressive progression.

Viral infections

Hepatitis may be a significant feature of viral infection other than with the classical hepatitis viruses. Thus, infection with cytomegalovirus, Epstein–Barr virus, herpesviruses, measles, rubella, coxsackievirus, adenoviruses, and echoviruses may all cause a significant hepatitis. Such viral infections (especially cytomegalovirus) are more common in immunosuppressed patients. The diagnosis is made serologically, but in some cases, such as with cytomegalovirus, herpes, and adenoviral infections, the liver histology may show characteristic features.

Pyogenic liver abscess

Pyogenic liver abscesses may occur as part of a systemic illness, or as a consequence of portal phlebitis. Abscesses are often associated with bowel sepsis, biliary tract disease, direct trauma, septicaemia, and in association with carcinoma of the colon or bacterial endocarditis. They most commonly arise out of portal phlebitis, with the primary focus being the appendix, colon, diverticular disease, or in the pelvis (Table 5). Although abscesses may occur in patients with inflammatory bowel disease, this is relatively rare. The patient presents with abdominal pain, pyrexia, nausea, and weight loss. However, fever is less common in children. Hepatomegaly may be present and the liver is sometimes tender. The serum albumin is often reduced and alkaline phosphatase elevated. There is usually marked neutrophil leucocytosis, but this is not invariable. The diagnosis is made on imaging of the liver. A chest radiograph may show elevation of the right hemidiaphragm with an associated pleural effusion or even lung consolidation. Ultrasound, computed tomography, and magnetic resonance imaging may define a hepatic abscess. With the increasing sensitivity of these techniques, radio-isotope scanning is now less important.

Treatment of a solitary abscess is by percutaneous drainage in the first instance. Under the guidance of ultrasound or computed tomography, a percutaneous drain should be established for single abscesses, and even in some cases of multiple abscesses. The abscesses should be drained to dryness, and antibiotics should be given according to the sensitivities of the organisms isolated. Pathogens are usually anaerobic or aerobic gut coliforms, especially *Streptococcus milleri*, but in children *Staphylococcus aureus* is common. The success rate of treatment with drainage and systemic antibiotics is 80 to 90 per cent. Fatality is high in children and the elderly, in those with coexisting disease such as diabetes mellitus, and in those with delayed diagnosis. Once the abscess has been drained, the primary source of infection must be sought and appropriate management instituted. Surgery may be required for patients with multiple abscesses or for those with abscesses that do not respond to simple drainage and antibiotic therapy. Liver abscess due to hydatid and amoebal infection is discussed elsewhere.

AIDS and liver disease

Liver disease in patients with human immunodeficiency virus (HIV) infection may be due to pre-existing hepatitis virus, opportunistic infections, or neoplasms. In some cases the abnormality of liver function may be due to virus itself. Such patients have non-tender hepatomegaly with anorexia, weight loss, and low-grade fever. Liver function tests show slight derangement with cholestasis. The liver biopsy shows non-specific features including Kupffer-cell hyperplasia, fat infiltration, non-caseating granulomas, and portal-tract inflammation; Mallory hyaline bodies may occasionally be present.

Other causes of hepatobiliary abnormality in patients with HIV include primary hepatic infection due to viral hepatitis.

Other causes of liver damage in AIDS

Many patients with AIDS are also at risk from hepatitis B, C, and D. As discussed elsewhere, these patients respond less well to interferon than do those who are HIV negative. Other infections that are more common in HIV-positive patients include cytomegalovirus, herpesvirus, cryptosporidiosis, and mycobacterial infections including tuberculosis and *Mycobacterium avium intracellurare*. Drug-induced liver damage must always be considered in HIV patients with abnormal liver tests, and it has been suggested that such patients are more susceptible to drug hepatotoxicity. Thus, many of the anticonvulsants, analgesics, and antimicrobials are associated with hepatocellular damage, and antibiotics may also be associated with cholestasis. Other abnormalities that may be of less significance clinically include peliosis hepatis and fatty infiltration.

The biliary tree may also be affected in HIV infection inducing a syndrome superficially resembling primary sclerosing cholangitis. This is characterized by a rapid elevation of the serum alkaline phosphatase, which may be associated with pain in the right upper quadrant and, later, jaundice. Ultrasonography may be unhelpful, although dilated and thickened walls of the bile duct may be seen. Otherwise, endoscopic retrograde cholepancreatography will show the characteristic changes of sclerosing cholangitis with bleeding, dilatation, and stricture. Both cryptosporidial and cytomegaloviral infections have been associated with this form of sclerosing cholangitis.

The liver may be affected by HIV in other ways. There is an association between AIDS and lymphomas, be they Burkitt's, large cell, or immunoblastic lymphomas. The liver and/or spleen may be the site of these tumours and hepatic infiltration may be present in up to a third of those with gastrointestinal lymphomas. Tumours may be microscopic or macroscopic. The hepatic masses are often asymptomatic but if large they may cause pain in the right upper quadrant, fever, jaundice, and abnormalities of serum liver tests, especially of the serum alkaline phosphatase. Kaposi's sarcoma may affect the liver and biliary tree but is often asymptomatic.

Liver and rheumatological disease

Liver abnormalities occur in patients with rheumatological disorders, although they rarely prove to be clinically significant. Hepatic disease may either be a consequence of treatment or occur in association with other autoimmune diseases. For example, those diseases assumed to have an autoimmune basis, such as autoimmune hepatitis or primary biliary cirrhosis, may be associated with extrahepatic rheumatological diseases such as the sicca syndrome.

Rheumatoid arthritis

Abnormalities of liver structure and function are uncommon in patients with rheumatoid arthritis, although minor abnormalities of liver function tests occur in 20 to 50 per cent of cases. Nodular regenerative hyperplasia may cause complications of portal hypertension.

Felty's syndrome

Felty's syndrome is characterized by the triad of splenomegaly, hypersplenism, and seropositive rheumatoid arthritis. Liver function tests tend to be more commonly deranged than in uncomplicated rheumatoid arthritis. Anti-inflammatory therapy may contribute to the abnormal liver tests. Histological examination of the liver shows lymphocytic infiltration and, rarely, an established cirrhosis. Nodular regenerative hyperplasia occurs in patients with Felty's syndrome, as with rheumatoid arthritis. Although portal hypertension and variceal haemorrhage may occur, jaundice is unusual.

Connective tissue disease

Systemic lupus erythematosus

Usually only minor abnormalities of liver function occur in patients with systemic lupus erythematosus, although spontaneous rupture of the liver has been described. The pattern of liver disease in patients with systemic lupus erythematosus varies from minimal change to chronic persistent hepatitis, chronic active hepatitis, and cirrhosis. In others, a granulomatous hepatitis has been identified.

Polyarteritis nodosa

In contrast to rheumatoid arthritis, liver injury in polyarteritis nodosa is relatively uncommon, although a hepatic arteritis may occur, leading to aneurysm. Rupture of an aneurysm is rare and is characterized by fever, pain in the right upper quadrant, and jaundice. In most cases, abnormalities of liver function are due to an associated hepatitis C virus infection.

Polymyalgia rheumatica

Abnormalities of liver function are well recognized in patients with polymyalgia rheumatica. These abnormalities (of elevation of serum alkaline phosphatase and aminotransferase activity) usually resolve with effective treatment. Histologically, the liver shows mild portal inflammation with occasional liver cell necrosis. Granulomas and steatosis may also be seen.

Sjögren's syndrome

Symptoms of sicca syndrome are common in patients with liver disease, particularly primary biliary cirrhosis, and abnormalities of salivary gland function have been described in all patients in some series. Sicca syndrome is also found in patients with cryptogenic cirrhosis and autoimmune chronic active hepatitis. In patients with Sjögren's syndrome there is often hepatomegaly and minor derangement of liver function tests, particularly serum alkaline phosphatase, in 25 per cent of cases. The liver may show non-specific inflammatory infiltration.

Amyloid (see Chapter 11.12.4)

The liver is infiltrated in both primary and secondary amyloidosis; liver disease is found in over 80 per cent of patients with either form. In general, however, liver amyloidosis has few significant clinical consequences, although jaundice, hepatitis, portal hypertension, and spontaneous hepatic rupture occur. Clinically, the liver is enlarged; the serum alkaline phosphatase is usually greatly elevated and jaundice is uncommon. It is believed that liver biopsy may be particularly hazardous in patients with amyloid because there may be an increased risk of bleeding after biopsy. There is no treatment other than that which addresses the underlying cause.

Cryoglobulinaemia

The reported incidence of liver disease in essential, mixed cryoglobulinaemia varies greatly. In 20 to 50 per cent of patients there is an association with hepatitis C viral infection. Up to half of patients have evidence of infection with hepatitis B virus, and up to 10 per cent have chronic active hepatitis or cirrhosis with jaundice.

The liver in malignancy

Although the liver may become infiltrated by metastatic cancer, abnormalities of liver tests can be seen in the absence of infiltration. This may be due to a systemic effect of tumour-derived cytokines, to the hepatotoxic effects of drugs, or to the effects of irradiation.

The liver in the sick patient

Abnormalities of serum liver tests are frequent in patients who are critically sick, and are associated with a poor prognosis. There are many causes of abnormal liver tests in this situation (Table 6). 'Intensive therapy unit' jaundice occurs in up to 10 per cent of patients so treated, usually in the context of sepsis or abdominal trauma. Hepatomegaly is often present, but the cutaneous features of chronic liver disease are absent. Encephalopathy is uncommon. The liver tests show a rise in serum bilirubin with a smaller and less consistent rise in serum alkaline phosphatase and aminotransferase levels. Blood coagulation tests are mildly deranged and blood sugar levels tend to be high rather than low. The cause of intensive therapy unit jaundice is unclear but factors such as ischaemia, hypoxia, and hepatocyte necrosis may occur.

Liver disease in pregnancy

Liver disease in pregnancy is discussed in Chapter 13.9.

Further reading

Birrer MJ, Young RC (1987). Differential diagnosis of jaundice in lymphoma patients. *Seminars in Liver Disease* **7**, 269–77.

Krowka MJ (2000). Hepatopulmonary syndromes. *Gut* **46**, I–4.

Lefkowitch JH (1990). Hepatic granulomas. *Journal of Hepatology* **30** (Suppl. I), 40–5.

Matsumoto T, *et al.* (2000). The liver in collagen diseases: pathologic study of 160 cases with particular reference to hepatic arteritis, primary biliary cirrhosis, autoimmune hepatitis and nodular regenerative hyperplasia of the liver. *Liver* **20**, 366–73.

Moseley RH (1997). Sepsis-associated cholestasis. *Gastroenterology* **112**, 302–6.

Sandhu IS, Jarvis C, Ererson GT (1999). Total parenteral nutrition and cholestasis. *Clinics in Liver Disease* **3**, 489–508.

Valla DC, Benhamou JP (2000). Hepatic granulomas and hepatic sarcoidosis. *Clinics in Liver Disease* **4**, 269–85.

14.22 Miscellaneous disorders of the gastrointestinal tract and liver

D. P. Jewell

This chapter describes conditions that do not fall naturally into any of the other major sections which deal with disorders of the gastrointestinal tract and liver.

Cystic disorders of the bowel

Diverse disorders of the small and large bowel are associated with cyst formation. All are rare but present major diagnostic and therapeutic problems.

Colitis cystica

There are several varieties of benign cystic lesions involving the colonic mucosa. Colitis cystica superficialis occurs in patients with pellagra and has also been reported in adult coeliac disease. The presenting feature is usually diarrhoea and the condition is characterized by the presence of small mucus-filled cysts that lie superficially to the muscular layer of the colon. The disease seems to respond to therapy for pellagra. In colitis cystica profunda the cysts occur below the muscular layer of the colon. The pathogenesis and aetiology of this condition are unknown. It is usually characterized by cramping lower abdominal pain, tenesmus, and diarrhoea associated with blood and mucus in the stools. The diagnosis is made by sigmoidoscopy, rectal biopsy, and barium enema examinations. Treatment consists of local surgical excision.

Pneumatosis cystoides intestinalis

This disorder is characterized by the presence of multiple, gas-filled cysts in the wall of the colon and, less frequently, the small intestine. The stomach and mesentery may also rarely be affected. The condition is usually seen in middle-aged patients and may be associated with obstructive airways disease or pyloric obstruction. It has also been found in association with a variety of other conditions, including mesenteric vascular disease, small-bowel tumours, and Whipple's disease, and after small-bowel surgery. It can also occur after sigmoidoscopy or colonoscopy.

The pathogenesis of the condition is unknown. The composition of the gas within the cysts is similar to atmospheric air with the addition of small quantities of methane and a higher concentration of hydrogen. Thus, the gas may diffuse into the lamina propria from the intestinal lumen. Suggestions that the gas is formed by excessive bacterial fermentation of carbohydrate or that, in patients with obstructive airways disease, alveolar air tracks down through the muscle planes of the posterior abdominal wall and out into the mesentery lack firm evidence.

Pneumatosis cystoides can be an incidental finding during radiological examination of the abdomen. It can present with symptoms that include lower abdominal pain, recurrent diarrhoea, rectal bleeding, and tenesmus. Obstructive symptoms can occur. The cysts can frequently be detected at sigmoidoscopy and rectal biopsy specimens may show the characteristic giant cells that frequently line the cyst walls. Double-contrast radiology is the best method to establish the diagnosis but computed tomographic scans may detect serosal or mesenteric cysts.

If the disease is detected incidentally and the patient has no abdominal symptoms, no treatment is necessary. However, treatment is indicated for symptomatic patients, especially those with severe obstructive symptoms. The most effective therapy is oxygen in high concentration. Most patients will benefit from a concentration of 60 to 70 per cent given by a face mask or a nasal catheter. Hyperbaric oxygen may rarely be indicated for resistant pneumatosis. The aim of oxygen therapy is to wash out the nitrogen from the cysts and replace it with oxygen, which can be resorbed, with collapse of the cysts. Other treatments that have been tried are antibiotics (metronidazole), a low-carbohydrate diet, and an elemental diet. Surgical resection may be necessary for patients who do not respond and for those with severe obstructive symptoms, especially if there is a suspicion of volvulus.

Lymphocytic and collagenous colitis

These two disorders are principally seen in middle-aged women presenting with a watery diarrhoea. Although claims have been made that lymphocytic colitis progresses to a collagenous colitis, this has not been universal experience; here the disorders will be described separately.

Collagenous colitis was first recognized by Lindstrom in 1976. Patients usually present in the fifth and sixth decade; mostly women, but the disorder can occur in young adults as well as in the elderly. A watery diarrhoea accompanied by abdominal cramps, wind, distension, and nausea are the usual symptoms. The diarrhoea can be severe and is often secretory in nature. There may be some mucus but bleeding is not a feature. Despite severe symptoms, these patients are usually well, with a good appetite, and do not lose weight. There are no abnormal physical signs on examination and, on sigmoidoscopy, the rectal mucosa can be normal but may be hyperaemic with some oedema and granularity. These endoscopic changes occur throughout the colon but are usually patchy and never severe. The diagnosis is made on the appearance of colonic biopsies. There is a thickened band of subepithelial collagen exceeding 15 μm compared with the normal thickness of 2 to 6 μm. The collagen band is of maximal thickness in the right colon and tends to become thinner more distally—there may rectal sparing, which may lead to misdiagnosis if the only histological material available is from the rectum. Immunohistochemical studies have shown that the abnormal tissue consists predominantly of collagen type III, but why the pericryptal fibroblasts synthesize abnormal collagen in this disorder but not in other inflammatory disorders is not understood. There is a patchy and variable inflammatory infiltrate in the lamina propria consisting of lymphocytes, plasma cells, and some neutrophils. The disease is confined to the colon and does not extend into the ileum. It is therefore distinct from collagenous sprue.

Lymphocytic colitis shares the same clinical features as collagenous colitis but the colonic mucosa always looks normal at colonoscopy. Nevertheless, histological examination of biopsy specimens shows a diffuse inflammatory infiltrate throughout the lamina propria with no architectural changes of the glands. The infiltrate is predominantly lymphocytic but there are also plasma cells and eosinophils. A characteristic feature is the marked increase in intraepithelial lymphocytes, which clearly separates this disease from ulcerative colitis in which the intraepithelial lymphocyte counts are normal or reduced. A lymphocytic colitis can be seen in some patients with untreated coeliac disease and in some patients receiving non-steroidal anti-inflammatory drugs. These possibilities need to be excluded before the diagnosis of lymphocytic colitis can be made.

Both of these disorders have a variable clinical course and the symptoms can spontaneously remit and relapse. However, in general, treatment is difficult, especially if a large-volume secretory diarrhoea occurs. Antidiarrhoeal agents, such as loperamide, should be used initially, but if no response is seen, then 5-aminosalicylic acid compounds or oral corticosteroids can be tried. If symptoms persist, dietary exclusion and even an elemental diet are sometimes helpful. Other treatments that have had some anecdotal benefit are metronidazole, cholestyramine, and mepacrine.

Patients may have intermittent symptoms over many years and, unless a rectal biopsy is obtained and correctly interpreted, the true diagnosis is not made. Many of these patients are thus labelled as having an irritable bowel syndrome. A few cases of lymphocytic colitis have been reported as progressing to a collagenous colitis, but this outcome appears to be infrequent. Neither disease progresses to a frank ulcerative colitis.

Miscellaneous vascular disorders of the bowel

Intramural bleeding

The most common cause of bleeding into the wall of the bowel is anticoagulant therapy. Occasionally, intramural haematomas form in patients with congenital coagulation defects such as haemophilia or in conditions such as vasculitis. The usual presentation is with abdominal pain and symptoms of intestinal obstruction and bleeding. The diagnosis is made by barium follow-through or enema examination and the condition can usually be treated conservatively with blood replacement and, if necessary, nasogastric suction.

Aortic aneurysm

Aneurysmal dilatation of the aorta is relatively common in elderly patients, the usual site being the segment distal to the origin of the renal arteries. Aneurysms larger than 6 to 7 cm or those that show increasing enlargement are probably best resected. Rarely, a spontaneous fistula into the duodenum may develop with catastrophic gastrointestinal haemorrhage. This condition is considered in more detail in Section 15.

In patients who have had prostheses inserted into the abdominal aorta or into other retroperitoneal arteries the complication of paraprosthetic-enteric fistula may occur. This is characterized by the abrupt onset of abdominal pain and shock. Treatment is surgical and the prognosis is extremely poor.

Vascular malformations

The vascular ectasias of the colon are described in Chapter 14.16. Haemangiomas of the small intestine are rare. They usually present as recurrent anaemia due to gastrointestinal bleeding. They are best diagnosed by angiography or, occasionally, by enteroscopy. Telangiectasia may also occur throughout the stomach and bowel as part of the Osler–Rendu–Weber syndrome.

Endometriosis

The term endometriosis describes the presence of extrauterine endometrial tissue and its clinical manifestations. The disorder occurs most frequently between the ages of 30 and 40 years and is rare below the age of 20. Intestinal involvement is most common in those parts of the bowel adjacent to the uterus and fallopian tubes, particularly the rectosigmoid colon. It is not certain how heterotopic endometrial tissue reaches the bowel and spread may be direct, by the bloodstream, or by means of the lymphatics. The mucosa of the bowel is seldom penetrated by ectopic endometrial tissue and therefore gastrointestinal bleeding is an unusual accompaniment of endometriosis of the bowel.

The usual symptoms of endometriosis include dysmenorrhoea, menorrhagia, sterility, and intermenstrual pelvic pain and backache. If the rectosigmoid region is involved, there may be cyclic pains in the rectum and, occasionally, mild diarrhoea and tenesmus. Implants in the small intestine may produce symptoms of obstruction or volvulus. However, endometriosis affecting the intestine is frequently asymptomatic and much of the pain and altered bowel habit attributed to this condition is more often caused by an irritable bowel syndrome.

The diagnosis of endometriosis requires a thorough pelvic examination to demonstrate tender nodules in the rectosigmoid region or in the rectovaginal area. The diagnosis may be suggested by radiological evidence of the presence of lesions in the rectosigmoid region, but ultimately depends on biopsy of these lesions. Laparoscopy is often helpful in making the diagnosis. The differentiation from carcinoma may be difficult and when in doubt surgical exploration should be carried out.

The management of this condition requires expert gynaecological assistance. Mildly symptomatic cases are probably best managed by analgesics and sedation. More severe cases may require hormonal therapy with danazole, gestrinone, or gonadorelin analogues. These drugs inhibit gonadotrophin release from the pituitary.

Malakoplakia

This is a rare granulomatous disease involving the urinary tract and occasionally the colon or stomach. Histologically the condition is characterized by the presence of histiocytes containing dark inclusions that stain positive with periodic acid–Schiff and seem to contain both calcium and iron. There appears to be an acquired abnormality of macrophage function associated with defective digestion of phagocytosed bacteria.

Colonic malakoplakia is usually discovered as an incidental finding in elderly debilitated individuals, quite often in association with a malignant disease of the colon. Occasionally it presents as a systemic illness characterized by fever, diarrhoea, and other gastrointestinal symptoms. It can only be diagnosed by histological examination of the bowel. There is no effective treatment, although there has been a recent report of improvement in otherwise unresponsive cases by the use of cholinergic drugs.

Isolated ulcers of the large intestine

Caecal ulcers

These occasionally occur but their aetiology is unknown. They can present with abdominal pain, either acute or chronic, and may be a cause of an acute abdomen. Laparotomy may show perforation or local abscess formation. More commonly, they present with bleeding. Similar ulcers may occur elsewhere in the colon, especially at the flexures.

Solitary rectal ulcer syndrome

This refers to the occurrence of an ulcer in the rectum, usually 4 to 10 cm from the anal verge. It is usually found in young adults, mostly in women, but can arise at any age. The ulcers are most commonly on the anterior wall, but they can be multiple and can be sufficiently extensive to encircle most of the rectum. There may be an associated anterior-mucosal prolapse. Symptoms include rectal bleeding, which is the most common mode of presentation, tenesmus, and abdominal and rectal pain. The bowel habit may be normal, but many of these patients have a history of constipation and straining. At sigmoidoscopy the ulcer is readily seen and usually has a greyish base. However, there may be just an area of inflamed mucosa or the ulcer can be a polypoid lesion simulating a carcinoma. In either case, a biopsy specimen should allow the correct diagnosis to be made. Histologically, there is often evidence of ischaemia, but the characteristic feature is hypertrophy of the muscularis mucosae with smooth-muscle fibres extending between the crypts towards the epithelium. There may be considerable fibrosis. Treatment is often difficult.

Correction of constipation with bulking agents (with or without lactulose) is important and patients should be warned not to strain. Topical treatment with 5-aminosalicylic acid or corticosteroids may be helpful, but there are no controlled data to confirm their efficacy. As these ulcers can remit spontaneously, it is difficult to be sure whether treatment has been effective in an individual patient. For patients with continuing and disabling symptoms, anorectal physiological measurements should be made because there may be evidence of denervation of the pelvic floor muscles. A defaecating proctogram should also be obtained to demonstrate whether the anorectal angle changes when the patient attempts to empty the rectum and to record the degree of mucosal prolapse. Surgical therapy such as an anterior and posterior rectopexy may be helpful in some patients.

Stercoral ulcers

These occur in association with faecal impaction and are most commonly found in the sigmoid–rectal area. Most patients are elderly but they can occur in any patient who is severely constipated, including patients with neurological causes of constipation (for example paraplegia or multiple sclerosis). The common symptoms are those associated with the constipation—nausea, abdominal distension and pain, and anorexia—and there may be overflow incontinence. The ulcers are frequently asymptomatic but may be a cause of anaemia from chronic blood loss. They are normally revealed because of an acute bleed or a perforation. Treatment of these acute complications requires surgery, but otherwise, the ulcers are treated by relieving the faecal impaction.

Acute colonic pseudo-obstruction (Ogilvie's syndrome)

This syndrome describes a massive and acute dilatation of the caecum and right colon in the absence of organic obstruction or inflammatory disease of the colon. It may occur following intra-abdominal surgery, including urological or gynaecological surgery, but can also happen in any sick patient. Hence, it can occur in association with severe systemic sepsis, or respiratory or cardiac disease. Most patients have a constant, rather dull pain with marked abdominal distension and they frequently vomit. There is constipation but usually patients continue to pass wind and some may actually have diarrhoea. On examination, there is a distended and tympanitic abdomen that is tender to the point of mild rebound tenderness. Bowel sounds are variable in pitch and frequency but are rarely absent. The diagnosis is made on a plain radiograph of the abdomen. The danger is that of a colonic perforation if treatment is not instituted immediately. Intravenous fluids and electrolytes are given together with nasogastric suction. Any drugs that might be implicated (such as tricyclic antidepressants, anticholinergics) are stopped. Rectal tubes and enemas are often used but there value is dubious. It is usually recommended that the patient is turned from side to side at regular intervals. Theoretically this should distribute the intracolonic gas and hence reduce a continuous high pressure in the right colon—again, evidence that this is effective in reducing the incidence of perforation is not clear-cut. If the dilatation does not resolve, decompression can sometimes be achieved by colonoscopy, but as it is dangerous to prepare the colon the procedure is difficult. Surgical decompression with a caecostomy may be needed, which may be accompanied by resection if there is an obviously ischaemic segment of colon.

Melanosis coli and related disorders

The term melanosis coli is used to describe black or brown discoloration of the mucosa of the colon. It results from the presence of dark pigment in large mononuclear cells or macrophages in the lamina propria of the mucosa. The coloration is usually most intense just inside the anal sphincter and is less dark higher up in the sigmoid colon. Similar pigment has been found in the appendix and mesenteric nodes. The condition is thought to result from faecal stasis and the use of anthraquinone cathartics such as cascara or senna. Chronic cathartic abuse may also cause radiological changes of the colon that go under the general heading of cathartic colon. The changes are characterized by loss of haustral markings and appearances resembling multiple strictures, although in fact these areas are capable of distension. These changes may involve all parts of the colon and the terminal ileum, and have to be distinguished from those of ulcerative colitis, Crohn's disease, and other inflammatory bowel diseases.

Miscellaneous vascular disorders of the liver

The important vascular disorders of the liver include underperfusion in conditions of shock and left ventricular failure, the wide variety of diseases that give rise to portal hypertension, and acute and chronic venous congestion due to cardiac failure, diseases of the pericardium, or rarely, primary obstruction of the hepatic veins. Portal hypertension is considered in Chapter 14.21.2.

Acute cardiac failure and shock

Reduced hepatic blood flow primarily causes ischaemia in Rappaport zone III, which leads to centrilobular necrosis. This occurs frequently in acute congestive heart failure or in severe hypotension from any cause.

In acute congestive cardiac failure, the liver may be enlarged and tender if the central venous pressure is elevated. There may be biochemical features of liver cell damage or intrahepatic cholestasis. Liver function may be interfered with sufficiently to cause a reduction of prothrombin synthesis and hence increased sensitivity to anticoagulant drugs.

In patients with severe hypovolaemic shock there may be marked biochemical changes of deranged liver function and, in severe cases, jaundice. These changes are transient and revert rapidly to normal following restoration of a normal blood pressure and perfusion.

Chronic venous congestion

A persistently elevated central venous pressure due to right-sided heart failure or constrictive pericarditis results in hepatic venous congestion and hepatomegaly. The associated histological changes are characterized by centrilobular congestion with surrounding fatty change (the typical 'nutmeg liver'). If the disorder is of long standing, there may be progressive fibrosis extending peripherally from centrilobular to portal areas, although regenerative nodules are not prominent.

Apart from the underlying cardiac lesion, this disorder is characterized by hepatic enlargement and signs of congestive cardiac failure. The serum bilirubin level is usually increased and there may be a slight elevation in the serum alkaline phosphatase and in the transaminases.

It should be emphasized that, although the changes outlined above are relatively common in patients with long-standing heart failure, true cirrhosis of the liver with regenerative nodules is very rare, as is the clinical picture of portal hypertension that accompanies other forms of cirrhosis of the liver; it is very unusual to find oesophageal varices in patients with cardiac cirrhosis.

Hepatic arterial occlusion

This is a rare condition. It usually follows surgical trauma but has been found in association with arteritis and bacterial endocarditis.

The condition is characterized by an acute onset of pain in the upper abdomen, tenderness over the liver, and progressive shock and liver failure. Most cases have a fatal outcome.

Hepatic arterial aneurysm

This condition is recognized by the triad of upper abdominal pain, jaundice, and haematemesis following rupture of the aneurysm into the stomach or duodenum. The diagnosis is made by hepatic angiography. Treatment is by surgical resection.

Septic venous thrombosis of the portal system

This condition results from infection anywhere in the abdominal cavity leading to pylephlebitis of the portal venous system. It may occasionally result from a systemic septicaemia or from inflammatory disorders of the bowel, such as ulcerative colitis.

The acute phase of the disorder is usually characterized by features related to the underlying abdominal sepsis. This is followed by an episode of high fever, worsening abdominal pain, and rigors. There may be obvious evidence of septic embolization to the liver. This may lead to abdominal pain and hepatic tenderness

with mild jaundice. All the systemic features of a severe infection develop and there is usually a polymorphonuclear leucocytosis and abnormal liver tests. Occasionally, multiple large intrahepatic abscesses may develop.

The condition should be suspected in any patient with abdominal sepsis who develops an acute systemic illness with abdominal pain and deranged liver tests. Management consists of intensive antibiotic treatment directed particularly towards Gram-negative organisms and micro-aerophilic streptococci. In some patients the clinical picture associated with portal venous thrombosis may develop after the acute phase has settled.

Protein-losing enteropathy

Rarely, hypoproteinaemia may result from excessive loss of plasma proteins into the gastrointestinal tract. All plasma proteins are affected; those showing the greatest reduction in concentration are the ones with the longest half-lives, including albumin. The resulting oedema is largely due to the low level of albumin, this being the main molecule responsible for the plasma colloid osmotic pressure. Although uncommon, this is an important condition to recognize because the resulting oedema or ascites may overshadow the intestinal symptoms and hence the underlying condition is easily missed.

This condition has been found in association with a wide variety of inflammatory or neoplastic disorders of the small bowel and abdominal lymphatic system. It is an almost inevitable consequence of intestinal lymphangiectasia, an inherited or congenital disorder caused by maldevelopment of the lymphatic system. It is also associated with allergic disorders involving the small bowel. n most of these conditions protein loss is mild and incidental. Severe protein loss occurs mainly in lymphatic disorders and in Ménétrier's disease, a curious condition characterized by the presence of giant gastric rugae.

Clinical features

When severe, the condition is characterized by peripheral oedema and occasionally by ascites and pleural effusions. There is marked hypoalbuminaemia in the absence of liver or renal disease. There may be associated steatorrhoea, particularly if the condition occurs in association with lymphoma of the bowel.

Diagnosis

The diagnosis is made by determining the rate of loss of protein into the intestine using a radioactive label, usually chromium(III) chloride ($^{51}CrCl_3$), which attaches to all plasma proteins, or as chromium-51 albumin, which redistributes the label, to some extent, to other proteins. ^{67}Copper-labelled caeruloplasmin is theoretically a better marker but has a short half-life. The proportion of radioactivity in the stool is measured during the succeeding 4 days; 0.7 per cent is taken as the upper limit of normal. In comparison with plasma radioactivity, more sophisticated measures of plasma clearance can be derived to give quantitative data. Alternatively, measurement of a_1-antitrypsin in the stool has been shown to be a useful marker of protein-losing enteropathy, without necessitating radioactive labelling.

The further diagnosis of the condition is directed towards determining the underlying cause.

Treatment

Treatment is directed towards raising the plasma albumin and correction of the underlying disorder. For example, cases associated with neoplasm of the stomach or colon require surgical resection. Those secondary to coeliac disease, sprue, Whipple's disease, or allergic gastroenteropathies should be treated appropriately.

Jejunoileal bypass

Most patients with massive obesity have fatty infiltration of the liver. After jejunoileal bypass, 55 per cent of them show further fatty change, although this is usually asymptomatic. The increase in fat is due entirely to an accumulation of triglyceride. There is frequently a mild elevation of liver enzymes in serum but this returns to normal once weight reduction has been achieved. The mechanism for the increased fatty change is unknown but it may be a result of protein–calorie malnutrition. An alternative possibility is that the steatosis may be secondary to bacterial overgrowth in the excluded loop of small intestine, as steatosis may diminish with metronidazole therapy.

Acute liver failure may develop in a few patients and is associated with considerable mortality. It is thought to be due to bacterial colonization of the included and excluded small intestine with the production of 'hepatotoxins', which are then absorbed. Treatment consists of intravenous amino acids and broad-spectrum antibiotics. If the condition recurs, further treatment should be given and the ileal bypass reversed.

A micronodular cirrhosis may develop 1 to 6 years after a bypass operation. Histologically, the liver often shows appearances similar to those induced by alcohol. If liver function deteriorates, small-bowel continuity should be restored. Patients may still progress to liver failure and death, but there are reports of the cirrhosis arresting or even reversing with complete recovery of the histology.

Parenteral nutrition and the liver

Abnormalities of serum liver enzymes and bilirubin are commonly seen in patients receiving total parenteral nutrition. Thirty to 60 per cent of patients will show a rise in at least one liver test of greater than 50 per cent of baseline, a rise in alkaline phosphatase being the most frequent abnormality. The changes occur towards the end of the first week and peak between 9 and 12 days. Patients receiving intravenous lipid are particularly at risk, but biochemical cholestasis can occur when no fat is given. Liver histology shows steatosis, mild portal inflammation and fibrosis, bile-duct proliferation, and bile plugs. The changes in serum liver tests and in liver histology are reversible once parenteral nutrition is discontinued, although persistent histological changes have been reported. The abnormal concentrations of liver enzymes may also return to normal if the calorie–nitrogen ratio is lowered by reducing the amount of dextrose given, or may even settle spontaneously if parenteral nutrition is continued without change.

The cause of the intrahepatic cholestasis is unknown. Direct toxicity of the intravenous solutions (especially those containing tryptophan), calorie excess, or a deficiency of essential fatty acids have been proposed as possible mechanisms. A more likely explanation is the possibility of an overgrowth of anaerobic bacteria in the intestine with subsequent production of endotoxin and lithocholic acid, both of which induce liver damage in animals with similar histological features to those seen in humans receiving total parenteral nutrition. Patients who develop a rise in serum transaminases and alkaline phosphatase have a high concentration of lithocholic acid in the bile compared with patients being parenterally fed who do not have abnormal liver tests. Furthermore, metronidazole has been shown to prevent cholestasis developing in these patients. Hence, the situation may be analogous to the cholestasis associated with a jejunolileal bypass (see above).

Peliosis hepatitis

This consists of venous lakes within the liver, which probably occur as a result of sinusoidal ectasia. It may be seen in association with oral contraceptive usage, terminal cachexia from carcinoma, and with androgenic steroid therapy. Clinically there are usually few symptoms, although hepatomegaly may be present. Mild to moderate increases in transaminases may occur. Diagnosis is usually made coincidentally on liver biopsy and the prognosis is that of the underlying condition.

Further reading

Adibi BA, Stanko RT (1984). Perspective on gastrointestinal surgery for the treatment of morbid obesity: the lesson learned. *Gastroenterology* **87**, 1381.

Baddeley RM (1980). Surgical management of severe obesity. In: Truelove SV, Kennedy HJ, eds. *Topics in gastroenterology*, Vol 8. Blackwell Scientific, Oxford.

Bolt RJ (1976). Disease of the hepatic blood vessels. In: Backus HL, ed. *Gastroenterology*, 3rd edn, p 471. Saunders, Philadelphia.

Chatel A *et al.* (1979). L'arteriographie dans la periartérite noueuse. *Journal of Radiology* **60**, 113–20.

Dockerty MB (1972). Primary malakoplakia of the colon. *Mayo Clinic Proceedings* **47**, 114.

Drenick EJ, Fisler J, Johnson D (1982). Hepatic steatosis after intestinal bypass—prevention and reversal by metronidazole, irrespective of protein–calorie malnutrition. *Gastroenterology* **82**, 535.

Lambert JR, Thomas SM (1985). Metronidazole prevention of serum liver enzyme abnormalities during total parenteral nutrition. *Journal of Parenteral Nutrition* **9**, 501.

Lazenby AJ *et al.* (1989). Lymphocytic ('microscopic') colitis: a comparative histopathologic study with particular reference to collagenous colitis. *Human Pathology* **20**, 18–28.

Long R, James O (1974). Polymyalgia rheumatica and liver disease. *Lancet* **i**, 77.

Ranney B (1975). The prevention, inhibition, palliation and treatment of endometriosis. *American Journal of Obstetrics and Gynecology* **123**, 778.

Runyon BA, La Brecque DR, Anuras S (1980). The spectrum of liver disease in systemic lupus erythematosus. *American Journal of Medicine* **69**, 187.

Sheldon GF, Peterson SR, Sanders R (1978). Hepatic dysfunction during hyperalimentation. *Archives of Surgery* **113**, 504.

Sherlock S (1981). *Diseases of the liver and biliary system*, 6th edn. Blackwell Scientific, Oxford.

Sleisenger MH, Fordtran JS (1993). *Gastrointestinal disease*, 4th edn. Saunders, Philadelphia.

Steer HD, Colin-Jones DG (1975). Melanosis coli. Studies of toxic effects of irritant purgatives. *Journal of Pathology* **115**, 119.

Whaley K, Webb J (1977). Liver and kidney disease in rheumatoid arthritis. *Clinics in Rheumatic Diseases* **3**, 527.

15.1.1.1

Introduction

Peter L. Weissberg

The scale of the problem

All organs require an adequate blood supply to survive and function normally, yet in the Western world we are in the midst of an epidemic of disease caused by atherosclerosis in which lipids are deposited in the subendothelial space of major arteries initiating a process that results in narrowing and occlusion of vessels and consequent ischaemia and necrosis of organs and limbs. Although atherosclerosis is manifested most dramatically as acute myocardial and cerebral infarction (heart attack and stroke), it also has a substantial impact on limb, gut, and renal function. Indeed, there is no other single disease that has such a potent and diverse impact on health. Coronary, cerebrovascular, and peripheral vascular disease together account for approximately 40 per cent of male and 30 per cent of female deaths in the United Kingdom in those under 75 years of age. In recent years there has been a gratifying fall in the incidence of deaths from cardiovascular causes in middle-aged adults in the so-called developed world. However, since most heart attacks and strokes occur in the elderly and life expectancy is increasing, the prevalence of cardiovascular disease is bound to increase. Also, there is clear evidence that the epidemic of cardiovascular disease is spreading rapidly, particularly in former Eastern Block countries and, alarmingly, in the Third World. It therefore follows that if atherosclerosis could be understood and overcome, the impact on world health would be substantial.

The atherosclerotic plaque

An atherosclerotic plaque comprises a subendothelial accumulation of oxidized lipid at its core with an inflammatory cell infiltrate covered by a fibrous cap consisting of modified vascular smooth muscle cells and their extracellular matrix. Plaques begin in early life as asymptomatic fatty streaks. Symptoms arise later on, either because a plaque has grown large enough to limit flow, or because a plaque erodes or ruptures causing platelet aggregation and evolution of an occlusive thrombosis within the vessel lumen.

Large plaques eventually declare themselves through the emergence of inducible ischaemia, manifested as angina pectoris (coronary disease, Fig. 1) or intermittent claudication (peripheral vascular disease), which resolves when oxygen demand is reduced, either by stopping the triggering activity or by drug therapy. Plaque erosion or rupture is sudden and usually precipitates an acute ischaemic event such as unstable angina (Fig. 1), myocardial infarction, transient cerebral ischaemia, or stroke depending on the vascular bed involved. Tissue viability then depends on whether or not flow can be restored and how quickly. In the coronary circulation removal of the offending thrombus is achieved either with drug therapy or by angioplasty. Pharmacological and mechanical strategies for restoring cerebral blood flow after ischaemic stroke are less well established.

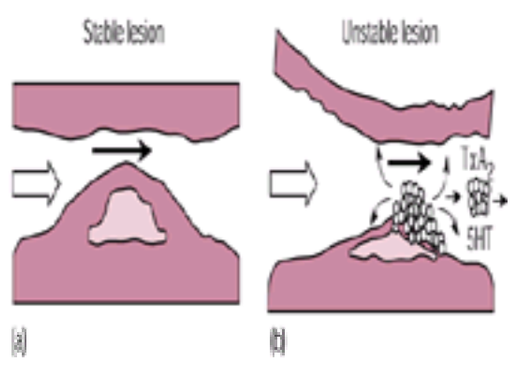

Fig. 1 Mechanisms by which atherosclerotic plaques induce symptoms. (a) Large stable atherosclerotic plaque limits blood flow in response to increased demand causing stable angina. (b) Platelet aggregation and thrombus formation on an unstable atherosclerotic plaque limit blood flow and cause vasoconstriction and embolization at rest, leading to unstable angina (TxA_2 thromboxane A_2, 5HT, 5-hydroxytryptamine). (Redrawn from Weinberg PL (1999). In: *Angina in clinical practice* (ed. P. Schofield). Martin Dunitz Ltd.)

The inflammatory basis of atherosclerosis

It is now recognized that atherosclerosis is a destructive inflammatory process, probably initiated and maintained by modified (oxidized) lipids, involving endothelial dysfunction, and mediated by activated macrophages, T cells, and possibly also mast cells. The inflammatory process is isolated from the circulation by a protective fibrous cap synthesized by intimal vascular smooth muscle cells. If the inflammatory process predominates, then plaque rupture and thrombosis results (Fig. 2). The inflammatory basis of atherosclerosis is reflected in a close correlation between circulating markers of inflammation, for example fibrinogen, C-reactive protein, serum amyloid A, and serum albumen (inverse) and clinical outcome in patients with known disease.

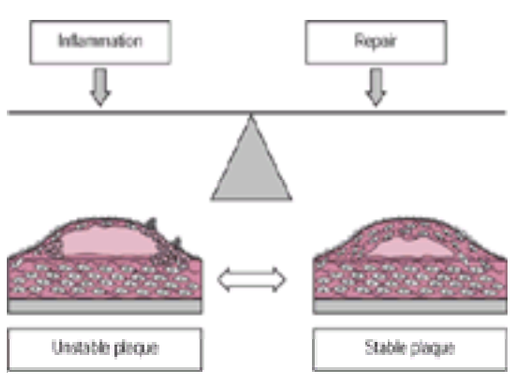

Fig. 2 The balance of atherosclerosis. The stability of atherosclerotic plaque is determined by the net effect of the destructive inflammatory process tending to plaque rupture and thrombosis balanced by the stabilizing influence of the fibrous cap synthesized by intimal smooth muscle cells. (Redrawn from Weinberg PL (1999). *Atherosclerosis*, **147** (suppl. 1), S3–S10.)

The changing treatment of atherosclerosis

Atherosclerosis has traditionally been regarded as a collection of end-stage organ-based diseases such as myocardial infarction, stroke, and limb ischaemia, treated by specialists from a variety of disciplines concerned more with its consequences than its cause. Therapy has therefore relied largely on mechanical procedures such

as bypass surgery, endarterectomy, and balloon angioplasty to relieve obstruction and improve blood flow. However, whilst these procedures have been very successful at relieving symptoms, they have had little impact on survival, except in a few specific patient groups. The main impact on survival has come from advances in drug therapy.

Since thrombosis is the commonest terminal event in atherosclerosis it is not surprising that fibrinolytic, antithrombotic, and antiplatelet drugs are all effective in acute events caused by plaque rupture, such as myocardial infarction, unstable angina, and transient ischaemic attacks. However, the most important recent therapeutic development has been the introduction of inhibitors of HMG Co-A reductase (the statins), which reduce low-density lipoprotein cholesterol and triglycerides, increase high-density lipoprotein cholesterol, and improve endothelial function. Treatment with statins reduces the risk of a vascular event (both heart attack and stroke) by about 30 per cent in patients with and without symptoms of atherosclerosis (secondary and primary prevention) and with high or 'normal' cholesterol levels, yet they produce little, if any, haemodynamically meaningful plaque regression. The optimistic message from these observations is that statins have a greater effect on plaque stability than on plaque size and that atherosclerosis is therefore a dynamic process that can be modified, even in advanced disease.

In the twenty-first century atherosclerosis is best viewed as a single dynamic disease in which complex interactions between endothelial cells, smooth muscle cells, inflammatory cells, and platelets dictate its course. The development of effective medical therapies for atherosclerosis argues for a common approach to its management regardless of its presentation. Thus all patients with established atherosclerotic disease should have the opportunity to benefit from treatment with lipid-lowering and antiplatelet therapies. Since it is now possible to prevent clinical events in asymptomatic patients with occult vascular disease, the challenge is to target expensive therapies to those at highest risk. In the future this is likely to include assessment of 'classical' risk factors (family history, smoking, diabetes, cholesterol levels) but also possibly indices of inflammation and measures of genetic variations (polymorphisms) in molecules known to play a central role in plaque progression and stability.

Further reading

LIPID Study Group (1998). Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. *New England Journal of Medicine* **339**, 1349–57.

Ridker P *et al*. (1997). Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *New England Journal of Medicine* **336**, 973–9.

Ridker P *et al*. (1998). Inflammation, pravastatin, and the risk of coronary events after myocardial infarction in patients with average cholesterol levels. Cholesterol and Recurrent Events (CARE) Investigators. *Circulation* **98**, 839–44.

Ross R (1986). The pathogenesis of atherosclerosis–an update. *New England Journal of Medicine* **314**, 488–500.

Ross R (1993). The pathogenesis of atherosclerosis: a perspective for the 1990s. *Nature* **362**, 801–9.

Ross R (1999). Atherosclerosis—an inflammatory disease. *New England Journal of Medicine* **340**, 115–26.

Ross R, Glomset J (1976). The pathogenesis of atherosclerosis. Part 1. *New England Journal of Medicine* **295**, 369–77.

Ross R, Glomset J (1976). The pathogenesis of atherosclerosis. Part 2. *New England Journal of Medicine* **295**, 420–8.

Sacks FM *et al*. (1996). The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. Cholesterol and Recurrent Events Trial investigators. *New England Journal of Medicine* **335**, 1001–9.

Scandinavian Simvastatin Survival Group (1994). Randomised trial of cholesterol lowering in 4444 patients with coronaryheart disease: the Scandinavian Simvastatin survival study (4S). *Lancet* **344**, 1383–9.

Shepherd J *et al*. (1995). Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *New England Journal of Medicine* **333**, 1301–7.

Treasure CB *et al*. (1995). Beneficial effects of cholesterol-lowering therapy on the coronary endothelium in patients with coronary artery disease. *New England Journal of Medicine* **332**, 481–7.

15.1.1.2Vascular endothelium: its physiology and pathophysiology

P. Vallance

A monolayer of endothelial cells lines the intimal surface of the entire vascular tree (Fig. 1) to form the largest endocrine/paracrine organ in the body. These cells are metabolically very active and exert a profound influence on vascular reactivity, thrombogenesis, and the behaviour of circulating cells. Abnormalities of endothelial function have been implicated in a wide variety of diseases ranging from atheroma and hypertension to acute inflammation and septic shock (Table 1). Recent advances in endothelial research have led to the development of new therapies and re-evaluation of those that exist. This section provides an introduction to the biology of the vascular endothelium and describes how endothelial dysfunction may contribute to cardiovascular disease.

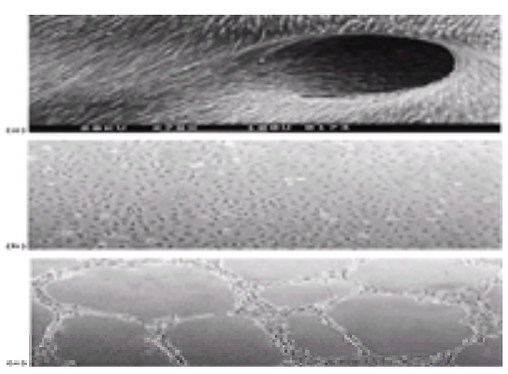

Fig. 1 Panel (a) shows a scanning electron micrograph of the endothelium of a human coronary vessel at a branch area (supplied by P. M. Rowles). Panel (b) shows endothelial cells grown in culture in the absence of angiogenic factors. With stimulation by the correct angiogenic factors, endothelial cells form a tube-like structure (Panel (c)).

Development of endothelium

During early development the endothelium forms the first layer of the circulatory system and extends to produce a network of interconnecting tubes; this ability of endothelial cells to form tube-like structures is retained even when they are grown *in vitro* (Fig. 1). *In vivo*, the endothelial tubes differentiate into arteries, arterioles, capillaries, veins, and lymph vessels, and regional differences in function and structure evolve such that the properties of endothelial cells vary between arterial and venous beds, between micro- and macrovasculature, between organs, and between different parts of individual organs—perhaps the most striking example being the specialized layer of endothelial cells that forms the blood–brain barrier. Heterogeneity of endothelial cell function undoubtedly has implications for physiology, pathophysiology, and therapeutics. However, endothelial cells from different vessels also have many features in common and a number of pathologies, including those causing premature vascular disease, are associated with widespread changes in the behaviour of endothelial cells.

Anatomy of endothelium

Each endothelial cell is between 25 and 50 μm long, 10 and 15 μm wide, and up to 5 μm deep, and lies with its long axis aligned in the direction of the blood flow (Fig. 1). The underlying smooth muscle cells lie radially, are about 5 to 10 μm wide, and taper at either end so that a single endothelial cell can communicate with many smooth muscle cells, and vice versa. The endothelium also comes into intimate contact with circulating cells, and the total area of the luminal surface of endothelium is in excess of 500 m^2. This thin layer of cells is particularly susceptible to injury, and changes in endothelial cell morphology and turnover occur in experimental hypertension, diabetes, and atheroma. Antibodies directed against endothelium can be found in a number of inflammatory and immune conditions.

Signal detection by endothelial cells

The endothelial cell membrane expresses a large number of receptors for circulating hormones, local mediators, and vasoactive factors released from blood cells. It can also sense local changes in pressure and flow; stretch of the cell membrane leads directly to opening of a cation channel permeable to calcium, and flow across the cell surface leads to opening of a potassium channel, which hyperpolarizes the cell. The precise mechanisms linking the various stimuli received to the response of the endothelial cell have yet to be determined, but calcium is undoubtedly important. Receptor occupation, stretch, or shear stress all lead to changes in the concentration of intracellular free calcium, and the profile of change influences which endothelial functions are activated and therefore which message is produced by the cell. In addition, it is clear that the endothelial cell can adjust both the expression and localization of certain key enzymes in response to physical or chemical stimuli. Translocation of enzymes from cytosol to cell surface or to specialized invaginations in the cell surface (caveolae) in response to stimuli can greatly alter the metabolic activity of the endothelial cell. The endothelium acts as a signal transducer and exerts a profound influence on the cardiovascular system by virtue of the way in which it alters its phenotype in response to signals.

Control of vascular tone

Endothelium extracts and inactivates circulating hormones, converts inactive precursors into active products, and synthesizes and releases a variety of vasoactive mediators (Fig. 2). Vasoconstrictor and vasodilator mediators are produced and allow the vessel to respond to changes in the local milieu, but the predominant background influence of the endothelium is dilator, with removal of the endothelium leading to vasoconstriction. Basal endothelium-dependent dilator tone seems to provide a physiological counterbalance to the continuous constrictor tone of the sympathetic nervous system.

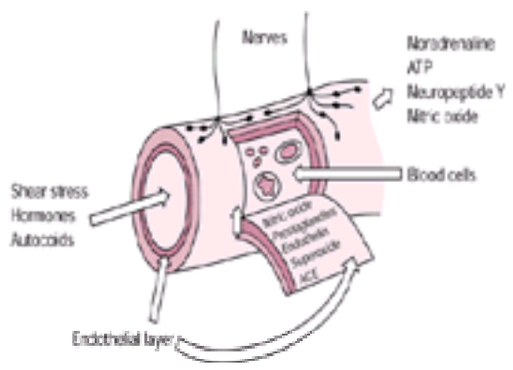

Fig. 2 Vascular endothelial cells lie at the interface between blood and the smooth muscle cells. They detect chemical and physical signals in the lumen of the blood vessel and adjust their output of biologically active mediators accordingly. This provides a mechanism of local regulation of vascular function. Rapid adjustment of vascular tone is probably achieved through a balance of endothelium-derived nitric oxide and neuronally derived noradrenaline. Endothelin provides a slowly modulating constrictor tone and angiotensin II has the capacity to fine-tune neuronal, endothelial, and smooth muscle function. Abbreviations: ACE (angiotensin converting enzyme), ATP (adenosine triphosphate).

Vasodilators

The endothelium produces at least three vasodilator mediators (Fig. 2): nitric oxide, prostanoids, and hyperpolarizing factor.

Nitric oxide

Physiology

The production of nitric oxide is responsible for basal endothelium-dependent dilator tone. This simple gas is a potent vasodilator: its synthesis and main actions through the second messenger cyclic GMP are described in Fig. 3. In addition, nitric oxide inhibits cytochrome c oxidase, initially in a reversible manner, but irreversibly under certain conditions. Inhibition of this enzyme decreases oxygen utilization, and the release of nitric oxide by endothelial cells appears to be an important determinant of oxygen consumption in the vasculature. It is possible that there are additional important targets for nitric oxide, including ion channels and other enzymes, but the physiological significance of these effects is not yet clear. Nitric oxide modifies the adhesiveness of the endothelial cell for circulating white cells, but rapid inactivation by haemoglobin prevents any significant downstream effect.

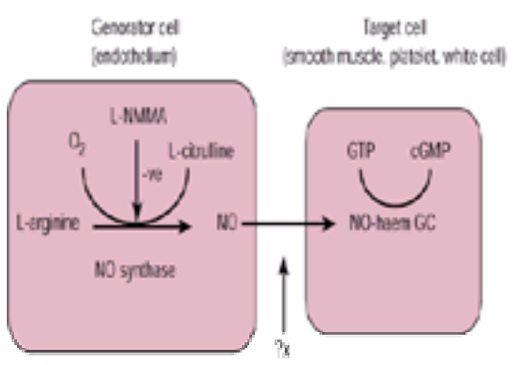

Fig. 3 Nitric oxide (NO) is synthesized from L-arginine by the action of nitric oxide synthase. Citrulline is the byproduct of the reaction. Nitric oxide, which has a half-life of only a few seconds, diffuses from the endothelial cell to reach a target cell—this may be vascular smooth muscle, a platelet, or a white cell. In the process of diffusion nitric oxide may interact with other molecules (X) which may stabilize or destroy the mediator. A major physiological target for nitric oxide is guanylyl cyclase (GC). Nitric oxide binds to the haem moiety of this enzyme and this leads to enzyme activation and generation of cyclic guanosine monophosphate (cGMP). Elevation of cGMP relaxes smooth muscle and inhibits platelet aggregation and adhesion. The platelet-derived mediators adenosine diphosphate (ADP) and 5-hydroxytryptamine (5-HT; serotonin) stimulate endothelial cells to synthesize nitric oxide and this provides a negative-feedback system to prevent activated platelets from causing vasospasm and further platelet adhesion and aggregation. The synthesis of nitric oxide can be inhibited by certain guanidino-substituted analogues of L-arginine such as N^G-monomethyl-L-arginine (L-NMMA) and this substance as well as asymmetric dimethylarginine (ADMA) occurs naturally and may contribute to nitric oxide deficiency in disease states. Nitric oxide synthase activity can be stimulated by calcium, provision of co-factors, phosphorylation of the enzyme, and by altered intracellular localization

Nitric oxide is a free radical (it has an unpaired electron in its outer orbit) and as such reacts readily with other free radicals and reactive oxygen species. The reaction between nitric oxide and superoxide anion (O_2^-) is extremely fast and can result in the formation of either the toxic product peroxynitrite ($ONOO^-$) or the inactive breakdown product nitrate (NO_3^-). Such interactions between radicals can greatly influence the overall behaviour of the wall of the blood vessel and can lead to an apparent defect in endothelial function even when the output of endothelial mediators is normal.

The arterial circulation of animals and humans is vasodilated continuously and actively by endothelium-derived nitric oxide, and inhibition of the synthesis of nitric oxide with certain guanidino-substituted analogues of L-arginine, including N^G-monomethyl-L-arginine, leads to vasoconstriction, hypertension, and sodium retention. Shear stress—the force caused by the viscous drag of flowing blood—is probably an important physiological stimulus for the continuous production of nitric oxide. As shear stress increases more nitric oxide is produced and the blood vessel relaxes, reducing the stress. This process of flow-mediated dilatation appears to be a homeostatic mechanism to prevent shear stress from increasing to levels that might initiate activation of platelets or other cells and may also help co-ordinate tissue perfusion. Flow-mediated dilatation is an autoregulatory property of blood vessels that tends to oppose classical myogenic autoregulation—the process by which a blood vessel constricts in response to an increase in intraluminal pressure.

Synthesis of nitric oxide is stimulated by acetylcholine, bradykinin, and substance P, and in many vessels release of nitric oxide accounts for the vasodilator actions of these mediators, which are known as 'endothelium-dependent vasodilators'. Circulating hormones, including insulin and oestrogens, may also act on receptors on or within the endothelial cell to stimulate the release of nitric oxide acutely or to alter the expression of endothelial nitric oxide synthase chronically.

Veins differ from arteries and arterioles, and do not seem to be actively dilated by continuous release of nitric oxide. Venous endothelium releases nitric oxide when stimulated by acetylcholine or bradykinin, but not under basal conditions. Furthermore, human veins do not release much nitric oxide in response to platelet-derived mediators. Indeed, aggregating platelets constrict veins, due to the unopposed action of the platelet-derived mediators on vascular smooth muscle. The reasons for the arteriovenous difference in nitric oxide production are not fully understood, but one consequence is that the guanylyl cyclase in venous smooth muscle is relatively upregulated and veins respond to maller amounts of nitric oxide than do arteries or arterioles. This is of therapeutic relevance; nitric oxide is the active moiety of glyceryl trinitrate and other nitrovasodilators, and the low basal synthesis of endogenous nitric oxide by venous endothelium accounts, in part, for the venoselective action of these drugs.

Pathophysiology

Loss of nitric oxide leads to arterial vasoconstriction, has the potential to enhance platelet and white cell adhesion, and in experimental models may enhance atherogenesis. Several clinical conditions, including atherosclerosis, hypertension, hypercholesterolaemia, and diabetes, are associated with a functional loss of nitric

oxide-mediated effects.

In the coronary vasculature, loss of nitric oxide predisposes to vasospasm and may contribute to the onset of anginal symptoms. Atherosclerotic coronary arteries constrict in response to the platelet-derived mediator serotonin (5-hydroxytryptamine), whereas healthy vessels are stimulated to produce more nitric oxide and dilate (Fig. 3). Flow-dependent dilatation is also lost in such vessels, and the response to sympathetic stimulation is converted from dilatation to unopposed constriction. Endothelial dysfunction precedes the development of overt atheroma and there is a relationship between risk factors for ischaemic heart disease and impaired responsiveness of coronary arteries to endothelium-dependent vasodilators. Furthermore, hypercholesterolaemia, even in the absence of angiographic evidence of atheroma in large vessels, is associated with abnormal endothelium-dependent vasodilatation in coronary and peripheral arterioles. Modified low-density lipoproteins appear to inhibit nitric oxide synthesis or speed its destruction, possibly by enhancing production of superoxide anion.

Basal endothelium-dependent dilatation is also impaired in patients with essential hypertension and the degree of impairment increases with increasing blood pressure. It is not known whether the defect is a consequence or a cause of the raised pressure, but the fact that endothelial function appears to be restored by antihypertensive therapy argues in favour of dysfunction being a response to raised pressure. Patients with diabetes show diminished endothelium-dependent dilatation, and this defect does not reverse with treatment. Thus patients with uncontrolled hypertension, diabetes, and hypercholesterolaemia all display defects of nitric-oxide-mediated vasodilatation and this could provide a common mechanism of vascular dysfunction in these diseases. Impaired endothelium-dependent dilatation associated with hypercholesterolaemia is partially reversed by supplementation with L-arginine (see below).

In addition to the effects of disease states on nitric oxide, genetic variation in endothelial nitric oxide synthase may predispose to cardiovascular disease. Common variations occur in promoter and coding regions of the gene encoding endothelial nitric oxide synthase and certain variants appear to be associated with excess cardiovascular risk, but the data are not yet conclusive. Nor is it clear how the genetic variations alter nitric oxide synthesis or endothelial cell function.

Overproduction of nitric oxide may also contribute to disease. Bacterial endotoxin, and certain cytokines, including interleukin 1 and interferon-g, induce expression of a second nitric oxide synthesizing enzyme which appears in the endothelium, vascular smooth muscle, and inflammatory cells invading the vessel wall. Unlike the constitutive enzyme present in healthy endothelium (endothelial nitric oxide synthase), this inducible isoform of nitric oxide synthase is not regulated by calcium and produces large amounts of nitric oxide. In these quantities nitric oxide, either alone or in combination with superoxide, may contribute to tissue damage in addition to causing profound vasodilatation and hypotension. Excess production of nitric oxide from endothelial nitric oxide synthase due to stimulation of certain essential cofactors including tetrahydrobiopterin may also contribute to these effects. The therapeutic potential of specific inhibitors of nitric oxide synthase as anti-inflammatory agents has not yet been established. In contrast to the general increase in nitric oxide seen in response to inflammation, certain proinflammatory cytokines (particularly tumour necrosis factor-a) may impair normal endothelium-dependent relaxation and nitric oxide synthesis. This might be an important mechanism linking infection or inflammation to increased risk of cardiovascular events including arterial or venous thrombosis.

Prostanoids

Nitric oxide appears to be the dominant vasoactive factor released from endothelial cells under basal conditions, but it is by no means the only mediator produced. The endothelium is a rich source of prostanoids, including the vasodilators prostacyclin and the prostaglandins E_2 and D_2. However, whereas inhibition of nitric oxide leads to profound and widespread changes in vascular tone, inhibition of prostanoid synthesis with aspirin, or other non-steroidal anti-inflammatory drugs, does not. Renal vasculature is an exception, and dilator prostanoids do appear to be important in the regulation of basal renal blood flow; aspirin and other non-steroidal anti-inflammatory drugs lead to vasoconstriction in the kidney, indicating tonic release of vasodilator prostanoids in this vascular bed. Furthermore, in the fetus and newborn, indomethacin leads to closure of the ductus arteriosus and a fall in cerebral blood flow suggesting a significant contribution of endothelium-derived prostanoids to tonic vasodilatation in these beds, at least during development. Cerebral blood flow in adults also falls in response to indomethacin, but not to aspirin and other cyclo-oxygenase inhibitors, and so the role of prostanoids is unclear. Vasodilator prostanoids are important in the vascular changes of inflammation, although whether the prostanoids derive exclusively from the endothelium is not known. A cytokine-inducible isoform of cyclo-oxygenase (cyclo-oxygenase II) has been identified and this probably contributes to the increased synthesis of vascular prostaglandins in inflammation.

Hyperpolarizing factor

An endothelium-derived hyperpolarizing factor has been identified in certain animal and human blood vessels. Hyperpolarization of vascular smooth muscle cells leads to a fall in calcium entry and vascular relaxation. Increasing evidence suggests that endothelium-dependent hyperpolarization may be particularly important in small arteries and arterioles. The chemical identity of endothelium-derived hyperpolarizing factor has not been clearly established, but products of activity of cytochrome P450, the cannabinoid anandamide, and the potassium ion have all been suggested as possible candidates. Only when a clear identify for endothelium-derived hyperpolarizing factor has been established, or a specific inhibitor produced, will it be possible to determine the precise role of this mediator in vascular physiology and pathophysiology.

Vasoconstrictors

Although the predominant background influence of the endothelium is dilator, important vasoconstrictor factors are also synthesized and released.

Endothelin

The endothelins are a family of potent vasoconstrictor peptides containing 21 amino acids that are closely related to the snake venom toxin of the Israeli burrowing asp. Three types of endothelin have been described, endothelin 1, 2, and 3, and there are at least two endothelin receptors in human blood vessels: endothelin A receptor and endothelin B receptor. Endothelins vasoconstrict and can promote the growth of vascular smooth muscle cells. Effects are mediated in part through stimulation of increases in calcium and in part through calcium-independent mechanisms including activation of protein kinases.

Endothelin 1 is synthesized from 'big endothelin' within human endothelial cells (Fig. 4). It is a potent and long-lasting constrictor of human blood vessels, and causes widespread vasoconstriction, hypertension, and sodium retention when infused into healthy volunteers. Antagonists of the endothelin A receptor cause vasodilatation when infused locally, and mixed endothelin A/B antagonists lower blood pressure in normotensive and hypertensive individuals. These findings indicate that there is a tonic synthesis and release of endothelin. A number of studies suggest that there may be important interactions between the sympathetic nervous system, the renin–angiotensin system, and the endothelin system and that these may act in concert to control constrictor tone, with the endothelin system providing a slowly modulating background constrictor tone.

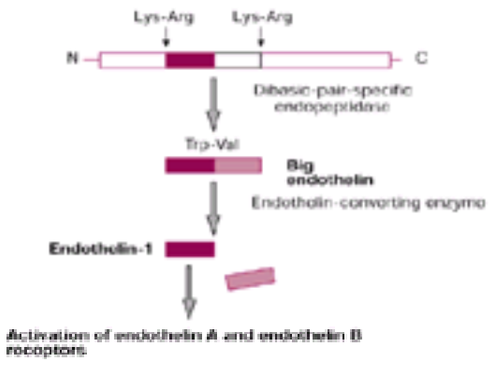

Fig. 4 Biosynthesis of the vasoconstrictor peptide endothelin 1. Endothelin 1 produces profound vasoconstriction, through activation of endothelin A receptors. Activation of endothelin B receptors on smooth muscle also causes vasoconstriction, but stimulation of endothelin B receptors on the endothelium stimulates nitric oxide and prostacyclin production.

Paradoxically, endothelin can also produce transient vasodilatation. This seems to be mediated by activation of endothelin B receptors. Although activation of

endothelin B receptors on vascular smooth muscle causes constriction, activation of endothelial endothelin B receptors leads to the generation of vasodilator prostanoids and/or nitric oxide. Binding of endothelin to endothelin B receptors also seems to be important to clear the peptide from the circulation. Stimuli for endothelin production include thrombin, insulin, cyclosporine, adrenaline, angiotensin II, cortisol, various proinflammatory cytokines, hypoxia, and shear stress.

The concentrations of endothelins circulating in plasma are low and may not reflect local concentrations achieved within the vessel wall. It is difficult to interpret the elevated values reported in many conditons. A role for endothelin in the pathogenesis of vasospasm associated with subarachnoid haemorrhage and some types of renal ischaemia is suggested by results from experiments in animals, and endothelin A/B antagonists produce short-term changes in haemodynamics in patients with heart failure that suggest a possible beneficial therapeutic effect. It seems likely that the precise role of endothelin in these diseases, and other vasospastic conditions, will become apparent as inhibitors of endothelin synthesis, and specific endothelin receptor blockers become more widely used in clinical trials.

Increased production of endothelin has been clearly implicated in the pathogenesis of a very rare form of secondary hypertension caused by malignant haemangioendothelioma, a vascular tumour characterized by intravascular proliferation of atypical endothelial cells. In this condition the degree of hypertension correlates with plasma levels of endothelin and when the tumour is removed blood pressure and plasma endothelin levels fall.

Angiotensin converting enzyme

Angiotensin converting enzyme is located primarily on the luminal surface of the endothelium. This enzyme converts angiotensin I to angiotensin II and also metabolizes bradykinin to inactive products (see Fig. 2). The pulmonary vasculature provides the largest area of endothelium, and is important in the regulation of circulating levels of angiotensin II, but activity of endothelial angiotensin converting enzyme in systemic vessels may be more important in determining the final concentrations of angiotensin II and bradykinin reaching the blood vessel wall. Furthermore, endothelial cells also have the ability to synthesize renin and its substrate. It therefore seems as though the enzymatic machinery for a complete renin–angiotensin system is present within the vessel wall.

The activity of the renin–angiotensin system is clearly important in cardiovascular diseases including hypertension and heart failure, but the relative importance of local compared with systemic regulation of angiotensin II production is not yet clear. Furthermore, the full clinical significance of bradykinin metabolism by endothelial angiotensin converting enzyme (see Fig. 2) has yet to be determined. It has been demonstrated that at least part of the vasodilator action of angiotensin converting enzyme inhibitors in certain isolated blood vessels is due to accumulation of bradykinin which stimulates nitric oxide synthesis.

Prostanoids

The endothelium synthesizes thromboxane and the unstable prostaglandin endoperoxides prostaglandin G_2 and prostaglandin H_2. Overproduction of constrictor prostanoids by the endothelium has been implicated in animal models of diabetes and hypertension, but the significance of these findings for human disease remains uncertain.

Superoxide

The superoxide anion (O_2^-) is synthesized within endothelial cells. There are several possible enzymatic sources including co-factor deplete nitric oxide synthase and cyclo-oxygenase. In neutrophils, NADH/NADPH oxidase is the major source of superoxide. Components of this system have been detected in endothelial cells and they are now assumed to be the major site of superoxide generation. Superoxide is usually destroyed by superoxide dismutase, but under certain conditions it seems as though it may act as an endothelium-derived contracting factor or interact with nitric oxide as described above.

Regulation of platelet function and haemostasis

The endothelium synthesizes and releases prothrombotic and antithrombotic factors. However, healthy endothelium presents a thromboresistant surface, indicating that the antithrombotic factors predominate under basal conditions.

Platelets

Endothelial cells inhibit the aggregation and adhesion of platelets, and disaggregate aggregating platelets. Two mediators are of particular importance: nitric oxide and prostacyclin (or prostaglandin E_2 in microvascular endothelium). They act synergistically through different second messenger systems: cyclic guanosine monophosphate for nitric oxide and cyclic adenosine monophosphate for prostacyclin.

Thiols and sulphydryl-containing molecules react with nitric oxide to produce more stable adducts, including nitrosocysteine, nitrosoglutathione, nitrosoalbumin, and even nitrosohaemoglobin. Some of these compounds are formed *in vivo* and may enhance the antiplatelet effects of endothelium-derived nitric oxide. Furthermore, interaction between nitric oxide and tissue plasminogen activator leads to the formation of nitrosotissue plasminogen activator, a molecule with fibrinolytic, antiplatelet, and vasorelaxant properties. It is not yet clear how important these nitric oxide adducts are in human physiology or pathophysiology.

Deficient production of nitric oxide has been implicated in a wide variety of cardiovascular diseases (see above) and abnormalities of prostanoid synthesis occur in experimental models of atherosclerosis and diabetes. In the presence of a quiescent healthy endothelium, loss of basal nitric oxide alone does not lead to significant systemic platelet activation. However, loss of nitric oxide and prostacyclin at sites of endothelial damage, dysfunction, or activation promotes the formation of platelet aggregates and may contribute to thrombosis and vessel occlusion. In animals, stenosed endothelium-denuded vessels lead to cyclical variations in flow as platelets stick to the vessel wall and release vasoactive and proaggregant mediators. If this also occurs in human vessels *in vivo*, it might be an important mechanism of vasospasm and thrombosis.

Under basal conditions the endothelium inhibits platelet activation, but in response to certain stimuli, proaggregant, proadhesive mediators may be synthesized and released. Unstable prostaglandin endoperoxides activate platelets, platelet activating factor may be produced, and von Willebrand factor, which is synthesized and stored within endothelial cells, increases platelet adhesion. These changes occur in response to inflammatory mediators and may also result from repeated endothelial 'injury'.

Coagulation

Heparan sulphate is a glycosaminoglycan closely related to heparin, but less potent, which is found on the surface of endothelial cells. Antithrombin III is also expressed on the endothelial cell surface and together with heparan sulphate provides a mechanism for binding and inactivating thrombin. In addition, endothelial cells participate in the activation of the anticoagulant protein C; protein S is secreted and thrombomodulin is found on the cell surface.

In the quiescent state, expression of anticoagulant factors predominates, but when activated the endothelium may promote coagulation. Receptors for clotting factors appear on the endothelial surface, von Willebrand factor is secreted, and tissue factor—the principal cellular initiator of coagulation—is expressed. Bacterial endotoxin, inflammatory cytokines, and glycosylated proteins activate the endothelium and shift the balance in favour of coagulation. This may occur in response to infection, inflammation, or endothelial injury. Circulating levels of von Willebrand factor are increased in certain patients with diabetes or hypertension.

Fibrinolysis

The endothelial cell surface has a fibrinolytic pathway. Urokinase and tissue plasminogen activator are secreted and there are specific binding sites for plasminogen activators and plasminogen. Thrombin, adrenaline, vasopressin, and stasis of blood may be physiological stimuli for the release of tissue plasminogen activator from human endothelium.

Plasminogen activator inhibitor 1 is also synthesized and bound by endothelium, providing a pathway for local inhibition of the fibrinolytic system. Under basal conditions fibrinolysis is dominant, but the balance may be altered by a variety of local and circulating factors, including inflammatory cytokines and the atherogenic particle lipoprotein (a), which inhibits plasminogen binding and hence plasmin generation. In the presence of atherosclerosis the fibrinolytic properties of endothelium are diminished.

Cellular adhesion

The resting endothelium prevents cells from adhering fully to the vessel wall but allows leucocytes to roll along its surface. The regulation of 'rolling', adhesion, and migration is governed largely by specialized glycoproteins known as cell adhesion molecules, which are expressed in varying amounts on the endothelial cell surface and interact with complementary adhesion molecules on circulating cells. Endothelial-leukocyte adhesion molecule 1 (also known as E-selectin), vascular adhesion molecule 1, intercellular adhesion molecule 1, and P-selectin (also known as GMP 140) are all expressed on cytokine-activated endothelium. The degree of expression and the type of adhesion molecules expressed determines the 'stickiness' of the endothelium for different cell types.

Expression of adhesion molecules is an important mechanism of cellular adhesion during inflammation and is also important in recruitment of T cells and monocytes in atherosclerosis. Increased expression of endothelial-leukocyte adhesion molecule 1 is seen in the coronary arteries of transplanted hearts and has been implicated in the rapid development of atherosclerosis in these vessels. Nitric oxide and prostacyclin inhibit the adhesion of white cells to endothelium and this effect may be mediated by changes in the expression or configuration of adhesion molecules. Certain endothelial cell adhesion molecules are shed into the plasma: changes in their concentration have been detected in a variety of cardiovascular diseases, but the significance of this is uncertain.

Proinflammatory cytokines

Cytokines are released from activated leucocytes in response to infection and immunological stimulation and are also produced by the vessel wall itself; interleukins 1, 6, and 8, and colony stimulating factors are synthesized by endotoxin-stimulated endothelial cells, and tumour necrosis factor by human smooth muscle cells. A large number of cytokines alter endothelial functions, upsetting the balance of vasoactive mediators, altering thrombotic activity and the expression of adhesion molecules, or initiating apoptosis (programmed cell death). Interleukin 1 and certain other proinflammatory cytokines alter the synthesis of nitric oxide (see above) and a variety of prostaglandins, enhance the generation of thrombin, platelet activating factor, von Willebrand factor, and plasminogen activator inhibitor, alter endothelial permeability, increase expression of intercellular adhesion molecule 1 and vascular adhesion molecule 1, and may also cause endothelial cell damage and death. These findings are of direct relevance to the vascular changes occurring in inflammation and sepsis, but might also provide a link between acute or chronic immunological stimulation (for example infection) and the development of cardiovascular disease including atherosclerosis or acute cardiovascular events.

Cell growth and angiogenesis

The endothelium of healthy differentiated vessels inhibits proliferation of the underlying smooth muscle. Endothelium-derived vasodilator, antiplatelet, and antithrombotic mediators (nitric oxide, prostacyclin) tend to inhibit the growth of vascular smooth muscle cells whereas vasoconstrictor and prothrombotic mediators (endothelin, angiotensin) tend to promote it. Thus the basal state of the endothelium, in which dilatation and thromboresistance predominates, also prevents the growth of smooth muscle. The heparin-like molecules prevent cell growth and molecules similar or identical to platelet-derived growth factor and fibroblast growth factor are endothelium-derived growth promoters. Others such as transforming growth factor-b produced by endothelial cells may either inhibit or promote cell growth, and the precise role of this molecule *in vivc* is unclear. The basal antiproliferative effects of the endothelium may retard the development of atherosclerosis and intimal proliferation.

In addition to affecting the growth of underlying smooth muscle, endothelial cells are essential for the formation of new blood vessels. The ability of endothelial cells to initiate the formation of new vessels (angiogenesis and vasculogenesis) is retained in adults, but the only place this occurs physiologically is in the female reproductive tract. However, angiogenesis occurs in a wide range of disease states including atherosclerosis, rheumatoid arthritis, and tumour growth and during wound healing or in response to ischaemia. Positive and negative regulators of angiogenesis have been identified and a wide variety of cytokines, growth factors, and local autacoids can act alone or in concert to promote endothelial cell growth, migration, and tube formation. Of particular interest is vascular endothelial growth factor, a growth factor produced by smooth muscle cells in response to hypoxia, inflammatory cytokines, and certain other growth factors. There is good evidence that vascular endothelial growth factor can promote angiogenesis in a variety of animal models and in humans. Intriguingly, it appears as though vascular endothelial growth factor can increase the production of nitric oxide by endothelial cells and this may be one of the effector molecules mediating some of the actions of this growth factor. In order to form tubes through tissues, endothelial cells must degrade matrix and they are capable of synthesizing and releasing a variety of matrix metalloproteinases. Some of these matrix metalloproteinases may in turn affect endothelial function by regulating cell attachment, proliferation, and migration. Failure of endothelial cells to initiate appropriate angiogenesis in response to ischaemia may lead to tissue hypoxia, whilst excessive or inappropriate angiogenesis may contribute to a sustained inflammatory response in the vessel wall, disrupt vessel wall architecture, or lead to haemorrhage into atherosclerotic plaques.

Transport and metabolism

The endothelium presents a permeability barrier for molecules in the bloodstream. Transfer of molecules from the bloodstream into the vessel wall across the endothelium can occur by transport through the endothelial cells or between them. The junctions between endothelial cells are maintained by specialized molecules, including cadherins, and are actively regulated. Transport between cells occurs when endothelial cells contract to leave intercellular gaps. This is an important mechanism for formation of localized oedema. Transport through cells occurs by transcytosis and is an important mechanism for the passage of certain macromolecules, including insulin. In addition, specialized channels for transport of water have been identified—the aquaporins.

The endothelium is intimately involved in lipid metabolism. Lipoprotein lipase is located on the endothelial cell surface and receptors for low-density lipoproteins are present in varying amounts. In quiescent endothelium lipoprotein lipase is active but there are few low-density lipoprotein receptors, indicating that healthy endothelium provides a barrier for entry of low-density lipoprotein into the vessel wall. However, under conditions in which low-density lipoprotein is taken into the endothelium, modification by oxidation occurs and this step may stimulate atherogenesis.

Therapeutic implications

The balance of mediators produced by quiescent healthy endothelium promotes vasodilatation, inhibits activation of platelets and white cells, prevents thrombosis, prevents the growth of smooth muscle cells, and does not support angiogenesis. However, the endothelium also has the capacity to constrict blood vessels, promote cellular adhesion, initiate thrombosis, stimulate the growth of smooth muscle cells, and initiate the formation of new vessels. Whilst all of these properties of the endothelium may be considered as appropriate in the correct physiological context, the term 'endothelial dysfunction' is usually taken to mean impairment of the usual vasodilator and thromboresistant properties of the endothelium and this seems to be a marker of enhanced atherogenesis and increased cardiovascular risk. Therapeutic implications of endothelial dysfunction are now emerging.

Certain therapeutic interventions cause endothelial damage. Antibodies directed against the endothelium are found after heart transplantation and endothelial dysfunction may contribute to the rapid development of coronary artery disease seen in transplant recipients. Balloon angioplasty leads directly to severe disruption of the endothelium and this has been implicated in the development of postangioplasty vasospasm, thrombosis, and restenosis due to the growth of smooth muscle. Venous coronary artery bypass grafts are more prone to occlusion than arterial grafts, and this may reflect differences between arterial and venous endothelium including reduced basal release of nitric oxide by venous endothelium or differential production of growth factors. Acute disruption of endothelial function may promote vasospasm, thrombosis and occlusion, while chronic changes enhance atherogenesis.

Drugs also affect endothelial function. Nitrovasodilators mimic endogenous nitric oxide: glyceryl trinitrate is metabolized to nitric oxide within the vessel wall while sodium nitroprusside liberates nitric oxide spontaneously. Like the endogenous mediator, certain nitrovasodilators inhibit platelet activation and this may provide an additional mechanism to explain the beneficial effects of these drugs in coronary artery disease. Other drugs of benefit in acute coronary artery disease may also replace endothelial mediators, or restore a healthy balance of mediators released by endothelium and blood-borne cells; heparin mimics heparan sulphate proteoglycans on the cell surface, plasminogen activators replace the endogenous molecule, and low-dose intermittent aspirin preferentially inhibits thromboxane synthesis in platelets while sparing endothelial prostanoid production. Furthermore, angiotensin converting enzyme inhibitors block the breakdown of bradykinin and this mediator stimulates the release of nitric oxide from endothelial cells. Recently it has been demonstrated that oestrogens modify endothelial function and enhance endothelium-dependent vasodilatation. Antioxidants may restore a redox balance within the vessel wall which reduces superoxide levels and protects and stabilizes nitric oxide. In experimental systems fish oils may enhance the generation of antiplatelet prostanoids and increase the generation of nitric oxide. Thus all of these interventions appear to work in part by affecting endothelial function and restoring the usual vasculoprotective balance.

There has been considerable interest in the possibility that the amino acid arginine might have therapeutic utility. Theoretically giving extra arginine should have no effect on the synthesis of nitric oxide since the intracellular concentrations of arginine are far in excess of the amounts needed for nitric oxide synthesis. However, in

practice several studies have shown that arginine supplementation restores endothelium-dependent relaxation towards normal in patients with certain conditions including hypercholesterolaemia and some types of renal failure. The discrepancy between biochemical theory and experimental observation is known as the 'arginine paradox'. Improvement of endothelium-dependent relaxation is a surrogate end-point of uncertain significance and clinical trials of the effects of arginine on clinically relevant end-points are currently under way.

A greater understanding of the normal protective functions of vascular endothelium and how these are altered by disease is bound to lead to therapies designed to modify endothelial function. New drugs based on nitric oxide, endothelin, adhesion molecules, and growth factors are in development and likely to enter clinical practice, and seeding of genetically altered endothelial cells on to blood vessels or genetic manipulation of the expression of enzymes that generate key endothelial mediators is a possibility. Indeed gene transfer experiments with endothelial nitric oxide synthase and vascular endothelial growth factor have already shown promise in studies in animals and appear to be technically feasible in humans.

Further reading

Feletou M, Vanhoutte PM (1999). The alternative: EDHF. *Journal of Molecular and Cellular Cardiology* **31**, 15–22.

Furchgott RF, Zawadzki JV (1980). The obligatory role of endothelial cells in the relaxation of arterial smooth muscle. *Nature* **288**, 373–6.

Gerlach H, Esposito C, Stern DM (1990). Modulation of endothelial hemostatic properties: an active role in the host response. *Annual Review of Medicine*.**41**, 15–24.

Hayden MR, Reidy M (1995). Many roads lead to atheroma. *Nature Medicine* **1**, 22–3.

Haynes WG, Webb DJ (1998). Endothelin as a regulator of cardiovascular function in health and disease. *Journal of Hypertension* **16**, 1081–98.

Isner JM, Asahara T (1999) Angiogenesis and vasculogenesis as therapeutic strategies for postnatal neovascularization. *Journal of Clinical Investigation* **103**, 1232–6.

Kinlay S, Libby P, Ganz P (2001). Endothelial function and coronary artery disease. *Current Opinion in Lipidology*, **12**, 383–9.

Krishnaswamy G *et al.* (1999). Human endothelium as a source of multifunctional cytokines: molecular regulation and possible role in humans disease. *Journal of Inteferon and Cytokine Research* **19**, 91–104.

Lüscher TF *et al.* (1992). Endothelium-derived contracting factors. *Hypertension* **19**, 117–30.

Mason JC, Haskard DO (1994). The clinical importance of leucocyte and endothelial cell adhesion molecules in inflammation. *Vascular Medicine Review* **5**, 249–75.

Maxwell AJ, Tsao PS, Cooke JP (1998). Modulation of nitric oxide synthase pathway in atherosclerosis. *Experimental Physiology* **83**, 573–84.

Noll G, Luscher TF (1998). The endothelium in acute coronary syndromes. *European Heart Journal*, **19** (Suppl. C), C30–C38.

Panes J, Perry M, Granger DN (1999). Leukocyte-endothelial cell adhesion: avenues for therapeutic intervention. *British Journal of Pharmacology* **126**, 537–50.

Papapetropoulos A, Rudic RD, Sessa WC (1999). Molecular control of nitric oxide synthases in the cardiovascular system. *Cardiovascular Research* **43**, 509–20.

Ross R (1999). Atherosclerosis—an inflammatory disease. *New England Journal of Medicine* **340**, 115–26.

Vallance P (1998). Nitric oxide in the human cardiovascular system. *British Journal of Clinical Pharmacology* **45**, 433–9.

Vallance P, Collier J, Bhagat K (1997). Infection, inflammation and infarction: does acute endothelial dysfunction provide a link? *The Lancet* **349**, 1391–2.

Vane JR, Bakhle YS, Botting RM (1998). Cyclooxygenases 1 and 2. *Annual Review of Pharmacology and Toxicology* **38**, 97–120.

15.1.1.3 Vascular smooth muscle cells

Peter L. Weissberg

Introduction

Normal human arteries comprise an intima made up of a single layer of endothelial cells and a few underlying vascular smooth muscle cells, separated from the tunica media by an internal elastic lamina. The media comprises only vascular smooth muscle cells and elastic lamellae arranged circumferentially. The number of layers of vascular smooth muscle cells varies, being fewest in small arterioles and greatest in large arteries like the aorta. On the outer boundary of the media is the external elastic lamina that separates the medial vascular smooth muscle cells from the adventitia containing connective tissue, small blood vessels (the vasa vasora), nerves, and adventitial myofibroblasts. The main function of the medial vascular smooth muscle cells is to contract and relax in response to exogenous stimuli, thereby altering the calibre of the arterial lumen and regulating vascular tone. However, unlike cardiac and skeletal myocytes, which are terminally differentiated and can only perform a contractile role, mature vascular smooth muscle cells are highly reactive and can respond to changes in their extracellular environment by dramatic alterations in gene expression, a process often referred to as phenotypic modulation. Thus vascular smooth muscle cells have the ability to migrate and proliferate and to change from a contractile cell to one that produces large quantities of extracellular matrix and other proteins (Fig. 1). This phenotypic plasticity means that vascular smooth muscle cells can play multiple roles in the pathogenesis and progression of vascular disease.

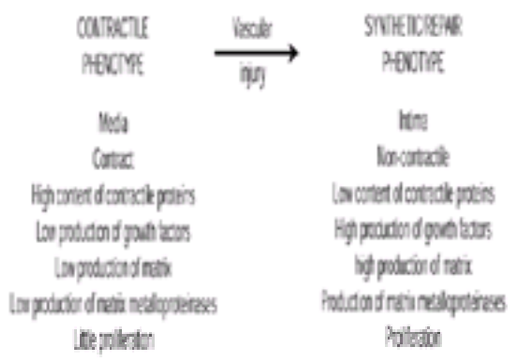

Fig. 1 Characteristics of vascular smooth muscle cell phenotypes.

Vascular development

The versatility of the vascular smooth muscle cell is best exemplified during vascular development. Most of what is understood about this process comes from observations in developing rodents, in particular rats and mice, where the vascular smooth muscle cell is the main cell responsible for vessel growth and development. Blood vessels begin to develop in very early fetal life as endothelial tubes that recruit cells from the surrounding tissues to form the vascular smooth muscle cells of the vessel media. Vascular smooth muscle cells in some vessels are derived from mesoderm and in others from neural crest tissue or both under the influence of local morphogens. It still remains to be determined whether the heterogeneous origin of vascular smooth muscle cells in different arteries has any influence upon development of disease, and in particular the propensity of some, particularly the coronary arteries, to develop atherosclerosis.

Expression of smooth muscle specific genes is first detectable in early fetal development, implying an early commitment to muscle development, but expression of contractile proteins remains low until well after birth. During late fetal and very early neonatal development vascular smooth muscle cells proliferate in the media as the vessels grow. However, shortly after birth, the proliferation of vascular smooth muscle cells decreases rapidly and thereafter vessel growth is achieved by a combination of vascular smooth muscle cell hypertrophy (enlargement) and, particularly, accumulation of extracellular matrix. As blood pressure increases after birth there is an increase in expression of the contractile proteins required to regulate vascular tone, such as smooth muscle myosin heavy chain and actin, and a corresponding decrease in production of matrix and basement membrane proteins. Thus, mature vascular smooth muscle cells in the adult vascular wall undergo little if any proliferation, contain abundant myofilaments, and produce only small amounts of extracellular matrix proteins. This 'contractile' phenotype is maintained thereafter by the combined influences of mechanical forces and extracellular matrix components, particularly sulphated proteoglycans and basement membrane proteins, which signal into the cell and dictate gene expression via receptors on the vascular smooth muscle cell surface.

Vascular smooth muscle cell phenotype

When adult vascular smooth muscle cells are removed from their extracellular environment and placed in cell culture, they immediately reduce production of contractile proteins and increase production of matrix proteins, in particular collagen and elastin. This change from a contractile phenotype to what has been called a 'synthetic' phenotype is characterized ultrastructurally by loss of contractile myofilaments and a dramatic increase in synthetic organelles. The cells also become more responsive to growth factors and gain the ability to take up lipids and to elaborate matrix-degrading enzymes (Fig. 1). This phenotypic change does not occur if the cells are maintained in a medium containing proteoglycans and basement membrane proteins that mimic their natural extracellular environment in the vessel wall, indicating that vascular smooth muscle cells posses an inherent tendency to 'default' to a phenotype resembling that of the developing vessel if not actively stimulated to do otherwise.

Electron microscopic studies have shown that intimal vascular smooth muscle cells in atherosclerosis contain a higher proportion of synthetic organelles and fewer myofilaments than medial vascular smooth muscle cells and therefore resemble the synthetic phenotype observed in cell culture. These observations contributed to the emergence of the 'response to injury' hypothesis of atherosclerosis, initially proposed in the 1960s and still widely quoted today. This hypothesis proposed that the initiating event in atherosclerosis was endothelial cell loss or injury leading to local platelet aggregation and recruitment of vascular smooth muscle cells into the intima where a switch to the synthetic phenotype facilitated proliferation and accumulation of lipids to form the atherosclerotic lesion. This paradigm therefore portrayed the vascular smooth muscle cell as being central to the initiation and maintenance of the atherogenic process, a view reinforced by the contemporaneous suggestion that restenosis after angioplasty was due to excessive proliferation of intimal vascular smooth muscle cells. However, over the past 20 years this view of the role of vascular smooth muscle cells has changed completely, such that vascular smooth muscle cells are now considered to be the main cell type in the plaque protecting against the thrombotic complications of atherosclerosis.

The response of the vascular smooth muscle cell to injury

When the intima of an artery is damaged, for example by a balloon catheter, there is a wave of DNA synthesis in medial vascular smooth muscle cells which is quickly followed by migration of vascular smooth muscle cells into the intima where they change phenotype to form a neo-intima comprising vascular smooth muscle cells and their extracellular matrix (Fig. 2). It remains unclear whether the neo-intimal population of vascular smooth muscle cells is derived directly from contractile medial vascular smooth muscle cells or arises from a clonal expansion of a small subpopulation of vascular smooth muscle cells with inherently different properties from normal adult medial cells. It is also possible that adventitial myofibroblasts may also migrate into the intima to contribute to neo-intima formation. The resulting endothelialized neo-intima 'heals' the damage and may be sufficiently large to narrow the vessel lumen, as occurs in restenosis following therapeutic balloon

angioplasty.

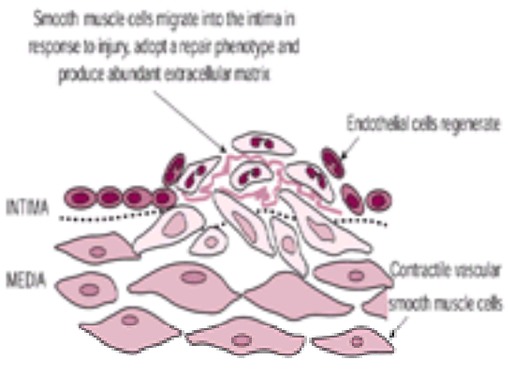

Fig. 2 The response of vascular smooth muscle cells to mechanical intimal damage.

Vascular smooth muscle cells in atherosclerosis

The recent recognition that atherosclerosis is an inflammatory process driven by the interaction between oxidized lipids and reactive inflammatory cells has brought about a complete re-evaluation of the role of vascular smooth muscle cells in its pathogenesis. Intimal vascular smooth muscle cells protect against plaque rupture and therefore the thrombotic consequences of atherosclerosis by migrating into the intima and changing phenotype to elaborate the matrix components of the all-important fibrous cap. Indeed, vascular smooth muscle cells are the only cells capable of making the fibrous cap, such that loss of vascular smooth muscle cells from the cap is one of the major determinants of plaque rupture. By reverting to a phenotype very similar to that of vascular smooth muscle cells in the neonatal blood vessel, they are adopting a beneficial 'repair' phenotype, in which the genes they express facilitate their reparative role. When reacting to either mechanical injury or the chemotactic influence of inflammatory cells they synthesize plasminogen activators and matrix metalloproteinases. These interact to digest the basement membrane of the vascular smooth muscle cells and thereby allow the vascular smooth muscle cells to migrate into the intima to form the fibrous cap by a combination of proliferation and matrix production, the latter being predominant (Fig. 3).

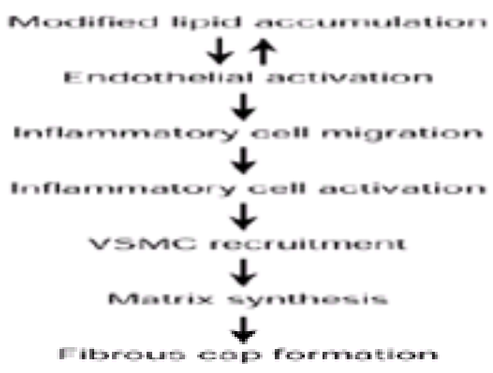

Fig. 3 Events leading to formation of the protective fibrous cap in atherosclerosis.

A paucity of vascular smooth muscle cells in the fibrous cap predisposes the lesion to plaque rupture. Vascular smooth muscle cells can be destroyed by inflammatory cells within the lesion, as discussed later in this chapter. Also, intimal vascular smooth muscle cells lose their capacity to regenerate and become highly susceptible to spontaneous 'suicide' by apoptosis (programmed cell death). Unless senescent vascular smooth muscle cells are replaced by those that are active, the capacity to repair and maintain the fibrous cap is lost, thereby tipping the balance in favour of plaque rupture. Indeed, senescence could be a feature of all vascular smooth muscle cells as they get older, possibly explaining why plaque rupture, leading to heart attacks and strokes, occurs increasingly frequently with age.

Atherosclerosis does, therefore, involve the response of vascular smooth muscle cells to injury, but it is not an initiating event, rather it is secondary to, and protective against, the destructive lipid-driven inflammatory process that leads to plaque instability, thrombosis, and patient death. However, vascular smooth muscle cells do contribute to some extent to plaque progression in as much as the formation of a new fibrous cap over a subclinical erosion or rupture necessarily increases the size of the lesion.

The role of vascular smooth muscle cells in restenosis

In approximately 30 to 40 per cent of patients undergoing successful balloon angioplasty for symptomatic coronary disease, symptoms will return because of restenosis at the site of the procedure. Initially this was thought simply to be due to excessive proliferation of intimal vascular smooth muscle cells in response to the injury. However, it has now become apparent that several factors in addition to the proliferation of vascular smooth muscle cells contribute to restenosis in man. Firstly, there is the early elastic recoil of the vessel wall that occurs soon after the balloon has been removed. Secondly, there is less vascular smooth muscle cell proliferation than anticipated from animal models and the bulk of the neo-intima comprises relatively few vascular smooth muscle cells scattered throughout an abundant extracellular matrix. Thus the production of vascular smooth muscle cell matrix is probably more important than the proliferation of vascular smooth muscle cells in determining neo-intimal bulk. Thirdly, eventual lumen diameter is determined by the capacity of the vessel to remodel. Angioplasty induces a vigorous adventitial response characterized by proliferation of adventitial myofibroblasts: these also express smooth muscle cell genes and synthesize a dense extracellular matrix that splints the vessel and prevents outward remodelling. As mentioned above, these cells may also migrate through the damaged media and contribute to the formation of neo-intima. Most of these consequences of balloon angioplasty can be abrogated by the deployment of an intravascular stent that prevents elastic recoil and negative remodelling. Although proliferation of intimal vascular smooth muscle cells and matrix synthesis still occur in stented vessels, their impact on the final diameter of the lumen is offset by the initial gain in diameter achieved by balloon dilatation and maintained by the stent (Fig. 4).

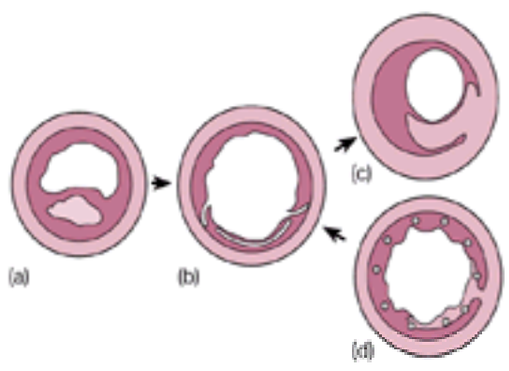

Fig. 4 Restenosis after balloon angioplasty. (a) Stenotic atherosclerotic lesion before angioplasty. (b) Lesion immediately after angioplasty showing expansion of the whole artery with fissuring through the media into the adventitia. (c) Restenosis occurs because of formation of neo-intima in response to balloon injury. The neo-intima arises in part through medial smooth muscle cell proliferation and matrix synthesis and invasion of adventital myofibroblasts. Note also that the adventitia has thickened and in effect 'splints' the vessel, preventing outward remodelling to compensate for the formation of neo-intima. The net result is lumen narrowing and restenosis. (d) Stent deployment after balloon dilatation prevents elastic recoil and negative remodelling induced by the adventitial response. Although there is still

neo-intima formation it is insufficient to induce significant lumen narrowing.

Summary

The vascular smooth muscle cell is highly adaptable and can serve multiple roles. In the normal vessel wall it acts principally as a contractile cell. However, in response to any form of vascular insult it rapidly assumes a wound-healing role and expresses the repertoire of genes necessary to repair the vessel wall. In the context of the chronic inflammatory stimulus of atherosclerosis, this response creates the protective fibrous cap responsible for conferring stability to the plaque and preventing plaque rupture. Following balloon angioplasty the reactive vascular smooth muscle cells act primarily to 'heal' the damage caused by the balloon. In so doing, they create a neo-intima that stabilizes the lesion but which may also cause restenosis if adequate compensatory remodelling does not occur.

Further reading

Bennett MR *et al*. (1997). Increased sensitivity of human vascular smooth muscle cells from atherosclerotic plaque to p53-mediated apoptosis. *Circulation Research* **81**, 591–9.

Bennett MR *et al*. (1998). Co-operative interactions between RB and p53 regulate cell proliferation, cell senescence and apoptosis in human vascular smooth muscle cells from atherosclerotic plaques *Circulation Research* **82**, 704–12.

Campbell GR *et al*. (1988). Arterial smooth muscle. A multifunctional mesenchymal cell. *Archives of Pathololology and Laboratory Medicine* **112**, 977–86.

Clowes AW, Reidy MA, Clowes MM (1983). Mechanisms of stenosis after arterial injury. *Laboratory Investigation* **49**, 208–15.

Glagov S *et al*. (1987). Compensatory enlargement of human atherosclerotic coronary arteries. *New England Journal of Medicine* **316**, 371–5.

Shanahan C, Weissberg P (1998). Smooth muscle cell heterogeneity—patterns of gene expression in vascular smooth muscle cells *in vitro* and *in vivo*. *Arteriosclerosis Thrombosis and Vascular Biology* **18**, 333–8.

Weissberg P, Clesham G, Bennett M (1996). Is vascular smooth muscle cell proliferation beneficial? *The Lancet* **347**, 305–7.

15.1.2.1 The pathogenesis of atherosclerosis

R. P. Naoumova and J. Scott

Introduction

Atherosclerosis, the underlying cause of heart attacks, strokes, and peripheral vascular disease, is one of the major killers in the world. The disease develops slowly over many years in the innermost layer of large and medium-sized arteries. It does not usually manifest before the fourth or fifth decade of life, but may strike with devastating suddenness. At least 30 per cent of individuals die from their first heart attack. In England and Wales coronary heart disease accounts for 31 per cent of all deaths in men and 23 per cent of all deaths in women, and morbidity from the disease is significant, with a profound impact on health-care services and on the industrial economy.

Epidemiology

The demonstration of coronary heart disease in an individual is taken as a reliable index of the presence of more general atherosclerosis. The highest death rates from coronary heart disease are found in Britain, northern Europe, the United States, Australia, and New Zealand. Deaths from coronary disease rose dramatically after the First World War, peaked in the late 1960s in the United States, and have since declined. In western Europe this peak and decline lagged behind the United States by some 10 years. Changes in diet, exercise, smoking, and affluence account for much of this decline. Better medical and surgical interventions have also been important. By contrast, the countries of eastern Europe and the former Soviet Union are showing a marked increase in the prevalence of coronary heart disease. This can be attributed to the influence of the risk factors that operate in the industrialized West (Table 1). Substantially lower death rates are found in southern Europe, Latin America, Japan, and China. For further discussion see Chapter 15.4.1.2.

The lesions of atherosclerosis

Autopsy studies show that in humans atherosclerosis begins in early life and develops slowly over many years (Fig. 1) before becoming symptomatic. Atherosclerosis is a focal intimal disease of arteries ranging in size from the aorta down to those of approximately 3 mm external diameter. The arteries most commonly involved with atherosclerosis include the aorta, coronaries, carotid, cerebral, and femoral. Branch points and curvatures, the sites of blood turbulence, favour the development of atherosclerotic lesions.

The earliest lesions of atherosclerosis are fatty streaks. These consist of an accumulation of lipid-engorged macrophages (foam cells), and T lymphocytes in the arterial intima. The fatty streaks progress to intermediate lesions (or transitional plaque), composed mainly of macrophage foam cells and smooth muscle cells which migrate into the intima from the media. With time these develop into raised fibrous (advanced) plaques, characterized by a dense fibrous cap of connective tissue and smooth muscle cells overlying a core containing necrotic material and lipid, mainly cholesteryl esters, which may form cholesterol crystals on histological section. The necrotic core is a result of apoptosis and necrosis, increased proteolytic activity and lipid accumulation. Fibrous plaques also contain a large number of macrophage foam cells, T cells, and smooth muscle cells. This collection of cells, surrounding the necrotic core, promotes plaque growth. The plaque undergoes vascularization and microvessels develop in connection with the artery's vasa vasorum. The new vessels provide a channel for the access of inflammatory cells and may also lead to intraplaque haemorrhage and thus weaken the plaque. Advanced atherosclerotic plaques frequently accumulate calcium, due to the presence of proteins specialized in binding calcium (osteocalcin, osteopontin, bone morphogenic proteins).

The advanced plaque is the substrate from which the complicated plaque develops, leading almost inevitably to clinical symptoms. The complicated plaque has a thin cap, especially at the shoulders or margins of the lesion, and may contain ulcerations, fissures, erosions, or cracks. These provide sites of platelet adherence, aggregation, and thrombosis. The thin fibrous cap may break or tear leading to haemorrhage into the necrotic core and thrombosis.

Arterial remodelling and clinical syndromes associated with atherosclerosis

Arterial remodelling is a clinically important feature in the evolution of the atherosclerotic lesion. It delays the development of significant luminal narrowing and is a compensatory process in human atherogenesis. During early phases of plaque formation the lesion grows away from the lumen, so affected vessels increase in diameter (compensatory enlargement) and the plaque will not cause flow-limiting stenosis. Most plaques of this type will not be visible angiographically.

When the plaque covers more than 40 per cent of the elastic lamina, the artery cannot compensate by dilatation and the lesion begins to intrude into the arterial lumen, becoming angiographically detectable. It may impede the blood flow to an organ, giving rise to ischaemia, the symptoms of stable angina, and intermittent claudication. If the atherosclerotic lesion undergoes superficial erosion of the endothelium with limited thrombosis, this may result in unstable angina or myocardial infarction, even when the lesion is not flow limiting. Deep fissuring or frank rupture of the plaque with complete sudden occlusion of coronary arteries may cause myocardial infarction or sudden death. In the cerebral circulation, the same process causes transient ischaemic attacks and completed stroke. In arteries weakened by the ageing process and complicated by atherosclerosis, aneurysmal dilatation and rupture may occur.

Plaque stability and plaque rupture

Rupture of atherosclerotic lesions can trigger the thrombosis that precipitates clinical events. However, plaques have different propensities to rupture. Plaques with dense extracellular matrix, relatively thick fibrous caps, and limited lipid cores are generally unlikely to initiate thrombosis followed by an acute vascular event.

Culprit lesions, causing myocardial infarction or unstable angina, characteristically have thin fibrous caps, large cores of extracellular lipids, and an abundance of macrophages and T lymphocytes at the site of plaque rupture. Plaques usually contain limited numbers of smooth muscle cells, leading to decreased synthesis of extracellular matrix and weakening of the plaque's fibrous cap. The integrity of the fibrous cap is also attacked by cytokines derived from activated macrophages, which can promote the expression of proteinases that can degrade the extracellular matrix. Plaque rupture usually occurs at the 'shoulders' of the plaque.

Angiography cannot accurately predict the stability of a lesion.

The pathogenesis of atherosclerotic lesions

Atherosclerosis develops as a healing response of the intima to repeated vascular wall injury, where risk factors (Table 1) operate by promoting chronic cycles of damage and repair. In the broadest terms, atherosclerosis is now recognized to be a chronic inflammatory process.

Endothelial dysfunction is the first step in the development of atherosclerosis. Modified low-density lipoprotein, toxins in tobacco smoke, the shear stress of hypertension, elevated plasma homocysteine concentrations, diabetes mellitus, and infectious micro-organisms can all cause dysfunction of endothelial cells. The dysfunctional endothelium undergoes a protective response that alters normal homeostatic properties due to expression of adhesion molecules, growth-promoting substances, and activation of the blood coagulation cascade. Monocytes and T lymphocytes adhere to the activated endothelium, become activated, and produce growth factors, cytokines, and chemoattractants. Adherent white blood cells migrate into the arterial intima and smooth muscle cells are recruited from the media into the intima. With repeated rounds of injury and repair, palisades of smooth muscle cells, matrix proteins, lipid-laden macrophages, and T lymphocytes accumulate to form atherosclerotic plaques (Fig. 1).

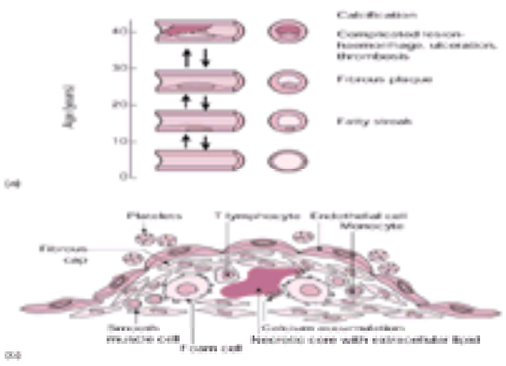

Fig. 1 (a) Natural history of atherosclerosis. (Reprinted with permission from McGill HC Jr, *et al.* In: Sandler S and Bourne N, eds. *Atherosclerosis and its origin*, Academic Press, New York, 1963). (b) Schematic diagram of an advanced atherosclerotic lesion. As fatty streaks progress to advanced lesions, they form a fibrous cap due to migration and proliferation of smooth muscle cells. This represents a type of healing response to injury. The cap covers a mixture of foam cells, T lymphocytes, lipids, and debris, forming a necrotic core—the result of apoptosis and necrosis.

Low-density lipoprotein has a central role in the pathogenesis of atherosclerosis. It may enter the intima through the damaged endothelium or, more commonly, by transcytosis across the intact endothelium, becoming 'trapped' in the subendothelial space. There, low-density lipoprotein undergoes low-grade modification by oxidative free radicals, secreted by cells of the artery wall, forming minimally modified low-density lipoprotein (Fig. 2). Although still recognized by the low-density lipoprotein receptor, minimally modified low-density lipoprotein can stimulate the release of macrophage colony-stimulating factor and monocyte chemoattractant protein 1 from endothelial cells: these facilitate monocyte recruitment and their differentiation into tissue macrophages. Minimally modified low-density lipoprotein adheres to matrix proteins of the arterial wall, where it undergoes more extensive oxidation. Free radicals are produced from macrophages and from nitric oxide derived from endothelial cells. This is compounded by the products of tobacco smoke and by homocysteine. Highly oxidized/modified low-density lipoprotein is characterized by changes not only of the lipid but also of the protein portion of low-density lipoprotein, leading to loss of recognition by the low-density lipoprotein receptor. Thus oxidized low-density lipoprotein becomes the major ligand for the scavenger receptor family (scavenger receptor A, scavenger receptor B, and others), expressed in the macrophages accumulating at the site of the injury to the vessel wall. This shift in receptor recognition leads to uptake of oxidized low-density lipoprotein by receptors, not regulated by the cholesterol content of the cell. The result is massive accumulation of cholesteryl esters in the macrophages, giving the cytoplasm its characteristic foamy appearance and transforming the macrophage into a foam cell.

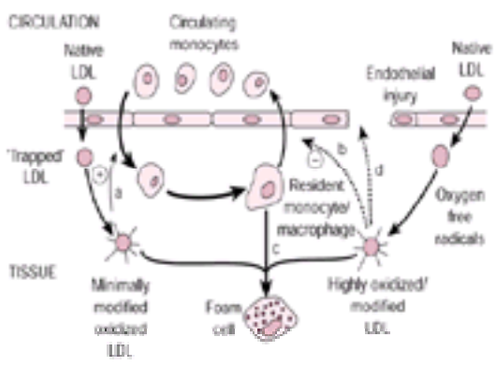

Fig. 2 Oxidation of low-density lipoprotein. The figure shows the mechanisms by which oxidized low-density lipoprotein contributes to atherosclerosis. (a) Oxidized low-density lipoprotein is chemotactic for circulating monocytes. (b) Oxidized low-density lipoprotein inhibits the movement of resident macrophages out of the arterial intima. (c) Resident macrophages generate free radicals and contribute to production of oxidized low-density lipoprotein, leading to the generation of foam cells. (d) Oxidized low-density lipoprotein is cytotoxic and this leads to endothelial cell damage and loss of integrity. (Reproduced from Quinn MT *et al*. (1987). Oxidatively modified low density lipoproteins: a potential role in recruitment and retention of monocyte/macrophages during atherogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 2995–8, with permission.)

Hypertension accelerates atherogenesis by activating genes in response to increased sheer stress, the products of which perturb vascular tone and promote the accumulation of smooth muscle cells. Hypertension also increases the formation of hydrogen peroxide and free radicals that worsen oxidative damage, reduces the formation of nitric oxide by the endothelium, and increases leucocyte adhesion. In diabetes mellitus hyperglycaemia may promote non-enzymatic glycation of low-density lipoprotein, which may initiate atherosclerosis in the same way as oxidatively modified low-density lipoprotein. A high plasma homocysteine concentration is toxic to endothelium, decreases the availability of nitric oxide, and has prothrombotic activity.

The cells of the atherosclerotic plaque

Endothelial cells

In the earliest stages of atherogenesis, damaged endothelial cells become dysfunctional. Sloughing of the endothelium occurs at a later stage, when plaques become complicated and split or fissure. Dysfunctional endothelial cells produce growth factors, cytokines, chemoattractants, clotting factors, and adhesion molecules. The result is recruitment and transformation of monocytes into macrophages, and recruitment and proliferation of smooth muscle cells and T cells. Thrombotic processes are activated. There is chronic alteration of vascular tone as a result of disordered nitric oxide production and signalling.

Monocyte/macrophages

The lipid-laden macrophage is the hallmark of atherosclerosis and is instrumental in its development. Monocyte conversion from a quiescent cell to a phagocytically active macrophage is associated with expression of scavenger receptors and oxidized low-density lipoprotein receptors which avidly take up highly oxidized low-density lipoprotein, the latter no longer being recognized by the low-density lipoprotein receptor. The cholesteryl ester released from low-density lipoprotein is

broken down in lysosomes and re-esterified in the cytoplasm.

Activated macrophages secrete a wide variety of growth-modulating substances and chemoattractants. Phagocytic macrophages produce free radicals and are induced to produce nitric oxide, which generates free radicals, promoting further oxidative damage to low-density lipoproteins (Fig. 2). Macrophages also secrete proteolytic enzymes (collagenase, elastase, stromelysin, and gelatinases): these contribute to the necrosis and liquefaction of the core of advanced fatty plaques and also render the plaque prone to rupture by thinning the fibrous cap.

Vascular smooth muscle cells

Smooth muscle cells in the walls of normal arteries mainly contain contractile proteins, such as actin and myosin, and are said to display a contractile phenotype. They respond to vasoregulatory substances—catecholamines, angiotensin II, prostaglandins, leukotrienes, endothelin, nitric oxide, and other regulatory compounds. However, under the influence of proinflammatory cytokines and growth factors, smooth muscle cells in the atherosclerotic plaque switch from a contractile to a secretory phenotype and produce extracellular matrix. In the media of normal arteries, the matrix consists of types I and III collagen, whereas in the atherosclerotic lesion it comprises largely proteoglycans, intermixed with loosely scattered collagen fibrils. The local release of growth factors, cytokines, and chemoattractants leads to autocrine and paracrine effects on growth and cell recruitment. Smooth muscle cells also express scavenger receptors and they, too, become lipid-loaded.

T lymphocytes

T cells (both CD4 and CD8) are present in the atherosclerotic lesion in all stages of the process. These are activated when they bind antigen processed and presented by macrophages, resulting in the secretion of proinflammatory cytokines, including interleukin 1, interferon-g, and tumour necrosis factor a and b, which amplify the inflammatory response and compound the atherosclerotic process by attracting further macrophages and T lymphocytes and perpetuating endothelial cell activation. T cells in the plaque become sensitized to new antigens in the lesion, such as modified low-density lipoprotein.

Platelets

Platelets undergo activation in response to agonists such as thrombin, ADP, adrenaline, and platelet activating factor. When activated, platelets release their granules, containing cytokines and growth factors. The activation is also triggered when peripheral blood is exposed to thrombogenic agents at the site of blood vessel damage. Here agonists such as collagen present in the extracellular matrix, exposed in the subendothelium along with von Willebrand factor and fibrinogen produced at the wound site, initiate the cascade of events that leads to platelet aggregation and the formation of a platelet plug. In this process platelet-specific integrins act as receptor tyrosine kinases (glycoprotein, GP IIb/IIIa), initiating the intracellular changes that mediate platelet activation and aggregation, and later the binding to fibrin and clot retraction.

Molecular and cell interactions

The formation of the atherosclerotic plaque is brought about by a complex series of cellular and molecular interactions. Substances expressed at the cell surface and secreted in response to cellular activation bring about these events. Intracellular co-ordinating mechanisms, such as the NF_B system, operate at the site of the lesion. NF_B is a ubiquitous transcription factor that can be activated by diverse proatherogenic stimuli and provides a potential common link to co-ordinate the expression of series of genes involved in atherogenesis.

Growth factors

Molecules controlling the proliferation of smooth muscle cells include platelet-derived growth factor, fibroblast growth factor, heparin-binding epidermal growth factor, insulin-like growth factor I, interleukin 1, tumour necrosis factor-a, transforming growth factors a and b, thrombin, and angiotensin II.

Platelets contain platelet-derived growth factor and other growth regulatory substances, such as epidermal growth factor, transforming growth factors a and b, insulin-like growth factor I, and thromboxane, which are released during platelet aggregation and activation. Platelet-derived growth factor is also produced from activated endothelial cells and secretory smooth muscle cells in response to the macrophage cytokines interleukin 1, tumour necrosis factor-a, and transforming growth factor-b. Fibroblast growth factor, the other potent mitogen for vascular smooth muscle cells, also has mitogenic activity for endothelial cells. Heparin-binding epidermal growth factor is a potent growth factor and chemoattractant for smooth muscle cells only. Transforming growth factor-b is a multifunctional cytokine, expressed with its receptor system on smooth muscle cells. It is a potent inhibitor of mitosis of smooth muscle cells and also stimulates elaboration of matrix proteins such as fibronectin and vascular collagen.

Vascular endothelial cell growth factor is a potent and specific mitogen for endothelial cells and also promotes permeability of small veins and venules. It is produced by endothelial cells and macrophages in response to ischaemia and acts as a potent chemoattractant for macrophages, also inducing endothelial cells to produce collagenase as well as urokinase-type plasminogen activator, tissue plasminogen activator, and their inhibitor plasminogen activator inhibitor 1.

Cytokines

Cytokines are multipotent mediators of inflammation and immunity with generalized action in host defence and pathology. They can affect key functions of vascular wall cells and may participate as autocrine and paracrine mediators in atherogenesis.

The cytokines interleukin 1, interleukin 2, tumour necrosis factor-a, g-interferon, and granulocyte–macrophage and macrophage colony stimulating factors are secreted from macrophages, T lymphocytes, activated endothelial cells, and secretory smooth muscle cells. Interleukin 1, tumour necrosis factor-a, and transforming growth factor-b induce endothelial cell activation, with the release of mitogens and chemoattractants for smooth muscle cells, and activate the coagulation cascade. Colony stimulating factors attract further inflammatory cells. Cytokines can also inhibit the proliferation of smooth muscle cells, and thus may either promote or retard atherogenesis.

Chemokines (chemoattractants)

Chemokines are members of a superfamily of small polypeptides that share the ability to induce migration, growth, and activation of subsets of leucocytes and other cells present in atherosclerotic lesions. They are also involved in regulating angiogenesis at inflammatory sites of the atherosclerotic plaque. More than 30 human chemokines have been identified. Interleukin 8 acts predominantly on neutrophils; monocyte chemoattractic protein 1 acts on lymphocytes, monocytes, mast cells, and eosinophils; whereas lymphotactin acts solely on lymphocytes. Chemokines mediate their actions via specific cell surface receptors, members of the seven transmembrane-spanning G-protein-linked molecules.

Adhesion molecules

Dysfunctional endothelial cells undergo activation, with the production of cell surface proteins that mediate the adherence of inflammatory cells. This inductive process is mediated by cytokines. The adhesion molecules induced include E-selectin, which is a membrane glycoprotein specific to endothelial cells that mediates the adhesion of neutrophils. It is a member of the selectin gene family, which also includes L-selectin and P-selectin. These molecules are implicated in the initial 'rolling step' of leucocyte extravasation.

Vascular cell adhesion molecule 1 is induced on endothelial cells by interleukin 1, tumour necrosis factor-a, lipopolysaccharide, and oxidized low-density lipoprotein. Vascular cell adhesion molecule 1 binds cells expressing integrins a4b1 (VLA4), such as monocytes and lymphocytes, but not neutrophils. Another adhesion molecule is intercellular adhesion molecule 1, a receptor for integrins, which binds all leucocytes. Thus vascular cell adhesion molecule 1 and intercellular adhesion molecule 1 serve to anchor activated leucocytes after the initial 'rolling step'.

Platelet endothelial cell adhesion molecule exists at the tight junctions of endothelial cells and is required for the transmigration of neutrophils and monocytes and for platelet adhesion.

Matrix proteins

Smooth muscle cells that have taken on a secretory phenotype are the primary source of extracellular matrix. Matrix proteins comprise collagens, elastin, proteoglycans, and microfibrillar protein. Their production is controlled by interleukin 1, tumour necrosis factor-a, and transforming growth factor-b, which mediate the switch between proliferative and secretory smooth muscle cell phenotypes, whereas g-interferon suppresses collagen expression.

In the normal artery both synthesis and degradation of matrix are very slow, whereas atherosclerosis and injury lead to increased synthesis of many matrix proteins. The degree of ongoing matrix degradation is a highly controlled and essential component of the homeostasis of the normal artery. Increased matrix degradation, due to high activity of proteases in the plaque, is common in unstable atherosclerotic lesions. Ageing is associated with a reduction of elastin in the extracellular matrix, leading to hardening of the arteries.

Cellular death

Advanced fibrous caps that have ruptured have twice as many macrophages, but only half as many smooth muscle cells, as unruptured fibrous caps. The relative decrease in the number of smooth muscle cells may result from growth inhibition due to g-interferon or cellular death from lytic injury and necrosis. Recent studies show that smooth muscle cells of the atheromatous plaque undergo programmed cell death, apoptosis.

Coagulation factors

The activation of endothelial cells initiates cell-surface assembly of the prothrombinase complex and subsequent deposition of fibrin and platelet activation. The process is initiated by plasma membrane expression of tissue factor, which activates factor VIIa and, in turn, factors IX and X. Thrombin is generated in the presence of endothelial cell factor V. Thrombin contributes to the inflammatory response by induction of the adhesion molecule P-selectin and platelet activating factor. Platelet arachidonic acid is released by activity of phospholipase C and phospholipase A2. The enzyme cyclo-oxygenase generates platelet endoperoxides, and the enzyme thromboxane synthase generates thromboxane, which in turn increases phospholipase C activity, stimulating platelet activation and degranulation. Together these substances promote neutrophil and platelet adhesion. Thrombin also induces plasminogen activator inhibitor 1, and increases tissue factor synthesis and expression of platelet-derived growth factor. E-selectin is induced and serves as a site for leucocyte attachment. Thus the control of coagulation by cytokines closely mimics that of the inflammatory response, indicating the interdependence of the two processes.

Restenosis

Surgical treatment of arteries narrowed by atherosclerosis is by arterial or venous bypass grafting or endarterectomy. Medical treatment is by percutaneous transluminal balloon angioplasty. Immediate complications of this procedure are thrombosis and arterial wall dissection. There is also a high failure rate due to restenosis of the arterial lumen (30–40 per cent within 3–6 months).

Restenosis is a complex reparative process involving the following sequence of events after angioplasty: recoil, remodelling, mural thrombus formation with subsequent organization by connective tissue, followed by smooth muscle cell activation, migration, proliferation, and increased synthesis of extracellular matrix. Growth factors originating from the thrombus, vessel wall, and circulating cells contribute to these events. After 2 to 6 months the stenotic region becomes organized and consists of a maturing scar, but little thrombus or lipid. The NF_B pathway plays a central role in triggering the transcription of genes encoding leucocyte adhesion molecules, chemokines, and enzymes that can influence extracellular martix metabolism, leading to restenosis. Satisfactory regimens for the prevention of restenosis have yet to be established.

Transplant atherosclerosis

Cardiovascular disease is emerging as the major cause of late morbidity and mortality in transplant patients, accounting for 45 per cent of deaths in renal transplant recipients and being the major factor limiting survival of cardiac allografts.

Recipients of organ transplant often have multiple classical risk factors before transplantation and contributing to accelerated atherosclerosis afterwards (Table 1). Immunosuppressive agents such as prednisolone and cyclosporine have adverse effects on lipid metabolism, whereas tacrolimus (FK506) affects lipid metabolism to a lesser extent. Cytomegalovirus infection and abnormal platelet aggregation also contribute to accelerated atherogenesis.

Specific immunological factors contribute to the development and progression of transplant atherosclerosis, especially in the accelerated form of coronary arteriopathy that plagues heart transplant recipients.

Future perspectives

Despite changes in lifestyle and the use of potent lipid-lowering agents, cardiovascular disease continues to be the major cause of death in western Europe and North America. Furthermore, end-point clinical trials using statins show at best a 30 per cent decrease in total mortality, with a 42 per cent decrease in coronary deaths. Clearly new therapeutic targets need to be pursued.

Serum levels of high-density lipoprotein cholesterol are inversely related to coronary heart disease. Reduced levels of high-density lipoprotein are found in half of the patients with coronary heart disease. The discovery of the pivotal role of the *ABC1* transporter gene, encoding the cholesterol-efflux regulatory protein, in the generation of high-density lipoprotein provides opportunities for new drug targets.

Better understanding of the pathogenesis of the initiation, progression, and complications of atherosclerotic lesions will provide new potential therapeutic approaches, different from plasma lipids. An understanding of the central role of oxidized low-density lipoproteins and of the macrophage in the pathogenesis of atherosclerotic lesions points to a new direction for prevention and treatment—antioxidants.

Since atherosclerosis is a multigenic disease, understanding the patterns of gene expression will shed light on differences in susceptibility to agents causing disease, on genetic variability in prediction of risk and response to therapy, and may provide clues for designing new therapeutic approaches.

Further reading

Davies MJ, Woolf N (1993). Atherosclerosis—what is it and why does it occur? *British Heart Journal* **69** (Suppl.), S3–S11.

Kiechl S, Willeit J for the Bruneck Study Group (1999). The natural course of atherosclerosis. Part II: vascular remodelling. *Arteriosclerosis, Thrombosis and Vascular Biology* **19**, 1491–8.

Krieger M (1997). The other side of scavenger receptors: pattern recognition for host defence. *Current Opinion in Lipidology* **8**, 275–80.

Libby P *et al*. (1998). Current concepts in cardiovascular pathology: the role of LDL cholesterol in plaque rupture and stabilization. *American Journal of Medicine* **104** (Suppl. 2A), 14S–18S.

McGill HC Jr, Geer JC, Strong JP (1963). The natural history of human atherosclerotic lesions. In: Sandler M, Bourne G, eds. *Atherosclerosis and its origins*, pp 396–405. Academic Press, New York.

Quinn MT *et al*. (1987). Oxidatively modified low density lipoproteins: A potential role in recruitment and retention of monocyte/macrophages during atherogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 2995–8.

Ross R (1993). The pathogenesis of atherosclerosis: a perspective for the 1990s. *Nature* **362**, 801–9.

Ross R. (1999). Atherosclerosis—an inflammatory disease. *New England Journal of Medicine* **340**, 115–25.

Scott J (1999). Good cholesterol news. *Nature* **400**, 816–19.

Seinberg D (1997). Oxidative modification of LDL and Atherogenesis. Lewis A. Conner Memorial Lecture. *Circulation* **95**, 1062–71.

15.1.2.2 The haemostatic system in arterial disease

T. W. Meade, P. K. MacCallum, and G. J. Miller

Introduction

General recognition of the thrombotic component in arterial disease, particularly coronary heart disease, is comparatively recent. The term 'coronary thrombosis' appears to have first been used by Herrick very early in the 1900s and it continued to be used until coronary heart disease became the preferred terminology after the Second World War, when the condition had reached epidemic proportions.

Epidemiological studies which started in the late 1940s—the work in the community of Framingham being the best known—began to establish the characteristics of those at particular risk of heart attacks. The main emphasis, however, was on the part played by lipid infiltration, which has tended to dominate thinking in North America ever since in comparison with a readier acceptance in Europe of a thrombotic component (as well as of atherogenesis). There was good reason for supposing that lipids play a major part: it was easy to demonstrate lipid-rich material, including cholesterol crystals, in the coronary arteries and it seemed logical to suggest that high-fat diets and blood cholesterol levels might contribute to atheroma. In the 1950s, however, J. N. Morris and his colleagues at the (now Royal) London Hospital showed very clearly that while advanced atheroma obviously contributed, it could not explain the whole of the coronary heart disease epidemic, although the implication that there must be another process involved—almost certainly thrombosis—was not fully recognized for another 20 years.

Interest in a thrombotic component to coronary heart disease started to re-emerge in the 1970s but was initially characterized by a rather sterile debate as to whether thrombosis causes or is a consequence of myocardial infarction. Evidence for the role of thrombosis in myocardial infarction was provided in convincing form when angiographic monitoring of the early use of thrombolytic therapy showed the development of occlusive thrombi preceding full manifestation of the clinical event. As for sudden coronary death, one reason for doubting the involvement of thrombosis had been the failure to demonstrate thrombi at autopsy in many cases. Apart from limitations in methods for detecting thrombi, the very striking increase in fibrinolytic activity associated with the agonal process of dying from a heart attack is likely to result in the dissolution of some thrombi that were nevertheless responsible for the event. In 1984, a particularly careful study comparing the prevalence of thrombosis in sudden death from coronary heart disease with sudden death from other causes gave the results summarized in Table 1, indicating that a degree of thrombosis is demonstrable in nearly all sudden coronary deaths. Other studies have generally confirmed this. It is now also recognized that the pathology of unstable angina pectoris is similar to that of myocardial infarction and of sudden coronary death in consisting of a significant thrombotic component and in responding to antithrombotic treatment (see Chapter 15.4.2.1).

Further evidence for the role of thrombosis in coronary heart disease came with recognition of the effects of aspirin in modifying the aggregation of platelets and with the results of observational studies and early trials showing the reduction in coronary heart disease attributable to aspirin. Morphological observations and striking cine film pictures of platelet aggregation at sites of vascular injury have put the role of platelets in the thrombotic process beyond any doubt for many years now. However, no tests of platelet behaviour have convincingly been shown to be associated with the subsequent risk of first events of coronary heart disease, although spontaneous platelet aggregation and increased platelet volume may help predict those at risk of recurrent episodes.

Until fairly recently, the contribution of the coagulation system to arterial thrombosis through fibrin formation was not considered to be of clinical significance. This was partly because platelet aggregation in response to vascular injury is very rapid, and hence possibly more relevant than the allegedly slower-acting coagulation system, and also because of the value of aspirin in reducing coronary heart disease. However, epidemiological studies of the coagulation system in thrombosis and coronary heart disease have now demonstrated its involvement and implications for the management and prevention of coronary heart disease. This section summarizes, first, the mainly epidemiological evidence regarding coagulability and coronary heart disease (also arterial disease at other sites), and secondly, implications for long-term management and prevention (the treatment of acute events being described in Chapter 15.4.2.3).

Epidemiological evidence

Population-based studies started from the general proposition that high levels of procoagulatory clotting factors and low levels of anticoagulatory factors would predispose to coronary heart disease. Sceptics argued that (other than in obvious deficiency conditions such as haemophilia) clotting factors circulate well in excess of concentrations required for haemostasis under normal conditions and that no associations with coronary heart disease would therefore be demonstrable. However, requirements for haemostasis may not be a reliable guide to the influence of different levels of clotting factors on thrombosis, where a high level of a procoagulatory factor might facilitate thrombosis and a major coronary heart disease event. By analogy with blood pressure, a certain level of pressure is necessary to maintain normal circulatory function while raised levels predispose to the pathological processes involved in coronary heart disease and stroke.

Studies that demonstrate associations between different characteristics and the risk of coronary heart disease have two main purposes. One is to contribute to our understanding of the pathogenesis of the condition. The other is to identify characteristics with reasonably clear implications either for screening purposes and/or for treatment and prevention.

Figure 1 shows the main features of the coagulation system. Here and in the text the letter 'a' signifies the activated form of the clotting factor. The coagulation process, resulting in the generation of thrombin, can be initiated either through the extrinsic system, so called because it depends on the availability of tissue factor which has not generally been considered a component of the circulating blood, or through the contact system which is not dependent on biochemical properties outside the circulating blood. Tissue factor becomes available when atheromatous lesions leak or rupture, and other evidence strengthens the conclusion that the extrinsic system predominates in coronary heart disease, though not to the exclusion of an influence of factors XII and VIIIa on the intrinsic system (see below). As well as its well-known function in converting soluble fibrinogen into insoluble fibrin, thrombin has numerous other properties, of which the principal ones are shown in Table 2. It is a potent platelet-aggregating agent and may exert this action at least as soon as (if not before) its action on fibrinogen, so that the coagulation system has a strong influence on platelet behaviour as well as on the deposition of fibrin. Other actions of thrombin include the activation of protein C, which together with protein S inhibits the coagulation process; it also has both activating and inhibiting effects on the fibrinolytic system (see Fig. 1 and Table 2).

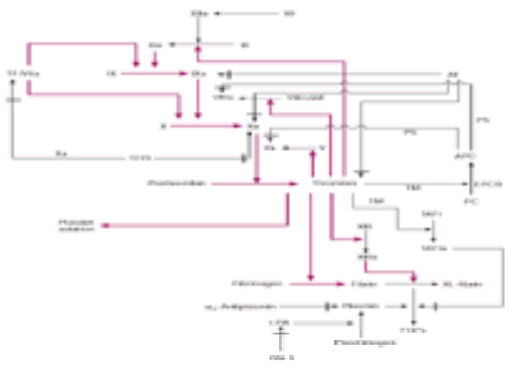

Fig. 1 Outline of the coagulation system and its regulation. The bold arrows represent the main pathways of generation of thrombin and thrombin's key role in fibrin formation and platelet activation: – reflects inhibitory activities; TF = tissue factor; TFPI = tissue factor pathway inhibitor; AT = antithrombin; PC = protein C; APC = activated protein C; PS = protein S; TAFI = thrombin activatable fibrinolysis inhibitor; XL-fibrin = cross-linked fibrin; FDPs = fibrin degradation products; t-PA = tissue-type plasminogen activator; PAI-1 = plasminogen activator inhibitor type 1; vWF = von Willebrand factor; EPCr = endothelial protein C receptor.

There have now been numerous cohort (prospective) and case-control or case-comparison studies of associations between different components of the coagulation system and the risk of coronary heart disease.

Fibrinogen

A recent overview of 18 cohort studies based on 4018 cases of coronary heart disease demonstrated that those in the top third of the fibrinogen distribution are at 1.8 times the risk of coronary heart disease compared with those in the bottom third, the findings being similar in the 12 studies concerned with first events and the other six with individuals known to have had previous episodes and followed for recurrences. The association of fibrinogen with the incidence of coronary heart disease is independent of other risk factors, is of a similar magnitude to the risk due to raised cholesterol, and is probably the same in women as in men. Besides first and recurrent events of coronary heart disease, high fibrinogen levels are also associated with the onset, recurrence, and progression of cerebrovascular and lower extremity arterial disease, with the incidence of graft occlusion following bypass surgery, and possibly with an increased risk of restenosis following coronary or lower limb angioplasty.

Factor VII

The possible involvement of factor VII activity in coronary heart disease is of theoretical as well as practical interest: when tissue factor is exposed it binds with factor VII, such that high levels of the latter might affect the amount of thrombin produced. However, the evidence is equivocal. Two cohort studies suggest that high levels increase risk, whilst others do not. Case-control and cross-sectional studies have also given conflicting results. Several assay techniques have been used and it has been established that these may vary in their sensitivity to factor VII activity, which might partly account for the differing results.

Factor VIII and von Willebrand factor

Factor VIII circulates in a complex with von Willebrand factor. The two proteins serve different haemostatic functions and have different sites of production, but they are closely correlated in a statistical sense so that independent contributions of the two proteins cannot easily be demonstrated, if at all. Four cohort studies have shown high levels of factor VIII to confer an increased risk of coronary heart disease. Haemophiliac patients appear to have a lower than expected incidence of coronary heart disease (though a considerable excess of cerebrovascular disease because of bleeding) and carriers of haemophilia have a reduced standardized mortality rate from coronary heart disease. Autopsy data show that haemophilia does not prevent the development of atheroma, suggesting that the decreased risk of coronary heart disease in patients with haemophilia is due to the effect of low levels of factor VIII on fibrin formation and thrombogenic potential. Case-control data show that elevated levels of factor VIII are associated with an increased relative risk of venous thrombosis, which also supports a prothrombotic role for factor VIII since vessel wall disease is not a consideration on the venous side. The association of factor VIII with coronary heart disease therefore appears to be due to a direct contribution of the level of factor VIII in circulating blood and not to a chronic phase response to atheromatous vessel wall changes.

Fibrinolytic activity

Several cohort studies, whether concerned with first events or in patients with previous episodes of coronary heart disease, have led to the conclusion that impaired fibrinolytic activity is an independent risk factor for coronary heart disease or its recurrence. Different studies have used different methods. These have included global tests of fibrinolytic activity such as the dilute clot lysis time, which takes account of both activators and inhibitors, and assays of specific components of the fibrinolytic system, principally plasminogen activator inhibitor type 1 and of D-dimer, the main degradation product of fibrinolysis. The principal determinants of the dilute clot lysis time are plasminogen activator inhibitor type 1 and, in men, activity of tissue-type plasminogen activator. Several studies have shown that fibrinogen does not make an independent contribution to measures of fibrinolytic activity so that the latter is not simply a reflection of the fibrinogen level. Fairly consistently, high levels of tissue-type plasminogen activator antigen have been associated with both coronary heart disease and stroke, which seems counterintuitive since high levels of tissue-type plasminogen activator would be expected to confer protection. The explanation may be that tissue-type plasminogen activator antigen, which is what the studies in question have measured, complexes with plasminogen activator inhibitor type 1 that is present in higher concentration and for which tissue-type plasminogen activator antigen is a surrogate marker. Raised levels of D-dimer, indicating increased fibrin turnover, have been found to be predictive of future cardiovascular events—again, independent of fibrinogen. The association of activity of plasminogen activator inhibitor type 1 with coronary heart disease is not seen after adjustment for the features of insulin resistance (body mass index, triglyceride, high-density lipoprotein cholesterol, blood pressure, and diabetes) so that the prognostic value of plasminogen activator inhibitor type 1 may be related chiefly to this syndrome.

Other factors

Whereas inherited deficiencies of the naturally occurring inhibitors of coagulation such as antithrombin, protein C, and protein S clearly increase the risk of venous thromboembolism, the contribution that alterations in the levels of these proteins makes to arterial thrombosis is unclear. Anecdotally, case reports and case series have described deficiencies of these inhibitors in patients who have sustained arterial events. More formal studies have given conflicting and inconclusive results. What evidence there is suggests that low antithrombin levels may predispose to coronary heart disease.

Controversy surrounds the role in arterial disease of two recently recognized clotting factor polymorphisms, the factor V Leiden mutation and the prothrombin G20210A mutation (a polymorphism in the 3' untranslated region of the gene associated with higher prothrombin levels), in contrast to their generally accepted role in venous thromboembolism (see Chapter 15.15.3.1). Resistance to activated protein C, the laboratory phenotypic abnormality which led to discovery of the factor V Leiden mutation at one of the activated protein C cleavage sites on factor Va, has been described in patients with arterial disease. The roles in coronary heart disease of particular polymorphisms of platelet surface glycoprotein receptors and circulating levels of thrombomodulin, an endothelial receptor that binds thrombin, thereby leading to activation of protein C (anticoagulant effect) and thrombin-activatable fibrinolysis inhibitor (antifibrinolytic effect), are also uncertain. The antiphospholipid syndrome, which is characterized by both venous and arterial thrombosis together with laboratory evidence of anticardiolipin antibodies and/or the lupus anticoagulant, is discussed in Chapter 13.14 and Chapter 18.10.2.

The contact system consists of factor XII, factor XI, prekallikrein, and high molecular weight kininogen. Regulation of the system is provided in part by the multifunctional inhibitor, C1 inhibitor. The step which initiates activity is the conversion of factor XII to its derivative enzyme, factor XIIa, in response to exposure to biological substances with a negatively charged surface, for example lipopolysaccharide and phosphotidylinositol. The generation of factor XIIa triggers several activating reactions along pathways concerned with the response to injury, including activation of factor XI (which activates factor IX in the intrinsic pathway of coagulation), the production of kallikrein (which cleaves high molecular weight kininogen to bradykinin), the production of plasmin and renin, degranulation of neutrophils, activation of collagenase, and activation of the first component of complement. This range of activities raises the possibility that the contact system plays a co-ordinating role in the response to injury. A high level of factor XIIa is associated with raised levels of a number of familiar risk factors for coronary heart disease,

including plasma triglyceride and systolic blood pressure, and it is an independent predictor of coronary heart disease.

Overall, the evidence shows that predisposition to thrombosis and coronary heart disease is associated with changes in several components of the coagulation system, as well as with platelet behaviour, and the question arises as to whether they represent causality. If so, what are the pathways involved and what are the implications for management and prevention through measures affecting the haemostatic system? There are three main ways in which these questions can be approached: first, detailed laboratory studies; secondly, the extent to which the associations of clotting factor with coronary heart disease are consistent with the effects of known risk factors such as smoking; and thirdly, the ability of agents used in randomized controlled trials (for whatever reason and with whatever clinical outcome) to affect particular pathways.

Laboratory studies

The effects of fibrinogen have been extensively studied and are summarized in Fig. 2. High fibrinogen levels make a substantial contribution to the viscosity of whole blood and plasma and to the amount of fibrin deposited when coagulation is initiated, they increase platelet aggregability, enhance the binding of leucocytes to platelets and endothelial cells, decrease clot deformability, and contribute to the atheromatous process—all of which have been shown to, or are likely to, increase the risk of thrombosis and thus of clinical events. The fibrinogen level itself is influenced by a range of characteristics, including smoking (increase), moderate alcohol consumption (decrease), and genetic characteristics. As an acute and chronic phase protein, fibrinogen levels also rise in response to inflammatory stimuli, of which underlying vessel wall pathology may be an example. This has sometimes led to the view that fibrinogen is no more than a marker of the risk of coronary heart disease, whereas the likelihood seems to be that, whatever the original explanations for raised fibrinogen levels may be, these will increase risk. If so, fibrinogen may be considered to be both a marker and a causal feature.

Fig. 2 Summary of determinants and thrombogenic pathways of fibrinogen in the pathogenesis of coronary heart disease.

A number of metabolic studies have shown associations of factor VII with dietary fat intake and with serum triglyceride concentrations. Factor VIIa is also associated with plasma levels of the activation peptide fragment 1 + 2, an indicator of thrombin generation that is released from prothrombin upon its conversion to thrombin. Dietary and other studies have shown a pivotal role for factor IXa on the level of blood coagulability. Binding of factor VIIa to tissue factor may also lead to intracellular signalling, the consequences of which may include augmented macrophage activation. Factor VIII has been shown to increase the rate of activation of factor X by factor IXa in a dose-dependent manner.

Consistency with known risk factors

Very generally, associations of haemostatic variables with coronary heart disease are similar to the effects of more familiar risk factors. This is best illustrated for fibrinogen in Fig. 2 in which many of the personal or lifestyle characteristics apparently influencing fibrinogen levels are known to be associated with the risk of coronary heart disease itself—for example the increased risk due to smoking and the protective effect of a moderate intake of alcohol. Indeed, the associations of smoking and alcohol with fibrinogen are likely to explain, at least in part, how these aspects of lifestyle affect coronary heart disease itself.

One feature absent from Fig. 2 is dietary intake, particularly of saturated fat, but diet undoubtedly exerts a major effect on factor VII activity. Obesity and the other features of the insulin resistance syndrome impair fibrinolytic activity, as does smoking. The general conclusion, therefore, is that the personal and lifestyle influences on the risk of coronary heart disease operate through effects on the haemostatic system and thrombotic tendency as well as through more familiar lipid pathways. The implications for pharmacological intervention are considerable (see below).

Other effects of haemostatic variables in arterial disease

Response to injury

It is becoming increasingly clear that besides coagulability the haemostatic system plays a major role in the response to injury (although in evolutionary terms these two functions serve the same common purpose of repair of injury). Fibrinogen, for example, serves as a cell–cell adhesion protein for binding between platelets, monocytes, neutrophils, and endothelial cells. Fibrin forms a temporary matrix at sites of vessel wall injury, providing a framework for infiltration of smooth muscle cells and fibroblasts. Factor Xa stimulates proliferation of smooth muscle cells, and by limited proteolysis of protease-activated receptors, thrombin induces many cellular inflammatory responses including the expression of cytokines and adhesion molecules. Factor XIIa and kallikrein activate neutrophils and trigger the degranulation reaction. Inflammatory products have actions on components of the haemostatic system. For example, leucocytosis with neutrophil activation is a feature of atherosclerotic disease. Neutrophil elastase and cathepsin-G have diverse actions (at least *in vitro*) on the haemostatic pathway including limited proteolytic cleavage of factor IX, factor VII, factor VIII, factor V, and of platelets, and degradation of fibrinogen, fibrin, antithrombin, and tissue factor pathway inhibitor. The significance of interactions such as these between the haemostatic system and the inflammatory/immune mechanism is poorly understood, but they are likely to be pivotal in ensuring an integrated response to injury and tissue defence and repair. In summary, the original hypothesis that the haemostatic system contributes to coronary heart disease simply by a direct effect on thrombogenic potential, and thus through 'hypercoagulability', probably now requires modification. Account must also be taken of other processes involved in the pathogenesis of coronary heart disease to which changes in the coagulation system may be a secondary response, although some of these changes may then contribute to thrombotic potential.

Inflammatory markers and arterial disease

Recent years have seen the widespread recognition that atherosclerosis is an inflammatory process and the emegence of evidence that circulating markers of inflammation might be used to predict the risk of coronary events.The most consistently observed association has been that of the acute-phase reactant, C-reactive protein (CRP), with coronary risk. In healthy, asymptomatic adults, a single CRP measurement (using a sensitive assay) in the high normal range is associated with an increased risk of angina, acute myocardial infarction and death. The association is independent of lipids and its strength is similar to that observed with other risk factors including cholesterol and fibrinogen. CRP is also associated with the risk of recurrent events in those with established coronary heart disease.

CRP is synthesized in the liver and, stimulated by cytokines (particularly interleukin 6) , its level rises more than 100-fold in response to severe infection. In this setting, it binds to phosphocholine on the surface of invading microbes and assists their killing by complement and phagocytes. It also diminishes adherence of leucocytes to the vascular endothelium, releasing marginated neutrophils to infected sites while preventing the accumulation of leucocytes in uninflamed tissues. By these and other mechanisms, CRP (together with other inflammatory markers) provides an important survival function.

By contrast with this beneficial effect, repeated or prolonged low-grade stimulation of the acute phase response might have harmful consequences (see below). In the context of arterial disease, debate continues as to whether CRP is merely a marker of the inflammation that characterizes atheroma or is directly involved in the pathogenesis of atherothrombosis. In support of the latter view, CRP is deposited in human atherosclerotic lesions and is capable of binding to both enzymatically-degraded, non-oxidized low-density lipoprotein and to the terminal complement complex, C5b-9, thereby promoting inflammation. It is chemotactic for

monocytes and may play a role in the recruitment of monocytes during atherogenesis. It may contribute to the expression on the endothelium of adhesion molecules such as intercellular adhesion molecule I (ICAM-I), vascular-cell adhesion molecule I (VCAM-I), and E-selectin, further enhancing the local inflammatory response within atheromatous plaques. CRP may also promote thrombosis by enhancing tissue factor expression by monocytes. Therefore, like fibrinogen, CRP may be both a marker of the underlying pathological process and a direct contributor to the development of atherothrombosis.

It is uncertain whether the acute phase stimulus for increased CRP levels comes from the atheromatous plaques themselves, from arterial infection, or from chronic extravascular stimuli. Causes of the latter may include smoking, chronic mucosal infections such as bronchitis, gastritis and periodontitis, and obesity, with strong associations having been made between the levels of body fat and inflammatory markers with adipose tissue increasingly recognized as a source of cytokines including interlukin 6. Indeed different mechanisms might operate within one individual. As yet little is know about the genetic determinants of CRP.

Statin therapy, administered to lower cholesterol levels in those at risk of coronary events, also lowers levels of CRP, an effect which appears to be independent of their effect on lipid levels. The possibility has therefore been raised that CRP might be added to an individual's risk factor profile when deciding whether or not to use a statin for the primary prevention of coronary heart disease but this requires confirmation in prospective trials.

Associations between many other inflammatory markers and coronary heart disease risk have been reported. Inflammatory markers are synthesized in a number of different sites including the liver (CRP, serum amyloid A), macrophages (lipoprotein-associated phospholipase A_2, soluble phospholipase A_2), the vessel wall (ICAM-I, VCAM-I, E-selectin), and adipocytes (cytokines including interleukin-lb, interleukin 6, tumour necrosis factor a). It is possible that each makes a contribution to the chronic process of atherosclerosis, although in prospective studies they have generally been less clearly associated independently with coronary heart disease than has CRP. However, this may be for assay-related rather than biological reasons and independent associations of soluble ICAM, interleukin 6, and lipoprotein-associated phospholipase A_2 with coronary disease have been reported.

Homocysteine

Although not itself a component of the haemostatic system, the sulphur-containing amino acid homocysteine has emerged in recent years as a potentially important risk factor for arterial disease (and also for venous thromboembolism), with postulated mechanisms of effect that may be mediated in part through components of the haemostatic system.

It was first recognized in the 1960s that premature atherothrombosis was often seen in individuals with the rare inherited metabolic disorder homocystinuria, in which the plasma level of homocyteine is very high. Within the past decade evidence has emerged from observational studies showing an association between homocysteine levels and the risk of vascular disease within the general population, although the causal nature of this association remains to be established.

Homocysteine is a byproduct derived from the metabolic demethylation of dietary methionine. Study of the metabolic pathway of homocysteine metabolism (Fig. 3) shows that it can be metabolized either by trans-sulphuration, with vitamin B_6 (pyridoxine) as a cofactor, or by remethyl-ation to methionine, with vitamin B_{12} as a cofactor and N^5-methyl-tetrahydrofolate (derived from dietary folate) as the methyl donor. In the liver, betaine can act as an alternative methyl donor.

Fig. 3 Homocysteine metabolism. MTHFR, methylenetetrahydrofolate reductase; CbS, cystathionine b synthase

Homocysteine is present in plasma in several forms: 70 to 80 per cent is disulphide-bound to plasma proteins, 20 to30 per cent combines either with itself to form the disulphide homocystine or with other thiols such as cysteine to form mixed disulphides, and about 1 per cent circulates as the free thiol. Homocysteine (usually as the combined total of the different forms) can be measured either in the fasting state or post- methionine-loading with a standard amount of methionine. The latter estimate may be particularly sensitive to disturbances in the trans-sulphuration pathway, thereby enabling additional cases of hyperhomocysteinaemia to be detected, but it is inconvenient for patients. Blood samples should ideally be centrifuged immediately because homocysteine is progressively released from blood cells with the passage of time. If this is not possible, samples should be placed on ice following collection and centrifuged as soon as possible. The fasting reference range in Western populations is approximately 5 to 15 µmol/1, reflecting levels 2 standard deviations above and below the mean, and higher levels are arbitrarily categorized as moderate (16–30 µmol/l), intermediate (31–100 µmo/l), and severe (>100 µmol/l) hyperhomocysteinaemia. However, although such a range can be defined statistically, there is no clear threshold effect in studies that have reported positive associations between homocysteine levels and atherothrombotic disease, an analogous situation to that observed with other CHD risk factors such as cholesterol and blood pressure where risk rises even within the 'normal range'.

Plasma homocysteine levels are influenced by a number of factors, both genetic and environmental. Nutritional deficiency of folate is the most common cause of hyperhomocysteinaemia and deficiencies of vitamin B_{12} and vitamin B_6 also contribute. Other acquired causes of hyperhomocysteinaemia include renal impairment, hypothyroidism, malignancy, severe psoriasis and drugs that interfere with folate or B6 metabolism. Levels rise with age and are higher in males than females and in smokers compared to non-smokers. The combined oral contraceptive pill and hormone replacement therapy appear to lower the concentration. Homocysteine should probably not be measured immediately after an occlusive vascular event as levels may be transiently depressed.

The most common genetic cause of moderate hyperhomocysteinaemia is a point mutation (C-to-T substitution at nucleotide 677) in the coding region of the gene for N^5, N^{10}-methylenetetrahydrofolate reductase (MTHFR) that results in a thermolabile variant of the enzyme with about half normal activity. The homozygous MTHFR polymorphism is present in 10 to 15 per cent of Caucasians and is associated with an increase in homocysteine levels particularly if folate status is suboptimal. Rarer inherited causes are covered in Chapter 11.1

The data linking homocysteine and atherothrombotic disease are somewhat inconsistent, perhaps inevitable given that the relationship has been examined in over12 000 patients in more than 100 observational studies. Data from case-control studies have mostly reported a positive association between homocysteine and arterial disease. Data from prospective cohort studies have been less consistent, although the association has been in a positive direction, even if not significantly so, in the majority. Moreover, the association appears to be independent of traditional CHD risk factors such as age, sex, smoking, blood pressure, and cholesterol. Critics of the association of homocysteine and atherothrombosis point to the lack of consistent findings in the prospective studies, the lack (so far) of an association between the common genetic marker of raised homocysteine (the thermolabile MTHFR genotype), and vascular disease despite the association of genotype with homocysteine level, and the possibility that homocysteine may be simply a marker of another causal risk factor. Overall, it seems probable that homocysteine is an independent risk factor and the results of further genetic analyses and intervention trials will hopefully resolve this uncertainty.

The mechanism(s) by which homocysteine might promote atherothrombosis remains speculative. A number of possible explanations have been put forward but have often been based on *in vitro* studies that have used higher levels of homocysteine than those typically found clinically and therefore should be interpreted with caution. They include effects on platelet adhesiveness or activation, activation of clotting factors V and X, and inhibition of fibrinolysis through enhanced binding of lipoprotein(a) to fibrin. Possible effects on the endothelium may result from oxidative damage and include increased tissue factor and decreased thrombomodulin expression, and inhibition of nitric oxide. Proliferation of smooth muscle may be enhanced.

Folic acid (pteroylmonoglutamic acid) is the single most effective treatment for hyperhomocysteinaemia. It is the synthetic version of dietary folate with twice the

bioavailability and it lowers homocysteine levels even in those who are not folate deficient. For most people the maximum reduction of about 25 per cent is seen with a dose of 0.8 mg daily although patients with renal impairment need much higher doses. Vitamin B_{12} produces a smaller 7 per cent decrease in homocysteine. Dietary sources of folate include green vegetables and fortified breakfast cereals. In the United States flour has been fortified with folic acid since 1998 in order to reduce the risk of neural tube defect by ensuring improved folate status in women of child-bearing age. The level of fortification is likely to produce an extra 0.1 mg at least of folic acid per day in the diet and therefore lead to a partial lowering of homocysteine in the general population. Discussions are ongoing as to whether similar measures should be adopted in the United Kingdom. Concerns have been expressed about increasing the risk of subacute combined degeneration of the cord in patients with undiagnosed B_{12} deficiency through correction of the haematological manifestations by administration of folic acid potentially masking development of the neurological condition. The extent of this theoretical risk is uncertain and is probably extremely small but no consensus has yet been reached on whether B_{12} deficiency should be excluded before starting higher doses of folic acid or whether vitamin B_{12} should be administered in conjunction with folic acid.

A number of clinical trials with therapy that lowers homocysteine (with folic acid alone or in combination with vitamin B_6 or B_{12}) are under way in an effort to prevent coronary and cerebrovascular disease and venous thromboembolism and the results should become available within the next few years.

Implications for clinical practice

Screening and diagnosis

Recent work on the haemostatic system has certainly led to improved understanding of the pathogenesis of coronary heart disease, but only some of the information gained so far has implications for clinical practice—in particular, attempting to identify those at increased risk of first or recurrent events. Measuring fibrinogen and assessing fibrinolytic activity are the two investigations that may be helpful. A high fibrinogen level is sometimes the only identifiable risk factor in a patient referred for investigation because of a strong family history of coronary heart disease, for example, or in some patients who have recovered from myocardial infarction and in whom there are no other obvious risk factors. It may also, of course, be an additional finding in those with other risk factors. Although there is only limited evidence on the value of lowering fibrinogen levels (see below), information about these may be useful in deciding whether, for example, to recommend low-dose aspirin for primary prevention—a measure that should almost certainly be taken only after much more careful consideration than is often the case—even though aspirin does not lower fibrinogen. Measuring the euglobulin lysis time and the activity of plasminogen activator inhibitor type 1 can also be helpful, though they are often to be explained by obesity and other features of the insulin resistance syndrome.

Fibrinogen and fibrinolytic activity should only be measured some time after an acute episode of coronary heart disease and in the absence of recent infection. As for other risk factors such as cholesterol and blood pressure, they should be measured several times before an individual's habitual level can be established with any certainty. There is now a World Health Organization standard for fibrinogen, which means that ranges and values can be much more confidently compared between different centres than previously. Assays of factor VII activity are difficult to perform and the interpretation of results is uncertain. There are no established interventions for lowering levels of activity of factor VIII. While there is some evidence that low levels of antithrombin increase the risk of coronary heart disease, there are no specific agents for raising them.

Prevention

Primary prevention

In contrast to the value of aspirin in reducing recurrent major vascular events and mortality after a first attack (secondary prevention, see below), its value in primary prevention seems on present evidence to be confined mainly to reducing non-fatal myocardial infarction. There is no evidence for a reduction in fatal episodes of coronary heart disease and it is possible that the risk of cerebral haemorrhage is actually increased. It is also possible that the reduction in non-fatal events is mainly confined to those who are normotensive. Thus, those with raised blood pressure may not only experience more cerebral haemorrhage but also be exposed to the risk of gastrointestinal haemorrhage while deriving no protection against coronary heart disease. These conclusions need to be strengthened or refuted by results from further studies, but they do justify considerably more thought than is commonly given to the use of aspirin in primary prevention: many middle-aged men take aspirin indiscriminately and not necessarily beneficially and safely.

One trial has evaluated low-dose aspirin (75 mg daily) and low-intensity oral anticoagulation with warfarin (aiming at an international normalized ratio of 1.5) either singly or in combination in men at increased risk of coronary heart disease. Both agents reduced the incidence of major coronary heart disease by about 20 per cent. However, in common with other evidence in primary prevention, aspirin achieved this as a result of reducing non-fatal events by just over 30 per cent (and, if anything, slightly increasing the risk of fatal episodes). Warfarin reduced fatal episodes by 39 per cent but had little effect on non-fatal events. The combination of both agents reduced events, whether fatal or non-fatal, by 34 per cent. Warfarin may also have reduced the onset of angina pectoris slightly. There was no demonstrable difference in serious bleeding between the three active treatment groups (warfarin and aspirin together, warfarin alone, and aspirin alone) or between them and the placebo group (although minor bleeding was clearly more frequent in those on active treatment regimens). The assumption that warfarin is intrinsically more dangerous than aspirin was therefore not supported. The disadvantage of the need for international normalized ratio and dose monitoring when using warfarin is balanced by the possibility of a substantial reduction in fatal events which needs to be considered alongside the inability to predict which first major events of coronary heart disease will be fatal and the high proportion of those experiencing their first major event who die.

There is still only limited agreement about the optimal dose of aspirin: the evidence mostly points to the need for no more than 75 mg and certainly no more than 300 mg daily, both for antithrombotic effect and safety.

Secondary prevention

Antithrombotic treatment in the early stages of myocardial infarction, in which a combination of aspirin and thrombolytic treatment is used, merges into the longer-term secondary prevention of further episodes. Aspirin in the early stages reduces further major vascular events (myocardial infarction, stroke, or vascular death) by some 25 per cent, the reduction in non-fatal episodes of coronary heart disease being somewhat more than for fatal outcomes. The proportional benefits of aspirin are similar in older and younger patients, in men and women, in normotensive and hypertensive patients (which contrasts with the possible difference in primary prevention), and in non-diabetic and diabetic patients. Absolute reductions, however, are greater in the higher-risk groups (for example older, hypertensive, or diabetic patients) because of their higher event rates. There is little or no formal evidence on how long aspirin should be taken after an initial event, but since those who have already experienced episodes of arterial disease are likely to remain at high risk indefinitely, antithrombotic treatment should probably also be continued long term.

Despite the clear value of aspirin in secondary prevention, the benefits of oral anticoagulation should not be overlooked. First, it is possible that anticoagulation confers slightly greater protection against recurrence than aspirin, perhaps because of the effect of thrombin, which is reduced by warfarin, on platelets as well as fibrinogen. Despite the disadvantages of anticoagulation, this extra benefit (if real) may still be worthwhile for a common condition with a high risk of recurrence. Secondly, the value of oral anticoagulants in modifying fibrin production may add to the value of aspirin in reducing platelet aggregability. The value of combined antithrombotic regimens has been well illustrated through the concurrent use of aspirin and thrombolytic therapy in early myocardial infarction and in the postoperative treatment with aspirin and warfarin of patients undergoing heart valve surgery. Aspirin with heparin followed by warfarin is also beneficial in the setting of acute coronary syndromes. Other trials have cast doubt on the possible benefit of combined antithrombotic therapy for recurrent coronary heart disease, but they used fixed or capped low-dose warfarin, whereas it is almost certainly necessary to give warfarin in a dose-adjusted manner, i.e. to achieve a target international normalized ratio. Combined regimens of agents modifying different aspects of platelet function such as aspirin and dipyridamole in the secondary prevention of stroke further illustrate the potential value of modifying more than one pathway at a time, provided the risk of serious bleeding is not unacceptably increased. An obvious question not so far tested in randomized trials is the potential value of the simultaneous use of antithrombotic and lipid-modifying agents. While intravenous antagonists of the platelet glycoprotein IIb/IIIa receptor are effective in the early postacute management of acute coronary syndromes, oral therapy does not appear to confer benefit.

Further reading

Antiplatelet Trialists' Collaboration (1994). Overview of randomised trials of antiplatelet therapy—I: Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. III: Reduction in venous thrombosis and pulmonary embolism by antiplatelet prophylaxis among surgical and medical patients. *British Medical Journal* **308**, 81–106, 235–46.

Banerjee AK *et al*. (1992). A six year prospective study of fibrinogen and other risk factors associated with mortality in stable claudicants. *Thrombosis and Haemostasis* **68**, 261–3.

Cairns JA *et al*. (2001). Antithrombotic agents in coronary heart disease. *Chest* **119 (suppl)**, 228S–252S.

Danesh J, Collins R, Peto R (1997). Chronic infections and coronary heart disease: is there a link? *Lancet* **350**, 430–6.

Danesh J *et al.* (1998). Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease. Meta-analyses of prospective studies. *Journal of the American Medical Association* **279**, 1477–82.

Davies MJ, Thomas A (1984). Thrombosis and acute coronary-artery lesions in sudden cardiac ischemic death. *New England Journal of Medicine* **310**, 1137–40.

Ernst E *et al.*, eds (1992). *Fibrinogen: a 'New' Cardiovascular Risk Factor*. Blackwell-MZV, Vienna.

Gillis S, Furie BC, Furie B (1997). Interactions of neutrophils and coagulation proteins. *Seminars in Hematology* **34**, 336–41.

Hankey GJ, Eikelboom JW (1999). Homocysteine and vascular disease. *Lancet* **354**, 407–13.

MacCallum PK, Meade TW, eds (1999). *Thrombophilia*, 2nd edn. *Baillière's Clinical Haematology* **12**, London.

Medical Research Council's General Practice Research Framework (1998). Thrombosis prevention trial: randomised trial or low-intensity oral anticoagulation with warfarin and low-dose aspirin in the primary prevention of ischaemic heart disease in men at increased risk. *The Lancet* **351**, 233–41.

Mennen LI *et al.* (1996). Coagulation factor VII, dietary fat and blood lipids. *Thrombosis and Haemostasis* **76**, 492–9.

Miller GJ *et al.* (1996). Activation of factor VII during alimentary lipemia occurs in healthy adults and patients with congenital factor XII and factor XI deficiency, but not in patients with factor IX deficiency. *Blood* **87**, 4187–96.

Morris JN (1951). Recent history of coronary disease. *Lancet* **1**, 1–7, 69–73.

Munford RS (2001). Statins and the acute-phase response. *New England Journal of Medicine* **344**, 2016–8.

Pulmonary Embolism Prevention (PEP) Trial Collaborative Group (2000). Prevention of pulmonary embolism and deep vein thrombosis with low dose aspirin: Pulmonary Embolism Prevention (PEP). *Lancet* **355**, 1295–302.

Rader DJ (2000). Inflammatory markers of coronary risk. *New England Journal of Medicine* **343**, 1179–82.

Ridker PM *et al*. (1997). Inflammation, aspirin, and the risk ofcardiovascular disease in apparently healthy men. *New England Journal of Medicine* **336**, 973–9.

Ridker PM *et al*. (2000). C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. *New England Journal of Medicine* **342**, 836–43.

Ross R (1999). Atherosclerosis – an inflammatory disease. *New England Journal of Medicine* **340**, 115–26.

Samis JA *et al.* (1998). Neutrophil elastase cleavage of human factor IX generates an activated factor IX-like product devoid of coagulant function. *Blood* **92**, 1287–96.

15.1.3.1 Physiological considerations: biochemistry and cellular physiology of heart muscle

P. H. Sugden, N. J. Severs, K. T. MacLeod, and P. A. Poole-Wilson

Introduction

The heart of a normal human weighs 250 to 300 g, contracts at a rate of 70 to 75 beats/min at rest, and pumps approximately 5 litres of blood/min. The cardiac content of the energy transducing molecule ATP is only sufficient to support contraction for a few beats (about five to ten) and the supply of endogenous fuels (for example glycogen, endogenous triglyceride) is limited given the amount of work the heart has to perform. In order to maintain fuel oxidation, ATP regeneration, and cardiac contraction, a highly developed coronary circulation and a maintained coronary blood flow are necessary to ensure adequate delivery of O_2 and fuels, and to remove the major product of fuel oxidation (CO_2). Thus, about 5 per cent of the cardiac output is used to perfuse the heart itself (about 1 ml/min/g of myocardium). On maximal exercise both heart rate and cardiac output can increase substantially—to 200 beats/min and 20 litres/min respectively—and these changes are accompanied by an almost immediate increase of coronary blood flow to about four times its normal amount. The magnitude of the potential increase is known as the coronary reserve.

The heart is made up of many different cell types. Cardiac myocytes (the contractile cells of the heart) constitute about 75 per cent of the ventricular mass but, as they are large cells, they account for only about 25 per cent of the cell number. Other types of myocytes are specialized for the initiation of the cardiac action potential (sinoatrial nodal cells) and its transmission in a regular and co-ordinated manner to the working ventricular myocardium (conduction myocytes of the atrioventricular node, the bundle of His, and the Purkinje fibres). In addition, the heart is innervated by neurones of the autonomic nervous system. Cardiac fibroblasts synthesize and maintain the extracellular matrix. The extensive vasculature contains smooth muscle cells and pericytes, and is lined by a monolayer of endothelial cells. The endothelium is more than simply a barrier lining the blood vessels and heart chambers. Release of signalling molecules such as endothelin and nitric oxide from the endothelial cells regulates vascular smooth muscle tone and the biological properties of the myocytes themselves (see Chapter 15.1.1.2).

Interrelationships between structure and function in the ventricular myocyte and myocardium

The adult ventricular myocyte is a large cell, approximately 100 to 120 µm long and 20 to 35 µm wide (Fig. 1). Each is physically joined to approximately 10 adjacent myocytes and lies close to an extensive capillary network. The ventricular myocyte contains a highly developed contractile apparatus and a large complement of mitochondria. These fit it for its major role *in vivo*, namely the rhythmic contraction that provides the force needed for the ejection of blood from the ventricles. Atrial myocytes have a less well developed myofibrillar apparatus than ventricular myocytes, in accordance with the lesser contractile demand on the atria compared with the ventricles.

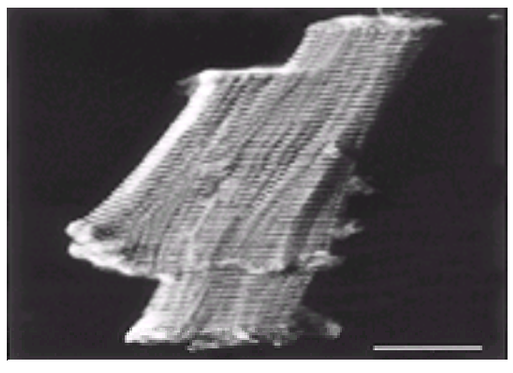

Fig. 1 Ventricular myocyte viewed by confocal microscopy. The image was prepared by combining a stack of serial optical sections through the cell. The striated myofibrils are visualized by the immunostaining of a-actinin, a component of the Z bands. The bar is 10 µm. (From Severs NJ (2000) *BioEssays* **22**, 188–99, with permission.)

Intercommunication between the myocytes and extracellular matrix

The myocardial cell is surrounded by the sarcolemma which comprises the plasma membrane (about 10 nm thick) and an outer layer (70 nm thick) called the surface coat or glycocalyx (Fig. 2). As with all biological membranes, the plasma membrane consists of a phospholipid bilayer with numerous intercalated and associated proteins. Transmembrane proteins include the channels, transporters, and pumps that allow passage of ions through the membrane, and receptor sites for hormones, pharmacologically active substances, and components of the extracellular matrix. The glycocalyx lies outside the plasma membrane (but is attached to it) and is made up of polysaccharides conjugated to protein or lipid.

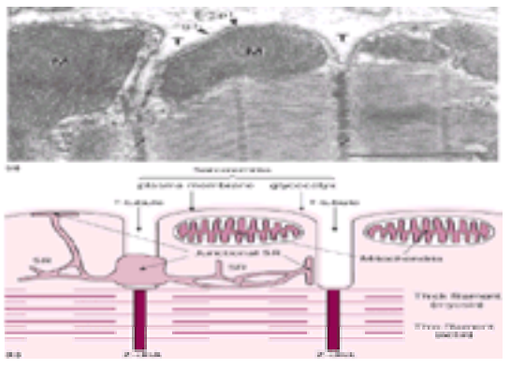

Fig. 2 Electron micrograph (a) and explanatory diagram (b) illustrating the organization of T tubules and sarcoplasmic reticulum. Labels on the micrograph as follows: T, T tubule; Z, Z disc of myofibril; M, mitochondrion; pl, plasma membrane; gl, glycocalyx. The bar is 1 µm.

At the ends of its long axis and at branches along its length, each myocyte makes contact with neighbouring myocytes through a characteristic area of plasma membrane called the intercalated disc. The intercalated disc contains three types of cell junction: the fascia adherens, the desmosome, and the gap junction. The fasciae adherentes link adjacent myocytes mechanically so that force can be transmitted between them. The contractile apparatus is anchored to the fascia adherens by a series of linking proteins such as a-actinin, filamin, and vinculin; these bind to adhesive proteins (cadherins) which are transmembrane proteins that bond across the extracellular space. The desmosomes form sites at which the intermediate filaments of the cytoskeleton attach to the plasma membrane, the cytoskeleton being important in the establishment and maintenance of myocyte shape as it provides an intracellular supporting lattice structure. As with the fasciae adherentes, bonding at the desmosome is mediated by proteins of the cadherin superfamily. The gap junctions are the sites in the intercalated disc where the membranes of adjacent myocytes come into intimate contact. Here, clusters of channels, each made of dodecamers of the protein connexin surrounding a central pore, permit myocyte-to-myocyte communication. The permeability of these pores to ions allows electrical impulses to pass easily between myocytes. The gap junctions also allow the passage of other small molecules (< 1 kDa) between the cytoplasmic compartments of adjacent cells.

The lateral (non-intercalated disc) plasma membrane is strengthened on its cytoplasmic aspect by a net-like skeleton composed of the structural proteins dystrophin and spectrin. As well as being responsible for the skeletal muscle abnormalities of Duchenne and Becker muscular dystrophy, mutations in dystrophin are responsible for the cardiomyopathy associated with these syndromes. The skeletal structure of the membrane is further reinforced by rib-like transverse bands of vinculin, termed costameres, which contain transmembrane proteins (for example integrins) that bind to components of the extracellular matrix, thereby allowing lateral transmission of the mechanical force of contraction from the cell to the matrix.

The extracellular matrix is important for the overall morphology of the heart and provides an anchoring structure against which the myocytes contract. It is made principally of collagen and fibronectin. Fibronectin fills the spaces between the cells and possesses binding sites for the integrins of the plasma membrane and for collagen. Collagen types I and III form a fibrous network that weaves around the myocytes, maintaining their alignment, preventing overstretching, and transmitting force. This network preserves the overall shape and architecture of the heart, and acts as a spring to store energy during systole. The normal heart contains about 5 per cent collagen but, in pathological conditions, this can increase to 25 per cent. This increases myocardial 'stiffness' and can impede contraction.

Contraction

The contractile apparatus is highly ordered, consisting of bundles of striated myofibrils (approximately 1 µm in diameter with around 150 per cell) running the length of the cell (Fig. 3). As in skeletal muscle, myofibrils (which are responsible for contraction) are made up of a repeating sarcomeric unit that is about 2 µm in length in the relaxed state. Each sarcomere consists of thick filaments that interdigitate with thin filaments (Fig. 4). The thick filament is a polymer of the protein myosin, a hexamer of two myosin heavy chains and four myosin light chains. The myosin heavy chain has an elongated rod-like domain and a globular head, to each of which two light chains are bound (Fig. 5). The rod-like regions of two myosin heavy chains intertwine to form the hexamer. The thin filaments consist of a double-beaded strand of the globular protein actin with which the rod-like protein tropomyosin and members of the troponin family (troponin I, troponin C, and troponin T) are associated (Fig. 5). The thin filaments within adjacent sarcomeres are linked at the Z line. The alignment and the overlapping of thick and thin filaments gives the myofibril its striated appearance in micrographs (Fig. 3 and Fig. 4).

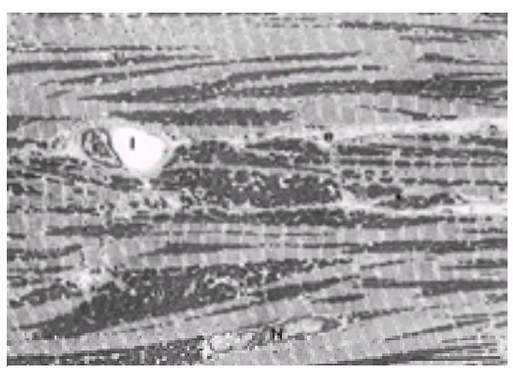

Fig. 3 Structure of ventricular myocardium by thin-section electron microscopy. The major components of the myocytes, the striated myofibrils and the mitochondria (m, seen as abundant dark-stained rounded objects), dominate the view. N, nucleus of myocyte; n, nucleus of endothelial cell; l, lumen of capillary; e, extracellular matrix. The bar is 10 µm.

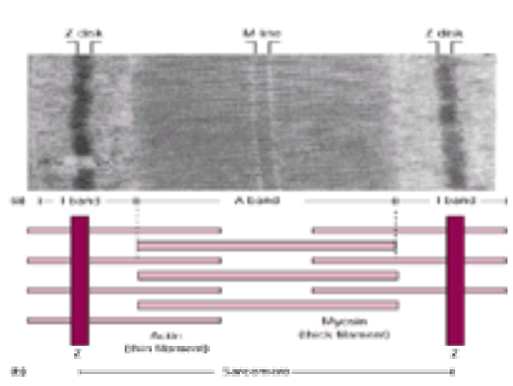

Fig. 4 High-power electron micrograph of a single sarcomere (a), the basic contractile unit of heart muscle, with explanatory diagram (b) of the organization of the thick and thin filaments, Z discs, and bands.

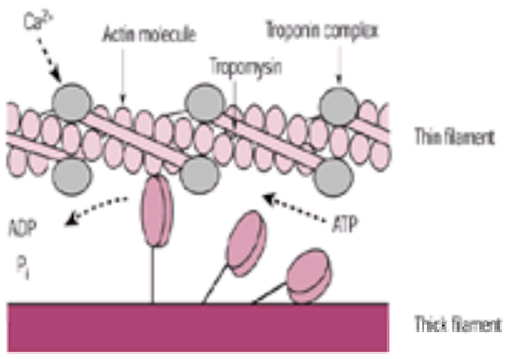

Fig. 5 Diagrammatic representation of myofilament structure and contraction. The thick filament is a polymer of myosin, the double beaded helix of the thin filament is a polymer of actin. Increases in intracellular Ca^{2+} ion concentrations are detected by troponin C of the troponin complex and this relieves the inhibition of the actomyosin ATPase by troponin I. The hydrolysis of ATP by the ATPase provides the energy for contraction. The globular myosin heads, which form crossbridges with the thin filaments, move as shown causing shortening of the sarcomere.

Myofibrillar contraction in the heart is essentially similar to that in skeletal muscle. The globular heads of myosin interact with actin, and the myofibrillar actomyosin adenosine triphosphatase (**ATPase**) transduces the chemical energy released by ATP hydrolysis into external mechanical work (Fig. 5). In the absence of Ca^{2+}, troponin I maintains the actomyosin ATPase in an inactive state. Binding of cytoplasmic Ca^{2+} ions to troponin C removes this restraint and contraction occurs by the movement of the myosin heads in the thick filaments along the actin beads in the thin filaments, with concomitant hydrolysis of ATP. Mutation of specific amino acid residues in the myosin heavy chain in particular (but also in other myofibrillar proteins) can give rise to an inheritable disease known generically as hypertrophic cardiomyopathy (see Chapter 15.8.2).

Since Ca^{2+} is intimately involved in the regulation of contractile activity, a well-developed apparatus controls the intracellular (subscript 'i') concentration of Ca^{2+} ions. The ventricular myocyte possesses an array of T tubules and an extensive sarcoplasmic reticulum, a specialized form of endoplasmic reticulum. The T tubules are finger-like invaginations from the cell surface with openings of up to 200 nm in diameter, spaced so that a T tubule lies alongside each Z disc of most (or even all) of the myofibrils (Fig. 2). Both the T tubule and the sarcoplasmic reticulum are involved in regulation of the movement of Ca^{2+}. The T tubules regulate the entry of extracellular (subscript 'o') Ca^{2+} into the myocyte, and the sarcoplasmic reticulum 'stores' Ca^{2+}_i during diastole. The regulation of movement of Ca^{2+} in the myocyte is described in more detail below.

Although much of the intracellular volume of the ventricular myocyte is occupied by myofibrils, about 30 per cent is taken up by mitochondria (Fig. 3). As in other oxidative tissues, these subcellular organelles oxidize metabolic fuels and convert the energy released by this oxidation to drive regeneration of ATP (from ADP and inorganic phosphate). In the myocyte, the majority of the energy released by ATP hydrolysis is used to power myofibrillar contraction, but there are also other essential processes that require energy (for example macromolecule synthesis and other biosynthetic pathways, ion transport, etc.).

The nucleus and the cell cycle

The mammalian ventricular myocyte is believed to lose its ability to divide during the perinatal period, hence most maturational growth occurs through cellular enlargement. However, complicating factors are that the myocyte may be multinucleate, and its nucleus may possess more that two chromosome pairs (polyploidy). Thus the arrest in the cell cycle appears to reside at the stage of cell division (cytokinesis) rather than nuclear division (karyokinesis). The reasons for the withdrawal of the myocyte from the cell cycle are not understood, and although some molecules which regulate the cell cycle are present in the ventricular myocyte, it has not yet been experimentally possible to drive the cell into division. This means that the myocardium cannot regenerate, and the loss of myocytes following, for example, an ischaemic insult is potentially disastrous. An ability to restore entry of ventricular myocytes into the cell cycle in a controlled manner might well have considerable clinical significance as it would be one step towards allowing the damaged heart to regenerate its contractile capacity. As it is, the myocyte can only increase its contractile capacity by cell enlargement (hypertrophy), and this leads to the clinical entity of cardiac hypertrophy.

Cardiac electrophysiology

Each heartbeat is initiated by a spontaneous electrical discharge in the sinoatrial node. The electrical signal passes across the atrium to the atrioventricular node, through the bundle of His, and down the Purkinje fibres to the ventricular myocardium. This incremental excitation of the heart provides a means of co-ordinating the contractile activities of the four chambers and is the basis for the electrocardiogram (see Chapter 15.3.2).

Electrical excitation of each myocyte involves the movement of ions through ion channels. These are 'excitable' macromolecules embedded in the plasma membrane which contain pores that open or close in response to a stimulus. This stimulus could be a change in membrane potential, a neurotransmitter or hormone, an intracellular second messenger or ion, or mechanical stretch of the membrane. When a channel opens, it becomes selectively permeable to a restricted series of ions, selectivity being determined by the interaction of the various ions with the channel pore. There are a large number of different types of channel, often named after the most important permeant ion they pass, for example the Na^+, Ca^{2+}, and K^+ channels. The groups of channels are functionally distinct and can be further divided into subgroups on the basis of amino acid sequence and tertiary structure. Ions move down their electrochemical gradients through the channels at high rates (> 10^6 ions/s), which distinguishes them from other ion transport proteins (for example, the Na^+,K^+-ATPase or pump, and the Na^+,Ca^{2+} exchanger, see below) which move ions across plasma membranes several orders of magnitude more slowly.

When a ventricular myocyte is at rest (diastole), there is a potential difference of – 80 mV across the plasma membrane, the inside of the cell being negative with respect to the outside. This is due to K^+ channels being open, making the plasma membrane more permeable to K^+ than any other ion. The concentration of K^+ is about 4 mmol/litre outside the cell and about 140 mmol/litre inside, so K^+ tends to leave the cell by diffusing down its concentration gradient, which results in the inside becoming negatively charged since there is no movement of anions to balance the K^+ loss. An equilibrium is established where the electronegative force retaining K^+ inside the cell balances its tendency to diffuse out of the cell down its concentration gradient. This is termed the equilibrium potential (E), and can be calculated from the Nernst equation. The calculated equilibrium potentials for important ions are shown in Table 1.

The actual transmembrane potential difference at rest and the calculated equilibrium potential for K^+ are rarely the same owing to a small leakage of other ions (mainly Na^+) into the cell. To counteract this leak of Na^+ down its concentration gradient and to maintain the concentration gradients of Na^+ and K^+ upon which the generation of the membrane potential depends, the sarcolemmal Na^+,K^+-ATPase uses energy derived from the hydrolysis of ATP to pump these ions against their concentration gradients. This process is electrogenic (three Na^+ are extruded for two K^+ entering) and generates 3 to 10 mV of the membrane potential.

When a myocyte is electrically excited, Na^+ channels open and allow Na^+ ions to enter the cell. Positive charge is taken into the cell, the membrane potential increases towards the equilibrium potential for Na^+ (Table 1), and the cell depolarizes (Fig. 6). This causes the rapid upstroke (phase 0) of the action potential. The rate of change of the potential is related to the propagation velocity of the action potential across the heart. The current (I) generated by the inward movement of Na^+ (I_{Na}), like the same current in nerve tissue, is inhibited by tetrodotoxin, lidocaine, and quinidine. Na^+ channels close very rapidly and so I_{Na} almost entirely inactivates within the first 4 ms of the action potential. A small proportion of Na^+ channels do not inactivate as rapidly and allow a small inward current to persist for up to 100 ms, i.e. during the plateau phase of the action potential (phase 2).

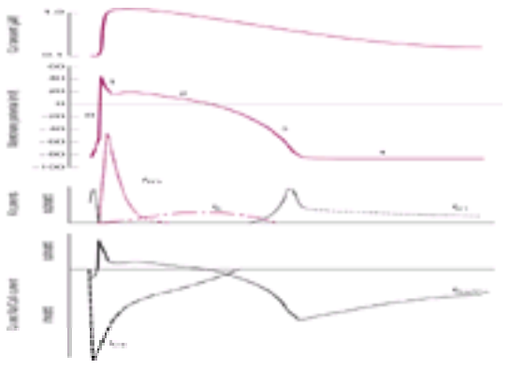

Fig. 6 Major ionic currents flowing during a ventricular myocyte action potential. Top trace: changes in cytoplasmic Ca^{2+} concentration during the action potential (Ca^{2+} transient). Second trace: the action potential recorded from a ventricular myocyte. Third and fourth traces: time courses and relative sizes of current flows during one beat. All K^+ currents (I_{TO}, I_K, and I_{K1}, seen in the third trace) repolarize the cell because of outward K^+ movement. Because of the inward movement of Ca^{2+}, Ca^{2+} current (I_{Ca}, seen in the fourth trace) is depolarizing. The Na^+,Ca^{2+} exchanger (Na/CaX, fourth trace) produces both outward and inward current ($I_{Na,Ca}$). Note that the inward Na^+ current that produces the rapid upstroke of the action potential is not shown: it is roughly eight to ten times the size of the Ca^{2+} current and has largely inactivated by the time the peak of the Ca^{2+} current is reached.

The characteristic notch observed in phase 1 of the action potential in ventricular myocytes (Fig. 6) (the notch is also particularly apparent in the Purkinje cell action potential; see Fig. 7) is caused by a transient outward current (I_{TO}), mainly carried by K^+ ions, that partially repolarizes the membrane. A number of different currents flow during phase 2 (the action potential plateau): the most important, from the point of view of the generation of contraction, is I_{Ca}. Ca^{2+} channels, which take longer to activate and inactivate than Na^+ channels, open within 3 ms of the start of the upstroke. The inward flow of Ca^{2+}, mainly through the L-type Ca^{2+} channel, maintains depolarization (Table 1 and Table 2) and can be inhibited by 'Ca^{2+} antagonists' such as verapamil and the dihydropyridines. The influx of Ca^{2+} initiate Ca^{2+}-induced Ca^{2+} release from the sarcoplasmic reticulum through the sarcoplasmic reticulum Ca^{2+}-release channels, and the increase in cytoplasmic Ca^{2+} concentration causes the myocyte to contract (see below).

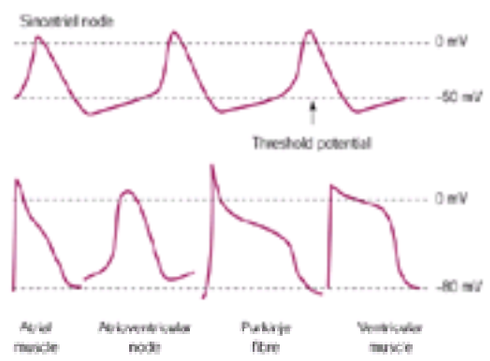

Fig. 7 Regional configuration of the action potential. In the sinoatrial and atrioventricular nodes, the cells spontaneously depolarize during diastole (phase 4 depolarization). When the membrane potential reaches a threshold value, the complete action potential is initiated. Because the sinoatrial nodal cells have the fastest phase 4 depolarization, they act as the cardiac pacemaker.

The plateau phase (phase 2) of the action potential (Fig. 6) is prolonged in ventricular myocytes because of the properties of several types of K^+ channel. The repolarizing current I_K flows through a channel that opens at positive membrane potentials and closes at negative potentials, akin to its counterpart in nerve. However, the kinetics of this channel are much slower than in nerve so that a much longer time is taken for it to start to repolarize, this being one of the reasons that a cardiac action potential is so much longer than a nerve action potential. In addition, ventricular myocytes possess another K^+ channel with peculiar characteristics. The current I_{K1} flows through a channel that first increases its conductance but then decreases it as the cell depolarizes away from E_K (anomalous rectification). The combined effect of these K^+ currents is that, despite the membrane potential approaching 0 mV during the plateau phase, a large outward K^+ current does not occur and the action potential is prolonged.

Repolarization (phase 3) starts to occur because of an increase in K^+ conductance via I_K and the termination of I_{Ca} (Fig. 6). As repolarization proceeds, the Na^+,Ca^{2+} exchanger responds to the increase in cytoplasmic Ca^{2+} concentration and produces an inward current ($I_{Na,Ca}$) through the exchange of three Na^+ entering the cell for one Ca^{2+} expelled. By producing an inward current, the Na^+,Ca^{2+} exchanger helps to prolong the plateau and slows repolarization. In ventricular myocytes, complete repolarization and a return to a negative membrane potential is eventually achieved by the current I_{K1}. The clinical consequences of the presence of I_{K1} are profound and result from this channel being acutely sensitive to the extracellular concentration of K^+. For example, shortly after myocardial infarction there is a loss of K^+ from cells and local K^+_o concentrations increase. Because K^+_o increases channel conductance and outward movement of K^+_i, I_{K1} increases accordingly. Thus, more outward current flows and the action potential duration shortens, which may lead to arrhythmia.

The configuration of the cardiac action potential differs regionally (Fig. 7) because of the presence or absence of different ionic currents. In the sinoatrial node (the pacemaker), I_{Na} is very small and the main current responsible for the depolarizing upstroke is I_{Ca}. The only repolarizing current is I_K. I_{K1} is absent and this partially explains why sinoatrial node cells have a more depolarized diastolic potential than ventricular myocytes. Sinoatrial node cells also depolarize spontaneously (phase 4), probably owing to the absence of I_{K1} and the presence of a current activated on hyperpolarization called I_f. Phase 4 is often termed the 'pre- or pacemaker potential' and is caused by the gradual decrease in I_K and increase in I_f (Fig. 6 and Fig. 7). Once the cell has depolarized to the point where Ca^{2+} channels open (the threshold), a more rapid depolarization takes place forming the upstroke of the sinoatrial node action potential. Atrial and ventricular myocytes do not have pacemaker potentials and spontaneously discharge only when injured or when there is abnormal intracellular Ca^{2+} balance. The longest action potential is in Purkinje fibres; this acts as a gate preventing retrograde activation by depolarization of adjacent ventricular myocytes. The action potential is longer in the epicardium than in the endocardium, and in the apex than in the base of the heart: the reason for this is not clear, but the discrepancy is the probable explanation for the upright T wave on the electrocardiogram.

When the cholinergic drive from autonomic neurones to the nodal cells is increased, the slope of the pacemaker potential is decreased (Fig. 8). Acetylcholine (**ACh**) opens another group of K^+ channels and activates $I_{K,ACh}$, which counters the decline in I_K and slows the rate of depolarization. I_f and I_{Ca} are also reduced and the overall effect is a reduction in the rate of production of action potentials. Conversely, upon stimulation of the sympathetic cardiac nerves, noradrenaline activates I_f and facilitates the opening of L-type Ca^{2+} channels, so increasing I_{Ca}. The net effect is to depolarize the membrane to the threshold level more quickly and increase the rate of production of action potentials. A summary of the ionic currents flowing during the cardiac action potential is given in Table 2.

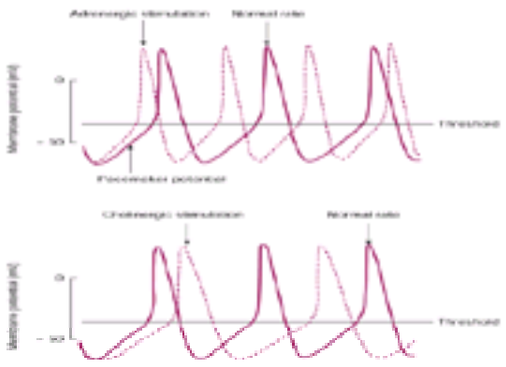

Fig. 8 Change in heart rate produced by altering the phase-4 slope of the pacemaker potential. Adrenergic stimulation increases and cholinergic stimulation decreases the slope, affecting the time taken to reach threshold.

Intracellular calcium ions—regulators of contraction

The electrical events throughout the heart initiate and regulate contraction (Fig. 9). The coupling of the electrical excitation of the heart to the production of contraction (called EC coupling) by Ca^{2+} ions involves the interaction of a number of cellular proteins involved in Ca^{2+} homeostasis. The T tubules allow the wave of depolarization of the action potential to reach deeply into the cell. The sarcoplasmic reticulum is an intracellular membranous lace-like structure surrounding the myofibrils, with swellings called junctional sarcoplasmic reticulum where the membrane of the sarcoplasmic reticulum comes close to the T tubules, (Fig. 2 and Fig. 10). During diastole, when cytoplasmic Ca^{2+} concentrations are low (around 0.1 µmol/litre), Ca^{2+} is sequestered by the Ca^{2+} buffering protein calsequestrin within the junctional sarcoplasmic reticulum. When opened, the L-type Ca^{2+} channels (also known as dihydropyridine receptors because of their sensitivity to the dihydropyridine Ca^{2+} channel antagonists) allow influx of Ca^{2+} across the sarcolemma (Fig. 10). This influx increases the local Ca^{2+} concentration around clusters of Ca^{2+} release channels in the sarcoplasmic reticulum (the ryanodine receptors, so-called because of their sensitivity to interference by the plant alkaloid ryanodine) sufficiently to open them, the number of channels activated in this way being mainly, though not exclusively, determined by the size of the Ca^{2+} current. This allows Ca^{2+} stored by the sarcoplasmic reticulum to be released into the cytoplasm (Ca^{2+}-induced Ca^{2+} release). The fluxes of Ca^{2+} combine to raise the cytoplasmic concentration of Ca^{2+} to between 5 and 10 µmol/litre, when contraction is initiated by the binding of Ca^{2+} to troponin C, relieving the inhibition of the actomyosin ATPase by troponin I.

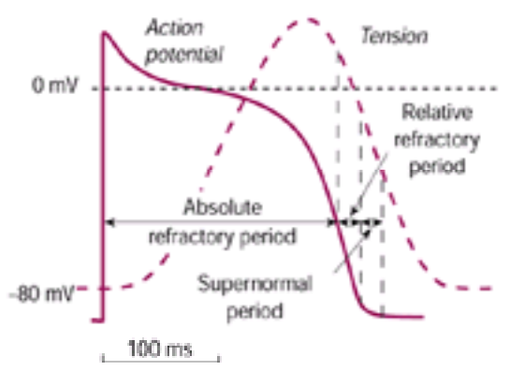

Fig. 9 The relationship between the action potential and the generation of force. The peak of force production is not achieved until near the end of the plateau phase of the action potential. This reflects the time required for Ca^{2+}-induced Ca^{2+} release. For a period between phase 0 and about halfway through phase 3, cardiac muscle cannot be excited with another stimulus no matter how strong. The muscle is in its absolute refractory period. Thus, tetanic contraction of the type seen in skeletal muscle cannot occur. When cardiac muscle is in its relative refractory period, an abnormally strong stimulus can initiate an action potential. In the supernormal period that follows, a slightly weaker stimulus that would normally fail to reach threshold can also initiate an action potential. The states of refractoriness are related to the ability of ion channels to recover from a stimulus. This recovery is both voltage- and time-dependent.

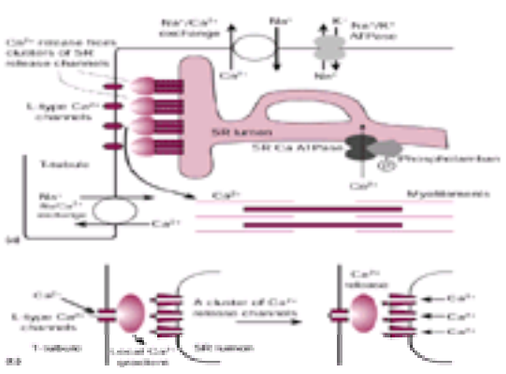

Fig. 10 EC coupling in the heart. Panel A: L-type Ca^{2+} channels allow Ca^{2+} influx across the sarcolemma and this creates I_{Ca}. This influx increases the local Ca^{2+} concentration around a cluster of Ca^{2+} release channels in the sarcoplasmic reticulum (SR) in sufficient amounts to open them (Ca^{2+}-induced Ca^{2+} release). Panel B: The opening of clusters of Ca^{2+} release channels in the sarcoplasmic reticulum allows Ca^{2+} stored by the sarcoplasmic reticulum to be released into the cytoplasm. The fluxes of Ca^{2+} combine to initiate contraction. The process is terminated by the sarcoplasmic reticulum Ca^{2+} ATPase (regulated by phospholamban) which removes Ca^{2+} from the cytoplasm and pumps it into the sarcoplasmic reticulum, and by the sarcolemmal Na^+,Ca^{2+} exchanger which expels Ca^{2+} from the cell. In steady-state conditions, the amount of Ca^{2+} leaving the cell (via the Na^+,Ca^{2+} exchanger) balances the amount entering (via the L-type Ca^{2+} channel).

Contraction is terminated by two principal mechanisms:

- By the activation of an ATP-requiring Ca^{2+} pump present in the sarcoplasmic reticulum membrane (the sarcoplasmic (endoplasmic) reticulum ATPase type 2 (**SERCA2**)), which catalyses ATP-dependent reuptake of Ca^{2+} into the sarcoplasmic reticulum.
- By the response of the sarcolemmal Na^+,Ca^{2+} exchanger to the increase in cytoplasmic Ca^{2+} (Fig. 10).

On a beat-to-beat basis, these are the main systems involved in removing Ca^{2+} from the cytoplasm and so inducing relaxation. Ca^{2+} is pumped back into the sarcoplasmic reticulum by SERCA2, which is regulated by the extent of phosphorylation of the SERCA2-associated protein, phospholamban. Hypophosphorylated phospholamban tonically inhibits SERCA2; as more phospholamban is phosphorylated (see below) the inhibition is removed and more Ca^{2+} is pumped into the sarcoplasmic reticulum. This, along with increased troponin I phosphorylation (see below), accounts for the more rapid relaxation of myocytes when the intracellular concentration of cyclic adenosine 3_,5_-monophosphate (**cAMP**) and the activity of cAMP-dependent protein kinase (see below) are increased by, for example, heightened sympathoadrenal tone.

Ca^{2+} is extruded from the cell by the sarcolemmal Na^+,Ca^{2+} exchanger that utilizes the energy associated with the concentration and electrical gradients for Na^+ to expel Ca^{2+} from the cell. It couples the transport of three Na^+ into the cell with the expulsion of one Ca^{2+}. It is thus electrogenic, and for every Ca^{2+} removed from the

cell, one positive charge enters. It can be predicted thermodynamically that the direction of ion movement mediated by the exchange can vary according to the membrane potential and the intracellular and extracellular concentrations of Na^+ and Ca^{2+}. The exchange is sensitive to the intracellular Na^+ concentration (normally about 7 to 10 mmol/litre in ventricular myocytes). When membrane potential is near diastolic levels and intracellular Na^+ concentration at normal physiological levels, the Na^+,Ca^{2+} exchanger will eject Ca^{2+} from the cell. However, if the intracellular Na^+ concentration increases by a few mmol/litre and the membrane potential becomes depolarized, the exchanger can reverse and mediate Ca^{2+} entry. Although ventricular myocytes possess other systems to decrease cytoplasmic Ca^{2+} concentrations (namely the sarcolemmal Ca^{2+} ATPase and mitochondrial Ca^{2+} uptake), these contribute less than 5 per cent towards relaxation of a normal twitch. SERCA2 and Na^+,Ca^{2+} exchange contribute about 70 and 25 per cent respectively towards relaxation, though these figures vary greatly between animal species. In steady-state conditions, the amount of Ca^{2+} leaving the cell via the Na^+,Ca^{2+} exchanger is the same as the amount entering (via I_{Ca}) to evoke Ca^{2+}-induced Ca^{2+} release, hence precise Ca^{2+} homeostasis is achieved.

Cardiac mechanics

Four key factors determine the contraction of isolated cardiac muscle preparations; these are also applicable to the intact heart. The first factor is the relationship between the initial fibre length and the force it produces: initial 'preload' stretches the sarcomeres, increasing the sensitivity of the myofilaments to Ca^{2+} and allowing them to produce force that is directly proportional to the preload (i.e. resting length). The second factor determining contraction is termed the 'afterload': if a mass is attached to a muscle just before it contracts isotonically, it represents a constant force (afterload) against which the muscle must work during contraction, and the amount and speed of contraction are inversely proportional to this. If preload and afterload are held constant, the maximum force and speed of contraction can be altered by changing the inotropic state of the muscle, which is generally brought about by chemical or hormonal influences (for example, increasing concentrations of catecholamines). In the whole heart, these three factors (preload, afterload, and inotropic state) influence the stroke volume (the volume of blood ejected by the ventricle during each systole), which, along with heart rate (the fourth factor), determines the cardiac output. These issues are discussed in Chapter 15.1.3.2.

Energy for contraction

As mentioned earlier, the heart is principally reliant on aerobic metabolism for its energy supply. In normal humans at rest, the heart extracts 60 to 65 per cent of O_2 passing through it. This corresponds to a rate of O_2 utilization of about 4.5 µmol/min/g wet weight (0.1 ml/min/g wet weight), which may increase by three- to fourfold during exercise. The comparable rates (in µmol/min/g wet weight) in other organs are: brain, 1.7; kidney, 7.1; liver 1.6; skeletal muscle at rest, 0.08; and skeletal muscle during exercise, 6.4. Thus, the maximal physiological O_2 uptake of the heart is higher than that of any other tissue.

In terms of metabolic fuels, the heart utilizes any fuel presented to it, within the constraints of metabolic regulation. Furthermore, because the heart is contracting continuously, an uninterrupted exogenous provision of fuels through the coronary circulation is essential. The major substrates for oxidation in man are lipid-derived fuels (long-chain fatty acids, principally palmitate, triglycerides, and ketone bodies (acetoacetate and 3-hydroxybutyrate)) and the carbohydrate-derived fuels (glucose, lactate, and pyruvate) (Fig. 11). Although the heart can oxidize amino acids, these probably represent a relatively minor fuel. The relative contribution of each substrate to cardiac fuel supply depends principally on the individual concentrations of substrates in the plasma (which are largely hormonally regulated).

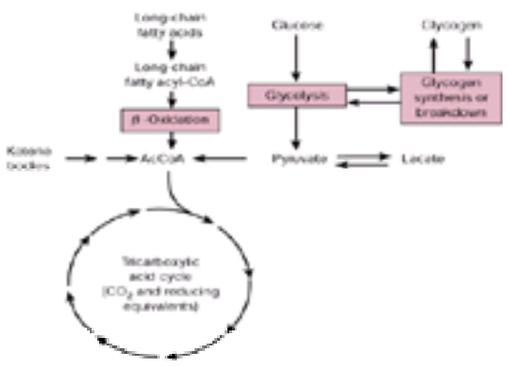

Fig. 11 Fuel utilization. Multistep processes are shown in boxes. In aerobic metabolism, the key intermediate acetyl CoA (AcCoA) is oxidized to CO_2 and water. Reducing equivalents produced mainly in the mitochondria by the tricarboxylic acid cycle and b-oxidation are passed between a series of carriers (the electron transport chain), the chemical energy released at each transfer step being used to drive regeneration of ATP from ADP and inorganic phosphate (oxidative phosphorylation). The ultimate electron acceptor for the reducing equivalents is O_2, which is reduced to water.

Carbohydrate metabolism

Glucose crosses the myocardial plasma membrane by two carrier-mediated mechanisms (the type 1 and the type 4 glucose transporters). The activity of the type 1 glucose transporter is largely independent of insulin but is controlled by the intra- and extracellular concentrations of glucose. Insulin can increase glucose uptake by recruiting intracellular type 4 glucose transporter to the plasma membrane. The principal use of intracellular glucose is to provide energy for contraction. In addition, the heart has a limited capacity to store carbohydrate as the polysaccharide glycogen. Whilst glycogen breakdown may not be quantitatively significant under normal conditions, it represents a fuel that can be used for a limited period in pathological conditions (e.g. during myocardial infarction, when the supply of fuels and O_2 is disrupted).

Each molecule of glucose is degraded through the exclusively cytoplasmic glycolytic pathway to pyruvate, the chemical energy released allowing the regeneration of two molecules of ATP (from ADP) per glucose molecule utilized and the concomitant reduction of the electron carrier nicotinamide adenine dinucleotide (**NAD+**) to NADH (Fig. 11). This pathway occurs anaerobically but is inefficient in terms of the quantity of ATP regenerated per glucose utilized. Under the aerobic conditions that normally exist in the heart, pyruvate is transported into the mitochondria and the glycolytically derived reducing equivalents enter on a shuttle mechanism (the malate/aspartate shuttle) regenerating cytoplasmic NAD^+. The remaining steps of carbohydrate metabolism take place in the mitochondria: pyruvate is oxidized to acetyl CoA and NADH (NAD^+ accepting the reducing equivalents) by the pyruvate dehydrogenase multienzyme complex, and acetyl-CoA then enters the tricarboxylic acid cycle, with the net result being the complete oxidation of the glucose molecule (glucose + $6O_2$ ® $6CO_2$ + $6H_2O$). The bulk of the mitochondrially generated ATP is then exchanged with cytoplasmic ADP, making it available for myofibrillar contraction and other processes.

In some pathological circumstances (for example in coronary artery disease), the heart may become intermittently hypoxic. When this happens, glucose is increasingly metabolized anaerobically (glucose ® 2 lactate + $2H^+$), and lactate and protons are released into the circulation. By contrast, under aerobic conditions exogenous lactate or pyruvate can also be utilized by the heart, entering oxidative metabolism in the same way as glycolytically derived pyruvate and NADH.

Metabolism of fatty fuels

Long-chain fatty acids, triglycerides, and ketone bodies are all capable of providing energy for the heart. Long-chain fatty acids (principally palmitate) are present in the plasma either non-covalently bound to albumin or covalently bound as triglycerides, which are in turn complexed with apolipoproteins. The albumin-bound long-chain fatty acids enter the ventricular myocyte by a carrier-mediated process that is still relatively ill-defined. Triglycerides are hydrolysed by the ectoenzyme lipoprotein lipase on the capillary wall to form long-chain fatty acids (and glycerol). Ketone bodies are synthesized hepatically from long-chain fatty acids and are (compared with long-chain fatty acids) a relatively soluble, readily diffusible, non-toxic fuel.

Long-chain fatty acids and ketone bodies can only be metabolized aerobically, and their catabolism takes place exclusively in the mitochondria (Fig. 11). Two-carbon fragments are successively removed from long-chain fatty acids (as long-chain fatty acid CoA) in a series of reactions known generically as b-oxidation to form acetyl CoA. Each turn of b-oxidation generates sufficient reducing equivalents to regenerate five ATP molecules. Acetyl CoA is then oxidized through the tricarboxylic acid cycle to regenerate more ATP, hence the energy yield from long-chain fatty acid oxidation is considerable. Lipid-derived fuels are preferentially used by the heart and their utilization diminishes the use of carbohydrate, thereby conserving glucose for tissues that are obligatorily dependent on carbohydrate as an energy source. This

is achieved by the inhibitory phosphorylation of pyruvate dehydrogenase multienzyme complex, which is stimulated by increased acetyl CoA and NADH concentrations, and by inhibition of the glycolytic pathway by intermediates of the tricarboxylic acid cycle, such as citrate.

Transmembrane and intracellular signalling pathways in the heart

Systemic stimuli in the form of neuroendocrine factors (for example catecholamines and insulin) impinge on the extracellular face of the sarcolemma of the myocyte. These molecules bind to transmembrane receptors that transfer information carried by the stimuli into the inside of the cell. Intracellular signalling pathways then transmit information from one part of the cell to another. Many of these pathways involve reversible protein phosphorylation (catalysed by protein kinases) and dephosphorylation (catalysed by protein phosphatases). Some hormones (for example thyroid hormone and oestrogen) interact directly with intracellular receptors. This group is mainly concerned with signalling at the level of transcription, which is achieved by the formation/presence of the signal–receptor complex in the nucleus.

Extraneous signals that utilize transmembrane receptors mediate their intracellular responses through signalling pathways that, although limited in number compared with the variety of receptors, produce diverse intracellular responses in different cellular contexts. End responses are cell specific because of the variety of proteins expressed in a given cell and the wide range of cell structures.

Intracellular signalling processes are discussed elsewhere. Those relevant to the heart are shown in Fig. 12.

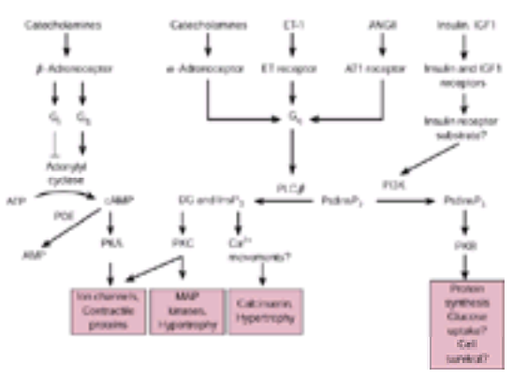

Fig. 12 Intracellular signalling pathways. Neurohumoral agonists bind to their transmembrane receptors and stimulate the formation of a variety of small molecule second messengers (cAMP, diacylglycerol (DG), inositol 1,4,5-trisphosphate ($InsP_3$), phosphatidylinositol 3,4,5,-trisphosphate ($PtdInsP_3$)). These then often activate enzymes, many of which catalyse reversible protein phosphorylation (protein kinases) or dephosphorylation (protein phosphatases) and alter the phosphorylation states of proteins which regulate biological processes, thereby modulating the rates of those processes. With the exception of the effect of the G_i protein on adenylyl cyclase, all stages shown are stimulatory. Other abbreviations: ANGII, angiotensin II; ET-1, endothelin-1; G_i, G_q, and G_s, heterotrimeric G_i, G_q, and G_s proteins; IGF 1, insulin-like growth factor 1; MAP kinases, mitogen-activated protein kinases; PDE, cyclic nucleotide phosphodiesterase; PI3K, phosphatidylinositol 3-kinase; PKA, PKB, and PKC, protein kinase A, B, and C, respectively.

G protein-coupled receptors

The ventricular myocyte is heavily dependent on signalling through transmembrane G protein-coupled receptors. The intracellular domains of these receptors interact with one or more of the large family of heterotrimeric (abg) guanine nucleotide-binding proteins (G proteins). In their inactive state, guanosine diphosphate (GDP) is bound to the G protein a subunit (a.GDP). The binding of agonists to the extracellular domain of their individual receptors and the receptor-G protein interaction that follows stimulates exchange of GDP for guanosine triphosphate (GTP) on the a subunit and dissociation of abg into a.GTP and bg. This dissociation is reversed by the innate GTPase activity of the a subunit and a(GDP). bg is reformed. a.GTP (and possibly bg) are effectors of membrane-bound enzymes that produce so-called 'second messengers'.

The archetypal second messenger is cAMP, which is formed from ATP by the membrane enzyme adenylyl cyclase following b-adrenergic receptor activation (Fig. 12). The interaction of the b-adrenergic receptor with the G_s protein causes formation of a_s.GTP which stimulates adenylyl cyclase and the cAMP formed activates cAMP-dependent protein kinase. The activation of this protein kinase is terminated by hydrolysis of cAMP to AMP by a group of phosphodiesterases. The b-adrenoceptor also activates G_i proteins (a_i.bg) which counteract the effects of a_s on adenylyl cyclase and act as a negative feed-back mechanism. In the heart, b-adrenergic agonists are positively inotropic, positively lusitropic (i.e. they increase the rate of relaxation), and positively chronotropic. The positive inotropism is the result of a cAMP-dependent protein kinase-catalysed phosphorylation of the L-type Ca^{2+} channel which enhances Ca^{2+} entry into the cell, and the direct activation of the L-type Ca^{2+} channel by a_s.GTP. The positive lusitropism involves increased phosphorylation of the sarcoplasmic protein phospholamban and the myofibrillar protein troponin-I. In its hypophosphorylated state, phospholamban is an inhibitor of SERCA2. Phosphorylation removes this inhibition and thus activates SERCA2, stimulating re-uptake of Ca^{2+} ions into the sarcoplasmic reticulum and increasing the rate of myofibrillar relaxation. Phosphorylation of troponin I stimulates dissociation of Ca^{2+} from troponin C, again increasing the rate of relaxation. The positive chronotropic effect of catecholamines is probably exerted at the level of the pacemaker and presumably also involves cAMP. The loss of b-adrenoceptor responsiveness in heart failure is due to increased inhibition of adenylyl cyclase by G_i and loss of cell surface b-adrenoceptors by an internalization pathway involving heightened activity of a b-adrenoceptor kinase and arrestin.

Receptor protein tyrosine kinases

Receptor protein tyrosine kinases (which include the insulin and the insulin-like growth factor 1 receptors) are a second group of transmembrane receptors. These possess an intracellular domain with a protein tyrosine kinase activity essential for signalling. The ventricular myocyte possesses receptors for both insulin and insulin-like growth factor 1 which mediate the stimulatory effects of these hormones on glucose uptake, protein synthesis and cell survival (Fig. 12). Activation of the insulin or insulin-like growth factor 1 receptor stimulates formation of another signalling molecule, phosphatidylinositol 3,4,5,-trisphosphate (**PtdInsP3**), which is formed in the membrane by phosphorylation of $PtdInsP_2$ by phosphatidylinositol 3-kinase (Fig. 12). This leads to the activation of the recently-identified protein kinase B (or Akt), which is intimately concerned with regulation of cell growth, protein synthesis, and cell survival (although its role in the ventricular myocyte has not yet been extensively investigated). However, activation of protein kinase B may account for the observed amelioration of heart failure by insulin-like growth factor 1.

Other signalling pathways

In addition to endothelin, the endothelial cells that line the coronary circulation and the endocardium produce nitric oxide, which induces the relaxation of smooth muscle (thus causing vasodilatation and increasing coronary blood flow) and is negatively inotropic. There is immense current interest in nitric oxide: see Chapter 15.1.1.2 for further information.

Positive and negative inotropes

The action of drugs on the myocardium can be due to an effect on the Ca^{2+} transient (upstream regulation) or on the sensitivity of the contractile proteins to Ca^{2+} (downstream regulation). No inotrope in general use in clinical practice increases the force of contraction by a direct effect on the myofibrils.

Upstream Ca^{2+} regulation—positive effects

The importance of increases in cytoplasmic Ca^{2+} concentrations and of b-adrenoceptor-mediated increases in cAMP concentrations in regulating myocardial contractility were described earlier. These two processes are the points of action of many useful drugs. Catecholamines (adrenaline and the pharmacological b-agonist isoprenaline) raise cAMP and protein kinase activity and are powerful positive inotropic drugs. Cyclic nucleotide phosphodiesterase inhibitors (for example caffeine, amrinone, and milrinone) raise cAMP concentrations by inhibiting its breakdown and thereby also activate protein kinase. Increased concentrations of cAMP

and increased protein kinase activity increase Ca^{2+} entry through L-type Ca^{2+} channels; relaxation is also augmented by phosphorylation of phospholamban and the ensuing activation of SERCA2, as described previously. The positive chronotropicity of these agents probably results from cAMP facilitating the opening of L-type Ca^{2+} channels and augmenting I_f in the conduction tissue.

Cardiac glycosides (for example digoxin) inhibit the Na^+,K^+-ATPase, preventing the extrusion of Na^+_i. This in turn inhibits Ca^{2+}_i extrusion through the Na^+,Ca^{2+} exchanger and may, when the cell is depolarized, augment Ca^{2+} entry by reversing the direction of the exchanger. Ca^{2+} may also be taken up in increased amounts by the sarcoplasmic reticulum, thereby increasing the cardiac Ca^{2+} pool and facilitating Ca^{2+}-induced Ca^{2+} release. The net effect is to increase the cytoplasmic concentration and availability of Ca^{2+} resulting in an increased force of contraction.

Upstream Ca^{2+} regulation—negative effects

In some circumstances, a decrease in myocardial contractility is desirable. This can be achieved with b-blockers that compete with catecholamines for occupancy of b-adrenoceptors. These lower cAMP and can exert an anti-ischaemic effect by lowering heart rate, increasing diastolic blood flow, and reducing myocardial contractility, thereby diminishing O_2 demand and improving O_2 delivery.

The 'calcium antagonists' (verapamil, nifedipine, diltiazem) inhibit Ca^{2+} entry through the L-type Ca^{2+} channel in both cardiac myocytes and vascular smooth muscle cells. They reduce myocardial contractility, relax smooth muscle, and reduce conduction in the sinoatrial and atrioventricular nodes. Therapeutically, their major effects are through their vasodilator activity (which increases blood flow and O_2 delivery to the myocardium).

Downstream Ca^{2+} regulation—positive effects

A group of drugs known as Ca^{2+}-sensitizing agents was initially believed to provide a fresh approach to the treatment of chronic/congestive heart failure. This heterogeneous group of positive inotropic agents mediate their effects by increasing the sensitivity of the contractile elements to Ca^{2+}, either by increasing the affinity of troponin C for Ca^{2+} or by direct effects on the actin–myosin complex. They were envisaged as being able to enhance myocardial contractility without changing the cytosolic Ca^{2+} concentration. Pimobendan, levosimendan, MCI-154, EMD-53998, and CGP-48506 have been studied as possible therapies for chronic/congestive heart failure: all have positive inotropic effects on isolated cardiac tissue, but their clinical usefulness has not yet been established.

Myocardial ischaemia

The three important consequences of myocardial ischaemia are failure of contraction, increased frequency of arrhythmias, and cell death. The general pathophysiology of this condition is discussed in Chapter 15.1.1.3 and the clinical aspects in Chapter 15.4.2.2 and Chapter 15.4.2.3: the following discussion is limited to the biochemical consequences of ischaemia for heart muscle.

Regulation of cardiac fuel and adenine nucleotide metabolism during hypoxia

In myocardial ischaemia the coronary blood flow to the affected area is insufficient to meet the demands for O_2 and fuels and to remove the products of metabolism. Maintenance of ATP regeneration is impossible, given the lack of O_2 supply and the fact that, given its high work output, the heart contains relatively little of the endogenous fuel polysaccharide glycogen. Anaerobic carbohydrate metabolism would have to increase about tenfold (not possible given the activity of the glycolytic pathway) and lactate would have to be removed (not possible in ischaemia). ATP concentrations eventually fall and contractile activity decreases. Two mechanisms operate in the short term to maintain ATP. First, ATP is buffered by the operation of the creatine phosphokinase equilibrium (phosphocreatine + ADP + H^+ ?2; creatine+ATP) which is driven to the right by proton production from glycolytic lactate. Second, the small amount of glycogen present is broken down (glycogenolysis) and glycolysis is increased.

In partial ischaemia, changes similar to those described above are also observed. Additionally, these hearts show increased accumulation of endogenous triglyceride droplets. These probably arise because, during ischaemia, the intermediates of fatty acid metabolism accumulate (for example long-chain fatty acid CoA and long-chain fatty acid carnitine, the long-chain fatty acid derivative which actually traverses the inner mitochondrial membrane). Combined with accumulation of glycerol-3-phosphate (formed from the reduction of the glycolytic intermediate 3-phosphoglyceraldehyde by accumulating NADH), this leads to synthesis of triglycerides. Long-chain fatty acid CoA and long-chain fatty acid carnitine are also powerful detergents. Their accumulation may therefore lead to disruption of membrane systems within the myocyte, to the detriment of cellular integrity.

Metabolic and mechanical consequences of ischaemia

Total ischaemia results in cessation of contraction within 60 s. Two important causes of the decline in contractility are the rapid development of intracellular acidosis through production of lactate and protons (see above), and an increase in intracellular concentrations of inorganic phosphate. The latter is caused by an inability to regenerate ATP from ADP and inorganic phosphate by fuel utilization. ATP concentrations are maintained in the short term (~ 60 s) because they are 'buffered' by creatine phosphate (see above). Thus, the inability to regenerate ATP by fuel utilization is reflected in a rapid decline in concentration of creatine phosphate. Both the fall in pH and the rise in inorganic phosphate decrease the maximum force that the myofilaments can produce, by shifting the Ca^{2+} sensitivity of the myofilaments to the right. The result is that higher concentrations of Ca^{2+} are required to produce the equivalent amount of force (Fig. 13). During ischaemia there is also a gradual increase in the resting (diastolic) level of cytoplasmic Ca^{2+}, probably because of a progressive failure of Ca^{2+} sequestration and efflux mechanisms. The effects of inorganic phosphate and acidosis would be larger if it were not for this increase in Ca^{2+} concentration. If ischaemia is prolonged, the sequestration and efflux mechanisms fail completely and this accounts for the cessation of the Ca^{2+} transient.

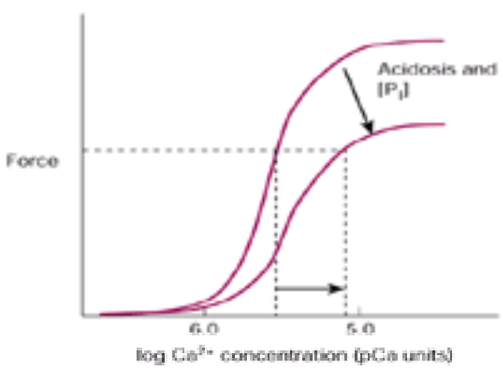

Fig. 13 Intracellular pH, inorganic phosphate, and myofilament Ca^{2+} sensitivity. The normal sigmoidal force/Ca^{2+} concentration relationship is shown by the line on the left. Acidosis or an increase in increased intracellular inorganic phosphate (P_i) decrease the sensitivity of myofilaments to Ca^{2+} and shift the relationship downwards and to the right. The maximum force produced is reduced.

In addition to intracellular buffering, there are other mechanisms to protect myocytes against acidosis. Two systems in the sarcolemma (the Na^+,H^+ exchanger and the Na^+,HCO_3^- symport) are activated by intracellular acidosis (Fig. 14). The Na^+,H^+ exchanger expels intracellular protons in exchange for extracellular Na^+ and is inhibited by the amiloride group of compounds. The Na^+,HCO_3^- symport transports HCO_3^- and Na^+ into the cell to buffer H^+.

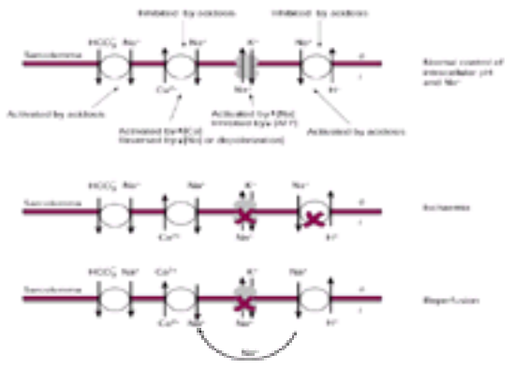

Fig. 14 Control of intracellular pH and Na^+ concentration. The membrane Na^+,H^+ exchanger and the Na^+,HCO_3^- symport counteract a fall in pH_i. Intracellular Na^+ concentration is controlled by the Na^+,K^+-ATPase and influences the direction of Ca^{2+} transport by the Na^+,Ca^{2+} exchanger. If ischaemia is prolonged, the decrease in extracellular pH inhibits the Na^+,H^+ exchanger and a decrease in intracellular ATP inhibits the Na^+,K^+-ATPase. On reperfusion, rapid normalization of extracellular pH reactivates the Na^+,H^+ exchanger and results in a large influx of Na^+ which cannot be expelled quickly owing to inhibition of the Na^+,K^+-ATPase. The increase in intracellular Na^+ concentration coupled with a depolarized membrane potential reverses the normal direction of the Na^+,Ca^{2+} exchange and mediates Ca^{2+} entry. Abbreviations: *o*, extracellular; *i*, intracellular.

Under normal resting conditions, the myocardium will recover almost completely after 10 to 15 min of ischaemia if adequate flow is restored. More prolonged periods of ischaemia cause the plasma membrane to become permeable to cations and recovery is reduced. If limited blood flow is present from collateral coronary arteries or from 'stuttering ischaemia' (periodic opening and closing of the native coronary artery), the onset of necrosis is delayed. After 60 to 90 min of ischaemia, the plasma membrane is destroyed. This may be attributable to low ATP concentrations, acidosis, activation of phospholipases, and lysosomal activity.

Reperfusion of ischaemic heart muscle results in further damage (reperfusion injury). This is characterized by an immediate swelling of the cell, release of intracellular enzymes (creatine phosphokinase, lactate dehydrogenase) and a large influx of Ca^{2+}. Ca^{2+} is taken up by the mitochondria and can be detected by electron microscopy as deposits of insoluble calcium phosphate. A large gain of Ca^{2+} is indicative of cell damage since it prevents the normal functioning of mitochondria and the regeneration of ATP. Many theories exist to explain the sudden influx of Ca^{2+}. A popular hypothesis is that the reintroduction of O_2 causes increased generation of reactive oxygen species, normal 'byproducts' of electron transport, which damage the plasma membrane through lipid peroxidation and render it permeable to Ca^{2+} in particular. The normal mechanisms within the cell for the removal of reactive oxygen species may be insufficient to protect against the sudden increased production of reactive oxygen species. Another hypothesis involves the Na^+,H^+ exchanger. If ischaemia is prolonged, there is a fall in extracellular pH that can often exceed the intracellular decline. Na^+,H^+ exchange is activated by intracellular acidosis, but is inhibited by extracellular acidosis. Thus, in prolonged ischaemia, the initial stimulation of the exchange will be followed by inhibition. On postischaemic reperfusion, there will be a rapid washout of extracellular protons that will reactivate the exchanger and result in a large influx of Na^+. The preceding ischaemia will have caused a decline in intracellular ATP so the Na^+,K^+ pump will be inhibited and thus the influx of Na^+ will lead to an increase in intracellular Na^+ concentration. As described earlier, the increase in intracellular Na^+ concentration coupled with a depolarized membrane potential reverses the normal direction of the Na^+,Ca^{2+} exchange and mediates excessive Ca^{2+} entry (Fig. 14).

Recovery from a period of ischaemia is slow. This is partly because the myocyte loses nucleotides. ATP is hydrolysed to ADP and operation of the adenylate kinase equilibrium (2ADP ?2; ATP + AMP) leads to AMP production. AMP is hydrolysed to adenosine (a vasodilator in its own right), but adenosine is rapidly broken down to inosine, thence oxidized to hypoxanthine and xanthine by xanthine oxidase, producing reactive oxygen species. Regeneration of nucleotides is slow and is the probable reason why, even if a cell does not die, total recovery is prolonged.

At present, treatments which are used in an attempt to reduce the size of a myocardial infarction act either by reducing ATP consumption (cardioplegic solutions, hypothermia, afterload reduction, negative inotropic agents) or by increasing coronary flow (afterload reduction, coronary vasodilators) through collaterals or the native coronary. The use of thrombolytic therapy in many patients with myocardial infarction reduces infarct size if the occlusion is due to thrombus, if the thrombus can be dissolved, and if the occlusion has not been present for more than 6 to 12 h. The only treatments that may benefit the ischaemic myocyte by mechanisms that act directly on the cell metabolism or cell structure are insulin, glucose and K^+ therapy, corticosteroids, and hyaluronidase.

Further reading

Bers DM (2001). *Excitation-contraction coupling and cardiac contractile force*, 2nd edn. Kluwer, Dordrecht.

Bolli R and Marban, E. (1999) Molecular and cellular mechanisms of myocardial stunning. *Physiological Reviews* **79**, 609–34.

Chien KR (1999). *Molecular basis of cardiovascular disease*. WB Saunders, Philadelphia.

Fozzard HA *et al.* (1991). *The heart and cardiovascular system. Scientific foundations*. Raven Press, New York.

Jennings RB, Steenbergen C Jr, Reimer KA (1995). Myocardial ischemia and reperfusion. *Monographs in Pathology* **37**, 47–80.

Kastor JA (1994). *Arrhythmias*. WB Saunders, Philadelphia.

Katz AM (1992). *Physiology of the heart*. Raven Press, New York.

Milnor WR (1990). *Cardiovascular physiology*. Oxford University Press, Oxford.

Nelson DL, Cox MM (2000). *Lehninger principles of biochemistry*. Worth, New York.

Newsholme EA, Leech AR (1983). *Biochemistry for the medical sciences*. Wiley, Chichester.

Opie LH (1995). *Drugs for the heart*. WB Saunders, Philadelphia.

Opie LH (1998). *The heart. Physiology, from cell to circulation*, 3rd edn. Lippincott Raven, Philadelphia.

Severs NJ (2000). The cardiac muscle cell. *BioEssays* **22**, 188–99.

Sheridan DJ (1998). *Left ventricular hypertrophy*. Churchill-Livingstone, London.

15.1.3.2Clinical physiology of the normal heart

D. E. L. Wilcken

Introduction

The function of the heart is to pump sufficient oxygenated blood containing nutrients, metabolites, and hormones to meet moment to moment metabolic needs and preserve a constant internal environment. The heart has two essential characteristics, contractility and rhythmicity. The nervous system and neurohumoral agents modulate relationships between the venous return to the heart, the outflow resistance against which it contracts, the frequency of contraction, and its inotropic state; there are also intrinsic cardiac autoregulatory mechanisms. This section describes normal cardiac function and discusses the principal mechanisms contributing to its regulation.

The cardiac cycle

Electrical events initiate the cardiac cycle with depolarization of the sinoatrial node in the upper right atrium near the orifice of the superior vena cava (Fig. 1). Cardiac muscle acts as a functional syncytium. Cell to cell conduction is possible because the intercalated discs offer a low electrical resistance. The action potential in an active cell causes current flow, which depolarizes the adjacent cells. The generated action potential spreads from the sinoatrial node across the functional syncytium at a speed of 1.0 to 1.2 m/s. The first mechanical response is atrial systole.

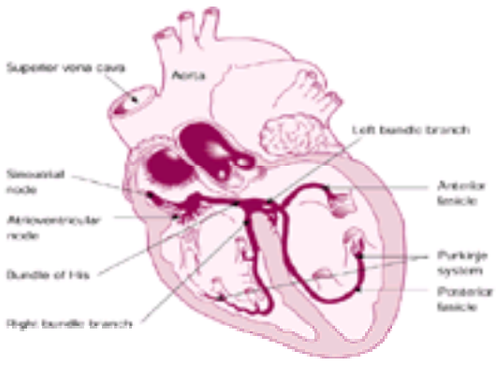

Fig. 1 Diagram of the heart showing the impulse-generating and impulse-conducting system. (Reproduced with permission from Junqueira LC, Carneiro J, Kelley RO (1998). *Basic histology*, 7th edn. Appleton and Lange, Norwalk, CT.)

The valvular attachments and connective tissue in the atrioventricular groove normally prevent cell to cell conduction of the electrical impulse from atrium to ventricle. This conduction occurs only through the specialized cells of the atrioventricular node (Fig. 1). The atrioventricular node is a region of slow conductance, from 0.02 to 0.1 m/s. This delays activation of the cells of the bundle of His and allows time for completion of ventricular filling. The conduction velocity in the bundle of His is from 1.2 to 2.0 m/s. The impulse passes via the right bundle branch and the two branches of the left bundle, and spreads rapidly (2.0 to 4.0 m/s) through the Purkinje fibres and each muscle cell to produce an orderly sequence of ventricular contraction (Fig. 1). Atrial and ventricular depolarization (P wave and a QRS complex) and repolarization (T wave) can be recorded on the electrocardiogram as the summation of the spread of the electrical potentials over all the cells of the heart (Fig. 2). Electrocardiography is considered in Chapter 15.3.2.

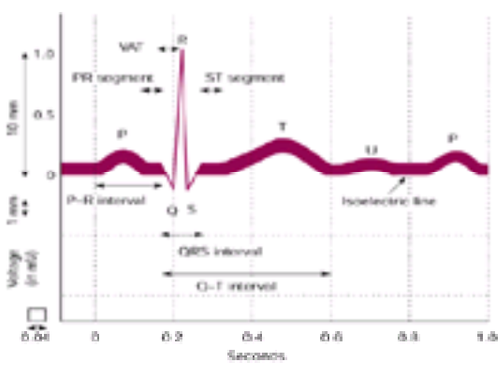

Fig. 2 Diagram of electrocardiographic complexes, intervals, and segments. VAT, ventricular activation time. (Reproduced with permission from Goldman MJ (1986). *Principles of clinical electrocardiography*, 12th edn. Lange, Los Altos, CA.)

The specialized cells of pacemaker tissue have an inherent rhythmicity which is shared by the sinoatrial node, the atrioventricular node, and Purkinje tissue. Unlike other myocardial cells these cells do not maintain a diastolic intracellular potential of about –90 mV but tend to depolarize spontaneously. Because the sinoatrial node has the fastest inherent discharge (depolarization) rate, and because there is a brief period after depolarization of the whole heart during which a further stimulus is ineffective—the absolute refractory period—the sinoatrial node is normally the pacesetter for the heart. However, if this does not occur, pacemaker tissue in the atrioventricular node, the bundle of His, or the Purkinje system, will assume this role. The heart rate is then considerably slower.

Mechanical events

The mechanical events following depolarization of the atrial and ventricular muscle and their timing in relation to the electrocardiogram, to pressure and flow changes, and to heart sounds are shown in five phases in Fig. 3. After the P wave, and coinciding with atrial systole, 'a' waves appear in left atrial and right atrial pressure tracings due to atrial contraction, and an 'a' wave can be seen in the jugular venous pulse. Atrial contraction increases ventricular filling by about 10 per cent (phase 1).

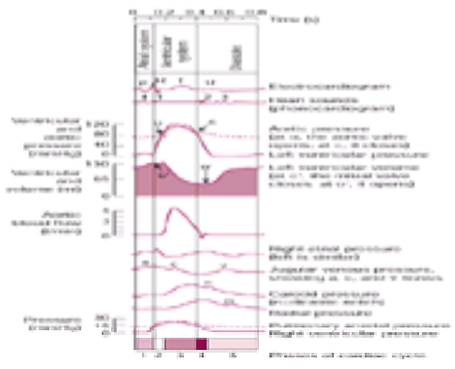

Fig. 3 Events of the cardiac cycle at a heart rate of 75 beats/min. The phases of the cardiac cycle identified by the numbers at the bottom are: (1) atrial systole; (2) isovolumetric ventricular contraction; (3) ventricular ejection; (4) isovolumetric ventricular relaxation; and (5) ventricular filling. Note that late in systole, aortic pressure actually exceeds left ventricular pressure. However, the momentum of the blood keeps it flowing out of the ventricle for a short time. The pressure relationships in the right ventricle and pulmonary artery are similar. The jugular venous pulse is similar in form to that seen in the right atrial pressure tracing. The 'c' wave interrupts the 'x' descent of the 'a' wave. The decline in pressure from the peak of the 'v' is the 'y' descent; the rate of decline reflects speed of ventricular filling. Atr. syst, atrial systole; ventric. syst, ventricular systole. (Modified with permission from Ganong WF (2001). *Review of medical physiology*, 20th edn. Appleton and Lange, Norwalk, CT.)

The onset of ventricular contraction coincides with the peak of the R wave of the electrocardiogram and there is a rapid rise in intraventricular pressure, which closes the mitral and tricuspid valves. The first heart sound is heard at the time of maximum displacement of these valves as they reach their closing positions. During this short isovolumetric period (phase 2 of Fig. 3) the pressure rises rapidly in the ventricle. When ventricular pressures exceed those in the pulmonary artery and aorta, the outflow valves open and ventricular ejection follows, with the highest flow rate occurring in early systole, and pressures in the aorta and pulmonary artery rise. Normally between 50 and 70 per cent of the ventricular volume is ejected during systole, and this can be seen in the volume curve included in Fig. 3 (phase 3).

The jugular venous pulse during ventricular contraction has a positive deflection in early systole, the 'c' wave, due to right ventricular contraction and bulging of the tricuspid valve into the right atrium. Descent of the tricuspid ring caused by ventricular contraction then produces a negative 'x' descent, but as atrial inflow continues the pressure rises in the atria and great veins, producing the 'v' wave. This reaches its peak just before the opening of the tricuspid valve, declining during early ventricular filling as the negative 'y' descent. The changes in the pulmonary veins and left atrium are similar.

As the strength of ventricular contraction declines, and coinciding with the end of the T wave of the electrocardiogram, the aortic and pulmonary valves close, producing the dicrotic notch seen on both aortic and pulmonary artery pressure tracings in Fig. 3. Aortic closure slightly precedes pulmonary closure, and together these are responsible for the two components of the second heart sound. A short period of further rapid decline in ventricular pressure ensues without change in the ventricular volume (the period of isovolumetric ventricular relaxation, phase 4) and at the end of this the mitral and tricuspid valves open. There is a pressure gradient from atrium to ventricle so that a period of rapid ventricular filling follows, which coincides with the timing of the third heart sound. The rapid ventricular filling is reflected in the shape of the ventricular volume curve, and is followed by a period of slower filling (phase 5) with a final sudden small increment from the next atrial contraction as diastole ends (phase 1).

Third heart sounds are audible with the stethoscope in normal children and young adults, but over the age of about 40 years this usually indicates elevation of ventricular end-diastolic pressure (most frequently in the left ventricle). This is probably because the myocardium and valvular structures become stiffer with ageing, and large increases in ventricular end-diastolic pressure are then required to tense valvular structures and generate audible vibrations. The hearing of a fourth heart sound almost always indicates abnormal ventricular function. The end-diastolic pressure in the affected ventricle (usually the left) is increased, and the already stretched inflow valve responds to atrial systole and further filling with oscillations, producing a low-pitched sound that is often palpable as well as audible at the cardiac apex. A fourth heart sound precedes the Q wave of the electrocardiogram and must be distinguished from a normal splitting of the two components of the first heart sound. The latter occurs after the Q wave (Fig. 2 and Fig. 3).

Normal volumes, pressures, and flows

The blood volume in normal adults is about 5 litres (haematocrit 45 per cent), and of this about 1.5 litres are in the heart and lungs—the central blood volume. The pulmonary arteries, capillaries, and veins contain about 0.9 litres, with only about 75 ml being in the pulmonary capillaries at any one instant. The volume of blood in the heart is about 0.6 litres. Left ventricular end-diastolic volume is about 140 ml, the stroke volume about 90 ml, so that the end-systolic volume is around 50 ml, and the ejection fraction (stroke volume/end-diastolic volume) is between 50 and 70 per cent. The right ventricular ejection fraction is similar.

Of the 3.5 litres in the systemic circulation most, at least 60 per cent of the total blood volume is in the veins. The term 'mean circulatory pressure' introduced by Guyton is useful and refers to the equilibrium pressure measured in the entire circulation within a few seconds of stopping the heart; in dogs this is about 7 mmHg. The systemic veins containing most of the blood volume are easily distensible, and input of blood into the contracting heart is associated with only small changes in venous pressure. By contrast, ejection of blood into the much less distensible arterial tree produces large pressure changes.

The normal values for pressures generated in the heart and great vessels during the cardiac cycle are shown in Table 1. Pressures are measured with reference to a zero pressure arbitrarily set at 5 cm below the sternal angle with the patient recumbent. 'Normal' arterial blood pressure is considered later (see below).

P>Cardiac output is the product of stroke volume and heart rate. It is related to body size and is best expressed as litre/min/m ^{2} of body surface area: the 'cardiac index'. The mean cardiac index under resting and relaxed conditions is 3.5 litre/min/m ^{2}, and values below 2 and above 5 are abnormal. The cardiac index declines with age. In persons of average size, resting oxygen consumption is about 240 ml/min, and the difference in oxygen content between arterial and mixed venous blood is about 40 ml/litre (arteriovenous oxygen difference), giving a basal cardiac output of 6 litre/min from the direct Fick equation. In normal subjects the arteriovenous difference in oxygen content at rest is maintained within narrow limits, from 35 to 45 ml/litre; values of 55 ml/litre and above are always abnormal.

Pulmonary or systemic vascular resistance is estimated by dividing the difference between mean inflow pressure (pulmonary artery or aortic) and mean outflow pressure (left atrial or right atrial) in mmHg by the flow in litre/min through the respective circulations. In normal subjects and patients without intracardiac shunts this flow is the cardiac output. Normal pulmonary vascular resistance is less than 2 mmHg/litre/min (160 dyn/s/cm ^{5}). Arterial blood pressure is the product of cardiac output and total peripheral resistance.

Stroke work is the integral of instantaneous ventricular pressure with respect to stroke volume, but is usually estimated as the product of stroke volume and mean ejection pressure. The orderly sequence of contraction in the normal cardiac cycle co-ordinates changes in instantaneous pressure and flow, so maximizing the transfer of energy to the circulation. Normal left ventricular work output at rest is about 6 kg/m ^{2}/min.

Myocardial mechanics

A more rational approach to the understanding of cardiac muscle contraction and altered performance in disease states has come from renewed interest in the results of classic experiments in skeletal muscle physiology. The three-component model for muscular contraction proposed by Hill in 1938 (Fig. 4) comprises, first, a contractile element which, when activated, develops force and shortens; second, a series elastic element that is stretched passively during shortening and produces a

dampening effect; and, third, a parallel elastic element which supports resting tension. The latter, together with the series elastic element, is responsible for the extensibility or compliance of relaxed muscle. It is not known which precise structures are responsible for the series elastic and parallel elastic components, but there is no doubt about their functional significance.

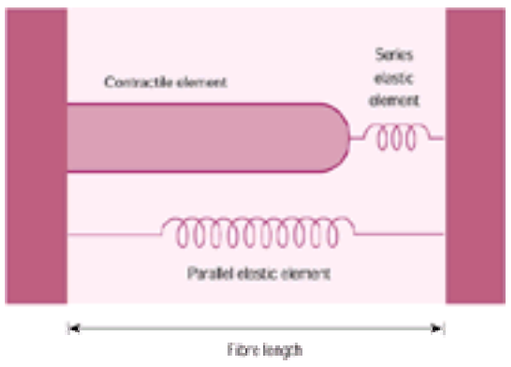

Fig. 4 A representation of the model used by A. V. Hill to illustrate the three mechanical components of functioning muscle.

When a muscle is activated to contract, it develops a potential for doing work. In isolated skeletal and heart muscle preparations the stretching force applied to the muscle, and therefore the length of the muscle, can be varied before contraction; this is the preload. The activated muscle will begin to shorten when it has generated a force sufficient to overcome that exerted by the attached weight or load against which it contracts. When the force exerted by the load is so arranged that it is not applied to the relaxed muscle and is applied only after the muscle has begun to develop tension it is termed the afterload. If this load is so large that the activated muscle is unable to overcome it, and so cannot shorten, the contraction produces tension only, and the contraction is isometric. When shortening does occur, external work is done. If the load is constant during the shortening, the contraction is said to be isotonic; if it changes it is auxotonic.

The tension produced by both skeletal and cardiac muscle during contraction depends on initial fibre length; during afterloaded isotonic contractions from a particular length, the amount and the speed of fibre shortening and the tension developed all depend upon the afterload. Over a range of loads the initial velocity of muscle shortening is most rapid and the most extensive shortening occurs when the load is smallest.

The inverse relationship between initial velocity of fibre shortening and load in an isotonic contraction is a fundamental one for both skeletal and cardiac muscle (Fig. 5). There is, however, a major difference between the two types of muscle in that the relationship at any one length is constant in a skeletal muscle, whereas in cardiac muscle there are variations in inotropic state that are accompanied by considerable changes in the relationship between force and velocity. A positive inotropic effect produces a more extensive contraction from the same initial length and afterload, and a faster maximum velocity of shortening (V_{max}). An increase in initial fibre length with no increase in inotropic state increases the force of contraction but does not, however, change the maximum velocity of shortening. This is illustrated in Fig. 5.

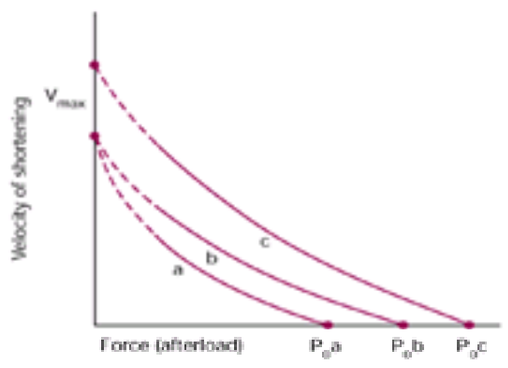

Fig. 5 Idealized relationships between velocity of fibre shortening and afterload or force developed during contraction of a strip of cardiac muscle under three different conditions. Curves a and b were obtained with the muscle in the same inotropic state but with a longer initial fibre length (greater preload) for curve b. Curves b and c were obtained with initial fibre length the same but with contractility increased in c by the addition of a drug producing a positive inotropic effect. The terms V_{max} and P_0 were used by Hill to describe, respectively, a hypothetical maximum shortening velocity in the absence of any load (hence the broken lines), and the force developed in an isometric contraction. An increase in initial fibre length increases P_0 but not V_{max}; a positive inotropic change increases both P_0 and V_{max}.

The contraction of the intact heart can be visualized as being similar mechanically to the afterloaded contraction of an isolated muscle strip. For the left ventricle, the preload is the distending force which stretches the muscle fibres in end-diastole, and the initial afterload is the force the ventricle must generate in order to open the aortic valve and eject blood. At the end of ejection, the ventricular muscle is isolated from the peripheral circulation, with the afterload then supported by the competent aortic valve, and the muscle relaxes against a comparatively small force. Relaxation of the heart is an active process due to withdrawal of calcium ions from the cytoplasm surrounding the myofibrils. 'Active' relaxation is still proceeding in the ventricular wall when the atrioventricular valves open, and, if it is delayed, as in the hypoxic heart, the slower relaxation increases the stiffness of the ventricular wall and reduces filling. Wall thickness is also a determinant of relaxation rate and compliance. For this reason filling pressures are higher for the thicker and stiffer left ventricle than for the thinner and more distensible right ventricle (Table 1). When the left ventricle is hypertrophied due to chronic pressure overload, as in systemic hypertension or aortic stenosis, it becomes stiffer and filling pressures may then be abnormally high.

Regulation of cardiac function

Four essential factors determine the performance of the heart:

1. venous return;
2. outflow resistance (afterload);
3. inotropic state or contractility;
4. heart rate.

Changes in cardiac performance are accomplished by mechanisms that alter these four determinants.

Venous return, preload, and the Frank–Starling relationship

The relationship described independently by Frank and Starling between end-diastolic fibre length and force of contraction is shown in Fig. 6. When the right or left ventricle ejects against a constant pressure, variations in venous return alter the degree of stretch of the muscle fibres in diastole, and this determines contraction strength and work output. The number of active force-generating sites in each fibre increases as it lengthens so that, within limits, the force of contraction and stroke work are positively related to end-diastolic fibre length. The relationship is curvilinear when stroke work is plotted against end-diastolic pressure as an index of preload, reflecting the exponential relationship between end-diastolic pressure and end-diastolic volume. When stroke work is plotted against end-diastolic volume the relationship between stroke work and preload is linear.

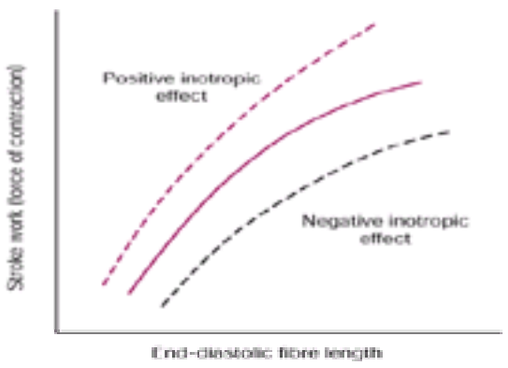

Fig. 6 The relation between left ventricular end diastolic fibre length and left ventricular stroke work showing displacement upward and to the left with an increase in contractility and downward and to the right with a reduction in contractility. Similar but not identical curves are obtained by plotting left ventricular stroke work as one measure of the force of contraction against ventricular end-diastolic pressure or volume (see text). Similar function curves may be obtained from both ventricles and both atria.

The response of the heart at any particular time depends upon:

1. The intrinsic state of the muscle, i.e. the nature of its own biochemistry and contractile machinery.
2. The prevailing neurohumoral state, i.e. increased sympathetic outflow produces a more forceful contraction at any end-diastolic fibre length and shifts the curve upward and to the left.
3. Extrinsic inotropic influences; drugs which have a positive inotropic effect also shift the curve upward and to the left, whereas myocardial depressants have a negative inotropic effect and shift the curve downward and to the right.

End-diastolic fibre length is determined by the force distending the ventricle at end-diastole, and end-diastolic pressure provides a reasonable indication of this force when the ventricle has normal distensibility or compliance; this is the preload. The systemic venous return and the elastic properties of the myocardium produce the end-diastolic distending pressure for the right ventricle, and the pulmonary venous return and myocardial elasticity that for the left ventricle. For clinical purposes it is convenient to equate venous return with preload because, as it changes from beat to beat, it adjusts the strength of the subsequent ventricular (and atrial) contraction by varying the force stretching the relaxed cardiac muscle and changing end-diastolic fibre length.

Outflow resistance or afterload

The pressure which the ventricle must develop to exceed that in the pulmonary artery and the aorta and open the pulmonary and aortic valves is determined largely by the pulmonary and systemic vascular resistances, as shown for the latter in Fig. 7. These resistances, together with an inertial component dependent upon the mass of blood within the vessels, the compliance (stiffness) of the vessels, and the physical characteristics of each vascular tree combined with the pulsatile nature of the flow, constitute the impedance to ventricular outflow. This is the load against which the ventricle must contract and shorten. As this load is not applied in diastole to the relaxed muscle, it then being supported by competent aortic and pulmonary valves, it is usefully described clinically as the afterload; it becomes applied to the muscle only after the ventricle has begun to develop tension.

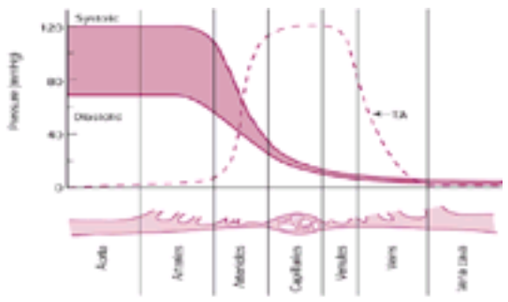

Fig. 7 Diagram of the changes in pressure as blood flows through the systemic circulation. TA, total cross-sectional area of the vessels. This increases from 4.5 cm ^{2} to 4500 cm^2 in the capillaries. The major resistance to flow is at the arteriolar level. (Modified and reproduced with permission from Ganong WF (2001). *Review of medical physiology*, 20th edn. Appleton and Lange, Norwalk, CT.)

Regulation of systemic arterial blood pressure

The regulation of the systemic circulation is well adapted to the vital function of maintaining constant, adequate cerebral perfusion. There is a need to maintain a relatively constant arterial blood pressure when there are changes in posture and circulating blood volume. The baroreceptors mediate rapid responses to alterations in aortic pressure, whilst a variety of hormonal and physical factors regulate the circulating blood volume.

Baroreceptors

The baroreceptor regulatory system comprises two groups of stretch receptors: one group in the carotid sinuses near the bifurcations of the common carotid arteries in the neck and a second group in the arch of the aorta. These respond to an increase in central arterial pressure by the firing of impulses, which pass by the glossopharyngeal and vagus nerves to the solitary tract nucleus in the medulla and inhibit sympathetic outflow. Efferent impulses from these central connections pass via the right vagus nerve mainly to the sinoatrial node, and via the left vagus mainly to the atrioventricular node. The effect is to decrease the heart rate and the force of atrial contraction. There is also attenuation of sympathetic discharge via the thororacolumbar sympathetic outflow to arteriolar smooth muscle in the limbs and visceral circulation, resulting in a release of peripheral arteriolar constriction and therefore peripheral vasodilatation. Thus the immediate response to a rise in arterial pressure is slowing of the heart rate, reduced force of atrial contraction, and reduced vascular resistance. The net effect of this negative feedback system is to offset the elevation in blood pressure. Conversely a lowering of blood pressure diminishes stimulation of the stretch receptors and reduces afferent traffic to the solitary tract nucleus resulting in reduced inhibition of sympathetic outflow. As a consequence there is a quickening of the heart rate and peripheral vasoconstriction so that the blood pressure increases. The changes in heart rate take place within 1 to 2 s and changes in vasomotor control within 5 or 6 s.

Baroreceptor mechanisms effectively modulate the responses of blood pressure to postural change. Additionally they adapt to maintain the normal circadian variation in blood pressure (see below). They also maintain elevated arterial blood pressure in systemic hypertension. Sensory input to the reflex is reduced in disorders of the autonomic nervous system, and in the prolonged weightlessness of space flight.

Blood volume

The circulating blood volume is relatively small and a large proportion is contained in the veins (Fig. 7) so that any change in blood volume will affect venous return and therefore cardiac output and blood pressure. When blood volume is large and the veins full there is little reduction in venous return on standing and cardiac output is maintained. However, when effective blood volume is reduced and the veins are relatively empty, on standing there is pooling of blood in the veins of the legs and a reduction in venous return and cardiac output so that arterial blood pressure falls. Baroreceptor responses become evident within a couple of beats, the

heart rate increases, and cardiac output and blood pressure are restored. Circulating blood volume is kept relatively constant by a combination of mechanisms which involve the actions of natriuretic peptides, the renin–angiotensin–aldosterone system, vasopressin, and osmolality.

The natriuretic peptides

The discovery of secretory granules in the atria of the heart and the demonstration in 1981 that they produce a natriuretic factor that inhibits the reabsorption of sodium in the distal tubule of the kidney enhanced understanding of the regulation of blood volume and cardiac performance. Three natriuretic peptides have subsequently been identified.

Atrial natriuretic peptide is present in the circulation and concentrations increase during volume expansion. The right atrium contains about two to four times as much activity as the left, and release of the hormone is mediated largely by atrial distension. The effect is to produce a diuresis and to reduce cardiac and circulating blood volume. Atrial natriuretic peptide also has a vasodilator action and opposes the vasoconstricting effects of noradrenaline and angiotension II.

The second natriuretic peptide was identified in brain tissue, and is referred to as brain natriuretic peptide. Large amounts were later shown to be in the ventricles of the heart and circulating levels are increased in venticular hypertrophy and cardiac failure. Brain and atrial natriuretic peptides have similar actions. The third to be identified was C-type natriuretic peptide. It is distributed widely in tissues, circulating concentrations are low, and it appears also to have actions similar to the other two peptides, but with a greater vasodilator effect on veins.

Thus these three peptides contribute to the regulation of cardiac and circulating blood volume and of blood pressure. The therapeutic potential of manipulating the effects of these peptides is currently being assessed.

The renin–angiotensin system

This system, which is both local and systemic, is of major importance in the regulation of circulating blood volume and the maintenance of normal blood pressure. Enhanced activity of systemic renin and angiotensin increases the production of aldosterone, which promotes reabsorption of sodium by the kidney and expansion of circulating blood volume. All components of the renin–angiotensin system are distributed widely throughout tissues, including the brain and the heart, and increased activation of the system increases the risk of cardiovascular events. Angiotensin II is a potent vasoconstrictor that also enhances the proliferation of smooth muscle cells. The angiotensin converting enzyme inhibitors in clinical use diminish angiotension II production locally and in the circulating blood. Both local and general effects appear important in mediating the benefits that accrue from the use of these drugs in the management of hypertension and congestive cardiac failure, and in the reduction in rates of recurrence of coronary events in ischaemic heart disease. The mechanisms mediating this latter effect in particular await clarification. It is yet to be determined whether the use of the more recently developed angiotensin II receptor blocking drugs will result in similar outcomes.

Ventricular volume and afterload

Ventricular volume also has a major effect on afterload, as pressure is equal to force per unit area. The force acting radially on the inner surface of the whole ventricle at any time during systole is the product of the intraventricular pressure and ventricular surface area at that time. If the left ventricle is assumed to be a sphere (surface area = pd^2), the force opposing ejection at any time during contraction is the product of the intracavity pressure and p d^2 at that time. Thus, a change in left ventricular diameter from a normal value of 5 cm to one of 10 cm would result in a fourfold increase in the force opposing ejection for the same intracavity systolic pressure; the ventricle would need to develop greatly increased wall tension to overcome that force. Because wall tension developed during systole is the major determinant of myocardial oxygen consumption, the contraction will clearly be much less efficient in the larger heart for the same stroke volume and ejection pressure (stroke work).

During a normal heartbeat the afterload is greatest at the beginning of ejection (rapid rise in pressure and maximum volume; Fig. 3), but thereafter decreases as the pressure reaches a plateau and then declines as the ventricle becomes smaller. There is therefore a matching of the afterload to the declining intensity of the contraction as it proceeds to completion, and fibres shorten at a relatively constant rate. This is less obvious in a large heart where the volume change during ejection is a smaller proportion of the total ventricular volume.

The end-diastolic volume is influenced by preload, afterload, circulating blood volume, the inotropic state of the ventricle, heart rate, and neurohumoral influences. For example, it is smaller in the erect than in the horizontal position because of reduced venous return, and it decreases with a moderate increase in heart rate because of an associated positive inotropic effect. The proportion of end-diastolic volume ejected during systole, the ejection fraction (normal 50 to 70 per cent), is a useful index of overall left ventricular function and is easily measured non-invasively by gated blood pool scanning and two-dimensional echocardiographic techniques. The ejection fraction increases with exercise and with positive inotropic interventions. Values for right ventricular ejection fraction are of the same order as those for the left side of the heart.

Myocardial contractility and inotropic state

Myocardial function is greatly altered by changes in inotropic state or contractility. Positive inotropic effects are thought to be mediated by activation of excitation–contraction coupling mechanisms and are associated with an increased influx of calcium ions into myocardial cells and a more powerful contraction. Changes in the intensity of excitation–contraction coupling are independent of the Frank–Starling mechanism. Increases in the intensity shift the curve upwards and to the left and decreases shift it downwards and to the right (Fig. 6). With a positive inotropic effect, the force of contraction, however measured, is increased for a given end-diastolic fibre length and, if the afterload is the same, the initial velocity of fibre shortening is also increased (Fig. 5); in the intact heart, there is more complete emptying during systole. Increased sympathetic stimulation, some drugs, and an increase in heart rate itself (the staircase or Bowditch phenomenon; postectopic potentiation, see below) have positive inotropic effects. Myocardial depressants, such as hypoxia and most anaesthetic drugs, have negative inotropic effects. Increased parasympathetic stimulation produces acetylcholine-mediated negative inotropic effects that are confined almost entirely to the atria because of the anatomical distribution of vagal endings in the myocardium.

It is difficult to measure inotropic changes accurately in the human heart because changes in the intensity of excitation–contraction coupling and changes in the Frank–Starling relationship, though separate, are nevertheless closely linked. Whilst Hill's classic model (Fig. 4) has been important conceptually, attempts to define contractility as predicted by the model—by deriving an extrapolated maximum velocity of fibre shortening which would obtain with the muscle contracting against zero load—have not been rewarding. The peak rate of change of intraventricular pressure (peak d *p*/d*t*) is a useful index of change in contractility provided that preload, afterload, and heart rate remain constant.

An approach that appears relatively insensitive to changes in both preload and afterload is that of Suga and Sagawa, using the ventricular pressure–volume loop diagram. There is an approximately linear relationship between end-systolic pressure (or wall stress) and end-systolic volume when measured over a narrow physiological range in the human left ventricle. Increased contractility shifts the relationship upward and to the left, as illustrated in Fig. 8, allowing the separation of enhanced from reduced contractility in the same heart, and poorly contracting from normally contracting ventricles. Stroke volume is shown on the abscissa as the difference between end-diastolic and end-systolic volumes. The efficacy of reduction in afterload in assisting reduced ventricular function is also easily explained from the diagram. With a reduced afterload, the aortic valve opens at a lower pressure and a greater stroke volume is ejected; a new end-systolic pressure–volume point is reached, which is shifted downwards on the same linear relationship. There has been no change in contractile state.

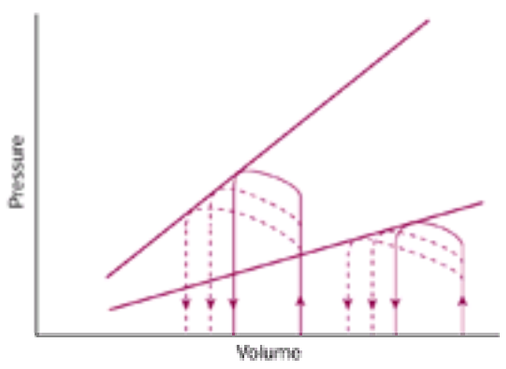

Fig. 8 Diagrammatic representation of intraventricular pressure and volume relationships during the cardiac cycle at two levels of myocardial contractility; three separate beats with the same end-diastolic volume are shown for each. The loops on the left of the diagram were obtained when contractility is increased and those on the right when it is reduced. There is a linear end-systolic pressure–volume relationship with different afterloads (pressures) for each level of contractility. The slope of the end-systolic pressure–volume relationship for any inotropic state is relatively insensitive (see text) to changes within physiological ranges in afterload and preload, although changes in preload are not shown in this diagram. The volume change seen on the horizontal axis for each beat is the stroke volume. This increases with reduction in pressure (afterload).

Heart rate

Frequency of contraction is the fourth essential determinant of cardiac performance. Heart rate during rest and exertion may vary from 45 to 200 beats/min in the healthy young adult. As changes can occur within seconds, an increase in heart rate is the usual and most effective way of producing a rapid increase in cardiac output. It plays the major role in the response to exercise, during which stroke volume does increase (more so in athletes and when in the erect rather than the supine position) but the changes are less marked than those of rate. In addition, an increase in contraction frequency itself produces a positive inotropic effect, whereby the force of contraction increases and reaches a new steady state within a few beats. This is termed the 'positive staircase', Treppe, or Bowditch effect. It may be a consequence of an augmented movement of calcium ions into myocardial cells with increased frequency of action potentials, combined with diminished time for outward movement of calcium between beats. More forceful contractions also follow premature beats—the phenomenon of postextrasystolic potentiation—and the mechanism is probably the same. The extrasystole occurring prematurely is a weak contraction because of decreased filling time and an unco-ordinated activation of the ventricle when the ectopic focus is within the ventricle. The next beat is delayed because of the refractory period of the extrasystolic beat, but is a more powerful contraction because of increased filling time and ventricular volume, and increased contractility. Calcium-dependent changes similar to those of the Bowditch effect are probably responsible for the latter.

Coronary blood flow

Coronary blood flow accounts for about 4 per cent of the cardiac output. The heart extracts most (70 per cent) of the oxygen carried in the coronary circulation; the arteriovenous difference for oxygen across the heart being about 110 ml/litre, whilst that for the whole body is only about 40 ml/litre under resting conditions. Therefore, large increases in myocardial oxygen requirements must be met largely by increases in coronary blood flow, and this may increase five- or sixfold during strenuous exercise. The greater part of this flow is to the left ventricle, of which at least two-thirds occurs during diastole because of the throttling effect systole has on myocardial perfusion. The main coronary arteries are on the superficial surface of the heart, and because of this, and the hindrance to coronary flow during systole, the subendocardial region of the left ventricle is more vulnerable to perfusion deficits in relation to oxygen need than the outer two-thirds of the muscle wall. Despite these mechanical problems, flow is normally evenly distributed throughout the myocardium so that, when regional coronary blood flow is measured using injected radioactive microspheres (in dogs), the ratio of endocardial to epicardial flow is approximately unity. In fact the inner layers of the heart probably receive slightly more blood (up to 10 per cent) than the outer layers. This is consistent with the subendocardium developing more tension than the subepicardium, and is evidence for a greater rate of myocardial oxygen consumption in the inner layers.

Myocardial oxygen requirements and coronary blood flow are finely adjusted. The regulation of coronary blood flow is described elsewhere.

The nervous system and the heart

The heart is richly supplied with adrenergic nerves. Terminals reach atrial and ventricular muscle fibres and impinge upon all pacemaker tissue including the sinoatrial and atrioventricular nodes and Purkinje fibres. Sympathetic stimulation leads to an increase in myocardial contractility and heart rate, and in the rate of spread of the activation wave through the atrioventricular node and the Purkinje system. This is mediated by local noradrenaline release, which interacts with b-adrenergic receptors. The key elements in these regulatory mechanisms are calcium ions and cyclic AMP. The activated b-receptor increases adenylcyclase activity and the conversion of ATP to cyclic AMP. Peptide cotransmitters released with noradrenaline and acetylcholine have recently been isolated and also influence autonomic function. Neuropeptide Y is a peptide of 36 amino acids that is collocated with noradrenaline in most sympathetic nerves and is released with sympathetic stimulation. It is a powerful pressor agent with direct arteriolar vasoconstrictor action and also potentiates the pressor action of noradrenaline.

The distribution of parasympathetic fibres is much more limited, being confined to the sinoatrial and atrioventricular nodes and the atria, with few if any fibres reaching the ventricles in humans, except perhaps in relation to coronary arteries and Purkinje tissue. The effects of parasympathetic nerve stimulation are mediated by local acetylcholine release, which slows the heart rate and speed of conduction through the atrioventricular node and Purkinje tissue, and depresses atrial contractility. The negative inotropic effects are associated with a lowering of the concentration of intracellular cyclic AMP.

The effect of the nervous system on the heart at any one time is the sum of the activities of these two opposing control systems. They usually vary reciprocally. Under resting conditions, vagal inhibitory effects predominate, maintaining a slow heart rate, there being virtually no sympathetic outflow. With exercise, there is withdrawal of vagal activity and an increase in sympathetic outflow. Afferents from stretch receptors in the carotid sinus and aortic arch—the baroreceptors—also have a considerable effect on cardiac performance, this effect being mediated via the adrenergic nervous system and vagal withdrawal. A fall in blood pressure reduces stretching in the carotid sinus and inhibitory afferent traffic so that the sympathetic outflow increases. As a consequence of this combined vagal and adrenergic effect there is a quickening of the heart rate within one or two beats, a positive inotropic effect, and also a constriction of veins and arterioles that increases preload and afterload. Elevation of pressure in the carotid sinus has the reverse effects. In cardiac failure there is a reduced variability in heart rate due to these autonomic mechanisms as there is then a predominance of adrenergic activity.

There are also mechanoreceptors in all four chambers of the heart (in dogs) and in the coronary vessels, which give rise to depressor reflexes. Their clinical relevance is uncertain, but they may contribute, for example, to the bradycardia and hypotension occurring in some patients with acute myocardial infarction and to the syncope that patients with critical aortic stenosis may experience with the onset of exercise when there is sudden left ventricular distension. Vagal afferents from reflexogenic areas in the infarcting left ventricle may be responsible for the bradycardia, gastric distension, nausea, and vomiting which frequently occur with the onset of inferior or posterior myocardial infarction, but not usually of anterior infarction, which is generally associated with a marked increase in sympathetic activity. The cardiac receptors connected to afferent fibres running in cardiac sympathetic nerves, however, are very important because they are responsible for the perception of cardiac pain. Receptors have also been identified (in animals) at the junction of pulmonary veins with the atrial wall. These respond to mechanical distension with increased sympathetic outflow to the sinus node and inhibition of secretion of antidiuretic hormone from the posterior lobe of the pituitary gland. The result is a quickening of the heart rate and diuresis, effects that could contribute to the regulation of cardiac volume.

Autonomic efferent activity

The autonomic outflow to the heart is controlled by multiple integrative sites within the central nervous system, with complex interactions between afferent and central inputs. Autonomic responses are mediated through the suprapontine and bulbospinal pathways, both those arising 'reflexly' and those arising from various types of volitional or central 'command'. Nevertheless, intrinsic mechanisms are sufficient for adequate cardiac function in the absence of autonomic control, as prolonged survival after cardiac transplantation has shown. But in the denervated heart there is blunting of the normally rapid physiological adjustments mediated by the autonomic nervous system.

Diurnal variation in autonomic function

Variations in vascular tone and control of blood pressure and of hormone secretion and platelet function occur in a predictable way throughout the 24-h cycle. In normal subjects there is a circadian rhythm of blood pressure changes that is not seen in patients after cardiac transplantation who have denervated hearts. There is a decline in blood pressure at night and an increase soon after wakening. This is due to a normal adrenergic surge in the early morning, which results in increased vascular tone and blood pressure. Increased forearm vascular resistance in the morning with a reduction in the afternoon and evening can be clearly identified in humans by assessing responses to a-adrenergic blockade. It is presumed that this occurs in coronary vessels as well. Measurable early morning increases in circulating catecholamines and in the propensity for platelets to aggregate can also be documented.

The circadian rhythm of autonomic function is correlated with a significant tendency for myocardial infarction and sudden cardiac death to occur more frequently in the morning soon after wakening. There is also an increase in the occurrence of angina pectoris in the early morning, independent of the level of physical activity.

Exercise and the heart: cardiac reserve

The heart responds to exercise with an increase in cardiac output, and values of 30 litre/min may be achieved in a trained athlete. Exercising muscles extract more oxygen from the blood perfusing them, but the response of the cardiac output is the ultimate determinant of delivery of oxygen to tissues and is the limiting factor for aerobic exercise.

The cardiac response to exercise involves all the mechanisms already discussed. Interaction within the central nervous system between higher and autonomic centres augments sympathetic discharge and there is a withdrawal of parasympathetic outflow. The heart rate increases immediately, and redistribution of peripheral flow increases venous return and preload. There is venoconstriction, particularly in the large-volume splanchnic circulation, and vasoconstriction and increased oxygen extraction in non-active parts. In active parts there is vasodilation. This is most evident in the vascular beds of the exercising skeletal muscles and of the heart. The overall effect is a marked lowering of total peripheral vascular resistance, which reduces afterload and encourages greater systolic emptying of the left ventricle. Stroke volume increases during exercise in the upright position. During light to moderate exercise (running or cycling), up to about 80 per cent of maximum exercise capacity, there is an almost linear relationship between work intensity and heart rate response, cardiac output, and oxygen uptake. With further exercise the heart rate and cardiac output responses level off whilst additional increases in oxygen consumption (about 500 ml) occur by increased oxygen extraction and a greater widening of the arteriovenous difference for oxygen.

The venous return increases in relation to the elevated cardiac output. Vasodilation in the working muscles that receive the bulk of the redirected blood permits high flow rates into the capacitance vessels. Because of adrenergically mediated venoconstriction the capacity of this system is reduced, so that blood moves rapidly into the right atrium. Venous return is also enhanced by the pumping action of the rhythmically contracting working muscles, by a decrease in intrathoracic pressure with forced inspiration, and by an increase in intra-abdominal pressure. The augmented pulmonary blood flow results in only slight increases in pulmonary artery pressure because of the distensibility of the large pulmonary arteries, an increased area of the pulmonary capillary bed due to the recruitment of more capillaries, and the low resistance offered by the normal pulmonary circulation (see Table 1).

The elevated cardiac output and larger stroke volume result in increased systolic blood pressure and pulse pressure even though the afterload itself is reduced. Enhanced neurohumoral activity from adrenergic stimulation of the heart and the suprarenal glands (increased circulating adrenaline and noradrenaline) effect positive inotropic changes, to which tachycardia also contributes because of the Bowditch effect. There is a shift in the Frank–Starling relationship to the left, increased speed and force of cardiac contraction, and elevated ejection fraction and stroke volume. Peak d p/dt is increased and there is a rapid rise in coronary blood flow to meet myocardial oxygen requirements that increase linearly with the product of systolic blood pressure and heart rate. During moderate exercise these changes together result in a decreased or unaltered end-diastolic volume and decreased end-systolic volume. With severe exercise, end-diastolic dimensions and end-diastolic fibre length are slightly increased and the Frank–Starling mechanism then operates and further augments the force of contraction.

The haemodynamic and ventilatory responses evoked by an increase to a new steady workload take about 2 to 3 min to equilibrate and adjust oxygen supply to the greater demand. Protocols for exercise testing are therefore usually based on work increments at 3-min intervals to allow time for a new 'steady state' to occur as, for example, in the standard Bruce Exercise Protocol. A steady state becomes progressively more difficult to maintain as maximal exercise capacity is approached. Glycogen is used by the working skeletal muscles as a source of stored energy and the anaerobic metabolism which ensues produces lactic acidosis and thereby further increases ventilation. As all cardiopulmonary transport mechanisms reach maximum levels, shortness of breath, fatigue, and muscle pain become limiting symptoms; motivation is then the final determinant of the duration of exercise. Ageing reduces the efficacy of cardiopulmonary transport mechanisms and, of course, exercise capacity. The heart rate response at peak exercise reflects this. In healthy individuals aged 20 years it is about 200 beats/min and at 65 years about 170 beats/min.

When exercise stops, the cardiopulmonary and metabolic changes return rapidly to resting levels, the rate following an exponential pattern in the first few minutes; the excretion and metabolism of lactate and other substances, and the dissipation of heat generated take longer (time constant of about 15 min or more). Reduced circulatory function slows the recovery rate.

Training effects

Regular exercise to about 60 per cent of maximal heart rate for 20 to 30 min three times a week is the minimum requirement for improved effort tolerance due to a training effect. The resting heart rate becomes slower whilst the cardiac output is maintained by an increased end-diastolic volume and ejection fraction, and therefore stroke volume. In a 'trained' exercising individual there is a reduced heart rate response to a standard submaximal work load, and systemic blood flow is more effectively distributed away from visceral and skin circulations to working muscles. Adaptive changes occur in muscle mitrochondria, permitting improved extraction of oxygen from perfusing blood so that maximum oxygen consumption increases. There is suggestive evidence that prolonged endurance training increases the calibre of coronary arteries and enlarges the capillary surface area relative to cardiac muscle mass (in animals). Myocardial protein synthesis increases. Adrenergic mechanisms appear to be involved in mediating this response. It should be noted that rhythmic exercise (such as running) and isometric exercise (such as weightlifting) have different physiological effects. The blood pressure rises disproportionately during the latter. The mechanisms are partly reflex and partly mechanical from the contracting muscles. Isometric exercise training is not recommended for cardiac patients because of the increased afterload it imposes on the heart.

Regular exercise may also partly prevent the now well documented endothelial dysfunction associated with ageing. The age-related reduction in availability of nitric oxide resulting from reduced activity of the L-arginine–nitric oxide pathway in the endothelium has now been established and is thought to be a consequence of oxidative stress. Regular exercise improves the availability of nitric oxide. Vascular effects related to nitric oxide are considered elsewhere.

There is now good evidence for exercise-induced mood changes resulting in feelings of well-being. Increased concentrations of circulating b-endorphin occur during exercise, and studies using the opiate receptor antagonist naloxone to block the effects of opioid peptides suggest that b-endorphin release may reduce exercise-induced adrenaline and noradrenaline responses. Regular exercise lowers blood pressure in normotensive and mildly hypertensive subjects, and modulation of catecholamine release by changes in endogenous opioid peptide secretion may be a possible contributing mechanism. There are other diverse exercise-induced hormonal changes, but one of particular clinical relevance is reduced glucose-stimulated insulin secretion. This is beneficial for type II diabetics, whose basal hyperinsulinaemia is the result of both hypersecretion and hypocatabolism of insulin, and for patients with insulin resistance and hyperinsulinaemia, obesity, hypertension, and dyslipidaemia—so-called syndrome X.

To summarize, changes in the four essential determinants of cardiac function—preload, afterload, heart rate, and contractility—combine to augment cardiac output and oxygen delivery during exercise. Measurement of the cardiovascular response to exercise is essential for the objective assessment of cardiac function.

Further reading

Braunwald E, Zipes DP, Libby P (2001). *Heart disease: a textbook of cardiovascular medicine*, 6th edn. WB Saunders, Philadelphia.

Ganong WF (2001). *Review of medical physiology*, 20th edn. McGraw Hill, New York.

Hill AV (1970). *First and last experiments in muscle mechanics*. Cambridge University Press, Cambridge.

Jones NL, Killian KJ (2000). Exercise limitation in health and disease. *New England Journal of Medicine* **243**, 632–41.

Suzuki T, Yamazaki T, Yazaki Y (2001). The role of natriuretic peptides in the cardiovascular system. *Cardiovascular Research* **51**, 489–94.

15.2.1Chest pain

J. R. Hampton

Introduction

Chest pain is one of the commonest causes of emergency admission to hospital (Table 1). In the hospital context the commonest cause of chest pain is myocardial ischaemia, but in primary care other causes predominate. The majority of those admitted with chest pain will have evidence of pre-existing coronary disease (a previous myocardial infarction or known previous angina) and their pain is likely to be a manifestation of this. Perhaps 20 per cent of this group will have a definite myocardial infarction. A misdiagnosis of chest pain can have serious consequences for both patient and doctor, but the number of patients seen in accident and emergency departments with chest pain is so large that it is impracticable for all to be admitted to hospital for investigations to be completed. Similarly, it is neither necessary nor sensible for every patient with chest pain seen in primary care to be referred to hospital.

The initial management of patients with chest pain depends on an accurate history, the physical examination and simple investigations being of lesser importance.

The patient's history

Myocardial ischaemia

Coronary artery insufficiency, usually the result of atheroma but occasionally the result of arterial spasm, leads to a spectrum of conditions which can be grouped together as 'myocardial ischaemia'. The common end result is chest pain.

Stable angina

Angina was first recognized by William Heberden in 1768, and his description of the pain of myocardial ischaemia can hardly be bettered. He wrote:

> The seat of it, and the sense of strangling and anxiety with which it is attended may make it not improperly called angina pectoris.
>
> They who are afflicted with it, are seized while they are walking (more especially if it be uphill, and soon after eating) with a painful and most disagreeable sensation in the breast, which seems as if it were to extinguish life, if it were to increase or continue; but the moment they stand still, this uneasiness vanishes....The pain is sometimes situated in the upper part, sometimes in the middle, sometimes at the bottom of the os sterni, and often more inclined to the left than to the right side. It likewise very frequently extends from the breast to the middle of the left arm.

Whether myocardial ischaemia is due to stable angina, myocardial infarction, or the acute coronary syndromes, the distribution of the pain is much as Heberden described it. It is central, or sometimes left sided; rarely it is felt only in the back. The most classical—but not the most common—radiation is to the front of the neck, and the lower jaw or the teeth. The most common radiation is to the left arm, though the pain of other chest problems may also radiate here. Radiation from the front of the chest to the back is not uncommon.

The nature of the pain is usually described as 'tight', 'crushing', 'squeezing', or 'heavy'. Many patients describe a 'sensation', or an 'ache' rather than a pain. The pain of myocardial ischaemia is seldom described as 'sharp'.

The duration of the pain in stable angina depends on whether or not the patient rests: typically the pain will disappear within a minute or two when exercise ceases. Sometimes the pain will disappear if exercise continues; this has been called 'walk through' angina.

Stable angina, as opposed to the other manifestations of myocardial ischaemia, can be recognized by factors that precipitate or relieve it. Stable angina is, above all, predictable. It occurs after a constant amount of exercise—often on a particular hill, after so many metres walking, on climbing so many stairs, or on hurrying. Pain that is unpredictable is unlikely to be stable angina, though it may be due to an acute coronary syndrome. The pain of stable angina is worse on exercise in cold or windy weather, and on exercise after a meal. It is often induced by sexual intercourse, or by any emotional stress. Any chest pain, whatever its distribution, that occurs with emotional stress is likely to be angina.

Stable angina is relieved rapidly by rest, and very rapidly by a tablet or sublingual spray of a short-acting nitrate. If the problem is not relieved within 3 or 4 min by a nitrate either it is not stable angina (it could be a myocardial infarction) or the nitrate has exceeded its shelf-life and has become inactive.

Chest pain, even with the characteristic distribution of angina, that occurs at the end of a busy day rather than during exercise, that lasts for hours, and which is not helped by nitrate is most unlikely to be angina.

Angina may be associated with breathlessness, and sometimes breathlessness on exertion is the dominant symptom. Table 2 gives a check list of the features that are characteristic of angina.

Myocardial infarction

The distribution of pain due to myocardial infarction is similar to that of stable angina. It is usually in the centre of the chest, and may radiate to the jaw, teeth, arms, or back. The nature of the pain is also similar—typically squeezing or crushing—but it is usually much more severe and can be one of the worse of all pains. It is often associated with a cold sweat, and with vomiting. The pain is frightening, and many will volunteer that they thought that they were going to die. The pain typically lasts a few hours. It may occasionally be associated with heavy exertion (particularly any activity that causes a sudden rise in heart rate and blood pressure) but usually there is no obvious precipitating cause. Nothing other than a strong analgesic such as diamorphine will relieve the pain, and nitrates are completely ineffective. Some patients need a second injection of diamorphine, but few need three, and if chest pain lasts more than 24 h a diagnosis other than myocardial infarction should be considered.

Myocardial infarction can be 'silent', meaning that it occurs without much in the way of pain. Population surveys of older people suggest that some 15 per cent of previous myocardial infarctions demonstrated on ECG are unrecognized. 'Silent' infarction is more likely to happen in diabetics.

The symptoms of myocardial infarction are, of course, modified by complications such as heart failure, arrhythmias, heart block, and pericarditis.

Unstable angina

The chest pain of unstable angina is difficult to describe because unstable angina is difficult to define. The term has been used to describe the first attack of what later proves to be stable angina, stable angina of increasing frequency and severity (sometimes called 'crescendo' angina), and angina at rest. The development of new markers of myocardial infarction such as the troponins has widened the possible definition of myocardial infarction, and unstable angina with a positive troponin test merges into 'non-Q-wave' infarction.

The distribution and intensity of the pain of unstable angina resembles that of stable angina, though patients with non-Q-wave infarction can have severe chest pain indistinguishable from that of Q-wave infarction.

Typically the chest pain of unstable angina will occur at rest. Many patients with unstable angina will gain relief from nitrates, but those with non-Q-wave infarction will not.

The pain of unstable angina is seldom associated with symptoms due to immediate complications such as heart failure, though unstable angina is associated with a relatively high risk of myocardial infarction and death in the next 3 months.

Prinzmetal's variant angina and syndrome X

There are two further varieties of angina that cause ischaemic chest pain, but both are somewhat nebulous concepts and are probably overdiagnosed. In Prinzmetal's 'variant' angina, chest pain characteristic of angina occurs either on exercise or at rest. The pain is believed to result from spasm of the coronary arteries. The diagnosis is only made if an ECG shows ST segment elevation, rather than the usual depression, during an attack of pain. The ST segment returns to normal as the pain disappears, which differentiates the ECG change from that of acute myocardial infarction.

'Syndrome X' is a term used to describe the occurrence of exercise-induced stable angina with a positive exercise test but a normal coronary angiogram. The problem is assumed to be in vessels too small to be demonstrated angiographically, and the term 'microvascular angina' is sometimes used. While the syndrome undoubtedly exists, the diagnosis should only be accepted after very careful investigation.

Dissection of the aorta

Perhaps one patient in a thousand of those who are admitted to hospital with chest pain has a dissection of the aorta. It can be very difficult to diagnose aortic dissection unless the possibility is considered in all of these patients. This is important: the treatments for myocardial infarction and aortic dissection are very different.

The typical patient has a pain similar to that of myocardial infarction, but the pain is more usually sudden in onset. Like infarction, it is in the centre of the chest. It is usually very severe and it often radiates to the back. It is often described by the patient as 'tearing', and it lasts much longer than the pain of myocardial infarction. The onset of pain is often associated with a 'collapse' or with sudden neurological deficit suggesting a stroke. The position of the pain gives a rough guide to the site of dissection; anterior chest pain correlates to some extent with proximal dissection, while pain in the back correlates with distal dissection. The pain may move from the front to the back of the chest as the dissection spreads distally.

If the dissection affects the ostium of a coronary artery, myocardial infarction with its own pain may result.

Unfortunately not all patients with aortic dissection give a typical story. Since the physical examination is not always helpful and the ECG may be normal, patients are not infrequently discharged from the accident and emergency department, or from the ward, with some such diagnostic label as 'myocardial infarction excluded'. Since patients with dissection who survive the first few hours have a reasonable chance of successful surgical repair, it is essential that the diagnosis should at least be considered (if only to be rapidly excluded) before any patient presenting with acute chest pain is discharged.

Pericarditis

The most common cause of pericarditis is myocardial infarction, and the pain of infarction is sometimes replaced by that of pericarditis. Viral infection is a common cause, and pericarditis should be suspected in anyone presenting with chest pain in the context of a 'flu-like illness, particularly if they are likely to be at low risk of coronary artery disease (young men or premenopausal women with few risk factors).

Pericardial pain has some features in common with those of myocardial ischaemia and aortic dissection. It is central, and may have the same radiation. It differs, however, in being (usually) less severe and lasting longer.

The most typical feature of pericardial pain is the effect of posture. The pain is worst when the patient lies on his or her back: the pericardial fluid drains to the back of the pericardial sac, leaving the inflamed anterior visceral and parietal pericardial surfaces to come into contact and so cause pain. Sitting up and leaning forward, which allows the fluid to drain to the front, separates the pericardial layers and relieves the pain.

Pericardial pain often has a pleuritic element, being worse on deep breathing.

Pleuritic pain

Pain from the pleura—the old term was 'pleurisy'—is usually on one side of the chest only. Whatever the cause (and the main causes are infection, pulmonary embolism, and pneumothorax) it is identified because it is worse on inspiration. As a result the patient will need to take shallow and therefore rapid breaths. Unlike the pain of myocardial ischaemia it is usually described as sharp or knife-like. It can be severe, though seldom as severe as myocardial infarction or aortic dissection. Its onset can be sudden (especially when due to pneumothorax or pulmonary embolism) or slow (infection). Pleuritic pain can often be identified because it is associated with symptoms of lung disease such as breathlessness, cough, sputum production, or haemoptysis.

Oesophageal pain

Pain originating in the oesophagus can be difficult to differentiate from the pain of myocardial ischaemia. Both are common, so patients not infrequently suffer from both, and cannot always tell one from the other.

Typical oesophageal pain is central and anterior, and is often described as 'burning'. It usually has some relation to eating. It is commonly due to oesophagitis caused by acid reflux from the stomach because of a hiatus hernia. The pain is often induced by bending, when the patient can be aware of an acid and bitter taste in the mouth. Oesophagitis can cause spasm which is itself painful, and the spasm may be relieved by nitrate, sometimes leading to confusion with angina. More commonly, oesophageal pain is relieved by an antacid.

Rupture of the oesophagus causes severe central chest pain very similar to that of a myocardial infarction. It always follows vomiting, as opposed to the vomiting which can accompany myocardial infarction, which occurs after the pain has become intense.

Musculoskeletal chest pain

Pain can arise in any of the structures of the chest wall, and can mimic all other causes of pain. Pain that is induced or relieved by postural change is likely to be musculoskeletal in origin, as is highly localized pain reproduced by pressure at the affected site. Nerve root compression due, for example, to vertebral disease (collapse, metastasis, abscess) can cause pain to radiate round the ribs. If the pain results from bony collapse it can be of sudden onset, but musculoskeletal pain seldom has the time course of ischaemic pain.

Pain to the left of the sternum, with tenderness over the costochondral junctions, is common in middle-aged men. Sometimes thought to have an inflammatory basis, it

is not associated with arthritis elsewhere or with a rise in inflammatory markers. It is sometimes called Tietze's syndrome.

Under the same heading comes the pain of herpes zoster. Patients, especially the elderly, may be admitted to a coronary care unit with a severe left-sided chest pain with tenderness, and the cause becomes obvious the following day as the characteristic rash appears.

Chest pain ? cause

After all the possible causes of chest pain have been excluded there remains a group of patients, usually middle-aged men, in whom no firm diagnosis can be made. It is entirely proper to label these as 'chest pain ? cause'. Making a diagnosis of 'musculoskeletal pain' in such patients is not only incorrect, but it may prevent the diagnosis being properly reassessed on a later occasion.

'Chest pain ? cause' is the discharge diagnosis in about 10 per cent of patients admitted to hospital with suspected myocardial infarction. The pain can be similar to that of acute ischaemia, though it is seldom very severe, and while it may radiate to the left arm it never spreads to the jaw or teeth. By definition, detailed investigation fails to explain the pain, which is not infrequently recurrent. Long term follow-up of such patients shows that their prognosis is essentially that of the healthy population.

'Chest pain ? cause' is also a proper diagnosis in patients with recurrent chest pain in whom the alternative is stable angina. Here the pain is nearly always left-sided, is not predictable, is not clearly related to exercise or emotional stress, and is never brought on by sexual intercourse. This sort of pain merges into Da Costa's syndrome, which was first identified in the American Civil War. Soldiers complained of sharp, lancinating or burning pain, often on a background of a duller pain with a feeling of 'uneasiness' around the heart. The patients described by Da Costa were also troubled by palpitation and hyperventilation. The same thing was observed in the First World War, and this type of pain is considered functional: older terms used to describe it were 'soldiers' heart' and 'cardiac neurosis'. However, non-specific 'chest pain ? cause' also occurs in people who are not apparently under stress.

Physical signs in patients with chest pain

In patients with severe pain the findings on examination may be dominated by those due to pain itself: pallor, cold and clammy extremeties, and a sinus tachycardia. The blood pressure may be high due to intense peripheral vasoconstriction, or low if there has been severe myocardial damage. After pain relief, or in between episodes of pain, there may be few physical abnormalities.

When the pain sounds like chronic stable angina it is important to look for possible causes other than coronary disease, including aortic stenosis, anaemia, and brady- or tachyarrhythmias. There may be evidence of peripheral vascular disease: absent peripheral pulses, or bruits over the carotid or femoral arteries, suggesting that coronary disease is likely. There may be signs of risk factors such as those associated with hypercholesterolaemia, and hypertension may be present.

All these physical signs should be sought in patients with persistent chest pain. In addition there may be evidence of myocardial dysfunction, including heart failure (raised jugular venous pressure, a gallop rhythm, pulmonary crackles), mitral valve regurgitation due to papillary muscle dysfunction or rupture, and ventricular septal defect.

Although there are classical signs of aortic dissection, these are by no means uinversal. About one-third of the patients will have high blood pressure; a third will have a murmur of aortic regurgitation due to distortion of the aortic root; and perhaps half will have a pulse missing or a different blood pressure in each arm. A pericardial friction rub or pericardial tamponade due to blood tracking backwards into the pericardium is uncommon, but when associated with aortic regurgitation the diagnosis of aortic dissection becomes almost certain.

Pericarditis is diagnosed from the presence of a pericardial friction rub, which is best heard with the patient lying flat. It may be associated with signs of tamponade which include a rise in the jugular venous pressure and fall in the arterial pressure on inspiration. Similarly, pleuritic pain can be identified from the pleural rub if one can be heard. Otherwise there may be sounds of pulmonary consolidation or pleural effusion that make associated pleurisy likely.

Patients with viral pericarditis are likely to have a fever. In those with pleurisy, a high fever (over 38.5 °C) makes pneumonia a more likely diagnosis than pulmonary embolism. Myocardial infarction causes a low-grade fever, but this not often seen until a day or two after the event.

There are few physical signs in patients with oesophageal pain, though there may be tenderness in the epigastrium. Musculoskeletal pain is suspected when there is bony tenderness (fractures, metastases) or when the pain is reproduced by local pressure or movement. 'Chest pain ? cause', virtually by definition, has no physical signs other, perhaps, than those of anxiety.

The immediate management of patients with chest pain

Patients seen in the accident and emergency department with chest pain need urgent assessment and treatment. Those with a possible acute coronary syndrome need rapid sorting into those who need immediate reperfusion of a blocked coronary artery (identified by raised ST segments on the ECG) and those with non-Q-wave infarction or unstable angina, who have less need for immediate treatment but who will need more detailed investigation and treatment over the next few days. All these patients should be admitted to hospital, ideally to a coronary care unit, though it is frequently impracticable to admit all patients seen in an accident and emergency department with chest pain.

If the initial ECG is normal but the patient has significant pain, the ECG should be recorded again after 1 and 2 h. Patients with persistent pain at that point will have to be admitted. Early plasma markers of myocardial necrosis—the creatine kinase and creatine kinase myocardial band enzymes, and the troponins—are helpful if positive, but will often not be elevated until about 6 h after the onset of pain. Although one possible treatment strategy is to hold patients in the accident and emegency department until 6 h have elapsed and the enzymes and troponins have been shown to be normal, this is not foolproof because the troponins may not become elevated for as long as 18 h.

Chest radiographs are seldom helpful in patients with chest pain, and if anteroposterior films are taken they can be positively misleading. Obtaining a posteroanterior film may mean transferring the patient to an X-ray department where close monitoring cannot easily be maintained, but using portable X-ray equipment in the accident and emergency department will almost inevitably produce a distorted cardiac and mediastinal shadow. It is under these circumstances that widening of the mediastinum, which may be the first indication of an aortic dissection, will be missed. Portable X-ray equipment is only reliable for assessing the lung fields, and to some extent for detecting heart failure. If aortic dissection seems at all possible, then a departmental posteroanterior chest radiograph is essential.

Patient management inevitably depends to some extent on the patient's perception of what has happened and whether, for example, the pain that has led to his or her hospital attendance is like or unlike previous angina. In general, however, if the patient becomes pain free and there are no important physical abnormalities, if an ECG repeated after 2 h shows no change from the initial recording, and if a departmental chest radiograph and plasma troponin or creatine kinase are normal, then it is not unreasonable to allow the patient to go home (with notification for their general practitioner and arrangements for further investigation, for example cardiac treadmill testing, as appropriate). Total safety of diagnosis and management requires hospital admission and observation for perhaps 18 h, but in the real world this is often impracticable.

The diagnosis depends on the synthesis of the patient's history, the physical examination, and simple investigations. A carefully taken history, coupled with a high index of suspicion for important problems such as an acute coronary syndrome or aortic dissection, is the key to a successful patient management.

Further reading

Bakker AJ *et al*. (1993). Failure of new biochemical markers to exclude acute myocardial infarction at admission. *The Lancet* **343**, 1220–2.

Bayliss RIS (1985). The silent coronary. *British Medical Journal* **290**, 1093–4.

Cannon RO (1993). Chest pain with normal coronary angiogram. *New England Journal of Medicine* **328**, 1706–8.

DeSanctis RW *et al*. (1987). Aortic dissection. *New England Journal of Medicine* **317**, 1060–7.

Hampton JR, Gray A (1998). The future of general medicine: lessons from an admissions ward. *Journal of the Royal College of Physicians of London* **32**, 39–42.

Ohman EM *et al*. (1996). Cardiac troponin T levels for risk stratification in acute myocardial ischaemia. *New England Journal of Medicine* **335**, 1333–41.

Ryan J *et al*. (1996). ACC/AHA Guidelines for the Management of patients with acute myocardial infarction. *Journal of the American College of Cardiology* **28**, 1328–1428.

Slater EE, DeSanctis RW (1976). The clinical recognition of dissecting aortic aneurysm. *American Journal of Medicine* **60**, 625–33.

Spittell PC *et al*. (1993). Clinical features and differential diagnosis of acute dissection: experience with 236 cases (1980 through 1990). *Mayo Foundation for Medical Education and Research* **68**, 642–51.

Thadani U *et al*. (1971). Pericarditis after acute myocardial infarction. *British Medical Journal* **2**, 135–7.

15.2.2 The syndrome of heart failure

Andrew J. S. Coats

Introduction

Heart failure is a common condition, carrying a high burden of disability and mortality. Many treatments have now been established that ameliorate, at least partially, its debilitating effects, but despite this it is increasing in both prevalence and cost in the developed and developing worlds as the population ages. Much disability remains, and there are many shortfalls between treatment possibilities and that which is achieved in everyday practice around the world. Major advances seem possible with the advent of greater understanding about the causes of cardiovascular disorders, including molecular mechanisms, and the development of newer, effective treatments, including surgical advances and gene therapies.

Definitions

'Heart failure' is an unfortunate term. It has negative connotations for the patient and describes imprecisely several different clinical situations. Left and right heart failure are quite distinct clinical syndromes, although they frequently coexist (biventricular failure). Historically heart failure has been further subdivided on the basis of presumed pathophysiological mechanisms into: (1)'forward' or 'backward' heart failure, depending on whether congestion or organ underperfusion was the predominant clinical feature; (2)'congestive' or 'non-congestive', depending on the presence or absence of oedema; and (3)'high-output' or 'low-output'. These subdivisions have not proved to be particularly useful. A more recent and more useful classification is dependent on the predominant pattern of left ventricular dysfunction, be it systolic, diastolic or mixed. Whatever the complexities of the ventricular pathophysiology that initiates events, a well-recognized clinical pattern is identifiable as 'heart failure' and has proved a useful description of a complex clinical syndrome for many years.

The important features of any definition of heart failure (of which there have been several) are that the clinical picture is:

1. initiated by a reduction in effective cardiovascular (usually left ventricular) functional reserve;
2. associated with symptoms either at rest or at an unexpectedly low level of exertion; and
3. associated with characteristic pathophysiological changes in many disparate organ systems.

These latter can include biochemical, hormonal, metabolic, or functional alterations. In simple terms heart failure is a syndrome in which a reduction in left ventricular function causes pathophysiology that produces symptoms and exercise limitation.

A clinical picture similar to that of heart failure can develop when ventricular function itself is normal, but where there is an extreme volume or pressure overload on the ventricle. These include volume overload conditions such as endotoxic high-output shock, severe anaemia, arteriovenous fistulas or shunts, and pressure overload conditions such as acute hypertensive crisis or prosthetic heart valve occlusion. It is appropriate both clinically and for research purposes to separate these from cases where the initiating cause is a reduction in ventricular function.

Acute and chronic heart failure

It is conventional, because of differences in assessment and management, to separate acute from chronic heart failure. Both are different stages of a single disease process, and in the clinical course of a patient with chronic heart failure acute exacerbations may be common, often described as 'acute decompensation' or 'acute on chronic' heart failure. Acute heart failure is typically a dramatic clinical presentation with an acutely dyspnoeic patient demonstrating visible signs of cardiovascular insufficiency such as tachycardia, pulmonary or peripheral oedema, and underperfusion of systemic organs. Chronic heart failure, by contrast, can be a subtle disorder, which if gradual in onset can be missed by both patient and physician. The salient features are the initiation and persistence of left ventricular dysfunction, and the pathophysiological changes in other organs that produce symptoms and which limit exercise. A persistent state of circulatory insufficiency can exist in severe chronic heart failure, with pulmonary and peripheral oedema and symptoms and signs of distress even at rest.

Epidemiology, aetiology, and pathogenesis

Epidemiology

Heart failure is a common condition with an estimated incidence of 20 to 30 per thousand of the adult population per year and an overall prevalence of about 1 per cent. The prevalence increases in frequency with increasing age, reaching 30 per cent in those aged over 80 years, and in developed countries the average age of patients with heart failure is now in excess of 75 years. It is one of the most expensive medical conditions, and is an increasing major healthcare cost. Because of its many debilitating symptoms, heart failure is a frequent cause for both acute hospital and long-stay residential care admissions, indeed it is the most common discharge diagnosis from hospitals in the developed world in people over the age of 65, and the second most common overall. Heart failure is a feature of the clinical

condition of approximately 5 per cent of patients in hospital at any time, and also the one with the greatest rate of hospital re-admission.

Paradoxically, improvements in the management of acute myocardial infarction and chronic coronary heart disease have led to more instances of heart failure rather than less, as more people survive to develop heart failure later in life. However, preventive therapies do work: multiple trials of antihypertensive treatment and the 4S trial of cholesterol reduction have shown a significant reduction in the incidence of new cases of heart failure in high-risk populations. Smoking reduction also reduces the number of new cases of heart failure, as does the selective use of angiotensin-converting enzyme inhibitors in high-risk patients after a myocardial infarction or for those with asymptomatic left ventricular systolic dysfunction. Appropriate use of thrombolytic therapy at the time of myocardial infarction will also reduce the incidence of new cases of heart failure.

Whilst there is reasonable consensus as to the prevalence of heart failure in younger and middle-aged populations, where most cases demonstrate significant deterioration in systolic function of an enlarged heart, in the elderly an increasing proportion of cases of clinically suspected cases of heart failure have small ventricular cavities and preserved systolic function. In these cases diastolic dysfunction can be frequently demonstrated, and many experts feel that the majority of cases in an older population will be due primarily to diastolic dysfunction. However, methods of assessing diastolic function are less developed than those for systolic function, and interventional trials have historically concentrated on systolic dysfunction as a cause of heart failure. As a result we know much less about how best to diagnose diastolic heart failure and how best to teat it once diagnosed. This remains a major challenge to the cardiological community in the twenty-first century.

Aetiology

Heart failure is a clinical syndrome, not a single diagnosis, and it can have many different aetiologies. In Western industrialized societies the most common underlying causes are ischaemic heart disease, hypertension, and idiopathic dilated cardiomyopathy (see Chapter 15.8.2). The Framingham study suggested that hypertension, especially when complicated by left ventricular hypertrophy, was by far the most common antecedent of heart failure. Recent intervention trials in heart failure have usually included a preponderance of patients whose heart failure was secondary to ischaemic heart disease. In recent cross-sectional studies of heart failure in the community, hypertension is cited to be a relatively minor cause of heart failure. This change has been attributed to better detection and treatment of hypertension, but it may also reflect re-labelling, with coronary artery disease more likely to be blamed for heart failure than hypertension in the many patients who have evidence of both. Some cases of hypertension may proceed to a dilated poorly functioning heart with an eventual normalization in arterial pressure. These cases may be labelled as idiopathic dilated cardiomyopathy, with the only clue to the correct underlying diagnosis being a greater than expected degree of left ventricular hypertrophy.

In industrialized societies previously common causes of heart failure such as nutritional deficiency disorders or chronic complications of rheumatic valvular disease are now rare. In less developed societies infective causes still underlie the majority of cases. Some disorders may be common in particular societies and these should always be borne in mind when assessing an individual patient: examples would include Chagas' disease in Central and Southern America, iron overload in certain tribes in southern Africa, and nutritional deficiency states in the world's poorest countries.

Classification of cause

More than one underlying cause of heart failure can coexist, such as hypertension and ischaemic heart disease. Table 1 lists the major causes of heart failure, subdivided according to the mechanisms by which ventricular disease leads to the clinical syndrome. Such a differentiation is important because of specific strategies available for certain diagnoses, such as nutritional support, cardiac valve or bypass surgery, endocrine therapy, and avoidance of a toxic agent.

Pathogenesis

There is no unique pathological finding in the heart or elsewhere that defines the presence of heart failure. Heart failure can be the result of a wide variety of cardiovascular disorders: anything that puts an excessive demand on the heart for a prolonged period can lead to myocardial failure. Alternatively, loss of myocytes or an abnormal myocardial interstitium can lead to loss of effective heart function and cause the clinical syndrome of heart failure.

Although the more severe the loss of myocyte number or reduction in cardiac pumping capacity, the more likely is clinical heart failure, there is no strict relationship between measures of global cardiac function and the presence or severity of the features of the heart failure syndrome. The severity of heart failure is measured by the severity of symptomatic limitation, and by the extent of pathophysiological abnormalities, which closely correlate with the reduction in survival. Important amongst these changes of heart failure are the body's responses such as neurohormonal overactivity, autonomic dysfunction, and immunological and metabolic derangements.

Oxygen and energy supply

Myocardial dysfunction can result from a deficient oxygen supply to the myocardium, whether caused by occlusive coronary artery disease or by a reduced blood-carrying capacity such as in anaemia or certain toxic states. In addition, endothelial dysfunction, raised ventricular myocardial tissue pressures, and reduced diastolic blood pressure and diastolic time intervals all contribute to a reduction in the net effective coronary flow. Energy metabolism is frequently abnormal within the myocardium in cases of chronic heart failure. This can be a primary defect in familial cardiomyopathies or an acquired defect, such as in the insulin resistance that complicates chronic heart failure.

Defects in myocardial contractile performance

There remains considerable controversy as to whether individual myocytes are functionally deficient in most cases of human chronic heart failure. There are isolated cardiomyopathies where such defects are likely, but in most cases the major defect is a loss of myocyte number with compensatory myocyte hypertrophy. This leads to dysfunction of myocyte relaxation, due in part to an intracellular accumulation of calcium, rather than deficient contraction. Isolated single gene defects can be the cause of rare familial cardiomyopathies and in these, and presumably in more cases in the future as we understand more of the genetic processes underlying the control of myocardial contraction, specific abnormalities of myocyte contraction can be implicated in the cardiac dysfunction evident at the organ level.

Defects in the control of myocardial function

Although in theory an abnormal control of contraction could lead to cardiac dysfunction sufficient to cause heart failure, examples of clinical syndromes demonstrating this pathophysiological mechanism are rare. Myocyte necrosis can occur in cases of persistent sympathetic overactivity such as phaeochromocytoma, and more commonly excessive blockade of sympathetic nerve endings can acutely remove this support to myocardial contractility leading to acute heart failure.

Changes in the interstitium of the heart

Excessive myocardial fibrosis, such as that seen in senile changes in the heart and as a complication of sustained hypertension and aortic stenosis, can reduce the effective myocardial performance despite individual myocyte hypertrophy. Similarly, rare cases such as endomyocardial fibrosis can cause a syndrome of cardiac failure despite individual myocytes being functionally normal if studied in isolation. The importance of the intracellular milieu has only recently been fully recognized, and this may be a target for future interventions to modify the processes of ageing of the myocardium and progression in the syndrome of chronic heart failure.

Pathophysiology

Cardiac

Structural changes

Structural changes in the heart are common, both at macroscopic and microscopic levels. The clinical picture usually includes enlargement of the left ventricular cavity (with the exception of diastolic dysfunction and restrictive or constrictive cardiomyopathies). The shape of the ventricle also changes, becoming more spherical. This can occur rapidly after a myocardial infarction via a passive process of stretching of the infarcted territory (infarct expansion), or more slowly over a period of weeks to months in a process termed 'remodelling' (see Chapter 15.4.2). A similar change in shape is seen in dilated cardiomyopathies, but not in the restrictive cardiomyopathies. The more spherical shape of the 'remodelled' and enlarged ventricle increases the stress of the myocardial wall and may thereby worsen myocardial ischaemia. The shape change may also disrupt the complex conformational change that normally occurs during the isovolumic contraction phase, in which

the apex of the ventricle constricts in a twisting motion and pushes the blood into the ventricular base. Where the ventricle is already spherical at rest, this intraventricular redistribution of blood during isovolumic contraction is not possible and the net effect is a reduction in the efficiency with which the blood is ejected.

Cardiac enlargement has long been known to be an adverse prognostic sign, even when estimated crudely as the cardiothoracic diameter on chest radiographs. More precise measurements of the internal dimensions of the left ventricle by echocardiography have confirmed the prognostic value of cardiac enlargement in patients recovering from myocardial infarction, even when accounting for the size of the myocardial infarct. Prevention of the late remodelling process was the theory behind the use of angiotensin-converting enzyme (**ACE**) inhibitors after myocardial infarction. These agents have been shown to reduce ventricular size and to reduce late mortality if given early after infarction, but whether this beneficial effect is directly related to any reduction in ventricular remodelling is not known.

Changes at the microscopic level

The failing heart also shows alterations in cardiac structure at microscopic and ultrastructural levels. There is an increase in the collagen content of the extracellular matrix, a process thought to be partly related to increased wall stress and partly due to neurohormonal activation, particularly aldosterone. This change reduces ventricular wall distensibility and may affect the efficiency with which active restorative forces can assist the diastolic filling process. Hence this microscopic structural change may help to explain the frequent coexistence of systolic and diastolic functional deterioration in an enlarging ventricle in chronic heart failure.

The enlargement of the ventricle is associated with thinning of the ventricular wall and, as there is believed to be no increase in the total myocyte population, there must be a realignment of the intercellular attachments between individual myocytes. This process, whereby there is a continual breaking and reforming of cell-to-cell junctions to allow remodelling, has been termed 'cell slippage', although exactly how this occurs has not been established. There are changes in the microscopic structure of the failing ventricle, with a reduced number of tight junctions between myocytes, and this may be involved in this process.

Overall circulatory function

The description of an objective measurement of systolic function in intact humans has proved difficult. In simplest terms, the left ventricle is a pump that generates both pressure and flow. It has a theoretical operating range from a pure pressure generator to a pure flow generator, although it always functions as a mixed pump. The function of this pump can be described in terms of the kinetic and potential energy it imparts to the blood ejected each beat, or in terms of the average power output of the circulation (flow multiplied by the mean pressure drop), assuming the left ventricle is the only significant power source in the circulation. Thus, overall ventricular function can be described as cardiac output multiplied by the pressure drop across the systemic circulation, a quantity described as cardiac power output. Cardiac power output is well preserved at rest even in severe heart failure, but the maximal reserve of cardiac power output is reduced progressively as heart failure progresses, and a significant reduction in maximal power output during inotropic stimulation is a poor prognostic sign.

The measurement of cardiac power output tells us little, however, of the mechanisms underlying any reduction in ventricular performance. This may be due to reduced ventricular filling, or emptying, or to wasted myocardial power such as in aortic stenosis. Hence, attempts have been made to define the components of ventricular function to explain the nature of a reduced overall circulatory function and to assist in monitoring a patient's clinical course and response to treatment.

Systolic dysfunction

Systole can be defined either clinically as the ejection phase between mitral valve closure and aortic valve closure, or in terms of ventricular dynamics as the phase of contraction of the myocytes within the ventricle. These two definitions do not coincide, for there is a period of isovolumic contraction at the onset of ventricular systole in which myocyte contraction generates a pressure increase within the ventricle and a conformational change in its shape, but during which no blood is ejected. Similarly during the latter phase of ventricular ejection, the blood is flowing out of the left ventricle passively and the myocardial elements may be already relaxing.

In clinical practice, systolic dysfunction is most easily recognized by direct haemodynamic measurements showing a reduced peak rate of pressure rise within the ventricle (positive d*P*/d*t* max), an increased filling pressure (left ventricular end-diastolic pressure, **LVEDP**), or by indirect measurement of ventricular volumes (see Chapter 15.3.6). If there is a reduction in myocardial contractile function an enlargement of the ventricle will develop, in which a greater preload will enhance ventricular emptying via the Frank–Starling mechanism. As a result the ventricle will operate at an increased end-diastolic and end-systolic volume. This can be measured by pressure and volume estimations, such as by ventriculography (either radiographic or radionuclear) or echocardiography. Although not a direct measure of ventricular performance, ejection fraction, being the fractional emptying of the ventricle with each beat, carries information about ventricular volumes and global ventricular function. This is only a poor predictor of the severity of symptomatic limitation, but it has been shown to be an important predictor of longevity in heart failure, independent of other measures of severity, and it has the advantage of simplicity. At the most simple level, therefore, systolic heart failure can be recognized by signs of cardiac insufficiency in the presence of an enlarged ventricle, and clinically is most conveniently estimated by the left ventricular ejection fraction.

Diastolic dysfunction

Diastole is the opposite of systole, the period of filling of the ventricle or the period of relaxation of the myocytes. Objective measurements of diastolic function are, however, more problematic than for systolic function. Whereas systole occurs rapidly and in one action, diastole is complex, with an initial rapid and active ventricular recoil producing filling of the ventricle, then a period of relative stasis as atrial and ventricular pressures equilibrate, followed by a second period of ventricular filling due to atrial contraction. These processes are affected by many factors including heart rate, atrioventricular delay, atrial contractility, active myocardial recoil, passive ventricular wall stiffness, the efficacy of ventricular systole, and the residual end-diastolic volume and pressure within the ventricle. As a result of all these interacting factors, it is not surprising that no simple measure of 'diastolic function' has been developed, and those measures that have been used clinically are profoundly affected by systolic function and heart rate. However, diastolic functional disturbance is important: there are cases of definite clinical heart failure in which the patient has a small heart, with normal or even increased left ventricular ejection fraction, and in whom the only demonstrable abnormalities of ventricular mechanics are those related to diastolic filling. These may include increased filling pressures, delayed pressure fall within the ventricle, and a greater than normal dependence on the effects of atrial contraction for ventricular filling. Such cases form the minority of cases of heart failure (estimates vary from a few per cent to about one-fifth of cases), but are seen with increasing frequency in older patients in whom senile myocardial fibrosis occurs more frequently as the major pathology underlying the heart failure. Other rarer causes include hypertrophic cardiomyopathy, infiltrative conditions such as amyloid heart disease, and the acute effects of ischaemia or the chronic effects of advanced hypertrophy in response to hypertension.

Diastolic dysfunction can be quantified by a variety of measurements: haemodynamic; echocardiographic; radionuclear; or ventriculographic. Those most commonly employed are the rate constant of isovolumic relaxation of the ventricle during early diastole (tau), the early to late peak filling velocity ratio (*E/A*) across the mitral valve on Doppler echocardiography, and the peak rate of ventricular filling on radionuclear gated acquisition (MUGA) scans in end-diastolic volumes per second. None of these parameters are independent of the loading conditions of the ventricle, nor of atrioventricular delay and heart rate, nor of the effect of systolic dysfunction.

Pure diastolic dysfunction is rare, as indeed is pure systolic dysfunction, as the two are almost inseparably interdependent. One can speak, however, of cases where the heart failure is predominantly due to systolic or diastolic impairment of the ventricle, and the simplest separation is via the size of the end-diastolic volume; if large, systolic dysfunction is likely to be the major abnormality; if small, diastolic. As will be discussed, this differentiation is important because of differing effects of treatment, in particular vasodilators, which may be less useful in diastolic dysfunction because of the requirement for high ventricular filling pressures in this condition.

Non-cardiac

General syndrome

Although initiated by ventricular dysfunction, in its chronic form heart failure is a multisystem disorder: the syndrome of chronic heart failure. The causes of many of the disparate organ pathologies that develop are poorly understood, as are the mechanisms by which these are (slowly) corrected by effective therapy, including transplantation of the heart. Much remains uncertain about this non-cardiac pathology and pathophysiology, including its genesis, symptomatic effects, and correct management. Evidence suggests, however, that non-cardiac factors become responsible both for the symptoms and the objective limitation of exercise capacity in chronic heart failure.

Specific organ systems

The microvasculature

Changes occur in the microvasculature in many organ systems and these may contribute to the organ underperfusion seen in this syndrome. There have been few reports of definite structural changes in the microvasculature, but functionally the endothelial-dependent vasomotor control systems are disordered. The endothelial-dependent vasodilator system is impaired both in the myocardial vessels and in the periphery. Tumour necrosis factor-alpha, which is elevated in some cases of chronic heart failure, has been implicated in impaired endothelial vasodilator function in addition to the enhanced activity of the endothelin vasoconstrictor system. This generalized endothelial dysfunction may contribute to some of the organ dysfunction described below, including renal, hepatic, and pulmonary vascular impairment. Specific treatments for endothelial abnormalities have not been established for heart failure, although promising results have been seen with improved endothelial function after localized exercise training or administration of the nitric oxide precursor L-arginine.

Large arterial function

In heart failure there is a reduction in large arterial compliance, which in turns leads to an increase in the impedance to ventricular outflow. Thus the efficiency of ventriculoaortic coupling is reduced, the impaired ventricular reserve is further stressed, and there is an increase in myocardial wall stress. The cause of the changes in large arteries probably relate to sympathetic and possibly local renin–angiotensin activation. In acute heart failure counterpulsation by intra-aortic balloon pumping probably helps forward aortic blood flow, at least partially by a mechanism involving improved ventriculoaortic coupling, in addition to the beneficial effects of enhanced diastolic coronary perfusion pressures.

The respiratory system

The lungs

Despite the frequency of dyspnoea as a central complaint of a patient with heart failure, relatively little is known of the role of the lung in chronic heart failure. In acute heart failure, changes within the lung are profound and easily explain much of the acute respiratory distress of the syndrome. With an acute reduction in left ventricular performance a rapid increase in left ventricular filling pressures, and hence pulmonary venous pressures, will lead to fluid accumulation in the lung parenchyma. Initially this will decrease the compliance of the lung, thereby reducing vital capacity and increasing the work of breathing. It may also, via oedematous swelling of the bronchial mucosa, cause a non-asthmatic bronchial constriction that can mimic asthma and further increase respiratory muscle work. With more severe pulmonary venous hypertension the alveolar membrane becomes thickened and oedematous and this may impair gas exchange, leading to an increase in the alveolar–arterial oxygen gradient and eventually arterial hypoxaemia. Eventually frank pulmonary oedema can form, further exacerbating the processes described above and leading to the clinical picture of gross dyspnoea, hypoxaemia, lung crepitations, and the production of copious quantities of pink frothy sputum (the alveolar oedema fluid itself).

In chronic heart failure, the patient is dyspnoeic but the changes in the lungs are far less marked. Pulmonary venous pressures may be normal if diuretic treatment is effective, and in well-diuresed and non-oedematous patients very few changes can be detected in lung histology. The changes of pulmonary siderosis seen with chronic untreated mitral stenosis are not seen in well-treated chronic heart failure cases. There have been reports of subtle changes in lung function in chronic heart failure, including a reduction in gas diffusing capacity, intermittent non-asthmatic bronchial constriction, and a purported increase in dead-space ventilation, but these are largely functional changes without an established anatomical cause. One pathophysiological change that can lead to respiratory distress is an alteration in the volume, structure, strength, and fatiguability of the respiratory musculature. The effects of these changes on the sensation of dyspnoea, or most appropriate therapy, are unknown. Similar changes are seen in skeletal muscle (see below).

Pulmonary oedema is not synonymous with heart failure. In addition to the acute respiratory distress syndrome (see Chapter 16.5.1) there are other causes of pulmonary oedema that need to be considered as differential diagnoses for heart failure: these are discussed in Section 15.15.

Respiratory control

The mechanisms of normal control of ventilation during exercise are not fully understood. It is not surprising, therefore, that the mechanisms underlying the abnormal respiratory response seen in chronic heart failure are also unclear. Patients with heart failure, even in the absence of pulmonary oedema, have an increased ventilatory response to exercise, whilst maintaining normal arterial blood gas tensions. They show reduced maximal oxygen consumption, an early dependence on anaerobic metabolism, and an increased ventilatory equivalent for carbon dioxide even at low work levels. This latter feature can be best appreciated by the plot of ventilation against the rate of carbon dioxide production (the $\dot{V}us.\dot{V}CO_2$ slope) during progressive exercise: this is significantly steeper (up to threefold) throughout both aerobic and anaerobic levels of exercise, and its steepness correlates closely with the reduction of maximal oxygen consumption. Although it is clear that this increased ventilation relative to the external work rate must indicate wasted ventilation, exactly why this occurs in non-oedematous patients is not certain. It has been assumed that there is a primary increase in dead-space ventilation due to a reduction in the ability of the right ventricle to perfuse adequately all lung regions, or the development of significant ventilation/perfusion mismatching within the lung, but these hypotheses have not been proven. An alternative hypothesis is that something other than the rate of carbon dioxide production causes the increased exertional ventilation in patients with heart failure, which is supported by the finding that, rather than being abnormal, arterial carbon dioxide during exercise is often lower than in normals, suggesting relative hyperventilation and the action of a non-carbon dioxide ventilatory stimulus. There are several candidate stimuli including an increased release of, or sensitivity to, known ventilatory stimuli such as lactate, arterial potassium, or adenosine. Skeletal muscle is abnormal in heart failure cases (see below) and releases metabolites earlier in exercise than age-matched normal controls.

There is also a neural pathway (the ergoreflex or metaboreflux) in the control of ventilation, utilizing group III and IV afferents from skeletal muscle. These are sensitive to the metabolic state of exercising muscle and transmit signals via the lateral spinothalamic tract to mediate reflex increases in ventilation as well as peripheral vasoconstriction and sympathoexcitation. Both this reflex control system and the arterial chemoreflexes that control ventilatory effort are abnormally active and oversensitive in chronic heart failure, possibly contributing to the excessive ventilation during exercise and playing a role in causing the subjective dyspnoea during low-level exercise seen in cases of chronic heart failure.

Airflow

Expiratory airflow can be restricted in patients with heart failure, even when all smokers and patients with a history of intermittent bronchospasm have been excluded. These patients can exhibit considerable dips in their peak expiratory flow rate on occasion, especially at night, which may lead to episodes of respiratory distress as well as adding to the work of breathing, and through that to the perception of dyspnoea. The mechanisms of this 'bronchoconstriction' are not known but they may involve oedema of the bronchial mucosa. Thus, a variety of lung factors can add together to contribute to dyspnoea in chronic heart failure, even in the absence of frank pulmonary oedema. Recently methoxamine in an opening of airways has been reported to improve peak flow rates and lead to an increase in exercise tolerance in these patients.

Gas exchange

Although arterial oxygen desaturation and carbon dioxide retention are rare in well-diuresed patients with heart failure, a more mild alteration in the gas exchange function of the lung can occur. These factors could reduce the rate of delivery of oxygen to the metabolizing tissues and act as a stimulus to increased ventilation. They may also explain the compensatory increase in arterial oxygen content seen in chronic heart failure if mild but intermittent hypoxia develops in this condition. This could also explain the beneficial effects of oxygen supplementation, even acutely, on exercise tolerance in patients with chronic heart failure.

It is not certain, however, that in chronic heart failure a reduction in diffusing capacity is either quantitatively important, or that oxygen supplementation works via increasing net oxygen delivery to the tissues. Alternative explanations are that the effect of a high inspired concentration of oxygen is non-specific in reducing peripheral chemoreflex drive and thereby relieving the sensation of dyspnoea, in a way akin to that produced by narcotic analgesics. Similarly, reduced gas exchange, especially for oxygen, may have more to do with inadequate expansion of the pulmonary capillary network and an inadequate time for gas transfer rather than to any alteration in the alveolar blood–gas barrier itself. The very low mixed venous oxygen saturations seen in chronic heart failure may mean that even a normal capillary transit time is inadequate for full oxygen exchange.

The sleep-apnoea syndrome

Nocturnal oxygen saturation monitoring of patients with chronic heart failure has demonstrated the presence of episodes of desaturation, often to below 80 to 85 per cent. These episodes coincide with and are caused by episodes of apnoea; they are also followed by semi-arousal from sleep and hyperventilation that may awaken and frighten the sleeping partner. In addition to obstructive sleep apnoea, which is common in patients with heart disease due to similar antecedent risk factors, the pattern is reminiscent of the Cheyne–Stokes respiratory patterns that are well recognized in patients with severe heart failure.

The mechanisms of both abnormalities of respiratory rhythm are incompletely understood. In some cases of nocturnal desaturations there is an obstructive element with obesity and pharyngeal occlusion by the tongue flopping back. In other cases there appears to be an alteration in the central sensitivity to carbon dioxide so that oscillating levels of respiratory drive, and hence of arterial oxygen saturation, develop. This second mechanism may be partly the cause for Cheyne–Stokes breathing as well. Another finding that may be related is that patients with chronic heart failure exhibit reduced total and high-frequency heart rate variability, but relatively enhanced variability of heart rate at very low frequencies (less than 0.01 Hz, or 1 cycle every 100 s). Although rhythmic variations in heart rate at higher frequencies are related to homeostatic mechanisms controlling blood pressure, in particular the vagal and sympathetic limbs of the arterial baroreflex, the genesis of this very low-frequency rhythm is not known. It does, however, have several features to suggest that chemoreflex activity may play a role in its genesis. First, this rhythm is particularly prominent in heart failure, where the circulation time is long; second, it is of similar frequency to the more obvious rhythm of Cheyne–Stokes breathing; and third, the chemoreflex loop has sufficient delay characteristics and possesses sufficient interactions with the baroreflexes and control of heart rate for a harmonic of oscillatory arterial gas concentrations to set up a similar harmonic oscillation in respiration, which would then entrain heart rate via an effect of the baroreflex. Finally, similar rhythms are particularly prominent in heart failure in pulmonary arterial pressure tracings, hence it may be that periodic sleep-apnoea (at least that which is not obviously obstructive), very low-frequency rhythms of heart rate variability, and Cheyne–Stokes respiration may all be reflections of harmonic oscillation of chemoreflex–baroreflex interactions. If this is the case then they may respond to therapies that alter chemoreflex gain or drive, and in this regard the promising reports of nocturnal oxygen supplementation and of nasal positive-pressure ventilation may be supportive for this theory. It is in any case surprising, and perhaps chastening, to note that in a condition so associated with dyspnoea that the state of the chemoreflex drive at rest, during sleep, or exercise is not known.

Musculoskeletal

Structure

Skeletal muscle biopsies in patients with moderate to severe chronic heart failure have shown a variety of pathological changes. These include individual fibre atrophy, a shift in the distribution frequency of types IIa and IIb fibres, and changes at an ultrastructural level, including a reduction in mitochondrial density, volume, and the number of cristae. It has proved impossible to define a specific pathological change characteristic of heart failure, partly because of the enormous variation in a control population, but also because muscle becomes abnormal in a limited number of ways in a variety of diseases associated with skeletal myopathy. It is also important to be sure of studying a specific heart failure-related change and not some subclinical skeletal myopathy as part of an inherited cardioskeletal myopathy. That such changes are seen with equal frequency and severity in ischaemic heart failure makes this unlikely to be the only explanation of the findings.

One of the most marked structural changes in peripheral muscle is the substantial reduction in total skeletal muscle bulk. Although it has long been recognized that in some cases of end-stage heart failure a catabolic wasting syndrome can develop (cardiac cachexia), it has only recently been stressed that more subtle evidence of muscle wasting may be both common and functionally important in chronic heart failure. If less muscle is available to do the work of the limb then each fibre will be more easily fatigued, be able to accept a lower total blood flow, and will appear metabolically more stressed and require anaerobic metabolism at an earlier point in exercise. These findings have all been taken as indicators of a deficiency in blood and oxygen delivery, rather than the alternative explanation that there simply is too little muscle for exercise to be performed efficiently.

Function

In chronic heart failure there is a reduction in the peak strength of both small and large muscle groups. In the case of the small muscles of the hand this clearly cannot be due solely to a reduced cardiac pumping capacity because of their tiny blood flow requirement. This suggests that there may be inherent defects in the quality of the muscle itself, but given the difficulty of exactly matching for active muscle bulk the difference may, however, be partly a reflection of muscle wasting.

In addition to reduced peak strength, there is an early fatiguability of muscle in heart failure. As a result, patients frequently complain that muscle fatigue is the major limitation to the performance of their daily tasks, and weakness and fatigue may both contribute to reduced physical activity that may induce physical deconditioning and further muscle wasting and dysfunction.

Metabolism

Skeletal muscle metabolism during exercise has been investigated by magnetic resonance spectroscopy. This technique allows an exploration of the rate of utilization of high-energy phosphate bonds associated with phosphocreatine and of intracellular pH, and through these the efficiency of aerobic and anaerobic metabolism within the muscle. These experiments have shown that there is an early depletion of phosphocreatine, an early acidification and accumulation of inorganic phosphate, a reduction in the rate of resynthesis of phosphocreatine, and in the removal of adenosine diphosphate (**ADP**). These changes cannot be explained by the acute effect of impaired blood flow, because the difference between normal controls and heart failure patients is seen even when both groups perform exercise in ischaemic conditions produced by regional circulatory occlusion. The metabolic abnormalities described by magnetic resonance spectroscopy probably reflect alterations in the oxidative enzymatic content of skeletal muscle described in biopsy studies. The causes of these metabolic changes are not understood, but it has been estimated that muscle wasting alone cannot explain them, because the half-time of ADP removal is independent of both the workload per unit muscle mass and the blood flow.

The only treatment that has been definitely shown to correct these metabolic abnormalities is physical exercise conditioning of the muscle, either localized or general. The time course of the possible correction of the muscle changes after cardiac transplantation has not been determined, nor has any definite effect of ACE-inhibitor treatment been described.

Autonomic and neuroendocrine systems

Much has been written about the importance of neuroendocrine activation in chronic heart failure, partly because of the established benefits of blocking two aspects of this with the ACE inhibitors and b-blockers. There is an undoubted activation of neuroendocrine systems involved in the 'fight or flight' reaction. These probably evolved, in a teleological sense, as a way of compensating for blood or fluid loss or sodium depletion, but in heart failure, although initially helping to support the circulation, continuous activation may be harmful. The neuroendocrine systems involved include the renin–angiotensin–aldosterone system, the sympathetic nervous system, and the vasopressin system, as well as that of the counteracting cardiac natriuretic peptides. Simultaneously with neuroendocrine activation there is a reduction in vasodilator influences and in vagal tone which, when maintained chronically, may be harmful. Adverse consequences have been described such as organ hypoperfusion, myocardial toxicity, an increased susceptibility to ventricular arrhythmias, and a possible progression of the underlying disease process, whether it be myocardial ischaemia or cardiomyopathy.

The renin–angiotensin–aldosterone system

In untreated heart failure there is mild activation of the renin system, which is dramatically augmented by the first use of diuretics in the treatment of the heart failure. After that there is a reasonable relationship between the severity of the heart failure and further increases in circulating renin and angiotensin II levels. In addition, all the components of the circulating renin–angiotensin system also exist in tissue sites and there is probably activation of these local tissue systems in the heart, kidney, brain, and blood vessel walls. The role and effects of these in health and in the progression of heart failure are unknown, but some of the beneficial effects of ACE inhibition described in other sections stress how important these systems may be in the syndrome of chronic heart failure.

At an organ level, the effects of elevated local and circulating angiotensin II can be very profound. In the kidney it can cause either a preservation of glomerular filtration rate (**GFR**) in the presence of low arterial pressure, or a reduced renal blood flow and GFR if the kidney is already dependent on angiotensin II-mediated efferent arteriolar constriction to maintain an adequate filtration pressure in the glomerulus. Such dependence can be seen in renal artery stenosis. In the heart, local increases in angiotensin II can cause coronary vasoconstriction and toxic effects on the myocytes, and in the periphery local angiotensin activation can elevate

systemic vascular resistance and thereby increase the afterload to the failing heart.

The clinical effects of inhibition of the renin–angiotensin–aldosterone system are dealt with more fully in other chapters, but it is important to note that we still do not know how they mediate their beneficial effects, whether by reduced circulating or tissue-based angiotensin II, or by augmentation of bradykinin or other kinin systems.

The autonomic nervous system

Early in the progression of heart failure from mild asymptomatic left ventricular dysfunction to the full clinical picture of chronic heart failure there is an activation of the sympathetic nervous system and a concomitant reduction in resting vagal tone. These changes are further enhanced by the administration of diuretics. There is no clear mechanism for either the activation of the sympathetic system in mild heart failure, or to explain why the activation should persist and progress in the chronic syndrome. In severe heart failure there may be a reduction in blood pressure, but this is often the result of aggressive therapy. By contrast, in asymptomatic left ventricular dysfunction, or mild heart failure, sympathetic activation commences at a stage when there is no perceptible change in blood pressure. It has been said that the activation is secondary to the withdrawal of the chronic sympathoinhibitory effects of the arterial baroreflexes, but there are flaws in this explanation. Even complete denervation of the baroreceptors does not lead to such persistent sympathoexcitation as seen in chronic heart failure, and in heart failure it also begs the question as to what caused the baroreceptor inhibition in the first place. If it is thought to be sympathetic activation, as seems likely, then we are left with a circular argument. No significant sympathoexcitatory influence has been demonstrated to underlie the very high levels of sympathetic tone in established heart failure. Two candidate mechanisms that have received little attention are the skeletal muscle ergoreceptor system and an interaction between the arterial chemoreflex and the baroreflex and cardiovascular autonomic centres. Both the ergoreflex and the chemoreflex cause sympathetic activation and may be abnormal throughout the progression of chronic heart failure.

The investigation of sympathovagal balance is limited by the lack of precise and quantifiable methods. Apart from a measurement of plasma norepinephrine (noradrenaline) levels there is no easily available clinical test for the activity of the sympathetic limb, and for the vagal limb the problem is even more difficult. Analysis of variations of heart rate variability has identified characteristic frequency harmonic oscillations in cardiovascular parameters, the relative oscillatory power of which show promise in the estimation of sympathovagal balance. The pattern in heart failure is very abnormal, with a dramatic reduction in total heart rate variability and a selective loss of the higher frequency (predominantly vagally mediated) rhythm characteristic of respiratory sinus arrhythmia, and a relative preservation of low- and very low-frequency rhythms which have their genesis more in the action of the sympathetic (low frequency) and renin–angiotensin or chemoreflex systems (very low frequency). Analysis of total heart rate variability, and in particular of individual frequency components, has shown that the pattern seen in heart failure is one associated with a high risk for the development of unstable ventricular arrhythmias and cardiac sudden death, although why this should be the case is not certain.

Beta-receptor function

With chronic sympathetic activation there is depletion of myocardial catecholamine stores and a downregulation of b1-receptors on the myocardium. There is also a decoupling of receptors from the postreceptor response, all of which lead to a loss of myocardial response to increased sympathetic drive. Clinically this manifests as chronotropic incompetence, loss of response to sympathomimetic stimulation, and a further impairment of exercise tolerance. Specific treatments are few, but there has been some improvement after b-blockade, ACE inhibition, and even very short-duration intermittent sympathomimetic stimulation.

The natriuretic peptide systems

The atria and ventricles contain granulated cells that release peptides, atrial natriuretic peptide (**ANP** or **ANF**) and brain natriuretic peptide (**BNP**), in response to stretch. In addition, the vasculature of the heart and other organs is the site of production of a closely related peptide, C-type natriuretic peptide (**CNP**), the physiological role and normal modulation of which is less clearly understood. These peptides are natriuretic agents that also relax peripheral vasculature and thereby oppose the actions of the sympathetic and renin–angiotensin systems. There is an increased release of these peptides in chronic heart failure associated with cardiac enlargement, but the significance of the increased plasma levels is uncertain. The use of exogenous ANP or neutral endopeptidase inhibitors, to increase endogenous levels by inhibiting the breakdown of these peptides, have produced only minor natriuretic and haemodynamic effects in heart failure, although the effects may be greater if administered on the background of inhibition of the opposing renin–angiotensin–aldosterone system. BNP has been shown to be quite accurate in determining the degree of systolic dysfunction of the heart in selected hospital series, and there is such a close relationship between myocardial stretch and BNP release that estimation of BNP has been shown to quite accurately predict in population surveys which patients are likely to be diagnosed as suffering from heart failure.

The vasopressin system

Elevated plasma concentrations of vasopressin, also known as antidiuretic hormone (**ADH**), are found in chronic heart failure, but the importance of vasopressin in the pathophysiology of the condition is not certain. Its actions are a combination of haemodynamic, with profound arteriolar vasoconstriction increasing peripheral resistance, and renal, with an action on the collecting duct to increase reduce free water reabsorption and thereby cause antidiuresis.

Other hormonal systems

Abnormalities have been described in several other hormonal systems in chronic heart failure, but the significance of these changes is uncertain. Thyroid hormone metabolism is deranged, with an increase in reverse T3 similar to that seen in the so-called 'sick cell syndrome'. Plasma insulin levels are increased in heart failure, whether of ischaemic, valvular, or idiopathic aetiology, and this is associated with a decreased sensitivity to the glucose transport effects of insulin. Alterations in sex hormones and growth factors are likely in advanced cardiac cachexia, but to what extent these are specific to the syndrome of chronic heart failure is uncertain.

The kidney

Control of fluid and electrolyte balance is impaired in heart failure. This is due to the reduced renal perfusion pressure in advanced disease, to the effects on intrarenal haemodynamics of the neuroendocrine activation described above, and to the effects and side-effects of commonly prescribed medications.

Fluid overload and hence oedema is common in heart failure, and electrolyte disturbances are both common and important. In mild heart failure fluid retention is due to the effects of aldosterone, vasopressin and catecholaminergic renal vasoconstriction. The kidney itself is partly a passive organ, responding to neuroendocrine activation outside its control, but it is also an active endocrine and autocrine organ responding to reduced renal perfusion pressure in heart failure.

The renin–angiotensin–aldosterone system

The juxtaglomerular apparatus, adjacent to the distal convoluted tubule, senses a reduction in the rate of delivery of sodium to the distal tubule and releases renin in response, thereby playing an important part in the activation of the circulating renin–angiotensin system described above. Activation leads, via aldosterone, to an increased reabsorption of sodium in the distal renal tubules, and an additional effect of angiotensin on thirst and possibly salt hunger completes the response, all encouraging the retention of sodium and water in the body.

All the components of the renin–angiotensin system also exist within the kidney and there can be local autocrine activation with important effects on intrarenal haemodynamics. These may either increase or decrease GFR, depending on the level of renal perfusion pressure and other factors operating on the kidney, such as the renal sympathetic nerves and circulating vasoactive factors.

The kallikrein–kinin system

This second autocrine system of the kidney is less well studied because of the short half-life of some of its active components and the difficulty in isolating them. In simple terms the kinin system appears complementary to the renin–angiotensin system, causing vasodilatation where the latter causes vasoconstriction. It is also thought to be involved in the control of renal tubular function, but its precise role in heart failure and the effects of ACE inhibitors (which also block the enzyme that breaks down bradykinin) are unknown.

The cardiorenal syndrome

Some of the causes of heart failure, most notably atherosclerotic arterial disease and hypertension, can have direct effects on the kidney. This is one reason for an increased coexistence of cardiac and renal failure. Other less common conditions that can cause both organs to fail include amyloid, sarcoid, and certain vasculitidies. A more common finding is that an apparently reasonably well-functioning kidney can progressively fail in the presence of severe heart failure. This is partly the effect of hypovolaemia and low blood pressure (prerenal azotaemia), but is also due to the circulating and intrarenal neurohormonal systems described above, and the renal effects of drugs used in treatment. The net effect is that renal failure is an extremely common and clinically important complication of severe heart failure. Aggressive diuretic therapy can precipitate a significant worsening of renal function that then blunts their effectiveness. The ACE inhibitors usually lead to a small increase in serum creatinine concentrations, but in a few patients they can precipitate clinically important renal failure.

Electrolyte disturbances

Electrolyte disturbances are not common in untreated heart failure, except when this is severe. There is an initial retention of sodium and a loss of potassium due to the effects of increased levels of aldosterone. Later, and especially after diuretics have been administered, there is a further depletion of potassium and also of magnesium. Both these disturbances may be important in generating cardiac arrhythmias, especially in digoxin toxicity. A dilutional hyponatraemia can develop in patients with severe heart failure: this can be both difficult to manage and is a poor prognostic sign.

The careful management of fluid and electrolyte balance in patients with heart failure can lessen renal complications and improve both symptoms and prognosis. There is no simple therapeutic regime that will ensure this, just careful repeated clinical examination and monitoring, judicious use of drugs, with knowledge of their potential side-effects.

Haematological system

Haemoglobin

Increased haemoglobin content has been described as an adaptive response in heart failure. This may be secondary to chronic tissue hypoxia, perhaps most importantly in the kidney in which it may stimulate an increase in erythropoietin production. It is doubtful if the increase in haemoglobin is very effective in increasing oxygen delivery to the tissues, for if the haematocrit increases too much then the resulting increased whole blood viscosity will reduce net tissue perfusion by increasing the resistance to blood flow. In more severe chronic heart failure anaemia becomes more common. It is similar to the anaemia seen in other chronic illnesses.

Other haematological changes

Impaired clotting factor production can result from hepatic dysfunction (see below) and resulting abnormalities in haemostatic function are not uncommon. The white blood cell count may be mildly elevated in heart failure as part of a more generalized, but poorly understood, immune activation.

Other organ systems

The liver

In heart failure the liver can be affected by an increased venous back-pressure, by an impaired arterial supply, and by the metabolic complications of the syndrome. The underlying process that leads to heart failure, for example alcohol excess or haemachromatosis, can also affect it.

The most common hepatic abnormality in chronic heart failure is congestion due to the effects of right heart failure on venous pressures. This leads to increased venous engorgement of the liver and can result in a noticeable increase in hepatic size, local tenderness, and minor derangements in liver function, causing modest increases in transaminase levels in its mildest form. In more severe cases, nausea and right hypochondrial discomfort develop, and in severe cases jaundice, impaired albumin and clotting factor production, and malabsorption of fats may result. These changes can have clinically important effects on clotting, especially as warfarin is commonly prescribed, and also on the hepatic metabolism of certain drugs. The nausea and malabsorption can worsen the catabolic state of the patient and can contribute to the wasting seen in cardiac cachexia. There is no specific treatment for this complication of chronic heart failure, other than the correct dosage of diuretics and the maximization of cardiac function with vasodilators.

Gastrointestinal tract

This is mainly affected by the increased venous pressure of right heart failure. Intestinal mucosal oedema can contribute to malabsorption and possibly nausea. Cardiac conditions are also associated with a higher rate of intestinal angiodysplasia: this can lead to recurrent blood loss, which can be a considerable management problem in the patient who requires anticoagulation.

Central nervous system

Certain conditions that cause heart failure can also produce neurological effects: these include alcoholism, amyloid, and heavy metal poisoning. Apart from the abnormalities of autonomic and neuroendocrine function described earlier, specific neural complications of heart failure are not common.

Immune function

Immune abnormalities in heart failure include both an excess of proinflammatory cytokines and a deficiency of inhibitory and immunomodulatory cytokines. Some cytokines are released in excessive amounts in both acute heart failure and in cases of relapsing or severe heart failure. Important amongst these is tumour necrosis factor-alpha (**TNF-a**), important because it has been shown to produce harmful effects that are common in chronic heart failure. These include endothelial dysfunction, myocyte depression, necrosis and apoptosis (programmed cell death), and loss of skeletal muscle mass. This cytokine has been implicated in the generation of severe wasting, both in cancer-related cachexia and in the syndrome of cardiac cachexia. Although the exact pathogenic mechanisms for the high cytokine levels in chronic heart failure are unknown, several theories have been proposed. These include the possibility that the failing heart itself is a site of production, or that recurrent bacterial loads due to bowel-wall oedema and bacterial-product translocation to the circulation cause an endotoxaemia that is known to be able to stimulate cytotoxic cytokine release. TNF-a has the properties to be harmful in heart failure and studies are underway to evaluate whether antitumour necrosis-factor strategies, such as the use of monoclonal antibodies or subantibody fusion-protein fragments, which neutralize tumour necrosis factor, would benefit the clinical course of patients with advanced chronic heart failure.

Cause of non-cardiac pathophysiology in heart failure

The changes described in different organ systems as part of the syndrome of chronic heart failure remain largely unexplained. We have proposed a 'muscle hypothesis' in which we explain the generation of many of these abnormalities via the combined effect of physical deconditioning and metabolic dysfunction, combining a release of catabolic factors with a loss of normal anabolic function. Figure 1 describes the general pathway by which these changes could lead to skeletal and respiratory muscle abnormalities, and via these to fatigue, dyspnoea, exercise limitation, and sympathoexcitation.

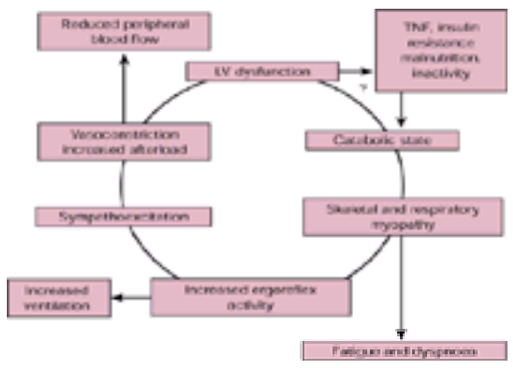

Fig. 1 The Muscle hypothesis in chronic heart failure. A proposal to explain the genesis and effects of several components of the non-cardiac pathophysiology of heart failure. As proposed by Coats *et al*. 1996. TNF, tumour necrosis factor.

Clinical assessment

The assessment of a patient with heart failure requires a careful history and examination, both at initial presentation and when assessing progress and response to treatment. Confirmation of heart failure can be aided by chest radiography, echocardiography, and, in selected cases, by cardiac catheterization, radionuclide ventriculography, and other imaging modalities. Invasive haemodynamic monitoring has a role in the assessment of acute severe heart failure. The use of a biochemical test (such as estimation of brain natriuretic peptide, BNP) to detect heart failure shows promise, but is not yet recommended in routine clinical practice.

Cardiopulmonary exercise testing with respiratory gas analysis can help to establish the cause of symptoms in patients with coexisting heart and lung disease, and to determine whether heart failure is causing the symptoms limiting the patient. When the respiratory exchange ratio (CO_2 produced per unit of oxygen consumed) exceeds 1.0, muscle metabolism has become anaerobic, indicating that a point of limiting cardiac reserve has been approached. If it does not exceed 1.0 at peak exercise then the true cardiac limitation cannot be assessed. Significant hypoxaemia and/or hypercapnia on exercise are rare in non-oedematous chronic heart failure, and when present suggests the limiting factor is pulmonary or (less commonly) a right to left shunt.

Regular assessment of the severity of heart failure is usually made by a history of symptomatic limitation and by clinical examination. The occasional use of chest radiography, echocardiography, and cardiopulmonary exercise testing can also help. Repeated haemodynamic monitoring has little role in chronic heart failure. Whether monitoring levels of plasma norepinephrine (noradrenaline) or atrial natriuretic peptides would materially assist in clinical management is uncertain: regular assessment of clinical biochemistry is certainly essential.

Symptoms

Heart failure is defined as symptomatic left ventricular dysfunction. As a result classical symptoms must be present to make the diagnosis. These most commonly are dyspnoea or fatigue, but initial presentation can be collapse, syncope, oedema, chest pain, or palpitations (Table 2).

Clinical examination

The physical examination of any patient with suspected heart failure is important. It can be a way of detecting the cause of the heart failure, such as the detection of aortic stenosis, leading to a definitive treatment option. Most cases of treated heart failure do not demonstrate florid signs such as seen in acute or decompensated heart failure, hence skilled clinical examination is necessary to obtain all the available clues as to cause, severity, and complicating factors (Table 3).

Investigations

Investigations may be necessary to confirm the diagnosis of heart failure, to establish the cause, and to stratify for the risk of complications or deterioration. A chest radiograph, an electrocardiogram, and an assessment of full blood count, urea, and electrolytes are essential in all cases of suspected heart failure. The electrocardiogram may give specific evidence of ischaemia/infarction, left ventricular hypertrophy, arrhythmia, and other causes of pathological Q-waves, and can combined with stress testing to detect reversible myocardial ischaemia.

Many studies have demonstrated the inaccuracy of diagnosis and assessment of the cause of heart failure in primary-care settings without the use of more specialized investigations, with up to one-third of cases being misclassified. As a result it is now a recommendation that all suspected cases of heart failure should have, if at all possible with local resources, an echocardiogram to assess the nature and extent of ventricular dysfunction and to detect those causes of heart failure identifiable by echocardiography. It can be used to detect and grade valvular disease, for the differentiation of a globally impaired left ventricle (for example, dilated cardiomyopathy) from segmental dysfunction (for example, ischaemic heart disease), and may also reveal ventricular hypertrophy, aneurysms, amyloid and other infiltrates, and specific forms of cardiomyopathy.

Specific blood tests can be used to check for certain rare causes of cardiac failure: hypocalcaemic cardiomyopathy; thyroid heart disease (hyper- or hypothyroidism); iron storage diseases; anaemia; heavy metal poisons; amyloid (serum electrophoresis); and sarcoid (serum ACE). Occasionally coronary angiography, and rarely ventricular biopsy, may be indicated in cases where the presentation is at a young age or specific features suggest a high likelihood of a treatable condition.

Certain investigations may be useful in assessing the severity of ventricular dysfunction in addition to or as alternatives to echocardiography, for example rest or exercise radionuclide ventriculography. Monitoring by 24-h Holter electrocardiography to detect ventricular arrhythmias and blood tests for associated renal or liver dysfunction or electrolyte disturbances can be used to give a rough guide to prognosis, although this is rarely reliable on an individual patient basis.

Other specialized investigations such as cardiopulmonary exercise testing, CT or MR scanning, cardiac catheterization, and myocardial biopsies are covered in other chapters to which the reader is referred for more detailed explanation.

Treatment

The major elements of treatment of heart failure are listed in Box 1 and Box 2. This has undergone a revolution in the last two decades, with substantial treatment efficacy established first for the ACE inhibitors and more latterly for b-blockers (a group previously considered absolutely contraindicated in heart failure) and the aldosterone antagonist spironolactone.

Box 1 Treatment of chronic heart failure

General	No added salt	Treat hypertension
	Maintain optimal weight	Detect alcohol abuse
	Stop smoking	Prevent coronary disease
	Encourage exercise	
Mild	Thiazide/loop diuretic	b-blocker

	ACE inhibitor	
	Digoxin if atrial fibrillation	
Moderate	Loop diuretic	b-blocker
	ACE inhibitor	
	Combine diuretics	
Severe	Increase loop diuretic	Spironolactone
	Combine diuretics	? b-blocker
	Metolazone	? digoxin
	ACE inhibitor	Transplant

Box 2 Options in the treatment of severe chronic heart failure

1. Drugs	Diuretics	Loop, thiazide, and potassium-sparing combination
	ACE inhibitors	
	Vasodilators	Nitrates, hydralazine
	Positive inotropes	Digoxin, IV intermittent inotrope
	Anticoagulants	
	b-blockers, calcium antagonists, or anti-arrhythmics	
2. Implantable cardiac defibrillator—ICD		
3. Haemofiltration		
4. Peritoneal dialysis or haemodialysis		
5. Aortic balloon pump or ventricular assist device		
6. Transplantation or cardiomyoplasty		

ACE, angiotensin-converting enzyme; IV, intravenous.

Diuretics

Diuretics remain the mainstay of the management of oedema in heart failure and are often the first agents to be used in new cases. This is not because their role for this indication has been proven, but rather they are used to treat what is frequently the first manifestation of heart failure, peripheral or pulmonary oedema. In acute heart failure intravenous loop diuretics lead to a dramatic and rapid improvement in condition, and in almost all patients with moderate or severe heart failure diuretics will be essential for adequate symptom control. Concern has been expressed about the potential adverse effects of diuretic agents, including activation of the sympathetic and renin–angiotensin systems, but until an alternative mechanism for the control of oedema fluid is achieved there remains no viable alternative.

In the modern era of evidence-based therapies diuretic use remains an example of the art of medicine surviving despite a marked lack of proof. No single placebo-controlled trial of a loop diuretic has proven a convincing improvement in either mortality or morbidity rates: the treatment was introduced and popularized before the modern era of randomized controlled trials. The only significant treatment trial of diuretics was with a low dose of the weak diuretic spironolactone 25–50 mg per day, which showed a significant reduction in total mortality in severe heart failure in the RALES study published in 1999. This effect was probably due to the inhibition of the harmful effects of the neurohormonal factor aldosterone than to the diuretic effect of spironolactone.

Initially the thiazide diuretics may be sufficient in patients with mild heart failure, but in more severe cases of heart failure one of the loop diuretics—furosemide (frusemide), bumetanide, or torasemide—will be necessary. These are very familiar agents, particularly furosemide, but there remains some confusion about the best mode of treatment. Furosemide and bumetanide both give an acute and relatively short-acting diuresis that some patients find disabling. Others actually prefer this, as they can time their outings to avoid periods of diuresis. The newer torasemide has a much more prolonged action over 24 h and the increase in urine flow is said to be much less obvious to the patient.

Initially 40 mg of furosemide or its equivalent (about 1 mg of bumetanide) may be sufficient to control oedema, but some patients will need much higher doses (e.g. furosemide 80 or 120 mg twice daily) for oedema control. A better alternative to increasing the dose of loop diuretic is to use combination diuretics by adding agents with different modes of action, such as amiloride, thiazides, or spironolactone, or all three together. This is often far more effective than even extremely high-dose intravenous loop diuretics. The potassium-sparing drugs have the added advantage of ameliorating the loss of potassium produced by the other agents, although this is now less of a problem because the majority of patients with heart failure will be taking an ACE inhibitor, which has potassium-sparing effects.

In an acute exacerbation, switching to intravenous administration can boost response, as can a short period of bed rest, or the use of a short period of positive inotropic therapy with dopamine or dobutamine, especially as these have some renal vasodilator action.

It should also never be forgotten that diuretic therapy should go hand in hand with sodium and fluid restriction. The 'Chinese take-away syndrome', due to the effects of an acute sodium load and water retention, describes an episode of acute pulmonary oedema occurring several hours after a high sodium meal in a previously stable patient with chronic heart failure. This should remind us of the effects of excessive sodium intake.

In patients with severe heart failure fluid restriction may be necessary, but care should be taken not to produce dehydration and further deterioration in renal function. In the long term, fluid restriction should not be to less than 1500 ml per day.

Metolazone is a thiazide-like diuretic with a profound diuretic action when given on the background of loop-diuretic therapy. This combination therapy can be very powerful, often considerably stronger than the intravenous administration of furosemide alone, even in high doses. As little as 2.5 or 5 mg of metolazone given in this way can lead to several litres of extra urine output; this should never be started for the first time in an outpatient, but should be reserved for specialist hospital use, except under special instruction. Profound electrolyte disturbance can accompany this diuresis. A small number of outpatients with severe oedematous heart failure require chronic administration of metolazone, but often only 2.5 mg on alternate days or once or twice a week may be needed. Care should be taken to monitor electrolytes, urea, and creatinine very carefully in patients so treated.

Digoxin

Digoxin is the oldest drug therapy available for heart failure, and it still retains a place as the only safe chronically administered positive inotropic agent. Its use in patients with sinus rhythm remains controversial, for although it is considered as first-line standard therapy in the United States and many European countries, cardiologists in the United Kingdom point to the lack of any data demonstrating its long-term benefits in a large prospective placebo-controlled trial. The only definitive

large-scale trial of digoxin against placebo, the DIG trial, showed no significant effect on mortality, despite a significant reduction in the number of hospital admissions. With its narrow therapeutic window the routine use of digoxin in patients with sinus rhythm remains a matter of debate. In severe heart failure, where the patient remains very symptomatic, it may be worth trying: some respond with an improvement in symptomatic status. (See Chapter 15.5.1 for further information.)

Direct-acting vasodilators

The major agents in this class in regular use for heart failure are the combination of hydralazine and isosorbide dinitrate. Although an advance at the time, their use has been largely supplanted by the more effective ACE inhibitors. Other agents such a prazosin have proved less effective, and the calcium-antagonist vasodilators, particularly the short-acting dihydropyridine group, such as nifedipine, are significantly negatively inotropic and can worsen heart failure. Longer acting agents that cause less reflex tachycardia appear to be preferable, with amlodipine in particular seeming to be at least safe in chronic heart failure if its use is needed for its antianginal profile. The routine use of vasodilators in chronic heart failure cannot be supported by clinical trial data. (See Chapter 15.5.1 for further discussion.)

Angiotensin-converting enzyme (ACE) inhibitors

In heart failure

The introduction of ACE inhibitors has had a profound effect on the treatment of heart failure. Their benefits are not restricted to patients with end-stage heart failure, but extend also to patients with mild to moderate heart failure, and even to modifying the progression of the disease in patients with asymptomatic left ventricular dysfunction or extensive left ventricular dysfunction after myocardial infarction. The ACE inhibitor enalapril was the first agent to be shown to dramatically reduce mortality in severe heart failure: 253 patients in the CONSENSUS I study were randomly allocated to enalapril or control and, with an overall 6-month mortality in the placebo group of 44 per cent, there was a significant 40 per cent reduction in mortality in those randomized to enalapril. These patients were in end-stage heart failure, symptomatic at rest or on minimal effort, and this treatment was a clear therapeutic breakthrough, so much so that almost immediately ACE inhibitors became standard therapy in this situation.

Within a few years of the CONSENSUS I study similar evidence of benefit was shown for ACE inhibitors in the treatment of mild to moderate heart failure (New York Heart Association, NYHA class II and III), patients in whom moderate exertion led to symptoms, but in whom the 1-year mortality rate was 10 to 20 per cent rather than the 50 per cent or more of the CONSENSUS I study patients. Despite the lower overall mortality rate in these patients, the ACE inhibitor enalapril produced further significant reductions in mortality. The SOLVD treatment study looked at 2569 patients with mild to moderate heart failure randomized to enalapril or control and showed a statistically significant 16 per cent reduction in mortality. The V.HefT II study looked at 804 patients randomized between enalapril and the vasodilator combination of hydralazine and isosorbide dinitrate and found a significantly lower mortality with enalapril.

The prevention limb of the SOLVD trial, despite not showing any significant reduction in overall mortality, did show a reduction in the rate of progression of disease, with less new diagnoses of clinical heart failure, and a reduction in the rate of hospital admissions for heart failure. This suggests there may be both clinical and economic gains from the use of ACE inhibitors in this clinical situation.

See Chapter 15.5.1 for further discussion of the use of ACE inhibitors in heart failure.

After myocardial infarction

After a myocardial infarction the area of infarcted myocardium does not form a stable scar immediately, but rather undergoes a complex series of changes over weeks to months and even years (see Chapter 15.4.2). Some of these changes are beneficial and some are not, and this process may hold some of the clues to why patients after a myocardial infarction can develop heart failure months or even years later, without any evidence of further infarction.

In the first few days after the infarction there is an increased load on the residual myocardium, which undergoes compensatory hypertrophy, partly stimulated by activation of some of the neurohormonal pathways described earlier. The infarcted area is not protected from these influences, and as it is under increased mechanical stress as well, the overstimulation of the adjacent myocardium by the sympathetic and renin–angiotensin systems can lead to an even greater mechanical stress on the freshly infarcted region. These processes can lead to infarct expansion, a process whereby the infarcted wall is stretched and thinned by the mechanical stress exerted upon it, and to an apparent increase in the extent of infarcted myocardium expressed in absolute size or as a percentage of the total left ventricular wall. This should be distinguished from infarct extension, where previously living myocardium at the fringes of the infarcted area itself infarcts, leading to an increased size of wall motion abnormality.

Over a period of weeks to months and possibly years, a second process develops where the wall of the whole myocardium becomes thinned and enlarged. This affects the residual myocardium and involves realignment between myocytes, a process called 'cell slippage', leading to a progressive alteration in the shape of the ventricle, not requiring any further episodes of infarction. The shape of the ventricle becomes more globular and enlarges and the ventricle is said to have 'remodelled'. Very good animal experimental data and observational data on patients have shown that this remodelling process precedes and predicts the development of heart failure, and that the administration of an ACE inhibitor could delay or prevent left ventricular remodelling and reduce mortality in this setting.

The first large trial reported orally was the CONSENSUS II study. This indicated a non-significant trend towards an adverse effect on mortality when enalapril was given to relatively high-risk patients recovering from infarction. Therapy was commenced with intravenous enalaprilat within 24 h of the infarct, and included subjects with quite low blood pressure at entry. The trial was terminated early with no definite effect on mortality. This trial was rapidly followed by the SAVE trial, which studied captopril in a target dose of 50 mg thrice daily in patients recovering from a myocardial infarction. Unlike the CONSENSUS II study the patients were recruited after the initial infarct-healing phase had been completed, after most infarct expansion and scar formation had begun, but before the later remodelling process had become established. The patients were all thoroughly investigated, including documentation of significantly impaired left ventricular function by a radionuclide ventriculogram ejection fraction of 40 per cent or less, and all had undergone correction of clinically important residual myocardial ischaemia by either angioplasty or bypass surgery prior to entry to the trial. A total of 2231 patients were randomized between 3 and 16 days after infarction to receive captopril or placebo and followed for 42 months. After the first 6 months of follow-up the survival curves for the two groups separated and at the end of the trial there was a significant 19 per cent reduction in total mortality; there was also a 22 per cent decrease in the rate of hospital admission for heart failure.

Following the SAVE trial there was rapid confirmation of its results. The AIRE study, co-ordinated from Leeds, recruited 2006 patients who had clinical evidence of transient heart failure after a myocardial infarction, including radiological evidence of pulmonary oedema or chest crepitations, or the presence of a third heart sound on auscultation. Patients were randomized to ramipril (5 mg twice daily) or placebo between 3 and 10 days postinfarction: there was a significant 27 per cent reduction in mortality at 15 months of follow-up. The survival benefit appeared to commence within the first few weeks of follow-up.

These beneficial effects have been confirmed with a number of ACE inhibitors in different trial settings. The overwhelming conclusion from these studies is that ACE inhibitors beneficially affect the recovery process after a myocardial infarction, and in the longer term reduce mortality by preventing the progression to heart failure. Based on available evidence, the vast majority of patients either with heart failure or at high risk of developing heart failure, should be on long-term ACE inhibitor therapy. These large trials have shown benefits postinfarction and in heart failure, with a variety of ACE inhibitors, including enalapril, captopril, ramipril, lisinopril, and trandolapril. It would seem a reasonable conclusion that the benefit of the ACE inhibitors is largely a class effect.

b-Blockers

This is an interesting therapeutic area because everything from total b-receptor antagonists (for example, metoprolol), through partial agonists (xamoterol), to totally positive agonists has been tried or suggested for the treatment of heart failure. Until recently b-blockers were routinely prohibited for patients with heart failure. However, the last few years have seen a sequence of well-designed randomized controlled trials that have demonstrated a profound reduction in mortality, improvement in left ventricular function, and a reduction in the need for hospital admission in patients with mild, moderate, and severe heart failure. This has been shown for metoprolol in a slow-release preparation, bisoprolol, and the combined b- and a-receptor antagonist carvedilol, but it is too early to say if the beneficial effects are a class effect.

The difficulty of using b-blockers in significant heart failure should not, however, be underestimated. Patients were carefully selected in the major trials, almost always stable outpatients at the time of treatment initiation, and with recent episodes of oedema or decompensation. b-Blockade was commenced at very low initial starting doses and increased slowly under careful observation. In addition, treatment was commenced and monitored by physicians expert in the care of patients with heart

failure. Those with stable heart failure should not be denied the benefits of b-blockade, but only specialists in the care of patients with heart failure should start this therapy, which can sometimes be difficult. (See Chapter 15.5.1 for specific guidance.)

Antiarrhythmic agents

Ventricular arrhythmias are extremely common in those with heart failure. As sudden death is a common mode of demise for these patients it is tempting to think that antiarrhythmic therapy, which can suppress the ventricular arrhythmias, may reduce the incidence of sudden death. Unfortunately this approach has not proved to be effective, and many agents that have been tried appear to induce more sudden deaths than they prevent. The most promising drug is amiodarone, despite its formidable list of side-effects. The GESICA trial in South America, which included a high proportion of patients with Chagas' cardiomyopathy, even suggested a net reduction in mortality in heart failure patients regardless of the presence of ventricular arrhythmias, but subsequent trials in ischaemic left ventricular dysfunction failed to confirm this promise. By and large, unless symptomatic, non-sustained episodes of ventricular tachycardia are best left alone. The implantable defibrillator is more effective at reducing mortality in patients resuscitated from sudden death or those with frequent potential fatal ventricular tachycardia, but costs are high and they are not uniformly available to all patients who might benefit from their use. (See Chapter 15.6 for further discussion.)

Oral, positive inotropic agents

This group of drugs, including the phosphodiesterase inhibitors and calcium sensitizers, were heralded as a major advance when introduced into practice, but trial after trial comparing them against placebo have not only shown a loss of effect with time, but have suggested an increased mortality. With the exception of digoxin there is no safe, chronically administered, positively inotropic agent.

Anticoagulants and antiplatelet agents

There is clear evidence for the benefits of aspirin or other antiplatelet agents in patients recovering from a myocardial infarction. As most people with heart failure in the developed world appear to have extensive ischaemic heart disease, it is likely the majority will be treated with aspirin to reduce the chance of coronary arterial occlusion. Another indication for aspirin is in the prevention of cerebral embolism in chronic atrial fibrillation in patients with significantly impaired left ventricular function. Several studies have shown a positive effect of aspirin in this situation, although it is probably less effective than full anticoagulation with warfarin, although there are some patients in whom it would be preferable. This indication for aspirin would incorporate patients with dilated cardiomyopathy as well as those with extensive ischaemic heart disease. However, some as yet unconfirmed fears have arisen that aspirin may interfere with some of the beneficial effects of ACE inhibitors in heart failure, so it is not considered routine to use aspirin in all patients with this condition. Full anticoagulation is usually reserved for those with heart failure who also have chronic or regularly recurrent atrial fibrillation, who have suffered a prior thrombotic or embolic stroke, or who suffer from transient ischaemic attacks. (See Chapter 15.5.2 for further discussion.)

Angiotensin-II receptor antagonists

Following the great success of the angiotensin-converting enzyme inhibitors in heart failure, much was expected of a group of agents that specifically blocked the harmful effects of angiotensin-II at its main site of action, the AT-1 receptor. However, the trials published to date have failed to prove that these agents are superior to ACE inhibitors in terms of reducing mortality, nor can it yet be said with confidence whether they are as good as the older agents. They do appear to be better tolerated, with less likelihood of producing cough as a side-effect, but their role, if any, in the management of patients with heart failure remains unclear. The largest trial to date, ELITE-II, of approximately 2000 elderly patients with chronic heart failure, compared captopril 50 mg three times a day with losartan 50 mg once a day. There was no significant difference in mortality between the two treatments. The trial was not large enough to prove that the angiotensin-II receptor antagonist was equivalent or an alternative to the ACE inhibitors. (See Chapter 15.5.1 for further information.)

Non-pharmacological treatments

Patient education

Patients and their families are often confused and bewildered by the term 'heart failure'. Alternatives such as 'weak heart', 'congestion', or 'large heart' may be better at giving the correct impression as to the nature of the condition. In can be extremely useful for long-term adherence to treatment recommendations to spend some time explaining to the patient and partner some simple physiology of left ventricular dysfunction, the body's compensatory mechanisms, and why these lead to symptoms and signs that the patient may have already noted. The patient will then be much more aware of the need for diuretics and vasodilators, and the effects of alterations in fluid and salt intake, intercurrent illness, etc. This could improve oedema control and lessen the frequency with which a patient needs to attend the outpatient department or be admitted to hospital for stabilization. Simple measures such as information on low-salt diets, fluid restrictions, and monitoring daily weight at home can significantly improve long-term heart failure management.

Specialist heart failure clinics and outreach nursing services

Recent trials have suggested that there can be a reduction in the need for emergency hospital admissions if patients are enrolled into specialist services to help them, their relatives, and their general practitioners manage their heart failure after discharge from hospital. Important amongst these services is adequate education about the correct way to take their heart failure medication. Specialized nursing services with home visits and improved liaison between the primary care and secondary care of patients with heart failure appears to be particularly helpful in improving the quality of care.

Rest and exercise

There is very good evidence in acute heart failure, or in an acute decompensation in chronic heart failure, that bed rest can improve renal blood flow and the response to diuretics. This is presumably via a reduction in the level of stimulation of the sympathetic and renin–angiotensin systems. Admission to hospital for a few days' rest is thus a common treatment for heart failure, and one with a very long history. Initial enthusiasm for the benefits of longer periods of bed rest (weeks to months) as a management strategy for chronic heart failure and cardiomyopathy have not been borne out; in fact this practice is accompanied by the considerable and well-known complications of prolonged bed rest. On the contrary, benefits have been shown after exercise training in carefully selected patients with chronic heart failure. Improvements are seen in exercise tolerance, skeletal muscle and respiratory function, and in autonomic balance. This raises the possibility that profound physical deconditioning may be contributing to some of the pathophysiological changes described in the sections above. In a patient with stable chronic heart failure, with no evidence of exercise-induced ventricular arrhythmias, regular exercise should be encouraged rather than prohibited. The reader is referred to Chapter 15.5.3 for a discussion of the benefits that can be obtained from a careful and selected use of exercise training programmes in patients with stable chronic heart failure.

Other treatments

Cardiac transplantation and mechanical assist device therapies are described in Chapter 15.5.4.

Prognosis

In severe heart failure, where patients are symptomatic at rest (NYHA class IV), the prognosis is dire, with survival expected to be 1 year or less. The prognosis remains poor even in mild heart failure (class II–III), being comparable to that of many solid tissue malignancies with a mortality rate of between 20 and 30 per cent per year. Although major treatment advances have been achieved in mild, moderate, and even severe heart failure, these have led to only a very partial correction of the excess mortality associated with this condition.

Prognostic factors and markers

Many different parameters have been described as having prognostic value in patients with heart failure. It is important to differentiate between prognostic factors, which have a direct functional link to increased mortality, and prognostic markers which merely reflect a worse prognosis, without themselves being involved in the mechanism. It can be dangerous to base treatments on the supposition that improving an adverse prognostic feature will improve outlook: treatment may improve a

marker but have either a neutral, or even a detrimental effect on survival.

The presence of non-sustained ventricular tachycardia on Holter monitoring is a sign of an increased mortality risk for sudden death. Class I antiarrhythmic agents can reduce the frequency of ventricular tachycardia, and as a result they were suggested for this purpose in heart failure. However, the Cardiac Arrhythmia Suppression Trial (**CAST**), a randomized controlled trial of three such agents in left ventricular dysfunction, showed that despite reducing the frequency of ventricular arrhythmias there was an increased rate of sudden death, presumably due to some proarrhythmic effect. Similarly, a low ejection fraction is an adverse prognostic sign in heart failure, and it was expected that agents that improve ejection fraction should increase survival. However, in controlled studies, positively inotropic agents such as milrinone (a phosphodiesterase inhibitor) increase ejection fraction but reduce survival. Hence, improving a risk marker should never be used as the justification for treatment, unless we have proof that in so doing we improve survival. The only justifications for treatment are to slow the progression of the underlying disease, to relieve symptoms that trouble the patient, or to use agents proven to improve survival.

Specific prognostic indicators

These can be divided into several relatively independent groups. The most important factors are:

1. the extent of the left ventricular dysfunction;
2. the degree of functional limitation;
3. electrolyte disturbances;
4. the degree of neurohormonal or autonomic dysfunction; and
5. certain electrophysiological or electrocardiographic indicators of ventricular arrhythmogenesis.

There are also general factors such as age or the presence of comorbidities. It has not been established whether estimation of these predictive variables materially improves patient management. An improved scheme for accurate risk stratification can be used to prioritize patients for more careful and regular medical follow-up, or to select patients for expensive or limited treatment options such as transplantation. Table 4 lists some of the established risk markers for poor survival in patients with chronic heart failure.

Further reading

Anon. (2000). Heart failure drugs: what's new? *Drug and Therapeutics Bulletin* **38**, 25–7.

Coats AJ (1996). The 'muscle hypothesis' of chronic heart failure. *Journal of Molecular and Cellular Cardiology* **28**(11), 2255–62.

Flather MD, *et al.* (2000). Long-term ACE-inhibitor therapy in patients with heart failure or left-ventricular dysfunction: a systematic overview of data from individual patients. ACE-Inhibitor Myocardial Infarction Collaborative Group. *Lancet* **355**,1575–81.

Kannel WB, Belanger AJ (1991). Epidemiology of heart failure. *American Heart Journal* **121**, 951–7.

McMurray JJ, Stewart S (2000). Epidemiology, aetiology, and prognosis of heart failure. *Heart* **83**(5), 596–602.

Westaby S (2000). Non-transplant surgery for heart failure. *Heart* **83**, 603–10.

(See Chapter 15.5.1 for more complete referencing of the medical treatment of heart failure.)

15.2.3 Syncope and palpitation

A. C. Rankin and S. M. Cobbe

Syncope and palpitation are symptoms that are commonly of cardiovascular origin, which may be related to abnormalities of cardiac rhythm. However, there are a number of aetiologies for both, and the prognosis for either can range from benign to life-threatening, hence a major priority in assessment is the identification of patients who may be at risk of dying. Treatment options may also range from reassurance with no therapy, to curative or lifesaving treatments for cardiac arrhythmia. The investigation, diagnosis, and management of cardiac arrhythmias are described in more detail in Section 15.6.

Syncope

Syncope is defined as a transient loss of consciousness with the loss of postural tone, and is most commonly due to cardiovascular mechanisms resulting in reduced cerebral perfusion. It is a common presentation, resulting in 1 to 2 per cent of emergency department visits and up to 6 per cent of hospital admissions. The cause is often initially uncertain and assessment must first differentiate syncope from other causes of loss of consciousness, in particular epileptic seizures. The next priority is to identify high-risk patients.

Syncope can be considered in three categories, namely (1) neurally-mediated, (2) cardiac, and (3) neurological or psychiatric (Table 1). The commonest cause is reflex-mediated or vasovagal syncope, which has a benign prognosis, whereas cardiac causes of syncope have been reported to have 1-year mortality rates as high as 18 to 33 per cent. Patients without underlying heart disease in whom no aetiology of syncope is established have a good prognosis, but they may have recurrent syncope.

Where a cause is eventually identified, the diagnosis is indicated by the initial clinical assessment—including history, physical examination, and an electrocardiogram (**ECG**)—in up to 50 per cent of cases. The history is most important and may strongly suggest a vasovagal origin or an epileptic seizure. However, the diagnosis may be complicated by an overlap in features, such as convulsive movements during a vasovagal episode due to anoxic convulsive seizures. It is increasingly recognized that many patients who attend clinics for epilepsy have been misdiagnosed and are suffering from recurrent syncope. Some of these patients have potentially lethal ventricular arrhythmias and should be receiving treatment.

Aetiology

Neurally mediated syncope

There are many disorders of autonomic control that can cause orthostatic intolerance and thereby syncope. To simplify, these can be considered as causing either reflex syncope, due to an increased sensitivity of normal reflex responses, or autonomic dysfunction, where abnormal neurovascular control results in orthostatic hypotension.

Vasovagal syncope

Vasovagal or neurocardiogenic syncope is the most common cause of syncope. It can affect all age groups and varies from infrequent episodes associated with obvious triggering factors to frequent unprovoked collapses, which may be debilitating. The pathophysiology most commonly involves the upright posture with venous pooling of blood and reduced venous return to the heart. Reduced cardiac output and blood pressure stimulate arterial baroreceptors with resultant increased sympathetic activity and catecholamine levels. The vigorous contraction of relatively empty ventricles results in the activation of mechanoreceptors that would normally respond to stretch in the left ventricular wall. Afferent nerve fibres conduct to the cerebral medulla and activate the reflex withdrawal of peripheral sympathetic tone and activation of vagal parasympathetic activity. The resultant vasodilatation and bradycardia cause reduced cerebral perfusion and loss of consciousness. However, there is debate about these mechanisms and other factors may be involved in the aetiology of syncope, as illustrated by the documentation of neurocardiogenic syncope, despite cardiac denervation, in orthotopic heart transplant recipients. Certainly, it is well recognized that vasovagal syncope can result from other stimuli, such as pain, emotional shock, or the sight of blood. In these instances, the reflex activation is central in origin.

The development of tilt-testing has allowed the study of the pathophysiology of neurocardiogenic syncope. The patient is strapped to a tilt-table and is tilted, head upright, usually at 70 degrees for up to 45 min. Protocols that use additional provocation with isoprenaline or nitrates are also commonly used. Blood pressure and cardiac rhythm are monitored throughout the tilt-test. In neurocardiogenic syncope, the patient classically maintains normal blood pressure initially, until the sudden onset of syncope is associated with severe hypotension and bradycardia, often preceded by tachycardia. These features resolve with return to the supine posture. Some patients have a mainly vasodepressor response, with hypotension and little change in heart rate, while others have a marked cardioinhibitory response, with severe bradycardia or asystole of several seconds' duration (Fig. 1). However, most patients exhibit a mixed response, and those patients with marked cardioinhibition also have a preceding vasodepressor response. This is an important observation when treatment is considered since permanent pacing to maintain cardiac rhythm may not cure all symptoms, because falls in blood pressure may still occur even when bradycardia is prevented.

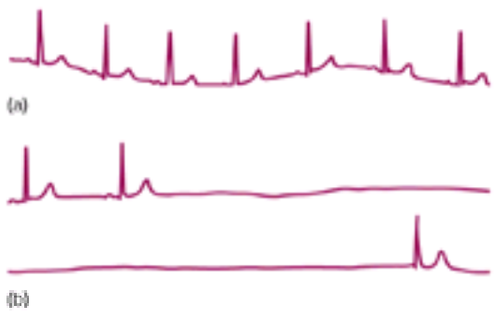

Fig. 1 Cardioinhibitory response to tilt-testing. (a) After 6 min of head-up tilting at 70° the patient complained of presyncope. Heart rate was 60 per min but blood pressure was 70 mmHg. (b) By 7 min the patient had lost consciousness, associated with an asystolic pause of 10 s duration and an unrecordable blood pressure. Recovery was rapid following the patient's return to the supine position.

Carotid sinus hypersensitivity

This is an abnormal sensitivity of a normal reflex that is responsible for syncope. Activation of the carotid sinus baroreceptors (for example by physical pressure, such as carotid sinus massage) results in sympathetic withdrawal and parasympathetic activation. Bradycardia is usually a prominent feature.

Situational reflex-mediated syncope

In susceptible individuals, similar abnormal reflex sensitivity can result in syncope in response to afferent activity from other mechanoreceptor activation. Syncopal responses to cough, micturition, defaecation, or swallowing have been reported.

Autonomic dysfunction

Hypotension may occur in patients in whom there are abnormalities in the autonomic control of cardiovascular function. Abnormalities of afferent or efferent pathways, or of peripheral vascular control, can result in low blood pressure in the upright posture, that is to say orthostatic hypotension. This may be diagnosed by a fall in systolic pressure of more than 20 mmHg, or to less than 90 mmHg, within 3 min of standing. During tilt-testing there may be an immediate drop in blood pressure with head-upright tilting, or a progressive fall may be observed in some patients, in contrast to those with reflex-mediated syncope in whom blood pressure is maintained until the sudden onset of symptoms. Orthostatic hypotension is more common in elderly patients, where it may be multifactorial, often exacerbated by drugs (Table 2). Nocturnal symptoms may occur, with a fall in blood pressure exacerbated by sudden rising from a warm bed.

Autonomic failure is an uncommon cause of syncope and patients may present with other features, including disturbances of bowel, bladder, or sexual function. Pure autonomic failure can be acute or chronic, primary (of unknown origin) or secondary to systemic disease. Multiple system atrophy is characterized by autonomic dysfunction, parkinsonism, and ataxia. Orthostatic hypotension may be a marked feature (the Shy–Drager syndrome), with additional parkinsonian features or cerebellar symptoms. Secondary autonomic failure can result from the central or peripheral involvement of certain diseases, including multiple sclerosis, a cerebral tumour, diabetes, and amyloidosis. A recently reported milder form of autonomic dysfunction, the postural orthostatic tachycardia syndrome (**POTS**), causes symptoms because of inappropriate tachycardia on standing, and occasionally syncope secondary to hypotension.

Cardiac syncope

Loss of consciousness of cardiac origin may result from abnormalities of heart rhythm, due to extremes of rate, either fast or slow, or from some major disturbance of cardiovascular function, with resultant reduced cerebral perfusion. The importance in establishing the diagnosis of cardiac syncope is the associated adverse prognosis, which may be improved with appropriate reatment. The probability of cardiac syncope is increased in the presence of structural cardiovascular disease identified from the history, clinical examination, or investigation.

Bradycardia

A sudden decrease in heart rate, onset of ventricular standstill, or asystole may be a cause of syncope. When due to sinoatrial dysfunction (Sick-sinus syndrome) this is not associated with a poor prognosis, but syncope due to intermittent complete atrioventricular (**AV**) block is. Syncope in a patient with a permanent pacemaker may indicate pacemaker malfunction.

Tachycardia

Syncope may be caused by tachycardia, most commonly ventricular, but supraventricular tachycardia can also be associated with loss of consciousness if it is very fast or in patients with structural heart disease (Fig. 2). Syncope, rather than cardiac arrest, may result from self-terminating ventricular tachycardia or from sustained tachycardia with hypotension at the onset, but with a subsequent recovery of blood pressure. Whether or not a tachycardia causes syncope is related to its rate, underlying left ventricular function, and to the patient's baroreceptor sensitivity. Structural heart disease, for example prior myocardial infarction, is the commonest substrate for ventricular tachycardia, but this may also occur in patients with structurally normal hearts. *Torsade de pointes* in a patient with the Long-QT syndrome is an important diagnosis to consider in young people with a history of loss of consciousness and possible epilepsy, in whom the episodes of collapse may be due to syncope caused by ventricular arrhythmia.

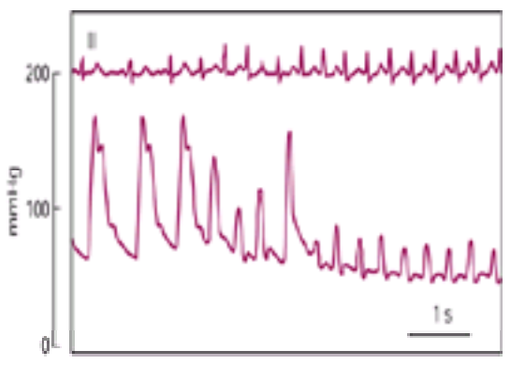

Fig. 2 Hypotension with the onset of supraventricular tachycardia. Surface ECG lead II and intra-aortic pressure are shown.

Structural cardiovascular disease

Aortic stenosis may be associated with syncope, particularly during sudden exertion when the demand for increased cardiac output cannot be met because of the mechanical obstruction. Hypertrophic cardiomyopathy may also be associated with syncope, either because of outflow obstruction or ventricular arrhythmia. Obstruction of blood flow through the mitral valve by an atrial myxoma or thrombus is an extremely uncommon cause of syncope, and a number of other cardiac diseases may be associated with loss of consciousness by a variety of mechanisms (arrhythmia, reflex-mediated or haemodynamic), including myocardial infarction, pulmonary embolism, congenital heart disease, or cardiac tamponade. Vascular diseases may also be involved, such as aortic dissection and extracranial vascular disease.

Neurological or psychiatric causes of syncope

When epilepsy is excluded, neurological aetiologies are rare causes of loss of consciousness, but possible causes include migraine, transient ischaemic attacks, vertebrobasilar vascular disease, and subclavian steal syndrome. A psychiatric origin of syncope implies the absence of neurally mediated, neurological or cardiac abnormalities, and may occur in association with anxiety, depression, and conversion disorders. For instance, apparent syncope may occur during tilt-testing but with normal pulse and blood pressure. Hyperventilation may be an associated mechanistic factor in psychogenic syncope.

History

The importance of the clinical history in assessing a patient with syncope cannot be overemphasized. If possible an eyewitness description of the patient during the syncopal event should always be obtained.

Provocative factors

Vasovagal syncope is classically associated with upright posture, often with aggravating circumstances such as prolonged standing, a hot environment, or hunger. However, episodes may also occur when seated, including when driving. Specific stimuli may be responsible for neurocardiogenic syncope in susceptible individuals.

Ventricular arrhythmia, in particular *torsade de pointes* in the Long-QT syndrome, may be provoked by sudden stimuli such as a noise, for example an alarm clock. Exertional syncope is a feature of aortic stenosis or hypertrophic cardiomyopathy.

Preceding symptoms

Sweating and feeling hot or nauseated may precede vasovagal syncope. Cardiac arrhythmia may be associated with palpitation, chest pain, or breathlessness. Bradycardia, such as intermittent complete heart block, may produce no preceding symptoms and may cause loss of consciousness without warning. Sinoatrial dysfunction is a cause of symptoms of dizziness and light-headedness in addition to syncope. A psychiatric origin may be suggested by multiple associated symptoms including hyperventilation, paraesthesiae in fingers and lips, palpitation and chest pain, which may precede syncope. Epilepsy may be preceded by a characteristic aura, which would strongly point away from syncope as the diagnosis.

The syncopal episode

The duration of loss of consciousness is usually short in syncope, with recovery after a few minutes. A longer duration of loss of consciousness would suggest an alternative diagnosis. An exception to this is when the patient has remained upright during the attack, possibly aided by well-meaning but misguided helpers. Incontinence is a feature of epileptic seizure but may also occur (uncommonly) with syncope. Description of the patient during the episode is of great value. The classic description of an episode of syncope due to cardiac arrhythmia, in particular sudden-onset severe bradycardia, is of a sudden loss of colour, becoming deathly pale, with flushing on recovery (Stokes–Adams attack). Cyanosis may be a feature of an arrhythmic origin of syncope. Convulsive movements during the episode would raise the possibility of epilepsy, but they also occur with syncope.

The recovery period

By contrast to the postictal phase following epilepsy, there is commonly a rapid recovery of cerebral function following syncope. Vasovagal syncope may be followed by persisting nausea or vomiting.

Family history

There are a few specific causes of syncope in which a family history of syncope or sudden death may have prognostic significance. Long-QT syndrome is hereditary and may be associated with sudden death. A family history of syncope is of adverse prognostic significance in hypertrophic cardiomyopathy.

Associated injury

In addition to concern about prognosis in cardiac syncope, there is the possibility of injury occurring with any cause of syncope. The exception to this is syncope of psychiatric origin when injuries are absent despite frequently recurring symptoms.

Investigation

The investigation of cardiac disease and arrhythmia are dealt with in the appropriate chapters, but the approach to the patient with syncope will be described briefly.

Electrocardiogram

An ECG should be performed on all patients with syncope. This may provide evidence of either aetiology of syncope, such as the Long-QT syndrome, or of structural heart disease, such as prior myocardial infarction or left ventricular hypertrophy. An arrhythmia may be documented if it is sustained. There may be evidence of arrhythmia, or sinoatrial disease or conduction system disease, such as trifasicular block, bundle-branch block, or first- or second-degree block. In the absence of carotid bruits, carotid sinus massage, with digital pressure to the carotid artery for up to 5 s, may cause marked bradycardia, with pauses of more than 3-s duration, in carotid sinus hypersensitivity.

Ambulatory monitoring

Documentation of cardiac rhythm during syncope is desirable but is difficult to obtain because of the intermittent and usually infrequent nature of the symptom. Holter monitoring is unlikely to record the rhythm during an episode but may provide evidence of lesser degrees of abnormality, which may support a diagnosis such as sinoatrial dysfunction. Real-time event-recorders are also of limited value in the investigation of syncope because they require a conscious patient to make the recording. Patient-activated loop-recorders, which can store the rhythm prior, during, and following an episode, may be of more value. Implantable loop-recorders are of value in difficult cases.

Tilt-testing

When the history is suggestive of vasovagal syncope, the tilt-test is of value in confirming the diagnosis, allowing reassurance of the patient and clarification of treatment options. However, if cardiac syncope is likely then tilt-testing may be deferred until cardiac investigations are completed. A negative tilt-test does not exclude neurocardiogenic syncope and repeating the test with provocation (isoprenaline or nitrate) may increase its sensitivity.

Electrophysiological testing

Abnormal sinus node function or evidence of atrioventricular conduction disease may be elicited by electrophysiological testing, but demonstrating bradycardia during ambulatory monitoring more reliably makes both these diagnoses. In patients with structural heart disease in whom arrhythmia is suspected, programmed electrical stimulation of the ventricles can induce sustained monomorphic ventricular tachycardia. This is a relatively specific response and shows that the patient is at risk of recurrent ventricular arrhythmia, and makes an arrhythmic origin of syncope likely. The diagnostic yield of electrophysiological testing is low in patients with a structurally normal heart.

Other investigations

Assessment for structural heart disease is important. Physical examination will detect most significant valve disease, but other diagnoses, for example hypertrophic cardiomyopathy or atrial myxoma, may produce little in the way of clinical signs. An echocardiogram is therefore worthwhile. A strong suspicion of diagnoses other than syncope should lead to other investigations, including EEG and brain imaging, but these have a low diagnostic yield in patients with syncope and should not be routine.

Treatment

Neurocardiogenic syncope may require no treatment other than reassurance and avoidance of provocative factors. Treatments of vasovagal syncope, bradycardia, and cardiac arrhythmia are discussed in Chapter 24.13.5 and Chapter 15.6. In up to one-third of patients the aetiology of syncope may not be found: these patients have a good outcome unless they have underlying heart disease.

Palpitation

The symptom of palpitation is defined as an awareness of one's heart beating. This may be due to an awareness of an abnormal heart rhythm but it may also be due to an abnormal awareness of normal rhythm. A careful and detailed history can provide a likely diagnosis. The most important aim in investigation is to correlate symptoms with cardiac rhythm.

History

A description of the symptom should include an estimate of heart rate, duration of symptom, regularity of rhythm, suddenness of onset and offset. It may be helpful to ask the patient to tap with their finger on a desk to describe their palpitation. Trigger factors, including exercise, and aggravating factors such as alcohol and caffeine should be detailed. The length of history may be of interest.

Sinus tachycardia

An awareness of a rapid heart rate of gradual onset and offset is often associated with feelings of alarm and panic in patients with anxiety.

Premature beats

Atrial and ventricular beats commonly occur in normal individuals and may be associated with symptoms. The patient may describe 'missed beats' or forceful beats. These symptoms relate to the pause that follows a premature beat. The premature beat produces a short diastolic filling interval and the low ventricular volume results in reduced ventricular contraction with a small stroke volume. However, the subsequent pause provides a long diastolic filling period and the resultant stretching of the ventricular walls is associated with an increased and forceful systolic contraction. The combination of the diminished premature beat and the enhanced postextrasystolic beat is responsible for the symptoms.

Atrial fibrillation

This common arrhythmia may produce a variety of symptoms depending on ventricular rate, irregularity, and persistence. Paroxysmal atrial fibrillation is typically associated with self-terminating episodes of atrial fibrillation when there is a rapid and irregular ventricular response. The patient is aware of an increased heart rate and often describes the irregular nature of the symptom. The variations in diastolic interval produce symptoms by similar mechanisms to that described above for premature beats, with 'missed' and 'forceful' beats. Patients with sinoatrial dysfunction may be most symptomatic on termination of the atrial fibrillation, which can be followed by sinus bradycardia or prolonged sinus pauses. Atrial fibrillation may be persistent or permanent, and the severity of symptoms will be related to the ventricular rate and irregularity.

Paroxysmal junctional re-entry tachycardia

A history of sudden onset, rapid, regular palpitation in a healthy patient with no underlying structural heart disease is suggestive of paroxysmal junctional re-entry tachycardia. It may stop spontaneously or with vagotonic manoeuvres, or the patient may have had to attend hospital for intravenous therapy. In addition to palpitation, patients commonly report fatigue, malaise, light-headedness, or dyspnoea, but because they have normal hearts such episodes of tachycardia are usually well tolerated. Polyuria is a common associated symptom, which results from the release of atrial natriuretic peptide secondary to atrial stretch.

Ventricular tachycardia

Ventricular arrhythmias can present with the symptom of palpitation, but more severe symptoms such as syncope or cardiac arrest also occur. Characteristically the symptom of palpitation would be of sudden onset and offset of a rapid regular heart rhythm. A history of structural heart disease should be sought.

Investigation

Electrocardiogram

The first aim is to document cardiac rhythm during symptoms. This may be possible with a standard ECG if the arrhythmia is sustained or persistent. Atrial or ventricular premature beats, or evidence of structural heart disease, for example myocardial infarction, may be documented. The presence of pre-excitation indicates the diagnosis of Wolff–Parkinson–White syndrome and suggests symptoms due to episodes of junctional re-entry tachycardia.

Ambulatory monitoring

The success of ambulatory monitoring in documenting the rhythm during symptoms will be dependent on the frequency of symptoms. If they occur daily then a 24- or 48-h Holter recording should suffice. However, palpitation is often infrequent and other patient-activated devices can be of more value. These include handheld, patient-activated event recorders that allow the transtelephonic transmission of recordings. These devices do not allow retrospective recording and require symptoms of sufficient duration to allow their use. Shorter episodes may be captured using loop-recorders.

Electrophysiological studies

Invasive studies are of most value in determining the mechanism of a previously documented tachyarrhythmia, particularly with a view to treatments such as radiofrequency catheter ablation.

Management

Documentation of the cardiac rhythm during palpitation allows appropriate management, with reassurance as the only treatment in those with sinus tachycardia or premature beats. The treatment of other cardiac arrhythmias is discussed in Chapter 15.6.

Further reading

Benditt DG, *et al*. (1999). Pharmacotherapy of neurally mediated syncope. *Circulation* **100**, 1242–8.

Fitzpatrick AP, *et al*. (1993). Vasovagal syncope may occur after orthotopic heart transplantation. *Journal of the American College of Cardiology* **21**, 1132–7.

Grubb BP (1999). Pathophysiology and differential diagnosis of neurocardiogenic syncope. *American Journal of Cardiology* **84**, 3Q–9Q.

Kapoor WN, *et al*. (1983). A prospective evaluation and follow-up of patients with syncope. *New England Journal of Medicine* **309**, 197–203.

Kenny RA, *et al*. (1986). Head-up tilt: a useful test for investigating unexplained syncope. *Lancet* **2**, 1352–4.

Linzer M, *et al*. (1997). Diagnosing syncope. Part 1: Value of history, physical examination and electrocardiography. *Annals of Internal Medicine* **126**, 989–6.

Linzer M, *et al*. (1997). Diagnosing syncope. Part 2: Unexplained syncope. *Annals of Internal Medicine* **127**, 76–86.

Muller T, *et al*. (1991). Electrophysiologic evaluation and outcome of patient with syncope of unknown origin. *European Heart Journal* **12**, 139–43**.**

Schatz IJ, Low P, Polinsky RJ. (1997). Disorders of the autonomic nervous system. *New England Journal of Medicine* **337**, 278–80.

Zaidi A, *et al*. (2000). Misdiagnosis of epilepsy: many seizure-like attacks have a cardiovascular cause. *Journal of the American College of Cardiology* **36**, 181–4.

15.2.4 Physical examination of the cardiovascular system

J. R. Hampton

Introduction

The cardiovascular system is perhaps more accessible to physical examination than any other. Examination is most helpful if it is approached logically: there is a group of general abnormalities that can be detected in cardiovascular disease, there is a group of signs associated with the arterial side of the circulation, a group associated with the venous side, and a group associated with the heart itself.

General examination

Pain due to cardiovascular disease (as in myocardial infarction or aortic dissection) may cause pallor, cold and clammy extremities, and a sinus tachycardia.

Breathlessness and an inability to lie flat (orthopnoea) may result from pulmonary congestion due to left heart failure. Pulmonary oedema will cause extreme breathlessness with the coughing up of frothy and sometimes bloodstained sputum, and peripheral vasoconstriction will cause cold and clammy extremities.

Central cyanosis (affecting the lips and tongue as well as the hands and feet) may indicate heart failure or pulmonary disease but may be due to polycythaemia. Central cyanosis, especially when associated with finger clubbing, is characteristic of a right to left shunt. Peripheral cyanosis indicates a high tissue oxygen extraction, and will be seen when the hands or feet are cold, or when there is peripheral arterial occlusion.

Infection of the heart valves (infective endocarditis) will cause fever, anaemia, weight loss, finger clubbing, 'splinter haemorrhages' under the finger and toenails, a large spleen, loss of arterial pulses and/or neurological abnormalities due to embolization of infected material, and skin lesions such as petechiae and nodules in the fingertips (Osler's nodes).

Physical signs in the arterial circulation

Peripheral pulses

A pulse can be felt whenever an artery is near the surface of the body. In fat or muscular people some pulses may be difficult to feel, but asymmetry of pulses is usually abnormal. Pulses may be lost through atheromatous disease, embolization, or injury. Pulses that are unusually easy to feel—for example in the abdomen, or the popliteal pulses—may be due to aneurysmal dilatation.

The pulses that should be checked are the superficial temporals, carotids, brachials, radials, the aorta, femorals, popliteals, dorsalis pedis, and posterior tibials. Tenderness over the superficial temporal pulses may indicate temporal arteritis. Auscultation over the carotid and femoral pulses, and over the aorta, should be routine as bruits will indicate narrowing, usually atheromatous.

Blood pressure

The blood pressure should be measured in the brachial artery using a cuff around the upper arm. It is essential to use a large cuff in fat people, because a small cuff will result in the blood pressure being overestimated. The diastolic pressure is taken as the point where the sound disappears (Karotkov V). For further discussion see elsewhere in Section 15.

In patients with chest pain, or if ever the radial pulses appear asymmetric, the pressure should be measured in both arms because a difference between the two may indicate aortic dissection.

Heart rate and rhythm

Although the heart rate is usually—and most conveniently—counted in the radial artery, this can be unreliable. In uncontrolled atrial fibrillation there may be a 'pulse deficit', with the true rate being faster than is apparent at the radial artery. The 'apex rate', counted by listening to or feeling the cardiac apex, is more reliable. Similarly, it is easier to distinguish irregularities of the heart rhythm by auscultation than by feeling a peripheral pulse.

Arterial waveform

Descriptions like 'strong' and 'weak' should not be used: these essentially reflect systolic pressure and are best indicated by the blood pressure itself.

The true 'waveform' in an artery can only be obtained by recording the intra-arterial pressure, and the 'pulse character', which describes the rate of rise and fall of pressure, is a crude reflection of this and an unreliable physical sign. It is best felt in large arteries such as the carotid or brachial. In aortic stenosis the rise is slow and the waveform feels flat—the 'plateau' pulse. In aortic regurgitation the arterial pressure falls rapidly, as it does in the left ventricle, causing a 'collapsing' or 'water hammer' pulse. These abnormalities are not easy to detect and are an unreliable guide to the severity of disease. Severe aortic regurgitation, classical of syphilitic aortitis, can cause the head to jerk with each pulse (de Musset's sign).

Abnormalities of the venous circulation

Peripheral veins

Venous thrombosis can occur in the arms, but is much more common in the legs. Thrombosis of an arm vein probably indicates local obstruction (perhaps by a tumour) or a thrombotic tendency. In the legs, superficial phlebitis causes local inflammation in the line of a vein, and the vein is usually palpable and tender. There may be associated swelling. Detecting deep venous thrombosis is much more difficult, but it must be suspected in any painful leg and particularly when there is

unilateral swelling. The 'classical' signs of deep vein thrombosis—calf or thigh tenderness, warmth, and Homan's sign (pain in the calf on dorsiflexion of the foot)—are all unreliable.

The jugular venous pulse

The abbreviation 'JVP' may be used for jugular venous pulse or jugular venous pressure. The importance of the jugular venous pressure is that it directly reflects the pressure in the right atrium and is the best clinical indication of heart failure.

A pulse can be detected in either the external or internal jugular vein. The external vein (running diagonally from the midpoint of the clavicle to the midpoint of the mandible) is usually easy to see. Unfortunately it is a less reliable guide to venous pressure than the internal jugular vein, because flow can be obstructed in the external vein as it passes through the fascia in the neck. The internal jugular vein runs with the carotid artery from the sternoclavicular joint to the lateral end of the mandible, crossing under the sternocleidomastoid muscle and the external jugular vein. The pulsation in the internal jugular is usually seen as a flickering movement under the skin, rather than as an obvious venous pulsation.

The jugular venous pressure is taken as the highest point at which a pulsation can be seen, measured vertically above the manubriosternal angle. The position of the patient does not affect this measurement, because the height of the manubriosternal angle above the right atrium varies little as position changes. The patient should therefore be placed at whatever angle allows the pulsation to be seen most clearly. A patient with a high jugular venous pressure might need to sit upright, whilst a patient with a low pressure might need to lie flat. The key to successful observation of the jugular venous pulse is careful positioning of the patient, and arranging a tangential light, which will make movement under the skin of the neck more obvious.

A jugular venous pulsation has characteristics that allow it to be differentiated from an arterial pulse:

- The position of the pulsation in the neck varies with the patient's posture, reflecting its constant vertical height above the manubriosternal angle.
- A venous pulsation has a complex waveform compared with the single peak of an arterial pulse. In the venous pulse there is the 'a' wave due to atrial contraction, which is followed quickly by the 'c' wave. This has been variously attributed to tricuspid valve closure or to transmission from the carotid artery, and while it can be detected with an external pressure transducer, it is seldom possible to see it. There is then the 'x' descent due to the downward movement of the tricuspid valve as the right ventricle contracts. This is followed by a rise called the 'v' wave, which corresponds with the late stage of atrial filling during ventricular systole, followed by the 'y' trough as the atrium drains into the right ventricle.
- A venous pulse cannot be felt (except sometimes in triscuspid regurgitation).
- A venous pulse can be obliterated by light pressure at the root of the neck.
- The height of the pulsation is affected by respiration, becoming less as intrathoracic pressure falls during inspiration.
- Pressure on the abdomen increases venous return and raises the jugular venous pressure (hepatojugular reflux).
- A prominent external jugular vein, which does not pulsate or move with respiration, may indicate superior mediastinal obstruction by a tumour.

The jugular venous pulse is abnormal in a number of disease states:

- The 'a' wave is lost in atrial fibrillation.
- The 'a' wave becomes more prominent (sometimes described as 'flicking') when the right atrial pressure is high, as in pulmonary hypertension.
- When the atrium contracts against a closed tricuspid valve, as occurs intermittently in complete heart block, large 'a' waves called 'cannon' waves are seen.
- If the tricuspid valve is incompetent the right ventricle expels part of its stroke volume through the right atrium and up the jugular veins, causing a 'cv' or 'systolic' wave. This is followed by a sudden and deep 'y' descent.
- If there is pericardial constriction the jugular venous pressure will rise rather than fall on inspiration.

The liver

If the right atrial pressure is high the liver will be distended and be palpable below the costal margin. It will be smooth and tender. With tricuspid regurgitation the liver will pulsate, and pulsation may also be seen in varicose veins in the legs. With marked and longstanding right heart failure there may also be splenic enlargement. Heart failure causes liver malfunction and—especially when tricuspid regurgitation is present—there may be jaundice.

Oedema

A combination of a high right heart pressure and hormonal changes cause fluid retention, with oedema fluid collecting in dependent areas. There will be symmetrical ankle and leg swelling, and if the patient is in bed for a prolonged period there will be oedema over the sacrum. Fluid can collect in all serous cavities causing pleural and pericardial effusions and ascites.

It is logical to think of the lungs at the same time as the veins: a rise in left atrial pressure will cause pulmonary congestion and eventually pulmonary oedema. These can be recognized by a soft wheeze ('cardiac asthma') and crackles at the lung bases.

Examination of the heart

Palpation

The apex beat is the furthest point outward from the midline, and downwards, where a cardiac impulse can be felt. It is important to accept this definition, and not to confuse the apex beat with the 'point of maximum impulse', because it provides the best guide to the size of the heart. The normal position of the apex beat is in the fifth rib interspace in the mid-clavicular line. The apex beat may be displaced by:

- left ventricular hypertrophy or dilatation
- right ventricular hypertrophy
- mediastinal shift.

Right ventricular hypertrophy is detected from a diffuse 'heaving' movement just to the left of the sternum; remember that the right ventricle forms the anterior surface of most of the heart. Mediastinal shift is identified if the trachea is moved to one side of the suprasternal notch. If the apex beat is displaced laterally and downwards but the trachea is central and there is no parasternal heave, then the left ventricle must be enlarged. A diffuse and abnormal movement medial to the apex may indicate a left ventricular aneurysm.

The precordium should be felt with the flat of the hand over the cardiac apex and the upper right sternal edge where thrills may be present. A thrill is simply a palpable murmur. A loud first sound may be felt as a tapping apex beat.

Heart sounds

There are four possible heart sounds and various extra sounds called clicks or snaps. These discrete noises arise from sudden movements in the circulating blood; they are not the result of valve cusps coming into contact with each other. Their pitch varies, and it is necessary to listen with the bell of the stethoscope for low-pitched sounds, and with the diaphragm for high-pitched sounds. Auscultation should be performed at the cardiac apex, to the left side of the top and bottom of the sternum and at the top of the right sternal edge. These positions are sometimes called the mitral, tricuspid, pulmonary, and aortic areas, but these descriptions are inappropriate because murmurs from these valves are not localized in these areas. The terms should be abandoned.

The first sound is associated with closure of the mitral and tricuspid valves at the beginning of systole. It is sometimes possible to hear separation, or splitting, of these two components but this is not a useful sign.

The second sound, best heard at the upper left sternal edge, is associated with closure of the aortic and pulmonary valves. The aortic valve closes first, and in

inspiration, when pulmonary closure becomes delayed due to the increased blood flow through the right heart, the two components separate more widely. Variable splitting of the second heart sound is heard almost invariably in the young, but becomes less marked with age. Fixed splitting of the second sound is associated with right bundle branch block on the electrocardiogram, and may indicate an atrial septal defect. Reverse splitting is associated with left bundle branch block on the electrocardiogram. Aortic valve closure is delayed in left bundle branch block, so pulmonary closure is heard first. On inspiration pulmonary closure becomes delayed, so the pulmonary sound moves back to join the sound of aortic closure and the second sound becomes single.

The third and fourth sounds are due to rapid ventricular filling. The third sound is in early diastole soon after the second sound, and the fourth sound is associated with atrial contraction and therefore comes just before the first sound. Both are low pitched and are best heard at the apex with the patient rolled slightly onto the left side. A third sound is normal in children and young adults, but after middle age it is a sign of heart failure. A fourth sound (which can be difficult to differentiate from a split first sound) is nearly always abnormal and is said to indicate reduced ventricular distensibility. Right ventricular third and fourth sounds can sometimes be heard at the lower left sternal edge; they are louder on inspiration.

An ejection click, occurring just after the first sound, indicates a deformed but mobile aortic or pulmonary valve. The click may be associated with 'doming' of the valve before it opens. Late systolic clicks are heard with mitral valve prolapse.

An opening snap is a diastolic sound, just after the second sound, and is associated with the opening of a stenosed but mobile mitral or tricuspid valve. Unlike a third sound it is high pitched, and an opening snap can often be heard towards the left sternal edge while a third sound is always localized to the apex.

Heart murmurs

Heart murmurs are the result of turbulent blood flow associated with stenosed, leaking, or incompetent valves (when the murmur is said to be 'regurgitant') and with abnormal connections between the pulmonary and systemic circulation. A loud murmur can sometimes cause a palpable thrill. Murmurs can be systolic or diastolic (referring to the ventricles). A systolic murmur occurs when the mitral and tricuspid valves are shut, and the aortic and pulmonary valves are open. Conversely, diastolic murmurs are heard when the aortic and pulmonary valves are shut and the mitral and tricuspid valves are open.

Most murmurs originate in the left heart. Right heart murmurs are usually quiet but may become louder on inspiration. It should be remembered that the loudness of a murmur is only an extremely crude guide to its haemodynamic significance.

Murmurs originating in the mitral valve are best heard at the cardiac apex with the patient rolled a little to the left side. Mitral regurgitation (or incompetence) causes a medium pitch 'pansystolic' murmur, which begins immediately after the mitral valve closes and persists throughout systole. Murmurs can be recorded and displayed pictorially by the old technique of phonocardiography: Figure 1 shows a murmur of mitral regurgitation.

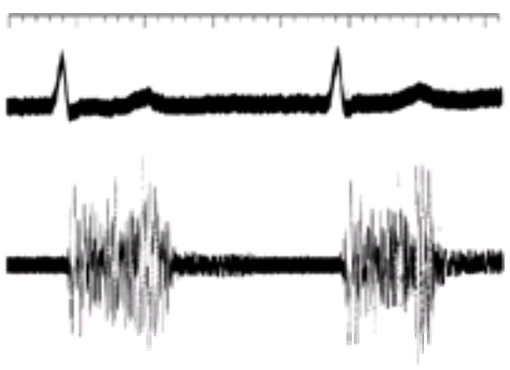

Fig. 1 Phonocardiographic recording of a patient with mitral regurgitation. The ECG recording (top line) determines that the murmur is systolic.

Murmurs spread, or radiate, in the direction of flowing blood and the murmur of mitral regurgitation radiates round the axilla and through to the back, as turbulent blood flows into the left atrium which forms the posterior part of the heart. When mitral regurgitation is due to mitral valve prolapse the murmur is still systolic, but it comes late in systole and is not pansystolic.

In mitral stenosis the opening snap is followed by a murmur, which rises to a peak and then falls, making a 'diamond' shape on a phonocardiogram (Fig. 2). This sort of murmur is called 'ejection'.

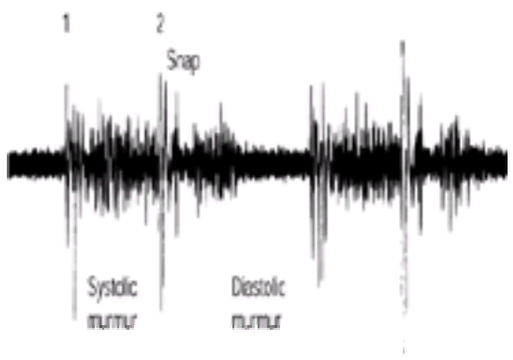

Fig. 2 Phonocardiographic recording of a patient with mixed mitral valve disease.

Tricuspid regurgitation causes a systolic murmur similar to that of mitral regurgitation, but it is loudest at the low left sternal edge, on inspiration. However, many patients with tricuspid regurgitation also have mitral regurgitation, and it can be difficult to differentiate the two. Tricuspid regurgitation is diagnosed from the large 'cv' wave with a steep 'y' descent in the jugular venous pulse, and from an enlarged and pulsating liver. Tricuspid stenosis can occur with congenital or rheumatic heart disease and causes a similar ejection diastolic murmur to mitral stenosis, but the murmur of tricuspid stenosis occurs at the lower left sternal edge and is loudest on inspiration. Such a murmur is also heard as a 'flow' murmur in patients with an atrial septal defect, because increased flow through the right heart causes turbulence even through a normal valve.

Aortic stenosis causes the classic ejection systolic murmur (Fig. 3) heard at the upper right sternal edge. It radiates up the carotids, and this radiation has to be differentiated from a bruit due to narrowing of the carotid arteries. Carotid artery bruits are usually asymmetrical. An aortic ejection systolic murmur may also be heard when the valve is abnormal (for example, when it is bicuspid) but does not impede blood flow, or where there is an increased flow through a normal valve as in anaemia or pregnancy.

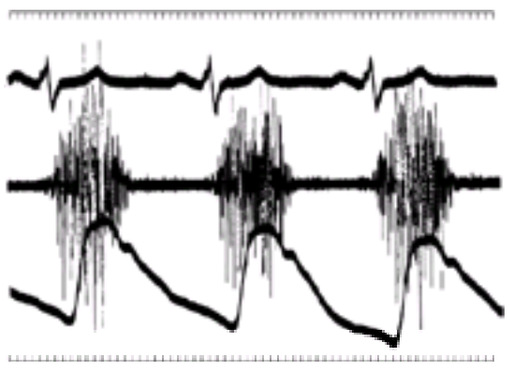

Fig. 3 Phonocardiographic recording of a patient with aortic stenosis. The ECG recording (top line) determines that the murmur is systolic. The bottom line records intra-arterial pressure.

Aortic regurgitation causes a high-pitched murmur that begins immediately after aortic valve closure and dies away, usually by mid-diastole. It is described as an 'early diastolic' murmur, and is best heard at the mid left sternal edge—hence the fallacy of calling the upper right sternal edge the aortic area. The murmur of aortic regurgitation is best heard with the patient leaning forward and breathing out. The regurgitant jet of aortic regurgitation may impinge on the anterior cusp of the mitral valve causing it to flutter. This can cause a murmur (the Austin Flint murmur) that is very similar to that of mitral stenosis, but there is no associated opening snap.

Pulmonary systolic and diastolic murmurs share many of the characteristics of aortic stenosis and regurgitation and are also heard at the left sternal edge. They are quieter and are heard best on inspiration. Pulmonary regurgitation, characteristic of pulmonary hypertension, is sometimes called the Graham Steell murmur.

Murmurs associated with congenital heart disease may be systolic, diastolic, or continuous. The systolic and diastolic murmurs depend on obstruction to flow or regurgitant flow. Continuous murmurs occur in both systole and diastole, with their greatest intensity at the time of the second heart sound. The most typical is the murmur of a patent ductus arteriosus, heard under the left clavicle, but similar sounds are made by artificial shunts such as a Blalock operation.

A ventricular septal defect causes a murmur at the low left sternal edge: it is typically pansystolic, but a small defect can cause a very loud ejection murmur that radiates up the left sternal edge. An atrial septal defect does not cause a murmur itself, but increased right heart blood flow can cause a pulmonary systolic and a tricuspid diastolic murmur.

The importance of the physical examination

An accurate diagnosis of cardiovascular disease depends on the history and on the identification of a group of physical signs. It is always more efficient to look for things rather than at things. In a patient with chest pain look for signs of risk factors (smoking, hypertension, hypercholesterolaemia), evidence of vascular disease elsewhere (absent pulses, arterial bruits), and cardiac damage (signs of heart failure, mitral regurgitation due to papillary muscle dysfunction, postinfarct ventricular septal defect). In patients with palpitations check the rhythm and look for evidence of cardiac disease (especially rheumatic disease) or of other diseases such as thyrotoxicosis. In patients with breathlessness look for signs of heart failure (a rapid heart rate, raised jugular venous pressure, distended liver, ankle swelling, a gallop sound at the cardiac apex).

Individually, almost all physical signs are fallible. Finger clubbing, the identification of which varies between individuals, can be a congenital abnormality totally unrelated to cardiac disease. Splinter haemorrhages may be due to trauma. Cyanosis may be due to polycythaemia rubra vera. A raised jugular venous pressure may due to a 'high output' state such as pregnancy. The loudness of a murmur gives little information about its importance—and so on. But grouped together (for example a mitral regurgitant murmur with a displaced apex beat due to left ventricular hypertrophy) and used intelligently they become much more reliable.

It is true that there is considerable variability in the identification of physical abnormalities, particularly between non-specialists. But to abandon physical signs in favour of the ECG, the echocardiogram, and the chest radiograph is impracticable and expensive. In appropriate patients all these and other investigations may be essential. But for the initial assessment of patients, and for monitoring progress of disease, physical examination is both essential and adequate. It maintains a holistic approach to patient care, and can be an extremely satisfying art.

Further reading

Butman SM *et al.* (1993). Bedside cardiovascular examination in patients with severe chronic heart failure: importance of rest or inducible jugular venous distension. *Journal of the American College of Cardiologists* **22**, 968–74.

Fletcher RH, Fletcher RW (1992). Has medicine outgrown physical diagnosis? *Annals of Internal Medicine* **117**, 786–7.

Ishmail AA *et al.* (1987). Interobserver agreement by auscultation in the presence of a third heart sound in patients with congestive heart failure. *Chest* **91**, 870–3.

Spiteri MA, Cook DG, Clarke SW (1988). Reliability of eliciting physical signs in examination of the chest. *The Lancet* **1**, 873–5.

Stevenson LW, Perloff JK (1989). The limited reliability of physical signs for estimating hemodynamics in chronic heart failure. *Journal of the American Medical Association* **261**, 884–8.

15.3.1 Chest radiography in heart disease

M. B. Rubens

Introduction

The chest radiograph is often abnormal in patients with congenital heart disease, but in most patients with acquired heart disease it is normal and rarely provides a precise diagnosis. Nonetheless, a chest radiograph remains part of the routine work-up of virtually all patients with known or suspected heart disease because it is relatively inexpensive and non-invasive and it provides a record of cardiac size and shape, sometimes also suggesting specific chamber enlargement. Abnormal cardiac calcification may be visible, and analysis of the pulmonary vessels may indicate particular physiological disturbances. Analysis of the skeleton may provide evidence of associated systemic disease or previous surgery, and abnormalities of situs may be apparent. Occasionally, unsuspected non-cardiac abnormalities are discovered.

A routine examination always includes a frontal view and sometimes a lateral view. Ideally, the frontal view is posteroanterior, with the patient upright and at end-inspiration. Patients who are too ill to be taken to the X-ray department may be examined with mobile equipment when an anteroposterior film is taken. In this projection the heart appears magnified because it is further from the film. A lateral film may give additional information on heart size and shape, and cardiac calcification is often best demonstrated in this view. Frontal and lateral films combined with a barium swallow may provide data on left atrial size and the presence of aberrant branches of the great vessels.

As in all areas of clinical examination, the chest radiograph should be analysed in a careful, systematic manner. It must be remembered that poor radiographic technique may produce spurious appearances. Moreover, the cardiovascular silhouette is such an obvious focus of attention that the bones, soft tissues, and upper abdomen may be overlooked. The cardiac shadow provides information about anatomy, but the lungs provide information about haemodynamics. A recommended order of analysis is as follows: technical factors, the bones, the upper abdomen, the lungs, and, finally, the cardiovascular silhouette. Discussion in this chapter will focus on the appearance of the heart and vessels and other changes on the chest radiograph produced by cardiovascular disease.

The normal chest radiograph

The cardiovascular silhouette

The right border of the cardiovascular silhouette (Fig. 1) comprises, from above downwards, the superior vena cava, the body of right atrium, and the inferior vena cava. The normal superior vena cava produces a low-density vertical shadow, just lateral to the spine. The azygos vein may be visible as a convex density above the origin of the right main bronchus and superimposed on the superior vena cava. The lower part of the right cardiovascular silhouette is convex and produced by the body of the right atrium. Occasionally, the inferior vena cava is visible as a short vertical shadow in the right cardiophrenic angle.

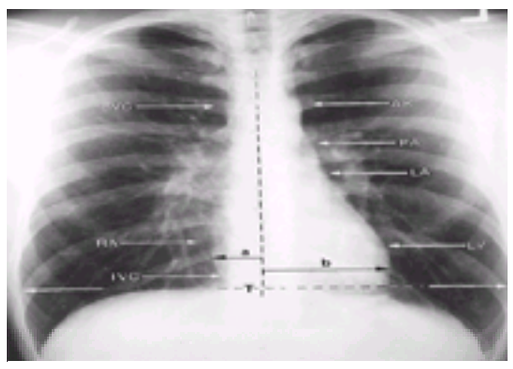

Fig. 1 Normal chest radiograph. The right border of the cardiovascular silhouette comprises superior vena cava (SVC), right atrium (RA), and inferior vena cava (IVC), and the left border comprises aortic knuckle (AK), pulmonary trunk (PA), left atrial appendage (LA), and left ventricle (LV). Transverse cardiac diameter=a+b. Cardiothoracic ratio=(a+b)/ T.

The left border of the cardiovascular silhouette comprises, from above downwards, the aortic knuckle, the pulmonary trunk, the left atrial appendage, and the left ventricle. The aortic knuckle is produced by the posterior part of the aortic arch. The proximal descending aorta may be visible as a vertical shadow, continuous with the knuckle and eventually merging with the left paraspinal shadow. The pulmonary trunk is situated below the aortic knuckle, and below this is a short segment of left atrial appendage. The bulk of the left heart border is formed by the body of the left ventricle. The left cardiophrenic angle may be occupied by a low-density shadow representing an apical pericardial fat pad. Less often, a fat pad is visible in the right cardiophrenic angle.

On the lateral film (Fig. 2) the heart is seen immediately posterior to the inferior half of the sternum. The anterior border of the cardiac shadow is formed almost entirely by the right ventricle, although in atrial diastole the right atrial appendage may come in contact with the sternum. The right ventricular outflow tract is continuous with the main pulmonary artery that arches posteriorly and continues into the left pulmonary artery. The branches of the right pulmonary artery cast an ovoid shadow anterior to the right bronchus. Part of the ascending aorta may be visible above the pulmonary trunk, and the aortic arch is seen passing posteriorly and then descending for a variable distance. The aortic arch is separated from the left pulmonary artery by the subaortic fossa. The upper part of the posterior aspect of the cardiac silhouette is formed by the body of left atrium and the pulmonary veins, and the lower part by the left ventricle. The inferior vena cava may be visible as a short, straight vertical shadow extending from the diaphragm and overlapping the posterior heart border.

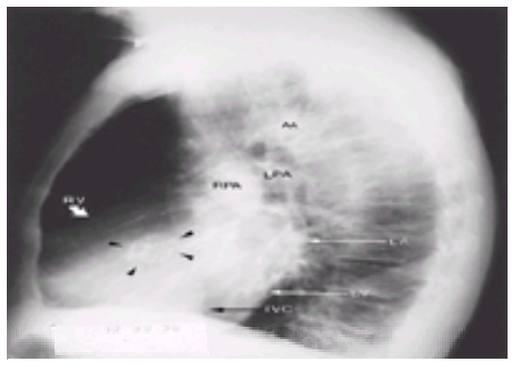

Fig. 2 Lateral chest radiograph. Patient with calcific aortic stenosis. The aortic valve is heavily calcified (arrowheads). AA=aortic arch; IVC=inferior vena cava; LA=left atrium; LPA=left pulmonary artery; LV=left ventricle; RPA=right pulmonary artery; RV=right ventricle.

The normal cardiovascular silhouette changes with age. In infancy, the thymus occupies much of the anterior mediastinum and may obscure the aorta, the pulmonary trunk, and the right ventricular outflow tract. By early adolescence, the thymus is no longer visible on the chest radiograph, and the normal pulmonary trunk is often prominent. In adulthood, the pulmonary trunk is less prominent, and with advancing years the left ventricular contour becomes more convex. In old age, the ascending aorta may become tortuous and project lateral to the superior vena cava, and the aortic knuckle and descending aorta may be increasingly prominent.

Heart size

The commonest methods of assessment of cardiac size using the chest radiograph are the measurement of the transverse cardiac diameter and the measurement of cardiothoracic ratio (Fig. 1).

The transverse cardiac diameter is measured on a posteroanterior film by adding the maximum distance of the right heart border from the mid-line to the maximum distance of the left heart border from the mid-line. The upper limit of normal is 16 cm for men and 15 cm for women. A change of 1.5 cm in cardiac diameter should be regarded as significant. Apparent increase in heart size may be due to a poor inspiration; on anteroposterior films geometric magnification of the heart shadow occurs.

The cardiothoracic ratio is the ratio of transverse cardiac diameter to maximum internal diameter of the thorax. There are racial differences in the normal ratio, which should not exceed 50 per cent in white subjects or 55 per cent in black subjects.

The pulmonary vasculature

The pulmonary trunk normally forms a short segment of the left cardiovascular silhouette. The pulmonary arteries are not visible on the chest radiograph until they emerge from the pericardium and are surrounded by aerated lung. The right pulmonary artery lies anterior to the right main bronchus and usually divides into upper lobe and descending arteries just before emerging from the pericardium. The descending branch is usually clearly seen lateral to the right heart border, where it forms the bulk of the right hilum. Its diameter should not exceed 15 mm in women and 16 mm in men. The left pulmonary artery arches posteriorly over the left main bronchus, and the left hilum is therefore higher than the right. The branching pattern of the pulmonary arteries is similar to that of the bronchi. As the pulmonary arteries pass peripherally, they taper smoothly and are not normally visible in the outer third of the lung.

The anatomy of the pulmonary veins is variable. The upper lobe veins run lateral to the corresponding pulmonary arteries and can often be identified crossing the pulmonary arteries at the hilum prior to the left atrium. The lower lobe veins run more horizontally and medially than the accompanying arteries.

On the frontal chest radiograph of a normal, erect subject the pulmonary vessels should be clearly visible and are larger in the lower zones than in the upper zones. The upper zone veins in the first anterior intercostal space should not exceed 3 mm in diameter. On a supine film, the upper zone and lower zone vessels appear similar in size.

The abnormal chest radiograph

Technical considerations

On an over-exposed chest radiograph the lungs appear blacker than usual and may mimic pulmonary oligaemia. Conversely, an under-exposed film may accentuate the pulmonary vascular pattern, or even suggest diffuse lung disease. A chest radiograph taken on expiration may show increased basal shadowing and suggest pulmonary oedema or other some other interstitial pulmonary abnormality, and the heart may appear enlarged. A rotated film may make some structures appear unusually prominent and others unusually small.

The bones

Deformity of the thoracic skeleton may alter the appearance of the heart. Sternal depression (pectus excavatum) usually displaces and rotates the heart to the left producing a characteristic straight left heart border). In the 'straight back syndrome', the anteroposterior diameter of the thorax is decreased and the heart may be compressed between sternum and spine producing a spurious appearance of cardiomegaly on the frontal chest radiograph. Severe scoliosis may not only alter the shape of the mediastinum but may actually cause cardiopulmonary disease.

Congenital deformity of the thoracic skeleton may be associated with cardiac disease. Both sternal depression and 'straight back' are associated with mitral valve prolapse. Many systemic diseases and congenital syndromes which involve the cardiovascular system may also have skeletal manifestations, for example in Down's syndrome there may be an atrioventricular septal defect and only 11 pairs of ribs.

Rib notching is usually associated with coarctation of the aorta, but may also be seen in pulmonary atresia and vena caval obstruction or following creation of a Blalock–Taussig shunt. Evidence of previous surgery, such as rib deformity, sternal sutures, and prosthetic valves, may be seen on the chest radiograph.

The upper abdomen

Rarely, patients presenting with chest pain have a hiatus hernia or gallstones that may be visible on the chest radiograph. In those with congenital heart disease, information about situs may be visible on the chest radiograph and it is worth noting if the stomach and liver are normally situated, also the visibility and location of the spleen. The best indication of atrial situs, however, is given by the tracheobronchial anatomy.

The lungs

The normal radiographic appearance of the lung is produced by pulmonary vessels outlined by aerated lung. Any opacity that is not a vessel should be carefully considered. Since smoking is an important aetiological factor in both cardiovascular and pulmonary disease, it is not surprising that the routine chest radiograph in a cardiac patient may uncover previously unsuspected lung disease.

The pulmonary vascular pattern may be normal, increased, decreased, or uneven. Normal pulmonary vascularity does not exclude significant myocardial disease, mild valvular disease, or a small intracardiac shunt.

Increased pulmonary vascularity

There are four distinct patterns of increased pulmonary vascularity:

1. pulmonary venous hypertension;
2. pulmonary arterial hypertension;
3. pulmonary over-circulation;
4. systemic supply to the lungs.

Any of these patterns may coexist.

Pulmonary venous hypertension

Pulmonary venous hypertension is most commonly caused by left ventricular failure, mitral valve disease, or aortic valve disease. Rarely, it is due to pulmonary venous obstruction. When pulmonary venous pressure rises, the upper lobe veins distend, becoming similar in size to the lower lobe veins and eventually larger. This phenomenon may be described as 'upper lobe blood diversion' (Fig. 3). When the pulmonary venous pressure exceeds the plasma osmotic pressure, fluid accumulates in the interstitial spaces of the lung. This appears radiographically as interstitial pulmonary oedema: the lower zone and hilar vessels may become indistinct (perihilar haze) and interstitial lines may appear. Kerley B lines are caused by fluid-filled interlobular septa and appear as fine, non-branching horizontal lines in the periphery of the lower zones (Fig. 4). Kerley A lines are less common and are longer, fine-line shadows that radiate from the hila into the mid and upper zones. They also represent distended interlobular septa. Excess interstitial fluid around bronchi may appear as peribronchial cuffing. A further rise in pulmonary venous pressure leads to accumulation of fluid in the alveolar spaces (Fig. 5). Classically, alveolar oedema is perihilar, but it may be patchy and asymmetric, or even nodular, and it may be indistinguishable from other forms of pulmonary consolidation. Pleural effusions are common in pulmonary venous hypertension.

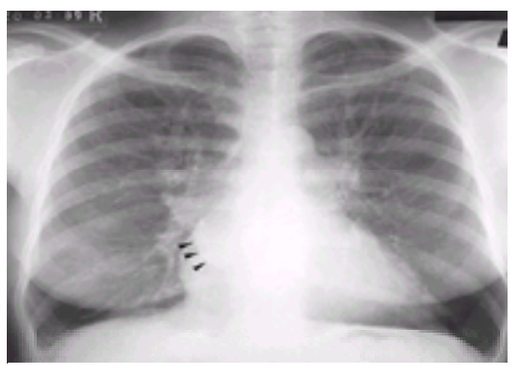

Fig. 3 Upper lobe blood diversion. Patient with mitral valve disease. The extra density over the right heart border (arrowheads) is due to left atrial enlargement.

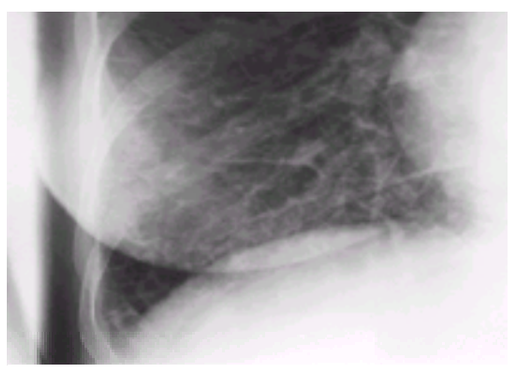

Fig. 4 Kerley B lines. Patient with interstitial pulmonary oedema due to mitral stenosis.

Fig. 5 Severe pulmonary venous hypertension. Patient with acute myocardial infarction. In addition to upper lobe blood diversion, perihilar haze and basal septal lines indicate interstitial oedema, and right lower zone consolidation indicates alveolar oedema.

In the untreated patient there is a fairly close correlation between the pulmonary capillary wedge pressure and the radiographic signs of pulmonary venous hypertension. A normal vascular pattern corresponds to a wedge pressure of less than 12 mmHg, redistribution of blood flow corresponds to 12 to 18 mmHg, interstitial oedema corresponds to 18 to 22 mmHg, and above 22 mmHg there is usually overt alveolar oedema. If a patient has received diuretic therapy, this correlation is less reliable.

In patients with long-standing pulmonary venous hypertension, radiographic signs of pulmonary arterial hypertension may also be present. Chronic pulmonary venous hypertension may also be associated with pulmonary haemosiderosis or pulmonary ossicles. The former appears as a fine nodular pattern throughout both lungs, and the latter as calcified basal nodules of up to 1 cm in diameter.

Although redistribution of blood flow and septal lines are most often a manifestation of pulmonary venous hypertension, there are other causes that should be considered. Redistribution of blood flow may occur in patients who have basal emphysema and no evidence of pulmonary venous hypertension. Septal lines may be seen in non-cardiogenic pulmonary oedema, lymphangitis carcinomatosa, sarcoidosis, and silicosis.

Pulmonary arterial hypertension

Pulmonary arterial hypertension may be defined as a pulmonary artery systolic pressure exceeding 30 mmHg. The commonest causes of this condition are chronic lung disease, pulmonary emboli, pulmonary venous hypertension, and intracardiac shunts. It may also be idiopathic. The typical radiographic appearances are enlargement of the central pulmonary arteries and attenuation of the peripheral arteries. In severe, long-standing pulmonary arterial hypertension, calcification may be

seen in the central pulmonary arteries.

An indication of the underlying cause may be present on the chest radiograph; for example there may be signs of chronic obstructive airways disease or pulmonary embolism. Bilateral hilar lymph node enlargement may mimic enlarged central pulmonary arteries, but usually lymphadenopathy is lobulated, whereas enlarged arteries have a smooth outline.

Pulmonary over-circulation

Pulmonary over-circulation or plethora implies increased blood-flow through the lungs. It is usually due to a left-to-right shunt, less commonly due to bidirectional shunting and rarely due to increased cardiac output. Small shunts may not be perceptible on the chest radiograph, but shunts with a pulmonary-to-systemic flow ratio of 2:1 or greater should be apparent unless there is coexisting heart failure. The central pulmonary arteries are larger than normal and peripheral pulmonary vessels are visible in the outer third of the lung (Fig. 6). Pulmonary plethora in a non-cyanosed patient indicates a left-to-right shunt, whereas in the presence of cyanosis it indicates bidirectional shunting.

Fig. 6 Pulmonary plethora. Infant with transposition of the great arteries. The pulmonary vascular pattern is accentuated; there is also cardiomegaly.

Systemic supply to the lungs

Systemic arterial supply to the lungs, which is sometimes referred to as 'bronchial circulation', develops in patients with severe right ventricular outflow obstruction. The pulmonary trunk is either small or absent, and the peripheral vessels are disorganized and may produce a reticular or nodular pattern that mimics diffuse lung disease.

Decreased pulmonary vascularity

Pulmonary oligaemia

Pulmonary oligaemia implies decreased blood flow through the lungs. It is usually due to right ventricular outflow obstruction in association with a right-to-left shunt, for example tetralogy of Fallot. The lungs appear to have fewer and smaller vessels than usual, and the pulmonary trunk may be small or inapparent. Pulmonary oligaemia due to restricted filling of the right heart, such as occurs in cardiac tamponade, is rarely perceptible on the chest radiograph.

Uneven vascularity

Uneven pulmonary vascularity is most commonly due to pulmonary disease. A previous lung resection will obviously alter the vascular pattern. Apart from pulmonary thromboembolism, cardiovascular causes of uneven vascularity are rare but include previous shunt operations for congenital heart disease, pulmonary artery stenoses, and pulmonary arteriovenous fistulae.

Abnormalities of the heart and great vessels

The systemic veins

Enlargement of the superior vena cava may be caused by either increased flow or increased pressure. Increased flow occurs in supracardiac anomalous pulmonary venous return. Increased pressure occurs in right heart failure, tricuspid valve disease, cardiac tamponade, and constrictive pericarditis. The superior vena cava may also dilate secondary to obstruction caused by mediastinitis or mediastinal tumour. The superior vena cava may be displaced laterally by a tortuous or dilated ascending aorta or a right-sided aortic arch.

The azygos vein may enlarge for the same reasons as enlargement of the superior vena cava. An enlarged azygos vein is also seen in superior vena caval obstruction, portal vein obstruction, and absence of the hepatic portion of the inferior vena cava in polysplenia.

The inferior vena cava may enlarge in secondary to tricuspid valve disease and right heart failure.

The right atrium

Right atrial enlargement rarely occurs in isolation, and is usually associated with right ventricular enlargement. Classically, right atrial enlargement produces increased prominence of the lower half of the right side of the cardiac shadow. It occurs in right heart failure, tricuspid valve disease, and in atrial septal defect and other shunts that enter the right atrium.

The right ventricle

The normal right ventricle is not a border-forming structure on the frontal chest radiograph. An enlarging right ventricle tends to displace the left ventricle laterally so that the cardiac apex becomes elevated. In gross right ventricular enlargement, the right ventricle may actually form the left heart border, and dilatation of its outflow tract may produce a bump just below the pulmonary trunk. On the lateral view, right ventricular enlargement may manifest as increased contact of the heart with the sternum. Right ventricular enlargement occurs in pulmonary arterial hypertension, tricuspid valve disease, pulmonary valve disease, left-to-right shunts, and tetralogy of Fallot.

The pulmonary trunk

Enlargement of the pulmonary trunk is due to increased pressure, increased flow, poststenotic dilatation, or idiopathic dilatation (Fig. 7). In pulmonary arterial hypertension, it may be associated with enlargement of the central pulmonary arteries and peripheral pruning. In situations of increased flow, it is associated with pulmonary plethora. In cases of poststenotic and idiopathic dilatation, it is usually associated with enlargement of the left pulmonary artery and normal peripheral vascularity.

Fig. 7 Patient with mitral stenosis. The transverse cardiac diameter size is normal, but the left atrial appendage is enlarged (curved arrow), and there is also prominence of the pulmonary trunk (arrowhead).

In corrected transposition of the great arteries, the pulmonary trunk is not visible on the chest radiograph. In tetralogy of Fallot and pulmonary atresia, the pulmonary trunk is small, producing an obvious pulmonary bay.

The left atrium

The body of the left atrium is situated beneath the carina and in front of the oesophagus. Enlargement superiorly may increase the angle between the left and right bronchi by elevating the left bronchus and displacing it posteriorly. Posterior enlargement may displace the oesophagus posteriorly. Enlargement to the right may produce an extra density over the right heart border (Fig. 3), and if grossly enlarged the left atrium may actually form the right heart border. Enlargement of the left atrial appendage causes straightening or convex bulging of the upper left heart border (Fig. 7). Left atrial enlargement occurs most obviously in mitral valve disease, but is seen in other forms of left heart failure, in shunts at ventricular and great vessel level, and in association with left atrial tumours.

The left ventricle

Left ventricular hypertrophy produces increased convexity of the left heart border, but not cardiac enlargement unless heart failure develops. Left ventricular dilatation causes displacement of the cardiac apex downward and to the left, and on the lateral view the heart shadow extends more posteriorly than usual. Left ventricular hypertrophy results from systolic overload, and dilatation from diastolic overload. In left ventricular aneurysm, a discrete bulge may develop on the left heart border.

The aorta

Selective enlargement of the ascending aorta is seen in poststenotic dilatation due to aortic valvar stenosis and in association with aneurysms. The aortic knuckle may be prominent due to aneurysm, patent ductus arteriosus, tetralogy of Fallot, and pulmonary atresia. In coarctation of the aorta, the knuckle always appears abnormal—it may be prominent, flat, high, low, or have an abnormal contour. In non-obstructing coarctation or pseudocoarctation, the arch appears elongated and kinked. Selective enlargement of the descending aorta may be due to aneurysm. Generalized prominence of the thoracic aorta may be part of the ageing process but is also seen in systemic hypertension and aortic regurgitation.

The aortic arch is usually left-sided, arching posteriorly over the left main bronchus. However, it can be right-sided, when it arches over the right bronchus and indents the right side of the trachea, the usual shadow of the left arch is absent, and the superior vena cava may be displaced laterally. A right arch with an aberrant left subclavian artery is not usually associated with heart disease, but if its branches are the mirror image of normal there is a high incidence of congenital heart disease. Tetralogy of Fallot, pulmonary atresia, truncus arteriosus, and ventricular septal defect may be associated with a right arch. An aberrant subclavian artery can be identified on a barium swallow.

The pericardium

Pericardial effusion may produce non-specific globular enlargement of the heart shadow and rapid increase in heart size on serial films is suggestive of this condition. Pulmonary vascularity is usually normal. A pericardial cyst may appear as a well-circumscribed, rounded opacity adjacent to the heart. Partial pericardial defects may allow herniation of the left atrial appendage causing a prominent bulge on the left heart border. In congenital absence of the pericardium there is usually displacement of the entire heart to the left.

Cardiac calcification

Calcification may occur in any cardiovascular structure, and is usually the result of inflammatory disease or infarction. Although cardiac calcification may be visible on the chest radiograph it is better demonstrated by fluoroscopy when movement of an abnormally calcified structure aids its detection, in contrast to the chest radiograph where movement causes blurring.

Myocardial and endocardial calcification most commonly occur in the left ventricle secondary to coronary artery disease. Curvilinear calcification may occur in the wall of left ventricular aneurysms, in thrombi, and in infarcts. Left atrial wall calcification may be due to rheumatic myocarditis, and left atrial thrombi may calcify.

Aortic valve calcification usually lies over the spine on the frontal chest radiograph and may, therefore, be obscured. It is best seen on the lateral view (Fig. 2), and tends to lie mostly above a line drawn from the carina to the anterior costophrenic angle. Mitral valve calcification usually lies to the left of the spine, and on a lateral view lies below the line drawn from carina to the anterior costophrenic angle. Calcification is rarely seen in the tricuspid and pulmonary valves, but commonly occurs in right ventricular outflow tract homografts.

Calcification is frequently seen in the aortic arch of older patients as part of the normal ageing process. Extensive aortic calcification is most likely to be due to atheroma, but it may be the result of an arteritis or syphilitic aortitis, which characteristically involves the ascending aorta. Calcification in a healed dissecting aneurysm may be seen in any part of the aorta. Chronic traumatic aneurysms in the region of the aortic isthmus may calcify. Coronary artery calcification indicates atheroma, but does not necessarily correspond to significant coronary artery narrowing. A patent ductus arteriosus may calcify, and calcification may develop in the central pulmonary arteries in long-standing, severe pulmonary arterial hypertension.

Pericardial calcification may be a sequel to pericarditis and haemopericardium, and it may be associated with pericardial constriction. Rare causes of cardiac calcification include tumours, hydatid disease, and coronary artery fistulae.

Further reading

Elliott LP (1991). *Cardiac imaging in infants, children and adults*. Lippincot, Phildelphia.

Elliott LP, Schiebler GL (1979). *X-ray diagnosis of congenital cardiac disease*, 2nd edn. Charles C. Thomas, Springfield.

Jefferson K, Rees S (1980). *Clinical cardiac radiology*, 2nd edn. Butterworth, London.

15.3.2 Electrocardiography

D. J. Rowlands

The resting 12-lead ECG

Historical introduction

The first electrocardiographic recording of the human heart was made in 1887 by A. D. Waller, who expressed the view that it was unlikely that such recordings would be of much use in clinical practice. This view was not shared by Einthoven, who noted the differences between recordings taken from healthy and from sick persons and was, in consequence, convinced that the technique would prove to be of great clinical value. Einthoven suggested the P, QRS, T, U terminology which is in universal use today, and recognized that a better recording device than the capillary electrometer (used by Waller) would be necessary. He modified and developed the string galvanometer for this purpose. The resulting instrument was unwieldy, weighing over a quarter of a ton (254 kg) and requiring five people for its operation, but it produced electrocardiograms of remarkable quality. The first commercially available machine for recording the electrocardiogram (ECG) was made in England in 1911 by the Cambridge Scientific Instrument Company and was delivered to Sir Thomas Lewis at University College Hospital in London. In the early 1900s Einthoven and Lewis were undoubtedly the pioneers who did most to advance the clinical study of electrocardiography. In the early 1930s the most productive research worker in this field was Frank N. Wilson of Ann Arbor, Michigan, who had been stationed in England during World War I in a rehabilitation hospital under the command of Sir Thomas Lewis. Wilson's early research work was, therefore, undertaken in England and, whereas Lewis had concentrated on the cardiac rhythm, Wilson's work was centred on the QRS complexes and T waves. It was Wilson who developed the unipolar recording system, by developing an 'indifferent' electrode that gave a stable reference potential with respect to which the potential at a single exploring electrode could be measured. This reference potential was obtained from a central terminal connected to the left arm, the right arm, and the left leg through equal resistors.

The development of electrocardiographic recording techniques continues (witness the rapidly expanding use of electrophysiological studies of the heart), but the standard 12-lead ECG still forms an essential part of any full clinical cardiological assessment more than 100 years after the first human ECG recording was made. It is estimated that in excess of 100 million 12-lead ECGs are recorded annually worldwide, a fact that would surely have astonished Waller. Electrocardiography developed empirically and its basic diagnostic criteria remain empirical. The criteria given in this chapter represent a reasonable compromise between sensitivity and specificity.

Normal ECG appearances

The basic ECG waveform

The basic ECG waveform consists of three recognizable deflections termed 'P wave', 'QRS complex', and 'T wave', by Einthoven (Fig. 1). The P wave is the surface electrocardiographic manifestation of atrial myocardial depolarization. Depolarization of the sinoatrial node is not recognizable on the surface ECG and can only be inferred from the shape and direction of the P wave. The QRS complex is the surface electrocardiographic manifestation of ventricular myocardial depolarization. The S–T segment and T wave represent ventricular myocardial repolarization. Atrial myocardial repolarization is indicated by the Ta wave, which is a small, asymmetrical negative wave following the P wave, usually obscured by the QRS complex which occurs at the same time. The Ta wave usually becomes easily recognizable during sinus tachycardia (especially during exercise), since it then increases in size and becomes a rounded negative wave beginning before the QRS complex and extending into the S–T segment. A prominent atrial repolarization wave occurring during an exercise stress test is frequently wrongly interpreted as S–T-segment depression. The key to avoiding this error is to recognize that the negativity begins before the QRS complex. The P wave and T wave have relatively simple shapes which exhibit few variations. The QRS complexes exhibit more readily recognizable differences in pattern in different leads within the same ECG.

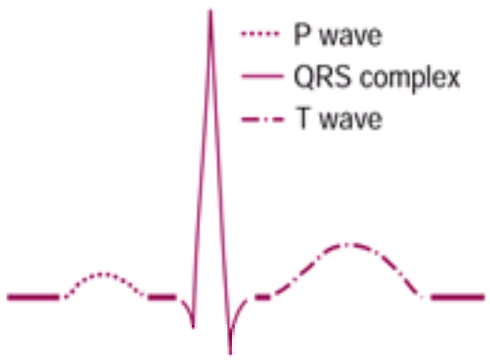

Fig. 1 The basic ECG waveform.

QRS waveform nomenclature

The QRS complexes usually have the largest voltages and virtually always the highest frequency components of the various ECG deflections, and typically consist of 'sharp', pointed deflections. The presence and relative size of the several possible components of the QRS complex may be indicated by a convention using combinations of the letters q, r, s, Q, R, S (Fig. 2). If a given component is considered to be large, an UPPER CASE letter is used, if it is considered to be small a lower case letter is used.

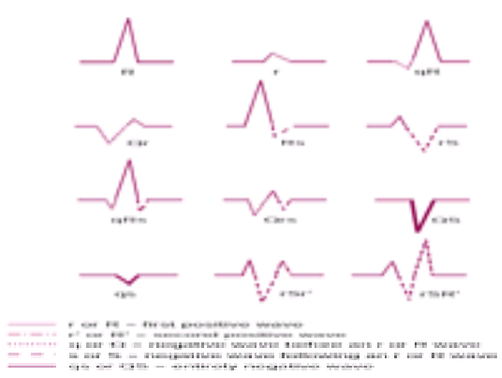

Fig. 2 QRS waveform nomenclature.

The 12 conventional ECG leads

Unipolar, bipolar, and augmented leads

Leads I, II, and III are the bipolar limb leads, introduced originally by Einthoven. The remaining three limb leads and the six precordial leads are unipolar (they involve the use of the central reference terminal of Wilson) and are termed V leads (the 'V' originally stood for 'voltage', to reflect the fact that these unipolar leads effectively measure the voltage at the location of the recording electrode). All currently available ECG machines use augmented (a) limb leads (that is to say they record aVR, aVL, and aVF as opposed to VR, VL, and VF; VR (right arm), VL (left arm), VF (left leg)) as a result of the use of a standard, but no longer necessary, modification of the original Wilson central terminal, designed to produce a 1.5-fold amplification in the recorded voltage.

The six limb leads (frontal plane leads)

The limb leads are remote from the heart and give (spatially) general rather than localized (spatially specific) information. In this respect they differ markedly from the precordial leads. The limb leads consist of the three bipolar leads (leads I, II, and III) and the three augmented, unipolar leads aVR, aVL, and aVF. The orientation around the heart of the six limb leads is illustrated in Fig. 3. The orientation of leads aVR, aVL, and aVF with respect to the heart is intuitively obvious since the limbs act as linear conductors (like wires). The left arm connection is therefore effectively 'looking at the heart' from the left shoulder (i.e. the left arm is acting as part of the wire connecting the ECG machine to the patient's left shoulder). Similarly, the right arm connection 'looks at the heart' from the right shoulder and the foot lead connection from the pelvic area. One practical consequence of the fact that the limbs act as linear conductors is that it does not matter whereabouts on any given limb the electrode is attached. The orientation of the bipolar leads with respect to the heart is not intuitively obvious (simply because they are bipolar leads), but may be worked out from the known polarities of the conventional connections used in the bipolar leads. Thus, for example, since lead I is recorded with the left arm connected to the positive and the right arm to the negative terminal of the recorder, the position of lead I is effectively that obtained by subtracting the right arm vector from the left arm vector. To subtract vector R from vector L one reverses the direction of vector R and adds it to vector L. Inspection of Fig. 3 reveals that if this is done the resulting 'direction' of lead I is effectively horizontally to the left of the heart. In a similar manner it can be shown that the effective orientations of leads II and III with respect to the heart are as shown in Fig. 3.

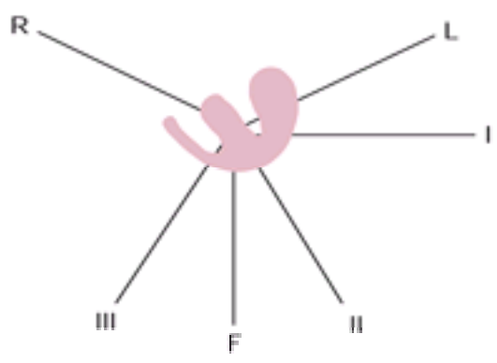

Fig. 3 The arrangement of the frontal plane leads. Note that leads II, III, and F are inferior to the heart, I and L are anterolateral to the heart, and R looks into the cavity of the heart.

The six precordial leads (chest leads)

For each precordial lead, the positive (recording) terminal is connected to an electrode at an agreed site on the chest wall. Since the connection to the negative terminal of the recorder is the 'indifferent' one formed by joining together leads R, L, and F, the chest leads are 'V' leads and are designated V_1, V_2, V_3, V_4, V_5, and V_6. Because the torso, unlike the limbs, acts as a volume conductor, the waveform obtained depends critically on the siting of the recording electrode. A standard anatomical siting of the precordial electrodes was agreed between the British Cardiac Society and the American Heart Association and is shown in Fig. 4. The important relationships of the precordial leads to the cardiac chambers are shown in Fig. 5.

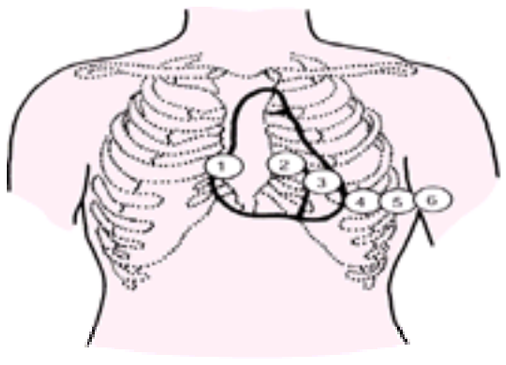

Fig. 4 The positions of the precordial leads. V1 is located at the right sternal margin in the fourth intercostal space, V_2 at the left sternal margin at the fourth intercostal space, V_4 at the intersection of the left midclavicular line and left fifth intercostal space, V_3 midway between V_2 and V_4, V_5 at the intersection of the left anterior axillary line with a horizontal line through V_4, and V_6 at the intersection of the left midaxillary line with a horizontal line through V_4 and V_5.

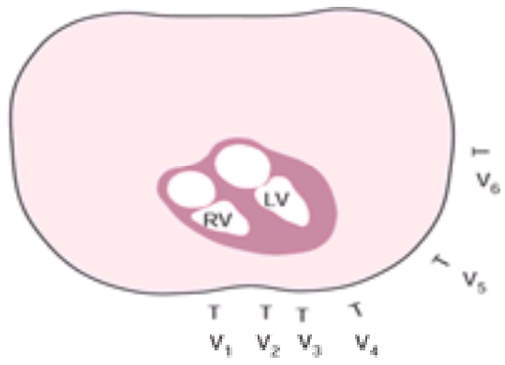

Fig. 5 The precordial leads and their important anatomical relationship to the main cardiac chambers.

The 12 conventional ECG leads

Figure 6 shows the relationship of the 12 conventional electrocardiographic leads to one another and to the heart.

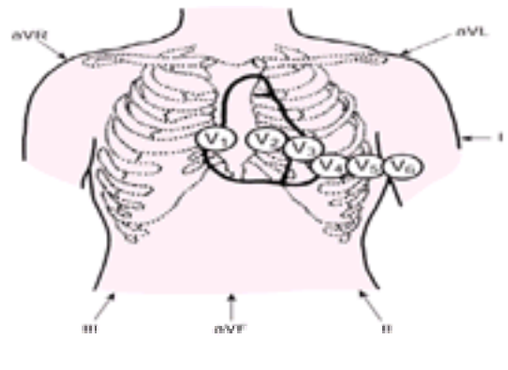

Fig. 6 The conventional 12 ECG leads and their relationships to the heart.

Recognizing the normal electrocardiogram

This is the most difficult and the most important aspect of understanding the electrocardiogram, which is recognized as being within or beyond normal limits by the normality or otherwise of the shape and dimensions of its various substituent deflections, and by the frequency of the deflections and their relationship in time to the deflections preceding and succeeding them. This introduction to the subject considers only morphological normality or abnormality. The presence of sinus rhythm will therefore be assumed. The criteria for normality of the P waves obtain in any rhythm where atrial depolarization is of sinus origin (sinus tachycardia, sinus bradycardia, sinus arrhythmia, first-, second-, or third-degree heart block). Those for the QRS complexes, S–T segments, and T waves obtain in any rhythm of supraventricular origin, provided the rate is not so rapid as to induce functional bundle-branch block. A supraventricular rhythm is one initiated at a site above the bifurcation of the His bundle.

All the criteria described below are dependent upon a normal (standard) calibration and a normal paper recording speed.

Precordial leads

Normal precordial P waves

The P waves are usually upright from V_4 to V_6. Upright or biphasic P waves may occur in V_1 and V_2. If the P waves are biphasic, the negative (terminal) component of the P wave should have an area no greater than the positive (initial) component.

Normal QRS appearances in the precordial leads

QRS morphology

The QRS complex in V_1typically shows a small initial positive wave followed by a larger negative wave, and in V_6a small initial negative wave followed by a large positive wave. In general, the size of the initial positive wave (r or R wave) increases progressively from $V_1$to V_6(Fig. 7(a)) The direction of the initial part of the QRS is generally upward (i.e. positive) in V_1 to V_3, and downward (i.e. negative) in V_4 to V_6That is, V_1 to V_3 show initial r waves and V_4 to V_6 initial q waves. Leads showing an rS complex are being primarily influenced by right ventricular myocardium, and leads showing a qR complex by left ventricular myocardium. The transition zone between right and left ventricular epicardial leads is seen (Fig. 7(b)) to be between V2 and V4. When the transition zone falls outside this region the heart is said to be rotated. If the transition zone occurs further to the left in the precordial series (for example, between V_5 and V_6) then the heart is said to be clockwise rotated. Conversely if the transition zone is moved to the right in the precordial series, the heart is said to be counter-clockwise rotated. Clockwise and counter-clockwise rotation refer to a normal state of variability between one subject and another and are not in themselves indicative of abnormality. More extensive works should be consulted for a detailed understanding of clockwise and counter-clockwise rotation. Although, as stated above, V_1usually shows an rS complex and V_6a qR complex, it is also possible for V_1 to show a QS complex and for V_6 to show a monophasic R wave, a QRS complex, or an Rs complex (Fig. 7(c)).

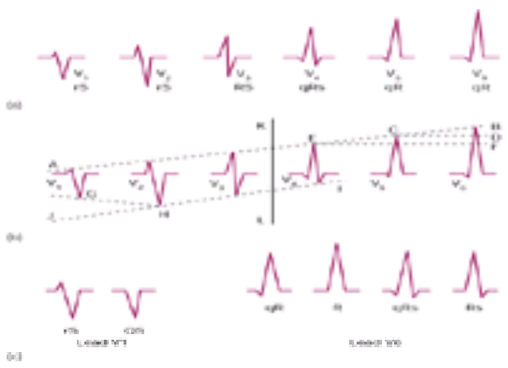

Fig. 7 Morphology of the precordial QRS complexes. (a) Typical normal QRS morphology of the precordial leads. (b) Normal variations of R-wave amplitude and S-wave depth in the precordial leads. The R wave in each precordial lead is usually larger than in the preceding lead in the series from V_1 to V_6 (line AB). However, it is quite normal for the R wave in V_6 to be smaller than that in V_5 (line CD) or for the R wave in V_5 to be smaller than that in V4, provided that the R wave in V_6 is also smaller than that in V_5 (line EF). The size of the S wave diminishes progressively across the precordial leads (line JI), although the S wave in V_2 is often greater than that in V_1 (line GHI). Leads before line KL have an initial deflection which is positive, while those after line KL have an initial deflection which is negative. This line marks the transition zone between right and left ventricular QRS configurations (i.e. between rS and qR configurations, respectively). (c) Possible normal QRS configurations in leads V1 and V6. Typically, V_1 has an rS configuration, but a QS configuration is also normal in this lead. Typically, V_6 has a qR configuration, but it is also normal for V_6 to show an R wave, a qRs complex, or an Rs complex.

QRS dimensions

The dimensions of the individual waves making up each part of the precordial QRS complexes are of crucial importance in determining normality or otherwise. Figure 8 shows how measurements within the QRS complexes are obtained. The criteria for normality of these individual waves are:

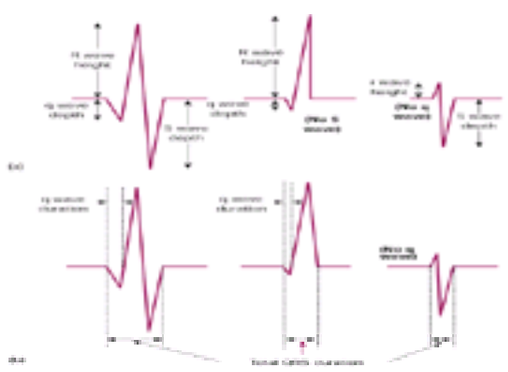

Fig. 8 The dimensions of constituent waves within QRS complexes. (a) Wave voltage measurements. (b) Wave duration measurements.

1. *minimum voltage*: at least one R wave in the precordial leads must exceed 8 mm in height;
2. *maximum voltage*: (a) the tallest R wave in the left precordial leads must not exceed 27 mm, (b) the deepest S wave in the right precordial leads must not exceed 30 mm, (c) the sum of the tallest R wave in the left precordial leads and the deepest S wave in the right precordial leads must not exceed 40 mm;
3. *maximum duration*: the total QRS duration in any one precordial lead must not exceed 0.10 s (2.5 small squares);
4. *q-wave criteria*: (a) no precordial q wave should equal or exceed 0.04 s (1 small square), (b) precordial q waves must not have a depth greater than a quarter of the height of the R wave in the same lead; and
5. *the ventricular activation time*, also known as 'intrinsic deflection time', in leads facing the left ventricle (i.e. showing qR complexes) must not exceed 0.04 s (one small square).

Normal precordial S–T segments

There is a single rule for normality of the S–T segment. It must not deviate by more than 1 mm above or below the isoelectric line in any precordial lead. The isoelectric line is that vertical position of the ECG recording when no part of the heart is being depolarized or repolarized (i.e. the interval between the end of one T wave and the beginning of the next P wave—the T–P interval).

Normal precordial T waves

The criteria given below for normality of the T waves are applicable to adults only.

T waves in lead V_1

In this lead 80 per cent of normal adults have upright T waves and 20 per cent have flat or inverted T waves. Therefore, the finding of an inverted T wave in Vl cannot be considered an abnormality (unless it was upright in a previous ECG).

T waves in lead V_2

About 95 per cent of normal adults show upright T waves and 5 per cent have flat or inverted T waves in V_2 Therefore there is a 1 in 20 possibility of inverted T waves in V_2 occurring by chance and not indicating an abnormality. However, if the T wave in V_2 is inverted when it was formerly upright, it is abnormal. Further, if there is T-wave inversion in V_2 with an upright T wave in V_l then the T wave in V_2 is abnormal.

T waves in leads V_3 to V_6

The T wave is normally upright in these leads. T-wave inversion in V_4, V_5, or V_6 is always abnormal. T-wave inversion in V_3, as well as in V_l and V_2, may (rarely) be found in healthy young adults.

T-wave size

There are no strict criteria for T-wave size. In general, the tallest precordial T wave is found in V_3 or V_4, and the smallest in V_1 and V_2,and, as a general rule, the T wave should not be less than one-eighth and not more than two-thirds of the height of the preceding R wave in each of the leads V_3 to V_6.

Limb leads

Normal limb lead P waves

The limb lead that normally best shows the P wave is lead II. In this lead the normal P-wave duration does not exceed 0.12 s and its height does not exceed 2.5 mm

Normal limb lead QRS complexes

Only three criteria need to be applied to the limb leads to determine the normality or otherwise of the QRS complexes:

1. *the size of any q waves* in aVL, I, II, or aVF;
2. *the size of the R waves* in aVL and aVF; and
3. *the electrical axis of the heart*.

Q waves

Any q wave present in lead I, II, or aVF must not exceed one-quarter the height of the ensuing R wave and must not equal or exceed 0.04 s in duration. Any q wave present in aVR or lead III should be ignored, irrespective of its size. Q waves present in aVL should fulfil the same criteria as those in leads I, II, or aVF unless the frontal plane QRS axis is more positive than +60°, in which case large q waves in aVL are acceptable, since aVL is then a cavity lead (and may therefore have a QS complex as aVR, which is virtually always a cavity lead, usually has).

R-wave size

The R wave in aVL must not exceed 13 mm, while that in aVF must not exceed 20 mm.

The frontal plane axis

The value of the axis is determined using the hexaxial reference system (Fig. 9) in which lead I is arbitrarily assigned the value 0°, with the convention that rotation clockwise from this point is denoted positive ('+') and rotation anticlockwise is denoted negative ('–'). The mean frontal plane QRS axis in the adult lies within the range from –30° to +90° (travelling clockwise). The frontal plane axis represents the dominant direction of ventricular myocardial depolarization on the frontal plane. The axis is closest to that frontal plane lead with the tallest QRS complex, but because of the uneven distribution of the limb leads the axis cannot reliably be determined by simple inspection of the limb lead QRS complexes.

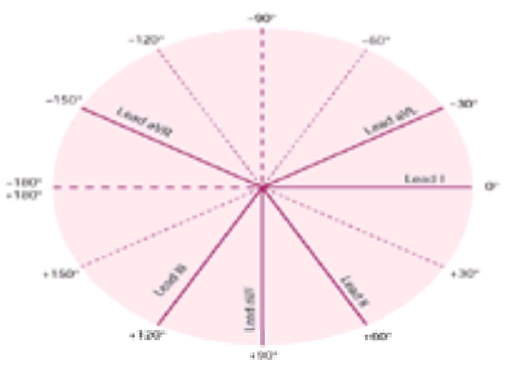

Fig. 9 The hexaxial reference system. The hexaxial reference system is constructed by taking each of the six frontal plane limb leads (continuous lines) and extrapolating them through the origin to create six additional lines (discontinuous), thereby producing 12 lines each at 30° to its neighbour on either side thus dividing the 360° of the frontal plane into equal (30°) sections.

Thegeneral direction of ventricular depolarization in the frontal plane in a normal record is typically towards lead II (60°), but can lie anywhere between aVL (–30°) and aVF (+90°). When the axis lies within the range –30° to +30° the heart is said to be horizontal and the tallest QRS complexes will be found in aVL and I. When the axis lies within the range +60° to +120° the heart is said to be vertical and the tallest QRS complexes will be found in aVF and II. With a typical normal axis of +60°there areupright QRS complexes in lead II, and to a lesser extent in aVF, and smaller QRS complexes in III and aVL. Lead aVR 'looks into' the cavity of the heart. Since depolarization of the ventricles is from endocardium (where the Purkinje tissue is) to epicardium, the cavity leads record depolarization travelling away from them and therefore have negative QRS deflections.

A knowledge of the axis is helpful in understanding the variation in appearances in the limb leads between different subjects with similar (normal or abnormal) precordial ECG appearances. Thus, for example, in a patient with left ventricular hypertrophy, if the heart is horizontal (axis lies in the region of –30° to +30°) appearances similar to those in V_5 and V_6 will be seen in aVL and I (Fig. 10), whereas if the heart is vertical (axis +60° to +120°) appearances similar to those in V_5 and V_6 will be seen in II and aVL.

Fig. 10 Left ventricular hypertrophy. There is also evidence of left atrial hypertrophy.

Determination of the mean frontal plane QRS axis

To determine the mean frontal axis to the nearest 30° requires two steps, which can be illustrated using Fig. 10 as follows:

1. Inspect the six frontal plane (limb) leads to find the lead in which the algebraic sum of the deflections in the QRS complexes approximates to zero (i.e. the lead in which the sum of all the positive deflections minus all the negative deflections gives a result closest to zero). In doing this one is looking for the smallest mean QRS deflection. The axis will always be approximately at right angles to the lead with the smallest QRS. Let the lead showing the smallest net QRS size be called lead 'x'. (In the example of Fig. 10 this is lead II.)
2. Using the hexaxial reference diagram (Fig. 9), decide which other lead (y) is at right angles to lead 'x'. (In Fig. 10 this is aVL.) There will always be one such a lead. The axis must either be: (a) in the same direction as or very close to lead y (in this example, aVL, i.e. –30°) or (b) directly opposite to y or very close to that

position (in this example, +150°), both of which possible positions are approximately at right angles to x. To determine which of these two possibilities indicates the correct orientation of the axis look again at the given ECG and inspect the QRS in lead y (in this case aVL). It must either have a large dominant positive wave, or a large dominant negative wave. If the former, the axis of the heart is along lead y (i.e. approximately –30°), if the latter the axis is directly away from lead y (i.e. approximately +150°). To the nearest 30°, therefore, the axis in the ECG of Fig. 10 is –30°.

3. To assess the axis to the nearest 15° it is necessary to take one further step
4. Look again at the QRS in that lead where the algebraic sum of QRS deflections is close to zero (II in Fig. 10). Assess whether that sum is (a) indistinguishable from zero, (b) close to zero but clearly slightly positive, or (c) close to zero but clearly slightly negative. If (a), then the current estimate of the axis is now correct to the nearest 15°; if (b), then the axis estimate must be 'bent' from the 30° accuracy estimate, 15° towards the lead where the QRS algebraic sum is close to zero; if (c), then the axis estimate must be 'bent' from the 30° accurate estimate, 15° away from the lead where the QRS algebraic sum is close to zero. In this case the QRS in II is clearly slightly positive and the 30° accuracy estimate (–30°) must be 'bent' 15° towards lead II, giving an axis of –15 (to a 15° accuracy). (Had the algebraic sum of QRS deflections in lead II, Fig. 10, been negative it would have been necessary to 'bend' the estimated axis 15° away from lead II, giving –45°).

Significance of the mean frontal plane QRS axis

The normal range for the frontal plane axis in adults is from –30°to +90° (travelling clockwise). Axes that are more negative than –30° are described as left-axis deviation (LAD). Axes more positive than +90° are described as right-axis deviation (RAD). These terms (LAD and RAD) are descriptive, in the same manner as the term 'hypertensive' is descriptive. Axes of –60°and –90° are both examples of left-axis deviation, but the difference is significant. It is therefore preferable to describe the axis in an individual case in degrees, just as it is preferable to give their measured blood pressure in mmHg. The other reason for preferring to describe the axis in degrees is that axes in the quadrant from –180° to –90°, which are uncommon, could be described either as extreme left- or as extreme right-axis shift.

The axis in the newborn tends to be around +120°. As the child grows the axis swings towards the left. In adult life the normal axis lies within the range –30° to +90°. In older age groups the axis typically lies between –30°and +30°. Tall, thin people have axes nearer the right end of the normal range and short fat people have axes nearer the left end of the normal range. The most common normal axis lies between 0° and +60°.

Deviation of the axis to the right, beyond +90 degrees (that is, RAD) occurs in cor pulmonale, right ventricular hypertrophy (pulmonary stenosis, pulmonary hypertension), and ostium secundum atrial septal defect.

Left-axis deviation occurs in left ventricular disease, when the superior division of the left bundle-branch system is blocked, in hyperkalaemia, in some cases of inferior infarction, and in ostium primum atrial septal defect.

Determination of the axis itself is primarily important in the diagnosis of right ventricular hypertrophy (RVH) and left anterior hemiblock (LAH). When the axis is in the range –30° to +30°, the heart is said to be 'horizontal'. When the axis is in the region of +60° to +120° the heart is said to be 'vertical'.

Normal limb lead S–T segments

Normal S–T segments do not deviate above or below the isoelectric line by more than 1 mm.

Normal limb lead T waves

In general, the T waves and QRS complexes in the limb leads are concordant: that is to say, when the QRS complexes are upright, the T waves are upright, and when the QRS complexes are negative, the T waves are negative. A normal T wave will always be negative in aVR and positive in I and II. T waves can be positive or negative in aVL, aVF, and II without necessarily indicating abnormality. A rough guide to assess normality of the T waves in the limb leads is:

1. in any lead in which the QRS is predominantly upright, the T wave must be clearly upright;
2. in any lead in which the QRS is predominantly negative, the T wave should be clearly negative;
3. in any lead in which the algebraic sum of QRS deflections is close to zero, the T wave may be positive or negative (though small in either case) or isoelectric (flat); and
4. the normal T wave is always upright in leads I and II.

Myocardial hypertrophy

Appreciable hypertrophy of the right or left ventricle produces characteristic changes in the electrocardiogram. Lesser degrees of hypertrophy may be present without electrocardiographic changes or with only minor, non-specific changes. This is more often true of right than of left ventricular hypertrophy.

Left ventricular hypertrophy

Left ventricular hypertrophy is not an all-or-none phenomenon, and the same is true of its recognition on the ECG. Several scoring systems have been devised. The most sensitive (and least specific) criteria consider the precordial lead voltages. The increased bulk of the left ventricle increases the voltage that is induced during left ventricular depolarization. This results in taller R waves in the left precordial leads and deeper S waves in the right precordial leads. The increased ventricular wall thickness also results in prolongation of the time taken for the depolarization wave to travel from endocardium to epicardium, that is to say it increases the ventricular activation time. In addition, secondary changes in depolarization occur, altering the S–T segments and T waves. The electrocardiographic criteria for left ventricular hypertrophy are:

1. at least one R wave in the left precordial leads exceeds 27 mm;
2. at least one S wave in the right precordial leads exceeds 30 mm:
3. the sum of the tallest R wave and the deepest S wave in the precordial leads exceeds 40 mm;
4. the largest positive or negative deflection in the limb leads exceeds 20 mm;
5. the intrinsic deflection time (ventricular activation time) exceeds 0.04 s;
6. S–T-segment depression and T-wave inversion may occur in the left precordial leads and in those limb leads that face the left ventricle.

The presence of one or more of the above abnormalities suggests the presence of left ventricular hypertrophy, only provided that the total QRS duration does not exceed 0.10 s. The greater the number of criteria fulfilled, the more confident one can be of the diagnosis. The voltage criteria have the greatest sensitivity, and the intrinsic deflection time the greatest specificity. However, one must exercise caution in diagnosing left ventricular hypertrophy on the basis of voltage criteria alone, especially if the patient is slim—and also note that the sensitivity for the diagnosis of left ventricular hypertrophy is very low in those who are obese. An example of clear-cut electrocardiographic changes of left ventricular hypertrophy is shown in Fig. 10.

Right ventricular hypertrophy

Increased bulk of the right ventricle gives rise to higher voltages during right ventricular depolarization, increasing the size of the positive deflection in the right precordial leads. In addition, it shifts the electrical axis towards the right and changes the S–T segments and T waves in leads facing the right ventricle, because of secondary changes in the repolarization process. The electrocardiographic criteria for right ventricular hypertrophy are:

1. a positive deflection equal to or greater than the negative deflection in V_1 (RS, Rs, qR, rR') in the presence of a normal total QRS duration;
2. a mean frontal plane QRS axis more positive than +90°; and
3. S–T-segment depression and T-wave inversion in right precordial leads.

The more features present, the more convincing is the electrocardiographic evidence for right ventricular hypertrophy, but, in general, the combination of a dominant positive deflection of the QRS in V_1 and an abnormal degree of right-axis deviation (axis more positive than +90°) establishes the diagnosis. Examples are shown in Fig. 11(a) and Fig. 11(b). In both examples there is abnormal right-axis deviation and a dominant R wave in V_1. Figure 11(a) shows an Rs complex in $V_1$and Fig. 11(b)

shows a qR complex. The more common finding is of an Rs in V_1. An initial q wave (qR complex) can appear in V_1 in right ventricular hypertophy, which results from hypertrophy of the right side of the upper part of the interventricular septum with resultant redirection of initial septal depolarization relatively posteriorly and therefore away from (and negative with respect to) V_1. Figure 11(b) also shows pronounced clockwise cardiac rotation, which often accompanies right ventricular hypertrophy.

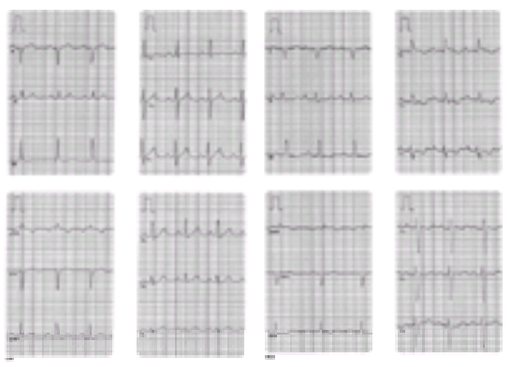

Fig. 11 Two examples of right ventricular hypertrophy. There is also right atrial hypertrophy. Fig. 11(a) shows an Rs complex in V_1;(b) shows a qR complex.

Atrial hypertrophy

The electrocardiographic changes produced by left atrial hypertrophy are those produced by an increase in the voltage and duration of the left atrial depolarization wave. Since the terminal part of the normal P wave is produced by left atrial depolarization, it follows that the total P-wave duration is prolonged in left atrial hypertrophy. In addition, the P wave tends to be bifid in lead II and biphasic in V_1. In $V_1$the area of the (terminal) negative component exceeds the area of the (initial) positive component (Fig. 10 and Fig. 12). Left atrial hypertrophy is most commonly seen in association with left ventricular hypertrophy and (in countries where rheumatic heart disease is rare) is most commonly due to systemic hypertension. The finding of left atrial hypertrophy in the absence of evidence of left ventricular hypertrophy suggests mitral valve obstruction (mitral stenosis or, very rarely, left atrial myxoma).

Fig. 12 Left atrial hypertrophy. Broad, bifid P waves in lead II. Biphasic P waves in V1, with dominant negative component. Since there is no evidence of left ventricular hypertrophy, the most likely cause of the left atrial hypertrophy is mitral valve obstruction. The patient had mitral valve stenosis.

The electrocardiographic change produced by right atrial hypertophy is an increase in the peak voltage of the P wave. This is usually best seen in lead II. In lead II the P-wave voltage is abnormal when it exceeds 3 mm (see Fig. 11(b)). Right atrial hypertrophy is most commonly seen in the presence of right ventricular hypertrophy. The finding of right atrial hypertrophy in the absence of right ventricular hypertrophy suggests tricuspid valve obstruction.

Bundle-branch block

Total failure of conduction in the right or left branches of the bundle of His (bundle-branch block) can only be diagnosed with confidence from the appearances in the precordial leads, although necessarily there are also changes in the appearances in the limb leads.

Right bundle-branch block

In right bundle-branch block, the primary change induced is a delay in depolarization in the right ventricular free wall. This results in the development of a second positive wave in the right ventricular leads (and a second negative wave in left ventricular leads), and prolongs the total QRS duration. The essential electrocardiographic features of right bundle-branch block are:

1. a total QRS duration of 0.12 s or more; and
2. the presence of a secondary positive wave in V_1 (rsR', rR').

In addition, secondary changes occur, but these are not in themselves essential for the definitive diagnosis. They include:

1. deep and slurred S waves in lead I, aVL, and V_4–V_6; and
2. secondary S–T, T changes in leads V_I–V_3.

An example of the appearances in right bundle-branch block is shown in Fig. 13.

Fig. 13 Right bundle-branch block. The total QRS duration is prolonged. There is an rsR' in V_1 and there are broad, slurred S vaves in V_3–V_6.

Left bundle-branch block

Left bundle-branch block induces changes in the ECG which are more extensive than those produced by right bundle-branch block. In left bundle-branch block not only is depolarization of the free wall of the left ventricle delayed (a precise corollary of the changes in right bundle-branch block), but also the direction of depolarization of the interventricular septum is from right to left instead of from left to right as in the normal electrocardiogram. This reversal of the direction of septal depolarization gives rise to widespread and major alterations in the QRS complexes in every lead of the electrocardiogram. The diagnostic criteria for left bundle-branch block are:

1. a total QRS duration equal to or in excess of 0.12 s; and
2. absence of the normal (septal) q waves in lead I, aVL, and V_4–V_6; and
3. absence of a secondary r wave in V_1.

This latter criterion is necessary to prevent confusion in cases of right bundle-branch block occurring in the presence of pronounced clockwise cardiac rotation, which gives a loss of q waves in left ventricular leads. (The finding of a secondary r wave in V_1 in the presence of an abnormally wide QRS complex indicates the presence of right bundle-branch block.) Secondary changes also inevitably occur, but these are not part of the diagnostic process. These include:

1. secondary S–T depression and T-wave inversion in leads I, aVL, and V_4–V_6;
2. broad QS waves in V_1–V_3;
3. notching of the R waves giving rise to rsR' or 'M-shaped' QRS complexes, and
4. broad, R waves in leads I, aVL, and V_4–V_6.

An example of the ECG appearances in left bundle-branch block is shown in Fig. 14. The changes in left bundle-branch block so disturb the normal pattern of the ECG that none of the usual criteria can be applied for determining any other abnormality of the QRS complexes, S–T segments, or P waves. When left bundle-branch block is present, a diagnosis of right or left ventricular hypertrophy, myocardial ischaemia or infarction, or of non-specific changes in the S–T segments and T waves cannot easily or reliably be made.

Fig. 14 Left bundle-branch block. The total QRS duration is prolonged and no q wave is seen in I, avL, and V_4–V_6. In addition there is no secondary R wave in V_1.

The hemiblocks

During normal intraventricular conduction the anterosuperior and the posteroinferior divisions of the left bundle-branch system conduct more or less simultaneously. This has the result that the dominant direction of depolarization of left ventricular myocardium lies midway between these two divisions and is therefore downwards and to the left, giving an axis in the range –30° to +90° and most typically in the region of +60°.

Left anterior hemiblock

When there is failure of conduction in the anterosuperior division of the left bundle branch ('left anterior hemiblock') the direction of the initial part of the QRS vector moves marginally to the right as a result of the loss of the opposing initial leftward voltage. However, the bulk of the QRS vector is slightly delayed and moves strongly to the left. This is because depolarization of the superior and left part of the left ventricular myocardium is dependent upon spread from the initially depolarized rightward and inferior aspects of the ventricle. The resulting changes in the ECG are therefore (1) a trivial increase in the QRS duration (but the QRS duration is not prolonged beyond normal limits) and (2) a shift of the axis to the left (more negative than –30°). The other common cause of abnormal left-axis deviation is inferior myocardial infarction. The ECG criterion for left anterior hemiblock is a mean frontal plane QRS axis more negative than –30° in the absence of abnormal q waves in aVF (which would indicate inferior infarction).

Left posterior hemiblock

Left posterior hemiblock, as would be expected from the above, gives rise to abnormal right-axis deviation (axis more positive than +90°). Unfortunately, there are numerous causes of abnormal right-axis deviation. It is therefore not possible, from the 12-lead ECG, to diagnose left posterior hemiblock. One can only raise the possibility of this in those situations in which there is an abnormal degree of right-axis deviation without any clear electrocardiographic or clinical explanation.

Ischaemic heart disease

ECG changes in ischaemic heart disease are very variable, depending on the site and severity of ischaemic damage. However, certain patterns are commonly produced.

Acute coronary syndromes (Q-wave infarction, non-Q-wave infarction, and unstable angina)

The terms 'transmural infarction' and 'Q-wave infarction' are often used interchangeably, as are the terms 'subendocardial infarction' and 'non-Q-wave infarction'. These two types of myocardial infarction are not separable on clinical grounds alone, and the distinction between the two depends entirely on the presence or absence of abnormal Q waves on the ECG. It is generally agreed that the terms 'Q-wave infarction' and 'non-Q-wave infarction' are preferable to 'transmural infarction' and 'subendocardial infarction', because autopsy data indicate that the ECG does not have sufficient sensitivity and specificity to guarantee reliable distinction between transmural and subendocardial infarction. However, the distinction between Q-wave and non-Q-wave infarction should not be used to determine management. The full spectrum of Q-wave infarction and unstable angina is now designated under the overall label of 'Acute coronary syndromes'.

Patients presenting clinically with an acute coronary syndrome may or may not have initial S-T elevation. Most (75 per cent) of those with S-T elevation subsequently develop abnormal Q waves and most of the remaining 25 per cent develop reduction in R wave height. Occasionally the ECG may return entirely to normal after initial S-T elevation. Those who present without initial S-T elevation do not usually develop abnormal Q waves. In such cases there may be (non-specific) S-T segment depression (flat, down sloping or up sloping), flattening or inversion of the T waves or a combination of non specific S-T and T changes and the final diagnosis can then be either non-Q-wave infarction or unstable angina. The distinction between these two depends entirely on whether or not serum markers of infarction are present.

In contrast to the typical, striking evolutionary changes of acute Q-wave infarction, the ECG findings in non-Q-wave infarction are less dramatic and less predictable. The electrocardiographic evidence consists of primary S–T-segment depression ('primary' implies in the absence of any detectable S–T elevation) or of deep,

symmetrical T-wave inversion without any change in the QRS complexes. Each of these changes can be produced by myocardial ischaemia without infarction, and the diagnosis of non-Q-wave infarction cannot be made on a single electrocardiogram alone. Either (1) a single such record accompanied by clinical or enzyme evidence of infarction, or (2) serial records that persistently show primary S–T depression or that show deep symmetrical T-wave inversion, is required. The persistence of non-specific S–T and/or T changes that were not formerly present, and which accompany clinical and enzymatic evidence of acute infarction, are the hallmark of non-Q-wave infarction. When the primary change is S–T depression it will often be visible in all or most leads, with the exception of the cavity leads (aVR is always a cavity lead, aVL is a cavity lead if the heart is vertical (axis +60° or more positive) and lead III is a cavity lead if the heart is (axis +30° or further to the left)). The cavity leads may also show reciprocal S–T elevation. The only exception to the rule 'that when S–T elevation and S–T depression are both present in the same recording, it is the elevation which is the primary change' occurs when, as in this case, the reciprocal S–T elevation occurs in cavity leads (which, by definition, all show QS complexes).

QRS changes of myocardial infarction

Two QRS changes are indicative of myocardial infarction. These are:

1. inappropriately low R-wave voltage in a local area; and
2. abnormal q waves.

These two changes represent parts of the same process. The development of increased negativity (abnormal q waves) and the reduction in the normal positivity (loss of R-wave height) of QRS complexes in the precordial leads each results from a loss of underlying viable muscle, with a consequent reduction in the normally generated positive voltage. When there is full-thickness (transmural) myocardial infarction in an area of myocardium underlying the precordial leads there is total loss of the positive deflection. In this situation a totally negative wave (QS complex) occurs. This totally negative wave occurs as a result of depolarization of the posterior wall of the ventricle travelling (posteriorly) from endocardium to epicardium in the normal way, which is no longer swamped by the usual simultaneous and dominant depolarization towards the exploring electrode of the anterior wall of the ventricle.

The normal precordial QRS complexes show a progressive increase in the R-wave height from V_1 to V_6 (Fig. 15). The positive (upgoing) part of the deflection in each precordial lead is predominantly the result of depolarization from underlying endocardium to epicardium. In the presence of infarction of part of the left ventricle, the positive waves overlying the necrotic area will be reduced in size (Fig. 16). Loss in R-wave height can only be used as a criterion for myocardial infarction if either: (1) larger, normal R waves are visible on both sides of the infarcted zone, or (2) previous ECGs are available demonstrating the normal R-wave height for that particular lead in that particular subject. If a major part of the thickness of the myocardial wall is infarcted, the positive wave generated by any remaining viable left ventricular myocardium underlying the electrode is insufficient to overcome the negative deflection induced by the normal depolarization of the interventricular septum, from left to right, and of the free wall of the right ventricle (or the posterior wall of the left ventricle) from endocardium to epicardium. In this situation an abnormal q wave will develop. In the precordial leads, a q wave is abnormal if its duration is equal to or in excess of 0.04 s or if its depth is equal to or greater than a quarter the height of the ensuing R wave in that lead. In Fig. 16, the q wave in V_4 satisfies this criterion. If the infarction involves the full thickness of the ventricular wall (transmural infarction), no R wave is generated at all and an entirely negative (QS) wave develops (Fig. 17). Figure 18 shows (in diagrammatic form) the appearances produced in the precordial leads when infarcts of varying thickness occur under each of three precordial electrodes. The QRS complex in V_3 is of QS type and indicates transmural infarction at this site. The appearances in V_4 indicate a substantial loss of myocardium underlying that electrode. The q wave is abnormal in duration and depth. The appearances in V_5 indicate a thinner zone of infarction. The q wave is not, in itself, abnormal, but the R-wave height is less than would be predicted from the height of the R waves present in V_2 and V_6.

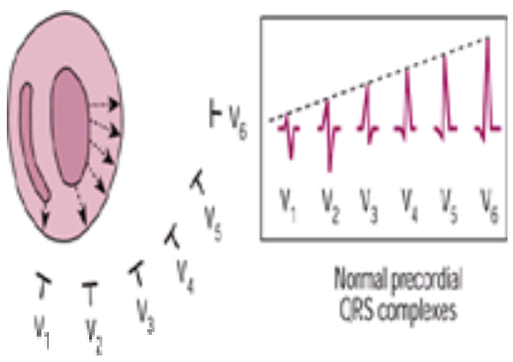

Fig. 15 Normal R-wave progression in the precordial series.

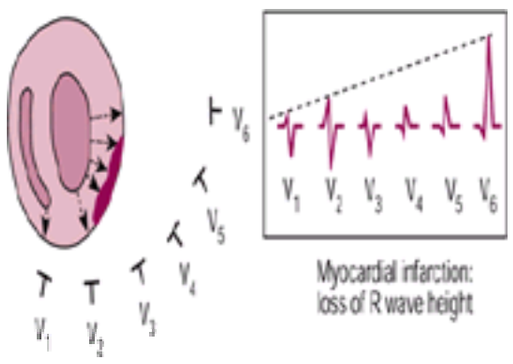

Fig. 16 Loss of R-wave height in myocardial infarction. The R-wave height is reduced in leads V3 to V5.

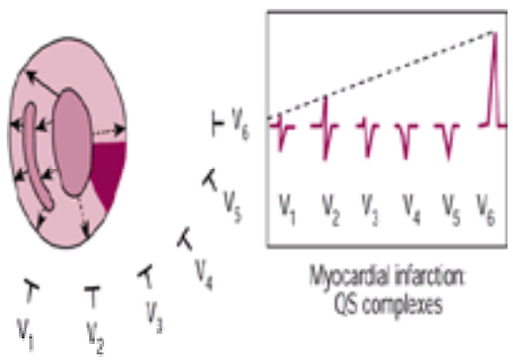

Fig. 17 Transmural myocardial infarction. QS complexes are seen from V_3 to V_5.

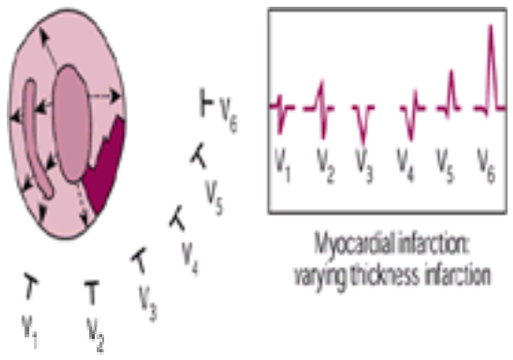

Fig. 18 Varying thickness infarction.

The diagnosis of myocardial infarction from the limb leads depends entirely on the presence of abnormal q waves. Q waves of any size may be seen in the normal ECG in aVR and in lead III. In leads I, II, and aVF, q waves which are equal to or greater than 0.04 s in duration or which have a depth in excess of a quarter the height of the ensuing R wave are abnormal and, unless a defect of intraventricular conduction is known to be present, indicate myocardial infarction. The same is also true of abnormal q waves in aVL, except when the mean frontal plane QRS axis is equal to or more positive than +60°, for in this situation aVL becomes a cavity lead like aVR.

S–T-segment changes of infarction

Only changes in the QRS complexes provide definitive electrocardiographic evidence of infarction, but an S–T-segment shift occurs in the acute stages of Q-wave infarction. Strictly speaking, this shift is evidence of injury to, rather than infarction of, the myocardium. Thus, although in the vast majority of cases the development of typical S–T-segment elevation is followed by the development of definitive QRS changes, occasionally the ECG with S–T-segment elevation of myocardial injury will revert to normal within hours. This does not happen if definitive QRS changes of infarction are also present. It is marginally more likely to occur following the use of thrombolytic therapy but is still relatively uncommon. The essential change of myocardial injury is deviation of the S–T segment above the isoelectric line. The S–T-segment shift must be in excess of 1 mm to be significant. Minor degrees of S–T-segment elevation in the right precordial leads are very common in normal ECGs, and S–T-segment elevation of up to 2 mm may be accepted as being within normal limits in V_1 and V_2. Significant S–T-segment elevation occurs in transmural and subepicardial infarction in leads facing the infarct. S–T-segment depression occurs in leads facing the infarct when it is subendocardial. Secondary S–T-segment depression also occurs as a reciprocal change (see below) in leads opposite to those showing the primary changes of acute infarction.

T-wave changes of infarction

A variety of T-wave changes occur in association with myocardial infarction. These include flattened, biphasic, and inverted (negative) T waves. None of these changes is specific. Whilst they are always abnormal in leads V_4 to V_6 and in those limb leads showing clearly upright QRS complexes, they may be caused by factors other than infarction or ischaemia, including electrolyte changes, digitalis effect, pericarditis, myocarditis, changes in body position, and changes in oesophageal temperature. T-wave changes are never, in themselves, reliable indicators of infarction, although characteristic T-wave changes do occur in relation to the latter. The most typical T-wave change associated with infarction is the development of deep, symmetrically inverted T waves (Fig. 19).

Fig. 19 Deep, symmetrical T-wave inversion. This is not truly specific but is typically found in association with myocardial ischaemia or infarction.

The sequence of ECG changes in Q-wave infarction

Any combination of the QRS, S–T-segment, and T-wave changes described above may occur in relation to acute infarction of the myocardium, but commonly a typical sequence of changes can be recognized in relation to Q-wave infarction (Fig. 20). Typically, S–T-segment elevation (which is convex upwards) appears within hours of the onset of symptoms. At this stage no change in the QRS complex can be recognized. Within 1 to 3 days, reduction in the R-wave height occurs, abnormally deep and broad q waves develop, some reduction in the extent of S–T-segment elevation occurs, and there is development of T-wave inversion. After the first few days the S–T-segment elevation disappears completely. The deep, symmetrical T-wave inversion typically persists for weeks before reverting to normal. The changes in the QRS complex are usually permanent. The QRS changes may occasionally disappear altogether if the infarct is small and the myocardial scar subsequently shrinks.

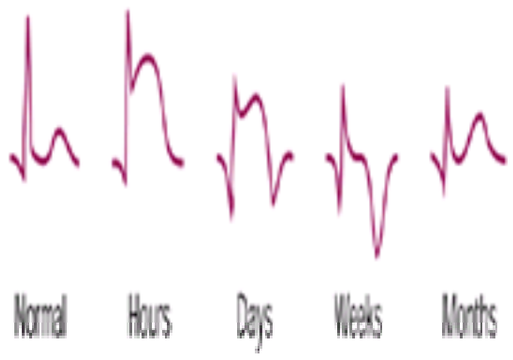

Fig. 20 Sequential changes in acute myocardial infarction.

Location of ECG changes in myocardial infarction

Primary ECG changes of the type described above occur in leads facing the infarct. It follows that the leads in which such primary changes occur indicate the location of the infarct (Table 1).

Reciprocal changes

In addition to the primary changes, 'reciprocal' changes occur in leads opposite those facing the infarct. Reciprocal changes are the inverse of primary changes (e.g. S–T-segment depression instead of S–T-segment elevation and tall, pointed T waves instead of symmetrical T-wave inversion). The inferior limb leads (II, III, and aVF) are reciprocal to the anterior leads (the precordial leads, lead I, and aVL) and vice versa. Examples of ECGs showing recent and old anterior and inferior infarctions are shown in Fig. 21, Fig. 22, Fig. 23, Fig. 24, and Fig. 25.

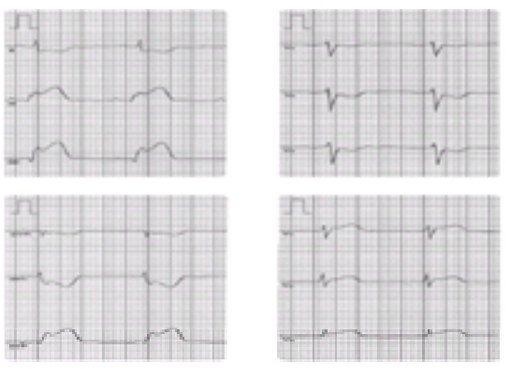

Fig. 21 Acute inferior myocardial ischaemic damage. Primary S–T elevation is visible in leads II, III, and aVF. S–T elevation is also visible in V_4–V_6, indicating that the damage extends to the lateral wall of the ventricle. There is reciprocal S–T-segment depression in I, aVL, aVR, and from V_1 to V_3

Fig. 22 Inferior myocardial infarction of intermediate age. The Q waves are abnormal in aVF (and also in III) and the q waves in II are borderline abnormal. There is T inversion in II, III, and aVF. The S–T segments are still minimally elevated in these leads. There is inversion of the terminal part of the T wave in V_5 and V_6, suggesting that the ischaemic area extends to the lateral wall of the ventricle. The T waves are strikingly tall in V_2 and V_3. This is not necessarily abnormal but could indicate true posterior ischaemia.

Fig. 23 Acute anteroseptal infarction. There is obvious S–T elevation in V_1–V_4 with minimal reciprocal S–T depression in III and aVF. There is obvious loss of initial R-wave height in V_2 and V_3.

Fig. 24 Old anterior myocardial infarction. There are QS complexes in $V_2$and V_3.

Fig. 25 Extensive anterior infarction. There are abnormally wide, abnormally deep Q waves from V_3–V_6 and in I. The T waves are of low voltage in most leads. This

latter abnormality is non-specific.

Pitfalls in the diagnosis of myocardial infarction

Left bundle-branch block

Left bundle-branch block so distorts the normal ECG that the usual criteria for the diagnosis of myocardial infarction are no longer applicable. It is sometimes possible to diagnose myocardial infarction in the presence of left bundle-branch block, but commonly the presence of left bundle-branch block obscures all other possible ECG diagnoses involving the QRS complexes, S–T segments, or T waves. The reason for this is that the two most important determinants of ventricular myocardial depolarization (and therefore also of repolarization) are radically altered in left bundle-branch block. These two determinants are (1) the direction of depolarization of the interventricular septum (which is reversed in left bundle-branch block), and (2) depolarization of the free wall of the left ventricle (which is appreciably delayed in left bundle-branch block). However, the changes of acute myocardial infarction can sometimes be recognized against a background of pre-existing left bundle-branch block, in which situation three independent criteria have been shown to have value in the diagnosis of acute infarction. These are: (1) S–T-segment elevation concordant with (i.e. in the same direction as) the QRS complex in a given lead; (2) S–T-segment depression of 1 mm or more in leads V_1, V_2, or V_3; and (3) S–T-segment elevation of 5 mm or more discordant with the QRS complex. Figure 26 shows an example of acute anterior infarction with pre-existing left bundle-branch block. There is in excess of 5 mm of S–T elevation concordant with the QRS in leads V_2 to V_4 and there is concordant S–T elevation in V_4 and V_5.

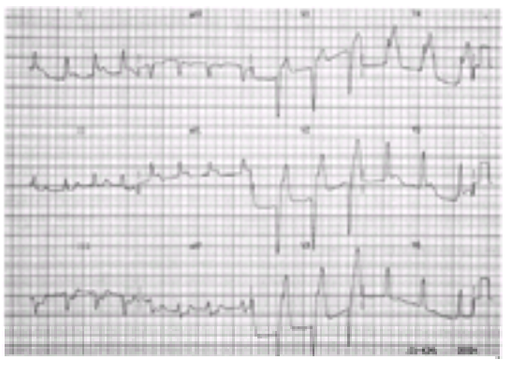

Fig. 26 Acute myocardial infarction with pre-existing left bundle-branch block. There is in excess of 5 mm of S–T elevation in leads V_2–V_4 and concordant S–T elevation in V_4 and V_5.

Right bundle-branch block

Right bundle-branch block does not interfere with the diagnosis of acute or of old myocardial infarction. (Fig. 27)

Fig. 27 Anteroseptal myocardial infarction with right bundle-branch block. Right bundle-branch block is diagnosed by the combination of (a) a total QRS duration of 0.12 s or more, plus (b) a secondary R wave in V_1. Anterior myocardial infarction is indicated by the presence of abnormally deep, abnormally wide Q waves in V_1–V_3. The presence of S–T elevation from V_2 to V_5 suggests that the infarct is recent.

Ventricular pre-excitation

When present, ventricular pre-excitation effectively precludes the diagnosis of left bundle-branch block and of myocardial infarction (see below). Very occasionally, substantial S–T elevation may provide an important clue to the diagnosis of acute infarction even when pre-excitation is seen.

Miscellaneous abnormalities

Abnormalities associated with the effects of drugs (including digitalis), hypokalaemia, hyperkalaemia, pericarditis, and hypothyroidism are discussed in other parts of this textbook. It is appropriate here, however, to mention ventricular pre-excitation, even though this is also dealt with in more detail elsewhere. The reason for this is that an incorrect diagnosis of bundle-branch block (right or left), ventricular hypertrophy (right or left), or myocardial infarction may easily be made if ventricular pre-excitation is present but not recognized.

Ventricular pre-excitation

Ventricular pre-excitation is found in approximately 0.1 per cent of the general population and failure to recognize it may lead to serious misdiagnosis. The term 'ventricular pre-excitation' implies that some part of the ventricular myocardium is depolarized (during normal sinus rhythm or any supraventricular rhythm) earlier than would be anticipated. This occurs as a result of the presence of one or more accessory (anomalous) pathways linking atrial and ventricular myocardium in such a way as to permit the depolarization wave descending through the atria from the sinoatrial node to bypass the atrioventricular node, partially or completely, intermittently or consistently. A variety of pathways exist which may, for example, pass (1) from atrial myocardium to ventricular myocardium, (2) from atrial myocardium to the distal part of the atrioventricular node, or (3) from the atrioventricular node to the ventricular myocardium. The commonest are those that link atrial and ventricular myocardium directly, and such pathways can exist at any point around the atrioventricular junction since, embryologically, the atrial and ventricular muscle masses were in continuity around the whole atrioventricular junction. These muscular remnants bear no resemblance to the junctional structures described by Kent; the use of the term 'Kent bundle' to describe the anomalous atrioventricular muscular connections that form the anatomical substrate for ventricular pre-excitation, though in widespread usage, is anatomically unjustifiable.

The presence of dual atrioventricular (AV) conduction routes (an accessory atrioventricular conduction pathway and the AV node) provides a substrate that facilitates the development of circus movement tachycardia. Patients with such pathways are at risk of developing paroxysmal tachycardia, but only a proportion do so: 10 per cent in the 20 to 39 age group and 36 per cent in the over-60s. In the presence of such a pathway, an appropriately timed atrial or ventricular premature beat may

initiate atrioventricular node re-entrant tachycardia (AVNRT).

Recognition of the presence and of the location of such pathways has assumed greater practical importance with the advent of the technique of radiofrequency ablation. This makes it possible to destroy, or at least to render non-functional, the accessory conduction pathway, thereby removing the anatomical substrate upon which the occurrence of AVNRT is dependent.

ECG appearances in the presence of anomalous AV pathways

Space constraints prevent a detailed discussion of this important topic, but the typical ECG appearances have the following features:

1. an abnormally short PR interval (0.11 s or less); and
2. an abnormally wide QRS complex (0.11 s or more), and
3. slurring of the initial 0.03 s of the QRS complex (a delta wave).

An example is shown in Fig. 28. This figure also illustrates the point that the presence of an accessory pathway does not guarantee that atrioventricular conduction will always occur via this route. Atrioventricular conduction can occur:

Fig. 28 Ventricular pre-excitation. (a) and (b) both show the essential features of ventricular pre-excitation. The PR interval is short in both and delta waves are seen in leads I, II, aVR, aVL, and V_2–V_6 in (a), and in leads I, II, III, aVR, aVL, aVF and V_2–V_6 in (b). In (a) the dominant R wave in V_1 and the 'pseudo RBBB' pattern indicates a left-sided AP. In (b) the small r waves in V_1 and the 'pseudo LBBB' pattern indicates a right sided AP.

1. always via the accessory pathway;
2. always via the AV node;
3. sometimes via the pathway and sometimes via the AV node;
4. simultaneously via the pathway and the AV node (giving 'fusion beats in respect of the QRS complexes.

When, in a given patient, conduction never occurs via the pathway, the 12-lead ECG never shows any evidence of ventricular pre-excitation and the pathway is said to be 'concealed'. Such pathways can sustain AVNRT in just the same way as revealed pathways: ventriculoatrial conduction in the re-entrant arrhythmia being via the pathway (which can therefore conduct backwards), and atrioventricular conduction occurring via the AV node.

When ventricular pre-excitation is seen on the 12-lead ECG no further interpretation of the electrocardiogram should be attempted in respect of the QRS complexes, S–T segments, or T waves, except in respect of assessment of the QRS complexes to try to determine the location of the accessory pathway. Ventricular pre-excitation may mimic left bundle-branch block, right bundle-branch block., left ventricular hypertrophy, right ventricular hypertrophy, and myocardial infarction.

Location of the accessory pathways

Accessory pathways may remain at any site around the atrioventricular junction, but the sites can be broadly categorized as:

1. left free wall (lateral) (55 per cent),
2. posteroseptal (25 per cent),
3. right free wall (15 per cent), and
4. anteroseptal (5 per cent).

The 12 lead ECG provides some information concerning the approximate location of any manifest accessory pathway (i.e. any pathway along which conduction is occurring at the time of the recording). Anterograde conduction along the pathway initiates early ventricular depolarization at the site of insertion of the pathway into the ventricular myocardium. Thus the initial part of the QRS complex (the delta wave in the case of a pre-excited beat) is determined by the ventricular location of the pathway and will be positive in V) in left sided accessory pathways (just as it is in right bundle branch block and in the case of left ventricular premature beats — two other situations in which left ventricular depolarization is initiated earlier than right ventricular depolarization). Similarly the initial part of the QRS in V, will be negative in right-sided accessory pathways (just as it is in left bundle branch block and in the case of right ventricular premature beats—two other cases which right ventricular depolarization is initiated earlier than left ventricular depolarization).

Numerous algorithms have been proposed to predict the location of accessory pathways from the QRS configuration. However, the majority of each QRS complex (i.e. excluding the delta wave) is determined by the relative contribution of (a) conduction along the accessory pathway and (b) conduction via the normal AV nodal route. This relative distribution varies greatly between subjects and can vary very significantly within subjects (being influenced, for example, by autonomic effects on AV nodal function). Furthermore the physical orientation of the heart within the chest, the presence of QRS abnormalities unrelated to pre-excitation, and the possibility of multiple accessory pathways are all potential confounding factors. It follows that the value of the surface ECG in determining the location of accessory pathways is limited.

As a general guide to the approximate location of an accessory pathway (AP), the following features are helpful. A left sided AP will typically give rise to a positive QRS in V_1 (i.e. a dominant R wave or an equiphasic RS complex) with a 'pseudo RBBB' pattern (Fig. 28(a)). A right sided AP will typically give rise to a small r wave in V_1 and V_2. There may be delta waves in V_6 with a 'pseudo LBBB' pattern (Fig. 28(b)). The smaller the r waves in the right precordial leads and the later in the precordial leads the RS transition (from dominant S in the right precordial leads to dominant R in the left precordial leads), the further to the right of the septum is the accessory pathway. The more posterior the location of the pathway, the more likely it is that there will be negative delta waves in the inferior leads and that there will be a superior QRS axis. The more anterior the pathway location the more likely it is that the QRS axis will be inferior.

The exercise ECG

Historical background

The present-day use of the exercise stress electrocardiogram in the diagnosis of coronary heart disease (in the form of the graded-exercise stress test—GXT) has evolved as a result of numerous observations and developments.

In 1908, Einthoven observed S–T depression after exercise but did not comment on it. In 1918, Blousfield recorded S–T-segment depression in leads I, II, and III during spontaneous angina. Feil and Siegel, in 1928, exercised patients known to have angina and observed S–T-segment and T-wave changes. Master and Oppenheimer, in 1929, developed an exercise test to assess 'circulatory efficiency' (using pulse and blood pressure) but did not use the ECG. In 1931, Wood and

Wolferth described S–T changes associated with exercise, but felt that the test was too dangerous to use in patients with coronary disease. In 1932, Goldhammer and Scherf reported S–T depression in 75 per cent of patients with angina—a figure indicating a remarkably similar false-negative rate to that of current-day studies. In 1941, Master and Jaffe suggested that the ECG recorded before and after exercise could be used to detect 'coronary insufficiency'. Paul Wood and colleagues, in 1950, at the National Heart Hospital in London, described their experience of a test in which the patients had to run up 84 steps adjacent to the laboratory. They showed an 88 per cent reliability (compared with 39 per cent in the Master's test) and emphasized that the amount of work required should be adjusted to the patient's physical capacity.

The era of modern, stress testing began in 1956 when Bruce reported his findings and established guidelines for a standardized GXT procedure. Subsequently, the application of Bayesian techniques of analysis; the addition of nuclear techniques (myocardial scintigraphy and cardiac blood pool analysis) and echocardiographic stress testing; and the use on non-exercise stress techniques (using dipyridamole, dobutamine, and adenosine) have all brought greater sophistication and applicability to cardiac stress testing.

This section will be confined to the use of the exercise stress ECG in the assessment of the heart and circulation and, in particular, to the role of the GXT in the detection and assessment of ischaemic heart disease.

Current usage

Although the exercise ECG may be used for several purposes, its commonest uses are in the diagnosis and assessment of ischaemic heart disease (IHD). In this respect, however, it is extremely important at the outset to recognize that the test has a significant false-negative rate, even in populations with an appreciable prevalence of IHD, and that the false-negative rate may be unacceptably high in populations with a low prevalence. The test is therefore of very limited value in screening low-risk, asymptomatic subjects. Most subjects who have undergone exercise stress testing as a screening procedure and who subsequently experience sudden cardiac death are found in retrospect to have had a normal exercise test result. A meta-analysis of 147 consecutive studies involving a total of 24 074 patients who had undergone both exercise stress testing and coronary angiography revealed sensitivities ranging from 23 to 100 per cent (mean 68) and specificities ranging from 17 to 100 per cent (mean 77). In patients with multivessel coronary disease the sensitivities ranged from 40 to 100 per cent (mean 81) and the specificities from 17 to 100 per cent (mean 66). For patients with single-vessel disease a positive GXT is most likely for lesions in the left anterior descending artery. Patients with lesions in the circumflex artery are least likely to give a positive result, while those with lesions in the right coronary artery occupy an intermediate position.

Exercise electrocardiography is also used in the estimation of prognosis in patients with known IHD, for risk stratification following myocardial infarction, for screening of professionals in high-risk situations (e.g. pilots and professional athletes), and in the assessment of some cardiovascular symptoms (e.g. palpitations, tachyarrhythmias, and syncope) when these are exercise related. The database for the evaluation of the usefulness of the technique in these situations is less well established than is the case in relation to its use in the assessment of IHD.

Exercise testing in females

The specificity of exercise testing is less in women than in men. It seems likely that this is, in part at least, related to their lower prevalence of IHD. However, biological differences might be relevant. It has been suggested that oestrogens (with certain chemical structural similarities to digitalis) contribute to S–T-segment depression, but it has also been pointed out that women secrete more catecholamines during exercise than men. Both of these postulated mechanisms have been thought possibly to act via coronary vasoconstriction.

Risks

High-level exercise carries a cardiovascular mortality risk, and a maximal-exercise stress ECG is, basically, supervised high-level exercise. Inevitably, therefore, a GXT carries a risk, but multiple studies have shown the risk to be remarkably low. In 1971 a survey of 73 medical centres summarized the risks in relation to approximately 170 000 stress tests. A total of 16 deaths were reported (mortality rate 0.01 per cent), and 0.04 per cent required admission within 24 h because of arrhythmia or prolonged chest pain. The risks are greater when the test is conducted soon after an ischaemic event. Even in this situation, however, the test is still remarkably safe. A survey of 151 941 tests undertaken within 4 weeks of acute myocardial infarction revealed a mortality rate of 0.03 per cent and a 0.09 per cent rate of non-fatal reinfarction or (successfully resuscitated) cardiac arrest.

Contraindications

Exercise stress testing is contraindicated to some extent whenever the pre-existing clinical state indicates a significantly increased risk of mortality or morbidity. In some situations the additional risk is so great as to constitute an absolute contraindication. In other situations the presenting clinical state indicates the need for more vigilant supervision than usual. Exercise, whilst not 'contraindicated', is of limited or negligible value in situations where abnormalities of the resting ECG make interpretation of the exercising record difficult or impossible.

Absolute contraindications

These include:

- acute ischaemic syndromes:
  - unstable angina,
  - suspected acute myocardial infarction,
  - known acute myocardial infarction within 5 days;
- known left main-stem stenosis;
- acute myocarditis;
- acute pericarditis;
- severe aortic stenosis;
- severe congestive cardiac failure;
- recent acute pulmonary oedema;
- current acute systemic illness;
- absence of trained supervisory staff or of resuscitation equipment;
- failure of the patient to understand the procedure or to give informed consent

Situations requiring intensive supervision

These include:

- known severe coronary disease;
- known moderate or mild aortic stenosis;
- severe or moderate systemic hypertension;
- severe or moderate pulmonary hypertension;
- severe impairment of ventricular function;
- known history of ventricular tachycardia;
- known history of supraventricular tachycardia;
- existing second- or third-degree atrioventricular block;
- hypertrophic cardiomyopathy;
- severe congestive cardiomyopathy;
- known hypokalaemia.

Situations where interpretation of the exercising record is difficult or impossible

Abnormalities of the resting ECG that preclude effective interpretation of the exercising record include:

- left bundle-branch block;
- ventricular pre-excitation;
- currently paced ventricular rhythm;
- widespread S–T,T changes;
- widespread QS complexes (especially across the precordial leads).

Procedures

Lead positioning

During exercise it is not possible to maintain adequate physical and electrical stability in relation to limb lead connections at their usual (for the standard 12-lead ECG) location. Instead, the 'limb' lead electrodes are positioned on the torso: with the right and left arm connections situated at the most lateral aspects of the respective infraclavicular fossa, and the right and left leg electrodes positioned halfway between the respective anterior iliac crest and the rib margin. This Mason–Likar modification of the standard 12-lead ECG results in a rightward shift of the axis, which is more marked in the standing than in the recumbent position. This rightward shift (typically giving an axis of +90° to +120°) sometimes results in the appearance of new q waves in aVL (but it should be noted that, whenever the mean frontal plane QRS axis is +90° or more positive, aVL becomes a 'cavity' lead and the finding of a q wave in a cavity lead is not abnormal).

Exercise protocols

Various exercise modalities can be used, including static or dynamic exercise, arm or leg exercise, and bicycle ergometry or treadmill procedures, but the commonest procedure by far is dynamic treadmill exercise. The most popular protocol is the Bruce protocol. This has a starting walking speed of 1.7 mph (1 km/h) at a 10 per cent slope, giving an oxygen consumption of about four metabolic equivalents, which in general use has proved very satisfactory. One major advantage of the Bruce protocol is that large diagnostic and prognostic databases exist for this test.

Exercise endpoints

Exercise is continued until one of the following endpoints is reached:

- subject wishes to stop (chest pain, dyspnoea, fatigue, leg weakness, light headedness, exhaustion, claudication);
- target endpoint is reached (target heart rate or exercise level);
- operator terminates the procedure:
  - early or severe (>2 mm) S–T depression,
  - S–T elevation,
  - ventricular tachycardia,
  - second- or third-degree heart block,
  - fall in heart rate (20 beats/min or more),
  - fall in blood pressure (20 mmHg or more),
  - perceived patient distress,
  - failure of monitoring equipment.

Assessment of the exercise electrocardiogram

As the heart rate increases with exercise, the PR, QRS, and QT intervals all reduce in normal subjects. The P-wave amplitude increases and the atrial repolarization wave (the Ta wave) increases in amplitude.

Atrial repolarization wave

Sinus tachycardia is associated with an increase in the depth and duration of the Ta wave. This gives a curved upsloping segment between the QRS complex and the T wave, often misconstrued as S–T-segment depression, and a common cause of an incorrect conclusion that an exercise test is positive. A Ta wave can be recognized when it is noted that back-extrapolation of a depressed S–T segment shows it to be continuous with downsloping depression in front of the QRS complex (Fig. 29).

Fig. 29 Negative exercise ECG. This ECG is the record taken at peak exercise in a maximal stress test. The heart rate is 180 beats/min. There is no S–T depression. A prominent atrial depolarization wave (Ta wave) is clearly seen in II, III, aVF, and V_4–V_6 (best in II). As would be expected aVR shows an 'upside down' Ta wave (since it is a cavity lead).

Standard criteria for a positive test

By definition, a positive test occurs when 1 mm (0.1 mV) of horizontal or downsloping S–T depression occurs during exercise (usually at peak exercise) or in the early recovery period. Upsloping S–T depression is less reliably predictive of the presence of coronary disease than flat or downsloping S–T depression. Greater (than 1 mm) degrees of S–T depression are more reliably predictive of coronary disease, as are S–T depression occurring earlier in the exercise period, more prolonged S–T depression, and a more widespread (within the ECG recording leads) S–T change. Figure 30 shows an example of significant (2 mm) S–T depression in the left precordial leads.

Fig. 30 Positive exercise ECG. This ECG is the record taken at peak exercise in a maximal stress test. The heart rate is only 136 beats/min. The test was stopped before the 'target' heart rate was reached because the patient had chest pain and the test was already clearly giving a positive result. There is 2 mm of S–T depression in the left precordial leads.

Sometimes the S–T depression is most marked or only occurs during the recovery period (Fig. 31).

Fig. 31 Positive exercise ECG. Record (a) was taken at peak exercise. It is not abnormal. It shows a prominent atrial repolarization wave. Record (b) was taken 6 min into the recovery period (after 9 min of exercise). Although no S–T depression occurred during exercise (a) there is clearly abnormal S–T depression during recovery. Coronary angiography confirmed the presence of significant disease in the anterior descending and diagonal branches of the left coronary artery.

An example of a negative stress test is shown in Fig. 29.

Interpretation of the test result

Positive or negative. Pre- and post-test probability. Bayesian analysis

The criterion for positivity of an exercise ECG is widely accepted as being 1 mm of flat or downsloping S–T segment depression during or early after exercise. The interpretation of a positive result is more problematical. Usually the question being asked is whether or not the test result indicates a high probability that the patient has coronary heart disease. Bayesian analysis of this problem indicates the enormous impact of the prevalence of coronary disease in the population group from which the subject is drawn (the prior probability of the condition) in answering this question. In essence, Bayes's theorem states the self-evident truth that interpretation of the future (probability of disease in the given subject) is helped by a knowledge of past experience (prevalence of the disease in the population from which the subject comes) as well as present observations (the test result).

Bayesian analysis expresses the probability that a subject with a positive exercise test result does actually have coronary heart disease, in terms of the sensitivity and specificity of the test and the prevalence of the disease, as follows:

$$\text{Probability} = [\text{prevalence} \times \text{sensitivity}]/[\text{prevalence} \times \text{sensitivity} + (1 - \text{prevalence})\ (1 - \text{specificity})\ ].$$

If one inserts reasonable (on the basis of published results of exercise testing) values for the sensitivity (say 0.8, i.e. 80 per cent) and specificity (say 0.9, i.e. 90 per cent) into this equation and then looks at the impact of variations in prevalence on the predictive value of a positive test, then the values shown in Table 2 are obtained. Clearly the false-positive rate is very high in low-prevalence populations (the healthy population) and this limits the value of exercise testing as a screening procedure in asymptomatic, presumptively healthy groups.

The likelihood that a subject with a positive stress-test result has coronary artery disease (the 'post-test or posterior probability') is therefore dependent on the prevalence of the disease in the population from which the subject is derived (the 'pretest or prior probability'). Equally, of course, the likelihood that a subject with a negative stress-test result does not have coronary artery disease (the 'post-test probability') is also dependent on the prevalence of the disease in the population from which the subject is derived (the 'pretest probability'). This concept is shown graphically in Fig. 32.

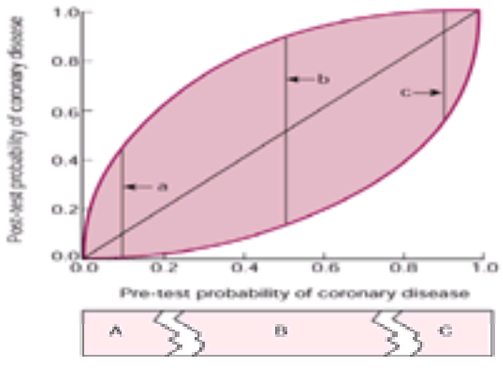

Fig. 32 Predictive value of positive and of negative exercise stress tests. The pretest probability of the condition (prevalence) is shown on the abscissa and the post-test probability (likelihood of the condition in the light of the test result and in the population from which the subject exercised was drawn) on the ordinate. The 45° line shows the impact of a completely non-predictive test result (for example, tossing a coin), the post-test probability being unchanged by the test result. Curves for clearly positive (= 2 mm flat or downsloping S–T depression) and clearly negative (no S–T depression at peak exercise with target heart rate achieved) test results are shown. The upper curve is for positive and the lower curve for negative results. The vertical distance between the curves shows the relative diagnostic 'benefit' (additional diagnostic probability) from the test result. This is greatest (b) where the pretest probability is intermediate (B). In a low-prevalence population (A), such as young, asymptomatic persons being screened, the pretest probability of the condition is (by definition) very low, and even the contribution of a strongly positive test result (upper curve) only results in a post-test probability of 40 per cent, i.e. the false-positive rate in this low-prevalence population is 60 per cent. In such a low-prevalence population a clearly negative result (lower curve), i.e. one concordant with the initial statistical probability, would be powerfully effective in confirming the initial likelihood. In this case a pretest probability of about 10 per cent would give way to a very low post-test probability of about 1 per cent. In a high-prevalence population (C) the pretest probability in this case is about 89 per cent and the post-test probability of a positive result would give about a 99 per cent probability of coronary artery disease. In such a population a clearly negative result would give about a 55 per cent probability of coronary disease, i.e. there would be a very

significant false-negative rate. In general, therefore, a positive exercise stress-test result in a high-prevalence population is likely to be a true-positive and a negative result in a low-prevalence population is likely to be a true-negative. Conversely, a positive result in a low-prevalence population and a negative result in a high-prevalence population are both significantly likely to be false results. The stress test will make its greatest diagnostic contribution (b) where the prevalence is intermediate (B), i.e. where the initial diagnostic position is unclear.

Degree of abnormality of the test result

The degree of abnormality of the stress-test result also has a powerful bearing on the predictive value of the result. Greater or lesser degrees of abnormality may be shown by:

1. the depth of the S–T depression;
2. the time of onset of the S–T depression;
3. the duration of the S–T depression;
4. the number of ECG leads showing significant S–T depression.

Only in respect of the depth of S–T depression, however, is there currently a large database of information. The effect of varying degrees of S–T depression on the predictive value of a positive test is shown in Fig. 33.

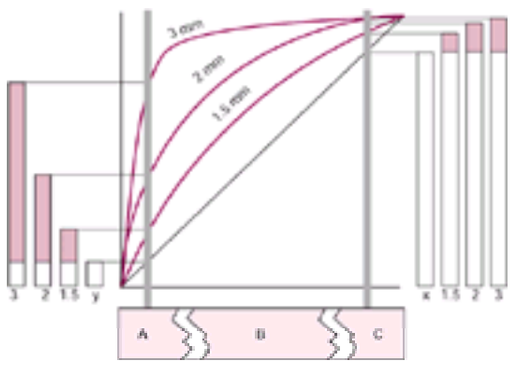

Fig. 33 Predictive value of different degrees of positivity of exercise stress tests. In a high-prevalence population (C) the pretest probability of coronary disease is high (×) and positive test results of increasing degree (1.5, 2, and even 3 mm S–T depression) only minimally increase the probability of the condition. In a low-prevalence population (A) a 1.5-mm and a 2-mm S–T depression result still leave less than a 50 per cent probability that the subject has significant ischaemic heart disease. In such a population only a really strikingly positive result (3-mm depression) would result in a significant (75 per cent) probability of the condition.

Confounding ECGs

Interpretation of the exercise ECG is dependent upon the assessment of the timing, duration, degree, and distribution of S–T depression occurring during exercise. When the pre-exercise ECG shows significant S–T-segment abnormalities (left bundle-branch block, ventricular pre-excitation, ventricular paced rhythm, non-specific S–T-segment depression, etc.), interpretation of changes in the S–T segments occurring during exercise is virtually impossible. In these situations the exercise stress ECG makes no useful contribution to the diagnosis of or to the exclusion of significant coronary artery disease.

Further reading

Bruce RA, *et al*. (1963). Exercise testing in adult normal subjects and cardiac patients. *Pediatrics* **32** (Suppl.) 742–56.

Casale PN, *et al*. (1987). Improved sex-specific criteria of left ventricular hypertrophy for clinical and computer interpretation of electrocardiograms: validation with autopsy findings. *Circulation* **75**, 565–72. [The Cornell gender-specific voltage criteria give the best correlation with left ventricular mass]

Chaitman BR (1997). Exercise stress testing. In: Braunwald E, ed. *Heart disease: a textbook of cardiovascular medicine*, pp 153–76. WB Saunders, Philadelphia. [The standard textbook of cardiovascular medicine]

Ellestad MH, *et al*. (1996). History of stress testing. In: Ellestad MH, ed. *Stress testing: principles and practice*, pp 1–9. FA Davies, Philadelphia. [An excellent reference textbook of exercise stress testing}

Fletcher GF, *et al*. (1995). Exercise standards. A statement for healthcare professionals from the American Heart Association. *Circulation* **91**, 580–615. [Standard protocols and procedures]

Gianrossi R, *et al*. (1989). Exercise-induced ST segment depression in the diagnosis of coronary artery disease: a meta-analysis. *Circulation* **80**, 87–98. [Comprehensive, authoritative meta-analysis]

Hamm LF, *et al*. (1989). Safety and characteristics of exercise testing early after acute myocardial infarction. *American Journal of Cardiology* B>63, 1193–7. [The largest report of the risks of exercise stress testing early after acute myocardial infarction]

Macfarlane PW, Lawrie TDV, eds (1989). *Comprehensive electrocardiography. Theory and practice in health and disease*. Pergamon Press, Oxford. [A three-volume encyclopaedia of electrocardiography.]

Rochmis P, Blackburn H (1971). Exercise test: a survey of procedures, safety and litigation experience in approximately 170,000 tests. *Journal of the American Medical Association* **217**, 1061–1066. [The first large survey of the risks of exercise stress testing]

Romhilt DW, *et al*. (1969). A critical appraisal of the electrocardiographic criteria for the diagnosis of left ventricular hypertrophy. *Circulation* **40**, 185–95.

Rowlands DJ (1978). The electrical axis. *British Journal of Hospital Medicine* **19**, 472–81. [A detailed description of the technique for estimating the clinical significance of the measurement]

Rowlands DJ (1991). *Clinical electrocardiography*. Gower, London. [A detailed explanation (extensively illustrated) of the basis of electrocardiography]

Schamroth L (1976). *An introduction to electrocardiography*. Blackwell Scientific, Oxford. [A simple, basic introduction to the ECG]

Sgarbossa EB, *et al*. (1996). Electrocardiographic diagnosis of evolving acute myocardial infarction in the presence of left bundle-branch block. *New England Journal of Medicine* **334**, 481–7.

Sokolow M, Lyon TP (1949). The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads. *American Heart Journal* **37**, 161–86.

Weinstein MC, Fineberg HV (1980). *Clinical decision analysis*. WB Saunders, Philadelphia.

15.3.3 Echocardiography

A. P. Banning

Introduction

Modern transthoracic echocardiography combines real-time two-dimensional imaging of the myocardium and valves with information about velocity and direction of blood flow obtained by Doppler and colour flow mapping. It is non-invasive and a complete examination can be performed in most patients in less than 30 min.

Doppler echocardiography has revolutionized the diagnosis and follow-up of patients with valvular heart disease. Serial cardiac catheterization to assess severity and progress of valvar stenosis has been almost completely superseded by echocardiography, and the role of invasive investigation is increasingly limited to assessment of the coronary arteries prior to corrective surgery.

Transoesophageal echocardiography is available in larger cardiac centres. Under sedation, an ultrasound probe is passed into the oesophagus to a position behind the heart producing excellent resolution of cardiac structures. It is used diagnostically in many emergency situations, including aortic dissection and suspected prosthetic mechanical valve dysfunction, and as an additional method of monitoring cardiac performance during cardiac and non-cardiac surgery.

The dramatic expansion in the availability of echocardiography has been accompanied by continuing technological development. Stress echocardiography can be used to detect occult coronary disease and predict cardiac risk, whilst the administration of contrast agents may allow visualization of myocardial perfusion. Although three-dimensional reconstruction is currently a research tool, real-time three-dimensional imaging is an increasingly realistic goal. In the future, these and other developments seem likely to ensure that echocardiography will maintain its central role in the diagnosis and management of most cardiac and many non-cardiac patients.

Principals of echocardiography

The transducer used for most echocardiographic examinations contains piezo-electric crystals that emit ultrasound frequencies of 2.5 to 5 MHz. Most of the sound energy is scattered or absorbed, but reflection occurs at interfaces between tissues of different acoustic impedance (e.g. between blood and muscle). The transducer collects these reflections and the time delay between emission and reception is calculated. This allows the depth of the reflection to be derived and its position to be displayed on a screen as a dot (pixel). The brightness of the dot is related to the magnitude of the reflected signal. In general, higher frequency transducers allow better discrimination between structures but more ready attenuation leads to reduced penetration.

There are three main echocardiographic techniques: cross sectional (two-dimensional), M-mode, and Doppler.

Cross sectional echocardiography (two-dimensional echo)

Cross sectional (two-dimensional) images are constructed as the ultrasound beam sweeps across the heart. Between 50 and 100 cross sections are presented each second and this gives the impression of a moving picture. These images are readily interpretable by an observer with a knowledge of cardiac anatomy and this technique is the cornerstone of modern echocardiography.

M-mode echocardiography

M-mode echocardiography preceded modern two-dimensional imaging. Unlike two-dimensional imaging, which uses a series of sweeps across the heart, M-mode uses a single static beam of very frequent ultrasound pulses. The narrow beam is analogous to a vertical mineshaft passing through various layers of rock. Displayed in real time this results in reflections from cardiac structures being displayed as horizontal lines with superficial structures at the top of the screen and the deeper structures at the bottom. These data are interpretable when one knows which structure each line represents and the technique has excellent spatial resolution. With the advent of two-dimensional and Doppler, M-mode is now principally used for measurement of cardiac chamber dimensions and observation of the relative movement of cardiac structures to each other, for example the relationship of the anterior leaflet of the mitral valve to the septum in hypertrophic cardiomyopathy.

Doppler echocardiography

The Doppler principal allows the velocity and direction of movement of an object (or moving blood in the case of cardiac ultrasound), to be calculated from the shift in the frequency of a reflected waveform relative to the observer. Cardiac imaging employs pulsed wave, continuous wave, and colour Doppler techniques.

Pulsed wave Doppler allows information about flow to be obtained from a particular point within the heart. The range of detectable velocities is limited, and it is used for sampling normal and low velocities, for example mitral valve flow.

Continuous wave Doppler measures the peak velocity encountered along the ultrasound beam and is particularly valuable for measuring high velocity jets, for example aortic stenosis. It is important to remember that failure to align the transducer exactly parallel to flow results in measurement of artefactualy low velocities and an underestimation of the valvular stenosis.

Colour Doppler allows a dynamic representation of the direction and velocity of flow to be superimposed onto a two-dimensional image of the heart. Velocities towards the transducer are coded in red and velocities away are coded in blue. Turbulent flow produces variable velocities and results in a mosaic pattern that is ideal for characterization of regurgitant lesions. This technique is now so sensitive that it can detect trivial regurgitation during the closure of many normal heart valves.

Imaging

Imaging is usually performed with the patient lying on their left hip in the left lateral position, with their left arm behind their head (Fig. 1). Ultrasound cannot travel through bone and thus cardiac imaging is performed in intercostal spaces to the left of the sternum and at the apex of the heart in the axillary line. These 'echo

windows' provide standard views described as the parasternal short and long axis and apical two, four, and five chamber. Useful additional views can be obtained from the subxiphoid and suprasternal approach in some patients.

Fig. 1 Patient in the standard position undergoing transthoracic echocardiography.

A standard echo examination involves two-dimensional imaging from the parasternal approach followed by M-mode measurements and colour Doppler. Apical two-dimensional views are followed by colour and continuous and pulsed Doppler interrogation.

Applications of echocardiography

Valvular heart disease

Transthoracic Doppler echocardiography is the investigation of choice for patients with suspected valvular heart disease. All four cardiac valves can be visualized and interrogated by Doppler. Concomitant abnormalities in ventricular performance can be assessed simultaneously.

Aortic stenosis

Two-dimensional echocardiography can usually image the aortic valve cusps and if they are thin and freely mobile it is unlikely that there is significant aortic stenosis. However, if the valve cusps are thickened and calcified, interrogation by continuous wave Doppler is mandatory. The severity of aortic stenosis is expressed as the peak pressure difference (or gradient) across the valve, and is calculated from the maximum flow velocity (V) using the modified Bernoulli equation (pressure gradient=$4V^2$). The gradient measured by continuous wave Doppler is not directly comparable with a gradient measured at cardiac catheterization, which can lead to confusion. Doppler measures the peak instantaneous gradient and is higher than peak-to-peak gradient measured by catheterization (Fig. 2). In patients with normal left ventricular systolic function a peak-to-peak catheter gradient of 50 mmHg suggests significant aortic stenosis (Fig. 3 and Plate 1): this corresponds to a peak instantaneous gradient measured by Doppler of about 70 to 80 mmHg.

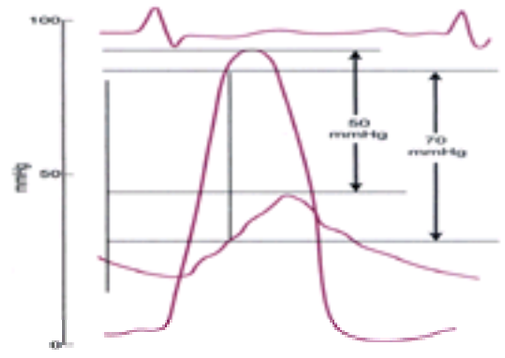

Fig. 2 Representative pressure traces of left ventricular pressure (lower line) and aortic pressure (upper line). The measured gradient at cardiac catheterization is the difference between the two peak pressures (peak-to-peak) and in this case it is 50 mmHg. Continuous wave Doppler measures the maximum difference between the pressures or the peak instantaneous gradient which in this case is 70 mmHg.

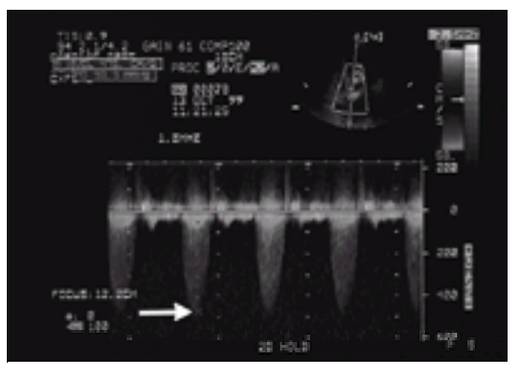

Fig. 3 Apical continuous wave Doppler across the aortic valve in a patient with severe aortic stenosis. The peak velocity (arrow) is greater than 4.5 m/s consistent with a peak instantaneous gradiant across the aortic valve of 90 mmHg. (See also Plate 1.)

When chronic critical outflow obstruction results in declining left ventricular function and reduced cardiac output, the gradient produced by any degree of valve obstruction also falls. Doubt about the severity of the stenosis can usually be resolved by calculating the valve area using the continuity equation which uses data from Doppler and two-dimensional echo. In experienced hands this provides valuable additional information, but accurate measurement of the left ventricular outflow tract diameter can be difficult and if the findings are not consistent with other data, the investigation should be either be repeated or the patient should be referred for cardiac catheterization.

Aortic regurgitation

Assessment of the mechanism and severity of aortic regurgitation requires a combination of all three echo modalities. M-mode may demonstrate fluttering of the anterior leaflet of the mitral valve and in the setting of acute severe aortic regurgitation, premature closure of the mitral valve. Two-dimensional will occasionally demonstrate prolapse of one more of the aortic cusps, but even severe aortic regurgitation can occur through an aortic valve that appears to be structurally normal.

The severity of aortic regurgitation can be estimated using continuous wave and colour Doppler, although assessment can be difficult as it is influenced by left ventricular function. Doppler-derived pressure half-time and measurement of regurgitant fraction and/or flow convergence zone are valuable when there is uncertainty

over lesion severity. M-mode and colour Doppler can be combined and when the regurgitant jet fills more than 50 per cent of the left ventricular outflow tract the regurgitation is classified as severe.

In patients with severe asymptomatic aortic regurgitation, a serial increase in left ventricular dimensions or a progressive fall in ejection fraction are indications for surgery. However, any increase in ventricular dimension should be at least 0.5 cm before it is regarded as significant given the limited reproducibility of echocardiographic parameters.

Mitral stenosis

Mitral valve stenosis is well visualized using either M-mode or cross-sectional echocardiography. Its severity can be determined by estimating the area of the valve orifice either by direct planimetry of the two-dimensional short axis image or from the Doppler pressure half-time (MVA=220/Pt $_{1/2}$). A valve area of less than 1.0 cm^2 indicates severe mitral stenosis. Transthoracic echocardiography is also used to assess the suitability of the mitral valve for balloon dilation, although transoesophageal imaging is necessary to exclude left atrial thrombus.

Mitral regurgitation

Transthoracic echocardiography will usually demonstrate the mechanism and severity of mitral regurgitation. Two-dimensional imaging identifies abnormalities of the valve leaflets and colour flow shows jet direction and area. Severe mitral regurgitation is suggested by increased left ventricular end-diastolic dimension and hyperdynamic wall motion due to volume overload. Precise quantification of the amount of regurgitation is demanding as it is influenced by left ventricular function, the direction of the jet, and left atrial size. Various algorithms have been devised to improve quantitation of mitral regurgitation, including measurement of the flow convergence zone and the PISA method, but most centres simply classify the extent of regurgitation as mild, moderate, or severe.

Pulmonary and tricuspid valve disease

In adults, two-dimensional imaging of the pulmonary valve may be difficult, particularly if there is lung disease. Despite this, accurate Doppler information is usually obtainable. Tricuspid stenosis is very uncommon but some degree of tricuspid regurgitation is detectable even in healthy individuals. Measurement of the peak velocity of tricuspid regurgitation (V) is valuable as in the absence of pulmonary valve disease it can be used to estimate pulmonary artery systolic pressure:

$$\text{PA systolic pressure (mmHg)} = 4V^{2+}\text{right atrial pressure (usually assumed to be 5–10 mmHg)}$$

Prosthetic valves

Transthoracic echo is commonly performed as part of the routine follow-up of prosthetic valves. It is usually able to assess biological valves accurately, but for mechanical mitral valve prostheses in particular, attenuation artefact produced by the metal may be problematic. Transoesophageal imaging is recommended when transthoracic imaging is suboptimal or if improved resolution is required, for example suspected prosthetic valve endocarditis.

Abnormal left ventricular function

In most patients a full transthoracic echo study will confirm or refute a clinical suspicion of left ventricular dysfunction and identify the likely aetiology of any abnormality. Systolic and diastolic left ventricular function can be assessed and a variety of methods can be used to derive an estimate of left ventricular ejection fraction. In patients with ischaemic heart disease, assessment of regional wall motion is valuable and may occasionally demonstrate evidence of aneurysm formation. Left ventricular hypertrophy is detected by echocardiography and a measurement of left ventricular mass can be derived.

Echocardiography is a pivotal investigation in suspected heart failure, particularly as an ejection fraction of less than 40 per cent has been used as an inclusion criteria for most therapeutic trials in this condition. Community studies have demonstrated that many patients treated with diuretics for mild heart failure (usually ankle swelling) do not have left ventricular systolic impairment, and screening with echocardiography was initially recommended before initiation of treatment with angiotensin-converting enzyme inhibitors. However, subsequent studies have shown that impaired systolic function is very unlikely when the ECG and clinical examination are normal, and conversely that angiotensin-converting enzyme inhibitors can be initiated without waiting for echocardiography in patients with evidence of ischaemic heart disease and radiographic evidence of pulmonary oedema, as the ejection fraction is almost invariably less than 40 per cent under these circumstances.

Transthoracic echo may detect mural thrombus, particularly in patients with impaired systolic ventricular function. However, differentiation between thrombus and myocardium can be problematic at the left ventricular apex and tissue harmonic imaging and/or contrast agents may be necessary.

Minor concentric left ventricular hypertrophy is common in patients with hypertension. In hypertrophic cardiomyopathy, two-dimensional imaging may demonstrate asymmetrical septal hypertrophy with disproportionate thickening of the interventricular septum compared with the left ventricular free wall, or dramatic concentric hypertrophy with left ventricular cavity obliteration. Other characteristic features of hypertrophic cardiomyopathy include systolic anterior motion of the mitral valve (Fig. 4 and Fig. 5 and Plate 2) and partial mid-systolic closure of the aortic valve, which usually correlates with the presence of outflow tract obstruction. In the absence of conditions that may induce ventricular hypertrophy, for example aortic stenosis, these findings are diagnostic of hypertrophic cardiomyopathy. Colour Doppler can demonstrate turbulence in the outflow tract and continuous wave Doppler may detect characteristic 'dynamic' gradients that increase in severity as systole progresses. Other associated echocardiographic abnormalities in hypertrophic cardiomyopathy include mitral regurgitation and severe diastolic dysfunction.

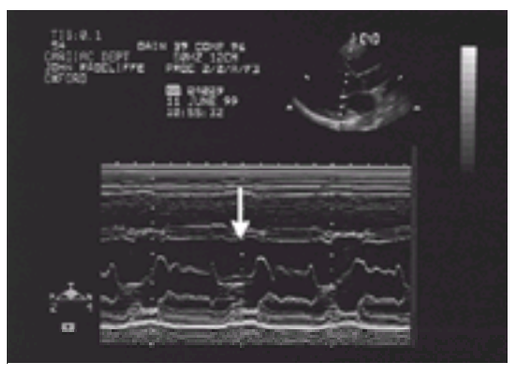

Fig. 4 M-mode echocardiogram through the mitral valve in a normal patient. Opening of the leaflets during ventricular diastole and closing during systole (arrow) can be observed. Contrast the behaviour of the mitral valve with that in Fig. 5. (See also Plate 2.)

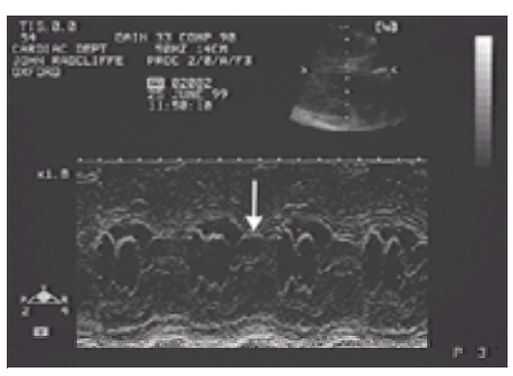

Fig. 5 M-mode echocardiogram through the mitral valve in a patient with obstructive hypertrophic cardiomyopathy. Closure of the valve during systole is accompanied by marked movement of the valve apparatus towards the left ventricular septum (arrow)—this is referred to as systolic anterior motion (SAM) of the mitral valve.

Atrial fibrillation

Echocardiography readily excludes a structural cause for atrial fibrillation (e.g. mitral stenosis) and facilitates thromboembolic risk stratification. It also allows measurement of left atrial dimension, which is valuable as cardioversion is less likely to be successful when this is large.

Following an embolic event/stroke

Echocardiography is the investigation of choice when a cardiac source of an embolus is suspected. It is mandatory in all patients presenting with embolic occlusion of a peripheral artery, or thromboembolic episodes in more than one vascular territory. Echocardiography should not, however, be performed in circumstances when the result is unlikely to influence patient management, but in patients with ischaemic stroke and a low likelihood of atheromatous arterial disease an echo can be considered as occasionally it will detect occult abnormalities such as atrial myxoma or cardiac thrombus.

Pericardial disease

Echocardiography readily diagnoses the presence of pericardial fluid and is useful when percutaneous drainage is attempted. Echocardiographic signs of pericardial tamponade include exaggerated respiratory variation in the mitral valve Doppler, presystolic closure of the aortic valve, and (particularly) right atrial and right ventricular diastolic collapse. Constrictive pericarditis is a difficult diagnosis to make using standard echocardiographic techniques and patients complaining of episodic breathlessness/fluid retention with characteristic abnormalities of the venous pressure require particularly careful interrogation by Doppler.

Pulmonary embolism

Echo can be useful in patients with pulmonary embolism as it can demonstrate right ventricular dilation and/or impaired right ventricular systolic function. Tricuspid regurgitant velocity can be used to estimate pulmonary artery systolic pressure, although it is unusual for this to be more than 70 mmHg acutely. Exceptionally, two-dimensional imaging may show thrombus within the right heart and/or the proximal pulmonary arteries. Although echocardiography is useful diagnostically when it demonstrates features consistent with pulmonary embolism, it cannot exclude the diagnosis.

Infective endocarditis

Echocardiography cannot be used to exclude endocarditis but is valuable when endocarditis is suspected clinically but there is insufficient data to make a formal diagnosis. Under these circumstances a typical vegetation detected by an experienced observer is regarded as a major criterion in the Duke diagnostic classification, and this may facilitate appropriate management. Transoesophageal echo should be performed when there is a suspicion of aortic root abscess, if prosthetic endocarditis is suspected, or occasionally in cases where there is persistent diagnostic doubt and the additional sensitivity and spatial resolution might be valuable.

Congenital heart disease

Echocardiography is the diagnostic modality of choice for patents with suspected congenital heart disease. Detailed transthoracic cardiac imaging is possible in co-operative babies and children but occasionally sedation or a short anaesthetic may be required. Rates of cardiac catheterization have been reduced by miniaturization of transoesophageal probes that facilitate diagnosis and follow-up of complex congenital heart disease. Fetal echocardiography is performed when surveillance obstetric ultrasound is abnormal or in cases where previous history suggests a possible cardiac problem.

Transoesophageal echocardiography

Transoesophageal echocardiography is available in cardiac centres and some smaller hospitals. The ultrasound probe is similar to the endoscope used for upper gastrointestinal investigation, except that there are no optical fibres. With the patient under sedation, the probe is manipulated into the oesophagus where its position behind the heart produces excellent resolution, particularly of posterior cardiac structures.

Transoesophageal echocardiography is an invasive procedure and the patient's written consent is required. After fasting for a minimum of 4 h, local anaesthetic spray (10 per cent lidocaine) is applied to the upper pharynx and the patient is usually sedated with a short acting benzodiazopine (e.g. midazolam). Blood pressure and oxygen saturation are monitored throughout and both resuscitation equipment and the benzodiazepine antagonist flumazanil should be readily available (Fig. 6).

Fig. 6 Patient in the standard position undergoing a transoesophageal echocardiogram. Non-invasive monitoring of pulse rate, oxygen saturation, and blood pressure is mandatory when intravenous sedation is being administered.

Even though transoesophageal echo is commonly performed in high-risk haemodynamically unstable patients, the rate of serious complications (aspiration and oesophageal rupture/tears) is less than 1 per cent. Absolute contraindications to transoesophageal echo include oesophageal tumours, strictures, varices, and diverticulae.

Who should have a transoesophageal echocardiogram?

The principal indications for transoesophageal echocardiography are listed in Table 1. Its principal advantages over transthoracic imaging are improved spatial resolution and the ability to image posterior structures such as the left atrium and descending aorta. It is valuable in a number of emergency situations including suspected aortic dissection, prosthetic mechanical valve failure, and possible endocarditis. It may be used to image patients in whom data from transthoracic imaging is unsatisfactory because of obesity, lung disease, or chest deformity. Other indications include screening for left atrial thrombus before cardioversion of atrial fibrillation and monitoring cardiac performance during cardiac and some non-cardiac surgery.

Valve disease

Patients with mitral stenosis are at particular risk of thromboembolism and transthoracic echo has limited sensitivity for the detection of left atrial thrombus (Fig. 7).

Transoesophageal echocardiography is recommended in those with mitral stenosis ifembolic events occur despite therapeutic anticoagulation and may demonstrate spontaneous echo contrast (smoke-like echoes produced by the interaction of erythrocytes and plasma proteins under conditions of stasis). This is an independent predictor of left atrial thrombus and/or cardiac thromboembolic events. Transoesophageal echo is also used to assess anatomy and exclude left atrial thrombus before balloon valvuloplasty in patients with mitral stenosis and to assess anatomy, severity, and suitability for surgical repair in patients with mitral regurgitation. In patients with mitral prostheses, reverberation artefact overlying the left atrium limits the ability of transthoracic imaging to detect paraprosthetic regurgitation. Transoesophageal imaging provides excellent visualization of the left atrium and is particularly recommended under these circumstances.

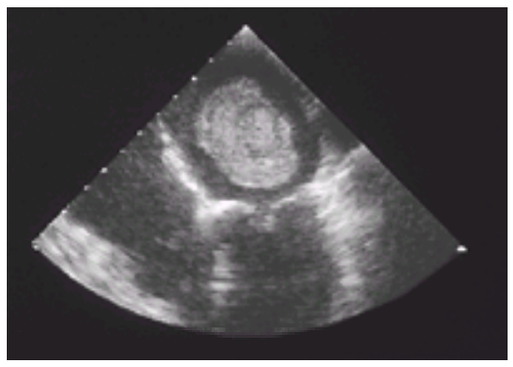

Fig. 7 Two-dimensional transoesophageal echocardiogram of the left atrium demonstrating a large mobile thrombus in a patient with mitral stenosis.

Endocarditis

Characteristic vegetations or evidence of abscess formation identified by echocardiography are increasingly used as diagnostic criteria in patients with possible endocarditis. The excellent spatial resolution (less than 1 mm) of transoesophageal echo makes it superior to transthoracic imaging for the detection of vegetations and its sensitivity may exceed 90 per cent. Transoesophageal echo should be considered when there is a high clinical suspicion of endocarditis but blood cultures are sterile and transthoracic imaging is not diagnostic, or under circumstances when the sensitivity of transthoracic imaging is particularly poor, for example prosthetic valves or calcific valvular disease. Transoesophageal echo is also recommended if there is a possibility of aortic root abscess formation as this complication is not easily identified using transthoracic imaging and surgery is usually necessary.

Aortic disease

Transthoracic imaging of the aorta is limited to the proximal aortic root and the arch in most patients. Using transoesophageal imaging most of the ascending and all of the descending thoracic aorta can be visualized and image quality is improved. This is particularly useful in patients with suspected acute aortic dissection and in many cases it is the only imaging necessary before emergency surgery. Large, mobile or pedunculated aortic atheroma in the descending aorta may be detected by transoesophageal echocardiography and several studies have suggested an association with ischaemic stroke.Transoesophageal imaging of the aorta has also been recommended in suspected cases of cholesterol embolization and to assess thromboembolic risk prior to cardiac intervention or surgery.

Thromboembolism

In patients with thromboembolism, there has been extensive debate over the value of imaging with transoesophageal echocardiography. Clinical examination, electrocardiography, and transthoracic echocardiography provide sufficient information to determine optimal management in the majority. However, transoesophageal echocardiography is indicated when embolic events occur in anticoagulated patients with native or prosthetic valvular heart disease, especially if endocarditis is suspected, or when transthoracic images are inconclusive.

In patients with unexplained or cryptogenic ischaemic stroke, wider use of transoesophageal echo has been advocated. Transthoracic echo and exclusion of alternative pathologies such as thrombophilia and carotid stenoses should precede the transoesophageal examination as under these circumstances minor cardiac structural abnormalities are more likely to be clinically relevant.

The transoesophageal approach is superior to transthoracic echocardiography for imaging the interatrial septum for atrial septal aneurysm (a redundant bulge in the area of fossa ovale, with respiratory movement greater than 10 mm) and assessing patency of the foramen ovale (Fig. 8). The clinical relevance of these atrial septal abnormalities can be questionable as the relationship to the thromboembolic event is commonly speculative. Currently, anticoagulation is the usual management following a otherwise unexplained single embolic event, but occasionally percutaneous or surgical correction of the defect is recommended.

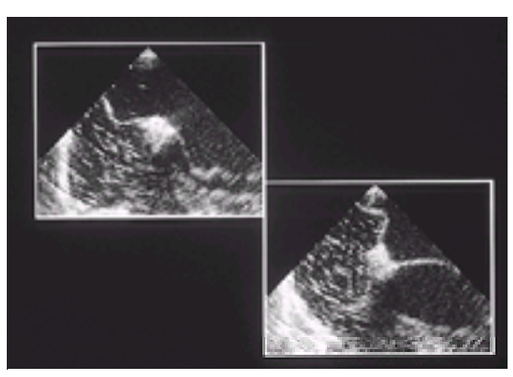

Fig. 8 Two-dimensional transoesophageal echocardiogram of the atria. In the upper panel contrast is principally contained within the right atrium although two bubbles can be seen in the left atrium. There is marked excursion of the atrial septum towards the left atrium during expiration (lower panal). These appearances are consistent a diagnosis of an atrial septal aneurysm with patent foramen ovale.

Stress echocardiography

Diagnosis of reversible ischaemic myocardial dysfunction is now possible using echocardiography. Imaging can be performed either during or immediately after exercise, but more commonly an intravenous infusion of dobutamine is used to mimic the cardiac response to exercise. Development of reversible systolic regional wall motion abnormalities suggests coronary artery disease and stress echo is used increasingly as a diagnostic test in patients with chest pain. Stress echo also has an increasing role in risk stratification prior to general surgical procedures and in assessing myocardial viability prior to revascularization.

Further reading

Cheitlin MD *et al*. (1997). ACC/AHA guidelines for the clinical application of echocardiography: executive summary. *Journal of the American College of Cardiology* **29**, 862–79.

Flachskampf FA, Decoodt P, FraserAG, Daniel WG, Roelendt JRTC (2001). Recommendations for performing transesophageal echocardiography. *European Journal of Echocardiography* **2**, 8–21.

Kerut EK, McIlwain EF, Plotnick GD. *Handbook of echo-Doppler interpretation*. Futura, New York. [A good integration of clinical and more technical echocardiography.]

Oh JK, Seward JB, Tajik AJ (1999). *The echo manual*, 2nd edn. Lipincott-Raven. [A well presented practical guide.]

Rimington H, Chambers J (1998). *Echocardiography: a practical guide for reporting*. Parthenon, London. [A short guide to reporting echocardiograms including normal ranges and how to interpret data.]

15.3.4 Nuclear techniques

H. J. Testa and D. J. Rowlands

Nuclear imaging of the heart plays an important role in the investigation of patients with heart disease, giving valuable information for the practical management of patients and contributing to the understanding of the physiology of myocardial perfusion. The techniques may be used to:

1. localize areas of ischaemia induced by exercise or by drugs;
2. demonstrate the extent and distribution of viable myocardium;
3. provide an assessment of global and of regional left ventricular function; and
4. demonstrate recent myocardial cell damage.

The investigations are scarcely invasive and of little discomfort or inconvenience to the patient.

In recent years, the use of nuclear imaging tests has increased dramatically. This increase is largely due to two factors: (i) development of new radiopharmaceuticals labelled with technetium to aid in the imaging process, and (ii) recognition of the value of pharmaceuticals for the induction of cardiovascular stress, permitting stress studies to be undertaken without exercise, which is particularly useful when exercise is difficult or contraindicated. There has also been considerable improvement in the quality of the images obtained thanks to substantial improvement in gamma-camera technology and to the widespread use of single photon emission computer tomography (SPECT). Recent estimates suggest that, in the United States, the annual number of procedures undertaken has doubled from 2.9 million in 1990 to 5.8 million in 1997. In our department, the number of cardiac studies has increased from 300 per year in 1995 to more than 1000 in 1999; the main area of increase being in the use of myocardial perfusion studies to investigate myocardial ischaemia.

Myocardial perfusion imaging in the recognition and assessment of myocardial ischaemia

Myocardial perfusion imaging is undertaken using an injection of a radiopharmaceutical that is taken up by the myocardium in proportion to the myocardial blood flow at the time of the injection and which remains in the myocardial cells during the period of the imaging. Studies are performed both at rest and under stress, in order to evaluate differing regional perfusion in these two states. The three compounds mainly used in current clinical practice are thallium ^{201}, technetium-99m-Cardiolite (sestamibi-MIBI), and technetium-99m-Myoview (tetrofosmin).

Thallium^{201} is an analogue of the potassium ion. It is taken up only by myocardial cells that are both viable and adequately perfused. A dose of 75 MBq, (approximately 2 mCi) is injected intravenously (i) at maximal chosen or achievable exercise, or (ii) after pharmacological stress, or (iii) in association with a combination of these procedures (exercise plus pharmacological stress). Pharmacological stress can be induced with powerful vasodilators (dipyridamole or adenosine) or with inotropic and chronotropic stimulators (dobutamine). The initial myocardial distribution of thallium ^{201} reflects regional blood flow. Redistribution of the isotope to all viable myocardial cells occurs approximately 3 to 4 h later. Images are taken immediately after exercise (to reflect blood flow) and 3 to 4 h later to reflect viable myocardium.

The patient is positioned supine on the SPECT couch with the left arm placed over the head to avoid attenuation artefacts. The gamma camera rotates around the patient through 180° from the right anterior oblique 30° to the left posterior oblique 30°, at 6° increments of 30 s duration. The orbit of rotation may be circular or non-circular; the latter is preferred. The acquired data is computer analysed and slice images are constructed using the techniques of filtering back projection at approximately 1-cm intervals. Short axis (SA), vertical long axis (VLA), and horizontal long axis (HLA) slices are reconstructed for clinical evaluation. An example of images obtained in this way is shown in Fig. 1.

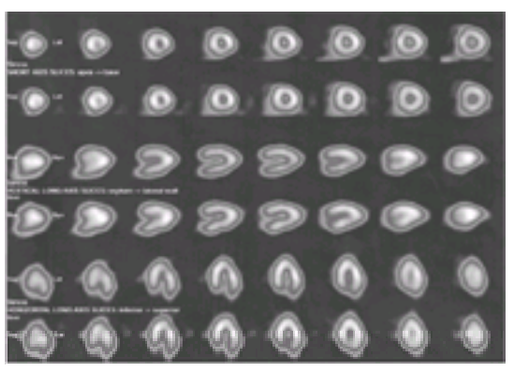

Fig. 1 Normal stress and rest study with thallium-201.

Any localized defects occurring in association with stress are indicative of non-viability (infarction) or of underperfusion (ischaemia). Defects of thallium uptake occurring at rest or after redistribution during a period of rest following stress (exercise or pharmacological) correlate well with areas of infarction. A defect in tracer uptake immediately post-exercise or stress that shows redistribution after rest is indicative of myocardial ischaemia. It is now known that redistribution can be very slow in hypoperfused but viable myocardium, and a further injection of thallium may help to improve the differentiation between reversible and fixed perfusion defects. Furthermore, the comparison of images taken immediately after injection at rest with those taken at rest 3 to 4 h later (rest-redistribution protocol) can be used in viability assessment and investigation of hibernating myocardium.

The drawback of thallium is its long physical half-life of 72 h (which limits the dose that can be used so as to minimize radiation exposure to the patient) and the low photon energy (60–90 keV) which cause significant attenuation and degradation of the images.

As a consequence of the physical limitations of thallium ^{201}, technetium-labelled radiopharmaceuticals have been developed. Technetium ^{99m} has a higher photon energy (140 keV) than thallium^{201} and a shorter physical half life of 6 h. Two compounds are used in routine practice—technetium-99m-MIBI and technetium-99m-Myoview. Both are lipophilic and, like thallium, are taken up by the myocardium in proportion to myocardial blood flow. The maximum dose injected to the patient for the study (stress/rest) is of 1000 MBq (approximately 25 mCi).

Technetium-99m-MIBI is an isonitrile complex which after intravenous injection diffuses from the blood into cardiac myocytes and it is retained by the mitochondria within the cell. Its clearance from the blood is rapid, but only 40 to 60 per cent is extracted on the first pass through the myocardium. However, at about 20 min the percentage uptake by the myocardium is similar to thallium. This compound does not have a significant clinical redistribution (less than 15 per cent from its initial uptake) and the concentration in the myocardium remains constant over a period of 4 or 5 h. Images can be taken several hours after injection, when they still represent the distribution of coronary blood flow in the myocardium at the time of injection. Because of this lack of redistribution it is necessary to inject the patient

twice, at maximum stress/exercise and at rest.

Technetium-99m-Myoview is also a lipophilic compound which, after intravenous injection, is rapidly cleared from the blood, taken into myocytes and retained in the mitochondria. Myocardial uptake is of the order of 1.2 per cent of the injected dose and reaches this level at about 5 min. It shows little redistribution and, as with technetium-99m-MIBI, two injections are necessary for a comparison of rest and stress/exercise uptake. Its advantage is that it has less liver uptake than technetium-99m-MIBI with resulting improvement in the interpretation of images in the inferior wall.

Several imaging protocols have been developed for the technetium compounds including rest and stress injections either on 1 or 2 days. Ideally the two studies should be carried out on separate days, but if they are performed on the same day at least a three times larger dose is given with the second injection in order to swamp activity from the first. It is also possible to use thallium ^{201} to obtain resting images and technetium compounds for stress studies. This later protocol permits stress and rest studies to be completed in about 2 h.

Interpretation of the images is, in practical terms, similar to that of thallium: any localized defects occurring in association with stress or exercise indicate either infarction or ischaemia. Defects of uptake occurring on the rest study are indicative of areas of infarction. A defect in tracer uptake immediately postexercise or stress that improves during the rest study is indicative of myocardial ischaemia. Figure 2 shows an example of a patient with myocardial ischaemia mainly affecting the apex.

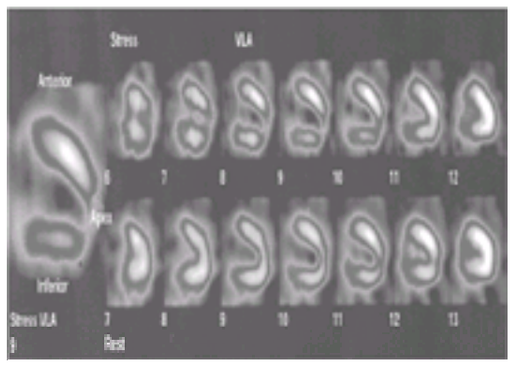

Fig. 2 Myocardial study carried out with Myoview showing, in the vertical long axis (VLA) view, an area of decreased tracer uptake at the apex on the images obtained after exercise (stress—row of seven images at the top of the figure) and improvement in this area at rest (row of seven images at the bottom of the figure). Views in other axes would be obtained simultaneously but are not shown.

The main clinical indications for myocardial perfusion scintigraphy, in patients without a clear diagnosis of coronary artery disease, include:

1. the assessment of patients with acute chest pain and non-diagnostic electrocardiograms;
2. use as a screening test in patients with
   a. familial hyperlipidaemia,
   b. a family history of coronary artery disease, and
   c. a perceived cardiac ischaemic risk in relation to proposed non-cardiac surgery.

In patients with known coronary artery disease the main indications include:

1. evaluating the functional significance of coronary stenoses detected by angiography;
2. assessment of the most significant functional stenoses in patients with multiple coronary lesions;
3. evaluation of postinfarction patients to establish the size of the infarct and the presence or absence of ischaemia in other areas of the myocardium.

It is also of value in the investigation of restenosis after revascularization with angioplasty or coronary artery bypass grafting.

The procedure is useful in relation to exercise stress testing (both for patients with known coronary disease and in those with no definite evidence of ischaemic disease) where the ECG in not able to yield useful information on stress-induced ischaemia (pre-existing left bundle branch block, ventricular pre-excitation, functioning ventricular pacemaker, or initial widespread abnormality).

SPECT perfusion studies for the diagnosis of coronary artery disease have a higher sensitivity and specificity than electrocardiography, particularly when state of the art systems are used (average sensitivity of about 90 per cent and specificity 80 per cent). They also provide better localization with respect to the vascular beds of the three coronary arteries. Sensitivity is higher for lesions occurring in the left anterior descending artery and lower for the right coronary artery. Sensitivity and specificity appear to be lower for female than for male populations. Stress tests or exercise tests have similar diagnostic performances, but exercise testing in patients who do not reach the target heart rate (220–age) has decreased sensitivity.

It is important to recognize that patients with left bundle branch block may have perfusion abnormalities in the septum that are not due to coronary artery disease; in these patients pharmacological stress studies improve the sensitivity of the test. Reversible perfusion defects have also been reported in patients with hypertrophic cardiomyopathy, aortic stenosis, and left ventricular hypertrophy.

Scintigraphic determination of ventricular function

The assessment of left ventricular function is one of the most important aspects of the evaluation of the cardiac status. Scintigraphy can provide the following information:

1. estimates of ventricular ejection fraction (the proportion of the ventricular end-diastolic volume ejected per beat, i.e. the stoke volume expressed as a percentage of the end diastolic volume), which is an overall measure of left ventricular performance;
2. estimates of regional ventricular performance by observation of the movements of the margins of the ventricle.

Two approaches can be taken: the 'first pass technique' and the 'gated equilibrium' method. Both involve the intravenous injection of technetium ^{99m}, which for first pass studies may be used in its ionic form as technetium-99m-pertechnetate, and for gated equilibrium studies by labelling the patient's own red blood cells. Both techniques (first pass and gated equilibrium) can be performed with a single injection of radioactive tracer.

In the first pass technique the first circulation of the radioactive bolus through the heart is studied and sequential images show the passage of the tracer through the right and left ventricles, separated in time. The single most useful view is probably the right anterior oblique (the projection of choice in single plane contrast radiography of the left ventricle). Images are taken at 1-s intervals and stored in the computer. A region of interest for the left ventricle is outlined and a high frequency activity–time curve for that region is plotted. Each point on the curve represents accumulated counts for a period of 0.04 s. The amount of radioactivity in the heart is proportional to the volume of blood in the cardiac cavities. Thus the change in the precordial count rates reflects the cyclical volume changes in the heart. A second region of interest is taken (usually as a horse-shoe-shaped region surrounded the left ventricular region of interest) to sample background activity variation with time. The background curve is 'normalized' to the left ventricular curve and then subtracted point-for-point from the high frequency left ventricular activity–time curve to give the corrected high frequency left ventricular activity–time curve (Fig. 3). It is usual to take two to six cycles around the peak of the left ventricular curve and then average the calculated ejection fraction for these cycles.

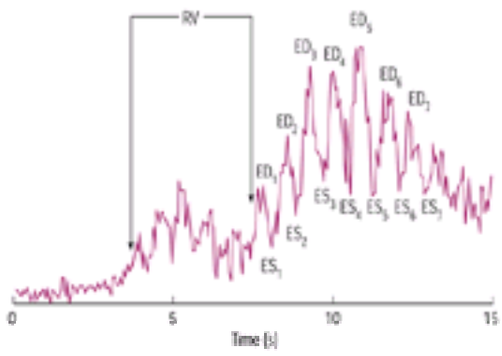

Fig. 3 High frequency activity—time curve recorded from the region of the left ventricle during the first passage of radioactive tracer through the central circulation. The time, in seconds, after the injection is shown on the abscissa. The ordinate displays the scintillation counting rate on a linear scale, after background correction. The early hump shows increased activity during the passage of the tracer through the right ventricle (RV). The later, larger hump shows the count rate during the passage through the left ventricle. Peaks and troughs are visible in relation to each cardiac cycle. Estimates of ejection fraction can be made for each cardiac cycle: (ED_1–ES_1/ED_1, ED_2–ES_2/ED_2, etc).

The gated equilibrium technique, referred to as gated cardiac blood pool imaging or as multigated acquisition (MUGA) imaging, depends upon complete mixing of the marker throughout the circulating blood volume and it therefore requires a marker that remains intravascular. The marker of choice is technetium-99m-labelled red blood cells. The cells do not have to be removed from the patient to be labelled: the ' *in vivo* labelling' technique may be used, which involves predisposing the patient's red cells to accept the technetium label by the administration, 30 min prior to technetium, of non-active stannous pyrophosphate. Subsequently, 600 to 800 MBq (approximately 15 to 20 mCi) of technetium-99m-pertechnetate are injected intravenously and imaging is begun after 10 min or so. As there is complete mixing of the marker throughout the blood volume, all four cardiac chambers are seen simultaneously and various degrees of superimposition of the chambers inevitable. Proper alignment of the gamma camera is therefore crucial for the optimal separation of the cardiac chambers. In general, maximal separation of the right and left ventricles is achieved in the left anterior oblique view, to which a caudal tilt of 15° may be added.

For the determination of ejection fraction, a region of interest over the left ventricle and a second region of interest for background correction are assigned in the same manner as for the first pass technique. A background-corrected activity–time curve of the left ventricular area is obtained: Fig. 4 (normal), Fig. 5 (abnormal). With the equilibrium technique this is usually displayed as a single cycle representative activity–time curve being produced as a composite of many (typically hundreds) consecutive cycles, synchronization of the cycles being achieved by means of the R wave of the electrocardiogram.

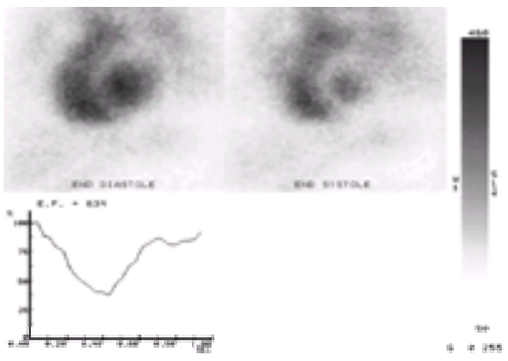

Fig. 4 Normal, resting multigated acquisition scan in the left anterior oblique 45° view. *Top view*, end-diastolic frame. The circular outline of the left ventribular cavity and the crescentic outline of the right ventricular cavity are seen. The curved zone of decreased activity between the two represents the interventricular septum. *Top right*, end-systolic frame. The left ventricular end systolic volume is clearly much smaller than in the end-diastolic frame and all regions of the left ventricule have contracted well. *Bottom left*, background corrected activity–time curve showing the overall count rate from the region of the left ventricle with the end-diastolic count rate normalized to 100 per cent. The ejection fraction is normal at 63 per cent.

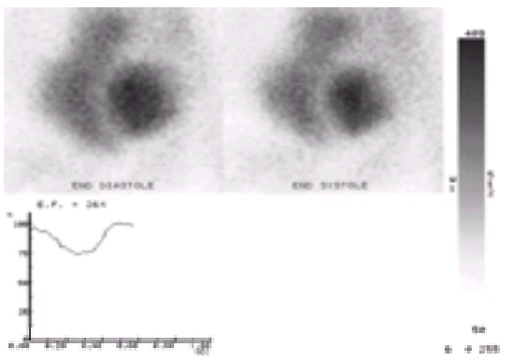

Fig. 5 Abnormal, resting multigated acquisition scan in the left anterior oblique 45° view. The format and layout are as in Fig. 4. There is very little difference between the end-diastolic and end-systolic left ventricular activities and, therefore, volumes (top left and top right respectively) and the reduction in left ventricular contractile performance is uniform. The background corrected activity–time curve shows an ejection fraction of only 26 per cent. The appearances are typical of a congestive cardiomyopathy.

Regional wall motion studies can also be carried out with this technique. Left ventricular images are collected for each of many short time intervals (typically 0.03 s, giving 25 images per cardiac cycle at a heart rate of 80/min) and images for corresponding parts of numerous cardiac cycles (typically several hundred) are summed to produce a composite. In this way 25 'frames' of a 'representative cine cycle' are produced. Figure 4 and Figure 5 show examples of end-diastolic and end-systolic 'frames' of such a representative cine-cycle. Comparison of the end-diastolic and end systolic ventricular boundaries permits the assessment of regional wall motion. In Fig. 4 this reveals normal contraction of those parts of the ventricular wall that are displayed in the view shown. In Fig. 5, from a patient with congestive cardiomyopathy, there is uniformly reduced myocardial contraction.

Currently available non-scintigraphic techniques for the assessment of ventricular function (echocardiography, angiography, etc.) are, in general, less satisfactory when applied to the right than to the left ventricle. The differences are less marked in respect of nuclear techniques, such that scintigraphic procedures currently offer one of the best approaches to the assessment of right ventricular function. The basic techniques involved are the same as for the left ventricle. First pass radionuclide angiography provides adequate temporal anatomical separation of activity within right-sided and left-sided cardiac structures and provides a valuable method for the assessment of right ventricular ejection fraction. The gated equilibrium technique is less useful in the assessment of right ventricular function because of the overlap between the two ventricles, but it is possible to obtain useful information concerning right ventricular size and regional wall motion (usually visually assessed) from this technique. The best approach is to combine the two techniques to obtain a gated first pass study.

Radionuclide imaging in the diagnosis of myocardial infarction

Two general approaches have been used for the detection of myocardial infarction:

1. Recent myocardial damage may be demonstrated using radiopharmaceuticals that concentrate selectively in acutely injured cells. This 'positive imaging', 'infarct-avid imaging', or 'hot spot scanning' can clearly only be applied when infarction has occurred recently (within several days).
2. Non-viable myocardium (i.e recent or long-standing infarction) may be demonstrated by the absence of uptake of several tracers such as thallium ^{201}, technetium-99m-MIBI, or technetium-99m-tetrofosmin as previously described (see above). This is called 'negative imaging' or 'cold-spot scanning'.

Infarct-avid imaging

Three main radiopharmaceuticals have been used for infarct avid imaging: technetium-99m-stannous pyrophosphate, indium-111-antimyosin and, more recently and still in process of investigation, technetium-99m-glucaric acid.

Technetium-99m-stannous pyrophosphate is the compound most extensively used to date. The observation that calcium is deposited in irreversibly damage myocardial cells led to the idea of using this tracer, a bone scanning agent, as a mean of demonstrating myocardial necrosis. The cellular death of myocardial infarction is accompanied by an influx of calcium ions, which are deposited in crystalline and subscrystalline form within the mitochondria, and it has been suggested that calcium accumulation in this way is an index of irreversible cell damage. However, it has also been suggested that the tracer is associated with cytoplasmic denatured macromolecules rather than with mitochondrial hydroxyapatite. Irrespective of the mechanism of tracer uptake, research and clinical work confirm that this radiopharmaceutical localizes in infarcted and severely injured myocardium. For a scan to be positive there must be both (i) significant myocardial necrosis (to give rise to myocardial uptake) and also (ii) persistent residual collateral coronary blood flow into the area of myocardial damage (to permit delivery of the tracer to the infarcted myocardium).

The time interval between the clinical onset of the infarction and scanning is critical. Scans are unlikely to be positive within the first 12 h and the optimum scanning time is 24 to 96 h, but scans can occasionally be positive 2 weeks after an isolated episode of infarction. Between 200 and 600 MBq (approximately 5 to 15 mCi) of technetium-99m-stannous pyrophosphate are given intravenously, with scanning undertaken 60 to 90 min later. Figure 6 shows a normal study and Fig. 7 one of a patient with myocardial infarction.

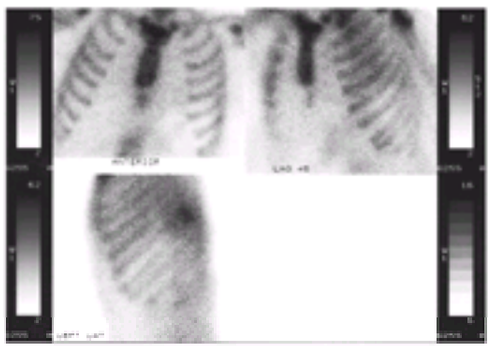

Fig. 6 Normal technetium-99m-stannous pyrophosphate scan, seen in the anterior 45°, and left lateral views. Normal uptake is seen in the sternum and ribs. There is no recognizable activity in the region of the myocardium.

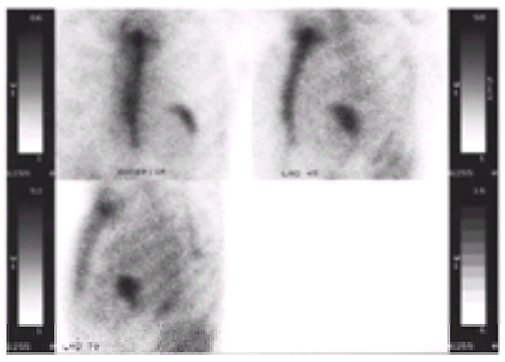

Fig. 7 Abnormal technetium-99m-stannous pyrophosphate scan, seen in the anterior, left anterior oblique 45°, and left lateral views. In addition to the normal uptake in the sternum (well seen) and ribs (less well seen), there is clearly a large area of localized uptake lateral and posterior to the sternum, in the region of the left ventricle. The patient had had an acute myocardial infarction.

False negative and false positive results occur. In 14 different series involving 562 patients with acute myocardial infarction, the false negative rate was 6 per cent. A further group of 15 different series involving 1083 patients with no evidence of acute infarction showed a false positive rate of 17 per cent. The 'efficiency' of the procedure (i.e. its overall ability correctly to classify patients as to whether or not they have acute infarction) is 86 per cent. False positive results have been described in patients with unstable angina, left ventricular aneurysms, cardiomyopathy, valvular calcification, myocardial contusion, persistent blood pool activity, rib fractures, breast tumours, calcified costal cartilages, skeletal muscle damage, and recent cardio-version (the latter giving either skeletal muscle or cardiac damage).

The drawback of this technique is the delay of 1 day before reliable diagnosis of infarction in the non-reperfused myocardial infarction can be made. It is clearly of no value in identifying patients who will benefit from thrombolytic therapy and this is one reason why the technique has not been commonly used in clinical practice.

Indium-111-labelled monoclonal antimyosin, which binds selectively to irreversibly damaged myocytes, has also been used for the investigation of acute myocardial infarction. Clinical studies have shown a sensitivity between 87 and 98 per cent and a specificity of 93 per cent. Because of the high sensitivity and specificity, the technique has been used in the investigation of equivocal myocardial infarction, and in the investigation of right ventricle myocardial infarction. The main drawback is again delay: 12 to 24 h between the administration of tracer and imaging, due to very slow clearance from blood.

More recently, some experience has been gained with the use of technetium-99m-glucaric acid, developed with the hope of achieving an early diagnosis of myocardial infarction. This is a natural dicarboxylic acid sugar which clears from the blood with a very short half-life, allowing images to be carried out within a few hours after injection. Uptake is due to its affinity for the histone of the necrotic myocytes. The clinical usefulness of this compound remains to be determined.

Functional images

Functional images of the heart can be obtained by applying the mathematical technique of Fourier analysis to the left ventricular volume–time curve. The time activity curve is fitted with the first Fourier harmonic, a cosine function with a period equal to the period of the cardiac cycle. The amplitude and phase of this function are adjusted to match the left ventricular curve optimally. The amplitude image reflects the change in ventricular volume through the cycle. It is similar to the stroke volume but may give a more reliable index of the change in chamber volume. The phase image represents the time at which the maximum contraction occurs, and gives information on the mechanical contractility of the heart. Phase data are also presented as histograms in which the number of pixels in an image with a particular phase is plotted against the phase.

Positron emission topography (PET)

This technique uses positron emitting radionuclides and emission computed axial tomography to produce tomographic images of coronary flow and cardiac metabolism. The instrumentation consists of detector systems working in coincidence to register the paired annihilation photons emitted from the radiopharmaceuticals.

PET devices record multiple slices (usually between three and 18) of the heart simultaneously. Perfusion studies can be carried out after injecting radiopharmaceuticals such as ammonia-13, $H_2{}^{15}O$, or rubidium-82. Because of the short half life of this radiopharmaceutical, it is necessary to inject the patient twice, at rest and during maximum exercise. Metabolic studies of the heart have been carried out using carbon-11-palmitate or glucose analogues such as fluorine-18-deoxyglucose. Although these procedures are now more commonly used, they are only performed in routine clinical practice in selected centres. The main drawback of these techniques is cost: they require expensive detectors and a cyclotron on site for the production of radionuclides.

Further reading

Rigo P (1998). Other cardiac applications. In: Maisey MN, Britton KE, Collier BD, eds. *Clinical nuclear medicine*, 3rd edn. Chapman and Hall, London.

Rigo P, Benoit T (1998). Myocardial ischaemia. In: Maisey MN, Britton KE, Collier BD, eds. *Clinical nuclear medicine*, 3rd edn. Chapman and Hall, London.

The heart (1995). In: Wagner HN, Szabo Z, Buchanan JW, eds. *Principles of nuclear medicine*. WB Saunders, USA.

Travin M, Wexler JP (1999). Cardiovascular nuclear medicine (Part 1). *Seminars in Nuclear Medicine* **24**.

Wexler JP, Travin M (1999). Cardiovascular nuclear medicine (Part 2). *Seminars in Nuclear Medicine* **24**.

15.3.5 Cardiovascular magnetic resonance and computed X-ray tomography

S. Richard Underwood, Raad H. Mohiaddin, and M. B. Rubens

Magnetic resonance

Cardiovascular magnetic resonance imaging (**MR**I) has an established clinical role, particularly for the assessment of congenital heart disease and diseases of the aorta and pericardium. However, its clinical impact is currently more limited than that of echocardiography, nuclear cardiology, and invasive investigation. As MRI technology evolves, rapid imaging techniques allow acquisition within a breath-hold and even in real-time. Such techniques promise to extend the capabilities of cardiovascular magnetic resonance to imaging of the coronary arteries and assessment of myocardial perfusion, when it may then play a more important clinical role. For information on the technical aspects of magnetic resonance imaging and spectroscopy, the reader is referred to other texts (see bibliography).

Congenital heart disease

Spin echo images in multiple contiguous slices and in several planes provide excellent anatomical information (Fig. 1). Magnetic resonance compares favourably with echocardiography and cardiac catheterization in providing a complete anatomical diagnosis in 90 per cent of cases, although congenital anomalies of the valves and small defects of the interatrial and interventricular septum are often difficult to visualize on spin echo images alone. Cine gradient echo images provide additional information such as ventricular function, particularly on the right. They also show turbulent blood flow in a manner similar to colour-coded Doppler, and this improves the detection of small ventricular and atrial shunts. The combination of cine imaging with velocity mapping provides further information, improving the detection of shunts and allowing flow to be measured in conduits, great vessels, and within the heart. Shunts can be measured either from the difference in stroke volumes of left and right ventricles or, more flexibly, by measuring flow directly in the aorta and pulmonary artery.

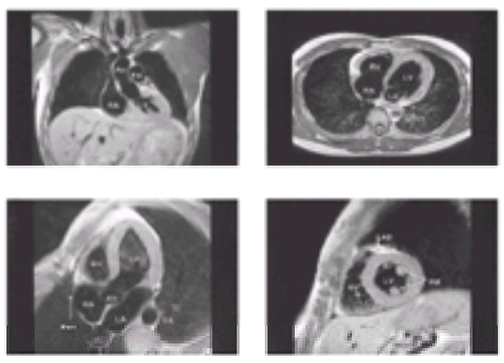

Fig. 1 Spin echo images showing normal anatomy. Top left, coronal; top right, transverse; bottom left, horizontal long axis; bottom right, short axis. LV, left ventricle; RV, right ventricle; LA, left atrium; RA, right atrium; Ao, aorta; DA, descending aorta; PA, pulmonary artery; AV, aortic valve; LAD, left anterior descending coronary artery; peri, pericardium; PM, papillary muscles.

Pulmonary arteries

Several studies have shown the ability of MRI to identify the central pulmonary arteries in patients with pulmonary atresia, which is particularly helpful for determining the feasibility of creating a shunt surgically and for monitoring the growth of the pulmonary artery after shunting. In these patients, magnetic resonance is also able to assess shunt patency accurately, although complete evaluation of systemic collateral arteries, particularly their distal connections, may require selective angiography. Peripheral pulmonary artery stenoses can be missed by MRI.

Pulmonary and systemic veins

Normal pulmonary veins can be identified in most patients and 95 per cent of pulmonary venous abnormalities can be diagnosed. This is superior to cardiac catheterization and to transthoracic echocardiography. Abnormalities of the systemic veins such as a left-sided superior vena cava and its drainage are clearly seen.

Transposition of the great arteries

Postoperative follow-up with MRI is a valuable addition to transthoracic echocardiography for the detection of superior vena caval obstruction in patients with a transposition that has been surgically repaired. Cine gradient echo imaging can improve the assessment by demonstrating abnormal flow patterns associated with residual ventricular septal defects, subpulmonary stenosis, obstruction of the pulmonary venous atrium, and baffle leaks. Because the right ventricle supports the systemic circulation in these patients, outcome is partly determined by right ventricular function and competence of the tricuspid valve. Magnetic resonance can be used to monitor both of these accurately and reproducibly. In a minority of patients, artefacts caused by sternal wires may preclude complete assessment of the right ventricle.

Surgical conduits

Obstruction of conduits between the right ventricle and the pulmonary circulation, such as in tricuspid atresia after the Fontan operation (Fig. 2), may be difficult to detect clinically because patients may be asymptomatic despite having significant obstruction. Magnetic resonance imaging can demonstrate the anatomy of the proximal and distal anastomoses, and of obstruction within the conduit caused by intimal proliferation or 'peel'. A pressure gradient can be determined by velocity mapping and invasive investigation avoided in many cases. By contrast, echocardiography often fails to visualize the conduit because of its position behind the sternum. Palliative systemic to pulmonary shunts are also difficult to assess by echocardiography, but MRI is able to provide anatomical and functional information in

most patients.

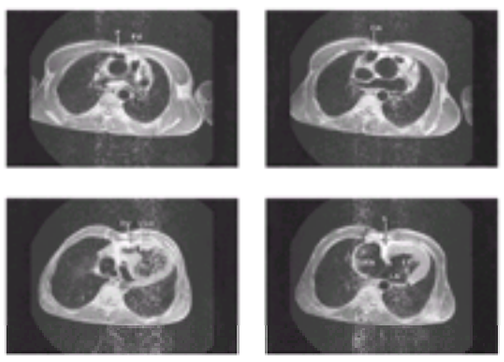

Fig. 2 Transverse spin echo images from superior (top left) to inferior (bottom right) in a patient with tricuspid atresia (S) and a Fontan conduit (F) connecting the right atrial appendage (RA) to the pulmonary artery (>PA). The right ventricle (RV) is hypoplastic and has a ventricular septal defect (VSD) connecting it to the left ventricle (LV). There is also an atrial septal defect connecting the right atrium and left atrium (LA).

Complex disease

Another group of patients in whom MRI has advantages over other techniques are those with complex congenital disease such as single or common ventricles. These anomalies are frequently associated with abnormal thoracic or abdominal situs, and abnormal venous and ventriculoarterial connections. Spin echo imaging is as effective as angiography in demonstrating ventricular morphology and size, the orientation of the septum relative to the atrioventricular valves, and the origins and relationships of the great vessels. Magnetic resonance imaging is superior to other imaging techniques for assessing thoracic and abdominal situs and systemic and pulmonary venoatrial connections.

The aorta

The advantages of magnetic resonance in imaging aortic dissection are its ability to image in oblique planes and the fact that it does not require contrast injection, but it is undoubtedly more difficult to image sick patients in the current generation of scanners than to use other techniques that might be applied in this clinical context (computed tomography (**CT**) or transoesophageal echocardiography). Dissection is readily detected and its extent can be seen, including the involvement of the arch and other vessels (Fig. 3). The ability to demonstrate aortic regurgitation and rupture into the pericardial space are important additional features when assessing these patients. Because the intimal flap is thin it may not always be revealed in spin echo images unless static blood in the false lumen leads to natural contrast with the true lumen. If there is any doubt, then the flap will be more easily seen using a gradient echo sequence, and velocity mapping will confirm the diagnosis by demonstrating the differential flow velocities in each lumen.

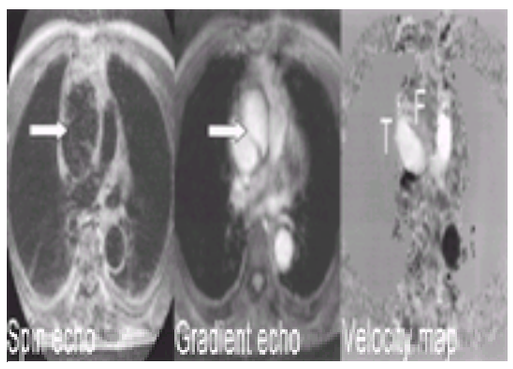

Fig. 3 Three images in the same plane in a patient with aortic dissection. The ascending aorta is dilated and there is an intimal flap (arrow), better seen in the gradient echo image. The velocity map shows rapid systolic flow in the true lumen (T—in white), and absent flow in the false lumen (F—in grey).

Comparisons of MRI with other imaging techniques have concluded that magnetic resonance should be the primary investigation in stable patients and transoesophageal echocardiography the primary investigation in patients who are too ill to be imaged by magnetic resonance. In most cases the investigation performed will depend upon practical issues such as local expertise and availability of equipment.

Other aortic abnormalities that can be seen by magnetic resonance are aneurysms and coarctation. The combination of anatomical imaging and velocity mapping to assess the gradient across the coarctation means that surgical decisions can be taken without invasive investigation in many cases. It is an ideal method for the long-term follow-up of patients following coarctation repair and those with Marfan's syndrome. A further application is in suspected myocardial or mediastinal abscess in postoperative patients with infection that is difficult to control. Echocardiography is often equivocal in such patients and magnetic resonance will usually produce a definitive answer.

Tumours

Magnetic resonance imaging can provide additional information in many patients with masses previously identified by echocardiography. Although it is not possible to identify the nature of a mass from its signal with certainty, the high signal of lipomas and the appearance of angiomas are often characteristic.

Gadolinium–diethylene-triamine-pentaacetic acid (**Gd-DTPA**) may be helpful for demonstrating vascularity and for distinguishing a myxoma from a thrombus. However, even in the absence of a typical signal, a diagnosis can often be made from the site and size of the tumour and from its involvement of neighbouring tissues.

Metastatic tumours are much more frequent than primary cardiac tumours. These can also be imaged successfully, whether as direct invasion of the heart, for example in carcinoma of the bronchus, or distant metastases as in melanoma. Involvement of the myocardium or pericardium can be identified, with the large field of view having a considerable advantage over echocardiography for determining the extent of a tumour.

Thrombus

Atrial or ventricular thrombus is easily identified by MRI, although transoesophageal echocardiography is also reliable in the left atrium, and the transthoracic approach often provides clear images in the ventricles. It is important to combine spin echo MRI with cine gradient echo imaging because it may be difficult to distinguish signal from thrombus and from slowly moving blood in spin echo images. Although the contrast is not as great in cine gradient echo images, it is more consistent, and the presence of a fixed filling defect is characteristic of a thrombus.

Pericardium

Pericardial thickening is readily demonstrated by magnetic resonance and by computed X-ray tomography: both techniques are more accurate than echocardiography. The commonest clinical question is to distinguish between pericardial constriction and myocardial restriction. Visualizing a thickened pericardium with the haemodynamic features of constriction makes the distinction reliably. Cine imaging shows immobility of the pericardium; additional features that indicate constriction are dilated atria and caval veins, small ventricles with retained systolic function, and a reduced diastolic caval flow peak, suggesting impaired right

ventricular filling. Magnetic resonance cannot detect calcification reliably and this may be a drawback if pericardectomy is planned.

Pericardial effusion is clearly seen on spin echo images but its appearance is variable (Fig. 4). Moving fluid gives no signal but static fluid gives a high signal, particularly if haemorrhagic. It can also appear with varying signal in cine gradient echo imaging because rapid through plane refreshment of fluid reduces magnetic saturation. Cine imaging is therefore particularly helpful to distinguish thickened pericardium from a pericardial effusion.

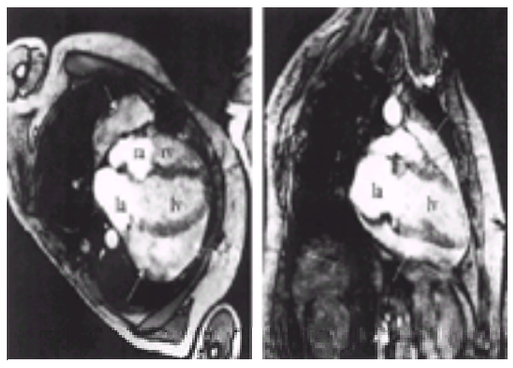

Fig. 4 End diastolic frames from a gradient echo cine acquisition in the horizontal (left) and vertical (right) long axis planes in a patient 3 weeks after heart transplantation. There is a large pericardal effusion (arrows). la, Left atrium; ra, right atrium; rv, right ventricle.

Myocardium

Hypertrophy

The measurement of myocardial volume and mass has been extensively validated in animal experiments and in humans and, because of its accuracy, magnetic resonance should now be the standard against which other techniques are judged. Increased muscle volume (and hence mass) can be observed in athletes and patients with left ventricular hypertrophy, and the regression of hypertrophy following treatment of hypertension can be monitored.

Hypertrophic cardiomyopathy

The location and severity of hypertrophic cardiomyopathy is readily assessed (Fig. 5). Many patients do not have the classical form of asymmetrical septal hypertrophy, and apical hypertrophy in particular is better shown by MRI than by echocardiography. Metabolic abnormalities have also been observed using phosphorus-31 spectroscopy, with a reduced ratio of phosphocreatine to adenosine triphosphate compared with control subjects and patients with dilated cardiomyopathy, and a lower myocardial pH. It is not yet clear whether these changes are specific and whether they will be helpful in distinguishing hypertrophic cardiomyopathy from other forms of hypertrophy.

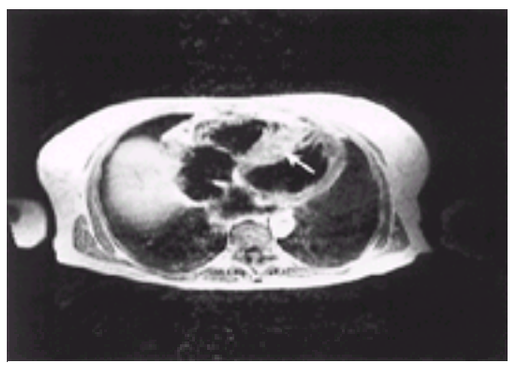

Fig. 5 Transverse spin echo image acquired at mid-ventricular level in a patient with hypertrophic cardiomyopathy. The interventricular septum is asymmetrically thickened (arrow).

Dilated cardiomyopathy

Chamber dilatation and impaired myocardial thickening are clearly demonstrated. Because of the reproducibility of MRI measurements even modest changes of systolic and diastolic function can be monitored serially. Metabolic abnormalities have been demonstrated by magnetic resonance spectroscopy, with a reduced ratio of phosphocreatine to ATP in patients with heart failure and an improvement with therapy.

Other myocardial disease

Non-coronary myocardial disease can manifest itself by abnormalities of global and regional left ventricular function or by abnormalities of relaxation times that lead to differential contrast within the myocardium. Myocardial sarcoidosis is an example of a condition where magnetic resonance may have a useful role because there is a high incidence of subclinical involvement: conventional methods of detection include electrocardiography, echocardiography, and thallium-201 or gallium-67 scintigraphy, but the sensitivity of all these techniques is limited; magnetic resonance can show active involvement either by an increased myocardial signal, indicating active inflammation, or by regional wall motion abnormalities.

Generalized thickening of the myocardium and valves is seen in advanced cardiac amyloidosis. Early involvement may not be apparent, although abnormal diastolic function may be suggestive. Other myocardial diseases in which abnormalities have been demonstrated include myocarditis, systemic lupus erythematosus, Pompe's disease, and Fabry's disease.

Endocardium

Magnetic resonance is not as good as echocardiography at demonstrating small moving structures such as thickened valves and vegetations, but it is able to detect complications of infective endocarditis such as aneurysms and abscesses. The interpretation of spin echo images is only minimally compromised by the presence of a prosthetic valve and, because infection of these valves is a frequent cause of perivalvular abscess, MRI should be used if there is any doubt after echocardiography.

Valve disease

Regurgitation

If only a single valve is regurgitant, comparison of the left and right ventricular stroke volumes allows the regurgitant fraction to be calculated. If single valves on both sides of the heart are regurgitant, the method can be extended by comparing ventricular stroke volumes with flow in the great vessels, the latter measured by magnetic resonance velocity mapping. The regurgitant fraction then compares well with the regurgitant grade assessed by Doppler echocardiography. The method still fails if both valves on one side of the heart are regurgitant, but flow studies in the proximal aorta (or pulmonary artery) can be used to measure aortic (or

pulmonary) regurgitation alone from the amount of retrograde diastolic flow in the artery, and it is then possible to assess even the most complex cases.

Regurgitation can also be detected using cine gradient echo imaging when a turbulent jet of regurgitation is seen as an area of signal loss (Fig. 6). The size of the jet can be used as a semiquantitative measure of regurgitation, although factors other than the size of jet can affect the area of signal loss.

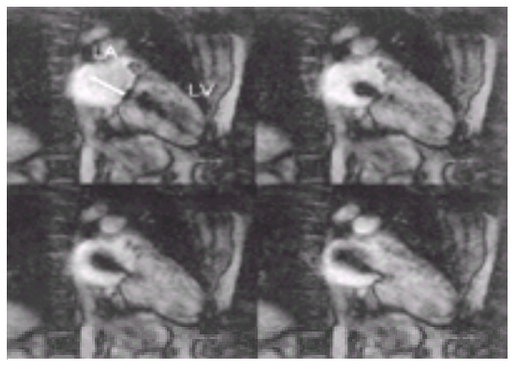

Fig. 6 Four systolic frames from a cine gradient echo acquisition in the vertical long axis plane in a patient with rheumatic mitral stenosis and regurgitation. The regurgitant jet from the left ventricle (LV) to the left atrium (LA) is seen by virtue of signal loss (black) from the turbulence. The size of the jet indicates that regurgitation is moderate.

Stenosis

As with regurgitant jets proximal to a valve, a turbulent distal jet can be used to detect potential stenosis, although abnormal valves that are not stenosed can also generate turbulence. The best method of assessing stenosis is therefore to use cine velocity mapping to measure the peak velocity within the jet. The modified Bernoulli equation, commonly used in Doppler echocardiography, can then be used to estimate the pressure gradient across the stenosis. A disadvantage of magnetic resonance is that it is not yet real-time and so careful alignment of the imaging plane is required in order to obtain an accurate measurement.

Ischaemic heart disease

Myocardial infarction

Magnetic resonance imaging can be used in a number of ways to detect and measure the extent of acute myocardial necrosis. The simplest methods are to use spin echo or cine gradient echo imaging to image the associated wall motion abnormality: the findings agree well with X-ray left ventriculography. Alterations in the myocardial signal can also be observed, an increase of signal in *T*2-weighted spin echo images occurring only a few hours after occlusion of a coronary artery. The changes are most likely related to oedema and the abnormal area may include viable as well as necrotic myocardium.

Abnormal signal can also be observed in *T*1-weighted images, but these changes follow a different time course to the changes of *T*2 and are maximal at 6 weeks, possibly corresponding to cellular infiltration and repair rather than to oedema. Intravenous contrast agents can highlight the abnormalities and have helped to distinguish reperfused from continuing ischaemia in animal models. The same has not been possible in humans.

Reversible ischaemia

Dynamic exercise is impractical within a scanner, but pharmacological intervention using dipyridamole or dobutamine is a suitable alternative. Regional function is assessed using cine gradient echo imaging and global function can also be measured from cine velocity mapping of aortic flow. New wall motion abnormalities imply myocardial ischaemia, and there is a close correspondence between these abnormalities and regional perfusion assessed by radionuclide perfusion imaging. The sensitivity of this approach for detecting coronary artery disease depends upon whether a vasodilator or a b-agonist is used. Because the heterogeneities of myocardial perfusion provoked by dipyridamole do not always cause myocardial ischaemia, sensitivity is not as high (60 per cent) as with dobutamine (91 per cent). The latter is therefore the preferred agent for provoking abnormalities when using a wall motion technique.

An alternative method of assessing reversible perfusion abnormalities is to study the transit of a bolus of magnetic resonance contrast medium through the myocardium. This is not possible using conventional triggered images, but with ultra-fast gradient echo (or echo planar) techniques, images can be acquired in more than 100 ms (or more than 50 ms). Images in each cardiac cycle show the arrival and transit of a bolus of contrast (Gd-DTPA) injected into a central vein. Territories supplied by diseased arteries have a delayed arrival of contrast and a reduced signal increase, and abnormalities can be provoked by dipyridamole vasodilatation. Such bolus-tracking studies of perfusion cannot yet replace radionuclide techniques, but they do have the advantage of higher resolution and the potential to provide measurements of myocardial perfusion in absolute terms.

Coronary vessels

Bypass grafts

Coronary artery bypass grafts can be imaged relatively easily using conventional spin echo or gradient echo techniques. In spin echo images, the appearance of a low intraluminal signal, contrasting with the high signal of the surrounding fat or other soft tissue, implies that the graft contains moving blood and is patent, particularly if it can be followed distally to its insertion. If a graft cannot be identified, or if its origin is seen but it cannot be followed distally, then it is likely to be occluded. Using these criteria, sensitivity and specificity for the detection of patent grafts in the region of 90 per cent can be achieved. Internal mammary artery grafts are more difficult to visualize than saphenous vein grafts, partly because of their smaller size and partly because they can be tortuous and therefore more difficult to follow through multiple slices.

Cine gradient echo imaging has been used with similar results and the cine technique is helpful for identifying a graft if there is doubt from the spin echo images alone. Metallic clips and sternal sutures produce larger artefacts in gradient echo than in spin echo images. They are not ferromagnetic and imaging is perfectly safe, but the artefacts can complicate image interpretation. More recently, contrast-enhanced magnetic resonance angiography has proved useful for demonstrating the anatomy of the tortuous bypass coronary graft.

Native arteries

The coronary arteries are small, tortuous, and rapidly moving: three properties that conspire against successful imaging. Despite this, the proximal vessels can nearly always be identified in conventional spin echo images and their appearance with a low intraluminal signal implies that they contain moving blood and are patent. Unfortunately, resolution is not sufficient to identify stenoses reliably. Rapid gradient echo techniques and acquisition within a single breath-hold provide much better images; moreover, the resolution is adequate for detecting atheromatous disease (Fig. 7). Spiral and echo planar imaging have also been used, all of which can be combined with velocity mapping. Rapid, magnetic resonance flow measurement techniques allow quantification of coronary blood flow reserve, but there is a need to improve the resolution and the reliability of coronary artery imaging and flow measurements non-invasively by magnetic resonance.

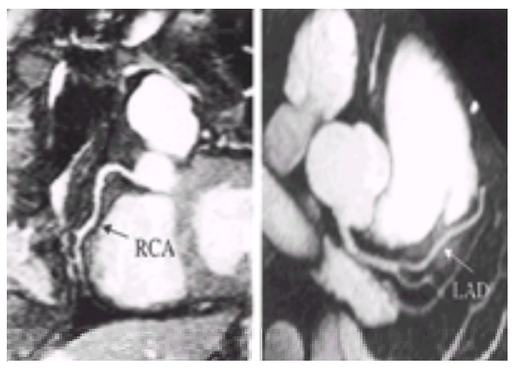

Fig. 7 Breath-hold, contrast-enhanced magnetic resonance angiogram of the right (a) and left (b) coronary arteries in a healthy volunteer. RCA, right coronary artery; LAD, left anterior descending artery.

Computed X-ray tomography

Computed X-ray tomography (CT) is widely available and has an established clinical role in imaging the heart, particularly the pericardium and great vessels (Table 1). Resolution was degraded on earlier generation scanners by cardiac and respiratory motion. State-of-the-art spiral scanners can image the entire heart volume within a single breath-hold, but even the fastest conventional scanner requires 300 ms to acquire a single slice. The best temporal resolution is achieved with electrocardiographic gating and an electron-beam CT scanner (**EBCT**). This uses an electron beam that is focused and deflected on to a stationary target to generate a fan of X-rays that pass through the patient (Fig. 8). In its fastest mode this scanner can acquire two contiguous 8-mm thick slices in 50 ms with a repetition rate of 34 images a second.

Fig. 8 Constrictive pericarditis. Electron-beam CT scan with contrast showing thickened and calcified pericardium, dilated atria, and normal-sized ventricles.

Pericardium

The normal pericardium is seen in 95 per cent of patients by conventional CT, particularly over the anterior surface of the heart. The posterior pericardium is most frequently seen at its caudal insertion into the central tendon of the diaphragm, where it may be 3 to 4 mm thick. It is less easily seen laterally and posteriorly because of the absence of epicardial fat. Pericardial cysts, thickening, calcification, and effusion can be readily identified. The presence of thickening can differentiate pericardial restriction from restrictive cardiomyopathy, although thickening is also seen in a variety of conditions without necessarily implying constriction (Fig. 8).

Aortic dissection

CT is sensitive (83 to 100 per cent) and highly specific (90 to 100 per cent) for the identification of thoracic aortic dissection. This is similar to MRI and transoesophageal echocardiography, but CT is more widely available. The true and false lumens are commonly seen separated by an intimal flap (Fig. 9). Other features that indicate dissection include differential opacification of the true and false lumens, compression of the true lumen by a thrombosed false lumen, inward displacement of intimal calcification, and intramural haemorrhage. Although CT is particularly successful in identifying the distal extent of dissection and the presence of a haemopericardium, it has limitations. Artefacts may create difficulties, and the intimal tear or flap may not be identified in all cases. Therefore, if findings are negative despite strong clinical suspicion of dissection, further investigation is necessary.

Fig. 9 Aortic dissection. Electron-beam CT scan with contrast showing a dilated ascending aorta with almost circumferential dissection. The descending aorta is also involved.

Intracardiac masses

The presence, location, and extent of a thrombus and tumour in the cardiac chambers can be defined with both CT and EBCT. In patients with right heart lesions appropriate windowing of the images will allow assessment of the lungs for pulmonary embolism, and in patients with malignant tumours for metastatic disease.

Cardiac structure and function

Using electrocardiographic gating and intravenous contrast medium, EBCT provides cine images at multiple contiguous levels in approximately eight cardiac cycles. This allows a qualitative assessment of ventricular function and morphology. Planimetry can then provide cavity, muscle area, and hence volume at any part of the cardiac cycle. Regional and segmental left ventricular wall motion can be assessed at rest and during pharmacological stimulation, and changes induced by exercise in patients with ischaemic heart disease can be measured.

Coronary calcification

Coronary artery calcification occurs only in the intima, and microcalcification detected by EBCT indicates the presence of atheroma (Fig. 10), arising when lipid pools first collect within the plaque and not necessarily a sign of advanced disease. A reliable and reproducible scoring system for quantifying calcium has been developed, and a significant relationship between the calcium score and the extent and severity of coronary artery disease as assessed by coronary angiography has been established. However, there is only a weak relationship between the presence of calcium and the severity of luminal stenosis at the same site. The absence of calcification on EBCT does not exclude the presence of coronary atheroma, but it does make significant stenosis unlikely.

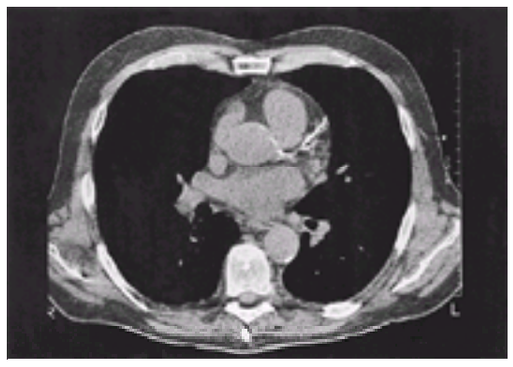

Fig. 10 Coronary calcification. Electron-beam CT scan without contrast. There is high-intensity calcification in the aortic root, the left main stem, left anterior descending, and diagonal coronary arteries.

EBCT offers the possibility of a non-invasive screening test for coronary artery disease, but this use is controversial because the implications of a positive scan are uncertain. It is likely that the incidence of future coronary events is related to the degree of calcification, but the evidence is conflicting. If a relationship between calcium and coronary risk were firmly established, then the technique could play an important role in detecting individuals at risk of coronary events, in selecting people who may benefit most from primary prevention such as lipid-lowering therapy, and in monitoring the progression of disease.

Although EBCT technology is expensive and not widely available, it is possible that modern mechanical CT scanners may also be used to detect coronary calcification, even if in a less sensitive and reproducible fashion.

Coronary angiography

When combined with intravenous X-ray contrast media, EBCT can be used to image the lumen of the coronary arteries and bypass grafts, and three-dimensional images of the coronary tree can be reconstructed from multiple contiguous tomograms. The patency of the arteries and grafts can be determined accurately, but individual stenoses are detected with lesser accuracy and so the technique does not currently provide an alternative to invasive angiography. It is roughly equivalent in terms of accuracy and capability to magnetic resonance coronary angiography. Disadvantages include incomplete visualization of the distal coronary arteries and their branches, and the lack of opportunity to proceed to an intervention if a suitable lesion is demonstrated.

Further reading

Axel L (1998). Physics and technology of cardiovascular MRI. *Cardiology Clinics* **2**, 125–33.

Higgins CB, Hricak H, Helms CA (1992). *Magnetic resonance imaging of the body*, 2nd edn. Raven Press, New York. [Covers the whole field of magnetic resonance in the body, but excludes neurological applications.]

Marcus ML, *et al.*, eds. (1991). *Cardiac imaging: a companion to Braunwald's heart disease.* WB Saunders, Philadelphia. [Reviews of all non-invasive imaging techniques in cardiology, including several chapters on magnetic resonance and computed X-ray tomography.]

Mohiaddin RH, Pennell DJ (1998). MR blood flow measurement. Clinical application in the heart and circulation. *Cardiology Clinics* **16**, 161–87.

Neubauer S, *et al*. (1998). The clinical role of magnetic resonance in cardiovascular disease. *European Heart Journal* **19**, 19–39. [Extensive literature review with recommended indications for MRI in clinical practice.]

Nienaber CA, *et al*. (1993). The diagnosis of thoracic aortic dissection by noninvasive imaging procedures. *New England Journal of Medicine* **328**, 1–9.

Shellock FC, Morisoli S, Kanal E (1993). MR procedures and biomedical implants, materials, and devices: 1993 update. *Radiology* **189**, 587–99. [An ideal entry to the literature on the safety of magnetic resonance imaging.]

Underwood SR, Firmin DN (1991). *Magnetic resonance of the cardiovascular system.* Blackwell Scientific, Oxford. [Textbook covering all aspects of cardiovascular magnetic resonance.]

Wexler L, *et al*. (1997). Coronary artery calcification: pathophysiology, epidemiology, imaging methods, and clinical implications. A statement for health professionals from the American Heart Association. *Circulation* **94**, 1175–92.

15.3.6 Cardiac catheterization and angiography

Edward D. Folland

Introduction

Invasive cardiac diagnosis by means of catheterization and angiography developed hand-in-hand with cardiac surgery throughout the twentieth century. It answered the need for precise information about cardiac physiology and anatomy, which arose in the 1940s when surgical techniques for the treatment of congenital and rheumatic heart disease first became available. A few years earlier, in 1929, Werner Forsman of Germany successfully and safely passed a filiform urinary catheter from a median basilic vein into the right atrium of his own heart and documented it on X-ray film. Although this feat cost him his own job, it enabled Andre Cournand and Dickenson Richards a decade later to use catheters for sampling blood, measuring pressure and flow, and injecting radio-opaque contrast medium (angiography) into the intact, beating human heart, ushering in the era of invasive cardiac diagnosis. Cournand and Richards later won the Nobel Prize for their important work. This chapter will review the diagnostic applications of cardiac catheterization and angiography.

Indications for cardiac catheterization and angiography

Because catheterization is expensive and entails some degree of risk and discomfort, patients should be carefully selected. In broadest terms, it is indicated for detailed evaluation of patients having coronary, valvular, and congenital heart disease once they have been identified as candidates for surgery or other forms of intervention. It may also be indicated for patients whose diagnosis is uncertain from non-invasive evaluation.

Coronary artery disease

The vast majority of patients presenting for cardiac catheterization have coronary artery disease. Angiography of the coronary arteries performed during cardiac catheterization is essential for patients in whom revascularization is indicated. In spite of the limitations discussed later in this chapter, no other imaging modality, including magnetic resonance imaging and computed tomography, can provide the detailed anatomy of the entire coronary circulation that is needed for planning revascularization procedures such as coronary artery bypass surgery and percutaneous intervention.

Coronary angiography is indicated for patients having chronic stable angina, which persists in spite of reasonable efforts at pharmacological therapy. It is also indicated for patients whose survival would be improved, regardless of symptoms. Such patients are those with severe stenosis of the main left coronary artery and those with severe two- and three-vessel coronary artery disease in combination with impaired left ventricular function. These patients may be identified by the following features of stress testing: ischaemia at low workload (especially in stage 1 of the Bruce Protocol), marked depression of the electrocardiographic ST segment (greater than 2 mm), failure to augment systolic blood pressure during exercise, and large exercise-induced defects or increased lung uptake during radionuclide perfusion imaging. In addition, patients having high-risk clinical presentations such as unstable angina and postmyocardial infarction ischaemia are candidates for angiography. Depending upon the availability of emergency revascularization, patients having acute myocardial infarction may be best served by immediate catheterization. The indications for emergency catheterization and percutaneous revascularization instead of thrombolytic therapy will be covered in more detail in Chapter 15.4.5. Finally, catheterization is sometimes indicated for obtaining a definitive diagnosis when non-invasive testing has yielded equivocal or inconsistent results.

Valvular disease

Catheterization was once considered essential prior to the surgical treatment of valvular heart disease. This is no longer the case because of advances in non-invasive testing using ultrasound and Doppler techniques. Nevertheless, catheterization is a frequently helpful technique for gathering the information needed to properly select patients for surgical therapy and to guide the surgeon in providing optimum treatment. The most common reason for catheterization in these patients is to assess the need for coronary artery revascularization, particularly amongst those with aortic stenosis since coronary artery disease is often present in the age group in which this disease commonly occurs. Haemodynamic study may also be necessary in cases where non-invasive diagnostic data are limited or equivocal. By contrast, it is often possible to avoid catheterization in young patients in whom non-invasive studies yield unequivocal conclusions and there is no evidence of coronary artery disease.

Congenital disease

Most patients with congenital heart defects can be definitively diagnosed by transthoracic or transoesophageal ultrasound. As in valvular disease, catheterization is most useful in cases where the abnormality is unusually complex, the non-invasive data incomplete, or the patient is suspected of having coronary artery disease. Catheterization is particularly useful in quantifying shunt flow and pulmonary vascular resistance, both of which are important considerations in the treatment of intracardiac defects. The physical passage of a systemic venous catheter across the atrial septum into a pulmonary vein or the left ventricle is diagnostic of an atrial septal defect.

Pericardial disease

Pericardial tamponade and constriction lend themselves particularly well to diagnosis by catheterization. Although ultrasound has superseded catheterization as a rapidly available method of confirming the clinical diagnosis of tamponade, it is usually inconclusive for patients with pericardial constriction. At catheterization, patients with both conditions usually demonstrate equalization of all intracardiac diastolic pressures. In addition, unique pressure waveforms are exhibited in the right atrium and right ventricle, which usually distinguish the two diagnoses (Fig. 1).

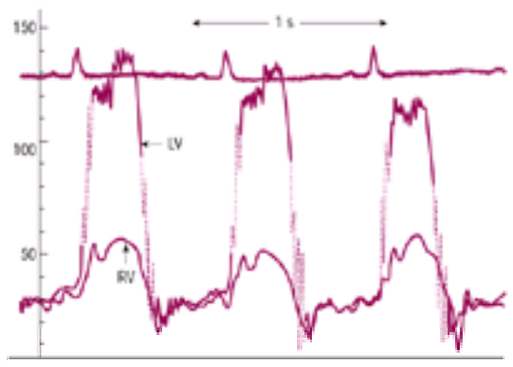

Fig. 1 Pericardial constriction. This is a tracing of simultaneous left ventricular (LV) and right ventricular (RV) pressure in a patient with pericardial constriction. Generally, the diastolic pressure of the left ventricle is higher than that of the right ventricle. For patients with a constriction, the pericardium determines the diastolic compliance of both chambers, causing the diastolic pressures to be equal. Note also the typical 'dip-plateau' pattern or 'square-root sign' of both chambers in diastole. Although diastolic ventricular pressures are also equal for patients having tamponade, the dip-plateau pattern is usually absent.

Congestive heart failure

The aetiology and pathophysiology of congestive heart failure are readily elucidated by catheterization. States of pressure and volume overload as well as systolic and diastolic dysfunction of the ventricles can be easily identified, as explained in detail later in this chapter. Furthermore, catheterization is uniquely suited for identifying transient or reversible causes of left ventricular dysfunction caused by ischaemia or myocardial hibernation due to underlying coronary artery disease. Sometimes exercise or other interventions are performed during a catheter study to elicit transient abnormal haemodynamic function. Myocardial biopsy performed during catheterization can often identify the aetiology of primary myocardial dysfunction.

Pulmonary vascular disease

Patients with primary pulmonary hypertension (see Chapter 15.15.2.1) should undergo catheterization to measure pulmonary vascular pressure and resistance. Certain vasodilating drugs may or may not benefit the patient, depending upon their effect on pressure and resistance during acute administration. Pulmonary angiography performed during right heart catheterization is still regarded as the most definitive test for pulmonary embolism, in spite of advances in radioisotope lung scanning and spiral computed tomography.

Preparing the patient for catheterization

Precatheterization evaluation should consist of a careful history and examination particularly aimed at eliciting details of prior cardiac procedures, reactions to contrast medium, renal function, peripheral vascular status, and haemostatic function. The patient should be carefully advised of the indications, alternatives, risks, discomforts, and expected benefits of the procedure. The skilled clinician does this while building the patient's confidence and avoids creating undue alarm. Following an uncomplicated diagnostic catheterization the patient should usually expect to go home the same day and to resume customary physical activities within a day or two.

Approaches to cardiac catheterization and angiography

Vascular access

The traditional approach to vascular access is via a cut-down near the antecubital fossa. Isolating and mobilizing the brachial or antecubital vein and the brachial artery for right and left heart catheterization may thereby achieve arterial and venous access. Following the procedure the arterial entry site is repaired by suture and the vein is usually tied off. This approach has the advantages of enabling earlier postprocedure ambulation and the security of direct arterial closure in anticoagulated patients. It has the disadvantage of being more time-consuming for most physicians and less cosmetic for the patient.

Percutaneous vascular access is achieved by direct puncture with a needle through which a flexible spring guidewire is passed into the vessel. Catheters may then be passed into the vessel over the guidewire. (The guidewire is placed in the catheter to maintain access during catheter exchanges.) Following the procedure haemostasis is achieved by applying pressure over the puncture site until bleeding stops. Percutaneous access is most frequently employed at the femoral site, although it may also be used at brachial, axillary, internal jugular, and radial locations. It has the advantage of speed, simplicity, and, when performed from the femoral vessels, frees the upper body and arms during angiographic filming. However, percutaneous access has the disadvantage of requiring several hours' immobilization of the catheterization site following the procedure, which is particularly troublesome after femoral puncture since 6 h of postprocedure bedrest is frequently required. Nevertheless, the percutaneous femoral approach has become the preferred choice in 90 per cent of cases, with the recent use of smaller catheters (4 and 5 French) and closure devices for the arterial puncture site enabling earlier ambulation. The percutaneous radial approach is becoming increasingly popular for outpatients.

Right heart catheterization

Right heart catheterization can be performed from any of the approaches described above. Although traditionally performed with a stiff, woven Dacron, end-hole catheter, it is often done with a flexible, balloon-tip, flow-directed catheter (Swan–Ganz) because this is safer and enables the measurement of cardiac output by thermodilution.

Catheterization of the right heart is indicated by itself for the study of pulmonary vascular disease and haemodynamic response to exercise or drug administration. It is indicated in combination with left heart catheterization for patients requiring haemodynamic study of valvular, congenital, or myocardial disease, and for patients being studied primarily for coronary artery disease who also have heart failure, valvular, or pulmonary disease.

Left atrial pressure can be measured indirectly via right heart catheterization by wedging the tip of the catheter in a pulmonary arteriole, or y occluding a pulmonary artery branch with the inflated balloon at the tip of a Swan–Ganz catheter. In either case, this creates a static column of blood from the tip of the catheter, through the pulmonary capillary bed, to the left atrium. This static column of blood has the effect of extending the tip of the catheter to the left atrium for pressure-measuring purposes. The resulting pressure is identical to the directly measured left atrial pressure, except that it is delayed temporally by approximately 80 milliseconds. This pressure, commonly known as the pulmonary (artery) capillary wedge (**PCW**) pressure, is very useful in the management of left heart failure and shock, and for estimating the diastolic gradient across the mitral valve in patients with mitral stenosis.

Left heart catheterization

Left heart catheterization is generally performed in conjunction with coronary angiography, but is specifically required for the assessment of left ventricular function and assessment of stenosis or regurgitation of the left-sided valves (mitral and aortic). It is most often accomplished by femoral or brachial arterial access, and by retrograde crossing of the aortic valve to enter the left ventricle. Left heart catheterization may also be achieved by controlled puncture of the interatrial septum with a catheter originating from the right femoral vein (trans-septal left heart catheterization): this can then be used to measure left atrial pressure directly, and be passed antegradely through the mitral valve to measure pressure and perform angiography of the left ventricle. Retrograde access of the left atrium from the left ventricle is technically difficult and seldom done. The left ventricle may also be entered via transthoracic needle puncture. This approach, known as direct left ventricular puncture, is occasionally necessary for studying patients who have mechanical prosthetic valves at both mitral and aortic positions. The passage of the needle into the left ventricle from the cardial apex is facilitated by echocardiographic guidance.

Information obtained from cardiac catheterization and angiography

Intracardiac pressures

Methodology

Pressure at the tip of the catheter is transmitted through the fluid inside the catheter (usually saline) to a device called a transducer, which converts the pressure signal to an electrical signal that can then be amplified and displayed on a television screen or on a strip-chart paper recording. Once calibrated, the pressure at the tip of the catheter can be read graphically from the recording screen or paper. The fidelity of recording depends upon the physical characteristics of the fluid-filled catheter, stopcocks, connecting tubing, and the pressure transducer itself. A fluid-filled system is usually capable of responding to transient pressure changes up to 20 or occasionally 30 Hz. This is sufficient fidelity to reproduce diagnostically useful pressure waveforms from the heart. However, it is not responsive enough to accurately reproduce the rate of rise of left ventricular pressure during the isovolumic phase of systole (d *P*/d*t*). This requires responsiveness to transient pressure changes of at least 60 Hz, of which fluid-filled catheter systems are not capable. For such applications catheter-tip manometers are available (Millar catheters) in which the transducer is placed at the tip of catheter, eliminating the need for an intervening column of fluid. These devices are expensive and are used only when such fidelity is required, usually in research applications.

Normal intracardiac pressures

The upper limits of all normal intracardiac pressures measurable from a right heart catheter are approximate multiples of six, hence they are easily remembered by 'The Rule of Sixes' (Table 1). For example, the mean right atrial pressure is 6 mmHg or less, mean left atrial pressure is 12 mmHg or less. A further aid to remembering normal pressures is the 'Corollary of Continuity', which means that contiguous chambers have a common pressure when the intervening valve is open. For example, the right ventricle and right atrium are essentially a common chamber when the tricuspid valve is open in diastole, therefore the upper limit of right ventricular end-diastolic pressure is the same as the upper limit of the normal right atrial pressure, or 6 mmHg. This assumes there is no significant stenosis or regurgitation across the tricuspid valve, and that the right ventricle has normal compliance. The same condition applies to the mitral valve in diastole and the pulmonic and aortic valves in systole. Another practical rule is that the pulmonary artery diastolic and pulmonary artery capillary pressures approximate each other in the absence of severe pulmonary vascular disease. Once this has been established for any given patient, the pulmonary artery diastolic pressure can be followed as a surrogate for pulmonary capillary wedge pressure in situations where a pulmonary artery catheter is used for intensive-care monitoring.

All intracardiac pressures rise and fall phasically with breathing due to transmission of shifting intrapleural pressure during respiratory effort. Usually this variation is no more than a few mmHg from inspiration to expiration, but it can be quite marked in patients with obstructive lung disease. Standards of normal pressure are based upon measurements taken during resting respiration, averaging several respiratory cycles. Pressures in the catheterization laboratory should be similarly measured: asking a patient to hold his or her breath may generate misleading data.

Waveforms

The shape of intracardiac pressure waveforms carries useful diagnostic information. Atria and ventricles have characteristic waveforms, the left-sided chambers normally demonstrating similar patterns at relatively higher pressures than right-sided chambers. The state of volume loading and the relative compliance or 'stiffness' of the respective ventricles during diastolic filling determines pressures in the right and left atria. The left ventricle is generally thicker, stiffer, and less compliant to the stretch of increasing volume than the right ventricle; hence the left atrial and left ventricular diastolic pressures are higher than the respective pressures in the right heart. Conditions such as pericardial constriction and tamponade alter this normal relationship (Fig. 1).

Cardiac flow and output

Measurement of cardiac output was one of the earliest applications of catheterization. Most methods entail application of the indicator dilution theory (the 'Fick' principle), summarized graphically in Fig. 2. Stated simply: the rate of flow can be measured if an indicator substance is added to the moving vehicle (for example, blood) at a known rate, and the concentration of the indicator is also known proximal and distal to the point where the indicator is added. The indicator can be any readily measured substance such as oxygen, indocyanine green dye, or saline, the temperature of which is known and different from that of the bloodstream.

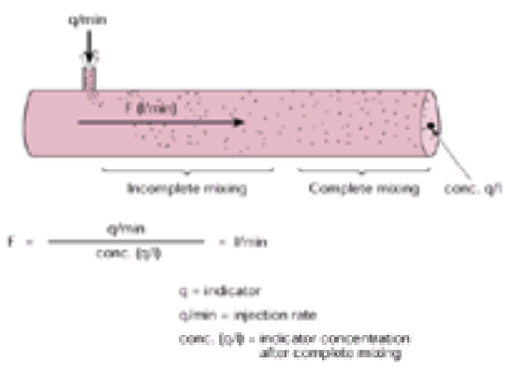

Fig. 2 The Fick principle. The flow rate (*F*) through a vessel (cardiac output, in this case) can be measured if an indicator is added to the flowing liquid at a known rate (*q/min*) and the concentration (*q/L*) of the indicator is measured after complete mixing has occurred.

Cardiac output by oximetry

In this method, commonly called the 'Fick method', the indicator is oxygen which is carried physiologically by the blood. The method requires that the subject be in a metabolic steady state where the use of oxygen is constant. Such a steady state exists at rest and also during exercise, provided that the workload is constant for at least 3 min. As seen in Fig. 3, the pulmonary blood flow can be calculated when the oxygen consumption rate is known and the oxygen contents of blood in systemic and pulmonary arteries are known. In the absence of intracardiac shunts the pulmonary blood flow equals the systemic blood flow, or cardiac output.

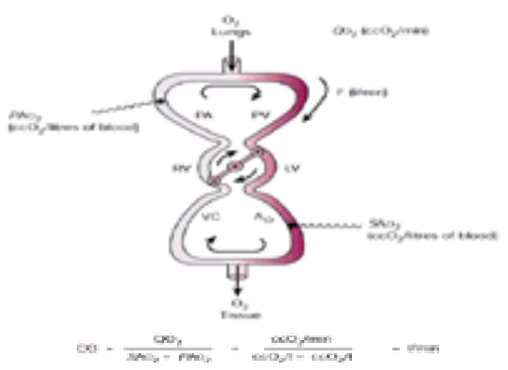

Fig. 3 Cardiac output measured by oximetry. This is an application of the Fick principle in which oxygen is the indicator carried by flowing blood. The patient's metabolism must be at steady state, a condition where oxygen consumption and utilization are matched. It requires three measurements: oxygen consumption rate (QO_2), systemic arterial oxygen content (SAO_2), and pulmonary arterial oxygen content (PAO_2). Other abbreviations: Ao, aorta; CO, cardiac output; LV, left ventricle;

PV, pulmonary vein; RV, right ventricle; VC, vena cava; cc, cm_3.

Dye dilution

This method entails the rapid injection of a known quantity of indocyanine dye into the pulmonary artery. Blood is then sampled by withdrawal at a constant rate from a systemic artery. The sampled blood passes through a spectrophotometer, which is calibrated to measure the concentration of dye. A concentration curve is inscribed when the injected bolus of dye passes the sampling point (Fig. 4). Dividing the quantity of dye injected by the area of the time–concentration curve (corrected for recirculation) yields the cardiac output.

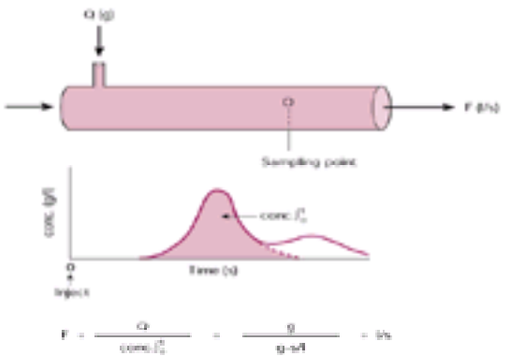

Fig. 4 Cardiac output measured by dye curve. The concentration curve of indocyanine green dye generated by sampling distal to an injection point can be analysed to yield cardiac output. See text for more details. Thermodilution cardiac output employs the same principle, except that temperature is the measured indicator. F, flow or cardiac output; Q, quantity of indicator injected.

Thermodilution

Measurement of cardiac output by thermodilution uses the same principle as dye dilution, with the indicator being 'negative calories' (the difference in caloric content of the injected bolus of cool saline compared to the caloric content of the same quantity of the subject's blood). The downstream 'concentration' of injected negative calories is measured as a transient drop in temperature by a thermistor at the tip of the injection catheter several centimetres from the point of injection. Dividing the negative calories injected by the area of the distal time–temperature curve yields cardiac output. The advantages of speed, automaticity, and repeatability of this method make it particularly suitable for serial measurements during different haemodynamic states.

Angiographic output

This is the only commonly used method that does not employ the indicator dilution or Fick principle. The left ventricular stroke volume calculated from quantitative angiography is multiplied by the heart rate to yield the left ventricular output. In the absence of valvular regurgitation this is the same as cardiac output. As explained in greater detail later in the chapter, this method is particularly useful in assessing mitral and aortic valvular regurgitation.

Quantitative angiography

Quantitative left ventricular angiography enables the measurement of left ventricular volume at instants throughout the cardiac cycle. Radiographic contrast medium is rapidly injected into the left ventricle and the shadow image of the opacified ventricle captured on film or electronically at a particular frame rate in any chosen projection. The most common projection is 30 degrees right anterior oblique at a filming rate of 30 frames per second. In this view the image of the left ventricle is parallel to its long axis, resembling an ellipse, or an American football/rugby ball. Arvidsson and Greene first suggested that the volume of the left ventricle could be calculated from the volume formula for an ellipsoid, the three-dimensional structure created by rotating an ellipse on its long axis. Dodge and Sandler improved upon this concept by deriving the minor hemi-axes from an idealized ellipse of the same length and area as the projected image of the ventricle. This method is still commonly used and is often referred to as the area–length method. Images captured at end-diastole and end-systole are analysed and corrected for magnification to yield end-diastolic and end-systolic volumes, the difference between these volumes being the stroke volume and the product of the stroke volume and heart rate, the angiographic left ventricular output. These indices are useful in the assessment of left ventricular function and valvular regurgitation as discussed later in this chapter.

Intracardiac shunts

The same methods of oximetry and indicator dilution utilized in measuring cardiac output can be employed for the detection and quantitation of intracardiac shunts. Under normal resting conditions, blood is approximately 75 per cent saturated as it returns from the body to the right heart and pulmonary artery. As it leaves the lungs in the pulmonary veins blood is 99 per cent saturated. Intracardiac shunts can be detected, localized, and quantified by measuring the oxygen saturation in various locations. Left to right shunts will cause a step-up in the saturation of the blood at the location of the shunt; for example, in a patient with an atrial septal defect the saturation will rise in the right atrium, whereas with a ventricular septal defect the saturation will rise in the right ventricle. A patient with Eisenmenger's syndrome (pulmonary hypertension and right to left shunting) will exhibit a drop in saturation at the location of the shunt, namely at the left atrium or ventricle in the case of atrial and ventricular septal defects, respectively. The degree of the change in saturation is proportional to the size of the shunt, and enables calculation of the shunt flow in either direction in litres per min. Figure 5 presents a scheme and formulas for calculating shunt volume.

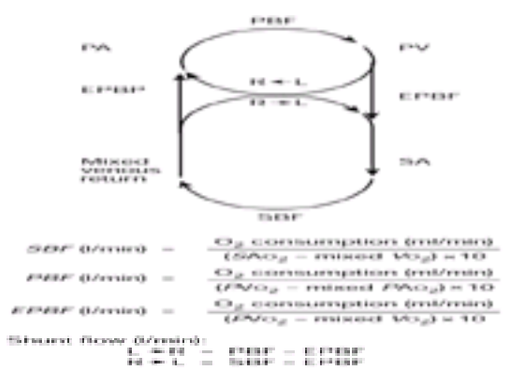

Fig. 5 Quantitation of intracardiac shunts. Shunts between the left and right sides of the heart due to septal defects can be quantified by oximetry using this scheme. PBF, pulmonary blood flow; SBF, systemic blood flow; EPBF, effective pulmonary blood flow, which is that part of the systemic venous return that actually passes through the lungs and is oxygenated; PAO_2, pulmonary artery oxygen content; PVO_2, pulmonary vein oxygen content; SAO_2, systemic artery oxtygen content; mixed VO_2, mixed systemic venous oxygen content.

Vascular resistance

Methodology

Blood flow through the pulmonary and systemic circulations can be compared to the flow of an electric current through a circuit. Pressure is the driving force analogous to voltage, flow rate is analogous to current, and the impediment to flow through the vascular bed is resistance. Pressure, flow, and resistance relate to each other in a fashion analogous to Ohm's law:

$$\text{Resistance} = \text{pressure/flow}.$$

In the above formula 'pressure' is the difference in mean pressure across the systemic vascular bed (systemic arterial pressure—right atrial pressure) or the pulmonary vascular bed (pulmonary artery pressure—left atrial pressure). In the absence of intracardiac shunts 'flow' is the same for both circulations and is measured as cardiac output by methods already described. In cases of intracardiac shunting the systemic and pulmonary flows will differ according to the degree of shunting, and can be calculated as described under the section on cardiac shunts. Normal values for pulmonary vascular and systemic vascular resistance are expressed either in dyne s cm^{-5} or Wood units and are displayed in Table 2. Total pulmonary resistance is a useful concept for expressing the total resistance against which the right ventricle must work, and includes not only the pulmonary vascular resistance, but also the resistance engendered by the static pressure in the left atrium. Hence, pulmonary vascular disease, left heart failure, or both, can increase the total pulmonary resistance.

Clinical application

Measurement of resistance is useful for assessing the state of the pulmonary circulation in congenital heart disease with intracardiac shunting: high pulmonary vascular resistance may preclude the safe correction of an intracardiac shunt, particularly if the shunt is from right to left. It is also useful in diagnosing the relative contribution of left heart failure and pulmonary vascular disease in patients with pulmonary hypertension, and is the best indicator of the effectiveness of vasodilating drugs for patients with pulmonary hypertension.

Valvular stenosis

Valvular stenosis is assessed by measuring the transvalvular pressure gradient and by calculating the valvular orifice area using a formula introduced in the late 1940s by cardiologist Richard Gorlin and his father, an engineer. The Gorlin formula for valve area was initially developed for patients with rheumatic mitral stenosis. It is based upon a study which utilized data from right heart catheterization alone, validated by relatively crude intraoperative estimates of valve area using the index finger of surgeon Dwight Harken during closed mitral commissurotomy operations at the Peter Bent Brigham Hospital in Boston, Massachusetts. In spite of this, the formula has stood the test of time and remains the standard for the haemodynamic assessment of valvular stenosis. In its generalized form it is expressed as follows:

$$\text{Valve area} = \text{transvalvular flow rate}/K \text{ Ögradient}.$$

In the above formula K is a constant unique to mitral or aortic valve analysis (38 and 44.5, respectively). The transvalvular flow rate (**TFR**) is cardiac output normalized for the time that the valve is actually open. In aortic valve applications TFR is the cardiac output divided by the product of heart rate and systolic ejection period. In mitral valve applications it is the cardiac output divided by the product of heart rate and diastolic filling period. Cardiac output is the effective systemic blood flow as determined by Fick, thermodilution, or dye dilution methods unless there is associated valvular regurgitation, in which case it is the total left ventricular output as determined by quantitative left ventricular angiography. Gradient is the mean pressure gradient in mmHg during the time when the valve is open.

Figure 6 shows tracings that demonstrate typical gradients from patients with aortic and mitral stenosis. The ranges of calculated valve area associated with various levels of stenosis for both aortic and mitral valves are displayed in Table 3. In general, procedures performed for the relief of anatomical stenosis are expected to be beneficial in symptomatic patients with severe valvular obstruction. However, many factors enter into such a decision and individual clinical judgement is required. Although patients with large transvalvular gradients generally experience the best result from intervention, the gradient by itself can be misleading due to its exponential relationship to cardiac output.

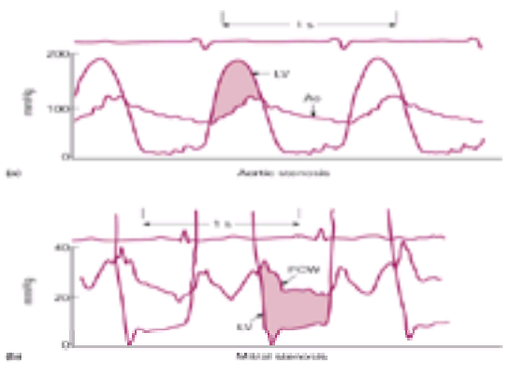

Fig. 6 Pressure gradients associated with valvular stenosis. The upper panel shows simultaneous tracings of left ventricular (LV) and ascending aortic (Ao) pressure in a patient with severe aortic stenosis. The mean systolic gradient across the aortic valve is 60 mmHg. The lower panel shows simultaneous tracings of left ventricular (LV) and pulmonary capillary wedge (PCW) pressure in a patient with severe mitral stenosis. The mean diastolic pressure gradient across the valve is 16 mmHg. The respective valvular gradients are cross-hatched.

Valvular regurgitation

Qualitative assessment

Regurgitation of all four cardiac valves can be qualitatively assessed by angiography. The downstream side of the valve in question is opacified by a rapid injection of radiographic contrast medium. Regurgitation is visualized as upstream leakage of contrast across the closed valve. In the case of mitral regurgitation systolic opacification of the left atrium occurs during injection of the left ventricle. In aortic regurgitation diastolic opacification of the left ventricle occurs during supravalvular injection of the aorta. The degree of regurgitation is graded on an arbitrary scale from mild (1+) to severe (4+).

Quantitative assessment

Aortic and mitral regurgitation can be quantified in terms of regurgitant flow in litres per min or regurgitant fraction as a percentage of left ventricular output. The method requires measurement of the total left ventricular output by the angiographic method and subtraction from that of the effective forward output measured by the Fick or indicator dilution methods (both described earlier). It is the best method for measuring the severity of regurgitation, provided that the left ventricular angiogram, which itself changes cardiac output, is performed soon after the Fick measurement. Furthermore, both measurements must be made with considerable care to ensure accuracy. Regurgitation is considered clinically severe when 50 per cent or more of the total left ventricular output is simply shuttling or regurgitating across the defective valve. The ability to quantify regurgitation across either valve is lost when both mitral and aortic valves are leaky.

Left ventricular function

Global function

Global function of the left ventricle is broadly described by its ability to generate pressure and flow under particular conditions of preload and afterload. Plotting the pressure and volume of the left ventricle at instants in time for a single cardiac cycle generates a pressure–volume loop displayed in Fig. 7. Most of the commonly

used indices of left ventricular function can be derived from such a loop, including end-diastolic volume, end-systolic volume, stroke volume, ejection fraction, end-diastolic pressure, and d*P*/d*t*. Of these, the ejection fraction is most useful because it correlates with prognosis in a variety of cardiac diseases.

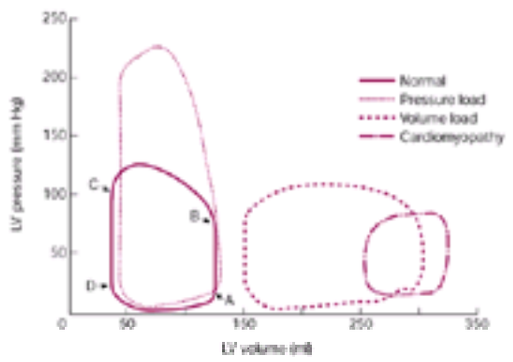

Fig. 7 Pressure–volume loops. Simultaneously plotting the instantaneous pressure and volume of the left ventricle throughout a single cardiac cycle produces these loops. The loop is a synthesis of most information relevant to left ventricular function. In this figure a loop from a normal patient is contrasted with those from patients with pressure load (hypertension or aortic stenosis), volume load (aortic or mitral regurgitation), and cardiomyopathy. Point A represents mitral valve closure; segment A–B, isovolumic contraction; point B, aortic valve opening; segment B–C, systolic ejection; point C, aortic valve closure; segment C–D, isovolumic relaxation; point D, mitral valve opening; and segment D–A, diastolic filling.

Grading angiographic wall motion in various segments of the left ventricle as normal, hypokinetic, akinetic, or dyskinetic assesses the regional function of the left ventricle. Regions of abnormal function generally correspond to locations of infarcted myocardium.

Contractility

This parameter is difficult to assess in the intact heart, because all pressure and volume indices are dependent upon preload and afterload. Although ejection fraction is clinically useful it can be misleading in situations of high afterload (for example, severe aortic stenosis) and low afterload (for example, severe mitral regurgitation). The concept of 'elastance' has gained favour as a useful index of intrinsic contractility, because it is relatively independent of loading conditions. Elastance is the slope of the line generated by plotting the end-systolic left ventricular pressure from a series of pressure–volume loops generated at differing afterloads created by the infusion of pressor or vasodilator drugs. The method is laborious and generally reserved for research applications.

Diastolic function

Diastolic function of the left ventricle is best appreciated from the slope of the pressure–volume loop during the period from mitral valve opening to its closure at the onset of systole. The curve becomes steeper as the left ventricle becomes less compliant due to the effects of hypertrophy, ischaemia, or infiltrative disease. In general, left ventricular end-diastolic pressure (**LVEDP**) rises as diastolic compliance falls, accounting for the high left atrial pressure and heart failure seen in diastolic left ventricular dysfunction.

Assessment of coronary arterial anatomy and function

Disease of the coronary arteries may be characterized at catheterization by both anatomical and functional assessment. Coronary angiography images the lumen of the vessel, which has been rendered radio-opaque by injection of radiographic contrast medium. It is a shadowing technique, which displays the impact of the lesion on the arterial lumen, but does not image the plaque *per se*. Intracoronary ultrasound provides a tomographic image of the vessel wall and is capable of demonstrating the thickness and sonic density of the vessel wall and any associated plaque. Angiography and intravascular ultrasound are complementary methods of assessing vascular anatomy. To learn the haemodynamic importance of a coronary lesion it may be necessary to analyse its effect on function by measuring pressure and flow in the affected vessel. All these anatomical and functional modalities may be accomplished by catheterization.

Coronary arteriography or angiography

Coronary arteriography or angiography is presently the single most essential application of cardiac catheterization. The anatomy of coronary arteries in living, conscious humans was first demonstrated by non-selective injection of the aortic root. In the early 1960s David Littmann developed a loop catheter that enabled the injection of contrast medium preferentially in the outer circumference of the aortic root, opacifying the left and right coronary arteries simultaneously. At the time it was commonly believed that selective injection of contrast material into a coronary artery would have fatal consequences. This changed when Mason Sones accidentally performed the first selective coronary angiogram without harm. He was intending to inject the left ventricle, but the catheter recoiled across the aortic valve and into the right coronary artery. Sones, a cardiologist by training, went on to develop a safe method of selective coronary angiography from the brachial artery cut-down approach using the flexible-tip catheter bearing his name. At the same time Melvin Judkins, a radiologist by training, was perfecting his own method of selective coronary angiography, using preshaped catheters, from a percutaneous femoral artery approach. Both methods have continued to be practised, although the percutaneous femoral, or Judkins' approach, has become most popular because of its speed and simplicity. In recent years there has been a return to the brachial and even radial artery approach using percutaneous methods, which enable more rapid patient ambulation.

Normal coronary anatomy is demonstrated in Fig. 8. A patient's anatomy is considered to be right- or left-dominant, depending upon whether the posterior descending artery arises from the right or left coronary artery, respectively. Approximately 80 per cent of humans are right-dominant.

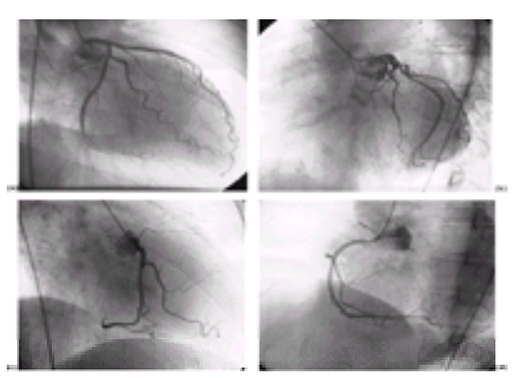

Fig. 8 Normal coronary anatomy. Left coronary angiogram showing main stem, left anterior descending, and left circumflex arteries from right anterior oblique (RAO) view (a) and left anterior oblique (LAO) view (b). Right coronary angiogram showing right coronary and posterior descending arteries from RAO view (c) and LAO view (d).

Atherosclerotic disease is manifest by lesions that encroach upon the opacified lumen of the coronary artery (Fig. 9). Various approaches are used to grade the severity of these lesions. Most commonly a visual estimate of the percentage of the diameter of stenosis is given to each lesion. Lesion severity may be quantified by comparing the minimal lumen diameter within a lesion to the diameter of the nearest normal segment of artery. This can be done manually using callipers or automatically using computer-based systems for edge detection and contrast densitometry. Quantitative coronary angiography is a complex subject because it

requires attention to many variables, such as selection of view and frame, and choice made from among several analytical techniques.

Fig. 9 Atherosclerotic coronary artery disease. The constrictions and blunt terminations seen in this patient's coronary angiogram represent atherosclerotic lesions.

Early work by Lance Gould determined that a lesion must impair coronary blood flow to be clinically important. Although flow at rest is usually not reduced until the diameter of the stenosis exceeds 90 per cent, flow under stress may be reduced when the diameter of the stenosis is 70 per cent. The clinical impact of a stenosis of any given severity is also dependent upon the degree of collateral flow into the vascular bed distal to the stenosis.

Flow and pressure may be directly measured in the coronary artery by means of special guidewires that have pressure transducers or Doppler flow transducers mounted near their tips. As mentioned above, the flow at rest may be normal across a particular coronary artery stenosis. Coronary flow normally increases after maximal vasodilatation induced by local vasodilators. The quotient of the vasodilated flow divided by the resting flow is called the coronary flow reserve, which is normally greater than two. If not, the lesion in question is considered to be haemodynamically important. Pressure can be measured in the coronary artery at a location distal to a lesion using a guidewire with a transducer at its tip. The quotient of pressure distal to a lesion compared to the proximal pressure during maximal vasodilatation is called the fractional flow reserve. A quotient less than 0.75 is considered to be clinically important.

Intravascular ultrasound

Intravascular ultrasound (**IVUS**) is accomplished by advancing a catheter over a guidewire previously placed into a coronary artery. The catheter has a miniature ultrasound transducer near its tip, which enables rotational Doppler imaging of the vessel wall in a plane perpendicular to its axis. IVUS is particularly useful for assessing the nature of angiographically questionable lesions, determining the true size of the vessel prior to stent deployment, and assessing the completeness of stent deployment. It is also probably the best method for serial studies of coronary anatomy during drug treatment trials, because it is able to image the plaque itself and is therefore a more sensitive method than angiography.

Complications

Although cardiac catheterization is a relatively safe procedure, it is nevertheless important for both the patient and the referring physician to recognize the nature and likelihood of potential complications. Table 4 lists the complications of bilateral heart catheterization including coronary, left ventricular, and aortic angiography in a prospective study of valvular heart disease from the United States Veterans Administration. Even though these data were collected over 20 years ago from a particularly high-risk group of patients, the frequency of complication is a realistic estimate of what should currently be expected. The rate of each particular complication will vary with the age and general health of the patient. For example, the risk of vascular complication is considerably increased by the presence of vascular disease, and the risk of renal failure due to contrast medium is particularly high in diabetic patients with pre-existing renal dysfunction. Therefore, in counselling the patient regarding the likelihood of untoward events it is important to give individualized advice based upon the patient's particular circumstances. Finally, the decision to recommend catheterization must be based upon the anticipation that its benefits justify its cost and risk.

Further reading

Baim DS, Grossman W, eds (2000). *Cardiac catheterization, angiography, and intervention*, 6th edn. Williams and Wilkins, Baltimore. [This is a standard textbook for the field of invasive cardiology. It covers all the subjects presented in this chapter in greater detail and gives detailed references to primary sources.]

15.4.1.1Influences acting *in utero* and early childhood

D. J. P. Barker

The fetal origins hypothesis

Over the past 10 years epidemiological studies have shown that people who had low birthweight, or who were thin or short at birth, are at increased risk of developing ischaemic heart disease and the related disorders stroke, hypertension, and non-insulin dependent diabetes (NIDDM). Associations between small size at birth and later disease, first recorded in Britain, have now been extensively replicated in studies in Europe and the United States. The associations extend across the whole range of birthweight and depend on lower birthweights in relation to the duration of gestation rather than the effects of premature birth. They are not the result of confounding variables acting in later life, such as low socio-economic status and smoking.

These observations have given rise to the 'fetal origins hypothesis', which proposes that cardiovascular disease originates through adaptations which are made by a fetus when it is under-nourished. Unlike adaptations made in adult life, those made during early development tend to have permanent effects on the body's structure and function—a phenomenon sometimes referred to as programming.

Fetal nutrition

In common with other living creatures, human beings are 'plastic' in their early life, and are shaped by their environment. Although the growth of a fetus is influenced by its genes, studies in humans and animals suggest that it is limited by the environment, in particular by the nutrients and oxygen the fetus receives from the mother. The fetus responds to undernutrition in a number of ways. It can redistribute its cardiac output to protect key organs, the brain in particular; it can alter its metabolism, for example by switching from glucose to amino acid oxidation; and it can change the production of, or tissue sensitivity to, hormones regulating growth, in which insulin has a central role. Slowing of growth is also adaptive because it reduces the requirement for substrate. Experiments show that even minor modifications to the diets of pregnant animals may be followed by life-long changes in the offspring in ways that can be related to human disease, for example raised blood pressure and altered glucose–insulin metabolism.

Birthweight serves as a marker of fetal nutrition and growth, but it is an imperfect one. The fetus can adapt to undernutrition and continue to grow at the same rate, but with permanently altered physiology and metabolism. Furthermore, the same birthweight may be the outcome of many different paths of growth. Where more detailed measurements of body size at birth are available they can give insights into adaptations that the fetus has made. For example babies that are thin, though within the normal range of birthweight, tend to be insulin resistant as children and adults and are therefore liable to develop NIDDM. It seems that the thin baby responds to undernutrition through endocrine changes.

Ischaemic heart disease

An important clue suggesting that ischaemic heart disease might originate during fetal development came from studies of death rates among babies in Britain during the early 1900s. The usual certified cause of death in new-born babies at that time was low birthweight. Death rates in the new-born differed considerably between one part of the country and another, being highest in some of the northern industrial towns and the poorer rural areas in the north and west. This geographical pattern in death rates was shown to closely resemble today's large variations in death rates from ischaemic heart disease, variations that form one aspect of the continuing inequalities in health in Britain. One possible conclusion suggested by this observation is that low rates of growth before birth are in some way linked to the development of ischaemic heart disease in adult life.

The subsequent studies that confirmed the association between ischaemic heart disease and small size at birth were based on the simple strategy of examining men and women in middle and late life whose body measurements at birth were recorded. In the first study of this kind, 16 000 men and women born in Hertfordshire, United Kingdom, during 1911 to 1930 were traced from birth to the present day. Death rates from ischaemic heart disease fell two-fold between those at the lower and upper ends of the birthweight distribution (Fig. 1). A study in Sheffield, United Kingdom, showed that it was people who were small at birth because they failed to grow, rather than because they were born early, who were at increased risk of the disease. The association between low birthweight and ischaemic heart disease has been confirmed in studies of men in Uppsala (Sweden), Helsinki (Finland), and Caerphilly (Wales), and among women in Helsinki and the United States. Among 80 000 women in the American Nurses Study there was a two-fold fall in the relative risk of non-fatal ischaemic heart disease across the range of birthweight. An association between low birthweight and prevalent ischaemic heart disease has recently been shown in a small study in South India. Among Indian men and women aged 45 years and over the prevalence of the disease fell from 18 per cent in those who weighed 5.5 lb (2.5 kg) at birth to 4 per cent in those who weighed 7 lb (3.2 kg) or more.

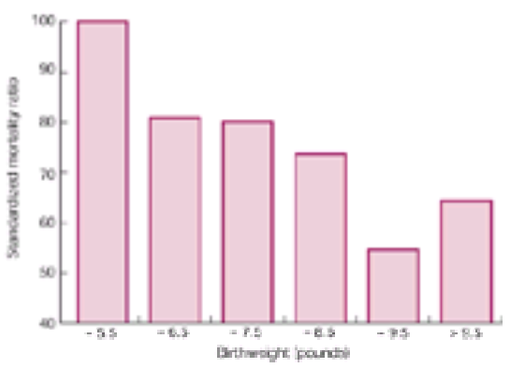

Fig. 1 Death rates from ischaemic heart disease in 15 726 men and women born in Hertfordshire according to their birthweights.

Some of these epidemiological studies included birth length, and other measurements of size at birth, in addition to weight. In Sheffield and in India, rates of ischaemic heart disease were higher in men who had short body length at birth. Thinness at birth, as measured by a low ponderal index (birthweight/length ^{3}), has also been found to be associated with ischaemic heart disease. Among men born in Helsinki, Finland, while low birthweight was associated with raised death rates for ischaemic heart disease, there was a stronger association with thinness at birth, especially in men born at term (Fig. 2). Among women in the same cohort, those who developed ischaemic heart disease also had low birthweight but were short at birth rather than thin. Since the men and women were born to the same group of mothers this difference may reflect intrinsic differences between the sexes in their paths of fetal growth. In the whole cohort, body proportions at birth differed in the sexes: the girls tended to be short while the boys tended to be thin. This may reflect differences in rates of fetal growth at similar levels of maternal nutrition. Female fetuses grow more slowly from an early stage of gestation and are therefore less vulnerable to undernutrition. The lower rates of ischaemic heart disease among women could be related to their slower rates of growth *in utero*.

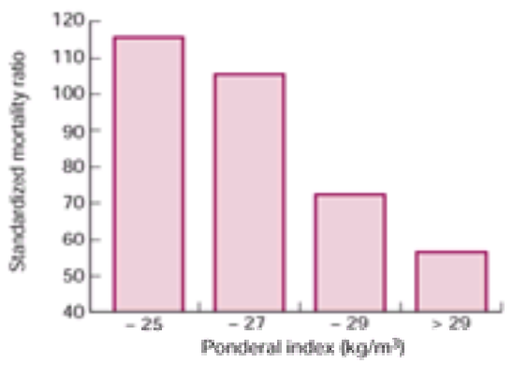

Fig. 2 Death rates from ischaemic heart disease in 3302 Finnish men born at term according to their ponderal indices at birth (birthweight/length ^{3}).

Stroke

The pattern of body proportions at birth which predicts stroke is different to that which predicts ischaemic heart disease. Whereas stroke is similarly associated with low birthweight it is not associated with thinness or shortness. Instead, the studies in Sheffield and Helsinki found increased rates among men who had a low ratio of birthweight to head circumference. One interpretation of this is that normal head growth was sustained at the cost of interrupted growth of the body in late gestation. 'Brain-sparing' patterns of growth can result from diversion of cardiac output to the brain at the expense of the abdominal viscera, importantly the liver. Preliminary evidence suggests that this has lasting effects on liver function including altered regulation of low density lipoprotein cholesterol and raised plasma fibrinogen concentrations, a known risk factor for stroke.

Hypertension

Studies of the mechanisms linking low birthweight with ischaemic heart disease have shown that the progressive fall in disease rates across the range of birthweight (Fig. 1) is paralleled by progressive falls in two of its major biological risk factors—hypertension and NIDDM. Associations between low birthweight and raised systolic and diastolic pressure in childhood and adult life have been extensively documented around the world. Averaged across 69 studies, the difference in systolic pressure associated with a 1-kg difference in birthweight is around 3.5 mmHg. In clinical practice this would be small, but it is a large difference between the mean values of populations. Available data suggest that lowering the mean systolic pressure in a population by 10 mmHg would correspond to a 30 per cent reduction in total attributable mortality. Although in these studies alcohol consumption and higher body mass were also associated with raised blood pressure, the associations between birthweight and blood pressure were independent of them. Nevertheless, body mass remains an important influence on blood pressure and, in humans and animals, the highest blood pressures are found in those who were small at birth but become overweight as adults.

Table 1 shows the systolic pressures of a group of 50-year-old men and women who were born at term in Preston, United Kingdom. The subjects are grouped according to their birthweights and placental weights. Consistent with findings in other studies, systolic pressure fell between subjects with low and high birthweight. In addition, however, there was an increase in blood pressure with increasing placental weight. Subjects with a mean systolic pressure of 150 mmHg or more, a level sometimes used to define hypertension in clinical practice, comprised a group who as babies were small in relation to the size of their placentas. A rise in blood pressure with increasing placental weight has also been found in children in Salisbury, United Kingdom, and Adelaide, Australia, but in studies of children and adults the association between placental enlargement and raised blood pressure or ischaemic heart disease has been inconsistent.

As yet, we know little about the mechanisms which underlie the association between low rates of fetal growth and raised blood pressure. One suggestion is that retarded fetal growth leads to a reduced number of nephrons which in turn leads to increased pressure in the glomerular capillaries and the development of glomerular sclerosis. Another hypothesis which is being actively investigated is that fetal undernutrition leads to life-long changes in the fetus' hypothalamic–pituitary–adrenal axis and these in turn reset homeostatic mechanisms controlling blood pressure. Excessive cortisone production, as occurs in Cushing's syndrome, is associated with raised blood pressure and people who were small at birth have elevated plasma cortisol concentrations within the normal range. A third hypothesis derives from the observation that men and women in Sheffield who were small at birth had reduced elasticity in the large arteries of the trunk and legs, and raised blood pressure. The elasticity of larger arteries depends on elastin, which is laid down *in utero* and during infancy and thereafter turns over slowly: its half-life is 40 years. The amount of elastin laid down *in utero* increases with blood flow. 'Brain-sparing' diversion of blood to the brain could therefore lead to permanent loss of elasticity in the large arteries of the trunk.

Non-insulin dependent diabetes

Both insulin resistance and deficiency in insulin production are thought to be important in the pathogenesis of NIDDM. There is evidence that both may originate during fetal life. Men and women with low birthweight and a low ponderal index have a high prevalence of the 'insulin resistance syndrome', in which impaired glucose tolerance, hypertension, and raised serum triglyceride concentrations occur in the same patient. The patients are insulin resistant and hyperinsulinaemic. A number of studies have shown that people who had low birthweight are already insulin resistant in childhood. A study of men and women who were *in utero* during the war-time famine in Holland provides direct evidence that maternal undernutrition can programme insulin resistance and NIDDM in the offspring. The 'Dutch famine' began abruptly in November 1944 and ended with the liberation of Holland in 1945. The official rations varied between 400 and 800 calories per day. Men and women exposed to the famine *in utero* had higher 2-h plasma glucose concentrations after a standard oral glucose challenge than those born before or conceived after it. They also had higher fasting plasma proinsulin and 2-h plasma insulin concentrations, suggesting insulin resistance.

Figure 3 brings together some of the ideas and findings about the mechanisms through which ischaemic heart disease may be programmed *in utero*. It is a working hypothesis and will need to be re-evaluated as more information becomes available.

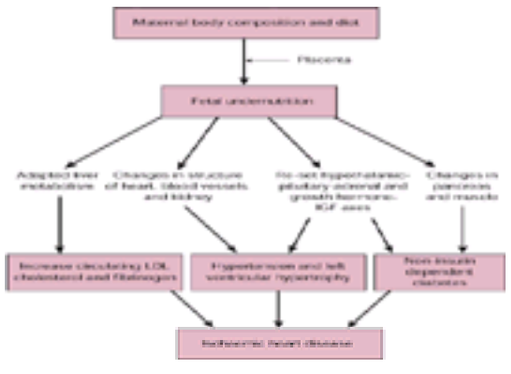

Fig. 3 Framework of possible mechanisms linking fetal undernutrition and ischaemic heart disease.

Ischaemic heart disease and childhood growth

As already described, babies that have low birthweight and are thin or short are at increased risk of ischaemic heart disease in adult life. We are beginning to learn how childhood growth modifies this risk. In the Helsinki study, ischaemic heart disease was commonest among men who were thin at birth, but who 'caught-up' in weight before the age of 7 years and had above average body mass index thereafter. Among women in the same cohort, those who developed ischaemic heart disease were short at birth but had accelerated growth in height in childhood. Table 2 shows that among men with the lowest ponderal indices at birth, but the highest

body mass indices in childhood, the risk of ischaemic heart disease was five times that of men with the highest ponderal indices but lowest body mass indices in childhood. It is not known why accelerated postnatal growth is detrimental. One speculation is based on the observation that restricted fetal growth leads to permanently reduced cell numbers in tissues such as the kidney, in which there is no further cell replication after birth. Accelerated postnatal growth could be deleterious either because overgrowth of a limited cell mass disrupts cell function or because large body size imposes an excessive metabolic demand on a limited cell mass.

Whatever underlies the association between death from ischaemic heart disease and accelerated growth in height and weight in early childhood, imbalances between prenatal and postnatal growth seem to be important in the genesis of adult disease. The effects of adult obesity are a further illustration of this. The highest prevalence of NIDDM is found in people who had low birthweight but become obese as adults. The Dutch famine had its greatest effect on the glucose tolerance of men and women who were overweight as adults.

Maternal nutrition

The nutrition of the fetus depends on the nutrition of the mother. In recent years 'maternal nutrition' has been equated with the diets of pregnant women. This is too limited a definition. Mellanby wrote in 1933 that 'it is certain that the significance of correct nutrition in child-bearing does not begin in pregnancy itself or even in the adult female before pregnancy. It looms large as soon as a female child is born and indeed in its intrauterine life'. Maternal nutrition defined in this way encompasses the nutritional experience of the mother from her own conception, through fetal life, childhood, and into adolescence and adult life. The Helsinki study shows that the mother's body composition before and during pregnancy is an important influence in programming the fetus. Figure 4 shows mortality from ischaemic heart disease according to the men's ponderal indices and their mothers' body mass indices (weight/height ^{2}) in late pregnancy. At any ponderal index, death rates were higher in men whose mothers had a high body mass, so that the highest rates were in men who were thin at birth but whose mothers had a high body mass. The effect of body mass was confined, however, to the offspring of mothers of below average stature (below 1.58 m). The processes by which high body mass in short mothers compounds the increased risk of ischaemic heart disease that is associated with thinness at birth are currently under investigation. The findings already described suggest that raised plasma glucose concentrations in overweight women, which necessarily lead to higher glucose intakes by the fetus, may be one influence.

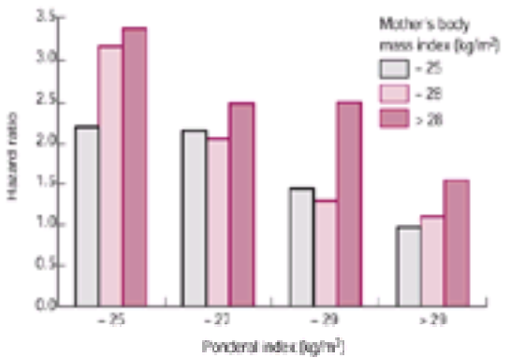

Fig. 4 Hazard ratios for ischaemic heart disease in Finnish men according to their ponderal indices at birth (birthweight/length ^{3}) and their mother's body mass indices (weight/height^2).

There is now a body of evidence suggesting that mothers who are thin also afford an unfavourable environment to their fetuses, leading to insulin resistance and raised blood pressure in the offspring. In the Dutch famine, for example, it was people born to mothers with the lowest weights in pregnancy who had the highest 2-h plasma glucose concentrations. Maternal thinness may have different consequences for the fetus depending on whether it is reflected in a low body mass index, low weight gain in pregnancy, or low skinfold thickness. Mothers' diet in pregnancy has been directly related to cardiovascular risk factors in the offspring during adult life in studies in Aberdeen, Scotland. The blood pressures of men and women were related to the balance of animal protein and carbohydrate in their mothers' diets in late pregnancy, while high intakes of fat and protein were associated with insulin deficiency. The findings of this small study are currently being examined in a larger study in Scotland.

Conclusion

Studies of programming in fetal life and infancy are now established in the agenda for medical research. They have refocused attention on maternal nutrition and fetal growth. The search for the environmental causes of ischaemic heart disease has hitherto been guided by a 'destructive' model. The causes to be identified act in adult life and accelerate destruction processes: the formation of atheroma, rise in blood pressure, and loss of glucose tolerance. There is now a 'developmental' model for the disease. The causes to be identified act on the baby. In responding to them the baby ensures its continued survival and growth at the expense of premature death from ischaemic heart disease.

Further reading

Barker DJP (1998). *Mothers, babies and health in later life*, 2nd edn. Churchill Livingstone, Edinburgh.

Bateson P, Martin P (1999). *Design for a life*. Jonathan Cape, London.

O'Brien PMS, Wheeler T, Barker DJP (1999). *Fetal programming: influence on development and disease in later life*. RCOG Press, London.

15.4.1.2 The epidemiology of ischaemic heart disease

A. R. Ness and G. Davey Smith

Introduction

Ischaemic heart disease (IHD) is defined by a joint International Society and Federation of Cardiology and World Health Organization task force as 'myocardial impairment due to an imbalance between coronary blood flow and myocardial requirements caused by changes in the coronary circulation.' In this chapter we will focus on the epidemiology of the clinical manifestations of IHD. These include angina pectoris, myocardial infarction, and coronary death.

Atherosclerosis is clearly an important underlying pathological process in IHD. Other non-coronary manifestations of atherosclerotic disease include stroke, peripheral vascular disease, and aortic aneurysm. These different conditions share some epidemiological features but show distinct patterns in other respects. For example there is little correlation between IHD and stroke mortality across countries, and within Britain aortic aneurysm mortality correlates negatively with both stroke and IHD mortality. It therefore makes more sense to consider the epidemiology of these conditions separately rather than together.

The process of atheroma deposition and arterial narrowing in the coronary vasculature cannot be observed directly in life without the use of invasive clinical procedures such as coronary angiography. The thickness of carotid arteries measured ultrasonically is currently under evaluation and may prove to be a useful marker of atheroma. Even so, the study of symptomatic disease rather than the underlying process of atheroma deposition may actually be more appropriate. The clinical disease is, after all, what is experienced and may represent the culmination of a number of pathological processes. Indeed, the fact that changes in atherosclerosis at post mortem over time do not mirror changes in clinical IHD rates suggest that it would be unwise to concentrate on the epidemiology of a single pathological process.

IHD is a—or the—leading cause of death in most developed countries. The rates of such deaths in men and women from the populations participating in the World Health Organization MONICA (monitoring trends and determinants in cardiovascular disease) study, which established arrangements to monitor the mortality and incidence of coronary disease (using comparable coding criteria) over a 10-year period in 37 defined populations (in 21 countries), are shown in Fig. 1. Large differences in disease rates between countries, evidence that risk of disease changes on migration, large differences within countries according to socioeconomic position and area of residence, and relatively rapid changes (both increases and decreases in rates of IHD mortality and incidence over time) suggest that the disease is preventable. In this chapter as we describe the epidemiology of IHD we will attempt to relate these findings to the potential for IHD prevention, considering medical therapy only to the extent that it informs our understanding of disease aetiology and prevention.

Fig. 1 Male and female IHD mortality rates in men and women aged 35 to 64 in 35 MONICA populations in the 1980s. Source: Lawlor DA, Ebrahim S, Davey Smith G (2001). *British Medical Journal*, **323,** 541–5.

The burden of ischaemic heart disease

Around one-quarter of all deaths amongst men and one-fifth of all deaths of women in Britain are due to IHD. Among women the proportion is relatively stable throughout the adult years, whilst in men it peaks among 55 to 64-year-olds, for whom IHD accounts for a third of all deaths. Around 6 per cent of 55 to 64-year-old men and 3 per cent of 55 to 64-year-old women report experiencing angina; this increases to 13 per cent and 9 per cent respectively for those aged 75 and over. The National Health Service in England deals with around 200 000 inpatient episodes due to IHD for men and 100 000 for women each year, representing around 5 per cent of all hospital inpatient episodes for men and 3 per cent for women. In addition, there are around 30 million work days lost due to certified incapacity for IHD amongst men and over 4 million amongst women each year in Britain.

A historical perspective

Ischaemic heart disease was until recently viewed largely as a twentieth-century epidemic. In retrospect it seems clear that IHD and IHD deaths occurred before the formal medical description of myocardial infarction in the early years of the twentieth century. William Heberden described the typical symptoms of angina pectoris and the fact that sudden death occurred as a complication in 1768. Later in the eighteenth century, Dr Samuel Black of Newry, County Down, described many cases of angina pectoris and produced a list of factors associated with increased and decreased susceptibility to angina (Table 1). While angina pectoris was occasionally recorded as a cause of death, less than 2 per cent of all deaths attributed to heart disease in Britain fell into this category even by the early part of the twentieth century.

In the middle of the nineteenth century, Richard Quain described a condition he called 'fatty disease of the heart'. A retrospective review of his case series suggests that 52 out of 83 cases probably suffered from IHD. The pathologist Carl Weigert and clinician Carl Huber in Leipzig described the myocardial lesions induced by acute ischaemia and speculated that myocardial infarction and angina both reflected underlying coronary artery disease. Before the turn of the century myocardial infarction and coronary thrombosis were thought of as terminal events. The first diagnosis of acute coronary thrombosis in the living patient is generally attributed to two Russian doctors, Obrastzow and Straschesko, in 1910. By 1918 Herrick was able to link clinical information with electrocardiogram patterns shown to reflect myocardial infarction in experimental canine studies, making the antemortem diagnosis of myocardial infarction easier.

Atherosclerosis has retrospectively been demonstrated in Egyptian mummies, but the first contemporary descriptions occurred in the sixteenth century. As IHD is generally associated with advanced coronary atherosclerosis the dramatic increase in IHD mortality in the first half of this century would be expected to be accompanied by increasing evidence of severe atheroma. However, studies of post-mortem records of the London Hospital suggested that there was no such increase between 1908 and 1949; in fact the degree of coronary atheroma declined. This was interpreted as indicating that an increase in the risk of thrombosis was responsible for the increase in IHD incidence and mortality. Later data comparing the coronary arteries at post mortem of United States soldiers who died in the Korean and Vietnam War show a higher prevalence of atheroma in young men in the early 1950s (77 per cent with atherosclerosis, 15 per cent with clinically significant narrowing of vessels) than the 1960s (45 per cent and 5 per cent, respectively), while IHD mortality in the United States was high and stable over this period. A study in the United States covering the period 1980 to 89—when IHD mortality was declining rapidly—found no reduction in the prevalence of atherosclerosis. Hence the relative contribution of atherosclerosis and thrombotic tendency to trends over time and differences between countries in IHD mortality remain difficult to elucidate.

Trends in ischaemic heart disease in Britain

Despite problems of changing definition, IHD rates in Britain clearly increased from the beginning of the century until the 1980s for men, the pattern of change in women being somewhat different (Fig. 2, and see 'Gender', below). Since the late 1970s, IHD mortality has declined steadily in both men and women.

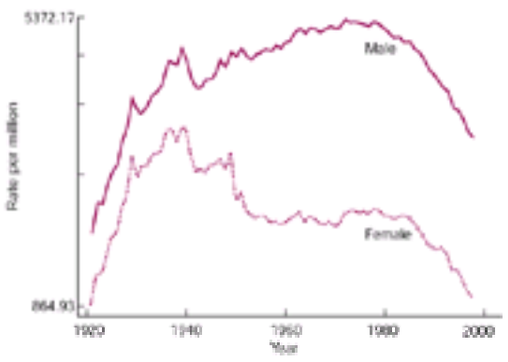

Fig. 2 Mortality from IHD, 1921 to 1998, ages 35 to 74, for England and Wales. Deaths per million population, age standardized to European population. Source: Lawlor DA, *et al*. (2001). *British Medicial Journal* **323**, 541–5.

In men, total circulatory disease mortality increased in the early decades of the century until the 1970s, although not as dramatically as IHD mortality. In women, total circulatory disease has tended to decrease over the century and the rise of IHD has been less consistent (with some decreases in the decades before the mid-century), and much less marked in women than men. Between the 1920s and 1960s the male to female ratio of IHD death rates increased from around 1.5 to around 6 for those under 55; males showed excess mortality at older ages but to a less extreme degree, with the ratio being around 2.5 for 65 to 74 year olds. Sex ratios have remained stable since the 1960s. The rapid rise and fall in rate of death attributed to IHD suggests that these differences were environmental rather than genetic and point to the scope for prevention.

International trends in ischaemic heart disease

Non-socialist, developed countries other than Britain have shown a similar pattern of rise and then fall in IHD mortality. Some, such as the United States and Australia, have experienced an earlier and more rapid fall in mortality than that seen in the United Kingdom. The United Kingdom (and in particular Scotland and Northern Ireland) has one of the highest death rates in the world from IHD, while rates in Japan and the Mediterranean countries of Europe, such as Greece, are low. These international time trends for men are illustrated in Fig. 3.

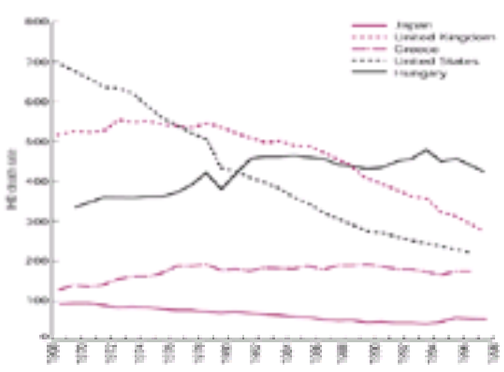

Fig. 3 Age-standardized death rates attributed to ischaemic heart disease from 1968 to 1997 in men aged 35 to 74 in selected countries. Source: Peterson S, Mockford C, Rayner M. *Coronary heart disease statistics 1999 edition.* British Heart Foundation, London, 1999.

More detailed data on recent trends in IHD incidence is available from the WHO MONICA study, which has shown that large differences between rates of death attributed to IHD do exist (Fig. 1), also that the recent trends in mortality are in part attributable to changes in event rates and in part to changes in case fatality. Other studies suggest that the recent falls in mortality attributed to IHD observed in the United States are largely due to reductions in case fatality rather than incidence. This suggests that interventions after myocardial infarction—such as aspirin, thrombolytics, b-blockers, and statin cholesterol lowering drugs—are reducing future mortality in groups with existing IHD.

Migrant studies

Studies of groups that migrate from a low-risk population to a high-risk population can provide valuable insights into the heritability of disease and the key stage of life at which risk is determined. Studies of Japanese migrants to Hawaii and mainland America suggest that as those of Japanese ancestry adopt Western lifestyles their risk of IHD increases. Thus the difference in risk between populations is at least in part environmental, rather than genetic. More puzzling is the fact that migrants to Australia from Greek islands, despite adopting many of the detrimental habits of their host country, are at lower risk of IHD. The acculturation process is clearly complex and it may be that Greek Australians adopt healthier lifestyles in other respects that counterbalance their increased fat consumption and smoking prevalence. Nevertheless, the experience of Greek migrants to Australia further emphasizes the ability of groups of individuals to modify their risk of subsequent IHD.

Overview of risk factors for IHD

There are a number of personal attributes, often described as risk factors, associated with the development of IHD. Some of these—such as age, sex, and family history—are fixed. Others, such as smoking and diet, are modifiable environmental exposures. Some, such as serum cholesterol and blood pressure, though modifiable, are really intermediate processes in the development of IHD and are a product of the interplay between an individual's genetic make-up and environmental exposures. A list of attributes associated with subsequent IHD, by no means exhaustive, is set out in Table 2. This list is not dissimilar to that proposed by Samuel Black over a 100 years earlier (Table 1).

Stamler suggests that the established major risk factors amenable to change are smoking, high blood pressure, high serum cholesterol, and diet. In a large study of middle-aged men in the United States, smokers in the highest quintiles for serum total cholesterol and systolic blood pressure were around 20 times as likely to die from IHD over the next 11.6 years as non-smokers in the lowest quintiles of serum cholesterol and blood pressure (Table 3). However, although these factors are powerful predictors of risk within a population, they only account for 23 per cent of the variance in IHD incidence in men and 14 per cent of the variance in women observed between 25 populations in the WHO MONICA study. Since these established major risk factors were described, many more risk factors have been suggested.

In the following sections we will discuss attributes associated with IHD, starting with fixed and relatively fixed risk markers, then covering the social environment and behavioural factors, finally considering physiological mediating processes.

Age

Age is the strongest risk indicator for IHD incidence and mortality. Compared to men aged 40, 50 year-old-men have five times the risk, 60 year-old-men have 15 times the risk, and 70-year-old men have over 40 times the risk of dying from IHD. A similar steep gradient with age is seen for women. Risk continues to increase into older age groups and recent evidence that the elderly benefit from risk factor control, blood pressure lowering, cholesterol lowering, and probably smoking cessation, combined with their much higher absolute level of risk than younger people, indicates that this is a group for whom sizeable absolute reductions in IHD risk can be achieved and therefore for whom therapies may be particularly cost effective.

Gender

Rates of death attributed to IHD in men are consistently three to four times higher than those in women across countries with differing background levels of disease. The rates of death attributed to IHD in women correlate closely with those in men, suggesting that environmental factors common to both sexes explain the international differences in ischaemic disease rates.

The temporal pattern of mortality attributed to IHD over the last 100 years in Britain has been different in women to that observed in men (Fig. 2). In men, mortality increased from the 1920s until the 1980s, when it began to decline. In women, the mortality rates have always been considerably lower than those observed in men. As with men, mortality in women increased in the early years of this century, but peaked around 1940. Mortality then declined until the 1960s, increased again until the mid-1970s, and has declined since then.

Despite the fact that women have been under represented in epidemiological studies of IHD, it nevertheless appears that classical risk factors such as serum total cholesterol, blood pressure, and smoking (shown to predict coronary disease in men) perform similarly in women. But as the disease is commoner in men the differences in absolute risk are much greater for men than for women.

Men are more likely to smoke than women are, but differences in smoking and other established coronary risk factor levels do not appear to fully explain the observed excess of IHD seen in men. The lower risk of coronary heart disease among women has understandably led to studies of sex hormones as potential protective or risk factors for IHD. There is, however, little empirical evidence to support an important role for sex hormones. It has often been stated that women are only protected against ischaemic heart disease premenopausally, and that their risk progressively increases towards that of men after the menopause. This supposition, which has recently been challenged, provided support for the notion that hormone replacement therapy taken after natural or artificial menopause would reduce the risk of IHD among women. Many observational studies (of women who had elected to take hormone replacement therapy) appeared to support this contention, but the preliminary findings of randomized, controlled trials (where women were allocated at random to hormone replacement therapy or placebo) suggest no such protection is given by hormone replacement therapy.

Further studies should allow more detailed study of the determinants of IHD in women. If these can also shed light on the reasons for the consistent sex difference in coronary disease they may, through uncovering important determinants of IHD risk, benefit both men and women.

Family history and genetic factors

A family history of IHD in first-degree relatives is associated with an increased risk of IHD over and above that produced by a shared environment. Currently identified, major, single gene conditions—for example familial hypercholesterolaemia—are of low prevalence, so that while they are associated with a large increase in relative risk for an individual, they contribute little to the overall population prevalence of IHD. Other polymorphisms with lower relative risks but greater prevalence are under investigation. The marked secular trends in IHD risk, change in risk on migration, and large differences between countries indicate that at the population level genetic factors must act in concert with environmental influences to produce population IHD rates, and that there is thus large scope for prevention. At the individual level, however, a strong family history of premature IHD means that an individual may have more to gain from intensive risk factor control than other members of the population.

Early-life influences

Until relatively recently the majority of epidemiological research on IHD focused on behavioural, physiological, socioeconomic, and psychological risk factors acting in adulthood. However, more recently it was noted that risk factors acting during adulthood could not account for all of the geographical and socioeconomic variation in IHD mortality, leading some to postulate that early-life influences could have a long lasting impact on IHD risk independent of adulthood risk factors. A series of studies have demonstrated that birth weight is inversely related to risk of IHD, suggesting that suboptimal intrauterine environments result in offspring who, many years later, are more likely to succumb to IHD. Furthermore, several conventional IHD risk factors have been shown to be related to birth weight—blood pressure, glucose tolerance, respiratory capacity, and (less consistently) haemostatic factors demonstrate associations with birthweight which would generate the observed inverse birthweight–IHD relationships. However, taking these risk factors into consideration does not appear to fully account for the influence of fetal development on IHD risk, which seems to involve other, as yet unspecified, pathways. It is possible that the birthweight–IHD associations, while they exist, are not causal and therefore not of public health significance. Genetic factors could underlie both low birthweight and increased disease risk in adulthood, and there is evidence that some polymorphisms could indeed act in this way. Whether modifying fetal development will modify later IHD risk is a matter that requires further investigation. These issues are discussed in detail in Chapter 15.4.1.1.

In addition to intrauterine influences, exposures acting in infancy, childhood, and adolescence may be of importance. A frequently replicated and long-standing observation is that taller individuals are at reduced risk of IHD. Growth occurs during infancy and childhood, being influenced by nutrition and infections, as well as genetic factors. Recently it has been shown that stature in childhood is inversely related to later IHD risk, indicating that the association between height and IHD mortality is not due to differential shrinkage occurring in adulthood amongst individuals most prone to IHD. Of the components of stature, leg length—rather than trunk length—is the one of importance, and this is particularly responsive to changes in nutrition and other environmental exposures acting during early childhood. Of the infections acquired in childhood *Helicobacter pylori* infection has been most widely investigated, but the overall evidence does not strongly point in the direction of a direct causal association between this infection and IHD risk. The role of nutrition and other infections acting in childhood requires more study if the intriguing associations between stature and IHD risk are to be understood.

Socio-economic position

In most industrialized countries, IHD risk is higher amongst people living in worse social circumstances. Whether the apparent gradients in the opposite direction seen during the early stages of the IHD epidemic in industrialized countries, and in newly industrialized countries currently, are genuine or due to differential classification of disease remains uncertain. Studies that have taken population samples and standardized measures of IHD prevalence or incidence—rather than simply using certified causes of death—have consistently shown a gradient of increasing risk with decreasing affluence. Social disadvantage can influence IHD risk in a wide

variety of ways, and act across an individual's entire lifetime. People born into worse social circumstances are likely to have lower birthweight, which appears to influence later IHD risk. Socio-economic deprivation in childhood is associated with poor nutrition and growth, and possibly higher incidence of childhood infections, which may increase later IHD risk. Behavioural patterns—in particular with relation to diet, physical activity, alcohol consumption, and smoking—are developed in childhood in a socially-patterned manner and track into adulthood. In most industrialized countries, smoking and dietary patterns associated with increased IHD risk are more common amongst poorer adults. Unemployment and job insecurity in adulthood may also increase IHD risk.

Given the wide array of exposures associated with socio-economic deprivation that could increase the risk IHD, it is not surprising that studies which have taken into account a limited number of these (generally smoking, blood cholesterol, body mass index, alcohol consumption, leisure time physical activity, and blood pressure) find they fail to account fully for the socio-economic distribution of IHD. The large (and in many countries increasing) socio-economic differentials in IHD risk provide both a model for testing aetiological hypotheses and also evidence of an important potential for public health interventions. Figure 4 shows the three-fold difference in IHD mortality between men in unskilled manual occupations in Britain compared to those in managerial and professional occupations, and the graded difference in risk between these occupational extremes. Clearly, reducing IHD risk amongst the less socio-economically advantaged to the same level as that of the most advantaged would have a dramatic influence on overall diseases rates in the population. A further implication is that since socio-economically disadvantaged people have a considerably increased risk of IHD, which is not purely dependent upon conventional risk factors, they have the most to gain from interventions, such as the statin cholesterol-lowering drugs that produce a consistent proportional risk reduction amongst different groups of people treated. A 30 per cent reduction in risk for an unskilled manual labourer in Britain is, in absolute terms, a considerably greater risk reduction than a 30 per cent risk reduction in a manager.

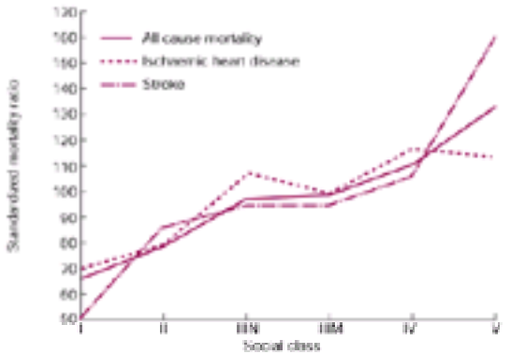

Fig. 4 Social class differences in mortality of men 15 to 64 between 1976 and 1989 from various causes of death. Source: Harding S. Social class differences in mortality of men: recent evidence from the OPCS Longitudinal Study. *Population Trends* 1995; **80**: 31–7.

Diet

The association of diet and IHD can be thought of on three levels: the overall dietary patterns, the foods eaten, and the constituents of the diet. The dietary features thought to be associated with IHD are summarized in Table 4.

Overall dietary patterns

Certain regional dietary patterns are associated with low risk of IHD, for example the Mediterranean diet, which contains more fish, fresh fruit, fresh vegetables, and olive oil than the English diet. One small, randomized trial allocated people with a recent myocardial infarction to receive either advice to eat a Mediterranean diet (more bread, more vegetables, more fruit, more fish, and less meat) and to replace butter and cream with rapeseed margarine, or to usual dietary advice. After 27 months the trial was stopped early because there was a marked reduction in the relative risk of death in those given advice to eat a Mediterranean diet. Of particular interest is the observation that these dietary changes did not alter blood cholesterol levels, implying that the protective effect was not mediated through an effect on cholesterol. The results of this study require confirmation in further trials, and while it is unlikely that the very substantial reduction in IHD risk seen in this trial will be seen in future studies, it is clear that such dietary advice could produce a worthwhile influence on IHD rates.

The dietary patterns in East Asia (the pacific diet) are probably one reason for the low IHD rates there. The broad characteristics of the United States, Japanese, and Mediterranean diets are summarized in Table 5. These so-called traditional diets have changed considerably over recent years so that, for example, the Japanese diet now contains more meat and dairy products and less salted foods. Such changes make it difficult to interpret comparisons of dietary differences and time trends in IHD between countries. Within countries, vegetarians have a lower risk of IHD, but are different in a number of other ways that could influence IHD risk from non-vegetarians.

Foods eaten

Various foods have been linked to risk of IHD. These include alcoholic beverages, fish, fruits, and vegetables. Moderate consumption of alcoholic drinks appears to protect against IHD, with higher rates of IHD in teetotallers and heavy drinkers. Ecological comparisons suggest that wine may exert additional protection, but cohort studies appear to show that the type of beverage is unimportant. Recent studies have looked at the pattern of alcohol consumption as well as the average weekly consumption. People who indulge in sporadic- or binge-drinking appear to be at increased (rather than reduced) risk of IHD.

The low reported rates of IHD among the Inuit (formally known as Eskimos) led to work on the effect of eating fish on IHD risk. Some studies have shown that regular consumers of fish are at reduced IHD risk, while others have not. Two large trials of advice to eat more fish or fish oil capsules in people following a myocardial infarction suggest that intake of dietary fish or fish oil reduces total mortality by around 20 per cent.

The regular consumption of fresh fruit and vegetables is widely believed to be protective against IHD, but the exact size and nature of any beneficial effect is unclear. It is even possible that the observed protective associations represent residual confounding by other lifestyle factors, since people who eat more fruit and vegetables differ in many ways from those who eat less.

Dietary constituents

A number of dietary constituents have been associated with risk of IHD. These include energy intake, intake of various fats, antioxidants, and cereal fibre. The more energy consumed per day the lower the risk of IHD. Though obese people under-report food consumption, reported energy intake also reflects energy expenditure—that is level of physical activity. Hence this finding may reflect higher participation in exercise protecting against IHD.

The Seven Countries Study compared the diets of individuals from populations with different rates of IHD. It showed a strong association between IHD mortality and both total dietary fat and saturated fat consumption. The results of this study are shown in Fig. 5. Studies comparing individuals within cohorts have mostly failed to confirm this relationship, perhaps because of the problem of measuring diet accurately. The relationship between fats and IHD may also be more complex than first thought, with different fatty acids increasing, decreasing or having no effect on disease risk. There is some evidence that high intake of trans fatty acids, produced when oils are solidified by hydrogenation to form margarine, increases risk of IHD.

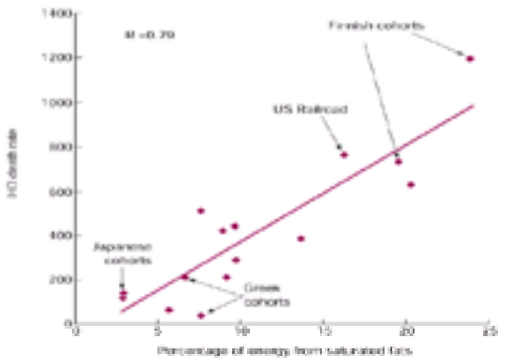

Fig. 5 Ecological comparison between percentage of calories from saturated fat consumed at baseline and IHD mortality over 15 years in 15 cohorts from seven countries recruited between 1958 and 1962. Source: Keys A, Menotti A, Karvonen MJ *et al*. The diet and 15-year death rate in the seven countries study. *American Journal of Epidemiology* 1986; **124**: 903–15

There is laboratory evidence that suggests that oxidation of cholesterol is an important step in atherosclerosis. This laboratory work has generated interest in the association between the intake of dietary antioxidants and rates of IHD. These include minerals such as copper, zinc, manganese, and selenium, vitamins (and provitamins) such as b-carotene, vitamin C, vitamin E, and other chemicals such as flavonoids. Though cohort studies have observed protective associations, the results of randomized trials—where they have been carried out—have not confirmed these observations.

Prospective observational studies of b-carotene intake showed a significantly lower pooled risk of cardiovascular death among those consuming more b-carotene. The results from the randomized trials, however, indicated a moderate adverse effect of b-carotene supplementation (Fig. 6). Various explanations have been put forward to account for these results, including the use of the wrong isomer of b-carotene, the wrong dose, and a detrimental effect of supplementation on levels of other carotenoids. It is more likely that the apparent protective association in observational studies represents confounding, as people who eat diets rich in b-carotene or who take supplements are more socio-economically advantaged and adopt a number of protective health-related behaviours.

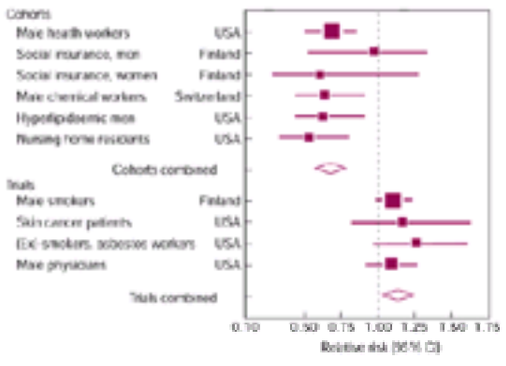

Fig. 6 Meta-analysis of the association between b-carotene intake and cardiovascular mortality: Results from observational studies indicate considerable benefit whereas the findings from randomized, controlled trials show an increase in the risk of death. Source: Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *British Medical Journal* 1998; **316**: 140–144.

Similarly, most observational studies of vitamin E intake and IHD have reported a protective association, but cardiovascular mortality in trials is essentially unchanged in those receiving vitamin E (Fig. 7). In the case of b-carotene it was suggested that inappropriate supplements were used and that the trials were too short. These arguments are difficult to sustain for vitamin E, as the protective observational association was observed in people taking supplements and in people who had only taken supplements for a few years. As with b-carotene, it is more likely that the protective association in observational studies represents confounding, but with vitamin E there is little evidence that supplements are harmful.

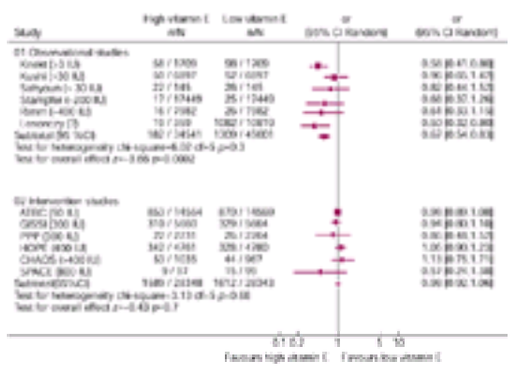

Fig. 7 Meta-analysis of the effect of high versus low vitamin E intake on cardiovascular mortality for observational and intervention studies. Source: Hooper L, Ness AR, Davey-Smith G (2001). *Lancet*, **357**, 1705. *n* = number of deaths; *N* = number at risk.

Cohort studies consistently report a protective association between increased intake of dietary fibre from cereals and reduced rates of IHD. One large trial randomized men with a recent history of myocardial infarction to receive advice to increase their dietary fibre intake (without making other changes to their diets) or no fibre advice. After 2 years there was no survival benefit in those randomized to receive fibre advice, indeed mortality was increased by about 20 per cent, though this increase was not statistically significant.

Summary of effect of diet on ischaemic heart disease

Dietary differences between populations and over time probably explain a considerable proportion of the temporal and geographical variation in IHD. The broad dietary patterns associated with reduced risk of IHD are not in dispute, but the relative importance of specific dietary features is less clear.

Smoking

Smoking was convincingly identified as a cause of lung cancer by German researchers in the 1930s and 1940s and suggestions about a detrimental effect on IHD were also made at this time, although backed by less elegant epidemiological evidence. Since the results of prospective epidemiological studies in the 1950s it has been clear that heavy cigarette smokers have an approximately two-fold elevation in risk of IHD incidence and death, which is reduced by giving up the use of cigarettes, the risk returning to close to that of non-smokers after about 10 years. Cigar smokers who have not previously smoked cigarettes do not have an increased risk of IHD, although amongst those who were previously cigarette smokers an increased risk is seen, presumably because cigarette smokers who switch to cigars continue to inhale some of the smoke.

The mechanism linking cigarette smoking and IHD remains unclear. Initially it was thought that nicotine was the important component of tobacco smoke, but the lack of elevated risk among primary cigar and pipe smokers who absorb nicotine through their buccal membranes suggests this is not the case. In Sweden (among other countries), raw or porous-bagged tobacco ('snuss') is widely used, applied to the buccal membrane this leads to high nicotine absorption but no increased risk of IHD.

Smoking accounts for some of the international variation in IHD rates, but it is noticeable that some countries with a high prevalence of smoking, such as Japan, have low IHD rates. This is borne out by data from the MONICA study that found that smoking alone (when included in models that adjust for cholesterol and blood pressure) accounts for only 5 per cent of the variance in rates of death attributed to IHD across countries. At a cross-national level a relatively high level of saturated fat intake and circulating cholesterol level may be required for smoking to translate into high population rates of IHD. Similarly some populations—for example women of Indian subcontinent family origin living in Britain—have high rates of IHD despite having low rates of smoking.

Obesity

For nearly 100 years insurance data have demonstrated that very obese individuals have an increased risk of death, including death from heart disease. However, studying the association between degree of overweight and IHD risk has produced conflicting findings. In most studies, body mass index (BMI) (weight (in kilograms) divided by height (in metres) squared) is used as the index of adiposity. A U-shaped association between BMI and all-cause mortality has been seen in many studies, with a similar pattern observed for IHD mortality. The reasons for this have been hotly debated, but it seems likely that the high mortality among thin individuals reflects a subset of people who have low BMI because they are ill, or because behavioural patterns—heavy smoking or heavy alcohol consumption—are associated with thinness. The greater the degree to which studies have been able to take this into account, the less evident is the elevated mortality amongst the thin.

Obesity is at least partially related to higher IHD risk because it is in turn associated with higher blood pressure and an unfavourable blood lipid profile (higher low-density lipoprotein (LDL) cholesterol and lower high-density lipoprotein (HDL) cholesterol). Reducing weight is accompanied by reductions in blood pressure and an improvement in blood lipid profile, therefore the risk factor profile associated with obesity should be considered to contribute to the mechanism linking overweight to IHD, rather than being seen as confounding factors. In some early epidemiological studies this issue was confused and investigators statistically adjusted the association between degree of overweight and IHD risk for these physiological measures, which is clearly inappropriate given that they are themselves influenced by the degree of overweight.

Obesity is associated with insulin resistance (which is covered in the next section) and this may also contribute to the elevated IHD risk amongst the obese. However, compared to a simple relative weight measure, such as BMI, the distribution of body fat is more strongly associated with insulin resistance. In epidemiological studies, the ratio of a waist measurement to a hip measurement (waist–hip ratio) has been used as an indicator of central adiposity. People with higher waist–hip ratios are at increased risk of IHD above and beyond the fact that people with high BMI tend to have higher waist–hip ratios; they also demonstrate adverse profiles of factors associated with insulin resistance. Prospective studies are currently examining other measures of adiposity such as impedance (which estimates percentage of body fat) and abdominal height (which estimates the amount of abdominal fat) to see if they are better predictors of IHD risk than BMI and waist–hip ratio.

Conventional hereditability models applied to studies of twins, adoptees, and siblings suggest that obesity is, to a remarkable degree, a genetic characteristic of individuals, with only a small proportion of the population variance being accounted for by environmental factors. This finding probably reflects the limitations of the methods available to study genetic contributions to population levels of risk factors. The substantial changes in the prevalence of obesity that have occurred within populations such as the United States or the United Kingdom over a relatively short period of time—during which there is essentially no genetic change of the population—indicate that environmental factors are of great importance. Low levels of physical activity and a high calorie intake to physical activity ratio are particularly implicated in the rising prevalence of obesity. Though mapping of the human genome may identify those at risk of obesity earlier and more precisely (and offers the possibility in the future of targeted treatments or genetic manipulation), tackling the current obesity epidemic requires measures to reduce population levels of physical inactivity and to encourage energy intake appropriate to activity levels.

Insulin resistance and diabetes

A spectrum of metabolic disorder running from frank diabetes to minor levels of glucose intolerance is of relevance when considering IHD risk. Examining fasting glucose levels or glucose levels after standardized intakes (e.g. in glucose tolerance tests) demonstrates this, with somewhat arbitrary cut-offs being used for 'impaired glucose tolerance (IGT)' and non-insulin dependant diabetes mellitus (NIDDM). This is not the case with insulin-dependant diabetes mellitus (IDDM), where in most instances people either have the disease or do not, but this is a considerably rarer condition that contributes substantially less to the population levels of IHD than NIDDM and IGT.

A cluster of physiological risk factors for IHD have been grouped into the insulin resistance syndrome (also known as the metabolic syndrome or syndrome X). These involve resistance to insulin-stimulated glucose uptake, high circulating levels of insulin, high triglyceride levels, low HDL cholesterol levels, and elevated blood pressure. The degree to which this truly constitutes a syndrome remains disputed, but it is clear that these risk indicators are correlated within populations and contribute to IHD risk.

The epidemiology of diabetes is discussed in detail elsewhere (see Chapter 12.11.1). By far the most important environmental factor influencing diabetes risk and level of glucose intolerance is obesity, in particular central obesity. Physical activity also appears to protect against NIDDM and IGT, while findings have been mixed with respect to alcohol intake and smoking, which have been both positively and negatively associated with disease risk.

Unlike risk factors such as circulating blood cholesterol or blood pressure, fasting glucose does not show a continuous association with IHD risk in prospective epidemiological studies. Only the top 5 or 10 per cent of the population levels of fasting glucose levels are associated with increased risk n studies in the United States and United Kingdom. Thus, the use of fasting glucose as a means of predicting risk in individuals is limited. The clear increased risk of IHD amongst people with frank NIDDM, which seems above and beyond their level of conventional IHD risk factors, indicates that these people have much to benefit from cholesterol and blood pressure lowering if they are eligible for these therapies.

Physical inactivity

In the early days of the coronary heart disease epidemic in industrialized countries it was thought that exercise may predispose to ischaemic heart disease through cardiac strain. However, research starting in the mid-twentieth century suggested that occupational physical activity was associated with reduced risk of IHD. With the increasingly sedentary nature of many occupations this has become a less important risk indicator for IHD and attention has shifted to physical activity in leisure time.

Clearly, health-related selection is a serious problem in studying the relationship between leisure time physical activity and ischaemic heart disease, since those with early signs of the disease will be less physically active. There is also considerable confounding as engagement in leisure time physical activity is associated with other important risk factors for IHD, such as smoking and socioeconomic position. However, the general picture suggests that increased physical activity does protect against IHD, although there is uncertainty about the type, intensity, and frequency of exercise that is required to confer such protection.

The key question has been whether the physical activity needs to be vigorous or not, with some studies suggesting that only vigorous exercise is protective, while others indicate that any form of increased physical activity is effective. This may reflect difficulties in measuring usual physical activity accurately, particularly at low levels. Differences in the findings between studies may also reflect the fact that the type of exercise that is vigorous for one group may not be vigorous for another. For example the elderly or unfit will find relatively low intensity physical activity results in increased cardiorespiratory fitness, while lower intensity activity will have no such training effect in fit young adults. A formulation that is consistent with the current data is that physical activity which is at a sufficient level to produce cardiorespiratory training reduces the risk of IHD, while the value of lower intensity physical activity is unclear.

Stress

Stress is widely considered by the general public to be an important cause of IHD. Indeed, several surveys of lay beliefs have shown that it is one of the most widely recognized risk factors for the disease. Stress is, however, difficult to define or measure. Epidemiological studies have examined a wide range of potential exposures that fall under the general heading of 'stress'. These include measures which are essentially of personality traits—such as the well-known Type A behaviour pattern—through to records of stressful life events or global measures of perceived stressfulness of daily activities. A broader conceptual category of 'psychosocial factors' is now widely used in the epidemiological literature. This includes aspects of social life such as the strength of social support networks or the level of control

that people have over their work.

The type A behaviour pattern was investigated from the 1950s, particularly in the United States, as a potential cause of IHD. Type A individuals—those who are involved in an incessant struggle to achieve more and more in less and less time—were found to have higher risk of IHD in several early prospective studies, and type A personality was included in official publications (produced by august bodies such as the American Heart Association) as a cause of IHD, along with smoking, high blood pressure, and high saturated fat diet. More recent studies have failed to find any association between type A behaviour and IHD. It may be that this association no longer exists because IHD is not now considered a disease of the wealthy, stressed, business man, who in the past was more likely to be diagnosed with the condition. Research on behavioural traits now focuses on those elements of the original type-A classification that are related to adverse social background, for example hostility: these may predict IHD risk because of this, rather than because of any causal link with the disease.

Measures of self-reported global stress and other self-reported indicators, such as low control at work or poor social networks, tend to be related to self-reported IHD symptoms (usually angina or severe chest pain indicative of a myocardial infarction). They less consistently relate to objective measures of IHD, such as ECG changes or IHD mortality. This may reflect an underlying reporting tendency, such that people who report higher levels of adversity in their lives also report higher levels of symptoms. Given the history of the association between type A behaviour and IHD it is also important to consider whether associations between stress and IHD are generated by confounding. People with low control over their work (e.g. shift workers in a factory) are, almost by definition, in less favourable social locations than those with higher control over their work (e.g. managers or senior academics). It is not surprising that the former have higher rates of IHD, given the strong social patterning of the disease. Whether these adverse psychosocial factors are one of the causes of the high IHD mortality in the less advantaged or merely another consequence of material inequalities is not clear.

More methodologically robust studies employing objective measures of both tress and IHD outcomes are required. In particular, it is remarkable that the supposed biological mediators between stress and disease (disturbed functioning of the hypothalamo–pituitary–adrenal axis and possible immune system outcomes of this) have not themselves been related to IHD risk in prospective observational studies, some 70 years after the basic concept was advanced by Hans Selye. The current controversy—with enthusiastic proponents of stress as a cause of IHD on one side (feeding into a popular propensity for accepting this view) lined up against the majority of academic and research cardiologists and epidemiologists who dismiss the association as spurious—will only be resolved when studies employing better methods report their findings.

Cholesterol

Circulating blood cholesterol is strongly and positively associated with ischaemic heart disease risk in men and women, both young and old. A series of large-scale, randomized, controlled trials of blood cholesterol reduction demonstrate that this process is reversible and risk reductions of the magnitude predicted from the observational data can be produced through lowering blood cholesterol.

One issue not directly relevant to ischaemic heart disease that led to controversy regarding cholesterol reduction was the suspicion that low circulating blood cholesterol caused an increased risk of morbidity and mortality from non-coronary causes, including cancer, psychiatric disease, and gastrointestinal and respiratory disease. Observational studies tended to report inverse associations between blood cholesterol and these conditions, though it was clear that these associations could be generated through early stages of ill-health or adverse health-related behaviours leading to lower circulating cholesterol levels. Findings from randomized, controlled trials of cholesterol reduction were initially ambiguous, in that there was evidence of elevation of non-coronary morbidity and mortality in some studies. This now appears to reflect specific adverse effects of certain cholesterol lowering drugs, in particular the fibrates. More recent studies in which circulating cholesterol levels were reduced more profoundly than in earlier studies, through the use of statins, suggest there are no detrimental effects of cholesterol lowering itself.

Early epidemiological studies only measured total circulating cholesterol, but it is evident that subfractions of cholesterol have differential effects on ischaemic heart disease risk. The adverse effects are restricted to the low-density lipoprotein cholesterol fraction, with high-density lipoprotein cholesterol (HDL) levels being inversely associated with ischaemic heart disease risk. Alcohol consumption increases HDL cholesterol and increased HDL levels are a potential mediator of the apparent protective effect of low to moderate alcohol consumption on IHD. Trials of raising HDL cholesterol are difficult to interpret as the agents employed also decrease trygliceride levels and, to an extent, LDL cholesterol. Current evidence suggests that there may be a protective effect of raising HDL cholesterol, but this requires confirmation.

Blood pressure

There is a strong, consistent dose–response relationship between increased casual blood pressure measured in middle age and increased risk of IHD. In observational studies a 10 mmHg increase in diastolic blood pressure is associated with a 37 per cent increase in the risk of coronary disease. Large, randomized trials of pharmacological blood pressure reduction in middle age have confirmed that blood pressure reduction reduces subsequent risk of coronary disease. Several observational studies have also shown that increased blood pressure in early adult life is associated with increased coronary mortality in later life, suggesting that risk trajectory may be set early.

Blood pressure is continuously distributed in populations and increases in blood pressure across the range of blood pressure measures are associated with increased risk of IHD. There is therefore no natural dichotomy between normotensives and hypertensives. Hypertension has to be defined by weighing up the risks of disease at a given level of blood pressure against the risks of treatment to lower blood pressure.

Blood pressure is determined both by environmental factors, such as increased sodium intake (increases blood pressure), increased alcohol intake (increases blood pressure), increased fruit and vegetable intake (lowers blood pressure), and genetic factors. While individual blood pressure response to environmental factors may be influenced by genetic factors, migrant studies illustrate the substantial and relatively rapidly acting effects of environmental factors. In one study of migrants from a rural community in western Kenya to urban Nairobi, mean diastolic blood increased by around 6 mmHg from that measured 10 months previously.

Not all environmental factors that increase blood pressure increase risk of coronary disease. For example increased alcohol intake results in increased blood pressure and people who drink heavily are at increased risk of coronary disease, but those that consume alcohol in moderation have lower risks of coronary disease than teetotallers. Though there are a number of explanations for this particular protective association, this example illustrates the more general point that effects on risk factors may not necessarily translate into changes in risk of IHD.

Haemostatic factors

Studies of changes in prevalence of atherosclerosis (discussed earlier) provide indirect evidence that another pathological process, such as thrombosis, influenced trends in symptomatic disease over the last century. Over the last 20 years a number of cohort studies have examined the association between haemostatic factors and coronary disease. There is a consistent, independent association between increased fibrinogen levels and increased risk of coronary disease. The subsequent risk of IHD comparing those in the highest third with those in the lowest third of the fibrinogen distribution at baseline is increased by around 80 per cent. Some prospective studies have reported increased coronary risk with factors VII and VIII but the data are less extensive and less consistent. Other studies have reported associations with tissue-type plasminogen activator (t-PA) and with plasminogen activator inhibitor-1 (PAI-1), while platelet function tests do not appear to predict subsequent coronary risk. These associations need to be confirmed. Smokers have higher fibrinogen levels than non-smokers and this may in part explain their excess coronary risk. Fibrinogen levels are also associated with higher levels of other acute phase reactants suggesting that chronic inflammatory or infective processes may increase fibrinogen levels and thus risk of symptomatic coronary disease. Alternatively, fibrinogen may merely be a marker for these underlying processes. These issues are discussed elsewhere.

The prevention of ischaemic heart disease

As we have seen, there are marked differences in the rate of death attributed to IHD over time, between countries, between regions, within countries, and between groups of individuals within countries. These differences point to the potential scope for prevention.

Attempts at prevention can be roughly classified as primary, secondary, or tertiary. Tertiary prevention is the treatment and rehabilitation of people with symptomatic disease: it clearly offers important opportunities for preventing further symptomatic episodes since many who experience an acute myocardial infarction will have known coronary disease. The recent results from the WHO MONICA study would seem to confirm that the widespread use of effective treatments can reduce the number of symptomatic episodes and deaths attributed to IHD. Secondary prevention is the identification and treatment of early, often asymptomatic, disease: this is

not currently (and may never be) a discrete option, because most people in Western populations have at least some atherosclerotic disease in their coronary arteries and there are no proven non-invasive tests that are able to detect reliably those with early disease before they develop symptoms. Primary prevention is the prevention of the onset of disease. In IHD the distinction between primary, secondary, and tertiary prevention is blurred: people identified as being at high risk of disease on the basis of risk factor profiles will comprise those with symptomatic disease, those with asymptomatic disease, and those without manifest disease.

There are two differing but complementary approaches to primary prevention: the high-risk approach and the population approach. The high-risk approach seeks to identify those at highest risk of developing disease with time to intervene to prevent disease. This approach has the advantage that it may be easier to encourage people at high risk to alter their lifestyles and to take tablets as these individuals have a considerable amount to gain personally from accepting change. Such an approach, however, requires that those at high risk be identified and ignores the many people who develop symptomatic disease who are at moderate rather than high risk. For example, in the Whitehall study, 42 per cent of the coronary deaths (and 36 per cent of all deaths) occurring over a 15-year follow-up occurred in the 20 per cent at highest risk (on the basis of smoking status, blood pressure, and plasma cholesterol). The majority of coronary deaths (58 per cent), however, occurred in those who were not at high risk (Table 6). Thus, even a universally effective package of interventions for people at high risk could only hope to prevent around 40 per cent of coronary deaths, and a (more plausible) package of interventions that halved coronary mortality in those at high risk could only hope to prevent around 20 per cent of coronary deaths. The alternative approach is the population approach, which seeks to alter risk-factor levels or behaviour across the whole population, its advantage being that for fairly modest changes in mean population risk-factor levels it is likely to produce larger overall improvements in the health of the population as a whole than the high-risk approach. The disadvantage is that the absolute benefit for each individual who makes a lifestyle change is likely to be small. This mismatch between the likely size of the benefits to the individual and population has been called the 'prevention paradox'.

In terms of IHD prevention, smoking cessation strategies are potentially of high impact. It has been shown that brief advice from physicians leads to small, but (in population terms) worthwhile reductions in smoking rates among their clients. Nicotine replacement gum can also aid smoking cessation. More problematic are attempts to reduce the initiation of smoking in adolescents. School-based antismoking programmes have shown disappointing results, which is perhaps not surprising given the nature of experimental behaviour among this age group. While smoking is an option it is likely that a high percentage of the population will experiment with the behaviour and strategies for early cessation—which also occurs amongst a high proportion of those who try the behaviour—should be built upon.

The detection and treatment of high blood pressure reduces the risk of IHD in the primary prevention setting and has the added benefit of also substantially reducing the risk of stroke. Similarly, lowering cholesterol levels with the statin drugs reduces IHD risk in primary prevention, although the absolute benefit is small for those who are not at increased risk of IHD for other reasons.

A series of multiple risk factor intervention trials have been carried out at both the community and individual participant level in which encouragement to modify diet, reduce smoking, increase physical activity, and (in some studies) increase medical treatment of elevated risk factors have been explored. Beyond the evident benefits of pharmacological reduction of blood pressure, the findings have been disappointing.

These results emphasize the degree to which health-related behaviours—such as dietary consumption, smoking, heavy drinking, and physical activity patterns—are strongly influenced by societal legislative and fiscal forces and are less to do with individual levels of knowledge and degree of willpower. This is clearly the pattern for smoking, where smoking levels in a country, smoking initiation rates, and the social distribution of smoking are all responsive to the profit-making incentives offered to tobacco companies. The diversification of tobacco companies to other products (within the 'home' advanced capitalist markets) and shifts to export to newly industrializing countries reflects the unfavourable economic environment that has been created for the companies in some countries. The prevention of smoking—one of the most important modifiable health risks globally—can only be ultimately successful if the world community accepts the need to restrict the profitability of growing and selling tobacco.

The differences between countries and over time in both behaviour and IHD disease rates clearly illustrate the capacity for populations to alter their behaviour and to adopt healthier lifestyles. While simple invocations to change have modest effects on behaviour and IHD risk, social change can result in profound improvements in health. Conversely, where cultures fail to adapt or social structures demise (as happened in Russia with the collapse of the Soviet Union) disease rates can increase. Effecting societal change is complex, but the potential public health benefits of even modest population change are profound. The concentration of IHD (and many other diseases) in those in society who are least advantaged suggest that policies will be more effective if they seek to improve the lot of the poor by improving their income, opportunities, and self-esteem. More inclusive and extensive programmes that seek to reduce inequalities in health and improve health for all offer real opportunities to reduce IHD incidence and mortality.

Summary

IHD is a globally important cause of morbidity and mortality. Epidemiological studies provide evidence that IHD is preventable and give some estimate of how much IHD might indeed be prevented. Current favourable trends in mortality in developed countries may not be maintained in the face of the current epidemic of obesity. Equally, the reductions in IHD case fatality (leading to reductions in mortality) achieved through use of effective treatments may not be viable solutions in less wealthy countries. While there are genuine areas of uncertainty that require further research, much is known about the aetiology and prevention of IHD. Invocations to individuals to adopt healthier lifestyles have produced disappointing results. At the population level, however, the will and capacity to modify behaviour exists. If coronary care units are to meet the same fate as sanatoriums, major societal and structural changes will be required to improve the material conditions of the least advantaged throughout their lives and to shape the diets, activity levels, alcohol consumption patterns, and smoking behaviour of us all.

Further reading

Anonymous. NIH Consensus development panel on physical activity and cardiovascular health (1996). *Journal of the American Medical Association* **276**, 241–6.

Barker DJP (1998). *Mothers, babies, and health in later life*. Churchill Livingstone, Edinburgh.

Burr ML, Fehily AM, Gilbert JF, *et al* (1989). Effects of changes in fat, fish, and fibre intakes on death and myocardial reinfarction: diet and reinfarction trial (DART). *Lancet* **ii**, 757–61.

Charlton J, Murphy ME, Khaw KT, Ebrahim SB, Davey Smith G (1997). Cardiovascular diseases. In: Charlton J, Murphy ME, eds. *The health of adult britain 1841–1994*, Vol. 2, pp. 60–75. Stationery Office, London.

Collins R, Peto R, Macmahon S, *et al* (1990). Blood pressure, stroke, and coronary heart disease. Part 2, short term-reductions in blood pressure: overview of randomised drug trials in their epidemiological context. *Lancet* **335**, 827–38.

Davey Smith G (1997). Down at heart: the meaning and implications of social inequalities in cardiovascular disease. *Journal of the Royal College of Physicians* **31**, 414–24.

de Lorgeril M, Renaud S, Mamelle N, *et al* (1994). Mediterranean alpha-linolenic acid-rich diet in secondary prevention of coronary heart disease. *Lancet* **343**, 1454–9.

Ebrahim S, Davey Smith G (1998). Health promotion for coronary heart disease: past, present and future. *European Heart Journal* **19**, 1751–7.

Ebrahim S, Davey Smith G, McCabe C, *et al* (1999). What role for statins? *Health Technology Assessment* **3**, 1–91.

Hart CL, Davey Smith G, Hole D, Hawthorne VM (1999). Alcohol consumption and mortality from all causes, coronary heart disease, and stroke: results from a prospective cohort study of Scottish men with 21 years of follow up. *British Medical Journal* **318**, 1725–9.

Hulley S, Grady D, Bush T, *et al* (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Journal of the American Medical Association* **280**, 605–13.

Isles CG, Hole DJ, Hawthorne VM, Lever AF (1992). Relation between coronary risk and coronary mortality in women of the Renfrew and Paisley survey: comparison with men. *Lancet* **339**, 702–6.

Jarrett RJ (1996). The cardiovascular risk associated with impaired glucose tolerance. *Diabetic Medicine* **13**, S15–S19.

Kuh D, Ben Shlomo Y (1997). *A lifecourse approach to chronic disease epidemiology*. Oxford University Press, Oxford.

Kuulasmaa K, Tunstall-Pedoe H, Dobson A, *et al* (2000). Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA project populations. *Lancet* **355**, 675–87.

Labarthe DR (1998). *Epidemiology and prevention of cardiovascular disease: a global perspective*. Aspen, Gaithersburg, Maryland.

Macmahon S, Peto R, Cutler J, *et al* (1990). Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* **335**, 765–74.

McKeigue PM, Marmot MG, Adelstein AM, *et al* (1985). Diet and risk factors for coronary heart disease in Asians in Northwest London. *Lancet* **ii**, 1086–90.

Meade TW, Mellows S, Brozovic M, *et al* (1986). Haemostatic function and ischaemic heart disease: principal results of the Northwick Park Heart Study. *Lancet* **ii**, 533–7.

Ness AR, Powles JW (1997). Fruit and vegetables and cardiovascular disease: a review. *International Journal of Epidemiology* **26**, 1–13.

Powles JW (1994). Greek migrants in Australia: Surviving well and helping their hosts. In: Marks L, Worboys M, eds. *Migrants, minorities and medicine: historical and contemporary perspectives*. Routledge, London.

Rimm EB, Klatsky A, Grobbee D, Stampfer MJ (1996). Review of moderate alcohol consumption and reduced risk of coronary heart disease: is the effect due to beer, wine, or spirits? *British Medical Journal* **312**, 731–6.

Rose G (1992). *The strategy of preventive medicine*. Oxford University Press, Oxford.

Steinberg D, Parthasarathy S, Carew TE, Khoo JC, Witztum JL (1989). Beyond cholesterol. Modifications of low-density lipoprotein that increase its atherogenicity. *New England Journal of Mediciane* **320**, 915–24.

Tunstall-Pedoe H, Kuulasmaa K, Mähönen M, *et al* (1999). Contribution of trends in survival and coronary-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA project populations. *Lancet* **353**, 1547–57.

Tunstall-Pedoe H, Vanuzzo D, Hobbs M, *et al* (2000). Estimation of contribution of changes in coronary care to improving survival, event rates, and coronary heart disease mortality across the WHO MONICA project populations. *Lancet* **355**, 688–700.

Ulbricht TLV and Southgate DAT (1991). Coronary heart disease: seven dietary factors. *Lancet* **338**, 985–93.

Willett WC (1994). Diet and health: What should we eat ? *Science* **264**, 532–7.

World Health Organization MONICA Project (1994). Ecological analysis of the association between mortality and major risk factors of cardiovascular disease. *International Journal of Epidemiology* **23**, 505–16.

15.4.2.1 The pathophysiology of acute coronary syndromes

Peter L. Weissberg

Introduction

An acute coronary syndrome arises when there is sudden total or partial occlusion of a coronary artery leading to myocardial ischaemia and its consequences. The syndrome therefore encompasses the clinical entities of unstable angina, acute myocardial infarction, and many cases of sudden cardiac death. In the vast majority of instances, an acute coronary syndrome is due to the thrombotic consequences of rupture or erosion of an atherosclerotic plaque. To understand the pathophysiology of acute coronary syndromes it is therefore important to understand the cellular and molecular events leading to the development and progression of an atherosclerotic plaque.

Atherosclerosis: stable and unstable plaques

Atherosclerosis

Endothelial function is crucially important in the development of atherosclerosis. Normal endothelial cells form a physical barrier between the thrombogenic matrix of the underlying vessel wall and the circulation. However, by producing a variety of antithrombotic and anti-inflammatory molecules the endothelium also protects against the development of atherosclerosis. In particular nitric oxide, which is constitutively produced from arginine by the action of endothelial nitric oxide synthase in normal endothelial cells, prevents accumulation of inflammatory cells and, along with prostacyclin, prevents activation and adhesion of platelets. Patients with established atherosclerosis have abnormal endothelial function, particularly if they smoke, as do apparently healthy individuals with high cholesterol levels.

Endothelial dysfunction is manifest in the coronary and peripheral circulations as reduced vasodilation in response to infused acetylcholine, a pharmacological stimulator of nitric oxide production, and in the brachial artery as reduced flow-mediated vasodilation in response to hyperaemic forearm blood flow. However, it remains unclear how these abnormalities of endothelial function relate to the development of atherosclerosis. Indeed, whilst it is accepted that the endothelium is abnormal in atherosclerosis, it is still unclear whether a primary endothelial abnormality predisposes to lipid accumulation and therefore the development of atherosclerosis, or whether endothelial dysfunction is secondary to hyperlipidaemia and/or the presence of subclinical atherosclerosis in apparently healthy subjects. The fact that the endothelium can respond to changes in shear stress by increasing or decreasing production of a number of molecules known to be involved in the atherogenic process suggests that subtle perturbations in endothelial function induced by local haemodynamic factors may contribute to the tendency of atherosclerosis to develop only at particular sites within the arterial tree.

The pathogenesis of the atherosclerotic plaque is discussed in detail in Chapter 15.1.2.1, but in brief, in the earliest atherosclerotic lesion, the fatty streak, there is subendothelial accumulation of oxidized lipid. This is associated with activation of the overlying endothelium and recruitment of inflammatory cells, predominantly monocytes and some T cells, into the subendothelial space. Once in the subendothelial space, the monocytes mature into macrophages and express a variety of surface molecules, in particular the scavenger receptors that allow them to bind and ingest lipid to become macrophage foam cells, the most abundant cell in the core of the atherosclerotic plaque. There is good evidence that oxidation of lipids is an essential step in the formation of foam cells and a mature atherosclerotic lesion. T cells are also activated and there is expression of major histocompatibility complex (MHC) molecules in surrounding vascular smooth muscle cells, some of which may also take up lipids to become foam cells. The activated inflammatory cells within the plaque produce a variety of cytokines that serve to recruit further inflammatory cells into the lesion. A crucial aspect is that some of these molecules also induce migration of vascular smooth muscle cells from the vessel media into the intima where they become incorporated into the atherosclerotic lesion. During the process of migration the vascular smooth muscle cells change from a contractile to a repair phenotype, as discussed in Chapter 15.1.1.3, which allows them to proliferate and elaborate the matrix proteins required to form a fibrous cap over the lipid core. Vascular smooth muscle cells are the only cells capable of synthesizing the fibrous cap, and their participation is therefore essential for plaque stability. Stable atherosclerotic plaques characteristically contain few inflammatory cells, large numbers of vascular smooth muscle cells, and have a thick fibrous cap that is resistant to rupture. They only cause symptoms if they are large enough to compromise flow through the artery, in which case they cause reversible ischaemia in the form of stable angina.

Plaque instability

Atherosclerotic plaques give rise to acute coronary syndromes when there is a sudden reduction in coronary blood flow. In most cases this is due to aggregation of platelets, with or without subsequent thrombosis. Spontaneous haemorrhage, presumably from immature microvessels within the plaque, causes rapid expansion of the plaque and a sufficient reduction in coronary blood flow to precipitate either myocardial infarction or unstable angina in a few cases. Most thrombotic events are due to rupture or fissuring of the fibrous cap with consequent exposure of the thrombogenic lipid core, but in some cases, variably reported to be between 25 and 44 per cent of thrombotic coronary events, thrombosis occurs because of accumulation of platelets at the site of endothelial erosion without obvious disruption of the fibrous cap or exposure of the lipid core.

Thrombosis due to endothelial erosion appears to be particularly common in female smokers dying suddenly with coronary disease, which may reflect a greater tendency to thrombosis in women than men rather than a difference in underlying plaque pathology. The mechanism of plaque erosion still remains to be determined, but it is possible that plaque inflammatory cells may produce cytokines that are toxic to the overlying endothelium, thereby inducing endothelial cell death and exposure of the underlying collagenous matrix of the atherosclerotic lesion. Alternatively, there may be a detrimental interaction between endothelial cells and underlying smooth muscle cells. Further studies are required to resolve the mechanism of endothelial erosion before therapies aimed at its prevention can be developed.

Plaque rupture, with the development of a thrombus that occludes the lumen and extends into the core of the lesion, is the commonest substrate for an acute coronary syndrome. In contrast to endothelial erosion, the pathophysiological mechanisms underlying plaque rupture are now beginning to be resolved. Modelling of the structural characteristics of atherosclerotic plaques has predicted that plaques with a large lipid core and a thin fibrous cap are subject to increased circumferential tensile stress, increasing the chance of rupture, whereas a thick fibrous cap confers structural stability by reducing tensile stress. The interaction between the physical properties of the plaque and local haemodynamic forces are therefore likely to play a part in determining stability and resistance to rupture, particularly in circumstances where the clinical event is related to physical activity, there being a well-recognized association between activities such as shovelling snow and the development of an acute coronary syndrome.

However, the most important determinant of plaque instability is the balance between the activity of inflammatory cells and the healing, fibrotic reaction of the smooth muscle cells in the fibrous cap. Plaque rupture occurs in lesions containing few vascular smooth muscle cells and abundant inflammatory cells, suggesting that inflammatory cells are responsible for the breakdown of the plaque: there are a number of ways in which they might weaken the fibrous cap. Firstly, they produce proinflammatory cytokines that inhibit vascular smooth muscle cell proliferation and matrix production. Secondly, inflammatory cytokines such as interleukin 1b, tumour necrosis factor-a, and interferon-g act synergistically to induce vascular smooth muscle cell death. Thirdly, activated macrophages can induce vascular

smooth muscle cell death by direct cell to cell contact. Fourthly, and probably most importantly, inflammatory cells, particularly macrophages, produce and activate a number of matrix metalloproteinases that digest the matrix of the fibrous cap. Thus inflammatory cells exert a potent negative influence on the turnover of matrix protein within the lesion. Furthermore, as discussed in Chapter 15.1.1.3, vascular smooth muscle cells in the fibrous cap become senescent and develop an inherent tendency to undergo apoptosis (programmed cell death). The overall effect is that inflammatory cell activity destroys the fabric of the fibrous cap and at the same time reduces the number and synthetic activity of intimal vascular smooth muscle cells, leading inevitably to weakening of the cap and eventual rupture under the stress of local haemodynamic forces (Fig. 1). The clinical importance of individual inflammatory cytokines and matrix metalloproteinases in the progression and rupture of plaque has yet to be established. However, evidence is beginning to emerge that subtle, genetically determined differences in production and activity of inflammatory mediators and matrix metalloproteinases, measured as polymorphisms in the genes coding for their production, may influence the rate of progression and outcome of atherosclerosis in different individuals.

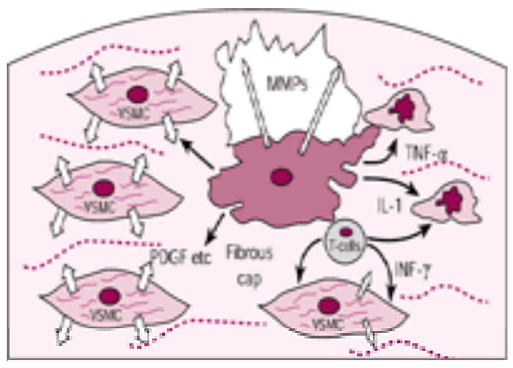

Fig. 1 Cellular interactions within the fibrous cap of an established atherosclerotic lesion leading to plaque rupture and an acute coronary syndrome. In early lesions cytokines produced by macrophages recruit vascular smooth muscle cells that produce the matrix proteins of the fibrous cap. In advanced lesions the cytokines produced by inflammatory cells inhibit vascular smooth muscle cell protein production and are cytotoxic. The vascular smooth muscle cells become senescent and die and macrophages produce matrix metalloproteinases that destroy the fibrous cap, leading to plaque rupture.

Modification of plaque stability

For many years symptomatic atherosclerosis was thought to represent the end stage of a slowly progressive, irreversible disease process that had developed over decades and which was therefore unlikely to be influenced favourably by medical therapy. Consequently, therapy for ischaemic heart disease focused either on abrogating the consequences of atherosclerosis, for example with antianginal, antiplatelet, and thrombolytic drugs, or relieving stenoses by interventions such as angioplasty or bypass surgery. However, as described above, it is now recognized that atherosclerosis is a dynamic inflammatory condition involving constant or cyclical recruitment and activation of inflammatory cells, with repeated subclinical episodes of plaque rupture and repair leading to episodic growth of individual plaques (Fig. 2). Encouragingly, results of several recently published large-scale trials of lipid-lowering agents, in particular the HMG CoA reductase inhibitors or 'statins', have shown that medical therapy can modify the pathophysiology of atherosclerosis.

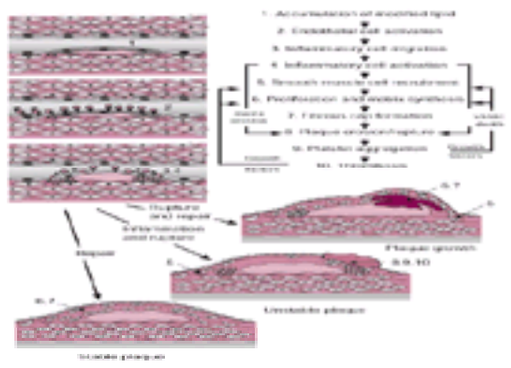

Fig. 2 Cellular interactions leading to the development and progression of atherosclerosis. (Reproduced from Weissberg PL (2000). *Heart* **83**, 247–52, with permission.)

Aggressive, sustained lipid lowering, by whatever means, reduces the risk of myocardial infarction. However, it is only since the introduction of statins that doctors have had access to powerful lipid-lowering drugs that are well tolerated and relatively free of toxicity. Angiographic studies in patients with angina have shown that statins reduce the rate of appearance of new lesions and reduce the incidence of spontaneous, clinically silent, vessel occlusions (occurring when there is plaque rupture and thrombosis in the context of a good collateral circulation that maintains viability of the tissue downstream of the occlusion). However, statin therapy has only a marginal and probably haemodynamically insignificant effect on the size of established stenotic atherosclerotic lesions. Despite this disappointing effect on plaque size, treatment with statins reduces mortality and the risk of a coronary event (unstable angina, myocardial infarction, or the need for surgical intervention) by about 30 to 40 per cent, both in asymptomatic individuals at high risk of a coronary event (primary prevention) and in patients who have already experienced a coronary event (secondary prevention). Results of theses studies suggest strongly that statins are stabilizing atherosclerotic plaques.

The inflammatory basis of atherosclerosis

The cellular and molecular interactions described above all point to an important role for inflammation in the development and progression of atherosclerosis. Evidence in support of this contention comes from recent clinical studies in which circulating levels of acute-phase proteins such as C-reactive protein, fibrinogen, and serum amyloid A have been shown to correlate closely with the risk of an acute coronary syndrome, or indeed stroke. This applies both in apparently healthy subjects without overt coronary disease and in those who have already survived a myocardial infarction. Recognition of this association led to the speculation that atherosclerosis or its consequences may be precipitated by infection, and several early studies demonstrated an association between serological evidence of recent infection with *Helicobacter pylori*, *Chlamydia pheumoniae*, and cytomegalovirus and coronary events. The case for *Chlamydia pheumoniae* was particularly plausible since these organisms can be found in vascular smooth muscle cells and macrophages within atherosclerotic lesions. However, recent more rigorous studies have suggested that these associations may be spurious, due to inadequate control for confounders, particularly for *Helicobacter pylori* infection, and do not indicate a strong association between infection and the development of coronary artery disease. The outcome of several ongoing studies on the effects of antibiotic therapy on coronary events should help resolve this important question. Evidence that elevated levels of C-reactive protein reflect inflammatory activity within atherosclerosis rather than infection comes from studies of inhibition of hepatic HMG CoA reductase with pravastatin. Compared with patients treated with placebo, whose C-reactive protein levels rose, patients treated with pravastatin, which lowers low-density lipoprotein cholesterol but has no known effect on bacteria, experienced a substantially reduced risk of a coronary event with an associated fall in level of circulating C-reactive protein.

The causes of acute coronary syndromes

Unstable angina versus myocardial infarction

The pathogenesis of unstable angina and myocardial infarction is similar in that both occur because of plaque erosion or rupture and subsequent aggregation of platelets. Lesions causing unstable angina are characteristically eccentric and ulcerated with associated non-occlusive thrombus, whereas angiographic studies have shown that myocardial infarction is usually due to complete thrombotic occlusion of the relevant vessel. Treatment of myocardial infarction therefore includes aspirin to prevent further platelet aggregation and a fibrinolytic drug to lyse the occluding thrombus. In unstable angina the accumulation and activation of platelets causes

local release of a number of peptides, such as thromboxane A2, 5-hydroxytryptamine, and platelet derived growth factor, all of which are potent vasoconstrictors. Thus, vasoconstriction superimposed on partial occlusion of the lumen is often responsible for myocardial ischaemia (Fig. 1). Although the thrombus in unstable angina may be non-occlusive, it provides the substrate from which embolization of platelet aggregates downstream can produce microinfarcts that are not apparent electrocardiographically, but which may be detectable through release of sensitive markers of myocyte necrosis, such as troponins I or T. Patients with unstable angina and an associated increase in circulating troponin levels are at particularly high risk of subsequent coronary events if not treated aggressively.

The logical result of this analysis is that the management of unstable angina includes inhibition of platelet aggregation with aspirin, heparin to prevent progression to a fibrin clot, and nitrates to offload the myocardium and relieve any local coronary artery spasm. Although aspirin continues to be the first-line antiplatelet drug, newer drugs such as ADP and glycoprotein IIb/IIIa antagonists are rapidly establishing their role in the management of unstable angina and other acute vascular syndromes such as transient cerebral ischaemic attack.

Plaque instability often occurs in small, haemodynamically insignificant atherosclerotic plaques that are clinically silent (not causing stable angina) and which may not be apparent at angiography. Approximately 70 per cent of coronary lesions that break down to cause thrombosis and subsequent myocardial infarction cause less than a 50 per cent stenosis of the relevant coronary artery. This is probably because arteries can remodel to accommodate large atherosclerotic lesions without reducing lumen diameter, but it explains why so many patients (about 50 per cent) suffer myocardial infarction without experiencing any prior symptoms of coronary disease. It also serves to emphasize the relatively greater importance of plaque composition than plaque size in determining the outcome of coronary disease.

Other causes of acute coronary syndromes

Not all acute coronary syndromes arise as a result of rupture of atherosclerotic plaque. Rarely, acute myocardial infarction can occur in the absence of significant coronary artery disease. Acute coronary artery dissection is a rare event that occurs more commonly in women than men and which can cause myocardial infarction and even sudden death. The underlying aetiology is unknown. Coronary artery spasm may also precipitate an acute coronary syndrome and even myocardial infarction, but the reasons for coronary spasm in the absence of atherosclerosis are poorly understood, although they almost certainly arise from endothelial dysfunction. Angina due to spasm, so-called Prinzmetal's variant angina, is characterized by marked ST segment elevation on the ECG, by contrast with ischaemia due to atherosclerotic plaque rupture or erosion that are characterized by ST segment depression. It usually responds to treatment with calcium channel blockers, particularly nifedipine. Finally, acute coronary syndromes can be precipitated by legitimate and illegitimate drug therapy. For example, symptoms of myocardial ischaemia can be provoked in some patients by the use of HT1 agonists to treat migraine or sildenafil for impotence, particularly when combined with nitrates. Recreational drug abuse, particularly cocaine derivatives, may also precipitate an acute coronary syndrome and should be considered when clinically appropriate.

Consequences of acute coronary syndromes

Sudden reduction in blood flow through a coronary artery inevitably leads to myocardial ischaemia unless the myocardium is supplied by an adequate collateral blood supply from another coronary artery. The three main clinical consequences of such ischaemia are arrhythmias, reduced contractile function, and, less commonly, myocardial rupture: these are discussed in Chapter 15.4.2.3. However, it is becoming clear that ischaemia has complex effects on the myocardium, and considerations of myocardial stunning and hybernation and of ischaemic preconditioning are likely to be increasingly important in the development of new therapies.

Myocardial stunning and hibernation

Contractile dysfunction after infarction does not always imply irreversible necrosis of myocytes. The myocardium may be either stunned or hibernating. Viable myocardium is said to be stunned when it fails to contract appropriately after the obstruction to coronary blood flow has been removed or bypassed. This may be due to what is called the 'no reflow' phenomenon in which there is a failure of flow through the previously ischaemic tissue, despite there being a patent feeding artery. The underlying mechanism for this phenomenon is unknown, but probably relates to local endothelial dysfunction. Stunned myocardium usually recovers normal or near normal function with time.

The term hibernating myocardium is used to describe viable myocardium that fails to contract properly because it has an inadequate blood supply. The importance of this concept lies in the fact that restoration of an adequate blood flow, by angioplasty or bypass surgery, usually results in recovery of contractile function. Identification of hibernating myocardium is crucial in determining which patients will experience an improvement in myocardial function if subjected to a revascularization procedure: revascularization of dead myocardium clearly confers no benefit. Techniques based on the echocardiographic measurement of left ventricular wall motion under pharmacological stress or which compare myocardial metabolism and blood flow with positron emission tomography (see above) can identify those with myocardial dysfunction due to hibernation.

Ischaemic preconditioning

Over recent years it has become clear that episodes of myocardial ischaemia may protect the myocardium from the consequences of a further ischaemic event. Thus it has been found that a transient interruption of myocardial blood flow delays the onset of infarction if blood flow is subsequently permanently interrupted. The biochemistry of this phenomenon is complex and incompletely understood and its clinical importance is currently unclear. However, its recognition heralds the possibility of future development of drugs that will protect the myocardium from the consequences of ischaemia.

Further reading

Arbustini E *et al*. (1999). Plaque erosion is a major substrate for coronary thrombosis in acute myocardial infarction. *Heart* **82**, 269–72.

Davies MJ (1995). Stability and instability—2 faces of coronary atherosclerosis—the Paul-Dudley-White-Lecture. *Circulation* **94**, 2013–20.

Fuster V, Fayad Z, Badimon J (1999). Acute coronary syndromes: biology. *Lancet* **353**, 5–9.

Libby P (1995). Molecular bases of the acute coronary syndromes. *Circulation* **91**, 2844–50.

Redwood SR, Ferrari R, Marber MS (1998). Myocardial hibernation and stunning: from physiological principles to clinical practice. *Heart* **80**, 218–22.

Yellon D *et al*. (1998) Ischaemic preconditioning: present position and future directions. *Cardiovascular Research* **37**, 21–33.

15.4.2.2 Management of stable angina

L. M. Shapiro

Introduction

Coronary artery disease is the predominant cause of death in the developed world, causing 300 000 deaths per year in the United Kingdom, and is increasing in importance in developing countries and the previous Communist bloc. Coronary artery disease is also a major cause of hospital admission and clinic consultation.

The syndrome of stable angina pectoris is clinically defined as consistent exertional- or stress-related cardiac symptoms, usually of chest pain and, less frequently, shortness of breath. In the last decade there have been major advances in the management of patients with stable symptoms and coronary artery disease: particularly in the areas of lifestyle modification, medical treatment, and revascularization. The management of patients with stable angina is the subject of this chapter.

General patient management

The diagnosis of angina pectoris in a patient should be accompanied by a detailed explanation of the disorder. In particular, that angina pectoris has an unpredictable nature, with the possibility of deterioration as well as stabilization or an improvement with treatment. The treatments available include lifestyle modification, pharmacological intervention, and revascularization. These need to be individualized and usually ameliorate symptoms, and in most circumstances improve prognosis.

In the initial management of patients with chronic angina pectoris, a search should be made for treatable conditions which increase myocardial oxygen demand or reduce oxygen delivery—for example, marked obesity, thyrotoxicosis, fever, anaemia, tachycardia, or aortic stenosis.

The medical management of stable angina pectoris then depends upon consideration of the following options to improve symptoms and/or prognosis:

1. control of risk factors and lifestyle modification;
2. pharmacological management;
3. revascularization.

Risk-factor management and lifestyle modification

Lifestyle modification, if sufficiently rigorous, has the advantage of providing a modest degree of symptom relief and a reduction in the rate of further development of coronary artery disease.

Hypertension

The relationship between the development of coronary artery disease and hypertension is well established. Hypertension increases myocardial oxygen demand and leads to ischaemia in patients with obstructive coronary artery disease. Elevation of left ventricular mass is a strong predictor of mortality due to coronary artery disease. Treatment of hypertension, especially in the elderly, has been shown to reduce the mortality from cardiovascular causes by nearly one-third.

Cigarette smoking

Not only is smoking a powerful risk factor for the development of coronary artery disease, it also leads to more severe and premature atherosclerotic plaques. Cigarette smokers with documented coronary artery disease have an increased 5-year mortality risk, and cessation of smoking lessens the risk of adverse cardiovascular events. Smoking may also lead to exacerbation of angina pectoris by increasing myocardial oxygen demand and reducing coronary blood flow by a direct effect on coronary artery tone. Passive smoking may also be important.

Hyperlipidaemia

Reduction of cholesterol by diet, and more especially by drugs therapy, has been shown in primary prevention trials to reduce the risk of the development of coronary disease. Treatment with the statin group of drugs in patients with established coronary artery disease (secondary prevention) has only a modest effect on angiographically documented, coronary artery obstruction, but significantly reduces the number of new cardiovascular events. The Scandinavian Simvastatin Survival Study treated patients with a cholesterol level in excess of 5.5 mmol/l, but a value of 4.8 mmol/l or less is currently seen as a treatment threshold.

Recent evidence from the AVERT study (atorvastatin versus revascularization treatment) suggests that, in patients with mild and stable angina pectoris, aggressive cholesterol lowering with atorvastatin had a similar effect in reducing ischaemic events as treatment with percutaneous transluminal coronary angioplasty, although the latter group had better symptom control.

Antioxidants

Oxidized low-density lipoproteins (**LDL**) may play an important role in the pathogenesis of atherosclerosis. Agents that prevent lipid peroxidation of LDL particles might therefore influence the development of atherosclerosis and its clinical consequences. Epidemiological data suggests that high vitamin E levels are protective of coronary artery disease. Giving b-carotene does not appear to confer an advantage, whereas vitamin E may show some benefits in secondary prevention in doses of 400 or 800 IU per day.

Aspirin

Activated, aggregating platelets play an important role in the development of acute coronary events. A meta-analysis of 300 studies—including 140 000 patients with chronic coronary heart disease, stroke, or previous bypass surgery—has shown aspirin to have a prophylactic benefit. Aspirin is therefore widely used in the dose

range of 75 to 150 mg per day in patients with chronic angina pectoris.

Physical inactivity

Regular aerobic exercise allows a greater workload to be performed for any level of oxygen consumption, allowing patients with coronary heart disease to increase their exercise tolerance. The physiological benefits of exercise training have largely been described from postmyocardial infarction rehabilitation. However, smaller studies have confirmed significant benefits in improved quality of life, effort tolerance, and possibly morphology of coronary artery lesions, from a graded supervised physical exercise programme in those with chronic stable angina.

Obesity

The presence of obesity most probably acts via increased blood pressure and serum cholesterol as coronary artery disease risk factors. However, it may lead to symptom development in chronic angina and weight loss may have a profound influence on symptoms.

Diabetes mellitus

This is a powerful and independent risk factor for the development of coronary artery disease. Control of blood glucose levels is vital in the management of patients with stable angina pectoris.

Pharmacological management

Basic treatment includes the use of aspirin, sublingual glycerol trinitrate (**GTN**), and b-blockade. Other antianginal agents may also be helpful, including calcium antagonists, long-acting nitrates, and potassium-channel openers.

Nitrates

In 1867, Brunton first described the clinical benefit of organic nitrates. These are prodrugs and are biotransformed by denitration, thereby liberating nitric oxide. This endothelium-derived relaxing factor (**EDRF**) exerts a vasodilatory effect, even in the absence of the endothelium, and also reduces platelet aggregation and adhesion. The antianginal and haemodynamic effects are mediated predominantly by vasodilatation of the venous system, leading to a fall in left ventricular preload and cardiac work, but also by vasodilatation of arteries, including the coronary arteries.

GTN remains the most commonly used preparation. Single doses of tablets or spray rapidly relieve angina pectoris and may be repeated every 5 min if symptoms persist. GTN is particularly effective when used prophylactically 2 to 5 min before activity. Many patients need no other antianginal medication if their angina is predictable and not particularly severe. Adverse effects of GTN are common and include flushing, headache, and hypotension. GTN is best used in the sitting or lying position to avoid hypotensive syncope, particularly for the first few doses.

Various nitrate preparations are widely used in the chronic treatment of angina pectoris, but are all limited by the development of nitrate tolerance. The mechanism leading to a reduction in clinical efficiency is not well understood, but its clinical effects can be overcome by intermittent nitrate dosing.

Nitrates can be given transdermally. Dermal absorption from ointment and patches has been shown to improve exercise duration. The ointment is applied in strips, often to the chest, and is particularly useful in those with nocturnal angina, or in immobile patients with severe symptoms. Transdermal application is also subject to nitrate tolerance.

Isosorbide dinitrate has a low bioavailability and marked variations in plasma concentration occur. Isosorbide-5-mononitrate is the active metabolite of the dinitrate and has excellent bioavailability, with a standard preparation yielding clinical effects for 4 to 8 h. Long-acting preparations, given at doses of 20 to 60 mg, are beneficial for up to 12 h. Single daily dosing does not induce tolerance, but such dosing regimes can lead to rebound myocardial ischaemia during the nitrate-free period. However, this is uncommon in clinical practice, particularly if the timing of the dose covers the period when the patient is physically active.

b-Blocking agents

These are the cornerstone of the pharmacological management of chronic angina pectoris. b-Blocking agents are well tolerated and reduce the frequency and duration of anginal episodes and improve exercise tolerance. They are also effective antihypertensive agents and prevent some arrhythmias. They act by competitively inhibiting catecholamine effects on the b-adrenergic receptor. This reduces heart rate and improves coronary perfusion (by prolonging diastole), thereby reducing an exercise-induced rise in blood pressure and contractility.

There are increasing numbers of b-blocking agents available. Factors that influence their usage include selectivity, elimination half-life, intrinsic sympathomimetic activity, and vasodilatory properties. However, for the standard treatment of patients with chronic angina pectoris, most agents will have a similar beneficial effect.

The b-receptor has two major subtypes: b1 and b2. The former predominates in the heart and the latter in the lungs. Non-selective b-blocking agents (propanolol, nadolol, pindolol, sotalol, and timolol) block both receptors, whereas selective agents (atenolol, bisoprolol, metoprolol) predominantly influence b1 receptors. These effects are relative and as doses rise selectivity becomes less prominent, so that bronchoconstriction may occur at effective antianginal doses.

Acebutolol, celiprolol, and pindolol have intrinsic sympathomimetic activity, inducing low-grade stimulation when sympathetic activity is low. The clinical significance of this is uncertain. Lipid-soluble agents such as propranolol and metoprolol are readily absorbed and have shorter half-lives.

Most agents are started in relatively small doses which are titrated against symptoms and markers of b-blockade such as heart rate, particularly on exercise. They are generally well tolerated, but bradycardia, atrioventricular block, heart failure, central nervous system effects (fatigue, depression, and nightmares) are often seen. Cold hands and feet, sexual dysfunction, and lethargy are also common. If b-blockers are to be stopped, this should be done gradually. Abrupt cessation can lead to worsening of angina pectoris with reflex tachycardia and anxiety.

The most appropriate recipient of b-blockade is an individual with exercise-induced angina, possibly with coexisting hypertension or arrhythmias. These patients should commence treatment with, for example, atenolol 50 to 100 mg once daily, metoprolol 50 to 100 mg twice daily, or propanolol 80 mg twice or three times per day. There is additional benefit after myocardial infarction. However, such agents are best avoided in the presence of reversible airways obstruction, diabetes, and impaired left ventricular function. Depression is often worsened, as is peripheral vascular disease.

Calcium antagonists

Calcium antagonists constitute a heterogeneous group of compounds with various degrees of effect on heart muscle, atrioventricular conduction, and peripheral and coronary vessels. They act by inhibiting calcium ion movement through slow channels in cardiac and smooth muscle membranes by non-competitive blockade of voltage-sensitive calcium channels. The effect of calcium antagonists in angina pectoris is related to a reduction in myocardial oxygen demand with some increase in oxygen supply. The latter is particularly important in patients with a vasoconstrictor component to their disease. While calcium antagonists may be effective on their own, some can be particularly useful when taken in combination with b-blocking agents. There are three main first-generation, calcium-channel blocking agents (verapamil, diltiazem, and nifedipine), which have quite diverse physiological actions and clinical effects.

Verapamil

This acts by slowing the heart rate and reducing myocardial contractility as well as dilating systemic and coronary vessels. It is markedly negatively inotropic, but this rarely causes clinical effects. Verapamil is started orally in the range of 40 to 80 mg three times daily. Adverse effects are hypotension, facial flushing, and constipation.

Diltiazem

The cardiac depressant effect of diltiazem is rather less than that of verapamil, but rather more than that of nifedipine. It is well tolerated. Although it causes little vasodilatation of coronary arteries, it does block exercise-induced coronary vasoconstriction and reduces afterload. It is usually started at a dose of 60 mg three times daily, but a number of long-acting preparations are now available (200–300 mg once daily).

Nifedipine

Nifedipine is a dihydropyridine derivative and there are a number of second-generation agents of a similar type (nicardipine, isradipine, and amlodipine). It is a more potent vasodilator than verapamil or diltiazem. Its beneficial effect in angina is due to its capacity to reduce myocardial oxygen requirement by afterload reduction and increase oxygen delivery through coronary vasodilatation. Nifedipine is usually started as 10 mg three times daily, but there are number of long-acting preparations that deliver 30 to 90 mg per day. Adverse effects are quite prominent and relate to vasodilatation, including headache, dizziness, palpitations, flushing, hypertension, and leg oedema. The adverse effects of nifedipine are reduced by the use of sustained-release preparations and short-acting formulations should probably not now be used. Nifedipine may increase in mortality. Second-generation dihydropyridine derivatives may have some advantages in side-effect profiles.

Other pharmacological agents

The potassium-channel opener nicorandil has been show to have effective antianginal properties. It is currently prescribed to patients with persisting symptoms despite 'maximal' medical therapy. Whether patients would benefit from an earlier introduction of nicorandil is yet to be determined.

Revascularization

Coronary angiography, with a view to revascularization, is recommended in patients who remain symptomatic or have documented ischaemia, despite maximal medical therapy. Other indications include the results of non-invasive testing suggesting poor prognosis despite milder symptoms (exertional hypotension, arrhythmias, and marked ischaemia), or for occupational reasons. While in younger, more active, patients, revascularization may be considered earlier in the disease course, age itself is not a restriction. However, whilst percutaneous transluminal coronary angioplasty can be performed safely in the elderly, the mortality from bypass surgery will rise unacceptably in very old patients with coexisting disease (see later). Coronary angiography is underutilized but clinically very useful in patients with diagnostic doubt as to the cause of chest pain, as normal findings considerably simplify management.

Exercise electrocardiography (or similar tests of ischaemia) give positive results in only 60 to 80 per cent of patients with coronary artery disease. The remainder are false-negative tests. Also, some normal individuals have an ischaemic response—these are false-positives. Such lack of sensitivity and specificity makes these tests too unreliable for screening normal individuals for coronary artery disease (for a fuller discussion of these important issues see Chapter 15.3.2).

Summary of the medical management of patients with chronic stable angina

1. Confirm the diagnosis by demonstrating myocardial ischaemia.
2. Control risk factors and modify lifestyle: in particular, weight loss may improve symptoms.
3. Treat with aspirin and GTN for both symptom relief and prophylaxis.
4. Add a b-blocking agent.
5. If angina is not controlled, add a long-acting nitrate or calcium-channel blocker.
6. If angina persists, and there are no contraindications, consider coronary revascularization.

Further reading

Diaz MN, *et al.* (1997). Antioxidants and atherosclerosis heart disease. *New England Journal of Medicine* **337**, 408–11. [Overview of the importance of antioxidants.]

O'Connor GT, *et al*. (1989). An overview of randomised trials of rehabilitation with exercise after myocardial infarction. *Circulation* **80**, 234–44. [Overview of the importance of exercise in coronary artery disease.]

Parker JD, Parker JO (1998). Nitrate therapy for stable angina pectoris. *New England Journal of Medicine* **338**, 520–6. [Important review of nitrate therapy and tolerance.]

Pitt B, *et al*. (1999). Aggressive lipid-lower therapy compared with angioplasty in stable coronary artery disease. *New England Journal of Medicine* **341**, 70–6. [First study to compare lipid-lowering therapy with PTCA]

Shuler G, Hambrecht R, Schlierf G *et al.* (1992). Myocardial perfusion and regression of coronary artery disease in patients with a regime of intensive physical exercise and low fat diet. *Journal of the American College of Cardiology* **19**, 34–8. [Effect of lifestyle modifications on symptoms.]

15.4.2.3 Management of acute coronary syndromes: unstable angina and myocardial infarction

Keith A. A. Fox

Introduction

Acute coronary syndromes comprise a clinical spectrum of conditions that extend from new-onset angina through unstable angina and minimal myocardial injury (enzyme release without diagnostic changes of infarction) to myocardial infarction based upon ECG and enzyme criteria. These different clinical presentations share important pathophysiological features. They occur in patients with underlying symptomatic or occult coronary artery disease and flow-limiting or non-flow-limiting atheromatous plaques in the coronary arterial wall.

The acute coronary syndrome is precipitated in an abrupt change in an atheromatous plaque, resulting in increased obstruction to perfusion and ischaemia or infarction in the territory supplied by the affected vessel. For discussion of the mechanisms involved, see Chapter 15.4.2.1. The clinical manifestations are dependent not only upon the degree of obstruction to perfusion, but also on the presence or absence of collateral perfusion, the extent and distribution of fragmented microthrombi, and myocardial oxygen demand in the perfused territory. Thus, the clinical consequences of plaque rupture can range from an entirely silent episode through to a development of abrupt occlusion with profound ischaemia and infarction or sudden death.

Rational management, including pharmacological treatment and percutaneous or surgical revascularization strategies, are critically dependent on the underlying pathophysiological mechanisms and on the extent and severity of myocardial ischaemia. Despite sharing key pathophysiological mechanisms with ST-segment elevation, acute myocardial infarction (**MI**), unstable angina, and non-ST elevation MI demand special attention. Whereas acute reperfusion strategies (thrombolysis or primary percutaneous coronary intervention, **PCI**) are of proven benefit in ST-segment elevation infarction (or that associated with new bundle-branch block), there is no evidence that thrombolytic treatment improves outcome in the remainder of the syndrome. For this reason a pragmatic division is made between acute coronary syndromes with ST-segment elevation, and those without.

Unstable angina/non-ST elevation MI

Outcome based upon trial data and large-scale observational registry studies

Imprecision in the definition and characterization of unstable angina or non-ST elevation MI has previously resulted in underestimation of the risk of this syndrome. The inclusion of patients with chest pain, but without diagnostic features of acute ischaemia, masked the true hazards of the syndrome.

Patients with acute coronary syndrome (without persistent ST elevation) are at substantial risk of subsequent cardiac events despite current therapy. Based on data from randomized trials and prospective registry studies, about 9 to 11 per cent suffer death or myocardial infarction at 6 months, with almost half of this risk within the first 7 days (GUSTO IIb, OASIS Registry, GRACE Registry). Between one-quarter and a third of patients suffer death, myocardial infarction, or readmission for unstable angina within 6 months (GRACE Registry, PRAIS Registry).

Clinical presentation and definition of the syndrome

Unstable angina may present *de novo* (new-onset angina) with episodes of typical ischaemic discomfort at rest (rest angina) or on minimal exertion. Alternatively, a previously stable pattern of angina may deteriorate abruptly or progressively, resulting in episodes of typical rest angina or angina provoked by minor exertion (crescendo angina). Although new-onset exertional angina is not generally recognized as part of the acute coronary syndrome, the outcomes are similar (7 per cent develop non-fatal MI and 4 per cent die, and a further 19 per cent require revascularization within 15 months) (Table 1).

As a clinical syndrome, unstable angina is conventionally diagnosed by the presence of new-onset angina or angina of worsening severity in terms of frequency or duration. The syndrome must be distinguished from non-cardiac pain, stable angina, and infarction. To improve the specificity of the diagnosis, and for the purposes of clinical trials, a more restricted definition has been employed which requires at least 15 to 20 min of typical, ischaemic discomfort, or two 5-min episodes at rest. The specificity is further improved when the definition requires objective evidence of ischaemia or evidence of underlying coronary artery disease. ST-segment depression on the electrocardiogram, especially in association with typical pain, is highly predictive, whereas the less specific ECG abnormalities including T-wave inversion are less strong predictors. Markers of myocardial damage (troponins or cardiac enzymes) are powerfully predictive. The ECG changes and the markers also indicate an adverse prognostic outcome. In the absence of such markers, documented evidence of underlying coronary artery disease (prior infarction or angiographically demonstrated coronary disease) helps to confirm the diagnosis.

Minimal myocardial damage—infarction without ST elevation or Q-wave development ('including non-Q-wave MI')—lies between unstable angina and Q-wave myocardial infarction in its prognostic significance (Fig. 1). It is best considered as part of the continuous spectrum of acute coronary syndromes rather than as a separate entity (Fig. 2). Management strategies are the same as for other higher risk patients with unstable angina. Minimal myocardial injury arises as a result of episodes of transient occlusion and/or embolization of thrombus into the distal circulation of the affected coronary vessel. Injury to myocytes results in the release of enzymes from the contractile apparatus (troponins) and cardiac enzymes, indicating irreversible injury (creatine kinase (CK) or CK-MB (**CK-MB**)). Thus, although the management of ST-*segment-elevation MI* differs, the remainder of the acute coronary syndrome should be managed as a continuous spectrum, but influenced by risk stratification.

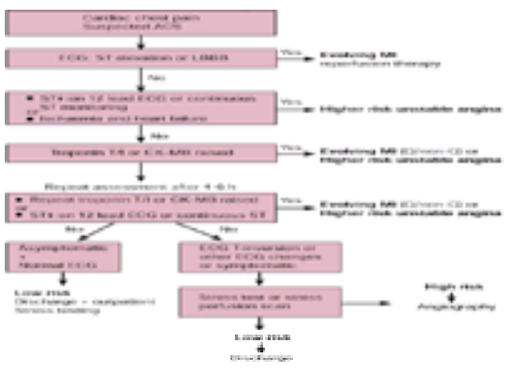

Fig. 1 Diagnostic triage for suspected acute coronary syndromes. Flow chart to illustrate the key diagnostic features for evolving myocardial infarction, for higher risk unstable angina and for low-risk patients.

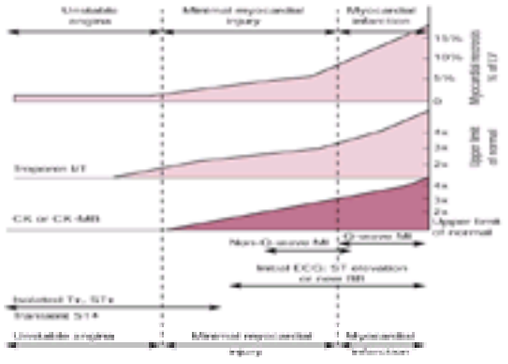

Fig. 2 Schematic of the spectrum of acute coronary syndromes. The relationship between the extent of myocardial necrosis (upper panel) and the release of cardiac enzymes (troponin I or T, or CK or CK-MB) and electrocardiographic changes. The schematic illustrates that unstable angina is associated with minor myocardial necrosis without the release of CK or CK-MB but with the release of troponin in approximately one-third of cases. With minimal myocardial injury, more extensive myocardial necrosis is associated with the release of cardiac enzymes and ECG changes (non-ST elevation-wave MI). With Q-wave MI the most extensive myocardial necrosis and release of cardiac enzymes occurs in association with the electrocardiographic changes.

Variant angina

A condition characterized by recurrent episodes of angina at rest, often accompanied by ST-segment elevation, was originally described by Prinzmetal. Such patients may have relatively a well-preserved exercise capacity and specific diurnal periodicity to the symptoms, with occasional clustering of symptomatic episodes and symptom-free periods. Coronary vasospasm has been implicated in the syndrome, and in some instances the coronary arteries appear angiographically normal, or with only minor occlusive disease. However, the vast majority of patients with symptomatic ischaemia and ST-segment elevation or depression have occlusive coronary artery disease with superimposed thrombosis, and they should be managed as such. Indeed, vasospasm very frequently accompanies plaque rupture and thrombosis. With the use of modern cardiac enzyme studies and coronary angiographic techniques, variant angina is a rare diagnosis that should only be considered if atherosclerotic acute coronary syndromes have been excluded. It is possible that erosion of a minor plaque, superimposed thrombosis and embolization may be responsible for a number of such clinical cases.

Risk characterization and outcome

In patients with acute coronary syndromes, risk can be separated into two components: prior risk and acute ischaemic risk. Based upon large observational studies it is possible to derive univariate and multivariate predictors of outcome. Prior risk is determined by patient characteristics (age and gender), prior ischaemic heart disease (myocardial infarction, heart failure, prior angina), and systemic factors that influence risk (hypertension, diabetes, and other life-threatening systemic disorders). These are largely unchangeable and can be considered as the background level of risk that the patient brings with them to the point of presentation. The extent of underlying coronary artery disease and prior heart failure require assessment and management.

The second component of risk, 'acute ischaemic risk' is determined by the severity of ischaemia induced by plaque rupture and thrombosis and the extent of the territory affected. Collateral perfusion and myocardial oxygen demand both influence the extent of ischaemia. Patients may experience transient complete occlusion, or severe subtotal occlusion complicated by distal embolization of fragments of a platelet-rich thrombus, also by changes in vascular tone in the distal territory. Clinical markers of acute ischaemic risk include electrocardiographic changes, enzyme release into the systemic circulation, and mechanical and arrhythmic complications of the ischaemic episode.

The first priority in the management of patients with acute coronary syndromes is to deal with acute ischaemic risk. Assessment of the extent of underlying coronary artery disease (for example with stress testing and/or angiography) and assessment of left ventricular function can take place later in the management of these patients. Simplistically, 'prior risk' can be regarded as the 'baggage' that the patient carries with them, and 'acute ischaemic risk' as an 'acquired hazard' arising from the new ischaemic event. The distinction is important because management strategies for prior risk aim to treat heart failure, underlying coronary and systemic disease, and risk factors. The management of acute ischaemic risk aims to reverse the impact of acute coronary obstruction and thrombosis.

The clinical syndrome and outcome

The Braunwald classification categorizes unstable angina according to the mode of onset and time course (Table 2). It was empirically based but has been validated by prospective studies. Patients with unstable angina at rest (Braunwald Class 3) have the highest risk of an adverse cardiac event (approximately 11 per cent in hospital event rate). Similarly, those with unstable angina following acute myocardial infarction are at an increased risk. Although useful, many of the patients that present with acute coronary syndromes are in Braunwald Class 3B and additional methods of risk characterization are required to optimize management.

A diagnostic triage system can be developed for patients with suspected acute coronary syndromes (Fig. 1). This is based upon electrocardiographic changes, enzyme release, and stress or perfusion testing. As a result, those with evolving infarction are identified and those with higher risk unstable angina are separated from those with lower risk. These categories of patients require different management strategies.

The electrocardiogram and outcome

The 12-lead electrocardiogram (performed on admission) provides direct prognostic information (Table 3).

The greatest risk of death and subsequent MI is seen in patients with simultaneous ST elevation and depression. The next highest risk is seen in those with transient ST-segment elevation or ST-segment depression. Isolated T-wave inversion carries a lower risk (Table 3). The number of leads demonstrating ST deviation also yields prognostic information: among those with ST deviation in the anterior leads a rate of death or myocardial infarction of 12.4 per cent was seen at 1 year, higher than seen with similar changes in other locations (TIMI III trial). Patients with a left main and three-vessel coronary artery disease may show a combination of ST-segment elevation and depression.

Holter ST-segment recording can identify patients with unstable angina and either silent or symptomatic myocardial ischaemia with an increased risk for major subsequent cardiac events. However, they provide off-line analysis and are not suitable for the prediction of imminent events. Computer-assisted, continuous,

multi-lead, ECG monitoring techniques have become available for real-time ECG and ST-segment monitoring. The occurrence and extent of ischaemic territory identified by such continuous recordings can provide additional prognostic information over and above the admission ECG. They can be combined with serial enzyme markers; recent studies have indicated that together they provide additional prognostic information (FRISC study).

Biochemical markers and outcome

Enzymes are gradually released into the systemic circulation following complete or transient occlusion of the coronary artery, or fragmentation of a thrombus and embolization. Following total occlusion of the vessel, creatine kinase (or more specifically CK-MB) will be released and detectable at clearly abnormal levels about 6 to 8 h after the event, unless there is extensive collateral perfusion. By convention, CK values greater than twice the upper limit of normal are associated with infarction, and this is categorized into those with Q-wave development and those without. However, a continuous spectrum of injury exists from unstable angina through non-Q-wave myocardial infarction to Q-wave infarction. The evolution of Q waves on the ECG (or none) provides prognostic information by the time of hospital discharge, but cannot be used to guide early treatment.

The measurement of myocardial isoforms of troponins in the blood are more sensitive markers of injury. The cardiac isoforms of troponin I and troponin T are exclusively expressed in cardiac myocytes and provide specific evidence of myocardial damage. Only a few patients with renal dysfunction will have falsely elevated troponin measurements. Following marked ischaemia or infarction, troponins are released from the cytosolic pool and first appear in the circulation in detectable concentrations between 3 and 4 h after the ischaemic event and reach diagnostic concentrations at 6 to 8 h. Troponin release may be regarded as evidence of myocardial injury and it carries a prognostic significance worse than that of patients without troponin release but less severe than those with acute infarction diagnosed by by a rise of more than twofold in CK-MB (Fig. 3) unless the concentrations in the blood are markedly elevated. The greater the troponin release the greater the risk of subsequent myocardial infarction and death.

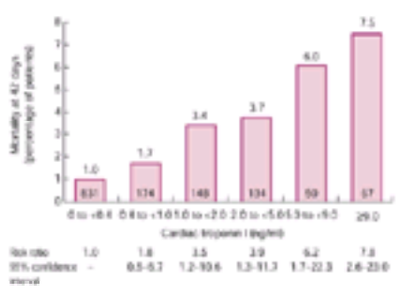

Fig. 3 Cardiac troponin I and subsequent risk of death among patients with unstable angina or non-ST-elevation MI. Mortality rates at 42 days (without adjustment for baseline characteristics) are shown for ranges of cardiac troponin I levels measured at baseline. The numbers at the bottom of each bar are the numbers of patients with cardiac troponin I levels in each range, and the numbers above the bars are percentages. p <0.001 for the increase in the mortality rate (and the risk ratio for mortality) with increasing levels of cardiac troponin I at enrolment. (Reproduced with permission from Antman *et al*. *New England Journal of Medicine* 1996, **335**, 1342–9.)

When should the cardiac enzymes be measured? The time course of the release of enzymes from myocardium is such that diagnostic concentrations may not be achieved until between 6 and 8 h after an ischaemic event. Thus, normal values for a patient on arrival do not exclude infarction or unstable angina. However, elevated values on arrival are highly predictive of subsequent infarction (for CK, or CK-MB, or troponins). The CK and CK-MB measurements should be repeated between 8 and 12 h later and also after any suspected ischaemic event. Troponins should be measured on arrival and at approximately 8 h: these provide the highest predictive accuracy.

Among those with persistently negative troponins and without significant ECG changes there is a very low risk of subsequent infarction and death (provided that severe underlying coronary artery disease is excluded). Ideally such patients should undergo predischarge stress testing; the most accurate test being that accompanied by myocardial perfusion scanning or stress echocardiography. Treadmill electrocardiograms on exercise are less accurate but more widely available (see Chapter 15.3.2).

Among patients in whom myocardial infarction is excluded by standard criteria, about one-quarter will have elevated troponin levels (minor myocardial damage) and these patients have the same frequency of cardiac events during follow-up as seen for conventionally diagnosed infarction.

Markers of inflammation

Inflammatory changes in the vessel wall promote plaque fissuring or erosion, and inflammatory changes also follow episodes of minor myocardial damage. In unstable angina there is evidence that inflammatory markers (C-reactive protein (**CRP**), interleukin-6, and interleukin-1) are independent predictors of adverse outcome. Only 50 to 70 per cent of patients with Braunwald class IIIB unstable angina have elevated CRP levels. After the acute phase, continuing inflammation—with, for example, elevated CRP—occurs in half of those whose levels are acutely elevated and identifies a category of patients at increased risk. Although inflammatory mechanisms are implicated in plaque growth and plaque destabilization, specific anti-inflammatory therapy of surface integrins (including the inhibition of surface integrins or inhibition of polymorph infiltration with anti-CD11/CD18 antibodies) has not yet been demonstrated to improve outcome.

Treatment

Antiplatelet therapy

Aspirin

Exposure of the contents of atheromatous plaque to circulating blood triggers platelet activation by several different pathways. Aspirin is a potent and irreversible inhibitor of platelet cyclo-oxygenase, blocking the formation of thromboxane A_2 and inhibiting platelet aggregation. Although the effects of aspirin can be overcome in the presence of potent thrombogenic stimuli, nevertheless the benefits of aspirin treatment in unstable angina are clearly defined and substantial. The Antiplatelet Trialists Collaboration demonstrated a reduction of 36 per cent in death or MI with antiplatelet treatment (predominantly aspirin) versus placebo in unstable angina trials. Aspirin treatment significantly reduces subsequent myocardial infarction, stroke, and vascular death, with the largest reductions seen amongst patients at highest risk. In patients with unstable angina, four key studies have demonstrated that aspirin significantly reduces the risk of cardiac death or non-fatal MI by approximately 50 per cent.

The efficacy of lower dose aspirin (75 mg day) therapy has been demonstrated in several studies, including those of Wallentin and colleagues where long-term effects were evaluated in men under 70 years of age with unstable coronary artery disease. After 6 and 12 months of aspirin treatment the risk of myocardial infarction or death was reduced by 54 per cent and 48 per cent, respectively (risk ratio 0.52 with 95 per cent confidence intervals 0.37–0.72). The strength of evidence and magnitude of benefit demonstrated with aspirin treatment in unstable angina or non-ST- *segment-elevation MI* is such that aspirin forms the reference standard against which alternative or adjunctive antiplatelet therapies are judged.

- Aspirin treatment is indicated in all patients with acute coronary syndromes unless there is good evidence of aspirin allergy.

Nevertheless, patients with acute coronary syndromes are at significant risk despite aspirin therapy. In prospective registry studies of unstable angina/non-ST-*segment-elevation MI*, and in spite of aspirin treatment in more than 80 per cent of patients, the risk of death or myocardial infarction is approximately 10 per cent at 6 months and the risk of death/myocardial infarction or refractory angina is approximately 22 to 33 per cent over the same period (OASIS Registry,

PRAIS Registry).

ADP antagonists (thienopyridines)

Ticlopidine and clopidogrel reduce thrombotic events following angioplasty and stenting. Clopidogrel has now replaced ticlopidine as it lacks the side-effect of thrombocytopenia.

Clopidogrel has now been tested in a large-scale trial of patients with unstable angina/non-ST-elevation MI (*N*=12 562 CURE trial). The agent was used on top of existing therapy and in addition to aspirin. It reduced death, non-fatal MI and stroke from 11.4 per cent to 9.3 per cent (95 per cent CI 0.72-0.90 *P* = <0.001). For every 1000 patients treated there were 28 fewer major cardiovascular complications but six more transfusions. Importantly, benefits were seen across risk groups (diabetics, hypertensives, CK or troponin elevation or not, revascularization or not). In a sub-study (PCI-CURE) clopidogrel also reduced death and myocardial infarction in those undergoing percutaneous revascularization (2.9 per cent clopidogrel versus 4.4 per cent for placebo). Thus, with the combination of clopidogrel and aspirin there is evidence of early and sustained reductions in the risks of death and myocardial infarction in patients that present with acute coronary syndromes, irrespective of their risk group and irrespective of baseline conditions. New guidelines are likely to incorporate this treatment in the management of the syndrome.

Glycoprotein IIb/IIIa inhibitors

Platelet adhesion is the initial step in haemostasis after disruption of an atheromatous plaque. It is triggered by damage to the vessel wall and exposure of the subendothelium, and is followed by platelet activation and aggregation. Regardless of the agonist, the final common pathway leading to the formation of a platelet aggregate is mediated by the glycoprotein (**GP**) IIb/IIIa receptor (up to 80 000 per platelet). GPIIb/IIIa receptor antagonists inhibit platelet aggregation irrespective of the agonist, and they prevent binding of fibrinogen to its receptor on the platelet surface.

To date, three GPIIb/IIIa inhibitors have been tested in large-scale randomized clinical trials. Abciximab is a chimeric human–murine monoclonal antibody that binds with high affinity to the receptor: it has a long biological half-life of 6 to 12 h, and low levels of receptor occupancy are detected even 2 weeks after treatment. Eptifibatide is a synthetic cyclic heptapeptide with high affinity for the arginine–glycine–aspartic acid ligand adhesion site of the IIb/IIIa receptor. It inhibits platelet aggregation in a dose-dependent manner and is readily reversible due to competitive binding and a short half-life of approximately 2.5 h. Tirofiban is a non-peptide tyrosine derivative which also binds to the arginine–glycine–aspartic acid site with high specificity. It inhibits platelet aggregation in a dose- and concentration-dependent manner and is rapidly reversible, with platelet function approaching normal levels in 90 per cent of patients within 4 to 8 h.

Although it is convenient to group glycoprotein IIb/IIIa receptor antagonists together, and undoubtedly there is evidence of a class effect, there are nevertheless biological and pharmacological differences between the agents.

Trials of GPIIb/IIIa inhibitors

More than 32 000 patients have been randomized in clinical trials involving GPIIb/IIIa inhibitors (16 trials). A highly significant (*p* <0.001) benefit is observed for the combined endpoint of death or MI at 48 to 96 h, 30 days, and 6 months. At 30 days the odds ratio is 0.76, or 20 fewer events per 1000 patients treated. Similarly, a highly significant benefit is observed for the combined endpoint of death/MI or revascularization at all time points. By contrast, mortality benefits are seen only at 48 to 96 h with no significant benefit at 30 days or 6 months. However, a pooled analysis of abciximab trials has revealed a net mortality benefit.

The effects of GPIIb/IIIa inhibitors may be greater in higher risk groups: trials involving percutaneous intervention (angioplasty with or without stent) have demonstrated significantly greater reductions in events with GPIIb/IIIa inhibitors in comparison with trials where angiography or intervention were not a prerequisite. In the interventional trials, death or MI was reduced by 27 fewer events per 1000 patients treated (odds ratio 0.64; confidence interval (95 per cent CI 0.51–0.80). In trials of GPIIb/IIIa inhibitors in acute coronary syndromes without mandatory intervention, there were 13 fewer events per 1000 patients treated (with an odds ratio of 0.88; 95 per cent CI, 0.81–0.97).

Subsequent analyses have been performed for those patients with elevated troponin levels and data from the CAPTURE study, the PRISM PLUS, and the PRISM study indicate that almost all the benefit is seen amongst those higher risk patients demonstrating troponin release.

Indications for treatment with GPIIb/IIIa inhibitors

Because of the variability in trial design, it is not yet feasible to assess differences, if any, in clinical benefit among the GPIIb/IIIa inhibitors. Head-to-head trials in acute coronary syndromes would be needed to resolve this issue. Nevertheless, robust evidence supports the following conclusions:

- Treatment with glycoprotein IIb/IIIa inhibitors results in improved outcome in unstable angina or non-Q-wave MI patients treated with aspirin and heparin. Most benefit is seen in high-risk patients (ST depression and/or troponin-elevation).
- Glycoprotein IIb/IIIa inhibitors result in improved outcome in patients requiring urgent percutaneous intervention for unstable angina or non-ST-*segment-elevation MI*.

Antithrombins

Unfractionated heparin

Unfractionated heparin has been adopted as standard antithrombin therapy in guidelines for the treatment of unstable angina/non-ST-elevation MI. However, the evidence upon which this is based is less robust than for other widely adopted treatment strategies. In practice, unfractionated heparin is difficult to control due to its unpredictable levels of binding to plasma proteins, and this may be amplified by the acute-phase response. In addition, heparin has reduced effectiveness against platelet-rich and clot-bound thrombin. In the absence of aspirin, heparin treatment is associated with a lower frequency of refractory angina/myocardial infarction and death (as a combined endpoint) compared to placebo.

Oler and colleagues have conducted a meta-analysis of the influence of adding heparin to aspirin in the treatment of patients with unstable angina. Only six randomized trials were available: there were 55 deaths or myocardial infarctions out of 698 in the aspirin plus heparin arm and 68 out of 655 in the aspirin-alone arm, giving a risk reduction of 0.67 and a 95 per cent confidence interval of 0.44 to 1.02. Thus, these results do not produce conclusive evidence of benefit from adding heparin to aspirin, but it must be stressed that appropriately powered, larger scale trials have not been conducted. Nevertheless, clinical guidelines have adopted unfractionated heparin treatment with aspirin as a pragmatic extrapolation of the available evidence.

Low-molecular-weight heparins versus placebo

The 1996 FRISC trial tested dalteparin against placebo in aspirin-treated patients with unstable angina/non-ST-elevation MI. Some 1506 patients were randomized to receive dalteparin (twice daily for the first 6 days and then once daily at a lower dose for approximately 6 weeks), and the trial showed a highly significant reduction in the frequency of death or new myocardial infarction at 6 days (1.8 per cent versus 4.8 per cent, with a risk ratio of 0.37). The effects were sustained to 42 days but were attenuated at 6 months, the differences no longer maintaining significance. Nevertheless, this trial clearly showed the benefit of low-molecular-weight heparin over placebo, in the presence of aspirin, and the feasibility of administering such treatment over a prolonged time.

Low-molecular-weight heparins possess enhanced anti-Xa activity in relation to anti-IIa (antithrombin) activity compared with unfractionated heparin. They also exhibit decreased sensitivity to platelet Factor 4, have more predictable anticoagulant effect, and lower rates of thrombocytopenia. In view of their enhanced bioavailability they offer the substantial practical advantage of subcutaneous administration, based on a dose per kilogram of body weight, and without the need for laboratory monitoring.

Low-molecular-weight heparin versus unfractionated heparin

Acute-phase treatment (approximately 2 to 8 days)

In the FRIC trial dalteparin was tested against unfractionated heparin in 1400 patients with unstable angina: it had limited power to show a difference, and no significant difference was seen between unfractionated heparin and dalteparin.

The ESSENCE trial was double-blinded and placebo-controlled and tested enoxaparin against unfractionated heparin. The treatments were given for 2 to 8 days (median 2.6 days) and the primary endpoints were death, myocardial infarction, or recurrent angina. Enoxaparin reduced the primary endpoint from 19.6 per cent to 16.6 per cent at 14 days (odds ratio 0.80 and confidence intervals 0.67–0.98). A similar and significant odds ratio was maintained at 30 days and 1 year. At 1 year there were 3.7 fewer events/100 patients (p = 0.022). The study was not powered for death/myocardial infarction alone but demonstrated corresponding trends for these endpoints.

The TIMI 11b trial was also double-blinded and tested enoxaparin *versus* unfractionated heparin, but additionally it examined 72 h of treatment *versus* 43 days of treatment. The results up to 14 days mirrored those seen in the ESSENCE trial: at 14 days the primary outcome occurred was 16.6 per cent (heparin) *versus* 14.2 per cent (enoxaparin), risk ratio 0.85 (p = 0.03). A combined analysis of ESSENCE and TIMI 11b in 1999 indicated an absolute reduction of 3.1 per 100 for death/MI/refractory angina, and showed a similar risk ratio of 0.79 (CI 0.65–0.96) for death and myocardial infarction. Taken together, these findings indicate that short-term treatment with enoxaparin results in about 3 per 100 fewer major cardiac endpoints compared to unfractionated heparin treatment, and this is achieved without additional major bleeding.

Prolonged outpatient treatment

The FRAXIS trial reported in 1999 tested fraxaparin, for 6 or 14 days, against unfractionated heparin. A total of 3468 patients were randomized within 48 h of symptom onset; no difference was seen at 6 days, 14 days, or 43 days, but there was a significant excess of major bleeds with longer term outpatient treatment. In TIMI 11b the curves remained separated over the succeeding treatment interval: at 43 days there were 19.6 per cent events (heparin) *versus* 17.3 per cent (enoxaparin) (p = 0.049), with no evidence of a further separation of the curves. However, only about 60 per cent of the patients entered the chronic-treatment phase of the study and it must be recognized that the study does not exclude a moderate treatment for more prolonged treatment. There was 1.4 per cent absolute excess in major bleeds over the chronic phase.

One component of the FRISC II trial compared long-term *versus* short-term low-molecular-weight heparin (dalteparin) treatment. After 5 days of open-phase treatment with dalteparin, patients were randomized to placebo or weight-adjusted dalteparin for a period of 3 months. The primary endpoint of death/MI occurred in 6.7 per cent of patients at 90 days in the dalteparin arm and 8 per cent in the placebo arm (a non-significant difference). The risk ratio was 0.82 but confidence intervals were between 0.6 and 1.11. The secondary analysis at earlier time points indicated a difference in favour of the low-molecular-weight heparin but this diminished by 3 months.

Conclusions from the low-molecular-weight heparin studies

There is convincing evidence in aspirin-treated patients that low-molecular-weight heparin is better than placebo (FRISC trial). The two trials using enoxaparin have provided consistent data in favour of low-molecular-weight heparin over unfractionated heparin when administered as an acute regimen. The other trials have produced a similar outcome for the acute phase of treatment and it can be concluded that acute treatment is at least as effective as unfractionated heparin. To date, the evidence to support longer term treatment with low-molecular-weight heparin is less convincing. Low-molecular-weight heparins offer significant practical advantages with simplicity of administration, more consistent antithrombin effects, lack of the need for monitoring, and a safety profile similar to that of unfractionated heparin. Evidence supports the following conclusions:

- Low-molecular-weight heparin is superior to placebo in aspirin-treated patients.
- Low-molecular-weight heparin is at least as effective as unfractionated heparin.
- Low-molecular-weight heparin can be used in place of unfractionated heparin and has practical advantages over unfractionated heparin.

Hirudin

Hirudin is a more potent and specific antithrombin than heparin, and large-scale trials have been conducted against unfractionated heparin. A combined analysis of the OASIS-1, OASIS-2, TIMI 9b, and GUSTO IIb trials indicates a 22 per cent relative-risk reduction in cardiovascular death or MI at 72 h, 17 per cent at 7 days, and 10 per cent at 35 days. This combined analysis is significant at 72 h and 7 days and the p value at 35 days is 0.057. Hirudin has specific indications for patients with heparin-induced thrombocytopenia. None of the hirudins are currently licensed in the United Kingdom for the treatment of acute coronary syndromes.

Anti-ischaemic therapy

Specific antithrombotic treatment will have an impact on limiting the progression of occlusion and improving perfusion, hence such treatment has an anti-ischaemic impact. In addition, other pharmacological treatments reduce myocardial oxygen demand and may induce coronary vasodilatation, thus reducing ischaemia. Mechanical revascularization (percutaneous intervention and coronary bypass surgery) also aims to relieve obstruction and reduce a patient's susceptibility to ischaemia and these interventions will be considered separately (see below).

Nitrates

Nitrates act by venodilatation, and in higher dose arteriolar dilatation, and hence reduce preload and afterload, thereby decreasing oxygen demand (see Chapter 15.4.2.2). In addition, nitrates can also induce coronary vasodilatation. They are effective in relieving symptoms of ischaemia. In the acute phase of the syndrome, where dose titration is required, they are most conveniently administered intravenously. Once dose titration is no longer required, buccal, oral, or topical administration is feasible.

The main limitation of continuous administration is the development of tolerance. Increased doses of nitrates may be required, with the dose adjusted on the basis of heart rate, blood pressure response, and relief of symptoms.

Large outcome trials have been conducted with nitrates in acute myocardial infarction but not in the remainder of acute coronary syndromes. However, patients without ST-segment elevation or bundle-branch block were randomized within the ISIS-4 trial. Their mortality was 5.3 per cent for nitrate treatment and 5.5 per cent for placebo treatment, a non-significant difference.

Following acute-phase treatment, patients may be switched to an outpatient oral administration of nitrates. However, if tolerance has been induced in the acute phase, such treatment may have reduced efficacy. Nevertheless, on the basis of current evidence:

- nitrates are effective in reducing ischaemia in the inhospital management of unstable angina/non-ST-elevation MI.

b-Blockers

b-Adrenoceptor antagonists reduce heart rate and blood pressure and myocardial contractility. They are primarily employed to reduce ischaemia in acute coronary syndromes. Large-scale trials have not been conducted in patients with unstable angina or non-Q-wave myocardial infarction. However, in the context of acute myocardial infarction b-blockers reduce mortality by approximately 10 to 15 per cent. They may act be reducing ventricular arrhythmias, reinfarction, and myocardial rupture. A meta-analysis of five trials involving 4700 patients with threatened MI (treated with intravenous b-blockers followed by oral therapy for approximately 1 week) resulted in a 13 per cent reduction in the risk of MI.

b-Blockers may exacerbate acute heart failure. By contrast, recent trials have produced strong evidence of a benefit for the gradual introduction of b-blockers in

ambulant patients with heart failure (see Chapter 15.2.2).

On the basis of current evidence:

- Patients with suspected acute coronary syndromes should be initiated on b-blocker therapy unless contraindicated in the individual case.

Calcium-entry blockers

These agents inhibit the slow inward current induced by the entry of extracellular calcium through the cell membrane, especially in cardiac and arteriolar smooth muscle. They act by lowering myocardial oxygen demand, reducing arterial pressure, and reducing contractility. Some agents induce a reflex tachycardia (nifedipine, nicardipine, amlodipine) and are best administered in combination with a b-adrenoceptor antagonist. By contrast, diltiazem and verapamil are suitable for patients who cannot tolerate a b-blocker because they inhibit conduction through the atrioventricular (**AV**) node and tend to cause bradycardia. All calcium antagonists reduce myocardial contractility and may aggravate heart failure. Calcium-entry blockers have been demonstrated to reduce the frequency of angina in patients with variant angina.

A meta-analysis of calcium-entry blockers in acute coronary syndromes indicates a non-significant trend towards a higher mortality in treated *versus* control patients (5.9 per cent *versus* 5.2 per cent, in 7551 patients). In individual trials, diltiazem has been compared with propranolol and both agents produced a similar reduction in anginal episodes. Subgroup analysis suggests that diltiazem is efficacious in the group with rest angina, but the clinician should always be cautious in extrapolating from subgroup analyses.

- Dihydropyridine calcium-entry blockers should be employed with b-blockers in acute coronary syndromes to avoid reflex tachycardia. In patients unable to tolerate b-blockers, a heart rate-slowing calcium antagonist may be appropriate. Short-acting dihydropyridines should not be used in isolation in acute coronary syndromes.

Potassium-channel activators

These agents (for example, nicorandil) have arterial and venous dilating properties but do not exhibit the tolerance seen with nitrates. They have been shown to be better than placebo in relieving the symptoms of angina, but little convincing evidence exists in comparison with other antianginal agents. Nicorandil possesses both potassium channel and nitrate properties and may be considered as an alternative to nitrate administration.

Conclusions: anti-ischaemic therapy

The following strategy is based upon available clinical and trial evidence:

- Patients with suspected acute coronary syndromes should be initiated on nitrate and b-blocker therapy unless there are contraindications to the use of b-blockers.
- In patients with contraindications to b-blockers, heart rate-slowing calcium antagonists should be employed.
- The combination of a calcium antagonist and b-blocker is superior to either agent alone.
- Angiography and revascularization should be considered in patients with recurrent ischaemia (with ECG abnormalities) or patients with troponin elevation (including non-ST elevation MI).

Revascularization

In chronic, stable angina strong evidence supports the use of surgical revascularization for the relief of symptoms and also for improved prognosis in patients with left main or three-vessel coronary artery disease (especially with left ventricular impairment). Percutaneous revascular *iza*tion (percutaneous intervention, PCI) is primarily employed for the relief of symptoms in chronic stable angina and in patients with one- or two-vessel coronary artery disease. By contrast, until 1999 evidence to support revascular*iza*tion in the acute coronary syndrome was inconclusive. The feasibility of PCI or coronary artery bypass grafting (**CABG**) had been established, but they were associated with an increased risk of complications in comparison with equivalent procedures performed in patients with chronic stable angina.

Observational studies

Large-scale observational studies have demonstrated wide variations between countries in the use of cardiac catheter *iza*tion and revascular*iza*tion for patients with acute ischaemic syndromes (OASIS Registry 1998, GRACE Registry 2001). Unsurprisingly, a direct correlation was demonstrated between the availability of revascular*iza*tion facilities and the frequency with which such procedures were performed. Thus, highest revascularization rates were demonstrated in the United States, with lower rates in Poland and Hungary. By contrast, no significant differences in the rates of death or myocardial infarction were seen, despite rates of invasive procedures of 59 per cent in high revascularization countries *versus* 21 per cent in low revascularization countries. Furthermore, the higher rates of revascularization were associated with an increased frequency of procedural complications, including stroke and major bleeding. These observational data highlighted the importance of performing randomized trials to resolve the role and timing of revascularization in patients with acute coronary syndromes, and to test the impact of adjunctive antithrombotic therapy.

Randomized trial data

Early comparisons of CABG and medical therapy for patients admitted with unstable angina were performed in two studies in the 1970s and 1980s, but these produced inconclusive results. The TIMI IIIB trial was conducted in the early 1990s: 1473 patients were randomized to an early invasive strategy or an early conservative strategy. Unfortunately, the trial was rather underpowered in size and was further underpowered by the high crossover rate from the conservative to invasive strategy (61 per cent revascularization in the invasive arm *versus* 49 per cent in the conservative arm). Mortality or myocardial infarction occurred in 7.2 per cent of patients randomized to the invasive strategy *versus* 7.8 per cent in those randomized to conservative strategy (at 6 weeks), and the corresponding rates at 1 year were 10.8 per cent *versus* 12.2 per cent. These differences were not significant, but the revascularization strategy was supported by a low frequency of hospital readmission. On the basis of this trial, guidelines have suggested that either strategy is acceptable.

The **VANQWISH** study (Veterans Affairs Non-Q Wave Infarction Strategies in Hospital) randomized 916 patients with evolving non- *ST-segment* elevation myocardial infarction. These patients had a high prevalence of comorbidity; moreover, the rate of death or reinfarction at 1 year was 24 per cent in the surgical group *versus* 19 per cent in the medical group (risk ratio of 1.29, $p = 0.05$). There was a high 30-day mortality in those undergoing surgical revascularization, but most of the deaths occurred amongst those randomized to revascularization but in whom the procedure was not performed. Furthermore, the study had a significant crossover rate, with 29 per cent crossing from the conservative to the revascularization arms within 30 days.

The FRISC-II trial compared an invasive strategy with a conservative strategy in patients who were initially stabilized with approximately 6 days of treatment with low-molecular-weight heparin. Coronary angiography was performed within the first 7 days and revascularization performed in 71 per cent of those in the invasive arm and 9 per cent of those in the non-invasive arm within 10 days. This was therefore the first trial to achieve substantial separations in strategy and to include an appropriately powered population. After 6 months, death or myocardial infarction occurred in 9.4 per cent of the invasive group compared with 12.1 per cent of the non-invasive group (a risk ratio of 0.78, $p = 0.031$) and the results remained significant at 1 year. Greatest benefits were demonstrated in higher risk patients.

Can the apparently discordant findings be resolved?

Early trials of revascularization predated modern techniques, and stenting was not performed in the TIMI IIIB study. The VANQWISH trial had no deaths among those undergoing percutaneous revascularization; however, it did demonstrate a high postoperative surgical mortality and a substantial death rate in those assigned revascularization but in whom the procedure was not performed. In addition, the strategy in the 'conservative arm' was more aggressive than in many other studies, in that it aimed to detect ischaemia with nuclear perfusion scanning and undertake revascularization where such tests revealed significant ischaemia. The FRISC-II trial demonstrated the feasibility of a revascularization strategy and a low surgical complication rate. It must be interpreted in the context of an initial stabilization of several days' infusion with low-molecular-weight heparin. Up until 30 days the invasive arm had an excess rate of death or myocardial infarction due to periprocedure

complications. Such complications may be reduced with the use of glycoprotein IIb/IIIa receptor antagonists (used in only 10 per cent of cases in FRISC-II). This strategy was tested in the TACTICS trial where all patients received a GPIIb/IIIa inhibitor (tirofiban) and no early excess hazard was observed in the intervention arm of the trial. The results support the findings of FRISC-II. As discussed above, GPIIb/IIIa antagonists reduce the frequency of peri and postprocedure myocardial infarction in patients with acute coronary syndromes and therefore their use is indicated, especially in those high-risk patients with positive troponins or marked *ST-segment* depression.

In conclusion, an invasive strategy of revascularization can result in a lower frequency of major cardiac complications when performed in patients who are initially stabilized with low-molecular-weight heparin treatment. The FRISC-II trial should not be interpreted as supporting very early revascularization in the absence of an initial stabilization period. The results of FRISC-II are supported by TACTICS indicating that an early invasive strategy is preferable to a conservative strategy in treating higher risk patients with acute coronary syndromes. Although very unstable patients have not been randomized in these trials (for example, those with profound ischaemia or haemodynamic complications) and emergency revascularization may provide their best therapeutic option. It is also important to note that the vast majority of patients in FRISC-II had evidence of ischaemia on the electrocardiogram and most had a positive troponin test: the results should not be extrapolated to low-risk patients, including those without clear-cut ischaemia or without troponin release.

An integrated approach to the patient with unstable angina/non-ST-elevation MI

The at-risk patient

Among patients presenting with an acute coronary syndrome approximately 40 per cent have evidence of prior coronary artery disease (**CAD**) (myocardial infarction, angiographically demonstrated CAD, documented angina with a positive stress test). Appropriate lifestyle, dietary, and non-smoking measures should be introduced for all such patients in addition to the prescription of long-term aspirin (75 mg per day). Implementation of secondary prevention drug treatment will also reduce the risk of subsequent acute coronary syndrome events and deaths, for example lipid-lowering therapy and angiotensin-converting enzyme inhibitors (**ACE** inhibitors).

Access to hospital care

Patients with acute coronary syndromes may present to primary care physicians or directly to emergency hospital services. In addition, 15 to 20 per cent of those presenting directly to chest-pain clinics may have acute coronary syndromes. Patients with previously documented coronary artery disease need specific advice about seeking emergency medical care for episodes of typical anginal pain that persist beyond 20 min at rest, especially if unrelieved by glyceryl trinitrate, or if symptoms are consistent with crescendo angina.

Emergency department triage

For the patient with chest pain, two issues must be resolved urgently:

- Is the chest pain/discomfort thought to be of cardiac origin? This is a clinical judgement and requires prompt and skilled assessment.
- In those with suspected cardiac pain, is there evidence of evolving infarction?

Patients with evolving infarction (*ST-segment* elevation or bundle-branch block and clinical features of infarction) require 'fast-track' reperfusion with thrombolysis or primary angioplasty (see Fig. 1 and below). The remaining patients can be triaged into low- or high-risk categories:

- Patients with typical clinical features of ischaemia and ST-segment depression or transient ST-segment elevation or with cardiac enzyme or troponinelevation are high-risk acute coronary syndrome. Those with elevated levels of cardiac enzymes (troponins or CK, CK-MB) are termed minimal myocardial injury or non-ST-elevation MI.
- Patients with clinical features of acute coronary syndrome and non-specific ECG changes (T-wave inversion, T-wave flattening, minor conduction abnormalities) have intermediate or low-risk acute coronary syndrome (unless enzymes/markers are elevated).
- Patients with a normal electrocardiogram and normal cardiac examination have a potentially low-risk acute coronary syndrome or alternative diagnosis.

Evaluation of patients at intermediate or low risk

Patients who initially seem to be at intermediate or low risk require further assessment to determine their risk status and management. They should be admitted to a cardiac or medical acute assessment area, where further clinical, electrocardiographic, and enzyme assessments will resolve them into relatively high- and low-risk groups.

- Patients with an indeterminate risk and those with a suspected evolving infarction require repeat 12-lead electrocardiography or continuous ST-segment analysis, ideally in a cardiac-care or intensive-care unit setting. Such patients require baseline and repeat troponin estimations to identify those at higher risk (see Fig. 1*).*
- Clinically stable patients with minor or non-specific ECG abnormalities can be separated into those at very low risk on the basis of negative troponins and the absence of diagnostic ECG changes on repeat evaluation. Such patients may nevertheless have significant underlying coronary artery disease. They require stress testing or perfusion scanning, ideally prior to discharge.

Patients with an indeterminate or low risk on clinical grounds can therefore be resolved into those that require further investigation and treatment for acute coronary syndromes (ECG evidence of ischaemia, positive troponins, or positive stress test) and those without (Fig. 4). Follow-up studies have demonstrated that those without significant ECG abnormality, without troponin elevation (at 12 h after the acute event), and with a low-risk stress test have a very low risk of subsequent cardiac events and prompt hospital discharge should be appropriate. In studies from Hamm and colleagues only one such patient out of 850 went on to have a cardiac event.

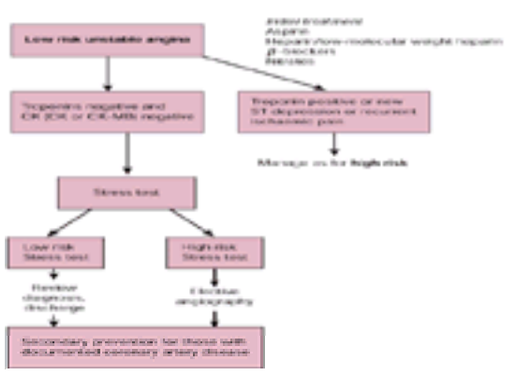

Fig. 4 Uncertain and low-risk unstable angina: initial treatment and diagnostic triage.

Management of patients at high risk

High-risk patients with acute ischaemia at initial presentation, and especially those with haemodynamic compromise, require emergency assessment for possible revascularization. Such patients should also benefit from glycoprotein IIb/IIIa inhibition (Fig. 5). Trial evidence also supports an improved outcome with glycoprotein IIb/IIIa inhibition amongst the remainder of patients with troponin positivity or *ST-segment* depression. Those proceeding to emergency revascularization should receive aspirin, unfractionated heparin, and glycoprotein IIb/IIIa inhibition. Large-scale safety studies have not yet been completed for the combination of low-molecular-weight heparin with glycoprotein IIb/IIIa inhibition, especially in the context of acute revascularization.

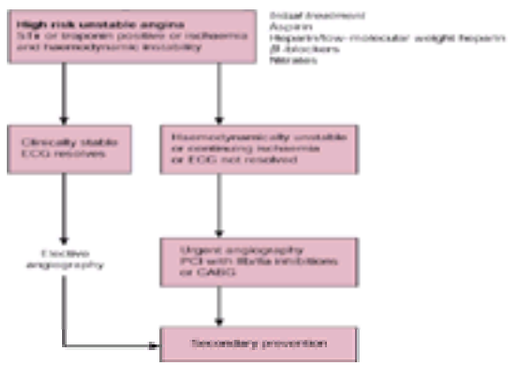

Fig. 5 High-risk unstable angina: initial treatment and diagnostic triage.

Coronary artery bypass surgery

As demonstrated by the FRISC II study, those with three-vessel or left main coronary artery disease and an acute coronary syndrome can be stabilized in the acute phase on low-molecular-weight heparin and aspirin, and can proceed to coronary artery bypass surgery (median 17 days) which carries a low perioperative and postoperative morbidity and mortality in experienced centres (2 per cent, 30-day mortality). A substantial part of the benefits seen in the FRISC II study were amongst those patients undergoing surgical revascularization.

Secondary prevention and rehabilitation

These issues are the same in those patients with unstable angina/non-ST-elevation MI as they are in those with ST-elevation MI (see below for details).

ST-segment-elevation MI

Introduction

In *ST-segment*-elevation myocardial infarction (MI), outcome is critically determined by the extent and severity of myocardial ischaemia. In addition, the eventual extent of irreversibly injured myocardium is influenced by residual myocardial perfusion, the duration of myocardial ischaemia, and cytoprotective mechanisms including preconditioning (see Chapter 15.4.2.1). As a result, the clinical consequences of abrupt coronary occlusion can range from an entirely silent episode, to profound ischaemia with major cardiac rhythm disturbances (ventricular fibrillation or asystole), to acute mechanical decompensation with heart failure. The outcome and management are influenced by the presence or absence of such complications, especially arrhythmias and acute heart failure.

The priorities in the management of *ST-segment*-elevation MI are to relieve acute distress and to limit the extent of infarction, mainly by reperfusion, and to treat complications. Beyond the acute phase, attention focuses on secondary prevention and rehabilitation.

Outcome in ST-segment-elevation MI

Community-based studies in various populations have demonstrated that the case fatality from acute MI is approximately 50 per cent by 1 month after the onset (MONICA studies). Approximately half of the deaths occur within the first 2 h. However, the risks of death, prior to hospitalization, vary with age: 80 per cent of those above 85 years die reaching hospital but only 40 per cent below 55 years. Prior to the introduction of cardiac care units in the 1960s, inpatient mortality was in the range of 25 to 30 per cent and in the 1980s, prior to the introduction of thrombolysis, inpatient mortality averaged approximately 18 per cent. More recently, the MONICA study from five cities has indicated that the 28-day mortality for patients admitted to hospital with a myocardial infarction ranged from 13 to 27 per cent, and other studies have provided figures of 10 to 20 per cent.

A marked discrepancy exists between mortality figures from randomized clinical trials and those from observational studies. Recent thrombolytic clinical trial data have consistently found that the 30-day mortality ranges from 6 to 8 per cent for those randomized within the trials. Some trials have analysed the outcome for individuals ineligible for inclusion and have demonstrated substantially higher death rates. Thus, although clinical trial data are accurate for the populations studied, they do not provide a comprehensive picture of outcome. This is the result of the exclusion of higher risk and complicated patients, including the elderly. In addition, the standards of care achieved in trial centres are not necessarily achieved in routine clinical practice. Although there is relatively little possibility of improving outcome amongst those eligible for randomization in clinical trials, substantial scope does exist in the remainder. Special attention needs to be drawn to the provision of acute resuscitation and defibrillation in the community.

Prehospital care

The priorities in prehospital care are to establish a prompt diagnosis of suspected acute infarction, to treat ventricular fibrillation, and to arrange emergency hospital admission for reperfusion therapy. In rural and other communities with more than a 30-min transfer time to hospital appropriate equipment and training facilities need to be established to allow prehospital thrombolysis to be administered safely and effectively.

The diagnosis of suspected infarction

A working diagnosis of suspected infarction is based upon typical severe chest discomfort of more than 15-min duration and which is unresponsive to glyceryl trinitrate. Characteristically, the pain may radiate to the neck, lower jaw, and arms and is often accompanied by autonomic features including sweating and pallor. Unless complications are present, physical examination may reveal no significant abnormalities, other than those associated with autonomic disturbance, but signs can include tachycardia or bradycardia, the presence of a third or fourth heart sound, and features of heart failure.

The initial electrocardiogram is seldom normal but may not show the classical features of *ST-segment* elevation or the development of Q-waves. Within minutes of the onset of ischaemia hyperacute T-wave changes can be present, and this may be followed by the evolution of characteristic ST-segment elevation, but minor or non-specific ECG abnormalities in conjunction with a characteristic history may signal the early stages of infarction. The working diagnosis relies heavily on the clinical history, and when this suggests myocardial infarction repeat electrocardiography within 30 to 60 min will frequently reveal the evolution of recognizable electrocardiographic changes.

In the prehospital setting a primary care physician may have to rely on the clinical findings to establish the working diagnosis and to initiate immediate treatment. Prompt relief of pain is important, not only for humanitarian reasons, but because pain is associated with sympathetic activation, vasoconstriction, and increased myocardial work. Effective analgesia is achieved by the titration of intravenous opioids, but paramedic crews only have access to non-opioid analgesia. Side-effects of analgesia include nausea and vomiting, hypotension, and respiratory depression. Antiemetics can be administered concurrently; hypotension and bradycardia will usually respond to atropine and respiratory depression to naloxone. Oxygen should be administered, especially to those who are breathless or those with any features of heart failure or shock (see Chapter 16.3 for information on basic and advanced life support in the management of cardiac arrest or ventricular fibrillation).

The logistics of providing acute care for patients with myocardial infarction depend upon the locally available facilities. Guidelines recommend integrated planning involving the emergency care system (ambulance and paramedic personnel), primary care physicians (general practitioners), and hospital-based specialists, including cardiologists and emergency care physicians. Within an urban setting, with relatively short transfer times, the shortest delays and the most prompt initiation of reperfusion occurs when the patient seeks an emergency medical ambulance and direct access to the hospital emergency department.

Prehospital versus inhospital thrombolysis

If patients initially call their primary care physician, this inevitably produces additional delays prior to reperfusion therapy. However, a general practitioner can administer intravenous opioids for the relief of pain. In the ideal scenario, the primary care physician and paramedic ambulance crew arrive together, analgesia is administered, acute complications managed, and the patient transferred rapidly to hospital. Telemetry of the electrocardiogram is possible, with physician-guided thrombolysis administered by paramedic and ambulance crews: the feasibility and safety of this approach has been established in The Netherlands. If a doctor is available in the ambulance, then after assessment and electrocardiography, thrombolysis can be initiated prior to transfer to hospital. In remote settings the feasibility and efficacy of prehospital thrombolysis administered by the general practitioner has been established (GREAT study).

To date, eight trials have been conducted comparing prehospital with inhospital administration of thrombolytic therapy. Depending upon the clinical setting, between 30 and 130 min are saved by prehospital thrombolysis (fibrinolytic drug plus aspirin). Overall, for the complete study population of 6607 patients, the 30-day mortality was 10.7 per cent for those receiving inhospital administration of thrombolysis and 9.1 per cent for those where it was administered prior to hospital admission. This amounts to a 17 per cent relative reduction in early mortality with a *p* value of 0.02 (1.6 per cent absolute reduction). Complication rates were similar for community-treated and hospital-initiated thrombolysis, although ventricular fibrillation occurred more frequently with community administration and necessitated well-trained staff and the availability of defibrillators. The greatest benefit is seen where prehospital treatment is applied in remote settings where transport delays are more than 1 h. Several studies have indicated that about 20 patients with chest pain require evaluation for each patient found to be eligible for thrombolytic therapy in the community. Nevertheless, with appropriate training and facilities prehospital care can provide a gain of approximately 20 lives per 1000 treated, amongst eligible patients.

Prehospital cardiac arrest

The management of prehospital cardiac arrest requires special attention. At least as many lives can be saved by prompt resuscitation and defibrillation as by prompt thrombolysis. For these reasons, emergency assessment of the patient with suspected infarction necessitates that the clinician or paramedic has access to a defibrillator and the skills to manage cardiac arrest promptly and effectively. The provision of basic or advanced life support training to paramedic ambulance crews, together with semiautomatic defibrillators, has resulted in a substantial increase in the number of patients surviving out-of-hospital cardiac arrest. Prior to the institution of such programmes successful resuscitations were opportunistic and often relied on the availability of a medical- or nursing-trained bystander. Nationwide figures indicate that resuscitation now achieves survival in 7 to 10 per cent of those patients found with cardiac arrest and in whom the initial rhythm is thought to be ventricular fibrillation. With effective integrated programmes higher success rates have been achieved. In the south-eastern region of Scotland about 14 per cent survive to reach hospital alive, and in Seattle, with a well-established community training and resuscitation programme, the figure exceeds 20 per cent. Of those reaching hospital alive, approximately half survive to be discharged home.

Emergency inhospital management and patient triage

The priority immediately after arrival at the hospital is to identify those patients with ST-elevation infarction for prompt reperfusion therapy (Fig. 6). The triage is usually performed in a casualty or similar emergency receiving department, but in some institutions patients with a high probability of infarction gain direct access to a cardiac-care assessment area. An integrated strategy involving the paramedic or ambulance system, the emergency physicians, and the cardiologists is required. 'Fast track' systems have been developed to minimize inhospital delay to thrombolysis: these are facilitated by specifically trained medical and nursing staff, with the aim of ensuring clinical assessment and electrocardiography within 15 min of arrival and the institution of thrombolytic therapy within 30 min. Audit programmes and continuous training are necessary for centres to achieve this 30-min median 'door to needle time'. Prior to the advent of 'fast track' systems, door to needle times of between 60 and 90 min were frequently recorded in clinical trials and in observational studies.

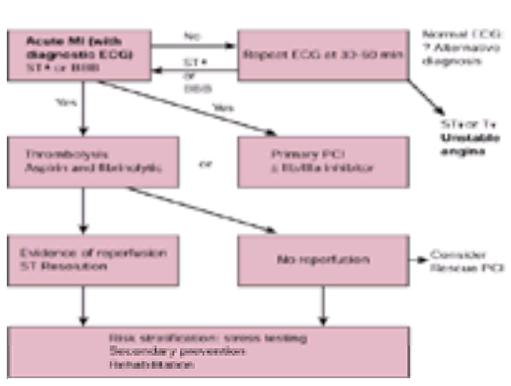

Fig. 6 Management of acute MI.

Definite versus suspected infarction

Rapid triage systems allow the identification of patients with clearly defined clinical and electrocardiographic features of infarction (characteristic symptoms of infarction which persist at rest and are not relieved by glyceryl trinitrate, in the presence of at least 1-mm ST-segment elevation in two or more contiguous leads, or the development of bundle-branch block). Clinical trials have employed ECG criteria of a 1-mm ST elevation for limb leads and 2 mm for chest leads. Although this definition improves specificity it is associated with reduced sensitivity.

Amongst those without diagnostic ECG changes a working diagnosis of suspected myocardial infarction or possible unstable angina can be established. Such patients require repeat clinical and electrocardiographic assessments to detect those with evolving infarction and to separate them from the remainder of patients with unstable angina or non-ST-elevation infarction (see Fig. 1 and Fig. 6). Patients with unstable angina or non-ST-segment elevation infarction do not benefit from thrombolytic treatment, and large-scale clinical trials and meta-analyses have demonstrated that they experience the hazards of bleeding complications from thrombolytic treatment with no evidence of improved survival.

The rationale for minimizing delays to thrombolysis

Experimental and clinical data demonstrate that the duration of ischaemia, prior to reperfusion, is a critical determinant of the eventual extent of myocardial damage. These data are supported by the improved outcome seen with prehospital *versus* inhospital thrombolysis and observational data from large clinical trials in which survival gain diminishes with each additional hour of ischaemia. The Fibrinolytic Trials Overview (Fig. 7) suggests about 1.6 additional deaths per hour of delay per 1000 treated.

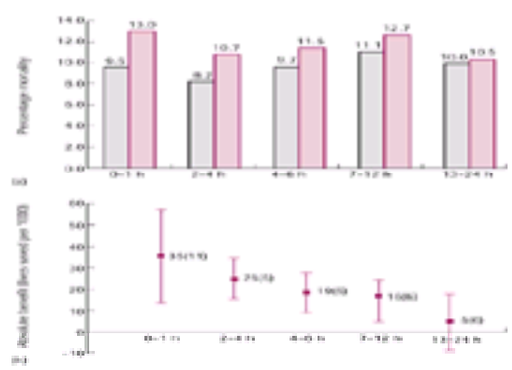

Fig. 7 Effects of thrombolytic therapy on mortality in subsets of patients with suspected MI, according to the time from the onset of symptoms. (a) Mortality rates in the fibrinolytic group (grey bars) versus control groups (pink bars); (b) absolute benefit (lives saved per 1000 treated, standard deviation in parentheses) by time to presentation. (Based on data from the FTT Collaborative Group and reproduced with permission.)

The relationship between the duration of ischaemia and the extent of infarction is non-linear: the greatest potential for salvage occurs when reperfusion is initiated within 60 min of the onset of infarction (Fig. 8). Under such circumstances, a proportion of patients (5 to 7 per cent) will have the infarction aborted and will not develop Q-waves or significant enzyme elevation despite characteristic ST elevation on the initial electrocardiogram. Minimizing the time delay is therefore critical in salvaging myocardium. Based on data from individual trials, and from the Fibrinolytic Trials Overview, most benefit occurs within the first 3 h of the onset of infarction, and highly significant benefits still occur at up to 6 h (Fig. 7). Statistically significant gains are still present at 12 h, but beyond 12 h the benefits are marginal. However, some patients present with a stuttering pattern and in the presence of persistent or intermittent ST-segment elevation and continuing symptoms of ischaemia, reperfusion beyond 12 h may salvage a significant proportion of ischaemic myocardium.

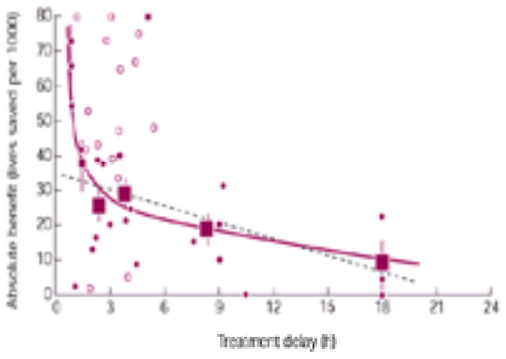

Fig. 8 Absolute benefit in lives saved per 1000 patients treated versus treatment delay. The relationship is non-linear, with most benefit occurring within the first 1 h of symptom onset. Small closed circles, information from trials included in the FTT analysis; open circles, information from additional trials; small squares, data beyond the scale of *x*/*y* cross. The linear (34.7 – 1.6*x*) and non-linear (19.4 – 0.6*x* +29.3x^{-1}) closed regression lines are fitted within these data, weighted by the inverse of the variance of the absolute benefit at each data point. The pink squares denote the average effects in six time-to-treatment groups (areas of the squares are inversely proportional to the variance of the absolute benefits described). (From Boeersma E, Maas ACP, Deckers JW, *et al*. Early thrombolytic treatment in acute myocardial infarction: reappraisal of the golden hour. *Lancet* 1996, **348**, 771–5 and reproduced with permission.)

Differential diagnosis

It is important to remember that thrombolytic therapy or angiography for anticipated primary angioplasty will be of no benefit to those who do not have myocardial infarction. Such patients suffer the dual hazards of thrombolysis or angiography in the acute phase of their illness and the delay in initiating appropriate treatment. Furthermore, those treated inappropriately with thrombolysis will experience the bleeding hazards of the drug (a net increase in intracerebral haemorrhage of approximately 0.5 per cent) and the disrupted coagulation system will render other emergency surgery (for example, for perforated peptic ulceration) more hazardous. Alternative cardiac diagnoses include unstable angina and non-ST- *segment-elevation MI*, myocarditis, pericarditis, and aortic dissection. Non-cardiac diagnoses include gastrointestinal pain of oesophageal, peptic, or biliary origin; pancreatitis; respiratory and musculoskeletal abnormalities.

Aortic dissection presents a particular problem when it extends proximally to the origin of the right coronary artery and produces inferior infarction. Computed tomography, magnetic resonance imaging, or transoesophageal echocardiography may be required to establish the diagnosis (see Chapter 15.14.1).

Transthoracic echocardiography can be valuable when infarction is suspected but characteristic electrocardiographic features are absent. Normal left ventricular function excludes significant infarction. Conversely, a regional contraction abnormality helps to confirm the diagnosis of ischaemia or possible infarction. However, in those with prior myocardial damage the differentiation of new from old mechanical dysfunction is complex and requires specialist assistance.

Cardiac enzymes are helpful when abnormal, but most patients present within 3 h of the onset of symptoms and insufficient time has elapsed to produce a diagnostic release of creatine kinase (CK) or CK-MB, or troponins. Patients with suspected infarction, but normal electrocardiograms, require further clinical electrocardiographic and enzyme estimations 4 to 6 h after the suspected event.

Among elderly and very elderly patients (over 90 years of age) the presentation of infarction is often atypical. They may not experience a typical pattern of symptoms and concomitant multisystem disorders may obscure the diagnosis. Myocardial infarction must be considered in the differential diagnosis of abrupt collapse, haemodynamic disturbance of sudden onset, or severe non-specific symptoms in elderly patients.

Treatment

Thrombolytic treatment

Thrombolytic treatment refers to the combination of antiplatelet therapy (usually aspirin) with fibrinolytic treatment. The fibrinolytic agent directly or indirectly converts plasminogen to plasmin and plasmin lyses fibrin in the clot. Crosslinked fibrin is more resistant to fibrinolytic drugs than a newly formed fibrin clot.

The combination of aspirin and a fibrinolytic agent has undergone extensive clinical testing in trials involving more than 100 000 patients. Additional trials have been conducted comparing one fibrinolytic agent with another. For patients presenting within 6 h of symptom onset, and with ST elevation or bundle-branch block, approximately 30 deaths are prevented per 1000 patients treated. For those presenting between 7 and 12 h, approximately 20 deaths are prevented per 1000 patients treated, and beyond 12 h the benefits are inconclusive.

The ISIS-2 trial demonstrated that the benefits of aspirin treatment were additional to those of fibrinolytic treatment, each achieving about 25 lives saved per 1000 patients treated (for the whole of the study population). Thus, in combination, about 50 lives are saved per 1000 patients treated, but the benefits are larger than this among those presenting within 3 h of infarction with ST-segment elevation or bundle-branch block. Overall, the largest *absolute* benefit is seen in patients at highest risk, although the proportional benefit may be similar for all. High-risk patients include those over 65 years of age, those with a systolic blood pressure below 100 mmHg, and those with anterior infarction or more extensive ischaemia (see primary angioplasty below). The absolute benefit in lives saved per 1000 treated is 11 ± 3 for those under 55 years of age; 18 ± 4 for those between 55 and 64; 27 ± 5 for those 65 to 74; and 10 ± 13 for those over 75 (Fibrinolytic Trials Overview). Similarly, for those patients with more extensive infarction or hypotension, the absolute benefit for the following systolic blood pressures is: 62 ± 18 for less than 100 mmHg; 18 ± 3 for 100 to 149 mmHg; 15 ± 4 for between 150 and 174 mmHg; and 11 ± 8 for those more than 174 mmHg (Fibrinolytic Trials Overview). For patients with bundle-branch block and acute MI, the absolute benefit in lives saved per 1000 treated is 49. For those with anterior ST elevation it is 37, and for inferior ST elevation it is 8. However, for ST depression there is a net hazard of 14 lives lost per 1000 treated, and for those with a normal ECG 7 lives lost per 1000 treated (Fibrinolytic Trials Overview). Thus, evidence supports thrombolysis treatment only for those patients with ST elevation or bundle-branch block.

Hazards of thrombolysis

Thrombolytic therapy is associated with a significant excess of haemorrhagic complications, including cerebral haemorrhage. Overall, about two non-fatal strokes occur per 1000 patients treated, and of these half are moderately or severely disabling. An additional two strokes per 1000 patients are fatal, and the net impact on mortality includes such patients. The risk of stroke increases with age, especially for those over 75 years of age, and for those with systolic hypertension. The excess

of non-cerebral bleeds is about 7 per 1000 treated, including those that require blood transfusion, or are life-threatening, or result in a longer hospital stay. Bleeding occurs at arterial and venous puncture sites, hence blood sampling or cannulation of vessels should be limited to sites where external compression can achieve haemostasis.

Streptokinase and other streptokinase-containing agents (anistreplase) can produce hypotension and, rarely, allergic reactions. Routine administration of hydrocortisone is not indicated. When hypotension occurs it can be managed by interrupting the streptokinase infusion, lying the patient flat or head down and by the administration of atropine, or intravascular volume expansion.

Comparison of thrombolytic agents

The most widely used thrombolytic agents are streptokinase and alteplase (tissue plasminogen activator, **tPA**). The GISSI International Trial and ISIS-3 international trial both failed to find a difference in outcome between streptokinase and tissue plasminogen activator. However, the **GUSTO** trial (Global Utilization of Streptokinase and Tissue plasminogen active for Occluded coronary arteries) employed an accelerated administration of alteplase over 90 min, and intravenous heparin adjusted using the activated partial thromboplastin time (Table 4). The GUSTO trial resulted in 10 fewer deaths per 1000 patients treated with alteplase compared with the streptokinase group. However, this was partially offset by 1 per 1000 additional non-fatal strokes (with residual neurological deficit).

Newer fibrinolytic agents

Novel agents have been developed with the aim of improved clot lysis and simpler administration. Such agents include mutants of native tPA (reteplase, lanoteplase, tenecteplase (TNK-tPA)), and a derivative of streptokinase with increased fibrin specificity and reduced antigenicity, staphylokinase.

Only reteplase, lanoteplase, and tenecteplase have been tested in large-scale comparative trials. Reteplase is administered as two 10-IU boluses given 30 min apart, and the mortality outcome is very similar to that of alteplase (0.23 per cent in favour of alteplase with confidence intervals of –1.11 per cent to +0.66 per cent). The impact on death or disabling stroke is also similar, with stroke occurring in 1.64 per cent of those treated with reteplase and 1.79 per cent of those treated with alteplase (results that are not significant). Lanoteplase or n-PA is a deletion mutant of alteplase that is administered as a single bolus. In the large In-TIME-2 trial 15 078 patients were randomized to lanoteplase versus alteplase, and the results were broadly similar with respect to mortality (6.77 per cent versus 6.60 per cent, respectively), but lanoteplase treatment was associated with slightly more bleeding events. Tenecteplase (TNK-tPA), administered as a single bolus, has been tested against the reference standard of accelerated tPA in a large-scale ASSENT-2 trial involving 17 000 patients. The 30-day mortality figures were virtually identical to those of alteplase (6.18 per cent versus 6.15 per cent, respectively); bleeding events were also similar (intracerebral haemorrhage 0.93 per cent versus 0.94 per cent and blood transfusions 4.25 per cent versus 5.49 per cent, respectively).

Neither staphylokinase nor prourokinase (saruplase) have been tested in large-scale trials, but patency studies have produced encouraging data for both agents.

Conclusion: comparison of thrombolytic agents

The current reference standard for the comparison of fibrinolytic agents is the accelerated infusion regimen of alteplase (tPA). However, streptokinase remains the most widely used fibrinolytic agent internationally. This is largely because the cost of alteplase is substantially higher than that of streptokinase. Newer agents are more convenient to administer than the rather complex infusion regimen of alteplase and have a lower frequency of hypotension and bradycardia than streptokinase. However, there is no evidence that any 'third generation' fibrinolytic agent has improved clinical outcome nor substantially different bleeding complications compared with alteplase. Their main advantage lies in ease of administration. Indeed, recent trials have been associated with a modest but progressive increase in intracerebral bleeding, which may be associated with more aggressive heparin anticoagulation.

In summary:

- The limit appears to have been reached in achieving reperfusion with fibrinolytic agents, and alternative strategies involving adjunctive glycoprotein IIb/IIIa inhibitor therapy have not demonstrated improved outcome or substantially improved safety profile.
- The limit in treating all potentially eligible patients with reperfusion therapy has not been reached. Between 45 per cent (Europe, ENACT study) and 60 per cent (United States, NRMI Registry) of all myocardial infarctions receive neither thrombolysis nor primary angioplasty (percutaneous coronary intervention, **PCI**).
- Thrombolysis is a very cost-effective treatment for acute MI. A sustained benefit on survival has been demonstrated 14 years after thrombolysis.

Combination of fibrinolytic agents with glycoprotein IIb/IIIa inhibitors

The importance of combining antiplatelet agents with fibrinolytic agents was demonstrated in the original ISIS-2 trial, where the benefits of aspirin treatment were additive to those of fibrinolysis. More potent and specific platelet inhibition with glycoprotein IIb/IIIa inhibitors offered the potential for enhanced thrombolysis, and dose-ranging studies were encouraging (TIMI-14).

The large scale randomized trials (ASSENT-S and GUSTO V) have not fulfilled the promise of the earlier studies. In GUSTO V patients were randomized to thrombolytic (reteplase) or the combination of half dose reteplase and abciximab. The combination was not superior to the thrombolytic alone. Nevertheless, there were reductions in secondary points including reinfarction. In ASSENT-S there were significantly fewer endpoints for the combination of abciximab plus thrombolytic (11.1 per cent versus 15.4 per cent) compared with thrombolytic and unfractionated heparin, but a similar benefit was achieved with low- molecular-weight heparin (enoxaparin) and thrombolytic: 11.4 per cent endpoints. There were fewer bleeding events with enoxaparin than with abciximab (major bleeds 3 per cent versus 4.4 per cent). Thus the glycoprotein Hb/IDas do not appear to offer any advantage over the combination of low-molecular-weight heparin and thrombolytic (TNKtPA). A combination ofTNKtPA and enoxaparin appears to provide the best combination of efficacy and safety based on the ASSENT-3 trial.

If the results of ASSENT-3 are confirmed in a separate study, the new standard of care for fibrinolysis will be aspirin plus a fibrinolytic agent plus antithrombin therapy with low-molecular-weight heparin (enoxaparin).

Primary angioplasty

In an attempt to overcome the limitations of fibrinolysis (bleeding hazards and delayed or incomplete reperfusion) studies of primary angioplasty have been undertaken. Primary angioplasty is defined as percutaneous coronary intervention (PCI) without concomitant fibrinolytic therapy. It requires a highly skilled interventional cardiology team with substantial experience of the procedure.

Patients are transferred as an emergency to the cardiac catheterization laboratory and angiography undertaken to establish coronary anatomy and the nature of the vessel occlusion. A flexible guidewire is then passed across the occluded lesion and balloon angioplasty (usually accompanied by stent implantation) performed, thereby restoring patency. Several moderate-sized comparative trials have demonstrated the feasibility of this approach, and in highly skilled centres the data suggest an improved outcome for primary angioplasty compared to conventional thrombolysis. The American College of Cardiology (**ACC**) and the American Heart Association (**AHA**) guidelines for the management of MI recommend primary PCI as an alternative to thrombolysis, provided it can be performed within 60 to 90 min of admission, and by individuals skilled in the procedure (those performing more than 75 cases per annum) and in a high-volume centre (more than 200 cases per annum). However, larger scale studies have been performed (GUSTO 2) and these reveal that the quality of the results achieved, in broad clinical practice, are not as high as those of very experienced interventional centres. In consequence, the differences in outcome between primary angioplasty and thrombolysis are less apparent in such large international comparisons. Nevertheless, primary angioplasty is effective in securing and maintaining coronary patency and avoids the intracerebral bleeding complications of thrombolysis. Randomized trials indicate that, in experienced centres, it is more effective in restoring patency and achieves better ventricular function, with trends towards improved clinical outcome (compared to thrombolysis). Particular gains are seen in haemodynamically compromised patients and those with cardiogenic shock. Thus:

- Primary percutaneous transluminal coronary angioplasty (**PCI**) is specifically indicated in individuals with a contraindication to thrombolytic therapy and in haemodynamically compromised patients.
- In highly experienced interventional centres primary **PCI** may provide an effective alternative to thrombolysis.

The choice of reperfusion strategy will clearly continue to depend upon available resources and the expertise of each centre. Unless primary angioplasty can be performed within approximately 1 h of hospital admission, the potential gains do not appear to outweigh the hazards when compared with thrombolysis.

Angioplasty combined with thrombolysis

The combination of angioplasty and thrombolytic therapy has proved disappointing in a number of trials, with a tendency to an increased risk of complications. Although many of these trials predate current instrumentation and drug therapy, the routine combination of angioplasty and thrombolysis is not recommended.

Rescue angioplasty

Patients in whom thrombolysis fails to achieve reperfusion may benefit from 'rescue' angioplasty. Such patients can be identified by the failure to resolve ST-segment elevation, in combination with persistent clinical features of ischaemia, with or without haemodynamic compromise. Limited experience from relatively small randomized trials suggests a trend towards an improved benefit from rescue angioplasty.

Coronary artery bypass surgery

In the acute phase of myocardial infarction the role of coronary artery bypass grafting (**CABG**) is limited to those patients with acute mechanical complications, such as ventricular septal defect or mitral regurgitation due to papillary muscle rupture. Unless such mechanical complications are present the hazards of acute bypass surgery are significantly increased compared to delayed revascularization in a stabilized patient. The Danish DANAMI study investigated the role of revascularization in those with ischaemia during the recovery phase of myocardial infarction. It suggested that, following infarction, individuals with symptomatic or electrocardiographic ischaemia on stress testing experience significant long-term benefit from surgical revascularization.

Later inhospital management

The main aims of later inhospital management are the:

- identification and treatment of acute complications of infarction;
- identification of patients at increased risk for subsequent cardiac events; and
- initiation of secondary prevention and rehabilitation.

The distinction between early treatment and later phase treatment is clearly arbitrary. Major complications may be apparent at the time of presentation and haemodynamic, arrhythmic, or ischaemic complications may be evident shortly thereafter. Nevertheless, in the period beyond the first 12 to 24 h it is appropriate to focus attention on the points listed above.

Identification and treatment of complications of infarction

Failure of reperfusion

Electrocardiographic markers of failed reperfusion are the persistence of ST-segment elevation together with clinical and haemodynamic features of continuing ischaemia. Continuous computed ST analysis allows the most accurate definition of ECG changes, but an approximation can be obtained with repeated 12-lead ECGs and measurement of ST-segment elevation.

- In those with successful reperfusion, ST segments decrease to less than 50 per cent of peak elevation within 60 min.

In addition, some patients exhibit reperfusion arrhythmias (ventricular tachycardia, idioventricular rhythm, and, rarely, ventricular fibrillation). Such arrhythmias are more common in the presence of marked ischaemia and prompt reperfusion within 60 to 90 min of occlusion.

The most effective treatment for failed reperfusion has not yet been validated in large-scale clinical trials. However in those without successful reperfusion:

- rescue angioplasty is feasible, consisting of mechanical recanalization of the occluded vessel with percutaneous intervention, often accompanied by stent implantation.

This strategy achieves an 'open artery' and may be associated with improved mechanical function and improved longer term prognosis.

Repeat thrombolysis

A thrombolytic agent may fail to achieve recanal*iza*tion as a result of an extensive organized thrombus (crosslinked fibrin), platelet-rich thrombi, and mechanical obstruction to flow due to intraplaque haemorrhage. In addition, among patients treated with streptokinase, previous exposure to a streptococcus or to streptokinase can induce an antibody response that can block at least half of the 1.5 million unit standard dose of streptokinase. Thus, previous exposure to streptokinase or known antistreptococcal antibodies are indications to use an alternative thrombolytic in the first instance. Feasibility studies have been undertaken of partial- or full-dose alteplase following (unsuccessful) streptokinase, but large-scale outcome trials are needed to establish the safety and efficacy of this strategy.

Cardiogenic shock, left ventricular dysfunction, and heart failure

Cardiogenic shock

In cardiogenic shock mechanical contractile abnormalities of the left ventricle or acute haemodynamic complications (papillary muscle rupture or ventricular septal defect) lead to reduced blood pressure and impaired tissue perfusion. Clinically, the condition is recognized by a systolic blood pressure of less 90 mmHg together with impaired tissue flow, as reflected by oliguria, impaired cerebral function, and peripheral vasoconstriction. Between 5 and 20 per cent of those patients admitted to hospital with acute MI demonstrate cardiogenic shock. There is evidence that its frequency has been reduced by thrombolytic therapy and primary PCI. The mortality rate when cardiogenic shock complicates an acute coronary event is in excess of 70 per cent if acute revascular *iza*tion is not possible.

Time delay is critically important in the management of cardiogenic shock: mortality rises progressively if more than 2 h have elapsed since its onset. Treatment aims to improve the recovery of acutely ischaemic myocardium (mechanical and surgical revascular *iza*tion) and to support the circulation with a combination of inotropes, vasodilators, and loop diuretics. Evidence suggests that the most important treatment may be to reopen the infarct-related artery with either thrombolysis or primary angioplasty. Once cardiogenic shock has developed observational studies of primary angioplasty have demonstrated improved outcome (PAMI trial). In addition, the SHOCK trial has demonstrated that aggressive treatment with intra-aortic balloon pumping (**IABP**) followed by surgical revascular*iza*tion may also significantly reduce mortality.

Aside from attempts to induce reperfusion, management of the patient with cardiogenic shock after myocardial infarction traditionally includes inotropes. Dopamine is commonly used, initially at a low 'renal dose' (1–5 μg/kg per min) that activates dopaminergic receptors (but also has an effect on the circulation), but if necessary at higher doses of 5 to 20 μg/kg per min that have positive inotropic and chronotropic effects. In doses above 20 μg/kg per min there is activation of a-adrenoceptors with undesirable peripheral vasoconstriction and a decline in renal perfusion. Dobutamine acts mainly as a b1-adrenoceptor agonist and is used in the range of 2 to 40 μg/kg per min. Phosphodiesterase inhibitors have both inotropic and vasodilator effects and, although they have produced favourable haemodynamic responses, the studies conducted have not shown an improvement in outcome.

The management of pulmonary oedema consists of opiates (to relieve distress and to reduce vascular resistance), oxygen, vasodilators, and diuretics. Vasodilators (including nitrates, salbutamol, and sodium nitroprusside) reduce venous and pulmonary arterial pressure, but tachycardia may be a limiting feature and their use is

limited in those who are profoundly hypotensive. Loop diuretics are employed in bolus intravenous doses or by infusion.

Left ventricular dysfunction and heart failure

Large-scale trials of angiotensin-converting enzyme (ACE) inhibitors have been conducted in patients with left ventricular dysfunction and those with clinical and radiological features of heart failure (see Chapter 15.5.3). Clear evidence demonstrates the improved short- and long-term outcome with ACE inhibitors in patients with heart failure and those with asymptomatic left ventricular dysfunction.

- In patients with left ventricular dysfunction or heart failure, ACE inhibitors improve the short- and long-term prognosis.

Caution must be exercised with the introduction of ACE inhibitors in patients with intravascular volume depletion: they can cause hypotension. ACE inhibition should commence with very small doses (for example, 6.25 mg of captopril) with dosages increased progressively in conjunction with clinical monitoring. They can provoke deterioration in renal function in patients with renal artery stenosis, which is not uncommon in those with atheromatous coronary disease. Hence, it is important to check serum electrolytes and creatinine during follow-up. Angiotensin receptor blockers appear to provide similar benefits to those seen with ACE inhibitors (ELITE 2 study), but most evidence exists for ACE inhibition.

Arrhythmias

A wide variety of arrhythmias can be seen in the context of acute myocardial infarction and its treatment. In the early days of coronary-care units, great emphasis was placed on the treatment of even minor rhythm disturbance. This is now thought to have been misplaced: antiarrhythmic agents are almost invariably negatively inotropic and they may also be proarrhythmic in the context of acute coronary ischaemia. An overview of randomized trials into the use of prophylactic lidocaine (lignocaine) showed that it increased mortality. Ventricular fibrillation should be treated with DC cardioversion, and treatment with lidocaine or other antiarrhythmics (for example, amiodarone) reserved for those who have had ventricular fibrillation or another symptomatic arrhythmia (see Chapter 15.6 for details of the diagnosis and treatment of arrhythmias).

Heart block

Heart block of any degree can occur after acute myocardial infarction. It is more common with inferior than anterior infarction because the right coronary artery supplies the atrioventricular node, and also because vagal reflexes are more likely from this area. It is often transient, and does not necessarily imply a large infarct, except when it occurs with anterior infarction, in which case the prognosis is grave. Temporary pacing is justified when bradycardia compromises the circulation, but not advocated 'prophylactically' for a first- or second-degree block.

Ventricular septal defect, papillary muscle rupture, and myocardial rupture

Rupture of the interventricular septum occurs in up to 3 per cent of acute infarctions and is responsible for about 5 per cent of deaths due to myocardial infarction. Rupture in the apical area may complicate anterior infarction and in the basal inferior area may complicate inferior infarction. Clinically, the condition is associated with the development of a new pansystolic murmur and clinical features of a left to right shunt with increased pulmonary congestion. Surgery should be undertaken as soon as possible: the outlook for those who are not operated upon is very bleak, few survive. However, a very few patients with small shunts survive the acute phase, but they may suffer the later consequences of the shunt.

Papillary muscle rupture occurs as a result of acute ischaemic damage due to obstruction of either the left anterior descending or circumflex coronary arteries. It causes the abrupt onset of severe mitral regurgitation and accounts for 5 per cent of deaths after acute MI. The complication generally occurs within the first week after infarction and may be recognized as the abrupt onset of acute pulmonary oedema. It is often accompanied by a new systolic murmur, but when the left atrial pressure rises acutely the murmur may be insignificant. The management is acute surgical repair with or without revascular *iza*tion (for further discussion of acute mitral regurgitation, see Chapter 15.7).

In the patient who deteriorates haemodynamically after myocardial infarction—with hypotension, pulmonary oedema, or both—it is important to consider the possibility of a ventricular septal defect or acute mitral regurgitation. However, it can be impossible to distinguish between the two on clinical grounds. Both classically produce a new pansystolic murmur, and although differences between the murmurs have been described, these are not robust enough to discriminate with certainty in the individual case. Acute mitral regurgitation is best diagnosed by echocardiography, but transthoracic echocardiography may be unable to detect a ventricular septal defect in a reliable manner. Transoesophageal echocardiography is better, as is the use of a contrast-enhanced technique. If unavailable, an alternative approach is to pass a flow-directed pulmonary catheter and take blood samples from the pulmonary artery, right ventricle, and right atrium. A step up in oxygen tension between the right atrium and the pulmonary artery indicates the presence of a left to right shunt and confirms the diagnosis of a ventricular septal defect.

Myocardial rupture may follow acute infarction, usually involving the free wall of the left ventricle. It is responsible for approximately 10 per cent of all deaths in acute MI. Half of the ruptures occur within the first week, and 90 per cent within 2 weeks. The location of rupture is usually within the infarcted area, but may be at the junction with adjacent normal myocardium. In most cases death is immediate and due to electromechanical dissociation. The patient is unresponsive to resuscitation measures but, rarely, with subacute rupture, patients can be supported until surgical repair is performed. The diagnosis is made on clinical and echocardiographic criteria with assessment for possible cardiac tamponade (see Chapter 15.9). In some patients, partial rupture of the free wall can result in the late development of a false aneurysm.

Left ventricular thrombus

A left ventricular thrombus can be detected using echocardiography in up to 40 per cent of patients with acute anterior MI. The thrombus is usually located at the apex in association with a dyskinetic or aneurysmal section of myocardium with impaired contractile function. The thrombus may be large and is associated with risks of embol*iza*tion (in 15 to 20 per cent of cases). Anticoagulation with heparin followed by warfarin is advised in patients with extensive infarction and those in whom apical aneurysms or mural thrombi are detected. Both thrombolysis and surgical repair have been successfully conducted. However, there is no clear evidence that either strategy is superior (provided there is no evidence of embol*iza*tion).

Pericarditis

Pericarditis is a frequent complication of transmural myocardial infarction and may be manifest clinically as a pericardial friction rub accompanied by pleuritic chest pain. A small pericardial effusion may be detected using echocardiography. Dressler's syndrome is associated with pericarditis between 2 weeks and 3 months after acute infarction and has an autoimmune basis, often accompanied by pleural and pericardial effusions. It is managed with salicylates or non-steroidal anti-inflammatory agents. The frequency of both pericarditis and Dressler's syndrome is reduced with acute reperfusion (see Chapter 15.9 for further discussion).

Shoulder hand syndrome

This is a syndrome of rheumatic pain in the left shoulder, with restricted movement, which can occur in the weeks after myocardial infarction. The pathogenesis is unknown. It is treated symptomatically and usually resolves spontaneously.

An integrated approach to the management of ST-segment-elevation MI

Management of the at-risk patient

Secondary prevention measures in patients with documented coronary artery disease reduce the frequency of subsequent infarction (aspirin, lipid lowering, ACE inhibitors, cessation of smoking, treatment of hypertension). Preventive strategies in those at risk of subsequent infarction can reduce the frequency of both MI and sudden cardiac death.

Prehospital management

In a patient with suspected acute infarction prehospital management aims to treat acute arrhythmic complications, including ventricular fibrillation and other forms of cardiac arrest, to provide analgesia and oxygen, and to minimize delays to reperfusion. Prehospital thrombolysis may be given where transfer times to hospital exceed 30 min, and appropriate facilities exist with trained paramedic crews.

Early inhospital management

Initial assessment involves the identification of those with clear-cut evidence of infarction (based on clinical and diagnostic ECG criteria). Such patients require immediate triage to reperfusion therapy (thrombolysis with a fibrinolytic agent plus an antiplatelet agent or primary PCI (Table 5). The remaining patients in whom the diagnosis of MI is suspected but the ECG criteria are not diagnostic should be managed in an intensive-care setting (in the Emergency Department or CCU) with repeat ECG evaluation at 30-min intervals (or ST-segment analysis). Cardiac enzymes may be elevated when the index episode of pain occurs more than 6 h prior to the obtained blood sample. Such patients may be resolved into those with evidence of non-ST elevation infarction (ECG and enzyme or troponin elevation) and those with unstable angina (T-wave inversion, ST-segment depression, or transient ST-segment elevation, without elevated cardiac troponins). Among those with minor or non-specific ECG changes and no enzyme elevation, re-evaluation should take place for alternative diagnoses, and stress testing performed subsequently to detect underlying coronary artery disease (Fig. 1 and Fig. 3).

Management of the later inhospital phase

During this phase the management of complications, initiation of secondary prevention, and early cardiac rehabilitation should take place. In high-risk patients (those with recurrent acute ischaemia or those with failure of ST-segment resolution and continuing pain) emergency angiography and possible revascular *iza*tion can be performed in appropriately equipped centres (Fig. 6).

Regular clinical and electrocardiographic assessments are required during the recovery phase to detect acute mechanical and arrhythmic complications, and to identify impaired contractile function in patients who will benefit from ACE inhibitor treatment. ACE inhibitor treatment is indicated in those with overt heart failure in the acute phase. Based on the HOPE trial, patients with documented vascular disease benefit substantially from ACE inhibition (ramipril 10 mg). Cardiovascular deaths are reduced from 8.1 per cent to 6.1 per cent and myocardial infarction from 12.3 per cent to 9.9 per cent (relative risk reductions 0.73 and 0.80 respectively, p <0.001 in both instances). Treating 1000 patients with ramipiril for 4 years prevents approximately 150 cardiovascular events in 70 patients. Thus, ACE inhibition is indicated for those with vascular disease irrespective of whether there is evidence of overt heart failure or impaired LV function in acute phase. Prior to discharge, patients also require assessment for lipid-lowering therapy (current evidence suggests that all patients with MI will benefit unless their total cholesterol concentration is below 4.8 mmol/l or their low-density lipoprotein (LDL) concentration is below 3.0 mmol/l). There is evidence to support management of diabetes with glucose and insulin during the in-hospital and early post-hospital phase (see Chapter 12.12.1).

All patients will benefit from smoking cessation and the management of hypertension (systolic pressure to less than 140 mmHg). Dietary advice is required for those with a basal metabolic index (**BMI**) above 25 kg/m^2.

Summary of secondary prevention measures in those with unstable angina and myocardial infarction

Following an acute coronary syndrome, patients require dietary and lifestyle advice including the support necessary to discontinue smoking (including nicotine replacement therapy). Lipids should be measured within the first 24 h of admission and evidence supports the use of lipid-lowering therapy with statins in almost all patients, excepting those with very low LDL or total cholesterol levels. Based on data from the HOPE study, individuals with documented coronary artery disease have reduced long-term risks of death and myocardial infarction if maintained on ACE inhibition. In addition, such patients may require antianginal therapy and all should receive long-term, low-dose aspirin.

Non-pharmacological interventions

Evidence supports the following non-pharmacological interventions in secondary prevention:

- cessation of smoking (including the avoidance of passive smoking);
- dietary modification;
- exercise;
- rehabilitation; and
- management of obesity.

Modification of high-risk conditions

Trial evidence supports therapeutic interventions to modify the following conditions:

- hyperlipidaemia;
- left ventricular dysfunction and heart failure;
- diabetes mellitus;
- hypertension.

Pharmacological interventions

Evidence (summarized in Table 6 and Table 7) supports the following therapies to reduce the risk of subsequent cardiovascular events:

- antiplatelet therapy (usually aspirin in a dose of 75 mg/day);
- b-blockers in those without contraindications;
- lipid lowering with 3-hydroxy-3-methylglutaryl coenzyme A (**HMG CoA**) reductase inhibitors (statins);
- ACE inhibitors in those with left ventricular dysfunction and heart failure, and based on the results of the HOPE study, in other patients with vascular disease (Table 6).

Anticoagulants

These are indicated in those with high risks of embolism due to left ventricular or atrial thrombus. There is evidence to support the use of anticoagulants in post-MI patients but no definitive evidence that such treatment is superior to aspirin therapy.

Hormone replacement therapy (HRT)

HRT is associated with a reduced risk of coronary heart disease in observational studies, but in the only randomized study (HERS) it resulted in no overall reduction in the risk of non-fatal MI or CHD death.

Calcium-channel blockers

An overview of data from 19 000 patients based on all randomized trials of acute infarction and unstable angina suggests that the available calcium-channel blockers are unlikely to reduce the rate of subsequent infarct development, infarct size, or subsequent infarction. They may, however, have indications for the relief of angina (especially heart-rate lowering calcium antagonists).

Antiarrhythmic agents

A review of the effects of antiarrhythmic agents (with the exception of b-blockers) does not demonstrate a beneficial impact on mortality.

Further reading

Antman EM, *et al*. (1996). Cardiac-specific troponin I levels to predict the risk of mortality in patients with acute coronary syndromes. *New England Journal of Medicine* **335**(18), 1342–9. [Troponin levels are key determinants of mortality risk in acute coronary syndromes.]

Antman EM, *et al*. (1999). Assessment of the treatment effect of enoxaparin for unstable angina/non-Q-wave myocardial infarction. TIMI IIB–ESSENCE meta-analysis. *Circulation* **100**, 1602–8. [Combined analysis of TIMI IIB and ESSENCE (enoxaparin) indicating 20 per 1000 fewer death/MIs compared with unfractionated heparin.]

ASSENT-2 Investigators (1999). Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. *Lancet* **354**, 716–22. [Tenecteplase and alteplase have almost identical outcomes in acute MI.]

ASSENT-3 Investigators (2001). Efficacy and safety of tenecteplase in combination with enoxaparin, abciximab, or unfractionated heparin: the ASSENT-3 randomised trial in acute myocardial infarction. *Lancet* **358**, 605–13.

Bode C, *et al*. (1999). Randomised comparison of coronary thrombolysis achieved with double-bolus reteplase (recombinant plasminogen activator) and front-loaded, accelerated alteplase (recombinant tissue plasminogen activator) in patients with acute myocardial infarction. *Circulation* **94**, 891–8. [Reteplase and alteplase have an almost identical outcome in acute MI.]

Braunwald E. (1989). Unstable angina: a classification. *Circulation* **80**, 410–14. [Classification of unstable angina based upon mode of presentation and time course (excludes ECG changes and predates troponins).]

Cannon CP, *et al*. (2001). Comparison of early invasive and conservative strategies in patients with unstable coronary syndromes treated with the glycoprotein IIb/IIIa inhibitor tirofiban. The TACTICS-Thrombolysis in Myocardial Infarction 18 Investigators. *New England Journal of Medicine* **344**, 1879–87.

CAPTURE Investigators (1997). Randomised placebo-controlled trial of abciximab before and during coronary intervention in refractory unstable angina: the CAPTURE study. *Lancet* **349**, 1429–35. [Reduced cardiac events with abciximab before and following PCI (mostly MI).]

Cox J, Naylor CD (1992). The Canadian Cardiovascular Society grading scale for angina pectoris: is it time for refinements? *Annals of Internal Medicine* **117**, 677–83. [A scoring system for the severity of angina pectoris.]

CURE (Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial) Investigators (2001). *New England Journal of Medicine* **345**, 494–502. [Key study demonstrating impact of clopidogrel, with aspirin, in reducing deaths, MI, and stroke.]

DANAMI—Madsen JK, *et al*. (1997). Danish multicentre randomised study of invasive versus conservative treatment in patients with inducible ischaemia after thrombolysis in acute myocardial infarction (DANAMI). *Circulation* **96**, 748–55. [Improved outcome with revascularization for spontaneous or exercise-induced ischaemia.]

ELITE II—Pitt B, *et al*. (1999). Effects of losartan versus captopril on mortality in patients with symptomatic heart failure: rationale, design, and baseline characteristics of patients in the Losartan Heart Failure Survival Study—ELITE II. *Journal of Cardiac Failure* **5**, 146–54. [Equivalent effects of angiotensin receptor blocker (losartan) versus ACE inhibitor (captopril).]

ESSENCE—Cohen MD, *et al.* (for the Efficacy and Safety of Subcutaneous Enoxaparin in Non-Q-Wave Coronary Events Study Group: ESSENCE) (1997). A comparison of low-molecular-weight heparin with unfractionated heparin for unstable coronary artery disease. *New England Journal of Medicine* **337**, 447–52. [First study to demonstrate the superiority of low-molecular-weight heparin (enoxaparin) over unfractionated heparin in unstable coronary artery disease.]

Fibrinolytic Therapy Trialists' (FTT) Collaborative group (1994). Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. *Lancet* **343**, 311–22. [Definitive combined analysis of fibrinolytic trials with more than 1000 patients and outcome.]

Fox KAA (1999). Comparing trials of glycoprotein IIb/IIIa receptor antagonists. *European Heart Journal Supplements* **1**(Suppl R), R10–R17. [Analysis of trials of glycoprotein IIb/IIIa antagonists in acute coronary syndromes.]

FRAX.I.S. Study Group (1999). Comparison of two treatment durations (6 days and 14 days) of a low molecular weight heparin with a 6-day treatment of unfractionated heparin in the initial management of unstable angina or non-Q wave myocardial infarction: FRAX.I.S. (FRAxiparine in Ischaemic Syndrome). *European Heart Journal* **20**, 1553–62. [Fraxiparine and unfractionated heparin have similar outcomes in unstable angina or non-Q-wave MI.]

FRISC Study Group—Lindhal B, Venge P, Wallentin L (1996). Relation between troponin T and the risk of subsequent cardiac events in unstable coronary artery disease. *Circulation* **93**, 1651–7. [Increased risks with troponin elevation in unstable coronary artery disease.]

FRISC II. FRagmin and Fast Revascularisation during inStability in Coronary artery disease (FRISC II) Investigators (1999). Invasive compared with non-invasive treatment in unstable coronary artery disease: FRISC II prospective randomised multicentre study. *Lancet* **354**, 708–15. [Revascularization after initial stabilization with low-molecular-weight heparin results in improved outcome compared with conservative management.]

Gandhi MM, Lampe FC, Wood DA (1995). Incidence, clinical characteristics, and short-term prognosis of angina pectoris. *British Heart Journal* **73**, 193–8. [Community-based study on the incidence and characteristics of new-onset angina pectoris.]

GISSI (Gruppo Italiano per lo Studio Della Streptochinasi Nell'Infart Miocardico) (1988). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *The Lancet* **i**, 397–402. [The first large-scale study of thrombolysis in acute MI: marked survival advantage with streptokinase treatment.]

GRACE—Foxkaa *et al*. (2001). Management of acute coronary syndromes. Variations in practice and outcomes: findings of the Global Registry of Acute Coronary Events (GRACE). *European Heart Journal* (in press).

GREAT—Rawles J, *et al.* (1994). Halving of mortality at 1 year by domiciliary thrombolysis in the Grampian Region Early Anistreplase Trial (GREAT). *Journal of the American College of Cardiology* **23**, 1–5. [Reduced mortality following domiciliary thrombolysis compared to delayed hospital thrombolysis (3 h time separation).]

GUSTO Investigators (1993). An international randomised trial comparing four thrombolytic strategies for acute myocardial infarction. *New England Journal of Medicine* **329**, 673–82. [Key large-scale international comparison of streptokinase and alteplase in acute MI: approximately 10 per 1000 survival advantage for alteplase; small excess of haemorrhage.]

GUSTO-IIb Investigators—Savonitto S, *et al.* (1997). Prognostic value of the admission electrocardiogram in acute coronary syndromes. Results from the GUSTO-IIb trial. *European Heart Journal* **18**(Suppl.), **335**, 5–82. [The importance of the admission electrocardiogram in determining prognosis in acute coronary syndromes.]

GUSTO-IIb—Armstrong PW, *et al*. (1998). Acute coronary syndromes in the GUSTO-IIb trial: prognostic insights and impact of recurrent ischemia. The GUSTO-IIb Investigators. *Circulation* **98**, 1860–8. [Prognostic factors based upon the GUSTO-IIb trial.]

GUSTO-V Investigators (2001). Reperfusion therapy for acute myocardial infarction with fibrinolytic therapy or combination reduced fibrinolytic therapy and platelet IIb/IIIa inhibition: the GUSTO V trial. *Lancet* **357**, 1905–14.

Hamm CW, *et al*. (1992). The prognostic value of serum troponin T in unstable angina. *New England Journal of Medicine* **327**, 146–50. [The important prognostic value of troponin in acute coronary syndromes.]

Held PH, Yusuf S, Furberg CD (1989). Calcium channel blockers in acute myocardial infarction and unstable angina: an overview of randomized trials. *British Medical Journal* **299**, 1187–92. [Combined analysis of calcium channel blockers.]

HERS–NHANES—Herrington DM (1999). *Erratum*, Comparison of the Heart and Estrogen/Progestin Replacement Study (HERS) cohort with women with coronary disease from the National Health and Nutrition Examination Survey III (NHANES). *American Heart Journal* **138**, 800. [First published in *American Heart Journal* 1998, **136**, 115–24]. [No advantage for HRT in this first randomized trial of women with coronary disease.]

HOPE Study Investigators (The Heart Outcomes Prevention Evaluation) (2000). Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on death from cardiovascular causes, myocardial infarction, and stroke in high-risk patients. *New England Journal of Medicine*, **342**, 145–53. [Marked and sustained impact of ACE inhibitor (ramipril) on survival and cardiac and cardiovascular events among patients with vascular disease. No benefits from vitamin E.]

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction. *Lancet* **ii**, 349–60. [Landmark thrombolysis trial of streptokinase and aspirin demonstrating a survival advantage with either but an additive benefit with both .]

ISIS-3 (Third International Study of Infarct Survival) Collaborative Group (1992). A randomised comparison of streptokinase *vs* tissue plasminogen activator *vs* anistreplase and of aspirin plus heparin *vs* aspirin alone among 41 299 cases of suspected acute myocardial infarction. *Lancet* **339**, 153–70. [Comparison of streptokinase versus tPA versus anistreplase in acute MI, similar outcome (tPA

regimen differed from that of GUSTO, duteplase versus alteplase and slower administration).]

ISIS-4 (Fourth International Study of Infarct Survival Collaborative Group) (1995). A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. *Lancet* **345**, 669–85. [Impact of ACE inhibitor in acute MI (modest benefit in 5 per 1000 treated). No benefit demonstrated with mononitrate nor with magnesium.]

Kong DF, *et al*. (1998). Clinical outcomes of therapeutic agents that block the platelet glycoprotein IIb/IIIa integrin in ischemic heart disease. *Circulation* **98**, 2829–35. [Pooled analysis of glycoprotein IIb/IIIa inhibitors in acute coronary syndromes, including those undergoing percutaneous intervention.]

Lewis WR, Amsterdam EA (1994). Utility and safety of immediate exercise testing of low-risk patients admitted to the hospital for suspected acute myocardial infarction. *American Journal of Cardiology* **74**, 987–90. [Benefits and safety of early exercise testing in low-risk patients with suspected MI.]

Maas ACP, *et al*. (1999). Sustained benefit at 10–14 years follow-up after thrombolytic therapy in myocardial infarction. *European Heart Journal* **20**, 819–26. [Sustained survival benefit following thrombolytic therapy.]

MONICA—Tunstall-Pedoe H, *et al*. (1996). Sex differences in Myocardial Infarction and Coronary Deaths in the Scottish MONICA Population of Glasgow 1985–1991. *Circulation* **93**, 1981–92. [Prospective study of myocardial infarction in death in a community-based population.]

OASIS—Yusuf S, *et al.* for the Organisation to Assess Strategies for Ischaemic Syndromes Registry Investigators (1998). Variations between countries in invasive cardiac procedures and outcomes in patients with suspected unstable angina or myocardial infarction without initial ST elevation. *Lancet* **352**, 507–14. [Outcome and clinical characteristics in a prospective registry of acute coronary syndromes.]

Oler A, *et al*. (1996). Adding heparin to aspirin reduces the incidence of myocardial infarction and death in patients with unstable angina. A meta-analysis. *Journal of the American Medical Association* **276**, 811–15. [Pooled analysis suggests benefit of adding heparin to aspirin in unstable angina, individual trial data not significant and overall result marginally significant.]

PAMI-I—Nunn CM, *et al*. (1999). Long-term outcome after primary angioplasty, Report from the primary angioplasty in myocardial infarction (PAMI-I) trial. *Journal of the American College of Cardiology* **33**, 640–6. [Long-term outcome of primary angioplasty.]

PCI-CURE Study—Mehta *et al*. (2001). (Clopidogrel in unstable angina to prevent recurrent events trial Investigators). *Lancet* **358**, 528–33.

Pocock SJ, *et al*. (1995). Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet* **346**, 1184–9. [Combined analysis of all the randomized trials (up to 1995) comparing coronary angioplasty and bypass surgery.]

PRAIS-UK Investigators—Collinson J, *et al.* (2000). Clinical outcomes, risk stratification and practice patterns of unstable angina and myocardial infarction without ST elevation: Prospective Registry of Acute Ischaemic Syndromes in the UK (PRAIS-UK). *European Heart Journal* **21**, 1450– 7. [12 per cent rate of death/MI and 30 per cent rate of death/MI, refractory angina at 6 months in those presenting with unstable angina or non-ST elevation MI.]

PRISM. The Platelet Receptor Inhibition in Ischemic Syndrome Study Investigators (1998). A comparison of aspirin plus tirofiban with aspirin plus heparin for unstable angina. *New England Journal of Medicine* **338**, 1498–505. [Tirofiban versus heparin in unstable angina.]

PRISM-PLUS. The Platelet Receptor Inhibition in Ischemic Syndrome Management in Patients Limited by Unstable Signs and Symptoms Study Investigators (1998). Inhibition of the platelet glycoprotein IIb/IIIa receptor with tirofiban in unstable angina and non-Q-wave myocardial infarction. *New England Journal of Medicine* **338**, 1488–97. [Improved outcome with tirofiban in unstable angina (high-risk population).]

Ravkilde J, *et al*. (1995). Independent prognostic value of serum creatine kinase isoenzyme MB mass, cardiac troponin T and myosin light chain levels in suspected acute myocardial infarction. Analysis of 28 months of follow-up in 196 patients. *Journal of the American College of Cardiology* **25**, 574–81. [Prognostic value of cardiac enzymes in suspected acute MI.]

Ryan TJ (1999). Early revascularisation in cardiogenic shock—a positive view of a negative trial. *New England Journal of Medicine* **341**, 687–8. [Beneficial trends for early revascularization in cardiogenic shock.]

Sivers F (1999). Evidence-based strategies for secondary prevention of coronary heart disease, 2nd edn. A&M Publishing, Guildford, Surrey. [Systematic and comprehensive analysis of evidence based strategies for secondary prevention in CHD.]

TIMI III—Braunwald E, *et al*. (1994). Effects of tissue plasminogen activator and a comparison of early invasive and conservative strategies in unstable angina and non-Q-wave myocardial infarction, results of the TIMI III trial. *Circulation* **89**, 1545–56. [No significant difference in death/MI for early invasive versus conservative strategy in unstable angina (but fewer hospital readmissions).]

TIMI-III—Cannon CP, *et al*. (1995). Prospective validation of the Braunwald classification of unstable angina: results from the Thrombolysis in Myocardial Ischemia (TIMI) III Registry. *Circulation* **92**, 1-19. [Outcome in relation to the Braunwald classification.]

TIMI-14—Antman EM, *et al*. (1999). Abciximab facilitates the rate and extent of thrombolysis: results of the thrombolysis in myocardial infarction (TIMI) 14 trial. *Circulation* **99**(21), 2720–32. [Abciximab plus half-dose tPA results in angiographic opening rates similar to primary angioplasty (approximately 70 per cent at 60 min).]

TRIM Study group—Luescher MS, *et al.* (1997). Applicability of cardiac troponin T and I for early risk stratification in unstable coronary disease. *Circulation* **96**, 2578–85. [Similar data for troponin T and troponin I in risk stratification.]

VANQWISH—Boden WE, *et al*. (1998). Outcomes in patients with acute non-Q-wave myocardial infarction randomly assigned to an invasive as compared with a conservative management strategy. Veterans Affairs Non-Q-Wave Infarction Strategies in Hospital (VANQWISH) Trial Investigators. *New England Journal of Medicine* **38**(25), 1785–92. [Increased mortality with an invasive strategy compared to a conservative strategy .]

White HD, Van de Werf FJ (1998). Thrombolysis for acute myocardial infarction. *Circulation* **97**, 1632–46. [Review of thrombolysis for acute myocardial infarction.]

Yusuf S, *et al*. (1985). b-blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Disease* **27**, 335–71. [Combined analysis demonstrating the benefits of b-blockers during and after acute MI.]

15.4.2.4 Percutaneous interventional cardiac procedures

Edward D. Folland

Introduction

The birth of interventional vascular medicine is generally credited to Charles Dotter, a radiologist from Portland, Oregon, who in 1964 first dared to relieve atherosclerotic stenosis of a patient's femoral artery by passage of a percutaneously introduced dilator. Although Dr Dotter had a few notable successes, which were widely publicized in the lay press, the scientific community scorned him. His radical concept lay dormant until a decade later when Andreas Gruentzig, a young German radiologist studying in Zurich, revived it. Dr Gruentzig was convinced that percutaneous dilatation of atherosclerotic stenosis was a sound concept and proposed that Dotter's solid dilator be replaced by a catheter with an inflatable cylindrical balloon at its tip. Using catheters he created in his own kitchen, he proceeded carefully and logically in applying his technique first to animal models, then to human peripheral vessels, and finally in 1977 to his ultimate goal, the human coronary artery. News of Gruentzig's percutaneous transluminal coronary angioplasty (**PTCA**) was quickly embraced by the medical community, and the era of percutaneous coronary intervention (**PCI**) was born. This chapter will deal with percutaneous approaches to treating coronary, valvular, and congenital heart disease.

Percutaneous coronary intervention

Percutaneous coronary intervention or PCI is the current general term applied to a variety of percutaneous, catheter-based procedures that accomplish revascularization either by angioplasty (enlargement of a vessel lumen by modification of plaque structure), stenting (deployment of an internal armature or stent), atherectomy (removal or ablation of plaque), or thrombectomy (removal of thrombus). Several different devices have been developed to perform these procedures. The interventional cardiologist chooses among these approaches to best suit the particular requirements of each individual patient.

Indications

The indications for percutaneous revascularization have expanded dramatically during the past 25 years. In the early days of PTCA it was indicated for subtotal, proximal occlusions of single vessels in patients with chronic, stable angina pectoris who had failed medical therapy. As experience grew and equipment improved, patients with unstable angina, total occlusions, bypass grafts, multivessel disease, and acute myocardial infarction were added to the list. Currently, the most common single indication for PCI is acute coronary syndrome (unstable angina or acute myocardial infarction).

PCI has traditionally been performed only in hospitals having cardiac surgical backup. However, as the procedure has become safer and the need for emergency bypass surgery less common (currently under 1 per cent of all cases), it has become more common, particularly in Europe, for these procedures to be performed in facilities where surgical backup is not on-site. Likewise, all patients undergoing PCI were once required to be potential candidates for bypass surgery in case of failure of the percutaneous procedure. Now some patients who are poor surgical candidates may undergo salvage intervention as their best or only avenue for revascularization. The choice of initial treatment (pharmacological, interventional, or surgical) for patients with each of the above coronary syndromes has been guided by evidence from a number of randomized clinical trials and will be treated in more detail in the later section headed 'Outcomes'.

Devices

Balloon angioplasty

Balloon angioplasty is the traditional, basic technique of coronary intervention, although it is now uncommonly employed as a stand-alone treatment. Nevertheless, it is fundamental to the deployment of coronary stents, which are currently the most widely utilized of the interventional devices. The equipment for angioplasty is displayed in Fig. 1 and consists of a coaxial array of guiding catheter, balloon catheter, and steerable guidewire. The procedure is accomplished by first engaging the left or right coronary orifice with the tip of the guiding catheter to access the vessel containing the target lesion and to provide backup support during advancement of the guidewire and balloon across the lesion (Fig. 2(a)). Next, the guidewire is advanced through the guide catheter into the appropriate vessel and across the lesion to be treated. Typical guidewires are 14-thousandths of an inch in diameter (about 0.36 mm) and have a flexible spiral coil tip that can be directed by rotating its proximal end outside the body. The balloon catheter is then advanced over the guidewire until the deflated balloon lies across the target lesion. Finally, the balloon is inflated with a solution of dilute contrast medium to a pressure sufficient to expand the cylindrical balloon to its nominal manufactured diameter (Fig. 2(b)). The balloon size is selected to match the estimated diameter of the nearest segment of normal vessel and the length of the target lesion. Sometimes intravascular ultrasound is used to assist in this choice. The balloon is then withdrawn and the result assessed by angiography and, occasionally, by ultrasound (Fig. 2(c)).

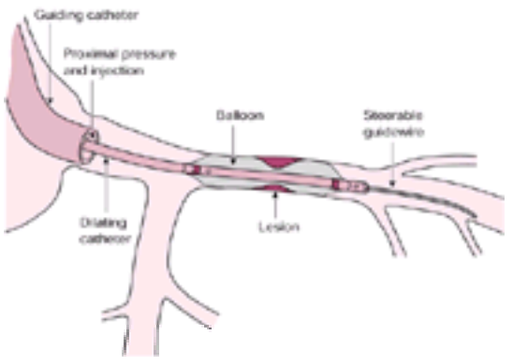

Fig. 1 Balloon angioplasty. The guiding catheter gives access to the coronary artery and provides a platform against which the dilating apparatus can be advanced. The steerable guidewire is passed down the vessel being treated and provides a rail over which the balloon catheter can be advanced. Once centred on the atherosclerotic lesion, the balloon is inflated under pressure to dilate the narrowed segment of artery.

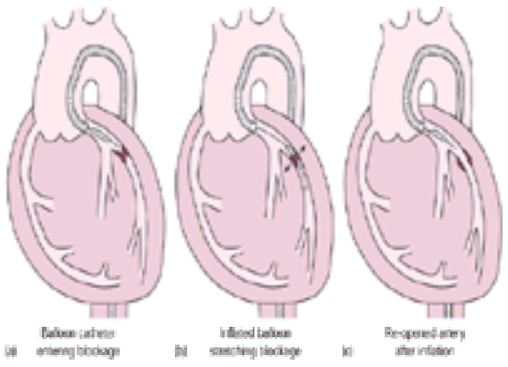

Fig. 2 A typical lesion before (a), during (b), and after (c) balloon angioplasty.

Traditional angioplasty now finds its chief application in deployment of balloon expandable stents. However, angioplasty may serve as a stand-alone interventional technique for the treatment of lesions of small vessels (less than 2.5 mm in diameter) and lesions located far distally or beyond tortuous segments where more rigid devices such as stents cannot reach. In experienced hands, with appropriate case selection, the initial success rate of balloon angioplasty should exceed 95 per cent. Abrupt closure of the vessel might be expected in about 3 per cent of cases (usually due to dissection), but the majority of these can be corrected by deployment of a stent, resulting in a need for emergency bypass surgery in less than 1 per cent of cases. The clinical consequence of vessel closure is often insufficient to justify surgery in vessels too small or distal for stenting.

The technology of guide, balloon, and guidewire systems has advanced to the point where few locations in the coronary anatomy are inaccessible. Totally occluded vessels can usually be successfully crossed with appropriate manipulation of the right guidewire, enabling successful angioplasty. The success rate for angioplasty of totally occluded vessels depends upon the age, length, and composition (thrombus versus plaque) of the occlusion; it is well over 90 per cent in cases of acute thrombotic occlusion, and over 50 per cent in cases of chronic occlusion (longer than 3 months).

The chief disadvantage of balloon angioplasty is the phenomenon of restenosis, which will be discussed in more detail later in this chapter, and which spurred the development of newer devices in the hope of circumventing restenosis.

Stenting

Stenting has become the intervention of choice in approximately 90 per cent of cases undergoing PCI. The term 'stent' is believed to originally derive from the name of an eighteenth-century British dentist who devised a compound for creating impressions of human teeth. A modern-day vascular stent is actually an armature, or internal skeleton, for restoring and maintaining the cylindrical structure of a diseased vessel. Most stents are made from a thin-walled, stainless-steel tube in which slots have been carved. The slotted tube is then mounted securely on a deflated angioplasty balloon and deployed at the target lesion of the coronary artery by inflating the balloon at high pressure with dilute contrast medium. When the balloon is deflated the stent remains expanded against the vessel wall, its slots stretched into diamond-shaped apertures (Fig. 3). Approximately 20 per cent of the vessel wall is covered by metal, the remainder being an intra-strut aperture. This accounts for the surprisingly high patency of side branches following stent deployment, and the ability to access these side branches when necessary for further intervention. A variation of the slotted-tube stent is a balloon-deployed coiled wire. Many recent stent designs are hybrids, which incorporate desirable properties of both the slotted-tube and coiled-wire designs.

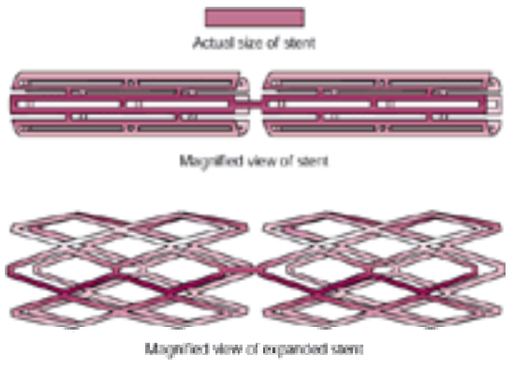

Fig. 3 A balloon-deployed coronary artery stent before (a) and after (b) deployment.

A somewhat different approach to stenting is the self-deploying, coiled-wire stent called the Wallstent. A coiled wire made from nitinol or another alloy with shape-retaining characteristics is compressed into a tubular delivery sheath, which is advanced over a guidewire across the target lesion. Once in its proper position the sheath is drawn back, allowing the stent to expand to its original size and shape (Fig. 4). As with slotted-tube stents, pre- or postdeployment dilation with a balloon may be necessary depending upon the nature of the lesion treated and the device used. Although it is one of the original stent designs, the self-expanding stent is used less commonly for coronary artery applications, but still finds use in many peripheral vascular cases.

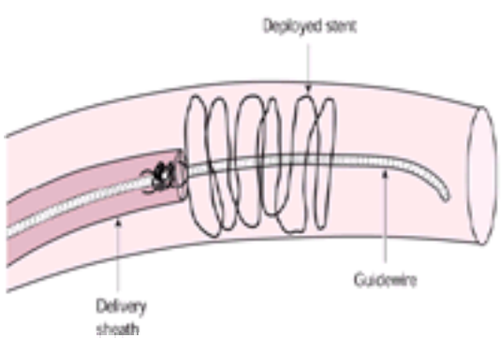

Fig. 4 A self-deploying coil stent. The stent unfurls as its delivery (containment) sheath is pulled back.

Stents have gained remarkable popularity mainly for three reasons. First, immediate complications are reduced because abrupt closure of the vessel due to dissection is less likely, emphasized by the fact that a stent is the best treatment for a balloon-induced dissection. Second, the immediate result is better in terms of the diameter and smoothness of the lumen, which turns out to be of more than cosmetic value because the early gain in lumen size relates directly to the late outcome. Finally, stents have been demonstrated in randomized clinical trials to be effective in reducing the likelihood of late restenosis.

However, stents do have some disadvantages, which include their propensity to subacute thrombosis, the persistence of some degree of restenosis (depending upon the size of the vessel and length of the lesion), and the fact that they cannot be deployed under some circumstances. Subacute thrombosis, a complication unique to stents, usually occurs within 10 days after stent deployment. By contrast to restenosis, which is a gradual phenomenon, stent thrombosis is usually sudden,

presenting as acute myocardial infarction and requiring emergency revascularization, usually by balloon angioplasty. The likelihood of subacute thrombosis has been reduced to less than 3 per cent by antiplatelet therapy with a combination of aspirin and clopidogrel or ticlopidine.

Directional coronary atherectomy

Directional coronary atherectomy (**DCA**) is achieved with a device illustrated in Fig. 5, which utilizes a rotating cylindrical blade that is advanced across an open aperture near the tip of a cone-shaped, guidewire-directed catheter. Opposite the aperture is an eccentric balloon, which when inflated compresses plaque of the opposite vessel wall into the aperture, where it is cut away by the rotating blade and pushed into the nose cone. The direction of the aperture can be rotated so that slices of plaque are removed in a radial fashion by multiple cuts taken at different locations around the circumference of the vessel. The catheter can then be withdrawn and the excised plaque removed from the nose cone. The catheter may be reintroduced, if necessary, for more atherectomy.

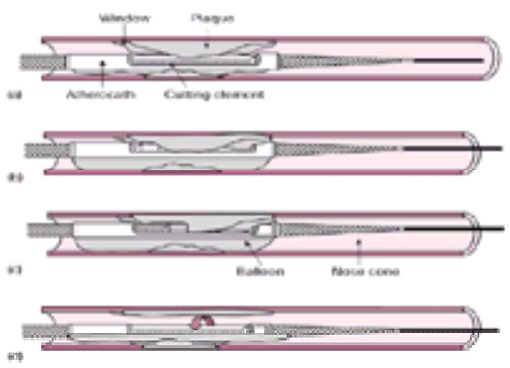

Fig. 5 Directional coronary atherectomy. The rotating cylindrical blade is advanced across the window of the housing and cuts away plaque, packing it into the nose cone.

Although DCA was originally devised with the hope of reducing the incidence of restenosis, it has failed to outperform balloon angioplasty in most circumstances. It has therefore assumed the role of a 'niche' technology, which is useful in particular situations such as very eccentric proximal lesions, and lesions involving the ostia of major side branches. Removal of plaque at branch points seems to reduce the likelihood of plaque shifting from one branch to another as the respective lesions are dilated with balloons or stents. DCA has the disadvantage of requiring a rather large, stiff device, limiting its application to proximal lesions of large vessels. Furthermore, the removal of plaque seems to have surprisingly little effect on restenosis. DCA is currently employed in less than 5 per cent of interventional cases.

Rotational ablation (Rotablator)

Rotational ablation (Rotablator) is a method of pulverizing plaque into particles smaller than the size of a capillary, which wash away with the circulating blood. This process is accomplished by means of a diamond-studded burr, which rotates at approximately 150 000 revolutions per min (Fig. 6) and is advanced along a guidewire into the plaque. The diamond studs on the forward face of the olive-shaped burr selectively cut into hard substances such as plaque and calcium, sparing the soft surface of normal tissue. During rotational atherectomy a vasodilating solution is infused into the artery proximal to the burr to prevent spasm and to maintain maximal coronary flow, which carries away particulate debris. Burrs are manufactured in sizes ranging from 1.5 mm to 2.5 mm in diameter. Atherectomy often requires the use of two or three burrs of progressively larger size until an adequate lumen size is achieved. Although occasionally used as a stand-alone procedure, rotational ablation is usually employed to 'debulk' lesions prior to final dilatation with a balloon or stent.

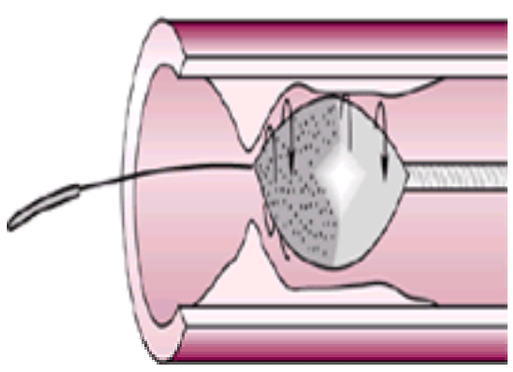

Fig. 6 Rotational atherectomy. The rotating burr pulverizes plaque as it is advanced over the guidewire into the lesion.

Like directional atherectomy, rotational ablation was originally conceived as a potential solution to the problem of postintervention restenosis. Unfortunately, it too has failed to outperform balloon angioplasty in this regard and has also assumed the role of a 'niche' device for special situations. It is most commonly used to treat restenosis within previously deployed stents, but it also finds application in the treatment of heavily calcified lesions that do not respond well to balloons and stents. It is also useful in treating diffuse, osteal, and bifurcating lesions. The frequency with which rotational ablation is employed varies by operator, but averages 5 to 10 per cent of most centres' cases.

Rotational ablation has the disadvantage of being a relatively expensive addition to other interventional modalities. It is unable to adequately increase the lumen of large vessels, and is contraindicated in lesions containing thrombus. Due to its tendency to transiently decrease contractility during the ablation process, it is also relatively contraindicated in patients whose left ventricular function is severely impaired.

Cutting balloon

The cutting balloon has several tiny longitudinally mounted blades which become erect when the balloon is inflated and create linear cuts along the vessel wall. This balloon is preferred by many cardiologists for initial treatment of stent restenosis. Its use is often followed by brachytherapy in order to reduce further the likelihood of another episode of stent restenosis.

Brachytherapy

The local, catheter based delivery of beta or gamma radiation is currently considered the best method for prevention of recurrent episodes of stent restenosis. Radiation is delivered with the assistance of a radiation therapist after initial treatment of stent restenosis with a cutting balloon, Rotablator, or conventional balloon. The benefit of brachytherapy appears to be limited to treatment of stent restenosis. It is not recommended following initial deployment of a stent. Brachytherapy also prolongs the period of risk for subacute thrombosis, making it necessary to treat patients with both aspirin and clopidogrel for at least 6 months following treatment.

Drug eluting stents

Polymer coated stents can be used locally to deliver drugs which inhibit the vessel's proliferative response to injury and therefore reduce the likelihood of stent restenosis. Drugs which have shown promise in preliminary trials include sirolimus (Rapamycin), taxol (and its derivative paclitaxel), and actinomycin.

Other devices

Distal protection devices are methods of capturing and collecting thrombus and other debris that may embolize distally from the target lesion during the use of many of the interventional tools mentioned above. They may be particularly beneficial during the treatment of old, degenerated vein grafts in which distal embolization is especially common.

Excimer laser coronary atherectomy (**ELCA**) employs a guidewire-directed fibreoptic catheter to deliver bursts of excimer laser energy to the plaque. Disintegrated plaque washes away in the circulation. However, ELCA has also failed to solve the restenosis problem, is used uncommonly at most centres, but remains the sole surviving member of a number of laser applications that have been tried and failed over the past 25 years. It finds its most frequent application in treatment of osteal lesions, stent restenosis, and diffuse calcified disease. Due to limitations of fibre size it is usually followed by balloon or stent treatment.

The transcutaneous excision catheter (**TEC**) device was developed at approximately the same time as directional coronary atherectomy and employs a rotating conical blade, which cuts away plaque and clot as it is advanced over a guidewire. The resulting debris is sucked back through the catheter into a reservoir outside the body. Although originally developed as an atherectomy device, it has found its chief application in treating clot-laden lesions. Nevertheless, it has not gained wide usage. More recently devised approaches to clot removal are the AngioJet and Excisor devices. The AngioJet uses the Venturi effect from a high-velocity jet of water, which draws thrombus into a window near the tip of a guidewire-directed catheter and propels it into a reservoir. The Excisor employs a helical screw at the end of a catheter, which breaks up the clot so that it can be withdrawn through the catheter. Both these devices currently find their chief application in the treatment of degenerated and clot-laden vein-graft lesions.

Complications

Percutaneous coronary intervention exposes the patient to all the potential complications of cardiac catheterization presented in Chapter 15.3.6. In addition, it carries the risk of other complications unique to interventional procedures. Most of these complications stem from four general processes, which account for most of the adverse outcomes from coronary artery intervention: abrupt closure, distal embolization, subacute stent thrombosis, and restenosis. When considering PCI for a patient, it is important to weigh the likelihood of these adverse outcomes against the expected chance of adverse events without intervention. The approximate frequencies of various specific complications from percutaneous coronary intervention are listed in Table 1. As in diagnostic catheterization, the likelihood of these complications is also dependent upon patient characteristics and operator skill.

Abrupt closure and distal embolization

Abrupt closure and distal embolization account for most of the immediate complications of PCI, especially acute myocardial infarction and emergency coronary artery bypass surgery. Dissection, spasm, and thrombosis are the leading causes of abrupt closure. The availability of stents has reduced the need for emergency bypass surgery to less than 1 per cent, because these are an effective treatment for acute dissection in most cases. Nevertheless, dissection sometimes extends with the addition of each stent, and occasionally the stent itself can be the cause of dissection at one of its edges. Acute thrombosis may occur in spite of routine prophylactic treatment with heparin and aspirin: glycoprotein IIb and IIIa inhibitors may stop this process and are often given prophylactically, especially in high-risk cases. Incomplete stent deployment seems to be a leading cause of thrombotic occlusion.

Distal embolization is surprisingly uncommon, except when patients have acute coronary syndromes or visible thrombus. It is especially troublesome for patients with degenerated or thrombus-laden vein grafts. Embolization may result in discrete occlusion of branch vessels or the phenomenon called 'no reflow', which is manifest by reduced flow without identifiable occlusion and thought to be due to capillary plugging from showers of microemboli.

In any case, either abrupt closure or no reflow often results in some degree of myocardial infarction. The frequency of this complication is a matter of how it is defined. Non-Q-wave infarction indicated only by a rise of creatine kinase enzyme is more common than Q-wave infarction.

Subacute thrombosis

Subacute thrombosis is a complication unique to stents, which occurs between 2 and 10 days following intervention and is manifest by acute myocardial infarction. It is a medical emergency that should be managed in a fashion similar to spontaneous acute infarction. Emergency reperfusion by balloon angioplasty is usually preferred, unless a catheterization laboratory is unavailable, in which case thrombolytic therapy is recommended. In the early days of stenting this complication occurred in over 3 per cent of cases in spite of vigorous anticoagulation including intravenous heparin and warfarin. This treatment required several days of hospital stay for the initiation of warfarin therapy and delayed the widespread acceptance of stenting. However, once the current treatment consisting of oral antiplatelet agents was proven to be superior, the length of hospital stay and local bleeding complications were reduced, and the use of stents grew rapidly. Subacute thrombosis now occurs in approximately 1 to 3 per cent of cases.

Restenosis

Restenosis remains the 'Achilles heel' of coronary intervention. In patients undergoing balloon angioplasty the likelihood of restenosis at 6 months following intervention is between 30 and 50 per cent if defined by angiographic stenosis of 50 per cent or greater, and approximately 25 per cent if defined by the clinical recurrence of symptoms. The use of stents has reduced the angiographic rate of restenosis to approximately 25 per cent and the clinical rate to as little as 10 per cent. The risk of restenosis varies depending upon the individual circumstances of each case: factors associated with restenosis include long lesions, small vessels, and suboptimal initial results.

Restenosis typically presents clinically as exertional angina at 1 to 6 months following intervention: if it is not present at 6 months, it is unlikely to occur. It is a gradual phenomenon, caused by the proliferation and migration of smooth muscle cells into the lumen of the treated vessel. Stents have been effective in reducing restenosis because they eliminate elastic recoil and generally result in a large lumen. However, smooth muscle cell migration is triggered by any form of vascular injury and takes place after stenting as well as other interventions. Attempts at reducing restenosis by pharmacological and mechanical means other than stenting have been largely unsuccessful. Brachytherapy, described above, is successful in reducing smooth muscle cell proliferation following repeat intervention. Other promising approaches under investigation include local gene therapy and local drug therapy delivered by coated stents or 'leaky' balloons.

Outcomes

Chronic stable angina

Randomized clinical trials have shown that patients with single- and double-vessel disease experience a more rapid and complete resolution of symptoms, and a greater improvement in treadmill exercise performance, when treated by balloon angioplasty rather than by pharmacological therapy for chronic stable angina pectoris. However, this comes at the price of a greater likelihood of repeat intervention or bypass surgery at 6 months, largely due to the need to treat restenosis. Nevertheless, the rate of bypass surgery becomes equal in both groups by 3 years.

When compared to coronary bypass surgery, angioplasty provides similar relief of symptoms and similar rates of mortality and myocardial infarction at 5 years' follow-up, with the exception of diabetic patients who have somewhat better 5-year survival rates when treated surgically. Otherwise, the main difference between patient groups randomly assigned to surgery or angioplasty is that repeat catheterization or revascularization is less frequent for those having surgery. Again, this difference is largely due to the effect of restenosis in the angioplasty group. Few of the interventionally treated patients in these trials received stents, so the likelihood of repeat procedures might be expected to be less using current devices.

Unstable angina

The choice between initial aggressive treatment (catheterization and revascularization) and initial conservative treatment (medical therapy with catheterization and revascularization only for those who have continued evidence of ischaemia) for patients with unstable angina has been controversial. However, one of the most recent

studies favours an aggressive approach to these patients.

Acute myocardial infarction

Percutaneous intervention has been shown to be an effective treatment for acute myocardial infarction, both as a salvage procedure after failed thrombolytic therapy and as a direct, initial approach to reperfusion. Randomized trials have shown that direct intervention is superior to initial thrombolytic therapy when performed in centres with expert interventionists and catheterization facilities that are available around the clock. In the general community setting this advantage has not yet been proven. In any case, direct PCI is the treatment of choice for patients in whom thrombolytic therapy is contraindicated and for patients who are haemodynamically unstable.

Economic considerations

The cost of equipment and supplies for percutaneous coronary procedures may become a limiting factor, particularly in developing countries and in healthcare systems with stringent budgets. Most catheters, guidewires, and other supplies are intended for one-time use. Expendable supplies alone cost approximately £500 ($US 800) for a simple balloon angioplasty procedure. That cost may be multiplied severalfold as multiple stents and Rotablator burrs are added to the list. Nevertheless, the cost of a single percutaneous revascularization procedure usually remains less than that of a comparable coronary bypass operation. However, when the added cost of repeat percutaneous revascularizations necessitated by restenosis is considered, the price difference between the two therapeutic approaches narrows.

Percutaneous balloon valvuloplasty

Allain Cribier of France developed the treatment of valvular stenosis by means of balloon catheters in the 1980s. The clinical utility of the procedure depends upon the valve treated and the age of the patient.

Mitral stenosis

Balloon valvuloplasty of the mitral valve has become the treatment of choice for selected patients with rheumatic mitral stenosis. The concept is similar to the aortic valvuloplasty illustrated in Fig. 7. The most common approach to the mitral valve is via trans-septal puncture of the left atrium from percutaneous access of the right femoral vein. After passing a stiff guidewire with a curved soft tip across the mitral valve, an appropriately sized balloon is centred on the valve and inflated with dilute contrast medium, tearing open the fused commissures and allowing the valve to open more normally. A dumbbell-shaped balloon, named after Dr Inoue, is often utilized, preventing the balloon from slipping off the valve during inflation. Clinical improvement, complications, and durability of the outcome from balloon mitral valvuloplasty have been shown to be comparable to surgical commissurotomy in appropriately selected patients. To be a candidate for balloon mitral valvuloplasty a patient must have no evidence of thrombus in the left atrium. Other features which auger poorly include immobility of the valve leaflets, severe calcification, thickening of the chordae tendineae, and more than mild regurgitation.

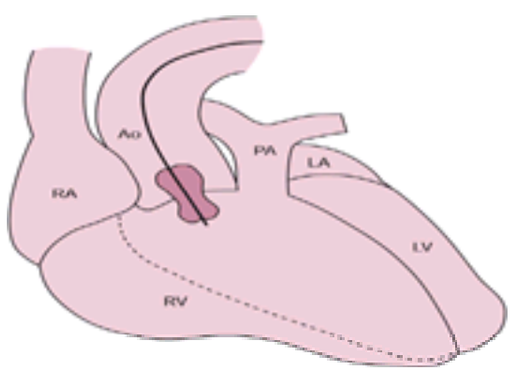

Fig. 7 Percutaneous balloon aortic valvuloplasty. The balloon is centred on the stenotic valve and inflated to tear open the fused commissures.

Aortic stenosis

Experience with balloon valvuloplasty for patients with aortic stenosis has been disappointing, largely due to an almost universal tendency for the stenosis to recur within 1 year. Consequently, the procedure is now performed only under unusual circumstances. It is occasionally used as a bridge to later surgery for patients who are initially too ill to safely undergo valve replacement. It is also sometimes performed for the temporary palliation of patients who are not candidates for valve replacement. In addition, it has a role for children with congenital aortic stenosis, where temporary treatment by valvuloplasty may allow the child to complete growth before requiring surgical valve replacement.

Pulmonic stenosis

Balloon valvuloplasty is the treatment of choice for patients with pulmonary stenosis. The majority are children whose valves respond well to this treatment. The advantage of avoiding surgery outweighs the moderate tendency for restenosis of these valves.

Percutaneous closure of cardiac defects

Atrial septal defects and patent ductus arteriosus can be closed percutaneously with catheter-delivered devices. One such device called a clamshell has been used for this purpose for a number of years, but has not yet gained wide acceptance.

Further reading

General reading

Topol EJ, ed (1999). *Textbook of interventional cardiology*, 3rd edn. WB Saunders, Philadelphia. [This is a standard textbook of interventional cardiology. It covers the topics discussed in this chapter in greater detail and gives complete references.]

Interventional versus medical therapy

Hartigan P, *et al*. (1998). Two to three year follow-up of patients with single vessel coronary artery disease randomized to percutaneous transluminal angioplasty or medical therapy (results of a VA Cooperative Study). *American Journal of Cardiology* **82**, 1445–50.

RITA-2 Investigators (1997). Coronary angioplasty versus medical therapy for angina: the second Randomised Intervention Treatment of Angina (RITA-2) trial. *Lancet* **350**, 461–8.

Interventional versus surgical therapy

Henderson R, *et al*. (1998). Long term results of RITA-1 trial: clinical and cost comparisons of coronary angioplasty and coronary-artery bypass grafting. *Lancet* **352**, 1419–5.

The BARI Investigators (2000). Seven-year outcome in the Bypass Angioplasty Revascularization Investigation (BARI) by treatment and diabetic status. *Journal of the Anmerican College of Cardiology* **35**, 1122–9. [These are two of several randomized trials yielding similar conclusions regarding the relative benefits of interventional versus surgical therapy.]

Unstable angina

FRISC II Investigators (1999). Invasive compared with non-invasive treatment in unstable coronary-artery disease: FRISC II prospective randomised multicentre study. *Lancet* **354**, 708–15. [The latest of several studies on this subject, and the first to show a clear advantage for invasive treatment.]

Acute myocardial infarction

Grines CL, *et al*. (1993). A comparison of primary angioplasty with thrombolytic therapy for acute myocardial infarction. *New England Journal of Medicine* **328**, 673–9.

Gibbons RJ, *et al*. (1993). Immediate angioplasty compared with the administration of a thrombolytic agent followed by conservative treatment for myocardial infarction. *New England Journal of Medicine* **328**, 685–91.

Zijlstra F, *et al*. (1993). A comparison of immediate coronary angioplasty with intravenous streptokinase in acute myocardial infarction. *New England Journal of Medicine* **328**, 680–4.

[These three studies all appeared in the same journal issue and presented similar data supporting immediate intervention over thrombolytic therapy. Such an advantage is yet to be proven in the community setting.]

15.4.2.5 Coronary artery bypass grafting

A. J. Ritchie and L. M. Shapiro

Introduction

Coronary artery bypass grafting (CABG) is one of the most effective therapeutic surgical interventions that can be undertaken. It can reliably be performed with a mortality of 1 to 3 per cent in complex groups of high risk patients, providing sustained and proven relief from the symptoms of angina, and high quality, event-free survival in up to 80 per cent of patients 10 years later. It is a unique procedure that has to be immediately effective if the patient is to survive: there is no room for recovery and rest of the organ as in almost all other forms of surgery. The role of CABG in the treatment if ischaemic heart disease is constantly evolving. Over the past decade, advances in all fields of cardiovascular medicine and surgery have taken place at an astonishing pace. This continues, such that discerning the role of the changing therapeutic options requires continuous reassessment. The purpose of this chapter is to define the role of CABG as it currently stands, discuss issues of particular relevance to physicians, to outline work currently under study, and to delineate likely future developments.

Historical perspectives of CABG

Cardiac surgery is the youngest surgical specialty. The earliest attempts by surgeons to treat ischaemic heart disease were aimed at avoiding operating on the heart altogether, for example endeavours to control the symptoms of angina by denervation with alcohol and phenol injections. Subsequently, it was claimed that 80 per cent of patients could have their angina abolished if the thyroid gland was removed, but the price was obvious—profound hypothyroidism.

In the early part of the twentieth century efforts were made to address the root cause, occlusive coronary artery disease, by indirect attempts at revascularization. These took the forms of attaching omentum to the heart, or inducing a pericarditis and inflammatory response. More directly, arterialization of the coronary sinus was achieved by connecting a vein graft from the aorta. In 1946, Vineberg implanted the bleeding end of the left internal mammary artery into the left ventricular wall.

Seminal to the development of modern cardiac surgery were technological advances that gave surgeons the time and opportunity to construct direct aortocoronary artery bypass grafts with the expectation of a successful outcome. These came in the form of cardiopulmonary bypass circuits and angiography, which allowed the identification and targeting of the epicardial arteries. Saphenous vein bypass conduits attached directly to coronary arteries by Johnson and Favaloro made direct coronary revascularization possible and this became achievable in routine operations, associated with low mortality and morbidity, and which were applicable to the vast majority of patients. They offered substantial benefits over previous therapeutic modalities for intractable angina pectoris. CABG passed from being possible to achievable.

Questions then arose about the applicability and affordability of CABG when compared to medical treatment, which was also evolving at the same time. The most important clinical trials that had a bearing on this were conducted in the 1970s and 1980s. The Veterans Administration Coronary Artery Surgery Study and the European Coronary Surgery Study were multicentre studies examining the efficacy of CABG versus medical management in different clinical situations. These landmark studies each had problems controlling variables and could not compensate for important improvements in treatment within each group and cross over from medical to surgical groups. In addition, different types of patient were enrolled in the various studies, making comparison between them difficult. However, in essence, they showed that surgery provided better relief of symptoms, improved functional capacity, and reduced the incidence of fatal myocardial infarction. They demonstrated that patients with left main stem and/or significant flow obstructing lesions in all three major coronary arteries (triple vessel disease) enjoyed a significant prognostic benefit compared with medical treatment. Those with impaired ventricular function benefited likewise from surgery. Hence, these flawed but crucial studies led to the development of broad clinical indications for CABG. These indications, together with others that are widely accepted in clinical practice, are summarized in Table 1.

The objectives of CABG

Rapid revascularization of ischaemic myocardium, no matter how achieved, dramatically improves outcome and survival. Complete reversibility of the changes that occur due to ischaemia can be achieved when perfusion is restored, providing the rational basis for thrombolytic therapy and mechanical revascularization, whether achieved by balloon angioplasty or CABG. The simplistic aim is to rematch oxygen supply with demand. However, recent studies indicate that where there is disease at the microvascular level, then epicardial revascularization alone is likely to be only of partial benefit. Advances in medical therapy that reduce mortality in ischaemic heart disease, such as aspirin, b blockade, aggressive treatment of hyperlipidaemia by HMG CoA reductase inhibition, and angiotensin-converting enzyme (ACE) inhibition were not available at the time of the early CABG versus medical trials outlined previously. These are likely to affect the microvasculature in addition to effects on disease of the main epicardial (coronary) arteries. We are therefore moving into an era where the control of risk factors may lead not only to a slowing of the progression of coronary disease, but even to its regression. This has major implications for defining the modern role of CABG, but the core objectives of mechanical revascularization remain the same:

1. revascularization of ischaemic areas of myocardium;
2. relief of anginal symptoms;
3. prolongation of patient survival;
4. prevention of myocardial infarction;
5. preservation of cardiac function;
6. improvement of quality of life, for example exercise tolerance.

The broad indications for CABG are to control symptoms where medication has failed and the patient remains unacceptably incapacitated by pain, and to improve prognosis when this is possible. Increased survival after coronary artery bypass surgery is seen in the following groups: left main stem stenosis, triple vessel disease, double vessel disease with left ventricular impairment, and left ventricular aneurysm.

The standard CABG operation

This involves the use of the left internal mammary artery and saphenous vein obtained from the leg as bypass conduits via a median sternotomy incision. The patient is then fully heparinized and cardiopulmonary bypass instituted, usually by cannulation of the ascending aorta and right atrium. When it is difficult to access these vessels safely, then femoral artery and vein cannulation provide a satisfactory alternative. The cardiopulmonary bypass machine provides continuous or pulsatile flow, oxygenates and removes carbon dioxide, and regulates the temperature of the blood. This leaves the heart isolated, allowing the operation to proceed. A range of techniques is then employed to reduce or stop heart movement and allow the construction of anastomoses. Inevitably these compromise the circulation to the heart and create myocardial ischaemia, creating a time limit within which the surgery must be completed. As directed by the angiogram, and in the knowledge that incomplete revascularization is a risk factor for premature death or recurrent angina, all technically suitable coronary arteries are bypassed distal to occlusive lesions. Where diffuse disease exists, or there is no lumen, an endarterectomy can be performed. As many bypass grafts as necessary can be constructed from available conduits using 'jump' grafts, but the usual number is three or four in patients with triple vessel disease. At the end of the procedure the patient is weaned from

cardiopulmonary bypass and returned to the intensive care unit.

Perioperative morbidity

Following a standard CABG procedure the patient must recover from the effects of cardioplegia and cardiopulmonary bypass as much as from the operation per se. The use of cardioplegia rarely causes perioperative myocardial infarction or more diffuse global myocardial ischaemia, thought to be due to reperfusion injury and resulting in myocardial stunning. This can result in a low cardiac output state, seen particularly in patients with impaired ventricular function or evolving infarction. Although this occurs in less than 2 per cent of cases, the changing case mix of the population undergoing cardiac surgery is resulting in operations on older, sicker, and higher risk patients (Table 2), such that these problems are likely to become more common. Several strategies are under investigation to improve cardioprotection, but in general the risk of complications rises with prolonged ischaemia and cardiopulmonary bypass times and in older patients; they are rare in routine elective cases.

The use of cardiopulmonary bypass results in local and systemic complications as outlined in Table 3. The most common form of arrhthymia after CABG is atrial fibrillation, occurring in up to 30 per cent of cases. This often delays patient discharge and carries extra risks for the patient, despite its usually transient nature. Ventricular arrhymias are much less frequently seen, but potentially much more serious, and usually indicate ongoing ischaemia.

Transient but diffuse cerebral injury resulting in short term memory and concentration loss are detectable by neurobehavioural comparison pre- and postsurgery in most patients undergoing CABG. By contrast, stroke is very rare, affecting less than 1 per cent of cases. Damage to other organs, in particular the development of acute renal failure, can follow cardiopulmonary bypass and the occurrence of a low output state.

Medical treatment and longer-term complications of patients who have undergone CABG

Routine medication in the immediate and long term almost always includes lifelong aspirin or other antiplatelet agent. There is little place for the use of warfarin, except in those patients who have undergone endarterectomy. Medications to control hypercholesterolaemia (statins) and hypertension (b blockers and ACE inhibitors) are increasingly being associated with improved survival and are usually lifelong treatments.

The patient has to recover from the operation before CABG can achieve its long-term goals. In the vast majority of cases there are no complications, when ambulation and return to exercise occurs within days. Initial stiffness and muscular aches can be expected, but there is no reason to advise against exercise or resumption of sexual activities for any specified period of time. By far the most frequent complaint that patients make is in relation to the saphenous vein harvest site. The ankle has a tendency to swell and the wound may be inflamed. This usually subsides quickly with rest and elevation. The increasing use of radial artery for conduits is associated with quickly healing, trouble-free wounds. A dull ache or burning pain in the chest can be associated with mammary artery harvest but usually recedes after 2 to 3 months.

Driving, riding bicycles, or other activities that put similar tensions on the chest should not be undertaken until the sternal base has a solid union, usually 6 to 8 weeks postoperatively. It is normally expected that patients who are working up until the time of their operation will be fit to resume work at 2 months postoperatively, although those employed in hard physical labour should be advised to recommence this carefully.

Other complications that can develop in the medium and long term are very rare. Postcardiotomy syndrome (Dressler's syndrome) presents as pain in the chest, which may be similar to angina but is associated with a pericardial friction rub and relieved by non-steroidal anti-inflammatory agents. Return of angina is the most sinister symptom, requiring further investigation and usually due to incomplete revascularization or occlusion of a bypass conduit. In diabetics this may be due to microvascular disease, the operation being conducted to confer prognostic benefit rather than relief of angina as its main aim.

Mortality

CABG is currently performed in over 1000 per million population in the United States and 300 to 700 per million population in Europe and Australia. It is the most documented and assessed operation ever performed. In the United Kingdom, mortality for first time CABG that includes elective and emergency cases has remained constant at 2 to 3 per cent over the last decade, with comparable figures in the United States. These data do not yet take account of risk stratification, the different nature, increased complexity, and age of patients currently referred, but there is clearly a range of outcomes, from less than 1 per cent mortality for a routine elective case to up to 10 per cent for rescue from acute percutaneous transluminal coronary angioplasty (PTCA) dissection.

Non-standard CABG procedures

Prior to the introduction of percutaneous transluminal coronary angioplasty (PTCA), CABG resulted in longer survival and better quality of life for patients with multivessel disease compared to medical treatment alone. Technological advances in PTCA and stenting now offer strategies in this group. The Bypass and Angioplasty Investigation (BARI) trial, a 5-year prospective comparison of CABG and PTCA in patients with multivessel disease, found no statistically significant difference in survival between the two groups (except for diabetic patients). However, the rate of reintervention or revascularization was 42 per cent in the PTCA group, compared with only 3 per cent in the CABG group, with 31 per cent of patients initially undergoing PTCA ultimately receiving CABG.

Standard CABG still has a major role to play in patients with multivessel disease, yet the invasive, expensive, and time-consuming nature of the operation are unattractive to physician and patient alike. While much has been done to limit the deleterious effects of cardiopulmonary bypass, reports in the last decade have continued to highlight the morbidity and mortality associated with its use. The introduction of PTCA has been seen as a way of avoiding these problems altogether and resulted in a dramatic shift of treatment paradigms, as well as driving new technical developments in surgery.

Total arterial revascularization

Cardiac surgeons no longer dispute the use of the left internal mammary artery as the choice for revascularizing the left anterior descending artery, pioneered in 1968. Critics doubted its ability to deliver enough flow and worried about morbidity associated with its harvest, particularly in diabetic patients. However, the use of single as well as double internal mammary artery grafts instead of saphenous vein has been convincingly demonstrated to improve survival and reduce recurrent angina and infarction without adding to morbidity or mortality. The outcome for the patient is longer survival and higher quality of life where revascularization is maintained by continued patency.

In the past, the issue of conduit patency was not accorded primary importance because progression of disease distal to the original site of anastamosis was thought to be the major determinant of outcome. However, this view has changed with appreciation of the time-related failure of saphenous veins due to accelerated atherosclerosis and the advent of a new era of medical interventions (aggressive lipid lowering strategies etc.). These may prevent progression of disease and stabilize acute plaque fissure, or even lead to disease regression, meaning that the long-term patency of the conduit becomes the major determinant of success or failure of CABG. The current trend in CABG is therefore to utilize a range of arterial conduits with similar biological properties, for example radial artery, gastroepiploic artery. Increased use of total arterial revascularization in combination with control of risk factors is likely to reduce significantly the requirement for redo CABG, with its attendant risk for the patient and cost to society.

Off pump coronary artery bypass

The initial work in coronary artery bypass grafting was done in an era before cardiopulmonary bypass was established. In 1967, Kolessov performed grafts to left anterior descending and circumflex through a left thoracotomy without bypass, and various reports over the following 7 years demonstrated the safety of off pump CABG in large, single centre cohorts of patients. However, the introduction and refinement of cardiopulmonary bypass made an extraordinary difference to our ability to provide definitive surgical management. Additionally, dramatic improvements in survival at and beyond surgery came with advances in cardioplegia, allowing a dry operative field and a protected myocardium. Such was the rapid advance in these technologies that the original method of CABG was abandoned, but off pump techniques are now being used again.

Minimal access CABG procedures, conducted through minimally invasive incisions, are capable of producing effective and long lasting anastomosis off pump in single vessel disease. The main shortcomings of this approach are the limited number of vessels that can be bypassed, that is primarily the left internal mammary to left

anterior descending arteries, and morbidity from minithoracotomy wounds. It is likely to be applicable to less than 10 per cent of the overall patient population, but because of the significant reintervention rate following PTCA and the associated problem of in-stent stenosis, the off bypass and minimal access CABG procedure has immense potential and may revolutionize coronary revascularization in a large group of patients.

Minimally invasive procedures in cardiology and surgery can have significant advantages for both patients and institutions: reduced recovery time, reduced requirement for intensive care, and shorter hospital stay. Costs can potentially be reduced in the short and long term, with quicker return to normal life. These procedures are patient and industry driven: audit data is disseminated on the internet, and many patients in the United States and Europe now actively seek off pump CABG procedures. Well-designed, randomized, prospective, controlled trials are needed in this area.

The future for CABG

The last decade has seen marked and rapid advances across all aspects of medical and surgical management of ischaemic heart disease, making it difficult to discern the optimal treatment strategy. Innovation in CABG surgery is likely to utilize total arterial conduits in short operations, possibly done without the use of cardiopulmonary bypass. It remains to be seen in whom these advances are best applicable, and how affordable they are in competition with percutaneous interventional techniques. However, single episode, short duration operations with unrivalled long-term patency provided by arterial conduits and backed up by preventative cariological medications are likely to be cheaper and more efficacious than currently available interventional alternatives.

Further reading

Bergsma TM *et al*. (1998). Low recurrence of angina pectoris after coronary artery bypass graft surgery with bilateral internal thoracic and right gastroepiploic arteries. *Circulation* **97**, 2402-5.

Buffalo E *et al*. (1996). Coronary artery bypass grafting without cardiopulmonary bypass. *Annals of Thoracic Surgery* **61**, 63–6.

Cooley DA (1998). Coronary bypass grafting with bilateral internal thoracic arteries and right gastroepiploic artery. *Circulation* **97**, 2384–5.

Loop FD *et al*. (1986). Influence of internal mammary artery graft on 10 year survival and other cardiac events. *New England Journal of Medicine* **314**, 1–6.

Pepine CJ, Deedwania PC (1998). How do we best treat patients with ischaemic heart disease? *Circulation* **98**, 1985–6.

Society of Cardiothoracic Surgeons of Great Britain and Ireland (1998). National adult cardiac surgical database report.

15.4.2.6 The impact of coronary heart disease on life and work

M. C. Petch

Introduction

Coronary heart disease is common and lethal (Table 1 and Table 2). In developed countries, heart attacks account for about a quarter of all deaths. Death is usually sudden. These facts are well known and have a profound influence on attitudes towards the victims of heart disease. Employers are reluctant to take back people who have lost time off work as a result of a heart attack. Spouses become overprotective. The survivors are acutely aware that they have received an intimation of their mortality; some fail to cope. The first manifestation of coronary heart disease, which is usually chest pain, prompts re-evaluation of the remainder of life and work. The spectre of cardiac pain and death hangs over many a middle-aged man, including employers, politicians, public health physicians, journalists, and others in positions of influence. In most developed countries there is therefore public pressure to prevent the development of coronary disease (primary prevention), to prevent a recurrence (secondary prevention), and to put in place measures which will reduce the risk of harm to the individual and others in the event of sudden incapacity/death of a worker in a 'safety-critical' job.

Women are not of course immune, but coronary heart disease does tend to strike later, often after usual retirement age. Nevertheless the impact of coronary heart disease can be as devastating: older women are often the most important carers in a family. Whilst there are minor differences between the sexes in the presentation and management of coronary heart disease, the comments in this chapter should be taken to apply to both sexes.

Life before coronary heart disease

Most people do not think about their health until it goes wrong. With advancing years people become aware that their contemporaries are suffering from mortal diseases. They then belatedly begin to look at their own lifestyle. Many believe the results of the latest research quoted in the press and attempt to adapt their habits by increasing their intake of vitamin E, or fish oil, or red wine, or by reducing the amount of coffee and animal fat that they consume, or by undergoing stress counselling, or by purchasing an exercise machine which they never use. Then along comes a new report which sets another fashion.

There are a few public health issues on which the medical profession can speak with authority. Cigarette smoking is the prime example. Doctors, nurses, and other health-care professionals have a duty to discourage this habit by example and by persuasion. No other habit enjoys such powerful evidence that mandates a lifestyle change. Regular exercise is to be commended. A prudent diet is capable of different interpretations, but the old adage 'a little of what you fancy does you good' dates back many generations to a time when coronary heart disease was much less common. Food can be one of life's great pleasures. The current political ambition to change national lifestyles is not heeded by those most at risk and has never been clearly shown to have lasting benefit.

A sensible compromise for most societies is to prevent smoking, encourage exercise, promote the sale of fruit, vegetables, and so on, and to reduce the availability of 'junk' foods in shops and workplace canteens, but not to go to such lengths that people feel guilt when faced with a delicious steak. The fact that this Epicurean attitude is shared by most doctors makes it all the more persuasive. The use of drugs such as aspirin and statins to reduce the risk of a coronary event can likewise only be justified in those individuals whose risk is especially high, as judged by their family history and other risk factors.

Health screening is another controversial topic. In (over)developed societies screening services have become very popular and assessment of cardiovascular risk in businessmen is a useful source of income for some clinics. Certainly the measurement of blood pressure can be supported and, in some circumstances, the estimation of serum lipids. Beyond that the advice that may be offered boils down to common sense—don't smoke, take more exercise, eat less.

Occasionally health screening can create extreme anxiety, for example when an electrocardiogram or exercise test suggests silent coronary disease. These investigations can only be justified when the individual is aware of the possible outcomes of screening and/or is in a safety-critical job.

Life and work with coronary heart disease

The risk of sudden disability and death through ventricular fibrillation is the major factor affecting work capacity amongst victims of coronary heart disease. The risk is greatest in the early days following the development of symptoms and in those with most myocardial damage.

Common sense and experience (i.e. clinical judgement) remain the best tools for assessing an individual's fitness to resume his life and work following the development of coronary heart disease. The onset of cardiac pain, or change in the nature of pain in someone with known ischaemic heart disease, should prompt rapid evaluation. Stable angina pectoris, preferably confirmed by exercise testing, usually requires no change in lifestyle: modern drug therapy is very effective and often comprises just aspirin, glyceryl trinitrate, and a statin. Unstable angina or myocardial infarction is a different matter and necessitates hospital admission, with further investigation. Even then clinical judgement remains the basis for advice about lifestyle changes, supplemented by 'non-invasive' tests.

The presence of myocardial failure and/or significant areas of ischaemia are the principal determinants of prognosis. The former may be identified by history, clinical examination, chest radiography, and echocardiography; the latter by the development of angina and electrocardiographic ST-segment shift on exercise testing. An exercise test may also reveal cardiovascular incapacity in other ways, for instance exhaustion, inappropriate heart rate and blood pressure responses, and arrhythmia.

Following myocardial infarction or unstable angina, assessment of prognosis along the lines outlined above is recommended: those with no complications and good exercise tolerance may return to work in about 4 weeks. This applies particularly to younger individuals whose employers need have little hesitation in taking them back to their former job, perhaps part-time initially. A few will take longer to recover, and some will need a change of job.

Limitation of working capacity and the risk of sudden incapacity can both be well judged in populations by specialist opinion, aided by the results of 'non-invasive' tests. However, the progression of coronary disease can be unpredictable, and individuals judged to be at low risk from further cardiovascular events can suffer recurrences. This difference between the individual and the population is not well understood by employers and employees and can be a source of misunderstanding and confusion. Nevertheless, individual exceptions do not invalidate the principles on which recommendations for individuals are made.

Coronary angioplasty/stenting

Patients with persistent angina, or those with a very abnormal exercise response, should undergo coronary arteriography with a view to myocardial revascularization. Coronary angioplasty is nowadays straightforward, safe, and effective in relieving angina. Resumption of normal activities, including work, is normally possible a few days afterwards. Recurrent angina is much less common with the more widespread use of stents, but it remains a problem in 10 to 20 per cent of patients and hence

regulatory authorities remain cautious about those individuals whose performance might be compromised by a return of cardiac pain.

Coronary artery bypass grafting

Coronary artery bypass grafting is also remarkably safe, with most centres reporting mortality rates of less than 1 per cent for elective operations. Recovery is rapid and most patients resume work within 2 to 3 months of surgery. Most are relieved of their angina. Patients who were able to work before surgery should generally be able to do so afterwards, and restrictions that may have been appropriate previously should no longer be relevant. However, since surgery is a dramatic event, it may prompt overprotective attitudes amongst family members, friends, employers, or even medical advisers. Many individuals who could and should return to work fail to do so for this reason, rather than because of continuing incapacity. No special restrictions are usually necessary after return to work. Coronary graft stenosis and occlusion leads to recurrence of angina at a rate of about 4 per cent per annum. This is generally less severe than previously but will affect long-term occupational planning. Unfortunately, waiting times for coronary arteriography and bypass surgery in some countries (such as the United Kingdom) are very long, so that many patients do not return to work.

Rehabilitation programmes

Rehabilitation programmes are now well established in many hospitals and communities. These enable patients to make a full physical and psychological recovery following a cardiac event such as myocardial infarction or coronary artery bypass grafting. An acceptable exercise response is a prerequisite for enrolment into a rehabilitation programme. The participants are thus the fittest survivors, selected for physical retraining on the strength of their satisfactory performance on the treadmill. Definite measures of benefit, such as reduction in recurrent myocardial infarction or death, are lacking.

However, the fashion for cardiac rehabilitation is undeniable and seems to owe much to the enthusiasm of the participants—patients and staff alike. This may be a comment on modern cardiology with its mechanistic approach, haste, and failure to recognize the psychological effects of heart attacks, with concomitant need for lifestyle advice. Sex, for example, is rarely discussed except in rehabilitation classes, yet for many patients it is a burning issue. The mechanistic view is that the physical effort required is equivalent to two flights of stairs or stage 3 of the Bruce protocol. The psychological aspects are probably better dealt with in a rehabilitation class (or perhaps the bar afterwards). With health-care budgets always under pressure, it is reasonable to suggest that the demand for rehabilitation might perhaps best be satisfied on a voluntary basis, and that scarce resources may be better directed elsewhere, for example towards the drug budget, where outcome benefits are measurable.

Risk evaluation: the 1 per cent rule

Workers whose sudden incapacity would place themselves and others at risk are described as being in 'safety-critical' jobs. The traditional approach to this dilemma was to exclude anyone with heart disease from working in such an environment. This may still be appropriate in certain occupations where any increased risk of incapacity is unacceptable: drivers of mainline passenger trains and captains of ocean-going vessels are two current examples. This blanket exclusion is patently unfair to some, and may waste the skills and experience of a valued employee. Also, no individual is totally free of risk of an incapacitating event, so a few accidents as a result of sudden illness are inevitable in apparently normal people. A better approach is therefore to define what level of risk is acceptable, and then decide whether the medical condition places that individual within the predetermined limits of acceptability. This has the great merit of objectivity and is a well-tried engineering practice.

The Civil Aviation Authority was the first to adopt this approach with what is now known as the '1 per cent rule'. Aircraft engineers have always recognized that a disaster may occur as a result of component failure, and have recommended design and safety features so that the risk of failure is 'extremely improbable' ($1/10^{-9}$ flying hours). This approximates to a risk of an incapacitating event in a pilot of 1 per cent per annum if a number of assumptions are made. A pilot with a medical condition may therefore be regarded as a component of aircraft safety and hold a licence if his risk of a cardiovascular event is comparable, that is, his risk is no greater than his peers or other parts of the aircraft.

There are a number of difficulties in applying the approach described above in other situations and in other industries. First, who decides an acceptable level? Second, the epidemiological data in cardiovascular medicine generally describe events such as death or heart attack, which may not be the relevant parameter. Heart attacks are a rare cause of road traffic accidents; more commonly the driver is found in his vehicle on the verge, 'slumped over the wheel with the engine still running'. Death may have been sudden in epidemiological terms, but it was not instantaneous; the victim had sufficient warning to pull over to the side of the road. Third, some incapacitating events, neurocardiogenic syncope for example, are clearly relevant to many safety-critical jobs, and yet there are scant data on which to base an objective decision. Fourth, cardiovascular event rates have fallen since the 1 per cent rule was formulated.

Risk evaluation: exercise testing

The data on exercise testing in coronary heart disease are the best established for evaluating the risk of incapacity in employees in safety-critical jobs, for example vocational driving. The guidelines relating to vocational drivers were developed in the United Kingdom and adapted by a Task Force of the European Society of Cardiology. They are now being applied more widely to other groups of workers whose occupation may involve an element of risk to themselves or others should that individual suffer cardiovascular collapse.

The protocol for which most information is available is that described by Bruce. He and Fisher examined strategies for risk evaluation of sudden cardiac incapacitation in men in occupations affecting public safety: 2373 men with clinically manifest coronary artery disease who had undergone exercise evaluation were followed up for a mean of 61 months; 300 sudden cardiac incapacitations (cardiac arrest or sudden cardiac death) occurred. Exercise testing in all age groups defined low- and high-risk populations with annual incapacitation rates of 1 and 3 per cent, respectively. The former were those who could reach stage 3 of the Bruce protocol with no chest pain, attain 85 per cent of age-predicted maximal heart rate, and manifest less than 1 mm of ischaemic ST-segment depression. A similar message came from the study of 4083 medically treated patients in the Coronary Artery Surgical Study registry. The 32 per cent of patients who could exercise into stage 3 of the Bruce protocol with less than 1 mm ST-segment depression on ECG (10 METS) had an annual mortality of 1 per cent or less. By contrast, the annual mortality rate of the 730 patients with 2 mm or greater ST depression was 3.6 per cent, ranging from 5.6 per cent for those patients achieving stage 1 or less of exercise to 2.0 per cent for those patients achieving stage 3. The study also confirmed the overriding prognostic importance of left ventricular function and the poor survival of patients with heart failure. An ability to exceed stage 3 of the Bruce protocol with less than 2 mm of ST-segment depression is the best criterion for identifying a population with an annual risk of death of less than 2 per cent.

Driving

Since decisions concerning fitness to drive should be objective and evidence-based whenever possible, a similar approach to that described for pilots is being adopted. An attempt is also being made to be consistent, so that all forms of illness that might cause sudden incapacity should be considered in comparable manner. One condition for which good data are available is epilepsy. Currently the agreed, annual, acceptable levels of risk in the United Kingdom are 2 per cent for vocational drivers and 20 per cent for ordinary drivers: a driver's licence entitlement can be determined by reference to well-validated tables of risk, for example following a head injury. Risks for drivers with cardiovascular disorders are less easy to quantify because of the poor relationship between the presence of the disease process and the risk of incapacity. However, some drivers with heart disease can be identified as being at an increased risk of an incapacitating event, and attempts are being made in the transport industry and elsewhere to provide objective criteria, which will be applicable across a range of disease processes.

The '2 per cent rule' may prove to be the correct criterion for vocational drivers and other workers in similar occupations who suffer from cardiovascular disorders. Society already accepts drivers with vocational licences up to the age of 80 years, by which time their annual risk of a cardiovascular event is 4 per cent. If the assumption is made that half of the events are incapacitating then the acceptable risk accords with the epilepsy criteria, so those drivers whose annual risk of a cardiovascular event is 4 per cent (or death 2 per cent) or greater should not be entitled to hold a vocational licence.

For ordinary drivers a 20 per cent annual risk also seems reasonable for cardiovascular disorders; such level of risk is in accord with existing guidelines, for instance shortly after a heart attack. Ordinary driving may be resumed 1 month after a cardiac event provided that the driver does not suffer from angina which may be provoked at the wheel. Vocational driving may be permitted at 6 weeks, subject to a satisfactory outcome from non-invasive testing. In the United Kingdom, ordinary driving licence holders do not need to notify the Driver and Vehicle Licensing Agency (**DVLA**), Swansea if they have made a good recovery and have no continuing disability, but vocational drivers must notify the DVLA. Insurance companies vary in their requirements, but most policies are temporarily invalidated by illness.

Special circumstances

Toxic substances

Work involving exposure to certain hazardous substances may aggravate pre-existing coronary heart disease and careful consideration should be given to patients who are returning to jobs involving exposure to chemical vapours and fumes. Methylene chloride, a main ingredient of many commonly used paint removers, is rapidly metabolized to carbon monoxide in the body and, in poorly ventilated work areas, blood levels of carboxyhaemoglobin can become elevated enough to precipitate angina or even myocardial infarction. A blood carboxyhaemoglobin level of 2 to 4 per cent has been shown to be associated with impairment of cardiovascular function in patients with angina pectoris. The World Health Organization recommends a maximum carboxyhaemoglobin level of 5 per cent for healthy industrial workers and a maximum of 2.5 per cent for susceptible persons in the general population exposed to ambient air pollution: this level may also be applied to workers whose jobs entail specific exposure to carbon monoxide, such as car park attendants and furnace workers. To ensure that the 2.5 per cent carboxyhaemoglobin level is not exceeded, the ambient carbon monoxide concentration should not be higher than 10 ppm over an 8-h working day: equivalent to exposure to the current occupational exposure standard (50 ppm) for no more than 30 min. Occupational exposure to carbon disulphide in the viscose rayon manufacturing industry is a recognized causal factor of coronary heart disease but the mechanism remains unclear.

Reports of sudden death from angina are well recognized in dynamite workers, particularly after a period of 36 to 72 h away from work and following re-exposure, an effect almost certainly related to direct action of nitroglycerine on the blood vessels of the heart or peripheral circulation. Persons with clinical evidence of coronary heart disease should avoid occupational exposure to these substances.

Solvents, such as trichloroethylene or 1,1,1-trichloroethane, may cause sudden death in workers receiving heavy exposure in poorly ventilated workplaces. The chlorofluorocarbon CFC-113 has been implicated in sudden cardiac deaths and CFC-22 has been reported to cause arrhythmias. Some industrial workers will need proper assessment of their workplace by an occupational physician and occupational hygienist so that they can be advised on their suitability for work handling chlorinated hydrocarbon solvents or involving exposure to gases.

There are no formal medical requirements for workers who have to enter confined spaces where there may be hazards of oxygen deficiency or a build up of toxic gases. Those with heart disease or severe hypertension may need to be excluded. Certain occupations may require the use of special breathing apparatus either routinely (e.g. asbestos-removal workers), or in emergencies (e.g. water workers handling chlorine cylinders). The additional cardiorespiratory effort required whilst wearing a respirator, combined with the general physical exertion that may be required, usually means that people with a previous history of coronary heart disease are excluded from such work.

Hot conditions

Working in hot conditions may prove difficult for some patients with heart disease. High ambient temperatures or significant heat radiation from hot surfaces or liquid metal, added to the physical strain of heavy work, will produce quite profound vasodilatation of muscle and skin vessels. Compensatory vascular and cardiac reactions to maintain central blood pressure may be inadequate and lead to reduced cerebral or coronary artery blood flow. The resulting weakness or giddiness could prove dangerous. Since many cardioactive drugs have vasodilating and negative inotropic actions, some reduction in dosage may be necessary.

Cold conditions

Cold is a notorious trigger of myocardial ischaemia and caution must therefore be exercised for individuals who suffer from coronary heart disease. Impaired circulation to the limbs will result in an increased risk of claudication, risk of damage to skin (frostbite), and poor recovery from accidental injury to skin and deeper structures.

Stress

The idea that psychological stress has a role in the aetiology of coronary heart disease is a persistent one, owing much to the work of Friedman and his colleagues who suggested that hectic work patterns marked by long hours, competitiveness, time urgency, and aggression (so-called type A behaviour) may predispose to the development of coronary heart disease. The results of other epidemiological and clinical studies have been conflicting, much of the difficulty arising from the criteria for identifying a type A personality. The idea that a stressful incident may trigger a heart attack is also well embedded in Western culture. Some studies have shown a relationship, for example after earthquakes. But generally psychological stress is not regarded as sufficient provocation to form the basis of a legal settlement. Such stress may, however, on occasion provoke angina in susceptible individuals. Patients and their relatives can almost invariably point to a stressful incident prior to a heart attack and may fail to appreciate that life is a series of stressful incidents, heart attacks are extremely common, and coincidences are inevitable. The only trigger of a heart attack that has withstood critical scrutiny (and legal cross-examination) so far is sudden unaccustomed vigorous effort within 2 h of the attack.

Travel

Following a cardiac event such as myocardial infarction, individuals should convalesce at home and not travel far for 4 to 6 weeks. Those with no evidence of continuing myocardial ischaemia or heart failure can then travel freely within their own country for pleasure, for example a holiday. Business and overseas travel is more problematical because the physical and psychological demands are greater. Additional difficulties for the overseas traveller include the uncertain provision of coronary care facilities in some countries and the justifiable reluctance of insurance companies to provide health cover. Such travel is best deferred until 3 months have elapsed and any necessary further investigations and treatment have been carried out to ensure cardiovascular fitness.

Overseas travel for those with continuing cardiovascular unfitness need not be ruled out. Utilizing the airport services for disabled travellers can ease a passenger through customs, passport control, and so on, at major airports. Modern aircraft can be very comfortable. The cabins are kept at a pressure equivalent to 2000 m so that those with angina are not likely to experience an attack. Businessmen with continuing cardiac disorders may therefore fly to Europe, North America, and other countries with good coronary care services with very little risk. But flights in unpressurized aircraft, work in undeveloped countries or in remote areas of the world, and work in a hostile environment (both climatic and political) are best avoided.

Aircrew are subject to guidelines drawn up by the Joint Aviation Authorities. In the United Kingdom the regulatory agency is the Civil Aviation Authority, whose advice should always be sought.

Cardiac deaths are uncommon in trekkers or workers at high altitude (2440 to 4570 m.). The increase in cardiac output at altitude will exacerbate symptoms in those who already experience symptoms at sea level, but asymptomatic individuals with coronary heart disease are unlikely to be at special risk.

Implanted devices

Cardiac pacemakers are generally implanted into older patients who have idiopathic degeneration of their conduction system. However, both heart block and sinoatrial disorder are well-recognized complications of coronary heart disease. Single- and dual-chamber (VVI and DDD in most) pacemakers are rarely subject to electromagnetic interference and no modification of lifestyle is necessary, with the exception that the device can trigger alarms at airports and elsewhere.

The implantable cardioverter defibrillator (**ICD**) has the capacity to detect and treat ventricular tachycardia and fibrillation, either by antitachycardia pacing or by a shock, in patients in whom a further cardiac arrest is anticipated. Both shock and arrhythmia are potentially incapacitating. In North America and Europe, patients with ICDs have restrictions placed upon them, for example driving. They commonly have severe underlying heart disease and may well not be able to work, but if they can do so, then this should be in an environment that is free from electromagnetic interference. There has been one report of ICD malfunction in the vicinity of an electronic antitheft surveillance system.

There has been considerable interest in the possibility that mobile telephones might interfere with pacemakers and ICDs. Studies have shown that this is a theoretical possibility and that reprogramming of a pacemaker can be achieved under exceptional circumstances if the telephone is held close (less than 20 cm) to the pacemaker. In practice no clinically significant interference has yet been reported, but individuals are advised to use the hand and ear furthest from the pacemaker

and not to 'dial' with the telephone near to the pacemaker.

Seafarers

The Merchant Shipping (Medical Examination) Regulations in the United Kingdom currently state that any manifestation of ischaemic heart disease renders the individual permanently unfit to return to sea. This regulation has been in force since 1983 and applies to all those seafarers who serve in vessels registered in the United Kingdom above a certain size (a small coaster upwards). This does not necessarily apply to vessels registered in other countries. These regulations will almost certainly change, the likely outcome of current discussions being that acceptable levels of risk will be defined, both for the individual and for the job.

Retirement and end of life

Despite modern treatments some patients will experience multiple coronary events which eventually lead to extensive ventricular damage and persisting symptoms of fatigue and dyspnoea, with signs of heart failure. Such individuals should be warned of their limited prognosis. Some should be advised to retire, which is never an easy decision.

There is often a discrepancy between the symptoms and the objective cardiac data. Some patients—typically the overweight, smoking, manual worker in his forties, who has always enjoyed robust good health—appear to be very symptomatic despite good ventricular function and no evidence of myocardial ischaemia. A heart attack proves devastating. One explanation for this is the profound psychological disturbance that sometimes follows the development of cardiovascular disease. At the other extreme some patients seem well and active despite appalling ventricular function. As always, common sense has to override the results of investigations, but the latter group are still liable to experience sudden cardiac death, when apparently 'so well'.

Patients (with their partners if appropriate) should be given the opportunity of a frank discussion about their prognosis, but some would rather not know and that attitude should be respected. However, most need to put their affairs in order: what to say exactly is one of the most difficult problems in cardiology. The victim's quality of life may be excellent. There is no point in advising a restricted lifestyle or retirement. The only lifestyle trigger of a heart attack, namely sudden unexpected vigorous exercise, should be avoided. Otherwise, normal activities should continue, with the knowledge that the chance of successful resuscitation following a coronary event are greater in developed countries, in fact better in most of Europe and North America than the United Kingdom.

Early retirement on grounds of ill health following a heart attack is sometimes seen as an attractive option. However, most permanent sickness policies contain a clause which states that benefit will only be payable if the subscriber is 'totally unable to follow his former occupation', which is often not the case after a heart attack or coronary artery bypass grafting. Advice about retirement should only be given after due consideration and a review of the job description.

Further reading

Baxter PJ, Petch MC (2000). Cardiovascular disorders. In: Cox RAF, Edwards FC, Palmer K, eds. *Fitness for work*, 3rd edn, pp. 349–70. Oxford Medical Publications.

Joy MD, ed. (1999). Second European Workshop in Aviation Cardiology. *European Heart Journal* Suppl D.

Petch MC (1998). Task Force Report: driving and heart disease. *European Heart Journal* **19**, 1165–77.

Taylor J (1995). In: *Medical aspects of fitness to drive*. Medical Commission on Accident Prevention, London.

15.5.1 Pharmacological management of heart failure

J. K. Aronson

There are three aims in treating heart failure: if possible, to remove causative factors and so reverse the condition; otherwise, to relieve symptoms and to improve survival. Examples of reversible causes include valvular disease, hypertension, anaemia, and hyperthyroidism. Symptomatic relief can be produced by the use of diuretics to relieve fluid retention, vasodilators to reduce the workload of the heart, and digitalis to increase cardiac contractility. Positive inotropic drugs other than digitalis (for example, inhibitors of phosphodiesterase type III) are not currently used in the long-term treatment of congestive heart failure, because they are associated with increased mortality. By contrast, reduction in mortality during long-term treatment of heart failure can be achieved with angiotensin-converting enzyme (**ACE**) inhibitors, spironolactone, beta-blockers, and the combination of hydralazine with a nitrate.

Mechanisms of action of drugs used to treat heart failure

The ways in which drugs affect the major pathophysiological abnormalities of heart failure are shown in Fig. 1, and a list of the drugs used is given in Table 1.

Fig. 1 The pathophysiology of cardiac failure and the sites and mechanisms of action of drugs used in its treatment. (Taken with permission from Grahame-Smith DG, Aronson JK (2001). *The Oxford textbook of clinical pharmacology and drug therapy*, 3rd edn. Oxford University Press, Oxford.)

The effects of heart failure and of some of the drugs used to treat it on the relation between cardiac output and ventricular end-diastolic pressure (the Frank–Starling curve) are shown in Fig. 2. In established heart failure the curve is displaced downwards. It may be possible to increase cardiac output (for example, by increased endogenous sympathetic drive), but that can only happen at the expense of an increased ventricular end-diastolic pressure, and eventually signs and symptoms of congestion occur. If cardiac output cannot be increased, the signs are of low output (for example in cardiogenic shock).

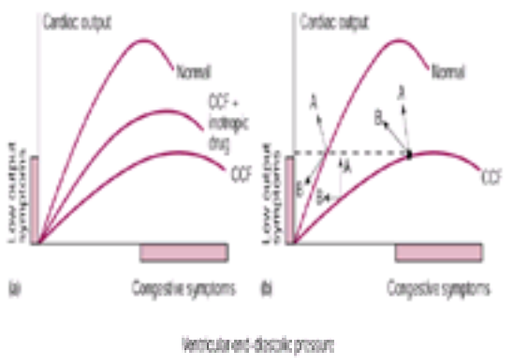

Fig. 2 The Frank–Starling curve, shown here as the relation between ventricular end-diastolic pressure and cardiac output. (a) The effect of positive inotropic drugs: inotropic drugs increase the cardiac output for any given value of end-diastolic pressure. (b) The effects of vasodilators: the effects of vasodilators depend on whether they are predominantly arterial vasodilators (A: for example hydralazine) or mixed vasodilators (B: for example ACE inhibitors). ((a) Adapted with permission from Mason DT (1973). *American Journal of Cardiology* **32**, 437–48. (b) Adapted with permission from Braunwald E (1980). *Heart disease*, p. 548. WB Saunders, Philadelphia.)

Diuretics

Sodium and water retention occur in heart failure through a combination of mechanisms, including reduced renal blood flow, increased ADH secretion, and increased renin secretion, leading to increased secretion of angiotensin and aldosterone. Diuretics reduce the body sodium and water content. Their sites of action in the renal tubule are shown in Fig. 3.

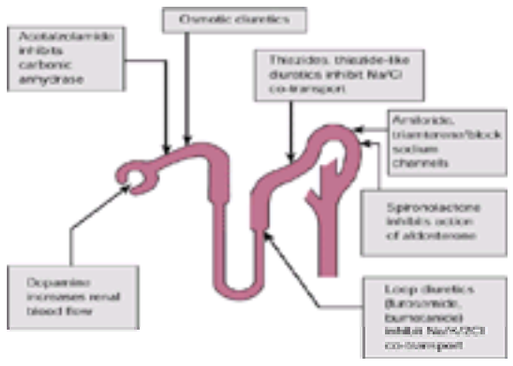

Fig. 3 The sites of action of diuretics in the nephron (Taken with permission from Grahame-Smith DG, Aronson JK (2001). *The Oxford textbook of clinical pharmacology and drug therapy*, 3rd edn. Oxford University Press, Oxford.)

The loop diuretics furosemide and bumetanide inhibit sodium and chloride reabsorption in the ascending limb of the loop of Henle, with a resulting increase in sodium excretion and a reduction in free-water clearance. They do this by inhibiting Na/K/2Cl cotransport. In addition potassium secretion in the distal convoluted tubule is increased, because of the exchange of potassium for sodium under the influence of aldosterone and an increased intraluminal sodium concentration. Note, however, that the effects of furosemide and bumetanide in the treatment of acute left ventricular failure occur more quickly than would be expected from the rate of onset of their diuretic actions; vasodilator effects may be involved in their acute actions.

The thiazide diuretics act by inhibiting sodium and chloride reabsorption in the distal convoluted tubule of the nephron, resulting in increased sodium and free-water clearance. The molecular mechanism of this effect is through the inhibition of a Na/Cl cotransport system. A secondary effect is the loss of potassium by increased secretion in the distal tubule in response to the increased intraluminal sodium concentration.

Spironolactone and its active metabolite canrenone are aldosterone receptor antagonists; they counteract the effects of hyperaldosteronism that can occur from heart failure itself, and as a secondary response to the natriuresis produced by other diuretics. In a low dosage (25 mg/day) spironolactone reduces mortality in congestive heart failure, presumably by inhibiting the effects of aldosterone, circulating concentrations of which are markedly increased in heart failure. Other potassium-sparing diuretics (amiloride and triamterene) do not interfere with the action of aldosterone; instead they inhibit sodium channels in the distal tubule. They are used only as potassium-sparing diuretics and have not been shown to affect mortality.

The effect of diuretics on the Frank–Starling curve is to lower the ventricular end-diastolic pressure.

ACE inhibitors and angiotensin receptor antagonists

ACE inhibitors, which reduce mortality in chronic heart failure, reduce the production of angiotensin II and prevent the breakdown of bradykinin, both of which are mediated by the angiotensin-converting enzyme (ACE). They are mixed arteriolar/venular vasodilators, mostly by their action on angiotensin production. Because angiotensin causes aldosterone release, the ACE inhibitors also reduce aldosterone production, which reduces sodium retention, leading to reduced blood volume and a fall in cardiac preload.

The ACE inhibitors do not completely prevent the effects of angiotensin on the myocardium, because some angiotensin II is produced by the action of a convertase that is not inhibited by ACE inhibitors. Angiotensin II receptor antagonists, which are selective for angiotensin II type 1 receptors, act beyond this point and prevent the effects of angiotensin II at its site of action. They do not affect the production of bradykinin.

b-Adrenoceptor antagonists

In chronic heart failure, particularly in patients with milder disease, the risk of sudden death is increased; b-blockers reduce that risk through an antiarrhythmic action. They also mitigate the effects of catecholamines, which are produced in excess due to increased sympathetic nervous system activity in heart failure. In some patients there is poor ventricular relaxation during diastole, which can be obviated by b-blockers. However, great care must be taken, because their negative inotropic effect can lead to worsening of heart failure. In addition to being a b-blocker, carvedilol is an a $_1$-adrenoceptor antagonist and therefore also a vasodilator.

Positive inotropic drugs

Positive inotropic drugs increase cardiac output at any given value of ventricular end-diastolic pressure, thus shifting the Frank–Starling curve upwards. Of drugs with positive inotropic effects only the cardiac glycosides are currently used in the long-term treatment of congestive heart failure. They act by inhibiting sodium transport out of cells, through inhibition of the Na+/K+ pump enzyme Na+/K+-ATPase. The resultant increase in the intracellular sodium concentration leads to altered calcium flux via the Na+/Ca2+ exchange mechanism, and thus to an increased intracellular calcium concentration. This leads to increased contractility through excitation–contraction coupling. The long-term use of other positive inotropic drugs increases mortality in heart failure; digoxin does not.

Other vasodilators

Vasodilators reduce the workload of the heart by dilating arterioles or venules, or both. Dilatation of arterioles results in a reduction in cardiac afterload; dilatation of venules results in a reduction in cardiac preload. A pure reduction in afterload increases the cardiac output at a given ventricular end-diastolic pressure, while a pure reduction in preload reduces the ventricular end-diastolic pressure and hence the cardiac output along the Frank–Starling curve. In practice, vasodilators cause both these effects. This is because a reduction in arterial resistance (reduction in afterload) increases ventricular emptying (which in turn reduces preload), and venous dilatation (reduction in preload) reduces ventricular volume (which in turn reduces afterload). In both cases cardiac output increases and ventricular end-diastolic pressure falls. However, the extent to which these two effects occur depends on whether the vasodilator acts predominantly on arterioles or venules. For example, nitrates are mixed vasodilators, but the venous element predominates, reducing filling pressure. The effects of different vasodilators on arterioles and venules are shown in Table 2.

Clinical pharmacology of drugs used to treat heart failure

Diuretics

All the thiazide diuretics are well absorbed and excreted unchanged by the kidney. They have a slow onset and long duration of action, with half-lives of about 8 to 12 h, and are given once a day. Hypokalaemia, hyponatraemia, and dehydration are their most important adverse effects, and hypomagnesaemia can also occur. Hypokalaemia due to thiazide diuretics potentiates the effects of cardiac glycosides and can cause ventricular arrhythmias, such as *torsade de pointes*, in patients taking antiarrhythmic drugs that prolong the QT interval. Hypercalcaemia can occur in susceptible patients, due to reduced urinary calcium excretion. Hyperglycaemia occurs but is usually not of clinical importance, although occasionally diabetes mellitus may be precipitated in a susceptible patient; increased doses of oral hypoglycaemic drugs may be required in diabetics. Hyperuricaemia occurs, but acute gout is uncommon. Erectile impotence can occur in men. Thiazide diuretics reduce the clearance of lithium by the kidney; lithium dosages should be halved initially and adjusted with careful serum concentration monitoring.

In contrast to the thiazide diuretics, furosemide and bumetanide have a rapid onset and short duration of action (about 6 to 8 h), with half-lives of between 1 and 2 h. Bumetanide is well absorbed after oral administration, but furosemide is not, and its absorption may be slowed in patients with congestive heart failure. For this reason patients who do not respond to oral furosemide should be given intravenous furosemide or oral bumetanide instead. Despite their short duration of action these drugs are usually given only once a day, partly to avoid night-time diuresis and partly because, at least in the case of furosemide, the kidney is refractory to further diuresis for about 8 h after an effective dose. If furosemide is given intravenously in doses of 80 mg or more it should be infused at a rate of 4 mg/min, partly to avoid ototoxicity and partly because it is more effective when infused slowly, for reasons that are not understood. The adverse effects and interactions of the loop

diuretics are similar to those of the thiazides, except that the loop diuretics are calciuric. Acute urinary retention can be precipitated by too rapid a diuresis in patients with prostatic hyperplasia. Encephalopathy can be precipitated in patients with hepatic insufficiency, particularly if hypokalaemia occurs. Rapid intravenous injection of high doses of furosemide can cause cochlear damage, which is usually reversible. Bumetanide can occasionally cause muscle cramps, independent of hypokalaemia.

Spironolactone is well absorbed. It has a short half-life (about 10 min) but is metabolized to the active compound canrenone (half-life 16 h), which is excreted by the kidney. Partly because of the long half-life of its metabolite, spironolactone has a long duration of action and its maximum effects develop over several days. Nausea and vomiting are common, but not with the low doses currently used to treat heart failure. Hyperkalaemia can occur and is dose-related. Gynaecomastia is common, even with low dosages, and is often painful. Other less frequent effects include menstrual disturbances, impotence, testicular atrophy, and peptic ulceration.

Amiloride is very poorly absorbed. It is almost completely excreted unchanged in the urine and has a half-life of 6 h. Triamterene is incompletely but fairly rapidly absorbed from the gastrointestinal tract. It is extensively metabolized before urinary excretion and its half-life is 2 h. It has variable biliary excretion. Both amiloride and triamterene commonly cause hyperkalaemia, dehydration, and hyponatraemia. The incidence of hyperkalaemia (about 5 per cent) is unaffected by the concurrent administration of potassium-depleting diuretics. Nausea and vomiting occur occasionally. Triamterene can cause crystalluria and rarely causes interstitial nephritis, particularly when it is used in combination with thiazide diuretics; renal prostaglandins may protect the kidney against this damage and this protection may be lost if patients also take non-steroidal anti-inflammatory drugs.

Angiotensin-converting enzyme inhibitors

The ACE inhibitors are variably absorbed: captopril and enalapril are well absorbed, ramipril is moderately well absorbed, and lisinopril is slowly and poorly absorbed (25 per cent or less); the absorption of captopril and enalapril is reduced to 50 per cent by food. Lisinopril is eliminated unchanged in the urine; captopril is 50 per cent excreted unchanged and 50 per cent metabolized to inactive compounds; and the other ACE inhibitors are prodrugs that are metabolized to active forms. The half-lives of the ACE inhibitors are long, because they bind to ACE in the plasma.

Hypotension can occur with ACE inhibitor overtreatment—particularly if the intravascular volume is depleted by concurrent diuretic therapy—and is most common after the first dose, which should therefore be low and taken whilst the patient is lying down. Cyclo-oxygenase inhibitors, such as indometacin, can reduce the hypotensive effects of the ACE inhibitors.

The ACE inhibitors can cause renal function impairment, particularly in those with renovascular disease and especially unilateral renal artery stenosis; this is because the perfusion pressure in the ischaemic kidney depends on the action of locally produced angiotensin. Proteinuria and the nephrotic syndrome are uncommon adverse effects; the latter is due to a membranous glomerulonephritis. ACE inhibitors inhibit the excretion of lithium.

The ACE inhibitors cause potassium retention by inhibiting aldosterone secretion, and potentiate the effects of other drugs that cause hyperkalaemia, for example, potassium-sparing diuretics or potassium chloride supplements.

Rashes are common (up to 10 per cent) and may be accompanied by fever and eosinophilia. Taste disturbance, which is usually transient, occurs in up to 5 per cent of patients. Cough is the commonest adverse effect requiring drug withdrawal. Angio-oedema occurs rarely. Although neutropenia is rare, it may progress to agranulocytosis.

b-Adrenoceptor antagonists (b-blockers)

Some properties of some commonly used b-blockers are shown in Table 3. Most are well absorbed; atenolol, nadolol, and sotalol, being relatively polar, are exceptions, at 50 per cent or less. During their first passage through the liver carvedilol, metoprolol, and propranolol are extensively metabolized to active compounds, which are excreted in the urine. Bisoprolol is 50 per cent metabolized, but without first-pass elimination. Atenolol and sotalol are mostly eliminated unchanged in the urine. Those b-blockers that are used to treat patients in heart failure (bisoprolol, carvedilol, and metoprolol) have durations of action that roughly correlate with their half-lives.

Blockade of b_2-adrenoceptors in the lungs in susceptible subjects can cause bronchoconstriction, which can lead to life-threatening acute severe asthma. Hence, non-selective b-blockers should not be given to asthmatics, and even relatively selective b-blockers should be used with caution, if at all, since none is completely devoid of some b_2-adrenoceptor antagonist activity.

b-Blockers have negative inotropic effects: careful dosage titration and monitoring of therapy is therefore important in patients in whom they are being used to treat heart failure. In patients with poor left ventricular function after myocardial infarction D-sotalol increased mortality from 3 to 5 per cent (the SWORD trial), probably because of cardiac arrhythmias secondary to prolongation of the QT interval.

Central nervous system effects (depression, hallucinations, sleep disturbances) are more common with lipophilic drugs, which enter the brain well (see Table 3). Peripheral vasoconstriction, resulting in Raynaud's phenomenon, which is particularly troublesome in cold weather, is a common complaint, the precise mechanism of which is still not understood.

If a b-blocker is to be withdrawn, this should be done slowly, since abrupt withdrawal can cause a rebound increase in anginal symptoms or frank myocardial infarction, possibly related to adaptive b-adrenoceptor supersensitivity in response to chronic blockade.

Carvedilol can cause postural hypotension because of its a-blocking action.

Drug interactions with b-blockers occur through a variety of mechanisms:

- The effects of insulin and oral hypoglycaemic drugs are potentiated by b-blockers, and hypoglycaemia can result. There is some evidence that this effect is more pronounced with non-cardioselective b-blockers. This interaction is distinct from the effect of b-blockers in blocking the peripheral clinical response to hypoglycaemia, except for sweating, which is a sympathetic nervous function not served by catecholamines.
- Cimetidine inhibits the first-pass metabolism of propranolol and metoprolol and the metabolism of bisoprolol.
- When b-blockers and verapamil are used concurrently there is an increased incidence of bradyarrhythmias and an increased risk of heart failure. There have also been reports of asystole attributed to the use of the combination.
- Sotalol should not be used in combination with other antiarrhythmic drugs that prolong the QT interval (for example, amiodarone, disopyramide, procainamide, quinidine), nor with the antimalarial drug halofantrine, which does the same.
- Monitoring the plasma digoxin concentration during carvedilol therapy is recommended, since carvedilol has been reported to increase plasma digoxin concentrations.

Cardiac glycosides

Digoxin is moderately well absorbed from tablets (about 67 per cent) and better from elixir (80 per cent) and encapsulated elixir (more than 90 per cent). It is mostly eliminated unchanged by the kidneys, with a half-life of about 40 h when renal function is normal, increasing to about 5 days in complete anuria. Dosages must therefore be reduced in renal insufficiency (see below). It has a fast onset of action after oral administration so that intravenous administration is rarely justified; intramuscular injection is painful and causes muscle necrosis and should be avoided.

Digitoxin, in contrast, is almost completely absorbed after oral administration, has a long half-life (about 5 days), and is eliminated by hepatic metabolism. This makes its effects less predictable than those of digoxin, because hepatic metabolism is more variable than renal excretion. However, some prefer it to digoxin, particularly when there is severe renal insufficiency. There are no advantages to using other cardiac glycosides (such as ouabain, acylated forms of digoxin, gitoformate, gitoxin, k-strophanthin, pengitoxin, and proscillaridin).

The adverse effects of the cardiac glycosides are dose-related. Common non-cardiac effects include: anorexia, nausea, vomiting, and diarrhoea; confusion and acute

psychiatric disturbances, particularly in old people; and visual disturbances (photophobia, blurring of vision, disturbances of colour vision). Virtually any cardiac arrhythmia can occur, the commonest being ventricular and supraventricular ectopic arrhythmias. Atrioventricular nodal conduction can be impaired, leading to heart block. The combination of an ectopic arrhythmia and heart block (for example, paroxysmal supraventricular tachycardia with block) is particularly suggestive of glycoside toxicity. Bradycardia occurs occasionally, but is often simply an effect of parasympathetic stimulation in a resting patient without glycoside intoxication.

The adverse effects of cardiac glycosides are enhanced by electrolyte disturbances, especially: hypokalaemia, hypercalcaemia, and hypomagnesaemia; hypoxia and acidosis; hypothyroidism; and old age (due to increased tissue sensitivity).

Drug interactions are common with digoxin. Hypokalaemia due to other drugs (for example diuretics) markedly enhances its effects and should be avoided. Drugs that inhibit P-glycoprotein, which mediates the renal tubular secretion of digoxin, increase plasma digoxin concentrations and the risk of toxicity; these include amiodarone, ciclosporin, quinidine, spironolactone, and many of the calcium channel blockers, notably verapamil. Quinidine also alters the tissue distribution of digoxin and reduces its non-renal clearance; this combination is better avoided. The antibiotics erythromycin, clarithromycin, and tetracycline increase the oral systemic availability of digoxin by inhibiting its breakdown by intestinal bacteria, mainly *Eubacterium glenum.* The metabolism of digitoxin is increased via enzyme induction by drugs such as rifampicin and barbiturates.

Cardiac glycoside plasma concentrations should be carefully monitored. This can be of value in individualizing therapy, in monitoring compliance, and in diagnosing digitalis toxicity. It is worth measuring the plasma (or serum) concentration during the initial stages of therapy to ensure that a reasonable target concentration has been achieved (1.0 to 2.0 nmol/l for digoxin, 10 to 20 nmol/l for digitoxin). A cautious increase in dosage is justifiable if there is still a poor response to treatment, but the risk of toxicity starts to rise markedly at plasma concentrations above 2.0 and 20 nmol/l respectively. If there are subsequent changes in the patient's condition, for example renal insufficiency in a patient taking digoxin, measurement of the plasma concentration may help in readjusting dosages. Toxicity is highly likely at plasma concentrations above 3.0 nmol/l (digoxin) or 30 nmol/l (digitoxin); at concentrations below 1.5 or 15 nmol/l respectively, toxicity is unlikely. However, toxicity can occur even with low concentrations and should particularly be suspected if there is hypokalaemia. Certain factors besides potassium depletion increase the risk of digitalis toxicity at a given plasma concentration (see above); these alter the interpretation of the plasma concentration and lower the threshold of suspicion of toxicity.

Other vasodilators

Hydralazine

Hydralazine is well absorbed and extensively metabolized, principally by acetylation. This has a bimodal distribution in the general population, but the half-life of hydralazine (about 4 h) does not differ much between people who are slow and fast acetylators. This is because acetylation occurs mainly during the first passage through the liver, hence the subsequent rate of clearance is not appreciably related to the rate of acetylation; however, patients who are slow acetylators are exposed to more of the parent compound.

Palpitation and tachycardia, nausea, vomiting, diarrhoea, and postural hypotension are all common adverse effects. An arthropathy resembling rheumatoid arthritis or a syndrome similar to that of systemic lupus erythematosus (so-called lupus-like syndrome) can occur with dosages over 200 mg/day, especially in those who are slow acetylators. Hydralazine-induced lupus is more common in patients with the HLA phenotype DR4.

Nitrates

In contrast to glyceryl trinitrate, which is completely metabolized in the liver after oral administration and cannot therefore be given orally, isosorbide dinitrate is absorbed from the gut and extensively metabolized to its active metabolites: especially isosorbide mononitrate, which is itself metabolized. Isosorbide dinitrate and isosorbide mononitrate are therefore active after oral administration, the half-life of isosorbide dinitrate being 1 h and that of isosorbide mononitrate being 4 h, which rate-limits the kinetics of isosorbide dinitrate.

Vasodilatation can cause throbbing headache, sinus tachycardia, and hypotension. Tolerance to the actions of the nitrates occurs with prolonged administration, for example if transdermal patches are left on the skin continuously, or if modified-release formulations are taken without a long enough gap between doses. This can be minimized or avoided by removing the patch for a few hours each day (for example, overnight) or by leaving at least 14 h between the night-time dose of a modified-release formulation and the next morning's dose. The mechanism of this tolerance is not known, but hypotheses include depletion of tissue sulphydryl groups (causing reduced transformation of organic nitrates to nitric oxide), desensitization of guanylyl cyclase, increased vascular production of superoxide anions, changes in plasma volume, and increased production of vasopressin. An attractive theory is that chronic vasodilatation in response to nitric oxide causes a compensatory increased production of the vasoconstrictor endothelin via activation of the renin–angiotensin system.

Patients who take sildenafil concurrently with a nitrate may experience profound hypotension; this combination should be avoided.

Angiotensin receptor antagonists

Candesartan esters are rapidly and completely de-esterified during absorption to candesartan, the systemic availability of which is low, and which is partly metabolized and partly excreted in the bile. Irbesartan is rapidly and completely absorbed and subject to only slight presystemic metabolism; it is metabolized by glucuronidation and oxidation by CYP2C9. Losartan is well absorbed but is subject to extensive presystemic metabolism by CYP2C9 and CYP3A4—one metabolite is active, and is inactivated by further metabolism. The metabolites are excreted in the bile and systemic availability is doubled in patients with liver disease. Valsartan has a systemic availability of about 0.25, which is reduced to about 0.15 by food. It is mostly excreted in the bile. The half-lives of these drugs or their active metabolites range from 6 to 12 h.

As with the ACE inhibitors, hypotension can occur with overtreatment, particularly if the intravascular volume is depleted by concurrent diuretic therapy. Similarly, the angiotensin receptor antagonists should be used with caution in patients with renal insufficiency, since they can cause impairment of renal function, particularly in those with renovascular disease. Hyperkalaemia can occur, but is uncommon. Angio-oedema is rare. Because they do not affect kinins the angiotensin receptor antagonists do not cause cough.

Practical management of heart failure

Acute left ventricular failure

Acute left ventricular failure producing pulmonary oedema is a medical emergency, requiring treatment with oxygen, morphine or diamorphine, a loop diuretic, and vasodilators if required.

Oxygen should be given in a high concentration by face-mask or nasal cannulae. Furosemide 40 mg, or bumetanide 1 mg, is given intravenously (**IV**), followed by 10 mg of morphine IV, via the same needle. If there is a poor response to this regimen, the dose of morphine is repeated and a higher dose of diuretic given. Intravenous vasodilators should be used in patients with severe left ventricular failure, for example glyceryl trinitrate or isosorbide dinitrate, but glyceryl trinitrate can be given sublingually (0.5 mg) or by transdermal patch (5 mg) as a stop-gap. Such treatment should ideally be monitored with the measurement of pulmonary artery wedge pressure using a Swan–Ganz catheter, but this may not be practicable, in which case careful monitoring of the systemic blood pressure is necessary to avoid a systolic pressure below 95 mmHg. Glyceryl trinitrate is given as an IV infusion through a syringe pump in a dosage of 10 to 200 µg/min—starting with no more than 10 µg/min, increasing as necessary, and monitoring the response. The dose of isosorbide dinitrate by IV infusion is between 2 and 10 mg/h.

Cardiac glycosides can also be used in the treatment of acute left ventricular failure, particularly when this is associated with fast atrial fibrillation. However, they increase the risk of cardiac arrhythmias after myocardial infarction.

In the rare circumstance of acute left ventricular failure due to acute severe hypertension, the blood pressure should be lowered rapidly (see Chapter 15.16.3).

Acute left ventricular failure due to iatrogenic fluid overload can be prevented by the use of a loop diuretic. For example, during blood transfusion in a patient with chronic anaemia and therefore a normal intravascular volume, furosemide 20 mg IV should be given immediately before each unit of blood, which should be infused

as slowly as possible.

Chronic heart failure

Diagnosis on the basis of symptoms and signs, assisted where possible by echocardiography, should be made as soon as possible, because of the improvement in prognosis promised by ACE inhibitors and spironolactone; such therapy should be begun as early as possible. Information on left ventricular function is immensely helpful in assessing severity, appropriate treatment, response, and prognosis. Appropriate action is indicated if a primary cause is selectively correctable, unsuspected aortic valve disease in the elderly being a common example.

A drug history is essential. Non-steroidal anti-inflammatory drugs cause sodium retention and can tip patients into heart failure. Many calcium channel blockers are negatively inotropic, and short-acting calcium blockers increase the risk of heart failure and should be avoided. Some antiarrhythmic agents are negatively inotropic. Most b-blockers (outside of their careful use, as described below) can be deleterious. Tricyclic antidepressants are best avoided.

Salt intake should be moderated. Hypertension must be treated, and ACE inhibitors can relieve both heart failure and hypertension. It is important that plasma electrolytes, urea, and creatinine be measured, so that renal function can be monitored during treatment.

Assuming that drug therapy is indicated and there is left ventricular dysfunction, then, although diuretics may bring symptomatic relief, an ACE inhibitor should also be used. These two drug categories are now the mainstay of therapy and can be supplemented by spironolactone, which also reduces mortality. In severe cases or when ACE inhibitors are contraindicated or poorly tolerated, there is a case for further vasodilator therapy with hydralazine plus isosorbide dinitrate or mononitrate. Cardiac glycosides as positive inotropic agents still have a place.

It is not yet known whether different types of heart failure require different types of pharmacological management. Currently there is not enough evidence to guide the selection of therapy in different patients, and choices among different types of drugs are generally made on the basis of contraindications and adverse effects rather than positive indications.

Diuretics

The choice of diuretic depends on the severity of heart failure. In mild heart failure a thiazide or thiazide-like diuretic is sufficient (see Table 1). The most commonly used of these diuretics in the United Kingdom are bendroflumethiazide (5 to 10 mg once daily, orally) and cyclopenthiazide (0.5 to 1.0 mg once daily, orally), but there is no particular advantage in using one thiazide rather than another.

Oral loop diuretics are used in cases of more severe heart failure; for example, furosemide 40 to 160 mg once daily or bumetanide 1 to 5 mg once daily. If there is a poor response to either a thiazide or a loop diuretic, the two types can be combined.

In the Randomized Aldactone Evaluation Study (**RALES**), 822 patients were randomly assigned to receive spironolactone 25 mg/day and 841 to receive placebo. There were significantly fewer deaths in the spironolactone group (284 versus 386), with fewer deaths from progressive heart failure and fewer sudden deaths from cardiac causes. There was also a reduced frequency of hospital admission for worsening heart failure and a significant improvement in the symptoms of heart failure. Most of the patients were taking an ACE inhibitor. Spironolactone 25 mg/day should be given routinely to all patients with congestive heart failure.

In using potassium-wasting diuretics care should be taken to avoid hypokalaemia, especially in old people and in patients taking cardiac glycosides. In patients who are also taking an ACE inhibitor and even a low dose of spironolactone, extra measures to conserve potassium are generally unnecessary. It is not known whether higher doses of spironolactone than 25 mg/day are also associated with a beneficial effect on mortality. If extra potassium-sparing is required it is probably wise to use potassium chloride supplements (to repair depletion) or another potassium-sparing diuretic (amiloride or triamterene) (to prevent further depletion).

ACE inhibitors

ACE inhibitors are now widely considered to be the first-line treatment for chronic heart failure, in combination with a diuretic. Several large studies have shown they improve symptoms and reduce morbidity and mortality in patients with left ventricular dysfunction. A systematic review of 32 randomized, controlled clinical trials in symptomatic heart failure showed that ACE inhibitors reduced mortality by 28 per cent, independent of the ACE inhibitor used. Treatment with an ACE inhibitor also reduced the number and duration of hospital admissions. There was a reduction in mortality of about 8 per cent in patients with asymptomatic heart failure (but see the results of the HOPE study mentioned below).

When starting treatment with an ACE inhibitor in a patient who is already taking a diuretic, particularly a high-efficacy (so-called 'high ceiling') loop diuretic, care must be taken not to cause serious hypotension. If a patient is in severe heart failure and taking a large dosage of a loop diuretic, or is hypovolaemic, or has hyponatraemia (plasma sodium concentration of 130 mmol/l or less), has renal impairment, is taking other vasodilator therapy, or is frail and elderly, it is wise to admit them to hospital for initiation of therapy. Diuretics should be stopped for 24 h, a low dose of a short-acting ACE inhibitor (preferably captopril) should be given, blood pressure should be monitored both when the patient is lying and standing, and the dosage of the ACE inhibitor should only be increased when one is satisfied that serious hypotension has not occurred. If the systolic blood pressure is less than 100 mmHg, if there is clinical heart failure, and if diuretic therapy has already been given, it is unlikely that ACE inhibitor therapy will be tolerated, but each case must be taken on its merits.

Despite all these cautions, it is possible to start some patients on an ACE inhibitor in the community. Diuretics should be stopped for at least 24 h; the first low dose should preferably be given in an environment in which blood pressure monitoring is possible; and the dosage should be increased very gradually.

The effects of ACE inhibitors in the treatment of heart failure are dose-related. The recommended maximal doses are those that have been shown to be efficacious and set as limits beyond which toxicity, namely hypotension, becomes a frequent and unacceptable problem. Thus, the maximal dose for captopril is 50 mg three times a day, for enalapril 20 mg/day, and for lisinopril 20 mg/day. Of these, the author prefers enalapril, which has a longer duration of action than captopril and is better absorbed than lisinopril. However, since the publication of the results of the Heart Outcomes Prevention Evaluation (**HOPE**) study, ramipril has become more widely used. In this study 9297 high-risk patients (aged 55 or over), with vascular disease or diabetes plus one other cardiovascular risk factor and who were not known to have a low ejection fraction or heart failure, were randomly assigned to receive ramipril (10 mg/day) or placebo. Ramipril reduced the death rates from cardiovascular causes (relative risk 0.74), myocardial infarction (0.80), stroke (0.68), death from any cause (0.84), revascularization procedures (0.85), cardiac arrest (0.63), heart failure (0.77), and complications related to diabetes (0.84).

b-Adrenoceptor antagonists (b-blockers)

Despite much evidence that b-blockers are beneficial in patients with chronic heart failure, particularly in preventing sudden death in patients with mild or moderate disease, there is understandable reluctance to use them, because of the risk of worsening heart failure through impaired myocardial contractility. Certainly they should not be used in patients with severe heart failure (New York Heart Association class IV); indeed, in patients with severe heart failure the b-blocker xamoterol increases mortality. However, if b-blockers are used in patients with milder forms of heart failure, and in very low initial doses with very gradual dosage increases, the evidence is that they are relatively safe and reduce all-cause mortality (risk ratio about 0.7) and cardiac deaths (risk ratio about 0.6). As yet, however, there is no evidence about their efficacy in direct comparison with ACE inhibitors and spironolactone, nor information about whether the combination of a b-blocker with such drugs produces further increases in benefit. Current trials are comparing different b-blockers with each other and b-blockers with ACE inhibitors. It is not known whether digoxin can mitigate the negative inotropic effect of b-blockers in patients with chronic heart failure (although one would expect it to do so), nor how such a combination would affect mortality.

In my view, b-blocker therapy should currently be initiated only by a specialist in hospital. It should be limited to patients with moderate heart failure at worst, despite optimal doses of diuretics, an ACE inhibitor, spironolactone, and digoxin, if indicated. For carvedilol the initial dose should be 3.125 mg and the patient should be observed for a few hours after the first dose. If there is no evidence of worsening heart function 3.125 mg can be given twice daily for 2 weeks, after which a further dosage increase to 6.25 mg twice daily can be attempted. The dosage can be further increased, no more often than every 2 weeks, to a maximum of 50 mg twice daily, and at each stage the patient should be carefully monitored for a few hours after the first dose for evidence of worsening cardiac function. Daily monitoring of body weight is also important, and if a patient's weight increases by 2 kg or more they should report to their doctor; an increase in the dosage of diuretic may help in such cases. The corresponding initial and maximum doses of other b-blockers of proven efficacy in heart failure are: bisoprolol 1.25 mg/day and 10 mg/day;

metoprolol 12.5 mg/day and 200 mg/day. The class III antiarrhythmic drug D-sotalol increases mortality after myocardial infarction, suggesting that the racemic form DL-sotalol, which also has b-blocking activity, should be avoided in such patients.

Cardiac glycosides

The positive inotropic effects of cardiac glycosides can be useful in reducing symptoms (mainly breathlessness) in patients already taking diuretics and ACE inhibitors. However, the beneficial effect is small and digoxin does not reduce mortality during long-term therapy. However, it does have a small effect in reducing hospital admission rates; there is also evidence that heart function deteriorates in about 25 per cent of patients after withdrawal, so there may still be a role for digoxin in a few patients who need extra symptomatic relief. It is also the drug of choice for treating atrial fibrillation in patients with congestive heart failure. Set against this is the difficulty in using it properly and the high risk of toxicity, particularly in older people, who have poor renal function and are prone to hypokalaemia, particularly if they are also taking diuretics.

Cardiac glycosides should not be used, or are ineffective, in the following conditions:

- left ventricular outflow obstruction (for example, aortic stenosis, hypertrophic obstructive cardiomyopathy), since they increase the force of contraction against a fixed obstruction;
- constrictive pericarditis, for an analogous reason;
- chronic cor pulmonale, because of reduced efficacy and an increased risk of toxicity, perhaps secondary to hypoxia and acidosis;
- hyperthyroidism, because of reduced efficacy and an increased risk of toxicity, although they may be useful in addition to a b-adrenoceptor antagonist in patients with atrial fibrillation and to some extent protect the heart against the negative inotropic effects of b-blockers;
- arrhythmias associated with accessory conduction pathways (for example Wolff–Parkinson–White syndrome), since they impair conduction through the normal conducting pathways without affecting the accessory pathways.

Digoxin is given orally in an initial loading dose of 15 µg/kg, preferably in three divided doses at 6-hour intervals, monitoring for evidence of toxicity (for example, cardiac arrhythmias or symptoms of nausea and vomiting) before the second and third doses. In severe renal insufficiency the initial loading dose should be reduced to 12 µg/kg and increased only in the face of a plasma concentration below the target range (see above). The subsequent maintenance dose should be based on renal function, as shown in Table 4. In patients with atrial fibrillation, in whom the response to digoxin can be easily measured by counting the ventricular rate, an extra dose of 5 µg/kg (in other words, a total loading dose of 20 µg/kg) can be given if necessary; in that case the daily maintenance dose should be increased proportionately. There is no advantage in giving digoxin intravenously; moreover, intramuscular injection is painful and causes muscle necrosis. If digoxin has to be given parenterally in a patient who cannot swallow, it should be infused intravenously over no less than 30 min to avoid the acute hypertension that can occur during rapid intravenous administration. Monitoring therapy by plasma concentration measurement is discussed above.

Other vasodilators

The combination of hydralazine (300 mg/day) with a nitrate (isosorbide dinitrate 160 mg/day) has beneficial haemodynamic effects, improves symptoms, and also reduces mortality, although not as markedly as ACE inhibitors. This combination is also less well tolerated than ACE inhibitors. It should be reserved for patients in whom renovascular disease militates against the use of ACE inhibitors and angiotensin receptor antagonists. It has been suggested that the addition of hydralazine and a nitrate to ACE inhibitor therapy should improve mortality even further, but this hypothesis has not yet been tested in a large clinical trial.

Short-acting calcium channel blockers increase cardiovascular mortality in patients with coronary heart disease and should be avoided in long-term treatment. Modified-release formulations of short-acting calcium blockers and long-acting drugs (such as amlodipine) seem to be safe in this regard, but there is as yet no evidence that they reduce mortality in patients with chronic heart failure. Amlodipine may delay the time to and reduce the number of admissions to hospital in connection with ventricular arrhythmias. However, there is currently no place for the use of calcium blockers in patients with congestive heart failure, except in the treatment of associated conditions such as hypertension or angina pectoris.

Alpha-blockers, such as prazosin, neither improve symptoms nor reduce mortality in chronic heart failure; they should not be used.

Angiotensin receptor antagonists

In patients who cannot tolerate ACE inhibitors because of adverse effects, such as cough or renal insufficiency, angiotensin receptor antagonists seem a logical alternative, and they are certainly better tolerated than ACE inhibitors. However, although there is some evidence that angiotensin receptor antagonists improve symptoms in congestive heart failure, there is currently no evidence that they are better than ACE inhibitors at reducing mortality. There is some early evidence that a combination of an ACE inhibitor with an angiotensin receptor antagonist may be more efficacious than either alone, but this hypothesis awaits proper testing. Until further information becomes available, angiotensin receptor antagonists should be reserved for patients who cannot tolerate ACE inhibitors. Typical once-daily doses are: candesartan 4 to 16 mg; irbesartan 75 to 300 mg; losartan 25 to 100 mg; and valsartan 40 to 160 mg.

Other positive inotropic drugs

Other positive inotropic drugs, including the phosphodiesterase inhibitors (such as amrinone, milrinone, and vesnarinone), ibopamine, and intermittent intravenous dobutamine, all increase mortality during long-term treatment of congestive heart failure and should not be used. However, there is evidence that short-term intravenous milrinone for a few weeks can help achieve haemodynamic stability and tide suitable patients over to heart transplantation. Milrinone has also been used to help wean patients off cardiopulmonary bypass.

Other antiarrhythmic drugs

If sudden death in chronic heart failure is due to cardiac arrhythmias, one would expect other antiarrhythmic drugs to be beneficial. However, the results of the few available studies of the class III antiarrhythmic drug amiodarone are not impressive. In one trial amiodarone 300 mg/day produced a small reduction in mortality, but this has not been confirmed. Some of the data suggest that amiodarone may be more beneficial in patients with non-ischaemic heart failure, but it should currently be reserved for patients with identified ventricular arrhythmias. Class I antiarrhythmic drugs increase mortality after myocardial infarction and should be avoided.

Anticoagulants

There is an increased risk of venous thrombosis in immobile patients who have severe heart failure and have swollen legs due to oedema. Prophylactic anticoagulation is therefore advisable, using either an oral anticoagulant (for example, warfarin) or low-dose subcutaneous heparin. Other patients with cardiomyopathy, severe left ventricular dilatation, or demonstrable intracardiac thrombus on echocardiography should be given anticoagulants to prevent systemic emboli. These issues are discussed in Chapter 15.5.2.

Monitoring therapy in heart failure

Drug therapy in heart failure should be monitored for evidence of therapeutic efficacy and drug toxicity.

Fluid and electrolyte balance and the response to diuretics should be monitored by body weight and serum electrolyte measurements. Potassium depletion can occur, even in patients taking potassium-sparing drugs (ACE inhibitors, angiotensin receptor antagonists, and spironolactone), in which case potassium chloride should be given to replace any deficit and additional amiloride or triamterene to prevent further potassium loss.

Renal function should be monitored before giving an ACE inhibitor or angiotensin receptor antagonist during the first few days or weeks of therapy, and whenever the dosage is increased. Worsening renal function will dictate dosage reduction or drug withdrawal.

Blood pressure should be measured at each visit. It will often fall after the first dose of an ACE inhibitor but will usually improve thereafter.

Serum or plasma digoxin concentration measurement is discussed above.

During anticoagulant therapy with warfarin the target International Normalized Ratio (**INR**) is 2.0 to 2.5 for the prevention of deep vein thrombosis, 2.5 to 3.0 for patients with atrial fibrillation, dilated cardiomyopathy, or mural thrombosis, and 3.5 for recurrent deep vein thrombosis or pulmonary embolism and in patients with prosthetic heart valves.

Further reading

Acute Infarction Ramipril Efficacy (**AIRE**) Study Investigators (1993). Effect of ramipril on mortality and morbidity of survivors of acute myocardial infarction with clinical evidence of heart failure. *Lancet* **342**, 821–8.

Australia/New Zealand Heart Failure Research Collaborative Group (1997). Randomised, placebo-controlled trial of carvedilol in patients with congestive heart failure due to ischaemic heart disease. *Lancet* **349**, 375–80.

Bart BA, *et al*. (1999). Contemporary management of patients with left ventricular systolic dysfunction. Results from the Study of Patients Intolerant of Converting Enzyme Inhibitors (SPICE) Registry. *European Heart Journal* **20**, 1182–90.

CIBIS-II Investigators and Committees (1999). The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial. *Lancet* **353**, 9–13.

Cohn JN, *et al*. (1986). Effect of vasodilator therapy on mortality in chronic congestive heart failure. Results of a Veterans Administration Cooperative Study. *New England Journal of Medicine* **314**, 1547–52.

Cohn JN, *et al*. (1991). A comparison of enalapril with hydralazine–isosorbide dinitrate in the treatment of chronic congestive heart failure. *New England Journal of Medicine* **325**, 303–10.

CONSENSUS Trial Study Group (1987). Effects of enalapril on mortality in severe congestive heart failure. Results of the Cooperative North Scandinavian Enalapril Survival Study (CONSENSUS). *New England Journal of Medicine* **316**, 1429–35.

Cruickshank JM (1993). The xamoterol experience in the treatment of heart failure. *American Journal of Cardiology*, **71**, 61C–64C.

De Vries RJ, Van Veldhuisen DJ, Dunselman PH (2000). Efficacy and safety of calcium channel blockers in heart failure: focus on recent trials with second-generation dihydropyridines. *American Heart Journal* **139**, 185–94.

Dickstein K, Kjekshus J (1999). Comparison of the effects of losartan and captopril on mortality in patients after acute myocardial infarction: the OPTIMAAL trial design. Optimal Therapy in Myocardial Infarction with the Angiotensin II Antagonist Losartan. *American Journal of Cardiology* **83**, 477–81.

Digitalis Investigation Group (1997). The effect of digoxin on mortality and morbidity in patients with heart failure. *New England Journal of Medicine* **336**, 525–33.

Doughty RN, *et al*. (1997). Effects of b-blocker therapy on mortality in patients with heart failure. A systematic overview of randomized controlled trials. *European Heart Journal* **18**, 560–5.

Doval HC, *et al.* (1994). Randomised trial of low-dose amiodarone in severe congestive heart failure. Grupo de Estudio de la Sobrevida en la Insuficiencia Cardiaca en Argentina (GESICA). *Lancet* **344**, 493–8.

Eichhorn EJ, Bristow MR (1997). Practical guidelines for initiation of b-adrenergic blockade in patients with chronic heart failure. *American Journal of Cardiology* **79**, 794–8.

Furberg CD, Psaty BM, Meyer JV (1995). Nifedipine. Dose-related increase in mortality in patients with coronary heart disease. *Circulation* **92**, 1326–31.

Garg R, Yusuf S (1995). Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. Collaborative Group on ACE Inhibitor Trials. *Journal of the American Medical Association* **273**, 1450–6.

Havranek EP, *et al*. (1999). Dose-related beneficial long-term hemodynamic and clinical efficacy of irbesartan in heart failure. *Journal of the American College of Cardiology* **33**, 1174–81.

Heart Outcomes Prevention Evaluation Study Investigators (2000). Effects of ramipril on cardiovascular and microvascular outcomes in people with diabetes mellitus: results of the HOPE study and MICRO-HOPE substudy. *Lancet* **355**, 253–9.

Heart Outcomes Prevention Evaluation Study Investigators (2000). Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *New England Journal of Medicine* **342**, 145–53.

Kober L, *et al*. (1995). A clinical trial of the angiotensin-converting-enzyme inhibitor trandolapril in patients with left ventricular dysfunction after myocardial infarction. Trandolapril Cardiac Evaluation (TRACE) Study Group. *New England Journal of Medicine* **333**, 1670–6.

Krum H (1999) Beta-blockers in heart failure. The 'new wave' of clinical trials. *Drugs* **58**, 203–10.

Massie BM, *et al.* (1996). Effect of amiodarone on clinical status and left ventricular function in patients with congestive heart failure. CHF-STAT Investigators. *Circulation* **93**, 2128–34.

Mehra MR, *et al*. (1997). Safety and clinical utility of long-term intravenous milrinone in advanced heart failure. *American Journal of Cardiology* **80**, 61–4.

MERIT-HF Study Group (1999). Effect of metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL Randomised Intervention Trial in Congestive Heart Failure (MERIT-HF). *Lancet* **353**, 2001–7.

Packer M, *et al*. (1993). Withdrawal of digoxin from patients with chronic heart failure treated with angiotensin-converting-enzyme inhibitors. RADIANCE Study. *New England Journal of Medicine* **329**, 1–7.

Packer M, *et al*. (1996). The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. US Carvedilol Heart Failure Study Group. *New England Journal of Medicine* **334**, 1349–55.

Packer M, *et al*. (1999). Comparative effects of low and high doses of the angiotensin-converting enzyme inhibitor, lisinopril, on morbidity and mortality in chronic heart failure. ATLAS Study Group. *Circulation* **100**, 2312–18.

Pennell DJ, *et al*. (2000). The Carvedilol Hibernation Reversible Ischaemia Trial, Marker Of Success (CHRISTMAS) study. Methodology of a randomised, placebo controlled, multicentre study of carvedilol in hibernation and heart failure. *International Journal of Cardiology* **72**, 265–74.

Pfeffer MA, *et al*. (1992). Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. Results of the survival and ventricular enlargement trial. The SAVE Investigators. *New England Journal of Medicine* **327**, 669–77.

Pitt B, *et al.* (1997). Randomised trial of losartan versus captopril in patients over 65 with heart failure (Evaluation of Losartan in the Elderly Study, ELITE). *Lancet* **349**, 747–52.

Pitt B, *et al*. (1999). Effects of losartan versus captopril on mortality in patients with symptomatic heart failure: rationale, design, and baseline characteristics of patients in the Losartan Heart Failure Survival Study—ELITE II. *Journal of Cardiac Failure* **5**, 146–54.

Pitt B, *et al*. (1999). The effect of spironolactone on morbidity and mortality in patients with severe heart failure. Randomized Aldactone Evaluation Study Investigators. *New England Journal of Medicine* **341**, 709–17.

SOLVD Investigators (1991). Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *New England Journal of Medicine* **325**, 293–302.

Swedberg K, *et al*. (1999). Candesartan in heart failure—assessment of reduction in mortality and morbidity (CHARM): rationale and design. Charm–Programme Investigators. *Journal of Cardiac Failure* **5**, 276–82.

Uretsky BF, *et al*. (1993). Randomized study assessing the effect of digoxin withdrawal in patients with mild to moderate chronic congestive heart failure: results of the PROVED trial. PROVED Investigative Group. *Journal of the American College of Cardiology* **22**, 955–62.

Xamoterol in Severe Heart Failure Study Group (1990). Xamoterol in severe heart failure. *Lancet* **336**, 1–6.

15.5.2Therapeutic anticoagulation in atrial fibrillation and heart failure

David Keeling

Atrial fibrillation

Atrial fibrillation is present in 5 per cent of the population over 65 years of age and in 10 per cent of those over 70 years. It increases the risk of stroke fivefold and is present in 15 per cent of all stroke patients. The overall risk of ischaemic stroke in atrial fibrillation without rheumatic heart disease is about 5 per cent per year, but this can be modified by other risk factors. The risk increases with age: those over 75 years old are at high risk; by contrast, those less than 65 years old with no other risk factors have a risk of stroke of 1 per cent per year. Other factors that increase the risk of stoke in atrial fibrillation have been well characterized (Table 1). It is not clear for how long one must successfully treat patients with hypertension before their risk of stroke decreases, and pending further studies treated hypertension should still be regarded as a major risk factor.

Warfarin decreases the risk of ischaemic stroke in atrial fibrillation by 68 per cent, compared with a reduction of approximately 21 per cent with aspirin. The risk of ischaemic stroke has to be balanced against the risk of intracranial haemorrhage on anticoagulant therapy. The risk of intracranial haemorrhage on warfarin is approximately 0.5 per cent per year, and although this increases with age, so does the risk of ischaemic stroke.

Long-term oral anticoagulation should be considered for those patients with atrial fibrillation who are at high risk of stroke. They should receive warfarin if they are over 75 years old or have one or more of the major risk factors in Table 1 and no contraindication. Oral anticoagulation is recommended instead of aspirin because of the large absolute risk reduction: when the risk of stroke is 7.5 per cent per annum only 20 patients need to be treated for 1 year to prevent a stroke. Patients between 65 and 75 years are at greater risk than those less than 65 years: if they have additional minor risk factors such as diabetes or coronary artery disease, they too should be offered warfarin. Patients between 65 and 75 years of age with no other risk factors and those less than 65 years old with only minor risk factors are at intermediate risk and the choice of oral anticoagulation or aspirin will be significantly affected by patient preference. Those less than 65 years old with no risk factors should receive aspirin (Table 2).

For warfarin the target INR should be 2.5 (range 2.0 to 3.0): lower intensity anticoagulation is not effective and higher intensity increases the risk of bleeding. For those on aspirin it is unnecessary to give more than 75 mg /day.

Anticoagulation after an acute ischaemic stroke should be delayed until most of the deficit has resolved, or for 2 weeks in the case of severe strokes.

Oral anticoagulation should be given for 3 weeks before elective cardioversion of patients who have been in atrial fibrillation for more than 48 h and continued until sinus rhythm has been maintained for 4 weeks. It is not needed if atrial fibrillation has lasted for less than 48 h.

Heart failure

The trials of oral anticoagulation (with high target INRs) after myocardial infarction showed reductions in recurrent myocardial infarction, ischaemic stroke, and all-cause mortality similar to those observed with aspirin. However, aspirin is recommended for long-term therapy in preference to warfarin because of its safety. Warfarin is given to survivors of myocardial infarction at high risk of systemic embolization because of severe left ventricular dysfunction, congestive heart failure, mobile mural thrombus, or atrial fibrillation. In these cases the target INR is 2.5 (target range 2.0 to 3.0). Although it is not clear if this lower intensity protects as effectively against vascular disease, aspirin is not given concurrently because of the increased risk of bleeding. Warfarin should also be considered in congestive heart failure due to causes other than myocardial infarction to reduce the high risk of thromboembolism.

Further reading

Laupacis A *et al.* (1998). Antithrombotic therapy in atrial fibrillation. *Chest* **114**, 579S–589S.

15.5.3Cardiac rehabilitation

Andrew J. S. Coats

Introduction

Cardiac rehabilitation constitutes the use of pharmacological and non-pharmacological treatment modalities to restore a patient to premorbid health, outlook, and activity. It is more than treating a disease: it is the systematic attempt to correct all factors limiting a subject from full participation in the healthy aspects of their predisease lifestyle.

The conventional four phases of rehabilitation constitute: the early interventions during hospital stay (phase 1); followed by the first 6 weeks after discharge (phase 2); then full participation in sport, work, and family life over the next 6 to 12 months (phase 3); and, last, the maintenance of the new healthier condition and lifestyle (phase 4). Secondary prevention of cardiovascular disease by the full use of available techniques is an important part of rehabilitation. The components of a successful cardiac rehabilitation programme, which needs to be well monitored and adequately staffed with a multidisciplinary team, are listed in Table 1.

It has been established that cardiac rehabilitation can increase quality of life, exercise capacity, and the chance of return to work in patients recovering from a myocardial infarction. Meta-analyses of published randomized controlled trials of cardiac rehabilitation programmes which include a structured exercise training component have shown that these are able to produce an approximately 25 per cent reduction in mortality. Rehabilitation programmes also give an opportunity for the effective implementation of secondary preventive measures to reduce the risk of subsequent cardiovascular events.

Rehabilitation in special populations

The benefits and risks of taking part in a cardiac rehabilitation programme after a myocardial infarction depend on the prior morbidity and prognosis of the individual. The lower the cardiovascular risk, and the better the exercise tolerance and state of health of the patient at the start, the lower the chance of any adverse events occurring during rehabilitation. However, such low-risk patients may have relatively little to gain from the programme if their degree of fitness and motivation are already high. By contrast, patients at a higher risk of reinfarction, or with a lower level of motivation and exercise tolerance, may achieve more substantial benefit. However, there is a tendency for rehabilitation to be offered to the lower risk younger patient rather than the higher risk older patient with more medical problems. But given the benefits that can be achieved in those at high risk, special procedures should be instituted to make sure these patients receive rehabilitation services.

The elderly

There is often an effective age restriction for entry to many cardiac rehabilitation programmes, but, physiologically and medically, there is no reason for age *per se* to be a contraindication to rehabilitation. The risk factors for cardiovascular disease may differ in the elderly, with dys- and hyperlipidaemias being quantitatively less important and systolic blood pressure being more important, but it can still be advantageous to identify and reduce these risk factors. Exercise tolerance is more markedly impaired at the outset, but it can be increased by training even in the very elderly. Other diseases frequently coexist in the elderly, and they are more likely to be taking multiple medications, to have a poorer memory, and to have less adequate social support networks for their greater needs. These and many other factors make rehabilitation programmes with an exercise component more problematic and associated with a higher rate of complications. This does not mean, however, that elderly patients should not be encouraged to participate.

The elderly are likely to need a longer and gentler introduction to rehabilitation, more frequent and prolonged contact throughout the programme, and closer attention from medical staff. They will often present with other non-cardiac problems, requiring assistance from various medical specialties and paramedical staff and social workers. It is wise to involve all relevant parties in the rehabilitation visits of the elderly patient, so that no conflict of advice ensues about, for example, activities and lifestyle. However, the benefits of recruiting elderly patients can be very great, enabling them to achieve a greater degree of independence and avoiding the need for long-term residential care.

Rehabilitation in those with specific cardiac conditions

Angina

Continuing angina frequently prevents a patient from participating in a formal rehabilitation programme. In some, participation can be delayed until completion of further investigation and revascularization procedures, but there remain many patients with persistent angina in whom revascularization procedures are either impossible or incompletely successful. These patients may gain benefit in terms of secondary prevention and in improving exercise tolerance by participating in a rehabilitation programme. There is good evidence for a modest antianginal effect of physical training, and risk factor reduction is of considerable importance for angina sufferers. Training in the presence of exercise-induced angina may also promote the development of new collateral vessels to the myocardium.

Heart failure

In the past, heart failure was frequently listed as an absolute contraindication to participation in cardiac rehabilitation. Whilst active myocarditis or acute heart failure with congestion remain contraindications to exercise training, research over the last decade has shown that carefully selected patients with stable chronic heart failure can achieve significant and worthwhile benefits from exercise training. This is now an important area in cardiac rehabilitation research.

Like the situation for the elderly, those with heart failure are significantly limited and in need of considerable medical care. They are also likely to benefit substantially from even modest improvements in their ability to perform exercise, as many daily tasks will stress them to close to their cardiopulmonary exercise reserve. These patients are frequently well motivated and co-operate fully with the rehabilitation programme. Several practical difficulties arise, however, including the need for closer supervision, more detailed preparticipation assessment, and a greater likelihood of complications including serious ventricular arrhythmia. Perhaps most importantly from a practical point of view, these patients may need a lifelong attachment to the programme for continuing benefit.

Research in specialist units has shown possible training benefits for patients with moderate and severe heart failure, provided their condition is stable. Improvements of between 20 and 25 per cent have been seen in exercise capacity, associated with reduced sympathetic tone, reduced breathlessness and exercise ventilation, and improved exercise haemodynamics. However, in each case the training exercise needs to be tailored specifically to the patient's reduced capacity. The level of exercise prescription may start at a very low level, such as 70 per cent of their existing maximal capacity for as little as 5 to 10 min a session. This is then gradually increased in duration and absolute intensity as the patient's maximal capacity increases.

It is recommended that patients with heart failure undergo cardiopulmonary exercise assessment in a specialist unit to establish accurately their exercise capacity prior to entry into a rehabilitation programme. Detailed evaluation is needed, such as the detection of ventricular arrhythmias either by 24-h ECG monitoring or on exercise testing. The presence of ventricular tachycardia is common in patients with moderate and severe heart failure, and may increase the risk during exercise. Whether these arrhythmias negate possible benefits of rehabilitation because of the risk of precipitating arrhythmias remains unknown.

P>No patient with heart failure should take part in an exercise programme in the presence of acute decompensation, such as with pulmonary or peripheral oedema, active myocarditis, or febrile illnesses. Although there is no lower limit on ejection fraction for the participation of those with heart failure, the patient must be

comfortable at rest and be able to exercise for 5 min at an exercise level of 2 **METS** (metabolic equivalents (of oxygen consumption)) or greater. A left ventricular ejection fraction of 20 per cent or less is still compatible with participation, and patients can still usefully participate when they are stable on a cardiac transplantation waiting list.

Further reading

Coats AJS, *et al*. (1992). Controlled trial of physical training in chronic heart failure: exercise performance, hemodynamics, ventilation and autonomic function. *Circulation* **85**, 2119–31.

Coats AJS. (1993). Exercise rehabilitation in chronic heart failure. *Journal of the American College of Cardiology* **22**(Suppl. A), 172A–177A.

Oldridge NB, *et al.* (1988). Cardiac rehabilitation after myocardial infarction. Combined experience of randomized clinical trials. *Journal of the American Medical Association* **260**, 945–50.

Todd IC, Ballantyne D (1990). Antianginal efficacy of exercise training: a comparison with beta blockade. *British Heart Journal* **64**, 14–19.

15.5.4 Cardiac transplantation and mechanical circulatory support

John H. Dark

Introduction

The basic surgical technique and the first clinical attempts at heart transplantation were described over 30 years ago, but poor results lead to a quiescent phase during the 1970s. This hiatus stimulated the development of a range of pumps intended to replace the function of the heart.

Effective immunosuppression, particularly with the introduction of ciclosporin A in 1981, enormously improved transplant results and there was a rapid expansion of activity over the next decade. This reached a plateau during the 1990s, with 3000 to 4000 heart transplants worldwide and up to 300 in the United Kingdom, although the total is now declining.

The size of the problem

The incidence of heart failure is increasing, not only because of an ageing population but also because of the longer survival of existing patients, benefiting from greatly improved medical management. In the United States, a conservative estimate is that 35 000 patients per year would benefit from cardiac replacement therapy, and enthusiasts put this figure above one million. By extrapolation, a figure of 5000 patients per year can be derived for the United Kingdom.

Transplantation brings with it unavoidable morbidity, both from immunosuppression and progressive graft failure, and long-term survival is limited. The limited number of donors means that it is only an option for a tiny proportion of those with congestive cardiac failure, and furthermore, transplantation is not always available when required, as demonstrated by a mortality rate of 15 to 40 per cent amongst those accepted on to transplant waiting lists. There is thus a huge need for an alternative to transplantation for patients with heart failure that is readily available, will provide an adequate cardiac output, and will have better long-term performance.

Mechanical circulatory support

Mechanical means of assisting or replacing the heart have evolved through short-term (days) postoperative support and medium-term (weeks to months) devices for 'bridge to transplant', to those on the verge of being accepted for long-term permanent implantation (Fig. 1).

Fig. 1 The Jarvik pump (by courtesy of Mr Steve Westaby).

Devices

Intra-aortic balloon pump

Counterpulsation with an intra-aortic balloon pump (**IABP**), usually placed via the femoral artery into the proximal descending aorta, was introduced in the late 1960s and is now used primarily for the short-term support of patients with postoperative low cardiac output or unresolved ischaemia. It is sometimes used as an adjunct to intravenous inotropes in potential transplant recipients who are deteriorating, but has no proven role except perhaps in those with ischaemic cardiomyopathy. The balloon pump has not found a role in the management of cardiogenic shock after myocardial infarction, unless there is a surgically remediable problem such as an acquired ventricular septal defect or papillary muscle rupture.

Centrifugal pumps

There are a number of paracorporeal pumps, developed as alternatives to roller pumps for cardiopulmonary bypass, that can be used for short-term circulatory support. Their use is limited to about 96 h, and there are significant problems with bleeding because of the need for partial heparinization. However, they are very inexpensive and have a role in support after cardiac surgery, or for assistance of the right ventricle.

Paracorporeal, pulsatile, ventricular assist devices

A variety of relatively simple, external, pneumatically powered pumps, connected to the heart with pipes traversing the skin, are available. The best known is the 'Thoratec' system, but others include the 'ABIOMED' pump, and in Europe, various forms of the 'Berlin' Heart. They can be used for either univentricular or biventricular support, for days or even weeks. The principal application is in 'bridging to transplant'. Morbidity and mortality is high, and complications include bleeding, thromboembolic events, and infection.

'Pusher-plate' implantable devices

This category includes the two most successful left ventricular assist devices, the 'Novacor' and the broadly similar 'HeartMate', which have the electrically powered pump lying deep to the abdominal wall. Connections are to the apex of the left ventricle and ascending aorta, with inflow and outflow valves. Power is supplied via a percutaneous cable and there is also a pipe to vent the air displaced with each beat. Transcutaneous transmission of electrical power is possible using induction

loops, but a reliable implantable expansion chamber to deal with the displaced air has yet to be perfected. Portable batteries and control systems have been developed for both devices to allow discharge of the patient from hospital. Some patients fitted with such a device have survived for several years.

The principal difference between these two pumps lies in their blood interface. In the Novacor there is a seamless polyurethane sac, which is rhythmically compressed by the pusher plate. Anticoagulation with both warfarin and antiplatelet agents is required. By contrast, the blood-contacting surfaces of the HeartMate are textured to encourage deposition of circulating cells. The resulting autologous tissue lining is resistant to thrombus formation and only aspirin is required.

Both these pumps are large and expensive—typically £50 000 for the implantable components, batteries, and controllers. They cannot be used in small adults and children, and there are risks of infection and thromboembolism. Typically, 40 per cent of patients fitted with the HeartMate will experience driveline infections, although these can be contained in the 'bridge to transplant' setting and do not preclude successful transplant. The Novacor may be more prone to thromboembolic complications: in one study, 17 per cent of 36 patients who had the device for more than a year had major emboli; in another group 24 strokes occurred in 36 patients treated for a mean of 200 days.

Implantable impeller pumps

The next stage in the evolution of mechanical support is represented by a variety of axial-flow pumps. These are electrically driven, but with much lower power requirements than their predecessors, and consist of a turbine, usually in a tube connecting the left ventricular apex with the aorta. There are no valves, and the whole of the moving parts are bathed in blood. The earliest reports are only just appearing, but these pumps, if laboratory reliability can be repeated in clinical practice, have considerable potential.

Aims of mechanical support

Apart from postoperative support, which will not be discussed further, there are three uses for mechanical pumps. The commonest is 'bridge to transplant', with the intention of stabilizing or improving a deteriorating patient until a transplant can be performed. In a small minority of patients, myocardial function improves to the extent that the device can be removed. This is termed 'bridge to recovery'. The longer term aim of all these devices, and particularly for the impeller pumps, is permanent implantation.

Bridge to transplant

Patients with heart failure who deteriorate rapidly may die before they can be transplanted, or they develop progressive end-organ failure that significantly worsens the outcome after transplantation. Mechanical pumps, principally the Thoratec device and the two implantable pusher-plate pumps, have been used in thousands of patients for periods of up to several years, before they eventually receive a transplant. Only 60 to 70 per cent of such patients will receive a transplant. Most of the remainder do not survive, either developing irreversible failure of other organs despite a good cardiac output, or succumbing to problems such as stroke, haemorrhage, or infection whilst on the pump.

The 1-year survival rate after transplantation for bridged patients is between 80 and 90 per cent, at least as good as for routine patients, and probably better than that for very sick patients who do not receive a pump but survive to transplant in a precarious state.

Many patients develop anti-HLA and other antibodies whilst on a pump, either as a result of multiple transfusions, or due to the inflammatory response that occurs secondary to prolonged exposure of the blood to foreign surfaces. This delays transplantation, sometimes for years, and makes the implantable systems and discharge home economically attractive despite the high initial cost.

Bridging does not result in more transplants, but biases the transplant population towards younger, sicker recipients, and away from older, more stable candidates. The costs are huge: sums of $300 000 to 400 000 per subsequent transplant have been quoted from centres in the United States. A cynical view is that it is a very expensive way of selecting a slightly different group of patients for transplantation, and it has not yet been adopted in the United Kingdom to any significant extent.

Bridge to recovery

The reduction in afterload when the left ventricle is completely decompressed by a mechanical pump results in shrinkage of the heart and an improved ejection fraction. In some patients this improvement is real (in other words, it is not just load-dependent), sustained, and the pump can eventually be removed. Such myocardial recovery occurs particularly after postpartum cardiomyopathy and in those supported for an acute myocarditis. However, recovery in idiopathic dilated cardiomyopathy can also result: an improvement in histological appearance is seen, with a more regular arrangement of myocytes accompanied by a reduction in their diameter to normal. Other markers of heart failure, such as levels of tumour necrosis factor-a (**TNF-a**) and anti-b-receptor antibodies, which may be involved in the aetiology of the cardiomyopathy, also improve.

Only a few patients recover to the point where transplantation is not needed, and such recovery is not always sustained. In the Berlin series, one of the largest, 7 out of 19 patients who had pumps removed had either died or had been transplanted within 12 months. The remaining 12 were alive 1 year after the device was removed. However, in the future, mechanical support combined with manoeuvres to reverse some of the causes of cardiomyopathy – control of inflammatory elements, reconstitution of b receptors – may be applicable to significant numbers of patients, avoiding the need for transplantation.

Permanent implantation

A number of patients have been sustained for years with various implantable, left-ventricular assist devices, and indeed such pumps were designed as an alternative to transplant. It is not known whether the current devices can be used in the very long term. A clinical trial of left-ventricular assist devices against medical treatment in patients deemed unsuitable for heart transplant (the REMATCH trial) reportED its results in November 2001. For these critically ill patients, there was a survival advantage in receiving an assist device but even for such patients the mortality at 2 years was 77 per cent. The implication is that the pusher plate type of pump used in this trial is not yet suitable for long-term implantation.

Cardiac transplantation

For a minority of patients with endstage cardiac failure, transplantation remains an excellent form of treatment. Crucial to a successful outcome are patient selection and donor management. Post-transplant, the avoidance of acute rejection, minimization of the side-effects of chronic immunosuppression, and steps to reduce late graft failure are important issues.

Patient selection

Almost all adults presenting for transplant will have either a dilated idiopathic or ischaemic cardiomyopathy. The latter predominates in older patients: other diagnoses in patients accepted for transplantation are shown in Table 1. Amongst children, dilated cardiomyopathy now exceeds congenital heart disease as a reason for transplantation. Young adults who underwent successful palliative procedures (such as Mustard or Senning operations for transposition of the great vessels) 20 or more years ago now need a transplant.

Regardless of the aetiology of the disease, referral for consideration for cardiac transplantation should only be made when conventional treatment, both medical and surgical, has been exhausted. Thus most patients with heart failure will have had the benefit of state-of-the-art medical therapy—angiotensin-converting enzyme inhibitors, b-blockade, diuretics including spironolactone, and often digoxin and amiodarone. Reversible ischaemia should be sought when there is coronary disease.

The outlook for patients with severe and deteriorating cardiac failure is clearly poor, indeed a proportion of these patients will either need urgent transplantation or a mechanical assist device. Consideration of the ambulant patient with controlled heart failure is more difficult: improvements in medical management have rendered obsolete many of the accepted markers of poor prognosis, and it has been suggested that transplantation only improves the 1-year survival rate of patients with the most severe heart failure.

Exercise testing with measurement of maximum oxygen uptake has been a useful objective test, with values below 14 ml/kg per min suggesting a 1-year survival rate of only 30 per cent. Its combination with other markers (Table 2) into a scoring system for likely survival has allowed the stratification of this group of patients, a system that has been validated both in North America and Europe.

Contraindications to transplantation

Careful screening of individuals with irreversible failure of other organs is essential and a list of standard exclusions is shown in Table 3. The absolute contraindications mainly relate to other life-threatening conditions that would not be reversed by the presence of a new heart. Active infection clearly has to be avoided in any patient who is about to be aggressively immunosuppressed. The relative contraindications are more difficult to judge, a good example being elevation of pulmonary vascular resistance. All patients with left-heart failure have some degree of pulmonary vasoconstriction and the chance of failure of the unprepared right ventricle of the donor heart, the risk of early death being correlated in a continuous fashion with the pulmonary vascular resistance or transpulmonary gradient. Higher values contribute the greatest risk but, of course, are more likely to be found in the sickest patients. An individual assessment has to be made in each case, considering not only the patient's need but also the best use of a limited number of donor organs.

Perioperative management

The cardiac donor

Although the heart continues to beat after brainstem death occurs, there is a progressive loss of homeostatic control. This includes hypotension with loss of vascular tone, compounded by a polyuria and hypothermia. Appropriate corrections include the administration of ADH analogues (usually intranasal **DDAVP**, 1-deamino-8-D-arginine vasopressin; desmopressin), intravenous fluids, and often peripheral vasoconstrictors as well.

The initial damage to the brain is usually intracerebral or subarachnoid haemorrhage, or head trauma. Positive serology for the human immunodeficiency virus (**HIV**) and hepatitis B, previous cardiac surgery, or a known history of ischaemic heart disease are contraindications to cardiac donation, as are prolonged periods of hypotension. Intravenous drug abuse (because of its association with HIV infection) is a contraindication, as is more than the occasional use of cocaine. Brainstem death following tricyclic antidepressant overdose or carbon monoxide poisoning is associated with specific damage to the heart and such donors would not usually be acceptable.

Older donors, into their fifties or even sixties, are acceptable if they have normal coronary arteries, but the use of such hearts is associated with a poorer long-term outcome since they are much less tolerant of longer ischaemic times. However, with falling mortality rates from trauma and an increasing demand for transplantation, donors are now more likely to be older and to die as a result of intracerebral haemorrhage, often being hypertensive with left ventricular hypertrophy and premature coronary artery disease.

Matching of recipient and donor

The donor should be blood-group compatible (but not necessarily identical) with the recipient, and size discrepancies of more than 20 to 30 per cent should be avoided. Although HLA matching, particularly for the DR antigens, may reduce rejection rates, time constraints make this impossible.

Some 5 to 10 per cent of recipients are presensitized to HLA antigens by blood transfusion, pregnancy, or a previous transplant. Moreover, preformed antibodies can cause immediate rejection, which is often fatal. Before a transplant is performed from a particular donor, recipient serum is tested against donor cells to exclude a positive or cytotoxic crossmatch in this group of patients.

The operation

After removal of the native heart, the standard implantation was, in the past, with anastomoses to the pulmonary artery, aorta, and cuffs of the left and right atrium. Whilst technically straightforward, this was somewhat clumsy and sometimes resulted in distortion of the interatrial septum. The standard approach now is to perform separate caval anastomoses, thus keeping the donor right atrium intact. Better right ventricular filling, fewer arrhythmias, and almost complete elimination of the need for permanent pacing has followed this modification.

Postoperative management is as for any other cardiac surgical procedure, with transfer from an intensive care unit to the recovery ward within a few days, and discharge from hospital after approximately 2 weeks.

Immunosuppression and rejection

A host's immune response is most vigorous early after the transplant and gradually diminishes with time. Immunosuppression is based principally on one of the calcineurin-inhibitor group of drugs, ciclosporin A or tacrolimus (FK506), in conjunction with corticosteroids and usually one other agent. There are few clinically important differences between the two calcineurin inhibitors: both are nephrotoxic and both have neurological side-effects, both require monitoring by measurement of trough drug levels. Tacrolimus does not cause the hirsuitism or gum hypertrophy seen with ciclosporin A and is therefore preferable in adolescents and females: it is, however, associated with a much higher rate of glucose intolerance.

Steroid treatment is initially with large intravenous doses of methylprednisolone, which is then converted to oral prednisolone, with dosage titrated down from 1 mg/kg to between 0.1 and 0.2 mg/kg over 2 to 4 weeks. Approximately 50 per cent of patients can be weaned off steroids entirely. However, in practice, continuation of a small dose (5–10 mg per day) allows lower doses of the other drugs with different toxicities to be tolerated.

The other component of so-called 'standard triple-drug regimens' is azathioprine in a dose of 1.5 to 2.5 mg/kg, adjusted against the white blood cell count. The principal side-effects are marrow toxicity and cholestatic jaundice, but azathioprine is effective, inexpensive, and easy to control. An alternative with broadly similar actions is mycophenolate mofetil (**MMF**). In comparative studies, patients treated with MMF have slightly lower rejection rates and there may be some reduction in the rate of progression of allograft vasculopathy (see below), but these advantages have not been translated into a demonstrable improvement in survival. Mycophenolate is much more expensive than azathioprine and its role in cardiac transplant patients has yet to be determined.

Monitoring rejection episodes

Acute rejection, a loss of the control of the immune response caused by drug therapy, leads to an infiltration of inflammatory cells and myocardial oedema. This causes diastolic rather than systolic dysfunction. Clinical signs may be sparse; a third sound, elevated filling pressures or atrial flutter, together with pyrexia and sometimes influenza-like symptoms can occur. The 'gold standard' for detecting rejection is transvenous endomyocardial biopsy, performed under local anaesthetic and radiological control. Non-invasive methods are not as sensitive or specific, although echocardiography may have a role in children. Biopsy is done weekly for the first month, then at decreasing intervals over the first year. Dense infiltrates with myocyte necrosis require augmented treatment, usually with intravenous steroids. Between 20 and 40 per cent of patients will have at least one such episode of acute rejection. Regular biopsies allow the titration of treatment against the histological picture, effectively tailoring therapy to the individual.

Infection after transplantation

The need for high levels of immunosuppression creates a substantial risk of infection, particularly during the first few months. Within the first few weeks this is typically bacterial (chest, urinary tract, intravenous catheters) and with commonplace rather than opportunistic organisms. Patients who receive augmented immunosuppression are at risk of fungal infections, of which aspergillosis is the most significant.

Cytomegalovirus (**CMV**) is the most important viral infection, occurring between 1 and 2 months' post-transplantation. This may be either a reactivation of a previous exposure (50 per cent of recipients are seropositive at the time of transplant), or acquired with the heart from a donor who was seropositive. For those at greatest risk, prophylaxis with oral ganciclovir has become routine. If an infection occurs with pyrexia, leucopenia, and organ involvement (for example, pneumonitis or gastritis), it

can be life-threatening and requires treatment with intravenous ganciclovir.

Patients with fever should be investigated aggressively to make a specific diagnosis, often with invasive means such as bronchoscopy. Except for life-threatening sepsis, empirical antibiotics should be avoided. Unusual, opportunistic organisms (for example, *Aspergillus* spp., *Pneumocystis carini*, and *Legionella* spp.) should all be considered.

For further discussion of infective complications after solid-organ transplantation, see Chapter 20.6.3 and chapters dedicated to particular pathogens.

Long-term complications

Despite excellent survival and functional rehabilitation, drug side-effects, the risks of continuous immunosuppression, and graft vasculopathy can lead to morbidity. The combination of these factors results in an attrition rate of 4 per cent per year after the first 12 months.

Side-effects of immunosuppressants

All the drugs used have relatively poor therapeutic profiles, and a degree of drug toxicity is almost invariable. For the calcineurin inhibitors (ciclosporin, tacrolimus) the most troublesome are hypertension and nephrotoxicity. Some 80 per cent of patients require an antihypertensive agent, with ACE inhibitors as the first choice, followed by nifedipine. Many other calcium-channel blockers interact with ciclosporin metabolism. Both ACE inhibitors and calcium-channel blockers may benefit graft vasculopathy.

Renal impairment is due to a combination of drug toxicity and pre-existing renal disease: by 10 years' post-transplant, between 5 and 10 per cent of patients need renal replacement therapy, but their prognosis is very poor.

Effects of chronic immunosuppression

The transplant patient remains at risk, diminishing with time, of opportunistic infections. Malignant change in the skin exposed to sunlight is very common, and appropriate precautions should be taken from the beginning.

Post-transplant lymphoproliferative disease (PTLD)

All immunosuppressants inhibit suppressor T cells, which usually exert immunological control over Epstein–Barr virus (**EBV**)-infected lymphocytes. Reduction of this control after transplantation may result in the proliferation of lymphocytes with a histological picture of B-cell 'lymphoma'. The clinical picture ranges from something akin to primary EBV infection (infectious mononucleosis) to a highly malignant, multifocal lymphoma. In most cases, reduction of immunosuppression results in restoration of control, and the 'lymphoma' is seen to shrink. Sometimes this is inadequate (or rejection of the heart occurs) and chemotherapy is required, when the outlook for the patient is much poorer.

PTLD affects about 2 per cent of transplant patient in the first year, and 1 per cent per year thereafter. It is much commoner (up to 40 per cent) if acquired as a primary infection from the donor organ. Use of polymerase chain reaction (**PCR**) technology to monitor viral load may have a role in predicting those at risk, and for allowing a pre-emptive reduction in immunosuppression.

Graft vasculopathy

There is continued immunological damage to the endothelium of the coronary arteries of the transplanted heart. (A similar process is seen in other solid-organ grafts.) Endothelial abnormalities can be detected as early as 6 weeks' post-transplant, and intravascular ultrasound (**IVUS**) can detect thickening of the subintimal layer over the first year. This process is accelerated in individuals with repeated rejection episodes and is worsened by hyperlipidaemia and hypertension. It is commoner in older donor hearts, even when free of disease at transplant. By 5 years, between 40 and 70 per cent of patients will have angiographically visible disease, and this is the commonest cause of death after the first year.

Control of risk factors, particularly hyperlipidaemia, is very important. The statin group of drugs are of proven benefit, and all adult patients with heart transplants should be given these agents (if tolerated).

Because the heart is denervated, clinical presentation is subtle. There is a spectrum of disease from exercise dyspnoea, through silent myocardial infarction, to sudden death. Detection is by surveillance angiography: non-invasive alternatives (for example, thallium scintigraphy and stress dobutamine echocardiography) have yet to prove themselves useful.

Treatment of established disease is unsatisfactory. Localized lesions may respond to angioplasty or coronary stenting, but graft vasculopathy is a diffuse process and restrictions to flow from small-vessel narrowing remain. The very occasional patient may require coronary grafting for a collection of proximal, focal lesions, but the only definitive treatment is retransplantation, which gives satisfactory results for carefully selected patients.

Results

Voluntary registries and publications from single institutions report a 1-year survival rate between 80 and 85 per cent, falling to 70 per cent at 5 years, and perhaps 50 per cent at 10 years (Fig. 2). Figures for a typical transplant programme may be a little worse. Rehabilitation is excellent, with 90 to 95 per cent of patients in **NYHA** (New York Heart Association) class I or II, although a disappointing proportion, perhaps only 20 per cent, return to work.

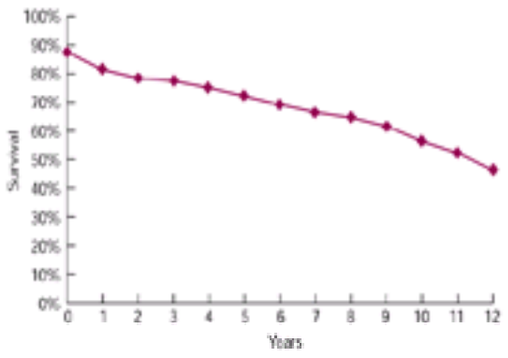

Fig. 2 Actuarial survival curve after cardiac transplantation.

Further reading

Aaronson KD, *et al*. (1997). Development and prospective validation of a clinical index to predict survival in ambulatory patients referred for cardiac transplant evaluation. *Circulation* **95**, 2660–7.

Hunt SA (1998). Current status of cardiac transplantation *Journal of the American Medical Association* **280**, 1692–8.

Hunt SA, Frazier OH (1998). Mechanical circulatory support and cardiac transplantation. *Circulation* **97**, 2079–90.

Kobshigawa JA, *et al*. (1995). Effect of pravastatin on outcomes after cardiac transplantation *New England Journal of Medicine* **333**, 621–7.

Perreas KG, *et al.* (2000). Donor management tactics for cardiothoracic transplantation. *Transplantation Reviews* **14**, 127–30.

Society of Thoracic Surgeons (2001). Fifth International Conference on Circulatory Support Devices for Severe Cardiac Failure. *Annals of Thoracic Surgery* **71**(Suppl.), S55–S222.

Westaby S, *et al*. (2000). First permanent implant of the Jarvik 2000 Heart. *Lancet* **356**, 900–3.

Young JB (2000). Perspectives on cardiac allograft vasculopathy *Current Atherosclerosis Reports* **2**, 259–71.

15.6 Cardiac arrhythmias

S. M. Cobbe and A. C. Rankin

General principles

Definition

The term cardiac arrhythmia (or dysrhythmia) is used to describe an abnormality of cardiac rhythm of any type. The spectrum of cardiac arrhythmias ranges from innocent extrasystoles to immediately life-threatening conditions such as asystole or ventricular fibrillation. Arrhythmias may occur in the absence of cardiac disease, but are more commonly associated with structural heart disease or external provocative factors.

Normal cardiac electrophysiology is discussed in Chapter 15.3.2. Abnormalities in cardiac impulse formation or propagation may give rise either to an abnormally slow heart rate (bradycardia) or fast heart rates (tachycardia).

Symptoms of cardiac arrhythmias

The symptoms produced by bradyarrhythmias depend on the extent of cardiac slowing. They may include sudden death, syncope (Stokes–Adams attacks), or dizziness (presyncope). Continuous bradycardia without asystolic pauses may produce symptoms of fatigue, lethargy, dyspnoea, or mental impairment.

The symptoms caused by tachyarrhythmias depend on a variety of factors including the heart rate, the difference between the rate during the arrhythmia and the preceding heart rate, the degree of irregularity of the rhythm, and the presence or absence of underlying cardiac disease. Symptoms of tachycardia include a feeling of rapid palpitation, angina or dyspnoea, syncope or sudden death. The differential diagnosis of palpitation and syncope is discussed in Chapter 15.2.3.

Investigation of arrhythmias

History taking must include a detailed description of the symptoms associated with the arrhythmia. Evidence should be sought for factors that may precipitate the arrhythmia (for instance, exercise, alcohol) and for the presence of underlying cardiac disease, in particular valvular heart disease, myocardial ischaemia/infarction, or congestive heart failure. Examination of the pulse will be unremarkable if the arrhythmia is intermittent. Physical examination for evidence of structural heart disease is essential. Further investigations to establish the presence of structural heart disease and to determine ventricular function may include 12-lead electrocardiography, chest radiography, echocardiography, exercise stress testing, and coronary arteriography.

Electrocardiography

The key to the successful diagnosis of cardiac arrhythmias is the systematic analysis of an electrocardiogram (**ECG**) of optimal quality obtained during the arrhythmia (Table 1). Ideally, this should comprise all 12 leads recorded on a multichannel recorder, which can allow the identification of P-waves in one lead while they may be absent or equivocal in another.

Ambulatory electrocardiography

Continuous monitoring is necessary for identification where arrhythmias are intermittent. This may involve monitoring in the cardiac care unit, particularly in the acute stages of myocardial infarction, or ambulatory (Holter) monitoring. Ambulatory electrocardiography is normally performed for periods of between 24 and 48 h using a portable recorder, which records every heartbeat on to magnetic tape or into solid-state memory. High-speed or automatic replay facilities enable the identification of intermittent arrhythmias, as well as the quantification of extrasystoles and assessment of parameters of heart rate variability.

Interpretation of recordings requires knowledge of possible artefacts, such as variations in tape speed in magnetic tape-based recorders, or movement artefact. It is important to allow for physiological variability in the sinus rate, also to appreciate that minor abnormalities such as extrasystoles or brief (3 to 4 beat) runs of supraventricular arrhythmias are usually of no significance. Ambulatory electrocardiographic recordings are of most value when they provide correlation between the patient's symptoms and the cardiac rhythm at that moment. Patients should therefore be issued with a diary card to report any symptoms suggestive of arrhythmia during the recording.

Conventional 24- to 48-h ambulatory electrocardiography is unlikely to yield useful results if the frequency of arrhythmic symptoms is less than every day or two. Alternative strategies include the use of a patient-activated recorder, which is applied and activated during symptoms, or an external or implanted 'loop' recorder. Loop recorders continually record the electrocardiographic signal, but only have sufficient buffer memory to retain a few minutes' data. In the event of symptoms, the patient activates the device, thus 'fixing' the previous few minutes' recording for analysis. Loop recorders are particularly useful in the diagnosis of infrequent brief, but disabling, symptoms, where the use of a patient-activated recorder is not feasible.

Cardiac electrophysiological study

More detailed investigation of cardiac arrhythmias is undertaken by invasive cardiac electrophysiological testing. Multipolar electrodes are inserted to record electrograms from the atrium, ventricle, His bundle, and commonly from the coronary sinus (Fig. 1). The site of conduction delays within the heart may be identified, or accessory pathways localized. Sustained arrhythmias may be initiated and terminated by extrastimuli (Fig. 2), and their pattern of activation in the heart studied in detail. Electrophysiological mapping is an essential part of radiofrequency ablation (see below).

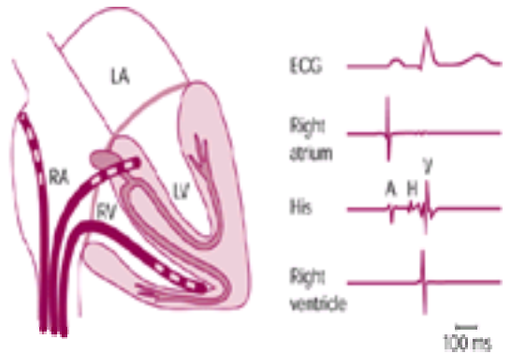

Fig. 1 Electrophysiological study. Illustration of lead placement (left). Quadripolar leads have been inserted from the femoral vein and the tips are shown positioned to allow recording and pacing from the high right atrium, the His bundle, and the right ventricular apex. Intracardiac electrograms (right) show recordings from atrium (A), His bundle (H), and right ventricle (V).

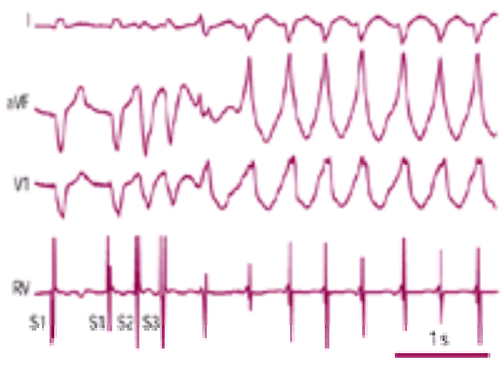

Fig. 2 Induction of ventricular tachycardia by programmed stimulation. Ventricular pacing stimuli (S1) at 100 beats per min are followed by two extra stimuli (S2, S3). Sustained monomorphic ventricular tachycardia is induced. Surface leads I, aVF, V_1, and the intracardiac electrogram from right ventricular apex (RV) are shown.

Bradycardias

Aetiology and mechanisms

Bradycardia is defined as a ventricular rate of less than 60 per min, and results from a reduction in the rate of normal sinus pacemaker activity, or from disturbances of atrioventricular (**AV**) conduction. Sinus bradycardia may be physiological—for example, during sleep, in athletes, and in young people. Pathological bradyarrhythmias can result from intrinsic degenerative disease of the sinus or atrioventricular node, or the conducting system. Changes may also be due to extraneous factors such as sympathetic withdrawal, vagal stimulation, drug effects, myocardial ischaemia/infarction, infiltration, or surgical trauma, also miscellaneous conditions such as hypothyroidism, hypothermia, jaundice, or raised intracranial pressure.

Specific disorders

Sinoatrial disease

Sinoatrial disease, often referred to as 'sick sinus syndrome', results in inappropriate sinus bradycardia, sinus pauses, or junctional rhythm (Fig. 3) in the absence of extrinsic factors. The condition is most commonly caused by idiopathic degeneration of the sinus nodal cells, particularly in the elderly, and is associated in about 20 per cent of cases with idiopathic bundle-branch fibrosis (see below). Occasionally, sinoatrial disease is caused by ischaemia due to obstruction of the right coronary artery. Conduction block may occur between the sinus node and the atrium (sinoatrial block), resulting in 'dropped' P-waves (Fig. 4). More prolonged suppression of sinus node activity results in periods of sinus arrest, which are terminated by an escape beat from the sinus node, atrioventricular junction, or ventricle (Fig. 5(a)). Where the sinus rate is permanently slower than the junctional rate, continuous AV junctional rhythm will be present. Patients with sinoatrial disease have an increased predisposition to atrial tachyarrhythmias (bradycardia/tachycardia syndrome), and prolonged pauses may follow termination of tachycardia (Fig. 5(b)).

Fig. 3 Sinus bradycardia. The heart rate is less than 40 beats/min, and the sinus rate is so slow that an escape junctional beat is seen (open circle), preceding the P-wave.

Fig. 4 Sinoatrial block. A pause occurred because of the absence of a P-wave (open arrow). The timing of the sinus beats, however, is not interrupted, indicating that the sinus node discharged but the impulse failed to excite the atria.

Fig. 5 Sinus arrest. (a) Pause of 4 s results from failure of the sinus node to discharge. (b) Termination of atrial fibrillation is followed by a sinus pause of 2.5 s due to sinus arrest in a patient with bradycardia/tachycardia syndrome.

Sinoatrial disease can cause symptomatic bradycardia, dizziness, or syncope, but may be asymptomatic. The diagnosis is normally made from 12-lead or ambulatory ECG recording. Investigation should focus on excluding extrinsic causes of bradycardia, and on demonstrating the correlation between bradycardia or pauses and symptoms.

Neurocardiogenic syncope

Conditions where patients suffer reflex-induced attacks of bradycardia or hypotension are described in Chapter 15.2.3.

Atrioventricular conduction disorders

Impairment of atrioventricular conduction may occur either within the atrioventricular node (intranodal) or within the His–Purkinje system (infranodal). Intranodal block is not associated with QRS abnormalities, while distal (infranodal) block is commonly associated with bundle-branch block.

First-degree atrioventricular block

The normal upper limit of the PR interval is 0.20 s, and if the value exceeds this then first-degree atrioventricular block is present (Fig. 6). A prolonged PR interval may be associated with bifascicular block (for example, right bundle-branch block plus left anterior hemiblock). This condition, termed trifascicular block, implies the presence of slowed conduction through the remaining fascicle.

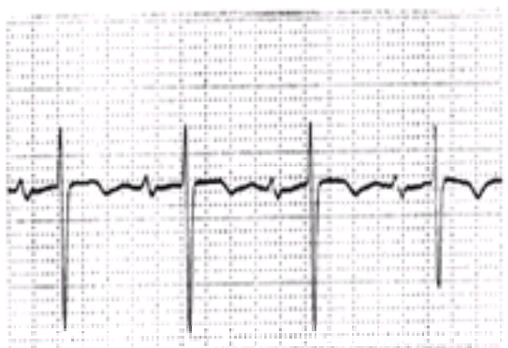

Fig. 6 First-degree heart block. The PR interval is prolonged (0.32 s).

Second-degree atrioventricular block

In second-degree atrioventricular block, there is intermittent failure of conduction from atrium to ventricle. In type I (Wenckebach) second-degree block, a characteristic pattern of increasing PR interval duration followed by a non-conducted P-wave is seen (Fig. 7). The QRS morphology is commonly normal. In type II second-degree AV block there is a sudden failure of conduction, without a preceding increase in the PR interval (Fig. 8). Regular non-conducted P-waves may result in high-degree block, with 2:1 or 3:1 conduction.

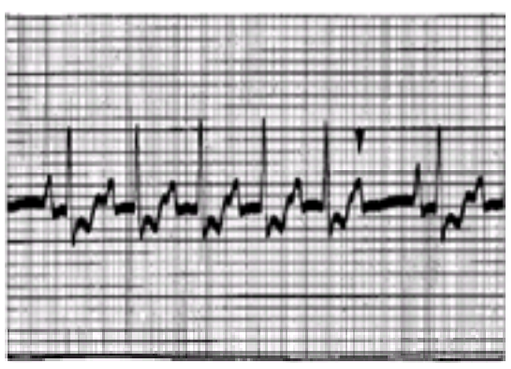

Fig. 7 Second-degree heart block, type I (Wenckebach). The PR interval progressively prolongs until there is a failure of conduction following a P-wave (arrow).

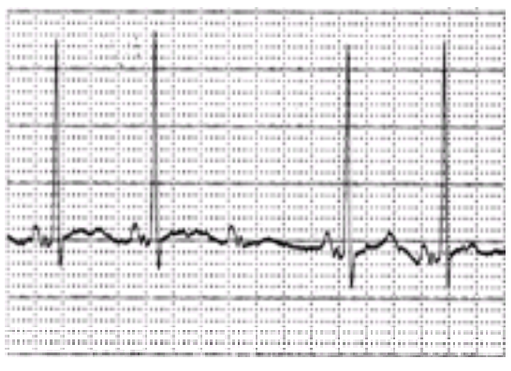

Fig. 8 Second-degree heart block type II. A non-conducted P-wave occurs without preceding prolongation of the PR interval.

Third-degree atrioventricular block

The characteristic feature of third-degree (complete) atrioventricular block is complete dissociation between atrial and ventricular activity (Fig. 9). The ventricular rate is regular and slower than the atrial rate. An escape rhythm arising above the bifurcation of the bundle of His will produce a narrow QRS morphology, commonly with a relatively rapid escape rhythm (50 to 60 per min). A more distal escape rhythm results in widened, bundle branch block morphology complexes with a slower escape rate (20 to 30 per min). When complete AV block coexists with atrial fibrillation, it is recognized by the presence of a slow, regular ventricular response.

Fig. 9 Third-degree (complete) heart block. Atrial activity does not conduct to the ventricles, and there is a regular escape rhythm of 35 beats per min.

High-degree AV block can be intermittent, and the resting ECG may be normal or only show evidence of mild conducting system disturbance such as first-degree AV block or bundle-branch block. If there is clinical suspicion, ambulatory ECG recording, for prolonged periods if necessary, is required to obtain evidence of higher degrees of AV block.

Aetiology of atrioventricular block

The causes of atrioventricular block are shown in Table 2: the commonest is idiopathic fibrosis of the His–Purkinje system, which occurs with increasing frequency from the seventh decade of life onwards, is associated with sinoatrial disease in up to 25 per cent of cases, and results in progressive impairment of atrioventricular conduction.

Atrioventricular block may occur acutely in myocardial infarction (Fig. 10). Inferior myocardial infarction predominantly affects atrioventricular nodal conduction by vagal overactivity, and possibly adenosine release from ischaemic myocardium. First-degree, type I (Wenckebach) second-degree block or third-degree atrioventricular block may occur, but these are commonly transient. Spontaneous recovery of normal conduction generally occurs within 7 to 10 days. By contrast, atrioventricular block secondary to anterior myocardial infarction is normally due to extensive infarction of the interventricular septum involving the left and right bundle branches after the division of the bundle of His. This may result in type II second-degree block or complete atrioventricular block, with a lower probability of recovery of normal conduction.

Fig. 10 Complete heart block in a patient with acute myocardial infarction. There is a narrow QRS-complex escape rhythm with ST segment elevation, ventricular rate 45 beats/min.

Any drug slowing atrioventricular conduction may potentially produce atrioventricular block. The risk is greater when such drugs are used in combination, the combination of intravenous verapamil in patients already receiving b-adrenoceptor blockers being particularly hazardous. Vagally mediated conduction disturbances occur as a physiological finding in highly trained athletes, or in neurocardiogenic syncope. Atrioventricular conduction disturbances arise in structural congenital heart disease such as endocardial cushion defects, but also as an isolated congenital abnormality, commonly in association with maternal systemic lupus erythematosus.

Management of bradycardias

The principal indications for active intervention in bradycardia are symptomatic (disturbances of consciousness, fatigue, lethargy, dyspnoea, or bradycardia-induced tachyarrhythmias) or prognostic (prevention of sudden cardiac death). Drugs interfering with sinoatrial or atrioventricular nodal function should be withdrawn if possible, although under certain circumstances (for example, tachybradycardia syndrome) it may be necessary to combine pacemaker implantation with continued drug therapy. Transient increases in sinus rate or the ventricular escape rate in complete atrioventricular block may be obtained with atropine or isoproterenol (isoprenaline). However, drug treatment is only of temporary value, and pacing is indicated for persistent bradycardia.

Management of specific disorders

Sinoatrial disease

Pacemaker implantation is indicated for the relief of symptoms (see below). Prognosis is not improved by pacemaker implantation in sinus nodal disease and thus pacemaker implantation in asymptomatic patients is not indicated.

Neurocardiogenic syncope

Patients with carotid sinus hypersensitivity and symptoms of presyncope or syncope should undergo permanent pacemaker implantation (see below). In vasovagal syndrome, the optimal treatment is uncertain: medical therapy with agents as diverse as a-agonists, b-blockers, vagolytic agents (disopyramide, hyoscine), ephedrine, or antidepressants is often tried, but the evidence base for the efficacy of drug therapy is weak, and spontaneous resolution of symptoms occurs in many patients.

Those whose symptoms persist despite drug therapy, particularly if bradycardia is a major component of the response to tilt-testing, are candidates for pacemaker implantation.

Atrioventricular conduction disturbances

First-degree AV block produces no symptoms and does not require treatment. The risk of progression from chronic bifascicular block to complete heart block is low, and patients with asymptomatic bifascicular block do not require prophylactic pacemaker implantation. The presence of trifascicular block implies advanced conducting system disease, and permanent pacing should be considered if there are symptoms or evidence of intermittent complete heart block.

Type I (Wenckebach) second-degree AV block is normally associated with a reliable subsidiary pacemaker and a low risk of progression to complete heart block. In the majority of instances active treatment is not necessary unless recurrent presyncope or syncope suggest the occurrence of an intermittent higher degree block, requiring consideration of pacemaker implantation. Type II second-degree AV block is generally indicative of extensive infranodal conduction abnormality, with a high risk of progression to complete AV block. Most authorities therefore recommend permanent pacemaker implantation even in the absence of symptoms. The presence of complete atrioventricular block, except in the context of an acutely reversible condition, should be regarded as an indication for permanent pacemaker implantation. This is urgent in cases where Stokes–Adams attacks are occurring, but even in asymptomatic patients the prognosis appears to be improved by permanent pacing. One exception to this general rule is congenital complete heart block, where the escape rhythm is often relatively fast (50 to 60 per min) with a narrow QRS morphology. Many patients remain asymptomatic well into adult life, although there is a small risk of syncope or sudden death. Pacemaker implantation should be considered if there are symptoms, or if the ventricular rate on ambulatory recording remains persistently below 50 beats/min, or in patients over 40 years of age.

Temporary pacemaker implantation is indicated where frequent Stokes–Adams attacks are occurring, or where the conduction disturbance is likely to be transient such as in cases of drug intoxication or inferior myocardial infarction. In the latter case, even the presence of complete heart block may be associated with an adequate ventricular rate and pacing need only be undertaken if there is haemodynamic compromise. Temporary pacing can only be used for a few days, owing to the risk of introducing infection along the electrode track.

Prognosis

The prognosis of patients with complete atrioventricular block having Stokes–Adams attacks is poor without pacemaker implantation, and is improved markedly by permanent pacing. Following pacing, the prognosis will depend on the nature and extent of underlying cardiac disease.

Asystole

The term asystole is used when the electrocardiogram shows a complete cessation of both atrial and ventricular activity: this appearance may be mimicked by disconnected ECG cables or other artefacts, but since asystole causes cardiac arrest the distinction is usually obvious. The management of asystole is discussed in Chapter 16.3.

Pacemaker therapy

Basic principles

The basis of pacemaker therapy is the local depolarization of the myocardium by an electric current passed through an electrode in contact with the heart (atrium or ventricle). Activation of the remainder of the atria or ventricles occurs by direct cell-to-cell conduction. The minimum current necessary to stimulate the heart during diastole is known as the pacing *threshold*. Pacemaker systems normally comprise one or more intracardiac catheter *electrodes*, introduced into the heart via the venous system, and a *pulse generator*, which contains the circuitry for generating and timing the pacing stimulus, as well as for sensing spontaneous cardiac depolarizations. The pacing stimulus is delivered between the active pole at the tip of the electrode catheter and an indifferent electrode. This is sited either on the same electrode catheter 1 to 2 cm proximal to the tip (bipolar pacing), or utilizes the can of the implanted pulse generator (unipolar pacing). An essential prerequisite for satisfactory pacing is that the electrode maintains a stable electrical contact with the myocardium. This is most likely to occur when the endocardial surface is trabeculated rather than smooth, and for this reason the standard sites for endocardial atrial and ventricular pacing are the right atrial appendage and the right ventricular apex, respectively (Fig. 11).

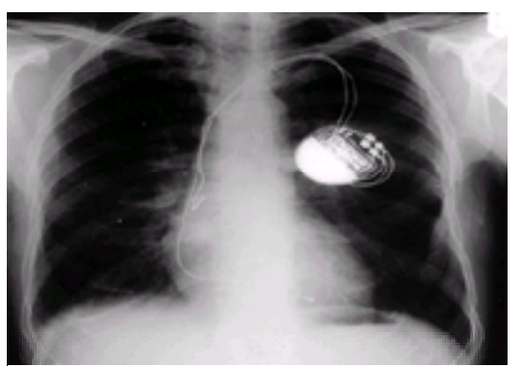

Fig. 11 Dual-chamber permanent pacemaker. Chest radiograph showing the pacemaker generator (in a subcutaneous pocket in the pectoral region) which is connected to electrodes that pass via the left subclavian vein and superior vena cava to the heart. The tips of the electrodes are in the right atrial appendage and the right ventricular apex.

The pulse generator is sited externally for temporary pacing. For permanent pacing, it is usually implanted deep to the subcutaneous fat layer in the prepectoral region (Fig. 11). The generator contains a timer set to deliver pacing stimuli at a preset *pulse interval* (for example, 1000 ms). The amplitude and duration of the pacing stimulus (*pulse width*) are usually set at nominal values (for example, 5 volts, 1 ms), but are adjustable and can be reduced to prolong the life of the battery, providing there is a sufficient safety margin between the pulse generator output and the pacing threshold. Most pulse generators are powered by lithium batteries and have a life of approximately 5 to 7 years, after which the generator is replaced. Pacemakers normally operate in the *demand* mode, whereby if spontaneous activation of the cardiac chamber is sensed via the electrode, the delivery of a pacing stimulus is *inhibited,* and the timer circuit of the generator is reset. Pacing in the *fixed rate* mode results in the delivery of stimuli regardless of the spontaneous activity of the chamber being paced.

Temporary ventricular pacing

Facilities for radiographic screening, continuous electrocardiographic monitoring, and defibrillation are necessary for the performance of temporary ventricular pacing. The temporary pacing electrode is introduced under aseptic conditions via an intravascular sheath inserted into the subclavian, internal jugular, or femoral vein and passed under radiographic control to the right atrium. The electrode tip is advanced across the tricuspid valve and impacted as distally as possible at the right ventricular apex. Non-sustained ventricular tachycardia, or occasionally ventricular fibrillation, may occur during catheter manipulation. Once the electrode is in an acceptable site, pacing is initiated, and the minimum output necessary to achieve stable ventricular capture is determined. The pacing threshold should normally be less than 1 volt, at a pulse width of between 0.5 and 2 ms. If the pacing threshold is unsatisfactory, the electrode is repositioned until an acceptable site is found. Care should be taken to determine that the electrode is stable by asking the patient to take deep breaths or to cough while pacing at threshold. The electrode is then secured at the site of insertion and the pulse generator set to an output of at least 3 volts above the pacing threshold.

Permanent pacemaker implantation

The technique of permanent pacemaker implantation is essentially similar to that of temporary pacing, except that the electrodes are much more flexible to minimize the risk of late myocardial perforation, and they are stiffened by a stylet during manipulation. Insertion is normally via the left subclavian or cephalic vein. Once the

electrode is in a satisfactory position, it is secured and connected to the implanted pulse generator. The rate, output voltage, pulse width, and other pacemaker functions can be modified non-invasively by means of telemetry via a transmitter/receiver placed on the skin over the pulse generator.

Pacing mode selection

The nomenclature used to describe pacing mode is given in Table 3, and electrocardiographic examples of the principal pacing modes are given in Fig. 12. Atrial demand (AAI) pacing is used for sinoatrial disease in the absence of atrioventricular block. Ventricular pacing (VVI) is the simplest and technically easiest mode of pacing, and is required for atrioventricular conduction disturbances. However, VVI pacing does not permit atrioventricular synchrony or an increase in pacing rate in response to an increase in sinus (atrial) rate. Dual-chamber (DDD) pacemakers have electrodes in both the right atrium and ventricle. If the sinus cycle length is greater than the pulse interval, atrial demand pacing occurs. Following the atrial stimulus, a programmable atrioventricular delay commences. If no spontaneous ventricular depolarization is sensed before the end of this interval, a pacing stimulus is delivered via the ventricular electrode. If the sinus cycle length is shorter than the pulse interval, no atrial stimulus is given, but the atrioventricular delay is *triggered*, followed by spontaneous or paced ventricular activation. By this means, the ventricular rate tracks the atrial rate up to a programmable maximum, allowing the heart to increase its rate in a physiological manner in response to metabolic demand. An alternative, and simpler, approach to achieve a rate response is the use of an activity sensor such as an accelerometer in the pulse generator. Such devices detect bodily movement and increase the pacing rate according to a programmable algorithm. Rate response can be utilized in either single- or dual-chamber pacemakers, and is designated by the suffix 'R' (e.g. AAIR, VVIR, DDDR).

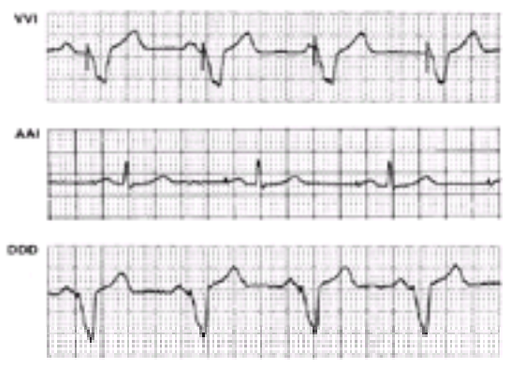

Fig. 12 Permanent pacemaker modes. Ventricular demand pacing, VVI (upper) with broad-complex ventricular complexes following the stimulus. Dissociated atrial activity can be seen. Atrial demand pacing, AAI (middle) with low amplitude bipolar pacing spike preceding the P-waves. Dual-chamber pacemaker, DDD (lower) with paced ventricular complexes following each P-wave (atrial tracking).

The advantage of dual-chamber (DDD) pacing over VVI pacing lies in the maintenance of atrioventricular synchrony and rate responsiveness, but this is achieved at the expense of increased complexity, complications, and cost. Observational studies have suggested that dual-chamber pacing reduces the risk of atrial fibrillation by virtue of pacing the atrium and avoiding retrograde atrial activation via the atrioventricular node. Additional suggested benefits include a lower incidence of the pacemaker syndrome (see below) and reduction in the risks of stroke, heart failure, and death. However, two recent large-scale randomized trials comparing DDD with VVI(R) pacing have failed to substantiate important prognostic or symptomatic benefits from DDD pacing, at least during follow-up periods of up to 3 years.

Complications

Complications of temporary or permanent pacemaker implantation include those of central venous cannulation (for instance, pneumothorax), perforation of the heart by the electrode tip leading to pericardial effusion and cardiac tamponade, and macroscopic or microscopic displacement of the electrode resulting in an increase in the pacing threshold or failure to capture. A chest radiograph should always be taken after pacemaker insertion to exclude pneumothorax and to confirm that the electrode position is satisfactory.

Permanent pacing may be complicated by the development of infection around the pulse generator, or by mechanical erosion of the generator through the skin. Once infection is established, or the skin is breached, it is almost never possible to eradicate infection with antibiotics: removal and replacement of the pacing system is required. Following electrode implantation, oedema and inflammation around the electrode tip result in a steady rise in the pacing threshold over the first few weeks. This may occasionally result in an increase of the pacing threshold such that capture is lost (Fig. 13(a)), although the process is normally mild and self-limiting. Demand pacemakers require an adequate intracardiac signal to recognize activation of the chamber in question, in order to inhibit output. If the intracardiac signal is of insufficient amplitude the pacing stimulus will not be suppressed (*undersensing*), resulting in inappropriate pacemaker firing (Fig. 13(b)). This phenomenon is commoner in atrial pacing, owing to the lower amplitude of atrial compared with ventricular electrograms. Alternatively, detection of extraneous electrical activity (for example, skeletal muscle activity) via the pacing electrode can result in inappropriate inhibition of the pacemaker output (*oversensing*) (Fig. 13(c)). Oversensing is commoner with unipolar than bipolar pacing modes because of the inclusion of the pulse generator can in the electrical circuit, and its proximity to the pectoral muscles. For the same reason, unipolar pacemaker systems are more prone to the problem of local skeletal muscle stimulation. Damage to the conductor or insulation of the pacing electrode may occur due to trauma at the site of ligation or to compression between the clavicle and first rib. This may result in oversensing, skeletal muscle stimulation, or to short-circuiting leading to premature battery depletion.

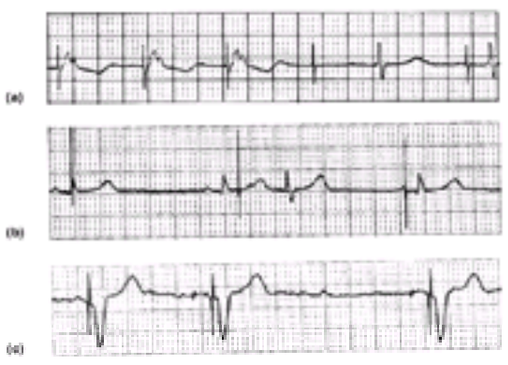

Fig. 13 Pacemaker malfunction. (a) Failure to capture. The fourth stimulus fails to capture the ventricle. (b) Undersensing. The atrial pacemaker has failed to sense the preceding atrial activity and therefore delivered the second stimulus. This has captured the atrium, with the P-wave in the ST segment, and subsequent conduction to the ventricle. (c) Oversensing. This dual-chamber pacemaker has sensed an electrical artefact through the ventricular lead and as a result has suppressed ventricular pacing, with the absence of ventricular activation following the third P-wave.

Patients receiving atrial demand (AAI) pacemakers may subsequently develop atrioventricular block, resulting in a recurrence of syncope and requiring upgrade of the pacing system to a dual-chamber (DDD) unit. A proportion of patients with ventricular demand (VVI) pacemakers, particularly those with sinoatrial rather than atrioventricular disease, will manifest retrograde ventriculoatrial conduction during ventricular pacing. This sometimes causes symptoms of fatigue, dizziness, or hypotension ('pacemaker syndrome'), which are associated with the presence of atrial cannon waves occurring as a result of simultaneous atrial and ventricular contraction. Upgrade of the system to a dual-chamber unit is necessary if symptoms are troublesome.

Follow-up

Patients with permanent pacemakers require follow-up by a pacemaker clinic. As well as detection of the complications described above, the function of such a clinic

is to assess the status of the pulse generator battery, and to maximize its life by programming the pulse generator output to the minimum consistent with a satisfactory safety margin. The design of pulse generators and the battery characteristics normally allow prediction of the expected replacement date several months if not years ahead. However, premature battery depletion or pacemaker failure does occur, and patients should therefore be assessed at least annually by the clinic. It must be borne in mind that many patients who have long-standing heart block treated by permanent pacing have no underlying cardiac rhythm, and that failure of the pacing system for whatever reason may be fatal.

Tachycardias

Mechanisms of arrhythmogenesis

The exact electrophysiological mechanism responsible for tachyarrhythmias is not known in all cases. There is a complex interaction between an underlying substrate such as previous myocardial infarction, a triggering event such as an extrasystole, and modulating influences, of which sympathetic stimulation and myocardial ischaemia are the most important. The principal mechanisms responsible for tachyarrhythmias are those of abnormal automaticity, triggered activity, and re-entry (Fig. 14).

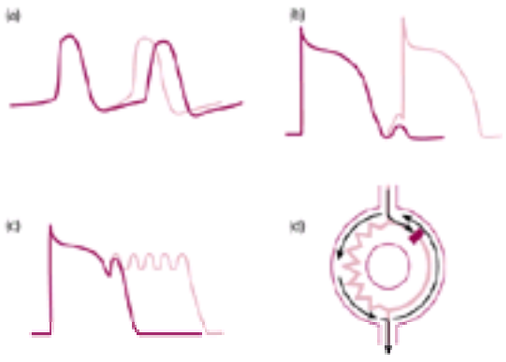

Fig. 14 Mechanisms of arrhythmia. (a) Increased automaticity. (b) Triggered activity due to early after-depolarizations. (c) Triggered activity due to delayed after-depolarizations. (d) Re-entry circuit. See text for details.

Automaticity

Abnormal automaticity is defined as an inappropriate increase in the rate of discharge of a tissue that has physiological pacemaker properties (namely, sinus node, atrioventricular node, or Purkinje fibres) (Fig. 14(a)). Such abnormalities are most commonly seen in the presence of ischaemia, sympathetic stimulation, or drug toxicity, especially digoxin. Automatic tachycardias are characterized by an absence of initiation by extrasystoles, either spontaneously or during electrophysiological testing.

Triggered activity

The term 'triggered activity' is used to define the appearance of automaticity immediately associated with a preceding action potential, and can be induced *in vitro* in tissues that do not demonstrate physiological automaticity. Two characteristic forms of depolarization may cause triggered activity:

1. *Early after-depolarizations* occur during the plateau phase of the action potential, prior to repolarization (Fig. 14(b)), and are more evident at slow heart rates, particularly in the presence of hypokalaemia and hypomagnesaemia. Mutations in cardiac Na^+ or K^+ channels, or drugs that prolong myocardial repolarization by inhibiting one or more components of the outward potassium current I_k (for example, class IA and class III antiarrhythmics, tricyclic antidepressants, antihistamines, organophosphorous insecticides, and many others) predispose to the appearance of early after-depolarizations *in vitro*. These changes are associated with the congenital and acquired long QT syndromes and the arrhythmia *torsade de pointes* (see below).
2. *Delayed after-depolarizations* are small subthreshold depolarizations occurring after full repolarization of the action potential (Fig. 14(c)). Their amplitude is increased by tachycardia or intracellular calcium overload, and may reach a level at which a spontaneous action potential is generated, potentially initiating a sustained tachycardia. Delayed after-depolarizations can be induced experimentally by digitalis overload, and are the likely mechanism of digitoxic arrhythmias.

Re-entry

The majority of the clinically important sustained tachycardias, whether of atrial, junctional, or ventricular origin, appear to arise on the basis of re-entry. The establishment of a re-entry tachycardia requires the presence of a potential circuit comprising two limbs with different refractoriness and conduction properties (Fig. 14(d)). A premature beat can be conducted in one limb of the circuit, but the other limb may still be refractory, resulting in unidirectional conduction block. If conduction is sufficiently slow, the tissue distal to the site of block in the refractory limb will have regained excitability before the arrival of the depolarizing wavefront, and can conduct the activity retrogradely. This results in reactivation of the initial conducting pathway and thus a circus movement tachycardia is established. Macro re-entry is defined as the occurrence of a re-entry circuit over a large area of the heart, such as in the presence of an accessory pathway (Fig. 15(a)). Micro re-entry occurs in a relatively small area of the heart, for example at the border zone of an old myocardial infarction, where effective conduction velocity is markedly slowed (Fig. 15(b)). The characteristic feature of a re-entrant tachycardia is that an appropriately timed extrastimulus can induce unidirectional block and initiate the arrhythmia. The tachycardia may be terminated by extrastimuli that depolarize the tissue ahead of the circulating wave front and thus interrupt the circus movement.

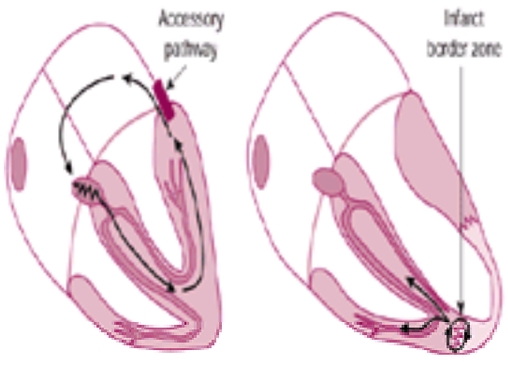

Fig. 15 Clinical examples of re-entry tachycardias. (a) Macro re-entry circuit involving an accessory pathway, which results in atrioventricular re-entry tachycardia. (b) Micro re-entry circuit at the border zone of a myocardial infarction.

Management of tachyarrhythmias

Differential diagnosis of tachycardias

The first and most important step in the diagnosis and management of tachycardias is to determine whether the arrhythmia arises within the atria and/or atrioventricular junction, or from the ventricles. An essential element in the differential diagnosis is to distinguish between tachycardias with normal QRS-complex

morphology and duration ('narrow-complex tachycardias'), and those where the QRS complexes are abnormal in morphology and increased in duration ('wide-complex tachycardias').

Narrow-complex tachycardias

Narrow-complex tachycardias arise through mechanisms that result in ventricular activation via the atrioventricular node and His–Purkinje system and therefore show normal QRS morphology and duration (£0.12 s) during tachycardia. The principal narrow-complex tachycardias and their characteristic ECG features are listed in Table 4. Careful study of all leads of the electrocardiogram is necessary to identify the presence of retrograde P-waves. Atrial flutter waves are most commonly evident in the inferior limb leads or in lead V1. Transient nterruption of atrioventricular nodal conduction by vagal stimulation or intravenous adenosine is of particular value in revealing the tachycardia mechanism. Atrial tachyarrhythmias will not normally be terminated by adenosine, but an increase in AV block reveals the underlying atrial tachyarrhythmia. By contrast, tachycardias utilizing the atrioventricular junction as part of the re-entry circuit will be terminated by transient AV block.

Wide-complex tachycardias

Few areas in cardiology cause more difficulty, or result in more mismanagement, than the diagnosis of wide-complex tachycardias. Whilst it is safe to assume that virtually all narrow-complex tachycardias have a supraventricular origin, wide-complex tachycardias (QRS duration ³0.12 s) may arise either from the ventricle or from supraventricular mechanisms, the latter occurring if there is pre-existing bundle-branch block in sinus rhythm or if functional bundle-branch block (aberration) occurs as a result of the tachycardia. An additional cause is activation of the ventricles via an accessory pathway. The electrocardiographic features of wide-complex tachycardias are described in Table 5.

Difficulties in diagnosis and management most commonly arise when ventricular tachycardia is not recognized and is misdiagnosed as 'SVT with aberration'. This usually happens as a result of a number of failings and misconceptions, the commonest being that the clinical context is not considered:

1. *The age of the patient*: middle-aged or elderly individuals presenting with a recent history of wide-complex tachycardia, and who give a history of myocardial infarction or congestive heart failure, are more likely to have ventricular than supraventricular tachycardia. Ventricular tachycardia can also arise in young patients.
2. *The haemodynamic status of the patient*: it is often assumed that ventricular tachycardia should cause haemodynamic collapse, whereas patients may in fact be haemodynamically stable if the rate is not excessively fast or if underlying cardiac function is good. Conversely, supraventricular tachycardias may cause syncope, hypotension, or shock if sufficiently rapid.
3. *The nature of the episodes of palpitation*: it is often not appreciated that ventricular tachycardia can present with a typical history of paroxysmal self-terminating episodes, just as in the case of supraventricular tachycardia.

In addition to attention to the history and 12-lead ECG, which must be analysed carefully during tachycardia, the response to transient AV nodal blockade with adenosine will assist considerably in diagnosis in many patients (Table 6). The importance of making a correct diagnosis in wide-complex tachycardia is twofold. First, inappropriate acute therapy of the tachyarrhythmia can be avoided. In particular, the use of verapamil in ventricular tachycardia misdiagnosed as supraventricular tachycardia is associated with a high risk of haemodynamic collapse as a result of the negative inotropic effect of verapamil, coupled with its lack of efficacy in terminating ventricular tachycardia. Adenosine is a safer diagnostic aid. Second, the correct diagnosis has prognostic implications. Although any wide-complex tachycardia can be terminated effectively by cardioversion, if the original arrhythmia has been misdiagnosed then the adverse prognostic significance of ventricular tachycardia will be overlooked. Appropriate investigation and long-term management may not be instituted.

Objectives of therapy

Many cardiac arrhythmias are benign and require no intervention. The main indications for treatment are the relief of symptoms, prevention of complications such as myocardial ischaemia, cardiac failure, or embolism, or an attempt to prevent arrhythmic sudden death. Precipitating factors such as myocardial ischaemia/infarction, infection, thyrotoxicosis, alcohol, electrolyte disorders, or drug toxicity must be sought and treated if possible. The type of therapy used will commonly be influenced by the presence of underlying structural heart disease such as myocardial ischaemia/infarction or left ventricular impairment.

Antiarrhythmic drug therapy

All antiarrhythmic drugs have potentially serious side-effects. They may worsen existing arrhythmias or produce new, possibly life-threatening ones (proarrhythmia), and the possibility of a proarrhythmic response should be borne in mind as part of the risk–benefit assessment whenever such drugs are prescribed. No classification exists that provides an accurate predication of the efficacy of a given drug for a given arrhythmia, thus therapy is initiated partly on the basis of trial and error, supported if necessary by more detailed investigation such as ambulatory ECG monitoring or cardiac electrophysiological testing.

The Vaughan Williams classification is based on the effects of anti-arrhythmic drugs in isolated tissue. The effects of the major classes of antiarrhythmic drug activity at the tissue level, and the associated electrocardiographic changes, are listed in Table 7. Individual drugs are described in Table 8.

Class I activity

Class I antiarrhythmic drugs act by inhibiting the rapid inward sodium current and have local anaesthetic activity. Class Ia agents (for example, quinidine, procainamide, and disopyramide) increase the action potential duration and have intermediate effects on the onset and recovery kinetics of the sodium channel and hence on intracardiac conduction. Class Ib agents (for example, lidocaine and mexiletine) shorten the cardiac action potential duration and have very rapid offset kinetics that result in minimal slowing of normal intracardiac conduction. Class Ic drugs (for example, flecainide and propafenone) have no major effect on action potential duration, but produce the most long-lasting effect on cardiac sodium channel kinetics and the most marked slowing of intracardiac conduction.

Class II activity

Class II activity is defined as antagonism of the arrhythmogenic effects of catecholamines. The commonest agents in this class are the competitive b-adrenoceptor blockers. Other agents such as propafenone have a weak b-receptor blocking activity, while amiodarone (see below) exhibits a non-competitive sympatholytic effect.

Class III activity

The class III mode of antiarrhythmic activity comprises lengthening of the cardiac action potential duration and hence of the effective refractory period. Drugs in this class possess a broad spectrum of activity against atrial, supraventricular, and ventricular arrhythmias. Currently available class III agents act by inhibiting the rapid component of the outward potassium current I_{kr}. Dofetilide and ibutilide are examples of drugs with 'pure' class III antiarrhythmic actions. Sotalol is a non-selective b-adrenoceptor antagonist that also possesses class III activity. Amiodarone possesses antiarrhythmic activity in all four Vaughan Williams classes.

Class IV activity

Class IV drugs (for example, verapamil and diltiazem) reduce the inward calcium current I_{Ca} in sinoatrial and atrioventricular nodal tissues. They are used to prevent or interrupt re-entry arrhythmias involving the atrioventricular node (for example, atrioventricular nodal re-entry tachycardia), or to slow the ventricular response in atrial fibrillation or flutter. The dihydropyridine calcium antagonists such as nifedipine have no antiarrhythmic action.

Digoxin

The antiarrhythmic activity of digoxin is not explained within the Vaughan Williams classification. Although its inotropic actions are based on inhibition of cardiac Na^+K^+ATPase, the antiarrhythmic activity appears to be mediated predominantly through vagal stimulation. Digoxin is used to slow ventricular rate in atrial fibrillation.

Adenosine

Adenosine, a naturally occurring purine nucleoside, is used pharmacologically to produce transient slowing or block of the sinus node or atrioventricular node, and is effective for the termination of re-entry arrhythmias involving the atrioventricular node. Adenosine is of particular value in view of its extremely short plasma half-life (about 2 s), which confers safety. It must be administered by rapid intravenous bolus injection, using incremental doses from 3 to 12 mg, to achieve the desired therapeutic effect.

Non-pharmacological therapy

Physical manoeuvres

Tachycardias involving the atrioventricular node may be terminated by manoeuvres that produce transient vagal stimulation, and patients with recurrent supraventricular tachycardias should be taught to perform these techniques in order to abort attacks and avoid the need for hospital treatment. The Valsalva manoeuvre, performed in the supine position, is the most effective technique.

Antitachycardia pacing

Re-entry tachycardias may be terminated by the delivery of appropriately timed extrastimuli that depolarize part of the re-entry circuit prior to the arrival of the wave front and interrupt the arrhythmia. Simple overdrive pacing can be effective in the termination of atrial flutter, AV nodal re-entry, AV (orthodromic) re-entry tachycardia, or sustained ventricular tachycardia (Fig. 16). The cardiac chamber in question is paced for brief periods at a rate just above that of the tachycardia, for example 6 to 12 beats, with repeated attempts sometimes necessary at gradually increasing rates. Overdrive atrial or ventricular pacing may result in degeneration into atrial and ventricular fibrillation, respectively, hence facilities for defibrillation must be available. Implantable antitachycardia pacing facilities are incorporated into implantable cardioverter-defibrillators (see below).

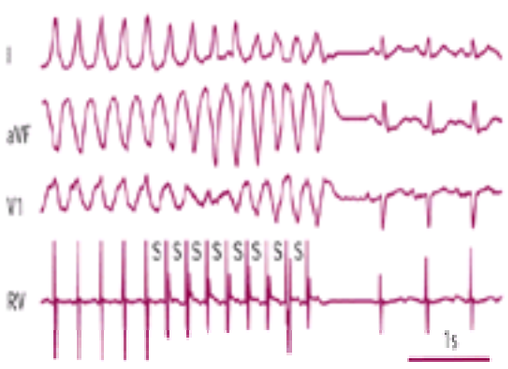

Fig. 16 Termination of ventricular tachycardia by overdrive ventricular pacing. During ventricular tachycardia a burst of eight stimuli (S) results in termination of the tachycardia and resumption of normal sinus rhythm. Surface leads I, aVF, V_1, and intracardiac electrograms from the right ventricular apex (RV) are shown.

External cardioversion/defibrillation

R-wave synchronized, direct current cardioversion under general anaesthesia or deep sedation is the most effective and immediate means of terminating sustained tachycardias, and is commonly used in the termination of atrial flutter, atrial fibrillation, or sustained ventricular tachycardia (Fig. 17). Although atrial flutter may respond to low-energy cardioversion (50 to 100 joules), the other arrhythmias normally require energies of 100 to 360 joules for termination. The use of non-synchronized DC shock in the termination of ventricular fibrillation is discussed in Chapter 16.3.

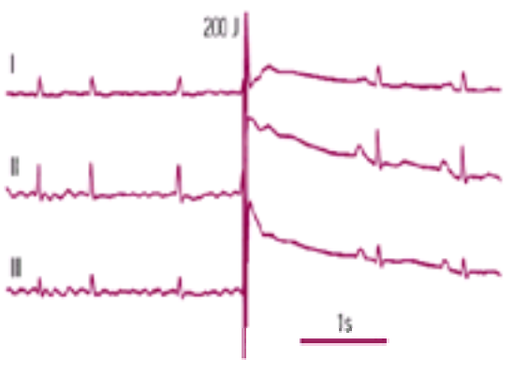

Fig. 17 Synchronized DC cardioversion of atrial fibrillation. A direct current shock, 200 joules, is delivered during atrial fibrillation to coincide with the R-wave of the QRS complex. This shock terminates the arrhythmia with restoration of normal sinus rhythm.

Internal cardioversion

Failure of external cardioversion of atrial fibrillation occurs in a significant proportion of patients as a result of various factors, including increased transthoracic impedance due to obesity, prolonged atrial fibrillation, left ventricular dysfunction, and left atrial dilatation. Internal cardioversion can still be achieved in a proportion of these patients. The procedure involves the introduction of specialized electrode catheters that permit DC-shock delivery between electrodes in the right atrium and the pulmonary artery or coronary sinus, providing a current field that achieves depolarization of both atria.

Implantable cardiovertor-defibrillators

Patients identified as being at high risk of sudden cardiac death, owing to a history of spontaneous or inducible sustained ventricular arrhythmias or out-of-hospital cardiac arrest, may be treated with an implantable cardioverter-defibrillator (**ICD**). However, these devices are expensive, complex, and require regular specialist follow-up. A transvenous rate-sensing/-shocking electrode is introduced via the subclavian vein to the right ventricular apex, with the generator implanted in the pectoral region (Fig. 18). The shock is delivered between the intracardiac shocking electrode and the generator can. Some devices also include a right atrial electrode to sense atrial activation. This improves the distinction between sinus or atrial tachyarrhythmias and ventricular tachycardia, and reduces the risk of an inappropriate shock being delivered. ICDs can be programmed to deliver initial antitachycardia ventricular pacing for slower tachycardias, with shock delivery available for faster rates or if pace-termination fails.

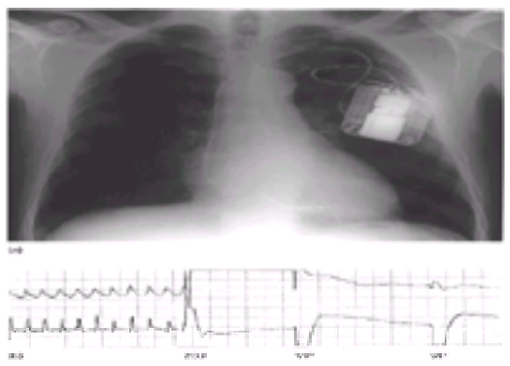

Fig. 18 Implantable cardioverter-defibrillator (ICD). (a) Chest radiograph showing the ICD generator in the left pectoral region, connected to a lead which passes via the left subclavian vein and superior vena cava to the heart. The tip of the lead is in the right ventricular apex. Cardiac rhythm is sensed from the electrodes at the tip of the lead, and shocks can be delivered between the metal casing of the generator and the right ventricular coil (thickened portion of lead). (b) Discharge from an ICD. A rapid polymorphic ventricular tachycardia is terminated by a 20-joule shock from the device. Electrograms shown are retrieved from the memory of the device, upper tracings from the shocking circuit (generator can to ventricular coil) and lower tracings from the sensing circuit (bipolar electrodes at the tip of the catheter in the right ventricle. The shock is followed by ventricular pacing (**VP**).

Implantable atrial defibrillators have been developed for use in those patients with frequent paroxysmal atrial fibrillation, but the value of these has not been fully evaluated.

Radiofrequency ablation

Selective ablation of part of a re-entry circuit, an arrhythmic focus, or of the atrioventricular node is used increasingly in the management of troublesome arrhythmias, and offers the opportunity of curative treatment. Radiofrequency energy is delivered between the tip of an intracardiac electrode positioned at the appropriate site and an indifferent surface electrode placed over the scapula. The energy produces a localized necrotic lesion 2 to 3 mm in diameter, which results in local conduction block. The success of the procedure depends on the accuracy of the placement of the lesion in relation to the re-entry circuit or arrhythmic focus. Current indications for radiofrequency ablation are listed in Table 9, and specific issues are discussed below in relation to individual arrhythmias.

Arrhythmia surgery

The 'Maze' procedure for atrial fibrillation involves creating a series of linear incisions in the left and right atria, which are then sutured, creating lines of conduction block. This prevents the development of atrial re-entry circuits while permitting atrioventricular conduction. Surgical management of recurrent ventricular tachycardia by mapping and resection of the re-entry circuit is occasionally performed, but is being superceded by ablation or ICD therapy.

Individual arrhythmias

Extrasystoles

The term extrasystole is used to describe a premature beat arising from a focus other than the sinus node. Extrasystoles are also described as premature beats, premature contractions, premature depolarizations, or ectopic beats.

Atrial extrasystoles

Atrial extrasystoles are recognized by a premature P-wave of different morphology from the sinus P-wave (Fig. 19(a)), which can be hidden within the ST segment or T wave of the preceding sinus beat. Premature atrial extrasystoles that occur before full recovery of the atrioventricular node will be followed by prolongation of the PR interval, or, if sufficiently premature, complete failure of conduction (Fig. 19(b)). Non-conducted atrial extrasystoles must be distinguished from sinus arrest or second-degree atrioventricular block.

Fig. 19 Atrial extrasystoles. (a) An atrial extrasystole, with an abnormal P-wave at the end of the preceding T wave, occurs following a sinus beat. (b) Blocked atrial extrasystoles. In the same patient, atrial extrasystoles occur following each sinus beat. They are earlier than those in (a), and the AV node is refractory because of the proximity of the atrial extrasystoles to the preceding beat, and conduction is blocked.

Atrial extrasystoles are a common finding in healthy people, particularly with increasing age, but are more frequent in the presence of increased atrial pressure or stretch such as in cardiac failure or chronic mitral valve disease. Patients should be reassured that the arrhythmia is benign and that drug treatment is rarely necessary. If treatment is required on symptomatic grounds, b-adrenergic blockers may be used, but class I antiarrhythmic drugs should be avoided in view of their proarrhythmic risk.

Junctional extrasystoles

Junctional extrasystoles are identified by the appearance of a premature, normal QRS complex in the absence of a preceding atrial extrasystole. The atria as well as the ventricles may be activated, resulting in an inverted P-wave simultaneous with the QRS complex, or inscribed within the ST segment. The significance and management of junctional extrasystoles are similar to those of atrial extrasystoles.

Ventricular extrasystoles

Ventricular extrasystoles are identified by the appearance of a bizarre, wide QRS complex not preceded by a P-wave (Fig. 20). There is commonly ST segment depression and T wave inversion. Ventricular extrasystoles may be intermittent, or occur with a fixed association to the preceding normal beats, that is 1:2, 1:3 (bigeminy or trigeminy). Where extrasystoles are of differing morphologies, the terms 'multifocal' or 'multiform' are used.

Fig. 20 Ventricular extrasystole (open circle). No retrograde atrial activation occurs, and the P-wave sequence is undisturbed (arrowed).

Ventricular extrasystoles occur in otherwise normal hearts, but are found particularly in the presence of structural heart disease. They occur commonly in the acute phase of myocardial infarction, but are also seen in the postinfarction phase, and in the presence of severe left ventricular hypertrophy or dysfunction of whatever cause. Extrasystoles may produce symptoms that require treatment in a minority of cases. The safest option is b-blockade.

Atrial arrhythmias

Atrial fibrillation

Atrial fibrillation is the commonest sustained tachycardia. The underlying mechanism is thought to be re-entry in most instances, with multiple wavelets (probably a minimum of six) circulating through the atria. Studies of patients with frequent paroxysmal 'lone' atrial fibrillation suggest that the arrhythmia may be triggered by one or more rapidly discharging foci, which are commonly situated in the pulmonary veins. Such patients often have frequent premature 'P-on-T' atrial extrasystoles on ambulatory ECG monitoring. Recent experimental and clinical studies have helped to explain the long-standing clinical observation that the longer the duration of atrial fibrillation, the more difficult it is to restore and maintain sinus rhythm ('atrial fibrillation begets atrial fibrillation'). Rapid atrial activation induces a process of electrical remodelling, resulting in shortening of the atrial refractory period and loss of the normal lengthening of the atrial refractory period at slower heart rates. The initial mechanism is thought to be intracellular Ca^{2+} overload, although more prolonged atrial tachyarrhythmias result in downregulation of Ca^{2+} entry and de-differentiation of atrial myocytes towards a fetal phenotype. Preliminary clinical data suggest that atrial electrical remodelling is reversible following cardioversion. The possibility exists that short-term treatment after cardioversion, during the period of regression of atrial electrical remodelling, could have long-term benefits in the prevention of relapse.

The characteristic ECG findings in atrial fibrillation of recent onset are of rapid, irregular 'f' waves at a rate of 350 to 600/min. These are associated with an irregular ventricular response because of variable conduction through the AV node (Fig. 21). With increasing duration of chronic atrial fibrillation, the amplitude of the 'f' waves diminishes until they are no longer visible. Under these circumstances, atrial fibrillation is diagnosed by the absence of P-waves and the irregular ventricular response (Fig. 21(b)).

Fig. 21 Atrial fibrillation. (a) Coarse atrial fibrillation of recent onset. (b) Fine atrial fibrillation in a patient with long-standing valvular disease. Surface V_1 leads are shown.

Clinical features of atrial fibrillation

The prevalence of atrial fibrillation increases with advancing age and may be as high as 5 per cent in the elderly. There are numerous causes of the arrhythmia (Table 10), but in many instances no obvious aetiological factor can be identified, and the individual is described as having 'lone' atrial fibrillation. Atrial fibrillation carries adverse prognostic significance, due in part to its association with organic heart disease. In addition, atrial fibrillation is an important risk factor for the development of stroke and systemic embolism as a result of stasis and thrombus formation in the left atrium. The risk of stroke is particularly high in patients with mitral stenosis or mitral valve replacement and chronic atrial fibrillation. The thromboembolic risk in non-rheumatic atrial fibrillation is related to age, previous left ventricular dysfunction, hypertension, and diabetes mellitus (Table 11).

Atrial fibrillation results in loss of the atrial contribution to left ventricular filling, which results in a modest reduction in cardiac output. In the presence of impaired ventricular function this can result in a worsening of heart failure. More commonly, symptoms and impairment of left ventricular function ('tachycardiomyopathy') arise as a result of a rapid uncontrolled ventricular rate. In addition, uncontrolled atrial fibrillation can cause further impairment of ventricular filling in mitral stenosis and conditions associated with left ventricular diastolic dysfunction, or the development of angina in patients with coexisting coronary artery disease.

Atrial fibrillation is classified into three patterns: paroxysmal, persistent, or permanent. In paroxysmal atrial fibrillation, spontaneously terminating attacks of palpitation last anything from a few seconds to a few days. The ventricular rate is often rapid and the patient may be severely symptomatic. The term 'persistent atrial fibrillation' is used to describe instances where the arrhythmia is not self-terminating, but where sinus rhythm can be restored by electrical or pharmacological cardioversion. In paroxysmal and persistent atrial fibrillation, the objectives of therapy are the restoration and maintenance of sinus rhythm. Permanent atrial fibrillation describes the situation where restoration of sinus rhythm is no longer possible, and the principal objective of therapy is control of the ventricular rate. At this stage, the ventricular rate is often slower and the patient may be unaware of the irregular pulse or of palpitations.

Management of atrial fibrillation

The management of a patient in atrial fibrillation depends upon the duration of the episode, the presence of organic heart disease, and any precipitating factors. Atrial fibrillation of recent onset (for example, less than 12 h) may terminate spontaneously. If fibrillation is persistent, an attempt to restore sinus rhythm should be made unless the arrhythmia is obviously long-standing or is associated with advanced organic heart disease. Underlying precipitating factors such as thyrotoxicosis should be corrected before attempting cardioversion. Chemical cardioversion may be achieved with class Ia, Ic, or III agents. Class Ia agents accelerate the ventricular rate by virtue of their anticholinergic action on the AV node and must be used in combination with digoxin. Traditionally, the commonest agent used was quinidine (1–2 g/day), but use of this drug has declined with the increasing recognition of adverse effects, in particular the risk of *torsade de pointes*. For patients without significant underlying heart disease, the current drugs of choice are the class 1c agents (for example, flecainide 2 mg/kg intravenously over 30 min). Class III drugs are somewhat less effective but are safer in the presence of left ventricular dysfunction or ischaemic heart disease. Options include sotalol (1.5 mg/kg intravenously over 30 min) or amiodarone (300 mg intravenously over 30 min followed by 1200 mg per 24 h until cardioversion). The pure class III agent ibutilide is approved for this indication in the United States. Normally, only one drug should be tried in any individual patient. If drug therapy fails, direct current cardioversion is commonly

effective.

Following successful cardioversion, or in the presence of paroxysmal atrial fibrillation, prophylactic therapy should be considered, particularly if multiple episodes have occurred. No drug is entirely satisfactory. Quinidine, the traditional mainstay of prophylaxis, increases mortality and is best avoided. Class 1c agents (flecainide or propafenone) are effective and safe in the absence of underlying ischaemia or left ventricular dysfunction. Sotalol (80 to 160 mg twice daily) is also effective and well tolerated. A randomized clinical trial comparing amiodarone, sotalol, and propafenone in the prophylaxis of atrial fibrillation showed amiodarone to be clearly superior to the other two drugs in the prevention of recurrent fibrillation following DC cardioversion. However, the side-effect profile is such that it is rarely indicated for long-term use unless the arrhythmia is troublesome and fails to respond to other drugs.

In permanent atrial fibrillation, restoration of sinus rhythm is not feasible or is unsuccessful and chronic management involves control of ventricular rate. The mainstay of treatment is digoxin, at a dose titrated to achieve adequate slowing in the ventricular rate at rest, with therapeutic plasma concentrations. Despite adequate rate control at rest, patients with atrial fibrillation commonly have an uncontrolled heart rate on exercise. Control of rate response with additional atrioventricular nodal blocking drugs such as verapamil or b-blockers does not improve exercise tolerance in the short term, but improved rate control reduces the risk of development of tachycardiomyopathy and chronic heart failure. Rate control is especially important if the duration of diastole is critical, as in mitral stenosis or ischaemic heart disease.

If drug therapy fails to control paroxysmal atrial fibrillation, particularly in the absence of underlying heart disease, consideration should be given to the possibility of a focal mechanism, which may be amenable to radiofrequency ablation. Atrioventricular nodal ablation and implantation of a permanent pacemaker may be indicated if such treatment fails, or in the case of permanent atrial fibrillation with uncontrolled ventricular rate. Ablation and pacing may improve symptoms and function, but it is wise to defer therapy if possible in view of the irreversible nature of the procedure and the need for lifelong permanent pacing. Furthermore, many patients with paroxysmal atrial fibrillation revert into permanent atrial fibrillation with a marked improvement in symptoms.

Prophylaxis against thromboembolism should be considered in all patients in atrial fibrillation. Cardioversion may be associated with embolism, hence patients who are scheduled to have elective chemical or electrical cardioversion should ideally be treated with warfarin for up to 4 weeks before admission. If the arrhythmia is known to have started within the previous 24 h, intravenous heparin for 24 to 48 h is acceptable. Once warfarin has been started, it should be continued for a minimum of 4 weeks after cardioversion since atrial mechanical function recovers slowly and there is a high risk of recurrent fibrillation.

Chronic anticoagulation with warfarin is indicated in patients in atrial fibrillation with mitral stenosis or regurgitation. Recent studies have shown that patients with non-rheumatic atrial fibrillation will also benefit from prophylaxis against thromboembolism. Meta-analysis of these trials shows that warfarin anticoagulation with a target range for the International normalized ratio (**INR**) of between 2.0 and 3.0 reduces the risk of thromboembolic events by about 60 per cent. Aspirin is a significantly less-effective alternative, achieving a risk reduction of around 20 per cent. The choice of antithrombotic prophylaxis depends on balancing the risk of thromboembolism (Table 11) against the risk of haemorrhagic complications, as well as the local facilities for anticoagulant control.

Atrial flutter

Atrial flutter is caused by a macro re-entrant circuit in the right atrium (Fig. 22), which produces a typical electrocardiographic 'saw tooth' pattern of atrial activity with a rate close to 300/min (Fig. 23). In the common form of the arrhythmia, flutter waves are negative in leads II, III, and aVF and positive in lead V1. Atrial flutter may be associated with either a regular or irregular ventricular response. Flutter with 2:1 atrioventricular conduction produces a regular tachycardia of 150/min and should always be considered in the differential diagnosis of a regular, narrow-QRS tachycardia of this rate. Occasionally, flutter occurs with 1:1 atrioventricular conduction producing a ventricular rate approaching 300/min. The flutter waves may not be seen easily with faster ventricular rates, and transient slowing of AV conduction may be necessary to make the diagnosis (Fig. 23).

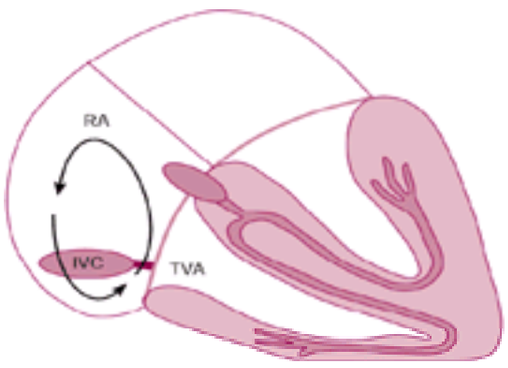

Fig. 22 Mechanism of atrial flutter. Typical atrial flutter results from a counter-clockwise re-entry circuit in the right atrium. The isthmus between the tricuspid valve annulus (**TVA**) and inferior vena cava (**IVC**) forms a critical part of this circuit, and linear ablation to create block can prevent recurrent atrial flutter.

Fig. 23 Atrial flutter with 1:1 AV conduction (above), 2:1 conduction (middle), and following adenosine administration (below) (6 mg intravenous injection 10 s previously).

The underlying causes of atrial flutter are the same as those of atrial fibrillation (Table 10). Although atrial flutter may last for many months or occasionally years, it usually degenerates into chronic atrial fibrillation unless cardioversion is undertaken. Atrial flutter also carries a risk of thromboembolism, and anticoagulation is indicated before and after cardioversion as for atrial fibrillation.

It is important to attempt to terminate atrial flutter since the ventricular rate is often poorly controlled by atrioventricular nodal blocking drugs. Termination may be achieved by chemical or electrical cardioversion as described above for atrial fibrillation. Bursts of atrial overdrive pacing at a rate approximately 10 per cent above the atrial flutter rate are also used: this may restore sinus rhythm or precipitate atrial fibrillation. Prophylaxis against atrial flutter is undertaken using the same agents as in paroxysmal atrial fibrillation, indeed the conditions often coexist and patients may manifest either flutter or fibrillation at different times.

Curative treatment of atrial flutter by radiofrequency ablation can be achieved by creating a line of conduction block between the tricuspid valve annulus and the inferior vena cava. This interrupts the isthmus through which the re-entry circuit must pass (Fig. 22).

Atrial tachycardia

Atrial tachycardia usually results in an atrial rate between 120 and 250/min. As in atrial flutter, there may be a degree of AV block, although 1:1 AV conduction may occur. The ECG shows regular P-waves which do not show the same 'saw tooth' appearance as in atrial flutter (Fig. 24). Atrial tachycardia may occur as a result of sinus node re-entry, when sudden paroxysms of tachycardia with a normal P-wave morphology will arise. Automatic atrial tachycardia manifests an abnormal P-wave morphology, commonly with a longer PR interval. The rate characteristically accelerates or 'warms up' before reaching a rate of 125 to 200/min. This arrhythmia is not started or terminated by atrial extrasystoles. Atrial tachycardia with atrioventricular conduction block is a manifestation of digitalis toxicity. Multifocal atrial tachycardia, in which rapid, irregular P-waves of three or four different morphologies are seen, may occur in severely ill elderly patients or in association with acute exacerbation of pulmonary disease.

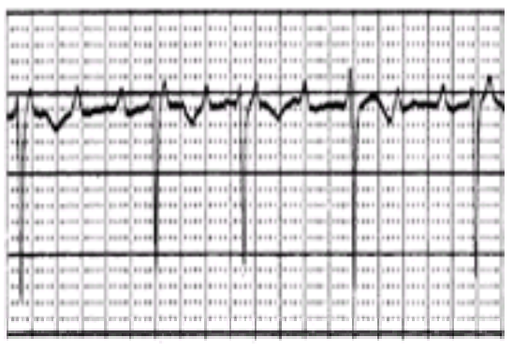

Fig. 24 Atrial tachycardia, with variable AV conduction. Lead V_1.

The approach to management is identical to that of atrial fibrillation. Focal atrial tachycardia can be treated by radiofrequency ablation.

Junctional re-entry tachycardias

The majority of regular narrow-complex tachycardias are junctional re-entry tachycardias, which involve the atrioventricular node in the re-entry circuit. Correct recognition of these arrhythmias has achieved additional importance with the development of effective curative measures.

Atrioventricular nodal re-entry tachycardia

This is the commonest cause of paroxysmal re-entry tachycardia manifesting regular, normal QRS complexes. The basis of the arrhythmia is the presence of two functionally distinct pathways in the region of the atrioventricular node (Fig. 25). The 'fast' pathway conducts more rapidly, but has a longer refractory period. The 'slow' pathway has slower conduction properties but a shorter refractory period. During sinus rhythm, atrioventricular nodal conduction occurs via the fast pathway with a normal PR interval (Fig. 25(a)). If a sufficiently premature atrial extrasystole arises, conduction in the fast pathway is blocked, but slow pathway conduction may continue, resulting in an abrupt increase in the AH interval as recorded in the His–bundle electrogram and corresponding to an increased PR interval on the surface ECG. Conduction down the slow pathway may be sufficiently tardy to allow the fast pathway to recover excitability before activation reaches the distal end of the pathways, allowing retrograde activation to occur via the fast pathway (Fig. 25(b)). The stage is then set for a re-entry circuit with anterograde conduction via the slow pathway and retrograde conduction via the fast pathway ('slow/fast atrioventricular nodal re-entry') (Fig. 25(c)). The arrhythmia circuit is functionally distinct from the atria and ventricles, which may be perturbed by extrastimuli without interruption of the tachycardia. Characteristically, anterograde activation of the ventricles and retrograde activation of the atria occur virtually simultaneously, resulting in the P-wave being 'buried' within the QRS complex, or producing a very small distortion of the terminal QRS, which requires careful comparison with the ECG during sinus rhythm (Fig. 26). The tachycardia is readily initiated by atrial premature stimulation, and terminated by appropriately timed extrastimuli or by overdrive pacing.

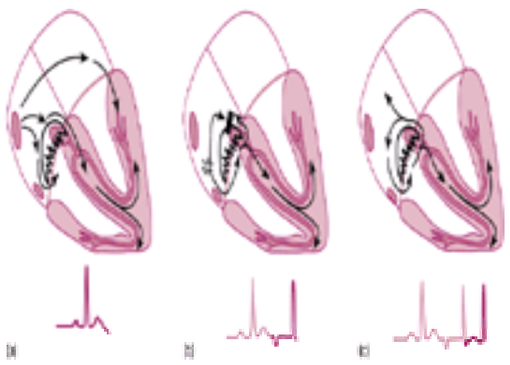

Fig. 25 Atrioventricular nodal re-entry tachycardia. Mechanism of initiation by atrial extrasystole. See text for details.

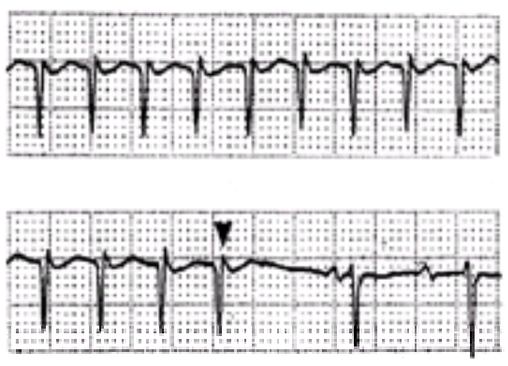

Fig. 26 Atrioventricular nodal re-entrant tachycardia. Rapid narrow-complex tachycardia with no apparent P-waves (upper) responding to 6 mg adenosine with restoration of sinus rhythm (lower). Close inspection reveals a positive deflection of the terminal QRS during tachycardia (arrow) which is absent during sinus rhythm. This is due to retrograde atrial activity coincident with ventricular activation. Lead V_1.

A less common variant of atrioventricular nodal tachycardia may arise where anterograde conduction during tachycardia is via the fast pathway with retrograde conduction via the slow pathway ('fast/slow atrioventricular nodal re-entry'). Under these circumstances, the atrium is activated well after the QRS complex, characteristically producing an inverted P' wave with the RP' interval greater than the P'R interval during tachycardia (Fig. 27). Occasionally, slow/fast and fast/slow tachycardias may coexist in the same patient.

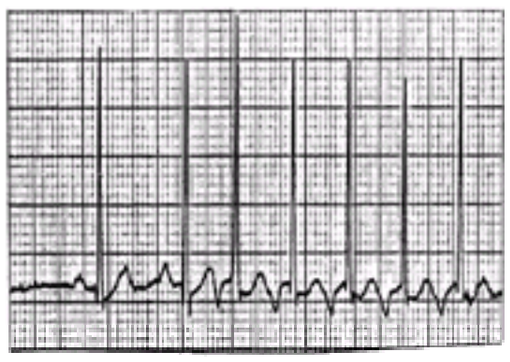

Fig. 27 Atypical atrioventricular nodal re-entry tachycardia (long RP'). Inverted P-waves precede the QRS complex during tachycardia (compare with preceding sinus beats).

Atrioventricular nodal re-entry tachycardia commonly presents for the first time in childhood or adolescence, although it may appear at any age. The natural history is of episodic paroxysmal tachycardia. Attacks occur at random intervals, although clustering of attacks may occur interposed with periods of relative freedom from symptoms. Atrioventricular nodal re-entry tachycardia has no specific association with other organic heart disease. Palpitations are normally well tolerated unless the tachycardia is particularly rapid, prolonged, or if the patient has other heart disease.

Management

Termination of an attack of atrioventricular nodal re-entry tachycardia is achieved by producing a transient AV nodal block. This may be achieved by vagotonic manoeuvres, by intravenous verapamil (5 to 10 mg), or by intravenous adenosine (3 to 12 mg) (Fig. 26). Drug prophylaxis of AV nodal re-entry tachycardia is undertaken with b-blockers, a combined b-blocker/class III agent such as sotalol, or with atrioventricular nodal blocking drugs such as verapamil or digoxin. Curative treatment of AV nodal re-entry tachycardia is readily achieved by ablation and is indicated if patients are refractory to drugs, intolerant of side-effects, or unwilling to take long-term medication. Radiofrequency energy is delivered to the 'slow' pathway, which lies between the compact atrioventricular node and the tricuspid annulus. Ablation at this site is normally curative but carries a small risk (1 to 2 per cent) of inducing complete heart block.

Pre-excitation syndromes (Wolff–Parkinson–White syndrome)

The term 'pre-excitation' refers to the premature activation of the ventricle via one or more accessory pathways that bypass the normal atrioventricular node and His–Purkinje system. The commonest of the pre-excitation syndromes is the Wolff–Parkinson–White syndrome, in which accessory pathways with electrophysiological properties of normal myocardium may lie at any point in the atrioventricular ring, the commonest sites being in the left free wall and the posteroseptal region. The characteristic electrocardiographic appearance is of early activation of the myocardium adjacent to the ventricular insertion of the accessory pathway (Fig. 28(a)). There is no atrioventricular delay, hence the PR interval is shortened, but slow intraventricular conduction results in slurred initiation of the QRS complex (the delta wave) (Fig. 29), although the remainder of the ventricle is excited via the normal His–Purkinje system. The ECG appearances of a delta wave occur in approximately 1.5 per 1000 of the population, but many individuals never experience paroxysmal tachycardias. The degree of pre-excitation during sinus rhythm is variable: it may be *intermittent* if the refractory period of the accessory pathway is close to the sinus cycle length (Fig. 29), or *inapparent* if the delta wave is obscured due to rapid AV nodal conduction. In such instances, transient slowing of AV nodal conduction (for example, by adenosine) will enhance the proportion of the ventricle excited by the accessory pathway and reveal pre-excitation.

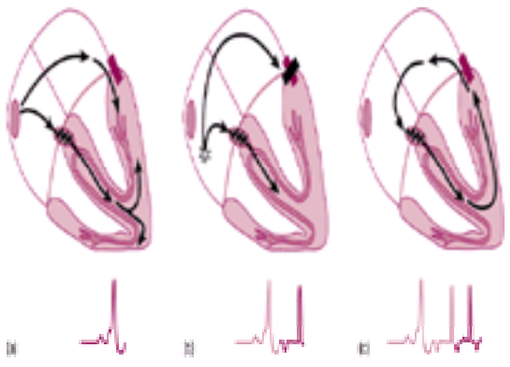

Fig. 28 Atrioventricular re-entry tachycardia. Mechanism of initiation by atrial extrasystole. See text for details.

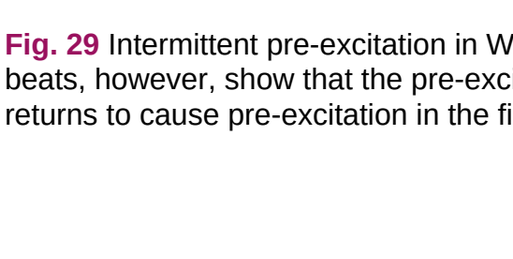

Fig. 29 Intermittent pre-excitation in Wolff–Parkinson–White syndrome. The first two beats show the characteristic short PR interval and delta wave. The middle two beats, however, show that the pre-excitation was intermittent. The pathway has become refractory, with normal PR interval and QRS morphology. Pathway conduction returns to cause pre-excitation in the final two beats.

In many instances, accessory pathways conduct only in the retrograde (ventriculoatrial) direction, and do not cause ventricular pre-excitation. Such pathways are termed *concealed*, since there is no clue to their presence on the resting ECG. These patients are not at risk of pre-excited atrial fibrillation (see below), but either overt or concealed accessory pathways can lead to episodes of atrioventricular re-entry tachycardia.

Atrioventricular re-entry tachycardia

The mechanism of orthodromic tachycardia, the common form of atrioventricular re-entry tachycardia, is illustrated in Fig. 28. A premature atrial extrasystole may find the accessory pathway refractory, but be conducted through the atrioventricular node to the ventricles (Fig. 28(b)). If sufficient delay has occurred by the time the ventricular insertion of the accessory pathway is depolarized, the pathway will have recovered excitability and allow retrograde activation from the ventricle to atrium, with the establishment of a re-entry circuit (Fig. 28(c)). Since the circuit involves activation of the ventricles via the His–Purkinje system, the QRS morphology during re-entry tachycardia is normal, unless a rate-related, bundle-branch block develops. Retrograde atrial activation can be identified by the presence of a characteristic

inverted P' wave early in the ST segment, an important diagnostic feature of atrioventricular tachycardia (Table 7, Fig. 30).

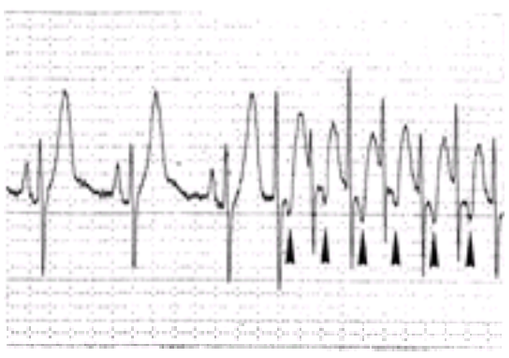

Fig. 30 Initiation of atrioventricular re-entry tachycardia. The third sinus beat is followed by the onset of narrow-complex tachycardia, initiated by an atrial extrasystole (obscured by T-wave). Retrograde atrial activation, with inverted P-waves in the ST segment (arrows), are seen during tachycardia.

Antidromic tachycardia

A rarer form of atrioventricular tachycardia has anterograde conduction via the accessory pathway and retrograde conduction via the atrioventricular node (antidromic tachycardia). The QRS morphology of this tachycardia is grossly abnormal with appearances dependent upon the site of insertion of the accessory pathway.

Other forms of pre-excitation include the Mahaim pathway, a direct atrioventricular or atriofasicular connection with slow conduction properties typical of AV nodal tissue. Evidence for direct atrionodal pathways associated with a short PR interval but no delta wave (Lown–Ganong–Levine syndrome) remains controversial and has not been established histologically.

Pre-excited atrial fibrillation

The major prognostic concern in Wolff–Parkinson–White syndrome is pre-excited atrial fibrillation. Conduction via an accessory pathway with a short refractory period, bypassing the normal AV nodal slowing, results in very rapid ventricular conduction that may degenerate into ventricular fibrillation (Fig. 31). The degree of pre-excitation during atrial fibrillation varies, giving a characteristic pattern of an irregular ventricular response with QRS morphology ranging from normal to fully pre-excited.

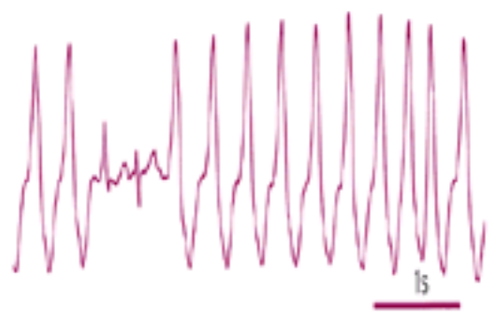

Fig. 31 Pre-excited atrial fibrillation. Conduction via an accessory pathway results in an irregular wide-complex tachycardia. The third and fourth beats show less pre-excitation, with activation mainly through the normal conducting system, with more normal QRS-complex morphology. Lead V_1.

Patients with symptomatic Wolff–Parkinson–White syndrome should be evaluated carefully for the risk of pre-excited atrial fibrillation. If pre-excitation is intermittent, this is commonly associated with a long accessory-pathway refractory period and a low risk of life-threatening tachycardias. Disappearance of the delta wave in response to administration of a class Ia or Ic antiarrhythmic drug also suggests a low risk. The risk of sudden death due to rapid pre-excited atrial fibrillation is very low among patients who have not had any symptomatic tachycardias, but is higher in symptomatic patients, particularly if episodes of pre-excited atrial fibrillation have been documented. Risk can be assessed by analysis of the shortest pre-excited RR intervals during spontaneous or induced episodes of pre-excited atrial fibrillation, a value of less than 250 ms identifying a higher risk group. Patients who have experienced atrial fibrillation with relatively slow ventricular rates are unlikely to develop faster ventricular responses subsequently. The general tendency is for accessory pathway conduction to become slower with increasing age, and spontaneous disappearance of conduction is well documented.

Management of patients with accessory pathways

Orthodromic re-entry tachycardia may be terminated by AV nodal blocking manoeuvres such as vagal stimulation, verapamil, or adenosine. Pre-excited atrial fibrillation requires particular care, since digoxin or verapamil may paradoxically accelerate the ventricular rate and are contraindicated. Electrical cardioversion is indicated in the presence of severe haemodynamic disturbance. Where patients are stable, agents such as intravenous flecainide, sotalol, or amiodarone, which both slow anterograde conduction through the accessory pathway and also restore sinus rhythm, are used. Drug prophylaxis is used to minimize the risk of recurrent orthodromic re-entry tachycardia or atrial fibrillation. Drugs acting only on the AV node, such as verapamil, are less effective than agents having additional action on the accessory pathway such as flecainide and sotalol. Although amiodarone is effective, its use is not desirable in otherwise fit young people who may require long-term drug therapy.

Patients with symptomatic Wolff–Parkinson–White syndrome are increasingly offered radiofrequency ablation as first-line therapy. This approach abolishes the risk of pre-excited atrial fibrillation as well as preventing further attacks of atrioventricular re-entry tachycardia. Careful electrode mapping of the atrioventricular annulus is necessary to identify the accessory pathway, the site at which the interval between the atrial and ventricular electrograms is at a minimum, ideally with a discrete accessory pathway potential between the signals. Passage of the radiofrequency current results in the disappearance of accessory pathway conduction within a few seconds (Fig. 32). The success rate of ablation varies according to the location of the pathway, but is usually over 90 per cent in experienced hands.

Fig. 32 Radiofrequency ablation of an accessory pathway. The patient had Wolff–Parkinson–White syndrome with evidence of ventricular pre-excitation on the surface electrogram during sinus rhythm (short PR interval, delta wave). One beat after switching on the radiofrequency (**RF**) current the QRS becomes normal, indicating successful ablation of the accessory pathway. This was a left-sided accessory pathway, as shown by the short interval between left atrial and left ventricular activation recorded from the coronary sinus (**CS**). This interval is prolonged following ablation of the pathway. Surface leads I, V_1, and intracardiac electrograms from CS and mapping catheter (**Map**) are shown.

Radiofrequency ablation is indicated in patients with tachycardias due to concealed accessory pathways if they are not well controlled on drugs, intolerant of side-effects, or unwilling to take long-term medication. Localization of the accessory pathway is performed as described above, except that mapping is performed during ventricular pacing or during stable atrioventricular tachycardia, and the earliest site of retrograde atrial activation at the atrioventricular annulus is the site of ablation.

Ventricular tachyarrhythmias

Definitions

Ventricular tachycardia is defined as the presence of three or more consecutive ventricular beats at a rate of 120 per min or greater. Ventricular tachycardia is considered *sustained* if an individual salvo lasts for 30 s or more, and *non-sustained* if the duration is between 3 beats and 30 s. *Monomorphic* ventricular tachycardia demonstrates a consistent QRS morphology during each paroxysm, although patients may have paroxysms of monomorphic ventricular tachycardia of different morphologies at different times. *Polymorphic* ventricular tachycardia demonstrates a constantly changing QRS morphology, often without discrete QRS complexes. Polymorphic ventricular tachycardia may degenerate into ventricular fibrillation and the electrocardiographic distinction between the two is difficult. *Torsade de pointes* is a characteristic type of polymorphic ventricular tachycardia with a typical undulating variation in QRS morphology as a result of variation in axis. The term is reserved for the arrhythmias arising in association with QT interval prolongation.

Sustained monomorphic ventricular tachycardia

Aetiology

Sustained monomorphic ventricular tachycardia commonly occurs in the presence of structural heart disease, but also arises in structurally normal hearts. It rarely occurs in the acute phase of myocardial infarction, but may be seen in the subacute phase (>48 h), or may arise many years after the index infarction, particularly in association with left ventricular dilatation and aneurysm formation. The arrhythmia also occurs in other conditions associated with ventricular dilatation or fibrosis such as dilated cardiomyopathy, hypertrophic cardiomyopathy, or previous ventriculotomy (for example, following repair of Fallot's tetralogy). Sustained monomorphic tachycardia can occur as a proarrhythmic response to antiarrhythmic drugs, particularly class I agents.

Although ventricular tachycardia normally occurs in individuals with overt heart disease, it is also seen in young, apparently healthy, subjects. Arrhythmogenic right ventricular cardiomyopathy (dysplasia) is an autosomal dominant condition associated with replacement of the right ventricular free wall with fat and fibrous tissue. These patients may have no symptoms or signs of cardiac disease, but typical ECG changes (T wave inversion in the right precordial leads) are associated with variable degrees of dilatation of the right ventricle demonstrated on echocardiography or magnetic resonance imaging. There remains a minority of patients with documented ventricular tachycardia in whom no structural heart disease is evident on clinical, ECG, or echocardiographic examination. The tachycardia may arise from the outflow tract of the right or, rarely, left ventricle, or from one of the fascicles of the left bundle branch.

ECG characteristics

The presence of atrioventricular dissociation is a particularly important feature to seek in a wide-complex tachycardia as it makes the diagnosis of ventricular tachycardia virtually certain (Table 8, Fig. 33(a)). A careful search for P-waves perturbing the QRS complex or T-waves is necessary, ideally using multichannel recordings. Occasionally, a fortuitously timed P-wave allows the development of a capture beat of normal QRS morphology without interrupting the tachycardia. A fusion beat occurs when activation of the ventricle is partly via the normal His–Purkinje system and partly from the tachycardia focus (Fig. 33(b)). Fusion and capture beats are diagnostic of ventricular tachycardia, but are commonly present only if the ventricular rate is relatively slow. Where dissociated P-wave activity cannot be recognized with certainty on the surface ECG, direct recording of atrial activity by an oesophageal or right atrial electrogram may aid the diagnosis. Although atrioventricular dissociation is diagnostic of ventricular tachycardia, it is not invariable. Retrograde ventriculoatrial conduction may occur, giving either 1:1 conduction or higher degrees of block (Fig. 33(c)).

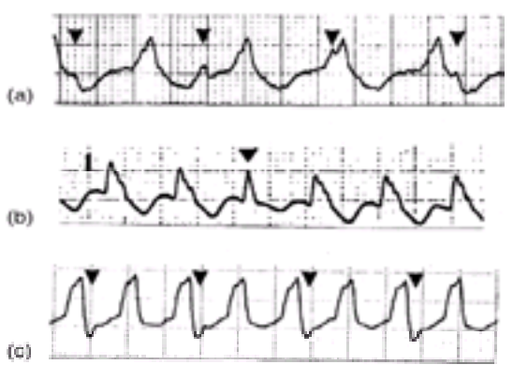

Fig. 33 Sustained monomorphic ventricular tachycardia. (a) Ventricular tachycardia with atrioventricular dissociation. P-waves (arrowed) are seen to have no relationship to the ventricular activation. Lead V_1. (b) Ventricular tachycardia with fusion beat (arrow). Lead V_1. (c) Ventricular tachycardia with 2:1 ventriculoatrial conduction. Lead III. Inverted P-waves (arrows) follow every second ventricular complex.

The QRS duration in ventricular tachycardia is commonly greater than 0.12 s, and values greater than 0.14 s are particularly suggestive of ventricular tachycardia. Although the QRS morphology may superficially resemble left or right bundle-branch block, the morphology is commonly atypical (Table 8). Ventricular tachycardia arising from the right ventricular free wall has a left bundle-branch block-like pattern, whilst left ventricular free wall tachycardias show right bundle-branch block morphology. The presence of concordant positive or negative QRS complexes across the chest leads is suggestive of ventricular tachycardia, as is the existence of extreme axis deviation.

Acute management of ventricular tachycardia

Rapid ventricular tachycardia may present with cardiac arrest, syncope, shock, anginal chest pain, or left ventricular failure, but slower tachycardias in patients with good cardiac function may be well tolerated.

Sustained ventricular tachycardia is a medical emergency. If the patient is pulseless or unconscious, immediate DC cardioversion is necessary. If the patient is conscious but hypotensive, urgent DC cardioversion under general anaesthesia or deep sedation is used. Haemodynamically tolerated tachycardias may be terminated by drug therapy. The commonest agent used is intravenous lidocaine (lignocaine) 100 mg, repeated if necessary after 5 min. Sotalol 1.5 mg/kg intravenously is more effective, but its use is restricted by its negative inotropic action. Second-line drugs for the termination of ventricular tachycardia include procainamide, disopyramide and amiodarone. Amiodarone normally has a slow onset of action but may be effective if the tachycardia is well tolerated. Flecainide is

contraindicated in view of the risk of developing incessant tachycardia. All antiarrhythmic drugs have significant negative inotropic actions that may further impair the haemodynamic status of the patient if sinus rhythm is not restored. For this reason, no more than one antiarrhythmic drug should normally be given before recourse to alternative therapy, usually DC cardioversion. Overdrive termination of ventricular tachycardia following insertion of a temporary pacing lead may be effective (Fig. 16), particularly if the tachycardia is relatively slow. Facilities for cardioversion must be available in view of the risk of acceleration or degeneration into ventricular fibrillation.

Prophylaxis of ventricular tachycardia

Ventricular tachycardia is a potentially life-threatening condition. Unless the acute episode was clearly precipitated by some transient or reversible factor, there is a high probability of recurrent attacks, which may result in sudden death rather than a sustained tachycardia. The 3-year cardiac survival rate varies from 80 per cent in patients in whom arrhythmia induction is suppressed by antiarrhythmic drug therapy, to 40 per cent in those in whom no effect of suppression is achieved and/or empirical therapy is used.

Clinical evaluation of the patient after restoration of sinus rhythm should be supported by electrocardiography, echocardiography, and/or radionuclide ventriculography. In those with ischaemic heart disease, exercise testing should be undertaken to identify the presence of reversible ischaemia, which may act as a trigger to ventricular tachycardia, and coronary arteriography to determine the extent of arterial disease. Particular attention should be paid to the possibility of right ventricular dysplasia in young patients.

Unless there is a clear precipitating factor such as drug toxicity, electrolyte abnormality, or acute ischaemia, patients who have had documented ventricular tachycardia require antiarrhythmic prophylaxis. The most reliable form of prophylaxis against arrhythmic sudden death or recurrent sustained ventricular tachycardia is provided by the implantable cardioverter-defibrillator (Fig. 18). The Antiarrhythmics versus Implantable Defibrillators Trial (**AVID**) showed defibrillators to be superior in preventing death from any cause, in comparison to drug therapy with amiodarone or sotalol in patients resuscitated from ventricular fibrillation or ventricular tachycardia causing haemodynamic compromise (Fig. 34). Those with non-sustained ventricular tachycardia (see below) and left ventricular dysfunction, in whom sustained tachycardia can be induced at electrophysiological testing, also have a better survival with defibrillator implantation compared with drug therapy. Indeed, the indications for such treatment are expanding to include a wide range of patients with sustained or non-sustained ventricular tachyarrhythmias who are at risk of sudden death.

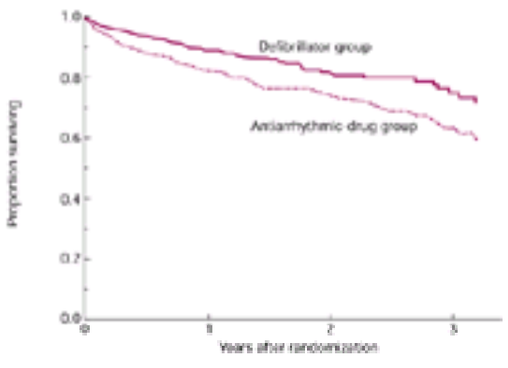

Fig. 34 Improved survival with the implantable cardioverter defibrillator compared to antiarrhythmic drugs in patients resuscitated from near-fatal ventricular arrhythmia. (Taken with permission from the AVID investigators (1997). *New England Journal of Medicine* **337**, 1576. Copyright 1997, Massachusetts Medical Society. All rights reserved).

In view of the cost and complexity of implantable defibrillator therapy, it is not affordable in all countries, and not appropriate for patients with advanced congestive heart failure or other conditions with a severely limited prognosis. Medical therapy is necessary for many patients, but is limited by a relative lack of randomized, controlled-trial evidence. Beta-adrenoceptor blockers have been shown to reduce the risk of sudden death in unselected survivors of myocardial infarction, and to be comparable to conventional antiarrhythmic agents in the prevention of recurrent ventricular tachyarrhythmias. Since b-adrenoceptor blockers are now shown to be of prognostic benefit even in the presence of severe left ventricular dysfunction, they should be used routinely in the prophylaxis of ventricular tachycardia if tolerated. Of the conventional antiarrhythmic agents, there is evidence that the class III drugs sotalol and amiodarone are superior to class I antiarrhythmic agents, which should no longer be used for this indication. However, sotalol and amiodarone have not been tested against placebo or conventional b-adrenoceptor blockers in randomized trials, although observational studies suggest they are of benefit in the prevention of arrhythmic death.

The efficacy of any given antiarrhythmic drug cannot be predicted, thus it is necessary to demonstrate antiarrhythmic drug efficacy and to exclude proarrhythmic responses in each patient. If episodes of ventricular tachycardia, sustained or non-sustained, are occurring frequently, it is sufficient to administer antiarrhythmic drugs under continuous electrocardiographic monitoring, and to demonstrate that salvoes of tachycardia have been completely suppressed. If episodes of ventricular tachycardia are infrequent, cardiac electrophysiological testing is indicated. The initial objective of such testing is to initiate a tachycardia of similar rate and QRS morphology to the spontaneous arrhythmia (Fig. 2): drug therapy is then administered and a repeat study is undertaken once stable plasma levels of the drug have been achieved. Suppression of inducibility of the tachycardia is associated with a reduced risk of arrhythmia recurrence and an improved prognosis in comparison with patients whose arrhythmia is not suppressed. Even if ventricular tachycardia is still inducible, the presence of marked slowing (increase in cycle length greater than 100 ms) associated with good haemodynamic tolerance appears to indicate a good long-term prognosis.

Radiofrequency ablation is used in the management of ventricular tachycardia, particularly in right ventricular outflow tract or fascicular tachycardia. Location and ablation of critical areas of slow conduction in ventricular tachycardias due to previous myocardial infarction are feasible but technically difficult, with lower rates of success than for other types of ablation.

Direct surgical management of recurrent ventricular tachycardia involves aneurysmectomy, endocardial mapping, and resection of the subendocardial area containing the micro re-entry circuit. The indications for surgery have been reduced considerably since the advent of the implantable cardioverter-defibrillator, since the surgical mortality is up to 10 to 15 per cent, compared with 0.5 per cent for defibrillator implantation. Where medically intractable ventricular tachyarrhythmias are associated with very poor left ventricular function, the only possible therapeutic option is cardiac transplantation.

Non-sustained ventricular tachycardia

Clinical features

The mechanism and causes of non-sustained ventricular tachycardia (Fig. 35) are similar to those of sustained ventricular tachycardia. There is often slight variation in the RR interval, particularly if the salvo involves only a few beats. Short salvoes of non-sustained ventricular tachycardia are often asymptomatic; more prolonged episodes may result in dizziness or presyncope, and occasionally in syncope. Apart from the instances where non-sustained ventricular tachycardia produces troublesome symptoms, the major clinical significance of the arrhythmia is as a risk marker for sustained ventricular tachycardia or sudden cardiac death. However, long-term follow-up of patients with non-sustained ventricular tachycardia in the absence of structural heart disease has indicated a good prognosis with no excess risk, although non-sustained ventricular tachycardia recorded by ambulatory ECG monitoring in the convalescent phase or after remote myocardial infarction is an independent risk factor for subsequent sudden cardiac death, especially if it is associated with impaired left ventricular function. The risk of arrhythmic death is particularly high if sustained ventricular tachycardia can be initiated in these patients by electrophysiological testing. Non-sustained ventricular tachycardia is also an adverse prognostic feature in patients with hypertrophic cardiomyopathy. Asymptomatic non-sustained ventricular tachycardia is commonly recorded in patients with advanced congestive heart failure: it is associated with an increased risk of cardiac death, but not selectively of sudden death.

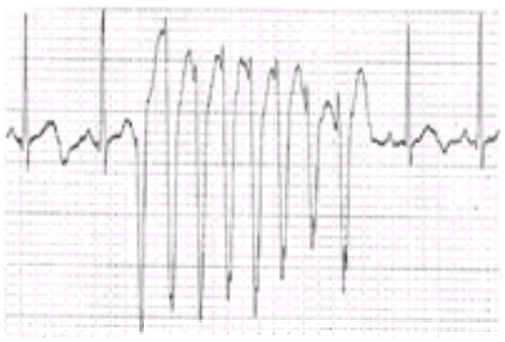

Fig. 35 Non-sustained ventricular tachycardia.

Management

The management of non-sustained ventricular tachycardia involves the identification of underlying organic heart disease, as described in the section on sustained monomorphic ventricular tachycardia. Patients should be evaluated non-invasively by echocardiography or radionuclide ventriculography. If no significant organic heart disease is present, and the patient is asymptomatic, no treatment is indicated. Treatment of symptoms in the absence of significant heart disease should be with b-blockers in the first instance to minimize the risk of proarrhythmic reactions. Calcium-channel blockers are effective occasionally and may be tried. Failing these, sotalol or a class I agent may be necessary.

Patients with structural heart disease but well-preserved ventricular function and a normal signal-averaged electrocardiogram are at low risk of sustained ventricular tachycardia and may be treated empirically with b-blockers. If non-sustained ventricular tachycardia is associated with impaired ventricular function, there is likely to be a substrate for sustained ventricular tachyarrhythmias. Patients in whom sustained ventricular tachycardia or fibrillation is inducible have an improved survival following defibrillator implantation compared with patients treated medically.

Low-dose amiodarone therapy has been recommended in the management of patients with hypertrophic cardiomyopathy and non-sustained ventricular tachycardia, although the evidence for its efficacy is based on comparison with historical controls rather than on a randomized prospective study.

Accelerated idioventricular rhythm

The term 'accelerated idioventricular rhythm' is used to describe a continuous ventricular rhythm with a rate less than 120/min. Idioventricular rhythm commonly occurs in the setting of acute myocardial infarction and appears to be a marker of successful thrombolytic therapy. No active treatment is necessary.

Polymorphic ventricular tachycardia

Polymorphic ventricular tachycardia is an unstable rhythm with varying QRS morphology. It is most commonly seen in the acute phase of myocardial infarction and is due to unstable re-entry circuits. As such, it commonly undergoes spontaneous termination, although it may degenerate into ventricular fibrillation. If episodes of polymorphic ventricular tachycardia are frequent in the early hours of myocardial infarction, they can be suppressed by intravenous lidocaine (lignocaine). However, short, infrequent episodes are commonly left untreated.

The Brugada syndrome is an autosomal dominant condition due to a mutation of one of the genes encoding the rapid sodium channel (SCN5a), causing partial inactivation. There is an unusual pattern of variable ST-segment elevation and partial right bundle-branch block in the right precordial leads, associated with a risk of polymorphic ventricular tachycardia and sudden death.

Torsade de pointes and the long QT syndromes

ECG characteristics

Torsade de pointes is an atypical ventricular tachycardia characterized by a continuously varying QRS axis ('twisting of points') (Fig. 36). Episodes of *torsade de pointes* are commonly repetitive and normally self-terminating, although they may degenerate into ventricular fibrillation. Paroxysms of *torsade de pointes* are associated in the preceding beats with evidence of marked QT prolongation, and frequently with morphological abnormalities of the T-waves such as T–U fusion, gross increases in T-wave amplitude, or T-wave alternans. Paroxysms of *torsade de pointes* in the congenital syndromes are often associated with increases in sinus rate, while in the acquired syndromes a slowing of the heart rate, and in particular a postextrasystolic pause, is often associated with initiation of the arrhythmia. This produces a characteristic 'short–long–short' sequence of initiation (Fig. 36). The combination of QT interval prolongation during sinus rhythm with intermittent *torsade de pointes* is described as the long QT syndrome.

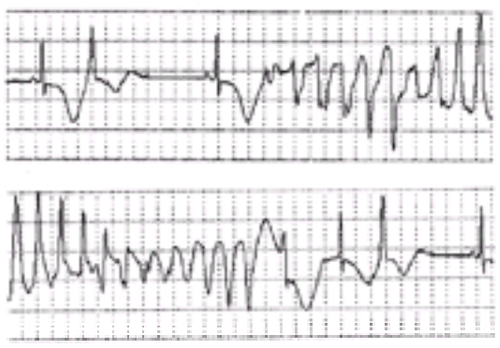

Fig. 36 *Torsade de pointes*. Note the marked QT interval prolongation in the sinus beats. Ambulatory monitoring recording is shown (continuous tracing).

Congenital long QT syndromes

The underlying arrhythmic mechanisms in the congenital syndromes involve mutations in genes encoding proteins in the ion channels conducting either the inward sodium current I_{Na}, or the rapid or slowly inactivating components of the outward potassium current (I_{kr} and I_{ks}, respectively). Multiple mutations have been described and the long QT syndromes are subclassified according to the underlying gene defect (Table 12). All the currently recognized subgroups show an autosomal dominant mode of inheritance. Lengthening of ventricular repolarization, and hence of the QT interval, occurs as a result either of increased duration of current flow via I_{Na}, or inhibition of outward current flow via I_{kr} or I_{ks}. The arrhythmias have characteristics consistent with triggered activity. A variety of congenital long QT syndrome phenotypes have been identified. In the Jervell–Lange–Nielsen syndrome, the long QT gene disorder is inherited as an autosomal dominant, but neural deafness as an autosomal recessive. The other long QT syndromes are not associated with deafness. Sporadic cases of idiopathic, presumed congenital long QT syndrome have been reported.

Attacks of *torsade de pointes* in the congenital syndromes are commonly associated with sympathetic stimulation such as exercise, waking, or fright. Paroxysms may produce syncope, which if prolonged may be complicated by convulsion, leading to misdiagnosis as epilepsy. A family history of recurrent syncope or sudden death

may be obtained. Sinus bradycardia is commonly seen in these syndromes.

Acquired long QT syndromes

Many drugs and other factors predispose to the development of the acquired long QT syndrome (Table 13). Although class Ia and III antiarrhythmic drugs are best known for this complication, it is important to recognize that a very large number of non-cardiac drugs inhibit the outward potassium current I_{kr}, and may cause significant lengthening of the QT interval singly or in combination. Episodes of *torsade de pointes* often occur as a result of a combination of factors, including prolongation of the QT interval in association with bradycardia or pauses, hypokalaemia, and hypomagnesaemia. All of these predispose to early after-depolarizations *in vitro* and this mechanism appears to be the likely cause of *torsade de pointes* in the acquired syndromes. It is increasingly recognized that there is a genetic predisposition to the development of acquired long QT syndrome in the face of predisposing factors, leading to the concept that patients developing acquired long QT syndrome have reduced 'repolarization reserve' as a result of a *forme fruste* of the congenital syndrome.

Acute management of torsade de pointes

The common clinical presentation is of recurrent dizziness or syncope, and the condition may easily be misdiagnosed as self-terminating polymorphic ventricular tachycardia or ventricular fibrillation unless the characteristic morphology of *torsade de pointes* and the associated QT interval prolongation is recognized. It is essential to discontinue predisposing drugs or other agents and to avoid empirical antiarrhythmic drug therapy, which may worsen the arrhythmia. Individual paroxysms of *torsade de pointes* are normally self-limiting, but if they are persistent, cardiac arrest will occur and emergency defibrillation is necessary. Intravenous magnesium sulphate (8 mmol over 10–15 min, repeated if necessary) is a safe and effective emergency measure for the prevention of recurrent paroxysms of tachycardia. If *torsade de pointes* is associated with bradycardia and pauses, the heart rate should be increased to between 90 and 100/min by atrial or ventricular pacing or isoproterenol (isoprenaline) infusion. Hypokalaemia should be sought and corrected if necessary.

Long-term management of long QT syndromes

The prognosis of untreated congenital long QT syndrome is poor, with a high incidence of sudden death in childhood. Patients with the congenital long QT syndrome presenting with attacks of syncope are initially treated with high-dose b-blockade for example, propranolol. If this is unsuccessful, selective high left stellate ganglionectomy has been employed successfully. Permanent pacing at rates of 70 to 80/min, in combination with b-blockers, may also be effective in reducing symptoms. Defibrillator implantation is necessary for resistant cases, and is commonly used as first-line therapy if episodes of *torsade de pointes* have resulted in cardiac arrest. Retrospective data from the International Registry have indicated that the 15-year survival in patients following their first episode of *torsade de pointes* has been improved from 50 per cent in untreated cases to 90 per cent following treatment with b-blockade and/or left stellate ganglionectomy. The prognosis of the acquired long QT syndromes are excellent, provided the underlying predisposing factors are identified and avoided.

Ventricular fibrillation

Ventricular fibrillation is defined as a chaotic, disorganized arrhythmia with no identifiable QRS complexes (Fig. 37). The mechanism is of multiple, unstable re-entry circuits. The electrocardiographic pattern depends on the duration of fibrillation: recent-onset fibrillation is described as 'coarse', with a peak-to-peak amplitude of around 1 mV (1 cm); with increasing duration of cardiac arrest, the amplitude of ventricular fibrillation diminishes and 'fine' ventricular fibrillation is less likely to be amenable to successful electrical defibrillation.

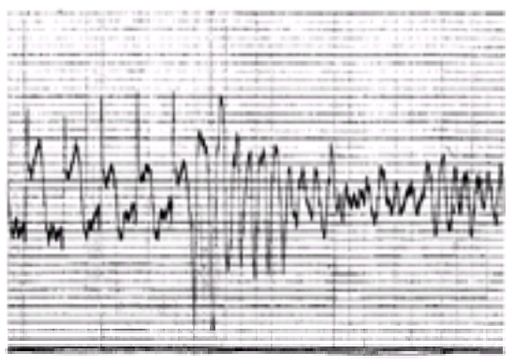

Fig. 37 Ventricular fibrillation complicating acute myocardial infarction. The arrhythmia is initiated by an 'R on T' ventricular extrasystole.

Ventricular fibrillation may represent the endpoint of cardiac disease of many aetiologies. Fibrillation may occur during acute myocardial ischaemia, and is the principal cause of death in the first 2 hours following acute myocardial infarction (Fig. 37). Ventricular fibrillation during myocardial infarction is subdivided into primary, occurring without warning in an otherwise stable patient, and secondary, where fibrillation occurs in the context of left ventricular failure and cardiogenic shock. In acute myocardial infarction, ventricular fibrillation is often initiated by an R on T extrasystole. Ventricular fibrillation occurring in chronic heart disease is most commonly a result of degeneration of rapid ventricular tachycardia, whose causes have been described above. Rarer causes of fibrillation are listed in Table 14.

Ventricular fibrillation is rarely self-terminating, and normally causes cardiac arrest with the rapid onset of pulselessness, unconsciousness, and apnoea. The management of cardiac arrest due to ventricular fibrillation is discussed in Chapter 16.3.

Management of survivors of ventricular fibrillation

Patients who survive an episode of ventricular fibrillation should be assessed carefully to determine the risk of recurrence. If ventricular fibrillation has occurred in the first few hours of a typical Q-wave myocardial infarction, the risk of recurrent cardiac arrest is low, and no specific prophylactic therapy other than conventional postinfarction b-blockade is indicated. In many instances ventricular fibrillation arises as a result of acute ischaemia in patients with known, extensive heart disease who have not sustained an acute infarction. These patients remain at high risk of recurrent ventricular fibrillation, and should be evaluated fully by exercise testing and coronary arteriography with a view to revascularization. Patients may sustain a cardiac arrest without any preceding chest pain, but in the presence of a known risk factor for ventricular tachycardia such as previous myocardial infarction. In these individuals, it is likely that ventricular fibrillation arose as a result of degeneration of rapid ventricular tachycardia. These patients are at risk of further cardiac arrest, and are managed with an implantable cardioverter-defibrillator or antiarrhythmic therapy as discussed in the section on ventricular tachycardia.

Further reading

Andersen HR, *et al*. (1997). Long-term follow-up of patients from a randomized trial of atrial versus ventricular pacing for sick sinus syndrome. *Lancet* **350**, 1210–16.

Atiga WL, Rowe P, Calkins H (1999). Management of vasovagal syncope. *Journal of Cardiovascular Electrophysiology* **10**, 874–86.

Atrial Fibrillation Investigators (1994). Risk factors for stroke and efficacy of anti-thrombotic therapy in atrial fibrillation: analysis of pooled data from five randomised controlled trials. *Archives of Internal Medicine* **154**, 1449–57.

Brugada J, Brugada R, Brugada P (1998). Right bundle-branch block and ST-segment elevation in leads V1 through V3: a marker for sudden death in patients without demonstrable structural heart disease. *Circulation* **97**, 457–60.

Connolly SJ (1999). Evidence-based analysis of amiodarone efficacy and safety. *Circulation* **100**, 2025–34.

Connolly SJ, *et al*. (2000). Effects of physiological pacing versus ventricular pacing on the risk of stroke and death due to cardiovascular cause. *New England Journal of Medicine* **342**, 1385–91.

Domanski MJ, Zipes DP, Schron E (1997). Treatment of sudden cardiac death. Current understandings from randomized trials and future research directions. *Circulation* **95**, 2694–9.

Drew BJ, Scheinman MM (1995). ECG criteria to distinguish between aberrantly conducted supraventricular tachycardia and ventricular tachycardia: practical aspect for the immediate care setting. *Pacing and Cardiac Electrophysiology* **18**, 2194–208.

Echt DS, *et al*. (1991). Mortality and morbidity in patients receiving encainide, flecainide, or placebo. *New England Journal of Medicine*, **324**, 781–8.

Fitzpatrick A, Sutton R (1992). A guide to temporary pacing. *British Medical Journal* **304**, 365–9.

Ginks W, Leatham A, Siddons H (1979). Prognosis of patients paced for chronic atrioventricular block. *British Heart Journal* **41**, 633–6.

Gregoratus G, *et al*. (1998). ACC/AHA guidelines for implantation of cardiac pacemakers and arrhythmia devices. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Pacemaker Implantation). *Journal of the American College of Cardiology* **31**, 1175–209.

Haïssaguerre M, *et al*. (1998). Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins. *New England Journal of Medicine* **339**, 659–66.

Kay GN, Plumb VJ (1996). The present role of radiofrequency catheter ablation in the management of cardiac arrhythmias. *American Journal of Medicine* **100**, 344–56.

Lamas GA, *et al*. (1998). Quality of life and clinical outcome in elderly patients treated with ventricular pacing as compared with dual-chamber pacing. *New England Journal of Medicine* **338**, 1097–104.

Lévy S, *et al*. (1998). Atrial fibrillation: current knowledge and recommendations for management. Working Group on Arrhythmias of the European Society of Cardiology. *European Heart Journal* **19**, 294–320.

Lip GYH (1999). Thromboprophylaxis for atrial fibrillation. *Lancet* **353**, 4–6.

Mehta D, *et al*. (1988). Relative efficacy of physical manoeuvers in the termination of junctional tachycardia. *Lancet* **i**, 1181–5.

Morady F (1999). Radio-frequency ablation as treatment for cardiac arrhythmia. *New England Journal of Medicine* **340**, 534–44.

Morley-Davies A, Cobbe SM (1997). Cardiac pacing. *Lancet* **349**, 41–6.

Moss AJ, *et al*. (1996). Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia. *New England Journal of Medicine* **335**, 1933–40.

Priori SG, *et al*. (1999). Genetic and molecular basis of cardiac arrhythmia; impact on clinical management. Study group on molecular basis of arrhythmias of the working group on arrhythmias of the European Society of Cardiology. *European Heart Journal* **20**, 174–95 (also published in *Circulation* **99**, 518–28, 674–81.)

Priori SG, *et al*. (2001). Task Force on Sudden Cardiac Death of the European Society of Cardiology. *European Heart Journal* **22**, 1374-450.

Rankin AC, Rae AP, Cobbe SM (1987). Misuse of intravenous verapamil in patients with ventricular tachycardia. *Lancet* **ii**, 472–4.

Rankin AC, *et al*. (1989). Value and limitations of adenosine in the diagnosis and treatment of narrow and broad complex tachycardias. *British Heart Journal* **62**, 195–203.

Roden DM (2000). Antiarrhythmic drugs: from mechanisms to clinical practice. *Heart* **84**, 339–46.

Roy D, *et al*. (2000). Amiodarone to prevent recurrence of atrial fibrillation. *New England Journal of Medicine* **342**, 913–20.

Task Force of the Working Group on Arrhythmias of the European Society of Cardiology (1991). The 'Sicilian Gambit'. A new approach to the classification of antiarrhythmic drugs based on their actions and arrhythmogenic mechanisms. *Circulation* **84**, 1831–51.

The Antiarrhythmic Versus Implantable Defibrillators (AVID) Investigators (1997). A comparison of antiarrhythmic-drug therapy with implantable defibrillators in patients resuscitated from near-fatal ventricular arrhythmias. *New England Journal of Medicine* **337**, 1576–83.

Shaw DB, Holman RR, Gowers JI (1980). Survival in sinoatrial disorder (sick sinus syndrome). *British Medical Journal* **280**, 139–41.

Viskin S (1999). Long QT syndromes and *torsade de pointes*. *Lancet* **354**, 1625–33.

Wijffels MCEF, *et al*. (1995). Atrial fibrillation begets atrial fibrillation: a study in awake chronically instrumented goats. *Circulation* **92**, 1954–68.

Zipes DP, Wellens HJJ (1998) Sudden cardiac death. *Circulation* **98**, 2334–51.

15.7 Valve disease

D. G. Gibson

Mitral stenosis

Aetiology

Chronic rheumatic heart disease is much the commonest cause of mitral stenosis, though there are a number of other well defined conditions in which blood flow across the mitral valve is limited to a variable extent.

1. Congenital mitral stenosis is a rare condition with thick, rolled leaflets and short chordae, with the spaces between them obliterated. The papillary muscles may be abnormally inserted, either directly from the free wall of the ventricle or from the septum. In parachute mitral valve, there is only one papillary muscle. Congenital mitral stenosis may be associated with left ventricular outflow obstruction, hypoplasia of the left ventricular cavity and the aorta, or endocardial fibroelastosis.
2. A calcified mitral valve ring may rarely cause mild mitral stenosis.
3. In infective endocarditis, bulky vegetations may occasionally interfere with transmitral flow.
4. Nodular rheumatoid arthritis may be associated with thickening of the valve cusps, but true mitral stenosis does not occur.
5. In systemic lupus erythematosus, treatment of Libman–Sachs endocarditis with steroids has led to fibrosis of the cusps with commissural fusion.
6. The combination of ostium secundum atrial septal defect and rheumatic mitral stenosis, Lutembacher syndrome, is probably fortuitous.

Rheumatic mitral stenosis

Incidence

The incidence of rheumatic mitral stenosis parallels that of acute rheumatic fever (see Chapter 15.10.1). It is thus much commoner, and presents earlier, in the Middle East, the Indian sub-continent, and the Far East than in the West.

Pathology

Rheumatic mitral stenosis is due to distortion of the normal mitral valve anatomy with fusion of the commissures. The cusps themselves become vascularized, thickened, and frequently develop thrombus on their atrial surfaces. The chordae become thickened and fused, and the papillary muscles scarred. Finally, the cusps may become calcified. The left ventricle is usually normal or small in pure mitral stenosis, but occasionally dilates. The left atrium is characteristically enlarged with scarring and disruption of muscle fibres. Mural thrombosis is common, particularly on the free wall just above the posterior mitral valve cusp (McCallum's patch). In long-standing cases, calcification of the left atrial wall may develop in plaques on its endocardial surface. In the lungs, the changes of pulmonary venous congestion, pulmonary hypertension, and haemosiderosis develop. These lead to dilatation and hypertrophy of the right ventricle with functional tricuspid regurgitation.

Pathophysiology

The main disturbance in mitral stenosis is due to left ventricular filling. When mitral valve area falls to around 2.5 cm^2, peak early diastolic ventricular filling rate falls and diastasis is lost. This does not matter at rest when the heart rate is slow and filling period relatively long, but during exercise as the heart rate increases, flow is maintained only by a pressure drop between atrium and ventricle. With a smaller valve area, a pressure drop is present at rest, and mean left atrial pressure rises. Patients with symptomatic mitral stenosis have a valve area of 0.75 to 1.25 cm^2, and a pressure drop as high as 20 to 30 mmHg across the valve during diastole. Cardiac output falls and pulmonary vascular resistance usually increases.

The subvalvular apparatus may interfere with left ventricular filling by restricting wall movement, so reducing stroke volume and increasing left atrial pressure in the absence of any diastolic pressure drop across the valve itself. Left ventricular cavity size, usually normal in young patients, may increase in the middle aged or elderly, and end-diastolic pressure may rise. A number of factors contribute to such left ventricular disease, including restriction of filling, coronary emboli, and distortion of the septum by right ventricular hypertrophy and overload. In addition, disturbed filling interferes in some way with systolic function, since after successful surgery cavity size usually falls, particularly at end-systole, as stroke volume increases.

Chronic left atrial hypertension causes a corresponding rise in pulmonary capillary pressure; clinical evidence of pulmonary congestion appears when it reaches around 25 mmHg. Further lung disease may be caused by active pulmonary hypertension, repeated pulmonary emboli or chest infections, haemosiderosis, or even bone formation.

Clinical features

Symptoms

The symptoms of mitral stenosis usually appear insidiously, and may have been present for several years before the patient seeks medical attention. They may be apparent within 3 or 4 years of the attack of acute rheumatic fever, or be delayed by up to 50 years. Less frequently, their onset is abrupt with an attack of acute pulmonary oedema, systemic embolism, or the onset of atrial fibrillation.

The commonest manifestation of mitral stenosis is a reduction in exercise tolerance by breathlessness, or less frequently, by fatigue or palpitation in patients in atrial fibrillation. Later in the disease, nocturnal dyspnoea occurs, though florid acute pulmonary oedema has become uncommon. Recurrent chest infections or winter bronchitis are very characteristic; the resulting infected pulmonary oedema responds poorly to antibiotics, and often leads to fluid retention and haemoptysis. Occasionally, massive or recurrent haemoptysis may be the presenting or only symptom of mitral stenosis. Systemic embolism from the left atrium is common in untreated mitral stenosis, particularly when atrial fibrillation is present. Any organ may be affected, but the commonest sites are cerebral, coronary, splenic, renal, mesenteric, or the arteries of the limbs. Salt and water retention is common in untreated mitral stenosis, and leads to peripheral oedema, ascites, pulmonary oedema, and pleural effusion.

Physical examination

Prolonged low cardiac output leads to weight loss and a malar flush.

The character of the pulse is normal, although its amplitude may be decreased and the rhythm irregular due to atrial fibrillation. The arterial pulses should always be checked in view of the possibility of previous arterial emboli.

The venous pressure is usually normal unless tricuspid regurgitation is present. An 'a' wave in the venous pulse of a patient with what appears to be pure mitral stenosis should always raise the possibility of additional tricuspid stenosis or severe pulmonary hypertension.

Palpation of the precordium at the apex may reveal a palpable first sound, previously called a 'tapping apex', and less frequently, a palpable opening snap. It may also be possible to feel pulmonary valve closure at the base of the heart if severe pulmonary hypertension is present. A left parasternal heave is usually due to right ventricular hypertrophy caused by pulmonary hypertension, but may also be due to tricuspid regurgitation, or increased prominence of a normal right ventricle secondary to an enlarged left atrium. In pure mitral stenosis, a sustained apex beat is unusual, but may be seen when the right ventricle is very considerably enlarged or more commonly, because of coexistent left ventricular disease.

On auscultation at the apex, the classic findings are a loud first sound, preceded by a presystolic murmur if the patient is in sinus rhythm, an opening snap, and a mid-diastolic murmur. A loud first sound is less specific for rheumatic mitral stenosis than a palpable one, since it also occurs in high cardiac output states, such as hyperthyroidism. A soft or absent first sound in mitral stenosis strongly suggests that the anterior cusp of the mitral valve is calcified or immobile. An opening snap is a very characteristic physical sign. It is usually loudest at the lower left sternal edge, less commonly the apex or the base. It is a sign of a pliable anterior cusp, and is absent if the valve structure is severely disorganized. The mid-diastolic murmur starts after the opening snap; it is low pitched and persists for a variable period throughout diastole. If the mitral stenosis is mild, the murmur is short, but if the murmur lasts throughout diastole at a normal ventricular rate, then the degree of stenosis is likely to be at least moderately severe. When the rate is rapid due to atrial fibrillation, the murmur may no longer be audible, although in these circumstances, the diagnosis can be suspected from the palpable first sound. However, there are some patients in whom no mid-diastolic murmur is audible even when the heart rate is controlled: so-called 'silent' mitral stenosis. These patients frequently either have severe pulmonary hypertension or a very disorganized and immobile mitral valve.

Investigations

Chest radiography (Fig. 1)

Fig. 1 Chest radiograph from a patient with pure mitral stenosis. Heart size is normal, but the left atrial appendage is enlarged. The upper lobe vessels are dilated and there are Kerley lines at both bases.

Heart size may be normal or increased. Selective enlargement of the left atrium is the commonest radiographic abnormality, 'selective' implying that the degree of enlargement is proportionately greater than that of the heart shadow as a whole. It appears on the penetrated posteroanterior film as a double outline on the right side of the heart shadow, with elevation of the left main bronchus, and enlargement of the left atrial appendix which forms that part of the left heart border just below the main pulmonary artery. Mitral valve calcification may be visible on the posteroanterior film just to the left of the spine.

In the lung fields, the upper lobe veins may be dilated with the patient in the erect position, indicating that left atrial pressure is raised. The size of the main pulmonary

artery can be increased due to pulmonary hypertension. Upper lobe blood diversion occurs when pulmonary vascular resistance is greatly increased: this can be recognized from decreased prominence of the vessels to the lower zones, while those to the upper zones are normal or increased. Pulmonary oedema occurs when the left atrial pressure reaches approximately 25 to 30 mmHg. It gives rise to lymphatic (Kerley B) lines in the lower zones, basal pleural effusions, generalized hazy shadowing, and finally, obvious interstitial oedema. Pulmonary haemosiderosis is due to long-standing pulmonary congestion. Bone formation, appearing as dense nodules of a few millimetres in diameter, is rarely seen.

ECG

The ECG is not very informative in mitral stenosis. Atrial fibrillation can be confirmed. If the patient is in sinus rhythm, then left atrial hypertrophy causes a bifid P wave in lead II and a dominant negative deflection in V1. ECG evidence of right atrial hypertrophy suggests tricuspid stenosis in addition to mitral stenosis. The electrical axis is usually vertical: right ventricular hypertrophy, if severe, is shown by a dominant R wave in V1.

Echocardiography

On M-mode echocardiography (Fig. 2) the mid-diastolic closure rate of the anterior cusp of the mitral valve is less than 50 mm/s in mild mitral stenosis and 0 to 20 mm/s in severe disease. Cusp fusion causes forward rather than backward movement of the posterior cusp during diastole.

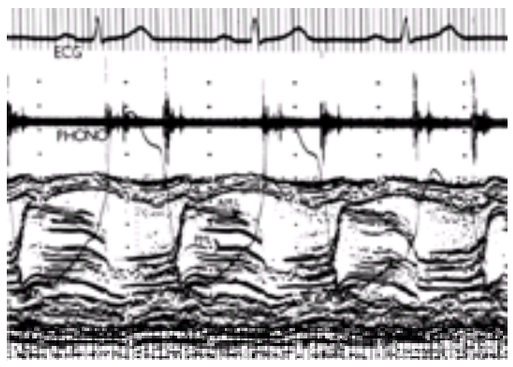

Fig. 2 M-mode echocardiogram from a patient with mitral stenosis. The anterior cusp (AML) is thickened, and its diastolic closure rate is reduced. The posterior leaflet (PML) moves forward during diastole. There is an opening snap on the phonocardiogram, coinciding with maximum forward motion of the anterior cusp.

On cross-sectional echocardiography the mobility of the anterior cusp is reduced, particularly near its tip (Fig. 3(a)). Valve area can be estimated semiquantitatively from the parasternal minor axis view (Fig. 3(b)), provided that the cusps are not calcified. The degree of subvalve involvement can be assessed and occasionally atrial thrombus can be detected (Fig. 4).

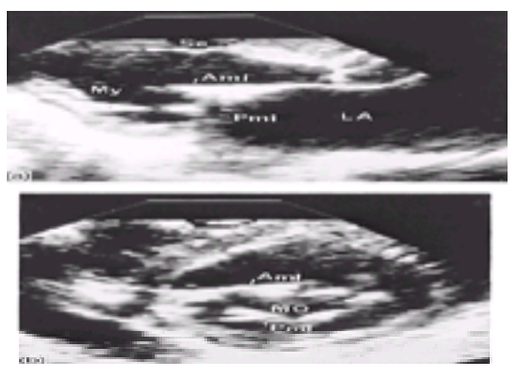

Fig. 3 (a) Two-dimensional echocardiogram from a patient with mitral valve disease; parasternal long-axis view taken in mid-diastole. The anterior cusp (Aml) of the mitral valve is thickened, and fails to open normally. LA, left atrium; Se, septum; My, myocardial echoes; whose intensity is increased due to scarring of the subvalve apparatus; Pml, posterior mitral leaflet. (b) Two-dimensional echocardiogram. Rheumatic mitral valve disease, parasternal minor-axis view during mid-diastole, at the level of the mitral valve orifice (MO), anterior cusp of the mitral valve (AML), and posterior mitral leaflet (Pml).

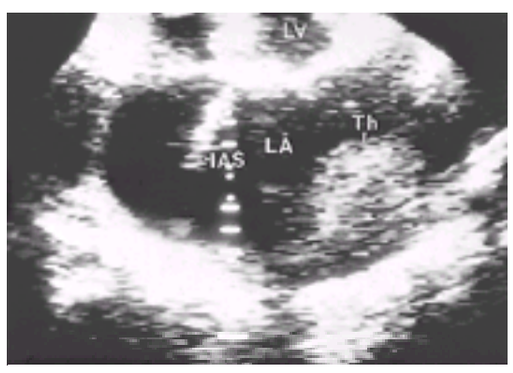

Fig. 4 Left atrial thrombus (Th) in a patient with mitral valve disease. LA, left atrium; LV, left ventricle; IAS, interatrial septum. The left atrium is considerably enlarged.

The diastolic pressure drop across the valve can be estimated by cross-continuous-wave Doppler. Peak right ventricular pressure is derived from the systolic velocity of tricuspid regurgitation. The aortic and tricuspid valves can also be checked.

Transoesophageal echocardiography is particularly useful for demonstrating thrombus in the body of the left atrium or in the left atrial appendix. Spontaneous contrast within the left atrial cavity is probably due to stasis resulting from a combination of atrial fibrillation, low forward flow, and increased cavity size. It indicates an increased risk of thrombus formation. Finally, the degree of thickening and calcification of the cusps and the extent to which the subvalve apparatus is involved can be assessed particularly well by this approach.

Cardiac catheterization

This is rarely necessary, either to make the diagnosis or to assess severity. It is performed only to determine the state of the coronary arteries in older patients and as a prelude to balloon valvuloplasty, or very occasionally in patients in whom diagnostic echocardiograms cannot be obtained.

Diagnosis

The diagnosis of mitral stenosis is usually straightforward on the basis of history, physical signs, and chest radiography, and can rapidly be confirmed by

echocardiography. When the ventricular rate is rapid, the diastolic murmur may be inaudible, but becomes apparent when the ventricular rate is controlled by digoxin. Silent mitral stenosis may mimic primary pulmonary hypertension, but the correct diagnosis is easily made by echocardiography. Mild mitral stenosis should be suspected as a source of systemic emboli and as a cause of unexplained atrial fibrillation, particularly in the elderly.

Differential diagnosis

1. Left atrial myxoma (see Chapter 15.11.1);
2. cor triatriatum (see Chapter 15.13);
3. pulmonary veno-occlusive disease (see Section 15.15.2); or
4. Austin–Flint murmur (see Aortic regurgitation section, below).

Treatment

Medical

In patients under 30 years of age, longer in some cases, penicillin prophylaxis against further attacks of acute rheumatic fever should be given (see Chapter 15.10.1 for further discussion).

Atrial fibrillation should be treated with a digitalis preparation to control ventricular rate. Anticoagulant therapy must be given to reduce the risk of systemic embolism to all patients with atrial fibrillation, unless there are very strong contraindications. It is also advisable to give anticoagulants to patients in sinus rhythm with mitral stenosis, particularly the middle aged and elderly: the incidence of embolism is significant, especially if they go into atrial fibrillation. The risk of embolism is particularly high when a patient with atrial fibrillation not receiving anticoagulants is admitted to hospital with a rapid heart rate and pulmonary oedema. In this situation intravenous heparin should be given until therapeutic anticoagulation with an oral agent is established.

Fluid retention associated with mitral stenosis responds well to treatment with diuretics. Chest infections should be treated promptly with appropriate antibiotics, and patients should be given a supply of antibiotic to take prophylactically at the start of a head cold. A diuretic is also useful because chest infections often precipitate, or may be precipitated by, fluid retention.

In all patients with valvular heart disease, prophylactic antibiotics should be given for all dental manipulations and other potentially septic hazards (see Chapter 15.10.2).

Mitral valvuloplasty

Rheumatic mitral stenosis results from fusion of the commissures between the two mitral cusps, which is susceptible to rupture by inflating a catheter-mounted balloon across the valve orifice (Fig. 5). Mitral valvuloplasty has the great advantage over surgery of avoiding thoracotomy. A catheter is introduced through the inferior vena cava to the right atrium. The atrial septum is crossed, and the catheter stabilized across the mitral valve, usually by a guidewire passed out through the aortic valve. The balloon itself may be single, often with a waist; less commonly, two balloons are placed simultaneously across the orifice. The balloon is inflated to a predetermined size for 20 s.

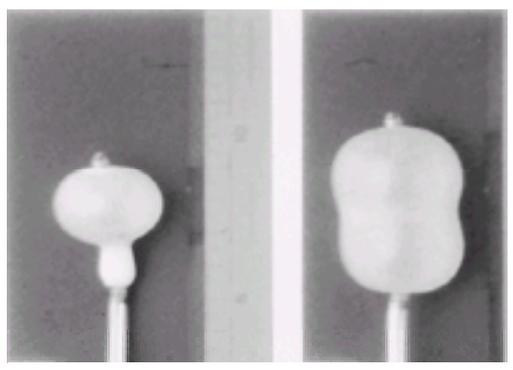

Fig. 5 The Inoue balloon catheter, as used for mitral valvuloplasty, partially (left) and completely (right) inflated.

Not all patients are suitable for valvuloplasty. It is unsatisfactory in those with calcified and immobile valve cusps, who constitute the majority with mitral stenosis in the developed world. There should be no more than minimal regurgitation. Ideally, the cusps should be thin and mobile, without calcification in the commissures, shortening of the chordae, or scarring of the papillary muscles. Clot in the left atrial appendix must have been excluded by transoesophageal echocardiography. In appropriately selected patients, valvuloplasty gives a satisfactory fall in transmitral pressure drop, maintained in the short and medium term. Mitral regurgitation may be provoked, sometimes severe enough to require valve replacement on an elective or even an emergency basis. The majority of patients are left with a small ASD at the site of passage of the catheter, but this is not of any haemodynamic significance.

Mitral surgery

Surgical procedures available include mitral valvotomy, open or closed, and mitral valve replacement. The choice of operation depends on the anatomy of the mitral valve determined on the basis of the physical signs and the echocardiogram, the age of the patient, and the surgical resources available.

Closed mitral valvotomy is a relatively simple procedure in terms of the resources that are required, although a satisfactory result presupposes considerable experience with the operation, experience that is now becoming rare in the developed world. It is particularly suitable in a Third World country, where the major radiographic and expensive disposables necessary for balloon valvuloplasty are not available. It is most effective in a young patient, in sinus rhythm, with evidence of a mobile anterior cusp. It can be regarded as a form of beating heart surgery, and when performed with intraoperative transoesophageal echocardiographic monitoring is particularly effective. Symptom-free follow-up of 40 years or more regularly occurs after this procedure.

Open valvotomy requires cardiopulmonary bypass but allows a more complete procedure to be undertaken, and in particular the subvalvular apparatus can be inspected and adherent chordae divided. If the results of valvotomy are found to be unsatisfactory, it is possible to proceed to valve replacement at the same operation. In general, significant mitral regurgitation is a contraindication to a conservative procedure; although the early results of repair may be excellent, replacement is usually necessary within 5 years.

Valve replacement is necessary when the valve cusps are greatly thickened or calcified. This operation should not be considered in patents in whom the haemodynamic disturbance caused by the valve disease is mild, since the prosthesis causes a resting diastolic pressure drop across it, as well as interfering with systolic and diastolic left ventricular function.

Indications for interventional procedures

It is difficult to lay down hard and fast indications for intervention in patients with mitral stenosis. If the clinical and echocardiographic evidence suggests that valvuloplasty is feasible, then the presence of definite limitation of exercise tolerance is an adequate indication, particularly in a young person. It may also be used in patients with asymptomatic but well developed mitral stenosis before pregnancy. Open valvotomy should be considered if there is any significant contraindication to valvuloplasty. Closed valvotomy is unfortunately no longer available in most developed countries, but it remains a most attractive possibility where medical and surgical resources are limited. In avoiding the use of either radiographic screening or cardiopulmonary bypass, it may be an attractive option for the patient who

develops acute pulmonary oedema during pregnancy, provided that local surgical expertise is available.

If valve replacement is likely to be required, then limitation of exercise tolerance should be more severe. In individual patients the decision is not usually difficult when there has been definite progression of symptoms. It is not often necessary to advise operation in a patient with normal exercise tolerance. Unless it is due to coronary artery disease, the presence of left ventricular dilatation is not a contraindication to operation, however severe it may appear to be in terms of increased cavity size or reduced amplitude of wall motion. The same applies to pulmonary hypertension, which is not a contraindication, since it increases the benefits of operation to a greater extent than the risks.

Prognosis

In the absence of surgical treatment, mitral stenosis is usually a progressive disease, although the rate is unpredictable. Unfavourable features include a gradual increase in the severity of the valve disease with disorganization of its structure and superimposed calcification, an increase in pulmonary resistance, and the development of functional tricuspid valve disease, with chronic elevation of the venous pressure leading to cardiac cirrhosis and impaired liver function.

Surgical treatment has considerably improved the prognosis, although conservative mitral surgery does not prevent progression of the rheumatic process, nor does it reduce the risk of infective endocarditis. It has also become clear that the life of biological mitral valve substitutes, particularly the porcine xenograft, is limited to no more than 10 years in the majority of patients above the age of 21, and considerably less than this in children. These valves should thus be confined to the very elderly, and to young women who wish to undertake pregnancy, knowing that repeat surgery will be needed. There are minor differences in haemodynamics between the different types of mechanical valve substitutes, but in individual cases, these are of little consequence.

Mixed mitral valve disease

Mixed mitral valve disease is nearly always rheumatic in origin. The mitral regurgitation is not usually severe in terms of the volume load that it imposes on the left ventricle, though the increased stroke volume increases the diastolic pressure drop across the valve. In general, it occurs in older patients than pure mitral stenosis, and the valve is more disorganized. It is more likely to be calcified with limited cusp mobility and scarred subvalve apparatus. The symptoms are similar. On examination, a pansystolic murmur is evident along with the mid-diastolic murmur. The first sound is not palpable or accentuated. The pansystolic murmur is usually loudest towards the axilla, reflecting the frequent scarring and retraction of the posterior cusp. Chest radiography (Fig. 6) may show more advanced changes than in pure mitral stenosis, and in particular, the left atrium may be very large indeed. ECG is unhelpful. Echocardiogram is likely to show thickened cusps whose motion is reduced as well as mitral regurgitation. Valvuloplasty or conservative surgery are both unsatisfactory, and when symptoms merit, mitral replacement is usually required.

Fig. 6 Chest radiograph from a patient with mixed mitral valve disease, showing gross cardiac enlargement, due mainly to dilation of the left atrium.

Mitral regurgitation

Aetiology

There are many causes of mitral regurgitation. Any of the components of the mitral valve apparatus may be involved (Table 1).

Degenerative mitral valve disease is the commonest condition to affect the cusps. It has been described under a number of other names, based either on its pathology or on its clinical features: degenerative mitral valve disease, mucinous or myxomatous degeneration, or floppy or ballooning mitral valve. It is a non-inflammatory process partially or completely affecting either cusp. Cusp area is increased, causing folding and upward doming into the left atrium during systole. The chordae may become elongated, tortuous, and thinned, predisposing to chordal rupture. Ulceration of the cusps may predispose to thrombosis on their surface and infective endocarditis. Ring circumference may increase. The papillary muscles are normal. Histologically, the centre of the cusp—the fibrosa—is abnormal, with large areas in which collagen bundles are fragmented or absent altogether, and a dense layer of laminated collagen forms on the atrial surface. There is no evidence of vascularization or of inflammatory cells in the absence of secondary infective endocarditis.

The cause of sporadic cases of floppy mitral valve is unknown. However, similar appearances may complicate Marfan's syndrome, pseudoxanthoma elasticum, Ehlers–Danos syndrome, and osteogenesis imperfecta. The incidence of the sporadic condition tends to rise with age, and individual case histories suggest that it can be a very benign and chronic process.

Infective endocarditis is a major cause of symptomatic mitral regurgitation (see Chapter 15.10.2).

Systemic lupus erythematosus can affect both mitral and aortic valves, causing thickening of the cusps and the appearance of sterile vegetations (Libman–Sachs endocarditis). Their appearance and severity fluctuates in individual patients, and does not correlate with other markers of activity of the underlying disease. They rarely give rise to significant haemodynamic disturbance, but may predispose to emboli and to infective endocarditis.

Pathophysiology

Pure mitral regurgitation increases left ventricular output. Since the pressure in the left atrium is lower than that in the aorta, the net force opposing left ventricular ejection is reduced, and stroke volume may be up to three times normal. Ejection begins almost immediately after the start of left ventricular contraction, and by the time the aortic valve opens, up to one-quarter of the stroke volume may already have entered the left atrium. Left atrial pressures are therefore increased, with the V or systolic wave sometimes reaching 50 to 60 mmHg. These high pressures shorten the phase of isovolumic relaxation and greatly increase the velocity of early diastolic left ventricular filling, thus causing the third heart sound. When mitral regurgitation is very severe indeed, left ventricular and left atrial pressure may equalize at mid-ejection. Left ventricular end-diastolic cavity size is not greatly increased, particularly when the history is short, but end-systolic size is considerably smaller than normal due to the low force opposing ejection. Resting left ventricular output is maintained by a sinus tachycardia that is nearly always present when mitral regurgitation is severe.

Clinical features

The clinical picture of pure mitral regurgitation depends on the underlying pathology, the severity of regurgitation, and whether or not the left ventricle is diseased. Different clinical patterns will be described separately, recognizing that they overlap and that the relation between the clinical picture and the underlying aetiology is not fixed.

Ruptured chordae tendineae

Ruptured chorda is a complication of degenerative mitral valve disease and often causes severe mitral regurgitation. It usually occurs spontaneously but may be caused by infective endocarditis. A murmur may have been heard in the past, often many years previously, and described at the time as 'innocent' or 'benign'. The onset of symptoms is usually gradual, but in a minority may be so sudden that patients are able to describe exactly what they were doing at the time. In such cases the symptoms are most severe at their onset, improving over the next few weeks as the ventricle adapts to the volume load. However, even in this more compensated phase, exercise tolerance may be severely limited by breathlessness or fatigue. When the regurgitation is only moderately severe, it can be tolerated remarkably well for many years with minimal symptoms. However, the most severe cases can present in intractable pulmonary oedema, requiring immediate intermittent positive-pressure ventilation.

Physical examination

Patients are usually in sinus rhythm until late in the course of the disease when mitral regurgitation is non-rheumatic. Sinus tachycardia is frequent, and the pulse 'jerky', implying that its amplitude is normal although the upstroke is rapid. The venous pressure is normal unless severe pulmonary hypertension or associated tricuspid regurgitation is present.

The precordial impulse at the apex is prominent and sustained, and may be double due to a palpable third sound. A systolic thrill may also be present. A left parasternal heave reflects systolic expansion of the left atrium rather than right ventricular hypertrophy.

On auscultation, the first sound is normal or reduced in intensity. The most prominent findings are a loud pansystolic murmur and a third heart sound. The third sound may be rather more high-pitched than that associated with left ventricular disease, reflecting the high early diastolic inflow velocity, and may be confused with the second sound. The murmur may thus be mistimed. This mistake can be avoided by starting auscultation at the base of the heart, where the true second sound can be appreciated, and 'inching' the stethoscope towards the apex, when the second heart sound can be heard to bury itself in the murmur as the third sound appears. If the mitral regurgitation is so severe as to cause left atrial and left ventricular pressures to equalize before the end of systole, the murmur stops early. In the most severe cases, presenting with acute pulmonary oedema and shock, the mitral valve is effectively absent and there may be no murmur at all. Unlike rheumatic mitral regurgitation, the position at which the amplitude of the murmur appears maximal is variable: it may be at the apex, down the left sternal edge, at the back, to the left of the spine, or even the top of the head.

Investigations

Chest radiography

The radiographic picture reflects the haemodynamic disturbance (Fig. 7). Overall heart size is normal or only moderately enlarged, with selective enlargement of the left atrium, though not to the same extent as in rheumatic mitral valve disease. The pulmonary vasculature reflects the increase in mean left atrial pressure. A chest radiograph taken soon after the onset of severe mitral regurgitation may show pulmonary oedema with a normal-sized heart. If the condition is severe and long-standing, considerable cardiac enlargement develops due to secondary left ventricular disease.

Fig. 7 Chest radiograph showing acute pulmonary oedema due to acute mitral regurgitation resulting from ruptured chordae tendineae.

ECG

The ECG usually shows sinus rhythm with only moderate left ventricular hypertrophy. There may, in addition, be evidence of left atrial hypertrophy. Frequent ventricular ectopic beats are characteristic of mild or moderate mitral regurgitation.

Echocardiography

M-mode echocardiography may show cusp prolapse, with cusp remnants visible in the left atrium during systole. The amplitude of left ventricular wall motion is increased. Initially the end-systolic cavity dimension is small, but it gets larger as the left ventricle adapts, increasing progressively when irreversible left ventricular disease supervenes.

Cross-sectional echocardiography confirms the presence of very active left ventricular wall motion. It allows a clearer view of the extent of systolic cusp prolapse into the left atrium, and the affected cusp is identified more reliably.

Continuous wave Doppler confirms the presence and timing of regurgitation, (Fig. 8) and the jet can be mapped within the left atrium by colour flow. Apparent jet area, whether or not normalized to left atrial cavity size, has proved a disappointing measure of the severity of the regurgitation.

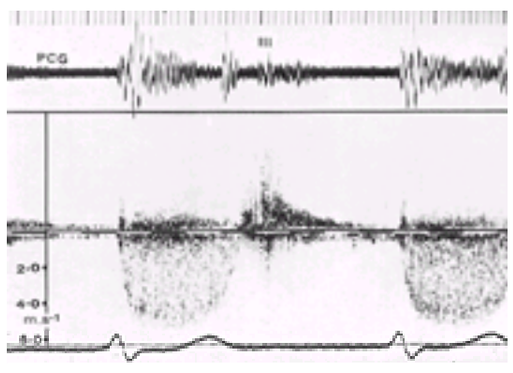

Fig. 8 Doppler cardiogram from a patient with pure mitral regurgitation, showing regurgitant flow as a downward deflection during systole. Diastolic flow velocity pattern is normal, with peak velocity coinciding with the third heart sound (III) on the phonocardiogram (PCG).

Transoesophageal echo may give more information about valve anatomy, allowing ruptured chordae and small vegetations to be seen in detail. Severe regurgitation causes retrograde flow in the pulmonary veins.

Cardiac catheterization

This is not usually necessary to make the diagnosis when the clinical features and echocardiography are typical, though many surgeons require views of the coronary arteries in older patients when planning operation.

Functional mitral regurgitation

Normal mitral closure depends on the integrity of the myocardium as well as that of the valve apparatus itself. In part, the position of the cusps during systole is maintained by contraction of the papillary muscles as the left ventricular cavity gets smaller. This mechanism can be disturbed in a number of ways. The papillary muscles themselves may be affected by ischaemic or other left ventricular disease, so that their ability to contract is impaired. If left ventricular cavity size is greatly increased, the relation between wall movement and papillary muscle shortening becomes abnormal. In hypertrophic cardiomyopathy (see Chapter 15.8.2), the greatly hypertrophied papillary muscles and abnormal cavity shape may contribute to the characteristic forward movement of the whole mitral valve apparatus during systole. Loss of support for the mitral ring itself may occur due to impairment of the function of circumferentially arranged myocardium at the base of the heart.

The term functional mitral regurgitation thus represents the combination of the regurgitation itself, a structurally intact valve apparatus, and left ventricular disease, usually cavity dilatation. It may also be referred to as papillary muscle dysfunction, although this mechanism has not been confirmed directly in humans.

The mitral regurgitation itself is usually mild, though in a minority it may be as severe as that due to ruptured chorda. However, even when mild it may last more than 500 ms, particularly when left bundle branch block is present, so that it limits the time available for ventricular filling when the heart rate is rapid. The clinical picture is therefore usually dominated by impaired left ventricular function. The presence of mitral regurgitation is demonstrated by either a late or a pansystolic murmur, which often varies in its intensity and timing from day to day, and which becomes softer with successful treatment as cavity size and left ventricular diastolic pressures fall.

Echocardiography shows a large cavity with poor wall movement, quite different from the picture seen in severe organic mitral regurgitation. The mitral regurgitation itself can be detected by continuous wave and colour flow Doppler. Cardiac catheterization confirms the presence of a raised left atrial pressure, secondary to a corresponding elevation of the left ventricular end-diastolic pressure. Left ventricular angiography shows a dilated and poorly functioning left ventricle with reflux of contrast into the left atrium, where it tends to accumulate due to poor forward flow. Coronary artery disease as the underlying cause can only be confirmed or excluded by coronary arteriography.

Ruptured papillary muscle

This is a rare and catastrophic complication of acute myocardial infarction, and is quite different from chordal rupture. Papillary muscle rupture is usually complete, though less commonly a single head may be involved. Rupture usually occurs 2 to 5 days after the infarct, and is rarely associated with survival for more than 24 h without very prompt surgical intervention. Whether partial or complete, papillary muscle rupture causes very severe mitral regurgitation, occurring on top of left ventricular impairment caused by the infarct itself. A pansystolic murmur may sometimes be audible at the apex. Death is due to cardiogenic shock and pulmonary oedema.

Partial rupture occurs rather later after the infarct than complete rupture, but similarly causes a striking deterioration in clinical state, along with the development of a pansystolic murmur. The posteromedial papillary muscle is involved more frequently than the anterolateral, both by partial and complete rupture. Since patients are likely to be ventilated, papillary muscle rupture is best diagnosed by transoesophageal echo (Fig. 9), which shows a very active left ventricle and an abnormally mobile mitral valve. The condition requires emergency mitral valve replacement, but even when this can be achieved the prognosis is much worse than that after chordal rupture due to associated left ventricular disease.

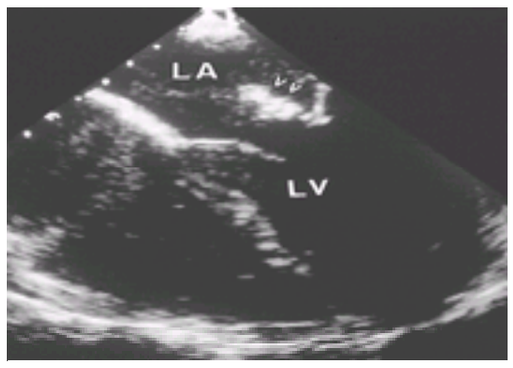

Fig. 9 Transoesophageal cross-section echocardiogram from a patient presenting with ruptured papillary muscle 48 h after acute myocardial infarction. Note the abnormal mobility of the papillary muscle head (indicated by arrows) attached to the anterior mitral valve cusp. LA, left atrium; LV, left ventricle.

Mitral prolapse

Mitral prolapse consists of systolic displacement of one or both mitral valve cusps into the left atrial cavity by 2 mm or more from the line joining the hinge points of the cusps as shown by cross-sectional echocardiography. There may or may not be associated thickening of the cusps themselves. The mid-systolic click occurs as the valve cusps move abruptly backwards into the left atrium during systole (Fig. 10).

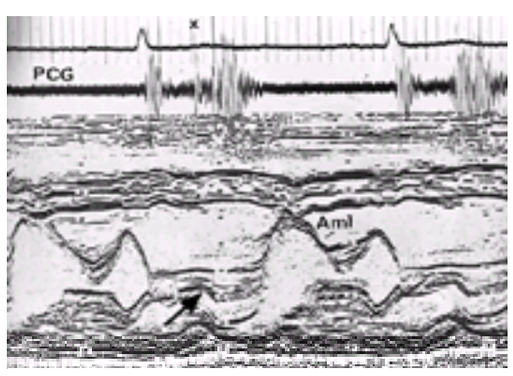

Fig. 10 Mitral valve prolapse, M-mode echocardiogram. Mid-systolic prolapse occurs, marked by the arrow. This is associated with a mid-systolic click (x) and late systolic murmur on the phonocardiogram (PCG).

The incidence of mitral prolapse, diagnosed on these strict criteria, is less than 2 to 3 per cent in a normal population. In the past, when the diagnosis was made on M-mode alone, or other less specific criteria, a much higher incidence was claimed, approaching 25 per cent in some populations. The majority of well documented

cases occur on the basis of mild degenerative mitral valve disease.

From the point of view of clinical manifestations, the main determining factor is the extent to which the margins of the cusps remain apposed during systole, that is, on the extent of secondary mitral regurgitation. It is usually insignificant and unrecognized until a mid-systolic click or late systolic murmur are heard at routine examination. Indeed, in the absence of both the diagnosis may be questioned.

In the past, it was thought that the condition was associated with non-specific chest pain, potentially life-threatening arrhythmias, and cerebral embolism. There is no good evidence for any of these clinical associations. However, mitral prolapse does predispose to infective endocarditis, so that antibiotic prophylaxis is essential. In addition, over the long term, simple mitral prolapse may gradually progress to more severe mitral regurgitation, either by chordal stretching or rupture, particularly when there is evidence of cusp thickening. This type of mitral valve disease is common in Marfan's syndrome and other connective tissue diseases.

Endomyocardial fibrosis

Endomyocardial fibrosis causes fibrosis of the endocardium and underlying myocardium of either or both ventricles. It also involves the papillary muscles, and thus causes secondary mitral regurgitation. It is common in Uganda and surrounding countries in East Africa, where it accounts for approximately 10 per cent of hospital admissions with heart disease, and also in Nigeria in West Africa. It occurs less commonly in south India and Sri Lanka. Occasionally, it is seen in European individuals who have lived in affected areas and very rarely in those who have never been to the tropics.

When the right ventricle is involved, fibrosis starts at the apex and spreads upwards towards the tricuspid valve, involving the papillary muscles and chordae, but sparing the outflow tract. In the left ventricle, the inflow tract, apex, and lower part of the outflow tract are usually involved, and also the posterior mitral valve cusp and its papillary muscle. In both ventricles there is involvement of the underlying myocardium and mural thrombosis. The result is atrioventricular valve regurgitation, an abnormally stiff ventricle, and very high atrial pressures.

The aetiology is unknown, but it does not appear to be related to rheumatic fever or any vector-borne virus, although the incidence increases with malnutrition.

The clinical picture is of progressive mitral or tricuspid insufficiency of insidious onset, together with restriction of ventricular filling by subendocardial scarring. When the tricuspid valve is mainly involved, there is gross fluid retention, whereas mitral or combined involvement leads to pulmonary oedema. Emboli from the right or left ventricle are common.

Medical treatment consists of high doses of diuretic and vasodilators, preferably an ACE inhibitor, if valvular regurgitation is severe. Decortication of the ventricular cavities may be possible surgically, along with replacement of mitral or tricuspid valve.

Mitral ring calcification

Heavy calcification of the mitral valve ring is a disease of the elderly, and is particularly common in women. Although it appears to be a degenerative condition, it occurs more frequently with left ventricular hypertrophy. It does not usually cause symptoms, and is detected incidentally by calcification in the mitral ring on chest radiography or on echocardiography. The central fibrous body may also be involved, with calcium spreading down the anterior cusp. However, the condition is not totally benign. The valve is a potential source of platelet emboli, and a focus for infective endocarditis. Approximately half the patients have abnormalities of conduction, including high-grade atrioventricular block, sinus node disease, or bundle branch block. Mild mitral regurgitation is common, but rarely is it severe enough to need valve replacement. Very occasionally the condition has been reported as causing mitral stenosis, with a diastolic pressure drop of up to 20 mmHg.

In the absence of complications, no treatment is required other than low-dose aspirin and prophylaxis against infective endocarditis. Complications are treated on their own merits.

Diagnosis of mitral regurgitation

The diagnosis of mitral regurgitation is usually straightforward on the basis of the physical signs, with a pansystolic murmur and third heart sound when it is haemodynamically significant, and a late systolic murmur when it is due to mitral prolapse and mild. Cardiac enlargement and pulmonary congestion on chest radiography reflect the extent both of the regurgitation itself and of associated left ventricular disease. Echocardiography is invaluable in assessing such cases. Abnormalities of mitral valve anatomy can be detected and associated abnormalities of left ventricular function quantified, whether due to volume overload or associated ventricular disease.

Difficulties in diagnosis may arise when mitral regurgitation is of acute origin and very severe. Patients may present with pulmonary oedema of sudden and unexplained onset with a chest radiograph showing a normal-sized heart shadow. Echocardiography demonstrates very active left ventricular wall movement, showing that the poor peripheral blood flow is due to valvular regurgitation rather than left ventricular disease, and in addition, one or both mitral valve cusps may be abnormally mobile, usually as the result of chordal or papillary muscle rupture. However, the Doppler echocardiogram may be atypical, showing an abbreviated regurgitant flow signal of low velocity reflecting near equalization of ventricular and atrial pressures. Such low blood velocities make colour flow Doppler particularly misleading.

In patients who present with more typical signs, the main diagnostic problem is to decide the relative contributions of valvular regurgitation and left ventricular disease to the overall clinical state. In such cases, intrinsic disease of the mitral valve apparatus suggests that ventricular disease is secondary to long-standing regurgitation, while normal mitral valve anatomy suggests the reverse.

Differential diagnosis

1. Ventricular septal defect (see Chapter 15.13).
2. Aortic valve disease—The ejection systolic murmur of aortic valve disease is frequently audible at the apex, where it may be louder than at the base, and have a slightly different quality. However, this is not an adequate basis for diagnosing additional mitral regurgitation and it is essential to establish that the timing of the murmur is pansystolic, either from its relation to the second heart sound, or in aortic regurgitation, from its relation to the start of the early diastolic murmur.
3. Tricuspid regurgitation—The pansystolic murmur of tricuspid regurgitation may be mistaken for that of mitral regurgitation, particularly when the right ventricle is greatly enlarged. The presence of tricuspid regurgitation can be suspected from an elevated venous pressure with systolic waves, and confirmed by Doppler echocardiography. In severe mitral regurgitation, however, additional tricuspid regurgitation may be present.

Treatment of mitral regurgitation

Mild or moderately severe mitral regurgitation is well tolerated and does not require treatment apart from prophylactic antibiotic for all dental manipulations and potentially septic hazards. If left ventricular size is increased either on chest radiograph or echocardiogram an ACE inhibitor should be added. Such patients should be followed up at annual intervals, since mitral regurgitation may be progressive, particularly when due to degenerative disease.

When mitral regurgitation is functional and mild, treatment is again medical and that of the underlying left ventricular disease. However, in a minority of patients with severe ischaemic ventricular disease, a more aggressive surgical approach may be warranted. Mitral regurgitation due to hypertrophic cardiomyopathy does not require specific treatment.

Severe mitral regurgitation, which causes significant symptoms in spite of medical treatment, is best managed by mitral valve surgery. This will involve either mitral valve replacement, or in suitable cases, mitral valve repair. The long-term prognosis following mitral repair is better than that after replacement, both with respect to survival and functional result. The exact reasons for the difference are not clear, but are probably because replacement involves insertion of a rigid mitral ring and section of the papillary muscles with very abnormal flow patterns into the ventricle. Mitral repair is thus to be preferred whenever possible, and in suitable cases can be recommended earlier in the course of the disease. It is particularly satisfactory with non-rheumatic regurgitation due to posterior cusp prolapse, and in an increasing number of patients with anterior cusp prolapse as surgical techniques develop.

The timing of operation is critical. After acute chordal rupture, it is often possible to treat the patients medically with rest, diuretics, and vasodilators for 1 to 2 weeks, while the left ventricle enlarges to compensate for the increased volume load. Clinical improvement may be striking, so that surgery becomes a less hazardous procedure than an emergency operation in the acute stage would have been. By contrast, those with a ruptured papillary muscle should undergo surgery at the earliest opportunity. Until this is possible, their pulmonary oedema is best treated by intermittent positive-pressure ventilation. Any benefit of pharmacological treatment or balloon counterpulsation is marginal at best, and probably only distracts attention from the main aim, which should be to get the patient to surgery as soon as possible.

The treatment of papillary muscle dysfunction is that of the underlying ventricular disease, with the particular aim of reducing the left ventricular diastolic pressures. In a minority of cases, hibernating myocardium will respond to vein grafting. Mitral valve replacement should be avoided whenever possible, since the deterioration in ventricular function caused by the valve replacement itself usually outweighs any benefit from correction of the regurgitation. A more promising approach is to insert an undersized mitral ring, leaving the mitral apparatus otherwise intact. This not only corrects the regurgitation but also may reduce basal left ventricular cavity size and thus systolic wall stress.

Aortic stenosis

Aortic stenosis represents a fixed obstruction to left ventricular ejection into the aorta. The obstruction is most commonly at the level of the valve itself, aortic valvar stenosis, but may also be immediately above the sinuses, supravalvar stenosis, or within the left ventricle, subvalvar stenosis.

Aetiology

Types of valvar aortic stenosis are summarized in Table 2. Valvar aortic stenosis is an important cause of cardiac disability, and though it is commonest in the elderly, it may present at any time of life. Congenital aortic stenosis, due to a valve with only a single commissure, is most frequent in infancy or childhood. Congenital bicuspid valve, consisting of fusion of one of the three commissures, is a much commoner abnormality. It may be detected as an incidental finding early in life, but does not usually give rise to significant haemodynamic abnormality unless it becomes calcified or involved by infective endocarditis. Rheumatic aortic stenosis develops as the result of commissural fusion in a tricuspid valve and may subsequently become calcified. Senile or degenerative aortic stenosis results from deposition of calcium in a tricuspid valve, initially on the aortic surface by a process similar to atherosclerosis. Calcification of a tricuspid valve is becoming an increasingly important cause of disability in the elderly. Very rarely, vegetations in infective endocarditis or lipid deposits in hyperlipidaemia may be bulky enough to cause significant left ventricular outflow tract obstruction.

Pathophysiology

Blood flow across a stenotic aortic valve causes a pressure drop between the left ventricular cavity and the aorta, which in symptomatic cases, may be greater than 60 mmHg at rest, and reach over 200 mmHg on exertion. Stroke work is therefore increased and left ventricular hypertrophy develops. Wall thickness increases although cavity size remains normal or even falls. A corresponding increase in the coronary vascular bed does not occur, predisposing to myocardial ischaemia, particularly in the subendocardial region. Hypertrophy, ischaemia, and associated fibrosis cause the diastolic stiffness of the myocardium to increase so that the end-diastolic pressure may rise causing pulmonary congestion. Increased left ventricular wall thickness also predisposes to ventricular arrhythmias. Late in the disease, when left ventricular involvement is severe, the cavity dilates and becomes more spherical. Calcification may spread from the aortic valve to the anterior cusp of the mitral valve or into the septum where it can involve the conducting system which runs nearby.

Aortic stenosis is most common in patients with a high incidence of ischaemic heart disease, so that obstructive coronary artery disease may contribute coincidentally to symptoms or impairment of left ventricular function.

Clinical features

Symptoms

The three characteristic clinical features of aortic stenosis are breathlessness, chest pain, and syncope.

Breathlessness in aortic stenosis is frequently associated with an elevated left ventricular end-diastolic pressure and occurs at first on exercise, but later at rest. Paroxysmal nocturnal dyspnoea and episodic pulmonary oedema—breathlessness persisting for 5 min or more after exercise—are both common in late stages of the disease and indicate the need for urgent treatment. The length of the history of breathlessness from its onset until it becomes severe is usually only of the order of 1 to 2 years, and thus considerably shorter than that in mitral stenosis.

Angina occurring in aortic stenosis is clinically indistinguishable from that due to coronary artery disease, which in many cases is the main cause. However, typical anginal pain can occur in patients in whom the large and medium-sized coronary arteries are normal. The mechanism for this is uncertain, but disproportion between muscle mass and coronary vascular bed, and the direct effects of abnormal myocardial relaxation in left ventricular hypertrophy are both likely to contribute.

There are several causes of syncope in aortic stenosis. In some patients it is clearly related to exertion and appears to be due to hypotension resulting from the combination of exercise-induced vasodilation and a fixed cardiac output. In other cases it results from transient complete atrioventricular block due to involvement of the atrioventricular node by calcification, carotid sinus hypersensitivity, or even from short periods of ventricular tachycardia or fibrillation.

Physical examination

The physical signs of well developed aortic stenosis are very characteristic.

1. The carotid pulse is slow rising with a reduced amplitude and an early notch on the upstroke, followed by a thrill.
2. The venous pressure is usually normal until late in the disease, but a small 'a' wave is frequently present. This cannot be taken as evidence of pulmonary hypertension, but appears to be related in some way to the presence of left ventricular hypertrophy (Bernheim 'a' wave).
3. The apex beat is sustained and is often double, due to an additional left atrial impulse.
4. On auscultation, the first sound is normal or soft, and may be preceded by a fourth heart sound. The second sound is single when the valve is calcified, due to lack of the aortic component. In younger patients with mobile valve cusps, aortic valve closure may be audible, but delayed, so that splitting of the second sound is reversed. When left ventricular disease is severe, pulmonary valve closure is accentuated. The characteristic ejection systolic murmur is maximal at the base of the heart, and is also audible over the right common carotid artery. It may seem longer than the ejection systolic murmur of, for example, anaemia or thyrotoxicosis because in aortic stenosis, ventricular systole is prolonged and aortic valve closure delayed. A soft early diastolic murmur is nearly always present, although this does not imply haemodynamically significant aortic regurgitation.

As ventricular disease progresses and stroke volume falls, these physical signs are modified. Pulse volume drops and loses its slow rising quality. The aortic murmur becomes shorter and softer, and a third heart sound and functional mitral regurgitation appear.

Investigations

Chest radiography

Heart size is normal in uncomplicated aortic stenosis. If it is increased, the underlying cause is likely to be unsuspected aortic regurgitation, left ventricular cavity dilatation, or very severe left ventricular hypertrophy, when the cavity may be normal in size, but the myocardium up to 50 mm thick. Increased left ventricular filling pressure may cause left atrial hypertension and thus dilatation of the upper lobe vessels as well as selective enlargement of the left atrium in the absence of organic mitral valve disease. The aortic root is nearly always dilated and the aortic valve calcified in older patients: this is best seen on the lateral chest radiograph or with screening.

ECG

The ECG characteristically shows changes of left ventricular hypertrophy, although it may be entirely normal, even in the presence of severe aortic stenosis. Left atrial hypertrophy is shown by a bifid P wave in lead II or a dominant negative deflection in V1. Conduction disturbances include left axis deviation, left bundle branch block, prolonged P–R interval, or complete heart block. Poor progression of R waves across the chest leads is common, and suggests septal hypertrophy rather than anterior myocardial infarction.

Echocardiography

Cross-sectional echocardiography demonstrates thickening and reduced mobility of the valve cusps. In young patients with a bicuspid valve, doming of the cusps during systole can be seen, while in older patients, a calcified aortic valve appears as an immobile mass (Fig. 11).

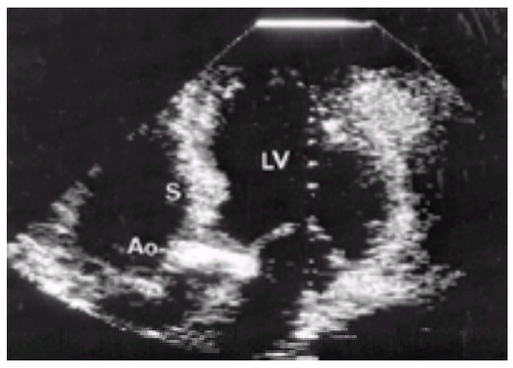

Fig. 11 Aortic stenosis, two-dimensional echocardiogram from apical four-chamber view, showing left ventricle (LV) and heavily calcified aortic valve (Ao). S, septum.

The pressure drop across the outflow tract can be reliably measured by continuous wave Doppler. Additional aortic regurgitation can also be detected. Left ventricular anatomy and function, both in terms of the extent of hypertrophy and cavity size and ejection fraction can be studied.

Transverse dimension and myocardial thickness can be measured by M-mode. Late in disease, the left ventricle becomes enlarged and its ejection fraction falls to values commonly seen in dilated cardiomyopathy.

Cardiac catheterization

The abnormal haemodynamics of aortic stenosis and associated left ventricular disease can usually be comprehensively demonstrated by echocardiography. The role of cardiac catheterization is thus to confirm the pressure drop across the valve in the minority of patients in whom this is not possible for technical reasons using continuous wave Doppler, and to display coronary artery anatomy.

Diagnosis

A complete diagnosis of aortic stenosis depends not only on establishing the anatomical abnormality, but also its severity and the degree of associated left ventricular disease. When cardiac output is normal, a peak pressure drop across the valve of greater than 60 mmHg indicates significant stenosis, and corresponds to a valve area of 0.9 cm^2 or less. The corresponding peak velocity, as measured by continuous wave Doppler is 4.0 m/s.

Mild aortic stenosis is associated with a normal carotid pulse and a short systolic murmur which stops well before the second sound, since a pressure difference between the left ventricular cavity and the aorta is present only during the first part of systole. In addition, both components of the second heart sound are audible and splitting is normal. An uncalcified bicuspid aortic valve causes mild stenosis, with an ejection click and systolic murmur, often followed by a short early diastolic murmur.

Left ventricular hypertrophy can be inferred from a sustained apical impulse with a palpable left atrial contraction. A raised left ventricular end-diastolic pressure can be deduced from accentuation of pulmonary valve closure, which forms the only component of the second sound and, in the late stages of the disease, a third heart sound. In such patients the stroke volume is low, the pressure drop across the valve falls, and significant stenosis may be associated with values of 30 to 40 mmHg. To allow for this, it is necessary to measure valve area.

Differential diagnosis

1. Hypertrophic cardiomyopathy—This can present with a history very similar to that of aortic stenosis. By contrast, the carotid pulse is normal or jerky rather than slow rising. The diagnosis is confirmed by echocardiography, which reveals a normal aortic valve and shows characteristic ventricular features (see Chapter 15.8.2).
2. Congestive cardiomyopathy—Patients with long-standing untreated aortic stenosis can present with severe breathlessness, a large heart on radiography, a small volume pulse with a normal upstroke, a third heart sound, and pansystolic murmur due to papillary muscle dysfunction. These features can all be found in congestive cardiomyopathy (see Chapter 15.8.2). In endstage aortic stenosis, the echocardiogram shows a calcified valve with a significant (more than 35 mmHg) pressure drop across it: in congestive cardiomyopathy it shows a dilated and poorly contractile ventricle, but the aortic valve is normal.
3. Fixed subaortic stenosis—This is usually discovered in asymptomatic children and young adults in whom a systolic murmur is detected on routine examination. An ejection click is absent, and a short early diastolic murmur usually heard. There is clinical and ECG evidence of left ventricular disease, which may be severe. The two-dimensional echocardiogram usually demonstrates the site and type of obstruction (see also Chapter 15.13).

Prognosis

The prognosis of symptomatic aortic stenosis is poor with a 50 per cent survival of only 1 to 2 years. Approximately half the deaths are due to relentless haemodynamic deterioration, and the remainder are 'sudden' and unexpected. The prognosis of asymptomatic but haemodynamically severe aortic stenosis is somewhat better. However, older patients with a peak velocity of 4 m/s or more across the aortic valve are likely to become symptomatic in a period of 2 years or less.

Exercise testing is as safe in aortic stenosis as it is in coronary artery disease, and using it a significant number of 'asymptomatic' patients can be shown to have reduced exercise tolerance. The truly asymptomatic patient, often below the age of 30 years, with normal ECG and without left ventricular hypertrophy or cavity dilatation on echocardiogram, can safely be watched. Regular follow-up of such patients is essential, and surgery should be considered with any evidence of deterioration.

Treatment

Medical treatment has little to offer in aortic stenosis: in mild cases it is unnecessary and in severe cases ineffective. However, it is essential that all patients with aortic stenosis, of whatever severity, have prophylactic antibiotic for any potentially septic hazard. Patients with severe left ventricular disease and fluid retention will benefit from a period of bed rest and treatment with a diuretic before operation is contemplated. ACE inhibitors are contraindicated.

Severe aortic stenosis requires intervention. Unfortunately, aortic balloon valvuloplasty, though satisfactory in infants and children, is either ineffective or harmful in adults in whom the cusps are calcified, and the procedure has been largely abandoned, even as a temporizing manoeuvre in very ill patients. Aortic valve

replacement for aortic stenosis is amongst the most effective of all surgical operations. In uncomplicated cases, it can be carried out with low mortality and morbidity, and should therefore be considered in all patients in whom the disease causes significant symptoms. It is likely to relieve breathlessness, angina, and syncope, whether due to ischaemic heart disease or to the aortic stenosis itself. Associated coronary artery disease is usually treated with bypass grafting at the same operation. Aortic valve replacement is also effective when significant aortic stenosis is complicated by severe left ventricular enlargement. Although the risks of surgery are greater, so are the benefits, and the remarkable improvement in symptoms and prognosis that may follow surgery for this combination of valve and ventricular disease is amongst the most gratifying in cardiology.

Mixed aortic valve disease

Mild to moderate aortic regurgitation often accompanies aortic stenosis, but does little to alter the overall clinical picture. The combination may result from a bicuspid aortic valve, chronic rheumatic heart disease, or be the result of conservative surgery or endocarditis on a stenotic valve. The main haemodynamic disturbance remains increased resistance to ejection rather than a volume load. In addition, superimposition of even a moderately increased stroke volume due to regurgitation on the small, stiff left ventricular cavity of pure aortic stenosis may lead to high filling pressures, left atrial enlargement, and even pulmonary hypertension. Breathlessness and chest pain thus remain the most prominent symptoms. The arterial pulse is bisferiens, with a notch half way up the upstroke, rather than slow rising. Atrial fibrillation usually points to a rheumatic basis; less commonly to high filling pressures in an incompliant cavity. The early diastolic murmur is audible down the left sternal edge. The extent of the volume load, the aortic pressure drop, and the presence or absence of rheumatic mitral involvement can all be determined by echocardiography. Treatment of symptomatic patients is likely to require valve replacement.

Aortic regurgitation

Aortic regurgitation increases stroke volume and when severe and uncorrected causes irreversible left ventricular disease. Its causes are summarized in Table 3.

Pathology

Chronic rheumatic involvement leads to a tricuspid valve whose cusps are thickened, with rolled edges and fused commissures. There may be superimposed calcification or thrombosis. Infective endocarditis may lead to cusp destruction or perforation and may spread to involve the sinus of Valsalva, the atrioventricular node and the interventricular septum, where abscess formation may occur. Organisms may also be carried to the anterior cusp of the mitral valve, where they cause 'jet lesions', localized aneurysms, or perforations. Dilatation of the aortic ring may cause aortic regurgitation with normal cusps. This can result from a 'flask-shaped' aneurysm of the ascending aorta, complicating Marfan's syndrome, or isolated medionecrosis. Syphilitic aortitis causes dilatation of the valve ring, with aneurysm formation of the ascending aorta and involvement of the coronary ostia. Dilatation of the ring may occur on its own or with connective tissue disease such as ankylosing spondylitis, rheumatoid arthritis, Reiter's syndrome, or relapsing polychondritis. Dissecting aneurysm involving the aortic root may separate the cusps from the valve ring; and the presence of a high ventricular septal defect or Fallot tetralogy may leave the cusps unsupported from below.

Pathophysiology

Aortic regurgitation is associated with an increase in left ventricular stroke volume and cavity size. Ventricular mass is therefore increased, but wall thickness is usually within normal limits. In moderately severe aortic regurgitation, the stroke volume is twice normal, and in severe cases, up to three or even four times normal. The characteristics of ejection are altered in that the end-diastolic pressure in the aorta is low, so that the resistance to ejection of blood by the left ventricle is reduced and ventricular systole is prolonged. These factors together with the large stroke volume, explain the characteristic rapid upstroke and large volume pulse. Peripheral vasodilatation also contributes to the large forward stroke volume. In long-standing cases, left ventricular cavity size increases out of proportion to the stroke volume, with loss of the normal myocardial architecture, so that the cavity becomes more spherical in shape; the walls become stiffer and the end-diastolic pressure increases.

Clinical features

Symptoms

Patients with aortic regurgitation remain asymptomatic for many years. When symptoms develop, they are those of left ventricular disease, with limitation of exercise tolerance by breathlessness or chest pain the most prominent one. Less commonly, the presenting symptom may be nocturnal dyspnoea, or an attack of acute pulmonary oedema. Retrosternal pain, aggravated by exertion, may develop in patients with aneurysms of the ascending aorta in whom the coronary arteries are normal. This seems to originate from the aortic root itself. Aortic dissection (see Chapter 15.14.1) may also cause severe central chest pain.

Physical signs

The physical signs of aortic regurgitation are characteristic.

1. The carotid pulse has a large amplitude and a rapid upstroke. Visible arterial pulsation in the neck (Corrigan's sign) excludes significant aortic stenosis. Other physical signs which depend on a large pulse volume and peripheral vasodilation include capillary pulsation, visible in the nail beds, and the de Musset sign, nodding of the head in time with the heart beat. The Durosiez sign, which is of greater clinical value, is elicited by compression of the femoral artery and listening proximally with the stethoscope for a diastolic murmur. It may be positive even when an aortic diastolic murmur is inaudible, and implies retrograde flow in the femoral artery due to aortic regurgitation that is at least moderately severe. The peripheral pulses should always be checked to exclude the presence of coarctation of the aorta.
2. The venous pressure is normal until late in the course of the disease, although the venous pulse may show a dominant 'a' wave (a Bernheim 'a').
3. The left ventricular impulse is sustained, indicating hypertrophy. A palpable 'a' wave is much less common than in aortic stenosis, and when present usually denotes additional left ventricular disease.
4. On auscultation, the characteristic finding is an early diastolic murmur, maximal down the left sternal edge. Less commonly it is loudest at the apex or even in the left axilla (the Cole–Cecil murmur). An ejection systolic murmur is nearly always present, due to the increased stroke volume, and not necessarily to additional stenosis. Aortic valve closure is usually inaudible, but P2 may be accentuated due to passive pulmonary hypertension. At the apex, a mid-diastolic murmur may be heard, indistinguishable from that of mitral stenosis (Austin–Flint murmur). This may continue throughout diastole, with presystolic accentuation and even a loud first heart sound, though the last is never palpable. With the development of ventricular disease, a soft mitral pansystolic murmur or a third sound may appear.

These classic signs of aortic regurgitation may be modified in a number of circumstances. If infective endocarditis has caused cusp perforation, then the early diastolic murmur may have a high-pitched musical quality, a 'seagull murmur'. In the presence of severe left ventricular disease, or less commonly, of rheumatic mitral stenosis or severe pulmonary hypertension, the collapsing pulse, and other evidence of aortic regurgitation may be lost, although the aortic diastolic murmur persists. It is worth noting, however, that Durosiez' sign frequently remains positive in these circumstances if the regurgitation is moderate or severe. In severe aortic regurgitation of rapid onset, usually due to infective endocarditis affecting the aortic valve, the patient presents with a low cardiac output state, normal or reduced pulse volume, and sinus tachycardia. On auscultation the main abnormality is a loud third sound, due to a very short period of forward flow across the mitral valve. A short early diastolic murmur may be audible. Durosiez' sign is usually positive.

Investigations

Chest radiography

Significant aortic regurgitation nearly always causes cardiac enlargement on chest radiography (Fig. 12). The aortic root is often dilated, but the aortic valve not necessarily calcified. The pulmonary vessels remain normal until severe left ventricular disease develops.

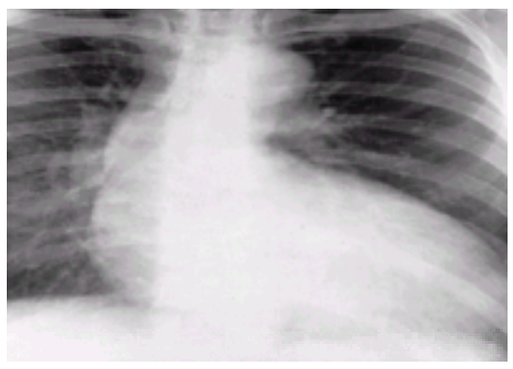

Fig. 12 Chest radiograph from a patient with chronic aortic regurgitation showing cardiac enlargement and dilation of the ascending aorta.

ECG

This usually shows left ventricular hypertrophy on voltage and T-wave criteria, with left atrial enlargement. The duration of the QRS complex increases and left bundle branch block may develop, indicating the presence of intrinsic left ventricular disease. A long P–R interval in association with aortic regurgitation is very suggestive of disease of the aortic root.

Echocardiography

The anatomy of the aortic valve and root can be determined. Dissection and aortic root abscesses can sometimes be detected; vegetations on the aortic valve are well. When infective endocarditis is suspected, transoesophageal echo may give further useful information, particularly if the P–R interval is prolonged. It is also the means of choice for demonstrating aortic root aneurysms. Colour flow Doppler allows the presence of regurgitation to be confirmed and a semiquantitative estimate made of its severity. Left ventricular cavity size is determined from the M-mode record (Fig. 13).

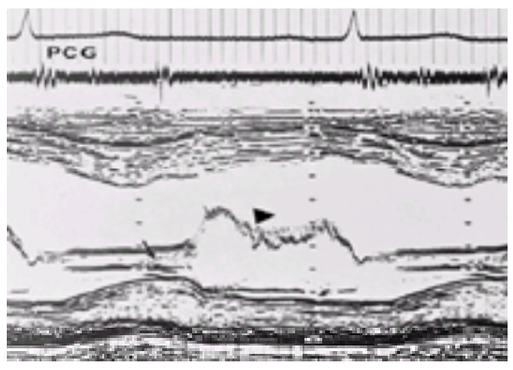

Fig. 13 Chronic aortic regurgitation. M-mode echocardiogram, showing dilatation of the left ventricular cavity; also 'flutter' on the anterior cusp of the mitral valve, marked by arrow, caused by the regurgitant aortic jet striking the cusp. PCG, phonocardiogram.

In acute aortic regurgitation, the mitral valve closes prematurely. This is the result of severe regurgitation into a relatively non-compliant left ventricle causing the cavity pressure to rise, closing the mitral valve in mid-diastole (Fig. 14). As ventricular diastolic pressure rises, the pressure drop from the aorta to the ventricle in diastole may fall to 20 mmHg or less, which is reflected in the continuous wave Doppler record across the valve in diastole. This is pathophysiologically significant, since the pressure difference between aorta and left ventricle during diastole supports coronary flow. Hence, as aortic regurgitation becomes more severe, coronary flow to a volume-loaded ventricle becomes progressively compromised.

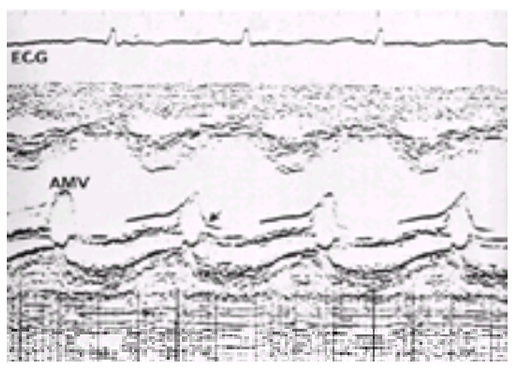

Fig. 14 M-mode echocardiogram showing premature mitral valve closure (arrow) in a patient with acute aortic regurgitation due to infective endocarditis.

Cardiac catheterization

It is not usually necessary to resort to cardiac catheterization to make the diagnosis of aortic regurgitation. Although an aortogram may give useful information, it should be avoided in seriously ill patients because radiographic contrast medium expands the plasma volume and depresses ventricular function. The anatomy of the aortic root is best demonstrated by MRI. Coronary arteriography is usually performed in older patients as a prelude to surgery, even in the absence of clinical evidence of significant coronary artery disease.

Diagnosis

As with aortic stenosis, it is not enough merely to establish the presence of aortic regurgitation; its severity must be estimated, and also the state of the left ventricle. In uncomplicated cases, the severity can be judged indirectly from the carotid pulse and from the heart size on chest radiography, but direct measurement of left ventricular cavity size and stroke volume by echocardiography, MRI, or angiography is more satisfactory. Left ventricular disease can be suspected clinically from accentuated pulmonary valve closure, and from chest radiography by the presence of pulmonary vascular congestion and inappropriate cardiac enlargement. However, left ventricular function is most satisfactorily assessed by direct measurement, the characteristic feature being enlargement of end-diastolic, and in particular, end-systolic cavity size out of proportion to stroke volume such that ejection fraction falls.

Acute aortic regurgitation may present difficulties in diagnosis when the classic physical signs are modified due to a low forward output. Echocardiography is particularly useful in making a definite diagnosis non-invasively, demonstrating aortic vegetations, a large left ventricular stroke volume, and premature mitral valve closure. It is also important to confirm or exclude other types of valve disease.

Coexistent aortic stenosis is often diagnosed on the basis of an ejection systolic murmur, but this does not constitute adequate evidence, and in order to confirm its

presence clinically, a bisferiens pulse should be present. Additional rheumatic mitral stenosis is best confirmed or excluded by echocardiography, although the presence of atrial fibrillation, a palpable first sound, or an opening snap makes its presence very likely on clinical grounds. Mitral regurgitation leads to an additional pansystolic murmur at the apex, which may sound continuous with the early diastolic murmur across the second sound. It is usually caused by a dilated valve ring in the absence of organic mitral valve disease, and thus indicates considerable left ventricular enlargement.

It is usually not necessary to establish the exact aetiology of aortic regurgitation, although it is important to exclude infection and look for the presence of disease of the aortic root and ascending aorta. This should be suspected if there is a history of chest pain that is not clearly anginal in nature, and also from excessive dilatation of the ascending aorta on chest radiography or a long P–R interval on ECG. Proximal aortic root diameter can be measured by echocardiography, and the anatomy of the whole aortic arch demonstrated by MRI.

Differential diagnosis

In the presence of severe pulmonary hypertension, the pulmonary artery may dilate, causing functional pulmonary regurgitation and a soft early diastolic murmur (Graham–Steell murmur). The carotid pulse is normal. Difficulty in diagnosis usually arises when the patient has mitral valve disease, pulmonary hypertension, and an early diastolic murmur. In these circumstances, aortic regurgitation may not necessarily cause an abnormal carotid pulse. In many cases, the differential diagnosis can only be made by Doppler echocardiography, but on clinical grounds, pulmonary incompetence is more likely when there is other evidence of severe pulmonary hypertension, and in particular, when chest radiography shows the main pulmonary artery to be appreciably dilated.

Aortic regurgitation should also be distinguished from other causes of aortic run-off (see Chapter 15.13):

1. persistent ductus arteriosus;
2. ruptured sinus of Valsalva aneurysm; and
3. coronary arteriovenous fistula.

These all cause an increase in pulse pressure, but a continuous rather than an early diastolic murmur down the left sternal edge.

Additional abnormalities that may give rise to confusion are the combination of aortic regurgitation and either mitral regurgitation or a ventricular septal defect. Here the combination of pansystolic and early diastolic murmurs may lead to a continuous quality.

Treatment

Chronic aortic regurgitation

Mild or moderately severe aortic regurgitation is well tolerated and requires no treatment other than prophylactic antibiotic to prevent infective endocarditis.

Severe aortic regurgitation in a symptomatic patient should be treated by aortic valve replacement.

In an asymptomatic patient with severe regurgitation, the decision is more difficult. Without treatment, the outlook is very favourable, with the clinical state remaining stable over many years. Premature operation should be avoided. The most reliable basis for recommending surgery is evidence of progression of disease, such as an increase in heart size on chest radiography, deterioration in the ECG, or an enlarging left ventricular cavity or aortic root on echocardiography. A policy of depending on the results of any single investigation or 'parameter', convenient as it may seem, is inflexible in practice and subject to the limited reproducibility of all cardiological measurements.

Prophylactic administration of nifedipine or ACE inhibitor has been shown to delay the necessity for operation by 2 to 3 years in asymptomatic patients with aortic regurgitation. Severe ventricular disease may become apparent over a period as short as 1 to 2 years, so patients being treated conservatively must be kept under regular review.

Acute aortic regurgitation

Acute aortic regurgitation is a surgical emergency. With a native valve, it is nearly always due to infective endocarditis, and blood cultures should be taken so that the organism can be isolated retrospectively, whilst antibiotics are started preoperatively.

One of the most useful criteria for emergency aortic valve replacement is premature mitral valve closure on the M-mode echocardiogram (Fig. 14). The aim should always be to transfer the patient as soon as possible to a centre capable of performing open heart surgery. As with mitral regurgitation, pharmacological treatment is usually a distraction, but intermittent positive-pressure ventilation is effective treatment for pulmonary oedema.

An increasing P–R interval is a sign of a septal abscess, and requires urgent pacemaker insertion, preferably before transfer, since bradycardia due to complete atrioventricular block can be fatal in severe aortic regurgitation. A prolonged preoperative course of antibiotics is contraindicated in such patients, since the valve is rarely sterilized, and the delay causes further deterioration in left ventricular function with a correspondingly poor outcome. The appearance of aortic regurgitation of any severity with acute dissection of the ascending aorta is also a very strong indication for surgery.

Acquired tricuspid valve disease

Tricuspid stenosis

Although functional tricuspid stenosis may occur with a large flow through the right heart such as occurs in an atrial septal defect, organic tricuspid stenosis is nearly always the result of chronic rheumatic heart disease. Rheumatic tricuspid stenosis virtually always coexists with rheumatic mitral valve disease, although its incidence is about one-tenth. The two conditions are similar both with respect to their pathology and to the functional disturbance that they cause. The valve cusps become thickened, and the commissures fused, so that the cross-sectional area of the orifice is reduced. The tricuspid subvalvar apparatus, though, is not usually involved, nor does calcification occur. The primary functional abnormality is obstruction to right ventricular filling associated with a diastolic pressure drop across the valve. In clinically severe tricuspid stenosis, however, this drop is smaller than it would be with clinically severe mitral stenosis, and is usually within the range of 3 to 10 mmHg. This causes a corresponding increase in right atrial pressure, which leads to ascites and peripheral oedema.

True acquired cusp fusion with severe tricuspid stenosis has recently been described in association with pacemaker catheters, when it presumably represents the effect of chronic trauma. Why it occurs so infrequently with simple pacing catheters, and whether its incidence will increase with the stiffer and more massive catheters needed for automatic implanted cardioverter devices is still not clear.

Clinical features

In patients with chronic rheumatic heart disease, the clinical problem is to recognize the presence of additional tricuspid stenosis in a patient known to have mitral and perhaps also aortic valve disease. This is not always possible on clinical grounds, but a number of indications may be sought. There are no specific findings in the history.

If the patient is in sinus rhythm, tricuspid stenosis is often associated with an 'a' wave in the venous pulse and with evidence of right atrial hypertrophy on ECG. These findings are unusual in the presence of pulmonary hypertension and mitral stenosis alone. The venous pulse is usually otherwise unremarkable.

On auscultation, a separate tricuspid mid-diastolic murmur may be audible. This is similar in timing to a mitral one, but it is higher in pitch, resembling an aortic diastolic murmur in this respect. It is maximal down the left sternal edge or in the epigastrium. A tricuspid opening snap may also be present; it is later than a mitral one and its timing with respect to pulmonary valve closure varies with respiration.

Investigations

Chest radiography may be suggestive, since right atrial enlargement causes the heart shadow to enlarge to the right of the midline. These appearances, however, are non-specific, and may be present with functional tricuspid regurgitation, or even a giant left atrium.

Echocardiography gives the diagnosis. Cross-sectional echocardiography shows doming of the tricuspid valve into the right ventricle during systole (Fig. 15) in the apical four-chamber view. The diastolic pressure drop can be estimated by continuous wave Doppler.

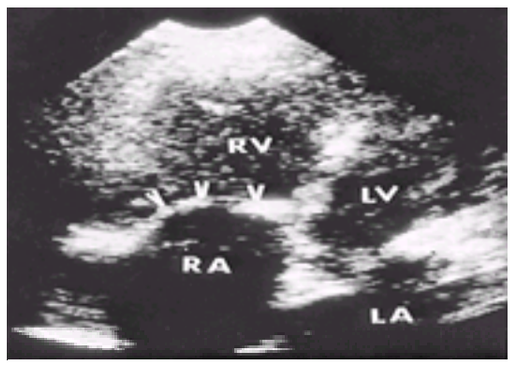

Fig. 15 Rheumatic tricuspid stenosis, apical four-chamber view showing doming and thickening of the tricuspid valve during diastole (arrows). LA, left atrium; LV, left ventricle; RA, right atrium; RV, right ventricle.

Cardiac catheterization is now rarely performed to diagnose tricuspid stenosis, but it may show a small diastolic pressure drop across the valve.

Cases of organic tricuspid stenosis may still reach operation for mitral or aortic valve disease without the diagnosis having been made. When they do so, the previously undiagnosed tricuspid stenosis may be unmasked by successful left-sided surgery, which allows cardiac output to increase. This causes a corresponding increase in the tricuspid diastolic pressure drop, which leads to salt and water retention, so that a patient thought to have had a satisfactory operation develops striking ascites or peripheral oedema afterwards.

Treatment

Medical treatment is not very satisfactory but consists of diuretic administration to control fluid retention. Prolonged administration of inappropriately large doses leads to potassium depletion. Definitive treatment is surgical, consisting of either valvotomy or repair at the time that the other valve lesions are dealt with. Isolated tricuspid stenosis, developing after mitral valve surgery, has been dealt with by balloon valvuloplasty. Tricuspid valve replacement is avoided whenever possible in view of the physiological significance on the right side of the heart of the diastolic pressure drop across all normally functioning prostheses. An additional procedure on the tricuspid valve increases the operative risk of mitral valve surgery, with greater postoperative incidence of jaundice and arrhythmias.

Tricuspid regurgitation

As with mitral regurgitation, a number of different pathological processes can cause tricuspid regurgitation (Table 4).

The tricuspid valve is much more liable to develop functional regurgitation than the mitral valve, and tricuspid regurgitation is often functional, occurring in association with dilatation of the right ventricular cavity. It is particularly common in patients with pulmonary hypertensive mitral valve disease, but may also occur with primary pulmonary hypertension, or in the terminal stages of many types of congenital heart disease, particularly those with a significant left to right shunt.

Severe, non-rheumatic tricuspid regurgitation is being increasingly recognized as occurring late after mitral valve replacement, in the absence of significant left-sided disease or pulmonary hypertension. Its cause is not clear, but there is no evidence of disease of the valve cusps or subvalve apparatus.

Organic tricuspid regurgitation may be congenital, as an isolated abnormality, or associated with the Ebstein anomaly. A cleft right-sided atrioventricular valve may also occur in ostium primum atrial septal defect. Acquired, organic tricuspid regurgitation may be rheumatic in origin, or result from infective endocarditis of a previously normal valve, which occurs particularly commonly in intravenous drug users. Right-sided endomyocardial fibrosis causes progressive obliteration of the right ventricular cavity with scarring and distortion of the tricuspid subvalvular apparatus.

Mid-systolic prolapse of the tricuspid valve can occur in exactly the same way as that of the mitral valve, and is common in Marfan's syndrome. Organic tricuspid regurgitation has been described as a long-term consequence of radiotherapy to the thorax, when it may be associated with features of pericardial constriction or restrictive myocardial disease, making its diagnosis difficult.

Clinical features

The clinical features of tricuspid regurgitation are those of severe and chronic elevation of the venous pressure, often in association with disease on the left side of the heart.

The symptoms are non-specific, although when tricuspid regurgitation supervenes in a patient with mitral stenosis, it is often associated with an increase in the prominence of fatigue as a factor limiting exercise tolerance instead of breathlessness. Symptoms may also be related to the development of oedema or ascites: hepatic enlargement may be associated with nausea and upper abdominal or epigastric pain aggravated by exercise.

The main physical sign of tricuspid regurgitation is a raised venous pressure with a prominent systolic wave, which is almost a *sine qua non* for the diagnosis. The mean venous pressure may be very high, greater than 15 cmHg, with pulsations visible in the retinal vessels or palpable in the femoral veins. The high venous pressure is also responsible for the protein-losing enteropathy that sometimes occurs in the same way as with constrictive pericarditis (see Chapter 15.9). In approximately two-thirds of patients, there is associated systolic expansile pulsation of the liver, which may be considerably enlarged and tender. In long-standing cases, hepatic fibrosis develops so that this physical sign disappears. The hepatic dysfunction may also cause mild jaundice, which with increased skin pigmentation can give these patients a very characteristic appearance.

In approximately one-third of cases, a tricuspid pansystolic murmur is present, audible down the left sternal edge. An increase in intensity during inspiration is difficult to demonstrate, so that the murmur is usually indistinguishable from that of functional mitral regurgitation.

Investigations

Chest radiographic findings depend mainly on other cardiac disease present, but, as with tricuspid stenosis, there may be enlargement of the heart shadow towards the right.

The ECG may show right atrial hypertrophy in isolated tricuspid regurgitation if the patient is in sinus rhythm, but otherwise is dominated by other cardiac disease present.

Echocardiography is the best way of making the diagnosis. Cusp disease, either rheumatic or 5-hydroxytryptamine (carcinoid, see later) induced, and right ventricular

function are assessed by the cross-sectional technique, while the extent of the regurgitation can be measured by continuous wave and colour flow Doppler. The nature and extent of left-sided disease and pulmonary hypertension can also be documented. Cardiac catheterization is rarely necessary either to make the diagnosis or assess its severity.

Treatment

Medical treatment with diuretics deals with associated fluid retention, and may even allow right ventricular cavity size to decrease, restoring competence to the tricuspid valve. Isolated tricuspid incompetence, unless very severe or accompanied by right ventricular disease, is reasonably well tolerated. Surgical treatment is avoided if possible. However, if regurgitation is very severe and fluid retention requires doses of diuretics large enough to cause significant metabolic consequences, then intervention may be considered. Unfortunately, repair and replacement of the tricuspid valve are unsatisfactory operations: the former does not usually control regurgitation and the latter leads to a very significant diastolic pressure drop between right atrium and right ventricle. In addition, the risks of surgery are high in these patients, and postoperative jaundice and renal failure are common.

When tricuspid regurgitation occurs in association with rheumatic heart disease involving the left side of the heart, it may subside spontaneously after the latter has been dealt with surgically, although there is a case for routine tricuspid valve plication or repair to prevent tricuspid regurgitation developing postoperatively.

Serotonin-induced heart disease

Increased levels of 5-hydroxytryptamine (**5HT**, serotonin) associated with metastatic carcinoid disease cause severe tricuspid and pulmonary regurgitation, and similar findings are associated with anorectic agents. Fenfluramine also causes 5HT release and thus has similar effects on the right-sided valves.

5HT is normally cleared by the lungs, so that in the absence of a right to left shunt, the left side of the heart is not usually affected in carcinoid disease. However, fenfluramine has been used in combination with phentermine, which blocks pulmonary uptake of 5HT, so that together these drugs can lead to involvement of the mitral and aortic valves as well.

The valves are thickened, with a glistening appearance, due to the deposition of plaques of fibrosis. Vascularization and cusp fusion do not occur. However, the cusps themselves become retracted, so that the dominant lesion is regurgitation through a central jet. Subendocardial thickening may also occur, so that the effects of reduced cavity compliance are superimposed on those of volume overload. On the left side of the heart, the regurgitation may be severe enough to require valve replacement.

Not all patients with carcinoid disease develop cardiac manifestations, but when they are present they may progress, even after removal of the tumour. In slowly progressive cases, tricuspid valve replacement has been successfully undertaken.

Pulmonary valve disease

Acquired pulmonary valve disease is unusual. The commonest form is that associated with severe pulmonary hypertension and dilatation of the pulmonary valve ring, causing mild regurgitation. This commonly occurs in association with pulmonary hypertensive mitral valve disease, causing a soft early diastolic murmur (the Graham–Steell murmur), but an identical picture may be present with severe pulmonary hypertension from any cause. Although the murmur itself is early diastolic in timing, there is no associated abnormality of the carotid pulse as would be expected in aortic regurgitation. Nevertheless, the differential diagnosis on clinical grounds can be difficult when mitral valve disease is severe. Mild pulmonary regurgitation is effectively a normal finding on colour flow Doppler; a more extensive jet in a patient with pulmonary hypertension being required to confirm the diagnosis. Aortic regurgitation can be confirmed or excluded by Doppler. Rheumatic pulmonary regurgitation is extremely rare, although it has been reported in populations living at high altitudes. However, even when present it contributes little to overall disability. Pulmonary regurgitation may also form part of the carcinoid syndrome, but its effect on the clinical picture is less than that of the tricuspid or left-sided valves. It may also be iatrogenic, following pulmonary valvotomy for pulmonary stenosis, when it contributes to the elevated venous pressure that can persist for a variable period after this operation. It is of no clinical consequence and requires no specific treatment.

Valve disease and pregnancy

For information of the effect of pregnancy on patients with valve disease, see Chapter 13.6.

Management of patients with valve prostheses

Many patients have received heart valve replacements over the past 40 years, with very significant improvement in their quality of life. However, although the haemodynamic performance of these prostheses is greatly superior to the diseased valves that they replaced, it is not normal, and survival with any valvular prosthesis is significantly less than that for age-matched normal individuals. Having a valve replacement should now be regarded as the commonest form of valve disease in Western society.

Valve prostheses can be mechanical or biological. Mechanical prostheses include the Starr–Edwards ball and cage prosthesis, which has the advantage of a 30-year follow-up and remarkable reliability; single tilting disc valves (Bjork–Shiley); but those inserted over the last 10 years are likely to be bileaflet (St Jude). The more modern types have a larger effective orifice area than the Starr–Edwards in relation to the size of their ring.

Biological prostheses usually consist of a plastic stent on which cusps made from some biological material are mounted. The cusps may be derived from porcine aortic valve or pericardium. More recently, unstented xenografts (e.g. the Toronto) have been used, where the aortic homograft is mounted on a ring of native aortic root. In the Ross operation, the patient's own pulmonary valve is inserted into the aortic position (pulmonary autograft) and replaced by an aortic homograft in the right ventricular outflow tract.

There are minor differences in performance between the various valve substitutes, but these are not of great clinical significance. Apart from the pulmonary autograft, all fall short of their natural counterpart *in vivo*. Under normal working conditions, pressure differences are present across mitral prostheses, which range from 4 to 5 mmHg for the Starr–Edwards to 2 to 4 mmHg for the others. In addition, all mitral valve substitutes have a rigid mitral ring which interferes with ventricular function. Apart from the autograft, and to a lesser extent the homograft, whose performance approximates to that of the native valve, systolic gradients across aortic prostheses are in the range 10 to 25 mmHg at rest, increasing on exercise.

The main factors guiding choice of valve prosthesis are durability and the likely incidence of thrombotic complications. Present operative mortality is in the region of 3 to 5 per cent for single valve replacement and approximately 10 per cent for double valve replacement. These values are higher if simultaneous coronary artery grafting is necessary. Long-term survival studies have shown that 10-year survival after single valve replacement is approximately 70 to 75 per cent, and after double valve replacement 50 to 65 per cent. For reoperation, mortality is higher, the exact figure depending on the circumstances in which surgery is performed.

Late complications of valve replacement

Thromboembolism

This is a major complication associated with all mechanical prostheses. Long-term anticoagulant therapy with a drug of the warfarin type is essential in all patients in whom these prostheses have been inserted, and even with satisfactory control (international normalized ratio, **INR**, between 3 and 4.5), an incidence of significant events including transient weakness, dysphasia, or visual disturbances of 1 to 2 per cent per annum can be expected. At the same time, anticoagulant therapy itself causes bleeding complications severe enough to require admission to hospital with an incidence of approximately 1 per cent per annum.

In a small minority of patients, emboli are frequent in spite of good anticoagulant control. Initially, such patients should be given an antiplatelet agent such a dipyridamole, and the anticoagulant dose adjusted accordingly. The possibility of some other cause for the neurological manifestations, such as cerebrovascular disease, must always be considered. However, frequent embolization may be associated with thrombosis of the prosthesis, and if this is proven beyond all question

and cannot be suppressed medically, reoperation and replacement with a biological prosthesis may be necessary.

The incidence of thromboembolic complications is much lower with biological prostheses, so that long-term anticoagulant therapy can be dispensed with in patients in sinus rhythm after aortic or mitral valve replacement. However, many surgeons recommend a short course of 2 to 3 months in such patients whilst suture lines become endothelialized. Patients with atrial fibrillation require standard long-term anticoagulant therapy.

In developing countries, mitral replacement may have to be performed in children under the age of 15 years, in whom biological valves are unsuitable. The use of a mechanical prosthesis might seem appropriate, but facilities for regular anticoagulant monitoring are not available, while uncontrolled administration of standard doses of warfarin is associated with unacceptable risk of haemorrhage. This therapeutic dilemma is, at present, unsolved.

Limited prosthetic function

In a minority of patients, valve replacement may give rise to severe haemodynamic disturbances, such that in extreme cases the condition of the patient may be worse after the operation than before. This usually arises when the valve ring or the ventricular cavity is very small, and a correspondingly small prosthesis was inserted. In the mitral position, resting diastolic pressure differences as high as 20 mmHg may be present, or of 50 mmHg across the aortic valve on this basis. A related problem is the insertion of a prosthesis that is too large, particularly of the ball and cage type. In the mitral position, the cage may impinge on the septum, and obstruct the left ventricular outflow tract, causing subaortic stenosis; whilst in the aortic position, obstruction may develop between the ball and the aorta. These complications are now avoided by the use of low-profile prostheses, such as the St Jude bileaflet prosthesis.

Infection

Patients with prostheses, mechanical or biological, are at greatly increased risk of infective endocarditis. The infecting organism may have been introduced at the time of operation, when it usually manifests within 2 months of surgery. Later infections are bloodborne. It is therefore essential that all patients receive full antibiotic prophylaxis immediately after surgery and subsequently for dental manipulations and other potentially septic hazards, rather than the single dose of single agent currently recommended for routine dental prophylaxis in those at lower risk (see Chapter 15.10.2).

Infective endocarditis is a very serious complication, and rarely responds to antibiotic therapy alone. A second valve replacement is nearly always required, often in a seriously ill patient in whom the valve ring may be infected and friable. The clinical features of endocarditis on a mechanical prosthesis differ very significantly from those on a native valve. Vegetations are uncommon. The infection is often confined to the tissue around the prosthesis, including the sewing ring, and may rot the sutures, so that the first manifestation is sudden death due to displacement of the valve to the aortic bifurcation. Partial dehiscence leads to severe regurgitation with a rocking prosthesis, demonstrable by simple screening or cross-sectional echocardiography. The main disturbance may be stenosis, due to ingrowth of infected clot. Abscess formation around the prostheses is particularly common in the aortic position, which may cause complete heart block or perforation to the right side of the heart. These difficulties are compounded by the acoustic properties of the mechanical valves themselves which limit the value of echocardiography. Para-aortic abscesses are well demonstrated by transoesophageal echo.

Prosthetic dysfunction

This is an important cause of morbidity in patients who have undergone valve replacement. There may be structural damage to the prosthesis itself, which is uncommon in mechanical valves, though occasional batches may undergo strut fracture due to metal fatigue, a well known example being the concavo-convex Bjork–Shiley valves inserted in the late 1970s.

Malfunction is much commoner with biological prostheses, when cusps may become calcified, perforated, or detached. Calcification of porcine bioprostheses regularly occurs within 1 to 2 years of insertion in children under the age of 15 years. This complication takes much longer to develop in the aortic homograft, usually 10 to 15 years. Once cusp degeneration has occurred, the valve should be regarded as unstable, since sudden cusp detachment or perforation can occur, and it is therefore inadvisable to adhere to haemodynamic guidelines appropriate to native valves when deciding on the timing of further surgery. It is an advantage of the homograft valve that sudden deterioration in haemodynamic function is much less common.

Mechanical prostheses are subject to thrombosis. This may take two forms. Deterioration in function may be insidious over a period of several months or years due to ingrowth of organized clot (pannus), usually from the atrial side (Fig. 16). This may be associated with an increased incidence of emboli in spite of adequate anticoagulant therapy. Alternatively, the prosthesis may clot acutely: this is particularly likely to occur with the Bjork–Shiley in the mitral position, and represents a surgical emergency. It can often be recognized clinically, the patient presenting with pulmonary oedema or in a low cardiac output state, and the closing click of the prosthesis no longer audible. Operation is required as soon as possible, since deterioration may occur within hours. If the condition of the patient is so poor as to preclude anaesthesia, then thrombolysis should be used, in spite of the risk of systemic embolism. Improvement may be expected within a few hours, when the thrombolysis can be neutralized and surgery undertaken.

Fig. 16 Thrombosed Star–Edwards prosthesis, removed at emergency operation.

Finally, paraprosthetic regurgitation may develop. In the aortic position, this may have been present since the original operation, because heavy calcification of the original valve extended into the valve ring. Paraprosthetic regurgitation that appears suddenly always raises the possibility that the prosthesis might have become infected. Mitral paraprosthetic regurgitation usually results from part of the valve sewing ring tearing away. Again, infection should always be considered, but regurgitation is well documented in its absence.

Recognizing prosthetic dysfunction

Stenosis or regurgitation associated with a prosthesis does not have the same physical signs as the corresponding lesion of the native valve. In general, it presents as deterioration in cardiac state, whose progress may be acute or chronic. The clinical picture is of 'heart failure'. On examination, the venous pressure is raised, the liver enlarged, and chest radiography shows that the heart has enlarged and pulmonary congestion appeared. When a mitral prosthesis is involved, there are characteristically no murmurs, other than those of tricuspid regurgitation; an aortic systolic murmur may be present, but its intensity and timing differs little from that of a normally functioning prosthesis. It is essential, therefore, that the possibility of a prosthesis-related complication is considered in all such patients in whom a diagnosis of 'heart failure' is entertained. The differential diagnosis is exacting, and requires echocardiography, Doppler, and possibly cardiac catheterization by an experienced operator. All patients presenting in this way should therefore be referred to a unit where these investigations can be performed reliably, and emergency surgery can be undertaken if necessary.

Haemolysis

All mechanical prostheses are associated with increased intravascular haemolysis. This rarely gives rise to clinical problems when the prosthesis is functioning

normally, and anaemia does not occur. The extent of haemolysis can be estimated from a peripheral blood film, which shows fragmented forms, from depression or absence of serum haptoglobin, and from an increase in lactate dehydrogenase levels.

Haemolysis may become significant with a normally functioning prosthesis when the patient has a compensated haemolytic state of some different aetiology, such as congenital spherocytosis or thalassaemia minor. In these circumstances, there is a risk of haemolysis becoming severe.

Mild paraprosthetic regurgitation whose severity is insufficient to give rise to any haemodynamic complications can cause clinically significant haemolysis. In such cases, it may be undesirable to expose the patient to the risk of reoperation, particularly if the original valve leak was due to some predictable cause such as heavy calcification of the valve bed, and so likely to recur. Provided that haemolysis is not severe, such patients can usually be treated medically on maintenance therapy with iron and folic acid. A requirement for transfusion, however, is a strong indication for reoperation.

Left ventricular disease

This is a major cause of morbidity and mortality after valve replacement. There is no single cause. In many patients, severe left ventricular disease was present preoperatively, and though some improvement frequently occurs with correction of the valve disease, function never returns to normal. Operation itself causes additional damage. Methods of myocardial preservation during the period of cardiopulmonary bypass have improved very considerably over the last 20 years with the general introduction of cold blood cardioplegia, but before then ischaemic arrest appears to have been associated with myocardial damage that may take several years to become manifest. A rigid prosthetic mitral ring invariably leads to abnormal function, and there is increasing evidence that section of the papillary muscles may have the same effect. Coronary emboli may arise from the prosthesis. Many patients are of an age to have additional coronary artery disease.

Whatever the cause, left ventricular disease after valve replacement presents its usual clinical features. There is progressive limitation of exercise tolerance and breathlessness due to reduction in cardiac output and pulmonary congestion. Venous pressure becomes raised, and the earliest clinical evidence may relate to right rather than left ventricular disease, with elevated venous pressure, fluid retention, and hepatic congestion. Auscultatory signs may be modified, and in particular third and fourth heart sounds are not audible in patients with mechanical mitral prostheses. Chest radiography shows an increase in heart size and pulmonary congestion. ECG may show Q waves, but their absence is of no significance.

The differential diagnosis of ventricular disease after valve replacement is prosthetic dysfunction, and it is essential that a correct diagnosis is established in any patient who fails to progress, or whose improvement after operation is not maintained. Echocardiography, transthoracic and transoesophageal, has proved of great value in such patients, since it allows the very active left ventricular wall motion that accompanies a paraprosthetic leak to be distinguished from the dilated cavity and poor shortening fraction of left ventricular disease. Continuous wave Doppler can be used to detect significant gradients across biological valves. However, unless the diagnosis is clear from non-invasive investigation, cardiac catheterization is required to settle the diagnosis beyond doubt.

The prognosis once clinically apparent left ventricular disease has developed is poor, usually being of the order of 1 to 2 years, so it is essential that no remediable cause is overlooked.

Follow-up of patients after valve replacement

Patients must be followed up after valve replacement for life. This must at least be at annual intervals, with regular chest radiography and ECG in an experienced clinic. Echocardiography should be performed early, not only to detect immediate postoperative complications such as a pericardial fluid collection, which can be potentially fatal if delayed tamponade occurs, but also to establish a baseline from which to detect future change. Deterioration must be detected early, and investigated in detail so that life-threatening complications are not missed. Dental prophylaxis is essential.

Valve disease and pregnancy

The circulatory changes associated with pregnancy modify the physiology of valve disease and thus its management. These changes are hormonally mediated, and result in an increase in cardiac output by 40 to 45 per cent above control values and a corresponding fall in peripheral resistance, maximal in the middle trimester. This elevation in cardiac output is mediated in approximately equal parts by increases in stroke volume and heart rate. The raised stroke volume, in particular, can lead to the development of a soft ejection systolic murmur, whose benign origin can normally be clarified by echocardiography, chest radiography being avoided where possible.

The effects of the circulatory changes of pregnancy on those of pre-existing valve disease are very predictable. Since the increase in cardiac output is brought about mainly by a fall in peripheral resistance, patients with valvular regurgitation do well. The main problems are with stenotic lesions. In mitral stenosis, the increased stroke volume and tachycardia combine to increase the diastolic pressure drop across the valve, thus predisposing to pulmonary oedema. Pulmonary oedema in a pregnant patient is a medical emergency, and prompt transfer to a surgical centre should be arranged if the symptoms are more than mild. Medical treatment with digoxin and b-blocking agents, aimed at slowing the heart rate, may be satisfactory, but surgery is often necessary. Mitral valve anatomy in young women is usually compatible with a conservative operation. Closed mitral valvotomy has a long and proven record, and is available if the surgical expertise is still available. In its absence balloon valvuloplasty, with appropriate screening to minimize radiation, is also satisfactory.

The combination of aortic stenosis and pregnancy is also difficult to manage, since an increase in the pressure drop across the valve combined with a fall in peripheral resistance is particularly liable to cause syncope as well as aggravating left ventricular disease, and leading to pulmonary oedema and death. Cardiopulmonary bypass for valve replacement during pregnancy is associated with a 20 to 30 per cent fetal loss, but there may be no alternative. In young female patients with mitral and particularly with aortic stenosis, therefore, there is a strong case for dealing with the valve lesion before pregnancy.

Prosthetic valves and pregnancy

Pregnancy

Clinical problems in pregnancy arise from the anticoagulant therapy needed for mechanical prostheses. Pregnancy is accompanied by an increase in coaguability and a decrease in fibrinolysis. Coumarin anticoagulants lead to an incidence of spontaneous abortion of approximately 30 per cent, particularly during the sixth to ninth weeks, while in the third trimester, they are associated with fetal haemorrhage and post-partum bleeding, even when the INR is well controlled. In the absence of large-scale trials, a commonly used regime is to give heparin for the first trimester, followed by warfarin until just before delivery, when heparin is substituted. Heparin does not cross the placenta, and can readily be discontinued or neutralized with protamine. There is recent evidence to suggest that the incidence of fetal abnormality is low if the warfarin dose can be maintained below 5 mg/day.

Further reading

Benjamin EJ *et al*. (1992). Mitral annular calcification and the risk of stroke in an elderly cohort. *New England Journal of Medicine* **327**, 374–9.

Blackstone EH, Kirklin JW (1992). Recommendations for prophylactic removal of heart valve prostheses. *Journal of Heart Valve Disease* **1**, 3–14.

Bulkley BH, Roberts WC (1976). The heart in systemic lupus erythematosis, and change induced in it by corticosteroid therapy. *American Journal of Medicine* **58**, 243–64.

Cohen DJ *et al*. (1992). Predictors of long-term outcome after percutaneous balloon mitral valvuloplasty. *New England Journal of Medicine* **327**, 1329–35.

P>Connolly HM *et al*. (1997). Valvular heart disease with flenfuramine-phentermine. *New England Journal of Medicine* **337**, 581–8.

Fowler N, van der Bel-Kahn JM (1979). Indications for surgical replacement of the mitral valve with particular reference to common and uncommon causes of mitral regurgitation. *American Journal of Cardiology* **44**, 157.

Freed LA *et al*. (1999). Prevalence and clinical outcome of mitral prolapse. *New England Journal of Medicine* **341**, 1–7.

Groves PH, Hall RJC (1992). Late tricuspid regurgitation following mitral valve surgery. *Journal of Heart Valve Disease* **1**, 80–6.

Hehoe JA, Carpenter DF, Golden A (1968). Cardiac valvular lesions in rheumatoid arthritis. *Archives of Internal Medicine* **122**, 141–6.

Leatham A, Brigden W (1980). Mild mitral regurgitation and the mitral prolapse fiasco. *American Heart Journal* **99**, 659–64.

Ling LH *et al*. (1996). Clinical outcome of mitral regurgitation due to flail leaflet. *New England Journal of Medicine* **335**, 1417–23.

Oakley CM, Burkhardt D (1993). Optimal timing of surgery for chronic mitral or aortic regurgitation. *Journal of Heart Valve Disease* **2**, 223–9.

Rahimtoola SH (1983). Valvular heart disease; a perspective. *Journal of the American College of Cardiology* **1**, 199–215.

Roberts WC *et al*. (1981). Congential bicuspid aortic valve causing severe, pure aortic regurgitation without superimposed infective endocarditis. *American Journal of Cardiology* **47**, 206–9.

Ruttley MST (1992). The chest radiograph in adult heart valve disease. *Journal of Heart Valve Disease* **2**, 205–17.

Selzer A (1987). Changing aspects of the natural history of aortic stenosis. *New England Journal of Medicine* **317**, 91–8.

Smith HJ *et al*. (1976). The natural history of rheumatic aortic regurgitation and the indications for surgery. *British Heart Journal* **38**, 147–54.

Smith N, McAnulty JH, Rahimtoola SH (1978). Severe aortic stenosis with impaired left ventricular function and clinical heart failure: results of valve replacement. *Circulation* **58**, 255–64.

Vijayaraghavan G *et al*. (1977). Rheumatic aortic stenosis in young patients presenting as combined aortic and mitral stenosis. *British Heart Journal* **39**, 294–8.

Wood P (1954). An appreciation of mitral stenosis. Part 1. Clinical features. *British Medical Journal* **i**, 1051–63.

Wood P (1954). An appreciation of mitral stenosis. Part II. Investigations and results. *British Medical Journal* **I**, 1113–24.

15.8.1

Myocarditis

Jay W. Mason

Introduction

Myocarditis is a disease that has captured the interest of clinicians and scientists. This interest is generated by its varied aetiology, its diagnostic and therapeutic challenges, the possibility that myocarditis may be the primary cause of dilated cardiomyopathy, and the availability of numerous, easily manipulated animal models of the disease.

Clinical features

Myocarditis affects young people. The average age of patients in the United States Myocarditis Treatment Trial was 42 years. There was a slight male predominance (62 per cent) in that trial, but other series have not demonstrated a sex predilection. The true incidence of myocarditis is unknown. Autopsy studies have reported up to a 3 per cent incidence, but varying histological criteria were used, and myocarditis may occur as an incidental complication of other fatal illnesses. About 10 per cent of patients with influenza infections have electrocardiographic abnormalities, but it is not know if these are the result of myocarditis. The incidence of fatal myocarditis was estimated in a retrospective review of United States Air Force recruits undergoing boot camp training. There were eight such deaths over 1 606 167 person days, which yields an estimate of 4/100 000 per year in people aged 17 to 28 years. This incidence is probably greater than would be expected in the general population in the United States, who would not be exposed to similar levels of intense exercise or high probability of transmission of viral illnesses.

In Europe and North America most cases of myocarditis present with congestive heart failure of unknown cause. In many cases there is a history of recent upper respiratory tract infection or of a 'flu-like' illness. This is followed by symptoms of cardiac decompensation, usually fatigue, breathlessness, and cough. Chest pain occurs in a substantial minority of patients. A small proportion of patients with myocarditis present with ventricular tachyarrhythmias and minimal or no cardiac dilatation. Typically, the duration of symptoms due to infection is brief, less than 1 month in approximately 50 per cent of patients and nearly always less than 1 year. Myocarditis should always be suspected when a patient presents with unexplained congestive heart failure with a rapid onset, especially if there is a viral prodrome.

Clinical examination typically reveals signs of cardiac failure. The ECG may show conduction abnormalities and ST/T-wave changes, or arrhythmias (atrial or ventricular). The chest radiograph shows cardiomegaly and pulmonary oedema. The echocardiogram reveals four-chamber dilatation and reduced contractility, and is notable for the fact that valvular disease is absent or minimal. Should coronary angiography be performed, the vessels are normal or show only minor abnormalities. The role of myocardial biopsy will be discussed later. CPK-MB elevation is common.

Although viruses are thought to be the most common cause of myocarditis, viral titres are rarely useful in diagnosis and treatment. Although the cardiotrophic enteroviruses, including echoviruses and coxsackieviruses, are the predominant aetiological agents, dozens of viruses have been implicated and many more, undoubtedly, cause myocarditis in humans. Thus, it is impractical to exclude all. In addition, patients usually present a substantial period of time after the viral infection has cleared, making it difficult or impossible to document an acute rise in titre. Knowledge of a specific virus, or any virus, as the cause in a given case of myocarditis has little, if any, therapeutic relevance. Even if virocidal therapy (which is not yet a proven treatment; see below) is being considered, negative titres for the common viral agents do not exclude a viral aetiology.

A small number of patients, perhaps about 10 per cent, present with a secondary form of myocarditis. These special presentations are discussed below.

Aetiology and pathogenesis Table 1)

The most common form of myocarditis in Europe and North America is known as lymphocytic myocarditis or non-specific lymphocytic myocarditis. Other frequently applied terms are viral or post-viral myocarditis, because an antecedent viral infection is common. Indeed, some experts believe that nearly all lymphocytic myocarditides are the result of viral infections, presumed to be subclinical in those patients with no awareness of a viral prodrome.

In animal models enteroviruses, such as coxsackie B3, can cause two phases of myocarditis. The first is the result of direct injury of myocytes by replicating virus and the resulting acute immune response. A delayed immune response brings about the second phase, and it is this which is thought to be the more common cause of overt congestive heart failure. The underlying mechanisms are complex and incompletely understood, but most hypotheses suggest that autoimmune phenomena play a major role. In some instances molecular mimicry may be involved, in which the similarity of a viral antigen to a myocardial protein triggers an autoimmune reaction. In others an autoimmune response to cellular proteins released during the viral replication phase may occur, and myosin has been implicated in this regard. Cytokines arising from immune activation and cellular necrosis probably play a role in some cases, bringing about further cellular damage. Although all of these mechanisms have been well delineated in murine models, they have not been proven to account for myocarditis in humans.

Myocarditis may result from a hypersensitivity reaction to a drug or other agent (see Table 1). In these cases eosinophils accompany the inflammatory lymphocytic infiltrate. A number of other specific causes of myocarditis, each with differing pathogenises and presentations, are discussed below.

Relationship to idiopathic dilated cardiomyopathy

Classic lymphocytic myocarditis usually resolves, with resultant improvement in cardiac function over weeks or months. In the United States Myocarditis Treatment Trial, the mean left ventricular ejection fraction improved during the year after initial presentation by more than 10 EF units (from 24 to 36 per cent; normal more than 55 per cent). However, residual cardiac dilatation and dysfunction were common, and mortality was high, reaching 55 per cent at 5 years. In those patients who do not recover fully, the ensuing clinical picture cannot be distinguished from that of idiopathic dilated cardiomyopathy. The possibility that myocarditis may occur without an obvious viral prodrome therefore raises the interesting possibility that viral myocarditis may be a common covert cause of idiopathic dilated cardiomyopathy. In the United States trial, only 10 per cent of patients with suspected myocarditis had positive biopsies. Hence, the fact that endomyocardial biopsy does not reveal myocarditis in patients with idiopathic dilated cardiomyopathy may be the result of timing of the biopsy. The lymphocytic infiltrate usually resolves spontaneously, and it may be that earlier biopsy might have detected myocarditis in a portion of the 90 per cent with negative biopsies. The presence of viral genomic material in a minority of these negative biopsies lends support to the viral aetiology hypothesis. Absence of viral genome in the rest of them does not eliminate post-viral autoimmune processes, proceeding despite complete viral clearing, as a possible aetiology.

Treatment of post-viral and non-specific lymphocytic myocarditis

As stated above, non-specific lymphocytic myocarditis is believed by most to have a viral aetiology, even in the absence of a clinically apparent viral prodrome. In the acute phase of viral myocarditis, the direct cytolytic effect of viral myocyte infection may lead to congestive heart failure, although this is uncommon. In this early phase, the immune response is likely, on balance, to be beneficial. Thus, antiviral therapy might be expected to be helpful, on theoretical grounds, but

immunosuppressive therapy would not. However, no antiviral therapies have been adequately tested in humans. Although hyperimmune globulin is thought to be effective on the basis of retrospective studies, its efficacy has not been proved in a prospective trial. In the second stage, thought to result from an adverse immune response to previous infection, immunosuppressive therapy has appeared to be beneficial in uncontrolled trials. However, no benefit was demonstrated in the United States Myocarditis Treatment Trial, the only prospective, randomized trial performed in patients with myocarditis defined histologically. In that trial the 'Dallas' criteria defined myocarditis histologically as a lymphocytic infiltrate with associated myocyte necrosis (Fig. 1). Treatment with prednisone combined with either cyclosporin or with azathioprine did not improve outcome, as defined by change in left ventricular ejection fraction.

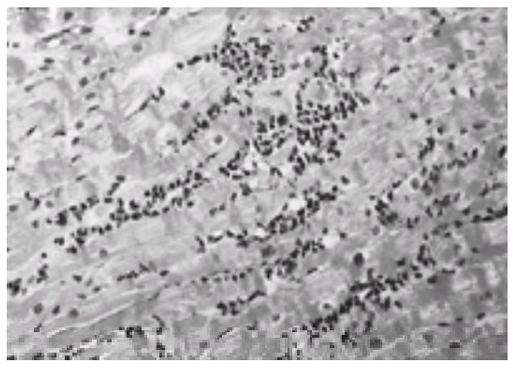

Fig. 1 An example of acute myocarditis, with lymphocytic infiltration adjacent to frayed myocytes.

An algorithm for the diagnosis and treatment of suspected myocarditis is shown in Fig. 2. Spontaneous improvement in left ventricular function can be anticipated in many patients with myocarditis. In most cases it is reasonable to use standard therapy for congestive heart failure, without performing a biopsy or administering steroids, and to observe the patient, using echocardiography to monitor left ventricular function. However, in patients who deteriorate, or who present in cardiogenic shock, an endomyocardial biopsy should be performed. If myocarditis is present, many would regard it as appropriate to administer immunosuppressive therapy, typically beginning with prednisone at 1.25 mg/kg per day, tapering to 0.15 mg/kg per day over 1 month. It must be admitted, however, that the efficacy of such treatment has not been proved.

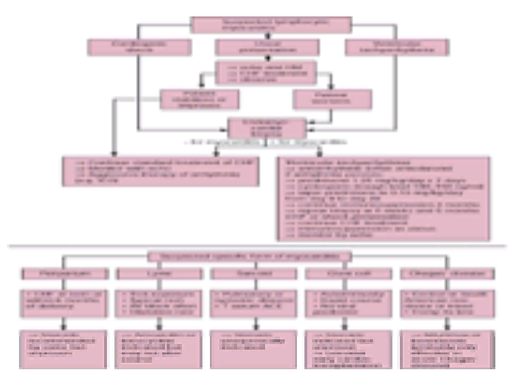

Fig. 2 Algorithm for diagnosis and treatment of suspected myocarditis. Abbrevations: CHF, congestive heart failure; ICD, implantable cardioverter device; ACE, angiotensin-converting enzyme; echo, echocardiogram; HM, Holter monitor.

Ventricular tachyarrhythmias

Lymphocytic myocarditis, with or without a viral prodrome, may present with ventricular tachyarrhythmias and little or no cardiac dilatation and dysfunction. An endomyocardial biopsy should be considered in all cases of ventricular tachycardia of recent onset if no aetiology is apparent, because the presence of myocarditis can substantially change treatment strategy. Since myocarditis is often a self-limiting disorder, the patient's risk of recurrent ventricular tachyarrhythmias may resolve, and it may be unnecessary to subject the patient to electrophysiological study and/or cardioverter-defibrillator implantation. If arrhythmia does not improve spontaneously, a trial of immunosuppressive therapy should be considered. In such cases it is difficult to know how long to continue with anti-arrhythmic drugs. The risks of ventricular arrhythmia should not be underestimated, but nor should those of long-term treatment with agents such as amiodarone. If 24-h ECG monitoring at 6 months shows no sinister abnormalities, then many would withdraw anti-arrhythmic treatment at that point, but others advocate repeat endomyocardial biopsy to document complete resolution of myocarditis before taking this step.

Specific forms of myocarditis

Peripartum myocarditis

Dilated cardiomyopathy developing during the last trimester of pregnancy or within 6 months of delivery is known as peripartum or postpartum cardiomyopathy (see also Chapter 13.6). In some series the dominant cause is myocarditis. When heart failure develops rapidly in the first few weeks after delivery, myocarditis is more likely to be found on endomyocardial biopsy than when the onset is insidious and delayed, and those with early, rapid onset are more likely to recover quickly and completely. While steroid therapy has been used and is recommended by some, its efficacy has not been proved, and spontaneous resolution of peripartum cardiomyopathy is well documented. The usual prohibition against future pregnancy has been debated; it is very clear that some women risk recurrent heart failure, while others do not. In those women in whom severe heart failure persists, cardiac transplantation is an appropriate therapy. After transplantation, successful pregnancies have occurred without recurrence of cardiomyopathy.

Lyme carditis (see Section 15.10 and Chapter 7.11.29)

Borrelia burgdorferi, a spirochaete, infects humans following *Ixodes* tick bites. Lyme disease, which results from this infection, has been reported in 48 of the 50 United States as well as in Europe and Asia. It is characterized by an erythema migrans rash and flu-like symptoms, followed by arthritis, carditis, and neurological disorders in some patients. Carditis is detected in approximately 8 per cent of cases. Both lymphocytic infiltration and the bacterium itself can be demonstrated by endomyocardial biopsy. The usual cardiac manifestation is varying degrees of atrioventricular block. Infrequently, cardiac dilatation occurs. Atrioventricular block is usually transient, though permanent complete heart block has been reported. The site of block appears to be the atrioventricular node in most cases, but block within the His bundle has been documented by electrophysiological study, and the common occurrence of intraventricular conduction delays suggests that bilateral bundle branch block may also occur. Temporary pacing is usually sufficient, though recovery of antegrade conduction may take a week or longer. Lyme carditis should be considered in any case of heart block of unknown cause, especially in young individuals.

Antibiotic therapy (see Chapter 7.11.29) is indicated in Lyme carditis, but it is not known if this alters the course of carditis and atrioventricular block.

Cardiac sarcoidosis

Less than 10 per cent of patients with pulmonary or systemic sarcoidosis have clinically manifest cardiac involvement, ranging from conduction disturbances and arrhythmias to cardiac dilatation. Endomyocardial biopsy reveals typical sarcoid granulomas. The most serious complications of cardiac sarcoidosis are complete heart block, ventricular tachyarrhythmias, and dilated cardiomyopathy. The relatively high incidence of sudden death in patients with sarcoidosis is thought to result from sudden complete heart block or ventricular fibrillation. Patients with sarcoidosis who develop significant conduction disease, arrhythmias, or congestive heart

failure should receive steroids. Occasionally, cardiac involvement will occur without detectable systemic manifestations of sarcoidosis. Thus, cardiac sarcoidosis is in the differential diagnosis of any undiagnosed ventricular arrhythmia, dilated cardiomyopathy, or atrioventricular block.

Giant cell myocarditis

Early recognition of this rapidly progressive form of myocarditis is required, as it has a prognosis considerably worse than that of non-specific lymphocytic myocarditis. The endomyocardial biopsy is distinguished by the presence of multinucleated giant cells and scattered lymphocytic infiltrates with eosinophils. The aetiology of giant cell myocarditis is unknown, but thought to be autoimmune, given its association with myasthenia gravis and other immune disorders, thymoma, and Crohn's disease. It should be suspected in patients, particularly those with a history of an autoimmune condition, who present with disease which progresses unusually rapidly, without viral prodrome, and who do not respond to standard therapy of congestive heart failure. Endomyocardial biopsy should be performed if giant cell myocarditis is suspected, because immunosuppressive therapy appears to be helpful, though not yet proved. Patients with giant cell myocarditis should be considered for early cardiac transplantation if they do not respond to therapy.

Chagas' disease (see Chapter 7.13.11)

Chagas' disease, caused by *Trypanosoma cruzi*, is the leading cause of myocarditis and dilated cardiomyopathy in some Central and South American countries, but uncommon in the United States. Overt acute myocarditis with congestive heart failure, arrhythmias, and conduction disease may develop, but cardiac involvement in early Chagas' disease is usually subclinical. Years later, chronic Chagas' disease may develop and may involve the heart. In the chronic phase, right bundle branch block and biventricular failure are present, and right heart failure predominates. Myocarditis occurs in both the acute and chronic phases, when immune mediation of myocyte injury is well documented. Antiprotozoal treatment with nifurtimox or benznidazole is beneficial in the acute phase. These agents are also indicated in the chronic phase, but, while they do reduce or eliminate serological immune markers of disease, it is not known if they improve outcome.

Further reading

Aretz HT *et al.* (1987). Myocarditis. A histopathologic definition and classification. *Cardiovascular Pathology* **1**, 3–14.

Cooper LT, Berry GJ, Shabetai R (1997). Idiopathic giant-cell myocarditis—natural history and treatment. *New England Journal of Medicine* **336**, 1860–6.

Gauntt CJ *et al.* (1995). Molecular mimicry, antcoxsackievirus B3 neutralizing monoclonal antibodies, and myocarditis. *Journal of Immunology* **154**, 2983–95.

McManus BM *et al.* (1993). Direct myocardial injury by enterovirus: a central role in the evolution of murine myocarditis. *Clinical Immunology and Immunopathology* **68**, 159–69.

McNamara DM *et al.* (1997). Intravenous immune globulin in the therapy of myocarditis and acute cardiomyopathy. *Circulation* **95**, 2476–8.

Mason JW *et al.* (1995). A clinical trial of immunosuppressive therapy for myocarditis. *New England Journal of Medicine* **333**, 269–75.

Matsumori A (1997). Molecular and immune mechanisms in the pathogenesis of cardiomyopathy. Role of viruses, cytokines and nitric oxide. *Japanese Circulation Journal* **61**, 275–91.

Midei MG *et al.* (1990). Peripartum myocarditis and cardiomyopathy. *Circulation* **81**, 922–8.

Rose NR, Hill SL (1996). The pathogenesis of postinfectious myocarditis. *Clinical Immunology and Immunopathology* **80**, S92–S99.

15.8.2 The cardiomyopathies: hypertrophic, dilated, restrictive, and right ventricular

William J. McKenna

Introduction

Heart muscle disease has traditionally been classified as idiopathic or specific. The former, termed the cardiomyopathies, are classified as hypertrophic, dilated, right ventricular, and restrictive. This descriptive classification is useful in relation to natural history, treatment, and prognosis, but recent discoveries of disease-causing mutations in genes encoding sarcomeric contractile proteins in hypertrophic, cytoskeletal proteins in dilated, and a desmosomal protein in right ventricular cardiomyopathy, all indicate that developing knowledge of aetiology/pathogenesis will ultimately require a new classification of the 'idiopathic cardiomyopathies'. The term specific heart muscle disease incorporates myocardial involvement as part of a systemic disease (such as sarcoidosis, systemic hypertension) or when the mechanism of myocardial damage is recognized (such as ischaemia).

Hypertrophic cardiomyopathy

Definition

Hypertrophic cardiomyopathy is defined clinically as an idiopathic heart muscle disorder that is characterized by a hypertrophied and non-dilated left ventricle in the absence of a cardiac or systemic cause. Such a diagnosis of exclusion often presents problems. Does the patient with moderate systemic hypertension and 1.5-cm left ventricular hypertrophy have one or two diseases? Is 1.5-cm hypertrophy a physiological response in a highly trained athlete? Diagnostic uncertainty in the presence of other causes of left ventricular hypertrophy highlights a major limitation of the current definition of hypertrophic cardiomyopathy.

Genetics

Hypertrophic cardiomyopathy is usually familial with autosomal dominant transmission. Clinical presentation with left ventricular hypertrophy under the age of 3 years is usually caused by metabolic or mitochondrial disorders (Table 1) and is unusual in autosomal dominant hypertrophic cardiomyopathy, where morphological and clinical features typically present during or following periods of childhood or adolescent growth. Clinical history reveals familial disease in 40 to 50 per cent, but when cardiovascular evaluation of first-degree relatives includes ECG and echocardiography, 90 per cent of patients have familial disease. Variable expression of the disease is common, even within families bearing the same gene defect.

Mutations in the DNA encoding cardiac b-myosin heavy chain, essential and regulatory myosin light chain, a-tropomyosin, cardiac troponin T and I, cardiac myosin binding protein C, and cardiac actin have been identified in families with hypertrophic cardiomyopathy. Mutations in these genes are found in 60 to 70 per cent of pedigrees evaluated, with those in the b-myosin heavy chain accounting for 20 to 30 per cent, defining hypertrophic cardiomyopathy as a disease of sarcomeric contractile proteins. Most mutations involve a single base-pair change, resulting in amino acid substitutions in exons encoding highly conserved regions. *De novo* mutations occur, but appear to account for 10 per cent of cases or less.

There is allelic heterogeneity with respect to penetrance, morphology, and prognosis. b-Myosin heavy chain mutations that are fully penetrant are associated with worse prognosis (such as Arg403Glu, Arg453Cys), while disease complications are uncommon in patients with mutations that cause mild or no clinical expression (such as Leu908Val). This contrasts with troponin T disease, which may cause premature sudden death in asymptomatic patients who have only minor ECG and/or echocardiographic abnormalities. Troponin T mutations cause 5 to 15 per cent of disease and are associated with incomplete penetrance, disease expression in adolescence, severe myocyte disarray despite mild hypertrophy, and sudden death in adolescents and young adults. Diagnostic difficulties in relatives of sudden death victims who had myocyte disarray with normal or near normal heart weights underscores the importance of implementing a DNA diagnostic test for troponin T. Mutations in myosin binding protein C cause 20 to 30 per cent of disease; most are major deletions rather than single base-pair changes. Myosin binding protein C disease differs in that disease expression occurs later in life, often associated with the recent onset of mild hypertension. However, once disease expression occurs (abnormal ECG and/or echocardiogram) patients may develop symptoms and are at risk from arrhythmia, emboli, and sudden death. Disease caused by the other five recognized gene abnormalities appears to account for less than 10 per cent of all cases of hypertrophic cardiomyopathy. Recently mutations have been identified in the g subunit of AMP dependent protein kinase in families with hypertrophic cardiomyopathy, premature conduction disease, and pre-excitation cosegregating to a locus on chromosome 7.

Pathology

Hypertrophic cardiomyopathy may involve the left or both ventricles. Hypertrophy in the left ventricle is usually asymmetric, involving the anterior and posterior septum and the free wall to a greater extent than the posterior wall. Right ventricular hypertrophy, which is usually symmetric, is seen in over 30 per cent of patients; isolated right ventricular hypertrophy has not been reported. Over 60 per cent of patients have structural abnormalities of the mitral valve, including increased leaflet area,

elongation of the leaflets, and malposition or anomalous insertion of the papillary muscles. Another common macroscopic finding is a patch of endocardial thickening just below the aortic valve, which results from contact of the septum with the anterior mitral leaflet in patients with mitral leaflet abnormalities and/or left ventricular outflow tract obstruction.

The histological findings in hypertrophic cardiomyopathy are distinctive and provide the basis for the pathological diagnosis (Fig. 1). Affected myocardium shows interstitial fibrosis with gross disorganization of the muscle bundles resulting in a characteristic whorled pattern. The cell-to-cell orientation of muscle cells is lost (disarray) and there is disorganization of the myofibrillar architecture within cells. Myocardial cells are wide, short, and often bizarre in shape. Foci of disorganized cells are often interspersed among areas of hypertrophied muscle cells that are otherwise normal in appearance. Such changes are not completely specific: small amounts of fibre disarray may be seen in congenitally abnormal hearts and in secondary left ventricular hypertrophy, and something similar is found at the junction of the septum with the anterior and posterior walls of the left ventricle in normal subjects. However, the extent of myocyte disarray in normal subjects rarely exceeds 5 per cent, whilst in hypertrophic cardiomyopathy up to 40 per cent of the myocardium may be involved. Extensive myocyte disarray is occasionally found in the macroscopically normal heart of a patient who experienced typical clinical features: this highlights the broader phenotype and suggests that hypertrophy may be a secondary rather than a primary abnormality.

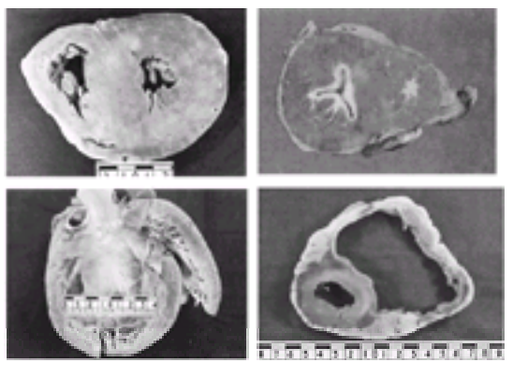

Fig. 1 Transverse short axis section through the ventricles from patients with cardiomyopathy. Upper left shows symmetrical left ventricular hypertrophy in hypertrophic cardiomyopathy. Upper right shows dense white fibrous tissue obliterating the apex of both ventricles in endomyocardial fibrosis. Lower left shows a globular, dilated left ventricle in a child with dilated cardiomyopathy. Lower right shows a grossly dilated right ventricle with adipose infiltration of the right ventricular free wall in arrhythmogenic right ventricular dysplasia. (Reproduced with permission from Davies MJ, 1986, *Colour atlas of cardiovascular pathology*, Oxford University Press.)

Pathophysiology

Disarray

Myocardial disarray and hypertrophy, hyperdynamic systolic function, and impaired diastolic function account for many of the clinical features of hypertrophic cardiomyopathy. The extent and distribution of myocardial disarray can only be determined at autopsy. It is probable that the disorganized architecture with abnormal myofibre and myofibrillar alignment provides a substrate for electrical instability and contributes to diastolic abnormalities, but the precise relationship between myocardial disarray and spontaneous arrhythmia, in particular the threshold for ventricular fibrillation, has not been established.

Diastole

Diastolic abnormalities are common but variable. Typically, left ventricular end-diastolic pressure and atrial pressures are elevated as consequences of abnormal left ventricular diastolic filling and reduced compliance. The isovolumic relaxation time is prolonged, filling is slow, and the proportion of filling volume that results from atrial systolic contraction (while still preserved) may be increased. Occasionally, there is rapid early filling with restrictive physiology that resembles the situation in patients with constrictive pericarditis or endocardial fibrosis (see Chapter 15.9). Altered diastolic function may be caused by myocardial hypertrophy, ischaemia, and architectural abnormalities including myocyte disarray and fibrosis. In an individual patient it is often difficult to identify the major determinant of diastolic disease.

Systole

Most young and some old patients have evidence of hyperdynamic systolic function with rapid, early, and near complete ventricular emptying. Approximately 30 per cent of patients with hyperdynamic systolic function have recordable gradients between the body and outflow tract of the left ventricle at rest; an additional 20 to 25 per cent develop such a gradient following manoeuvres that increase myocardial contractility or result in a decrease in ventricular volume with reduced afterload or venous return. The presence and magnitude of a gradient is determined not only by systolic contractile performance, but also by left ventricular outflow tract size and geometry, which are determined by the extent of upper septal hypertrophy, mitral leaflet morphology, and papillary muscle size and position. The conventionally accepted mechanism of the gradient is that Venturi forces from increased ejection velocity in the narrowed outflow tract draw the anterior and posterior mitral leaflets (which are often large and redundant) toward the septum. However, the significance of such gradients has been controversial. Many workers have claimed that the development of a left ventricular gradient in close temporal association with the development of systolic anterior motion (**SAM**) of the mitral valve and a fall in peak aortic velocity represents impediment or obstruction to left ventricular emptying, but another interpretation of these findings is that they are generated by a dynamic left ventricle that has almost completely emptied. Assessment of the significance of a left ventricular gradient in an individual patient is aided by knowledge of the relative volume ejected by the onset of the gradient. In most patients with resting left ventricular gradients (i.e. 30 to 70 mmHg), at least 70 per cent of stroke volume has already been ejected by the onset of the gradient. By contrast, patients with larger gradients usually have a significant residual volume in the left ventricle at the onset of SAM-septal contact and can be considered to have obstruction.

Ischaemia

Myocardial ischaemia despite normal epicardial coronary arteries is common and caused by several features that relate to myocardial hypertrophy (Table 2). Evidence of ischaemia, however, is not limited to those with severe hypertrophy, and abnormalities of the intramural arteries and of coronary vasomotor behaviour may also be important.

Diagnosis

Hypertrophic cardiomyopathy has been described in Western, African, and Asian populations. The prevalence in adults is estimated to be 1:500 of the population. The diagnosis of hypertrophic cardiomyopathy is based upon the demonstration of unexplained myocardial hypertrophy, which is best done using two-dimensional echocardiography. The diagnosis requires that measurements of wall thickness exceed two standard deviations for gender-, age-, and size-matched populations. In practice, in an adult of normal size, the presence of a left ventricular myocardial segment of 1.5 cm or greater in thickness, in the absence of a recognized cause, is usually considered to be diagnostic. Less stringent criteria should be applied to first-degree relatives of an affected individual, where the probability of carrying the disease gene drops from 1:500 to 1:2. Unexplained symptoms or minor ECG or echocardiographic abnormalities have a high probability of representing disease expression when there is a 50 per cent chance of carrying the gene defect. Modified diagnostic criteria are applied in the context of proven familial disease (Table 3).

In children and adolescents the diagnostic features, particularly the ECG and echocardiographic manifestations of myocardial hypertrophy, often develop during or following growth spurts. The finding of a normal ECG and echo in a child or adolescent reduces the probability of their developing hypertrophic cardiomyopathy, but re-evaluation—ideally annually during adolescence—is warranted because disease expression may not occur until adolescent growth has been completed. In adults, however, the *de novo* development of unexplained left ventricular hypertrophy has so far only been seen with myosin binding protein C disease. Hence an asymptomatic adult with familial disease of onset before age 30 years who has a completely normal ECG and two-dimensional echocardiogram is at very low risk of developing hypertrophic cardiomyopathy and in practical terms does not need re-evaluation in this regard. By contrast, when there is a family history of disease presentation in later life, the possibility of late-onset disease caused by a myosin binding protein C mutation warrants re-evaluation every 3 to 5 years, or sooner if

symptoms or ECG changes develop. This represents a clinical situation where a diagnostic DNA test could be cost-effective and aid clinical management.

Problems in diagnosis often arise in highly trained athletes and in patients with mild hypertension in whom the hypertrophic response appears greater than expected from the apparent stimulus. Competitive athletes normally have an increase in myocardial mass, with maximum increase of 2 to 3 mm in left ventricular wall thickness. The determinants of the hypertrophic response in a patient with hypertension are unknown, but at least in part racially determined; the Afro-Caribbean response appears to be greater than the Caucasian one. In athletes and hypertensive subjects, the diagnosis or exclusion of hypertrophic cardiomyopathy is dependent on the whole clinical picture. An athlete who has 1.5-cm left ventricular hypertrophy with either a small left ventricular cavity or a family history of hypertrophic cardiomyopathy probably does have the condition, whereas an athlete who has negative family history and normal or increased left ventricular cavity dimensions probably does not.

There is the potential for molecular genetic evaluation to provide a 'gold standard' when the clinical diagnosis is equivocal, or when preclinical diagnosis would be of value, for instance in families where there have been multiple sudden deaths in children and adolescents (namely troponin T disease). The finding of a mutation in one of the identified contractile protein genes would confirm disease, but absence of mutation would not exclude the possibility. Identification of the remaining gene(s) for hypertrophic cardiomyopathy would increase the potential value of DNA diagnostic testing.

Though less common than in young children (under 3 years), hypertrophic cardiomyopathy in adults may also be caused by metabolic and mitochondrial diseases. Fabry's disease may account for up to 2 per cent of adult hypertrophic cardiomyopathy: the distinguishing cardiovascular features of this autosomal recessive disorder remain to be fully elucidated, but reduced a-galactacidase enzyme activity is diagnostic. Enzyme replacement therapy is feasible and potentially may improve symptoms and decrease cardiac hypertrophy. Occasionally fatigue or extreme limitation of exercise that is out of proportion to the morphological and haemodynamic severity of the cardiomyopathy are the predominant symptoms. These are characteristics of mitochondrial myopathies and in the absence of severe cardiac failure, sleep apnoea, chronotrophic incompetence, and/or excessive b-blockade should lead to appropriate investigations. In Friedreich's ataxia the cardiac manifestations may precede neurological features by years, necessitating mutation analysis of the frataxin gene for diagnosis. Reduced peak oxygen consumption and early acidification during maximal exercise testing with metabolic gas exchange measurements suggest a coexistent skeletal myopathy. The coexistence with the hypertrophic cardiomyopathy of other somatic features including conduction disease, accessory pathways, eye abnormalities, and diabetes are also suggestive of mitochondrial disease. However, routine neurological examination often fails to elicit abnormalities and muscle biopsy with enzyme analysis is often required for diagnosis.

Clinical features

History

Symptomatic presentation may be at any age with breathlessness on exertion, chest pain, sustained palpitation, syncope, or sudden death. Hypertrophic cardiomyopathy is occasionally found at autopsy in a stillborn baby or presents during infancy with cardiac failure, which is usually fatal. In children and adolescents the diagnosis is most often made during screening of siblings and offspring of affected family members. Paroxysmal symptoms or mild impairment of exercise tolerance are often present, but in the absence of a murmur may not elicit a diagnostic cardiac evaluation. Approximately 50 per cent of adults present with symptoms; in the remainder the diagnosis is made during family screening or following the detection of an unsuspected abnormality on physical, electrocardiographic, or echocardiographic examination.

In adults dyspnoea is common (over 50 per cent) and thought to be a consequence of elevated left atrial and pulmonary capillary wedge pressures resulting from impaired left ventricular relaxation and filling. Approximately 50 per cent of patients complain of chest pain, which is exertional, atypical, or both in similar proportions of patients. Atypical pain may have no obvious precipitant; more commonly it follows exercise or anxiety-related tachycardia, when it persists for up to several hours after the stress has been removed without enzymatic evidence of myocardial damage. Approximately 15 to 25 per cent of patients have experienced syncopal episodes, but in only a minority are there findings suggestive of an arrhythmia or evidence of overt conduction disease: in most patients the mechanism cannot be determined. Patients rarely present with symptoms attributable to left or right heart failure, for example recurrent paroxysmal nocturnal dyspnoea, ascites, or peripheral oedema. Thus, there is a wide spectrum of clinical presentation in hypertrophic cardiomyopathy, from severe cardiac failure in infancy to an incidental finding at any age.

Physical examination

In most patients with hypertrophic cardiomyopathy the physical examination is unremarkable and the detection of abnormalities is dependent on the elucidation of subtle physical signs. There is usually a rapid upstroke arterial pulse, best felt in the carotid area, which reflects dynamic left ventricular emptying. Most patients also have a forceful left ventricular cardiac impulse, best appreciated on full-held expiration in the left lateral position. In about one-third of patients the jugular venous pulse may demonstrate a prominent 'a' wave, reflecting diminished right ventricular compliance secondary to right ventricular hypertrophy. The first and second heart sounds are usually normal, and—unless patients are in atrial fibrillation—there is either a loud fourth heart sound, reflecting increased atrial systolic flow into a non-compliant ventricle, or a palpable atrial beat reflecting forceful atrial systolic contraction that may or may not be associated with significant forward flow of blood.

The most obvious physical sign in hypertrophic cardiomyopathy is an ejection systolic murmur present in those patients (20 to 30 per cent) who have a resting left ventricular outflow tract gradient. This murmur starts well after the first heart sound and ends well before the second. It is best heard at the left sternal border, radiating towards the aortic and mitral areas but not into the neck or the axilla. The intensity varies with changes in ventricular volume; it can be increased by physiological and pharmacological manoeuvres that decrease afterload or venous return (amyl nitrate, standing, Valsalva), and decreased by manoeuvres that increase afterload and venous return (squatting, phenylephrine). Occasionally, ejection systolic murmurs are associated at their onset with an ejection sound.

Most patients with a left ventricular gradient also have mitral regurgitation, which may be difficult to distinguish by auscultation. Doppler examination reveals that mitral regurgitation usually begins just before (30 to 40 ms) the onset of the gradient and continues for the duration of systole. Radiation of the systolic murmur to the axilla is often the best auscultatory clue to the presence of coexistent mitral regurgitation, which may be moderate to severe, either alone or in association with a left ventricular outflow tract gradient. A mid-diastolic rumble may sometimes result from increased transmitral flow in patients with severe mitral regurgitation; more commonly it occurs in isolation, presumably reflecting inflow tract turbulence.

Early diastolic murmurs of aortic incompetence may develop following surgical myotomy/myectomy or infective endocarditis involving the aortic valve. Although such murmurs are rare in the absence of such complications, they appear to occur more commonly than would be expected by chance and may reflect traction on the non-coronary cusp of the aortic valve by the septum. An ejection systolic murmur in the pulmonary area, reflecting right ventricular outflow tract obstruction, is also rare; when present it is usually associated with severe biventricular hypertrophy in the young or in those with coexistent Noonan's syndrome and a dysplastic pulmonary valve (see Chapter 15.12).

Prognosis

Patients with hypertrophic cardiomyopathy experience slow progression of symptoms, gradual deterioration of left ventricular function, and a significant incidence of sudden death, which occurs at all ages. Referral centre data from the 1970s and 1980s reveal an annual mortality from sudden death of 2 to 3 per cent in adults and 4 to 6 per cent in children and adolescents. It is even greater in young patients with recurrent syncope or a family history of 'malignant' hypertrophic cardiomyopathy. Although the mortality figures from non-referral hospitals are lower, the risk of sudden death is still present. More recent data in managed patients reveal annual mortality rates related to sudden death and disease of 1 and 2 per cent, respectively.

Symptomatic deterioration is usually slow. However, severe symptoms may develop in association with progressive myocardial wall thinning, presumably reflecting myocyte necrosis or fibrosis and severe reduction in left ventricular systolic performance and/or diastolic filling. Patients who experience such deterioration occasionally present with a clinical picture resembling restrictive cardiomyopathy with grossly enlarged atria, signs of right heart failure, and relative preservation of left ventricular systolic performance.

Atrial dilatation and the development of atrial fibrillation/flutter are important features in the clinical course, representing a risk of embolic stroke as well as of acute and/or chronic deterioration. Earlier onset of atrial fibrillation was considered to be an ominous development but is part of the evolution of patients with diastolic dysfunction and with appropriate management need not represent a major cause of morbidity or mortality. The largest study of patients with atrial fibrillation revealed that their 5-year survival was similar to that of age- and sex-matched patients who remained in sinus rhythm and, if the ventricular response was controlled,

symptomatic status remained stable.

Left ventricular hypertrophy develops during childhood and adolescence but is not progressive in adults. The trigger and other determinants of disease expression in late-onset myosin binding protein C disease are uncertain, but (as with the other disease-causing genes) left ventricular hypertrophy does not appear to be progressive beyond the limited phase of disease expression.

Investigations

Cardiological evaluation of patients with hypertrophic cardiomyopathy is performed to confirm or make the diagnosis, to characterize the functional and morphological features to guide symptomatic therapy, and to assess the risk of complications, particularly that of sudden death.

Electrocardiography

The 12-lead electrocardiogram is the most sensitive diagnostic test, although occasionally normal (around 5 per cent), particularly in the young. Five to ten per cent are in atrial fibrillation at the time of diagnosis. Many patients have an intraventricular conduction delay, 20 per cent have left axis deviation, while complete right bundle or left bundle branch block are uncommon (less than 5 per cent). The latter may develop following surgery and is occasionally seen in the elderly. ST-segment depression and T-wave changes are the most common abnormalities and are usually associated with voltage changes of left ventricular hypertrophy and/or deep S waves in the anterior chest leads V1–V3. Isolated repolarization changes or giant negative T waves are occasionally seen. Voltage criteria for left ventricular hypertrophy are rare in the absence of repolarization changes. Approximately 20 per cent of patients have abnormal Q waves, either inferiorly (II, III, and aVF), or less commonly in leads V1–V3. P-wave abnormalities of left and/or right atrial overload are common. The distribution of the P–R interval is similar to that in the normal population, but occasionally a short P–R interval may be associated with a slurred upstroke to the QRS complex, similar to that seen in the Wolff–Parkinson–White syndrome. At electrophysiological study such changes are not usually associated with evidence of pre-excitation, although patients with hypertrophic cardiomyopathy and accessory pathways have been described. Despite the many electrocardiographic abnormalities, there is no electrocardiogram that is typical of hypertrophic cardiomyopathy; a useful rule is to consider the diagnosis whenever the electrocardiogram is bizarre, particularly in younger patients.

The incidence of arrhythmias during 48-hour ambulatory electrocardiographic monitoring increases with age (Fig. 2). Non-sustained ventricular tachycardia is detected in 20 to 25 per cent of adults and, although usually asymptomatic, is associated with an increased risk of sudden death. Supraventricular arrhythmias are also common in adults: these are poorly tolerated if sustained (more than 30 s)—unless the ventricular response is controlled—and they carry an increased risk of embolism. By contrast, most children and adolescents are in sinus rhythm and arrhythmias during ambulatory electrocardiographic monitoring are uncommon (Fig. 2). The increased incidence of supraventricular arrhythmias with age is not surprising: their development is related to increased echocardiographic left atrial dimensions and increased left ventricular diastolic pressure, both of which increase with age. The aetiology of ventricular arrhythmias is not known, but may relate to myocyte loss and myocardial fibrosis, which appear to be related to age. Documented sustained ventricular tachycardia is uncommon, but a recognized complication in patients with an apical outpouching or aneurysm, which may develop as a consequence of mid-ventricular obstruction.

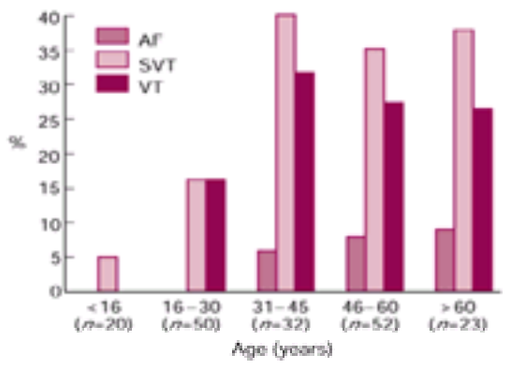

Fig. 2 The relation of arrhythmias and age in hypertrophic cardiomyopathy. SVT, supraventricular tachycardia; VT, ventricular tachycardia; AF, atrial fibrillation.

Imaging

Chest radiography

The chest radiograph may be normal or show evidence of left and/or right atrial or left ventricular enlargement; if left atrial pressure has been chronically elevated, there may be evidence of redistribution of blood flow to upper lung zones. Mitral valve annular calcification is seen, particularly in the elderly.

Echocardiography

The extent and severity of myocardial hypertrophy is best evaluated with two-dimensional echocardiography/Doppler (Fig. 3). Left ventricular hypertrophy may be symmetric or asymmetric and localized to the septum or the free wall, but most commonly to both the septum and free wall with relative sparing of the posterior wall. 'Apical' hypertrophic cardiomyopathy appears to be common in Japan, but is rare in the West, although approximately 10 per cent of patients have left ventricular hypertrophy that is maximal in the distal ventricle from the level of the papillary muscles down to the apex. Approximately one-third of patients also have hypertrophy of the right ventricular free wall, the presence and severity of which is strongly related to the severity of left ventricular hypertrophy. Typically, left ventricular end-systolic and end-diastolic dimensions are reduced, and the left atrial dimension is increased. Indices of systolic function such as ejection fraction may be increased, but systolic function is often impaired, which may be best appreciated by measurement of long axis rather than short axis function.

Fig. 3 An echocardiogram (parasternal long axis view) of a patient with hypertrophic obstructive cardiomyopathy demonstrating hypertrophy of the interventricular septum (IVS), enlargement of the left atrium (LA), and systolic anterior motion of the mitral valve, bringing it into contact with the septum (arrow).

Colour Doppler provides a sensitive method of detecting left ventricular outflow tract turbulence (Fig. 4 and Plate 1), and when combined with continuous wave Doppler the peak velocity (*V*max) of left ventricular blood flow can be measured and left ventricular outflow tract gradients calculated. Doppler-calculated gradients (pressure gradient (mmHg) = $4V\text{max}^2$) are seen in 20 to 30 per cent of patients and correlate well with those measured invasively. Systolic anterior motion of the mitral valve is usually present when the calculated outflow tract gradient is more than 30 mmHg. Early closure or fluttering of the aortic valve leaflets and Doppler evidence of mitral regurgitation are often seen in association with systolic anterior motion of the mitral valve. A posteriorly directed mitral regurgitant jet is seen in association

with and related to the magnitude of the outflow tract gradient (Fig. 4). An anterior regurgitant jet or mitral regurgitation in the absence of obstruction suggest the coexistence of structural mitral valve abnormalities.

Fig. 4 Colour flow Doppler image (parasternal long axis view) of the same patient as shown in Fig. 3, demonstrating left ventricular outflow tract (LVOT) turbulence and mitral regurgitation (MR) with a posteriorly directed jet. (See also Plate 1.)

Other imaging techniques

There is no role for routine magnetic resonance or computed tomographic imaging, but these modalities may be helpful when echocardiography fails to visualize cardiac structures adequately, particularly the left ventricular apex.

Cardiac catheterization

Two-dimensional echo/Doppler evaluation has replaced invasive haemodynamic measurements and angiography as the method of assessing left ventricular structure and function in hypertrophic cardiomyopathy. Cardiac catheterization is not necessary for diagnosis and is rarely indicated unless symptoms are refractory and direct measurement of cardiac pressures is potentially informative, particularly in assessing the severity of mitral regurgitation. Coronary arteriography may be necessary to exclude coexistent coronary artery disease in older patients who have significant angina or ST-segment changes during exercise. The left coronary arteries are usually large in calibre. The left anterior descending and septal perforator arteries may demonstrate phasic narrowing during systole in the absence of fixed obstructive lesions, but such changes do not appear to relate to symptoms.

Left ventricular angiography is rarely indicated, but recognition of the abnormally shaped ventricle, which typically ejects at least 75 per cent of its contents in association with mild mitral regurgitation, may provide a valuable diagnostic clue when hypertrophic cardiomyopathy was not suspected before catheterization.

Exercise testing

Maximal exercise testing in association with respiratory gas analysis provides useful functional and prognostic information, which can be monitored serially. Peak oxygen ventilatory capacity (peak $\dot{V}O_2$) is often moderately and occasionally severely reduced, even in patients who claim their exercise tolerance is not limited. Continuous measurement of the blood pressure during upright treadmill or bicycle exercise reveals that approximately one-third of younger patients (less than 40 years) have an abnormal blood-pressure response, with either drops of more than 20 mmHg from peak recordings or a failure to rise by 20 mmHg or more despite an appropriate increase in cardiac output. Such changes are usually asymptomatic but are associated with an increased risk of sudden death. The mechanism of the hypotensive response during exercise in hypertrophic cardiomyopathy is uncertain, but may relate to myocardial mechanoreceptor activation and altered baroreflex control causing inappropriate drops in systemic vasculature resistance despite maintenance of an appropriate cardiac output. ST-segment depression of 2 mm from baseline is documented in 25 per cent of patients. The relation of such changes to metabolic markers of ischaemia requires further evaluation and their prognostic significance has yet to be determined.

Electrophysiological studies

Electrophysiological studies may occasionally be necessary in patients with sustained, rapid palpitation to identify associated accessory pathways or aid management of sustained monomorphic ventricular tachycardia. Conventional programmed ventricular stimulation does not aid the identification of high-risk patients (see below).

Management

Pharmacological

The goal of therapy is to improve symptoms and prevent complications, in particular sudden death. b-Adrenoceptor blockers, particularly propranolol, and calcium antagonists, especially verapamil, are the mainstay of symptomatic pharmacological therapy. Both drugs have several potentially beneficial actions, including a decrease in the determinants of myocardial oxygen consumption and blunting of the heart-rate response during exercise, providing increased time for filling at equivalent workloads in those with poor relaxation and slow filling. Both agents exert a negative inotropic effect, thereby reducing hyperdynamic systolic function and left ventricular gradients; it is also claimed they improve diastolic filling, verapamil by improving relaxation and propranolol by increasing compliance. The side-effects of propranolol are rarely serious, but the suppressant effect of verapamil on atrioventricular nodal conduction may cause problems in patients with unsuspected pre-existing conduction disease, and its vasodilatory and negative inotropic effects have resulted in acute pulmonary oedema and death. In practice, both drugs are effective but it is safer to use propranolol. If this is ineffective, verapamil can then be tried, but it should be started in hospital in patients with conduction abnormalities, resting or provocable gradients, or impaired systolic function.

Surgical

Surgery is a therapeutic option in patients with obstruction and/or mitral valve abnormalities. The conventional indication for surgery has been a resting left ventricular outflow tract gradient of more than 50 mmHg in patients refractory to medical therapy, and the commonest operation has been to remove a segment of the upper anterior septum (myotomy/myectomy) via a transaortic approach. Transventricular approaches have been used, but these are associated with a higher incidence of late complications, particularly of cardiac failure. Mitral valve 'repair' and papillary muscle repositioning or remodelling may be required, and mitral valve replacement has also been advocated; excellent results have been achieved, particularly in elderly patients with severe mitral regurgitation. Specialist hypertrophic cardiomyopathy centres report perioperative mortality of 2 per cent or less, with 80 per cent success in abolishing gradients and improving symptoms. It is unlikely, however, that such excellent results are obtained in centres with more limited experience of the medical and surgical aspects of the condition. Obstruction is a function of several features including septal hypertrophy, outflow tract dimensions, ventricular geometry and flow patterns, mitral valve/papillary muscle anatomy and position, and left ventricular contractile performance. The approach to gradient reduction needs to be individualized to a greater extent than has been recognized, perhaps contributing to mortality rates of 5 to 10 per cent in early operative series.

Pacing

The pacing option was promoted following recognition that alteration of the ventricular activation sequence, with optimization of filling characteristics by DDD pacing, may result in reduction of gradients and filling pressures and improved symptoms in selected patients. The role of DDD pacing in symptomatic management of obstruction was evaluated in three randomized multicentre trials, demonstrating symptomatic improvement and gradient reduction (50 per cent), but no change in exercise capacity. However, the placebo effect of the procedure was considerable: 40 per cent reported significant symptomatic improvement with the pacemaker programmed to a standby mode. Overall, the initial enthusiasm for DDD pacing has not been substantiated by greater experience and trials. Nevertheless, pacing offers a therapeutic option in patients with obstruction that is refractory to drug treatment, and in whom surgery is either not acceptable or inappropriate. It appears

that elderly patients with localized septal hypertrophy and without significant free wall involvement or mitral regurgitation may do particularly well.

Other techniques

Injection of alcohol into the septal artery that supplies the 'obstructing' septal muscle has been developed as a percutaneous, non-pharmacological approach to gradient reduction. Most experienced centres have reported excellent results. As for surgery and DDD pacing, patient selection—in particular regarding the mechanism of the gradient—and technical considerations are important determinants of outcome. The major complication has been the need for a pacemaker in 5 to 10 per cent, and concerns remain about long-term left ventricular function and arrhythmia risk from the 'controlled myocardial infarction'. At present alcohol septal ablation offers a therapeutic option in older patients with suitable anatomy who are refractory to drugs.

Particular symptoms

Dyspnoea

Dyspnoea most often occurs in patients who also experience chest pain or discomfort. Treatment depends on the predominant mechanism. In patients with dyspnoea who have slow filling that continues throughout diastole, b-blockers and verapamil are appropriate. Conversely, those with rapid early filling may benefit from a relative tachycardia and do better without negative chronotrophic agents. When dyspnoea is associated with significant obstruction, that is, at least 50 per cent of stroke volume in the left ventricle at the onset of the gradient, b-blockers, disopyramide, and—failing this—myotomy/myectomy or the other non-pharmacological options may be beneficial. Disopyramide should be used in the maximum tolerated dose (anticholinergic side-effects may limit higher doses) in conjunction with a conventional b-blocker. Occasionally, dyspnoea is associated with severe mitral regurgitation and responds well to mitral valve replacement. Endocarditis is a rare complication of hypertrophic cardiomyopathy; it occurs in patients with left ventricular outflow tract turbulence and/or mitral regurgitation, may involve the mitral and/or aortic valve, and is usually associated with increased dyspnoea. Antibiotic prophylaxis is important in appropriate patients, such as those with intracardiac transvalvular turbulence.

Chest pain

When chest pain is severe, associated with significant ST-segment changes during exercise, or refractory to therapy, the performance of coronary arteriography is warranted to exclude coexistent coronary artery disease. The results of coronary artery bypass grafting in hypertrophic cardiomyopathy are good, even when additional procedures such as myectomy/mitral valve replacement are performed. Exertional chest pain usually responds to therapy with propranolol or verapamil, and when refractory has responded to very high doses of these agents (propranolol at 480 mg daily, verapamil at 720 mg daily). Short-acting nitrates, diuretics, and high-dose verapamil may be useful in selected patients, perhaps by reducing filling pressures and improving coronary flow to subendocardial layers. Atypical chest pain may persist long after the initial stimulus has been removed.

Arrhythmia

Arrhythmias are a common complication of hypertrophic cardiomyopathy. Treatment with anticoagulants and digoxin, with or without verapamil or b-blockers, is appropriate once atrial fibrillation is established. The aim of therapy is to control the ventricular response and prevent emboli. Most patients who develop atrial fibrillation during electrocardiographic monitoring are unaware of changes from sinus rhythm to atrial fibrillation as long as the ventricular response is well controlled. However, in a few cases the loss of atrial systolic contribution to filling volume is important, when electrical cardioversion can be facilitated by prior therapy (4 to 6 weeks) with amiodarone (300 mg daily) if pharmacological cardioversion does not occur first.

Sustained (more than 30 s) episodes of paroxysmal atrial fibrillation or supraventricular tachycardia occur, representing a risk of haemodynamic collapse and emboli. Low-dose amiodarone (1000 to 1400 mg weekly) is effective in suppressing such episodes and also provides control of the ventricular response should breakthrough occur. If episodes persist, the threshold for anticoagulation should be low as embolic complications are common, even when atrial dimensions are only moderately increased.

Non-sustained episodes of supraventricular arrhythmia are common. Though often asymptomatic they are a marker (albeit of low positive predictive accuracy) for the subsequent development of established atrial fibrillation. If they occur in the presence of atrial enlargement, the threshold to introduce amiodarone with or without anticoagulation should be low. Episodes of non-sustained ventricular tachycardia are common but are rarely symptomatic: therapy is warranted only if it can be shown to improve prognosis (see below).

Prevention of sudden death

Sudden death is probably a consequence of multiple interacting mechanisms. The histological abnormalities, particularly myocyte disarray, small vessel disease, and replacement scarring, contribute to the underlying substrate. Events may be triggered by haemodynamic alterations, myocardial ischaemia, and arrhythmias, including ventricular tachycardia, atrial fibrillation, atrioventricular block, and rapid conduction of a supraventricular arrhythmia via an accessory pathway. Intense physical exertion may also contribute to the above triggers. The interaction of triggers and substrate may be modified by inappropriate peripheral vascular responses and the development of ischaemia.

Risk factor stratification

Prevention of sudden death relies on risk factor stratification to identify the high-risk cohort. Several adverse features, which can be elicited from the clinical history and non-invasive evaluation, have been identified (Table 4). Their relative importance varies with age; for example, the finding of non-sustained ventricular tachycardia on 24-h electrocardiographic monitoring in children and adolescents is uncommon (less than 5 per cent), but is associated with an eightfold increased risk of sudden death, whereas in adults this arrhythmia is common (20 to 25 per cent), but in isolation confers only a twofold increased risk.

In the young (less than 25 years) the finding of non-sustained ventricular tachycardia, severe and diffuse left ventricular hypertrophy, unexplained syncope (particularly if recurrent or exertional), or a family history where a high proportion of affected individuals experienced premature (less than 40 years) sudden death warrants prophylactic treatment. Such patients usually also exhibit abnormal exercise blood-pressure responses, indeed the finding of a normal exercise blood-pressure response appears to identify the low-risk younger (less than 40 years) patient (negative predictive accuracy 97 per cent), allowing appropriate reassurance that is also clinically important. In adults aged 25 to 60 years the positive predictive accuracy for each of the risk factors is lower (15 to 20 per cent): in general prophylactic treatment is reserved for those with two or more risk factors who will have a predicted risk of sudden death of at least 2 per cent per year.

It is important to consider risk in all patients, even those who are asymptomatic or who have mild echocardiographic features of hypertrophic cardiomyopathy. Though children and adolescents with severe congestive symptoms may be at greater risk, the data reveals that the severity of chest pain, dyspnoea, and exercise limitation are not reliable predictors of the risk of sudden death in adults. In addition it is recognized that most patients who die suddenly have mild (1.5 to 2.0 cm) or moderate (2.0 to 2.5 cm) left ventricular hypertrophy, while some genetic defects (for instance cardiac troponin T) may cause sudden death in the absence of symptoms or hypertrophy. The presence of a left ventricular outflow tract gradient is not associated with sudden death, although data on patients with large gradients (more than 100 mmHg) are limited. Diastolic impairment with abnormal Doppler filling patterns and atrial enlargement is associated with symptomatic limitation and poor prognosis, but not with premature sudden death.

Some investigators have suggested that the induction of sustained ventricular arrhythmias during programmed electrophysiological stimulation is associated with a higher risk of sudden death. However, the predictive accuracy is low, and as most high-risk patients can be identified using non-invasive clinical markers, the inherent risks and inconvenience associated with programmed stimulation dictate that it should not be used routinely to assess risk in hypertrophic cardiomyopathy.

Dilated cardiomyopathy

Definition

Dilated cardiomyopathy is characterized by unexplained dilatation and impaired contractile performance of the left ventricle. Potential causes of ventricular

dysfunction, particularly coronary artery disease and systemic hypertension, must be excluded for the diagnosis to be made. Typical angina pectoris, fluctuating ST and T-wave changes, and regional abnormalities on two-dimensional echocardiography or thallium scintigraphy, which reflect damage to a specific vascular territory, suggest ischaemic heart disease. Renal and ocular hypertensive changes may provide a useful marker of previous systemic hypertension, but are often unremarkable in the decompensated phase in the normotensive patient. Calcific aortic stenosis may be overlooked as a cause of heart failure, particularly when the murmur is soft or absent.

Specific heart muscle disorders should also be considered in differential diagnosis (see Chapter 15.8.3). A primary cardiac presentation of diabetes mellitus, connective tissue disorders, and neuromuscular disease is rare, but arrhythmias or progressive conduction disturbance with mild left ventricular dysfunction may provide the earliest evidence of cardiac sarcoidosis.

Since the definition of dilated cardiomyopathy is a diagnosis of exclusion, it is likely that structural and functional abnormalities result from heterogeneous pathogenic processes. In North America and Europe symptomatic dilated cardiomyopathy has an incidence and prevalence of 20 and 38 per 100 000, respectively, and is the commonest indication for cardiac transplantation.

Genetics

Pedigree analysis reveals familial disease in at least 25 per cent of cases and an additional cohort (10 to 20 per cent) with mild abnormalities of left ventricular performance who possibly have early presymptomatic dilated cardiomyopathy. Inheritance is usually autosomal dominant with incomplete penetrance, although families with X-linked transmission have been reported. Different patterns of disease expression are recognized. Disease progression appears to be slow (over decades) in most cases, and conduction disturbance is a late complication related to disease severity. However, in some families (less than 20 per cent) the early stages are characterized by progressive conduction disease, and left ventricular dilatation and impairment are later manifestations, which typically occur in the 4th to 6th decade. Families are also recognized in whom dilated cardiomyopathy develops in later decades in individuals who have had sensorineural hearing loss since childhood, or in association with skeletal myopathy.

Penetrance is age dependent and has been estimated to be 10 per cent in those aged less than 20 years, 34 per cent in young adults aged 20 to 30 years, 60 per cent in adults aged 30 to 40 years, and 90 per cent in those over 40 years. Familial evaluation is recommended: guidelines have been proposed, based on the identification of major and minor criteria for the diagnosis (Table 5). It has been proposed that the diagnosis of familial dilated cardiomyopathy is fulfilled in a first-degree relative of a proband in the presence of one major criterion, or left ventricular dilatation plus one minor criterion, or three minor criteria.

Disease-causing genes have been reported in dystrophin (X-linked), taffazin (X-linked Barth syndrome), metavinculin (X-linked), cardiac actin (two small autosomal dominant families), lamin A/C (families with premature conduction disease), and desmin (1 of 44 probands with gene-positive affected family members). Lamin A/C mutations may also cause Emery–Dreifuss and limb girdle muscular dystrophy and familial partial lipodystrophy, desmin may cause conduction disease, and dystrophin mutations cause childhood (Duchenne) and adult (Becker) forms of muscular dystrophy. Mutations in these genes appear to account for less than 5 per cent of cases of dilated cardiomyopathy, but they provide potentially valuable clues in relation to aetiology and pathogenesis. The function of taffazin is unknown, but the other genes all encode cytoskeletal proteins that are involved in force transmissions between cells.The recent identification of mutations in genes which encode for sarcomere proteins (TropT, bMHC) involved in force generation may change this paradigm.

Pathogenesis

Based on the molecular genetic findings described above, dilated cardiomyopathy is hypothesized to be a disease of the cardiac cytoskeleton, but pathogenesis is poorly defined. Macroscopic examination of hearts with dilated cardiomyopathy taken at autopsy or explanted reveals dilated cardiac chambers (Fig. 1), mural thrombi, and platelet aggregates with normal extra- and intramural coronary arteries. Histology shows features consistent with healed myocarditis—patchy perimyocyte and interstitial fibrosis, various stages of myocyte death, as well as myocyte hypertrophy and rare isolated inflammatory cells (see Chapter 15.8.1). These postinflammatory findings are non-specific and do not suggest a particular pathogenesis.

The myocardial depressant effects of alcohol in normal and diseased myocardium are established. Alcohol, like pregnancy, may precipitate cardiac failure in predisposed individuals, but an additional specific aetiological or pathogenetic role remains uncertain. Viral involvement is supported by viral myocarditis progressing to dilated cardiomyopathy in specific genetic strains of a murine model, as well as in isolated rare patients, also by an association with abnormal coxsackievirus serology and hybridization studies that show non-replicating enteroviral genome in a variable proportion (0 to 30 per cent) of dilated cardiomyopathy hearts. The potential for immune pathogenesis is supported by development of autoimmune murine myocarditis and by the findings of a cardiac- and disease-specific autoantibody in over 30 per cent of patients with dilated cardiomyopathy and their first-degree relatives, inappropriate MHC class II expression on endothelial cells from cardiac tissue, and a weak HLA DR4 association.

In summary, pathogenesis of dilated cardiomyopathy remains controversial, with resolution hampered by clinical presentation at 'endstage' when pathogenesis may be largely completed. A reasonable working hypothesis proposes an immune pathogenesis, with or without a viral trigger, in genetically predisposed individuals who carry a cytoskeletal or sarcomere gene mutation.

Clinical features

Dilated cardiomyopathy has been described in Western, African, and Asian populations, affecting both genders and all ages. Initial presentation is usually with symptoms of cardiac failure (fatigue, breathlessness, decreased exercise tolerance), but an arrhythmia (atrial fibrillation, ventricular tachycardia, atrioventricular block), a systemic embolus, or the finding of an electrocardiographic or radiographic abnormality during routine screening may prompt earlier diagnosis. Symptoms, physical signs, and chest radiographic changes are those of cardiac failure (see Chapter 15.2.2) and depend on the stage of the disease.

Physical examination may be entirely normal or may reveal evidence of myocardial dysfunction with cardiac enlargement and signs of congestive heart failure. Systolic blood pressure is usually low with a narrow pulse pressure and a low volume arterial pulse. In patients with severe left ventricular failure, pulsus alternans may be present and the jugular veins may be distended, with a prominent V wave reflecting tricuspid regurgitation. In such patients the liver is often engorged and pulsatile, and there is usually peripheral oedema and ascites. The precordium often reveals a diffuse and dyskinetic left and occasionally a right ventricular impulse. The apical impulse is usually displaced laterally, reflecting ventricular dilatation. The second heart sound is usually normally split, but paradoxical splitting may be present when there is left bundle branch block, which occurs is approximately 15 per cent of patients. With severe disease and the development of pulmonary hypertension, the pulmonary component of the second heart sound may be accentuated. Characteristically, a presystolic gallop or fourth heart sound is present before the development of overt cardiac failure. However, once cardiac decompensation has occurred, ventricular gallop or third heart sound is often present. When there is significant ventricular dilatation, systolic murmurs are common, reflecting mitral and (less commonly) tricuspid regurgitation.

The development of unexplained cardiac failure during pregnancy or within the 3 months following parturition is often labelled as peripartum cardiomyopathy. Unrecognized pre-eclamptic heart disease may also present with cardiac failure and should be excluded by careful examination of the antenatal records; this has a different prognosis and recurs with increasing severity during subsequent pregnancies unless treated. Antecedent cardiac evaluation is often absent in those with peripartum cardiomyopathy, and there is usually uncertainty whether the cardiac failure is acute (for instance potentially myocarditic) or chronic and exacerbated by the haemodynamic stress of pregnancy and labour (for instance dilated cardiomyopathy). When the heart failure is acute and there is persistence of left ventricular chamber dilatation or impaired systolic performance, the diagnosis of peripartum cardiomyopathy can legitimately be made. The mechanism and true natural history is uncertain, though it is probable that the adverse prognostic effect of subsequent pregnancies is less important than the literature would suggest, particularly in those with only mild residual abnormalities of left ventricular structure and function. For further discussion of cardiac disease in pregnancy, see Chapter 13.6.

Arrhythmia

Atrial arrhythmias and particularly atrial fibrillation are common and are associated with the severity of symptoms, left ventricular dysfunction, and poor prognosis. Atrial fibrillation is a marker of disease severity, but not an independent predictor of disease progression or sudden death. Occasionally, however, focal atrial tachycardia or atrial fibrillation may cause a tachycardia that results in gradual deterioration in left ventricular function, resembling dilated cardiomyopathy. Systolic function usually returns to normal with control of the arrhythmia.

Ventricular arrhythmias are also common and like supraventricular arrhythmias are markers of disease severity. Non-sustained ventricular tachycardia during ECG

monitoring is seen in approximately 20 per cent of asymptomatic or mildly symptomatic patients and in up to 70 per cent of those who are severely symptomatic. The prognostic significance of this arrhythmia is controversial: its presence early in the course of disease, when left ventricular function is relatively preserved, is probably an independent marker of sudden death risk, whereas in general, markers of haemodynamic severity (such as ejection fraction, left ventricular end-diastolic dimension, filling pressures) are more predictive of disease-related mortality and sudden death. Sudden death risk in patients with severe disease (NYHA Class III, IV) increases approximately threefold when syncope is present.

Prognosis

The prognosis of dilated cardiomyopathy is uncertain because the diagnosis is usually not made until clinical features, which are late manifestations of the disease, become obvious. However, recent follow-up of a large cohort of asymptomatic first-degree relatives suggests that disease progression is insidious over decades. An upper respiratory tract infection or a salt or fluid load (pregnancy) often precipitates clinical presentation. Symptoms develop when filling pressures rise or stroke volume diminishes sufficiently to cause salt and water retention and oedema. Once clinical symptoms of impaired ventricular performance are apparent, prognosis is poor and related to the degree of left ventricular dilatation and impaired contractile performance. Data from adult and paediatric referral centres in the 1970s and 1980s indicate 50 per cent mortality from progressive heart failure or its complications in the 2 years following referral diagnosis. Survival will undoubtedly be improved by recognition of asymptomatic family members as well as of others with early/mild disease, and by modern management including the early introduction of ACE inhibitors, b-blockade, aggressive treatment of arrhythmias, and the availability of cardiac transplantation. Treatment will usually stabilize or improve the patient's condition once symptoms develop, with a reduction in cardiac dimensions and improvement in myocardial performance. However, conventional evaluation of cardiovascular structure and function does not permit accurate prediction of outcome and there is an annual mortality of up to 4 per cent, predominantly from sudden death, even in those who improve or stabilize.

The recent recognition of the familial nature of dilated cardiomyopathy indicates that a correct diagnosis is of practical importance for family members as there is now the potential to identify individuals with dilated cardiomyopathy at an early or preclinical stage.

Investigation

Cardiological evaluation of patients with dilated cardiomyopathy is performed to confirm the diagnosis, to determine objective measurements of functional capacity as a guide to symptomatic treatment, and to assess risk of complications, particularly progressive deterioration, arrhythmias, and sudden death.

By the time of diagnosis in a referral centre, a normal ECG is rare (less than 5 per cent) and most patients show features consistent with diffuse myocardial abnormalities. Twenty per cent are in established atrial fibrillation and paroxysmal supraventricular and ventricular arrhythmias during 24-h ECG monitoring are common.

Patient perception of functional limitation relates to many factors, including the time course of the illness, and is very variable. Maximal exercise testing, ideally with respiratory gas analysis, provides a simple reproducible measure of functional capacity and is also useful to exclude ischaemia and assess risk of arrhythmia. Similarly two-dimensional echocardiography is important in providing an easily repeated measure of cardiac cavity dimensions and systolic performance, assessment of regional wall motion as well as mural and intracavitary thrombi (Fig. 5 and Fig. 6 and Plate 2). In the young patient the origins of the right and left main coronary arteries can often be visualized to exclude mainstem coronary anomalies as the cause of myocardial dysfunction. Symptoms, exercise testing, and two-dimensional echocardiography provide the basis for assessment of treatment and monitoring of disease progression.

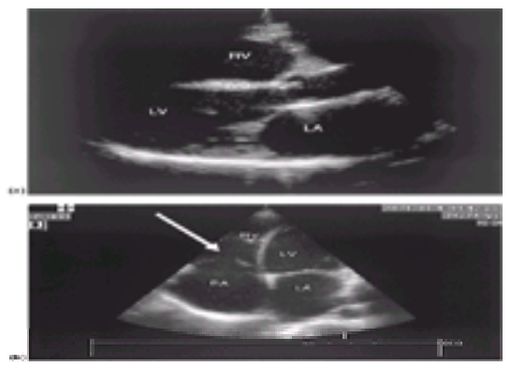

Fig. 5 Echocardiographic appearances of a young patient with familial dilated cardiomyopathy. Panel A: parasternal long-axis view showing significant left atrial (LA) and biventricular dilatation with a thin intraventricular septum (IVS). Panel B: apical four-chamber view demonstrating dilatation of all four chambers. There is failure of the tricuspid leaflets to coapt in systole (arrow). LV, left ventricle; RA, right atrium; RV, right ventricle.

Fig. 6 Colour flow Doppler image of the same patient as shown in Fig. 5 showing a regurgitant tricuspid jet (TR). (See also Plate 2.)

Many of the systemic diseases that are associated with heart muscle disorders have typical clinical, immunological, and biochemical features (see Chapter 15.8.3), and in the absence of evidence to suggest a systemic disease an exhaustive 'routine screen' is probably not cost-effective. There are, however, several potential reversible secondary causes of heart muscle disorder that may simulate dilated cardiomyopathy, and basic screening tests should include serum phosphorus (hypophosphataemia), serum calcium (hypocalcaemia), serum creatinine and urea (uraemia), thyroid function tests (hypothyroidism), and serum iron/ferritin (haemochromatosis).

Electrophysiological testing

Programmed electrical stimulation is of limited clinical value in the identification of high-risk patients. Polymorphic ventricular tachycardia is inducible in up to 30 per cent of cases, but this is a non-specific endpoint. Approximately 10 per cent of patients have inducible sustained monomorphic ventricular tachycardia; about one-third of these die suddenly, but most (75 per cent) who die in this way do not have inducible ventricular tachycardia during programmed stimulation.

Inducible ventricular tachycardia usually arises from diseased myocardium. However, bundle branch re-entry ventricular tachycardia may occur, and in one selected series was seen in up to 40 per cent of patients. This tachycardia is typically rapid (mean cycle length 280 ms) and uses a macrore-entrant circuit that involves the His Purkinje system, usually with right bundle branch anterograde conduction and left bundle branch retrograde conduction. Differentiation from myocardial ventricular tachycardia is confirmed by the presence of a His or right bundle branch potential preceding each QRS: diagnosis is important since catheter ablation of either the left

or right bundle branch usually is curative.

Cardiac catheterization/biopsy

Coronary arteriography should be performed if doubt remains regarding potential ischaemic aetiology of left ventricular dysfunction. Cardiac catheterization is also warranted for measurement of pulmonary vascular resistance in those with very severe or rapidly progressive disease in whom cardiac transplantation may be required. However, if coronary artery disease can be confidently excluded and transplantation is not an imminent consideration, then cardiac catheterization is not required for a diagnosis or symptomatic management. Useful prognostic information regarding cardiac enlargement and systolic performance can be more readily provided from echocardiographic or radionuclide studies.

The role of endomyocardial biopsy is controversial. It is warranted to exclude myocarditis and specific heart muscle disorders and to characterize patients for the presence of viral genome and markers of immune activation, but these evaluations require specialist expertise in cardiac pathology that may not be readily available and the therapeutic implications of findings are uncertain.

Management

Management in dilated cardiomyopathy is aimed at improving symptoms, attenuating disease progression, and preventing arrhythmia, stroke, and sudden death. Such non-specific therapy is unsatisfactory and will remain so until the aetiology and pathogenesis of dilated cardiomyopathy are better delineated.

Pharmacological

Symptomatic therapy is the treatment of heart failure with reliance on ACE inhibitors, b-blockers, and diuretics (see Chapter 15.2.2 and Chapter 15.5.1). Evidence-based therapy of cardiac failure has, with few exceptions, not specifically been evaluated in dilated cardiomyopathy in a randomized fashion, but is likely to be as effective as in others with this condition. Vigorous exercise and significant alcohol intake are proscribed. Moderate- to high-dose ACE inhibition is probably the goal, although in advanced disease this may be limited by hypotension and take time to achieve. The use of low-dose b-blockade with gradual dosage augmentation as tolerated over months is increasingly supported by trial evidence, with metoprolol (6.25 to 50 mg twice daily), bisoprolol (1.25 to 10 mg once daily), and carvedilol (3.125 to 50 mg twice daily) proven to be beneficial. Note, however, that rapid increase in dosage of b-blockers or their sudden withdrawal after chronic administration may precipitate deterioration and dosage changes should be made gradually. Diuretics should be reserved for patients with congestive symptoms as their inappropriate use will limit the ability to achieve optimal dose of ACE inhibitor and b-blocker.

If sustained or symptomatic arrhythmias are documented during 24-h ECG monitoring or exercise testing, conventional treatment is warranted (see Chapter 15.6). Amiodarone (100 to 400 mg daily) has no negative inotropic effect and is effective in suppressing both supraventricular and ventricular arrhythmias. Most class I antiarrhythmics will depress left ventricular function and are likely to be poorly tolerated. The use of low-dose amiodarone (200 mg od) with low to moderate dose of a b-blocker is often effective, but the combination of sotalol and amiodarone represents a proarrhythmic risk because of their additive effects on repolarization. If drug treatment is unsuccessful, the threshold for implantation a cardioverter defibrillator should be low, although concomitant antiarrhythmic therapy may still be required.

Non-pharmacological

Non-pharmalogical alternatives for the treatment of heart failure are increasingly available. Permanent pacing can correct two important intracardiac conduction abnormalities. First, a small subset of patients who have marked P–R interval prolongation (more than 220 ms), usually secondary to atrioventricular nodal disease, experience deleterious effects on left ventricular haemodynamics with reduction in diastolic ventricular filling time and the development of end-diastolic tricuspid and mitral regurgitation. Correction of P–R interval prolongation with short atrioventricular delay dual-chamber pacing may increase stroke volume and blood pressure, and decrease mitral regurgitation with dramatic clinical improvement. Second, patients with marked intraventricular conduction delay (for example left bundle branch block greater than 150 ms) experience asynchronous contraction of the left ventricular free wall and interventricular septum (which may decrease ejection fraction) and late activation of the anterolateral papillary muscle (which may increase functional mitral regurgitation). Biventricular or left ventricular pacing with specialized leads via the coronary sinus can correct both problems and early anecdotal reports of dramatic amelioration of symptoms have been confirmed in randomized trials with subjective and objective evidence of clinical improvement. In addition, the resultant increase in blood pressure and pacemaker maintenance of the desired minimum heart rate permits use of higher doses of b-blockade and ACE inhibition with potential secondary benefit.

Surgical removal of non-viable (Dor procedure) and/or viable myocardium (Batista procedure) to improve haemodynamics by reducing left ventricular volume has been advocated. Though mitral valve repair may occasionally be helpful, these and other surgical volume reduction procedures (partial left ventriculotomy) probably have no role in dilated cardiomyopathy.

Cardiac transplantation has provided a lifeline in those with progressive deterioration. However, the improvements in the pharmacological and non-pharmacological treatments of heart failure in dilated cardiomyopathy appear to be attenuating the progression to endstage disease requiring transplantation. In addition, improvements in left ventricular assist devices and artificial heart technology provide alternatives that are now reasonably seen as viable future treatment options. These issues are discussed in Chapter 15.5.4.

Prevention of disease progression, thromboembolism, and sudden death

With awareness of the importance of familial disease, patients with asymptomatic mild left ventricular dysfunction are increasingly being recognized. Though there is no proof that early treatment will attenuate or prevent disease progression, the recognition that dilated cardiomyopathy is characterized by insidious progression during the asymptomatic phase, and that analogous patients who were in the treatment limb of the Studies of Left Ventricular Dysfunction (SOLVD) prevention trial had improved survival, suggests a role for the introduction of ACE inhibitors at a presymptomatic stage.

Systemic and pulmonary emboli are common. In the retrospective series from the Mayo Clinic, 25 per cent of patients experienced a documented embolic event during 5 years of follow-up. Precise guidelines for anticoagulation are not established. However, patients with mural or intracavitary thrombi and those with established or paroxysmal atrial fibrillation should be fully anticoagulated. Those with severe left ventricular dysfunction (ejection fraction less than 20 per cent) or atrial dilatation (more than 40 mm) should also be anticoagulated, with the INR (international normalized ratio) maintained at the lower end of the therapeutic range (1.5 to 2.5).

The other major complication is sudden death, which may occur in those who are stable or improving as well as in those who are deteriorating. The mechanism is probably ventricular arrhythmia, although bradyarrhythmias may be more likely in those who are severely ill, such as those awaiting cardiac transplantation. Myocyte loss and replacement by fibrous tissue is common in dilated cardiomyopathy, creating a milieu for anisotropic conduction and re-entrant arrhythmias. At autopsy, extensive subendocardial scarring in the left ventricle is seen in approximately one-third of patients, with multiple patchy areas of replacement fibrosis in the majority. Catecholamine excess in such patients may result in several maladaptive responses, including down-regulation of b-adrenergic receptors, inappropriate sinus tachycardia, increased transmural dispersion of refractoriness, and enhanced automaticity of both atrial and ventricular ectopic foci, all of which may increase the risk of both atrial and ventricular arrhythmias.

Recent large, prospective, randomized trials have redefined the role of b-blockade in the treatment of patients with dilated cardiomyopathy and congestive heart failure, demonstrating a substantial reduction not just in sudden death rates but in total mortality, heart failure mortality, and rates of admission to hospital for heart failure. These studies underscore the importance of b-blockers as first-line therapy for symptoms and prognosis in patients with dilated cardiomyopathy.

Though ACE inhibition and angiotensin II (AT_1) receptor antagonism also improve prognosis, the effects in reducing sudden death are less consistent. ACE inhibition usually results in only a transient decrease in aldosterone concentrations. The recent Randomised Aldactone Evaluation (RALES) study, which enrolled approximately 800 patients with severe dilated cardiomyopathy (ejection fraction less than 35 per cent and NYHA class IV symptoms), showed that the addition of low-dose spironolactone (25 mg) to conventional heart failure treatment reduced sudden death as well as progressive heart failure deaths.

Primary prevention of sudden death has been limited by the absence of an accurate risk factor stratification algorithm to identify the appropriate high-risk patients and by the complications of antiarrhythmia therapy. Earlier trials of class I antiarrhythmics were associated with increased mortality from sudden death, presumably due to ventricular proarrhythmia and possibly also related to drug-related worsening of ventricular performance. Recent survival trials (involving only small numbers of

patients with dilated cardiomyopathy) provide reassurance that amiodarone is safe and possibly effective and, though unproven, it is reasonable to treat those with frequent episodes of asymptomatic, non-sustained ventricular tachycardia during ECG monitoring, as well as those with symptomatic or documented ventricular arrhythmia, with low-dose amiodarone (200 to 300 mg/day). Ongoing prospective trials are examining the role of implantable cardioverter defibrillators in the primary prevention of sudden death. They will probably be shown to be effective: the hallenge, however, is to identify those patients whose disease-related life expectancy is adequate and whose arrhythmia risk is sufficient to warrant an implantable cardioverter defibrillator.

Restrictive cardiomyopathy

Definition

Restrictive cardiomyopathy, the least common of the cardiomyopathies, is characterized by restrictive filling of one or both ventricles. This is usually caused by endomyocardial fibrosis. Two variants with similar pathology are recognized. Tropical endomyocardial fibrosis is more common and accounts for 10 to 20 per cent of deaths from heart disease is Africa. Endomyocardial fibrosis in temperate countries is rare and typically associated with hypereosinophilia. The pathology is similar in advanced cases with or without hypereosinophilia, and both variants are considered to be different manifestations of the same disease process. Idiopathic myocardial restrictive cardiomyopathy occurs with and without myocardial fibrosis, but is rare. Myocardial infiltrative diseases (amyloid, sarcoid, Gaucher's), storage diseases (haemochromatosis, glycogen, and Fabry's), and endomyocardial disease associated with malignancies (metastases, carcinoid, radiation, anthracycline toxicity) may also have restrictive physiology and mimic restrictive cardiomyopathy.

Pathology

In endomyocardial fibrosis the cardiac pathology is distinctive, with endocardial fibrosis and overlying thrombosis involving the inflow tracts and the apices, but sparing the outflow tracts of one or both ventricles. Necrotic, thrombotic, and fibrotic stages have been defined in patients with endomyocardial fibrosis and hypereosinophilia. In the necrotic stage there is an acute inflammatory reaction characterized by eosinophilic abscesses in the myocardium, with associated necrosis and arteritis. The endocardium is often thickened and mural thrombi may develop. The thrombotic stage is characterized by endocardial thrombus formation that may be severe, with massive intracavitary thrombosis causing restriction to ventricular filling and a low-output state with high filling pressures. There is a risk of systemic emboli. During the necrotic and thrombotic stages the disease may mimic a hyperacute rheumatic carditis (see Chapter 15.10.1). If the patient survives, healing by fibrosis with hyaline fibrous tissue occurs. There is no further evidence of inflammation and the impact of the disease is caused by the effect of the dense fibrous tissue on ventricular filling volume and atrioventricular valve function.

Clinical features and investigation

Disease onset is usually insidious. Clinical presentation relates to endomyocardial fibrosis: left-sided disease may present with symptoms of pulmonary congestion and/or mitral regurgitation, right-sided disease with raised jugular venous pressure, hepatomegaly, ascites, and tricuspid regurgitation. Radiographic and electrocardiographic appearances are non-specific, showing evidence of raised left and/or right atrial pressure and cardiomegaly with left ventricular hypertrophy. Pulmonary infiltrates, non-specific repolarization changes, and fascicular blocks may occasionally develop.

Two-dimensional echocardiography provides the best non-invasive means of confirming the diagnosis, allowing visualization of the structural abnormalities involving the endocardium and atrioventricular valves as well as demonstration of the abnormal physiology with restriction to filling. There may be intracavitary thrombus with apical cavity obliteration, or bright echoes from the endocardium of the right or left ventricle with tethering of the chordae and reduced excursion of the posterior mitral valve leaflet. Typically, ventricular dimensions and wall thickness are normal, whereas the atria are grossly enlarged. Left ventricular filling terminates early and is followed by a plateau phase coincident with the third heart sound.

The principal haemodynamic consequence of endomyocardial scarring is a restriction to normal filling. Early diastolic pressures are normal, but there is a rapid mid-diastolic rise (square root sign), which plateaus and is not associated with impairment of systolic performance. A similar functional haemodynamic abnormality is seen in pericardial constriction (see Chapter 15.9), but in the latter condition end-diastolic pressures are usually closely similar within the two ventricles, whereas in endomyocardial fibrosis there is usually inequality of the end-diastolic pressures. Mitral and tricuspid regurgitation may be severe and both ventricles appear abnormal in shape on angiography due to obliteration of the apices. This may be particularly marked in the right ventricle in which the infundibulum is hypertrophied and hypocontractile. In addition, the fibrotic process results in smoothing of the internal architecture of the ventricle with loss of the normal trabeculas. The presence of intracavitary thrombi in the left ventricle may give rise to the erroneous diagnosis of a cardiac tumour.

The structural and physiological abnormalities that can be demonstrated with two-dimensional echocardiography or during cardiac catheterization result from the thrombotic and fibrotic stages of the disease. During the early acute phase the appearances of the left and right ventricle are far less abnormal and the diagnosis can best be confirmed at this stage by endomyocardial biopsy. In later stages, diagnosis should be apparent and the risk of biopsy is excessive.

Management

Medical treatment of advanced disease is not particularly effective and the prognosis is poor, with 35 to 50 per cent 2-year mortality. Congestive symptoms from raised right atrial pressure can be improved with diuretics, though too great a reduction in ventricular filling pressure will lead to a reduction in cardiac output. Arrhythmias are common, but their prognostic significance is uncertain and they should therefore not be treated unless they are sustained or associated with symptoms. Antiarrhythmic drugs that significantly slow the heart rate may be deleterious because of the small stroke volume. Digoxin may be helpful to control the ventricular response in atrial fibrillation, but cannot be expected to improve congestive symptoms as systolic function is usually well preserved. Anticoagulants may help to prevent venous thrombosis and systemic emboli; both warfarin and antiplatelet drugs are advised.

Surgery with either mitral and/or tricuspid valve replacement, with or without decortication of the endocardium, has been carried out in some patients. Good long-term symptomatic results have been obtained, but there is significant perioperative mortality (15 to 20 per cent).

Arrhythmogenic right ventricular cardiomyopathy

Definition

Arrhythmogenic right ventricular dysplasia or cardiomyopathy has only recently been recognized. It is characterized pathologically by fibrofatty replacement of the right ventricular myocardium and by clinical presentation with arrhythmia and sudden death. The prevalence is unknown, but estimated to be between 1:1000 and 1:5000. It occurs worldwide, but the high incidence of disease recognized in the Veneto region of Northern Italy raises the possibility of a founder effect. In young athletes (under 25 years) who die suddenly a cardiovascular cause is identified in over 80 per cent, most commonly hypertrophic cardiomyopathy and arrhythmogenic right ventricular cardiomyopathy.

Genetics

The disease is often familial (at least 30 per cent) with autosomal dominant inheritance and incomplete penetrance. Six loci have been reported in autosomal dominant families. The identification of gene abnormalities has been slow, perhaps because of problems in diagnostic ascertainment caused by age-related penetrance, which is seen even in the later decades. Mutations in the ryanidine receptor gene have been found in families with adrenergically mediated ventricular tachycardia which may overlap with the disease phenotype. In an autosomal recessive family from the Greek island of Naxos, in whom palmoplantar keratoderma and woolly hair cosegregated with arrhythmogenic right ventricular cardiomyopathy, a disease-causing mutation has been identified in the plakoglobin gene. Plakoglobin has signalling and intracellular adhesion properties and is an important constituent of the cell-to-cell junction.

Aetiology/pathogenesis

The identification of the plakoglobin gene abnormality in an autosomal recessive form of arrhythmogenic right ventricular cardiomyopathy provides a candidate gene pathway for evaluation. Whether the paradigm that arrhythmogenic right ventricular cardiomyopathy is a disease of the cell-to-cell junction is correct, analogous to hypertrophic cardiomyopathy as a disease of the sarcomere, remains to be determined. Segmental disease is usual in arrhythmogenic right ventricular

cardiomyopathy, with involvement of the diaphragmatic, apical, and infundibular regions of the right ventricular free wall. Evolution to more diffuse right ventricular involvement and left ventricular abnormalities with heart failure are more common than the earlier literature suggested. The fibrofatty replacement of the myocardium may be focal or widespread, usually involves the subepicardial layer of the right ventricular free wall, and when severe may appear transmural. Two morphological patterns are recognized: lipomatous replacement of the myocardium without fibrosis is usually seen with preservation of normal right ventricular free wall thickness in the absence of an inflammatory infiltrate, whilst the fibrolipomatous pattern is characterized by replacement myocardial fibrosis with thinning and discrete bulges of the right ventricular free wall, often in association with lymphocytic infiltrates surrounding degenerating or necrotic myocytes. Animal and *in vitro* studies support the hypothesis that mutations in plakoglobin or analogous genes involved in cell adhesion may cause myocytes under mechanical stress to detach and die, with subsequent fibrofatty replacement.

Clinical presentation and management

Clinical manifestations of the disease include structural and functional abnormalities of the right ventricle, electrocardiographic depolarization/repolarization changes, and presentation with sudden death or arrhythmias of right ventricular origin. Structural and functional evaluation of the right ventricle is problematic. There is no ideal method: reliance on invasive angiography, two-dimensional echocardiography, radionuclide angiography, computed tomography, and/or magnetic resonance imaging will depend on local expertise and facilities. Quality imaging usually reveals segmental dilatation or localized aneurysm(s) of the right ventricular free wall with minimal but occasionally severe left ventricular impairment. The typical ECG presents inverted T waves in right ventricular precordial leads (V1–V3(4)) (Fig. 7) and ventricular postexcitation 'epsilon waves'. These waves are the surface ECG manifestation of late potentials that are found on the time domain signal-averaged electrocardiogram in 40 to 50 per cent of patients who present with arrhythmia. They occur as a consequence of the inhomogenous and delayed right ventricular depolarization.

Fig. 7 A 12-lead ECG from a young woman showing the most common electrocardiographic abnormality found in arrhythmogenic right ventricular cardiomyopathy, T-wave inversion in the precordial leads V1–V4.

Symptomatic presentation is usually with palpitation and/or syncope from sustained ventricular arrhythmia. Ventricular tachycardia is of left bundle branch block morphology suggesting a right ventricular origin. Sudden death related to exercise may be the initial manifestation, especially in the young.

The diagnosis of right ventricular dysplasia is based on histological demonstration of fibrofatty replacement of right ventricular myocardium at either autopsy or surgery. Diagnosis based on biopsy specimens from the right ventricular endomyocardium is inherently difficult because the segmental nature of the disease causes false negatives (Fig. 8), and because the amount of tissue usually obtained is insufficient to differentiate fibrofatty replacement from islands of adipose tissue that are not infrequently seen between myocytes in the right ventricle of normal subjects. Nevertheless, the positive finding of fibrofatty replacement of myocytes on biopsy can be a valuable diagnostic clue.

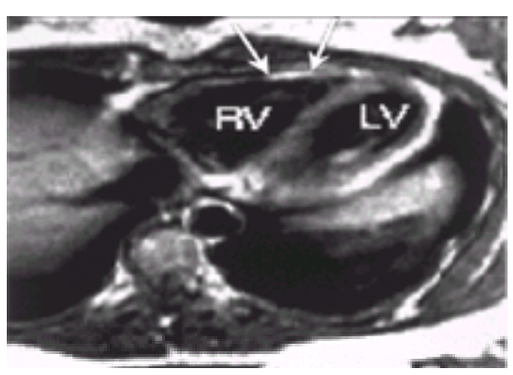

Fig. 8 A transverse plane spin-echo MRI in a young woman with arrhythmogenic right ventricular cardiomyopathy demonstrating a circumscribed area of enhanced MR signal intensity in the right ventricular (RV) free wall (arrows) due to fatty infiltration. (By courtesy of Dr Dudley Pennell, Cardiovascular Magnetic Resonance Unit, Royal Brompton Hospital, London.)

Diagnostic criteria are based on evidence of familial disease and on the clinical demonstration of structural, functional, and electrophysiological abnormalities that are caused by or reflect the underlying histological changes (Table 6). The presence of two major or one major and two minor or four minor features provides specific though possibly insensitive diagnostic criteria.

The natural history of arrhythmogenic right ventricular dysplasia is uncertain because patients at autopsy and/or those presenting with sustained ventricular arrhythmias bias published series. In the absence of sustained ventricular arrhythmia most patients will be asymptomatic. Progression from localized (with no or minor symptoms) to more diffuse right ventricular involvement with features of right ventricular failure has been reported. The left ventricle may be involved in long-standing disease, making differentiation from dilated cardiomyopathy with biventricular involvement difficult.

Management aims to identify those at risk of sustained ventricular arrhythmia and to prevent sudden death. Assessment of asymptomatic patients should include exercise testing and Holter monitoring for detection of occult arrhythmia. Antiarrhythmic treatment guided by electrophysiological studies is warranted in patients with palpitation, syncope, or documented sustained ventricular arrhythmia, and should also be considered in those with a markedly abnormal signal-averaged electrocardiogram who are at increased risk. The morphology of ventricular arrhythmia may vary, suggesting multiple sites of origin. Arrhythmias are usually progressive and therapy—whether pharmacological, ablation, or surgical—is not usually definitive. Implantable cardioverter defibrillators are the treatment of choice in patients resuscitated from haemodynamically compromising ventricular tachycardia or ventricular fibrillation.

Further reading

Hypertrophic cardiomyopathy

Braunwald *et al.* (1964). Idiopathic hypertrophic subaortic stenosis. I. A description of the disease based upon an analysis of 64 patients. *Circulation* **30**(Suppl IV), 3–119.

Elliott PM *et al.* (2001). Relation between the severity of left ventricular hypertrophy and prognosis in patients with hypertrophic cardiomyopathy. *Lancet* **357**, 420–4.

Maron BJ *et al*. (2000). Efficacy of the implantable cardioverter-defibrillator for the prevention of sudden death in hypertrophic cardiomyopathy. *New England Journal of Medicine* **342**, 365–73.

McKenna WJ *et al*. (1985). Improved survival with amiodarone in patients with hypertrophic cardiomyopathy and ventricular tachycardia. *British Heart Journal* **53**, 412–16.

Spirito P *et al.* (1997). The management of hypertrophic cardiomyopathy. [Review.] *New England Journal of Medicine* **336**, 775–85.

Teare D (1958). Asymmetrical hypertrophy of the heart in young adults. *British Heart Journal* **20**, 1–8.

Thierfelder L *et al.* (1994). a-Tropomyosin and cardiac troponin T mutations cause familial hypertrophic cardiomyopathy: a disease of the sarcomere. *Cell* **77**, 1–20.

Vosberg H-P, McKenna WJ (2002). Cardiomyopathies. In: Rimoin DL, Connor JM, Pyeritz RE, eds. *Emery and Rimoin's principles and practice of medical genetics*, 4th edn. Churchill Livingstone, New York, in press.

Watkins H *et al*. (1992). Sporadic hypertrophic cardiomyopathy due to *de novo* myosin mutations. *Journal of Clinical Investigation* **90**, 1666–71.

Wigle ED *et al.* (1985). Hypertrophic cardiomyopathy. The importance of the site and the extent of hypertrophy. A review. *Progress in Cardiovascular Diseases* **28**, 1–83.

Dilated cardiomyopathy

Baboonian C, Treasure T (1997). Meta-analysis of the association of enteroviruses with heart disease. *Heart* **78**, 539–43.

Caforio ALP *et al.* (1994). Autoimmunity in dilated cardiomyopathy: evidence from family studies. *Lancet* **344**, 773–7.

Mestroni L *et al*. (1999). Collaborative research group of the European human and capital mobility project on familial dilated cardiomyopathy. Guidelines for the study of familial dilated cardiomyopathy. *European Heart Journal* **20**, 93–102.

Michels VV *et al.* (1992). The frequency of familial dilated cardiomyopathy in a series of patients with idiopathic dilated cardiomyopathy. *New England Journal of Medicine* **326**, 77–82.

Noutsias M *et al.* (1999). Expression of cell adhesion molecules in dilated cardiomyopathy: evidence for endothelial activation in inflammatory cardiomyopathy. *Circulation* **99**, 2124–31.

Pauschinger M *et al*. (1999). Dilated cardiomyopathy is associated with significant changes in collagen type I/III ratio. *Circulation* **99**, 2750–6.

Tracy S *et al.* (1990). Molecular approaches to enteroviral diagnosis in idiopathic cardiomyopathy and myocarditis. *Journal of the American College of Cardiology* **15**, 1688–94.

Vosberg H-P, McKenna WJ (2002). Cardiomyopathies. In: Rimoin DL, Connor JM, Pyeritz RE, eds. *Emery and Rimoin's principles and practice of medical genetics*, 4th edn. Churchill Livingstone, New York, in press.

Restrictive cardiomyopathy

See Chapter 15.8.3.

Arrhythmogenic right ventricular dysplasia

Corrado D *et al.* (1997). Spectrum of clinicopathologic manifestations of arrhythmogenic right ventricular cardiomyopathy/dysplasia: a multicenter study. *Journal of the American College of Cardiology* **30**, 1512–20.

Marcus FI *et al.* (1982). Right ventricular dysplasia: a report of 24 adult cases. *Circulation* **65**, 384–98.

McKenna WJ *et al.* (1994). Diagnosis of arrhythmogenic right ventricular dysplasia/cardiomyopathy. *British Heart Journal* **71**, 215–18.

McKoy G *et al.* (2000). Identification of a deletion in plakoglobin in arrhythmogenic right ventricular cardiomyopathy with palmoplantar keratoderma and woolly hair (Naxos disease). *Lancet* **355**, 2119–24.

15.8.3 Specific heart muscle disorders

William J. McKenna

Cardiac manifestations of musculoskeletal and connective tissue diseases

The cardiac manifestations of musculoskeletal and connective tissue diseases often go undetected. Every anatomical structure in the heart may be involved, there usually being no correlation between the extent of systemic disease and cardiac involvement. For details of the cardiac manifestations of musculoskeletal and connective tissue diseases, see Table 1 and Table 2.

Systemic lupus erythematosus

Systemic lupus erythematosus is a multisystem immune disorder characterized by the formation of autoantibodies to numerous organ systems. The prevalence of cardiovascular involvement is reported to be greater than 50 per cent. The pericardium is most commonly affected, with as many as 30 per cent of patients with lupus having clinical pericarditis at some stage, and up to 66 per cent affected at autopsy. Progression to constrictive pericarditis or tamponade remains extremely rare.

Myocardial involvement occurs less frequently, although reported in up to 30 per cent of patients at autopsy: signs and symptoms are uncommon, but patients may occasionally present with heart failure or arrhythmias.

As many as one-third of patients have systolic murmurs, but these usually represent hyperdynamic flow states due to other causes. The classic verrucous vegetations adherent to the endocardium described by Libman and Sachs in 1924 can be identified in up to 30 per cent of patients at autopsy. These lesions most commonly affect the mitral valve but rarely become clinically significant, although thromboembolism, valvular incompetence, and infective endocarditis are all described.

Conduction abnormalities are common: various degrees of heart block and bundle branch block can be seen, but complete heart block is rare. Arrhythmias such as atrial fibrillation and flutter may also occur, particularly in association with pericarditis. Myocardial infarction is very rarely reported in patients with systemic lupus erythematosus, atherosclerosis usually being implicated in the pathogenesis, although arteritis and steroid use may also play a role.

Death from the cardiac complications of lupus is rare. Mild pericardial disease may respond to non-steroidal anti-inflammatory drugs, heart failure is treated conventionally, and conduction defects may require pacing. Corticosteroids are thought to be useful in patients with coronary vasculitis and myocarditis, but there is no evidence for the use of other immunosuppressants.

Antiphospholipid syndrome

The antiphospholipid syndrome is recognized both in patients without (primary) and with systemic lupus erythematosus. It is a thrombophilic disorder characterized by arterial and venous occlusions, recurrent fetal loss, thrombocytopenia, and increased maternal complications of pregnancy. It is associated with persistently raised titres of anticardiolipin antibodies or the Lupus anticoagulant. Involvement of the mitral and aortic valves is particularly common and dramatic response to prednisolone has been described (see Chapter 13.14 for further information).

Systemic sclerosis

Heine in 1926 and Weiss *et al*. in 1943 first described cases of myocardial involvement: it has now been shown that up to 60 per cent of patients have cardiac involvement at autopsy. Gradual obliteration of the microvasculature leads to a 'Raynaud's' type of phenomenon in the heart, which is thought to be responsible for ischaemia that may then progress to patchy myocardial fibrosis. This pathological process most commonly affects the left ventricle. Right ventricular dysfunction is usually secondary to pulmonary vascular disease or pulmonary hypertension, but an associated right ventricular cardiomyopathy may coexist.

Dyspnoea is the most common symptom and usually attributable to pulmonary involvement, though it may be secondary to left ventricular systolic and/or diastolic dysfunction. Atypical chest pain may be secondary to pulmonary hypertension or fibrosis, pericarditis, or oesophageal reflux.

Diffuse narrowing of intramural coronary arteries may be associated with myocardial infarction or angina. Clinical pericarditis is reported in 15 per cent of patients and up to 70 per cent demonstrate pericardial involvement at autopsy.

ECG abnormalities are reported in 75 per cent of patients with scleroderma. Conduction abnormalities are seen in as many as 50 per cent of those with cardiac disease and may present with palpitations, syncope, or sudden death. The presence of left bundle branch block or bifascicular block suggests significant myocardial involvement: a septal infarct pattern or interventricular conduction abnormalities may also be seen. Ambulatory electrocardiograms frequently reveal a high prevalence of atrial and ventricular premature beats: supraventricular tachycardias are seen in 30 per cent of patients and ventricular tachycardia (which may predict future sudden death) in up to 15 per cent. The echocardiogram typically shows features of dilated or restrictive cardiomyopathy. Resting thallium perfusion abnormalities are seen in the majority of affected individuals, whilst arteriography usually reveals normal coronary arteries. Endomyocardial biopsy is rarely performed.

Treatment for pulmonary involvement may involves prostacyclin, long-term oxygen therapy, and anticoagulation. Treatment for heart failure is along conventional lines; calcium channel antagonists, nitrates, and vasodilators may improve resting myocardial perfusion.

The major cause of death in scleroderma is pulmonary hypertension, with a 5-year mortality rate of 60 per cent. Although nothing has been shown to alter the cardiac manifestations, D-penicillamine and isoretinoin have been used with some success. Symptomatic cardiac involvement predicts a poor prognosis, with a 2-year mortality of approximately 60 per cent.

Rheumatoid arthritis

Cardiac involvement is found in up to 60 per cent of patients on echocardiography, but only in 10 to 15 per cent clinically. The presence of cardiac disease tends to correlate with the severity of joint disease and the presence of rheumatoid nodules. Histological changes consist of a non-specific inflammatory infiltrate, myocyte necrosis, and fibrosis affecting any part of the heart. Rheumatoid nodules may accompany this, and the heart may also be affected by secondary amyloidosis. Myocarditis is reported in up to 20 per cent at autopsy, but symptoms are uncommon. Pericarditis occurs more frequently, and up to 40 per cent of patients have an

effusion on echocardiography, but progression to constrictive pericarditis or tamponade is rare. Acute vasculitis involving the larger epicardial arteries has been reported but is uncommon. Non-specific valvitis may affect the mitral and particularly the aortic valve: this may eventually lead to scarred, hyalinized, and even incompetent valves. Rheumatoid nodules may occasionally deform the mitral valve and lead to valvular incompetence. Conduction disturbances may be secondary to infiltration by rheumatoid nodules: the commonest ECG abnormality is first-degree heart block, but left bundle branch block and complete heart block are also described. Although pericarditis is usually responsive to steroids, it is unclear whether steroids or disease-modifying drugs alter the other cardiac manifestations.

Polymyositis and dermatomyositis

Cardiac involvement in polymyositis or dermatomyositis is present in up to 15 per cent of patients clinically and as many as 55 per cent on echocardiography. Histological changes consist of non-specific inflammatory cell infiltrate, myocyte necrosis, and fibrosis, involving particularly the cardiac conducting tissue and leading to various degrees of conduction block and arrhythmias.

There is echocardiographic evidence of pericardial disease in up to 25 per cent of patients, but this usually remains asymptomatic. Myocarditis may be present in 25 per cent of patients at autopsy, although the development of clinically apparent heart failure is uncommon. The ECG is often abnormal, particularly in children. Abnormalities consist mostly of non-specific 'ST' and 'T' wave changes and conduction delays. Treatment is based on symptomatology.

Seronegative arthropathies

This group of disorders is characterized by the absence of rheumatoid factor and includes ankylosing spondylitis, Reiter's syndrome, and psoriatic and gastrointestinal arthropathies. These may all be associated with cardiac involvement, in particular pancarditis, proximal aortitis, aortic incompetence, and varying degrees of conduction abnormalities. They may also result in amyloid deposition. On occasion cardiac disease may present before joint disease. Treatment is empirical and based on symptomatology.

Neuromuscular diseases

The muscular dystrophies are a group of disorders characterized by progressive skeletal and cardiac muscle involvement (Table 3). Dystrophic effects on skeletal muscle result in fibre necrosis, followed by fibrosis and fatty replacement. The heart is commonly affected and cardiac disease tends to be progressive. The structural and functional changes, which occur in the ventricles, can lead to the development of cardiomyopathy, in particular dilated cardiomyopathy and heart failure. The effect on the specialized conducting tissue may lead to bradyarrhythmias, conduction defects, malignant arrhythmias, and sudden death.

Duchenne and Becker muscular dystrophy are progressive disorders arising from abnormalities (deletion, duplication, or point mutation) in the genes involved in the manufacture of the extra-sarcomeric cytoskeletal protein dystrophin. In addition to defects in dystrophin, other defects that might be responsible for muscular dystrophy and dilated cardiomyopathy include those affecting the genes for the intracellular proteins, emerin and laminin. Emerin is a transmembrane protein that is embedded in the inner nuclear cell membrane. Laminin A–C are filament-like proteins that form a proteinaceous mesh underlying and attached to the inner nuclear membrane. The exact mechanism by which alterations in these proteins may lead to cardiomyopathy remains unclear.

In general, treatment of the cardiomyopathy of neuromuscular disorders is empirical and based on symptomatology and evidence of arrhythmia or conduction block. Should advances in the treatment of neuromuscular disorders by gene therapy or other means result in prolonged survival, then cardiac failure may become the limiting factor.

Amyloid

Amyloidosis describes a group of diverse protein-deposition diseases. The biochemical nature of the proteinaceous deposits and the aetiology of the underlying associated diseases differ (Table 4).

As many as 50 per cent of patients with systemic AL (primary) amyloidosis have cardiac involvement and this will manifest clinically in up to half of these. Systemic AA (secondary) amyloidosis is almost never associated with clinical cardiac amyloidosis. The heart is frequently involved in familial amyloid polyneuropathy, which is the most common type of hereditary amyloidosis and caused by more than 70 mutations in the transthyretin gene. Senile amyloidosis is extremely common, indeed almost all individuals over the age of 80 years will have scattered deposits of amyloid, particularly affecting the aorta. Clinical involvement is variable, depending on the extent of deposition, but tends to be unimportant in senile amyloidosis.

The extracellular deposition of amyloid results in a firm, thickened, non-compliant myocardium. Deposition occurs throughout the atrial and ventricular muscle. Conducting as well as nodal tissue may be affected: fibrosis of these structures may occur. Valvular function is rarely affected, although deposition in and thickening of cardiac valves is common. Intramural coronary arteries and veins frequently contain deposits, which can occasionally compromise the lumina of these vessels.

Amyloid heart disease most frequently mimics hypertrophic cardiomyopathy with restrictive physiology. The reduced compliance of the myocardium produces the characteristic diastolic dip and plateau (square root sign) in the ventricular pressure waveform that may simulate constrictive pericarditis. An impaired rate of early diastolic filling is characteristic and systolic dysfunction may also occur, leading to congestive heart failure.

Progressive infiltration of the autonomic nervous system results in orthostatic hypotension in 10 per cent of cases. Arrhythmias are common, in particular ventricular premature beats and atrial fibrillation. Complex ventricular arrhythmias may be harbingers of sudden death.

The chest radiograph may show cardiomegaly in patients with systolic dysfunction but is often normal in those with restrictive cardiomyopathy, although pulmonary congestion may be prominent. The ECG shows diminished voltages in approximately 50 per cent of patients, and loss of 'R' waves in precordial leads; the presence of 'Q' waves in the inferior leads may simulate myocardial infarction. Echocardiography reveals an increased thickness of the ventricular walls with small ventricular chambers, dilated atria, intra-atrial septal thickening, left ventricular dysfunction, and a characteristic 'sparkling' appearance to the myocardium. Asymmetrical septal hypertrophy has also been recognized. Scintigraphy with technetium-99-pyrophosphate may be strongly positive. CT and MRI may also be helpful, as may endomyocardial biopsy.

Symptomatic heart disease typically presents late in the course of amyloidosis and the presence of clinical signs is an ominous feature with mortality approaching 100 per cent at 2 years. Treatment is supportive in combination with measures to suppress the underlying amyloidogenic condition. This ranges from myeloma-type chemotherapy in AL amyloidosis to liver transplantation in familial amyloid polyneuropathy. Digoxin and calcium channel antagonists should be used with caution as they selectively bind to amyloidal fibrils, enhancing their effect. Patients with symptomatic conduction system disease require a pacemaker. Diuretics and vasodilators should be used cautiously as they may aggravate hypotension. Transplantation is feasible in selected cases but is a palliative procedure without treatment of the underlying process.

Inherited infiltrative disorders causing cardiomyopathy

Various disorders may lead to infiltration of the myocardium by an abnormal metabolic product, resulting in abnormal systolic and/or diastolic function of the heart. The disorders include glycogenoses, the mucopolysaccharidoses, Fabry's disease, and Gaucher's disease.

Fabry's disease (angiokeratoma corporis diffusum universale)

An X-linked disorder present in 1 in 40 000 live-born babies in which an inherited deficiency of the enzyme a-galactosidase results in the intracellular accumulation of a glycolipid substrate in numerous organs, including the myocardium. Most patients eventually develop symptomatic cardiovascular manifestations including hypertension, mitral valve prolapse, and congestive heart failure. The electrocardiogram often shows left ventricular hypertrophy, 'P' wave abnormalities, conduction defects, and arrhythmias. Echocardiography usually demonstrates increased thickness of the left ventricle, which may simulate hypertrophic cardiomyopathy. Differentiation from other hypertrophic or restrictive processes may require MRI or endomyocardial biopsy. A low leucocyte a-galactosidase activity is diagnostic.

Gaucher's disease

A deficiency of the enzyme b-glucosidase results in the accumulation of cerebrosides in the spleen, liver, bone marrow, lymph nodes, brain, and myocardium. Diffuse interstitial infiltration of the left ventricle leads to a reduction in left ventricular compliance and cardiac output. Clinical evidence of cardiac involvement is uncommon, but when present is characterized by left ventricular dysfunction, haemorrhagic pericardial effusion, thickened left ventricle, and calcification of left-sided valves.

Sarcoid

Sarcoid is a multisysytem granulomatous disorder of unknown aetiology. Myocardial involvement is seen in 20 to 30 per cent of patients at autopsy but is clinically apparent in less than 10 per cent of cases. Primary cardiac involvement is extremely rare.

Non-caseating granulomas may involve any region of the heart, although the left ventricular free wall and interventricular septum are the most commonly affected sites. The granulomas can be localized or widespread, and healing may result in the formation of scars. The ventricular muscle eventually becomes increasingly non-compliant; this can lead to defects in contractile function as well as wall motion. Replacement of large portions of the ventricle by sarcoid tissue may lead to aneurysm formation. Granulomas and fibrosis may also extend to involve nodal or conducting tissue. Isolated pericardial involvement is rare, although pericardial effusions are commonly seen on echo. Valvular dysfunction occurs in fewer than 5 per cent of patients and may be the result of infiltration of papillary muscles or direct valvular involvement, which is less common.

Clinical manifestations of myocardial sarcoidosis are shown in Table 5. Chest pain has been described in up to 28 per cent of patients, and since about half of these will have abnormal thallium perfusion scans despite arteriographically normal coronary arteries, this is thought to be secondary to microvascular spasm.

Sudden death is one of the most common and feared manifestations of myocardial sarcoidosis, occurring in about 65 per cent of affected patients. It is thought to be predominantly secondary to arrhythmias, including ventricular tachycardia and fibrillation. The presence of a ventricular aneurysm may be associated with resistant ventricular arrhythmias and necessitate its resection. Conduction disturbances such as complete heart block are a frequent occurrence and may also predict sudden death. The electrocardiogram is frequently abnormal with 'T' wave abnormalities and varying degrees of interventricular or atrioventricular block. Pathological 'Q' waves may simulate myocardial infarction when myocardial involvement becomes extensive. Echocardiography most commonly shows features of restrictive or occasionally dilated cardiomyopathy. Systolic and/or diastolic dysfunction as well as regional wall motion abnormalities may also be seen. Gallium or technetium pyrophosphate scanning and MRI have all been used to detect affected areas of myocardium. Endomyocardial biopsy can be diagnostic but is rarely done due to the patchy nature of the disease.

Steroids have been shown to lead to improvements in symptoms as well as electrocardiographic and echocardiographic features and myocardial perfusion defects, although there is a theoretical risk of increased aneurysm formation. Amiodarone may be of benefit in resistant arrhythmia and the insertion of an implantable defibrillator may protect against sudden death in susceptible patients. Transplantation may improve prognosis and quality of life in patients who remain symptomatic despite these measures, although recurrence has been documented. The average survival from the onset of symptomatic cardiac involvement has been reported as 1 to 2 years.

Haemochromatosis

Hereditary haemochromatosis is the most common single-gene disorder in people of northern European origin, where approximately 3 to 5 persons per 1000 are homozygous for the disease. It results in excessive and inappropriate mucosal absorption of iron, which is then deposited predominantly in the heart, liver, gonads, and pancreas. Deposition in the heart results in thickening of the ventricular walls together with dilatation of the ventricular chambers and heart failure. Histopathologically, myocardial degeneration and fibrosis occur over time and may extend to involve the conducting system of the heart.

The ECG most commonly reveals 'ST' and 'T' wave changes. Supraventricular arrhythmias are also characteristic, with atrioventricular conduction defects and ventricular arrhythmias being less common. Echocardiography typically shows a mixed dilated and restrictive cardiomyopathy with thickened ventricular walls, ventricular chamber enlargement, systolic and/or diastolic dysfunction. Endomyocardial biopsy may be useful to confirm the diagnosis but cannot rule it out. Treatment involves repeated phlebotomy and/or desferrioxamine.

There is evidence that the type of the inherited mutation may determine the development of cardiomyopathy. Two new mutations have recently been described which may predispose to the development of hereditary haemochromatosis and dilated cardiomyopathy. These affect the haemochromatosis gene on chromosome 6p (termed the HPE gene) and probably act as disease-modifying genes in dilated cardiomyopathy, although having little effect on iron status. The pathogenesis of cardiomyopathy here may be unrelated to excessive iron.

Diabetes

A man with diabetes has a relative risk of developing heart failure that is 2.4 times higher than that of a man without diabetes, and the equivalent relative risk for a woman is 5:1. The risk has been shown to be independent of age, systolic blood pressure, serum cholesterol, and weight. People with diabetes have been shown to have elevated end-diastolic pressures, reduced ejection fractions, left ventricular dilatation, and hypertrophy, even in the absence of coronary artery disease. Diastolic dysfunction as well as a diffuse hypokinesis of the myocardium has also been demonstrated. Implicated mechanisms include small vessel disease and autonomic neuropathy.

The most prominent histopathological finding is that of myocardial fibrosis. Occasionally a picture resembling restrictive heart disease is seen, with a small left ventricular chamber and reduced compliance of the left ventricle.

The treatment of heart failure is the same as in patients without diabetes, although b-blockers with intrinsic sympathomimetic activity are preferred. Preload and after-load reducing agents should be used cautiously because of autonomic dysfunction. It is unclear whether tight glucose control affects the progression of diabetic 'cardiomyopathy', but it is clearly prudent for other reasons to optimize control as well as to reduce obesity and control hypertension.

Hyperthyroidism

In general, excess thyroid hormone results in a high output state with tachycardia, increased cardiac contractility, and peripheral vasodilatation. In the long term this can result in ventricular hypertrophy and an increase in ejection fraction. However, some patients may develop a low output state with symptoms of heart failure and echocardiographic demonstration of dilated cardiomyopathy and systolic dysfunction. These changes may be a result of long-standing tachycardia and increased cardiac work, but thyroxine itself may directly alter the expression of certain cardiac proteins involved in cardiac function, and there is also some evidence that direct autoimmune attack on the myocardium may occur in Graves' disease.

Typical symptoms of hyperthyroidism include angina-like chest pain, fatigue, palpitations, and exertional dyspnoea. Cardiac findings include sinus tachycardia and atrial flutter or fibrillation in 17 to 20 per cent. These may be complicated by thromboembolism in up to 40 per cent; also by congestive heart failure. Mitral valve prolapse has been reported in patients with Graves' disease.

Control of the ventricular rate in atrial fibrillation may be obtained with digoxin, b-adrenergic antagonists, or calcium channel antagonists. The increased metabolic clearance of digoxin may necessitate a higher maintenance dose. Cardioversion should generally be deferred until euthyroid. b-Adrenergic antagonists offer prompt control of sympathomimetic manifestations. The presence of an already dilated vascular bed means that diuretics should be used with caution and vasodilators are generally contraindicated. Treatment of hyperthyroidism *per se* is discussed elsewhere.

Hypothyroidism

Patients suffering from hypothyroidism, whether in its mild form or full-blown myxoedema, present a wide variety of symptoms. Complaints of fatigue, lethargy, mental slowness, and cold intolerance usually dominate. Less frequently, symptoms suggestive of cardiac dysfunction such as dyspnoea on exertion, syncope, or angina-like

chest pain may be prominent. The most common cardiac abnormality is pericardial effusion, which is usually asymptomatic but reported in at least 30 per cent of untreated patients. Heart failure generally represents exacerbation of pre-existing cardiac disease by the superimposed haemodynamic consequences of thyroid deficiency—bradycardia, diminished myocardial contractility, and increased peripheral vascular resistance.Rarely, hypothyroidism alone can closely resemble cardiomyopathy severe enough to cause heart failure. Echocardiographic evidence of asymmetric thickening of the interventricular septum as well as reduced left ventricular outflow tract dimensions has been reported. The characteristic ECG findings are sinus bradycardia, prolongation of the QT interval, and a reduction in voltages if there is an associated pericardial effusion.

The management of heart failure involves the identification of any primary cardiac disease that may coexist; both ischaemic heart disease and aortic stenosis may be exacerbated by thyroid replacement. L-Thyroxine (T_4) significantly enhances myocardial performance within 1 week. It is generally used as first-line treatment of hypothyroidism, but in those with known or suspected coronary artery disease it should be initiated at a lower dose than usual, typically 25 µg/day, and increased slowly at 4- to 6-week intervals until the thyroid-stimulating hormone is within the normal range. Tri-iodothyronine (T_3) may be preferable in severe cases as clinical improvement occurs sooner. b-Blockade can be used prophylactically or added if treatment with L-thyroxine exacerbates ischaemic heart disease.

Further reading

Benson MD (1997). Aging, amyloid, cardiomyopathy. *New England Journal of Medicine* **336**, 502–4.

Braunwald E, ed. (1998). *Heart disease: a textbook of cardiovascular medicine*, 5th edn, pp 1427–35. WB Saunders, Philadelphia.

Cox GF, Kunkel LM (1997). Dystrophies and heart disease. *Current Opinion in Cardiology* **12**, 329–42.

Landerson PW (1990). Recognition and management of cardiovascular disease related to thyroid dysfunction. *American Journal of Medicine* **88**, 638–41.

Shabina H, Isenberg DA (1999). Autoimmune rheumatic diseases and the heart. *Hospital Medicine* **60**, 95–9.

Shammas RL (1993). Sarcoidosis of the heart. *Clinical Cardiology* **16**, 462–72.

Topol EJ (1998). *Comprehensive cardiovascular medicine*, volume 1, chapter 27, pp. 690–726.

15.9 Pericardial disease

D. G. Gibson

Anatomy of the pericardium

The normal pericardium consists of serous and fibrous components. The fibrous pericardium is thick and unyielding, separating the heart from surrounding organs. It fuses with the central tendon of the diaphragm below and with the great vessels above, 1 to 2 cm beyond their origins. Inside the fibrous pericardium is the serous pericardium, in which the heart is invaginated. It has two layers, a parietal layer which lines the inner aspect of the fibrous pericardium and a visceral layer, sometimes called the epicardium, which covers the surfaces of the heart and the origins of the great vessels. The pericardial cavity is the potential space between the visceral and parietal layers of the serous pericardium. It normally contains only a few millilitres of fluid, but it has a considerable capacity where fluid may accumulate.

Physiology of the pericardium

The normal pericardium is not essential to life: the pericardial space is often obliterated after open heart surgery, and both layers may be removed in patients with constrictive pericarditis without apparent ill effect. Whether restraint by the normal pericardium is of any pathophysiological importance as a mechanism limiting stroke volume in disease remains uncertain.

Congenital abnormalities of the pericardium

Congenital abnormalities of the pericardium are uncommon with an incidence of 1 to 2 per 10 000 autopsies. Congenital absence of the pericardium may be partial or complete. A partial defect, involving the left side of the pericardium is about four times as common as the complete form. Either type may be associated with additional congenital anomalies in about one-third of cases, including Fallot tetralogy, atrial septal defect, or sequestered pulmonary segments. Clinical features include non-specific chest pain and sinus bradycardia, with an ECG showing right axis deviation. A chest radiograph is characteristic, with a shift of the heart to the left and prominence of the main pulmonary artery. Heart size is increased and the lower border of the cardiac shadow ill-defined. Echocardiography shows increased right ventricular size and reversed septal motion, as occurs in atrial septal defect. If the defect is partial, the left atrial appendix may herniate through it, and even strangulate leading to a clinical picture suggesting acute pericarditis. If the defect is larger, the left ventricle may herniate and undergo torsion. Alternatively, lung may become trapped in the pericardial space. These complications are treated surgically by enlarging the defect.

Pericardial cysts

These are rare, have a variety of embryological origins, and may be continuous with the pericardium or separate from it. They do not usually cause symptoms, but can be discovered on chest radiographs taken for any reason. Their nature becomes obvious with CT scanning or MRI, but the exact diagnosis is often established only when they are removed surgically.

Mulibrey nanism

Mulibrey (**mu**scle, **li**ver, **br**ain, **ey**e) nanism is an autosomal recessive condition characterized by growth failure, a triangular face, often with a hydrocephaloid skull, hypotonia, a peculiar voice, large liver, and yellowish dots and pigment dispersion in the optic fundi. The majority of cases have pericardial constriction due to congenital thickening of the pericardium, and this may be responsible for some of the clinical features. Histologically, the pericardium shows simple fibrosis. Considerable improvement follows pericardiectomy.

Acquired pericardial disease

Diseases of the pericardium may be considered from two points of view. The first is aetiological, the second is in terms of the physiological and clinical disturbances that result. There is no fixed relation between the two, so that an account will be given of the different diseases affecting the pericardium and then of the three main syndromes: acute pericarditis, pericardial tamponade, and pericardial constriction.

Aetiology

Diseases affecting the pericardium are given in Table 1.

Acute idiopathic pericarditis

Acute idiopathic pericarditis is a disease occurring in young adults, usually sporadically. Prospective studies have suggested a viral basis for around half. Coxsackie B is most commonly involved, but others including ECHO type 8, rubella, hepatitis B, mumps, and influenza have also been identified. Epidemics can occur, with

approximately equal numbers of patients developing pericarditis and myocarditis.

The commonest clinical feature is chest pain, but 'flu-like' symptoms, palpitations, orchitis, encephalitis, and radiographic appearances of pneumonitis or pleural effusion have all been reported. The condition is usually self-limiting, but a minority of cases follow a relapsing course over the succeeding 6 to 12 months. A virus may be identified from paired blood samples taken 2 weeks apart, or recovered from throat or rectal swabs, but in many cases no clear cause is identified and positive virology is not necessary for the diagnosis.

HIV infection

Pericardial effusion can occur in AIDS, usually as a late manifestation with poor prognosis. For further information see Chapter 15.10.4.

Pyogenic infection

Pyogenic infection of the pericardium is uncommon. It is usually due to bloodborne infection of a previously sterile pericardial effusion, or the result of direct spread from the lungs or pleural space. The organisms most commonly involved are staphylococci, pneumococci, or streptococci. Bacterial infection of the pericardium is not usually an isolated event and occurs more frequently in an immunologically compromised patient.

Tuberculous infection

Tuberculous infection is an important cause of pericardial disease, particularly in the Third World. It may take the form of acute pericarditis, pericardial effusion, or constriction. Acute pericarditis appears to be a 'primary' response, and can be regarded as an exudative lesion whose main basis is allergic. Chronic pericardial effusion and constriction both reflect granulomatous disease, often with fibrosis and calcification in the late stages. Both parietal and visceral layers of the pericardium may be involved, and spread of the disease to the myocardium follows. In the first instance, treatment is with antituberculous drugs. In both effusion and constriction in HIV-negative patients, adding steroids for the first 11 weeks reduces the need for pericardiectomy, increases the speed with which heart rate and venous pressure fall to normal, and expedites return to work. In the absence of specific contraindication, therefore, steroids should be added to standard antituberculous chemotherapy.

In Sub-Saharan Africa, patients with tuberculous pericarditis now have a more than 80 per cent chance of being HIV positive. Pericardial constriction severe enough to require surgery is rare, possibly reflecting depressed ability to form dense fibrosis in HIV infection. Treatment is with standard antituberculous drugs which, apart from thiacetazone, are well tolerated and effective. It is uncertain whether there is an adjuvant role for additional steroid in such patients. Long-term survival is shorter than in HIV-negative patients, due to other opportunistic infections rather than recurrence of tuberculosis.

Fungal infection

Fungal pericarditis is uncommon, but infection with actinomycosis, coccidioidomycosis, and histoplasmosis have all been recorded, the last leading to constriction and calcification. Pericardial calcification by hydatid disease is increasingly recognized in areas where the disease is endemic and may require surgical treatment if cardiac compression occurs.

Myocardial infarction

Evidence of acute pericarditis can be found in up to 15 per cent of patients in the first 24 to 72 h after acute myocardial infarction. This may take the form of a friction rub when infarction was transmural. It seldom gives rise to symptoms other than dull retrosternal pain, which differs from that due to the infarction itself by varying with posture and respiration. Patients with pericarditis have more extensive ST segment changes and a slightly higher risk of supraventricular arrhythmia. There is no evidence to suggest increased risk of complication with thrombolysis. Echocardiography may demonstrate a small pericardial effusion, but this is unlikely to require treatment.

Post-cardiotomy syndrome

Pericardial involvement is an important component of the post-cardiotomy syndrome. This is an acute febrile illness occurring up to 1 year after cardiac surgery. The onset is usually sudden, with pleural or precordial pain and pyrexia of up to 40°C. Chest radiography may show an enlarged heart or a pleural effusion. ECG is unaffected. The condition is usually self-limiting, but can recur. Diagnosis is by excluding, in particular, infective endocarditis or cytomegalovirus infection from blood transfusion. Treatment is with aspirin or indomethacin. Rarely a large pericardial effusion may develop, requiring surgical drainage.

Dressler's syndrome

Dressler's syndrome is similar to post-cardiotomy syndrome. It follows 2 to 4 weeks after acute myocardial infarction, in 3 to 4 per cent of cases. It is a self-limiting febrile illness, accompanied by pericardial or pleural pain, and by pneumonitis in more severe cases. Like the post-cardiotomy syndrome, it responds to aspirin, indomethacin, and (if necessary) steroids.

Rheumatic fever

A small pericardial effusion accompanies virtually all cases of acute rheumatic fever, where it is associated with epicardial inflammation and sometimes acute pericarditis. Less commonly, the effusion may be large enough to cause cardiac enlargement on chest radiography and so suggest myocardial disease. Healing is virtually complete, although rheumatic pericarditis may be responsible for adhesions found at the time of subsequent valve replacement, and it has also been invoked as a cause of subsequent constriction. The diagnosis is made echocardiographically, and the condition must be distinguished from myocarditis or severe valve disease.

Autoimmune rheumatic disorders

Pericardial involvement can be a serious manifestation of rheumatoid disease, particularly in male patients with positive serology. Transient pericardial pain, symptomatic pericardial effusion, and particularly pericardial constriction may all occur. Pericardial involvement is also common in systemic lupus erythematosus, whether spontaneous or precipitated by procainamide or hydralazine. Pericardial pain, asymptomatic effusion, and chronic constriction have all been reported. Pericardial effusion can also be seen in association with scleroderma, polyarteritis nodosa, and the Churg–Strauss syndrome, when it may accompany myocardial involvement and functional mitral regurgitation.

Renal failure

Pericarditis, commonly fibrinous and associated with a bloody effusion, is often seen in untreated or inadequately treated chronic renal failure. The usual presentation is with pericardial pain and a rub, both of which subside if an effusion develops. Tamponade is common in untreated cases. Collagenous thickening of the epicardium is less common, but may give rise to myocardial constriction. Either of these complications may need surgical relief.

Hypothyroidism

Clinically silent pericardial effusion is common in untreated hypothyroidism. The effusion itself has a high cholesterol content which may produce an unusual secondary pericarditis with cholesterol deposits of 'gold paint' appearance. The pericardial effusion very rarely needs to be treated in its own right, and subsides when thyroid replacement therapy is given.

Malignancy

Malignant involvement of the pericardium may be due to a primary tumour, or much more commonly to secondary involvement. The least rare primary tumours are mesothelioma or myosarcoma. Clinical manifestations of malignant involvement include supraventricular arrhythmias or atrial fibrillation as well as pericardial

tamponade or constriction. Malignant effusion is a very common cause of tamponade, and likely to need drainage. Positive diagnosis is best made by a limited surgical approach, allowing open biopsy and the fashioning of a window into the pleural cavity to prevent recurrence.

Irradiation

Pericarditis can be caused by irradiation. This is usually asymptomatic and a rub is unusual, but transient cardiac enlargement and minor ECG changes occur. It may occur at the time of the irradiation or at any time thereafter, and can be large enough to require drainage. The clinical picture needs to be distinguished from recurrence of malignancy. At operation, the pericardium is found to be thickened with fibrosis and dense adhesions. In a small minority of patients, pericardial constriction can develop up to 40 years after irradiation.

Haemorrhage

Haemorrhage into the pericardium is an important cause of tamponade. It may occur with aortic dissection involving the ascending aorta. If the leak is large it causes pericardial tamponade and death, but a small volume of blood is not uncommon with dissection. It can be detected by echocardiography, and may be responsible for ST segment changes on the ECG.

Pericardial haemorrhage may be the result of stab wounds or blunt injury, or may occur after cardiac surgery. It may also be induced by excessive anticoagulant therapy, or follow invasive procedures such as myocardial biopsy or pacemaker insertion.

Symptoms can occur at the time of bleeding, or may be delayed by 2 to 3 weeks, possibly because autolysis of clotted blood increases the volume of fluid within the pericardial space. Delayed tamponade causes a characteristic syndrome of elevated venous pressure, fluid retention, and low cardiac output that resembles myocardial disease. Haemorrhage into the pericardial space may also be the basis of delayed pericardial constriction occurring up to 10 years after open heart surgery.

Clinical syndromes associated with pericardial disease

Acute pericarditis

Clinical findings

There are three main components to the clinical syndrome of acute pericarditis: chest pain, pericardial rub, and ECG changes. The pain is usually retrosternal, continuous, and sharp or 'raw' in character. It is frequently aggravated by sudden movements or deep inspiration, and is relieved by sitting up. Less commonly it may resemble angina pectoris, or may be mild and 'atypical'. Painful breathing causes dyspnoea. The onset of the pain is usually sudden, but in idiopathic pericarditis, it may have been preceded by several days' malaise or other non-specific symptoms.

On examination, the main abnormality is a pericardial rub, audible in any position over the precordium. In patients in sinus rhythm it has two components, corresponding to atrial and ventricular systole. Rubs are frequently evanescent, and may vary with posture. They are often louder in inspiration. An irregular pulse due to supraventricular ectopic beats is common, particularly in patients with renal failure or after cardiac surgery. Atrial fibrillation or flutter are also seen.

The third clinical feature of the syndrome of acute pericarditis is an abnormal ECG. Symmetrical elevation of the ST segments by 1 mm or more in all leads other than aVr is seen in over 90 per cent of patients in whom the diagnosis is confirmed. Early in the illness, the T waves are upright, but over the next 2 to 3 weeks, they become flattened and inverted as the ST-segment changes regress. These T-wave changes are variable in incidence, direction, and extent. They usually resolve completely, but a minority of patients may be left with minor T-wave inversion, only to be detected many years later at a routine ECG.

Chest radiography is usually uninformative. It may show cardiac enlargement, but it is not possible to tell whether this is due to pericardial fluid, an increase in wall thickness, or enlargement of one or more cardiac chambers.

Echocardiography is the method of choice for detecting pericardial effusion (see section xxx). Cross-sectional echo may also detect pericardial adhesions that are responsible for the rub. However, acute pericarditis can occur without demonstrable pericardial effusion.

Diagnosis

The is usually straightforward, although it is possible that either the late systolic murmur of mitral prolapse or the systolic 'scratch' of Ebstein's anomaly may be mistaken for a pericardial rub. An underlying cause for acute pericarditis should always be sought, though it may not be found, and a final diagnosis of idiopathic pericarditis is probably the commonest outcome. The possibility of additional myocarditis should always be considered.

Treatment

Idiopathic acute pericarditis is usually self-limiting, requiring simple analgesics only. Since additional myocarditis is possible, the patient should rest until pain has subsided. Pericarditis due to Dressler's or the post-cardiotomy syndrome responds well to aspirin, or if more severe, to a non-steroidal anti-inflammatory drug. When symptoms are severe, repeated, or prolonged, steroids may be given empirically. Associated pericardial effusion is treated on its own merits: only rarely does it need to be drained. Supraventricular arrhythmias are treated in the standard way. When pericarditis is part of a generalized disease, this should obviously be treated appropriately.

Pericardial tamponade

Pathophysiology and clinical features

Pericardial tamponade occurs when the pressure of fluid within the pericardial cavity becomes high enough to interfere with ventricular filling. The volume of fluid needed to cause tamponade varies considerably between patients. If the fluid has collected slowly, 1 to 2 litres may be present, but if it has collected rapidly, or the pericardium is rigid, a much smaller volume will cause tamponade.

When pericardial pressure increases, right and left atrial pressures must necessarily rise to allow cardiac filling. The presence of fluid in the pericardium also reduces the volume of blood that can be accommodated in the cardiac chambers, such that stroke volume becomes small and fixed. Patients with cardiac tamponade therefore present with a high jugular venous pulse and clinical evidence of low cardiac output: the skin is cold, the pulse rate rapid but the volume small, and urine flow reduced, though systolic arterial pressure may be above 100 mmHg.

An important and characteristic physical sign in cardiac tamponade is arterial 'pulsus paradoxus'. Arterial pressure normally falls on inspiration by up to 10 mmHg. This fall is more obvious in patients with obstructive lung disease. Arterial paradox is unfortunately named: it is an accentuation of the normal response and not paradoxical. What is abnormal is the extent to which the arterial pressure falls. In severe tamponade the reduction in pulse pressure can readily be palpated at the radial artery, and with critical circulatory embarrassment the pulse may disappear altogether on inspiration. In milder cases, arterial paradox is sought using the sphygmomanometer.

The mechanism of pulsus paradoxus is still uncertain. Direct measurement of the pericardial pressure shows it to rise during inspiration, probably because the cavity is distorted by downward motion of the diaphragm. This increase is accompanied by a corresponding rise in right atrial and central venous pressure so that filling of the right side of the heart is maintained. By contrast, on the left side of the heart there is no corresponding increase in pulmonary venous pressure, which can fall to a low level compared with that in the pericardium and compromise left ventricular filling. During inspiration the interventricular septum then shifts from right to left, right ventricular stroke volume is maintained only by almost complete obliteration of the left ventricular cavity, left ventricular stroke volume drops dramatically, and profound inspiratory hypotension results. Finally, as pericardial pressure rises higher, there is diastolic collapse of right atrium and right ventricle.

Abnormal right ventricular filling is reflected in the venous pulse. The pressure is always raised: if it is not, the diagnosis of tamponade must be questioned. Usually it is very high, and it may be difficult to see the top. If a central venous line is in place, a further increase occurs with inspiration (Kussmaul's sign). This is a much less specific finding that arterial paradox and merely reflects the inability of the right heart to deal with an increase in stroke volume. It is seen in a variety of conditions including right ventricular disease and pulmonary hypertension. Although X and Y descents are visible, their amplitude is small, since the main disturbance is an increase in mean venous pressure. Unlike pericardial constriction, therefore, abnormalities in the form of the venous pulse are not particularly helpful in making the diagnosis.

Investigations

Chest radiography shows a large globular heart (Fig. 1), similar to that seen in dilated cardiomyopathy. More useful in making the diagnosis, therefore, is the absence of any evidence of pulmmonary congestion, which would be expected if myocardial disease were the main abnormality. Pulmonary oedema is most unusual in pure tamponade: if it is present, it suggests additional myocardial disease.

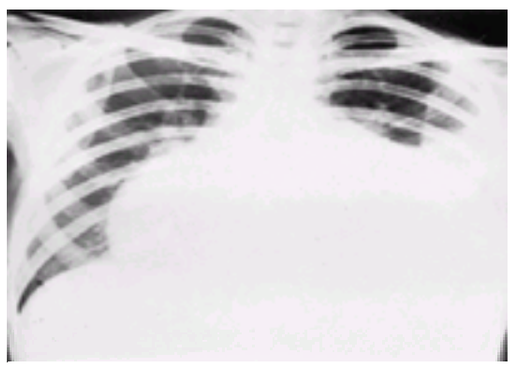

Fig. 1 Posteroanterior chest radiograph of a patient with a large pericardial effusion. The heart shadow is greatly enlarged and globular in configuration. The lung fields are normal.

ECG shows tachycardia, often with low-voltage QRS complexes, but without Q waves or conduction disturbances. If the effusion is large, electrical alternans is present, when alternate QRS complexes show differing morphology (Fig. 2), because the heart swings to and fro in a large (therefore usually malignant) effusion.

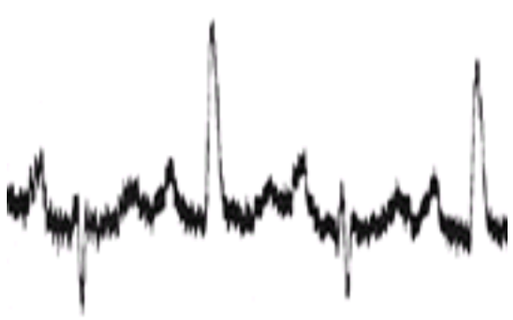

Fig. 2 Electrocardiogram from a patient with massive malignant pericardial effusion showing electrical alternans. Note that all are sinus beats with the same PR interval, but that the QRS axis alternates.

Echocardiography is a most important investigation since it allows rapid and unequivocal diagnosis of pericardial effusion, which is usually large with tamponade (Fig. 3). Evidence for circulatory embarrassment is diastolic collapse of the right ventricle or right atrium (Fig. 4), and a striking increase in the amplitude of septal motion with respiration. If electrical alternans is present, motion of the heart within the pericardium can be confirmed.

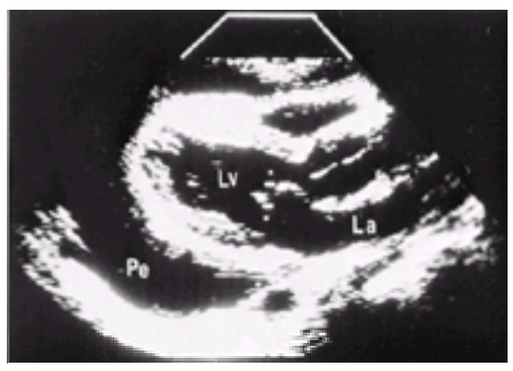

Fig. 3 Two-dimensional echocardiogram, parasternal long-axis view, showing a large pericardial effusion (Pe) posterior to the left ventricle (Lv). La, left atrium.

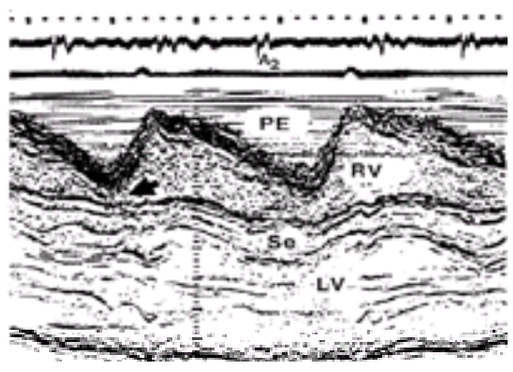

Fig. 4 M-mode echocardiogram showing diastolic collapse of the right ventricle (marked by arrow) in a patient with a large pericardial effusion. Note the minimum dimension of the right ventricle occurs at the end of the diastole. PE, pericardial effusion; RV, right ventricle; Se, interventricular septum; A2, aortic valve closure on phonocardiogram; LV, left ventricle (time marker = 200 ms)

Arterial pulsus paradox can be confirmed or excluded from a simultaneous trace of respiration and peripheral arterial Doppler (Fig. 5).

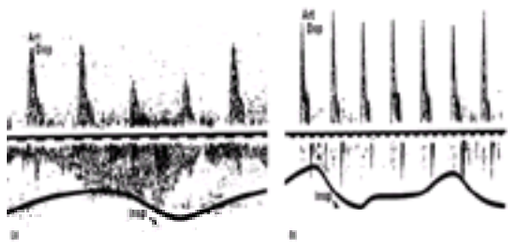

Fig. 5 (a) Arterial Doppler variation with respiration recorded from a patient with a large pericardial effusion. Note that peak arterial velocities drop to approximately half their peak values with inspiration. Art Dop, arterial Doppler; insp, inspiration. (b) The same patient after aspiration of the pericardial effusion. Note that the arterial pulse no longer varies with respiration (time marker = 100 ms).

The circulatory embarrassment occurring with pericardial effusion varies in its exact nature between cases. 'Tamponade' does not therefore represent a uniform diagnosis. In the small minority of patients in whom an echocardiographic diagnosis of pericardial effusion cannot be made for technical reasons, some other imaging method such as CT or MRI may have to be used. Cardiac catheterization is no longer necessary.

Differential diagnosis

The main step in making the diagnosis of pericardial tamponade is to think of it in a patient presenting with clinical evidence of a low cardiac output. The condition must be distinguished from severe ventricular disease, massive pulmonary embolism, hypovolaemia, or overwhelming sepsis. Hypovolaemia is ruled out by the high venous pressure, whilst the absence of added heart sounds and pulmonary congestion makes severe ventricular disease unlikely. Massive pulmonary embolism is accompanied by a right ventricular third sound and characteristic ECG abnormalities.

An echocardiogram should be obtained early in all patients with low cardiac output for which the cause is not apparent, and if there is a large pericardial effusion the diagnosis of tamponade becomes very likely. It is essential for the occasional echocardiographer to distinguish pericardial effusion from pleural effusion. This is done by locating the high-intensity echo from the fibrous pericardium posterior to the left ventricle on the left parasternal view. A pericardial effusion is inside this structure, and a pleural effusion outside. Rarely, a large pleural effusion may compress the heart and cause a clinical picture very similar to tamponade in the absence of any pericardial fluid. This seems to occur when the pleural effusion is under pressure, and haemodynamics rapidly return to normal with pleural drainage.

Treatment

Pericardial tamponade is a medical emergency. It needs urgent treatment, particularly if there is obvious arterial paradox, or if the effusion is of recent onset and fluid is collecting rapidly. Pericardial aspiration should be performed in an area where resuscitation facilities are available. Echocardiography is used to determine where to insert the needle, and to get some idea of the direction and depth. Subcostal or apical routes are possible, but the former is preferable if the heart is accessible in this way, since damage to the anterior descending coronary artery is possible from the apex. The depth of the pericardial fluid can usually be confirmed when the local anaesthetic is inserted. A larger needle or polythene cannula is then introduced into the effusion and a pig-tail catheter inserted over a guide wire. A maximum of 500 ml of fluid is removed initially and relieves any haemodynamic problem: rapid withdrawal of larger volumes can provoke cardiovascular collapse. Continuous drainage is then instituted and the remainder of the effusion drained over 12 to 24 h.

Many pericardial effusions, particularly malignant ones, are heavily bloodstained. They can be distinguished from blood associated with puncture of a chamber by their colour, since they are very desaturated, and by their failure to clot, since they are defibrinated. If necessary, the haematocrit of the fluid can be compared with that of blood taken simultaneously.

Aspiration of a pericardial effusion is necessary when there is any suspicion of tamponade. It should also be considered when the volume is large, even in the absence of specific evidence of resting circulatory embarrassment, since exercise tolerance is commonly limited before overt tamponade develops. In addition, such patients are unstable and tamponade can develop quickly with the accumulation of a relatively small additional volume of fluid.

Aspiration is not necessarily the best way of definitively managing pericardial effusion. It does not prevent recurrence, and it is not usually possible to make a diagnosis from the pericardial fluid alone. The most satisfactory line of treatment, therefore, is to undertake limited thoracotomy, either through the fifth interspace, or subcostally, the latter operation being possible with local anaesthetic. This allows an adequate specimen of pericardium to be removed under direct vision for histology, and drainage of the pericardial space can be assured by making a window to the pleura. It is also possible to deal with a loculated effusion and to remove blood clots whose presence can give rise to delayed tamponade.

Pericardial constriction

Pericardial constriction is the haemodynamic disturbance caused when ventricular filling is limited by the pericardium. The pericardium itself is usually, but not always, thickened. The myocardium may also be involved, particularly in its subepicardial layers, by atrophy and fibrosis. Constriction usually affects both ventricles symmetrically, but in rare cases it may be localized. The majority of cases, particularly in the developed world, show no evidence of inflammation, acute or chronic, so 'pericardial constriction' is a better name than 'constrictive pericarditis'.

Pathophysiology and clinical features

Pericardial constriction prevents cardiac filling in late diastole. Since the two sides of the heart are usually affected symmetrically, right and left atrial filling are equally compromised. Early diastolic ventricular pressure is normal, but since the pericardium is effectively indistensible, a normal or reduced stroke volume causes a striking increase in filling pressure. End-diastolic pressures are equal to within 1 to 2 mmHg in all four cardiac chambers. This persists with respiration or even with fluid loading, and is the main criterion on which the invasive diagnosis of constriction is based. The ventricular pressure trace during filling is also characteristic. It rises rapidly in early diastole, and then stops rising abruptly, often with a slight rebound, and remains constant for the remainder of diastole. This pattern is often referred to as the 'square root sign' from a fancied resemblance to the mirror image of the mathematical symbol for a square root. Abnormal early diastolic filling is also reflected in the transmitral Doppler trace, which shows a rapid early diastolic deceleration, and reduced or absent flow across the valve during atrial systole.

The jugular venous pulse is also characteristic. Overall pressure is raised, with the dominant descent during systole, the X descent (Fig. 6). This descent is independent of right atrial systole, occurring later in the cardiac cycle than the A wave and persisting with atrial fibrillation. Flow towards the heart in the superior cava is also systolic, meaning that right atrial volume must also be increasing at this time. The unexpected combination of an increase in right atrial volume with a simultaneous fall in right atrial pressure is caused as follows. In pericardial constriction, an increase in the transverse dimension of the ventricle is limited by the pericardium, but increase in the longitudinal axis is not. Long-axis changes are brought about mainly by motion of the atrioventricular ring, and during ventricular ejection both atrioventricular rings move towards the cardiac apex. This enlarges the capacity of the atria, draws blood in from the vena cavae, and manifests as a dramatic X descent in the jugular venous pulse.

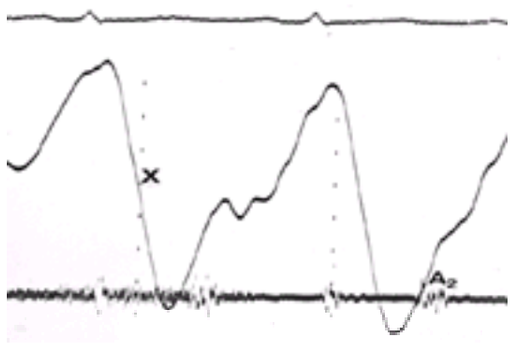

Fig. 6 Jugular venous pulse recording showing a dominant X descent in a patient with pericardial constriction. A2, aortic valve closure; (time marker = 500 ms).

The clinical picture of pericardial constriction is dominated by obstruction to right ventricular filling. The jugular venous pressure is always raised, in well developed cases by 15 cm or more, showing abrupt systolic, and to a lesser extent, early diastolic descents whether or not the patient is in atrial fibrillation. Tachycardia and atrial fibrillation are common. The precordial impulse is not usually palpable, and on auscultation the heart sounds are soft. There may be an early diastolic sound, whose timing corresponds to the end of rapid filling, and which should therefore be classified as a third sound, though it is sometimes referred to as a ventricular 'knock'. It is often earlier than the classic third heart sound, but only because rapid filling ends earlier in constriction than in uncomplicated ventricular disease. The liver is enlarged, and in patients with long-standing disease, there may be wasting and jaundice. Ascites is often more prominent than peripheral oedema, particularly when the patient is stabilized on large doses of diuretic.

Investigations

The chest radiograph is usually normal, but the heart may be enlarged and it is important to look for pericardial calcification. This appears as multiple plaques or, more frequently, as a rim covering the diaphragmatic and anterior surfaces of the heart.

The ECG often shows atrial fibrillation, low-voltage QRS complexes and non-specific T-wave abnormalities. There are no diagnostic features.

CT scanning or MRI can demonstrate the extent and distribution of pericardial thickening. While this does not make the diagnosis of constriction, it is often very useful to know that the pericardium is actually abnormal in a patient in whom this diagnosis is suspected on clinical grounds. M-mode and cross-sectional echocardiography are unhelpful in making the diagnosis of constriction. Doppler may be useful in demonstrating abnormalities of ventricular filling.

Unless the diagnosis is very obvious, cardiac catheterization is still usually performed. To establish the diagnosis, three features should be present:

1. a difference of less than 5 mmHg between the equal end-diastolic pressures in the two ventricles, persisting with respiration;
2. a peak right ventricular pressure of less than 50 mmHg; and
3. a ratio of end-diastolic to peak right ventricular pressure of more than 0.33.

It is still uncertain whether the normal pericardium can ever cause constriction in humans. The pericardium can stretch and accommodate a gradual increase in heart size. If constriction were to occur, it would probably be in the setting of rapid increase in ventricular size, for example in myocarditis or valvular regurgitation of acute onset. In these circumstances it would be difficult to dissociate from the primary manifestations of ventricular disease.

Alternative clinical presentations

A number of less common clinical presentations have been described.

1. Localized constriction may compress the outfow tract of the right ventricle or may mimic mitral or tricuspid stenosis.
2. A greatly raised venous pressure may lead to the clinical features of a protein-losing enteropathy or classic nephrotic syndrome.
3. The possibility that occult pericardial constriction might exist has been raised. In these patients, the resting venous pressure is normal and the symptoms are non-specific, including mild limitation of exercise tolerance or fatigue. There may be a history of previous acute pericarditis. The diagnostic haemodynamics of constriction can be unmasked by rapid volume infusion.

Differential diagnosis

The main differential diagnosis of pericardial constriction is restrictive myocardial disease, where the passive properties of the myocardium itself are abnormal, usually as the result of fibrosis or infiltration. The haemodynamics are very similar in the two conditions: ventricular early diastolic pressure is normal, but that at end-diastole is greatly increased. The differential diagnosis is important: both are debilitating and life-threatening conditions, but whereas constriction can frequently be treated effectively by surgery, restrictive myocardial disease cannot (other than by transplantation). Approaches used to distinguish between the two conditions include the following.

Anatomical

If the pericardium is thickened or calcified, a diagnosis of constriction is very likely. Similarly, if echocardiography shows the characteristic appearances of amyloid then it is very likely that restrictive myocardial disease is present. The same applies when only one ventricle is greatly dilated with a reduced ejection fraction. In typical restrictive disease, however, left ventricular end-systolic cavity size is normal, although stroke volume may be reduced. Tricuspid regurgitation severe enough to lead to a clinical picture resembling either condition can readily be diagnosed by echocardiography.

Haemodynamics

Raised ventricular filling pressures, with a square root sign on the pressure pulse, increased early diastolic filling velocities, and shortened early filling periods do not distinguish between the two. All three catheter criteria mentioned above should be present before a definitive diagnosis of constriction is made. A dominant systolic X rather than Y descent on the jugular venous pulse is also very characteristic of constriction, probably because longitudinal as well as circumferential filling is impaired in restrictive myocardial disease.

Clinical progress

With diuretic treatment, the venous pressure usually drops in patients with restrictive myocardial disease, albeit at the cost of causing fatigue and hypovolaemia. It is very rare to be able to bring the venous pressure down to normal in a patient with well developed constriction.

This discussion is based on the assumption that constriction and restriction are independent conditions, and that a patient has either one or the other. However, this is not always the case. When the visceral pericardium (epicardium) is involved, fibrosis can spread to involve the myocardium and such cases may show features of both constriction and restriction. This state of affairs is analogous to that seen with subendocardial thickening as occurs in eosinophilic heart disease, which is usually classified as a form of restrictive cardiomyopathy. For this reason, in particular, the differential diagnosis between constriction and restriction may not be clear even after extensive investigation, and in a minority of cases it is not possible to avoid an exploratory thoracotomy. This enables the diagnosis of constriction to be made and treated accordingly, or to be definitively excluded, so that the patient can be reconciled to medical treatment, unsatisfactory as it may be.

Pericardial constriction must also be distinguished from other causes of raised venous pressure. Superior caval obstruction is excluded by the presence of venous

pulsation. Right ventricular inflow may be obstructed by tricuspid stenosis or, very rarely, by a right ventricular tumour. Elevation of the venous pressure is also a feature of selective right atrial compression by blood clot occurring in the postoperative period (see later). Severe tricuspid regurgitation may occur on its own, or because of right ventricular disease. It seems to be becoming increasingly common as a long-term complication following mitral valve replacement. These possibilities can all be excluded by echocardiography.

Treatment

Mild pericardial constriction can usually be managed by diuretics. Although the venous pressure does not fall to normal, fluid retention can often be controlled. However, if fluid retention persists, or if an excessive dose of diuretic is needed, as shown by an increase in blood urea or impaired exercise tolerance due to fatigue, then surgery should be considered. The thickened pericardium must be removed from the anterior and inferior surfaces of the heart, and from the atrioventricular sulci. Cardiopulmonary bypass is usually needed to expose and decompress the heart satisfactorily. The operation is often a long and difficult one, particularly when the pericardium is calcified or when there is fibrosis of the myocardium. In many patients, the venous pressure is as high after the operation as it was before, though the X descent is lost and the Y descent becomes dominant. However, with digitalis and diuretic treatment, the pressure gradually falls over the succeeding weeks, as the condition of the patient improves.

Other manifestations of pericardial disease

Postoperative pericardial disease

A modified type of pericardial tamponade occurs after open heart surgery due to blood clots within the pericardium, particularly behind the left atrioventricular sulcus. Clinically this presents as a fall in urine flow and cardiac output, a reduction in skin temperature and finally hypotension. The atrial pressures may be normal or raised, and the classic arterial and venous pulse abnormalities are absent. Chest radiography and ECG show no specific abnormality. The transthoracic echocardiographic window is often poor immediately after surgery, but transoesophageal echo may show clot alongside the heart, compressing one or more chambers. The condition should be suspected in a patient who may have bled rather heavily after operation, particularly when the blood flow from the chest drains suddenly falls, and is most satisfactorily diagnosed by reopening the chest and removing the blood clots.

Clot may also compress the right atrium. This characteristically occurs towards the end of the first postoperative week, after the chest drains have been removed and when the patient is being mobilized. The main clinical features are fluid retention and elevation of the venous pressure. The diagnosis can usually be made by transthoraic echocardiography, which demonstrates distortion of the right atrial cavity by blood clot and sometimes increased right atrial filling velocities. If the precordial window is poor, transoesophageal echocardiography is required. Treatment is by drainage.

Pericardial constriction is increasingly being recognized as a long-term complication of cardiac surgery. It has a major effect on the clinical course of 1 to 2 per cent of patients; minor degrees are probably rather more common. It presents as chronic elevation of the venous pressure, and is often diagnosed as postoperative 'heart failure'. The diagnosis is suspected from the absence of any intracardiac cause of the syndrome of heart failure, such as ventricular disease, valvular regurgitation, or pulmonary hypertension, and from pericardial thickening demonstrated by CT or MRI. It can usually be controlled by a small dose of diuretic, but in a minority of cases, pericardial surgery may be needed.

Recurrent acute pericarditis

Recurrent acute pericarditis is an uncommon but clinically demanding form of pericardial disease to manage. It occurs at any time up to 10 years after an apparently uncomplicated episode of acute pericarditis of any aetiology. The commonest manifestation is chest pain, although rarely it may present as recurrent pericardial effusion. ECG changes and echocardiographic evidence of effusion occur in about half of cases. As with the original attack, immunological studies are likely to be indecisive. The clinical problem is that repeated episodes can become debilitating to the patient, particularly as they occur after what was represented as a self-limiting disease. Constriction and myocardial disease are significant complications. Management consists of maintaining a positive outlook and controlling the manifestations of acute pericarditis. Simple analgesia with aspirin is the most satisfactory means, but this may not always be adequate. Non-steroidal anti-inflammatory agents or corticosteroids may be required, the latter sometimes in large enough doses to lead to Cushingoid manifestations. There is no evidence to suggest that immunosuppressive agents have a therapeutic role. Pericardiectomy may be necessary, but is not necessarily effective, presumably because all pericardium cannot be removed. The overall prognosis of the condition is good.

Tuberculous pericardial constriction in the Third World

This runs a very different course from that seen in developed countries. In the absence of HIV infection, it occurs early in the disease and may be the presenting feature. Alternatively, it may supervene after an effusion has been drained. Patients present with sinus tachycardia rather than atrial fibrillation, a very high venous pressure, ascites, and weight loss. The venous pressure often does not show the characteristic pattern of systolic dip, and a third heart sound is present in about 50 per cent of cases. Chest radiography shows a normal-sized heart, but characteristically a 'shaggy' left heart border. There is no pericardial calcification. ECG shows sinus tachycardia and non-specific T-wave abnormalities. Cross-sectional echocardio-graphy is very helpful, showing the two layers of pericardium separated by amorphous echoes often enclosing small loculated pockets of fluid (Fig. 7). This pattern is sometimes referred to as 'effusive-constrictive' pericarditis. The amplitude of ventricular wall motion is reduced, and the pericardial surface shows a very characteristic 'frozen' appearance. Treatment is with antituberculous chemotherapy. Added steroids help, with heart rate and venous pressure returning more rapidly to normal. They also reduce the risk of death and the requirement for operation. When possible, fluid retention should be controlled by diuretics. During the early subacute phase, surgery is demanding and unsatisfactory. It may be needed, however, in a minority of seriously ill patients who cannot be held on medical treatment. In the absence of AIDS, ultimate prognosis is excellent, being indistinguishable from that of the population at large whether treatment is medical or surgical.

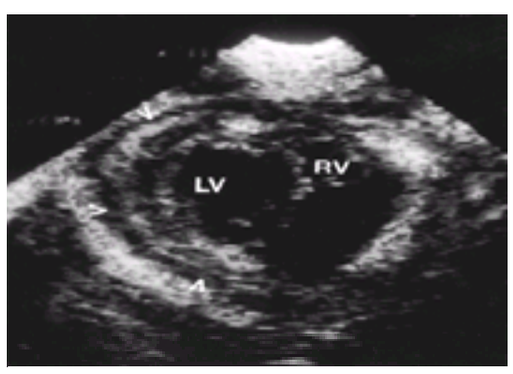

Fig. 7 Cross-sectional echocardiogram showing effusive-constrictive pattern of pericardial involvement (marked with arrow) in a patient with tuberculous pericarditis.

Further reading

Baldwin JJ, Edwards JE (1976). Uremic pericarditis as a cause of tamponade. *Circulation* **53**, 896–901.

Caird R, Conway N, McMillan IKR (1973). Purulent pericarditis followed by early constriction in young children. *British Heart Journal* **35**, 201–3.

Carty JE, Deverall PB, Losowsky MS (1975). Retrosternal pain, widespread T wave inversion and collapse of left lower lobe with effusion, strangulated atrial appendix. *British Heart Journal* **37**, 98–100.

Dresler W (1959). The post-myocardial infarction syndrome. A report of 44 cases. *Archives of Internal Medicine* **103**, 28–20.

Fowler NO, Harbin III AD (1986). Recurrent acute pericarditis: follow-up study of 31 patients. *Journal of the American College of Cardiology* **7**, 300–5.

Hatle LK, Appleton CP, Popp RL (1989). Differentiation of constrictive pericarditis and restrictive cardiomyopathy by Doppler echocardiography. *Circulation* **79**, 357–70.

Heidenreich PA *et al*. (1995). Pericardial effusion in AIDS. *Circulation* **92**, 3229–34.

Kahn AH (1975). Pericarditis of myocardial infarction. *American Heart Journal* **90**, 788–94.

Martin RG *et al*. (1975). Radiation induced pericarditis. *American Journal of Cardiology* **35**, 217–20.

Perheentupa J *et al*. (1973). Mulibrey nanism, an autosomal recessive syndrome with pericardial constriction. *Lancet* **ii**, 351–5.

Spodick DH (1974). ECG in acute pericarditis. *American Journal of Cardiology* **40**, 470–4.

Strang JIG (1984). Tuberculous pericarditis in Transkei. *Clinical Cardiology* **7**, 667–70.

Tubbs OS, Yacoub MH (1968). Congenital pericardial defects. *Thorax* **23**, 598–607.

Vaitkus PT, Kussmaul WG (1991). Constrictive pericarditis versus restrictive cardiomyopathy: a reappraisal and update of diagnostic criteria. *American Heart Journal* **122**, 1431–41.

Watters DAK (1997). Surgery for tuberculosis before and after human immunodeficiency virus infection: a tropical perspective. *British Journal of Surgery* **84**, 8–14.

Wood P (1961). Chronic constrictive pericarditis. *American Journal of Cardiology* **7**, 48–55.

15.10.1 Acute rheumatic fever

Jonathan R. Carapetis

Introduction

Acute rheumatic fever is an immunologically mediated, multisystem disease induced by recent infection with group A streptococcus. Most medical practitioners in industrialized countries will rarely, if ever, see a case. However, the dramatic decline in incidence of acute rheumatic fever in industrialized countries during the second half of the twentieth century is not replicated in many developing countries, or among some indigenous and other populations living in poverty in industrialized countries. Moreover, acute rheumatic fever has recently returned as an important public health problem in some middle-class regions of the United States. Rheumatic heart disease remains the most common acquired heart disease of childhood in the world.

Epidemiology

The highest reported annual incidence of acute rheumatic fever, more than 500/100 000, occurs in the Aboriginal population of the Northern Territory of Australia. Populations in developing countries commonly have incidence rates between 50 and 200/100 000 per year. There have been dramatic declines in recent decades in many Latin American and Asian countries with improving economic and living conditions. In most populations with high incidence rates, the predisposing conditions are those that promote endemicity and high levels of transmission of group A streptococci: these include overcrowded housing, poor personal and community hygiene, poor access to medical services and, in some circumstances, widespread skin infection and scabies infestation.

Outbreaks of acute rheumatic fever occurred in middle-class areas of the United States during the 1980s and 1990s. These outbreaks arose because of the emergence of virulent strains of group A streptococci, particularly belonging to M serotypes 1, 3, and 18. By contrast, outbreaks of acute rheumatic fever have rarely, if ever, been described from developing countries; most cases appear to arise from the ongoing circulation of pathogenic group A streptococcal strains in the population.

Recurrent episodes are almost as common as primary episodes in many populations with high incidence rates of acute rheumatic fever, and account for approximately 40 per cent of all episodes among the Aboriginal population of northern Australia. Recurrences may lead to accumulated cardiac valvular damage and are therefore responsible for many cases of rheumatic heart disease, yet they are almost entirely preventable using secondary prophylaxis (see later).

In many developing countries females are affected more than males, usually in the ratio between 1.3 and 2 to 1. In affluent countries males and females appear to be affected equally. The gender association is stronger for rheumatic heart disease (especially mitral stenosis) than acute rheumatic fever; this may reflect a greater tendency to recurrences among females. Any female preponderance may relate to inherited characteristics, to greater exposure to group A streptococci because of the increased involvement of girls and young women in child-rearing in most cultures, or to reduced access by females to primary and secondary prophylaxis.

The maximum incidence of acute rheumatic fever is between the ages of 5 and 15 years in all populations. Approximately 5 per cent of cases occur in children younger than 5 years, but very rarely are children younger than 3 years affected. This age distribution parallels that of group A streptococcal pharyngitis, and supports the hypothesis that all cases of acute rheumatic fever follow this condition. However, it may be that cases do not occur in infants or very young children because of the need for maturity of the immune system (particularly of cellular immunity), or sensitization of the immune response by prior streptococcal infections. New cases occur occasionally up to age 30, but rarely beyond. Hypotheses to explain the reduced incidence in adulthood include development of non-type-specific immunity to primary group A streptococcal infections, further maturation of immune responses, or reduced sensitization by recurrent streptococcal infections.

Pathogenesis

Despite a century of research, the pathogenesis of acute rheumatic fever remains incompletely understood. The presumed pathogenetic pathway is summarized in Fig. 1.

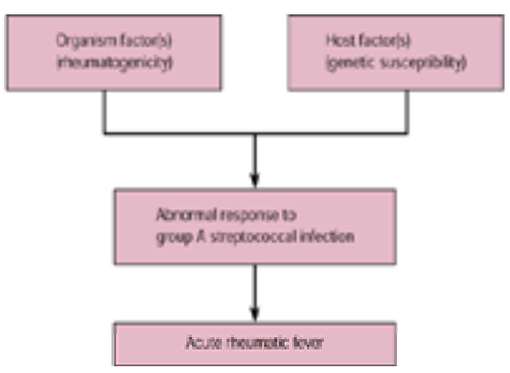

Fig. 1 Simplified approach to understanding the pathogenesis of acute rheumatic fever. (Adapted with permission from Carapetis JR, Currie BJ, Good M (1996). Editorial review: Towards understanding the pathogenesis of rheumatic fever. *Scandinavian Journal of Rheumatology* **25**, 127–31.)

Host factors

Epidemiological evidence suggests that less than 5 or 6 per cent of people have the potential to develop acute rheumatic fever after relevant streptococcal exposure, and that this proportion does not vary substantially between populations. Attack rates of acute rheumatic fever after untreated group A streptococcal pharyngitis vary from less than 1 to 3 per cent. Genetic susceptibility to acute rheumatic fever was first suggested by its familial aggregation and by a greater concordance in monozygotic than in dizygotic twins. The mode of inheritance is uncertain; autosomal recessive or autosomal dominant with partial penetrance have been suggested.

The basis for genetic susceptibility is not known. Recent work suggests an association of rheumatic heart disease with certain HLA class II alleles. A B-cell alloantigen (D8/17) is expressed in a high percentage of B cells from patients with acute rheumatic fever and their family members in many populations. However, D8/17 may not predict susceptibility in all populations; recent studies in India suggest that different B-cell alloantigens may identify patients with acute rheumatic fever there.It is not yet clear whether these putative markers are genetic or induced by streptococcal infection as part of the pathogenesis of acute rheumatic fever.

Organism factors

The observation that outbreaks of pharyngitis due to certain serotypes of group A streptococcus resulted in high attack rates of acute rheumatic fever, whereas no cases occurred after infection with other serotypes, led to the concept of 'rheumatogenicity'—that only some strains of group A streptococcus have the potential to cause acute rheumatic fever. M serotypes 1, 3, 5, 6, 14, 18, 19, 24, 27, and 29 have been most frequently implicated. However, there may be substantial genetic diversity among strains belonging to a particular M serotype, and not all strains of 'rheumatogenic serotypes' appear to cause acute rheumatic fever. Therefore, rheumatogenicity may be strain specific rather than serotype specific; that is, any group A streptococcus may acquire the potential to cause acute rheumatic fever.

The pathogenic factor(s) are not known. Parts of the organism have immunological cross-reactivity with human tissue; there is close homology between regions of the M protein and human myosin, tropomyosin, and keratin. Other components of group A streptococci, including the hyaluronic acid capsule, the cell-wall associated group-specific carbohydrate, and the cell membrane, cross-react with a variety of human tissues damaged in acute rheumatic fever, including components of heart muscle and valves, joints, and brain. Acute rheumatic fever-associated strains of group A streptococcus also tend to be heavily encapsulated with hyaluronic acid, and not to express opacity factor. Group A streptococci possess components which act as superantigens, selectively stimulating subsets of T cells without the need for antigen presentation. Their role in acute rheumatic fever pathogenesis is not yet clear.

Site of infection

Although it is widely accepted that acute rheumatic fever may result from group A streptococcal infection of the upper respiratory tract, but not of the skin, there is some evidence that this may not always be the case. Upper respiratory tract infection certainly accounts for most, if not all, episodes of acute rheumatic fever in countries with a temperate climate. However, in tropical countries where streptococcal impetigo is highly endemic but group A streptococcal pharyngitis less common, it may be that skin infection accounts for many cases of acute rheumatic fever, either *de novo* or after subsequent throat infection. Determining whether group A streptococcal skin infection may have a role in pathogenesis of acute rheumatic fever would have enormous public health implications, as it may redirect present approaches to primary prevention (see later).

The immune response

The finding of immunological cross-reactivity led to initial enthusiasm for the role of humoral immunity in the pathogenesis of acute rheumatic fever. This was supported by the finding of anti-group A streptococcal antibodies cross-reactive with heart, joint, and brain in the sera of patients with acute rheumatic fever. Immunoglobulin and complement deposits have also been demonstrated in damaged heart tissue.

More recently, patients with acute rheumatic fever were found to have elevated levels of most markers of cellular immune activation, including circulating CD4 lymphocytes, interleukins (**IL**)-1 and -2, IL-2 receptor-positive T cells, neopterin, tumour necrosis factor-a receptors, natural killer cell cytotoxicity, T-cell responsiveness to group A streptococcal antigens, and others. T-cell and histiocytic-cell infiltrates are also present in valvular and myocardial tissue during acute rheumatic fever. This has led to theories that the primary damage in acute rheumatic fever may be due to cell-mediated immune responses and that the humoral response may be secondary to antigens released from already damaged tissues. Cross-reacting antigens of group A streptococci may be presented to the immune system abnormally, or they may be abnormally recognized by helper T cells, resulting in uncontrolled activation of cellular immunity. The resulting damage targets those tissues for which the inducing strain has sequence mimicry.

Clinical manifestations

There is always a latent period between group A streptococcal infection and the development of acute rheumatic fever. This varies from 1 to 5 weeks in most cases (usually about 3 weeks), but may be shorter in recurrences. Chorea may occur up to 6 months after the precipitating streptococcal infection. The preceding infection is asymptomatic in about two-thirds of cases.

The tissues most commonly affected are the heart, joints, and brain. Although the symptoms due to each can be disabling in the short term, only cardiac damage may be permanent and progressive. Therefore, the focus in controlling or treating acute rheumatic fever is always to prevent the development of rheumatic heart disease.

The frequency with which the various clinical manifestations have occurred in recent descriptions of acute rheumatic fever is listed in Table 1.

Carditis

Although inflammation in acute rheumatic fever may affect the pericardium (causing pericardial rubs and occasionally pleuritic chest pain) or the myocardium (sometimes causing cardiac failure, and evident on biopsy with pathognomonic Aschoff bodies), endocardial inflammation is the most important cause of cardiac damage. If acute cardiac failure or chronic cardiac disease occur, they are almost always due to damage to the cardiac valves.

A murmur is the most common evidence of acute valvular disease, usually the apical pan-systolic murmur of mitral regurgitation, with or without a low-pitched mid-diastolic (Carey–Coombs) murmur. Occasionally an aortic regurgitant murmur may be heard, mainly in older adolescents or young adults. Murmurs of tricuspid or pulmonary regurgitation are rare and are usually secondary to increased pulmonary venous pressures resulting from mitral regurgitation or stenosis. Sinus tachycardia or gallop rhythms may also be present in acute carditis.

Valves affected by rheumatic carditis may have a characteristic appearance or pattern of regurgitation on Doppler echocardiography (when interpreted by experienced technicians), which may be found even in the absence of a cardiac murmur. This may be useful for diagnosis when other clinical manifestations are not definitive. However, echocardiographic criteria have not yet been standardized, and it is difficult to distinguish acute carditis from previous rheumatic valve damage.

Mitral or aortic stenosis may develop as later complications of severe and/or recurrent acute carditis due to scarring and contraction following the acute inflammatory process. Rarely, mitral stenosis may occur in young children with acute rheumatic fever—so-called 'juvenile mitral stenosis'—the reasons for the development of this condition are not clear.

Damage to the electrical conduction pathways may result in prolongation of the P–R interval on electrocardiography. Although a subset of healthy people may have this finding, the presence of a prolonged P–R interval that resolves over the ensuing few days to weeks may be a useful diagnostic feature in cases where the clinical

manifestations are not clear. Occasionally, in the acute phase, second- or third-degree heart block or a nodal rhythm may be present (Fig. 2).

Fig. 2 Electrocardiographic changes in a young adult with acute rheumatic fever, showing evolution over 18 days from complete heart block, to second-degree (Wenckebach) block, to first-degree block, and then to normal sinus rhythm. (Reproduced with permission from Bishop W *et al*. (1996). A subtle presentation of acute rheumatic fever in remote northern Australia. *Australian and New Zealand Journal of Medicine* **26**, 241–2.)

Arthritis

The characteristic joint manifestation of acute rheumatic fever is severe, large-joint, migratory polyarthritis. The knees, ankles, wrists, and elbows are most commonly involved; only rarely, and usually only when the patient is untreated for several days, are the hips or small joints of the hands or feet inflamed. One joint characteristically becomes exquisitely painful and inflamed as another is waning. Most patients have only one or two joints affected at any one time, and each joint may be involved for just a few hours or up to 1 or 2 days. The arthritis is so responsive to non-steroidal anti-inflammatory medication that its persistence more than 1 or 2 days after commencing high-dose aspirin should lead one to consider alternative diagnoses.

Arthritis of a single large joint may occur in acute rheumatic fever, although other causes (including septic arthritis) should first be excluded. Arthralgia (joint pain without objective evidence of inflammation) is usually migratory and affects large joints, and like the arthritis of acute rheumatic fever is very responsive to anti-inflammatory medication.

Sydenham's chorea

In 1686 the English physician Thomas Sydenham described rheumatic chorea, initially naming it 'St Vitus' dance'. It is the most intriguing manifestation of acute rheumatic fever, particularly as it commonly occurs in the absence of other manifestations, usually follows a prolonged latent period (up to 6 months) after the precipitating group A streptococcal infection, and occurs most commonly in females (and almost never in post-pubertal males). The rapid, jerky, involuntary movements affect predominantly the upper limbs and face, may be asymmetrical, and may be sufficiently severe to render the patient unable to eat, drink, walk, or perform other activities of daily living. Mild chorea can sometimes be detected by having the patient join palms above the head to reveal occasional twitches of the arms or the head. Typical signs include the 'milk-maid's grip' (rhythmic squeezing when the patient grasps the examiner's fingers), spooning of extended hands (caused by flexion of the wrists and extension of the fingers), darting of the protruded tongue, and the 'pronator sign' (the arms and palms turn outwards when held above the head). As with other forms of chorea, the disorder usually becomes more evident with anxiety or purposeful movements (such as drinking or writing). Movements may appear semi-purposeful, and symptoms subside during sleep. Sydenham's chorea often is associated with excessive emotional lability or personality changes: these may precede the abnormal movements.

Most patients can be reassured that Sydenham's chorea will resolve completely and leave no long-lasting effects, usually within 6 weeks and almost always within 6 months, but rarely lasting up to 3 years.

Subcutaneous nodules and erythema marginatum

Both of these manifestations are found in less than 2 per cent of patients with acute rheumatic fever, although they were described in up to 10 to 20 per cent of patients in earlier studies from the United States and the United Kingdom. Subcutaneous nodules are firm, painless lumps, usually between 0.5 and 2 cm in diameter, commonly found in crops of three or more, and usually appear 2 to 3 weeks after the onset of acute rheumatic fever. They occur mainly over extensor surfaces or bony protuberances, particularly the hands, feet, occiput, and back. The nodules are similar, though often smaller, to those found in rheumatoid arthritis, and are most likely to be associated with severe carditis. Nodules usually last from a few days to 2 or 3 weeks.

The characteristic rash, erythema marginatum, appears as a light pink macule that spreads outwards with a serpiginous, well-demarcated edge, while the central portion clears. It appears, disappears, or moves before the observer's eyes. Multiple areas are often involved, usually over the trunk, occasionally over the proximal portions of the limbs, but rarely, if ever, the face. It usually appears together with the other initial symptoms of acute rheumatic fever, but may recur intermittently for weeks or even months. This does not indicate ongoing rheumatic inflammation, and patients can be reassured that the rash will eventually disappear without complications.

Fever

With the exception of those with pure chorea, 90 per cent of patients will have a temperature at presentation higher than 37.5°C. Although it has been reported that the temperature usually exceeds 39°C, others have found only 25 per cent of confirmed cases with fever to that level. Any temperature above 37.5°C should be considered a minor manifestation. As with arthritis, fever is very sensitive to anti-inflammatory medication, usually resolving completely within 1 or 2 days of commencing high-dose salicylates.

Elevated acute-phase reactants

Almost all patients, except those with pure chorea, have a dramatically elevated erythrocyte sedimentation rate or serum C-reactive protein. There appears little difference between these measurements in their diagnostic usefulness. The C-reactive protein may return to normal more rapidly than the sedimentation rate when rheumatic activity subsides. Mild to moderate peripheral leucocytosis is common, although this is a less sensitive marker of rheumatic inflammation.

Other features

Severe, central abdominal pain is found at presentation in a small proportion of patients. It may be associated with other features of acute rheumatic fever; if not, these features usually appear within 1 or 2 days. The pain responds quickly to anti-inflammatory medication. Epistaxis was reported frequently in historical accounts of acute rheumatic fever, but does not feature prominently in recent descriptions. Pulmonary infiltrates may be found in patients with acute carditis; this has been labelled 'rheumatic pneumonia' although it is not clear whether the infiltrates represent rheumatic inflammation or another process. There may be microscopic haematuria, pyuria, or proteinuria; also mild elevations of liver transaminases: these are non-specific and not usually severe.

Associated post-streptococcal syndromes

Post-streptococcal reactive arthritis has been differentiated from rheumatic fever by some authors because it has a shorter incubation period after streptococcal infection, sometimes follows non-group A b-haemolytic streptococcal infection, may have a different pattern of arthritis (including small joint involvement), and is less responsive to anti-inflammatory medication. Because of the lack of cardiac involvement, these patients are said not to require secondary prophylaxis. However, a few patients who have subsequently developed carditis have led other authors to question the distinction between post-streptococcal reactive arthritis and rheumatic fever. If post-streptococcal reactive arthritis is diagnosed, secondary prophylaxis should be prescribed for at least 1 year and discontinued if there is no evidence of

carditis. In populations with high incidence rates of acute rheumatic fever, it may be prudent to treat all cases of possible post-streptococcal reactive arthritis as acute rheumatic fever.

The frequent finding of emotional lability, motor hyperactivity, and occasional obsessive–compulsive symptoms in patients with Sydenham's chorea led to the observation that group A streptococcal infections may precipitate or exacerbate other disorders of the basal ganglia. These include tic disorders, Tourette syndrome, and obsessive–compulsive disorder, and the term **PANDAS** (paediatric autoimmune neuropsychiatric disorders associated with streptococcal infections) has been coined. Patients with PANDAS appear not to be at risk of developing carditis. There is evidence that these patients, and some children with autism, have high proportions of circulating B cells expressing D8/17 antigen, which is a proposed marker of rheumatic fever susceptibility. It is not yet clear whether these syndromes are linked with acute rheumatic fever.

Diagnosis

Because of the diversity of symptoms and signs, and the non-specific nature of most of them, Dr T. Duckett Jones developed a set of criteria to aid in the diagnosis of acute rheumatic fever in 1944. The Jones criteria have subsequently been revised and updated a number of times to improve their positive and negative predictive values. The most recent version, the 1992 update, is shown in Table 2. The manifestations are divided into: major, those which are most predictive of acute rheumatic fever, and minor, those which are commonly found in acute rheumatic fever but are less specific.

The diagnosis requires the presence of either two major, or one major and two minor criteria, plus the demonstration of a current or recent group A streptococcal infection. Evidence of group A streptococcal infection is not required for chorea, where the onset may be delayed up to 6 months after streptococcal infection, and late-onset carditis, when low-grade inflammation may persist for prolonged periods after the precipitating infection.

The 1992 updated Jones criteria are to be used only for the diagnosis of the initial episode of acute rheumatic fever. Patients with a previous history of acute rheumatic fever or rheumatic heart disease need have only one major or two minor manifestations, have evidence of recent group A streptococcal infection, and have no other plausible explanation for their symptoms. The Jones criteria are a guideline, but cases not fulfilling the criteria should only be diagnosed as acute rheumatic fever once all other possible diagnoses have been excluded.

Proof of a recent group A streptococcal infection can include demonstrating the organism in the upper respiratory tract, either by culture or rapid antigen techniques. However, most children with acute rheumatic fever no longer have a group A streptococcus detectable by these methods, and up to 15 to 25 per cent of normal children in temperate climate countries may carry the organism in their throats . Therefore, serological techniques are most commonly used, particularly the anti-streptolysin O, anti-DNase B, or anti-hyaluronidase titres. One of any two of these tests will be positive in well over 90 per cent of recent streptococcal infections. Their usefulness is increased by performing more than one serological test, or by demonstrating rising titres in paired sera. Serology is of limited value in regions with high prevalence rates of streptococcal impetigo, where children may have positive anti-streptococcal titres most of the time. The diagnosis of acute rheumatic fever in these circumstances can be very difficult. There is therefore a need for a better diagnostic test of recent streptococcal infection, or an objective diagnostic test for acute rheumatic fever itself.

The most common clinical presentation, that of a child with fever and polyarthritis, raises multiple differential diagnoses that will vary by region. Table 3 lists some alternative diagnostic possibilities for the three most common major manifestations.

Treatment

If untreated, acute rheumatic fever lasts on average for 3 months. Except in the case of life-threatening acute carditis, there is no evidence that presently available treatments alter the outcome. Most treatments are designed to provide symptomatic relief or are based on theoretical (but unproven) approaches to attenuating the long-term damage.

All patients with acute rheumatic fever should be admitted to hospital (if practical) to confirm the diagnosis, to perform baseline investigations to ascertain the status of the heart, to provide adequate treatment for the acute phase, to commence secondary prophylaxis, to allow communication of details to personnel responsible for long-term follow-up of the patient, and to begin education of the patient and family. The mainstays of treatment are bed rest, penicillin, and salicylates.

Bed rest

Previous recommendations that children with acute rheumatic fever be rested in bed until all signs of active inflammation abated were probably more extreme than is necessary. Once symptoms of arthritis have subsided and any cardiac failure is controlled, the child may begin gentle mobilization, which may be increased as tolerated. There is no evidence that bed rest beyond the period where mobilization leads to exacerbation of pain or cardiac failure has any long-term benefit.

Penicillin

All patients with acute rheumatic fever should be given penicillin to eradicate the group A streptococcus that precipitated the attack. This is based on an early finding that, in some cases, prolonged group A streptococcal infection led to more severe acute rheumatic fever. Although in most cases the precipitating organism cannot be cultured, a treatment course of penicillin is prudent in case the strain remains present in low numbers, and to prevent its transmission to other contacts. As the aim is eradication of group A streptococcal infection, penicillin may be administered either as a single intramuscular injection of benzathine penicillin G at a dose of 1.2 million units (600 000 U for patients less than 30 kg) into the gluteal or quadriceps muscles, or as a 10-day course of oral phenoxymethyl penicillin (V) at a dose of 500 mg (adolescents and adults) or 250 mg (children) given either two or three times daily. In the case of penicillin allergy, the present recommendation is to use oral erythromycin at 20 to 40 mg/kg per day given two to four times daily for 10 days, although in some regions levels of erythromycin-resistance among group A streptococci are increasing.

Salicylates

Children with arthritis or severe arthralgia should be treated with non-steroidal anti-inflammatory medication; salicylates have been most widely used. Aspirin at a dose of 80 to 100 mg/kg per day (4 to 8 g/day in adults) usually results in defervescence and resolution of arthritis and arthralgia within 1 to 2 days. Sometimes these doses lead to nausea or vomiting; this can be minimized by increasing from lower starting doses. When the diagnosis is uncertain, salicylates should be withheld for a day or two to observe for the development of characteristic migratory polyarthritis. In such cases, codeine can be used to control pain until the diagnosis is confirmed.

There is no evidence that salicylates reduce the severity of acute carditis or the risk of chronic cardiac valve damage. Nevertheless, many clinicians administer salicylates until acute-phase reactants have returned towards normal in the belief that this may reduce the risk of long-term cardiac damage. After 2 weeks, the dose is often reduced to 60 to 70 mg/kg per day for the remaining 2 to 4 weeks. Arthritis or arthralgia may return up to 2 to 3 weeks after discontinuation of therapy; this is usually a brief and mild recrudescence, often associated with increased erythrocyte sedimentation rate or C-reactive protein, and can be managed either with rest and reassurance or a short course of lower-dose anti-inflammatory medication.

Corticosteroids

For many years, corticosteroids have been used in acute rheumatic fever, particularly for patients with severe carditis. As with salicylates, the evidence that they reduce either the severity of acute carditis or the risk of long-term valve damage is conflicting. Many clinicians continue to use them, commonly oral prednisone or prednisolone at a dose of 40 to 60 mg/day, tapering after 2 or 3 weeks.

Treatment of cardiac failure

Although the use of corticosteroids is controversial, there is no doubting the need to treat cardiac failure. Diuretics, angiotensin-converting enzyme inhibitors (especially in aortic regurgitation), and fluid restriction are most commonly employed. Digoxin is usually restricted to cases where atrial fibrillation coexists with cardiac failure, often found in older patients with established mitral stenosis.

If medical therapy fails, cardiac surgery should be considered, even during the acute phase. In populations where fulminant acute carditis is relatively common (e.g. South Africa), mitral valve repair or replacement can be life saving and surgeons have developed techniques for undertaking these procedures despite friable, acutely inflamed valvular and perivalvular tissues. In recent years, there has been a greater tendency to undertake valve repair rather than replacement, or to use homografts or xenografts rather than mechanical prostheses. This is to avoid high rates of thromboembolic complications associated with mechanical prostheses, particularly in populations where compliance with anticoagulation chemotherapy is suboptimal and there are difficulties in monitoring coagulation indices.

Treatment of chorea

Sydenham's chorea always resolves, and if mild there may be no need for specific treatment. However, medications may reduce abnormal movements in moderate or severe chorea. Haloperidol is commonly used as a first-line treatment. Other medications employed include sodium valproate, pimozide, chlorpromazine, or benzodiazepines. Occasionally, low doses of minor tranquillizers are necessary for associated anxiety and emotional lability. All of these medications should be used sparingly and only for defined periods. Salicylates and steroids have no role in treatment of chorea. Psychotherapeutic interventions have little role in the short to medium term, and may increase the stigma of this self-limited organic disease. However, if longer-term behavioural abnormalities persist (e.g. emotional lability, obsessive–compulsive traits), behavioural therapy should be considered.

Newer therapies

Because of the autoimmune nature of acute rheumatic fever, immunomodulatory therapies have been tried. Intravenous immune globulin (IVIG) has been given in some small trials. One study showed no apparent benefit on rate of improvement of clinical, laboratory, or echocardiographic parameters of acute carditis, but another suggested that it may accelerate recovery from chorea. Other therapies have yet to be formally assessed.

Prognosis and follow-up

The most important prognostic factors are the severity of the acute carditis and the number of recurrences. Overall, approximately 30 to 50 per cent of patients with a first episode of acute rheumatic fever will develop chronic rheumatic heart disease. This increases to more than 70 per cent in patients with severe carditis at the first episode, or in those who have had at least one recurrence.

Any patient with acute rheumatic fever requires long-term follow-up. Follow-up assessments should focus on cardiac status, adherence to secondary prophylaxis, early treatment of group A streptococcal pharyngitis, and prevention of streptococcal pyoderma (including hygiene and treatment or prevention of scabies infestation). Patients with evidence of cardiac valve damage should be assessed regularly by specialist physicians and considered for cardiac surgery before substantial left ventricular dysfunction occurs. Vasoactive drugs, particularly angiotensin-converting enzyme inhibitors, may delay the need for operation in asymptomatic patients with chronic aortic regurgitation. Regular echocardiography may be useful to follow the progress of rheumatic heart disease, especially in populations where follow-up may be irregular or in whom communication or cultural differences make clinical assessment difficult.

Recurrences

Approximately 75 per cent of all recurrences occur within 2 years of an episode of acute rheumatic fever. The reasons for this are not known, but are thought to relate to a time-dependent sensitization of the immune response. The clinical features of recurrences tend to mimic those present at the initial episode, particularly in the case of chorea. However, this rule is not absolute, and the risk of developing other manifestations increases with each recurrence. For example, in the Australian Aboriginal population 40 per cent of patients without carditis at the initial episode of acute rheumatic fever developed it at the first recurrence, and 70 per cent developed carditis at either of the first two recurrences. The practical implication of this is that the absence of carditis at the first episode does not help to identify patients who may not need secondary prophylaxis.

Prevention of acute rheumatic fever

Secondary prophylaxis

Every patient with acute rheumatic fever should immediately commence secondary prophylaxis: long-term, regular antibiotics to prevent primary group A streptococcal infections. This strategy is proven to reduce the incidence of recurrences and the risk of developing chronic rheumatic heart disease.

The optimal regimen is 1.2 million units of intramuscular benzathine penicillin G every 3 weeks, and this is commonly given in populations with high incidences of acute rheumatic fever and programmes in place to support the regimen. Higher doses (1.8 or 2.4 million units) given every 4 weeks may have similar effect, but further evidence is needed before such regimens can be recommended routinely. An alternative strategy is to use oral penicillin V at a dose of 250 mg twice daily; this is almost as effective as using benzathine penicillin G, but adherence is usually less reliable.

The most effective strategy for patients proven to be allergic to penicillin is to attempt desensitization using an approved protocol. If this is unsuccessful, the present recommendation is to use oral erythromycin at a dose of 250 mg twice daily. Recent trials have shown newer oral cephalosporins to be effective at eliminating upper respiratory tract carriage of group A streptococci. However, none of these antibiotics have been evaluated for their ability to prevent acute rheumatic fever.

The duration of secondary prophylaxis is dictated by the reducing risk of recurrence with increasing age, with time since the last episode, and the possible consequences of recurrences. Secondary prophylaxis should continue for at least 5 years following the most recent episode or until age 21 years, whichever comes last. However, in patients with substantial valvular disease (e.g. moderate or severe mitral or aortic regurgitation, or any mitral or aortic stenosis), secondary prophylaxis should be continued longer—to age 30 or 35 years in most populations. If the damage is severe, or in patients who have had valve surgery, the possibility that recurrence might be catastrophic mandates that secondary prophylaxis be continued for life.

Primary prophylaxis

A full course of penicillin treatment commencing within 9 days of the onset of symptomatic group A streptococcal pharyngitis will prevent the subsequent development of acute rheumatic fever in most cases. After the diagnosis has been confirmed by a throat culture or rapid antigen diagnostic test, the treatment of choice is penicillin, administered either as a single intramuscular injection of benzathine penicillin G (600 000 U for children who weigh less than 30 kg, or 1.2 million U for larger children and adults) or as a full 10 days of oral (phenoxymethyl) penicillin V (250 mg for children or 500 mg for adults given two to three times daily). The importance of completion of the 10-day course, even if symptoms abate quickly, should be stressed to patients and parents. Shorter courses of oral penicillin treatment are associated with higher risks of acute rheumatic fever. There has never been a clinical isolate of group A streptococcus that is resistant to penicillin; therefore, the use of other antibiotics for primary prophylaxis should be restricted to patients who are allergic to penicillin.

In the case of penicillin allergy, a 10-day course of an oral macrolide such as erythromycin is recommended. First-generation oral cephalosporins also may be considered. However, these agents have not been evaluated in populations with high incidences of acute rheumatic fever. Shorter courses (e.g. 5 days) of some later-generation oral cephalosporins appear to be effective in eradicating carriage, but because of their expense and broader spectrum of antimicrobial activity they should be considered as second-line agents.

It is not possible to predict which episodes of group A streptococcal pharyngitis will precipitate acute rheumatic fever, so this treatment must be offered in all cases to be effective. Unlike prevention of recurrent episodes, which is virtually complete using secondary prophylaxis, penicillin treatment of streptococcal pharyngitis will prevent only the one-third or so of cases of acute rheumatic fever that follow a sore throat. However, this important intervention may arrest the spread of pathogenic group A streptococci in the community. Penicillin treatment of group A streptococcal pharyngitis should begin as early as possible in patients with a history of acute rheumatic fever, should they not be taking secondary prophylaxis, but even then may not prevent a recurrence, hence the need for secondary prophylaxis.

In recent years the use of primary prophylaxis has been questioned in some industrialized countries where acute rheumatic fever is now rare. It is argued that the strategy prevents few cases of acute rheumatic fever but contributes to overuse of antibiotics. Similar arguments were raised in the United States during the 1970s, but faded somewhat with the resurgence of acute rheumatic fever in that country during the 1980s. Any country considering abandoning primary prophylaxis should

first have in place effective surveillance to detect changes in the epidemiology of primary group A streptococcal infections and the appearance of cases of acute rheumatic fever.

Primary prophylaxis is unsuccessful in many developing countries. It requires trained health workers, microbiology laboratories, transportation and communication infrastructure, the availability of penicillin, and a population likely to seek and adhere to treatment for sore throats. In some high-risk populations, all patients with sore throats receive intramuscular benzathine penicillin G without further attempts at diagnosis; the cost-effectiveness of this strategy has not been fully determined. Clinical algorithms to identify patients with group A streptococcal pharyngitis without resorting to laboratory tests have not been validated sufficiently for them to be recommended universally. Even if primary prophylaxis were to be instituted effectively in developing countries, acute rheumatic fever would not disappear, as most cases do not follow a sore throat.

Other methods of primary prevention are clearly needed in developing countries. Improved living standards and access to primary health care appear years or decades away in many places. Although streptococcal skin infections may be linked to acute rheumatic fever pathogenesis, there are no trials of impetigo control programmes to prevent acute rheumatic fever. There is a current focus on attempts to develop a group A streptococcal vaccine. Clinical trials of prospective vaccines are imminent, but the process will take many years, and recent experience suggests that new vaccines are often beyond the financial reach of most developing countries. For the foreseeable future at least, acute rheumatic fever prevention in many developing countries will depend on improving adherence to secondary prophylaxis and developing new strategies for primary prophylaxis.

Further reading

Anonymous (1995). Strategy for controlling rheumatic fever/rheumatic heart disease, with emphasis on primary prevention: memorandum from a joint WHO/ISFC meeting. *Bulletin of the World Health Organization* **73**, 583–7. [Recommendations for prevention in developing countries.]

Bach JF *et al.* (1996). Ten-year educational programme aimed at rheumatic fever in two French Caribbean islands. *Lancet* **347**, 644–8. [Demonstrates dramatic impact of comprehensive public health approach in developing countries.]

Bisno AL (1991). Group A streptococcal infections and acute rheumatic fever. *New England Journal of Medicine* **325**, 783–93. [Concise summary, including pathogenesis.]

Carapetis JR, Currie BJ, Kaplan EL (1998). The epidemiology and prevention of group A streptococcal infections: acute respiratory tract infections, skin infections and their sequelae at the close of the twentieth century. *Clinical Infectious Diseases* **28**, 205–10. [Comparison of epidemiology and public health approaches in industrialized and developing countries.]

Committee on Rheumatic Fever, Endocarditis, and Kawasaki Disease of the Council on Cardiovascular Disease in the Young, the American Heart Association (1995). Treatment of acute streptococcal pharyngitis and prevention of rheumatic fever: a statement for health professionals. *Pediatrics* **96**, 758–64. [Updated recommendations for prophylaxis.]

Kaplan EL (1993). Global assessment of rheumatic fever and rheumatic heart disease at the close of the century. Influences and dynamics of populations and pathogens: a failure to realize prevention? *Circulation* **88**, 1964–72. [Summary of epidemiological aspects, and their contribution to understanding pathogenesis.]

Martin DR *et al*. (1994). Acute rheumatic fever in Auckland, New Zealand: spectrum of associated group A streptococci different from expected. *Pediatric Infectious Diseases Journal* **13**, 264–9. [Important study suggesting a link between skin streptococci and rheumatic fever.]

Quinn RW (1989). Comprehensive review of morbidity and mortality trends for rheumatic fever, streptococcal disease, and scarlet fever: the decline of rheumatic fever. *Reviews of Infectious Diseases* **11**, 928–53. [Exactly as the title suggests.]

Stollerman GH (1975). *Rheumatic fever and streptococcal infection*. Grune & Stratton, New York. [Landmark review of all aspects of rheumatic fever, with comprehensive clinical information.]

Stollerman GH (1997). Rheumatic fever. *Lancet* **349**, 935–42. [Excellent summary of recent advances in rheumatic fever research.]

15.10.2 Infective endocarditis

W. Littler and S. J. Eykyn

Historical background

Lazerous Riverius recorded the first case of what is now known as infective endocarditis in 1723. He described a French magistrate with an irregular pulse, oedema, and congestion, who at autopsy had fleshy masses 'the size of hazelnuts' obstructing the aortic ostia. Fifty years later Morgani (1769) made the link between infection (fulminating gonorrhoea) and 'whitish polypus concretions on the upper part of the aortic valve near its borders'.

The clinical picture of endocarditis was first described by Jean Baptiste Bouillard in 1835: 'fever, an irregular pulse, cardiomegaly (by percussion) and a bellows murmur in the heart'. He gave the disease the name 'endocarditis' or an inflammation of the inner membrane of the heart and fibrous tissues of the valve and was the first to use the term 'vegetations' for the valvular lesions.

Winge used the term 'mycoses endocardi' for the groups of micro-organisms that he saw when he examined vegetations under the microscope in 1870. In 1886 Wyssecokowitch cultured *Staphylococcus aureus* from an endocardial vegetation. Lenthartz in 1901 was the first to use blood cultures in the diagnosis of endocarditis. 'Infective endocarditis' was the term used by Thomas Horder in 1901 to describe the syndrome consisting of (i) the presence of valvular disease, (ii) the occurrence of systemic embolism, and (iii) the discovery of micro-organisms in the bloodstream.

Epidemiology

Infective endocarditis was universally fatal before the advent of antibiotic therapy. Since 1944 deaths from endocarditis have fallen by 80 per cent: about 200 deaths are recorded each year in the United Kingdom, which is probably an underestimate. The true incidence of the condition is unknown, but it is at least 25 cases per 1 000 000 of population. The incidence is greater in men, in those over 65 years of age, and in those with prosthetic heart valves. About 20 per cent of patients with infective endocarditis die and they account for about 0.1 per cent of the total deaths from diseases of the circulatory system (ICD codes 390–429).

Endocarditis does occur in children but is rare, especially in the first decade of life. In the older literature tetralogy of Fallot was the commonest cardiac problem associated with infective endocarditis, but nowadays cardiac surgery is the most likely predisposing cause.

Pathogenesis

Normal vascular endothelium is resistant to microbial infection and very few patients potentially at risk actually develop infective endocarditis. Since low-grade bacteraemia occurs frequently in everyone, a defence mechanism must exist that can eradicate microbes adherent to vegetations. Platelets play a pivotal role in the antimicrobial host defence mechanism and human platelets have been found to contain at least 10 different bactericidal proteins or 'thrombocidins'.

Damage to the endothelial surface of the heart or blood vessels induces platelet and fibrin deposition, producing a sterile thrombotic vegetation; infective endocarditis is initiated by the binding of microbes, discharged into the general circulation from a peripheral site, to these vegetations. These microbes become encased in further depositions of platelets and fibrin and multiply.

The pathogenesis of infective endocarditis involves complex interactions between microbes and the host defence mechanisms, both circulating and at the site of endothelial damage. An essential step is the activation of the clotting system and the formation of a fibrin clot on the endothelial surface. Experimental evidence suggests that the main pathogens in infective endocarditis (streptococci and staphylococci) can bind to endothelial cells and induce functional changes within these cells, causing monocyte adhesion. The combination of damaged endothelial cells, bacteria, and endothelial-bound monocytes results in the induction of tissue factor-dependent procoagulant activity which initiates clot formation. Polymorphonuclear leucocytes that are recruited to the infected endothelial site subsequently may be involved in the disease progression: probably the contents of lysosomes released by the activated leucocytes cause softening and separation of valve tissue leading to its destruction.

In endocarditis the vegetations are found predominately on the left side of the heart (95 per cent). In a large autopsy series of more than 1000 cases reported over 50 years ago the mitral valve was involved in 86 per cent, the aortic in 55 per cent, the tricuspid in 20 per cent, and the pulmonary valve in only 1 per cent. The predominance of left-sided lesions led to the belief that the higher pressures and velocities encountered in the left side of the heart and the proximal aorta must impose a greater mechanical stress on the valves and endocardium, which in turn leads to local damage.

Endocarditis is classically associated with 'jet lesions', where blood flowing from a high pressure area through an orifice to an area of lower pressure produces a high velocity jet. Vegetations are usually found in the lower pressure area, for example on the atrial surface of the mitral valve in mitral regurgitation, or the ventricular surface of the aortic valve in aortic regurgitation. This particular deposition of vegetations has been explained on the basis of the Venturi effect.

Once a vegetation is established it determines the subsequent clinical picture by four basic processes: bacteraemia, local tissue destruction, embolization, and the formation of circulating immune complexes.

Clinical features

Early reports of infective endocarditis described a low-grade febrile illness caused by viridans streptococci from the patient's mouth in those with chronic rheumatic heart disease. Night sweats, anorexia, and weight loss were followed by the development of splinter haemorrhages and Osler nodes, finger clubbing, and splenomegaly. The infection progressed relentlessly with increasing cachexia and the patient died from cardiac failure or a major embolic episode. The term 'subacute bacterial endocarditis' was used to describe this illness. 'Acute or malignant endocarditis' described an aggressive form of the disease usually caused by *S. aureus*, or other virulent bacteria.

During the past 50 years there has been a striking change in the pattern of endocarditis. The dramatic decrease in rheumatic fever in developed countries, the use of antibiotics, and the emergence of antibiotic-resistant organisms, together with surgical advances have all contributed to many clinical variants and modes of presentation.

The proportion of patients in developed countries with endocarditis with no known pre-existing cardiac lesion has risen to over 50 per cent. This change is related to both the decline in rheumatic heart disease and to the increase in extracardiac predisposing factors including intravenous narcotic abuse, haemodialysis, and the use of intravascular devices. Prosthetic heart valves are an important predisposing factor and cardiac surgery for complex congenital lesions has increased the lifespan of patients who would previously have died prematurely. The longevity of the populations in developed countries has resulted in an increasing age of patients with infective endocarditis. The mean age has risen from under 40 years before 1940 to between 60 and 70 years today.

Features of a bacteraemic illness

Discharge of the infecting agent into the circulation produces constant bacteraemia, which may present as pyrexia, rigors, malaise, anorexia, headache, confusion, arthralgia, and anaemia. However, some cases of endocarditis may present without fever, particularly in the elderly.

Features of tissue destruction

Endocarditis initially affects valve cusps, leaflets, or chordae tendineae. Tissue destruction results in valvular incompetence, cusp perforation, or rupture of the chordae producing an appropriate cardiac murmur that may change in character during the course of the illness. Large vegetations rarely obstruct a native valve, but mechanical obstruction of prosthetic valves is more common and clinically more difficult to detect. As the infective process progresses it may extend beyond the valve into the paravalvular structures. This is more common in native aortic valve endocarditis than in mitral valve infection. Aortic root abscess is a serious complication and a destructive lesion. When the abscess extends through the aortic wall into other tissues or cavities a fistula may be formed or pseudo-aneurysms produced. Involvement of the conducting tissue leads to heart block. Infection of a mechanical valve involves the sewing ring and may lead to valve dehiscence. Endocarditis involving an aortic mechanical valve is often localized to the junction between the sewing ring of the aortic valve and the aortic annulus: a large false aneurysm may develop in this area.

Features of systemic or pulmonary emboli

Fragments of an infected vegetation may be dislodged into the general or pulmonary circulation, depending on the site of the vegetation, producing the emboli that are reported in 20 to 40 per cent of cases; a higher incidence (50 per cent) has been reported in autopsy series. Emboli may lodge in any part of the circulation and present as a cerebrovascular accident, arterial occlusion of a limb, myocardial infarction, sudden unilateral blindness, or infarction of the spleen or a kidney. In right-heart endocarditis, recurrent septic pulmonary emboli may be misinterpreted as 'pneumonia'. Mycotic aneurysms arise from embolism of the vasa vasorum weakening the arterial wall: they have been reported in almost 3 per cent of clinical cases, but are found in up to 15 per cent of cases at autopsy. In the cerebral circulation such aneurysms may produce subarachnoid haemorrhage. The popliteal artery is a common site for mycotic aneurysms.

Emboli are characteristic of *S. aureus* infections and large emboli are a feature in HACEK (see below) and fungal endocarditis. Emboli usually occur before or within the first few days after starting antimicrobial therapy. The risk of emboli decreases with time during appropriate antimicrobial treatment. There is no significant difference between mitral valve and aortic valve vegetations with respect to embolization. Vegetation size does not predict systemic embolization, but large vegetations (greater than 10 mm) are associated with a poor outcome overall.

Features of circulating immune complexes

The infected vegetation contains antigens that trigger an immune response. The length of the illness seems to determine the extent of this response; chronic antigenaemia stimulates generalized hypergammaglobulinaemia, so that after several weeks of infection a variety of autoantibodies can be detected. Immune complex deposition may cause many of the extracardiac manifestations of infective endocarditis, but these classic signs are relatively uncommon and are often absent in individual patients.

Splinter haemorrhages

These are found in the nail bed of the fingers, less commonly the toes, and are linear in form (Fig. 1 and Plate 1).

Fig. 1 Splinter haemorrhages in a case of infective endocarditis. (See also Plate 1.)

Osler nodes

These transient painful erythematous nodules are found at the ends of fingers and toes and the thenar and hypothenar eminences. An alternative explanation is that Osler nodes are due to minute infected emboli.

Janeway lesions

These irregular painless erythematous macules are found in roughly the same distribution as Osler nodes. They tend to blanche with pressure.

Vasculitic rash

Immunoglobulin and complement deposits are found in the walls of skin capillaries (Fig. 2 and Plate 2). Vasculitis may account for some of the neurological findings in

infective endocarditis.

Fig. 2 Vasculitic rash on lower limb of a patient with infective endocarditis. (See also Plate 2.)

Roth spots

Boat-shaped haemorrhages in the retina are often called Roth spots, but true Roth spots are white retinal exudates that may be surrounded by haemorrhage. They consist of perivascular collections of lymphocytes.

Splenomegaly

Clinical splenomegaly is now less common than reported in the earlier literature. CT scanning of the abdomen shows the spleen to be enlarged in at least 50 per cent of cases and often demonstrates splenic infarcts.

Nephritis

Immune complexes can cause glomerulonephritis, with immunofluoresence demonstrating deposition of immunoglobulins and complement in irregular granular deposits in the glomerular basement membrane and mesangium. Proteinuria, haematuria, and cellular urinary casts may be present.

Arthralgia

The joint manifestations of infective endocarditis may result from immune complex deposition in the synovial membrane.

Other features

Up to 30 per cent of patients with endocarditis present with neurological symptoms; these are most common in staphylococcal infection, in which one-third present with the clinical features of meningitis. Headaches, confusion, and toxic psychosis can be present as well as encephalomyelitis. Cerebral embolism, which may produce a stroke as a result of cerebral infarction, is more characteristic of viridans streptococcal and enterococcal endocarditis. Mycotic aneurysms may rupture causing subarachnoid or intracerebral bleeding. Septic embolism may result in the formation of a cerebral abscess. It is not certain whether some of these neurological manifestations arise from repeated small emboli or from a vasculitic process within the cerebral circulation resulting from immune complex deposition. The cerebrospinal fluid can show an increase in white cells, but is usually sterile on culture, although very occasionally positive in staphylococcal infection.

Immune-mediated glomerulonephritis has been regarded as the typical lesion of infective endocarditis, but this assumption was based on small series pre-dating modern treatment regimens. More recent work indicates that the commonest renal histological finding is infarction, usually septic. Glomerulonephritis is usually vasculitic. Acute postinfective glomerulonephritis and membranoproliferative glomerulonephritis are less common. Circulatory compromise can cause severe renal impairment as a result of acute tubular necrosis or (very rarely) renal cortical necrosis.

Finger clubbing is one of the classic features of infective endocarditis, usually seen after 1 or 2 months of the illness. It is seldom seen now, but remains a useful sign since it rarely occurs in conditions with which infective endocarditis is confused.

Specific types of endocarditis

Prosthetic valve endocarditis

Patients with prosthetic heart valves have a small but constant risk of infective endocarditis, estimated at 0.2 to 1.4 events per 100 patient years. The incidence of prosthetic valve endocarditis is about 3 per cent in the first postoperative year, with the highest risk during the first 3 months. Prosthetic valve endocarditis is five times more common with aortic than mitral prostheses, and may involve mechanical, xenograft, and homograft valves.

Prosthetic valve endocarditis has been classified as early or late according to its temporal relationship to the time of surgery. Early prosthetic valve endocarditis accounts for 30 per cent of cases and usually occurs within 60 days of open heart surgery. It is caused either by contamination of the prosthetic valve at implantation or by perioperative bacteraemia. The commonest organisms are usually coagulase-negative staphylococci.

Late prosthetic valve endocarditis accounts for 70 per cent of cases and usually occurs 60 days or more after surgery. The pathogens are usually those seen in native valve endocarditis with a preponderance of viridans streptococci and staphylococci, but with a higher incidence of other organisms. Some patients with late prosthetic valve endocarditis will have acquired the infection at the time of surgery, but a bacteraemia is usually the principal cause.

Bacteraemia in a patient with a prosthetic valve must always be taken seriously, but it may not always be the result of endocarditis. The clinical picture of prosthetic valve endocarditis is usually fever, malaise, and weakness, but the more classic signs are usually absent. The condition is often insidious and difficult to diagnose clinically. A new murmur may appear and heart failure and embolic phenomena result in a high mortality (20 to 50 per cent). Infection in a mechanical valve is located in the sewing ring; the infection can spread into the host tissues producing annular abscesses, paravalvular leak, and prosthetic dehiscence. Myocardial abscesses can develop as a consequence of an annular abscess with xenograft or homograft valves. Infection usually involves the valve leaflets, resulting in destruction or perforation and consequent valvular incompetence. The infection involves the valve annulus less commonly than with a mechanical prosthesis. Vegetations may cause obstruction with all forms of prosthetic valve.

The diagnosis of prosthetic valve endocarditis requires a high index of clinical suspicion, blood cultures, and transoesophageal echocardiography. This technique is far superior to the transthoracic approach for finding vegetations and identifying periprosthetic spread of the infection. Vegetations are more difficult to identify in patients with mechanical valves than those with bioprostheses.

Right-sided endocarditis

Right-sided infective endocarditis accounts for only 5 per cent of cases overall, but centres that treat large numbers of intravenous drug users will have a higher incidence. The clinical picture differs significantly from left-sided disease. It is usually associated with intravenous drug addiction or indwelling intravascular devices, and in the former is found particularly in a younger population. *S. aureus* is the commonest pathogen and the tricuspid valve is more commonly affected that the pulmonary. Fever is almost always present and a cardiac murmur is found in 80 per cent of cases. Right-sided endocarditis is associated with septic pulmonary emboli, and the resultant pulmonary infarcts may cavitate. Symptoms include cough, haemoptysis, and pleuritic chest pain, while a chest radiograph shows pulmonary

infiltrates often misdiagnosed as 'patches of pneumonia' (Fig. 3). Renal involvement has been described in over half the cases; most commonly abscess formation or diffuse pyelonephritis. Myocarditis is more common in right-sided involvement than left. Peripheral stigmas of infective endocarditis, splenomegaly, and central nervous system involvement are rare, being described in 5 per cent or less of cases. Death is most commonly due to sepsis, rarely to heart failure.

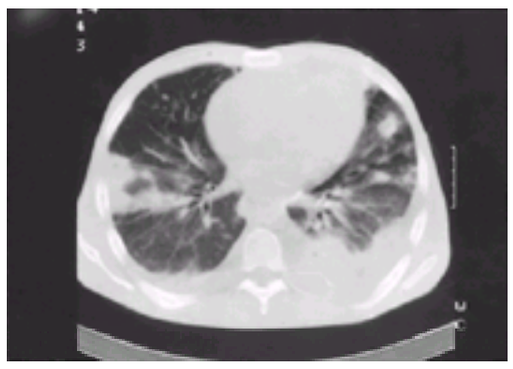

Fig. 3 CT scan of the chest showing multiple pulmonary infarcts in a case of right-sided endocarditis of the tricuspid valve in an intravenous drug user.

Endocarditis in intravenous drug users

Endocarditis is a serious complication of intravenous drug abuse. The right side of the heart is affected most commonly, but the left may also be involved in a substantial number of patients (37 per cent), and both right and left side in a minority (7 per cent). On the right side the tricuspid valve is affected in 80 per cent of cases, while the mitral and aortic valves are equally infected in left-sided disease. A history of previous heart disease is only found in some 25 per cent of cases. *S. aureus* is responsible for 40 per cent of all cases. Gram-negative bacilli are the next most frequent, with *Pseudomonas aeruginosa* and *Serratia marcescens* accounting for the majority of these. Candida can cause endocarditis in intravenous drug users and polymicrobial endocarditis accounts for 5 per cent of cases.

The skin is the commonest site from which pathogens enter the bloodstream via needles. Gram-negative bacilli are rarely recovered from needles or the drug itself and it has been suggested that these organisms come from tap water, sinks, or lavatory pans.

The clinical picture of drug-associated endocarditis depends on which side of the heart is affected. Right-sided disease is associated with fever, a murmur of tricuspid incompetence, and pulmonary infiltrates on the chest radiography. Left-sided disease behaves like that seen in cases not associated with intravenous drugs, with a high incidence of heart failure, arterial embolism, central nervous system involvement, and peripheral stigmas.

The overall mortality depends on when the patient presents: it is high if they present late and reflects, among other things, the difficulty in dealing with addicts because of their poor compliance and reluctance to discontinue their drug habit. The principles of management are similar to those for patients who do not abuse drugs. The duration of intravenous antibiotics should be at least 4 weeks, but it is usually impossible to do this in practice; while in right-sided endocarditis simple removal of the valve without replacement appears to be the best strategy.

The diagnosis of infective endocarditis

Laboratory diagnosis

Blood culture

This is the most important laboratory investigation in the diagnosis of endocarditis. Isolation of the pathogen enables an effective antibiotic treatment regimen to be devised. Blood cultures should be taken before antibiotics are given; if they have already been given, cultures should still be done, and if possible the giving of further antibiotics delayed for a few days. However, previous antibiotics may render the blood sterile for some time and the chances of recovering the pathogen, particularly when it is a viridans streptococcus, are very low. Much mystique has been attached to the number and timing of blood cultures in cases of suspected endocarditis. What is known is that the bacteraemia is usually constant and that whenever the blood is obtained for culture, and however many sets are taken, in most cases all bottles will grow the pathogen. There are of course rare exceptions when only a small proportion of bottles cultured are positive, and this is one reason why it is conventional to take two or three sets. Another reason for several cultures is to assess the relevance of the common skin contaminants, particularly the coagulase-negative staphylococci but also *Corynebacterium* spp., which can cause endocarditis.

In most laboratories blood culture systems are automated, with continuous monitoring which flags up growth for further investigation. Most cultures become positive within 48 h and after this the chances of isolating the pathogen recede, with the exception of fastidious organisms of the HACEK group (see below) that may take much longer to recover from the blood. In most laboratories blood cultures are incubated for 5 to 7 days, but this may not be long enough for the rare fastidious slow grower. The onus is on the clinical microbiologist or clinician to request prolonged incubation of blood cultures from patients in whom endocarditis is strongly suspected on clinical grounds and echocardiography who have not had previous antibiotics and whose blood cultures are sterile after a week's incubation.

Blood tests

In infective endocarditis an elevated erythrocyte sedimentation rate and C-reactive protein are almost invariable and these inflammatory markers are used most commonly to monitor the activity of the disease. A normochromic normocytic anaemia is often present and a polymorphonuclear leucocytosis is found in the majority of cases. Hypergammaglobulinaemia and a low serum complement may be present, together with a false-positive rheumatoid factor. Circulating immune complexes may be detected.

Dipstick testing of the urine may reveal the presence of proteinuria or haematuria, indicating renal involvement. When haematuria is present the pellet of a centrifuged specimen of urine should be resuspended and examined microscopically for the presence of red cell casts, which clinch the diagnosis of glomerulonephritis in this context.

Serology

Serum antibodies are used to diagnose *Coxiella burnetii* (Q fever), bartonella, and chlamydia endocarditis and should be done in any patient with convincing evidence of endocarditis and negative blood cultures. Candida antibodies are of no diagnostic value.

Echocardiography

In suspected cases of endocarditis echocardiography should be performed as soon as possible and interpreted by an experienced cardiologist. Its principal role is to detect vegetations. Echocardiography is not sufficiently sensitive to allow the clinician to exclude the diagnosis confidently on the basis of a negative result. The sensitivity depends on the size of the vegetations and the time course of the disease. Echocardiography can resolve vegetations as small as 1 to 2 mm, but it is more difficult with prosthetic than native valves, and more difficult with mechanical than biological prostheses.

Vegetations appear as thick, ragged, non-uniform echoes oscillating on or around a cardiac valve or in the path of a regurgitant jet. They do not usually restrict leaflet mobility and exhibit valve-dependent motion. On native valves vegetations are usually attached to the ventricular side of the aortic valve and the atrial side of the mitral and tricuspid valves.

Two-dimensional echocardiography should be employed initially in all cases of suspected endocarditis (Fig. 4). Transoesophageal echocardiography has improved

the rate of diagnosis of infective endocarditis over that of transthoracic echocardiography, particularly in the presence of a prosthetic valve. Transoesophageal echocardiography has made it easier to recognize many complications of prosthetic valve endocarditis, such as abscesses, fistulas, and paravalvular leak (Fig. 5). In addition to vegetations, echocardiography may demonstrate indirect signs of valvular integrity, such as excessive systolic expansion of the left atrium in mitral incompetence or fluttering of the anterior leaflet of the mitral valve in aortic incompetence. Ventricular size and contractility are both easily assessed. The diagnosis of right-sided endocarditis has been greatly facilitated by echocardiography, particularly transoesophageal echocardiography. Vegetations, which in general tend to be larger on the right side, can be demonstrated in 80 to 100 per cent of cases.

Fig. 4 Two-dimensional echocardiogram showing a large vegetation involving the posterior leaflet of the mitral valve and prolapsing into the left ventricle. LA, left atrium; LV, left ventricle.

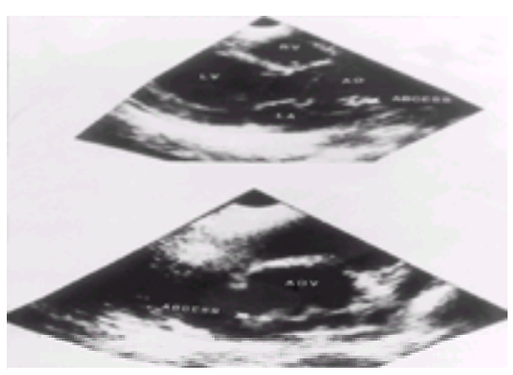

Fig. 5 Transoesophageal echocardiogram showing a large abscess communicating with aortic root. RV, right ventricle; LV, left ventricle; LA, left atrium; AO, aorta; AOV, aortic valve.

Vegetations need to be differentiated from other conditions that produce echo density on cardiac valves, including calcification, myxomatous degeneration, and atrial myxoma. Echocardiography does not provide direct information on blood flow, but Doppler echocardiography complements the technique and adds significantly to the diagnosis of valvular function. It is able to diagnose valvular regurgitation with great accuracy.

Criteria for the diagnosis of infective endocarditis

In 1994 Durack and his colleagues introduced criteria for the diagnosis of infective endocarditis that have been accepted as 'the Duke Criteria' (Table 1). These include two major criteria (typical blood culture and positive echocardiogram) and six minor criteria (predisposition, fever, vascular phenomena, immunological phenomena, suggestive echocardiogram, and suggestive microbiological findings). Application of these criteria is used to define three diagnostic categories: definite, possible, or rejected cases of infective endocarditis.

Modifications of the Duke Criteria to increase their sensitivity have been suggested by others. These include the following additional minor criteria: the presence of newly diagnosed clubbing, splenomegaly, splinter haemorrhages and petechias, a high erythrocyte sedimentation rate or a high C- reactive protein, and the presence of central non-feeding lines, peripheral lines, and microscopic haematuria.

Microbiology

While almost any micro-organism can cause infective endocarditis, particularly when this involves a prosthetic valve, certain species do so much more commonly than others and the predominant species involved in the infection have not changed significantly in their incidence in the past three decades. Overall, viridans streptococci and staphylococci account for about two-thirds of all cases. However, endocarditis cannot be considered as a microbiologically homogeneous entity as the incidence of any specific organism depends: (i) on the patient—whether an intravenous drug user or not; (ii) on the valve—whether native or prosthetic; if native whether previously abnormal or not, and if prosthetic whether mechanical or a bioprosthesis, and whether the infection was acquired early or late; and (iii) where (and how) the infection was acquired—whether in the community or (and increasingly these days) in hospital, usually via an infected intravascular device. The more common species encountered will be considered individually.

Streptococci

The genus *Streptococcus* includes species of differing virulence and pathogenicity as well as differing normal habitat in man. It has undergone numerous taxonomic revisions over the past decade or more and the previous dependence on haemolytic activity on blood agar and serological reactions has been superseded in many cases by molecular and chemotaxonomic approaches. Examples of such taxonomic change include the assignment of the faecal streptococci to the genus *Enterococcus*, of *Streptococcus morbillorum* to *Gemella morbillorum*, and of the nutritionally dependent streptococci previously known as *Streptococcus adjacens* and *Streptococcus defectivus* to the genus *Abiotrophia.* There are many other examples, but taxonomic change is of limited interest to clinicians and has no bearing on the management of infection.

Viridans streptococci

For many years it has been conventional to refer to a group of streptococci that produce greening (a-haemolysis) on blood agar as viridans streptococci, indeed many still refer (inaccurately) to a microbe 'Streptococcus viridans'. While most of these streptococci are virtually specific to the normal oropharyngeal flora and are rarely encountered at other sites, some are not found in the oropharynx at all, for example *S. bovis*, and others are found at many sites including the oropharynx, for example the milleri group of streptococci. The viridans streptococci are the commonest cause of community-acquired native valve endocarditis and community-acquired late-onset prosthetic endocarditis. The commonest species of the viridans streptococci that are specific to the oropharynx are *S. sanguis*, *S. oralis*, and *S. mutans*, but there are others. Dextran formation may be a virulence factor in these streptococci. Contrary to popular belief they do not require a dental extraction to enter the bloodstream and cause frequent bacteraemias after chewing, tooth brushing, and so on. They are organisms of low virulence and thus usually only infect previously abnormal heart valves. Whereas *S. oralis* and *S. sanguis* are occasionally isolated from blood cultures of patients who do not have endocarditis, the isolation of *S. mutans* from the blood is virtually synonymous with endocarditis.

Streptococcus bovis

This streptococcus, which may appear 'viridans' on blood agar, is part of the normal intestinal flora but may initially be mistaken for an oral streptococcus. In common

with the enterococci it bears the Lancefield group D antigen and thus can also be mistaken for *Enterococcus faecalis*, though it is sensitive to penicillin whereas the latter is resistant. There is a significant association between *S. bovis* bacteraemia (and hence endocarditis) and colonic pathology, and any patient with *S. bovis* endocarditis thus warrants ppropriate investigation for this. *S. bovis* endocarditis is much less common than that caused by oral streptococci.

Pyogenic streptococci

These organisms, often referred to as b-haemolytic streptococci, cause endocarditis less frequently than the viridans streptococci, but are more aggressive microbes and are likely to affect (and often rapidly destroy) a previously normal valve. The commonest pyogenic streptococcus to cause endocarditis is the Lancefield group B b-haemolytic streptococcus (**GBS**) sometimes referred to as *S. agalactiae*. This organism is found as normal flora in the genital and gastrointestinal tracts. As with *S. aureus*, any patient with community-acquired GBS bacteraemia should be assumed to have infection in bone, joint, or on a heart valve until proved otherwise. Groups C and G b-haemolytic streptococci occasionally cause endocarditis and group A even more rarely. The milleri group of streptococci are best regarded as pyogenic streptococci. These streptococci form part of the normal flora of all mucous membranes and occasionally cause endocarditis, though much more often abscesses at many different sites. The milleri group consists of three species, *S. constellatus*, *S. intermedius*, and *S. anginosus*. Interestingly these streptococci can bear the Lancefield antigens A, C, G, or F (or none); all group F streptococci are milleri but not all milleri are group F.

Streptococcus pneumoniae (pneumococcus)

Pneumococcal endocarditis accounted for about 10 per cent of cases of endocarditis in the preantibiotic era, but is now rarely seen, although it is sometimes diagnosed at autopsy of patients with fatal pneumococcal infection. The pneumococcus is a virulent pathogen and attacks normal heart valves. Patients with endocarditis generally have pneumonia and sometimes meningitis, the organism originating in the upper respiratory tract.

Enterococci

Enterococci form part of the normal gastrointestinal flora. They are more virulent than viridans streptococci and more resistant to antibiotics. The past decade has seen an increase in enterococcal endocarditis, particularly in the elderly, but this infection is still much less common than that caused by viridans streptococci. While there are many species of enterococci, those causing endocarditis are usually *E. faecalis* and occasionally *E. faecium*. Most cases are community acquired but the infection can be acquired in hospital, sometimes as a result of urological instrumentation. Any patient admitted from the community with *E. faecalis* in the blood should be investigated for endocarditis.

Staphylococci

Staphylococci now account for about a third of cases of community-acquired endocarditis and are the commonest cause of hospital-acquired endocarditis. Most of these staphylococci are *S. aureus*, but an increasing proportion is now due to coagulase-negative staphylococci. All staphylococci are skin organisms and patients become infected from their own skin flora, or in the case of methicillin-resistant *S. aureus* (**MRSA**) from that of others by cross-infection.

Staphylococcus aureus

S. aureus is an important and aggressive pathogen in community-acquired native valve endocarditis. Sometimes a trivial skin lesion can be identified as the source of the organism, but there is often no obvious lesion. *S. aureus*, and increasingly now MRSA, is the commonest cause of hospital-acquired endocarditis. Prosthetic valves can become infected with *S. aureus* both early as result of sternal wound sepsis and late as with native valves. *S. aureus* is the commonest pathogen causing endocarditis in intravenous drug users.

Coagulase-negative staphylococci

Although still regarded by many as pathogens of prosthetic rather than native valves, coagulase-negative staphylococci also cause native valve infection and this has become more common, or certainly more commonly recognized, in the last two decades. The infecting species is most often *S. epidermidis* (*sensu stricto*) but in many reports the designation *S. epidermidis* tends to be used for any unspeciated coagulase-negative staphylococcus. Many other species have been reported in native valve endocarditis including *S. lugdunensis*, *S. simulans*, *S. warneri*, *S. capitis*, *S. caprae*, and *S. sciuri*. Coagulase-negative staphylococci are normal skin flora and different species vary in their distribution throughout the body. As in community-acquired *S. aureus* endocarditis, there is sometimes a presumptive predisposing skin lesion. Most patients have a pre-existing cardiac abnormality. Many of these staphylococci can be as virulent as *S. aureus* and actually share some of the same virulence factors.

Other organisms

A wide variety of organisms account for the small percentage of cases of endocarditis that are not caused by streptococci, staphylococci, or enterococci. Only a few warrant a specific mention here.

HACEK group

These are fastidious slow-growing species that are oropharyngeal commensals and have a predilection for heart valves, such that their presence in blood cultures is virtually synonymous with this infection. The group consists of *Haemophilus aphrophilus/paraphrophilus*, *Actinobacillus actinomycetemcomitans*, *Cardiobacterium hominis*, *Eikenella corrodens*, and *Kingella kingae*. *A. actinomycetmecomitans* in particular seems more likely to infect prosthetic than native valves. The large vegetations thought to be characteristic of HACEK organisms in native valve infection may be the result of diagnostic delay and prolonged illness rather than any inherent property of the microbes *per se*.

Organisms that cannot be cultured by routine techniques

Endocarditis is a rare (and late) sequel of acute *Coxiella burnetii* (Q fever) infection. Most infections occur in middle-aged men with pre-existing valve disease. The reservoir of the organism is usually sheep or cattle, but the source and mode of transmission in many human cases is unknown. The diagnosis is usually made serologically, although *C. burnetii* can be recovered from the blood and excised valves by special techniques. The disease is almost certainly underdiagnosed and some cases are labelled 'culture negative' endocarditis.

Bartonella quintana endocarditis was first recognized in 1995 in homeless alcoholic patients; *Bartonella henselae* infection may be associated with cat or cat flea contact and other species of bartonella have also been described causing endocarditis. Bartonella infection is usually diagnosed by serology, although these bacteria can also be recovered from the blood and excised valves by special culture techniques and their presence detected by polymerase chain reaction (PCR). False-positive serology for *Chlamydia* spp. has been reported with bartonella infections, but *Chlamydia* spp.—and particularly *C. psittaci*—can also cause endocarditis (very rarely); it is possible that some cases attributed to chlamydia in the past on the basis of serology may have been caused by bartonella.

Fungi

Fungal endocarditis is very rare and more likely to occur on prosthetic than native valves, except in intravenous drug users. Most infections are acquired in hospital, when infection at intravascular access sites and broad-spectrum antibiotics predispose to candida infections. *Candida* spp., usually *C. albicans*, are the commonest fungi, but *Aspergillus* spp. and more exotic genera have also been reported. Blood cultures are only likely to be positive with *Candida* spp., and then often only intermittently; for other fungi the diagnosis must be made by serology and culture of the fungus from the excised valve or detection on valve histology.

Blood culture negative endocarditis

The possibility that the illness is not endocarditis should always be entertained when blood cultures are repeatedly negative. However, in 5 to 10 per cent of definite cases of endocarditis the blood cultures will be negative. The commonest explanation for this is previous antibiotics. In a few cases the pathogen will be recovered from another site, including the excised valve, excised emboli, or specifically in right-sided endocarditis, respiratory specimens. Other causes of negative blood

cultures are infection with organisms that cannot be grown by conventional blood culture methods and infections that are diagnosed by serology such as *C. burnetii*, *Bartonella* spp., and *Chlamydia* spp.

Treatment

Initial therapy

In those patients who have been chronically unwell for many weeks, antibiotic treatment can be deferred until the blood cultures are positive and the pathogen known. In patients who are acutely ill, antibiotic treatment should be started after taking blood cultures, using a broad-spectrum combination that can be adjusted when the pathogen is known. However, in many who are acutely ill with native valve infection, endocarditis is often not suspected initially—there may be no obvious signs of this—and the antibiotics are started for 'septicaemia'. There are many possible combinations for acutely ill patients, but intravenous vancomycin and gentamicin will encompass most possible pathogens. When methicillin-resistant staphylococci (whether *S. aureus* or coagulase-negative staphylococci) are likely pathogens, vancomycin or teicoplanin are an essential component of any combination.

Definitive therapy

There are various national guidelines for the treatment of specific organisms. It is important to realize that very few are based on clinical trials that show efficacy of any particular regimen. It is possible to conduct such trials in endocarditis caused by viridans streptococci, but well nigh impossible in cases caused by virulent organisms such as staphylococci as the patients are seldom comparable, with many needing surgery after varying periods of antibiotic treatment. It is conventional to estimate the minimum inhibitory concentration (**MIC**) of the antibiotic for the pathogen, though in practice routine disc sensitivity tests are quite satisfactory in many cases. Although it is widely believed that prosthetic endocarditis requires a longer duration of antibiotic treatment than native valve infection, there are few data to support this. Recommendations for the commonest causative organisms will be given.

Penicillin-sensitive streptococci (MIC < 0.1 mg/l)

It was shown 30 years ago that native valve endocarditis caused by sensitive streptococci could be treated effectively with 2 weeks of intravenous penicillin and an aminoglycoside (originally streptomycin but now gentamicin). The purpose of the aminoglycoside is to achieve synergy, so a full therapeutic dose is not given. This regimen is seldom used in practice in the United Kingdom. Patients allergic to penicillins should be given vancomycin or teicoplanin and gentamicin.

Streptococci with reduced sensitivity to penicillin (MIC > 0.1 mg/l)

The regimens given above should be continued for 4 weeks

Enterococci

Enterococci are rather more sensitive to amoxicillin and ampicillin than penicillin and thus these agents are recommended rather than penicillin. Many enterococci are still relatively sensitive to gentamicin and this drug is given with amoxicillin (for synergy, not a full therapeutic dose) for 4 weeks. Patients allergic to penicillin should be given vancomycin and gentamicin. Some enterococci are now resistant to high levels of gentamicin and for such strains gentamicin should not be given. Some gentamicin-resistant strains are sensitive to streptomycin and, if so, this can be used instead of gentamicin. If not, high-dose amoxicillin should be given for 4 to 6 weeks. Unfortunately some enterococci (usually *E. faecium*) are resistant to amoxicillin, gentamicin, and vancomycin and for them expert help should be obtained.

Staphylococci

The same antibiotic regimens should be used whether the staphylococcus is *S. aureus* or a coagulase-negative strain—it is the antibiotic sensitivity that matters not the infecting species. Strains that are sensitive to penicillin should be treated with this, those that are penicillin resistant but methicillin sensitive should be treated with flucloxacillin, and methicillin-resistant strains with vancomycin. There is no evidence that the addition of gentamicin (for gentamicin-sensitive strains) to the b-lactam antibiotic improves cure rates, but it may result in more rapid defervescence and clearance of bacteraemia.

Practical treatment recommendations are shown in Table 2 and Table 3.

Monitoring of treatment

Serum bactericidal titres against the infecting organism are no longer recommended. There was always great variation in the monitoring methods used for these tests and in the interpretation of their results. At best they could only predict bacteriological not clinical cure and bacteriological failure is very rare. The most useful laboratory test for monitoring the response to treatment (which is usually obvious clinically) is serial C-reactive protein estimation. This is of much more use than the erythrocyte sedimentation rate, which is much slower to fall.

Prevention and prophylaxis

While antibiotic prophylaxis in 'at-risk patients' is accepted as reasonable, there are many uncertainties about its value and data confirming its effectiveness are lacking. The rationale for antibiotic prophylaxis depends on indirect data from *in vitro* studies, experimental animal models, and clinical bacteraemia studies. Despite this uncertainty, all authorities continue to recommend antibiotic prophylaxis to cover certain procedures associated with a predictable and significant bacteraemia in patients known to be at risk, but accept that prophylaxis may fail, even with the recommended regimens, and that adverse reactions to the antibiotics are important even if relatively uncommon.

An international consensus group has recently undertaken a comparative analysis of the published national guidelines, which in the main are quite similar, though the antibiotic regimen for a given procedure may vary according to the perceived cardiac risk. Controversial areas include fibreoptic bronchoscopy, colonoscopy, vaginal hysterectomy, and vaginal delivery. Based on their analysis the consensus group have proposed universal guidelines for cardiac (Table 4 and Table 5) and procedural (Table 6) risks. Prophylactic regimen are shown in Table 7.

Surgical treatment of infective endocarditis

Surgery will be required in about 30 per cent of cases during the acute phase (first 4 months) of endocarditis and 20 to 40 per cent of cases thereafter (Fig. 6 and Plate 3). Since surgery may be required at any time during an episode of endocarditis, it is essential to involve a cardiac surgeon in the overall management from the outset: in practical terms this means transferring the patient to a centre with cardiac surgery wherever possible. Even so, surgery for endocarditis carries a risk of 10 to 25 per cent mortality, and up to 25 per cent of patients develop a paravalvular leak requiring a further operation. The main predictive factors for mortality associated with surgery are prosthetic valve endocarditis, infections due to staphylococci or candida, perioperative shock, or late referral. The timing of surgery is all important and demands experience and clinical judgement, which is best achieved by a team approach with cardiologists, cardiac surgeons, and microbiologists.

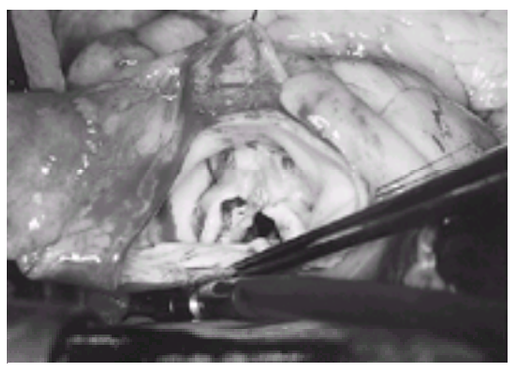

Fig. 6 A large vegetation on the aortic valve of a patient with infective endocarditis as seen at the time of surgery. (See also Plate 3.)

The main indications for surgery are haemodynamic instability and persistent infection. In such cases surgery should never be delayed, even if only hours or days of antibiotic treatment have been given. The primary goals of the surgeon are to remove all infected material and to reconstruct the heart and/or restore valvular function at the lowest operative risk. An understanding of the surgical anatomy of infective endocarditis is a precondition for surgical success, which means the involvement of an experienced surgical team. Wherever possible surgeons now strive to preserve the native valve, either by removal of the vegetation(s) or valve repair. In prosthetic valve endocarditis removal of all foreign material is mandatory.

There are two unresolved issues with regard to the surgical treatment of endocarditis. The first concerns the timing of surgery in patients who have had a cerebrovascular accident either as a result of an embolic stroke or from haemorrhage due to a ruptured mycotic aneurysm. As a general rule, if haemorrhage is detected by CT scanning, delay of at least 1 week is suggested; if there is no haemorrhage, surgery can be undertaken within 72 h.

The second issue concerns the duration of antibiotic treatment postoperatively. If the excised valve is sterile it is doubtful whether further antibiotics are of any benefit. If the pathogen is isolated from the excised valve, antibiotics should be given for a further 2 weeks. If debridement is incomplete, whatever antibiotics are given may fail.

Further reading

Amoury RA, Bowman EO, Malm JR (1966). Endocarditis associated with intracardiac prostheses. Diagnostic management and prophylaxis. *Journal of Thoracic and Cardiovascular Surgery* **51**, 36–48. [One of the earliest papers setting out the problems of endocarditis associated with prosthetic heart valves.]

Baine RJI *et al*. (1988). Impact of a policy of collaborative management on mortality and morbidity from infective endocarditis. *International Journal of Cardiology* **19**, 47–54. [This paper demonstrates the benefit of a 'team approach' to the management of infective endocarditis.]

Birmingham GD, Rahko PS, Ballantyne R (1992). Improved detection in infective endocarditis with transoesophageal echocardiogram. *American Heart Journal* **123**, 774–821. [This paper describes the benefits of using the transoesophageal approach to echocardiography in the diagnosis of infective vegetations.]

Cohen PS, Maguire JH, Weinstein L (1980). Infective endocarditis caused by Gram-negative bacteria: a review of the literature 1945–1977. *Progress in Cardiovascular Diseases* **22**, 205–41. [Even after 20 years this is still one of the best reviews of this particular aspect of endocarditis.]

Durak DT, Lukes AS, Bright DK (1994). New criteria for diagnosis of infective endocarditis: utilisation of specific echocardiographic findings. *American Journal of Medicine* **96**, 200–9. [This paper describes the application of criteria to increase the number of definite diagnoses of infective endocarditis.]

Durak DT (1995). Prevention of infective endocarditis. *New England Journal of Medicine* **332**, 38–44. [An excellent review of prophylaxis against endocarditis.]

Gutschik E and The Endocarditis Working Group of the International Society of Chemotherapy (1998). Microbiological recommendations for the diagnosis and follow-up of infective endocarditis. *Clinical Microbiology and Infection* **4**, 3S10–3S16. [A comprehensive review of investigations currently available for the diagnosis of infective endocarditis.]

Hoen B *et al*.(1995). Infective endocarditis in patients with negative blood cultures: analysis of 88 cases from a one year nationwide survey in France. *Clinical Infectious Diseases* **20**, 501–6. [An excellent review of the problems involved in culture-negative endocarditis. How the problem might be tackled.]

Report of a Working Party of the British Society of Antimicrobial Chemotherapy (1998). Antibiotic treatment of streptococcal, enterococcal and staphylococcal endocarditis. *Heart* **79**, 207–10. [Recommendations for the treatment of the common causes of infective endocarditis in the United Kingdom.]

15.10.3 Cardiovascular syphilis

B. Gribbin and I. Byren

Introduction

Cardiovascular syphilis is a feature of tertiary syphilis and no longer a prominent cause of heart disease: even in specialized cardiac units it is now a rarity. Left untreated, about 12 per cent of patients with syphilis will eventually develop cardiovascular complications. Although gummas can occur in the pericardium, myocardium, and endocardium, and have been the cause of Stokes–Adams attacks when present in the atrioventricular node or the bundle of His, the characteristic lesion is an aortitis. This follows spirochaetal infection of the aortic wall, leading to an endarteritis and periarteritis of the aortic vasa vasorum, initially in the adventitia and subsequently in the media. Lymphocytes and plasma cells surround these small feeding vessels and obliterative changes result in the loss of medial smooth muscle and elastic fibres, occasionally with frank necrosis and eventually with fibrous tissue replacement. This causes scarring of the aortic wall and weakening of its structure. Macroscopically the intima becomes thickened in a gelatinous patchy fashion and fibrosis produces an irregular linear thickening that has been termed the tree-bark appearance. Intimal scarring may involve the ostia of the coronary arteries, which are susceptible to further narrowing by accelerated and superimposed atheroma.

The ascending aorta is involved in about half of all cases, the arch is next in frequency, and the descending aorta in only 10 per cent, with changes virtually limited to that part of the vessel lying above the renal arteries. As the aortic wall structure weakens, so dilatation occurs, resulting in aneurysm formation, which in turn leads to further dilatation and the risk of rupture. The major branches of the aorta may also be affected, especially the innominate artery.

Enlargement of the aortic root and separation of the cusp commissures causes aortic reflux, and although thickening and retraction of the leading edges of the cusps also occurs, this is thought to be a secondary change due to abnormal turbulence rather than a consequence of direct syphilitic involvement of cusp tissues.

Clinical features

Because cardiovascular syphilis can take up to 40 years after primary infection to become apparent, most patients are middle aged or elderly, although tertiary syphilis has been known to occur within a year or two of primary infection. Men are more often affected. Patients with aortitis can present in four main ways: asymptomatic aortitis (most common), aneurysm formation (10 per cent), aortic reflux (25 per cent), and lastly as the result of coronary artery ostial stenosis (25 per cent). The last three are not mutually exclusive and aortic reflux plus ostial stenosis may coexist with aneurysm formation. Gummatous disease of the myocardium is extremely rare.

Aortitis in asymptomatic patients is usually diagnosed as the result of radiographic findings of a dilated ascending aorta with calcification in the wall (Fig. 1). Although aortic calcification is common, particularly in the elderly and hypertensive population, it is then virtually limited to the aortic knuckle and descending aorta. More generalized thoracic aortic calcification is sometimes seen in patients with widespread atheromatous disease, but when visible in the ascending aorta, particularly in a linear fashion, syphilitic aortitis should come to mind and supporting evidence, such as mild aortic reflux, be sought. Serology is likely to be positive.

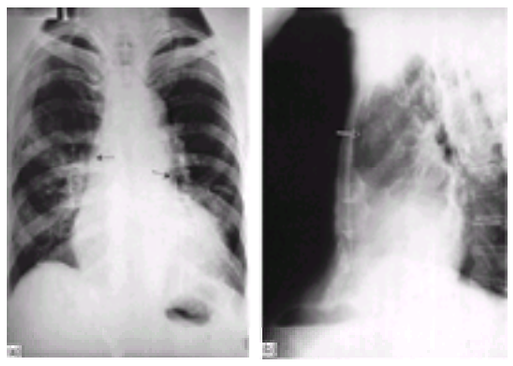

Fig. 1 Posteroanterior and lateral radiographs showing evidence of syphilitic aortitis. A line of calcification (arrowed) is visible in the wall of the dilated ascending aorta.

Syphilitic aortic aneurysms tend to be saccular rather than fusiform and occur most commonly in the ascending aorta, also in the arch, and with increasing rarity down the descending aorta. This is in contrast to atherosclerotic aneurysms, which tend to involve the distal aorta below the renal arteries. The clinical features vary depending on the site of the aneurysm, its size, and whether or not compression and even erosion of adjacent structures occurs. A large aneurysm may cause no symptoms, but pain can be a prominent feature, often sustained and boring in nature, influenced by position, and exacerbated by impending rupture. Pain due to ascending aortic aneurysms is felt in the upper chest wall to the right of the sternum, and with large aneurysms a bulge may appear at this site and erosion of ribs and even sternum may be apparent on radiographs. Aneurysms of the arch produce pain over the upper sternum and occasionally in the throat and there may be visible arterial pulsation in the root of the neck with tracheal deviation. Pressure on upper mediastinal structures can produce superior vena caval obstruction, dysphagia, stridor, and a tracheal tug in time with the pulse. Involvement of the upper descending aorta causes pain between the scapulae or to the left of the spine, and there is a risk of hoarseness from pressure on the left recurrent laryngeal nerve and complications arising from compression of the left main bronchus. Rupture can occur into the bronchus, into the left pulmonary artery, or the left pleural space, and erosion of vertebrae may result in chronic and debilitating pain.

Syphilitic aortic reflux may have a number of features to help distinguish it from the more usual varieties (see Chapter 15.7). Radiographic or clinical evidence of aneurysmal dilatation of the ascending aorta is one, and explains the fact that the early diastolic murmur may be heard better at the right, rather than the more usual left, sternal edge position. Furthermore, an ejection click may be audible, probably as a result of sudden distention of the dilated aortic root by the large stroke volume. However, these auscultatory signs are not entirely specific and may be found in patients with annuloaortic ectasia, now a more common condition, and characterized by a flask-like dilatation of the proximal ascending aorta. Even severe aortic reflux may be tolerated well for years, but eventually the volume overload of the left ventricle leads to cardiac failure: this carries a poor prognosis without, and sometimes despite, surgical intervention.

Coronary ostial stenosis is not restricted to cases of syphilitic aortitis and may occur as a variant of the more usual atheromatous coronary artery disease. It presents as angina, the true cause of which may be missed unless a thin line of calcification is noted in the ascending aorta, or there is other evidence of syphilitic disease in the cardiovascular system or elsewhere. Myocardial infarction may occur as a complication.

Diagnosis

The diagnosis of cardiovascular syphilis is usually made by detecting a positive serum antibody test in a patient who may give a history of past syphilitic infection, and who has evidence of aortitis or one of its complications. It is important to note that 10 to 25 per cent of patients with cardiovascular syphilis also have central nervous system involvement.

Non-specific antibody tests such as the Venereal Diseases Research Laboratory (**VDRL**) or rapid plasma reagin (**RPR**) may be negative in cardiovascular syphilis. If positive in high titre, they may indicate active untreated disease and can be used to gauge response to treatment. However, they have drawbacks: false positive results are not uncommon, and they are invariably positive in patients with endemic non-venereal treponematosis. Specific tests such as the fluorescent treponemal antibody absorption test (FTA-ABS) and *Treponema pallidum* haemagglutination test (TPHA) are almost always positive. They remain positive despite treatment and therefore cannot be used to assess response to treatment. Sexual and vertical transmission does not occur with cardiovascular syphilis, but nevertheless it is prudent to consider testing of known sexual partners and the children of women with cardiovascular syphilis. For more detailed information see Chapter 7.11.33.

Treatment

It is generally accepted that antibiotic treatment is indicated for patients with cardiovascular syphilis who have not received effective antibiotic treatment in the past. There is, however, no evidence that this reduces the severity of aortitis or improves prognosis. All patients with active cardiovascular syphilis should be examined for neurosyphilis and treated accordingly (see Chapter 24.14.4). If absent, treatment consists of 2.4 million IU of benzathine penicillin intramuscularly at weekly intervals for a total of three doses, alternatively 600 000 IU of procaine penicillin daily by intramuscular injection for 21 days. For patients known to be allergic to penicillin some practitioners desensitize and then use penicillin and others use doxycycline (200 mg orally, twice a day, for 28 days). There are insufficient data to unequivocally recommend the use of other drugs such as ceftriaxone and azithromycin.

Following treatment patients should be reviewed and those with positive non-treponemal tests (VDRL/RPR) should be checked serologically at 6-monthly intervals, with a declining titre used to confirm the adequacy of treatment.

There has been a concern that provocation of a Jarisch–Herxheimer reaction might lead to inflammatory swelling of the aortic wall with the risk of rupture or further critical narrowing of ostial stenoses. However, large numbers of patients with cardiovascular syphilis have been given penicillin without untoward effects, and whereas the Jarisch–Herxheimer reaction can occur rarely, it has never been shown to cause life-threatening changes in the aortic wall.

Surgery may be required to deal with the complications of aortitis. Symptoms of ischaemic heart disease caused by severe ostial stenosis have been relieved successfully by endarterectomy of the coronary orifices, although coronary bypass grafting is the usual treatment of choice. Theoretically, coronary stenting may offer another therapeutic option, although the degenerative changes apparent in the surrounding aortic wall make this seem even less attractive than when atheromatous disease is the cause. Aortic valve replacement has been carried out successfully for severe aortic reflux. Saccular and the less common fusiform aneurysms have been excised and scarred aortic tissue replaced by grafts. Indications for the latter form of surgery are based on: the need to relieve pain; to prevent rupture, the risk of which is considerable when the aneurysm reaches 6 to 7 cm in diameter; and the need to decompress adjacent organs such as the left main bronchus, pulmonary artery, or oesophagus.

Further reading

Augenbraun MH, Rolfs R (1999). Treatment of syphilis, 1998: nonpregnant adults. *Clinical Infectious Diseases* **28** (Suppl 1), S21–8. [Evidence-based recommendations for treatment of syphilis.]

Frank MW *et al.* (1999). Syphilitic aortitis. *Circulation* **100**, 1582–3. [Contemporary imaging of syphilitic aortitis with histology.]

Heggtveit HA (1964). Syphilitic aortitis. A clinicopathologic autopsy study of 100 cases, 1950 to 1960. *Circulation* **29**, 346–55.

Jackman JD, Radolf JD (1989). Cardiovascular syphilis. *American Journal of Medicine* **87**, 425–33. [A case report of a patient with syphilitic aortitis and a literature review of the condition.]

Vlahakes GJ, Hanna GJ, Mark EJ (1998). Case records of the Massachusetts General Hospital. *New England Journal of Medicine* **338**, 897–903. [Syphilitic aortitis with coronary ostial stenosis.]

15.10.4 Cardiac disease in hiv infection

N. Boon

Some form of heart disease is demonstrable at autopsy in approximately 40 per cent of patients with AIDS, and by echocardiography in approximately 25 per cent. However, many of these lesions are mild and heart disease probably causes symptoms in less than 10 per cent and death in only 1 to 2 per cent of all patients infected with HIV.

The common cardiac manifestations of HIV infection are listed in Table 1. Although, tuberculous pericarditis is a major problem in Africa (see section 15.09), heart muscle disease is the most important cardiac complication of AIDS in the Western world.

Pericardial effusion

HIV-related pericardial effusions are usually exudates and tend to occur in patients with advanced disease. The annual incidence of pericardial effusion in AIDS is approximately 10 per cent and the prevalence in all forms of HIV infection is approximately 20 per cent. The development of an HIV-related pericardial effusion is an independent risk factor for early death and when this complication occurs in AIDS the median survival is less than 6 months.

Less than 5 per cent of HIV-related pericardial effusions are associated with pericardial pain or friction rubs. Breathlessness, however, is common and approximately 10 per cent of effusions are associated with the symptoms and signs of tamponade.

Small subclinical effusions do not usually have an identifiable cause and often resolve spontaneously; they can therefore be managed conservatively. By contrast, moderate and large effusions are often symptomatic and are frequently due to specific opportunist infections or malignancy: aspiration is usually advisable. A wide variety of viral, bacterial, and fungal pericardial infections have been reported and conventional antimicrobial therapy combined with drainage procedures can produce good results in those with early disease.

Pericardial tuberculosis is a frequent complication of HIV infection and has become the commonest cause of pericardial effusion in many parts of the world, particularly sub-Saharan Africa. The clinical features vary widely but most patients present with prolonged fever, breathlessness, and pleuropericardial pain. Antituberculous chemotherapy is usually effective but it is not clear whether the benefits of corticosteroids seen in HIV-negative tuberculous pericarditis extend to patients who are HIV positive (see Chapter 15.9).

Malignant effusions are usually due to non-Hodgkin's lymphoma.

Malignant cardiac tumours

Kaposi's sarcoma, which is associated with human herpesvirus-8, is the most common cardiac tumour in HIV disease. The heart is involved in approximately 25 per cent of patients with disseminated Kaposi's sarcoma and there have been a few reports of primary cardiac disease. The tumour shows a predilection for subepicardial adipose tissue and seldom infiltrates the myocardium. Pericardial effusion is a surprisingly rare complication and the diagnosis is usually made at autopsy.

Non-Hodgkin's lymphoma can also involve the heart and is more likely to cause cardiac symptoms and signs. These tumours are usually derived from B cells and may be associated with Epstein–Barr virus or human herpesvirus-8. They are usually metastatic, although primary cardiac tumours do occur, and tend to invade the epicardium. In contrast to Kaposi's sarcoma, this tumour often causes symptomatic pericardial effusion and can also infiltrate the myocardium, provoking fatal arrhythmias including venticular fibrillation and all forms of heart block.

Heart muscle disease

HIV-related heart muscle disease tends to occur in the late stages of HIV infection and usually presents with heart failure or otherwise unexplained cardiomegaly. Left ventricular systolic dysfunction (dilated cardiomyopathy) is present in approximately 15 per cent of patients with AIDS and overt heart failure will develop in approximately 25 per cent of these patients.

The symptoms and signs of heart muscle disease include breathlessness, fatigue, tachycardia, a high jugular venous pressure, a gallop rhythm, and crepitations at the lung bases; they can be subtle and are sometimes mistakenly attributed to anaemia or chest infection. The ECG is usually abnormal, but changes are non-specific and seldom aid diagnosis; increased ventricular ectopic activity, a variety of repolarization changes, and features of left atrial hypertrophy, left ventricular hypertrophy, and left bundle branch block have all been documented. Although chest radiographs may reveal cardiomegaly and pulmonary venous congestion, the diagnosis usually depends on echocardiography, which typically shows global left ventricular dysfunction with enlargement of all the cardiac chambers.

HIV-related dilated cardiomyopathy carries an exceptionally poor prognosis with a median survival of approximately 100 days compared with 500 days for patients with similar disease and normal hearts (Fig. 1); death is often due to progressive heart failure or arrhythmia. Conventional therapy for heart failure is given, but many patients tolerate vasodilators poorly, possibly because peripheral vascular resistance tends to be low due to recurrent sepsis.

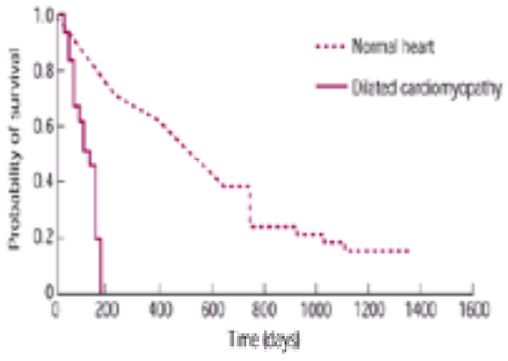

Fig. 1 Kaplan Meir survival curves for patients with HIV-related dilated cardiomyopathy and patients with structurally normal hearts and otherwise similar HIV disease (CD4 count < 20 × 10^6 cells per litre). (Modified from Currie *et al*. (1994). *British Medical Journal* **309**, 1605–7, with permission.)

The aetiology of HIV-related heart muscle disease has not been established beyond doubt and is almost certainly complex and multifactorial. However, it seems likely that an autoimmune lymphocytic myocarditis is the usual substrate for HIV-related dilated cardiomyopathy. There are intriguing parallels between HIV-related heart muscle disease and idiopathic dilated cardiomyopathy (see Chapter 15.8.1 and Chapter 15.8.2), and the pathogenesis of the two conditions may be very similar.

Autopsy and endomyocardial biopsy studies have shown that some form of myocarditis is present in approximately 40 per cent of patients with AIDS and up to 80 per cent of patients with HIV-related heart muscle disease. The Dallas criteria for the diagnosis of myocarditis are seldom satisfied but it can be argued that these are not appropriate in the presence of marked immunodeficiency. Specific forms of myocarditis (e.g. *Toxoplasma gondii* and penicillin hypersensitivity) have been described but these are rare and a non-specific focal lymphocytic myocarditis appears to be the underlying problem in the majority of patients with HIV-related heart muscle disease. The inflammatory infiltrates are composed mainly of CD8+ lymphocytes with increased MHC (major histocompatibility complex) class I antigen expression. In some cases there are excess circulating cardiac autoantibodies.

A variety of molecular techniques have demonstrated that a few transcripts of HIV-1 are present in the myocardium of many patients with heart muscle disease. Myocardial damage is unlikely to be due to direct HIV toxicity because the viral load is low and there is no CD4 receptor on the myocyte; nevertheless, it is possible that the presence of the virus could trigger an autoimmune reaction. Other factors that might contribute to or amplify cardiac damage include increased oxidative stress due to micronutrient (particularly selenium) deficiency, coinfection with cardiotropic viruses (e.g. cytomegalovirus), and direct cytokine-mediated injury.

The antiretroviral drug zidovudine can cause a specific dose-related reversible skeletal myopathy by inhibiting mitochondrial g-DNA polymerase; it can also damage cardiac muscle in rats and may be implicated in some cases of HIV-related heart muscle disease.

Infective endocarditis

Non-bacterial thrombotic (marantic) endocarditis is a disease of unknown aetiology, in which friable clumps of platelets and red blood cells adhere to the cardiac valves. The condition is sometimes complicated by systemic embolism and is associated with a variety of debilitating illnesses. It is a recognized complication of AIDS, but is infrequent and seldom causes clinical problems.

HIV infection is not associated with an increased incidence of infective endocarditis. However, intravenous drug use is an important risk factor for infective endocarditis in this patient group. *Staphylococcus aureus* is the most common pathogen in this setting and the clinical features of the condition appear to be identical in HIV-positive and HIV-negative drug users; the tricuspid valve is usually involved and infected pulmonary emboli may occur. Survival rates are around 80 per cent.

There have been surprisingly few reports of infective endocarditis in HIV-positive individuals who do not use intravenous drugs. Nevertheless, infections with a wide variety of unusual organisms (e.g. *Aspergillus fumigatus* and *Pseudoalleschira boydii*) have been described and it has been suggested that severe immunodeficiency may modify the presentation and course of disease. Salmonella endocarditis, which is usually associated with devastating cardiac complications in HIV-negative patients, appears to carry a surprisingly good prognosis in HIV-positive patients and it is conceivable that, in some situations, even minor changes in the inflammatory reaction may impair vegetation formation and limit valvular damage. On the other hand, there is evidence that other forms of infective endocarditis run a more fulminant course when they occur in the late stages of HIV infection.

Pulmonary hypertension

Significant pulmonary hypertension occurs in approximately 5 per cent of patients with advanced HIV infection. This usually presents with increasing breathlessness and right heart failure. The ECG typically shows right ventricular hypertrophy and the diagnosis is often established by echocardiography which shows characteristic dilatation of the right ventricle with flattening of the interventricular septum (Fig. 2).

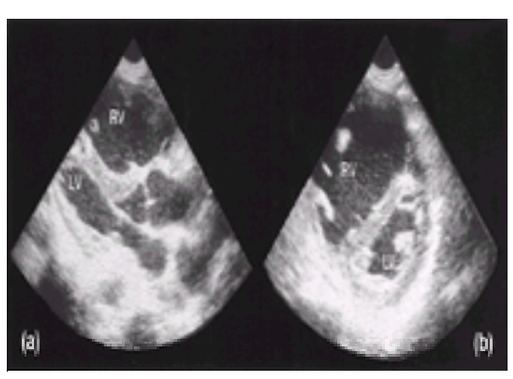

Fig. 2 (a) Long-axis and (b) short-axis parasternal view of a two-dimensional echocardiogram from an HIV-positive intravenous drug user with idiopathic pulmonary hypertension illustrating dilatation of the right ventricle and flattening of the interventricular septum.

The common causes of HIV-related pulmonary hypertension are recurrent chest infections, thromboembolism, and idiopathic or primary pulmonary vascular disease. Patients with left heart failure do not usually live long enough to develop pulmonary hypertension.

The symptoms and signs of right heart failure in patients with cor pulmonale due to recurrent chest infections often improve with appropriate antibiotic therapy and oxygen supplementation. Pulmonary thromboembolism is particularly common in active intravenous drug users, who can sometimes obliterate their pulmonary vascular bed by inadvertently injecting themselves with particulate material.

Approximately 2 per cent of patients with advanced HIV infection develop idiopathic pulmonary vascular disease. The clinical and histological features of this condition are indistinguishable from those of primary pulmonary hypertension (see Chapter 15.15.2.1) and it has been suggested that both conditions may have an autoimmune basis. Treatment is usually ineffective and the outlook is extremely poor.

Sudden death

Sudden death due to ventricular arrhythmias or heart block is a rare but recognized complication of HIV infection. It is usually attributed to a combination of factors including autonomic dysfunction, heart muscle disease, and drug effects. Autonomic dysfunction is a common and disabling complication of AIDS and may cause syncope, presyncope, symptomatic hypotension, and a broad range of arrhythmias, including ventricular tachycardia and ventricular fibrillation. Many potentially proarrhythmic drugs are used in the treatment of HIV infection; these include pentamidine and ganciclovir, which prolong the QT interval and can cause the form of ventricular tachycardia known as Torsades des Pointes (see Chapter 15.6).

Further reading

Barbaro G *et al.* (1998). Incidence of dilated cardiomyopathy and detection of HIV in myocardial cells of HIV-positive patients. *New England Journal of Medicine* **339**, 1093–9.

Chen Y *et al.* (1999). Human immunodeficiency virus-associated pericardial effusion: report of 40 cases and review of the literature. *American Heart Journal* **137**, 516–21.

Currie PF *et al.* (1994). Heart muscle disease related to HIV infection: prognostic implications. *British Medical Journal* **309**, 1605–7.

Heidenreich PA *et al.* (1995). Pericardial effusion in AIDS: incidence and survival. *Circulation* **92**, 3229–34.

Lipshultz SE, ed. (1998). *Cardiology in AIDS*. Chapman & Hall, New York.

Nahass RG *et al*. (1990). Infective endocarditis in intravenous drug users: a comparison of human immunodeficiency virus type 1-negative and -positive patients. *Journal of Infectious Diseases* **162**, 967–70.

Pisani B, Taylor DO, Mason JW (1997). Inflammatory myocardial diseases and cardiomyopathies. *American Journal of Medicine* **102**, 459–69.

Petitpretz P *et al*. (1994). Pulmonary hypertension in patients with human immunodeficiency virus infection: comparison with primary pulmonary hypertension. *Circulation* **89**, 2722–7.

15.11.1 Cardiac myxoma

Thomas A. Traill

Introduction

Cardiac myxomas are benign, typically golfball-sized tumours that grow in the lumen of the atria, usually the left, attached by a stalk to the atrial septum. They are not common, but are important because they can present in a number of ways to general physicians, and because most can easily and permanently be removed by heart surgery. They are easily demonstrated by conventional transthoracic echocardiography, and it is usually the echocardiographer who makes the diagnosis; seldom has the patient been referred with this possibility in mind. Estimates of the prevalence of such a rare condition are necessarily approximate and range from 1 to 5 per 10 000 in autopsy series, or 2 per 100 000 in the general population, with a sex ratio of 2:1 in favour of women. As a cause of left atrial obstruction, myxomas are 200 to 400 times less common than mitral stenosis. The majority of patients are between 30 and 60 years, but there are reports of tumours occurring in infants and in the elderly.

Most myxomas are sporadic, unassociated with other diseases, but there is at least one mendelian syndrome involving myxoma, best named the Carney complex. This is characterized by lentiginosis, multiple myxomas (most of them cardiac), skin fibromas, and various kinds of endocrine overactivity, which has included Cushing's syndrome caused by pigmented adrenocortical hyperplasia, acromegaly, and Sertoli cell tumour. Unlike the usual kind of atrial myxoma, myxomas in Carney's syndrome may arise anywhere in the heart, are commonly multiple, and frequently recur. Inheritance of this rare disease is autosomal dominant, with centrofacial freckling as the most obvious outward marker of the phenotype. This freckling often involves unusual areas, for instance the lips, conjunctiva, and vulva.

Pathology

Cardiac myxomas are benign. Local invasion is unknown and metastatic growth is exceptional, despite the lesions' situation in the bloodstream. They take the form of polypoid masses arising from a stalk, ranging in size from 3 cm to as much as 10 cm or more, with a smooth or lobulated surface and gelatinous consistency. They are frequently covered with more or less adherent thrombus. More than 75 per cent occur within the left atrium, with the base of the pedicle arising from the fossa ovalis or its rim. Occasionally they arise from the base of the mitral valve leaflets, from the posterior part of the left atrium, or from within the right atrium. Sometimes they grow in both atria, in the form of a dumb-bell. Ventricular myxomas are exceptional and seen almost exclusively as part of Carney's syndrome. Because they are in the systemic circulation, left atrial myxomas usually draw attention to themselves at a size smaller than those on the right side.

The histology is that of a loosely woven, sparsely cellular, connective tissue tumour with very infrequent mitotic figures. Several cell types are identifiable, including undifferentiated stellate and polygonal cells, as well as smaller numbers of fibroblasts, smooth muscle cells, and endothelial cells. Among these are found macrophages and plasma cells, and rarely other mesodermal tissues, including bone. Cytogenetic studies fit with the general presumption that these indolent masses are indeed neoplastic, but immunohistochemical studies of differentiation markers do not clearly define the cell type of origin. It is suggested that the source is a primitive multipotential mesenchymal cell and that the predilection of these tumours for the atrial septum reflects the abundance of such cells in this region.

Clinical features

Presentation

Although the wide availability of echocardiography has made the diagnosis of atrial myxoma straightforward, it remains true that the prerequisite for recognizing this rare lesion is to include it in the differential diagnosis of patients presenting with symptoms and signs of much more common conditions. Left atrial myxomas may mimic mitral stenosis and cause left atrial obstruction. They may be the source of emboli to the systemic circulation, and occasionally they may present as an obscure constitutional illness with fever. Right atrial myxomas seldom cause symptoms until they are very large, when they cause right atrial obstruction with elevated systemic venous pressure, splanchnic congestion, and oedema.

Left atrial obstruction

The most common symptoms and signs mimic those of mitral stenosis, with left ventricular inflow obstruction as the chief pathophysiological change. The presenting symptoms are progressive breathlessness, orthopnoea, paroxysmal nocturnal dyspnoea, fluid retention, and atrial arrhythmias. Examination suggests rheumatic heart disease, and before the routine use of ultrasound a few such patients were referred for mitral valve surgery and the lesion was first diagnosed at operation. Some patients may develop pulmonary hypertension before the diagnosis becomes apparent.

Systemic embolism

Systemic emboli occur in about 40 per cent of patients and are frequently the first manifestation of disease. By contrast to mitral stenosis, such emboli often occur while patients are in sinus rhythm. Emboli may be sizeable, large enough even to occlude the aortic bifurcation, and besides thrombus they frequently contain tumour material, so that histological examination may be diagnostic. Thus, when systemic emboli are removed from patients they should always be sent for histological analysis. Typically, patients with systemic embolism are referred for echocardiography, and the diagnosis is then easily made.

Constitutional effects

Constitutional effects of the neoplasm predominate in a few patients. These include fever, weight loss (which is more conspicuous than in mitral stenosis and often occurs without severe left atrial obstruction), Raynaud's phenomenon (rare), finger clubbing (rare), a raised erythrocyte sedimentation rate (present in about 60 per cent of patients), and abnormal serum proteins with elevated immunoglobulin levels. These changes are usually attributed to abnormal proteins secreted by the tumour, although the nature of these has not been determined. Other haematological abnormalities include anaemia, which may be due to mechanical haemolysis, polycythaemia, associated particularly with right atrial tumours, leucocytosis, and thrombocytopenia. Such constitutional changes may prompt an initial diagnosis of infective endocarditis in patients who have heart murmurs, or lead to the suspicion of collagen vascular disease or occult cancer.

Physical signs

In many patients specific cardiovascular signs of myxoma are inconspicuous or absent. In others they vary from a prominent first heart sound to obvious changes similar to those of mitral valve disease. These include apical systolic murmurs, somewhat more common than diastolic rumbles, and signs of pulmonary hypertension with accentuated pulmonary closure and tricuspid regurgitation in some patients. Some may have an audible 'tumour plop' in early diastole, analogous to a mitral opening snap, but this is often heard best only after echocardiographic diagnosis. On combined echocardiographic and phonocardiographic recordings the plop is

seen to coincide with the end of the tumour's downward movement into the ventricle, usually a short time after mitral valve opening. A rare but specific feature of the condition is variation of the auscultatory findings with change in posture; this may be particularly obvious in right atrial tumours.

Investigations

Chest radiography and electrocardiography do not help to distinguish myxoma from mitral valve disease. Left atrial enlargement is common but seldom marked and signs of pulmonary venous hypertension are infrequent. Calcification within the tumour is rarely demonstrable.

Echocardiography

While the first account of left atrial myxoma diagnosed during life was not until 1951, it is now exceptional for the diagnosis to be made first at autopsy. This is chiefly attributable to the wide availability of echocardiography, which has proved itself both reliable and specific for recognizing these tumours. The characteristic pattern of left atrial myoma is easily recognized, and it is no accident that the echocardiographic appearance of these lesions was among the first clinical reports by ultrasonographers, in 1959. Figure 1 illustrates a typical two-dimensional echocardiogram from a patient with left atrial myxoma. This 'four-chamber view' shows the characteristic dense mass of echoes from the tumour lying just above the mitral valve orifice. A video recording would demonstrate the mobility of the mass as it flops to and fro within the atrium, restrained only by its peduncle. Trans-oesophageal echocardiography affords the opportunity to examine the tumour and its attachment with great precision; generally this extra clarity is unnecessary, but on occasion the trans-oesophageal technique is helpful if there is difficulty in differentiating tumour from an atrial thrombus.

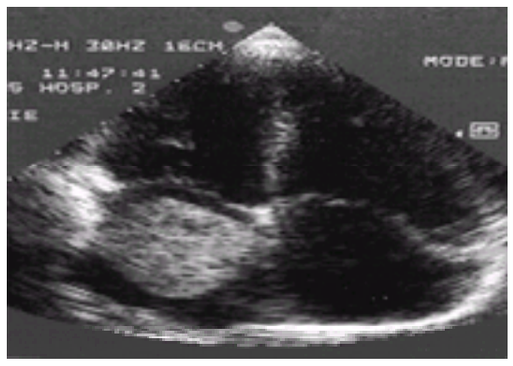

Fig. 1 Echocardiogram in the four-chamber view showing a myxoma occupying much of the left atrium.

The differential diagnosis of left atrial myxoma is seldom difficult. Large masses may occasionally be difficult to distinguish from left atrial ball thrombus, a lesion that is even rarer than myxoma. Smaller left atrial masses may be papillary fibroelastomas or infective vegetations caused by endocarditis. These can usually be distinguished by their clinical context. Masses in the right atrium may also represent thrombus, sometimes propagated from the inferior vena cava, or occasionally venous extension of abdominal cancers, particularly renal cell cancer. In a few patients abundant strands of the Chiari network of right atrial trabeculation may give rise to similar echocardiographic appearances.

Cardiac catheterization

The echocardiographic appearance is so characteristic that angiography no longer has a role in diagnosis of myxoma. The only time to undertake it is in an older patient in whom there is fear of occult coronary artery disease.

Treatment and prognosis

Atrial myxoma is treated by urgent surgical removal. The risk is low, comparable with that of surgery for mitral valve disease. It is important to ensure complete removal of the base by excising a full-thickness button of the atrial septum. The resulting defect is repaired with a small patch.

Functional results of surgery are good. Some patients are left with mitral regurgitation, but this is seldom severe. Recurrence is uncommon, provided excision has been complete, except in Carney's syndrome. In these patients regular echocardiographic follow-up is required, at intervals of 6 months. The rare occurrence after excision of the usual kind of myxoma generally occurs within the first 2 years; thereafter follow-up can safely be infrequent.

Further reading

Casey M *et al*. (2000). Mutations in the protein kinase A r1alpha regulatory subunit cause familial cardiac myxomas and Carney complex. *Journal of Clinical Investigation* **106**, R31–8.

Greenwood WF (1968). Profile of atrial myxoma. *American Journal of Cardiology* **21**, 367–75.

Krikler DM *et al*. (1992). Atrial myxoma: a tumour in search of its origins. *British Heart Journal* **67**, 89–91.

Murphy MC *et al*. (1990). Surgical treatment of cardiac tumors: a 25-year experience. *Annals of Thoracic Surgery* **49**, 612–18.

Schaff HV, Mullany CJ (2000). Surgery for cardiac myxomas. *Seminars in Thoracic and Cardiovascular Surgery* **12**, 77–88.

15.11.2Other tumours of the heart

Thomas A. Traill

While each individually is rare, taken together the other tumours of the heart have an incidence that roughly equals that of myxoma. They include benign lesions, seen especially in children, sarcomas, and secondary involvement by metastasis or direct tumour extension. They are generally first recognized or suspected during echocardiography. Magnetic resonance imaging, or occasionally echo-directed transvenous biopsy, usually yield the diagnosis.

Benign cardiac tumours

Papillary fibroelastoma

The most common tumour seen in adult patients is the papillary fibroelastoma, a small pedunculated mass that hangs off one of the left-sided valve leaflets, usually the mitral valve. Its echocardiographic appearance is very characteristic. The size of the mass and presence of a peduncle distinguish this small tumour from the usual kind of Lambl's excrescence, but histologically they are identical and, like Lambl's excrescences, papillary fibroelastomas probably arise through organization of fibrinous material that collects at the trailing edges of the valve leaflets. Their importance lies in the fact that they have been labelled as a potential source of systemic embolism, and that some authors have recommended that they should be removed as a matter of routine. The evidence to support this view is thin, and this author's recommendation is to remove them only if they have been discovered in the search for a source of otherwise unexplained embolism. If they are an incidental echocardiographic finding then it is safe to leave them alone; aspirin treatment may be recommended.

Fibroma, rhabdomyoma, hamartoma, haemangioma

These are tumours of childhood, rhabdomyoma being the characteristic cardiac tumour in patients with tuberous sclerosis. By contrast to myxomas and fibroelastomas they grow within the myocardium, not into the lumen of the heart. Rhabdomyomas are usually asymptomatic, and when they are they should be left alone, since most regress spontaneously. Fibromas and hamartomas are both very rare, and may present with arrhythmias (particularly ventricular hamartomas, or Purkinje cell tumours) or with haemodynamic abnormalities caused by their mass effect. They require surgical excision, and when this is feasible the long-term results of treatment are very good. Haemangiomas, also very rare, tend to grow, and to develop multiple feeding vessels, so that surgical excision is usually recommended.

Cardiac sarcoma

Primary cardiac sarcomas can have one of several cell types. They are found more often in the right heart than in the left. Haemangiosarcoma is the most common, and typically develops in the right atrium. Rhabdomyosarcoma may develop in the ventricular septum or in the right ventricular outflow tract, as may the still rarer osteosarcoma, or tumours that are undifferentiated. Since these tumours often present with mechanical effects, typically obstruction at the atrial or outflow tract level, surgical resection is often attempted. However, recurrence and metastasis are common, and long-term results are very poor.

Cardiac involvement by other malignancies

Microscopic secondary deposits within the myocardium can often be found in patients who die of metastatic cancer, but intramyocardial secondaries of a size large enough to be of clinical importance are very rare. By contrast, pericardial involvement by lymphoma, or by cancers of the lung, breast, pancreas, and other tumours is not uncommon, and may sometimes be the first presentation of the tumour (see Chapter 15.9). Treatment is analogous to that of malignant pleural effusions, with drainage, creation of a window, or intrapericardial chemotherapy depending on the rest of the clinical situation.

Intraluminal spread of cancer, by direct extension up the inferior vena cava, is a particular feature of renal cell cancers. Diagnosis by echocardiography is generally obvious as the tumour has a very characteristic appearance as it waves, like seaweed in the right atrium and even dangles through the rest of the right heart. It may prove possible to resect the cava, along with the kidney and the tumour mass, under circulatory arrest.

Further reading

Case Records of the Massachussetts General Hospital (1999). *New England Journal of Medicine* **341**, 1217–24.

Olinger GN, ed (2000). Cardiac neoplasms. *Seminars in Thoracic and Cardiovascular Surgery* **12**, 76–129.

15.12 Cardiac involvement in genetic disease

Thomas A. Traill

Introduction

Singling out a few of the more prominent mendelian disorders seen by cardiologists may seem a somewhat arbitrary basis for a chapter, especially in an age when we are exploring the molecular genetic basis for so many more of the common heart diseases. However, this is a grouping that works in practice. Many clinicians find themselves faced from time to time with a patient who has a family history of a known disorder, such as Marfan's syndrome, or who has non-cardiac features that suggest a syndrome, perhaps Noonan's. They may wonder how to make the diagnosis, what else to look for, and how to screen family members.

Two important familial heart diseases are covered elsewhere in this book, namely familial hypertrophic cardiomyopathy (see Chapter 15.8.2), and the long QT syndromes (see Chapter 15.6). Both are genetically heterogeneous: familial hypertrophic cardiomyopathy is caused by mutations in any of a number of the sarcomere proteins responsible for contraction, and the long QT syndromes are caused by mutations of ion channels that affect the cardiac action potential.

The first part of this section deals with syndromes that include congenital cardiac structural defects and is restricted to a few relatively common disorders seen in adult patients. The second part describes the two common connective tissue disorders—Marfan's and Ehlers–Danlos syndromes. A number of other heritable diseases that affect the heart are listed in a table, without discussion in the text. Haemochromatosis (Chapter 11.3) and Friedreich's ataxia (Chapter 24.13.12) are discussed elsewhere in this book; the others, though important to other organ systems, offer little opportunity to the cardiologist for diagnosis or management.

Syndromic congenital heart disease

Aneuploidy disorders

The two commonest chromosomal disorders in adult patients are Down's and Turner's syndromes, and each includes characteristic cardiac abnormalities. A third, Klinefelter's syndrome, does not. Twenty-five to fifty per cent of patients with Down's syndrome have congenital heart disease. The characteristic lesion, present in about half of the affected hearts, is atrioventricular canal defect. This ranges from the relatively simple primum atrial septal defect to the complete type, in which the defect involves both the atrial and ventricular septa, between which there lies a single atrioventricular valve ring. In other patients, ventricular septal defect, tetralogy of Fallot, and persistent ductus arteriosus are seen in roughly equal numbers. Some suspect that patients with Down's syndrome are especially prone to develop pulmonary vascular disease, and hence Eisenmenger reaction, but growing experience with surgical repair for affected children seems to show that this suspicion is ill-founded. Patients with Down's syndrome undergo surgery most easily when they are infants, and the tendency has shifted away from a nihilistic approach to operating on patients with serious cardiac malformations early in life.

Turner's syndrome causes abnormalities of the aorta. The two principal lesions are coarctation and congenital abnormalities of the aortic valve, usually a bicuspid valve. Most patients with coarctation have a bicuspid aortic valve as well, and patients with a bicuspid valve frequently have some degree of anuloaortic ectasia. In some patients with Turner's syndrome the whole aorta is abnormal, either hypoplastic or weakened by the presence of cystic medial necrosis. Aortic dissection may occur, and aortic surgery, to repair coarctation for example, can sometimes be very difficult. Other congenital heart abnormalities are not common in Turner's syndrome, except for anomalies of pulmonary venous return.

Mendelian 'single-gene' disorders causing congenital heart disease

Noonan's syndrome

Noonan's syndrome is the most common heritable syndrome that characteristically causes congenital heart disease. The syndrome shares a number of features with the Turner phenotype, and the two were confused between 1930 and Noonan's studies in the 1960s, which coincided with the advent of karyotyping. In 1963 Noonan described a small series of patients with pulmonary stenosis who shared a characteristic facial appearance. Since then the phenotype has been well described, associated with a normal karyotype and autosomal dominant inheritance. Cardiac involvement has been recognized to include not only pulmonary stenosis, but a wide variety of other lesions, much wider than in Turner's syndrome. A locus has been mapped to chromosome 12.

Patients with Noonan's syndrome are of short stature, with a facies that is variously described as elfin or triangular. There is ocular hypertelorism, and the palpebral fissure may slope downwards (the antimongoloid slant), which may be emphasized by ptosis or an epicanthal fold. The ears are set low and rotated forwards so that the lobes are prominent, and there is characteristic webbing of the neck, the most obvious of the features that may lead to confusion with Turner's syndrome. Pectus deformities are common, as are other miscellaneous skeletal abnormalities, including cubitus valgus. Patients with Noonan's syndrome are prone to develop keloid scars. Cryptorchidism is common, as is delayed sexual maturation, but not infantilism as in Turner's syndrome. Unlike Turner's syndrome, a proportion of patients with Noonan's syndrome have a degree of mental retardation, but this is quite variable. Among this author's patients with Noonan's syndrome are a physician, an architect, a certified accountant, and a high-school mathematics teacher.

The frequency of cardiac involvement in Noonan's syndrome is unknown, since the diagnosis is so easily missed in the absence of congenital heart disease. The most characteristic lesion is pulmonary stenosis, but in contrast to the almost stereotypical cardiovascular findings in Turner's syndrome, the range of congenital heart abnormalities in Noonan's syndrome is broad. In many patients the stenotic pulmonary valve leaflets are not simply fused, as in non-syndromic pulmonary stenosis, but may be dysplastic, thickened, and immobile, unsuitable for simple balloon or surgical valvotomy. Other congenital lesions found in Noonan's syndrome are ventricular and atrial septal defects, tricuspid atresia, single ventricle, and abnormalities of the left ventricle including congenital mitral stenosis, subaortic stenosis, and a combination of these two lesions. The electrocardiogram often shows a superior axis (left axis deviation), even when there is pulmonary stenosis and right ventricular hypertrophy.

The most ominous complication of Noonan's syndrome is cardiomyopathy, taking the form of myocardial hypertrophy complicated by progressive fibrosis. This leads over the course of 5 to 15 years to low cardiac output with very high ventricular diastolic pressures—the pathophysiology of restrictive cardiomyopathy. Since the valvular abnormalities are for the most part correctable, this hypertrophic restrictive cardiomyopathy is the main factor limiting life-expectancy.

Williams' syndrome and familial supravalve aortic stenosis

Williams' syndrome is a contiguous gene phenomenon, caused by a macrodeletion that includes the elastin gene on chromosome 7. A loss-of-function mutation or hemizygosity of the elastin gene alone causes familial supravalve aortic stenosis, inherited as a dominant trait, in which a tight constriction develops in the aorta just above the sinuses of Valsalva. The aortic lesion is generally accompanied by a similar abnormality of the left and right pulmonary arteries, leading to peripheral pulmonary artery stenosis. In Williams' syndrome more far-reaching effects caused by deletion of contiguous genes accompany these cardiac abnormalities. The full syndrome comprises a characteristic facial appearance, with round blue eyes, a distinctive stellate pattern of the irises, depression of the nasal bridge, outwards tilting of the nostrils, abnormal dentition, and big lips, together with small stature, mental retardation, and a history of infantile hypercalcaemia. Mental retardation in Williams' syndrome takes on very individual forms, the patients often being articulate and socially adept: several purported idiots savants have had

Williams'syndrome. Supravalve aortic stenosis in either of these syndromes can lead to severe left ventricular outflow obstruction, with left ventricular failure or even sudden death: surgical treatment may be required.

DiGeorge and velocardiofacial syndromes (chromosome 22 deletion syndrome)

DiGeorge syndrome, described in 1965, comprises abnormalities of the parathyroid glands, absence or hypoplasia of the thymus, and conotruncal abnormalities of the heart such as pulmonary atresia and severe forms of tetralogy of Fallot. A number of affected patients have learning disabilities or schizophrenia. It was recognized soon after the original description that the syndrome is generally caused by deletions in a region of chromosome 22.

Velocardiofacial syndrome, or Shprintzen's syndrome, described in 1981, comprises similar cardiac abnormalities along with cleft palate, a characteristic facies, and learning difficulty. It has since proved to be caused by deletions in the same region of chromosome 22, now often referred to as the DGCR (DiGeorge critical region). A third syndrome, known as 'conotruncal anomalies face', also linked to this site has been described.

With a broad spectrum of phenotypic variation, and deletions that are often quite large, it was suspected for some time that these syndromes are related manifestations of a contiguous gene phenomenon, just as in Williams' syndrome. However, it has emerged that the size of the deletion does not predict the extent of the phenotype, and that within a family the same (presumably stable) deletion can be the cause of a wide range of phenotypes. Mutations that may well account for the entire range of phenotypes have been recently been discovered in a single gene (*Ufd1*) within the region, in which case renaming as a single gene syndrome will become appropriate.

Heart–hand syndromes

The two commonly recognized heart–hand syndromes are Holt–Oram syndrome and Ellis–van Creveld syndrome.

Holt–Oram syndrome

Holt–Oram syndrome, inherited as an autosomal dominant trait, was described in 1960, comprising a secundum atrial septal defect and skeletal abnormalities, principally affecting the upper limbs and shoulder girdle, never the legs, and more pronounced in the left arm. Within a family, affected individuals may have skeletal abnormalities, congenital heart disease, or both. The limb abnormalities cover a wide spectrum from just a triphalangeal thumb to phocomelia. Abnormalities of the hand and forearm always involve the radial side and thumb (in contrast with Ellis–van Creveld syndrome). The characteristic cardiac abnormality is fossa ovalis (secundum) atrial septal defect, but affected patients may have other relatively simple lesions, for example ventricular septal defect or pulmonary stenosis.

Holt–Oram syndrome has been mapped and cloned. The mutation is in a transcription factor known now as TBX5, a close homologue of a transcription factor seen as phylogenetically far back as the fruit fly. Mutations of the homologous gene in the fruit fly produce abnormalities of the wing.

Ellis–van Creveld syndrome

Ellis–van Creveld syndrome is inherited as a recessive trait, hence the more complete clinical descriptions have come from studies in genetically circumscribed communities, notably the Old Order Amish of Pennsylvania where thanks to a founder effect the gene is common and homozygotes abound. The syndrome, described in 1940, includes dwarfism, caused mainly by shortening of the forearms and lower legs, and symmetrical polydactyly affecting the ulnar side with accessory sixth and even seventh digits attached to or beyond the little finger. Cardiac involvement is very common, present probably in three-quarters of homozygotes. The characteristic lesion is common atrium—a lesion that has the appearance on echocardiography and to the surgeon of a very large primum atrial septal defect. A few patients have more complete forms of atrioventricular canal defect, and, at least among the Amish, there is a high perinatal mortality rate among affected infants, suggesting the possibility of still more extensive cardiac involvement. The gene has been mapped to chromosome 4 and sequenced, but the protein's function is unknown.

Connective tissue disorders

Marfan's syndrome

Thanks principally to the work of McKusick and his collaborators, beginning in 1955, Marfan's syndrome has become the paradigm for the clinical and genetic investigation of the heritable disorders of connective tissue. The importance of the syndrome is heightened by the fact that its recognition and treatment have had a dramatic impact on survival among those affected. In 1896 Marfan described a patient with what he termed arachnodactyly. In the century since, it has been appreciated that the syndrome is mendelian and pleiotropic, involving several apparently unrelated organs whose common feature proves to be the importance of elastic tissue to their structural integrity. Ocular involvement, with the lens subluxed because of failure of its suspensory ligament, was recognized early in the twentieth century. Cardiovascular involvement was noted incidentally in the 1940s, and studied systematically from the 1950s onwards. Skeletal involvement includes—besides long limbs and arachnodactyly—scoliosis and other abnormalities of the thoracic cage. The sternum may be pushed outwards or inwards by the abnormally long ribs, hence pectus carinatum and/or excavatum, often asymmetrical. Skin involvement is identified by light-coloured striae, which should be looked for over the deltopectoral groove and the flanks. Less common findings are dural ectasia, which can sometimes be so marked as to cause radicular symptoms, and spontaneous pneumothorax or apical blebs.

The characteristic cardiovascular findings in Marfan's syndrome are aneurysmal dilatation of the aorta, and occasionally other large arteries, and floppy mitral valve. The former was recognized in the 1920s but not really addressed until McKusick showed that it was the principal cause of early death in the disease. Shortly afterwards, echocardiography became available to identify and follow these abnormalities, and surgical techniques were developed by Bentall and Gott to repair the aneurysms. Until then, median life expectancy for men with Marfan's syndrome had been 45 years, for women a year or two longer.

Fibrillin gene mutations

The syndrome is caused in almost all patients by mutations of the fibrillin gene on chromosome 15. Fibrillin is a pleated protein laid down in sheets, and a mutation functions as a dominant negative by coding for a misshapen protein that interferes with this polymerization. This is a common mechanism among the connective tissue disorders, seen in Ehlers–Danlos syndrome and in osteogenesis imperfecta. It appears from animal models that in Marfan's syndrome the stage is set for abnormal arterial wall development early in fetal life.

The fibrillin molecule is large and most of the disease-causing mutations have yet to be described, hence genetic diagnosis by screening for known mutations is seldom possible and diagnosis usually depends on applying clinical criteria. There are a number of polymorphisms within the gene, so in some kindreds it is possible by tracking particular alleles to determine which is associated with the disease, and therefore contains the pathogenetic mutation. This has allowed diagnosis of the syndrome in individual family members in whom the clinical findings were uncertain, and has been used for prenatal diagnosis. Furthermore, the technique makes it possible to infer the existence of a fibrillin mutation in kindreds where the phenotype has not met clinical criteria for Marfan's syndrome; if aortic ectasia segregates with a particular copy of the fibrillin gene, then the chances are high that a fibrillin mutation somewhere in that copy is the pathogenetic mechanism.

Diagnostic criteria

The clinical diagnosis of Marfan's syndrome rests on major and minor criteria. In an index case, involvement of three organ systems is required, with major criteria in two. Major criteria can be aortic aneurysm, lens subluxation, characteristic skeletal abnormalities, or dural ectasia. Minor criteria can be striae, mitral valve prolapse, joint laxity, the facies, or moderate pectus excavatum. Characteristic skeletal abnormalities can be: arachnodactyly (encircling the wrist with the thumb and little finger, the 'wrist sign', and making a fist with a protruding thumb, the 'thumb sign'), marked pectus deformity, increased wing-span to 5 per cent more than the height, and scoliosis. In the relative of an index case, the positive family history becomes another major criterion.

In clinical practice, determining whether a patient satisfies these criteria may be fairly subjective and requires experience with the syndrome. Often it is enough to know whether or not there is cardiovascular involvement, and there are numerous families with aortic aneurysms or ectasia whose full phenotype does not satisfy clinical criteria for Marfan's syndrome, yet whose long-term management is identical. Equally, a lanky patient who has a normal aorta needs only infrequent follow-up,

even though there may be a suspicion that he has a mild case of the syndrome.

Clinical management

Patients with Marfan's syndrome should be followed up with annual or 6-monthly echocardiograms to examine the aortic root. If there is reason to suspect that the aorta may be dilated above the echo plane then CT scanning or MRI is required at least once to validate the echo measurement. When the maximum measurement across the aorta reaches 5 cm, we generally recommend surgical replacement of the aortic root, to prevent aortic dissection (see Chapter 15.14.1), which becomes a real risk once the dimension reaches 6 cm. The traditional and very successful approach is with the composite graft: a mechanical aortic valve prosthesis to which is indissolubly attached a tubular vascular prosthesis is used to replace the entire aortic root and annulus. The coronary artery ostia are excised from the native aorta and reattached to the prosthetic root. Recently, to avoid anticoagulation in certain patients, there has been interest in a valve-sparing technique of root replacement in which a vascular prosthesis is fitted snugly over the aortic valve commissures, with the native leaflets suspended in their normal anatomical arrangement. Long-term success with this approach will depend on the degree to which the valve leaflets themselves degenerate because of the connective tissue abnormality. The Ross (pulmonary autograft) procedure is not appropriate in Marfan's syndrome. After surgery, and especially in patients whose surgery was done as an emergency for dissection, follow-up is with periodic imaging by CT or MRI to keep the remaining aorta under surveillance. Management of mitral prolapse and regurgitation in Marfan's syndrome is the same as in other patients. Surgery is required for severe or symptomatic regurgitation; mitral valve repair has proved surprisingly successful.

It is usual to treat patients who have aortic involvement with b-blockers, to slow the progression to aneurysm. We generally advise against excessively demanding sports, particularly competitive basketball, but in all affected children it is important to balance the risks of aortic disease against the importance of normal psychological development. Pregnancy is not contraindicated in all women with Marfan's syndrome, but genetic counselling should be offered, and it is advised that people not become pregnant if the aorta is enlarged to over 4 cm.

Ehlers–Danlos syndromes

In the early part of the twentieth century Ehlers and Danlos described an association between hyperextensibility of the skin, atrophic scarring, and hypermobility of the large joints. Several different mendelian Ehlers–Danlos syndromes are now defined, caused by mutations of different collagen molecules (and some others besides collagen). Since phenotypes may be highly variable, even within a single family, the details of classifying these diseases according to their pathogenesis are still not completely worked out.

Three principal cardiovascular manifestations of the Ehlers–Danlos syndromes are recognized, namely vascular fragility and rupture, mitral valve prolapse, and aortic root ectasia. The first of these is the potentially fatal feature of the vascular type of Ehlers–Danlos (formerly type IV), inherited as a dominant trait and caused by mutations in the type III procollagen molecule. Patients with this form are prone to spontaneous rupture of large and medium-sized arteries. The aorta does not dilate as in Marfan's syndrome and dissection is less common than simple through and through tearing. Less common, but also potentially fatal complications of the vascular type, are spontaneous perforations of the bowel: with severe vascular fragility, treatment of this or other surgical emergencies can be difficult or impossible.

The classic form of Ehlers–Danlos syndrome (formerly types I and II), inherited as a dominant trait, is characterized by marked skin extensibility, joint laxity, and characteristic wide, atrophic ('cigarette paper') scars at the sites of previous injury or surgery. Patients frequently have mitral valve prolapse, as do many people with joint laxity who do not have diagnosable Ehlers–Danlos syndrome, but only a few progress to develop severe mitral reflux or to the point of requiring surgery. Enlargement of the aortic sinuses of Valsalva may occur, but this is seldom severe or progressive. Surgical replacement of the aortic root, as is performed in Marfan's syndrome, is exceptional in Ehler–Danlos syndrome.

Other heart-related connective tissue and metabolic disorders

Osteogenesis imperfecta causes aortic and mitral regurgitation, as do several of the mucopolysaccharidoses (Table 1). It is striking, particularly in the case of osteogenesis imperfecta, how healing is almost non-existent where there is foreign material. If the opportunity arises, even years later, to inspect the operative result in a patient who has undergone valve replacement, the sutures look as though they had only just been placed, with minimal endothelial reaction and scar tissue formation.

Inherited disorders of heart muscle

All three principal groupings of cardiomyopathy—dilated, hypertrophic, and restrictive—include mendelian forms, but familial hypertrophic cardiomyopathy is by far the best studied (see Chapter 15.8.2). Mutations in the actin molecule have been implicated in some families with dilated cardiomyopathy, as have inborn errors of metabolism affecting high-energy phosphate production, for example carnitine deficiency. Other types of familial dilated cardiomyopathy are seen in several striated myopathies, notably Becker's muscular dystrophy, and in mitochondrial dystrophies. The severity of cardiac involvement in muscular dystrophy is quite variable, and the converse, so that some cases involving both kinds of myopathy may be missed. For example, muscle wasting caused by limb-girdle dystrophy may be ascribed to cardiac cachexia in a case of severe cardiomyopathy, and cardiomegaly in a patient with endstage muscular dystrophy may be wrongly attributed to cor pulmonale.

Arrhythmogenic right ventricular dysplasia is a familial disease, inheri-ted as a dominant trait, that affects almost exclusively the right ventricle. The pathology consists of replacement of islands of right ventricular myocardium with fatty and fibrous connective tissue. Generally these islands are small and the disease seldom leads to any detectable mechanical deficit. However, the areas of fibrous replacement create the substrate for ventricular arrhythmias and the clinical presentation is with palpitations, syncope, and even sudden death. Exercising may unmask the tendency to arrhythmias, so the disease is particularly recognized among athletes. Echocardiography or MRI usually establishes the diagnosis. Treatment is by management of the arrhythmia; severely affected patients may require implantation of an automatic defibrillator.

Further reading

Dietz HC *et al*. (1991). Marfan syndrome caused by a recurrent *de novo* missense mutation in the fibrillin gene. *Nature* **352**, 337–9.

Gott VL *et al*. (1999). Replacement of the aortic root in patients with Marfan's syndrome. *New England Journal of Medicine* **340**, 1307–13.

Lowery MC *et al*. (1995). Strong correlation of elastin deletions, detected by FISH, with Williams syndrome: evaluation of 235 patients. *American Journal of Human Genetics* **57**, 49–53.

McKusick VA (2000). Ellis–van Creveld syndrome and the Amish. *Nature Genetics* **24**, 203–4.

Noonan JA (1999). Noonan syndrome revisited. *Journal of Pediatrics* **135**, 667–8.

Pepin M *et al*. (2000). Clinical and genetic features of Ehlers–Danlos syndrome type IV, the vascular type. *New England Journal of Medicine* **342**, 673–80.

Pyeritz RE (1983). Cardiovascular manifestations of heritable disorders of connective tissue. *Progress in Medical Genetics* **5**, 191–302.

Yamagishi H *et al*. (1999). A molecular pathway revealing a genetic basis for human cardiac and craniofacial defects. *Science* **283**, 1158–61.

15.13 Congenital heart disease in adolescents and adults

*S. A. Thorne and P. J. Oldershaw**

Introduction

Doctors in all areas of medicine and surgery will encounter the growing number of patients with congenital heart disease who survive beyond childhood. It is therefore important that all doctors have an understanding of the principles of congenital heart disease and enough knowledge to know when to refer such patients to a specialist centre.

The future size of the population of long-term survivors will be influenced by a number of factors. In the era before paediatric and neonatal cardiac surgery, 70 per cent of the approximately 4 per 1000 live-born babies diagnosed as having congenital heart disease died before their 10th birthday. However, with advances in medical and surgical care during childhood the majority can now expect to survive into adulthood, increasing the numbers of patients with operatively modified disease. The advent of echocardiography has allowed less severe lesions that previously presented later in life to be diagnosed in infancy, so that the true incidence of congenital heart disease is around 10 per 1000 live-born babies. In societies where termination of pregnancy is available, prenatal diagnosis by fetal echocardiography may result in a reduced incidence of live-born infants with severe congenital heart disease, but its full impact on the population of long-term survivors remains to be determined.

Many of those who survive to adulthood do so with surgically modified lesions, and the continuing evolution of new surgical techniques creates a population with different residual lesions, long-term complications, and survival than earlier generations with the same initial diagnosis. Careful follow-up is therefore crucial, not only to provide high standards of clinical care, but also to provide feedback about late results in order to inform initial management in infancy. For example, as a result of such long-term follow-up information, the operation of choice for transposition of the great arteries is now the arterial switch, because of the late problems encountered in patients who had undergone interatrial repair with the Senning or Mustard operations.

The concepts of congenital and acquired heart disease are arbitrary, lesions may change and develop during a patient's lifetime either as part of the natural history, or in response to surgical intervention. For example, aortic regurgitation may be acquired in the presence of a subaortic ventricular septal defect because the aortic valve which forms the roof of the defect is unsupported and becomes incompetent as a result of the Venturi effect 'sucking' one of the valve leaflets into the defect. In double-inlet left ventricle with ventriculoarterial discordance, the aorta arises from the rudimentary right ventricle via a ventricular septal defect that is usually non-restrictive and hence does not limit aortic flow. However, age-related or surgically induced ventricular hypertrophy and interventions that reduce the volume load on the ventricle may reduce the size of the ventricular septal defect, causing subaortic stenosis. Changing lesions can also be observed *in utero*, further challenging the division between congenital and acquired disease. In some cases of pulmonary atresia with intact interventricular septum, serial fetal examinations may show pulmonary stenosis evolving into pulmonary atresia, with concomitant failure of development of an initially normal-looking right ventricle.

Molecular genetics in congenital heart disease

Advances in molecular genetics are changing our understanding of congenital heart disease, not only in terms of recurrence risks, but also in embryogenesis and in genetic–environmental interactions. Recently, single gene abnormalities have been identified in Holt–Oram syndrome (skeletal and cardiac anomalies, especially atrial septal defect) and in non-Down's atrioventricular septal defect. New technologies have begun to explain the overlapping phenotypes of di George syndrome and velocardiofacial syndrome which comprise a variety of defects of neural crest-derived tissues, and are also known by the acronym, CATCH 22. In both syndromes there is deletion or microdeletion of a region of chromosome 22q11, the phenotypic spectrum being dependent on the degree and position of the deletion. This discovery has implications for genetic counselling: if a patient has tetralogy of Fallot with 22q11 microdeletion, the chances of their offspring inheriting the microdeletion and a phenotypic abnormality are higher than if there is no microdeletion.

Some cardiac defects have a clear environmental cause, such as patent arterial duct and peripheral pulmonary arterial stenosis in maternal rubella and ventricular and atrial septal defects in maternal alcohol abuse. However, in up to 80 per cent of congenital heart defects, no clear single genetic or environmental factor is implicated, that is, the cause is multifactorial and due to interactions between gene(s) and the environment. Candidate genes for specific stages in embryogenesis are beginning to be identified. For example, septation of the ventricular outflow tract appears to be dependent on the HIRA gene regulating migration of neural crest cells; attenuated expression of this gene may play a role in di George and velocardiofacial syndromes.

Cyanotic congenital heart disease

Cyanosis and pulmonary vascular disease are common problems in congenital heart disease and are discussed in general terms below.

Eisenmenger syndrome: defects with secondary pulmonary vascular disease

The Eisenmenger reaction describes the pathophysiology of patients who have pulmonary hypertension at systemic level as a result of high pulmonary vascular resistance with a reversed or bidirectional shunt. The shunt is usually a non-restrictive communication between the systemic and pulmonary circulations and may occur at atrial, ventricular, or arterial levels. Pulmonary vascular disease is established early in life when the shunt is at ventricular or arterial level. A continuing large left to right shunt at systemic pressure causes the pulmonary vascular resistance to rise progressively until it exceeds systemic vascular resistance and the shunt reverses, establishing Eisenmenger physiology. By comparison, the pulmonary hypertension associated with non-restrictive shunts at atrial level usually occurs later in life.

Clinical findings

Whatever the underlying defect, some examination findings are shared. Patients have cyanosis and clubbing and may be plethoric. There is a right ventricular heave and the pulmonary component of the second heart sound is palpable and loud. A pulmonary ejection click may be audible, also a soft early diastolic murmur of pulmonary regurgitation (Graham–Steell murmur). A soft systolic flow murmur may be heard from the dilated pulmonary artery. No systolic murmur can be heard from the lesion responsible for the pulmonary vascular disease since the chambers on both sides of it are at equal pressures. Thus the presence of a loud systolic murmur brings in to doubt a diagnosis of Eisenmenger syndrome, although the murmur of associated tricuspid regurgitation may occasionally be loud.

It is frequently possible to distinguish between the common lesions associated with the Eisenmenger syndrome on clinical grounds (Table 1). The patient with an Eisenmenger duct has differential cyanosis and clubbing since fully saturated blood from the left ventricle supplies the aortic arch and its branches before mixing occurs with desaturated pulmonary arterial blood via the patent duct. The right hand may therefore be pink with no clubbing, the left may be slightly more cyanosed because of the origin of the left subclavian artery opposite the duct, and the toes are more deeply cyanosed and clubbed (see Taussig–Bing anomaly for reversed differential cyanosis). The second heart sound may be closely or normally split. By contrast, cyanosis and clubbing is uniform when the right to left shunt occurs at atrial, ventricular, or ascending aortic (as in truncus arteriosus) levels. The second sound is single in ventricular septal defect, atrioventricular septal defect, and truncus but may be split in an atrial septal defect.

Natural history and complications

Survival into adulthood with the Eisenmenger syndrome is common. Symptoms of breathlessness relate to the degree of hypoxia; many patients feel worse in hot weather or after a hot bath because the resulting systemic vasodilation is not accompanied by a reduction in pulmonary vascular resistance, so the right to left shunt is enhanced and the patient becomes more hypoxic.

The patient with Eisenmenger syndrome is prone to all the complications of cyanotic heart disease discussed below. In addition, exercise-induced syncope may occur and is exacerbated by hot weather and dehydration. Haemoptysis is common and may be fatal. It is usually due to rupture of small hypertensive intrapulmonary vessels, or more rarely to thrombosis *in situ* and pulmonary infarction. All patients with haemoptysis should be admitted and the systemic pressure kept low by bed rest and b-blockade; the pulmonary artery pressure is the same as that measured in the brachial artery. Any non-steroidal anti-inflammatory agents should be stopped and vasodilators should not be used. If the haemoptysis is massive, diamorphine should be administered and consideration given to selectively intubating the non-bleeding lung. Fresh frozen plasma or cryoprecipitate may be given. Bronchoscopy has no role and may worsen the haemorrhage. Spiral CT differentiates pulmonary artery thrombosis from intrapulmonary haemorrhage. Few data exist to direct management of patients with pulmonary arterial thrombus. Warfarin may increase the risk of bleeding whilst failing to reduce the thrombus, and aspirin should be avoided as it may exacerbate haemorrhage associated with thrombocytopenia.

Right ventricular failure may be precipitated by atrial arrhythmia and usually occurs after the age of 30 years. Decline may be heralded by the onset of right ventricular failure, supraventricular arrhythmia, and haemoptysis. Death is sudden in about 30 per cent of patients and results from arrhythmia or massive haemoptysis. In some patients death appears to follow progressive hypoxia, terminating in bradycardia and asystole from which resuscitation is impossible.

Pregnancy (see below) and non-cardiac surgery pose major risks. The latter is particularly dangerous when carried out without the benefit of expert cardiological anaesthetic and perioperative care. A sound understanding of the pathophysiology and the importance of avoiding vasodilators, dehydration, hypotension, and air emboli are vital.

Investigations

The chest radiograph shows a dilated pulmonary trunk because of high pulmonary blood flow in earlier life, but the lung fields are oligaemic (Fig. 1 and Fig. 2). Unless cardiac failure intervenes, the cardiac silhouette is usually normal, the effects of volume overload having regressed as pulmonary vascular resistance increased and the left to right shunt diminished and disappeared. The electrocardiogram shows p pulmonale and biventricular hypertrophy. The echocardiogram should establish the site of the shunt and allow an estimation of pulmonary arterial pressure and ventricular function.

Fig. 1 Chest radiograph of a 35-year-old woman with Eisenmenger secundum atrial septal defect. The aortic knuckle is small and the central pulmonary arteries enlarged, indicating pulmonary arterial hypertension; the lung fields are clear. The cardiac silhouette is not enlarged.

Fig. 2 Chest radiograph of a 45-year-old woman with Eisenmenger arterial duct which is calcified and fills the indentation between the aortic knuckle and main pulmonary artery (arrow). There has been an exploratory left thoracotomy.

Cardiac catheterization is unnecessary and potentially dangerous for patients with established pulmonary vascular disease. The only indication is for those patients

whose pulmonary vascular disease is suspected to be reversible and who would be considered for surgical repair if reversibility can be confirmed. This situation is rarely encountered in the adult population. Histologically, pulmonary vascular disease progresses from medial hypertrophy through intimal proliferation with migration of smooth muscle cells, to progressive fibrosis and obliteration, dilatation, the development of angiomas, and finally fibrinoid necrosis. Those who have developed fibrotic and obliterative changes are likely to have irreversible pulmonary vascular disease. Routine lung biopsy is not recommended; it carries a high risk in the adult with pulmonary hypertension and is unlikely to show reversible pathology. In addition, thoracotomy scars are a relative contraindication to heart–lung transplantation.

Treatment options in the Eisenmenger syndrome

Avoiding unnecessary intervention is the mainstay of management. Heart failure and arrhythmia should be treated with care to avoid overdiuresis and vasodilation. The limited role of phlebotomy for symptomatic hyperviscosity is discussed below.

Heart–lung transplantion, or lung transplantion with repair of the cardiac defect, are often seen as the ultimate options. However, current donor shortages and the high risk of transplanting patients with long-standing cyanosis who are prone to excessive haemorrhage, renal failure, and technical surgical difficulties due to previous thoracotomy, mean that many patients never receive a transplant. Whether chronic therapy with oxygen or with pulmonary vasodilators, such as prostaglandin or nitric oxide, has a role in improving morbidity or mortality in adults remains to be seen. Given the shortage of organ donors, it may not be realistic to consider such approaches as a temporary bridge to transplantation.

Cyanotic heart disease: a multisystem disorder

Cyanotic heart disease is a multisystem disorder; its manifestations are listed in Table 2.

Secondary erythrocytosis

Chronic hypoxia is the stimulus to the increased red blood cell mass and high haematocrit found in cyanotic heart disease. This physiological response increases the oxygen carrying capacity of the blood and may improve tissue oxygenation sufficiently to reach a new equilibrium at a higher haematocrit. However, adaptive failure occurs if the increase in blood viscosity brought about by the high haematocrit impairs oxygen delivery and negates the beneficial effects of erythrocytosis.

The secondary erythrocytosis of cyanotic heart disease is a physiological response, often associated with thrombocytopenia. It is fundamentally different to the generalized increase in all haemopoietic stem cell lines found in the malignant disease, polycythaemia rubra vera. Failure to differentiate between these two phenomena has contributed to the persistent mismanagement of erythrocytosis in cyanotic heart disease. Three misconceptions lead to inappropriate venesection in cyanotic heart disease:

- Misconception 1—volume replacement is not necessary. If venesection is performed without simultaneous volume replacement, the sudden fall in systemic blood flow, oxygen delivery, and cerebral perfusion may result in cardiovascular collapse. Simultaneous infusion of an equal volume of 0.9 per cent saline or colloid should be given.
- Misconception 2—venesection is performed to reduce the risk of stroke. The risk of stroke in adults with cyanotic heart disease does not relate to the haematocrit, but rather to microcytosis and iron deficiency brought on by injudicious venesection.
- Misconception 3—venesection should be done routinely to keep the haematocrit less than 65 per cent. The only indication for venesection is for the temporary relief of symptoms of hyperviscosity in hydrated, iron-replete individuals with a haematocrit greater than 60 to 65 per cent (Table 3 and Table 4). If the patient does not gain symptomatic improvement then further venesection is unlikely to be beneficial. Any dehydration should be corrected before assessing the need for venesection. Some patients reach a stable equilibrium with a haematocrit greater than 70 per cent; venesection is not indicated if there are no symptoms of hyperviscosity. The only exception is the preoperative patient with thrombocytopenia and a high haematocrit, when venesection may cause a temporary rise in platelet count and a reduction in perioperative bleeding.

Microcytic iron-deficient erythrocytes have a reduced oxygen carrying capacity and are less deformable than biconcave iron-replete cells and so increase blood viscosity, negating any beneficial effect of venesection in reducing the haematocrit. Iron deficiency also causes muscle weakness and myalgia independent of its effect on blood viscosity. If standard doses of iron supplements are given, uncontrolled erythropoiesis occurs and the haematocrit rises rapidly, resulting in a cycle of excessive venesection and iron deficiency, leaving the patient symptoms of hyperviscosity induced by both the haematocrit level and iron deficiency. Low-dose iron replacement (ferrous sulphate, 200 mg daily) combined with close monitoring of the blood count so that iron therapy is withdrawn as soon as the haematocrit rises (often within a week) should allow the gradual recovery of iron stores and the avoidance of counterproductive venesection and further iron deficiency. Hydroxyurea is an antitumour agent that may have a role in suppressing the erythrocytotic response to iron therapy in patients with a high haematocrit. It also causes thrombocytopenia and neutropenia, so should be used with caution.

Disorders of coagulation

Why patients with cyanotic disease are at increased risk of haemorrhage and thrombosis is poorly understood. There is often a mild thrombocytopenia that may be partly due to shortened platelet survival time, and the large multimeric forms of von Willebrand factor and other clotting factors may be depleted. Bleeding may be minor and mucocutaneous, but major haemorrhage can occur during surgery, or from pulmonary haemorrhage (see above). Coagulation testing may yield spurious results in patients with a haematocrit greater than 55 per cent unless the amount of citrate anticoagulant is reduced, according to the following equation:

$$\text{Volume of anticoagulant per ml blood} = 100 - \text{haematocrit}595 - \text{haematocrit}$$

Other complications of cyanotic heart disease

Right to left shunting creates a risk of paradoxical embolism causing stroke and cerebral abscess as well as air emboli from venous lines not fitted with filters. Patients who require transvenous pacing should be formally anticoagulated with warfarin to prevent paradoxical thromboembolism from pacing leads.

Despite the high incidence of hyperuricaemia, attacks of acute gout are uncommon and asymptomatic hyperuricaemia does not require treatment. Acute attacks should be treated with colchicine, avoiding non-steroidal anti-inflammatory agents because of their detrimental effects on haemostasis and renal function. As in primary hyperuricaemia, allopurinol is useful in preventing recurrence. The renal abnormalities outlined in Table 2 are rarely associated with abnormal baseline renal function. However, renal failure may be precipitated by hypotension and dehydration, especially in combination with radiographic contrast media or non-steroidal anti-inflammatory agents. Acne is a common complaint in adolescents and adults with cyanotic disease and may be widespread and psychologically debilitating. When severe it may also increase the risk of bacteraemia and endocarditis.

Digital clubbing is almost universal in cyanotic heart disease, and some degree of hypertrophic osteoarthropathy of the long bones may occur in up to one-third of patients. Symptoms include aching and tenderness of the long bones of the forearms and legs. There is oedema and cellular infiltration, causing lifting of the periosteum that is visible radiographically, and new bone formation and resorption. Localized activation of endothelial cells by an abnormal platelet population, with the ensuing release of fibroblast growth factors, may play a central role in the pathogenesis of both phenomena.

Cyanotic patients become more hypoxic during air travel as the partial pressure of oxygen in a pressurized aircraft is lower than that at sea level. However, such travel seems to be well tolerated and supplemental oxygen should not normally be necessary. Travellers should be warned to avoid dehydration and to plan their journeys to avoid having to carry baggage long distances within large airports.

Classification of congenital heart disease

A classification according to pathophysiological groups allows discussion of the basic physiological principles of congenital heart disease (Table 5 and Table 6). A full morphological classification and discussion of segmental sequential analysis is beyond the scope of this chapter.

Specific lesions

Anomalies of pulmonary venous drainage

Total anomalous pulmonary venous drainage

Total anomalous pulmonary venous drainage occurs in 1 per 17 000 live-born babies. All four pulmonary veins drain into the right atrium either directly, or via a common vein into a systemic vein. The anomalous veins may follow:

1. a supracardiac course draining to the superior vena cava, azygos, or brachiocephalic veins;
2. a cardiac course, draining to the right atrium directly or to the coronary sinus directly or via a persistent left superior vena cava connection; or
3. an infradiaphragmatic course, draining to the portal vein or inferior vena cava.

The presence of pulmonary venous obstruction is the most important predictor of a poor outcome. Associated anomalies include an obligatory right to left shunt, nearly always at atrial level.

The condition presents in infancy and 98 per cent of patients reaching the adolescent or adult clinic will have survived corrective surgery in early life. Unless there is residual pulmonary hypertension most such adults should be asymptomatic, have a normal cardiovascular examination, and an excellent prognosis. In the long term, atrial arrhythmias may develop, probably with a similar incidence to that following repair of secundum atrial septal defect. Patients who are still growing may develop obstruction of the redirected pulmonary venous pathway and present with dyspnoea, signs of pulmonary oedema, evidence of pulmonary venous congestion on the chest radiograph, and an obstructive Doppler flow signal at the site of the stenosis.

The rare patient who reaches adulthood without an operation is likely to have survived because of a large atrial septal defect and unobstructed pulmonary venous drainage. They will be cyanosed, have developed pulmonary vascular disease, and be at risk of atrial tachyarrhythmias and right heart failure. The chest radiograph has the appearance of a large atrial septal defect with a small aortic knuckle, cardiomegaly, and a dilated main pulmonary artery. In addition the anomalous veins may cause an abnormal vascular shadow.

Partial anomalous pulmonary venous drainage

There is anomalous drainage of some of the pulmonary veins to the right atrium. In 90 per cent of cases the anomalous pulmonary venous connection is between the right upper or middle pulmonary vein to the superior vena cava or right atrium, usually in association with an atrial septal defect. Ten to fifteen per cent of all atrial septal defects and 80 to 90 per cent of superior vena cava-type sinus venosus atrial septal defects are associated with partial anomalous pulmonary venous connection.

Partial anomalous pulmonary venous drainage may present in adult life with signs of a left to right shunt at atrial level, and the pathophysiological consequences are the same as for an atrial septal defect with an equivalent shunt. When in coexistence with an atrial septal defect, the clinical findings of the two lesions are inseparable. If the atrial septum is intact, physiological splitting of the second heart sound enables anomalous venous drainage to be distinguished from atrial septal defect.

The chest radiograph may reveal the abnormally draining pulmonary vein. Transthoracic echocardiography may be indicative of a shunt at atrial level, but in adults it may not be possible to image the pulmonary veins and a transoesophageal approach is likely to be necessary. The identification of all the pulmonary veins is crucial in assessing the suitability of a secundum atrial septal defect for transcatheter device closure, this technique being contraindicated in the presence of anomalous pulmonary veins (see later).

The indications for surgical repair are the same as those for repair of an atrial septal defect. In the most common variant of right pulmonary venous connection to the superior vena cava in association with a sinus venosus defect, the patch closing the atrial septal defect is placed to direct the anomalous vein into the left atrium.

Scimitar syndrome (Fig. 3)

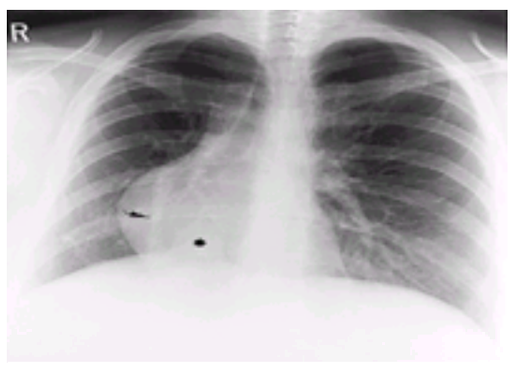

Fig. 3 Chest radiograph of a 25-year-old woman with scimitar syndrome. The heart is shifted into the right hemithorax because the right lung is small. The 'scimitar' shadow (arrow) is produced by the anomalous descending venous channel which drains into the dilated inferior vena cava(*).

Partial anomalous pulmonary venous drainage also occurs as part of the rare familial 'scimitar syndrome' in which part or all of the right pulmonary venous drainage is to the inferior vena cava below the diaphragm. The affected lung lobes are usually hypoplastic and are supplied with arterial blood from the descending aorta. Recurrent infection and bronchiectasis may develop in the hypoplastic lung. Magnetic resonance imaging demonstrates the abnormal arterial supply and venous drainage of the affected lung segment, and may obviate the need for diagnostic cardiac catheterization. Surgical repair may be complicated by difficulty in maintaining perfusion to the affected lung, and lobectomy may by required. In view of this it should be remembered that patients presenting with scimitar syndrome for the first time in adult life have a good prognosis without an operation, similar to that of a small atrial septal defect.

Anomalies of systemic venous drainage

These anomalies frequently form part of a more complex lesion, particularly atrial isomerism.

Superior caval vein anomalies (Fig. 4)

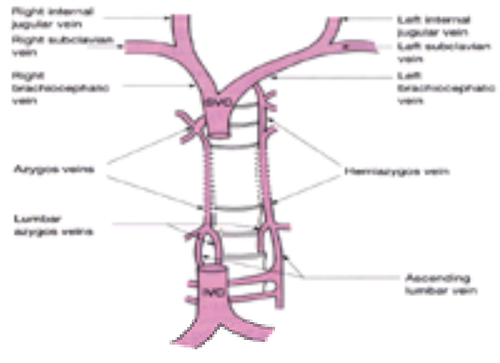

Fig. 4 Schematic diagram of normal systemic venous drainage.

A persistent left-sided superior vena cava occurs in 0.3 per cent of the general population, around 3 per cent of patients with congenital heart disease, and 15 per cent of those with tetralogy of Fallot. The left superior vena cava may be visible on the chest radiograph; it drains to the right atrium via the coronary sinus which is seen to be dilated on two-dimensional echocardiography. A right-sided superior vena cava is usually also present and the two cavae do not usually communicate via the brachiocephalic vein. This common anomaly should be sought routinely at cardiac catheterization; although it does not have any haemodynamic significance, it may cause technical difficulties during transvenous pacemaker insertion and cardiac surgery (Fig. 5).

Fig. 5 Chest radiograph of a 56-year-old man with bicuspid aortic valve, aortic regurgitation, and coarctation. A left superior vena cava draining via the coronary sinus to the right atrium is marked by the path taken by the transvenous pacing leads, inserted for complete heart block.

Other superior vena cava anomalies are the following.

1. An absent right superior vena cava is associated with arrhythmias including atrioventricular block, sinus node dysfunction, and atrial fibrillation.
2. The left, or rarely the right superior vena cava may connect directly to the left atrium, causing an obligatory right to left shunt and cyanosis. It is associated with isomerism of the atrial appendages.

Inferior caval vein anomalies

Azygos continuation of the inferior vena cava occurs in 0.6 per cent of patients with congenital heart disease. The infrahepatic portion of the inferior vena cava is absent and continues to the superior vena cava via an azygos vein; the hepatic veins drain directly into the right atrium. It is often associated with complex lesions, particularly left atrial isomerism. The chest radiograph reveals an absence of the inferior vena cava at the junction of the diaphragm with the right heart border and a dilated azygos vein (Fig. 6). Direct connection of the inferior vena cava to the left atrium is rare; the patient is cyanosed, as in the superior vena cava to left atrium connection.

Fig. 6 Chest radiograph of a 50-year-old man with abdominal situs inversus (*) and laevocardia. Left atrial isomerism is inferred from the symmetrical long bronchi. The inferior vena cava is absent at the level of the diaphragm (small arrow), and the azygos vein receiving inferior caval venous blood is prominent (large arrow).

Atrial arrangement and isomerism of the atrial appendages

Atrial situs solitus is the term used to describe normal atrial arrangement, that is, a right atrium with right-sided morphology and a left atrium with left-sided morphology. Atrial situs inversus is a mirror image arrangement: the right-sided atrium has left morphology and that on the left is a morphological right atrium. The term isomerism refers to abnormal symmetry of paired structures which are normally asymmetrical and have laterality. In right isomerism, both atrial appendages are of right morphology; both are of left morphology in left isomerism. A full description is beyond the scope of this book, but the major features of different atrial arrangements are outlined in Table 7 and illustrated in Fig. 6 and Fig. 7.

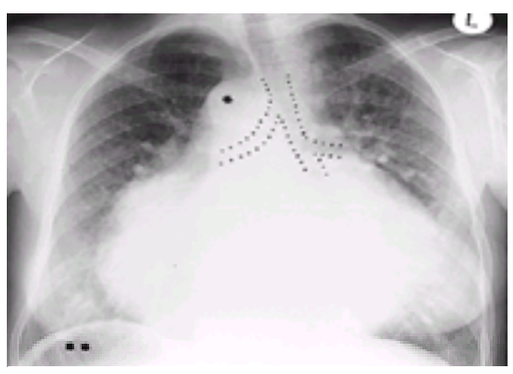

Fig. 7 Chest radiograph of a 21-year-old woman with abdominal situs inversus (**), bronchial and inferred atrial situs inversus, mesocardia, and right aortic arch (*). She has tetralogy of Fallot with pulmonary atresia, palliated with an aortopulmonary shunt via a left thoracotomy.

The key for the physician is to be alerted by the presence of isomerism to the coexistence of complex associated lesions, including a variety of abnormalities of venous connections which may cause technical difficulties at cardiac catheterization and permanent pacemaker insertion. Right isomerism is commoner in males and left in females. Survival to adulthood with right isomerism is uncommon because of associated asplenia and severe cyanotic heart disease, including obstructed

anomalous pulmonary venous drainage (the pulmonary venous confluence is a left atrial structure), complete transposition of the great arteries, atrial septal defect, atrioventricular septal defect, absent coronary sinus (a left atrial structure), severe pulmonary stenosis or atresia, and univentricular heart. The lesions associated with left isomerism tend to produce left to right shunts and little if any cyanosis. They include atrioventricular block, atrioventricular septal defect, common atrium, and left ventricular outflow tract obstruction.

Atrial septal defects

Atrial septal defects account for approximately 10 per cent of congenital heart disease. The exact figure depends on the definition of atrial septal defect, since small or probe-patent foramen ovale occurs in around 10 per cent of the population. The sites of the various types of atrial septal defect are shown in Fig. 8.

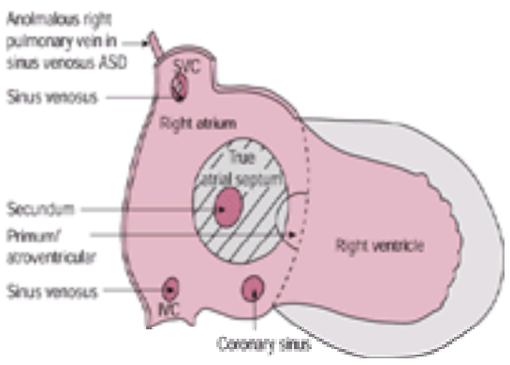

Fig. 8 Sites of atrial septal defects. The shaded area delineates the true atrial septum. Sinus venosus and coronary sinus defects are therefore not strictly atrial septal defects although they permit shunting at atrial level.

Ostium secundum atrial septal defect

Secundum atrial septal defect accounts for 40 per cent of left to rights shunts in adults over 40 years of age. It is commoner in females, with a sex ratio of 2:1, and may be familial. Atrial septal defect may be an incidental finding in a geriatric patient at autopsy and diagnosis in life may be delayed well into adulthood because of the absence of symptoms and subtlety of clinical signs. However, the natural history of this lesion is not benign, only 50 per cent of those with non-restrictive atrial septal defects surviving without operation to the age of 40 years, and 10 per cent beyond 60 years of age.

Presentation in adulthood may be with symptoms or as a result of incidental clinical or radiographic findings. Patients over the age of 30 years with unrepaired atrial septal defect commonly develop paroxysmal and eventually chronic atrial arrhythmia, but also flutter and ectopic atrial tachycardia. Most patients over the age of 60 years are symptomatic with exertional dyspnoea and palpitation. A left to right shunt at atrial level predisposes to paradoxical embolus since simple manoeuvres such as the Valsalva are sufficient to increase right atrial pressure and reverse the shunt. Patients with unoperated atrial septal defect are therefore at risk of embolic stroke, and should not dive because of the risk of paradoxical gas embolism. An age-related reduction in left ventricular compliance augments the left to right shunt and is one of the causes for the progression of symptoms with age. In addition, modest pulmonary arterial hypertension increases with age so the right ventricle is exposed to pressure as well as volume overload and may eventually fail.

Clinical signs

If the defect is non-restrictive, the a and v waves of the jugular venous pulse tend to be equal. In older patients with reduced left ventricular compliance, the left and therefore right atrial pressure is raised, reflected in an elevated jugular venous pressure. A right ventricular heave may be felt at the left sternal border and the dilated pulmonary artery may be palpable in the left second intercostal space. The first sound is loud due to increased diastolic flow across the tricuspid valve. The second heart sound is widely split and fixed, and there is loss of normal sinus arrhythmia if the left to right shunt is equal to 2:1 or greater. There may be a pulmonary flow murmur at the upper left sternal edge. Only if the atrial septal defect has a high gradient across it will it generate a murmur itself, usually a soft continuous murmur. This is the case if the defect is small and restrictive and the left atrial pressure high, for example if there is associated mitral stenosis. If the patient has pulmonary vascular disease, the signs will be the same as for pulmonary hypertension with right to left shunt (see above).

Associations

Acquired disease may coexist and interact with congenital heart disease, especially in the ageing patient. Left ventricular dysfunction due to coronary artery disease and systemic hypertension may increase the left to right interatrial shunt, resulting in a more rapid clinical deterioration than would be expected. Similarly, mitral regurgitation increases the effective interatrial shunt and mitral valve abnormalities may be acquired secondary to the effects of a secundum atrial septal defect. There may be distortion of the anterior mitral valve leaflet with fibrotic shortened chordae due to the abnormal position of the interventricular septum as a result of chronic right ventricular overload. Concomitant pulmonary stenosis may be overestimated in the presence of an atrial septal defect, since Doppler velocities are increased in the presence of a left to right shunt.

Lutembacher's syndrome is the association of mitral stenosis with secundum atrial septal defect. The presence of mitral stenosis increases the left to right shunt at atrial level, causing an overestimation of the significance of the atrial septal defect and an underestimation of the severity of mitral stenosis. Repair of the atrial septal defect alone may unmask severe mitral stenosis.

Pulmonary vascular disease and atrial septal defect

Only around 10 per cent of atrial septal defects develop a right to left shunt secondary to pulmonary vascular disease, and a causal relationship between atrial septal defect and the Eisenmenger reaction remains controversial. In atrial septal defect, unlike other lesions which may cause the Eisenmenger reaction such as large ventricular septal defect, the pulmonary vasculature is not exposed to increased flow at systemic pressure.

Atrial septal defect with a right to left shunt due to pulmonary vascular disease and pulmonary hypertension occurs most commonly in young women and in some cases may be due to primary pulmonary hypertension with an incidental atrial septal defect. In this combination, the prognosis may be better than for primary pulmonary hypertension with intact atrial septum, the septal defect protecting the right heart from pressure overload by allowing right to left shunting. Persistence of the fetal pulmonary vascular pattern may be implicated in the development of pulmonary hypertension in some young patients with atrial septal defect. Patients living or born at high altitude have a higher incidence of pulmonary vascular disease. In older patients with atrial septal defect there may be a relationship with *in situ* pulmonary arterial thrombosis and the development of pulmonary hypertension.

Investigations

The electrocardiogram may show sinus node dysfunction and, less commonly, prolongation of the P–R interval. The QRS axis is usually vertical. The QRS complex may be prolonged with rSr' in lead V1: this does not represent incomplete right bundle branch block, but occurs because the last part of the myocardium to depolarize is the right ventricular outflow tract, which is enlarged and thickened due to volume overload. The sinus node may be damaged when the superior vena cava is cannulated during surgery: occasionally perioperative sinus bradycardia and sinus pauses persist and a permanent pacemaker is required. The P–R interval may return to normal as right atrial size decreases. Macro-reentry circuits at the site of atrial surgery may result in postoperative ectopic atrial tachycardias.

The typical chest radiograph shows dilated proximal pulmonary arteries with a small aortic knuckle, plethoric lung fields, and cardiomegaly secondary to dilatation of the right atrium and ventricle.

Transthoracic echocardiography demonstrates the volume overloaded right atrium and ventricle. The size of the shunt can be estimated and colour flow Doppler

facilitates the detection of the site of the shunt. If transcatheter device closure is considered, a transoesophageal approach is necessary to define the site and size of the atrial septal defect precisely and to identify the pulmonary veins.

Cardiac catheterization is only indicated to calculate pulmonary vascular resistance if there is a suspicion of pulmonary hypertension, or to exclude coexisting congenital or acquired cardiac pathology such as coronary artery disease.

Indications for closure of atrial septal defect

Surgical repair carries low mortality and morbidity. However, vigilance is required to detect postoperative pericardial effusions, which appear to be more common than following other operations. Closure of an atrial septal defect is indicated if there is exertional dyspnoea, if the left to right shunt is greater than 1.5:1, if there is right heart volume overload, or in order to prevent recurrent paradoxical embolism. Repair of an isolated secundum atrial septal defect by the third decade results in a normal life expectancy, between the ages of 25 and 41 years it results in a good but shorter than normal life expectancy, and beyond the age of 41 years, morbidity and mortality remain significantly higher than normal. None the less, functional status and longevity are improved following repair over the age of 40 years, 5- and 10-year survival being estimated as 98 and 95 per cent, respectively, for patients who underwent repair, and 93 and 84 per cent for those treated medically. Surgical repair in older patients does not reduce the risk of late atrial arrhythmia, particularly if there is right ventricular dysfunction, elevated pulmonary artery pressure, or pre-existing atrial arrhythmia. Whether the incorporation of a modified maze procedure or cryoablation into the surgical repair of atrial septal defect will reduce the long-term incidence of existing or *de novo* atrial arrhythmia remains to be determined.

Secundum atrial septal defects up to 3 cm in stretched diameter may be closed by transcatheter devices as long as the surrounding rim of atrial septal tissue is sufficient. Criteria for device closure of secundum atrial septal defect are: size less than 3 cm; a situation away from the atrioventricular valves, pulmonary and caval veins; and normal pulmonary venous drainage. Following closure, antiplatelet or anticoagulant therapy is recommended for 3 months. Device closure is the procedure of choice for patent foramen ovale complicated by paradoxical embolism.

Sinus venosus defect

Sinus venosus defects account for 2 to 3 per cent of atrial septal defects and have an equal sex incidence.

They are not truly defects of the atrial septum, but since they allow shunting at atrial level, they are included in the classification of atrial septal defects. The inferior border of the more common superior vena cava type of sinus venosus defect is made by the superior limbus of the fossa ovalis, and the upper border comprises the junction of the superior vena cava with the atrial mass. The superior caval vein overrides the atrial septum, connecting to both atria, and the right upper pulmonary vein usually drains anomalously into the superior vena cava. There may be an ectopic atrial pacemaker because the defect is located in the area of the sinoatrial node. This may be reflected by a leftwards p-wave axis and an inverted p wave in lead III.

Coronary sinus defect

The rarest form of atrial septal defect, this defect is at the site of entry of the coronary sinus to the right atrium. The unroofed coronary sinus is a variation of coronary sinus defect in which the partition between the coronary sinus and the left atrium is absent as the coronary sinus runs posteriorly along the floor of the left atrium. In this condition, a left superior vena cava commonly connects directly to the left atrium, producing a right to left shunt and cyanosis.

Atrioventricular septal defect

The sites of ostium primum and atrioventricular septal defects are shown in Fig. 8, and the leaflet patterns in atrioventricular septal defects in Fig. 9.

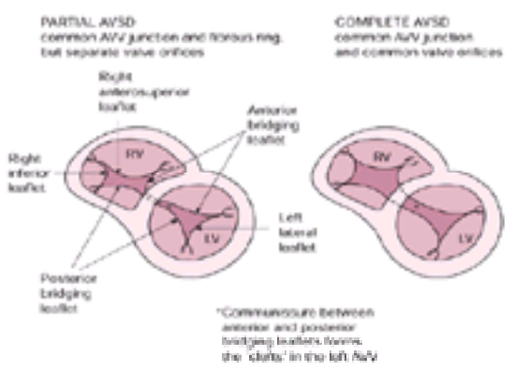

Fig. 9 Leaflet patterns in atrioventricular septal defect. Diagrams are in short axis, as imaged by two-dimensional echocardiography. (Modified from Anderson RH *et al*., 1983, *Morphology of congenital heart disease*. University Park Press, Baltimore.)

Ostium primum defect describes the atrial component of the atrioventricular septal defect, previously termed endocardial cushion defect or atrioventricular canal. The atrioventricular septum is absent and the atrioventricular valves share a common junction and fibrous ring, with a five leaflet atrioventricular valve. Since they share common leaflets, the valves are not correctly called mitral and tricuspid valves, but left and right atrioventricular valves. As a consequence the normal offsetting of the right atrioventricular valve towards the right ventricular apex is absent. In addition, the aorta is 'unwedged' from its normal position between the left and right atrioventricular valves, with loss of the normal fibrous continuity between the 'mitral' and aortic valves. The left ventricular outflow tract is therefore elongated ('gooseneck') and has the propensity to develop obstruction. 'Cleft mitral valve' refers to the commissure between the anterior and posterior bridging leaflets which renders the left atrioventricular valve potentially regurgitant. The left ventricular papillary muscles are abnormally placed anteriorly and posteriorly instead of in the normal anterolateral and posteromedial positions. A partial atrioventricular septal defect has a common atrioventricular junction, but the right and left atrioventricular valves have separate orifices and the ventricular component of the defect is usually small or absent. There are both a common atrioventricular junction and a common valve orifice in a complete atrioventricular septal defect, and the ventricular component of the defect is usually large.

Atrioventricular septal defect occurs with equal sex incidence. The complete form of the defect is most commonly associated with Down's syndrome. A single gene defect may be responsible for atrioventricular septal defect with normal chromosomes; the recurrence risk is high: around 10 per cent if the mother has an atrioventricular septal defect, less if the father is affected.

The physiological consequences of an atrioventricular septal defect are the same as for other conditions with left to right shunting at atrial or ventricular level, but may be complicated by left atrioventricular valve regurgitation or left ventricular outflow tract obstruction. If the ventricular septal defect is large and non-restrictive, pulmonary vascular disease may develop. Patients with Down's syndrome are at particular risk of this complication; coexisting upper airway obstruction and sleep apnoea, and abnormal pulmonary parenchyma may be contributory factors.

Investigations

The electrocardiogram is distinctive, with a left and superior QRS axis and notching of S waves in the inferior leads. The chest radiograph appearances depend on the degree of interatrial shunting and left atrioventricular valve regurgitation; the former producing cardiomegaly due to left heart dilatation and the latter, left atrial enlargement. There may be increased pulmonary vascularity, particularly in young patients with complete atrioventricular septal defect and high pulmonary blood flow. Transthoracic echocardiography reveals the detailed anatomy of the defect and establishes the site and degree of shunting, the presence of left ventricular outflow tract obstruction, and the function and anatomy of the atrioventricular valves, including the classic'fish mouth' deformity of the left atrioventricular valve during diastole. The indications for cardiac catheterization are the same as for secundum atrial septal defect, namely to exclude inoperable pulmonary vascular disease. In addition, useful information may be obtained regarding the severity of left atrioventricular valve regurgitation and left ventricular outflow tract obstruction.

Clinical course

First presentation may occur in adulthood if the left to right shunt is small and the left atrioventricular valve is competent. Physical signs are the same as in other atrial septal defects: there may also be an apical pansystolic murmur. Paradoxical embolism is less common than in secundum atrial septal defect, since the position of the primum defect low in the interatrial septum avoids the streaming of inferior vena cava blood which is most likely to carry emboli and is directed towards the mid-portion of the septum.

Most adult patients have undergone surgery to repair the defect and left atrioventricular valve. Others have survived without an operation and may have developed pulmonary vascular disease.

Whether or not it has previously been repaired, the abnormal left atrioventricular valve may become regurgitant in later life, particularly in response to changes in the left ventricle due to ageing, ischaemia, or systemic hypertension. The risk of endocarditis relates largely to the abnormal left atrioventricular valve. Atrial arrhythmias occur in the same way as in secundum atrial septal defect, however there is a higher incidence of postoperative atrioventricular block in atrioventricular septal defect because of the proximity of the atrioventricular node to the site of repair.

Lesions affecting ventricular inflow

Cor triatatrium

This is a very rare defect in which one of the atria (nearly always the left) is partitioned by a fibromuscular membrane into an upper chamber that receives the pulmonary veins, and a lower chamber connecting with the atrial appendage and mitral valve. The membrane usually inserts into the atrial septum at the fossa ovalis, where an atrial septal defect coexists in around 50 per cent of cases, allowing communication between the right and left atria. The membrane may be intact, or pierced by one or more holes that are usually restrictive, causing supramitral stenosis. First presentation in adulthood is unusual unless the membrane is non-restrictive or coexists with a large atrial septal defect. Patients may have signs of an atrial septal defect or mitral stenosis. The diagnosis is made by echocardiography. The chest radiograph is also characteristic, showing signs of pulmonary venous congestion, but not the left atrial appendage enlargement that accompanies valvar mitral stenosis, since the appendage lies in the low pressure atrial chamber. The lateral chest radiograph may show enlargement of the pulmonary venous compartment of the left atrium. Treatment is surgical resection and the postoperative prognosis is good.

Congenital mitral valve anomalies

These anomalies are rare and frequently coexist with other lesions. A supramitral ring often coexists with congenital mitral stenosis. It differs from cor triatatrium in that the ring is sited inferiorly to the os of the appendage and lies immediately above the mitral valve.

Shone syndrome comprises four levels of left heart obstruction: supramitral ring, parachute mitral valve, subaortic stenosis, and coarctation of the aorta. Parachute mitral valve exists when the two papillary muscles are fused or there is hypoplasia or absence of one papillary muscle; the valve and its apparatus are often additionally dysplastic. Obstruction occurs at the level of the abnormal papillary muscles. The parachute mitral valve may also be regurgitant if the chordae are elongated and not significantly fused.

Isolated cleft mitral valve differs from the 'cleft' seen in an atrioventricular septal defect in being in the anterior (aortic) leaflet, directed towards the aortic outflow tract, rather than being in the space between the bridging leaflets and pointing towards the septum. The isolated cleft can be readily repaired to resemble a competent normal mitral valve.

Ebstein's anomaly of the tricuspid valve

This rare condition occurs in 1 per 20 000 live-born babies and affects both sexes equally. The risk may be increased by maternal exposure to lithium during the first trimester. In the normal heart, the mitral and tricuspid valves are offset so that the tricuspid valve is displaced up to 1.5 cm towards the right ventricular apex. In Ebstein's anomaly, the anterior leaflet usually inserts normally at the atrioventricular junction, but the attachments of the septal and sometimes mural (posterior) leaflets are apically displaced, causing atrialization of the proximal part of the right ventricle and reducing the size of the functional right ventricle. In addition the movement of the septal and mural leaflets is usually limited either by being thickened and fibrotic, or by being tethered by short chordae to the septum. The major haemodynamic effect is usually tricuspid regurgitation, but occasionally a muscular shelf or fused anteromedial commissure causes tricuspid stenosis.

Associated abnormalities

A patent foramen ovale or atrial septal defect is present in most cases. Left heart abnormalities occur as a consequence of alterations in left ventricular geometry due to leftwards displacement of the interventricular septum: for example, mitral valve prolapse may occur as a result of relatively long chordae in a left ventricle of reduced cavity size. Twenty per cent of patients have coexistent Wolff–Parkinson–White syndrome, usually with a right-sided or multiple pathway(s).

Clinical presentation and course

There is a broad spectrum of severity, ranging from intrauterine death to presentation in late adulthood. Mortality is influenced by age at presentation, the condition of the tricuspid valve, the cardiac rhythm, and the functional capacity of the right ventricle, including the severity of right ventricular outflow tract obstruction and the size of the right atrium in relation to the other cardiac chambers.

Cyanosis may develop in adulthood if there is an associated atrial septal defect or patent foramen ovale; as the right ventricular filling pressure increases there is a parallel rise in right atrial pressure, and a right to left interatrial shunt is established. These patients are at risk of paradoxical embolism. The risk of endocarditis is low, because the tricuspid regurgitant jet is of low velocity.

Heart failure may intervene as a result of the combination of severe tricuspid regurgitation and the onset of atrial fibrillation or flutter. These atrial arrhythmias may be particularly troublesome if a coexistent accessory pathway allows a rapid ventricular response rate. The onset of atrial fibrillation is a predictor of death within 5 years, and may account for the increased death rate in the fifth decade.

Physical signs

The patient may be acyanotic or cyanosed and clubbed. Even when tricuspid regurgitation is severe the jugular venous pressure may not be particularly high, nor the v wave prominent, because of the capacity of the right atrium and thin-walled atrialized right ventricle to accommodate the low pressure regurgitant volume. Once right ventricular failure develops the jugular venous pressure rises further and the a and v waves become more prominent. In the uncommon situation of tricuspid stenosis, the a wave is increased and may be giant. The first heart sound is widely split with a delayed tricuspid component, due to the extra distance that the large anterior leaflet has to travel to reach the limit of its systolic excursion. The second heart sound may be single because low pressure in the right ventricular outflow tract renders the pulmonary component inaudible, or it may be widely split reflecting right bundle branch block. The systolic murmur of tricuspid regurgitation varies from inaudible to loud enough to generate a thrill, but is classically decrescendo and scratchy.

Investigations

The chest radiograph is characteristic (Fig. 10). The electrocardiogram typically shows a superior axis and right atrial enlargement, with or without right bundle branch block. The p wave may be peaked and the P–R interval prolonged, reflecting the prolonged conduction in the large right atrium, or there may be evidence of pre-excitation. Right bundle branch block may occur due to abnormal activation and conduction in the atrialized right ventricle.

Fig. 10 Chest radiograph of a 43-year-old woman with classic cardiac silhouette of Ebstein's anomaly due to right atrial enlargement. The aortic knuckle and pulmonary arteries are inconspicuous and the lung fields oligaemic.

Two-dimensional echocardiography with colour flow Doppler establishes the diagnosis and severity of Ebstein's anomaly. The atrialized and functional portions of the right ventricle can be identified, as can the precise attachments and degree of tethering of the anterior leaflet of the tricuspid valve. Echocardiography is the investigation of choice in planning surgical intervention, tethering and restricted motion of the anterior leaflet and a small right ventricle being strong predictors of the need for tricuspid valve replacement rather than repair. Cardiac catheterization is only necessary if specific haemodynamic questions remain after non-invasive assessment.

Treatment

Once atrial arrhythmias develop, patients should be anticoagulated, particularly if there is an atrial septal defect. If re-entry tachycardias cannot be controlled with antiarrhythmic drugs, radiofrequency ablation of accessory pathways may be performed. However, ablation may be made difficult by the size and abnormal shape of the right atrium and abnormal position of the accessory pathway or pathways.

Surgery is indicated when the patient's clinical status deteriorates and aims to reduce the size of the right atrium and repair the tricuspid valve so that valve function and right ventricular geometry are improved. The best haemodynamic results are achieved if the valve can be repaired rather than replaced. The addition of a right-sided maze procedure may reduce the long-term risk of developing atrial fibrillation. In a subset of patients with a small right ventricle, the addition of a bidirectional Glenn anastomosis to offload the ventricle may facilitate repair.

Other right ventricular anomalies

Uhl's anomaly and arrhythmogenic right ventricular dysplasia (right ventricular cardiomyopathy) are rare sporadic or familial conditions affecting the right ventricle. Table 8 lists the key distinguishing features.

Ventricular septal defect

With the exceptions of bicuspid aortic valve and mitral valve prolapse, ventricular septal defect is the commonest congenital cardiac malformation, occurring in around 3 per 1000 live-born babies. Defects may exist in isolation, in association with other lesions such as coarctation of the aorta, or as an integral part of lesions such as tetralogy of Fallot. This section deals with isolated ventricular septal defects.

Morphology and classification

The ventricular septum is mostly muscular, with a small fibrous membranous portion. Ventricular septal defects are classified according to their borders, seen from the right ventricular aspect. There are three types: muscular, perimembranous, and doubly committed subarterial (Table 9). The position of muscular and perimembranous ventricular septal defects may be inlet, trabecular, or outlet, depending on which part of the right ventricle they open into.

Perimembranous ventricular septal defect is the commonest type of defect, only 5 to 7 per cent of ventricular septal defects in Europe and North America are doubly committed subarterial defects, whereas they account for up to 30 per cent of defects in Asian patients. Outlet perimembranous ventricular septal defects usually occur due to a malalignment between the outlet and trabecular septa with overriding of either the aortic or pulmonary valve. If the malalignment of the outlet septum is towards the right ventricle, the aorta overrides the defect and tends to cause subpulmonary obstruction; this is the type of defect typical of tetralogy of Fallot. Malalignment towards the left ventricle may cause subaortic obstruction and may be associated with hypoplasia of the aortic arch.

Clinical presentation and complications of unoperated ventricular septal defect

The grades of unoperated ventricular septal defect are shown in Table 10. Adults with unoperated restrictive ventricular septal defects are usually asymptomatic. There is a high risk of endocarditis in small ventricular septal defects due to the high velocity jet from left to right ventricle, particularly if the jet is directed on to the tricuspid valve. There is a small increased incidence of sudden death and ventricular tachycardia in unoperated small ventricular septal defect, but longevity is otherwise normal. The adult with an isolated unoperated restrictive ventricular septal defect is acyanotic with normal arterial and jugular venous pulses. There may be a thrill at the left sternal border, the left ventricular apex may be thrusting if the defect is large enough to cause volume overload, and a dilated pulmonary artery may be palpable. The second heart sound is usually normally split. There is a loud harsh pansystolic murmur at the left sternal edge; the murmur being softer and shorter (early systolic) in very small defects.

Larger ventricular septal defects rarely present for repair in adulthood since the large left to right shunt is unlikely to allow unoperated survival unless pulmonary vascular disease developed. Non-restrictive defects are not associated with the classic ventricular septal defect murmur since left and right ventricular pressures are equal.

Investigations

Investigation should determine the type and number of ventricular septal defects, the size of the defect (restrictive or non-restrictive), estimate the size of the shunt (Qp:Qs), pulmonary artery pressure and resistance, and assess left and right ventricular function, and volume and pressure overload. Associated lesions that may alter management should be identified, especially aortic regurgitation, subaortic stenosis, and right ventricular outflow tract obstruction.

The chest radiograph is normal if the defect has been small from birth. If the ventricular septal defect is, or has been, larger, the left ventricle, left atrium, and pulmonary trunk may be dilated and there may be increased pulmonary vascularity. The electrocardiogram shows a normal QRS axis unless there are multiple defects, in which case there may be left axis deviation. In the presence of a large left to right shunt the p wave may be broad and there may be evidence of left ventricular hypertrophy. Two-dimensional echocardiography identifies the number and site of defects as well as describing the morphology and associated defects. Doppler is used to estimate the size and direction of the shunt, and right ventricle to left ventricle pressure difference, but this may not be accurate if there is an obliquely lying muscular ventricular septal defect. Cardiac catheterization is important to measure the size of shunt and pulmonary vascular resistance with reversibility studies if baseline resistance is high.

Indications for repair and postoperative sequelae

Repair of a ventricular septal defect is indicated in the presence of symptoms, if Qp:Qs is greater than 2:1, or if there is ventricular dysfunction with right ventricular pressure overload or left ventricular volume overload. Repair should also be undertaken if there are coexisting lesions such as significant right ventricular outflow tract obstruction, more than mild aortic regurgitation, or aortic valve prolapse in the presence of an outlet ventricular septal defect. A second episode of endocarditis may

also be considered as an indication for ventricular septal defect closure. If the pulmonary artery pressure is greater than two-thirds systemic pressure, repair should only be considered if Qp:Qs is greater than 1.5:1 or if there is evidence of reversibility in response to pulmonary vasodilators such as oxygen and nitric oxide.

The surgical approach to repair aims at avoiding damage to important structures such as the conducting tissues that are especially vulnerable in perimembranous defects. Transatrial repair reduces the risk of postoperative ventricular arrhythmias by avoiding a right ventriculotomy. Transient postoperative complete heart block is associated with an increased risk of late high-degree block. Permanent pacemaker implantation is indicated in the 1 to 2 per cent of patients in whom complete heart block persists, even if they are asymptomatic, because of the significant risk of sudden death. The prognosis after ventricular septal defect repair in the early years of life is good. However, left ventricular dilatation may persist and systolic function be impaired if repair is delayed into late childhood. Long-term postoperative survival depends on the presence of pulmonary hypertension, left ventricular dysfunction, and complications such as aortic regurgitation and endocarditis.

Transcatheter device closure of ventricular septal defects is possible providing that valvar apparatus can be avoided. The approach is most suited to muscular ventricular septal defects, especially multiple defects that are difficult to close surgically and it may be combined with a surgical approach for inaccessible muscular defects.

Double-outlet right ventricle

In double-outlet right ventricle more than half the circumference of both great vessels arises from the morphological right ventricle. A complete or partial muscular infundibulum lies beneath each arterial valve. This definition includes variants of tetralogy of Fallot in which the aorta overrides the ventricular septal defect by more than 50 per cent.

The degree of pulmonary stenosis and the relation of the ventricular septal defect to the great vessels determine the haemodynamics. Eighty per cent of subaortic defects have pulmonary stenosis and Fallot-like physiology.

The Taussig–Bing anomaly accounts for less than 10 per cent of double-outlet right ventricle and describes a subpulmonary defect without pulmonary stenosis. There is transposition-like physiology with cyanosis and high pulmonary blood flow. As the pulmonary vascular resistance rises, pulmonary blood flow falls and cyanosis increases. Unoperated survival to adulthood is uncommon but occurs occasionally if the pulmonary vascular resistance establishes adequate, but not excessive, pulmonary blood flow. If such a survivor also has a patent arterial duct, there will be reversed differential cyanosis. Deoxygenated blood selectively enters the aorta to supply the arch vessels, whereas oxygenated blood enters the pulmonary artery and supplies the descending aorta via the duct: thus the fingers are more cyanosed and clubbed than the toes.

Tetralogy of Fallot

Tetralogy of Fallot is the commonest cyanotic defect, occurring in 1 per 3600 live-born babies. The four hallmarks of tetralogy of Fallot are:

- subvalvar pulmonary stenosis
- ventricular septal defect
- aortic valve overrides the ventricular septal defect
- right ventricular hypertrophy.

The fundamental abnormality is anterocephalad deviation of the outlet septum that both creates the subpulmonary stenosis and accounts for the aortic valve overriding the muscular septum. There is great anatomical variation, ranging from minimal aortic override to double-outlet right ventricle, and from minimal pulmonary stenosis to pulmonary atresia. The ventricular septal defect is perimembranous and there is usually additional pulmonary valvar stenosis.

Defects associated with tetralogy of Fallot include a right-sided aortic arch in 16 per cent, a left superior vena cava in around 15 per cent, additional ventricular septal defects in 5 per cent, and a secundum atrial septal defect ('pentalogy' of Fallot) in 8 per cent. The most important associated coronary anomaly is the crossing of the right ventricular outflow tract by a left anterior descending coronary artery arising anomalously from the right coronary sinus and vulnerable to damage during surgical repair via a right ventricular approach.

Unoperated natural history and management

Without surgical intervention, only 2 per cent of patients survive to their 40th year. Those that do survive may represent a select group in whom subpulmonary stenosis was not severe in early life, but progressed with advancing age. Such patients may rarely live into their eighth decade; one of our patients survived 77 years. Unoperated patients are at risk of the complications of cyanosis, endocarditis, atrial and ventricular arrhythmias, progressive ascending aortic dilatation, and aortic regurgitation which causes volume overload of both ventricles and subsequent biventricular failure. Systemic hypertension adds additional pressure overload to the work of both ventricles and further contributes to the onset of biventricular failure.

There is cyanosis and clubbing, a right ventricular heave, and sometimes a thrill over the right ventricular outflow tract. A right-sided aorta may be palpable to the right of the sternum. The second heart sound is usually single, and there is a loud pulmonary ejection murmur. There may be aortic regurgitation.

The electrocardiogram shows right axis deviation, right ventricular hypertrophy, and the QRS duration may be prolonged in older patients. The classic cardiac silhouette is a 'coeur en sabot', that is, a clog-shaped heart, but it is more likely to be seen in tetralogy with pulmonary atresia (see below). The heart size is usually normal and pulmonary vascularity reduced. There may be a right-sided aortic arch indenting the right of the trachea and also a prominent dilated ascending aorta. Two-dimensional echocardiography reveals infundibular stenosis with or without pulmonary valve stenosis, right ventricular hypertrophy, the typical ventricular septal defect, and varying degrees of aortic override. There may be evidence of left ventricular volume overload, aortic root dilatation, and aortic regurgitation. Cardiac catheterization should be performed prior to radical repair in adults. The anatomy of the right ventricular outflow tract obstruction (Fig. 11) and pulmonary arteries is defined, and pulmonary vascular resistance assessed. Selective coronary angiography demonstrates any anomalous origin and course as well as acquired coronary disease. Aortography shows aortic root dilatation and any aortopulmonary collaterals.

Fig. 11 Right ventricular angiogram (lateral projection) of a 45-year-old woman with unoperated tetralogy of Fallot. The right ventricle is entered via the aorta overriding the ventricular septal defect. There is severe muscular infundibular stenosis (small arrows), the pulmonary valve is thickened and doming (large arrow), and there is right ventricular hypertrophy.

Palliated history

Helen Taussig first suggested palliative surgery in 1943, and the first Blalock–Taussig shunt was performed in 1945 (Fig. 12 and Table 11). Nowadays, palliative shunts are usually performed as a staging procedure in small infants; however, occasional patients reach the adult clinic having had palliation without subsequent

radical repair. They are cyanosed and clubbed and have a continuous murmur under the clavicle and over the scapula on the side of the shunt. In a classic Blalock–Taussig shunt, the ipsilateral radial pulse is diminished or absent and the hand often small. Late complications of systemic to pulmonary artery shunts include infective endarteritis, acquired pulmonary atresia, aortic regurgitation, biventricular failure, increasing cyanosis, bronchopulmonary collateral development if the shunt blocks or is outgrown, and pulmonary vascular disease if the shunt is too big.

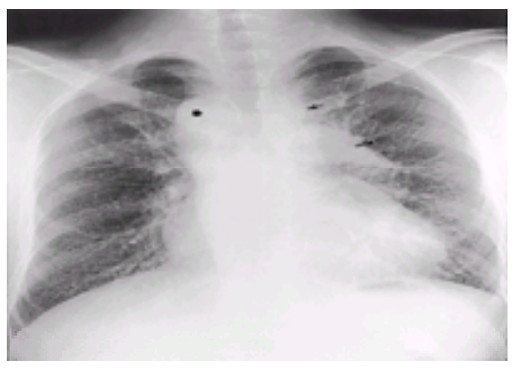

Fig. 12 Chest radiograph of a 36-year-old man with tetralogy of Fallot palliated by a classic left Blalock–Taussig shunt (small arrow). There is secondary dilatation of the left pulmonary artery (large arrow) and a right aortic arch (*).

Follow-up after radical repair

Radical repair involves patch closure of the ventricular septal defect with infundibular resection with or without pulmonary valvotomy or replacement. Eighty-six per cent of patients who undergo radical repair survive to 32 years of age, and these represent the majority of patients with tetralogy seen in the adult clinic. However, they remain at risk of late complications including pulmonary regurgitation and stenosis, arrhythmia, sudden death, endocarditis, and aortic regurgitation. Those repaired beyond late childhood have a higher morbidity and mortality than those repaired by the age of 12 years.

Free pulmonary regurgitation may be present since surgery if the valve was removed, or become progressively more severe, particularly if a monocusp valve or transannular patch was placed. Although well tolerated for many years, pulmonary regurgitation may result in progressive right ventricular dilatation and increased risk of atrial and ventricular arrhythmias. A progressive prolongation of the QRS duration may reflect these changes and a QRS greater than 180 ms is a marker for patients at particular risk of ventricular arrhythmia. The timing of replacement of a regurgitant pulmonary valve remains difficult, but is indicated if there is dyspnoea, palpitation, and progressive right ventricular enlargement or dysfunction. Pulmonary regurgitation is worsened in the presence of pulmonary arterial stenosis that can occur at the site of a previous shunt. Right ventricular outflow tract obstruction may recur, especially if a valved right ventricular to pulmonary artery conduit was placed; this may be due to excessive formation of neointima (peel) in the conduit or to calcification of the valve. Valve calcification is readily seen on the lateral radiograph and can be followed as it encroaches on to the valve cusps to cause stenosis (Fig. 13).

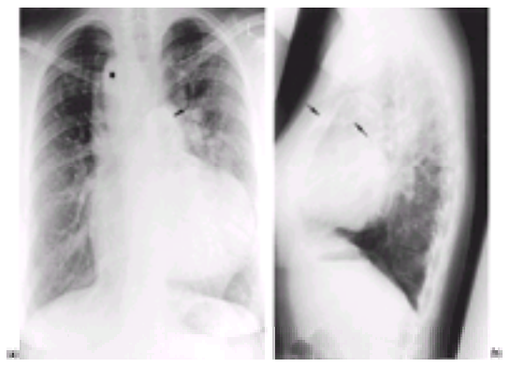

Fig. 13 Chest radiographs, (a) posteroanterior and (b) lateral, of a 30-year-old woman with tetralogy of Fallot and pulmonary atresia who underwent repair with a valved homograft conduit from right ventricle to pulmonary artery and ventricular septal defect closure 10 years previously. There is a right aortic arch (*) and a 'coeur en sabot' cardiac silhouette. The calcification in the homograft (arrows) is more clearly seen on the lateral radiograph. The abnormal pulmonary vasculature reflects persisting aortopulmonary collaterals.

The majority of patients have right bundle branch block after repair (Fig. 14). Bifasicular block and transient postoperative complete heart block carry a risk of developing late complete heart block. Atrial arrhythmias occur in 30 per cent of long-term survivors and are a major cause of morbidity. Those with left-sided volume overload and left atrial dilatation secondary to residual ventricular septal defect or previous shunts are at particular risk of atrial flutter and fibrillation. Rapidly conducted atrial flutter is particularly poorly tolerated and is likely to be responsible for a proportion of sudden deaths. Ventricular arrhythmias occur in up to 45 per cent of patients. However, the incidence of late sudden death is only 1 to 5 per cent, so not all patients with ventricular arrhythmias are at risk. Sustained monomorphic ventricular tachycardia is likely to be a significant risk factor for sudden death, as are atrial arrhythmias and heart block. Right ventricular risk factors may include right ventricular dilatation, outflow tract obstruction, hypertrophy, aneurysm, impaired myocardial blood flow, and pulmonary regurgitation. Surgical risk factors for late sudden death include transventricular as opposed to transatrial repair, large ventriculotomy scar, residual ventricular septal defect, previous complex or multiple operations, impaired left ventricular function, older age at operation, and length of follow-up.

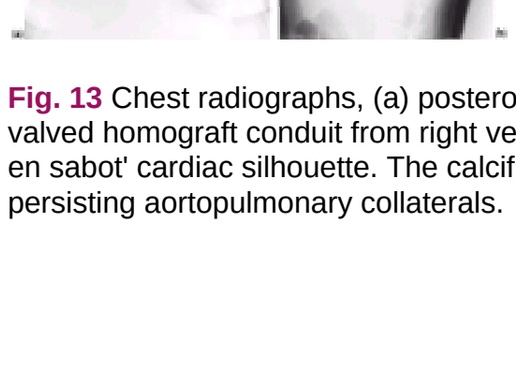

Fig. 14 Electrocardiograms of a 35-year-old woman who underwent radical repair of tetralogy of Fallot. Preoperatively (a) there is right ventricular hypertrophy, postoperatively (b) there is right bundle branch block, due to damage to the right bundle as it runs in the floor of the ventricular septal defect.

Tetralogy of Fallot with pulmonary atresia (pulmonary atresia with ventricular septal defect)

This condition represents the extreme end of the spectrum of tetralogy. There is considerable anatomical variation, including acquired pulmonary atresia with well-developed confluent pulmonary arteries, hypoplastic confluent pulmonary arteries that may not supply all segments of the lungs, non-confluent pulmonary

arteries, and complete absence of central pulmonary arteries. The right ventricular outflow tract is blind-ended and the pulmonary blood supply is derived entirely from three types of systemic vessels: a large muscular duct that resembles a collateral, a diffuse plexus of small 'bronchial' arteries arising from mediastinal and intercostal arteries, or from large tortuous systemic arterial collaterals. These large collaterals arise directly from the descending aorta, from its major branches (usually the subclavian artery), or from bronchial arteries. They may connect with central pulmonary arteries or supply whole segments or lobes of lung independently. This variation has also been termed complex pulmonary atresia.

Examination findings are similar to those of unoperated Fallot without pulmonary atresia, except that there are continuous collateral murmurs and often a collapsing pulse. Coexistent chromosome 22q11 deletion is more common than in tetralogy without pulmonary atresia. There may be dysmorphic facies, hypertelorism, narrow eye fissures, puffy eyelids, a small mouth, deformed earlobes, and sometimes a cleft palate.

The chest radiograph shows a right aortic arch in 25 per cent of cases and has a typical appearance (Fig. 15). The pulmonary collateral vessels may follow a bizarre pattern. Colour flow Doppler may identify collateral vessels, but conventional angiography is required to delineate precisely their origin, degree of ostial stenosis, and intrapulmonary course (Fig. 16). Coronary–pulmonary collaterals are a frequent finding (Fig. 17). Three-dimensional magnetic resonance angiography is likely to prove a useful tool in imaging complex pulmonary vasculature.

Fig. 15 Chest radiograph of a 21-year-old woman with tetralogy of Fallot and pulmonary atresia, no central pulmonary arteries, and multiple aortopulmonary collaterals which create an abnormal pulmonary vascular pattern. The typical 'coeur en sabot' silhouette is due to right ventricular hypertrophy and the pulmonary bay where the pulmonary artery should be (arrow).

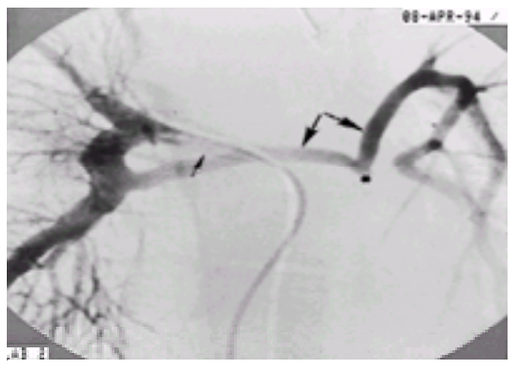

Fig. 16 Aortopulmonary collateral angiogram of a 24-year-old woman with tetralogy of Fallot and pulmonary atresia. There is a stent (small arrow) at the origin of the collateral from the descending aorta. Only after dilation and stenting of the collateral was its connection with confluent main pulmonary arteries (large arrows) apparent. The atretic main pulmonary artery is indicated (*).

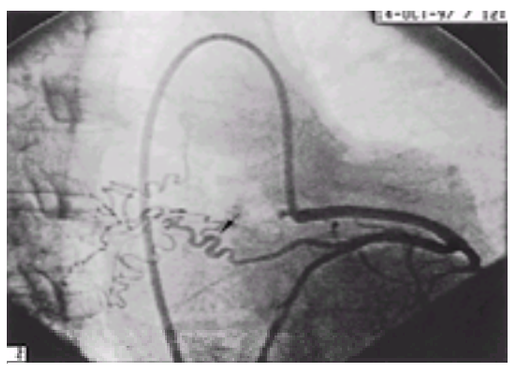

Fig. 17 Selective left coronary angiogram in same patient as Fig. 11 showing a circumflex coronary artery to pulmonary collateral (arrow).

Overall survival, including the effects of operation, is around 25 per cent at 20 years. Late complications in unoperated survivors include increasing cyanosis due either to the development of pulmonary vascular disease in lung segments perfused at systemic pressure through non-stenosed collaterals, or to the progressive stenosis of collateral vessels. In the latter, good symptomatic relief may be obtained from stenting the stenotic vessel. The aortic root may become markedly dilated and aortic regurgitation develop, resulting in biventricular volume overload and failure. Aortic valve endocarditis is a particular risk. Surgery to repair or replace the aortic valve in an unrepaired patient with pulmonary atresia is particularly hazardous because of the aortopulmonary collateral vessels. When cardiopulmonary bypass is instituted, aortic blood flows into the low resistance pulmonary collateral vessels so that it is not possible to maintain an adequate systemic perfusion pressure or to control pulmonary venous return to the left atrium.

Surgery to unifocalize collateral vessels (i.e. disconnect them from the aorta and anastomose them to the pulmonary artery, maintaining pulmonary blood supply by means of a shunt) may need to precede radical repair. Suitability for radical repair depends on the size of the pulmonary arteries and the proportion of lung they supply. The ventricular septal defect is closed and the right ventricular outflow tract connected to the pulmonary artery via a valved conduit. Right ventricular hypertension may follow surgery if the pulmonary arteries are small or distal pulmonary vessels inadequate. Both these factors and the number of aortopulmonary collaterals decrease long-term survival.

Tetralogy of Fallot with absent pulmonary valve syndrome

This variation accounts for around 3 per cent of cases of tetralogy. There is a ring-like malformation, usually stenotic, with failure of development of the pulmonary valve cusps. The central pulmonary arteries are usually hugely dilated or aneurysmal (Fig. 18).

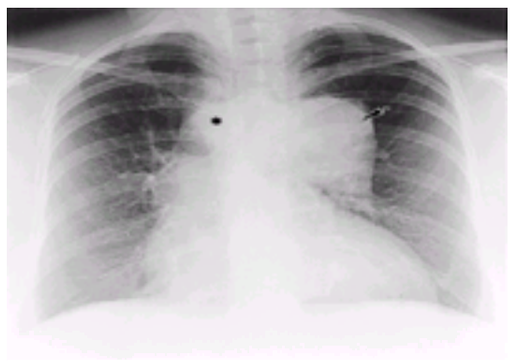

Fig. 18 Chest radiograph of a 54-year-old woman with repaired tetralogy of Fallot and absent pulmonary valve syndrome. The central pulmonary arteries are hugely dilated (arrow) and there is a right aortic arch (*).

Other right-sided obstructive lesions

Isolated pulmonary valve stenosis

Pulmonary stenosis is discussed in detail elsewhere. Cyanosis may be present if severe stenosis coexists with an atrial septal defect or patent foramen ovale. Noonan's syndrome is associated with valvar and infundibular stenosis as well as with hypertrophic cardiomyopathy.

Lone infundibular stenosis and double-chambered right ventricle

Abnormally placed muscle bands cause either infundibular obstruction, or if placed more inferiorly, subinfundibular obstruction and a double-chambered right ventricle. The degree of obstruction may be mild in childhood, but progress into adult life and cause symptoms as the right ventricle hypertrophies. A perimembranous ventricular septal defect usually coexists and may close spontaneously. Treatment is by surgical resection of the obstructing muscle bands.

Pulmonary atresia with intact septum

A full discussion of this complex lesion is beyond the scope of this chapter, since patients do not survive unoperated beyond infancy. Those currently in the adult clinics represent the mild end of the spectrum of this condition and are likely to have had a pulmonary valvotomy or valved right ventricular to pulmonary artery conduit.

Left ventricular outflow tract obstruction

Bicuspid aortic valve

The commonest congenital cardiac anomaly, occurring in 1 to 2 per cent of the population, bicuspid aortic valve is four times as common in males as females. In 20 per cent of cases it is associated with other lesions such as patent arterial duct and coarctation. There is also an association with aortic root dilatation and dissection. Aortic stenosis is discussed in detail elsewhere.

Supravalvar aortic stenosis

In this least common form of left ventricular outflow tract obstruction, there is a localized narrowing of the aorta immediately above the aortic sinuses. Fibromuscular thickening of the aortic wall at the site of obstruction may encroach into the coronary ostia or on to the aortic valve leaflets and adversely influence prognosis. Unlike other forms of left ventricular outflow obstruction, the coronary arteries lie proximal to the obstruction and so are exposed to high left ventricular pressures, resulting in premature atherosclerosis. The condition may be associated with Williams' syndrome, when the prognosis may be worse since there is diffuse arterial involvement that may involve the pulmonary and renal arteries.

Subaortic stenosis

Subaortic stenosis is due either to a discrete fibromuscular ridge or ring, or a long muscular tunnel. It may exist in isolation or as part of another lesion such as atrioventricular septal defect where the aorta is 'unwedged' and the left ventricular outflow tract elongated, or mitral valve anomalies where abnormal insertion of the mitral valve causes obstruction. Whether discrete or tunnel-like, subaortic stenosis tends to progress and may recur following surgical resection. It may result in functional disruption of the aortic valve and secondary aortic regurgitation, which may progress, even after resection of subaortic stenosis.

Coarctation of the aorta

One of the commonest congenital cardiac lesions, occurring in 1 per 12 000 live-born babies with a male to female ratio of 3:1, aortic coarctation is a narrowing of the aorta usually sited near the ligamentum arteriosum. There is considerable variation in anatomy and severity, ranging from a mild obstruction to interruption of the aorta, and from a discrete fibromuscular shelf to hypoplasia of the arch. Coarctation is most strongly associated with bicuspid aortic valve; other associations are ventricular septal defect, patent ductus arteriosus, subaortic ridge and mitral valve abnormalities. It is a frequent finding in Turner's syndrome and is also associated with congenital aneurysm of the circle of Willis.

Unoperated history

Most patients present in infancy, but some survive into adulthood before being diagnosed at routine examination or during investigation for hypertension, leg claudication (uncommon unless there is coexisting abdominal aortic coarctation), angina, heart failure, or cerebral haemorrhage. More than 75 per cent of patients with unoperated coarctation die by age 50 years, from coronary disease, stroke, or aortic dissection.

Clinical findings include upper body hypertension: the leg blood pressure is lower, as is that in the left arm if the subclavian artery is involved in the coarctation. If there is a good collateral supply, femoral arteries may be easily palpable, but they are usually reduced, with radiofemoral delay. Intercostal collaterals may be both visible and palpable over the patient's back. There is an ejection systolic murmur from the site of coarctation, and systolic collateral murmurs may be heard. Fundoscopy shows a typical corkscrew appearance of the retinal vessels and there may be evidence of hypertensive retinopathy.

There may be electrocardiographic evidence of left ventricular hypertrophy. The chest radiograph (Fig. 19) has a typical appearance. Transthoracic echocardiography may show left ventricular hypertrophy, but the coarctation site may not be visualized on two-dimensional imaging, although the severity of coarctation can be assessed using Doppler mode from the suprasternal notch. A peak gradient of gretaer than 20 mmHg is significant, especially if accompanied by a diastolic tail. Angiography allows full haemodynamic and anatomical data to be obtained from both the coarctation site and related vessels, as well as assessing secondary ischaemic myocardial disease. Magnetic resonance imaging provides excellent non-invasive haemodynamic data and two- and three-dimensional images of the coarctation site and related vessels. It may obviate the need for angiography unless coronary disease is suspected.

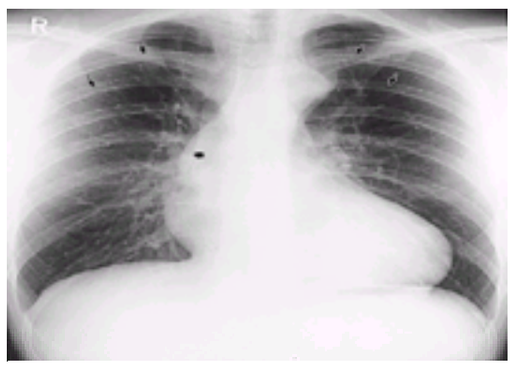

Fig. 19 Chest radiograph of an 18-year-old man with unoperated coarctation of the aorta and bicuspid aortic valve. There is bilateral rib notching (arrows), a dilated ascending aorta (*), and a prominent deformed aortic knuckle.

Repair of coarctation

Surgical repair is the conventional approach. There is a 0.4 per cent incidence of perioperative spinal cord ischaemia and paraplegia: patients without an abundant collateral circulation may be most at risk. Those with well-developed collaterals are at risk of significant intraoperative haemorrhage. Early postoperative hypertension is common and may be difficult to control, and postoperative intestinal ileus may persist for several days.

Transcatheter balloon dilatation and primary stenting of native coarctation in adults are reported, but data are limited and the procedures should still be considered experimental. Stenting has the hypothetical advantage of supporting the dilated segment of aorta which may sustain a significant intimal tear, and preventing aortic rupture or aneurysm formation.

Follow-up after coarctation repair

Follow-up after repair of coarctation should be life-long, since late complications are frequent: recoarctation, aneurysm formation, persistent hypertension despite adequate repair, premature atherosclerotic disease, and progression of associated lesions such as bicuspid aortic valve. Older age at repair is the main risk factor influencing longevity. Late survival is 92 per cent for patients repaired in infancy, 25-year survival is 75 per cent for those repaired between the ages of 20 and 40 years, but 15-year survival is only 50 per cent for those repaired at age 40 years or more.

Recoarctation may be diagnosed when the resting arm–leg systolic blood pressure gradient is 20 mmHg at rest and 50 mmHg postexercise. It occurs most commonly in neonatal repair by end-to-end anastomosis. Recoarctation should be sought when there is new or persisting hypertension. Blood pressure should be recorded in both arms of all such patients; spuriously low readings may be obtained if one of the subclavian arteries (usually the left) is involved in the repair or recoarctation. Magnetic resonance imaging is the investigation of choice for both recoarctation and aneurysm formation after coarctation repair (Fig. 20). Balloon angioplasty with or without stent insertion is used to relieve the majority of recoarctations, but reoperation is required for some patients with complex anatomy.

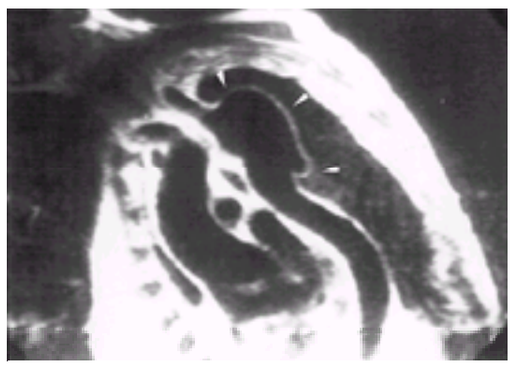

Fig. 20 Magnetic resonance spin echo image showing aneurysm of the descending aortic arch (arrows) following previous Dacron patch repair of coarctation

Hypertension is a major risk factor for atherosclerotic disease and may persist despite an apparently good result from surgical repair. Continuing hypertension is thought to relate in part to older age at time of surgery. None the less, even if repaired in adulthood, systolic hypertension becomes less marked and easier to control, b-blocking agents being the antihypertensives of choice.

The 14-year incidence of aneurysm formation at the site of repair is up to 27 per cent, it occurs most commonly in adults and in those with Dacron patch repair (Fig. 20). The aneurysm may rupture into the bronchial tree: any patient with a history of coarctation who presents with haemoptysis should undergo emergency non-invasive diagnostic imaging (preferably MRI) and surgical repair. Bronchoscopy and conventional angiography are contraindicated since they may cause further damage to the ruptured area.

Congenitally corrected transposition of the great arteries (atrioventricular and ventriculoarterial discordance)

This rare condition accounts for less than 1 per cent of all congenital heart disease. Both atrial and arterial connections to the ventricles are discordant, so pulmonary venous blood passes through the left atrium, through the right ventricle, and into an anteriorly lying aorta. Similarly, systemic venous blood reaches the pulmonary trunk via the left ventricle. The circulation is therefore physiologically 'corrected', but the morphological right ventricle and tricuspid valve support the systemic circulation.

More than 95 per cent of cases have associated anomalies, most commonly ventricular septal defect and pulmonary stenosis, but also Ebstein anomaly of the systemic (tricuspid) atrioventricular valve, aortic stenosis, atrioventricular septal defect, abnormalities of situs, and coarctation. Congenital complete heart block occurs in around 5 per cent of patients and may develop at any stage of life, particularly following surgery to the atrioventricular valve.

Presentation depends on associated lesions. Patients with isolated congenitally corrected transposition of the great arteries may remain asymptomatic and undiagnosed into old age, but failure of the systemic ventricle, systemic atrioventricular valve regurgitation, or the onset of complete heart block and atrial arrhythmias usually result in presentation with symptoms from the fourth decade onwards. Those with ventricular septal defect and pulmonary stenosis may be cyanosed, and those with ventricular septal defect alone may present with pulmonary hypertension.

A parasternal heave is usually palpable from the pressure-loaded anteriorly lying systemic right ventricle; this may be especially prominent if it is also volume-loaded by systemic (tricuspid) atrioventricular valve regurgitation. There may be a prominent aortic pulsation in the suprasternal notch and the aortic component of the second heart sound may be palpable and loud. The pulmonary component is soft or inaudible due to the posterior position of the pulmonary artery.

The electrocardiogram may show varying degrees of atrioventricular block or evidence of pre-excitation due to accessory pathways (associated with Ebstein-like anomalies of the systemic atrioventricular valve). There may be left axis deviation. The right and left bundles are inverted, so the initial septal activation is right-to-left, resulting in Q waves in V1 to 2 and an absent Q in V5 to 6; this pattern is often wrongly interpreted as a previous anterior myocardial infarction. The chest radiograph has a typical appearance (Fig. 21). Echocardiography confirms the discordant relations and assesses ventricular and systemic (tricuspid) atrioventricular valve function as well as other associated lesions. Ebstein's anomaly may be diagnosed if the tricuspid valve is apically displaced more than 8 mm. Cardiac catheterization

is indicated to assess the haemodynamic importance of associated lesions.

Fig. 21 Chest radiograph of a 23-year-old woman with congenitally corrected transposition of the great vessels. There is a narrow pedicle due to the abnormally related great arteries (small arrow) and the left heart border is straight (large arrow) due to the abnormal position of the left-lying anterior ascending aorta.

ACE inhibitors may be useful when there is systemic ventricular dysfunction or atrioventricular valve regurgitation, but there are no trial data to support their use. Transvenous atrioventricular sequential pacing is indicated for complete heart block: active fixation ventricular leads are required because of the absence of coarse apical trabeculations in the morphologically left subpulmonary ventricle. If there are associated intracardiac shunts, patients should be formally anticoagulated to reduce the risk of paradoxical embolism, or epicardial pacing should be considered.

The conventional surgical approach to systemic atrioventricular valve regurgitation is tricuspid valve replacement (repair is rarely successful), but if systemic ventricular function is poor (ejection fraction less than 40 per cent), transplantation may be the only option. Where there is coexistent ventricular septal defect and pulmonary stenosis, classic repair involves closure of the ventricular septal defect and insertion of a valved conduit between the left ventricle and pulmonary artery; the right ventricle continuing to support the systemic circulation.

Anatomical repair, so that the morphological left ventricle supports the systemic ventricle, has achieved short-term success in children with systemic atrioventricular valve regurgitation and systemic ventricular dysfunction. For patients with an associated non-restrictive ventricular septal defect the left ventricle is at systemic pressure and therefore 'pre-trained' to support the systemic circulation. If there is no pulmonary stenosis, a 'double switch' is performed, combining an intraatrial repair (usually Senning operation) with an arterial switch operation. If there is also pulmonary stenosis, the Senning operation is combined with a Rastelli-type repair, in which the ventricular septal defect is closed so that the left ventricle is tunnelled to the aorta, and a right ventricular to pulmonary artery conduit is placed. The regurgitant tricuspid valve and right ventricle are therefore placed in the pulmonary circulation. For children with corrected transposition whose left ventricle is at low pressure, a period of left ventricular 'training' is required before a double switch operation can be performed. Training is achieved by placing a pulmonary artery band to increase left ventricular pressure and induce hypertrophy. Pulmonary artery banding *per se* may improve symptoms, since the increased left ventricular pressure causes the interventricular septum to move towards the systemic ventricle, reducing systemic atrioventricular regurgitation. The long-term outcome of these anatomical approaches to corrected transposition is not yet known; complications relating to conduit replacement, neo-aortic valve regurgitation, and arrhythmia may become significant. There are reports of adults with ventricular septal defect and pulmonary stenosis having successfully undergone Mustard–Rastelli repair. However, whether it is possible to 'train' the adult left ventricle that has been at low pressure for many years remains to be seen, so this approach remains experimental in older patients.

Complete transposition of the great arteries (atrioventricular concordance, ventriculoarterial discordance)

Complete transposition of the great arteries accounts for around 5 per cent of congenital cardiac malformations and is four times more common in males than females. Associated anomalies such as ventricular septal defect and pulmonary stenosis occur in about one-third of patients. Desaturated systemic venous blood passes through the right atrium into the right ventricle and then to the aorta, and oxygenated pulmonary venous blood passes into the left atrium, through the left ventricle and back into the lungs. Once the arterial duct closes, survival depends on an intracardiac communication (non-restrictive patent foramen ovale or ventricular septal defect) allowing mixing of blood between the two separate circuits. Without intervention, 30 per cent of patients die within the first week and only 10 per cent survive their first year. A prostaglandin infusion maintains patency of the arterial duct until a balloon atrial septostomy is performed. The neonate remains cyanosed, but there is usually adequate mixing to allow him or her to thrive until definitive surgery.

Most patients in the adult clinic have survived intra-atrial repair (Senning or Mustard operation), or for those with associated ventricular septal defect and pulmonary stenosis, a Rastelli operation. Intra-atrial repair involves excision of the atrial septum and placement of a saddle-shaped patch ('baffle') to direct pulmonary venous blood into the right atrium, right ventricle and then to the aorta, and systemic venous blood into the left atrium, left ventricle and then into the pulmonary artery. The right ventricle supports the systemic circulation. In the Rastelli operation, the ventricular septal defect is closed so that the left ventricle carrying oxygenated blood empties into the aorta, and a conduit is placed between the right ventricle and pulmonary artery. The left ventricle supports the systemic circulation.

Since the late 1970s anatomical correction by the arterial switch operation began to supersede intraatrial repair as the operation of choice for most patients with transposition of the great arteries. Blood is redirected at arterial level by switching the aorta and pulmonary arteries so that the left ventricle becomes the subaortic ventricle supporting the systemic circulation. The coronary arteries are reimplanted into the neo-aortic root.

'Palliative' Mustard or Senning operations are performed for patients with transposition of the great arteries, ventricular septal defect, and pulmonary vascular disease to improve mixing of blood and oxygenation. The ventricular septal defect is left open. These patients should be treated as other patients with the Eisenmenger syndrome.

Follow-up

After intra-atrial repair of transposition of the great arteries, the systemic right ventricle causes a parasternal heave, the aortic component of the second heart sound may be palpable and loud, and the second sound single, due to the anterior-lying aorta. The presence of cyanosis suggests a baffle leak allowing right to left shunting between the systemic and pulmonary venous atria. Systemic venous pathway obstruction may be associated with elevation of the jugular venous pressure and hepatomegaly. The chest radiograph may show a dilated azygos vein indicative of systemic venous pathway obstruction, with run-off from the obstructed to the unobstructed pathway. Pulmonary interstitial fluid may reflect pulmonary venous pathway obstruction. Although late results of intraatrial repair are good, failure of the systemic right ventricle is exacerbated as the atrioventricular valve becomes regurgitant. In addition, baffle obstruction causing systemic or pulmonary venous obstruction may require transcatheter intervention or reoperation. Atrial arrhythmias, especially atrial flutter, occur in up to 10 per cent of patients 10 years postoperatively. They may be poorly tolerated and are associated with an increased risk of sudden death. Late sinus node dysfunction and complete heart block may also occur: active fixation leads are required to pace both the systemic venous atrium and left ventricle. As in congenitally corrected transposition of the great arteries, the conventional surgical approach for systemic atrioventricular valve regurgitation is tricuspid valve replacement. It is uncertain whether pulmonary artery banding to train the left ventricle and reduce regurgitation, followed by the arterial switch operation, will benefit adults in this situation.

The major late complication following the Rastelli operation is the need for conduit replacement. Late results from the arterial switch operation are awaited. It is likely to have long-term advantages over intra-atrial repair, since it restores normal connections so that the left ventricle supports the systemic circulation without the need for a conduit. Follow-up is needed to detect late pulmonary arterial stenosis, neo-aortic regurgitation, and coronary ostial stenoses.

Hearts with univentricular atrioventricular connection (double-inlet left ventricle and tricuspid atresia)

Also known as univentricular or single ventricle hearts, these hearts are defined by the connection of both atria to one ventricle, or by the absence of one of the atrioventricular connections. There is one dominant ventricle, with a second rudimentary and incomplete ventricle. When the rudimentary ventricle is of right morphology, it nearly always lies anteriorly. Less commonly, there is a posteriorly lying morphologically left rudimentary ventricle, and rarely there is solitary ventricle

of indeterminate morphology.

This section considers double-inlet left ventricle and tricuspid atresia. Less common variants such as double-inlet right ventricle, mitral atresia, and more complex univentricular hearts are beyond the scope of this chapter. Hearts with a large ventricular septal defect and two fully formed ventricles are not correctly termed univentricular and are also excluded. None the less the pathophysiological principles are similar.

In double-inlet left ventricle, ventriculoarterial discordance (transposed great arteries) usually coexists; the aorta arises from the rudimentary right ventricle via the ventricular septal defect. Double-inlet left ventricle with concordant ventriculoarterial connection (Holmes heart) is rare. Tricuspid atresia has a number of morphological variations, but the different morphologies do not affect the basic haemodynamics: the only route out of the right atrium is via an atrial septal defect into the left atrium and then to the ventricular mass. The ventriculoarterial connection is most commonly concordant.

Pathophysiology

For all univentricular hearts, key factors determining the mode of presentation and survival are the presence and degree of pulmonary stenosis, presence and degree of subaortic stenosis, and the morphology of the dominant ventricle.

In hearts with concordant ventriculoarterial connections, pulmonary stenosis may be both valvar and subpulmonary, caused by a restrictive ventricular septal defect, since the pulmonary artery arises from the rudimentary ventricle. Absence of pulmonary stenosis results in high pulmonary blood flow and eventual pulmonary vascular disease. Severe pulmonary stenosis or atresia causes marked cyanosis but protects against pulmonary vascular disease. In hearts with ventriculoarterial discordance, a restrictive ventricular septal defect causes subaortic stenosis, since the aorta arises from the rudimentary ventricle. Subaortic stenosis may be acquired (see Introduction), resulting in preferential pulmonary blood flow and reduced cardiac output. If the dominant ventricle is of left morphology it is better able to adapt to its abnormal geometry and chronic volume overload than a right ventricle.

The outcome is most favourable for patients with left ventricular morphology, moderate pulmonary stenosis, and no subaortic stenosis. Unoperated survival into adulthood is uncommon, 50 per cent of patients with double-inlet left ventricle die before 14 years, 50 per cent with double-inlet right ventricle die by 4 years of age. None the less, rare patients with balanced circulation reach their sixth decade without surgical intervention.

Clinical signs in the unoperated adult

There is cyanosis and clubbing. A giant 'a' wave may be present in the jugular venous pulse in tricuspid atresia. An absent right ventricular impulse and prominent left ventricular impulse are characteristic of double-inlet left ventricle and tricuspid atresia. There may be a precordial thrill from pulmonary stenosis, particularly if the pulmonary artery lies anteriorly. If there are discordant ventriculoarterial connections, the aortic pulsation of the anteriorly lying aorta may be prominent in the suprasternal notch. The second heart sound is usually single and there may be a pulmonary ejection systolic murmur radiating laterally, also a pansystolic murmur of mitral regurgitation.

If pulmonary vascular disease has developed there will be additional signs of pulmonary hypertension. Signs of congestive heart failure may be present in the ageing patient, particularly with the onset of atrial arrhythmia: the venous pressure may be raised, with hepatomegaly and peripheral oedema.

Investigations

The chest radiograph shows cardiomegaly due to chronic ventricular volume overload. If ventriculoarterial connections are discordant, there is a narrow pedicle and the ascending aorta forms a straight edge along the left heart border. Pulmonary vascularity reflects the pulmonary blood flow. The main pulmonary arteries are small where there is significant pulmonary stenosis. Large main pulmonary arteries indicate high pulmonary blood flow, either past or present.

In tricuspid atresia the electrocardiogram usually shows right atrial hypertrophy, normal P–R interval, small or absent right ventricular forces, and left axis deviation. There is left axis deviation and large left ventricular forces in double-inlet left ventricle. If the rudimentary chamber lies to the right, the P–R interval is usually normal, but if it lies to the left, the P–R interval may be prolonged or there may be complete heart block.

Two-dimensional echocardiography and colour flow Doppler allow detailed assessment of the anatomy and physiology, including ventricular morphology and pulmonary and subaortic stenosis. Cardiac catheterization is required to assess pulmonary artery pressure and resistance and to detail pulmonary artery anatomy.

Operations for hearts with univentricular atrioventricular connections

All surgical approaches are palliative, since a biventricular repair is not possible. The aim of definitive surgery is to separate the systemic and pulmonary circulations. Early procedures such as the Glenn anastomosis partly achieved this aim, and it remains a useful staging procedure in children prior to definitive surgery, with the advantage of offloading the ventricle and perfusing the lung at low pressure. However, the relative contribution of the superior vena cava reduces as the child grows, so progressive cyanosis develops. It is therefore inadequate as the sole source of pulmonary blood supply in adults. Pulmonary arteriovenous fistulas are a late complication of Glenn anastomoses and are thought to relate to the exclusion of hepatic venous blood from the pulmonary circulation.

The Fontan operation creates an atriopulmonary connection so that systemic venous blood passes directly to the pulmonary artery, thus bypassing the ventricle, completely separating the systemic and pulmonary circulations and abolishing cyanosis. The ventricle supports the systemic circulation and systemic venous blood flows passively into the pulmonary arteries. The evolution of the Fontan operation and its successor, the total cavopulmonary connection is outlined in Fig. 22. Many variations of the Fontan operation have been developed that are beyond the scope of this chapter.

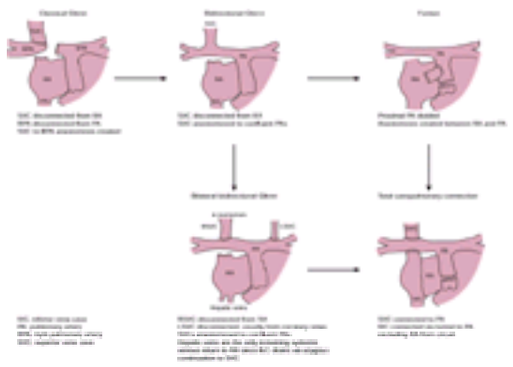

Fig. 22 Evolution of the Fontan operation and total cavopulmonary connection.

The Fontan-type circulation is one of a low cardiac output that can only be increased by increasing the heart rate. This lack of reserve and dependence on adequate systemic venous return both limits exercise tolerance and renders the patient susceptible to subsequent anaesthesia and surgery. Meticulous fluid balance and avoidance of hypovolaemia and excessive vasodilatation are required to prevent cardiovascular collapse.

Survival after Fontan-type surgery is dependent on patient selection and ranges from 81 per cent at 10 years for 'perfect candidates' to 60 to 70 per cent for all patients. Patients with preoperative adverse risk factors for Fontan-type surgery (Table 12) are also more likely to develop long-term complications (Table 13). Paroxysmal atrial flutter (intra-atrial re-entry tachycardia) is a major cause of morbidity in long-term survivors. Right atrial distension, high atrial pressures, atrial suture lines, and obstruction to the Fontan circuit contribute to the development of arrhythmias. Transoesophageal echocardiography and magnetic resonance imaging may visualize the Fontan connection, but cardiac catheterization is usually required to quantify the degree of obstruction. Since the velocities within the Fontan circuit are

low (less than 1 m/s), a pressure drop of only 2 to 3 mmHg between the pulmonary artery and right atrium may indicate haemodynamically important obstruction. Atrial flutter is poorly tolerated so cardioversion should not be delayed. The arrhythmia may become increasingly difficult to control, and if ventricular function is impaired, amiodarone may be the most effective and best tolerated antiarrhythmic. However, amiodarone-induced thyrotoxicosis occurs in up to 40 per cent of Fontan survivors and may precipitate further tachyarrhythmias and heart failure, so thyroid function should be monitored carefully. Whether by excluding the atrium from the systemic venous–pulmonary artery circuit the total cavopulmonary connection will reduce atrial arrhythmia remains to be seen. There is early evidence that conversion of the Fontan to a total cavopulmonary connection, in combination with arrhythmia circuit cryoablation, may reduce the incidence of arrhythmia and improve functional class.

There is usually echocardiographic evidence of spontaneous contrast in Fontan survivors with a dilated right atrium, indicating sluggish flow and a high risk of thrombus formation. Patients with atrial arrhythmia are at particular risk and should be formally anticoagulated. A deficiency of anticoagulation factors in patients after Fontan procedures may also contribute towards thromboembolism.

Protein-losing enteropathy is one of the most debilitating complications of Fontan-type surgery, occurring in up to 13 per cent of late survivors, with a 5-year survival after its onset of less than 50 per cent. It is thought to result from the effects of chronically elevated systemic venous pressure on the lymphatic system causing gastrointestinal protein loss, with malnutrition, oedema, effusions, and ascites due to hypoalbuminaemia, as well as infections secondary to hypogammaglobulinaemia. The diagnosis is confirmed by a low serum albumin and high faecal a_1-antitrypsin. A high protein, low fat, high medium-chain triglyceride diet has been advocated, but is unpalatable. Treatment with corticosteroids, unfractionated heparin, or transcatheter fenestration of the atrial septum may be beneficial. Surgical relief of any Fontan obstruction may be successful, but carries a high mortality and cardiac transplantation may be the only option, although protein-losing enteropathy may recur.

Most patients in the adult clinic have undergone Fontan-type surgery in childhood or adolescence. However, those patients with univentricular heart that survive to adulthood without an operation or with previous shunts are often considered for Fontan-type surgery. Such patients represent a highly selected group of survivors with a well-balanced circulation, and the long-term complications of cyanosis and ventricular volume overload should be weighed carefully against the risk of Fontan surgery in an adult and its long-term complications.

Other arterial anomalies

Persistent patent ductus arteriosus

The pathophysiological consequences of a patent arterial duct in adulthood depend on the size of the shunt. Small ducts are of no haemodynamic significance and are associated with a low risk of infective endarteritis. Moderate-sized ducts may cause left heart volume overload and late atrial fibrillation and ventricular dysfunction. A large non-restrictive duct may cause pulmonary vascular disease (see Eisenmenger syndrome).

If a duct is clinically detectable, that is there is a machinery murmur in the left subclavicular area, then closure is usually recommended to avoid long-term haemodynamic complications. Ducts up to about 8 mm in diameter are usually suitable for transcatheter device closure, but calcification and aneurysmal dilatation around the area of the duct may necessitate surgical repair. In large ducts, pulmonary vascular disease should be excluded before surgical repair is undertaken.

Truncus arteriosus

A single great artery arises from the heart and gives rise to the coronary arteries, aorta, and pulmonary arteries. There is a single semilunar 'truncal' valve that has three or more leaflets, and a subtruncal ventricular septal defect. Figure 23 shows the different patterns by which the pulmonary arteries arise, type 1 is the most common.

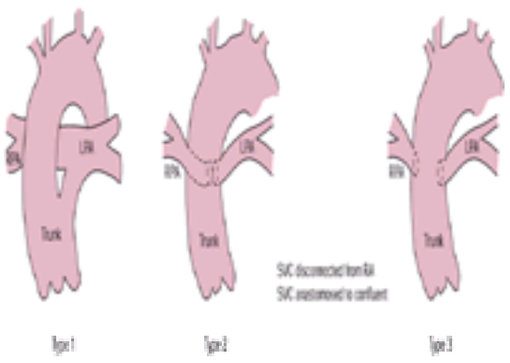

Fig. 23 Truncus arteriosus. Type 1 has a short main pulmonary artery arising from the common arterial trunk. Types 2 and 3 can be considered together, the pulmonary arteries have separate origins from the common trunk. LPA, left pulmonary artery; RPA, right pulmonary artery.

Most patients present in infancy with heart failure. If left unoperated, pulmonary vascular resistance rises, cyanosis becomes more marked, and the Eisenmenger reaction becomes established. Repair before pulmonary vascular disease develops involves closure of the ventricular septal defect, detachment of the pulmonary arteries from the common arterial trunk, and placement of a valved right ventricular to pulmonary artery conduit. The truncal valve then functions as the aortic valve. Late complications include truncal regurgitation and the need to replace stenotic conduits.

Aortopulmonary window

In this rare condition there is a direct communication between adjacent portions of the proximal ascending aorta and pulmonary artery. The communication is usually large and the physiological consequences are the same as for a patent arterial duct. Rare patients surviving without operation into adulthood are likely to have developed the Eisenmenger reaction. If pulmonary vascular resistance is low at the time of childhood repair, long-term postoperative survival is good.

Coronary artery anomalies

The importance of congenital coronary anomalies lies in their potential to impair myocardial blood flow and cause ischaemia and sudden death. Evidence of ischaemia is the main indication for repair. The major types of coronary anomaly are shown in Table 14.

Anomalous origin of the coronary arteries from an inappropriate aortic sinus

Ischaemia is particularly associated with an anomalous proximal coronary course between the aorta and pulmonary trunk, an intramural proximal segment of the coronary artery inside the aortic wall, and acute angulation between the origin of an anomalous coronary artery and the aortic wall.

Anomalous origin of the left coronary artery from the pulmonary artery

This rare condition usually presents in infancy with myocardial ischaemia and left ventricular failure when pulmonary vascular resistance decreases. However, 10 to 15 per cent survive into adulthood because an adequate intercoronary collateral circulation is established. Adults may be asymptomatic or present with myocardial ischaemia or mitral regurgitation due to papillary muscle dysfunction. Survival following surgical repair depends on the amount of ischaemic myocardial damage and degree of mitral regurgitation.

Congenital coronary arteriovenous fistulas

The coronary arteries arise normally from their aortic sinuses, but a fistulous branch communicates directly with the right ventricle in 40 per cent of cases (Fig. 24), the right atrium in 25 per cent, pulmonary artery in 15 per cent, or rarely the superior vena cava or pulmonary vein. Survival to adulthood is usual, but life expectancy is reduced and depends on the size of the fistulous connection and the presence of myocardial ischaemia resulting from any coronary steal phenomenon. Symptoms increase with age and there is a risk of endocarditis, heart failure, arrhythmia, myocardial ischaemia and infarction, and sudden death. Surgical repair is recommended unless there is a trivial isolated shunt. Some smaller fistulas are suitable for transcatheter device occlusion.

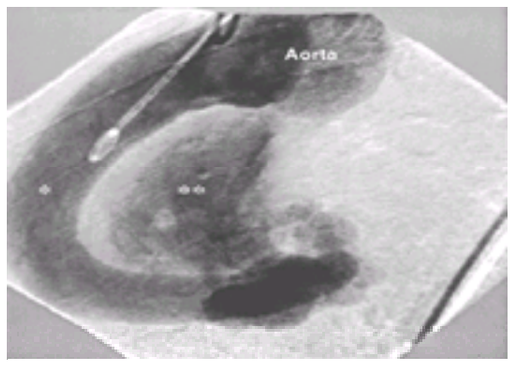

Fig. 24 Angiogram (left anterior oblique) of a 32-year-old man with a huge right coronary artery fistula (*) to the right ventricle (**).

Ruptured sinus of Valsalva aneurysm

Any of the three aortic sinuses of Valsalva may become aneurysmal and rupture. The right and non-coronary cusps are most often affected; rupture of the non-coronary sinus aneurysm is nearly always into the right atrium and of the right coronary sinus into the right ventricle or atrium. Involvement of the left coronary sinus is rare. Rupture usually occurs in early adulthood and may be precipitated by endocarditis. If sudden, it is accompanied by tearing chest pain, breathlessness, and congestive cardiac failure with a loud continuous murmur. Small perforations may remain asymptomatic for many years. The diagnosis and site of the rupture is confirmed angiographically prior to surgical repair. Transcatheter closure has also been reported.

Pregnancy in women with congenital heart disease

In general, the combined oral contraceptive pill is contraindicated in cyanosis, pulmonary hypertension, and following Fontan-type surgery. The progesterone-only pill and progesterone depot injections are safe alternatives, but the former is less reliable. Intrauterine devices are not ideal for many women because of the risk of endocarditis and menorrhagia. Women with pulmonary hypertension may wish to consider sterilization: although a delicate issue, sterilization of the male partner is not recommended as he is likely to outlive the patient and may wish to start a family later. Ventricular dysfunction and arrhythmias tend to deteriorate during pregnancy: regurgitant lesions are better tolerated than stenotic lesions. Pregnancy is contraindicated in pulmonary hypertension, severe systemic ventricular dysfunction (ejection fraction (EF) less than 20 per cent), and severe unoperated left-sided obstruction.

Unoperated atrial septal defect

Atrial septal defect is usually well tolerated in pregnancy; the major risk is of paradoxical embolism, so meticulous leg care and low-dose aspirin are important to avoid venous thrombosis. Haemorrhage may cause a significant increase in left to right shunting by reducing systemic venous return and increasing systemic vascular resistance. Women older than 30 years are at higher risk of developing atrial arrhythmias and right ventricular dysfunction as a result of the increased blood volume of pregnancy.

Tetralogy of Fallot after radical repair

Pregnancy is usually well tolerated after radical repair, provided there is little outflow tract gradient and ventricular function is good.

Ebstein anomaly

The risks relate to an inability of the functional right ventricle to accommodate the increased blood volume of pregnancy. Cyanosis and the risk of paradoxical embolism may present for the first time in pregnancy because of a rise in right ventricular pressure.

Systemic right ventricle

The ability of patients with congenitally corrected transposition of the great arteries or previous Mustard operation to tolerate pregnancy depends on the function of the systemic ventricle, systemic atrioventricular regurgitation, pulmonary venous baffle obstruction (patients with a Mustard procedure), and atrial tachyarrhythmias, all of which may worsen as pregnancy progresses. Patients with complete heart block may tolerate pregnancy better after the insertion of a permanent pacemaker.

Coarctation of the aorta

Pregnancy is a risk factor for aortic dissection and rupture, especially if there is an aneurysm at the site of repair or a coexisting bicuspid aortic valve. Histology of any dissected area shows cystic medial necrosis. Meticulous control of blood pressure is important and the second stage of labour should be kept short to reduce arterial wall stress.

After Fontan operation

Despite the limited ability to increase cardiac output, successful pregnancy is possible in patients with a good postoperative result who have not developed major late complications. There is a high incidence of early miscarriage.

Women with cyanosis but no pulmonary hypertension

Maternal mortality is considerably less in women with cyanosis but normal or low pulmonary artery pressures than in those with pulmonary hypertension. Maternal morbidity and fetal outcome are determined by ventricular function and cyanosis. The reduction in systemic vascular resistance and increase in cardiac output during pregnancy increases the right to left shunt, resulting in a fall in SaO_2. If resting SaO_2 is less than 85 per cent, the risk of miscarriage is around 80 per cent, with additional risks of low birth weight and prematurity. Heart failure may be precipitated by the increased blood volume, particularly if there is aortic regurgitation. The hypercoaguable state of pregnancy combined with a high haematocrit increases the risk of thrombosis and paradoxical embolism. Patients are also at risk of haemorrhage during delivery, particularly if a section is performed. Labour and delivery must be managed carefully, avoiding hypovolaemia and vasodilatation which may precipitate intense cyanosis, syncope, and death.

Eisenmenger syndrome

Maternal mortality remains unchanged at about 40 per cent. Women should be counselled strongly against pregnancy, given adequate contraceptive advice, and be advised to undergo therapeutic termination if contraception fails.

Management of pregnant women with the Eisenmenger syndrome is difficult and requires expertise, not least because data to support treatment decisions are sparse.

The fixed pulmonary vascular resistance and reduction in systemic vascular resistance enhance the right to left shunt, increasing maternal and fetal hypoxia. The patient should be admitted for bed rest during the second trimester and meticulous care given to avoid deep venous thrombosis, dehydration, and vasodilation. Oxygen therapy, nitric oxide, and heparin may be given, although there is no firm evidence that any are beneficial, and heparin increases the risk of haemorrhage. Abrupt changes in systemic vascular resistance or blood pressure during labour and delivery may induce intense cyanosis and death. Opinion is divided as to whether vaginal delivery with forceps assistance or caesarean section should be performed. Vaginal delivery may be safer, causing less rapid haemodynamic changes and less blood loss. The risk of sudden death continues for the first 2 weeks after delivery, either from deteriorating haemodynamics or pulmonary infarction.

Bacterial endocarditis

Endocarditis is discussed elsewhere; the risks for specific congenital lesions are outlined in Table 15.

---

*We acknowledge the pioneering work of Jane Somerville in the field of adult congenital heart disease.

Further reading

Reviews and books on congenital heart disease

Anderson RH, Becker AE (1997). *Controversies in the description of congenitally malformed hearts.* Imperial College Press, London.

Ho SY *et al.* (1995). *Colour atlas of congenital heart disease. Morphological and clinical correlations.* Times Mirror Publications Mosby-Wolfe, London.

Kirklin JW, Barratt-Boyes BG (1993). *Cardiac surgery*, 2nd edn. Churchill Livingstone, New York.

Perloff JK (1994). The clinical recognition of congenital heart disease. In: Perloff JK, ed. *Congenital heart disease in adults*, pp 293–380. WB Saunders, Philadelphia.

Perloff JK, Child JS (1998). *Congenital heart disease in adults*. WB Saunders, Philadelphia.

Stark J, de Leval MR, eds (1994). *Surgery for congenital heart defects*. WB Saunders, London.

Warnes CA (1997). Cyanotic congenital heart disease. In: Oakley C, ed. *Heart disease in pregnancy*, pp 83–96. British Medical Journal Publishing Group, London.

Introduction and genetics

Burn J *et al*. (1998). Recurrence risks in offspring of adults with major heart defects: results from first cohort of British collaborative study. *Lancet* **351**, 311–6.

Digilio MC *et al*. (1997). Recurrence risk figures for isolated tetralogy of Fallot after screening for 22q11 microdeletion. *Journal of Medical Genetics* **34**, 188–90.

Farrell MJ *et al*. (1999). HIRA, a di George syndrome candidate gene, is required for outflow tract septation. *Circulation Research* **84**, 127–35.

Kelly D *et al.* (1993). Confirmation that the velo-cardio-facial syndrome is associated with haplo-insufficiency of genes at chromsome 22q11. *American Journal of Medical Genetics* **45**, 308–12.

Li QY *et al.* (1997). Holt–Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nature Genetics* **15**, 21–9.

MacMahon B, McKeown T, Record RG (1953). The incidence and life expectation of children with congenital heart disease. *British Heart Journal* **15**, 121–9.

Nora JJ (1968). Multifactorial inheritance hypothesis for the etiology of congenital heart diseases: the genetic–environmental interaction. *Circulation* **38**, 604–17.

Trainer AH *et al.* (1996). Chromosome 22q11 microdeletion in tetralogy of Fallot. *Archives of Disease in Childhood* **74**, 62–3.

Eisenmenger syndrome

Ammash N, Warnes CA (1996). Cerebrovascular events in adult patients with cyanotic congenital heart disease. *Journal of the American College of Cardiology* **28**, 768–72.

Bowyer JJ *et al.* (1986). Effect of long term oxygen treatment at home in children with pulmonary vascular disease. *British Heart Journal* **55**, 385–90.

Harinck E *et al.* (1996). Air travel and adults with cyanotic congenital heart disease. *Circulation* **93**, 272–6.

Martïnez-Lavín M (1997). Hypertrophic osteoarthopathy. *Current Opinion in Rheumatology* **9**, 83–6.

Maurer HM *et al*. (1975). Correction of platelet dysfunction and bleeding in cyanotic congenital heart disease by simple red cell volume reduction. *American Journal of Cardiology* **35**, 831–5.

Rosenzweig EB, Kerstein D, Barst RJ (1999). Long term prostacyclin for pulmonary hypertension with associated congenital heart defects. *Circulation* **99**, 1858–65.

Thorne SA (1998). Management of polycythaemia in adults with cyanotic congenital heart disease. *Heart* **79**, 315–16.

Wood P (1958). The Eisenmenger syndrome: or pulmonary hypertension with reversed central shunt. *British Medical Journal* **ii**, 701–9, 755–62.

Specific forms of congenital heart disease

Anomalies of venous drainage and atrial arrangement

Dupuis C *et al.* (1992). The 'adult' form of the scimitar syndrome. *American Journal of Cardiology* **70**, 502–7.

Van Mierop LHS, Eisen S, Schiebler GL (1970). The radiographic appearance of the tracheobronchial tree as an indicator of visceral situs. *American Journal of Cardiology* **26**, 432–5.

Van Mierop LHS, Gessner IH, Schiebler GL (1972). Asplenia and polysplenia syndromes. *Birth Defects* **8**, 36–44.

Atrial septal and atrioventricular canal defects

Campbell M (1970). Natural history of atrial septal defect. *British Heart Journal* **32**, 820–6.

Cherian G *et al.* (1983). Pulmonary hypertension in isolated atrial septal defect. *American Heart Journal* **105**, 952–7.

Clapp S *et al.* (1990). Down's syndrome, complete atrioventricular canal and pulmonary vascular obstructive disease. *Journal of Thoracic and Cardiovascular Surgery* **100**, 115–21.

Cooney TP, Thurlbeck WM (1982). Pulmonary hypoplasia in Down's syndrome. *New England Journal of Medicine* **307**, 1170–3.

Dalen JE, Bruce RA, Cobb LA (1962). Interaction of chronic hypoxia of moderate altitude on pulmonary hypertension complicating defect of the atrial septum. *New England Journal of Medicine* **266**, 272–7.

Gatzoulis MA *et al.* (1999). Atrial arrhythmia after surgical closure of atrial septal defects in adults. *New England Journal of Medicine* **340**, 839–46.

Helber U *et al*. (1997). Atrial septal defect in adults: cardiopulmonary exercise capacity before and 4 months and 10 years after defect closure. *Journal of the American College of Cardiology* **29**, 1345–50.

Konstanides S *et al.* (1995). A comparison of surgical and medical therapy for atrial septal defects in adults. *New England Journal of Medicine* **333**, 469–73.

Murphy JG *et al*. (1990). Long term outcome after surgical repair of isolated secundum atrial septal defect: follow up at 27–32 years. *New England Journal of Medicine* **323**, 1645–50.

Schamroth CL *et al.* (1987). Pulmonary arterial thrombosis in secundum atrial septal defect. *American Journal of Cardiology* **60**, 1152–6.

Lesions affecting ventricular inflow

Gentles TL *et al.* (1992). Predictors of longterm survival with Ebstein's anomaly of the tricuspid valve. *American Journal of Cardiology* **69**, 377–81.

Shiina A *et al.* (1983). Two-dimensional echocardiographic–surgical correlation in Ebstein's anomaly: preoperative determination of patients requiring tricuspid valve plication vs replacement. *Circulation* **68**, 534–44.

Theodoro DA *et al*. (1998). Right-sided maze procedure for right atrial arrhythmias in congenital heart disease. *Annals of Thoracic Surgery* **65**, 149–54.

Van Arsdell GS *et al.* (1996). Superior vena cava to pulmonary artery anastomosis: an adjunct to biventricular repair. *Journal of Thoracic and Cardiovascular Surgery* **112**, 1143–8.

Other right ventricular anomalies

Blake RS *et al.*(1982). Conduction defects, ventricular arrhythmias, and late death after surgical closure of ventricular septal defect. *British Heart Journal* **47**, 305–15.

Campbell M (1971). Natural history of ventricular septal defect. *British Heart Journal* **33**, 246–57.

Chaturvedi RR, Shore DS, Redington AN (1996). Intraoperative apical ventricular septal defect closure using a modified Rashkind double umbrella. *Heart* **76**, 367–9.

Kumar K, Lock JE, Geva T (1997). Apical muscular ventricular septal defects between the left ventricle and the right ventricle infundibulum: diagnostic and interventional considerations. *Circulation* **95**, 1207–13.

Lue HC (1986). Is subpulmonic ventricular septal defect an Oriental disease? In: Lue HC, Takao A, eds. *Subpulmonic ventricular septal defect*, 1st edn, pp 3–8. Springer-Verlag, Tokyo.

Moe DG, Guntheroth WG (1987). Spontaneous closure of uncomplicated ventricular septal defect. *American Journal of Cardiology* **60**, 674–8.

Tetralogy of Fallot

Blalock A, Taussig HB (1945). Surgical treatment of malformations of the heart in which there is pulmonary stenosis or pulmonary atresia. *Journal of the American Medical Association* **128**, 189–202.

Bricker JT (1995). Sudden death and tetralogy of Fallot. *Circulation* **92**, 162–3.

Bull K *et al.* (1995). Presentation and attrition in complex pulmonary atresia. *Journal of the American College of Cardiology* **25**, 491–9.

Gatzoulis MA *et al.* (1995). Mechanoelectrical interactions in tetralogy of Fallot. *Circulation* **92**, 231–7.

Kirklin JW *et al.* (1988). Survival functional status and reoperations after repair of tetralogy of Fallot with pulmonary atresia. *Journal of Thoracic and Cardiovascular Surgery* **96**, 102–16.

Murphy JG *et al.* (1993). Long-term outcome in patients undergoing surgical repair of tetralogy of Fallot. *New England Journal of Medicine* **329**, 593–9.

Redington AN, Somerville J (1996). Stenting of aortopulmonary collaterals in complex pulmonary atresia. *Circulation* **94**, 2479–84.

Left ventricular outflow tract obstruction and aortic coarctation

Campbell M (1999). Natural history of coarctation of the aorta. *British Heart Journal* **32**, 633–40.

Carvlho JS *et al*. (1990). Continuous wave Doppler echocardiography and coarctation of the aorta: gradients and flow patterns in the assessment of severity. *British Heart Journal* **64**, 133–7.

Koller M, Rothlin M, Senning Å; (1987). Coarctation of the aorta: review of 362 operated patients. Long term follow up and assessment of prognostic variables. *European Heart Journal* **8**, 670–9.

Roberts CS, Roberts WC (1991). Dissection of the aorta associated with congenital malformation of the aortic valve. *Journal of the American College of Cardiology* **17**, 712–16.

Wells WJ *et al.* (1996). Repair of coarctation of the aorta in adults: the fate of systolic hypertension. *Annals of Thoracic Surgery* **61**, 1168–71.

Transposition of the great arteries

Cochrane AD, Karl TR, Mee RBB (1993). Arterial switch for late failure of the systemic right ventricle. *Annals of Thoracic Surgery* **56**, 854–61.

Imai Y *et al*. (1994). Ventricular function after anatomic repair in patients with atrioventricular discordance. *Journal of Thoracic and Cardiovascular Surgery* **107**, 1272–83.

Reddy VM *et al.* (1997). The double switch procedure for anatomical repair of congenitally corrected transposition of the great arteries in infants and children. *European Heart Journal* **18**, 1470–7.

Redington AN *et al*. (1998). *The right heart in congenital heart disease*. Greenwich Medical Media, London.

Yagihara T *et al.* (1994). Double switch operation in cardiac anomalies with atrioventricular and ventriculoarterial discordance. *Journal of Thoracic and Cardiovascular Surgery* **107**, 31–8.

Univentricular atrioventricular connection and the Fontan operation

Cromme-Dijkhuis AH *et al.* (1990). Coagulation factor abnormalities as possible thrombotic risk factors after Fontan operations. *Lancet* **336**, 1087–90.

de Leval M *et al.* (1988). Total cavopulmonary connection: a logical alternative to atriopulmonary connection for complex Fontan operations. *Journal of Thoracic and Cardiovascular Surgery* **96**, 682–5.

Driscoll DJ *et al.* (1992). Five to fifteen year follow-up after Fontan operation. *Circulation* **85**, 469–96.

Feldt RH *et al.* (1996). Protein-losing enteropathy after the Fontan operation. *Journal of Thoracic and Cardiovascular Surgery* **112**, 672–80.

Fontan F *et al.* (1990). Outcome after a 'perfect' Fontan operation. *Circulation* **81**, 1520–36.

Fontan F, Baudet E (1972). Surgical repair of tricuspid atresia. *Thorax* **26**, 240–8.

Glenn WW (1958). Circulatory bypass of the right heart. IV. Shunt between superior vena cava and distal right pulmonary artery—report of clinical application. *New England Journal of Medicine* **259**, 117.

Mavroudis C *et al.* (1998). Fontan conversion to cavopulmonary connection and arrhythmia circuit cryoablation. *Journal of Thoracic and Cardiovascular Surgery* **115**, 547–56.

Moodie DS *et al.* (1984). Long term follow up in the unoperated univentricular heart. *American Journal of Cardiology* **53**, 1124–8.

Thorne SA *et al*. (1999). Amiodarone-associated thyroid dysfunction in adults with congenital heart disease. *Circulation* **100**, 149–54.

Coronary artery anomalies

Roberts WC (1986). Major anomalies of coronary arterial origin seen in adulthood. *American Heart Journal* **111**, 941–62.

Pregnancy and congenital heart disease

Canobbio MM *et al.* (1996). Pregnancy outcomes after Fontan repair. *Journal of the American College of Cardiology* **28**, 763–7.

Clarkson PM *et al.* (1994). Outcome of pregnancy after the Mustard operation for transposition of the great arteries with intact ventricular septum. *Journal of the American College of Cardiology* **24**, 190–3.

Connelly HM, Grogan M, Warnes CA (1999). Pregnancy among women with congenitally corrected transposition of the great arteries. *Journal of the American College of Cardiology* **33**, 1692–5.

Presbitero P *et al.* (1994). Pregnancy in cyanotic congenital heart disease. *Circulation* **89**, 2673–6.

Yentis SM, Steer P, Plaat F (1998). Eisenmenger's syndrome in pregnancy: maternal and fetal mortality in the 1990s. *British Journal of Obstetrics and Gynaecology* **105**, 921–2.

15.14.1 Thoracic aortic dissection

B. Gribbin and A. P. Banning

Introduction

Acute dissection of the thoracic aorta is uncommon: approximately 20 to 40 cases may be seen in a specialist cardiac unit each year. Unrecognized and untreated it carries a mortality of up to 2 per cent per hour and 90 per cent within the first few weeks.The catastrophic and potentially lethal nature of the dissection process means that quick recognition and treatment are fundamental. Non-invasive diagnostic imaging must be rapid, safe, and accurate. For patients with confirmed dissection of the ascending aorta, emergency surgery may be lifesaving, but when the ascending aorta is spared, aggressive control of blood pressure is the usual initial management, with surgery being considered if there is evidence of further progression of dissection or ischaemic complications.

Despite advances in aortic imaging, surgery, and anaesthesia, survivors of acute dissection have an uncertain long-term prognosis. Careful medical follow-up is recommended and serial aortic imaging can provide important prognostic information, but optimal management of survivors remains controversial.

Pathogenesis

The aortic wall is composed of three layers: a thin intimal lining; a thicker medial layer, largely composed of elastin fibres that provide strength; and a thinner adventitial outer layer from which small blood vessels, the vasa vasorum, arise to nourish the outer layers of the media. Dissection occurs when a breach in the integrity of the intima allows blood at high pressure to penetrate the media. Through this tear, pulsatile blood flow can then propagate distally, parallel to the lumen, often spiralling and splitting the arterial wall into an inner (intima–medial) and outer layer (media–adventitial). This process of tearing within the wall results in the formation of a false lumen, parallel to the original true lumen, and commonly of similar size (Fig. 1 and Plate 1). Further communication(s) between the lumens (or re-entry tears) can occur and may reduce the pressure within the false lumen thus limiting propagation of the dissection. However, the process often extends along the entire length of the aorta to the common iliac artery, threatening the origins of branch vessels which may be avulsed or narrowed by the mass effect of the false lumen, leading to ischaemia in the dependent vascular territories. When dissection extends retrogradely towards the heart it can cause occlusion of a coronary artery and distortion of the aortic valve causing aortic regurgitation. It may also rupture into the pericardial space causing tamponade. The weakened aortic wall can rupture at any point along its length: this is usually fatal.

Fig. 1 Post-mortem specimen of aortic dissection. The intimal/medial flap is pulled back with a retractor to show the false lumen parallel to the true lumen. (See also Plate 1.)

Classification

The commonest sites for thoracic aortic dissection to begin are in the ascending aorta just above the sinuses of the aortic valve and in the upper descending aorta just beyond the origin of the left subclavian artery. Two classifications are used to describe the extent of aortic involvement. De Bakey and colleagues described a classification of dissection that is primarily anatomical and involves three groups: type 1—involving the ascending aorta and arch, with or without involvement of the descending aorta; type 2—involving the ascending aorta alone, without involvement of the arch or descending aorta; and type 3 —involving only the descending aorta. When dissection involves the ascending aorta, emergency surgery is the usual treatment, whereas medical treatment is the initial treatment for patients with uncomplicated dissection sparing the ascending aorta. Therefore, using the De Bakey classification, types 1 and 2 dissection would be considered for surgery.

The Stanford group have subsequently proposed a classification that is directly linked to patient management. Dissection involving the ascending thoracic aorta is classified as type A and demands immediate surgery, whereas dissection which spares the ascending aorta is classified as type B and initial management is usually medical. This classification is recommended as it is unambiguous (Fig. 2).

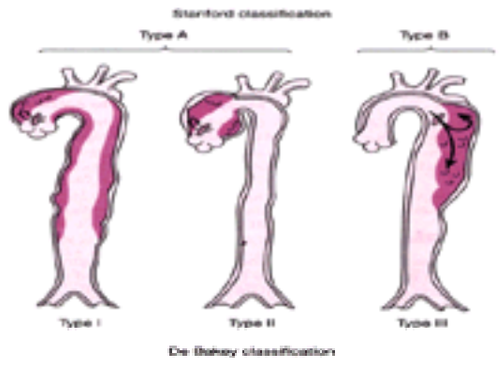

Fig. 2 Diagram of the Stanford and De Bakey classifications of aortic dissection.

Aetiology

The most common predisposing factor is hypertension, which is present in the great majority of patients with aortic dissection. Although the processes involved in the initiation of dissection remain incompletely understood, medial haemorrhage from rupture of vasa vasorum appears to be important. When this process is self-limiting and there is no expansion of the resultant haematoma by recurrent bleeding, healing may occur with reabsorption of the haemorrhage. Alternatively, and particularly when the bleeding is extensive or recurrent, a large intramural haematoma may form around the circumference of the aorta. This alters the distribution of tensile stresses within the aorta, with much of the redistributed stress affecting the intima/endothelium overlying the mass. An intimal tear may then result in splitting and separation of the media, propagation of a false lumen, and dissection.

Patients with the Marfan syndrome may present with aortic dissection or aortic root dilatation (see Chapter 15.12). It is proposed that abnormal fibrillin within the aortic media results in intimal instability, particularly when aortic dilation leads to increased wall stress. Although the absolute risk of dissection rises with increasing size of the ascending aorta, it is important to remember that all patients with Marfan syndrome are at risk, particularly when there is a family history of dissection. Patients with Ehlers–Danlos syndrome are also at risk of spontaneous dissection, not only of the aorta but of its principal branches, including the coronary arteries.

Patients with coarctation of the aorta and those with bicuspid aortic valves appear to be at increased risk of dissection: it is uncertain if this association is related to a defect in aortic wall composition. Dissection may also occur in patients with Turner syndrome, Noonan syndrome, and in the later stages of pregnancy, particularly in patients with Marfan syndrome. In high-risk patients with Marfan syndrome with dilated aortas or a family history of dissection, deferring pregnancy until after elective aortic root replacement may be advisable.

Clinical presentation

Most patients present with characteristic symptoms and clinical findings, in which case the diagnosis of dissection can be made with reasonable assurance. However, a minority present atypically and it is worth considering the possibility of aortic dissection in any patient who is haemodynamically unstable without satisfactory explanation.

The pain of acute dissection of the aorta can be described in terms of (i) its instantaneous onset, (ii) its cataclysmic severity, (iii) its pulsatile and tearing quality, (iv) its location either in the anterior thorax or back, and (v) its migration as it follows the course of the dissection through the thorax. Careful interrogation about the presence of these five features will usually allow differentiation from other causes of chest pain. The instant onset, tearing/pulsatile quality, and migratory pattern contrasts particularly with the pain of cardiac ischaemia, which is usually gradual in onset, tight or crushing, and more unchanging in its distribution in the anterior chest. Syncope shortly after the onset of typical pain is not common but is another characteristic presentation of dissection, often caused by rupture of the false lumen into the pericardial cavity. Other uncommon modes of presentation include stroke and limb ischaemia with or without pain and very occasionally congestive heart failure resulting from severe aortic regurgitation.

Although patients with dissection usually appear shocked, their blood pressure may be normal or raised and their heart rate relatively slow. The distribution of the abnormalities detected by physical examination usually reflects the region of the aorta involved in the dissection and pressure on adjacent structures. Signs of aortic regurgitation or tamponade are likely to be found in a patient with dissection involving the ascending aorta, whereas absent upper limb pulses and cerebral abnormalities suggest involvement of the aortic arch. Expansion of the arch may compress venous return and cause engorgement of one or both jugular veins. Similarly, hoarseness and Horner's syndrome can follow pressure on the left recurrent laryngeal nerve and superior cervical ganglion, respectively. Tenderness over a carotid artery may be due to dissection extending up the artery from the arch. Involvement of the descending aorta can result in visceral and lower limb ischaemia.

Although traditional teaching emphasizes the relevance of blood pressure discrepancy between the arms, this is not a particularly sensitive sign, particularly when dissection spares the ascending aorta and arch. However, evidence of new aortic regurgitation or development of pulse deficits are specific signs of dissection and should be actively sought by the examining physician.

Abnormalities of the chest radiograph and electrocardiogram are not uncommon in patients with dissection, but neither investigation is diagnostic and further imaging is always necessary. Potential abnormalities on the chest radiograph include tracheal deviation, left pleural effusion, a widened mediastinum, and the 'calcium sign' (Fig. 3).

Fig. 3 Chest radiograph showing calcium in the aortic knuckle, which is displaced medially (arrows).

Non-specific ST-segment and T-wave changes on the electrocardiogram are often found, as are changes related to previous hypertension. Actual involvement of a coronary artery is relatively uncommon although the right coronary artery is more likely to be affected. An atypical distribution of ischaemic changes (i.e. generalized acute changes, not consistent with just one coronary territory) is more usual and should always alert the physician to the possibility of a diagnosis other than acute myocardial infarction and the danger of the inadvertent administration of thrombolytic treatment.

Emergency management

Lowering systolic blood pressure and limiting shear stress reduces the likelihood of progression of dissection. Every patient with a clinical suspicion of dissection should therefore receive effective pain relief (intravenous morphine is usually required) and antihypertensive medication pending a definitive diagnosis by imaging. Patients should be cared for in a high dependency area with continuous monitoring of the electrocardiogram and regular blood pressure and urine output measurement. Ideally, systolic blood pressure should be maintained below 110 mmHg, using intravenous labetalol (initial dose 1 mg/min) or esmolol. Both of these agents produce a rapid and titratable reduction in blood pressure, with b-blockade particularly appropriate in this context as it reduces the force of cardiac contraction and the rate of rise of the arterial pressure. If blood pressure control remains suboptimal, an additional infusion of sodium nitroprusside may be used (0.5 to 8 µg/kg.min).

Optimal management of patients with dissection requires close liaison between district hospitals and cardiac surgical centres, using local guidelines for investigation that should reflect the available expertise and surgical opinion. Patients with a low clinical index of suspicion of dissection who are in a stable cardiovascular state should undergo prompt investigation in their local hospital, using a nominated non-invasive technique—usually CT scanning. Unless non-invasive imaging is available immediately, unstable patients with a high clinical index of suspicion should receive medical treatment and be transferred immediately to a surgical centre for both diagnostic imaging and management. This approach minimizes delay, a critical aspect of the management of acute aortic dissection.

Imaging

The priorities when imaging a patient with suspected dissection are to confirm the diagnosis and to decide if the ascending aorta is involved (Stanford type A) as this

will determine whether or not surgery is required. The surgeon wants to know the entry site of the dissection, if the aortic valve is competent, if there is a pericardial effusion or tamponade, and if there is involvement of the coronary arteries. Several diagnostic techniques are available.

Historically, aortography was the investigation of choice, but it has several disadvantages. These include delay during the assembly of the catheter laboratory team, the risk of aortic rupture during catheter manipulation, and the nephrotoxicity of radiological contrast media when renal function may already be compromised by hypotension or renal artery involvement. Computed tomography (**CT**), magnetic resonance imaging (**MRI**), and echocardiography all have proven advantages over aortography.

Contrast-enhanced CT is non-invasive, but requires the use of radiological contrast medium. Its sensitivity and specificity is at least equivalent to aortography, but its accuracy is inferior to both MRI and transoesophageal echocardiography, although improved diagnostic accuracy has been demonstrated by ultrafast and spiral CT.

MRI is non-invasive and provides excellent images of the whole aorta. Its sensitivity and specificity for dissection is up to 100 per cent in some series, and the addition of cardiac gated and 'cine' techniques can give information on luminal blood flow and valvular regurgitation. MRI is therefore the investigation of choice for most diseases affecting the aorta, but it has several limitations in patients with suspected acute dissection of the aorta. These include the requirement for patient transfer to the scanner, with attendant delays, restricted access to the patient during scanning, and the high degree of patient co-operation required to obtain artefact-free images.

Transthoracic echocardiography cannot exclude aortic dissection as it has limited sensitivity and specificity. However, in some patients dissection of the ascending aorta can be confidently diagnosed using parasternal and suprasternal imaging, mandating urgent transfer to a surgical centre where additional information can be obtained by transoesophageal echocardiography in the anaesthetic room.

Transoesophageal echocardiography provides detailed anatomical information about the morphology of a dissection and can also demonstrate the consequences of proximal extension, including the presence of aortic regurgitation, pericardial effusion, and involvement of the coronary artery ostia, thus making complementary investigations such as angiography unnecessary. It can be performed rapidly by a single operator with nursing assistance, in an environment where the patient can be monitored and remains accessible to medical staff. This approach minimizes delay and allows rapid transfer to the operating theatre, making transoesophageal echocardiography the ideal diagnostic tool for the emergency situation (Fig. 4and Plate 2).

Fig. 4 Transoesphageal transverse two-dimensional and colour Doppler echo images of the ascending aorta showing a dissection membrane partitioning the true (TL) and false lumen (FL). Upper left panel shows systolic flow in the true but not the false lumen. (See also Plate 2.)

Surgery for dissection of the ascending aorta

When the dissection involves the ascending aorta (Stanford type A), immediate surgery is required as there is a high risk of proximal extension causing dissection of the coronary arteries, incompetence of the aortic valve, and rupture into the pericardium. Surgery usually involves excision of the intimal tear in the ascending aorta and interposition of a dacron graft. This procedure protects the lower ascending aorta and valve from progressive dissection and prevents distal extension by reducing pressure within the false lumen. The false lumen may subsequently thrombose, or in cases with multiple intimal tears, may remain patent but decompressed.

Replacement of the aortic valve is usually performed only when resuspension of the valve is not possible. However, in patients with Marfan syndrome the ascending aorta and valve are usually replaced with a composite graft to prevent subsequent annular dilatation. In cases where dissection extends into the aortic arch, some surgeons advocate that the arch and great vessels should be included in the initial repair as arch involvement is a strong predictor of a requirement for repeat surgery. However, this extended surgery increases the duration of the operation and the risk of central nervous system damage, hence inclusion of the arch in dissection repair is generally restricted to centres with particular expertise.

Management of descending aortic dissection

Proximal extension towards the heart is less likely when the dissection begins distal to the left subclavian artery (Stanford type B). These patients tend to be older than those with ascending aortic involvement and are more likely to have comorbidity. Diligent blood pressure management is the usual initial treatment as surgery upon the descending thoracic aorta carries significant mortality and morbidity, including impaired blood supply to the spinal cord and paraplegia.

This approach is not universally accepted, however, and some centres recommend elective surgery (after several weeks) in selected patients with Marfan syndrome, in younger patients with dissection associated with large aneurysms, and if thrombosis of the false lumen fails to occur. In addition, surgery for type B dissection should always be considered if there is evidence of proximal extension of the dissection, progressive aortic enlargement threatening external rupture, or ischaemic complications from involvement of major arteries. For example, the prognosis is extremely poor when ischaemia occurs in the territory of a major abdominal artery, in which case emergency surgical fenestration of the intimal flap can be lifesaving.

Encouraging results have recently been achieved using endovascular stenting for patients with complicated dissection starting distal to the left subclavian artery. Using vascular access from a groin incision, a covered stent can be delivered to cover the intimal tear. In suitable cases this obliterates flow into the false lumen, relieving branch ischaemia and preventing further aneurysmal dilatation.

Follow-up of patients with aortic dissection

Strenuous efforts to control blood pressure are indicated for all patients who have survived aortic dissection. b-Blocking drugs are the agents of choice for most, with other agents added as required.

Despite advances in the medical and surgical management of patients with aortic dissection, their long-term prognosis remains uncertain, with several adverse risk factors identified. These include new or progressive dissection, aneurysm formation at the site of surgical anastomosis, and persistence of flow in the false lumen, the last seemingly related to the presence of multiple intimal tears that allow communication between the lumens. Management is controversial, but it has been suggested that in the chronic situation, endovascular stenting may have a role in sealing these communications, thus allowing thrombosis of the false lumen.

Imaging at least once a year is recommended, using the modality with which there is most local expertise. Increased frequency of imaging is recommended following any acute event, for example severe chest pain, and for some patients with Marfan syndrome.

Modern imaging techniques have shown that variants of acute aortic dissection occur. They present in much the same way as classic dissection and may be considered part of the acute aortic syndromes. They include spontaneous intramural haematoma and penetrating atherosclerotic ulcer.

Spontaneous intramural haematoma

Spontaneous intramural haematoma was described by pathologists in 1920. It occurs when the small arterioles (vasa vasorum) which run in the outer media of the aorta rupture and bleed. As it is a medial/adventitial event, the intima remains intact and there is no false lumen. The clinical presentation may mimic that of dissection and the diagnosis can only be made by exclusion of an intimal tear or a penetrating atherosclerotic ulcer. The intramural haematoma is not readily identifiable by aortography, but using non-invasive imaging, a circular or crescentic thickening of the aortic wall of more than 0.7 cm in depth ssociated with central displacement of any intimal calcification supports the diagnosis (Fig. 5).

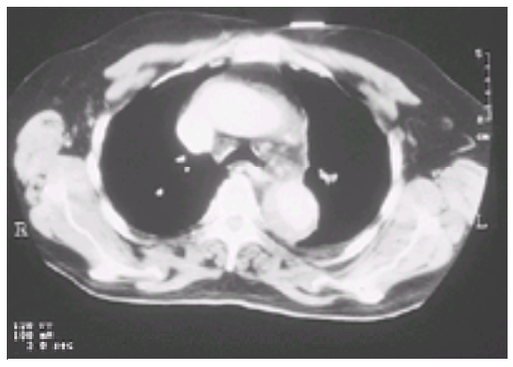

Fig. 5 CT of the thorax showing crescentric thickening of the posterior wall of the descending aorta (adjacent to the vertebra) without compromise of the aortic lumen. Transoesophageal echocardiography showed no intimal flap. The diagnosis is spontaneous haematoma of the aortic wall.

As outlined earlier, there is increasing evidence that spontaneous intramural haematoma may be a precursor of aortic dissection. Clinical studies have supported this assertion: despite aggressive blood pressure control, up to 50 per cent of patients with an intramural haematoma develop dissection or aortic rupture and many now regard this condition as an indication for surgery when the ascending aorta is involved.

Penetrating atherosclerotic ulcer

Penetrating atherosclerotic ulcer presents with similar symptoms to aortic dissection, usually in patients with disseminated atheroma. Intimal disruption caused by atheroma results in perforation and secondary haemorrhage into the media. Imaging demonstrates an out-pouching from the lumen into the aortic wall with localized haemorrhage and evidence of diffuse atheroma.(Fig. 6). Rarely, this can cause a localized dissection, but the main threat is the high incidence of rupture. Treatment is currently surgical, but in the future endovascular stenting may be useful.

Fig. 6 Transoesophageal two-dimensional echo image of a penetrating ulcer (arrow).

Further reading

Dake MD *et al*. (1999). Endovascular stent graft placement for the treatment of acute aortic dissection. *New England Journal of Medicine* **340**, 1546–52. [Original paper describing this novel treatment for dissection.]

Davies MJ, Treasure T, Richardson PD (1996). The pathogenesis of spontaneous arterial dissection. *Heart* **75**, 434–5. [Review of the pathological processes which may result in arterial dissection.]

Erbel R *et al*. (1993). Effect of medical and surgical therapy on aortic dissection evaluated by transoesophageal echocardiography. *Circulation* **87**, 1604–15. [Original paper which outlines relationship between echocardiographic appearances and long-term prognosis.]

Khandheria BK (1993). Aortic dissection; the last frontier. *Circulation* **87**, 1765–8. [Editorial discussing imaging and management of dissection.]

Kouchoukos NT, Dougenis D (1997). Surgery of the thoracic aorta. *New England Journal of Medicine* **336**, 1876–88. [Authoritative surgical review of the literature.]

Miller DC (1993). The continuing dilemma concerning medical versus surgical management of patients with acute type B dissections. *Seminars in Thoracic and Cardiovascular Surgery* **5**, 33–46. [Editorial which outlines the principles and controversies of management of type B dissection.]

Nienaber CA *et al*. (1993). The diagnosis of thoracic aortic dissection by non-invasive imaging procedures. *New England Journal of Medicine* **328**, 1–9. [Original paper comparing different diagnostic imaging techniques in dissection.]

Robbins RC *et al*. (1993). Management of patients with intramural haematoma of the thoracic aorta. *Circulation* **88**, 1–10. [Original paper which describes the diagnosis of intramural haematoma and outlines management strategies.]

Vilacosta I *et al*. (1998). Penetrating atherosclerotic ulcer: documentation by transoesophageal echocardiography. *Journal of the American College of Cardiology* **32**, 83–9. [Original paper which describes the pathology, diagnosis, and treatment of atherosclerotic ulcer.]

Wooley CF, Sparks EH, Boudoulas H (1998). Aortic pain. *Progress in Cardiovascular Diseases* **40**, 563–89. [Detailed review of the clinical presentation and management of aortic pathology.]

15.14.2 Peripheral arterial disease

Janet Powell and Alun Davies

Introduction

Peripheral arterial disease, defined for the purpose of this chapter as diseases of the abdominal aorta and its branches, has risk factors and features that overlap with, but can be distinguished from, those of coronary artery disease. The two conditions often coexist, but patients with coronary disease are almost always referred directly to physicians, whilst those with peripheral arterial disease are referred directly to vascular surgeons, particularly in regions where angiology is a poorly developed specialty, since there are few effective medical therapies. Vascular surgeons also manage patients with arterial disease in the carotid vessels and upper limbs. These aspects will receive only passing mention in this chapter: for discussion regarding the clinical features and management of carotid artery disease, see Section 15.14.

The most common presentations of peripheral arterial disease are intermittent claudication and abdominal aortic aneurysm. Most peripheral arterial disease remains asymptomatic. It is not a new disease that results from a modern Westernized lifestyle. Atherosclerotic disease, partially occluding the peripheral arteries, has been described in the mummies of ancient Egypt. Life as a cavalry officer was associated with an increased risk of popliteal aneurysm, a condition treated by ligation by John Hunter, the pioneering eighteenth-century surgeon. Albert Einstein died of a ruptured abdominal aortic aneurysm.

Techniques for repairing abdominal aortic aneurysms were not developed until the middle of the twentieth century. This was the golden era for the development of vascular surgery as a specialty, with the increasing use of bypass surgery hat has minimized the need for amputation. Today newer, less invasive approaches are being used—angioplasty and endovascular stenting—but few medical therapies are on the horizon.

Aetiology and epidemiology

Peripheral arterial disease may occur in the young but the prevalence increases sharply with age. Both young and old may suffer from occlusive (stenosing) disease of the peripheral arteries or dilating (aneurysmal) disease, while vasospastic disease is uncommon. However, the underlying causes of peripheral arterial disease in those below and above 50 years of age tend to be very different.

Peripheral arterial disease in patients less than 50 years old

In younger patients the cause of disease is most likely to be genetic, congenital, immunological, infectious, or traumatic. Patients with familial hypercholesterolaemia and related inherited disorders of lipid metabolism may present with peripheral limb ischaemia. There are also congenital causes of early-onset leg ischaemia. These include aortic hypoplasia, which occurs during the embryonic fusion of the distal aortas, and popliteal entrapment, where the popliteal artery takes an unusual course through the head of the gastrocnemius muscle, with exercise involving knee flexion causing intermittent occlusion of the artery and calf pain that resembles intermittent claudication. A fierce immunological inflammatory response to smoking causes Buerger's syndrome, which involves the artery, vein, and associated nerves in both the legs and the arms. This disease, seen principally in men, is particularly prevalent in the Indian subcontinent, and may resolve if the patient stops smoking. Sudden thrombotic occlusion of the iliac and distal arteries may occur in those below 50 years of age, suggesting the presence of an inherited thrombotic disorder. Embolic occlusion from a proximal source is also possible.

Marfan syndrome may sometimes be confirmed (mutation in the fibrillin-1 gene) only after a patient has presented with a ruptured abdominal aortic aneurysm. In some variants of Ehlers–Danlos syndrome, patients with mutations in type III collagen present with visceral artery aneurysms. In South Africa (and elsewhere) aneurysms of the abdominal, femoral, or popliteal arteries in those under 50 years have been attributed to infectious causes, from HIV to tuberculosis. Syphilitic aneurysms, which used to affect principally the thoracic aorta, are now rare.

Peripheral arterial disease in patients over 50 years old

For patients over 50 years of age, the principal risk factor for peripheral arterial disease—stenosing, aneurysmal, or vasospastic—is smoking. The pathology is atherosclerotic change with superimposed thrombosis. Of patients who present with peripheral arterial disease, less than 5 per cent have never smoked. For this reason, more men than women presented with peripheral arterial disease in the past, but recently more women have been presenting with the condition, perhaps a reflection of the increasing number of women who smoke. Nevertheless, unlike Buerger's disease, cessation of smoking is not associated with an immediate dramatic improvement in symptoms and it may take several years without smoking to improve prognosis.

Diabetes is another important risk factor for stenosing peripheral arterial disease. Other risk factors include hypertension, raised levels of plasma fibrinogen, and hyperlipidaemia, with elevated plasma triglycerides being a common finding. The risk factors for dilating arterial disease are similar, with the exception of diabetes, which is rare.

For aortic aneurysms, although strong familial clustering has been observed, no specific genetic mutations associated with aneurysmal disease have been identified and atherosclerotic change is commonplace. Caucasian and northern European populations appear to be at higher risk of aneurysmal disease than black populations. Stenosing and aneurysmal disease are associated with degenerative changes of the artery wall, the prevalence of both diseases increasing sharply with age (Table 1). Epidemi-ological studies also indicate a difference between stenosing and aneurysmal disease, with death from aneurysmal disease (aortic aneurysm) being more common amongst those of higher social classes and in affluent geographical areas.

Clinical features of leg ischaemia

The terms acute and chronic relate purely to the length of time that symptoms have been present and must not be confused with terms related to severity, such as

critical limb ischaemia.

Critical leg ischaemia

Critical leg ischaemia is defined as gangrenous change, ulceration, or rest pain lasting for 2 weeks with an absolute ankle pressure of less than 50 mmHg, although patients with diabetes are difficult to include in this classification.

Acute leg ischaemia

The incidence of acute leg ischaemia, which presents as a painful, pale, and pulseless limb, is 1 in 12 000 patients per annum. It can be due either to an embolic event or thrombosis of an atherosclerotic stenosis. The commonest cause of a peripheral embolus used to be rheumatic heart disease in a patient with atrial fibrillation, but this is becoming less common, and other sources of emboli, such as an aortic aneurysm, must be considered. The development of a thrombosis at the site of an atherosclerotic stenosis, either in the superficial femoral artery or popliteal artery, is undoubtedly now the commonest cause of acute leg ischaemia. However, it should be stressed that, whatever the cause, there is no difference on clinical examination of the acutely ischaemic limb.

Arterial trauma, due to road traffic accidents and knife or gunshot wounds is becoming commoner, as is iatrogenic trauma following the insertion of intra-arterial catheters for diagnosis or therapy. A rare but dramatic cause of acute leg ischaemia is phlegmasia cerulea dolens, in which massive thrombosis of all the major veins of the limb occurs with gross swelling that obstructs the arterial supply.

Patients with a thrombosis of a popliteal aneurysm may present with classic symptoms of pain, paralysis, loss of power, paraesthesia, pallor, lack of pulse, and perishing cold. If the blood supply is not restored, fixed blue staining of the skin is a further sign of irreversible ischaemia, as is a tense calf with plantar flexion. However, the majority of patients presenting with acute ischaemia have symptoms that are less severe.

Chronic leg ischaemia

Chronic leg ischaemia is much more common than acute ischaemia (Table 1) and its main cause is atherosclerosis. In the young patient one should also consider cystic adventitial disease, entrapment of the popliteal artery, and occasionally fibromuscular hyperplasia of the iliac arteries, particularly in women.

Symptoms are pain on walking, claudication affecting the calf and thigh, rest pain, ulceration, and gangrenous change. Less commonly patients may present with buttock claudication and impotence (Leriche's syndrome). Whilst the differential diagnoses of the acutely ischaemic limb are few, in the chronically ischaemic limb pain may be due to nerve root compression or arthritis of the hip or knee. Classically the patient with claudication will complain of cramp-like pain in the calf, appearing after walking a particular distance, relieved by a few minutes rest, and recurring again at the same distance if the patient resumes walking. Failure of the pain to disappear on resting, or its reappearance after a shorter distance after each rest, suggest a possible musculoskeletal cause, particularly if distal pulses are present on examination. However, it should also be remembered that distal pulses may be felt at rest in the limbs of patients with claudication due to peripheral vascular disease, but disappear on exercise to the point of pain.

Investigation of the patient with an ischaemic leg

The main diagnostic method used to confirm the diagnosis of peripheral arterial disease is Doppler ultrasonography (duplex scanning), an example of which is shown in Fig. 1 and Plate 1. The ratio of systolic blood pressure at the ankle and in the arm, ankle–brachial pressure index (**ABPI**), provides a physiological measure of blood flow at the level of the ankle. At rest, in a normal leg, the ABPI lies between 1.0 and 1.4. As the blood flow in the leg is compromised, the ABPI falls sharply and values below 0.9 are considered abnormal and likely to confirm the diagnosis of peripheral vascular disease. To emphasize the important overlap between this condition and coronary artery disease, a reduction in ABPI nearly always signals the presence of coronary artery disease, which is the cause of death in the majority of patients with peripheral arterial disease.

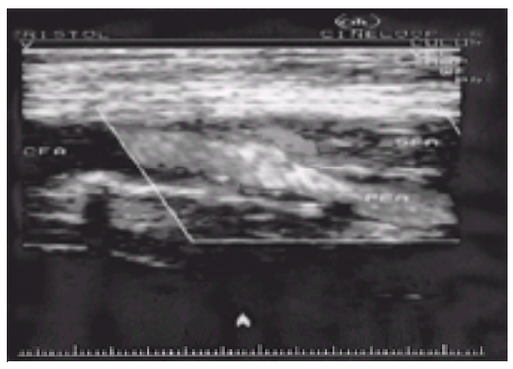

Fig. 1 Occlusion of the superficial femoral artery demonstrated by colour-coded duplex ultrasonography. On the left, the common femoral artery (CFA) lies outside the colour box. In the colour box antegrade flow through the profunda femoris artery (PFA) is shown in blue. The red flash represents rebound flow against the occluded origin of the superficial femoral artery (SFA). (See also Plate 1.)

Exercise testing provides an objective method of assessing walking distance and helps with the identification of disease processes such as angina that may be limiting. It only needs to be used in those people who have a history of claudication but have normal resting ankle–brachial pressure indices, and it can be used as a way of eliminating or suggesting other diagnoses.

In addition to establishing the diagnosis of peripheral arterial disease, duplex ultrasonography is able to determine the site of disease and to indicate the degree of stenosis or length of an occlusion and hence aid in the planning of interventional treatment. Angiography is only required as an adjuvant to endovascular treatment, for surgical planning in some circumstances, or in the management of the acutely ischaemic limb.

Attention to risk factors, in particular smoking and blood pressure, are important issues.

Management of critical and acute limb ischaemia

Critical limb ischaemia requires administration of analgesia and rapid surgical intervention. The severity of ischaemia will determine the treatment options considered. However, all patients with a severely ischaemic limb should be given adequate analgesia and 5000 units of heparin intravenously. Many will be old and frail, with significant medical comorbidities. These issues must be considered in deciding whether or not surgical intervention is appropriate for any individual case, with action taken to improve those aspects of the patient's medical condition that can be improved before surgery, or as part of continuing medical management.

For a patient with irreversible ischaemia (fixed skin staining and tense muscles), the main decision is whether a primary amputation or palliative care should be offered. If severe but potentially reversible ischaemia is present (white leg), surgery is usually the treatment of choice. Delay while thrombolytic therapy is tried is not advisable in this group. For patients with moderate limb ischaemia, where there is no paralysis and only mild sensory loss, arteriography with a view to thrombolysis should be performed. However, it should be remembered that thrombolysis is associated with numerous potential complications, most notably gastrointestinal haemorrhage and stroke. If the limb is salvageable, it may be possible to offer the patient an endovascular procedure, such as an angioplasty (with or without stenting). Surgical treatment can involve simple embolectomy, but may require a bypass procedure or endarterectomy, and in the severely ischaemic limb fasciotomies may be needed to treat or prevent a compartment syndrome. For at least 10 per cent of patients, it will not be possible to offer revascularization. Such patients may benefit from the use of a prostacyclin analogue (Iloprost), which has been shown to reduce amputation rates and alleviate pain. Limb salvage rates for

patients presenting with critical limb ischaemia are variable, probably 50 to 60 per cent at 2 years, dependent on the severity of disease.

In a patient presenting with acute leg ischaemia the outlook is poor with only about 60 per cent leaving hospital with an intact limb. The 30-day mortality for this group of patients can be as high as 30 per cent, the main cause of death being cardiac disease. The strategy for management is described in Fig. 2. The controversies that exist in the treatment of acute leg ischaemia are mainly related to the role of arteriography and which technique of thrombolysis is the safest and most cost effective.

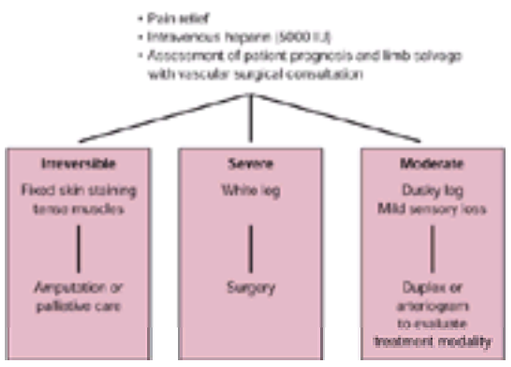

Fig. 2 Management of the acutely ischaemic leg.

In the patient who has had an embolic event, long-term anticoagulation should not be forgotten. Nor should a search for the source of embolus: if the patient is not in atrial fibrillation, has normal cardiac enzymes and 12-lead ECG, then they should have an echocardiogram to exclude any valvular lesion, a 24-h ECG to look for arrhythmia, an ultrasound to exclude abdominal aortic aneurysm, and a screen for thrombophilia.

Management of the chronically ischaemic leg

In chronic limb ischaemia management depends upon the severity of the disease. The vast majority of patients present with claudication, which is relatively benign. Only about 5 per cent of those who have claudication will go on to lose a limb, but claudication identifies patients with a threefold increased risk of death from either heart disease or cerebrovascular disease compared with age- and sex-matched controls. It is important when planning treatment that all the potential risk factors are covered. The general advice is to stop smoking and to exercise, with the use of structured exercise programmes shown to be of benefit. Surgical intervention is not usually required. Over 80 per cent of patients do not require any form of interventional procedure, and at least a third will have improvement of symptoms with simple medical treatment. The management of patients with chronic lower leg pain is shown in Fig. 3.

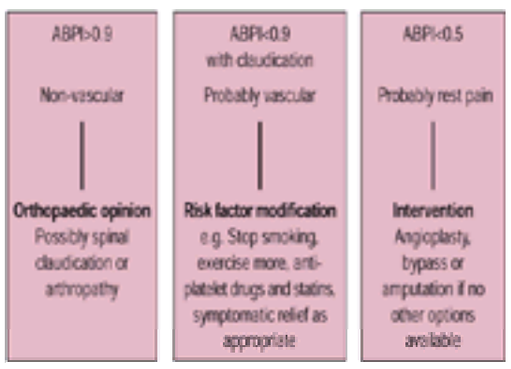

Fig. 3 Management of the patient with chronic lower leg pain.

General management

Careful attention must be paid to the cleanliness of ischaemic feet to avoid infection, and particular care should be given to the cutting of toe nails. In many patients this is best done by a careful younger relative or chiropodist, since apparently minor lacerations can lead to ulcers, infection, and gangrene. Walking to the point of claudication is not harmful, and may improve collateral circulation with beneficial results.

Smoking is by far the most significant risk factor for occlusive arterial disease and every effort should be made to encourage smokers to stop. If patients undergo surgical treatment, then the long-term patency rate following arterial reconstruction is four times greater in smokers who stop than in those who persist.

Medical treatment

Since coronary artery disease is the main cause of death in those with peripheral arterial disease, low-dose aspirin therapy (75 to 325 mg/day) should be recommended for all patients. If aspirin cannot be tolerated, ADP-receptor antagonists, such as clopidrogel, are equally effective in reducing the risk of cardiovascular events (stroke, myocardial infarction, and vascular deaths) and particularly effective for patients with peripheral arterial disease.

Secondary prevention trials have demonstrated the benefits of statin therapy following myocardial infarction. It is likely that similar benefits would be seen for patients with stenosing atherosclerotic disease of the peripheral arteries, particularly in those with elevated serum cholesterol concentrations. The new generation of fibrates, which lower both plasma fibrinogen and triglycerides, may also be effective in reducing cardiovascular events in patients with peripheral arterial disease, but randomized trials have not yet reported. Chelation therapy offers no benefits.

Few patients with peripheral arterial disease are prescribed b-blockers for control of hypertension or angina. However, when these patients require surgery, perioperative cover with a b-blocker is likely to minimize myocardial ischaemia and diminish postoperative morbidity and mortality associated with major vascular surgery.

Surgical treatment

In general, surgeons are becoming more conservative with respect to interventional treatment for patients with claudication, despite a possible early benefit for those having an endovascular procedure. However, in the patient who has severe claudication, with symptoms that significantly affect their quality of life, it is certainly possible and appropriate to offer interventional treatment.

Several issues in the management of chronic limb ischaemia are the subject of ongoing clinical trials. These include comparison of endovascular angioplasty and bypass surgery. For infrainguinal bypass, good-quality autologous vein is the conduit of choice. However, reasonable results can be obtained with synthetic grafts, particularly where the distal anastomosis is above the knee. Below the knee an adjuvant vein interposition in the form of either a Miller cuff or Taylor patch is used. The role of stenting in the leg vessels is contentious, and it may not be of value. The role of exercise therapy compared with angioplasty in the treatment of mild to moderate claudication continues to be debated.

Ischaemia of the arm

Ischaemia of the arm is usually a result of embolism from the heart. Occasionally the subclavian artery is diseased or has suffered traumatic injury or radiation damage following radiotherapy. The basic principles of investigation and management are the same as for the leg. However, it should be noted that the upper limb has multiple interconnection of collateral vessels, hence occlusion of the major arterial supply may still leave a viable limb. The other disease process that needs to be considered is the thoracic outlet syndrome, which gives rise to symptoms in the arm as a result of arterial, venous, or neurological compression caused by an additional cervical rib or by scalene bands. Management may require surgical intervention, either cervical rib excision or thoracic outlet decompression with the removal of the first rib.

Mesenteric ischaemia

Mesenteric ischaemia is uncommon. Over one-third of cases of acute mesenteric ischaemia are due to arterial embolism, with emboli lodging at the ostium of the superior mesenteric artery in many cases. Patients with acute mesenteric artery thrombosis have often had symptoms of mesenteric ischaemia prior to the acute episode. Chronic mesenteric ischaemia typically presents with weight loss and abdominal pain on ingestion of food, the classic story being that the patient is constantly hungry, but frightened to eat. Other causes of acute mesenteric ischaemia include venous thrombosis and non-occlusive ischaemia secondary to hypoperfusion.

Patients with acute mesenteric ischaemia will usually present with abdominal pain, but the abdominal physical signs may be much milder than would be anticipated from the subsequent clinical course. Suspicion of the diagnosis should be heightened in the presence of atrial fibrillation or widespread atheromatous vascular disease. Patients may deteriorate suddenly and present in shock.

The diagnosis of acute mesenteric ischaemia is difficult to make. In the acute situation clues to look for include leucocytosis, hyperamylasaemia, and unexplained acidosis. Liver function tests are usually normal. Radiological imaging is rarely able to make a positive diagnosis, although it can be very useful in excluding other possibilities. Angiography is not always accurate. CT scanning can be helpful in the diagnosis of mesenteric venous thrombosis.

Intensive resuscitation to replace fluids is essential. Surgery is usually necessary for the patient to survive, and the possibility of acute mesenteric ischaemia remains one of the dwindling number of reasons for requiring an emergency diagnostic laparotomy. Depending on the findings, resection of small bowel may suffice, but formal arterial surgery may be necessary, and in some unfortunate instances the extent of irreversible ischaemia can preclude an attempt at resection or revascularization. In cases where the surgeon is unsure of the viability of bowel remaining after resection, a second laparotomy may be planned to assess the situation a few days later. Repeat laparotomy may also be required to examine, and if necessary resect, more bowel in the patient who is not 'doing well' postoperatively. The prognosis for patients who present with acute mesenteric ischaemia is poor.

For patients who present with chronic mesenteric ischaemia, the aim of treatment is to improve blood flow and to act as a prophylactic procedure to prevent the catastrophic disaster of arterial occlusion. The potential options, having identified the site of the disease process by duplex scanning and angiography, include angioplasty, endarterectomy, reimplantation, or a surgical bypass procedure.

Abdominal aortic aneurysm

Definition

There is no fixed definition of an abdominal aortic aneurysm. It is a localized dilatation of the abdominal aorta, usually fusiform, with dilation starting distal to the renal arteries. One definition is when the maximum aortic diameter is more than 1.5 times the diameter of the undilated proximal aorta. Manual palpation to detect abdominal aortic aneurysms is unreliable, unless undertaken by a specialist on non-obese patients. The most convenient method of screening for the presence of these aneurysms is ultrasonography, measuring the anterior–posterior diameter. Since the reproducibility of ultrasound measurements of the suprarenal aorta is poor, a convenient working definition of an abdominal aortic aneurysm is when the maximum diameter exceeds 3 cm, which in most people is more than 1.5 times the diameter of the undilated proximal aorta. In practice, it is only aneurysms of 4 cm or greater in diameter that have been of clinical concern.

Epidemiology

Population screening studies in northern Europe have shown that the disease is usually without symptoms, much more common in men than in women (Table 1), and is strongly associated with smoking. The prevalence of large aneurysms (greater than 5 cm in diameter) detected by screening is only about 1 per cent in men and the large majority of these screen-detected aneurysms are 3 to 5 cm in diameter. The natural history of abdominal aortic aneurysms is progressive enlargement (with the diameter increasing by 2 to 5 mm each year) without symptoms, until the aortic wall is so weakened that it ruptures. Rupture is a catastrophic event.

The infrarenal aorta is by far the most common site of aneurysmal dilatation, and usually the abdominal aorta is the only site of dilatation. When patients present with aneurysms of the iliac, femoral, or popliteal arteries, abdominal aortic aneurysm is often present and screening for this is mandatory. This emphasizes the tendency of some patients to have a more generalized form of dilating arterial disease.

Ruptured aneurysms

The symptoms of a ruptured abdominal aortic aneurysm are collapse (shock) and severe back or abdominal pain.

Rarely a ruptured aneurysm will present with gastrointestinal bleeding from an aortoduodenal fistula or high-output cardiac failure from an aortocaval fistula.

Less than 20 per cent of patients with a ruptured abdominal aortic aneurysm reach hospital alive, and even among those that undergo emergency surgical repair almost half will die within 30 days of surgery. With this bleak prognosis and the very significant costs associated with emergency repair following rupture, it has been suggested that widespread screening to detect those with the largest aneurysms, at highest risk of rupture, would be cost-effective. A screening trial is in progress to assess the benefit of such a policy.

Management of ruptured aneurysms requires:

1. access lines, cross-matched blood, and resuscitation;
2. confirmation of diagnosis—ultrasound (to show aneurysm), CT scan or experienced vascular surgeon (to confirm diagnosis of rupture);
3. rapid assessment of whether patient would benefit from emergency repair; and
4. if yes, immediate surgical repair.

Aneurysms detected before rupture

Abdominal aortic aneurysms are commonly symptomless but rupture is catastrophic. However, elective repair of an abdominal aortic aneurysm, a major surgical procedure, is not without risk. Traditionally, larger aneurysms have been repaired by cross-clamping of the aorta and insertion of a Dacron inlay graft at open surgery. This is a durable procedure and effectively 'cures' the patient. However, although some specialized surgical centres report an operative mortality of less than 2 per cent associated with this elective procedure, on a population basis the mortality is more likely to be 5 to 8 per cent. This very significant surgical mortality is an important reason for avoiding surgery in those with small aneurysms.

Recently, the United Kingdom Small Aneurysm Trial showed that for aneurysms of 4.0 to 5.5 cm in diameter the policy of early elective surgery conferred no long-term survival benefit, and surgery should not be recommended. For such patients surveillance, with measurement of ultrasound diameter every 6 months, is a safe policy which engenders little patient anxiety, and the risk of aneurysm rupture is very low—1 per cent per year. By contrast, for patients with aneurysms greater than 6 cm in diameter the risk of rupture may be as high as 25 per cent per year, and in most such cases elective repair is recommended.

Repair is recommended also when symptoms are attributed to the aneurysm, whatever its size, the commonest being back or abdominal pain, or tenderness to palpation. It is assumed that such aneurysms are at high risk of rupture and need early repair. As the aneurysm dilates, layers of laminated thrombus deposit in the lumen, like onion skins, leaving a blood flow channel of approximately normal aortic diameter. These layers of thrombus are very stable and only in rare circumstances are the source of emboli to the legs. The aneurysms which most often provoke symptoms have very thick, inflamed, fibrotic walls, which entrap nerves and may become adherent to other tissues. These are known as inflammatory aneurysms and the thickened wall can often be detected by CT scan or magnetic resonance imaging. They are technically demanding to repair. There is no convincing evidence that a course of preoperative corticosteroids is beneficial. In the Japanese population inflammatory aneurysms have been associated with active cytomegalovirus infection.

A strategy for the management of abdominal aortic aneurysms detected before rupture is shown in Fig. 4. Patients with small aneurysms should stop smoking and have their blood pressure controlled. Since screening detects mainly small aneurysms, it would clearly be beneficial if a treatment to limit aneurysm growth were available. Although propranolol has proved effective in limiting the dilation of the proximal aorta in patients with Marfan syndrome, as yet there is no evidence that it is effective for abdominal aortic aneurysms. Furthermore, many patients with abdominal aortic aneurysm have impaired lung function, perhaps through smoking, and b-blockers often are poorly tolerated. However, effective control of blood pressure and cessation of smoking are both likely to minimize the rate of aneurysm growth and the risk of rupture. Pathophysiological studies suggest that inflammation and proteolysis are important processes driving aneurysm growth, but there is no firm evidence that either anti-inflammatory or antiproteolytic therapy is effective. Surgery remains the only available treatment for aneurysms larger than 5.5 cm in diameter.

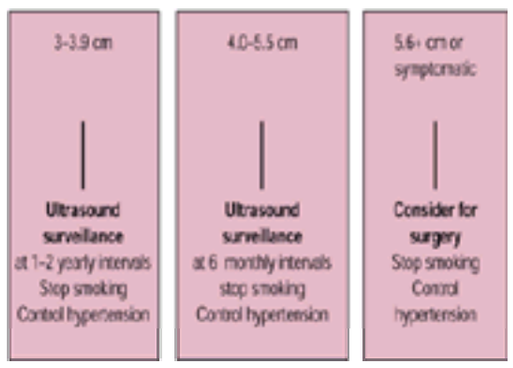

Fig. 4 Management of unruptured abdominal aortic aneurysms depending on their size as demonstrated by ultrasonography.

Recently a less invasive repair procedure has been developed, where an aortic graft is inserted via the femoral artery. This technique remains under development, its durability is being investigated, and several randomized trials are comparing this new endovascular approach with the traditional surgical one.

Conventional surgical management

Preoperative evaluation requires CT or magnetic resonance imaging to define the anatomy and extent of the aneurysm. Cardiac, pulmonary, and renal function should always be assessed, poor renal and lung function being associated with an increased risk of postoperative morbidity and mortality.

The most common surgical approach to an abdominal aortic aneurysm is through a transperitoneal incision under general anaesthesia. The retroperitoneal approach, which avoids bowel manipulation and permits a more rapid return to oral diet, has similar cross-clamp, operating, and recovery times. The transperitoneal approach offers the advantage of exploring the abdominal cavity for other pathology. In this approach, after the bowel has been removed from the operative field, the aorta is exposed anteriorly from the left renal vein to the bifurcation. The infrarenal neck of the aneurysm is exposed anteriorly and laterally so that an occluding clamp may be applied. Both common iliac arteries are exposed for the placement of the distal occluding clamps. With the clamps in place, the aneurysm is opened longitudinally on the anterior surface and the remainder of the procedure performed from inside the aneurysm cavity, usually following a small dose of intravenous heparin. Clot and debris are evacuated and any back-bleeding lumbar or mesenteric arteries ligated. A Dacron prosthesis is then sutured, end to end, to the normal-diameter aorta above the aneurysm. This anastamosis is tested for leaks before the graft is trimmed to appropriate length and sutured in place above the aortic bifurcation. The aneurysm sac is closed over the prosthesis, before replacement of abdominal contents. Such tube grafts are the most common type, but when the iliac arteries are dilated or diseased a bifurcated prosthesis is used. The cross-clamp time should be less than an hour and the whole procedure completed within 2 to 4 h. The longest procedures involve inflammatory aneurysms and cases where the proximal aneurysm neck lies above the renal arteries. The patient should be ready to leave hospital 7 to 12 days after the operation, with a durable repair.

Endovascular aneurysm repair

The technique of endovascular repair was introduced in the early 1990s and the technology is still evolving. The procedure may be performed under general or epidural anaesthesia. This flexibility allows endovascular repair in patients where pulmonary or cardiac function is too poor to consider open repair, and the avoidance of aortic cross-clamping is an additional benefit for those with limited cardiac reserve.

Preoperative investigation to evaluate the extent and size of the aneurysm (spiral CT or magnetic resonance imaging) is of critical importance. The length of the aneurysm neck below the renal arteries, angulation of the aorta, and tortuosity of the iliac arteries must be evaluated precisely so that the correct size of graft can be placed via the femoral artery. The insertion of the graft is performed under fluoroscopic control. This requires the use of significant amounts of contrast material, which may underlie the unfavourable results reported in patients with high serum creatinine. The proximal end of the graft is held in place either by hooks or balloon-expandable stents.

The length of the procedure and the transfusion requirements are similar to those for open surgical repair, but the patient recovers rapidly and is ready to leave hospital within 2 to 5 days. The long-term success of the procedure depends on the successful exclusion of the aneurysm sac and the security of the proximal attachment to prevent graft migration. Endoleaks may develop when the aneurysm is not completely excluded or there is back-bleeding from lumbar vessels or the inferior mesenteric artery into the aneurysm sac. These are associated with an important risk of continued aneurysm expansion and rupture. For these reasons continued vigilance and repeated evaluation of the aneurysm with duplex or CT scanning is necessary at 6-monthly intervals. Currently, neither the durability of endovascular grafts nor their use in patients unfit for open repair has been properly evaluated in clinical trials.

Further reading

CAPRIE Steering Committee (1996). A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events. *Lancet* **348**, 1329–39.

Fowkes FGR (1988). The epidemiology of atherosclerotic arterial disease in the lower limbs. *European Journal of Vascular Surgery* **2**, 283–91.

Mangano DT *et al.* (1996). Effect of atenolol on mortality and morbidity after noncardiac surgery. *New England Journal of Medicine* **335**, 1713–20.

Meijer WT *et al.* (1998). Peripheral arterial disease in the elderly: The Rotterdam Study. *Arteriosclerosis, Thrombosis, and Vascular Biology* **18**, 185–92.

Perkins JMT *et al.* (1996). Exercise training versus angioplasty for stable claudication. Long and medium term results of a prospective, randomised trial. *European Journal of Vascular and Endovascular Surgery* **11**, 409–13.

Tetteroo E *et al*. (1998). Randomised comparison of primary stent placement versus primary angioplasty followed by selective stent placement in patients with iliac artery occlusive disease. Dutch Iliac Stent Trial Group. *Lancet* **351**, 1153–9.

UK Small Aneurysm Trial Participants (1998). Mortality results for randomised controlled trial of early elective surgery or ultrasonographic surveillance for small abdominal aortic aneurysm. *Lancet*

352, 1649–55.

Whyman MR *et al.*(1996). Randomised controlled trial of percutaneous transluminal angioplasty for intermittent claudication. *European Journal of Vascular and Endovascular Surgery* **12**, 167–72.

15.14.3 Cholesterol embolism

C. R. K. Dudley

The clinical features of cholesterol embolism mimic a number of conditions, particularly systemic vasculitis, and if misdiagnosed can result in the inappropriate use of powerful immunosuppressive drugs.

Introduction

When atheromatous plaques ulcerate and become denuded of their endothelial covering, the underlying cholesterol-rich extracellular matrix can become detached and embolize. If the dislodged atheroma is sufficiently large, occlusion of a major systemic artery results in infarction of the organ or ischaemia of the limb supplied. This has been termed 'atheroembolism'. By contrast, cholesterol-crystal embolism occurs when much smaller and more numerous particles, composed principally of cholesterol crystals, lodge in a number of small arteries simultaneously. The presence of a collateral circulation usually prevents infarction and the event frequently passes unrecognized by the patient or their physician. However, tissue damage in a number of organs can result from multiple showers of emboli. Because severe ulcerative atherosclerosis is most frequently present in the abdominal aorta, cholesterol embolism commonly affects the lower limbs, gastrointestinal tract, and kidneys. The condition usually presents as a complication of vascular surgery or angiographic procedures, when mechanical dislodgement of crystals from ulcerated plaques occurs. Anticoagulant and thrombolytic use has also been implicated as a predisposing factor. The clinical features are those of a systemic disorder with renal failure that can mimic vasculitis.

Epidemiology

The incidence of cholesterol-crystal embolism found at postmortem is high: 77 per cent after aortic surgery, 30 per cent after aortography, and 25.5 per cent after cardiac catheterization. By contrast, the clinical syndrome of cholesterol-crystal embolism is rare, complicating less than 2 per cent of cardiac catheterizations.

Since the condition occurs in patients with severe atheromatous disease, it is most often seen in older male patients with obvious risk factors (hypertension, diabetes mellitus, smoking) and overt vascular disease (ischaemic heart disease, abdominal aortic aneurysm, cerebrovascular disease, etc.). Although spontaneous cholesterol embolism can occur, it is much more common after vascular surgery or invasive radiology including aortography, angiography, and angioplasty. Under these circumstances direct trauma to the vessel may result in detachment of atheromatous material from a ruptured plaque, or denude the endothelial lining of the vessel exposing the underlying atheroma for subsequent embolization. Anticoagulant use has been associated with cholesterol embolism, and it has been proposed that by preventing thrombosis of ulcerating atheromatous plaques, anticoagulants favour the dissemination of atheromatous material. However, a causal relationship is unproven and many patients with widespread atherosclerosis coincidentally receive anticoagulants for a variety of reasons. Cholesterol embolism following the use of thrombolytic agents has been rarely reported.

Clinical features

Symptoms are often non-specific with fever, weight loss, and myalgia. The clinical features are otherwise determined by the pattern of organ involvement and are usually referable to the gastrointestinal tract, kidneys, and lower limbs. Bilateral skin changes over the lower extremities are the commonest physical finding and include livedo reticularis, a purpuric rash, 'trash feet', blue toes (acral cyanosis), and focal digital necrosis (Fig. 1 and Fig. 2 and Plate 1 and Plate 2). Ulceration, nodules, and petechiae have also been described. Despite these skin changes and the presence of calf claudication (or frank myositis), pedal pulses may be felt easily, emphasizing that small vessels are occluded in this disorder. Carotid and femoral bruits are frequently heard, reflecting widespread and generalized atherosclerosis.

Fig. 1 Livedo reticularis and vasculitic-like erythematous nodules on the leg of a patient in whom cholesterol-crystal embolization occurred after coronary angiography. (See also Plate 1.)

Fig. 2 Purpuric spots and acral cyanosis of the toes from cholesterol embolism after aortic aneurysm repair. (see also Plate 2.)

Abdominal pain, gastrointestinal bleeding, and pancreatitis may occur and embolism to the stomach, small bowel, colon, gallbladder, and spleen have all been reported. The most frequently involved of these sites is the colon.

Because of their large blood supply and proximity to the abdominal aorta, the kidneys are commonly affected. This usually manifests as a subacute stepwise

deterioration in renal function over 2 to 6 weeks, invariably accompanied by a worsening of pre-existing hypertension that can be labile and difficult to control. Cardiac failure with pulmonary oedema is a common accompaniment. Acute renal failure with necrotizing glomerulonephritis and crescent formation on renal biopsy has been described but is rare. Thus a typical case is an elderly man presenting after angiography with progressive renal failure accompanied by a low-grade fever, abdominal pain, livedo reticularis of the lower body, and purpura over the feet with focal digital ischaemia of the toes.

Transient ischaemic attacks, amaurosis fugax, and strokes can occur when embolism is from the carotid arteries or aortic arch. Retinal cholesterol-crystal emboli may be observed on ophthalmoscopy as bright refractile plaques within the retinal arterioles, especially at their bifurcation. Spinal cord infarction has also been reported.

Investigations

Laboratory findings are non-specific, but frequently include a raised erythrocyte sedimentation rate (**ESR**), plasma viscosity, and C-reactive protein (**CRP**). Leucocytosis and a transient eosinophilia are common and may be pronounced. Depending on the tissue involvement, an elevation in creatine phosphokinase, amylase, lactate dehydrogenase (**LDH**), serum aspartate aminotransferase (**AST**), and alkaline phosphatase may all be seen. Hypocomplementaemia is rare and usually mild. Antineutrophil cytoplasmic antibodies (**ANCA**) have been reported, and their presence may further confuse the diagnosis with a multisystem vasculitic process. Mild proteinuria is generally present and nephrotic-range proteinuria has been reported. Urine microscopy may be bland or reveal red cells, white cells (particularly eosinophils), and hyaline and granular casts. Renal failure is frequently non-oliguric.

Histology

The definitive histological diagnosis of cholesterol-crystal embolism can usually be made from biopsies of kidney, skin, or muscle (including clinically uninvolved areas), although sampling error may miss the lesion due to its patchy distribution. Antemortem histological diagnoses have also been made from other tissues, including a gastric biopsy, prostatic currettings, and a bone marrow biopsy.

The diagnostic feature is of biconvex, needle-shaped cholesterol clefts within the lumen of arteries or arterioles that remain after the crystals have dissolved during routine histological preparation (Fig. 3 and Plate 3). In fresh samples, the crystals can be identified by birefringence under polarized light or by specific histochemical staining of cholesterol. In the kidneys, the typical finding is occlusion of small arteries and arterioles of between 150 and 200 µm in diameter, such as the arcuate and interlobular arteries, resulting in patchy areas of ischaemia and small areas of infarction. Crystals can also be seen within the glomeruli. In chronic cases, ischaemia produces a wedge-shaped lesion involving all components of the renal cortex radiating towards the capsule. The glomeruli appear ischaemic and sclerosed and the tubules become atrophic and separated by interstitial fibrosis. Grossly, the kidneys may be reduced in size with a rough granular surface and wedge-shaped scars.

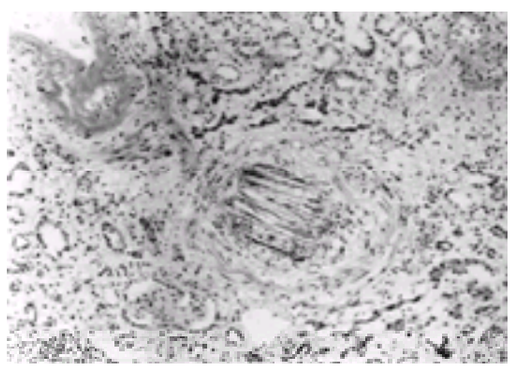

Fig. 3 Renal biopsy demonstrating the characteristic needle-shaped cholesterol clefts occluding a medium-sized renal arteriole with surrounding inflammatory cell infiltration, intimal proliferation, thickening, and concentric fibrosis. There is extensive autolysis (postmortem sample). (See also Plate 3.)

Based on animal studies involving the injection of atheromatous material, the presence of cholesterol crystals in the vascular lumen is thought to trigger a localized inflammatory and endothelial vascular reaction. Inflammatory cells (mainly macrophages and eosinophils) infiltrate, and multinucleated giant cells engulf the cholesterol crystals, but these are resistant to the scavenger effects of macrophages and may persist for many months. The inflammatory phase is followed by marked intimal thickening with concentric fibrosis and occlusion of the vessel. Depending on the extent of organ involvement, these pathological changes result in ischaemia, infarction, or, rarely, necrosis of the distal tissue.

Differential diagnosis

The diagnosis is frequently missed during life. A high index of clinical suspicion is therefore required, particularly in elderly patients with evidence of atherosclerotic disease who develop renal failure after arteriography or following aortic or cardiac surgery; cholesterol embolism should also be considered in the differential diagnosis of a multisystem disease in elderly patients. Spontaneous cholesterol-crystal embolism associated with renal failure, fever, rash, and eosinophilia may not surprisingly be misdiagnosed as a vasculitic illness such as Wegener's granulomatosis, microscopic polyangiitis, Churg–Strauss syndrome, polyarteritis nodosa, or bacterial endocarditis (see Section xxxx, Rees). A false-positive ANCA test may further compound the diagnostic difficulty. Under these circumstances, renal biopsy is mandatory to make the correct diagnosis.

Clinical course and management

Mortality is high due to the coexistence of cardiac and vascular disease with renal failure in an elderly patient. Renal impairment may remain stable, but frequently progresses such that dialysis is required, although partial recovery has been reported, even after several months of dialysis. The mechanism of this recovery is uncertain.

There is no effective therapy. Steroids, aspirin, dipyridamole, and low molecular weight dextran have all been tried, but without any clear effect. There are anecdotal reports of a response to human menopausal gonadotrophin coenzyme A (**HMG CoA**)-reductase inhibitors (theoretically inducing plaque stabilization), but recovery may have been spontaneous. Anticoagulants are of no proven benefit and should be avoided given their potential role in the pathogenesis of the disorder. Encouraging results with iloprost have recently been reported although these observations require replication.

Computed tomography (**CT**) scanning of the aorta has been used to identify the precise source (for instance, aortic aneurysm, localized aortic plaque) of cholesterol emboli, and surgical replacement of the diseased vessel with a graft has been advocated. However, major surgery in elderly vasculopaths with renal impairment carries significant risks and is generally avoided.

Supportive therapy is directed at the control of hypertension and appropriate management of renal failure. Prevention is important, particularly with the increasing number of older patients submitted to invasive angiography. Non-invasive methods of arterial imaging such as CT or magnetic resonance (**MR**) angiography are to be preferred in patients with diffuse atherosclerosis. When invasive angiography is unavoidable, careful attention must be paid to the angiographic technique, including the arterial approach (brachial instead of femoral for cardiac catheterization), use of softer, more flexible catheters and reduced catheter manipulation.

Further reading

Belenfant X, Meyrier A, Jacquot C (1999). Supportive treatment improves survival in multivisceral cholesterol crystal embolism. *American Journal of Kidney Disease* **33**, 840–50. [Recent study reporting good (87 per cent) 1-year patient survival with aggressive protocol-based supportive care.]

Case Records of the Massachusetts General Hospital (Case 34–1991). *New England Journal of Medicine* **325**, 563–72. [Classic clinicopathological exercise in the best tradition.]

Elinav E, Chajek-Shaul T, Stern M (2002). Improvement in cholesterol emboli syndrome after iloprost therapy. *British Medical Journal* **324**, 268–9. [New therapeutic approach requiring replication

elsewhere.]

Fine MJ, Kapoor W, Falanga V (1987). Cholesterol crystal embolization: a review of 221 cases in the English literature. *Angiology* **38**, 769–84. [Excellent review.]

Hyman BT, *et al*. (1987). Warfarin-related purple toes syndrome and cholesterol microembolization. *American Journal of Medicine* **82**, 1233–7. [Association of cholesterol-crystal embolization with anticoagulant use.]

Keen RR, *et al*. (1995). Surgical management of atheroembolization. *Journal of Vascular Surgery* **21**, 773–81. [Retrospective series of patients (45 per cent had cholesterol embolism alone) in whom the source of embolism was removed surgically.]

Mannesse CK (1991). Renal failure and cholesterol crystal embolization: a report of 4 surviving cases and a review of the literature. *Clinical Nephrology* **36**, 240–5. [Excellent review.]

Moolenaar W, Lamers CBH (1996). Cholesterol crystal embolisation to the alimentary tract. *Gut* **38**, 196–200. [Clinicopathological report of cholesterol crystal embolism to the gastrointestinal tract using the Dutch National Pathology database to identify cases.]

15.14.4 Takayasu arteritis

Fujio Numano

Introduction

Takayasu arteritis is a systemic chronic vasculitis that mainly involves the aorta and/or its major branches as well as the coronary and pulmonary arteries. Chronically progressive inflammation induces arterial stenosis and/or occlusion due to thrombus formation, resulting in the characteristic clinical picture of weak or absent arterial pulses (Fig. 1). By contrast, acute inflammation causes dilatation of vessel walls and/or aneurysm formation, which can lead to serious problems such as aneurysmal dissection, aortic regurgitation due to dilatation of the ascending aorta, or even aortic rupture. Clinical manifestations will clearly depend on which arteries are involved. The disease generally has a chronic progressive course, often presenting with non-specific inflammatory symptoms such as intermittent fever, fatigue, and malaise that may exist for months to years prior to the onset of full-blown vasculitis.

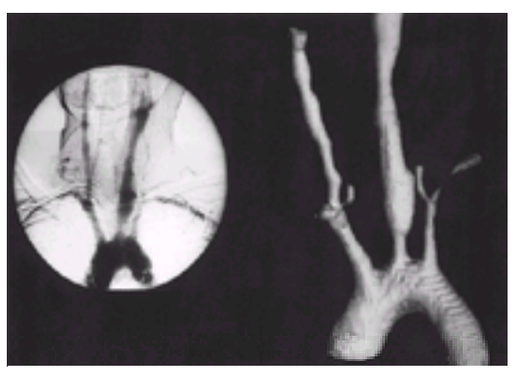

Fig. 1 Digital subtraction angiography and three-dimensional CT of the aortic arch and its branches in a patient with Takayasu arteritis.

The first case of this disease was reported in 1908 at the Japan Ophthalmology Society Meeting by Mikito Takayasu who described interesting fundal findings in a 21-year-old woman suffering from pulmonary tuberculosis, characterized by coronary anastomosis of central retinal arteries (Fig. 2 and Plate 1). Similar ocular appearances were corroborated at the same meeting by K. Ohnishi and T. Kogoshima. Furthermore, both pointed out that they were associated with absence of one or both radial pulses. Today this condition is known to be a vasculitis and the characteristic ophthalmic condition thought to be the result of ischaemia of the retinal circulation due to stenosis or obstruction by arteritis of cervical arteries.

Fig. 2 Typical coronary anastomosis of retinal vessels in Takayasu arteritis. (See also Plate 1.)

The first detailed pathological study was reported by K. Oota in 1940 of a 25-year-old woman with Takayasu arteritis who was admitted to hospital for visual loss and frequent syncopal attacks. He confirmed that this patient had systemic vasculitis involving the aorta, cervical arteries, and pulmonary arteries, and that inflammatory changes were seen not only in the media but in both intima and adventitia, thus calling this condition 'panarteritis'. In 1951, studying the clinical features of 31 cases, K. Shimizu and K. Sano applied the name 'pulseless disease' to the triad of clinical features comprising pulselessness, coronary anastomosis in the retinal vasculature, and accentuated carotid sinus reflex due to ischaemia of the cervical circulation.

Epidemiology

Takayasu arteritis is frequently encountered in Asian countries such as Japan, Korea, China, Singapore, Thailand, Vietnam, India, Israel, and Turkey; also on the American continent in countries such as Peru, Mexico, Brazil, and Colombia. The disease is rarely seen in Caucasian populations. There are geographical differences in sex-specific prevalence. In Japan more than 5000 patients have been treated, of which more than 80 per cent are women with a peak age in the twenties. A recent international comparative analysis revealed a decline of this female preponderance as one moves westwards from Japan; in Israel the sex ratio is almost equal.

Aetiology

The aetiology of Takayasu arteritis is still unknown. The suggestion that it might be due to tuberculosis is now discounted by most authorities, at least in Japan where tuberculosis has drastically decreased, whereas Takayasu arteritis is actually slightly increasing in number. Another hypothesis argues that vasa vasoritis due to virus infection may be the initiating phenomenon, and hyperoestrogenism is still discussed as one of the major causative factors because of the high prevalence of Takayasu arteritis in young women. However, clinical and laboratory findings together with a favourable response to steroid therapy strengthen the argument that autoimmune mechanisms are important.

HLA and other genetic factors

Familial occurrence of Takayasu arteritis has been reported, including three pairs of monozygotic twin sisters in Japan, India, and Brazil, strongly suggesting that some genetic factor(s) is involved in pathogenesis.

There is a close association of Takayasu arteritis with HLA B52 and DR2 in Japan, B5 in India, and B52, DR2, and DQ2 in Korea. In Japan, patients with Takayasu arteritis carrying haplotype A24-B52-DR2 show rapid progression and are resistant to steroid therapy when compared with patients of other haplotypes. Furthermore, patients carrying this haplotype sometimes have other coexisting autoimmune disorders such as Behçet's disease, systemic lupus erythematosus, or ulcerative colitis. Analysis of the MIC gene, which is located near the HLA B locus, also exhibits a close association with Takayasu arteritis. It may suggest that genes involved in pathogenesis may be located between the MIC and HLA B gene locus.

Pathology

Nasu classified Takayasu arteritis histologically into three types—granulomatous inflammation, diffuse productive inflammation, and fibrotic type—which chronologically characterize the progression of this disease. Hochi stressed elastophagia as an important characteristic feature of Takayasu arteritis, progressing segmentally and from the adventitial side towards intima. Inflammation originates around vasa vasorum in the outer side of the media and/or its neighbouring adventitia. The infiltrating cells are mainly T cells, which later are mixed with macrophages. Progression to vasculitis is characterized by destruction of medial smooth muscle cells, elastophagia, and fibrosis. Fibrocellular thickening of intima follows and atherosclerotic changes may accelerate intimal thickening, a complication that makes the diagnosis of Takayasu arteritis difficult in older patients. Skipped lesions composed of a mixture of involved and non-involved areas were once deemed to be a characteristic feature of Takayasu arteritis, but with early diagnosis and treatment this is no longer the case.

Clinical features

Some patients are totally free of symptoms and are diagnosed incidentally as having Takayasu arteritis during regular health examinations because of pulselessness, difference in blood pressure between the arms, or an elevated erythrocyte sedimentation rate. However, Takayasu arteritis can present with non-specific symptoms or, depending upon which vessels are involved, with a wide variety of symptoms, hence the diagnosis is sometimes very difficult to make.

Japanese patients usually have involvement of the ascending aorta and cervical vessels (Fig. 1 and Fig. 3), with the main complaints relating to ischaemia of cerebral, ophthalmological, and/or upper extremity circulation—for example dizziness, syncope, visual disturbance, weak pulse, or pulselessness—as well as inflammatory symptoms such as general malaise, neck pain, and palpable cervical lymph nodes. These inflammatory signs in young women are sometimes misdiagnosed as tuberculosis, viral infection, or rheumatoid arthritis. A 'bird face' due to atrophy of facial muscles (Fig. 4), intermittent claudication of jaw muscles, and perforation of the nasal septum due to long-term cervical circulatory disturbance are helpful in establishing the diagnosis. Today it is becoming rare to find patients whose fundi show a typical retinal coronary anastomosis due to early diagnosis and early treatment; indeed the American College of Rheumatology Subcommittee has excluded this rarely seen ophthalmic condition from the diagnostic criteria of Takayasu arteritis.

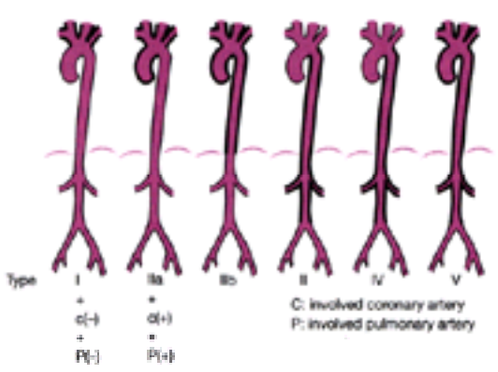

Fig. 3 New classification of the angiogram in Takayasu arteritis, according to the International Conference on Takayasu Arteritis, 1994. Type I involves branches of the aortic arch. Type IIa involves type I plus the ascending aorta and aortic arch. Type IIb involves type IIa plus the thoracic descending aorta. Type III involves the thoracic descending aorta, abdominal aorta, and/or renal arteries. Type IV involves only the abdominal aorta and/or renal arteries. Type V involves the whole aorta and its branches.

Fig. 4 'Bird face' of Takayasu arteritis: hollow cheeks and eye sockets.

Headaches and dizziness associated with hypertension are the most common symptoms of Takayasu arteritis in Asian countries except Japan. These patients mainly have involvement of the abdominal aorta, inducing renovascular hypertension, and they show hypertensive retinal changes. Although rare in Japan, intermittent claudication is another characteristic symptom in China, India, and Thailand.

The commonest finding on physical examination is a weak or absent pulse in one or both brachial, radial, and/or ulnar arteries, which is noticed in almost 80 per cent of Japanese patients. These abnormalities are found more often in the left arm than the right, perhaps because the left axillary artery comes directly from the aortic arch whereas the right one arises from the brachiocephalic artery, thus making the left side more prone to inflammation.

Cardiac manifestations have become the most important cause of death in patients with Takayasu arteritis in Japan. Careful evaluation is required, particularly of aortic regurgitation due to a dilated ascending aorta, which can sometimes present as an emergency with congestive heart failure. The manifestations of ischaemic heart disease, including chest pain, arrhythmia, and/or congestive heart failure, have increased among patients with Takayasu arteritis. This may be due to involvement of the coronary arteries and/or complicated coronary atherosclerosis, which results from the increased longevity of many patients with Takayasu arteritis.

Respiratory symptoms such as dyspnoea, haemoptysis, and pleurisy due to involvement of the pulmonary artery or arteries are not uncommon. Easy thrombus formation predisposes to pulmonary infarction, but this may be clinically silent requiring a pulmonary scintigram in all patients diagnosed with Takayasu arteritis. More than 70 per cent of patients exhibited segmental and/or non-segmental pulmonary infarction in Japanese subjects.

Renal involvement in Takayasu arteritis is usually characterized by renovascular hypertension and renal dysfunction. In India, almost half of all patients with Takayasu

arteritis exhibit renal artery involvement, including the ostia and a variable length of proximal renal artery. Severe proteinuria and hypercholesterolaemia (i.e. nephrotic syndrome) are characteristic of the renal dysfunction that arises in long-standing cases. It is believed that non-specific ischaemic glomerular lesions and mesangial deposition of immunoglobulin during the extended lifespan of patients receiving modern treatment have created this new renal complication of Takayasu arteritis and that it will become a major cause of death.

Hypertension and/or easy thrombus formation cause cerebral vascular accidents in many patients. Although these are no longer the dominant cause of death of patients with Takayasu arteritis in Japan, fatal stroke, hemiplegia, sensory disturbance, and aphasia are still frequently encountered despite well controlled blood pressure.

Laboratory findings and imaging modalities

The inflammatory process of Takayasu arteritis is expressed as an elevated erythrocyte sedimentation rate and C-reactive protein as well as hypergammaglobulinaemia and leucocytosis, as shown in Table 1. Changes of these indicators are well correlated with inflammatory activity and response to steroid therapy. Two national surveys in Japan conducted 10 years apart demonstrated almost equally high frequencies of accelerated erythrocyte sedimentation rate (greater than 20 mm/h) and positive C-reactive protein (greater than 20mg/l). Anaemia is often seen, probably due to chronic inflammation, and total T-cell count is significantly elevated. Other common findings are an increased antistreptolysin O titre and positive rheumatoid factor, which may suggest a common mechanism with rheumatic diseases.

There is a remarkable thrombogenic tendency during the acute inflammatory stage. Accelerated platelet aggregation, hyperfibrinogenaemia, expression of adhesion molecules, and accelerated coagulability give valuable information for assessing the clinical condition of these patients. HLA analysis also provides an additional clue to the diagnosis and is helpful in selecting patients for steroid treatment.

Calcification of the aorta on the chest radiograph in young women sometimes points to the diagnosis in patients free of subjective complaints. A definitive diagnosis can be established by angiography, which provides precise information about the vessels and sites involved, as well as about changes of the inner surface of blood vessels. For example, determining the affected site in the carotid artery is important, but not always straightforward, since it is possible that an easily palpable vessel is aneurysmal, whereas a poorly palpated one is normal. Digital subtraction angiography has become particularly popular because it does not require arterial puncture (Fig. 1). Imaging modalities such as CT, MRI, and /or magnetic resonance angiography (MRA) are also very useful in confirming Takayasu arteritis, even at an early stage, without seriously burdening the patient. Stenosis, dilatation, aneurysmal formation, and thrombosis can be well documented (Fig. 1). These procedures are also good for following the therapeutic effects of steroid treatment on the vasculature. In particular, as aortic regurgitation caused by dilatation of the ascending aorta is a serious complication that determines the prognosis, follow-up of the diameter of the ascending aorta by echocardiogram, MRI, or CT is critically important.

Figure 4 shows an angiographic classification of Takayasu arteritis. An international comparative study demonstrated that half of the patients in every country are type V; many Japanese patients show mainly involvement of cervical vessels and the aortic arch (type I, II); whilst many Indian and Thai patients have involvement of the abdominal aorta (type II, IV) (Table 2).

Diagnosis

Unless the physician is well aware of the disease, the diagnosis of Takayasu arteritis is frequently delayed due to its non-specific presentation. Table 3 includes the guidelines for clinical diagnosis of Takayasu arteritis summarized by the Committee of Takayasu Arteritis of the Ministry of Health and Welfare of Japan. Another set of criteria published by the American College of Rheumatology in 1990 lists six characteristics:

1. age less than 40 years old;
2. claudication of the arm;
3. decreased brachial arterial pulse;
4. greater than 10 mmHg difference in systolic blood pressure between the arms;
5. bruit over subclavian arteries or aorta; and
6. angiographic evidence of narrowing or occlusion of the aorta or its primary or proximal branches.

Presence of three out of six criteria is required for diagnosis, but by these criteria patients in several Asian countries whose abdominal aorta is predominantly involved would elude diagnosis. Although ophthalmological findings and/or symptoms are excluded, in Japan approximately 35 per cent of patients show abnormal ophthalmological findings including microaneurysm, retinal haemorrhage, cataract, or glaucoma.

Therapy

Steroid and antiplatelet therapies are essential in addition to symptomatic treatments such as antihypertensive and vasodilating drugs. The disease requires long-term observation, even after patients are completely free from symptoms, because vascular changes can progress silently and recurrence of vasculitis is not rare.

Steroids and immunosuppressants

Significant improvement can be achieved by steroid treatment, particularly when the disease is at an acute or active stage, starting with 0.5 to 1.0 mg/kg per day of prednisolone, then reducing by 5 mg/day every 2 to 3 weeks, depending upon the response of the erythrocyte sedimentation rate and C-reactive protein as well as symptoms. The target maintenance dose is 5 to 10 mg/day, gradually tapering before being discontinued, which in our experience is possible in 50 per cent of cases. If patients show a haplotype of HLA-B52-DR2 or B*3902, a larger dose (30 to 40 mg/day) of steroid and a longer tapering period may be necessary. Close observation is required because vasculitis can easily flare up during a common cold or other infection, requiring repeated steroid therapy or increased dose.

Immunosuppressive therapy is sometimes combined with steroid, allowing a lower dose and a shorter tapering period to be achieved, which is especially important for patients carrying an HLA B52-DR2 or B*3902 haplotype whose disease is relatively steroid resistant. Conventionally, 100 mg of cyclosporin is administered daily or every other day together with 10 to 20 mg/day of steroid.

Antithrombotic therapy

Antiplatelet therapy should be given to all patients for protection against subsequent thrombus formation. A small dose of acetylsalicylic acid (child bufferin) is the most popular drug used in Japan. Other antiplatelet drugs such as dipyridamole, ticlopidine, or cilostazol are sometimes employed, either alone in patients who cannot tolerate aspirin or combined with child bufferin. Thrombus formation is easily induced on the roughened surface of the arterial wall and active inflammation accelerates thrombus formation even more. Anticoagulant therapy with or without fibrinolytic therapy is necessary when thrombus formation is proceeding or already complete.

Percutaneous vascular intervention

Angioplasty and/or stenting can be effective procedures for some vascular complications of Takayasu arteritis, especially for coronary atherosclerosis. Several authors have reported successful percutaneous transluminal angioplasty (PTA) treatment of patients with renovascular hypertension, but this procedure seems most effective for atherosclerotic disorders involving mainly the intima and far less useful for Takayasu arteritis characterized by thickened adventitia and fibrous intima.

Surgery

The option of surgical treatment (usually bypass procedures) should always be considered for complications of Takayasu arteritis that are beyond medical treatment, for instance renovascular hypertension, coarctation of the aorta, severe ischaemia of the cerebral circulation, severe aortic regurgitation, progression of aneurysm, and dissecting aneurysm. Surgery should be performed, if possible, when inflammation is reasonably under control.

Prognosis

By early diagnosis and early initiation of treatment the prognosis of this disease has been improved. The clinical characteristics of 897 Japanese patients with Takayasu arteritis were studied in 1998: 71 per cent were well controlled and enjoyed an almost normal healthy lifestyle, many receiving no treatment. By contrast, 25 to 30 per cent suffered from severe complications such as aortic regurgitation, congestive heart failure, renal failure, or low visual acuity. In Japan the complications that determine prognosis are heart failure due to aortic regurgitation, thrombus formation leading to cerebrovascular accident, pulmonary infarction, and myocardial infarction. Cerebral haemorrhage due to hypertension is the main cause of death in Asian countries other than Japan.

Further reading

Arend WP *et al*. (1990). The American College of Rhematology: Criteria for the determination of Takayasu arteritis. *Arthritis and Rheumatism* **33**, 1129–34.

Hashimoto Y *et al*. (1992). Aortic regurgitation in patients with Takayasu arteritis—assessment by color Doppler echocardiography. *Heart Vessels* **Suppl 7**, 111–15.

Hata A, Numano F (1995). Magnetic resonance imaging of vascular changes in Takayasu arteritis. *Journal of Cardiology* **52**, 45–52.

Hochi M (1992). Pathological studies on Takayasu arteritis. *Heart Vessels* **Suppl 7**, 11–17.

Kimura A *et al*. (1996). Comprehensive analysis of HLA genes in Takayasu arteritis in Japan. *International Journal of Cadiology* **54**, 61–9.

Kiyosawa M, Baba T (1998). Opthalmological findings in patients with Takayasu disease. *International Journal of Cardiology* **66**(Suppl 1), 141–7.

Nagasawa T (1996). Current status of large and small vessel vasculitis in Japan. *International Journal of Cardiology* **54**(Suppl), 75–82.

Numano F, Kakuta T (1996). Takayasu arteritis—five doctors in the history of Takayasu arteritis. *International Journal of Cardiology* **54**(Suppl), 1–10.

Numano F, Kobayashi Y (1996). Takayasu arteritis: clinical characteristics and the role of genetic factors in its pathogenesis. *Vascular Medicine* **1**, 227–33.

Numano F (1999). Takayasu arteritis beyond pulselessness. *Journal of Internal Medicine* **38**, 226–32.

Numano F (2000). Vaso vasoritis, vasculitis and atherosclerosis. *International Journal of Cardiology* **75** (suppl.), 1–8.

Numano F (2001). Vascular manifestation in Takayasu arteritis. In: Asherson RA, Cervera R, eds. *Vascular manifestations of systemic autoimmune disease*, pp.251–72. CRC Press, Boca Raton.

Sekiguchi M, Suzuki J (1992). An overview on Takayasu arteritis. *Heart Vessels* **Suppl 7**, 6–10.

Takayasu M (1908). A case with peculiar changes of the retinal central vessels. *Acta Societatis Ophthalmologicae Japonicae*, **12**, 554–5. [In Japanese.]

15.15.1 The pulmonary circulation and its influence on gas exchange

Tim Higenbottam, Eric Demoncheaux, and Tom Siddons

Introduction

The main functions of the lungs are the uptake of oxygen and elimination of carbon dioxide. These are driven by the respiratory demands of active cells in the body. The lungs also have important metabolic and endocrine functions. Gas exchange is achieved by a reciprocating cycle of airflow into and out of the lungs through a complex branching system of airways, the bronchi. The gas exchange takes place in the alveoli, the peripheral airspaces that are surrounded by circulating pulmonary capillary blood. This capillary blood is delivered continuously from the right ventricle and passes on to the left atrium. The total cardiac output circulates through the lungs, at flow rates that can vary from 5 litre/min at rest, to 24 litre/min during exercise. The functional anatomy of the airways and lungs are described in detail in Chapter 17.1.2; this section will focus on the pulmonary circulation.

The pulmonary circulation in health

Functional anatomy

The lungs receive blood from the pulmonary and bronchial arteries. The pulmonary arteries deliver the major blood flow, the bronchial blood supply normally being limited to a tiny fraction of the total, 1 to 2 per cent of the cardiac output, but in disease this can expand and disturb gas exchange.

The main pulmonary trunk and the large pulmonary arteries are responsible for connecting the right ventricle to the pulmonary circulation. They accompany the bronchi into the lungs and within the bronchovascular bundles branch dichotomously alongside the bronchi down to the level of the terminal bronchioles, before breaking up into the alveolar capillary bed. The close association of the dense capillary network containing deoxygenated blood to the alveoli offers an ideal environment for gas exchange to take place. Pulmonary venules collect the capillary blood and drain laterally to the periphery of the lung lobules, returning the blood to the left atrium by four branches of the pulmonary veins.

The vessels of the pulmonary circulation have thin walls and can be classified according to their calibre: those of greater than 1000 µm are considered as elastic arteries, those between 100 and 1000 µm as muscular arteries, and smaller ones with endothelial lining and a single elastic lamina but no muscular lining as pulmonary arterioles. The pulmonary capillaries form a dense network around the alveoli and are lined only by endothelial cells. Their diameter is about 7 µm and their walls fuse directly to the alveolar wall (in most places). There are rich anastomotic channels between pulmonary and bronchial vessels.

At rest the pulmonary capillary blood flow is about 5 litre/min. The pulmonary circulation has a pressure drop (from pulmonary artery to left atrium) of only 10 mmHg across it, as against 100 mmHg for the systemic circulation. The resistance of the pulmonary circulation is thus about 10 per cent of the systemic, making it a high-flow, low-resistance circuit. The pulmonary vascular bed has a high reserve capacity to adapt to increased blood flow, adaptation occurring in both large and small pulmonary arteries. The large vessels are dilated by a nitric oxide-dependent mechanism; blocking the production of endothelial nitric oxide limits adaptation. We as yet cannot explain the mechanism by which the precapillary arteries adapt to increased blood flow: these are responsible for the recruitment of previously unperfused capillaries, resulting in a change in gas exchange.

It has been suggested that the alveolar capillary bed is similar to a thin sheet of blood, interrupted in various places by posts, much akin to a large sprawling room with pillars. The oxygenated blood is collected from the capillary network by small pulmonary veins that anatomically run between the lobules of the lung, converge as four pulmonary veins (in humans), and drain back into the left atrium to be pumped by the left ventricle around the systemic circulation.

The lung also derives a very minor part of its circulation from bronchial arteries, which usually arise from the descending aorta and distribute oxygenated blood at systemic pressure to many different sites in the lungs, including the pleura, the nerves, walls of the pulmonary vessels, the intrapulmonary lymph node, and the bronchi down to the terminal bronchiole. Most of the blood supplied by the bronchial arteries drains in the pulmonary veins, thereby contributing to a limited degree of desaturation of arterial blood, a 'physiological shunt', which is present in normal healthy individuals. Small amounts of blood drain in the bronchial veins that enter the azygos and hemiazygous veins.

In pulmonary embolic disease blood flow from the bronchial arteries can sustain the lung tissue and prevent infarction of the lung. The bronchial arteries may undergo hypertrophy in chronic pulmonary inflammation and bronchial neoplasia. Major haemoptysis such as bronchiectasis or aspergilloma arises from hypertrophied bronchial rather than pulmonary arteries and may be treated with therapeutic bronchial artery embolization.

Physiology

The pulmonary circulation offers much lower resistance to flow and operates at lower perfusion pressure than the systemic circulation. The normal mean pulmonary artery pressure is in the region of 15 mmHg with a systolic of 25 mmHg and diastolic of 8 mmHg compared with the systemic pressures of 120/80 mmHg.

At the resting systolic pressure of the right ventricle, 15 mmHg, and in the upright position, gravity causes preferential flow to the basal regions of the lungs rather than the apices. Another mechanism that affects the distribution of blood flow is vasoconstriction of the resistance pulmonary arteries, which are localized just before the alveolar capillaries. Hypoxia in the alveoli and mixed venous blood (blood from the central veins) acts directly on smooth muscle cells in these vessels to cause vasoconstriction, leading to diversion of blood flow from regions of the lung with low oxygen levels towards those regions of higher oxygen tension. This acts to autoregulate the 'matching' of distribution of blood flow to distribution of ventilation.

The relationship between ventilation and perfusion

In the upright lung, after a full expiration at residual volume, the intrapleural pressure at the base of the lung (+3.5 cmH_20) is higher than the atmospheric pressure, but the apex (–4 cmH_20) still remains below atmospheric pressure. During tidal breathing, however, the region of the lung that experiences the greatest change in volume is the dependent part, the bases, when we are in an upright position. This is a result of the sigmoidal shape of the pressure-volume curve of the lung.

The effects of gravity mean that the rate of blood flow is not uniform throughout the lungs (Fig. 1), being higher in the basal regions than the apices. The different hydrostatic pressures within the blood vessels of the lung can explain this uneven distribution. An upright lung may be divided into three zones—1, 2, and 3. The blood flowing in the lung will also act like a continuous column due to the effect of gravity. In the uppermost parts, zone 1, alveolar air pressure (P_A) is usually greater than both arterial (*P*a) and venous (*P*v) pressures. This causes the 'squashing' of capillary beds and restriction of flow.

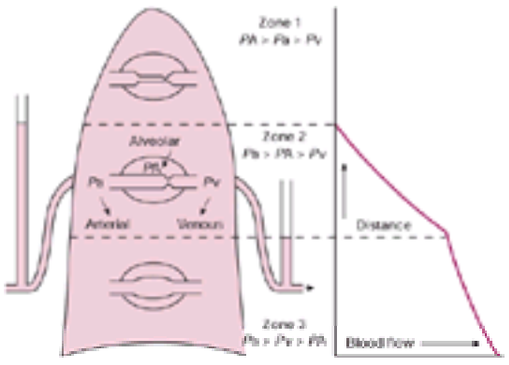

Fig. 1 Diagram of West showing the distribution of blood flow through the lung, which explains the uneven distribution of blood flow in different regions of the lung. *P*a, arterial pressure; P_A, alveolar pressure; *P*v, venous pressure. (Reproduced from West J (2000). *Respiratory physiology*, 6th edn. Lippincott Williams and Wilkins, with permission.)

In the intermediate parts of the lung, zone 2, pulmonary artery pressures are increased due to the hydrostatic effect, but venous pressures (*P*v) may still not be high enough to overcome alveolar pressures and blood flow is only slightly increased compared with the apex (zone 1) regions of the lung.

In the basal part of the lung, zone 3, the hydrostatic pressures generated in both the arterial and venous vessels due to the effect of gravity are greater than alveolar pressures. The flow here is chiefly determined by arterial–venous pressure difference and is therefore higher than in the intermediate and apical regions.

In summary, tidal ventilation is higher at the bases of the lung. Perfusion follows the same distribution to the bases. Ventilation–perfusion ratio (V/Q) values are high at the apex and diminish towards the base of the lung. In a normal healthy individual the overall V/Q ratio is closely matched to allow for efficient gas exchange to take place. Some small areas in the lung may still exist with good perfusion but poor ventilation (or vice versa). In many lung diseases it is likely that gross mismatch of V/Q ratios in many areas impair the process of gas exchange despite the individual having almost normal total ventilation and normal total pulmonary blood flow.

The three-compartment model of ventilation–perfusion and gas exchange

The lungs may be partitioned into a number of functional units in terms of ventilation–perfusion: these do not have discrete anatomical definitions but provide reference points for analysis of gas exchange.

In the absence of a diffusion barrier, the partial pressures of O_2 and CO_2 of the air in each alveolus are the same as those of the end-capillary blood draining it. For the lung as a whole this situation does not arise due to lack of uniformity of ventilation–perfusion within its various units, resulting predominantly from gravitational effects and pleural pressure gradients as described in the preceding section. To obtain the V/Q ratio of such units, the alveolar composition of O_2 and CO_2 would need to be known. Direct sampling of the alveoli is technically not feasible. Sampling at the mouth after deep maximal exhalation is flawed by the dependence of the technique on the speed of exhalation and the baseline state of lung inflation. Alveolar partial pressures of O_2 and CO_2 vary in different parts of the lungs and at different times in the respiratory cycle, hence samples obtained at the mouth are likely to be influenced by the rate and time of emptying of the various lung units.

The three-compartment model of Riley visualizes the lung perfusion as comprising (i) a physiological shunt (Qs), perfusing an area with no ventilation; (ii) alveolar capillary blood equilibrating with ideal alveolar air; and (iii) physiological deadspace (*V*d). Detailed calculations are beyond the scope of this text, but in a resting man up to 70 per cent of the pulmonary ventilation is distributed to the well perfused areas of the lung and up to 95 per cent of the pulmonary capillary blood flow supplies lung units that are well ventilated. These areas are well matched for ventilation and perfusion and constitute the functional part of the lungs in terms of respiratory gas exchange. Approximately 30 per cent of the lung units are ventilated but unperfused, and this constitutes physiological deadspace. Similarly, at rest, 5 per cent of the lung units are well perfused but not ventilated and they constitute the physiological shunts.

There are methods by which the 'matching' between ventilation and perfusion can be assessed: these are still underdeveloped and need refinement, but are none the less providing improved understanding. One such method is the multiple inert gas excretion technique (MIGET). A physiological saline solution of the six inert gases is injected into a peripheral vein. The mixed arterial concentration of each of the gases is measured and taken as the blood-flow weighted mean for the various compartments of the lung. Mixed expired levels are measured and taken as the ventilation weighted mean of the compartmental values. Cardiac output and minute ventilation are also measured and used to calculate the corresponding mixed venous and alveolar concentrations. Based on mass balance considerations a lung ventilation–perfusion distribution is mathematically derived, which is compatible with the arterial and alveolar concentrations of all the inert gases concurrently (Fig. 2).

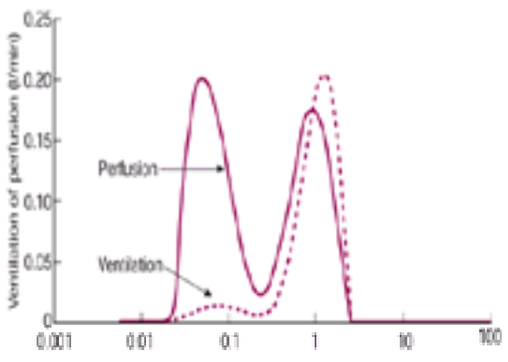

Fig. 2 A representative ventilation–perfusion distribution obtained from a patient with severe airways obstruction at rest. Large areas of perfusion are insufficiently ventilated resulting in hypoxaemia.

Further reading

Deffebach M *et al.* (1987). The bronchial circulation. Small but a vital attribute of the lung. *American Review of Respiratory Disease* **135**, 463–81.

Florence A, Attwood D (1998). *Physicochemical principles of pharmacy*. MacMillan Press, London.

Nunn J (1993). *Applied respiratory physiology*. Butterworth, London.

Silverman E, Gerritsen M, Collins T (1997). *Metabolic function of the pulmonary endothelium*. Raven Press, New York.

Singhal S, Henderson R, Horsfield K (1973). Morphometry of the human pulmonary arterial tree. *Circulation Research* **33**, 190–7.

Wagner P, Naumann P, Laravuso R (1974). Simultaneous measurement of eight foreign gases in blood by gas chromatography. *Journal of Applied Physiology* **36**, 600–5.

West J (1963). Distribution of gas and blood in the normal lungs. *British Medical Bulletin* **19**, 53–8.

West J (1977). Ventilation–perfusion relationships. *American Review of Respiratory Disease* **116**, 919–43.

West J, Wagner P, Derks C (1974). Gas exchange in distributions of VA/Q ratios: partial pressure–solubility diagram. *Journal of Applied Physiology* **37**, 533–40.

Williams S *et al*. (1979). Methods of studying lobar and segmental function of the lung in man. *British Journal of Disease of the Chest* **73**, 97–112.

15.15.2.1 Primary pulmonary hypertension

Tim Higenbottam and Helen Marriott

Introduction

This chapter will consider the nature of primary pulmonary hypertension and how it relates to the other types of pulmonary hypertension. It will also consider the impact of pulmonary hypertension on survival in both primary pulmonary hypertension and that associated with other disease. The causes of the disease and effects of current treatments will be reviewed.

The publication of reports from the National Institutes of Health (**NIH**) sponsored registry of primary pulmonary hypertension provided a description of the clinical features and prognosis, also a clear diagnostic pathway for investigation. The introduction of heart–lung transplantation in 1981 by Bruce Rietz in Stanford University, California, offered the first real hope of long-term survival for those with primary pulmonary hypertension. However, the realization that transplant surgery could only be offered as a treatment for a tiny minority of patients provided the stimulus to consider alternative medical treatments. This led to the introduction of long-term intravenous infusion of prostacyclin (prostaglandin I_2) in 1984, first considered as a bridge to lung transplantation, but subsequently demonstrated in controlled studies to improve survival and enhance quality of life. Inhaled nitric oxide was later recognized to act as a selective vasodilator of the pulmonary circulation (unlike prostacyclin) and has recently gained United States Food and Drug Administration approval as a therapy, specifically for neonatal pulmonary hypertension.

Finally, the quest to understand familial primary pulmonary hypertension has identified causal mutations of the gene for the bone morphogenetic protein receptor II. This is a member of the superfamily of transforming growth factor-b receptors, and mutations appear to increase the sensitivity of cellular responses to this important growth factor. The hope for new therapies that alter the structure rather than physiology of the pulmonary circulation could now be realized.

What is pulmonary hypertension?

In all its forms, pulmonary hypertension is characterized by fibromuscular intimal hypertrophy that results in narrowing and obliteration of the lumen of blood vessels, most commonly the precapillary arteries (Fig. 1 and Plate 1). By the time of presentation with primary pulmonary hypertension up to 80 per cent of lung vessels have been 'lost' (Fig. 2).

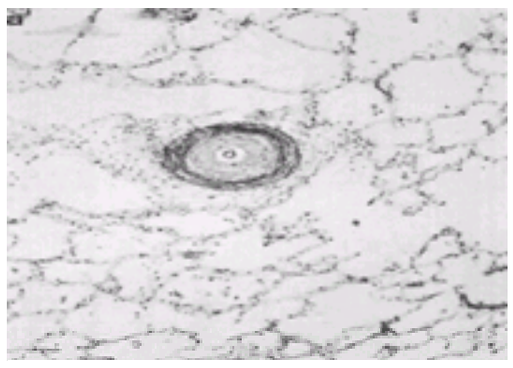

Fig. 1 Intimal thickening of a pulmonary artery in pulmonary hypertension (Chazova I *et al.*, 1995. Pulmonary artery adventitial changes and venous involvement in primary pulmonary hypertension. *American Journal of Pathology* **146**, 389–97). (See also Plate 1.)

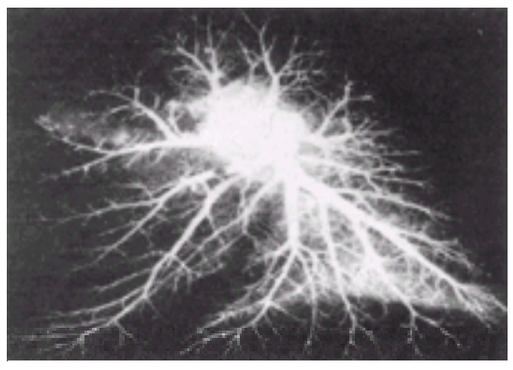

Fig. 2 Loss of precapillary arteries in the lung of a patient with primary pulmonary hypertension.

The physiological consequences of the dramatic loss of small arteries are that the pulmonary vascular resistance is increased, as is the ability of the pulmonary circulation to adapt to the increased blood flow associated with exercise. The resting pulmonary artery pressure is raised; by definition pulmonary hypertension is defined as a mean pulmonary artery pressure in excess of 25 mmHg at rest. More importantly, during exercise there is a rapid rise in pulmonary artery pressure as the pulmonary blood flow increases with cardiac output (Fig. 3).

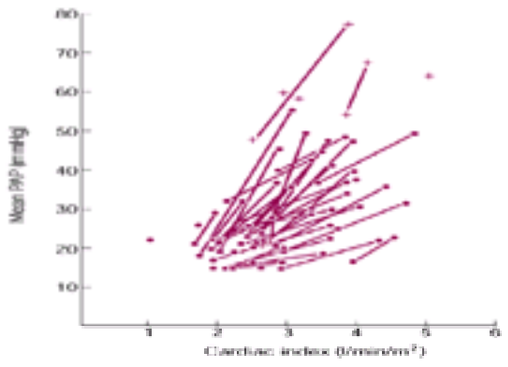

Fig. 3 The increase in pulmonary artery pressure (PAP) with increase in cardiac output (index).

Exercise limitation as a result of breathlessness is the most common symptom in primary pulmonary hypertension and probably caused by acute right ventricular failure. Because the cardiac output fails to rise appropriately with exercise, the mixed venous oxygen saturation ($S\bar{v}O_2$ per cent) falls through increased tissue extraction of oxygen and the arterial oxygen saturation (SaO_2 per cent) also falls. In the later stages of disease the SaO_2 per cent falls even at rest. Right ventricular failure is the most common cause of death.

Clinical features

Pulmonary hypertension should be considered when there is unexplained breathlessness, with no obvious heart or lung disease. In the later stages of the disease symptoms of right ventricular failure become obvious, including syncope, angina-like chest pain, and peripheral oedema. General malaise and cachexia of cardiac failure are endstage symptoms.

In 85 per cent of patients a loud second heart sound is heard. The ECG shows right ventricular strain and RBBB pattern. Also in 85 per cent of patients chest radiography shows large pulmonary arteries.

The screening test is transthoracic echocardiography with Doppler estimation of the tricuspid valve regurgitant flow velocity, which allows the systolic pulmonary artery pressure to be estimated.

At this point it is appropriate to refer all patients with severe pulmonary hypertension—where exercise limitation is a result of pulmonary hypertension—to a specialist centre because of the complexities of diagnosis and the difficulties in managing long-term treatments (Fig. 4). All require a ventilation and perfusion lung scintigraphy followed by a diagnostic right heart catheter. At this the right atrial pressure is measured along with the pulmonary artery pressure, the pulmonary wedge pressure, the cardiac output, and the $S\bar{v}O_2$ per cent. A pulmonary angiogram is undertaken in those patients whose V/Q shows unventilated perfusion defects, although as an alternative investigation it is becoming common to use high-speed spiral CT with an injection of contrast media to assess proximal pulmonary artery obstruction in chronic thromboembolic pulmonary hypertension.

Fig. 4 The investigation and treatment pathway for a patient with pulmonary hypertension.

Classification of pulmonary hypertension

In 1973 the World Health Organization (**WHO**) sponsored the first meeting on pulmonary hypertension. This considered in detail the pathology of the condition and presented the first approach to the classification of the disorder. Unexplained or primary pulmonary hypertension was considered separately from secondary pulmonary hypertension, where an additional disease could be identified in association with pulmonary hypertension. Twenty-five years later the second WHO sponsored meeting focused more upon the common features of the different forms of pulmonary hypertension, in particular their response to treatments such as prostacyclin and the common histopathological changes. Five types of pulmonary hypertension were recognized (Fig. 5).

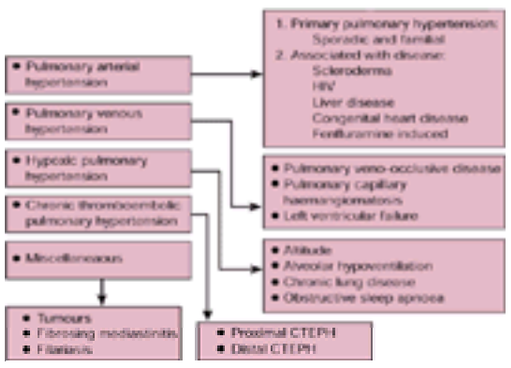

Fig. 5 Types of pulmonary hypertension. CTEPH, chronic thromboembolic pulmonary hypertension (Rich SE, 1998. Primary pulmonary hypertension: executive summary from the world symposium—Primary Pulmonary Hypertension. Available from the World Health Organization via the internet http://www.who.int/ncd/cvd/pph.thml).

Pulmonary arterial hypertension

Pulmonary arterial hypertension was defined as the type of disease where the intimal proliferative changes are found in the arteries alone. In addition there is medial hypertrophy of the muscular arteries and, in advanced disease, changes also include plexiform lesions that it has been suggested are neovascular structures which 'bypass' the obstructed precapillary arteries. Primary pulmonary hypertension is the exemplar and is divided into the sporadic and the familial forms. A necrotizing arteritis has also been described in primary pulmonary hypertension. Peripheral arterial thrombi are often present. Pulmonary arterial hypertension associated with

other diseases includes that seen with HIV infection, liver disease, congenital heart disease (Eisenmenger's syndrome), scleroderma, and other connective tissue diseases. Up to 60 per cent of the pulmonary vascular bed can be obstructed before the symptoms of pulmonary hypertension develop.

The progressive reduction in number and narrowing of the lumen of precapillary arteries initially causes a loss of capacity of the pulmonary circulation to accommodate the increased pulmonary blood flow of exercise. Whilst the resting pulmonary artery pressure may not be raised initially, with exercise the pressure rises rapidly. Right ventricular failure, defined as an inability to sustain the required cardiac output, contributes to the exercise intolerance and the patient experiences breathlessness.

As the disorder advances the pulmonary artery pressure begins to rise, even at rest. The higher systolic and diastolic pressures in the right ventricle can limit diastolic filling of the right ventricular coronary arteries: angina can result, even in the absence of coronary artery disease. Sudden failure of the right ventricle accounts for the exercise-induced syncope seen in the late stages of the disorder. Palpitations on exercise are not uncommon and cardiac dysrythmias are the cause of sudden death in advanced disease.

With failure of the right ventricle at rest the right atrial pressure rises and failing venous return results in the development of peripheral oedema and ascites. Gross hepatic engorgement may occur which impairs the liver's metabolic role.

In addition to these predominately vascular effects, many patients with primary pulmonary hypertension have impaired gas exchange. There is evidence of extensive ventilation/perfusion mismatch, in the main resulting from perfusion of regions of the lungs that are normally poorly ventilated, but also from ventilation of poorly perfused regions. Furthermore, as the cardiac output falls, the mixed venous oxygen level falls as a result of greater extraction of oxygen in the peripheries of the body. Selective flow of this profoundly deoxygenated blood through a limited number of perfused areas of the lungs leaves insufficient time for oxygen uptake to occur, particularly during exercise when the arterial oxygen level can fall profoundly.

In about 15 per cent of patients with primary pulmonary hypertension the foramen ovale of the atrial septum opens as a result of high right ventricular pressure, leading to a right-to-left intracardiac shunt of blood. This adaptation has been shown to prolong survival, as does the fashioning of an atrial septal defect in advanced primary pulmonary hypertension, particularly in infants. The mechanism of this effect is the opportunity for excessive right ventricular pressures during exercise to be released through the defect into the left side of the circulation. This means that the right ventricle does not fail acutely, but the cost is further reduction of arterial oxygen content. However, whilst being more hypoxic than their counterparts without septal defects, these patients can undertake more exercise and live longer.

Pulmonary venous hypertension

Pulmonary venous hypertension is where the intimal proliferation is found in the veins rather than the arteries. There are three main forms—two rare and one very common. Pulmonary veno-occlusive disease and pulmonary capillary haemangiomatosis are the rare conditions, whilst pulmonary venous hypertension from left ventricular failure and left-sided valvular heart disease are common. Any increase in left atrial pressure, pulmonary vascular resistance, and capillary blood flow can lead to a chronic increase in pulmonary venous pressure.

Distinction between pulmonary arterial and pulmonary venous hypertension can be clinically challenging. However, in pulmonary venous hypertension the pulmonary wedge pressure measured with a Swan/Ganz balloon catheter is elevated. On high-resolution computed tomography the lung fields show extensive interstitial lines in the interlobular septi, an appearance akin to left ventricular failure.

Hypoxic pulmonary hypertension

Acute pulmonary hypoxic vasoconstriction

The precapillary arteries of the lungs are sensitive to falls in both alveolar and capillary oxygen partial pressure. At values of oxygen partial pressure below 7.2 kPa, their vascular smooth muscle cells contract, dependent on the activity of voltage-dependent potassium channels, ensuring a matching of the distribution of perfusion of the capillaries to ventilation of the alveoli.

Chronic hypoxic pulmonary hypertension

Whilst transient pulmonary hypertension is often a feature of acute lung disease, normal distribution of perfusion is restored when the alveolar oxygen levels are returned to normal. However, in chronic lung disease (see Chapter 17.6 and Chapter 17.7) the structure of the walls of the precapillary arteries changes. As in pulmonary arterial hypertension the intima proliferates and there may be smooth muscle hypertrophy and adventitial fibrosis. These changes are extensive and not simply localized to those regions of the lungs that are predominantly under-ventilated. In other words the chronic changes in structure of the pulmonary arteries in hypoxic lung disease do not necessarily facilitate close matching between the distribution of ventilation and perfusion.

The increased pulmonary vascular resistance in patients with chronic hypoxic lung disease is usually less severe than in primary pulmonary hypertension, but it does contribute to ill health. There is evidence on echocardiographic studies that right ventricular function is impaired: this is exacerbated by exercise and could contribute to reduced exercise tolerance.

Chronic thromboembolic pulmonary hypertension

Chronic thromboembolic pulmonary hypertension is where the peripheral or proximal pulmonary arteries are occluded by thrombus and emboli, causing widespread segmental and subsegmental defects in lung perfusion. These areas are usually normally ventilated and so gas exchange is impaired. Patients compensate with increased rates of ventilation: $P\mathrm{aO_2}$ is often maintained with a lower than normal $P\mathrm{aCO_2}$ level. As in all forms of pulmonary hypertension, resting cardiac output may be reduced or is reduced when exercise is undertaken.

Miscellaneous

Finally there is the miscellaneous type of pulmonary hypertension. This includes pulmonary hypertension with fibrosing mediastinitis, pulmonary artery tumours, and obstructions associated with protozoal and nematode infestation, such as filariae and schistosomes.

Prognosis

The development of pulmonary hypertension shortens life. This is best seen in primary pulmonary hypertension, where the untreated 3-year post-diagnosis survival is less than 35 per cent (Fig. 6), but pulmonary hypertension also reduces survival when it is associated with other disease such as chronic obstructive bronchitis. This is a result of alveolar hypoxia, as survival is enhanced when the partial pressure of oxygen is restored to normal by long-term oxygen therapy (LTOT).

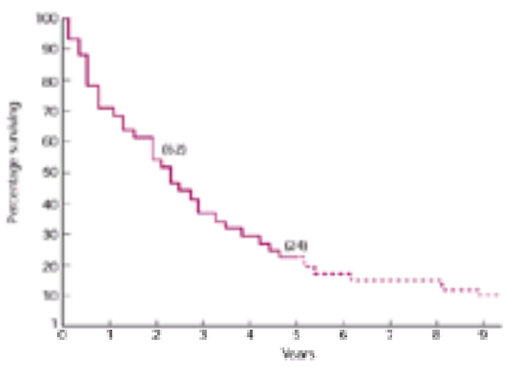

Fig. 6 Survival in primary pulmonary hypertension (Fuster V *et al.*, 1984. Primary pulmonary hypertension: natural history and the importance of thrombosis.

Circulation **70**, 580–7).

The cause of death in pulmonary hypertension depends on the severity of right ventricular failure. In patients with primary pulmonary hypertension it is possible to predict survival from haemodynamic measurements at right heart catheter studies: the presence of a cardiac output less than 2.5 litre/min, mean right atrial pressure more than 10 mmHg, and SvO_2 less than 63 per cent all being poor prognostic factors. For ease it is valuable to use the New York Heart Association Grade III or IV and pulmonary vascular resistance greater than 15 units to predict the chance of survival to be very low without treatment (Fig. 7). A very similar prediction can be made on the basis of haemodynamic measurements in patients with Eisenmenger's syndrome. Sudden death from right ventricular failure is responsible for 47 per cent of the deaths from primary pulmonary hypertension.

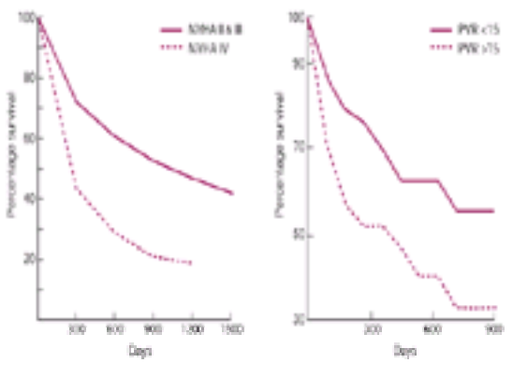

Fig. 7 Survival of patients with primary pulmonary hypertension in the NIH registry according to their New York Heart Association (NYHA) Grade and pulmonary vascular resistance (PVR) (D'Alonzo GE *et al*., 1991. Survival in patients with primary pulmonary hypertension. Results from a national prospective registry. *Annals of Internal Medicine* **115**, 343–9).

Treatment

All patients with pulmonary arterial hypertension should receive anticoagulation treatment as uncontrolled studies have shown overall survival to be improved by their use (Fig. 8). The treatment of chronic thromboembolic pulmonary hypertension with proximal obstructions is surgical thromboendarterectomy: for other types of pulmonary hypertension—with the exception of pulmonary venous hypertension—the use of vasodilators should be considered.

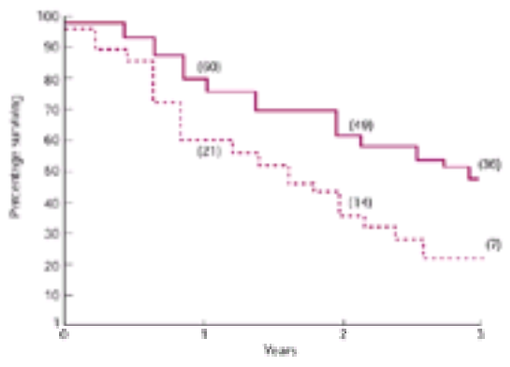

Fig. 8 Survival of patients with primary pulmonary hypertension treated (solid line) or not treated (broken line) with anticoagulation (Fuster V *et al*., 1984. Primary pulmonary hypertension: natural history and the importance of thrombosis. *Circulation* **70**: 580–7).

Early disease

In early/mild disease (cardiac output greater than 2.5 litre/min, mean right atrial pressure less than 10 mmHg, and SvO_2 greater than 63 per cent) decisions about other treatment are made following an acute vasodilator trial, most commonly using inhaled nitric oxide, intravenous prostacyclin, or adenosine. Only if there is evidence of a greater than 20 per cent drop in pulmonary vascular resistance are oral vasodilators considered. This response is found in less than 25 per cent of patients with primary pulmonary hypertension, and for these nifedipine, diltiazem, or amlodopine are used in therapeutic doses. Survival is improved, although the studies are uncontrolled (Fig. 9).

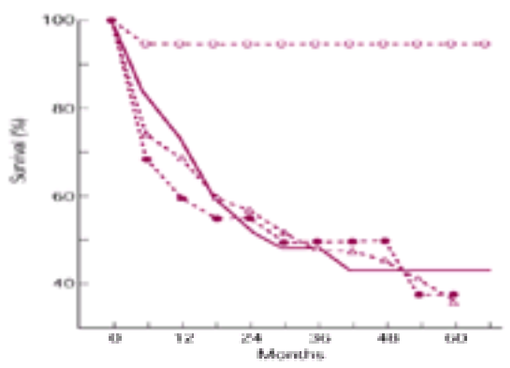

Fig. 9 Kaplan–Meier estimates of survival among patients who responded to treatment (open circles), those who did not respond (solid line), patients enrolled at the NIH registry who were treated at the University of Illinois (solid circles), and the NIH registry cohort (triangles).

Severe disease

In patients with severe disease (cardiac output below 2.5 litre/min, right atrial pressure greater than 10 mmHg, or SvO_2 less than 63 per cent), long-term continuous infusions of prostacyclin improve survival and enhance the quality of life. Actuarial survival at 2 years has been increased to 80 per cent, which exceeds that of untreated patients and the survival following lung transplantation (Fig. 10). Of special interest is the recent report that prostacyclin may, when used long-term, reverse the disease to some extent, with pulmonary vascular resistance falling below the lowest level achieved during the initial right heart catheter study (Fig. 11). This has raised the hope that medical treatments may be able to reverse the disease process.

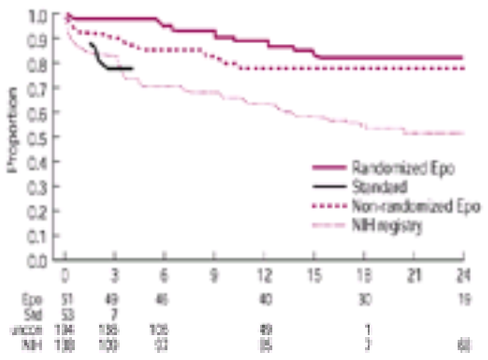

Fig. 10 Long-term continuous infusions of prostacyclin improve survival. Epo, epoprostanol; std, standard; uncon, uncontrolled non-randomized Epo; NIH, National Institute of Health Registry.

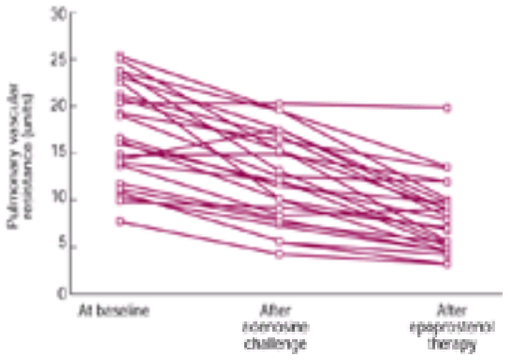

Fig. 11 Long-term prostacyclin may reverse pulmonary hypertension (McLaughlin VV *et al*. (1998). Reduction in pulmonary vascular resistance with long-term epoprostenol (prostacyclin) therapy in primary pulmonary hypertension. *New England Journal of Medicine* **338**, 273–7).

Primary pulmonary hypertension and Eisenmenger's syndrome have the poorest outcomes of any diseases after lung transplantation, the 3-year actuarial survival being only just in excess of the untreated patients. However, the introduction of prostacyclin has delayed the need for lung transplantation by on average 17.5 months and has improved the success of subsequent lung transplantation, with 1-year survival rate in excess of 80 per cent.

Use of vasodilators

Prostacyclin

In pulmonary hypertension the phenotypes of many cells that make up the blood vessels are changed: endothelium, the cells of the matrix tissues, and the vascular smooth muscle cells. In patients with primary pulmonary hypertension, urinary excretion of prostacyclin metabolites is reduced and endothelial cells from the pulmonary arteries express a reduced level of prostacyclin synthase, the enzyme that forms prostacyclin, hence it would seem appropriate to supplement the reduced endogenous production of prostacyclin with exogenous therapy.

The success of intravenous prostacyclin in primary pulmonary hypertension has led to its use in other forms of pulmonary arterial hypertension. In isolated pulmonary hypertension from scleroderma, prostacyclin also improves the quality of life, and in pulmonary arterial hypertension associated with congenital heart disease it improves survival. In those patients with chronic thromboembolic pulmonary hypertension who are not suitable for surgery, prostacyclin also achieves an equivalent improved survival chance to that seen in patients with primary pulmonary hypertension. By contrast, in pulmonary venous hypertension—that is, pulmonary veno-occlusive disease, pulmonary capillary haemangiomatosis, and left ventricular failure—long-term intravenous prostacyclin worsens the survival chances and should not be used. The increased perfusion leads to pulmonary oedema, respiratory failure, and death.

Analogues for prostacyclin have been developed. Iloprost, which is more stable, is of equivalent effect to prostacyclin when given intravenously and has also been used orally and by inhalation. Beroprost is another oral analogue, whilst UT-15 was introduced for continuous subcutaneous infusion. All are in clinical trials, with encouraging early results. There is evidence that inhaled iloprost achieves the same degree of 'reversal' of pulmonary hypertension as intravenous delivery when used long term.

Nitric oxide

As with prostacyclin synthase, expression of nitric oxide synthase and release of nitric oxide is also reduced in the pulmonary arteries in both primary and other types of pulmonary hypertension. Basal nitric oxide release seems normal from pulmonary hypertensive lungs, but stimulated release of the gas is reduced, particularly in hypoxic pulmonary hypertension. In advanced primary pulmonary hypertension there are clusters of new vessels, called plexiform lesions, branching from precapillary arteries. Nitric oxide synthase is overexpressed in these lesions, which develop in the region of the 'obstructed' precapillary pulmonary arteries, perhaps bypassing the obstruction and providing a route from pulmonary artery to the pulmonary veins.

In 1987 inhaled nitric oxide was shown to be a selective pulmonary vasodilator (Fig. 12). However, a problem of handling and using nitric oxide is its speed of oxidation to form nitrogen dioxide, which can injure the epithelial lung surfaces even at low concentrations of 5 parts per million (ppm). This rate of reaction is dependent on the concentration of the gas and the concentration of oxygen, hence it is necessary to mix the nitric oxide (which is stored in nitrogen) carefully with air or oxygen just before it is inhaled. This is straightforward when the patient is receiving mechanically assisted ventilation, but difficult when ventilation is spontaneous. It may be achieved, however, using a 'spiked' delivery system that adds nitric oxide as a small bolus of gas at the beginning of the breath using a patient triggered device. This can reduce the pulmonary vascular resistance by the same amount as when the breath is filled with 40 ppm of nitric oxide, but with only a fortieth of the dose with each breath.

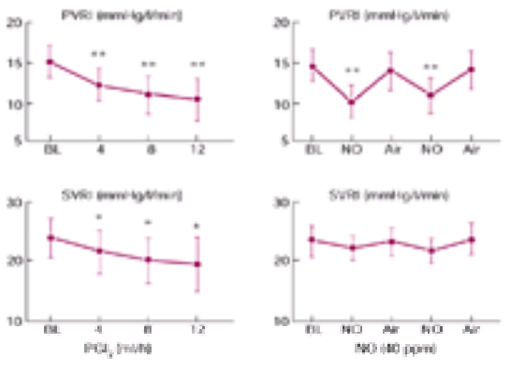

Fig. 12 Inhaled nitric oxide is a selective pulmonary vasodilator. PVRI, pulmonary vascular resistance index; SVRI, systemic vascular resistance index; ppm, parts per million (of inhaled nitric oxide); PGI_2, prostaglandin I_2; *, significantly different from baseline (Pepke Zaba J *et al.* (1991). Inhaled nitric-oxide as a cause of selective pulmonary vasodilatation in pulmonary-hypertension. *Lancet* **338,** 1173–4).

Inhaled nitric oxide is about half as effective as a vasodilator as prostacyclin. It has not yet found a place in the treatment of primary pulmonary hypertension in general, although in some case reports it seems to 'reverse' established primary pulmonary hypertension when given long term. In chronic obstructive pulmonary disease, the pulmonary vascular resistance is reduced in patients receiving long-term oxygen therapy when nitric oxide is administered for 3 months, and in this group inhaled nitric oxide may have a part to play in reducing the pulmonary hypertension and right ventricular dysfunction seen during exercise. Long-term, randomized, controlled trials are awaited. As a therapy inhaled nitric oxide has recently been granted a licence for the treatment of pulmonary hypertension of the neonate, here avoiding the use of extracorporeal oxygenation.

Phosphodiesterase inhibitors

In the lungs the breakdown of cGMP, formed in vascular smooth muscle cells by nitric oxide, is predominately by the phosphodiesterase-5-enzyme that is especially expressed in lung and in the vascular cells of the corpus cavernosum. By use of the inhibitor sildenafil (Viagra) it is possible to induce pulmonary vasodilatation. Given orally long-term this reduces pulmonary vascular resistance and improves exercise tolerance in patients with pulmonary arterial hypertension.

Endothelin-1 antagonists

There have been numerous reports that endothelin-1 is expressed to a greater degree than normal in most types of pulmonary hypertension and that circulating levels of endothelin-1 are also elevated, especially in acute hypoxia of altitude and in chronic hypoxic lung disease. These findings have led to the suggestion that endothelin receptor antagonists might be effective treatments for pulmonary hypertension, and some 32 are in development. One agent, bosentan, has been shown to reverse structural abnormalities in experimental models of pulmonary hypertension. Anxieties about teratogenicity and liver injury have limited the widespread use of these compounds in clinical practice, but encouraging results have been observed in reports of series of patients with isolated pulmonary hypertension and scleroderma. Long-term, randomized, controlled trials have now reported improved exercise tolerance.

Particular causes of pulmonary arterial hypertension

Pulmonary hypertension in fenfluramine and dexfenfluramine users

In the late 1980s a physician in Paris, Francois Brenot, noticed an increasing number of patients with primary pulmonary hypertension being referred who had been treated with the anti-obesity drug fenfluramine or its isomer dexfenfluramine. This observation led to a case–control study of all patients with primary pulmonary hypertension diagnosed over a period of 18 months. For each of 192 patients with this condition three controls were found who were matches for gender, age, and district of residence. An independent panel of doctors checked the entry criteria for each patient with primary pulmonary hypertension. All had a detailed questionnaire administered on past and present therapies, including anti-obesity treatments. It was found that an excess number of patients had taken one of the two drugs, for at least 3 months. Indeed, 1 in 8000 was the calculated risk for developing primary pulmonary hypertension when either fenfluramine or dexfenfluramine was taken for more than 1 year, suggesting a causal relationship. By contrast, the appetite suppressant phentermine was not associated with pulmonary hypertension. Of note, in the 1960s a similar association had been found between the use of another anti-obesity drug, aminorex, and the development of primary pulmonary hypertension.

What is the mechanism by which fenfluramine/dexfenfluramine might cause primary pulmonary hypertension? Both accumulate in the lungs through uptake by the serotonin transporter for which they have a high substrate affinity, and both are metabolized by the cytochrome P450 enzyme CYP2D6 that is expressed in the liver, lungs, and right side of the heart. In part the susceptibility of certain individuals to develop primary pulmonary hypertension when taking fenfluramines can be accounted for by polymorphism of the CYP2D6 enzyme. About 8 per cent of Caucasian and 1 per cent of Asian people express little or no active enzyme, hence they will have little ability to metabolize these drugs within the lungs, with local lung toxicity a possible explanation for the development of pulmonary hypertension. Within the population of patients with primary pulmonary hypertension who have taken dexfenfluramine and fenfluramine there is a higher proportion of poor metabolizers compared with controls (20 per cent compared with 7 per cent). High concentrations of dexfenfluramine and fenfluramine are able to induce vascular smooth muscle contraction in humans as a result of inhibition of membrane potassium channels. In addition, fenfluramine at high concentration can inhibit expression of the 1.5Kv subunit of the voltage-dependent potassium channel of vascular smooth muscle cells.

Familial primary pulmonary hypertension

About 6 per cent of all patients with primary pulmonary hypertension have a family history of the condition, exhibiting an autosomal dominant pattern of inheritance with incomplete penetrance. The histopathology of the pulmonary vascular disease is identical to that found in sporadic primary pulmonary hypertension.

In 1998 two laboratories identified a region on chromosome two associated with familial primary pulmonary hypertension after detailed linkage analysis of extended families. Further work led to identification of mutations of a gene encoding for bone morphogenetic protein receptor II (*BMPR 2*) as the cause. This is a member of the superfamily of transforming growth factor-b receptors.

Transforming growth factor-b is an important cytokine regulator of pulmonary angiogenesis. It inhibits cell growth, cell differentiation, and stimulation of collagen deposition. In the inherited illness hereditary haemorrhagic teliangectasia, new vessels in the lungs and mucosa lead to bleeding. Mutations of the gene encoding endoglin, which is a transforming growth factor-b receptor-complex accessory protein, and of the transforming growth factor-b type 1 receptor ALK-1, have been identified in this disease. A 'failure' of signalling by transforming growth factor-b could account for the new vessel formation. By contrast, mutations of *BMPR 2* appear to enhance the effects of transforming growth factor-b, perhaps promoting vessel 'loss' and synthesis of collagen. Much needs to be learnt about the key ligands, receptors, and downstream signalling pathways. However, this lead from familial primary pulmonary hypertension could have identified a common mechanism for many different types of pulmonary hypertension. Overexpression of transforming growth factor-b is often found in both experimental and human forms of pulmonary hypertension. Lessons learnt from genetics could offer useful clues as to how we might 'fully' reverse pulmonary hypertension in the future.

Further reading

Abenhaim L *et al*. (1996). Appetite-suppressant drugs and the risk of primary pulmonary hypertension. *New England Journal of Medicine* **335**, 609–16.

Anderson E, Simon G, Reid L (1973). Primary and thromboembolic pulmonary hypertension: a quantitative pathological study. *Journal of Pathology* **110**, 273–93.

Barst RJ *et al*. (1996). A comparison of continuous intravenous epoprostenol (prostacyclin) with conventional therapy for primary pulmonary hypertension. The Primary Pulmonary Hypertension Study Group. *New England Journal of Medicine* **334**, 296–302.

Bourdillon P, Oakley C (1976). Regression of primary pulmonary hypertension. *British Heart Journal* **38**, 264–70.

Brenner O (1935). Pathology of the vessels of the pulmonary circulation. Part I. *Archives of Internal Medicine* **56**, 211–37.

Channick *et al.* (2001). Effects of the dual endothelin-receptor antagonist bosentan in patients with pulmonary hypertension: a randomised placebo-controlled study. *Lancet* **258**, 1119–23.

Chazova I *et al.* (1995). Pulmonary artery adventitial changes and venous involvement in primary pulmonary hypertension. *American Journal of Pathology* **146**, 389–97.

Conte JV *et al.* (1998). The influence of continuous intravenous prostacyclin therapy for primary pulmonary hypertension on the timing and outcome of transplantation. *Journal of Heart and Lung Transplantation* **17**, 679–85.

D'Alonzo GE *et al*. (1991). Survival in patients with primary pulmonary hypertension. Results from a national prospective registry. *Annals of Internal Medicine* **115**, 343–9.

Deng Z *et al*. (2000). Familial primary pulmonary hypertension (gene PPH1) is caused by mutations in the bone morphogenetic protein receptor-II gene. [In process citation.] *American Journal of Human Genetics* **67**, 737–44.

Fuster V *et al.*(1984). Primary pulmonary hypertension: natural history and the importance of thrombosis. *Circulation* **70**, 580–7.

Heath D, Whitaker W, Brown J (1957). Idiopathic pulmonary hypertension. *British Heart Journal* **19**, 83–92.

Heath D, Segel N, Bishop J (1966). Pulmonary veno-occlusive disease. *Circulation* **34**, 242–8.

Higenbottam T *et al.*(1984). Long-term treatment of primary pulmonary-hypertension with continuous intravenous epoprostenol (prostacyclin). *Lancet* **i**, 1046–7.

Higenbottam TW *et al*. (1999). Subjects deficient for CYP2D6 expression (poor metabolisers) are over-represented among patients with anorectic ssociated pulmonary hypertension. *American Journal of Respiratory and Critical Care Medicine* **159**(3 SS), A165.

Hoeper MM *et al*. (2000). Long-term treatment of primary pulmonary hypertension with aerosolized iloprost, a prostacyclin analogue. *New England Journal of Medicine* **342**, 1866–70.

Kay J, Smith P, Heath D (1971). Aminorex and the pulmonary circulation. *Thorax* **26**, 262–70.

Lane K *et al*. (2000). Heterozygous germline mutations in *BMPR2*, encoding a TGF-b receptor, cause familial primary pulmonary hypertension. *Nature Genetics* **26**, 81–4.

Lawson R., Higenbottam T (1999). Primary pulmonary hypertension. In: Grassi C *et al.* eds. *Pulmonary diseases*, pp 373–9. McGraw-Hill, London.

Loyd J, Primm R, Newman J (1984). Familial primary pulmonary hypertension: clinical patterns. *American Review of Respiratory Disease* **129**, 194–7.

Mosser K *et al.* (1983). Chronic thrombotic obstruction of major pulmonary arteries; results of thromboembolectomy in 15 patients. *Annals of Internal Medicine* **99**, 299–305.

Pepke Zaba J *et al.* (1991). Inhaled nitric-oxide as a cause of selective pulmonary vasodilatation in pulmonary-hypertension. *Lancet* **338**, 1173–4.

Rich S *et al*. (1987). Primary pulmonary hypertension. A national prospective study. *Annals of Internal Medicine* **107**, 216–23.

Rich SE (1998). Primary pulmonary hypertension: executive summary from the world symposium—Primary Pulmonary Hypertension. Available from the World Health Organization via the internet http://www.who.int/ncd/cvd/pph.thml

Rozkovec A, Montanes P, Oakley C (1986). Factors that influence the outcome of primary pulmonary hypertension. *British Heart Journal* **55**, 449–58.

Wagenvoort C, Wagenvoort N (1970). Primary pulmonary hypertension: a pathologic study of vessels in 156 clinically diagnosed cases. *Circulation* **42**, 1163–84.

Wang J *et al.* (1998). Action of fenfluramine on voltage-gated K+ channels in human pulmonary- artery smooth-muscle cells. [Letter.] *Lancet* **352**, 290.

15.15.2.2Pulmonary oedema

J. S. Prichard (revised by J. Firth)

Introduction

Acute fulminant pulmonary oedema is a terrifying event in which patients literally drown in their own body fluids. Much more commonly, the clinician is called to treat pulmonary oedema in its less acute form, for breathlessness disturbs the patient long before serious alveolar flooding has begun.

Because pulmonary oedema is very commonly seen as a manifestation of left-sided heart disease—where its relief by diuretics is so effective—there is a temptation to forget the very wide range of other causes. Indeed, it is prudent to make the diagnosis of hydrostatic pulmonary oedema of cardiac origin only when other manifestations of heart disease are present, and to consider wider possibilities in all other circumstances. Pulmonary oedema has many possible causes, which occur in combination more often than is usually recognized (see Table 1). Only by careful and clear analysis of clinical and pathophysiological data can the contributing factors be identified and the clinical situation fully understood.

Physiological and experimental aspects of pulmonary oedema

Fluid balance between the capillaries and the interstitial space

The continuous movement of water from the lung capillaries into the interstitium is regulated by the permeability of the endothelium to water and protein and by the imbalance of hydrostatic and osmotic forces across the membrane. The Starling hypothesis suggests that perturbation of any one of five factors could lead to oedema (Fig. 1 and Fig. 2). These are capillary hydrostatic pressure (P_{cap}), interstitial tissue pressure (P_{int}), plasma colloid osmotic (oncotic) pressure (P_{cap}), endothelial permeability (expressed by k and s), and lymphatic function. Abnormalities in the first four will cause oedema by increasing water entry to the interstitial space, whilst impaired function of the last will diminish drainage. Interstitial colloid osmotic pressure (P_{int}) has not been included as an independent variable as it is determined by the plasma protein concentration and endothelial permeability.

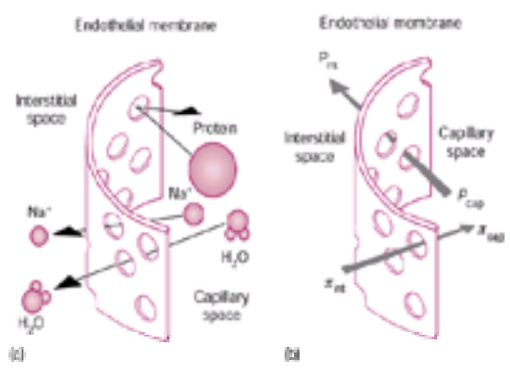

Fig. 1 (a) The lung endothelial membrane is permeable to water and electrolytes but less permeable to macromolecules. (b) The Starling equation: $Q_1 = K(P_{cap} - P_{int}) - Ks(p_{cap} - p_{int})$, where Q_1 is the net fluid filtration rate, K is the filtration coefficient, s is the reflection coefficient, ($P_{cap} - P_{int}$) is the hydrostatic pressure gradient from the capillary lumen to interstitial space, and ($p_{cap} - p_{int}$) is the oncotic pressure difference across the capillary membrane.

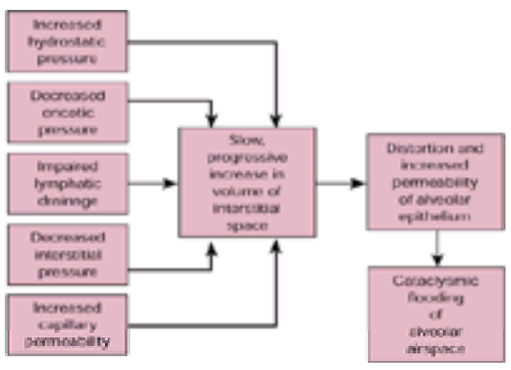

Fig. 2 The initiation of pulmonary oedema and the sequence of development.

Experimentally, the development of pulmonary oedema may be characterized by the relationship between tissue water and microvascular hydrostatic pressure (Fig. 3). In the normal lung, the water content rises only slowly until the capillary pressure reaches 25 to 30 mmHg: thereafter, the rise is rapid. The curve is shifted leftwards by decreased interstitial pressure, increased endothelial permeability, decreased plasma oncotic pressure, or impaired lymphatic drainage. Figure 3 illustrates the interactions between these factors: at low and normal hydrostatic pressures, changes in oncotic pressure, permeability, and lymphatic drainage do not readily cause oedema but, at higher hydrostatic pressures, their effect is much more dramatic. The fact that pulmonary capillary pressure may be raised to 25 to 30 mmHg before there is any significant accumulation of water in a normal lung is a considerable 'safety factor', due principally to the behaviour of the lymphatic system. In response to faster transcapillary water flux from whatever cause (see below), the lymphatic system can increase its activity so much that flow accelerates to between three and ten times the basal level before the drainage becomes overwhelmed. The situation in which the lung water content has increased only little whilst

the transcapillary and lymph fluxes have increased considerably emphasizes that pulmonary oedema is a dynamic phenomenon, in which tissue swelling is but the endstage reached when lymphatic drainage capacity is exceeded. Only then does fluid accumulation begin—slowly at first in the interstitial space, but then rapidly as alveolar flooding begins.

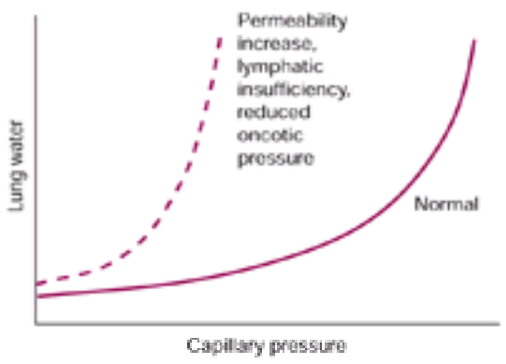

Fig. 3 Lung water content and capillary pressure. In the normal lung tissue, the water content does not begin to increase until the capillary pressure is approximately 30 mmHg. Where colloid osmotic pressure (e.g. plasma protein concentration) is reduced, endothelial permeability is increased or the lymphatic pump is impaired, the whole curve is shifted to the left. (Reproduced from Prichard JS (1982). *Edema of the lung*. Charles C. Thomas, Springfield, Illinois, with permission.)

Hydrostatic pulmonary oedema

Any increase in capillary hydrostatic pressure, whether from cardiac failure, fluid overload, or pulmonary venous occlusion, speeds the rate of water flow into the interstitium. Provided the increase in pressure is not too great, this process will be self-limiting. Thus, molecular sieving, by allowing water to enter the interstitial space more readily than macromolecular solutes, will reduce p_{int}. Increased interstitial water increases the interstitial hydrostatic pressure P_{int} and decreases the macromolecular exclusion volume—again increasing P_{int}. So, as long as lymphatic pumping can keep pace, the tissue water will expand only slightly. However, once the capacity of the lymphatic drainage is exceeded, accumulation of an oedema fluid with a low protein content begins. This starts in the lower parts of the lung (because it is here that hydrostatic pressures are greatest) and is associated with a characteristic redistribution of blood flow away from the lung bases.

The activity of lung lymphatics is critical in determining the onset and extent of hydrostatic oedema, and therefore it is not surprising to find that, in conditions where pulmonary vascular pressures are chronically elevated, the lymphatics undergo hypertrophy as a protective mechanism. Consequently, acute elevations of pulmonary vascular pressure will produce acute life-threatening oedema at levels that, when reached chronically, cause little distress and are registered clinically only by the characteristic radiological changes of lymphatic hypertrophy.

High permeability pulmonary oedema

Endothelial damage speeds water flux into the interstitial space. But, unlike hydrostatic oedema, there is also an increase in protein flux so that the oedema fluid has a high protein content. This has four consequences:

1. The oncotic pressure of the interstitial fluid increases and one of the major mechanisms for limiting the progress of oedema becomes unavailable.
2. Much of the protein reaching the tissue and alveoli is fibrinogen, which coagulates. Initially, the damage from interstitial coagulation is limited by fibrinolysis by plasminogen, but this defence is soon exhausted and mobilization of the coagulum ceases.
3. The residual coagulum impairs lymphatic drainage.
4. The residual coagulum becomes the skeleton on which lung fibrosis develops.

By far the most common cause of high permeability pulmonary oedema is the acute respiratory distress syndrome, which is discussed in Chapter 16.5.1. Less common causes include toxic gases and fumes (Chapter 17.11.17) and drugs (Chapter 17.11.19).

Pulmonary oedema and reduced plasma oncotic pressure

A reduction in plasma oncotic pressure increases fluid transudation into the lung and leads to pulmonary oedema at lower hydrostatic pressures than would otherwise be expected. Although this is readily demonstrable experimentally, it is frequently overlooked in clinical practice, where it may be of importance following myocardial infarction, after transfusion of crystalloids, and in adult respiratory distress syndrome. A useful clinical guide to the danger is the difference between pulmonary wedge pressure (measured by a Swann–Ganz catheter) and colloid osmotic pressure (the **COP–PAW** gradient). The normal lower limit of this index is about –12 mmHg, but at levels below –9 mmHg the risk of oedema is considerably enhanced. A practical problem in applying this method has been the difficulty in standardizing and maintaining protein oncometers. The alternative of using serum protein measurements is valuable but slower.

Lymphatic oedema and the role of the lung lymphatics

The lymphatic system provides the lung with its major 'safety factor'. It is capable of increasing the tissue clearance rate at least 10-fold before becoming overwhelmed. In chronic venous and capillary hypertension, as in mitral stenosis, even larger lymph flows occur because of lymphatic hypertrophy.

Oedema soon develops when lymphatic drainage is occluded experimentally. This has clinical relevance for patients with lung transplants, whose lung lymphatic pathways are severed and in whom initial alveolar flooding is common. Lymphatic oedema also plays a part in pulmonary oedema from lymphangitis carcinomatosa and in facilitating oedema in patients with silicosis and malaria.

Reduced interstitial pressure and pulmonary oedema

Tissue pressure within the interstitial space is one of the determinants of transendothelial fluid movement. It can be altered independently of intravascular events by changes in the intrapleural pressure. Thus, when extreme negative intrapleural pressures occur, the interstitial perialveolar tissue pressure can fall considerably below its normal subatmospheric level and accelerate the rate of fluid movement into the interstitium. Oedema will appear if the rate of fluid entry exceeds the rate at which it can move through the interstitium and be removed by the lymphatics.

The sequence of oedema accumulation

When oedema fluid begins to accumulate in lung tissue—irrespective of the underlying cause—it does so first around fissures, blood vessels, and airways because these tissues are 'loose' and swell easily without great change in tissue pressure. When this 'sump' has become near maximally dilated, swelling and thickening of the alveolar wall begin. Finally, after a phase of progressive alveolar wall thickening, fluid begins to accumulate in the alveoli themselves. This final phase begins at a point where total lung water has increased by about 30 per cent (Fig. 4).

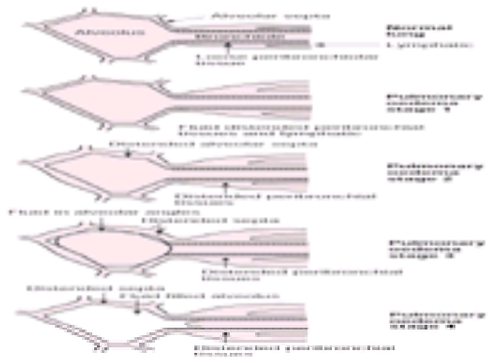

Fig. 4 Stages in the development of pulmonary oedema. Stage 1: peribronchial swelling; stage 2: distended alveolar septa; stage 3: limited accumulation of fluid in alveolar angles; stage 4: alveolar flooding. (Reproduced from Prichard JS (1982). *Edema of the lung*. Charles C. Thomas, Springfield, Illinois, with permission.)

At first, the fluid in the alveoli is confined to the alveolar angles. Subsequently, complete flooding of individual alveoli occurs. A striking feature of the microscopic appearance at this stage is the way in which alveoli are either completely filled with fluid or else have only minimal accumulation in the angles. There are no half-filled alveoli: flooding is a 'quantal' event, with flooded alveoli scattered at random throughout the affected area. Atelectasis is uncommon and air is rarely trapped, although the volume of each alveolus is smaller when fluid-filled than when air-filled.

The quantal nature of alveolar flooding arises from the interaction of surface and tissue forces (Fig. 5). The immediate precipitating factor is probably an increase in alveolar epithelial permeability caused by the distortion and swelling of the alveolar wall, which allows water to flood from the interstitium into the air space. An alternative, less likely, hypothesis is that fluid entry occurs via pores in the epithelium of the terminal airways. Irrespective of the route, the ease of fluid entry now makes the relationship between pressure and volume inverse and unstable, as explained in the legend to Fig. 5.

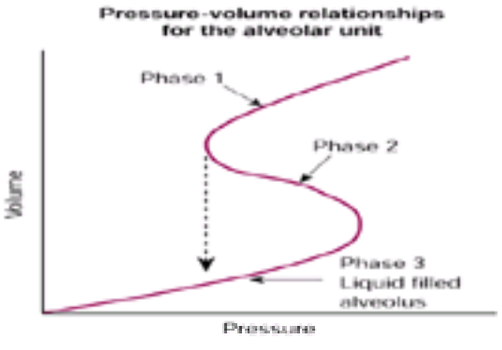

Fig. 5 Pressure–volume relationships in the alveolus. Phase 1 represents the normal alveolus lined by surfactant. Tissue elasticity, osmotic balance, and the presence of surfactant combine to produce a direct, mechanically stable relationship between pressure and volume. Phase 2 represents the situation in which alveolar permeability has increased. Any influx of fluid into the alveolus decreases the overall surface area. At these lower dimensions, surfactant is inoperative and surface tension is independent of area. The relationship between volume and pressure is that of an air bubble in liquid ($P = 2T/R$) and is unstable, a situation only resolved when the alveolus has flooded. The air volume therefore shrinks as air is expelled until phase 3 is reached. Here the remaining air is a 'bleb' at the bronchiolar orifice.

The resolution of pulmonary oedema

The extent and rapidity of resolution of pulmonary oedema depend upon its cause. Hydrostatic oedema and that due to low oncotic pressure can resolve completely and rapidly, but this is rarely the case with permeability oedema, where slow disappearance and permanent lung damage are the rule.

Resolution of hydrostatic oedema occurs in two phases: return of the capillary pressure towards normal, and then lymphatic and osmotic resorption of tissue and alveolar fluid. In cardiac failure, the shift of blood from the pulmonary to the systemic circulation by sitting up is the most powerful method of reducing capillary pressure, but other mechanisms have also been suggested, including:

1. progressive hypovolaemia (from fluid extravasation into the lung);
2. increasing plasma oncotic pressure from the relatively greater transendothelial loss of water than of plasma protein;
3. hypoxic vasoconstriction of the muscular pulmonary arteries causing a fall in capillary pressure;
4. exhaustion of sympathetic neurotransmitter in the systemic circulation with reduction in venomotor tone and left heart afterload.

The first three are all known to occur, but quantitatively their contributions are uncertain. The fourth is conjecture.

In hydrostatic oedema, once hydrostatic pressure has been reduced, fluid is removed from the interstitial space by lymphatic drainage, which can be increased for as long as 24 h after an acute episode in experimental models. Oncotic resorption into the circulation can also play a significant part. However, the mechanism of alveolar clearance is not well understood. Much fluid is removed by coughing and ciliary drainage and final resorption seems to occur as a result of active sodium ion transport, although it is uncertain whether this takes place in the alveolar or terminal airway epithelium.

The clearance mechanisms in high permeability oedema are considerably less efficient than in hydrostatic oedema because fibrin has coagulated in the interstitium, lymphatics, and alveoli, and because the tissues have often been damaged. Regeneration of epithelium and endothelium is frequently necessary. In the case of the alveolar epithelium, cell replacement is by transdifferentiation from type II pneumocytes and this is rarely complete. Whether the transdifferentiated cells are able to play an efficient part in active transport and fluid removal is unknown.

Clinical aspects

Causes of pulmonary oedema

Table 1 lists the main causes of pulmonary oedema classified according to the predominant pathophysiological mechanism. However, the clinician should never forget that more than one cause may be operating (Table 2), and must not neglect one remediable factor at the expense of another, or allow therapy itself to intensify the problem. For example, overvigorous fluid replacement following pulmonary endothelial damage may be the very factor that accelerates water and protein flow into the interstitium and provokes oedema.

Descriptions of the clinical manifestations and management of the more common diseases listed in Table 1 are provided elsewhere, but certain aspects need particular comment.

Pulmonary oedema in heart failure

Heart failure, discussed in detail in Chapter 15.2.2, is the cause of the commonest form of pulmonary oedema, but two features deserve comment. The first is the symptom of orthopnoea in which the oedema either first appears or, if already present, intensifies after a period of lying down. The cause is a shift of blood from the

systemic to the pulmonary circulation, which occurs because of the change in posture. This leads to an increase in intracapillary hydrostatic pressure, which in turn triggers oedema. The symptom is at its most dramatic in paroxysmal nocturnal dyspnoea. The second feature—and one that frequently causes confusion because it is contrary to expectation—is the tendency of the blood pressure to rise as the patient progresses into left heart failure. The cause is probably increasing sympathetic activity and circulating catecholamines, which lead to intense systemic arterial and venous vasoconstriction, thereby increasing afterload and central venous pressure inappropriately and so intensifying the development of oedema.

Loculated constrictive pericarditis

Loculated constrictive pericarditis, predominantly involving the left ventricle, can occur in patients with chronic renal failure who are undergoing dialysis. Echocardiography is helpful in diagnosis, which may be difficult because the characteristic signs of pericardial tamponade may be missing. When located posteriorly, the fluid is difficult to aspirate percutaneously, but open drainage is rarely necessary because strict attention to fluid regulation during and between dialyses usually leads to resolution.

Pulmonary venous thrombosis

This is a rare condition that is difficult to diagnose. It may be idiopathic or may be a manifestation of conditions such as polyarteritis nodosa, other vasculitic disorders, and occult neoplastic disease. The idiopathic condition is most common in middle-aged women: symptoms of increasing lassitude and breathlessness, sometimes with a low-grade fever, are the presenting symptoms. Gross-effort dyspnoea and pulmonary oedema, usually with pleural effusions, develops later. Signs of pulmonary hypertension are present. Difficulty in obtaining a clear pulmonary artery wedge pressure tracing and normal left atrial pressure (measured directly by the trans-septal route) should alert suspicion. Pulmonary artery angiography demonstrates poor segmental drainage in the regions affected by thrombosis. Open lung biopsy will confirm the diagnosis, but this is dangerous and should be undertaken only when there is real fear of missing an alternative cause of the oedema.

Left atrial myxoma, ball thrombus of the left atrium, and cor triatriatum

These are rare, but must not be missed as they are remediable by surgery. Their clinical presentation may be very similar to that of pulmonary venous thrombosis. All three conditions also enter into the differential diagnosis of tight mitral stenosis. Echocardiography is the key investigation.

High-altitude oedema

Some apparently normal people who ascend rapidly to high altitude experience acute pulmonary oedema. The condition develops only in that minority of individuals who have an exaggerated acute pulmonary arterial pressor response to hypoxia. These develop pulmonary hypertension at high altitude, with oedema possibly resulting from transarterial fluid leakage, but alternatively due to inhomogeneity of vasoconstriction and consequent extreme hyperperfusion of those areas not vasoconstricted. A further contribution may arise from the effects of vasoactive amines on the contractile filaments of endothelial cells leading to separation of endothelial junctions.

Pulmonary oedema with pulmonary arterial hypertension

Pulmonary arterial hypertension secondary to high output states, such as large shunts, can occasionally be associated with pulmonary oedema (possibly from transarterial leakage), especially following exercise. This is usually avoided because acute breathlessness is such a prominent early symptom.

Pulmonary oedema following acute intracranial lesions

A large variety of intracranial lesions may occasionally be associated with acute pulmonary oedema. It is probable that damage to the nucleus of the tractus solitarus and the hypothalamus lead to severe systemic vasoconstriction ('sympathetic storm'), which shifts blood to the pulmonary circulation, causing an extreme paroxysm of pulmonary hypertension. In addition, there is evidence to suggest that pulmonary venoconstriction also occurs, thus causing a rise in pulmonary capillary pressure even in excess of that predicted from the pulmonary arterial pressure measurements. The extreme high blood pressure in the capillaries first induces hydrostatic oedema and if sufficiently severe also damages the endothelium, leading to a less easily resolved permeability oedema.

Pulmonary thromboembolism

This may occasionally lead to florid pulmonary oedema, for which two hypotheses have been proposed: (1) local overperfusion caused by diversion of blood flow away from the occluded site; and (2) humoral alteration of permeability. It is possible that both mechanisms may play a part, and also that the causes may be different in micro- and macroemboli. Evidence for the overperfusion mechanism originates from experiments in which balloon occlusion of the major pulmonary vessels leads to oedema and increased flow of low protein lymph from other areas. As in high-altitude oedema, the site of fluid transudation is unclear: the arterial vessels have been proposed, but without strong evidence. In favour of permeability change is the observation that the lymphatic fluid following experimental microembolization is of high protein content, even when the microemboli are pharmacologically inert glass microspheres.

Expansion pulmonary oedema

Pulmonary oedema after expansion of a collapsed lung is rare, but more likely when the lung (or lobe) has been collapsed for some time. The likelihood may be reduced by ensuring that negative pressure in the pleural space during re-expansion does not exceed 10 cmH2O, that the procedure is terminated if cough develops, and that not more than 1500 ml of fluid is aspirated at any one time when collapse is related to an effusion. The mechanism is uncertain. Permeability change is likely as high protein oedemas have been found in both clinical and experimental situations. The mechanism of damage may be from toxic oxygen free radicals, as in cardiac reperfusion injury. Additional contributing factors could be loss of surfactant during the period of collapse and increased negativity of interstitial pressure during re-expansion.

Postobstructive pulmonary oedema

The initiating event is a markedly negative intrapleural pressure generated by forceful inspiratory effort against an obstructed upper airway, which is then transmitted to the pulmonary interstitial space. During normal breathing, intrapleural pressures rarely fall below –5 cmH_2O, but in upper airway obstruction the value may be as low as –50 cmH_2O. Postobstructive oedema should therefore be suspected wherever there is the rapid onset of dyspnoea, cyanosis, frothy pink sputum production, and radiological pulmonary infiltrates after the rapid relief of upper airway obstruction. The onset is usually immediate but, occasionally, delays of up to 2 h have been reported. The chronic form occurs in patients with obstructive sleep apnoea, in whom negative intrapleural pressures as low as –100 cmH_2O have been recorded.

Lymphatic oedema and lymphatic obstruction

Although, in one sense, all pulmonary oedema can be thought of as lymphatic failure, surprisingly little is known of pulmonary lymphatic failure in clinical practice. Lymphatic occlusion underlies the oedema and dyspnoea of lymphangitis carcinomatosa. In cases where cardiac failure and pneumoconioses coexist it has been found that oedema develops at lower capillary pressures than would be expected, and this has been attributed to lymphatic blockage. Mechanical lymphatic disruption is probably a contributing factor to the ease with which lungs develop oedema immediately after transplantation.

Disorders of capillary permeability

Many of the conditions associated with adult respiratory distress syndrome (see Chapter 16.5.1) can also be associated with less dramatic degrees of oedema. It is a good rule always to consider the possibility that a permeability abnormality might exist as an associated cause in all cases of pulmonary oedema. The history can be particularly helpful, particularly with regard to possible infections, use of drugs, and occupational chemicals. The possibility of oxygen toxicity should be borne in mind in all patients in intensive care.

Unilateral oedema

This frequently causes diagnostic confusion. Unilateral oedema on the same side as pre-existing lung abnormalities (ipsilateral oedema) may arise from posture (lying on one side during oedema development), increased perfusion of one lung secondary to a systemic to pulmonary shunt, unilateral venous occlusion (either from unilateral veno-occlusive disease or from extrinsic compression), or unilateral lymphatic pathology such as lymphangitis carcinomatosa. Contralateral oedema is seen where the pre-existing pathology protects that lung. Instances include congenital unilateral pulmonary artery, wyer–James–McLeod syndrome, unilateral thromboembolism, and unilateral fibrosis causing unilateral hypoxia and vasoconstriction.

The diagnosis of pulmonary oedema

The diagnosis of pulmonary oedema is by clinical observation and chest radiography.

Clinical features

The characteristic symptom of pulmonary oedema is breathlessness, probably generated by an awareness of inappropriate respiratory effort and by firing of 'J' (juxta-alveolar) receptors. This dyspnoea comes on more or less acutely in the first instance, often following exercise. Later, paroxysmal nocturnal dyspnoea develops because of postural hydrostatic factors. Only then are signs of diminished breath sounds at the bases and fine lung crepitations found. The crepitations (crackles) characteristic of pulmonary oedema are intermittent explosive sounds that each last for less than 20 ms. They are probably caused by the sudden opening of a succession of small airways, the acoustic wave being produced either by equalization of downstream and upstream pressures or by sudden alterations in the tension of the airway walls. They thus relate to the 'all-or-none' features of alveolar flooding observed physiologically. The rhonchi (musical sounds) that are sometimes heard, and which may cause considerable diagnostic confusion in the dyspnoeic patient, can arise either from bronchiolar wall oedema or from vagally mediated reflex bronchospasm.

Pulmonary oedema is never a static condition, but always either developing or regressing. The observations of Altschule, who over 40 years ago recorded the sequence of events as a patient progressed into ever greater left heart failure, are worth recalling.

1. *premonitory*: anxiety, pallor, tachycardia, raised blood pressure, cold sweaty skin;
2. *interstitial oedema*: dyspnoea, orthopnoea, cyanosis, congested neck veins, wheezing and rales;
3. *intra-alveolar oedema*: crackling rales progressing to general bubbling; cough, sputum—becoming frothy then blood-stained;
4. *shock*: clouding of consciousness;
5. *terminal*: cardiac and respiratory arrhythmias.

The clinical features of the non-haemodynamic oedemas are not dissimilar but are generally less florid.

Chest radiography

The chest radiograph is a sensitive and easily available tool for spotting early pulmonary oedema (Fig. 6 and Fig. 7). The majority of radiographical studies have been made during cardiogenic oedema where changes of oedema are necessarily superimposed on other circulatory alterations. Three successive and overlapping phases can be identified.

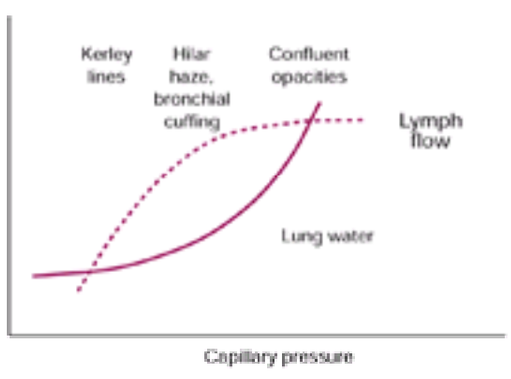

Fig. 6 Radiological signs and pulmonary pathophysiology. Kerley lines are a particularly useful radiological sign, as they occur at a stage where lymph flow and transinterstitial water flow have both increased but where appreciable tissue swelling has not yet appeared. (Reproduced from Prichard JS (1982). *Edema of the lung*. Charles C. Thomas, Springfield, Illinois, with permission.)

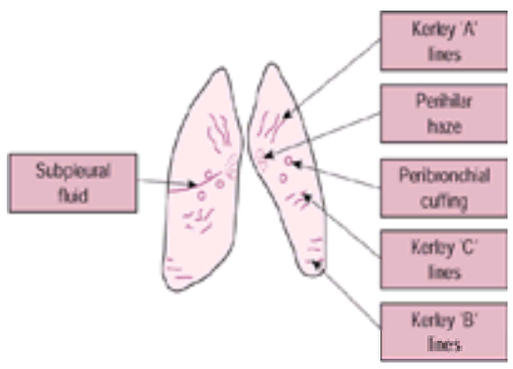

Fig. 7 Characteristic radiological appearances in interstitial oedema (see text). (Reproduced from Prichard JS (1982). *Edema of the lung*. Charles C. Thomas, Springfield, Illinois, with permission.)

Preoedema

This reflects cardiac and circulatory changes and the increased flow of fluid that occurs through the lymphatics before swelling of the tissue takes place. Usually, the cardiothoracic ratio on a posteroanterior film is more than 0.5 (>0.57 for an anteroposterior film is standard when geometry is preserved). Distension and engorgement of blood vessels occur, particularly in the upper zone with inverse changes at the bases leading to reversal of the usual pattern. Distended lymphatics subsequently become identifiable as septal lines, perilobular lines, and rosettes. Septal lines were originally identified by Kerley: type A lines are ragged, unbranched, and run centripetally towards the hilum; type C lines are fine, interlacing, and seen most easily in the central and perihilar regions; type B lines are the best known and most commonly seen. They are short, sharp, horizontal, and found in the costophrenic angles. They occur most often in pulmonary oedema due to chronic pulmonary venous hypertension. Indeed, there is excellent correlation between the density of Kerley B lines and left atrial pressure in mitral stenosis. The lines are rarely seen below a mean left atrial pressure of 13.5 mmHg, are commonly found in the region of 22 mmHg, and are invariably present when the left atrial pressure exceeds 30 mmHg. Perilobular lines and rosettes are found on close inspection in about 3 per cent of radiographs and probably represent the lymphatics running around the respiratory acini.

Interstitial oedema

This first appears in areas of 'loose' connective tissue when wedge pressure begins to rise above 15 mmHg (see sequence of oedema accumulation, above). Visible

interlobar and accessory lung fissures are the first manifestations. They are followed by perivascular and peribronchial cuffs, which contribute, respectively, to the homogeneous circular shadows formed by the already distended vessels and to the 'ring' shadows around bronchi seen close to the hilum. Micronoduli consist of small, round densities (<3 mm) arising from the accumulation of fluid around the smaller blood vessels. Blurring and hazing of the hilar regions represent the beginning of true alveolar septal interstitial oedema and, in hydrostatic oedema, begin at wedge pressures of around 20 mmHg. A diffuse increase in lung density (clouding) represents the final phase.

Alveolar oedema

This starts when wedge pressure reaches 25 to 28 mmHg. It is seen as a 'fluffy' loss of lucency, either around the hila in a 'butterfly' or 'batswing' pattern, or predominantly in the lower zones, usually reflecting a 'gravitational' distribution. Associated changes are the development of effusions and a loss of lung volume caused by a fall in lung compliance.

Radiologically, the permeability oedemas follow the pattern of the hydrostatic except that: (1) the distribution of alveolar oedema tends to be patchy; and (2) the characteristic vascular and cardiac changes are not present.

The presence of pre-existing lung disease—particularly chronic obstructive pulmonary disease—may modify the radiological appearance of pulmonary oedema considerably. Hyperinflation may render the silhouette of a large heart unremarkable; with the onset of interstitial oedema, a hyperinflated lung may shrink to normal size; the distribution of oedema shadowing may be patchy and only evident where parenchyma is sufficiently preserved; Kerley lines may be difficult or impossible to identify.

Other investigations

Computed tomography (**CT**) scanning is not necessary for the diagnosis of pulmonary oedema, but the appearances are characteristic, with thickening and increased visualization of interlobular septa and associated thickening of subpleural and peribronchial interstitial spaces. Alveolar oedema leads to varying degrees of alveolar consolidation. As in the plain radiograph, there may be associated heart and vascular changes and pleural effusions.

Pulmonary function in oedema of the lung

The oedematous lung shows a mixture of restrictive and obstructive defects, although the former dominate. The restrictive component arises from decreased compliance, which is a result of vascular congestion (in cardiogenic and fluid overload oedema), interstitial oedema, and surfactant washout. Of these, the interstitial oedema contributes surprisingly little, so that, to start with, restrictive changes are indicative of an engorged vascular system and later, of alveolar flooding.

Sometimes, airflow resistance may cause easily audible rhonchi and a reduction in forced expiratory volume in 1 s/forced vital capacity (**FEV1/FVC**), but, more usually, it is difficult to detect by simple methods because it occurs predominantly in the small airways of 1 to 2 mm diameter, which contribute relatively little to overall resistance. There could be a number of causes for such airflow obstruction. In the preoedematous phase of heart failure the smallest airways may be compressed by the distension of adjacent vessels in the bronchovascular bundle. In frank interstitial oedema it has been suggested that perivascular cuffing could do the same, but this has not been substantiated. However, submucosal oedema and vagally mediated reflex bronchoconstriction are proven and probably responsible for most of the effect. Restriction and obstruction may combine to reduce vital capacity, and serial measurements of this can be a good index of severity of and recovery from pulmonary oedema.

Tachypnoea is a prominent feature of all forms of pulmonary oedema. This is associated with a low tidal volume, but total ventilation (V_E)—both at rest and during exercise—is high relative to the prevailing level of carbon dioxide consumption. Most of this increase is accounted for by deadspace ventilation hence, unless the patient is progressing into severe alveolar oedema (see below), he or she remains normocapnic. The mechanism underlying this tachypnoea is uncertain. Hypoxic effects upon the central carbon dioxide chemostat do not appear to be an explanation, and recent evidence using perialveolar local anaesthetic suggests that the 'J' (juxta-alveolar) receptor—an unmyelinated nerve ending in the vicinity of the alveoli, which responds to interstitial swelling and distension—is only involved at more severe levels of oedema. Respiratory muscle fatigue is a possibility.

In acute, severe oedema the usual blood gas abnormalities are hypocapnia and hypoxia. The hypoxia is a result of ventilation/perfusion mismatching. The hypocapnia is accounted for by the reflex tachypnoea leading to an increased alveolar ventilation, which more than compensates for the increased pulmonary deadspace (volume of deadspace/volume of tidal air ratios, V_D/V_T ratio). However, in about 20 per cent of severe cases, hypercapnia (with respiratory acidosis) is seen, even when no chronic airflow disease coexists. A number of mechanisms have been proposed, including uncontrolled oxygen administration accompanied by low central carbon dioxide sensitivity, respiratory muscle fatigue, and severe ventilation perfusion imbalance. In acute oedema, blood gas abnormalities are usually accompanied by a mild metabolic acidosis, but occasionally the base excess may exceed –15 mmol/l. This frank metabolic acidosis is most likely in patients with severe oedema who already have carbon dioxide retention.

In the more chronic permeability oedemas—as in adult respiratory distress syndrome—the overwhelming problem is continuing severe hypoxaemia. Three mechanisms have been proposed: (1) diffusion impairment; (2) low ventilation/perfusion $_2$ ($_A$/) values; and (3) shunt ($_A$/ <0.005). Using both the arterial oxygen response to changing fractional inspired oxygen concentration (**FiO2**) and the inert gas technique, it has been shown that diffusion impairment plays little part. Shunt and ventilation/perfusion mismatch are more important, but their contribution varies greatly from patient to patient.

Treatment of pulmonary oedema

Pulmonary oedema may result from increased microvascular hydrostatic pressure, decreased tissue interstitial pressure, decreased plasma colloid oncotic pressure, increased microvascular permeability, or impaired lymphatic drainage. Treatment of each form should include measures designed to reverse the specific cause. However, with the exception of a reduction of elevated hydrostatic pressure and relief of upper airway obstruction, these are rarely available and the clinician has to rely upon general supportive measures combined with meticulous attention to fluid balance and monitoring of plasma oncotic pressure. (See Chapter 16.1 and Chapter 16.5.2 for discussion of the clinical approach to the severely ill and breathless patient.)

Acute cardiogenic and fluid overload pulmonary oedema

By far the most common causes are acute and chronic left-sided heart disease, although the overenthusiastic use of intravenous fluid regimen containing normal saline are frequently an additional factor. The patient is most comfortable in the 'trunk up, legs down' position to help pool blood in the dependent parts and reduce central venous pressure. Oxygen, diuretics, intravenous nitrates, and morphine are the fundamentals of treatment. Thigh cuffs inflated to occlude venous return can help as a form of a bloodless phlebotomy and venesection, and removal of 200 to 500 ml of blood is an effective treatment when other measures are not available.

Hypoxia is relieved with a standard face mask, nasal prongs, or reservoir bag, delivering oxygen at high flow rates—up to 10 litres/min—providing an inspired concentration of up to 60 per cent. The fractional inspired oxygen concentration given should be as high as is necessary to keep the arterial partial pressure of oxygen near to the normal level—and no more—because, in permeability oedema: (1) high oxygen levels may lead to absorption atelectasis in areas of low ventilation/perfusion; (2) oxygen toxicity may become a problem where prolonged administration is necessary. If oxygen administration at normal airway pressure cannot maintain the arterial partial pressure of oxygen and/or hypercapnia develops, then application of a continuous positive airway pressure (**CPAP**) mask, non-invasive ventilation, or tracheal intubation and intermittent positive pressure ventilation will need to be considered (see Chapter 16.5.2).

A bolus dose of furosemide (frusemide) (or other loop diuretic), administered intravenously, is usually given. This acts both as a venous dilator, and as a diuretic. Diuretics are at their most valuable where pulmonary oedema is a component of congestive cardiac failure and where the volume of extracellular fluid is generally increased. By contrast, when left ventricular failure has come on acutely, significant fluid retention has not occurred and pulmonary oedema is a result of fluid shift from the systemic circulation, then overvigorous use of diuretics runs the risk of causing hypovolaemia. If the patient is *in extremis* or heavily sedated, it is wise to catheterize the bladder for, as a result of the diuresis, bladder distension may induce intense reflex systemic vasoconstriction leading, on occasion, to disastrous cardiac overload.

Intravenous nitrates, in particular isosorbide dinitrate at a dose between 2 and 20 mg/h, can effectively reduce venous pressure and alleviate pulmonary oedema.

Arterial hypotension is the effect that usually limits dosage, the combination of heart failure with low blood pressure and pulmonary oedema being difficult to treat and of grave prognosis (see Chapter 16.2).

Morphine acts centrally to relieve the distress of dyspnoea and also dilates the systemic venous system. This reduces venous filling pressure to the heart and shifts blood from the pulmonary to the systemic circulation. It is best administered by slow intravenous injection in a total dose of 2 to 10 mg at a rate of 2 mg/min, together with an appropriate antiemetic.

Aminophylline has diuretic, bronchodilator, cardiac inotropic, and respiratory muscle inotropic effects. Its use would seem logical, but it is now scarcely ever given because of its capacity to induce arrhythmias and the availability of other effective treatments. If it is to be administered, a dose of 250 to 500 mg should be given intravenously in not less than 20 min.

Other types of pulmonary oedema

A reduction in oncotic pressure may contribute to pulmonary oedema, as in crystalloid fluid overload, hepatic failure, or nephrotic syndrome. In fluid overload, the most appropriate therapy is the use of diuretics, for these not only reduce the extracellular and blood volumes but also return osmotic pressure towards normal. It is more difficult to be certain about therapy in true hypo-oncotic states. Even where there is no evidence of endothelial damage, the effects of salt-free albumin and plasma concentrate are disappointing.

In high-permeability pulmonary oedema the best form of management would be to block the inappropriate activation and progress of the cascades that are responsible for the condition (see Chapter 16.5.1). Unfortunately, there is no therapy that allows this at present and management, aside from aiming to treat any precipitating disorder, is supportive (see Chapter 16.5.2).

---

*It is with regret that we report the death of Dr J. S. Prichard. Much of his chapter in the third edition has been retained here.

Further reading

Anonymous (1986). Adult respiratory distress syndrome. *Lancet* **i**, 301–3.

Anonymous (1986). The enigma of breathlessness. *Lancet* **i**, 891–2.

Artigas A, *et al*. (1992). *The adult respiratory distress syndrome*. Churchill Livingstone, Edinburgh.

Egan EA (1983). Fluid balance in the air filled alveolar space. *American Review of Respiratory Disease* **127**, 37–9.

Guyton AO, Lindsey AW (1959). Effect of elevated left atrial pressure and decreased plasma protein concentration upon the development of pulmonary oedema. *Circulation Research* **7**, 649.

Kreiger BP, de la Hoz RE (1999). Altitude related pulmonary disorders. *Critical Care Clinics* **15**, 265–80.

Morgan PW, Goodman LR (1991). Imaging of diffuse lung diseases: pulmonary oedema and adult respiratory distress syndrome. *Radiologic Clinics of North America* **29**, 943–63.

O'Brodovich H (1990). When the alveolus is flooding, it is time to man the pumps. *American Review of Respiratory Disease* **142**, 1247–8.

Pang D, *et al.* (1998). The effects of positive pressure airway support on mortality and the need for intubation in cardiogenic pulmonary oedema: a systematic review. *Chest* **114**, 1185–92

Prichard JS (1982). *Edema of the lung*. Charles C. Thomas, Springfield, IL.

Sacchetti AD, Harris RH (1998). Acute cardiogenic pulmonary edema. What's the latest in emergency treatment? *Postgraduate Medicine* **103**, 145–7, 153–4, 160–2.

Simon RD (1993). Neurogenic pulmonary edema. *Neurologic Clinics* **11**, 309–23.

Szidon PS (1989). Pathophysiology of the congested lung. *Cardiology Clinics* **7**, 39–48.

Trimby J, *et al*. (1990). Mechanical causes of pulmonary oedema. *Chest* **98**, 973–9.

Veeraraghavan S, Koss MN, Sharma OP (1999). Pulmonary veno-occlusive disease. *Current Opinion in Pulmonary Medicine* **5**, 310–13.

Wiedemann HP, Matthay MA, eds (2000). Adult respiratory distress syndrome. *Clinics in Chest Medicine* **21**, 401–620.

15.15.3.1 Deep venous thrombosis and pulmonary embolism

Paul D. Stein and J. Firth

Introduction

Pulmonary embolism is a complication of deep venous thrombosis. Among patients with pulmonary embolism, the thrombi in 80 per cent or more originate in the legs. Deep venous thrombosis and pulmonary embolism are sometimes described together, using the term 'thromboembolism'. Strategies of management have been developed which are based on the diagnosis of either pulmonary embolism or deep venous thrombosis, provided the patient has good respiratory reserve. Treatment with anticoagulants is the same for both. Some believe, however, that patients can be managed better if it is known whether acute pulmonary embolism is present, even if a diagnosis of deep venous thrombosis is already established.

Mortality of untreated deep venous thrombosis and pulmonary embolism

The frequency of fatal pulmonary embolism in patients with untreated deep venous thrombosis has diminished as diagnostic tests have made it possible to diagnose deep venous thrombosis before it becomes extensive. In 1955, prior to the use of sensitive non-invasive tests for the early detection of deep venous thrombosis, the risk of fatal pulmonary embolism in untreated patients with clinically apparent deep venous thrombosis was 37 per cent. With the use of radioactive fibrinogen scintiscans, the risk of fatal pulmonary embolism in patients with untreated deep venous thrombosis, most of which was subclinical, was approximately 5 per cent.

Early diagnosis has also reduced the risk of death in those with pulmonary embolism. In the early 1960s, the mortality in untreated patients with acute pulmonary embolism, diagnosed on the basis of clinical features, was 26 to 37 per cent. An additional 36 per cent died of recurrent pulmonary embolism. In recent years, among patients with mild pulmonary embolism who inadvertently escaped treatment, the mortality was 5 per cent. The mortality of untreated patients with mild pulmonary embolism was comparable with the mortality from fatal pulmonary embolism in untreated patients with subtle deep venous thrombosis.

Deep venous thrombosis

Incidence and pathology

Deep venous thrombosis is often silent. In one study of those with deep venous thrombosis detected by screening with ^{125}I-labelled fibrinogen scans, clinical evidence was present in 49 per cent. The percentage of patients with acute pulmonary embolism who have clinically detectable deep venous thrombosis decreases with decreasing severity of pulmonary embolism, which implies more subtle disease. Of patients who died from acute pulmonary embolism, 53 per cent had clinically identified deep venous thrombosis. In patients with massive or submassive acute pulmonary embolism, most of whom survived, 34 per cent had clinically identifiable deep venous thrombosis. Among patients with mild as well as severe acute pulmonary embolism, only 15 per cent had clinically apparent deep venous thrombosis.

Proximal deep venous thrombosis was found at autopsy in 22 per cent of patients who died in a tertiary care hospital. Thrombosis of the leg veins usually occurs without inflammation. Inflammation of the walls of the veins, when it occurs, is usually secondary to the thrombus. No clear evidence indicates that inflammation of the veins prevents embolization, or that embolization is more frequent in those patients with thrombi not associated with venous inflammation.

In the past, patients who had thrombosed leg veins accompanied by signs of inflammation were diagnosed as having thrombophlebitis, based on the presumption that the primary event was inflammation of the walls of the veins. Patients with no clinical signs in the lower extremities who had thrombosis that resulted in pulmonary embolism were said to have phlebothrombosis. Histological investigations have not supported a distinction between the clinical diagnoses of thrombophlebitis and phlebothrombosis. Thrombus can induce inflammation in the underlying wall of the vein, and this inflammation in some patients is extensive enough to produce pain, tenderness, swelling, and fever.

The valve pockets are a frequent site of origin of thrombi. At autopsy, clinically unsuspected deep venous thrombosis is often extensive, causing collateral circulation around occlusions and dilatation of collateral veins. When the veins of the thigh and the calf are thrombosed in continuity, the thrombi in the calf are older than those in the thigh, which suggests that the thrombosis extended from the veins of the calves to the veins of the thighs.

Clinical diagnosis

Patients may complain of pain or swelling of the leg, but physical examination remains the means by which attention is usually drawn to the potential diagnosis of deep venous thrombosis.

Deep venous thrombosis sometimes, but not always, leads to swelling of the leg. If restricted to the popliteal and calf veins, swelling is confined to below the knee, but if thrombosis involves the femoral and pelvic veins (or inferior vena cava), then swelling of the thigh is also expected. A difference of circumference of the calves of greater than 1 cm, measured 10 cm below the tibial tuberosity, is abnormal. It is important to repeat the measurement of diameter of the calves and thighs at frequent intervals: proximal extension of a thrombus is likely to cause increased swelling. To allow repeated measurements to be made from a fixed point, it is good practice for the position of the first measurement to be marked indelibly on the patient's skin.

Homans' sign is positive when active and/or passive dorsiflexion of the foot is associated with any of the following: (i) pain, (ii) incomplete dorsiflexion (with equal pressure applied) to prevent pain, or (iii) flexion of the knee to release tension in the posterior muscles with dorsiflexion. This sign was present in 44 per cent of patients with deep venous thrombosis of the lower leg, and in 60 per cent of patients with femoral venous thrombosis. The elicitation of pain with inflation of a blood pressure cuff around the calf to 60 to 150 mmHg has been recommended as a test for deep venous thrombosis. This test, however, has not been shown to be more helpful than the assessment of direct tenderness or leg circumference.

In one study the sensitivity of oedema, erythema, calf tenderness, palpable cord, or Homans' sign alone, or greater than 1 cm calf asymmetry alone was 55 to 80 per cent, but the specificity only 49 per cent. The combination of one of these signs plus greater than 1 cm ipsilateral calf asymmetry increased the specificity to 87 per cent, but decreased the sensitivity to 15 to 33 per cent. The specificity increased to 91 per cent with one of these signs in combination with greater than 2 cm calf asymmetry. Only 3 to 10 per cent of patients had one or more qualitative signs plus greater than 3 cm ipsilateral calf asymmetry: in these, the specificity for deep venous thrombosis was 96 per cent.

Other clinical features of deep venous thrombosis, whose sensitivity and specificity have not been tested, include increased temperature on the affected side, cyanotic discoloration of the limb, and persistent engorgement of superficial veins. Superficial varicose veins almost always empty when the patient lies down: if they remain engorged, this suggests problems with drainage through the deep veins. In very rare cases, tense venous oedema can cause arterial compression and venous gangrene.

The clinical diagnosis of deep venous thrombosis is not always straightforward. Many of the findings described above can also be found in those with muscular strains and bruising, ruptured Baker's cyst or plantaris tendon, superficial thrombophlebitis, cellulitis, and other traumatic conditions. Given the sinister nature of untreated deep venous thrombosis, it is important to confirm or refute (so far as is possible) the diagnosis with appropriate investigations whenever clinical suspicion is aroused, unless the general condition of the patient makes this inappropriate.

Investigation

Detection of the physical presence of thrombus in leg veins

The 'gold standard' is contrast venography, but this can be unpleasant for patients, time consuming for radiology departments, and is expensive. This has driven the search for acceptable non-invasive methods of diagnosis. Among patients with deep venous thrombosis proven by contrast venography, B-mode ultrasonography using compression showed a 95 per cent sensitivity in symptomatic patients. In asymptomatic patients who were evaluated because of a high risk of deep venous thrombosis, venous compression ultrasound showed a sensitivity of only 67 per cent. Regarding veins of the calves, venous compression ultrasound was 93 per cent sensitive in symptomatic patients, but only 26 per cent sensitive in asymptomatic high-risk patients with deep venous thrombosis. In all instances, specificity was 97 to 99 per cent. Impedance plethysmography was 86 to 94 per cent sensitive for detection of deep venous thrombosis of the thighs, but the sensitivity was only 25 per cent for the veins of the calves. Specificity was high, 97 per cent. In most centres contrast venography and impedance plethysmography have been replaced by B-mode ultrasonography as the preferred first-line diagnostic technique.

Venous phase contrast enhanced spiral computed tomography (CT) appears promising for imaging the veins of the pelvis and thighs. Spiral CT imaging of the veins of the lower extremities and pelvis in combination with spiral CT imaging of the chest potentially offers a comprehensive study for thromboembolism. The sensitivity and specificity of contrast enhanced spiral CT, however, are still being evaluated.

Magnetic resonance imaging, tested in small numbers of patients, was 100 per cent sensitive for veins of the thighs and pelvis and somewhat less sensitive (85 per cent) for veins of the calves. Specificity in all regions was 95 to 100 per cent. Its problem is cost and availability.

Fibrinogen-uptake radionuclide scanning was used extensively in the 1960s. It is more sensitive for deep venous thrombosis in the calves than in the thighs. In view of the greater risk of pulmonary embolism with deep venous thrombosis of the thighs than of the calves, the value of fibrinogen-uptake scanning is limited.

Lately, technetium apcitide labelling of glycoprotein IIb/IIIa receptors expressed on activated platelets has permitted radionuclear imaging of acute proximal deep venous thrombosis and pelvic vein thrombosis. Its sensitivity has been incompletely evaluated.

Detection of evidence of thrombus within the circulation: D-dimer

D-Dimer is a specific degradation product released into the circulation by endogenous fibrinolysis of a cross-linked fibrin clot. A D-dimer measured by enzyme-linked immunosorbent assay (ELISA) below a cut-off of 300 to 540 ng/ml (the values differ slightly from one study to another) make the diagnosis of deep venous thrombosis (or pulmonary embolism) unlikely. A concentration of D-dimer above any particular cut-off level is not useful for making a positive diagnosis because of the large number of false positive tests. However, conventional ELISA assays are cumbersome and not suited for emergency use, which limited the practical utility of D-dimer measurements until the development of rapid ELISA assays. These provide the best balance of sensitivity and specificity among the various assays for the safe diagnostic handling of patients with suspected deep venous thrombosis and pulmonary embolism.

Strategy for diagnosis

For reasons given above, subjecting all patients who might have a deep venous thrombosis to contrast venography is not an attractive option for patients, physicians, radiologists or those who pay for health care. Much effort has therefore been expended in trying to develop management algorithms that will identify those at very low risk of deep venous thrombosis (or pulmonary embolism), who can then be spared invasive tests. These algorithms typically use scoring systems to stratify the clinical probability that the particular patient has a deep venous thrombosis (or pulmonary embolus), and then proceed to D-dimer testing of those with low probability. Patients with a low clinical probability and a negative D-dimer test are not investigated further. Patients with either a high clinical probability or a low clinical probability but elevated D-dimer proceed to tests for the presence of thrombus in the leg veins, typically by ultrasonography. Examples of a pre-test scoring system and management algorithm are shown in Table 1.

Prevention, treatment, and complications of deep venous thrombosis

The prevention of deep venous thrombosis is critical in the prevention of pulmonary embolism. Deep venous thrombosis itself carries extensive morbidity irrespective of pulmonary embolism. Severe postphlebitic syndrome (venous ulcer or combinations of pain, cramps, heaviness, pruritus, paraesthesia, pretibial oedema, induration, hyperpigmentation, venous ectasia, redness or pain with calf compression) occurs in 9 per cent of patients by 5 years after a 3-month course of treatment with anticoagulants. There is some evidence that the likelihood of this problem developing can be reduced by use of elastic stockings at the time of acute deep venous thrombosis and afterwards.

Proximal deep venous thrombosis leads to pulmonary embolism more frequently than deep venous thrombosis confined to the calf. Even so, symptomatic isolated calf vein thrombosis, limited to the calves and diagnosed by non-invasive testing, should be treated with anticoagulation for 3 months. If anticoagulation cannot be given, serial non-invasive studies of the leg veins should be performed over the next 7 to 14 days to assess for proximal extension of the thrombus.

Risk factors for deep venous thrombosis are almost certainly the same as those for pulmonary embolism (see later). Recommendations for the prevention of deep venous thrombosis are shown in Table 2, Table 3, Table 4 and Table 5, and for treatment in Table 6.

Acute pulmonary embolism

Incidence

Acute pulmonary embolism is the third most common cardiovascular problem after coronary heart disease and stroke. The incidence of objectively diagnosed acute pulmonary embolism in a tertiary care hospital is probably higher than in most short-stay hospitals, but in one such centre a definitive diagnosis of pulmonary embolism was made in 0.3 to 0.4 per cent of patients. Inclusion of extrapolated data from autopsy studies and the estimated frequency of pulmonary embolism in patients with non-high probability lung scans increases the calculated incidence of acute pulmonary embolism to 1.0 per cent of patients admitted to hospital. Silent pulmonary embolism was not included in these calculations, and this has been reported in 38 to 51 per cent of patients with deep venous thrombosis.

In one major study, the incidence of acute pulmonary embolism was linearly related to age: more than half of patients were between 65 and 85 years of age, while fewer than 5 per cent were under age 24. Occasionally, however, young adults or adolescents had pulmonary embolism. In patients 50 years of age or older, the incidence of pulmonary embolism was higher among women. By contrast, the incidence was comparable in men and women under age 50 years, suggesting that

childbirth and oral contraceptives had little impact.

In autopsy studies, when the pathologist judged that pulmonary embolism contributed to death or caused death, the diagnosis was unsuspected ante-mortem in 70 per cent. This was true in several series, encompassing university as well as non-university hospitals. Some of the unsuspected pulmonary embolism was in patients who died of malignancy, in whom a diagnosis of pulmonary embolism may (appropriately) not have been actively pursued. But the time-honoured point remains as valid today as ever: a high index of suspicion is necessary to reduce the number of patients with unsuspected pulmonary embolism.

Predisposing factors

Immobilization, irrespective of the cause, is the most frequent predisposing factor (Table 7). Immobilization of even 1 or 2 days may predispose to pulmonary embolism and most patients with pulmonary embolism are immobilized for less than 2 weeks.

Whether obesity is a predisposing factor is controversial. Most patients with pulmonary embolism had smoked at one time or continued to smoke at the time of their pulmonary embolism.

Thromboembolic events have been linked to high oestrogen content in oral contraceptives (greater than 50 µg), but this association has been questioned. Irrespective of whether the risk ratio is increased among women who take oral contraceptives, the absolute risk of venous thromboembolism is low. The United States Food and Drug Administration (**FDA**) in 1980 recommended the use of the lowest possible dose of oestrogen for birth control. Childbearing and oral contraceptives did not result in a higher incidence of pulmonary embolism among women under 50 years of age compared with men. The combination of surgery and oral contraceptives may increase the risk of thromboembolism. This is true even with oral contraceptives that have a low oestrogen content. For further discussion of these and other risks associated with oestrogen use, see Chapter 13.19 and 13.20.

There has been much recent interest in the subject of genetic predisposition to thromboembolism. Heterozygosity for the Factor V Leiden mutation increases susceptibility three- to eightfold in a variety of circumstances. Other genetic and acquired thrombophilic factors include protein C deficiency, protein S deficiency, antithrombin deficiency, prothrombin 20201A, high concentration of factor VIII, hyperhomocystinaemia, heparin cofactor II deficiency, dysfibrinogenaemia, decreased levels of plasminogen, decreased levels of plasminogen activators, antiphospholipid antibodies, heparin-induced thrombocytopenia, and myeloproliferative disorders. For full discussion of this and related issues, see Chapter 15.15.3.2.

Patients with the nephrotic syndrome are known to be at particularly high risk of deep venous thrombosis and pulmonary embolism.

Syndromes of acute pulmonary embolism

Pulmonary embolism can present in diverse ways. The syndrome of pleuritic pain or haemoptysis, in the absence of circulatory collapse, is the most frequent mode of presentation of acute pulmonary embolism. It occurred in 60 per cent of patients recruited in a collaborative investigation, the Prospective Investigation of Pulmonary Embolism Diagnosis (**PIOPED**). A syndrome of dyspnoea in the absence of haemoptysis or pleuritic pain or circulatory collapse occurred in 25 per cent. Circulatory collapse (systolic blood pressure less than 80 mmHg or loss of consciousness) was an uncommon mode of presentation, occurring in 15 per cent. Recognizing that patients with circulatory collapse may not be candidates for recruitment into trials of diagnostic investigations or therapies, and patients with circulatory collapse often die within the first few hours, it may be that the incidence of circulatory collapse as determined from such investigations is falsely low. Patients with pulmonary infarction have less severe pulmonary embolism than those with isolated dyspnoea, and those with circulatory collapse probably have the most severe of all. The clinical characteristics of patients with acute pulmonary embolism often reflect the severity of the syndrome.

Symptoms of acute pulmonary embolism

The clinical characteristics of acute pulmonary embolism have been derived from prospectively acquired data of patients recruited in trials of diagnostic investigations or therapies. Such trials clearly only include those in whom there was sufficient clinical suspicion to lead physicians to obtain diagnostic tests: whether subtle pulmonary embolism was overlooked is undetermined. The specificity of signs, symptoms, and ordinary clinical tests, among patients with suspected pulmonary embolism in whom the diagnosis was eventually excluded by pulmonary angiography, was low. The specificity of such tests when evaluated in a normal population, however, would be higher.

To characterize the diagnostic features of acute pulmonary embolism, it is useful to evaluate patients in whom the diagnosis is not confused by pre-existing cardiac or pulmonary disease. When manifestations related to coexistent disease are excluded, dyspnoea is the most common symptom, occurring in 73 per cent (Table 8). Pleuritic chest pain (66 per cent of patients with pulmonary embolism) occurred much more often than haemoptysis (13 per cent of patients with pulmonary embolism).

Cough was common (37 per cent) among patients with pulmonary embolism and could be non-productive, or productive of purulent or bloody sputum. When haemoptysis occurred, the sputum typically was blood-streaked, but can be pure blood or blood-tinged. Purulent sputum was present in 7 per cent of cases.

The angina-like pain that occurred in a few (4 per cent) patients with pulmonary embolism did not radiate to either arm or to the jaw, which can assist in distinguishing it from true angina. It was usually located in the anterior chest and it was described as heavy.

Signs of acute pulmonary embolism

Tachypnoea (respiratory rate 20/min or greater) was the most common sign of acute pulmonary embolism among patients with no prior cardiac or pulmonary disease (70 per cent of patients) (Table 9). Tachycardia (heart rate greater than 100/min) occurred in 30 per cent; the pulmonary component of the second sound was accentuated in 23 per cent; and deep venous thrombosis was clinically apparent in 11 per cent. A right ventricular lift, third heart sound, or pleural friction rub were uncommon, each occurring in 4 per cent or less of patients with pulmonary embolism.

Most patients with pulmonary embolism who had rales (crepitations) had pulmonary parenchymal abnormalities, atelectasis, or a pleural effusion on the chest radiograph. Rales, therefore, appeared to relate to the effects of pulmonary infarction or atelectasis.

Among patients with pulmonary embolism and no other source of fever, a temperature below 39.9°C was present in 12 per cent and fever of 39.9°C or higher occurred in 2 per cent. Fever in patients with pulmonary haemorrhage/infarction was not more frequent than among those with no pulmonary haemorrhage/infarction. Clinical evidence of deep venous thrombosis was often present in patients with pulmonary embolism and otherwise unexplained fever.

Combinations of signs and symptoms

Dyspnoea or tachypnoea (respiratory rate 20/min or greater) were present in 90 per cent of patients with acute pulmonary embolism. Dyspnoea or tachypnoea or pleuritic pain were present in 97 per cent. Dyspnoea or tachypnoea, pleuritic pain, radiographic evidence of atelectasis, or a parenchymal abnormality were present in 98 per cent of patients. The remaining 2 per cent had either deep venous thrombosis or an unexplained low PaO_2. In the absence of dyspnoea or tachypnoea or pleuritic pain, pulmonary embolism was rarely diagnosed.

Accuracy of clinical assessment

To emphasize the point that the diagnosis of pulmonary embolism is difficult to make, senior staff physicians and postgraduate fellows taking part in the PIOPED study were uncertain of the diagnosis in the majority of patients. Using individual clinical judgement without any specific predetermined criteria, senior staff were correct in the diagnosis in 88 per cent of cases when their clinical assessment indicated a high probability of pulmonary embolism. When their clinical assessment indicated a low probability of pulmonary embolism, senior staff correctly excluded pulmonary embolism in 86 per cent. Postgraduate fellows, on the basis of clinical assessment, were more accurate in excluding pulmonary embolism than they were in making the diagnosis.

Differential diagnosis of pulmonary embolism

The commonest presentation of acute pulmonary embolism is with dyspnoea and/or pleuritic chest pain. There are, however, several other possible causes of these symptoms, the commonest being musculoskeletal pain and pneumonia. Musculoskeletal chest pain can be very similar to that caused by pleurisy, and splinting of the chest can lead to a perception of breathlessness, which may be exacerbated by anxiety. If there is an obvious history of local trauma to the chest, then patients will rarely present to the physician, but it is worthwhile to ask specifically whether there has been any trauma or unaccustomed physical activity, whether the pain can be brought on by particular movements, and to examine carefully for local tenderness of ribs, muscles, or costal margins. Tenderness can sometimes be found in cases of pleurisy, but with appropriate history clearly supports a diagnosis of musculoskeletal pain. Pneumonia complicated by pleurisy can cause dyspnoea and chest pain. Important features to look for in the history include preceding systemic upset ('flu-like' symptoms), high fever, and rigors; and on examination, high fever, 'toxic' appearance, and chest signs of pneumonic consolidation. If a positive diagnosis of another cause of dyspnoea and/or pleuritic chest pain cannot be made, then the default position should be to assume that the patient has pulmonary embolism until proven otherwise.

Investigation

Simple laboratory tests

Among patients with pulmonary embolism in whom a possible or definite cause for leucocytosis was eliminated, 80 per cent had a normal white blood cell count, 6 per cent had a count of 10 100 to 11 900/mm^3, and 13 per cent had a count of 12 000/mm^3 or greater. None had a white blood cell count that was 20 000/mm^3 or greater. Leucocytosis was not more frequent in patients with the pulmonary haemorrhage/infarction syndrome than in other patients with acute pulmonary embolism.

Electrocardiogram

Electrocardiographic abnormalities are described among patients with pulmonary embolism who had no prior cardiopulmonary disease (Table 10). A normal electrocardiogram was shown in 30 per cent of patients with acute pulmonary embolism. Only 5 per cent of the patients with acute pulmonary embolism had atrial fibrillation or atrial flutter. Atrial flutter or atrial fibrillation in patients with acute pulmonary embolism is nearly always limited to individuals with prior heart disease.

Abnormalities of the ST segment and T wave are by far the most frequent electrocardiographic manifestation of acute pulmonary embolism. Non-specific ST-segment or T-wave changes were observed in 49 per cent of patients in whom the severity of pulmonary embolism ranged from mild to severe.

Electrocardiographic manifestations of acute cor pulmonale (S1Q3T3, complete right bundle branch block, P pulmonale, or right axis deviation) were less common than ST-segment or T-wave changes. One or more of these abnormalities occurred in 26 per cent of patients with submassive or massive acute pulmonary embolism not associated with cardiac or pulmonary disease (32 per cent of patients with massive pulmonary embolism).

The electrocardiogram may simulate inferior infarction with Q waves and T-wave inversion in leads II, III, and aVF or anteroseptal infarction characterized by QS or QR waves in V1 and T-wave inversion in the right precordial leads. The development of Q waves and extensive T-wave inversion in the anterior and lateral leads has also been observed. A pseudoinfarction pattern, however, was seen in only 3 per cent of patients with no prior cardiopulmonary disease who had pulmonary embolism that ranged in severity from mild to massive.

New leftward shifts of the frontal plane axis in pulmonary embolism are frequent. Among patients with acute pulmonary embolism, left axis deviation was more frequent than right axis deviation. Low-voltage QRS complexes were observed in 3 per cent.

Inversion of the T waves was the most persistent electrocardiographic abnormality. Inversion of the T wave disappeared in only 22 per cent of patients 5 or 6 days after the pulmonary embolism was diagnosed, although it resolved in 49 per cent by 2 weeks. Depression of the ST segment tended to resolve somewhat faster. Abnormalities of depolarization resolved more quickly than abnormalities of repolarization. Well over half of the electrocardiograms that showed pseudoinfarction, S1S2S3, S1Q3T3, right ventricular hypertrophy, or right bundle branch block, no longer showed these abnormalities 5 or 6 days after the diagnosis was made.

Patients with ST-segment abnormalities, T-wave inversion, pseudoinfarction patterns, S1Q3T3 patterns, incomplete right bundle branch block, right axis deviation, right ventricular hypertrophy, or ventricular premature beats had larger perfusion defects on the lung scan or larger defects on the pulmonary arteriogram than those with normal electrocardiograms. Such patients had higher pulmonary arterial pressures and in general had low partial pressure of oxygen in arterial blood. Acute ventricular dilatation is speculated to be the most likely cause of the electrocardiographic changes.

Chest radiograph

The findings on the plain chest radiograph, when used together with the history, physical examination, electrocardiogram, and simple laboratory tests assist in identifying a syndrome of pulmonary embolism. The chest radiograph, when normal in a patient who is dyspnoeic, may hint that pulmonary embolism is a diagnostic possibility. Abnormalities on the plain chest radiograph may suggest a need for further diagnostic evaluation.

Among patients with pulmonary embolism who had no prior cardiopulmonary disease a normal chest radiograph was shown in 16 per cent (Table 11). Atelectasis or a pulmonary parenchymal abnormality were the most frequent abnormalities (68 per cent). When present, pleural effusions were usually small. The majority were limited to blunting of the costophrenic angle. In some studies, an elevated hemidiaphragm was the most frequent abnormality. The Westermark's sign (prominent central pulmonary artery and decreased pulmonary vascularity) was identified by radiologists in only 7 per cent of patients with pulmonary embolism.

In cases of pulmonary embolism, those with a normal plain chest radiograph had the lowest pulmonary artery mean pressures. The highest pulmonary artery mean pressures were in patients with a prominent central pulmonary artery or cardiomegaly.

Arterial blood gases and alveolar–arterial oxygen difference

A low partial pressure of oxygen in arterial blood (PaO_2) is typical of acute pulmonary embolism and supports the diagnosis, but patients with acute pulmonary embolism can have a normal PaO_2. Among patients with acute pulmonary embolism and no prior cardiopulmonary disease who had measurements of the PaO_2 while breathing room air, 24 per cent had a PaO_2 of 80 mmHg (10.5 kPa) or higher. Even among patients with submassive or massive acute pulmonary embolism, 12 per cent had a PaO_2 of 80 mmHg or higher. A normal alveolar–arterial oxygen difference (alveolar–arterial oxygen gradient) does not exclude acute pulmonary embolism. No value of the alveolar–arterial oxygen difference was diagnostic of pulmonary embolism, and no value excluded pulmonary embolism.

D-Dimer

As when considering the diagnosis of deep venous thrombosis, a 'negative' D-dimer test is useful for excluding pulmonary embolism in patients who are clinically thought to be at low risk, but a 'positive' result does not establish the diagnosis. Hence, when used in the appropriate clinical context, D-dimer testing is useful in defining a group of patients with suspected pulmonary embolism who do not require further investigation.

Ventilation–perfusion lung scans

The ventilation–perfusion lung scan in pulmonary embolism, if high probability, indicates pulmonary embolism in 87 per cent of patients (Table 12 and Fig. 1). If normal, pulmonary embolism is excluded. If intermediate probability, the scan contributes no useful diagnostic information, pulmonary embolism being present in about 30 per cent. If low probability, pulmonary embolism is present in 14 per cent. A low probability ventilation–perfusion scan, therefore, by the criteria used in the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) does not exclude pulmonary embolism, and intermediate and low probability interpretations may be grouped as 'non-diagnostic'. Criteria for a very low probability lung scan (positive predictive value less than 10 per cent) have been developed since the conclusion of PIOPED.

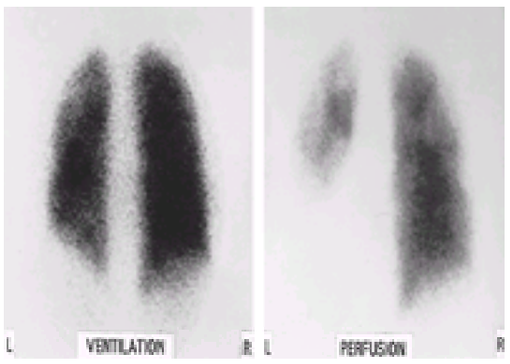

Fig. 1 Ventilation lung scan (left) and perfusion lung scan (right), posterior views. Left (L) and right (R) lungs are indicated. Ventilation scan, equilibrium phase, shows nearly normal ventilation. Perfusion scan shows absent perfusion in the left lower lobe and mismatched perfusion defects in the left upper lobe. Perfusion defects (grey areas) are also shown in the right lung. The ventilation–perfusion lung scan was interpreted as high probability for pulmonary embolism.

Prior clinical assessment in combination with interpretation of the ventilation–perfusion scan improves the diagnostic validity (Table 12). If the ventilation–perfusion scan is interpreted as high probability for pulmonary embolism, and if the clinical impression is concordantly high, then the positive predictive value for pulmonary embolism is 96 per cent. If the ventilation–perfusion scan is low probability and the clinical suspicion is concordantly low, then pulmonary embolism is excluded in 96 per cent of patients.

The probability of pulmonary embolism can be determined based on the number of mismatched defects. A further refinement of probability can be made if the ventilation–perfusion scan is interpreted after being stratified according to prior cardiopulmonary disease. Fewer mismatched perfusion defects are required to diagnose pulmonary embolism among patients with no prior cardiopulmonary disease. Adding clinical assessment to the stratification results in a more accurate evaluation.

Repeat ventilation–perfusion lung scanning

A residual abnormality of perfusion 1 year after acute pulmonary embolism is more frequent among patients with prior cardiopulmonary disease than among patients with none. It is useful to obtain a post-therapy baseline ventilation–perfusion lung scan for use in the event of suspected recurrent pulmonary embolism. This will assist in determining if abnormalities on a ventilation–perfusion scan are new or residual from prior pulmonary embolism.

Pulmonary angiography

Pulmonary angiography is associated with serious complications in about 1 per cent of patients. When needed, pulmonary angiography is useful and remains the diagnostic reference test for pulmonary embolism (Fig. 2, Fig. 3, and Fig. 4). Patients in whom the risk of complications of pulmonary embolism are greatest are patients referred for angiography from the medical intensive care unit. Frequently such patients are on respiratory support and in an unstable condition. The presence or absence of pulmonary embolism and the magnitude of pulmonary hypertension did not relate to the frequency of morbidity from angiography. Elderly patients (70 years or older) are at greater risk of renal impairment than younger patients as a result of the injection of contrast material. A retrospective analysis of complications, among patients in whom angiography was performed with non-ionic low-osmolar contrast material, showed fewer (0.1 per cent) major complications.

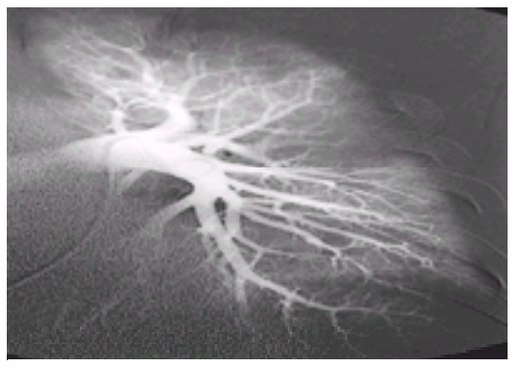

Fig. 2 Normal selective digital subtraction pulmonary angiogram of the left pulmonary artery. Vessels fill completely, taper gradually, and show numerous fine branches. This film and other pulmonary angiograms were supplied by Dr P.C. Shetty, Hurley Medical Center, Flint, Michigan, United States.

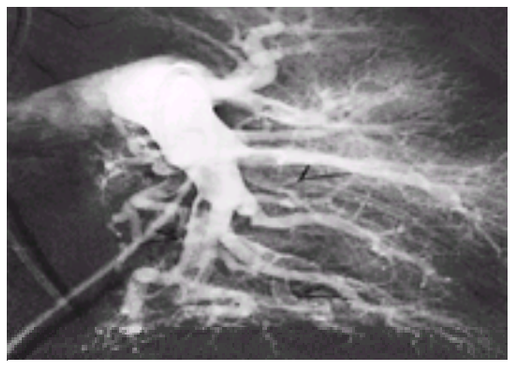

Fig. 3 Selective digital subtraction pulmonary angiogram of the left pulmonary artery showing multiple intraluminal filling defects indicative of pulmonary thromboemboli. Some of these have been identified by arrowheads.

Fig. 4 Selective pulmonary angiogram of the left pulmonary artery showing a large intraluminal filling defect with a saddle embolus at the bifurcation.

Contrast-enhanced spiral computed tomography

The results of imaging with contrast-enhanced spiral computed tomography (**CT**) (Fig. 5) compared with pulmonary angiography or autopsy are sparse. Results of individual small case series varied widely, with reported sensitivities ranging from 50 to 93 per cent. These studies all utilized 5-mm CT sections. With newer equipment, thinner sections with less breathing artefact are available. The sensitivity and specificity should, therefore, improve. However, the sensitivity and specificity of contrast enhanced spiral CT with multidetector scanners have not yet been determined. Investigation is in progress.

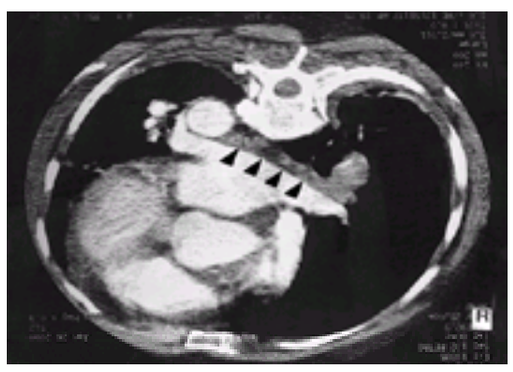

Fig. 5 Contrast-enhanced spiral computed tomogram showing a large intraluminal filling defect (arrowheads).

Magnetic resonance imaging

Magnetic resonance imaging, including magnetic resonance angiography and gadolinium-enhanced magnetic resonance angiography for acute pulmonary embolism are 'evolving' techniques. Preliminary observations suggest that gadolinium-enhanced three-dimensional magnetic resonance angiography during a single breath hold shows promise as a potentially useful imaging technique. Among the potential advantages are rapidity, avoidance of nephrotoxic iodinated contrast agents, and minimal invasiveness. Potential disadvantages include lack of sensitivity in detecting subsegmental pulmonary embolism and the fact that some patients may have a contraindication to magnetic resonance imaging.

Echocardiography

Echocardiography may show right ventricular dilatation and evidence of pulmonary hypertension, which, in the proper clinical setting, may strengthen the clinical impression that pulmonary embolism has occurred. Transoesophageal echocardiography sometimes may show proximal pulmonary emboli, but it has limited value in this regard.

Clues for the diagnosis of pulmonary embolism

With increasing severity of pulmonary embolism, from pulmonary infarction to isolated dyspnoea to circulatory collapse, trends suggest that the prevalence of a high probability ventilation–perfusion lung scan increases, as does the pulmonary artery mean pressure, while the PaO_2 decreases. However, making the diagnosis of pulmonary embolism is difficult, and depends on consideration of clinical, laboratory, and imaging data. Clues that can assist the physician in assessing the possibility of pulmonary embolism and avoiding inadvertent exclusion of the diagnosis are as follows.

1. Some patients with pulmonary embolism and circulatory collapse do not have dyspnoea, tachypnoea, or pleuritic pain.
2. Rales (crepitations) are common among patients with pulmonary infarction, but less so in those with isolated dyspnoea or circulatory collapse. They occur in those with radiographic evidence of a parenchymal abnormality.
3. A normal electrocardiogram is frequent in patients with the pulmonary infarction syndrome, but uncommon in those with isolated dyspnoea.
4. Abnormalities on the chest radiograph, although more common among patients with pulmonary infarction, are often observed in those with isolated dyspnoea.
5. Patients with circulatory collapse may have a normal chest radiograph.
6. A high probability interpretation of the ventilation–perfusion scan occurs in a minority of patients with the pulmonary infarction syndrome, but it is found in the majority of patients with the isolated dyspnoea syndrome.
7. A low probability ventilation–perfusion scan can occur in patients with pulmonary embolism and circulatory collapse.
8. A PaO_2 higher than 80 mmHg (10.5 kPa) is not uncommon in patients with the pulmonary infarction syndrome, but such levels are uncommon in those with the isolated dyspnoea syndrome.

Strategy for diagnosis

There are clear parallels with the situation when the diagnosis of deep venous thrombosis is considered. Subjecting all patients who might have a pulmonary embolus to complex, expensive and/or invasive tests is best avoided, such that management algorithms have been developed to identify those at very low risk, who can then be spared invasive tests. These algorithms typically use scoring systems to stratify the clinical probability that the particular patient has a pulmonary embolus, proceeding to D-dimer testing of those with low probability. Patients with a low clinical probability and a negative D-dimer test are not investigated further. Patients with either a high clinical probability or a low clinical probability but elevated D-dimer proceed to tests for the presence pulmonary emboli, typically by ventilation-perfusion lung scanning or contrast-enhanced spiral computed tomography. Examples of a pre-test scoring system and management algorithm are shown in Table 13.

Management algorithms of the type described in Table 13 are now employed in many hospitals, but a note of caution is appropriate. Their use has not been well validated. When there is genuine clinical doubt and a ventilation perfusion lung scan is reported as of low or intermediate probability, then there is a considerable diagnostic problem. In many centres contrast-enhanced spiral CT has become the diagnostic method of choice for suspected pulmonary embolism, or is used as the next line of investigation if ventilation perfusion scanning does not produce a clear-cut result. However, some would argue that the method has not been tested sufficiently and the physician should weigh the value of angiography (if available) against its hazards in the context of the whole clinical situation. The finding that 57 per cent of positive angiograms in the PIOPED study occurred in those with intermediate and low probability scans strengthens the case for an angiogram. An alternative strategy is to image the veins in the legs if ventilation perfusion scanning does not allow confident exclusion of the diagnosis of pulmonary embolism. If imaging reveals thrombus, then treatment is given. If imaging is negative, then this does not exclude the diagnosis of venous thromboembolism (the thrombus may have gone into the lungs), but the patient can be followed safely with serial non-invasive imaging of the leg veins. The risk of pulmonary embolism is low in patients in whom serial investigations of the legs show no deep venous thrombosis.

Treatment

General measures

All patients who are hypoxic should be given supplementary oxygen at high concentration (enough to restore normal PO_2), excepting those few with coincident chronic chest disease where carbon dioxide retention is problematic. In the early stages, continuous monitoring of arterial oxygen tension by pulse oximetry is advised.

Resuscitation

Most patients with acute pulmonary embolism do not have substantial circulatory compromise, but those presenting with massive pulmonary embolism may have circulatory collapse. They may look as though they are about to die, with cool peripheries, cyanosis, profound hypotension, and marked elevation of the jugular venous pulse. Features typical of long-standing pulmonary hypertension (palpable right ventricular heave, right ventricular gallop, loud P2, hepatomegally, ascites, peripheral oedema) are unlikely to be present. This dramatic haemodynamic picture may not be simply due to the direct anatomical effects of occlusion of main pulmonary vessels (the same picture is not seen after pneumonectomy, when one pulmonary artery is tied off completely), but also secondary to pulmonary neurogenic reflexes and local release of vasoactive substances, including 5-hydroxytryptamine and thromboxane from activated platelets.

Even though the jugular venous pulse is markedly elevated in acute massive pulmonary embolism, volume expanders should be administered rapidly to increase right ventricular filling pressure still further. The aim is to support the circulation until measures designed to deal with the embolus (usually thrombolysis—see below) can be applied and take effect.

Antithrombotic treatment

Unless there are serious concerns about the potential side-effects of anticoagulation or imaging is immediately available, it is common and sensible practice to begin anticoagulant treatment as soon as the diagnosis of pulmonary embolism is suspected. The antithrombotic regimen is the same as for deep venous thrombosis: see Table 5 and Chapter 15.15.3.2. This treatment is effective: the mortality from acute pulmonary embolism or recurrent pulmonary embolism during the first 3 months among 297 patients treated only with anticoagulants in the PIOPED study was 1.7 per cent. An additional 2.0 per cent suffered non-fatal recurrent pulmonary embolism during the first 3 months. Among all patients, irrespective of the treatment, the mortality from pulmonary embolism during the first 3 months was 2.4 per cent and an additional 3.5 per cent suffered non-fatal recurrent pulmonary embolism.

Thrombolytic therapy

Thrombolytic therapy is not indicated for the routine treatment of pulmonary embolism. Hypotension and continuing hypoxemia while receiving high fractions of inspired oxygen (FiO_2) are indications for intervention. Right ventricular dysfunction on the echocardiogram may also be an indication.

More rapid lysis of pulmonary thromboemboli occurs with thrombolytic agents than occurs spontaneously in patients treated only with anticoagulants. However, pulmonary reperfusion, as shown on perfusion lung scans, is similar after 2 weeks in patients treated with thrombolytic agents and those given anticoagulants.

A large prospective randomized trial in 1973 using urokinase showed no improvement of mortality and no difference of the rate of recurrence of pulmonary embolism among stable patients treated with thrombolytic therapy as opposed to anticoagulants. There have been no subsequent prospective randomized trials which contradict these results. A trend suggesting a lower rate of recurrent pulmonary embolism has been shown among patients with right ventricular dysfunction who were treated with tissue plasminogen activator.

Thrombolysis has risks. The frequency of major bleeding from tissue plasminogen activator among patients with pulmonary embolism diagnosed by angiography, based on pooled data, is 13 per cent. All investigations excluded patients at a high risk of bleeding, such as those with recent surgery, recent biopsy, peptic ulcer disease, blood dyscrasia, or severe hepatic or renal disease. The reported patients, therefore, had a low risk of bleeding. The risk of intracranial haemorrhage with tissue plasminogen activator (2 per cent) was higher among patients with pulmonary embolism than among patients who received tissue plasminogen activator for myocardial infarction.

Regimens of thrombolytic therapy

Regimens approved by the United States FDA for treatment of acute pulmonary embolism are:

1. streptokinase, 250 000 IU over 30 min followed by 100 000 IU/h for 24 h;
2. urokinase, 4400 IU/kg/h over 10 min followed by 4400 IU/kg/h for 12 to 24 h; or
3. tissue plasminogen activator, 100 mg(50 million IU)/2 h.

Potentially advantageous regimens of thrombolytic therapy, not fully evaluated, are:

1. urokinase, 15 000 IU/kg over 10 min; or
2. tissue plasminogen activator, 0.6 mg/kg (max. 50 mg) over 2 min.

It is recommended that heparin be discontinued during thrombolytic therapy and reinstituted upon discontinuation of thrombolytic therapy. None of the FDA approved regimens utilize concomitant heparin.

Inferior vena cava occlusion

An inferior vena cava filter is recommended in a patient with proximal deep venous thrombosis or pulmonary embolism if:

1. anticoagulants are contraindicated;
2. pulmonary embolism has recurred while on adequate anticoagulant therapy; or
3. pulmonary embolism is severe (right ventricular failure on physical examination, hypotension) and any recurrent pulmonary embolism may be fatal.

Insertion of an inferior vena cava filter is also strongly recommended in patients following pulmonary embolectomy.

Routine insertion of an inferior vena cava filter is not indicated only on the basis of a continuing predisposition for deep venous thrombosis. In special circumstances, however, this may be the best approach. Prophylactic insertion of vena cava filters may be considered for high-risk patients with deep venous thrombosis, severe pulmonary hypertension, and minimal cardiopulmonary reserve.

A number of vena cava filters have been designed for percutaneous insertion. They differ in outer diameter of the delivery system, maximal caval diameter into which they can be inserted, hook design, retrievability, biocompatability, and filtering efficiency. Filter migration, thrombosis, and cava wall perforation occur. Anticoagulant therapy after insertion of a filter is recommended. The filter alone, however, may be effective.

Symptomatic occlusion of the inferior vena cava is the most frequent complication, occurring in about 9 per cent of patients. Pulmonary embolism after insertion of an inferior vena cava filter is uncommon (1 per cent), and fatal pulmonary embolism is rare. Possible mechanisms that can explain pulmonary embolism after filter insertion are: (i) ineffective filtration, especially with tilting of the filter; (ii) growth of trapped thrombi through the filter; (iii) thrombosis on the proximal side of the filter; (iv) filter migration; (v) filter retraction from the caval wall; (vi) embolization through collaterals; (vii) embolization from sites other than the inferior vena cava; and (viii) incorrect position of the filter.

Complications of vena cava filters include filter deformation, filter fracture, insufficient opening of the filter, and improper anatomical placement of the filter. Filter-related complications include migration, angulation of the filter, caval stenosis, caval occlusion, erosion of the caval wall, and leg oedema. Complications at the site of insertion of the catheter do not differ from complications observed locally with other catheter techniques. Deep venous thrombosis at the puncture site generally has been reported in 8 to 25 per cent of cases.

Catheter interventions

Catheter-tip devices for the extraction or the fragmentation of embolus have the potential of producing immediate relief from massive pulmonary embolism. Such interventions may be particularly useful in patients in whom there is a contraindication to thrombolytic therapy. A suction-tip device for extraction of pulmonary embolism has been used in some patients. Thrombus fragmentation with a guide wire, angiographic catheter or balloon catheter, or specially designed devices have

been reported in small case series or case reports. The release of fragmented thromboemboli into the distal pulmonary arterial branches is not a problem. A registry of management strategies used by hospitals throughout Germany showed use of thrombus fragmentation in 1.3 to 6.8 per cent of patients with pulmonary embolism, depending on the severity. Catheters also have been developed that deliver high-velocity jets in the region of the thrombus, causing the thrombus to be sucked into the adjacent low pressure zone and undergo fragmentation due to the powerful mixing forces.

Pulmonary embolectomy

Medical therapy is likely to give better results than embolectomy. The operative and perioperative mortality related to surgical pulmonary embolectomy ranges between 28 and 74 per cent. A candidate for pulmonary embolectomy should meet the following criteria: (i) massive pulmonary embolism, angiographically documented if possible; (ii) haemodynamic instability (shock) despite heparin therapy and resuscitative efforts; and (iii) failure of thrombolytic therapy or a contraindication to its use.

Chronic pulmonary thromboembolic hypertension

In a very few patients with extensive embolization, the emboli fail to resolve and undergo fibrovascular organization, causing chronic obstruction to pulmonary arterial blood flow. Subsequently this can result in chronic pulmonary thromboembolic hypertension. This occurs in about 0.1 to 0.2 per cent of survivors of acute pulmonary embolism. In most of these patients, a procoagulant abnormality or defect in the fibrinolytic system cannot be shown. In many patients, both acute and chronic emboli are simultaneously present. A proliferation of fibrous connective tissue and small blood vessels penetrate a variable distance into the thrombus and attach it to the intimal surface of the vessel wall. Pulmonary hypertension develops as the result of a critical reduction of cross-sectional area for blood flow. The relatively high flow through non-occluded vessels may cause secondary hypertensive changes in the resistive or precapillary vessels. Patients with mean pulmonary artery pressures over 30 mmHg had a 5-year survival of 30 per cent and if the mean pulmonary artery pressure was over 50 mmHg, the 5-year survival was 10 per cent.

Most patients do not have an obvious history of venous thrombosis or pulmonary embolism. Dyspnoea is present in virtually all, but in the early stages may occur only on exertion. Compensatory right ventricular hypertrophy develops, and there may be a period of months to years when symptoms remain stable, but ultimately right ventricular function deteriorates. Symptoms then include worsening dyspnoea, fatigue, presyncope, syncope (rarely), pleuritic pain, angina-like pain, abdominal swelling, and peripheral oedema. Signs of respiratory and right ventricular failure develop, with cyanosis, grossly elevated jugular venous pulse, palpable right ventricular heave, right ventricular gallop, loud P2, hepatomegaly, ascites, and peripheral oedema. Bruits may be heard over the pulmonary arteries.

Ventilation–perfusion lung scans typically show one or more mismatched segmental or larger perfusion defects, and most patients have several bilateral mismatched perfusion defects. Pulmonary angiography is the most definitive diagnostic test. It shows narrowed segmental pulmonary arteries, sometimes accompanied by post-stenotic dilatation, irregularity of the intima, luminal narrowing of the central arteries, and oddly shaped vessels. Pulmonary fibreoptic angioscopy is useful to define surgical accessibility.

Pulmonary thromboendarterectomy is the treatment of choice. The procedure is highly specialized and postoperative management is difficult and complex. Such surgery is performed only at a limited number of centres. At the most experienced centre, the operative mortality is 8.7 per cent.

Medical therapy is adjunctive and should not delay the assessment for surgery. Medical therapy fails to address the underlying problem of fixed pulmonary vascular obstruction. In patients awaiting surgery, it is important to prevent further emboli and to treat cor pulmonale. Anticoagulation should be initiated immediately and maintained for life. Insertion of an inferior vena cava filter is generally advisable. There is no role for thrombolytic therapy. Angioplasty has not been successful. The potential role of pulmonary vasodilator therapy is speculative.

For fuller discussion of the management of pulmonary hypertension see Chapter 15.15.2.

Further reading

Bates SM, Grand'Maison A, Johnston M, Naguit I, Kovacs MJ, Ginsberg JS (2001). A latex D-dimer reliably excludes venous thromboembolism. *Archives of Internal Medicine* **161**, 447–53.

Collaborative Study by the PIOPED Investigators (1990). Value of the ventilation/perfusion scan in acute pulmonary embolism—results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *Journal of the American Medical Association* **263**, 2753–9. [Landmark investigation of ventilation/perfusion lung scans.]

Geerts *et al*. (2001). Prevention of venous thromboembolism. *Chest* **119**(Suppl), 132S–75S. [Detailed and authoritative review of methods for prevention of venous thromboembolism.]

Hull RD *et al*.(1990).Noninvasive strategy for the treatment of patients with clinically suspected pulmonary embolism. *Archives of Internal Medicine* **154**, 289–97.

Hull RD, Raskob GE, Pineo GF, eds (1996). *Venous thromboembolism: an evidence-based atlas*. Futura Publishing, New York. [Comprehensive review of pulmonary embolism.]

Hyers TN *et al*. (2001). Antithrombotic therapy for venous thromboembolic disease. *Chest* **119**(Suppl), 176S–93S. [Authoritative and in-depth review of treatment with recommendations .]

Kearon C *et al.* (2001). Management of suspected deep venous thrombosis in outpatients by using clinical assessment and D-dimer testing. *Annals of Internal Medicine* **135**, 108–11.

National Cooperative Study (1973). The Urokinase Pulmonary Embolism Trial. *Circulation* **47**(Suppl II), II-1–II-108. [Basic investigation of thrombolytic therapy.]

Stein PD (1996). *Pulmonary embolism.* Williams & Wilkins, Media, Pennsylvania. [Detailed comprehensive review.]

Stein PD *et al*. (1991). Clinical, laboratory, roentgenographic and electrocardiographic findings in patients with acute pulmonary embolism and no pre-existing cardiac or pulmonary disease. *Chest* **100**, 598–603. [Shows detailed results of clinical findings in patients with pulmonary embolism.]

Stein PD, Hull RD, Pineo G (1995). Strategy that includes serial noninvasive leg tests for diagnosis of thromboembolic disease in patients with suspected acute pulmonary embolism based on data from PIOPED. *Archives of Internal Medicine* **155**, 2101–4. [Gives background for validity of strategy that includes serial non-invasive leg tests.]

Wells PS *et al*. (1997). Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet* **350**, 1795–8.

Wells PS *et al.* (2000). Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. *Thrombosis and Haemostasis* **83**, 416–20.

Wells PS *et al*. (2001). Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department using a simple clinical model and D-dimer. *Annals of Internal Medicine* **135**, 98–107.

Viner SM *et al.* (1994). The management of pulmonary hypertension secondary to chronic thromboembolic disease. *Progress in Cardiovascular Diseases* **37**, 79–92. [In-depth review of chronic pulmonary thromboembolic hypertension.]

15.15.3.2 Therapeutic anticoagulation in deep vein thrombosis and pulmonary embolism

David Keeling

Introduction

Deep vein thrombosis and pulmonary embolism are aspects of the same disease, venous thromboembolism. Forty per cent of patients with deep vein thrombosis without clinical evidence of pulmonary embolism have evidence of emboli on lung scanning. The principles of therapeutic anticoagulation are the same for both. In proximal deep vein thrombosis and pulmonary embolism this involves immediate anticoagulation with heparin followed by a period of anticoagulation with warfarin (or other oral vitamin K antagonist). Distal deep vein thrombosis can be managed in the same way, but an alternative strategy is to use serial non-invasive testing (e.g. ultrasound), which only reliably detects proximal thrombosis, to ensure that suspected distal thrombosis does not extend above the knee, withholding treatment if it does not.

There is clear evidence that heparin is needed in the initial phase and that anticoagulation with oral anticoagulants alone is inadequate. Warfarin can be commenced on the first day and heparin is continued for 5 days or until the international normalized ratio (**INR**) is greater than 2.0 for 2 consecutive days, whichever is the longer. Extending the period of heparinization from 5 to 10 days is not more effective and increases the risk of heparin-induced thrombocytopenia. However, for massive pulmonary embolism or severe ileofemoral thrombosis a longer period of heparin therapy may be considered.

Heparin

Heparin, a glycosaminoglycan, is composed of alternating uronic acid and glucosamine saccharides that are sulphated to a varying degree. Its mode of action is to potentiate the activity of the serine protease inhibitor (serpin) antithrombin, whose main mode of action is to inhibit thrombin, but which also inhibits several other coagulant proteases such as factor Xa. A specific pentasaccharide sequence determined by the sulphation pattern along the heparin chain binds to antithrombin and causes a conformational change, giving it full activation against factor Xa but only partial activation against thrombin. Heparins of 18 saccharides (MW 5400) or more can extend across the intermolecular gap and also bind to thrombin giving full antithrombin activity, which is lost if the chains are shorter. Unfractionated or standard heparins are a mixture of chains of different lengths (MW 5000 to 35 000, mean 13 000) and low-molecular-weight heparins (MW 2000 to 8000, mean 5000) are derived from them by enzymatic or physicochemical cleavage. Low-molecular-weight heparins have, with good reason, largely replaced unfractionated heparin for the treatment of venous thromboembolism, but the use of the latter will be discussed first.

Anticoagulation with unfractionated heparin

Unfractionated heparin has most often been given by continuous intravenous infusion, the rate of which has to be adjusted, usually by measuring the activated partial thromboplastin time (**APTT**). An inadequate APTT response in the first 24 h may increase the risk of recurrence of thromboembolism, though this does not seem to be critical if the starting infusion rate is at least 1250 IU/h. A validated regimen is to give a bolus dose of 80 IU/kg and to start the infusion at 18 IU/kg.h, performing the first APTT estimate after 6 h. The dose is then usually adjusted to maintain the APTT between 1.5 to 2.5 times the average laboratory control value. With older APTT reagents this corresponded to a therapeutic heparin level of 0.2 to 0.4 IU/ml by protamine titration or 0.3 to 0.6 IU/ml by anti-Xa assay. However, many current APTT reagents show an increased sensitivity to unfractionated heparin and with these higher ratios should be aimed for. The local laboratory should advise on the appropriate therapeutic range with its reagent. When the dose is therapeutic the APTT should be checked daily.

An alternative is to give unfractionated heparin subcutaneously once every 12 h, and a meta-analysis suggested that this might be more effective and at least as safe as continuous intravenous infusion. A reasonable starting dose is 250 IU/kg, adjusting the dose according to the mid-interval APTT.

Anticoagulation with low-molecular-weight heparin

Although much is made of the greater anti-Xa to antithrombin ratio of the low-molecular-weight heparins, their key clinical property is that they produce a much more predictable anticoagulant response than unfractionated heparin. This, combined with the fact that they have very high bioavailability after subcutaneous injection, means that the dose can be calculated by body weight and be given subcutaneously without any monitoring or dose adjustment. The actual dosage used differs slightly with the different low-molecular-weight heparins and the manufacturers' recommendations should be followed, but a typical dose is 200 IU/kg once a day. They are at least as effective and at least as safe as unfractionated heparin, even when given once a day. Their widespread use has enabled many patients with deep vein thrombosis to be managed as outpatients. Low-molecular-weight heparin is renally excreted so should be used with caution in renal failure (anti-Xa levels can be checked if necessary).

Complications of heparin treatment

If a patient on intravenous unfractionated heparin is excessively anticoagulated, it is usually sufficient simply to stop the infusion, the half-life being 1 to 2 h. If bleeding is severe the heparin can be neutralized with protamine sulphate, giving 1 mg for every 100 IU that have been infused over the previous hour. The reversal of low-molecular-weight heparin is more problematic. Although protamine sulphate may not neutralize the smaller chains it is often clinically effective, though estimating an appropriate dose is more difficult (the maximum dose is 50 mg, so this is often given if the subcutaneous injection was recent).

Heparin-induced thrombocytopenia is a feared complication, but much less common now short courses of low-molecular-weight heparin are used. It is due to the development of an antibody to the heparin–platelet factor 4 complex and can be associated with serious venous and arterial thrombosis. Patients on heparin for 5 or more days should have their platelet count checked. If heparin-induced thrombocytopenia is suspected, then heparin must be stopped and an alternative substituted (danaparoid or hirudin being most suitable).

Long-term treatment, for example in pregnancy, is associated with osteopenia, but this may be less of a problem with low-molecular-weight heparin.

Warfarin

The oral vitamin K antagonists are the mainstay of long-term anticoagulant therapy. Warfarin is the commonest vitamin K antagonist given, though the shorter acting nicoumalone and the longer acting phenprocoumon are also used. Phenindione is used less commonly because of a high incidence of skin rashes. The procoagulant factors II, VII, IX, and X (and the anticoagulants protein C and protein S) need vitamin K for the g-carboxylation of the glutamic acid residues that form their gla domains. Without this post-translational modification they cannot bind calcium, and as a consequence cannot bind to anionic phospholipid surfaces, such that assembly of the key coagulation complexes is disrupted.

Warfarin takes about 4 days to become effective, during which period heparin is given. When warfarin is started the vitamin K-dependent factors fall according to their half-lives. Factor VII and protein C have the shortest half-lives, so that despite a prolongation of the INR due to factor VII deficiency, warfarin may initially be

procoagulant. This is the mechanism for the rare problem of warfarin-induced skin necrosis, most often described in those with protein C deficiency.

Anticoagulation with warfarin

Initiation and monitoring

Monitoring of warfarin treatment is by the international normalized ratio (INR). This is a manipulation of the prothrombin time (**PT**) to allow for the different sensitivities of various laboratory reagents to the warfarin-induced coagulopathy. The INR equals (PT/MNPT) ^{ISI} where MNPT is the (mean normal) control PT and ISI is the international sensitivity index of the thromboplastin used in the assay. For the treatment of deep vein thrombosis and pulmonary embolism the target INR should be 2.5 (target range 2.0 to 3.0). If a recurrence occurs despite an INR of 2.0 to 3.0, then the dose is usually increased to a target INR of 3.5 (target range 3.0 to 4.0).

If the initial coagulation tests are not prolonged it is usual to give 10 mg of warfarin on the first evening and check the INR the following morning. Warfarin dose is adjusted according to the daily INR results until the patient is stable. Stable anticoagulation is more quickly and safely achieved if a dosing algorithm is followed (Table 1).

When patients are stable they may go for up to 8 weeks between INR checks. If the INR is unstable, patients are seen more frequently, but it should be noted that with warfarin it takes approximately 1 week (five times the half-life of 36 h) to reach a new steady state after dose adjustment and more frequent dosage alteration is inadvisable.

How long should the patient take warfarin?

It is a difficult clinical decision to decide how long to continue warfarin, a matter of balancing the risks of recurrence against the risks of warfarin. The latter are well known, 1 to 2 per cent of people on warfarin have a major bleed each year and 0.5 per cent suffer an intracranial bleed, of which 50 per cent die, giving a fatality rate of 0.25 per cent per annum. However, warfarin is highly (90 to 95 per cent) effective at preventing recurrence. The risk of a recurrent venous thromboembolism after a first deep vein thrombosis is approximately 5 per cent per year, when a case–fatality rate of 5 per cent (and some estimates for recurrent venous thromboembolism have been this high, though the overall rate for all venous thromboembolism is probably 1 to 2 per cent) would also give a fatality rate of 0.25 per cent per year. Other factors can be taken into account: the risk of recurrence is higher for the first 6 months, it is higher for proximal deep vein thrombosis and pulmonary embolism than for distal deep vein thrombosis, and it is lower if a transient risk factor was present (e.g. recent surgery, use of the contraceptive pill). Six months of anticoagulation has been shown to be more effective than 6 weeks of anticoagulation in all subgroups. Whether an inherited thrombophilia should influence the long-term management is not clear. Although they predict first events, the commoner defects (factor V Leiden and prothrombin G20210A) may not predict recurrence. Taking all this into account a reasonable approach is indicated in Table 2.

Complications of warfarin treatment

The only major complication of warfarin treatment is bleeding. Risk factors for bleeding are an age of 65 years or more, a history of stroke, a history of gastrointestinal bleeding, anaemia, renal impairment, diabetes, and recent myocardial infarction. A major problem in control is the starting and stopping of other medication. Many drugs interact with warfarin (see Table 3 for those with the most evidence) and patient education and constant vigilance is essential. Close monitoring of the INR is advised when concomitant medication is altered.

The approach taken to reverse over-anticoagulation with warfarin depends on the circumstances (see Table 4). Prothrombin complex concentrates, unlike fresh frozen plasma, reliably and rapidly correct the defect and should be used in life-threatening situations such as intracranial bleeding. Small doses of phytomenadione (vitamin K_1) can lower a high INR without making subsequent anticoagulation difficult, as is the case if high doses are given.

Fibrinolysis

Thrombolytic agents dissolve thrombi by directly or indirectly activating the zymogen plasminogen to plasmin. Plasmin then degrades fibrin to soluble peptides, but cannot distinguish fibrin in pathological thrombi from fibrin in haemostatic plugs and it may also degrade and so deplete plasma fibrinogen. The use of thrombolytic agents for venous thromboembolism requires careful individual assessment. It is rarely given in deep vein thrombosis though its use can be considered in massive ileofemoral thrombosis. Although thrombolytic therapy for pulmonary embolism achieves more rapid resolution than heparin alone, there is no clear evidence of lasting benefit. Patients with pulmonary embolism who survive long enough to have the diagnosis made and treatment with heparin begun have an excellent prognosis, unless they have associated severe medical disease. Thrombolytic therapy, which carries a much greater risk of bleeding, is therefore reserved for those cases of massive pulmonary embolism with haemodynamic instability threatening the patient's life (see Chapter 15.15.3.1).

Streptokinase (which forms a complex with plasminogen that then activates free plasminogen), urokinase, and tissue plasminogen activator (**tPA**) have all been used. For pulmonary embolism, streptokinase is recommended as a 250 000 IU loading dose followed by an infusion for 24 h at 100 000 IU/h. Urokinase is given as a 4400 IU/kg loading dose followed by 2200 IU/kg.h for 12 h. Following the success of rapid fibrinolytic regimens in myocardial infarction, tPA given as 100 mg over 2 h has been used for pulmonary embolism, and the use of more rapid regimens with the other two agents has been suggested (see Chapter 15.15.3.1 for further discussion).

Treatment of venous thromboembolism in pregnancy

Heparin does not cross the placenta and can be used in pregnancy, as described above, but higher doses of unfractionated heparin are sometimes needed to achieve therapeutic levels. As pregnant women are excluded from clinical trials, experience with low-molecular-weight heparin is limited. However, the evidence seems to indicate that low-molecular-weight heparin, with all its logistical advantages, can be used effectively and safely in pregnancy.

The real problem is warfarin, which crosses the placenta and can cause an embryopathy if given between 6 and 12 weeks' gestation. At any time it can cause fetal bleeding and has been associated with central nervous system abnormalities. The usual treatment recommended for venous thromboembolism in pregnancy is to continue with full-dose heparin until term (long-term treatment with unfractionated heparin can be given by subcutaneous injection once every 12 h, and with low-molecular-weight heparin by injection once daily). As heparin may cause osteopenia (possibly less of a risk with low-molecular-weight heparin) some would consider the use of warfarin after the first trimester, switching back to heparin at 36 weeks. Warfarin can be used for the 6 weeks of the puerperium, and women taking warfarin can breast feed.

Further reading

Anonymous (1998). Guidelines on oral anticoagulation: third edition. *British Journal of Haematology* **101**, 374–87.

Anand S *et al.* (1996). The relation between the activated partial thromboplastin time response and recurrence in patients with venous thrombosis treated with continuous intravenous heparin. *Archives of Internal Medicine* **156**, 1677–81.

Leizorovicz A *et al.* (1994). Comparison of efficacy and safety of low molecular weight heparins and unfractionated heparin in initial treatment of deep venous thrombosis: a meta-analysis. *British Medical Journal* **309**, 299–304.

15.16.1.1 Prevalence, epidemiology, and pathophysiology of hypertension

C. G. Isles

Definitions of hypertension

The simplest and most widely accepted definition of hypertension in an adult is that hypertension is present when clinic systolic pressure exceeds 140mmHg and/or clinic diastolic pressure exceeds 90mmHg. The American and World Health Organization International Society of Hypertension definition of hypertension, reproduced in Table 1, gives a more detailed classification of blood pressure for adults aged over 18 years. Any definition of abnormality that is based on a measurement distributed within the population as a continuous variable, such as blood pressure, serum cholesterol, or height, must necessarily be somewhat arbitrary, and so it is with hypertension.

Prevalence

United Kingdom studies

Estimates of prevalence vary. The number of measurements made, the method used, and the circumstances in which measurements are made all influence prevalence. Also important here are age, gender, race, and socio-economic status. In a representative population sample of 10359 Scottish men and women aged 40 to 59 years between 1984 and 1986, and based on the average of two measurements at a single screening visit in a primary care setting, 1262 (25 per cent) men and 1061 (20 per cent) women were considered to be hypertensive, defined in this study as a blood pressure of 160/95mmHg or higher or receiving antihypertensive drug treatment. A more recent survey of 12116 English adults aged 16 years or more in 1994 yielded similar results. Based on the average of the second and third of three readings at a single screening visit in the respondent's home, 19 per cent of men and 20 per cent of women had a blood pressure of 160/95mmHg or higher or were receiving antihypertensive drug treatment. Both surveys probably overestimate the prevalence of hypertension in the United Kingdom because they were based on the average of readings at a single screening session.

United States studies

The value of repeated measurements at different visits in determining prevalence of hypertension is apparent in a study of 158906 mixed-race individuals aged 30 to 69 years examined in their homes or work places in the United States in the early 1970s. Twenty-five per cent had a diastolic blood pressure of 90mmHg or higher at the first screen. When these subjects were rescreened in a clinic setting, there was a substantial fall in blood pressure such that only 25 per cent (6.4 per cent of the original cohort) remained hypertensive. The majority (75 per cent) were mildly hypertensive with a diastolic blood pressure in the range 90 to 104mmHg. Put another way, in this large population survey most individuals with hypertension identified had a mild to moderate elevation of blood pressure, with less than 2 per cent of those screened having sustained hypertension with a diastolic blood pressure higher than 105mmHg after two measurements.

The importance of age as a determinant of prevalence of hypertension is evident from data collected during the 1988 to 1991 National Health and Nutrition Examination Survey. The proportion of the United States population having a systolic blood pressure of 140mmHg or higher, or a diastolic blood pressure of 90mmHg or higher, based on the average of six measurements at two visits, or currently being treated with an antihypertensive drug, was 4 per cent for young adults aged 18 to 29 years, increasing to 65 per cent for those over 80 years (Fig. 1).

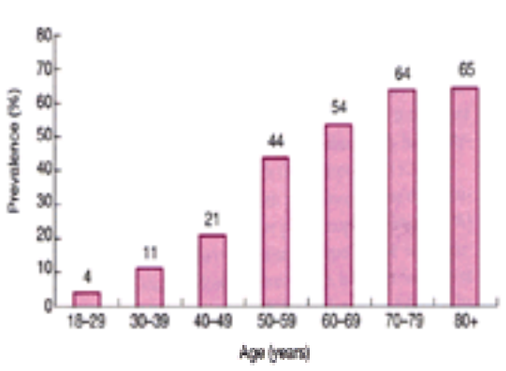

Fig. 1 Prevalence of hypertension, defined as a systolic blood pressure of 140mmHg or higher, or a diastolic blood pressure of 90mmHg or higher in the United States. Prevalence increases with advancing age. Adapted with permission from NHANES III.

Recent data from Framingham, United States suggest that the increasing use of antihypertensive medication has resulted in a decline in the prevalence of hypertension. In an analysis of 10333 subjects aged 45 to 74 years, examined over a 40-year period from 1950 to 1989, using definitions of 160mmHg systolic or 100mmHg diastolic on or off treatment, and based on an average of two separate measurements at each visit, the prevalence of hypertension decreased from 18.5 to 9.2 per cent for men, and from 28.0 to 7.7 per cent for women (Fig. 2).

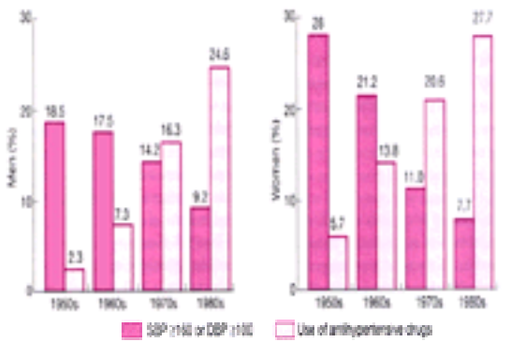

Fig. 2 Age-adjusted temporal trends in prevalence of raised blood pressure and use of antihypertensive drugs among men and women 45 to 74 years of age in Framingham, United States. Adapted with permission.

Measuring blood pressure

If blood pressure is to be measured precisely, a device with validated accuracy that is properly maintained and calibrated must be used. For a long time the mercury sphygmomanometer has been the gold standard in clinical practice but—for health and safety reasons related to mercury—is likely to be replaced by aneroid, semi-automatic and automated devices in the near future. Information on the accuracy of the alternatives to mercury sphygmomanometers can be obtained from the British Hypertension Society Information Services (e-mail: bhsis@sghms.ac.uk; website: www.hyp.ac.uk/bhs/).

Cuff size

Whatever method is employed, a cuff with an appropriately sized bladder should be applied to the upper arm, leaving adequate space for the bell of a stethoscope to be positioned in the antecubital fossa. The arm should be supported at the level of the heart and unrestricted by tight clothing. For most patients an appropriately sized bladder is 14 by 35cm, sometimes known as the alternative adult cuff. Smaller bladders will over-read blood pressure in fat arms, while larger bladders as used in thigh cuffs are too unwieldy for routine use. It is good practice to measure the blood pressure in both arms at the first visit and then to select the arm with the higher reading for future measurements; also to record pressure both seated and standing initially to exclude a significant postural fall, more likely in elderly and diabetic subjects.

Clinic/doctor's office (surgery)

If a mercury sphygmomanometer is being used, the correct procedure for measurement is to deflate the cuff at 2mm/s and read blood pressure to the nearest 2mmHg, recording diastolic pressure at the disappearance of the sounds (phase V). Patients in whom the sounds never disappear probably have arterial disease of their upper limb vessels causing turbulent flow and should have their pressure recorded as systolic/diastolic phase IV (muffling)/0mmHg. At least two measurements should be made at each visit. Because blood pressure tends to fall with repeated measurements, patients whose pressure is found to be greater than 140/90mmHg should be brought back for further measurements at intervals that may vary from 1 day to 1 month, according to the level of their pressure, before a diagnosis of sustained hypertension is made.

Home/ambulatory

Interest in home and ambulatory blood pressure measurement continues to grow. Both techniques permit the recording of numerous values in settings that are closer to daily life than the clinic or office. By averaging up to 60 readings taken at intervals of 20 to 30min over the course of 24h, ambulatory blood pressure measurement improves precision and reproducibility in blood pressure measurement. It also correlates more closely than clinic blood pressure with risk, as judged by evidence of target organ damage, but lacks the authority of an outcome trial, which means that it is not possible to recommend ambulatory blood pressure measurement over clinic pressure in routine practice.

White coat hypertension

Despite the reservations expressed above, ambulatory blood pressure measurement is widely used and can be valuable in a number of circumstances. Foremost among these is the evaluation of the patient with suspected white coat hypertension, also known as isolated clinic or office hypertension. White coat hypertension is present in up to 10 per cent of the hypertensive population and is usually defined as persistent clinic or office hypertension greater than 140/90mmHg in the face of consistently normal readings less than 135/85mmHg at home. Because white coat hypertension carries much less risk of cardiovascular disease, the decision to recommend or withhold drug treatment must be based on the overall risk profile. Subjects whose risk is low require continued monitoring only, whereas patients who already have vascular disease or are at high risk of developing vascular disease, for instance because they have diabetes or target organ damage, should be given antihypertensive drug therapy.

Other indications for ambulatory blood pressure measurement

Ambulatory blood pressure measurement may also be indicated in a number of other circumstances: when clinic blood pressure is unusually variable; when patients complain of postural symptoms related to their drug therapy; and in resistant hypertension, defined here as blood pressure greater than 150/90mmHg despite lifestyle measures and three or more antihypertensive drugs. Whatever the indication for ambulatory blood pressure measurement, the average daytime pressure is recommended for decisions on treatment, rather than the average 24-h pressure or the percentage of readings that lie above a certain threshold. It must also be recognized that ambulatory blood pressure is systematically lower than clinic or office blood pressure by an average of at least 10/5mmHg. This means that treatment thresholds and targets for achieved blood pressure should be adjusted downwards when using ambulatory blood pressure measurement. Pending the outcome of further prospective observational studies, average daytime ambulatory pressure of less than 135/85mmHg should be regarded as probably normal, and average daytime ambulatory pressure of greater than 140/90mmHg as probably abnormal and requiring drug treatment.

Epidemiology

Renfrew Paisley survey

Hypertension is an important contributor to morbidity and mortality from cardiovascular disease, particularly when present in combination with other cardiovascular risk factors. This is apparent from any one of a number of observational studies. The data shown in Fig. 3 are from the Renfrew and Paisley Survey in Scotland and show quite clearly that in this Western population, the relative risk of diastolic pressure for both coronary heart disease and stroke is approximately two to three in both sexes, when comparing high-risk with low-risk subjects; also that the rates of coronary heart disease are higher than those of stroke at all levels of blood pressure; and that coronary deaths are more common in men than in women, while gender differences in stroke mortality are much less apparent.

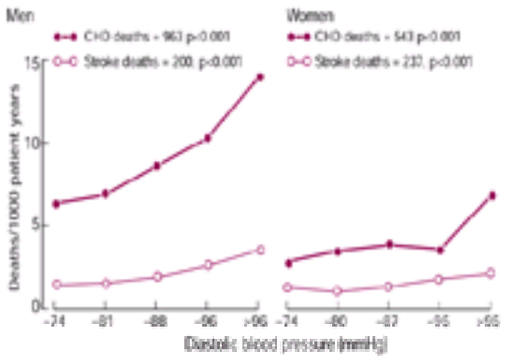

Fig. 3 Deaths per 1000 patient years among men and women aged 45 to 64 years after 17 years of follow-up in Renfrew and Paisley, Scotland. *P* value indicates significance of trend across quintiles of diastolic blood pressure. Reproduced with permission.

Framingham

Some authorities have suggested that the combined effects of risk factors are multiplicative rather than additive. Thus in Framingham, the probability of a 40-year-old male with a systolic pressure of 195mmHg developing coronary heart disease over 6 years is said to be 10-fold greater if he smokes and has plasma cholesterol of 9mmol/l with glucose intolerance and electrocardiographic left ventricular hypertrophy. The problem with data such as these is that very few subjects in Framingham (or elsewhere) possess so many risk factors. This means that the estimates of risk must necessarily be based on mathematical models rather than on actual numbers of events: indeed the authors of the Framingham study concede that the models used assume a degree of non-linearity (Kannel WB, personal communication).

Oxford meta-analysis

The most powerful and persuasive study of the association between blood pressure and risk remains the Oxford meta-analysis of 843 strokes and 4856 coronary heart disease events in 420000 individuals followed in nine major prospective observational studies for an average of 10 years (Fig. 4). The results of this meta-analysis were corrected for a phenomenon known as regression dilution bias by using repeated measurements to determine an individual's long-term average blood pressure, and showed that the relation between diastolic blood pressure, stroke, and coronary heart disease was at least 60 per cent stronger than had previously been thought. Within the range of diastolic blood pressure 70 to 110mmHg, the risk of stroke and coronary heart disease was positive, continuous, and independent of other risk factors, even when diastolic pressure was supposedly normal, namely 70–90mmHg. This raises the intriguing possibility, recently confirmed by the results of PROGRESS, that lowering blood pressure in 'normotensive' high-risk subjects might reduce their risk of vascular disease.

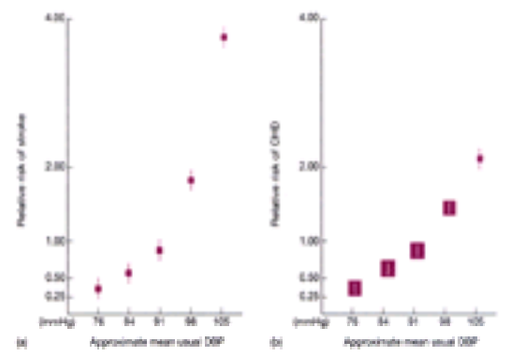

Fig. 4 Relation between diastolic blood pressure, stroke, and coronary heart disease during an average of 10 years follow-up. (a) Stroke and usual diastolic blood pressure. Seven prospective observational studies: 843 events. (b) Coronary heart disease and usual diastolic blood pressure. Nine prospective observational studies: 4856 events. (Reproduced with permission.)

Systolic hypertension and pulse pressure

Systolic blood pressure rises in a linear fashion with age, whereas diastolic pressure increases until the age of 50 then levels off and even begins to fall. This leads to an important transition in the form of hypertension with age. Isolated diastolic hypertension is more common in younger subjects, whilst isolated systolic hypertension emerges as the most common form of hypertension in the elderly. The underlying pathological process is loss of arterial elastic tissue, which means that the pressure wave created by left ventricular contraction can no longer be damped by the aorta and major vessels. Although clinicians have tended to focus on the diastolic component of blood pressure in the past, systolic blood pressure is a better predictor of cardiovascular risk and isolated systolic hypertension is now recognized to be an independent risk factor of cardiovascular disease. A wide pulse pressure has a similar influence on prognosis (Fig. 5).

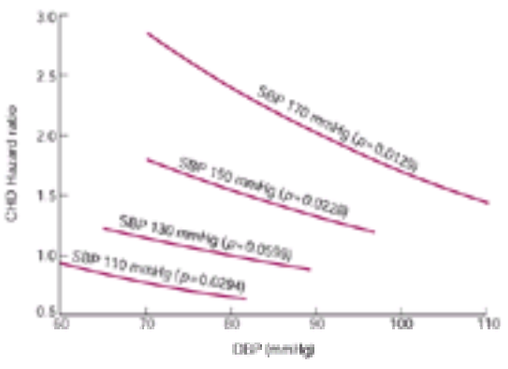

Fig. 5 Adjusted CHD risk in 1924 Framingham men and women between 50 and 79 years with no clinical evidence of CHD at baseline. At any given level of systolic pressure, risk was higher in those whose diastolic pressure was lower, highlighting the predictive power of pulse pressure..

Likely benefits of treatment

The relative risk of high blood pressure for heart attack and stroke is higher among younger subjects and decreases with age. Absolute risk, by contrast, increases with age. This means that attributable risk, which is the number of events that would be avoided if blood pressure were lower, is likely to be higher in the elderly. A similar logic can be applied to the likely impact of interventions that reduce mean blood pressure by even a few millimetres of mercury. The steepness of the slope relating diastolic blood pressure and stroke in the Oxford meta-analysis suggests that more strokes might be avoided than coronary heart disease events. However, the attributable risk of diastolic blood pressure for coronary heart disease, that is the number of coronary heart disease events that would be avoided if diastolic blood pressure were lower, is likely to be as great simply because coronary heart disease is more common than stroke in most Western populations.

Heart failure, renal disease, and recurrent events

Blood pressure is a powerful determinant not only of heart attack and stroke but also of heart failure and renal failure. A non-linear or J-shaped association between blood pressure and recurrent events has been reported for patients with previous myocardial infarction. Despite concerns that this might represent an adverse consequence of treatment, it is is now considered more likely to reflect the effect of disease on blood pressure—the bigger the myocardial infarction the bigger the fall in pressure—than the effect of blood pressure or its treatment on the disease.

Hypertension in black populations

Ethnic variations in the incidence, pathophysiology, and complications of hypertension are well described. Hypertension is more common in black than white populations, and more common in urban than rural black populations. Black individuals have a higher incidence of salt-sensitive hypertension than white individuals, and retain more sodium leading to expanded plasma volumes and lower plasma renin activity. The complications of hypertension also tend to be different in black populations with a higher incidence of left ventricular hypertrophy, stroke, and renal failure, and lower risk of coronary heart disease. The increased frequency of left ventricular hypertrophy, stroke, and renal failure may relate to the severity of hypertension in black individuals, and their relative lack of coronary heart disease may be due to more favourable lipid profiles.

Gender differences

Hypertension is an important risk factor for cardiovascular disease in women. Although premenopausal women have lower blood pressure than age-matched men, the prevalence of hypertension is higher in women than men after the age of 65. Obesity is significantly more common in middle-aged and older women, and is likely to contribute to the crossover in prevalence. Oral contraceptive use increases the risk of hypertension in younger women. Hormone replacement therapy (HRT) does not raise blood pressure in women who are normotensive at the start of treatment, but more research is required to determine whether blood pressure rises when hypertensive postmenopausal women are given HRT.

Pathophysiology

When asked to define hypertension, the clinician may reply 'that level of blood pressure above which treatment does more good than harm'. The response of the pathophysiologist is likely to be more complex, reflecting the fact that we don't fully understand the mechanisms involved. Most individuals with hypertension have essential hypertension, which is best thought of as a progressive rise in pressure with age, as a result of an interplay between genetic factors, environmental influences, and blood pressure control mechanisms. A much smaller number of patients (certainly less than 5 per cent of the hypertensive population) will have a renal or adrenal cause for their high blood pressure (see Chapter 20.10.2).

Genetic factors

The fact that high blood pressure runs in families does not tell us whether the genes or environment are responsible, because families usually share both. Helpful in sorting this out are studies of adopted children within a family, and of twins living apart. The twin studies suggest that approximately 50 per cent of the blood pressure variability between individuals is related to inheritable factors. The genetic component of the development of high blood pressure may not itself necessarily cause hypertension. Rather there may be a genetic predisposition to develop raised blood pressure in response to various environmental factors. For further information on the genetics of hypertension see Chapter 15.16.1.2.

Environmental influences

Environmental influences are more easily studied than genetic factors. Evidence of their importance comes from a number of sources, including studies of migrants, comparison between different communities, prospective population studies, and randomized trials of behaviour modification.

Migrant studies

Adults who migrate adopt the level of blood pressure, frequency of hypertension, and coronary heart disease risk of their destinations. Primitive Kenyan Luo tribespeople moving in search of work from their villages in rural Kenya to the urban slums of Nairobi show marked increases in blood pressure and body weight within a month of migration. Similar findings have been reported for Australian Aborigines migrating from the bush to Melbourne.

Comparison between different communities

In developed countries such as the United Kingdom, the rise in blood pressure that occurs with age is well recognized. The mechanisms responsible for this are not physiologically inevitable because primitive rural populations, most notably the Yanamamo Indians of Brazil and certain New Guinea tribes, show only a tiny rise in blood pressure with advancing age. These observations suggest that the rise in blood pressure seen with advancing age in urban societies must be due to some very powerful environmental factors.

Prospective population studies

A classic study compared Italian women entering a nunnery with a control group of women from the same town. In the control group, blood pressure rose normally with age whereas the nuns showed no rise in blood pressure over 20 years of follow-up. In another example, blood pressure did not rise as expected with age in long-term residents of a psychiatric hospital who entered with normal blood pressure. Studies such as these also provide evidence of which environmental influences might be important.

Specific environmental factors

The best known environmental influences on blood pressure are obesity, alcohol, and salt. Early nutritional deficiency may be important, and recent evidence suggests that psychosocial factors are likely to play a role in the development of essential hypertension. A small socio-economic gradient of blood pressure has been observed: inverse for developed and positive for developing countries. This probably reflects the higher prevalence of obesity, alcohol, and salt intake among those of lower socio-economic status in developed countries, and higher socio-economic status in developing countries. Recent evidence, to be discussed later, suggests that diets rich in fruit and vegetables with low total and saturated fats may protect against hypertension. Low calcium intake, although associated with hypertension in population studies, is now considered to play no part in pathogenesis.

Obesity

Fat people have higher blood pressures than thin people. This is not merely a consequence of cuff artefact, which is the tendency to overestimate blood pressure in fat arms when small cuffs are used, because the relation persists after correcting for arm circumference. Most studies have used body mass index (normal range 20 to 25kg/m^2) as a measure of obesity, although it is probable that high blood pressure correlates more closely with an android (apple-shaped) rather than a gynoid (pear-shaped) fat distribution. This suggests that at least some of the association between blood pressure and obesity is due to sex hormones, which are an important determinant of these body features.

Epidemiologically there is an association between high blood pressure, obesity, impaired glucose tolerance, and dyslipidaemia (particularly low high-density lipoproteins with high triglycerides). The possibility that obesity causes insulin resistance which leads in turn to the other metabolic disturbances in the so-called insulin resistance syndrome is the subject of a great deal of research, and may explain why fat people are more prone to heart disease. The cause and effect relationship between obesity and hypertension has been confirmed in randomized trials, which show that blood pressure falls when hypertensive obese patients are given calorie restricted diets. The degree of blood pressure reduction is variable, but a 1mmHg fall in diastolic blood pressure for each kilogram reduction in body weight might reasonably be anticipated from the data in these trials.

Alcohol

Epidemiological data have consistently shown an association between alcohol intake and blood pressure, while intervention trials confirm that blood pressure falls when alcohol is withdrawn from heavy drinkers. Moderate drinking (2 to 3 units daily where 1 unit is equivalent to 8 to 10g of ethanol) does not appear to exert a pressor effect, and is likely to be beneficial in coronary heart disease. The mechanism of the pressor effect in heavy drinkers is unknown. In a prospective study of 490000 men and women in the United States, the relative risk of death from cardiovascular disease in moderate drinkers compared with non-drinkers was 0.7 for men and 0.6 for women.

Salt

The role of dietary sodium intake in the pathogenesis of essential hypertension, and of dietary sodium restriction in its treatment, is probably more controversial now than it was at the time of the last edition of this textbook. The failure to observe a rise in blood pressure with ageing and the absence of essential hypertension in some non-Westernised cultures has been attributed to very low sodium intake. However, there are other reasons why blood pressure in primitive societies may differ from blood pressure in the United Kingdom, United States, and Europe. Opinions on the merits or otherwise of sodium restriction are similarly polarized.

Those who favour the sodium hypothesis point to the results of randomized controlled trials of sodium restriction in hypertensive subjects. These suggest that sodium intake can be halved to less than 100mmol/day with a fall in blood pressure of approximately 4/2mmHg, leading to a worthwhile reduction in the number of individuals requiring antihypertensive therapy. By contrast, those who argue against the sodium hypothesis can claim that a systolic pressure reduction of less than 1mmHg in normotensive subjects does not support a general recommendation to reduce sodium intake; and that a low sodium diet may have adverse effects on plasma renin, aldosterone, noradrenaline, total cholesterol, and low-density lipoprotein cholesterol.

Ultimately the health effects of a low sodium diet will only be resolved by trials relating nutrition to morbidity and mortality. Until such time as these have been completed, public policy recommendations regarding sodium intake for the general population should probably be withheld. Moderate salt restriction can still be recommended as a supplementary therapy in hypertensive individuals requiring drug treatment.

Fetal and infant growth

Babies who are small at birth have higher blood pressures during adolescence and are more likely to be hypertensive as adults. The relationship strengthens with advancing age such that subjects in their 60s show a decrease of approximately 5mmHg for every 1kg increase in birth weight. This has been interpreted as evidence that differences in blood pressure are initiated *in utero* and then amplified during adult life, and may explain a number of important clinical findings, including the higher prevalence of hypertension in black individuals, who are more likely to have small babies than white individuals, and also the increased risk of coronary heart disease seen in adults of low birth weight. For further discussion of this and similar issues see Chapter 15.4.1.1.

Psychosocial stress

It is well known that acute stress can raise blood pressure acutely—the act of taking blood pressure for example can increase systolic by up to 75mmHg—and it has long been suspected that chronic stress may be a risk factor for hypertension. However, the role of chronic stress has been difficult to assess, partly because stress means different things to different people, and partly because stress is not easy to measure.

Recently, using sophisticated measures of assessment including ambulatory monitoring, it has been shown that in men, but not in women, job strain is associated with an elevated blood pressure, not only at work but also while at home and during sleep. Job strain is defined here as the result of a highly demanding job with low control, as in a shop-floor worker. Subjects in demanding jobs who are able to exert control over their work patterns (e.g. doctors !), do not show the same elevation of blood pressure. The effect of job strain on blood pressure is independent of other environmental influences, and is as strong as that of obesity.

Blood pressure control mechanisms

The third components of the triad leading to essential hypertension are the blood pressure control mechanisms. Necessarily, these become involved as hypertension develops. The challenge for the researcher is to know whether abnormalities of the regulatory mechanisms are cause, consequence, confounder, or coincidental change. By comparison with our knowledge of environmental influences, this is an area fraught with difficulty.

Importance of the kidney

Transplantation experiments in genetically hypertensive rats, and observations following renal transplantation in humans, suggest that hypertension must result at least in part from renal mechanisms. For example, if a kidney from a genetically hypertensive rat is given to a control rat then that animal develops high blood pressure. Conversely, transplantation of a control kidney into a genetically hypertensive rat prevents the development of hypertension. Further studies have suggested that the genetic abnormality in the kidney expresses itself as a difficulty in handling sodium.

The resistance vessels

Blood pressure (BP) is a haemodynamic variable that depends on two other haemodynamic factors— cardiac output (CO) and total peripheral resistance (TPR), where BP = CO × TPR. The hallmark of essential hypertension is increased peripheral resistance with a normal cardiac output, and because of this any discussion on the pathogenesis of essential hypertension must centre on the resistance vessels. These are not the large arteries or capillaries, but the small arterioles.

The walls of small arterioles contain smooth muscle cells that respond to both circulatory and local hormonal influences, and to neural input through the sympathetic nervous system (Fig. 6). There is evidence that increased pressure within these small vessels leads to structural changes in vascular morphology, particularly an increase in the ratio of media thickness:lumen diameter. Recent data suggest this occurs mainly as a result of vascular remodelling (the rearrangement of existing material around a small lumen) and to a lesser extent by myocyte hypertrophy.

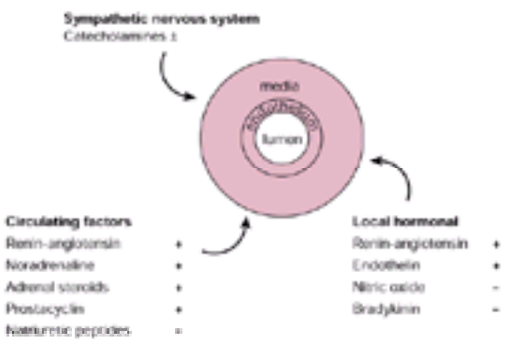

Fig. 6 Arteriolar tone is determined by circulatory and local hormonal influences, also by neural input through the sympathetic nervous system. + vasoconstriction, - vasodilation.

The possibility that these structural changes might act as vascular amplifiers in the hypertensive circulation, both contributing to and maintaining the rise in pressure, has been the subject of much debate. While undoubtedly attractive, the amplifier theory is not supported by results of studies that show no increase in pressor responsiveness when vasoconstrictor stimuli are infused into chronically hypertensive rats; or by restoration of normal pressure in rats when a pressor stimulus is

withdrawn, before structural changes have reversed. The clinical significance of the structural changes in small blood vessels is therefore uncertain.

Atrial natriuretic peptides

A bewildering number of homeostatic mechanisms interact in a complex fashion to maintain blood pressure and adjust it in response to changing circumstances. Atrial natriuretic peptide (**ANP**) is one of a family of natriuretic peptides whose other members are brain natriuretic peptide (**BNP**) and C-type natriuretic peptide (**CNP**). ANP is secreted primarily by the right atrium when the atrial wall is stretched. BNP was identified initially in the brain but is also present in the ventricles and circulates in the blood at approximately one-fifth of the plasma level of ANP. CNP is produced by vascular endothelial cells and in the kidney.

The actions of ANP, which are mediated by its attachment to a specific receptor on the cell membrane, the natriuretic peptide receptor A, include natriuresis, diuresis, decreased secretion of renin and aldosterone, vasodilatation, and a modest fall in blood pressure. These effects raise the possibility that ANP may be involved in the development and maintenance of high blood pressure in essential hypertension. ANP is indeed implicated in the regulation of arterial pressure, but a lack of ANP is unlikely to be the cause of essential hypertension. From a therapeutic viewpoint, however, it is now possible to reduce blood pressure using a new class of drugs that combine neutral endopeptidase inhibition (which blocks the breakdown of ANP) and ACE inhibition.

Endothelial-based systems

The endothelium is known to play an important role in the regulation of vascular tone. Endothelial cells form nitric oxide from L-arginine via the activity of nitric oxide synthase. Nitric oxide, formally known as endothelium-derived relaxing factor or EDRF, is a powerful local vasodilator that also inhibits platelet aggregation and vascular smooth muscle cell proliferation. The action of nitric oxide is opposed by endothelin, a powerful vasoconstrictor peptide also secreted by endothelial cells, and a host of other vasoconstrictor influences. For further discussion see Chapter 15.1.1.2.

Using an analogue of arginine called L-NMMA which inhibits the action of nitric oxide synthase, it has been possible to show a decrease in nitric oxide production in the vasculature of patients with high blood pressure. However, the evidence suggests this is more likely to be a consequence than a cause of hypertension in humans. Equally, the demonstration that endothelin receptor antagonists reduce blood pressure does not prove that an excess of endothelin is the cause of hypertension.

Renin–angiotensin systems

Renin is an enzyme produced by the juxtaglomerular apparatus of the kidney in response to falls in renal perfusion pressure, sodium depletion, and increased sympathetic nerve activity. Renin acts on its substrate angiotensinogen to produce the decapeptide angiotensin I, which in turn is cleaved by angiotensin-converting enzyme to give angiotensin II. Angiotension II is the effector component of the system, with a number of important actions on blood vessels (contraction), heart (hypertrophy), kidney (glomerulosclerosis), and drenal cortex (release of aldosterone). In health, aldosterone feeds back on the kidney to cause sodium retention and potassium excretion, and in this way homeostasis is maintained (Fig. 7).

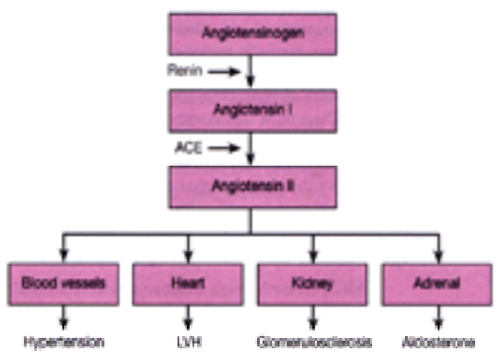

Fig. 7 The renin–angiotensin system.

Recent interest in the renin–angiotensin system has focused on the demonstration that renin may be synthesized in a number of tissues apart from the kidney, including adrenal, heart, the blood vessel wall, and brain. Tissue renin–angiotensin systems have been implicated in the regulation of mineralocorticoid secretion by the adrenal cortex, left ventricular hypertrophy, resistance vessel hypertrophy, and central nervous control of blood pressure.

The role of the renin–angiotensin system in the pathogenesis of essential hypertension is unclear. Plasma renin levels vary widely in essential hypertension from low (30 per cent), to normal (50 per cent), to high (20 per cent). Hypertensive individuals with low renin are usually considered to have volume-dependent hypertension (that is their renin levels are suppressed), whereas high renin in hypertensive individuals may reflect increased levels of sympathetic nervous system activity. Black individuals and the elderly have a high prevalence of low renin hypertension, which may be the reason these groups respond best to diuretics. In everyday clinical practice, however, baseline renin measurements are hardly ever requested or needed.

Sympathetic nervous system

The sympathetic nervous system is known to be involved in the regulation of arteriolar resistance (Fig. 6), also of cardiac output, renin release by the kidney, and catecholamine and mineralocorticoid release by the adrenal gland, and as such might reasonably be expected to have a role in the pathophysiology of essential hypertension. There is no evidence however of sustained overactivity of the sympathetic nervous system in essential hypertension, and even though drugs that block the sympathetic nervous system can lower blood pressure, this does not prove that overactivity of the system is the cause of the disease. In summary, it seems unlikely that a defect in any one of the regulatory mechanisms is directly responsible for the rise in blood pressure in essential hypertension. More probably essential hypertension is a consequence of interactions between several mechanisms, the exact nature of which remain to be determined.

Further reading

Appel LG *et al*. (1997). A clinical trial of the effects of dietary patterns on blood pressure. DASH Collaborative Research Group. *New England Journal of Medicine* **336**, 1117–24. [New evidence that diets rich in fruit and vegetables with low total unsaturated fats may protect against hypertension.]

August P, Oparil S (1999). Hypertension in women. *Journal of Clinical Endocrinology and Metabolism* **84**, 1862–6. [Up-to-date review.]

Burt VL *et al*. (1995). Prevalence of hypertension in the US adult population. Results from the Third National Health and Nutrition Examination Survey 1988–1991. *Hypertension* **25**, 305–13. [Large United States cross-sectional survey of hypertension prevalence, treatment, and control.]

Colhoun HM, Dong W, Poulter NR (1998). Blood pressure screening, management and control in England: results from the Health Survey for England 1994. *Journal of Hypertension* **16**, 747–52. [Includes prevalence data for hypertension in a contemporary English population.]

Flack J *et al*. for the Multiple Risk Factor Intervention Trial Research Group (1995). Blood pressure and mortality among men with prior myocardial infarction. *Circulation* **92**, 2437–45. [A J-curve analysis which supports the view that the main risk associated with blood pressure in survivors of myocardial infarction is due to high rather than low blood pressure.]

Franklin SS *et al*. (1999). Is pulse pressure useful in predicting risk for coronary heart disease? The Framingham Heart Study. *Circulation* **100**, 353–60. [An analysis of Framingham data showing the predictive power of pulse pressure.]

Gibbons GH (1998). The pathophysiology of hypertension: the importance of angiotensin II in cardiovascular remodelling. *American Journal of Hypertension* **11**, 177S–181S. [Emerging data on the role of the tissue renin–angiotensin system.]

Gibbs CR, Beevers DG, Lip GYH (1999). The management of hypertensive disease in black patients. *Quarterly Journal of Medicine* **92**, 187–92. [Up-to-date review.]

Graudal NA, Galoe AM, Garred P (1998). Effects of sodium restriction on blood pressure, renin, aldosterone, catecholamines, cholesterols and triglyceride: a meta-analysis. *Journal of the American Medical Association* **279**, 1383–91. [The results of this meta-analysis do not support a general recommendation to reduce sodium intake]

Guidelines Sub-Committee (1999). 1999 World Health Organization International Society of Hypertension Guidelines for the Management of Hypertension. *Journal of Hypertension* **17**, 151–83. [The latest edition of the only international guideline.]

Heggarty AM (1997). Significance of structural changes in small arteries in hypertension. *Blood Pressure* **6**(Supp 2), 31–3. [A review of the changes that occur in small blood vessels and their possible clinical significance.]

Isles C (1995). Blood pressure in males and females. *Journal of Hypertension* **13**, 285–90. [An analysis of blood pressure in the Renfrew and Paisley Survey, one of the few United Kingdom population surveys to include both men and women.]

Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure (1997). The sixth report of the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Blood Pressure (JNC-VI). *Archives of Internal Medicine* **157**, 2413–46. [The latest United States guideline.]

Klag MJ *et al*. (1996). Blood pressure and end stage renal disease in men. *New England Journal of Medicine* **334**, 13–18. [An analysis of the MRFIT database showing that high blood pressure is a strong independent risk factor for renal failure.]

Law CM, Shiell AW (1996). Is blood pressure inversely related to birth weight? *Journal of Hypertension* **14**, 935–41. [Strength of evidence from a systematic review of the literature.]

Levy D (1999). The role of systolic blood pressure in determining risk for cardiovascular disease. *Journal of Hypertension* **17**(Supp 1), S15–S18. [All you ever wanted to know about systolic pressure in a single review.]

MacMahon S *et al*. (1990). Blood pressure, stroke and coronary heart disease. Part I, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* **335**, 765–74. [Important meta-analysis of studies showing relation between diastolic pressure and risk of vascular disease.]

Mancia G, Zanchetti A (1996). White coat hypertension: misnomers, misconceptions and misunderstandings: what should we do next? *Journal of Hypertension* **14**, 1049–52. [Unravels the mysteries of ambulatory blood pressure monitoring.]

Mosterd A *et al*. (1999). Trends in the prevalence of hypertension, antihypertensive therapy and left ventricular hypertrophy from 1950–1989. *New England Journal of Medicine* **340**, 1221–7. [A recent analysis from Framingham which supports the view that increasing use of antihypertensive medication has resulted in a reduced prevalence of high blood pressure and left ventricular hypertrophy.]

O'Brien E, Staessen JA (1999). What is 'hypertension'? *Lancet* **353**, 1541–3. [Editorial which discusses the confusion still surrounding the definition of hypertension.]

O'Brien E *et al.* (2001). Blood pressure measuring devices: recommendations of the European Society of Hypertension. *British Medical Journal* **322**, 531–636. [Includes alternatives to the mercury sphygmomanometer.]

Owens P *et al*. (1998). Ambulatory blood pressure in the hypertensive population: patterns and prevalence of hypertensive subforms. *Journal of Hypertension* **16**, 1735–43. [Study confirming that isolated systolic hypertension is the most common form of hypertension in the elderly.]

PROGRESS Collaborative Group (2001). Randomised trial of a perindopril based blood pressure lowering regimen among 6105 individuals with previous stroke or transient ischaemic attack. *Lancet* **358**, 1033–41.

Ramsay LE *et al*. (1999). Guidelines for management of hypertension: report of the Third Working Party of the British Hypertension Society. *Journal of Human Hypertension* **13**, 569–92. [The latest British guideline.]

Schnall PL *et al*. (1998). A longitudinal study of job strain and ambulatory blood pressure: results from a three year follow up. *Psychosomatic Medicine* **60**, 697–706. [New evidence supporting the hypothesis that job strain is an occupational risk factor in the aetiology of essential hypertension.]

Thun MJ *et al*. (1997). Alcohol consumption and mortality among middle aged and elderly US adults. *New England Journal of Medicine* **337**, 1705–14. [The definitive meta-analysis of the protective effects of moderate alcohol intake.]

Van den Hoogen PCW *et al.* (2000). The relation between blood pressure and mortality due to coronary heart disease among men in different parts of the world. *New England Journal of Medicine* **342**, 1–8. [Prospective observational study showing similar relative risk but widely differing absolute risk of blood pressure for coronary heart disease in seven countries.]

15.16.1.2Genetics of hypertension

N. J. Samani

Historical perspective

The concept that genetic factors may be involved in causing hypertension goes back more than two hundred years and predates the ability to measure blood pressure. In 1769, Morgagni observed that the father of a patient who had died of cerebral haemorrhage had himself died of 'apoplexy' (stroke). The history of the genetics of hypertension is marked by a celebrated debate in the 1950s and 1960s between Platt and Pickering, two doyens of British medicine. On the basis of a finding of a bimodal distribution of blood pressures in some families of patients with hypertension, and evidence of hypertension transmitted over three generations in a few pedigrees, Platt argued that hypertension was a distinct genetic disorder with a likely autosomal dominant mode of inheritance. By contrast, Pickering and colleagues showed that in the general population there was no obvious discontinuity of blood pressure distribution and that the familial resemblance of blood pressure spanned the whole range of blood pressures, and was not different for those with hypertension. Thus, Pickering argued that blood pressure, like height and weight, was a quantitative trait, and that although there was a significant genetic contribution, this was polygenic and that hypertension represented one extreme of the trait but was not a distinct disorder, except perhaps for rare monogenetic forms embedded in the blood pressure distribution curve. Today, the overwhelming mass of evidence supports the Pickering concept.

Genetic epidemiology of blood pressure

The extent of familial aggregation of blood pressure has been studied in diverse ethnic groups living in different places, ranging from Polynesians to Middle Americans. A remarkably consistent level of correlation of around 0.2 between first-degree relatives has been found, that is, if the blood pressure of one member of the family deviates from the norm by +10 mmHg, the first-degree relative will deviate +2 mmHg on average. Studies in children and infants suggest that the familial resemblance in blood pressure starts very early and is maintained throughout the rest of life.

Attempts to partition the familial resemblance of blood pressure between shared genes and shared environment have been made through studies of adoptees and twins. In the Montreal Adoption Study, correlations between natural siblings compared with adoptive siblings and between parents and natural children compared with parents and adopted children were at least twice as great. Similarly, several studies have documented much higher correlations in blood pressure between monozygotic twins (0.55 to 0.85) compared with dizygotic twins (0.25 to 0.50), although the results from twin studies have to be viewed with some caution as there is substantial evidence of excess sharing of sociocultural environments by twin pairs, especially monozygotic twins.

However, taken together the epidemiological data suggest that genetic factors account for about 30 to 35 per cent of the population variability of blood pressure, common household environment for about 10 to 15 per cent, and non-familial factors for the remaining 50 to 55 per cent.

Genetic predisposition to hypertension

Although determination of familial correlations of blood pressure provides an overall view of the impact of heredity in determining blood pressure, a more relevant measure of the importance of genetic factors in determining susceptibility is relative risk. This is the ratio of the risk of an individual developing the condition given its presence in a first-degree relative compared with the overall population risk. For relatively rare monogenetic conditions such as cystic fibrosis, relative risk is as high as 500. For common and complex polygenic disorders, relative risk tends to be much lower. For hypertension, relative risk estimates vary between 2 and 5 depending on the criteria used to define family history. Values are highest when both parents have hypertension before the age of 55 years.

Apart from the increased familial risk, two other observations provide support for an important contribution of genetic factors in the pathogenesis of hypertension. First, spontaneous as well as salt-dependent hypertension can be inbred into animal strains. Second, there are a number of rare monogenetic forms of hypertension and hypotension, where the presence of a single defective gene is sufficient to cause altered blood pressure (Table 1). The molecular basis of several of these disorders has now been elucidated.

Mendelian forms of hypertension

Hypertension and hypokalaemia are features of 11b-hydroxylase and 17b-hydroxylase deficiency—two rare recessive gene disorders of adrenal steroid-synthesizing enzymes that, among others, cause congenital adrenal hyperplasia (see Chapter 12.7.2). 11b-Hydroxylase deficiency usually presents in infancy or early childhood with virilization of both sexes, while presentation of 17b-hydroxylase deficiency may be delayed until adolesecence or adulthood. Hypertension due to a phaeochromocytoma may also be a feature of multiple endocrine neoplasia type 2 (Sipple's syndrome), which when familial is inherited in an autosomal dominant pattern, or rarely be a feature of neurofibromatosis (von Recklinghausen's disease).

Apart from these conditions, several other mendelian disorders (glucocorticoid remediable aldosteronism, syndrome of apparent mineralocorticoid excess, Liddle's and Gordon's syndromes) where hypertension is the predominant manifestation have now been characterized at the molecular level. Although diverse in their molecular basis, all of them, interestingly, impact ultimately on the homeostatic role of the kidney in maintaining sodium balance.

Glucocorticoid-remediable aldosteronism (GRA)

GRA is a form of mineralocorticoid hypertension. It is inherited in an autosomal dominant fashion. The hypertension is accompanied by hypokalaemia (not invariably), a tendency to metabolic alkalosis, an elevated plasma aldosterone level, and a suppressed renin level. The hypertension often responds to thiazides or spironolactone. Patients are usually suspected of having primary aldosteronism (Conn's syndrome), although the age of onset, usually in the first two decades of life, is younger than typical of primary aldosteronism. The two hallmark features of GRA are the presence of large amounts of two abnormal steroids—18-hydroxycortisol and 18-oxocortisol—in the urine, and the paradoxical response of the hypertension, with return of plasma aldosterone to a normal level and disappearance of the abnormal steroids following treatment over a few days with a low dose of exogenous glucocorticoid, such as 0.5 to 1.0 mg of dexamethasone per day (hence the name).

The molecular basis of GRA was solved by Lifton and coworkers in 1992. Patients with GRA have a chimeric gene due to an unequal crossing-over event at meiosis between two adjacent and highly homologous genes involved in adrenocorticosteroid synthesis—aldosterone synthase (CYP11B2) (involved in aldosterone synthesis and normally regulated by angiotensin II) and 11b-hydroxylase (CYP11B1) (involved in glucocorticoid synthesis and normally regulated by ACTH). In the chimeric gene, the regulatory elements of CYP11B1 have become attached to the aldosterone synthase coding region of CYP11B2 (Fig. 1(a)). Thus, the gene produces

aldosterone (and the other abnormal hormones), but under the control of ACTH and hence is suppressible by glucocorticoids, thereby explaining the clinical behaviour.

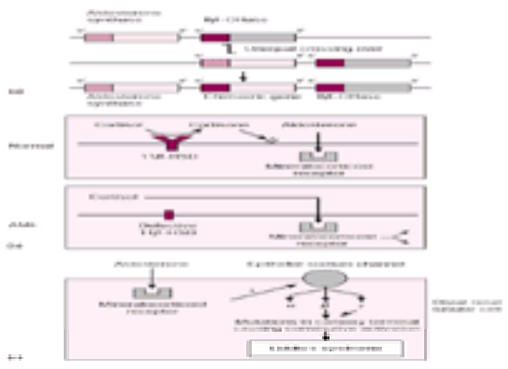

Fig. 1 Mechanisms underlying three forms of monogenetic hypertension. (a) Glucocorticoid-remediable aldosteronism—In GRA an unequal crossing-event leads to a chimeric gene where the coding region of aldosterone synthase (light pink bar) becomes attached to the regulatory region for 11b-hydroxylase (magenta bar). The chimeric gene produces excess amounts of aldosterone under the regulation of ACTH. (b) Syndrome of apparent mineralocorticoid excess—Normally, the mineralocorticoid receptor in the distal renal tubule is protected from stimulation by cortisol by the activity of the 11b-hydroxysteroid dehydrogenase enzyme. In AME, mutations in the enzyme allow cortisol to gain access to the receptor. (c) Liddle's syndrome—The trimeric epithelial sodium channel mediates sodium reuptake in the distal renal tubule under regulation by the mineralocorticoid receptor. In Liddle's syndrome, mutations in the b- and g-subunits of the channel render the channel constitutively active.

Syndrome of apparent mineralocorticoid excess (AME)

AME is an autosomal recessive disorder that usually presents in childhood with hypertension, hypokalaemia, and low renin activity. Despite the clinical features of mineralocorticoid excess, levels of all known mineralocorticoid hormones are low, yet the hypertension responds to spironolactone or amiloride. Patients with the disorder cannot metabolize cortisol to its inactive metabolite cortisone normally, resulting in a prolonged half-life of cortisol and a characteristic increase in urinary cortisol compared with cortisone metabolites.

Elucidating the defect causing AME first required the solution of another paradox—why cortisol, which circulates at a level several-fold greater than aldosterone, does not overwhelmingly activate the renal mineralocorticoid receptor *in vivo*, despite the two having equal affinity *in vitro* for the receptor. The reason relates to the enzyme 11b-hydroxysteroid dehydrogenase (**11b-HSD**), which has two isoforms. Type 1 11b-HSD is located in the liver, adipose tissue, and gonad and converts cortisone to cortisol. Type 2 11b-HSD is expressed in the mineralocorticoid target tissues—kidney, colon, and salivary gland—and inactivates cortisol to cortisone. In the kidney, the enzyme plays the crucial role of protecting the mineralocorticoid receptor on the distal tubule from activation by cortisol. In subjects with AME, a variety of disabling mutations in the type 2 11b-HSD gene cause a deficiency of the enzyme, allowing cortisol access to the mineralocorticoid receptor (Fig. 1(b)).

AME resembles the syndrome observed in subjects ingesting large amounts of liquorice or taking the now redundant antiulcer drug carbenoxolone. Both liquorice and carbenoxolone contain glycyrrhetinic acid, which inhibits type 2 11b-HSD. This, therefore, explains the hypertension and hypokalaemia observed with these compounds. Spillover access of cortisol to the mineralocorticoid receptor may also, at least in part, explain the hypertension accompanying some forms of Cushing's syndrome and glucocorticoid resistance.

Liddle's syndrome

Liddle described a family in which the siblings were affected by early-onset hypertension and hypokalaemia, but with low renin and aldosterone levels. The clue to the nature of the molecular defect underlying this autosomal dominant disorder came from the observation that the hypertension does not respond to spironolactone, the mineralocorticoid receptor antagonist, but does respond to direct inhibitors (such as triamterene or amiloride) of the epithelial sodium channel which mediate the effects of activation of the mineralocorticoid receptor. This indicated that the defect lay downstream of the mineralocorticoid receptor, and Lifton and coworkers subsequently showed that the syndrome arises due to activating mutations in the b- or g-subunits of this trimeric channel (Fig. 1(c)). All mutations identified so far cause an alteration or deletion of a proline-rich (PY) motif in the carboxy-terminal cytoplasmic tails of the subunits. This motif is necessary for regulatory proteins such as Nedd4 to bind and internalize the channel, and when its function is impaired the channel remains constitutively active at the cell surface.

Gordon's syndrome

Pseudohypoaldosteronism type II (PHA-II, also known as Gordon's syndrome), is an autosomal dominant disorder characterized by hyperkalaemia despite normal renal glomerular filtration, hypertension, and correction of physiological abnormalities by thiazide diuretics. Mild hyperchloraemia, metabolic acidosis, and suppressed plasma renin activity are variable associated findings. Genes for PHA-II have been mapped in different families to chromosomes 17, 1, and 12. Recently, the causative genes on chromosomes 12 and 17 have been identified as two members, WNK1 and WNK4, of the WNK family of serine-threonine kinases. Both proteins localize to the distal nephron and gain-of-function mutations are thought to be responsible for causing the abnormalities, although the precise mechanisms leading to the disturbance in electrolyte transport remain to be determined.

Other monogenetic forms of hypertension

A missense mutation in the ligand-binding domain of the mineralocorticoid receptor (MR) has been found to cause an autosomal dominant form of hypertension that is markedly accelerated in pregnancy. The mutation, MR S810L, causes partial, aldosterone-independent, activation of the receptor, causing carriers to develop hypertension before age 20. More interestingly, compounds such as progesterone that normally bind but do not activate MR are all potent agonists of the mutant receptor. Since pregnancy is accompanied by a 100-fold rise in progesterone, MR S810L carriers have dramatic acceleration of hypertension during pregnancy. Although the MR S810L mutation is extremely rare, the finding does raise the question of whether related mechanisms may underlie other forms of hypertension in pregnancy.

A gene causing autosomal dominant hypertension in conjunction with type E brachydactyly in a large Turkisk kindred has been mapped to chromosome 12p. The hypertension in this syndrome, unlike most of the disorders described above, closely resembles essential hypertension with no evidence of volume expansion or electrolyte imbalance. Elucidation of the genetic defect is keenly awaited.

Genetic defects causing hypotension

Just as single-gene disorders have been identified that cause hypertension, a number of mendelian syndromes where hypotension is a feature have recently been characterized at a molecular level (Table 1). Many are mirror images of the genetic abnormalities causing the mendelian forms of hypertension described above. Pseudohypoaldosteronism type 1 (PHA-I) occurs in two forms, autosomal recessive and autosomal dominant. Both are characterized by life-threatening dehydration in the neonatal period, hypotension, salt wasting, hyperkalemia, metabolic acidosis, and marked elevation of renin and aldosterone. The autosomal recessive form is due to inactivating mutations (compare with Liddle's syndrome) in any of the subunits of the epithelial sodium channel, while the autosomal dominant form is due to loss-of-function mutations in the mineralocorticoid receptor.

Gitelman's syndrome is also an autosomal recessive disorder. It is characterized by hypotension, neuromuscular abnormalities, hypokalaemia, metabolic alkalosis, and an activated renin–angiotensin system. It arises due to inactivating mutations in the gene encoding the renal thiazide-sensitive NaCl cotransporter.

Bartter's syndrome is distinguished from Gitelman's syndrome by hypercalciuria and presentation in the neonatal period with life-threatening hypotension. This

disease is caused by mutations in one of several genes that are required for normal salt absorption in the thick ascending loop of Henle.

Does my patient have a recognized form of monogenetic hypertension?

Finding that a patient has GRA, AME, Liddle's syndrome, or Gordon's syndrome has important consequences for treatment (see above) and family screening. However, all of these syndromes are extremely rare and suspicion will usually go unrewarded. Phenotypic expression is highly variable. Features that may suggest a diagnosis of mendelian hypertension include a young age of onset, moderate to severe hypertension, strong family history, and electrolyte abnormalities, particularly of potassium (although this is not invariable). A good starting point is the measurement of plasma renin activity (suppressed in all three syndromes) and plasma aldosterone. If the aldosterone is significantly elevated then the differential diagnosis lies between the various forms of Conn's syndrome and GRA. Diagnosis of GRA would be supported by the finding of elevated 18-hydroxycortisol and 18-oxocortisol in the urine, and can now be relatively easily confirmed by finding a chimeric gene fragment with DNA testing. If the aldosterone level is suppressed, then finding an increased ratio of cortisol/cortisone metabolites in the urine would support a diagnosis of AME. The presence of hyperkalaemia, hyperchloraemia, and metabolic acidosis would suggest a diagnosis of Gordon's syndrome. No biochemical abnormalities specifically support a diagnosis of Liddle's syndrome. Ultimately, diagnosis of AME, Liddle's syndrome, and Gordon's syndrome also requires DNA confirmation, but this is not as straightforward as it is with GRA since several different mutations can give rise to each syndrome.

Progress towards identifying genes that increase susceptibility to essential hypertension

The genetic contribution to essential hypertension (and the population variability in blood pressure) is polygenic. However, little is known about the number of genes involved, their mode of transmission, their quantitative effect on blood pressure, their interaction with other genes, or their modulation by environmental factors. Further, the impact of genetic factors may be dependent on age, gender, and ethnicity, and may only influence specific blood pressure phenotypes such as systolic or diastolic blood pressure. Such complexities have contributed to the difficulty in elucidating the genetic basis of essential hypertension.

Despite evidence that the same genes that increase susceptibility to essential hypertension most likely also contribute to normal blood pressure variation, much of the work to date has focused on the former. Two main approaches have been used: (i) association studies in which the frequency of specific alleles are compared in hypertensive and normotensive subjects, (ii) sibling-pair linkage studies, which test whether specific alleles or markers are shared more often by hypertensive sibling pairs than would be expected by chance. Association studies are most suited to study candidate genes, while with affected sibling pairs the whole genome can also be scanned using anonymous microsatellite markers to identify areas of linkage.

Association studies

Variants in over 30 candidate genes have been associated with hypertension (MEDLINE search under hypertension x genetics), although for the majority the findings are as yet far from robust. However, data for a number of genes are sufficiently persuasive, or otherwise interesting, to merit specific mention.

Angiotensinogen

Cleavage of angiotensinogen (**AGT**) by renin is the rate-limiting step in the generation of angiotensin II, the effector molecule of the renin–angiotensin system. Evidence for the involvement of the AGT gene comes from both linkage of the AGT locus to hypertension in Caucasian and Afro-Carribean hypertensive sibling pairs and association of a polymorphism (M235T) in the gene with hypertension. In a meta-analysis of 32 case–control studies corresponding to 13 760 patients, the TT genotype conferred a 31 per cent increased risk of hypertension compared with the MM genotype. Other estimates suggest that mutations in the AGT gene might be predisposing factors for hypertension in 3 to 6 per cent of subjects with onset before the age of 60. Data showing that the TT genotype is also associated with higher plasma and possibly tissue AGT levels, with its potential impact on angiotensin generation, provides a plausible explanation for the association of the AGT gene with hypertension.

Angiotensin- converting enzyme

Angiotensin converting enzyme (ACE) cleaves angiotensin I to form the vasoactive peptide angiotensin II. A common variant in the ACE gene, the insertion/deletion (I/D) polymorphism, is strongly associated with differences in plasma ACE levels. Although initial linkage and association studies were negative, recent analyses have shown significant association of the I/D polymorphism with blood pressure at least in males. Given the widespread and increasing use of ACE inhibitors and angiotensin receptor antagonists in treating hypertension and related cardiovascular diseases, these observations, if confirmed, could have pharmacological implications.

a-Adducin

Adducin is a ubiquitous a/b heterodimeric cytoskeletal protein that promotes the assembly of actin with spectrin. In the Milan hypertensive rat, point mutations in the adducin a- and b-subunits affect actin assembly and Na–K pump activity and possibly explain up to 50 per cent of the blood pressure difference compared with the Milan normotensive rat. In humans a polymorphism changing glycine for tryptophane at codon 460 in the a-adducin gene has been found to be more common in individuals with hypertension in Italian, French, and Japanese populations, although other studies have not shown an association. More persuasive are data showing a direct functional impact of the 460Trp variant. In an acute salt-sensitivity test, where change in blood pressure from a state of salt-loading to one of salt-depletion was measured, hypertensive individuals heterozygous for the 460Trp allele had a much greater decrease in mean arterial pressure than hypertensive individuals homozygous for the 460Gly allele (15.9 [SE 2.0] compared with 7.4 [SE 1.3] mmHg). Similarly, heterozygous hypertensive individuals showed a much greater fall in mean arterial pressure in response to 2 months' treatment with hydrochlorthiazide than did 460Gly homozygote individuals (14.7 [2.2] compared with 6.8 [1.4] mmHg). Consistent with these observations, the 460Trp variant was associated with a reduced-slope (more salt sensitive) pressure–natriuresis curve. Further confirmation is required, but these findings suggest that variation at the adducin locus may contribute to salt sensitivity and through it to susceptibility to hypertension in some populations.

G protein b3 subunit

Increased activity of the pH-regulating transporter system, the sodium-proton exchanger, is seen in some patients with essential hypertension. Siffert and coworkers found that this was due to enhanced intracellular signal transduction via pertussis toxin-sensitive heterotrimeric G proteins. Sequencing revealed a polymorphism (C825T) in the gene encoding the b3 subunit (GNB3), which although itself silent, appears to cause a biologically active splice variant, Gb3-s, missing 41 amino acids in those carrying the T allele. In an initial study of 426 hypertensive and 427 normotensive subjects, there was a significant association of carriage of the T allele with hypertension (53 compared with 44 per cent). In a further study, where hypertensive subjects were recruited on the basis of a very strong family history of premature hypertension affecting both parents, the association of hypertension with the T allele was even stronger. These early findings suggest that in some patients hypertension may be related to inherited dysfunction of G proteins.

Epithelial sodium channel

Attempts to show that less severe mutations in the genes responsible for monogenetic forms of hypertension (see above) may underlie essential hypertension have by and large met with disappointment. One exception may be the epithelial sodium channel involved in Liddle's syndrome, where a variant in the C-terminus of the b-subunit (T594M) was found to be four times more frequent in hypertensive black subjects resident in London compared with normotensive black subjects (8 compared with 2 per cent). Subjects carrying the mutation demonstrate increased activity of the nasal epithelial sodium channel (a possible surrogate marker of the renal channel) and lower plasma renin activity, supporting the notion of increased sodium reabsorption and volume-dependent hypertension. In turn, this could at least partly explain the generally poor response of black hypertensive subjects to angiotensin-converting enzyme inhibitors.

Epistatic interactions

Current findings (see above) suggest that individually, gene variants will, at best, only contribute a small amount to the risk of hypertension. Interest is therefore turning to additive and epistatic interactions. There are only a few publications, but in one recent study by Staessen and coworkers, possession in combination of the ACE DD genotype, the a-adducin Trp allele, and the aldosterone synthase CC genotype (at the –344C/T polymorphism), increased the risk of developing hypertension by 252 per cent compared with other genotypes. Over a median follow-up of 9.1 years the cumulative incidence rates for hypertension were 71.0 cases

per 1000 person-years in those carrying the three risk genotypes compared with 20.2 cases per 1000 person-years in those without. If confirmed, the findings could have implications for prediction and primary prevention

Linkage studies

These have lagged behind association studies. However increasing numbers are being reported. The most consistent data have been found for a region on chromosome 17 where linkage has been found, not only to hypertension analysed as a qualitative trait in a panel of French and United Kingdom affected sib pairs (Julier and coworkers), but also to blood pressure analysed as a quantitative trait in subjects from the Framingham Heart Study (Levy and coworkers). Involvement of a gene in this region in blood pressure regulation is further supported by the fact that the syntenic interval is also linked to blood pressure in several strains of genetically hypertensive rats. Interestingly, the region overlaps with the chromosome 17 locus for Gordon's syndrome where a mutation in the WNK4 gene has recently been identified to be the cause. Whether other mutations in WNK4 are responsible for the reported linkage to essential hypertension and blood pressure variability remains to be determined.

Future perspectives

Recent progress suggests that in the next few years much more will be understood about the nature of individual genetic factors that influence susceptibility to hypertension and the environmental/genetic context in which they exert their effect on blood pressure. It may prove possible to sub-categorize patients with essential hypertension on the basis of molecular mechanisms. Although this information will not help with diagnosis (we already have an accurate tool– the sphygnomanometer), there could be several important applications. Preventive measures, such as salt restriction, could be better targeted if findings, for example with a-adducin, are confirmed. The information would almost certainly influence choice of antihypertensive medication (pharmacogenetics). Several studies have shown that individuals show significant variation in their response to different classes of antihypertensive agents. At present the choice of drug is largely empirical. The early findings with a-adducin and the epithelial sodium channel gene illustrate how the presence of a particular genetic variant could influence this choice. Finally, there is increasing experimetal as well as some clinical data to show that genetic factors not only contribute to the development of hypertension but also influence the development of end-organ damage. The ultimate goal of hypertension management is not just to lower blood pressure but to prevent complications, and strategies based on a more refined assessment of risk will undoubtedly be more cost-effective.

Further reading

Baker EH *et al.* (1998). Association of hypertension with the T594M mutation in the b subunit of epithelial sodium channel in black people resident in London. *Lancet* **351**, 1388–92. [The same gene as that involved in Liddle's syndrome may cause essential hypertension in some ethnic groups.]

Cusi D *et al.* (1997). Polymorphisms of a-adducin and salt sensitivity in patients with essential hypertension. *Lancet* **349**, 1353–7. [Key paper on the role of adducin.]

Geller DS *et al*. (2000). Activating mineralocorticoid receptor mutation in hypertension exacerbated by pregnancy. *Science* **289**, 119–23. [Describes an activating mutation in the mineralocorticoid receptor which makes it responsive to progesterone.]

Jeunemaitre X *et al.* (1992). Molecular basis of human hypertension: role of angiotensinogen. *Cell* **71**, 169–80. [Seminal paper describing involvement of the angiotensinogen gene in essential hypertension.]

Julier C *et al.* (1997). Genetic susceptibility for human essential hypertension in a region of homology with blood pressure linkage on rat chromosome 10. *Human Molecular Genetics* **6**, 2077–85. [Identification of the first locus for essential hypertension via methods not based on a candidate gene.]

Levy D *et al*. (2000). Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. *Hypertension* **36**, 477–83. [Findings from a quantitative trait approach supporting the presence of a gene on chromosome 17 that influences blood pressure.]

Lifton RP *et al.* (1992). A chimaeric 11 b-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension. *Nature* **355**, 262–5. [Describes the discovery of the molecular mechanism underlying glucocorticoid-remediable aldosteronism.]

Mune T *et al.* (1995). Human hypertension caused by mutations in the kidney isozyme of 11b-hydroxysteroid dehydrogenase. *Nature Genetics* **10**, 394–9. [Describes the discovery of the molecular mechanism underlying apparent mineralocorticoid excess.]

Shimkets RA *et al.* (1994). Liddle's syndrome: heritable human hypertension caused by mutations in the b subunit of the epithelial sodium channel. *Cell* **79**, 407–14. [Describes the discovery of the molecular mechanism underlying Liddle's syndrome.]

Siffert W *et al.* (1998). Association of a human G-protein b3 subunit variant with hypertension. *Nature Genetics* **18**, 45–8. [Findings suggesting that hypertension may be a G-protein related inherited disorder in some patients.]

Staessen JA *et al*. (2001). Effects of three candidate genes on prevalence and incidence of hypertension in a Caucasian population. *Journal of Hypertension* **19**, 1349–58. [One of the first studies demonstrating epistatic interactions between genes in increasing risk of hypertension.]

Swales JD (1985). *Platt verus Pickering: an episode in recent medical history.* Keynes Press, London. [Describes a celebrated debate about the nature of the genetic basis of essential hypertension.]

Ward R (1990). Familial aggregation and genetic epidemiology of blood pressure. In: Laragh JH, Brenner BM, eds. *Hypertension: pathophysiology, diagnosis and management*, pp 81–100. Raven Press, New York. [Comprehensive review of the evidence indicating a significant genetic contribution to hypertension.]

Wilson FH *et al*. (2001). Human hypertension caused by mutations in WNK kinases. *Science* **293**,1107–12. [Describes the identification and characterisation of mutations at two loci causing Gordon's syndrome.]

15.16.1.3 Essential hypertension

*J. Swales**

Pathology

Introduction

High blood pressure induces changes in the heart and blood vessels. These are partly the direct effect of cyclic stress on the arterial wall and also due to indirect trophic effects. In addition, turbulence and shear stress influence endothelial function. Secondary changes may then occur in the organs served by these vessels, giving rise to the clinical features of hypertension. However, not all the pathological changes observed in the cardiovascular system are the result of pressure (Table 1). Hypertension is associated with risk factors for atheroma such as dyslipidaemia and insulin resistance. Smooth muscle growth is enhanced in some animal models of hypertension, probably as a result of genetic factors. There may be other cellular influences at work. Most of the powerful vasoconstrictors such as angiotensin II, catecholamines, vasopressin, and endothelins are also mitogens and may therefore have trophic effects.

Blood vessel changes

Aorta and large arteries

Recurrent pulsatile stress produces uncoiling, disruption and calcification of elastic fibres. At the same time, relatively inelastic collagen is increased. This is a result of ageing as well as hypertension: both processes therefore cause loss of the normal elastic reservoir function of the aorta and large arteries. The effects are additive, so that changes occur at an earlier age in hypertensive than in normotensive subjects. Another reason for the loss of arterial compliance in hypertensive patients is that, as pressure increases, the elastic fibres become fully stretched, causing the inelastic collagen fibres to bear the load. As a result of these changes, the pressure wave generated by left ventricular contraction is no longer buffered by the aorta and proximal arteries, but is transmitted into the arterial tree with greater amplitude. This is manifested clinically as increased pulse pressure, with higher systolic and lower diastolic pressures. In addition, loss of compliance produces increased pulse wave velocity. Pulse waves are reflected back from peripheral sites, normally returning in late diastole as a secondary wave visible on the aortic trace. When the arterial tree is stiffened the return of these waves amplifies the aortic late systolic wave. This is not demonstrable when brachial artery pressure is measured, but throws an additional load on the left ventricle. In addition, lower diastolic pressure reduces coronary artery perfusion.

Large artery changes in hypertension were ignored for many years. However, it is now clear that pulse pressure is an important risk factor in cardiovascular disease. This explains one curious feature of elderly hypertensive patients. Diastolic blood pressure in patients with isolated systolic hypertension is inversely related to prognosis, that is, for any given systolic blood pressure, the lower the diastolic, the worse the risk.

Decreased compliance of the large arteries also has an important effect on the carotid and aortic baroceptors which normally buffer rapid changes in blood pressure. These become less sensitive. As a result, circulatory adaptation to rapid changes in posture may be impaired. This rarely causes problems, except in elderly hypertensive subjects, when age and blood pressure have additive effects. The clinical manifestation is as postural hypotension or post-prandial hypotension in elderly patients whose blood pressure has been overtreated.

Medium-sized arteries

These arteries perform a conduit rather than reservoir function, reflected in their lower elastin content. The predominant pathological change is wall thickening caused by increased deposition of collagenous material.

Resistance vessels

The characteristic structural change in the smaller arteries and arterioles responsible for peripheral vascular resistance is an increase in wall:lumen ratio. This has important functional consequences. The vessels can still dilate in response to stimuli such as warmth or drugs, but maximal vasodilatation is reduced. This change is a response to pressure; it can be prevented by protecting vessels from the increased pressure load in hypertension by means of a mechanical constriction. There may be a genetic factor in hypertension which regulates the structural response, the evidence for which is largely based on animal studies.

In recent years it has become clear that what was thought to be a trophic response is largely if not entirely due to rearrangement of smooth muscle cells around a smaller lumen. There is little if any evidence of hypertrophy and no evidence of hyperplasia (i.e. increased number of cells). The mechanism of this is entirely unknown. Local growth factors, such as angiotensin II have been implicated. Thus lowering blood pressure by angiotensin-converting enzyme (**ACE**) inhibition is better at reversing remodelling than lowering it by b-blockade.

Atheroma in hypertension

The increased prevalence of atheroma in hypertensive patients reflects the cumulative effect of several contributory factors. The importance of local mechanical consequences of increased pressure and turbulence on the arterial wall are demonstrated by the distribution of lesions and the absence of atheroma in the pulmonary vessels of hypertensive patients. Endothelial dysfunction is probably the first stage in initiating a complex chain of local cellular processes leading to the formation of the atheromatous plaque. Increased adhesion molecule expression, increased permeability, and cell migration, accumulation, and proliferation are all involved. Most studies in patients with essential hypertension have shown impairment of endothelium-dependent vasodilatation, although animal studies have shown more complex changes in endothelial function.

Systemic factors are also of importance in the development of atheroma in hypertensive patients. Atheroma cannot usually be produced in hypertensive rabbits unless their lipid levels are raised by feeding them cholesterol. A combination of high perfusion pressure above induced aortic constriction and a high cholesterol diet induces atheroma above, but not below the constriction. A low cholesterol diet prevents it. This is analogous to the situation in humans, where the presence of insulin resistance or diabetes in a hypertensive patient further increases the risk of atheromatous complications.

It has also been postulated that the renin–angiotensin system plays a role in atherogenesis. Renin and angiotensin II levels are often raised in severe hypertension and in hypertension associated with increased sympathetic nervous system activity. In laboratory studies angiotensin II enhances the smooth muscle proliferative response observed in atheroma. High renin levels have been shown to be associated with a worse prognosis in some studies, although this finding is still controversial. Since drugs may either block or activate the renin–angiotensin system, this debate has important therapeutic implications which will only be resolved

when the relevant trials report.

Vessels in malignant hypertension

The characteristic pathological lesion in malignant hypertension is fibrinoid necrosis (necrotizing arteriolitis). The normal structure of the vessel wall is lost and replaced with fibrin-like material. A variable cellular reaction takes place. Fibrinoid necrosis is usually associated with focal areas of vasodilatation and increased permeability. These changes are probably a primary mechanical effect. The endothelium of the dilated segments is disrupted and the vessel wall becomes permeable to particles as large as colloidal carbon. The increased permeability permits exudation of plasma into the media and local tissue destruction. The intima may become massively thickened by concentric collagenous rings as a result of locally released growth factors until the lumen is almost obliterated.

Specific organ changes in hypertension

The heart

Angina and myocardial infarction in the hypertensive patient are usually due to coronary atheroma. Coronary perfusion may also be lowered as a result of reduced diastolic pressures in patients with isolated systolic hypertension (see above). In a minority of patients, anginal pain occurs in the absence of significant coronary atheroma ('syndrome X'). A possible explanation is luminal narrowing of the coronary vessels associated with an increased wall:lumen ratio. Other explanations which have been put forward are increased vasomotor tone and increased pressure on the left ventricular wall causing subendocardial ischaemia. The most likely explanation for syndrome X in hypertensive patients is impairment of endothelial-dependent relaxation in the coronary vascular tree.

Left ventricular hypertrophy is demonstrable in about 50 per cent of untreated hypertensive patients when echocardiography is used, and in 5 to 10 per cent with electrocardiography using conventional criteria. Histologically, it is caused by an increase in size of cardiomyocytes with an increase in intercellular matrix. Its role as a powerful independent risk factor for cardiovascular disease and death has led to intensive research into its causes and reversal by treatment. Pressure load on the left ventricle is unquestionably important, as evidenced by the observation that ambulatory monitoring of blood pressure is much better correlated with left ventricular hypertrophy than clinic measurements of pressure. Furthermore, lowering blood pressure consistently produces regression of left ventricular hypertrophy. However, local trophic factors may also play a role. The sympathetic nervous system and renin–angiotensin–aldosterone system have been implicated by both clinical and experimental studies. Meta-analyses of clinical trials in which blood pressure was lowered by different classes of antihypertensive agent have suggested that ACE inhibitors have a greater effect in reversing left ventricular mass, even when other factors such as duration of therapy and degree of blood pressure reduction are taken into account. However, there have been no adequately sized trials in which agents have been formally compared, and this clinically very important possibility still remains to be demonstrated convincingly.

The bad prognosis carried by left ventricular hypertrophy in the hypertensive patient could have several explanations. These include:

1. Arrhythmias—there is increased prevalence of simple and complex ventricular arrhythmias in hypertensive left ventricular hypertrophy;
2. Relative ischaemia produced by increased muscle mass;
3. Decrease in ventricular compliance causing pulmonary oedema; and
4. Ventricular hypertrophy is also a marker for integrated exposure of the circulation to high blood pressure over a prolonged period.

Central nervous system

Cerebral (atherothrombotic) infarction in a hypertensive patient is usually attributable to atheroma of one of the larger cerebral arteries (usually the middle cerebral artery) and accounts for about 80 per cent of the strokes which these patients suffer. Intracerebral haemorrhage accounts for 10 to 15 per cent, usually the result of rupture of a small intracerebral degenerative microaneurysm (Charcot–Bouchard aneurysm). These lesions develop in the small (less than 200 μm diameter) perforating arteries in the region of the basal ganglia, thalamus, and internal capsule. Hyaline degeneration (lipohyalinosis) occurs in the aneurysmal wall with a defect in the media at the neck of the aneurysm. The incidence of Charcot–Bouchard aneurysms is closely correlated with age and blood pressure, the two factors acting additively so that lesions are rarely if ever seen in younger normotensive people. The remaining strokes in hypertensive patients are due to subarachnoid haemorrhage. Transient ischaemic attacks due to disease of extracranial vessels are also more frequent in hypertensive subjects.

More diffuse changes account for the cognitive decline that may occur, particularly in untreated hypertension and in older patients. Functional imaging studies have shown relative reductions in blood flow in parietal and forebrain areas in hypertensive patients during memory tasks and areas of cortical and subcortical hypometabolism. More advanced vascular disease gives rise to multiple, punctate, hyperintense white matter lesions on MRI. These are due to focal ischaemia, either as a result of lipohyalinosis or microatheromatous disease, tortuosity, and narrowing of the perforating arteries. All degrees of impairment of cognitive performance may occur as a result of these lesions, ranging from effects only detectable with sensitive psychometric testing to lacunar strokes and Binswanger's disease.

Hypertensive encephalopathy

The cerebral vessels usually constrict in the face of increased pressure and dilate in the face of decreased pressure to maintain a constant flow (autoregulation). Resistance vessel remodelling and hypertrophy seem to have a protective function in this respect, so that the autoregulatory range is raised in long-standing hypertension. When blood pressure rises above the autoregulatory range, however, focal areas of vasodilatation and localized perivascular oedema and fibrinoid necrosis occur. Focal haemorrhages, ischaemia, and infarction may result, giving rise to the clinical picture of encephalopathy.

The kidney

In non-malignant hypertension, glomerular filtration rate is well preserved. However, filtration fraction is increased since efferent glomerular arteriolar resistance increases more than afferent resistance, causing a rise in intraglomerular capillary pressure. The long-term renal damage produced by glomerular hypertension probably accounts for progressive glomerulosclerosis in essential hypertension. Thus, the decline in glomerular filtration rate with age is more rapid in hypertensive than normotensive subjects. This phenomenon is not usually significant in mild to moderate essential hypertension where endstage renal disease in the absence of any other lesion is unusual.

Hypertension-induced glomerulosclerosis is much more important, however, in severe and malignant hypertension and in the presence of intrinsic renal disease due to (for instance) diabetes or glomerulonephritis. Effective control of blood pressure arrests or retards the process, and acute hypertension-induced renal failure can be partially or completely reversed by early treatment in many cases. Hyaline degeneration is particularly observed in the afferent arterioles of the kidney in association with ageing and hypertension. Involvement of the juxtaglomerular baroceptor may account for the decline in renin secretion which is demonstrable in elderly and hypertensive populations. In malignant hypertension, glomerular hypertension and vascular necrosis produce proteinuria, haematuria, and progressive renal failure.

Atheromatous renal vascular disease much more commonly causes renal impairment in elderly hypertensive subjects than younger patients with treated mild to moderate hypertension.

Clinical features

Symptoms

Elevated blood pressure is usually asymptomatic until organ damage occurs. However, most patients labour under the illusion that any concurrent symptom is attributable to high blood pressure or its treatment. In some cases, the knowledge that a patient has high blood pressure creates a fertile soil for the growth of functional symptoms. Thus, patients who have been told that they are hypertensive have a much higher incidence of headache than hypertensive patients who are unaware of the fact. In some studies, 'labelling' a patient as hypertensive has led to an increased absenteeism from work, although no target organ damage had occurred. It is a common lay fallacy that a patient can recognize when their blood pressure is elevated, usually on the basis of such symptoms as plethoric features, palpitations, dizziness, or a feeling of tension. A screening survey in the United States examined the frequency of such symptoms as headache, epistaxis, tinnitus, dizziness, and fainting in healthy subjects. None of these symptoms was more prevalent in subjects with diastolic blood pressures over 100 mmHg.

In spite of such evidence, it should be borne in mind that target organ damage can occur and may not be clinically obvious. This may be particularly relevant in the elderly patient with diffuse cerebrovascular disease, whose fairly non-specific symptoms may be dismissed. Functional imaging has shown this to be much more frequent than was once believed (see above). Additionally, higher blood pressure levels may be responsible for some symptoms, such as headache.

Headache

The classic hypertensive headache is present on waking in the morning, situated in the occipital region of the head, radiating to the frontal area, throbbing in quality, and wears off during the course of the day. Most headaches in hypertensive patients are tension headaches not directly related to blood pressure at all. The incidence of such symptoms rises when patients become aware of the diagnosis. Nevertheless, effective treatment of hypertension reduces the incidence of headache. How far this is a specific consequence of blood pressure lowering and how far it is due to reassurance is uncertain. Morning headaches in obese hypertensive patients may be due to sleep apnoea.

Epistaxis

Whilst epistaxis is not associated with mild hypertension, it is much more common in moderate to severe hypertension. When patients present with epistaxis and high blood pressure, it is particularly important to dissociate hypertension as a cause of epistaxis from a pressor response to an alarming episode.

Nocturia

This is one of the most frequent clinically apparent consequences of blood pressure elevation resulting from reduction in urine-concentrating capacity.

Impotence

Erectile dysfunction occurs frequently in hypertension, but is usually not spontaneously mentioned by the patient or enquired about by the physician. It is often attributed by both patient and doctor to drug therapy. Although this may be the case with some classes of drug, untreated hypertension has been associated with an increased incidence of erectile dysfunction in the few studies that have specifically addressed this issue. It is probably a consequence of structural change in the peripheral vasculature limiting the capacity for acute increase in penile perfusion.

Symptoms associated with target organ damage

Cardiovascular system

Effort dyspnoea and orthopnoea suggest cardiac failure. Increased left ventricular mass is associated with decreased compliance and impaired cardiac output response to exercise. This is more likely in elderly patients whose cardiac reserves are less. Claudication suggests peripheral atheromatous vascular disease and is usually associated with atheroma elsewhere, such as the renal or carotid arteries. Angina of effort is also usually due to atheroma, although the coronary vascular tree may be free of plaques in a few cases.

Central nervous system

Scotomas suggest fundal haemorrhages or exudates, whilst blurring of vision is associated with papilloedema. These symptoms therefore deserve particular attention. Decline in cognitive performance detectable only by formal psychometric testing is more common than was once believed. It occurs particularly in long-standing untreated hypertension and in the elderly. More clinically apparent failure in concentration and memory may be due to more advanced cerebrovascular disease, depression, or centrally acting antihypertensive drugs. Extensive disease of the perforating arteries may give rise to a lacunar state characterized by progressive pseudobulbar palsy and dementia. The presence or absence of diffuse cerebrovascular disease can have important consequences for the development of dementia in Alzheimer's disease. Patients without such vascular lesions are less likely to show cognitive impairment than those with vascular lesions in the presence of the characteristic pathology of Alzheimer's disease.

Renal system

Haematuria or haematospermia suggest the malignant phase of hypertension in the absence of any other cause. Advanced renal failure in the absence of malignant hypertension suggests bilateral atheromatous renovascular or other forms of renal disease.

Clinical examination

The objectives of clinical examination are to assess blood pressure, any consequences of its elevation, and any associated disease which might modify its treatment.

Blood pressure measurement

Direct blood pressure measurement

The most accurate way to measure blood pressure is by direct arterial cannulation. Portable recording devices enable continuous arterial blood pressure measurements to be made with this technique, which antedates the modern indirect ambulatory devices. The advantages are precision, independence of arterial wall changes, the absence of any 'white coat effect' as a result of cuff inflation, and the ability to observe beat by beat variability in blood pressure and pulse. Until recently, this was not possible with indirect instruments. However, the inconvenience and morbidity caused by indwelling arterial catheters have prevented their use in routine clinical practice. The only role for direct blood pressure measurement by arterial cannulation in essential hypertension is when calcification of the arterial wall in elderly patients is suspected of causing spuriously high systolic blood pressure measurements. It is also used to monitor blood pressure in severely hypertensive patients during parenteral therapy.

Indirect manual measurement

The 'gold standard' for clinical measurement is still the mercury sphygmomanometer, using the Korotkoff sounds. However, concerns about the toxicity of mercury and the greater reliability of more recent electronic devices is now leading to their increasing usage.

The manual auscultatory technique is based upon the sounds described by Korotkoff in 1905 and uses the inflatable air-filled cuff constructed by Riva-Rocci in 1897. The brachial artery is occluded by inflating the cuff above the pressure at which the radial pulse disappears to palpation. Pressure in the cuff is estimated by a mercury or aneroid manometer. The pressure is then slowly lowered through the valve on the inflating bulb, whilst listening for the Korotkoff sounds.

Although frequently employed, auscultatory measurement of blood pressure is often carried out badly: inter- and intraobserver variability is often unacceptably high. Training videos and CDs are available for doctors and nurses and have been shown to improve measurement technique. Important points to note include the following.

Bladder and cuff

If the cuff is too small (usually as a result of an obese arm) inadequate pressure will be applied and blood pressure will be overestimated. The length of the bladder should be at least 80 per cent and the width of the bladder 40 per cent of the arm circumference. The optimal size is 26 × 12 cm for normal-sized arms and 40 × 12 cm for obese arms. Minor overlap of the ends of the bladder on thinner arms does not significantly influence readings.

Manometers

Mercury manometers should be vertical and should read zero when no pressure is applied to the cuff. Aneroid manometers require calibration every few months.

Bulb and tubing

These require checking for significant leaks (i.e. more than 1 mmHg/s) and for smooth working of the valve and inflation systems.

Technique of blood pressure measurement

The patient should be seated or supine and allowed a minimum of 2 to 3 min rest. On initial assessment or when excessive postural falls are suspected, standing pressures should be measured as well. The patient should be relaxed and the arm supported in the horizontal position with the cuff at heart level. The arm should not be constricted by tight clothing. The bladder mid-point should be placed over the brachial artery. Blood pressure should be recorded in both arms on initial examination and the arm found to have the higher pressure subsequently used. Systolic blood pressure is initially determined by palpation and then the stethoscope is lightly placed over the brachial artery and the cuff pressure raised to approximately 30 mmHg above the point at which the radial pulse disappears and then released at the rate of 2 to 3 mm/s. Both systolic and diastolic pressures should be read to the nearest 2 mm mark. The point of disappearance of sounds (Korotkoff phase V) is preferable to the point of muffling (phase IV). The reasons for this are as follows.

1. Direct arterial blood pressure measurements indicate that the phase of muffling is 5 to 10 mmHg higher than actual diastolic pressure. Korotkoff phase V is 3 to 7 mmHg higher.
2. There are fewer observer errors in identifying the disappearance of sounds.
3. Most of the epidemiological data and multicentre trials of treatment used phase V as the criterion for diastolic blood pressure.

In some clinical situations where blood flow through the brachial artery is high (immediately after exercise, in hyperthyroidism, pregnancy, and anaemia), sounds can be detected down to zero cuff pressure. Under these circumstances, the fourth phase should be recorded, with a note of the fact. Blood pressure in infants and neonates should be measured using a small cuff and Doppler ultrasound as a detection device.

Pseudohypertension in elderly people is due to an incompressible brachial artery wall. Suspicions should be raised when a high systolic pressure is associated with little in the way of target organ damage (particularly echocardiographic left ventricular hypertrophy), postural symptoms on treatment, and a firm radial artery despite cuff occlusion (Osler's phenomenon). Proof depends upon measurement of blood pressure by a method which does not depend upon arterial compression, such as arterial cannulation, finger volumetric, or automatic oscillometric devices.

Home blood pressure monitoring

Patients can usually be taught to measure their own blood pressure by electronic devices which use the Korotkoff sounds or oscillometry. A large range of cheap devices is now available. These are of varying reliability. Only those tested by specialist or consumer bodies and approved should be used. They should also be checked against manual blood pressure measurements. Home blood pressure monitoring has three advantages:

1. Blood pressure can be measured at the end of a dosing interval of an antihypertensive drug, even when this occurs in the evening or early morning.
2. 'White coat effects' as a result of measurement by a doctor or nurse are avoided, although occasional patients find measurement stressful even when carried out by themselves.
3. By encouraging participation in treatment, self-recorded blood pressures are useful both for encouraging compliance and giving reassurance.

Readings should be taken at a consistent time each day. After work in the evening is usually optimal. Blood pressures measured during stressful work are difficult to interpret and provide no useful guidance for treatment. Home blood pressure monitoring gives somewhat higher readings than 24-h ambulatory monitoring, which includes blood pressures during sleep. However, they are usually substantially lower than clinic or office blood pressures. Current evidence suggests that readings of 135/85 mmHg or higher should be considered elevated.

Indirect ambulatory blood pressure monitoring

These devices automatically inflate a cuff at set intervals during the day and night. A recording device is suspended on a belt or sling and the record subjected to computer analysis and print-out at the end of the recording period. One device ('Portapress') measures volumetric change in the finger with each heart beat and therefore offers a means to measure beat to beat variability without requiring large artery compression. At present its use is confined to research, but potentially it offers a novel approach to blood pressure evaluation.

Blood pressure falls during sleep, hence it is preferable to analyse daytime and night-time readings separately using fixed times. Alternatively, an activity meter can detect when the patient is actually asleep. The patient should keep a diary during the day so that blood pressures can be correlated with activity. Night-shift workers rapidly reverse their diurnal rhythm. A reduced or absent nocturnal fall in pressure is associated with a worse cardiovascular prognosis. Pooled analysis of population studies has yielded a figure of 138/87 mmHg for the upper 95 percentile level for daytime blood pressures. The difference between ambulatory daytime pressures and clinic blood pressures is, on average 12/7 mmHg, but individual variability is great. Suggested criteria for abnormality are shown in Table 2.

Ambulatory blood pressures have been shown repeatedly to be more closely correlated with target organ damage than clinic blood pressures. This is true of left ventricular hypertrophy, carotid wall thickness, fundal vessel changes, and microalbuminuria. There is also increasing evidence that ambulatory blood pressure monitoring is better at predicting cardiovascular events when clinic and ambulatory blood pressures diverge. However, this may partly be due to the fact that many more readings contribute to the calculated average. Multiple measurements of blood pressure have more prognostic value than a few.

Indirect ambulatory blood pressure monitoring is labour intensive and costly. Its place in routine practice therefore lies in situations where it adds information to clinic blood pressures which influences treatment. Where the decision to treat is based upon cardiovascular complications or target organ damage, or where the overall cardiovascular risk is below the level for drug treatment, there is no indication for carrying it out. On the other hand, where clinic blood pressures are sustained at high levels (such as 160/100 mmHg or higher) and treatment is not justified by high cardiovascular risk, ambulatory monitoring may be useful in confirming the presence of blood pressure elevation meriting therapy in its own right. Normal readings suggest 'white coat hypertension' under these circumstances. Other indications for ambulatory monitoring are to confirm resistance to antihypertensive medication in the absence of target organ damage, to assess 24-h control of blood pressure and episodic hypertension, and to investigate hypotensive symptoms in treated patients with no supporting evidence from clinic blood pressures.

'White coat hypertension'

Some patients have consistently elevated clinic blood pressures in the treatment range but unequivocally normal ambulatory blood pressure readings. Clearly the proportion depends critically on the criteria used. However, it is probable that about 20 per cent of patients who would be treated on the basis of high clinic blood pressures have normal ambulatory blood pressures. Almost all patients show blood pressure falls with repeated clinic measurement, so the need for sustained elevation of clinic blood pressures is critical. The prognostic significance of 'white coat hypertension' (some authorities use the term 'isolated office hypertension') is uncertain. Some of the cardiovascular changes observed in hypertension have been described in some studies and not others. There is little in the way of end-point data. Such as there is suggests that the cardiovascular risk is low compared with patients who have elevated ambulatory and clinic blood pressures. Where 'white coat hypertension' is diagnosed, the best advice is to monitor blood pressure and target organ damage and not treat unless ambulatory blood pressures become elevated.

Fundal examination

Fundal appearances provide vital information on vascular pathology and prognosis in hypertension. The Keith Wagener classification is still frequently used, although it has serious shortcomings. Chief amongst these is that grade I and II changes are produced by arterial wall thickening as a result of ageing as well as high blood

pressure. Clinically, the main requirement is to differentiate between malignant and non-malignant hypertension on fundal appearances.

Non-malignant hypertension

The earliest effects of blood pressure elevation are generalized reduction in arterial calibre with consequent reduction in arteriovenous ratio. Focal arterial narrowing is seen less often, usually when an acute rise in blood pressure has occurred. The remaining changes are frequently seen in older patients with arteriosclerosis, where their value in assessing hypertensive organ damage is small. They only have significance when seen in younger patients. The light reflex from the arterial wall is increased as a result of thickening. Nipping of the retinal veins occurs largely as a result of the optical effect of the thickened arterial walls preventing visualization of the columns of blood within the veins. Thus the veins appear to taper until they disappear before actually being crossed by the arteries. The veins may also be displaced laterally or posteriorly. Venous obstruction is much less common.

Malignant hypertension

Flame-shaped haemorrhages are superficial and owe their character to constraints imposed by nerve fibres. Dot and blot haemorrhages are deep to nerve fibres and so are not limited in the same way. Haemorrhages usually disappear after a few weeks of effective blood pressure control. There are two types of exudates. Hard or waxy exudates represent the end result of fluid leakage into the fibre layers of the retina from damaged vessels. Fluid is resorbed leaving a protein–lipid residue that is slowly removed by macrophages. Soft exudates or cotton-wool patches are aetiologically and ophthalmoscopically quite different. They are usually larger than hard exudates and have a woolly, ill-defined edge. They are not true exudates, but nerve fibre infarcts caused by hypertensive vascular occlusion. Unlike hard exudates, these lesions disappear within a few weeks of establishing adequate antihypertensive therapy.

Papilloedema is associated with raised pressure in the disc head secondary to severe vascular damage and increased permeability. Venous distention is followed by increased vascularity of the optic disc, which has a pink appearance with blurring of the disc margins and loss of the optic cup. Raising of the optic disc with anterior displacement of the vessels occurs later. The surrounding retina often shows oedema, small radial haemorrhages, and cotton-wool exudates.

The presence of haemorrhages and exudates (grade III), or papilloedema (grade IV) in essential hypertension all carry the same prognosis and the terms 'accelerated' and 'malignant' hypertension should therefore be considered synonymous.

Other physical signs

Clinical evidence of left ventricular hypertrophy and a loud aortic second sound indicate moderate or severe hypertension. Other physical signs indicate target organ damage to the cardiovascular, renal, or central nervous systems.

Investigations

Concentrations of urea, electrolytes, creatinine, and uric acid are usually normal in essential hypertension unless renal damage has occurred. Severe hypertension may be associated with elevated plasma renin and aldosterone levels, which can give rise to a modest hypokalaemic alkalosis. Serum sodium is usually low normal or low under these circumstances. This is an important differentiating point from primary aldosteronism, in which hypokalaemic alkalosis is usually associated with a high or high normal serum sodium concentration.

Microalbuminuria (20 to 200 µg/min or 30 to 300 mg/24 h) in hypertensive patients is prognostic of target organ damage. Other urinary changes, such as urinary casts, haematuria, and proteinuria, usually indicate that hypertension has entered the malignant phase or reflect primary renal disease.

Electrocardiographic left ventricular hypertrophy in the absence of any other cause indicates moderate or severe hypertension and is a valuable independent risk factor in hypertension. Echocardiography is much more sensitive in detecting early changes of left ventricular hypertrophy and provides the best independent evidence of severity in mild to moderate hypertension. Chest radiography is insufficiently sensitive as a measure of left ventricular hypertrophy. 'Unfolding of the aorta' is often observed in moderate and severe hypertension.

Management of essential hypertension

A number of national and international expert bodies have drawn up guidelines for the management of hypertension. Although these differ in detail, the overall structure of the recommendations is similar (Fig. 1). Initial clinical assessment is followed by a period of observation and monitoring depending upon the level of cardiovascular risk. During this period treatment of other cardiovascular risk factors and non-pharmacological measures are undertaken. If, at the end of this period, the patient is still at sufficient risk, antihypertensive drugs are recommended.

Fig. 1 Treatment plan for hypertension.

Assessment

The patient with essential hypertension may present in one of three ways:

1. As an asymptomatic individual whose blood pressure has been measured at routine examination for employment, insurance, or as a result of screening or preoperatively;
2. As a patient presenting with an unrelated disorder; or
3. Much less commonly, as a result of symptoms produced by hypertension or by the complications of hypertension.

Clinicians who deal with hypertension are at a great disadvantage. Whilst treatment of most symptomatic conditions leads to subjective improvement, drug treatment of hypertension may create unpleasant symptoms in an individual who was previously, to the best of their knowledge, perfectly well. It is imperative, therefore, to explain the significance of high blood pressure at the earliest opportunity. It is important to point out that in most cases it does not have a single cause. Many patients find difficulty in grasping the concept of blood pressure variability. Often they are alarmed by the inevitable occasional high reading. Discussion of the rationale for evaluation and treatment and an explanation of the nature of high blood pressure and its very high prevalence serves to reassure patients and improve compliance. Much literature is now available to help.

Establishing the diagnosis

History and examination usually provide few positive features in the uncomplicated hypertensive patient. The usually quoted range for age of onset is 35 to 55 years,

but this of course reflects the arbitrary criteria for diagnosis, and many patients who subsequently develop unequivocal hypertension have a blood pressure in the upper part of the 'normal range' below the age of 35. Moderate or severe hypertension first occurring outside the 35 to 55 age range suggests a secondary cause. The presence of hypertension in parents or siblings is of modest value in making the diagnosis. Often a negative family history simply reflects ignorance or failure to diagnose hypertension, particularly in the previous generation when health checks were less common. In addition, a positive family history can often be obtained fortuitously for a condition of very high prevalence such as essential hypertension. Positive indications of a cause for hypertension are of more value, for example a history of oestrogen-containing contraceptive pill exposure or exposure to other medications which elevate blood pressure, previous renal disease or clinical features suggestive of phaeochromocytoma, renal disease, or primary aldosteronism. Specific enquiries should be made regarding heavy alcohol intake.

Clinical evaluation

This provides essential information in the decision to treat. A history of smoking, diabetes, dyslipidaemia, or cardiovascular disease should always be sought. The presence of cardiovascular target organ damage in the form of coronary artery, cerebrovascular, or peripheral vascular disease, generalized or focal fundal arterial narrowing, or clinical left ventricular hypertrophy, all place the patient in a high-risk category. Fundal changes of malignant hypertension, left ventricular failure, encephalopathy, and hypertensive renal failure place the patient in a very high-risk category. Obesity is important as a factor in hypertension and is an independent cardiovascular risk factor.

The presence of conditions such as obstructive airways disease, diabetes, or gout may play a role in the selection of antihypertensive drug.

Investigations

When there is no clinical suspicion of secondary hypertension, extensive investigation for a primary cause is unnecessary, because the prevalence of secondary hypertension is so low. Measurement of serum urea, sodium, potassium, and creatinine, urinary microscopy, and dip-stick testing for protein are sufficient. Risk factor assessment should include an ECG, which is an important but insensitive indicator of left ventricular hypertrophy. Echocardiography is preferable as a more sensitive but expensive alternative. Where left ventricular hypertrophy is present, it offers an excellent means of excluding white coat hypertension. Fasting glucose and total cholesterol (preferably together with high-density lipoprotein cholesterol) should always be measured to define the patient's cardiovascular risk profile.

Initial advice

Except for patients who are at very high risk, or present as hypertensive emergencies, repeated measurements of blood pressure on several occasions should be carried out. The plan should be explained to the patient and advice given to improve cardiovascular risk and lower blood pressure before antihypertensive medication is considered. Cessation of smoking is probably the most important feature of advice at this stage. Other advice is directed at non-pharmacological blood pressure lowering by lifestyle modification, including dietary and behavioural measures. Dietary methods include weight reduction, sodium restriction, reduced alcohol intake in heavy drinkers, and increased fruit and vegetable intake. Behavioural methods include physical training, biofeedback, and relaxation. Successful non-pharmacological treatment may allow patients with milder degrees of hypertension to avoid drug treatment and enable lower doses of antihypertensive medication to be used in others.

Weight reduction

The epidemiological relationship between weight and blood pressure is reflected in a number of trials which have shown that dietary weight reduction produces a useful fall in blood pressure (Table 3). The mechanism for the fall in blood pressure is debated, but the most likely explanation is a fall in sympathetic efferent output to the cardiovascular system.

Sodium restriction

The average intake of sodium in Westernized cultures is 120 to180 mmol/day. Severe salt restriction (less than 10 mmol/day) produces substantial blood pressure lowering and, together with increased potassium intake, was probably responsible for the efficacy of the Kempner rice–fruit diet, used in the treatment of severe hypertension before the modern drug era. Long-term sodium restriction of this degree is not feasible and carries significant risks. More moderate sodium restriction (70 to 80 mmol/day) can be achieved by abstaining from adding salt at the table, avoiding salt in cooking, and avoiding heavily salted processed foods. As sole therapy, such moderate salt restriction produces a modest reduction in blood pressure, particularly in older subjects and black individuals (Table 3). The individual response is variable: patients showing a more substantial blood pressure fall have been classified as 'salt sensitive'. The reproducibility of 'salt sensitivity' is poor, however, and although a genetic factor has been postulated, the only way of identifying such individuals is by empirically testing the blood pressure response to salt restriction. Salt restriction enhances the blood pressure-lowering action of ACE inhibitors, angiotensin receptor blockers, b-blockers, and diuretics. Curiously, it is ineffective in patients treated with calcium antagonists.

Increased fruit and vegetable intake

Adoption of a vegetarian diet produces a modest fall in blood pressure and increased fruit and vegetable intake has been part of combined regimens of non-pharmacological blood pressure control involving reduction in weight and sodium restriction. However, a recent large trial (DASH—Dietary Approaches to Stop Hypertension) has shown that increased intake of fruit and vegetables can have an important blood pressure-lowering action, equivalent to the effect of a single antihypertensive drug. This was independent of any change in weight or salt intake (Table 3). These effects were produced by doubling the average American intake of 4.3 servings of fruit and vegetables a day. Larger effects were observed when this diet was combined with reduction in total and saturated fats. One contributory factor was probably the increase in potassium intake which occurred. Potassium supplementation has a significant blood pressure-lowering effect, partly at least through natriuresis. There is, however, no justification for potassium supplementation as an independent form of treatment unless the patient is potassium depleted. Although other features of the high fruit and vegetable diet, such as increased fibre and magnesium content, have been claimed to have a blood pressure-lowering action, these actions have not been persuasively demonstrated.

Reduced alcohol intake

The elevated blood pressures shown by heavy drinkers (more than 6 units of alcohol a day) are lowered by withdrawal. This is not related to changes in weight or electrolyte intake. This useful clinical effect has to be differentiated from the pressor response sometimes exhibited by chronic alcoholics on abstention from alcohol, which is mediated by sympathetic overactivity. The optimal intake is 2 to 3 units/day (2 units/day in women). Although some epidemiological studies have shown slightly lower blood pressures in moderate drinkers compared with total abstainers, the individual effect is likely to be small, has not been tested by intervention studies, and it would seem undesirable to advise moderate alcohol intake to teetotal hypertensive patients.

Fish oil and other dietary manoeuvres

Large increases in dietary fish oil (more than 3 g/day of omega-3 fatty acids) lower blood pressure modestly. The effect is only seen in those whose consumption of fish is low. Many patients find the ingestion of such amounts unacceptable, either in the form of oily fish or capsules. Olive oil has been shown to have a very small blood pressure-lowering action in some studies, but there is little evidence to support a therapeutic effect of other unsaturated fatty acids. Claims have also been made for calcium, garlic, protein, and vitamin C, but these are not persuasive.

Physical exercise

Regular aerobic exercise lowers blood pressure independently of any weight loss. Although the overall reduction in blood pressure is impressive, the figures quoted in Table 3 may reflect design flaws in some of the trials, since proper controls are difficult to achieve and the true effect is usually less. Moderate physical activity equivalent to 40 to 60 per cent of maximal oxygen consumption is optimal. This may take the form of, for instance, 30 to 45 min of brisk walking daily. This is also often associated with an improvement in well being attributable to endorphin release.

Other behavioural manoeuvres

Theoretically, interventions which reduce sympathetic efferent output and the alerting response should lower blood pressure. Although superficially attractive as a means of avoiding drug therapy, training in these procedures is labour intensive for professionals and carrying them out is time consuming for patients. A number of controlled trials have claimed such effects, but most have been flawed in design and there is no consistent evidence of efficacy.

Use of non-pharmacological therapy

All patients should receive relevant lifestyle advice. A diet high in fruit and vegetables, and low in salt and saturated fats, together with a recommendation of regular exercise, provide a core management strategy additional to medication, where that is necessary. Weight reduction, even in only marginally overweight patients, is probably the most efficacious manoeuvre of all. Some patients may find it feasible to increase their intake of oily fish. Patients should be advised to reduce heavy alcohol intake.

Although clinical trials have now demonstrated that these various manoeuvres can produce significant blood pressure lowering under rigorous conditions, the effects are often quite small in clinical practice, except perhaps in the case of weight reduction. Some of the dietary advice is expensive for patients (if not for drug budgets), and it is important that those asked to modify their lifestyle substantially are not disappointed by the outcome, or that the uncritical medical enthusiasm which occasionally invests this field does not imply that an aberrant life style is to blame for the development of essential hypertension.

Observation period

A period of observation extending from 1 to 2 weeks up to 1 year before antihypertensive drugs are prescribed is imperative in most patients. There are three reasons for this:

1. Multiple readings on a number of occasions provide a much better estimate of overall risk than readings on one occasion only;
2. Blood pressure usually falls owing to diminution in the alerting response as a result of repeated measurement; and
3. Non-pharmacological management may produce additional blood pressure falls.

The latter two effects may enable substantial numbers of patients with mild hypertension to avoid drug treatment. The observation period is omitted or shortened only if the patient is at high risk. The level of overall cardiovascular risk also determines whether a patient requires antihypertensive medication or not. It is therefore central to the management of essential hypertension.

Cardiovascular risk profile

The epidemiological risks associated with hypertension are not only correlated with sustained systolic and diastolic pressures, but also with:

1. Irreversible factors such as age, race, male gender, family history, and ethnicity;
2. Associated, potentially reversible factors, such as smoking, dyslipidaemia, diabetes, and obesity;
3. The presence of asymptomatic target organ damage such as left ventricular hypertrophy, atheromatous plaques on imaging, or microalbuminuria; and
4. Clinical complications such as ischaemic heart disease, heart failure, cerebrovascular disease, symptomatic peripheral arterial disease, renal disease, and malignant hypertensive fundal changes.

Higher absolute risk conferred by these other factors is translated to greater potential benefit from blood pressure lowering. The blood pressure levels at which drug treatment is initiated should therefore be lower in the presence of additional risk factors.

It is possible to calculate risk using the Framingham epidemiological data. Some guidelines, such as those issued in Britain and New Zealand, recommend this using a printed risk chart (Fig. 2) or computer program. The complexity of the data requires access either to a computer or printed table. Other guidelines—such as those issued by the American Joint National Committee (**JNC**), World Health Organization (**WHO**), and International Society of Hypertension (**ISH**)—use summary tables to assign risk to a limited number of categories (Table 4). The disadvantage of this approach is that it takes no account of differential weighting of risk factors. Age, for instance, becomes a dominant risk factor in the elderly compared with, say, dyslipidaemia.

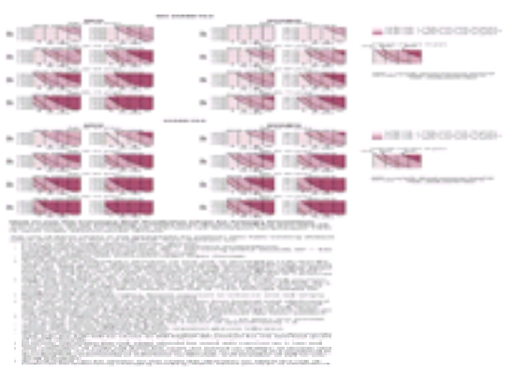

Fig. 2 Coronary risk prevention chart recommended for risk assessment in the British Hypertension Society guidelines. (Reproduced from British National Formulary (2001), **42**, with permission.)

All guidance suggests that the uncomplicated hypertensive patients with no other risk factors above a sustained threshold value for systolic and diastolic pressure should be treated with antihypertensive medication. The British Hypertension Society guidelines recommend values of 160 and 100 mmHg, respectively. The WHO–ISH thresholds are 150 and 95 mmHg, whilst the JNC thresholds are 140 and 90 mmHg.

The presence of added risks modifies both the period of observation and, in some cases, the blood pressure threshold for drug treatment. The British Hypertension Society guidelines reduce the thresholds to 140 and 90 mmHg in the presence of a 15 per cent or more 10-year risk of coronary heart disease (approximately equal to a 20 per cent 10-year risk of coronary heart disease and stroke). Patients with target organ damage, complications, or diabetes associated with hypertension reach this level without further risk-factor calculation. These higher-risk patients and those with high sustained levels (200/110 mmHg or higher) should be observed for shorter periods of 1 to 4 weeks, whilst lower-risk patients should be observed for 12 weeks according to these criteria. The WHO–ISH recommend treating medium-risk patients after observation for 3 to 6 months if blood pressures are 140/90 mmHg or higher, while low-risk patients should be observed for 6 to 12 months before treatment. The JNC risk advice suggests observation periods of up to 12 months in low-risk individuals (most of whom would not be treated on the basis of the other recommendations); other non-urgent patients are treated after two follow-up visits (Table 5).

Patients at the highest risk should be treated more promptly. Accelerated hypertension and hypertension associated with acute complications such as left ventricular failure require immediate treatment. The WHO–ISH recommends treating patients in the 'high' and 'very high' risk groups within a few days, after repeated measurement of blood pressure. The JNC guidelines imply a similar policy.

Adoption of the JNC or WHO–ISH recommendations will result in many more patients receiving drug treatment than the British guidelines. The 10-year level of cardiovascular risk in those who would be given drug treatment according to both sets of recommendations falls below 15 per cent. This is not an issue which clinical trials can resolve. It reflects a balanced judgement of the benefits of treatment set against the cost and inconvenience of life-long therapy, both to the individual and to society. The final decision depends upon the values of both. Often the language of guidelines in the treatment of hypertension has a prescriptive tone. It should not be forgotten that their only objective is to inform the professional so that appropriate options can be presented to the patient.

Treatment of hypertension

Target blood pressure

The one clinical trial which has addressed the issue of target blood pressure suggested no significant differences in outcome when blood pressure control was aimed at different strata below 140/90 mmHg, although the lowest number of events occurred around pressures of 138/83 mmHg. There was no indication of a worse outcome when pressures were reduced to lower levels than this (the so-called 'J-shaped curve'). Rigorous blood pressure control is particularly important in diabetic hypertensive patients, where pressures should be kept below 140/80 mmHg.

Impact of drug treatment

A number of important, large multicentre trials have demonstrated the impact of antihypertensive treatment on cardiovascular disease. Patients recruited have differed, and in particular some trials have been confined to elderly subjects. In addition, drug regimens and protocols have been widely disparate. Nevertheless, the conclusions have shown an impressive degree of concordance. Meta-analysis of the data has demonstrated that in these trials, for a drug-induced fall in systolic blood pressure of 12 to 13 mmHg, coronary heart disease is reduced by 21 per cent, stroke by 37 per cent, and cardiovascular mortality by 25 per cent. A fall in diastolic pressure of 5 to 6 mmHg produced figures of 16, 38, and 21 per cent, respectively. The benefits of treatment were seen in patients with isolated elevation of systolic or diastolic blood pressure. The epidemiological risk of stroke associated with these blood pressure differences was almost identical, so that, at least over the period of the trials, the risk of stroke attributable to hypertension was totally reversed. This does not imply that drug treatment abolishes the life-long risk of stroke, but it does indicate that pharmacological lowering of blood pressure reduces the short-term risks both of atherothrombotic and haemorrhagic strokes. The impact of treatment on coronary events is also significant, but probably falls short of complete reversibility. The reasons for this are still controversial.

Selection of therapy

A number of different classes of hypertensive drug are available and widely used.

Thiazide and related diuretics

Diuretics were used in several of the end-point trials which demonstrated efficacy in treating hypertension. Thiazides are ineffective in patients with a glomerular filtration rate below 40 ml/min. The more potent short-acting loop diuretics have less antihypertensive efficacy over 24 h since rebound sodium retention occurs at the end of the diuretic and natriuretic effect. They are more inconvenient for patients and should not be used except where sodium balance has to be controlled. Their major indication is to control sodium retention associated with hypertension, for example due to renal or cardiac failure. Potassium-retaining diuretics are commonly used in combination with thiazides or, in larger doses, in the treatment of primary aldosteronism. A recent study showing the beneficial effect of spironolactone on morbidity and mortality in heart failure is likely to increase use of this agent. If it, or other potassium-retaining agents, are given to patients with renal impairment, then close monitoring is required because of the danger of hyperkalaemia. This may also occur in patients who are receiving treatment which blocks the renin–angiotensin–aldosterone system. An incidental advantage of thiazides may be reduction in osteoporosis as a result of calcium retention.

Only low doses of diuretics should now be used for uncomplicated hypertension. These have a degree of patient acceptability similar to placebo in more recent studies. The dose–response curve is flat and dose titration is not required. The adverse metabolic effects on potassium balance, plasma lipids, and insulin resistance have attracted attention, since these worsen the cardiovascular risk profile. They are minimal on low-dosage regimes, even in patients with diabetes. Urate retention may precipitate gout in predisposed patients. The incidence of erectile dysfunction (which is elevated in untreated hypertension) is increased. In long-term cohort studies the risk of renal cell carcinoma has been reported as doubled, although the absolute incidence remains very low. Sodium and fluid depletion with prerenal failure is very unusual with low-dose thiazide therapy, but is much more common with loop diuretics, particularly in elderly subjects.

Dosage regimens of selected diuretics are shown in Table 6.

b-Blockers

b-Blockers have also been shown to reduce cardiovascular events in trials of the treatment of hypertension.

There are clinically important differences between the pharmacological properties of different members of the class. b_1-Selective blocking agents have less action upon bronchial and vascular b_2-adrenergic receptors, and so are less likely to cause bronchospasm or vasoconstriction in susceptible patients (for example patients with chronic obstructive airways disease or Raynaud's phenomenon). There is, however, still significant risk as tissues have mixed populations of receptors. Some b-blockers have partial agonist action at the b_1-receptor (intrinsic sympathomimetic activity) so that there is less slowing of the heart, although the heart rate response to exercise is still blunted. The clinical significance of this is uncertain. Heart rate is a risk factor for cardiac mortality and slowing might therefore be thought to be advantageous. However, bradycardia is associated with increased stroke volume and pulse pressure, which constitutes an adverse risk factor. There have been no comparative trials to address this issue in hypertension, although prevention of secondary infarction has been observed only in trials of agents without intrinsic sympathetic activity.

b-Blockers are contraindicated in hypertensive patients with asthma, chronic obstructive airways disease, heart block greater than grade 1, and sick sinus syndrome. They should be used with great caution in cardiac failure, although low-dose b-blockade with carvedilol or metoprolol has reduced morbidity and mortality in end-point trials. The combination of b-blockade with the calcium-channel blockers verapamil and diltiazem may have additional depressant actions on the sinatrial and atrioventricular node and have negative inotropic effects in patients with cardiac disease, and therefore should be avoided.

b-Blockers can have a large number of side-effects. Exercise capacity is decreased and fatiguability increased, most frequently causing problems in athletes and enthusiasts for regular physical training. They probably cause erectile dysfunction, although the incidence is not as high as that reported with diuretics where this has been recorded in trials. Non-selective b-adrenergic blockers inhibit b_2-induced vasodilatation. This may cause cold extremities and worsening of symptoms in peripheral vascular disease and Raynaud's disease. This action may also become important when there are high circulating concentrations of noradrenaline, for instance in patients with phaeochromocytoma, or after clonidine withdrawal, or in patients treated with sympathomimetic medication. Under these circumstances, non-selective b-blockade worsens hypertension. Serum triglycerides are slightly elevated and high-density lipoprotein cholesterol is reduced. This effect is not seen with drugs having intrinsic sympathomimetic activity and is less marked with cardioselective agents. b-Blockade delays the recovery from hypoglycaemia in diabetic patients and may worsen glucose intolerance. These effects are less with cardioselective drugs. b-Blockade may also mask symptoms of hypoglycaemia due to adrenaline. These are relative disadvantages in patients with diabetes and glucose intolerance that have to be balanced against benefits, for example in patients with ischaemic heart disease.

Some members of the class—such as propranolol, metoprolol, and timolol—are lipid soluble: entry into the brain appears to be associated with nightmares and sleep disturbances. These effects are unusual with water-soluble agents such as atenolol, nadolol, celiprolol, and betaxolol. Sudden discontinuation of b-adrenergic blockers in patients with cardiac disease has been associated with sudden death and it is preferable to tail off treatment in such patients over a week or two.

Pharmacological properties and dosage regimens of b-adrenergic blocking drugs are shown in Table 7.

Calcium antagonists

This class of drugs has been extensively used in treating hypertension since the 1970s. There is evidence of reduction in cardiovascular disease in end-point trials of systolic hypertension in elderly patients.

The dihydropyridine channel blockers act mainly upon vascular smooth muscle, causing vasodilatation. The non-dihydropyridines include diltiazem, a benzothiazepine, and verapamil, a phenylalkylamine. These cause less marked vasodilatation but have more pronounced effects upon cardiac contractility and atrioventricular conduction. Short-acting dihydropyridines (such as nifedipine) cause reflex baroreceptor-mediated tachycardia. If vasodilatation is not maintained, the baroreceptors do not reset, and so there may be a sustained increase in pulse rate, with other features of sympathetic activity such as sweating, palpitations, and

headache. These side-effects occur in a substantial number of patients. Sustained release preparations, or dihydropyridines with prolonged action (such as amlodipine), do not usually produce these effects. Flushing and ankle oedema occur frequently with all dihydropyridines. Oedema reflects local increase in pressure distal to dilated precapillary resistance vessels causing transudation of fluid into the tissues.

Considerable controversy was generated in 1995 by a case–control study and a meta-analysis of clinical trials which suggested that calcium-channel blockade in hypertension was associated with an increased incidence of coronary events. Although this finding was hotly disputed, it does seem likely that the first-generation, short-acting dihydropyridines were producing an adverse outcome through repeated sympathetic activation. For this reason they are best avoided in treating hypertension. Other associations with increased incidence of cancer have not been confirmed, and an association with gastrointestinal haemorrhage is still debated.

Verapamil may precipitate cardiac failure in predisposed patients and exacerbate conduction disorders. It should therefore be avoided in patients with sick sinus syndrome and atrioventricular block. Although these dangers are less with diltiazem, it is also best avoided in such patients. An action upon colonic smooth muscle causes constipation with verapamil: this is usually the only adverse effect that interferes with the quality of life with this drug. This side-effect is less commonly seen with diltiazem, although headache is more frequent.

Dosage regimens of calcium-channel blockers are shown in Table 8.

Angiotensin-converting enzyme (ACE) inhibitors

These agents have been used since the late 1970s in treating hypertension. Only one rather inconclusive end-point trial in hypertension has so far been reported, but there has been impressive evidence for reduction of morbidity and mortality in cardiac failure. Their major advantage is that they are very well tolerated with the exception of dry, unproductive cough, observed in about 20 per cent of patients. Since this does not occur with angiotensin-receptor blockers, it is probably due to bradykinin potentiation. Where patients find it unacceptable, there is no alternative to discontinuing ACE inhibitors and substituting another class of drug.

There are two serious side-effects, which are uncommon and avoidable in most cases. First, angiotensin II maintains renal glomerular hydrostatic pressure and filtration fraction by efferent arteriolar constriction. Blockade may therefore reduce glomerular filtration rate, although renal blood flow is increased. This is clinically important when renal blood flow is critically reduced, as for instance in bilateral renal artery stenosis or stenosis of the artery supplying a single kidney. It may also occur when renal perfusion is reduced through salt and water depletion, congestive cardiac failure, or renal microvascular disease. In all these situations, ACE inhibitors may precipitate acute renal failure. When glomerular filtration rate is reduced in the ischaemic kidney in patients with unilateral renal artery stenosis, the effect may not be diagnosed without imaging studies since routine tests of renal function may not be affected. Second, administration of ACE inhibitors to patients with high circulating renin levels due to salt and water depletion may precipitate severe hypotension. Where there is this clinical possibility, a small dose, preferably of a short duration agent such as captopril (6.25 mg), should be administered under supervision.

A small elevation of plasma potassium consistently occurs as a result of reduction in aldosterone secretion. This may be clinically important in those with renal failure. For this reason, ACE inhibitors should not normally be combined with potassium-sparing diuretics. Angioneurotic oedema has occurred rarely.

Dosage regimens of ACE inhibitors are shown in Table 9.

Angiotensin II receptor antagonists

The newest major class of drugs used in treating hypertension are the angiotensin II subtype 1 receptor (AT(1)) antagonists. No hard end-point trials in hypertension have so far been reported. Experience to date indicates a very low incidence of side-effects and very good patient acceptability.

Although, like ACE inhibitors, their prime target is the renin–angiotensin system, they are different in potentially important ways. Thus, while blocking the AT(1) receptor, they stimulate the angiotensin II subtype 2 (AT(2)) and other receptors as a result of high angiotensin II levels. The role of these other receptors is uncertain, and the major actions of angiotensin receptor antagonists reflect inhibition of the 'classic' effects of the renin–angiotensin system mediated by the AT(1) receptor on the blood vessels, heart, adrenal gland, kidneys, brain, and sympathetic nervous system. The actions are in this respect identical with those of ACE inhibitors. The same adverse effects in sodium-depleted patients and in patients with critical reduction of renal blood flow are therefore seen.

Addition of an AT(1) receptor antagonist to an ACE inhibitor has, in some trials, produced additional blood pressure lowering, perhaps due to more complete blockade of the renin–angiotensin system. They have no bradykinin-potentiating activity, accounting for the absence of drug-induced cough and giving these agents an important place as substitutes for ACE inhibitors when unacceptable cough occurs. Losartan is unique in this class of drugs in having a uricosuric effect.

Dosage regimens of AT(1) receptor antagonists are shown in Table 10.

a-Adrenergic blocking drugs

These drugs have been available for over 40 years, but early members of the class were never used routinely in the treatment of hypertension because of severe postural hypotension and the development of tolerance. Prazosin was initially introduced as a direct-acting vasodilator in 1976, but was subsequently found specifically to inhibit post-synaptic a_1-receptors. Initially, the recommended dosage was too high and postural hypotension and syncope proved serious problems which retarded the acceptance of this class of drugs, although the use of lower doses and the development of longer-acting agents has largely overcome this problem. Blockade of sphincteric receptors produces improvement in symptoms in patients with benign prostatic hyperplasia. Occasionally, the sphincteric effects worsen symptoms in patients with stress incontinence. Improvement in penile blood flow is associated with some benefit in patients with erectile dysfunction. Priapism has been reported rarely. Indoramin has additional central effects, causing dry mouth, nasal congestion, and extrapyramidal syndromes. Uniquely amongst antihypertensive drugs, the a_1-antagonists produce favourable changes in plasma lipids, with a reduction in total and low-density lipoprotein cholesterol and triglycerides, and an increase in high-density lipoprotein cholesterol.

Dosage regimens of a_1-adrenoreceptor blocking drugs are shown in Table 11.

Centrally acting sympatholytic drugs

Methyldopa was originally developed in the late 1950s and for many years it was one of the mainstays of antihypertensive therapy. However, it frequently causes sedation, impaired psychomotor performance, dry mouth, and erectile dysfunction. Its unfavourable impact upon quality of life caused it to be replaced by equally effective drugs in the 1970s, although it is still used extensively in the management of hypertension of pregnancy.

The withdrawal syndrome is an occasional but potentially dangerous feature of these drugs. It is most common with clonidine, when discontinuation results in a rebound rise in catecholamines with features that may resemble phaeochromocytoma, such as severe hypertension, tachycardia, and sweating. This is exacerbated when patients are also receiving non-selective b-blockers such as propranolol, which inactivates peripheral b-vasodilator adrenoceptors. The syndrome is treated by readministering the drug and then gradually discontinuing. A combined adrenoceptor inhibitor such as labetolol can be used to control blood pressure in an emergency situation. These central side-effects are shared by guanabenz, guanfacine, and clonidine. Recently, more specific agents such as moxonidine, directed at the imidazoline receptor, have been developed. These have a lower incidence of central side-effects.

Dosage regimens of centrally acting drugs are shown in Table 12.

Direct vasodilators

Hydralazine was extensively used as part of a stepped care regimen. The main disadvantages were sympathetic activation and the development of a lupus-like syndrome, particularly in patients with the slow acetylator genotype. These disadvantages, together with the need for multiple daily dosage, have resulted in the replacement of hydralazine by other agents, except for occasional use in severe hypertension and hypertension associated with pregnancy. No end-point trials have been carried out.

Usage of the more potent vasodilator minoxidil is confined to severe, resistant hypertension because of its side-effects, which include hypertrichosis and severe fluid retention. For this reason, combination with a potent loop diuretic is almost always necessary. T-wave changes and S–T depression may occur in the early phases of treatment with minoxidil due to increased cardiac work as a result of generalized vasodilatation. On these grounds, it is preferable to combine minoxidil with a b-blocker unless contraindicated.

Dosage regimens of oral vasodilators are shown in Table 13.

Drug regimens

Indications and contraindications for specific classes of drugs are shown in Table 14.

Monotherapy on average reduces systolic pressure by 7 to 13 mmHg and diastolic pressure by 4 to 8 mmHg. There is, however, marked heterogeneity in response among individuals to particular drugs. Treatment should normally commence with a low dose of the drug selected. If an adequate response is not obtained, which certainly applies if the blood pressure remains above the level at which treatment was deemed to be indicated in the individual (Table 4 and Table 5), then a number of strategies can be pursued. Firstly, the dose of the initial drug can be titrated upwards against blood pressure, except in the case of diuretics, where a single dose is used. Secondly, a small dose of a second drug can be used either separately or as a combination tablet as a means of limiting dose-related side-effects. Thirdly, the initial drug can be stopped and another class of antihypertensive agent started in the hope of greater efficacy. One study found that a rotational policy of four agents tried sequentially increased the chance of successful control of blood pressure (defined as < 140/90 mmHg) with monotherapy from 39 per cent to 73 per cent. This study also found that the responses to ACE inhibitor (A) and b-blocker (B) were correlated, as were those to calcium-channel blocker (C) and diuretic (D), hence an 'AB/CD' rule was proposed. If the initial drug used did not produce a satisfactory response, it could be substituted with one of the drugs in the other pair of treatments, thus abbreviating the rotation for use in routine practice.

In any situation a drug that is poorly tolerated must be substituted, and when a satisfactory response cannot be produced by a single agent, another class of drug must be added. Three or even four classes of drug may be needed to control more resistant hypertension. However, some combinations of drugs, which share mechanisms of antihypertensive action, have less than additive effects and are best avoided, unless the drugs are indicated on other grounds. Such combinations include a b-blocker and ACE inhibitor or angiotensin receptor antagonist, and a calcium antagonist and diuretic. Effective combinations include:

1. diuretic and b-blocker;
2. diuretic and ACE inhibitor or angiotensin receptor antagonist;
3. dihydropyridine calcium antagonist and b-blocker;
4. a-blocker and b-blocker; and
5. a-blocker with any other class of agent as third-line treatment.

Drug selection

The strong end-point trial data, patient acceptability, and low cost make diuretics or b-blockers the preferred first-line therapy in hypertensive patients who do not have indications for other drugs. The Medical Research Council Trial of Treatment in Hypertension in the Elderly suggested that diuretics were associated with a better outcome in terms of ischaemic heart disease and they should therefore normally be the first-line therapy in elderly patients. There are powerful reasons for using other drugs in the presence of comorbidity, for instance ACE inhibitors in the patient with cardiac failure and b-blockers in the patient with angina. Diuretics or dihydropyridine calcium antagonists are recommended for isolated systolic hypertension on the basis of end-point trials, but it is probable that benefit is common to most classes of drug. The only possible exception to this are rate-limiting b-blockers and calcium antagonists, where cardiac filling and stroke volume may be increased with a consequent rise in pulse pressure.

Follow-up

It is essential that patients are monitored regularly. It is important that they understand the need for this, or default is more likely. The interval between clinic visits may be short initially, usually varying between 1 and 4 weeks. When blood pressure is controlled, it is probably not necessary to see patients more often than once every 6 months.

Other treatment

Other risk factors may need control, such as serum lipids, obesity, and glucose intolerance, all of which are more prevalent in hypertensive patients. Advice about smoking is of paramount importance, since the risks of this habit exceed those of mild hypertension in many patients. Low-dose aspirin (75 mg/day) has been shown to reduce the incidence of myocardial infarction in higher-risk patients over 50 years old and this should be offered routinely to patients who fall in this category and who do not have contraindications. In view of the increased incidence of haemorrhage, it is probably not indicated in lower-risk hypertensive patients.

Resistant hypertension

In the absence of evidence of target organ damage, 'white coat hypertension' should be excluded by 24-h ambulatory monitoring. However, in some cases, blood pressures measured in the clinic, at home, or by ambulatory monitoring may remain high despite therapy with four, or occasionally five, classes of drug in optimal dosage. Minoxidil, in combination with a diuretic and preferably a b-blocker, should be reserved for such cases, titrating the dose against blood pressure, and adjusting the dose of diuretic to control oedema. Possible explanations of resistance should be sought if such measures fail to control blood pressure. These are:

1. Secondary hypertension (e.g. renovascular or endocrine);
2. Ingestion of drugs which may raise blood pressure (e.g. non-steroidal anti-inflammatory agents);
3. Heavy alcohol intake;
4. Sodium and fluid retention as a result of inadequate diuretic therapy; and
5. Poor patient compliance.

Poor compliance is often difficult to detect in hypertensive patients. Clues are provided by an initial reluctance to take medication, absence of expected pharmacological effects (such as bradycardia with some b-blockers, oedema with minoxidil), or evidence of failure to consume tablets, as revealed by tablet counts or prescription frequency. Other forms of poor compliance, such as incorrect dosing intervals, are more frequent but only detectable by electronic pill counting. In some cases it may be necessary to admit a patient to hospital and supervise administration of treatment. Where compliance is obviously poor, a number of manoeuvres can help to improve it. The regimen should be kept as simple as possible, using once daily drugs and combination tablets. A carer needs to be involved in administering medication to those who are confused. Whenever possible, effective communication with full information and involvement of the patient in his or her treatment is essential. Nurses, pharmacists, and other health professionals can play a vital role in this process.

Hypertension in specific groups of patients

Hypertension in Afro-Caribbean patients

Hypertension is more prevalent in black Afro-Caribbean patients and carries a worse prognosis. Meticulous blood pressure control therefore assumes greater importance than normal. Black patients as a group tend to respond better to diuretics, calcium antagonists, and dietary salt restriction than white patients. ACE inhibitors, angiotensin receptor antagonists, and b-blockers are, as a rule, less effective, although there is substantial patient variability in responsiveness. Activation of the renin–angiotensin system by diuretics may restore blood pressure responsiveness to ACE inhibitors or angiotensin receptor antagonists.

Hypertension in the elderly

The elderly are a very high-risk group. Inevitably, therefore, more elderly patients will meet the criteria for antihypertensive medication than will those in younger age

groups. A number of surveys have shown that doctors consistently underestimate the risks of hypertension in the elderly and therefore under-treat. However, the elderly present some particular problems as a result of the changed physiology of ageing. Thus:

1. Arterial wall stiffness gives rise to systolic hypertension and increased pulse pressure (isolated systolic hypertension). This also causes impaired baroreflex sensitivity with increased risk of orthostatic hypotension.
2. Renal conservation of sodium and fluid in the face of depletion is impaired. Elderly patients are therefore more subject to dehydration as a result of diuretic therapy or dietary restriction.
3. Clearance of drugs and their active metabolites is decreased as a result of declining hepatic and renal function.
4. Cardiac compliance and reserve are reduced and patients are therefore much more likely to develop cardiac failure. End-point trials of hypertension treatment have consistently shown reductions in morbidity and mortality from cardiac failure.
5. Comorbidity is much more common.
6. Communication and compliance may be difficult with decline in cognitive function. Some evidence from clinical trials suggests that this decline may be retarded by antihypertensive treatment.

Despite these important considerations, there is no fundamental difference in the approach to treating hypertension in the elderly patient. As a general rule, drug regimens should be as simple as possible and dosages increased more gradually. The greatest danger results from lowering pressure too much and too rapidly. Although trial evidence is limited in the very old (i.e. those over 80), there is no reason to manage these patients any differently from those who are not as old. Biological rather than chronological age should be the deciding factor in initiating antihypertensive treatment.

Essential hypertension in children

Although secondary hypertension is more common in children than in adults, no specific cause is found for hypertension in the majority of adolescents. The criteria for drug treatment, however, have to be modified because of the lower blood pressure range. The American Joint National Committee guidelines recommend that blood pressures above the 95th percentile taking into account age, height, and sex should be considered elevated. In principle, regimens are the same as those recommended for adults, with appropriate dose adjustment.

---

*It is with regret that we report the death of Professor John Swales during the preparation of this edition of the *Oxford Textbook of Medicine*.

Further reading

Appel LJ *et al*. (1997). A clinical trial of the effects of dietary patterns on blood pressure. *New England Journal of Medicine* **336**, 1117–24. [The DASH trial of increased fruit and vegetable intake producing substantial blood pressure lowering.]

British Cardiac Society, British Hyperlipidaemia Association, British Hypertension Society (1998). Joint British recommendations on prevention of coronary heart disease in clinical practice. *Heart*, **80**(Suppl 2), S1–S29. [Contains important risk chart for multiple risk factor profiling.]

Dickerson JE *et al*.(1999). Optimisation of antihypertensive treatment by crossover rotation of four major classes. *Lancet* **353**, 2008–13.

Fagard RH (1993). Physical fitness and blood pressure. *Journal of Hypertension* **11**(Suppl 5), S47–S52. [A meta-analysis of the effects of training on blood pressure.]

Graudal NA, Galoe AM, Garred P (1998). Effects of sodium restriction on blood pressure, renin, aldosterone, catecholamines, cholesterols and triglyceride: a meta-analysis. *Journal of the American Medical Association* **279**, 1383–91. [Important meta-analysis of the effects of salt restriction in hypertensive and normotensive subjects.]

Hansson I *et al*. (1998). Effect of intensive blood pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) trial. *Lancet* **351**, 1755–62. [The only trial specifically to address the question of optimal target blood pressure for treatment.]

Joint National Committee on Detection, Evaluation and Treatment of High Blood Pressure (1997). Sixth Report (JNC VI). *Archives of Internal Medicine* **157**, 2413–46. [The latest United States guidelines.]

Kaplan NM (1998). *Clinical hypertension*, 7th edn. Williams & Wilkins, Baltimore. [One of the most comprehensive recent texts from the United States perspective, fully referenced.]

Keil U *et al.* (1998). Alcohol, blood pressure and hypertension. In: *Alcohol and cardiovascular disease*, *Novartis Foundation Symposium* **216**, pp 125–51.Wiley, Chichester. [Systematic review of alcohol and hypertension.]

Medical Research Council Working Party (1985). MRC trial of treatment of mild hypertension: principal results. *British Medical Journal* **291**, 97–104. [One of the major and most influential trials.]

Medical Research Council Working Party (1992). MRC trial of treatment of hypertension in older adults: principal results. *British Medical Journal* **304**, 405–12. [A major trial of treatment in elderly patients which compares b-blockers and diuretics.]

Packer M *et al*. (2001). Effect of carvedilol on survival in severe chronic heart failure. *New England Journal of Medicine* **344**, 1651–8.

Peto R, Collins R (1994). Anti-hypertensive drug therapy: effects on stroke and coronary heart disease. In: Swales JD, ed. *Textbook of hypertension*, pp 1156–64. Blackwells, Oxford. [Important meta-analysis of the effects of antihypertensive medication in the large end-point trials.]

Pitt B *et al*. (1999). The effect of spironolactone on morbidity and mortality in patients with severe heart failure. Randomized Aldactone Evaulation Study Investigators. *New England Journal of Medicine* **341**, 709–17.

Psaty BM *et al*. (1997). Health outcomes associated with anti-hypertensive therapies used as first line agents: a systematic review and meta-analysis. *Journal of the American Medical Association* **277**, 739–45. [Comparison of different medications used in the large end-point trials.]

Ramsay LE *et al*. (1999). Guidelines for management of hypertension: report of the third working party of the British Hypertension Society. *Journal of Human Hypertension* **13**, 569–92. [The most recent British guidelines.]

Ramsay LE *et al*. (1999). British Hypertension Society guidelines for hypertension management 1999: summary. *British Medical Journal* **319**, 630–5.

SHEP Cooperative Research Group (1991). Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. Final results of the Systolic Hypertension in the Elderly Program (SHEP). *Journal of the American Medical Association* **265**, 3255–64. [An important trial showing the benefits of treating isolated systolic hypertension.]

Staessen J, Fagard R, Amery A (1988). The relationship between body weight and blood pressure. *Journal of Human Hypertension* **2**, 207–17. [Meta-analysis of weight reduction trials and blood pressure.]

Swales JD (2000). *Manual of hypertension*. Blackwells, Oxford. [British textbook with emphasis on clinical aspects of hypertension.]

Vasan RS *et al*. (2001). Assessment of frequency of progression to hypertension in non-hypertensive participants of the Framlingham Heart Study: a cohort study. *Lancet* **358**, 1682–6.

World Health Organization–International Society of Hypertension (1999). Guidelines for the management of hypertension. *Journal of Hypertension* **17**, 151–83. [The only international guidelines. Comprehensive and with a good literature review.]

15.16.2.1 Hypertension—indications for investigation

Lawrence E. Ramsay

Introduction

Sustained hypertension is very common in the general population, with a prevalence of 10 to 20 per cent of adults depending on the definition used. Most of these people will have their hypertension managed entirely in primary care, without referral for specialist investigation or treatment. About 90 per cent have idiopathic or essential hypertension, meaning that no specific cause can be identified. Perhaps 5 per cent of hypertension may be caused by drug therapy (Table 1), particularly non-steroidal anti-inflammatory drugs or oestrogen-containing oral contraceptive preparations; a further 5 per cent may be caused by renal or renovascular disease; while phaeochromocytoma, Conn's and Cushing's syndromes, coarctation, acromegaly, and other even rarer conditions together account for less than 1 per cent of all hypertension.

Note that an identifiable cause does not equate with 'curable' hypertension. In many cases, for example those with bilateral parenchymal renal disease, the cause cannot be rectified. Even when the underlying cause can be corrected, hypertension persists in about 30 per cent of cases regardless of whether the original aetiology was renal, renovascular, or endocrine. Thus, curable hypertension is very uncommon, so much so that routine extensive investigation of all patients with hypertension is unjustifiable. National and international guidelines for hypertension management agree that the investigations for all patients with hypertension should be limited in number and simple, so that they can be done readily in primary care. Detailed investigation should be reserved for patients who have specific indications in the clinical evaluation, and these patients should generally be referred for specialist opinion. This policy for investigation is logical but often not followed: a proportion of patients with hypertension do not receive even very basic tests such as creatinine and electrolyte measurement, while in hospital practice investigations are often done that are unnecessary, costly, inconvenient, and even invasive or potentially harmful.

Clinical evaluation

One consequence of reserving detailed investigation for selected patients is that the detection of uncommon but important cases of curable hypertension relies heavily on thoroughness and clinical acumen in the initial clinical evaluation. The aims of clinical evaluation are to elicit and document:

1. causes of hypertension, such as renal disease and endocrine causes;
2. contributory factors, such as obesity, high salt intake, and excess alcohol;
3. complications of hypertension, such as previous stroke and left ventricular hypertrophy;
4. cardiovascular risk factors, such as smoking, family history, sex, and age; and
5. contraindications to specific drugs, such as asthma (b-blockers) and gout (thiazides).

Those aspects of the clinical evaluation pertinent to detecting possible causes of hypertension are described in more detail in Table 2.

Routine investigations

There is agreement in recent guidelines that routine investigation should be limited to:

- urine strip test for protein and blood
- serum creatinine and electrolytes
- blood glucose
- serum total:HDL cholesterol
- electrocardiogram.

Some guidelines, such as the World Health Organization/International Society of Hypertension and the United States JNC VI guidelines, add 'optional' investigations to those above, implying that these are at the discretion of individual doctors. The British Hypertension Society guidelines do not endorse 'optional' investigations, because any additional investigations performed should be justifiable by evidence that there is some useful influence on clinical management or outcome.

Note that only two of these routine investigations are aimed primarily at detecting underlying causes for hypertension, namely urinalysis (renal causes) and creatinine and electrolytes (for renal causes and for mineralocorticoid excess such as Conn's syndrome). The other routine tests, namely glucose, total:HDL cholesterol, and electrocardiogram for left ventricular hypertrophy, are performed to assess cardiovascular risk, and are combined with other major risk factors (age, sex, smoking, and family history) to estimate cardiovascular or coronary risk formally using a chart, table, or computer program based on the Framingham risk function. Formal cardiovascular or coronary risk assessment is central to decisions on antihypertensive treatment for people with mild hypertension who have no cardiovascular complications, and also central to decisions on aspirin or lipid-lowering drug therapy.

Indications for further investigation

Common indications for more detailed investigation in hypertension are:

1. any evidence of an underlying cause in the history or examination (Table 2);
2. proteinuria, haematuria, or elevated serum creatinine;
3. hypokalaemia not caused by diuretics;
4. accelerated (malignant) hypertension;
5. documented recent onset or recent worsening of hypertension;
6. resistant hypertension (uncontrolled by a regimen of three antihypertensive drugs); or
7. young age, meaning any hypertension in patients less than 20 years old, or hypertension needing treatment in patients aged 20 to 35 years.

In those with accelerated hypertension, a recent onset or worsening of hypertension, or resistant hypertension, the prevalence of secondary hypertension is about 25 per cent, and all should have screening tests for phaeochromocytoma, renal disease, and renovascular disease. Investigation is readily justified in these patients because their hypertension is more difficult to manage, and they often have an impaired prognosis. The chance of curing hypertension is relatively small, but underlying conditions such as phaeochromocytoma, obstructive uropathy, or some parenchymal renal diseases may require treatment in their own right. The 'enrichment' of the patient population that is investigated so as to yield positive findings in about 25 per cent is important for the performance of the screening tests widely used. Tests to screen for renovascular disease and phaeochromocytoma are imperfect, meaning that their sensitivity and specificity are less than 100 per cent. Sensitivity and specificity do not depend on the prevalence of the abnormality sought, but the predictive values of the tests are highly dependent on the underlying prevalence. Screening tests that are valuable in this selected 'enriched' population of patients are often misleading if used to screen all patients with hypertension

indiscriminately.

Hypertension at a young age is generally accepted as an indication for detailed investigation, but note that the yield of underlying causes for hypertension and curable hypertension is disappointingly small when young age is the only indication for investigation. The main justification for more aggressive investigation in young patients is that detection of curable hypertension is more valuable when the alternative is many decades of antihypertensive treatment.

Positive screening investigations will often prompt more definitive tests. These are summarized in Table 3, and discussed in more detail in the chapters dealing with renal and renovascular hypertension (Chapter 15.16.2.2), phaeochromocytoma (Chapter 15.16.2.4), Conn's syndrome (Chapter 15.16.2.3), Cushing's syndrome (Chapter 12.2), and coarctation of the aorta (Chapter 15.16.2.5).

Individual investigations

Renal and renovascular disease

Investigations for renal and renovascular disease are discussed in more detail elsewhere (Chapter 15.16.2.2), but some difficulties surrounding this topic are mentioned here. Policies for renal investigation in hypertension are not uniform, and indeed few institutions or even individual clinicians seem to agree as regards the use of (for example) renal ultrasound, intravenous urogram, isotope renogram (with or without captopril), intravenous digital subtraction angiography, renal artery Doppler, magnetic resonance angiography, or renal arteriography. This unfortunate situation has arisen because those who evaluate different diagnostic methods are not always aware of, or do not address, the diagnostic problem facing clinicians. Investigation of appropriately selected patients with hypertension turns up a wide range of renal problems, among which the most common are renovascular disease, renal scarring (previously called chronic pyelonephritis), and obstructive uropathy. The clinical evaluation does not usually indicate which of these renal pathologies is present, although there are exceptions to this. For example a young patient with hypertension and a family history of polycystic kidney disease is likely to have polycystic kidneys, and the investigation of choice is clearly ultrasound. However, patients with resistant or accelerated hypertension may have any of the renal abnormalities mentioned above. The clinician does not want a test that is specific and highly accurate for renovascular disease, or for scarring, or for obstructive uropathy, but rather needs a screening test or sequence of tests that is general purpose, that is, capable of detecting all the renal abnormalities commonly found in hypertension. Used singly some tests are clearly unsuitable for this purpose, for example the isotope renogram. The best imaging policy now is renal ultrasound, followed by magnetic resonance angiography with gadolinium enhancement where this is available; when magnetic resonance angiography is not available, ultrasound followed by arteriography; and where resources are limited the rapid-sequence intravenous urogram remains a valuable general purpose investigation.

Routine measurement of creatinine clearance is not useful, but is indicated when serum creatinine is elevated and there is uncertainty whether this is related to large muscle mass or a renal abnormality. Measurement of microalbuminuria has no proven value in non-diabetic patients.

Tests for primary aldosteronism

Measurement of the aldosterone:renin ratio to screen for primary aldosteronism (Conn's syndrome) is usually triggered by finding hypokalaemia. However, patients with primary aldosteronism may have intermittent hypokalaemia or even persistent normokalaemia, and primary aldosteronism may be more common than is generally believed. The question arises whether the aldosterone:renin ratio should be measured more often, or even routinely, even without hypokalaemia. From a practical point of view the priority is to detect primary aldosteronism caused by an aldosterone-secreting adrenal adenoma, because surgical removal of the tumour may cure the hypertension and biochemical disturbance. Patients with adrenal hyperplasia causing aldosterone excess are managed medically, and have no prospect of cure of hypertension. The severity of biochemical disturbance relates to the adrenal pathology, so that patients with adrenal adenomas have more marked hypokalaemia. From a pragmatic point of view investigation only of those patients who have hypokalaemia is likely to detect those with surgically curable Conn's syndrome. However, there may be a case for measuring the aldosterone:renin ratio in patients with resistant hypertension, even when hypokalaemia is absent, or for trying the effect of spironolactone in such patients. These issues are discussed in more detail elsewhere (Chapter 15.16.2.3).

Other blood tests

Routine measurement of full blood count is sometimes advocated because of the link between hypertension and polycythaemia, but its value is doubtful. The relation between primary hyperparathyroidism and hypertension might suggest that serum calcium should be measured routinely, but treatment of the primary hyperparathyroidism does not influence the blood pressure or cardiovascular risk. Serum uric acid is often elevated in untreated hypertension, and related to male sex, obesity, alcohol use, and renal impairment. Might routine measurement allow avoidance of diuretic treatment and reduce the risk of precipitating gout? In fact there is little relation between pretreatment uric acid and the risk of gout, and most hyperuricaemic patients do not develop gout on diuretics. Many patients would be denied these valuable drugs unnecessarily with this policy, and routine measurement of serum uric acid is not recommended. Fasting serum triglycerides make no important contribution to cardiovascular risk assessment provided HDL-cholesterol is measured, and need not be measured routinely.

Ambulatory blood pressure measurement

Ambulatory blood pressure measurement should not be used routinely or indiscriminately in the management of hypertension according to current guidelines. Specific indications for ambulatory blood pressure measurement (**ABPM**) are:

1. extreme variability of blood pressure at different visits or in different situations;
2. symptoms suggesting hypotensive episodes;
3. hypertension resistant to a three-drug regimen; or
4. sustained clinic, surgery, or office hypertension in people otherwise at low cardiovascular risk.

The last category is most important because it concerns the phenomenon termed white-coat hypertension (or isolated clinic hypertension). It is not necessary or feasible to perform ABPM to exclude white-coat hypertension in all patients with hypertension. It is not indicated in those who are at high cardiovascular or coronary risk, including patients who already have target organ damage or cardiovascular complications, and those who have an estimated coronary risk of 15 per cent or higher over 10 years. In these patients treatment decisions should be based on clinic pressures rather than ABPM, as was the case in outcome trials of hypertension treatment. ABPM is also unnecessary in patients with mild hypertension (140 to 159/90 to 99 mmHg) who have no target organ damage, no cardiovascular complications, and an estimated 10-year coronary risk of less than 15 per cent. These patients can be left untreated without using ABPM, but should be followed up. ABPM is indicated when the average clinic blood pressure is 160/100 mmHg or higher, but there is no target organ damage or cardiovascular complication, and the estimated 10-year coronary risk is less than 15 per cent. Here elevated blood pressure is the only indication of high cardiovascular risk and for antihypertensive treatment, and normal blood pressure by ABPM may alter the treatment decision. However, any decision to withhold treatment in such patients should be based on appropriately adjusted normal values for ambulatory measurement, and should be confirmed by a second ABPM record because of within-patient variability and limited reproducibility.

Echocardiography

Echocardiography is more 'sensitive' than the electrocardiogram for detecting left ventricular hypertrophy, but this does not mean that it is better or useful. Left ventricular hypertrophy is used to estimate cardiovascular risk, and the relevant question is whether echocardiography is superior to the electrocardiogram for estimating cardiovascular risk. In fact it does enhance the accuracy of risk estimation, but only very slightly, and the gain does not justify routine echocardiography. It is indicated in patients who have 'voltage criteria' for left ventricular hypertrophy, but no T-wave abnormalities. Voltage criteria alone are very unreliable, particularly in young men, and should not be used to diagnose left ventricular hypertrophy without confirmation by echocardiography.

Further reading

British Cardiac Society, British Hyperlipidaemia Association, British Hypertension Society, endorsed by the British Diabetic Association (1998). Joint British Recommendations on prevention of coronary heart disease in clinical practice. *Heart* **80**(Suppl 2), S1–S29. [British guidelines including policy for investigation and method of coronary risk assessment.]

Cameron HA *et al.* (1992). Investigation of selected patients with hypertension by the rapid-sequence intravenous urogram. *Lancet* **339**, 658–61. [Detailed outcome of investigating selected patients with hypertension related to different indications for investigation.]

Guidelines Subcommittee (1999). 1999 World Health Organization–International Society of Hypertension guidelines for the management of hypertension. *Journal of Hypertension* **17**, 151–83. [International guidelines including policy for investigation.]

Haq IU *et al.* (1995). Resistant hypertension. In: Kendall MJ, Kaplan NM, Horton RC, eds. *Difficult hypertension*, pp 97–115. Martin Dunitz, London. [Review of clinical assessment, investigation, and treatment of resistant hypertension.]

Joint National Committee (1997). The sixth report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. *Archives of Internal Medicine* **157**, 2413–46. [United States guidelines including policy for investigation.]

Lever AF, Swales JD (1994). Investigating the hypertensive patient: an overview. In: Swales JD, ed. *Textbook of hypertension*, pp 1026–30. Blackwell, Oxford. [Rationale for policy of selective investigation in hypertension.]

Ramsay LE *et al.* (1999). Guidelines for management of hypertension: report of the third working party of the British Hypertension Society. *Journal of Human Hypertension* **13**, 569–92. [British guidelines including policy for investigation.]

Wallace EJ *et al.* (2000). Coronary and cardiovascular risk estimation for primary prevention: validation of a new Sheffield table in the 1995 Scottish health survey population. *British Medical Journal* **320**, 671–6. [Principles, practice, and accuracy of cardiovascular risk assessment in mild hypertension.]

15.16.2.2Renal and renovascular hypertension

Lawrence E. Ramsay

Introduction

Renal or renovascular abnormalities are present in about 5 per cent of all hypertensives, and 25 per cent of those selected appropriately for detailed investigation (see Chapter 15.16.2.1). Some renal lesions are incidental to hypertension, some cause the hypertension but are uncorrectable, some cause the hypertension and are correctable, but without cure of the hypertension, and finally a few cause the hypertension, and are correctable with cure of hypertension. Unfortunately this last category, curable hypertension, is very uncommon. Fibromuscular renal artery stenosis is the only form of renal hypertension that is 'usually' curable, meaning that more than 50 per cent of patients are cured.

Investigation of hypertension often uncovers renal diseases that need management or monitoring in their own right, such as polycystic kidneys, glomerulonephritis, chronic pyelonephritis, or obstructive uropathy. However, policies for investigation and particularly for intervention should be tempered by the knowledge that renal or renovascular hypertension can rarely be cured. Intervention should generally be reserved for patients with a compelling indication, for example severe and resistant hypertension, declining renal function, or 'flash' pulmonary oedema. The outcome is generally best with younger age (under 60 years), normal renal function, and a short history of hypertension.

Causes and mechanisms

Renal abnormalities that may be found in hypertensive patients are shown in Table 1, but note that the relation between hypertension and the renal lesion differs among these. In bilateral parenchymal disorders such as glomerulonephritis, interstitial nephropathy, or polycystic kidneys the prevalence of hypertension in the early stages is about 30 per cent, although it varies with the pathology. For example, hypertension is more common in mesangiocapillary type 1 glomerulonephritis (40 per cent) than in minimal change nephropathy (16 per cent). Again, hypertension is three times more likely in the common form of polycystic kidneys associated with mutations at the *PKD1* locus than in non-*PKD1* disease. The hypertension early in these conditions is probably renin dependent. With progression to renal failure the prevalence of hypertension increases to 80 to 90 per cent, and hypertension then reflects imbalance between volume and vasoconstriction, with the emphasis on volume dependence. Volume-dependent hypertension that is usually curable also occurs in bilateral obstructive uropathy, caused for example by bladder neck obstruction.

The relation of unilateral renal and renovascular abnormalities (Table 1) to hypertension also varies for different entities. Unilateral chronic hydronephrosis is no more common in hypertensive than normotensive subjects and relief of obstruction does not cure the hypertension. It is therefore an incidental finding and should be managed entirely on its own merits. Simple renal cysts or neoplasms are also generally incidental, but very rarely they can cause curable renin-dependent hypertension, either through renin production or by compression of renal tissue leading to renin release. Unilateral chronic pyelonephritis (renal scarring, reflux nephropathy) commonly causes hypertension, yet nephrectomy rarely cures the hypertension. The reason is that the contralateral kidney is usually abnormal also, due either to scarring or to the effects of hypertension itself. Fibromuscular renal artery stenosis is the most convincing cause of reversible hypertension, because correction by angioplasty cures hypertension in at least 50 per cent of cases. The relation of atherosclerotic renal artery stenosis to hypertension is complex. Correction of the stenosis cures hypertension in fewer than 10 per cent of cases, but it is significantly 'treatment sparing'. Thus intervention sometimes improves but rarely cures hypertension. The relation of atherosclerotic renal artery stenosis to hypertension probably differs between patients: in some, renal artery stenosis may be coincidental to, or even a complication of, hypertension; in others, renal artery stenosis may have caused hypertension, which has then become irreversible because of ischaemia or atheroembolic disease (cholesterol embolism); and in a minority (less than 10 per cent) renal artery stenosis is the cause of reversible hypertension. Unfortunately, no tests can distinguish between these. Investigations such as renal vein renin or split function measurements having no useful predictive value.

Clinical evaluation

Pointers to a possible renal abnormality include documented recent onset or worsening of hypertension, onset before 35 years of age, accelerated phase hypertension, resistance to drug therapy, or renal failure precipitated by ACE inhibitor treatment. The physician should enquire about present or past urinary symptoms or loin pain, a family history of renal disease or polycystic kidneys, clues to systemic conditions that may cause renal disease, such as diabetes or vasculitis, and ingestion of drugs that may be nephrotoxic. Examine for palpable kidneys (polycystic kidneys, neoplasm) or bladder (obstructive uropathy), and auscultate the epigastrium and renal angles for a vascular bruit. Systolic bruits are common and have low specificity, whereas continuous or systolic–diastolic bruits are rare but highly suggestive of renovascular disease.

Investigation

Routine investigation for renal disease is limited to serum creatinine and glucose (to exclude diabetes), and a urine stick test for protein and blood. When clinical evaluation and these routine tests are normal, further investigation is not indicated. If they suggest renal abnormality, further investigation may include:

- microscopy and culture of midstream urine; but note that absence of proteinuria, red cells, and casts does not exclude glomerular or interstitial disease with certainty
- quantitation of proteinuria over 24 h to help distinguish between glomerular and interstital disease; proteinuria of more than 1 g/24 h also signals the need for a lower blood pressure target and ACE inhibitor treatment
- renal ultrasound.

Depending on the results of these tests, additional investigation might include tests for systemic causes of renal disease ((for example antinuclear factor (ANF), DNA antibodies, antineutrophil cytoplasmic antibodies (ANCA)), renal biopsy if glomerulonephritis or interstital nephropathy is suspected, or imaging for renovascular disease (see below). However, it is often appropriate to watch rather than investigate further at this stage. For example, a patient with severe hypertension, proteinuria less than 1 g/day, mild renal impairment, and normal renal ultrasound most likely has renal damage caused by previous accelerated hypertension. It is entirely reasonable to control the hypertension, monitor the renal function closely, and investigate further only if the renal function declines.

Unilateral renal disease

Unilateral renal abnormalities should generally be treated on their own merits because nephrectomy rarely cures hypertension. However, lesions such as renal cysts or radiation nephropathy very rarely do cause curable hypertension, and patients therefore have to be considered individually. Nephrectomy should be considered only when:

- hypertension is severe and difficult to control
- hypertension is of recent onset
- the affected kidney has no or very little function

- the contralateral kidney is completely normal on detailed investigation
- serum creatinine is normal
- the patient is young and generally fit
- the patient is willing to accept a small chance of cure or improvement from operation.

In practice these criteria are rarely satisfied, and hypertension is managed medically in almost all patients with unilateral renal disease. Lateralizing tests such as renal vein renin measurements have no useful predictive value in this situation.

Renovascular disease

Aetiology

Only 20 per cent of patients with renovascular disease have fibromuscular dysplasia, but it is important that cases are recognized because patients are often young and hypertension is often curable. The most common dysplastic pathology is medial fibroplasia, in about 70 per cent (Fig. 1), with stenotic lesions that are generally distal, multifocal, rarely cause occlusion, and may affect other vessels such as the carotid or mesenteric arteries. Other forms of medial dysplasia, and perimedial fibroplasia or adventitial fibroplasia, are uncommon. Some of these are unifocal, proximal, and can cause occlusion, and are therefore readily mistaken for atherosclerotic renovascular disease. Fibromuscular dysplasia is five times more common in women than men, bilateral in about 25 per cent of cases, and much more common on the right.

Fig. 1 Selective right renal angiography in a patient with hypertension caused by fibromuscular dysplasia of the common medial fibroplasia type. Note that the lesions involve the distal part of the renal artery, are multifocal, and show the 'string of beads' appearance of alternating stenoses and poststenotic dilatations.

About 80 per cent of renovascular disease is atherosclerotic (Fig. 2). This is strongly associated with vascular disease elsewhere, particularly peripheral vascular disease, and with major risk factors for atherosclerosis including male sex, old age, diabetes, hyperlipidaemia, hypertension itself, and, particularly, cigarette smoking. Atherosclerotic stenotic lesions are usually proximal, often at the ostium of the renal artery, and bilateral in 25 per cent of cases. Atherosclerotic renovascular disease is progressive, leading to arterial occlusion in about 2 per cent of patients per year, and progression to high-grade stenosis in about 10 per cent of patients per year. Loss of about 70 per cent of the artery lumen is necessary for haemodynamic significance, but radiological assessment of the degree of stenosis is imprecise.

Fig. 2 Aortography in a 65-year-old man with severe atherosclerotic renovascular disease. There is total occlusion of the right renal artery, tight stenosis of the left renal artery (arrow), and extensive aortic atheroma.

Less common causes of renovascular disease include neurofibromatosis, transplant renal artery stenosis, aortic or renal artery dissection, embolism, Takayasu's arteritis, arteriovenous fistula, and radiation.

Diagnosis

Patients who have treatment-resistant hypertension or a decline in renal function with ACE inhibitor treatment should be assessed by a clinical prediction method that uses nine simple variables and estimates the probability that renovascular disease is present (Table 2, Fig. 3). This is at least as accurate as many of the non-invasive screening tests in wide use, such as renin measurements or isotope renography with or without an ACE inhibitor. The probability estimate should be considered together with the clinical circumstances to decide on further investigation. For example, a young patient with severe hypertension that is difficult to control, and with side-effects from drugs, would certainly be considered for investigation even if the probability of renovascular disease was only 10 per cent. By contrast, an elderly patient who was well controlled by three drugs, entirely comfortable on treatment, and had normal renal function, would be managed conservatively even if the probability of renovascular disease was much higher.

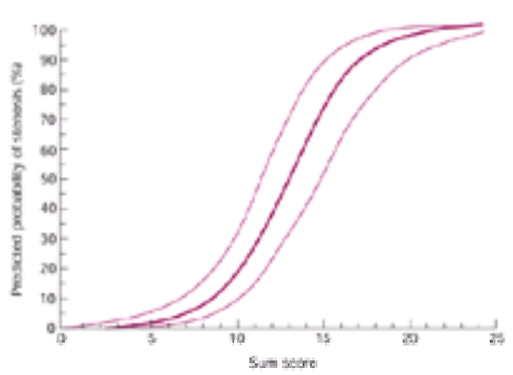

Fig. 3 Predicted probability of renal artery stenosis in patients with drug-resistant hypertension as a function of the sum score. The sum score is derived from the prediction rule in Table 2. The thin lines represent 95 per cent confidence intervals. From Krijnen *et al*. (1998) with permission.

When the probability of renovascular disease and clinical circumstances warrant further investigation, renal ultrasound is done first to exclude other abnormalities. The non-invasive investigation of choice then is magnetic resonance angiography of the renal arteries with gadolinium enhancement. Spiral computed tomography angiography is equally accurate but needs contrast medium and ionizing radiation. Doppler ultrasound examination of the renal arteries is also accurate, but highly operator dependent. Intra-arterial digital subtraction angiography is the gold standard, but is invasive. Magnetic resonance angiography, spiral computed tomography, or Doppler are all valuable methods when they are available: when they are not, ultrasound should be followed by intra-arterial digital subtraction angiography.

Management

Interventional

Stenosis caused by fibromuscular renovascular disease can be corrected completely by angioplasty in 90 per cent of cases, and does not usually require stent insertion. Hypertension is cured completely in around 50 per cent of patients and complications are uncommon. Patients who are not suitable for angioplasty should be considered for surgical bypass or reconstruction in a centre experienced in these techniques.

Atherosclerotic renovascular disease often cannot be corrected completely by angioplasty alone, but stent insertion has greatly increased the technical success and long-term patency rates. However, stenting has not improved the disappointing outcome, with hypertension cured in less than 10 per cent of patients. Complications of the procedure are common and sometimes serious. The effect of angioplasty with stent insertion on renal function in atherosclerotic renovascular disease is unclear, and currently being studied in controlled trials. Limited evidence suggests that renal function improves in one-third, declines in one-third, and remains stable in one-third, and that angioplasty with stent insertion may slow the decline in renal function in some cases. Many patients with atherosclerotic renovascular disease have a very limited prognosis because of their widespread vascular disease. Given these considerations, angioplasty with stent insertion should be considered for atherosclerotic renovascular disease in the following circumstances:

- Hypertension that is severe and uncontrollable by several drugs in combination, including high doses of a loop diuretic.
- 'Flash' pulmonary oedema—patients with critical renovascular disease, meaning severe bilateral stenosis or tight stenosis to a single functioning kidney, may develop fulminant pulmonary oedema even when left ventricular function is normal or near normal. This condition can respond dramatically to correction of atherosclerotic renovascular disease.
- Progressing renal failure in a patient with bilateral renovascular disease, despite adequate blood pressure control, and with no other cause. Intervention hoping to prevent progression is justified pending the outcome of controlled trials in progress.

Surgical correction of renovascular disease should be considered for these same indications when angioplasty with stent insertion is technically impossible, although morbidity and mortality are daunting because of extensive cardiovascular disease.

The combination of hypertension, renal impairment, and one small kidney is a common presentation in elderly patients, and is caused by renal artery thrombosis with stenosis in the contralateral renal artery. This is usually best managed medically. However, in appropriate cases revascularization of a kidney with renal artery thrombosis can restore renal function, and should be considered if the kidney length is more than about 8 cm, the thrombosis is fairly recent, and renal biopsy shows no irreversible fibrosis or glomerular loss.

Medical

Patients with hypertension and atherosclerotic renovascular disease or renal failure usually have severe hypertension that is resistant to drug therapy. As a rule they will need several antihypertensive drugs in combination. Thiazide diuretics are ineffective or insufficiently effective in patients with resistant hypertension or renal impairment, and a loop diuretic is needed, often at high dosage. ACE inhibitors are a two-edged sword in patients with hypertension and renal disease. In critical renovascular disease, defined as severe bilateral renal artery stenosis or tight stenosis to a single functioning kidney, glomerular filtration is entirely dependent on increased efferent arteriolar tone, which is maintained by angiotensin II. Treatment with ACE inhibitors or angiotensin II antagonists abolishes the increased efferent arteriolar tone, stops glomerular filtration entirely, and causes acute renal failure. ACE inhibitors and angiotensin II antagonists should be avoided, or used with extreme caution and close monitoring of renal function, in patients known to have renovascular disease or who may have renovascular disease, for example those with peripheral vascular disease or unexplained renal impairment. On the other hand, ACE inhibitors are renoprotective and positively indicated in patients who have hypertension, renal impairment, and proteinuria over 1 g/day.

Hypertensive patients who have atherosclerotic renovascular disease or renal impairment are at very high cardiovascular and coronary risk. In addition to good blood pressure control they often need treatment with low-dose aspirin, and with a statin if the serum cholesterol is equal to or more than 5 mmol/litre.

Further reading

Aitchison F, Page A (1999). Diagnostic imaging of renal artery stenosis. *Journal of Human Hypertension* **13**, 595–603. [Review of non-invasive methods for diagnosing renovascular disease.]

Cameron HA *et al*. (1992). Investigation of selected patients with hypertension by the rapid-sequence intravenous urogram. *Lancet* **339**, 658–61. [Renal abnormalities in consecutive hypertensive patients investigated appropriately related to indication for investigation.]

Caps MT *et al*. (1998). [Prospective study of atherosclerotic disease progression in the renal artery. *Circulation* **98**, 2866–72. Natural history of atherosclerotic renovascular disease.]

Harden PN *et al*. (1997). Effect of renal-artery stenting on progression of renovascular renal failure. *Lancet* **349**, 1133–6. [Effect of angioplasty and stent insertion on renal function in atherosclerotic renovascular disease.]

van Jaarsveld BC *et al*. (2000). The effect of balloon angioplasty on hypertension in atherosclerotic renal-artery stenosis. *New England Journal of Medicine* **342**, 1007–14. [Largest and best randomized controlled trial of angioplasty for atherosclerotic renovascular disease.]

Krijnen P *et al*. (1998). A clinical prediction rule for renal artery stenosis. *Annals of Internal Medicine* **129**, 705–11. [Probability of renovascular disease predicted from simple clinical and biochemical variables.]

Pickering TG *et al*. (1988). Recurrent pulmonary oedema in hypertension due to bilateral renal artery stenosis: treatment by angioplasty or surgical revascularisation. *Lancet* **1**, 551–2. [First description of 'flash' pulmonary oedema related to critical renovascular disease.]

Ramsay LE, Waller PC (1990). Blood pressure response to percutaneous transluminal angioplasty for renovascular hypertension: an overview of published series. *British Medical Journal* **300**, 569–72. [Overview of outcome of angioplasty in fibromuscular and atherosclerotic renovascular disease.]

Robertson JIS (1992). Unilateral renal disease in hypertension. In Robertson JIS, ed. *Handbook of hypertension, Vol. 15: Clinical hypertension*, pp 266–325. Elsevier, Amsterdam. [Excellent review of unilateral renal and renovascular disease and hypertension.]

van de Ven PJG *et al*. (1999). Arterial stenting and balloon angioplasty in ostial atherosclerotic renovascular disease: a randomised trial. *Lancet* **353**, 282–6. [Controlled trial showing increased arterial patency with stent insertion, but no advantage on blood pressure control.]

Whitworth JA (1992). Renal parenchymal disease and hypertension. In Robertson JIS, ed. *Handbook of hypertension, Vol. 15: Clinical hypertension,* pp 326–56. Elsevier, Amsterdam. [Excellent review of bilateral renal disease and hypertension.]

15.16.2.3 Primary hyperaldosteronism (Conn's syndrome)

M. J. Brown

Introduction

Conn's syndrome is the eponymous term that embraces the various causes of primary aldosteronism. Although the current trend in medicine is away from eponymous nomenclature and towards names that reveal more of a disease's pathogenesis, there are good arguments for retaining the eponym when there is a need to ensure much wider recognition, and diagnosis, of the syndrome. All drugs receive two names, a generic and brand name: the former is more informative but often forgettable (or unpronounceable), encouraging the use of the memorable brand name after patent life expires. 'Patent life' on Conn's description of primary hyperaldosteronism as a cause of hypertension, in 1966, long since expired, but controversy remains regarding many aspects of the syndrome—prevalence, pathogenesis, treatment—and look-alikes continue to appear.

Within the continuum of hypertension are a number of so-called secondary syndromes, meaning that the hypertension is due to a specific, recognizable cause. The search for secondary causes is sometimes motivated by the aim of finding a curable cause, but it is better to think of the aim as finding the optimal treatment.

The three different types of Conn's syndrome embrace, and illustrate, the spectrum of hypertension. Firstly, adrenal adenoma is the only curable type, but the overlap with hyperplasia contributes to the hazards of predicting cure. Secondly, bilateral hyperplasia can be hard to differentiate from the low-renin end of the spectrum of essential hypertension and on recognition is not curable, but diagnosis is still rewarding because of the usually excellent blood pressure response to spironolactone. Thirdly, the only definite genetic cause of Conn's syndrome identified to date, glucocorticoid remediable aldosteronism, is clearly incurable, but the genetic test provides an infallible diagnosis and predicts a reversal of the hypertension by an otherwise ineffective treatment.

Physiological background

The zona glomerulosa, where aldosterone—the principal salt-retaining hormone—is synthesized, is the outermost of the three secretory zones of the adrenal cortex and the usual site of the tumours or hyperplasia in Conn's syndrome. The zonas glomerulosa and fasciculata, where cortisol is synthesized, are distinguished by their respective expression of the closely related genes, *CYP11B1* encoding 11b-hydroxylase, and *CYP11B2* encoding aldosterone synthase. The secretion of aldosterone is regulated by angiotensin II, whose concentration is determined by that of circulating renin. However, aldosterone secretion also responds in some degree to ACTH, acting both directly on zona glomerulosa cells and through release of endothelin from adrenal endothelial cells and of 5-HT from mast cells. It is these alternative stimuli that may be responsible for aldosterone-dependent hypertension in low-renin patients.

The main receptor for aldosterone, the mineralocorticoid receptor, is a nuclear hormone receptor in the distal tubules and collecting duct of the kidney. Stimulation of the receptor leads to activation of the epithelial sodium channel on the apical (luminal) surface of the tubular cells, through the action of a serine/threonine kinase called serum glucocorticoid kinase, which disappears from the renal tubules after adrenalectomy. As an apparently passive consequence of the increased apical Na^+ flux into the cells, there is also enhanced activity of the basolateral Na^+,K^+-ATPase, which pumps Na^+ into the peritubular interstitium and thereby into the blood.

Of the two main adrenal steroids, cortisol is much the more abundant, by 100- to 1000-fold. Since both steroids have a similar affinity for the mineralocorticoid receptor, it used to be a mystery why aldosterone is the physiological agonist. The explanation became clear with the discovery, in the same distribution as mineralocorticoid receptors, of the enzyme 11-hydroxysteroid dehydrogenase: this enzyme inactivates cortisol to cortisone, preventing access to the receptor. The enzyme is inhibited by liquorice, or the old antiulcer drug carbenoxolone, and is congenitally deficient in homozygotes with the rare syndrome of apparent mineralocorticoid excess (see Chapter 15.16.1.2). The enzyme is also inhibited (or, rather, saturated) by very high plasma concentrations of cortisol, such as occur in patients with the ectopic ACTH syndrome (see Chapter 12.7.1). These patients develop the clinical and biochemical features of Conn's syndrome before (or without) becoming floridly cushingoid. Indeed, the lowest levels of plasma K^+ (less than 2.5 mmol/litre) in the presence of plasma Na^+ over 145 mmol/litre should suggest the diagnosis of ectopic ACTH rather than primary hyperaldosteronism.

Incidence

This is contentious. Until recently, the figure was considered to be 1 to 2 per cent of hypertensives, but this is based on the incidence in patients with the typical electrolyte pattern described below, which is now recognized to be absent in many patients. Selected series of hospital patients have suggested figures as high as 15 per cent of hypertensives. Our own survey of plasma aldosterone to renin ratios in 800 unselected hypertensive patients suggests that the true prevalence is at least 5 per cent, most of whom do not have adenomas. The increased recognition of Conn's syndrome is mainly among patients with plasma K^+ levels within the normal range. Previous algorithms designed to distinguish adenomas from bilateral hyperplasia have emphasized the absolute level of aldosterone as a guide to adenoma: it seems likely that patients with adenomas will therefore have more instantly recognizable plasma electrolyte abnormalities, unless these have been masked by treatment with a calcium-channel blocker.

Clinical characteristics

Adenomas are said to occur more commonly in women and bilateral hyperplasia more often in men, but the differences are too slight to be helpful diagnostically. Except for the rare monogenic syndrome of glucocorticoid remediable aldosteronism, Conn's syndrome is not a cause of childhood hypertension. The main, and essential, clinical feature of Conn's syndrome is hypertension, any other clinical features being secondary to hypertension or hypokalaemia, but the majority of patients are asymptomatic.

Two important features differentiate patients with Conn's syndrome from those with secondary hyperaldosteronism, in which increased aldosterone secretion is driven by elevated levels of renin and angiotensin. First, patients with Conn's do not develop oedema; it is assumed that secretion of a natriuretic hormone, such as atrial natriuretic hormone, leads to escape from the salt-retaining effect of aldosterone (the 'escape phenomenon'). Secondly, and of diagnostic value, the plasma Na^+ concentration is within—or just above—the upper part of the normal range, usually more than 140 mmol/litre, whereas in secondary hyperaldosteronism the reduced free water clearance caused by angiotensin II results in some dilution of the plasma Na^+, whose concentration is therefore less than 140 mmol/litre. Very rarely, primary hyperaldosteronism is associated with phaeochromocytoma, primary hyperparathyroidism, or acromegaly.

Investigation

Who requires investigation?

In patients with hypertension the diagnosis of Conn's syndrome should be suspected in two main circumstances. The conventional one is the presence of hypokalaemia and high normal plasma sodium concentration. Because in general practice K^+ measurements can be unreliable if samples have stood for some hours before separation, routine measurement of bicarbonate is recommended in the initial sample from a hypertensive patient. Conn's patients typically have hypokalaemic

alkalosis because stimulation of the epithelial sodium channel by aldosterone causes exchange of Na^+ for both K^+ and H^+ ions. It is important to have increased suspicion when the plasma K^+ falls (or the bicarbonate rises) substantially on diuretic treatment: the low doses of thiazide diuretics commonly used nowadays in the treatment of hypertension do not usually lower K^+ by more than 0.5 mmol/litre.

The second circumstance under which Conn's syndrome should be suspected is when patients appear resistant to conventional antihypertensive treatment and the electrolytes are 'in the direction' of Conn's, without necessarily being outside the normal range (plasma Na^+ more than 140, K^+ less than 4.0 mmol/litre).

Establishing the diagnosis of Conn's syndrome

When the diagnosis of Conn's syndrome is suspected, it should be pursued by estimation of plasma aldosterone and renin, seeking evidence of elevated aldosterone secretion in the absence of elevated renin production. An aldosterone to renin ratio of over 850 is usually diagnostic, and should at least trigger a trial of spironolactone therapy and an adrenal scan. Application of the ratio, rather than consideration of the absolute level of aldosterone, is useful in encouraging measurement of the hormones under more everyday conditions than is recommended for either hormone alone. Thus, a patient whose aldosterone secretion is elevated by physical activity will have a similar elevation in plasma renin activity. A further important practical point is that most antihypertensive drugs can be continued, provided that the clinician is aware of some potential for interference.

Effects of antihypertensive drugs on plasma renin and aldosterone

ACE inhibitors and angiotensin receptor antagonists markedly reduce the aldosterone to renin ratio in most non-Conn's patients by interrupting the negative feedback inhibition of renin secretion by angiotensin II, but do not prevent detection of an elevated ratio in low-renin patients with autonomous aldosterone secretion (i.e. Conn's). b-Blockers elevate the ratio by suppressing renin secretion more than aldosterone; in patients whose ratio but not absolute level of aldosterone is elevated, a repeat estimation of b-blockade may be worthwhile. Calcium blockers cause variable suppression of aldosterone and renin secretion, sometimes sufficiently to mask the diagnosis of Conn's syndrome. It is not necessary to stop a calcium blocker in order to measure aldosterone except in patients whose electrolyte abnormalities have been corrected by the calcium blocker. Diuretics increase renin and aldosterone in parallel, and the measurement of the aldosterone to renin ratio therefore readily distinguishes Conn's from diuretic induced hypokalaemia.

Further investigations in patients with an elevated aldosterone to renin ratio

Once the probable diagnosis is established, the next question is whether the patient has a unilateral adenoma or a bilateral hyperplasia. The answer clearly has a major influence on the choice of long-term treatment, although it is important to remember that Conn's adenomas are always benign and that some patients will opt for long-term medication in preference to surgery. A number of algorithms have been devised to help distinguish the two conditions. Some of these are based on the relatively greater response of aldosterone secretion from adenomas and hyperplasia to ACTH and angiotensin II, respectively. Thus, there is a greater diurnal rhythm in aldosterone levels in patients with adenomas, but greater response to posture in hyperplasia. The problem with these algorithms is that they usually depend on multiple measurements, requiring admission of patients to hospital, and the result is still only a probability of one or other diagnosis, insufficiently strong to make a decision about surgery. The practicalities of outpatient investigation, coupled with the quality of modern imaging techniques, dictate a more empirical approach. The clinician wants to know the following:

- Is there an operable adenoma?
- Is the contralateral adrenal anatomically normal (i.e. no adenoma)?
- Is aldosterone secretion unilateral and from the side with the adenoma?

A good magnetic resonance or computed tomography scan will usually answer the first two questions, and where imaging reveals an adrenal mass 1–2 cm in size, the probability of adenoma is high enough to justify surgery without further troublesome investigations (Fig. 1). Above this size, the possibility of an adrenal carcinoma should be considered. Below 1 cm, the anatomical scan cannot with certainty exclude non-functioning myelolipomas ('incidentalomas') or large nodules within a hyperplastic gland. Magnetic resonance imaging has the slight edge over computed tomography on specificity, but the choice between the two scans can reasonably depend on local availability.

Fig. 1 Computed tomography scan of a 2 × 1 cm left adrenal adenoma (arrowed) in a patient with Conn's syndrome.

There are two tests for lateralization, neither of which is ideal. The better is measurement of the aldosterone to cortisol ratio in samples from the adrenal veins, with a reference sample from the vena cava. A 'perfect' result is a ratio in blood from the adenoma that is ten times greater than the reference sample, with the suppressed contralateral adrenal having a ratio lower than reference. However, suppression is not always demonstrable, and a serious problem with the test is that cannulation of both adrenal veins is technically demanding; few radiologists can claim a greater than 75 per cent success at cannulating the right adrenal vein, which drains directly into the back of the inferior vena cava. A hooked catheter with side holes should be used.

The alternative test is the radio-isotope scan using selenium cholestenol as a precursor for adrenal steroids. This is less invasive than venous sampling, but in practice can inflict more discomfort and is less accurate. Because most steroid synthesis takes place in the zona fasciculata of the adrenal, this needs to be blocked by pretreatment with dexamethasone for at least a week. In addition, several of the main antihypertensives, including those most likely to be effective in Conn's syndrome, can interfere with the scan and need to be stopped. Scanning is usually performed at 3, 7, and sometimes 14 days after radionuclide administration. For reasons that are not clear, the scan is often misleading, failing to detect any uptake in some patients, and failing to lateralize in some patients with adenomas.

Treatment

Medical

The medical treatment of choice is spironolactone, which is a competitive antagonist of the aldosterone receptor. It can be started once the diagnosis is suspected from the biochemical results, and the blood pressure response is valuable in confirming the presence of increased aldosterone secretion. Most patients will have a substantial fall in blood pressure, and normalization of plasma electrolytes after a month's treatment at a dose of 50 mg daily. This relatively low dose has the advantage of reducing the risk of adverse reactions, of which dyspepsia is the main short-term and gynaecomastia the main long-term problem. However, 50 mg is too small a dose for larger patients, and it is worth prescribing as near as possible to 1 mg/kg (using the available 50 mg and 25 mg size tablets). The dose can sometimes be reduced after prolonged administration.

Not all patients with Conn's syndrome tolerate or respond adequately to spironolactone alone. Usually a dose is tolerated sufficient to permit control of hypokalaemia; if not, high doses (20–40 mg) of amiloride may be required. As mentioned above, the calcium blockers can suppress aldosterone secretion and are a logical addition. However, patients with Conn's syndrome can develop quite resistant hypertension, which is a reason for trying harder to make the diagnosis at an early stage in the

development of hypertension. Spironolactone should be stopped a week before any lateralization tests, because of the risk that suppression of the contralateral adrenal is removed.

Surgical

This is the treatment of choice for patients with the unilateral adenomas. Because the tumours are always small, they lend themselves well to laparoscopic surgery, although patients should always be warned about the possible need to proceed to an open operation. Patients should be on spiro-nolactone (if tolerated) in the period running up to surgery, but there are few risks of uncontrolled hypertension or postoperative hypotension in the surgery of Conn's syndrome.

No steroid replacement is required after surgery, when most patients are able to discontinue all antihypertensive therapy. This may not be possible in older patients, perhaps because the adenoma arose on a background of essential hypertension. There is no need for long-term follow-up.

It seems likely that Conn's is familial more often than has been recognized, or can be currently explained by any known genetic variant, and it is therefore worth considering whether siblings with hypertension should have their plasma renin and aldosterone measured.

Further reading

Barzon L *et al.* (1999). Risk factors and long-term follow-up of adrenal incidentalomas. *Journal of Clinical Endocrinology and Metabolism* **84**, 520–6.

Brown MJ, Hopper RV (1999). Calcium-channel blockade can mask the diagnosis of Conn's syndrome. *Postgraduate Medical Journal* **75**, 235–6.

Dluhy RG, Lifton RP (1999). Glucocorticoid-remediable aldosteronism. *Journal of Clinical Endocrinology and Metabolism* **84**, 4341–4.

Ganguly A (1998). Primary aldosteronism. *New England Journal of Medicine* **339**, 1828–34.

Gordon RD *et al.* (1994). High incidence of primary aldosteronism in 199 patients referred with hypertension. *Clinical and Experimental Pharmacology and Physiology* **21**, 315–18.

Nomura K *et al.* (1992). Plasma aldosterone response to upright posture and angiotensin II infusion in aldosterone-producing adenoma. *Journal of Clinical Endocrinology and Metabolism* **75**, 323–7.

Stewart PM (1999). Mineralocorticoid hypertension. *The Lancet* **1353**, 1341–7.

Stewart PM *et al*. (1996). Hypertension in the syndrome of apparent mineralocorticoid excess due to mutation of the 11 beta-hydroxysteroid dehydrogenase type 2 gene. *The Lancet* **347**, 88–91.

Stowasser M *et al.* (1995). Plasma aldosterone response to ACTH in subtypes of primary aldosteronism. *Clinical and Experimental Pharmacology and Physiology* **22**, 460–2.

15.16.2.4

Phaeochromocytoma

M. J. Brown

Introduction

Phaeochromocytoma is a rare tumour. Estimates of incidence are unreliable because none has been undertaken in an unselected group of patients, but it is probably in the range 0.1–1 per cent of hypertensives. During a study of prevalence of hypertension, we measured blood pressure in 30 000 healthy subjects in general practice, selected only for the absence of known hypertension or vascular disease: 8 per cent were found to have a systolic blood pressure of more than 150 mmHg, and two were subsequently found to have a phaeochromocytoma, giving an incidence of about 0.1 per cent of hypertensives. This fits our parallel experience in 750 patients referred over 10 years by general practitioners following their own diagnosis of hypertension and before initiation of treatment; only one of these patients was found to have a phaeochromocytoma. Both these series will have missed those cases in whom typical symptoms led to correct diagnosis soon after presentation with hypertension, but it is reassuring to know that for a potentially malignant tumour (unlike the situation in Conn's syndrome, see Chapter 15.16.2.3) the typical picture is unlikely to be just the tip of an iceberg.

Despite its rarity, phaeochromocytoma justifies the disproportionate interest and awareness of the condition that exists among physicians. Like a few other rare conditions which share this position, such as infective endocarditis or Addison's disease, phaeochromocytoma combines the potential for being lethal if not diagnosed and treated, and for cure in most patients if diagnosed. The diagnosis of phaeochromocytoma offers the best chance of a cure of all the secondary causes of hypertension (especially those presenting in the second half of life), and avoidance of the need for lifelong antihypertensive therapy.

The need for maintaining a high awareness of the condition is emphasized by the small number of deaths each year, in both anaesthetic and obstetric practice, due to undiagnosed phaeochromocytoma.

Catecholamine biochemistry

An understanding of the tests used to diagnose phaeochromocytoma requires reference to an outline of both the synthetic and degradative pathways of catecholamine metabolism. The term catechol refers to a phenyl ring with hydroxyl groups at adjacent carbons (conventionally, the 3' and 4' positions). The precursor essential amino acid, phenylalanine, is not itself a catechol; neither is tyrosine which has only the 3' hydroxyl. This amino acid is the substrate for the rate-limiting step in the biosynthetic pathway, tyrosine hydroxylase, which yields L-dopa, the first catechol and still an amino acid. Decarboxylation of L-dopa yields the first catecholamine in the pathway, dopamine. This can occasionally be the principal catecholamine secreted by phaeochromocytomas, or more often by childhood neuroblastomas, but in the chromaffin tissue from which phaeochromocytomas originate dopamine is usually further hydroxylated, in the sidechain bearing the amine group, to noradrenaline. The final step in the biosynthetic pathway is the *N*-methylation of noradrenaline to adrenaline, the prefix 'nor' being used for substances that are *N*-demethylated, a common step in degradative metabolism.

N-methylation usually occurs in only two sites in the body: the adrenal medulla and certain hindbrain nuclei involved in blood pressure control. The enzyme responsible, phenylethanolamine-*N*-methyltransferase, may differ between these sites, as outside the central nervous system it is dependent for induction on glucocorticoids, which are provided in the adrenal through the portocapillary circulation. The clinical importance of this is threefold. First, extra-adrenal phaeochromocytomas rarely produce adrenaline, most of the reports to the contrary being in older literature when the methodology was less satisfactory for separating adrenaline and noradrenaline. Secondly, the normal adrenal produces mainly adrenaline, and accounts for less than 2 per cent of circulating noradrenaline concentrations. Thirdly, when a tumour is present in the adrenal, the disruption of the portocapillary circulation causes a reversal of the normal adrenaline to noradrenaline ratio. The relevance of these to the clinical features and diagnosis of phaeochromocytoma will become apparent.

The metabolic breakdown of catecholamines is due to two principal enzymes, monoamine oxidase and catechol-*O*-methyltransferase. The metabolism of catecholamines is different from normal in phaeochromocytoma in that adrenaline and noradrenaline are liberated directly into the bloodstream rather than mainly into the synaptic gap around sympathetic nerve endings. Noradrenaline released into these gaps is largely recaptured by neuronal and extraneuronal uptake, being metabolized before any free amine escapes into the bloodstream. Consequently, the proportion of parent amine to metabolite is usually higher in blood and urine in the presence of a phaeochromocytoma than in any other cause of elevated catecholamine production.

The most abundant product of the action of monoamine oxidase and catechol-*O*-methyltransferase (acting in sequence, in either order) is vanillylmandelic acid. Normetanephrine and metanephrine are produced by catechol-*O*-methyltransferase from noradrenaline and adrenaline, respectively. The products of monoamine oxidase alone are less often used in diagnosis, but in specialized laboratories the ratio of one of these, dihydroxyphenylglycol, to noradrenaline is a useful clue to the origin of noradrenaline, as dihydroxyphenylglycol arises mainly intraneuronally, and relatively little is therefore formed from noradrenaline liberated directly into the bloodstream as in phaeochromocytoma.

Laboratory diagnosis of phaeochromocytoma

Measurements of catecholamine metabolites in 24-h urine samples are the most appropriate screening tests because they offer an integrated measure of total catecholamine release over this period and can be performed in most routine laboratories. Vanillylmandelic acid is the product least prone to interference, L-dopa being the only drug that can crossreact in measurement through its equivalent metabolite, homovanillic acid. Although, for the reasons discussed, quantitation of vanillylmandelic acid is less sensitive than other measures, it remains very rare that a patient with a secreting phaeochromocytoma has a 24-h vanillylmandelic acid result that is normal. High doses of a-methyldopa reduce, and occasionally normalize, vanillylmandelic acid levels by competitively inhibiting catechol-*O*-methyl-transferase. However, there is a problem of distinguishing the true positive from the relatively large number of hypertensive patients with results in the 'grey zone'. Metanephrines are arguably more sensitive than vanillylmandelic acid, but their assay is more prone to interference by drugs, especially by b-blockers.

The measurement of free catecholamines is a more specialized procedure. Assays can be made in plasma or urine, the former generally being more accurate and also allowing variation in secretion to be assessed because of the very short half-life in plasma (around 1 min) of catecholamines. The assay generally used is based on high-performance liquid chromatography separation of the catecholamines, followed by electrochemical or fluorometric detection. However, this technique does not eliminate the possibility of interference, especially in the adrenaline peak, and it is necessary to be particularly suspicious of any result showing a higher adrenaline than noradrenaline concentration. A few centres still undertake the gold-standard radioenzymatic assay in which the catecholamines are converted to their [^{3}H]-methylated derivative in the presence of catechol-*O*-methyltransferase and a [^{3}H]-methyl donor, a double-isotope technique usually being used to ensure accurate quantification.

Pathology

Phaeochromocytomas arise in chromaffin tissue, and their anatomical distribution closely parallels the sites where this tissue is present at the time of birth. These

tumours, like the normal sympathoadrenal tissue, are of neuroectodermal origin. The term phaeochromocytoma reflects the dusky colour of the cut surface of the tumour, whereas the term chromaffin refers to the brownish colour caused by contact with dichromate salts, which oxidize the catecholamines. Much has been written about pathological differences between extra-adrenal phaeochromocytomas at various sites, but this is not relevant to clinical practice except as a possible explanation for the failure of some head and neck phaeochromocytomas to accumulate the noradrenaline analogue used as a radionuclide in scanning, as discussed later.

The pathogenesis of most phaeochromocytomas is unknown. However, at least 10 per cent of patients have inherited, as an autosomal dominant, their susceptibility to phaeochromocytoma. Two syndromes are recognized, both with germline mutations in a tumour suppressor gene. In multiple endocrine neoplasia type 2 syndrome, phaeochromocytoma is most commonly associated (not necessarily at the same time) with medullary carcinoma of the thyroid. In von Hippel–Lindau syndrome, multiple other tumours can occur, most commonly retinal angiomas, but also hypernephroma and central nervous system haemangioblastomas. There is evidence of somatic mutations in the same genes as responsible for these syndromes in some of the sporadic phaeochromocytomas. Given the large size of the von Hippel–Lindau gene, it is likely that many more sporadic phaeochromocytomas have such mutations than have been identified. Familial phaeochromocytoma without other tumours can be caused by mutations in the succinate dehydrogenase complex.

Most phaeochromocytomas are benign. However, the pathologist can rarely provide a clear distinction between benign and malignant phaeochromocytomas: benign tumours can appear to be invading the capsule of the tumour, which is often ill defined, and malignant tumours may show no mitoses because of their slow rate of division.

Clinical features

Hypertension is the most common presentation of phaeochromocytoma in clinical practice; other rare presentations include unexplained heart failure or paroxysmal arrhythmias. Increasingly, small and asymptomatic phaeochromocytomas are detected through regular screening of patients with a genetic diagnosis of multiple endocrine neoplasia or von Hippel–Lindau syndrome.

In the hypertensive patients, a spontaneous history or direct enquiry will usually reveal at least one of a group of characteristic symptoms. The most common are headache, sweating, and palpitations. Less frequent are episodes of pallor, a feeling of 'impending doom', and paraesthesiae. Examination rarely reveals useful signs, but an exception is a Raynaud's-type of discoloration over the extremities and the larger joints in the limbs. This is due to ischaemia and occasionally progresses to atrophic ulceration over pressure points.

A rare initial presentation, pathognomonic of phaeochromocytoma, is with wildly swinging blood pressure, between extremes of hypertension and hypotension, in combination with other signs of retroperitoneal haemorrhage. Even prompt recognition of the diagnosis, spontaneous haemorrhage and infarction of the tumour, is not always sufficient to save patients presenting in this way. By contrast, it is important to emphasize that some patients with large tumours, causing significant hypertension, may be asymptomatic.

Many of the symptoms of phaeochromocytoma can be readily ascribed to the expected effects of the excess catecholamine, and disappear rapidly on initiation of appropriate treatment. Some remain more difficult to explain, including the sweating whose control in healthy subjects is usually ascribed to cholinergic sympathetic innervation. Because large tumours secrete principally noradrenaline, even when arising within the adrenal gland, tachycardia is usually only modest, and can be replaced altogether by reflex bradycardia when episodes of hypertension are triggered by release of noradrenaline alone. The author once treated a 'cardiac arrest' with phentolamine when the arrest call was for an episode of apparent asystole that was in reality a vagal reaction to an arterial pressure of 300/160. Severe bradycardia is also recorded in response to the paradoxical rise in blood pressure when a patient with a phaeochromocytoma is inadvertently given a non-selective b-blocker such as propranolol. In a few patients, excess catecholamine can cause myocardial necrosis, which is probably due to a mixture of a-receptor mediated vasoconstriction and a b-receptor direct toxic effect on the cardiomyocytes. These are rare presentations and it is important to recognize that clinical features are usually less impressive than expected, possibly because the adrenoceptors have been downregulated by years of exposure before the diagnosis is first entertained. Indeed, hypertensive patients who complain of symptoms suggestive of excess catecholamines are more likely to be found to be suffering from side-effects of a vasodilator drug activating the baroreflex than to have a phaeochromocytoma.

Establishing the diagnosis

The diagnosis is not usually difficult once the possibility of phaeochromocytoma has been entertained, and it is important to exclude the diagnosis in patients who have clinical and/or biochemical features of catecholamine excess due to sympathetic overactivity rather than phaeochromocytoma. There are two distinct questions to ask: 'Does the patient have a phaeochromocytoma?' and 'Where is it?'. The tests required to answer the first question are mainly biochemical, as described above, whereas the second is answered by radiological investigation. A golden rule, which saves false positives and negatives, and therefore a large number of unnecessary investigations and sometimes operations, is that the first question should be answered before proceeding to the second. No single radiological investigation is sufficiently accurate to detect more than 80 to 90 per cent of phaeochromocytomas, whilst computed tomography scanning of the adrenal glands can detect non-functional myoleiomas that should not lead to further investigation in the absence of biochemical abnormalities.

The symptoms of a functioning phaeochromocytoma are remarkably variable and may be absent. To avoid missing the diagnosis, therefore, an average size general hospital might expect each year to screen several hundred patients with hypertension, tens of patients with unexplained heart failure, and all their known patients with multiple endocrine neoplasia or von Hippel–Lindau syndromes. Although the diagnosis is often postulated in other patients with isolated features of catecholamine excess, for example patients without hypertension but complaining of palpitation, headaches, sweating, or panic attacks, the chance of such patients having a phaeochromocytoma is very, very small. In a 15-year period during which the author investigated more than 100 phaeochromocytomas and more than 1000 patients referred with a possible phaeochromocytoma, none has proven to have phaeochromocytoma in the absence of hypertension, heart failure, or a genetic syndrome.

The next question is how should screening be performed? There is no single perfect or 'best' test. It is important to recognize the diversity of analyses in use, quite different from the position for most standard endocrine analyses, and reflecting the difficulty of achieving an entirely reliable method in routine laboratories. Our own practice is to use 24-h urine vanillylmandelic acid as the initial screen in most patients, supplemented when necessary with the specialized catecholamine analyses. An entirely normal 24-h urine vanillylmandelic acid measured in a good hospital laboratory is most unlikely in the presence of a phaeochromocytoma. Conventionally, patients are asked to avoid vanilla-containing foods during the collection for assay of vanillylmandelic acid and to undertake three collections in order to exclude the diagnosis of phaeochromocytoma. Both of these precautions are unnecessary in the majority of cases: those with phaeochromocytoma have become relatively insensitive to the effects of catecholamines, hence a patient with 'significant' hypertension (diastolic blood pressure over 100 mmHg) must have a several-fold elevation of catecholamine secretion. Although the urinary vanillylmandelic acid is not proportionally elevated, for the reasons discussed earlier, the elevation is still sufficient to ensure an abnormal result provided that this is correctly measured. A vanilla-free diet is unnecessary because the dietary contribution to vanillylmandelic acid excretion is small compared with that derived from noradrenaline, and is unlikely to push the vanillylmandelic acid excretion into an abnormal range.

Most patients whose vanillylmandelic acid excretion is more than twofold above the upper limit of normal will prove to have a phaeochromocytoma, and a threefold elevation is almost always diagnostic. Patients who need further biochemical analyses are those with a less than twofold elevation of vanillylmandelic acid excretion, of whom only a very small proportion (less than 5 per cent) will have a phaeochromocytoma. Here, the single most helpful investigation is measurement of plasma noradrenaline, which will be at least twofold elevated in those with a phaeochromocytoma, whereas a single resting plasma noradrenaline will often be normal in those without.

Suppression tests

If the urinary vanillylmandelic acid analysis and assay of resting plasma noradrenaline does not resolve whether or not a patient has a phaeochromocytoma, there are two further useful investigations. The most widely used is a pharmacological suppression test, in which physiological elevations of noradrenaline release are temporarily suppressed by administration of either the ganglion-blocking drug pentolinium, or centrally acting a $_2$-agonist, clonidine. The former is more widely used in the United Kingdom and has three advantages:

1. It is most effective at suppressing noradrenaline release in the problem patients—namely those with elevated sympathetic nervous activity but without a phaeochromocytoma.
2. It also suppresses release of adrenaline from the adrenal medulla.

3. It has a short half-life (of approximately 20 min) so that the test can be completed in the outpatient clinic.

Clonidine has the supposed advantage of suppressing release of noradrenaline even when the basal level is normal, but when this is the case a suppression test is rarely necessary. An exception is in patients with von Hippel–Lindau syndrome found to have a small adrenal mass on their annual abdominal computed tomography (**CT**) or magnetic resonance scan: they may have a normal plasma noradrenaline concentration and, unlike the multiple endocrine neoplasia patients with phaeochromocytoma, the tumour secretes little adrenaline. In the multiple endocrine neoplasia patients, by contrast, an elevated plasma adrenaline concentration is the first biochemical abnormality, and biochemical diagnosis is likely to precede radiological diagnosis in multiple endocrine neoplasia, where annual scans are not indicated.

Only patients with normal or near normal renal function (serum creatinine less than 150 µmol/litre) are suitable for the pentolinium test, as this agent is entirely excreted by the kidneys. After the patient has rested supine for 15 to 30 min, plasma catecholamines are measured in two samples taken 5 min apart from an intravenous cannula, and in two further samples taken 10 and 20 min after an intravenous bolus of pentolinium 2.5 mg. They should remain supine for a further 60 min, and their erect arterial pressure should be checked before they are allowed to leave the clinic. A normal response to pentolinium is a fall of both plasma noradrenaline and adrenaline concentrations into the normal range or by 50 per cent from baseline. It should be noted that since ganglion-blocking drugs are less effective at low rates of sympathetic nerve discharge there may be little fall in plasma catecholamine values when the basal levels are already within the normal range.

Another test that can sometimes be helpful is to assay a plasma sample for dihydroxyphenyl glycol, the deaminated metabolite synthesized principally in sympathetic nerve endings. The ratio of dihydroxyphenyl glycol to noradrenaline is reversed from normal (more dihydroxyphenyl glycol than noradrenaline) in patients with phaeochromocytoma, allowing an alternative method to a suppression test for distinguishing patients with borderline noradrenaline results.

Localization of phaeochromocytomas

It is helpful to measure plasma catecholamines even in patients with unequivocal elevation of their 24-h urine vanillylmandelic acid as the adrenaline level is a most useful clue to the location of a phaeochromocytoma. Most adrenal phaeochromocytomas do secrete adrenaline, although the proportion of noradrenaline to adrenaline is reversed from that in normal subjects, whilst it is exceptional for extra-adrenal phaeochromocytomas to secrete adrenaline because of the lack of cortisol stimulation.

Although a major clue to localization can be provided by measurement of plasma adrenaline, CT scanning is the method of choice. The adrenal gland, where 90 per cent of phaeochromocytomas arise (Fig. 1), is easy to visualize, and imaging is able to distinguish cortical tumours such as a Conn's tumour from medullary tumours. These differences should, however, not be used as a basis for diagnosis: mistakes will be made if the differentiation between these tumours is attempted radiologically rather than biochemically. It should also be emphasized that both of these tumours account for a minority of adrenal tumours identified by CT, the majority of which are non-functional adenomas of no significance.

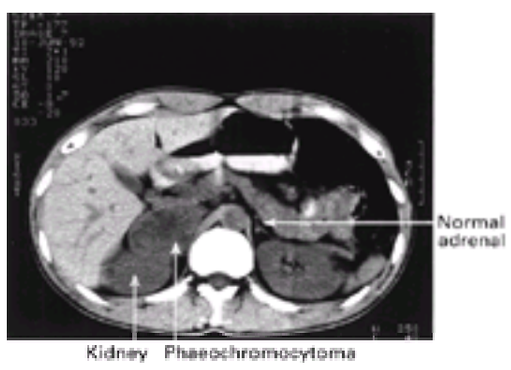

Fig. 1 CT scan of a right adrenal phaeochromocytoma. The phaeochromocytoma has the typical non-homogeneous appearance due to areas of haemorrhage and infarction. The normal left adrenal has the typical tricornuate appearance with concave borders.

While modern CT is capable of whole body imaging at high resolution, it is preferable to withhold CT for extra-adrenal phaeochromocytomas until the radiologist can be given some clue as to where to direct their activities. In about 85 per cent of patients, this can be achieved by radio-isotopescanning, using the iodinated analogue of noradrenaline, *m*-iodo-benzylguanidine. There is a case for undertaking such scanning in addition to CT, even for patients found to have an adrenal phaeochromocytoma, to identify extra-adrenal secondary deposits when tumours are malignant, and because there may be coexisting adrenal and extra-adrenal phaeochromocytomas.

If these investigations fail to localize a phaeochromocytoma diagnosed by biochemical assays, the next step is to undertake selective venous sampling. In this procedure, about 25 samples of blood for estimation of catecholamine concentration are collected under fluoroscopic guidance from various sites in the vena cava and the veins that drain into it. An arterial sample taken at the end of the procedure is invaluable for interpretation of data, as it enables sites with a positive venoarterial difference to be readily detected. Although invasive, venous sampling is free of significant hazard, but it is important that the radiologist is not tempted to undertake a venogram of the phaeochromocytoma, since this can cause immediate infarction of the tumour with release of the stored catecholamines and catastrophic consequences. The procedure is more helpful in the diagnosis of phaeochromocytoma than of other endocrine tumours because of the very short half-life of catecholamines in the circulation (about 1 min), such that most is removed during one passage round the circulation, hence the concentration at the tumour site is usually several fold greater than concentrations elsewhere. This procedure should not usually be used for adrenal phaeochromocytomas because the concentration of catecholamines is much higher than elsewhere in veins draining normal adrenals, and because CT scanning should have already rendered their imaging unnecessary. An exception, once again, is in patients with von Hippel–Lindau syndrome with small adrenal masses. As discussed above, these are the patients in whom all other biochemical tests may be normal, and the diagnosis of phaeochromocytoma is suggested by a reversal of the normal excess of adrenaline to noradrenaline in the adrenal vein. Because in von Hippel–Lindau syndrome the adrenal phaeochromocytomas are frequently bilateral, but asymmetrical, venous sampling may be required to determine whether one or both adrenals need to be removed.

The place of angiography has been much diminished but not entirely removed by the advent of CT scanning. As phaeochromocytomas are vascular tumours, they provide a good tumour blush, and angiography should resolve equivocal CT scans. However, by contrast to venous sampling, this procedure can provoke an outpouring of catechols. Patients must be fully a- and preferably also b-blocked prior to angiography, and their blood pressure, pulse rate, and ECG must be monitored during the procedure with phentolamine and atenolol readily available to treat arterial hypertension or tachycardia.

In some centres, magnetic resonance imaging may be tried before angiography to determine the nature of lesions of doubtful significance on CT. However, the semi-infarcted nature of some phaeochromocytomas can make it difficult to interpret magnetic resonance scans, and in our experience such imaging has only helped with a few head and neck phaeochromocytomas that were not detected by *m*-iodobenzylguanidine or CT scanning.

Other investigations

It is important to check blood glucose in every patient as there may be a mediated inhibition of insulin release prior to effective treatment. All patients should be screened for an associated medullary carcinoma of the thyroid by plasma calcitonin estimation.

Apart from the catecholamines, most phaeochromocytomas also secrete one or more neuropeptides, especially neuropeptide Y (NPY), which is a normal cotransmitter of noradrenaline and adrenaline. There is no need routinely to measure other neurotransmitters that may be cosecreted with the catecholamines, but unusual symptoms may indicate that a gut peptide screen should be undertaken. Although very rare, it is essential to detect (especially preoperatively) coexisting ectopic adrenocorticotropic hormone syndrome that manifests as gross hypokalaemia. A particular catch is that the excess secretion of catecholamines may suppress release

of other peptides until treatment with a-blockade is initiated, hence it is important to recheck the electrolytes after a few days of a-blocking treatment.

Even where there is no suggestive family history, routine slit-lamp examination of the fundi has resulted in a more frequent diagnosis of von Hippel–Lindau syndrome, sometimes as a *de novo* occurrence.

Treatment

The definitive treatment is surgical removal of the tumour or tumours. Even the small number of phaeochromocytomas that can be recognized to be malignant preoperatively (e.g. by the presence of bone or liver metastases) may still benefit from resection of the primary tumour. The task for the physician is to make the surgery safe. The mainstay of medical treatment is a-blockade, but not all patients—especially those without elevated plasma adrenaline levels—require b-blockade. The objective of this treatment is not solely control of blood pressure but also the expansion of blood volume, which is always reduced in those with a phaeochromocytoma. The a-blocker of choice is phenoxybenzamine, the principal reason for this being that it is irreversible, actually destroying the a-receptor by alkylation. More modern a-blockers, such as prazosin, doxazosin, and the mixed a- and b-blocker, labetalol, cause competitive blockade, which can be overcome by a surge of noradrenaline release from the tumour. An additional advantage of phenoxybenzamine is that it will block both a_1- and a_2-receptors. Blockade of the latter is considered disadvantageous in essential hypertension since the main a_2-receptors outside the central nervous system are presynaptic and may serve a useful role in damping neuronal release of noradrenaline, whereas in phaeochromocytoma patients a_2-receptor blockade may be advantageous because a small population of extrasynaptic a_2-receptors mediate direct vasoconstriction by circulating (non-neuronal) catecholamines. The diabetogenic effect of catecholamines is also an a_2-mediated response.

The starting dose of phenoxybenzamine depends on the degree of catecholamine excess, but is usually 10 mg twice daily. The effect of irreversible antagonists is cumulative, and the effect of the drug—and each subsequent dose increment—takes several days to reach maximum. It is reasonable to aim for a diastolic blood pressure of between 90 and 100 mmHg during outpatient treatment, and to admit patients for 5 days preoperatively, during which time the dose is increased until there is at least a 10 mmHg postural fall in blood pressure and little if any variability in arterial pressure. An important objective of preoperative a-blockade is expansion of intravascular volume. Surgery should not take place until a new steady-state weight has been achieved, which usually requires about a month.

The need for b-blockade is indicated by tachycardia, which may become apparent only after treatment with phenoxybenzamine. Lower doses of b-blocking drugs are necessary than used generally in the treatment of hypertension, and it is usually better to use a selective b_1-selective agent so that the peripheral vasodilatation mediated by b_2-receptors is not affected. The reason for using as low a dose as possible is that immediately upon removal of the phaeochromocytoma there may be a period of hypotension despite the preoperative preparation that has been outlined. This is due to the withdrawal of any a-mediated vasoconstriction, and should normally be offset by the ability to mount a tachycardia. It is important to note that if hypotension does occur, it should not be treated with pressor agents; the correct treatment is by volume replacement, supplemented if necessary by b-agonists. Most vasoconstrictor drugs are unlikely to be effective, because of the previous treatment with phenoxybenzamine. Angiotensin is no longer available as a pharmaceutical preparation.

The treatment of malignant phaeochromocytomas remains uncertain and unsatisfactory. As is the case for many endocrine cancers, the rate of growth is usually slow, but outcome can vary from local recurrence at intervals of many years to rapid demise, sometimes precipitated by surgery. These tumours are not particularly sensitive to either chemotherapy or radiotherapy. There has been interest in the use of therapeutic doses of *m*-iodobenzylguanidine, as a means of targeting high doses of radioactivity to the tumour, and some patients show considerable regression after such treatment. Long-term results are less certain. If the primary tumour has been removed or debulked, it is rare for the pharmacological effects of the tumour to be the principal problem. However, if this is the case, then high doses of phenoxybenzamine are greatly preferable to a-methyltyrosine, occasionally used as an inhibitor of noradrenaline synthesis, but which also depletes noradrenaline in the brain, causing sedation and depression.

Prognosis

Ninety per cent of phaeochromocytomas are benign. For adrenal phaeochromocytomas, the proportion is probably even higher, whereas extra-adrenal phaeochromocytomas have a greater than 10 per cent likelihood of proving malignant. However, because of the difficulties already described in ascertaining malignancy, all patients with a phaeochromocytoma should be followed indefinitely with at least an annual measurement of arterial pressure and analysis of one of the indices of catecholamine secretion.

The removal of a phaeochromocytoma cures most patients of their hypertension, especially the younger ones. In only 13 out of a personal series of 76 patients with phaeochromocytoma was the blood pressure greater than 140/85 at 6 and 12 months postoperatively (compared with an average 172/114 at presentation), and these 13 were all aged over 50 (compared with an average age of 37 for all 76 patients).

Further reading

Allison DJ *et al.* (1983). Role of venous sampling in locating a phaeochromocytoma. *British Medical Journal* **286**, 1122–4.

Brown MJ *et al.* (1981). Increased sensitivity and accuracy of phaeochromocytoma diagnosis achieved by use of plasma adrenaline estimations and a pentolinium suppression test. *The Lancet* **i**, 174–7.

Col V *et al.* (1999). Laparoscopic adrenalectomy for phaeochromocytoma: endocrinological and surgical aspects of a new therapeutic approach. *Clinical Endocrinology* **50**, 121–5.

Manger WM (1997). *Pheochromocytoma.* Springer, Berlin.

Richards FM *et al.* (1998). Molecular genetic analysis of von Hippel–Lindau disease. *Journal of Internal Medicine* **243**, 527–33.

Sisson JC, Shulkin BL (1999). Nuclear medicine imaging of pheochromocytoma and neuroblastoma. *Quarterly Journal of Nuclear Medicine* **43**, 217–23.

15.16.2.5Aortic coarctation

Lawrence E. Ramsay

Further reading

Coarctation is a congenital narrowing of the aorta near the junction with the ligamentum arteriosum and usually distal to the left subclavian artery. About 20 per cent of coarctations present in adolescent or adult years, often with hypertension, but coarctation is very rare in unselected hypertensive patients. It is very easy to overlook, but should come to mind in hypertensive patients who are young, male (four times commoner than in females), have isolated or disproportionate systolic hypertension, have a prominent murmur over the precordium or back, have disproportionate left ventricular hypertrophy, or have Turner's syndrome. About 70 per cent have other congenital abnormalities, most commonly a bicuspid aortic valve, but also ventricular septal defect or patent ductus arteriosus. Palpate routinely for delayed or weak femoral pulses in all hypertensive patients. When coarctation is suspected, seek other clinical features which include palpable collaterals in the back, displaced apex beat and systolic thrill, systolic murmurs over the coarctation, collaterals, or aortic valve, and an ejection click or loud aortic valve closure. Measuring blood pressure in the legs with a sphygmomanometer is a nightmare, and Doppler measurement of the ankle–brachial pressure index is much more accurate. A chest radiograph may show rib notching (Fig. 1), cardiomegaly, and abnormal aortic configuration. The diagnosis is confirmed by aortography (Fig. 2) or magnetic resonance imaging (Fig. 3). Echocardiography reveals the state of the aortic valve and left ventricle, and aortography is done before intervention to define the anatomy and haemodynamics.

Fig. 1 Chest radiographic appearances of rib notching in a patient with coarctation of the aorta. (By courtesy of Dr N. Boon.)

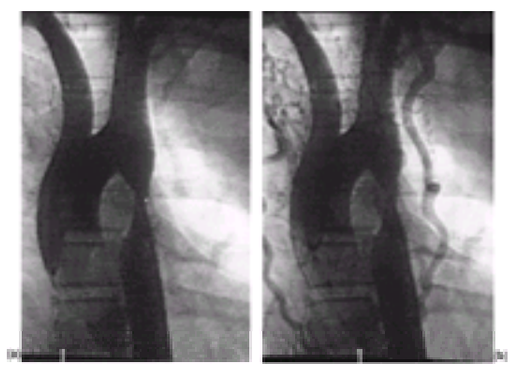

Fig. 2 (a) Digital aortogram showing typical appearances of a coarctation of the aorta. (b) A later frame showing marked dilatation of the internal mammary arteries due to increased collateral flow. (By courtesy of Dr N. Boon.)

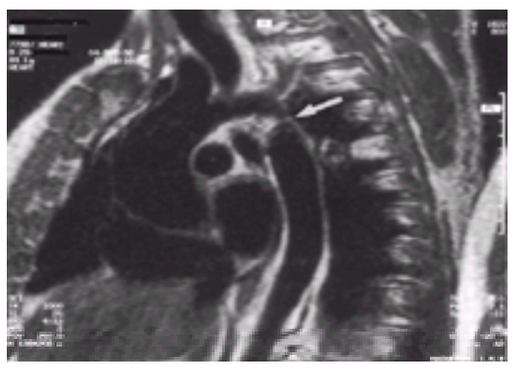

Fig. 3 Sagittal spin echo magnetic resonance image scan showing a well defined coarctation (arrow) in the typical position. (By courtesy of Dr N. Boon.)

Untreated coarctation has a bad prognosis because of heart failure, aortic rupture or dissection, cerebral haemorrhage related to berry aneurysms, bacterial endocarditis or endarteritis, and other complications of hypertension. Correction of coarctation improves survival, but not to a normal life expectancy. The prognosis after correction is affected adversely by correction at an older age, and persistent hypertension after correction. Coarctation should therefore be dealt with as soon as it is diagnosed, provided that this is technically possible and the patient is otherwise fit. Surgical repair may be complicated by postoperative paradoxical hypertension, and mesenteric or spinal ischaemia. Balloon angioplasty has lower immediate morbidity and mortality, and is possible as a primary procedure in about 90 per cent of cases, and also for recoarctation. There is uncertainty about the long-term outcome after angioplasty, particularly as regards development of aneurysms, but data to 5–10 years postangioplasty are reassuring. Nevertheless, indefinite follow-up and monitoring with repeat magnetic resonance scans are needed to detect possible recurrence or aneurysm development. About 30 per cent of patients still need treatment for hypertension after correction, with a higher probability as the age at correction increases. Patients need advice on prophylaxis against bacterial endocarditis before and after correction.

Further reading

Fawzy ME *et al*. (1997). 1–10 year follow-up results of balloon angioplasty of native coarctation of the aorta in adolescents and adults. *Journal American College of Cardiology* **30**, 1542–6.

Jenkins MP, Ward C (1999). Coarctation of the aorta: natural history and outcome after surgical treatment. *Quarterly Journal of Medicine* **92**, 365–71.

de Leeuw PW, Birkenhäger WH (1992) Coarctation of the aorta. In: Robertson JIS, ed. *Handbook of hypertension, Vol.15: Clinical hypertension*, pp 236–65. Elsevier, Amsterdam.

15.16.2.6Other rare causes of hypertension

Lawrence E. Ramsay

Further reading

Identifiable causes of hypertension are commonly renal, renovascular, drug-induced, primary aldosteronism, or a phaeochromocytoma. There are numerous other causes of, or associations with, hypertension (Table 1). Some are rare conditions that usually cause hypertension (for example, liquorice excess, renin-secreting tumour); some are rare conditions that rarely cause hypertension (for example, carcinoid syndrome); some are common conditions that rarely present as hypertension (for example, pregnancy presenting as pre-eclampsia); and some are associations that may not be causal (such as hypothyroidism, acromegaly).

Mineralocorticoid excess is suggested by hypertension, hypokalaemia, and high or high-normal serum sodium levels. Primary aldosteronism (Conn's syndrome, see Chapter 15.16.2.3) is the commonest cause, but rare causes should be suspected (Table 1) if the aldosterone level is low or normal rather than high. Most important among these is excess liquorice ingestion, because asking one simple question may avoid complex and costly investigations and rapidly resolve the hypertension and biochemical abnormalities. Liquorice excess causes hypertension, hypokalaemia, and increased levels of free cortisol in urine through an acquired 11-dehydro-genase deficiency.

Table 1 includes several conditions that can cause paroxysmal hypertension, sometimes with additional symptoms that may closely mimic a phaeochromocytoma. Some also show excess urinary or plasma catecholamines, particularly posterior fossa tumours near the fourth ventricle, which may be tiny and virtually undiagnosable. These causes of paroxysmal hypertension should be considered when investigation of a patient with features suggesting phaeochromocytoma fails to localize a phaeochromocytoma or other catecholamine-secreting tumour.

Further reading

Laragh JH, Brenner BM, eds (1990). *Hypertension. Pathophysiology, diagnosis and management*. Raven Press, New York.

Swales JD, ed (1994). *Textbook of hypertension*. Blackwell Scientific, Oxford.

15.16.3 Hypertensive emergencies and urgencies

Gregory Y. H. Lip and D. Gareth Beevers

Introduction

Hypertensive emergencies occur when severe hypertension is associated with acute end-organ damage. These can take a variety of forms and can occur at any age. They may be acute life-threatening medical conditions, and are associated with either severe hypertension or sudden marked increases in blood pressure (Table 1). Symptomatic patients with complications such as aortic dissection and hypertensive encephalopathy require parenteral antihypertensive therapy to reduce the blood pressure promptly, but in a controlled manner and with careful monitoring. However, over-rapid treatment may itself be hazardous, leading on occasions to ischaemic complications such as stroke, myocardial infarction, or blindness. Thus, in patients who have severe hypertension but are asymptomatic, slower controlled reduction in blood pressure should be achieved with oral antihypertensive agents, making such situations hypertensive 'urgencies' rather than 'emergencies'.

In general, there has been a decline in the incidence of hypertensive emergencies over the past 20 years in the Western world, which may possibly be the result of the more effective detection, diagnosis, and treatment of mild to moderate hypertension.

If patients with hypertensive emergencies are not recognized or treated appropriately, the mortality and morbidity can be very high, with the 1-year mortality being 70 to 90 per cent, and the 5 year mortality 100 per cent. With adequate blood pressure control, the 1-year and 5-year mortality rates decrease to 25 and 50 per cent, respectively. The mortality of untreated malignant phase hypertension (a hypertensive urgency rather than emergency) is around 80 per cent at 2 years, and if managed inappropriately there is also a high rate of progression to renal dysfunction, necessitating long-term dialysis, in addition to strokes and heart failure.

Clinical presentations

Hypertensive emergencies occur most commonly in patients with previous hypertension, especially if inadequately managed. Nevertheless, some patients can present with hypertensive emergencies *de novo*, without any previous history of hypertension.

Very severe and malignant hypertension are more likely to be associated with underlying causes such as renovascular disease, primary renal diseases, phaeochromocytoma, and connective tissue disorders, but malignant hypertension complicating primary hyperaldosteronism (Conn's syndrome) is very rare. Approximately 50 per cent of patients with malignant hypertension have an underlying cause.

Pathophysiology

The common denominator in hypertensive emergencies is intense peripheral vasoconstriction, resulting in a rapid rise in blood pressure and a vicious circle of events, including ischaemia of the brain and peripheral organs. This ischaemia stimulates neurohormone and cytokine release, exacerbating vasoconstriction and ischaemia, further increasing blood pressure and resulting in target organ damage. In addition, myointimal proliferation in the vasculature may exacerbate the situation, as can disseminated intravascular coagulation. Also, renal ischaemia leads to activation of the renin–angiotensin system, causing further rise in blood pressure and microvascular damage.

With mild to moderate elevation of blood pressure, the initial response of the vasculature is arterial and arteriolar vasoconstriction. Thus autoregulation maintains tissue perfusion at a relatively constant level and prevents the raised blood pressure from damaging the smaller, more distal blood vessels. Later, arteriolar hypertrophy also minimizes the transmission of pressure to the capillary circulation. In chronic hypertension, the lower limit of autoregulation of cerebral blood flow is shifted towards higher blood pressures, with impairment of the tolerance to acute hypotension. In normotensive subjects, the upper limit of autoregulation can be a mean arterial pressure of 120 mmHg (equivalent to 160/100 mmHg), but in individuals whose vessels are hypertrophied by long-standing hypertension, this upper limit is substantially higher (Fig. 1 and Plate 1). However, with rapid and severe rises in blood pressure, this process of autoregulation fails, leading to a rise in pressure in the arterioles and capillaries, causing vascular damage. Disruption of the endothelium allows plasma constituents (including fibrinoid material) to enter the vessel wall, narrowing or obliterating the lumen in many tissue beds. The level at which fibrinoid necrosis occurs is dependent upon the baseline blood pressure. In the cerebral circulation this can lead to the development of cerebral oedema and the clinical picture of hypertensive encephalopathy.

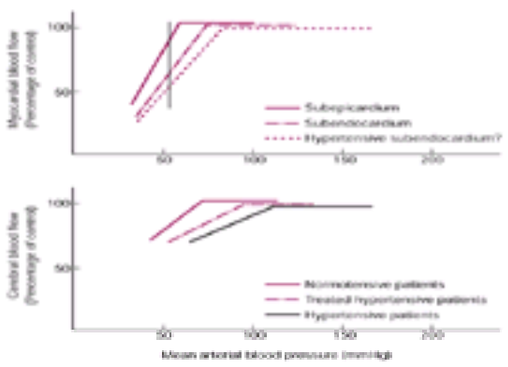

Fig. 1 Autoregulation of myocardial and cerebral blood flow in normotensive and hypertensive patients. (Reproduced from Strandgaard S, HaunsØ S (1987). Why does antihypertensive treatment prevent stroke but not myocardial infarction? *Lancet* **ii**, 658–60, with permission.)

In addition to protecting the tissues against the effects of hypertension, autoregulation maintains perfusion during the treatment of hyper-tension via arterial and arteriolar vasodilatation. However, falls in blood pressure below the autoregulatory range can lead to organ ischaemia, and the arteriolar hypertrophy induced by chronic hypertension means that target organ ischaemia will occur at a higher pressure than in previously normotensive subjects.

Malignant hypertension

The malignant phase of hypertension is a rare condition characterized by very high blood pressures, with bilateral retinal haemorrhages and/or exudates or cotton wool spots, without the added requirement for papilloedema (Fig. 2). This clinical definition of malignant hypertension includes both Keith, Wagener, and Barker grades 3 and 4 retinopathy, previously designated as 'accelerated' and 'malignant' hypertension, respectively. Differentiation between Keith, Wagener, and Barker grades 3 and 4 retinopathy has been shown to be unhelpful, as the presence of papilloedema is an unreliable sign. Furthermore, both categories carry an equally bad prognosis and should therefore be considered the same disease—'malignant' hypertension.

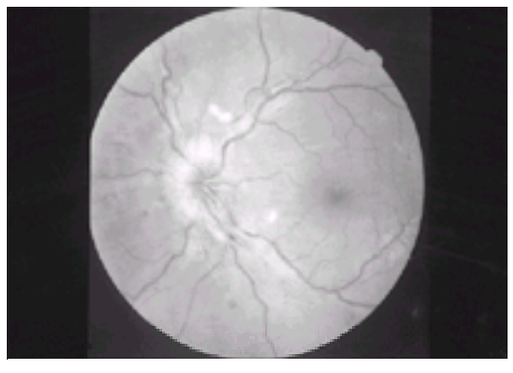

Fig. 2 Ocular fundus in hypertension, showing papilloedema, exudates, and a few haemorrhages. (See also Plate 1.)

The pathophysiological definition of malignant hypertension is based on the histological hallmark of fibrinoid necrosis of arterioles in many tissues, particularly the kidney. The histological changes are broadly similar to those seen in the haemolytic–uraemic syndrome or scleroderma. Mucoid intimal proliferation in renal interlobular arteries and ischaemic collapse of the glomerular tufts may also be seen. In black patients, myointimal hyperplasia is a common finding. The consequent intrarenal vascular disease leads to ischaemia of the juxtaglomerular apparatus and activation of the renin–angiotensin system with further vasoconstriction and wall damage, as well as exacerbation of the hypertension.

Epidemiology

Malignant hypertension has been reported to be becoming rarer in some countries, particularly amongst white populations. However, malignant hypertension still remains a common problem in the Third World and in other populations with health and social deprivation, where it is an important cause of endstage renal failure. Furthermore, in west Birmingham in the United Kingdom, the incidence of malignant hypertension was found to be around 1 to 2 per 100 000 population per year, with no clear reduction between 1970 and 1993 in the number of new cases seen, the mean duration of known hypertension before presentation, presenting blood pressures, or the number of antihypertensive drugs that were being used.

Whilst essential hypertension is usually the most common underlying cause of malignant hypertension in adults, secondary causes are more prevalent among younger patients. In children (aged less than 16 years) with malignant hypertension, parenchymal renal disease is the commonest cause (63 per cent), with 33 per cent having renovascular hypertension (aortoarteritis and fibromuscular dysplasia), and only 5 per cent with essential hypertension.

There is an association between cigarette smoking and malignant hypertension, which remains unexplained. Very rarely, the oral contraceptive pill may be implicated, consistent with the well-recognized increase in blood pressure in some women taking the 'combined' oestrogen/progesterone oral contraceptive pill. It is uncertain whether oral contraceptives di-rectly cause hypertension, or whether they simply exaggerate a tendency in women who already have a propensity to raised blood pressure. Malignant hypertension may also occur in the elderly, and is more common in Afro-Caribbean than white Caucasian and Indo-Asian populations. Possible reasons for the higher proportions of black and Asian people include the relative resistance of black patients to some antihypertensive therapies and perhaps poorer drug compliance.

One reason for the failure of malignant hypertension to decline in some centres may be inadequate medical screening facilities among poorly educated people with a limited understanding of the nature of the disease and the need to comply with antihypertensive therapy. Any reduction in the incidence of malignant hypertension may be because of the increasing use of drug therapy in the milder grades of hypertension preventing progression to the malignant phase. Nevertheless, it is possible that there has been no real decline in malignant hypertension, but merely a failure to recognize this life-threatening condition.

Clinical features

Some patients with malignant hypertension remain asymptomatic, but others present at a late stage of their disease. This proportion ranged from 10 to 75 per cent in one series from Nigeria. The predominant presenting symptom is visual disturbance with or without headaches. In the west Birmingham series, the presenting mean systolic and diastolic blood pressures have remained surprisingly similar over the 24 years surveyed (average blood pressure 228/142 mmHg), despite improvements in antihypertensive therapy. Heart failure, angina, or myocardial infarction are complicating features in approximately 20 per cent of patients with malignant hypertension, and ECG shows a high proportion of patients to have cardiomegaly and left ventricular hypertrophy. Nevertheless, some patients do have normal chest radiographs, ECGs, or echocardiograms despite very high blood pressure, suggesting that hypertension may have been of acute onset.

Investigation of malignant hypertension

All patients with malignant hypertension need a detailed clinical history and examination and investigation with blood tests (full blood count, serum biochemistry—including electrolytes and renal function), 12-lead ECG, chest radiography, and urinalysis. Fundoscopy and retinal photography are mandatory. The kidneys should be imaged by abdominal ultrasound to assess renal size and appearance, with a low threshold for proceeding to renal angiography to look for renal artery stenosis if the kidneys are asymmetric. A 24-h urine collection is necessary for urine catecholamines and protein excretion in all patients. These initial screening tests serve to identify patients in whom additional investigations may be appropriate to detect a secondary cause of hypertension.

The full blood count and film may reveal the anaemia of chronic renal failure or occasionally a microangiopathic haemolytic anaemia, with red cell fragmentation and intravascular coagulation, possibly related to the degree of arteriolar fibrinoid necrosis. Serum urea or creatinine should initially be measured daily, and if stable, creatinine clearance may be measured to give a more precise estimate of renal function. Raised serum creatinine or urea, indicating renal impairment, may have significant prognostic implications. Mild hypokalaemia may be present, due to secondary hyperaldosteronism. This usually resolves after control of the hypertension. Only very rarely does hypokalaemia indicate primary hyperaldosteronism (Conn's syndrome), but if it is extreme or persists despite good blood pressure control, then the characteristic findings of low renin levels but high aldosterone concentrations may be present. More commonly, both plasma renin and aldosterone levels are high in malignant hypertension, usually attributed to juxtaglomerular ischaemia. Urinalysis may demonstrate proteinuria and haematuria, even in the absence of primary renal disease, but the presence of proteinuria is a poor prognostic sign. Inflammatory markers (erythrocyte sedimentation rate and C-reactive protein) are often modestly elevated in malignant hypertension, but measurement of autoantibodies (antinuclear antibodies and antineutrophil cytoplasmic antibodies) can be used to discern uncommon cases due to vasculitis. Renal biopsy is required to make a specific diagnosis in some instances, but should not be performed until after blood pressure is controlled. The chest radiograph may show cardiomegaly and the presence of pulmonary oedema. Cardiomegaly and the presence of left ventricular hypertrophy can also be assessed using echocardiography.

Retinopathy in malignant hypertension

The most widely used classification of hypertensive changes in the fundus is that of Keith, Wagener, and Barker (Table 2). The strength of this classification was the correlation between clinical findings and prognosis, where grades 3 and 4 had a poor prognosis compared with grades 1 and 2. Although widely used, this grading system has some limitations, and has been made obsolete by advances in the understanding of the pathophysiology of arterial hypetension and the availability of

effective antihypertensive therapy. This and other traditional grading systems have therefore become less applicable to clinical practice than previously. The ophthalmoscopic grading can be simplified into two workable groups: grade A (non-malignant)—arteriolar narrowing and focal constriction, which also correlate with age and general cardiovascular status as well as blood pressure; and grade B (malignant)—linear flame-shaped haemorrhages, and/or exudates, and/or cotton wool spots with or without disc swelling.

Similar retinal appearances with haemorrhages and papilloedema can occur in severe anaemia, connective tissue disease, and infective endocarditis. Benign intracranial hypertension may cause bilateral papilloedema, but is usually self-limiting and minimally symptomatic. Nevertheless, severe hypertension and lone bilateral papilloedema may be a variant of malignant hypertension, with similar clinical features and prognosis. The retinal features of malignant hypertension regress over a period of 2 to 3 months if good blood pressure control is achieved.

The kidney in malignant hypertension

Renal involvement in malignant hypertension has been referred to as malignant nephrosclerosis, manifest as haematuria, proteinuria, and (sometimes) acute renal failure. Renal failure was the commonest cause of death in the west Birmingham series, and presenting urea and creatinine levels were independent predictors of survival. The Aberdeen Hypertension Clinic also found that serum creatinine at referral was an independant predictor of survival.

When antihypertensive therapy is initiated and blood pressure control achieved, the effect on renal function is variable. In the short term, renal function stabilizes in 10 per cent of cases, deteriorates progressively in 30 per cent, and deteriorates transiently before improving over a matter of weeks in the remainder. Isles and coworkers have suggested that the renal outcome of patients with malignant hypertension can be considered in three groups, each with a different renal prognosis: (i) patients whose serum creatinine is less than 300 µmol/l at presentation, who do well with effective antihypertensive therapy; (ii) patients with chronic renal failure (serum creatinine greater than 300 µmol/l) who do not require renal dialysis immediately, but are unlikely to maintain or recover renal function, except possibly in the short term, and commonly progress to endstage renal failure; and (iii) a small group with acute renal failure. Some of these patients may have post-streptococcal acute nephritic syndrome, characterized by retinopathy, fluid retention, and usually complete renal recovery. Thus, in the long term, some patients with mild renal impairment at first presentation may improve or even regain normal renal function as fibrinoid necrosis heals, especially with good blood pressure control. This is unlikely to occur in patients with more severe renal impairment at presentation.

There are varying reports of the frequency of renovascular disease in malignant hypertension, and this variation may be due to the frequency of renal angiography. In older patients, renal artery stenosis is more likely to be due to atheromatous disease which itself may be a consequence of chronic hypertension and chronic hyperlipidaemia, as well as cigarette smoking. In younger patients and particularly in women, renal artery stenosis may be due to fibromuscular dysplasia of the renal arteries with the characteristic 'string of beads' appearance on renal angiography. The value of surgical or angioplastic correction of atheromatous disease is debatable, possibly producing no better results than effective blood pressure control with antihypertensive drugs. In patients with fibromuscular dysplasia, however, renal angioplasty is worthwhile and may sometimes lead to a normal blood pressure level.

Practical guidelines for management of malignant hypertension

All patients with malignant hypertension should be admitted for assessment, investigation, and commencement of therapy under supervision. The initial aim of treatment is to lower the diastolic pressure to about 100 to 105 mmHg over a period of 2 to 3 days, with oral therapy and dose escalation at daily intervals if necessary. The maximum initial fall in blood pressure should not exceed 25 per cent of the presenting value. Blood pressure should be measured 4 hourly. Gradual reduction will allow adaptation of disordered tissue autoregulation and avoid target organ ischaemia. More aggressive antihypertensive therapy is both unnecessary and dangerous as it may reduce the blood pressure to below the autoregulatory range, leading to ischaemic events such as strokes, heart attack, or renal failure.

The first-line oral antihypertensive agent is either a short-acting calcium antagonist (such as nifedipine) or a b-blocker (such as atenolol). An appropriate dose of nifedipine is 10 to 20 mg of the tablet formulation, which can be repeated or increased as necessary to bring about gradual reduction in blood pressure. Nifedipine is not absorbed from the oral mucosa, and there have been reports of complications including visual loss, cerebral infarction, and myocardial infarction with nifedipine therapy using the short-acting sublingual capsules. Sublingual nifedipine produces unpredictable falls in blood pressure and should never be used. b-Blockers are useful alternatives, but should be avoided in patients with asthma or where there is a high suspicion of an underlying phaeochromocytoma. It is sensible to start with small doses, such as 25 mg of atenolol, increasing as necessary. The combination of oral atenolol and nifedipine is often a well tolerated and effective regime.

Diuretics should be restricted to those with evidence of fluid overload. Some patients are volume depleted, presumably secondary to a pressure-related diuresis and activation of the renin–angiotensin system. Captopril and the other ACE inhibitors can produce rapid and dangerous falls in blood pressure, particularly in patients with hypokalaemic secondary hyperaldosteronism and hyponatraemia secondary to juxtaglomerular ischaemia or renovascular disease, which may be unrecognized in the acute situation. Over a period of about 1 to 2 weeks, further antihypertensive drugs should be added in to achieve a gradual reduction of blood pressure to less than 140/85 mmHg. Triple or quadruple drug regimens are invariably necessary in the long term.

Prognosis

If malignant hypertension is left untreated, around 80 per cent of patients die within 2 years, hence the name. The importance of early diagnosis is emphasized as patients tend to develop clinical symptoms only at a late stage of their disease. Black male patients with malignant hypertension have a poor prognosis when compared with other ethnic groups or women. These patients also present with more severe hypertension and greater renal damage, which represent independent predictors of outcome and explain the poorer prognosis.

The advent of effective and tolerable antihypertensive drug therapy has meant that this prognosis is greatly improved. For example, in the west Birmingham malignant hypertension series, median survival times for the patients presenting before 1970, between 1970 and 1979, and between 1980 and 1989 were 39.2, 68.6, and 144.0+ months, respectively, suggesting an improvement in prognosis with more recently diagnosed patients (Fig. 3). The series by Scarpelli and coworkers reported a 12-year survival rate of about 69 per cent, although patients with malignant hypertension diagnosed after 1980 had a 100 per cent survival rate. Whatever the cause, progressive renal impairment is still a common complicating factor in malignant hypertension, with many patients needing dialysis in the long term.

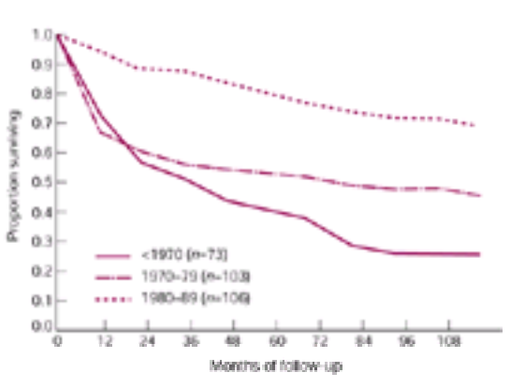

Fig. 3 Prognosis in 282 patients with malignant hypertension, divided into cohorts depending on year of presentation. (Reproduced from Lip GYH, Beevers M, Beevers DC (1995). Complications and survival of 315 patients with malignant-phase hypertension. *Journal of Hypertension* **13**, 915–24, with permission.)

Hypertensive left ventricular failure

Hypertension causes heart failure by a number of mechanisms: these include pressure overload on the heart due to the raised peripheral vascular resistance, reduced left ventricular compliance (for example, in left ventricular hypertrophy), an increased risk for coronary artery disease and the precipitation of cardiac arrhythmias (such as atrial fibrillation). Severe hypertension results in a significant increase in afterload and may result in decompensation of the failing heart.

In addition to the conventional management with opioids and loop diuretics, in very severe hypertension with marked pulmonary oedema, intravenous sodium nitroprusside may be necessary to reduce preload and afterload. Nitrates may also be used, but are less potent than sodium nitroprusside. ACE inhibitors should be considered in these patients only after stabilisation. These drugs are well-established to be life-saving in patients with left ventricular systolic impairment, and in addition they lead to long-term regression of LVH, which may also improve heart failure secondary to diastolic dysfunction.

Hypertensive encephalopathy

Hypertensive encephalopathy refers to the presence of signs of cerebral oedema caused by breakthrough hyperperfusion following severe and sudden rises in blood pressure. There is failure of autoregulatory vasoconstriction with focal or generalized dilatation of small arteries and arterioles. This leads to high cerebral blood flow, dysfunction of the blood–brain barrier, and the formation of brain oedema, which is thought to cause the clinical symptoms. The condition is very rare, and it is essential to ensure that this hypertensive emergency is distinguished from other neurological syndromes associated with high blood pressure, including intracerebral or subarachnoid haemorrhage, ischaemic stroke, or lacunar infarction.

Hypertensive encephalopathy is usually associated with a history of hypertension which has been inadequately treated, or where previous treatment has been discontinued. The condition is characterized by the insidious onset of headache, nausea, and vomiting, followed by visual disturbances and fluctuating non-localizing neurological symptoms such as restlessness, confusion, and, if the hypertension is not treated, seizures and coma. Severe retinopathy is frequently present, but not always.

The cerebrospinal fluid is usually normal but is at an increased pressure. The electroencephalogram may show variable transient, focal, or bilateral abnormalities. The CT scan or MRI may demonstrate white matter oedema: one of these tests is mandatory to exclude cerebral haemorrhage or infarction. Indeed, the increased use of CT scanning has demonstrated that almost all patients who appear to have hypertensive encephalopathy have cerebral infarction or haemorrhage with surrounding oedema and space-occupying cerebral symptoms.

Sodium nitroprusside is the drug of choice for genuine hypertensive encephalopathy but is not usually given if there is a cerebral infarct or haemorrhage. Parenteral labetalol and nitrates have also been used successfully. Rarely, diazoxide and hydralazine have been used, but they can cause precipitate and life-threatening acute falls in blood pressure. Sublingual nifedipine capsules should never be used (see above).

Severe pre-eclampsia and eclampsia are discussed in detail elsewhere. They may present with clinical features similar to hypertensive encephalopathy and treatment is broadly similar with antihypertensive drugs, magnesium sulphate, and early delivery of the fetus.

Hypertension with unstable angina or acute myocardial infarction

In a patient presenting with unstable angina or acute myocardial infarction and severe hypertension, a 'true' hypertensive emergency, such as aortic dissection, must first be ruled out. The risk of bleeding and stroke is significantly increased if anticoagulation with heparin or thrombolytic therapy is administered.

The appropriate initial treatment of patients with severe hypertension (greater than 180/110 mmHg) and an acute coronary syndrome should include the initiation of intravenous nitrates, with intravenous labetalol, sodium nitroprusside, or nicardipine as alternatives. The reduction of blood pressure should not be too abrupt, and (as with malignant hypertension) a gradual reduction is recommended, so that further myocardial or brain ischaemia is avoided. Sublingual nifedipine, which was once considered as a first-line drug, should not be used, in view of the negligible oral absorption and unpredictable hypotensive effects from later gastric absorption. Once the blood pressure is adequately controlled (less than 180/110 mmHg), anticoagulation or thrombolytic therapy can then be administered.

Hypertension with acute stroke

It is common to find modestly elevated blood pressure in patients admitted to hospital following an acute stroke. Cerebral autoregulation is commonly disturbed in this situation and excessive antihypertensive treatment may only serve to worsen the cerebral damage resulting from intracerebral infarction or haemorrhage. Such treatment should only be administered for severe elevation of blood pressure (diastolic blood pressure greater than 130 mmHg). In these cases, oral therapy with small doses of nifedipine or atenolol may be required. Parenteral treatment or sublingual nifedipine is almost always contraindicated. The calcium antagonist nimodipine has beneficial effects on cerebral vasospasm following subarachnoid haemorrhage, but these effects are not related to the small fall in blood pressure with this drug.

Severe hypertension after a stroke is a risk factor for further stokes and long-term treatment is worthwhile. It is unclear whether the immediate treatment of mild hypertension is of benefit. The role of antihypertensive medication before, during, and after a stroke can therefore be summarized as follows.

1. Before a stroke: it is of benefit to have blood pressure reduced to below 140/85 mmHg, as stroke prevention can be achieved.
2. During a stroke: it is detrimental to have hypertension treated aggressively, in view of the disordered cerebral autoregulation.
3. After a stroke: the role of antihypertensive medication remains unanswered, as the value of long-term antihypertensive therapy in mildly hypertensive patients following a stroke remains uncertain. Ongoing trials will answer this question. Nevertheless, drug therapy should be prescribed if blood pressure exceeds 160/100 mmHg.

The management of blood pressure in a patient with aortic dissection

The detailed presentation, diagnosis, and treatment of aortic dissection is discussed elsewhere. On suspicion of the diagnosis, whether or not surgery is indicated, all patients should be treated pharmacologically to reduce the systolic blood pressure to around 110 mmHg and the heart rate to 60 to 70 beats/min, thus reducing the force of systolic ejection to reduce aortic shear stress and limit the size of the dissection. Labetalol is an effective agent, or alternatively, sodium nitroprusside in conjunction with a b-blocker may be used. Patients should have haemodynamic monitoring with an arterial line and a Swan–Ganz catheter in position. Diagnostic tests are then performed on an urgent basis to confirm the dissection, identifying whether the ascending aorta is involved and defining any vascular abnormalities resulting from the dissection.

Summary of drug treatment options for hypertensive urgencies and emergencies

In uncomplicated malignant hypertension, where acute target organ damage is absent, and in uncomplicated severe hypertension, immediate blood pressure reduction with parenteral drugs is not indicated and blood pressure should be gradually reduced with oral agents (Table 3). Thus malignant hypertension, pre-eclampsia, and very severe hypertension without end-organ damage can be classified as hypertensive 'urgencies' rather than emergencies. Parenteral drugs to lower blood pressure may be dangerous and sublingual or capsular nifedipine should never be used.

In hypertensive crises the high blood pressure is directly responsible for a pressing clinical problem (hypertensive encephalopathy, left ventricular failure, or aortic dissection) and controlled reduction with parenteral treatment over a matter of hours is needed (Table 4). All such patients should be admitted to a high dependency or intensive care unit for monitoring. Blood pressure should be reduced by 25 per cent over several hours, depending on the clinical situation, usually with a target diastolic blood pressure of less than 100 to 110 mmHg. Thus, the goal of blood pressure reduction in those with hypertensive emergencies as well as urgencies is not to return blood pressure to a normal value immediately.

While nitroprusside is used in most hypertensive emergencies, its metabolism to cyanide, possibly leading to the development of cyanide or rarely thiocyanate toxicity, may be a limitation, especially in children. Toxicity is manifest by clinical deterioration, altered mental status, and lactic acidosis, and can be fatal; the risk of toxicity is increased with prolonged treatment (more than 24 to 48 h), underlying renal insufficiency, and high doses (greater than 2 µg/kg.min). An infusion of sodium thiosulphate can be used in affected patients to provide a sulphur donor to detoxify cyanide into thiocyanate.

Other parenteral agents such labetolol, hydralazine, and diazoxide are alternatives. Phentolamine is used only in patients with severe hypertension due to increased catecholamine activity, such as that seen in phaeochromocytoma, or after tyramine ingestion in a patient being treated with a monoamine oxidase inhibitor. Direct vasodilators such as diazoxide or hydralazine require concurrent b-blocker administration to minimize reflex sympathetic stimulation and are rarely used. ACE

inhibitors are best avoided in the early stage as they may, even in very low dose, cause precipitate falls in blood pressure and life-threatening reduction in cerebral perfusion. These rapid falls occur when patients are fluid depleted due to diuretic therapy or in renal artery stenosis.

Final summary

Hypertensive emergencies and urgencies carry a poor short- and long-term prognosis unless adequately managed. Initially, reduction of blood pressure to a normal value is dangerous, but in the long term, blood pressure should be reduced to 140/85 mmHg.

Further reading

Ahmed MEK *et al.* (1986). Lack of difference betwen malignant and accelerated hypertension. *British Medical Journal* **292**, 235–7.

Bloxham CA, Beevers DG, Walker JM (1979). Malignant hypertension and cigarette smoking. *British Medical Journal* **i**: 581–3.

Clough CG, Beevers DG, Beevers M (1990). The survival of malignant hypertension in blacks, whites and Asians in Britain. *Journal of Human Hypertensionn* **4**, 94–6.

Elliot JM, Simpson FO (1980). Cigarettes and accelerated hypertension. *New Zealand Journal of Medicine* **91**, 447–9.

Gudbrandsson T *et al.* (1979). Malignant hypertension. Improving prognosis in a rare disease. *Acta Medica Scandinavica* **206**, 495–9.

Harvey JM *et al.* (1992). Renal biopsy findings in hypertensive patients with proteinuria. *Lancet* **340**, 1435–6.

Isles C *et al*. (1979). Excess smoking in malignant-phase hypertension. *British Medical Journal* **i**: 579–81.

Isles CG, McLay A, Boulton Jones JM (1984). Recovery in malignant hypertension presenting as acute renal failure. *Quarterly Journal of Medicine* **212**, 439–52.

Islim IF *et al.* (1993). Prevalence of electrocardiographic left ventricular hypertrophy in malignant hypertension and its correlation with renal function. *Journal of Hypertension* **11** (Suppl 5), S106–S107.

Jhetam D *et al.* (1982). The malignant phase of essential hypertension in Johannesburg blacks. *South African Medical Journal* **61**, 899–902.

Kadiri S, Olutade BO (1991). The clinical presentation of malignant hypertension in Nigerians. *Journal of Human Hypertension* **5**, 339–43.

Keith NM, Wagener HP, Barker NW (1939). Some different types of essential hypertension: their course and prognosis. *American Journal of the Medical Sciences* **196**, 332–43.

Kincaid Smith P (1985). What has happened to malignant hypertension? In: Bulpitt CJ, ed. *Handbook of hypertension*, Vol 6, *Epidemiology of hypertension*, pp 255–65. Elsevier, Amsterdam.

Kincaid-Smith P (1991). Malignant hypertension. *Journal of Hypertension* **9**, 893–9.

Kumar P *et al.* (1996). Malignant hypertension in children in India. *Nephrology, Dialysis, Transplantation* **11**, 1261–6.

Ledingham JGG, Rajagopalan B (1979). Cerebral complications in the treatment of accelerated hypertension. *Quarterly Journal of Medicine* **189**, 25–41.

Leishman AWD (1959). Hypertension—treated and untreated: a study of 400 cases. *British Medical Journal* **i**: 1361–3.

Lim KG *et al.* (1987). Malignant hypertension in women of childbearing age and its relation to the contraceptive pill. *British Medical Journal* **294**, 1057–9.

Lip GYH *et al.* (1995). Severe hypertension and lone bilateral papilloedema: a variant of malignant phase hypertension. *Blood Pressure* **4**, 339–42.

Lip GYH *et al.* (1995). Malignant hypertension in the elderly. *Quarterly Journal of Medicine* **88**, 641–7.

Lip GYH, Beevers M, Beevers DG (1997). Does renal function improve following diagnosis of malignant phase hypertension? *Journal of Hypertension* **15**, 1309–15.

Mamdani BH *et al.* (1974). Recovery from prolonged renal failure in patients with accelerated hypertension. *New England Journal of Medicine* **291**, 1343–4.

McGregor E *et al.* (1986). Retinal changes in malignant hypertension. *British Medical Journal* **292**, 233–4.

Pitcock JA *et al*. (1976). Malignant hypertension in blacks. Malignant intrarenal arterial disease as observed by light and electron microscopy. *Human Pathology* **7**, 333–46.

Scarpelli PT *et al.* (1997). Accelerated (malignant) hypertension: a study of 121 cases between 1974 and 1996. *Nephrology* **10**, 207–15.

Shapiro LM, Mackinnon J, Beevers DG (1981). Echocardiographic features of malignant hypertension. *British Heart Journal* **46**, 374–9.

Strandgaard S, Paulson OB (1996). Antihypertensive drugs and cerebral circulation. *European Journal of Clinical Investigation* **26**, 625–30.

Veriava Y *et al.* (1990). Hypertension as a cause of end-stage renal failure in South Africa. *Journal of Human Hypertension* **4**, 379–83.

Webster J *et al.* (1993). Accelerated hypertension—patterns of mortality and clinical factors affecting outcome in treated patients. *Quarterly Journal of Medicine* **86**, 485–93.

Zampaglione P *et al*. (1996). Hypertensive urgencies and emergencies. Prevalence and clinical presentation. *Hypertension* **27**, 144–7.

15.17

Lymphoedema

Peter S. Mortimer

Introduction

Lymphoedema is swelling due to the accumulation of lymph within the tissues and results from a failure of lymphatic drainage. Lymphoedema differs clinically from other forms of chronic oedema by its altered skin texture and the brawny quality of the subcutaneous tissues, which limit pitting. There may be no distinguishing clinical features, particularly in the early stages of swelling.

It is the essential function of the lymphatic system to return to the plasma any proteins which escape from the blood circulation. Impairment of lymph drainage therefore causes the accumulation of protein as well as fluid since both enter the tissues in substantial amounts. The lymphatics act as a safety valve or buffer in the event of fluid overload in the tissues. Therefore, theoretically, oedema should not arise as long as lymph flow can respond to the increased 'lymph load'.

In addition, the lymph drainage pathways are responsible for the removal and processing of foreign organic material, such as microbes, as well as for the trafficking of immunologically active cells such as lymphocytes. Consequently, a predisposition to infection, which can be relapsing, is a common occurrence with lymphoedema. The elimination of dying and mutant cells is via the lymphatic system, but the mechanism by which lymph vessels attract and channel cancer cells is not understood.

Aetiology/pathophysiology

In practice, the clinician is faced with a problem of oedema, usually a swollen limb, for which the cause may not be obvious. Rather than assemble a list of specific medical conditions, such as heart failure and venous oedema, as differential diagnosis, it is more sensible to consider causes of oedema from physiological principles as there may be more than one reason for the swelling.

Oedema

Oedema is swelling due to the excessive accumulation of fluid within tissues. Interstitial fluid volume must increase by over 100 per cent before oedema is clinically detectable. Oedema develops when the capillary filtration rate exceeds the lymph drainage rate for a sufficient period. All oedema, whatever the cause, results from an imbalance between capillary filtration and lymph drainage. Therefore it follows that the pathogenesis of any oedema results from either a high filtration rate or a low lymph flow or a combination of the two.

Elevation of capillary pressure is usually secondary to chronic elevation of venous pressure caused by heart failure, fluid overload, or deep vein thrombosis. A rise in blood (arterial) pressure alone should not raise capillary pressure sufficiently to cause oedema. Reduced plasma colloid osmotic pressure, for example in hypoproteinaemia, raises net filtration rate. Changes in capillary permeability, for example due to inflammation, increase the escape of proteins into the interstitium and water follows osmotically. Any change to the Starling forces governing fluid exchange can influence oedema formation (Box 1).

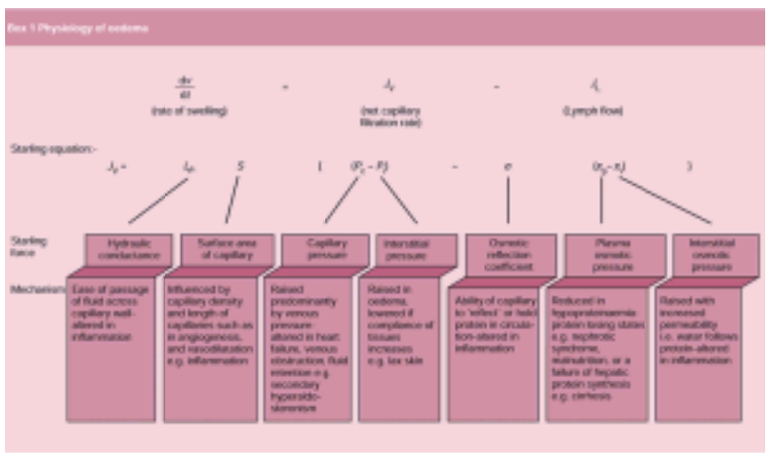

Most oedemas arise from increased capillary filtration overwhelming lymph drainage. To some extent any oedema incriminates the lymphatic system through its failure to keep up with demand (capillary filtration). Lymphoedema, however, is strictly oedema arising principally from a failure of lymph drainage.

Lymphoedema

Lymph drainage may fail either because of a defect intrinsic to the lymph conducting pathways (primary lymphoedema, Box 2(a)) or because of irreversible damage from some factor(s) originating from outside the lymphatic system (secondary lymphoedema, Box 2(b)).

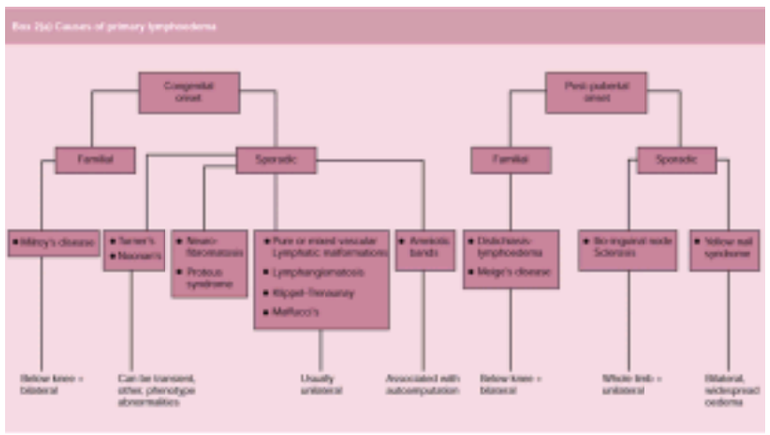

Physiologically there are only a limited number of ways that lymphatics can fail. They may be reduced in number (aplasia/hypoplasia), obliterated (lymphangitis), obstructed, lose contractility (pump failure), or become incompetent (valvular reflux). A lack of sensitive methods for investigation makes it difficult to distinguish between these mechanisms.

Primary lymphoedema

True congenital lymphoedema, which presents at or soon after birth, is rare. Milroy's disease is both congenital and familial: both features are required for the diagnosis. The fault is a hypoplasia of peripheral lymphatic vessels. The gene for Milroy's disease is vascular endothelial growth factor C receptor, which is known to be involved with lymphangiogenesis.

Congenital lymphoedema may present as a result of a developmental malformation either in pure form, such as lymphangiectasia, or in combination with a congenital vascular syndrome. Lymphoedema may coexist with cardiovascular malformations in other genetic syndromes such as Turner's and Noonan's.

Most forms of primary lymphoedema present at or after puberty with distal lower limb swelling, usually bilateral. Familial forms frequently occur (Meige's syndrome) in which lymphangiograms usually demonstrate a reduction in size and number of peripheral leg lymphatic (main collecting) vessels. Familial post-pubertal lymphoedema, associated with an extra row of ingrowing eyelashes (distichiasis–lymphoedema syndrome), seems to be associated with hypertrophy of leg lymphatics (increase in number and size of vessels). Mutations in *FOXC2*, a forkhead family transcription factor, have been found responsible.

Genetic forms of lymphoedema are usually (ultimately) symmetrical with predominantly distal limb swelling. Lymphoedema of the proximal obstructive type with unilateral whole leg swelling is sporadic in type. Lymphangiograms of this form of lymphoedema demonstrate obstruction at the iliac nodes, but no cause can be identified. It is of paramount importance to exclude serious pathology such as tumour or thrombosis. Whole leg swelling can sometimes be bilateral due to megalymphatics. Grossly dilated (varicose) lymphatics with valvular incompetence give rise to lymph and sometimes chylous reflux, indicating the primary fault lies with abdominal lymphatics.

Protein-losing enteropathy and hypogammaglobulinaemia may arise from intestinal lymphangiectasia. The yellow nail syndrome is characterized by yellow, non-growing, over-curved nails, pleural effusions, bronchiectasis, chronic sinusitis, and peripheral oedema, but the mechanism is unknown.

Secondary lymphoedema

Damage to lymph-conducting pathways may be secondary to outside influences. In developed countries, surgical removal or radiation (or both) of lymph nodes as a necessary part of cancer treatment probably represents the largest group. Trauma to lymphatics either from elective surgery or by accident usually needs to be extensive to induce lymphoedema. The experimental production of lymphoedema is extremely difficult owing to the efficient regenerative powers of lymphatics. It is probably the failure of lymphatics to regenerate or regrow through scarred or irradiated tissue which is responsible for lymphoedema following cancer therapy. Cancer rarely presents with lymphoedema, except in advanced disease, but relapsed tumour (after first-line treatment) frequently produces swelling because of infiltration of collateral lymphatic routes which hitherto have prevented oedema.

Filariasis is probably the most common cause of lymphoedema worldwide. Microfilariae are introduced into the skin by mosquitoes and migrate to the lymphatics. Adult worms develop within the main lymphatic vessels proximal to the nodes. Animal studies have shown that within days of infection, vigorous movement by adult worms directly impacts on the endothelial lining of the lymphatic trunks and indirectly distorts the local lymph-node architecture. The result is enlarged lymph nodes and lymphatics with thickened walls and valves, thrombus formation, and perilymphangitis.

Lymphangitis or cellulitis probably only cause lymphoedema when the lymphatics are perilously vulnerable. Any patient suffering recurrent lymphangitis/cellulitis in the same region is likely to have impaired lymph drainage. Proving which came first—the cellulitis or the lymphatic insufficiency—is difficult. Recurrent acute inflammatory episodes frequently lead to a stepwise deterioration in swelling.

Podoconiosis is a form of endemic non-filarial lymphoedema caused by particles of silica dust which penetrate the feet during barefoot walking. The microparticles are taken up by lymphatics and result in fibrosis.

Cellulitis (acute inflammatory episodes)

Evidence supporting the role of the lymphatic route in immunosurveillance was provided during the early days of organ transplantation. Lymphatics convey antigens to lymph nodes. Without intact lymphatics a primary immune response cannot develop. T lymphocytes constantly recirculate between central organs, blood, and tissues for the purpose of immunosurveillance. Disturbances in host defence mechanisms within the lymph drainage area in lymphoedema probably predispose to infections such as cellulitis. However, in some patients even prophylactic antibiotics fail to control recurrent attacks, thus questioning infection as the sole explanation.

Epidemiology

Filarial lymphoedema is a major health problem in tropical and subtropical areas where the disease is endemic. The number of infected people worldwide is estimated to be 750 million. Morbidity is high. Fever ('mumu' attacks) is the earliest symptom followed by enlarged lymph nodes. Damage to central abdominal nodes results in reflux of lymph or chyle into lower limbs and genitalia with swelling and seepage of lymph on to this skin surface (lymphorrhoea). Chyluria may also occur.

The cumulative prevalence of arm swelling following curative treatment for breast cancer has been reported at 28 per cent. Similar values are quoted for incidence in hospital series if surgery and radiotherapy are combined.

Primary lymphoedema is perceived as uncommon, yet chronic ankle oedema—particularly in the elderly—is considered common. Lymphatic insufficiency s likely to be a contributing cause because of the decline in lymph drainage function with age. Lymphoedema can be a difficult diagnosis in mild or early cases and may be underdiagnosed. Primary lymphoedema tends to affect females more than males. In less than 10 per cent of cases is swelling present at birth and most cases present at or soon after puberty. It is estimated that 80 per cent will present before the age of 35 years (lymphoedema praecox) and 10 per cent after the age of 35 years (lymphoedema tarda). In a study of 1000 normal young adults, 8 per cent of women demonstrated signs of lymphoedema in the lower limb.

Clinical features

Lymphoedema most commonly affects the extremities, particularly the lower limb. Midline swelling affecting head and neck or genitalia can be an isolated finding. Truncal oedema is often observed in the adjoining quadrant of the trunk to an affected limb because of the shared lymph drainage route.

Cells, proteins, lipids, and debris accumulate giving a 'solid' as well as a 'fluid' component to the swelling, hence the brawny texture to lymphoedema. Pitting occurs in

most patients and many early cases will pit readily.

History

Swelling frequently develops rapidly—overnight—but may be mild and intermittent at first. Pain may feature initially, prompting diagnoses such as deep vein thrombosis and soft tissue injury. With time, oedema becomes more permanent and painless, although discomfort, aching, and heaviness are common symptoms. Functional impairment is slight until swelling becomes more severe. The major problem is disfigurement (Fig. 1).

Fig. 1 Lymphoedema of the left lower limb exhibiting characteristic skin changes and loss of shape with folds developing around the ankle.

Lymphoedema does not usually respond to elevation or diuretics, except in the early stages or when it is compounded by increased capillary filtration. Chronic oedema that does not reduce significantly overnight is likely to be lymphatic in origin.

Clinical signs

Most swelling occurs in the subcutaneous layer, but it is the skin which exhibits most changes. It becomes thicker, as demonstrated by the Kaposi–Stemmer sign (a failure to pick up or pinch a fold of skin). Skin creases become enhanced and a warty texture develops (hyperkeratosis). Accumulation of stagnant lymph in the dermis can produce surface bulges resembling cobblestones that feel firm to the touch (papillomatosis). The resemblance of the skin texture to elephant hide explains the term elephantiasis. Dilated lymphatics (lymphangiectasia) can also bulge on the surface like a blister from which lymph can leak.

Lymphatic insufficiency has three major consequences: (i) swelling (oedema); (ii) a predisposition to infection, in particular cellulitis; and (iii) the uncommon complication of malignancy arising with the lymphoedema (Stewart–Treves syndrome).

Oedema

Most cases of primary lymphoedema present with bilateral but asymmetrical oedema of the feet, ankles, and lower legs (distal hypoplastic type). However, it may take months to years before oedema manifests in the contralateral limb. Whole limb swelling usually indicates a problem at the level of the regional lymph nodes such as following cancer treatment or in filariasis. Oedema that begins proximally and spreads distally suggests an obstructive process (proximal obstructive type). Abnormalities of central abdominal and thoracic lymphatics may produce lower limb swelling, often bilateral, with lymph or chyle reflux. Lymph and chyle effusions can be observed in the pleural, peritoneal, and pericardial cavities and rarely in joints.

Cellulitis (acute inflammatory episodes)

Acute inflammatory episodes describe the attacks of apparent infection, simulating cellulitis, which afflict patients with lymphoedema. A typical attack starts rapidly, often without warning. Fever, rigors, headache, and vomiting can occur, but patients usually feel as if influenza is starting. A feeling of heat, redness, and increased swelling occurs within the lymphoedematous area. Pain may precede the rash. Areas of clearly demarcated erythema with a migrating border, as seen in classic cellulitis, may not be observed, presumably because of rapid dissemination throughout the oedematous tissue. Sometimes the condition may 'grumble' in a rather chronic manner for days or weeks and only on complete recovery after prolonged antibiotics and rest can the diagnosis be made. A characteristic of lymphoedema is for acute inflammatory episodes to recur. Intervals may be more than 12 months or as short as 3 weeks.

Lymphangiosarcoma

Lymphangiosarcoma (Stewart–Treves syndrome) is the most serious but rarest complication of lymphoedema. Chronic lymphoedema is a major predisposing factor in the development of malignant vascular tumours irrespective of the cause of the lymphoedema. The clinical appearance is usually of a fixed purple or bruise-like discoloration within the lymphoedematous skin. Infiltrated plaques and nodules later appear.

Differential diagnosis of the swollen limb Box 3)

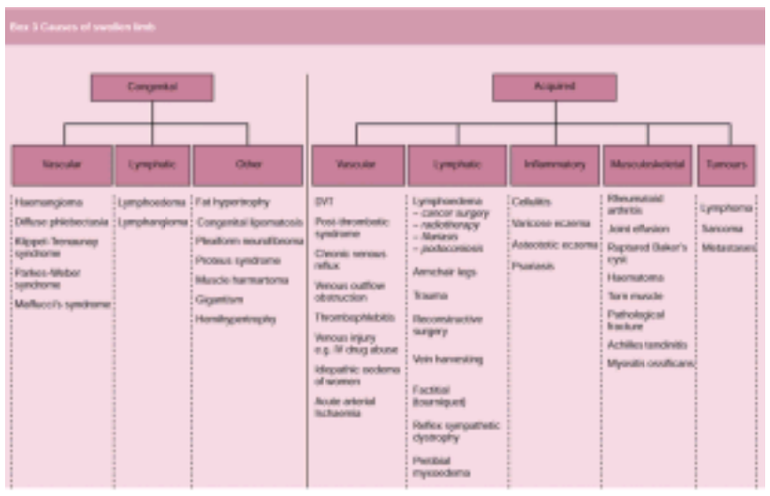

Both excessive capillary filtration and compromised lymph drainage frequent coexist.

'Venous' oedema

Most cases of chronic venous disease do not manifest oedema because of increased lymph flow in response to increased capillary filtration. This suggests that the development of oedema in post-thrombotic syndrome and venous ulceration is as much a failure of lymph drainage to compensate as it is due solely to overwhelming filtration. The expansion of the venous pool in the leg due to dilatation of veins will also contribute to an increase in limb girth independent of oedema.

'Armchair' legs

This syndrome refers to those patients who sit in a chair night and day with their legs dependent. Immobility results in minimal lymph drainage and 'functional lymphoedema' ensues, compounded by increased capillary filtration from gravitational forces. Predisposed are those patients suffering cardiac or respiratory failure

who cannot lie flat, those paralysed from strokes or spinal damage including spina bifida, and those with arthritis, particularly rheumatoid arthritis.

Lipoedema (lipidosis, lipodystrophy)

Frequently misdiagnosed as lymphoedema, lipoedema is peculiar to females with onset at or after puberty. A 'fatty', non-pitting swelling affects legs, thighs, and hips. Characteristic inverse shouldering occurs above the ankle because of sparing of the foot. The skin is soft, tender, and bruises easily. Pain may feature. Lipoedema is not influenced by dieting and is distinct from morbid obesity.

Investigation

The investigation of choice for confirming that oedema is primarily of lymphatic origin is lymphoscintigraphy (isotope lymphography). Conventional direct-contrast X-ray lymphography is now rarely undertaken to investigate lymphoedema. MRI, in preference to CT, is of value in identifying a cause for lymphatic obstruction.

Lymphoscintigraphy

A radiolabelled protein or colloid is administered via a subcutaneous injection and its movement through lymphatic vessels to nodes is monitored. The dynamics of lymph flow, as depicted by tracer removal from injection site and/or trapping in regional nodes, can be captured using a scintiscanner or g-camera. Off-line calculation of time–activity curves from regions of interest permit quantitative analysis of lymph drainage. A normal lymphoscintigram is shown in Fig. 2 and an abnormal one in Fig. 3.

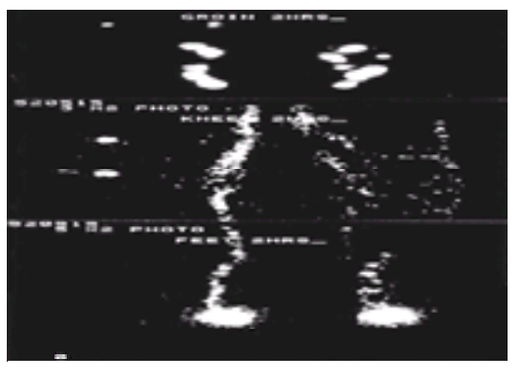

Fig. 2 A normal lymphoscintigram apart from some collateral lymph drainage in the left thigh. Following a web space injection of radiolabelled colloid (^{99m}Tc–antimony sulphide colloid) the transport of radioactivity is imaged by a g-camera. Measurement of radioactivity over the ilio-inguinal nodes for a given time interval can quantify lymph drainage function.

Fig. 3 A lymphoscintigram in a patient with Milroy's disease (congenital familial lymphoedema) demonstrates no migration of tracer in the affected right leg. The left leg (clinically normal) shows normal ilio-inguinal nodal uptake but extravasation of tracer where lymphatics are abnormal.

Direct-contrast X-ray lymphography (lymphangiography)

The technique requires the identification of a peripheral lymphatic (usually in the foot) by subcutaneous injection of a vital dye, such as patent blue. The oily contract medium Lipiodol is then administered into the canulated lymphatic, with subsequent imaging by X-ray as the contrast passes through the main limb lymphatics. The failure to identify a lymphatic with vital dye suggests lymphoedema with distal hypoplasia of vessels, particularly if the dye persists in the tissues for days. Lymphography is an invasive procedure which provides little functional information, but it remains the gold standard for delineating lymph vessel and node anatomy.

Magnetic resonance imaging

MRI (and CT) demonstrates a thicker skin and a characteristic 'honeycomb' pattern in the subcutaneous compartment. Following deep vein thrombosis in the leg the muscle compartment is enlarged, but this remains unchanged in lymphoedema.

Treatment

Physical therapy

This first-line approach is designed to stimulate lymph drainage within main or collateral lymph routes. Intermittent changes in tissue pressure are normally responsible for moving materials and fluid from tissue spaces into initial lymphatics. Lymphatic collecting vessels that possess smooth muscle then actively pump the lymph downstream. In lymphoedema, exercise, through dynamic muscle contractions undertaken while compression is applied to the skin surface, encourages movement of lymph through lymph vessels in a matter akin to external cardiac massage. In practice, this means fitting of compression hosiery or bandages and instruction on exercise. Overexertion and excessive static exercises, such as gripping, increase blood flow and can therefore increase oedema. Compression has the added benefit of opposing excessive capillary filtration should it coexist. Manual lymph drainage, a specific form of lymphatic massage, is added to stimulate lymph flow and redirect the lymph towards the functioning lymph nodes in an unaffected lymphatic basin.

In moderate to severe lymphoedema an intensive period of treatment using multilayer bandaging, exercise, and manual lymph drainage is used to reduce swollen limbs so that subsequent maintenance treatment with hosiery and exercise is more effective at controlling the condition.

Pneumatic compression therapy, such as Flowtron, softens and reduces limb volume during treatment but it is doubtful that there is any long-term benefit compared with hosiery (and exercise) alone.

Infection

Prevention of acute inflammatory episodes (cellulitis) is crucial to the control of lymphoedema. Care of the skin, good hygiene, treatment of any dermatitis or fungal infection, and antisepsis following minor wounds are important. Prompt administration of antibiotics, such as co-amoxiclav in a dose of 625 mg three times daily at the

onset of an attack, is mandatory (Box 4). For relapsing cellulitis, prophylactic antibiotics, preferably phenoxymethylpenicillin in a dose of 500 mg twice daily, are indicated for an indefinite period. Control of the oedema may help to reduce antibiotic requirements.

Box 4 Stepladder approach to treatment of acute inflammatory episodes (cellulitis)

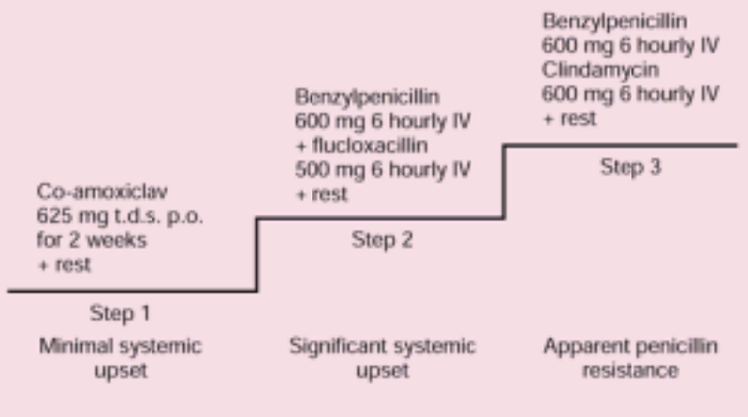

Drug therapy

Diuretics remain the most commonly used treatment but have very little benefit in established lymphoedema because their main action is to reduce capillary filtration. Improvement with diuretics suggests that the predominant cause of the oedema is not lymphatic. Diuretics, such as spironolactone, may be necessary in circumstances where there is substantial truncal lymphoedema, particularly if venous hypertension coexists.

The benzopyrone group of drugs, for instance rutosides and coumarin, have been advocated, but the clinical effect is minimal.

Surgery

Surgery can involve removal of excess tissue (reducing/debulking operations or liposuction) or bypassing of local lymphatic defects. The defective region may be bridged with omentum or an isolated segment of gut. Microsurgery such as lymphovenous anastomoses remains experimental. All surgery is followed by continued compression therapy, such as hosiery.

Prevention

Primary lymphoedema usually develops spontaneously without warning. A diagnosis of lymphoedema should be entertained if oedema is persistent. Lymphoscintigraphy is the investigation of choice at this stage as clinical signs are few in the early stages. In the future, identification of the genes programming for inherited lymphoedema may help predict those at risk.

Secondary lymphoedema can, in theory, be prevented by avoiding the cause or limiting the damage. Less intervention to lymph nodes for staging and treatment of cancer would help. Patients at significant risk of lymphoedema should receive advice from the appropriate professional, such as a breast-care nurse, and any oedema should be treated early, at a stage when it is more responsive to intervention and when it may even be reversible.

Further reading

Kaipainen A *et al*. (1995). Expression of the fms-like tyrosine kinase 4 gene becomes restricted to lymphatic endothelium during development. *Proceedings of the National Academy of Sciences, USA* **92**, 3566–70. [The first indication of the genetic basis of lymphangiogenesis.]

Karkkainen MJ *et al*. (2000). Missense mutations interfere with VEGFR-3 signalling in primary lymphoedema. *Nature Genetics* **25**, 153–9. [The first gene mutation found for lymphoedema.]

Kinmonth JB (1982). *Lymphatics, lymphology and disease of the chyle and lymph systems*, 2nd edn. Edward Arnold, London. [The original definitive clinical reference book.]

Ko DSC (1998). Effective treatment of lymphedema of the extremities. *Archives of Surgery* **133**, 452–8. [Open study but with robust data on physical therapy for lymphoedema.]

Levick JR (1995). *An introduction to cardiovascular physiology*, 2nd edn. Butterworth-Heinemann, Oxford. [Physiology made simple and meaningful.]

Roddie IC (1990). Lymph transport mechanisms in peripheral lymphatics. *News in Physiological Science* **5**, 85–9. [Understanding how lymph drainage operates.]

Stewart G *et al*. (1985). Isotope lymphography: a new method of investigating the role of lymphatics. *British Journal of Surgery* **72**, 906–9.

Yoffey JM, Courtice JM (1970). *Lymphatics, lymph and the lymphomyeloid complex*. Academic Press, New York. [A classic scientific review.]

15.18 Idiopathic oedema of women

J. Firth

Definition and diagnosis

In some women fluid retention occurs in the absence of any clear explanation and is termed idiopathic oedema. Since the condition typically fluctuates in severity from one time to another it is sometimes called cyclical or periodic oedema, but these terms mislead; first, because there is rarely any recognizable periodicity, and second, because the condition is not related to menstrual periods. Most women retain fluid just before the menses and lose this fluid immediately afterwards. Idiopathic oedema occurs most commonly in women aged 20 to 40 years, but has no clear relationship with the menstrual cycle and can persist after the menopause or oophorectomy.

The diagnosis of idiopathic oedema depends on the exclusion of other causes of oedema, including cardiac, hepatic, renal, allergic, or hypoproteinaemic disease, venous or lymphatic obstruction, and use of some medications. The role of diuretics, causally or in treatment, is contentious, as discussed below. However, it is always unsatisfactory when a diagnosis is made by exclusion of other conditions rather than on the basis of 'positive' criteria. Such criteria for the diagnosis of idiopathic oedema have not been universally agreed, although both Thorn and McKendry (see Kay *et al*. for discussion) have made proposals. These require evidence of substantial weight gain during the course of the day from morning to evening, with a figure of more than 1.4 kg often quoted, although this does not provide a clear-cut separation from normal. They also demand the presence of emotional or psychological factors. Many authors comment on the aggravation of swelling by prolonged sitting or standing, but this does not feature in the diagnostic criteria mentioned.

Clinical features

The patient's complaint is of swelling, which usually waxes and wanes but can be constant. In the morning the face and eyelids feel swollen and heavy. By the end of the day the areas worst affected are the hands, breasts, trunk, abdomen, thighs, ankles, and feet. Rings no longer fit the swollen fingers and undergarments and clothes can feel uncomfortably tight such that they have to be removed or replaced with something larger. The feet and ankles may be relatively spared, hence the disposition of oedema tends to be different from that in most other oedematous states, where it begins distally in the feet and ankles and progresses proximally.

Episodes or exacerbations of fluid retention often occur unpredictably, but obesity, emotional stress, and consumption of high-carbohydrate food are thought to be triggers in some. Sufferers are often mentally and physically lethargic during periods of fluid retention, frequently expressing the view that they feel bloated and ugly, even though this may not be apparent to the observer. Many appear to be emotionally labile or anxious and some are depressed, invariably (and perhaps correctly) claiming that this is secondary to the fluid retention. Other common symptoms include carpal tunnel syndrome, non-articular rheumatism, palpitations, non-ulcer dyspepsia, and headaches.

Aside from oedema, which may or may not be present at the time of medical assessment, examination is unremarkable, as are routine investigations for the cause of oedema. Those patients that have used diuretics may have a hypokalaemic hypochloraemic metabolic alkalosis.

Pathophysiology

The cause of idiopathic oedema is not known (by definition). Diurnal weight fluctuation of more than 1.4 kg is required for diagnosis, but weight may fluctuate from day to day by up to 4 or 5 kg. During periods of weight gain the patient may be oliguric, passing low volumes of urine in which there is little sodium (less than 20 mmol/l). The blood vessels of women with idiopathic oedema are more permeable to albumin, the fractional catabolic rate of albumin is increased, both intravascular and total body albumin pools are smaller, and the plasma volume decreases by more on standing than in normal controls. Activation of the sympathetic nervous system, renin–angiotensin–aldosterone system, and high levels of ADH in the plasma that are consistent with intravascular volume depletion have all been reported, as has reduction in dopaminergic activity. These changes provide a plausible explanation for why the kidney retains salt and water in idiopathic oedema, but the prime mover remains uncertain. They also form the background to postural water-loading or sodium-loading tests that have been advocated as diagnostic tools, although these are not used routinely in clinical practice. After similar loading on two separate occasions, patients with idiopathic oedema who remain upright throughout the test excrete less water or sodium than they do if they remain supine.

Many patients seen in hospital practice will already be taking diuretics or have taken them in the past, and some will be consuming large doses of loop agents every day. One influential study reported 10 such patients who started to take diuretics because of concern about swelling or their body weight and who continued to take them because cessation provoked rapid weight gain, facial bloating, and abdominal distension. When prevailed upon to stop diuretics they each gained weight (up to 5 kg), reaching a maximum in 4 to 10 days, but by 20 days 7 of the 10 had fallen to below their previous weight, and 9 of the 10 remained free of oedema over a long period of follow-up without taking diuretics. This led the authors to suggest that diuretic abuse might be the cause of all cases of idiopathic oedema. This view is not held by most with experience in the field, but rebound oedema on diuretic withdrawal can undoubtedly be an exacerbating feature, and it is appropriate to look for evidence of diuretic abuse if the patient denies taking such drugs and yet routine biochemical testing of blood and urine suggests the possibility (see Chapter 20.2.2 for further discussion).

Management

Women with idiopathic oedema frequently complain that doctors have not taken their condition seriously, and there is no doubt that it is a frustrating disorder for both patients and their physicians. Sympathetic explanation of the nature of the problem helps management.

If the patient is obese, then they should be given advice as to how to lose weight, and—independent of any effect on weight—some find that reducing dietary carbohydrate helps. They should be advised to avoid long periods of standing or sitting and to wear loose-fitting clothing, although most will have discovered these things for themselves. Avoidance of an excessive dietary intake of sodium is a sensible recommendation. On theoretical grounds the use of elastic stockings would also seem appropriate, since these might reduce the postural reduction in plasma volume seen in idiopathic oedema. However, few find that their benefits outweigh their disadvantages and it is difficult to get most patients to persist with them for long enough to see whether or not they really would be of help.

Diuretics are a real problem. It seems intuitively obvious to most patients and to many doctors that someone who is retaining fluid would benefit from a diuretic, hence many patients with idiopathic oedema end up on very large doses of loop agents, often combined with amiloride or spironolactone. Rather than helping, these may worsen symptoms of tiredness, lethargy, weakness, and dizziness by exacerbating intravascular volume depletion, and attempts to stop typically lead to rebound oedema. Explanation is the key here in that if patients recognize rebound oedema for what it is and relieve oedema with supine rest rather than renewed consumption of high doses of diuretics, then there is a reasonable chance that they can be weaned off diuretics with benefit.

Levodopa, carbidopa, bromocriptine, captopril, and a variety of other agents have been tried in idiopathic oedema, but none is of proven benefit.

Further reading

Kay A, Davis CL (1999). Idiopathic edema. *American Journal of Kidney Diseases* **34**, 405–23.

MacGregor GA, *et al*. (1979). Is 'idiopathic' edema idiopathic? *Lancet* **i**, 397–400.

Marks AD (1983). Intermittent fluid retention in women. Is it idiopathic edema? *Postgraduate Medicine* **73**, 75–83.

Sabatini S (2001). Hormonal insights into the pathogenesis of cyclic idiopathic edema. *Seminars in Nephrology* **21**, 244–50.

Streeten DH (1995). Idiopathic edema. Pathogenesis, clinical features, and treatment. *Endocrinology and Metabolism Clinics of North America* **24**, 531–47.

16.1 The clinical approach to the patient who is very ill

J. Firth

Introduction—recognizing the problem

As a young doctor, I vividly remember watching a senior physician at a teaching hospital endeavouring to take a history from a middle-aged man who looked grey and very unwell. The man was not giving lucid answers and the conversation seemed increasingly unlikely to lead to a useful conclusion. After a period of silence his breathing became extremely laboured, and within a minute he suffered a cardiac arrest from which he could not be resuscitated.

The first priority in the management of patients who are very ill is to recognize that this is the situation. It is a sensible discipline when dealing with emergency admissions (and sometimes in other contexts) to ask yourself the question: Is this patient well, ill, very ill, or nearly dead? Some physicians will recognize this intuitively, others will have to make a conscious effort, lest they make the sort of error described in the previous paragraph. In general the approach to the patient begins with the history, followed by the examination, and sometimes the ordering and appraisal of the results of investigations, before a diagnosis is reached and treatment commences. This approach can be fatal in those who are very ill or nearly dead at the start.

Key features—summarized as airway, breathing, circulation (ABC)—to assess immediately in the patient who is very ill or worse are shown in Table 1. If in any doubt, the most important of the questions to answer is: 'Do you think that this patient can keep breathing like this for the next 10 min?' If the answer is no, then you need to get help immediately: this will usually involve summoning someone directly from the intensive care unit, or putting out a 'cardiac arrest' call. It is better to do this 10 min before the heart stops than 2 min afterwards, a strategy emphasized by replacing the 'cardiac arrest team' with a 'medical emergency team' in one study. One doctor, with or without one nurse, is not enough to deal optimally with the patient who is *in extremis*.

Immediate management of airway and breathing

The immediate treatment priorities for the patient with cardiorespiratory collapse are shown in Table 2.

Airway and oxygen

If a patient is having problems breathing, then is there a difficulty in maintaining the upper airway? If a head tilt/chin lift manoeuvre is beneficial, then a gentle attempt to insert an oropharyngeal airway should be made. This should not be done against resistance—a fight is much more likely to do harm than good—and if the patient spits the airway out it almost certainly means that it is not necessary.

All patients who are extremely ill should be given oxygen in as high a concentration as possible by face mask. A fraction of inspired oxygen (FiO_2) of around 60 per cent can be obtained using a standard face mask with a reservoir bag and an oxygen flow rate of 10 litre/min. An obvious concern about this recommendation is that it may induce carbon dioxide retention in those who are prone to this (usually patients with chronic obstructive pulmonary disease), but the downside of denying someone oxygen can be substantial: hypoxia kills, hypercarbia merely intoxicates. If blood gas analysis shows that the patient is retaining carbon dioxide, the FiO_2 can gradually be reduced or elective ventilation can be used.

If a patient with respiratory difficulty has received opioids within the past 48 h, or could have done so, then give intravenous naloxone (0.2 to 0.4 mg in those who have received opioids; 0.8 to 2.0 mg repeated to a maximum of 10 mg in case of overdosage). This sometimes produces a dramatic response.

Elective ventilation

The patient should be electively intubated and ventilated if breathing seems to be failing despite the measures indicated above, although in some circumstances non-invasive ventilation may be an appropriate alternative. There is no substitute for wise clinical judgement in deciding when this should be done, and it is easy for the inexperienced to be led astray. Too soon is better than too late. A 'normal' respiratory rate of, say, 12 breaths/min may indicate normality, but is also compatible with near death in the patient with a severe respiratory problem who is becoming exhausted. A blood gas level that 'doesn't seem too bad', meaning perhaps a PO_2 of 9 kPa and PCO_2 of 5.5 kPa, which may not be a cause for any concern at all in a patient who is comfortable and breathing room air, is not at all reassuring if the patient is breathing 60 per cent oxygen and looks very tired.

The work of breathing accounts for up to one-third of the body's oxygen consumption, hence taking this burden from patients by sedating, paralysing, and ventilating them can have a dramatically beneficial effect, whatever the reason for their predicament. However, one note of caution: whilst being ventilated can be very helpful, the minute or two when the patient is being sedated, paralysed, and intubated is a time of very high risk, since the pharmacological agents used can, to varying degrees, induce profound hypotension culminating in a 'crash'. The chances of this happening can be reduced by giving the patient a bolus of fluid (a rapid infusion of 500 ml of 0.9 per cent saline or colloid) immediately before induction, with dilute adrenaline (1:10 000; not 1:1000) given as 1-ml intravenous pushes (up to one push/min) in the event of a dramatic fall in blood pressure. Whilst the anaesthetist is attending to the airway, the match-hardened physician will stand by a site of intravenous access with such a syringe in their hand and not skulk off into a corner.

Tension pneumothorax

If the patient has a tension pneumothorax, then this should be decompressed immediately by inserting a large-bore venous cannula into the chest in the second intercostal space in the mid-clavicular line and then withdrawing the stylet. The response to this is dramatic and satisfying. A chest drain with underwater seal can then be inserted at (relative) leisure.

The prospect of performing chest decompression is daunting for many junior physicians, and even more so for most of their senior colleagues. Remember that the physical signs are not subtle, the patient with '? minor shift of the trachea' as the only relevant sign does not have a tension pneumothorax. If the patient is blue and can't breathe, one side of the chest looks blown up and there are no breath sounds over it, then (after attending to the airway and oxygen and calling the cardiac arrest team) stick in the cannula. There is much to be gained from doing so, and little to be lost if it does not lead to improvement.

Immediate management of the circulation

The patient who is volume depleted

In a patient who is very ill, if the pulse rate is less than about 60 to 70/min, or above 120/min other than with a sinus tachycardia, then manoeuvres to speed it up (atropine, isoprenaline, pacing) or slow it down (DC cardioversion or antiarrhythmics, but avoiding any of the latter that are negatively inotropic excepting in rare

circumstances) are likely to improve the circulation. (See Chapter 16.3 and Chapter 15.6 for further information.)

Most patients who are very ill will have a sinus tachycardia, when there is no advantage in attempting to alter pulse rate and rhythm, indeed there is much to be lost from ill-advised attempts to do so. The key question then becomes: Is the filling pressure optimal? Does the patient need to be given fluid, or is there too much fluid on board? Those who do not see many very ill patients might think that it should be easy to tell the difference, but the answer is not always obvious, and yet the physician must decide rapidly, often without anything other than clinical judgement to guide them. Although they may yearn for a measurement of central venous or pulmonary capillary wedge pressure, or for a chest radiograph to see whether or not there is pulmonary oedema, to delay management might be fatal. The clinical features to look for are listed in Table 3, and the appropriate responses discussed in Table 4.

Obtaining venous access

The need is to insert a cannula into a decent-sized vein quickly, and with as low a risk of complication as possible. Try initially for a peripheral vein in the forearm or antecubital fossa, but if these are constricted and cannot be cannulated and the patient is *in extremis*, then go for the femoral vein, which lies medial to the artery in the groin (NAVY, nerve artery vein Y-fronts). The procedure is easiest if the patient can lie flat, but can be performed with the patient propped up if respiratory difficulty means that lying down is impossible. Feel for the femoral pulse just below the crease of the groin and, after giving local anaesthetic, insert the needle (with the bevel pointing forwards) one finger breadth medial to the point of maximum pulsation at an angle of about 60° to the skin and parallel to the long axis of the leg. The only significant complication of this procedure is inadvertent arterial puncture (Table 5), the consequences of which are much less likely to be severe in the groin than in the neck or below the clavicle.

An attempt to insert a central venous cannula into the internal jugular or subclavian vein of a patient who is *in extremis* has led to more deaths than such catheters have prevented. If the patient's intravascular volume is depleted, then the veins are constricted and small, cannulation is very difficult, and the procedure tends to degenerate into what is known in the trade as the 'sewing machine technique', where multiple stabs culminate in something being hit, often not the vein that was being (increasingly loosely) targeted. If patients are volume overloaded and in respiratory difficulty, then they will not tolerate being laid flat for the cannulation attempt: if you try to make them do so, they won't lie still for long, and if they do become still it might be because they have died.

When the patient who is volume depleted has had intravascular volume restored, or when the situation of the patient who is volume overloaded and breathless has been rendered safe (perhaps by intubation and ventilation), then the insertion of a central venous catheter can be helpful in diagnosis (for example to allow passage of a right heart catheter for measurement of the pulmonary capillary wedge pressure), monitoring (for example fall in the central venous pressure indicating further gastrointestinal haemorrhage), and treatment (for example infusion of inotropes, other drugs, or parenteral nutrition). The approaches for internal jugular and subclavian vein cannulation are shown in Fig. 1 and Fig. 2, with details of safety and reliability in Table 5.

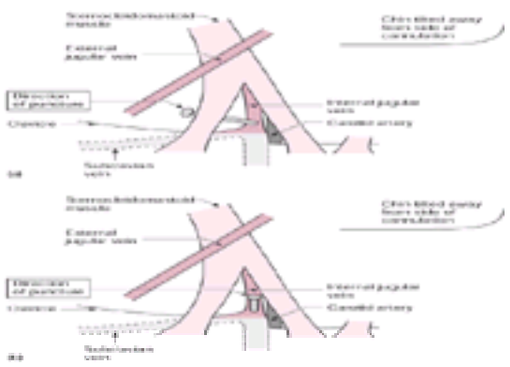

Fig. 1 The low lateral (a) and axial (b) approaches to the internal jugular vein. (a) The patient is supine with the head turned away from the side of the puncture. A towel may be placed under both shoulders to extend the neck. After preparation of the skin and drapes, and insertion of local anaesthetic, the bed is tilted to a 25° head down position. The needle is inserted just lateral to the posterior border of the clavicular head of the sternocleidomastoid muscle, about one finger breadth above the clavicle. It is then advanced parallel to the line of the clavicle and just behind the sternocleidomastoid muscle. The internal jugular vein, which lies superficially at this point, is cannulated close to its junction with the subclavian vein. As soon as the vein is entered the needle is angulated caudally to ease cannulation, the guidewire passing directly into the innominate vein. The risk of complications was lower with this technique than for any other method of central venous cannulation used in one large series (see Table 5). (b) The patient is positioned as described for the low lateral approach to the internal jugular vein. The needle is inserted in the centre of the triangle defined by the sternal and clavicular heads of the sternocleidomastoid muscle and the clavicle itself. It should be angulated caudally, at about 60° to the skin, and in a line pointing towards the ipsilateral anterior superior iliac spine. See Table 5 for details of complications using this approach.

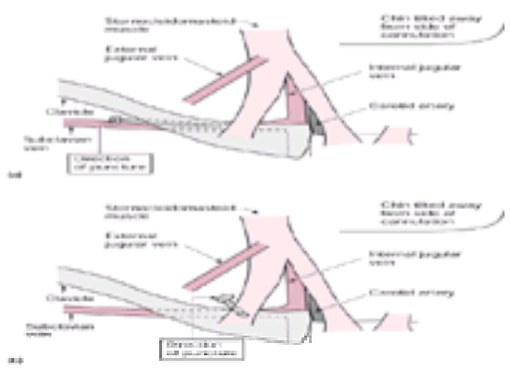

Fig. 2 The infraclavicular (a) and supraclavicular (b) approaches to the subclavian vein. (a) The patient is positioned as described for the low lateral approach to the internal jugular vein (Fig. 1(a)), excepting that instead of a towel being placed under both shoulders it should be positioned under the spine, allowing the shoulders to retract to reduce the risk of pneumothorax. The needle enters the skin below the mid-point of the lower border of the clavicle and is advanced under the clavicle towards the upper edge of the junction of the clavicle with the manubrium. See Table 5 for details of complications using this approach. (b) The patient is positioned as described for the infraclavicular approach to the subclavian vein. The needle is inserted into the angle between the superior border of the clavicle and the posterior border of the clavicular head of the sternocleidomastoid and advanced caudally, medially, and ventrally. See Table 5 for details of complications using this approach.

What fluid should you give, and how much?

A great deal of heat, but little light, has been generated in the literature on the subject of when to replace fluid, what to give, and how much. For patients with penetrating trauma to the torso it has been shown that delayed resuscitation, where venous access is established but fluid is not given until the patient is in the operating theatre, is preferable to immediate resuscitation, but there is no obvious analogy between this situation and that of the vast majority of patients with circulatory collapse and the rule, in general, remains that resuscitation should start as soon as possible.

It is logical that the fluid given to the patient whose intravascular volume is depleted should be one that remains substantially within the intravascular compartment. Solutions based on dextrose (with zero or low concentration of sodium) are most certainly not appropriate, since they partition throughout the body water and relatively little remains in the intravascular compartment, but beyond this it is not possible to make any firm recommendation. If the patient has lost blood, then it would seem sensible to give blood, and most physicians would recommend this. Whether isotonic crystalloid (usually 0.9 per cent saline), hypertonic crystalloid (usually saline), or various types of colloid are best in other situations is not clear. Cochrane reviews have failed to find significant differences between the use in critically ill patients of crystalloid or colloid, between isotonic or hypertonic fluids, and between different types of colloid solution, although there is evidence that albumin infusion

may increase the risk of death in this situation.

At the outset it is not possible to judge precisely how much fluid will be needed to resuscitate a patient. The only way to determine this is by frequent clinical examination as fluid is given. In the patient who is very unwell and clearly volume depleted, standard practice is to give 500 ml of blood or plasma expander (as appropriate and as available) as fast as the giving set and venous cannula will allow (applying pressure to the bag by manual or mechanical compression if the patient is *in extremis*). A second 500 ml infusion is commenced whilst checking peripheral perfusion, pulse rate, blood pressure, and the jugular venous pressure. Rapid infusion is continued until there is clear evidence that the situation is beginning to improve, as manifest by warming of the peripheries, slowing of the pulse rate, and rise in blood pressure. Interpretation of change in the height of the jugular venous pressure requires some care. It rises as fluid starts to be given, but may then fall for two reasons: first, if there is further fluid loss, most typically haemorrhage; and second, as venoconstrictor tone diminishes in the patient who is 'warming up' with adequate resuscitation. Their different effects on peripheral perfusion, pulse rate, and blood pressure easily distinguish these two eventualities.

As soon as it is clear that the patient's circulation is beginning to improve, the rate of fluid infusion should be slowed so as not to risk precipitating pulmonary oedema by forcing very high hydrostatic pressures in a circulation that is still 'tight' due to the effect of endogenous vasoconstrictors. Hence the patient who has lost, say, 2 litres of blood, may be optimally treated by receiving the first litre as quickly as possible, followed by the second litre over the next 2 h or so as the circulation 'relaxes'.

When resuscitation is complete—meaning that peripheral perfusion, pulse rate, blood pressure, and jugular venous pressure have all returned to acceptable levels—fluid input should then be given with regard to fluid output. At this stage it is good practice to insert a urinary catheter into any patient who has presented with severe cardiorespiratory disturbance. Urinary flow rate reflects renal perfusion, which is a marker of the overall state of the circulation, and accurate measurement of urinary output is essential to judge continuing fluid requirement. If patients have developed acute tubular necrosis with oliguria as a result of hypotension, they will not be well served by a 'standard' prescription of 3 litres of fluid per day to follow on from that given to resuscitate. This will inevitably lead to pulmonary oedema if continued in the face of diminished urinary output. A daily input equal to the last 24-hours' output plus 500 to 1000 ml for insensible losses is appropriate in these circumstances.

The patient who is volume verloaded

Acute volume overload manifests as pulmonary oedema, which can be a most terrifying condition. When severe, patients cannot get their breath and, with good reason, think they are going to die. As for patients in extreme respiratory difficulty, they sit up and use their accessory muscles. They are sweaty, cool peripherally, tachycardic, hypertensive (usually), centrally cyanosed, the jugular venous pressure is raised, and there is a gallop rhythm, although this may be hard to appreciate amidst the widespread crackles and wheezes in the chest. In extreme forms, frothy pink oedema fluid may come from the mouth. For further information see Chapter 15.15.2.2. The main features of treatment are shown in Table 4.

The underlying condition

The initial management of patients who are desperately ill does not depend on making a precise diagnosis of the cause of their predicament. However, as soon as resuscitation is underway, attention must turn towards making a diagnosis. Although the naive might think that the more severe the illness, the more obvious the cause should be, the opposite is often the case. When dead, all patients look identical, and the same is true just before they die. Patients who are *in extremis*, whether due to profound hypoxia or with next to no blood pressure, are not lucid historians, and it may be that the only question that they can usefully answer is: 'Do you have any pain?' If they indicate their chest or their abdomen, this might be a helpful clue.

The pragmatic approach to making a diagnosis in the patient with cardiorespiratory collapse is to use a 'surgical sieve' technique, looking systematically for features on examination and investigation to nail the diagnosis of conditions that can kill (Table 6). Details of the management of the many specific disorders listed in Table 6 can be found in the relevant sections of this book, but there is one general point: if initial investigations do not give any clear diagnostic lead, then do not be afraid to start treatment 'on suspicion', especially for those disorders that cannot reliably be diagnosed or excluded by clinical examination or with those tests that are rapidly available. In particular, consider pulmonary embolism and sepsis.

If the clinical context makes pulmonary embolism likely, for instance the patient has collapsed after an operation a week or so ago, then—in the absence of other explanation for the problem—it would not be unreasonable to start anticoagulation with intravenous heparin (which can be reversed if necessary) pending definitive imaging, but it would be unwise to give thrombolytic agents until the diagnosis was established. I began this chapter with a paragraph describing a failing of another doctor, so it is fair and reasonable that I also include details of one of my own. When I was a medical registrar 'on call' 15 years ago, I was asked to see a woman in her 60s from a long-stay psychiatric ward because the nurse looking after her thought that 'she wasn't her usual self'. She would not speak to me, but this was not abnormal since she had not spoken to anyone for many years. I examined her with some difficulty, because she did not co-operate, and could find nothing wrong, except that her systolic blood pressure was 60 mmHg. ECG and chest radiography were unremarkable, blood tests were taken, and I arranged to 'review with the results'. She died 4 h later. The next day the microbiologist phoned to say that streptococci had been grown from all blood cultures. Even awkward patients should have a blood pressure, and serious disease is likely if they don't. Give broad-spectrum parenteral antibiotics as soon as blood cultures have been taken if the cause of cardiorespiratory collapse is not apparent. And with regard to sepsis, ask if patients have travelled recently: if they have, the diagnosis may be malaria.

Communication

Once resuscitation and diagnostic endeavours are underway, and certainly when you have a clear idea what the diagnosis is, do not forget to speak to the patient's relatives and record in the medical notes what you have told them.

Further reading

Alderson P *et al*. (2000). Colloids versus crystalloids for fluid resuscitation in critically ill patients. *Cochrane Database System Review* CD000567.

Bickell WH *et al*. (1994). Immediate versus delayed fluid resuscitation for hypotensive patients with penetrating torso injuries. *New England Journal of Medicine* **331**, 1105–9.

Bristow PJ *et al*. (2000). Rates of in-hospital arrests, deaths and intensive care admissions: the effect of a medical emergency team. *Medical Journal of Australia* **173**, 236–40.

Bunn F *et al*. (2000). Colloid solutions for fluid resuscitation. *Cochrane Database System Review* CDOO1319.

Bunn F *et al*. (2000). Human albumin solution for resuscitation and volume expansion in critically ill patients. The Albumin Reviewers. *Cochrane Database System Review* CDOO1208.

Bunn F *et al*. (2000). Hypertonic versus isotonic crystalloid for fluid resuscitation in critically ill patients (Cochrane Review). *Cochrane Database System Review* **4**, CD002045.

Pittiruti M *et al*. (2000). Which is the easiest and safest technique for central venous access? A retrospective survey of more than 5400 cases. *Journal of Vascular Access* **1**, 100

16.2 The circulation and circulatory support of the critically ill

David F. Treacher

Introduction

This section considers global and regional oxygen delivery, the concept of shock, assessment and monitoring of the circulation, and the use of vasoactive drugs and in the critically ill patient.

The aphorism 'prevention is better than cure' should guide the management of the circulation, since early identification of circulatory derangement and the prompt diagnosis and treatment of underlying pathology provides the best chance for the rapid and complete recovery of the patient. However, with the limited monitoring and supervision available outside critical care units, early detection may be difficult. Hypotension is widely considered to be the cardinal sign of circulatory dysfunction, but other global features such as poor peripheral perfusion, persistent tachycardia, restlessness, confusion, tachypnoea, respiratory distress, hypoxaemia, and progressive metabolic acidaemia frequently occur earlier. These reflect the activation of powerful homeostatic mechanisms that maintain pressure at the expense of flow.

Features of regional circulatory failure include:

- oliguria with deteriorating renal function;
- confusion and impaired conscious level;
- chest pain, dysrhythmias, and ischaemic changes on electrocardiography (**ECG**);
- nausea, vomiting, and impaired gastrointestinal function: these symptoms of splanchnic ischaemia are often an early manifestation, but their significance is frequently overlooked.

Recent studies have demonstrated that evidence of an impending critical illness is often present—but not acted upon—for over 24 h in many of those patients on general wards who are subsequently admitted to intensive care units (**ICU**) following a cardiorespiratory arrest or with progressive organ failure. Such delay results in mortality rising almost exponentially as the number of organs that fail increases, leading to recommendations that 'rapid-response' or 'patient-at-risk' teams should be created with the aim of ensuring that experienced evaluation is rapidly available and timely admission to a critical care unit is arranged.

In high-risk surgical patients, preoperative 'optimization' of the circulation that ensures an oxygen delivery greater than 600 ml/min per m^2 through fluid resuscitation and, if necessary, inotrope/vasodilator therapy, has reduced mortality in controlled studies from over 20 per cent to less than 5 per cent, and the length of hospital stay by over 30 per cent.

Oxygen delivery and shock

Circulatory problems prompting a patient's admission to an ICU are cardiac arrest, pulmonary oedema, systemic hypotension, metabolic acidaemia, and organ failure, all situations that either threaten or have resulted from a failure of global or regional oxygen delivery. These patients are frequently described as suffering from 'shock'. The term has the advantage of brevity but little else, since it is imprecise and is applied to circulatory states that differ profoundly when analysed pathophysiologically. It can therefore guide neither management nor prognosis, but is irreversibly part of critical care terminology. Shock may be defined as inadequate cellular perfusion and oxygen uptake with consequent tissue hypoxia and organ failure. The terms 'early' and 'late' shock reflect the association between the duration and severity of the circulatory derangement and the prospect for recovery.

The major categories of shock, with examples, are:

- *Hypovolaemic*: haemorrhage, burns, gastrointestinal fluid loss;
- *Cardiogenic*: myocardial infarction, myocarditis, valve dysfunction;
- *Obstructive*: pulmonary embolus, cardiac tamponade, tension pneumothorax;
- *Anaphylactic*: drugs, blood transfusion, insect sting;
- *Septic*: bacterial infection, non-infective inflammatory conditions, e.g. pancreatitis, burns, trauma.

The primary problem in hypovolaemic, cardiogenic, and obstructive shock is a progressive decline in cardiac output and global oxygen delivery (DO_2), which—if not corrected—leads to secondary failure of the peripheral circulation and progressive organ dysfunction. By contrast, in sepsis and anaphylaxis, the primary problem is a failure of control of the peripheral circulation with disruption of the regional distribution of cardiac output, hence the alternative description of peripheral or distributive shock.

The initial physiological goal in the circulatory resuscitation of the critically ill patient is to restore global oxygen delivery while appropriate specific treatment is instituted.

Oxygen delivery

The delivery of oxygen from the external environment via the lungs to the mitochondria within individual cells is summarized in Fig. 1, with typical values quoted for normal activity in a 70-kg individual. Global oxygen delivery (DO_2) is calculated from the cardiac output, the oxygen saturation, and the haemoglobin concentration of arterial blood (Table 1).

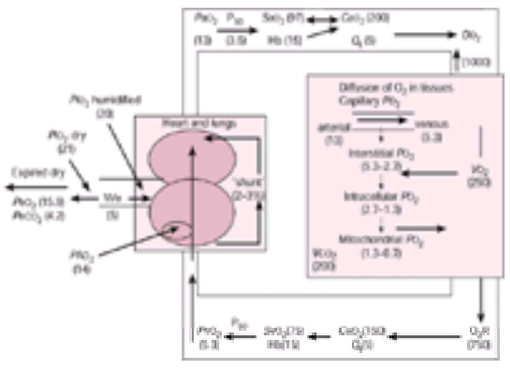

Fig. 1 Oxygen transport from the atmosphere to the mitochondria. Values in parentheses are for a normal 70-kg individual (body surface area (BSA), 1.67 m^2) breathing air (fractional inspiratory oxygen concentration (FiO_2), 0.21) at standard atmospheric pressure (PB), 101 kPa). Partial pressures of O_2, CO_2 in kPa; saturation in per cent; contents (CaO_2,CvO_2) in ml/litre; Hb in g/100 ml; blood/gas flows (Qt, Vi/e) in litres/min; oxygen transport (DO_2, O_2R), VO_2, and VCO_2 in ml/min. Abbreviations: SO_2, oxygen saturation (%); PO_2, partial pressure of oxygen (kPa); PiO_2, partial pressure of inspired O_2; PeO_2, partial pressure of mixed expired O_2; $PeCO_2$, partial pressure of mixed expired CO_2; PAO_2, partial pressure of alveolar O_2; PaO_2, partial pressure of arterial O_2; SaO_2, arterial SO_2; SvO_2, mixed venous SO_2; Qt, cardiac output; Hb, haemoglobin; CaO_2, arterial O_2 content; CvO_2, mixed venous O_2 content; VO_2, oxygen consumption; VCO_2, CO_2 production; O_2R, oxygen return; DO_2, oxygen delivery; Vi/e, minute volume-inspiration/expiration.

Apart from the adequacy of overall DO_2, the regional distribution both between and within organs is important in maintaining normal organ function. If the myocutaneous bed receives a disproportionately high blood flow, but the flow to the splanchnic bed is poor, then the gastrointestinal tract and liver will become ischaemic despite a high DO_2.

The final parts of the oxygen cascade depend on diffusion, determined by the PO_2 gradient and the distance from capillary to cell, and also upon the integrity of cellular metabolic function. Increased levels of DO_2 cannot compensate for either the disruption of regional distribution or impaired diffusion between capillary and cell, or for primary metabolic failure within the cell as occurs in cyanide poisoning and sepsis. This explains why strategies that improve global oxygen delivery may even adversely affect cellular oxygen status—for example: (1) vasoactive agents altering the distribution of DO_2 between and within organs; (2) excessive volume loading producing tissue oedema and impairment of diffusion; (3) inotropes that increase cellular oxygen requirements (for instance, epinephrine (adrenaline) and dobutamine) may increase cellular oxygen debt.

Although achieving and maintaining an adequate global DO_2 is undoubtedly important, particularly in early resuscitation, the ability to measure and control regional distribution is now the challenge in the circulatory management of the critically ill patient.

Relationship between oxygen delivery and consumption

The oxygen extraction ratio (**OER**) is the percentage of the oxygen delivered that is extracted by the tissues, which is normally approximately 25 per cent. As metabolic demand (VO_2) increases or supply diminishes, the OER increases to maintain aerobic metabolism. However, as demonstrated by point B in Fig. 2, there is a maximum extraction ratio (slope AB) that can be achieved. This is around 60 to 70 per cent for most tissues, beyond which a further increase in VO_2 or decline in DO_2 must lead to an inadequate availability of oxygen, at least in certain tissue beds, thus leading to anaerobic metabolism and increased lactic acid production. The dotted line, DEF, represents the altered relationship that may exist during critical illness. The slope of maximum OER falls, reflecting the reduced tissue extraction of oxygen, but the relationship does not plateau as in the normal relationship, that is to say oxygen consumption continues to rise with increasing delivery.

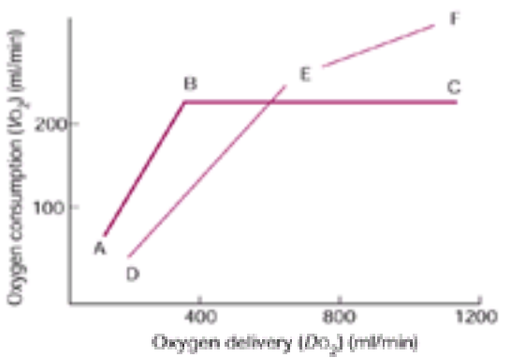

Fig. 2 Relationship between oxygen delivery and consumption. The effects of changing oxygen delivery on consumption. The solid line, ABC, represents the normal relationship and the fine line, DEF, the relationship believed to exist in critically ill patients.

An enthusiasm for achieving supranormal levels of DO_2, so called 'goal-directed' therapy, arose from the concept of shock as a failure of adequate global oxygen delivery and the belief that oxygen consumption had become 'supply-dependent'. Observations of postoperative patients in the early 1980s led to a decade of ICU practice that focused on such goal-directed therapy. Based on studies relevant to postoperative patients, many critically ill patients with established 'shock' in ICU were managed with the aim of achieving oxygen deliveries *above* 600 ml/min per m^2 using vigorous volume loading and inotropic support, frequently with dobutamine, in the belief that this would increase VO_2, relieve tissue hypoxia, prevent multiorgan failure, and improve prognosis. Several major studies have now failed to demonstrate any benefit from this approach in the treatment of established shock, where such aggressive volume loading can be detrimental. In these studies, the patients who failed to achieve the desired DO_2 levels despite volume loading and dobutamine had a poorer outcome than those with a higher DO_2, whether achieved spontaneously or following intervention, as did patients with low levels of DO_2 in the control group. This 'physiological reserve' or ability to increase DO_2 in response to the stress of critical illness is an important prognostic marker, which progressively diminishes with the length of time that the state of shock persists and as the circulatory state changes from early to late or 'irreversible/unresponsive' shock.

Although the relationship DEF of Fig. 2 may be unproven for established or late shock, the concept that a point B exists beyond which a further reduction in DO_2 causes progressive tissue hypoxaemia is valid early in the evolution of shock. This is particularly so for the low cardiac output/high vascular resistance causes (hypovolaemia, cardiogenic, obstructive) and also in early septic shock with intravascular depletion and low flows. In these situations the appropriate treatment is to increase DO_2 promptly by restoring the intravascular volume, relieving obstruction, or augmenting cardiac function. However, in the later stages of established shock from any cause the problem becomes peripheral at the microcirculatory level, with failure of autoregulation, abnormal regional distribution of flow within the organs, and direct cellular toxicity preventing oxygen uptake and utilization despite high levels of global DO_2.

Aside from variations in DO_2, tissue oxygenation may also be improved and aerobic metabolism sustained by reducing oxygen demand—achieved by controlling those factors that increase metabolic rate, such as: sympathetic activation from pain, agitation, shivering, and various interventions (nursing procedures, physiotherapy, visitors), drugs, and pyrexia. For each degree Celsius rise in temperature, oxygen consumption increases by 10 to 15 per cent.

Assessment of global circulatory performance

Cardiac output (Qt) depends upon:

1. ventricular 'preload', i.e. atrial filling pressures (right atrial pressure (**RAP**), left atrial pressure (**LAP**));
2. ventricular 'afterload', i.e. mean pulmonary and systemic arterial pressures (**PAP**, **SAP**);
3. ventricular contractility;
4. heart rate and rhythm.

Preload

Ventricular preload, traditionally assessed from the atrial filling pressures (RAP, LAP), determines the end-diastolic ventricular volume. This, according to Starling's law and depending on ventricular contractility, will in turn determine the work generated by the next cardiac contraction, the resulting stroke volume depending on the afterload.

Both ventricular contractility and afterload will affect the atrial filling pressures, but the predominant factor influencing preload is venous return, which is determined by the intravascular volume and the venous 'tone'. The systemic venous 'tone' or compliance is controlled by the autonomic nervous system and circulating catecholamines and can vary from 30 to over 300 ml/mmHg, such that the normal 70-kg person can compensate through venoconstriction for an intravascular volume loss of up to 1.5 litres without developing overt circulatory disturbance.

If the preload is low and either blood pressure or cardiac output is inadequate, the treatment priority is volume loading to restore the intravascular volume and venous return.

A high preload reflects one or more of the following: (1) high intravascular volume; (2) impaired myocardial contractility; or (3) increased 'afterload'. If treatment to reduce the preload is warranted, the options are therefore: to remove volume from the circulation (diuretics, venesection, haemofiltration) or to increase the capacity of the vascular bed with venodilator therapy (e.g. glyceryl trinitrate, morphine); to improve contractility; or to reduce the resistance of the relevant arterial bed.

In interpreting atrial pressures as measures of preload, two points must be considered. First, if the intrathoracic pressure (**Pt**) is raised, the intravascular pressure (**Pv**) may be misleading as a measure of preload since the true ventricular distending pressure is the transmural pressure (Pv – Pt). This is particularly relevant in situations where there is significant alveolar gas trapping, as in asthma or ventilation with high end-expiratory pressure and an increased ratio of inspiratory to expiratory time. Second, when the ventricle is dilated and poorly compliant the end-diastolic pressure-volume relationship is not necessarily linear, and ideally volume rather than pressure preload should be measured.

Afterload

Afterload influences the tension developed in the ventricular wall during systole, and is determined by the resistance to ventricular outflow from valvular abnormalities and the peripheral vascular resistance. The systemic and pulmonary vascular resistances (**SVR**, **PVR**) are calculated by analogy with Ohm's law as the pressure drop across the resistance bed divided by the flow, making the considerable assumption that flow in the circulation is linear and non-pulsatile.

Appropriate circulatory management requires an appreciation of the relationship between pressure, resistance, and flow: for a constant ventricular stroke work, the cardiac output will be inversely related to arterial pressure and resistance. The drugs used to manipulate pulmonary and systemic resistances are listed in Table 2.

Myocardial contractility

The ventricular stroke work is the external work performed by the ventricle with each beat and is calculated from the stroke volume and the pre- and afterload pressures (Table 1). The relationship between filling pressure, stroke work, and the resulting stroke volume for a constant afterload defines myocardial contractility. Consideration of ventricular work is important since circulatory management involves achieving the necessary pressures and flows to maintain satisfactory perfusion and oxygen delivery to all organs at maximum cardiac efficiency, that is for the least ventricular work, so that myocardial ischaemia is avoided.

Myocardial contractility is frequently reduced in critically ill patients due either to pre-existing cardiac disease, most often ischaemic heart disease, or as a consequence of the disease process, particularly sepsis. If the cardiac output is inadequate and myocardial contractility is poor, as defined by a 'flattened' stroke work/filling pressure equation, the available options are to:

- *reduce afterload* using an arteriolar dilator (nitrates, a-receptor blocker, angiotensin-converting enzyme (**ACE**) inhibitor). However, this may be limited by the consequent fall in systemic pressure.
- *increase preload*. Although appropriate if the preload is low or the intrathoracic pressure is high, the filling pressure will often already be raised and any further increase may not only fail to augment stroke volume but could increase ventricular wall tension, compromise ventricular blood supply, particularly to the endocardium, and potentially cause further impairment of contractility as well as pulmonary oedema.
- *increase myocardial contractility*, either by removing negatively inotropic influences (acidaemia, hyperkalaemia, drugs (for example, b-receptor blockers), or by using an inotrope (Table 2). The possible adverse effects of vasoactive agents on regional distribution of flow must also be considered.

Heart rate and rhythm

If the heart rate is low (<65 beats/min) with a low cardiac output, then increasing the rate either with a b_1-receptor agonist (for example, isoprenaline) or by pacing should be considered. Atrial pacing, or atrioventricular sequential pacing, has the advantage of maintaining co-ordinated atrial contraction that increases ventricular end-diastolic volume and hence stroke volume without increasing myocardial irritability.

Heart rates above 120 beats/min, other than sinus tachycardia, should be controlled either by drugs and/or DC cardioversion after ensuring that low plasma potassium and magnesium levels have been corrected.

Monitoring

The purpose of monitoring the critically ill patient is both to alert staff if a physiological variable is not maintained within preset limits and to guide therapy. Critically ill patients require the following investigations: continuous monitoring of heart rate and rhythm; intravascularly measured blood pressure and RAP; arterial oxygen saturation by oximetry; intermittent blood gas analysis, which usually also provides electrolyte, glucose, and lactate concentrations; and hourly fluid balance measurements.

Pulmonary artery catheterization

Of the six primary circulatory variables measured in an intensive care unit, three (RAP, mean arterial pressure (**MAP**), and heart rate) are routinely available even outside an ICU. Conventionally, the other three (PAP, LAP, and Qt) are measured by the insertion of a pulmonary artery catheter. Apart from the derivation of vascular resistances, ventricular work, and oxygen delivery, the mixed venous oxygen saturation can be measured and used to calculate oxygen consumption (Table 1).

Sampling from the pulmonary artery is necessary to ensure good mixing of the venous blood, since the saturation of venous blood from different organs does vary considerably. Hepatic venous saturation may only be 40 per cent, whereas renal venous saturation may exceed 80 per cent, reflecting the considerable difference between these organs in oxygen delivery compared to their metabolic requirements. Provided the peripheral distribution of DO_2 and cellular metabolic function are normal, a value of SvO_2 above 65 per cent indicates that oxygen delivery is satisfying tissue oxygen requirements.

The focus on 'goal-directed therapy' led to the widespread use of pulmonary artery (**PA**) catheters, with an annual consumption of over 2 million in the United States.

However, their indiscriminate use has been challenged by:

1. evidence that goal-directed therapy, although probably appropriate for under-resuscitated perioperative patients, may be harmful for other critically ill patients;
2. a multicentre case-controlled study of the use of the PA catheter, which suggested that those patients managed with a PA catheter had a poorer prognosis than those managed without such an intervention;
3. doubts about the appropriateness of using pressure preload as reflected by the pulmonary artery occlusion pressure as a surrogate for the volume preload of the left ventricle.

Although the PA catheter is undoubtedly associated with complications, particularly infection (see Table 3), and the measurements subject to significant error, the results of this study probably reflected poor training in the use of the catheter and a failure to respond appropriately to the data obtained. Insertion of a PA catheter should still be considered:

- if the RAP is raised and the relationship with the LAP is uncertain due to a recent myocardial infarction, valvular abnormalities, or high pulmonary vascular resistance—the most useful finding is a low 'wedge' pressure, demonstrating that further volume is indicated, but a high value does not necessarily exclude the need for further volume;
- if the patient's condition fails to improve with initial management;
- to measure cardiac output to guide the appropriate choice of vasoactive drug, particularly when high doses are being used;
- if continuous monitoring of PA pressures and evaluation of right ventricular (**RV**) function is indicated

The PA catheter debate has, however, led to considerable interest in alternative, less invasive methods of assessing cardiac output and the adequacy of left ventricular preload.

Measurement of global cardiac output

Table 4 summarizes the features of the alternative methods available for measuring cardiac output, with an assessment of those aspects other than accuracy that should be considered. Any additional information that is provided, such as continuous mixed venous oxygen saturation with a PA catheter, should be considered when selecting the most appropriate monitor.

The most widely used method for measuring cardiac output remains the thermodilution technique using a PA catheter. Although generally viewed as the 'gold standard', the error is at least 10 per cent and there are potentially serious complications, particularly infection.

Clinical assessment of systemic vascular resistance and pulse volume allows cardiac output to be estimated with sufficient accuracy to direct initial management. This approach is rapid, requires no invasive monitoring, may be performed outside the intensive care unit, and is relevant for use in less-well resourced countries. If the patient does not improve with management based on this initial assessment, more invasive monitoring is indicated.

With the oesophageal Doppler technique an ultrasound transducer probe is positioned in front of the descending aorta and blood flow velocity measured, from which cardiac output is derived using an estimate of the aortic cross-sectional area calculated from the patient's height, weight, and age. Intravascular volume status can also be assessed from changes in the size and shape of the aortic velocity waveform. It can be inserted rapidly and left *in situ* for up to a week without serious complication and is notionally non-invasive. However it is expensive, operator-dependent, and cannot easily be used in patients who are not intubated.

The recently reported technique of lithium dilution cardiac output can be performed with a peripheral venous and arterial line and compares favourably with the thermodilution method.

Pulse contour analysis relies on analysis of the arterial pressure waveform to provide a beat-by-beat estimation of the stroke volume. Regular calibration by thermal or lithium dilution is necessary but only requires a central venous injection of the indicator, and from this both intrathoracic blood volume and lung water can also be derived to provide an assessment of the adequacy of intravascular fluid resuscitation.

Measurement of regional blood flow

Appropriate management of the circulation requires information about regional as well as global flow. Urine output is sensitive to changes in renal perfusion, provided the kidneys have not developed acute tubular necrosis (**ATN**) or been poisoned with drugs, particularly diuretics. The lower limit is 0.5 ml/kg per min, but twice this rate is appropriate in the catabolic patient.

Peripheral perfusion and changes in cardiac output and its distribution can be assessed from the gradient between the peripheral temperature, usually measured over the dorsum of the foot, and the central or core temperature (rectal, oesophageal, or possibly tympanic).

Splanchnic blood flow, which is particularly sensitive to hypovolaemia and vasoconstricting inotropes and which is important in the aetiology of multiple organ failure, can be assessed by tonometry. The gastric tonometer is a nasogastric tube with a second channel connected to a balloon that can be inflated with saline or air. When the PCO_2 within the gastric mucosal cells (PCO_{2i}), stomach lumen, and the balloon have equilibrated, a sample is withdrawn from the balloon to measure PCO_2. From this measurement and the arterial bicarbonate concentration, the intracellular pH (pHi) can be calculated. Some perioperative studies in patients have shown that pHi is valuable both prognostically and in guiding treatment. Its role in the established critically ill patient is less clear. The gastric mucosal and alveolar PCO_2 gradient is probably a more appropriate measure. However, problems remain with obtaining useful data when the patient is being enterally fed, and there is also uncertainty whether gastric PCO_{2i} reflects the energy and oxygen status of other parts of the bowel. Although tonometers can be placed in the small bowel and sigmoid colon, the technique remains predominantly a research tool.

The new technologies of magnetic resonance spectroscopy and positron emission tomography allow more detailed study of the regional circulation, tissue oxygenation, and cellular bioenergetics and will soon become available for clinically based studies in the critically ill patient.

Echocardiography is useful in assessing ventricular volume preload and will also provide information about cardiac function and structure. It is particularly valuable in the diagnosis of obstructive shock and aortic dissection. The transthoracic approach may be difficult in ventilated patients due to a poor 'echo window', but the transoesophageal approach has been a major advance.

Acid–base balance and serum lactate are useful metabolic indices of the circulation and both are sensitive to intravascular volume depletion, particularly in the presence of inotropic support. Progressive metabolic acidaemia requires diagnosis and effective treatment. The serum lactate level reflects the balance between increased production in hypoxic tissues and metabolic clearance, which is predominantly hepatic. High lactate levels often indicate splanchnic ischaemia when there is both increased production and reduced clearance due to hepatic ischaemia. A single measurement is difficult to interpret, but the trend is valuable in assessing the response to treatment.

Key points in monitoring the circulation

- Regular clinical examination should never be forgotten: the symptoms and signs already discussed are as important as impressively displayed numbers on expensive monitors.
- Trends are more useful than a single observation.
- Pressure is no guarantee of flow.
- Although the jugular venous pressure (**JVP**) is traditionally measured from the sternal angle, in ICU pressures are measured from the mid-axillary line in the 5th intercostal space. From this reference point, in the supine position, the normal RAP is between +4 and +8 mmHg and the LAP or 'wedge' between +10 and +14 mmHg.
- If monitor data and clinical judgement conflict, check for common sources of error in the monitoring such as catheter blockage or malposition, failure to re-zero after postural change.
- Critical care monitoring is invasive, potentially hazardous (Table 3), expensive, and mostly of no proven benefit. The device should be used while *in situ*, but it

should be removed as soon as the information obtained is no longer required.
- Outcome benefit from the use of any monitor depends on the data being displayed accurately (zero and calibration), observed promptly, and interpreted appropriately and on an effective intervention being available and rapidly instituted.

Management of the circulation in the critically ill

General principles

The circulation should be assessed as shown in Table 5, which provides examples of the circulatory abnormalities in various conditions. The severity of the condition and pre-existing cardiorespiratory disease will affect the precise figures obtained in individual cases. It is not always necessary to use invasive monitoring to obtain precise measurements, but the discipline of estimating the key variables commits the clinician to a logical analysis of the problem and an awareness of which measurements need to be confirmed, invasively if necessary, if the patient's condition fails to improve.

The six key questions in managing the circulation are:

1. *What are the appropriate targets for blood pressure and cardiac output?* —Despite the emphasis placed on blood flow to the tissues, adequate perfusion pressure is also necessary to achieve the appropriate distribution of cardiac output and oxygen supply. The target for mean arterial pressure should be 65 mmHg but adequate splanchnic and renal perfusion may require higher pressures, particularly in the elderly patient with pre-existing hypertension or widespread atheroma. In treating a patient with cerebral oedema the target pressure should be 65 mmHg *above* the intracranial pressure.
2. A minimum cardiac output of 2.8 litres/min per m^2 should ensure that tissue hypoxia is not due to inadequate global oxygen supply. Thereafter, management of organ dysfunction should concentrate on the regional distribution.
3. *Has sufficient fluid been given?*—Conventionally this is based on measurement of the atrial filling pressures (RAP and LAP, or pulmonary artery occlusion/'wedge' pressure). Before starting inotropic support, fluid should be given to achieve an RAP up to 12 mmHg or a 'wedge' pressure of 18 mmHg. This assumes that the relationship between the atrial filling pressures, the permeability of the pulmonary capillary membranes, and the intrathoracic pressure are all normal, but none of these assumptions is necessarily valid in the critically ill patient. A low value (RAP <10 mmHg, LAP <14 mmHg) is helpful since further volume will improve cardiac work. Higher levels are more difficult to interpret, particularly in the ventilated patient, since, although an inadequate volume preload is not necessarily excluded, further fluid may result in pulmonary oedema.
4. Intravascular volume depletion is suggested by hypotension precipitated by sedation, analgesia, postural change, or during the inspiratory phase of positive-pressure ventilation. Brief disconnection from the ventilator causes the blood pressure to rise and venous pressure to fall: the measurement 'off' the ventilator more accurately reflects the ventricular end-diastolic transmural pressure. This manoeuvre is relatively contraindicated in patients with acute respiratory distress syndrome (**ARDS**), since loss of positive end-expiratory pressure may cause widespread alveolar collapse.
5. The difficulty in interpreting the absolute levels of RAP/LAP can be resolved by a fluid challenge: 200 ml of colloid is administered and the impact on blood pressure, flow, and preload observed. In the volume-depleted patient, blood pressure and flow will increase with only a small, transient increase in filling pressures. While pulmonary gas exchange remains satisfactory there is less anxiety about giving further colloid. Sufficient volume will have been given when either the target pressures are achieved and the evidence of poor peripheral perfusion and organ dysfunction has resolved, or when there is a sustained rise in filling pressures to a level above which there is a risk of pulmonary oedema developing.
6. *Which fluids should be used?*—The previous day's crystalloid and colloid balance should be reviewed, both intravascular and extravascular compartments should be assessed, and the volume of fluid to be given over the following 24 h should be decided. The crystalloid balance should include the planned enteral intake, fluid for central lines and drug infusions, urine output, and correction for both 'insensible' losses (sweat, diarrhoea) and the state of hydration of the extravascular tissue space. A daily target balance ranging from –1.5 to >+3 litres may be appropriate, but typically it will be between –0.5 and +1.5 litres.
7. If the intravascular space is underfilled, the rate of infusion of normal saline should be increased by up to 100 ml/h, but, acutely, the extra volume required to reach the preload target should be given as colloid. With an active haemorrhage or if the haemoglobin concentration is less than 8 g/dl, blood should be used. Traditionally, the target haemoglobin has been 10 g/dl, since this was believed to represent the balance between oxygen content and viscosity that achieved optimum tissue oxygen delivery. However, a prospective randomized study to assess the impact of the haemoglobin level on outcome demonstrated that, for patients without significant coronary artery disease, survival was improved if the haemoglobin was maintained between 7 and 9 g/dl rather than between 10 and 12 g/dl.
8. After appropriate blood transfusion, synthetic colloid rather than albumin should be used. A much-debated meta-analysis comparing the use of albumin with crystalloid or synthetic colloid concluded that, in the critically ill, albumin was associated with an increased mortality. Certainly, attempting to correct a low serum albumin level in such patients exhibiting a significant inflammatory response is futile, since their vascular endothelium will be freely permeable to albumin. There is relatively little evidence on which to base the choice of synthetic colloid (starch or gelatin) but the increase in intravascular volume is sustained for longer with the starch solutions, which also provide a wider range of molecular weight products and a sodium-free option.
9. *Are there metabolic factors that require correction?*—Metabolic acidaemia with a pH below 7.20 or a base deficit above 10 mmol/l should be corrected, as myocardial contractility increases linearly with rising pH to values above 7.40. The suggestion that sodium bicarbonate will produce a damaging paradoxical intracellular acidosis is misleading, since the experiments demonstrating this effect were performed *in vitro* with unphysiological solutions, within a 'closed system' that allowed no correction for any rise in carbon dioxide concentration, and the sodium bicarbonate was given by bolus injection rather than infusion. Sodium bicarbonate, given as a physiological infusion at between 1 and 2 mmol/min, improves myocardial contractility and cardiac output, as demonstrated by several clinical studies and shown by the benefits of bicarbonate haemofiltration in lactate-intolerant, critically ill patients.
10. Hyperkalaemia, hypocalcaemia, and hypophosphataemia also impair myocardial contractility and should be corrected.
11. *Is ventilatory support indicated?*—Hypoxaemia should be corrected promptly to ensure that oxygen saturation is at least 92 per cent, both to prevent myocardial ischaemia and to ensure maximum oxygen delivery. The work of breathing may account for up to one-third of oxygen consumption and, if provided mechanically, circulatory demands will be reduced significantly. Non-invasive ventilatory support should first be considered: both continuous positive airway pressure (**CPAP**) and pressure support ventilation can be provided by a face or nasal mask. If formal mechanical ventilation with intubation becomes necessary, colloid volume should be given and a vasoconstricting inotrope infusion should be available to prevent the potentially catastrophic hypotension that may result from the use of sedative anaesthetic drugs in the volume-depleted patient with high endogenous catecholamine levels and raised SVR.
12. *Which vasoactive agents should be used?*—If the target systemic pressure and cardiac output/oxygen delivery are not achieved with appropriate intravascular filling, a suitable vasoactive agent or combination must be selected (Table 2). The impact of such treatment in individual patients will be influenced by their baseline circulation state (that is, if it is either intensely constricted or dilated the same drug will potentially produce different effects on pressure, flow, and its distribution). The initial choice of vasoactive agent will depend on the mean arterial pressure (**MAP**), cardiac output, and derived systemic vascular resistance (**SVR**). If, for example:
    - cardiac output and MAP are both low with a high SVR, an inotropic and dilating (inodilator) effect is required and epinephrine (adrenaline) with glyceryl trinitrate or dobutamine would be appropriate. If cardiac output rises but MAP falls, as may happen when dobutamine is given, a constricting agent such as norepinephrine (noradrenaline) is required.
    - MAP and SVR are low with a high cardiac output, as frequently occurs in sepsis after volume resuscitation, then arteriolar constriction with norepinephrine is needed.
    - MAP is at or above target but the cardiac output is low with a raised SVR, a dilating agent (glyceryl trinitrate) or an inodilator is appropriate treatment.
13. When the PVR and RAP are acutely raised, a pulmonary vasodilator to 'offload' the right ventricle and maintain cardiac output is required. A nitrate or b-receptor agonist, such as isoprenaline, would be appropriate, but hypotension may result from arteriolar dilatation and hypoxaemia due to increased ventilation–perfusion mismatch.
14. A vasoactive drug that could influence regional flow would be valuable. The belief that low-dose dopamine selectively improves renal blood flow has resulted in its widespread use. However, the evidence to support this belief is scanty, and a study of its use in perioperative patients suggested that it may even be harmful. Its undoubted effect in increasing urine output is probably due to an improvement in MAP and cardiac output together with a natriuretic effect, rather than to any specific effect on renal blood flow. When these effects are required, provided it is not used as a substitute for adequate volume replacement, it remains useful. Dopexamine, with both dopamine (DA1, DA2)- and b-receptor effects, is used to improve splanchnic blood flow. However, despite reported benefits when used with volume loading in perioperative patients, there is little evidence of outcome benefit in the treatment of established shock.
15. Phosphodiesterase inhibitors, such as milrinone, offer a theoretically attractive approach to improving myocardial contractility by increasing intracellular cAMP when there is reduced responsiveness to b-agonists. Useful increases in cardiac output can be achieved, but these agents are powerful vasodilators and hypotension may limit their use and results in the need for a norepinephrine infusion.

Management in specific conditions

Myocardial infarction/pulmonary oedema

The aim should be a dilated circulation to achieve the target cardiac output for the minimum cardiac work, but with a diastolic pressure above 50 mmHg to protect coronary artery perfusion. The preload should be reduced with nitrate therapy, both to reverse pulmonary oedema and to avoid ventricular overdistension and increased wall tension that reduces epicardial to endocardial flow. An ACE inhibitor should be started early, although hypotension and impairment of renal function may cause delay. The intra-aortic balloon pump has the advantage of augmenting cardiac work, improving coronary artery perfusion, and reducing the left ventricular afterload and hence the need for inotropic drugs.

Haemorrhage

During a major haemorrhage, sympathetic activation with intense venoconstriction prevents hypotension developing until 30 per cent of the circulating blood volume is lost, which explains the response to subsequent transfusion. As volume is lost, venous 'tone' increases to maintain venous return and only late is there a marked fall in venous pressure, cardiac output, and finally arterial pressure. If the rate of fluid replacement does not exceed the speed of resolution of the reflex increase in sympathetic tone, the atrial filling pressures will not rise excessively. However, rapid transfusion may produce very high atrial pressures and even pulmonary oedema, even though the volume replaced is less than the volume lost from the circulation.

Major pulmonary embolism

Oxygen delivery should be maintained by oxygen administration, volume expansion, and inotropic stimulation, while the obstruction is relieved by heparinization with thrombolysis or embolectomy.

The RAP should be raised to at least 15 mmHg using colloid to increase right ventricular work, and the resulting expansion of the pulmonary vascular bed may reduce the pulmonary vascular resistance by up to 50 per cent. Hydraulic pulmonary oedema does not occur since the atrial pressure relationship is reversed. An initial intravenous bolus of heparin (15 000 units) should be given to initiate anticoagulation and to reverse the reflex pulmonary vasoconstriction that occurs in the unobstructed pulmonary vascular bed. Digitalization improves right ventricular contractility and is sensible prophylaxis against supraventricular tachycardias.

Sedation, diuretics, haemorrhage, induction of anaesthesia, vena caval ligation, and the administration of contrast material during angiography may produce a disastrous fall in RAP and cardiac output. To prevent such changes, venous return must be maintained with volume expanders and by venoconstriction with a-agonists such as norepinephrine or phenylephrine.

Septic shock

In septic shock the circulatory changes are primarily peripheral, with cytokine release and activated white cells producing both direct cellular toxicity and microcirculatory chaos. As a result of the 'shunting' of blood through the tissues and the failure of cellular oxygen utilization, the tissues remain hypoxic despite high DO_2. This results in a low oxygen extraction ratio, a raised lactate level, and a reduced arteriovenous oxygen difference with a paradoxically high SvO_2. Secondary failure of the global circulation is caused by:

- volume depletion due to the increased leakiness of the vascular endothelium and sequestration secondary to venodilatation; and
- toxic effects on the myocardium impairing both systolic contraction and diastolic relaxation.

The typical circulatory profiles before and after volume resuscitation are shown in Table 5.

Despite a low SVR, the patient may feel cold peripherally and present as having unexplained hypotension with a speed of onset that suggests a major pulmonary embolus or myocardial infarction. However, palpation of a surprisingly large volume yet a rapid central pulse reveals that the cardiac output must be high, effectively ruling out these other causes of shock and indicating the true diagnosis.

Deciding on the 'appropriate' fluid volume to give in sepsis can be difficult. Frequently, the decision represents a balance between giving sufficient volume to prevent the use of excessive doses of constricting inotropes and giving excessive amounts with consequent tissue oedema and deterioration in pulmonary gas exchange.

Norepinephrine and epinephrine infusions will often be necessary, with doses adjusted as previously described to achieve pressure and flow targets. If high doses (>0.15 µg/kg per min) are required, splanchnic ischaemia is a major concern and should be suspected if the patient develops abdominal tenderness, large nasogastric aspirates, and a lactic acidosis. Patients with severe sepsis develop b-receptor desensitization with reduced intracellular cAMP levels, which makes the use of phosphodiesterase inhibitors such as milrinone logical. However, their use may exacerbate hypotension and result in the dose of vasoconstrictor drugs being increased.

Earlier studies suggested there was no role for steroids in the treatment of septic shock. However, there is now evidence that, at a lower dose (100 mg hydrocortisone, three times per day) than previously used, they can reduce the marked vasodilatation associated with a persistent and excessive inflammatory response, so that the doses of constricting inotropes can be reduced with potential benefit for regional perfusion.

Several large studies of anticytokine therapy and of antithrombin III supplementation in patients with severe sepsis have all failed to improve outcome. Similarly, a trial of *N*-monomethyl-L-arginine (**l-NMMA**)—an antagonist of the effects of inducible nitric oxide synthase, an enzyme responsible for the excess production of nitric oxide that is central to the severe peripheral vasodilatation associated with septic shock—had to be stopped prematurely because of an adverse outcome in the treatment group. However, a study in which activated protein C was given as an infusion over 4 days to patients with severe sepsis has now shown a significant reduction in mortality at 28 days.

Further reading

Bernard GR, *et al*. (2001). Recombinant human protein C. Worldwide Evaluation in Severe Sepsis (PROWESS) Study Group. *New England Journal of Medicine* **344**, 699–709.

Bradley RD (1977). *Studies in acute heart failure*. Edward Arnold, London.

Cochrane Injuries Group Albumin Reviewers (1998). Human albumin administration in critically ill patients: systematic review of randomised controlled trials. *British Medical Journal* **317**, 235–40.

Connors A, *et al*. (1996). The effectiveness of right heart catheterisation in the initial care of critically ill patients. *Journal of the American Medical Association* **276**, 889–97.

Consensus Conference (1996). Tissue hypoxia: how to detect, how to correct, how to prevent. *American Journal of Respiratory and Critical Care Medicine* **154**, 1573–8.

Gutierrez G, *et al*. (1992). Gastric intramucosal pH as a therapeutic index of tissue oxygenation in critically ill patients. *Lancet* **339**, 195–9.

Hayes MA, *et al*. (1994). Elevation of systemic oxygen delivery in the treatment of critically ill patients. *New England Journal of Medicine* **330**, 1717–22.

Hebert PC, *et al*. (1999). A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. *New England Journal of Medicine* **340**, 409–17.

Iberti TJ, *et al*. (1990). A multi-centre study of physicians' knowledge of the pulmonary artery catheter. *Journal of the American Medical Journal* **264**, 2928–32.

Leach RM, Treacher DF (1992). Oxygen transport: the relation between oxygen delivery and consumption. *Thorax* **47**, 971–8.

Schierhout G, Roberts I (1998). Fluid resuscitation with colloid or crystalloid solutions in critically ill patients: a systematic review of randomised trials. *British Medical Journal* **316**, 961–4.

Shippy CR, *et al*. (1984). Reliability of clinical monitoring to assess blood volume in critically ill patients. *Critical Care Medicine* **12**, 107–12.

Shoemaker WC, *et al*. (1988). Prospective trial of supranormal values of survivors as therapeutic goals in high-risk surgical patients. *Chest* **94**, 1176–87.

Wilson J, *et al*. (1999). Reducing the risk of major elective surgery: randomized, controlled trial of preoperative optimisation of oxygen delivery. *British Medical Journal* **318**, 1099–103.

16.3 Cardiac arrest

C. A. Eynon

Introduction

Cardiac arrest is a clinical syndrome consisting of unresponsiveness, absence of a detectable pulse, and either apnoea or agonal respiration. Cessation of cardiac activity is common to all causes of death and it is important to remember that cardiopulmonary resuscitation (**CPR**) was only developed to treat potentially reversible causes.

From the earliest recorded times it has been recognized that there is a period in which it is possible to reverse the transition from life to death. In Ancient Egyptian mythology, Isis, the goddess of healing, revived her husband Osiris by breathing into his mouth. In the Old Testament, Elijah successfully resuscitated an apparently dead child. The techniques that comprise modern CPR were all described before the end of the 19th century. Accounts of mouth-to-mouth ventilation appeared in the recommendations of the Dutch Humane Society in 1767, and in 1775 Abildgaard described the effects of electricity in first killing, and then resuscitating chickens. The earliest reports of closed-chest compression were in the late 1800s. It was only from the mid-1950s, however, that these individual modalities were combined and evolved into the concept of the 'chain of survival' (Fig. 1).

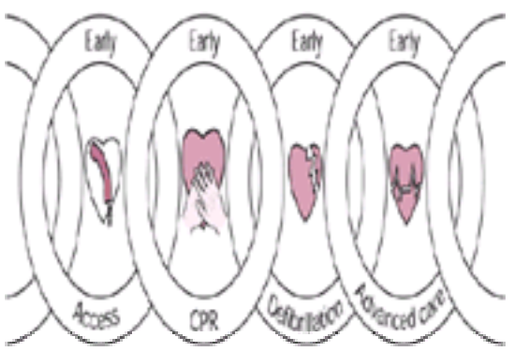

Fig. 1 The sequence of events in emergency cardiac care is displayed schematically by the 'chain of survival' metaphor. (From Cummins RO *et al*., 1991. Improving survival from sudden cardiac arrest: the 'chain of survival' concept. *Circulation* **83**, 1832–47.)

Aetiology

In the adult population, sudden cardiac arrest commonly results from ischaemic heart disease, 30 per cent of victims having evidence of recent myocardial infarction at autopsy. Other cardiac causes include primary arrhythmias, cardiomyopathies, and structural abnormalities. A wide range of non-cardiac conditions can precipitate a secondary cardiac arrest.

Three cardiac arrest rhythms are recognized: ventricular fibrillation (**VF**) or pulseless ventricular tachycardia (**VT**), asystole, and pulseless electrical activity (**PEA**). VF is characterized by a chaotic, uncoordinated waveform on the electrocardiogram (**ECG**). Asystole occurs when no ventricular electrical activity is present. Although atrial and ventricular asystole usually coexist, ventricular asystole may precede atrial asystole by a short time, producing an ECG in which there are isolated p waves. PEA occurs when there are clinical features of cardiac arrest despite an ECG rhythm that would normally be associated with a palpable pulse. Outside hospital, up to 75 per cent of cardiac arrests are due to VF/VT. In hospital, asystole and PEA are more common. This may reflect the greater prevalence of comorbid conditions in the hospital population.

Epidemiology

Sudden cardiac arrest remains the leading cause of unexpected death in the Western world, with 500 000 cases annually in the United States. With resuscitation the rate of return of spontaneous circulation is approximately 30 per cent for patients suffering an in-hospital cardiac arrest, but only around 15 per cent survive to discharge. Patients who suffer an out-of-hospital cardiac arrest have a worse outcome, with 8 to 22 per cent surviving to admission and 2 to 8 per cent being discharged. This variation in outcomes results from differences in emergency medical systems and study methodology. The main determinant of outcome is the initially documented cardiac rhythm. The survival rate for patients in VF/VT is 10 to 15 times that for asystole or PEA.

Pathophysiology

During cardiac arrest there is global ischaemia. Therapy is directed at maintaining perfusion of vital organs and re-establishing organized myocardial activity. Release of endogenous catecholamines and other vasoactive peptides causes redistribution of blood flow to the brain and heart and away from other organs. The brain is the organ most susceptible to ischaemia, and the rate of neurologically intact survival decreases rapidly to virtually zero at 20 min following cardiac arrest.

Myocardial ischaemia causes maximal coronary vasodilatation and myocardial blood flow becomes dependent on the coronary perfusion pressure: the pressure gradient between the aorta and right atrium during diastole. A coronary perfusion pressure of 15 mmHg during resuscitation appears to be the threshold for successful return of spontaneous circulation. Under normal conditions, coronary blood flow is autoregulated and coronary artery stenoses of up to 70 per cent do not compromise flow. However, during cardiac arrest autoregulation is lost and relatively insignificant lesions may reduce distal perfusion.

Following resuscitation there is a period of reversible organ dysfunction. Neurological impairment often persists for 12 to 24 h. Thereafter, two patterns emerge:

progressive improvement, or persisting coma and stable neurological deficit. Myocardial dysfunction contributes to the high mortality from arrhythmias and heart failure in the hours and days after resuscitation. Other organ systems are relatively resistant to periods of ischaemia compatible with successful resuscitation. Although multiple organ support may be necessary, permanent damage of organs other than the brain is uncommon in survivors.

Management

The current European Resuscitation Guidelines for adult basic and advanced life support are shown in Fig. 2 and Fig. 3. The most important factors affecting outcome are the times before institution of CPR, defibrillation, and advanced care. For VF/VT, survival rates correlate with the time to defibrillation. For asystole and PEA, the aim of treatment is to maintain organ perfusion until remediable causes for cardiac arrest can be identified and treated.

Fig. 2 European Resuscitation Council guidelines for adult basic life support. (From Handley AJ *et al*., 1998. The 1998 European Resuscitation Council guidelines for adult single rescuer basic life support. A statement from the Working Group on Basic Life Support, and approved by the executive committee of the European Resuscitation Council. *Resuscitation* **37**, 67–80. © ERC. Published by Elsevier Science Ireland Ltd.)

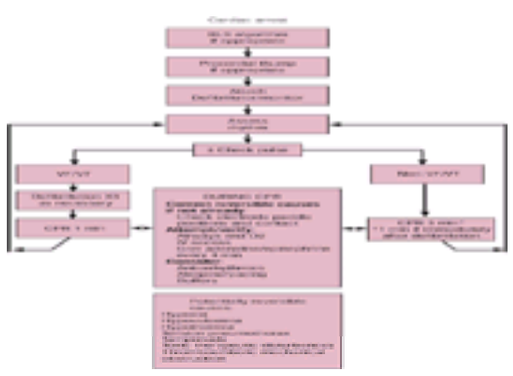

Fig. 3 European Resuscitation Council guidelines for adult advanced life support. (From Robertson C *et al*., 1998. The 1998 European Resuscitation Council guidelines for adult advanced life support. A statement from the Working Group on Advanced Life Support, and approved by the executive committee of the European Resuscitation Council. *Resuscitation* **37**, 81–90. © ERC. Published by Elsevier Science Ireland Ltd.)

Early access/assessment

Immediate activation of the emergency medical systems is essential once it has been determined that the patient is unresponsive. Campaigns have highlighted the need for witnesses of cardiac arrest to notify the emergency services correctly. Awareness of the appropriate telephone number is impaired by the use of different numbers in different countries. The Council of the European Communities has recommended that the number '112' should be used throughout Europe, but this has not been widely implemented, and telephone codes to summon the cardiac arrest team within hospitals have also not been standardized.

After activation of the emergency medical systems, assessment of the patient continues with **ABC** (airway, breathing, and circulation). Airway obstruction commonly results from loss of muscle tone in the tongue and jaw muscles, allowing the tongue and epiglottis to occlude the airway. In the absence of head or neck trauma, the airway should be opened using the head-tilt, chin-lift manoeuvre. The rescuer should then look for chest movement, listen for breath sounds, and feel for air movement on their cheek. Absence of breathing should be confirmed over 10 s. The patient should then be examined for signs of circulation for up to 10 s.

Basic life support

The objective of basic life support is to generate sufficient flow of oxygenated blood to the heart and brain until definitive therapy can be applied and spontaneous circulation re-established. Basic life support at least doubles the chances of survival if applied between the time of collapse and first defibrillation.

Mouth-to-mouth ventilation

Maintenance of the airway and ventilation may be performed concurrently using the mouth-to-mouth method. The rescuer's lips are placed around the patient's mouth and the patient's nose is sealed by pinching the nostrils together. The recommended tidal volume is 400 to 600 ml. Current recommendations are for two ventilations to 15 chest compressions. Normal expired air contains 15 to 17 per cent oxygen and 4 per cent carbon dioxide. Hyperventilation increases the oxygen concentration to 18 per cent and reduces the carbon dioxide concentration to 2 per cent. Enough oxygen can be administered in this manner to maintain an adequate arterial oxygen saturation in the early stages of cardiac arrest.

Closed-chest compression

Closed-chest compressions are performed by placing the heel of one hand over the lower sternum, two fingerbreadths from the xiphisternal junction. The heel of the second hand is placed over the first. The sternum is compressed 4 to 5 cm in the adult, maintaining compression for 50 per cent of the compression–relaxation cycle. The rate should be 100 compressions per minute. Properly performed closed-chest compressions can produce peak systolic arterial pressures of 60 to 80 mmHg, but diastolic arterial pressure and coronary perfusion pressure fall rapidly after the first few minutes. Cardiac output during closed-chest compression ranges from a quarter to a third of normal.

The mechanism leading to blood flow during closed-chest compressions is still debated. The cardiac pump hypothesis proposes that direct compression of the heart between the sternum and the paraspinal structures results in ejection of blood. The aortic and pulmonary valves open during compression, whilst closure of the mitral and tricuspid valves prevents regurgitation of blood. As compression is released, intracardiac pressures fall and the mitral and tricuspid valves open, promoting ventricular filling. In contrast, the thoracic pump hypothesis proposes that the entire thorax is the pump, with the heart being a passive conduit and the pulmonary vascular bed acting as a reservoir for blood. This theory is supported by the finding that patients in cardiac arrest can maintain prolonged consciousness by forceful coughing. Coughing increases intrathoracic pressure which results in antegrade blood flow. Also favouring this theory are studies showing no significant reduction in left ventricular dimensions and patent heart valves, during closed-chest compression.

The major complication from closed-chest compression is trauma. Rib and sternal fractures occur in up to 30 per cent of patients, even with well-trained rescuers.

Incorrectly performed compressions may injure the thoracic or abdominal contents.

Adjuncts to standard CPR

The airway may be maintained using either nasopharyngeal or oropharangeal airways. These hold the base of the tongue away from the posterior pharynx. Pocket masks for mouth-to-mask ventilation reduce possible spread of pathogens and can incorporate supplementary oxygen. Bag–valve–mask devices with a reservoir bag can deliver an inspired oxygen concentration of over 90 per cent.

Alternative methods of closed-chest compression

The poor survival rate following closed-chest compression has driven the search for more effective methods. Interposed abdominal counterpulsation aims to increase venous return to the heart prior to chest compression. Active compression–decompression CPR uses a device applied to the chest wall to reduce intrathoracic pressure and increase venous return during decompression. Vest CPR uses a circumferential vest around the thorax. Sequential inflation and deflation alters intrathoracic pressure resulting in blood flow. Although promising in animal trials and limited human studies, no survival benefit has been seen in larger studies for any of these methods.

Open-chest cardiac massage

The ease of closed-chest compression led to it quickly supplanting open-chest cardiac massage without trials to demonstrate its superiority. Experimental evidence, coupled with reports of patient survival using open-chest cardiac massage after prolonged periods of unsuccessful closed-chest compression, have led to a resurgence of interest in the open-chest method. The American Heart Association recommends open-chest cardiac massage for cardiac arrest associated with penetrating trauma, hypothermia, pulmonary embolus, pericardial tamponade, abdominal haemorrhage, or where chest or vertebral anomalies prevent effective closed-chest compression.

Advanced life support

Defibrillation

The majority of survivors of cardiac arrest are in VF/VT. Defibrillation is the only effective method of terminating these rhythms and its effectiveness falls rapidly with time (Fig. 4). Over 80 per cent of patients successfully resuscitated from VF/VT are resuscitated by one of the first three shocks. A precordial thump may terminate VF/VT and should be considered for witnessed arrests when a defibrillator is not immediately available.

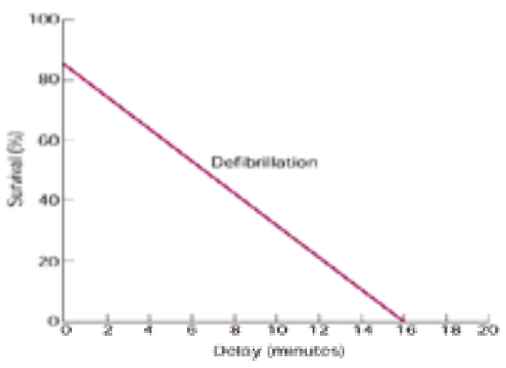

Fig. 4 Effect of time on success of defibrillation. (From Colquhoun MC, Handley AJ, Evans TR, 1998. *ABC of resuscitation*, 3rd edn. © BMJ Books.)

Mechanism of defibrillation

The mechanism of defibrillation is unknown. The critical mass hypothesis suggests that defibrillation occurs when sufficient current passes through the heart to depolarize approximately 75 per cent of the myocardium. This causes arrest of the activating wave and allows a normal rhythm to be re-established. The upper limit of vulnerability theory proposes that fibrillation terminates only when the strength of shock is greater than the threshold limit for the whole myocardium. Sub-threshold shocks induce refractoriness in susceptible tissue but lead to new activation waveforms from regions excited by the shock. This second theory is supported by the finding of a period of electrical silence after a countershock, followed by resumption of fibrillation with a different morphology.

As little as 4 per cent of the current applied using external defibrillation actually crosses the heart. The current applied depends on the energy selected and the impedance of the thorax. Transthoracic impedance is altered by electrode size, the use of couplants (usually preformed gel pads) between electrodes and chest wall, multiple shocks, the phase of respiration, and the pressure with which the electrodes are applied.

The recommended energy level for the initial shock is 200 joules. This aims to produce a high success rate whilst minimizing the risk of damage to the heart. If a second shock is required it should also be 200 J. Transthoracic impedance reduces with successive shocks and each shock should deliver more energy to the myocardium. Subsequent shocks are delivered at 360 J, which is the maximum available on most defibrillators. The most commonly used electrode placement is apex–anterior with one placed to the right of the upper sternum below the clavicle, and the other over the apex of the heart. Alternative placements may result in successful defibrillation in some patients.

Current-based defibrillation uses a microprocessor to measure transthoracic impedance. The operator selects the required current and the defibrillator charges a capacitor to the energy required to deliver that current. Using biphasic waveforms rather than the standard damped sinusoidal waveform may reduce energy requirements. Both of these methods have shown promise in early clinical trials.

Automated external defibrillators

It is important that the first responder at a cardiac arrest can defibrillate. Even small reductions in the time to first shock can dramatically improve survival. Interpretation of ECG rhythms is difficult even for medical professionals and can delay defibrillation. Automated external defibrillators are capable of analysing the ECG rhythm and advise that a countershock be given if VF is present. Fully automated machines deliver the shock after offering an audible warning. By simplifying the training required, they may be used by a wide range of non-medical personnel.

Early advanced care

Advanced care is required if CPR and defibrillation fail to achieve or sustain a spontaneous circulation.

Advanced airway support

Endotracheal intubation remains the gold standard of airway management. Intubation maintains and protects the airway, permits administration of additional oxygen and certain drugs, and allows adjustment of ventilation. Successful intubation requires a high level of skill and regular practice. Laryngeal mask airways are commonly used in anaesthesia and increasingly for airway management during cardiac arrest. Their use requires a lower level of training than does intubation.

Drug therapy

Peripheral venous cannulation is the quickest method of administering medication during cardiac arrest. Peak drug levels are lower and circulation times are longer using peripheral rather than central routes. If there is delay in obtaining venous access, adrenaline, atropine, and lignocaine may be administered via the endotracheal tube. They should be given at 2 to 3 times the normal intravenous dose and diluted in 10 ml of normal saline. Drug absorption and pharmacodynamics are less predictable when given via the endotracheal route.

Vasopressors

Adrenaline has not been proved to improve survival from cardiac arrest. The standard dose of 1 mg originates from studies using 0.1 mg/kg in 10 kg dogs. This was translated into human use without evidence of comparable efficacy. In a dose range of 0.045 to 0.2 mg/kg (3 to 14 mg in a 70 kg man), adrenaline improves the arterial pressures generated during CPR and increases myocardial and cerebral blood flow. High-dose adrenaline has failed to show any survival benefit compared with standard dosage. Currently, the majority of studies support the use of adrenaline for patients who remain in cardiac arrest after CPR and defibrillation.

No other adrenergic agonist has been shown to have advantages over adrenaline in the treatment of cardiac arrest. The vasoconstrictor vasopressin has been used in limited clinical trials with encouraging results.

Antiarrhythmics

The advantage of routine use of antiarrhythmics in cardiac arrest is unproved. Although bretylium and lignocaine are effective in suppressing arrhythmias after myocardial infarction, they have not shown clear benefit in the setting of cardiac arrest. Amiodarone has been used in trials of persistent ventricular arrhythmias with reported improvement. Magnesium sulphate is the treatment of choice for torsade de pointes. Empirical magnesium supplementation has not, however, been shown to be of benefit in cardiac arrest.

Atropine is of value in haemodynamically compromising bradycardia, but the recommendation for its use in asystole is based on limited data. It is hypothesized that excessive vagal tone may inhibit cardiac action. Small-scale studies suggest that atropine has a favourable effect on cardiac rhythm.

Buffers

Cardiac arrest causes a mixed respiratory and metabolic acidosis due to retention of carbon dioxide and lactic acid. Low blood flow during CPR often causes differences between venous and arterial samples with a mixed venous acidosis despite arterial hypocarbic alkalosis. No clinical data support correction of the acidosis by means other than adequate ventilation and tissue perfusion. Addition of bicarbonate causes further increases in venous carbon dioxide and venous respiratory acidosis. Sodium bicarbonate also has the theoretical disadvantages of hyperosmolarity and hypernatraemia. Alternative buffering agents have shown no benefit. Buffers remain of use in cases of cardiac arrest associated with tricyclic antidepressant overdose and pre-existing metabolic acidosis.

Pacing

External pacing of asystole may be beneficial in the small group of patients who arrest shortly before pacing is instituted. For the majority, application of external pacing is too late.

Post-resuscitation care

Following return of spontaneous circulation the cause of the cardiac arrest should be ascertained and tissue perfusion optimized. Emergency thrombolysis or angioplasty may offer benefit for patients with acute myocardial infarction. Prolonged coma following resuscitation is associated with poor outcome. Neurological assessment after 72 h can predict the likelihood of long-term survival. Trials to evaluate treatment that might reduce cerebral damage have been disappointing. Myocardial dysfunction accounts for the majority of early deaths after return of spontaneous circulation. The only method of minimizing post-resuscitation sequelae is to strengthen the chain of survival.

Training in life support measures

Most sudden cardiac arrests occur in the community. For these patients, the highest rates of survival have been achieved when CPR is initiated within 4 min. Well-performed basic life support is more successful than poorly performed basic life support, and some basic life support is better than none. Training in the technique is well established in many communities. The use of automated external defibrillators is now included in the American Heart Association basic life support programme.

Ethics of resuscitation

When resuscitation is not attempted, death is virtually inevitable. Only with evidence that resuscitation is not indicated should it be omitted.

Advanced directives or living wills are becoming more common, especially in the United States. These convey the patients' wishes with respect to specific treatments when they are unconscious or incapacitated. If possible, decisions not to attempt resuscitation (DNR orders) should be made after discussion with the patient. If the patient is not competent to make the decision, factors to be considered include the quality of life prior to the illness, the expected quality of life assuming recovery, and the likelihood of successful resuscitation.

Although doctors are not required to offer a treatment if it is deemed futile, determinations of futility vary greatly between physicians. Resuscitation has been deemed futile for the following reasons.

1. An adequate trial of basic and advanced life support has already been attempted without return of spontaneous circulation.
2. No physiological benefit from basic or advanced life support can be expected because the patient's vital functions are deteriorating despite maximum therapy.
3. No survivors after cardiopulmonary resuscitation have been reported under the given circumstances.

The presence of sepsis, disseminated cancer, or major organ failure is associated with very poor outcome. Lower survival rates in the elderly result from a higher presence of comorbidity rather than from age itself. In practice, doctors must determine the merit of attempting CPR for each patient. If resuscitation is attempted but return of spontaneous circulation does not occur promptly, consideration must be given to termination of the attempt. This decision depends on a number of factors in addition to those cited above.

1. Time intervals to initiation of basic life support and defibrillation—delays of over 5 min to basic life support or 30 min to defibrillation are associated with extremely poor prognosis.
2. Evidence of cardiac activity—termination of resuscitation effort after 15 to 20 min has been suggested for non-VF/VT arrests.
3. Protective factors—hypothermia or ingestion of sedatives, hypnotics, or narcotics confer a measure of protection from the sequelae of cardiac arrest.

Relatives and resuscitation

The presence of relatives in the resuscitation room remains controversial. Whilst common in paediatric practice it is not routine in adult cardiac arrest. However, if properly managed witnessing resuscitation does no harm and probably aids the grieving process should the patient not survive.

Practical skill training using the recently dead

Mannikins have largely replaced the use of the recently dead for training in the practical skills used in resuscitation.

Future developments

End-tidal carbon dioxide levels ($ETCO_2$)

$ETCO_2$ meters are widely used to confirm correct endotracheal tube placement. During cardiac arrest, $ETCO_2$ levels depend primarily on the cardiac output generated by resuscitation. A value of 10 mmHg at 20 min can distinguish patients who may survive from ones who will not. Use of $ETCO_2$ may clarify when resuscitation attempts can be deemed futile.

Invasive support

Cardiopulmonary bypass

Whilst cardiopulmonary bypass is a proven treatment for cardiac arrest associated with hypothermia, its role in routine arrests is likely to be limited by cost as well as the time required to achieve bypass.

Minimally invasive direct cardiac massage

In this technique a plunger device is placed directly on to the pericardium via a small thoracotomy. Manual compression of the ventricles provides an artificial circulation.

Anstadt cup

This is a biventricular assist device placed around the heart. Its role is limited by the requirement for significant surgical skill and equipment.

Retroaortic perfusion

This involves insertion of an aortic occlusion balloon into the descending aorta and infusion of oxygenated blood or blood substitutes into the proximal aorta.

Further reading

Advanced Life Support Working Group of the International Liaison Committee on Resuscitation (1997). Early defibrillation. An advisory statement by the Advanced Life Support Working Group of the International Liaison Committee on Resuscitation. *Resuscitation* **34**, 113–15.

Becker LB (1996). The epidemiology of sudden death. In: Paradis NA, Halperin HR, Nowak RM, eds. *Cardiac arrest: the science and practice of resuscitation medicine*, pp 28–47. Williams & Wilkins, Baltimore.

British Medical Association (1993). Decisions related to cardiopulmonary resuscitation. A statement from the BMA and RCN in association with the Resuscitation Council (UK). BMA, London.

Handley AJ *et al*. (1998). The 1998 European Resuscitation Council guidelines for adult single rescuer basic life support. A statement from the Working Group on Basic Life Support, and approved by the executive committee of the European Resuscitation Council. *Resuscitation* **37**, 67–80.

Paradis NA *et al*. (1989). Simultaneous aortic, jugular bulb, and right atrial pressures during cardiopulmonary resuscitation in humans: insights into mechanisms. *Circulation* **80**, 361–8.

Paradis NA *et al*. (1990). Coronary perfusion pressure and the return of spontaneous circulation in human cardiopulmonary resuscitation. *Journal of the American Medical Association* **263**, 1106–13.

Robertson C *et al*. (1998). The 1998 European Resuscitation Council guidelines for adult advanced life support. A statement from the Working Group on Advanced Life Support, and approved by the executive committee of the European Resuscitation Council. *Resuscitation* **37**, 81–90.

16.4 Anaphylaxis

Anthony F. T. Brown

Introduction

The term anaphylaxis was introduced by Richet and Portier in 1902, literally meaning 'against protection'. It is currently used to describe the rapid, generalized, and often unheralded IgE immunologically mediated events that follow exposure to some foreign substances in those who have previously been sensitized. An identical clinical syndrome involving release of similar potent mediators but not triggered by IgE, and thus not necessarily requiring previous exposure, is termed an anaphylactoid reaction. Despite aetiological differences of vital importance to subsequent follow-up and prevention, the clinical term anaphylaxis is often used to describe both of these syndromes (as in this chapter).

Anaphylaxis is potentially the most severe of the immediate-type hypersensitivity reactions. It is the quintessential medical emergency that may be mild or severe, gradual in onset or fulminant, involve multiple organ systems or cause isolated wheeze or shock. It often occurs without warning in otherwise healthy people, and since there is no immediate confirmatory laboratory test it mandates prompt clinical recognition and treatment to prevent death from hypoxia or hypotension in severe cases.

Epidemiology

The true incidence and prevalence of anaphylaxis are unknown. The literature is heterogeneous, retrospective, and utilizes variable definitions. Emergency department presentations range from 1 in 440 to 1 in 1500 attendances, representing an incidence from 1 per 3400 to 1 per 10 000 catchment population per year. Hospital discharge data suggest 10 per 100 000 admissions are due to anaphylaxis. No age, sex, race, or locality is exempt, with cases occurring from infancy to old age.

Atopy increases the risk of idiopathic anaphylaxis and that caused by ingested antigens, latex, radiocontrast media, and exercise. Asthma increases the risk of dying from anaphylaxis. Patients taking b-blocking drugs develop more severe reactions which may be resistant to therapy, and those taking angiotensin-converting enzyme (ACE) inhibitor drugs are particularly prone to angioedema that can be life threatening. The newer angiotensin-II receptor antagonists cause fewer reactions but are also implicated.

Deaths from anaphylaxis are rare, although penicillins, hymenopteran stings, and radiocontrast media account for the majority.

Aetiology

IgE-mediated anaphylaxis

A vast array of IgE immune-mediated triggers are recognized (Table 1), the most important being b-lactam antibiotics including the penicillins and cephalosporins, hymenopteran stings such as wasps and bees, and foods.

Antibiotics

Penicillins are the most common cause of anaphylaxis in adults, occurring after 1 in 5000 parenteral doses, and are the most frequent cause of death. True cross-sensitivity to the cephalosporins is considerably less than 10 per cent and largely confined to the first-generation cephalosporins. Fatalities to oral penicillins are exceedingly rare.

Hymenoptera

Reactions to hymenopteran venom, such as from wasps, bees, hornets, and fire ants, are second only to antibiotics in frequency, occurring in up to 3 per cent of the population. Large local reactions, toxic reactions, and late serum sickness-like, non-IgE reactions may also occur following a sting.

Foods

Peanuts, other legumes, true nuts, shellfish, milk, and eggs cause the largest number of food-related cases, particularly in children. Reactions occur rapidly and may recur several hours later (biphasic reaction). Cross-reactivity to seemingly unrelated plants is seen. Mislabelling and contamination at the manufacturing stage or in the home cause inadvertent exposure.

Latex

Allergy to latex (rubber) is seen particularly in hospital personnel, children with spina bifida and genitourinary abnormalities, and in workers with occupational exposure. It may occur by contact, parenteral administration, and aerosol transmission and is one cause of perioperative anaphylaxis, along with thiopentone, neuromuscular blocking drugs, antibiotics, protamine, and anaphylactoid reactions to opiates, colloids, or blood.

Non-immune, anaphylactoid reactions

Various less well understood, non-immunological mechanisms lead to the release of similar inflammatory mediators in anaphylactoid reactions. Agents responsible include aspirin, non-steroidal anti-inflammatory drugs (**NSAIDs**), radiocontrast media, and opiates, and in the case of infusions may relate to their rate, concentration, and volume delivered (see Table 2).

Radiocontrast media

These were the most common cause of anaphylactoid reactions, but the incidence has declined to less than 1 in 200 patients receiving the newer low osmolality, non-ionic agents. Asthma, atopy, and patients on b-blockers or with prior reactions are at greatest risk.

Aspirin and NSAIDs

Bronchospasm following these is common in patients with nasal polyps and reactive airway disease, but systemic anaphylactoid reactions may occur in their absence following cyclo-oxygenase inhibition. Cross-reactivity occurs between different NSAIDs and rare reactions have occurred to the newer COX-II (cyclo-oxygenase II) inhibitors.

Pathogenesis

Irrespective of the triggering event, two main groups of mediators are released by mast cells and basophils. These groups include the preformed, granule-associated mediators histamine, neutrophil and eosinophil chemotactic factors, enzymes such as tryptase and b-glucuronidase, and proteoglycans such as heparin. They also include newly synthesized mediators from arachidonic acid metabolism via the cyclo-oxygenase pathway, such as prostaglandin D_2 and thromboxane A_2, and via the lipoxygenase pathway, such as the leukotrienes. In addition, platelet-activating factor and cytokines such as the interleukins are also rapidly formed.

All these mediators act by inducing vasodilatation, increasing capillary permeability and glandular secretion, causing smooth muscle spasm (including bronchoconstriction), and by attracting new cells to the area (see Chapter 5.2 for further discussion).

Clinical features

Anaphylaxis is typically a disease of those who are fit and is rarely seen or described in patients who are critically ill or shocked other than those suffering from asthma. The speed of onset of symptoms and signs is related to the severity of the process, with life-threatening reactions occurring in minutes to parenteral antigen exposure, although deaths have been associated with oral, topical, or cutaneous triggers. Most symptoms occur within 30 min, but may be delayed for some hours, particularly following oral or topical exposure.

Over half the deaths from anaphylaxis occur within the first hour of onset. Seventy five per cent result from asphyxia due to upper airway oedema and hypoxia from severe bronchospasm. The remaining 25 per cent are due to circulatory failure with shock related to vasodilatation and hypovolaemia from plasma volume losses, cardiac arrhythmias, pulmonary hypertension, and (possibly) decreased myocardial activity in the absence of cardiac disease.

Ninety five per cent of patients have cutaneous features, but these can sometimes be absent, particularly in cases presenting with the rapid onset of laryngeal oedema or circulatory shock, or resolve spontaneously or in response to prehospital treatment.

Cutaneous features

A premonitory aura, tingling, or warm sensation may precede generalized erythema, urticaria with pruritus, local oedema or angioedema, and occasionally pallor. Rhinitis and conjunctivitis are also seen.

Respiratory system features

Sudden cough, hoarseness, throat tightness, or the sensation of a 'lump' may precede shortness of breath, dyspnoea, stridor, wheeze, and cyanosis due to oropharyngeal or laryngeal oedema and bronchospasm.

Cardiovascular and neurological features

Apprehension, light-headedness, dizziness, or syncope may precede or accompany cardiovascular collapse with tachycardia, hypotension, and cardiac arrhythmias. Hypoxia, hypoperfusion, and the direct effect of mediators lead to incontinence, confusion or coma, and myocardial or cerebral ischaemia or infarction.

Gastrointestinal and miscellaneous features

Cramping abdominal pain, nausea, vomiting, and diarrhoea are common, but the more obviously life-threatening features described above usually overshadow these gastrointestinal manifestations. In rare instances there may be back pain, watery vaginal discharge, pulmonary oedema, and even disseminated intravascular coagulation.

Differential diagnosis

The protean manifestations of anaphylaxis allow a potentially vast differential diagnosis, although the rapid onset, accompanying cutaneous feature, and relationship to a potential trigger suggest the diagnosis of anaphylaxis in most cases. However, the following alternative diagnoses may be considered.

Facial swelling or oedema

These may result from bacterial or viral infection, although fever and pain should predominate. Traumatic or spontaneous bleeding, particularly in patients on warfarin, usually causes recognizable bruising.

Wheeze and difficulty breathing

Bronchial asthma, cardiogenic pulmonary oedema, foreign body inhalation, irritant chemical exposure, and tension pneumothorax should be distinguished on the basis of associated presenting features.

Light-headedness and syncope

A vasovagal reaction must be considered in the context of a painful procedure such as an injection or local anaesthetic infiltration: bradycardia, pallor, and rapid response to a recumbent position are usual. Panic attacks are common in allergic patients confronted by an unexpected, potential allergen exposure.

Angioedema

Angioedema in the absence of urticaria may be caused by actual or functional C_1 esterase inhibitor deficiency. This may be hereditary with autosomal dominant inheritance or acquired related to lymphoproliferative disorders. A family history, the absence of pruritus, prominence of abdominal symptoms, and recurrent attacks suggest the hereditary cause. C_1 esterase inhibitor concentrate or fresh frozen plasma is used to treat recalcitrant cases.

Investigation

The diagnosis of anaphylaxis is clinical. No immediate laboratory or radiological tests confirm the process and investigation must not delay immediate management.

Disease progress may be monitored by pulse oximetry, arterial blood gases (looking for metabolic or respiratory acidosis), haematocrit level (may rise with extravasation of fluid), and measurement of electrolytes and renal function. These tests together with blood glucose level, chest radiograph, and ECG are necessary if there is a slow response to therapy or when there is doubt about the diagnosis.

Mast cell tryptase

The only direct marker of mast cell activation is an elevated serum tryptase level. Tryptase is released in both anaphylactic and anaphylactoid reactions, beginning to rise within 30 min and remaining high for up to 6 h. It is of value when the diagnosis of anaphylaxis is uncertain clinically, particularly during anaesthesia, and may also be useful after death. Frozen serum specimens must be sent to a specialist laboratory.

Management

Patients with anaphylaxis may present directly to their family doctor or to the emergency department, or the reaction may start in hospital in the radiology department, theatre, on the ward, or even in the outpatient department. Cutaneous features of generalized mediator release may be the first signal, followed by more serious systemic symptoms or signs.

Any causative agent, such as an intravenous drug or infusion, must be stopped immediately. The patient should initially be managed in a monitored resuscitation area, or equipment including at least a pulse oximeter, non-invasive blood pressure device, and ECG monitor brought to them.

After a brief history of possible allergen exposure is obtained, a rapid assessment of the extent and severity of the reaction must be made, particularly looking for upper and lower respiratory tract involvement and the early signs of shock.

Oxygen, adrenaline, and fluids are the mainstay of treatment, with antihistamines and steroids only utilized as second-line agents once the cardiorespiratory status has been stabilized (see Table 3 and Table 4). A more detailed exposure history may be taken at this time.

First-line treatment

Oxygen and airway patency

Oxygen by face mask should be given to all patients, aiming to maintain an oxygen saturation above 92 per cent. Allow the patient to remain upright unless shocked, and call urgently for experienced anaesthetic help if there are signs of impending airway obstruction such as worsening hoarseness or stridor, or rapidly progressive respiratory failure with tachypnoea and wheeze. Cyanosis and exhaustion indicate imminent respiratory arrest, but sedative or muscle relaxant drugs should not be given unless the physician is competent in the management of a difficult airway, since endotracheal intubation and mechanical ventilation may be extremely difficult. As a last resort, a surgical airway via the cricothyroid membrane should be established before hypoxic cardiac arrest occurs.

Adrenaline

Adrenaline is the drug of choice and should be given in all but the most trivial cases, particularly to patients with airway swelling, bronchospasm, or hypotension. It has beneficial a-, b_1-, and b_2-adrenergic effects, including a rise in intracellular cyclic AMP that inhibits further mast cell and basophil mediator release.

When anaphylaxis is treated early, is mild or progressing slowly, if venous access is difficult or delayed, or in an unmonitored patient, give 0.3 to 0.5 ml of 1:1000 adrenaline (0.3 to 0.5 mg) intramuscularly in adults. This has advantages in terms of safety and is usually rapidly effective. The dose may be repeated every 5 to 10 min or longer according to response, and is preferred to the subcutaneous route.

However, in serious or quickly progressing cases, particularly in the presence of vascular collapse and shock, marked airway compromise, or severe bronchospasm, intravenous adrenaline is essential to achieve more rapid and reliable delivery. Although 1:10 000 adrenaline containing 100 µg/ml is readily available, for instance as the Min-I-Jet preparation, it is difficult to give this slowly in small quantities titrated to response. An alternative is to prepare a 1:100 000 dilution containing 10 µg/ml, and to give 1 to 2 ml (10 to 20 µg) per minute at an initial dose of 0.75 to 1.5 µg/kg (see Table 3). Intravenous adrenaline must only be given under ECG control in a monitored resuscitation area by doctors experienced in its use.

Whilst parenteral adrenaline is being prepared, patients may be given nebulized 1:1000 adrenaline from 1 to 4 ml (1 to 4 mg) via an oxygen-driven system. This may dramatically improve upper airway oedema or bronchospam.

Fluids

A large-bore intravenous cannula should be inserted as soon as possible in patients with shock to administer a 10 to 20 ml/kg fluid bolus. Gelatin preparations such as Haemaccel or Gelofusine are best, although normal saline in larger quantities is suitable and there are no outcome data favouring one infusion over another. Further fluid boluses are indicated by the clinical response, although it is important to give adrenaline in addition.

Second-line treatment

Once the cardiorespiratory status and tissue oxygenation have been improved with oxygen, adrenaline, and fluids, other drugs may be given in a supporting role.

Antihistamines

Although popular, antihistamines must never be relied upon as sole therapy. They are only indicated: when the reaction is not life threatening; when it is progressing slowly with predominant cutaneous features such as angioedema and urticaria; to prevent later recrudescence of symptoms or signs; as pretreatment to prevent reactions to radiocontrast media or volume expanders; and finally as prophylaxis during anaesthesia.

Newer, second-generation 'non-sedating' H_1-antihistamines such as loratadine and cetirizine are available orally, but still may cross the blood–brain barrier to cause drowsiness, particularly with alcohol. H_2-antihistamines have been used successfully in protracted anaphylaxis with shock, and increasingly now in combination with H_1-antihistamines. H_1- and H_2-antihistamines given together are supported by improved outcome in both prevention and treatment of acute anaphylaxis, combined with steroids in some instances.

Steroids

Steroids are of limited value despite many theoretical beneficial effects on mediator release and tissue responsiveness. They may prevent or shorten protracted

reactions, particularly those associated with bronchospasm, and may reduce the likelihood of relapse of symptoms, particularly after discharge. They are, however, essential in the management of recurrent idiopathic anaphylaxis.

Salbutamol, aminophylline, and glucagon

Nebulized salbutamol may be given for bronchospasm and has the advantage of familiarity. Aminophylline intravenously has additive effects to adrenaline in refractory bronchospasm. Glucagon may be used, particularly in patients already taking b-blockers, who at the same time may be both therapeutically resistant yet overly sensitive to unopposed a-mediated effects of adrenaline.

Admission, observation, and follow-up

Patients with cutaneous features alone who remain well may be discharged after a short 3- to 4-h period of observation, dependent on the nature and time of the suspected allergen exposure and their response to treatment.

Patients with significant systemic anaphylactic reactions, including all those receiving adrenaline, must be observed for a minimum of 6 to 8 h after apparent full recovery, as late deterioration may occur in 1 to 5 per cent of cases (biphasic response).

Those patients with unstable vital signs or with protracted or resistant anaphylaxis should remain monitored and be admitted to an intensive care area. An adrenaline infusion at 1 to 10 µg/min may be needed as a temporizing measure, and can be given by adding 1 ml of 1:1000 adrenaline to 100 ml of normal saline and infusing this at 6 to 60 ml/h.

Discharge medication

Following a successful period of observation, patients discharged from the emergency or outpatient department should be given combined oral H_1-and H_2-antihistamines and steroids for 2 to 3 days, such as chlorpheniramine at 4 mg 6-hourly, ranitidine at 150 mg 12-hourly, and prednisolone at 50 mg once daily in adults.

Patients who have had a life-threatening reaction, particularly to a food or insect sting, should also be prescribed a self-administered adrenaline syringe. Several devices are available, including the EpiPen that delivers 300 µg intramuscularly via a pressure-activated, spring-loaded needle, or the EpiPen Jr delivering 150 µg. It is essential that the patient and their immediate family are taught how and when to use the device before they are discharged from hospital, and appropriate early follow-up with an allergy specialist must be arranged.

Allergy referral

All patients who have suffered a severe anaphylactic reaction and those with significant attacks, particularly if they are recurrent or the stimulus is unknown or unavoidable, should be referred to an allergy specialist. This must include all those prescribed an EpiPen, and patients suitable for immunotherapy following a wasp or bee sting. Expert knowledge is required to determine whether skin testing, *in vitro* IgE tests, and/or challenge tests should be used to confirm the suspected cause.

A detailed letter of the nature and circumstances of the anaphylactic reaction, the treatment given, and the suspected causative agent(s) should accompany the patient home. If the cause of the reaction was unclear, ask the patient to write a brief diary of events of the immediate 6 to 12 h preceding the reaction, including all foods ingested, drugs taken (including non-proprietary), cosmetics used, and so on, and all activities performed outside as well as indoors. Later recall of events will be flawed unless documented.

Prevention

Education

Patient education about the nature and cause of the reaction and the relevance of carrying an EpiPen is fundamental. Individualized antigen elimination measures must be carefully explained, including hymenopteran avoidance, and recognizing hidden or unexpected sources of antigen such as salicylate in over-the-counter preparations and trace food elements such as nuts, plus possible cross-reactions to unrelated antigens. Latex-sensitive patients may require all future medical care in a latex-controlled environment.

An alert bracelet such as Medic-Alert should be worn, particularly following a severe reaction that may recur and render the patient unable to give a history. This should highlight drug or vaccine allergy to avoid inadvertent iatrogenic exposure.

Pretreatment

Pretreatment is only helpful in limited situations, for example prednisolone at 50 mg orally, with or without an antihistamine, 12 and 2 h prior to radiocontrast media in patients with asthma or those with previous reactions, and 1:1000 adrenaline at 0.25 ml subcutaneously prior to polyvalent snake antivenom. In the latter situation antihistamines are of no proven value.

Skin testing and short-term desensitization

In some clinical circumstances, such as when a penicillin is considered essential and there is a history of possible penicillin allergy, skin testing followed—if positive—by short-term desensitization over several hours with increasing doses at 15-min intervals may be instituted under strict medical control in a monitored area.

Long-term desensitization (immunotherapy)

Hyposensitization immunotherapy is principally reserved for wasp and bee allergy, as these preventable reactions may become life threatening. b-Blockade is a contraindication, and asthma or ACE inhibitor use require careful risk–benefit evaluation. Therapy needs to be continued for at least 3 to 5 years.

Drug avoidance

Wherever possible give therapy orally, and if intravenous use is chosen, always administer a drug slowly. Avoid drugs known to predispose to reactions, particularly aspirin and NSAIDs, as well as b-blockers and ACE inhibitors. Patients at risk of anaphylaxis with hypertension or ischaemic heart disease should ideally be taken off b-blockers, and care taken not to substitute an ACE inhibitor.

Conclusion

Anaphylaxis is a common clinical and diagnostic challenge for physicians. Prompt treatment with oxygen, adrenaline, and fluids to restore cardiorespiratory stability is followed with second-line therapy such as antihistamines and steroids. Proactive discharge planning with allergy referral when appropriate protects against further unheralded or potentially avoidable attacks.

Further reading

Bochner BS, Lichtenstein LM (1991). Anaphylaxis. *New England Journal of Medicine* **324**, 1785–91.

Brown AFT (1998). Therapeutic controversies in the management of acute anaphylaxis. *Journal of Accident and Emergency Medicine* **15**, 89–95.

Fan HW *et al*. (1999). Sequential randomised and double blind trial of promethazine prophylaxis against early anaphylactic reactions to antivenom for bothrops snake bites. *British Medical Journal* **318**, 1451–3.

Hollingsworth HM, Giansiracusa DF, Upchurch KS (1991). Anaphylaxis. *Journal of Intensive Care Medicine* **6**, 55–70.

Joint Task Force on Practice Parameters (1998). The diagnosis and management of anaphylaxis. *Journal of Allergy and Clinical Immunology* **101**, S465–S528.

Krause RS. Anaphylaxis. http://www.emedicine.com/emerg/topic25.htm (accessed Aug 7th, 2001).

Lin RY *et al*. (2000). Improved outcomes in patients with acute allergic syndromes who are treated with combined H_1 and H_2 antagonists. *Annals of Emergency Medicine* **36**, 462–8.

O'Brien J, Howell JM (2000). Allergic emergencies and anaphylaxis: How to avoid getting stung. *Emergency Medicine Practice: An Evidence-Based Approach to Emergency Medicine* **2**(4), 1–20.

Premawardhena AP *et al.* (1999). Low dose subcutaneous adrenaline to prevent acute adverse reactions to antivenom serum in people bitten by snakes: randomised, placebo controlled trial. *British Medical Journal* **318**, 1041–3.

Project Team of the Resuscitation Council (UK) (1999). The emergency medical treatment of anaphylactic reactions. *Resuscitation* **41**, 93–9.

Project Team of the Resuscitation Council (UK) (2000). Update on the emergency medical treatment of anaphylactic reactions for first medical responders. *Resuscitation* **48**, 241–3.

Simons FER *et al.* (1998). Epinephrine absorption in children with a history of anaphylaxis. *Journal of Allergy and Clinical Immunology* **101**, 33–7.

16.5.1 Pathophysiology and pathogenesis of acute respiratory distress syndrome

C. Haslett

Introduction

The acute respiratory distress syndrome (**ARDS**) is a form of acute inflammatory lung injury, initiated as part of an injurious, generalized, systemic inflammatory microvasculature response that may also result in failure of other organs (multiple organ failure). It occurs in an unpredictable fashion after a 'latent' period of several hours or days following a wide range of predisposing events, which may injure the lung directly (for instance, severe pneumonia, acid inhalation, toxic gas inhalation, near-drowning), or indirectly (for example, multiple trauma, sepsis, pancreatitis). Despite the varied causes, there appears to result a common clinical picture of non-cardiogenic alveolar oedema and a common histopathological picture of 'diffuse alveolar damage'.

Although subclinical forms of ARDS are likely to be common, many patients require mechanical ventilation. Mortality is high in this group, with around 50 per cent of patients dying, usually as a result of multiple organ failure, nosocomial infection in the injured lung, and episodes of septicaemia. The only treatment available is supportive. Recent advances in the understanding of pathophysiology, particularly in the role of inflammatory processes, leads to the hope that we will soon see novel mechanism-driven therapies that could perhaps be applied in the early stages of disease, before the full complexity of lung injury has evolved.

Pathophysiology

Oedema

Gas-dilution studies have shown that only one-third to one-half of the total lung volume is gas-filled in patients with acute lung injury. The use of computed tomography (**CT**) scanning has revealed that most of the alveolar oedema fluid is distributed to the dependent parts of the lungs. Sometimes changing the position of the patients to the prone position can improve gas exchange. Other studies have demonstrated that oedema distribution is more uniform in patients who are ventilated by high-frequency jet ventilation, suggesting that the mode of ventilation may also have an important influence on how oedema fluid is distributed.

Changes in surfactant

Other pathophysiological consequences may result from disturbance of surfactant production in acute lung injury, which is disordered both in volume and quality, perhaps the result of dysfunction of type II alveolar epithelial cells. Qualitative changes in surfactant may be sensitive markers of early alveolar injury, and in some studies surfactant alterations during the risk period and early stages of ARDS correlate with the severity of lung function changes in full-blown disease. Progressive alterations in the percentage composition of certain phospholipids during the later stages of the natural history of ARDS have been observed, but their functional significance is uncertain. Surfactant function may also be detrimentally influenced by reactive oxygen intermediates and phospholipases that are released locally by neutrophils and other inflammatory cells, and it is also likely that the high protein concentration in the inflammatory exudate markedly impairs surfactant function. These qualitative and quantitative changes in surfactant composition and adverse influences on its function undoubtedly make a major contribution to the evolution of atelactasis, the reduced functional residual capacity, reduced compliance, and increased shunt found in established ARDS.

Other changes in surfactant function may relate to other aspects of lung disease in ARDS. Little is known about alterations in the protein composition and function of surfactant in patients with ARDS. These surfactant proteins (SpA, B, C, and D) have recently been subjected to detailed study (see Chapter 17.1.3) and have been found to possess a number of properties, including bacterial opsonization, and to exert effects on inflammatory cells including macrophages. It is possible that changes in these proteins could have secondary influences, particularly on host defence in the damaged lung, which we know is particularly prone to secondary and often devastating infections with Gram-negative and other bacteria.

Pulmonary hypertension

Pulmonary hypertension is a common complication of the early stages of ARDS, contributes to the generation of alveolar oedema, and is associated with increased mortality. It probably arises as the result of the release of vasoconstrictor mediators, but in the late stages of ARDS it may occur as a result of pulmonary thromboembolism or remodelling of the injured lung. Pulmonary hypertension may contribute to right ventricular dysfunction, although poorly characterized circulating factors also directly depress the contractility of both the right and left ventricles.

At different stages of ARDS both augmented hypoxic pulmonary vasoconstriction and loss of the normal hypoxic vasoconstrictor response are thought to play a role in the development of pulmonary hypertension and increased right to left shunting, respectively.

Decreased cardiac output in patients with ARDS may lead to impaired oxygen delivery to tissues, even in the presence of the normal arterial PO_2. This problem may be compounded by impaired tissue oxygen uptake that is particularly common in patients with sepsis, and which may occur as a result of tissue oedema, microembolization of capillaries, or loss of local microvascular control mechanisms. Finally, to make matters worse, in many of these patients there is an increased tissue oxygen demand as a result of fever, inflammation, and repair processes.

Pathology

The non-specific acute alveolar injury that characterizes ARDS was first described in detail by Liebow and given the term 'diffuse alveolar damage' (**DAD**). Like the clinical situations in which diverse predisposing conditions appear to result in the common clinical picture described by Asbaugh *et al.*, diffuse alveolar damage can be induced by a wide variety of noxious stimuli. In its early stages the alveoli may show atelactasis and the lung microvessels may appear engorged. The alveolar septa are oedematous with inflammatory exudate and extravasated erythrocytes, and a proteinaceous inflammatory exudate may flood the alveoli in some areas of the lung.

The initiation of lung injury is likely to occur within the pulmonary capillaries (see below). Increased numbers of neutrophils may be seen in these and in the interstitial spaces and, if obtained at the earliest stages, bronchoalveolar lavage fluid shows large numbers of neutrophils. Neutrophil numbers in bronchoalveolar lavage and the concentration of neutrophil secreted products therein appear to correlate with ARDS and its severity. These observations draw attention to the likely role of neutrophils early in ARDS pathogenesis. On ultrastructural examination there is clear evidence of endothelial and epithelial injury: this may be extensive. Hyaline membranes, which are the light-microscopical hallmark of diffuse alveolar damage, are likely to be derived from layers of necrotic epithelial cells.

It is important to recognize that these pathological events do not occur in a strictly ordered sequence, and may appear to have reached different stages in different

parts of the lungs at the same time. It is not unusual to find evidence of continued inflammatory injury concurrent with alveolar type II cell proliferation or other evidence of attempts at repair. In those patients who die after just a few days of mechanical ventilation there is often evidence of a pronounced fibroproliferative response, including fibroblast migration into injured alveoli, and fibroblast proliferation and collagen deposition that within 2 weeks can achieve quite remarkable proportions. Nevertheless, in those patients who survive the initiating condition and who are mechanically ventilated, death is not usually due to progressive respiratory failure. Most patients die from the failure of other organs involved in multiple organ failure that are less easy to support than the lung, septicaemia, or secondary nosocomial infections in the injured lung, although barotrauma from mechanical ventilation of poorly compliant lungs can be a major problem in some patients.

Since case reports describing lung biopsies that have been repeated for clinical indications are extremely rare, very little is known about the pathology of the recovery phase of ARDS. Those that have been done suggest that some pulmonary remodelling had occurred. The remarkable examples of patients with severe ARDS, yet who nevertheless appear to regain virtually normal lung function, suggest that some forms of inflammatory lung injury, and perhaps even fibrosis, have the capacity to resolve and/or become significantly remodelled.

Pathogenesis—cellular and humoral mechanisms

Neutrophils and other inflammatory cells

Neutrophils have long been recognized in the lung tissues of necropsy specimens obtained early in the course of ARDS, and bronchoalveolar lavage cytology shows a high percentage of neutrophils and their products, such as myeloperoxidase, which correlate with the development and severity of ARDS. Other potentially injurious neutrophil products including neutrophil elastase and collagenase are also found; and, of importance, peripheral blood levels of neutrophil elastase in patients at risk of ARDS correlate with subsequent ARDS development. External imaging of radiolabelled neutrophils has demonstrated neutrophil accumulation in the lungs of patients with ARDS. Recent studies have suggested that the neutrophil chemokine interleukin-8 (**IL-8**) may be specifically related to the development of this condition. Studies in animal models of acute lung injury, using stimuli of relevance to the pathogenesis of ARDS, critically implicate the neutrophil.

This is not to suggest that other inflammatory cells are unimportant in the pathogenesis of ARDS. Although this condition has been described in neutropenic patients, histology does show the presence of some neutrophils in the lungs, and it is uncertain how much of a neutrophil 'load' may be required: neutrophil replenishment experiments in neutropenic animals replace only a small proportion of the total neutrophil complement. Nevertheless, studies of neutropenic patients raise the possibility that other cells, perhaps monocytes, play an important ancillary role or may even substitute for granulocytes under some circumstances: these cells possess most (if not all) of the potentially injurious mechanisms and capacity of neutrophils.

It is now generally believed that one of the earliest initiation mechanisms in ARDS is damage to the capillary endothelial cells and the airway epithelial cells that form the delicate alveolar gas-exchange membrane. This is caused by toxic products of inflammatory cells that have become sequestered and activated as a result of inflammatory mediators generated as a consequence of the initiating insult (see Fig. 1). It is uncertain why lung injury is so prominent and often the first clinically obvious event in the multisystem microvascular injury of multiple organ failure. However, this may partly be due to the fact that most of the 'marginating pool' of neutrophils resides in lung capillaries, and even in the healthy state neutrophils (average diameter 7.5 µm) have to squeeze through lung capillaries (mean diameter 5.5 µm), thus presenting a massive surface area of contact between this potentially injurious cell and the at-risk gas-exchange membrane (see below). With regard to this interaction, neutrophils cannot injure cells or degrade matrix proteins without extremely close apposition, and it is essential for us to understand the kinetics and adhesion mechanisms that relate to this critical interaction to gain a full appreciation of the pathogenesis of acute lung injury. Finally, neutrophil secretion is not necessarily an all-or-nothing phenomenon: for maximum release of reactive oxygen intermediates or granule enzymes the neutrophil needs to be exposed to agents that 'prime' the cell, together with those that trigger it.

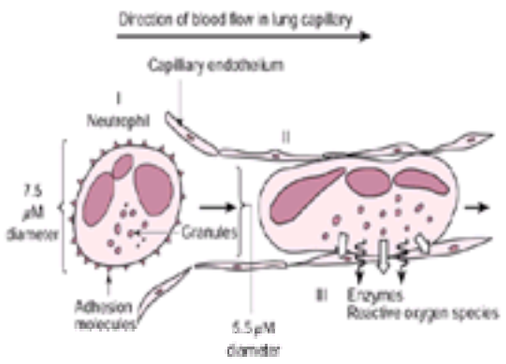

Fig. 1 Initiating insults and toxic products of inflammatory cells that are involved in ARDS. I: Neutrophils became 'primed' by circulating mediators (lipopolysaccharide (**LPS**), platelet-activating factor (**PAF**)) to become 'stiff', 'sticky', and very responsive to secretagogues. II: Primed neutrophils sequester abnormally in the pulmonary capillaries. III: Sequestered primed neutrophils are acted upon by circulating secretagogues (e.g. IL-8) to release injurious reactive oxygen species (**ROS**) and enzymes.

Mediators

A very large number of mediators, indeed several mediator cascades, have been implicated in the pathogenesis of ARDS.

Endotoxin

Endotoxic lipopolysaccharide (**LPS**) is the main 'active' ingredient of the cell walls of *Escherichia coli* and other Gram-negative organisms that cause sepsis syndrome, septic shock, and ARDS. When injected into animals, LPS causes many of the pathophysiological features of septic shock. Low concentrations of LPS may enter the circulation of patients with circulatory shock by the process of 'translocation' through the compromised gut lining. LPS exerts a number of direct influences on neutrophils. In the presence of serum, very low concentrations of LPS (pg/ml) are required to cause enhanced expression of neutrophil surface adhesion molecules that bind the activated endothelium (see Chapter 4.4 and Chapter 5.1) and cause a direct reduction in neutrophil deformability, both of which promote excessive and prolonged neutrophil sequestration in pulmonary microvessels. LPS also causes macrophages to release cytokines, including IL-1 and tumour necrosis factor-a (**TNF-a**), which play key roles in the initiation of inflammation and activate other cascades, including the complement, coagulation, and kinin cascades, as well as generating systemic cytokines including IL-6 that regulate the acute-phase response.

Peptide mediators

The complement peptide C5a is an effective neutrophil chemotaxin and secretagogue *in vitro*, and is found in the blood of patients at risk of ARDS as well as in those with established disease. However, most workers in this field believe that chemokines, especially IL-8, play a more important role in neutrophil chemoattraction to the lung in ARDS. There has been much less study of the role of the contact system, which activates bradykinin, and the clotting and fibrinolytic cascades, but these are also likely to be important in the pathogenesis of ARDS. Activated kinins are vasoactive and cause increased vascular permeability: they can also act as secretagogues for neutrophils and other inflammatory cells.

Cytokines

Cytokines are a diverse group of soluble, hormone-like polypeptides produced by leucocytes and also by some constitutive tissue cells, especially macrophages, endothelial cells, epithelial cells, and fibroblasts. It is likely that cytokines (see Chapter 4.4 and Chapter 5.1) play key roles at all stages of the evolution of ARDS. TNF-a and IL-1 are generated by alveolar macrophages on exposure to LPS and other stimuli of relevance to ARDS pathogenesis, and are likely to play key initiator roles by stimulating other resident cells, particularly epithelial and microvascular endothelial cells, to release IL-8 and other potent neutrophil chemokines. They also act on capillary endothelial cells to induce the expression and activation of adhesion molecules necessary for neutrophil sequestration and for the creation of an

'injury-promoting' microenvironment. Other cytokines—for example, platelet-derived growth factor (**PDGF**), fibroblast growth factor (**FGF**), transforming growth factor-b (**TGF-b**)—are likely to play an important part in the vascular remodelling, fibroblast chemotaxis, and fibroblast proliferation and collagen synthesis that characterize the poorly understood fibroproliferative or 'chronic' phase of ARDS.

Chemokines

This expanding family of small molecular weight peptides has received much recent attention. Depending on the position of cysteine in the molecule they have been subdivided into the 'C–X–C' and 'C–C' chemokines. The C–X–C subgroup contains a variety of peptides that are powerful neutrophil chemoattractants and activators, whereas the C–C group exert their chemotactic influences on monocytes and/or eosinophils.

It is now generally agreed that IL-8 and its family members are responsible for neutrophil attraction to the lung in ARDS. IL-8 is an 8.0 kDa polypeptide that is a potent neutrophil chemoattractant and a powerful stimulus for angiogenesis. IL-8, of a plethora of candidates, was the only mediator to correlate with subsequent ARDS development in studies of patients at the earliest stage of the risk period for ARDS. Other chemokines are likely to play important roles in subsequent monocyte emigration, but these are less well characterized.

Membrane phospholipid derivatives

Membrane-derived phospholipid products, including platelet-activating factor (PAF), leukotrienes, prostaglandins, and prostacyclin, may influence inflammatory cells (PAF is an important neutrophil priming agent, for example), but they also exert major influences on local blood vessels and promote the generation of oedema fluid. Thromboxanes can cause marked pulmonary vasoconstriction and may be partly responsible for the pulmonary hypertension that characterizes the early stages of ARDS.

An injury-promoting microenvironment between abnormally sequestered neutrophils and pulmonary capillary endothelial cells

Neutrophils do not injure endothelial cells *in vitro* without there being direct contact. It is likely that stimulated neutrophils interact with endothelium in a fashion that leads to the formation of a specialized intercellular microenvironment, within which concentrations of histotoxic agents (such as enzymes and reactive oxygen intermediates) would reach high levels, whereas their high molecular weight inhibitors would be relatively excluded (Fig. 2). Furthermore, many potent neutrophil enzymes, such as elastase, are preferentially located on the 'leading surface of the cell' and would need close apposition in order to cause effects. Finally, some of the most potent reactive oxygen intermediates are so labile that they are likely to have a very short distance of activity in tissues. This concept, of a restricted intercellular microenvironment necessary for neutrophil-mediated injury, is supported by experiments showing that matrix degradation occurs only in areas where stimulated neutrophils are tightly adherent and continues in the presence of the large molecular weight antiproteinase, a-1 proteinase inhibitor. The creation of such a microenvironment between neutrophils and endothelial cells in lung capillaries is likely to occur by a combination of adhesion mechanisms and a reduction in the ability of neutrophils to deform.

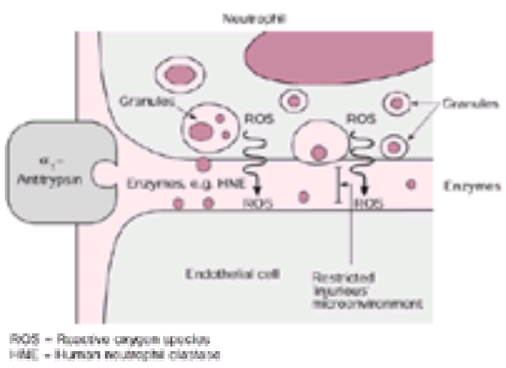

Fig. 2 The neutrophil and the injurious microenvironment in ARDS.

Neutrophil–endothelial surface adhesive molecules

Adhesion between neutrophils and endothelial cells *in vitro* is greatly enhanced within minutes of the addition of inflammatory mediators. Much of this enhanced adhesion can be abolished by monoclonal antibodies directed against the CD11/CD18 group of adhesive leucoproteins on the neutrophil surface. Similarly, endothelial cells that have been activated by LPS or cytokines express adhesion molecules on their surface, and others that are already expressed become activated. *In vivo*, it is likely that neutrophil adhesion to endothelial cells in microvessels occurs by a complex process that involves at least two phases. In the first transient phase of adhesion, which is nevertheless necessary for the second phase, interactions between molecules of the selectin family on neutrophils and endothelial cells are particularly important. In the second phase of 'tight' adhesion, which is necessary for the creation of an injurious microenvironment and also for capillary transmigration of neutrophils, integrin molecules play the central role.

Reduced neutrophil deformability

As described previously, neutrophils are normally required to 'squeeze' through the narrow lung capillaries. Hence, any factors that reduce neutrophil deformability would significantly increase their time of sequestration, and thereby increase the time of contact between activated neutrophils and the delicate gas-exchange membranes. Although it was generally believed that abnormal sequestration occurs mainly by upregulation of neutrophil and endothelial surface adhesive molecules, alteration in neutrophil deformability is also likely to play an important role in the pulmonary circulation. Neutrophils treated with relevant inflammatory mediators demonstrate markedly reduced deformability *in vitro*, and when injected intravenously have prolonged residence time in the pulmonary microcirculation. The molecular mechanisms controlling neutrophil deformability are much less well understood than those governing the expression of surface adhesion molecules.

Neutrophil priming, triggering, and secretion of injurious products

Even when neutrophils are tightly apposed to matrix proteins or 'target' endothelial cells *in vitro*, the induction of neutrophil-mediated injury is not an all-or-nothing phenomenon. When neutrophils are prepared by stringent methods that avoid their exposure to ubiquitous LPS or other agents that might influence their function, stimulation with secretagogues causes little or no release of reactive oxygen intermediates or enzymes unless they are previously exposed to low concentrations of priming agents such as LPS. These priming and triggering phenomena have a number of implications for tissue injury. First, the presence *per se* of neutrophils in tissue does not equate with injury—it is likely that they need to be primed and triggered to achieve a maximal secretory state. Second, when examined in this context, many of the mediators implicated in ARDS pathogenesis exert different effects: for instance, LPS, TNF-a, and PAF are poor secretagogues but highly effective priming agents, whereas C5a, IL-8, and leukotriene-4 (**LTB_4**), together with other neutrophil chemotaxins, are potent secretagogues for primed cells. Therefore, rather than seeking a 'single common mediator' it is perhaps more important to define how certain key mediators act together to influence neutrophil secretion and other critical mechanistic events.

The vast array of potentially injurious neutrophil products represents another example of the remarkable redundancy of the inflammatory response. Most of these products have probably evolved to assist the neutrophil in its rapid passage to the inflamed/infected site, and in its effective killing of bacteria, but in neutrophil-mediated tissue injury and disease processes the difficulty of identifying centrally important toxic agents cannot be exaggerated. Over the years, much circumstantial evidence has accrued to support a role for neutrophil-generated reactive oxygen intermediates in ARDS. However, studies in the early risk period for ARDS suggest that neutrophil elastase is also an important agent.

Lung scarring in ARDS

Whether or not there has been a critical level of epithelial injury in the primary inflammatory damage to the lung seems to be a key factor in determining whether excessive scarring occurs. Most pathologists now believe that the lung can tolerate a certain degree and extent of injury to type I alveolar epithelial cells without the necessity for excessive scarring. In these circumstances it is thought that gaps in the epithelium are repaired by the division of type II epithelial pneumocytes to form a new monolayer of type I cells. However, if there is extensive disruption of the epithelium, and particularly if the basement membrane is severely damaged and loses its architectural integrity, it appears that a scarring response is more likely to result. The inflammatory response can impinge at two levels on the scarring process (Fig. 3). First, on the degree of epithelial injury caused by the inflammatory process; second, by inflammatory cells producing agents that can induce fibroblasts to proliferate and deposit scar-tissue matrix proteins.

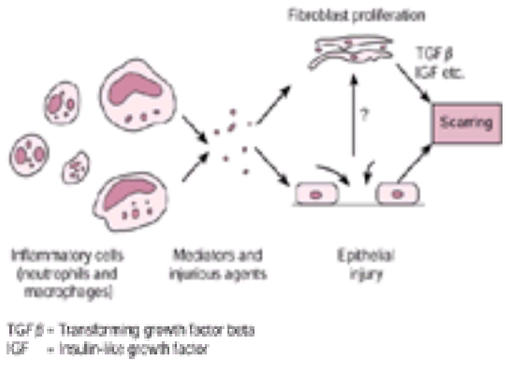

Fig. 3 Inflammation and scarring in ARDS.

Granulation tissue (the precursor of scar tissue) is a very cellular and dynamic tissue, particularly in the lung. It is composed of proliferating fibroblasts that lay down scar-tissue matrix proteins such as collagen. Collagen production by fibroblasts is normally a highly regulated process, with controls being exerted at several levels. All but 30 per cent of the extracellularly secreted collagen is normally degraded, mainly by the effect of fibroblast-derived collagenase. The effect of collagenase is under a further internal control mechanism whereby it is kept in check by inhibitors, for example by tissue inhibitors of metalloproteinases (**TIMPs**). Fibroblast activity is also under the control of external factors including cytokines and growth factors, many of which can be secreted in large quantities by local cells, particularly inflammatory macrophages. Most of these factors, including PDGF, TGF-b, FGF, and insulin-like growth factor (**IGF**), have been shown to exert permissive or stimulatory effects on fibroblast growth and secretion. External factors exerting negative influences must also exist. Although these have received less attention, prostaglandin E_2 represents a good example of a factor with mainly inhibitory effects on fibroblast function.

Further reading

Baggiolini M, Walz A, Kunkel SL (1989). Neutrophil-activating peptide-1/interleukin-8, a novel cytokine that activates neutrophils. *Journal of Clinical Investigation* **84**, 1045–9.

Biondi JW, *et al.* (1990). Mechanical heart–lung interaction in the adult respiratory distress syndrome. *Clinics in Chest Medicine* **11**, 691–714.

Donnelly SC, *et al*. (1993). Interleukin-8 and development of adult respiratory distress syndrome in at-risk patient groups. *Lancet* **341**, 643–7.

Donnelly SC, *et al.* (1994). Role of selectins in development of adult respiratory distress syndrome. *Lancet* **344**, 215–19.

Erzurum S, *et al*. (1992). Cell mechanics of neutrophils: induction of stiffness and actin by lipopolysaccharide. *Journal of Immunology* **149**, 154–62.

Fowler AA, *et al*. (1983). Adult respiratory distress syndrome: risk with common predispositions. *Annals of Internal Medicine* **98**, 593–7.

Guthrie LA, *et al.* (1984). The priming of neutrophils for enhanced release of oxygen metabolites by bacterial lipopolysaccharide: evidence for increased activity of the superoxide-producing enzyme. *Journal of Experimental Medicine* **160**, 1656–71.

Haslett C, *et al.* (1985). Modulation of multiple neutrophil functions by preparative methods or trace concentrations of bacterial lipopolysaccharide. *American Journal of Pathology* **119**, 101–10.

Snow RL, *et al*. (1982). Pulmonary vascular remodelling in adult respiratory distress syndrome. *American Review of Respiratory Disease* **126**, 887–92.

Warshawski F, *et al.* (1986). Abnormal neutrophil–pulmonary interaction in the adult respiratory distress syndrome: qualitative and quantitative assessment of pulmonary-neutrophil kinetics in humans with *in vivo* indium-111 neutrophil scintigraphy. *American Review of Respiratory Disease* **133**, 792–804.

16.5.2The management of respiratory failure

Christopher S. Garrard

Acute respiratory failure: intensive care

Respiration is a complex process involving ventilation, pulmonary gas exchange, oxygen delivery by the circulation, and oxygen utilization by the tissues for the production of cellular high-energy phosphate. By convention, respiratory failure is used in a clinical context to mean failure of ventilation and/or pulmonary gas exchange. Accordingly, the treatment of acute respiratory failure is directed at:

1. establishing and maintaining the airway;
2. administering oxygen;
3. maintaining adequate ventilation, using mechanical ventilation if necessary;
4. enhancing oxygen 'on-loading' by the lungs, that is, improving the efficiency of getting oxygen into the blood for any given concentration of inspired oxygen;
5. identifying and treating the underlying cause (not the subject of this chapter);
6. monitoring SaO_2 (the oxygen saturation of arterial blood—pulse oximetry), ECG, and vital signs; and
7. staged withdrawal of respiratory support (weaning) as the underlying disease process resolves.

Establishing and maintaining the airway

Simple manoeuvres to re-establish and clear the airway must always be followed. These include positioning and maintaining the head and neck in the 'sniff position', inspection of the oropharynx, suctioning, and if necessary, the insertion of an oral or pharyngeal airway.

Endotracheal intubation

If a reliable or adequate airway cannot be established by the above means, endotracheal intubation must be performed. Orotracheal intubation is particularly suited to the emergency situation. Nasotracheal intubation requires a little extra time and should be avoided in those with coagulation defects or thrombocytopenia because of the risk of serious haemorrhage. Whatever technique is selected, intubation should be performed in a safe and expeditious manner by the most experienced clinician available. Neuromuscular relaxant drugs to facilitate intubation should only be used by experienced personnel.

The complications of endotracheal intubation are due to occlusion or displacement of the tube, and airway trauma. The appropriate endotracheal tube size for most adult men is 8 to 9 mm in internal diameter, and for women 7 to 8 mm. For children, a rough calculation using the child's age in years divided by 4, plus 4.0 will provide the tube internal diameter in millimetres.

It is essential that the endotracheal tube be securely anchored and the cuff inflation pressure restricted to less than 30 cmH_2O. High-volume, low-pressure cuffed tubes are generally recommended and cuff inflation pressures should be checked periodically using an aneroid manometer and adjusted accordingly. Using higher cuff pressures does not improve airway protection against aspiration but may damage the tracheal mucosa and risk later subglottic stenosis. Smaller tubes, as used for children, are typically uncuffed.

Difficulties with endotracheal intubation can be encountered in those with a short 'bull' neck or receding lower jaw. Any patient with restricted neck and jaw movements (rheumatoid arthritis or cervical spine injury) or who has abnormal oropharyngeal anatomy (tumour or trauma) should also be regarded as a potential problem. Several options can be considered in these situations. Inhalational anaesthesia by face mask can facilitate intubation, but under no circumstances must muscle relaxants be given unless satisfactory airway access can be ensured. Awake intubation can be performed with topical anaesthesia. Blind nasal intubation or intubation using a fibreoptic bronchoscope or laryngoscope requires considerable skill and training but may be the safest option.

Tracheostomy

Tracheostomy should only replace endotracheal intubation for specific indications and not merely after the elapse of a predefined time interval. Using modern endotracheal tubes and techniques, endotracheal intubation can be tolerated without permanent harm to the airway for months if necessary: the greater part of mucosal damage is done in the first week of intubation with little additional change thereafter.

The common indications for replacement of endotracheal intubation by tracheostomy include the need for chronic or permanent ventilation, to help weaning after previously failed attempts at extubation, to facilitate oral nutrition, or the presence of upper airway complications of endotracheal intubation.

Most tracheostomies for patients in intensive care units are now performed by a percutaneous Seldinger technique: this can be done at the bedside, avoiding the need for moving the patient to the operating theatre. Although percutaneous tracheostomy 'kits' have made the procedure rapid and safe, the clinician needs to be

aware of serious complications such as the formation of false tracks and perforation of structures adjacent to the trachea.

The same principles of cuff pressure management apply to tracheostomy tubes as to endotracheal tubes. Tracheostomy is associated with fewer but more serious complications than endotracheal intubation. These include tube displacement, pneumothorax, severe haemorrhage, and wound infection.

Minitracheostomy

Some patients with an ineffective cough or neurological impairment require continued suctioning of airway secretions without the need for formal endotracheal intubation or tracheostomy. In such cases a 3.5- to 4.0-mm diameter, cuffless minitracheostomy tube can be inserted percutaneously, under local anaesthesia, through the cricothyroid membrane. These tubes cannot be used for conventional ventilation, may result in local haemorrhagic complications, and can be the source of infection.

Cricothyroidotomy

A cricothyroidotomy may be needed in life-threatening, upper airway obstruction where endotracheal intubation is not feasible and there is insufficient time to perform tracheostomy. A full-sized tracheostomy tube (6 to 8 mm internal diameter) can be inserted under local anaesthesia to allow mechanical ventilation.

The administration of oxygen

Hypoxia should never be tolerated through a concern over oxygen toxicity; although this is a recognized complication of prolonged administration of a high concentration of oxygen, the use of 50 or 100 per cent oxygen for less than 24 h is usually considered acceptable. Oxygen can be delivered by a variety of means depending upon the concentration desired and the patient's minute ventilation (Table 1).

Oxygen should be given in such concentrations as to prevent hypoxia with the caveat that controlled (limited) oxygen concentrations should be administered, usually by a 'Venturi type' of face mask, to patients with chronic obstructive lung disease. The response to oxygen therapy can best be measured continuously by pulse oximetry (SaO_2) or by intermittent gas sampling of arterial blood. Once oxygen therapy has been initiated, sudden or abrupt removal of oxygen supplementation runs the risk of severe hypoxia, with the risk of neurological impairment, arrhythmias, or even cardiac arrest.

Mechanical ventilation

Indications for intubation and mechanical ventilation

Failure to intervene promptly can clearly have catastrophic consequences for the patient, but mechanical ventilation is not to be undertaken lightly since it is associated with much morbidity and some mortality.

The indications for mechanical ventilation fall into two broad categories: (i) inadequate alveolar ventilation with increasing PCO_2 and (ii) inadequate gas exchange with increasing $D(A\text{-}a)O_2$ and arterial hypoxaemia. Guidelines for mechanical ventilation in acute respiratory failure are shown in Table 2: the physician should exercise clinical judgement in the interpretation of these and anticipate problems before they arise. For example, one of the simplest criteria for mechanical ventilation is a respiratory rate of 35 breaths/min or more. If, with a respiratory rate of 30 breaths/min a patient is clearly fatiguing, then early elective intubation is clearly preferable to an emergency procedure an hour or so later. Similarly, a progressive fall in vital capacity in a patient with myasthenia gravis receiving full medication may need ventilatory support although the critical value of less than 15 ml/kg is not reached.

The treatment of hypoventilatory respiratory failure consists of assisting ventilatory function, usually by mechanical external means. Figure 1 shows a flow diagram outlining the decision process involved in the assessment of patients who may require mechanical ventilation.

Fig. 1 Respiratory failure algorithm.

Features and applications of a mechanical ventilator

The principles of mechanical ventilation using simple mechanical ventilators need to be understood before moving on to consideration of the complex and sophisticated mechanical ventilators that offer a bewildering range of features. Those not familiar with the field run the risk of being overwhelmed by an overabundance of studies claiming superiority of certain techniques over others. Fortunately, the application of common sense and sound physiological principles will serve better than devotion to attractive technical innovation.

Most adult patients are supported on volume/time-cycled, pressure-limited ventilators (volume ventilator or flow generator). These deliver preset tidal volumes regardless of changes in lung compliance or impedance. The price paid for this desirable characteristic is that the inflation pressure will rise to overcome any mechanical load. A limit must be set to protect the patient from inappropriately high pressure: when this limit is reached the ventilator terminates inspiration regardless of the volume delivered and triggers an alarm.

Neonates and infants can be satisfactorily ventilated using time-cycled, pressure-limited devices (pressure ventilator or pressure generator). The pressure-limited paediatric ventilator offers simplicity and reliable ventilation, although the delivered tidal volume is difficult to measure. In the premature neonate this is not a serious limitation and pressure-limited ventilation is the preferred technique.

Specifically designed, compact, lightweight ventilators, driven by cylinder oxygen and utilizing fluid logic circuits are available for transporting ventilator-dependent patients. These are pressure generators and can be used for both adults and children. By entraining air, a choice of either 60 or 100 per cent oxygen is available.

Modes of ventilation

Depending upon the underlying pathophysiology, the clinician must select the mode of ventilation, choose the ventilation parameters, and set the ventilator alarms. The most commonly available ventilator modes include:

1. control mechanical ventilation (**CMV**);
2. assist control (triggered ventilation, volume cycled);
3. pressure support (triggered ventilation, pressure cycled);

4. intermittent mandatory ventilation (**IMV**, or if synchronized, **SIMV**)—volume or pressure controlled;
5. bi-level positive airway pressure (**BiPAP**); and
6. others.

Control mechanical ventilation (CMV)

This provides time- and volume-cycled, pressure-limited breaths at a preset rate, but does not allow the patient to breathe spontaneously. This mode is suitable for the paralysed or heavily sedated patient.

Assist control

Assist control or triggered ventilation synchronizes the ventilator to the patient's own respiratory rhythm, delivering a volume-preset, pressure-limited tidal volume. A trigger sensitivity is selected, usually –0.5 to –2.0 cmH_2O, by which the patient can initiate volume-preset breaths. As a safety requirement, a high respiratory rate alarm is needed and a 'back-up' ventilation rate must be set in the event of apnoea. Assist control is better tolerated than CMV and the patient requires less sedation, but does have a tendency to hyperventilate.

Pressure support

Pressure support uses a triggering facility to deliver, not a volume-preset breath as in assist control, but a pressure-limited breath (i.e. as with paediatric pressure ventilation). The inspiratory flow rate is usually high so as to minimize phase lag and the work of breathing. Pressure support may be used alone or in conjunction with SIMV when it assists spontaneous breaths. Pressure support provides an efficient maintenance and weaning mode that is well tolerated by the patient. The trigger mechanism is usually a negative pressure threshold, but some ventilators trigger on changes in circuit gas flow that potentially could be more sensitive and reduce the work of breathing.

Intermittent mandatory ventilation (IMV)

This was originally devised for weaning but is now widely adopted as a maintenance mode. It provides the opportunity for the patient to breathe spontaneously and supplement the positive-pressure minute ventilation. In the standard IMV mode there is a theoretical risk of stacking a ventilator breath on top of a spontaneous breath. More modern ventilators utilize the triggering or assist facility to synchronize the IMV breaths with the patient's own spontaneous breathing pattern (synchronized IMV, SIMV). The IMV mode can provide complete or partial ventilation support and, with the patient taking spontaneous breaths, IMV is better tolerated than CMV. Compared with CMV, SIMV results in lower mean airway pressures, has less effect on the cardiovascular system, and allows patients to regulate their own PCO_2 to at least some degree.

Bi-level positive airway pressure (BiPAP)

This is a form of pressure-controlled ventilation based upon raising airway pressure from a lower setting (equivalent to constant positive airway pressure/positive end-expiratory pressure, **CPAP/PEEP**) to a higher setting (equivalent to the peak airway pressure). The lower pressure level should be set at a point above the inflection point in the respiratory pressure–volume curve. A potential advantage of this mode is that it allows spontaneous ventilation to occur at both the lower and higher positive airway pressures, which may be better tolerated by the patient. BiPAP has been successfully adopted as a method of delivering non-invasive positive-pressure ventilation via a nasal or full face mask.

High frequency ventilation

This may have potential advantages in supporting patients with acute respiratory distress syndrome. In randomized studies, peak airway pressures are lower but mean airway pressure and mortality are unchanged. High frequency ventilation can be delivered by several techniques including 'jet' ventilation and high frequency oscillation: both are capable of sustaining adequate oxygenation and carbon dioxide clearance. These techniques have been promoted by enthusiasts for over three decades and yet have failed to establish themselves in routine management of respiratory failure. However, high frequency ventilation may be useful following reconstructive laryngeal, tracheal, or bronchial surgery, or for patients with bronchopleural or bronchocutaneous fistulas, but even in these applications the advantage, if any, over conventional modes of ventilation seems marginal.

Mandatory minute ventilation (MMV)

This is an innovative mode whereby the combined spontaneous and mechanical ventilation must reach a minimum preset level. As the patient's spontaneous ventilation increases the mechanically assisted breaths become fewer. Individual ventilators vary in their ability to achieve successful MMV.

Setting ventilator parameters

Once a ventilation mode has been selected (at least temporarily), ventilatory parameters must be set before attaching the patient to the ventilator. The ventilator parameters include:

1. tidal volume;
2. ventilation rate;
3. inspiratory/expiratory (**I:E**) ratio;
4. flow waveform;
5. inspired oxygen concentration (FiO_2, 0.21 to 1.0);
6. pressure limit (if using volume cycling);
7. trigger threshold (if triggered mode selected)—-pressure trigger (–0.5 to –5 cmH_2O) or flow trigger (3 to 5 l/min); and
8. positive end-expiratory pressure (PEEP) or constant positive airway pressure (CPAP) (0 to 20 cmH_2O).

Tidal volume and respiratory rate

The delivered, inspiratory tidal volume may be set at 10 to 12 ml/kg body weight. This should be reduced if the patient has restrictive lung disease or has undergone lobectomy or pneumonectomy. Using respiratory rates of more than 10 breaths/min with such tidal volumes will provide full ventilatory support. If the patient is breathing spontaneously, an IMV mode will be preferred at rates of between 4 and 8 breaths/min. If assist control or pressure support is chosen, the respiratory rate will be the patient's spontaneous rate.

Respiratory rate, I:E ratio, inspiratory flow rate

These variables are linked and often affect one another. The ratio of inspiratory to expiratory time (I:E ratio) generally ranges from 1:2 to 1:4, allowing sufficient time for full passive exhalation. The higher the set respiratory rate the shorter expiration becomes and the I:E ratio falls. In patients with obstructive lung disease, failing to allow adequate time for exhalation results in air-trapping and hyperinflation, the extra pressure remaining in the alveoli at end-expiration being referred to as auto- or intrinsic-PEEP. This can lead to the paradoxical situation in the patient with chronic obstructive pulmonary disease where the PCO_2 rises as the ventilator rate is increased (usually at rates higher than 20 breaths/min).

The I:E ratio can be adjusted is several ways depending upon the make of ventilator: in some a ratio can be selected directly, whilst in others the inspiratory flow rate determines the duration of inspiration. An acceptable range for inspiratory flow rates is between 30 and 60 litre/min (0.5 to 1.0 litre/s)

Minute ventilation

Some ventilators require the minute ventilation to be set as a primary variable (e.g. Servo 900C). Setting the respiratory rate then effectively determines the tidal volume.

Inspiratory waveforms

Many volume- and time-cycled (flow generator) ventilators allow the choice of several waveforms. Although there is little evidence to favour one over another, a square waveform delivers the tidal volume in the least time and with higher peak pressures. A decelerating flow pattern results in lower peak pressures, longer inspiratory intervals, and lower I:E ratios.

Inspired oxygen concentration (F_iO_2)

This should be constantly adjusted to provide adequate arterial oxygenation without hyperoxia. Too high a F_iO_2 may risk oxygen toxicity and is frequently the cause of failure to wean patients with chronic obstructive pulmonary disease from a mechanical ventilator (by depressing 'hypoxic respiratory drive').

Pressure limit

In volume-cycled modes**,** a pressure limit about 10 cmH_2O above the peak pressure reached during each ventilator cycle protects the patients against inadvertently high pressures experienced during coughing or straining. Hitting the pressure limit terminates inspiration and triggers an audible alarm.

PEEP/CPAP

Maintaining airway pressure above barometric pressure in a spontaneously breathing patient is called constant positive airway pressure (CPAP, Fig. 2). The same pressure applied to a patient on intermittent positive-pressure ventilation is called positive end-expiratory pressure (PEEP). PEEP/CPAP is used to increase lung volume (functional residual capacity) in conditions where this is reduced, for example acute respiratory distress syndrome or cardiogenic pulmonary oedema. It may also be beneficial in patients with flail chest segments by splinting the chest wall. The terms PEEP and CPAP can be used interchangeably provided that the differences regarding spontaneous and assisted ventilation are recognized.

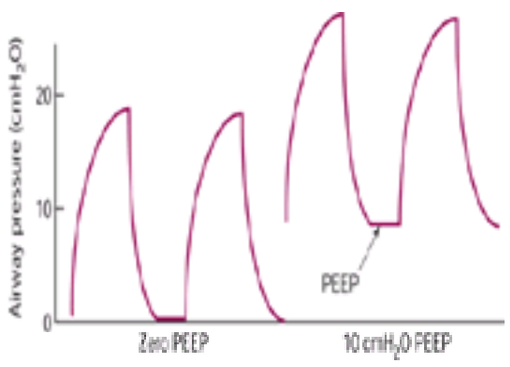

Fig. 2 Schematic of airway pressure measured without PEEP and following the addition of 10 cmH_2O PEEP. By preventing end-expiratory pressure from falling to zero, reinflation and recruitment of alveolar units is encouraged.

PEEP/CPAP is achieved by the inclusion of a resistance at the expiratory end of the breathing circuit. Ideally this resistance should be as close to a threshold resistor as possible, such as an underwater column. In practice most of the valves produce some flow-dependent retardation of expiration that increases the work of breathing in spontaneously breathing patients.

Sighs

Before the advent of high tidal-volume ventilation and PEEP/CPAP, sighs were added to ventilation protocols to prevent progressive atelectasis. Each sigh was delivered 2 to 6 times/h and was equivalent to about twice the conventional tidal volume. The risks of barotrauma probably outweigh the theoretical benefits.

Ventilator-induced lung injury and permissive hypercapnia

Animal and human studies indicate inhomogeneous reinflation of the lung in patients ventilated for acute respiratory distress syndrome. Some parts of the lung are therefore exposed to excessive shear stresses, resulting in a process known as 'volutrauma' that may lead to permanent fibrocystic changes. In the early 1990s evidence from randomized studies suggested that survival from acute respiratory distress syndrome could be improved if inflation pressures were kept below 40 mmHg. As a consequence there was an inevitable degree of hypoventilation and CO_2 retention (P_aCO_2 8 to 12 kPa). This technique of 'permissive hypercapnia' has been widely adopted and may reduce the incidence of ventilator-induced lung injury.

Ventilator monitors and alarms

The ventilation monitors and high/low alarms that must be set and maintained include the following:

- exhaled tidal volume (V_T)
- exhaled minute ventilation
- spontaneous respiratory rate
- airway pressure and circuit disconnect
- peak airway pressure
- inspired oxygen concentration (F_iO_2)
- inhaled gas temperature

The importance of ventilator alarms cannot be overemphasized. The modern, microprocessor-based ventilator is not only more efficient but is significantly safer than its predecessors. An audible alarm alerts the nurse or doctor to an adverse event, while an alarm section of the ventilator control panel indicates the exact event. Some ventilators have progressive alarms that change in character, sounding more urgent, until the alarm is cancelled or the alarm condition rectified.

Clinical monitoring of mechanical ventilation

An essential aspect of monitoring is regular clinical examination of the patient, and inspection of the ventilator and ventilator circuit. Expansion of the chest should be symmetrical with each ventilator-cycled breath (CMV, SIMV), assisted breath (assist control or pressure support), or unassisted spontaneous breath (SIMV). Auscultation should confirm air entry and detect any added sounds. The patient should be sat up or rolled side to side to allow inspection of the whole of the chest. The endotracheal tube should be secure and as comfortable as possible for the patient, and the endotracheal cuff pressure should be adjusted to less than 30 mmHg or so that a small air leak becomes audible with a stethoscope on the side of the neck with each ventilator cycle. The ventilator circuit should feel warm but be free of significant amounts of condensed water. The humidifier temperature and water level should be checked.

The pulse oximeter has contributed significantly to the monitoring and safety of patients on mechanical ventilation. Not only does it provide a continuous measurement

of oxygenation but it also reduces the need for arterial blood gas sampling.

Much can be appreciated from watching the ventilator pressure gauge with each cycle. In addition to evaluating peak inspiratory pressure the clinician will be able to judge whether the patient is 'fighting' the ventilator. Comparing inspiratory and expiratory tidal volumes may indicate a leak in the circuit, either at circuit connections or at the endotracheal tube cuff. When peak pressures are high, the internal compliance of the ventilator and circuit (about 2 to 2.5 ml/cmH_20) may account for much of the volume loss. An assessment of the compliance of the respiratory system can be made from the peak inflation pressures and the resulting exhaled tidal volume during SIMV-delivered breaths. Normal values approximate 50 to 60 ml/cmH_2O, while in severe acute respiratory distress syndrome, effective static respiratory system compliance may fall to 10 ml/cmH_2O.

Methods of enhancing oxygen on-loading

Positive end-expiratory pressure/constant positive airway pressure (PEEP/CPAP)

PEEP/CPAP is the most commonly adopted method of enhancing oxygen on-loading (other than oxygen administration) and is often used in the presence of refractory hypoxaemia due to acute lung injury. Trends in the application of PEEP/CPAP in patients with respiratory failure have changed over the years. The use of maximum tolerated levels of PEEP ('super-PEEP') and 'best' PEEP have generally been replaced by the employment of 'least' or 'enough' PEEP to allow adequate arterial oxygenation with an FiO_2 less than 0.6. However, it is still not uncommon to have to consider levels of PEEP greater than 10 or 15 cmH_2O, particularly in patients with acute respiratory distress syndrome.

An indication of the optimal PEEP level can be determined from the pressure–volume curve during lung inflation. An inflection point at the lower end of the curve suggests that areas of the lung are being allowed to collapse and therefore require higher pressures for their re-expansion. PEEP should be set above the inflection point to prevent collapse/re-expansion stresses.

Care must be exercised to ensure that oxygen delivery is not impaired in the unbridled pursuit of a higher level of arterial oxygenation. Monitoring continuous cardiac output or SvO_2 using a specially developed pulmonary artery catheter is particularly useful in detecting adverse effects of PEEP. If oxygenation is considered inadequate or the FiO_2 is greater than 0.6, increases in PEEP above 10 cmH_2O can be attempted in increments of 2.5 to 5 cmH_2O. If there are adverse effects, such as hypotension or reduced cardiac output (or SvO_2), then intravenous volume loading or circulatory support with an inotrope or pressor agent such as adrenaline may stabilize the patient.

Mask CPAP

CPAP can be applied without resorting to endotracheal intubation in the treatment of selected patients with acute respiratory failure. Close-fitting CPAP masks, which are very similar to standard anaesthetic masks, are widely available together with disposable circuitry and gas supply/pressure regulator mechanisms to ensure the safe delivery of air/oxygen mixtures.

The patient must be fully alert and co-operative since there is a major risk of aspiration should vomiting occur. Mask CPAP is particularly suited to patients with diffuse, reversible, interstitial processes such as cardiogenic pulmonary oedema or interstitial pneumonia (e.g. *Pneumocystis carinii* pneumonia). Recovery should be expected within 1 or 2 days since it is difficult for the patient to tolerate a tight-fitting CPAP mask for much longer periods, and CPAP levels above 10 to 15 cmH_20 should not be employed. In general, patients with established acute respiratory distress syndrome are unsuitable if a long or protracted period of treatment is envisaged.

Suitable patients must be carefully selected and managed in a clinical area where appropriate observation and monitoring can be assured, usually an intensive care, high dependency, or respiratory unit. Continuous assessment by the clinical team is essential so that endotracheal intubation can be substituted for the mask system if necessary.

PEEP/CPAP in unilateral lung disease

The use of PEEP/CPAP in patients with unilateral or irregularly distributed lung disease may not result in improved oxygenation. Indeed, paradoxical falls in oxygenation may occur as a result of the shunting of blood from areas of well matched V/Q to parts of the lung that are poorly ventilated. Unilateral lung consolidation, lung collapse, massive pleural effusion, and pulmonary infarction may therefore not benefit from PEEP/CPAP. Treatment should be directed at correction of the specific lung pathology whenever possible.

Weaning of PEEP

Reduction or weaning of PEEP/CPAP should be conducted carefully and gradually once the underlying pathology (e.g. sepsis) has resolved. Ideally, the FiO_2 should have been reduced to 0.4 or less to maintain a PaO_2 of 10 kPa (80 mmHg) and the PEEP/CPAP level should not have changed for at least 12 h. Even when these criteria are satisfied, a significant proportion of patients will require the return of PEEP/CPAP to previous or even higher levels. Reduction of PEEP/CPAP in 2.5 or 5.0 cmH_2O decrements should be carried out at 10- to 15-min intervals after clinical, pulse oximeter, and blood gas assessment. Even after resolution of lung pathology the retention of low levels of PEEP/CPAP of 3 to 5 cmH_2O ('physiological PEEP') up to the time of extubation may help maintain normal lung volumes and improve gas exchange.

Prone positioning

A feature of the patient with acute respiratory distress syndrome is the extensive gravity-dependent lung collapse that is best visualized by CT scan. To a large degree this is a consequence of nursing the patient for prolonged periods in the supine position. Sitting the patient as upright as possible minimizes the volume of dependent lung. In recognition that dependent regions of the lung exhibit severe atelectasis and loss of volume, it is logical to alternate the patient between prone and supine positions every 8 to 12 h. Up to half of patients with severe acute respiratory distress syndrome show improved oxygenation with such manoeuvres. The major drawback is that vascular access sites and airway access can be dislodged during the process of turning the patient over. However, of all the methods of enhancing oxygen on-loading, only prone positioning actively addresses the issue of dependent lung atelectasis.

Nitric oxide therapy

The addition of low concentrations of nitric oxide (2 to 10 parts per million NO) to the inspired gas mixture will improve oxygenation in about 50 per cent of patients with acute respiratory distress syndrome, but there is no evidence from randomized studies that survival is better. Care must be exercised with delivery systems to ensure correct doses and avoid excessive production of toxic metabolites (nitrogen dioxide and methaemoglobin).

Extracorporeal membrane oxygenation

Extracorporeal membrane oxygenation in premature infants with persistent respiratory distress significantly improves survival. Its use for respiratory failure in adults was largely abandoned over two decades ago following controlled, randomized investigation. With the availability of improved oxygenators there has been renewed interest in combining extracorporeal partial CO_2 removal and low frequency conventional ventilation. Whether this approach offers significant advantages over techniques of permissive hypercapnia remains unproven, although there are enthusiasts who promote the use of extracorporeal membrane oxygenation in adults. Both techniques are aimed at reducing the deleterious effect of positive pressure upon the alveolar epithelium and may therefore hasten recovery from acute respiratory distress syndrome.

Surfactant

Following the dramatic benefit from the use of surfactant in neonatal respiratory distress, attempts to achieve similar effects in adults have been explored. Despite

encouraging anecdotal reports, surfactant has not produced consistent improvement in acute respiratory distress syndrome, particularly when resulting from sepsis syndrome. This failure may stem in part from uncertainty regarding the type of surfactant (synthetic or animal derived) and the dose of surfactant required.

Liquid ventilation

Alveolar instability and collapse in acute respiratory distress syndrome stems, in part, from the high surface tensions at the alveolar air/liquid interface. Von Neerguard demonstrated in the 1920s that a liquid to liquid interface in the lung would remove surface tension effects. Application of this observation can be found in liquid lung ventilation, using a liquid with a high carrying capacity for both oxygen and carbon dioxide instead of air/oxygen mixtures. Perfluorocarbons (carbohydrate molecules with the hydrogen elements replaced by fluorine) are inert liquids that are capable of sustaining gas exchange when instilled into the lung. By partially filling the lungs with perfluorocarbon a conventional ventilator can be used to oxygenate the liquid, which in turn transfers the oxygen to the alveolar membrane (partial liquid ventilation). Most experience has been gained in infants initially sustained by extracorporeal membrane oxygenation. Much greater experience with this technique is required before more widespread use can be justified.

Weaning of mechanical ventilation

More than 80 per cent of patients who are ventilated postoperatively can be weaned simply by clinically evaluating their spontaneous ventilation on a 'T-piece' or similar circuit. The remainder require a progressive reduction in ventilatory support until measurement of ventilation parameters can be made, including the negative inspiratory force and vital capacity. A negative inspiratory force greater than –25 cmH_2O or a vital capacity greater than 10 ml/kg usually indicates sufficient ventilatory reserve for spontaneous ventilation. However, these parameters cannot be applied reliably to patients with severe chronic obstructive pulmonary disease, when blood gases have to be followed with each reduction in ventilation support. Modes of ventilation such as SIMV, IMV, BiPAP, and pressure support are very suitable for weaning since they allow gradual and progressive reduction in support. Regular clinical and physiological assessment after each reduction in ventilation support is essential. Failure to wean a patient successfully from mechanical ventilation should prompt the questions addressed in Table 3.

Complications of mechanical ventilation

Several complications of mechanical ventilation can be attributed to the local effects of the endotracheal tube upon the airway. These include airway obstruction due to tube displacement and pressure necrosis leading to vocal cord injury and subglottic stenosis. The risk of nosocomial pneumonia is increased in the intubated patient.

Many complications are the direct consequence of positive-pressure ventilation. Haemodynamic effects such as reduced cardiac output, reduced renal perfusion, salt and water retention are primarily the result of mechanical, neuroreflex, and humoral factors. The greatest concern relates to the risk of pneumothorax, pneumomediastinum, pneumopericardium, or subcutaneous emphysema (barotrauma). Pneumothorax is the most feared of these because it is associated with rapid deterioration unless dealt with quickly. Signs of tension pneumothorax include arterial desaturation (pulse oximeter), sudden rise in peak airway pressure, asymmetry of chest wall movement, hypotension and tachycardia, and finally circulatory collapse.

Tube thoracostomy is mandatory for pneumothorax in the ventilated patient since progression to a tension pneumothorax is very likely. However, prophylactic thoracostomy tubes are not recommended, even in the presence of pneumomediastinum. Emergency decompression with a 14-gauge cannula is essential in tension pneumothorax, can produce temporary relief of a pneumothorax not under tension, and may have a diagnostic role, but tube thoracostomy should be performed without delay and without radiographic confirmation if necessary. Blunt dissection through the parietal pleura with forceps and digital exploration of the pleural space prior to insertion of the thoracostomy tube is essential if lung damage is to be avoided. Thoracostomy tubes with rigid metal stylets must not be used under any circumstances.

Non-invasive methods of ventilation support

Positive pressure non-invasive ventilation

Non-invasive assisted ventilation may offer an alternative to endotracheal intubation in selected patients with acute respiratory failure. Current evidence suggests that positive pressure non-invasive techniques support the respiratory muscles and avoid upper airway obstruction better than negative pressure techniques. Non-invasive positive pressure ventilation can be applied with volume cycled ventilation, bi-level positive airway pressure (BiPAP), and pressure support modes delivered via face and nasal masks.

Although the principle of the treatment is straightforward, with the application of positive pressure being used to assist ventilation, the practice can sometimes be difficult. Masks must fit tightly or else leakage prevents the generation of adequate positive pressure, and they can be uncomfortable, particularly if not expertly fitted. Indeed, failure to tolerate the mask, sometimes with the development of pressure damage to the face or nose (up to and including necrosis), often prevents effective treatment. If patients are going to respond well to non-invasive ventilation, benefit is apparent within the first 60 min. The maximal duration of successful ventilatory support using non-invasive positive pressure ventilation is usually less than 7 days. A specialist unit with expertly trained nursing care is essential for best results.

Although suitable for those with a variety of causes of respiratory failure, the main use of non-invasive ventilation is in the management of patients with acute exacerbation of chronic obstructive pulmonary disease, and an increasing number of studies testify to its efficacy in this disorder. A prospective randomized study compared non-invasive pressure support ventilation delivered through a face mask with standard treatment in selected patients with acute exacerbation of chronic obstructive pulmonary disease admitted to five intensive care units. This found that non-invasive ventilation reduced the need for endotracheal intubation (11 of 43 (26 per cent) versus 31 of 42 (74 per cent)), the frequency of complications (16 versus 48 per cent), the mean hospital stay (23 days versus 35 days), and the in-hospital mortality rate (9 versus 29 per cent). In a similar population non-invasive ventilation has also been shown to be better than continued invasive pressure support ventilation by an endotracheal tube in a randomized study of weaning from mechanical ventilation, both reducing the duration of mechanical ventilation (from 17 to 10 days) and the time spent in the intensive care unit (15 versus 24 days), and improving survival at 60 days (92 versus 72 per cent).

The use of non-invasive ventilation has also been examined in a general respiratory ward setting. A prospective randomized controlled study in 14 United Kingdom hospitals compared this treatment with standard therapy in patients with acute exacerbation of chronic obstructive pulmonary disease and mild to moderate acidosis. The use of non-invasive ventilation reduced the need for intubation as defined by objective criteria from 27 per cent (32/118) to 15 per cent (18/118) and reduced in-hospital mortality from 20 per cent to 10 per cent.

Negative-pressure ventilation (NPV)

Negative-pressure ventilation is achieved by applying subatmospheric pressures to the surface of the thorax during inspiration and provides an alternative means of assisting ventilation without intubation. Expiration is usually accomplished passively by the elastic recoil pressure of the lungs and chest wall. Negative pressures at predetermined frequencies and depths can be applied to the whole body (iron lung), chest wall, or chest and abdominal wall (cuirass). Some devices can maintain a negative pressure at end-expiration to maintain lung recruitment in a negative-pressure equivalent of PEEP. The haemodynamic effects of NPV are variable and depend on whether negative pressure is applied to the whole body or only to the thorax.

NPV using 'iron lungs' has been used mostly in patients with neuromuscular disorders, such as poliomyelitis or muscular dystrophy, and in those with chronic obstructive pulmonary disease. NPV with negative end-expiratory pressure (NEEP) has been used in patients during hypoxaemic respiratory failure due to acute respiratory distress syndrome and *Pneumocystis carinii* pneumonia.

The negative intratracheal pressures generated during NPV may adduct the vocal cords producing laryngeal obstruction. As a result NPV is not ideal for diseases that may be complicated by upper airway obstruction, such as Guillain–Barré syndrome and myasthenia gravis. For similar reasons, NPV may be ineffective in patients with sleep apnoea. By contrast, the application of nasal CPAP (positive pressure) to such patients is recognized as reliably preventing upper airway obstruction (see Chapter 17.8.1).

Recent developments of the cuirass type of negative-pressure ventilator (Hayek®) permit high frequency oscillations to be used, resulting in enhanced gas exchange.

Specific strategies in ventilator management

Restrictive lung disease

Patients with restrictive lung diseases such as sarcoidosis or fibrosing alveolitis should be ventilated with small tidal volumes of between 5 to 8 ml/kg at rates of 15 to 20 breaths/min. Oxygen need not be restricted in the manner recommended for patients with chronic obstructive pulmonary disease.

Chronic obstructive pulmonary disease

Low rate SIMV (6 to 8 breaths/min) or low pressure levels of pressure support are ideal for patients suffering from chronic obstructive pulmonary disease with acute or chronic respiratory failure. The $PaCO_2$ should be reduced very slowly towards—but not to—normal levels. The FiO_2 rarely needs to be higher than 0.35. High ventilator rates (more than 16/min) are associated with incomplete expiration and air trapping, and the $PaCO_2$ may rise paradoxically if ventilator rates are increased in an attempt to increase minute ventilation. To avoid this, the I:E ratio should be maintained at 1:2 or more.

Weaning can begin as soon as the precipitating cause of respiratory failure has been corrected. Weaning will be unsuccessful if there is any underlying metabolic alkalosis or the patient receives sedative or analgesic agents. The $PaCO_2$ can be allowed to rise slowly to above normal levels provided sufficient time is given for the blood pH to correct and the FiO_2 is kept below 0.35. Carbon dioxide production can be minimized by providing balanced nutrition with calories from both lipid and carbohydrate.

Asthma

Probably less than 1 per cent of acute severe asthma attacks require mechanical ventilation. However, some patients suffer cardiac arrest and die every year because intubation and mechanical ventilation was not performed in time. Hypercarbia alone is generally insufficient as an indication for ventilation, but a combination of a rising $PaCO_2$, fatigue, failure of conservative measures, or arrhythmias does call for elective intubation and mechanical ventilation. Adequate oxygenation must be ensured by the administration of unrestricted and high concentrations of oxygen, in contrast to the patient with CO_2-retaining chronic obstructive pulmonary disease for whom controlled oxygen (24 to 28 per cent) is generally indicated.

Patients with asthma may be difficult to ventilate initially and often require high inflation pressures. Hypoxia may persist despite the use of high concentrations of oxygen and is probably the result of mucus plugging of the airways. A philosophy of 'permissive hypercapnia' or 'controlled hypoventilation' should be adopted with the $PaCO_2$ remaining at elevated levels (7 to 8 kPa, 50 to 60 mmHg). This allows lower tidal volumes and respiratory rates; lower inspiratory flow rates result in lower peak pressures and reduced risk of barotrauma. Deaths in ventilated asthmatic patients are rare but are usually the result of barotrauma, hypotension in volume-depleted patients, arrhythmias, or lung infection.

Maximal bronchodilator therapy including corticosteroids should be continued throughout the period of mechanical ventilation, supplemented if necessary with inhalational anaesthetics such as isoflurane or the intravenous anaesthetic ketamine. Both of these agents are potent bronchodilators. Rehydration and adequate humidification of inspired gases will ordinarily mobilize secretions and mucous plugs; if not, bronchoalveolar lavage may be indicated.

The use of extracorporeal membrane oxygenation and CO_2 removal has been reported in acute asthma. These must be considered exceptional cases and such techniques cannot be generally recommended.

Further reading

Brochard L, *et al.* (1995). Noninvasive ventilation for acute exacerbations of chronic obstructive pulmonary disease. *New England Journal of Medicine* **333**, 817–22.

Cameron PD, Oh TE (1986). Newer modes of mechanical ventilatory support. *Anaesthesia and Intensive Care* **14**, 258–66.

Downs JB *et al.* (1973). IMV: A new approach to weaning patients from mechanical ventilators. *Chest* **64**, 331–5.

Downs JB, Block AJ, Vennum KB (1974). Intermittent mandatory ventilation in the treatment of patients with chronic obstructive pulmonary disease. *Anesthesia and Analgesia* **53**, 437–43.

Dreyfuss D, Saumon G (1998). Ventilator-induced lung injury: lessons from experimental studies. *American Journal of Respiratory and Critical Care Medicine* **157**, 294–323.

Garrard CS (1992). Mechanical ventilation support in severe asthma. *Care of the Critically Ill* **8**, 201–11.

Gattinoni L *et al*. (1980). Treatment of acute respiratory failure with low frequency positive-pressure ventilation and extracorporeal removal of CO_2. *Lancet* **ii**, 292–4.

Hess DR (1999). Noninvasive positive pressure ventilation for acute respiratory failure. *International Anesthesiology Clinics* **37**(3), 85–102.

Hickling KG, Henderson SJ, Jackson R (1990). Low mortality associated with low volume pressure limited ventilation with permissive hypercapnia in severe adult respiratory distress syndrome. *Intensive Care Medicine* **16**, 372–7.

Hill NS (1993). Noninvasive ventilation: does it work, for whom, and how? *American Review of Respiratory Disease* **147**, 1050–5.

Kirby RR *et al*. (1975). High level PEEP in acute respiratory insufficiency. *Chest* **67**, 156–63.

Kumar A *et al*. (1970). Continuous positive-pressure ventilation in acute respiratory failure. *New England Journal of Medicine* **283**, 1430–6.

Nava S, *et al*. (1998). Noninvasive mechanical ventilation in the weaning of patients with respiratory failure due to chronic obstructive pulmonary disease. A randomized, controlled trial. *Annals of Internal Medicine* **128**, 721–8.

Patel RG, Petrini MF (1998). Respiratory muscle performance, pulmonary mechanics, and gas exchange between the BiPAP S/T-D system and the Servo Ventilator 900C with bilevel positive airway pressure ventilation following gradual pressure support weaning. *Chest* **114**, 1390–6.

Plant PK, *et al*. (2000). Early use of non-invasive ventilation for acute exacerbations of chronic obstructive pulmonary disease on general respiratory wards: a multicentre randomised controlled trial. *Lancet* **355**, 1931–5.

Rabatin JT, Gay PC (1999). Noninvasive ventilation. *Mayo Clinic Proceedings* **74**, 817–20.

Slutsky AS (1999). Lung injury caused by mechanical ventilation. *Chest* **116**(Suppl), 9S–15S.

Smith RA, Desautels DA, Kirby RR (1985). Mechanical ventilators. In: Kirby RR, Smith RA, Desautels DA. *Mechanical ventilation*, pp 327–474. Churchill Livingstone, New York.

Suter PM, Fairley HB, Isenberg MD (1975). Optimum end-expiratory pressure in patients with acute pulmonary failure. *New England Journal of Medicine* **292**, 284–9.

Sykes MK (1985). High frequency ventilation. *Thorax* **40**, 161–5.

Tobin MJ (1988). Predicting weaning outcome (Editorial). *Chest* **94**, 227.

Tuxen DV (1989). Detrimental effects of positive end-expiratory pressure during controlled mechanical ventilation of patients with severe airflow obstruction. *American Review of Respiratory Disease* **140**, 5–9.

Wood LH, Prewitt RM (1981). Cardiovascular management in acute hypoxemic respiratory failure. *American Journal of Cardiology* **47**, 963–72.

16.6.1 Sedation and analgesia in the critically ill

G. R. Park and B. Ward

Sedation and analgesia are used to increase patient comfort by minimizing the pain and anxiety produced by illness and its treatment. Factors contributing to patient discomfort are shown in Fig. 1.

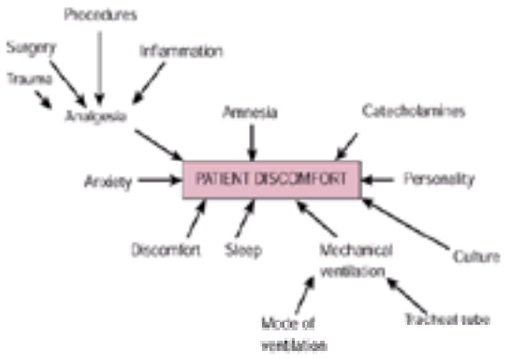

Fig. 1 Factors contributing to patient discomfort.

The relief of pain is an obvious part of being comfortable, but the role of sedation is more complex. The term sedation covers a broad range of conscious states, from almost wide awake to deeply unresponsive. The 'ideal' level of sedation for most patients is at ease, without signs of anxiety or agitation and easily rousable from light sleep. Sedation is needed for a variety of reasons, including:

1. reduction of anxiety caused by fear, inability to communicate, loss of control, or unfamiliar environment;
2. allowing patients to tolerate treatment—e.g. stops them pulling out the tracheal tube;
3. allowing patterns of ventilation to be imposed which do not synchronize with a normal breathing pattern;
4. prevention of awareness when neuromuscular paralysis is used;
5. minimizing distress during uncomfortable procedures;
6. allowing sleep; and
7. control of fits.

Patients will usually tolerate a tracheal tube without the need for paralysis if the ventilator is properly set and they are properly sedated. The indications for neuromuscular relaxation in the critically ill are listed below: the use of muscle relaxants is otherwise avoided.

1. Acute respiratory distress syndrome (ARDS)—paralysis allows the patient to tolerate unusual ventilatory modes, e.g. reverse ratio ventilation;
2. raised intracranial pressure—paralysis prevents coughing and straining; and
3. status asthmaticus—paralysis can reduce risks of barotrauma to lungs.

The intravenous route is used almost exclusively for the administration of analgesia and sedation in the critically ill, as it is faster and more reliable than other routes. Drugs can be given either as repeated bolus doses, or as a continuous infusion. Although a continuous infusion has the advantage of avoiding peaks and troughs associated with bolus doses, there is also an increased risk of inadvertent overdose or accumulation.

The analgesic needs of most patients can best be met with regular bolus doses of analgesic titrated against repeated assessment of the pain. A patient- or nurse-controlled syringe pump driver will deliver a bolus of a predetermined amount of drug when triggered to do so. There is usually a predetermined 'lockout' safety period during which further requests for bolus doses will be ignored. Morphine is the drug most commonly given in this manner, but diamorphine, pethidine, and fentanyl can also be used. A loading dose may be needed before starting.

Hazards of sedation and analgesia

The use of drugs for sedation and analgesia involves risks to the patient. These include:

1. over-sedation or a prolonged sedative effect caused by poor elimination in the critically ill;
2. hypotension/myocardial depression;
3. antitussive effects leading to failure to clear pulmonary secretions;
4. hypoventilation, delaying weaning;
5. toxic effects due to accumulation of sedative/analgesic agents or their metabolites; and
6. expense, both of the drugs and their adverse effects.

There are many reasons why the behaviour of drugs administered to the critically ill patient may be abnormal. These include:

1. hepatic failure leading to poor metabolism or biliary excretion of the drug;
2. renal failure leading to decreased excretion of the drug or its metabolites;
3. haemofiltration/dialysis may have unpredictable effects on clearance of the drug or its metabolites;
4. reduced plasma protein levels (e.g. albumin) may lead to increased free (active) drug levels;
5. volume of distribution may be affected by oedema, ascites, or hyper/hypovolaemia;
6. interactions between drugs; and
7. solvent toxicity.

The risks of using drugs can me minimized by a knowledge of their routes of breakdown and excretion. Agents that are unlikely to accumulate should be chosen when possible. Drugs with more than one site of metabolism, or those which can undergo non-organ-based breakdown are preferred. The risk of accumulation of a sedative drug can be reduced by stopping it every 24 h whenever possible and letting the patient recover from its effects. If the patient wakes or becomes restless, the drug can be restarted knowing that accumulation has not occurred.

To avoid under- or over-sedation, drugs need some assessment of their effects. Because of the many components which are involved in sedation, no simple method

exists. Although work is progressing on physical methods of assessing the level of sedation (e.g. spectral analysis of electroencephalogram waveforms), the most commonly used methods rely on bedside observations. We use a scoring system comprising several different elements (Fig. 2) (see below). The key to avoiding under- or over-sedation is regular assessment of the patient and adjustment of the sedation regimen accordingly.

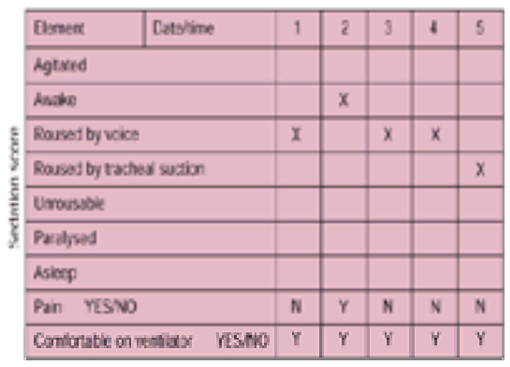

Sedation score	Element	Date/time	1	2	3	4	5
	Agitated						
	Awake			X			
	Roused by voice		X		X	X	
	Roused by tracheal suction						X
	Unrousable						
	Paralysed						
	Asleep						
	Pain YES/NO		N	Y	N	N	N
	Comfortable on ventilator YES/NO		Y	Y	Y	Y	Y

Fig. 2 The Addenbrooke's Sedation Score.

Psychological disturbances

Severe illness, the intensive care environment, and drugs usually prevent patients from sleeping normally. Deprivation of sleep, especially if prolonged, combined with the fear of dying may make some patients psychotic. Close attention to environment (e.g. normal day/night light levels, noise etc.) may help. Drugs may be of some benefit, but can cause prolonged sedation. If the patient has a prolonged recovery phase then depression is common. Antidepressants are rarely of value and can have toxic effects.

Drug treatment

Before using drugs, causes of pain and agitation such as a full bladder or rectum should be excluded.

Sedative drugs

There are two main types of drugs, those principally sedative and those mostly analgesic. The agents most commonly used for sedation are the benzodiazepine midazolam and the anaesthetic agent propofol. These, and other agents commonly used for sedation in the intensive care unit, are described below.

Midazolam

Midazolam is a water-soluble benzodiazepine, which can be given peripherally without causing thrombophlebitis or pain. Like all benzodiazepines it has sedative, amnesic, anxiolytic, and anticonvulsive properties. It has a rapid onset, short half-life (approximately 2 h), and is commonly used in combination with morphine in order to achieve both analgesia and sedation. Midazolam is primarily metabolized by the liver, and accumulation occurs in liver failure. The (phase I) metabolic product, l-hydroxymidazolam has around 10 per cent of the activity of the parent drug. In renal failure, accumulation of l-hydroxymidazolam glucuronide (the phase II metabolic product) can cause prolonged sedation or coma.

Lorazepam

This has been used as an alternative to midazolam. It undergoes metabolism only by glucuronidation to render it water soluble. This makes it less likely for the parent drug to accumulate. It is dissolved in propylene glycol.

Diazepam

Diazepam is rarely used in the critically ill, having been replaced by midazolam. It has a much longer duration of action and has many metabolites with significant activity of their own. This increases the risk of accumulation.

Propofol

Propofol (2,6-di-isopropylphenol) was introduced as an anaesthetic agent but is widely used for sedation in the critically ill as a continuous infusion. Emergence from sedation is rapid and without hangover effect. Propofol is a respiratory depressant, and prolonged apnoea can occur after bolus doses. Hypotension associated with propofol use is common in the critically ill and is dose related. Although metabolized primarily in the liver, extrahepatic breakdown does occur. There are no active metabolites and propofol does not accumulate in hepatic or renal failure to a significant extent. However, because it is formulated in soya bean extract, prolonged infusion (more than 48 h) can lead to hyperlipidaemia. Propofol is expensive, and its use is often limited to those patients who require short-term sedation only.

Dexmedetomidine

Dexmedetomidine is a potent, highly selective, a_2-adrenoceptor agonist. It has sedative, anxiolytic, amnesic, and sympatholytic effects. In addition, dexmedetomidine appears to reduce requirements for opioid analgesia. These effects are mediated centrally at post-synaptic a_2-receptors. In contrast to the agents already discussed, dexmedetomidine does not seem to cause respiratory depression, and exhibits remarkable cardiovascular stability. Because of these features, there is currently great interest in the use of this agent.

Thiopentone

The intravenous anaesthetic agent thiopentone retains certain specialized indications, for example use in status epilepticus or to reduce raised intracerebral pressure. Thiopentone has a half-life of 11 h, and prolonged infusion (i.e. > 24 h) is usually associated with extremely prolonged action.

Combinations of agents

Sedative drugs often act via differing mechanisms and so have slightly different actions. This difference can be used to advantage. For example propofol is mostly an hypnotic, whilst midazolam is a good anxiolytic and amnesic agent as well as producing hypnosis. In combination they are synergistic.

Analgesic drugs

Opioid drugs remain the mainstay of analgesic treatment in the critically ill, and morphine is the most common choice. Some properties of the opioid drugs used in the critically ill are listed in Table 1.

Morphine

Morphine is a cheap and effective analgesic agent and is the opioid against which others are judged. It has both analgesic and sedative effects, although an excessive dose would be required to produce adequate sedation by its use alone. It is often given with a benzodiazepine, such as midazolam, to achieve analgesia

and sedation. It is the standard agent for use in patient- and nurse-controlled syringe pumps. Morphine is metabolized in the liver, forming two major metabolites—morphine 3-glucuronide (**M3G**) and morphine 6-glucuronide (**M6G**), both of which are active. M6G is a potent analgesic, whilst M3G is thought to be antianalgesic.

Pethidine

Pethidine is a synthetic compound and was originally developed as an anticholinergic agent. It does tend to cause anticholinergic effects, such as dry mouth, blurred vision, and tachycardia. It is claimed that pethidine induces less constriction of the biliary sphincter than morphine, and perhaps the only indication for its use is in patients with biliary pathology. It is metabolized in the liver to form norpethidine, pethidinic acid, and pethidine- *N*-oxide. These metabolites are excreted by the kidneys, and in renal failure significant amounts of norpethidine may accumulate, leading to grand mal convulsions.

Fentanyl

Fentanyl is approximately 100 times as potent as morphine, and has a rapid onset of action (3 min). In low doses the analgesic effect of fentanyl ends after about 20 min by its rapid redistribution around the body. With larger doses, tissues may become saturated and drug action is prolonged, termination depending on the slow process of *N*-demethylation in the liver. The major metabolite, norfentanyl, is excreted by the kidneys, and its accumulation may cause toxic delirium in patients with renal failure. Accumulation of fentanyl itself may occur in hepatic failure, causing prolonged effect. Fentanyl has a potent apnoeic effect, and in large doses, fentanyl can produce muscle rigidity, particularly of the chest wall.

Alfentanil

Alfentanil is approximately 10 to 20 times as potent as morphine, and has a very fast onset time (1 min). The effects of alfentanil are short lived (approximately 10 to 15 min), ending by redistribution to tissues. Because of this, alfentanil is unsuitable for use in patient-controlled syringe pumps, and it is administered by continuous infusion. Elimination takes place almost exclusively in the liver, and alfentanil is the current drug of choice in severe renal impairment. It can accumulate in hepatic failure, cirrhosis, or when hepatic enzyme inhibitors such as cimetidine are used.

Remifentanil

Remifentanil is a relatively new agent which may prove to have pharmacological properties useful in critically ill patients. It has a fast onset of action and a very short half-life (10 to 21 min). Remifentanil has an ester linkage within its structure, which is broken down by a non-specific, non-saturable enzyme system present in plasma. This breakdown pathway means that accumulation does not occur, and the drug wears off rapidly even after prolonged infusions and in renal or hepatic failure. Remifentanil must be given by constant infusion, indeed the effects wear off so rapidly that even small delays, such as the time taken to make up a new syringe, can leave the patient without analgesia.

Sedative and analgesic antagonists

When accumulation of a drug or its metabolite is suspected as the cause of prolonged sedation, the diagnosis can be confirmed with the use of antagonists. Naloxone will quickly (but temporarily) reverse the effects of opiates, whilst flumazenil is a benzodiazepine antagonist. Their use is not recommended in patients suffering from head injury. Large doses of either antagonist given quickly can produce sudden arousal, causing agitation. When using naloxone, the sudden reversal of analgesia can cause a massive outpouring of catecholamines and precipitate arrhythmias.

Regional and epidural anaesthesia

For analgesia after certain surgical procedures or trauma, regional and epidural techniques can be extremely effective. Lumbar or thoracic epidurals can prevent hypoventilation and diaphragmatic splinting caused by pain after abdominal or thoracic procedures and fractured ribs, whilst avoiding the side-effects of high-dose opioids. The problem of correct placement of regional blocks in critically ill patients is a considerable one, and complications (such as pneumothorax following intercostal block) must be carefully considered. Epidural analgesia, although desirable, may be contraindicated in the critically ill patient because of coagulopathy or sepsis.

Further reading

Bion JF, Oh TE (1997). Sedation in intensive care. In: Oh TE, ed. *Intensive care manual*, pp 672–8. Butterworth Heineman, Oxford. [An overview of the principles and practice of sedation in intensive care.]

Burns AM, Shelly MP, Park GR (1992). The use of sedative agents in critically ill patients. *Drugs* **43**, 507–15. [A full review of the drugs used to sedate critically ill patients.]

Carrupt PA *et al*. (1991). Morphine 6-glucuronide and morphine 3-glucuronide as molecular chameleons with unexpected lipophilicity. *Journal of Medical Chemistry* **34**, 1272–5. [An important paper that describes how metabolites that should be inactive change their configuration to become active.]

Park GR (1996). Molecular mechanisms of drug metabolism in the critically ill. *British Journal of Anaesthesia* **77**, 32–49. [Describes the problems of drug elimination, solvent toxicity, and makes brief mention of protein binding in the critically ill.]

Park GR, Sladen RN, eds (1995). *Sedation and analgesia in the critically ill*, pp 18–50. Blackwell Science, Oxford. [A multinational book that describes sedation in various diseases, rather than looking at the use of individual drugs.]

Shapiro BA *et al*. (1995). Practice parameters for intravenous analgesia and sedation for adult patients in the intensive care unit: an executive summary. *Critical Care Medicine* **23**, 1596–600. [An American consensus document on how to provide sedation and analgesia in the critically ill.]

Shelly MP, Pomfrett CJD (1999). Assessment of sedation and analgesia and muscle relaxation in the intensive care unit. *Current Opinion in Critical Care* **5**, 269–73. [A paper reviewing clinical as well as experimental methods of assessing sedation and analgesia.]

Tryba M, Kulka PJ (1993). Critical care pharmacotherapy. *Drugs* **45**, 338–52. [Interesting review looking at propofol, isoflurane, clonidine, and sufentanil for sedation. Also reviews H_2- receptor antagonists and sucralfate against gastrointestinal bleeding.]

Venn R *et al*. (1999). Monitoring the depth of sedation. *Clinical Intensive Care* **10**, 81–9. [A review on how to measure sedation and analgesia in the critically ill.]

16.6.2 Management of raised intracranial pressure

David K. Menon

Introduction

The normal intracranial pressure (**ICP**), measured at the level of the foramen of Monro, is between 5 and 15 mmHg in supine subjects. Intracranial hypertension (ICP >20 mmHg) is a common accompaniment of many central nervous system (**CNS**) diseases, when it is often the most important cause of symptoms and modulator of outcome, and—in fatal cases—frequently the immediate cause of death.

Pathophysiology

The cranial cavity contains brain (80 per cent), blood (10 per cent), and cerebrospinal fluid (10 per cent). These incompressible contents are contained in a rigid skull with a fixed capacity, hence an increase in volume of any of these contents, or the presence of any space-occupying pathology, results in an increase in ICP unless one of the other constituents can be displaced or its volume decreased (Fig. 1). This principle is referred to as the Monroe–Kelley doctrine. Increases in intracranial volume may be caused by:

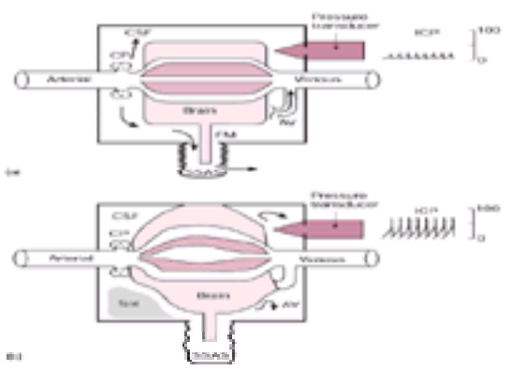

Fig. 1 Schematic diagram showing intracranial contents in the normal brain (a) and with elevated intracranial pressure (b). Note that cerebrospinal fluid (CSF) produced by the choroids plexus (CP), circulates freely, passing through the foramen magnum (FM) into the spinal subarachnoid space (SSAS), before absorption by arachnoid villi (AV) in the cerebral venous sinuses. Increases in ICP may be due to brain oedema, vascular engorgement, space-occupying lesions (SOL), or impaired CSF circulation or absorption. Compensatory mechanisms include translocation of CSF to the SSAS, and compression of cerebral vascular beds. The ICP trace shows a higher mean value, and the inability of the non-compliant brain to cope with increased blood during each systole results in an increased pulsatility of the ICP waveform.

1. *Brain oedema*, which may have different pathogenic mechanisms:
   - cytotoxic oedema occurs as a result of cell swelling, most commonly due to ischaemic energy depletion and rises in intracellular sodium and water;
   - vasogenic oedema results from an increased permeability of the blood–brain barrier with an expansion of the extracellular fluid compartment;
   - interstitial oedema occurs in the context of hydrocephalus, where increased intraventricular cerebrospinal fluid (**CSF**) pressures result in permeation of CSF into adjacent brain, typically in the frontal periventricular regions.
2. *Vascular engorgement*, which results from an increased cerebral blood volume. This may be due to the vasodilatation that accompanies normal or abnormal (for example, epileptiform) neuronal activity. In other situations vasodilatation may be due to the loss of vasoregulation, either due to disease (vasoparalysis), or to the effect of potent physiological (carbon dioxide) or pharmacological (nitrates and other nitric oxide donors) cerebral vasodilators.
3. *Hydrocephalus*, which may be non-communicating (where an obstruction prevents the ventricular system communicating with the subarachnoid space), or communicating (where there is a defect in CSF reabsorption).
4. *Space-occupying lesions* (**SOLs**), which may be chronic (for example, intracranial tumours) or acute (for example, intracranial haematomas associated with trauma).

Temporal patterns of ICP change

Initial increases in intracranial volume are buffered by the displacement or reduction in volume of other contents. Thus, cerebral oedema may result in compression of the ventricles, with translocation of CSF to the spinal subarachnoid space, and compression of cerebral vasculature. Over longer periods, normal brain may be compressed and CSF production diminished. The relationship between intracranial volume (**ICV**) and ICP is commonly depicted as a hyperbolic curve, with an initial flat part during which compensatory mechanisms are effective, moving after their progressive exhaustion to a steep phase when even small increases in intracranial volume produce large increases in ICP. However, the extent and efficiency with which these mechanisms buffer increases in volume depend on the speed of disease progression, and given these considerations it is more appropriate to depict the evolution of pathophysiology as a family of curves, with variable rates of progression (Fig. 2). It is important to make three further points in this context:

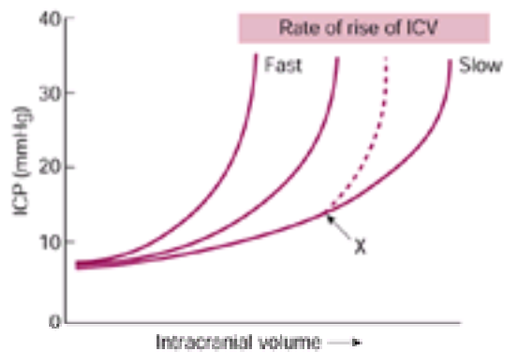

Fig. 2 Intracranial volume/pressure curves. Increases in intracranial volume (ICV) are initially buffered by compensatory mechanisms, but eventually result in elevation of intracranial pressure (ICP). The ability to buffer increases in ICV depends on the speed at which pathology develops. Gradually progressive ICV increases (such as those produced by a slow growing tumour) may be well compensated, until a precipitating factor (e.g. the development of hydrocephalus, denoted by X in the diagram) shifts the relationship to a steeper curve.

1. A precipitating factor may suddenly increase the speed of progression of a relatively slow pathophysiological process, and be the proximate cause of symptomatic decompensation.
2. Acute changes in cerebrovascular physiology are an important cause of such deterioration. Both hypoxia and hypercarbia can cause cerebral vasodilatation and elevate ICP, and whilst *severe* hypertension may result in cerebral oedema, it is far more common to find that relatively minor reductions in mean arterial pressure compromise cerebral perfusion and trigger reflex vasodilatation and a secondary increase in ICP. Such haemodynamic instability may be the underlying cause of phasic increases in ICP (Fig. 3).

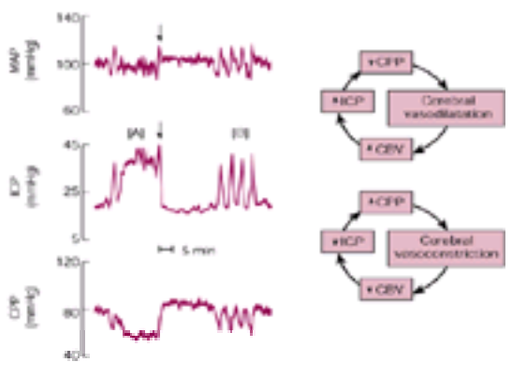

Fig. 3 Intracranial pressure (ICP) traces show phasic variations that may last several minutes (Lundberg A waves (A)) or be more transient (Lundberg B waves (B)). Elevations of ICP are often initiated by reductions in mean arterial pressure (MAP), which reduce cerebral perfusion pressure (CPP) and thereby trigger compensatory vasodilatation and increase cerebral blood volume (CBV) and ICP. This vicious cycle may be terminated by spontaneous hypertension associated with a Cushing response (arrow in MAP and ICP traces), or by therapeutic elevation of MAP, which triggers compensatory cerebral vasoconstriction and reductions in ICP. Note that a period of stable MAP greater than 100 mmHg is associated with a low, stable ICP. (Figure modified with permission from Rosner MJ (1993). Pathophysiology and management of increased intracranial pressure. In: Andrews BT, ed. *Neurosurgical intensive care*, p 75. McGraw-Hill, New York.)

3. Finally, since patients with significant intracranial hypertension operate on the steep part of the ICV/ICP curve, even small decreases in intracranial volume (for example, a 5-ml decrease in cerebral blood volume produced by mild hyperventilation) can have a gratifyingly large effect on ICP.

Why treat intracranial hypertension?

Brain perfusion depends on the difference between mean arterial pressure (**MAP**) and ICP, termed 'cerebral perfusion pressure' (**CPP**). While the normal brain autoregulates cerebral blood flow across a large range of CPP values, the lower limit of such autoregulation is about 50 mmHg in healthy subjects, but may be significantly higher (60–70 mmHg) in disease. CPP reductions below the lower limit of autoregulation result in cerebral ischaemia, and even minor reductions in CPP may trigger reflex vasodilatation and increase ICP in a non-compliant intracranial cavity.

An expanding focal mass can generate pressure gradients within the intracranial cavity. Moreover, the resulting displacement of brain against rigid structures and protrusion (herniation) of brain through narrow openings between intracranial compartments can press on vital structures and result in death (Fig. 4).

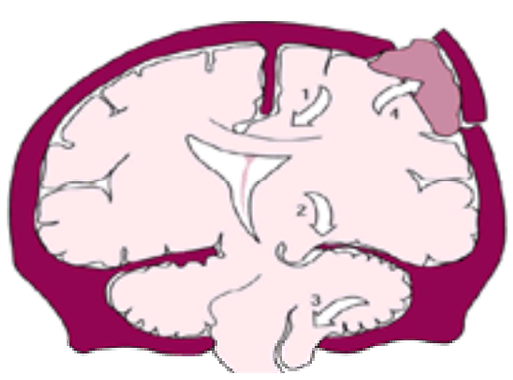

Fig. 4 Cerebral herniation may be: (1) subfalcine (beneath the falx cerebri); (2) transtentorial (through the tentorial hiatus with compression of the midbrain and posterior cerebral artery); (3) tonsillar (where the cerebellar tonsils herniate through the foramen magnum and compress the lower brainstem upper and cervical cord); or (4) transcalvarial (through a traumatic or surgical defect in the roof of the cranial cavity). (Modified with permission from Fishman RA (1975). *New England Journal of Medicine* **293**, 706.)

Prolonged intracranial hypertension may result in permanent damage to critical structures. Thus, benign intracranial hypertension rarely results in herniation syndromes, but frequently causes optic atrophy if left untreated.

Diagnosis

Symptoms

The symptoms that accompany ICP elevation are non-specific and insensitive. The cardinal feature is headache, which may be described as severe ('worst ever') and explosive in its onset in the setting of intracranial haemorrhage. By contrast, the headache of an intracranial tumour is often progressive, worst on awakening (possibly due to ICP elevations associated with the supine position and $PaCO_2$ elevation during sleep), and exacerbated by coughing and straining. However, it may

be indistinguishable from a common tension headache, and dangerous intracranial hypertension may occur without headache. The headache is often accompanied by vomiting, which is classically described as projectile and not preceded by nausea. Visual disturbances are common and may be attributable to optic or ocular motor nerve compression, with accompanying visual failure or diplopia, respectively. There may be alterations in mental function or conscious state, ranging from impaired concentration, through increased irritability, impaired cognition and memory, and altered personality to increased somnolence and deep coma.

Signs

Papilloedema is the classical sign associated with ICP elevation, but is not seen with acute intracranial hypertension and may be absent even where there are large intracranial masses. Pressure on the cranial nerves may result in weakness of ocular movement. The abducens nerve is often involved in such a process due to its long intracranial course; the resultant diplopia provides the classical example of a false localizing sign. Lesions that irritate the meninges of the posterior fossa can produce neck stiffness.

Progressive rises in ICP result in bradycardia and hypertension, which constitute the Cushing response, and signify stimulation of brainstem autonomic nuclei. Worsening brainstem compression and/or ischaemia result progressively in Cheyne–Stokes respiration, central neurogenic hyperventilation, and irregular respiratory patterns ('ataxia of breathing'). Both neurogenic pulmonary oedema and the adult respiratory distress syndrome have been associated with intracranial hypertension.

Severe ICP elevation may result in herniation of the temporal lobe through the tentorial notch (Fig. 4). This produces clinical features due to pressure on the ipsilateral oculomotor nerve (ipsilateral pupillary dilatation), pyramidal tract (contralateral weakness), and brainstem (Cushing response and abnormal respiratory patterns followed by circulatory collapse and respiratory arrest). The posterior cerebral artery is frequently compressed by the herniating temporal lobe, and successful resuscitation from threatened or early transtentorial herniation may leave a patient with an ipsilateral occipital infarction and cortical blindness.

Imaging

Tomographic imaging now provides most diagnostic information in intracranial hypertension. Computed tomographic scanning may reveal subarachnoid or intracerebral blood, contusions, or a tumour. In addition, cerebral oedema may be manifest by a loss of sulci, compression of the third and lateral ventricles, and effacement of the perimesencephalic and suprasellar cisterns. Unilateral lesions may result in a midline shift, compression of the ipsilateral lateral ventricle, and, in some cases, dilatation of the contralateral ventricle due to obstruction of the foramen of Monro. It is important to recognize that overt ventricular dilatation may be absent when hydrocephalus coexists with cerebral oedema. Indeed, the presence of normal-sized ventricles in the context of intracranial hypertension should suggest the possibility of coexisting hydrocephalus and trigger the consideration of CSF drainage as a means of therapy.

Magnetic resonance imaging may provide better definition of underlying pathology, particularly in the posterior fossa, and its multiplanar capability may allow a better appreciation of the extent of space-occupying lesions. Modern imaging methods can also detect patients who may have relatively normal ICP but are at high risk of severe intracranial hypertension, for example a patient with a middle cerebral artery territory infarction is at high risk of severe brain swelling if more than 50 per cent of the middle cerebral artery territory is hypodense.

Lumbar puncture

A lumbar puncture offers the opportunity to directly measure CSF pressure, and can be the defining investigation in meningitis, subarachnoid haemorrhage, or benign intracranial hypertension. However, in the context of clinical features that suggest intracranial hypertension, a lumbar puncture **must be preceded** by CT scanning, and **avoided** if the basal cisterns are effaced by cerebral oedema. Removal of CSF from the lumbar subarachnoid space under these circumstances can markedly increase the pressure differential in the supratentorial compartment, and precipitate transtentorial herniation.

Monitoring intracranial pressure

The clinical evaluation of intracranial hypertension is difficult due to its non-specific clinical picture and phasic variations. Management may therefore be greatly facilitated by direct monitoring of ICP using intraparenchymal or ventricular monitoring devices. Such monitoring is mandatory in patients with severe intracranial hypertension and in those who are sedated or deeply unconscious, in whom changes in clinical signs do not provide an alternative means of assessing progress and response to therapy.

Management

Management focuses on four areas.

Monitoring disease progression and response to therapy

The approach to monitoring will depend on the clinical context. Repeated clinical examination with regular charting of the Glasgow Coma Scale may suffice in many cases. Patients with benign intracranial hypertension may require a regular visual field assessment, whilst those with head injury, intracranial haemorrhage, or severe cerebral oedema may benefit from direct ICP monitoring. The value of ICP monitoring may be substantially enhanced by the use of other monitoring modalities such as jugular bulb oximetry.

Maintenance of stable physiology and removal of precipitating factors

Hyponatraemia and low plasma osmolality will tend to worsen cerebral oedema by favouring water entry into the brain: they should be corrected vigorously. Maintenance of cerebral perfusion pressure with fluid resuscitation and vasoactive agents will prevent cerebral ischaemia. Comatose patients should have their arterial blood gas levels measured, with intubation and ventilatory support provided if airway protection is required or gas exchange is impaired. Whilst hyperventilation has been widely used to control ICP in the past, there is increasing concern regarding the induction of critical cerebral ischaemia by hypocapnic vasoconstriction. Current recommendations suggest that near-normal $PaCO_2$ levels (4.5–5 kPa) should be maintained, with moderate hyperventilation ($PaCO_2$ 4.0–4.5 kPa) guided by jugular bulb oximetry, and reserved for the control of acute episodes of severe intracranial hypertension. Attention should also be paid to treating epilepsy and significant pyrexia, both of which can precipitate rises in ICP, and to discontinuing or reversing the action of drugs such as opiates, which may be responsible for physiological derangements that precipitate ICP elevation.

Treatment of the underlying condition

Early neurosurgical evaluation and operative therapy may be lifesaving if a patient has an acute intracranial haematoma, a large tumour, or established hydrocephalus. Specific antimicrobial therapy may be required for meningitis, encephalitis, or a brain abscess. Systemic arterial hypertension commonly accompanies intracranial hypertension: it should generally not be treated since it may be needed to preserve cerebral perfusion. It is best to avoid nitric oxide donors such as nitrates if therapy is needed for extreme hypertension or for hypertensive encephalopathy: these can cause cerebral vasodilatation and further increase ICP.

Specific treatment of intracranial hypertension

Several therapies can be used to reduce intracranial pressure: their application will depend on the cause and severity of ICP elevation. Commonly used interventions and their indications are outlined in Table 1, but it must be pointed out that few of these have been assessed by good-quality outcome studies.

Treatment pathways for the emergency management of an unconscious patient with suspected intracranial hypertension are outlined in Table 2.

Further reading

Brain Trauma Foundation. The American Association of Neurological Surgeons. The Joint Section on Neurotrauma and Critical Care (2000). Guidelines for the treatment of severe head injury. *Journal of Neurotrauma* **17**(6–7), 449–554. [Series of articles in a special issue.] Also on http://www.braintrauma.org/index.nsf/Pages/Guidelines-main

Kimelberg HK (1995). Current concepts of brain edema: review of laboratory investigations. *Journal of Neurosurgery* **83**, 1051–9.

Maas AIR, *et al*. (1997). EBIC guidelines for management of severe head injury in adults. *Acta Neurochirugica (Wien)* **139**, 286–94.

Menon DK (1999). Cerebral protection in severe brain injury. Physiological determinants of outcome and their optimisation. *British Medical Bulletin* **55**, 226–58.

Menon DK (2000). Cerebral circulation. In: Priebe H-J, Skarvan K, eds. *Cardiovascular physiology*, pp 240–77. BMJ Books, London.

Plum F, Posner JB (1992). *Diagnosis of stupor and coma*, 3rd edn. FA Davis, Philadelphia.

Roberts I, Schierhout G, Alderson P (1998). Absence of evidence for the effectiveness of five interventions routinely used in the intensive care management of severe head injury: a systematic review. *Journal of Neurology, Neurosurgery, Psychiatry* **65**, 729–33.

Rosner MJ (1993). Pathophysiology and management of increased intracranial pressure. In: Andrews BT, ed. *Neurosurgical intensive care*, pp 57–112. McGraw-Hill, New York.

16.6.3 Brainstem death and organ donation

M. J. Lindop

Introduction

The statement by the Conference of Royal Medical Colleges and their Faculties in 1976 led to the establishment of the concept of brainstem death in British practice. Similar procedures took place in many other countries, such as The President's Guidelines in the United States in 1981. Although the motive was to clarify the practice of organ donation for transplantation, the concept has proved useful in determining appropriate care for many patients in intensive care who are certainly not suitable as organ donors.

The concept of brainstem death

Brain death in the United States is defined as the 'irreversible cessation of all functions of the entire brain, including the brainstem... that are clinically ascertainable'. In the United Kingdom the focus has been on brainstem function since it is argued that, in the absence of brainstem function, there will be no activity of the reticular formation and the capacity for consciousness is lost. Deep unconsciousness results from damage bilaterally to a circumscribed area in the tegmentum of the mesencephalon and the rostral pons. In the determination of brainstem death, there is no testing of the function of other areas of the brain, such as electroencephalography of the cerebral cortex. The diagnosis can be made on clinical signs alone, and half of those who fulfil the necessary clinical criteria will have a cardiac arrest despite intensive treatment within 24 h, and this happens to almost all within 72 h.

Managing the patient who is potentially brainstem dead (Fig. 1)

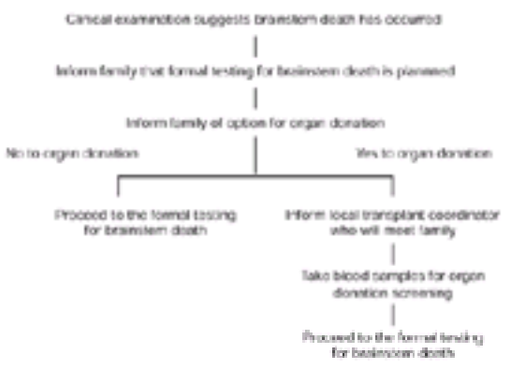

Fig. 1 Management of brainstem death diagnosis.

The patient with severe brain damage will be unconscious and on a ventilator. There will have been testing to chart progress showing that it is likely that the clinical signs of brainstem death will be present. The admitting consultant or the intensive care consultant should see the family to discuss the severity of the brain damage. They are told that there will be formal clinical testing of brainstem reflexes that will determine whether there is any prospect of recovering consciousness. A timetable for this testing is proposed. This interview is an opportunity to discuss the option for organ donation with the family. Organ donation when brainstem death occurs is now well known to the public, and often the family will be the first to raise the matter. In the United Kingdom there is no legal obligation for the doctor to raise the subject, but many families will feel cheated if they have had no chance to offer organ donation: they feel comfort when donation is seen as the only good that can come out of the disaster of an unexpected death. A transplant co-ordinator will be available to meet the family and discuss the process of organ donation. This is the time to check whether the patient fulfils the criteria for acting as an organ donor. The family may decide which organs should be available for donation, but they may not influence or make conditions regarding the choice of recipient.

Diagnosis of brainstem death

Planning the tests (Fig. 2)

Fig. 2 Organization of brainstem death testing after first family interviews.

The diagnosis must be confirmed by two medical practitioners that are competent in neurological examination and have been registered with the General Medical Council for more than 5 years. At least one should be a consultant, and neither can be a member of the transplant team. They can conduct the tests either separately or jointly. A second testing is done at a later time to remove the risk of observer error. The interval between the tests is not fixed, but usually between 1 and 6 h: it should be based on clinical judgement that will reassure all those concerned with the care of the patient that there has been a measured assessment. The legal time of death is the completion of the first set of tests that reveals no brainstem function. The declaration of death and, if organs will not be donated, the stopping of the

ventilator take place after the second testing. Most intensive care units now have a proforma which guides the process of testing.

Performance of the tests

There are three essential components to the clinical testing of brainstem function prior to the declaration of brainstem death—preconditions, exclusions, and clinical criteria.

Preconditions

1. The diagnosis must give an aetiology that confirms that the damage is irreversible.
2. The patient must be in unresponsive coma, though spinal reflexes do not exclude the diagnosis.

Exclusions

Reversible causes of coma must be excluded with certainty.

1. Drug activity, such as narcotics, muscle relaxants, or hypnotics. Due attention must be made to the possibility of prolonged action from previous overdosage, or from metabolism that could be impaired by hepatic or renal failure.
2. Metabolic or endocrine causes of coma—such as hypoglycaemia, hyperglycaemia, hyponatraemia, hepatic failure, uraemia, myxoedema, or Reye's syndrome.
3. Hypothermia—there is no fixed recommendation, but testing should be done at higher than 35°C.

Clinical criteria

These tests show absence of brainstem reflexes, and absence of spontaneous respiration. It is helpful to remember them by relating them to the relevant cranial nerves.

1. No pupillary response to light (II, III).
2. Absent corneal reflexes (V, VII).
3. Absent vestibulo-ocular reflex (VIII)—no nystagmus with installation of 20 ml of cold fluid into the unblocked ears.
4. No motor response within cranial nerve distribution with pain stimulus to face, trunk, or limbs. No limb response to painful pressure over supraorbital notch (V, VII).
5. Absent gag reflex (IX).
6. Absent cough reflex (X).
7. Absence of spontaneous respiration—the apnoea test. At the beginning of reflex testing, the ventilator should be set to deliver 100 per cent oxygen to denitrogenate the lungs (for more than 10 min). A blood gas may be taken. The patient is disconnected from the ventilator and oxygen is insufflated via a catheter into the tracheal tube. Although less than 1 litre/min is absorbed, a flow of 6 litre/min is usually recommended. In apnoea the $PaCO_2$ rises at between only 0.5 and 1 kPa/min. Careful observation of the patient for respiratory movements during the disconnection continues until a blood gas shows that the $PaCO_2$ has risen to more than 7 kPa (just over 50 mmHg). Oxygenation is usually well maintained. An alternative arrangement is to connect the patient to an anaesthetic breathing circuit with a reservoir bag. Respiratory movements may be seen more easily, but small cardiac pulsations can be transmitted and these can be mistaken for breathing activity by inexperienced staff. The patient is usually reconnected to the ventilator once the target $PaCO_2$ is reached if this is the first testing, or if organ donation is planned.

The patient who is brainstem dead and will not become an organ donor

After the completion of the apnoea test in the second testing, the oxygen catheter is removed from the tracheal tube, the patient is not reconnected to the ventilator and death is pronounced. The heart will stop over the next 15 min.

The patient who is brainstem dead and will become an organ donor

Acceptability as an organ donor

Transplant co-ordinators will ensure that the criteria in Table 1 are met.

Clinical management of the organ donor on the intensive care unit

The management of the patient now changes dramatically. Previously all therapy has been directed at maintaining cerebral perfusion to preserve brain function. Now the emphasis changes to care of the potential donor organs. The transplant co-ordinator is informed that the tests have now been completed and have confirmed brainstem death, and the donor operation is arranged as soon as possible.

The common problems of the organ donor are hypotension, cardiac arrhythmias, diabetes insipidus, pulmonary oedema, and disseminated intravascular coagulation.

Cardiovascular problems

Cardiovascular instability is the commonest problem in the organ donor. Brainstem death leads to high catecholamine levels that may increase heart rate, blood pressure, cardiac output, and systemic vascular resistance. Widespread myocardial ischaemic damage can occur associated with defective oxidative metabolism. These changes in autonomic tone, combined with myocardial ischaemia and metabolic and electrolyte instability, all lead to a high incidence of cardiac rhythm disturbance. The ECG may show atrial and ventricular arrhythmias, atrioventricular conduction blocks, and widespread ST-segment and T-wave changes.

An initial hypertensive phase is followed by hypotension in 80 per cent of patients, the common causes of which are shown in Table 2.

Cardiovascular management

Instability is so common that direct arterial and central venous pressure monitoring are essential. A pulmonary artery catheter often proves useful. Targets for management are:

1. mean arterial pressure at least 70 mmHg;
2. central venous pressure about 10 mmHg; and
3. urine output at least 1 ml/kg per hour.

Fluid replacement should be with blood or colloid, though urinary losses should be replaced with electrolyte solutions (nasogastric water or intravenous 5 per cent dextrose in uncontrolled diabetes insipidus). Likely electrolyte imbalance such as hypokalaemia must be sought by regular blood electrolyte measurements. Anti-arrhythmic drugs, inotropic agents, and vasopressors may be required. Inotropes and vasopressors are used as sparingly as possible since the direct action of large doses will threaten perfusion and function of the donor organs. Their use should only be considered when cardiac output studies confirm an indication; for example, when a profound fall in systemic vascular resistance with high cardiac output may justify use of noradrenaline in a minimal dose of 0.02 µg/kg per min. Some protocols use dopamine at 2 µg/kg per min routinely to improve renal, mesenteric, and coronary blood flow. Dopexamine at 1 µg/kg per min may give similar benefit. Use of a hormone replacement regimen may improve cardiovascular stability (see below).

Tight control of fluid balance is important as overload can impair organ function, particularly in lung transplantation.

Endocrine problems

Pituitary damage leads to failure of endocrine homeostasis. Tri-iodothyronine (T_3) and thyroxine (T_4) levels fall. Loss of antidiuretic hormone (ADH) leads to diabetes insipidus in up to 65 per cent of donors: Table 3 shows the characteristics of this condition. Cortisol production may fall, though this does not seem to correlate with cardiovascular changes. Blood sugar control is often defective and an insulin infusion is commonly required.

Endocrine management

The routine use of endocrine supplements is not established. Animal studies show no correlation between hormone deficiencies and cardiovascular instability, but endocrine supplements should be used where there is significant cardiovascular instability with substantial inotrope requirements. A suitable regimen is:

1. tri-iodothryonine as a 4 µg bolus with infusion of 3 mµg/h;
2. desmopressin or vasopressin (see Table 3);
3. insulin at 1 unit/ml, with continuous infusion of 1 to 10 ml/h titrated to keep blood sugar at 6–8 mmol/l; and
4. hydrocortisone supplements at 100 mg every 2 h (there is no risk of toxicity during the short period prior to donation).

Respiratory problems

The organ donor commonly has some pulmonary dysfunction. Possible causes are pneumonia, aspiration of gastric contents, neurogenic pulmonary oedema, and direct contusion.

Respiratory management

Controlled ventilation is used to achieve $PaCO_2$ in the normal range (4.5 to 5.5 kPa) to avoid hypocapnic reduction of peripheral oxygenation and disruption of regional blood flows. Sufficient oxygen is given to achieve a PaO_2 of 11 to 13 kPa. A large tidal volume (12 to 15 ml/kg) delivered at a low ventilatory rate promotes gas exchange and reduces atelectasis. Modest positive end-expired pressure (**PEEP**), about 5 cmH_2O, is useful, but higher levels may impair cardiac output, and hepatic and renal blood flow.

'Neurogenic' pulmonary oedema occurs in about 20 per cent of donors. The causes of this are not known and there is no specific therapy. The usual approach is to monitor the circulation closely and endeavour to optimize left ventricular function. PEEP often helps, but even at high levels (up to 15 or 20 cmH_2O) is not always effective.

Unless there is pulmonary oedema, regular tracheal toilet to prevent accumulation of secretions and atelectasis is important. Strict asepsis must be maintained as the lungs may be implanted into a recipient prone to infection. If PEEP is being used, then tracheal suction should be used sparingly and via a closed system.

Coagulation abnormalities

Disseminated intravascular coagulation can be precipitated in 30 per cent of donors by release of tissue thromboplastin, fibrinolytic substances, and plasminogen activators in severe head injury. Characteristic changes are thrombocytopenia with fall in fibrinogen levels and the appearance of D-dimer fibrin degradation products. Effective haemostasis is needed during the donor operation to reduce blood loss and maintain cardiovascular stability. Fresh frozen plasma and platelets should be given to correct the deficiencies. Antifibrinolytics such as epsilon aminocaproic acid must be avoided in case they provoke microvascular thrombosis in donor organs.

Temperature control

Temperature regulation is impaired. Heat production falls with low metabolic rate and muscle inactivity. Vasodilatation promotes heat loss. Cooling can lead to impaired oxygen delivery to tissues, aggravation of cardiac arrhythmias, increased diuresis, and impaired platelet function. There should be active warming (higher than 35°C) by use of limitation of exposure to the environment, warming blankets, fluid warming, and proper humidification of inspired gases.

Clinical management of the organ donor operation

The operation lasts 3 to 6 h. Four units of blood should be cross-matched to compensate for blood loss. Although anaesthesia is not required, an anaesthetist will be required to supervise cardiovascular monitoring and maintenance of a stable circulation. Reflex hypertension can be a problem and small doses of vasodilating isoflurane or intravenous vasodilators are often required.

Further reading

Chase TN, Moretti L, Prensky AL (1968). Clinical and electroencephalographic manifestations of vascular lesions of the pons. *Neurology* **18**, 357–68. [Consequences of loss of function of brainstem.]

Council of Europe (1997). Standardisation of organ donor screening to prevent transmission of neoplastic diseases. ISBN 92-871-3485-5. [Detailed recommendations where organ donors have neoplastic disease.]

Honorary Secretary of the Conference of Royal Medical Colleges and their Faculties in the United Kingdom (1976). Diagnosis of brain death. *British Medical Journal* **ii**, 1187–8. [The initial formal statement in the United Kingdom.]

Mackersie RC, Bronsther OL, Shackford SR (1991). Organ procurement in patients with fatal head injuries. The fate of the potential donor. *Annals of Surgery* **213**, 143–50. [Survival of patients with head injuries who fulfil brain death criteria.]

Mollaret P *et al*. (1959). Coma dépassé et nécroses nerveuses centrales massives. *Revue Neurologique* **101**, 116–39. [Original description of catastrophic brain damage with loss of autonomic control which led to concept of brain death.]

Morgan G, Morgan V, Smith M (1999). *Donation of organs for transplantation*. Intensive Care Society, Tavistock House, London WC1H 9HR. [A clear comprehensive account of current practice.]

Wheeldon DR *et al.* (1993). Transplantation of unsuitable organs. *Transplantation Proceedings* **25**, 3014–15. [Argument for use of endocrine supplements in unstable patients.]

16.6.4 The patient without hope

M. J. Lindop

The nature of the problem

A long-standing dilemma for doctors has been to judge the appropriateness of further treatment for patients who are already gravely ill. There has commonly been a reluctance to embark on major treatment, such as mechanical ventilation, for fear that it will be more difficult to withdraw this treatment than to avoid its introduction in the first place. Decisions are made on a constantly changing background—the views of society on the ethics of medical management, and the efficacy of new medical techniques are two prime factors.

Who is the patient without hope?

The public finds increasing difficulty with the concept that death is inevitable. There can be a pressure to prolong life for its own sake. Most people would cite as abilities that were important features for an adequate quality of life:

1. an ability to interact with others;
2. an awareness of his or her own existence with a pleasure in the fact of that existence; and
3. an ability to achieve some purposeful or self-directed action, or some self-set goal.

Where it is possible to know the patient's own wishes and values, it may be possible to infer whether he or she would consider life-prolonging treatment to be beneficial.

How certain is the outcome of a medical treatment?

Many treatments in intensive care will prolong life (mechanical ventilation or haemofiltration) but may not have a high likelihood of allowing complete recovery. Scoring systems, such as **APACHE** (Acute Physiology And Chronic Health Evaluation), and later APACHE II and III, have been developed to describe the severity of initial illness and have had some success in predicting hospital mortality. However, they have proved of little value in making decisions on individual patients.

Is the decision to withdraw treatment different from the decision not to institute therapy?

Although it is emotionally easier for the doctor to avoid embarking on treatment than to withdraw it once started, there are no legal or moral differences between these options. The patient will never have the chance of benefit if the treatment is untried. Precedent is not useful in determining the appropriateness of treatment. A treatment is reviewed by looking forward to whether the patient will gain benefit from it. Prolongation of life itself is not necessarily a benefit unless it is associated with the aforementioned qualities. If no benefit can be argued, a treatment should be withdrawn. Knowing that it is possible to withdraw a treatment can give the confidence to embark on that treatment where its outcome is uncertain.

Ways of tackling the problem

How is the decision to withhold or to withdraw treatment made?

The patient

Some patients, such as those with particular types of advanced neurological disease, will be able to participate in decision making. This may be in the form of an advance directive made before the moment of decision about life-prolonging treatment. In this situation the patient's view is paramount and limits the need for further discussion, but it is important that a full account of treatment options and their implications is given and that the patient is judged to understand the issues involved. The challenge for the physician is to embark on these discussions: they must not be avoided, but must take place at a time that has been chosen with the advice of family and nursing staff.

The family

Much time may need to be spent with the family to gain information about the quality of previous lifestyle, and the likelihood that the patient will see life-prolonging treatment as a benefit. They should understand and support any discontinuation of treatment, but should not be asked to make the decision to stop treatment, or be put in a position where they think that they are being asked to do so. This is rarely a problem if full discussions have taken place throughout the course of the illness. In rare instances, families can have complex structures and it will be clear that they are not able to put the interest of the patient first. In this situation further discussions may be required and the help of social workers and religious advisers can be useful in orchestrating dialogue. Neither the next of kin nor those with enduring power of attorney have any legal right to determine treatment. This responsibility remains with the doctor assisted by the health-care team, and, very rarely, by the courts of law.

The medical, nursing, and paramedical team

Much important information can be gained by talking to the patient's own doctor (general practitioner). The nurses caring for a patient over a prolonged period will be able to provide much useful information and should be consulted. Several physicians may have cared for patients with complex problems. In an intensive care unit there will be a team of consultants, and a formal arrangement should be made to consult all these doctors in the process of making the decision. Their opinions should be carefully documented. Their help will be needed to answer the essential questions:

1. Is the diagnosis secure? Are further investigations required before a decision can be made?
2. Is the benefit of further treatment to the patient clear?
3. Is the invasiveness of any treatment justified in the circumstances?

How can treatment be withdrawn?

Once a decision is made a drug or a feeding regimen can simply be stopped. An intermittent therapy such as haemofiltration can be omitted. However, patients who are mechanically ventilated need more careful management. Despite the discussions about terminal weaning of patients in the 1980s, a survey of critical care physicians in 1994 revealed widespread disparity of practice.

Terminal weaning is a protocol that allows death with dignity as mechanical ventilation is discontinued without causing distress to the patient or his family and carers. This is conducted as follows:

1. Stop vasoactive and antibiotic drugs.
2. Stop any paralysing drugs.
3. Continue sedatives and analgesics to avoid distress.
4. Continue physiological monitoring and recording of observations and medical actions.
5. Change the mode of ventilation to synchronized intermittent mandatory ventilation (**SIMV**), which allows the patient to breathe but superimposes a defined number of breaths per minute.
6. Halve the SIMV rate every 30 min until less than 6. Then discontinue SIMV.
7. Use morphine to control dyspnoea and benzodiazepines to control restlessness.
8. If breathing has become stable, allow the patient to breathe spontaneously, and consider lying the patient on the side for extubation.
9. Usually the patient will have died by this stage, but it may be necessary to transfer them to a general ward area for basic care. This can be very disruptive for the family and should be avoided if possible.

Decisions not to escalate treatment, and not to resuscitate a patient

Careful discussions as described above, which should be clearly documented in the medical notes, are a prerequisite of making decisions not to escalate treatment, or not to resuscitate a patient. The decisions must be reviewed each day by the consultant in charge of the patient to ensure their continuing relevance.

Do not escalate

If further complications supervene it can be decided that new treatment is unlikely to give the patient real benefit. A 'do not escalate' order is made so that no therapies will be added. Typically it may be decided not to increase the inotrope dose, not to use haemofiltration, not to transfuse blood or blood products, or not to implement or to increase ventilatory support. In summary, treatment continues only whilst the patient continues to show a beneficial response.

Do not resuscitate

In similar circumstances it may be appropriate to decide that in the event of unexpected circulatory arrest no resuscitation will be attempted. If the documentation is not clear, a resuscitation attempt will be necessary and this can be very distressing to the family at the bedside if it seems inappropriate. Where circulatory arrest has occurred as a result of a drug administration error, a tension pneumothorax, or a complication of therapy, it may be appropriate that a limited (5 min) attempt at resuscitation is made despite the existence of a 'do not resuscitate' order.

Does basic care include provision of fluids and nutrition?

Basic care provides warmth, shelter, hygiene, and comfort to the patient by relieving pain and distress. It includes the regular offer of oral fluid and nutrition. Fluid and nutrition provided 'artificially' by intravenous infusion, or by tube (whether nasogastric or percutaneous endoscopic gastrostomy—PEG), is considered a form of treatment and as such can be withdrawn (although in the specific instance of the persistent vegetative state in England and Wales, review by a court of law is needed).

Summary

Key points in the management of patients who are severely ill and in whom escalation of treatment may not be kind or sensible are:

1. early anticipation of possible outcomes;
2. continuing review of whether treatments remain beneficial;
3. establishment of local guidelines for limiting or withdrawing treatment;
4. good communication skills within the health team;
5. good communication skills with the patient and the family; and
6. clear documentation of decisions.

Case report

An 80-year-old patient with many severe chronic health problems is admitted shocked with acute abdominal pain. A laparotomy is required to establish the diagnosis. The patient may be admitted to an intensive care unit for stabilization prior to surgery.

1. The bowel is extensively infarcted with no hope of survival. The patient is extubated at the end of surgery, and is allowed to die peacefully with appropriate analgesia and sedation, perhaps in the post-anaesthesia care unit or the general ward.
2. There is extensive peritoneal sepsis that could respond to definitive surgery and antibiotic therapy. The patient is admitted to an intensive care unit where monitoring and initial treatment is aggressive. Limits are set beyond which treatment will not escalate. These may be a maximum dose of adrenaline of, say, 0.25 µg/kg per min, no haemofiltration in the event of renal failure, and no resuscitation and no continuing mechanical ventilation after 48 h if there has been no improvement.

Further reading

British Medical Association (1999). *Withholding and withdrawing life-prolonging medical treatment*. BMJ Publishing Group, London. [A guidance for decision making—the outcome of widespread consultation through the British Medical Association's Medical Ethics Committee in 1998.]

Faber-Langendoen K (1994). The clinical management of dying patients receiving mechanical ventilation. *Chest* **106**, 880–8. [A survey of problems of inconsistent intensive care practice persisting over 10 years.]

Grenvik A (1983). Terminal weaning: discontinuance of life-support therapy in the terminally ill patient (Editorial). *Critical Care Medicine* **11**, 394–5. [An influential editorial for intensive care practice.]

Knaus WA *et al.* (1991). The APACHE III prognostic system. *Chest* **100**, 1619–36. [An example of a physiological scoring system.]

17.1.1 The upper respiratory tract

J. R. Stradling

The upper respiratory tract extends from the anterior nares to the larynx. This part of the respiratory tract has to cope with specific problems: it is exposed to incoming air and has to double as the entry to the digestive system. This has led to specific evolutionary adaptations which are not always perfect.

The nose

The anterior nares, which includes the nasal valve just inside the nose, is usually the narrowest part of the respiratory tract and accounts for about 40 to 50 per cent of the total respiratory resistance. In normal subjects the resistance in the lower airways is small (less than 25 per cent) compared with the larynx and nose. This anterior nasal resistance is actively controlled by the levator alae nasi and procerus muscles, which flare the nostrils, and the compressor naris muscle, which narrows the nasal valve further. During mild exercise these muscles (combined with sympathetic nasal mucosal vasoconstriction) can halve the nasal resistance and allow minute ventilations up to 30 litre/min before conversion to oral breathing is necessary. These muscles receive a phasic inspiratory signal, to brace open the nares with each breath, just in advance of diaphragmatic activity.

Occasionally, owing to deformity of the anterior nasal cartilages, the anterior nares are very narrow and limit inspiration, particularly during sleep when the dilator muscle activity is reduced. This is one of the rarer causes of snoring which is amenable to treatment.

The main function of the nose is as first-line defence against problems with the incoming air. In this respect it acts as a coarse particle filter and a conditioner (temperature and humidity) of the air, and with the sense of smell helping to detect noxious substances that are best avoided. The turbinates in the nose present a surface on to which large inhaled particles, such as pollen grains and house dust mite faecal particles, will be retained, with the potential for an allergic response producing allergic rhinitis. Debris arriving on the mucosal surfaces is wafted backwards to be swallowed eventually. Without this so-called 'mucociliary carpet' there is decreased resistance to infections (usually a generalized respiratory problem and not just in the nose) with pooling of mucopurulent material, and recurrent sinus infections. This mucociliary function can be tested by placing a saccharine tablet on the anterior floor of the nasal cavity and timing the period that elapses before it can be tasted in the oral cavity. The normal interval is about 15 to 20 min, but when ciliary defects exist this can extend to an hour or more.

The turbinates fill such a large proportion of the nasal cavity that minor swelling produces large changes in nasal airflow resistance (Fig. 1). There are several rich vascular beds at different depths in the nasal mucosa, providing a large surface area to warm and humidify incoming air. These are supplied by the sphenopalatine branch of the maxillary artery, with venous drainage passing back into the cavernous sinus around the carotid artery. The volume of fluid in these vascular beds is controlled via the vidian nerve (containing sympathetic vasoconstriction and parasympathetic vasodilation) acting on both arterioles and venules. The overall blood flow and total volume of blood in the sinusoids determines the degree of mucosal congestion, which undergoes a cyclical reciprocal change across the two sides of the nose over 2 to 4 h, that is, as the mucosa on one side is congesting, that on the other side is shrinking. This cycle, usually only obvious to individuals with already narrowed nasal passages (when blockage can occur intermittently), can be interrupted by a reflex mediated by pressure on the side of the thorax or in the axilla. Thus, in the decubitus position, the upper nostril becomes clearer and the lower more congested, with the two sides swapping within a minute or two of turning on to the other side. The purpose of this nasal cycle is not known, but using the upper rather than the lower nostril when lying on one's side may lessen the chance of inhaling particulate matter. In addition to this effect, there is a general increase in nasal congestion on lying down due to a hydrostatic rise in capillary pressure.

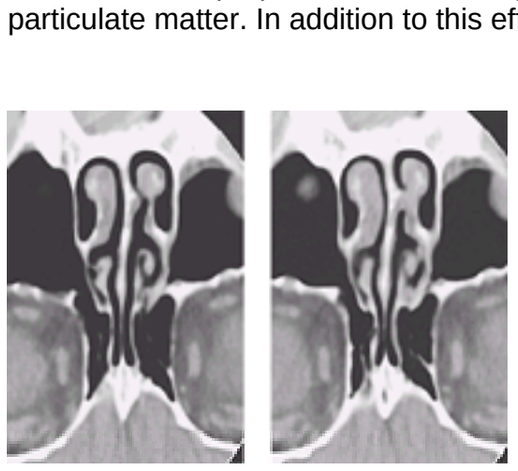

Fig. 1 Coronal sections of human maxillary sinuses and the turbinates in the nose. The view in the panel on the left is taken after ephedrine drops and shows mucosal shrinkage. The consequent small increase in the size of the lumina was attended by a large increase in maximum nasal airflow. (By courtesy of Dr F. Gleeson.)

The volume of fluid needed to humidify the incoming air is considerable, but is reduced by condensation of some of this moisture back on to the cooler nasal mucosa during exhalation. Of course, this conditioning is lost during oral breathing, which has important implications for exercise-induced asthma, which is due to cooling and drying of intrathoracic airways.

Nasal secretions come mainly from submucosal glands that are stimulated by parasympathetic (cholinergic) fibres. There is some evidence that sympathetic activity can also stimulate secretions, but of higher viscosity.

The sensory fibres from the nose travel in the maxillary nerve (mainly the ophthalmic branch) and are the afferent limb of some interesting reflexes. Airflow is sensed and can itself influence breathing pattern. Nerves containing substance P in the epithelium seem to be responsible for sensations leading to sneezing. Sneezing is like coughing in that an explosive expiration is generated in an attempt to expel foreign matter. Coughing involves closure of the larynx until pressure builds up, whereas sneezing involves closure of the pharynx. Unlike coughing, sneezing is never voluntary. Sensory fibres from much of the upper airway, nose, and face are also involved in the diving reflex. This reflex is of great importance to diving mammals when the combination of facial stimulation by cold water, apnoea, and hypoxaemia produce intense peripheral, splanchnic, renal, and muscular vasoconstriction. This diverts blood to the brain and conserves oxygen (producing a heart–lung–brain circulation that prolongs diving time), with the rise in blood pressure limited by a marked vagally induced bradycardia. This vestigial reflex in humans can be utilized in the control of some cardiac arrhythmias, when a brisk increase in vagal tone can be produced by applying ice-cold water to the face.

Nasal irritation can lead to either bronchoconstriction or bronchodilation. The bronchoconstriction can be prevented by atropine and is presumably vagally mediated. This reflex may be important in provoking bronchospasm in some asthmatics. Negative pressure in the nasal cavities can also be sensed, producing a reflex increase in upper airway dilator action (see the following section on the pharynx).

Olfaction depends on recognition of molecules by mucosal receptors at the very top of the nose. These olfactory cells have central axons that pass through multiple tiny holes in the skull (cribriform plate) to the brain. At this point they are very vulnerable to shearing forces during a blow to the head, leading to anosmia (loss of ability to smell).

The pharynx

The pharynx is divided into the nasopharynx, oropharynx, and laryngopharynx or hypopharynx—behind the soft palate, the back of the oral cavity down to the tip of

the epiglottis, and the tip of the epiglottis down to the cricoid cartilage, respectively. Thus the top end is level with the base of the skull and the bottom end is about level with the sixth cervical vertebrae, giving an overall length of about 12 cm. When being used to breathe through, the pharynx has to be a rigid tube (like the trachea), but during swallowing it has to be a collapsed tube capable of peristalsis (like the oesophagus). This combination of functions is achieved by having a muscular tube that can constrict to propel food, but also has external muscles whose function is to brace open the pharynx when required. Figure 2 shows the enormous complexity of the pharyngeal musculature, supplied mainly by the hypoglossal nerve (XII). The pharyngeal constrictors (superior, middle, and lower) are the main peristaltic muscles; the lower part of the inferior constrictor also functions as a sphincter to the top of the oesophagus, preventing air entry during inspiration. Most of the other pharyngeal muscles work in concert to hold open the pharynx. For example, the genioglossus pulls forward the tongue, the geniohyoid together with the strap muscles (sternothyroid, thyrohyoid, etc.) pulls forward the hyoid (enlarging the oropharynx), and the stylopharyngeus probably pulls sideways on the lateral pharyngeal walls. The palatopharyngeus will hold open the pharynx if supported by the levator palati, but will also pull forward the palate to open the nasopharynx. The upper pharyngeal muscles (tensor palati and levator palati) also close off the nasal cavity during swallowing to prevent regurgitation of fluids into the nose. To prevent aspiration, closure of the larynx and the false cords above is co-ordinated with swallowing. Some of these actions require sensory information about the exact location and consistency of any food being swallowed, carried via the glossopharyngeal and vagus nerves (IX and X). Sensory branches of these nerves also supply the ear, which explains why pharyngeal lesions may present with pain in the ear.

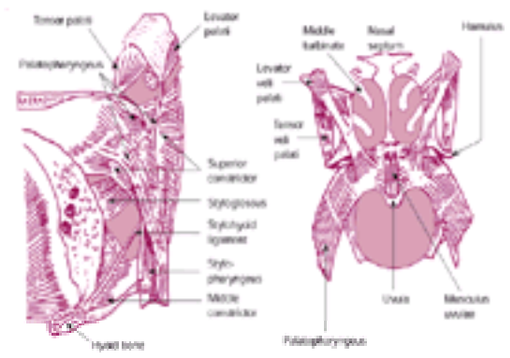

Fig. 2 Two views of the pharyngeal muscles: from inside the pharynx looking laterally, and from high up on the posterior pharyngeal wall looking anteriorly. These muscles act in concert and the physical effect of their contraction depends on which other muscles are simultaneously activated.

Given the complexities of pharyngeal function, it is not surprising that severe swallowing difficulties with aspiration of food and drink are often seen following cerebrovascular accidents in the brainstem involving the control of pharyngeal muscles and the sensory pathways.

Powerful mechanisms are available to maintain patency of the pharyngeal airway during breathing. As with the alae nasi, the pharyngeal dilator muscles receive a respiratory input in time with diaphragm activation. The diaphragm receives a gradually increasing level of phrenic activity to overcome elastic recoil as tidal volume increases, whereas the pharyngeal activation follows more of a 'square wave'. This makes teleogical sense, since the collapsing force is dependent on inspiratory flow and this is roughly constant throughout inspiration. In addition, if pharyngeal patency is threatened, the dilator activity increases. Figure 3 shows the increase in genioglossus tone in response to a fall in intrapharyngeal pressure. This negative pressure will pull in the pharyngeal walls, and there are thought to be 'distortion' receptors of some kind mediating this reflex. Snoring occurs when the pharynx narrows enough to vibrate, and there is some evidence that this vibration itself can also activate pharyngeal dilators, thus warding off full collapse. The factors predisposing to pharyngeal collapse during sleep are discussed in Chapter 17.8.2.

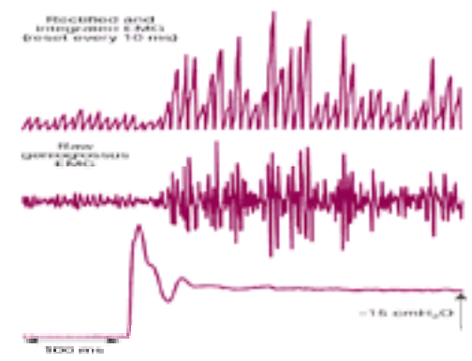

Fig. 3 Response of the genioglossus muscle in a conscious human to a sudden fall in intrapharyngeal pressure. The time delay (about 50 ms) is too short to be due to a cortical response and is presumably a spinal cord reflex. (Reproduced from Horner 1991, with permission.)

Sets of lymphoid tissue (Waldeyer's ring), comprising the adenoids, the palatine tonsils, and the lingual tonsils (back of tongue), are situated in the pharynx. These subepithelial collections of lymphoid tissue are ideally suited to process inhaled and swallowed antigens. Unfortunately, if they hypertrophy too much in response to recurrent infections, they are also positioned such that they obstruct the airway, particularly in small children. This is usually first apparent during sleep, but may become severe enough to provoke inspiratory stridor, even while awake. Adenoidal enlargement, by blocking nasal airflow, will force mouth breathing which, if it occurs early enough (perhaps under 18 months of age), retards development of the lower jaw (the so-called 'adenoidal facies'). This probably leads to overcrowding of the teeth and a narrower retroglossal space (this is further discussed in Chapter 17.8.2).

The larynx

The larynx (Fig. 4) has three important functions: communication, protection of the airway, and dynamic control of lung volume.

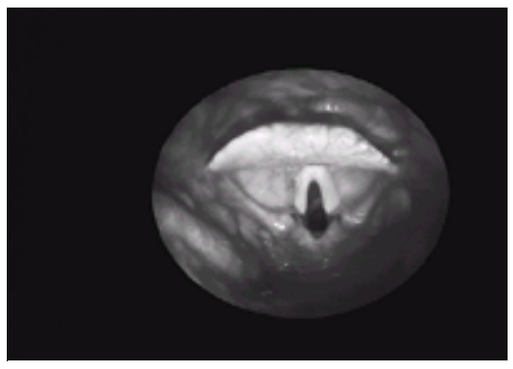

Fig. 4 Bronchoscopic view of the larynx from above. The top of the picture is the anterior. (By courtesy of Dr P. Stradling.)

A minority of the intrinsic and extrinsic muscles of the larynx (e.g. cricothyroid, posterior cricoarytenoid) open (abduct) or brace the vocal cords, whereas the majority (e.g. thyroarytenoid, transverse and oblique arytenoids) close (adduct) the cords. The recurrent laryngeal nerve (from the vagus) supplies all the muscles apart from

the cricothyroid (supplied from the superior laryngeal nerve, which is also a branch of the vagus). The left recurrent laryngeal nerve comes off the vagus and passes under the aortic arch before running up close to the thyroid gland to the larynx. This means that it can be damaged by a tumour at the left hilum and surgically during a thyroidectomy. The right recurrent laryngeal nerve passes under the right subclavian artery where it can be damaged by a right-sided apical lung tumour.

Complete paralysis of the recurrent laryngeal nerve gives permanent hoarseness of the voice, and the affected cord assumes a position midway between full abduction and adduction. The cord is floppy and can be moved passively very easily, being 'sucked' towards the mid-line during inspiration and blown open during expiration. The unparalysed cord may eventually compensate to some degree and move nearer the paralysed cord, improving the voice. If paralysis of the recurrent laryngeal nerve is incomplete, the affected cord may take up the adducted position, presumably because fibres running to the abductors are amaged first. When there is bilateral damage to the recurrent laryngeal nerves, loss of adequate abduction causes inspiratory stridor as the cords are passively drawn together.

As mentioned earlier, there are reflexes initiated by supralaryngeal sensory fibres (mainly via the internal branch of the superior laryngeal nerve) designed to protect the airway. Fluid or food landing on or near the vocal cords will provoke coughing and/or laryngeal closure. During sleep, irritation of the cords tends to produce apnoea and laryngeal adduction, and coughing occurs only when wakefulness supervenes.

One of the less well-known functions of the larynx is to brake expiratory flow and thereby control lung volume. In some species, and in neonates, laryngeal expiratory braking is very important, acting rather like positive end-expiratory pressure to maintain end-expiratory lung volume above the passive functional residual capacity, thus preventing atelectasis. In adults there is no good evidence that the rate of expiration is under active laryngeal control, but this mechanism may come into action during respiratory illnesses (such as pneumonia), especially if there is marked hypoxaemia. If the upper airway is bypassed, for instance by tracheostomy or intubation, then other mechanisms come into play to maintain end-expiratory lung volume, such as postinspiratory contraction of the diaphragm (thus delaying expiration) and shortening of expiratory time (thus starting inspiration again before lung volume has fallen too far). Figure 5 is from a tracheotomized dog with areas of atelectasis. This shows how once laryngeal braking is denied to the animal, expiration proceeds faster, lung volume falls, and expiratory time is shortened to produce tachypnoea. This reflex was not present when the areas of atelectasis had resolved. The clinical correlate of this is sometimes seen as an expiratory grunt in babies who have a respiratory illness. Intubation may worsen gas exchange in this situation unless positive end-expiratory pressure is also applied.

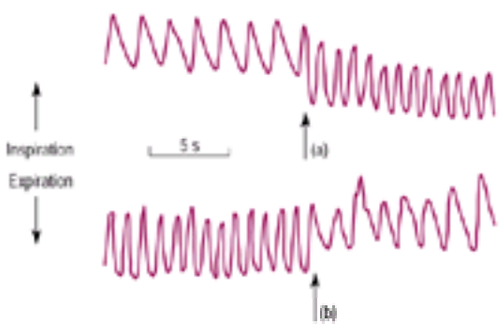

Fig. 5 Recorder tracings in a dog with atelectasis showing the effect of switching from upper airway to tracheostomy breathing (arrow at (a)) and from tracheostomy to upper airway breathing (arrow at (b)). The signal is from an inductive plethysmograph measuring movement of both the rib cage and abdomen which represents lung expansion and contraction.

Further reading

Brouillette RT, Thach BT (1979). A neuromuscular mechanism maintaining extrathoracic airway patency. *Journal of Applied Physiology* **46**, 722–9.

Gautier H (1973). Control of the duration of expiration. *Respiration Physiology* **18**, 205–21.

Horner RL (1991). Evidence for reflex upper airway dilator muscle activation by sudden negative airway pressure in man. *Journal of Physiology* **436**, 15–29.

Matthew OP, Sant 'Ambrogio GS (1988). Respiratory function of the upper airway. In: *Lung biology in health and disease*, Vol. 35. Marcel Dekker, New York.

Remmers JE, Bartlett D (1977). Reflex control of expiratory airflow and duration. *Journal of Applied Physiology* **42**, 80–7.

17.1.2 Structure and function of the airways and alveoli

Peter D. Wagner

The organ of gas exchange

The lung is the organ of gas exchange, providing the means of transferring oxygen (O_2) from the air to the blood for subsequent distribution to the tissues. At the same time it enables removal of metabolically produced carbon dioxide (CO_2) from the blood, which is then exhaled to the atmosphere. Not just in health, but also in lung disease, the volumes of O_2 taken up and CO_2 removed by the lung per minute must equal the rate of O_2 consumption and CO_2 production by the aggregate tissues of the body.

The lung will also exchange any other gas that is presented to it, but the principles involved—passive diffusion—mirror those for O_2 and CO_2. Quantitative but not qualitative differences occur in how such gases (e.g. anaesthetic agents, carbon monoxide, toxic gases inhaled by accident) are handled by the lung. These differences stem from the means by which any particular gas is transported in the blood; whether in simple physical solution alone, or also in some chemical combination with molecules such as haemoglobin.

The principles are similar for gas uptake into blood and elimination from the blood. In fact, because gas exchange occurs by passive diffusion, whether a gas is taken up from the air into the blood or eliminated from the blood into the air depends simply on the partial pressures of the gas on each side of the blood–gas barrier, the 0.3-µm thick tissue layer separating alveolar gas from pulmonary capillary blood.

For transfer of a gas from the environment to the blood to occur, the gas in question must first be brought to the alveolar blood–gas barrier by the process of ventilation. Diffusion across this barrier then occurs at a rate proportional to: (i) the alveolar surface area available, (ii) the partial pressure difference between alveoli and blood, and inversely proportional to the thickness of the barrier, in concordance with the rules of simple passive diffusion. The gas molecules, now present dissolved physically in plasma, also distribute into the red cells. Depending on the gas, chemical associations may occur—with haemoglobin in the case of O_2, CO_2, carbon monoxide (CO) and nitric oxide (NO), and through transformation to bicarbonate ion for CO_2. The last element of the exchange process now occurs—the transport of the gas in blood pumped by the heart through the systemic circulation to the tissues of the body.

This chapter focuses on the first two of these three steps in gas exchange—ventilation and diffusion. A separate chapter deals with the third step—the pulmonary circulation (see Chapter 15.15.1). The structural basis of ventilation and diffusion, and the associated functional consequences, will be presented with particular emphasis on implications for disease. Lung diseases of many types commonly affect each of the steps involved in gas exchange, and the clinical consequences can usually be readily understood if the structure–function relationships are known.

Basic airway and alveolar design

In essence, the lung is a balloon undergoing cyclical inflation and deflation (ventilation, or tidal breathing) around some partially inflated state; the main anatomical elements are shown in Fig. 1. The gas-filled interior of the balloon corresponds to the alveolar gas spaces of the lung. The thin wall of the balloon may be likened to the blood–gas barrier, with the pulmonary capillary network imagined as covering the balloon's surface, separated from the interior gas by the elastic material making up the balloon's wall. The balloon is inflated through its neck (trachea) with each inspiration, thus bringing fresh air (21 per cent O_2, no CO_2) to the balloon's interior. This fresh gas is rapidly mixed with resident gas already present. This resident gas is partially depleted of O_2 by ongoing diffusion of O_2 into the capillaries, whilst at the same time, CO_2 is evolved into the gas from the capillary blood. Each inflation, by bringing fresh air into the alveoli, slightly increases alveolar P_{O_2} and decreases alveolar P_{CO_2}. Each deflation moves some of this alveolar gas back to the environment. This rids the lung of some CO_2, but also removes some O_2, albeit at lower concentrations than in room air. In normal quiet breathing, alveolar O_2 concentration averages about 16 per cent over a respiratory cycle, whilst that of CO_2 is about 5 per cent, and a long-term steady-state of gas exchange is achieved.

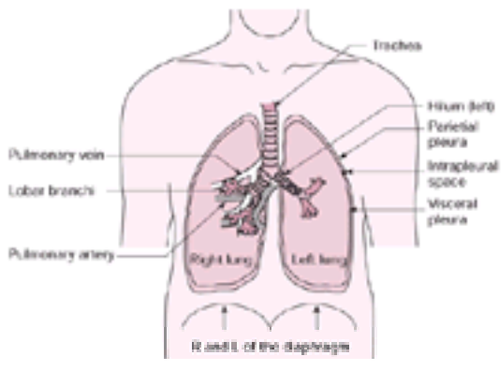

Fig. 1 The right and left lungs are separately encased within the thorax, and each is covered by a visceral pleural membrane. This is continuous with the parietal pleural membrane which lines the interior thoracic cavity, and the thin fluid-filled space between the visceral and parietal pleuras constitutes the intrapleural space. The hila of the two lungs contain the main stem bronchi, and accompanying pulmonary arteries and veins. The main stem bronchi join at the carina to form the trachea. The pulmonary arteries emanate from the right ventricle; the pulmonary veins empty into the left atrium. Within the lungs, the airways and blood vessels continue branching for approximately 20 generations. The major muscle of inspiration, the diaphragm, consists of two domes upon which the right and left lungs sit, and which separate the thoracic and abdominal contents.

Because the process of gas exchange depends on simple, passive diffusion, a large area of contact between alveolar gas and capillary blood is required to ensure sufficient gas flux across the blood–gas barrier to meet metabolic demand. The balloon analogy, while useful as an initial concept, thus exhibits a major difference from how the real lung is configured. The real lung has its total gas volume constituted not as a single balloon-like gas chamber, but as a very large number (about 300 million) of very small (radius, r, about 150 µm) almost spherical balloons or alveoli. Since the volume (V) of a sphere is $V = (4/3) \times p \times r^3$, while its surface area (A) is $A = 4 \times p \times r^2$, dividing a lung of a given volume (given because the lung must fit within the thoracic cage no matter whether as one large alveolus or 300 million small alveoli) into many small alveoli allows a much larger total surface area than if the lung were indeed a single large chamber, as seen from dividing the expression

for surface area by that for volume: $A/V = 3/r$. Since total volume V is fixed by the thoracic cage, total surface area A increases as the number of alveoli is increased because radius of each alveolus must be reduced if volume is to remain constant. A typical value for V is 4000 ml. A single sphere of this volume would thus have a radius of about 10 cm, and a surface area of about 1200 cm^2, only slightly more than 1 square foot. By contrast, 300 million alveoli, each with a radius of 150 µm, would by the above formulae have the same total volume, thereby fitting equally as well inside the chest, but have a total surface area of about 800 000 cm^2. This approximates the area of a tennis court, and is about 650 times larger than the 1200 cm^2 area were the lung a single chamber. Since the laws of diffusion state that diffusive gas transport is proportional to surface area, and since the maximal rate of O_2 uptake in a normal man is about 4000 ml/min in heavy exercise, maximal pulmonary O_2 exchange would be insufficient for life were the lung a single chamber, at 4000/650 or 6 ml/min. Normal resting O_2 utilization by the tissues is about 300 ml/min. These calculations serve to highlight the critical importance of dividing the lung into a large number of small alveoli, and set the stage for discussing the structural configuration of a lung that must find a way to ventilate simultaneously 300 million separate, miniscule units of gas exchange.

The lungs inside the thoracic cavity

As with a balloon, the lung cannot inflate itself (although, as an elastic structure, once inflated it is capable of unassisted deflation just like a balloon). Inflation requires creation of a pressure difference between the outside and inside of the lung, pressure being higher inside. This may be accomplished in one of only two ways. One is by positive pressure inflation, typical of most clinical ventilators that are connected to the trachea and produce inflation by mechanically increasing intratracheal airway pressure. Spontaneous breathing throughout normal life does not happen in this way, and so the only possibility of normally achieving lung inflation is by the second option—that of decreasing the pressure around the lungs below that of the surrounding air. This is accomplished by encasing the lungs within the closed thoracic cavity, and having the muscles in the wall of this cavity (the intercostal muscles and the diaphragm) contract when inflation is desired. Contraction of these muscles moves the diaphragm caudally and expands the rib cage in both anteroposterior and lateral dimensions. As a result, the pressure inside the thoracic cavity but external to the lungs (that is, within the intrapleural space) is reduced to below that of the air. Since the alveolar tissue is extremely thin and easily deformable, the pressure within the alveolar gas spaces is also reduced to below that of the air, and thus inflation occurs as a result of a hydrostatic pressure gradient from the mouth to the alveoli.

Inflation in the course of normal tidal breathing usually commences from a state of partial lung inflation that reflects that particular volume of the lung at which its own elastic recoil tendency to collapse is exactly balanced by the opposite, natural tendency of the rib cage to expand outwards. This volume is known as the functional residual capacity (**FRC**) and because it reflects recoil balance between lung and chest wall, it is the only volume which can be maintained without muscular effort. Thus, to inhale above FRC or to exhale below FRC both require respiratory muscle contraction, but the return to FRC from either higher or lower volumes can be passive, stored elastic energy provided by respiratory muscle contraction from the preceding active volume change being used to reverse the transpulmonary pressure difference and enable gas flow from the alveoli to the mouth.

Clinical significance

Elastic properties and lung volume

If the elastic properties of either the lungs or the chest wall are altered by disease, FRC will change. Should the lungs become less elastic, typically seen in emphysema due to disorganization of the elastin and collagen fibres making up much of the alveolar wall structure, the tendency for the lung to collapse is less, and the lung/chest wall recoil balance shifts to a higher lung volume, thus increasing FRC. By contrast, diseases characterized by proliferation of alveolar wall elements, collagen in particular, render the lung more elastic and thus collapsible, shifting FRC to lower values. These changes in FRC may be used to aid in diagnosis and in following the natural history and response to treatment of such diseases, since FRC is readily measured in the pulmonary function laboratory by either plethysmography or helium dilution methods. Changes in FRC also have important implications for lung function, discussed later in this chapter.

Whilst FRC is a key volume upon which to focus, the lung can normally be inflated to well above FRC, and also deflated to considerably below FRC. At maximal inflation, lung volume is referred to as the total lung capacity (**TLC**), while at maximal deflation, lung volume is called the residual volume (**RV**). Of major significance, RV is well above zero volume. As will be apparent, if all alveoli could be fully emptied of gas, they would be very difficult to reinflate to allow resumption of gas exchange, due to surface tension. The difference between TLC and RV is called the vital capacity (**VC**). As with FRC, each of these volumes is readily measured during routine pulmonary function testing, and together they provide a simple yet informative profile useful in characterizing many lung diseases and their progress. Unlike some physiological variables such as arterial pH or haemoglobin concentration, all of the above volumes depend to a major extent on body size. They also depend to a lesser degree on gender (smaller in females), age (deterioration with ageing), bodily habitus (often smaller in the obese), and ethnicity. Many tables of normal values have been published, and interpretation must allow for all of the determinants mentioned above.

Trachea, main bronchi, and pleura

For all 300 million alveoli to participate in the gas exchange process, each must be connected to the environment by an air pathway. The analogy now changes from a balloon to a tree. Imagining an upside-down tree, the main trunk represents the trachea, the single common airway segment through which inhaled and exhaled gas from all alveoli must pass. The upper end of the trachea begins at the lower margin of the larynx. The trachea lies anteriorly in the neck and chest, passing caudally in the midline retrosternally to the level of about the sternal attachment of the second rib. There it divides into left and right mainstem bronchi, each smaller and shorter than the trachea. These two airways angle caudally and laterally within the upper mediastinum to enter the left and right lungs at the left and right hilar regions, respectively, and they divide into the lobar bronchi, three on the right to feed the right upper, middle, and lower lobes, and two on the left to feed the two left lobes, upper and lower.

Note that the two hilar regions are the only normal points of actual connection of the left and right lungs to any thoracic structures, and also contain the large pulmonary arteries and veins, lymphatics, and nerves. The entire remaining lungs, while opposed against the chest wall, are not connected to it and are able to slide easily over the inner chest wall surface. This inner surface is covered by the parietal pleural membrane, and the outer surface of the lungs is similarly covered by the visceral pleural membrane. These two pleural membranes are joined at the hilar regions to form a fully enclosed sac that separates the lung and chest wall. The left and right pleural sacs do not communicate with each other, and normally contain only a very thin layer of plasma-like fluid and no gas at all. This arrangement may be pictured by imagining a sealed, but empty, plastic sandwich bag from which all air has been expelled and which contains a very small volume of water. If one's right hand is balled into a fist and invaginates this bilayered bag against the cupped left hand, we have the analogy to the right (or left) lung and chest wall. The balled right fist is the lung; the right wrist and forearm represent the hilar structures. The cupped left hand is the chest wall, and the two layers of the closed sandwich bag form the pleural membranes.

Clinical significance

Mainstem bronchial branching angles

The mainstem bronchial branching from the trachea is not quite symmetrical. The right mainstem bronchus continues caudally a little more directly in line with the trachea above it than does the left, which angles laterally more sharply. As a result, accidentally inhaled foreign bodies more frequently lodge in the right than left lungs. For similar reasons, advancing an endotracheal tube too deeply may cause it to lodge in the right mainstem bronchus rather than where intended—the trachea. This will result in lack of ventilation of the left lung, and if not recognized, hypoxaemia from continued perfusion of this unventilated lung with venous blood, and ultimately left lung collapse (over minutes to hours).

The intrapleural space and pneumothorax

The pressure within the pleural space (i.e. between visceral and parietal pleural surfaces) is normally subatmospheric because of the above-mentioned counterbalancing inward lung and outward chest wall recoil forces. This prevents lung collapse. Disruption of either the visceral or parietal pleura (i.e. pneumothorax) allows air to enter the pleural space, increasing the intrapleural pressure back to atmospheric. This results in collapse of the lung, with abolition of ventilation even if chest wall muscle contraction continues. Gas exchange therefore ceases, with life threatened. In humans, since the right and left lungs are encased in separate pleural sacs, if one side suffers pneumothorax, gas exchange can usually be maintained by the other. Pneumothorax can occur from rupture of lung surface alveoli in predisposed individuals, or from chest wall trauma in anyone. Whether the source of the intrapleural air is alveolar gas as in the former case or room air in the latter makes no difference. However, depending on conditions, intrapleural air pressure may actually rise above that of room air. This situation, the tension pneumothorax, can arise whenever air enters the pleural space via a valve-like mechanism, when the patient's respiratory effort or that of a mechanical ventilator can lead to intrapleural pressure rising well above atmospheric. The lung collapses, but the (increasingly desperate) respiratory effort or mechanical ventilator keeps pumping air

into the pleural space via the torn lung surface. This is a true emergency requiring immediate needle puncture of the chest wall of the affected side to relieve the built-up pressure. If this is not done, the high intrathoracic pressure compresses and distorts the mediastinum and vena cavae, impeding venous return. Both pulmonary gas exchange and the circulation fail, and death follows rapidly.

Mediastinal shifts

The separation of the right from left pleural spaces provides for lateral movement of the mediastinum should there be a difference in mechanical properties of the right and left lungs or their associated pleural spaces or chest wall structures. For example, fibrosis of the right lung, or alternatively its collapse from complete airway obstruction, will reduce the volume of intrathoracic contents and therefore pressure on that side, and mediastinal contents will shift towards the right, visible on chest radiography. In fact, the trachea may also be shifted from its normal midline location in this direction, evident on clinical examination of tracheal position just above the suprasternal notch. Conversely, a pleural effusion on the right or a right pneumothorax (see below) may raise intrathoracic pressure above that on the left, and have the opposite effects on mediastinal and tracheal position.

The bronchi and bronchioles

After the mainstem bronchi have arisen from the trachea, the airways continue an essentially dichotomous branching pattern until the alveoli are reached. Thus, successive branching yields of the order of 50 000 to 100 000 airways (called terminal bronchioles) that constitute the 16th generation ($2^{16} = 65\,536$). The entire collection of airways from the trachea to these last bronchioles before alveoli begin forms a system of connected conducting pipes needed to deliver gas between the alveoli and the environment during ventilation (Fig. 2).

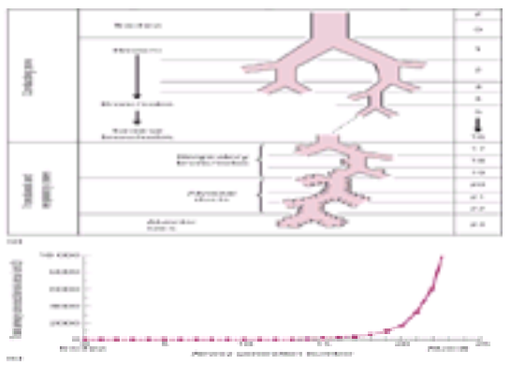

Fig. 2 (a) This shows a stylized model of the branching of the airways from trachea to alveoli encompassing some 23 generations of branching. The first 16 generations contain no alveoli and are purely conducting airways, but the next 7 generations contain progressively more alveoli in the airway walls and serve the dual -purpose of conducting air to the alveolar sacs and also providing gas exchange. (b) Total cross-sectional area of each generation shown in (a). This is obtained by multiplying the average cross-sectional area of a single airway by the number of airways in the particular generation. Cross-sectional area is small throughout the conducting zone (first 16 generations), but then increases exponentially in the respiratory zone. The implications are that the forward velocity of inspired gas falls dramatically in the respiratory zone such that diffusion becomes the faster mode of molecular movement. In addition, this diagram implies that during flow between the mouth and alveoli, most of the airways resistance resides in the first 15 generations. (Adapted from Weibel ER, 1984, with permission.)

As with the branching of a tree, both the diameter and the length of each successive branch falls. The trachea typically is 12 cm long and 2 cm in diameter. By contrast, the typical terminal bronchiole is just 1 to 2 mm long and 0.6 mm in diameter. Airflow is normally mostly laminar (except for that in the upper airways) and therefore is governed by Poiseuille's law of fluid dynamics. The essence of this law is that resistance to airflow depends inversely on the fourth power of the airway radius, but varies only in direct proportion to airway length. As airways become both narrower and shorter with increasing branching, it is evident that resistance of a single airway increases dramatically because of the dominating effect of the fourth power of the radius. However, if one asks how the entire system behaves by plotting how airway pressure must fall from trachea to generation 16 during, for example, steady inspiratory flow, one must allow for the fact that all airways of any single generation are arranged in parallel with one another. Because branching is essentially dichotomous, there are twice as many airways in any given generation as in the one before. Thus, total airway resistance of any one generation is diminished in proportion to the exponentially increasing number of airways as branching continues. This actually overcomes the fourth power disadvantage of Poiseuille's law, such that most of the pressure drop, or put another way, most of the system airway resistance, is associated with the first few generations despite their large individual airway size.

Another way to understand this somewhat counterintuitive result is to consider the sum total of the cross-sectional areas of all airways in a single generation. This is of course the area of a typical airway multiplied by the number of airways in that generation. That number is low for the first few generations, but then rises dramatically because of the exponentially increasing number of airways in each generation. Airway resistance of a generation therefore falls from the first few generations to the terminal bronchioles. The summed total volume of gas contained within all 16 generations of these conducting airways is only about 150 ml despite their prodigious number.

Clinical significance

Dead space

The interposition of airways between the mouth and the alveoli creates a volume of gas (about 150 ml as mentioned) called the anatomical dead space. The gas in this dead space simply passes back and forth during inspiration and expiration without contributing to gas exchange since the conducting airways contain no alveoli in their walls. It constitutes a penalty since it adds an obligatory 150 ml volume requirement to every breath taken. This is of no importance in health, but in patients with severe lung disease such as chronic obstructive lung disease or fibrosis, the energy cost of overcoming either high resistance in obstructed airways or low compliance of fibrotic lung tissue, and of thus mounting adequate ventilation, may be greatly increased. Then, the need to breathe some 150 ml more per breath than actually required for alveolar gas exchange can be clinically important as a factor contributing to respiratory failure. Recognition of this has led to the use of transtracheal insufflation of air, which permits the anatomical dead space of at least the upper airways to be circumvented and reduces the ventilation necessary for any given activity.

Particle deposition

Ventilation involves breathing some 6 to 10 litres of air every minute of our lives. Air contains much particulate matter of very small size. Depending on particle size, rate of gas flow in the airways, and airway geometry, such particles may move harmlessly in and out with the next breath or they may be deposited somewhere on the epithelial surface in the bronchial tree. To the extent that they do deposit and are chemically or physically harmful to tissue, they can be responsible for disease. Pneumoconioses, chronic obstructive pulmonary disease, bacterial and viral infections, asthma, and other diseases may all be initiated and/or afftected by such mechanisms. The dividing airway structure described above combines ever-diminishing individual airway diameter with ever-diminishing gas velocity (due to increasing summed cross-sectional airway area of all airways in a generation) as branching continues. As airways narrow and flow velocity falls, the chance of airborne particles being deposited on airway walls increases. It is for this reason that coal dust, for example, settles mostly in the terminal bronchiolar region deep within the branching system. Thus, the basic nature of gas exchange, demanding the branching network of airways described, leads to intrinsic vulnerability to disease from airborne particulate matter.

Mucociliary function

As seen commonly in evolutionary responses to deleterious phenomena, a protective system has been developed to mitigate the consequences of particle deposition in the airways. This is the mucociliary apparatus. It has several components. There are submucosal glands in the walls of the conducting airways that secrete mucus into the airway lumen when stimulated by irritant signals. These glands are supported by other secretory cells in the epithelium of the airways such as goblet cells. The epithelial cells that line the entire conducting airway system are ciliated, and they function in a co-ordinated manner, beating rhythmically to move the secreted

mucus upward from smaller to larger airways. The primary purpose of the mucus is to trap inhaled particulates before they can reach and damage the airway and lung tissues themselves. This upwardly transported mucus is clinically evident as sputum.

The volumes of sputum produced normally are so small as to be unnoticeable, and are usually swallowed. However, inhalation of toxic irritants, infectious agents, and other particles will rapidly increase the volume of sputum to noticeable levels, and chronic airway inflammation from, for example, cigarette smoking will produce chronically increased amounts of mucus that give rise to the syndrome of chronic bronchitis. It is especially noteworthy that in asthma, not only is the volume of mucus increased, probably from airway inflammation, but its composition is altered, rendering it much more tenacious and difficult to eliminate by the ciliary system. Mucus thus accumulates in the airway lumina, particularly those of the smaller conducting bronchioles, creating mucus plugs that cause obstruction to airflow and marked reduction in ventilation of alveoli lying distal to them. When this occurs, asthma is often refractory to usual pharmacological therapy, and patients dying from asthma universally exhibit widespread airway mucus plugging.

Dynamic airway compression

Another intrinsic physiological problem of the branching airway system within the chest is related to the mechanical nature of respiration—the need for inflating and deflating the lung by altering the pressure around it—combined with the fact that the airways are not rigid tubes. The airways are thus susceptible to expansion and compression (and therefore to collapse) on inspiration and expiration, respectively. The intrapleural pressure may be transmitted to the conducting airways, and while reduction in this pressure on inspiration will only distend the airways, allowing air to flow more freely, opposite effects during expiration may not be innocuous. Passive expiration, that is, expiration fuelled only by the elastic energy stored in the lung tissue from the previous inspiration, without active expiratory muscle effort, does not compress the airways because the intrapleural pressure remains subatmospheric. However, active expiratory muscle contraction as occurs during a forced expiratory manoeuvre and during heavy exercise leads to compression of the airways because intrapleural pressure is raised to above atmospheric. In fact, the greater the expiratory effort made, the greater the increase in intrapleural pressure and the degree of airway compression. Because of this, flow rates uring forced expiration cannot be increased by making a greater muscular effort: any greater driving pressure for expiratory flow is balanced by the increased resistance resulting from more compression. As a result, even in normal subjects, expiratory flow of air under these conditions is limited by this phenomenon, known as dynamic compression, which is illustrated in Fig. 3.

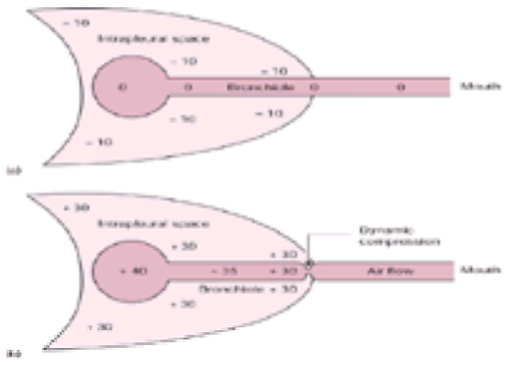

Fig. 3 Diagram to explain dynamic compression during expiration. (a) This depicts intrapleural, alveolar, and airway pressures while breath holding with an open glottis at total lung capacity. Due to lung elasticity, intrapleural pressure is negative (–10 cmH2O), but because of breath holding there is no flow, and pressure in the airways and alveoli equals that at the mouth, 0 cmH_2O. Immediately after commencing a forced expiration from total lung capacity (b), intrapleural pressure is high due to expiratory muscle contraction (+30 cmH_2O). Alveolar pressure is even higher due to 10 cmH_2O of lung elastic recoil pressure. However, due to flow resistance, pressure falls from +40 gradually to +30 as shown. At this point, intrapleural pressure equals intraluminal pressure and immediately downstream dynamic compression occurs as airway pressure falls even further and is now less than intrapleural pressure.

The loss of elastic recoil in emphysema, mentioned above in the context of its effects on FRC, also has a major influence on dynamic compression. The airways are much more susceptible to dynamic compression (discussed below), such that even breathing at rest with just small increases in intrapleural pressure from active expiratory muscle contraction may be subject to flow limitation by this mechanism. When this problem is compounded by the separate phenomenon of increased airway luminal mucus from chronic inflammation induced by cigarette smoking, it is easy to understand how chronic obstructive lung disease (emphysema and chronic bronchitis) has airway obstruction as its major disturbance.

In the consideration of dynamic compression it is important to note that the alveoli are not physically independent of one another or of the conducting airways, which run within the lung parenchyma from the lobar bronchi all the way out to the terminal bronchioles. The alveoli share walls in their mutual attachments, and the alveoli beside any intrapulmonary conducting airway are physically connected to the outside of that airway wall. A good analogy for how the alveolar parenchyma is configured comes from examining the cut surface of a sponge where the myriad air cells are surrounded by thin tissue walls. Every wall serves two adjacent air cells, and the overall structure is solid (rather than like the leaves of the tree which are physically independent of each other even while being connected to the same dividing network of branches). The net result of this matrix of alveolar and airway connections is that when the lung is inflated, the elastic tension in the parenchyma exerts radial traction on the conducting airways, increasingly so as the lung is further inflated. This stiffens the airway walls and acts to oppose dynamic compression during active expiration. That maximal expiratory flows are greater at high than low lung volumes is explained by the greater radial traction at high volumes as the alveoli are stretched more.

The walls of the larger conducting airways (the trachea and first few generations of bronchi) are reinforced with cartilage rings that further help to counter the forces favouring dynamic compression. However, the smaller conducting airways do not enjoy this protection, and it is in the smaller airways that dynamic compression usually has its major effects.

Airway smooth muscle

All generations of conducting airways contain smooth muscle cells. When stimulated to contract, their concentric arrangement leads to reduction in airway lumen size, and airway obstruction results. While not a significant effect in normal individuals, patients with asthma have hyperresponsive airway smooth muscle that contracts in response to the inflammatory reaction usually present in the asthmatic airway walls. This is a major mechanism of airway obstruction in asthma, and is the basis of the mainstay therapy in this disease—bronchodilators. For reasons that remain unclear, smooth muscle contraction does not occur to the same degree in all airways of the asthmatic lung: there are different degrees of obstruction both with respect to airway generation number and among airways of a given generation. Ventilation of alveoli is thus very uneven, with many alveoli being very poorly supplied with air, yet others well-supplied. Gas exchange becomes inefficient as a result, and arterial hypoxaemia is seen.

Airway smooth muscle also contracts when local CO_2 concentrations fall. This happens commonly in pulmonary thromboembolism, when vascular obstruction results in focal areas of hypoperfusion that remain relatively overventilated, such that their local alveolar CO_2 tension falls. This, possibly in concert with bronchoactive inflammatory mediators released in association with the embolic event, can produce local airway smooth muscle contraction and airway obstruction. This might tend to better matching of local ventilation with blood flow, but the benefit is generally small, and local bronchoconstriction can manifest as wheezing, which should not be mistaken for asthma.

Dynamic tests of airflow

All of the consequences of the branched structure of the airways and their interconnectedness need to be integrated if one is to understand common pulmonary function tests. How the 'static' lung volumes (FRC, TLC, VC, and RV) are affected by changes in elastic recoil are discussed above, but such measures form only a part of standard pulmonary function testing. Usually included are 'dynamic' tests that measure expiratory and inspiratory gas flow rates, conventionally during manoeuvres wherein the patient is asked to make a maximal inspiratory or expiratory muscle effort. These are discussed in Chapter 17.3.2.

Distribution of ventilation

The extremely large number of very small respiratory bronchioles creates an environment in which alveoli distal to each bronchiole become susceptible to impaired ventilation. Small intrinsic or pathological reductions in airway diameter of such bronchioles can impair distal ventilation substantially. When the effects of variation in mucus secretion, bronchial smooth muscle tone, and radial traction are added to this inherently vulnerable system, it is surprising that the distribution of ventilation to the 300 million alveoli is as uniform as it is. Were it not, there would probably be considerable hypoxaemia, even in health. This topic is discussed further below.

The parenchyma distal to the terminal bronchioles

The terminal bronchioles (16th generation airways) are the final divisions of the wholly conducting airways. They are completely lined with ciliated epithelium, and function primarily as simple conduits for gas, linking the air around us to the alveoli where gas exchange occurs. The next few divisions of the airways result in transitional airways called respiratory bronchioles, so named because they serve a dual role—as continued gas conduits and as the first locations for gas exchange. Respiratory bronchioles are partly lined with ciliated epithelium, but also have small alveolar outpouchings opening directly into the airway lumen. With continued branching of these bronchioles, more and more of the luminal surface is given to the alveolar outpouchings, and less and less to ciliated epithelium. After about three generations of respiratory bronchioles, the airways, whilst still essentially tubular in shape, are made up entirely of alveolar tissue capable of gas exchange, and are called alveolar ducts. These alveolar ducts branch even further into collections of alveoli whose distal end is blind, known as alveolar sacs, the end of the line of the airway branching system. A diagram of the functional lung unit is shown in Fig. 4. With some 7 orders (or division points) of branching between the terminal bronchioles and the final alveoli, together with 16 orders of branching in the conducting airway segment, the whole airway tree consists of about 23 orders or branch points. Because, after the final branch point, the alveolar sacs are blind, the process of ventilation must occur as a tidal (back and forth) event, alternately adding air to, and removing alveolar gas from, each alveolus with each breath.

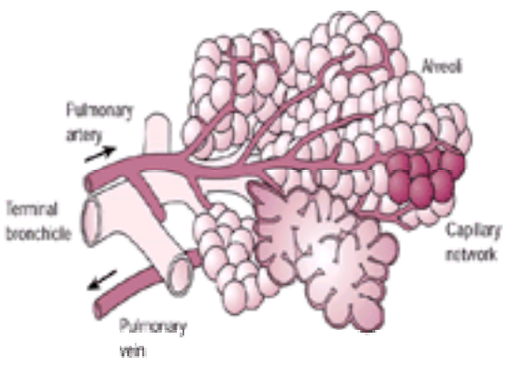

Fig. 4 Diagram of the functional lung unit. The collection of alveoli and associated pulmonary arteries and veins distal to the terminal bronchiole constitutes a functionally homogeneous unit of gas exchange. Mixing of gas amongst alveoli and of blood in the capillary networks of the alveoli in the unit is sufficiently rapid that gas concentrations are in effect uniform throughout. This unit, also called the acinus, corresponds approximately to generations 17 to 23 of Fig. 2(a).

The transport of gas in either direction between the trachea and the last conducting airway takes place principally by convective flow, much as water flowing in a pipe depends on the pressure difference between the two ends of the pipe and the flow resistance of the pipe. Since flow is mostly laminar, velocity profiles are largely parabolic, flow being highest in the centre of the lumen and lowest at the airway wall, just as is the velocity profile across a quietly flowing river. There are, however, minor additional influences of diffusive movement at the interface between the convective front of each inspiration and residual gas from the previous breath. These interactions, and eddies that develop at each branch point, may assist gas mixing, but their effects are physiologically small. Of much more significance is the fact that the total luminal cross-sectional area of each generation increases exponentially as the airways divide. Since total volumetric flow of gas is the same in each generation, average gas velocity falls reciprocally with the increase in area.

By the time inspired gas reaches the first alveoli, forward velocity has dropped to such a low level that random, thermally fuelled molecular motion (i.e. diffusion) becomes a more important mechanism of gas transport than convection. The small size of the alveoli, about 150 μm in radius, means that diffusive mixing of each new breath with gas resident in the alveoli from prior breaths is nearly instantaneous. Although careful physiological studies can show that low-molecular-weight gases mix slightly faster than those of high molecular weight, this turns out to be of essentially no quantitative significance to gas exchange. Even in emphysema, where many alveolar spaces are enlarged, there is evidence that diffusive mixing in alveolar gas is functionally complete and does not pose a gas exchange threat.

Of more concern for gas exchange is whether all alveoli receive a similar share of each breath. It was pointed out above that the intrinsic structure of the lungs makes it vulnerable to ventilatory inequality, and that this has the potential to disrupt gas exchange. Indeed, recent studies of the structural influence on gas distribution reveals that there are sometimes substantial differences in the ventilation of different alveoli. One property of the system that lessens the negative effects of such inhomogeneity on gas exchange is the finding that individual alveoli do not maintain gas exchange differences from closely adjacent alveoli. In fact, a fairly large number of connected alveoli are normally able to function as a single homogeneous unit of gas exchange. This is no doubt due partly to the rapid diffusive movement of molecules throughout the alveolar gas mentioned above, but it is also facilitated by the rich capillary network lying in the wall of each alveolus. The density of capillaries is so great that should flow fall in one, its neighbour can seamlessly take over its gas exchange role without any resultant inefficiency. It turns out that the functional unit of gas exchange, known as the acinus, corresponds approximately to all the alveoli distal to the last terminal bronchiole.

Clinical significance

The functional lung unit

Pathological events, in either the alveoli or the capillaries, occurring at a scale smaller than that of the functional lung unit will not *per se* have much impact on gas exchange. Thus, a large number of tiny pulmonary emboli each lodging in one capillary of different functional lung units will not impair gas exchange function, whilst a single large embolus of the same total mass obstructing one much larger vessel may. However, if enough microvessels within functional units become obstructed, their summed effects may become considerable.

Surface tension and mechanical instability of the lung

Another consequence of the branched nature of the lungs resulting in so many very small alveoli is inherent mechanical instability. The alveolar wall, where it interfaces with alveolar gas, forms a roughly spherical air–liquid interface. In this context, the alveoli may be likened to a mass of soap bubbles lying together. All air–liquid interfaces are subject to surface tension, which in this case will act to minimize the surface area of each bubble. For an enclosed bubble, this tension increases the pressure inside the bubble, with the relationship between the tension and the interior pressure given by the law of Laplace: pressure = 2 × surface tension/radius. Thus, pressure inside a small bubble exceeds that inside a larger bubble, and if two such unequal bubbles are in contact and their interiors become connected, the small bubble will collapse into the larger. This process of bubble accretion may continue until the many small soap bubbles have collapsed into a single large one. Based on the opening premise of this chapter, if small alveoli had this tendency to collapse into larger neighbours due to surface tension effects, the end result would be disaster for gas exchange. There would be massive alveolar collapse, and with loss of surface area, sufficient O_2 exchange to support metabolic needs would not be possible. Only if all alveoli were identical in both size and surface tension would this problem be avoided, but when 300 million alveoli exist, it is impossible to imagine them all being identical, and indeed they are not.

The lung avoids this dilemma through two quite separate but complementary mechanisms of stabilization. The first, already mentioned above in a different context, is the interconnected nature of the whole alveolar structure. Any tendency for one alveolus to collapse would have to increase the tension on all its immediately connected neighbours. This tension from surrounding alveoli will automatically serve to splint open the alveolus in question, thus opposing its tendency to collapse. This concept, termed alveolar interdependence, is felt to be of considerable importance in maintaining alveolar stability. The second mechanism is the presence of phospholipid molecules that reduce surface tension in the alveolar air–liquid interface. Termed surfactant, and produced in conjunction with proteins from alveolar

type II epithelial cells lying free in the alveolar spaces against alveolar walls, this material reduces surface tension severalfold (Fig. 5). Thus, whilst that of water is some 75 dyn/cm, surface tension of the alveolar lining fluid is only about 10 dyn/cm. Moreover, probably due to molecular realignment of surfactant molecules, surface tension is even lower when lung volume is reduced. Based on the law of Laplace given above, this can be seen to be even more advantageous for evening out surface tension differences among alveoli of different size.

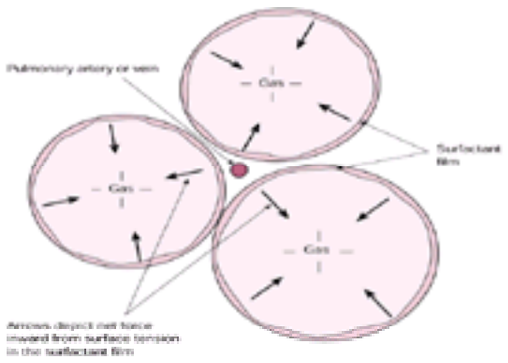

Fig. 5 Diagram to indicate potential effects of surface tension on lung structure and function. Three gas-filled alveoli are shown, each lined by a thin film of surfactant. A pulmonary artery or vein is shown in the corner formed where the three alveoli come together. Arrows show the net inward force produced by surface tension, tending to reduce alveolar gas volume and promote atelectasis. In addition, the pressure in the perivascular space around the corner vessel shown will be reduced by these inward surface forces, increasing the pressure difference from inside to outside the vessel lumen and thereby promoting fluid movement from plasma to interstitial space. The presence of surfactant reduces the magnitude of surface tension forces and therefore stabilizes the alveoli against atelectasis and reduces the transmural pressure difference, attenuating transvascular fluid movement.

Surfactant is thought to have another crucial role that promotes efficient gas exchange between alveolar gas and capillary blood. Given that adjacent alveoli share a common wall, the tendency for surface reduction in each alveolus will create a force that tends to reduce the interstitial tissue pressure around capillaries in the alveolar wall between the adjacent alveoli. From the Starling relationship that governs water escape out of capillaries in any tissue (based on the transcapillary differences in both hydrostatic and oncotic pressures), reducing pressure around the capillary will lead to increased water escape into the alveolar wall. This could have several deleterious consequences. First, the affected alveolar walls would become stiffer and harder to inflate, tending to reduce lung volume. Second, the tissue separating gas from capillary blood would become thicker, directly impairing diffusive transport between gas and blood. Third, this water would find its way into the pulmonary lymphatics, which begin in the alveolar interstitium and run along the large airways and vessels to the hilar regions, before exiting the lungs and emptying into the superior vena cava. Extra water frequently accumulates in the peribronchial and perivascular spaces and results in their partial compression, reducing distal ventilation and/or blood flow of subtended alveoli, causing maldistribution of either or both, and rendering gas exchange inefficient. The presence of surfactant is thought to reduce the rate of transcapillary water exchange, and therefore to contribute to efficient gas exchange.

Clinical significance

Impaired surfactant activity

When surfactant is not present, when its rate of renewal is insufficient, or when it is inactivated rapidly, pathological changes can be severe. Best known is the infant respiratory distress syndrome, occurring in otherwise normal premature infants born before the late-maturing surfactant system is functional. Without exogenous surfactant replacement therapy the condition may be fatal because of alveolar collapse and pulmonary oedema. Surfactant activity is also compromised in the adult respiratory distress syndrome and may compound the disturbances of pulmonary function arising from the primary cause of the pulmonary disease.

Gravity and lung function

Causes of potential unequal distribution of ventilation or blood flow to the alveoli extend beyond those associated with the intrinsic branching structure of the lungs discussed above. In particular, the presence of gravity influences lung function because key components of the lungs have significant weight. The weight of the parenchyma itself, plus the blood within the alveolar capillaries, feeding arteries and draining veins, together cause the lungs to sag toward the diaphragm in the upright lung sitting at FRC. The upper pole of the lungs is still applied to the parietal pleural surface of the chest wall—there is no pleural airspace created by this gravitational stress. Rather, the rest of the lung is displaced caudally, sagging much like a heavy sweater pegged to a clothes line. As expected, this creates stress in the alveolar walls, more in the uppermost than lowermost alveoli. A good analogy is the toy Slinky—a coiled spring that when hanging vertically under its own weight shows wider separation between adjacent coils at its top than at its bottom. Correspondingly, the uppermost alveoli in the upright lung are larger than the lowermost alveoli. The lowermost alveoli are thus more compliant—that is, able to be further inflated more per unit transpulmonary pressure—than the uppermost alveoli, because the latter are stretched almost to their limit. Accordingly, normal ventilation from FRC results in greater ventilation of the lung bases than of the lung apices. Much the same effect is seen for blood flow: apical blood flow is less than that at the base of the upright lung. In this case it is the weight of the blood itself that is responsible: perfusion depends on pulmonary arterial pressure, which falls linearly with height up the lung.

The apex to base differences in perfusion exceed those of ventilation, such that the ratio of ventilation to blood flow is higher at the apex than at the base. The local ventilation/perfusion ($\dot{V}_A/\dot{Q}$) ratio determines local alveolar P_{O_2} and P_{CO_2}, P_{O_2} increasing and P_{CO_2} falling as the $\dot{V}_A/\dot{Q}$ ratio increases. Thus, P_{O_2} at the apex is higher, and P_{CO_2} lower than at the base. If the $\dot{V}_A/\dot{Q}$ ratio everywhere was the same, so too would be P_{O_2} and P_{CO_2}, and the exchange of O_2 and CO_2 would be maximally efficient. However, the presence of a range of $\dot{V}_A/\dot{Q}$ ratios (no matter what its cause) results in gas exchange inefficiency and arterial hypoxaemia.

Clinical significance

Effects of gravity on lung function in disease

Although gravity creates $\dot{V}_A/\dot{Q}$ maldistribution, common disease processes are in large part randomly distributed in the lungs, and their effects on $\dot{V}_A/\dot{Q}$ matching generally much greater than those of gravity. Thus, whilst the effect of gravity on $\dot{V}_A/\dot{Q}$ mismatching in normal lungs is barely measurable, $\dot{V}_A/\dot{Q}$ mismatching based on non-gravitational influences in many diseases leads to profound gas exchange disturbances. However, the presence of gravity must not be discounted in several disease states.

Emphysema, and even the normal ageing process, often causes tissue breakdown in the apical lung regions because, as in the Slinky analogy, the alveolar wall stresses are largest there. When mechanical failure occurs, it is most likely to happen in the regions of greatest stress, and as a result the alveolar wall breakdown so typical of emphysema, and to a much lesser extent normal ageing, is often exaggerated in the apices. An important gravitational influence occurs in patients in intensive care with severe lung disease. In any body position, both blood flow and alveolar fluid collection tend to be concentrated in dependent regions (e.g. posteriorly in the supine patient). Those regions with high blood flow may also have little or no ventilation if their alveoli are filled with fluid and cell debris. The blood flowing through such regions can therefore pick up little or no O_2, and hypoxaemia may be severe. This has led some intensive care staff to rotate their patients from supine to lateral to prone and back. The argument being that the gravitational influences on blood flow are essentially instantaneous, whilst those on alveolar fluid collection may take hours to respond to body positional changes. Thus, for a time after rotating a patient, the dependent region may enjoy high flow but not yet be fluid filled and thus still be well ventilated. Gas exchange is therefore enhanced, and arterial hypoxaemia is mitigated. Such behaviour may also explain positional influences on gas exchange in patients with unilateral lung disease such as pneumonia, effusion, or atelectasis.

Further reading

Crystal R, West JB (1994). *The lung: scientific foundations*. Raven Press, New York.

Weibel ER (1963). *Morphometry of human lung*. Springer-Verlag, Berlin.

Weibel ER (1984). *Pathway for oxygen: structure and function in the mammalian respiratory system.* Harvard University Press, Cambridge, Massachusetts.

West JB (1990). *Respiratory physiology, the essentials*, 4th edn. Williams & Wilkins, Baltimore, Maryland.

17.1.3 'First-line' defence mechanisms of the lung

C. Haslett

Introduction

In their critically important service as our central gas exchange organs the lungs are continuously exposed to more than 7000 litres of air per day, but their membranes are delicate and require to be kept moist and protected from the daily bombardment of particles including dust, pollen, and pollutants, together with viruses and bacteria. These agents have the potential to cause lung injury or to invade the lung and generate potentially life-threatening infections. That these problems rarely occur is because the lung possesses very effective local 'primary' protective mechanisms, which are the focus of this chapter. If an infectious agent is able to penetrate these defences and set up a 'bridge-head', highly effective and complex 'secondary' responses, including the inflammatory and classic immune responses, can be recruited rapidly. In the event that immune or inflammatory responses should be initiated, the lung also has mechanisms by which it can protect itself from their potentially detrimental local side-effects. The inflammatory and immune responses themselves will only briefly be alluded to; detailed treatment of these processes is beyond the scope of this chapter (see Chapter 16.5.1 for further discussion).

The respiratory tract is protected by different mechanisms at its various levels. In general terms, physical mechanisms, including cough, are particularly important in the large airways; the lower airways are protected by complex mucociliary clearance mechanisms; and the gas exchange units at the alveolar level are protected by surfactant and by 'patrolling' alveolar macrophages. The lung lining fluids (mucus in the airways and surfactant in the gas exchange units) contain a variety of proteins that are particularly important in host defence. In this section we will therefore consider physical defences, mucociliary clearance mechanisms, surfactant and important defensive proteins in the lining fluid of the lung, and how the 'second-line defences'—the classic inflammatory and immune responses—can be initiated in the lungs.

Physical defences

The nose makes an important contribution to the physical defences of the upper airway. It comprises a stack of fine aerodynamic filters of respiratory epithelium arranged over the turbinate bones. These remove most large particles from inspired air. Their filtering effect is greatly enhanced by fine hairs in the anterior nares and by mucociliary action that, apart from a small area anterior to the inferior turbinates, is directed posteriorly such that trapped particles are swallowed or expectorated. The larynx acts as a sphincter during cough and expectoration and is an essential protective mechanism for the lower airways during swallowing and vomiting.

Particles with a size greater than 0.5 µm that survive passage through the nose will be trapped by the lining fluid of the trachea and bronchi, to be cleared by the mucociliary clearance mechanism, which has been called the 'mucociliary escalator'.

Mucociliary clearance

Cilia

The mucociliary escalator works by a complex interaction between cilia, which are a series of projections on the bronchial epithelial cells, and mucus, forming a 'raft' on top of the cilia, which then sweep this raft in a cephalad direction. The combined effect of this interaction can readily be appreciated by scanning electron microscopy (Fig. 1). There are about 200 cilia on each of the pseudostratified columnar epithelial cells lining the bronchi and it has been calculated that these can carry weights of up to 10 g/cm without slowing, working with a ciliary beat frequency of 12 to 14 beats/s. Their motility depends upon the contraction of longitudinal fibrils that contain the contractile protein tubilin arranged as nine outer and two central microtubular pairs, and their effectiveness in sweeping mucus in the cephalic direction is enhanced by small 'claws' in their tips, which penetrate the overlying mucus sheet. The actual driving mechanism of the cilia is uncertain, but involves dynein, an ATPase protein that forms a major part of the cilium. This appears to derive energy from ATP along the cilium and convert it into the forces that are generated by the contractile proteins.

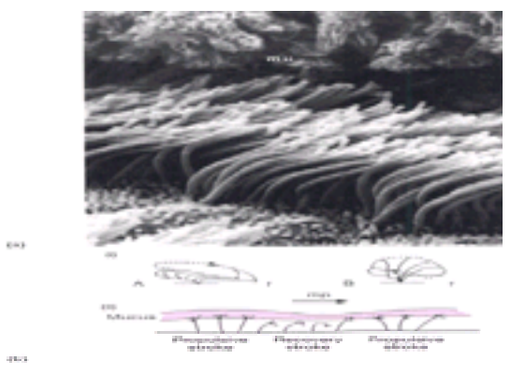

Fig. 1 (a) Scanning electron micrograph (× 14 000) of the 'mucociliary escalator' of the bronchial epithelium. The cilia and the overlying 'raft' of mucus (mu) are clearly seen. (By courtesy of Dr P.K. Jeffery.) (b) Beat cycle of a tracheal cilium seen from the side: (i) In a single cilium during the recovery stroke (A) the cilium starts from its rest position r and unrolls backwards; in the propulsive stroke it remains extended to reach its rest position (B). Mucus is propelled to the right (mp). (ii) Groups of cilia move in co-ordinated waves with clusters in their propulsive phase bordered by those in recovery. (Adapted from Sleigh MA, Blake JR, Liron N (1988). The propulsion of mucus by cilia. *American Review of Respiratory Disease* **137**, 726.)

It is generally believed that ciliary beating occurs by the microtubules, which provide the 'skeleton' of the cilia, sliding over each other, much like the sliding fibre theory of muscle contraction. Since not all microtubular pairs move at the same time, their co-ordinated shortening leads to reduction in length of some microtubules relative to those at the opposite side of the cilium, and with the skeletal rigidity that is provided by the radial pairs of the microcilial 'spokes' and their basal anchoring system, this causes the cilium to bend in the direction of shortening. The ciliary beat itself can be divided into two phases: the forward mucus-propulsive stroke, and a slower recovery stroke (much like the forward and recovery actions of a whip, Fig. 1(b)). Where this occurs in a co-ordinated fashion in the ciliated epithelium, the wave-like motion of numerous cilia effected through their terminal claws propels the mucus raft in a cephalad direction.

Ciliary function can be assessed in a number of ways, facilitated by the fact that cilia can survive freezing for up to a month and may beat for several hours after death

of the host. Thus, it is possible to assess their motility directly in cytological specimens from nasal and bronchial brushings and to perform detailed photometry and determine ciliary beat frequency in epithelial specimens sampled by biopsy. Cilial structure can also be assessed by electron microscopy. In a simple and practical clinical test, the time taken for saccharine placed in the anterior nares to cause a sweet taste in the mouth (around 11 min normally) can be used as a convenient clinical measure of ciliary function, which is informative because in most examples of ciliary disease the nasal cilia are also affected. Other more complex methods of assessing mucociliary clearance *in vivo* include cinebronchography and assessment of the rate of clearance of radio-aerosols.

Mucus

Mucus is secreted by the goblet cells and submucosal glands of the first few bronchial generations. Secretion is under the control of a variety of chemical mediators: neuropeptides including substance P, vasoactive intestinal polypeptide, and bombesin, in addition to vagal stimulation and acetylcholine, will cause discharge. In health, mucus is composed of 95 per cent water, the mucus glycoproteins or mucins, and a variety of other proteins (see below), which although present in low concentration, probably play an important part in defence of the bronchial tree. The function of mucus is to trap and clear particles, to dilute noxious influences, to lubricate the airways, and to humidify respired air. The viscoelastic or rheological properties of mucus are likely to be controlled by the concentration of different mucins and are probably critical in determining adequate mucociliary transport.

Factors that affect mucociliary clearance

A number of external factors may reduce mucociliary clearance by interfering with ciliary function or by causing direct ciliary damage. These include pollutants, cigarette smoke, local and general anaesthetic agents, bacterial products, and viral infection. In severe asthma it is thought that eosinophil products including major basic protein may have detrimental effects on ciliary function. Thus there are a number of diseases in which mucociliary clearance may be adversely affected in a secondary fashion. There is also an autosomal recessive condition (occurring with a frequency of about 1 in 30 000 population), called primary ciliary dyskinesia, in which defects in cilial dynein may be associated with male infertility and situs inversus (Kartagener's syndrome). Primary ciliary dyskinesia is associated with repeated sinusitis and respiratory infections that often progress to persistent lung suppuration and severe bronchiectasis, thus underlining the importance of cilia in antibacterial lung defences.

There is now a great deal of interest in abnormal properties of mucus and deranged mucociliary clearance in cystic fibrosis. In this condition, mucus is abnormally viscous with grossly altered rheological properties resulting in markedly retarded mucociliary clearance (see Chapter 17.10).

Surfactant and surfactant 'collectins'

As is discussed in more detail elsewhere (Chapter 17.1.2), surfactant is a complex surface-active material that lines the alveolar surface to reduce surface tension and prevent the lung from collapsing at resting transpulmonary pressures. It also provides a simple mechanism for alveolar clearance since, at end expiration, surface tension decreases and the surface film moves from the alveolus towards the bronchioles, thus carrying small particles and damaged cells towards the mucociliary transport system.

Surfactant is synthesized and secreted by the alveolar type II pneumocytes. It comprises phospholipids, neutral lipids, and at least four different specific proteins, termed surfactant proteins A, B, C, and D (**SP-A**, **SP-B**, **SP-C**, and **SP-D**). It is now recognized that in addition to promoting the surface-active properties of surfactant these proteins have important roles in host defence. SP-B and SP-C are likely to have the major surface-active roles, since both accelerate the adsorption of lipids to an air–liquid interface. Although SP-A can act co-operatively with SP-B in the formation of a surface film, there is no significant derangement of surfactant function or metabolism in SP-A gene-targeted ('knockout') mice, whereas genetically engineered SP-D deficient mice demonstrate abnormal accumulation of surfactant, suggesting a previously unsuspected role for SP-D in surfactant homeostasis.

While it has been recognized for decades that, in addition to its surface-active properties, surfactant can modulate host defence responses, interest in this area has recently intensified with the discovery that SP-A and SP-D, which are secreted by Clara cells in the epithelial lining of the lungs, are members of the collectin (collagen-like lectins) family of proteins. Members of this family in the serum are known to possess important bacterial binding and opsonization properties, and also to influence inflammatory cell function, including stimulation of chemotaxis and the production of reactive oxygen species and cytokines by immune cells. Moreover, mutations of the serum collectin mannose-binding lectin are linked to repeated infections in neonates and children. The collectin family now includes mannose-binding lectin, conglutinin, and CL-44 in blood, and SP-A and SP-D in lung secretions. The collectins are also known as the group II C-type lectins, since their binding to carbohydrate requires calcium.

SP-A and SP-D have been shown to bind to a wide variety of pathogenic microbes (Table 1) and *in vitro* will stimulate the uptake of many of these by macrophages. More recently SP-A, which has been better studied than SP-D, has been shown to possess a number of other properties to suggest that it represents an important component of the lungs armamentarium in the first line of defence against infection (Table 2). Gene-targeted mice deficient in SP-A show marked reduction in the ability to clear *Streptococcus pneumoniae*, *Pseudomonas aeruginosa*, and respiratory syncitial virus from their lungs. The relevance of these observations to human disease remains to be clearly defined.

It has long been recognized that surfactant lipids suppress a variety of immune cell functions, including lymphocyte proliferation, whereas SP-A (and probably SP-D) mainly enhance immune cell functions, suggesting the intriguing possibility of counter-regulatory effects of changes in lipid/protein ratios that might be important in regulating the immune status of the lung. Levels of SP-A (and SP-D) can vary greatly in human disease, with reduced concentrations in HIV-positive patients and markedly reduced concentrations in adult respriratory distress syndrome (**ARDS**) and severe pneumonias. While it is tempting to speculate that low levels of SP-A predispose to infection and immune dysfunction in these settings, it is not certain whether these changes are 'cause' or 'effect' (for example simply reflecting local consumption). There has been less study of SP-D, but it may well possess many of the immune modulatory functions as well as the bacterial-binding properties of SP-A; studies in SP-D ' knockout mice' suggest an important role in influenza neutralization *in vivo*.

Other protective proteins of the lining fluid of the respiratory tract

These may be derived from plasma (e.g. albumin, transferrin, a_2-antiplasmin, a_2-macrogrobulin), by secretion from local epithelial cells, macrophages, or inflammatory cells (e.g. lysozyme, lactoferrin, and defensins), or by selective epithelial transport (e.g. IgA). Clearly the local availability of plasma-derived proteins increases greatly during the exudative phase of any inflammatory process, thus adding more complement, antiproteinases, immunoglobulins, and proteins, including cytokines derived from inflammatory cell secretion.

Defensins and other antibacterial proteins and peptides

Some of the principal antimicrobial molecules isolated from human pulmonary secretions are outlined in Table 3.

Lactoferrin

Lactoferrin was first recognized as a high-affinity iron chelator in human milk, but in the 1960s it was found to possess bacteriostatic properties that were thought to be exerted by deprivation of iron, which is essential for bacterial growth. It occurs in the primary granules of neutrophils and mammalian exocrine secretions, including lung fluids. More recently it has been shown to exert direct membrane effects on some bacteria and to have important modulatory effects on the inflammatory process. Some of these effects may be mediated by its high-affinity binding to bacterial lipopolysaccharide and include the inhibition of cytokine release from cells of the monocyte/macrophage series and modulation of the proliferation and differentiation of immune cells.

Lysozyme

Lysozyme is a 1,4-b-N-acetylmuraminidase that enzymatically degrades a glycosidic linkage of bacterial membrane peptidoglycan. Acting alone, human lysozyme can lyse and kill a variety of Gram-positive micro-organisms, but most Gram-negative organisms are resistant to its direct effects. Like lactoferrin, lysozyme is a major component of the specific granules of neutrophils and is found in the mucosal secretions of the respiratory tract; it has recently been suggested that they may collaborate in killing Gram-negative bacteria such as *Escherichia coli* by disrupting the cell membrane.

Defensins

Defensins are small-molecular-weight, cationic, cysteine-rich peptides that are able to kill a wide range of micro-organisms. They are membrane active and are believed to aggregate in order to 'punch' pores or channels in microbial cell membranes. The a-defensins are major constituents of the neutrophil a granule and are also found in airway secretions. In humans, the a-defensins HNP 1 to 4 account for about 5 per cent of the total cellular protein of neutrophils and are active against staphylococci, *E. coli*, *Pseudomonas aeruginosa*, *Cryptococcus neoformans*, and some enveloped viruses such as herpes simplex virus 1 and vesicular stomatitis virus. The b-defensins are derived from the epithelial lining and were first discovered in cattle as antimicrobial peptides of airway cells, for instance tracheal antimicrobial peptide. At least three human b-defensins have been identified. HBD-1 and HBD-2 are expressed in airway epithelial cells from patients with and without cystic fibrosis. HBD-2 protein is found in lung secretions from patients with cystic fibrosis, patients with inflammatory lung disease, and also from healthy volunteers, whereas HBD-1 protein is not detected in healthy controls but is found in lung secretions from patients with cystic fibrosis and inflammatory lung diseases. Together with the observation that the pro-inflammatory mediator interleukin 1b stimulates epithelial generation of HBD-2 but not HBD-1 *in vitro*, these observations suggest that in the lung HBD-2 is induced by inflammation whereas HBD-1 may serve as a lung defence in the absence of lung inflammation. HBD-1 has also recently been implicated in the athogenesis of cystic fibrosis where it appears to be inactivated by the high salt milieu.

Cathelicidins

Cathelicidin peptides contain a highly conserved signal sequence ('cathelin') but show substantial heterogeneity in the C-terminal domain. The first human cathelicidin to be characterized—LL-37/hCAP-18—is expressed in airway epithelium and displays antibiotic activity against a range of Gram-negative and -positive organisms.

Small-molecular-weight antiproteinases

Agents such as secretory leukoproteinase inhibitor and elafin, previously thought to exist in lung secretions solely as part of the 'antiproteinase protective shield', have now been shown to possess important additional antibacterial properties. Both secretory leukoproteinase inhibitor and elafin display major activity against *Pseudomonas aeruginosa* and *Staphylococcus aureus*, which resides in a molecular position distinct from their powerful antineutrophil elastase activities.

In summary, a growing number of antibiotic peptides are now shown to demonstrate a marked selectivity for prokaryotic and eukaryotic micro-organisms, thus providing an effective but simple method whereby diverse classes of micro-organisms can be recognized and destroyed as 'non-self'. Many more peptides and proteins like these are likely to emerge, displaying properties that combine antibacterial actions with important roles in modulating the key cells responsible for our host defences, perhaps providing important links between primary 'innate' defence mechanism and the generation of secondary inflammatory and immune responses. Some of these agents may be of value in future therapeutic regimes, particularly in situations where the lungs are injured and at risk of secondary infection, such as cystic fibrosis and the adult respiratory distress syndrome.

Immunoglobulins

Normal lung secretions contain all the immunoglobulins present in plasma, but in different proportions. In the absence of disease, immunoglobulins are produced locally, with IgA greatly in excess and only small contributions from IgG and IgM. It is thought that B lymphocytes and plasma cells are particularly important in producing secretory IgA in the upper airways by a collaborative mechanism involving the epithelial cells as follows. Dimeric IgA is assembled in the plasma cells from two monomeric IgA molecules and joined by another protein, the J chain. Dimeric IgA then binds to the secretory component on the surface of epithelial cells, forming a dimeric IgA–secretory component complex that is pinocytosed, transported through the epithelial cell, and released from its luminal surface into the airways. The secretory component appears to protect IgA from enzymic attack during bacterial infection and inflammation in the host. IgA is produced in very high concentrations in the upper airways and is therefore likely to serve a number of important roles, but these are not fully understood. IgA deficiency is associated with local defects in immunity.

IgG concentrations in lung secretions are quantitatively similar to plasma IgG concentrations and may be particularly important in the lower airways where IgG may act as a very effective opsonin and activator of complement. IgG deficiency is associated with recurrent respiratory tract infection, suggesting that it provides an important local defence mechanism.

Complement proteins, proteinase inhibitors, etc.

Most of the proteins involved in the complement system have been identified in lung secretions. Most are probably derived by plasma exudation during inflammation, and C3a, C3b, and C5a may be secreted by alveolar macrophages. Patients with C3 deficiency have recurrent upper and lower respiratory tract infections, particularly with *Streptococcus pneumoniae* and *Haemophilus influenzae*. C3 is likely to play a key role in opsonization (via C3bi) of bacteria.

Lung secretions also contain a variety of antiproteinases, including the large molecules a $_2$-macroglobulin and a$_1$-proteinase inhibitor as well as moieties of lower molecular weight, such as secretory leukoproteinase inhibitor and elafin. Antiproteinases are probably secreted in higher concentrations by local epithelial cells and alveolar macrophages during inflammatory and injurious processes, at which time there will be additional contribution to local defences from leakage of plasma protein-derived antiproteinases. It is likely that these antiproteinases play an important part in the antiproteinase 'shield' which is necessary to protect the healthy local tissues against damage from the release of proteinases by inflammatory cells.

The alveolar macrophage and other alveolar cells

Alveolar macrophages are highly differentiated cells that have matured in the lung from bloodborne, bone marrow-derived monocytes. They normally 'patrol' the alveoli (Fig. 2), where they exist with a half-life of several weeks. The technique of bronchoalveolar lavage, whereby fluid is instilled into the small airways via a fibreoptic bronchoscope and fluid and harvested cells (normally greater than 95 per cent alveolar macrophages in healthy individuals) are returned by suction, has greatly facilitated the *ex vivo* study of the various functions of these versatile cells. It was quickly recognized that alveolar macrophages possessed marked phagocytic ability, with the capacity to ingest and destroy pathogenic bacteria, but only recently has their capacity to generate mediators of central importance in the initiation of inflammation and to present antigen in the initiation of the immune response been fully recognized. The alveolar macrophage could therefore be considered as a 'microcomputer', sampling and sensing via a vast array of receptors (Table 4) the external environment in the alveolar spaces and subsequently determining whether inflammatory or immune responses should be generated. It is also likely to assist the inflammatory monocyte-derived macrophages in the scavenging roles required during the aftermath of infections and the resolution of inflammation, and may play a further important role in the processes whereby inflammatory tissue injury is repaired, since it can produce a number of proteins involved in tissue repair processes and can generate a variety of cytokines that influence fibroblast function (Table 5).

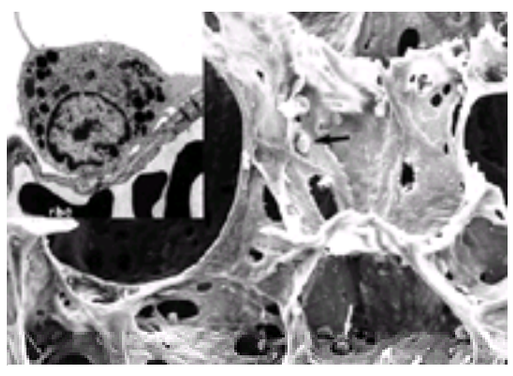

Fig. 2 Scanning electron micrograph (× 350) showing alveolar macrophages (arrows) 'patrolling' the alveolar airspaces. A high power view of a single alveolar macrophage adhering to the alveolar lining (arrows illustrate points of adhesion) is shown in section on the inserted (top left) transmission electron micrograph. (By courtesy of Dr P.K. Jeffery.)

Phagocytosis and bacterial killing

Macrophages can recognize and ingest opsonized (via their surface CR3 or FcR) or non-opsonized particles by a variety of receptors (Table 4). Within the phagolysosome, ingested particles are subjected to the combined destructive forces of reactive oxygen intermediates generated via the metabolic burst and a wide range of degradative enzymes that have the capacity to digest proteins, lipids, and carbohydrates. It appears that the local generation of nitric oxide is an important defence mechanism against a variety of micro-organisms. Activated macrophages form nitrite (NO_2), nitrate (NO_3), and nitric oxide (NO). *In vitro* experiments suggest that these products, particularly NO and the peroxynitrite anion, contribute to antifungal, antiparasitic, and tumoricidal activity of macrophages. Macrophages may 'call in antibacterial reinforcements' of other phagocytic cells including neutrophils, monocytes (which mature into inflammatory macrophages), and eosinophils by the generation of specific chemotaxins (see below). They may also generate a local immune response by presenting antigen and producing a variety of lymphokines.

Despite the availability of such powerful mechanisms, it is clear that not all phagocytosed particles are effectively destroyed. For example, asbestos, silica, and a number of micro-organisms including tuberculosis, some strains of *H. influenzae*, and trypanosomes at various stages of their lifecycle are able to resist destruction within macrophages.

Generation of the inflammatory response

Macrophages can secrete a number of chemotactic proteins in the chemokine family as well as mediators in the 5-lipoxygenase and cyclo-oxygenase pathways, all of which can exert profound pro-inflammatory effects. Neutrophil chemotaxins include interleukin 8, leukotriene B_4, and NAP-2. Peptides chemotactic for monocytes include MCP-1 and MIP-1. Other macrophage-derived cytokines may have important secondary pro-inflammatory effects through their influences on other cells. For example, tumour necrosis factor and interleukin 1 act not only on endothelium to stimulate the expression of adhesion molecules necessary for inflammatory cell emigration, but may also act on local fibroblasts to produce interleukin 8 that exerts neutrophil chemotactic effects. Thus, macrophages not only generate chemoattractants for inflammatory cells, but they can also recruit other local cells such as fibroblasts to help in the initiation of inflammation, thereby governing graded levels of amplification of the inflammatory response.

Generation of the immune response

Alveolar macrophages are effective antigen-presenting cells and can display partially degraded antigens on their surface to interact with recirculating T and B cells, generating clonal expansion and initiating the immune response.

Tissue remodelling and repair

Alveolar macrophages can secrete proteins, including fibronectin, vitronectin, and laminin, which are important in tissue repair. Macrophages can also produce a number of cytokines including PDGF, TGF-b, and interleukin-1 that can influence the behaviour of other cells, particularly fibroblasts that are critically involved in the repair process (see Chapter 16.5.1).

The pulmonary marginated pool of neutrophils

The neutrophil is the archetypal acute inflammatory cell, equipped with a variety of mechanisms that make it a very effective agent in host defences against bacteria such as streptococci. After release from the bone marrow, mature neutrophils exist in the vascular compartment with a half-life of about 6 h. Unlike red blood cells, up to half of the neutrophils in the vascular compartment do not circulate at any given time, but form a 'marginated pool', which is in dynamic equilibrium with the 'circulating pool' of vascular neutrophils. The marginated pool can be released into the circulating pool by exercise or adrenaline. The vascular beds of the lung and spleen appear to make the most important contribution to the marginated pool, which may serve as a source of rapidly releasable neutrophils in times of stress or injury. The presence of large numbers of neutrophils in the pulmonary microvascular bed is likely to increase the mobilization and effectiveness of local lung defences in response to their inevitable exposure to inhaled micro-organisms or toxins. The mechanisms underlying the formation of the lung marginated pool are uncertain. It could be formed as the result of low-grade adhesive interactions between neutrophils and lung capillary endothelial cells, but it is likely that the rheological properties of neutrophils in pulmonary capillaries are more important in the physiological margination of neutrophils in the lung. The mean diameter of the pulmonary capillary is 5.5 µm, whereas that of the neutrophil is 7.5 µm, hence neutrophils are normally required to squeeze through the pulmonary capillaries and minor changes in their deformability or alterations in the fluid pressure gradient across the lung capillary bed would be expected to influence markedly the size of the pulmonary marginated pool. The presence of this pool of neutrophils in the lung may have advantages for host defence, but, paradoxically, it could also partly explain why the lung appears to be such an important target in conditions such as the adult respiratory distress syndrome, which may result from systemic or distant insults, such as Gram-negative septicaemia, multiple trauma, or pancreatitis (see Chapter 16.5.1).

Further reading

Clark HW, Reid KBM, Sim RB (2000). Collective and innate immunity in the lung. *Microbes and Infection* **2**, 273–8.

Ganz T *et al.* (1985). Defensins. Natural peptide antibiotics of human neutrophils. *Journal of Clinical Investigation* **78**, 1427–35.

Hancock RE (1997). Peptide antibiotics. *Lancet* **349**, 418–22.

Singh PK *et al*. (1998). Production of b-defensins by human airway epithelia. *Proceedings of the National Academy of Sciences, USA* **95**, 14961–6.

Sleigh MA, Blake JR, Liron H (1988). The propulsion of mucus by cilia. *American Review of Respiratory Disease* **137**, 726–41.

Van Wetering S *et al.* (1999). Defensins: key players or bystanders in infection, injury, and repair in the lung? *Journal of Allergy and Clinical Immunology* **104**, 1131–8.

Wright JR (1997). Immunomodulatory functions of surfactant. *Physiological Reviews* **77**, 931–62.

17.2 The clinical presentation of chest diseases

D. J. Lane

The presenting symptoms of chest diseases are few, but the structural and functional disturbances that these symptoms reflect are numerous and the underlying disease entities are many. The symptoms of lower respiratory tract disease can be grouped under just three headings: cough, breathlessness, and chest pain.

Cough may or may not produce sputum. Patients occasionally report the expectoration of sputum while denying they have a cough. This seems to be a socially determined separation of the act of 'clearing the throat' to expel sputum, from a non-productive cough which, perhaps because it appears to have no purpose, is regarded as more sinister. Breathlessness itself is a complex symptom. Wheezing and stridor, which are audible accompaniments to the act of breathing, are rarely reported without breathlessness and so they will all be considered together. Discussion of chest pain will include mention of chest tightness.

In the analysis of symptoms it is important to recognize and differentiate between the pathology or disordered physiology likely to be responsible for the symptoms, and the clinical diagnoses associated with that symptom. The investigation of mechanism, although superficially of little clinical relevance, can be the key to symptomatic treatment, creating opportunities for relief when the underlying condition is untreatable. Knowledge of the clinical significance of symptoms is largely empirical, but forms the essential diagnostic base of clinical medicine. Thus research into the mechanisms of breathlessness, for example, will continue to be a proper concern of clinicians as long as disabling and irreversible conditions such as chronic airways obstruction exist. By contrast, knowing the mechanism of dyspnoea in pleural effusion is unimportant compared with knowing how to relieve the symptom by draining the effusion, and being able to diagnose the underlying clinical condition.

Cough

Coughing is a defensive reflex designed to clear and protect the lower respiratory tract. The act of coughing is essentially a forced expiratory effort against a transitorily closed glottis, which then opens allowing a sudden expulsion of air from the lungs. Except when the cough arises from laryngeal irritation, there is an initial deep inspiration which allows the respiratory muscles to act to greater mechanical advantage, although this could draw any offending material deeper into the bronchial tree. The pressure that builds up behind the closed glottis can reach as much as 40 kPa and, if often repeated in a sequence of coughs, can seriously impede venous filling of the heart. The consequent drop in cardiac output is responsible for the well-described 'cough syncope'.

Mechanism

The cough reflex can be initiated by the stimulation of irritant receptors in the pharynx, larynx, trachea, and major bronchi. These receptors respond to mechanical irritation by mucus, dust, or foreign bodies, and to chemical irritation by fumes and toxic gases. Mechanical events within the thorax, such as sudden and large changes in airway calibre or lung collapse, can also stimulate cough receptors. The afferent fibres run in the branches of the superior laryngeal nerve and the vagus to the medulla, where the resultant efferent activity of virtually the whole of the respiratory musculature is co-ordinated. The explosive action of the respiratory muscles produces laryngeal air velocities that can approach the speed of sound and is accompanied by bronchial constriction, mucus secretion, and a transient systemic hypertension.

Causes of cough

An epidemiological analysis of cough would reveal an acute, viral, upper respiratory tract infection affecting the pharynx, larynx, or postnasal space as the most common cause of short-lived cough at all ages, with smoking being the main cause of chronic cough in adults. Half of those smoking 20 or more cigarettes a day can expect to have a persistent cough. Children exposed to passive smoking from their parents are twice as likely to cough as children in non-smoking families. Asthma is the next commonest cause at all ages, with chronic upper or lower respiratory tract infection also important. Tuberculosis heads the list in the developing world. Of the more sinister causes, carcinoma of the lung is the most important. It remains the commonest neoplasm in men and is second only to carcinoma of the breast in women, and must feature high in the differential diagnosis of a new presentation of cough or a change in the character of cough in a middle-aged smoker. Less usual causes are endobronchial sarcoidosis and pulmonary fibrotic conditions. Both beta-blockers and angiotensin-converting enzyme (**ACE**) inhibitors can cause an irritating and persistent cough. Although rare, an inhaled foreign body must not be forgotten.

Clinical features

The clinical description of cough relies on its sound, its timing, and whether or not there is expectoration. A dry cough with an irritative barking quality, short and often repeated, is heard in pharyngitis, tracheobronchitis, and early pneumonia. With laryngitis the sound is harsh and hoarse ('croup'). The long inspiratory sound that gives whooping cough its name is also produced by tracheal and laryngeal inflammation. Abductor paralysis of the vocal cords creates a cough that is prolonged and lowing like the sound of cattle, and hence is described as 'bovine'. The usual cause is pressure on the left recurrent laryngeal nerve by lesions in the thorax: carcinoma of the bronchus or oesophagus, enlarged (usually neoplastic) hilar nodes, or (now very rarely) aortic aneurysm. If similar lesions press on the trachea but spare the nerve, the cough has a hard metallic quality described as 'brassy'. Unilateral abductor palsy of the larynx does not affect the voice, and even with additional abductor palsy the voice often remains good. Complete paralysis of both cords gives aphonia and a weak ineffectual cough. Weakness of the thoracic muscles, as in polyneuritis or the muscular dystrophies, will lessen the expulsive force in coughing, as will the general weakness of prostration, toxaemia, or the deeper states of unconsciousness. Cough may be suppressed when there is severe thoracic or upper abdominal pain.

Certain aspects of the timing of coughing may give useful diagnostic clues. A cough that awakens the patient in the small hours of the night suggests asthma; wheezing need not be evident. Cough with expectoration on rising in the morning is characteristic of chronic bronchitis, although it may also be reported by

asthmatics. A bout of coughing with food or when lying down after a meal points to oesophageal, pharyngeal, or neuromuscular disease, causing aspiration into the lungs. Changes of posture can also set off coughing in the bronchiectatic; and free expectoration of sputum at any time of day is common in these patients. A dry cough that persists over many weeks can signify a neoplasm, but a non-productive barking cough that has lasted for years is more likely to be a nervous habit often perpetuated by psychogenic factors.

A cough may fail to produce expectoration because there is nothing to produce, because secretions are swallowed (as is almost universal in children), because there is severe airways obstruction, because of weakness (as outlined above), or because the secretions are too viscid. In the last four instances the sound quality of the cough differs from that of a dry cough, in the sense that secretions can be heard moving in the major airways. This type of cough and the cough productive of sputum can be described as 'moist' or 'loose'.

Phlegm and sputum

Phlegm, the secretions of the lower respiratory tract, is admixed with nasal and pharyngeal secretions as well as saliva to give expectorated sputum. It has been very difficult to study the natural secretions of the healthy tracheobronchial tree in man, for only about 100 ml is produced daily and most of this is swallowed. In disease the quantity of secretions is often sufficient to swamp contamination from the upper respiratory tract, so that valid observations can be made, but it may be necessary to obtain lung secretions by induced coughing or by bronchoscopy.

Mucus is viscoelastic. Its viscosity or stickiness influences the effect of forces applied to it in coughing. Initially it resists flow, but then as increasing force is applied it becomes more and more liquid, returning to its original state when the flow stops. The elasticity of sputum appears to alter with the rate of application of stress to it, and this may be important in relation to the rate of beating of the bronchial epithelial cilia. Intrabronchial mucus exists in two layers: one of low viscosity and high elasticity touching the cilia, and above this a more viscous layer which, in disease, carries globules of mucus.

Airway mucus is 95 per cent water and derives its distinctive physical characteristics from the glycoprotein content. Two components of these glycoproteins, sialic acid and sulphate, enable airway mucus to be chemically analysed and identified *in situ* in histological sections. At least four glycoproteins have been identified in human bronchial mucus, produced in various combinations by different mucous-cell types. Serous fluid is produced from other cells in the bronchial glands, and with water, lipids, and proteins makes up a transudate component. Although bronchial secretions do not show diagnostically distinctive changes in disease, there is, for example, a shift towards greater glycoprotein production in chronic bronchitis and greater transudate formation in asthma. In infection both components increase, and the breakdown of leucocytes and of bronchial mucus increases the DNA content of sputum, making it less viscid. The accumulated debris of cells and micro-organisms imparts a yellow colour to infected sputum, and the subsequent action of verdoperoxidase derived from leucocytes gives a green colour.

The distinction between infected and non-infected is one of the most obvious descriptive features of sputum that is relevant in clinical medicine. Non-infected mucoid sputum is variously described as clear, white, or like jelly. Viscid mucoid sputum is sometimes seen in asthma, and the patient may report seeing pellets or even branching plugs of mucus that are presumed to be casts of small bronchi. In bronchopulmonary aspergillosis, similar pellets or casts have a dark brown colour. In city dwellers, and those in dusty occupations, mucoid sputum can be various shades of grey. Coal miners may produce jet-black sputum (melanoptysis) if an area of fibrosis breaks down and is expectorated.

In most lower respiratory tract infections pus is admixed with mucus to produce mucopurulent sputum. Pure pus can be expectorated from a lung abscess or from stagnant bronchiectatic cavities. An offensive smell to the sputum, particularly in these last two conditions, often comes from infection with anaerobic organisms. A rarely seen but distinctive brown discoloration ('anchovy sauce') is seen with pus from an amoebic lung abscess, usually secondary to hepatic amoebiasis.

Apart from its appearance, the only other macroscopic attribute to sputum is its quantity. Excessively large quantities of sputum are found in bronchiectasis, particularly where this is widespread, as in cystic fibrosis, and in alveolar-cell carcinoma where large quantities of watery mucus can occasionally be produced. The amount of sputum in both chronic bronchitis and asthma is very variable, but can be excessive. Briefly, severe pulmonary oedema leads to the production of a large quantity of frothy sputum.

Haemoptysis

Patients rightly regard the presence of blood in the sputum as sinister. Despite this, a definite cause of haemoptysis is only found in about half of cases in most series. In the assessment of haemoptysis it is important to establish first that the blood-stained material has come from the chest and not from the gastrointestinal tract. Some patients find this difficult. Haemoptysis is produced with a 'cough' not a 'retch'. Accompanying features of an appropriate disease are usually present, but it is worth remembering that in haemoptysis there is usually froth due to admixed air, and the blood is bright red, not dark brown. Gastric contents should be acid; bronchial contents should be alkaline. Another trap for the unwary is contamination with blood from the nose or upper respiratory tract.

It is unwise to attribute haemoptysis simply to 'bronchitis' or infection. In bronchiectasis, however, haemoptysis not uncommonly mixes with mucopurulent sputum. In the early stages of pneumococcal pneumonia a 'rusty' staining of mucoid sputum is quite characteristic. In tuberculosis, frank blood in otherwise mucoid sputum is well recognized. Sudden haemoptysis is a hallmark of pulmonary embolism with infarction. In bronchial neoplasia there may be streaking of the sputum with blood or more substantial bleeding with clots, often observed daily. Recurrent blood-staining of the sputum is seen in idiopathic pulmonary haemosiderosis and also, although usually over a shorter time span, in Goodpasture's syndrome, which are both uncommon conditions. Cardiac conditions associated with blood in the sputum are pulmonary oedema, with pink frothy sputum, and mitral stenosis. The recurrent haemoptyses of the latter condition are infrequently seen today. In a general context, it may be necessary to consider thoracic trauma, endometriosis, or a blood coagulation disorder as causes of haemoptysis.

In the investigation of haemoptysis the chest radiograph will often indicate a probable diagnosis, for example an apical tuberculous infiltrate or a neoplastic hilar mass, but this must be backed up by appropriate microbiological or cytological examination of the sputum. Old, presumably healed and calcified, tuberculous lesions may be reactivated and may be a sufficient cause for haemoptysis simply due to local bronchiectasis, and invasion by mycetoma must be considered. Bronchiectasis and pulmonary infarction may not be evident on a plain radiograph, but both should give a suggestive history. High-resolution computed tomography (HRCT) will diagnose bronchiectasis, and ventilation–perfusion scanning or spiral CT will diagnose a pulmonary embolism.

If examination of the sputum and radiology (plain or specialized) yields no obvious cause for haemoptysis, then bronchoscopy must be considered. After a single haemoptysis in a young person with a normal chest radiograph, this can be deferred for a month. A recurrence of haemoptysis or a single episode in an older person, particularly a smoker, are indications for early bronchoscopy.

Laboratory examination of sputum

Expectorated sputum should be subjected to microscopic and microbiological investigation as appropriate (see Chapter 17.3.3). Sputum eosinophilia is a good guide to airway allergy. The cytological examination of sputum for malignant cells can only be done by an expert, but in skilled hands it is invaluable and time-saving.

Investigating cough

The cause of a cough of recent onset will usually be obvious enough. Infection tops the list. Carcinoma must be excluded in the smoker and radiology will detect parenchymal disease. In the case of persistent cough without apparent cause in a non-smoker, occult asthma should first be eliminated. Assessment of airflow variability using either diurnal peak flows or a histamine reactivity test are first-line investigations (see Chapter 17.4.1.1). It is worth looking for a chronic sinonasal infection or allergy and then for gastro-oesophageal reflux before proceeding to bronchoscopy, which statistically is not very rewarding in the context of cough as a lone symptom.

Treatment

Once diagnosed, the cause may well be treatable, even if the only appropriate advice is to stop smoking. If the condition is not treatable or if no cause can be found, symptomatic measures need to be considered. Two lines of approach are open: to suppress the cough or, accepting the cough as inevitable, to make expectoration easier.

All cough suppressants in common use act centrally. Most are opiate derivatives. Codeine and pholcodine have a weak antitussive action, but when made into a

sweet syrup they seem to have a soothing effect. Methadone and the stronger opiates are more powerful in suppressing cough, but they depress respiration and also cause constipation. In terminal bronchial carcinoma they are invaluable. Attempts to suppress cough by a peripheral action on bronchial afferent receptors have not been successful. Inhaled local anaesthetic can be helpful, and its effect on cough may long outlast its anaesthetic action. Drugs that act on the production of bronchial mucus will lessen cough if its purpose is the expectoration of that mucus. Atropine is used to this end preoperatively, but rarely in disease. Corticosteroids can diminish mucus production in asthma and in alveolar-cell carcinoma.

Agents claimed to increase sputum quantity or accelerate expectoration include the volatile oils such as menthol and inorganic salts such as potassium iodide. The movement of particles up the mucociliary escalator of the bronchial tree has been charted using radioisotope techniques and shown to increase under the influence of ingested guaifenesin (present in several 'cough medicines'), inhaled beta-adrenergic agonists, and hypertonic (1.2 M) saline.

Attempts to decrease the viscosity of sputum using mucolytic agents have been clinically disappointing, despite definite *in vitro* evidence of activity. Inhaled acetylcysteine works more convincingly as a mucolytic but has the great disadvantage of inducing bronchoconstriction. In cystic fibrosis, DNase has been used with modest success.

Most patients with haemoptysis require no more than treatment appropriate to their underlying condition, but occasionally haemoptysis is massive and life-threatening. The recorded mortality of 50 per cent with haemoptysis of 200 ml or more includes patients with initially poor respiratory reserve, as well as those who asphyxiate. At bronchoscopy it may be difficult to locate the source of bleeding, but local endoscopic measures may be applicable (topical epinephrine (adrenaline) application, balloon tamponade, and cold saline lavage). An open surgical approach (lobectomy or pneumonectomy) carries a mortality of up to one-third, and if operative intervention is contemplated, bronchial arteriography should be considered. The source of bleeding is from the bronchial arteries, so that embolization of the appropriate bronchial artery has successfully been used to control massive haemoptysis in a high proportion of patients who are actively bleeding.

Breathlessness

This major symptom of pulmonary, cardiovascular, and other systemic diseases suffers much because it is so frequently referred to by physicians as dyspnoea. Whilst patients sometimes speak of difficulty in breathing, they more frequently use the terms 'breathlessness', 'short of breath', or 'out of breath'. It is usually only on direct questioning that specific features reveal clues that are likely to be useful clinically. Despite the often quoted statement of Comroe that dyspnoea is not tachypnoea, hyperpnoea, or hyperventilation, but difficult, laboured, or uncomfortable breathing, patients are quite unaware of these fine distinctions. Rapid breathing, the necessary increase in ventilation in response to exercise, and ventilation in excess of metabolic requirements are all at times described by patients as breathlessness. Just what degree or quality of awareness of respiratory movement deserves to be called breathlessness is probably indefinable; awareness undoubtedly varies from patient to patient and even within the same subject from time to time. However, the implication of most terms used by patients to describe this type of pulmonary sensation is that in some way the performance of the respiratory apparatus ('breath-') is not meeting ('-less') a demand placed on it.

Pathophysiology

The respiratory muscles are supplied by motor nerve fibres from cervical and thoracic anterior horn cells, from C3 to T12. Like all other anterior horn cells, the respiratory motor nerve cells are served by pyramidal fibres from the motor cortex in the precentral gyrus. Directives from the cortex enable respiratory movement to be modulated to serve such functions as talking, holding the breath, voluntary hyperventilation, and the performance of lung function tests. This pathway will also be responsible for the conscious, and perhaps unconscious, transmission of anxiety or a calming influence on respiratory performance. However, to an extent that is unparalleled in other mammalian skeletal muscle, the respiratory motor neurones are under dual control, the second component being the motor output from the brainstem respiratory centre responsible for involuntary or automatic respiratory movement. This is the movement necessary to satisfy metabolic requirements for oxygen supply and carbon dioxide removal.

For the purposes of understanding breathlessness, respiratory centre activity can be seen as being under the influence of chemical and neurogenic stimuli. The chemical stimuli of hypoxia and acidaemia are relevant to the breathlessness of high altitude and diabetic coma. The general traffic of neurogenic stimuli impinging on the reticular formation from all sources maintains a certain level of activity in the medullary respiratory neurones irrespective of more specific stimuli. The modest quietening of this activity in sleep is associated with a small drop in minute ventilation, and the dramatic curtailment of spinal ascending information that sometimes occurs following high spinal tractotomy (usually for intractable pain) can completely abolish automatic medullary respiratory activity. There is no clear-cut association between increased reticular formation activity causing hyperventilation and states of breathlessness, but the increase in ventilation at the very onset of exercise is thought to be neurogenic in origin, possibly originating from the exercising muscles.

The respiratory centre receives information from the lungs through the vagus nerve. This originates in bronchial epithelial irritant receptors, stretch receptors, and interstitial J receptors within the alveolar/capillary interface. Stimulation of all these receptors will produce reflex effects, amongst which (for example) tachypnoea will make up a component of breathlessness in an appropriate setting. Whether the afferent information travelling up the vagus itself reaches the sensorium, or whether it merely modulates some other afferent pathway, is not clear. Afferent information that undoubtedly reaches consciousness is that concerning the rate and degree of thoracic cage movement and, quite accurately, a sense of lung volume (degree of lung inflation/deflation). This information comes from joint, tendon, and muscle receptors in the chest wall, and for sense of movement (of air) perhaps also from the oropharyngeal mucosa. It seems evident that information from these latter sources is part of natural and healthy sensation; it also seems likely that the same channels will signal an increased rate or depth of movement which, if excessive, will be described as breathlessness. In exercise the description 'breathless' often comes at the point where the smooth linear relation between ventilation and oxygen consumption is disturbed. Ventilation becomes excessive for metabolic requirements. How this becomes described as a shortness or loss of breath is not clear.

Two 'unnatural' respiratory sensations that can only be inadequately mimicked in a healthy individual are those associated with abnormal lung mechanics (for example in airways narrowing) and muscle paralysis. An obvious parallel for the first is breathing through an external resistance, and this technique has been widely used by those investigating dyspnoea. The useful finding that may have some bearing on the clinical situation is that in resistance breathing the ability to detect an increased load depends not on the absolute magnitude of the load, but on the ratio of that load to pre-existing loading of the system. Thus, a given absolute increase in airways resistance will be much more obvious to an individual with near-normal airways function than to one already suffering from pathological airways narrowing. The sensations of those few normal individuals who have undergone muscle paralysis for experimental purposes include phrases such as 'choking' and 'I would give anything to be able to take one deep breath'. These are similar to the reported symptoms of patients with paralytic diseases affecting the respiratory muscles. The element of inadequate performance is stressed. Whether this sensation can be simulated by the voluntary withholding of respiratory movement, as in breath-holding, is very doubtful. This much studied experimental model undoubtedly gives sensations, most of which probably arise from the diaphragm twitching ineffectually, and it is difficult to see where this fits into a clinical setting.

If any common thread can be drawn between these examples, it is at the level of an interaction between the drive to breathing and achievement—a drive that fails to achieve because of poor performance or mechanical loading of the respiratory system. The neurophysiological implications of this hypothesis are that there should be monitoring systems for both drive and performance. It is obviously feasible for the brain to assess and summate the various drives to breathing, and so monitor motor output. The muscle spindle has been proposed as the probable detector of achievement, but several objections exist to this proposal, not least that there are relatively few muscle spindles in the diaphragm, which is the most important muscle of respiration.

When it comes to the application of these principles to disease, some parallels are obvious, but an example from one common condition—acute bronchial asthma—will illustrate that the situation is not straightforward, and frequently multifactorial. In acute asthma there may be excessive drive from bronchial irritant receptors, and often cortical drive expressing itself in anxiety, even panic, as well as hyperventilation. There is clearly poor performance because of the increased resistance of narrowed airways, but in addition it seems likely that the respiratory muscles will act at a mechanical disadvantage because of lung hyperinflation.

The clinical analysis of breathlessness

Whilst bearing these neurophysiological points in mind, when it comes to devising symptomatic measures for the relief of breathlessness the clinician must still largely rely on an empirical approach to the analysis of this symptom. Such an analysis will rely on four characteristics of breathlessness: its quality, its timing, its severity, and the circumstances that precipitate or relieve it.

Quality

Breathlessness is difficult to describe. Most patients can go no further than saying that they are 'short of breath'. The asthmatic will generally recognize the quality of

wheeze and, contrary to the opinion of physiologists, usually finds it more difficult to breathe in than out. An asthmatic who develops more persistent breathlessness between attacks often recognizes this as 'different from my asthma'. A sense of suffocation is a feature of massive pleural effusions and of pulmonary oedema. Phrases such as 'I can't fill my lungs properly' and 'I need to take a big breath' suggest the possibility of psychogenic breathlessness, but muscle weakness must be carefully excluded.

Timing

Of the greatest value in separating out conditions likely to be associated with breathlessness is noting its rate of onset. There are five categories. Breathlessness may be of dramatic onset (over minutes), acute onset (over hours), subacute onset (over weeks), or chronic onset (over months or years), or it may be intermittent. Table 1 gives a guide to conditions falling into these categories. The subdivisions are not rigid. Asthma again provides an example. About half of all acute attacks of asthma build up in less than 24 h, but some asthmatics slowly deteriorate over a week or so and, occasionally, they can be transformed from being asymptomatic to having desperate breathlessness and unconsciousness within 15 min; in addition, asthma is also intermittent. Likewise, left ventricular failure, although usually developing over hours, may be dramatic in, for example, aortic valve rupture, or more persistent in long-standing hypertension. Pulmonary embolism can also present in a very variable manner, ranging from the dramatically sudden to the chronic.

The pattern of breathlessness in certain disorders depends on the stage of the disease and the structural or functional changes that it causes. Thus in early sarcoidosis, diffuse infiltration of the lungs can cause the quite rapid development of breathlessness over a week or so; by contrast, the late fibrotic stage of sarcoid will be associated with relentlessly progressive breathlessness as pulmonary reserve diminishes. Breathlessness in a condition such as carcinoma of the lung will be determined by the pattern of structural change—whether there is, for example, bronchial stenosis, collapse, or pleural effusion.

Severity

The severity of breathlessness is traditionally gauged on scales relating to activity. Many scales have been devised. All have two faults. The first is that there is a temptation to assume that the grading system with which one is familiar is universally known. It is not. Thus it is preferable to describe the amount of exercise limitation on an individual patient basis. Second, no scale suggests a convention for dealing with variable breathlessness. Very few patients have a consistent level of severity of breathlessness, but any attempt to introduce a range of severity will have to be accompanied by some assessment of the time extent of each grade, which is an almost impossible task.

A refinement of the scaling technique can be used to record the degree of breathlessness during exercise. During a standard exercise test a record is made of breathlessness on a visual analogue scale concurrently with minute ventilation. The sort of data produced can be used to assess the benefits of therapeutic intervention. The Borg scale (Table 2) can be similarly adapted and has been used with simple exercise tests. Tests such as the distance walked in a specified time (for example, a 6- or 12-min walk), are influenced by training and motivation. Patients may differ in their approach to a treatment benefit. One might walk no further and be relieved to achieve the same distance with less breathlessness, whereas another might extend the walking distance by being prepared to become just as breathless as before.

Occurrence

The circumstances under which breathlessness is experienced can give important diagnostic clues. Only psychogenic breathlessness bears no relation to exertion or is experienced only at rest. Many patients with organic diseases are breathless at rest as well as on exertion, this being an expression of severity. Breathlessness made worse by lying flat (orthopnoea) is characteristic of left ventricular failure and is also experienced by patients with diaphragmatic paralysis. Nocturnal awakening with suffocating breathlessness and frothy sputum production (paroxysmal nocturnal dyspnoea) is a more serious manifestation of left ventricular failure and can be relieved by sitting or standing up. The asthmatic also awakens in the small hours of the night with breathlessness accompanied by coughing and wheezing, or these symptoms may be delayed until the normal waking hours. Any sputum produced by the asthmatic under these circumstances is likely to be sticky and mucoid. Postexertional breathlessness and the immediate triggering of an episode of wheezing breathlessness by non-specific irritants (dust and fumes) or specific allergic stimuli (pollen, animal danders, etc.) also characterize the asthmatic. In occupational asthmas, breathlessness will bear a temporal and circumstantial relation to the working environment. In byssinosis the first day at work is characteristically troublesome (Monday morning tightness). Patients with type III hypersensitivity reactions such as bronchopulmonary aspergillosis or extrinsic allergic alveolitis (Chapter 17.11.11) will notice breathlessness 4 to 6 h after exposure. An intercurrent respiratory tract infection will worsen breathlessness in patients with any form of diffuse airway or parenchymatous lung disease.

Spontaneous improvement occurs in most breathless patients with rest or the removal of trigger factors. The postexertional breathlessness of the asthmatic is an important, though temporary, exception. Patients with pulmonary hypertension, even with severe exertional breathlessness, improve dramatically quickly as soon as they sit down.

The investigation of the breathless patient

The clinical history may immediately suggest a probable cause. Beyond this the two most helpful pointers are simple lung function tests and chest radiology. Spirometric testing will define three groups: normal, an obstructive pattern, and a restrictive pattern (see Chapter 17.3.2). The chest radiograph will be of most value in furthering the diagnosis in conditions giving a restrictive pattern. The further investigation of the patient with airflow obstruction is dealt with in Section 17.4.

Breathlessness in a patient with normal spirometric testing and a clear chest radiograph presents special problems. In this situation four categories should be considered. Is there intermittent disease? Are the tests being used too crude to pick up significant abnormalities? Is there extrathoracic disease? Is this psychogenic breathlessness?

Many asthmatics reviewed in a clinic will have normal lung function. The value of serial recordings of lung function over several days in these patients cannot be overemphasized. Conditions affecting the heart and pulmonary circulation can also be intermittent, but more often the problem is that conventional tests of lung function do not seem to demonstrate significant abnormalities when there is quite considerable dyspnoea. Pulmonary embolism is a good example of this: imaging to assess the integrity of the pulmonary vascular bed is required. Tests of muscle power will pick up neuromuscular conditions weakening respiratory movement. The hyperventilation of acidosis, as in uraemia or diabetic coma, is not often described by patients as breathlessness and is otherwise easily diagnosed. However, hyperthyroidism and anaemia should not be forgotten as causes of breathlessness. Some 60 per cent of patients with a haemoglobin level of less than 8 g/dl will have this symptom.

Psychogenic breathlessness is diagnosed by exclusion, although there may be clues in the history and examination. The quality of the breathlessness has been described above. The sighing and irregular breathing will be readily noticeable to a keen observer. Associated complaints directly related to the hyperventilation are paraesthesiae in the hands and perhaps feet, dizziness, and collapse. Apparently non-specific features such as fatigue, insomnia, weakness, or chest pains may all be part of the syndrome. Depression and anxiety may both be aspects of the underlying psychiatric state. By definition, in pure psychogenic breathlessness, the chest radiograph and lung function are normal. However, some patients may develop breathlessness because they have been told they have a 'shadow on the lung' or through anxiety exhibit a degree of breathlessness disproportionate to a mild functional abnormality. The latter patients tend to have an obsessional personality or may be looking for compensation for supposed 'lung damage' due to injury or occupation.

Treatment

The relief of breathlessness is best achieved by treating the underlying condition. This may mean the removal of 'mass' lesions (pneumothorax, pleural effusion), or the treatment of pneumonia, airflow obstruction, or alveolitis. Loss of muscle power is occasionally treatable, as in myasthenia gravis, or may recover spontaneously. An attempt to relieve breathlessness should not be neglected, even when the underlying condition is untreatable (pleural effusion in carcinoma of the bronchus, or a reversible steroid-responsive component in a patient with chronic airflow obstruction).

The symptomatic treatment of breathlessness is far from satisfactory, but can usefully be considered in terms of the physiological disorder(s) that can be responsible for the symptom. Excessive respiratory drive may be dampened, for example by oxygen for the hypoxic. A direct approach to the vagal afferent system has met with little success: local anaesthetic to the airways gives a short-lived effect but may itself be irritant. In a select few with intense breathlessness due to diffuse infiltrative disease, vagotomy in the thorax has given some relief. Psychogenic breathlessness may be helped with b-blockers (but asthma must be excluded with absolute certainty). Dihydrocodeine reduces breathlessness in chronic obstructive lung disease but is very constipating and, like the more powerful opiate sedatives, can

dangerously depress respiration. There has been a sad failure to find opiate derivatives with more selective action on breathlessness. Diazepam and promethazine have given subjective relief to some patients disabled by breathlessness from severe emphysema. The use of rehabilitation measures is considered elsewhere (Chapter 17.7).

Chest pain

The greater part of the lower respiratory tract is insensitive to pain and most parenchymal lung disorders proceed to an advanced state without becoming painful. However, the parietal pleura is exquisitely sensitive to painful stimuli and unpleasant sensations can arise from the tracheobronchial tree.

Pleurisy

Typical pleural pain has a sharp, stabbing, and knife-like character; is aggravated by respiration and coughing; and leads to rapid, shallow breathing and a suppressed cough. The pain is likely to be due to stretching of the inflamed parietal pleura and can be relieved by splinting the chest wall.

Afferent pain fibres from the parietal pleura pass through the intercostal nerves. Those from the central portion of the diaphragm run in the phrenic nerve to the cervical cord (C3/4). Central diaphragmatic pleurisy is thus referred to the lateral side of the neck and shoulder tip; indeed, local anaesthesia to the shoulder trigger area can relieve diaphragmatic pleurisy. The outer portions of the diaphragm are served by intercostal nerves (T7–12), causing referred pain to be felt in the lower thorax, lumbar region, and upper abdomen.

Most conditions giving rise to pleuritic pain are acute and inflammatory, either infective (usually pneumonia) or infarctive, as in pulmonary embolism. The immunologically based pleurisies (as in systemic lupus erythematosus or rheumatoid disease) give pain less frequently. Recurrent pleurisy at the same site should suggest bronchiectasis; at different sites it suggests embolism or bronchopulmonary aspergillosis. Sudden chest pain occurs at the onset of a pneumothorax. It is pleural in origin, due to the inrush of air from the lungs and, sometimes, the tearing of adhesions.

If pleurisy progresses to pleural effusion, the sharp pain largely disappears and is replaced by a dull and more constant ache or heaviness. Pleural fibrotic disease is rarely painful, but pleural neoplasia frequently is. The severity and quality of pain depends on the extent of the tumour, and particularly on spread into the chest wall. A superior sulcus tumour of bronchial origin (Pancoast's tumour) infiltrating the brachial plexus gives very severe and persistent pain in the shoulder and in the distribution of C8, T1, and T2.

Pain from the chest wall

Chest-wall pain can mimic pleurisy, and conditions in the chest wall provide its most important differentials. Pain due to strain or tearing of thoracic muscles can be quite sharp, and since it is likely to be caused or exacerbated by coughing and lead to shallow respiration, it can easily be confused with pleurisy. However, there is always local tenderness over the affected muscle and none of the ancillary investigations for pleurisy prove positive. Patients with persistent cough or distressing breathlessness, particularly due to asthma, may complain of muscular pain around the lower rib cage.

Epidemic myalgia or Bornholm's disease is a bothersome manifestation of Coxsackie B infection giving fever and recurrent muscle pain. If the intercostal muscles are involved (pleurodynia), the associated breathlessness and tachypnoea can exactly mimic pleurisy, as can the pre-eruptive stage of thoracic herpes zoster, which gives a stabbing pain in the distribution of the affected nerve. Costal cartilage pain is generally not inflammatory. In Tietze's disease there is a painful protuberance of one or more costal cartilages, usually the second to fourth, probably due to asymmetrical growth of the rib cage. Osteoarthritis and dislocation of the costosternal joints can give chronic pain. Rib fractures rarely present diagnostic problems, but cough fracture in osteoporotic bone should not be forgotten. Thrombophlebitis of chest-wall vessels after surgery or trauma gives anterior chest pain and a tender palpable vascular cord. Most primary chest-wall tumours are not painful, but the more common metastatic disease of bone frequently is, and may be symptomatic before radiological change is evident.

Fleeting transient chest pains are often part of chronic somatized anxiety states, and when this is the case tend to be accompanied by tachycardia, palpitations, and features indicating hyperventilation. Perhaps the commonest chest pain of all is left inframammary pain. This is a transitory sharp but quite severe pain, felt over the apex of the heart at rest or on mild activity. It lasts up to a few minutes and may cause a catching of the breath or shallow breathing. Its cause is unknown but it seems to be totally benign.

Central chest pain

Sensations arising from the tracheobronchial tree are less easy to characterize as painful, although some are exceedingly unpleasant. Instrumentation of the trachea causes pain referred to the anterior chest wall. This is usually abolished by vagotomy and is most likely to be perceived from irritant receptor discharge. Tracheal inflammation, as in infective tracheobronchitis or following the inhalation of toxic vapours, causes a raw painful retrosternal sensation. It is difficult to say how much or how often sensations arising from the main airways are describable as pain. There is often a component described as tightness, and this is a common complaint of patients with generalized airflow obstruction, although it is probably naïve to think that the sensation is a direct appreciation of airways narrowing. Further complicating the interpretation of sensation in these conditions is the almost universal association with coughing which, if persistent, can itself lead to soreness in the upper airways and trachea.

Finally, the mediastinal structures of the thorax are responsible for a multitude of pains, the majority of which are dealt with elsewhere in the chapters on cardiology and gastrointestinal disease. Few central pulmonary lesions give mediastinal pain. Only neoplasia is a common culprit. A central bronchial carcinoma or hilar nodes associated with it can be responsible for a deep dull aching pain in the centre of the chest. Similar pain can sometimes occur in the early stages of sarcoidosis with hilar lymphadenopathy and in lymphoma.

Other symptoms in pulmonary diseases

Patients or their relatives on their behalf may complain of noisy breathing, generally using the word 'wheeze'. A harsh inspiratory wheezing sound arising from obstruction in the larynx or major airways is termed stridor. There may be accompanying hoarseness or features of intrathoracic disease. Wheeze is the externally audible counterpart of the sounds heard with the stethoscope in asthma and obstructive bronchitis. It is a term frequently used by asthmatic patients to describe their respiratory symptoms.

When airflow obstruction is suspected, specific enquiries should be made for the features of bronchial irritability. In response to changes in atmospheric conditions (particularly temperature) or to the inhalation of dusts, fumes, or vapours, the patient with irritable bronchi will respond with a variety of symptoms: cough, tightness in the chest, wheeze, or breathlessness.

Rarely, patients may complain that they are blue (cyanosed), although their carers may do so more often. This and finger clubbing are more often elicited as physical signs (see below).

General history

A full history is essential, emphasizing the following features:

1. cardiac disease as a cause or aggravating factor in breathlessness;
2. the legs for ankle oedema as a result of lung disease or deep venous thrombosis;
3. the upper respiratory tract for infectious, allergic, or vasculitic disorders;
4. the skin for eczema, urticaria, erythema nodosum, or vasculitis;
5. features of rheumatoid or collagen-vascular disease;
6. the nervous system for disease that might impair ventilatory control; and
7. pointers to metastatic spread or the non-metastatic manifestations of malignant disease.

The past history may reveal atopy, tuberculosis, or other serious infectious disease, particularly in childhood. It is always worth asking about previous chest radiographs which may be obtainable for comparison.

A full smoking history is essential. A detailed drug history is essential because of potential toxic effects on the lungs (see Chapter 17.11.19).

A complete occupational and environmental history is of the utmost importance. Whilst the mining industries will be obvious, many other occupations that create dusts of both inorganic and organic materials are now recognized as presenting hazards to the chest (see Chapter 17.11.7). Certain working environments may lead to exposure to organisms likely to cause pulmonary infection: *Chlamydia psittaci* from contact with domestic or wild birds, *Coxiella burnetti* in slaughterhouses, and tuberculosis through working with susceptible groups.

Finally, certain disorders have a familial predisposition. These include asthma and other atopic diseases (see Chapter 17.4.1), cystic fibrosis (see Chapter 17.10), Kartagener's syndrome, familial fibrocystic pulmonary dysplasia (a form of fibrosing alveolitis) (see Chapter 17.11.2), pulmonary lymphangiomyomatosis (see Chapter 17.11.10), and alveolar microlithiasis (see Chapter 17.11.16). A family or personal contact history of tuberculosis should be noted, as should any record of previous tuberculin testing or bacille Calmette–Guérin (**BCG**) vaccination.

Physical signs in pulmonary disease

Inspection of the chest

The pattern of breathing and the configuration of the chest must be observed. The normal respiratory rate when the subject believes him or herself to be unobserved is around 10 to 14 per min. Higher rates than this are commonly recorded in the healthy, but a rate above 20 per min is abnormal. Pneumonia, many interstitial lung disorders, and abnormal drives to breathing, including anxiety, will increase the rate. If the chest is free to move, the tidal volume will also increase, but this is not the case with restrictive disease or painful conditions of the thoracic cage or upper abdomen. An abrupt stop to inspiration when there is pain can be seen. The frequency of deep sighs, normally 8 to 10 per hour in quiet breathing, is greatly increased in psychogenic breathlessness, when there may be an irregular pattern including phases of rapid breathing and relative apnoea. A regular alternation of apnoeic periods of 5 to 30 s with a period of increasing and then decreasing ventilation characterizes Cheyne–Stokes respiration. This and several other irregular breathing patterns are usually associated with brainstem or cerebral lesions, but they can also be a feature of severe heart failure.

In observing respiratory movement, particular attention must be paid to expansion. Poor movement of the chest on one side only always indicates pathology on that side. Generally poor expansion is seen in the hyperinflated chest of the patient with severe airflow obstruction and in the fixed thoracic cage of advanced ankylosing spondylitis. In airflow obstruction two other features may be observed: an indrawing of intercostal spaces during inspiration (reflecting the negative intrapleural pressure necessary to draw air into the lungs) and abnormal movement of the lower chest. Normally, the lower chest moves outwards during inspiration, but in gross hyperinflation the diaphragm is flat and its contraction merely causes the lower thoracic cage to move inwards. In the same patients the anterior abdominal wall may also move inwards during inspiration instead of outwards, and this asynchrony of movement carries a poor prognosis.

Abnormalities of the shape of the chest are well recognized. An increased anteroposterior diameter to give a 'barrel chest' is as often a sign of the kyphosis that accompanies senile osteoporosis as it is of the hyperinflation of emphysema and chronic airflow obstruction. Pectus carinatum (pigeon chest), an outward protuberance of the sternum, may reflect severe attacks of asthma in childhood when it may be accompanied by bilateral indrawing of the anterior portions of the lower ribs (Harrison's sulci); it is now rarely due to rickets. The opposite, pectus excavatum (depressed sternum), is a congenital anomaly. Scoliosis is important because of the severe impairment of respiratory movement that it causes, and it can lead to respiratory failure (see Chapter 17.13). Localized collapse and fibrosis may draw in the adjacent rib cage (which will also move poorly) and, if severe, unilateral fibrosis of the whole lung can cause a scoliosis with its curvature towards the affected lung.

Palpation of the chest

Palpation is used to confirm the observed patterns of chest expansion and to identify the position of the trachea and apex beat. The trachea should be localized in the suprasternal notch with the index finger. With the patient looking directly forwards, any deviation of the trachea from the mid-line should be assessed using a combination of touch and vision. Aside from tension pneumothorax, deviation of the trachea to one side is due to either apical fibrosis pulling it to the affected side or a mass in the neck (for example, goitre) or upper mediastinum pushing it to the opposite side. As with the trachea, the position of the apex beat can reflect pressure against or traction on mediastinal structures, but due consideration must be given to displacement of the apex beat due to intrinsic cardiac disease.

The detection of the transmission of vocal sounds by the placing the palm of the hand on the chest (vocal fremitus) should be abandoned in favour of vocal resonance (listening with the stethoscope for voice sounds), except for a simultaneous comparison of the two sides of the chest.

Percussion of the chest

In properly performed percussion, the examiner listens for the pitch and loudness of the percussed note, and both listens and feels for the postpercussive vibrations that give the note its resonance. The sides of the chest must be compared from identical sites. A dull note lacks resonance and is higher in pitch and softer than a normal percussion note; it signifies the presence of solid tissue or fluid underneath the percussed area. 'Stony' dullness with a complete lack of any vibrations coming back from the lung is heard and felt over pleural effusions. It is important to delineate the surface markings of any dullness: pneumonic consolidation and collapse will follow the distribution of the affected lobe, whereas the upper limit of a pleural effusion will be determined by the effects of gravity. It requires a fine ear to pick up Ellis's S-shaped line—a slightly higher level of dullness in the axilla when the patient is in the sitting position. Large effusions which displace mediastinal contents may produce an area of dullness at the opposite base close to the mid-line (Grocco's sign).

A hyper-resonant note is lower in pitch and louder than normal, and occurs over hyperinflated lung as in emphysema or an air-filled space, that is to say a large bulla or a pneumothorax. It is more difficult to be certain about hyper-resonance than dullness, particularly in thin subjects.

Auscultation of the chest

There are three types of sound that can be heard coming from the lung: breath sounds, adventitious sounds, and voice sounds.

Breath sounds

Normal breath sounds are better termed 'normal' rather than 'vesicular'. They are certainly not generated in the vesicles or alveoli of the lung where air flow is too low, but probably reflect turbulent flow in major bronchi. The pattern and intensity of breath sounds reflects regional ventilation. Thus, in the normal upright lung, breath sounds are loudest at the apex in early inspiration and at the bases in mid-inspiration. Breath sounds are quietened over areas of atelectasis. During expiration normal breath sounds rapidly fade out, probably due to the decreasing air-flow rate.

Bronchial breathing is heard over airless lung as in consolidation, atelectasis, or dense fibrosis. There is some resemblance to the sounds heard over the normal trachea, but, by comparison with normal breath sounds, bronchial breathing is higher in pitch and more blowing in quality. It does not have to be loud. Bronchial breath sounds are classically heard throughout both inspiration and expiration. Very quiet breath sounds are heard over hyperinflated lungs as in emphysema or when breath sounds are prevented from reaching the chest wall by a layer of air, fluid, or fibrosis.

Adventitious sounds

The terminology of adventitious sounds is confused. This arises because, whereas Laennec originally used the term rales (rattle) to embrace all added sounds, Latham, introducing the classification dry and moist sounds in 1876, applied rale exclusively to the former and rhonchi to the latter. Until recently the established convention in the United Kingdom was to drop rale altogether and to call interrupted non-musical sounds crepitations and continuous musical sounds rhonchi. The move to replace the term crepitations with crackles and the term rhonchi with wheezes has now gained widespread acceptance. Crackles may be coarse or moist

when they are due to the movement of sputum in large airways, or fine when they are probably created by small airways snapping open as pressure equalizes in the distal lung compartment. Coarse early inspiratory and expiratory crackles are often heard in respiratory tract infection, particularly in patients with chronic obstructive lung disease, whilst fine late inspiratory crackles are characteristic of pulmonary oedema and fibrosing alveolitis. Occasionally a single mid to late inspiratory 'squawk' is heard in patients with a variety of pulmonary fibroses.

Wheezes signify obstruction in airways. A sound of single pitch (monophonic) in inspiration and/or expiration, which cannot be altered by coughing to shift mucus, signifies a localized obstruction in a major airway. Several sounds of varying pitch (polyphonic) heard randomly in inspiration and expiration are typical of the widespread airways obstruction of asthma and chronic obstructive bronchitis. A polyphonic wheeze on forced expiration signifies diffuse airflow obstruction and can be a useful sign when tidal breathing is free of added sounds.

A pleural rub is the diagnostic added sound of pleurisy. It is a superficial grating or rasping sound synchronous with late inspiration and early expiration, best heard at the bases and rarely at the apices. A soft friction rub may be mistaken for crepitations, but is not altered by coughing and can be made louder by pressure with the stethoscope. Inflammation of the pleura close to the heart can give a friction rub that synchronizes with the heart beat but will cease if the breath is held.

Voice sounds

A long sound such as 'ninety-nine' is favoured for detecting voice sounds that are transmitted by normal lung, but not by air space or fluid, and pass through solid lung with undue clarity, even allowing whispered sounds to be heard (whispering pectoriloquy). Certain physical characteristics of a solid lung allow low frequency sounds to be filtered out, leaving a sound of bleating or nasal quality (aegophony); this is particularly noticeable over a collapsed lung adjacent to a pleural effusion.

The relevance of the general examination in respiratory disease

Clues to the diagnosis of respiratory disease and critical extrathoracic manifestations of primary lung conditions must be sought in the general examination.

Overall appearance

Obesity places an added burden on the respiratory system, sometimes sufficient in itself to cause a degree of exertional breathlessness and potentially a cause of obstructive sleep apnoea (see Chapter 17.8.2). Truncal obesity with moon facies and skin bruising is an unfortunate complication of oral corticosteroid therapy which may have to be given for several pulmonary diseases.

Weight loss is a feature of emphysematous obstructive lung disease and, of course, malignancy, with the late stages of bronchial carcinoma and pleural mesothelioma often being characterized by a distressing cachexia. Malabsorption can result in weight loss in cystic fibrosis if inadequately managed.

Body habitus can alert to possible respiratory complications, particularly the more severe degrees of kyphoscoliosis, which can lead to hypoventilation, and also rarer disorders such as Marfan's syndrome (associated with pneumothorax) or ankylosing spondylitis.

Cyanosis

Cyanosis is the blue discoloration imparted to the nailbeds, lips, and tongue by hypoxaemic blood. Peripheral cyanosis due to a sluggish peripheral circulation, as in cold weather, will leave the tongue still pink, whereas in central cyanosis the tongue will be blue and the peripheries blue yet often warm. The frequently repeated statement that 5 g of reduced haemoglobin is required before cyanosis can be detected is false. Most patients with a saturation of 90 per cent or less will appear cyanosed. This represents just 1.5 g of reduced haemoglobin if the total haemoglobin is 15 g. Cyanosis is less marked in severe anaemia and more obvious in polycythaemia. The curious phenomenon of orthocyanosis (hypoxia occurring only in the upright position) is generally associated with pulmonary arteriovenous malformations.

Clubbing of the fingers

Loss of the natural angle between the nail and the nailbed in a properly manicured finger, and a boggy fluctuation of the nailbed are cardinal signs of clubbing (Fig. 1). An increased curvature of the nail and enlargement of the end of the finger develop later. The toes may also be affected. The differential diagnosis of clubbing of the fingers includes many extrathoracic conditions but, as far as the lungs are concerned, three categories deserve consideration: (1) suppurative disease, particularly bronchiectasis of long-standing but also acute lung abscess and empyema, but not uncomplicated bronchitis; (2) fibrosing alveolitis and asbestosis, but rarely other diffuse fibrotic diseases; (3) malignant disease, particularly carcinoma of the bronchus and also pleural malignancy. If finger clubbing is associated with hypertrophic pulmonary osteoarthropathy, a painful osteitis of the distal ends of the long bones of the lower arms and legs, malignancy is associated in 95 per cent of cases.

Fig. 1 Clubbing of the fingers.

There is no satisfactory explanation for clubbing and hypertrophic pulmonary osteoarthropathy. Pathologically, there is abnormal vascularity and new bone formation in the peripheries, and evidence of abnormal bronchopulmonary anastomoses in the lungs. The latter may be under vagal control since vagotomy has sometimes abolished clubbing in lung cancer patients. These intrathoracic channels may allow substances normally detoxified by the lungs, which could be responsible for the peripheral changes, to enter the systemic circulation.

The skin and eyes

Eczema and urticaria point to an atopic diathesis and hence possible asthma. Erythema multiforme can accompany mycoplasma pneumonia, rarely other pneumonias, and pulmonary blastomycosis. Erythema nodosum, tender nodules fading to a bruised purple on the shins and occasionally the forearms, is a classical presentation of sarcoidosis, frequently associated with hilar lymphadenopathy and less often with pulmonary infiltrates. It may also be found in primary tuberculosis and rarely in other chest infections.

Several other dermatological, ocular, arthritic, and internal manifestations may also alert to the diagnosis of sarcoid. Skin and eyes are also the site of lesions in Wegener's granulomatosis, systemic lupus erythematosus, systemic sclerosis, and dermatomyositis, each of which has potential pulmonary manifestations.

Patients with diffuse neurofibromatosis and tuberous sclerosis can both develop a severe pulmonary fibrosis with late-stage destructive emphysema. Rarely, hereditary haemorrhagic telangiectasia, with its characteristic lesions on the lips, face, and mouth, can extend to the lungs with pulmonary haemangiomas which give haemoptysis as well as the more commonly found gastrointestinal lesions. The latter give anaemia, and this, whatever its cause, can cause breathlessness and so

must be checked for, as must jaundice by looking at the eyes.

The skin is the site of secondary deposits from carcinoma of the lung in a small percentage of cases, although usually late in the disease when the diagnosis is all too obvious. Other carcinomas may spread to skin and lungs, and Kaposi's sarcoma is a cutaneous manifestation of disseminated HIV infection (see Chapter 7.10.21).

Head and neck

Signs of upper respiratory tract disease are relevant in pointing to a site and source of infection that could track down to the lungs, and for the ways in which they signify allergy. Furthermore, neurological disease of the pharynx or structural abnormalities of the larynx encourage aspiration and repeated respiratory tract infection. A short thick neck, retrognathia, and a large uvula can all predispose to sleep apnoea. A goitre may be large enough to compress the trachea and cause stridor. It could be associated with hyperthyroidism and so breathlessness; or hypothyroidism and hence hypoventilation; or even be a source of a carcinoma that could metastasize to the lungs. Even more important are signs in the neck that represent primary intrathoracic malignancy. Hard enlarged cervical lymph nodes are a well-recognized metastatic site for carcinoma of the lung, but can signify lymphoma or primary cancer from elsewhere (stomach, breast). The local extension of central bronchogenic carcinoma gives superior vena caval thrombosis. There is fixed elevation of jugular venous pressure, a large neck, a congested face and head, and, in severe cases, exophthalmos and impaired vision.

Cardiovascular system

The pulse is of poor volume in pulmonary hypertension, bounding and full in hypercapnic respiratory failure, and waxes and wanes in acute severe airflow obstruction (pulsus paradoxus). In the last of these the pulse can actually disappear at the height of the negative intrapleural pressure swing (mid to late inspiration).

The pulmonary hypertension that accompanies severe hypoxic cor pulmonale is manifested on physical examination by a raised jugular venous pressure with a prominent A wave, peripheral pitting oedema, and cardiac signs of right ventricular heave, a loud P2 wave, and, in failure, a gallop rhythm. Oedema alone should add differential diagnoses such as liver or renal disease, both of which can have pulmonary manifestations.

17.3.1Thoracic imaging

Susan Copley and David M. Hansell

Despite recent technological advances, chest radiography remains the cornerstone of thoracic imaging. The chest radiograph is justifiably regarded as an integral part of the examination of the patient in respiratory medicine. Because of the wealth of information available from chest radio-graphy, careful interpretation of the chest radiograph remains a necessary clinical skill. Advances in cross-sectional imaging have had a great impact in improving the diagnosis of thoracic pathology, not only for the assessment of mediastinal disease but also in the evaluation of patients with suspected diffuse lung disease. Nevertheless, a chest radiograph should always be obtained and looked at carefully before submitting a patient to more sophisticated imaging techniques. The expense and radiation burden, in the case of computed tomography, is an important consideration.

Techniques in thoracic imaging

Chest radiography

The first chest radiograph was taken over 100 years ago and chest radiography is now the most frequently requested radiological investigation worldwide. The technique has changed surprisingly little over the years, although digital technology has recently been used to overcome some of the shortcomings of conventional film-based radiography.

Technical considerations

An ideal chest radiograph is taken with the patient standing erect, suspending respiration at total lung capacity, and with the X-ray beam traversing the thorax from back to front (the posteroanterior (**PA**) or frontal view). Because of the wide range of densities within the chest (soft tissues of the mediastinum through to aerated lung), perfect exposure of every part of the chest radiograph is impossible. The resulting suboptimal exposure of the denser part of the chest can be partially overcome with a high-kilovoltage technique (120 to 150 kVp). With this technique there is greater penetration of the mediastinum, which improves visualization of the trachea and main bronchi. A disadvantage of high-kilovoltage radiography is the relatively poor demonstration of calcified structures so that rib fractures and calcified pulmonary nodules or pleural plaques are less conspicuous. Even with an optimal technique, nearly a third of the lungs are partially obscured by overlying mediastinum, diaphragm, and ribs.

Automatic exposure devices have been developed to expose accurately the various parts of the chest. One of these, the Advanced Multiple Beam Equalisation Radiography (**AMBER**) system, produces chest radiographs which greatly improve the demonstration of both mediastinal anatomy and pulmonary abnormalities (see Fig. 16). Another approach to this problem is phosphor-plate computed radiography, which uses digital technology. This is ultimately expected to replace conventional film radiography, and has already done so in some centres. A phosphor plate is handled in a conventional cassette (which does not contain film) and is exposed in the normal way. The energy of the incident X-ray beam is stored as a latent image. The phosphor plate is then scanned with a laser beam and the light emitted from the excited latent image is detected by a photomultiplier. Thereafter this signal is processed in digital form, and the image may either be viewed on a television monitor or laser-printed on to film. The advantage of phosphor-plate computed radiography is that it can retrieve an image of diagnostic quality from an imperfect exposure, which would otherwise result in a non-diagnostic conventional film radiograph. Manipulation or processing of the digital image data can enhance certain features of the radiograph to improve diagnosis.

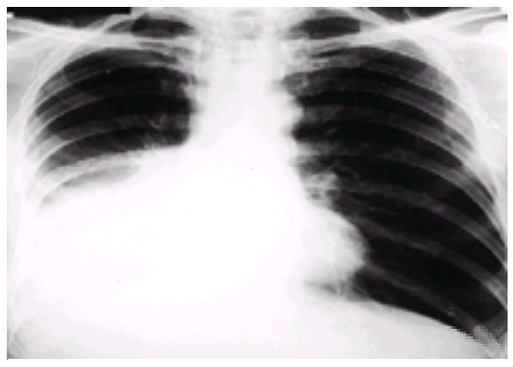

Fig. 16 Right lower lobe collapse.

Standard radiographic views of the chest

The posteroanterior projection is the standard view (see Fig. 11(a)). The patient is positioned with his anterior chest wall against the film cassette and his arms are abducted to rotate the scapulas away from the posterior chest. Chest films taken in the anteroposterior (**AP**) projection are usually taken when the patient is too ill to stand for a formal PA radiograph. A consequence of this view is that the heart is magnified because it lies further from the film. Moreover, the shorter X-ray tube-to-film distance, which is inevitable when a portable AP radiograph is taken, causes further magnification that must be taken into account when assessing the

heart size on an AP chest radiograph.

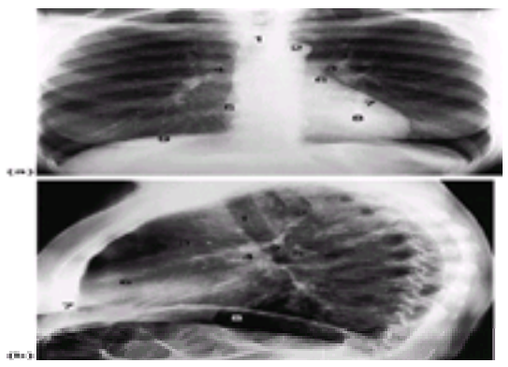

Fig. 11 Normal radiographic anatomy on (a) PA and (b) lateral chest radiographs. (a) 1, Trachea; 2 aortic arch; 3, left main pulmonary artery; 4, right main pulmonary artery; 5, right atrial border; 6, left atrial appendage; 7, left ventricular border; 8, right ventricle; 9, right dome of the diaphragm. (b) 1, Trachea; 2, scapulas; 3, anterior aortic arch; 4, right pulmonary artery; 5, left pulmonary artery; 6, right ventricle; 7, breast shadows; 8, gastric bubble under the left hemidiaphragm; 9, left main bronchus.

The lateral radiograph is obtained by placing the patient at right angles to the film cassette (see Fig. 11(b)). The lateral projection provides the third dimension and helps to determine the site of a lesion identified on the PA projection (although it is surprising how often an opacity clearly seen on the PA radiograph is invisible on the lateral radiograph). As well as allowing accurate localization of lesions, the lateral radiograph may reveal concealed abnormalities that lie behind the heart or diaphragm. Furthermore, with some experience, evaluation of the hilar structures and major airways is aided by the lateral radiograph (see later section on the anatomy on the lateral chest radiograph).

Over the years, a number of supplementary projections have been developed to provide information about areas that are not easily seen on the standard PA and lateral radiograph. With the advent of cross-sectional imaging, notably computed tomography (**CT**), many of these extra views have become obsolete. However, even with access to CT, some of these views supply extra anatomical detail readily and inexpensively and these will be considered briefly.

The lateral decubitus projection is sometimes useful for the demonstration of small pleural effusions. For this view the patient lies on his side, with the side in question downwards; the film is positioned behind the patient and the X-ray beam traverses the patient horizontally. Small quantities of pleural fluid (50 to 100 ml), which are not detectable on a PA chest radiograph, can be demonstrated tracking up the lateral chest wall, but ultrasonography is increasingly being used as a reliable technique for demonstrating small pleural effusions. Other supplementary projections, for example apical and lordotic views, improve visualization of the lung at the extreme apices. These are now rarely performed, CT being much more effective at showing pathology in these difficult areas.

The technique of screening the patient with fluoroscopy has the advantage of allowing 'real time' radiographic examination. It allows localization of lesions by the use of unusual oblique projections, for example to distinguish a small pleural plaque from an intrapulmonary nodule. Fluoroscopy is also the quickest method of evaluating diaphragmatic movement and diagnosing air-trapping in a child who is suspected of inhaling a foreign body.

Ultrasonography

High-frequency sound waves do not traverse air and are completely reflected at interfaces between soft tissue and air. The use of this technique in the chest is therefore limited by normally aerated lung. However, fluid can be readily detected and the main use of ultrasound is for the localization of small or loculated pleural effusions. Furthermore, ultrasound can differentiate between pleural fluid and pleural thickening in cases in which radiography cannot make this distinction. Ultrasonography is an extremely useful technique for guiding percutaneous needle biopsy of masses arising from the chest wall or pleura, or peripheral pulmonary masses or consolidation, and for aiding the accurate placement of a chest drain within a pleural collection. Ultrasonography may show numerous septations within an exudative pleural effusion (Fig. 1), but thoracocentesis may be required to distinguish between a parapneumonic effusion and an empyema.

Fig. 1 Ultrasonography showing a pleural effusion. Fibrinous septations traverse the pleural space.

Computed tomography

Computed tomography (CT) depends on the same basic principle as conventional radiography, namely the differential absorption of X-rays by tissues of disparate densities. However, CT is much more sensitive to differences in attenuation of X-rays by various tissues. A CT machine consists of an X-ray source and an array of detectors which surround the patient. The X-ray source rotates around the patient and the resulting attenuated beam is measured by the detectors. The signals from the detectors are used to construct an image by a mathematical technique. The reconstructed images are transverse (axial) cross-sections of the patient and are viewed as if from the patient's feet (i.e. on the image, the patient's right side is to the viewer's left). Each CT section is a matrix of three-dimensional elements (voxels) containing a measurement of X-ray attenuation, arbitrarily expressed as Hounsfield Units (**HU**): water measures 0 HU, air –1000 HU (so that lung parenchyma is approximately –600 HU), fat –80 HU, soft tissue 40–80 HU, and bone 800 HU. If a voxel is completely occupied by a tissue of uniform density (most frequently the case with narrow sections) then the HU will be truly representative of that tissue. If the section contains tissues of two different densities (more likely with thicker sections), for example half lung and half dome of diaphragm, then the attenuation value will be a weighted average of the two components: the so-called 'partial volume' effect.

Because of the cross-sectional nature of CT it can accurately localize lesions seen on only one view on chest radiography. The superior contrast resolution of CT gives exquisite detail of the various components of mediastinal anatomy (for example, lymph nodes and vessels) and density differences (for example, calcifications within a pulmonary nodule). Different image settings are needed to view the soft tissue structures of the mediastinum and the aerated lung parenchyma (Fig. 2).

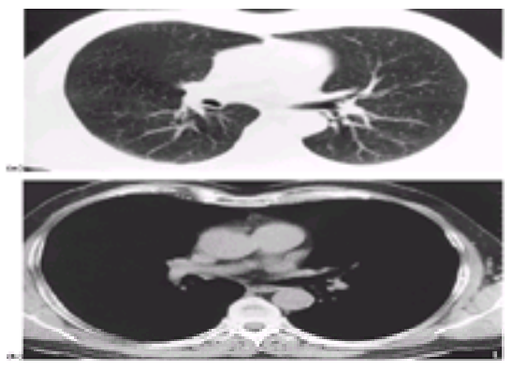

Fig. 2 Computed tomography section through the mid-thorax. The window settings have been adjusted to show details of (a) the lungs and (b) the soft tissues of the mediastinum.

Spiral (helical) CT

The principle of spiral CT involves the continuous rotation of the X-ray beam and detectors around the patient while the table moves into the gantry. Markedly reduced scan times are possible with the introduction of spiral CT, allowing the entire thorax to be imaged in a single breath-hold. An examination of sufficient diagnostic quality can be obtained in dyspnoeic patients and young children during quiet respiration. It also allows an intravenous injection of contrast medium to be accurately timed to give the optimum opacification of, for example, the pulmonary arteries, thus enabling the detection of pulmonary emboli (Fig. 3). In many centres, spiral CT has supplanted ventilation–perfusion radionuclide imaging in the investigation of patients with suspected pulmonary embolism and, with optimum technique, the accuracy approaches that of pulmonary arteriography for the diagnosis of central, lobar, and segmental pulmonary emboli. However, radiation dose and availability are important considerations. The technique is most useful in patients with coexisting lung disease which would result in an inconclusive ventilation–perfusion radionuclide study.

Fig. 3 Computed tomography of a patient with acute bilateral pulmonary emboli. Filling defects in contrast media opacification are seen within the right main pulmonary artery and the left interlobar pulmonary artery (arrows). Note the right-sided pleural effusion.

Computer software can perform multiplanar two- and three-dimensional image reconstructions of spiral CT images, which provide novel views of the bronchial tree that can aid interventional techniques such as bronchial stent placement (Fig. 4).

Fig. 4 Reconstructed three-dimensional image from spiral CT showing the carina and the right and left main bronchi viewed from above.

High-resolution computed tomography

High-resolution computed tomography (**HRCT**) uses very thin sections (1–3 mm) and a high spatial frequency reconstruction algorithm to produce highly detailed sections of the lung parenchyma. Both conventional and spiral CT scanners can produce thin sections, and the terms 'spiral CT' and 'high-resolution CT' should not be confused. Submillimetre structures can be resolved with this technique, and the subtle and sometimes complex morphology of interstitial lung diseases can be shown with great clarity (Fig. 5). Since the mid-1980s, the development of high-resolution computed tomography has changed the radiological approach to the diagnosis of diffuse lung disease. High-resolution CT images of the lung correlate closely with the macroscopic appearances of pathological specimens, so that HRCT represents a substantial improvement over chest radiography in terms of sensitivity, specificity, and diagnostic accuracy. Furthermore, CT samples a far greater volume of lung than even the most generous lung biopsy, making it less prone to sampling errors.

Fig. 5 High-resolution computed tomography of a patient with lymphangioleiomyomatosis showing thin-walled cysts throughout the lungs. These cysts were not apparent on chest radiography.

High-resolution CT has also been shown to provide useful information regarding prognosis and response to treatment in some diffuse lung diseases. Nevertheless, despite the increased confidence with which a specific diagnosis of diffuse lung disease can be made with HRCT, open lung biopsy is still required to achieve a definitive histological diagnosis in difficult cases. The extent of diffuse lung disease can be estimated precisely on HRCT and, when a biopsy is indicated, the distribution of disease will indicate whether a transbronchial biopsy or an open lung biopsy is more likely to obtain a representative specimen.

The disadvantages of CT are its relatively high cost and increased radiation exposure to the patient, particularly in comparison with chest radiography. For these reasons, CT should not be regarded as a routine investigation and examinations should always be tailored to solve questions not answered by less sophisticated investigations. The commonest indications for thoracic CT are summarized in Table 1.

Magnetic resonance imaging

The physical principles of magnetic resonance imaging (**MRI**) are very different to those governing CT scanning. An MR image is obtained by placing an individual in a strong magnetic field which polarizes some of the ubiquitous hydrogen protons (which can be thought of as behaving like randomly oriented bar magnets) in the body so that they have the same alignment. The application of radiofrequency wave pulses of specified lengths and repetition (pulse sequences) displace the protons and some of this transmitted energy is absorbed by them. With the cessation of the radiofrequency pulse, the protons return to their initial alignment and in so doing they emit, as a weak signal, some of the energy they have absorbed; this signal is received and then amplified and handled in digital form, and is subsequently reconstructed into an image.

The advantages of MRI include its ability to obtain sections in any plane (Fig. 6), the improved contrast resolution between different soft tissues compared with CT scanning, and the use of special sequences which give functional information, for instance the velocity of blood flow. Another important advantage of MRI is the lack of any known hazard to the patient, in contrast to CT scanning with its small attendant risk from ionizing radiation. Disadvantages of MRI include the long scan time (although this is continually being shortened), reduced spatial resolution compared with CT, the inability to image calcium, its reduced acceptability to patients because of the claustrophobic bore of the magnet, and important contraindications such as permanent cardiac pacemaker devices and ferromagnetic intraocular foreign bodies.

Fig. 6 Magnetic resonance image (coronal section) showing the relationship of an apical bronchial carcinoma to the chest wall and adjacent mediastinum. There are enlarged subcarinal lymph nodes and a metastatic deposit in the right adrenal gland.

In many respects the imaging of the mediastinum by CT scanning and MRI are comparable. However, MR images of the lungs are currently markedly inferior to CT scanning. This is because of the very low water (and therefore proton) content of the lungs: the signal produced by a normal lung is therefore small and not visualized by conventional sequences.

Radionuclide imaging

Ventilation–perfusion radionuclide scanning is an effective non-invasive method of providing both anatomical and physiological information about the lung. It is the commonest radionuclide study of the lungs, and is most frequently used to confirm or exclude the diagnosis of suspected pulmonary embolism.

Regional pulmonary capillary perfusion can be assessed following the intravenous injection of a bolus of particles which have been labelled with technetium-99m. The minute particles are microspheres or macroaggregates of human albumin (between 15 μm to 70 μm in diameter). These particles are evenly dispersed by the time they reach the pulmonary circulation and become temporarily lodged in a very small fraction (less than 0.5 per cent) of the precapillary arterioles and capillaries of the lungs. There is a small theoretical risk of compromising the pulmonary vascular bed in patients with severe pulmonary hypertension, but this is not an absolute contraindication to the examination. The distribution of gamma-ray emission from the technetium-labelled particles is directly proportional to the regional pulmonary flow, and a significant defect in perfusion is usually readily detected. It is important to appreciate that such defects may be due to a variety of conditions other than a pulmonary embolism, including any cause of hypoxic vasoconstriction such as an area of subsegmental collapse or space-occupying lesions not supplied by the pulmonary circulation. However, in these cases the affected area of lung will be neither ventilated nor perfused, in contrast to acute pulmonary embolism in which there is no corresponding defect of ventilation. Thus, to improve the specificity of the diagnosis of a pulmonary embolism, ventilation scintigraphy is usually performed at the same time as perfusion scanning.

Evaluation of ventilation of the lungs depends on filling the distal air spaces with a gamma-ray emitting radionuclide. The radionuclides suitable for inhalation are the inert gases xenon-133 and krypton-81m or a technetium-99m aerosol (Technegas). While krypton-81m gives the highest quality images, Technegas is being increasingly used because of its ready availability. The characteristic abnormality of pulmonary embolism is the so-called mismatched defect in which a regional defect in perfusion is not matched by a defect in ventilation (Fig. 7(a) and Fig. 7(b)). However, the picture in pulmonary embolism may not always be clear-cut, particularly when pulmonary infarction has occurred, when there will be a matched defect of both ventilation and perfusion. Because of the importance of establishing a correct diagnosis of pulmonary embolism, ventilation–perfusion scans should always be interpreted in the light of current chest radiographs and clinical information. Even then a proportion of ventilation–perfusion scans remain indeterminate, hence the increasing use of CT angiography in the management of these patients. Due to the decreased radiation burden, V/Q scanning remains a reasonable first-line investigation in young patients with no pre-existing lung disease and a low pretest probability for pulmonary embolism.

Fig. 7 A ventilation–perfusion radionuclide study (oblique views). The perfusion scan (a) shows a defect in the left mid-zone which is not matched on the corresponding view of the ventilation scan (b). The so-called mismatched defect is characteristic of a pulmonary embolus.

Positron emission tomography

Positron emission tomography (**PET**) relies on the tissue uptake of radioisotopes which decay by positron emission. Detectors located around the patient map the site of origin of the two resultant photons emitted at 180 degrees from each other. The most widely used isotope for the detection of pulmonary malignancy is [^{18}F]fluorodeoxyglucose (**FDG**), a D-glucose analogue. The increased uptake and retention of glucose by malignant cells allows differentiation of benign from malignant pulmonary masses, detection of lymph node involvement by tumour, and identification of distant metastases (Fig. 8). Limitations of the technique include false-positive results caused by granulomatous infection or acute inflammation, and false-negative results with certain tumours (e.g. bronchioloalveolar carcinomas and carcinoid tumours). The technique is not widely available at present.

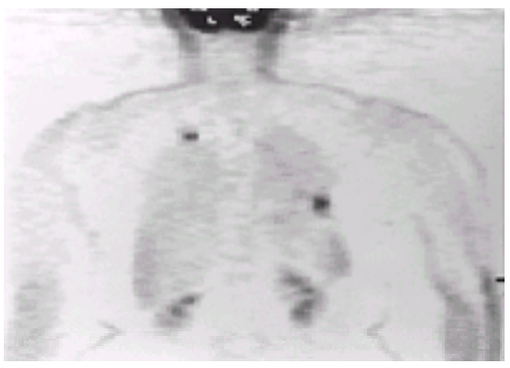

Fig. 8 Positron emission tomography image showing increased uptake of [^{18}F]fluorodeoxyglucose (FDG) in the left lower zone corresponding to a primary bronchial carcinoma. Note the pulmonary metastasis at the right apex.

Pulmonary and bronchial arteriography; superior vena cavography

The 'gold standard' for identifying emboli within the pulmonary arteries has traditionally been pulmonary arteriography (Fig. 9). This requires the puncture of an antecubital, jugular, or femoral vein, with the catheter guided through the right heart under fluoroscopic control. While the complication rate is low, it is a time-consuming procedure requiring an experienced angiographer.

Fig. 9 A digital subtraction pulmonary arteriogram showing abrupt termination of the vessels supplying the right upper lobe caused by a pulmonary embolus.

The bronchial arteries which supply the airways become hypertrophied in chronic inflammatory pulmonary disease, notably in bronchiectasis. Rupture of these vessels can cause severe and life-threatening haemoptysis. The bronchial arteries can be selectively catheterized by the passage of a catheter via the femoral artery and aorta. Having identified the abnormally hypertrophied bronchial arteries (Fig. 10), they can be therapeutically embolized. This technique is usually successful in abating massive haemoptysis in patients unable to undergo immediate surgical treatment.

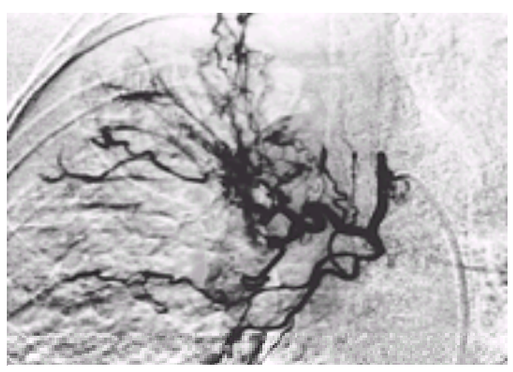

Fig. 10 Abnormally hypertrophied bronchial arteries supplying the right upper lobe shown on a selective digital subtraction bronchial arteriogram. The patient had cystic fibrosis and had had a massive haemoptysis; these bronchial arteries were subsequently embolized.

Superior vena cavography is usually performed to evaluate the exact site of narrowing in patients with symptoms of obstruction of the superior vena cava: it is not generally required to confirm the diagnosis, which is usually evident from the clinical signs alone. Patients with symptoms of superior vena cava obstruction, most frequently due to neoplastic involvement of mediastinal lymph nodes, can be successfully palliated by radiotherapy or the insertion of an expandable metallic wire stent at the site of the narrowing.

Percutaneous lung biopsy

Percutaneous needle biopsy of a pulmonary lesion or mediastinal mass is usually performed in patients in whom a bronchoscopic biopsy has failed to produce a histological specimen, or if a thoracotomy to resect the lesion is deemed inappropriate. It should not be regarded as a routine procedure in the investigation of all solitary pulmonary nodules, and should only be performed after considering the risks to the patient and whether the information forthcoming from the procedure will direct management.

Many different types of needles have been developed, and the frequency of complications, mainly pneumothorax and haemoptysis, is partly related to the diameter of the needle and the depth of the lesion. Percutaneous biopsy is performed under local anaesthesia with CT guidance: ultrasound guidance can be used if the mass

abuts the pleura. Contraindications to the procedure include any patient with poor respiratory reserve who would be unable to withstand a pneumothorax, and pulmonary arterial hypertension.

Normal radiographic anatomy

The mediastinum

On a PA chest radiograph (Fig. 11(a)) the mediastinal structures are superimposed on one another and thus cannot be distinguished individually. The mediastinum is conventionally divided into superior, anterior, middle, and posterior compartments. The practical use of these arbitrary divisions is that specific mediastinal pathologies show a definite predilection for individual compartments (for example, a superior mediastinal mass is most frequently due to intrathoracic extension of the thyroid gland, a middle mediastinal mass is usually due to enlarged lymph nodes). However, it should be borne in mind that the position of a mass within one of these compartments is no guarantee of a specific diagnosis, nor do these boundaries preclude disease from spreading from one compartment to the next.

Because only the outline of the mediastinum and the air-containing trachea and bronchi are clearly seen on a PA chest radiograph, the mediastinal anatomy will be considered in more detail in the description of CT anatomy. On a chest radiograph, the right superior mediastinal border is formed by the right brachiocephalic vein and superior vena cava. The mediastinal border to the left of the trachea above the aortic arch represents the sum of the left carotid and left subclavian arteries together with the left brachiocephalic and jugular veins. The left cardiac border comprises the left atrial appendage which merges inferiorly with the left ventricle. The cardiac silhouette is always sharply outlined: any blurring of the border denotes replacement of the aerated lung immediately adjacent to the heart, usually by collapse or consolidation (see Silhouette sign in Common radiological signs of disease).

The density of the cardiac shadow to the left and right of the vertebral column should be identical, and any difference signals pulmonary pathology (for example, consolidation in a lower lobe). A density with a convex lateral border is often seen through the right heart border on a well-penetrated film: this apparent mass is due to the confluence of the pulmonary veins as they enter the left atrium and is of no pathological significance.

The trachea and main bronchi are visible through the upper and middle mediastinum. The trachea is rarely straight and is often to the right of the mid-line at its mid-point. In elderly patients, the trachea may appear dramatically displaced by a dilated aortic arch. The angle of the carina is usually somewhat less than 80 degrees. Splaying of the carina is a sign of gross disease, either in the form of massive subcarinal lymphadenopathy, or a markedly enlarged left atrium. A more sensitive sign of a subcarinal mass is obliteration of the azygo-oesophageal line which is usually visible on a well-penetrated chest radiograph. The origins of the lobar bronchi, where they are projected over the mediastinal shadow, can usually be made out, but the segmental bronchi within the lungs are not generally seen on plain radiography.

The hilar structures

The hilar shadows on a chest radiograph are a complex summation of the pulmonary arteries and veins, with virtually no contribution from the overlying bronchial walls or normal-sized lymph nodes. The hila are approximately the same size, with the left hilum always lying between 0.5 cm and 1.5 cm above the level of the right hilum. The size and shape of the hila in normal individuals show remarkable variation so that subtle abnormalities are difficult to detect. In detecting a mass at the hilum, at least as important as an abnormal contour is a discrepancy in density between the two hila. Both hilar shadows, at equivalent points, will be of equal density, and a mass at the hilum (or an intrapulmonary mass projected over the hilum) will be evident as increased density of that hilum.

The pulmonary fissures, vessels, and bronchi

The lobes of each lung are surrounded by visceral pleura: the upper and lower lobes of the left lung are separated by the major (or oblique) fissure. The upper, middle, and lower lobes of the right lung are separated by the major (or oblique) and minor (horizontal or transverse) fissures. The minor fissure is visible in about 60 per cent of normal PA chest radiographs. In normal individuals, this fissure runs horizontally and any deviation from this course represents a loss of volume of a lobe. The major fissures are inconstantly identifiable on lateral radiographs. Other fissures are occasionally seen, for example in the left lung a minor fissure can occur which separates the lingula from the remainder of the upper lobe.

All the branching structures seen within the lungs on a chest radiograph represent either pulmonary arteries or veins. The larger pulmonary vessels can be traced back to the hila and mediastinum. The pulmonary veins can sometimes be differentiated from the pulmonary arteries: the superior pulmonary veins have a distinctly vertical course, but in practice it is often impossible to distinguish arteries from veins in the outer two-thirds of the lung. On a chest radiograph taken in the erect position, there is a gradual increase in the diameter of the vessels, at equidistant points from the hilum, travelling from lung apex to base: this is a gravity-dependent effect and is abolished if the patient is supine or in cardiac failure.

The lobes of the lung are divided into segments, each of which is supplied by its own segmental bronchi. The walls of the segmental bronchi are rarely seen on the chest radiograph, except when lying parallel with the X-ray beam when they are seen end-on as ring shadows measuring up to 8 mm in diameter.

The diaphragm and thoracic cage

The interface between aerated lung and the domes of the diaphragm is sharp, and in general the highest point of each dome is medial to the mid-clavicular line. The right dome of the diaphragm is higher than the left by up to 2 cm in the erect position, unless the left dome is temporarily elevated by air in the stomach. Laterally, the diaphragm dips steeply downwards to form an acute angle with the chest wall. Filling in or blunting of these costophrenic angles usually represents pleural disease, either pleural thickening or an effusion.

Localized humps on the dome of the diaphragm are common and represent minor weaknesses or defects. Similarly, interposition of the colon in front of the right lobe of the liver is a frequently seen normal variant.

Deformities of the thoracic cage may cause distortion of the normal mediastinum and so simulate disease. One of the commonest deformities is pectus excavatum, which, by compressing the heart between the depressed sternum and vertebral column, causes displacement of the apparently enlarged heart to the left and causes blurring of the right heart border.

High-kilovoltage chest radiographs often allow the vertebral bodies to be seen through the cardiac shadow. However, the ribs, particularly their posterior parts, are often rendered invisible by this technique.

Anatomy on the lateral chest radiograph

It is useful to get accustomed to viewing a lateral film (Fig. 11(b)) in the same orientation, whether it is a right or left lateral projection. Familiarity with the same orientation improves the viewer's ability to detect deviations from normal.

The trachea is angled slightly posteriorly as it runs towards the carina and the posterior wall of the trachea is always visible as a fine stripe. Furthermore, the posterior walls of the right main bronchus and the right intermediate bronchus are outlined by air and are also seen as a continuous stripe on the lateral radiograph. The spines of the scapulas are invariably seen running almost vertically in the upper part of the lateral radiograph and they should not be confused with intrathoracic structures. Further spurious shadows are formed by the soft tissues of the outstretched arms which are projected over the anterior and superior mediastinum. Although the carina is not visible on the lateral radiograph, the two transradiancies projected over the lower trachea represent the right main bronchus (superiorly) and the left main bronchus (inferiorly).

More lung is obscured by overlying structures on a lateral radiograph than on the PA view. The unobscured lung in the retrosternal and retrocardiac regions should be of the same transradiancy. Furthermore, as the eye travels down the dorsal spine, the viewer should be aware of a gradual increase in transradiancy. The loss of this phenomenon suggests the presence of disease in the posterobasal segments of the lower lobes (sometimes not visible on the frontal radiograph).

The two major fissures are seen as diagonal lines, often incomplete and of a hair's breadth, running from the upper dorsal spine to the anterior surface of the

diaphragm. Care must be taken not to confuse the obliquely running edges of ribs with fissures. The minor fissure extends horizontally from the mid right major fissure. It is often impossible to distinguish the right from the left major fissures with confidence. Similarly, although the two hemidiaphragms may be identified individually (especially if the gastric bubble is visible under the left dome of the diaphragm), the distinction between the right and the left is often impossible. A helpful sign is the relative heights of the two domes: the dome furthest from the film is usually higher because of magnification.

The summation of both hila on the lateral radiograph generates a complex shadow. However, there are some generalizations which aid the interpretation of this difficult area. The right pulmonary artery lies anterior to the trachea and right main bronchus, whereas the left pulmonary artery hooks over the left main bronchus so that a large part of it lies posterior to the major bronchi. As a result, any mass identified on a PA and lateral radiograph that lies anterior to the left hilum or posterior to the right hilum is not vascular in origin and is most likely to represent enlarged hilar lymph nodes.

A band-like opacity is often seen along the lower third of the anterior chest wall behind the sternum. This represents a normal density and occurs because there is less aerated lung in contact with the chest wall because the space is occupied by the heart: it should not be confused with pleural disease.

Normal CT anatomy of the mediastinum

As computed tomography provides unique information about the anatomy of the mediastinum, it is often used to provide further information about abnormalities which are seen merely as a deformity of the mediastinal contour on chest radiography. The normal structures that are always identified on a CT of the mediastinum are the blood vessels (which make up the bulk of the superior mediastinum), the major airways, the oesophagus, and mediastinal fat. An appreciation of the relationship of these structures to each other is crucial for the correct interpretation of CT scans: four important levels are shown in Fig. 12(a), Fig. 12(b), Fig. 12(c) and Fig. 12(d)).

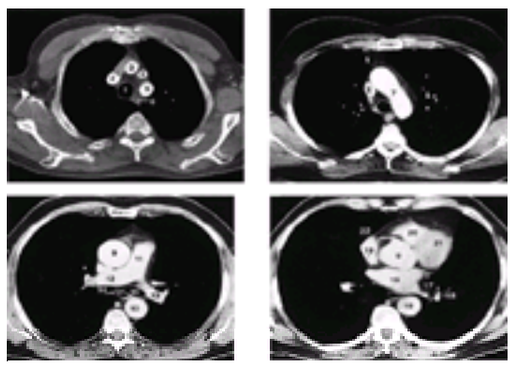

Fig. 12 (a–d) Computed tomography with contrast enhancement to show the normal anatomy at four levels through the mediastinum. 1, Trachea; 2, superior vena cava; 3, brachiocephalic artery; 4, left common carotid artery; 5, left subclavian artery; 6, oesophagus; 7, aortic arch; 8, azygos vein; 9, ascending aorta; 10, descending aorta; 11, main pulmonary artery; 12, right pulmonary artery; 13, left pulmonary artery; 14, right main bronchus; 15, left main bronchus; 16, left atrium; 17, left inferior pulmonary vein; 18, segmental bronchi of the left lower lobe; 19, right atrium; 20, right ventricular outflow; 21, left ventricle.

Normal lymph nodes surrounded by fat may be identified throughout the mediastinum. Many schemes have been devised to map their precise locations, but they can be broadly divided into: (1) anterior mediastinal; (2) posterior mediastinal; and (3) tracheobronchial. The latter can be further subdivided into the following regions: (a) right and left paratracheal; (b) subaortic; (c) pretracheal; (d) subcarinal. It is important to appreciate that the absolute size of lymph nodes identified on CT (or by direct inspection at mediastinoscopy) should not be regarded as a foolproof criterion for significant disease, particularly in the context of lung cancer. Although markedly enlarged lymph nodes, greater than 2 cm in diameter, almost invariably signify important pathology, moderate enlargement of mediastinal lymph nodes may represent reactive hyperplasia of little clinical significance. Conversely, small-volume lymph nodes or lymph nodes not identified by CT may sometimes contain micrometastases from a distant primary neoplasm.

The thymus gland occupies a large part of the anterior mediastinum in children. In adult life the remnants of the normal thymus are normally inconspicuous on CT.

Points in the interpretation of a chest radiograph

Even when there is an obvious radiographic abnormality, there is much to recommend a careful and systematic method in reviewing a chest radiograph. Such an approach will allow an appreciation of normal variations of anatomy to be built up with time. With increasing experience an appreciation of a deviation from normal appearances becomes more rapid, which leads quickly to a directed search for related abnormalities.

Before interpreting a chest radiograph, it is vital to establish whether there are any previous radiographs for comparison: the sequence and pattern of change is often as important as the identification of a radiographic abnormality. Information gained from preceding radiographs, particularly the lack of serial change, will often prevent needless further investigation. Demographic details, particularly the age and racial origin of the patient, should be noted, since this information may increase the probability of a differential diagnosis which is based on the radiographic findings alone.

A quick check that the radiograph is of satisfactory quality includes an estimation of the radiographic exposure, depth of inspiration, and position of the patient. As a general rule, the intervertebral disc spaces of the entire dorsal spine should be visible on a correctly exposed radiograph, with the mid-point of the right hemidiaphragm lying at the level of the anterior end of the sixth rib if the patient has taken a satisfactory breath in. The patient is axially rotated if the medial ends of the clavicles are not equidistant from the spinous process of the thoracic vertebral body at that level.

The order in which the structures on a chest radiograph are analysed is unimportant. A suggested sequence is to start with a scrutiny of the position of the trachea, mediastinal contour (which should be sharply outlined in its entirety), and then the position, outline, and density of the hilar shadows. Only then are the lungs examined, taking into account their size, the relative transradiancy of each zone, and the position of the horizontal fissure (and any other indirect signs of volume loss—see later section on lobar collapse). Pulmonary vessels are seen as far as the outer third of the lung and the number of vessels should be roughly symmetrical on the two sides. Next, the position and clarity of the hemidiaphragms should be noted, followed by an assessment of the ribs and soft tissues of the chest wall. Special care should be taken to look for pleural thickening along the lateral chest walls which can easily be overlooked.

Before saying that a chest radiograph is normal, it is worth reviewing areas which are either poorly demonstrated or often misinterpreted. These include: (1) the central mediastinum, where even a large mass may be barely visible on the PA view; (2) the areas behind the heart and hemidiaphragms; (3) the lung apices, often obscured by overlying clavicle and ribs; and (4) the lung and pleura just inside the chest wall.

Once a radiographic abnormality has been detected it should be considered in terms of gross pathology. Both the site and the radiographic characteristics of the lesion will allow the observer to proceed to, at the very least, a generic diagnosis. A precise (unique) diagnosis can only rarely be achieved from the radiographic appearances alone without knowledge of the clinical context.

Common radiological signs of disease

Pulmonary consolidation

Consolidation is a pathological description of the state of the lungs when the normal air-filled spaces, distal to the bronchi, are occupied by the products of disease (for example, water, pus, or blood). The most important radiographic signs of pulmonary consolidation are: (1) an area of increased opacification in the lungs which obscures the underlying blood vessels and has a poorly defined margin—unless it is bounded by a fissure; (2) an 'air bronchogram'; and (3) the 'silhouette sign' (Fig. 13). The air bronchogram is a distinctive and certain sign of intrapulmonary pathology and is seen as a radiolucent (grey) branching structure of the bronchi against a more opaque (white) background of an air-less lung. Although an air bronchogram is seen almost invariably in consolidation, a lung which has become collapsed and

air-less, for example due to a large surrounding pleural effusion, may also show an air bronchogram. The silhouette sign is seen when the normally clear border of a structure is lost because the air-filled lung outlining the border is replaced by fluid or a mass. Recognition of this sign can help to localize the area of abnormality within the lungs, for example consolidation in the lingula will make the left heart border indistinct. As with the air bronchogram sign, the silhouette sign may be seen in either pulmonary consolidation or collapse, for example loss of a clear right heart border may be due to right middle lobe consolidation with or without lobar collapse: the common feature is loss of normal aeration of the affected lung. The causes of widespread pulmonary consolidation are numerous but may be broadly divided into five categories shown in Table 2.

Fig. 13 Widespread pulmonary consolidation in a patient with alveolar proteinosis. The right heart border is obscured, confirming that a large part of the consolidation is in the right middle lobe (the silhouette sign).

Pulmonary collapse

This is the term used to describe the loss of aeration and therefore inflation in part or all of a lung. Depending on the cause, collapse may occur at any level from small, subsegmental areas of lung through to an entire lung. Small areas of subsegmental collapse occur very commonly in debilitated and postoperative patients, where they are seen as linear, usually horizontal, opacities. At the other end of the spectrum, collapse of an entire lung, usually due to an endobronchial lesion or inhaled foreign body, has a dramatic radiographic appearance with complete opacification of the affected lung and loss of volume of that hemithorax. At the lobar level, the signs of collapse of an individual lobe are characteristic, but, depending on the lobe, may be very subtle. Recognition of the collapse of individual lobes is important and these are described in detail.

Collapse of individual lobes

- *Right upper lobe*: on the frontal radiograph there is elevation of the minor fissure and of the right hilum. If the collapse is complete the non-aerated lobe is seen as a density alongside the superior mediastinum (Fig. 14). On the lateral view the minor fissure moves upwards and the major fissure moves forwards. The retrosternal area becomes progressively more opaque and the anterior margin of the ascending aorta becomes obscured.

Fig. 14 Right upper lobe collapse.

- *Right middle lobe*: on the frontal radiograph the lateral part of the minor fissure moves down. There is blurring of the normally sharp right heart border, which may be a subtle abnormality that is easily overlooked (Fig. 15). On the lateral view the minor fissure moves downwards and the lower half of the major fissure moves forwards, giving rise to a triangular shadow with its apex at the hilum and its base behind the lower sternum.

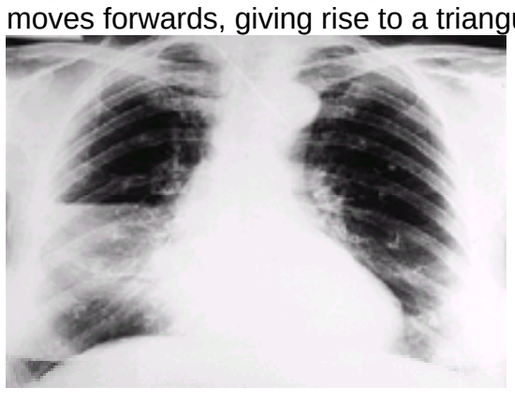

Fig. 15 Right middle lobe collapse.

- *Right lower lobe*: there is an increase in density overlying and obscuring the medial portion of the right hemidiaphragm, and the right hilum is displaced inferiorly on the frontal radiograph (Fig. 16). By contrast to right middle lobe collapse, the right heart border usually remains sharply defined since this is in contact with the aerated right middle lobe. On the lateral view the major fissure moves backwards and downwards; with increasing collapse there is a loss of definition of the posterior part of the right hemidiaphragm as well as increased density overlying the lower dorsal vertebral column.
- *Left upper lobe*: the main finding on the frontal radiograph is a veil-like increase in density, without a sharp margin (quite unlike right upper lobe collapse), spreading outwards and upwards from the elevated left hilum (Fig. 17). The outlines of the aortic knuckle, left hilum, and left heart border become ill-defined. As the collapse increases, the lobe moves centrally and the apical segment of the left lower lobe expands to fill the space left by the collapsed upper lobe: this is the cause of the relatively transradiant lung apex. With complete left upper lobe collapse, a sharp border may return to the aortic arch because it is surrounded by the hyperinflated apical segment of the lower lobe. On the lateral view the major fissure moves superiorly and anteriorly while remaining relatively vertical and roughly parallel to the anterior chest wall.

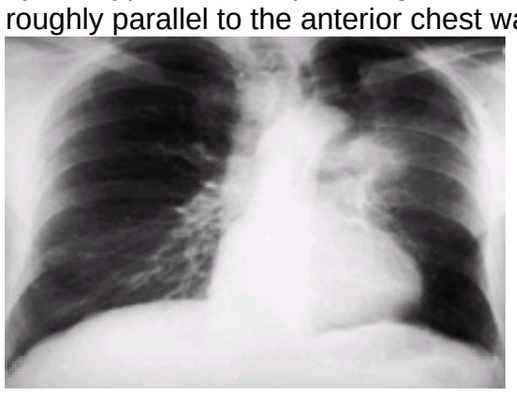

Fig. 17 Left upper lobe collapse.

- *Left lower lobe*: on the frontal radiograph there is a triangular density behind the heart with loss of the medial part of the left hemidiaphragm (Fig. 18). Even on a properly exposed radiograph it may be difficult to appreciate the collapsed lobe behind the heart. Supplementary signs include inferior displacement of the left hilum, loss of volume and increased transradiancy of the left hemithorax. On the lateral view there is posterior displacement of the major fissure. As with right lower lobe collapse, there is increased density over the lower dorsal vertebral column and the posterior part of the left hemidiaphragm is effaced.

Fig. 18 Left lower lobe collapse (an AMBER chest radiograph which improves exposure in the mediastinal region).

Complete opacification (or a white-out) of a hemithorax is generally due to either collapse of a lung or a large pleural effusion or tumour. Shift of the mediastinum to the affected side implies that volume loss, that is to say collapse of the lung, has occurred. By contrast, a pleural effusion or soft tissue mass which is large enough to cause complete opacification of a hemithorax will almost invariably displace the mediastinum away from the side that is opacified. An important exception is an advanced mesothelioma which may encase one lung and 'freeze' the mediastinum and prevent a contralateral mediastinal shift. Occasionally, when there is no obvious shift of the mediastinum, it is surprisingly difficult to differentiate between these two completely different causes of an opacified hemithorax. In these instances, ultrasonography and computed tomography allow the distinction to be made with confidence and may give further information about the underlying disease.

Increased transradiancy of a hemithorax

There are many causes of an increased transradiancy (darkening) of one lung, ranging from a loss of soft tissues of the chest wall (for example, a mastectomy) through to reduced perfusion of one lung due to hypoxic vasoconstriction resulting from underventilation of the lung because of an inhaled foreign body, or a tumour in a main bronchus. It is surprisingly easy to overlook this important radiographic abnormality, especially when the density difference between the two lungs is slight. A subtle discrepancy in density between the two hemithoraces is more readily appreciated by viewing the radiograph from a distance of at least 1.5 metres. The commonest causes of a relatively transradiant hemithorax are shown in Table 3. Close scrutiny of the chest radiograph will usually indicate which one of the categories of causes is responsible for this radiographic sign. If there is any clinical suggestion that the cause of the increased transradiancy is due to an obstructing lesion in a central airway, a chest radiograph taken in full expiration will accentuate the increased transradiancy and will show that the lung fails to empty.

Once it has been established that the difference in density of the lungs is not due to a technical problem, for example rotation of the patient, points to look for are: (1) loss of symmetry of the soft tissues of the chest wall; (2) discrepancy in the volumes and vascular pattern between the two lungs; (3) a visceral pleural edge (denoting a pneumothorax). The identification of a pneumothorax on an erect chest radiograph is usually straightforward because of the appearance of the collapsed lung which is clearly demarcated by the fine edge of the visceral pleura. However, in the supine patient, such an edge is often not seen because air in the pleural space drifts anteriorly to the least dependent part of the chest. In this situation, a pneumothorax is only seen as a vague area of increased transradiancy over the lower zone of the chest. It is vital to recognize when the pressure of the air trapped in the pleural space exceeds alveolar pressure, a so-called tension pneumothorax. The typical signs are of a contralateral mediastinal shift with straightening and flattening of the ipsilateral dome of the diaphragm (Fig. 19).

Fig. 19 A left-sided tension pneumothorax in a patient with cystic fibrosis. Note the mediastinal shift and straightening f the left hemidiaphragm.

The pulmonary mass

Many pulmonary masses are discovered incidentally on a chest radiograph. Whenever possible, previous films should be obtained so that the growth rate of the lesion can be estimated. The growth rate is a more reliable indicator of the likely nature of a pulmonary mass than any one of its radiographic features: if a lesion doubles in volume (increases in diameter by approximately 25 per cent on serial chest radiographs) in less than 1 week or more than 18 months, it is very unlikely to be malignant. The doubling time of most malignant lesions is between 1 and 6 months.

Over the years much importance has been attached to the radiological characteristics of a solitary pulmonary mass in an attempt to make the crucial distinction between benign and malignant lesions. With the possible exception of heavy calcification within the lesion (most commonly seen in ancient granulomas), no radiological appearance will reliably differentiate a benign from a malignant mass. Although generalizations can be made (for example, that bronchial carcinomas have irregular and spiculated margins, whereas benign lesions are more likely to have smooth outlines), in the individual patient it is not safe to rely on these radiographic features alone to make the distinction between a benign and malignant lesion.

After the discovery of a pulmonary mass on chest radiography, further imaging and other investigations of a patient will depend on the symptomatology, age, and smoking history of the patient. Computed tomography is valuable in evaluating extension of a central mass into the mediastinum (Fig. 20); for demonstrating the presence or absence of enlarged mediastinal lymph nodes, which may (but not invariably) indicate local tumour spread; and also for the detection of distant metastases, for example to the contralateral lung, adrenal glands, and liver. It is usually the overall pattern and extent of disease on a staging CT examination, rather than any single abnormality, which indicates whether a patient with bronchial carcinoma, who is otherwise fit, is likely to be suitable for surgical resection. Local invasion of the chest wall by an adjacent bronchial carcinoma is not always demonstrated by CT: MRI, because of its ability to image in different planes, may be useful. When surgery is not indicated and a histological diagnosis is needed, percutaneous needle biopsy of central lesions can be performed safely under CT guidance. Similarly, smaller peripheral lesions that are not accessible by bronchoscopy may be biopsied under CT or, if abutting the pleura, ultrasound guidance.

Fig. 20 Computed tomography of a central cavitating bronchial carcinoma showing direct extension of the tumour into the subcarinal region of the mediastinum.

Cavitating pulmonary lesions

The radiological definition of cavitation is a lucency, representing air, within a mass or area of consolidation. The cavity may or may not contain a fluid level or an intracavitary body, and is surrounded by a wall of variable thickness. The two most likely diagnoses in an adult presenting with a cavitating pulmonary mass on chest radiography are bronchial carcinoma (central, large, and often squamous in type) (Fig. 21) or a lung abscess (usually peripheral and sometimes multiple). Cavitation is recognized in a variety of bacterial pneumonias, particularly those due to tuberculosis, staphylococcus, anaerobes, and klebsiella infections. Less commonly, cavitation is seen within pulmonary infarcts and in areas of pulmonary contusion due to trauma. Long-standing cavities in lungs scarred by previous tuberculosis infection predispose to the formation of mycetomas; once these fungus balls occupy most of the cavity, a characteristic translucent 'air-crescent sign' may be seen between the upper surface of the fungus ball and the margin of the cavity on chest radiography (Fig. 22).

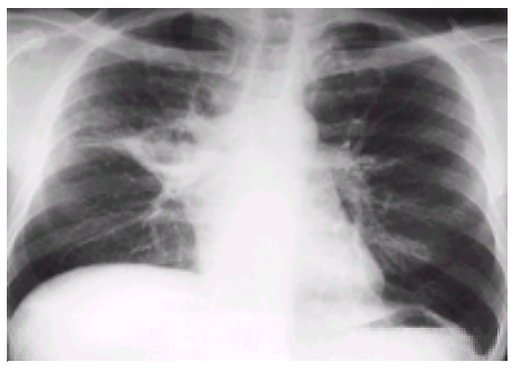

Fig. 21 Chest radiograph of a large cavitating squamous-cell bronchial carcinoma adjacent to the right hilum. The right hemidiaphragm is raised because of phrenic nerve invasion by the tumour.

Fig. 22 An air crescent (arrow) around a fungus ball at the left apex. This had developed in a tuberculous fibrotic cavity.

Multiple pulmonary nodules

Many conditions are characterized by multiple small pulmonary nodules (Fig. 23). Only by combining the relevant clinical information with a precise description of the size and distribution of the nodules can the differential diagnosis be narrowed. In the United Kingdom, metastatic deposits are by far the commonest cause of multiple pulmonary nodules of varying sizes in an adult. In some parts of the southern United States, histoplasmosis is endemic and multiple granulomatous nodules are commoner than those due to disseminated malignancy. In the absence of a known malignancy and when clinical findings and laboratory investigations are inconclusive, biopsy of one of the nodules may be the only means of establishing a diagnosis.

Fig. 23 Multiple pulmonary nodules of varying sizes typical of metastatic disease.

A myriad of small nodules, less than 5 mm in diameter, produces a pattern which is often described as miliary. A list of causes of miliary shadowing is given in Table 4. An important diagnosis to consider in any patient with this radiographic pattern is miliary tuberculosis. Other differential diagnoses in an asymptomatic patient with numerous pulmonary nodules include sarcoidosis, metastatic disease, or, if there is a relevant occupational history, a pneumoconiosis. As always, comparison with previous radiographs will give invaluable information about the rate of progression and thus the likely nature of the pulmonary nodules. To a lesser extent the distribution of nodules is a consideration in refining the differential diagnosis of multiple pulmonary nodules: for example, the small nodules of pulmonary sarcoidosis

tend to be mid-zone and perihilar, whereas haematogenous metastases are generally of varying sizes and have a predilection for the lower lobes (probably because of increased blood flow to these regions).

The density of nodules sometimes provides conclusive evidence that the nodules are of benign aetiology, for example the heavily calcified nodules which are seen following histoplasmosis or chickenpox (varicella) pneumonia. The majority of multiple pulmonary nodules are of soft tissue density, and it may be extremely difficult to judge whether small nodules are of calcific or soft tissue density because their apparent density depends so critically on the radiographic technique used.

Numerous poorly defined nodules of low density of approximately 8 mm in diameter may be seen around areas of pulmonary consolidation. In other areas they may be confluent and so make up a larger poorly defined opacity; occasionally these nodules will be uniformly distributed through out the lungs. At a pathological level these nodules correspond to individual acini which are full of the products of disease, such as pulmonary oedema, an inflammatory exudate, or haemorrhage.

Further reading

Armstrong P, *et al*. (2000). *Imaging of diseases of the chest*, 3rd edn. Mosby Year Book, St Louis.

Austin JHM, *et al*. (1996). Glossary of terms for CT of the lungs: recommendations of the Nomenclature Committee of the Fleischner Society. *Radiology* **200**, 327–31.

Engeler CE (200q). Interpreting the chest radiograph. In: Grainger RG, Allison DJ, Adam A, Dixon AK, eds. *Diagnostic radiology: a textbook of medical imaging*, 4th edn, pp. 303–14. Churchill Livingstone, Edinburgh.

Fleischner Society (1984). Glossary of terms for thoracic radiology: recommendations of the Nomenclature Committee of the Fleischner Society. *American Journal of Roentgenology* **143**, 509–17.

Goodman LR (1999). *Felson's principles of chest roentgenology*, 2nd edn. WB Saunders, Philadelphia.

Heitzmann ER (1988). *The mediastinum: radiologic correlations with anatomy and pathology*, 2nd edn. Springer-Verlag, Berlin.

Lowe VJ, Naunheim KS (1998). Current role of positron emission tomography in thoracic oncology. *Thorax* **53**, 703–12.

Mathieson JR, *et al*. (1989). Chronic diffuse infiltrative lung disease: comparison of diagnostic accuracy of CT and chest radiography. *Radiology* **171**, 111–16.

Müller NL (1991). Clinical value of high resolution CT in chronic diffuse lung disease. *American Journal of Roentgenology* **157**, 1163–70.

Naidich DP, *et al*. (1999). *Computed tomography and magnetic resonance of the thorax*, 3rd edn. Lippincott-Raven, Philadelphia.

Proto AV, Speckman JM (1979). *The left lateral radiograph of the chest. Medical Radiography and Photography* **56**, 38–64.

Rémy-Jardin M, Rémy J (1996). *Spiral CT of the chest*. Springer-Verlag, Berlin.

Webb RW, Müller NL, Naidich DP (2001). *High-resolution CT of the lung*, 3rd edn. Lippincott Williams and Wilkins, Philadelphia.

17.3.2 Respiratory function tests

G. J. Gibson

Scope of respiratory function tests

The clinical roles of respiratory function tests include diagnosis, assessment of severity, monitoring the effects of treatment, and assessing prognosis of various respiratory conditions. In the diagnosis of specific diseases, respiratory function tests, like functional tests of other organs, inevitably have limitations. Their use as a diagnostic tool is in recongizing that different patterns of abnormality characterize particular types of disease. More often they are used to quantify the severity of functional disturbance or to locate the likely anatomical site(s) of disease, such as airways, alveoli, or chest wall. The results should be compared with reference values obtained in healthy populations (see Appendix) and should always be evaluated in the light of clinical and radiographic information. The commonly applied tests are most conveniently classified as tests of (1) respiratory mechanics, (2) pulmonary gas exchange (and acid–base balance), and (3) exercise. Measurements made during sleep are described elsewhere (see Chapter 17.8.2).

Tests of respiratory mechanics

The volume of air in the lungs at the end of tidal expiration (functional residual capacity— **FRC**) represents the 'neutral' volume of the thorax, that is, the volume pertaining when the respiratory muscles are inactive (as during anaesthesia with muscle paralysis). Expansion of the lungs above FRC is achieved by contraction of the inspiratory muscles (predominantly the diaphragm), which results in a negative (subatmospheric) alveolar pressure. Normal resting tidal expiration is essentially passive with the driving force provided by elastic recoil of the lungs. The main expiratory muscles are those of the abdominal wall, contraction of which increases abdominal pressure which in turn is transmitted to the thorax. In health these muscles become active when ventilation is increased markedly, as on exercise, or during coughing, when a high intrathoracic pressure aids the clearance of airway secretions.

Measurements of tidal breathing (tidal volume, respiratory frequency) are rarely made in the resting awake subject, other than conventional recording of respiratory rate, which is part of clinical examination. Measurement of ventilation is of more importance in patients receiving ventilatory support (such as in intensive care units), during detailed exercise testing, and during sleep investigations. During exercise testing, ventilation is usually obtained by electrical integration of airflow measured at the mouth, but this approach is impracticable for prolonged monitoring (such as during sleep) and the application of a mouthpiece and nose clip may itself disturb the pattern of resting breathing. An alternative, less intrusive method is to measure external movement of the chest wall (rib cage and abdomen), but the estimates of ventilation obtained are at best semiquantitative.

In principle, the mechanical function of the respiratory system can be characterized by the compliance of the lungs and chest wall and the resistance of the airway. In practice, however, neither of these is commonly measured directly in clinical testing. Measurement of pulmonary compliance (an index of the 'stiffness' of the lungs) requires measurement of oesophageal pressure, which equates to pleural pressure and allows calculation of the pressure required to distend the lungs. In clinical investigation, the elastic properties of the lungs are usually inferred from measurements of lung volumes, because lungs which are unusually stiff and poorly compliant (as in pulmonary fibrosis) are usually shrunken and reduced in volume, while lungs with abnormally high compliance are easily distensible and are associated with increased total lung capacity (as in emphysema). The traditional subdivisions of lung volume are illustrated in Fig. 1.

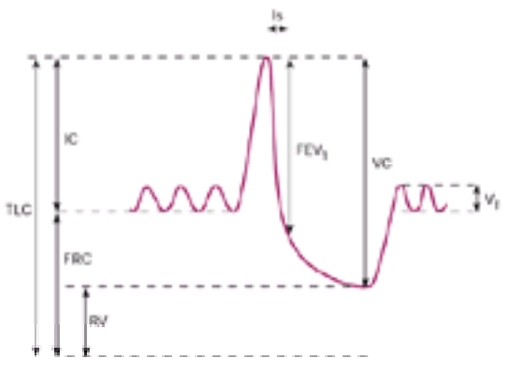

Fig. 1 Subdivisions of lung volume illustrated by spirometric recording of volume against time during tidal breathing for three breaths, followed my maximal inspiration and then maximal forced expiration, before returning to tidal breathing in a normal subject. RV, residual volume; FRC, functional residual capacity; IC, inspiratory capacity; TLC, total lung capacity; FEV_1, forced expiratory volume in 1 s; VC, vital capacity; V_T, tidal volume. Note that TLC = FRC + IC = VC + RV.

Measurement of airway resistance requires estimation of the pressure difference along the airway—between the alveoli and mouth. The various techniques available for obtaining alveolar pressure include oesophageal pressure monitoring, body plethysmography, and transient interruption of airflow, with mouth pressure during occlusion taken to equal alveolar pressure. None of these is widely used in clinical testing. An alternative method, which has recently gained popularity, involves superimposition of a small oscillating pressure at the mouth during tidal breathing; the resulting pressure and flow information is used to calculate airway resistance. In practice, however, airway function is most commonly assessed by tests based on forced expiration.

Measurements of lung volume

A spirometer measures only the air which can be displaced from the lungs and does not give an indication of their absolute volume because the unmeasured residual volume (**RV**) remains in the lungs after full expiration. The vital capacity (**VC**) is the maximum volume expired after a full inspiration (or inspired after a full expiration) and the total lung capacity (**TLC**) represents the volume of air in the lungs after full inspiration— the sum of VC and RV (Fig. 1).

Two main clinical methods are used for measurement of absolute lung volume: inert gas dilution and whole body plethysmography. With the former, the subject breathes from a closed circuit a gas mixture containing an inert marker gas, usually helium. The helium equilibrates gradually with the gas in the lungs so that its concentration falls progressively and stabilizes once mixing is complete. In a healthy individual this occurs in 5 to 10 min, but in patients with diffuse airway disease such as asthma or chronic obstructive pulmonary disease (**COPD**), equilibration is much slower due to unequal ventilation, and the end point may be much less definite. The lung volume measured is that in the lungs when the subject was connected to the circuit (usually FRC). After disconnection from the rebreathing circuit

the subject inspires fully and the volume inspired (inspiratory capacity, **IC**) added to FRC gives TLC (Fig. 1). With moderate or severe airway disease the uneven distribution of the inspired gas and poor mixing in the lungs results in underestimation of lung volumes.

In the alternative plethysmographic technique, the subject sits within a large air-tight chamber and makes gentle breathing efforts against a shutter, which closes the airway at the mouth. Since the pressure within the rigid plethysmograph changes as lung volume changes according to Boyle's law (pressure × volume = a constant), this allows calculation of thoracic gas volume, from which total lung capacity and residual volume are derived by full inspiration and expiration immediately on opening the shutter. This method measures the volume of any air spaces within or without the lung which share pressure changes during breathing efforts, so that poorly ventilated (or even totally unventilated, such as a bulla) areas of lung are included.

Some increase in TLC occurs in most patients with symptomatic diffuse airway obstruction. A large increase is characteristic of emphysema but is not specific for this condition. Increases are seen in asthma, even when the condition is in relative remission. A pathological reduction in TLC occurs in several conditions (Table 1), not only lung diseases such as pulmonary fibrosis, but also extrapulmonary diseases affecting the pleura, thoracic skeleton, or respiratory muscles, conditions which all potentially impede full lung expansion.

Tests of forced expiration

The strengths of tests of forced expiration include the simplicity of both the manoeuvre and equipment required, and also the relative independence of the measurements on the effort applied by the patient. Forced expiratory tests are effort dependent to the extent that a preceding full inspiration is required, but during forced expiration the larger intrathoracic airways are subject to dynamic compression by the surrounding pleural pressure. The net result is that, provided a modest effort is applied, increasing the effort merely compresses the airway further and produces no increase in flow. This effort independence is more marked as forced expiration proceeds and is also more marked in patients with airway obstruction than in healthy subjects. At higher lung volumes (i.e. closer to full inflation), maximum expiratory flow is more dependent on effort. Since peak expiratory flow (**PEF**) is attained very rapidly at the start of forced expiration, it is more effort dependent than the forced expiratory volume in 1 s (**FEV1**), which effectively integrates flow over a large proportion of the volume range. PEF is measurable with a very simple peak flow meter which has the advantage of portability and is used routinely, particularly in asthma, to monitor respiratory function at home. Although more effort dependent, PEF is reproducible by most patients after a few practice efforts.

The most commonly used index of mechanical function of the lungs in hospital is the 1 s forced expiratory volume (FEV_1)—the volume expired forcefully in 1 s following complete inspiration (Fig. 1). This is usually obtained together with the forced vital capacity (**FVC**). In healthy subjects the FVC is effectively the same as VC, but in patients with airway disease the FVC is often appreciably less than the true ('relaxed') VC obtained when the subject is encouraged to expire completely without excessive initial effort.

The characteristic feature of diffuse airway obstruction is slowing of the rate of expiration so that the ratio of FEV_1 to FVC (or FEV_1 to VC) is reduced. This defines an 'obstructive' ventilatory defect as opposed to the 'restrictive' pattern in which both FEV_1 and FVC are reduced in approximate proportion. Although in patients with diffuse airway obstruction the FVC and VC are reduced, at least in patients with symptomatic disease, the reduction is proportionally less than the reduction in FEV_1. The ratio of FEV_1 to FVC indicates the presence of airway obstruction but is a poor guide to severity, which is better assessed by comparing the FEV_1 alone with its predicted value. An obstructive spirometric pattern is seen in asthma, COPD, and bronchiectasis, while a restrictive spirometric pattern is seen in all those conditions which are also associated with a reduced TLC (Table 1). A further feature of diffuse airway obstruction is an increase in RV and in the ratio RV/TLC, but this is less specific than the spirometric pattern as it also occurs in some patients with cardiac disease or with respiratory muscle weakness. With dual pathology, combined obstructive (low FEV_1/VC) and restrictive (low TLC) defects are seen. Sometimes in this situation total lung capacity may be within the normal range due to opposing influences with, for example, lung fibrosis tending to shrink the lungs and airway obstruction tending to produce hyperinflation.

Spirometric measurements such as FEV_1 are less sensitive to localized narrowing of the central airway than to the diffuse airway narrowing of COPD or asthma. If upper airway obstruction is suspected, it is helpful to visualize the information obtained during forced expiration (and also inspiration) in a different manner as the maximum flow–volume curves, which relate instantaneous flow to volume expired and inspired (Fig. 2). The expiratory curve has a characteristic shape with an early peak (equivalent to the PEF obtained with a peak flow meter). Maximum expiratory flow then declines progressively as volume is expired. In young healthy subjects (Fig. 2(a)), the descending limb of the curve approximates a straight line, whilst in older normal subjects (Fig. 2(b)), maximum expiratory flow decreases, particularly at lower lung volumes so that the curve appears concave. In patients with diffuse intrathoracic airway obstruction (such as COPD or asthma) this ageing appearance is greatly exaggerated so that expiratory flow is reduced more markedly as lung volume declines (Fig. 2(d)). The shape of the flow–volume curve does not distinguish between different causes of diffuse airway narrowing, that is, it does not allow the distinction of asthma from COPD or emphysema.

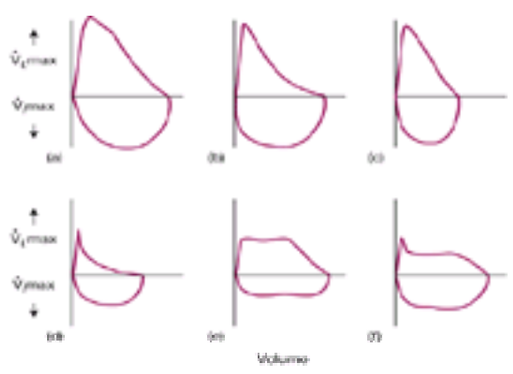

Fig. 2 Schematic maximum expiratory and inspiratory flow–volume curves in: (a) normal young adult; (b) normal older adult; (c) patient with fibrosing alveolitis and reduced FVC; (d) patient with moderately severe COPD showing markedly reduced $_E$max, particularly at lower lung volumes; (e) patient with subglottic (extrathoracic) tracheal stenosis showing markedly reduced $_I$max at all volumes and reduced $_E$max at higher volumes; (f) patient with central intrathoracic (carinal) tracheal narrowing showing similar plateaus of flow to (e) but greater reduction of $_E$max than of $_I$max.

The maximum inspiratory flow–volume curve has a more symmetrical appearance than the expiratory curve. In patients with diffuse airway narrowing there is an overall reduction in inspiratory flow, but little change in shape (Fig. 2(d)). In patients with a restrictive ventilatory defect caused, for example, by pulmonary fibrosis, the volume displaced (FVC) is reduced but absolute flows are little affected (Fig. 2(c)).

Characteristic flow–volume curves are seen in patients with localized narrowing of the proximal airway, with the pattern depending on whether the narrowing is extra- or intrathoracic. Extrathoracic narrowing (Fig. 2(e)), such as occurs with subglottic tracheal stenosis or tracheal tumours, has a relatively greater effect on inspiratory than expiratory flow (which corresponds to the predominantly inspiratory timing of the stridor of upper airway narrowing). It also affects maximum expiratory flow but (unlike COPD or asthma) the effects are most marked at higher lung volumes, often producing a virtual 'plateau' of expiratory flow in the first part of forced expiration. If, on the other hand, the central airway is narrowed within the thorax (for example the lower trachea, carina) a similar plateau of expiratory flow, often with a small initial peak, may be seen, but maximum inspiratory flow is less affected than with narrowing of the extrathoracic airway (Fig. 2(f)). These patterns of abnormality of maximum flow can be quantified in terms of various ratios, such as the ratio of maximum expiratory to inspiratory flow at 50 per cent VC, or the ratio of PEF (markedly reduced with upper airway obstruction) to FEV_1 (proportionally less reduced). Usually, however, it is essential to visualize the curves and interpret such derived indices in the light of the overall contour.

The 'plateau' of maximum expiratory flow over a significant proportion of the FVC which occurs with upper airway obstruction has implications for the shape of the more commonly recorded forced expiratory spirogram in this situation. Since flow on the spirogram (relation of volume to time) is given by the instantaneous gradient of the curve, a plateau on the flow–volume curve implies a 'straight' (rectilinear) spirogram over the same volume range. This appearance should therefore alert the investigator to the likelihood of narrowing of the central airway rather than the more common diffuse airway obstruction seen with asthma and COPD (Fig. 3).

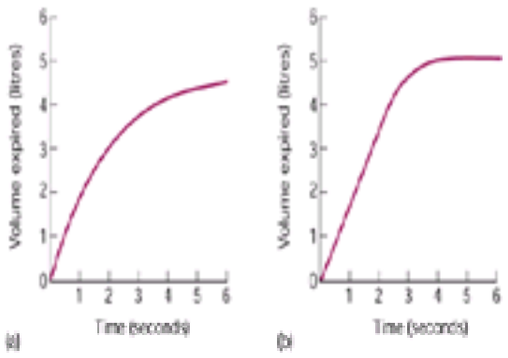

Fig. 3 Spirograms of two patients with airway obstruction and similar FEV_1. (a) Diffuse intrathoracic airway narrowing (COPD or asthma). Note that forced expiration is continuing after 6 s. (b) Upper airway narrowing with 'straight' spirogram which corresponds to plateau of flow in earlier part of expiration.

Respiratory muscle function

The simplest method of measuring respiratory muscle strength is for the subject to perform forcible static inspiratory and expiratory efforts against a closed airway. This provides values of maximum expiratory and inspiratory pressures (***P*Emax**, ***P*Imax**). In general the expiratory (predominantly abdominal) muscles perform most effectively at high lung volumes and the inspiratory muscles (predominantly the diaphragm) at lower volumes. $P_E$max is therefore usually measured after full inspiration and $P_I$max at either FRC or RV. Unfortunately the normal ranges for these tests are wide and some subjects find difficulty in performing the manoeuvres (which are, of course, by definition completely effort dependent). An alternative method for assessing inspiratory muscle strength is by making the measurement during a forceful sniff, with the pressure measured in the nose via an occluded nostril (sniff nasal inspiratory pressure—SNIP). Many patients find this easier than performing maximum static manoeuvres so that the sniff technique tends to give more reproducible results.

These measurements all assess the global strength of the inspiratory or expiratory muscles. More specific information on diaphragmatic function requires measurement of transdiaphragmatic pressure using pressure sensing devices in both oesophagus and stomach, a specialized investigation available in only a few centres. A simple indirect index of disproportionate diaphragmatic weakness or paralysis is a large (more than 25 per cent) reduction in VC in the supine compared with the erect posture. However, isolated bilateral diaphragmatic paralysis or severe weakness is very unusual and most patients with respiratory muscle weakness have disease affecting all the muscles. The causes include not only primary neuromuscular diseases such as myopathies, muscular dystrophy, motor neurone disease, and myasthenia gravis, but also drug treatment (corticosteroids), several endocrine and connective tissue disorders, and cachexia from whatever cause. Respiratory muscle weakness is often an important factor preventing weaning from assisted ventilation.

Measurements of respiratory muscle function are indicated in evaluation of patients with various neuromuscular diseases. They are also helpful in confirming or excluding muscle problems in those with otherwise unexplained dyspnoea and in patients with a restrictive ventilatory defect in whom the cause of the lung volume reduction is not apparent on clinical and radiographic grounds. Interpretation of results may be complicated in patients with airway obstruction (such as COPD or asthma) because the associated hyperinflation of the lungs (increased FRC) itself impairs inspiratory muscle function simply because of the distorted thoracic mechanics. Consequently an apparently impaired maximum inspiratory pressure in such patients may not necessarily reflect true muscle weakness. Maximum expiratory pressure is not significantly affected by hyperinflation, however, and can be used as a guide to the presence of true muscle weakness in this situation.

Tests of pulmonary gas exchange

Carbon monoxide uptake

Carbon monoxide (CO) diffusing capacity or transfer factor (*TL*CO) is used widely as a simple test of the integrity of the alveolar capillary membrane and of the overall gas exchanging function of the lungs. It has good sensitivity but poor specificity, as impairment can result from a variety of pathological processes (Table 2). The subject takes a full inspiration of a gas mixture containing a very low concentration of CO and the rate of uptake of gas is measured during breath holding for 10 s. The test was introduced originally as a method of assessing diffusion of gas across the alveolar capillary membrane, but thickening of the diffusion pathway for carbon monoxide in disease is quantitatively less important than other mechanisms, the most important factor being the effective surface area of alveoli available for gas exchange. Consequently *TL*CO is reduced when this area is diminished, for example after resection of lung or with widespread emphysema, in which normal alveoli are replaced by much larger air spaces. *TL*CO is also reduced when there is loss of the 'effective' alveolar volume (V_A). The latter is measured simultaneously from the dilution of helium which is also included in the test breath. The 'effective' V_A is reduced if there is maldistribution of ventilation as this causes some alveoli to receive little or none of the inspired gas. Other factors affecting the *TL*CO include the availability of haemoglobin and disease of the pulmonary capillaries.

The transfer coefficient (*K*CO), which is obtained along with the *TL*CO, represents the uptake of CO per litre of 'effective' alveolar volume, that is, $K\text{CO} = TL\text{CO} / V_A$. To a large extent, *K*CO allows correction for any real or effective reduction of alveolar volume, tending to be normal after lung resection, when both *TL*CO and V_A are reduced approximately to the same degree. *K*CO is usually normal (or even mildly increased) in asthma, where any reduction in *TL*CO is due only to maldistribution of ventilation secondary to airway narrowing. By contrast, in widespread emphysema, *TL*CO is reduced due not only to maldistribution of inspired gas, but also because the gas exchanging surface area is diminished even in the relatively better ventilated parts of the lung. Consequently, there is an associated reduction in *K*CO. Diseases associated with reductions in *TL*CO and *K*CO are listed in Table 2.

In some conditions *K*CO and, less commonly, *TL*CO may increase (Table 3). The latter usually results from an increase in red blood cells in the lungs due to greater blood flow, haemorrhage, or polycythaemia. *K*CO is similarly increased in these conditions, as it is if, at full inflation, the density of pulmonary capillaries per unit alveolar volume is greater than normal. This occurs most commonly in patients with extrapulmonary volume restriction, when the density of pulmonary capillaries is unusually high in relation to the (restricted) lung volume at which the measurement is made.

Arterial blood gases

The primary measurements made by modern blood gas analysers are the arterial partial pressures of oxygen ($P\text{aO}_2$) and carbon dioxide ($P\text{aCO}_2$), and pH. The alternative commonly used method of assessing oxygenation is by pulse oximetry, which estimates arterial oxygen saturation ($S\text{aO}_2$). An oximeter has the advantage of allowing continuous monitoring but it provides no information on $P\text{CO}_2$. The general relation between oxygen pressure and saturation is defined by the oxygen–haemoglobin dissociation curve (Fig. 4). The position of this curve is influenced by the prevailing pH, temperature, and $P\text{CO}_2$. In addition, several rare abnormal variants of the haemoglobin molecule cause the curve to shift either to the right (reduced oxygen affinity) or the left (increased affinity). Approximate values for normal arterial and resting mixed venous $P\text{O}_2$ and saturation are shown in Fig. 4. Another clinically useful 'landmark' is a saturation of 90 per cent which, with a normally positioned curve, represents a $P\text{O}_2$ of approximately 8 kPa (60 mmHg). Also shown in Fig. 4 is the P_{50}, that is, the $P\text{O}_2$ at a saturation of 50 per cent, which for normal adult haemoglobin is approximately 3.5 kPa (27 mmHg). This is essentially an *in vitro* measurement which is used to characterize abnormal haemoglobin molecules associated with increased (low P_{50}) or decreased (high P_{50}) affinity for oxygen.

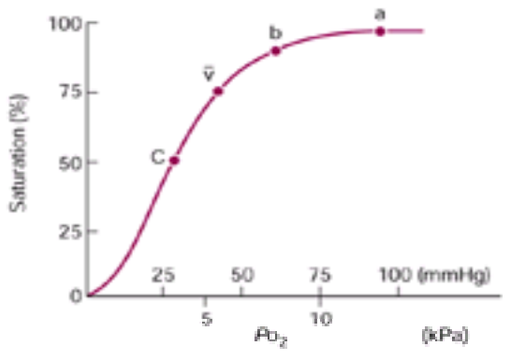

Fig. 4 Normal haemoglobin–oxygen dissociation curve relating saturation to PO_2. Point a represents normal arterial values (PO_2 90 mmHg, 12 kPa; SaO_2 98 per cent) and v normal resting mixed venous values (PvO_2 40 mmHg, 5.3 kPa; SaO_2 75 per cent). Also shown are the PO_2 (~ 60 mmHg, 8 kPa) corresponding to 90 per cent saturation (point b) and the P_{50} (point c), i.e. PO_2 corresponding to 50 per cent saturation (~ 27 mmHg, 3.5 kPa).

A reduction in PaO_2 can occur by various mechanisms (Table 4). In disease, the commonest is mismatching of alveolar ventilation ($_A$) and perfusion (). Even in healthy lungs, distribution of both ventilation and perfusion is uneven. In normal subjects this results mainly from the effects of gravity. In the upright posture, both ventilation and perfusion increase towards the lung bases, but the effects on perfusion are relatively greater, so that the ratio of ventilation to perfusion ($_A$/) is higher towards the apices and lower towards the bases. In disease, these relatively small gravitational effects are outweighed by unevenly distributed pathological changes affecting the distribution of ventilation or perfusion or both. Alveoli with greater than average $_A$/ have higher than average local PO_2 and lower PCO_2, that is, closer to those of inspired air. Conversely, those with lower than average $_A$/ have lower PO_2 and higher PCO_2, that is, closer to the values in mixed venous (pulmonary arterial) blood. The gas tensions in the draining pulmonary capillaries essentially reflect those of the alveoli which they subtend as, within a single alveolus, complete equilibration of local gas tensions usually occurs. For CO_2 the effects of high $_A$/ and low $_A$/ areas on the final arterial value approximately cancel each other out, that is, the $PaCO_2$ is close to the average value in all the capillaries draining the alveoli with a variety of local PCO_2 values. However, for oxygen the situation is different as blood draining alveoli with high $_A$/ (and relatively high local PO_2) cannot compensate for the areas with low $_A$/ (and low PO_2) because of the shape of the oxygen dissociation curve. The relatively flat upper part of the curve means that increasing PO_2 adds very little to oxygen saturation and therefore to oxygen concentration. Consequently, mixed pulmonary venous (and therefore systemic arterial) blood has an appreciably lower PO_2 than would be found in mixed alveolar air.

An approximate assessment of the overall effects of $_A$/ mismatching on arterial oxygenation and PaO_2 is given by calculation of the alveolar to arterial oxygen pressure gradient ($P(A\text{-}a)O_2 = PAo_2 - PaO_2$). This requires estimation of the average alveolar PO_2 (PAO_2) which is determined by the inspired PO_2 (PIO_2) and the average alveolar PCO_2 ($PACO_2$). For the reasons discussed above, alveolar and arterial PCO_2 (unlike PO_2) are virtually the same and the alveolar PO_2 is given simply by:

$$PAO_2 = PIO_2 - PaCO_2/0.8$$

The PIO_2 breathing room air at sea level is 20 kPa (150 mmHg). In normal young subjects the upper limit for $P(A\text{-}a)O_2$ is 2.5 kPa. It rises with age and in healthy subjects aged 60 to 70 years may be up to 4.7 kPa (35 mmHg). Unfortunately interpretation of the $P(A\text{-}a)O_2$ is complicated by the fact that its relation to the severity of $_A$/ mismatching is not constant. For a given degree of $_A$/ mismatching, the $P(A\text{-}a)O_2$ increases as the alveolar PO_2 increases. It therefore increases if the inspired oxygen is increased or if $PaCO_2$ falls (see Equation 1).

Alternative indices which relate more predictably to the degree of $_A$/ mismatching are the ratios of arterial to alveolar PO_2 (a/A PO_2) and of arterial PO_2 to the inspired oxygen fraction (PaO_2/FIO_2). The former is normally greater than 0.75 and changes little as FIO_2 increases, whereas the more traditional $P(A\text{-}a)O_2$ difference increases. The ratio of PaO_2/FIO_2 is widely used in assessment of patients with severe problems of oxygenation. For example, in acute lung injury a value greater than 300 (PaO_2 in mmHg, FIO_2 as a fraction) indicates relatively mild hypoxaemia, whilst a value of less than 100 represents very severe disturbance of gas exchange.

The dependence of $P(A\text{-}a)O_2$ on inspired oxygen is exemplified by the effects of breathing pure oxygen. This is used as a test for the presence of anatomical right to left shunting, since the effects of $_A$/ mismatching on PaO_2 are effectively eliminated by breathing pure oxygen. Even in diseased lungs, nitrogen is gradually 'washed out' of all the alveoli and the only remaining cause of arterial hypoxaemia is the anatomical shunt via channels which bypass the lungs, or through the capillaries supplying any alveoli which are totally unventilated. Although prolonged breathing of 100 per cent oxygen may itself encourage alveolar atelectasis which would exaggerate the shunt, in practice the technique is useful in investigation of causes of hypoxaemia. The usually quoted normal upper limit for the 'anatomical' shunt measured in this way is 5 per cent of the cardiac output. In terms of the PaO_2, a value greater than 500 mmHg (more than 73 kPa) is usually achieved, representing a $P(A\text{-}a)O_2$ of more than 100 mmHg, which greatly exceeds the normal upper limit when breathing room air. At these levels of PaO_2 haemoglobin is virtually fully saturated with oxygen and increasing the PaO_2 above 200 to 300 mmHg results in greater oxygen carriage by simple solution only. Consequently, increases in oxygen content and PO_2 become linearly related on the 'flat' part of the dissociation curve. As a rule of thumb, with a PaO_2 greater than 300 mmHg, each 20 mmHg of $P(A\text{-}a)O_2$ represents an anatomical shunt of 1 per cent.

Respiratory failure

Respiratory failure is defined conventionally in terms of the arterial blood gas tensions as a reduction in PaO_2 below 8 kPa (60 mmHg) at sea level, either without ('type I') or with ('type II') CO_2 retention. Hypercapnic (type II) respiratory failure is also known as ventilatory failure. The causes of type I respiratory failure are legion and include virtually all diseases which can affect the alveoli or the airways, either primarily or secondarily (as in cardiac failure). Hypercapnic (type II) respiratory failure is most commonly due to severe chronic airway disease and less often to reduced ventilation as, for example, with severe respiratory muscle weakness or scoliosis. The mechanisms of elevation of $PaCO_2$ in type II respiratory failure are twofold. Sustained 'pure' hypoventilation, that is, a reduction in overall ventilation resulting in hypercapnia, is rare. It is seen with inadequate performance of the respiratory 'bellows', for example in neuromuscular disease or because of reduced drive to breathe in the unconscious subject. Much more commonly in chronic airway disease, the 'effective' alveolar ventilation is reduced as a consequence of mismatching of ventilation and perfusion. In this situation there is often a considerable amount of ineffectual or wasted ventilation ('physiological dead space') and consequently in such patients the total ventilation is often greater than normal, even in the presence of hypercapnia.

Acid–base balance

The carriage of CO_2 by the blood and its excretion by the lungs constitute one of the two homeostatic mechanisms for regulating the acid–base status of the body. Because of the ease with which CO_2 excretion can normally be increased, the lungs are able to adjust acid–base balance much more rapidly than the kidneys.

The concentrations of hydrogen ions, bicarbonate, and carbonic acid in the blood are linked inevitably by the carbonic acid association/dissociation equation:

$$CO_2 + H_2O \text{ ?2; } H_2CO_3 \text{ ?2; } H^+ + HCO_3^-$$

This defines the chemical relation between the three variables, pH, PCO_2, and $[HCO_3^-]$ and if two are measured, the third is readily calculated. Abnormal acid–base disturbances are classified in terms of these variables as one of four classic types (Table 5 and Fig. 5), but combined disturbances are frequently seen. The commoner causes of acid–base disturbance are given in Table 6.

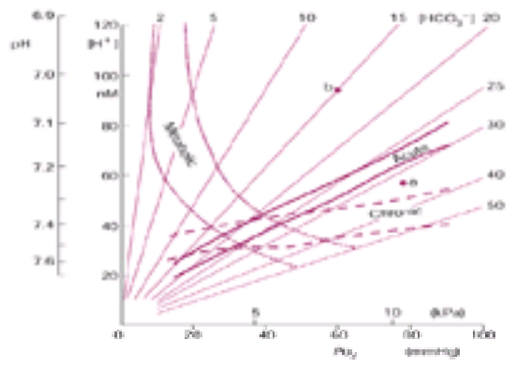

Fig. 5 Relations of pH and [H^+] to PCO_2 in acid–base disorders. Bands indicate the expected ranges in uncomplicated respiratory disorders (acute and chronic) and in metabolic acidosis and alkalosis. Isopleths represent corresponding estimates of arterial [HCO_3^-] (mmol/l). Values outside these bands indicate intermediate or combined disturbances. For example: patient a with an acute exacerbation of COPD has an 'acute on chronic' respiratory acidosis ($PaCO_2$ 10.6 kPa, pH 7.24, [HCO_3^-] 34 mmol); patient b with both respiratory and circulatory failure has a combined respiratory and metabolic acidosis ($PaCO_2$ 8 kPa, pH 7.04, [HCO_3^-] 15 mmol).

Respiratory acidosis and alkalosis

In respiratory acidosis the prime event is accumulation of CO_2 due to inadequate or ineffective ventilation. This causes the equilibrium of Equation 2 to shift to the right, generating both hydrogen and bicarbonate ions. The immediate increase in bicarbonate concentration is dictated by this chemical relationship and not by the physiological response, which occurs later. The vast majority of hydrogen ions produced are buffered by proteins with the result that the measured rise in [HCO_3^-] is actually very much greater than the measured increase in hydrogen ion concentration. (The logarithmic pH scale is deceptive in that both the hydrogen ion concentration in the blood and the changes which occur in disease are extremely small. For example, a normal pH of 7.4 represents a [H^+] of 40×10^{-9} mol, whereas [HCO_3^-] is measured to 10^{-3} mol, i.e. the hydrogen ion concentration is approximately 1 millionth of the concentration of bicarbonate). Conventionally the effects of acute respiratory acidosis are distinguished from the 'chronic' respiratory acidosis which results after several hours or days. This follows renal retention of even more bicarbonate, which in turn tends to correct the pH towards normal (Fig. 5).

In respiratory alkalosis the primary event is an increase in CO_2 excretion resulting from increased ventilation, so that both [HCO_3^-] and hydrogen ion concentrations fall (pH rises). Again, most of the reduction in hydrogen ion concentration is buffered. It is less useful to distinguish acute and chronic respiratory alkalosis than it is to distinguish acute and chronic forms of respiratory acidosis.

Metabolic acidosis and alkalosis

In metabolic acidosis, [H^+] rises (pH falls) and [HCO_3^-] falls. The physiological response is so rapid that acute and chronic phases are not distinguishable. Any tendency for $PaCO_2$ to rise (equilibrium of Equation 2 shifted to the left) is more than offset by the increased drive to breathe resulting from production of acid, and the measured effect is a reduction in $PaCO_2$.

In metabolic alkalosis there is an increase in [HCO_3^-] and a reduction in [H^+] (pH increases). The measured result is somewhat variable as opposing influences are involved; any increase in $PaCO_2$ tends to stimulate breathing but the reduced acidity tends to inhibit it. In subjects with healthy lungs, the net effect is often maintenance of $PaCO_2$ in the high normal range, unless the alkalosis is profound (as, for example, is seen with pyloric stenosis and severe depletion of acid). However, in patients with chronic airway disease and either pre-existing or incipient hypercapnia, a more marked increase in $PaCO_2$ is frequently seen. This is particularly relevant to patients with COPD receiving treatment with diuretics and corticosteroids, both of which tend to produce a metabolic alkalosis.

Several other indices of acid–base status have their advocates. Standard bicarbonate, base excess and deficit, and total buffer base are often derived when blood gases are measured by automated equipment. They are obtained by titration of the blood *in vitro* to specified standard values of pH and/or PCO_2. As such, they are open to the very real objection that the results differ from those which would be obtained if the same titration could be performed *in vivo*, where the extracellular fluid, and not just the blood, participates in buffering. Indices such as standard bicarbonate and base excess are used mainly to distinguish 'respiratory' and 'metabolic' components of an acid–base disturbance, but in this context the 'metabolic' component includes renal compensation for a primary respiratory disturbance. Consequently, in a respiratory acidosis an increased standard bicarbonate indicates some degree of chronicity.

One further frequently available index of acid–base status is the venous 'bicarbonate' (strictly total CO_2 content) which is often obtained routinely when electrolytes are measured. A raised value is seen with primary metabolic alkalosis, but in patients with respiratory disease it may be a useful clue to unsuspected ventilatory failure.

Exercise testing

Exercise tests allow observation of patients and their performance at a time when symptoms are present. This can be useful in assessing breathlessness as the meaning of the term varies among patients: to some it means excessive ventilation, to others difficulty in breathing because of airway narrowing, while some interpret the sensation of myocardial ischaemia as breathlessness. Formal quantification of exercise performance also provides objective assessment of disability and an exercise test may give useful information on the likely factors limiting exercise in that individual.

In healthy subjects, ventilation and cardiac output increase progressively with oxygen consumption. Oxygen uptake (O_2) increases with work rate, but at high levels of exercise anaerobic respiration increases with generation of lactate. Initially, CO_2 output increases in proportion to oxygen consumption until increasing anaerobic metabolism results in disproportionate production of CO_2. Measurement of an 'anaerobic threshold' is favoured by some investigators, but the criteria used for its identification are not universally agreed.

The maximum oxygen consumption or maximum aerobic capacity of a healthy subject is determined by the ability of the circulation to supply oxygen to exercising muscle, rather than by the maximum ventilation which can be achieved. In patients with pulmonary disease, particularly airway disease, the maximum attainable ventilation is reduced, approximately in proportion to the abnormality of pulmonary mechanics. This may then determine exercise capacity, although circulatory factors and deconditioning also contribute in many patients and dominate in some.

Exercise tests vary considerably in complexity and in the number and types of measurements made. Simple self-paced tests of walking distance, most commonly in 6 min, aim to mimic the real life situation and are widely used for global assessment of disability. However, such tests are insensitive to mild disease and there is a significant learning effect, as well as dependence on motivation and encouragement. In the shuttle walk test the subject increases his walking speed each minute, giving results which are more reproducible and closer to laboratory-based tests of maximum performance. More formal testing involves exercise on a bicycle ergometer or treadmill. Usually the workload is increased by a constant amount, with periods of 1 to 3 min at each level. Measurements include heart rate, ventilation, and gas exchange (O_2 and CO_2) and oxygen saturation by pulse oximetry. The subject exercises at increasing loads until no longer able to continue because of discomfort, or until stopped by the investigator. The maximum oxygen consumption (symptom limited O_2max) is a useful indicator of overall exercise capacity. Comparison of the maximum ventilation and heart rate at the end of progressive exercise with those predicted from spirometric measurements and age, respectively, gives some indication of the likely factor(s) limiting performance. The level of breathlessness at each workload in an incremental test can also be usefully assessed using simple self-rating scales (visual analogue scale or Borg scale). Arterial oxygen desaturation is seen particularly in interstitial lung disease and pulmonary vascular disease and this may be helpful in predicting which patients are likely to benefit from use of ambulatory oxygen.

A common reason for performing an exercise test is to evaluate the main cause of breathlessness and, in particular, to determine whether this is due predominantly to

cardiac or ventilatory abnormalities. If a patient achieves the predicted maximum heart rate during a progressive test (as is seen in normal subjects), it is reasonable to conclude that the limit to further exercise is set by the cardiovascular system. In most respiratory diseases, patients cease exercise with a lower heart rate as more often the limit is set by the maximum ventilation achievable.

The identification of exercise-induced asthma has rather different requirements. During exercise, most subjects with asthma show some degree of bronchodilatation, and in those who develop exercise-induced asthma, bronchoconstriction develops after exercise. Many patients with asthma, of course, become breathless during exercise, but this is not necessarily due to bronchoconstriction. The intensity of exercise necessary to provoke asthma is relatively high and for this reason exercise-induced asthma is relevant mainly to children and young adults. Optimally it is demonstrated after exercise for at least 5 min at a constant rate, chosen to increase ventilation to around 50 per cent maximal or to increase heart rate to around 80 per cent maximal. FEV_1 or peak flow should be measured beforehand and for up to 30 min afterwards.

Miscellaneous tests

Analysis of expired air has traditionally been limited to oxygen and carbon dioxide, but recently attention has turned to other gases which are present in very low concentrations. The concentration of exhaled carbon monoxide has been used for some years as a guide to its inhalation and as a valuable method for confirming non-smoking claims. The measurement can now be made very simply with a portable analyser. Breath carbon monoxide is also increased in non-smoking subjects with asthma, where it appears to be released as a result of airway inflammation. In similar fashion, expired nitric oxide concentration is increased as a consequence of airway inflammation and it has been proposed as a non-invasive way of assessing airway inflammation and its treatment, particularly in those with asthma. Care needs to be taken to avoid contamination of expired air from the bronchial tree with that from the nose and nasal sinuses, which contain higher concentrations.

Further reading

American Thoracic Society and European Respiratory Society (2001). Statement on standardization of respiratory muscle tests. *American Journal of Respiratory and Critical Care Medicine*, in press.

Clark JS *et al*. (1992). Non-invasive assessment of blood gases. *American Review of Respiratory Disease* **145**, 220–32.

Gibson GJ (1996). *Clinical tests of respiratory function*, 2nd edn. Chapman & Hall, London.

Hughes JMB, Pride NB, eds (1999). *Lung function tests: physiological principles and clinical applications*. Saunders, London.

Kharitonov S, Alving K, Barnes PJ (1997). ERS Task Force Report: Exhaled and nasal nitric oxide measurements: recommendations. *European Respiratory Journal* **10**, 1683–93.

Roca J, Whipp BJ, eds (1997). Clinical exercise resting. *European Respiratory Monograph* **2**(6).

West JB, Wagner PD (1997). Ventilation–perfusion relationships. In: Crystal RG, West JB, eds. *The lung: scientific foundations*, 2nd edn, pp 1693–709. Lippincott-Raven, Philadelphia.

Sources of normal reference values

Cerveri I *et al*. (1995). Reference values of arterial oxygen tension in middle-aged and elderly. *American Journal of Respiratory and Critical Care Medicine* **152**, 934–41

Cotes JE (1993). *Lung function: assessment and application in medicine*, 5th edn. Blackwell, Oxford.

European Respiratory Society (1993). Standardised lung function testing. *European Respiratory Journal* **6**(Suppl 16).

Jones NL, Summers E, Killian KJ (1989). Influence of age and stature on exercise capacity during incremental cycle ergometry in men and women. *American Review of Respiratory Disease* **140**, 1373–80

Appendix

Normal values of lung volumes and ventilatory flows vary considerably with age and height. Table 7 is modified from that produced by the European Respiratory Society (1993). Standardised lung function testing. *European Respiratory Journal* **6**, Suppl 16.

Summary equation for lung volumes and ventilatory flows for Caucasian adults aged 18 to 70 years*. The lower 5 or upper 95 percentiles are obtained by substracting or adding the figure in the last column from the predicted mean.

17.3.3Microbiological methods

*Robert Wilson**

Microbiological investigations in clinical practice

Microbiological investigations are an important part of the management of infected patients, since few aetiological agents produce diagnostic clinical features and other investigations are not specific. However, treatment of a severe infection should not be delayed while awaiting the results of laboratory tests because this can be fatal. Clinicians need to know the types of respiratory infection that are prevalent and the likelihood of antibiotic resistance to enable them to select appropriate empirical treatment. However, the level of microbiological investigation needed to provide this information for surveillance purposes usually exceeds that required in clinical practice.

Once the clinical diagnosis of a respiratory infection has been made the physician must decide whether to perform any investigations before starting treatment. The type of patient and the severity of the illness will guide this decision. However, even with extensive testing, it is recognized that the causal pathogen may not be identified in over 50 per cent of patients with community-acquired pneumonia, a condition in which a bacterial aetiology is most likely. The proportion of negative results rises steeply if the patient has received an antibiotic before the microbiological samples are taken. Other explanations for negative bacteriology results include a viral infection being the cause, the presence of non-infectious conditions mimicking pneumonia, the presence of unusual pathogens that go unrecognized (for example, fungi), and the presence of pathogens that are currently not identified or recognized.

This chapter describes the available microbiological methods, and Table 1 indicates the clinical conditions for which they should be used. Some patients should receive more intensive investigation. These include those patients with more serious illness; those with underlying medical problems that put them at higher risk of serious illness (for example, those who are immunocompromised); those at risk of exposure to more unusual pathogens (for instance, nursing-home residents), or of nosocomial infections; and those not responding to treatment. Decisions regarding more invasive investigations, which might have a greater likelihood of giving a positive result, need to be balanced against the risks of any procedure.

Interpretation of results

The interpretation of positive microbiological results may call for fine judgement, and careful consultation between the clinician and medical microbiologist. There are two situations that commonly cause difficulty: the isolation of a species which is part of the commensal flora, and the isolation of an opportunistic pathogen. In these circumstances particularly, but true always, the microbiological results should be considered together with the clinical information and the results of non-microbiological investigations.

The mucosal surfaces of the mouth, nose, pharynx, larynx, and trachea are colonized by a complex variety of bacterial species that make up the commensal flora, whereas the middle ear cavity, the paranasal sinuses, and the lower airways distal to the first bronchial division are usually sterile. The commensal flora confer a level of protection against infection by occupying a niche within the body that might otherwise be colonized by species with greater pathogenicity, and by providing non-specific and specific (via crossreactive antigens) stimulation to the immune system. Some of the commensal species, for example *Streptococcus pneumoniae* and unencapsulated non-typable *Haemophilus influenzae*, may, under permissive conditions, be pathogenic and are amongst the most common causes of respiratory infection.

Many respiratory samples are obtained via routes that are naturally colonized by commensal flora (for example, expectorated sputum), so there is always some uncertainty whether the bacterium has been cultured from the putative site of infection in the bronchial tree or has contaminated the sample during its passage through the oropharynx. A significant proportion of exacerbations of chronic bronchitis have a non-bacterial aetiology, but clinical information can be used as a guide, since patients who have an increased sputum volume which is purulent and increased breathlessness are more likely to have a bacterial infection.

Opportunistic pathogens do not infect patients with intact normal host defences, but do so if the host defences are impaired, either by a humoral or cellular defect, or when the defences are breached artificially, for example by an endotracheal tube. Opportunistic pathogens, for example *Pseudomonas aeruginosa*, have relatively low pathogenicity. However, chronic infection commonly occurs for many years in conditions such as cystic fibrosis and other forms of bronchiectasis. Patients have acute exacerbations intermittently when their symptoms increase and the level of lung inflammation is greater. The sputum bacteriology during these exacerbations is usually the same as when the patient is in a stable state, although bacterial numbers may be greater. Other features, such as the white cell count and C-reactive protein level are helpful in differentiating an acute exacerbation. *P. aeruginosa* can also colonize the bronchial tree of patients who are being ventilated, without causing a significant deterioration in their condition, but this bacterium is also a major cause of ventilator-associated pneumonia, a condition with high mortality. An increase in temperature and the appearance of a new infiltrate on the chest radiograph, as well as a rise in the inflammatory markers, signal the onset of pneumonia.

Direct investigations

Non-invasive tests

Upper respiratory tract samples

Nose and throat swabs provide no useful information about the likely pathogen in patients with sinusitis and otitis media. Throat swabs should be performed during investigation of suspected complications of group A b-haemolytic streptococcal pharyngitis (acute rheumatic fever or glomerulonephritis). Virus isolation is dependent on the presence of an adequate number of infected epithelial cells and high-quality samples are imperative; nasopharyngeal washes or aspirates may provide a better sample than swabs in paediatric cases.

Sputum culture

The information gained from this frequently performed test is limited, unless careful steps are taken to improve specificity. Often, the patient cannot produce a sample to order, which limits its usefulness for outpatients or when empirical treatment is to be commenced quickly. The sample should be transported to the laboratory within 2 h to avoid overgrowth by rapidly growing species. The yield of positive cultures is higher if the sample is purulent; a good sample should have fewer than 10 squamous epithelial cells, indicating the lack of significant contamination from the upper respiratory tract, and more than 25 neutrophils per low-power field (at 100 × magnification). As bacteria are unevenly distributed in sputum, the sample is first homogenized by vigorous agitation in Ringer's solution or by the addition of a commercially available digestion agent, and then is diluted so that a quantitative assessment can be made of the bacteria present.

Gram stain of sputum has been advocated as a rapid diagnostic test, such as when Gram-positive diplococci are seen indicating a pneumococcal pneumonia. However, interpretation of the stain can be subjective and this test should not be performed by an inexperienced observer. In addition, it has to be kept in mind that respiratory infections can be mixed, so focused therapy based on the result of a Gram stain might not cover co-infection with an atypical pathogen such as *Mycoplasma pneumoniae*.

Sputum culture is an important non-invasive investigation in patients with pneumonia of sufficient severity to lead to hospital admission. However, it is rarely useful in the community; nor is it usually helpful in chronic bronchitis, when the results are predictable and unlikely to influence the choice of antibiotic. It might be considered if

a patient fails empirical therapy, when culture may reveal a b-lactamase-producing strain, or occasionally in severe chronic obstructive pulmonary disease an unexpected pathogen such as *P. aeruginosa*. Results of routine sputum culture performed in a cystic fibrosis outpatient clinic (and in other patients with chronic infection, for example bronchiectasis) may be used to guide empirical treatment when the patient presents with an acute exacerbation; sensitivity testing is also useful in these situations to monitor the development of resistance. A new pathogen that would alter management, for example *Burkholderia cepacia*, might also be identified by routine screening of this type of patient.

A range of culture media are inoculated: blood agar, chocolate (heated blood) agar to aid the isolation of *Haemophilus* species, and MacConkey's agar for some Gram-negative bacilli and coliforms. Special culture medium can be used for other bacteria, for example *Legionella* species, various fungi, and acid-fast bacilli. Although many species can be identified to guide the choice of antibiotic after overnight culture, full identification and determination of antibiotic sensitivities take a further 24 h. For some patients (for example, cystic fibrosis with mixed infections), a range of selective media can be used to encourage the growth of some species whilst suppressing others. Additional special staining techniques can be used in appropriate clinical circumstances, for example for acid-fast bacilli and *Pneumocystis carinii*. Culture of *Mycobacterium tuberculosis* on standard media such as Lowenstein–Jensen used to take 6 to 8 weeks before antibiotic sensitivities were available, but nowadays more rapid automated liquid cultures, for example BACTEC®, provide a result in 2 to 4 weeks.

A considerable proportion of patients cannot produce a sample of sputum even with the help of a physiotherapist. This difficulty led to the development of a technique to induce sputum, which has been particularly useful in human immunodeficiency virus (**HIV**)-infected patients with suspected *P. carinii* or mycobacterial infection. The patients brush their teeth and gums, gargle with water, and hydrate themselves with a couple of glasses of water. They then inhale nebulized 3 per cent saline for 20 min, and every 5 min are encouraged to cough. A b2-agonist can be given before the procedure, but the technique has been limited in its application because of the severity of coughing and bronchospasm that may be produced. The procedure should not be performed in an open ward or an area where there are other immunocompromised patients, because of the danger of spreading pathogens.

Several new diagnostic tools have been developed or are in development to examine sputum and other specimens. Broadly speaking, these detect antigens or other products of micro-organisms directly by immunological techniques based on monoclonal antibodies; or they use molecular techniques to identify the organism's DNA or RNA, either directly using a probe or following amplification by the polymerase chain reaction (**PCR**). At the present time, molecular techniques are most commonly used in selected patients to identify *M. tuberculosis*, and in particular isolates carrying the antibiotic-resistance gene for rifampicin. Probes for other resistance genes will follow in time and provide a powerful tool in the diagnosis of multidrug-resistant tuberculosis. Molecular techniques are also being introduced for the detection of cytomegalovirus from cases of pneumonia in immunodeficient patients. It is in this area that new microbiological methods in the diagnosis of respiratory infections are likely to appear, and they will be particularly useful if problems of sensitivity and specificity can be solved so that they can be applied to readily obtained samples such as sputum.

Other non-invasive investigations

In patients with pneumonia two sets of blood cultures should be taken before antibiotics are started. Bacteraemia is intermittent, so ideally samples for culture should be taken at least 1 h apart using the inoculum volume recommended by the supplier, but treatment should not be delayed unless the patient's condition allows it. The presence of bacteraemia increases the risk of complications from pneumonia, so a positive result has prognostic as well as diagnostic implications. However, the sensitivity of the investigation is only 10 to 20 per cent overall, with *S. pneumoniae* being the most common pathogen identified by this method.

A significant pleural effusion should always be aspirated and the following investigations requested: white cell count and differential; measurement of protein, glucose, lactate dehydrogenase, and pH; Gram stain and staining for acid-fast bacilli; culture for bacteria, mycobacteria, and fungi.

Pneumococcal antigen detected in the urine by counter-immunoelectrophoresis has an acceptable sensitivity, which is even higher in pleural fluid, but this is a cumbersome test to perform in the laboratory and is rarely used. *L. pneumophila* antigen in the urine is now used routinely and identifies serotype-1 infection, which is the most common serotype causing pneumonia.

The isolation of respiratory viruses requires sensitive cell-culture systems. Incubation time is very variable from 24 h with herpes simplex to 14 days for cytomegalovirus, but in most cases it is too long for the test to be clinically useful. Viral infection of exfoliated cells can be diagnosed rapidly using immunofluorescent techniques incorporating monoclonal antibodies. Conjugated monoclonal antibodies are available for a range of viruses including respiratory syncytial, influenza, parainfluenza, adeno- and cytomegaloviruses, as well as other microbes (for example *L. pneumophila* and *P. carinii*). Other viruses may require an indirect method using unlabelled mouse antibody and a second step with anti-mouse conjugated antibody.

Invasive tests

A number of invasive diagnostic techniques have been developed to obtain specimens directly from the lower airways that are relatively uncontaminated by oropharyngeal flora. In all cases, the yield is greater if antibiotics have not been commenced prior to the procedure; if the patient is already receiving antibiotics they should not have been changed for several days before the test is performed.

Transtracheal aspiration

Although still performed in some centres this approach is not recommended due to poor patient tolerance and low specificity.

Bronchoscopic protected brush catheter (PBC)

This technique employs a double-catheter brush system which is inserted via the bronchoscope. A distal wax plug in the catheter is dislodged by advancing the inner catheter only when the bronchoscope is in the correct position to take a sample from the identified area. The brush is advanced to take the specimen and then retracted before withdrawing the whole catheter from the bronchoscope. The brush is aseptically cut into a vial containing Ringer's solution, or its equivalent, and agitated to ensure all bacteria are removed, then quickly transported to the laboratory where quantitative cultures are performed. Care should be taken when interpreting the results obtained from patients with chronic obstructive pulmonary disease who can have lower airway bacterial colonization without parenchymal infection.

Bronchoscopic bronchoalveolar lavage (BAL)

This technique also uses the bronchoscope to obtain samples from distal airways and the alveolar space, but it is more likely to be contaminated by nasopharyngeal commensals during insertion of the bronchoscope in the non-ventilated patient. The bronchoscope is wedged into a distal segment of the identified area and sterile normal saline (about 50 to 100 ml) is instilled and aspirated to provide about 10 ml for investigation. The fluid from an initial aliquot may be discarded to try to reduce any contamination with bacteria from the upper airways. Squamous epithelial cells signify upper airway contamination, while the presence of intracellular organisms in phagocytic cells indicates true bacterial infection. Quantitative cultures are again recommended. A certain level of bacterial growth is required to be regarded as significant in both PBC and BAL, usually 10^3 colony-forming units/ml for the former and 10^4 colony-forming units/ml for the latter. However, detection of some microbial species should be considered significant whatever their concentration—for example, *P. carinii*, *Toxoplasma gondii*, *Legionella* species, *M. tuberculosis*, respiratory syncytial virus—whereas isolation of fungi and environmental *Mycobacteria* species need to be correlated with clinical and radiographic findings.

Percutaneous fine-needle aspiration

This may be guided by computed tomography (**CT**) scanning. Complications are infrequent in centres experienced in this technique.

These invasive procedures are rarely used in patients with community-acquired pneumonia, particularly since retrospective data has shown that outcome is not improved by establishing a specific aetiology in those patients with a severe illness. However, bronchoscopy is useful in patients who have failed empirical therapy. This may reveal resistant or unusual pathogens or a mechanical factor delaying resolution, for example an obstructing endobronchial lesion.

Invasive procedures are used more commonly in the immunocompromised patient with pneumonia, when the range of pathogens is much larger, and consequently the choice of empirical therapy much more difficult. In addition, the likelihood of a non-infectious cause of 'pneumonia' is greater. The bronchoscopic techniques have reasonable sensitivity and specificity when performed correctly, carry less risk of complications, and are usually more acceptable to patients. A transbronchial biopsy

can be taken during the bronchoscopy to obtain lung tissue for histology and culture. Direct histological examination of the lung or pleura is important in several situations: detection of herpes simplex virus or cytomegalovirus is not an accurate indicator of pneumonitis without histological confirmation; cytomegalovirus pneumonitis is clinically very similar to acute rejection in transplant patients; granulomas suggest mycobacterial or fungal infection and acid-fast bacilli or fungi with characteristic features may be seen.

In ventilated patients there is less agreement about the role of invasive tests. Culture of endotracheal aspirates should be performed routinely. Failure to culture bacteria from a patient not being given antibiotics has a high negative predictive value for ventilator-associated pneumonia. There may be a higher percentage of false-positive results compared to bronchoscopic techniques, due to bacterial colonization; but quantitative cultures taken together with clinical information and the results of other investigations usually indicate the significance of a positive culture. In ventilator-associated pneumonia this approach is simpler than bronchoscopy; moreover, studies have failed to demonstrate that the information obtained from invasive techniques reduces mortality. Also, bronchoscopy may not be readily available in some hospitals, and by sampling a limited area of the lung it may be less sensitive.

Indirect investigations

Respiratory infections caused by a range of pathogens can be detected late in the infection or retrospectively by serological tests. The delay required until the antibody response occurs means that the results are rarely clinically relevant and these investigations therefore do not need to be performed routinely.

Serological methods are commonly used for viral infections and the atypical bacterial pathogens that are difficult to culture: *L. pneumophila*, *M. pneumoniae*, and *Chlamydia pneumoniae*. Seroconversion takes 3 to 6 weeks, but in elderly patients legionella can take up to 14 weeks. In several rarer infections (for example, histoplasmosis, coccidiomycosis, filariasis), a positive antibody result suggests the presence of active infection. Several serologic methods are available. For many years the most widely used, because of its flexibility, was the complement fixation assay. This detects primarily IgG antibody and, therefore, requires a fourfold rise in antibody levels between acute and convalescent serum samples to demonstrate a new infection. The complement fixation assay has now been replaced for some species by enzyme-linked immunosorbent assays (**ELISA**) that detect specific IgM, which is predictive of a recent or active infection in a sample collected 10 days or more after the onset of symptoms.

*I thank Maureen Chadwick and Paul Taylor from the Microbiology Department at Royal Brompton Hospital for their helpful comments on the manuscript.

Further reading

American Thoracic Society (2001). Guidelines for the management of adults with community-acquired pneumonia: diagnosis, assessment of severity, antimicrobial therapy, and prevention. *American Journal of Respiratory and Critical Care Medicine* **163**, 1730–54

Blasi F, Costentini R (1997). Non-invasive methods for the diagnosis of pneumonia. In: Torres A, Woodhead M, eds. *Pneumonia. European Respiratory Monograph*, pp. 157–74. European Respiratory Society Journals, Sheffield.

Davidson M, Tempest B, Palmer DL (1976). Bacteriologic diagnosis of acute pneumonia. *Journal of the American Medical Association* **235**, 158–63.

Roberts DE, Cole PJ (1980). Use of selective media in bacteriological investigation of patients with chronic suppurative respiratory infection. *Lancet* **i**, 796–7.

Sanchez-Nieto JM, *et al.* (1997). Impact of invasive and noninvasive quantitative culture sampling on outcome of ventilator-associated pneumonia. *American Journal of Respiratory and Critical Care Medicine* **156**, 1–6.

Wilson R (1999). Bacterial infection and chronic obstructive pulmonary disease. *European Respiratory Journal* **13**, 233–5.

Wimberley N, Faling SJ, Bartlett JG (1979). A fiberoptic bronchoscopy technique to obtain uncontaminated lower airway secretions for bacterial culture. *American Review of Respiratory Diseases* **119**, 337–43.

17.3.4 Diagnostic bronchoscopy, thoracoscopy, and tissue biopsy

M. F. Muers

Introduction

Diagnostic bronchoscopy, thoracoscopy, and tissue biopsy are an integral part of the investigation of respiratory disease. They should be regarded as complementary to, rather than substitutes for, simpler and cheaper tests.

The introduction of the flexible fibreoptic bronchoscope by Ikeda in 1974 and the subsequent improvements in instrumentation, together with a widening number of applications, have revolutionized the practice of respiratory medicine worldwide. By contrast, rigid bronchoscopy—although essential in some circumstances—has become much less common. Thoracoscopy, the examination of the pleural cavity by a percutaneously introduced instrument, was first performed in 1913 by the Swedish physician, Jacobaeus. The technique has remained similar ever since, but has in recent years been substantially improved with the introduction of video-assisted equipment—the **VATS** (video-assisted thoracoscopic surgery) approach. With respect to other methods of tissue sampling, the major change in recent years has been the improved accuracy and safety of percutaneous needle-biopsy techniques by the simultaneous use of cross-sectional imaging or ultrasound.

Bronchoscopy

Indications (Table 1)

Bronchoscopy is mainly used to investigate or confirm the possibility of carcinoma, and/or to obtain histological and cytological confirmation of a clinical diagnosis, and to provide evidence about operability. A diagnosis at bronchoscopy does not just depend on tissue sampling as many abnormal appearances are characteristic. It is particularly useful in excluding endobronchial abnormalities as a cause for persistent symptoms in the presence of a normal radiograph. The use at bronchoscopy of imaging and flexible instruments allows sampling of distal bronchi or lung parenchyma that cannot be seen directly.

Techniques

Fibreoptic bronchoscopy

Fibreoptic bronchoscopy is usually an outpatient procedure, done with local anaesthesia and sedation. The list of preoperative requirements is shown in Table 2. The procedure causes a fall in PaO_2 of about 2.5 kPa. For this reason, and particularly because many patients have impaired lung function, oxygen supplementation by nasal cannulation to maintain the SaO_2 at 90 per cent or greater is recommended, with oxygen saturation monitored by pulse oximetry and oxygen supplementation continued postoperatively.

The bronchoscope is best inserted through the nose, but if the nasal passages are too narrow, the instrument can be inserted through the mouth with an appropriate guard. The nose, oropharynx, vocal cords, and bronchial tree are anaesthetized with lidocaine (lignocaine) aerosol or gel. Care must be taken not to over use lidocaine: a maximum total dose <7 mg/kg should be ensured.

There have been four large retrospective studies of the safety of fibreoptic bronchoscopy involving between 4000 and 48 000 patients. Reported mortality rates have ranged between 0 and 0.04 per cent and major complication rates between 0.08 per cent and 0.5 per cent. It is accepted that the risk of complications is greater when transbronchial biopsy is performed, in the presence of coagulopathies, and in patients who are frail or very ill.

Postprocedure infection is very rare, occurring in one in 2500 procedures in a large study. There have been no reported cases of the human immunodeficiency virus (**HIV**) being transmitted by bronchoscopy. However, contamination of bronchoscopes is important for another reason, namely that organisms introduced into the lungs and sampled may give a false impression of infection. This is a particular difficulty with non-tuberculous mycobacteria, which can grow in contaminated rinse water and be resistant to some disinfectants. The most common disinfecting agent is gluteraldehyde. This has to be handled with care: it is a potent sensitizer of the respiratory tract and aerosol contamination of the working environment can put staff at risk of contracting occupational asthma.

Rigid bronchoscopy

Rigid bronchoscopy is performed under general anaesthesia with oxygen Venturi ventilation. The procedure is indicated if previous fibreoptic bronchoscopy has failed to make a diagnosis and there is still a suspicion of pathology, if there is anxiety about uncontrolled bleeding, or when foreign body removal is being contemplated. Most surgeons will re-inspect the bronchial tree before a planned resection to reassess operability, and rigid bronchoscopy offers better conditions for some difficult therapeutic procedures such as laser therapy and the insertion of stents—although both of these can be performed with the fibreoptic instrument. Rigid bronchoscopy

is preferable for children.

Diagnosis at bronchoscopy

The standard 5-mm diameter fibreoptic bronchoscope allows inspection of all the lobes to subsegmental level; smaller 3-mm diameter paediatric bronchoscopes extend this range of vision but cannot be used for biopsies. Approximately 70 per cent of bronchial carcinomas are within visible and sampling range. Many diagnoses can be made without biopsy: for example, paralysis of the left recurrent laryngeal nerve causing vocal cord paresis; endobronchial pus from infected segments; distortion due to lung collapse or metastatic carcinoma; or a large endobronchial tumour (Fig. 1 and Plate 1). Although the appearance of many tumours is quite characteristic, it must be borne in mind that the differential diagnosis of primary lung cancer at bronchoscopy includes adenomas, endobronchial metastatic deposits, for example from breast cancer, or more rarely endobronchial tuberculosis or sarcoidosis.

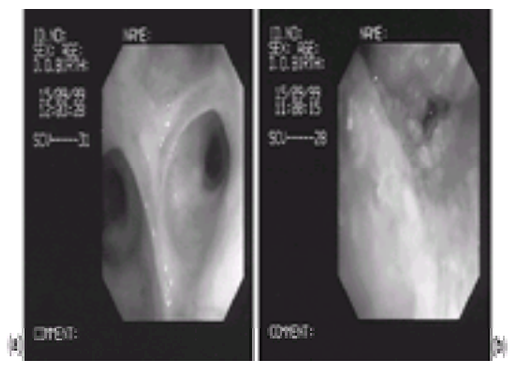

Fig. 1 Appearances at bronchoscopy. The normal thin mucosa and sharp interlobar carinae of the normal left side (a) are in contrast to the irregular exophytic appearance of an advanced non-small cell tumour of the right main bronchus (b) in the same 73-year-old patient. (See also Plate 1.)

Endobronchial brushing, biopsy, and 'washing'

The majority of endobronchial lesions are carcinomas and best sampled by a combination of brushing of the surface for cytology, forceps biopsy for histology, and (where sampling has been difficult) bronchial 'wash' for cytology. Sheathed disposable nylon brushes can safely be applied to most tumours and the resulting specimens rubbed on to slides and air- or ethanol-fixed for cytology. Flexible biopsy forceps provide adequate samples from most endobronchial tumours, but it is advisable to take up to four or five biopsies of each lesion. For a diagnostic bronchial wash, 20 to 40 ml of normal saline is injected over the endobronchial lesion into the peripheral lung and the residual fluid rapidly aspirated in to a trap. A combination of brushing, biopsy, and washing gives the highest yield for malignancy, and this should be well over 80 per cent. If the tumour appears vascular, it is wise to brush first, and proceed to biopsy if significant bleeding does not occur.

A bronchial wash is also useful for microbiological tests particularly for mycobacteria or fungi—for example, if there is a suspicion of tuberculosis in an upper lobe, but no sputum.

Bronchoalveolar lavage

This technique provides a sample of cells from the peripheral airways and alveoli of a lung segment or lobe. It is occasionally useful for the diagnosis of diffuse lung disease, and has been used extensively as a research tool. A bronchoscope is wedged into a segment of lung, either guided by a radiographic abnormality or if not, usually the right middle lobe. Between 150 and 300 ml of buffered normal saline at 37 °C are instilled by syringe pressure in 50 to 60 ml aliquots. Low-pressure suction is continuously applied after each aliquot and the aspirated fluid collected in a trap, the average return being approximately 60 per cent of the injected volume. Particular care has to be taken when the procedure is undertaken in patients with an FEV_1 of less than 1.5 litres (**FEV_1**, forced expiratory volume in 1 second). Supplementary oxygen is nearly always necessary, and after 10 per cent of lavages there is transient pyrexia lasting for between 4 and 8 h.

The lavage specimen is centrifuged down and subjected to cytological examination, when absolute yields and differential cell counts as well as functional studies can be made (Table 3). Lavage in normal non-smoking subjects shows a preponderance of alveolar macrophages with less than 20 per cent of the cells being lymphocytes, polymorphonuclear neutrophils, and eosinophils. Cell counts are increased approximately threefold in smokers and altered in many diffuse lung diseases, particularly sarcoidosis and allergic alveolitis. Although characteristic profiles can be recognized, lack of specificity is a major problem, and in routine clinical practice they provide only supportive evidence for most diagnoses. There are, however, a few pathognomonic appearances (see Table 3). Lavage specimens are particularly useful in the investigation of diffuse lung shadowing in patients who are immunocompromised (see below).

Transbronchial biopsy

This technique enables specimens of lung parenchyma (namely small bronchi, bronchioles, alveoli, and their associated vessels) to be examined. It may be useful in the diagnosis of diffuse parenchymal lung disease (**DPLD**) and occasionally in the diagnosis of localized lesions, for example persistent consolidation.

Closed, flexible, bronchial biopsy forceps are advanced to within about 1 cm of the pleura and then opened, next they are moved gently backwards and forwards two or three times to ensure full opening, before being firmly advanced more peripherally, whilst the patient is asked to breathe out (Fig. 2). The forceps are then closed and withdrawn, usually with a perceptible 'tug'. The bronchoscope is left in position to check there is no appreciable bleeding. The specimens, which are approximately 1 to 2 mm^3, are put in formal saline for histology, or in normal saline for microbiological culture. The diagnosis of diffuse lung disease needs two or three specimens, and more are required for the accurate diagnosis of focal lesions. The diagnostic rate depends very much on the lung pathology being sampled. In fibrosing alveolitis biopsies are often unsatisfactory, but in a condition such as lymphangitis carcinomatosa or sarcoidosis, diagnostic rates may approach 80 per cent or more. For peripheral tumours, the diagnostic rate depends upon the size, but it is characteristically about 50 per cent. An alternative procedure here is a percutaneous lung biopsy.

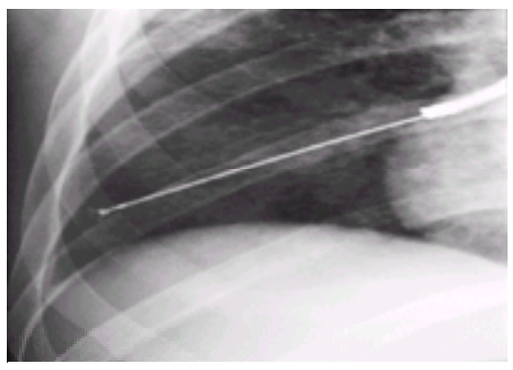

Fig. 2 Transbronchial biopsy through the fibreoptic bronchoscope. The bronchoscope is wedged in the right lower lobe bronchus. Flexible biopsy forceps have been advanced to within 1 cm of the chest wall under screening. The diagnosis: miliary tuberculosis in a renal transplant recipient.

Complications are more common in immunocompromised patients or in the presence of coagulopathies. Mild haemorrhage occurs in 5 to 10 per cent of patients, and pneumothorax in 1 to 5 per cent, with a mortality rate of approximately 0.1 per cent. Bleeding can be reduced by the bronchoscopic injection of 1 to 5 ml of 1:10 000

epinephrine (adrenaline). Because of the risk of pneumothorax, transbronchial biopsy is usually restricted to one lung, although bilateral samples are taken in cases of heart lung transplantation for the surveillance of transplant rejection. A postprocedure radiograph is usual: if pneumothorax is not present 1 h afterwards, it is very unlikely to develop later. There is debate as to whether fluoroscopic screening is required routinely: it is certainly an advantage when the technique is being learned, but in skilled hands the yield and complication rate is similar whether or not screening is used. It is possible to take transbronchial biopsies from patients on intermittent positive-pressure ventilation, but under these circumstances bronchoalveolar lavage alone or an open lung biopsy is probably safer and the latter has a far higher chance, usually, of achieving an accurate diagnosis.

Transbronchial needle aspiration

This technique can be used to sample abnormal bronchial mucosa, peripheral lesions, and occasionally peribronchial lymph nodes. A sampling needle, in a flexible sheath, is passed through the bronchoscope and the needle advanced into lung tissue. Suction is applied to obtain a cell sample.

The technique can be used in addition to, but not usually as a substitute for, the assessment of endobronchial tumours by forceps biopsy, brush, and wash. It is likely to be more useful if a carcinoma has spread submucosally, when bronchial biopsies are often difficult. In this instance the needle is inserted at an angle to the bronchial wall.

Transbronchial needle aspiration is an alternative to transbronchial forceps biopsy for the assessment of peripheral lesions under fluoroscopic guidance. For small peripheral lesions less than 2 cm in diameter, the sensitivity is similar to that of forceps biopsy, about 35 per cent, but may be up to 75 per cent for larger lesions.

Transbronchial needle aspiration can be used to stage lung cancer by obtaining samples from peribronchial nodes. The technique used is similar to that for submucosal sampling, but the needle has to be inserted more perpendicular to the bronchial wall, between bronchial cartilage rings. There is good evidence that directing sampling in the light of a previous thoracic computed tomography (**CT**) scan improves the yield. Under these circumstances about 65 per cent of true-positive nodes may be sampled. The technique is relatively easy when subcarinal nodes are sampled but is much more difficult if the nodes are paratracheal. Probably for this reason, most centres still prefer mediastinoscopy or mediastinotomy.

Bronchoscopic bronchography

High-resolution CT (**HRCT**) is now the diagnostic methods of choice for the investigation of possible bronchiectasis. Bronchography is reserved for the uncommon indication of a focal lung lesion in which HRCT has given equivocal results.

Percutaneous needle biopsy

In 1883 Leyden made a diagnosis of pneumonia by needle aspiration, and in 1886 it was used by Menetier to diagnose lung cancer. The development of fine-bore needles and screening techniques means that this is a very widely used method of obtaining lung tissue. The indications for needle biopsy are usually to confirm a diagnosis of lung cancer, particularly where there is a peripheral lesion not easily accessible to bronchoscopy; occasionally to prove that a lesion is benign; to obtain micro-organisms from an area of consolidation or abscess; and to diagnose a mediastinal mass.

Percutaneous fine-needle aspiration biopsy (FNAB)

This produces samples for cytological examination, but lung architecture is not preserved. Screening is required: fluoroscopic, CT, or (if the lesion abuts the pleura) real-time ultrasonography. A fine spinal 22- to 25-gauge needle is advanced perpendicular to the skin under local anaesthesia until the tip lies within the lesion (Fig. 3). Suction is applied and the needle tip is moved slightly to increase the tissue yield. Aspirated material is expressed on to slides for cytology, or sent for culture. The procedure can be repeated, but multiple passes increase the chance of a pneumothorax. A postprocedure radiograph is mandatory.

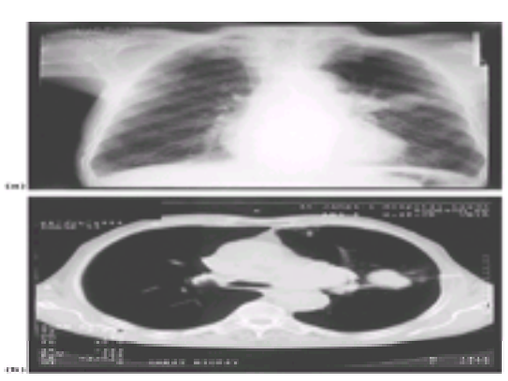

Fig. 3 Chest radiograph (a) showing a mass in the left upper lobe with hilar node enlargement; (b) fine-needle aspiration biopsy under CT guidance. The diagnosis: squamous-cell lung cancer.

There are no absolute contraindications to needle biopsy, but an uncooperative patient, severe emphysema or pulmonary hypertension, a coagulopathy, or a contralateral pneumonectomy can substantially increase the risk. A positive diagnosis can be obtained in patients with cancer in about 90 per cent of cases. The rate is influenced by the size and depth of the lesion and the experience of the operator. False-positive rates (namely a diagnosis of cancer where none is present) are extremely low at well under 2 per cent, but the false-negative rate is much higher, probably 30 per cent to 40 per cent. This means that it is wiser to attempt a specific histological diagnosis of a benign lesion rather than rely on the reported absence of malignant cells. Cell typing of a carcinoma is more difficult from needle aspirates than from histological specimens. The most common complications of FNAB are pneumothorax and minor haemoptysis (about 10 per cent). A large haemoptysis, haemothorax, and implantation of tumour in to the needle track are rarer complications.

Percutaneous cutting-needle biopsy

This technique uses larger gauge needles to obtain a specimen suitable for histology. A 20-gauge or more Trucut biopsy needle or a spring-loaded instrument such as the Biopty are used. The sampling technique is similar to FNAB, but multiple biopsies are accompanied by a much higher complication rate. The indications for using a cutting needle are usually when there is an area of pleural thickening or mass, or when an accessible lesion is thought, prebiopsy, to be benign and requires a specific diagnosis. Theoretically cutting needles can sample diffuse lung disease, but because of the high complication rate transbronchial biopsy or a thoracoscopic biopsy are probably safer.

Pleura and pleural fluid sampling

Most unilateral and some bilateral pleural effusions need samples to be taken for diagnosis. Larger effusions can be sampled by needle aspiration using the physical signs on the chest radiograph to direct the needle into the intercostal space above the area of maximum dullness to percussion. Smaller effusions require fluoroscopic or ultrasound guidance. A 21-gauge venepuncture needle fitted to a 20 to 50 ml syringe can be used in the clinic or on the ward, usually without local anaesthesia. Diagnostic information is obtained from the appearance of the fluid, for example whether it is blood-stained, pus, or chylous, and from various microbiological, cytological, and biochemical tests, as indicated in Table 4. Measurement of pleural fluid atrial natriuretic factor (ANF) is very occasionally helpful in making the diagnosis of systemic lupus erythematosus (SLE). For malignancy, simple aspiration cytology has a diagnostic sensitivity of about 60 per cent, rising to about 75 per cent after repeated aspirations.

Pleural biopsy

Percutaneous pleural biopsy using an Abram's needle is usually done at the same time as a first or repeat aspiration of pleural fluid, when local pleural disease (as opposed to organ failure) is suspected as the cause, and when the diagnosis has not been obtained by simple aspiration. It is necessary to use an aseptic technique, to give adequate local anaesthesia down to the pleural surface (commonly 20 ml of 2 per cent lidocaine (lignocaine)), and to verify the presence of an effusion by prebiopsy aspiration of fluid. A deep incision is made above a rib so that the puncture biopsy needle can be introduced without undue effort in to the pleural space. Multiple samples should be taken, avoiding the inferior surface of the rib above. These should be sent for histological examination and for microbiological culture, particularly if tuberculosis is suspected. The technique is not easy, although samples are highly specific for tuberculosis and malignancy. Routine biopsy after repeatedly negative fluid cytology increases the diagnostic yield for neoplasia by about 10 per cent. False-negative biopsies are common, particularly in mesothelioma. If there is a prebiopsy suspicion of localized pleural tumour, particularly mesothelioma, cutting-needle biopsy under screening is a preferred technique.

Thoracosopy and diagnostic thoracotomy

Thoracoscopy allows direct inspection and biopsy of the pleural cavity. The use of video-assisted equipment in recent years has allowed the expansion of this technique to include more complex procedures (Table 5).

Thoracoscopy

This is indicated if a pleural effusion remains undiagnosed after percutaneous aspiration and needle biopsy. It can be done under sedation and local anaesthesia, but more commonly a general anaesthetic is employed. The patient lies on the contralateral side, a small stab wound is made in the mid-axillary line in the 6th or 7th interspace, and after blunt dissection a rigid 9-mm thoracoscope is used to enter the pleural space. A flexible fibreoptic thoracoscope has been developed, but its use is not widespread.

Pleural fluid is drained, then 200 ml of air is introduced to collapse the lung and allow the pleural surface to separate. The thoracoscope is manipulated to allow inspection of the whole of the pleural surface, and biopsies of abnormal pleura can then be done under direct vision. Adhesions can be broken down, and, if needed, a pleurodesis can be achieved either by using talc powder or a slurry of talc in saline. A chest drain is placed postoperatively. Complications are rare and death from the procedure extremely uncommon. The sensitivity of thoracoscopy for the diagnosis of malignant pleural effusion or tuberculosis is more than 90 per cent.

Video-assisted thoracic surgery (VATS)

In recent years the technique of video-assisted minimally invasive surgery has been applied to the thorax. This avoids the postoperative pain and morbidity associated with many thoracotomies. It allows lung biopsy or resection in patients who might be at high risk because of poor lung function for an open procedure, but the technique is difficult and adequate training is mandatory. Under general anaesthesia the ipsilateral lung is collapsed with the use of a standard double-lumen endotracheal tube, and a stab incision with adjacent instrumentation ports is made in the 6th or 7th intercostal space in the mid-axillary line (Fig. 4). In other respects the technique is similar to standard thoracoscopy, and postprocedure pulmonary tube drainage is required. Patients require a shorter hospital stay than after a standard thoracotomy.

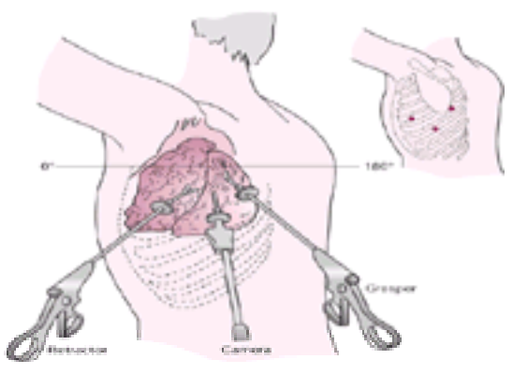

Fig. 4 The arrangement of ports for video-assisted thoracic surgery (VATS). Note the principal access is in the mid-axillary line. (Taken with permission from Landreneau RJ, *et al*. (1992). *Annals of Thoracic Surgery* **54**, 425.)

VATS can be used for diagnosis and treatment. Control of malignant pleural effusions can be achieved by simple talc pleurodesis, providing the lung remains flexible and the visceral and pleural surfaces can be apposed. Alternatives are mechanical pleurodesis by abrading the visceral pleural surface, or total or partial pleurectomy. However, effusions accompanied by contraction and trapping of the lung by tumour or fibrous tissue cannot be effectively pleurodesed. Although small peripheral lesions up to 3 cm in diameter and in the outer third of the lung parenchyma can be relatively easily removed at VATS, greater skill is needed to achieve lobectomy. VATS is an excellent technique for the treatment of persistent pneumothorax, since peripheral bullae can be recognized and excised and a pleurodesis undertaken. Rarer indications include mediastinal sampling—an alternative to percutaneous techniques—and the treatment of malignant or benign pericardial disease by pericardiectomy.

Open lung biopsy

Introduced by Klassen in 1949, the sampling of lung tissue under direct vision through a small 7- to 10-cm thoracotomy under general anaesthesia was the final arbiter in difficult cases, particularly of diffuse lung disease. It has largely been superseded by VATS biopsy. Bilateral diffuse lung disease is best sampled by a right submammary incision allowing samples to be taken from the right upper, middle, and lower lobe. Sampling from an upper lobe lesion, particularly the apical segments, requires a much larger incision. Surgeons are advised not to simply sample the most visibly affected areas, but also the less abnormal, where active pathology rather than fibrosis is more often to be found. As with VATS it is important that CT scanning is obtained to direct the biopsy.

Adequate material for histology is nearly always obtained and a specific diagnosis achieved in more than 90 per cent of cases. However, indications for open lung biopsy, particularly for the confirmation of fibrosing alveolitis, are decreasing due to the combination of high-resolution CT scanning and transbronchial biopsy.

Mediastinal sampling

Mediastinal sampling is required when the clinical problem is either the diagnosis of a mediastinal mass or the assessment of operability of lung cancer.

Needle biopsy of the mediastinum

Mediastinal masses can be diagnosed by percutaneous needle biopsy. The techniques used are similar to those described above for pulmonary sampling by a cutting needle. Biopsy under fluoroscopy was introduced in the late 1970s but has now been largely superseded by real-time ultrasound, which allows sampling of anterior and posterior mediastinal masses that abut the chest wall, or CT scanning which allows anterior middle and posterior mediastinal compartment masses to be sampled.

As with pulmonary lesions, fine-needle aspiration biopsy is both sensitive and highly specific for a diagnosis of cancer, but much less satisfactory if the prebiopsy diagnosis is considered likely to be a benign lesion, a cyst, lymphoma, or a thymic tumour. If this is the case, a cutting-needle biopsy is required. Thus, fine-needle aspiration biopsy has a sensitivity of approximately 85 per cent for any malignancy, but allows an accurate histological diagnosis in only about 60 per cent of cases. Diagnostic sensitivity of a cutting-needle biopsy approaches 90 per cent.

Complications are rare, but include pneumothorax and bleeding, which should occur in much less than 10 per cent of cases.

Surgical mediastinal sampling

The usual indication is for the staging of lung cancer. It is also used when needle biopsy has failed to produce an accurate diagnosis of a mediastinal mass.

Mediastinoscopy

This was introduced by Karlens in 1959, who developed a rigid cervical mediastinoscope. Under general anaesthesia a small 3- to 4-cm transverse excision is made 1 to 2 cm above the suprasternal notch. Blunt dissection approaches the pretracheal fascia and the trachea is followed downwards, again by blunt dissection. The mediastinoscope is inserted and the anterior mediastinum can be dissected and sampled. Complications occur in less than 2 per cent of cases. Hilar nodes, and on the left side the aortic nodes, are best reached by anterior mediastinotomy, using a 6 cm incision in the 2nd intercostal space to allow direct inspection of the mediastinal and hilar structures below.

Special cases

Children

General anaesthesia is needed for both rigid and fibreoptic bronchoscopy. The 3.5-mm diameter paediatric fibreoptic bronchoscopes do not easily allow biopsy, although small forceps are now available. Other sample techniques are similar to those for adults.

The elderly

With appropriate attention to sedation and oxygenation, fibreoptic bronchoscopy and other biopsy techniques are safe and effective in the elderly.

The intensive care unit

Fibreoptic bronchoscopy is easily performed through an endotracheal tube with appropriate attention to oxygenation. Transbronchial biopsy is also possible, although pneumothorax is more likely. In difficult cases it is often better to request an urgent open lung biopsy through a mini-thoracotomy if non-invasive tests, including bronchoalveolar lavage, are non-diagnostic.

Diagnostic clinical applications

This section discusses the use of the techniques described above to assist in the diagnosis and management of different, common respiratory conditions.

Perihilar lesions

In modern adult practice the most common and important diagnosis is lung cancer. Usually, if simple investigations are inconclusive, fibreoptic bronchoscopy should be considered unless there are technical contraindications or good clinical reasons why further information is not needed. The advantage of bronchoscopy over information derived from further imaging is that a tissue diagnosis may be obtained and some aspects of operability can be assessed—such as the proximity of a tumour to the carina. However, recent evidence has suggested that if facilities exist, a better algorithm may be to request a spiral CT scan with contrast before bronchoscopy. This has been shown to reduce the number of negative bronchoscopies as the technique allows some benign diagnoses, can demonstrate that needle biopsy would be better for some patients, and it can direct bronchoscopy to a particular area of interest. This approach depends on having rapid access to scanning, and for most units bronchoscopy is much more easily available.

If plain radiology shows a perihilar lesion but the bronchoscopy is entirely normal, then most physicians would proceed to a conventional CT scan followed by appropriate sampling, usually by percutaneous needle aspiration biopsy or surgical approach. If an endobronchial lesion is seen, a biopsy is unhelpful, but then the options are a repeat bronchoscopy with more biopsies and transbronchial needle aspiration or rigid bronchoscopy.

Solitary pulmonary nodule

When these are detected on plain radiographs, the immediate concern is usually whether or not the nodule represents an early (therefore curable) primary lung cancer. Whether the policy should be one of immediate removal, biopsy, or observation depends upon a careful assessment of the probabilities of a particular diagnosis in any one case. For example, the probability of cancer would be very high in a heavily smoking elderly man with a recent haemoptysis, and it would be lower if the lesion appeared to be calcified or, for example, it had a very smooth edge and was growing slowly. Algorithms exist to assist physicians in what can be a complex decision. It is sensible in most cases to request a thoracic CT scan. This usually gives much helpful additional information such as the density of the lesion, whether it is truly solitary or multiple, whether there is associated lymphadenopathy, and it allows very precise localization. At the same time, if the probability of tumour is high and immediate resection is not planned, the scan can be combined with a fine-needle aspiration biopsy. The majority of nodules are probably better sampled by this technique than by directed bronchoscopy using screening, since the diagnostic sensitivity of bronchoscopic sampling of peripheral lesions that are not visible is only about 50 per cent. This might be an appropriate approach, however, if the patient was not thought to be able to tolerate a pneumothorax or percutaneous biopsy was not available.

An alternative in some cases might be a VATS procedure or a mini-thoracotomy and removal of the nodule, with immediate frozen-section examination to determine whether it is malignant. If so, a decision can be made by the surgeon as to whether to limit the resection to segmentectomy or to proceed to formal thoracotomy and lobectomy. Needle biopsy is not appropriate if the prebiopsy diagnosis is likely to be a vasculitis or other complex disease. A larger sample is required for an accurate diagnosis and a VATS or mini-thoracotomy biopsy should be obtained.

Diffuse parenchymal lung disease

Under this heading are both widespread bilateral interstitial alveolar shadows and also similar shadows confined to one lobe or segment of the lung, for instance a persistent 'pneumonia'. The role of tissue biopsy in the diagnosis and management of patients with these shadows is difficult to clarify, because published series do not necessarily give adequate answers to the questions of when and how the lung should be biopsied.

Practical points are as follows:

1. Biopsies should not be considered until an adequate history has been taken and there has been a careful physical examination, looking particularly for evidence of systemic disease, and the patient has had a full set of pulmonary function tests and a high-resolution CT scan. After these investigations there will be a high probability of a particular diagnosis in many cases, and in some the HRCT scan will show pathognomonic appearances, such as lymphangitis carcinomatosa or bronchiectasis, rendering a tissue diagnosis unnecessary.
2. Biopsy should not be considered if the tissue diagnosis is almost certain not to result in any change of management, increased precision of diagnosis, or a more accurate prognosis.
3. Any biopsy should be performed by an experienced operator, or under their immediate supervision, and the possible complications should be explained beforehand to the patient.
4. Biopsy should not be performed if the occurrence of such a complication, particularly a pneumothorax in the presence of poor lung function, would endanger the patient.

For a fuller discussion of the diagnostic approach to the patient with diffuse parenchymal lung disease, see Section 17.11.

Pleural disease

Most pleural effusions can be diagnosed confidently with a combination of basic clinical information and needle aspiration. This should always be the first approach. For the remainder the usual problem is to decide whether a persistent exudative effusion is or is not due to malignancy. If a first aspiration fails to provide a diagnosis then it should normally be repeated with closed punch biopsies and at least one of these sent for microbacterial culture and the others for histology. This approach is reasonable if there is no evidence from the plain radiographs that diffuse pleural thickening is present. If this is the case it is probably wiser to obtain a CT scan, and consider the next step in the light of the findings. The CT scan has a large advantage over plain radiology in that it can indicate the most appropriate point for pleural biopsy. Localized pleural thickening can be guided by this information and a percutaneous cutting-needle biopsy (for example, a Trucut or Biopty needle) is then superior to bedside biopsy with the Abram's punch.

An increasingly common problem in industrialized countries is the appearance of a pleural effusion due to mesothelioma in a previously well patient. This is notoriously difficult to diagnose. Early scanning is required, and if there is no obvious target for percutaneous needle biopsy, early thoracoscopy is recommended. This has the advantage that multiple samples can be taken, and if a frozen section demonstrates a malignancy, an immediate pleurodesis can be performed. However, some mesotheliomas have a florid fibrous stroma and all biopsies are negative, so that in a small proportion of cases confirmation of the diagnosis remains elusive and observation has to be advised.

There is a small risk that percutaneous procedures will be followed by tumour nodules as a result of seeding along the biopsy-needle tract. Trials have shown that this possibility is much reduced by a postprocedural, localized, short course of radiotherapy. It is the author's opinion that the advantages to the patient of a precise diagnosis and the possibility of better management of his/her effusion outweighs this risk. Bronchoscopy is usually unrewarding if the only radiographic abnormality is a small to moderate effusion.

Tuberculosis

Further sampling is often required in cases where tuberculosis is suspected on clinical grounds, but where conventional sputum specimens are negative or absent, and further information is thought necessary before treatment begins. Samples can be obtained either by induced sputum or at bronchoscopy. The former technique uses 3 per cent (hypertonic) saline, usually in volumes of 70 to 100 ml in an ultrasonic nebulizer, to induce coughing and sputum. At bronchoscopy a 40- to 60-ml wash of the affected segment is usually combined with brushing for microscopy, and occasionally with transbronchial biopsy. Comparative studies have shown that the diagnostic yield in patients with focal radiographic abnormalities is similar. Where miliary tuberculosis seems likely and sputum is absent, bronchoscopy is the preferred technique, with a diagnostic sensitivity of about 80 per cent. Florid endobronchial tuberculosis is a comparatively rare disease, but it can be mistaken for tumour. Biopsy shows profuse organisms.

Tuberculosis is often an important differential diagnosis of large pleural effusions, not only in younger patients where primary disease is likely, but also in the elderly where reactivation may have occurred. Pleural fluid sampling alone is much less satisfactory than combining this with closed pleural biopsies. Multiple samples should be sent both for histological examination and for culture. In parts of the world where tuberculosis is common, a high level of pleural fluid adenosine deaminase can be a strong indication that this is the underlying diagnosis, but false-positives occur in empyemas and sometimes in cancer.

Mediastinal disease

For details of the preoperative assessment of the patient with lung cancer, see Chapter 17.14.1.

A tissue diagnosis is required for most mediastinal masses. Sampling under CT guidance is best. Ultrasound is equally good if the mass is anterior or posterior and abuts the chest wall. If the prior working diagnosis is carcinoma, then fine-needle aspiration biopsy is recommended. If this test is negative (no malignant cells) or there is an indication on the smear that the diagnosis may be thymoma or lymphoma, or if the prior working diagnosis is either of these, then a cutting-needle biopsy should be preferred. It is unwise to diagnose thymoma or lymphoma on the results of a fine-needle aspiration. In all other cases, open surgical biopsy is required usually at mediastinotomy or mediastinoscopy.

Biopsy outside the thorax

It is always important to look at the whole patient, and not just the thorax or chest radiograph. Abnormal tissue outside the thorax may be considerably easier to biopsy than tissue within it. A good example would be enlarged supraclavicular nodes, easily accessible to fine-needle aspiration biopsy in cases of suspected cancer. On occasion, putative liver metastases may be easier to sample than a small thoracic primary.

Therapeutic clinical applications

Bronchoscopic suction and saline lavage through either a fibreoptic bronchoscope or a rigid bronchoscope can be used to relieve obstruction due to viscid secretions. Rigid bronchoscopy is necessary if these are inspissated or very tenacious, as in many cases of non-asthmatic mucus impaction in the elderly. Bronchoscopy should be considered for this reason in cases of 'resistant asthma' in this age group.

Carcinoma

Impressive palliation of distressing symptoms due to endobronchial tumour, such as stridor, breathlessness, or cough, can be obtained in a number of ways. Immediate relief can be produced by deploying endobronchial stents, or using diathermy, cryotherapy, or laser therapy to obliterate tumours. Insertion of stents usually requires rigid bronchoscopy, although techniques have been described allowing placement by physicians using fibreoptic bronchoscopes. In laser therapy, a plastic catheter containing an optical fibre is passed through the instrument channel of the bronchoscope and directed at a tumour. Pulses of high-energy light, usually from a neodynium–YAG laser, cause superficial vaporization and charring of tumour tissue whilst small blood vessels are sealed, providing a relatively dry field. Treatment of haemoptysis in this way is easy, but tumour ablation is a longer and more difficult procedure. Nevertheless, palliation is provided in about 60 to 80 per cent of cases.

Brachytherapy (endobronchial radiotherapy) can complement external beam radiation and has been shown to produce adequate palliation either as initial treatment or if further external beam treatment cannot be given. At fibreoptic bronchoscopy, a catheter is inserted into the narrowed bronchus and the tip passed peripherally. Under fluoroscopic guidance a marker wire is inserted to allow the prescription of treatment, the bronchoscope is withdrawn, and using a remote control device a radioactive source is advanced through the catheter, delivering a high dose of radiotherapy endobronchially, for example 10 Gy at a distance of 1 cm. A single treatment suffices. Relief is not immediate. Combining brachytherapy with other techniques appears to predispose patients to severe haemoptysis.

Photodynamic therapy utilizes the fact that previously injected haematoporphyrins are selectively taken up by tumour cells and, when activated by appropriate laser light, release active oxygen radicals which destroy these cells. There have been reports of treatment of early tumours by this technique, and in skilled hands it can be used to palliate, in a similar fashion to laser therapy. Disadvantages are that repeat bronchoscopy may be needed to remove tumour debris, and the requirement for the patient to remain out of bright light for a period after treatment.

New developments

Early cancer

It is difficult for even an experienced bronchoscopist to detect early tumours such as carcinoma *in situ*. Bronchial epithelium passes through a number of malignant stages before invasive carcinoma develops—these are metaplasia, dysplasia, and then carcinoma *in situ*. It is known that when the bronchial epithelium is illuminated with laser light 405 to 442 nm in wavelength, there is progressive reduction in fluorescence intensity as tissue becomes more abnormal. Thus, if at bronchoscopy the bronchial walls are illuminated with laser light, endoscopists are able to detect and biopsy areas that may be abnormal. Experience with this technique is limited at the moment, but it is likely that with greater refinement it will have an important role in the surveillance of those at very high risk, or in the investigation of 'difficult cases',

for example repeated minor haemoptyses with a normal routine white light examination.

Different degrees of epithelial abnormality are accompanied by an increase in number and a change in the nature of chromosomal abnormalities in exfoliated cells. Examples are loss of heterozygosity or the presence of DNA adducts, or the amplification of proto-oncogenes. With the ability of induced sputum techniques to obtain representative samples of bronchial epithelial cells, even in patients who do not routinely produce sputum, there is considerable interest in the possibility of using such techniques in screening programmes. At present, however, there is insufficient discrimination between minor abnormalities unlikely to develop in to cancer and those that are more likely to.

Tuberculosis

At present, if a sample is smear-negative, confirmation of a diagnosis of tuberculosis may wait upon a culture result taking up to 6 to 8 weeks to mature. Accelerated culture methods are now available, and in addition the introduction of the polymerase chain reaction (**PCR**), although beset by the problems of false-positives, does allow a more confident diagnosis to be made quickly in such cases.

Further reading

Abolhoda A, Keller SM (2000). Surgical staging of the mediastinum. In: Pass HI, *et al*., eds. *Lung cancer: principles and practice*, 2nd edn, pp. 628–48. Lippincott-Raven, Philadelphia.

Anderson C, Inhaber N, Menzies D (1995). Comparison of sputum induction with fibreoptic bronchoscopy in the diagnosis of tuberculosis. *American Journal of Respiratory and Critical Care Medicine* **152**, 1570–4.

British Thoracic Society Guidelines (1998) The diagnosis, assessment and treatment of diffuse parenchymal disease in adults. *Thorax* **54**, S1.

British Thoracic Society Guidelines on Diagnostic Flexible Bronchoscopy (2000) *Thorax* **56 Suppl**., 1–21.

Burgess CJ, *et al.* (1995). Use of adenosine deaminase as a diagnostic tool for tuberculous pleurisy. *Thorax* **50**, 672–4.

Flower CDR, Schneerson JM (1984). Bronchography via the fibreoptic bronchoscope. *Thorax* **39**, 260–3.

Hansell DM (1995). Interventional techniques. In: Armstrong P, Wilson AG, Hansell DM, eds. *Imaging of diseases of the chest*, 2nd edn, pp. 894–912. Mosby, St Louis. (Includes a general account of needle biopsy.)

Haplin DMG, Colins J (1995) Bronchoscopy and lavage. In: Brewis RAL, *et al.* eds. *Respiratory medicine*, 2nd edn, pp. 362–74. WB Saunders, London. (A general account of bronchoscopy.)

Hernandez P, *et al*. (1995). High dose rate brachytherapy for the local control of endobronchial carcinoma following external irradiation. *Thorax* **51**, 354–8.

Hetzel MR, Smith SGT (1991). Palliation of tracheobronchial malignancies. *Thorax* **46**, 325–33.

Klech H, Hunter C, eds. (1990) Clinical guidelines and indications for bronchoalveolar lavage (BAL): Report for the European Society of Pneumonology Task group on BAL. *European Respiratory Journal* **3**, 937–74.

Lamb S, MacAuley CE (1998). Endoscopic localisation of pre-neoplastic lung lesions. In: Martinet Y, *et al*., eds. *Clinical and biological basis of lung cancer prevention*, pp. 231–8. Birkhauser Verlag, Berlin.

Muers MF (1994). How much investigation? In: Thatcher N, Spiro S, eds. *New perspectives in lung cancer*, pp. 77–104. BMJ Publishing Group, London. (An account of the application of bronchoscopy and biopsy techniques for the diagnosis of lung cancer.)

Simpson, FG, *et al*. (1986). Postal survey of bronchoscopic practice by physicians in the United Kingdom. *Thorax* **41**, 311–17.

Vansteenkiste J, *et al*. (1994). Transcarinal needle aspiration biopsy in the staging of lung cancer. *European Respiratory Journal* **7**, 265–8.

Wilcox PA, *et al*. (1986). Rapid diagnosis of sputum negative miliary tuberculosis using the fibreoptic bronchoscope. *Thorax* **41**, 681–4.

17.4.1 Asthma: genetic effects

J. M. Hopkin

Introduction

Asthma is a heterogeneous syndrome characterized clinically by labile airflow obstruction. There are distinct pathological features of prominent eosinophilic inflammation, additional T-lymphocyte infiltration, mucous gland hypertrophy and hypersecretion, smooth muscle hypertrophy and hyper-reactivity, and, in a long-established case, epithelial basement membrane thickening. Asthma's closest correlate is atopy, the state of allergic response to common environmental antigens mediated by the antibody IgE, and both conditions are characterized by exuberant **TH2** (subset 2 of T-helper cells) immune mechanisms. Atopy is present in more than 90 per cent of people between 5 and 30 years of age with asthma, but only 30 per cent of those with atopy develop asthma. In older life asthma may arise in non-atopics exposed to isocyanates and other substances at work. Some develop the syndrome as a result of aspirin sensitivity. In others no 'trigger' is demonstrable—so-called intrinsic asthma, arising as a result of both environmental and genetic influences.

Environment and genetics

Besides the clear involvement of reaction to the extrinsic agents noted above, there has been a surge in the prevalence of atopic disorder, and with it asthma, in recent decades. This is centred on developed communities, pointing to important environmental determinants of atopy and asthma that relate to socioeconomic development.

Epidemiological and experimental findings suggest that the rise of atopy and asthma in developed countries may be due to changing patterns of microbial exposure in early childhood. Natural exposure to *Mycobacterium tuberculosis*, measles, *Helicobacter pylori*, *Toxocara carnis*, and hepatitis A all predict less subsequent atopy. Oral antibiotic receipt in early childhood predicts more subsequent atopy, perhaps because of the deletion of gut microflora. Hence it has been proposed that exposure to certain microbes in early childhood may strongly promote natural immune restraint mechanisms and that these, through the actions of TH3 and TR1 cells (and IL-10 and TGF-b), cause repression of TH2 mechanisms against allergens. Inoculations of mycobacteria do effectively prevent experimental allergy in mice.

Despite the evident importance of environmental factors, there are also clear indications of important genetic determinants. Both atopy and asthma aggregate in families, though there is no consistent pattern of inheritance. Twin studies, comparing concordance rates for asthma and IgE sensitization in monozygotic and dizygotic twins, indicate that genetic variables have accounted for approximately 50 per cent of atopy and asthma syndromes in recent decades.

Genome screening studies

The recognition of the genetic contribution to atopy and asthma has resulted in a clutch of genome screening studies over the past decade. The results have been predictably complex. The use of affected sibling-pairs and microsatellite DNA markers has identified linkages at a number of chromosomal locations, with a significant impression that the linkages may vary between racial groups. To date the following chromosomes have shown repeatable linkages—2q, 5q, 6p, 11q, 12q, 13q, 14q, and 16p—with either asthma *per se* or IgE levels.

However, the complex mix of heterogeneous genetic factors and environmental input places important limitations on exact gene mapping. Direct candidate gene approaches, based on the pathophysiology of asthma and allergy and allied to chromosomal approaches, are providing advance.

HLA

Genetic linkage studies and direct genetic association studies demonstrate a clear relationship between HLA variants on the long arm of chromosome 6 (that is, chromosome 6q) and clinical atopy or asthma. HLA variants relate to allergic responses to distinct antigens or epitopes, e.g. HLA-DR2 and one component of the ragweed antigen (Amb a V) and HLA-DR3 and acid-anhydride sensitization asthma. DR variants have also been associated with allergic aspergillosis.

Interleukin (IL)-4 and IL-13 signalling

Because of the prominent involvement of these cytokines in bronchial inflammation and IgE production they have been targets for direct genetic investigation, and a number of important relationships are emerging. Figure 1 illustrates the signalling pathways of IL-4 and IL-13 and emphasizes that the receptors for both include IL-4Ra, and that IL-13 in comparison with IL-4 has the greater actions in promoting bronchial mucosal pathology.

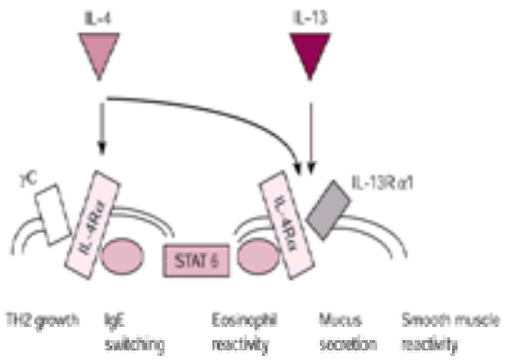

Fig. 1 Interaction between IL-4 and IL-13 signalling in overactive TH2 immunity in bronchial asthma. IL-4Ra is common to the receptors for both IL-4 and IL-13.

A number of coding variants of IL-4Ra (encoded on the short arm of chromosome 16; namely chromosome 16p) have been identified, and these relate to atopic disorder in genetic association studies. For instance, an extracellular variant—ILe50Val—is substantially more prevalent in its homozygous state in young Japanese asthmatics compared with controls. In transfection experiments it appears to upregulate receptor response to IL-4, with enhanced signalling transduction (through STAT 6 activation (**STAT**, signal transducers and activators of transcription)), increased TH2-cell growth and increased cellular IgE production. Intracellular variants of IL-4Ra, for example Gln551Arg, associate with total serum IgE levels, atopic eczema, asthma, and the rare hyper-IgE syndrome. Gln576Arg may also have an action on STAT-6 binding. Thus variants at the IL-4Ra locus enhance atopy *per se*, and have an important effect on predicting disease, including asthma, in the diverse populations.

A charge-changing variant (Gln110Arg) in the helical tail portion of IL-13 ligand (encoded on chromosome 5q) associates with asthma, both allergic and non-allergic, in Caucasian and Asian populations. Molecular modelling studies indicate that position 110 of IL-13 plays a crucial role in ligand receptor interaction (Fig. 2) and

suggest that the amino acid substitution results in increased affinity between ligand and receptor.

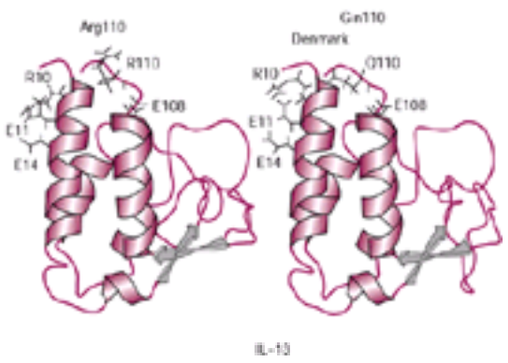

Fig. 2 The Gln110Arg variant of IL-13 is in the helical tail of the ligand, a position that critically mediates receptor binding.

One relatively common variant of IL-13Ra1 shows association with both high IgE levels and asthma. Of note, IL-13Ra1 is encoded on the X chromosome and the risk of atopic disorder and asthma for this variant is confined to young males.

In summary, relatively common variants of IL-4 and IL-13 signalling have been recognized which relate significantly to atopy and asthma in epidemiological and experimental studies. The existence of such TH2-promoting variants in human populations may relate to their potential to enhance protective TH2 mechanisms against certain phases of helminthic infestation—a regular threat to mankind in certain environments now, and almost ubiquitously in the past.

The IgE receptor

One of the first linkages noted for atopy was on chromosome 11q, a linkage that is variably present in different Caucasian and Asian populations. The principal candidate locus here is the b-subunit of the high-affinity IgE receptor. This is an important amplifier of signals through the receptor, and any genetic variants that would enhance its expression or its activity might have important effects on the allergen–IgE interaction and the triggering of mast cells in the bronchial mucosa to release proinflammatory mediators. A number of variants of FCeR1-b, both coding and non-coding, show association with high IgE levels and asthma. No functional effect has been demonstrated as yet, and studies continue.

Effector mechanisms

Although IL-4 and IL-13 signalling, and the secretion of IgE, are primary mediators of the asthma and atopy syndromes, their actions require the adjunctive activity of a whole set of effector molecules. Genetic variants also play some role here in predisposing to disease. A variant within the promoter of the 5-lipoxygenase gene (encoding for the synthesis of inflammatory leukotriene mediators) associates with the response to the antileukotriene, anti-asthma pharmacological agents. Variants of the leukotriene C4 synthase promoter associate with the increased risk of aspirin-sensitivity asthma. Variants of the b $_2$-adrenergic receptor, involved in smooth muscle activity and other inflammatory mechanisms, modulate receptor activity and associate with adverse clinical response to long-acting b-adrenergic bronchodilators seen in people with asthma.

Conclusion

Genetic and environmental factors play equally important parts in the development of asthma. Genetic factors are complex and heterogeneous. In keeping with the TH2-driven bronchial pathology typical of asthma, common variants pertinent to these mechanisms have an important impact on the risk of disease—including those in IL-4 and IL-13 signalling, and IgE binding to mast cells.

Further reading

Collaborative Study on the Genetics of Asthma (1997). A genome wide search for asthma susceptibility loci in ethically diverse populations. *Nature Genetics* **15**, 389–92.

Renauld JC (2000). New insights into the role of cytokines in asthma. *Journal of Clinical Pathology* **54**, 577–89.

Shirakawa T, *et al*. (2000). Atopy and asthma: genetic variants of IL-4 and IL-13 signalling. *Immunology Today* **21**, 60–4.

17.4.2Allergic rhinitis ('hay fever')

S. R. Durham

Introduction

Rhinitis refers to inflammation of the nasal mucosa. In clinical terms it may be defined as symptoms of nasal itching, sneezing, discharge, or blockage, which occur for more than 1 h on most days. Although frequently trivialized, allergic rhinitis remains a common cause of morbidity and social embarrassment. Estimates have suggested that 10 to 15 per cent of the population of the United Kingdom have perennial and/or seasonal rhinitis. Furthermore, the prevalence of hay fever appears to be increasing: in the United Kingdom: in 1955–56 there were 5.1 consultations with general practitioners for hay fever per 1000 population; in 1981–82 there were 19.8.

The lining of the nose and paranasal sinuses is in continuity with the lower respiratory tract and diseases of the upper and lower airways frequently coexist. The nose provides an accessible 'window' for studying allergic disorders and other diseases that can affect the lower airways. Nasal disease may also be the presenting feature of systemic disorders. The aetiology and pathogenesis of allergic rhinitis are described in this chapter, followed by practical guidelines for the diagnosis and management of allergic rhinitis.

Aetiology

Seasonal allergic rhinitis

Pollens of importance include tree pollens in the spring and grass pollens during the summer (Fig. 1). Weed pollens and mould spores predominate in the latter part of the summer and early autumn. Grass pollen counts above 50/m^3 are considered high and represent the threshold level at which most hay fever sufferers experience symptoms.

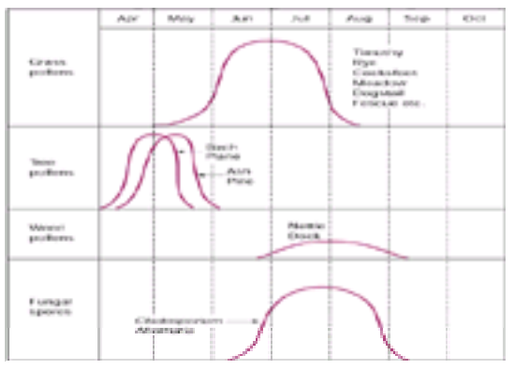

Fig. 1 Calendar of common seasonal aeroallergens (by courtesy of Professor A. B. Kay, Imperial College School of Medicine, London).

Perennial allergic rhinitis

By far the commonest cause of perennial allergic symptoms is the house dust mite (*Dermatophagoides pteronyssinus*, *Dermatophagoides farinae*, and *Euroglyphus maynei*). Mites are found in almost every home, where they live in the dust that accumulates in carpets, bedding, fabrics, and furniture. They live on shed human skin scales and thrive in temperatures of between 15 and 20 °C and a relative humidity of 45 to 65 per cent, which are typical of many modern centrally heated homes. The major allergen of the house dust mite (Der p1) is a digestive enzyme present in the gut and excreted in high concentrations in the mite faeces.

Domestic pets are the second important cause of perennial allergy, identifiable in up to 40 per cent of children with asthma and/or rhinitis. The major allergen (Fel d1) is a salivary protein, which is preened on to the fur and released on very small particles (less than 2.5 μm diameter) which remain airborne for many hours, explaining why a sensitized person can experience symptoms almost immediately upon entering a home containing a cat without being directly exposed to the animal. Dog allergens are less well characterized (Can F1). Recently, cockroaches have been described as a cause of perennial allergic symptoms, particularly in inner-city areas.

Food allergy is unusual as a cause of rhinitis in the absence of other organ involvement. However, rhinitis may be one component of IgE-mediated food-induced symptoms commonly due to egg, milk, and nuts in children; nuts, fish, shellfish, and fruit in adults. Preservatives such as tartrazine, benzoates, and sulphites may provoke symptoms of rhinitis. Important drugs that can trigger rhinitis include beta-blockers, aspirin, and (occasionally) angiotensin-converting enzyme (**ACE**) inhibitors.

Occupational rhinitis

Occupational rhinitis refers to rhinitis caused by an agent inhaled in the workplace. Like other causes of seasonal and perennial rhinitis, occupational rhinitis may also be associated with bronchial asthma. Occupations at risk include laboratory animal workers (rats, guinea-pigs, mice), bakers (flour, grain mites), agricultural workers (cows, pollens, fungal spores), electronic solderers (colophony), and health workers and users of rubber gloves (latex).

Pathophysiology

Immediate symptoms of allergic rhinitis occur as a consequence of allergen crosslinking adjacent IgE molecules on the surface of mast cells in the nasal mucosa (Coombs' classification type-1 immediate hypersensitivity). This results in the release of a range of granule-derived mediators, including histamine and tryptase, and the generation of bradykinin. IgE-dependent activation of mast cells also results in the release of newly formed membrane-associated mediators derived from arachidonic acid associated with the membrane lipid. These include leukotriene C4 (LTC4), LTD4 and LTE4 and prostaglandin D2.

In patients with allergic rhinitis, eosinophils are prominent in nasal washings and in biopsies of the nasal mucosa. The mechanism of this tissue eosinophilia is largely unknown. Chemotactic factors released following mast-cell activation may play a role. Recent evidence suggests that peptide messengers (cytokines) released predominantly from T lymphocytes—but also from mast cells, eosinophils and other cell types—may be important.

A hypothesis for the pathogenesis of allergic rhinitis is shown in Fig. 2. Helper (CD4+) T lymphocytes may be subdivided according to their profile of cytokine release: 'TH_1-type' cells producing predominantly interleukin-2 (**IL-2**) and interferon-g (**IFN-g**) and 'TH_2-type' cells producing mainly IL-4 and IL-5. The biological properties of TH_2-type cytokines suggest their involvement in allergic rhinitis, and increases in cells expressing these cytokines have been detected in the nasal mucosa during the 'late' nasal responses that occur in sensitized subjects between 6 and 24 h following experimental nasal provocation with allergen. IL-4 is the major cytokine responsible for switching B-cell immunoglobulin production from IgM and IgG to predominantly IgE. IL-3 is a growth factor for mast cells. IL-3, IL-5, and granulocyte-macrophage colony-stimulating (**GM-CSF**) factor are important in the proliferation of eosinophils from bone marrow precursors and their maturation, activation, and prolonged survival in tissues. IL-5 promotes the selective adhesion of eosinophils to vascular endothelium prior to diapedesis. **VCAM-1** (vascular cell-adhesion molecule-1) is selective for the ligand **VLA-4** (very late activation antigen-4; an integrin) expressed on eosinophils, and is upregulated by IL-4 and IL-13, an alternative pathway of eosinophil recruitment.

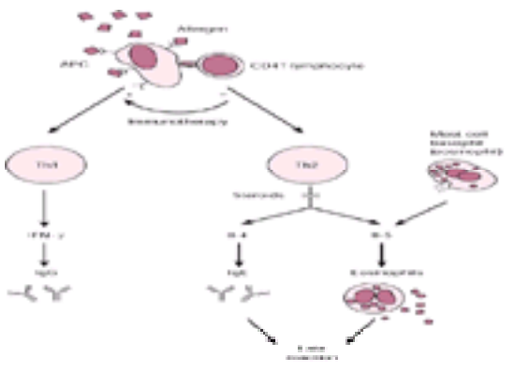

Fig. 2 Hypothesis: pathogenesis of allergic rhinitis and influence of treatment. TH2 cells are predominantly T lymphocytes, although mast cells, basophils, and eosinophils represent alternative sources of TH2-type cytokines. Topical corticosteroids downregulate the production of TH2-type cytokines from T lymphocytes and other cells. Allergen-immunotherapy alters the TH2/TH1 T lymphocyte balance, either by inducing immune deviation, TH2®TH1 responses and/or by inducing T-cell unresponsiveness (anergy) of TH0/TH2-type responses to specific allergens.

Mechanism of effect of treatments

Topical nasal corticosteroids and allergen injection immunotherapy (desensitization) are highly effective treatments for allergic rhinitis (see later). They appear to act by distinct mechanisms. Topical corticosteroids have multiple anti-inflammatory effects. They reduce the number of antigen-presenting cells (**APCs**; Langerhan's cells) within the nasal mucosa during natural pollen exposure. They also inhibit the recruitment of basophils and mast cells to the nasal epithelium and inhibit the production of TH2-type cytokines such as IL-4 and IL-5 from T lymphocytes and possibly other cell sources, including basophils. By contrast, immunotherapy acts by altering the TH2/TH1 balance in favour of the production of inhibitory cytokines such as IFN-g, which downregulate IL-4-induced B-cell switching in favour of IgE and IgG production. This shift may occur either as a consequence of immune deviation of TH2 responses (TH2®TH1) or induction of T-cell unresponsiveness (anergy) of TH0/TH2-type responses to aeroallergens (Fig. 2).

Clinical diagnosis

An approach to the diagnosis of patients presenting with nasal symptoms is summarized in Fig. 3. The diagnosis of allergic rhinitis is usually straightforward. However, the differential diagnosis should be considered in every case: frequently more than one cause coexists.

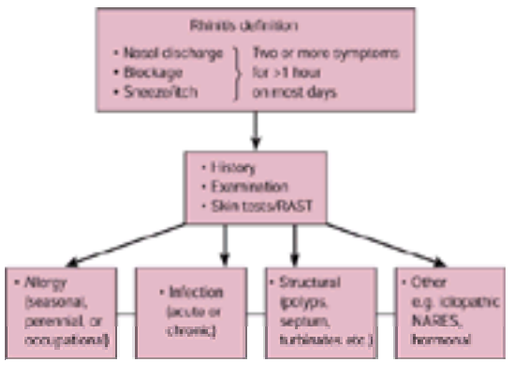

Fig. 3 Diagnostic approach to patients presenting with nasal symptoms. A careful history, clinical examination, and skin-prick tests and/or measurement of serum allergen-specific IgE (RAST or ELISA) should be performed in every case. More than one cause may be present. 'Other' causes include hormonal (pregnancy, premenstrual), drugs (aspirin, b-blockers, ACE inhibitors, cocaine abuse, and atrophic, postsurgical, and ageing). Idiopathic rhinitis refers to nasal hyper-reactivity of unknown cause, manifest as an exaggerated response to non-specific stimuli such as changes in temperature, tobacco smoke, domestic sprays, etc. The differential diagnosis includes vasculitis (Churg-Strauss syndrome), granulomatous conditions (Wegener's, sarcoidosis), atrophic (old age, surgical), and, rarely, tumours of the nose and paranasal sinuses.

History

A careful history is essential both to establish the diagnosis of rhinitis and to assess the severity of symptoms. An allergic aetiology is suggested by dominant itching, sneezing, and watery nasal discharge. Associated eye or chest symptoms (asthma) also point to an allergic cause, and a history of potential allergic triggers should always be sought. However, in addition to provoking immediate nasal symptoms, allergen may also cause late symptoms several hours after exposure, and these may not be recognized as being related. A history of potential allergic triggers includes enquiry into the seasonality of symptoms and whether symptoms are work-related (in other words, do they occur at work or in the evening following work, with improvement at weekends and during holiday periods). The home environment, including the presence of domestic pets, birds, fitted carpets, central heating, and the use of blankets on beds should be established. A personal or family history of atopy is extremely common in patients with allergic rhinitis.

There are many alternative causes of rhinitic symptoms. It is common for there to be more than one cause, and important to consider the differential diagnosis (Table 1). The presence of facial pain, fever, systemic upset, and mucopurulent discharge suggests infection. Nasal obstruction, which alternates with the nasal cycle, is common to both allergic and infective causes. Nasal crusting and/or bleeding may occur in granulomatous disorders, atrophic rhinitis, or, rarely, tumours (particularly if associated with persistent unilateral symptoms). Impaired taste and/or smell may occur with many forms of rhinitis. It is particularly common with nasal polyposis and may occasionally follow trauma (olfactory nerve damage).

The presence of infertility and recurrent respiratory infections (including bronchiectasis) should raise the possibility of mucus abnormalities (Young's syndrome or cystic fibrosis) or ciliary dysfunction (primary ciliary dyskinesia, Kartagener's syndrome). Recurrent respiratory infections or a history of chronic rhinosinusitis should

also raise the possibility of immune deficiency disorders including hypogammaglobulinaemia and acquired immune deficiency syndrome (AIDS).

Hormonal imbalance (premenstrual symptoms, pregnancy, hypothyroidism, or acromegaly) may be associated with rhinitis. A history of trauma or previous nasal surgery should be sought.

Enquiry regarding associated chest disease is important. Rhinitis and asthma frequently coexist and recognition and appropriate treatment of rhinitis may improve asthma control. The efficacy, frequency, and regularity of previous treatments should also be considered, as should the patient's perception of possible side-effects of treatment, a frequently missed cause of poor compliance.

Examination

Local examination may be performed with a head mirror and speculum or an auroscope. Allergic rhinitis is accompanied by a pale bluish 'boggy' appearance of the nasal mucosa only if the patient has current symptoms. A red inflamed appearance with pus suggests an infective cause. A granular appearance with fine pale nodules is diagnostic of sarcoidosis. Enlarged turbinates may be confused with polyps by the unwary. If doubt exists, further examination with a rigid and/or flexible endoscope should be performed. The identification of structural abnormalities such as polyps, deflected nasal septum, or enlarged turbinates is important: surgical treatment may be indicated (a major advance has been the development of minimally invasive endoscopic sinus surgery).

Examination of the nose should also include tests of smell and examination of the ears, eyes, mouth, and throat. Examination of the chest and a general examination should be performed when indicated, in view of the common association of nasal disease with lower respiratory and systemic conditions.

Investigations

Skin-prick tests

In the presence of a clear history, particularly of seasonal hay fever symptoms, skin-prick testing is not essential. However, skin-prick tests are useful for several reasons (Table 2). They should only be interpreted in conjunction with the clinical history, and not performed when the patient is taking antihistamines, if 'dermographism' (wealing in response to pressure) is present, or in the presence of severe eczema. In these circumstances measurement of serum IgE antibodies by radioallergosorbent test (**RAST**) or enzyme-linked immunosorbent assay (**ELISA**) is indicated. A useful basic skin-prick testing kit should include the following:

1. a positive control (histamine 10 mg/ml);
2. negative control (allergen diluent solution);
3. house dust mite (*D. pteronyssinus*);
4. grass pollen;
5. cat fur;
6. *Aspergillus fumigatus*.

Skin-prick tests should be performed with a sterile 23-gauge needle or lancet, which is lightly inserted through the epidermis without inducing bleeding. Responses are recorded as the mean weal diameter at 15 min. A positive prick test is defined as a weal diameter 3 mm or more greater than that of the negative control test.

Treatment

Treatment for allergic rhinitis involves the avoidance of provoking allergens where possible and the use of topical corticosteroids and H1 selective antihistamines. Allergen immunotherapy has a place in patients who do not respond to these measures. The approach is summarized in Table 3.

Allergen avoidance

It is impossible to avoid pollens, although sensible advice includes wearing sunglasses and keeping car windows tightly shut. All windows should be kept closed, particularly in high buildings. Walking in parks and wide open spaces should be avoided, particularly during the late afternoon or evening when pollen counts are highest. A holiday by the sea or abroad during the peak pollen season may be helpful.

House dust mite control and avoidance measures should be undertaken in the homes of sensitive individuals with disease. Precise advice concerning the bedroom can be provided, with avoidance of non-synthetic bedding, restriction of soft toys, which should be washable, the use of mattress covers, changes to vinyl or cork flooring, and thorough vacuum cleaning and damp-dusting at least once weekly. A leaflet entitled *House dust mites: avoidance measures for allergy sufferers* is available from the British Allergy Foundation, Deepdene House, 30 Bellgrove Road, Welling, Kent DA16 3BY. Treatment for adults should be concentrated on the bedroom and living room, while measures for mite-sensitive children should be extended to all parts of the home. Measures to eradicate mites and allergens should be undertaken only following proper diagnosis and with appropriate medical supervision. There is no firm evidence to recommend the additional use of air conditioners, air ionizers, or acaracides. A recent meta-analysis has questioned the value of avoidance measures in mite-allergic patients. Where animal exposure is relevant, there is frequent resistance to advice to remove a family pet. However, patients can be advised to avoid replacing animals, to confine them to the kitchen or outdoors where possible, and to avoid contact with them or contaminated clothing. Recent evidence suggests that washing both cats and dogs may be effective in reducing pet allergen exposure.

Pharmacotherapy

The availability of potent specific histamine H1 receptor antagonists with a low potential for anticholinergic side-effects and a low sedative profile has been a major advance. Antihistamines are particularly effective for sneezing, itching, and rhinorrhoea, but unlike topical corticosteroids they have less effect on nasal blockage. They are also effective for eye and throat symptoms.

A rare but important complication of terfenadine is prolongation of the QT interval on the electrocardiogram (**ECG**). This only occurs when doses in excess of those recommended are employed, or in the presence of hepatic impairment or concomitant use of ketoconazole or erythromycin, both of which modify the hepatic metabolism of terfenadine. Astemizole may have the same effect in overdose. Acrivastine, loratadine, des-loratidine, cetirizine, ebastine, fexofenadine, and mizolastine are effective second-generation H1 antihistamines with an extremely low (or absent) potential for cardiac side-effects. H1-selective antihistamines can also be given as a topical nasal spray (levocabastine, azelastine). Antihistamines should be avoided during pregnancy.

Topical corticosteroids are highly effective in the majority of hay fever sufferers. Preparations include beclometasone, budesonide, fluticasone, triamcinolone, and mometasone. Aqueous formulations are better tolerated and have a better local distribution in the nose. Treatment should begin before the hay fever season for maximal effect, and the importance of regular treatment, even when symptoms are absent, should be emphasized. Side-effects are minor. Systemic effects are virtually absent at conventional doses, but caution should be exercised in children, particularly those receiving additional corticosteroids by other routes (for example, for associated asthma and/or eczema).

The topical anticholinergic agent ipratropium bromide is a potent inhibitor of glandular secretion and may be effective where watery nasal discharge is the dominant symptom, uncontrolled by the measures described above.

Sodium cromoglycate is available as a topical nasal spray for use four times daily. It is less effective than topical corticosteroids. Topical cromoglycate eye drops are effective for allergic eye symptoms in the majority of patients. Topical nedocromil sodium eyedrops have the advantage of a longer duration of action, allowing twice daily administration.

In the small proportion of patients whose symptoms are not otherwise controlled, there is a place for a short course of prednisolone (20 mg daily for 5 days). This approach may unblock the nose, thereby improving access for topical corticosteroids, which may then be more effective.

Topical decongestants (oxymetazoline) are effective in treating nasal blockage, although they should only be used for short periods (no more than 2 weeks) in view of

the risk of tachyphylaxis and rebound persistent nasal blockage (so-called rhinitis medicamentosa).

Allergen immunotherapy

In patients with severe summer hay fever unresponsive to topical corticosteroids and antihistamines and in those reluctant to take long-term medication, immunotherapy (desensitization) is an alternative treatment option. Immunotherapy involves the subcutaneous injection of increasing concentrations of allergen (standardized pollen extract) at weekly intervals for 6 to 12 weeks, followed by monthly injections of a maintenance dose for 3 to 5 years. It should only be given by those who are properly trained, and with epinephrine (adrenaline) and facilities for cardiopulmonary resuscitation immediately available. Patients should be kept under medical observation for at least 60 min following injections.

Recent controlled studies have confirmed the efficacy of immunotherapy, particularly for patients with grass pollen-induced summer hay fever (WHO guidelines). It is less effective in those with perennial rhinitis and asthma, where the disease is frequently heterogeneous with multiple allergic sensitivities and/or other causes of ongoing symptoms. The risk/benefit is less favourable in patients with chronic bronchial asthma in whom the risks of systemic adverse reactions are greater. Recent data suggests that pollen immunotherapy may confer long-term benefit. In patients who had received 3 to 4 years' treatment, clinical improvement was maintained for at least 3 years following discontinuation. This suggests that allergen immunotherapy, unlike pharmacotherapy, confers long-term benefit and has the potential to modify the course of the disease.

Future prospects for immunotherapy include the use of alternative safer routes such as local nasal or sublingual immunotherapy. Low molecular weight allergen peptides represent an alternative strategy for vaccine development, with the potential to modify T-cell responses with clinical benefit without the potential for IgE crosslinking and attendant risk of serious IgE-mediated side-effects.

Further reading

Abramson M, Puy R, Weiner J. (1999). Immunotherapy in asthma: an updated systematic review. *Allergy* **54**, 1022–41.

Akdis CA, Blaser K (1999). IL-10-induced anergy in peripheral T cell and reactivation by microenvironmental cytokines: two key steps in specific immunotherapy. *FASEB Journal* **13**, 603–9.

Bousquet J, Lockey RF, Malling HJ (1998). WHO position paper: allergen immunotherapy: therapeutic vaccination for allergic diseases. *Allergy* **53**, Supplement 44.

Colloff MJ, *et al*. (1992). The control of allergens of dust mites and domestic pets: a position paper. *Clinical and Experimental Allergy* **22**(Suppl. 2), 1–28.

Durham SR, Till SJ (1998). Immunologic mechanisms associated with allergen immunotherapy. *Journal of Allergy and Clinical Immunology* **102**(2), 157–64.

Durham SR, *et al.* (1999). Long term clinical efficacy of grass pollen immunotherapy. *New England Journal of Medicine* **341**, 468–75.

Fleming DM, Crombie DL (1987). Prevalence of asthma and hayfever in England and Wales. *British Medical Journal* **294**, 279–83.

Gotzsche PC, Hammarquist C, Burr M (1998). House dust mite control measures in the management of asthma: meta-analysis. *British Medical Journal* **317**, 1105–10.

Lund V, *et al*. (1994). International consensus report on the diagnosis and management of rhinitis. *Allergy* **49**(Suppl. 19), 1–34.

17.4.3Basic mechanisms and pathophysiology of asthma

Tak H. Lee

Introduction

Bronchial asthma is characterized by episodic wheezing, airways obstruction that is reversible, either spontaneously or with therapy, bronchial hyper-responsiveness to non-specific stimuli, and airways inflammation.

Risk factors for the development of asthma

Atopy is the strongest risk factor for the development of asthma. The most important mechanism by which IgE determines the expression of atopy is through its binding to high-affinity receptors (Fce-RI) expressed on the surface of tissue mast cells and basophils, and to lower-affinity receptors (Fce-RII or CD23) on macrophages, eosinophils, and platelets. Crosslinkage of IgE with a specific allergen results in the non-cytotoxic release of an array of preformed and newly generated mediators of inflammation.

The domestic house dust mite (**HDM**, *Dermatophagoides pteronyssinus*) is the major allergenic cause of perennial asthma. To date, seven groups of HDM allergens have been identified, the first four of which are known to exhibit proteolytic or other enzymatic activities. For example, Der p1, the major allergen of *D. pteronyssinus,* is a cysteine protease derived from the mite's gastrointestinal tract, whereas Der p2 is a lysozyme and Der p3 a chymotryptic enzyme. The potent biological activities of other allergens might explain why these particular proteins are able to penetrate epithelial surfaces so easily and lead to specific sensitization.

Maternal smoking, both before and after birth, is a risk factor for developing respiratory disease early in life. Exposure to environmental tobacco smoke has been shown to increase IgE in adults, and some studies have suggested that maternal smoking in pregnancy increases cord-blood IgE and the subsequent risk of atopic disease.

Other factors implicated in the early-life origins of asthma include respiratory-tract virus infections, particularly with the respiratory syncytial virus (**RSV**), possibly through damage to the bronchial epithelium, thereby augmenting the penetration of the airway mucosa by inhaled allergens. Exposure to environmental air pollutants, such as ozone, sulphur dioxide, and oxides of nitrogen, has recently been shown to enhance allergen sensitization of the lower respiratory tract.

Pathology

Histological examination of the airway tissue in an asthmatic lung shows the presence of an inflammatory reaction, with extensive remodelling of the airway wall. The inflammation is a multicellular process; even in the mildest of asthmatic individuals there is *in vivo* evidence of infiltration of the bronchial mucosa with mast cells, mononuclear cells, and granulocytes, of which the eosinophil is prominent. The tenacious plugs that fill the lumen are an exudate of plasma and inflammatory cells, particularly eosinophils, which have migrated into the lumen, as well as epithelial cells that have sloughed from the airway surface, often leaving the basement membrane denuded. The basement membrane appears thickened when viewed under the light microscope, which is due to an increase in the amount of collagen deposition at this site, but the lamina densa, which forms the true basement membrane, is normal when observed with the electron microscope.

The changes in the lumen and wall of the airways of asthmatic lungs are reflected in the cytological examination of sputum from patients with the disease. Early studies of the sputum showed the presence of Creola bodies (sloughed epithelial clumps), Charcot-Leyden crystals (remnants of eosinophils), and Curschmann spirals (casts of the airway formed by the exudate). More recent studies have established that the eosinophil is the prominent cell found in the sputum of asthmatic patients.

There is a clear increase in wall thickness throughout the bronchial tree in patients with asthma, and all the tissue layers participate in this generalized increase in airway wall thickness. The volume of the bronchial microvasculature is also increased in both the submucosa and the adventitia of the airways. It has been calculated that the thickening of the airway wall caused by asthma has only a minor effect on the lumen of a fully dilated airway, but when a modest increase in wall thickness is associated with smooth-muscle shortening, the two factors acting together produce markedly increased airway resistance.

Inflammatory cells

Bronchial inflammation is orchestrated by a network of cytokines and growth factors, including those encoded by the **GM-CSF/IL-4/IL-5** (granulocyte–macrophage colony-stimulating factor/interleukin-4/-5) gene cluster on chromosome 5, derived from both inflammatory and structural cells in the airways.

Mast cells

Mast cells are clearly important in initiating the acute bronchoconstrictor responses to allergen, and probably to other indirect stimuli such as exercise and hyperventilation (via osmolarity or thermal changes). When sensitized subjects inhale specific allergen it causes both early (5–15 min—early airway response, **EAR**) and late (2–6 h—late airway response, **LAR**) bronchoconstrictor responses, which last approximately 60 min and between 12 and 24 h, respectively. The LAR is accompanied by an acquired increase in bronchial responsiveness to such stimuli as inhaled histamine and methacholine (Fig. 1).

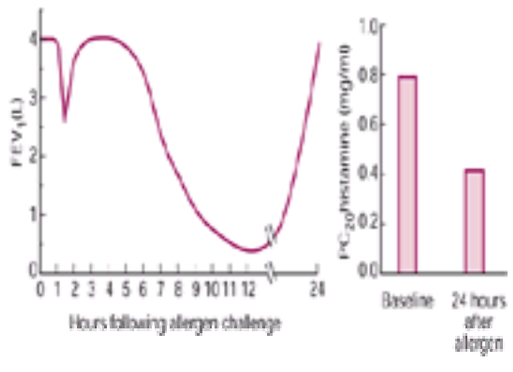

Fig. 1 Changes in **FEV_1** (forced expiratory volume in 1 second) and histamine airway responsiveness after allergen inhalation challenge in an asthmatic subject. PC_{20} histamine is the concentration of histamine producing a 20 per cent decrease in FEV_1. It is a measure of 'non-specific' airway responsiveness.

Measurement of mediators in the peripheral blood and bronchoalveolar lavage fluid together with their metabolites in urine has shown that the EAR is a mast cell-dependent response resulting from the IgE-dependent secretion of constrictor substances. The type of mast cell involved in this reaction contains predominantly tryptase as its neutral protease. Among its biological actions, tryptase is able to produce a prolonged increase in microvascular permeability, upregulate adhesion molecules, attract and activate eosinophils, and augment epithelial and fibroblast proliferation. Histamine exerts most of its airway effects via H1 receptors, which are present both on airway smooth muscle and on the microvasculature, whilst prostaglandin D2 (**PGD2**) contracts airway smooth muscle by interacting with thromboxane (TP1) receptors. Once released from mast cells, LTC_4 is rapidly metabolized to LTD_4 and subsequently to LTE_4, the three cysteinyl leukotrienes responsible for the smooth-muscle contractile and vasoactive properties of the biological activity previously described as **SRS-A** (slow-reacting substance of anaphylaxis) (Fig. 2).

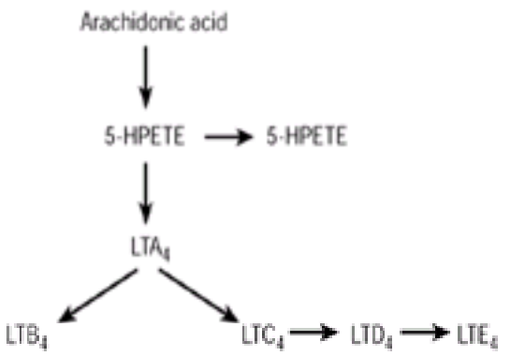

Fig. 2 The 5-lipoxygenase pathway. 5-HPETE = 5-hydroperoxy-eicosatetraenoic acid, 5-HETE = 5-hydroxy-eicosatetraenoic acid. LT = leukotriene.

Eosinophils

The late airway response has an inflammatory basis. During this response there is an increased bronchoalveolar lavage eosinophilia, suggesting the selective recruitment of these cells into airway tissue from the microvasculature. Eosinophil infiltration is a characteristic feature of asthmatic airways and differentiates asthma from other inflammatory airway conditions. There is a close relationship between eosinophil counts in peripheral blood or bronchoalveolar lavage (**BAL**) fluid and airway hyper-responsiveness.

Eosinophil migration may be due to the effects of lipid mediators, such as leukotrienes, or to the effects of cytokines such as GM-CSF, IL-5, RANTES, and eotaxin (**RANTES**, regulated upon activation, normal T-cell expressed and secreted chemokine). Several of these molecules might also be very important for the survival of eosinophils in the airways and may 'prime' eosinophils to exhibit enhanced responsiveness. There appears to be a co-operative interaction between IL-5 and chemokines, so that both cytokines are necessary for the eosinophilic response in airways. Once recruited to the airways, eosinophils require the presence of various growth factors, of which GM-CSF and IL-5 appear to be the most important. In the absence of these growth factors eosinophils undergo programmed cell death (apoptosis).

Within the airways, active eosinophils secrete a wide array of preformed and newly generated inflammatory products. These compromise the toxic granule components of the eosinophil (major basic protein, eosinophil cationic protein, and eosinophil-derived neurotoxin) and a range of lipid products, including leukotrienes and platelet activating factor (**PAF**). The administration of LTD_4 antagonists prior to allergen provocation of sensitized airways produces marked inhibition of both the EAR and LAR and attenuation of the acquired increase in bronchial hyper-responsiveness. Although PAF was at one time regarded as a prime mediator of the late-phase inflammatory response and bronchial hyper-responsiveness, investigation of the orally active PAF receptor antagonist WEB 2086 has failed to reveal any inhibitory effect on either early- or late-phase, allergen-induced airway events. Eosinophils are also involved in the generation of oxygen-derived free radicals.

Other leucocytes

The mechanism(s) by which leucocytes move into the airway and become activated have attracted considerable interest. At 6 h following allergen challenge, there is marked upregulation of E-selectin, whose ligand on neutrophils and other leucocytes is sialyl-Lewis × and intercellular adhesion molecule-1 (**ICAM-1**), a member of the immunoglobulin superfamily. One ligand for ICAM-1 is lymphocyte function-associated antigen-1 (**LFA-1**) (an integrin heterodimer CD11a–CD18) expressed on a large number of leucocytes, but especially on lymphocytes, neutrophils, and eosinophils. Another member of the immunoglobulin superfamily, vascular cell-adhesion molecule-1 (**VCAM-1**), is expressed in the airway microvasculature at a low level, but this is not increased within the time frame of 6 h following allergen provocation. A positive correlation has been observed between the extent of ICAM-1 expression and leucocyte infiltration, and more specifically between E-selectin and the increase in neutrophil numbers, suggesting an important role for these molecules in the allergic inflammatory process.

The initial expression of P-, L-, and E-selectins, which contain lectin binding regions that interact with carbohydrate ligands and leucocytes (e.g. sialyl-Lewis ×) results in the rolling of leucocytes along the endothelial cell, whereas upregulation of ICAM-1 and VCAM-1 arrests the leucocytes, thereby facilitating transendothelial migration. In non-human primates naturally sensitized to *Ascaris* antigen, blocking antibodies directed to E-selectin and ICAM-1 abrogate the LAR and the resultant bronchial hyper-responsiveness following allergen challenge, in parallel with a reduction in neutrophils and eosinophils.

The role of neutrophils in human asthma is less clear. They are found in the airways of patients with chronic bronchitis and bronchiectasis who do not have the degree of airway hyper-responsiveness found in patients with asthma. Neutrophils are rarely seen in the airways of patients with chronic asthma, but large numbers are seen in those who die suddenly of asthma. This may reflect the rapid kinetics of neutrophil recruitment compared with eosinophil inflammation.

Macrophages, which are derived from blood monocytes, traffic into the airways and may orchestrate the inflammatory response. Macrophages may both increase and decrease inflammation, depending on the stimulus. Alveolar macrophages normally have a suppressive effect on lymphocyte function and may play an important role in preventing the development of allergic inflammation, but this may be impaired in asthma after allergen exposure. Macrophages can also act as antigen-presenting cells (**APCs**), processing allergen for presentation to T-lymphocytes, but alveolar macrophages are far less effective in this respect than macrophages from other sites. By contrast, dendritic cells, which are specialized macrophage-like cells in the airway epithelium, are very effective antigen-presenting cells and might play a very important role in the initiation of allergen-induced responses in asthma.

T lymphocytes play a very important role in co-ordinating the inflammatory response in asthma through the release of specific patterns of cytokines, resulting in the recruitment and survival of eosinophils and the maintenance of mast cells in the airways. T lymphocytes are coded to express a distinctive pattern of cytokines, which may be similar to that described in the murine TH_2 type of T lymphocyte that characteristically express IL-4, IL-5, and IL-13. There appears to be an imbalance of TH cells in asthma, with the balance in favour of TH2 cells. The balance between TH1 cells and TH2 cells may be determined by locally released cytokines such as IL-12, favouring the expression of TH1 cells, and IL-4, favouring TH2 cells. There is a suggestion that early infections might encourage TH1-mediated responses to predominate and that a lack of infection in childhood may favour TH2-cell expression and thus atopic diseases.

Structural cells

Structural cells of the airways, including epithelial cells, fibroblasts, and airway smooth muscle cells, are also important sources of inflammatory mediators, such as cytokines and lipid mediators, in asthma. In addition, epithelial cells may play a key role in translating inhaled environmental signals into an airway inflammatory

response and are probably a major target cell for inhaled glucocorticoids.

Inflammatory mediators

Many different mediators have been implicated in asthma. Those such as histamine, prostaglandins, and leukotrienes contract airway smooth muscle, increase microvascular leakage, increase airway mucus secretion, and attract other inflammatory cells. The availability of specific receptor antagonists has defined the critical role of certain mediators. For instance, the cysteinyl leukotrienes LTC_4, LTD_4, and LTE_4, are potent constrictors of human airways and have been reported to increase airway hyper-responsiveness, recruit eosinophils, and possibly play an important role in asthma. Potent LTD_4 antagonists protect against exercise-, aspirin-, and allergen-induced bronchoconstriction, suggesting that cysteinyl leukotrienes contribute to bronchoconstrictor responses. Chronic treatment with cysteinyl leukotriene antagonists improves lung function and symptoms in asthmatic patients.

Cytokines

Cytokines play a central role in orchestrating the type of inflammatory response. Many inflammatory cells (macrophages, mast cells, eosinophils, and lymphocytes) are capable of synthesizing and releasing these proteins, and structural cells such as epithelial cells and endothelial cells may also release a variety of cytokines and therefore participate in the chronic inflammatory response. While inflammatory mediators like histamine and leukotrienes may be important in the acute and subacute inflammatory responses and in exacerbations of asthma, it is likely that cytokines play a dominant role in chronic inflammation.

As atopy is a major risk factor for the development of asthma, it is essential to understand how IgE synthesis is regulated. IL-4 interacts with B cells via specific cell-surface receptors, and is a key cytokine involved in the isotype-switching of B cells from synthesizing IgM and IgG to IgE. An important accessory signal for IgE-switching is provided by CD40 on T cells signalling through the CD40 ligand on B cells. IL-13, which exhibits 30 per cent homology with IL-4, has also been shown to mediate IgE isotype-switching through its own specific receptors but, unlike IL-4, it is also a differentiation factor for dendritic cells. Switching of B cells to IgE synthesis is potently inhibited by interferon-gamma (IFN-g) from TH_1 cells and monocytes/macrophages.

In vivo evidence for cytokine involvement in asthma

Bronchoalveolar lavage (BAL)

Increased proportions of cells positive for IL-4 and IL-5 mRNA are found using the technique of *in situ* hybridization in atopic asthmatic subjects. Symptomatic asthmatic subjects have greater proportions of cells positive for IL-3, IL-4, IL-5, and GM-CSF mRNA in BAL fluid than asymptomatic asthmatics. No differences between the groups in the numbers of cells expressing IL-2 and IFN-g mRNA are detected. In addition, there are significant associations between the number of cells expressing mRNA for IL-4, IL-5, and GM-CSF and baseline airflow obstruction, airway hyper-responsiveness, and asthmatic symptoms. Increased levels of tumour-necrosis factor-a (**TNF-a**), GM-CSF, and IL-6 have been found in BAL fluid of symptomatic as compared to asymptomatic asthmatic subjects.

Studies on alveolar macrophages of asthmatic subjects showed that IL-1b expression is upregulated and that the level of IL-1b in BAL fluid of subjects with symptomatic asthma is higher than that of normal subjects or subjects with asymptomatic asthma. Additionally, bronchoalveolar lavage fluid from symptomatic non-allergic asthmatic subjects contains elevated levels of IL-1.

Evidence from challenge models further supports cytokine involvement in asthma. Increases have been shown in the number of cells expressing mRNA for the TH2 cytokines IL-4, IL-5, and GM-CSF but not IL-3, IL-2, or IFN-a after allergen challenge. Furthermore, IL-4 and IL-5 mRNAs transcripts were associated with activated CD4+ T cells.

Bronchial biopsies

Immunohistochemical analysis of bronchial biopsies from symptomatic and asymptomatic atopic and non-atopic asthmatics reveals increased immunoreactivity with IL-1/2/3/4/5, GM-CSF, **MCP-1** (monocyte chemotaxis protein-1), and TNF-a as opposed to asymptomatic control subjects.

Mucosal biopsies obtained from atopic asthmatics after allergen challenge demonstrate an increased influx of activated eosinophils and activated CD4+ T cells, and an increase in the number of cells expressing mRNA for IL-5 and GM-CSF. There is a significant inverse correlation between the numbers of cells expressing mRNA for IL-4 and IFN-a. This supports the hypothesis that allergen-induced late asthmatic responses are accompanied by T-cell activation, cytokine mRNA expression for IL-5 and GM-CSF, and local recruitment and activation of the eosinophils in the bronchial mucosa.

Effects of inflammation

The chronic inflammatory response may alter the structure and function of critical target cells in the airways.

Airway epithelium

Airway epithelial shedding may be important in contributing to airway hyper-responsiveness, and may explain how several different mechanisms (for example, ozone exposure, certain virus infections, chemical sensitizers, and allergen exposure) can lead to the development of this condition. All these stimuli may lead to epithelial disruption, which may contribute to airway hyper-responsiveness by the loss of barrier function to allow penetration of allergens, loss of enzymes (such as neutral endopeptidase) that normally degrade inflammatory mediators, and exposure of sensory nerves, which might lead to reflex neural effects on the airway.

Airway smooth muscle

An abnormality in smooth muscle function is thought to be the basis of the bronchial hyper-responsiveness in asthma, but studies of isolated human bronchi from asthmatic subjects have failed to demonstrate a clear consensus for the presence of a functional abnormality. Of eight reports of individuals who are hyper-responsive to histamine and methacholine *in vivo*, half have shown increased force generation by asthmatic airways *in vitro*, compared to control airway preparations; the remainder report no difference in force generation by asthmatic or normal isolated airway preparations. The reasons for this anomaly are unclear. In general, studies have been carried out using airways from mild asthmatics and differences in optimal length for maximum force generation and smooth muscle content have not been accounted for. Similarly, passive sensitization studies have produced conflicting data. In three of four studies, exposure of airways removed from non-atopic patients to serum from patients with high IgE levels confers increased responsiveness to specific allergen and hyper-responsiveness to non-specific stimuli such as histamine and neuropeptides.

Other functional changes may be important. Reduced responsiveness to b_2-adrenoceptor agonists has been reported in both postmortem and surgically removed bronchi from asthmatic patients, although the number of b-adrenoceptors is not reduced, suggesting that these receptors have in some way become functionally uncoupled. This aspect of altered airway smooth muscle function can be modelled *in vitro*, where treatment of airway smooth muscle cells with cytokines induces a similar refractoriness to b_2-adrenoceptor agonists.

Changes in compliance of the extracellular connective tissue components between individual muscle cells and in the stiffness of parallel elastic elements in the extracellular compartment may reduce the tethering loads of the muscle and allow greater shortening. Indeed, treatment of human airway smooth muscle preparations with the matrix degrading enzyme collagenase enhances force generation and shortening.

It seems likely that the involvement of the smooth muscle in excessive airway narrowing in asthma involves several mechanisms. These include an increased content of smooth muscle in the airway wall, or a combination of increased content and altered function. It is now recognized that, in addition to contractile function, airway smooth muscle can undergo hyperplasia and/or hypertrophy, leading to apparently irreversible changes in wall structure and contributing to the development of persistent airway obstruction and non-specific airway hyper-responsiveness. Numerous studies have characterized some of the proinflammatory mediators and growth factors that elicit proliferation of smooth muscle. Other trophic factors such as altered mechanical stress and reactive oxygen species have also been

identified. Components of the extracellular matrix that are increased in asthma may also impact on the proliferation of airway smooth muscle.

Proliferating airway smooth muscle cells undergo phenotypic modulation from a contractile to synthetic–proliferative state, where additional functions of airway smooth muscle such as cytokine/chemokine and extracellular matrix secretion become more apparent. This is likely to be of particular relevance in the diseased asthmatic lung where the content of airway smooth muscle as a fraction of the total cells in the airway wall is already increased. Human airway smooth muscle releases several chemokines important for the activation of eosinophils. These include chemotaxins such as RANTES and eotaxin, as well as GM-CSF. Airway smooth muscle cells also interact by direct contact with immunocytes such as T lymphocytes through the expression of cell-adhesion molecules and the induction of myocyte DNA synthesis. T lymphocytes derived from bronchoalveolar lavage fluid following antigen challenge of atopic subjects adhere to human airway smooth muscle cells in culture, inducing ICAM-1 and HLA-DR expression on the smooth muscle cells.

As additional functions of airway smooth muscle are described, a view (Fig. 3) is emerging that airway smooth muscle cells may adopt an immunoeffector role in chronic asthma by proliferating, secreting cytokines, expressing adhesion molecules, and by interacting with various immunocytes. This may involve changes in the phenotypic status of airway smooth muscle, and, as a result, these cells may play an active role in perpetuating and orchestrating airway inflammation in the remodelled airway. These novel insights question previously held paradigms of altered neural activity and inflammation in asthma, and emphasize the importance of the response of the end-organ, namely the smooth muscle cell.

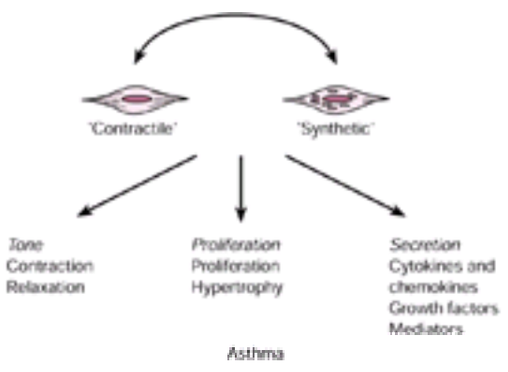

Fig. 3 The role of airway smooth muscle in chronic severe asthma (by courtesy of S. Hirst).

Vascular responses

Vasodilatation occurs in inflammation, yet little is known about the role of the airway circulation in asthma, partly because of the difficulties involved in measuring airway blood flow. The bronchial circulation may play an important role in regulating airway calibre, since an increase in the vascular volume may contribute to airway narrowing. Increased airway blood flow may be important in removing inflammatory mediators from the airway, and may play a role in the development of exercise-induced asthma.

Microvascular leakage is also an essential component of the inflammatory response. It may increase airway secretions, impair mucociliary clearance and enhance mucosal oedema, thereby contributing to airway narrowing and increased airway hyper-responsiveness.

Neural effects

Non-adrenergic, non-cholinergic (**NANC**) nerves and several neuropeptides have been identified in the respiratory tract, in addition to those involved in classical cholinergic and adrenergic mechanisms. It is unclear whether abnormalities of autonomic function in asthma are secondary to the disease or primary defects. It is recognized that inflammation may interact with autonomic control by several mechanisms. For instance, neuropeptides such as substance P, neurokinin A, and calcitonin gene-related peptide may be released from sensitized inflammatory nerves in the airways and may amplify the inflammatory response. There is evidence for an increase in substance P-immunoreactive nerves in the airways of patients with severe asthma, although this has not been confirmed in patients with mild disease. There may also be a reduction in the activity of enzymes such as neutral endopeptidase which degrade neuropeptides such as substance P. There may also be increased expression of the receptor that mediates the inflammatory effects of substance P.

Further reading

Arm JP, Lee TH (1993). Chemical mediators II: Leukotrienes and eicosanoids. In: Weiss EB, Stein M, eds. *Bronchial asthma: mechanisms and therapeutics*, 3rd edn., pp. 112–34. Little, Brown Co.

Cockcroft DW, Murdock KY (1987). Changes in bronchial responsiveness to histamine at intervals after allergen challenge. *Thorax* **42**, 302–8.

Israel E, *et al.* (1993). The pivotal role of 5-lipoxygenase products in the reaction of aspirin-sensitive asthmatics to aspirin. *American Review of Respiratory Disease* **148**, 1447–51.

Jeffery PK, *et al.* (1989). Bronchial biopsies in asthma: an ultrastructural quantification study and correlation with hyperreactivity. *American Review of Respiratory Disease* **140**, 1745–53.

Johnson SR, Knox AJ (1997). Synthetic functions of airway smooth muscle in asthma. *Trends in Pharmacological Science* **18**, 288–92.

Lambert RK, *et al*. (1993). Functional significance of increased airway smooth muscle in asthma and COPD. *Journal of Applied Physiology* **74**, 2771–81.

Lee TH (1998). Cytokine networks in the pathogenesis of bronchial asthma: implications for therapy. *Journal of the Royal College of Physicians of London* **32**, 56–64.

Moreno RH, Hogg JC, Pare PD (1986). Mechanisms of airway narrowing. *American Review of Respiratory Disease* **133**, 1171–80.

Panettieri RA (1998). Cellular and molecular mechanisms regulating airway smooth muscle proliferation and cell adhesion molecule expression. *American Journal of Respiratory and Critical Care Medicine* **158**, S133–S140.

Rabe KF (1998). Mechanisms of immune sensitisation of human bronchus. *American Journal of Respiratory and Critical Care Medicine* **158**, S161–S170.

Reiss TF, *et al.* (1998). Montelukast, a once-daily leukotriene receptor antagonist, in the treatment of chronic asthma: a multicenter randomized, double-blind trial. *Archives of Internal Medicine* **158**, 1213–20.

Robinson DS, *et al.* (1992). Evidence for a predominant Th_2-type bronchoalveolar T-lymphocyte population in atopic asthma. *New England Journal of Medicine* **326**, 298–304.

Seow CY, Schellenberg RR, Pare PD (1998). Structural and functional changes in the airway smooth muscle of asthmatic subjects. *American Journal of Respiratory and Critical Care Medicine* **158**, S179–S186.

Shirakawa T, *et al.* (1997). The inverse association between tuberculin responses and atopic disorder. *Science* **275**, 77–9.

Stewart GA (1994). The molecular biology of allergens. In: Holgate ST, Busse W, eds. *The mechanisms of asthma and rhinitis*, pp. 898–932. Blackwell Science, Boston.

Wegner CD, *et al.* (1996). ICAM-1 in the pathogenesis of asthma. *Science* **247**, 416–18.

17.4.4

Asthma

A. J. Newman Taylor

Introduction

Asthma is a chronic inflammatory disease of the bronchial airways, which is characterized by a desquamative eosinophilic bronchitis (Fig. 1). The defining clinical characteristics of asthma—reversible airway narrowing and increased airway responsiveness to non-specific provocative stimuli—are manifestations of the underlying chronic inflammatory process. Definitions of asthma that have focused on these clinical characteristics, to distinguish it from diseases associated with predominantly irreversible airway narrowing, have emphasized the intermittent nature of asthma rather than the persistence of the underlying inflammation, with potentially inappropriate implications for treatment.

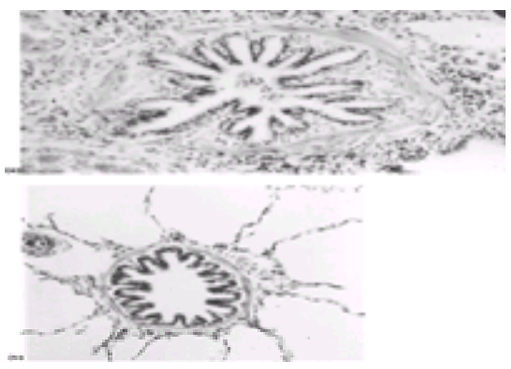

Fig. 1 The defining pathology of asthma: (a) desquamative eosinophilic bronchitis in patient with asthma, in comparison with (b) normal histological appearances.

The recognition that asthma is a chronic inflammatory disease implies that, in addition to identifying and avoiding inducing causes, such as domestic pets and occupational sensitizers, disease control is likely to require long-term anti-inflammatory treatment. Appreciation of the inflammatory nature of asthma has also led to recognition of the associated injury and damage to the airway wall—airway remodelling—which may lead to irreversible loss of function, and be preventable by the early institution of anti-inflammatory treatment.

Asthma and airway hyperresponsiveness: inducers and provokers

The distinguishing abnormalities of lung function in bronchial asthma are:

1. reversible airway narrowing and
2. airway hyperresponsiveness to non-specific provocative stimuli.

Airway responsiveness describes the ease with which acute airway narrowing can be provoked by a variety of stimuli. Non-specific provocative stimuli include exercise, inhalation of cold dry air, inhaled respiratory irritants such as sulphur dioxide, and pharmacological agents such as histamine and methacholine (Table 1). Provocation of asthma by specific allergens can induce hyperresponsiveness to non-specific stimuli, when smaller doses of such stimuli provoke acute airway narrowing. Inhaled non-specific provocative stimuli, such as histamine or methacholine, incite airway narrowing that usually resolves within minutes; exercise provokes asthma within minutes that resolves within 1 h.

The degree of airway responsiveness can be expressed as the dose or concentration of the stimulus which provokes a specified fall in the forced expiratory volume in 1 s (**FEV1**), commonly the dose or concentration of histamine or methacholine which provokes a 20 per cent fall in FEV_1—**PD20** or **PC20**, histamine or methacholine.

While provokers of asthma incite acute airway narrowing in individuals with hyperresponsive airways, inducers of asthma increase the magnitude of airway hyperresponsiveness and the clinical manifestations of asthma by increasing the severity of the underlying airway inflammation, which can persist for days or weeks. The principal inducers of asthma are inhaled allergens, low molecular weight chemicals encountered at work, and viral respiratory tract infections (Table 1).

Allergen inhalation tests are a good model of the airway response to an inducer and demonstrate the interrelationship between airway inflammation, airway narrowing, and airway hyperresponsiveness. Inhalation of an allergen by an individual allergic to it with asthma will provoke:

1. an immediate fall in FEV_1, which develops within minutes and usually resolves spontaneously within 1 to 1.5 h; and
2. a subsequent late fall in FEV_1, which develops in about 50 per cent of cases 1 h or more after the inhalation test and persists for several hours, on occasions for days.

The immediate fall in FEV_1 is IgE dependent and is due to airway smooth muscle contraction and airway wall oedema provoked by mediators, such as histamine,

released from mast cells resident in the airways. It is not associated with an increase in airway responsiveness. The late fall in FEV_1 is the outcome of recruitment to the airways of inflammatory cells, particularly T_{H2} lymphocytes and eosinophils, reducing airway calibre. It is associated with an increase in airway responsiveness (manifest as a reduction in PC_{20}) that can persist, with associated increased diurnal variation in airway calibre, for several days after resolution of airway narrowing (Fig. 2).

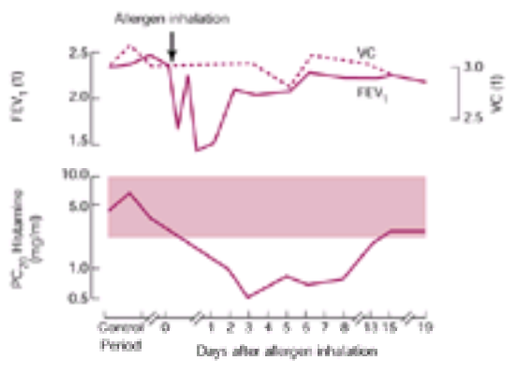

Fig. 2 Increased airway responsiveness associated with late asthmatic reaction provoked by inhalation of ragweed pollen.

Prevalence of asthma

Asthma is a common disease, frequently disabling and, uncommonly, the cause of death. In the Western world it now has an estimated prevalence of more than 10 per cent in children and more than 5 per cent in adults. In the United Kingdom in 1999 it was the cause of 86 000 hospital admissions and the certified cause of death of 1520 people.

The prevalence of asthma in children and young adults has increased markedly in the Western world during the past 20 to 30 years (Fig. 3). Although in part this increase may reflect a greater awareness of and tendency to diagnose asthma, repeat cross-sectional studies of children in the United Kingdom, using identical methods of ascertainment at different time points, have shown a definite increase in disease prevalence. A study of Aberdeen school children found the prevalence of wheeze and of diagnosed asthma had increased 2.5-fold between 1964 and 1989. A similar study in South Wales, made at two time points 15 years apart, found a history of reported asthma to have doubled from 6 to 12 per cent. This study also reported a similar increase of reported hay fever and eczema, and in the proportion of children in whom exercise provoked asthma.

Fig. 3 Increasing prevalence of asthma in Finnish military recruits.

Comparison of the prevalence of asthma in different parts of the world suggests that the high and increasing prevalence in the Western world is associated with urbanization and material prosperity. A study of school children in Zimbabwe found asthma to be uncommon in those living in a rural area, more common in poor urban dwellers, and most common in affluent urban dwellers, equally in the black and white population, in Harare. In Europe the reunification of Germany allowed comparison of the prevalence of asthma and associated conditions in cities in former East and West Germany. The prevalence of asthma, hay fever, eczema, and atopy (identified as immediate skin test responses to common inhalant allergens) was greater in school age children living in the West German city of Munich than in the East German cities of Leipzig and Halle. This greater prevalence of asthma, atopy, and atopic diseases in Western compared with Eastern Europe has been replicated in comparison studies between other countries (such as Finland compared with Estonia). Interestingly, the prevalence of atopy (particularly skin test responses to pollens) and hay fever, but not asthma, has subsequently increased in children now living in reunified Germany who had lived the first 5 years of their lives in Leipzig.

Many explanations have been advanced to explain these observations. These include increased indoor allergen exposure (particularly house dust mite and cat), increased exposure to vehicle exhaust pollution, increased tobacco smoking by women of childbearing age, changing diet, and reduced infection rates in childhood. Of these, outdoor air pollution has, until recently, attracted most public attention, although there is no substantive evidence in its support: the prevalence of asthma in urban parts of the United Kingdom is no greater (and possibly less) than in rural parts, including Skye, where measured levels of air pollutants are the lowest in the United Kingdom. Similarly, there is little evidence that the increased prevalence of asthma and other atopic disease has been caused by increased indoor allergen exposure or tobacco smoking, although the increase in asthma has been parallelled by an increase in cigarette smoking by women of childbearing age. Several dietary explanations have also been advanced, including increased salt and reduced antioxidant intake. The most plausible explanation advanced to date is that the increase in atopy and asthma is a consequence of a reduction in rates of infection during childhood. The evidence is both indirect and direct, although not yet conclusive. The most consistent observation, now from several studies, is of an inverse relationship between family size, and birth order, and the risk of atopy and hay fever. This has been interpreted as being consistent with the age at which a child encounters infectious agents having a decisive influence on the development of atopy and associated diseases: children in large families and those with older siblings are more likely to encounter infections earlier in life, reducing their risk of becoming atopic. Several, although not all, subsequent studies of the relationship of atopy and atopic disease to childhood infection have supported this hypothesis. In a group of children studied in Guinea-Bissau, West Africa the risk of being atopic (and in particular having specific IgE to *Dermatophagoides pternonyssinus*) was inversely related to having had measles in early childhood. Similarly, the risk of atopy, hay fever, and asthma was inversely related to having been infected with *Mycobacterium tuberculosis* in Japanese school children, and to having serum antibodies to hepatitis A, considered an indicator of hygiene in early life, in Italian military recruits.

Risk factors for asthma

The geographical variation and considerable increase in recent decades in the prevalence of asthma indicate important environmental influences on the development of the disease. There is also evidence for genetic susceptibility. The risk of asthma is associated with atopy, for which genetic influences are strong. As with many common complex diseases, such as diabetes mellitus, asthma is probably the outcome of multiple genes and their interaction with the environment.

Atopy

Atopy is defined as the production of specific IgE antibody to common inhalant allergens, such as grass pollen, house dust mite, and cat. Atopy may be identified by the presence of immediate skin prick test responses (or of specific IgE in serum) to extracts of common inhalant allergens and has a prevalence of some 40 per cent in the adult United Kingdom population. The risk of developing asthma as well as eczema and hay fever is increased in atopic individuals. In a random population sample in the south-west United States, a close relationship was found at all ages between skin test responses to local inhalant allergens and the prevalence of

asthma and allergic rhinitis. Similarly, in Canadian university students the prevalence of airway hyperresponsiveness to inhaled histamine correlated significantly with the degree of atopy. For further discussion see Chapter 17.4.1.1 and Chapter 17.4.1.2.

Genetic susceptibility

Asthma and atopy show clear indications of genetic susceptibility: the frequency of disease in family members is greater than in the population as a whole and is greater in identical than non-identical twins.

In one study the population prevalence of asthma was estimated at between 5 and 10 per cent; the risk for the child of a parent with asthma and atopy was 14 per cent when one parent was affected and 29 per cent when both parents were affected. The risk for the child of a parent with asthma but not atopy was little more than the population frequency of asthma; whereas atopy increased the risk of a child developing asthma some threefold. The reasons for genetic susceptibility are discussed in section 17.4.1.1, but it is important to recognize that genetic influences alone cannot explain the striking geographical variation and recent secular change in the prevalence of asthma.

Allergen exposure

In people with asthma, natural allergen exposure induces asthma and airway hyperresponsiveness. Both the severity of asthma and airway responsiveness increased in patients with asthma who were allergic to ragweed pollen during the ragweed pollen season. Similarly, avoidance of relevant allergen exposure is associated with an improvement or resolution of asthmatic symptoms, improved lung function, and decreased airway responsiveness. Patients with asthma who were allergic to house dust mite have shown considerable symptomatic and objective improvement when avoiding house dust mite for several months at altitude in Davos in the Swiss Alps; also for several weeks in a London hospital. In south-east United States, asthma deaths in patients allergic to the mould *Alternaria alternata* increase during the months of the year when *Alternaria* spore counts are highest.

Although there is clear evidence that natural exposure to allergens to which they are allergic can induce asthma in patients (and avoidance can improve it), the influence of allergen exposure on the development of asthma is less clear. Studies comparing populations born and living in different environments and climates, and therefore exposed to different allergens in childhood, demonstrate allergy and associated disease in relation to allergens present in the particular environment. Comparison of children born and living in Marseilles (humid and at sea level, encouraging the growth of house dust mite) and in Briancon, the highest town in the French Alps (not conducive to the growth of house dust mite, but encouraging to the growth of many flowering plants), showed that allergy to house dust mite was considerably more prevalent in Marseilles than in Briancon, whereas allergy to pollens was considerably more prevalent in Briancon than Marseilles. The prevalence of asthma was similar in both environments, although the associated allergies differed. The introduction of a new allergen into an environment can cause the development of allergy and asthma, particularly among adults. Unloading soya beans in Barcelona harbour caused 'epidemic days', when the number of hospital admissions with asthma increased several-fold: these continued for 7 years until the cause was identified and a filter placed on the silos to prevent the release and dissemination of soya bean during unloading.

Respiratory virus infections

While respiratory virus infections have long been suspected to be the major cause of exacerbations of asthma, it is only with the development and use of polymerase chain reaction in controlled studies that the true proportion of virus-induced exacerbations of asthma has become clear. In studies of asthma in the community, 85 per cent of asthma attacks in children and 44 per cent in adults were induced by upper respiratory tract infections, of which the great majority were caused by rhinoviruses. Seasonal patterns of respiratory infections are strongly correlated with hospital admissions for asthma. In school children the major determinant of paediatric admission is school attendance, with the peaks of respiratory infections and asthma admissions both occurring at the start of the school term.

Exacerbations of asthma provoked by respiratory infections are often severe and can be prolonged and associated with increased airway responsiveness. In school children, peak flow measurements can remain abnormal for several weeks after a respiratory tract infection.

Drugs

Relatively few drugs exacerbate asthma. Of those which do, b-blockers and non-steroidal anti-inflammatory drugs (**NSAIDs**) are the most important.

b-Blockers

Precipitation or worsening of asthma was first reported with propanolol, but subsequently found to occur with all non-selective b-adrenoceptor antagonists, implying adrenergic bronchodilator tone in asthmatic airways. The severity of the airway narrowing provoked by b-blockers is not predictable, nor is it closely related to the severity of airway hyperresponsiveness. The dose provoking asthma can be low: severe asthma can be precipitated by timolol eye drops, a non-selective b-blocker used to treat glaucoma. Selective b-antagonists, such as atenolol, acebutalol, and metoprolol, provoke less severe reactions than non-selective b-blockers. However, although the fall in lung function provoked by a b-blocker may be reversed by an inhaled b_2-agonist, the severity of airway narrowing is unpredictable; alternative drugs are available for their indications (such as hypertension and angina), and patients with asthma should avoid b-blockers including b_1-selective antagonists. Although ACE inhibitors can cause cough and occasionally rhinitis, they have not been associated with the provocation of asthma and are not contraindicated in asthma.

Aspirin and NSAIDs

Aspirin and other NSAIDs which inhibit cyclo-oxygenase 1, can provoke severe attacks of asthma in some 10 per cent adults with asthma, more frequently in women than men. Aspirin-induced asthma may be part of a well-recognized association of aspirin intolerance, asthma, and rhinitis with nasal polyps (Samter's triad), which is characterized by severe mucosal eosinophilic inflammation of the nose and airways. The onset is usually in the third or fourth decade, with chronic nasal congestion, discharge, and nasal polyps, followed by development of asthma and aspirin-induced asthma. Ingestion of aspirin or a NSAID then characteristically provokes acute severe asthma within 1 h, accompanied by profuse nasal discharge, peri-orbital oedema, conjunctival infection, in some cases with flushing of the head and neck and, on occasions, with vomiting and diarrhoea. Aspirin-induced asthma can provoke life-threatening asthma resistant to bronchodilators: in one survey, 25 per cent of 145 patients requiring mechanical ventilation for acute severe asthma had aspirin-induced asthma although aspirin ingestion had not necessarily provoked this attack.

Despite avoidance of aspirin and NSAIDs, severe asthma and rhinitis with nasal polyps usually persist, associated with a raised blood eosinophil count and intense eosinophil infiltration of the nasal and airway mucosa. The most plausible explanation of aspirin-induced asthma is that it occurs as a consequence of specific inhibition in respiratory cells of intracellular cyclo-oxygenase enzymes. NSAIDs which inhibit cyclo-oxygenase activity provoke asthma in patients with aspirin-induced asthma; NSAIDs which do not inhibit cyclo-oxygenase activity do not provoke asthma. The potency of NSAIDs to inhibit cyclo-oxygenases correlates with their ability to provoke asthma in individuals with aspirin-induced asthma; and cross-tolerance to NSAIDs that inhibit cyclo-oxygenase occurs after desensitization to aspirin. Cross-tolerance involving such chemically distinct moieties argues strongly against aspirin-induced asthma being an immunological reaction.

The intense tissue eosinophilia is accompanied by overproduction of cysteinyl leukotrienes. In aspirin-induced asthma, cyclo-oxygenase inhibition is associated with release of cysteinyl leukotrienes that are important mediators of nasal inflammation and asthma. Cysteinyl leukotrienes, continuously synthesized in patients with aspirin-induced asthma, even in the absence of aspirin ingestion, are released into nasal and bronchial secretions and can be collected in urine. Aspirin-provoked nasal and asthmatic reactions are attenuated by leukotriene antagonists, both cysteinyl leukotriene receptor antagonists (zafirlukast, montelukast, and pranlukast) and 5-lipoxygenase inhibitors (zileuton).

Patients with aspirin-induced asthma should avoid all aspirin-containing products and other analgesics that inhibit cyclo-oxygenase (Table 2). Patients with aspirin-induced asthma can usually, although not always, take paracetamol. Selective inhibitors of cyclo-oxygenase 2 (celecoxib and rofecoxib) should theoretically be safe, but have not yet been investigated for cross-reactions with aspirin.

Tolerance to aspirin and NSAIDs can be induced in patients with aspirin-induced asthma by the ingestion of increasing doses of aspirin over 2 to 3 days, until 400 to 650 mg of aspirin can be tolerated. Daily doses of between 80 and 325 mg of aspirin can maintain tolerance, allowing aspirin and other cyclo-oxygenase inhibitors to be taken safely. A dose of aspirin of 650 mg twice daily can provide improvement in asthma and particularly nasal inflammation. One report has suggested that

regular aspirin treatment after sinus surgery for polypectomy may delay recurrence of nasal polyps, on average by 6 years. However, aspirin desensitization requires daily maintenance of high-dose aspirin, which may not be well tolerated. Furthermore, omission of aspirin for 2 to 3 days can result in complete loss of tolerance, in which case the initial desensitization protocol needs to be repeated. It is also not clear whether aspirin desensitization has the potential to modify the long-term course of asthma. For these reasons, aspirin desensitization has not been widely adopted.

Clinical features of asthma

The development of asthma

Knowledge of the way that asthma develops has been hindered by the lack of a clear workable definition that includes all cases (sensitive) and excludes non-cases (specific), and by the relative paucity of longitudinal data on well-defined community cohorts, which include a representative group of cases of asthma and are not limited to those coming to medical attention. None the less, there is now sufficient information to allow a reasonable view of the situation.

The relationship between wheezing in preschool children and asthma in school-age children has been clarified by a number of overlapping studies. Wheezing and cough in children aged less than 2 to 3 years is common and typically associated with viral respiratory infections. The important risk factors are reduced lung function at birth, prematurity or low birth weight, and maternal smoking during pregnancy. The prognosis for such children is good, with remission in the majority by school age and normal lung function in adult life. 'Wheezy bronchitis' in preschool years does not occur more frequently in school-age children with asthma, whose risk factors are different, suggesting the two disorders are independent. The peak prevalence of asthma occurs between the ages of 5 and 10 years, is associated with eczema in infancy, and evidence of sensitization to common inhalant allergens (identified either by skin test responses or by increased total IgE).

The outcome for children who develop asthma has been the subject of several general-practice and hospital-based reports, which of necessity will describe the prognosis of more severe cases. The outcome for cases identified in random population samples has been reported from Australia and the United Kingdom. The Australian study found that the risk of asthma persisting at ages 21 and 28 years was associated with the frequency of wheezing at ages 7 and 14 years. Children who wheezed infrequently in childhood and adolescence were least likely to have continuing asthma as young adults: more than half of those with asthma before the age of 7 years that had remitted by the age of 14 years remained symptom free aged 21 years. However, less than 20 per cent of those with persistent symptoms in childhood were symptom free in adolescence, and frequent attacks in this group continued to the age of 28 years. Some two-thirds of those without symptoms in adolescence remained free of asthma at the age of 28 years. The United Kingdom study described the incidence of wheezing from birth to age 33 years. The incidence of wheezy illness at all ages was related to a history of eczema and hay fever. One-quarter of children with a history of asthma or wheezy bronchitis by the age of 7 years continued to have symptoms when aged 33 years. Asthma developing in adult life was strongly associated with cigarette smoking and a history of hay fever.

In both the United Kingdom and Australian studies, asthma recurred in adult life after a period of remission in adolescence. More than one-half of those in the United Kingdom study who had wheezed before the age of 7 years and reported wheezing aged 33 years had been free of symptoms for 7 years between the age of 16 and 23 years. Similarly, in the Australian study, wheezing had recurred in 30 per cent of those who were free of wheezing aged 21 years. In both studies asthma recurred in some individuals with mild symptoms in childhood, which were frequently not recalled, and who would otherwise have been labelled as having 'adult-onset' asthma.

Symptoms and signs

The symptoms of asthma are non-specific: shortness of breath, wheezing, chest tightness, and cough. These are manifestations of airway narrowing, which is usually variable in severity over short periods of time, but can be persistent, and of airway hyperresponsiveness. Asthma as the cause of these symptoms is suggested by the variability in their severity and distinguished by their periodicity (such as daily, weekly, monthly, or seasonal), their provocation by specific (such as allergen) and non-specific stimuli, and their reversibility with bronchodilators or corticosteroids.

Patients with asthma can be categorized, at any one time, by whether their symptoms are intermittent or persistent, and by the severity of their symptoms and underlying airway narrowing (measured by lung function tests. Even those with mild asthma–intermittent or persistent–can develop severe asthma.

1. Mild intermittent asthma—symptoms occur less than weekly with normal or near normal lung function between episodes.
2. Mild persistent asthma—symptoms occur more than weekly but less than daily with normal or near normal lung function between episodes.
3. Moderate persistent asthma—symptoms occur daily with mild to moderate variable airflow limitation.
4. Severe persistent asthma—symptoms occur daily and interfere with normal activities. There is frequent nocturnal waking and moderate to severe variable airflow limitation.
5. Severe asthma—severe distressing symptoms prevent sleep. Severe airflow limitation responds poorly to inhaled bronchodilators and can be life-threatening.

Symptoms of asthma are typically worse at night, waking the affected individual in the early hours of the morning (on occasion several times), and on first waking in the morning, when chest tightness may be the dominant symptom. Asthmatic symptoms may also be provoked by non-specific stimuli such as exercise and cold air, and by specific allergens such as domestic animals, particularly cats.

Respiratory viral infections that occur predominantly in the winter months are the most important precipitating causes of exacerbations of asthma. In patients allergic to pollens or moulds, asthmatic symptoms occur or worsen during the relevant season (in the United Kingdom—late spring: tree pollen; May and June: grass pollen; late summer months: mould spores). In those with asthma induced by occupational sensitizers, symptoms characteristically increase in severity during the working week and improve when away from work on holidays of 1 week or more, if not at weekends (see Chapter 17.4.1.5). In some women asthma has a monthly periodicity, becoming increasingly severe during the days before menstruation, improving with its onset.

Although breathlessness and wheeze are often considered the most characteristic symptoms of asthma, cough can be the dominant and on occasions, particularly in children, the only symptom of asthma. Nocturnal cough suggests asthma, although in community studies isolated nocturnal cough has been found to be a poor predictor of the condition. 'Cough-variant asthma' is occasionally seen in adults who do not have airway narrowing and in whom cough and eosinophil-rich sputum are the only manifestations of the disease.

The characteristic symptoms of asthma are manifestations of variable airway narrowing and airway hyperresponsiveness. Patients with chronic severe asthma have more persistent airway narrowing and are limited in their day to day activities by breathlessness. They may have less symptomatic evidence of spontaneous variability of airway narrowing, although they can be awoken by asthma at night as well as having symptoms provoked by inhalation of cold air or by laughter.

Patients with acute severe asthma are usually distressed by severe shortness of breath with wheezing, unable to sleep and to complete sentences in one breath because of the severity of the airway narrowing.

The physical signs of mild or moderate asthma may be limited to expiratory wheezes audible over the lungs. Because of the variable nature of the airway narrowing, some patients have normal lung sounds, although expiratory wheezes would be anticipated in patients with persistent symptomatic asthma. Patients with chronic persistent asthma can develop hyperinflated lungs.

In acute severe asthma, patients are usually severely short of breath, sitting up or leaning forward using their accessory muscles of respiration. With increasingly severe airway narrowing, expiration becomes increasingly prolonged and alternates with short inspiratory gasps, impairing speech. Tachycardia and pulsus paradoxus often accompany acute severe asthma, but pulsus paradoxus is not a reliable indicator of severity. Airway narrowing may become sufficiently severe for no wheeze to be audible (the 'silent chest') and gas exchange sufficiently impaired to cause detectable cyanosis. Patients with asthma of this severity are usually distressed, hyperventilating, anxious, apprehensive, and can be confused because of hypoxia. Exhaustion ultimately leads to inadequate ventilation and a rising PCO_2, the two cardinal features that indicate the need for transfer to an intensive care unit in the event that assisted ventilation is required.

Diagnosis of asthma

Although asthma is now defined by characteristic pathological changes in the airways, it is usually identified by its pathophysiological manifestations, variable or reversible airway narrowing and airway hyperresponsiveness. In some patients the presence of eosinophils in sputum or a raised eosinophil count in the blood can be

a valuable diagnostic pointer.

Asthma is usually diagnosed by the demonstration of airflow limitation that varies spontaneously over short periods of time, or which reverses after inhalation of short-acting b-agonists or, over a more prolonged period, in response to inhaled or oral corticosteroid. In a small minority of patients, provocation tests using exercise, or pharmacological agents such as histamine or methacholine can be valuable. In suspected cases of occupational asthma, inhalation tests with the specific agent may be indicated (see Chapter 17.4.1.4); inhalation tests with common inhalant allergens are rarely indicated in clinical practice.

Airflow limitation

The most clinically useful measurements of airflow limitation are: (i) forced expiratory volume in 1 s (FEV_1), which may be expressed as a proportion of the forced vital capacity (**FVC**) as FEV_1/FVC per cent, and (ii) peak expiratory flow rate (**PEFR**). Both tests require the patient to provide a reproducible maximal forced expiratory manoeuvre using tested and validated equipment. FEV_1 has the advantage of a visible tracing of the expelled volume of air over time, which allows the observer to determine whether reproducible maximal forced expiratory manoeuvres have been made. PEFR testing does not provide this opportunity. However, peak flow meters employed to measure PEFR, unlike spirometers required to measure FEV_1, can be used regularly by patients to monitor their lung function, indicating the need for altered treatment at an early stage. Whether abnormality of FEV_1 and PEFR should be expressed in absolute or proportional terms remains undecided. Expression as an absolute difference from the average value anticipated for an individual of given age, gender, and height has more physiological validity, but the majority of lung function laboratories in the United Kingdom continue to define as abnormal, values of FEV_1 or PEFR of 20 per cent or more below the mean predicted value.

The difference of the measured from the predicted mean value for an individual can be conveniently expressed as a Z value, which is the number of standard deviations the measured value is from the predicted mean:

$$Z = \frac{\text{predicted mean value} - \text{measured value}}{\text{standard deviation (c.0.5)}}$$

The distribution of FEV_1 around the mean predicted value is similar at all ages (from 25 to 70 years) with a standard deviation of some 0.5 litres. The disadvantage of using a proportional rather than an absolute difference can best be appreciated by comparing a patient whose predicted FEV_1 is 5 litres with one whose predicted FEV_1 is 2.5 litres. In the first patient a 20 per cent reduction equates to a loss of 1 litre (2 standard deviations); in the second a 20 per cent reduction equates to a loss of 0.5 litres (1 standard deviation).

Variability and reversibility of airflow limitation

Serial measurements of PEFR in most patients with asthma show spontaneous variability. The most characteristic pattern is of a circadian variation, with airflow limitation most severe on waking in the morning (and during the night if awoken) with improvement occurring during the morning after waking (Fig. 4). A small circadian variation in PEFR or FEV_1 is seen in normal individuals; in asthma a difference of 20 per cent or more between the highest and lowest values may be found. Other patterns of variation in severity of airflow limitation may be imposed on this circadian rhythm, such as falls in PEFR provoked by exercise or exposure to an allergen or occupational sensitizer, which resolve after avoidance of the stimulus. While variations of 20 per cent or more in FEV_1 or PEFR are commonly regarded as indicating asthma, in patients with severe airflow limitation, with an FEV_1 of 1 litre, 20 per cent variability equates to 200 ml, a level of spontaneous variation observed in people without asthma.

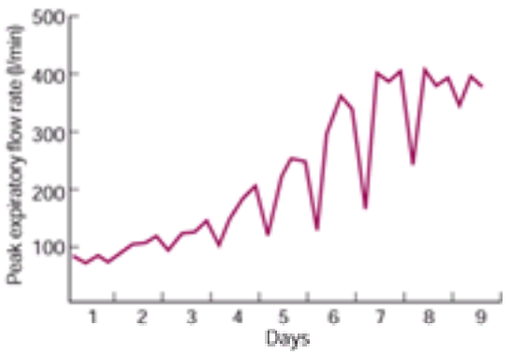

Fig. 4 Circadian rhythm in peak expiratory flow rate in a patient with asthma recovering from an acute attack.

The most commonly used means to identify asthma in clinical practice is an improvement after 15 to 20 min in airflow limitation, identified by FEV_1 or PEFR, after inhalation of a bronchodilator, usually a short-acting b-agonist such as salbutamol in a dose of 200 µg—evidence of asthma is usually regarded as an improvement in FEV_1 or PEFR of 20 per cent or greater. However, it is important to appreciate that the absence of a significant improvement in lung function after inhalation of a bronchodilator does not exclude a diagnosis of asthma (that is, it is a more specific test than it is sensitive). Rapid reversibility of airflow limitation is more readily seen in young adults with mild or moderate asthma than in older patients with more severe airflow limitation. Reversibility cannot be tested in a patient whose lung function is normal at the time of testing.

An increase of 20 per cent or more from baseline FEV_1 or PEFR is generally accepted as diagnostic of asthma. Expressing change as a proportion of baseline will exaggerate the degree of improvement in those with a low initial FEV_1 or PEFR. A 20 per cent increase in FEV_1 in a patient with a baseline FEV_1 of 4 litres is 800 ml, but only 200 ml in a patient whose baseline FEV_1 is 1 litre. Studies of short-term (20 min) variability in FEV_1 in patients with airflow limitation have found that the increase in FEV_1 needed to exclude natural variability with 95 per cent confidence was 160 ml. This value did not differ significantly from the value in normal individuals, in whom an absolute increase in FEV_1 of 190 ml was needed to exclude a chance increase with 95 per cent confidence. In both normal individuals and those with an airflow limitation, expression of variability as an absolute difference was similar at all levels of FEV_1, whereas when expressed as a percentage change, the degree of variability decreased with increasing FEV_1. Selecting a specific percentage change in FEV_1 (or PEFR) to define asthma will necessarily include a greater proportion of patients with lower prebronchodilator FEV_1. Equally, patients with a higher baseline FEV_1 need to achieve a greater absolute increase to fulfil the defined criterion. Expression of variability as an absolute change has more biological and statistical validity, and an increase of more than 200 ml in FEV_1 has a probability of less than 5 per cent of occurring by chance. As with expression of lung function, it is unlikely that the use of results based on absolute values, although biologically more valid, will be adopted. It should be appreciated, however, that in patients with a low FEV_1 a 20 per cent increase in FEV_1 may have occurred by chance, and in those with a high FEV_1 an increase of more than 200 ml is unlikely to have occurred by chance.

In some patients with asthma, particularly those with severe airflow limitation, inhalation of a short-acting bronchodilator does not provide significant improvement in FEV_1 or PEFR. In these circumstances the diagnosis of asthma and differentiation from less reversible causes of airflow limitation, such as chronic bronchitis and emphysema, can be made with a 'trial' of treatment with corticosteroids. Significant improvement in airflow limitation both implies a diagnosis of asthma and demonstrates that corticosteroids (inhaled or oral) are effective treatment. However, corticosteroids can also improve exercise tolerance by enhancing mood and outlook, hence the benefit of a trial of steroids has to be judged by its effect on lung function. Although there is no formally agreed protocol for a steroid trial, a generally acceptable trial would be oral prednisolone taken in a dose of 0.6 mg/kg, (40 mg/day in a 70 kg male) for 3 weeks, with measurement of spirometry made on at least two separate occasions, once before and once at the end of the trial supplemented by three times daily home peak flow measurements. Symptomatic improvement with an increase in FEV_1 or PEFR of 20 per cent or more during the trial is generally considered as evidence of asthma and an indication for treatment with corticosteroids, inhaled or oral.

Tests of airway hyperresponsiveness

Airway hyperresponsiveness—an exaggerated response to non-specific provocative stimuli—is a cardinal feature of asthma. Tests of airway responsiveness to exercise and to inhaled histamine or methacholine, which can provoke acute airway narrowing in a dose-dependent fashion, can be of value in the diagnosis of asthma, particularly in patients with symptoms suggestive of the condition but in whom measured lung function is normal or, if abnormal, shows no reversibility with inhaled bronchodilators. These tests are required in only a few patients and each has its limitations: exercise testing can be insensitive (false negatives) and tests of airway reactivity to inhaled histamine or methacholine non-specific (false positives), although the provocation of a 20 per cent fall in FEV_1 by histamine at 4 mg/ml or less (or equivalent) occurs uncommonly in those without asthma.

Exercise testing

Acute airway narrowing provoked by exercise is a common feature of asthma, particularly in children. Testing for exercise-provoked asthma requires continuous exertion for 5 min. This is most conveniently undertaken in a lung function laboratory by running on a treadmill or exercising on a cycle ergometer, although free running is more likely to provoke an asthmatic reaction. Measurements are made of FEV_1 or PEFR 5 min before, during, and at 5 min intervals for 30 min after the test. A normal individual will have a less than 5 per cent increase in FEV_1 or PEFR during and a less than 10 per cent fall after exercise. Depending on the level of baseline, patients with asthma can have a greater than 5 per cent increase during exercise and greater than 10 per cent fall from pretest value afterwards. Exercise is a valid and reproducible test for asthma, although it can have false negatives, particularly when undertaken by methods other than free running.

Airway reactivity to inhaled histamine or methacholine

Acute airway narrowing can be provoked in a dose-dependent manner by the inhalation of increasing doses of a bronchoconstrictor, of which histamine or methacholine are the most commonly used. The test as described by Cockcroft and colleagues consists of tidal breathing of doubling doses of histamine, with measurement of FEV_1 6 min after each inhaled dose. The percentage change in FEV_1 from a post-saline baseline after each concentration of inhaled (histamine) can be plotted, the test being terminated when either a 20 per cent or greater fall in FEV_1 is provoked or the maximum concentration (usually 16 or 32 mg/ml) is reached. The level of airway reactivity is usually expressed as the concentration of histamine that provokes a 20 per cent fall in FEV_1 (PC_{20} histamine), which can be identified by linear interpolation. The lower the PC_{20}, the more reactive the airways. The test is usually repeatable within one doubling dose, but may not be consistent in any individual, PC_{20} falling, for instance, after exposure to allergen or occupational sensitizer.

In population studies the major determinants of airway reactivity have been atopy (in children and young adults) and smoking in older adults (probably reflecting reduced FEV_1). Airway responsiveness can be increased in atopic children with rhinitis and in healthy adults after a viral respiratory tract infection. Hence, evidence of measurable airway reactivity is not necessarily evidence of asthma, but it is uncommon for individuals without asthma to have a PC_{20} histamine of methacholine of less than 8 mg/ml. Measurement of airway reactivity to histamine or methacholine is more sensitive than exercise testing, although a less specific test for asthma. Like exercise testing its value in clinical practice is primarily in symptomatic patients with normal or near normal FEV_1 and without evidence of spontaneous variability or reversibility. A negative test in a symptomatic patient suggests that current asthma is unlikely to be the cause of their symptoms.

Imaging in asthma

Imaging of the chest is not commonly of diagnostic value in asthma, but can be important in identifying its complications. In patients in whom asthma develops over the age of 30 years, the chest radiograph is usually normal. However, about one-quarter of children and one-fifth of adults show changes of hyperinflation on the chest radiograph. These changes include a low diaphragm (below the sixth intercostal space anteriorly) and an increased retrosternal space. In some children with chronic persistent asthma, the length of the lung becomes greater than the width of the thorax, with the posterior ends of the ribs becoming more horizontal.

The most commonly observed radiographic sign in asthma is of thickened bronchial walls due to eosinophilic infiltration of the airways: these are visible on the chest radiograph as parallel lines ('tram lines') or as a thick-walled ring shadow when seen end on.

Complications of asthma that can be seen on the chest radiograph include pneumothorax, pneumomediastinum, pulmonary collapse, and eosinophilic pneumonia. The physical signs of pneumothorax can be difficult to detect in an asthmatic attack, but its detection can be lifesaving. Pneumomediastinum is of less clinical importance. Plugging of the airways by a mucus plug characteristically occurs in allergic bronchopulmonary aspergillosis, but also in asthmatic patients without this condition: in both it can cause atelectasis, which is usually lobar or segmental. Eosinophilic pneumonia is characterized by consolidation on the chest radiograph accompanied by a raised blood eosinophil count. This can be a manifestation of several conditions that include allergic bronchopulmonary aspergillosis, helminth infections, and drug reactions, as well as being of unknown cause—acute and chronic eosinophilic pneumonia. Of these, allergic bronchopulmonary aspergillosis and chronic eosinophilic pneumonia (which can be a manifestation of Churg–Strauss syndrome—allergic granulomatosis) are the most common causes of eosinophilic pneumonia in patients with asthma. (See Chapter 17.11.5, Chapter 17.11.9 and Chapter 17.11.11 for further discussion of these conditions.)

Differential diagnosis of asthma

Asthma needs to be differentiated from:

1. localized airways obstruction;
2. other causes of generalized airways obstruction; and
3. other causes of intermittent breathlessness.

Localized airways obstruction

Upper airways obstruction, of the larynx or trachea, causes a monophonic inspiratory wheeze (stridor) audible over the trachea with a characteristic abnormality of the flow volume loop—decreased inspiratory flow rates. In a child, wheezing can be caused by an inhaled foreign body (classically a peanut), which should be particularly suspected when the problem develops suddenly in a previously healthy individual. The chest radiograph may show the foreign body if it is opaque, or distal atelectasis, consolidation, or air trapping on an expiratory film (which may not be possible to obtain in small children). However, it can be normal and, if foreign body inhalation is suspected, bronchoscopy should be undertaken to identify and remove it or to exclude the possibility. In adults, localized airway narrowing is more likely to be due to a tumour—benign or malignant—which may occasionally cause a unilateral monophonic wheeze. The tumour may be visible on the chest radiograph, but definitive diagnosis will require bronchoscopy and biopsy.

Generalized airways obstruction

The major causes of generalized airways obstruction from which asthma needs to be distinguished are chronic bronchitis and emphysema, although in some cases these may coexist with asthma. Other causes such as obliterative bronchiolitis are less common. In general, chronic bronchitis and emphysema cause breathlessness on exertion that increases slowly in severity over years and only uncommonly causes breathlessness before the age of 40 years. Nocturnal waking by respiratory symptoms is uncommon in chronic bronchitis and emphysema, although not universal in asthma. Chronic severe asthma responsive to corticosteroids, but without significant reversibility to inhaled bronchodilators, may have similar radiographic and spirometric abnormalities. In both, the lungs may be hyperinflated on the chest radiograph, but in asthma, unlike emphysema, there is no associated loss of vascular markings. Lung function tests in both asthma and emphysema can show airflow limitation with reduced FEV_1, reduced FEV_1/FVC ratio, and hyperinflated lungs with increased total lung capacity. However, while factor transfer ($TLCO$) and gas transfer coefficient (KCO) are reduced in emphysema, in asthma KCO is normal or increased. Differentiation from chronic bronchitis can be difficult because, like asthma, there is no loss of vascular markings on the chest radiograph or reduction of KCO. If present, sputum (and blood) eosinophilia suggests asthma, but differentiation in these circumstances often depends on the outcome of a trial of steroids.

In young children asthma needs to be differentiated from wheezing episodes associated with viral respiratory tract infections, and in children and adolescents from cystic fibrosis. Cystic fibrosis is suggested by disproportionate production of (usually discoloured) sputum, weight loss, and an abnormal chest radiograph. The

presence of staphylococci in sputum and development of nasal polyps in childhood is very suggestive of cystic fibrosis.

Other causes of intermittent breathlessness

The most important causes of intermittent breathlessness from which asthma should be differentiated are left ventricular failure, pulmonary emboli, extrinsic allergic alveolitis, hyperventilation, and vocal cord dysfunction.

Left heart failure sufficient to cause breathlessness will usually be apparent on clinical examination, chest radiograph, ECG, and echocardiogram. The heart is clinically and radiographically enlarged, with the exception of pulmonary venous hypertension caused by mitral stenosis. Inspiratory crackles are usually audible at the lung bases and the jugular venous pressure may be elevated. In addition to an enlarged heart the chest radiograph may show upper lobe venous distension, Kerley 'B' lines, and pleural effusion. Echocardiography will usually show evidence of left ventricular disease, or in the case of mitral stenosis, left atrial enlargement. Identification of the cause of breathlessness can be difficult when left heart failure is provoked by an intermittent arrhythmia.

Pulmonary embolism causes breathlessness that can occasionally be associated with wheezing; the diagnosis is suggested by associated pleuritic pain and haemoptysis. The chest radiograph and CT scan may show pleural-based 'humpback' opacities and pleural effusion. The diagnosis is most securely made angiographically, but more usually from a ventilation perfusion scan or spiral CT scan. A normal ventilation perfusion scan makes all but the smallest emboli unlikely, although interpretation can be difficult in patients with widespread ventilatory disease. A normal spiral CT scan excludes pulmonary emboli to subsegmental level.

Extrinsic allergic alveolitis

Extrinsic allergic alveolitis can provoke recurrent episodes of breathlessness which characteristically develop 4 to 8 h after exposure to the cause (usually mouldy hay or birds—pigeons or budgerigars). Breathlessness in extrinsic allergic alveolitis is usually not accompanied by wheeze but with fever, flu-like symptoms, and a neutrophil leucocytosis. The chest radiograph often shows widespread nodular or groundglass shadowing and the CT scan discrete areas of groundglass opacification. Lung function tests show a proportionate reduction in FEV_1 and FVC, which may be accompanied by a reduced *TL*CO and *K*CO. (See Chapter 17.11.11 for further discussion.)

Hyperventilation

Episodes of hyperventilation may be difficult to distinguish symptomatically from asthma, and can in some cases complicate asthma. The diagnosis should be suspected in a patient who complains of breathlessness that occurs without identifiable cause (for example while sitting reading), may be associated with pins and needles in the fingers and dizziness (attributable to hypocapnia), and does not disturb sleep. The symptoms complained of can often be reproduced by a short period of voluntary overbreathing; 20 deep breaths is usually sufficient. The reason why some patients present with hyperventilation is unknown: Howell has suggested that it develops in obsessional (perfectionist) personalities and is usually precipitated by one of three events: bereavement, resentment, or concern about illness. However, it is important to recognize that asthma is characteristically a variable condition and a diagnosis of hyperventilation should not be made on the basis of absent physical signs or normal lung function at the time of consultation, but based on the characteristics described above.

Vocal cord dysfunction

Vocal cord dysfunction is easily misdiagnosed as asthma and may coexist with asthma. Wheezing, in vocal cord dysfunction, is caused by paradoxical adduction of the anterior two-thirds of the vocal cords on inspiration, and does not occur during sleep. The diagnosis is best made by direct examination of the cords during an attack. Management can be difficult, but recognition of the condition can allow high-dose oral corticosteroid treatment for 'uncontrolled asthma' to be avoided.

Hyperventilation and vocal cord dysfunction can each occur in patients with underlying asthma, frequently in association with underlying psychosocial problems. A critical point can be to determine the relevant life events associated in time with the onset of deterioration in the patient, who has often had previously well-controlled asthma. Vocal cord dysfunction is more common in women and in those engaged in health care provision.

Management of asthma

Objectives

The objectives of treating patients with intermittent or persistent asthma are:

1. to prevent troublesome symptoms (such as cough, shortness of breath) at night or with exercise;
2. to enable patients to achieve normal levels of activity;
3. to maintain normal or near normal lung function; and
4. to prevent recurrent episodes of acute severe asthma and minimize the need for emergency hospital treatment.

These objectives are most likely to be achieved by treatment that reduces airway inflammation, either by avoidance of its inducing cause or by drugs with anti-inflammatory activity. The risk of side-effects of asthma treatment should be appreciated and minimized, patients' concerns about the potential side-effects of long-term treatment should be recognized, and relevant information provided to them.

Treatment selection

Randomized controlled trials of asthma treatments published in the past decade have shown the magnitude of benefit of different treatment interventions in patients with asthma of varying severity. This information has provided a secure basis for deciding which treatment is likely to be most effective in individual patients. Of particular importance has been the broadening of the indications for the use of inhaled corticosteroids. Inferences from these studies on the optimal treatment for asthma of differing degrees of severity has informed the published guidelines for asthma management in the United Kingdom, United States, and elsewhere.

The targets for effective asthma treatment in individual patients are:

1. normal daytime activities, such as going to work and to school, as well as the ability to enjoy physically demanding activities (such as sport);
2. sleeping through the night without awakening by respiratory symptoms;
3. use of 'rescue' medication with inhaled b_2-agonists needed less than once per day;
4. normal or near normal PEFR and FEV_1, with less than 20 per cent variability between best and worst values; and
5. avoidance of drug side-effects.

Targets 1, 2, and 5 are of the most interest to the patients, who will primarily be seeking improvement in their quality of life from treatment. Zealous pursuit of restoration of normal lung function in a patient whose quality of life has already been restored by treatment is of questionable value.

Asthma, except where caused by a dominant and avoidable agent (such as a domestic pet or occupational sensitizers), is not usually curable, but current treatment offers the great majority of patients the opportunity to enjoy a normal life. Most asthma is mild: in one community survey only 15 per cent of patients had persistent asthma of moderate severity (step 3 of British Thoracic Society (**BTS**) guidelines or worse—see below). However, some 5 per cent of patients have severe asthma that responds poorly to conventional treatment. These patients suffer most, both from their disease and from the side-effects of its treatment, and are at highest risk of admission to hospital and death from asthma.

Treatments for asthma

Allergen avoidance

The identification and, where feasible, the avoidance of relevant allergens at home or at work should be considered an essential part of the management of asthma. It enables patients to recognize important causes of their asthma and to take responsibility for their avoidance. Allergen avoidance should be regarded as complimentary to drug treatment of asthma, with the advantage in some cases (where a single allergen is the dominant cause) of providing a cure with avoidance of the possible side-effects of drugs. Complete avoidance of exposure to house dust mite, domestic pets, and occupational causes of asthma has been associated with marked improvement in respiratory symptoms, lung function, and airway hyperresponsiveness. Avoidance of exposure is most clearly indicated and usually most feasible when the cause of asthma is an agent inhaled at work (see Chapter 17.11.1.5). Removal of a pet, particularly a cat from the home, is most effective when accompanied by thorough cleaning and washing of the house to remove residual allergen, which can otherwise persist in concentrations sufficient to provoke asthma for many months. Avoidance of exposure to the house dust mite, *D. pternonyssinus*, by spending several months in the Alps or in a hospital, has been shown to provide symptomatic and functional improvement. However, house dust mites are ubiquitous in many environments, including much of the United States, United Kingdom, and Europe, and it can be difficult to eliminate mites sufficiently from the home so that exposure to the relevant allergens (such as Der p1) is reduced to concentrations which do not continue to induce airway inflammation. The issue with house dust mite avoidance is the feasibility of securing an effective intervention.

Drug treatment

The drugs primarily used to treat asthma are the progeny of two hormones secreted by the adrenal glands: cortisol and adrenaline. Pharmaceutical research in the past 50 years has led to the development of selective b $_2$-agonists, both short and long acting, and lipid-soluble topically active inhaled corticosteroids. b $_2$-Agonists and corticosteroids account for nearly 90 per cent of prescriptions for asthma in the United Kingdom. Other drugs used include sodium cromoglycate and nedocromil sodium amongst the prophylactic agents, and ipratropium bromide and theophyllines amongst the bronchodilators. The core treatment for mild and moderately severe persistent asthma is inhaled corticosteroids and inhaled b $_2$-agonists. Other agents are used as additional treatments when these alone are not sufficient to provide control. Leukotriene receptor antagonists and 5-lipoxygenase inhibitors have been introduced recently; their place in the treatment of asthma is currently being assessed.

Corticosteroids

Corticosteroids are the most effective treatment for asthma. Systemic corticosteroids were introduced for this purpose in the 1950s, but their use was limited by serious unwanted side-effects, which stimulated research into the development of equally effective but safer alternatives. The introduction of topically active corticosteroids that can be administered by inhalation and are free of the systemic side-effects of oral corticosteroids at therapeutically effective doses has revolutionized the treatment of asthma.

Corticosteroids suppress airway inflammation, with improvement in airway hyperresponsiveness, lung function, and associated respiratory symptoms. Although the mechanism of action of steroids in asthma continues to be debated, they inhibit the formation of cytokines relevant to asthmatic inflammation, such as interleukin 4 (**IL-4**), IL-5, IL-13, and granulocyte–macrophage colony-stimulating factor, by lymphocytes and macrophages, by inhibition of transcription of cytokine genes.

Corticosteroids suppress the chronic inflammation in the asthmatic airways, but do not cure the disease. To be effective they must therefore be taken continuously—oral steroids usually daily and inhaled steroids usually twice daily.

Oral corticosteroids

Oral corticosteroids—prednisolone and prednisone—are rapidly absorbed from the gut, achieving peak plasma levels at 1 to 2 h. Prednisone is biologically inactive but rapidly and completely converted in the liver to the active form, prednisolone. Some 20 per cent of prednisolone is inactivated in the liver by first-pass metabolism leaving 80 per cent of the oral dose bioavailable. The plasma half-life of prednisolone is usually 2 to 3 h. Corticosteroids are inactivated in the liver by conjugation. Hepatic enzyme inducers such as rifampicin, barbiturates, and phenytoin can reduce the half-life of prednisolone by 50 per cent. To counter the consequent reduction in anti-inflammatory activity, the dose of oral prednisolone should be doubled in patients concurrently receiving these treatments.

Oral corticosteroids effect detectable improvement in airflow limitation in patients with asthma within 6 to 12 h of administration. In cases of severe asthma, maximum improvement can take several days, probably reflecting the time to reverse the inflammatory changes in the airways.

The early use of oral corticosteroids in the treatment of asthma was severely limited by the high risk of unwanted effects, which include osteoporosis, hypertension, diabetes mellitus, cataract formation, and (in children) growth suppression. The introduction in the 1970s of inhaled corticosteroids revolutionized the treatment of asthma by providing local anti-inflammatory activity in doses that did not cause these limiting systemic side-effects.

Inhaled corticosteroids

Inhaled corticosteroids are highly lipophilic and rapidly enter cells within the airways. They combine high topical potency with low systemic bioavailablilty of the swallowed dose and rapid metabolic clearance of any corticosteroid reaching the systemic circulation, conferring a high benefit:risk ratio. Although 80 to 90 per cent of a dose from a metered dose inhaler is deposited in the oropharynx, swallowed, and absorbed, more than 80 per cent of beclomethasone, 90 per cent of budesonide, and 99 per cent of fluticasone is inactivated by first-pass metabolism in the liver. The 10 to 20 per cent of the inhaled dose deposited in the airways is available for absorption into the systemic circulation. For fluticasone and budesonide, devices that increase lung deposition (such as large volume spacer and Turbohaler) therefore increase the dose available for systemic absorption.

At present, three inhaled corticosteroids are generally available: beclamethasone diproprionate, budesonide, and fluticasone proprionate. Beclomethasone and budesonide are equipotent; fluticasone is twice as potent, requiring half the dose to achieve the same benefit. In general, most of the benefit of inhaled corticosteroids in the treatment of asthma is effected at low doses, where there is little evidence of adverse effects. By contrast, adverse effects develop at higher inhaled doses, where there is little evidence of greater benefit.

The clinical effects and side-effects of inhaled corticosteroids have been the subject of considerable clinical investigation. A systematic review of five randomized controlled trials comprising 141 adults with mild, persistent asthma that compared inhaled steroids with b $_2$-agonists found significant improvement in lung function in those receiving inhaled steroids. A randomized controlled trial that compared inhaled budesonide at 1200 µg/day with regular inhaled terbutaline in adults found improvement in the budesonide group during the 2 years of the study in respiratory symptoms, morning peak flow rates, airway hyperresponsiveness to inhaled histamine, and requirement for symptomatic b$_2$-agonists. In a subsequent study of the same population, no deterioration in symptoms or lung function, including airway responsiveness to inhaled histamine, occurred in three-quarters of those randomly assigned to reduce their dose of budesonide from 1200 to 400 µg/day, whereas deterioration in symptoms and lung function occurred in two-thirds of those assigned to discontinue budesonide and take placebo. Symptoms and lung function improved in those patients who had received terbutaline during the initial study and were subsequently assigned to budesonide at 1200 µg/day. However, the extent of improvement was less than in those originally assigned budesonide 2 years earlier, suggesting possible benefit from the earlier institution of inhaled corticosteroid treatment.

Several recent studies have compared the benefit of increasing the dose of inhaled steroid with the addition of a long-acting b $_2$-agonist (salmeterol or formoterol) in patients whose asthma was not controlled by conventional doses of inhaled steroids. Two randomized controlled trials found that the addition of twice daily inhaled salmeterol provided more benefit in terms of symptom-free days and nights, improvement in FEV_1 and PEFR, and need for rescue b$_2$-agonists than doubling the dose of inhaled steroids (Fig. 5). A third randomized controlled trial of 852 patients with moderately severe asthma investigated (i) the addition of formoterol 12 µg twice daily to budesonide 100 µg, (ii) increasing the dose of budesonide to 400 µg twice daily, and (iii) the addition of formoterol 12 µg twice daily to budesonide 400 µg twice daily. Symptoms of asthma and lung function were improved with both higher dose budesonide and with formoterol, but greater improvements were obtained with the additions of formoterol.

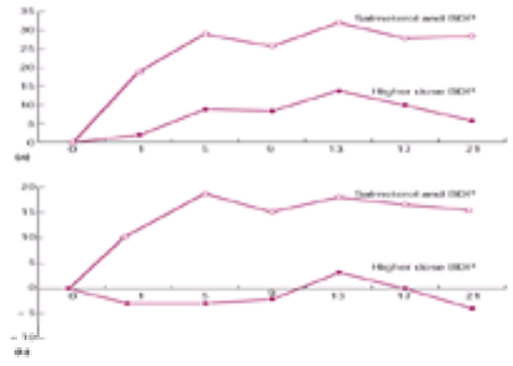

Fig. 5 Comparison of addition of salmeterol compared with doubling dose of inhaled steroid (BDP).

Inhaled corticosteroids are as effective in the treatment of asthma in children as in adults. Comparison in a randomized controlled trial of inhaled budesonide (600 µg/day) with an inhaled placebo for an average of 22 months in 116 children aged between 7 and 16 years also taking inhaled salbutamol at 200 µg/day showed marked benefit in the group taking inhaled budesonide: 26 children (45 per cent) in the placebo group withdrew from the study compared with only 3 (5 per cent) in the budesonide group. Symptoms and lung function (FEV_1, PEFR, and airway response to inhaled histamine) improved significantly in the budesonide group. Cessation of inhaled budesonide treatment in a randomized controlled trial in children aged 11 to 18 years previously treated for 28 to 36 months was associated with: recurrence of symptoms, in some of sufficient severity to need treatment with prednisolone; increased use of inhaled b_2-agonists; and deterioration in lung function, both FEV_1 and airway response to inhaled histamine. Such deterioration did not occur in the group continuing to take budesonide at 600 µg/day.

Side-effects of inhaled corticosteroids

Local side-effects

The severe adverse effects of systemic steroids and the widening indications for the use of inhaled corticosteroids has led to close scrutiny of the side-effects of inhaled corticosteroids. Local side-effects are well recognized and are dose dependent. These are oropharyngeal candidiasis (thrush) and dysphonia. Oropharyngeal candidiasis occurs in about 5 per cent of patients and can be a problem, particularly in the elderly. The risk of its development can be reduced by the use of a large volume spacer and rinsing the mouth out after each inhaled dose. Dysphonia occurs in at least one-third of patients and can cause particular problems for public speakers and professional singers. It is believed to be due to a myopathy of the laryngeal muscles and reverses when treatment is stopped. Inhaled corticosteroids do not cause atrophy of the airway epithelium after 10 years of treatment and are not associated with an increased risk of pulmonary infection, including tuberculosis.

Systemic side-effects

Concern about systemic side-effects of inhaled steroids stems from the need for their regular use for prolonged periods (years or decades) both in adults and children. Because many patients who take inhaled corticosteroids have also required oral corticosteroids, distinguishing the adverse systemic effects of inhaled corticosteroids can be difficult. In general, current evidence indicates that inhaled corticosteroids do not cause important side-effects in doses of up to 400 µg/day in children and 800 µg/day in adults. There is some evidence of side-effects at higher doses, more with beclomethasone than with budesonide or fluticasone. However, systemic absorption and the risk of systemic side-effects can be reduced by the use of a spacer with metered dose inhalers and by rinsing the mouth after inhalation of a dry powder inhaler, which should be recommended when doses of 400 µg/day or more in children and 800 µg/day or more in adults are prescribed.

Two important risks of inhaled corticosteroids that have been the subject of recent concern are osteoporosis in adults and growth suppression in children. A cross-sectional study of 81 patients with asthma aged between 20 and 40 years compared bone mineral density in 47 (19 men) who had taken inhaled steroids in doses of between 100 and 3000 µg/day (mean dose 620 µg/day) for at least 5 years with 34 (19 men) who had never taken inhaled or oral corticosteroids. Bone mineral density was not different in the two groups, but cumulative inhaled steroid dose was associated with a reduction of bone mineral density in the lumbar spine of 0.1 standard deviations per 1000 µg inhaled corticosteroid per year. In a subsequent study of adults aged between 20 and 40 years, who had taken inhaled steroids on average for 6 years, bone mineral density of the lumbar spine and femoral neck was inversely correlated with cumulative inhaled corticosteroid dose: doubling of inhaled corticosteroid dose was associated with an estimated decrease in bone mineral density of 0.16 standard deviations. Although unlikely to be associated with fracture in this age group, these results imply that bone mineral density would become significantly decreased in adults taking inhaled corticosteroids by their fifth and sixth decades, potentially increasing the risk of osteoporotic fracture substantially in later life.

Inhaled corticosteroids also increase the risk of posterior subcapsular cataract, glaucoma, and easy skin bruising.

Both asthma and oral corticosteroids can impair growth in children, but longitudinal studies have found no evidence of growth impairment in children taking inhaled corticosteroids in doses of up to 800 µg/day. A meta-analysis of 21 studies, including more than 800 children, found no effect of inhaled beclamethasone on height. None the less, the potential of inhaled corticosteroids taken in high doses for prolonged periods to impair growth remains, and regular monitoring of the growth of children taking inhaled steroids should be undertaken.

Beclomethasone and budesonide in doses of more than 160 µg/day taken by a metered dose inhaler cause a dose-dependent reduction in morning plasma cortisol and 24-h urinary cortisol excretion. However, 2000 µg/day of these inhaled corticosteroids when taken via a spacer has no effect on 24-h urinary cortisol. In children, beclomethasone in daily doses of 800 µg had no effect on 24-h urinary cortisol excretion.

The evidence of side-effects caused by inhaled corticosteroids, particularly osteoporosis, is now sufficient to mean that the lowest dose of inhaled corticosteroid which is clinically effective should be prescribed in both children and adults, and particularly in patients taking topical corticosteroids by other routes (such as nose or skin), and the dose tapered to the minimum necessary when symptomatic and functional improvement is achieved.

b_2-Adrenoreceptor agonists

b-Agonists are sympathomimetic amines that include catecholamines—both naturally occurring adrenaline, noradrenaline, and dopamine and synthetic isoprenaline—and non-catecholamines, both short-acting, such as salbutamol and terbutaline, and long-acting, such as salmeterol and formoterol. Catecholamines have been replaced in the treatment of asthma by b_2-selective non-catecholamines, which have a longer half-life than catecholamines because they are not subject to catecholamine uptake mechanisms and not broken down by catechol-*O*-methyl transferase. The duration of bronchodilatation after inhalation of non-catecholamines is longer; the effects of salbutamol and terbutaline persisting for 3 to 6 h and of salmeterol and formoterol for up to 12 h.

The actions of b-agonists in asthma are the result of stimulation of b-adrenoreceptors that are located in the airways, on airway epithelium, submucosal glands, airway and vascular smooth muscle. b-Receptors in the airways are entirely $b_{2,}$ with the exception of some b_1 -receptors on submucosal glands.

b_2-Agonists can influence airways function through several mechanisms:

1. relaxation of bronchial smooth muscle by direct effect on b_2-receptors;
2. inhibition of mast cell mediator release; and
3. enhanced mucociliary clearance.

Inhalation of a b_2-agonist by a patient with asthma increases airway calibre and reduces airway hyperresponsiveness. b_2-Agonists also cause tachycardia and increased cardiac output, systemic vasodilatation, and increased muscle blood flow. The tachycardia and increased cardiac output are the results of both stimulation of cardiac b-adrenoreceptors and a reflex response to peripheral vasodilatation. In addition b_2-agonists cause tremor and have metabolic actions, of which

hypokalaemia is probably the only potentially important clinical effect.

Inhaled, selective, short-acting b $_2$-agonists reverse mild acute airway narrowing and are sufficient treatment, alone, for mild intermittent asthma causing occasional symptoms. (Step 1 of BTS guidelines: Table 3.)

Studies comparing regular with as-needed inhaled b $_2$ -agonists in patients with asthma not taking inhaled corticosteroids have shown that regular treatment confers no benefit over as-needed inhalation and can have adverse consequences. A randomized controlled trial in 255 patients with mild intermittent asthma, comparing salbutamol taken as-needed with regular treatment, found no difference at 16 weeks in respiratory symptoms, airway function, or frequency of exacerbations. However, those taking regular salbutamol took more salbutamol, showed more variability in peak flow rates, and had increased airway responsiveness to inhaled methacholine. Short-acting b $_2$-agonists should, in general, be reserved to provide reversal of acute airway narrowing, taken as-needed or prior to exercise in patients with exercise-provoked asthma, except in cases of severe asthma not controlled with maximal doses of inhaled corticosteroids and additional long-acting b $_2$-agonist (step 4 BTS guidelines), when regular inhaled short-acting b $_2$-agonists can be added.

By contrast, long-acting b $_2$-agonists—salmeterol and formoterol—are taken regularly and their addition to inhaled corticosteroids has been shown to be more effective than doubling the dose of the inhaled corticosteroids in improving symptoms and lung function in patients not controlled by low-dose inhaled corticosteroids. Regular treatment with long-acting b $_2$-agonists, when taken with inhaled corticosteroids, has not been associated with deterioration in asthma control.

Two epidemics of asthma deaths, the first in the 1960s in six countries which followed the introduction of Isoprenaline Forte, the second in the mid-1970s in New Zealand after the introduction of fenoterol, have led to concerns about the safety of inhaled b-agonists. Case–control studies have also identified an association between asthma deaths and overuse of inhaled b $_2$-agonists. However, the evidence for cause and effect is confounded because overuse of b $_2$-agonists to treat frequent symptoms is more likely to occur in patients with severe uncontrolled asthma who are at high risk of a fatal attack. The evidence for cause and effect in the asthma epidemics is stronger: the increased death rates that followed the introduction of the particular inhaled b-agonists fell rapidly after recognition of the association and no other plausible explanation has been advanced. Isoprenaline is a non-selective b-agonist and fenoterol is less selective than salbutamol and terbutaline. Both drugs were marketed in high dose and are cardiotoxic in the presence of hypoxia, hence both epidemics may have been due to the acute cardiac effects of b-agonists inhaled in high dose by hypoxic patients with acute severe asthma. The evidence that selective b $_2$-agonists formulated in lower doses have a similar cardiotoxic effect and cause asthma deaths outside these epidemics is limited to associations in case–control studies, from which it is not possible to infer cause and effect. However, a small effect can be difficult to detect and, as pointed out by Tattersfield, if a fatal arrhythmia occurred in 1 in 8000 patients treated with b-agonists each year this would account for 50 per cent of asthma deaths in patients under 65 years, but its detection would require observation of many thousands of patients.

Methylxanthines

Theophylline is the pharmacologically active methylxanine most usually employed in clinical medicine, because of its greater bronchodilator activity, less erratic absorption, and longer half-life. More predictable theophylline absorption can be obtained by slow-release formulations and the addition of ethylene diamine to theophylline (aminophylline) provides the increased solubility required for intravenous administration. None the less, theophylline has a relatively narrow 'therapeutic window' for a safe and effective dose, with wide differences among individuals in its metabolism, which can also be adversely affected by several extrinsic factors to cause clinically important side-effects. The most common side-effects are 'caffeine-like': anorexia, nausea, and vomiting, followed by headache and insomnia. At higher concentrations, potentially fatal fits and arrhythmias can occur.

Theophylline relaxes bronchial smooth muscle and, like b-agonists, is a functional antagonist that causes bronchial muscle relaxation irrespective of the constrictor stimulus. Its action was thought to be mediated via phosphodiesterase inhibition increasing intracellular cyclic AMP. However, the intracellular concentration necessary for theophylline to achieve this is some 20 times greater than its therapeutic plasma levels. More recently, anti-inflammatory activity in 'subtherapeutic' concentrations has been suggested as a possible mechanism of action in asthma.

In addition to bronchodilatation, theophylline increases the force and rate of heart contraction and causes vasodilatation. In toxic doses it can cause arrhythmias, which may be fatal. It is also a central nervous system stimulant, causing increased alertness and, in toxic doses, confusion, irritability, and fits.

Theophylline is metabolized to inactive products by cytochrome P-450 enzyme-dependent pathways in the liver. The variation in its metabolism among individuals is large, and the half-life of theophylline can vary between 4 and 24 h. This may in part reflect the wide range of exogenous factors that influence hepatic metabolism of the drug: its half-life is increased by several drugs—cimetidine (but not ranitidine), erythromycin, ciprofloxacin, and oral contraceptives—and decreased by rifampicin, barbiturates, and carbamazepine (Table 4).

Bronchodilatation increases linearly with increase in serum theophylline concentration. Toxic effects also show a similar linear relationship, but at higher concentrations, although there are considerable differences between individuals in the serum concentration at which side-effects occur, in some occurring at low serum concentrations. Serum concentrations of between 10 and 20 $\mu g/ml^{-1}$ combine substantial bronchodilatation with a low risk of side-effects.

Safe and effective theophylline treatment requires monitoring of plasma concentration at the start of treatment to ensure a concentration within the therapeutic window, and subsequently to ensure its maintenance. This can be measured by immunoassay, when in patients on regular, twice daily, maintenance treatment the difference between peak and trough levels is usually between 5 and 10 $\mu g/ml^{-1}$, although greater in smokers, who may require treatment three times daily.

Theophyllines are now most commonly used as an additional treatment in asthma that is inadequately controlled by inhaled corticosteroids. Comparison in a randomized controlled trial of budesonide at 400 µg twice daily plus theophylline (250 or 375 mg twice daily) with budesonide at 800 µg twice daily for 3 months in 62 patients, whose asthma was not controlled by the lower dose of inhaled steroid, found the combination of low-dose inhaled corticosteroid and theophylline provided the greater improvement in lung function, peak flow variability, and b $_2$-agonist use. In those receiving it, median theophylline concentration was 8.7 µg/ml, hence this additive effect was achieved at doses lower than those conventionally considered therapeutic (10 to 20 µg/ml) and similar to that provided by inhaled salmeterol. This suggests that oral theophylline may therefore be an appropriate alternative to inhaled salmeterol at stage 3 of BTS guidelines.

Sodium cromoglycate

Sodium cromoglycate is a bischromone that has prophylactic but not bronchodilator activity in asthma. Originally available as a dry powder (mixed with lactose), it is now also formulated as a metered dose inhaler and as a nebulizer solution for children.

In inhalation tests, sodium cromoglycate inhibits asthmatic reactions provoked by inhaled allergen, exercise, and other provocative stimuli including sulphur dioxide and adenosine, although it is less effective in a dose of 20 mg than salbutamol at 200 µg in preventing asthma provoked by exercise. The major benefit of sodium cromoglycate is its safety. However, it is less effective than inhaled corticosteroids, needs to be taken four times daily, and its use is now generally reserved for children with mild asthma and taken immediately prior to exercise to prevent exercise-induced asthma. Sodium cromoglycate is no longer recommended in the new BTS Guidelines.

Nedocromil sodium

Nedocromil sodium has a similar activity profile to sodium cromoglycate. It is available as a metered dose inhaler and needs to be taken four times a day.

Its activity is equivalent to low-dose inhaled corticosteroid and it can be used either in place of inhaled corticosteroid or to reduce the dose of inhaled corticosteroid. Nedocromil sodium is an alternative to inhaled corticosteroids if inhaled steroids cannot be used in step 2 of BTS guidelines.

The 'stepped' approach to treatment of asthma

The purpose of treatment of asthma varies from the reversal of occasional mild symptoms to the restoration of normal life in a patient with severe disabling ill health. Treatment needs will vary greatly among different patients and underlie the 'stepped' approach to treatment, which is the basis of current guidelines for asthma

management, including the British Thoracic Society (BTS) guidelines published in 1997. In the stepped approach, asthma severity is defined by the treatment step (1 to 5) needed to achieve and maintain good control. Inhaled corticosteroids form the mainstay of maintenance treatment for the majority of patients. In deciding the starting treatment step, a 'start high–go low' policy has been recommended, recognizing that initial control of airway inflammation is likely to need a higher dose of inhaled corticosteroid than will subsequent maintenance of disease control. Sodium cromoglycate or nedocromil sodium may be used as alternative inhaled anti-inflammatory drugs, but are less efficacious than inhaled corticosteroids and are now primarily used for the treatment of mild asthma in children.

Inhaled b_2-agonists are used primarily for symptomatic relief. There is good evidence that regular treatment with short-acting b_2-agonists alone is less effective than regular inhaled corticosteroids, and provides less good control of asthma—both symptomatically and of lung function—when taken regularly than as-needed.

Steps 1 to 5 of the BTS guidelines identify the treatment requirements for asthma of increasing severity (Table 3). Failure to achieve treatment targets at any step implies the need to increase treatment to a step that provides good control. Figure 6 shows the proportion of patients on each of the BTS guideline steps.

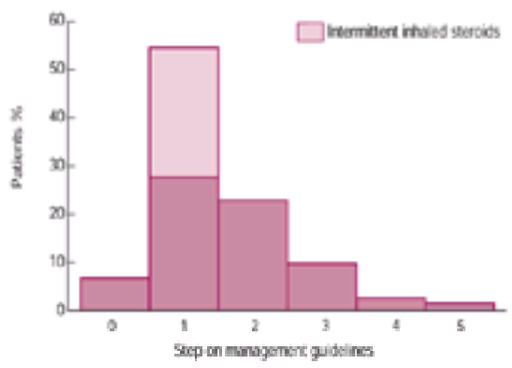

Fig. 6 Proportion (per cent) of 3372 patients with asthma on each of the BTS guideline steps.

Step 1

This comprises patients with mild intermittent asthma whose asthma is controlled by the use of an inhaled shorter-acting b_2-agonist (such as salbutamol or terbutaline) less than once a day. Requirement for more regular treatment implies the need for regular anti-inflammatory treatment (that is, a higher step).

Step 2

This comprises patients with mild persistent or intermittent asthma that is of sufficient frequency to require regular anti-inflammatory treatment. Inhaled corticosteroids are the most effective and commonly used anti-inflammatory drugs, with sodium cromoglycate and nedocromil sodium as alternatives. Treatment with an inhaled corticosteroid should be started at a dose of 400 µg of beclomethasone twice daily or equivalent. This dose should be continued for at least 3 months, the period when most benefit of the inhaled steroid is obtained, before reducing the dose to the minimum required for maintenance of good control. This can be achieved by reducing the dose by 25 to 50 per cent every 1 to 3 months. Short-acting b_2-agonists are used as required for symptomatic relief.

Step 3

- This comprises patients with moderate persistent asthma that is not controlled despite adherence to inhaled steroid (or equivalent prevention) treatment and correct inhaler technique. The recommended treatment is a regular long-acting b_2-agonist (salmeterol or formoterol) added to low-dose beclomethasone (less than 800 µg/day) or equivalent.

Two randomized controlled trials have shown that the addition of salmeterol to a low-dose inhaled corticosteroid was more effective than doubling the dose of inhaled corticosteroid in improving symptoms and lung function. In a third randomized controlled trial the addition of formoterol increased the benefit provided by both low-dose and high-dose budesonide. These studies suggest that the addition of a long-acting b_2-agonist to low-dose inhaled corticosteroid is the preferred option for improvement of day to day control of asthma that is not controlled by low-dose inhaled corticosteroid, whereas reducing the frequency of exacerbations of asthma was best achieved by increasing the dose of inhaled steroid and additional long-acting b_2-agonist.

If there is no benefit, stop long-acting b_2-agonist and consider:

a. trial of: slow-release theophylline added to low-dose inhaled corticosteroid. (One randomized controlled trial of 133 patients found the addition of oral slow-release theophylline to beclomethasone at 400 µg/day for 6 weeks was as effective in improving symptoms and lung function as doubling the dose of beclomethasone to 800 µg/day.)
   or
b. b_2-agonist tablet
   or
c. leucotriene receptor antagonist.

Step 4

This involves a combination of the alternatives in step 3—high-dose inhaled corticosteroid (up to 2000 µg/day of beclomethasone or equivalent) together with a long-acting b_2-agonist or alternative bronchodilators, such as an oral b_2-agonist or slow-release theophylline or leucotriene receptor antagonist, if these were not tried at Step 3.

Step 5

Patients who fail to respond to these combinations of step 4 treatments will require the addition of an oral corticosteroid. This is an important decision, which should be made in consultation with a respiratory physician and requires continuous monitoring to identify, and where possible avoid, the associated side-effects. The risk of osteoporosis can be reduced by taking regular exercise, and by not smoking. Bone density should be monitored regularly and where appropriate calcium, vitamin D, and bisphosphonates given. Postmenopausal women can benefit from hormone replacement therapy.

Most patients with asthma in the community have disease of severity indicated for steps 1 and 2 of the BTS guidelines; 'difficult' asthma, requiring treatment equivalent to step 5, constitutes less than 5 per cent of patients. A recent community study of five large general practices in south Nottinghamshire (a population of 38 865) found a prevalence of asthma of 9 per cent, with a peak of 17 per cent in 10 to 14 year olds, falling to less than 6 per cent in adults aged over 70 years. Most of those diagnosed with asthma were either not receiving treatment (8 per cent) or receiving treatment equivalent to steps 1 and 2 (76 per cent); 11 per cent were on step 3, and some 5 per cent on steps 4 and 5. The authors endeavoured to assess the effectiveness of asthma treatment in this population by measuring the proportion of patients who during a 1-year period required oral corticosteroid courses or were prescribed 10 or more short-acting b_2-agonist inhalers: 12.5 per cent of patients not taking them regularly had been prescribed one or more courses of oral corticosteroids, 1.6 per cent on three or more occasions; 13.6 per cent of patients had been prescribed 10 or more short-acting b_2-agonist inhalers. Both were increasingly more frequent in patients on steps 3 or higher of the BTS guidelines. However, because only a minority of patients (15 per cent) were in these categories, more than half of the patients who required either oral corticosteroids or 10 or more b_2-agonist inhalers were on steps 1 or 2, indicating continuing significant morbidity among some patients with asthma receiving either low-dose or no anti-inflammatory treatment.

Difficult asthma

Difficult asthma is asthma not controlled by maximum doses of inhaled treatment, including inhaled corticosteroids in doses of beclomethasone of up to 2000 µg/day or equivalent, with additional treatment such as long-acting b $_2$-agonists. It occurs uncommonly, probably in less than 5 per cent of patients with asthma, but is important. The severity of disease and associated disability is considerable, the risk of near fatal and fatal asthma is high, and the adverse consequences of treatment are severe, and only worthwhile if treatment is demonstrably effective.

Failure to respond to maximal inhaled treatment can result from several causes (Table 5). It is clearly important to confirm the diagnosis of asthma and exclude other diseases that may be mistaken for it. Demonstration of spontaneous variability or reversibility of airflow limitation is important to avoid treatment of irreversible airflow limitation, due either to localized obstruction or chronic obstructive pulmonary disease, with ever increasing doses of oral corticosteroids. Assessment of reversibility may require a formal trial of oral prednisolone in a dose of 30 to 40 mg taken each morning for 1 month to determine whether this provides a significant improvement in airway function.

The conditions most easily mistaken for asthma were considered earlier (see differential diagnosis). Once the diagnosis of asthma has been confirmed, it is important to ensure good inhaler technique and adherence to prescribed treatment: failure to take treatment properly is a common reason for failure to respond. This may reflect lack of understanding that preventive treatment needs to be taken regularly and not 'as needed', or poor inhaler technique. Patients may take preventive treatment irregularly because, unlike short-acting b $_2$-agonists, it does not provide immediate symptomatic relief. Others may be inappropriately concerned about potential side-effects or resent the need to take regular treatment. In patients taking oral corticosteroids, blood eosinophil count is markedly reduced and often reported as zero, hence a blood eosinophil count above 0.3 in a patient prescribed oral steroids suggests that these are not to being taken regularly, or alternatively another disease, particularly Churg–Strauss syndrome, may accompany the asthma.

One study, using a computerized timing device in a dry powder inhaler, found that only 18 per cent of patients took inhaled steroids as prescribed. Adherence to inhaled treatment is difficult to monitor, and poor adherence to treatment may be suspected as a cause of difficult asthma in patients whose asthma improves when treatment, although unchanged, is supervised. Patient understanding of the effectiveness of regular treatment may also be reinforced by this means.

Unidentified provoking factors include allergens, commonly domestic pets, in particular cats, whose allergens can be present in sufficient concentrations to cause asthma for several months after cats have left the home. Sensitizing agents encountered at work can also cause asthma that is poorly controlled by inhaled treatment. Early identification and avoidance of the cause is important to minimize the risk of development of chronic asthma. Aspirin, NSAIDs, and b-blockers can also be important provoking factors.

Rhinitis commonly accompanies asthma and its treatment can be assocaited with improvement in asthma and airway hyperresponsiveness. The explanation for this association is unclear, but may be a consequence of inflammatory mediators in post-nasal drip increasing airway responsiveness and provoking cough. Similarly gastro-oesophageal reflux can provoke cough and worsen asthma; where this is suspected a trial with a proton pump inhibitor such as omeprazole should be instituted, but objective improvement in asthma with such treatment is uncommon.

Uncommonly asthma may be a manifestation of systemic disease, particularly a systemic vasculitis, Churg–Strauss syndrome, when asthma, which can be difficult to control, is accompanied by a high blood eosinophil count (usually more than 1.5). Other manifestations including eosinophilic pneumonia, pleural and pericardial effusions, and mononeuritis multiplex. Effective treatment requires high-dose oral corticosteroids and in some cases immunosuppressant treatment, initially with cyclophosphamide and subsequently azathioprine (see Chapter 17.11.5).

Nocturnal asthma can persist in some patients despite treatment with inhaled corticosteroids, which provides good daytime control. This may be improved by the addition of a long-acting b $_2$-agonist or slow-release theophylline.

Premenstrual deterioration of asthma is not uncommon, and can in some patients be severe and unresponsive to corticosteroid treatment. Symptoms characteristically increase, and PEFR falls, 2 to 5 days before the menstrual period, improving with the onset of menstruation. This coincides with the fall in progesterone secretion and increase in oestrogen:progesterone ratio. Asthma in some patients is improved by treatment with intramuscular, but not oral, progestogen during the week before menstruation. Patients with severe premenstrual exacerbations can require hospital admission, in some cases ventilation, and may (very rarely) experience improvement only by surgical removal of the ovaries.

Brittle asthma is characterized by widely varying peak flow rates uncontrolled by maximum inhaled treatment. Two patterns of brittle asthma have been distinguished:

1. type I—persistent daily chaotic variability in peak flow (usually greater than 40 per cent diurnal variation in PEFR more than 50 per cent of the time); and
2. type II—sporadic sudden falls in PEFR against a background of usually well-controlled asthma with normal or near normal lung function.

Treatment of brittle asthma of both types is difficult. Type I brittle asthma, not responding to inhaled long-acting b $_2$-agonists, can be improved by subcutaneous terbutaline administered via an insulin infusion pump, usually in a dose of between 3 and 12 mg in 24 h: this treatment is limited by side-effects, of which the most important is muscle cramp associated with increased levels of serum creatinine kinase.

Type II brittle asthma requires immediately available treatment for what can be catastrophic falls in peak flow. The speed of onset of attacks requires immediately injected bronchodilator and such patients should have preloaded adrenaline syringes (such as Epi-pen) available at all times and wear a Medic-alert bracelet. Potential provoking factors, such as foods, should be sought and avoided.

Asthma in a very few patients is only controlled with continuous oral corticosteroids, often in high doses; reduction in dose is followed by worsening of asthma. The term corticosteroid-dependent asthma has been used for the condition in these patients. They differ from those with corticosteroid-resistant asthma in their response to oral corticosteroids. Patients with corticosteroid-resistant asthma show no response to oral corticosteroids, even in high dose, but do show spontaneous variability of peak flow and reversibility with inhaled bronchodilators. Corticosteroid-resistant asthma is very uncommon, estimated at between 1 in 1000 and 1 in 10 000 patients. It probably forms the end of a spectrum of resistance to the anti-inflammatory activity of corticosteroids, to which corticosteroid-dependent asthma also belongs. Treatment of corticosteroid-resistant asthma is difficult, but should include stopping oral corticosteroids, which still cause side-effects, and relying on other forms of treatment, including long-acting b $_2$-agonists.

Treatment of acute exacerbations of asthma

Asthma exacerbations are episodes of progressively worsening airway narrowing, associated with increasing shortness of breath, cough, wheezing, chest tightness, or some combination of these. Exacerbations can vary in severity from episodes which patients are able to manage themselves by following an agreed treatment plan, to severe and potentially life-threatening episodes that require medical attention and hospital admission. Severe attacks also vary in their speed of onset, ranging from deterioration over days to episodes that progress rapidly and can become life-threatening within minutes or hours. In about one-half of cases of fatal asthma the attack lasted more than 24 h, in one-quarter less than 1 h. Fatal or near fatal attacks of asthma are associated with the following.

1. Attacks may occur in patients who have previously required hospital admission for severe asthma and who require regular oral steroid treatment.
2. The doctor may fail to recognize the severity of the asthma. This can be minimized by making appropriate objective measurements of respiratory, heart, and peak flow rates to assess severity.
3. Patients may fail to recognize the severity of their asthma. Those with long-standing asthma can become accustomed to their symptoms and not appreciate an important increase in their severity, which may persist for days or weeks. This may also be associated with psychosocial problems and poor adherence to treatment.
4. The asthma attack may have been undertreated or given inappropriate treatment. Failure to use oral corticosteroids in adequate doses early in an exacerbation is probably the single most remediable factor. The use of sedatives or anxiolytics to reduce the anxiety or agitation that can often accompany acute severe asthma is absolutely contraindicated.

Many of these problems can be overcome by improved patient understanding, allowing them to have control over their illness supported by a jointly agreed

management plan. A systematic review of 22 studies comparing self management education for asthma with usual care found that self management reduced hospital admissions and days off work by nearly one-half and emergency room visits by one-quarter. Results were best when self management included a written action plan.

Moderate exacerbations

Exacerbations of asthma with increased symptoms, both during the daytime and at night, frequently follow an upper respiratory tract infection, allergen exposure in allergic individuals, or a reduction in anti-inflammatory treatment. Increase in symptoms associated with deterioration in peak flow can often be treated adequately by the patient increasing their frequency of inhaled short-acting bronchodilators, doubling the dose of inhaled steroids, or taking a short course of oral steroids. Several studies have shown that a short course of oral steroids taken at the start of an acute exacerbation reduces the need for hospital admission, the frequency of relapse, and the need for b_2-agonists. One recent overview of seven randomized controlled trials in 320 persons found that systemic corticosteroids, taken at the onset of an acute exacerbation, reduced hospital admissions in both children and adults by 65 per cent in the first week compared with placebo, an effect maintained for 21 days. No difference was observed between the use of oral and intramuscular corticosteroids.

Acute severe asthma

Acute severe asthma is a potentially life-threatening increase in the severity of asthma that can develop over minutes, hours, or usually days and which has often failed to respond to conventional, inhaled bronchodilator treatment. It is usually the outcome of airways increasingly narrowed by the consequences of chronic inflammation, resulting in increasing resistance to airflow identified as a reduction in PEFR and FEV_1, hyperinflated lungs, ventilation–perfusion inequality, and hypoxia, which is the most serious consequence of severe asthma. Initially this stimulates alveolar hyperventilation, but with increasing airway narrowing and exhaustion, arterial PO_2 continues to fall while arterial PCO_2 rises to normal, and subsequently increases steeply, due to alveolar hypoventilation: PCO_2 rises into the normal range when FEV_1 is some 25 per cent and PEFR 30 per cent of predicted normal values.

The clinical features of importance in identifying acute severe asthma and assessing its severity are shown in Table 6. Patients are usually extremely breathless and unable to complete sentences in one breath. A rapid respiratory rate and heart rate are good markers of severity of asthma and hypoxia. Although anxiety and increased use of b_2-agonists can increase heart rate, tachycardia should not be ignored by attributing it to these factors. An objective measure of airflow limitation should be made because severity is difficult to assess clinically with accuracy. Although PEFR is an effort-dependent measurement, it can usually be obtained from patients with severe asthma: a value of less than 50 per cent of predicted or of the recent best value in an adult aged less than 50 years indicates severe asthma, a value of less than 33 per cent, a potentially life-threatening attack.

Blood gas measurements should be made in adults seen in hospital as an important guide to the severity of asthma; children can often safely be managed by measurement of SaO_2 alone. Most patients admitted to hospital with acute severe asthma are hypoxic, of whom about one-third will have a PO_2 of less than 8 kPa (60 mmHg). PCO_2 is reduced in patients with moderately severe asthma, but with increasingly severe airways obstruction and fatigue, PCO_2 falls and subsequently rises in parallel with a falling PO_2. A normal PCO_2 in a hypoxic patient with acute severe asthma indicates impending hypoventilation, with a rapidly increasing PCO_2, falling PO_2, acidosis, narcosis, and death.

Treatment of acute severe asthma

The aims of the treatment of acute severe asthma are to reverse the hypoxia, airflow limitation, and airway inflammation with oxygen, bronchodilators, and corticosteroids (Table 7).

Oxygen

Oxygen relieves the hypoxia present in most patients with acute severe asthma. High concentrations of inspired oxygen are safe in patients with asthma, and certainly in those aged less than 50 years; a high $PaCO_2$ in acute severe asthma reflects fatigue and the severity of airways obstruction and is not a contraindication for a high concentration of inspired oxygen. Oxygen can be administered by nasal cannulas or by facemask in the highest available concentration (usually FiO_2 between 40 and 60 per cent). The aim is to increase SaO_2 to above 92 per cent or PaO_2 to above 9 kPa (80 mmHg).

Bronchodilators

The purpose of bronchodilator treatment in acute severe asthma is to reverse airway narrowing due to smooth muscle contraction, before the onset of the anti-inflammatory action of corticosteroids, which usually takes 6 to 12 h from administration.

Inhaled high-dose b_2-agonists, salbutamol and terbutaline, are now used as initial treatment, usually from a nebulizer. The benefit of the nebulizer is that it allows inhalation of bronchodilator to be driven by a high flow of oxygen. This can be important in severe asthma as b_2-agonists may increase ventilation–perfusion inequality and consequently arterial hypoxia, hence in hypoxic patients b_2-agonists should not be administered without oxygen. Nebulized salbutamol in a dose of 5 mg, or terbutaline in a dose of 10 mg, driven by 6 litre/min oxygen can be given safely by trained ambulance crews during transfer to hospital. However, nebulizers are inefficient and widely variable in their performance, which has led to the suggestion of large volume spacers as alternative delivery systems. In adults and children with severe but not life-threatening asthma treated outside hospital, inhalation of b_2- agonist by nebulizer has not been found to provide additional bronchodilatation compared with use of a metered dose inhaler via a spacer, but it is important to note that the studies on which this is based are of patients without life-threatening asthma who did not require hospital admission. Spacers do not easily allow concurrent administration of oxygen and require patient co-operation, which can be difficult in severely breathless patients.

The available intravenous bronchodilators are b_2-agonists and theophylline. The theoretical advantage of intravenous b_2-agonists is access to peripheral airways so narrowed that they cannot be reached by inhalation. However, inhaled salbutamol is rapidly absorbed from the lungs, reaching a peak concentration within 10 min of inhalation. The major disadvantage of intravenous b_2-agonists, by comparison with inhalation, is the greater frequency of systemic side-effects. What is relevant is whether intravenous b_2-agonists provide additional improvement in bronchodilator response to inhaled b_2-agonists and corticosteroids. In adults with acute asthma, intravenous salbutamol at 12 µg/min given 4 hourly after an initial dose of 5 mg of nebulized salbutamol and intravenous hydrocortisone provided greater bronchodilation compared with three further doses of nebulized salbutamol given during 2 h, although the patients receiving intravenous salbutamol had a greater increase in heart rate. Similarly, in a study of children with acute severe asthma, the addition of salbutamol (15 µg/kg) in a 10-min infusion to nebulized salbutamol and intravenous hydrocortisone was associated with a reduced period of need for inhaled salbutamol, a decreased requirement for oxygen, and earlier discharge from the emergency department.

The use of intravenous aminophylline in the treatment of asthma has decreased with the recognition that it does not provide additional benefit to repeated nebulized b_2-agonist bronchodilators in the initial hour of emergency treatment. This, together with its narrow therapeutic window, need for drug monitoring, and interactions with other drugs has led to its replacement as first-line bronchodilator treatment of asthma by inhaled b_2-agonists. However, it is recommended as additional therapy for patients not responding to initial treatment with inhaled b_2-agonists and corticosteroids, and as initial treatment in the very severely ill patient with a normal or high PCO_2. In those who have not been taking theophylline prior to admission, a loading dose of 5 mg/kg body weight over 20 min should be followed by a maintenance dose of 0.5 to 0.9 mg/kg body weight per hour until a serum level of 10 to 20 µg/ml is obtained. The loading dose should be omitted in patients currently taking theophyllines, in whom the serum concentration should be measured. The infusion rate should be decreased in patients with liver or heart failure or those taking cimetidine, macrolide antibiotics, or ciprofloxacin (Table 4). Toxic side-effects are increasingly common in patients with serum levels exceeding 25 µg/ml, ranging from gastrointestinal symptoms to fits and cardiac arrhythmias.

The BTS guidelines for the treatment of acute severe asthma recommend the administration of intravenous b_2-agonist or aminophylline to patients not responding to oxygen, inhaled b_2-agonists, and corticosteroids after the addition of inhaled ipratropium, or as initial treatment for those with life-threatening features.

Antimuscarinics

The purpose of antimuscarinic treatment is to reverse airway narrowing caused by increased vagal tone and not responsive to high-dose inhaled b $_2$-agonists. Several studies have suggested the addition of an inhaled antimuscarinic provides additional benefit in the treatment of acute severe asthma. One multicentre study found the addition of inhaled ipratropium to inhaled salbutamol increased bronchodilation by 10 to 20 per cent compared with inhaled salbutamol alone. However, other studies have failed to demonstrate additional benefit from the addition of ipratropium to inhaled high-dose b $_2$- agonist. The evidence is consistent with considerable individual variation in response to inhaled antimuscarinics, maximal response occurring in those with exaggerated cholinergic tone. Inhaled antimuscarinics have few side-effects, can provide benefit in some patients, and are recommended in the guidelines as additional treatment in a dose of 0.25 to 0.5 mg for patients not responding to initial high-dose inhaled b $_2$-agonists.

Corticosteroids

Systemic corticosteroids are given in acute severe asthma to reverse the underlying airway inflammation. The anti-inflammatory action requires 6 to 12 h from administration for demonstrable bronchodilatation to occur. Steroids may also reverse b $_2$-receptor desensitization induced by regular b $_2$- inhalation within 1 h of their administration. The value of corticosteroid treatment in acute severe asthma was first demonstrated in a randomized controlled trial in 1956 and their value in the treatment of acute severe asthma has since been generally accepted. They are usually given by intravenous administration, but other than in life-threatening asthma and in patients vomiting or unable to swallow, there is no demonstrable advantage of intravenous over oral administration. When indicated, intravenous doses initially of 200 mg every 4 to 6 h can be followed by oral prednisolone in a dose of 40 to 60 mg/day. The duration of treatment with oral prednisolone will depend on the severity and rate of recovery of the acute episode. In general, oral prednisolone should be continued until resolution of the acute episode with return to usual daytime activities, resolution of nocturnal symptoms, and PEFR within 80 per cent of the patient's predicted or best values. Short courses of oral corticosteroids (taken for less than 3 weeks) do not need to be tapered, provided patients are taking an appropriate dose of inhaled corticosteroid. Although some studies in patients with relatively mild exacerbations of asthma (PEFR greater than 60 per cent of predicted or best) have suggested that high-dose inhaled steroids are an effective alternative to oral corticosteroids, these results should not be extrapolated to acute severe asthma where the recommended guideline is that all patients should be treated with systemic corticosteroids.

Intensive care and intermittent positive-pressure ventilation

Most attacks of acute severe asthma respond to treatment with high-concentration inspired oxygen, systemic corticosteroids, and inhaled b $_2$-agonists. However, this treatment is insufficient in a few patients who require intensive care and, on occasion, intermittent positive-pressure ventilation (**IPPV**). This occurs in two particular situations: patients who have a catastrophic hyperacute attack, and patients whose asthma progressively increases in severity despite maximal bronchodilator and corticosteroid treatment. The indications for intensive care and IPPV are given in Table 8. Patients with increasing drowsiness or who lose consciousness with hypoxia and worsening hypercapnia require IPPV, as do those who suffer a respiratory arrest. However, because of the high inflation pressures needed to overcome the high airway resistance and poorly compliant hyperinflated lungs and chest wall, IPPV in acute severe asthma can be difficult and hazardous. High inflation pressures can cause barotrauma with pneumomediastinum and, on occasion, pneumothorax. In addition up to one-third of patients can develop clinically significant hypotension, requiring inotropic support.

Further reading

Barnes PB (1998). Current issues for establishing inhaled corticosteroids as the anti-inflammatory agents of choice in asthma. *Journal of Allergy and Clinical Immunology* **101**, 5427–33.

Barnes PJ, Pederson S, Busse WW (1998). Efficiency and safety of inhaled corticosteroids. *American Journal of Respiratory and Critical Care Medicine* **157**, S1–S53.

British Thoracic Society (1997). The British guidelines on asthma management. *Thorax* **52**(Suppl 1), S1–21.

Drazen JM *et al*. (1996). Comparison of regularly scheduled with as needed use of albuterol in mild asthma. Asthma clinical research network. *New England Journal of Medicine* **335**, 841–7.

Evans DJ *et al*. (1997). A comparison of low dose inhaled budesonide plus theophylline and high dose inhaled budesonide for moderate asthma. *New England Journal of Medicine* **337**, 1412–18.

Garbelt JF *et al.* (1997). Nebulised salbutamol with and without ipratropium bromide in the treatment of acute asthma. *Journal of Allergy and Clinical Immunology* **100**, 165–70.

Greening AP *et al.* (1994). Added salmeterol versus higher dose corticosteroid in asthma patients with symptoms on existing inhaled corticosteroid. *Lancet* **344**, 219–24.

Haahtela T *et al*. (1991). Comparison of a b_2 agonist terbutaline with an inhaled corticosteroid budesonide in newly detected asthma. *New England Journal of Medicine* **325**, 388–92.

Haahtela T *et al*. (1994). Effects of reducing or discontinuing inhaled budesonide in patients with mild asthma. *New England Journal of Medicine* **331**, 700–5.

Marquette CH *et al*. (1992). A 6 year follow up study of 145 asthmatic patients who underwent mechanical ventilation for near-fatal attack of asthma. *American Review of Respiratory Diseases* **146**, 76–81

Newman Taylor AJ (1995). Environmental determinants of asthma. *Lancet* **345**, 296–9.

Pauwels RA *et al*. (1997). Effect of inhaled formoterol and budesonide on exacerbations of asthma. *New England Journal of Medicine* **337**, 1405–11.

Walsh LJ *et al*. (1999). Morbidity from asthma in relation to regular treatment: a community based study. *Thorax* **54**, 296–300.

Woolcock AJ *et al.* (1996). Comparison of addition of salmeterol to inhaled steroids with doubling of the dose of inhaled steroids. *American Journal of Respiratory and Critical Care Medicine* **153**, 1481–8.

17.4.5Occupational asthma

A. J. Newman Taylor

Occupational asthma is asthma induced by an agent inhaled at work. Agents inhaled at work can aggravate pre-existing asthma, but the term occupational asthma is usually restricted to asthma initiated or induced by such agents.

Asthma may be initiated or 'switched on' either by respiratory irritants inhaled in toxic concentrations—irritant-induced asthma—or as the outcome of an acquired specific hypersensitivity response. Hypersensitivity-induced occupational asthma is considerably more common than irritant-induced asthma and is important to recognize because in the majority of cases it improves or resolves with the avoidance of further exposure. Furthermore, the earlier further exposure is avoided, the more probable is complete resolution of the asthma. Identification and avoidance of the specific occupational cause therefore provides one of the few opportunities to cure asthma in adult life.

Causes of occupational asthma

The known causes of irritant-induced asthma are relatively few and include well-recognized respiratory irritants such as chlorine and ammonia, as well as others such as toluene di-isocyanate inhaled in toxic concentrations. However, any respiratory irritant inhaled in concentrations toxic to the epithelial cells in the airways is a potential cause of reactive airways dysfunction syndrome.

By contrast, the number of reported causes of hypersensitivity-induced occupational asthma is considerable, and with the rapid development of biotechnology and the continuous introduction of newly synthesized organic chemicals is likely to increase. The causes described include proteins of animal, vegetable, and microbiological origin, naturally occurring organic chemicals, synthetic chemicals, and inorganic chemicals, particularly metal salts. Some of the more important are listed in Table 1.

Agents and occupations associated with occupational asthma

The proportion of cases of asthma in the general population that are attributable to an occupational cause is not known, although estimates have varied between 2 and 15 per cent. A recent community-based study in Spain of adults aged between 20 and 44 years estimated that the risk of asthma attributable to occupation was between 1 in 20 (5 per cent) and 1 in 15 (6.7 per cent). The highest risks occurred in spray painters, bakers, and laboratory technicians. During the past 10 years a voluntary reporting scheme for registering new cases of occupational lung disease seen by respiratory and occupational physicians in the United Kingdom (SWORD) has estimated the incidence of occupational asthma in different occupations and provided information about the relative importance of its causes. Both have remained remarkably stable during 10 years of reporting to the scheme. Organic agents, such as flour/grain, colophony, wood dust, and laboratory animals, and chemicals, isocyanates, and glutaraldehyde, account for some two-thirds of the reported cases (Table 2). The relative frequency of the different causes has also remained similar during this period, with the exception of some reduction in the proportion of cases attributed to isocyanates and an increase in the number of cases attributed to latex (Fig. 1).

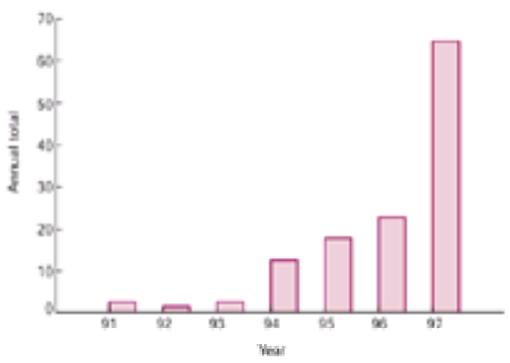

Fig. 1 Annual reports to SWORD of occupational asthma attributable to latex (1991–7).

The estimated annual incidence by occupation varied from 1380 per million per year in coach and other spray painters to 12 per million per year in transport and storage workers (i.e. a range of two orders of magnitude). The estimated annual incidence in high-risk occupations between 1992 and 1997 is shown in Table 3.

Determinants of occupational asthma

Four major determinants of occupational asthma have been identified, which vary in their importance in relation to different causes: exposure intensity, atopy, smoking, and HLA genotype. Intensity of exposure is the single most important determinant of irritant-induced occupational asthma. Evidence for exposure–response relationships has also been found for many of the causes of hypersensitivity-induced occupational asthma, both for proteins, including laboratory animal proteins and enzymes, and for low-molecular-weight chemicals, including acid anhydrides and complex platinum salts. The risk of asthma is increased in atopics exposed to many of the protein causes of occupational asthma such as enzymes, latex, and laboratory animals, and in cigarette smokers exposed to low-molecular-weight chemicals that induce IgE production, such as complex platinum salts and acid anhydrides. HLA DR3 has been associated with an increased risk of developing specific IgE and asthma in those exposed to acid anhydrides and complex platinum salts. It should be appreciated, however, that the prevention of occupational asthma is more effectively achieved by reducing the intensity of exposure to its causes in the workplace than by exclusion of 'susceptible' individuals who are atopic, cigarette smokers, or HLA DR3 positive.

Pathology and pathogenesis

The pathological changes in the airways of patients with asthma of occupational cause are no different in any important way from those in patients with asthma of other or unknown cause: a desquamative eosinophilic bronchitis with infiltration of the airway wall by eosinophils and lymphocytes, accompanied by desquamation of bronchial epithelial cells.

In common with asthma caused by allergy to proteins encountered in the general environment, hypersensitivity-induced occupational asthma is probably the outcome of T_{H2} lymphocyte stimulation, and the pathological features observed are primarily the consequence of T_{H2} lymphocyte–eosinophil interaction. The evidence for T_{H2} lymphocyte stimulation is in part direct, but primarily comes from evidence of specific IgE antibody to many, although not all, of the causes of occupational asthma. In a few cases specific IgG antibodies can also be detected. These seem to reflect exposure, whereas IgE is more closely associated with disease. In general, specific IgE has been identified with the protein causes of occupational asthma, but only with a minority of the non-protein causes. Whilst it is likely that the majority of the low-molecular-weight chemicals that cause occupational asthma do so by binding to body proteins and acting as haptens, the difficulties of preparing the relevant hapten–protein conjugate *in vitro* have limited demonstration of this process, other than with chemicals such as acid anhydrides and reactive dyes that form stable conjugates with human serum albumin.

Clinical features

Irritant-induced asthma

Asthma caused by the inhalation of an irritant chemical in toxic concentrations is usually one manifestation of general tissue injury to exposed mucosal surfaces—eyes, nose, throat, and bronchial airways. The onset of symptoms follows a single identifiable exposure to a toxic chemical. Running, swelling, and discomfort of the eyes, running and obstruction of the nose, and painful throat usually occur within minutes of the exposure, and symptoms of asthma (shortness of breath, wheezing, chest tightness, and cough) develop within a few hours—certainly within 24 h—of inhalation of the chemical in toxic concentrations. Respiratory symptoms often have the characteristic circadian pattern characteristic of asthma: they are more severe during the night and on waking than during the daytime. In most cases asthma resolves spontaneously within a few weeks, but can occasionally persist for several years, if not indefinitely.

Hypersensitivity-induced asthma

In the commoner cases of hypersensitivity-induced occupational asthma respiratory symptoms develop insidiously and do not follow a single identifiable exposure to its cause. Asthma develops after an initial symptom-free period of exposure, commonly within 1 year of starting a new job or changing duties at work, although in some cases asthma may not develop until there have been several years of exposure. The onset of asthma may have been preceded or be accompanied by 'hay fever'-like symptoms of the nose and eyes. Characteristically, symptoms become increasingly severe during the working week and improve during absences from work during holidays and at weekends. However, the patient may not appreciate the relationship of respiratory symptoms to work, particularly when symptoms develop during the second half of the day and are most severe, as is characteristic of asthma, in the evenings, during the night, and on waking in the morning. Asthmatic symptoms can also persist for several days after avoidance of exposure, in which case appreciable symptomatic improvement at weekends does not occur, but improvement is usually sufficient to be appreciated by the end of a 2-week holiday or deterioration to be recognized on return to work. With continuing exposure asthma can become chronic and the relationship between symptoms and periods at work less clear, although even in these circumstances it is usual for some symptomatic improvement to occur on avoidance of exposure, although this may take several weeks.

The findings on clinical examination depend upon the severity of the asthma at the time of the examination. There may be no abnormal findings if seen when away from exposure. During a period of symptomatic exposure the patient will have typical signs of airflow limitation with breathlessness and wheeze, and depending on severity, other signs described in Section 17.4.

Diagnosis

The diagnosis of occupational asthma should be considered in any adult who develops asthma or whose asthma has deteriorated in working life. In the case of irritant-induced asthma the association of the onset of asthma with inhalation of a toxic chemical is usually clear. The association of asthma caused by a specific hypersensitivity reaction is often less apparent, and the diagnosis is based on the following:

1. Exposure to a sensitizing agent at work.
2. Characteristic history of:
   a. onset of asthma after an initial symptom-free period of exposure;
   b. deterioration in symptoms during periods at work and improvements during absence from work.
3. Results of objective investigations:
   a. lung function tests
   b. immunological tests
   c. inhalation tests.

Objective investigations

Lung function tests

The most commonly used criterion for diagnosing asthma—improvement in airflow limitation (usually measured as forced expiratory volume in 1 s (**FEV1**) or peak expiratory flow (**PEF**)) after inhalation of broncho-dilator—is often absent in cases of occupational asthma because lung function may be normal when the patient is seen away from work and, if present, does not identify a work relationship.

The measure of lung function most commonly used to identify work-related asthma is serial self-recorded PEF. A patient with suspected occupational asthma is asked to record his or her PEF at intervals of 2 to 3 h for a month from waking to sleeping, and at night if awoken, both during periods at and absences from work. The results can be summarized in a graphical display that records the best, worst, and average values for each day, allowing comparison of PEF during days at work with days away from work (Fig. 2). The diagnostic value of the test depends on the reproducibility of the patient's forced expiratory manoeuvres and their honesty and compliance. Concurrent treatment can influence the results, particularly when treatment is systematically increased during periods at work and reduced during absences from work. When possible treatment should be kept constant during the period of testing, or at least recorded. Comparisons with the results of inhalation testing as the 'gold standard' have shown that serial self-recorded PEF measurements are a sensitive and specific index of work-related asthma. Patients with evidence of work-related asthma on PEF records reacted on inhalation testing to a specific agent inhaled at work and had occupational asthma; patients who did not show evidence of asthma on PEF records (i.e. less than 20 per cent within-day variability) did not react in inhalation tests. The major diagnostic difficulties were patients with evidence of asthma on PEF records without a work relationship, of whom a proportion were eventually shown to have occupational asthma; the commonest reason for this false-negative response was insufficient time away from work for significant improvement to have occurred.

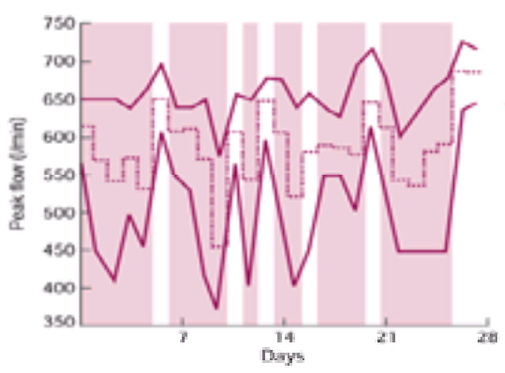

Fig. 2 Serial peak expiratory flow results in a baker sensitive to flour. The best, worst, and average values are plotted for each day. Shaded areas are periods at work; unshaded areas are periods away from work. Peak flows are consistently worse in each work period and improve during each period away from work.

Immunological tests

The presence of specific IgE antibody, identified either by immediate skin test response to a soluble protein extract or a hapten–protein conjugate or by immunoassay in serum (usually radio-allergosorbent testing) is evidence of sensitization to a specific agent. The diagnostic value of a positive test depends upon its predictive value in cases of asthma among those exposed to the specific agent. Specific IgE can be identified in most, if not all, protein causes of occupational asthma, and in a small number of low-molecular-weight chemical causes of asthma, notably complex platinum salts, acid anhydrides, and reactive dyes. No reliable immunological test has been developed for sensitivity to the other important causes of asthma such as isocyanates and colophony. The diagnostic value of a positive test has been formally examined for few of the causes of occupational asthma, and in these cases has been found to be significantly associated with asthma caused by both proteins and low-molecular-weight chemicals inhaled at work.

Inhalation testing

The objective of an inhalation test is to expose the individual under single-blind conditions to the putative cause of their asthma in circumstances that resemble as closely as possible the conditions of exposure at work. The different test methods used depend upon the physical state of the test material, which can be water soluble (most proteins) and inhaled in solution, a volatile organic liquid inhaled as a vapour, or a dust. Any change in lung function in both airways calibre (usually measured as FEV_1 or PEF) and airways responsiveness to inhaled histamine or methacholine, measured as PC_{20}, (concentration of inhaled histamine or methacholine which provokes a 20 per cent fall in FEV_1) is compared with results on appropriate control days. The patterns of airways response provoked by specific inhalation tests have been distinguished by their time of onset and duration. Immediate asthmatic responses occur within minutes of the test exposure and usually resolve spontaneously within 1 to 2 h (Fig. 3). Late asthmatic responses develop 1 h or more after the test exposure and can persist for 24 to 36 h (Fig. 4). Late asthmatic (but usually not immediate) responses are accompanied by an increase in non-specific airways responsiveness 3 h and, less reliably, 24 h after the test inhalation. An immediate response followed by a late response has been called a dual response.

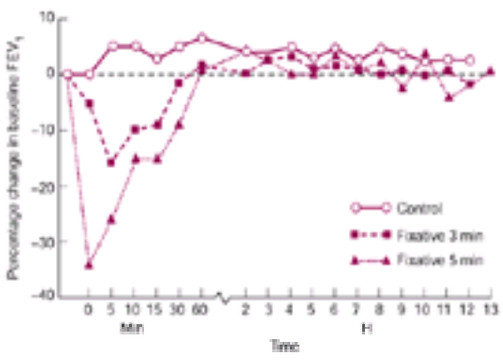

Fig. 3 Immediate asthmatic reactions in a radiographer provoked in inhalation tests of 3 and 5 min with X-ray fixative material, but not by the control test.

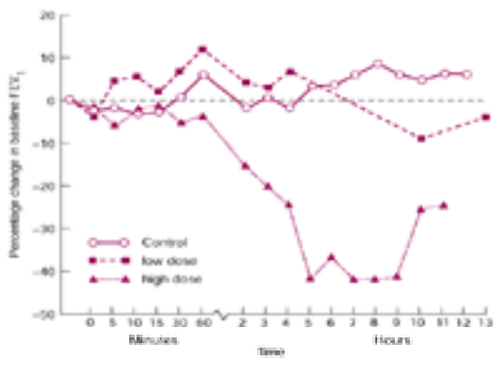

Fig. 4 Late asthmatic reaction in a platinum refiner provoked by exposure to the complex platinum salt ammonium hexachloroplatinate in a concentration of 10 mg (high dose) but not 1 mg (low dose) in 250 g of lactose (the control material).

Inhalation testing allows the identification of specific causes of asthma in individuals exposed to them. Provided that the agent being tested is not a non-specific mucosal irritant and does not provoke an immediate asthmatic response in patients with hyper-responsive airways, for example sulphur dioxide, histamine, or exercise, the provocation of an asthmatic response by an occupational agent implies that it is a cause of asthma. This causal relationship is strengthened if the agent reproducibly provokes a late asthmatic response and increases non-specific airways responsiveness.

There are four major indications for inhalation testing in the diagnosis of occupational asthma:

1. Where the agent considered responsible for causing asthma has not previously been reliably shown to do so.
2. Where an individual with occupational asthma is exposed at work to more than one potential cause that cannot be distinguished by other means.
3. Where asthma is of such severity that further uncontrolled exposure at work is unjustifiable.
4. Where the diagnosis or cause of occupational asthma remains in doubt after other investigations, including serial PEF and immunological tests where applicable, have been completed.

Inhalation tests should be undertaken only for clinical purposes, to provide information important for future management advice. Inhalation tests undertaken solely for medicolegal purposes are not justified.

Differential diagnoses

The diagnosis of occupational asthma requires differentiation:

1. from other causes of similar respiratory symptoms, in particular chronic airflow limitation and hyperventilation;
2. of occupational from non-occupational cause;
3. of asthma initiated by an agent inhaled at work from pre-existing or incidental asthma aggravated by non-specific provocative stimuli encountered at work, such as sulphur dioxide, exercise, and cold air.

Prognosis

Asthma initiated by an agent inhaled at work and caused by toxic damage to the airway epithelium (irritant-induced asthma) or as the outcome of a hypersensitivity response may become chronic and persist for several years, if not indefinitely. Chronic asthma induced by a hypersensitivity response has been reported most frequently in cases caused by low-molecular-weight chemicals such as isocyanates, colophony, plicatic acid from western red cedar, and acid anhydrides. Continuing asthmatic symptoms and airways hyper-responsiveness have been reported in 50 per cent or more of patients several years after avoidance of exposure to the initiating cause. Chronic asthma has also been reported in snow crab process workers in Canada in whom airways responsiveness improved during the first 2 years of

avoidance of exposure but subsequently reached a plateau.

The only important determinant for developing chronic asthma identified to date has been the duration of symptomatic exposure to the initiating cause after the onset of asthma: those who remain exposed to the cause are more likely to develop chronic asthma.

Management

Patients who develop occupational asthma in whom a specific cause is identified should be advised to avoid further exposure to that cause. This seems particularly important where low-molecular-weight chemicals such as isocyanates, plicatic acid, or anhydrides are the cause, as continuing symptomatic exposure to these is particularly associated with the development of chronic asthma and airways hyper-responsiveness.

Avoidance of further exposure may require a change or loss of job that, for social or financial reasons, may not be possible. A change of occupation can be particularly difficult for those who are highly trained, such as experimental scientists whose livelihood depends on their knowledge and experience of working with laboratory animals. Such individuals and others sensitized to biological dusts who are unable to change their job, at least in the short term, should be advised to minimize exposure to the cause of their asthma and to wear adequate respiratory protection, most conveniently laminar-flow equipment, when in contact with the organic dust. In addition, background prophylaxis such as sodium cromoglycate can minimize the risk of the provocation of asthma by indirect allergen contact, such as dust on colleagues' clothing. None the less, it should be emphasized that such measures are temporary, and in the long term means should be sought to avoid exposure to the cause of asthma.

When individuals remain in employment exposed to the cause of their asthma, either directly or indirectly, the effectiveness of relocation or of respiratory protection needs to be monitored. This can be conveniently done by serial self-recordings of PEF to determine whether or not asthma is continuing and, if so, whether it is work related.

Compensation

Statutory compensation in the United Kingdom

Occupational asthma is a prescribed disease for 'employed earners'. The terms of prescription have recently been broadened considerably. They now include asthma caused by exposure to 22 specified groups of agents (listed below) as well as a 'z' category, which specifies 'any other sensitizing agent inhaled at work':

a. isocyanates
b. platinum salts
c. acid anhydride and amine hardening agents
d. fumes arising from the use of rosin as a soldering flux
e. proteolytic enzymes
f. animals including insects and other arthropods or their larval forms used for the purposes of research, education, in laboratories, pest control, or fruit cultivation
g. dusts arising from barley, oats, rye, wheat or maize, or meal or flour made from such grain
h. antibiotics
i. cimetidine
j. wood dusts
k. ispaghula
l. castor bean dust
m. ipecacuanha
n. azodicarbonamide
o. glutaraldehyde
p. persulphate salts or henna arising from their use in the hairdressing trade
q. crustaceans or fish or products arising from these in the food processing industry
r. reactive dyes
s. soya bean
t. tea dust
u. green coffee bean dust
v. fumes from stainless steel welding
w. and
x. any other sensitizing agent inhaled at work.

Byssinosis

In the United Kingdom byssinosis occurs most commonly in cotton mill workers, usually after some 20 to 25 years of exposure to cotton dust. It is probably a response to agents inhaled in the cotton bract and is characterized by chest tightness on the first day of the working week, which usually develops some 3 to 4 h after the start of a work shift. Typically, the chest tightness improves on subsequent working days, despite continuing exposure to cotton dust. The symptoms are often, although not always, accompanied by changes in lung function and the majority of cases of byssinosis have hyper-responsive airways. Cotton dust also provokes acute airway narrowing in about one-third of persons exposed to an extract of cotton bract for the first time; this reaction is probably an important contributory factory in the high turnover in the early months of employment in cotton mills. Whether byssinosis causes long-term respiratory impairment and disability remains controversial. Several studies have failed to find an increase in mortality from respiratory causes, which has been interpreted as suggesting that exposure to cotton dust does not cause chronic lung disease. However, in one survey of a community which included ex-cotton workers, a reduction in FEV_1 of between 2 and 8 per cent was observed in the ex-cotton textile workers and a loss of lung function in those with 15 years' heavy exposure to cotton dust that was equivalent to that observed in light and ex-smokers.

Byssinosis should probably be considered as a form of occupational asthma: the characteristic symptoms are associated with acute reductions in FEV_1 and patients with byssinosis commonly have hyper-responsive airways.

Further reading

Bernstein IL *et al.*, eds (1999). *Asthma in the workplace*, 2nd edn. Marcel Dekker, New York.

Fishwick D and Pickering CAL (1992). Byssinosis—a form of occupational asthma. *Thorax* **47**, 401–3.

McDonald JC, Keynes H, Meredith S (2000). Reported incidence of occupational asthma in the United Kingdom 1989–1997. *Occupational and Environmental Medicine* **57**, 823–9.

Newman Taylor AJ (2000). Asthma. In: McDonald JC, ed. *Epidemiology of work related diseases*, 2nd edn, pp 149–74. BMJ Books, London.

17.5.1 Upper respiratory tract infections

P. Little

Introduction

Acute upper respiratory tract infections (URTIs) include acute pharyngitis/tonsillitis and acute rhinitis. Acute sinusitis, acute otitis media, and influenza also come under the umbrella of infections of the upper respiratory tract. Otitis media and influenza will be discussed elsewhere: this chapter will concentrate on acute pharyngitis/tonsillitis, acute rhinitis, and acute sinusitis.

Acute URTIs are the commonest reason for patients to seek medical advice in the United Kingdom, and nearly all cases are managed in primary care. Respiratory tract infections are also the commonest reason for antibiotics to be prescribed. A serious concern is that the inappropriate use of antibiotics for predominantly self-limiting conditions will foster the development of antibiotic resistance, with the danger that serious infections will become untreatable. Thus it is currently a national priority in the United Kingdom, and should be in other countries, to discourage the use of antibiotics unless there is good evidence of their efficacy. The evidence for the effectiveness of treatments for URTI in this chapter comes from a search of the Cochrane Library databases of systematic reviews and randomized controlled trials.

Pharyngitis/tonsillitis

Clinical presentation

Pharyngitis is caused by both bacterial and viral organisms, and has been somewhat arbitrarily divided into nasopharyngitis (with nasal symptoms, that is to say rhinitis), and pharyngitis or tonsillopharyngitis (without nasal symptoms). Causal organisms include group A beta-haemolytic streptococcus (**GABHS**); adenoviruses; influenza A and B; parainfluenza 1,2,3; Epstein–Barr virus (**EBV**); enteroviruses; *Mycoplasma pneumoniae*; and *Chlamydia pneumoniae*. In addition to a sore throat, pharyngitis is often accompanied by fever, headache, nausea, vomiting, anorexia, and sometimes abdominal pain, with or without enlarged and tender cervical lymph nodes, tonsillar erythema and exudate. Scarlet fever has a characteristic 'scarlatiform' rash caused by GABHS exotoxins. Infectious mononucleosis due to the Epstein–Barr virus may present with or without exudative tonsillitis, cervical or general lymphadenopathy, palatal petechiae, splenomegaly, rhinitis, and cough.

Throat swabs, rapid tests, and clinical algorithms

Antibiotics can be targeted to those patients who have positive throat swabs for group A streptococcus, a positive rapid streptococcal test, or clinical characteristics associated with a positive throat swab (for example, the 'Centor' criteria of fever, tonsillar exudate, anterior cervical adenopathy, and absence of cough). However, the throat swab is not a very good test: in both unselected and clinically selected populations in primary care practice it is neither particularly sensitive nor specific when compared to a rise in Anti Streptolysin O Titres (**ASOT**) or Anti DNAse B titres. A rise in ASOT or Anti DNAase B are better indicators of true infection and predict complications, but are not suitable for clinical diagnosis. The results of throat swabs take days to return to the clinic, and they greatly increase the costs of managing what is mostly a self-limiting condition. Furthermore, evidence suggests that in practice clinicians do not use the results, even of rapid tests, and that the overall accuracy of decision-making is little changed when they are used.

Attempts to derive algorithms or clinical-decision rules based on the throat swab have the same limitations of validity as the throat swab itself. Although clinical scoring methods may provide a crude method of identifying patients at a higher risk of complications (see below), better evidence is needed about how well each clinical sign correlates with proof of infection.

Treatment

Antibiotics for symptoms

The Cochrane review of the efficacy of antibiotics for the treatment of sore throat indicates that antibiotics have modest benefit in reducing the duration of symptoms—by a few hours to half a day—but may have a role in preventing complications (acute otitis media, sinusitis, rheumatic fever, and quinsy). This marginal benefit of antibiotics in resolving symptoms suggests that, for patients who are not unwell systemically, the physician should either not prescribe, or use a delayed prescribing approach, advising the patient to wait for several days before collecting or using their prescription. Both these approaches have been shown in a large randomized controlled trial to be acceptable, change attitudes and behaviour, and not to delay symptom resolution appreciably.

In the context of a likely streptococcal infection, trial evidence suggests that delaying the prescription results in 20 per cent fewer recurrences than the immediate prescriptions of antibiotics, presumably because antibiotics modify local or systemic immune mechanisms. Thus, any marginal symptomatic benefit from an immediate prescription of antibiotics for the current illness must be weighed against the disadvantage that the patient is more likely to suffer symptoms from a recurrence.

Antibiotics to prevent complications

The Cochrane review of antibiotics for treating a sore throat supports the use of antibiotics to prevent complications, but the evidence is limited by both clinical importance and generalizability. For the commoner complications, for example otitis media, 30 children and 140 adults would have to be treated to prevent one case of a self-limiting illness, in other words it is not important clinically. For the rarer complications the evidence is not generalizable; for instance, evidence of efficacy in rheumatic fever is based largely on trials where intramuscular penicillin was used in barracked military personnel after the Second World War. This evidence cannot be sensibly applied to healthier modern settings where the attack rate is much lower, and oral antibiotics are used.

The commonest complication of practical importance to the health service is quinsy; this is relatively uncommon, about 1 in 400 following presentation in primary care with sore throat. The Cochrane systematic review, which demonstrates that antibiotics prevent quinsy, relies on data from patients with tonsillitis who were systemically unwell enough to be admitted to hospital shortly after the Second World War, when the prevalence of quinsy in untreated patients was very high (1:18). Clearly, this data cannot be extrapolated to patients presenting from healthier modern populations who are not systemically unwell, treated with oral antibiotics, and where the prevalence of quinsy is much lower. Quinsy following sore throat is possibly slightly more common (1:60) in those who are unwell, with three out of four Centor criteria, most of whom have fever. Rigorously conducted placebo-controlled trials in patients with these criteria suggest quinsy may be prevented by oral penicillin. However, in routine clinical practice, where compliance is not assessed, the preventive benefit of penicillin is not likely to be 100 per cent, as reported in the trials where compliance was assured. Limited routine data suggests that many patients who develop quinsy after being seen in primary care do this despite being given penicillin. Whether using the clinical Centor criteria is better than the primary care physician's assessment of how unwell patients are is unclear: 20 per cent of those considered 'not to be very unwell systemically' by the physician will still have three out of four of the Centor criteria, and the criteria do not necessarily predict the very few individuals who will develop quinsy. Thus, where the primary care physician judges the patient to be both systemically unwell and/or have three out of

four of the Centor criteria, it would be reasonable to treat with penicillin or at least discuss with patients the likely risks of non-treatment.

Which antibiotic and for how long?

If an antibiotic is to be prescribed, then it is probably preferable to use one of the narrow-spectrum agents (for example, penicillin V) to minimize both side-effects and the risk of resistance. There are arguments for using a large dose, given the variable absorption of penicillin V (e.g. 2 g per day for adults and 1 g per day for children). The length of course is debatable: 10 days may better eradicate streptococcus microbiologically, but the clinical significance of this in affluent western populations is unclear. Longer courses also have the theoretical disadvantage of poorer compliance, and concerns about antibiotic resistance. There is preliminary evidence that a twice-daily dosing of penicillin V, compared to the same amount spread over four doses, results in better compliance, and also better clinical and microbiological outcomes. Amoxicillin and ampicillin are effective against streptococcus but cause a rash in patients with glandular fever: this can be very severe, such that amoxicillin or ampicillin should not be used to treat sore throat, and erythromycin used where penicillin allergy has been documented.

Treatment of patients with rheumatic fever

Patients who have had one attack of rheumatic fever are at a higher risk from new infections since they are likely to develop recurrent attacks of rheumatic fever and complications. Although most of the evidence for the prevention of rheumatic fever comes from old trials in unusual settings, it seems reasonable to treat patients with a past history who are at a high risk of recurrence and secondary complications **,** since what evidence there is suggests penicillin prevents rheumatic fever. (See Chapter 15.10.1 for further discussion of the issues involved.)

Other medical treatments

Treatment with aspirin in children is contraindicated due to the small but avoidable risk of Reye's syndrome. There are several trials of the use of non-steroidal anti-inflammatory drugs (**NSAIDs**) in providing effective relief of pain and fever in tonsillitis and pharyngitis. However, only one trial has made a key clinical comparison of NSAIDs with standard treatment (paracetamol), demonstrating no superiority of NSAIDs. Limited trial data suggests that other useful analgesic adjuncts include caffeine, and benzydamine hydrochloride gargle.

Recurrent attacks

A Cochrane review identified only one published randomized trial of tonsillectomy in children (with serious methodological problems) and none in adults. Further evidence is needed before surgery can be advocated firmly for recurrent tonsillitis. There is preliminary trial evidence for the use of a-streptococci spray, immune stimulants, and pneumococcal vaccination, but further confirmation is required.

Nasal congestion and rhinorrhoea

Nasal symptoms are a common reason for attending the doctor. They may be due to a variety of causes, commonly acute viral infection (common cold), allergic rhinitis and sinusitis, vasomotor rhinitis and rhinitis medicamentosa, and less commonly atrophic rhinitis, hormonal rhinitis, and mechanical/obstructive rhinitis. Colds are responsible for significant morbidity: on average there are 0.4 episodes and 1.2 days of restricted activity per person per year for the common cold.

Acute rhinitis

Symptoms are acute nasal congestion and rhinorrhoea, mild malaise, sneezing, sore throat, variable loss of taste and smell, and usually last from 1 to 2 weeks unless sinusitis is present. Examination reveals a hyperaemic and oedematous mucosa, with or without purulent secretions.

Treatment

Symptomatic treatment

Trial evidence supports the use of both oral and topical decongestants for the symptoms of rhinitis. Intranasal ipratropium bromide is also effective symptomatic treatment, but is only available (in the United Kingdom) on prescription. However, topical decongestants should probably not be used for more than a maximum of 7 days: rhinitis medicamentosa starts to develop at 10 days. Due to their moderate systemic effects, care should be taken with oral decongestants in patients with heart disease and hypertension. Saline drops are commonly advocated, but saline or medicated nose drops have been shown to be ineffective in trials in both children and adults. A Cochrane review suggests that steam may provide some relief of symptoms.

Antibiotics

The use of antibiotics for the common cold has been assessed in a Cochrane systematic review and shown not to be helpful.

Other treatments

Reviews of trials indicate little benefit from antihistamines, nor from zinc lozenges. A Cochrane review of the herb echinacea demonstrated positive results in most studies, but there was not enough evidence to recommend the use of a specific product. Trials in adult volunteers with URTI indicate that intranasal sodium cromoglycate may help to relieve symptoms, and limited evidence suggests that NSAIDs may improve both symptoms and mucociliary clearance. There is also preliminary evidence of promise for the use of immune stimulants in preventing recurrent URTIs.

Acute sinusitis

Diagnosis

Acute sinusitis, usually defined as an infection that lasts for less than 3 weeks, is an uncommon complication of coryzal illness and pharyngitis. There is no absolute standard against which symptoms and signs can be compared for accuracy of diagnosis: aspiration by sinus puncture is probably the definitive investigation, since it indicates the presence of infecting organisms, but for obvious reasons this is rarely performed, and contamination by commensal organisms can occur.

The four-view radiographs show acceptable agreement with aspiration and culture, although only moderate interobserver agreement. The US Agency for Health Care and Policy Research (**AHCPR**) has reviewed the diagnosis and treatment of sinusitis: combining all studies comparing sinus radiographs with sinus puncture demonstrated a sensitivity of 73 per cent and specificity of 80 per cent. A history of purulent nasal discharge, maxillary toothache, purulent secretions on examination, poor response to decongestants, and abnormal illumination of the sinuses, are all predictive of sinusitis defined using four-view radiographs as the standard: four or more symptoms or signs giving a likelihood ratio of a positive test of six. A problem with sinus illumination as a diagnostic tool in primary care is that it performs differently in different settings, probably due to operator sensitivity. There is preliminary evidence comparing symptoms with computed tomography (**CT**) as the 'standard', which is justified since the presence of fluid and total opacification of the sinuses on CT predicts antibiotic response. Purulent rhinorrhoea, purulent secretion in the cavum nasae, a history of 'double sickening' (getting better, then getting worse again), and an erythrocyte sedimentation rate (**ESR**) of greater than 10 are predictive of a CT diagnosis of sinusitis—three of these features giving a likelihood ratio of a positive test of 1.8.

However, using a four-item clinical risk score—of purulent rhinorrhoea with unilateral predominance, local pain with unilateral predominance, bilateral purulent rhinorrhoea, and presence of pus in the nasal cavity—is as sensitive and specific as any other method in predicting the results of sinus puncture. Thus, for acute sinusitis, diagnostic tests are not currently indicated, and until valid near-patient tests are available clinical targeting probably performs as well as any other method.

Treatment

Antibiotics

A Cochrane review of all controlled trials suggests that the absolute benefit for symptom resolution is moderate, and must be balanced against the disadvantages of prescribing antibiotics. Furthermore, this review does not include all the trials from primary care, which show moderate or no effect, and thus both the effectiveness and cost-effectiveness of antibiotic treatment of acute sinusitis in primary care is questionable for most patients.

Other treatments

Preliminary trial evidence shows that decongestants are unlikely to be helpful. There is limited evidence that antihistamines may be helpful for patients with a history of allergic rhinitis who develop sinusitis, and some evidence that proteolytics (e.g. bromelain) and mucolytics may help.

There is mixed trial evidence for the benefit of topical steroids. Although trials of NSAIDs suggest they are helpful, they may not be significantly more effective than paracetamol.

Further reading

Cochrane database of systematic reviews. Cochrane Library, January 2000. (sore throat: antibiotics, tonsillectomy; colds: antihistamines, zinc, echinacea, steam, antibiotics; sinusitis: antibiotics)

Cochrane database of randomised controlled trials. Cochrane Library, January 2000.

Systematic review of diagnosis of sore throat

Del Mar C (1992). Managing sore throat: a literature review. I: Making the diagnosis. *Medical Journal of Australia* 156, 572–5.

Antibiotics and recurrent sore throat, the 'medicalizing' effect of prescribing antibiotics, and the use of delayed prescriptions

Little PS, *et al*. (1997). An open randomised trial of prescribing strategies for sore throat. *British Medical Journal* **314**, 722–7.

Little PS, *et al*. (1997). Reattendance and complications in a randomised trial of prescribing strategies for sore throat: the medicalising effect of prescribing antibiotics. *British Medical Journal* **315**, 350–2.

Use of the 'Centor' criteria to target antibiotic prescribing for sore throat

Zwart S, *et al*. (2000). Penicillin for acute sore throat: randomised double blind trial of seven days versus three days treatment or placebo in adults. *British Medical Journal* **320**, 150–4.

Diagnosis and treatment of sinusitis

US Department of Health and Human Services (1999). *Evidence report/technology assessment number 9: diagnosis and treatment of acute bacterial rhinosinusitis*. AHCPR, Rockville, MD.

Diagnosis of rhinitis

Canadian Rhinitis Symposium (1994). Proceedings of the Canadian Rhinitis Symposium 1994. Assessing and treating rhinitis: a practical guide for Canadian physicians. *Canadian Medical Journal* **15**(Suppl. 4), 1–27.

17.5.2.1 Pneumonia—normal host

John G. Bartlett

Introduction

Pneumonia is an acute or chronic infection involving the pulmonary parenchyma. Most cases are caused by microbial pathogens, including bacteria, viruses, fungi, and parasites. Pneumonia may also refer to inflammation involving the pulmonary parenchyma due to non-microbial causes such as chemical pneumonia. Other modifying terms are used as follows: pneumonia may be acute, subacute, or chronic, depending on the duration of symptoms; it may be described as bronchopneumonia, consolidated (lobar) pneumonia, or interstitial pneumonia based on chest radiography changes; or it may be named after the putative agent, for example pneumococcal pneumonia, mycoplasma pneumonia, *Pneumocystis carinii* pneumonia, etc. Pneumonia is also identified by the place of acquisition—as community-acquired, nursing home-acquired, or hospital-acquired. This chapter will be restricted to community-acquired pneumonia in the adult immunocompetent host.

Aetiology

Although the list of microbes that can cause pneumonia is legion, only a relatively small number are frequent pathogens, for example: *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Mycoplasma pneumoniae*, *Chlamydia pneumoniae*, *Legionella* spp., anaerobic bacteria, and viruses. Less common pathogens are *Moraxella catarrhalis*, *Streptococcus pyogenes*, *Acinetobacter* spp., *Chlamydia psittaci*, *Coxiella burnetii*, *Neisseria meningitidis*, *Staphylococcus aureus*, and enteric Gram-negative rods. In most reported series, each of these generally accounts for less than 1 to 2 per cent of cases. The relative frequencies of different pathogens causing community-acquired pneumonia in two large studies are summarized in Table 1. However, important limitations of these studies should be acknowledged: all the cases in the review conducted by the British Thoracic Society were inpatients, as were the great majority of those reviewed in the meta-analysis. Most studies of pneumonia show that only 20 to 30 per cent of patients are sufficiently sick to require hospitalization. Furthermore, nearly all studies, including those that use extensive diagnostic resources, only identify a likely aetiological agent in 40 to 60 per cent of cases. This suggests that fastidious microbes are under-represented and that many cases of pneumonia may be caused by, as yet, unidentified organisms.

Epidemiology

Pneumonia is the most important infectious disease in terms of morbidity and mortality. It is estimated that in the United States there are four million cases of pneumonia per year (45 000 deaths), and worldwide there are 4400 million cases per year (4 million deaths). In the United States, data suggest that between 20 and 30 per cent of all patients with a diagnosis of pneumonia are hospitalized, and that the mortality rate for this subpopulation is about 14 per cent. The crude death rate from influenza and pneumonia in the United States for 1994 was 31.8 deaths per 100 000 of the population; this represents a 59 per cent increase over the 20.0 deaths per 100 000 recorded in 1979, suggesting that the frequency of lethal pneumonia in the United States is increasing. Those aged 65 or older accounted for 89 per cent of the deaths in 1994, suggesting that increases in longevity account for most of this increase in mortality rate.

Those pathogens associated with specific epidemiological and underlying conditions are summarized in Table 2.

When an aetiological agent is identified, just three microbial agents account for the majority of lethal cases of community-acquired pneumonia. Influenza accounts for an average of 20 000 deaths per year in the United States: the majority involve influenza A, occur in patients over 65 years of age, and most deaths are due to complications of influenza rather than influenza *per se.* The second common cause of lethal pneumonia is pneumococcal pneumonia; risk factors for a fatal outcome include: bacteraemia, advanced age, and concurrent alcoholism. Legionella is the third agent, with associated mortality rates generally reported between 15 and 25 per cent for patients with community-acquired infections.

Nearly all studies show that the risk of death with pneumonia is strongly associated with age extremes. Concurrent conditions that contribute to increased mortality rates include neoplastic disease, hepatic failure, congestive heart failure, cerebrovascular disease, and renal disease.

Pathogenesis

As with nearly all infectious diseases, the probability of disease depends on the virulence of the organism, the inoculum size, and the status of host defences. The normal tracheobronchial tree and lung parenchyma is sterile below the level of the larynx, so that agents of pneumonia must reach this site from external or adjacent sources, usually either by aspiration or inhalation. Organisms may also reach the lung by haematogenous seeding, direct extension from infection in a contiguous structure, or by activation of dormant organisms in the lung. These mechanisms are pathogen-specific, as summarized in Table 3.

Most pneumonias are probably caused by aspiration, which is defined as the abnormal entry of endogenous secretions or exogenous substances into the lower airways. There is a problem here with semantics because most cases of pneumonia are probably due to aspiration as classically described, but 'aspiration pneumonia' probably accounts for only 5 to 10 per cent of cases. The explanation is presumably quantitative, 'aspiration' generally referring to the abnormal entry of relatively large volumes in patients who are so predisposed due to dysphagia or a compromised level of consciousness. The alternative form is presumed to be 'microaspiration', involving the aspiration of very small numbers of microbes, a process that commonly takes place in healthy patients during sleep and with no apparent sequelae.

Clinical features

The classic presentation of pneumonia is of a cough and fever with the variable presence of sputum production, dyspnoea, and pleurisy. Most patients have

constitutional symptoms such as malaise, fatigue, and asthenia, and many also have gastrointestinal symptoms. Although patients with pneumonia usually possess these characteristic clinical features, there can be major differences in presentation based on the host and the aetiological agent, as summarized below.

Pneumococcal pneumonia

Streptococcus pneumoniae is nearly always the most commonly identified pathogen in patients hospitalized with a community-acquired pneumonia. A meta-analysis of 122 reports of community-acquired pneumonia by Fine *et al*. for the period 1966 to 1995 showed that *S. pneumoniae* accounted for 65 per cent of all cases where a microbial pathogen was defined and 66 per cent of all bacteraemic cases (referenced in the Further reading list). Studies using transtracheal aspiration or transthoracic aspiration, methods that avoid the problem of expectorated sputum contamination, show the presence of *S. pneumoniae* in 50 to 80 per cent of cases.

The classic presentation is of a previously healthy adult with an upper respiratory tract infection who then develops a rigor followed by fever, dyspnoea, pleurisy, and a cough that usually becomes productive with a purulent, blood-streaked or 'rusty' sputum. However, many patients show variations in this pattern, including one of a more subtle onset. Moreover, atypical presentations are particularly common in elderly patients. Chest radiography invariably shows an infiltrate, and lobar consolidation specifically suggests this diagnosis (Fig. 1). A pleural effusion is present in about 25 per cent of patients, but only 1 to 2 per cent have an empyema.

Fig. 1 Chest radiograph showing a left lower lobe pneumonia. Blood cultures grew *S. pneumoniae.*

Important observations over the past decade include the declining frequency of cases where this organism is identified and the increasing resistance of *S. pneumoniae* to penicillin and a variety of other antibiotics. The declining frequency is commonly ascribed to a general decline in the quality of microbiological testing currently performed in cases of pneumonia in general, and decreased antibiotic susceptibility is thought to reflect antibiotic abuse. The question still remains as to why, when these drugs were introduced in the 1940s but penicillin resistance was not really encountered until the 1990s, has there been such a long delay in the development of resistance?

Poor prognostic findings in patients with pneumococcal pneumonia include advanced age, bacteraemia, alcoholism, and multiple lobe involvement.

The preferred antibiotics are amoxicillin for oral treatment and ceftriaxone or cefotaxime for parenteral treatment; penicillin-resistant strains may be treated with fluoroquinolones, vancomycin, or linazolide.

Haemophilus influenzae

This organism was originally described in 1892 by Pfeiffer who erroneously thought it was the agent of influenza; it was sometimes referred to as 'Pfeiffer's bacillus'. *H. influenzae* is always the second most common agent (behind *S. pneumoniae*) when an identified bacterial pathogen is found in community-acquired pneumonia. However, the diagnosis is difficult owing to problems with its recognition by direct Gram stain, the fastidious growth requirements of the organism, and with interpretation—even when it is recovered—because it commonly colonizes the upper airways, leading to contamination of expectorated specimens. Type-B *H. influenzae* is a well-established pathogen primarily in infants and young children, but is a relatively rare cause of disease in adults or anyone who has received *H. influenzae* vaccine. *H. influenzae* strains causing pneumonia in adults are usually non-typable.

The clinical features are rather non-specific and include fever, cough, purulent sputum, leucocytosis, and radiographic evidence of pneumonia—usually in a bronchopneumonic pattern, but it may occasionally be lobar. Patients with chronic obstructive lung disease often harbour *H. influenzae* in their lower airways and, allegedly, are prone to pneumonia caused by this organism, although supporting data for the association are not strong. Bacteraemia with *H. influenzae* in adults is infrequent. Most patients simply have a non-specific pneumonia, with *H. influenzae* as the only potential pathogen identified in expectorated sputum.

About 30 to 45 per cent of strains produce b-lactamase so that penicillin and amoxicillin are often ineffective. When *H. influenzae* is suspected or established the preferred agents are second- and third-generation cephalosporins, any combination of a b-lactam–b-lactamase inhibitor, azithromycin, or a fluoroquinolone.

Anaerobic bacteria

These organisms are the dominant components of the microbial flora in the upper airways and average 10^{12}/ml in the gingival crevice. Anaerobes are the major pathogens identified in aspiration pneumonia and its sequelae, lung abscess and empyema. The major pathogens in this group are *Peptostreptcocci* spp., *Bacteroides* spp. (other than *B. fragilis*), *Prevotella* spp., and *Fusobacterium nucleatum*.

The typical patient has gingival crevice disease combined with a predisposition for aspiration that is usually due to a suppressed level of consciousness or dysphasia. The clinical presentation is usually more subtle than that for pneumococcal pneumonia in that the infection evolves over a period of many days, weeks, or even months. Chest radiographs usually show infection in a dependent segment (usually the superior segments of the lower lobes or posterior segments of the upper lobes since these are dependent in the recumbent position), fever, sputum that is often putrid, and evidence of chronic disease with weight loss or anaemia. Putrid discharge is very characteristic and diagnostic of anaerobic bacterial infection (Fig. 2).

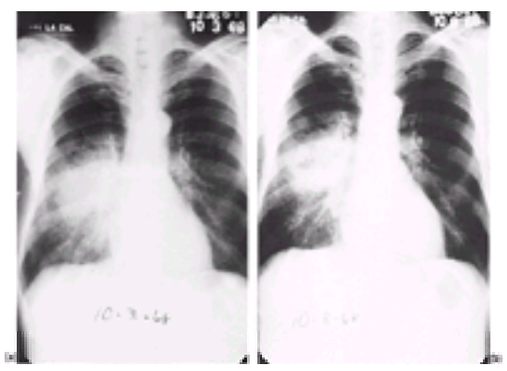

Fig. 2 (a) Consolidated pneumonia in the superior segment of the right lower lobe in a 56-year-old alcoholic. A transtracheal aspirate showed numerous anaerobic bacteria including *Prevotella melaninogenicus* and *Peptostreptococcus* spp. (b) A follow-up radiograph five days later showing cavitation.

Aspiration pneumonia may also be due to chemical insults from gastric acid or other toxins, or may reflect the aspiration of foreign bodies or fluids (victims of drowning). However, the most common sequel to aspiration is bacterial infection involving the anaerobes that normally colonize the upper airways, and such bacteria account for 60 to 80 per cent of cases of aspiration pneumonia, lung abscess, and, in many case series, empyemas. Although the bacterial aetiology can be identified from anaerobic cultures of uncontaminated specimens, these are generally not obtained except in the case of pleural fluid in the presence of empyema; even then the cultures are often falsely negative due to inadequate techniques used to recover oxygen-sensitive bacteria. Thus, the aetiological diagnosis is usually based on the clinical features—where key clues are the chronicity of the infection, associated conditions suggesting aspiration, tissue necrosis with abscess formation, or a bronchopleural fistula leading to empyema and/or putrid discharge.

The preferred drugs are clindamycin or a b-lactam–b-lactamase inhibitor.

Mycoplasma pneumoniae

This organism is one of the most common causes of lower airways' infection in young adults, and it is now more frequently recognized in older adults. The original appellation was 'primary atypical pneumonia', a term applied in the 1930s to a relatively benign form of pneumonia to distinguish it from pneumococcal pneumonia. Early work showed that it was associated with a serum factor that agglutinated erythrocytes in the cold; furthermore, Eaton reported that the infection was transmissible from person to person by intracheal inoculations. Thus, atypical pneumonia, cold-agglutinin pneumonia, and Eaton-agent pneumonia were found to be synonymous.

The typical patient is usually a young adult who experiences a respiratory tract infection accompanied by headache, myalgia, cough, and fever and with a chest radiograph that shows bronchopneumonia. The cough is often non-productive, but when sputum is obtained it is mucoid, shows predominantly mononuclear cells, and no dominant organism. A characteristic feature is the relatively high frequency of extrapulmonary complications such as rash, neurological syndromes (aseptic meningitis, encephalitis, neuropathies), myocarditis, pericarditis, and haemolytic anaemia. The diagnosis should be suspected in those patients with a relatively mild form of pneumonia, particularly in previously healthy young adults.

Most laboratories do not cultivate mycoplasma due to the effort needed to recover the organism, the long time required, and the ease of empirical treatment. Serological tests may be used, but their merits are disputed. Polymerase chain reaction (**PCR**) and other rapid diagnostic tests are under development.

With regard to treatment, the pathogen lacks a cell wall and hence is not susceptible to penicillin, cephalosporins, or other cell-wall active antibiotics. The usual therapeutic agents are macrolides (such as erythromycin, clarithromycin, or azithromycin) or doxycycline; fluoroquinolones are also active.

Chlamydia pneumoniae

This relatively recently identified pathogen is now thought to account for about 5 to 10 per cent of all community-acquired cases of pneumonia, often in young adults who present in a fashion quite similar to that of patients with a mycoplasma pneumonia. *C. pneumoniae* continues to be regarded as a relatively benign agent of pneumonia: most patients have an upper airways' infection with this organism, laryngitis is relatively common, bronchitis is less common, and pneumonitis is an infrequent complication. *C. pneumoniae* plays a role in exacerbations of asthma, and the organism may also be involved in some chronic conditions such as cardiovascular disease.

The diagnosis of chlamydia pneumonia is difficult. The organism is cultivated like a virus using tissue cultures, but few laboratories offer this test. Serology is difficult to interpret; the usual titres for IgM or serial changes with acute and convalescent sera are arbitrary. Like mycoplasma, this is an organism that is often suspected, infrequently proven, and easily treated empirically.

The usual treatment is doxycycline, a macrolide (erythromycin, clarithromycin, or azithromycin), or a fluoroquinolone.

Legionella spp.

Legionnaires' disease was originally described during the American Legion Convention in Philadelphia in 1976, with the putative agent reported the following year. Legionella cause two major syndromes: the pneumonic form or legionnaires' disease, referring to the American Legion Convention epidemic, and a benign influenza-like illness called 'Pontiac fever' in reference to an outbreak in 1967 in Pontiac, Michigan. Although legionnaires' disease is often grouped with mycoplasma and chlamydia infection as being an 'atypical pneumonia', it is a quite different pulmonary infection because it occurs primarily in older adults, is a serious and often lethal form of pneumonia, and most hospital laboratories have diagnostic resources to establish the aetiology. Legionnaires' disease is defined as pneumonia caused by any species of the genera *Legionella*, but the great majority of cases are caused either by *L. pneumophila* (80 to 90 per cent of cases) or *L. mcdadei* (5 to 10 per cent). This disease may be epidemic or sporadic. Epidemics usually occur in buildings, especially hotels and hospitals, and they reflect legionella contamination of the potable water or cooling systems of air conditioners. Predisposing factors include exposure to environmental sources of legionella (there is no patient-to-patient transmission), age over 40 years, smoking, or reduced cell-mediated immune responses as with organ transplantation, cancer chemotherapy, or chronic corticosteroid usage; patients with AIDS do not seem to be uniquely susceptible.

There are no remarkable features of the clinical presentation, except that patients are almost invariably quite sick and may be critically ill. In addition to the typical symptoms of pneumonia with cough and dyspnoea, most present with a profound systemic illness with high fever and myalgias, often with gastrointestinal and neurological symptoms.

The diagnosis can be established with a urinary antigen assay for the detection of *L. pneumophila* serogroup I, culture of respiratory secretions on selective media, serology, or direct fluorescent stain (**DFA**) of sputum. All these tests are quite specific, but none are sufficiently sensitive to exclude the diagnosis when they are negative.

The drugs of choice are fluoroquinolone or a macrolide, or one of these given with rifampicin. However, the mortality rate is generally reported to be 5 to 15 per cent even with proper therapy.

Staphylococcus aureus

Staphylococcus pneumonia was classically described as a complication of influenza during the 1918 epidemic of 'Spanish Flu'. This organism continues to be a potentially important superinfecting pathogen in influenza, and is the most common form of embolic pulmonary infection with injection-drug use and tricuspid valve endocarditis. Staphylococcal pneumonia may be acute or chronic and, despite common impressions to the contrary, pulmonary abscess is a relatively unusual complication. Most patients simply have bronchopneumonia; lobar pneumonia is rare. Those patients with embolic pneumonia show multiple embolic infiltrates that are diffusely scattered and which often cavitate.

The organism can usually be recovered in blood cultures and in respiratory secretions. However, care must be exercised when interpreting respiratory secretion cultures that yield *S. aureus* since this may be a contaminant, and it is particularly common as a contaminant in those patients who have received previous antibiotic treatment.

The treatment should be based on *in vitro* susceptibility tests, usually an antistaphylococcal penicillin (flucloxacillin, oxacillin, or nafcillin), a first-generation cephalosporin (cefazolin), or vancomycin (for methicillin-resistant strains and for patients with severe penicillin allergy).

Gram-negative bacilli

Klebsiella pneumoniae was originally described in 1882 by Friedlander, who believed it was the cause of pneumococcal pneumonia. This organism has continued to be a rare but important cause of community-acquired pneumonia, accounting for about 0.5 to 1.5 per cent of all cases. The classic presentation of 'Friedlander's pneumonia' was a serious pneumonia in an alcoholic patient with a chest radiograph that showed upper lobe involvement and the 'bulging fissure sign' (indicating abscess formation) and sputum that resembled currant jelly. This form of klebsiella pulmonary infection is rarely encountered now, although klebsiella infection is

occasionally implicated in community-acquired pneumonia.

Other Gram-negative bacilli may also cause pneumonia, but the frequency in immunocompetent hosts is very low. A possible exception is *Acinetobacter* spp., which may cause pneumonia in otherwise healthy adults. *Pseudomonas aeruginosa* is a rare pulmonary pathogen, but should be suspected when recovered in respiratory secretions from patients with specific predisposing conditions including structural lung disease, neutropenia, cystic fibrosis, or advanced AIDS. Gram-negative bacteria are commonly encountered in cultures of respiratory secretions, but care must be exercised in interpretation because they are often contaminants reflecting upper airway colonization, especially in patients who have previously received antibiotics.

Treatment should be based on *in vitro* sensitivity tests.

Viruses

Viral infections of the lower airways account for pneumonia in 10 to 15 per cent of inpatients, and probably a substantially larger number of those managed as outpatients. The most frequent pathogens are influenza, parainfluenza, and respiratory syncytial virus (**RSV**). Influenza infections with bronchitis occur in epidemics, but influenza pneumonia is rare. More common in influenza patients with chest radiographs showing infiltrates is bacterial superinfection, most frequently with *S. pneumoniae* or *S. aureus*; less common superinfecting pathogens in this setting are *N. meningitis* and group A streptococcus. The diagnosis of influenza can be made by the combination of an established epidemic and typical influenza symptoms, especially fever. The alternative is to establish the presence of the organism by one of several rapid tests for influenza-A or influenza-B antigen. These rapid tests provide results that are available in about 20 minutes and have a sensitivity of about 70 to 80 per cent.

Clinical features of influenza are generally well known and include cough, fever, purulent sputum, and myalgias. Patients with bacterial superinfections will usually have typical flu-like symptoms, improve, and then deteriorate after 1 to 2 weeks.

Infections involving influenza A may be treated with amantidine or rimantadine; influenza A or B may be treated with these agents, or the neuraminidase inhibitors Relenza (zanamivir) or oseltamivir. If given within 48 h of the onset of symptoms, these anti-influenza drugs reduce the duration of typical symptoms by 1 to 1.5 days and are more effective in seriously ill patients. However, their role in primary influenza pneumonia or in patients with complications of influenza is unknown.

RSV has usually been considered a pathogen in paediatric patients, but is now recognized with increasing frequency in adults, especially the elderly. The diagnosis is easily established with a DFA stain of respiratory secretions for RSV. Ribivirin is active against RSV and is sometimes used by inhalation therapy in children, but there is no treatment with established merit for adult cases.

Laboratory diagnosis

Laboratory tests are used to establish the diagnosis, evaluate the severity, and identify the aetiological agent (Table 4).

Tests to establish the diagnosis and evaluate severity

Chest radiography

The chest radiograph is a pivotal test for the confirmation of pneumonia. It is impossible to make this diagnosis in the absence of a new infiltrate, with four possible exceptions:

1. Although severe dehydration allegedly accounts for a false-negative chest radiograph, animal studies do not support this contention and clinical evidence is weak.
2. Severe neutropenia may give rise to a false-negative radiograph due to the patient's inability to generate an acute inflammatory response. While this is theoretically possible, data suggest that the frequency is low.
3. Some patients have a normal radiograph early in the course of disease, and physicians in the pre-penicillin era claimed they could detect rales before the radiograph showed an infiltrate. This appears to be true, but it applies only to the first 24 h and even then is uncommon.
4. *Pneumocystis carinii* pneumonia may present with a completely normal chest radiograph in 20 to 30 per cent of cases, and probably represents the only form of pneumonia commonly seen nowadays where false-negative routine chest radiographs are common.

Most patients with symptoms of pneumonia and a negative chest radiograph have acute bronchitis, which is generally caused by viral pathogens that do not respond to antibiotic treatment. Thus, the importance of the chest radiograph is in confirming pneumonia, which is a critical feature in avoiding antibiotic abuse. Additional advantages of the chest radiograph is that it provides assistance in identifying the aetiological agent, establishes a baseline for subsequent evaluation, provides prognostic information, and permits the detection of underlying or associated conditions such as a neoplasm.

Other laboratory tests

The most useful additional laboratory tests to determine the severity of illness and need for hospitalization are evaluation of blood oxygenation with pulse oximetry or arterial blood gas determination, blood chemistries (glucose, blood urea nitrogen, and serum sodium levels), and a full blood count. Patients who are hospitalized should generally undergo HIV serology, provided their informed consent is given.

Studies to determine microbial aetiology

Laboratory studies for identifying pulmonary pathogens are among the most controversial issues in pneumonia management. The American Thoracic Society guidelines endorse a nihilistic approach, with the conclusion that microbial studies in pneumonia cases are usually negative, are not cost-effective, and are largely unnecessary since empirical treatment is generally successful. The guidelines from the Infectious Diseases Society of America emphasize the conduct of studies to identify the microbial pathogen(s) in order to promote pathogen-specific treatment and reduce antibiotic abuse, and to identify epidemiologically important organisms such as penicillin-resistant *S. pneumoniae*, *Legionella* spp., influenza, or Hantavirus. It is acknowledged that many recent reports show a low yield of such organisms and fail to document a benefit in terms of cost or outcome; nevertheless, this is attributed to the serious decline in the quality of microbiological standards in recent years.

While empirical therapy is generally advocated for outpatients, routine microbiological testing to identify the aetiological agent of pneumonia is generally only recommended for inpatients. Such tests include blood cultures (from blood samples taken prior to the initiation of antibiotic treatment), which yield a pathogen in about 12 per cent of cases. In general, the only additional test commonly performed to identify an aetiological agent is an expectorated sputum Gram stain and culture. Practice standards for this process include the following:

- The specimen should be obtained by deep cough and should be grossly purulent. It should be collected before antibiotic therapy, preferably in the presence of a physician or nurse.
- The specimen should be promptly transported to the laboratory for processing and incubation within 2 to 5 h.
- A qualified technician should select a purulent portion for Gram stain and culture.
- Cytological screening should be done under low-power magnification (× 100) to determine cellular composition as a contingency for culture.
- The sample should be cultured using standard techniques, with results reported by semiquantitative assessment; most pathogens are recovered in 3 to 4 plus growth, indicating more than five colonies in the second streak.
- Interpretation should be based on the correlation of the Gram stain, semiquantitative culture results, and clinical observations.

The aetiological agent of pneumonia is considered to be clearly established if a likely pulmonary pathogen is recovered from an uncontaminated specimen such as blood culture, pleural fluid, transtracheal aspiration, or transthoracic aspirate. Alternatively, the very presence of a likely pathogen recovered from respiratory secretions is tantamount to a diagnosis; organisms in this category include *Legionella* species, *Mycobacterium tuberculosis*, most viruses other than the herpesvirus group (influenza virus, respiratory syncytial virus, Hantavirus, parainfluenza virus, and adenovirus), and certain fungi (*Histoplasma capsulatum*, *Coccidioides immitis*,

Blastomyces dermatitidis, and *P. carinii*). Organisms such as *S. pneumoniae*, *M. catarrhalis*, *H. influenzae*, and *S. aureus* may be pulmonary pathogens, but interpretation is problematic due to possible contamination with specimens from the upper airway flora. Organisms that virtually never represent pulmonary pathogens include *S. epidermidis*, *Enterococcus* spp., *Neisseria* spp. other than *N. meningitidis*, *Candida* spp., and Gram-positive bacilli other than *Nocardia* spp. or actinomyces.

Transtracheal aspiration was once a popular method of obtaining specimens from the lower airways that avoided upper airway contamination, but the technique requires a skilled clinician and is generally thought to be too invasive for routine use. Transthoracic aspiration has the same limitations, and furthermore seems to give a relatively large number of false-negative results. Bronchoscopy is an attractive method for obtaining respiratory secretions directly from the lower airways; however, the procedure is complicated by contamination with instrument passage through the upper airways so that routine cultures of bronchoscopic aspirates are no better than expectorated sputum. These results may be substantially improved with quantitative cultures of bronchoalveolar lavage specimens or quantitative brush specimens, but many laboratories do not offer this type of analysis, and many pulmonary services cannot provide the samples in a timely fashion.

Most hospital laboratories offer diagnostic tests for detecting the atypical causative agents of pneumonia, for instance of *Legionella* spp. The preferred tests for legionella detection are the urinary antigen assay and culture. Urinary antigen testing is advocated because it is rapid, simply performed, and highly specific; disadvantages include the fact that it only detects *L. pneumophila* serogroup 1, although this accounts for 70 per cent of cases. The alternative test for detecting *Legionella* spp. is culture, which has the advantage of detecting all species of *Legionella*; but the disadvantage is that it requires three days, requires specialized media, and is technically demanding. Most laboratories do not offer diagnostic tests to detect *M. pneumoniae* or *C. pneumoniae*, despite their presumed frequency. This reflects the lack of an acceptable test that is easily performed, provides adequate sensitivity and specificity, and can provide results in a timely fashion.

Treatment

Critical components of initial treatment may include intravenous hydration, oxygenation, and/or intubation and mechanical ventilatory support. Pleural effusions should be sampled to exclude empyema and, when the effusions are large, drained to improve oxygenation. Most authorities feel that expectorants, cough suppressants, and chest physiotherapy are of little value.

Antibiotic therapy

Antibiotics are the mainstay of therapy. Suggestions for specific agents according to microbial pathogen are summarized in Table 5. Most of these are relatively non-controversial and demonstrate the advantage of establishing an aetiological agent. However, as noted above, no pathogen can be detected in 40 to 60 per cent of cases despite arduous attempts to do so; even when an agent is found, this information is usually not available when initial therapeutic decisions are needed. For this reason, most patients are treated empirically, at least initially, whilst microbiological results are pending. Recommendations for empirical treatment are summarized in Table 6. These options are selected on the basis of predicted activity against the most likely pathogens and extensive clinical trials. Nevertheless, this is one of the most controversial areas in medicine based on concerns for antibiotic abuse, increasing resistance of *S. pneumoniae* to many antimicrobials, and sharp geographical differences in the rates of *S. pneumoniae* resistance.

Of major concern in recent years is *S. pneumoniae*, the most common identified agent of pneumonia, because it shows escalating rates of resistance to penicillin, other b-lactams, macrolides, trimethoprim–sulfamethoxazole (**TMP–SMX**), clindamycin, and tetracycline. The only drug that is virtually always active is vancomycin, but the use of this agent in pulmonary infections is discouraged due to a somewhat limited published experience of using vancomycin to treat pneumonia, and concern for the possible promotion of vancomycin-resistant enterococci. Fluoroquinolones with enhanced activity against *S. pneumoniae* include levofloxacin, moxifloxacin, clinafloxacin, trovafloxacin, and gatifloxacin. Most strains of *S. pneumoniae* are susceptible, although some laboratories are reporting increasing resistance to the fluoroquinolones as well as the other antibiotics noted. Multiple large therapeutic trials are underway of patients randomized to receive macrolides, fluoroquinolones, or b-lactams, nearly all of which show therapeutic equivalence. However, there are also reports of anecdotal cases of failure correlating with *in vitro* resistance to b-lactams, macrolides, and fluoroquinolones.

Timing of antibiotic therapy

A large retrospective trial showed that mortality increased with a progressive delay in the time taken to initiate antibiotic therapy after patients had been evaluated. The increase in mortality became statistically significant when the delay exceeded 8 hours. This observation is not surprising since pneumonia is a potentially lethal infection that usually responds to antibiotics, so any delay in treatment would be expected to have deleterious effects. As a result of these observations, many hospitals in the United States are now audited to determine their compliance with antibiotic administration, where necessary, within 8 hours of a patient's admission to the emergency room or admission to the hospital.

Monitoring response to therapy

Subjective responses are usually noted within 3 to 5 days of initiating treatment. Objective parameters to monitor include fever, oxygen saturation, peripheral leucocyte count, and changes on serial chest radiographs. The most carefully documented responses are mortality rates, time to defervesence, and duration of hospital stay. With regard to fever, the temperature in young adults with pneumococcal pneumonia usually drops within 2 to 3 days, patients with bacteraemic pneumococcal pneumonia usually require 6 to 7 days, and elderly patients often respond more slowly. Blood cultures in bacteraemic patients are usually negative within 24 to 48 h. Cultures of sputum will usually show eradication of bacterial pathogens within 24 to 48 h, a major exception being *P. aeruginosa*. Radiographic appearances are slow to improve and much less useful than clinical observations for evaluating response. Follow-up radiographs are generally not recommended, except for patients who are over 40 years of age or are smokers, and the suggested time to do this is 7 to 12 weeks after initiating treatment. Patients who are initially treated with intravenous antibiotics can usually be changed to receive oral agents when they are able to take oral medications and show clinical improvement, such as a temperature below 38 °C for 24 h, a respiratory rate of less than 24/min, and when the PO_2 has returned to normal.

Failure to respond

The major considerations in patients who fail to respond according to the guidelines noted above are:

- The disease is too far advanced at the time of treatment, or treatment is delayed for too long: this is most commonly seen with pneumonia caused by *Streptococcus pneumoniae* or *Legionella* spp.
- The wrong antibiotic was selected: but this is uncommon.
- An inadequate antibiotic dosage is given: this is most common with the aminoglycosides
- The wrong diagnosis is made: for example, a non-infectious disease such as pulmonary embolism with infarction, congestive failure, Wegener's granulomatosis, sarcoidosis, atelectasis, chemical pneumonitis.
- The wrong microbial diagnosis is made.
- The patient may be debilitated, have a severe associated disease, or be immunosuppressed: or there may be other host inadequacies.
- There may be a complicated pneumonia with undrained empyema, metastatic site of infection (meningitis), or bronchial obstruction (foreign body, carcinoma).
- There may be a pulmonary superinfection: most patients in this category respond and then deteriorate with a new fever.

Prognosis

The overall mortality for patients who are hospitalized with community-acquired pneumonia, according to a meta-analysis of 122 reports, is 14 per cent. Risk factors for lethal outcome were well described in the pre-penicillin era, when extremes of age were probably the most important factor. Other risks included bacteraemia, the concentration of bacteria according to quantitative blood cultures in those who were bacteraemic, the extent of changes on chest radiography, alcohol consumption, and the extent of leucocytosis. More recent studies have continued to show that these factors, especially age, are major risk factors for morbidity and mortality. Investigators from the Pneumonia Patient Outcomes Research Team (**PORT**) have developed a prediction rule using a cumulative point score obtained from five categories comprising 19 variables (Table 7). This prediction rule was applied retrospectively to 38 039 inpatients and showed a direct correlation between numerical score and mortality, the authors concluding that these factors predict outcome and can also be used to determine the need for hospitalization.

With regard to specific pathogens, the major agents of community-acquired pneumonia associated with high mortality rates are bacteraemic pneumococcal pneumonia and Legionnaires' disease. Influenza is directly or indirectly implicated in about 20 000 deaths per year in the United States, but primary influenza

pneumonia is relatively rare and most of the influenza-associated deaths are of elderly patients who succumb to complications of influenza. It should also be noted that pneumonia is an extremely common terminal event in patients who die of other conditions, presumably because of aspiration in the terminal stages. Thus, pneumonia is a common autopsy finding when other medical conditions are actually the major cause of death.

Prevention and control

The major preventive measures are influenza and *S. pneumoniae* vaccines. The components selected for the influenza vaccine each year are based on the anticipated strains for the forthcoming season, a prediction that has been quite accurate in 10 of the 11 influenza seasons from 1988 to 99 through 1999 to 2000. Protective efficacy is generally 60 to 70 per cent in the general population when there is a good match between the vaccine strains and the epidemic strain; it is less in elderly vaccinees, but those who develop influenza after vaccination usually have attenuated courses with significant reductions in mortality. The current recommendation is for vaccination between October and November of patients living in the Northern hemisphere. Targeted populations are summarized in Table 8. Amantadine, rimantadine, zanamivir, and oseltamivir may be used to prevent influenza in unvaccinated patients who are so exposed.

The 23-valent vaccine for *S. pneumoniae* contains capsular polysaccharide from 23 serogroups that are responsible for 80 to 85 per cent of bacteraemic pneumococcal infections. Studies of this vaccine suggest a 60 per cent efficacy in preventing bacteraemic pneumococcal infection in immunocompetent adults in the United States, but efficacy is reduced or negligible in immunosuppressed hosts. There is a newly developed 7-valent protein-conjugated pneumococcal vaccine that has the advantage of stimulating a good antibody response in children under 2 years of age, but this vaccine has not been extensively tested in adults.

Controversies

There are probably few diseases in medicine that have been better studied than pneumonia, but with such extraordinary controversy in management guidelines. The major controversies are the utility of microbiology studies, the empirical selection of antibiotics, and the use of pneumococcal vaccine.

Studies of microbial aetiology

Culture and Gram stain of expectorated sputum is the time-honoured method for determining the microbiology of community-acquired pneumonia. Nevertheless, there is substantial controversy regarding the worth of this exercise and a wealth of medical reports with highly divergent findings that simply fuel the debate. In general, the optimal results were achieved in the pre-penicillin era when sputum bacteriology was an art and many patients underwent transthoracic needle aspiration. The reason being that the only available therapy was type-specific antisera for *S. pneumoniae*, thus requiring retrieval of the specific strain. High-quality laboratory technology persisted through the mid-1980s, when the yield of *S. pneumoniae* in expectorated sputum samples for inpatients with community-acquired pneumonia was generally reported at 40 to 70 per cent. The more recent experience is much different, in that the yield of *S. pneumoniae* in expectorated sputum by either Gram stain or culture is only 10 to 20 per cent in most series. Occasional investigators report a higher yield, but they generally employ an antigen-detection method using blood, urine, or respiratory secretions for pneumococcal polysaccharide, which is controversial based on disagreements about specificity. An additional concern is that antigen detection fails to identify other microbial agents of pneumonia, so that culture is generally required to determine *in vitro* sensitivity test results in an era of escalating resistance. Arguments favouring sputum microbiology are the benefits of pathogen-directed therapy that restrains antibiotic abuse, limits side-effects, and reduces cost. In addition, this permits the identification of epidemiologically important organisms, knowledge of which provides the database for empirical therapy recommendations. Arguments against microbiological studies include the facts that: this procedure, as currently performed in most laboratories, shows a low yield; the information is infrequently available when therapeutic decisions are made; empirical treatment usually works (Fig. 3); and, even if a pathogen is recovered, there is no good way to exclude the presence of a copathogen.

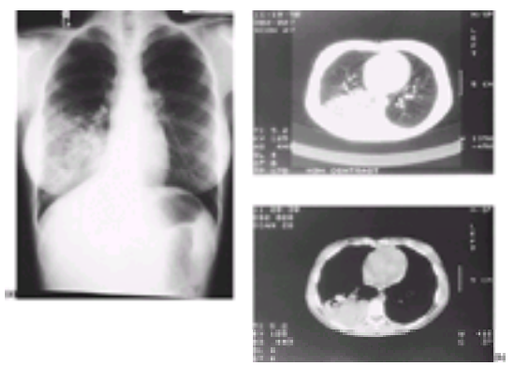

Fig. 3 Chest radiograph (a) and CT scan (b) showing pneumonitis in the right lower lobe of an 18-year-old college student. The patient responded to empirical antibiotic treatment.

Antibiotic selection

Recommendations for the empirical selection of antibiotics include drugs that are active against the major pathogens, including *S. pneumoniae*, *H. influenzae*, and the atypical agents. The favoured classes are tetracyclines, macrolides, and fluoroquinolones. When *S. pneumoniae* is identified or strongly suspected, many authorities conclude that b-lactams are the preferred drugs, despite the fact that b-lactam resistance is 10 to 20 per cent, is substantially higher in some regions, and is increasing in virtually all areas. *S. pneumoniae* is also increasingly resistant to tetracyclines and macrolides, and to a lesser extent, fluoroquinolones. As noted above, large therapeutic trials comparing these four categories of drugs have shown no differences, but there are multiple reports of anecdotal cases with failures involving resistant strains. The controversy regarding empirical therapy is largely explained by geographical differences in resistance patterns, and the concern for fluoroquinolone abuse. With regard to geography, the rates of resistance to the various categories of drugs are highly variable in different parts of the world and within different parts of individual countries. Thus, doxycycline or a macrolide may be rational options in one area, but not another. With regard to the fluoroquinolones, these drugs are active *in vitro* against more than 98 per cent of *S. pneumoniae* strains as well as virtually all atypical organisms, but there is concern that extensive use will result in increasing resistance, a concern that is already being witnessed in some parts of the world. The attractive feature of the fluoroquinolones is the nearly uniform activity against *S. pneumoniae*, the ease of once-daily administration, the availability of both parenteral and oral forms, extensive therapeutic trials to confirm efficacy, and good *in vitro* activity against virtually all treatable pulmonary pathogens.

Pneumococcal vaccine

The polysaccharide vaccine has established merit in young adult African men, but some retrospective analyses and most prospective, randomized, controlled trials have failed to show a significant benefit in terms of reducing the rates of pneumonia, rates of pneumococcal pneumonia, or rates of bacteraemic pneumococcal pneumonia. Most reports of a beneficial effect of vaccination have been based on statistical analyses of the serotype of patients with pneumococcal infection, which demonstrated higher rates of vaccine strains in unvaccinated patients. Even these studies failed to show a benefit in the highest risk group, namely the elderly and the immunosuppressed. The need for a pneumococcal vaccine is widely appreciated due to the extent of morbidity and mortality caused by *S. pneumoniae* and the increasing difficulty caused by resistance in treating these infections. Many authorities feel that the best solution to the dilemma is a better pneumococcal vaccine.

Further reading

Bartlett JG, Mundy L (1995). Community-acquired pneumonia. *New England Journal of Medicine* **333**, 1618–24.

Bartlett JG, *et al*. (2000). Community-acquired pneumonia in adults: guidelines for management. *Clinical Infectious Diseases* **31**, 347–82.

British Thoracic Society (1993). Guidelines for the management of community-acquired pneumonia in adults admitted to hospital. *British Journal of Hospital Medicine* **49**, 346–50.

Chen DK, *et al*. (1999). Decreased susceptibility of *Streptococcus pneumoniae* to fluoroquinolones in Canada. *New England Journal of Medicine* **341**, 233–9.

File TM, *et al*. (1997). A multicenter, randomized study comparing the efficacy and safety of intravenous and/or oral levofloxacin versus ceftriaxone and/or cefuroxime axetil in treatment of adults with community-acquired pneumonia. *Antimicrobial Agents and Chemotherapy* **41**, 1965–72.

Fine MJ, *et al*. (1996). Prognosis and outcomes of patients with community-acquired pneumonia. *Journal of the American Medical Association* **275**, 134–41.

Fine MJ, *et al*. (1997). A prediction rule to identify low-risk patients with community-acquired pneumonia. *New England Journal of Medicine* **336**, 243–50.

Gleason PP, *et al*. (1999). Associations between initial antimicrobial regimens and medical outcomes for elderly patients with pneumonia. *Archives of Internal Medicine* **159**, 2562–72.

Heffelfinger JD, *et al*. (2000). Management of community-acquired pneumonia in the era of pneumococcal resistance. *Archives of Internal Medicine* **160**, 1399–408.

Heffron R. (1939). *Pneumonia: with special reference to pneumococcus lobar pneumonia. A Commonwealth Fund Book*. Copyright 1939. The Commonwealth Fund. (Reprinted by Harvard University Press, Cambridge, MA in 1979.)

Marrie TJ, *et al*. (2000). A controlled trial of a critical pathway for treatment of community acquired pneumonia. *Journal of the American Medical Association* **283**, 749–55.

Meehan TP, *et al*. (1997). Quality of care, process and outcomes in elderly patients with pneumonia. *Journal of the American Medical Association* **278**, 2080–4.

Niederman MS, *et al*. (1993). Guidelines for the initial empiric therapy of community-acquired pneumonia: Proceedings of an American Thoracic Society Consensus Conference. *American Review of Respiratory Disease* **148**, 1418–26.

Pallares R, *et al*. (1995). Resistance to penicillin and cephalosporin and mortality from severe pneumococcal pneumonia in Barcelona, Spain. *New England Journal of Medicine* **333**, 474–80.

Pinner RW, *et al*. (1996). Trends in infectious diseases mortality in the United States. *Journal of the American Medical Association* **275**, 189–93.

17.5.2.2Nosocomial pneumonia

John G. Bartlett

Introduction

Nosocomial pneumonia is generally defined as a new pulmonary infiltrate on chest radiography, combined with evidence of infection expressed as fever, purulent respiratory secretions, and/or leucocytosis, with onset at least 72 h after admission. These infections account for only about 15 per cent of all nosocomial infections, but they are the most frequent, lethal, nosocomial infection. The bacteriology and management are different than that of community-acquired infections of the lung (see Chapter 17.5.2.1).

Aetiology

The microbiology of nosocomial pneumonia shows that Gram-negative bacteria account for 50 to 70 per cent of cases; other common pathogens include *Staph. aureus*, anaerobic bacteria, *H. influenzae*, and *Streptococcus pneumoniae* (Table 1). Legionella accounts for about 4 per cent of all nosocomial pneumonia, but the frequency may be much higher when it is epidemic or endemic within a hospital. Viruses are implicated in 10 to 20 per cent, primarily influenza and respiratory syncytial virus and, in the immunocompromised host, cytomegalovirus. Tuberculosis is rare, but important to remember. Fungi are also rare, with the exception of *Aspergillus* in selected immunocompromised patients.

Epidemiology

Most reports indicate that 0.5 to 1 per cent of all hospitalized patients develop nosocomial pneumonia. The rates in intensive care units are generally reported at 15 to 20 per cent, and among patients who are mechanically ventilated, the rate is 20 to 50 per cent. However, it should be noted that some of these incidence statistics are disputed due to the lack of precision in the diagnosis of nosocomial pneumonia: other processes that may cause pulmonary infiltrates with variable presence of fever, purulent respiratory secretions, and/or leucocytosis include congestive heart failure, pulmonary embolism, atelectasis, adverse drug reactions, pulmonary haemorrhage, and the acute respiratory distress syndrome.

The epidemiology of the pathogens in nosocomial pneumonia is highly variable. Some organisms become endemic, especially in intensive care units; the major pathogens in this setting being *Acinetobacter*, *Serratia*, *Xanthomonas*, *Pseudomonas*, *Enterobacter*, and methicillin-resistant *S. aureus*. Another important nosocomial pathogen is *Legionella*, which may cause outbreaks of Legionnaire's disease in hospitals that can sometimes be traced to water supplies with distribution via air conditioning cooling systems or showerheads. In these cases, the same species and serogroup found in the nosocomial cases should be found in the epidemiologically-linked water supply. Aspergillosis may occur as epidemics among vulnerable patients with compromised cell-mediated immunity, neutropenia, or both. Influenza is highly contagious, and patients with influenza are commonly hospitalized, hence it is now recommended that all patients with suspected influenza have confirmation of this diagnosis by rapid influenza testing, and the preference is for a single room when this is feasible.

Pathogenesis

The relatively high rates of pulmonary infections among patients who are hospitalized reflects (Table 2):

1. clustering of highly vulnerable patients;
2. patients rendered particularly vulnerable by violations of the integrity of the upper airways by intubation or tracheostomy;
3. many patients who are prone to aspiration due to compromised consciousness caused by associated medical conditions and anaesthesia;
4. patients rendered susceptible due to organ transplantation, cancer chemotherapy, and AIDS.w

The dominant pathogen in nosocomial pneumonia are Gram-negative bacteria, which reach the lung by aspiration of gastric contents or by microaspiration of upper airway secretions. The best explanation for this association between bacteriology and pathogenesis is the observation that patients with serious illness commonly have abnormal colonization of the upper airways by Gram-negative bacteria. Thus, throat cultures show that Gram-negative bacteria are found in only 2 to 3 per cent of healthy persons, psychiatric patients, physicians, and medical students, whereas the rate of colonization in patients who are moderately ill is 30 to 40 per cent, and in intensive care units the rate is 60 to 70 per cent. These colonization rates are independent of antibiotic administration. It can also be shown that buccal epithelial cells from patients who are seriously ill have enhanced attachment by Gram-negative bacteria *in vitro*. The frequency of positive throat cultures for Gram-negative bacteria and the number that attach to respiratory cells is directly correlated with the severity of the associated disease. The usual mechanism of Gram-negative bacillary pneumonia in most hospitalized patients is aspiration of these organisms in the upper airways, or aspiration of these organisms from gastric contents after they are swallowed.

Pathogenesis of other organisms is quite different. *Legionella*, tuberculosis, influenza, and *Aspergillus* are inhaled, the usual source being environmental (*Legionella* or *Aspergillus*) or another patient (influenza or tuberculosis).

Clinical features

The classic presentation for pneumonia is cough and fever, usually with purulent respiratory secretions. The diagnosis of pneumonia requires the demonstration of a pulmonary infiltrate on chest radiography. These same symptoms may be present in patients with acute bronchitis, which is virtually always a viral infection that does not merit antibacterial treatment. A notable exception is patients who have violation of the airways with endotracheal tubes or tracheostomies who may have 'febrile tracheobronchitis' due to bacterial infection, most frequently at the tip of the tube, the site of the cuff, or the site of insertion. As noted previously, many patients who satisfy the definition for nosocomial pneumonia based on a pulmonary infiltrate accompanied by fever and purulent respiratory secretions have alternative diagnoses when studied by reliable microbiological techniques using bronchoscopy with quantitative cultures of a bronchial-protected brush or bronchoalveolar lavage (BAL).

Laboratory diagnosis

Tests to establish diagnosis and evaluate severity

The chest radiograph is critical for the confirmation of pneumonia. Major causes of false-negative radiographs in the presence of nosocomial pneumonia are severe neutropenia and pneumonia caused by *P. carinii*. CT scans may reveal infiltrates that are not present on plain films, but it is not clear that this distinguishes a group that requires antibiotic treatment. Thus, the chest radiograph is generally viewed as adequately sensitive for detection of nosocomial pneumonia.

It is important to monitor blood gases to determine severity of illness and to monitor respiratory support.

Studies to determine microbial aetiology

Blood cultures are positive in 2 to 6 per cent of patients with nosocomial pneumonia and clearly identify the causative agent. Some patients will have empyemas, and thoracentesis is necessary for both diagnosis and treatment. Again, this represents an uncontaminated source for culture, providing definitive evidence of the responsible pathogen. Empyema is an infrequent complication of nosocomial pneumonia, excepting in patients who have undergone thoracotomy who often have an empyema as a complication of chest tube placement.

Legionella, *M. tuberculosis,* and respiratory viruses (influenza, parainfluenza, and respiratory syncytial virus) represent definitive pathogens when recovered in respiratory specimens since these organisms do not colonize the normal respiratory tract.

The majority of patients with nosocomial pneumonia do not have bacteraemia, empyema, or the pathogens that do not colonize the normal airway. In these cases, the physician usually must rely on routine bacterial cultures of respiratory secretions or resort to invasive diagnostic tests using bronchoscopy with quantitative cultures of BAL specimens or of the protected brush. Multiple studies have tested the validity of these techniques for distinguishing contaminants and pathogens. The results are somewhat variable, but often dependent on the precision of methodology. The use of these techniques has also resulted in substantial controversy in the management of nosocomial pneumonia, especially in intensive care units where the stakes are high due to high mortality rates. Arguments in favour of invasive diagnostic studies with bronchoscopy are the facts that the technology is well studied, about one-half of patients with suspected pneumonia have negative results and antibiotics can be avoided in this population, and the clear definition of pathogens permits pathogen-specific antibiotic treatment. Others argue that the invasive methods are unrealistic or unnecessary because many hospital laboratories do not provide an adequate microbiology service, patients with the characteristic clinical features will be treated with antibiotics regardless of the bronchoscopy results, or because of the perception that semiquantitative cultures of tracheal aspirates are cheap, easy, and provide information that is equally valid.

Regardless of the method to obtain respiratory secretions for microbiology studies, it is usually beneficial to examine the specimen cytologically. Cultures should be reported with either quantitative or semiquantitative results. For quantitative results, the usual threshold for significance with the protected brush is 10^3/ml, and for BAL specimens it is usually 10^3 or 10^4/ml. With semiquantitative techniques, moderate or heavy growth usually indicates 'significant concentrations.' The major pathogens are summarized in Table 1: *S. epidermidis*, diphtheroids, *H. parainfluenza*, *Entero-coccus*, and a-haemolytic *Streptococcus* are generally regarded as contaminants, regardless of concentrations. Anaerobic bacteria are frequently neglected pulmonary pathogens, but it is difficult to obtain specimens valid for anaerobic cultures, and many laboratories struggle with anaerobic microbiology even when the right specimens are obtained. The diagnosis of anaerobic pneumonia should be suspected when Gram stains show mixed bacteria, especially when there are morphotypes suggesting anaerobes, and specimens obtained by tracheal aspirate or bronchoscopy should be examined for these organisms. Putrid drainage always indicates anaerobic infection.

Treatment

The major management issues are antibiotic selection and respiratory support. The optimal method to select antibiotics is to base this decision on results of Gram stains and cultures (Table 3). When empiric decisions are necessary in seriously ill patients, agents are directed against Gram-negative bacteria, and the favoured drugs in this context are ceftazidime, cefepime, imipenem/meropenem, piperacillin/piperacillin–tazobactam, ticarcillin/ticarcillin–sulbactam, or ciprofloxacin. For *S. aureus*, vancomycin is often added on the basis of Gram stain results or the perceived need to cover this pathogen. Treatment for *P. aeruginosa*, the predominant Gram-negative bacillus in nosocomial pneumonia in intensive care units, should be based on *in vitro* sensitivity tests. Anaerobic bacteria are well treated with imipenem/meropenem or any b-lactam–b-lactamase inhibitor; clindamycin can be used if these organisms are suspected and the alternatives are not used for other pathogens. The role of aminoglycosides in pulmonary infections involving Gram-negative bacilli is controversial because these agents appear to have a marginal activity at the concentrations achieved in the lung and at the pH of pulmonary secretions. Nevertheless, they appear to work well in animal experiments. The recommendation is that if they are used, they are always combined with a second agent that should have activity against Gram-negative bacilli as summarized above, and that there is assurance of adequate levels with either once-daily dosing or with monitoring of levels with thrice-daily dosing.

It should be emphasized that cultures of respiratory secretions obtained after the inception of antibiotic treatment have substantially reduced validity. This observation emphasizes the importance of pretreatment cultures and caution with therapeutic decisions based on post-treatment cultures other than those of blood and plural fluid.

Outcome

Nosocomial pneumonia is associated with a mortality rate reported at 8 to 20 per cent for all cases. The mortality rate for infections acquired in the intensive care unit is 20 to 40 per cent, with a mean of 25 per cent. In the latter group the attributable mortality is 30 to 33 per cent, meaning that associated conditions are the major factors in causing death.

Prevention

The frequency of nosocomial pneumonia and high mortality rate, especially in intensive care units, has prompted extensive studies of prevention. The methods that have withstood the test of time and have proven meritorious are the use of the semiupright position to reduce the risk of aspiration, and hand washing between patients to prevent transmission of nosocomial pathogens.

The concern for patient positioning is based on marker studies showing that stomach contents are displaced to the lower respiratory tract with high frequency in patients in the recumbent position, and this can be easily corrected by use of an upright or semiupright position. The assumption is that nosocomial pneumonia is frequently due to bacteria that reside in the stomach as a result of oral colonization. With regard to hand washing, this is a time-honoured method to reduce nosocomial infection that is commonly neglected by hospital personnel. It appears to be particularly important in the transmission of *S. aureus*, and is often important in organisms that are endemic or epidemic within hospital units such as *Acinetobacter*, *Serratia*, *Xanthomonas*, *Pseudomonas*, and *Enterobacter*.

A common practice in intensive care units is prophylaxis to prevent peptic ulceration of the stomach, but neutralization of gastric acid eliminates the gastric barrier, the defence mechanism that prevents colonization of the stomach by bacteria, including Gram-negative bacteria from the upper airways. As a result, sucralfate is commonly advocated in place of H2 agonists or antacids. Another approach to dealing with colonization of the upper airways and stomach is 'selective decontamination' to interrupt the cycle of colonization of the colon by Gram-negative bacteria followed by colonization of the upper airways by the same organisms. The goal of selective decontamination is elimination or reduction in Gram-negative bacteria in the gastrointestinal tract with antibiotics that also preserve the anaerobic bacteria in the flora, since these are largely responsible for population control in the colon. Drugs that are commonly used are oral preparations of polymyxin, aminoglycosides, poorly absorbed fluoroquinolones, aztreonam, trimethoprim–sulfamethoxazole, or cephalosporins. These drugs are given orally with the expectation that they will have a major impact on the colonic flora, and they are sometimes also incorporated into paste formulations for application to the upper airways as well. Extensive trials with selective decontamination show that they achieve a substantial reduction in nosocomial pneumonia, but do not seem to influence mortality due to nosocomial pneumonia. Major concerns are:

1. the failure to reduce mortality rates;
2. excessive costs of the regimens; and
3. the perception of antibiotic abuse with encouragement of resistance.

Topical antibiotics have also been tested for utility in prophylaxis. The method is installation of drugs (usually polymyxin or aminoglycosides) through tracheostomies, endotracheal tubes, or by aerosolization. Extensive therapeutic trials with this tactic have shown that they are sometimes successful in interrupting epidemics due to susceptible bacteria, especially *P. aeruginosa*, but mortality rates have generally remained unchanged, and there is concern about the evolution of resistant bacteria. Topical antibiotics are generally not recommended, excepting for some patients with cystic fibrosis.

Interruption of epidemics involving *Legionella* and *Aspergillus* requires different tactics because these organisms are inhaled. For *Legionella* and *Aspergillus*, the goal is to eliminate the environmental source. Influenza is transmitted from person-to-person, so the goal is to eliminate this type of contact, which must include removal of health-care workers with influenza from jobs that require patient contact. All hospital personnel should have influenza vaccine as a method to protect patients, and

hospital personnel with jobs that require patient contact must be furloughed if they have suspected or established influenza.

Further reading

American Thoracic Society (1995). Hospital-acquired pneumonia in adults: diagnosis, assessment of severity, initial antimicrobial therapy and preventative strategies. A consensus statement. *American Journal of Respiratory and Critical Care Medicine* **153**, 1711–25. [Guidelines for managing nosocomial pneumonia, including diagnostic studies and antibiotic selection.]

Fagon J-Y, Chastre J, Wolff M, *et al*. (2000). Invasive and noninvasive strategies for management of suspected ventilator-associated pneumonia. *Annals of Internal Medicine* **132**, 621–30. [A multicenter, randomized trial; there were significantly fewer deaths at 14 days in the group that had invasive bacteriological methods.]

Fagon JY, Chastre J, Hance AJ, *et al*. (1988). Detection of nosocomial lung infection in ventilated patients: use of a protected specimen brush and quantitative culture techniques in 147 patients. *American Review of Respiratory Disease* **138**, 110–16. [Over half of the patients who had standard criteria for ventilator-associated pneumonia actually had an alternative diagnosis.]

Fagon JY, Chastre J, Vuagnat A, Trouillet JL, Novara A, Gibert C (1996). Nosocomial pneumonia and mortality among patients in intensive care units. *Journal of the American Medical Association* **275**, 866–9. [A review of mortality data for nosocomial pneumonia.]

Johanson WG Jr, Woods DE, Chaudhuri T (1979). Association of respiratory tract colonization with adherence of gram-negative bacilli to epithelial cells. *Journal of Infectious Diseases* **139**, 667–73. [Gram-negative bacilli stuck to buccal epithelial cells better if the source was seriously ill. This is thought to be the explanation for high rates of pharyngeal colonization in these patients.]

Johanson WG, Pierce AK, Sanford JP (1969). Changing pharyngeal bacterial flora of hospitalized patients: emergence of gram-negative bacilli. *New England Journal of Medicine* **281**, 1137–40. [Colonization of the pharynx by Gram-negative bacilli correlates with severity of associated illness. These strains are thought to be the precursor of Gram-negative bacillary pneumonia due to aspiration.]

Morehead RS, Pinto SJ (2000). Ventilator-associated pneumonia. *Archives of Internal Medicine* **160**, 1926–36. [A review of the topic. The authors claim that invasive diagnostic methods require extensive resources, but may save costs. Prompt use of antibiotics is stressed. Prevention is reviewed without endorsement of selective decontamination.]

17.5.2.3 Pulmonary complications of HIV infection

Mark J. Rosen

Since 1995, the relentless increase in death rates from AIDS in developed nations has reversed due to the use of highly active antiretroviral therapy (HAART), and there is now hope that AIDS may not be uniformly fatal. Nevertheless, thousands of people are infected with HIV each year, and thousands more still develop opportunistic infections and HIV-associated neoplasms because they do not know they are infected, have limited access to these treatments, choose not to use them, or the treatments are unsuccessful. Antiretroviral therapy is too expensive to be available to millions of people with AIDS worldwide, and the AIDS epidemic will be devastating in Africa, Asia, and South America.

Spectrum of pulmonary disorders

Lung diseases have been important causes of illness and death in AIDS since the beginning of the epidemic. The first cases of AIDS were described in homosexual men in Los Angeles who had *Pneumocystis carinii* pneumonia, without a known reason for immunodeficiency. The incidence of *Pneumocystis carinii* pneumonia is declining because of the widespread use of prophylaxis and HAART, but it is still the most common AIDS-defining opportunistic infection in the United States and Western Europe. Despite the importance of *Pneumocystis carinii* pneumonia as an AIDS-defining disorder, clinicians should not assume that most HIV-infected persons with pulmonary disease have *Pneumocystis carinii* pneumonia, because there is a wide range of pulmonary infections, neoplasms, and inflammatory disorders associated with HIV infection (Table 1). These range from mild abnormalities in pulmonary function unaccompanied by respiratory symptoms to fatal opportunistic infections.

Early investigations of the types of pulmonary disorders that occur in patients with HIV infection were limited by analysis of data from single sites, and by restricting the scope of analysis to patients who had an opportunistic infection or neoplasm. This approach systematically underestimated the incidence and importance of pulmonary disorders that occur in early HIV infection. In the Pulmonary Complications of HIV Infection Study, 1353 subjects were followed prospectively in six American cities to determine the prevalence, incidence, and types of lung diseases that occur in persons in selected HIV transmission categories. After 18 months of follow-up, the most frequent respiratory diagnoses in the HIV-seropositive subjects were upper respiratory infection (33.4 per cent), acute bronchitis (16 per cent), and acute sinusitis (5.3 per cent). Although these disorders were also common in a control group of HIV-seronegative gay men and injecting drug users, bronchitis and sinusitis were reported more frequently in the HIV-infected subjects. The types and frequencies of lower respiratory tract disorders diagnosed in the first 18 months of follow-up are listed in Table 2. After 64 months of follow-up, the incidence of bacterial pneumonia remained higher than that of *Pneumocystis carinii* pneumonia.

Pulmonary tuberculosis and fungal infections are also common in patients with AIDS, and are addressed in other sections of this textbook. Non-infectious disorders may also involve the lung in HIV-infected individuals: these include neoplasms (especially Kaposi's sarcoma), non-specific and lymphocytic interstitial pneumonitis, primary pulmonary hypertension, and bronchiolitis obliterans organizing pneumonia. These disorders may vary in severity from asymptomatic to life threatening.

Diagnostic approach

When a homosexual man or an injecting drug user presents with a respiratory illness, most clinicians will suspect an HIV-related disorder. However, the incidence and prevalence of HIV infection is increasing rapidly among patients who acquired HIV infection by heterosexual contact, and should be considered in all patients with pneumonia, especially in communities where the prevalence of HIV infection is high. Conversely, many patients with known HIV infection have disorders that are common in the community.

The initial approach to patients with suspected HIV-related pulmonary disorders is the same as for any other patient: the clinician will take a careful history, perform a physical examination, and determine whether or not to perform diagnostic tests. In HIV-infected patients, careful consideration of risk factors for specific pulmonary diseases and the chest radiograph are especially important in formulating a differential diagnosis.

Patients with HIV-associated pulmonary disorders typically present with non-specific symptoms such as cough, dyspnoea, sputum production, and wheezing. Pulmonary symptoms are more common in HIV-infected people than in the general population; in the Pulmonary Complications of HIV Infection Study, 15 per cent of 1171 HIV-infected persons who did not have a diagnosed pulmonary disorder or an AIDS diagnosis complained of cough and dyspnoea at the time of enrolment. Both dyspnoea and cough were more prevalent in injecting drug users than others, and were strongly associated with cigarette smoking. Respiratory symptoms may be unrelated to the HIV infection, and disorders that are common in the general population, such as asthma or bronchitis, must be considered.

Risk factors for specific pulmonary disorders

The risk of developing particular HIV-associated disorders is strongly influenced by the patient's immune status, demographic characteristics, current or prior place of residence, use of antiretroviral agents, and prophylaxis against common HIV-associated infections. Genetic factors are undoubtedly important, but are less precisely defined.

Immune status

The severity of immunosuppression probably has the strongest influence on the risk of specific AIDS-associated disorders, and the CD4+ lymphocyte count is the best surrogate marker for immune function and predictor of the probability of developing specific diseases. In early HIV infection, when the immune system is not severely compromised, respiratory disorders are similar to those that affect the general population, while opportunistic infections occur only with severe immunodeficiency. The CD4+ lymphocyte count, together with quantitative assay of plasma HIV RNA (a surrogate marker of HIV replication), provides the best predictor for disease progression and death.

The association between the CD4+ lymphocyte count and the risk of developing specific diseases was explored in a survey of more than 18 000 HIV-infected subjects who received care in 10 American cities as part of the Adult/Adolescent Spectrum of HIV Disease surveillance system (Table 3). Common problems like sinusitis, bronchitis, and pharyngitis occurred at all strata of CD4+ cell counts. With lower counts, different pulmonary infections occurred with increasing frequency. Bacterial pneumonia and pulmonary tuberculosis occur with relatively intact CD4+ lymphocyte counts, while opportunistic infections like *Pneumocystis carinii* pneumonia and disseminated *Mycobacterium avium* complex and fungi are likely to occur only with severe immunosuppression. Conversely, if the CD4+ lymphocyte count has a sustained increase to more than 200 cells/μl following HAART, the risks of developing *Pneumocystis carinii* pneumonia and other opportunistic infections decline to the point that prophylaxis can be discontinued. Hence knowing the CD4+ count is very helpful in formulating a differential diagnosis in a patient with known or suspected HIV infection. Respiratory problems like sinusitis and bronchitis may occur at any level of CD4+, and bacterial pneumonia and tuberculosis often occur before AIDS-defining opportunistic infections and neoplasms. Declining immune function increases the risk for all HIV-associated respiratory disease, except perhaps for mild upper respiratory tract infections.

Demographic factors

The demographic characteristics of those infected with HIV influence the incidence of specific pulmonary disorders, and the changing demographics of HIV infection in the United States and Europe are accompanied by a changing spectrum of disease. Injecting drug users are at increased risk of developing bacterial pneumonia and pulmonary tuberculosis, and HIV-infected drug users are at especially high risk.

Race and ethnicity may also influence the risk of developing bacterial pneumonia and tuberculosis, but these associations are confounded by differences in access to health care, the higher prevalence of tuberculosis in minority communities, and disproportionately high numbers of injecting drug users who are Black or Hispanic. Nevertheless, the risk of tuberculosis is higher in Blacks and Hispanics than Whites, while Whites have a higher risk of HIV-associated malignancies and cytomegalovirus disease.

Residence

Geographical considerations influence the types of diseases that HIV-infected individuals are at risk of developing. In the United States, the incidence of HIV-associated tuberculosis is highest in the northeast. The high incidence of *Pneumocystis carinii* pneumonia in the United States and Europe contrasts sharply with most regions in Africa, where it is much less common. It is still unknown whether inherited differences in susceptibility to *Pneumocystis carinii* or environmental factors account for the lower incidence of *Pneumocystis carinii* pneumonia in Africa.

The geographical distribution of endemic fungi is also a strong determinant of risk of those infections; disseminated histoplasmosis and coccidioidomycosis are common in patients with AIDS who live in endemic areas. These infections may also occur as reactivation disease after HIV-infected persons move to other areas and develop immunocompromise. For example, cases of disseminated histoplasmosis were reported in New York in patients with AIDS who had relocated from Puerto Rico years before.

Antiretroviral therapy

The availability of potent drugs that inhibit HIV replication led to the development of combination regimens that can accomplish prolonged and near complete suppression of viral replication, with improvement in immune function and clinical outcomes. The standard care of those infected with HIV now includes use of combination antiretroviral therapy with a goal of suppressing HIV replication below the limits of detection, guided by monitoring plasma HIV RNA levels and CD4+ lymphocyte counts. Since the introduction of HAART into clinical practice, death rates from AIDS and AIDS-related opportunistic disorders have fallen dramatically. In a study of 1255 patients with severe immune deficiency, defined by at least one CD4+ count of less than 100 cells/µl, morbidity, mortality, and the incidence of opportunistic infections declined as the use of combination antiretroviral therapy including protease inhibitors increased. These trends were not explained by patient characteristics (sex, age, race/ethnicity, HIV risk) or increasing use of prophylaxis against opportunistic infections. Rather, improving outcomes were attributable to the intensity of antiretroviral therapy, especially with protease inhibitors.

The new antiretroviral agents, and especially the protease inhibitors, interact with many other drugs, and clinicians must investigate possible drug interactions when prescribing new medications for patients taking antiretrovirals. These drug interactions have particular importance in the treatment of tuberculosis in patients with HIV infection, discussed in Chapter 7.10.21 and Chapter 7.11.22.

Pneumocystis carinii prophylaxis

Even before the introduction of HAART, anti-*Pneumocystis* prophylaxis had a profound impact on the spectrum of HIV-associated pulmonary diseases by reducing the incidence and mortality rate due to that infection, and the prognosis for survival after developing immunosuppression or an AIDS-defining diagnosis also improved, largely attributable to that treatment. However, mortality from non-tuberculous mycobacterioses, cytomegalovirus disease, bacterial infections, non-Hodgkin's lymphoma, tuberculosis, and other opportunistic infections increased.

Chest radiography

The chest radiograph is extremely useful in evaluating symptomatic HIV-infected patients, because different radiographic patterns are associated with specific disorders. These are shown in Table 4. A normal radiograph may be seen with common disorders like bronchitis or asthma, but also with *Pneumocystis carinii* pneumonia or tuberculosis. Patients with normal films who are suspected of having one of these infections should proceed to CT scan or pulmonary function tests with carbon monoxide diffusing capacity: if either of these studies is normal, the diagnosis of *Pneumocystis carinii* pneumonia is extremely unlikely. An abnormal film will prompt a diagnostic evaluation based on the pattern of the abnormality, and influenced by the clinician's perception of the relative risk of different disorders. For example, focal infiltrates are usually caused by bacterial pneumonia or tuberculosis, but fungal infection is possible in endemic areas. Diffuse opacities are usually associated with *Pneumocystis carinii* pneumonia, but may also be seen with other disorders. Diagnostic algorithms based on the radiographic findings in patients with suspected HIV-associated pulmonary disorders are summarized in Table 5.

Specific pulmonary diseases

Bacterial pneumonia

HIV infection impairs humoral as well as cellular immunity, increasing the risk of developing bacterial infections, including sinusitis and pneumonia. lthough a first episode of bacterial pneumonia usually occurs before the diagnosis of AIDS, the risk of developing pneumonia increases as the CD4+ lymphocyte count declines. Drug users are at higher risk than other groups, and neutropenia is an independent risk factor. Patients who use trimethoprim-sulfamethoxazole to prevent *Pneumocystis carinii* pneumonia have a reduced risk of developing bacterial pneumonia.

Streptococcus pneumoniae, *Haemophilus influenzae*, and *Staphylococcus aureus* are the most common bacterial pathogens, but disturbing patterns of antimicrobial resistance are emerging, and new aetiologies of pneumonia are recognized more commonly, especially in patients with advanced HIV disease. For example, pneumonia caused by *Pseudomonas aeruginosa,* which was rarely diagnosed in the 1980s, is a relatively common pathogen in severely immunocompromised patients, typically with CD4+ counts of less than 50 cells/µl, and especially following another opportunistic infection. Although pneumonia due to 'atypical' pathogens like *Mycoplasma*, *Chlamydia*, and *Legionella* is reported, these pathogens are relatively uncommon. *Rhodococcus equii*, an aerobic Gram-positive acid-fast bacillus, may cause focal consolidation, endobronchial disease, and cavitation, usually in patients with advanced HIV disease. *Nocardia asteroides* may cause nodules, consolidation, cavitation, pleural effusions, empyema, and intrathoracic lymphadenopathy in patients with HIV infection.

Bacterial pneumonia usually presents with the same symptoms as those who are HIV seronegative, with fever, chills, productive cough, and localized areas of consolidation on chest radiography. While this clinical picture strongly suggests bacterial pneumonia, it may also occur with tuberculosis and fungal infection, and patients with bacterial pneumonia sometimes have diffuse pulmonary opacities that resemble *Pneumocystis carinii* pneumonia.

The approach to diagnosis and treatment is the same as that for HIV-seronegative patients (see section 17.5.2.1). Polyvalent pneumococcal vaccine is recommended for all HIV-infected people, although they may not mount an adequate antibody response. The vaccine against *Haemophilus influenzae* type b is not used in patients with HIV infection because most *Haemophilus* infections are with strains not covered by this vaccine. Although annual influenza vaccine is recommended, there are no data indicating that patients with HIV infection are at increased risk of contracting influenza, or that the illness is more severe than in the general population.

Pneumocystis carinii pneumonia

Clinical features

Pneumocystis carinii pneumonia typically presents with gradually increasing dyspnoea and cough over weeks, but sometimes as an acute illness with rapid deterioration over a few days. The chest radiograph usually has diffuse ground glass opacities, which strongly suggests the diagnosis, but sometimes shows nodular opacities, lobar consolidation, or a normal film. All of these radiographic patterns may also occur with other infections and neoplasms. Cystic abnormalities and

spontaneous pneumothoraces in patients with known or suspected HIV infection are usually caused by *Pneumocystis carinii* pneumonia.

The diagnosis of *Pneumocystis carinii* pneumonia may be supported by adjunctive tests. *Pneumocystis carinii* pneumonia is unlikely in a patient who had a CD4+ cell count above 200 cells/µl in the preceding 2 months in the absence of other HIV-associated symptoms. Approximately 90 per cent of patients with *Pneumocystis carinii* pneumonia have an elevated serum lactic dehydrogenase, but this may occur with other pulmonary diseases. Oxygen desaturation with exercise is a relatively sensitive and specific test in patients suspected to have *Pneumocystis carinii* pneumonia, but is not diagnostic. Gallium-67 and indium-111 lung scans are highly sensitive indicators of *Pneumocystis carinii* pneumonia, but isotope uptake also occurs in other pulmonary infections, so they are seldom useful in a diagnostic evaluation.

Microbiological diagnosis

Pneumocystis carinii cannot yet be cultured *in vitro*, so the diagnosis of *Pneumocystis carinii* pneumonia can be confirmed only by demonstrating organisms in a lung-derived specimen. The least invasive diagnostic test is the analysis of sputum induced with 3 per cent saline delivered by ultrasonic nebulization. Using modified Giemsa, methenamine silver, or immunofluorescent staining, and depending on the experience of the laboratory, *Pneumocystis carinii* can be identified in up to 80 per cent of cases. Other pathogens, particularly *Mycobacterium tuberculosis* and fungi, may also be found using appropriate staining and culture techniques.

It is controversial whether to routinely proceed with fibreoptic bronchoscopy to confirm the diagnosis of *Pneumocystis carinii* pneumonia in patients suspected of having the disease, but who have non-diagnostic sputum specimens. Some prefer to treat empirically for *Pneumocystis carinii* pneumonia, and establish a diagnosis only if there is no clinical response within 5 days. Proponents of this approach hold that a presumptive diagnosis of *Pneumocystis carinii* pneumonia is usually accurate, and that the procedure carries unnecessary inconvenience, risk, discomfort to patients, and expense. Proponents of early bronchoscopy maintain that using an empirical approach will subject many patients to treatment and its attendant toxicity for a disease that they do not have, and non-responders may be too ill to undergo bronchoscopy after several days of inappropriate therapy. Furthermore, coinfection with other pathogens is common, may not be diagnosed in patients treated empirically, and adjunctive corticosteroid therapy may transiently improve symptoms in patients with other pulmonary disorders, contributing to the emergence of other opportunistic infections such as aspergillosis and cytomegalovirus.

Patients with suspected *Pneumocystis carinii* pneumonia who have non-diagnostic sputum studies should have fibreoptic bronchoscopy with bronchoalveolar lavage as the next procedure in the diagnostic evaluation. The complication rate is very low, and the yield is over 90 per cent in most centres. This yield is optimized by performing lavage in more than one lobe, and is higher in the upper lobes than the lower. All lavage specimens should be examined for the presence of acid-fast bacilli, fungi, and viral cellular inclusions, since patients with suspected *Pneumocystis carinii* pneumonia may have another infection, or may be coinfected with other pathogens. The role of bronchoalveolar lavage in the diagnosis of bacterial pneumonia in HIV-infected persons is not established. Some clinicians also perform bronchoscopic lung biopsies routinely during bronchoscopy, as this procedure increases the diagnostic yield for *Pneumocystis carinii* and other disorders. Others reserve biopsy for patients who have no diagnosis after bronchoalveolar lavage. Bronchoscopic biopsy is contraindicated in the presence of bleeding disorders, and the high risk of pneumothorax usually precludes biopsy in patients undergoing mechanical ventilation. Diagnosing *Pneumocystis carinii* pneumonia by video-assisted thoracoscopy or an open procedure is rarely necessary.

Respiratory failure caused by *Pneumocystis carinii* pneumonia

Despite HAART, antipneumocystis prophylaxis, and declining mortality rates from *Pneumocystis carinii* pneumonia, it is still the most common cause of respiratory failure and admission to intensive care units in patients with AIDS. When treatment of *Pneumocystis carinii* pneumonia is postponed or ineffective, a clinical syndrome develops that resembles the acute respiratory distress syndrome, with severe hypoxaemia, intrapulmonary shunt, reduced pulmonary compliance, and the appearance of diffuse radiographic opacities. Just as severe *Pneumocystis carinii* pneumonia clinically resembles acute respiratory distress syndrome, the supportive treatment, including intubation, mechanical ventilation, and application of positive end-expiratory pressure, is similar. Continuous positive airway pressure delivered by mask may improve gas exchange without endotracheal intubation, but its usefulness is limited in patients with severe disease. However, it may afford the patient and physician more time to consider whether mechanical ventilation is desirable.

As changes in therapy have modified the prognosis, three distinct 'eras' of critical care for patients with *Pneumocystis carinii* pneumonia and respiratory failure can be identified. Initially, the outlook for survival was dismal, at around 15 per cent, and in some centres admissions to intensive care units declined because physicians did not recommend aggressive interventions, and patients were more likely to decline them. After 1986, mortality rates seemed to improve to around 50 per cent, attributed to selection of patients with a better prognosis and to the benefits of adjunctive corticosteroid therapy. We are now in a third 'era' of outcomes, when patients who require mechanical ventilation for *Pneumocystis carinii* pneumonia again have a high mortality rate since they are more likely to have failed prophylaxis, anti-*Pneumocystis* treatment, or adjunctive corticosteroid therapy, and therefore are expected to have a poor prognosis. Recent studies show that when respiratory failure follows several days of appropriate therapy for *Pneumocystis carinii* pneumonia, the probability of survival is only around 20 per cent.

The prospects for long-term survival following *Pneumocystis carinii* pneumonia and respiratory failure are more hopeful than earlier in the epidemic. Prolonged survival is probably related to the selection of patients with a better prognosis for mechanical ventilation, and to the use of HAART and prophylaxis against subsequent infections.

Airway diseases

For unknown reasons, patients with advanced HIV infection may have chronic bronchitis and bronchiectasis, even if they do not smoke. These patients usually have severe immunodeficiency, with CD4+ counts of less than 100 cells/µl. Standard antimicrobial agents are usually effective, but symptoms are likely to recur, especially when *Pseudomonas aeruginosa* is isolated from the sputum. The role and efficacy of bronchodilators and antiinflammatory agents in HIV-associated airway diseases have not been studied systematically. HIV infection also predisposes to early emphysema, possibly related to enhanced pulmonary cytotoxic T-lymphocyte activity.

Kaposi's sarcoma

Kaposi's sarcoma, probably caused by human herpevirus 8, is the commonest malignancy in people with HIV infection, and the skin is the major site of involvement. This virus infects many healthy adults, and may be isolated commonly in saliva, prostatic tissue, and semen. It is probably transmitted by sexual contact, and causes disease when activated by HIV-associated immunosuppression. This hypothesis helps to explain why Kaposi's sarcoma is much more common among HIV-infected gay men than in other transmission groups.

Kaposi's sarcoma may involve many organs, including the lung. Patients with pulmonary Kaposi's sarcoma usually have obvious mucocutaneous lesions, but the lung may be the only site of disease in up to 15 per cent of cases. Involvement of the airways, parenchyma, pleura, and intrathoracic lymph nodes causes a diverse range of symptoms and radiographic findings. The majority of patients with pulmonary Kaposi's sarcoma diagnosed antemortem have cough, dyspnoea, and fever.

Kaposi's sarcoma lesions in the airways do not usually cause symptoms, but sometimes lead to obstruction or haemoptysis. The finding of typical lesions on inspection of the airways is usually considered diagnostic. Histological diagnosis may be difficult because the yield of forceps biopsy is low, and some authors believe that forceps biopsy of Kaposi's sarcoma lesions places the patient at significant risk of bleeding, but this is controversial.

Parenchymal involvement with Kaposi's sarcoma is suggested by bronchial wall thickening, nodules, Kerley B lines, and coexisting pleural effusions, especially in patients with cutaneous disease. Patients may be bronchoscoped to determine whether diffuse radiographic opacities are caused by Kaposi's sarcoma or an opportunistic infection. The yield of bronchoscopic lung biopsies in the diagnosis of Kaposi's sarcoma is low, and even open lung biopsy is non-diagnostic in approximately 10 per cent of cases because of the focal distribution of lesions. The diagnosis of pulmonary parenchymal Kaposi's sarcoma is therefore usually inferred in patients with cutaneous disease, chest radiographs that suggest this disorder, visual confirmation of airway lesions, and no evidence of opportunistic infection on bronchoalveolar lavage or bronchoscopic lung biopsy. Patients with parenchymal opacities who have typical lesions in the airways and no identified pulmonary infection are assumed to have parenchymal Kaposi's sarcoma.

When Kaposi's sarcoma involves the pleura, effusions are usually exudative and sanguinous, but cytological examination is non-diagnostic. Closed pleural biopsy is rarely positive due to the focal nature of pleural lesions and predominant involvement of the visceral rather than parietal pleura. Since establishing a diagnosis usually necessitates a thoracoscopic or open pleural biopsy, the presence of pleural involvement with Kaposi's sarcoma is usually inferred in a patient with cutaneous

disease and a serosanguinous effusion without a reasonable alternative explanation.

Lymphoma

Non-Hodgkin's B-cell lymphoma is associated with HIV infection. Although pulmonary involvement is usually clinically innocuous, the lung is a common site of extranodal disease. Even in patients with an established diagnosis of lymphoma, lung involvement is usually a late feature of disease. If symptoms occur, they are usually late in the course of HIV disease, and simulate common opportunistic infections, presenting with lobar consolidation, nodules, reticular opacities, and masses. Intrathoracic lymphoma usually presents with lymphadenopathy, pleural effusions, or pleural thickening. Airway involvement may cause atelectasis, and pulmonary involvement may be seen in the form of nodules or consolidation. The diagnosis is established by lung or lymph node biopsy, or by cytological analysis of pleural fluid.

Carcinoma of the lung

A few series report an increased incidence of lung cancer in those infected with HIV, and that these cancers are more aggressive, diagnosed at a more advanced stage, and are associated with shorter survival than lung cancer in HIV-negative individuals. The link between HIV infection and lung cancer is supported by genomic differences between lung cancers in patients with and without HIV infection. However, lung cancer is still very rare compared with opportunistic infections and Kaposi's sarcoma.

Further reading

Afessa B, Green B (2000). Bacterial pneumonia in hospitalized patients with HIV infection. *Chest* **117**, 1017– 22. [Shows that *Pseudomonas aeruginosa* is a common cause of pneumonia in persons with AIDS.]

Batungwanayo J *et al.* (1994). Pulmonary disease associated with human immunodeficiency virus in Kigali, Rwanda: a fiberoptic bronchoscopic study of 111 cases of undetermined etiology. *Amercian Review of Respiratory and Critical Care Medicine* **149**,1591– 6. [Shows that *Pneumocystis carinii* pneumonia is uncommon in Central Africa, even among patients with severe immunocompromise.]

Burack JH *et al.* (1994). Microbiology of community-acquired bacterial pneumonia in persons with and at risk for human immunodeficiency virus type 1 infection. *Archives of Internal Medicine* **154**, 2589– 96. [Documents the pathogens that cause bacterial pneumonia in HIV infection, showing that resistant organisms are becoming more common.]

Carpenter CCJ *et al.* (2000). Antiretroviral therapy in adults. Updated recommendations of the International AIDS Society—USA Panel. *Journal of the American Medical Association* **283**, 381– 90. [Latest recommendations.]

Centers for Disease Control (1980). *Pneumocystis pneumonia*—Los Angeles. *Morbidity and Mortality Weekly Review* **30**, 250– 2. [First report of patients with AIDS.]

Centers for Disease Control and Prevention (1999). 1999 USPHS/IDSA guidelines for the prevention of opportunistic infections in persons infected with human immunodeficiency virus: U.S. Public Health Service (USPHS) and Infectious Diseases Society of America (IDSA). *Morbidity and Mortality Weekly Review* **48**, (No RR-10), 5– 6. [Latest recommendations.]

Diaz PT *et al.* (2000). Increased susceptibility to pulmonary emphysema among HIV-seropositive smokers. *Annals of Internal Medicine* **132**, 369– 72. [Demonstrates that emphysema is common in HIV-infected persons, and possibly related to enhanced cytotoxic T-lymphocyte activity.]

Fauci AS (1999). The AIDS Epidemic. Considerations for the 21st century. *New England Journal of Medicine* **341**, 1046–50.

Hanson DL *et al.* (1995). Distribution of CD4+ T lymphocytes at diagnosis of acquired immunodeficiency syndrome-defining and other human immunodeficiency virus-related illnesses. *Archives of Internal Medicine* **155**, 1537– 42.

Hirschtick RE *et al.* (1995). Bacterial pneumonia in patients infected with human immunodeficiency virus. *New England Journal of Medicine* **333**, 845– 51. [Elucidates the risk factors for developing bacterial pneumonia and the pathogens identified.]

Hoover DR *et al.* (1993). Clinical manifestations of AIDS in the era of *Pneumocystis* prophylaxis. *New England Journal of Medicine* **329**, 922– 1926. [As the incidence of *Pneumocystis carinii* pneumonia declined due to prophylaxis, the rates of other opportunistic infections increased.]

Huang L *et al.* (1996). Presentation of AIDS-related Kaposi's sarcoma diagnosed by bronchoscopy. *American Journal of Respiratory and Critical Care Medicine* **153**, 1385– 90. [Describes the clinical features of Kaposi's sarcoma involving the lung.]

Jones JL *et al.* (1999). Surveillance for AIDS-defining opportunistic illnesses, 1992– 1997. *Morbidity and Mortality Weekly Review* **48**, (No SS-2), 1– 22. [Documents the declining incidence of *Pneumocystis carinii* pneumonia since 1994.]

Markowitz N *et al.* (1997). Incidence of tuberculosis in the United States among HIV-infected persons. *Annals of Internal Medicine* **126**,123– 32. [Relates the risk of tuberculosis with CD4+ lymphocyte count and place of residence.]

Mocroft A *et al.* (1998). The incidence of AIDS-defining illnesses in 4883 patients with human immunodeficiency virus infection. *Archives of Internal Medicine* **158**, 491– 7. [These studies relate the risk of developing specific HIV-associated disorders with CD4+ lymphocyte counts.]

Palella FJ *et al.* (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine* **338**, 853– 60. [Improving survival and reduced incidence of opportunistic infections are related to the increased use of HAART.]

Rosen MJ (1999). Intensive care of patients with HIV infection. *Seminars in Respiratory Infection* **14**, 366– 71. [Comprehensive review.]

Salzman SH (1999). Bronchoscopic techniques for the diagnosis of pulmonary complications HIV infection. *Seminars in Respiratory Infections* **14**, 318– 26. [Detailed review.]

Verghese A *et al.* (1994). Bacterial bronchitis and bronchiectasis in human immunodeficiency virus infection. *Archives of Internal Medicine* **154**, 2086– 90. [Describes the clinical features of airways disease in patients with advanced HIV infection.]

Wallace JM *et al.* (1993). Respiratory illness in persons with human immunodeficiency virus infection. *American Review of Respiratory Diseases* **148**, 1523– 9. [Eighteen-month follow-up of a multicenter prospective study of HIV-seropositive persons, showing that bacterial pneumonia is more common than *Pneumocystis carinii* pneumonia.]

White DA, Matthay RA (1989). Noninfectious pulmonary complication of infection with the human immunodeficiency virus. *American Review of Respiratory Diseases* **140**, 1763– 87. [Comprehensive review.]

Wistuba II *et al.* (1998). Comparison of molecular changes in lung cancers in HIV-positive and HIV-indeterminate subjects. *Journal of the American Medical Association* **279**, 1554– 9. [Supports the hypothesis that HIV infection increases the risk of developing lung cancer.]

17.6 Chronic obstructive pulmonary disease

William MacNee

Introduction

Chronic obstructive pulmonary disease (**COPD**) produces considerable morbidity and mortality: it is the sixth commonest case of death worldwide, and set to become the third commonest cause by the year 2020. It is a slowly progressive condition characterized by airflow limitation that is largely irreversible.

Definitions

The group of conditions characterized by airways obstruction that is incompletely reversible have no universally accepted definition. The term COPD has become accepted in recent years, but is not truly a disease, rather a group of diseases. A major problem in defining COPD is the difficulty in differentiating this condition from asthma, particularly the persistent airways obstruction of older chronic asthma sufferers that is often difficult or even impossible to distinguish clinically from that in COPD, although a history of heavy cigarette smoking, evidence of emphysema by imaging techniques, decreased diffusing capacity for carbon monoxide, and chronic hypoxaemia favour a diagnosis of COPD.

Chronic bronchitis is defined clinically as the presence of a chronic productive cough on most days for 3 months, in each of two consecutive years, in a patient in whom other causes of chronic cough have been excluded. Chronic bronchitis can be classified into three forms: simple bronchitis, defined as mucus hypersecretion; chronic or recurrent mucopurulent bronchitis in the presence of persistent or intermittent mucopurulent sputum; and chronic obstructive bronchitis when chronic sputum production is associated with airflow obstruction.

Emphysema is defined as abnormal, permanent enlargement of the distal airspaces, distal to the terminal bronchioles, accompanied by destruction of their walls and without obvious fibrosis. As with chronic bronchitis the definition of emphysema does not require the presence of airflow obstruction. Thus emphysema is defined pathologically.

Bronchiolitis or small airways disease results from inflammation, squamous cell metaplasia, and/or fibrosis in airways less than 2 mm in diameter. These changes are amongst the earliest to appear in cigarette smokers but are difficult to detect by physiological measurements. Although relatively little is known of the natural history of this condition, it is considered to contribute increasingly, as it progresses, to the airways obstruction of COPD.

The relative contribution made by airway abnormalities or distal airspace enlargement to the airways obstruction in an individual patient with COPD is difficult to determine. Thus in the United States the term COPD was introduced in the early 1960s to describe patients with largely irreversible airways obstruction, due to a combination of airways disease and emphysema, without defining the contribution of these conditions to the airways obstruction.

In their statement on the Standards for Diagnosis and Care of Patients with COPD, the American Thoracic Society defined COPD as 'a disease state characterized by the presence of airflow obstruction due to chronic bronchitis or emphysema; the airflow obstruction is generally progressive, may be accompanied by airway reactivity, and may be partially reversible'. The European Respiratory Society has adopted a similar definition—'a disorder characterized by reduced maximum expiratory flow and slow forced emptying of the lungs, features which do not change markedly over several months'. The definition produced by the British Thoracic Society is similar— 'a slowly progressive disorder characterized by airways obstruction (reduced FEV_1 and FEV_1/VC ratio), which does not change markedly over several months. Most of the lung function impairment is fixed, although some reversibility can be produced by bronchodilator (or other) therapy'. Recently, the global initiative on obstructive lung disease introduced the concept of COPD as an inflammatory disease by suggesting in their definition that COPD was characterized by 'an abnormal inflammatory response in the lungs to inhaled particles or gases'.

In clinical practice a diagnosis of COPD is usually associated with:

1. a history of chronic progressive symptoms (cough, wheeze, and/or breathlessness), with little variation;
2. usually a cigarette smoking history of more than 20 pack years (1 pack year is 20 cigarettes per day for 1 year); and
3. objective evidence of airways obstruction, ideally by spirometry, that does not return to normal with treatment.

The term COPD excludes a number of specific causes of chronic airways obstruction, such as cystic fibrosis, bronchiectasis, and bronchiolitis obliterans (for example associated with lung transplantation or chemical inhalation). The differentiation of COPD from asthma remains a problem, particularly as a large proportion of patients with COPD show some reversibility of their airflow obstruction with bronchodilators.

Aetiology

Chronic mucus hypersecretion

Population studies of respiratory symptoms show a much higher prevalence of cough and sputum among smokers than among non-smokers. A survey in urban and rural populations in the United Kingdom found that a history of chronic bronchitis was present in 17.6 per cent of males aged 55 to 64 years who were heavy smokers, in 0.9 per cent of light smokers, 4.4 per cent of ex-smokers, and was absent in non-smokers. Stopping smoking produces cessation of the sputum production in 90 per cent of cases. Pipe and cigar smokers have a much lower prevalence of chronic bronchitis and less impairment of respiratory function, possibly reflecting lower rates of smoke inhalation.

The 'British hypothesis' suggested that chronic airflow obstruction resulted from the development of chronic mucus hypersecretion as a result of recurrent bronchial infection. This hypothesis was tested in the landmark studies of Fletcher and Peto in working men in London followed up between 1961 and 1969, which showed that smoking accelerated the decline in forced expiratory volume in 1 s (**FEV1**), but failed to show a correlation between the degree of mucus hypersecretion and an accelerated decline in FEV_1 or mortality. By contrast, mortality was strongly related to the development of low FEV_1.

More recent data, from a study of 15 000 adults from the general population in Copenhagen followed up between 1976 and 1994, suggested that mucus hypersecretion was not such an innocent phenomenon, since it was associated with increased risk of hospital admission and accelerated decline in FEV_1. Moreover, as the FEV_1 decreases, the association between mucus hypersecretion and mortality becomes stronger. Differences in the degree of airflow obstruction between the populations in these two studies may explain the different findings.

Cigarette smoking

Cigarette smoking is the single most important identifiable aetiological factor in COPD. However, only 10 to 20 per cent of smokers develop clinically significant COPD, whilst approximately half will never develop a clinically significant physiological deficit. In general, the greater the total tobacco exposure, the greater the risk of developing COPD. However, for any exposure there are clearly individual variations in susceptibility to the effects of tobacco smoke (Fig. 1). Although smoking is the dominant risk factor, COPD does occur in non-smokers, such as in patients with a_1-antitrypsin deficiency.

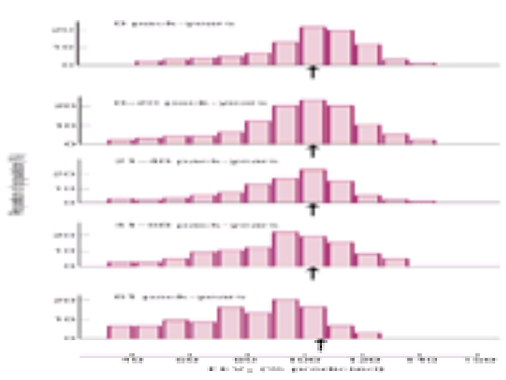

Fig. 1 Effect of increasing cigarette smoke consumption on FEV_1. Although the mean FEV_1 falls as smoking consumption increases, there is a wide variation in this effect, suggesting variable susceptibility to the effects of tobacco smoke. (Adapted from Burrows B *et al*., 1977, *American Journal of Respiratory and Critical Care Medicine* **115**, 195–205.)

The most important evidence linking smoking and mortality from bronchitis comes from a study of 40 000 medical practitioners in the United Kingdom who recorded their smoking habits. In male doctors, mortality from chronic bronchitis fell between 1953 and 1967 by 24 per cent, compared with a fall of only 4 per cent in other men in the United Kingdom of the same age. This difference was attributed to the decrease in smoking in doctors compared with an overall increase in smoking in the general population.

Passive smoking

There is a trend to an increased relative risk of the development of chronic airflow obstruction from passive smoking, but not powerful enough to demonstrate statistical significance. Cumulative lifetime exposure to environmental tobacco smoke during childhood is associated with significantly lower peak levels of FEV_1 in adulthood. Maternal smoking is associated with low birth weight, and smoking by either parent is associated with an increased incidence of respiratory illnesses in the first 3 years of life.

Air pollution

The introduction of the clean air acts (1956, 1965) led to a reduction in smoke and sulphur dioxide levels during the 1960s, which produced less discernible peaks of pollution related to morbidity and mortality, compared with the 1950s. More recent studies show an association between respiratory symptoms in patients with airways disease, general practitioner consultations, and hospital admissions in patients with airways diseases at levels of particulate air pollution below 100 $\mu g/m^3$, levels that are currently experienced in many urban areas in Europe and the United Kingdom. Furthermore, levels of particulate air pollution are associated with deaths from all

causes, particularly cardiorespiratory deaths.

Although there have been associations between exacerbations of airways diseases and photochemical air pollutants, such as nitrogen dioxide and ozone, this association has been largely confined to patients with asthma. There are a few longitudinal studies on the effects of air pollution on decline in lung function, but the data are conflicting. Indoor air pollution, for example as a result of the use of biomass fuel for cooking is associated with the development of COPD in low income countries, particularly in women.

Protease inhibitor deficiency

a_1-Antitrypsin or a_1-protease inhibitor is a polymorphic glycoprotein that is responsible for most of the antiprotease activity in the serum. Laurell and Eriksson in 1963 were the first to describe the association between a_1-antitrypsin deficiency and the development of early onset emphysema, and that the abnormality was transmitted as an autosomal recessive gene. Since the discovery of the deficiency, over 75 biochemical variants have been described relating to their electrophoretic properties, giving rise to the phase inhibitor (**Pi**) nomenclature. The commonest allele in all populations is PiM and the most common genotype is PiMM, which occurs in around 86 per cent of the United Kingdom population. PiMZ and PiMS are the next two commonest genotypes and are associated with a_1-protease inhibitor levels of between 50 and 75 per cent of mean levels of PiMM subjects, as is the much less common PiSS type. The homozygous PiZZ type, in which serum levels are 10 to 20 per cent of the average normal value, is the strongest genetic risk factor for the development of emphysema. The most important other type is PiSZ, where basal levels are 35 to 50 per cent of normal values. A few rare variants that result in complete functional absence of a_1-protease inhibitor account for the remainder of the severely deficient patients.

In the United States, screening of adult blood donors identified a 1 in 2700 prevalence of PiZZ subjects, of which most had normal spirometry. Around 1 in 5000 children in the United Kingdom are born with the homozygous deficiency (PiZZ). However, the number of subjects identified with disease is much less than predicted from the known prevalence of the deficiency. Therefore it is by no means inevitable that all individuals with a homozygous deficiency develop respiratory disease. Indeed a few PiZZ individuals live beyond their sixth decade and escape the development of progressive airways obstruction. Prospective follow-up of PiZZ subjects has shown a greatly accelerated decline in FEV_1, but with large variations among individuals. There is a clear interaction with cigarette smoking, but this cannot entirely account for the variation in the decline in FEV_1 observed. Life expectancy of subjects deficient in a_1-protease inhibitor is significantly reduced, especially if they smoke.

Occupation

It is generally accepted that there is a causal link between occupational dust exposure and the development of mucus hypersecretion. Cigarette smoke has been a confounding feature since the prevalence of smoking remains disproportionately high in many workers who are exposed to dust. Longitudinal studies on workforces exposed to dusts show an association with dust exposure and a more rapid decline in FEV_1 and increased mortality.

The accumulating evidence for an association between coal dust exposure and the development of COPD led recently to the establishment of COPD as a disease that is considered for compensation in miners in the United Kingdom. A small, but significant effect of exposure to welding fumes on the development of COPD was shown in a study of shipyard workers. Workers exposed to cadmium have an increased risk of emphysema.

Chronic bronchopulmonary infection

Studies in the 1960s and 1970s in men with chronic bronchitis demonstrated that prophylactic antibiotics to prevent recurrent infective exacerbations did not slow the decline in lung function. However, acute bronchopulmonary infection was associated with an acute decline in lung function that may persist for several weeks, but which usually recovers completely.

Cough and sputum production between the ages of 20 and 36 years is more commonly reported in those with a history of chest illness in childhood. The association between childhood respiratory illness and ventilatory impairment in adulthood is probably multifactorial. Several factors such as low economic status, greater exposure to passive smoking, poor diet and housing, and residence in areas of high pollution may all contribute to this finding.

Growth and nutrition

Several recent studies have suggested that mortality from chronic respiratory diseases and adult ventilatory function correlate inversely with birth weight and weight at 1 year of age. Thus, impaired growth *in utero* may be a risk factor for the development of chronic respiratory diseases.

One study of British adults has shown that there is a correlation between consumption of fresh fruit in the diet and ventilatory function, a relationship that held both in smokers and in those who had never smoked. Dietary factors, particularly a low intake of vitamin C and low plasma levels of ascorbic acid, were related to a diagnosis of bronchitis in the United States National Health and Nutrition Examination Survey.

Atopy and airway hyperresponsiveness

In the 1960s Dutch workers proposed that smokers with chronic, largely irreversible airways obstruction and subjects with asthma shared a common constitutional predisposition to allergy, airway hyperresponsiveness, and eosinophilia—the 'Dutch hypothesis'. Numerous studies have shown that smokers tend to have higher levels of IgE and higher eosinophil counts than non-smokers, but the levels are not as high as those in individuals with asthma. Studies in middle-aged smokers with a degree of impairment of lung function show a positive correlation between accelerated decline in FEV_1 and increased airway responsiveness to either methacholine or histamine. However, atopic status, as defined by positive skin tests, does not differ between smokers and those who have never smoked.

Whether airway hyperresponsiveness is a cause or consequence of COPD is still a matter of debate.

Epidemiology

Prevalence

The symptom of mucus hypersecretion has been extensively studied in general population surveys over the last 40 years. In these studies, usually in middle-aged men, the prevalence of chronic cough and sputum production ranges between 15 and 53 per cent, with a lower prevalence of between 8 and 22 per cent in women, being more prevalent in urban than in rural areas. A study in the late 1980s showed a decline in the prevalence of chronic cough and phlegm in middle-aged men to 15 to 20 per cent, with little change in women.

Prevalence studies of COPD are normally based on values of percentage of predicted FEV_1, which defines individuals with and without airways obstruction. In a survey in 1987 of a representative sample of 2484 men and 3063 women in the United Kingdom, in the age range 18 to 64 years, 10 per cent of men and 11 per cent of women had an FEV_1 that was greater than 2 standard deviations below their predicted values, the numbers increasing with age, particularly in smokers. In current smokers in the age range 40 to 65 years, 18 per cent of men had an FEV_1 greater than 2 standard deviations below normal and 14 per cent of women, compared with 7 and 6 per cent of male and female non-smokers, respectively.

Studies from the United States, which used a lower limit of normal for FEV_1 of less than 65 per cent of the predicted value, show the prevalence of COPD in men falling from 8 per cent in the 1960s to 6 per cent in the late 1970s, whereas in women the prevalence of 3 per cent did not change over a similar period. National surveys of consultations in British general practices have shown a modest decline in the number of middle-aged men consulting their doctor with symptoms suggestive of COPD and a slight increase among middle-aged women. These trends are confounded by changes over the years in the application of the diagnostic labels for this condition, particularly the overlap between COPD and asthma.

Mortality

There are large international variations in the death rate for COPD, which cannot be entirely explained by differences in diagnostic patterns, labels, or by differences in smoking habits (Fig. 2). COPD is often a contributory factor to the cause of death, and thus figures from death certification underestimate the mortality from COPD. Most of the mortality from this condition occurs in the over 65 years age group. Within the United Kingdom, age-adjusted death rates from chronic respiratory diseases vary by a factor of 5 to 10 in different geographical locations. Mortality rates tend to be higher in urban areas than in rural areas.

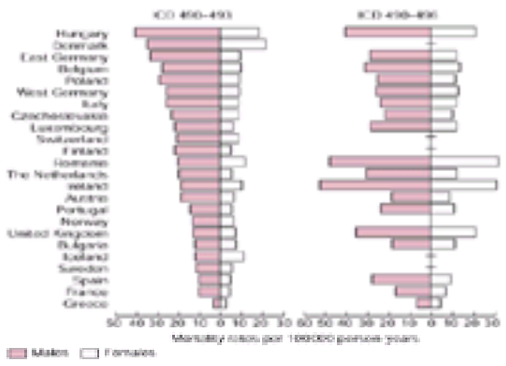

Fig. 2 Age standardized mortality rates in Europe, 1988 to 1991, for chronic obstructive pulmonary disease (COPD). ICD 490–496 = COPD and similar conditions. Closed bars to the left indicate mortality rates in males; open bars to the right in females. (Adapted from Siafakas NM *et al*., 1995, *European Respiratory Journal* **8**, 1398–420.)

Mortality from chronic respiratory disease (ICD 9 490–493 and 496) in males aged 55 to 84 years has been falling, except in the group over 75 years of age. Similar trends have been recorded in American men, whereas in women (whose mortality is one-third that of men) the decline in mortality which was recorded until 1975 has since shown a slight increase in those over the age of 65 years. These trends presumably relate to the later time of the peak prevalence of cigarette smoking in women compared with men.

In the United Kingdom there are around 30 000 deaths per year from COPD. These accounted for 6.4 per cent of all male deaths and 3.9 per cent of all deaths in females in 1994.

Morbidity/use of health resources

COPD places an enormous burden on health-care resources. An estimate of the annual workload in primary and secondary care attributable to COPD and its associated conditions in an average United Kingdom health district is shown in Table 1. It has been calculated that airways diseases (chronic bronchitis and emphysema, COPD, and asthma) account for 24.4 million lost working days per year in the United Kingdom, which represents 9 per cent of all certified sickness absence among men, and 3.5 per cent of the total among women. Respiratory diseases in the United Kingdom rank as the third commonest cause of days of certified incapacity; COPD accounting for 56 per cent of these days lost in males and 24 per cent in females.

History and prognosis

Severe airways obstruction occurs in susceptible smokers as a result of years of an accelerated decline in FEV_1. In non-smokers the FEV_1 declines at a rate of 20 to 30 ml/year (Fig. 3); this occurs at a faster rate in smokers, reported changes in FEV_1 in patients with COPD being more than 50 ml/year. Fletcher and colleagues found a relationship between the initial level of FEV_1 and the annual rate of decline in FEV_1 over a follow-up period of 8 years in working men in London. From these data they suggested that susceptible cigarette smokers could be identified in early middle age by a reduction in the FEV_1 (Fig. 3). They also suggested that there was a tracking effect, whereby individuals in the highest or lowest FEV_1 percentiles remained in the same percentiles over subsequent years. Support for the tracking effect comes from a study of 2718 working men, whose pulmonary function was assessed in the 1950s and subsequently followed up over 20 years. In those whose initial FEV_1 was more than 2 standard deviations below predicted values, the risk of death from chronic airways obstruction was 50 times greater than those whose initial FEV_1 was above average. There is a tendency for annual rates of decline in FEV_1 to be slower in advanced than in mild disease.

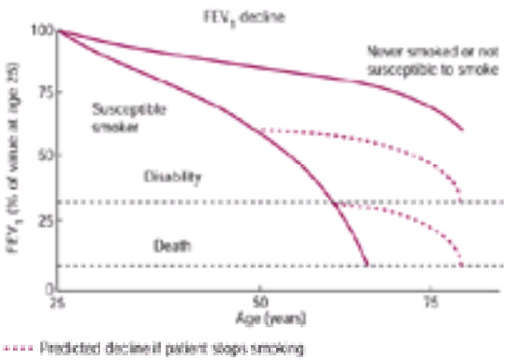

Fig. 3 The effect of age upon airflow obstruction in normal subjects and susceptible cigarette smokers. Cessation of smoking (dotted curved lines) returns the rate of decline to normal.

The strongest predictors of survival in patients with COPD are age and baseline FEV_1. Less than 50 per cent of patients whose FEV_1 has fallen to 30 per cent of predicted are alive 5 years later. There is an even stronger relationship between survival and the post-, rather than prebronchodilator FEV_1. Other unfavourable prognostic factors include severe hypoxaemia, raised pulmonary arterial pressure, and low carbon monoxide transfer, which become apparent in patients with severe disease. Factors favouring improved survival are stopping smoking and a large bronchodilator response. A reduced FEV_1 is also an important additional risk factor for lung cancer, independent of age or cigarette smoking.

Pathology

The pathological changes in patients with COPD are complex and occur in both the large and small airways, and in the alveolar compartment. The relative contributions that the pathological changes in the airways and those of emphysema make to airways obstruction have been the subject of considerable study. In general, pathological changes correlate rather poorly with both clinical and functional patterns of the disease. As a result there is still no clear consensus on whether the fixed airway obstruction in COPD is largely due to inflammation and scarring in the small airways, resulting in narrowing of the airway lumen, or to loss of support for the airways due to loss of alveolar walls, as in emphysema. Although the pathology of COPD is complex, it can be simplified by considering separately the three sites described above in which pathological changes could, in smokers, produce a clinical pattern of largely fixed airways obstruction. The clinicopathological picture is complicated by the fact that these three entities, or any combination of the three, may exist in an individual patient.

Chronic bronchitis

The pathological basis of hypersecretion of mucus is an increase in the volume of the submucosal glands, and an increase in the number and a change in the distribution of goblet cells in the surface epithelium. Submucosal glands are confined to the bronchi, decrease in number and in size in the smaller, more peripheral bronchi, and are not present in the bronchioles.

In healthy subjects who have never smoked, goblet cells are predominantly seen in the proximal airways, and decrease in number in more distal airways, being normally absent in terminal or respiratory bronchioles. By contrast, in smokers, goblet cells not only increase in number, but extend more peripherally.

The use of bronchoscopy to obtain airway cells by bronchoalveolar lavage and bronchial tissue samples by biopsy has added new insights into the role of inflammation in COPD. Bronchial biopsy studies confirm those of resected lung material, which show bronchial wall inflammation in this condition. As in asthma, bronchial biopsies in patients with chronic bronchitis reveal that activated T lymphocytes are prominent in the proximal airway walls. However, in contrast to asthma, macrophages are also a prominent feature and the CD8 suppresser T-lymphocyte subset, rather than the CD4 subset, predominates.

Bronchial biopsies from limited studies in patients during exacerbations of chronic bronchitis show increased numbers of eosinophils in the bronchial walls, although their numbers are small compared with exacerbations of asthma and, by contrast to those in asthma, these cells do not appear to have degranulated.

Bronchoalveolar lavage, or more recently studies of spontaneously produced or induced sputum, have shown increased intraluminal airspace inflammation in patients with chronic bronchitis, with or without airways obstruction, and predominantly neutrophils and macrophages in bronchoalveolar lavage studies. There is also evidence that airspace inflammation in patients with chronic bronchitis persists following smoking cessation if the production of sputum persists.

These studies of sputum and bronchial biopsies in chronic bronchitis have mainly sampled the proximal airways. Recent studies suggest that inflammatory changes present in the large airways may reflect those present in the small airways, and perhaps even in the alveolar walls.

Emphysema

Airspace enlargement can be identified macroscopically on the cut surface of an inflated lung when the airspace size reaches 1 mm. Two major types of emphysema are recognized, according to the distribution of enlarged airspaces within the acinar unit (Fig. 4):

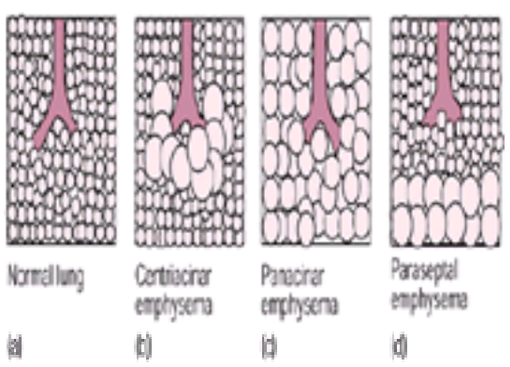

Fig. 4 Diagramatic representation of the distribution of abnormal airspaces within the acinar unit in different types of emphysema. (a) This represents the acinar unit in a normal lung, (b) shows the focal enlargement airspaces around the respiratory bronchioles in centriacinar emphysema, (c) shows the confluent even involvement of the acinar unit in panacinar emphysema, and (d) shows the peripherally distributed enlarged airspaces abutting the pleura in paraseptal emphysema. (Adapted from Lamb (1995). In: Calverley P, Pride N, eds. *Chronic obstructive pulmonary disease*, pp.9–34. Chapman and Hall, London.)

1. centriacinar (or centrilobular) emphysema, in which enlarged airspaces are initially clustered around the terminal bronchiole; and
2. panacinar (or panlobular) emphysema, where the enlarged airspaces are distributed throughout the acinar unit.

Centriacinar emphysema is more common in the upper zones of the upper and lower lobes: panacinar emphysema may be found anywhere in the lungs, but is more prominent at the bases, and is associated with a_1-protease inhibitor deficiency. Both types of emphysema can occur alone or in combination in a patient with emphysema. There is still debate over whether centriacinar and panacinar emphysema represent different disease processes, and hence have different aetiologies, or whether panacinar emphysema is a progression from centriacinar emphysema. There is a clearer association between centriacinar emphysema and cigarette smoking than with panacinar emphysema. Smokers with centriacinar emphysema have more small airways disease than those patients with predominantly panacinar emphysema.

Periacinar (or paraseptal or distal acinar) emphysema describes enlarged airspaces along the edge of the acinar unit, but only where it abuts against a fixed structure such as the pleura or a vessel. This is less common and usually of little clinical significance except when extensive in a subpleural position and may be associated with pneumothorax.

Scar or irregular emphysema is used to describe enlarged airspaces around the margins of a scar, unrelated to the structure of the acinus. This lesion is excluded from the current definition of emphysema.

A bulla represents a localized area of emphysema that has locally overdistended; conventionally greater than 1 cm in size.

Absence of fibrosis is a prerequisite in the most recent definition of emphysema. However, fibrosis occurs in the terminal or respiratory bronchioles as part of a respiratory bronchiolitis in asymptomatic cigarette smokers. Furthermore, there is an increase in collagen in the lung parenchyma in smokers compared with non-smokers.

The bronchioles and small bronchi are supported by attachment to the outer aspect of their walls of adjacent alveolar walls. This arrangement maintains the tubular integrity of the airways. It has been suggested that loss of these attachments may lead to distortion and irregularity of airways, which results in airflow limitation (Fig. 5).

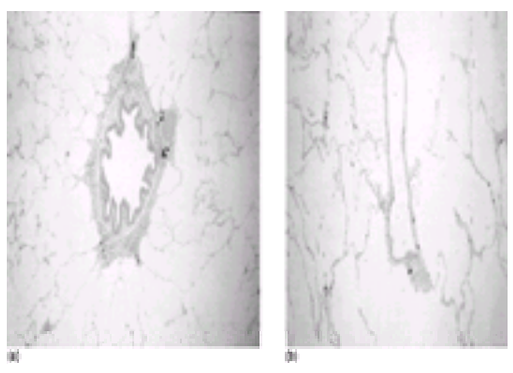

Fig. 5 (a) Cross-section of a normal small peripheral bronchiole, showing a circular outline supported by adjacent alveolar walls. (b) A small bronchiole at the same

magnification in a patient with emphysema. The loss of alveolar supporting walls results in an eliptical airway.

Bronchiolitis/small airways disease

Hogg, Macklem, and Thurlbeck introduced the concept of 'small airways disease' in studies using a retrograde catheter in which they showed that the increased flow resistance in the lungs in patients with COPD largely occurred in the small airways at the periphery of the lungs. Several pathological changes are found in small airways (Fig. 6):

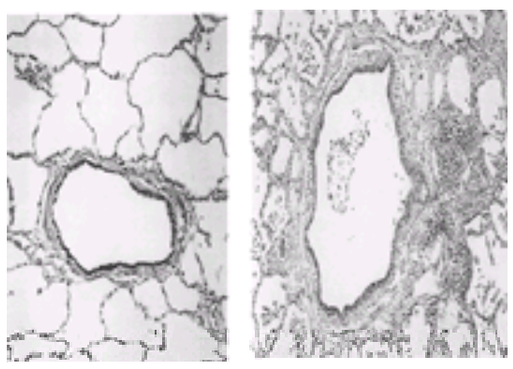

Fig. 6 Normal bronchiole (left) and a bronchiole of a patient with small airways disease/bronchiolitis (right), showing marked increase in inflammatory cells in the walls and in the airway lumen.

- inflammatory infiltrate in the airway wall
- mucus and cells in lumen
- goblet cell hyperplasia
- fibrosis in the airway wall
- squamous cell metaplasia
- mucosal ulceration
- increased amount of muscle
- pigmentation.

The pathological changes in the pulmonary vasculature and the right ventricle are described in Section 15.15.2.

Pathogenesis of COPD

Important to the pathogenesis of COPD were the observations of an association between a $_1$-antitrypsin deficiency and the development of early onset emphysema, and the development of emphysema following instillation of the proteolytic enzyme papain into rat lungs. These two important observations form the basis of the protease/antiprotease theory of the pathogenesis of emphysema. This hypothesis states that under normal circumstances the release of proteolytic enzymes from inflammatory cells that migrate to the lungs to fight infection does not cause lung damage because of inactivation of these proteolytic enzymes by an excess of inhibitors. However, in conditions of excessive enzyme load, or where there is an absolute or a functional deficiency of antiproteases, an imbalance develops between proteases and antiproteases in favour of proteases, leading to uncontrolled enzyme activity and degradation of lung connective tissue in alveolar walls, resulting in emphysema (Fig. 7).

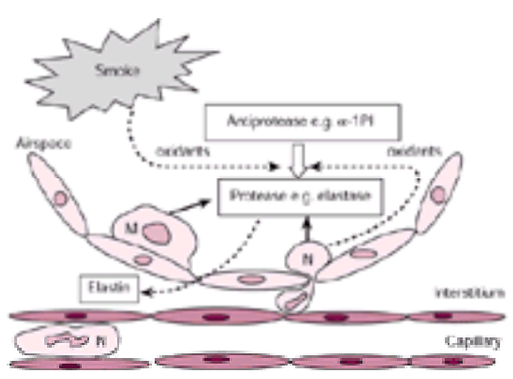

Fig. 7 Simplified protease/antiprotease theory of the pathogenesis of emphysema. Neutrophils (N) sequester in the pulmonary capillaries, initially as a result of the oxidant effect of cigarette smoke that decreases neutrophil deformability. Activated neutrophils adhere to the endothelial cells and subsequently migrate into the airspaces. Oxidants either directly from cigarette smoke or released from activated airspace neutrophils inactivate antiproteases such as a $_1$-protease inhibitor (a_1-PI) reducing its ability to bind to and hence inactivate proteases, particularly elastase. This allows active elastase to enter the lung interstitium and bind to and destroy elastin, causing destruction and enlargement of the distal airspace walls. This simplified protease/antiprotease theory is complicated by the presence of other antiproteases such as antileucoprotease, and other proteases such as metalloproteases released from macrophages (M). There is also the potential for neutrophils to be activated and release elastase while sequestered in the pulmonary capillaries without the need to migrate.

a_1-Antitrypsin/a_1-protease inhibitor

a_1-Antitrypsin is a potent inhibitor of serine proteases, and has greatest affinity for the enzyme neutrophil elastase. It is synthesized in the liver and increases from its usual plasma concentration of approximately 2 g/l as part of the acute-phase response. The activity of the protein is critically dependent on the methionine–serine sequence at its active site.

The average a_1-antitrypsin plasma levels for the more common phenotypes are shown in Table 2. The Z deficiency state (PiZ) is associated with periodic acid–Schiff (PAS)-positive inclusion bodies in the liver, which represent accumulations of the a $_1$-antitrypsin protein. Although liver and mononuclear cells from PiZ patients can manufacture normal amounts of mRNA, and the protein can be translated, there is little secretion of the protein. It is now recognized that the Z a $_1$-antitrypsin gene is normal except for a single point mutation, resulting from substitution of a glycine nucleotide for adenine in the DNA sequence that codes for the amino acid at position 342 on the molecule. This results in spontaneous polymerization of the protein. Large polymers of a $_1$-antitrypsin accumulate in the liver and are unable to pass through the rough endoplasmic reticulum. Their accumulation impairs a $_1$-antitrypsin secretion.

Pathogenesis of emphysema in patients without a_1-antitrypsin

The pathogenic mechanisms of emphysema and also of the small airways disease in COPD are clearly more complex than in patients with a $_1$-antitrypsin. In this case

the clearest association is with cigarette smoking, but since only 15 to 20 per cent of cigarette smokers develop COPD, the question of susceptibility has to be considered. The development of pulmonary emphysema in smokers is thought to occur as a result of several mechanisms:

- increased protease burden
- oxidant/antioxidant imbalance
- decreased antiprotease function
- decreased synthesis of elastin.

Proteases other than neutrophil elastase and antiproteases other than a $_1$-protease inhibitor may be involved in the protease–antiprotease imbalance in emphysema (Fig. 7).

Pathophysiology

Lung mechanics

The characteristic physiological abnormality in COPD is a decrease in maximum expiratory flow. Two major factors can reduce forced expiratory flow:

1. loss of lung elasticity; and
2. an increase in airways resistance in small and/or large airways.

In healthy young subjects significant airway closure only occurs below functional residual capacity (**FRC**). However, enhanced airway closure occurs in the early stages of COPD, which can be measured by plotting the nitrogen concentrations against expired volume following a single vital capacity breath of 100 per cent oxygen. The 'closing volume' measures the lung volume at which expired nitrogen concentrations increase abruptly during slow deflation from total lung capacity (**TLC**) and is determined by the lung volume at which some lung units close their airways and hence stop emptying. In healthy young non-smokers, closing volume is about 5 to 10 per cent of vital capacity (**VC**), rising to 25 to 35 per cent of VC in old age. Compared with non-smokers, young asymptomatic adult smokers have an increase in closing volume. As airways disease progresses the ability to define a closing volume decreases and therefore the test is not useful in established disease. Asymptomatic smokers who develop a reduced FEV_1, initially had an abnormal single-breath nitrogen washout test. Conversely, many subjects who have an abnormal single-breath nitrogen washout test do not develop an abnormal FEV_1.

In comparison with non-smokers, asymptomatic smokers also show frequency dependence of lung compliance, implying increased inequality of time constants in the lungs, resulting from changes in the compliance and resistance of parallel lung compartments. The pathological changes that occur in the peripheral airways in COPD are thought to be reflected in changes in maximum flow at lung volumes below 50 per cent of VC.

Tests of overall lung mechanics such as the FEV_1 and airways resistance are usually abnormal in patients with COPD when breathlessness develops. Residual volume, FRC, and (in some cases) TLC increase. Maximum expiratory flow–volume curves (MEFV) show a characteristic convexity towards the volume axis, initially with preservation of peak expiratory flow (Fig. 8).

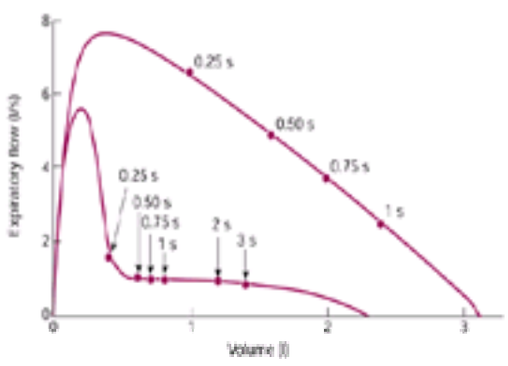

Fig. 8 Maximum flow–volume curves in a healthy subject (FEV_1 2.4 litres) and a subject with COPD and airways obstruction (FEV_1 0.8 litres). The development of convexity of the expiratory curve in mild obstruction is characteristic, as is the relative preservation of peak expiratory flow in the patient with COPD compared to the FEV_1. The figures on the flow–volume curve represent time in seconds for both curves.

The uneven distribution of ventilation in advanced COPD causes a reduction in 'ventilated' lung volume and thus the carbon monoxide transfer factor (*TL*CO) is almost always reduced, although the *TL*CO normalized to ventilated alveolar volume (*K*CO) may remain relatively well preserved in those without emphysema.

The characteristic changes in the static pressure/volume (*P*/*V*) curve of the lungs in COPD are an increase in static compliance and a reduction in static transpulmonary pressure at a standard lung volume (Fig. 9). These changes are generally thought to indicate emphysema.

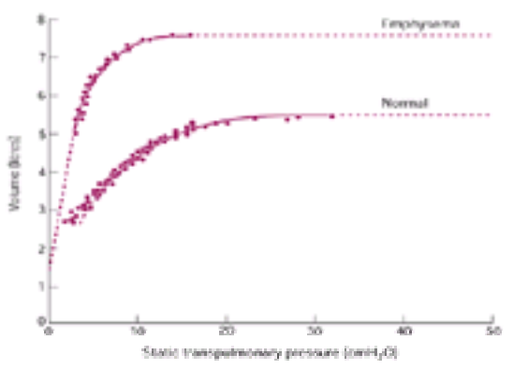

Fig. 9 Static expiratory pressure–volume curves of lungs in a subject with severe emphysema compared with a normal subject. The broken lines represent extrapolation of the curve to infinite pressure and to the volume axis at zero pressure.

The major site of the fixed airway narrowing in COPD is in the peripheral airways less than 2 to 3 mm in diameter. In addition, loss of lung elastic recoil pressure is also important in terms of airways obstruction, particularly in those with severe emphysema, as a result of a reduction in the distending force on all of the intrathoracic airways. Dynamic expiratory compression of the airways is enhanced by loss of lung recoil and by atrophic changes in the airways and loss of support from the surrounding alveolar walls, allowing flow limitation at lower driving pressures and flows.

Pulmonary gas exchange

Ventilation–perfusion (*V*/*Q*) mismatching is the most important cause of impaired pulmonary gas exchange in COPD. Other causes such as alveolar hypoventilation, impaired alveolar–capillary diffusion to oxygen, and increased shunt are of much less importance. The distribution of ventilation is very uneven in patients with COPD.

A reduction of blood flow is produced by several mechanisms including local destruction of vessels in alveolar walls as a result of emphysema, hypoxic vasoconstriction in areas of severe alveolar hypoxaemia, and passive vascular obstruction as a result of increased alveolar pressure and distension.

Respiratory muscles

In patients with severe COPD a combination of pulmonary overinflation and malnutrition, resulting in muscle weakness, reduces the capacity of the respiratory muscles to generate pressure over the range of tidal breathing. In addition, the load against which the respiratory muscles need to act is increased, due to the increase in airways resistance. Overinflation of the lungs leads to shortening and flattening of the diaphragm, thus impairing its ability to generate pressure in order to lower pleural pressure. During quiet tidal breathing in normal subjects, expiration is largely passive and depends on the elastic recoil of the lungs and the chest wall. Patients with COPD increasingly need to use their rib cage muscles and inspiratory accessory muscles, such as the sternomastoids, even during quiet breathing. During exercise, this pattern may be even more distorted and result in paradoxical motion of the rib cage.

Patients with COPD have impaired values of global function of the respiratory muscles such as maximum inspiratory mouth pressures ($P\mathrm{e}_{\mathrm{max}}$), although these measurements are very effort dependent. Diaphragmatic function can be assessed during inspiration by measurement of transdiaphragmatic pressure (Pdi), using balloon-tipped catheters with small transducers placed in the oesophagus and stomach. Measurements of Pdi are reduced in patients with COPD.

Cor pulmonale

Pulmonary arterial hypertension occurs late in the course of COPD with the development of hypoxaemia ($P\mathrm{aO}_2$ less than 8 kPa) and usually also hypercapnia. It is the major cardiovascular complication of COPD and is associated with the development of right ventricular hypertrophy ('cor pulmonale') and poor prognosis. Further details can be found in Section 15.15.2.

Clinical features

Symptoms

Patients with COPD characteristically complain of the symptoms of breathlessness on exertion, sometimes accompanied by wheeze and cough, which is often, but not invariably, productive. Breathlessness is the symptom that commonly causes the patient to seek medical attention and is usually the most disabling problem. Patients often date the onset of their illness to an acute exacerbation of cough with sputum production, which leaves them with a degree of chronic breathlessness. However, close questioning will usually reveal the presence of a 'smoker's cough', with the production of small amounts of mucoid sputum, often predominating in the morning, for many years.

A smoking history of at least 20 pack years is usual before symptoms develop, commonly in the fifth decade, following which there is progression through the clinical stages of mild, moderate, and severe disease. Breathlessness, usually first noticed on climbing hills or stairs, or hurrying on level ground, heralds the development of moderate impairment of airway function and patients may adapt their breathing pattern and their behaviour to minimize the sensation of breathlessness. The perception of breathlessness varies greatly for individuals with the same impairment of ventilatory capacity. However, when the FEV_1 has fallen to 30 per cent or less of the predicted values, breathlessness is usually present on minimal exertion. Severe breathlessness is often affected by changes in temperature and occupational exposure to dust and fumes. Some patients have severe orthopnoea, relieved by leaning forward, whereas others find greatest ease when lying flat. Breathlessness can be assessed on the Medical Research Council, Borg, and Visual Analogue scales.

A productive cough occurs in up to 50 per cent of cigarette smokers, may precede the onset of breathlessness, and many patients may dismiss this as simply a 'smoker's cough'. The frequency of nocturnal cough does not appear to be increased in stable COPD. Paroxysms of coughing in the presence of severe airway obstruction generate high intrathoracic pressures, which can produce syncope and 'cough fractures' of the ribs. Wheeze is common, but not specific to COPD, since it is due to turbulent airflow in large airways from any cause.

Chest pain is common in patients with COPD, but is often unrelated to the disease itself, and may be due to underlying ischaemic heart disease or gastro-oesophageal reflux. Chest tightness is a common complaint during exacerbations of breathlessness, particularly during exercise, and this is sometimes difficult to distinguish from ischaemic cardiac pain. Pleuritic chest pain may suggest an intercurrent pneumothorax, pneumonia, or pulmonary infarction.

Haemoptysis can be associated with purulent sputum and may be due to inflammation or infection. However, this symptom should be treated seriously and the need for investigations for bronchial carcinoma should be considered.

Weight loss and anorexia are features of severe COPD and thought to result from both decreased calorie intake and hypermetabolism. Psychiatric morbidity, particularly depression, is common in patients with severe COPD, reflecting social isolation and the chronicity of the disease. Sleep quality is impaired in advanced COPD, which may contribute to the impaired neuropsychiatric performance.

Occupation/smoking history

A detailed smoking history is important in patients with COPD since the disease is rare in lifelong non-smokers. Although there is, in general, a dose–response relating the number of cigarettes smoked and the level of the FEV_1, there are huge individual variations, reflecting variations in the susceptibility to cigarette smoke (Fig. 1). Occupational exposure to dusts has an additive effect on the decline in lung function, as has been shown in coal miners, where both smoking and years of dust exposure contribute to the decline in FEV_1, although the contribution of smoking was three times as great as that of the dust exposure.

Clinical signs

General examination

The physical signs in patients with COPD are not specific, and depend on the degree of airflow limitation and pulmonary overinflation. The sensitivity of physical signs to detect or exclude moderately severe COPD is poor. Tachypnoea may be present in patients with severe COPD and prolonged forced expiratory time (more than 5 s) can be a useful indicator of airway obstruction. The breathing pattern in COPD is often characteristic, with a prolonged expiratory phase, some patients adopting pursed lipped breathing on expiration, which reduces expiratory airway collapse. Use of the accessory muscles of respiration, particularly the steromastoids, is often seen in advanced disease and these patients often adopt the position of leaning forward, supporting themselves with their arms to fix the shoulder girdle, allowing the use of the pectorals and the latissimus dorsi to increase chest wall movement.

Tar-stained fingers emphasize the smoking habit. In advanced disease cyanosis may be present, indicating hypoxaemia, but may be diminished by anaemia or accentuated by polycythaemia, and is a fairly subjective sign. The flapping tremor, associated with hypercapnia, is neither sensitive nor specific, and papilloedema associated with severe hypercapnia is rarely seen. Weight loss may also be apparent in advanced disease, as well as a reduction in muscle mass. Finger clubbing is not a feature of COPD and should suggest the possibility of complicating bronchial neoplasm or bronchiectasis.

Examination of the chest

In the later stages of COPD the chest is often barrel-shaped with a kyphosis and an apparent increased anterior/posterior diameter, horizontal ribs, prominence of the sternal angle, and a wide subcostal angle. Due to the elevation of the sternum, the distance between the suprasternal notch and the cricoid cartilage (normally three finger-breadths) may be reduced. These are all signs of overinflation. An inspiratory tracheal tug may be detected, which has been attributed to the contraction of the low flat diaphragm. The horizontal position of the diaphragm also acts to pull in the lower ribs during inspiration—Hoover's sign. Widening of the xiphisternal angle and abdominal protuberance occur, the latter due to forward displacement of the abdominal contents, giving the appearance of apparent weight gain. Increased intrathoracic pressure swings may result in indrawing of the suprasternal and supraclavicular fossas and of the intercostal muscles.

On percussion of the chest there is decreased hepatic and cardiac dullness, indicating overinflation. A useful sign of gross overinflation is the absence of a dull

percussion note, normally due to the underlying heart, over the lower end of the sternum. Breath sounds may have a prolonged expiratory phase, or may be uniformly diminished, particularly in the advanced stages of the disease. Wheeze may be present both on inspiration and expiration, but is not an invariable clinical sign. Crackles may be heard particularly at the lung bases, but are usually scanty and vary with coughing.

Cardiovascular examination

The presence of emphysema or overinflation of the chest produces difficulty in localizing the apex beat and reduces the cardiac dullness. The characteristic signs that indicate the presence or consequences of pulmonary arterial hypertension may be detected in advanced cases. The heave of right ventricular hypertrophy may be palpable at the lower left sternal edge. Heart sounds are generally soft, although the pulmonary component of the second heart sound may be exaggerated in the second left intercostal space. A gallop rhythm may be detectable, with a third sound audible in the fourth intercostal space to the left of the sternum. The jugular venous pressure can be difficult to see in patients with COPD as it swings widely with respiration and is difficult to discern if there is prominent accessory muscle activity. When the fluid retention of cor pulmonale occurs there may be evidence of functional tricuspid incompetence, producing a pansystolic murmur at the left sternal edge. The liver may be tender, pulsatile, and a prominent 'v' wave may be visible in the jugular venous pulse. The liver may also be palpable below the right costal margin as a result of overinflation of the lungs.

Peripheral vasodilatation accompanies hypercapnia, producing warm peripheries with a high-volume pulse. Pitting peripheral oedema may also be present as a result of fluid retention.

Investigation

Physiological assessment

The most important disturbance of respiratory function in COPD is obstruction to forced expiratory airflow. The degree of airflow obstruction cannot be predicted from the symptoms and signs and therefore an assessment of the degree and the progression of airways obstruction should be made. At an early stage of the disease conventional spirometry may reveal no abnormality, since the earliest changes in COPD affect the alveolar walls and small airways, producing a modest increase in peripheral airway resistance that is not reflected in spirometry measurements. Tests of small airway function, such as the frequency dependency of compliance and closing volume may be abnormal. These tests are difficult to perform, have high coefficients of variation, and are only valid when lung elastic recoil is normal and there is no increase in large airway resistance. They are therefore not recommended in normal clinical practice.

Spirometry

Spirometry is the most robust test of airflow limitation in patients with COPD. A low FEV_1 with an FEV_1/VC ratio below the normal range is a diagnostic criterion for COPD. The rate of decline of the FEV_1 can be used to assess susceptibility in cigarette smokers, progression of the disease, and to test reversibility of the airways obstruction. It is important that a volume plateau is reached when performing the FEV_1, which can take 15 s or more in patients with severe airways obstruction. If this manoeuvre is not carried out the FVC can be underestimated. The FEV_1 as a percentage of the predicted value can be used to assess the severity of the disease (Table 3).

Flow volume loops

Expiratory flows at 75 or 50 per cent of vital capacity have been used as a measure of airflow limitation. These measurements are less reproducible than spirometry, so that abnormal values must fall to below 50 per cent of the predicted values. Flows at lung volumes less than 50 per cent of vital capacity were previously considered to be an indicator of small airways function, but probably provide no more clinically useful information than measurements of FEV_1.

Peak expiratory flow

Peak expiratory flow can either be read directly from the flow volume loop or measured with a hand-held peak flow meter; the latter are relatively easy to use and are particularly useful in subjects with asthma for revealing variations in serial measurements. However, in COPD many variations are often within the error of the measurement. The peak expiratory flow may underestimate the degree of airflow obstruction in COPD (Fig. 8).

Lung volumes

Static lung volumes such as total lung capacity (TLC), residual volume (RV), and functional residual capacity (FRC) are measured in patients with COPD to assess the degree of overinflation and gas trapping. Dynamic overinflation occurs particularly during exercise and may be an important determinant of the symptom of breathlessness.

The standard method of measuring static lung volumes, using the helium dilution technique during rebreathing, may underestimate lung volumes in COPD, particularly in those patients with bullous disease, where the inspired helium does not have time to equilibrate properly in the airspaces. Body plethysmography uses Boyle's law to calculate lung volumes from changes in mouth and plethysmographic pressures. This technique measures trapped air within the thorax, including poorly ventilated areas, and therefore gives higher readings than the helium dilution technique.

Reversibility to bronchodilators

Assessment of reversibility to bronchodilators is recognized as an essential part of the investigation and management of patients with COPD. Reversibility tests are important:

1. to help distinguish those patients with marked reversibility who have underlying asthma;
2. to aid with future management; and
3. since the post-bronchodilator FEV_1 is the best predictor of survival.

There is, however, no agreement on a standardized method of assessing reversibility, which is usually recorded as change in FEV_1 or peak expiratory flow. However, there may be changes in other lung volumes after bronchodilators, such as inspiratory capacity: this may explain why symptoms improve in some patients following a bronchodilator without change in spirometry. An improvement in FEV_1 in response to a bronchodilator does not necessarily predict a symptomatic response.

The European Respiratory Society and the British Thoracic Society guidelines both recommend that changes should only be considered significant if they exceed 200 ml and represent a 15 per cent improvement over the baseline value. A change over baseline of greater than 12 per cent of the predicted normal value is regarded as a significant bronchodilator response by the American Thoracic Society. A third approach, which has received less support, is to express the change in FEV_1 as a percentage of the potential possible change, which is the predicted value minus the baseline value. Daily variations in airway smooth muscle tone may affect the response to bronchodilators in patients with COPD. Thus, when airway smooth muscle tone is higher, and thus FEV_1 is lower, a response to bronchodilators may be more likely to be achieved than when muscle tone is lower and FEV_1 is higher.

It is usually recommended that the response to a bronchodilator be assessed with a large dose, either using repeated doses from a metered dose inhaler, or by the nebulized route, since this produces a larger number of patients with a significant response. In some cases the addition of a second drug, such as an anticholinergic drug to a b-agonist, will produce a further increase in FEV_1.

Reversibility to corticosteroids

Whether all patients with symptomatic COPD should have a formal assessment of steroid reversibility remains controversial. The commonest regimen is the administration of 30 mg of prednisolone for a period of 2 weeks. Those patients who have previously shown a response to nebulized bronchodilators are more likely to

show a response to steroids. However, it is not possible to predict the response to corticosteroids in an individual patient.

An alternative approach is to assess the response to inhaled steroids, usually over a 6-week period, measuring the FEV_1 before and after the average of the first 5 days and the last 5 days measurements of peak expiratory flow.

Gas transfer for carbon monoxide (*TL*CO)

A low *TL*CO is present in many patients with COPD. Although there is a relationship between the *TL*CO and the extent of microscopic emphysema, the severity of the emphysema in an individual patient cannot be predicted from the *TL*CO, nor is a low *TL*CO specific for emphysema. The commonly used method is the single-breath technique, which uses alveolar volume calculated from helium dilution during the single-breath test. This will underestimate alveolar volume in patients with severe COPD, producing a lower value for the *TL*CO.

Arterial blood gases

Arterial blood gases are needed to confirm the degree of hypoxaemia and hypercapnia in patients with COPD. It is essential to record the inspired oxygen concentration when reporting blood gases. It is also important to note that it may take at least 30 min for a change in inspired oxygen concentration to have its full effect on the PaO_2, because of long time constants for alveolar gas equilibration in COPD, particularly during exacerbations. Pulse oximetry is increasingly used to measure the level of oxygenation, but should not replace an assessment of blood gas tensions, since measurements of $PaCO_2$ are often required. Acid–base status can also be assessed from the arterial pH (hydrogen ion concentration) and the bicarbonate. Increases in $PaCO_2$, which can occur rapidly, can be compensated by renal conservation of bicarbonate ions, which is a relatively slow process. Acid–base status, particularly mixed respiratory and metabolic disturbances, can be characterized by plotting values on an acid–base diagram (Fig. 10).

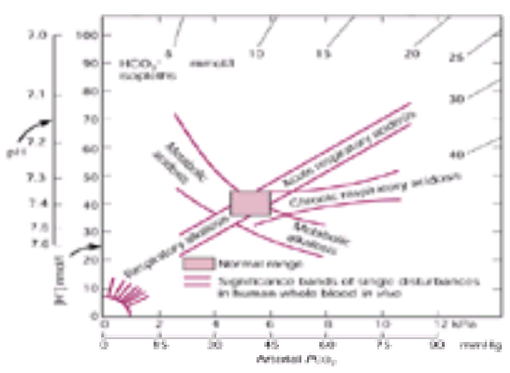

Fig. 10 A non-logarithmic acid–base diagram derived from the measured acid–base status of patients within the five abnormal bands illustrated and of normal subjects (hatched box). This plot of carbon dioxide tension against hydrogen ion concentration (pH) allows the likely acid–base disturbance and calculated bicarbonate value (obtained from the relevant isopleth) to be determined, whilst changes during treatment can be plotted serially for each patient. (Adapted from Flenley DC, 1971, *Lancet* **i**, 961–5.)

Exercise tests

Exercise increases oxygen consumption and carbon dioxide production from skeletal muscle. Patients with COPD have the same oxygen consumption for a given workload as normal subjects, but because their dead-space ventilation is higher, a larger minute ventilation is needed to maintain a constant carbon dioxide level. Since in many patients expiratory airflow is limited within the tidal volume range, the only way to increase minute ventilation is to increase inspiratory flow and/or shift the end expiratory position. Both of these manoeuvres are problematic in patients with COPD and require more work from already compromised inspiratory muscles, or result in progressive overinflation, which increases both the work of breathing and symptoms. Metabolic acidosis develops at lower work rates in patients with severe COPD. In patients with COPD, progressive cycle exercise is limited by dyspnoea in 40 per cent and by leg fatigue in 25 per cent, probably reflecting general debility.

Three forms of exercise test can be performed, which provide useful information.

Progressive symptom-limited exercise

In this test the patient is encouraged to maintain exercise, on a treadmill or a cycle, until symptoms prevent them from continuing. A maximum test is usually defined as a heart rate of greater than 85 per cent of predicted or ventilation greater than 90 per cent predicted. The results are useful to assess whether coexisting cardiac or psychological factors contribute to exercise limitation.

Self-paced exercise

These tests are easy to perform. The 6-min walk is the most commonly used test and has a coefficient of variation of around 8 per cent. Shortening the walk to 2 min decreases the reproducibility. The test is only useful in patients with moderately severe COPD (FEV_1 less than 1.5 litres) who would be expected to have an exercise tolerance of less than 600 m in 6 min. There is a weak relationship between walking distance and FEV_1.

Steady-state exercise

This involves exercise at a sustainable percentage of maximum capacity for 3 to 6 min, during which blood gases are measured, enabling calculation of dead space:tidal volume ratio (*V*D/*V*T) and shunt. This assessment is seldom required in patients with COPD.

Sleep studies

Hypoxaemia occurs during sleep, particularly rapid eye movement sleep, in patients with COPD. However, measurement of nocturnal hypoxaemia does not provide any further prognostic or clinically useful information in the assessment of patients with COPD, unless coexisting sleep apnoea syndrome is suspected.

Non-physiological assessments

Identifying polycythaemia is important in patients with severe COPD since it predisposes to vascular events. Polycythaemia should be suspected when the haematcrit is greater than 47 per cent in women and 52 per cent in men, and/or the haemoglobin is greater than 16 g/dl in women and 18 g/dl in men, provided other causes of spurious polycythaemia, due to decreased plasma volume, such as caused by dehydration or diuretics, can be excluded.

There is no indication for measuring blood biochemistry routinely in patients with clinically stable COPD. a $_1$-Antitrypsin levels and phenotype should be measured in all patients under the age of 40 years, and in those with a family history of emphysema at an early age.

Routine electrocardiography is not required in the assessment of patients with COPD and is an insensitive technique in the diagnosis of cor pulmonale.

Radiology

Plain chest radiography

The features on a plain posterior–anterior chest radiograph are not specific for COPD and are usually those of severe emphysema. There may be no abnormalities, even in patients with very appreciable disability. Bronchial wall thickening, shown as parallel line opacities on plain chest radiography, has been described, but this finding may relate to coincidental bronchiectasis. The most reliable radiographic signs of emphysema can be divided into those due to overinflation, vascular changes, and bullae.

Overinflation of the lungs results in the following.

1. There is a low flattened diaphragm (Fig. 11). Low means that the border of the diaphragm in the mid-clavicular line is at or below the anterior end of the sixth or seventh rib. In a flattened diaphragm the maximum perpendicular height from a line drawn between the costal and cardiophrenic angles to the border of the diaphragm is less than 1.5 cm.

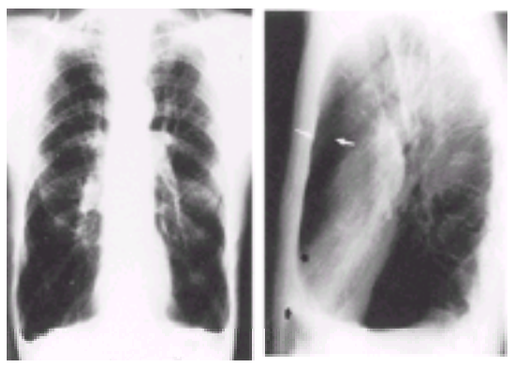

Fig. 11 Generalized (panacinar) emphysema on plain chest radiograph. The diaphragm is low (below the anterior ends of the seventh ribs) and flat. The lower zones are transradient because of oligaemia and there are obtuse costophrenic angles. On the lateral chest radiograph the diaphragm is mildly inverted, the retrosternal transradiency is wide (white arrows) and inferiorly it closely approaches the diaphragm (black arrow).

2. An increased retrosternal airspace occurs when the horizontal distance from the anterior surface of the aorta to the sternum exceeds 4.5 cm on the lateral film at a point 3 cm below the manubrium.
3. There is an obtuse costophrenic angle on the posterior–anterior or lateral chest radiograph.
4. The inferior margin of the retrosternal airspace is 3 cm or less from the anterior aspect of the diaphragm.

The vascular changes associated with emphysema result from loss of alveolar walls and appear as:

1. a reduction in size and number of pulmonary vessels, particularly at the periphery of the lung;
2. vessel distortion, producing increased branching angles, excess straightening, or bowing of vessels; and
3. areas of transradiency.

A general increased transradiency may be due to an overexposed chest radiograph. Focal areas of transradiency surrounded by hairline walls represent bullae. These may be multiple, as part of a generalized emphysematous process, or localized. An 'increase in lung markings' rather than areas of increased transradiency has often been described in patients with COPD: the cause of these changes is unknown, but may at least be contributed to by non-vascular linear opacities due to scarring.

The accuracy of diagnosing emphysema on the plain chest radiograph increases with the severity of the disease and has been reported as being 50 to 80 per cent accurate in patients with moderate to severe disease. However, the sensitivity has been reported as being as low as 24 per cent in patients with mild to moderate disease.

Computed tomography (CT)

CT scanning has been used since the early 1980s to detect and quantify emphysema. Studies using CT scanning can be divided into those that use visual assessment of low-density areas of the CT scan, which can be either semiquantitative or quantitative, and those that use CT lung density to quantify areas of low x-ray attenuation. These studies roughly divide into those that measure macroscopic or microscopic emphysema, respectively.

A visual assessment of emphysema on CT scan (Fig. 12(a) and Fig. 12(b)) reveals:

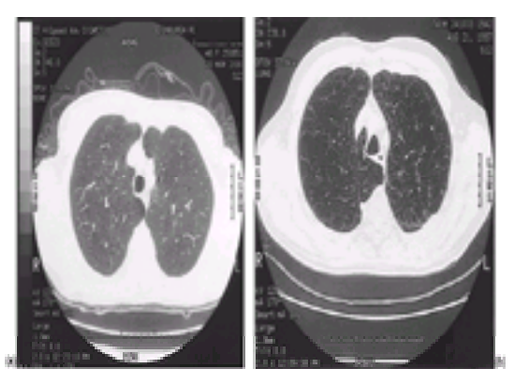

Fig. 12 High resolution CT scan of (a) normal lung, (b) panacinar emphysema.

1. areas of low attenuation without obvious margins or walls;
2. attenuation and pruning of the vascular tree; and
3. abnormal vascular configurations.

The sign that correlates best with areas of macroscopic emphysema is an area of low attenuation. However, visual assessment of the extent of macroscopic emphysema by CT scanning is insensitive, subjective, and has a high intra- and interobserver variability. Thus, in most of this type of study CT scanning tends to underestimate the severity of the disease, with centrilobular lesions smaller than 5 mm particularly likely to be missed.

It is possible using high-resolution CT to distinguish the various types of emphysema, particularly when the changes are not severe, depending on the distribution of the lesions.

A more quantitative approach of assessing macroscopic emphysema is by highlighting pixels within the lung fields in a predetermined low-density range, between –910 and –1000 Houndsfield units, the so-called 'density mask' technique. The choice of the density range is fairly arbitrary. A good correlation has been shown

between pathological emphysema scores and CT 'density mask' score, but this technique may still miss areas of mild emphysema.

Microscopic emphysema can be quantified by measuring CT lung density. CT density is expressed on a linear scale in Houndsfield units (water = 0; air = –1000). In this range, CT lung density is a direct measure of physical density and is determined by the relative mix of air, blood, and interstitial fluid in tissue. Thus, as emphysema develops, a decrease in alveolar surface area would occur as alveolar walls are lost, associated with an increase in distal airspace size, which would decrease lung CT density in association with a decrease in lung function.

More studies are required before CT lung density can be used as a standardized technique to quantify microscopic emphysema. It is particularly important to define the range of normality, and to standardize the calibration of CT scanners and the lung volume at which scans should be performed. However, at present, CT scanning is the most sensitive and specific imaging technique for assessing emphysema in life and can detect mild emphysema in symptomatic patients with a normal chest radiograph.

Pulmonary hypertension/cor pulmonale

Right ventricular hypertrophy or enlargement produces non-specific cardiac enlargement on the plain chest radiograph, the most widely used measurement to assess the presence of pulmonary hypertension being the width of the right descending pulmonary artery, measured just below the right hilum, where the borders of the artery are delineated against air in the lungs laterally and the right main-stem bronchus medially. The upper limit of the normal range of the width of the artery in this area is taken as 16 mm in males and 15 mm in females. Other studies have suggested an upper limit of normal ranging between 16 and 20 mm, which gives a sensitivity of detecting a pulmonary arterial pressure greater than 20 mmHg of 68 to 95 per cent, with a specificity of 65 to 88 per cent. Although these measurements can be used to detect the presence or absence of pulmonary arterial hypertension, they cannot accurately predict the level of the pulmonary artery pressure, and can therefore only be used as a screening test.

Emphysematous bullae

A bulla has been defined arbitrarily as an emphysematous space greater than 1 cm in diameter. On the plain chest radiograph a bulla appears as a localized avascular area of transradiency, usually separated from the rest of the lung by a curvilinear hairline wall. Marked compression of the surrounding lung may be seen, and bullae may also depress the diaphragm. CT scanning is much more sensitive than plain chest radiography at detecting bullae and can be used to determine their number, size, and position. Ventilation of the bullae can also be assessed using inspiratory/expiratory images. It is also possible to estimate the volume of bullae by measuring the area of the bullae in each CT lung slice.

Prevention

Since tobacco smoking is the major aetiological factor in COPD, the disease is theoretically preventable, with cessation of cigarette smoking the single most important way of affecting the outcome. Other important aetiological factors such as atmospheric pollution are also preventable. In the United Kingdom around 31 per cent of men and 29 per cent of women are current cigarette smokers, and around 80 to 90 per cent of patients with COPD have been regular smokers at sometime in their life. At least 90 per cent of smokers are aware of the adverse health effects of cigarette smoking, 70 per cent wish to give up smoking, and the majority of these have made a serious attempt to quit. However, only 40 per cent of regular smokers have succeeded in quitting cigarette smoking by age 60. Nicotine in tobacco smoke is addictive and regular smokers who reduce or cease their nicotine intake experience the characteristic withdrawal syndrome resulting from nicotine craving, manifest as anxiety, lack of concentration, irritability, restlessness, and increased appetite. Nicotine addiction develops rapidly and withdrawal symptoms can be shown to occur even in adolescent smokers. Thus a critical preventive measure is to reduce the number of children starting smoking.

Smoking cessation

Smoking cessation reduces the subsequent decline in lung function (Fig. 3), therefore smoking cessation is the single most important step that can be taken to prevent the progression of the disease. This is particularly true during the early stages of COPD, where both symptoms and lung function may improve. In advanced disease, quitting smoking may not improve pulmonary function, but symptoms such as cough may still improve.

Every patient who smokes should have a discussion of the implications for their future health. Asking about smoking habit in every patient may have a positive reinforcing effect against starting smoking in non-smokers. The reported success rates of smoking cessation interventions come mainly from studies conducted in a primary care setting, and vary between 10 and 30 per cent. A recent review of the literature suggests that in those who request extra help to stop smoking, and when this is given in the form of nicotine replacement or even contact with a support group, the success rate can be up to 25 per cent.

Although it would seem logical, as in other addictions, to suggest a reduction in nicotine levels by a gradual reduction in the number of cigarettes smoked, so as to reduce the severity of withdrawal symptoms, it has been shown that patients who gradually cut down the number of cigarettes smoked tend to inhale more to maintain their usual blood nicotine levels. It has also been shown that those who are unable to quit abruptly are not successful in reducing their consumption of cigarettes over the long term.

The intensity of the strategy employed in a cessation programme should depend on the motivation of the patient to give up smoking. There is no difference in the success rates between regimes involving brief intervention and those with more prolonged intervention in unselected smokers, whereas it is clear that those who are motivated to attend smoking cessation clinics have a better chance of long-term cessation than those who have a brief intervention by the general practitioner. It is therefore better to put time and effort only into those patients who are motivated to give up, and offer only a brief intervention in those with less motivation.

It is important that patients are given a clear strategy for smoking cessation and that the success rates are measured by corroboration with carbon monoxide measurements in breath, or urinary cotinine levels. Meta-analysis of randomized controlled trials of nicotine gum found a clear benefit in terms of abstinence rates at 1 year (23 compared with 13 per cent) in a smoking cessation clinic, but no effect in a general practice setting (11 compared with 12 per cent). Similar abstinence rates at 1 year have been quoted in a general hospital study in the United Kingdom.

Nicotine skin patches allow a slow infusion of nicotine, which creates plasma nicotine levels up to half of those produced by smoking. Trials carried out with nicotine patches indicate that similar success rates to nicotine chewing gum can be achieved. Recent studies using the antidepressant drug bupropion have also shown quit rates similar to those of nicotine replacement therapy in smokers. Based on a recent review of the literature, a strategy for smoking cessation has been suggested (Table 4).

Management of stable COPD

Bronchodilators

Bronchodilator therapy is the cornerstone of treatment in patients with COPD to reduce symptoms and increase exercise tolerance. By contrast with bronchial asthma, the effects are small in patients with COPD, due to structural changes within the airways. The principal bronchodilators—b_2-agonists, anticholinergic drugs, and theophylline derivatives—relax airway smooth muscle as their primary action and hence decrease airway resistance. However, these drugs may also reduce overinflation of the lungs, allowing the lungs to empty more completely. It should be emphasized that relatively small changes in airway dimensions can have major effects on respiratory mechanics, which may be translated into improvement in symptoms and exercise capacity.

b-Agonists

Inhaled b_2-agonists are preferred to oral preparations, since they are as efficacious in much smaller doses and have fewer side-effects. They have a relatively rapid onset of action and are therefore used for symptomatic relief, and can also increase exercise tolerance in patients with COPD. There is no evidence that the response to a b-agonist diminishes with time and patients with COPD should be told to take them as required, although those with severe disease may prefer to take regular doses three to four times daily to obtain symptomatic relief.

There is limited information on the effects of long-acting b_2-agonists in patients with COPD. In randomized placebo-controlled studies there was an improvement in

symptoms and a small improvement in spirometry, without any significant change in exercise capacity but with an improvement in symptoms and quality of life. There is little evidence to support the use of sustained-release oral b $_2$-agonists in patients with COPD.

Anticholinergics

Like b$_2$-agonists, anticholinergics affect both central and peripheral airways and also reduce functional residual capacity (FRC). They have a 30- to 60-min time to peak effect in most patients with COPD, which is slower than b $_2$-agonists, but act for longer than b$_2$-agonists (6 to 10 h).

Optimal bronchodilatation occurs with 80 µg of ipratropium and 200 µg of oxitropium bromide, and studies comparing these treatments suggest no difference in the peak or duration of bronchodilatation. Thus 80 µg of ipratropium should be used in patients with COPD, rather than the customary 40 µg, to produce maximum effect. Tiotropium bromide is a newly developed anticholinergic agent that appears to have a longer time course of action than ipratropium.

Some studies found an increase in maximum exercise and a reduction in oxygen consumption at any given workload with both ipratropium bromide and oxitropium bromide. In a large group of patients with COPD the Lung Health Study showed that treatment with ipratropium bromide produced a small but significant beneficial effect on FEV$_1$ during treatment, but had no other effect on the decline in FEV $_1$ over a 5-year period.

Clinical studies of anticholinergic drugs show that they are at least as efficacious as b $_2$-agonists in patients with COPD and some report a more prolonged bronchodilator response.

Theophyllines

Theophyllines, or methylxanthine derivatives, produce a modest bronchodilator effect in patients with COPD. Their effect on symptoms and on exercise tolerance is variable and often occurs at the top of the therapeutic range. Long-term treatment with theophyllines is limited to the oral route, resulting in a slower onset of action compared with inhaled bronchodilators. Improvement in the phamacokinetics of oral theophyllines has occurred with the production of long-acting formulations.

The bronchodilator action of theophyllines is relatively limited in patients with COPD. Exercise tolerance in patients with COPD changes little with theophylline treatment, showing no or little improvement. Any improvement in exercise tolerance has been thought to result from an effect on respiratory muscles or a fall in trapped gas volume, but these mechanisms are still the subject of debate. Other non-bronchodilator effects of theophylline, such as improving right ventricular performance or anti-inflammatory actions, are of questionable clinical significance.

Theophyllines have a narrow therapeutic index and patients often experience side-effects within the therapeutic range. Other factors that are common in COPD, such as smoking, hypoxaemia, and infection, all alter theophylline clearance and make the control of theophylline dosage difficult, requiring measurement of plasma theophylline levels (Table 5). The possible beneficial effects of theophyllines have to be balanced against their potential side-effects and the fact that a similar benefit may be achievable with inhaled bronchodilators, hence theophyllines are reserved for patients in whom other treatments have failed to control symptoms adequately.

Combination therapy

Studies of combination therapy are difficult to assess owing to problems of suboptimal dosing. Some studies suggest that drug combinations such as salbutamol and ipratropium or salbutamol and aminophylline produce improvement in exercise tolerance in the face of trivial changes in spirometry. It is unclear whether higher doses of salbutamol could have achieved a similar effect. Thus, combinations of bronchodilator drugs should only be used if single drugs have been tried and have failed to give adequate symptomatic relief. Combination therapy should only be continued if there is good subjective or objective benefit.

Drug delivery devices

Compliance with inhaled treatment is poor. In the Lung Health Study the overall compliance with therapy was 65 per cent. Since many patients with COPD are elderly, the difficulties encountered with standard metered dose inhalers (MDI) are exaggerated. These problems can often be overcome by dry powdered formulations or by a spacer device. However, patients with severe COPD are only able to achieve low inspiratory flow rates, and rates as low as 40 litre/min may cause failure of the one-way valve in a spacer device to open.

Home nebulizer therapy

There is controversy over the use of home nebulizer therapy in patients with COPD. Using end-points such as spirometry and corridor walking exercise, it has been shown that nebulized salbutamol is no more effective in patients with COPD than lower doses of the same drug given through a spacer device. However, patients appear to prefer nebulized bronchodilator therapy. This may be because the total dose of the drug delivered by nebulizer therapy is higher, and the facial cooling that occurs with the nebulized solution itself may have an effect on dyspnoea, independent of any effect on airway calibre.

Acute improvement in spirometry with nebulized bronchodilator therapy does not necessarily predict a long-term response and only a minority of patients are likely to obtain benefit from high-dose bronchodilator therapy. Patients should only be supplied with a nebulizer if they have been fully assessed by a respiratory physician who is able to assess the risk/cost benefit. This assessment should include ensuring that optimal use is made of a simple metered dose inhaler or dry powdered device and that some assessment is made of the patient's response to nebulizer therapy, including a home trial with peak expiratory flow measurements. Dosage regimes must be tailored to individual patient's needs and their side-effects monitored.

Corticosteroids

The chronic inflammation that occurs in the large and small airways provides a rationale for the use of corticosteroids in COPD. However, the use of corticosteroids in patients with COPD remains contentious, particularly the prediction of which patients will respond to this treatment.

Oral corticosteroids

A subgroup of patients respond to corticosteroids. A meta-analysis of trials of oral corticosteroids indicates that a significant improvement in FEV $_1$ (over 15 per cent and greater than 200 ml improvement) occurs in 10 to 20 per cent of patients with clinically stable COPD. There are no reliable predictors of which patients will respond. Furthermore, the response to high doses of oral prednisolone in short-term studies does not necessarily predict continued FEV $_1$ response to long-term inhaled steroids. Data from uncontrolled retrospective studies of oral corticosteroids suggest that long-term treatment may slow the decline in FEV $_1$, although 10 mg of prednisolone per day was required to prevent the decline in FEV $_1$ and patients with asthma may have been included. This treatment cannot be recommended in view of adverse side-effects.

Inhaled corticosteroids

Two large controlled trials of the effects of inhaled corticosteriods in patients with mild COPD have now been published. Both the Copenhagen City Lung study and the EUROSCOP study showed no effect of budesonide at a dose of 800 µg twice daily on the rate of decline in FEV $_1$over a follow-up period of 3 years. In patients with a mean FEV$_1$ of 77 per cent of predicted (i.e. mild airways obstruction) who continued to smoke, the EUROSCOP study showed an initial improvement in FEV $_1$ over the first 6 months (at a rate of 17 ml/year—Fig. 13), which was maintained over the 3-year follow-up period. This was not the case in the Copenhagen study. A third trial in moderate COPD (FEV$_1$ 50 per cent of predicted) in a mixed group of smokers and ex-smokers also showed a similar small improvement in FEV$_1$ over 3 months in a group treated with fluticasone at a dose of 1000 µg/day, but no significant effect on the rate of decline in FEV $_1$. However, there was a significant benefit in health status, and a reduction in exacerbation rates by 25 per cent. Since exacerbations of COPD have an adverse effect on health status, these two effects may be linked.

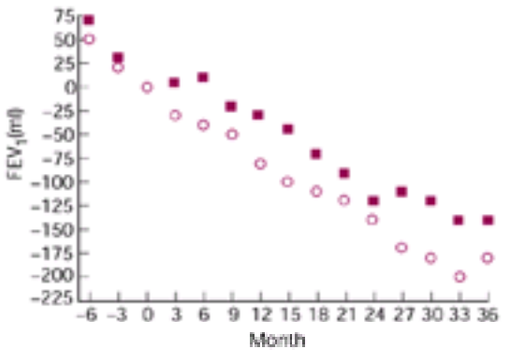

Fig. 13 Effects of budesonide at 800 µg twice daily (closed squares) or placebo (open circles) on the rate of decline in FEV $_1$ in patients with COPD and mild airways obstruction. There is a small decrease in the rate of decline in FEV $_1$ over the first 3 months of treatment in the budesonide but not in the placebo group, but no subsequent difference in the rate of decline in either group.

Based on the results of these large-scale trials there appears to be no effect of inhaled corticosteroids on the prognosis of mild to moderate COPD as assessed by the decline in FEV_1. However, there may be an effect on health status and exacerbation rates in moderate COPD. Inhaled corticosteroids may therefore be of benefit to patients with moderate to severe COPD who have frequent exacerbations, as well as those who show reversibility of their airway obstruction. Further studies are required to distinguish the subpopulation of patients with COPD who show a response to inhaled corticosteriods.

Other therapeutic agents

There is no evidence for the use of anti-inflammatory drugs such as sodium cromoglycate, nedocromil sodium, or antihistamines in patients with COPD. Although used widely in continental Europe, mucolytic drugs are rarely used in the United Kingdom. There is no evidence to support the use of continuous or intermittent prophylactic antibiotics in patients with COPD.

British Thoracic Society guidelines suggest an approach to management of patients with COPD based on the severity of disease (Fig. 14).

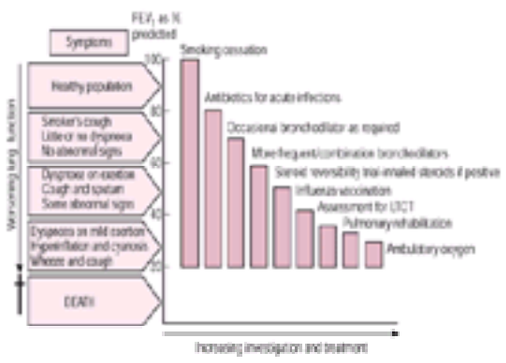

Fig. 14 The chronic obstructive pulmonary disease escalator: as lung function declines the treatments need to be increased. LTOT, long-term oxygen therapy. (Adapted from BTS Guidelines, 1997, *Thorax* **52**, S1–S18.)

Domiciliary oxygen therapy

The only treatment that improves the long-term prognosis in patients with COPD is long-term domiciliary oxygen therapy, given for at least 15 h per day, as shown by two multicentre trials, one conducted by the Medical Research Council (**MRC**) in the United Kingdom, and the other by the nocturnal oxygen therapy trial (**NOTT**). These are discussed in Chapter 17.7.

The reasons for the improvement in survival with oxygen therapy in patients with COPD are still uncertain, but are not clearly related to improvements in pulmonary haemodynamics. As usual, survival is related to the level of pulmonary arterial hypertension in patients who receive long-term oxygen therapy. In the MRC trial, there was no significant improvement in pulmonary arterial pressure following oxygen therapy, but the increase of 3 mmHg per year in pulmonary arterial pressure in the control group did not occur in those who were treated. Overnight oxygen therapy, which abolishes nocturnal desaturation, also decreases pulmonary arterial pressure. However, since the changes in pulmonary haemodynamics produced by long-term oxygen therapy are small, it seems unlikely that these have a major influence on survival.

In addition to the improvement in survival, a number of studies have examined other effects of supplementary oxygen therapy. The effects of oxygen therapy on breathlessness remain unclear, but several studies have shown that oxygen therapy can lead to an improvement in exercise endurance in patients with COPD, associated with a reduction in ventilation at a given submaximal work rate, and an improvement in walking distance and in ability to perform daily activities. Assessment of patients taking part in the NOTT study showed that they have marked disturbances in mood and quality of life: after 6 months of oxygen therapy, 42 per cent showed evidence of an improvement in cognitive function, but little change in mood or quality of life. As in all studies of patients with COPD, the FEV $_1$ is the strongest predictor of survival in patients receiving long-term oxygen therapy, but this does not influence the decline in FEV $_1$.

Long-term oxygen therapy has been shown to affect the polycythaemia that occurs in patients with chronic hypoxaemia, by reducing both the haematocrit and the red cell mass. The clinical relevance of these haematological changes produced by oxygen therapy remains unclear. Continued cigarette smoking should be a relative contraindication to long-term oxygen therapy.

Criteria for the prescription of domiciliary oxygen therapy

There are three forms of domiciliary supplemental oxygen therapy:

1. long-term controlled oxygen therapy for at least 15 h per day in patients with chronic respiratory failure;
2. portable oxygen therapy for exercise-related hypoxaemia; and
3. short-burst oxygen therapy, as a palliative treatment for the temporary relief of symptoms.

The criteria for the prescription of long-term oxygen therapy are based on the clinical parameters of those patients with COPD who showed an improved survival in the two controlled trials of long-term oxygen therapy. Central to the prescription criteria is the demonstration of significant hypoxaemia in a patient with COPD breathing room air, measured when clinically stable (Table 6).

In the United States, long-term oxygen therapy can be prescribed based on pulse oximetry and in patients whose PaO_2 lies between 7.3 and 7.9 kPa, provided there is evidence of cor pulmonale or polycythaemia.

Long-term oxygen therapy is usually prescribed in the United Kingdom in the form of oxygen concentrators, although the majority of home oxygen therapy is given to patients with COPD as cylinder oxygen therapy for the relief of breathlessness. Adherence to the criteria for the prescription of long-term oxygen therapy is less than

optimal, applying to around 40 per cent of patients. Data from the NOTT study showed that 43 per cent of patients who were initially shown to fit the criteria for long-term oxygen therapy were no longer eligible when reassessed 4 weeks later. It is therefore essential that clinical stability is demonstrated, with no exacerbation of COPD for at least 4 weeks, before a decision is made to prescribe long-term oxygen therapy, and that other treatments such as bronchodilators and inhaled steroids are optimized before the prescription of long-term oxygen therapy. Furthermore, reassessment is recommended to ensure that the patient remains significantly hypoxaemic and still fits the criteria for long-term oxygen therapy and to ensure that adequate oxygenation is achieved while breathing oxygen. A PaO_2 of 8 kPa is desirable, and this can usually be achieved by nasal prongs at flow rates between 1 and 3 litre/min. Precipitation of increasing hypercapnia with long-term oxygen therapy is seldom a problem in clinically stable patients. Long-term oxygen therapy should be prescribed continuously during sleeping hours, which prevents episodes of oxygen desaturation at night, and improves sleep quality.

Portable oxygen therapy

There are no established criteria for the prescription of portable oxygen in the form of small light-weight cylinders or liquid oxygen therapy (available in the United States and continental Europe, but not currently in the United Kingdom). Patients who desaturate during exercise by 5 per cent or more may be suitable for this treatment, although exercise capability may improve irrespective of arterial oxygen desatuation. The use of portable oxygen therapy to enhance mobility should be encouraged in patients to help to prevent the downward spiral of immobility and physical deconditioning.

Controlled oxygen is typically delivered by means of nasal prongs, or by mask in patients who are intolerant of nasal cannulas owing to local irritation and dermatitis. Patient compliance with masks is generally less than with nasal prongs. In patients in whom there is refractory hypoxaemia, oxygen can be delivered by the trans-tracheal route. This can reduce the resting flow rate requirements by 25 to 50 per cent compared with nasal cannulas, resulting in considerable savings, particularly if liquid oxygen is the supply mode. However, there are complications, including the formation of mucus balls in 25 per cent of cases, cough, infection, and catheter dislodgement. Reservoir devices have also been developed to reduce total oxygen requirement and cost: these work on the basis that the reservoir fills during the patient's exhalation and supplies oxygen only during inspiration.

Travel

Commercial aircraft cabins are pressurized to the equivalent of an altitude of no greater than 2600 m, producing a cabin oxygen tension of around 100 mmHg. Worsening hypoxaemia may exacerbate the symptoms of breathlessness, particularly in patients who are already hypoxaemic with a PaO_2 less than 8 kPa, in which case the airline should be contacted by letter by the patient's respiratory physician, recommending the use of oxygen: most will provide oxygen throughout the flight.

Pulmonary rehabilitation

The aim of pulmonary rehabilitation is to prevent or reverse the deconditioning that occurs with lack of exercise and immobility due to dyspnoea, thereby allowing the patient to cope with his/her disease. Before considering rehabilitation, it is vital that investigations and therapy are directed towards any reversible component of the airflow limitation and that this treatment is optimized. Patients with moderate to severe COPD should be considered for pulmonary rehabilitation programmes and each rehabilitation programme should be tailored to fit individual patient's needs, depending on the factors that are deemed to limit exercise.

Pulmonary rehabilitation programmes

Establishing a pulmonary rehabilitation programme requires a multidisciplinary approach and appropriate health-care resources. Exercise training programmes have taken two approaches. The first is to attempt to improve cardiorespiratory fitness by aerobic exercises of 20 to 30 min duration at least three times per week. However, patients with COPD may be unable to achieve the necessary increase in oxygen uptake to produce the required 'training effect' because of breathlessness, hence this approach is usually restricted to those with mild to moderate exercise limitation. The second approach, used in patients unable to sustain sufficient exercise to improve anaerobic fitness, aims to improve mobility. This can be achieved by providing regular exercise sessions so that the patient works to his/her maximum tolerable ventilatory limit. In patients with very severe COPD there are no established guidelines for pulmonary rehabilitation programmes, but carefully supervised exercise conditioning in the hospital setting, with oxygen supplementation, should be considered in those who develop hypoxaemia during exercise.

Respiratory muscle training

A meta-analysis of 17 randomized trials of respiratory muscle training in patients with COPD showed that although training, using either resistance breathing or isocapnic hyperventilation, improved respiratory muscle strength and endurance, there was no overall improvement in exercise tolerance.

Numerous *in vitro* studies have shown that methylxanthines, such as theophyllines (in doses not possible in humans), potentiate the response of fresh and fatigued muscle strips to an electrical stimulus. Studies *in vivo* in humans have been less convincing, reporting either little or no improvement in the transdiaphragmatic pressure (*P*di) generated during maximal voluntary effort.

Some uncontrolled studies have suggested that resting the respiratory muscles, by long-term nocturnal use of either negative pressure, applied to the chest wall, or intermittent positive pressure ventilation (IPPV) using a nasal mask, results in some improvement in respiratory muscle function. However, a large controlled study failed to show any benefit on respiratory muscle function in patients with COPD.

The results of controlled trials on the effects of nutritional supplementation on respiratory muscle function in malnourished patients with COPD have shown no consistent benefit. Those studies that have achieved positive results have done so in association with an increase in weight.

Controlled breathing techniques have been used as part of a pulmonary rehabilitation programme to diminish breathlessness by training patients to breathe efficiently. This treatment aims to:

1. restore the diaphragm to a more normal position and function;
2. decrease the respiratory rate by using a breathing pattern that diminishes air trapping;
3. diminish the work of breathing; and
4. reduce dyspnoea and patient anxiety.

Techniques such as pursed lip breathing have been employed and some studies have shown an improvement in blood gases.

The outcome of a pulmonary rehabilitation programme is usually assessed by measuring improvement in lung function or exercise tolerance, but benefit may not always be apparent in these variables. In general, studies show a favourable effect on exercise tolerance and a reduction in symptoms such as breathlessness during exercise. Since social factors contribute to the disability in COPD, assessment of quality of life should be included in a rehabilitation programme (measured by a health profile questionnaire), as should that of compliance, longevity, and cost–benefit.

Other aspects

Nutrition

Weight loss is common in patients with COPD, particularly those with severe airways obstruction, and is associated with high mortality. However, studies that have addressed this issue have produced variable results, although those who do show weight gain may have improved survival. Obesity should be discouraged by appropriate dietary advice in patients with COPD to avoid additional strain on the cardiorespiratory system.

Depression

Mood disturbances, particularly depression, are very common in patients with advanced disease and often contribute to an enhanced perception of symptoms, particularly breathlessness, and to social isolation. Antidepressant drugs can often produce encouraging results in these patients.

Vaccination

Influenza vaccination is recommended for patients with COPD, although specific evidence is lacking. The rationale relates to other studies in elderly patients, not specifically with COPD, where a 70 per cent reduction in mortality from influenza can be demonstrated.

Management of acute exacerbations of COPD

Exacerbations of COPD occur on a background of established disease and are amongst the commonest acute respiratory problems presenting to either primary or secondary care. Many patients can be managed in the community.

Antibiotics

Infection is a common precipitating feature in exacerbations of COPD, although only 50 per cent of patients with severe exacerbations with associated respiratory failure will have a positive sputum culture for a bacterium. The commonest organisms are *Haemophilus influenzae*, *Streptococcus pneumoniae*, and *Moraxella catarralis*. However, patients with COPD are often chronically colonized with common bacterial pathogens, hence culture of one of these organisms during an acute exacerbation does not imply that this organism is responsible for the exacerbation. Viral infections have been shown to be responsible for up to 30 per cent of all exacerbations of COPD.

There is limited information from controlled trials on the effects of antibiotics in exacerbations of COPD. In a trial of 173 patients with 362 exacerbations of COPD, patients received either a 10-day course of sulphamethoxazole, amoxycillin, doxycycline, or placebo: relief of symptoms within 21 days was achieved in 68 per cent of the antibiotic-treated group and in 55 per cent of the placebo-treated exacerbations. Peak expiratory flow recovered faster with antibiotics, although the differences were small, and treatment failures were twice as common with placebo. The difference in successful outcome between antibiotic and placebo were significant if two of the following symptoms were present—increase in dyspnoea, increase in sputum volume, and increase in sputum purulence: antibiotics are recommended if two of these symptoms are present.

In view of the limited range of bacteria present in the sputum of these patients, broad-spectrum antibiotics such as amoxycillin at a dose of 250 mg three times daily or clarithromycin at 250 to 500 mg twice daily—as an alternative in patients with penicillin allergy—are recommended. Prescription of antibiotics should take into account local bacteriological sensitivity patterns, particularly the prevalence of b-lactamase-positive *H. influenzae*, which is around 20 per cent in most areas, and *Moraxella catarralis*, of which 90 per cent are b-lactamase positive. If the patient is known to have had b-lactamase -positive organisms previously in sputum, or fails to respond to amoxycillin, then co-amoxiclav should be considered. Antibiotics should be given orally unless there is a specific indication for intravenous treatment.

Bronchodilators

The use of nebulized bronchodilators in acute exacerbations of COPD is recommended. These should be given as soon as possible on admission and at 4- to 6-hourly intervals thereafter, or more frequently if required. In patients with COPD, particularly in those with an elevated $PaCO_2$, the nebulizer should be driven by compressed air and not by oxygen, to avoid a further rise in $PaCO_2$. Oxygen can be given by nasal prongs at 1 to 2 litre/min during nebulization. b-Agonists (salbutamol at 2.5 to 5 mg, or terbutaline at 5 to 10 mg) or an anticholinergic drug (ipratropium bromide at 0.5 mg) are the drugs commonly used. In acute exacerbations of COPD, no difference has been shown between these drugs given alone or in combination in nebulized form.

A response to a nebulized bronchodilator in an acute exacerbation does not imply long-term benefit and assessment for a home nebulizer should be made when the patient is in a stable condition. Several studies have shown no difference in the degree of bronchodilatation achieved when the same dose of bronchodilator is given by a metered dose inhaler, with or without a spacer device, or via a nebulizer, even in patients with an acute exacerbation of airways obstruction. However, patients with respiratory failure have been excluded from these studies and hence nebulized bronchodilators are still recommended, but in most cases these should only be necessary for 24 to 48 h and a change to a metered dose inhaler, or a dry powder device should be made 24 to 48 h before discharge.

If a patient is not responding to nebulized bronchodilators during an exacerbation, then intravenous methylxanthines should be considered. However, a small randomized placebo-controlled trial of intravenous aminophylline showed no differences in spirometry, arterial blood gases, or the sensation of dyspnoea between the aminophylline and placebo groups over a period of 72 h following admission with exacerbation of COPD. Thus, the prescription of theophyllines has no clear role in management of acute exacerbations of COPD and the possible benefits should be weighed against the side-effects, particularly in patients with COPD who have hypoxaemia, infection, and are receiving antibiotics, all of which can affect theophylline clearance. Thus the dose must be carefully individualized and the serum level maintained within a narrow therapeutic range (10 to 20 mg/l). The usual loading dose is 6 mg/kg of aminophylline with maintenance dosage of 0.5 mg/kg.h.

Corticosteroids

There are now several controlled trials showing benefit of oral corticosteroids in patients with acute exacerbations of COPD. A placebo-controlled study in hospital patients without hypercapnic respiratory failure showed improvement in FEV_1 and reduction in days in hospital in those treated with 30 mg prednisolone daily. A further study of exacerbations treated with prednisolone in the community also showed a positive result.

There is therefore good evidence that corticosteroids are beneficial in exacerbations of COPD, the usual regime being 30 mg prednisolone daily for 2 weeks. The lowest dose that produces benefit requires further study.

Diuretics

In patients with fluid retention as a result of respiratory failure and cor pulmonale, diuretics should be used judicially, as they have the potential to reduce right ventricular end-diastolic volume considerably and hence cardiac output.

Anticoagulants

Pulmonary emboli are probably are under-recognized in severe COPD. It is difficult to diagnose pulmonary emboli in such patients: ventilation/perfusion abnormalities are often present and can lead to false-positive reports of pulmonary thromboembolic disease. Prophylactic subcutaneous heparin is often given to patients with exacerbations of COPD, particularly those who have respiratory failure.

Physiotherapy

There is very little evidence to support the use of physiotherapy to improve expectoration in patients with acute exacerbations of COPD, although some studies suggest that there is some benefit in patients producing large amounts of sputum.

Assessment of recovery from acute exacerbations of COPD

Arterial blood gases should be checked while breathing air, which gives a guide to the need for later formal reassessment for long-term oxygen therapy. Antibiotics need not usually be given for more than 7 days. Respiratory failure in COPD is dealt with in Chapter 17.7.

Surgical treatment in COPD

Bullous emphysema

Exertional dyspnoea is the usual presenting feature in patients with bullous disease, although a single bulla of moderate size is unlikely to produce symptoms when the remaining lung is normal. Bullae may present as a chance finding on a chest radiograph or as a pneumothorax. Occasionally they become infected, in which case

there may be a fluid level, sometimes with surrounding consolidation. Such infection may result in closure of the bronchial connection, shrinkage, or even obliteration of the bulla.

Respiratory function tests may be non-specific and simply reflect COPD. Almost always there is some degree of airway obstruction, which may result from concomitant diffuse emphysema or airways disease, or as a result of the loss of lung elastic recoil that accompanies large bullae. Overinflation is typically present, but is underestimated if measured by the helium dilution technique rather than by plethysmography. Gas exchange is usually impaired as shown by a reduced TL_{CO}. The K_{CO} may reflect the quality of the non-bullous lung if the bullae are non-ventilating, which may be helpful in making a decision concerning surgery.

Treatment

The only treatment that is considered for large bullae is surgical obliteration. This should not be offered to patients who are asymptomatic, since the operation does have an appreciable risk. The principal indication is progressive dyspnoea, but in those with airflow limitation it has been difficult to determine which patients benefit from bullectomy. Many techniques have been used in the past to assess patient's suitability for this procedure, such as bronchography and pulmonary angiography, which have now been essentially replaced by CT scanning. A critical feature is the quality of the non-bullous lung: airflow limitation is determined by the degree of emphysema in the non-bullous lung rather than the extent of the bullous disease. Quantitative perfusion scanning may demonstrate retained perfusion in collapsed peribullous lung, which may improve after operation. Patients with bullae occupying less than 50 per cent of the hemithorax, with an FEV_1 of less than 1 litre, or hypercapnia carry a high risk of a poor response to surgery.

The aims of surgery are to obliterate the bullous space and restore the elastic integrity of the lung. Several techniques have been described, including excision, plication, marsupialization, and intracavity drainage. Most operations are performed by a conventional lateral thoracotomy, but superficial bullae have also recently been dealt with using thoracoscopic and laser techniques. The perioperative mortality in published series ranges from 0 to 20 per cent in patients with a wide range of disability and hence operative risk.

The best functional results occur in younger patients with mild symptoms, with large bullae, relatively well-preserved pulmonary function, and normal surrounding lung. Those with small bullae (less than 1 litre or less than 50 per cent of the hemithorax), poor overall lung function (FEV_1 less than 1 litre), and diffuse emphysema have least functional improvement and in these the improvement has to be weighed against the risk of surgery. Studies of the long-term follow-up of patients after surgery indicate that giant bullae do not recur.

Lung transplantation

It was originally considered that patients with endstage COPD were not suitable for a single lung transplant since perfusion would be preferentially directed towards the transplanted lung because of its lower pulmonary vascular resistance, producing profound ventilation/perfusion mismatch. The current success of a single lung transplant, despite the presence of abnormal mechanics in the native lung, is due in part to improved patient selection, lung preservation, and anaesthetic management. There are problems if residual infection is present in the native lung, and large bullae in the native lung may show gross hyperinflation in the early postoperative period, causing mediastinal shift and compression of the transplanted lung. Hence those patients with recurrent pulmonary infection or bilateral large bullae are considered for heart–lung transplantation or bilateral sequential lung transplantation. For detailed discussion of lung transplantation, see Chapter 17.16.

Lung volume reduction surgery

The growing number of patients with emphysema on waiting lists for lung transplantation has led to a recent re-examination of previous surgical techniques that might give symptomatic relief, particularly the technique of lung volume reduction surgery (pneumonectomy or pneumoplasty). The rationale for this technique is to reduce the volume of overinflated emphysematous lung by 20 to 30 per cent, with the aim of improving the elastic recoil of the lungs, diaphragm configuration, chest wall mechanics, and gas exchange. Persistent postoperative air leaks were overcome by the use of strips of bovine pericardium to buttress the stapling line. The technique is usually performed via a median sternotomy, without the need for cardiopulmonary bypass. Thoracoscopic laser pneumoplasty has been developed as an alternative technique to the more conventional excisional surgery. Careful patient selection is necessary on the basis of a distended thorax, predominantly upper lobe disease demonstrated by CT scanning, and severe functional disability in spite of a programme of pulmonary rehabilitation.

The early results are encouraging: improvements in lung function that have occurred up to 6 months after surgery are impressive and better than can be produced by conventional medical treatment with bronchodilators or corticosteriods. Controlled trials of this technique are underway but have not yet been published.

A recent study compared lung volume reduction surgery with single lung transplant for emphysema: disease severity was greater in the lung transplant group, and their increase in FEV_1 and FVC was greater, but the increase in 6-min walking distant was similar in both groups.

Many questions concerning this technique require answers from future studies, particularly knowledge of the duration of the benefit, the best selection criteria for patients, and the mechanism of the improvement.

Further reading

American Thoracic Society (1995). Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine* **152**, S77–S120.

Anthonisen NR (1994). The Lung Health Study; effects of smoking intervention and the use of an inhaled anticholinergic bronchodilator on the rate of decline of FEV_1. *Journal of the American Medical Association* **272**, 1497–505.

British Thoracic Society (1997). British Thoracic Society guidelines for the management of chronic obstructive pulmonary disease. *Thorax* **52**(Suppl 5), S1–S28.

Burrows B (1991). Predictors of loss of lung function and mortality in obstructive lung disease. *European Respiratory Journal* **1**, 340–5.

Calverley P, Pride N (1996). *Chronic obstructive pulmonary disease*. Chapman & Hall, London.

Davies L, Calverley PMA (1996). Lung volume reduction surgery in chronic obstructive pulmonary disease. *Thorax* **51**(Suppl 2), S29–S34.

Jeffrey PK (1996). Bronchial biopsies and airway inflammation. *European Respiratory Journal* **9**, 1583–7.

Lange P *et al*. (1990). The relation of ventilatory impairment and of chronic mucus hypersecretion to mortality from obstructive lung disease and from all causes. *Thorax* **45**, 579–85.

MacNee W (1994). State of the Art: pathophysiology of cor pulmonale in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine* **150**, Part I—833–52, Part II—1158–68.

MacNee W (2000). Chronic bronchitis and emphysema. In: Seaton A, Seaton D, Leitch AG, eds. *Crofton and Douglas's respiratory diseases*, 5th edn, pp.616–95. Blackwell Science, Oxford.

MacNee W (2000). Respiratory failure. In: Seaton A, Seaton D, Leitch AG, eds. *Crofton and Douglas's respiratory diseases*, 5th edn, pp.696–717. Blackwell Science, Oxford.

Medical Research Council Working Party (1981). Long term domiciliary oxygen therapy in chronic cor pulmonale complicating chronic bronchitis and emphysema. *Lancet* **i**, 681–6.

Nagai A *et al.* (1985). The National Institutes of Health Positive-Pressure Breathing Trial: Pathology Studies. I. Inter relationship between morphologic lesions. *American Review of Respiratory Diseases* **132**, 937–45.

Nagai A, West WM, Thurlbeck WM (1985). The National Institutes of Health Intermittent Positive-Pressure Breathing Trial: Pathology Studies. II. Correlations between morphologic findings, clinical findings and evidence of expiratory airflow obstruction. *American Review of Respiratory Diseases* **132**, 946–53.

Nocturnal Oxygen Therapy Trial Group (1980). Continuous or nocturnal oxygen therapy in hypoxemic chronic obstructive pulmonary disease: a clinical trial. *Annals of Internal Medicine* **93**, 391.

Pauwels RA *et al.* for the European Respiratory Society Study on Chronic Obstructive Pulmonary Disease (1999). Long-term treatment with inhaled budesonide in persons with mild chronic obstructive

pulmonary disease who continue smoking. *New England Journal of Medicine* **340**, 1948–53.

Saetta M (1997). Airway pathology compared with asthma. *European Respiratory Review* **45**, 211–15.

Siafakas NM *et al*. (1995). ERS Consensus Statement. Optimal assessment and management of chronic obstructive pulmonary disease (COPD). *European Respiratory Journal* **8**, 1398–420.

Vestbo J, Prescott E, Lange P and the Copenhagen City Heart Study Group (1996). Association of chronic mucus hypersecretion with FEV_1 decline and chronic obstructive pulmonary disease morbidity. *American Journal of Respiratory and Critical Care Medicine* **153**, 1530–5.

Vestbo J *et al*. (1999). Long-term effect of inhaled budesonide in mild and moderate chronic obstructive pulmonary disease: a randomised controlled trial. *Lancet* **353**, 1819–23.

17.7 Chronic respiratory failure

P. M. A. Calverley

Introduction

Although respiration is ultimately a biochemical process involving the generation of ATP, the term respiratory failure is used more loosely to describe the failure of gas exchange within the lung to maintain arterial blood gas homeostasis. Defining normal blood gas tensions is harder than it may appear initially as P_aO_2 falls with age and the extent of this is debated. The most commonly applied formula to describe this is:

$$P_aO_2 \text{ (kPa)} = 13.86 - [0.036 \times \text{age (years)}]$$

Thus a P_aO_2 of 10.6 kPa may be abnormal in a man of 24 years but a 'normal' value in a woman of 80. Subnormal levels of arterial oxygenation are described as hypoxaemia, whilst arterial CO_2 tensions, which do not show similar age dependence, are considered to be hypercapnic when they exceed 6.0 kPa (45 mmHg).

Respiratory failure is defined primarily in terms of hypoxaemia and is arbitrarily considered to be present when the arterial P_O_2 (at sea level) is less than 8.0 kPa (60 mmHg). It need not be accompanied by hypercapnia, but when this develops it leads to acidosis due to the accumulation of carbonic acid by the Henderson–Hasselbalch equilibrium. If the acidosis is not rapidly progressive, and in the presence of intact renal compensatory mechanisms that generate bicarbonate ions, it becomes 'chronic'—a compensated state where the arterial pH returns to normal.

In summary, chronic respiratory failure describes a clinical state when the arterial P_O_2 breathing air is less than 8.0 kPa, which may or may not be associated with hypercapnia, but is accompanied by a normal arterial pH and has been present for several days or more. This definition emphasizes the physiological determinants of gas exchange that characterize the problem.

Unlike other forms of organ system failure, such as cardiac or hepatic failure, the clinical symptoms and signs of chronic respiratory failure are relatively undramatic, but its development is equally significant, both as a marker of disease progression and in producing serious complications beyond those normally seen with the underlying disease. This chapter will review the causes, clinical features, and assessment of chronic respiratory failure as well as specific means of treatment. However, to do so logically requires some understanding of the principles underlying the development of this condition, as well as the factors relevant to the selection of the threshold values used in defining this state.

Physiological determinants of blood gas tensions

In health there is a predictable fall in the partial pressure of oxygen from that in the room air to that in mixed venous blood. This reflects the effect of diluting room air with resident gas in the alveoli, the efficiency of pulmonary oxygen exchange, and the consumption of oxygen by metabolizing tissues. Conversely, there is a predictable increase in the amount of CO_2 added to the circulation and subsequently removed from the lungs during expiration. This simple system is reliant on a range of physical processes that differ somewhat for O_2 and CO_2. Within the lungs gas transport is largely by convective bulk transport and in the alveoli by diffusion. In the blood oxygen combines with haemoglobin, which augments transportation to the tissues where diffusion is the final process involved. By contrast, CO_2 transport begins with diffusion from relatively high tissue concentrations and is buffered in solution in the blood. This complex mechanism can be deranged in a number of predictable ways that are discussed below.

In the last 20 years the analysis of pulmonary gas exchange has been revolutionized by the use of the complex multiple inert gas elimination technique in research laboratories around the world. This gives a relatively complete description of the distribution of gas exchange abnormalities within the lungs. However, for an understanding of the general principles involved in disease states the traditional three-compartment model is easier to follow. This assumes that alveolar air within the lungs is either ideally matched to pulmonary arterial blood flow within the pulmonary capillary bed or is totally mismatched, meaning that either the ventilation–perfusion ratio is unity, that is, ventilation without perfusion (physiological dead space, V_D), or this ratio is zero, that is, perfusion without ventilation (venous admixture effect). The physiological dead space includes a component due to dilution of the resident gas in the airways, the anatomical dead space, while the shunt fraction incorporates the very small amount of cardiac output (less than 1 per cent) not passing through the pulmonary capillary bed.

Hypoxaemia

The principal mechanisms leading to arterial hypoxaemia are shown in Table 1. Individuals resident at altitude, for example the high Andes and Himalayas, experience significantly lower inspired oxygen tensions than those at sea level and even individuals with normal lungs can develop clinically significant hypoxaemia, especially during sleep. Even minor degrees of respiratory impairment in these circumstances can produce dramatic changes in blood gas tensions and the early onset of cor pulmonale. Conversely, people with established hypoxaemia at sea level can occasionally experience worsening symptoms when travelling by air where cabin pressurization is 75 per cent of atmospheric. However, in clinical practice, this is relatively infrequent.

Four processes cause arterial hypoxaemia due to inefficient pulmonary gas exchange:

- ventilation–perfusion (V/Q) mismatch
- hypoventilation
- diffusion limitation
- true shunt.

Much the most important of these is ventilation–perfusion mismatching. In many diseases where minute ventilation is increased, the additional inspired gas is distributed to well-perfused areas of the lungs, but when the opposite occurs and perfusion exceeds effective ventilation (low V/Q states), arterial P_aO_2 falls. At first this might seem surprising as most diseases associated with ventilation–perfusion imbalance are of patchy distribution and compensation from areas of high V/Q ratios might be expected. However, this does not occur owing to an important feature of the oxyhaemoglobin dissociation curve (Fig. 1), whose sigmoid shape means

that well-perfused parts of the lung cannot increase the arterial oxygen saturation of the blood leaving them beyond 100 per cent, hence the saturation of the pulmonary venous blood must fall as long as low *V*/*Q* areas are present.

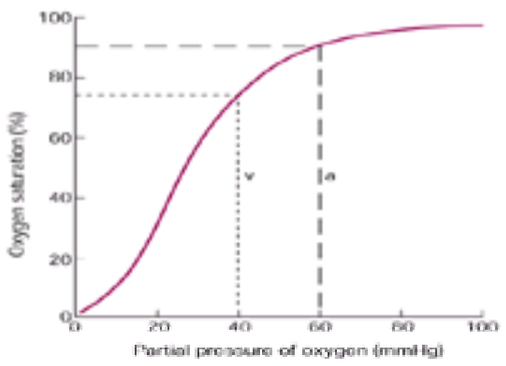

Fig. 1 The haemoglobin–oxygen dissociation curve. a, partial pressure of oxygen of 8 kPa (60 mmHg), which is the definition of arterial hypoxia. v, partial pressure of oxygen of 5.3 kPa (40 mmHg), which is typical of mixed venous blood. Note that once the PaO_2 falls below 8 kPa small further falls dramatically decrease the arterial oxygen saturation.

The second important mechanism contributing to arterial hypoxaemia is alveolar hypoventilation, where the supply of fresh oxygen is globally reduced because of generally inadequate minute ventilation. This process often coexists with ventilation–perfusion mismatching and tends to exacerbate it. In some situations, such as during exercise, total minute ventilation may lie within the normal range but can still be inappropriately low for the subject's metabolic requirements, thereby leading to hypoxaemia.

Two less important mechanisms are anatomical shunting and diffusion limitation. The former occurs predominantly with intrapulmonary arteriovenous malformations. Congenital cardiac anomalies such as ventricular septal defects with reversed flow are often lumped in with this problem, although technically they are extrapulmonary in origin. The failure to increase PaO_2 to greater than 40 kPa (300 mmHg), even when exposed to 100 per cent oxygen, is diagnostic. Diffusion limitation was initially believed to be important in many diseases, the assumption being that passive diffusion of oxygen was reduced to the point where equilibration with haemoglobin during red cell transit of the pulmonary capillaries was incomplete. Detailed studies with modern techniques of gas exchange analysis have largely discredited this, except for small falls in arterial oxygen tension at maximum levels of performance in elderly athletes and possibly during exercise in some forms of interstitial lung disease.

Although not the sole explanation of arterial hypoxaemia, the degree of hypoxaemia can be worsened when the mixed venous arterial oxygen tension is significantly reduced as occurs in low cardiac output states or conditions where peripheral oxygen consumption is increased.

Hypercapnia

Analysis of the pulmonary causes for changes in arterial CO_2 tension is much simpler, the relevant relationship being:

$$PaCO_2 = K \times VCO_2/V_A$$

where VCO_2 is the CO_2 production by the body, V_A is the alveolar ventilation, and K is a constant.

It is easy to see that inadequate alveolar ventilation, due to either low total alveolar ventilation or an inability to increase V_A in response to an increase in metabolic CO_2 production, will increase the arterial CO_2. Alveolar ventilation is influenced by a range of factors, reflecting the balance of the intrinsic capacity of the ventilatory pump and the demands placed on it (Fig. 2).

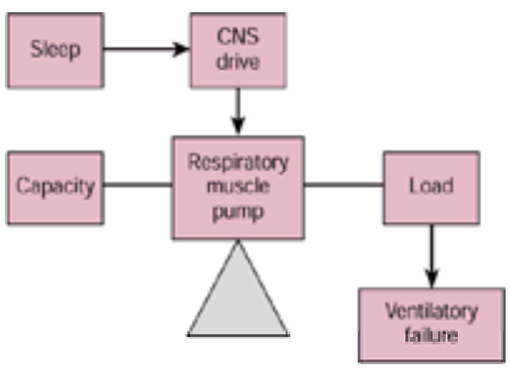

Fig. 2 Alveolar ventilation reflects the balance of the intrinsic capacity of the ventilatory pump and the demands placed on it. A reduced respiratory drive, particularly during sleep, reduces alveolar ventilation but does not produce significant hypercapnia.

The second important mechanism is ventilation–perfusion abnormality, although here the important component is the increased physiological dead space. This can be seen by a rearrangement of the equation above as shown below:

$$PaCO_2 = K \times VCO_2/ V(1–V_D/V_T)$$

where V_D/V_T is the ratio of the physiological dead space to the tidal volume and V is the total minute ventilation.

An increase in V_D occurring when ventilation–perfusion ratios are high can lead to an increase in CO_2 tension. Rather surprisingly, low ventilation–perfusion units are much less important in producing CO_2 retention than they are in producing hypoxia since CO_2 transport from the blood to the alveolar gas is linear (Fig. 3). This means that in areas of normal ventilation–perfusion ratios an increase in overall minute ventilation will increase CO_2 elimination and compensate for the CO_2 that is not excreted from areas of reduced perfusion.

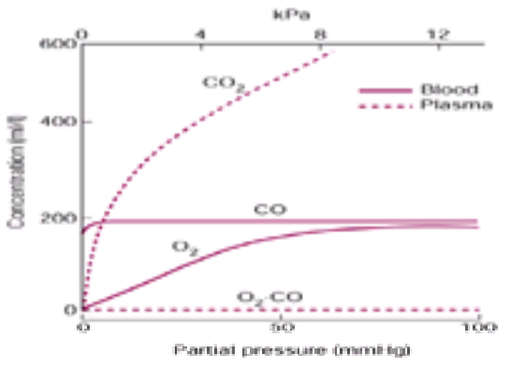

Fig. 3 Concentration of oxygen (O_2), carbon monoxide (CO), and carbon dioxide (CO_2) in blood and plasma at differing partial pressures of these gases.

In most cases of chronic respiratory failure with CO_2 retention, both of these processes operate and the patient is unable to sustain the high overall levels of ventilation needed to maintain CO_2 tensions within the normal range. An important compensatory mechanism in the trade-off between the increased chemical drive to breathing and the mechanical limitations on ventilation is the breathing pattern. In both chronic obstructive and restrictive lung disease a rapid shallow breathing pattern is adopted to minimize respiratory discomfort whilst maintaining minute ventilation. However, the relative fall in tidal volume further worsens the V_D/V_T ratio and can itself contribute to CO_2 retention. Some of these problems are resolved when the buffering capacity of the blood rises as compensation for respiratory acidosis occurs.

Special circumstances

As already noted, residence at altitude and exercise pose particular problems for gas exchange and may induce temporary respiratory failure. There is now a wealth of data indicating that similar changes can occur during sleep. All healthy people show an approximately 15 per cent reduction in minute ventilation in the transition from wakefulness to stable non-REM (rapid eye movement) sleep, and this may be greater still in phasic REM sleep. The ventilatory responses to both hypoxia and hypercapnia decline as sleep deepens and upper airway resistance rises, especially in those who snore. Despite this the blood gas tensions vary little in health during sleep, but dramatic abnormalities can develop during periods of repetitive upper airway obstruction (see Chapter 17.8.2) or when coexisting neuromuscular weakness leads to excessive dependence on muscle groups whose activity declines with sleep (see below).

Gas transport to the tissues

Oxygen delivered to the tissues depends on the oxygen saturation of arterial blood (SaO_2), the haemoglobin concentration (**Hb**), and the cardiac output (**C.O.**), related as follows:

$$\text{Oxygen delivery } (DO_2) = \text{C.O.} \times (\text{Hb} \times 1.34) \times (SaO_2/100)$$

This is influenced only indirectly by the effectiveness of gas exchange. Since oxygen delivery is the clinically relevant outcome of oxygenation, decisions about when and how much to intervene therapeutically will be influenced by this variable. Small changes in saturation become clinically more important in individuals with impaired cardiac function and/or reduced haemoglobin concentration, and a higher SaO_2 should be maintained. In general, there is little to be gained by increasing SaO_2 to the high 90s, especially as this may cause secondary carbon dioxide retention in some diseases. As is clear from Fig. 1, desaturations below 90 per cent only occur when the arterial oxygen tension is below 8.0 kPa (60 mmHg) and this is also influenced by a number of other factors that determine the position of the dissociation curve (see Table 2). This provides the rationale for the choice of 8.0 kPa as the cut-off point for the onset of respiratory failure.

Causes of chronic respiratory failure

The principal causes of chronic respiratory failure are summarized in Table 3. This list is extensive, but the commonest causes are discussed below.

Chronic airflow limitation

This term covers the most important cause of chronic respiratory failure, chronic obstructive pulmonary disease (**COPD**), but is also relevant to diseases such as chronic bronchial asthma, which is now excluded from the definition of COPD, and bronchiectasis, where airflow obstruction is a frequent finding as the disease advances. In all these cases there is a reduction in the forced expiratory volume in 1 s (**FEV_1**) to forced vital capacity (**FVC**) ratio below 70 per cent and a reduction in the FEV_1, which is commonly below 35 per cent predicted before chronic respiratory failure is noted clinically.

In all these cases, hypoxaemia is the first abnormality, which is largely due to ventilation– perfusion mismatching. Early attempts at relating these changes to structural patterns of airway and alveolar disease in COPD have proved unsuccessful. As lung mechanics worsen (commonly when FEV_1 falls below 1.5 litres or 35 per cent of the predicted value), arterial CO_2 increases. This has been related to the development of inspiratory threshold loading (PEEPi) with the onset of chronic hyperinflation, but the degree of CO_2 retention varies between subjects suggesting that individual variations in chemoresponsiveness/perception of ventilatory load contribute to this process. There is no predictable relationship between the severity of impaired lung mechanics below the thresholds indicated and the degree of hypoxaemia or hypercapnia, and many patients who maintain arterial CO_2 tensions within the normal range develop acute CO_2 retention during exacerbations of their disease. These changes can be relatively short lived and the hypercapnia resolves by the time of discharge. Coexisting left ventricular impairment reduces cardiac output and increases venous admixture, which can cause severe hypercapnia and acidosis, which none the less respond rapidly to appropriate treatment.

Patients with COPD in association with persistent hypercapnic respiratory failure have a worse prognosis than those with intermittent hypercapnia during exacerbations (Fig. 4). The pattern in chronic asthma and bronchiectasis appears similar to COPD indicating that lung mechanics rather than individual pathology dictates the severity of the gas exchange disorder.

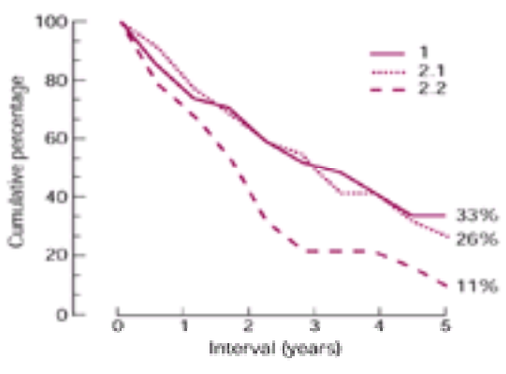

Fig. 4 Survival after index admission in three groups of COPD patients with similar initial spirometry. Group 1 never exhibited CO_2 retention, group 2.1 retained CO_2during the admission but this resolved, while group 2.2 had persistent arterial hypercapnia. (Based on data from Costello *et al*. 1997.)

Interstitial lung disease

Despite the wide range of primary pathologies covered by the term interstitial lung disease they present with a relatively stereotyped physiological picture. A restrictive physiological disorder (FEV_1/FVC greater than 75 per cent with a reduced absolute FEV_1 and FVC) is usual, although patients with sarcoidosis commonly show airways involvement and can present with severe airflow limitation or a mixed physiological pattern. Near normal spirometry can be seen with significant exercise limitation and exercise-induced oxygen desaturation in some patients where COPD and interstitial lung disease coexist. Typically, resting gas exchange is relatively preserved in interstitial lung disease until late in the course of disease, whereas exercise-induced desaturation is an early finding, often seen when spirometric changes are unimpressive. Studies using the multiple inert gas technique have described a bimodal pattern of ventilation–perfusion distribution, with some areas of lung having normal *V*/*Q* relationships and others relatively little ventilation, that is, increased physiological shunting. This situation worsens during exercise. A small number of patients with severe interstitial lung disease develop CO_2 retention in the terminal phase of their illness and cor pulmonale. The physiological mechanisms underlying this are poorly studied but are probably similar to those in COPD.

Chest wall and neuromuscular disease

Here the underlying lung structure and potential for gas exchange are unimpaired, but the ability to maintain adequate alveolar ventilation is reduced. This can be due to increased chest wall stiffness as in kyphoscoliosis, or reduced inspiratory muscle force as in neuromuscular disease. In this latter group the reduction in maximum inspiratory pressure can be global, such as in Duchenne muscular dystrophy, or more specific, such as isolated diaphragmatic weakness, where gas exchange abnormalities may only be present during specific sleep stages. Significant abnormalities of gas exchange at rest only occur with advanced disease and not in every patient. Alveolar hypoventilation is the dominant mechanism of both hypoxaemia and hypercapnia, although secondary changes such as pulmonary microatelectasis may contribute an element of *V*/*Q* mismatching. Assessing exercise hypoxaemia is difficult in these patients due to their generalized muscle weakness. However, sleep-related oxygen desaturation, particularly during REM sleep when the inspiratory system is most dependent on diaphragm function, is a common finding in patients with mild daytime hypoxaemia due to chest wall problems or neuromuscular diseases. Occasionally these changes are dramatic, but in boys with muscular dystrophy the presence of transient hypoxaemic episodes was no better guide to prognosis than was measurement of the vital capacity (Fig. 5). Arterial CO_2 tensions often lie in the high normal range, daytime hypercapnia only being seen in advanced disease.

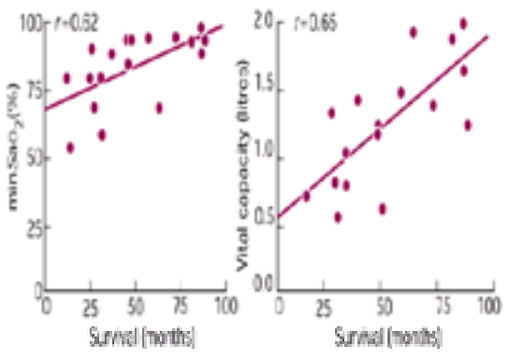

Fig. 5 Survival of boys with respiratory failure due to neuromuscular disease plotted against minimum arterial oxygen saturation recorded during sleep and vital capacity. (Based on data from Phillips *et al*.1999.)

Non-pulmonary disorders

Patients with stable congestive cardiac failure often show mild reductions in PaO_2 and a normal or low $PaCO_2$ due to premature airway closure secondary to pulmonary oedema. Some patients with severe liver cirrhosis develop the so-called hepatorenal syndrome, with otherwise unexplained hypoxaemia due to *V*/*Q* mismatching and true anatomical shunting through arteriovenous communications in the pulmonary circulation. Morbidly obese individuals can develop hypoxaemia and hypercapnia due to profound nocturnal hypoventilation and chemoreceptor resetting. Rather more common are the problems of patients with severe obstructive sleep apnoea who develop daytime hypoxaemia and hypercapnia secondary to recurrent nocturnal upper airway obstruction and oxygen desaturation. Careful review of these 'Pickwickian' patients often shows coexisting hypothyroidism or obstructive lung disease. This diagnosis should be suspected in any patient with COPD with significant respiratory failure and an FEV_1 greater than 1.5 litres. Correction of the sleep apnoea by nasal continuous positive airway pressure can produce significant improvement in daytime blood gases, but in the great majority of patients with obstructive sleep apnoea no significant abnormalities of waking gas exchange are seen.

Pulmonary vascular disease

This is an uncommon cause of hypoxaemia and at the time of diagnosis CO_2 retention is rare. Rather variable changes in D_LCO are reported, but as pulmonary hypertension becomes more advanced, exercise and resting hypoxaemia develops, a significant component being secondary to the reduced cardiac output and increase in mixed venous oxygen tension.

Assessment of chronic respiratory failure

This is relatively straightforward. The diagnosis of mild/moderate hypoxaemia rests on an awareness of the possibility rather than any specific clinical finding. When the arterial PO_2 is below 8.0 kPa, impairment of concentration and memory can be demonstrated, but these features are extremely non-specific. Although tempting to ascribe to hypoxaemia, the principal cause of breathlessness in these patients is usually the underlying disease. Reduction of peripheral chemoreceptor activity by supplementary oxygen can be beneficial, but this is usually secondary to a fall in minute ventilation rather that to any specific 'dyspnogenic' effect of hypoxia itself.

Hypercapnia is equally non-specific, with headache the most commonly attributed symptom. There are no good data to support this in compensated respiratory failure, although a generalized degree of vasodilation is seen in some patients with CO_2 retention, which may be accompanied by a large volume pulse and warm peripheral extremities.

On examination central cyanosis may be apparent as a bluish discoloration of the mucous membranes associated with an increase in the reduced circulating haemoglobin to approximately 5 g/dl. It is an unreliable clinical sign in some ethnic groups and in the presence of artificial illumination. An increased facial colouring due to secondary polycythaemia can occur, whilst the jugular venous pressure may be elevated and ankle swelling can develop in the face of worsening CO_2 retention.

The principal diagnostic steps are listed in Table 4. Measurement of arterial blood gases, preferably breathing air, is the most reliable way of diagnosing chronic respiratory failure. It is essential to know whether the patient was using oxygen when the sample was taken, and if so, at what inspired concentration. Patients with chronic airflow limitation treated with bronchodilators nebulized in oxygen may show unexpectedly high PaO_2 for some time after this treatment. Non-invasive measurement of arterial oxygen saturation using pulse oximetry can be used to screen individuals at risk of chronic respiratory failure and to monitor patients in hospital or overnight, but it is no substitute for assessing blood gas tensions to make the diagnosis correctly.

Treatment of chronic respiratory failure

Managing stable chronic respiratory failure involves several steps:

1. making a firm diagnosis;
2. correction of the underlying disorder;

3. increasing the inspired oxygen concentration; and
4. increasing alveolar ventilation.

Making a firm diagnosis

This is essential for rational management. It is important to remember that more than one process may contribute to the development of chronic respiratory failure; for example, poor left ventricular function due to cardiac disease and chronic obstructive pulmonary disease together. The relative importance of each factor should be determined.

Correction of the underlying disorder

In general, treatment of the primary pathology improves both *V*/*Q* relationships and hence oxygenation, and respiratory system mechanics, which increases ventilatory capacity and lowers the $PaCO_2$. In patients with COPD this usually involves administration of inhaled bronchodilators and corticosteroids (see Chapter 17.6), but marked improvement is the exception rather than the rule in patients where chronic respiratory failure has developed. Medical therapy tends to be ineffective by the time chronic respiratory failure has developed in interstitial lung disease and the neuromuscular disorders.

Specific pulmonary vasodilator treatment has been used to treat pulmonary hypertension, with most evidence of improvement seen after infusion of prostacyclin in primary pulmonary hypertension. Attempts to improve gas exchange in secondary pulmonary hypertension by the use of inhaled nitric oxide, a specific pulmonary arterial vasodilator, have been disappointing and in general, resting gas exchange has deteriorated after this treatment rather than improved.

Increasing the inspired oxygen concentration

Hypoxaemia secondary to *V*/*Q* mismatch or global hypoventilation is relatively easily corrected by supplementary oxygen. In chronic airflow limitation and especially COPD, where respiratory time constants for gas exchange are long, it may take 30 minutes before a new steady state is reached when breathing relatively low concentrations of oxygen. Monitoring of blood gases should be adjusted accordingly.

In the chronic stable state, treatment with oxygen is given to prevent or reverse the chronic consequences of hypoxaemia. The benefits of regular oxygen treatment on breathlessness are marginal and there are no data to suggest that the severity or subsequent progression of breathlessness is influenced by chronic oxygen treatment. Almost all data about oxygen therapy in chronic respiratory failure are based on observations in hypoxaemic COPD, treatment in other conditions being offered by analogy with this more common problem.

Two well-performed randomized clinical trials have shown that regular treatment of patients with COPD and stable hypoxaemia (PaO_2 less than 55 mmHg) prolongs life (Fig. 6(a)). These data suggest that patients using more oxygen (the 'continuous' limb of the Nocturnal Oxygen Therapy Trial Group) do better than either the United Kingdom Medical Research Council treatment group or the North American patients using oxygen only at night. A more recent Polish study found no benefit when patients with COPD with a PaO_2 of 7.3 to 8.8 kPa were treated with oxygen at home for 15 h/day (Fig. 6(b)), emphasizing that chronic oxygen therapy is only of value when the oxygen saturation falls below 90 per cent.

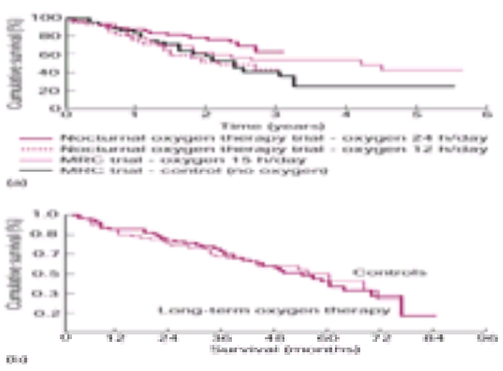

Fig. 6 The effect of regular domicilliary oxygen on survival in COPD. Panel (a) combines data from the MRC and NIH oxygen trial in the United States. Survival is greatest in those receiving oxygen for the whole 24-h day. Panel (b) is based on the study of Gorecka *et al*. for COPD patients with a PaO_2 between 7.3 and 8.5 kPa who were treated with oxygen or normal medical therapy. There was no survival benefit in the oxygen treated group, confirming the importance of the 7.3 kPa threshold in selecting patients for this therapy.

These studies showed that progression of secondary pulmonary hypertension can be halted by regular oxygen treatment and secondary polycythaemia can be corrected. However, secondary polycythaemia in COPD is influenced by the amount of carboxyhaemoglobin from cigarettes and patients who continue to smoke do not show a fall in red cell mass or packed cell volumes with oxygen treatment. Neuropsychological effects of chronic hypoxaemia have been described and may be improved by regular oxygen treatment, although the evidence for this is limited.

Giving oxygen during exercise increases performance and particularly endurance in patients with COPD who are relatively normoxaemic, as well as those with resting hypoxaemia. Again carbon monoxide from cigarette smoking reduces this response, and whether oxygen desaturation during exercise is necessary for the benefit to occur has not been conclusively established, although it is used as a reimbursement criterion for portable oxygen in North America.

Oxygen concentrators are the most cost-effective way of delivering oxygen for near continuous use. These devices have proved reliable and safe and use the ability of zeolite cells to separate nitrogen from room air and so generate an oxygen-enriched inspirate. Liquid oxygen has the advantage of allowing refilling of portable oxygen units relatively easily for use during exercise. Oxygen masks are the most accurate way of delivering oxygen and a range of inspired concentrations (24, 28, 35 per cent) are available. However, they are easily dislodged during sleep and plastic oxygen nasal prongs with a long extension pipe offer an easier system for use in the home. Occasional patients, especially those with severe interstitial lung disease, may have difficulties obtaining a PaO_2 greater than 8.0 kPa with these systems. Transtracheal oxygen delivery may have a role here, but early enthusiasm for this has been tempered by problems with cannula occlusion, infection, and bleeding. A variety of oxygen-conserving devices that deliver oxygen only during inspiration have been developed which increase the time between refills of portable oxygen equipment as well as having financial advantages in some health-care systems.

Improving alveolar ventilation

Mechanical

This is a valuable way of reducing arterial CO_2 in disorders like COPD and can increase arterial oxygen tension as well, especially in conditions such as neuromuscular disease where hypoventilation predominates. The use of tank respirators in neuromuscular weakness has now been superseded by the development of non-invasive nasal positive-pressure ventilation (**NIPPV**), which is normally only needed at night. This therapy is used increasingly in the management of acute on chronic respiratory failure in patients where the primary problem is ventilatory without coexisting pneumonia/acute lung injury. Its chronic use arose from the belief that respiratory muscle fatigue was an important cause of CO_2 retention in COPD and the empirical observation that gas exchange and survival were better in patients with kyphoscoliosis treated with night-time cuirass ventilation. Newer studies have shown that respiratory muscle function is well preserved in COPD when allowance in made for the muscle shortening secondary to pulmonary hyperinflation. Several trials of NIPPV in stable hypoxaemic but normacapnic COPD have reported relatively unimpressive results. By contrast, NIPPV mainly given at night improved blood gases in patients with hypercapnic COPD, as well as leading to benefits in health status. No good randomized trial of this therapy has yet been reported and the role of NIPPV in the treatment of hypercapnic COPD is still controversial.

By contrast, significant symptomatic and blood gas improvements have been demonstrated in patients with kyphoscoliosis, but again no randomized clinical trial data

are available. At present, it appears unlikely that these will be performed given the significant and sustained symptomatic benefits seen clinically. The only study to report prospective data on muscular dystrophy found no effect of regular NIPPV on survival in normocapnic patients, but use of this therapy as supportive treatment in the terminal phases of advanced muscular dystrophy appears to be associated with prolonged survival. It is always important in the face of progressive disease such as muscular dystrophy or motor neurone disease that the patient should be fully informed of the complications of NIPPV and the fact that it is unlikely to influence the underlying progression of the condition. Provided a good dialogue between patient, carer, and physician is established, then reasonable decisions about the use of this ethically difficult treatment are still possible.

Although volume-cycled ventilation was initially preferred, most patients are now managed with a bilevel pressure-cycled patient triggered device. This may have some advantages in obstructive lung disease where small amounts of PEEP can be added to reduce static PEEPi. Adequate peak inspiratory pressure generation, preferably in excess of 20 cmH_2O, is needed in both COPD and kyphoscoliosis, where total respiratory system compliance is reduced. Patient–mask interfaces remain a major problem, especially in patients with unusual craniofacial structure where getting a comfortable mask fit without excessive tightness can be difficult. Progress should be assessed by regular blood gas measurements, and overnight monitoring of oxygenation and CO_2 tensions is useful at the start of therapy. Patience and trained respiratory therapists are the best way of ensuring long-term compliance with treatment.

Specific pharmacological therapy

Whilst mechanical ventilatory support is effective it is also cumbersome, uncomfortable, and restricting, hence a simple drug treatment would be invaluable. Although medroxyprogestrone acetate has non-specific ventilatory stimulant effects and can produce small falls in CO_2 tension in patients with COPD, its oestrogen-like side-effects limit its use. Methylxanthines like theophylline have some chemoreceptor stimulant effects, but are mainly of use for their bronchodilator and anti-inflammatory properties. Almitrine bismethylate is an interesting specific peripheral chemoreceptor stimulant drug, which also modifies intrapulmonary *V*/*Q* matching and increases arterial oxygen while reducing CO_2 tensions in patients with resting hypoxaemia. These properties have led to its use in parts of Europe, but it is associated with the development of peripheral neuropathy and possibly increasing pulmonary artery pressure during exercise, which has limited its more widespread application.

Despite the attractions of a pharmacological approach, concerns over the precipitation of inspiratory muscle fatigue and the recognition that, in most diseases, the central drive to breath is already high, mean that treatment with respiratory stimulant therapy is likely to have only limited clinical application.

Further reading

Anonymous (1980). Continuous or nocturnal oxygen therapy in hypoxemic chronic obstructive lung disease: a clinical trial. Nocturnal Oxygen Therapy Trial Group. *Annals of Internal Medicine* **93**, 391–8.

Anonymous (1981). Long term domiciliary oxygen therapy in chronic hypoxic cor pulmonale complicating chronic bronchitis and emphysema. Report of the Medical Research Council Working Party. *Lancet* **i**, 681–6.

Calverley PM, Leggett RJ, Flenley DC (1981). Carbon monoxide and exercise tolerance in chronic bronchitis and emphysema. *British Medical Journal* **283**, 878–80.

Calverley PM *et al*. (1982). Cigarette smoking and secondary polycythemia in hypoxic cor pulmonale. *American Review of Respiratory Disease* **125**, 507–10.

Costello R *et al*. (1997). Reversible hypercapnia in chronic obstructive pulmonary disease: a distinct pattern of respiratory failure with a favorable prognosis. *American Journal of Medicine* **102**, 239–44.

Doherty MJ *et al*. (1997). Cryptogenic fibrosing alveolitis with preserved lung volumes. *Thorax* **52**, 998–1002.

Gorecka D *et al*. (1997). Effect of long-term oxygen therapy on survival in patients with chronic obstructive pulmonary disease with moderate hypoxaemia. *Thorax* **52**, 674–9.

Haluszka J *et al.* (1990). Intrinsic PEEP and arterial P_{CO_2} in stable patients with chronic obstructive pulmonary disease. *American Review of Respiratory Disease* **141**, 1194–7.

Meecham JD *et al*. (1995). Nasal pressure support ventilation plus oxygen compared with oxygen therapy alone in hypercapnic COPD. *American Journal of Respiratory and Critical Care Medicine* **152**, 538–44.

Phillips MF *et al*. (1999). Nocturnal oxygenation and prognosis in Duchenne muscular dystrophy. *American Journal of Respiratory and Critical Care Medicine* **160**, 198–202.

Raphael JC *et al*. (1994). Randomised trial of preventive nasal ventilation in Duchenne muscular dystrophy. French Multicentre Cooperative Group on Home Mechanical Ventilation Assistance in Duchenne de Boulogne Muscular Dystrophy. *Lancet* **343**, 1600–4.

17.8.1 Upper airways obstruction

J. R. Stradling

Definition

The trachea and carina are usually included in discussions of upper airways obstruction because many of the conditions that can completely block off the main airway can affect the trachea, presenting in a similar way to those affecting the larynx and pharynx. For convenience, the causes of upper airways obstruction are divided into acute (within minutes or hours) and non-acute, although there is not quite such a clear distinction in clinical practice. Many of the causes of upper airways obstruction (particularly infection) are more common in children, but this section deals with the problem mainly from an adults' physician's perspective.

At resting levels of minute ventilation, the main airway can be reduced to a diameter of 3 mm or so before respiratory distress and stridor occur. Little more narrowing is required to precipitate complete asphyxia. Hence, when upper airways obstruction is suspected, assessment of severity, diagnosis, and treatment must be regarded as a medical emergency. The causes are listed in Table 1 and discussed in more detail below.

Diagnosis

Diagnosis of upper airways obstruction requires a high degree of clinical suspicion. Not all that wheezes is asthma. If upper airways obstruction develops gradually, then it is most likely to be misdiagnosed as asthma or chronic airways obstruction, particularly if, for example, a carcinoma of the trachea coexists with chronic airways obstruction. This is not uncommon, since both are usually caused by smoking. Clues in the history will be a more rapid onset than might be expected for chronic airways obstruction and no previous history of a similar problem. The progression is usually relentless, without fluctuations, although a course of steroids prescribed for 'asthma' may produce temporary tumour shrinkage. At first, stridor or noisy breathing will only be heard on exercise, but it will gradually appear at lower and lower levels of activity. Sometimes the patient is well aware that the blockage is 'somewhere in the neck' and such a complaint should be taken seriously, as should associated haemoptysis. A non-productive cough is often present. A change in the voice in association with shortness of breath indicates the possibility of laryngeal obstruction. Upper airways obstruction is sometimes more symptomatic on lying down.

Examination

In pure upper airways obstruction, the noisy breathing will localize to the airway and tends to be monophonic and stridulous. Stridor may be absent if there is a long segment of obstruction. The only sound at the periphery on auscultation of the chest will be the transmitted noise of the stridor. However, as mentioned above, some patients will have coincidental lower airways obstruction, which should not discourage further investigation of a suggestive history. If upper airways obstruction is extrathoracic, stridor will tend to be worse on inspiration, and the converse may be true when the lesion is intrathoracic. The reasons for this are discussed below.

Tests of lung function

During a forced expiration from total lung volume down to residual volume there is a progressive fall in expiratory flow rate. This is largely due to the fact that the airways become narrower as the lungs become smaller, and this progressively restricts maximum flow rate regardless of the effort made. This can be displayed graphically as a plot of expiratory flow against the volume exhaled from total lung capacity down to residual volume, the so called 'flow–volume' plot or loop (Fig. 1(a)). This fall-off in maximal flow rates with falling lung volume is called 'volume dependence of flow'. However, if a fixed resistance is introduced (such as tracheal stenosis), then the maximal flow rate possible is independent of the lung volume. High flow rates, usually seen at larger lung volumes, cannot be generated and, instead of the normal triangular appearance of the flow–volume plot, it has a squared appearance. At lower lung volumes the normal intrinsic airways resistance may again exceed the abnormal upper airways resistance so that the flow–volume plot may once again follow the normal path (Fig. 1(a)).

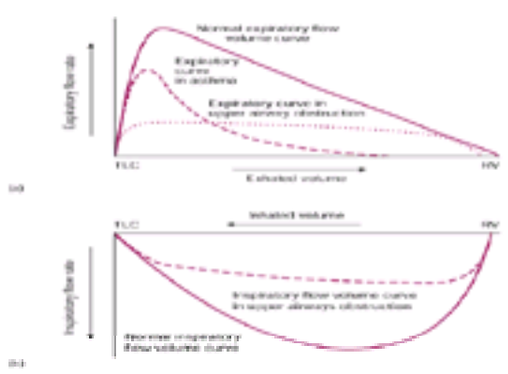

Fig. 1 (a) Expiratory flow–volume loop. The subject exhales with maximum effort from total lung capacity (TLC) until residual volume (RV) is reached. Normally, maximum flow (vertical axis) is reached early on and the flow falls almost linearly with lung volumes thereafter. In lower airways obstruction (e.g. asthma), all flows are reduced, but particularly at lower lung volumes. In upper airways obstruction, the maximum flow is clipped and roughly constant across most of the manoeuvre. (b) Inspiratory volume loop. The subject inhales maximally from residual volume (RV) up to total lung capacity (TLC). Normally, maximum flow is reached at about half-way when there is the best combination of airways size and muscle strength (see text). In upper airways obstruction, maximum flow is determined by the size of the remaining orifice and is roughly constant across the manoeuvre.

If the apparatus required to measure flow–volume plots or loops is not available, a peak expiratory flow (PEF) meter and spirometry plot may be useful. Because the fixed extra expiratory resistance clips the high flow rates predominantly, then the PEF rate will be reduced disproportionately to the forced expiratory volume in 1 s (FEV_1). This is because the FEV_1 is a measure over a longer time period, which includes lower flow rates because of the falling lung volume. This gives rise to a simple index of upper airways obstruction; FEV_1 (ml) divided by the PEF rate (l/min). Normally the value will be less than 10, but as the PEF rate is preferentially clipped (which does not happen when there is increased diffuse airways obstruction such as asthma) it may rise above 10 in upper airway obstruction. However, an index of less than 10 does not exclude upper airways obstruction because the lesion may not be rigid and, if it is intrathoracic, it may also narrow a little as lung volume falls. Another spirometric clue to upper airway obstruction is the shape of the FEV_1 curve. Normally this is curved because flow rate falls with time, but it becomes straighter if flow rate is fixed (Fig. 2(a)).

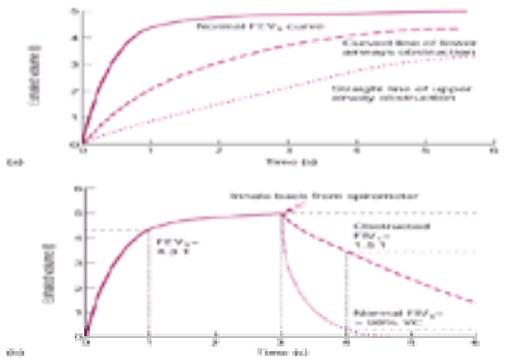

Fig. 2 (a) Curves of forced expiratory volume (FEV) against time. Normally exhalation is rapid, and more than 75 per cent of the final volume (vital capacity (VC)) is exhaled in 1 s (FEV_1). In lower airways obstruction, flows are slower and thus less air is exhaled in 1 s; the line is still curved because flows are falling. In upper airways obstruction, because flows are roughly constant at a low level set by the remaining orifice, the line is nearly straight and FEV_1 is also low. (b) Following a forced exhalation manoeuvre into a spirometer, a forced inhalation can be made. Normally inspiration is fast and the forced inspiratory volume in 1 s (FIV_1) is almost the vital capacity (VC). If there is upper airways obstruction, particularly extrathoracic such as at the vocal cords, then inspiration will be very limited and FIV_1 will be small.

If flow–volume plotting apparatus is available, inspiratory patterns can also be examined. The normal inspiratory limb of the flow–volume loop is almost semicircular (Fig. 1(b)). This is because at residual volume the airways are small and limit flow; towards total lung capacity the inspiratory muscles are reaching their full contraction and power is falling off. Hence maximum flows are achieved in the midrange of lung volume (Fig. 1(b)). Again, if upper airways obstruction is present, this pattern may be replaced by a squarer shape owing to the imposition of a lower maximum flow rate by the fixed resistance (Fig. 1(b)).

Comparison of the inspiratory and expiratory limbs of the flow–volume loop may give a clue as to the location of an upper airways obstruction. If the lesion is extrathoracic and has any variability to its lumen, it will tend to be narrowest during inspiration (walls sucked together) and widest during expiration (blown apart). Conversely, an intrathoracic lesion will tend to be squashed on expiration by the raised intrathoracic pressures, thus presenting a higher resistance than during inspiration when it will tend to be pulled open. Although in theory these statements are correct, in practice flow–volume loops are not always sufficiently characteristic to allow a confident diagnosis about the exact site and presence of an upper airways obstruction. They may be more useful as a tool to follow changes, such as in response to treatment.

Vocal cord paresis due to bilateral recurrent laryngeal nerve damage is often very much worse on inspiration. Simple spirometry can be diagnostic here. The expiratory tracing will be normal as the cords are blown apart. If the patient immediately inhales back from the spirometer (make sure that a new in-line filter is present first) the inspiratory rate will be tortuously slow (Fig. 2(b)). The forced inspiratory volume in 1 s (FIV_1) will often be much smaller than the FEV_1, whereas normally the reverse is true (Fig. 2(b)).

The effect of breathing a low-density gas mixture, such as 21 per cent oxygen in helium (Heliox), on airways resistance has also been used to try to differentiate lower from upper airways obstruction. Flow in small airways is largely laminar and not affected by the density of the intraluminal gas. However, flow at a tight stenosis will be turbulent, and then it becomes dependent on the gas density. Thus airflow resistance will fall on breathing Heliox if the obstruction is in the upper airways, but will remain unchanged when the increased resistance is due to peripheral airways narrowing. This test may be particularly useful when there is evidence of both lower and upper airways obstruction, but proof of a significant contribution from an upper airways lesion is required before considering any further intervention.

For simple monitoring of progress of a patient with known upper airways obstruction on the ward, for example during treatment, sophisticated tests are not required, and measurement of the PEF rate is probably adequate for most purposes.

Specific causes

Acute causes of upper airway obstruction

Aspiration

Upper airways obstruction due to aspiration is usually due to the object lodging in the larynx, since this is the narrowest portion of the airway until two or three divisions down the bronchial tree. The usual culprit is a piece of food, and thus this condition has been colourfully called the 'café coronary'. The patient suddenly becomes distressed, is unable to talk, and apparently unable to breathe. He may point to his throat trying to indicate the problem. Inspiration to provide the air necessary for a good expulsive cough may not be possible. Indeed, lung volume may 'ratchet' down to residual volume.

The Heimlich manoeuvre was invented for this circumstance and its principles should be taught to first-aid workers. If the patient is still upright, then the helper stands behind with his arms clasped around the upper abdomen. A very forceful pull, backwards and upwards, will drive the diaphragm upwards and should provide enough expired air to shift the aspirated food from the cords (Fig. 3). The manoeuvre can be repeated, of course, but a forceful first try is likely to be the most successful. If the Heimlich manoeuvre fails, then it may be possible to dislodge the lump of food with a finger once the patient has become unconscious. The only alternative is an emergency cricothyrotomy which requires a hole to be made in the cricothyroid membrane just below the Adam's apple of the thyroid cartilage and above the cricoid cartilage. Even a small hole (2 mm or so) will allow sufficient ventilation to keep the patient alive. Emergency cricothyroidotomies have been attempted with everything from penknives to ball-point pens: special large-bore curved-needle kits are available for the purpose and are safer than an unskilled attempt at a tracheostomy.

Fig. 3 The Heimlich manoeuvre for the emergency treatment of acute pharyngeal or laryngeal obstruction due to a bolus of food. Two or three sharp thrusts in the direction of the arrow may cause the food to be ejected. (Reproduced with permission from Flenley, 1990.)

Oedema

Acute oedema of the larynx or pharynx is usually either due to allergy (atopic or non-atopic), a hereditary abnormality in the complement pathway, or inhalation of noxious gases.

Allergic oedema

Episodes of upper airways and facial oedema sometimes have no known cause and appear without warning. Often there will be an atopic history with a specific allergy. Some allergic reactions are not based on atopy and IgE, but may occur through IgG or direct activation of other inflammatory pathways. Allergies to nuts, strawberries, etc. may involve this latter mechanism rather than IgE. Insect stings usually produce pharyngeal and glottic oedema via IgE mechanisms.

Treatment of these allergic causes of upper airways obstruction consists of subcutaneous (or intramuscular) adrenaline (1 ml of 1:1000) with antihistamines and steroids (see Section xxx on acute anaphylaxis). Aerosolized adrenaline may also be useful, using 10 ml of 1:10 000 in an ordinary nebulizer.

Hereditary and acquired angioedema

Hereditary and acquired angioedema are due to deficiency of plasma C1 esterase inhibitor. This defect may be caused by impaired synthesis, due to a genetic defect causing failure of protein production, or production of inactive protein (hereditary angioedema), or by increased catabolism (acquired angioedema). The absence of C1 esterase inhibitor means that the enzyme activating the first component of the complement pathway is unchecked, allowing abnormal increase in activity of the whole pathway and production of vasoactive products.

Hereditary angioedema typically presents in infancy. An episode of oral and facial swelling is often precipitated by local trauma, such as a tooth extraction or a blow to the face, and lasts about 48 to 72 h. The skin manifestations do not itch in the way that allergic oedema does. Colicky abdominal pain due to intestinal oedema is an alternative presentation, when pooling of fluid in oedematous bowel can be sufficient to cause hypotension and shock. Diagnosis hinges on the clinical presentation, a positive family history (autosomal dominant), and low levels of C1 esterase inhibitor. In the form where normal amounts of inactive C1 esterase inhibitor are present, it is necessary to demonstrate low C4 levels during an attack.

Acquired angioedema presents in adult life. It is not familial. It may be associated with lymphoproliferative or other malignant diseases (type I) or more commonly with the presence of autoantibodies to C1 inhibitor (type II).

The disease is serious: 25 to 30 per cent of sufferers of hereditary angioedema die from asphyxia. As for cases of allergic upper airway oedema, treatment of the acute attack consists of adrenaline and steroids, but the response is much less satisfactory. Emergency tracheostomy or cricothyrotomy may be necessary. C1 esterase inhibitor plasma concentrate can be used for severe attacks, but very large doses are frequently needed for those with an acquired defect. Fresh-frozen plasma has been reported to work sometimes, but it contains larger amounts of the substrates C4 and C2 which might theoretically provoke worsening oedema.

Attenuated androgens, for example danazol, raise C1 esterase inhibitor levels within a few weeks, probably by increasing hepatic synthesis, and can usually prevent attacks in hereditary angioedema. Acquired angioedema generally fails to respond to this treatment, but can be treated prophylactically with antifibrinolytic agents. If episodes such as tooth extractions are triggers, then C1 esterase inhibitor can be given beforehand.

Smoke inhalation

Inhalation of hot smoke can burn the upper airways and contributes significantly to deaths due to fires. Upper airways obstruction due to heat injury and mucosal swelling usually develops within 24 h of exposure, but stenosis due to scarring can develop later. A hoarse voice, stridor, severe conjunctivitis, burnt nasal hairs, and falling peak flow all suggest significant upper airways damage. Bronchoscopy is then the best tool to establish whether there is significant oedema or mucosal ulceration obstructing the airways.

Management usually consists of simple measures such as elevating the head of the bed and inhaling cool moist air with added oxygen. If peak flow continues to fall, then transfer to an intensive care unit and bronchoscopy with the capability to perform an intubation, guided by direct vision, is the correct approach.

Infections

Upper airway infections rarely cause obstruction in adults, but can do so in infants and young children. Whilst they can present dramatically with upper airways obstruction, prodromal symptoms usually occur. Streptococcal pharyngitis, tonsillitis, and retropharyngeal abscesses are amongst the most important. Croup (due to respiratory syncytial, parainfluenza, and other viruses) is very common, with narrowing of the subglottic trachea, sometimes with a thick purulent coating over the larynx and trachea. Treatment consists of cool mist and supplemental oxygen, with careful monitoring of upper airways function.

Although again more common in children, acute epiglottitis, usually due to *Haemophilus influenzae*, can affect adults. Pyrexia, drooling, hoarse voice, difficulty in breathing, intense sore throat, and stridor are the usual presenting symptoms. Compared with croup, there is usually a faster onset and course. The diagnosis may be missed initially but lateral neck radiographs show swelling of the epiglottis (Fig. 4). Attempts to examine the back of the throat may precipitate further obstruction, particularly in children. Even tipping the head back for a lateral neck radiograph can provoke complete obstruction and be disastrous. Thus, if there is evidence of breathing difficulty with stridor and the clinical diagnosis is epiglottitis, then the correct management for children is immediate transfer to intensive care and intubation for 48 to 72 h whilst the infection is controlled by ampicillin or chloramphenicol. In adults, close monitoring in intensive care is probably adequate and prophylactic intubation is not routinely practised. General use of *H. influenzae* vaccines should make this problem increasingly rare.

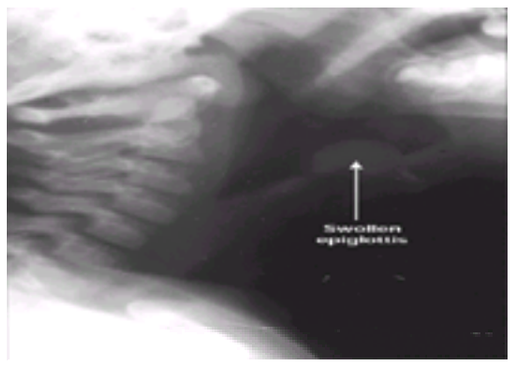

Fig. 4 Lateral neck radiograph of a child with epiglottitis. Note the swollen epiglottis overlying the glottis.

Non-acute causes of upper airway obstruction

Tumours

Laryngeal, and less commonly tracheal, tumours are usually seen in smokers. The dominant cell type is squamous. Spread of a primary bronchial carcinoma into the base of the trachea is probably the commonest cause of upper airways obstruction in pulmonology practice.

Laryngeal tumours nearly always present with hoarseness, or voice change, and cough. Large airways tumours are commonly not diagnosed until far advanced. This is because they mimic lower airways obstruction, as mentioned earlier, and chest radiography is often normal. Tumours may also respond to asthma therapy, showing temporary shrinkage with steroids, which may further mask the real diagnosis.

If history, examination, and lung function tests suggest an upper airways obstruction, then some form of imaging is required. CT is the least invasive approach and therefore least likely to disturb the airway and make matters worse, but will provide no histology. Plain films (posteroanterior and lateral) may show tracheal narrowing

but can be very deceptive. Direct visualization is usually necessary for diagnosis, biopsy, and to aid future therapy.

There is some disagreement as to whether fibreoptic or rigid bronchoscopy should be the investigation of first choice. Rigid bronchoscopy requires anaesthesia and sometimes this precipitates acute obstruction, when the bronchoscope then has to be passed quickly and forced through the obstructing tumour. This 'core-out' can reduce tumour bulk, with control of haemorrhage possible under direct vision, and improvement in the airway may buy time while other treatments such as radiotherapy are employed. Flexible fibreoptic bronchoscopy may be possible without disturbing the tumour, although coughing and increased secretions can precipitate complete occlusion. Direct application of adrenaline may help as an initial emergency treatment, and in theory cocaine (a vasoconstrictor) would be preferable to lignocaine as a local anaesthetic. If stenosis reduces tracheal diameter to less than 4 mm or so, it is best left alone during flexible bronchoscopy and should certainly not be biopsied (Fig. 5). In cases that are likely to be difficult, it can be helpful to pass an endotracheal tube over the flexible bronchoscope before it is introduced. This allows a guided intubation in an emergency, using the bronchoscope as a guidewire.

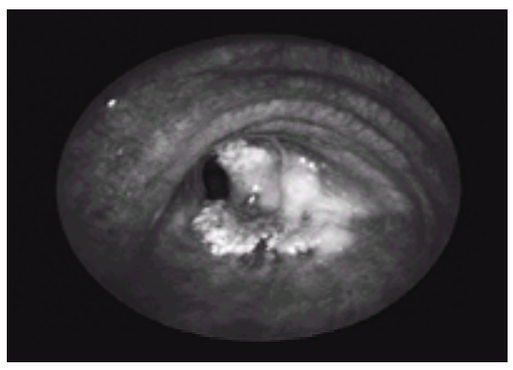

Fig. 5 Bronchoscopic view of a tracheal carcinoma blocking most of the lumen. (By courtesy of Dr P. Stradling.)

Aside from intubation or tracheostomy (when appropriate), emergency treatment of tumours compromising the upper airway consists of dexamethasone (12 mg daily), nebulized adrenaline (10 ml of 1:10 000 up to six times daily), humidification of inspired air, and breathing Heliox (21 per cent oxygen in helium). Improvement in the airway may then be achieved by treatment of the tumour with chemotherapy or radiotherapy. Sometimes, however, these can initially provoke swelling of the tumour, so that steroids are usually prescribed first, with emergency treatments kept close to the hand (Heliox, adrenaline). If these therapies do not help, then palliation can be achieved with the use of bronchoscopically guided laser therapy or cryotherapy, which literally either burn or freeze away tumour tissue with a low incidence of serious haemorrhage. Laser therapy is a laborious procedure, currently only available in a few specialist centres. Cryotherapy is quicker and safer, and can be performed down a flexible bronchoscope. These techniques are only of use with intraluminal tumours and cannot be applied when the narrowing is due to external compression. Another approach is the use of silicon or metal endobronchial stents. Some of these can be inserted via a flexible bronchoscope; others require surgery. They are particularly useful when external compression is present, and can produce dramatic resolution of symptoms. It is rarely appropriate to 'debulk' a malignant tumour at thoracotomy in an attempt to improve large airway patency.

Unfortunately, upper airways obstruction from a tumour becomes a terminal event in many cases. Powerful sedation is indicated to make the patient unaware that he or she is asphyxiating and choking to death.

Some rare, non-malignant tumours can obstruct the trachea, and rarely granulomatous conditions such as sarcoid and Wegener's granulomatosis may mimic tumour.

Tracheal stenosis

Tracheal stenosis usually develops following prolonged intubation or after a tracheostomy has been allowed to close following tube removal (Fig. 6). This scarring may appear some time after the initiating event. Again, radiology or bronchoscopy will usually confirm the diagnosis, already strongly suspected from the history. Temporary relief may be obtained by dilating the stricture at rigid bronchoscopy. Definitive treatment involves resection of the stenosed portion and reanastomosis.

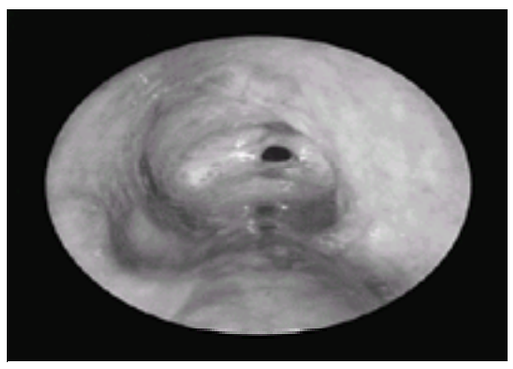

Fig. 6 Bronchoscopic view of a post-tracheostomy tracheal stricture. The remaining hole is about 2 to 3 mm in diameter. (By courtesy of Dr P. Stradling.)

Tracheal compression

Tracheal compression (Fig. 7) may be due to malignant or non-malignant conditions. External compression by malignant tumour (primary or secondary) has been covered in the previous section. Non-malignant causes include thyroid enlargement, aortic aneurysm, sclerosing mediastinitis, mediastinal neurofibroma, and Castleman's disease. If definitive treatment is not possible, then stenting the airway is the only option available. When thyroid enlargement leads to tracheal obstruction, surgical removal may not solve the problem completely. Prolonged pressure on the trachea can lead to tracheomalacia, so that the tracheal wall collapses when unsupported by the thyroid. Temporary use of an endotracheal stent is then appropriate.

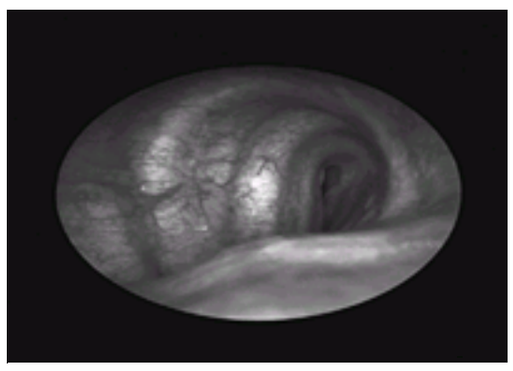

Fig. 7 Bronchoscopic view of external tracheal compression by right-sided paratracheal malignant nodes. (By courtesy of Dr P. Stradling.)

Tracheal abnormalities

Tracheomalacia may be secondary to prolonged external compression (see above) or a primary abnormality that presents in childhood. It is essentially a weakness or deficiency of the supporting cartilages. It is sometimes seen secondary to a long history of chronic airways obstruction. Normally the anteroposterior diameter of the trachea decreases by up to about 10 per cent during a cough. In tracheomalacia, collapse during coughing is over 50 per cent and sometimes is complete. The symptoms of this are usually stridor, shortness of breath, and paroxysms of coughing. In addition, inefficient coughing can lead to recurrent pneumonia and bronchiectasis.

A 'scabbard trachea' is said to be present when the lateral dimensions of the trachea are significantly narrower than the anteroposterior dimensions. This deformity, usually present along the whole intrathoracic trachea, is normally associated with chronic airways obstruction. It rarely causes severe upper airways obstruction and on a plain chest radiograph there is obvious tracheal ring calcification.

Tracheobronchiomegaly (or Mounier–Kuhn disease) is probably an inherited structural abnormality of the trachea presenting in adult life as apparent chronic airways obstruction. The trachea is dilated from the larynx to the second- or third-generation airways. There is atrophy of both cartilage and muscle. It is usually misdiagnosed as chronic airways obstruction, but the presence of prolonged but ineffectual coughing and harsh upper airway sounds should lead to lung function tests which then show evidence of an expiratory (intrathoracic) upper airway resistance. Radiological examination will show the dilated airways.

Tracheobronchopathia osteochondroplastica is a very rare condition characterized by cartilaginous and bony excrescences growing into the large airway lumina. This can lead to significant upper airways obstruction, but is more often a post mortem finding which is unsuspected in life.

Relapsing polychondritis is an 'autoimmune' systemic disorder affecting cartilage all over the body (ribs, trachea, ear lobes, nose, joints) and may be associated with systemic lupus erythematosus, Wegener's granulomatosis, and cryptogenic liver cirrhosis. Large airways involvement is a frequent cause of death in this condition. There is irregular narrowing of the trachea and main airways, with flaccidity of the tissues sometimes allowing marked collapse on expiration. The diagnosis can be hard to make: involvement of other cartilaginous sites is the critical feature to look for. The condition is often difficult to treat. Aside from local surgical or stenting procedures, steroids and other immunosuppressive therapies are usually given.

Laryngeal dysfunction

Damage to one recurrent laryngeal nerve usually causes a weak voice that improves a little with time as the opposite cord 'learns' to compensate by moving slightly across the midline to improve apposition. As the recurrent laryngeal nerve is invaded or compressed (usually by tumour at the left hilum), differential effects on abductors and adductors may be seen. For example unopposed adduction may occur prior to complete paralysis.

Bilateral recurrent laryngeal nerve paralysis produces flaccid cords that lie passively midway between full abduction and adduction. Abduction is very poor, such that rapid inspiration will draw the cords together and produce stridor, thus limiting exercise tolerance. Inspiratory stridor may initially be present only during sleep, when general decrease in muscle tone reduces any residual laryngeal abductor activity. Although this may be labelled as snoring, careful questioning of a witness will identify whether snoring or the machinery-like screech of inspiratory stridor is present (particularly if the physician can imitate the two noises!).

The usual clinical history is of voice change following thyroidectomy some years before. This may have been quite subtle, such as difficulty in singing, but with speech relatively unaffected. Nocturnal stridor and reduction in exercise tolerance then develop over a period of years. Eventually the obstruction at night can be sufficient to produce obstructive apnoea and respiratory failure. These events may be due to involvement of the previously damaged recurrent laryngeal nerves in scarring at the thyroidectomy site. Sometimes bilateral paralysis can occur for no apparent reason, and it is assumed that the aetiology is similar to Bell's palsy or the diaphragmatic palsy of neuralgic amyotrophy. A very rare differential diagnosis is an Arnold–Chiari malformation causing brain stem compression and presenting with sleep-related stridor in association with ventilatory failure.

Laryngeal surgery could prevent inspiratory cord closure, but would do so at the expense of the voice. Thus a tracheostomy with a speaking tube is the usual approach. However, if the night-time obstruction is the main problem (with sleep disruption and daytime sleepiness), nasal continuous positive airway pressure therapy will usually keep the cords apart during sleep.

Damage to the superior laryngeal nerves supplying the cricothyroid (a vocal cord tensor) causes only a weak voice. Speech is still possible because the main adductors still function

Apart from laryngeal paralysis, laryngeal destructive conditions such as rheumatoid arthritis can lead to poor abduction with inspiratory stridor, particularly at night. In Parkinson's disease with autonomic involvement (Shy–Drager syndrome) or more generalized brain atrophy (multisystem atrophy) there can be a fairly specific wasting of the laryngeal abductors. This also presents with inspiratory stridor (or apnoea) at night and can progress to respiratory failure.

Functional laryngeal abnormalities, with narrowing during inspiration and/or expiration, can occur. These may be due to psychological problems, but the syndrome blends with reflex laryngeal dysfunction in patients with asthma. Expiratory laryngeal wheezing can occur in response to emotional pressure even in well-controlled asthma. In this situation, the laryngeal component of the increased airways resistance can be considerable. Inhalations of histamine can sometimes mimic this, which might therefore be due to a reflex originating from afferent receptors. Why this should happen is not clear, but it may be activation of the laryngeal braking mechanism to help raise functional residual capacity.

Functional inspiratory stridor is not particularly related to asthma, but may follow a respiratory tract infection. There is some evidence that techniques used by speech therapists can help with this problem.

Further reading

Cicardi M, Bergamaschini L, Cugno *et al.* (1998). Pathogenetic and clinical aspects of C1 inhibitor deficiency. *Immunobiology* **199**, 366–76.

Empey DW (1972). Assessment of upper airways obstruction. *British Medical Journal* **3**, 503–5.

Flenley DC (1990). *Respiratory medicine*. Baillière Tindall, London.

Fraser RG, Paré JAP, Paré PD, Fraser FS, Genereux GP (1990). *Diagnosis of diseases of the chest*, Vol. 3. W.B. Saunders, Philadelphia, PA.

Goldman J and Muers M (1991). Vocal cord dysfunction and wheezing. *Thorax* **46**, 401–4.

Valsecchi R, Reseghetti A, Pansera B, Di Landro A (1997). Autoimmune C1 inhibitor deficiency and angioedema. *Dermatology* **195**, 169–72.

17.8.2 Sleep-related disorders of breathing

J. R. Stradling

Introduction

This chapter discusses the disorders of breathing that appear only during sleep. This rapidly expanding subspecialty within respiratory medicine now provides about 10 to 15 per cent of the speciality's referrals. Following its first proper description in 1967 as a medical curiosity, obstructive sleep apnoea and its variants are now thought to significantly impair the functioning of about 0.5 to 1 per cent of the population. Most general hospitals will have some form of monitoring system for the diagnosis of sleep apnoea syndromes, although tertiary centres tend to provide most of the treatment. The diversity of symptoms produced by these disorders means that all physicians need to have an understanding of them and are likely to come across many cases during their professional life.

Normal physiology of breathing during sleep Table 1)

Sleep can be divided into two very different states. The dominant sleep stage is non-rapid eye movement (NREM) sleep (Fig. 1 and Fig. 2). This phase of sleep, which is preferentially reclaimed following sleep deprivation, appears to be when the brain shuts down and is necessary for maximum daytime alertness and continuing cognitive function. NREM sleep shows a continuum from drowsy down to very deep sleep, arbitrarily subdivided into stages 1, 2, 3, and 4. The awake electroencephalogram (EEG) is characterized by low voltage, high frequency activity, with the only dominant frequency being the so-called alpha activity (approximately 10 Hz), present when the eyes are closed. As sleep supervenes, the alpha activity disappears, overall EEG frequency falls, muscle tone (usually measured from a chin electromyogram (EMG)) falls, and the eyes begin to roll from side to side. This transition phase is called stage 1. Stage 2 is defined by the appearance of K complexes (isolated large waves) and sleep spindles (bursts of about 13 Hz activity). As sleep deepens further, increasing amounts of large, slow waves (approximately 1 Hz) appear. These stages are called 3 and 4, or slow-wave sleep.

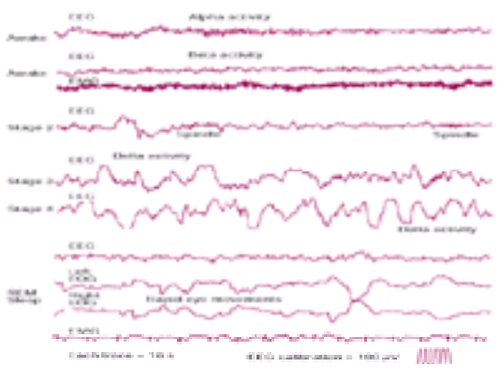

Fig. 1 Examples of electrical brain activity (EEG), eye movements (EOG), and chin muscle tone (EMG) during wakefulness and the different sleep stages.

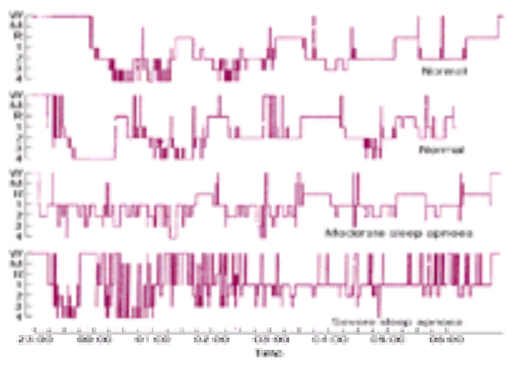

Fig. 2 Examples of all-night hypnograms (based on 20-s epochs) in two normal subjects and two patients with sleep apnoea. Note the reduced deep sleep (stages 3 and 4) in the patients, but no indication that they are waking up hundreds of times a night. W, awake; M, movement (awake); R, REM sleep; 1 to 4, stages 1, 2, 3, 4 of non-REM sleep.

The other main phase of sleep is rapid eye movement (REM) sleep or dreaming sleep. This stage is characterized by a return of the EEG to a pattern resembling wakefulness. The EMG tone falls to very low levels and there are bursts of rapid eye movements, mainly from side to side, under closed eyelids. Effectively the cortex is 'awake' again, processing images and able to integrate outside noises or other stimuli into complex dreams. The fall in EMG tone is because the rest of the body's muscles have been 'cut off' from the brain and paralysed, hence the fall in EMG tone. This paralysis (or atonia) is under active control from a centre in the pons that hyperpolarises the lower motor neurones via inhibitory reticulospinal pathways. Cats in whom this centre has been destroyed no longer show atonia during REM sleep, and as a consequence they may get up and walk around or appear to chase phantom birds, presumably reflecting their dream content. The function of this atonia centre may therefore be to prevent the dreaming brain from influencing the rest of the body. Paralysis during REM sleep occurs dominantly in muscles that normally have a tonic postural activity; thus the diaphragm is spared although pharyngeal, intercostal, and accessory muscles are all affected to differing extents.

The normal pattern of the oscillation between NREM and REM sleep is shown in Fig. 2. This 'hypnogram', as it is called, is constructed by classifying successive 30-s epochs from tracings of EEG, EMG, and eye movement data into either awake, movement, REM sleep, or stages 1–4; thus 960 epochs are obtained in an 8-h night.

During wakefulness breathing is influenced by a variety of pathways, some conscious and voluntary, others entirely automatic and involuntary. Classic responses to hypoxia, hypercapnia, and vagal afferents (integrated in the brainstem) can be overruled by cortical signals to subserve functions such as talking. These two types of control are separate and can be damaged separately by disease processes. The presence of wakefulness itself provides an input to the respiratory centre, almost equivalent to the amount of ventilation seen at rest. Thus, following a period of hyperventilation, a normal subject will go on breathing at just below normal levels, despite hypocapnia and hyperoxia, until the carbon dioxide rises, when normal ventilation is re-established. This is not true during NREM sleep, when hypocapnia will produce apnoea until the $PaCO_2$ rises back to a critical threshold level.

Another component of wakefulness is the high muscle tone that holds the body in the required posture. This 'awake' input into the anterior horn cells means that other inputs, such as those from the respiratory centre, can further activate muscles, including the intercostals and pharyngeal. The withdrawal of this 'awake' tone with the onset of sleep means that a certain respiratory centre output to the relevant anterior horn cells is less able to raise membrane potentials to firing threshold, such that respiratory muscle activity falls with sleep onset, minute volume typically reduces by about 10 to 15 per cent, and $PaCO_2$ rises by 3 to 8 mmHg. Reduction of pharyngeal muscle tone narrows the lumen, and thus there is normally a rise in upper airway resistance. This reduction in ventilation has trivial effect on the arterial oxygen saturation (SaO_2) in the normal circumstance when SaO_2 is on the flat part of the haemoglobin dissociation curve, but dramatic falls in saturation will be apparent when SaO_2 starts below 92 per cent, the steep part of the curve. If ventilatory responses to carbon dioxide or hypoxia are measured during NREM sleep, the slopes are flatter and right shifted, indicating a reduced overall sensitivity. Exactly why this occurs is not known, but reduced tone of the respiratory muscles, the withdrawal of awake drive, increased upper airway resistance, and (probably) true reduction in central sensitivity to carbon dioxide or $[H^+]$ could all contribute. If, as a consequence of respiratory disease, compensatory mechanisms are already employed to cope with the extra work of breathing, then these seem to be particularly reduced during sleep as well.

During REM sleep, overall ventilation stays much the same as in NREM sleep, but the breath-to-breath variability increases considerably, sometimes with apnoeas during the actual periods of eye movements, and compensatory increases in between. Sensitivities to carbon dioxide and hypoxia were originally thought to be further reduced, but they are hard to measure in the presence of spontaneously variable breathing and more recent evidence suggests they may not change much at all. What is far more important is the atonia of postural muscles. The hyperpolarization of the anterior horn cells greatly reduces the efficacy of respiratory signals to the intercostal, accessory, and pharyngeal muscles. This will not matter in a normal subject with an efficient diaphragm and a non-compromised pharynx. However, if the subject is dependent on muscles other than the diaphragm for breathing, or has a narrow compromised pharynx, then REM sleep may powerfully interfere with ventilation with consequent hypoxaemia and hypercapnia.

Also of relevance to breathing during REM sleep are the reduced arousal responses to respiratory stimuli compared with non-REM sleep. The arousal responses to some ventilatory stimuli (hypoxia, hypercapnia, extra resistive load) are believed to be mediated mainly by the perception of the ventilatory effort made in response, rather than the specific ventilatory stimulus itself. If a ventilatory response to hypoxia is measured during REM sleep, then the subject will usually tolerate a much lower SaO_2 before arousing compared to NREM sleep. Furthermore, if the drive to sleep is high, such as after sleep deprivation, arousal will be delayed still further.

It can be seen from the above that, although sleep is not a problem for those with normal respiratory systems, once abnormalities are present there is potential for a damaging interaction between sleep and breathing, particularly during REM sleep.

Obstructive sleep apnoea

Definition

Sleep apnoea was first properly documented in neurophysiological sleep laboratories using techniques that had been developed for the investigation of conditions such as insomnia, narcolepsy, and depression. It was realized that hundreds of episodes of breath cessation, or apnoea, usually due to upper airway obstruction with associated snoring, were related to marked sleep disturbance. Because simple oronasal flow detectors were used, the critical event was defined as an episode of apnoea. Because it was easy to measure, an arbitrary definition was made, and breath cessation for longer than 10 s became an official apnoea. Early work suggested that normal, young people rarely had more than about 30 apnoeas per night, so that the standard definition of 'sleep apnoea syndrome' became more than 35 apnoeas per night, or more than five per hour of sleep, each lasting for 10 s or longer. This definition has existed long beyond its clinical usefulness: it is quite clear that recurrent partial obstruction to the upper airway can fragment sleep just as severely with no actual apnoeas or hypopnoeas developing at all. A more pragmatic and clinically useful definition might now be 'a sleep disruption syndrome that is due to a respiratory problem engendered by sleep itself, sufficient to cause symptoms when awake. Usually this is upper airway incompetence during sleep, but may also be due to problems of respiratory drive'. As the pathogenesis of sleep apnoea is explained, this shift in emphasis, with the inclusion of symptoms, will become clear.

Aetiology (Table 2)

The upper pharyngeal airway has to serve two functions, swallowing and breathing, which require different design features. When used for swallowing the pharynx has to behave like the oesophagus, and when used for breathing it has to remain an open tube like the trachea. These dual functions are achieved by having a floppy and collapsible muscular tube that is also capable of being held open rigidly by dilator muscles. The muscles responsible for this dilator function are discussed in the section on the structure and function of the upper respiratory tract (see Chapter 17.1.1). All these muscles have reduced activation during sleep, so that some pharyngeal narrowing occurs normally. There are, then, additional factors that determine whether this reduction leads to significant upper airflow obstruction in a particular individual. There are various theories as to these additional factors, but essentially they divide into two groups. Firstly, there may be abnormalities of the activation of the pharyngeal dilator muscles, perhaps due to defective or unstable central control. Secondly, there may be anatomical abnormalities that allow significant obstruction to occur even with the normal sleep-related reduction in muscle tone.

Neuromuscular function

Early investigations of EMG activity in pharyngeal muscles found reductions in tone with sleep during obstructive apnoeas. However, it was very difficult to show that these reductions were truly abnormal. Recent evidence shows that there is in fact an increase in activity of these muscles, both awake and asleep, in response to factors provoking pharyngeal collapse. In some patients with primary neuromuscular problems (from brainstem lesions to myopathies) there can be associated obstructive sleep apnoea, and pharyngeal muscle involvement seems a probable explanation. However, the majority of patients with obstructive sleep apnoea do not show evidence of any other neuromuscular problems.

During inspiration, pharyngeal dilator activity has to be synchronized with diaphragmatic activity and be adequate to overcome negative intrapharyngeal pressure. It has been suggested that a lack of co-ordination between diaphragmatic and pharyngeal activation may allow the pharynx to collapse. For example normal subjects breathing against an inspiratory resistance can be made to have a few obstructive apnoeas by artificially inducing periodic breathing during sleep. The gradual return of respiratory drive, following the nadir of ventilation, seems to activate the diaphragm first, leaving the pharynx unbraced. The presence of an inspiratory resistance then 'challenges' the pharynx and allows collapse for a few breaths before pharyngeal tone returns and restores patency.

Although instability of respiratory control during sleep has been postulated as a cause of obstructive sleep apnoea, following treatment with nasal continuous positive airway pressure therapy (see later) there is no evidence of a premorbid underlying respiratory instability, nor does altering respiratory drive have a useful effect. More convincing is the suggestion that there may be failure of normal reflex protective mechanisms in the pharynx, whereby receptors in the pharynx detect falls in pressure that distort the airway and provoke protective increases in pharyngeal dilator tone (see Section 17.1.1). Snoring itself may also be one of the stimuli that activate this dilator reflex, and it is conceivable that interruption of this reflex arc can occur, perhaps through years of pharyngeal trauma from snoring, mucosal oedema, or toxic agents such as cigarette smoke and alcohol.

Anatomical causes

Anatomical abnormalities influence pharyngeal function in a variety of ways. Simple encroachment of the pharyngeal lumen, for example with tonsillar hypertrophy, means that the normal fall in pharyngeal dilator tone with sleep can lead to critical narrowing and obstruction. Alternatively, there are abnormalities which 'load' the

upper airway, requiring increased dilator muscle action that is then lost during sleep (for example high nasal resistance or increased external compression from neck obesity). Finally, there may be mechanical problems such that muscular activity fails to dilate the pharyngeal lumen effectively.

There are many case reports of obvious anatomical abnormalities provoking obstructive sleep apnoea, for example tonsillar hypertrophy, pharyngeal oedema, tumours, acromegaly, mucopolysaccharidoses, and mandibular or maxillary underdevelopment. These reports show that pharyngeal narrowing (asymptomatic whilst awake) can provoke obstructive sleep apnoea, but such diagnoses represent only a small proportion of cases.

The majority of patients with obstructive sleep apnoea are overweight. In many clinics the average obesity index is well over 30 kg/m ^{2}, equivalent to being about 30 per cent overweight, for example 95 kg (15 stone) at a height of 1.78 m (5 ft 10 in). Weight loss can certainly cure obstructive sleep apnoea, and all studies identifying risk factors have found obesity to be dominant, accounting for up to 40 per cent of the variance in severity.

Most groups have found neck circumference to be a better predictor of severity of obstructive sleep apnoea than obesity itself, suggesting that it is neck obesity and external pharyngeal loading that is important. Animal studies have shown that only a small amount of extra external pressure over the pharynx is required to collapse it during sleep, and recent imaging studies have suggested that quite small amounts of extra fat do occur either side of the pharynx in patients with obstructive sleep apnoea, together with larger amounts subcutaneously.

Although general obesity is related to neck obesity, the overall correlation is only about 0.75. This is because fat distribution varies considerably between individuals. The 'female' distribution tends to be in the lower body and the 'male' distribution is more central. Thus a man who is not particularly overweight can have a large neck and vice versa.

As mentioned earlier, there is evidence that some of the upper airway dilator muscles (e.g. genioglossus) of obese patients with obstructive sleep apnoea are actually working harder than normal, perhaps as compensation for the added external loading from neck obesity. Compensations by the respiratory system for other types of extra loading have been shown to be much less active during sleep.

In summary, overall the evidence suggests that most obstructive sleep apnoea in adults is due to loading of the upper airway caused by obesity. This external loading can be fended off during wakefulness but not during sleep, when the withdrawal of postural muscle tone allows the pharyngeal dilators to be overwhelmed, leading to excessive narrowing or collapse of the airway, with consequent apnoea.

However, not all adult sleep apnoea can be explained by obesity or intrapharyngeal anatomical abnormalities. The significance of marked retro- or micrognathia for obstructive sleep apnoea was recognized early on, particularly in children (Pierre–Robin syndrome). Careful cephalometric studies of facial and skull morphology have revealed that some patients with obstructive sleep apnoea have longer faces, retropositioning of the mandible (measured as a more acute angle between the sella to nasion and nasion to supramentale planes), a downward movement of the hyoid, elongation of the soft palate, and a narrower anteroposterior distance behind the tongue. Some, or all, of these changes may be secondary to many years of sleep apnoea rather than part of the cause. However, the retropositioning of the mandible may be contributory, and surgery to advance the mandible may be curative in carefully selected cases.

P>Retropositioning of the mandible may be a legacy from childhood. There is good evidence that nasal blockage and mouth breathing very early in life alter facial development (the so-called 'adenoidal facies'), and one feature of this is mandibular retropositioning. Following early adenoidectomy and resumption of nasal breathing, the mandible can return to its normal position. One theory is that mandibular underdevelopment and obesity are two relatively common independent risk factors for obstructive sleep apnoea that together may be synergistic.

Other factors provoking obstructive sleep apnoea

Alcohol is a potent reducer of muscle tone, and can further reduce pharyngeal dilator muscle tone during sleep. It is well known that alcohol worsens snoring, but it can also convert snoring to full apnoea. Other sedatives, such as the benzodiazepines, barbiturates, and opiates, can do the same, and this has important consequences for anaesthesia in such patients. Sleep deprivation itself can reduce upper airway muscle tone during subsequent sleep, provoking a vicious circle, whereby apnoea causes sleep disruption, causing worsening apnoea.

Nasal blockage can contribute to the tendency of the pharynx to collapse by lowering intrapharyngeal pressure. If extra effort has to be made to inspire through a high nasal resistance, there will be a greater vacuum effect in the pharynx, increasing its tendency to collapse. Once collapse occurs, flow ceases, pharyngeal pressure returns to atmospheric, the lumen opens, and the cycle repeats. This certainly leads to snoring, but may no longer be very important when there is full apnoea. However, nasal obstruction may contribute long term to sleep apnoea by damaging the pharynx through years of snoring, making it more collapsible, but improving nasal patency rarely cures obstructive sleep apnoea.

Hypothyroidism is associated with obstructive sleep apnoea, but the mechanism s not clear. It may be through weight gain or through tissue or fluid deposition in the pharynx. Alternatively, a low thyroxine level may interference directly with muscle function.

Immediate consequences of sleep apnoea

Upper airway narrowing, sometimes with complete apnoea, usually commences as sleep passes from awake to stage 2. Once significant obstruction occurs there will be increasing respiratory effort to try and overcome it. The length of such events is highly variable, ranging from only a few seconds to well over a minute. At some point arousal occurs, with an improvement in upper airway resistance, resolution of any asphyxia, and then a return to sleep, whereupon the cycle repeats (Fig. 3, Fig. 4, and Fig. 5). Hypoxaemia and mild hypercapnia usually accompany these periods of obstructed breathing. If there is complete apnoea, the rate of fall of $Sa{O_2}$will depend mainly on the amount of oxygen stored in the lungs. This depends on the functional residual capacity since apnoeas occur at end-expiration, preventing inspiration. The length of the apnoea also determines how low the $Sa{O_2}$ will fall, and varies considerably between patients. The consequences of such hypoxaemia and hypercapnia are not clear. Because the blood gas derangements are so transient they may do little harm, unless there is already ischaemic heart disease, for example.

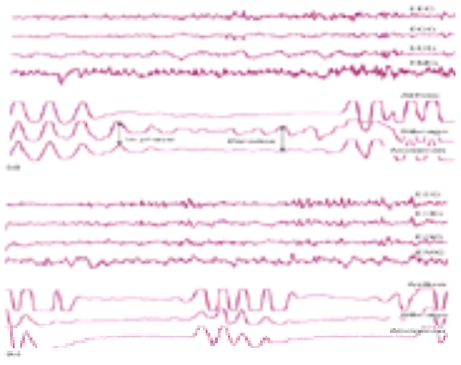

Fig. 3 Obstructive (a) and central (b) apnoeas (16-s traces): (a) airflow ceases, but rib-cage and abdominal movements persist and become paradoxical; (b) rib-cage and abdominal movements cease as well as airflow.

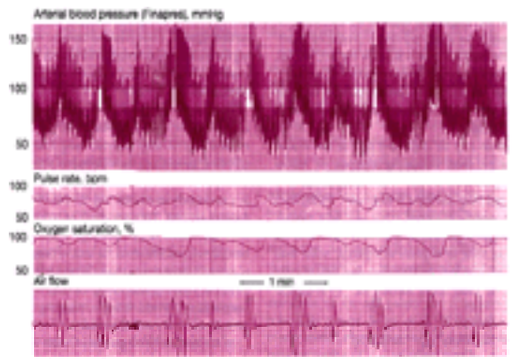

Fig. 4 Five-min tracing from a patient with obstructive sleep apnoea. The rises in blood pressure (top trace) and heart rate (second trace) coincide with the cessation of each apnoea and an arousal. During each apnoea (evident from the bottom airflow trace) each frustrated inspiratory effort is accompanied by a fall in blood pressure (pulsus paradoxus).

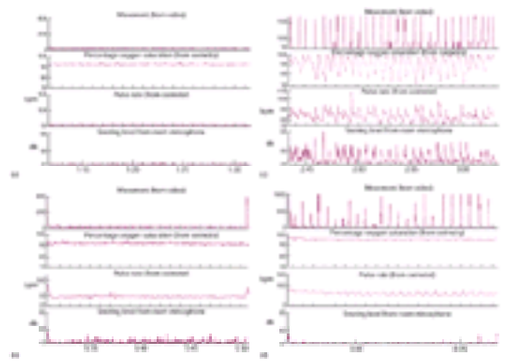

Fig. 5 Short sleep tracings of body movement, SaO_2, pulse rate, and snoring level in four different subjects. (a) Normal subject (no fluctuations in any signals), 20 min. (b) Patient with continual low level snoring and almost no arousals, 20 min. (c) Patient with classical obstructive apnoeas, evident from the snoring–silence–snoring pattern together with movements and oscillations in the pulse and SaO_2, 20 min. (d) Patient with periodic movements of the legs during sleep, recurrent arousal (oscillations in pulse and body movements), but no evidence of a respiratory cause (no snoring or SaO_2 dips), 10 min. A video recording of the whole night is always available and can be viewed when the exact cause of abnormal signals is not immediately obvious.

Hypoxaemia was believed to play an important part in the arousal response that saves the patient from continuing asphyxia. In animal models, removal of the carotid body abolishes significant ventilatory response to hypoxaemia during sleep and there is no arousal. Giving extra added oxygen does prolong apnoeas to a small extent and delay arousal. However, recent evidence suggests that the main arousal stimulus is the actual respiratory effort being made in response to asphyxia, rather than the asphyxia per se. Normal subjects tend to wake when they have to make respiratory efforts about three times above the normal (10–20 cmH_2O pleural pressure swings). This degree of effort is easily reached in obstructive sleep apnoea, when pressures down to –80 cmH_2O can be recorded during the frustrated inspiratory efforts. Such pressures can also be reached by heavy snorers, even if they do not develop hypoxaemia, and this also leads to arousals.

In terms of symptoms, the most important consequence of sleep-induced upper airway narrowing is sleep fragmentation. The original methodology of sleep analysis, using coarse 30-s epochs to stage sleep, effectively glossed over the multitude of transient arousals that are the main consequence of obstructive sleep apnoea. Superficially, a sleep hypnogram in a moderately severe case (Fig. 2) could look almost normal despite hundreds of arousals. The importance of trying to measure these has recently been appreciated, and technology to measure arousals automatically is being developed. However, the level of sleep disruption (number and 'size' of arousals) necessary to cause daytime symptoms is not known. There is a clear, but variable, relationship between increasing sleep disruption and deteriorating daytime function, meaning that there is no clear cut off between normality and abnormality.

In addition to blood gas disturbances and sleep disruption, there are many other consequences of obstructive sleep apnoea. During the apnoea there is activation of the diving reflex that produces bradycardia, particularly when there is associated hypoxaemia. Upon arousal there is a sudden pulse rate and blood pressure rise, probably due to activation of the sympathetic nervous system as part of the arousal process itself. During the actual frustrated inspiratory efforts, blood pressure falls with each reduction in intrathoracic pressure (pulsus paradoxus) and, in conjunction with the blood pressure rise on arousal, produces a very characteristic trace (Fig. 4). As well as increased nocturnal catecholamine secretion in patients with obstructive sleep apnoea, there is also a suppression of growth hormone and testosterone levels. There is marked polyuria during sleep (a reversal of the normal relative oliguria), but the mechanism is not clear. It may be related to the recurrent arousals, or to increased atrial natriuretic peptide (ANP) production following right atrial distension due to large inspiratory efforts.

It will be clear from this account that there are grey areas of uncertainty regarding definition and measurement of significant aspects of sleep apnoea. We discussed earlier that original definitions centred on the actual obstructive event. There has now been a shift towards trying to look more closely at the most important result—sleep fragmentation. This is particularly necessary now we know that 10-s apnoeas are not the only result of upper airway narrowing during sleep that can provoke multiple arousals and daytime sleepiness. Since heavy snorers can have considerable sleep fragmentation without significant falls in SaO_2, examining blood gas abnormalities (for example with an oximeter) is not always good enough either. The implications for this in terms of investigations are discussed later. A considerable amount of effort is being put into establishing which variable, measured during a sleep study, best defines the severity of the disorder. At present there is no clear answer and therefore there remain many different approaches to diagnosis and management.

Symptoms and presentation

The main symptom of obstructive sleep apnoea is daytime hypersomnolence, and this correlates broadly with the degree of sleep disruption. Early in the development of the disorder the daytime sleepiness is little more than often experienced by normal people after a few disturbed nights. Whilst occupied there is little difficulty in concentrating and staying awake, but once activities become more boring, unwanted sleepiness intervenes. Initially this may be viewed as normal, such as falling asleep in front of the television every evening. As the sleep disruption worsens there will be interference with an increasing number of activities. Of particular importance is sleepiness whilst driving. Sleepiness can be devastating, particularly on long motorway journeys after dark, when sensory stimulation is low. Initially there will be lane wandering with sudden arousal and correction. Accidents involving driving off the road, or driving into vehicles in front, are more common in patients with obstructive sleep apnoea. Sleepiness also impinges greatly on work performance and home life. The patient will develop a reputation for slothfulness and lack of interest.

It is important to ask the right questions to assess sleepiness. It is not the same as tiredness, which is a lack of energy or desire to get up and do anything, without a desire to sleep. Because of the insidious onset of obstructive sleep apnoea, any sleepiness may be regarded as normal by the patient, and thus situational questions need to be asked such as 'how often do you have to pull off the road whilst driving owing to sleepiness?' rather than just 'are you sleepy?' A well validated and simple way to do this is with the Epworth Sleepiness Scale (Fig. 6). Objective sleepiness can be assessed in the sleep laboratory by measuring how long the patient takes to fall asleep on a number of occasions across the day. This is useful for research purposes but adds little to the clinical management of such patients. A list of other symptoms seen in obstructive sleep apnoea is given in Table 3. It is sad to say, but the corrosive effect of sleepiness on all aspects of a patient's life has often been present for years before someone (usually not a doctor) tumbles to the diagnosis.


EPWORTH SLEEPINESS SCALE

Fig. 6 Questionnaire scale to assess subjective sleepiness. The scores for each answer (0–3) are summed to give a range from 0 (no sleepiness at all) to 24 (maximally sleepy). The upper limit of normal is about 9, and most patients with symptomatic obstructive sleep apnoea are in the middle teens.

A typical case history would be that of a middle-aged man complaining of increasing daytime sleepiness. It is usually some specific event that prompts initial consultation, such as falling asleep whilst driving, operating machinery, or during an important board meeting. There will be a long history of gradually worsening snoring with apnoeas, possibly witnessed by the spouse, who will probably have moved out of the bedroom owing to the noise. There is likely to have been a weight gain over the last few years with an obesity index of greater than 30 kg/m^2 and a collar size of 17 inches or more. There is usually a history of fairly high alcohol intake and smoking. On examination there may be nasal stuffiness, evidence of a small lower jaw (such as teeth crowding or several extractions for this problem), and a small pharynx with mucosal bogginess and wrinkling. Of course, it should be stressed that not all these features are likely to be present in one individual.

Part of the history and examination of patients with possible obstructive sleep apnoea should be directed towards precipitating factors such as hypothyroidism and acromegaly. Other diagnoses such as mucopolysaccharidosis, pharyngeal tumours, tonsillar hypertrophy, neurological disorders, and significant retrognathia will be more obvious.

Diagnosis

Following the history and examination, further outpatient tests may be appropriate, for example thyroxine or growth hormone estimations. Blood gases and simple lung function tests may be necessary if associated diurnal respiratory failure is suspected. A raised haemoglobin may also signify diurnal respiratory failure, as will a raised venous bicarbonate. Obstructive sleep apnoea tends to go with the findings that constitute the so-called 'syndrome X', namely hypertension, obesity, and insulin resistance. Blood pressure and blood sugar should be measured.

Unless the presenting problem turns out not to be sleep related, some form of sleep study will be required. In the past, the usual procedure was to employ full polysomnography, which measured sleep state and respiratory variables (Fig. 2 and Fig. 3). This investigation and its analysis is expensive and time consuming, particularly if all recurrent arousals are documented. The primary requirement is to assess sleep fragmentation, establish if a respiratory problem is responsible, and decide if upper airway obstruction is the primary cause. Full polysomnography, properly interpreted, will usually allow this, with the EEG and EMG giving good information on sleep disruption, and aspects of respiration deduced from rib-cage/abdominal movement transducers, oronasal airflow, and snoring and continuous oximeter recordings. However, there is considerable signal redundancy in such recordings, and the essential derivatives—sleep disruption and respiration—can be assessed in much simpler ways (Fig. 5). Most clinical respiratory sleep laboratories have abandoned routine, conventional polysomnography because of its unnecessary expense.

Sleep fragmentation can be inferred from a variety of signals. The most sensitive appears to be autonomic markers of brainstem activation, such as blood pressure and pulse rate rises. In addition, since most abnormal respiratory events will end in some form of arousal, counting body movements provides some guide to the degree of sleep fragmentation, and may be most predictive of daytime symptoms. Upper airway obstruction can be inferred from snoring, a particular inspiratory pattern on a nasal flow tracing (flow limitation), paradoxical ribcage/abdominal movements, and from pulsus paradoxus visible on a beat to beat blood pressure tracing (now easily obtainable non-invasively, Fig. 4). Many simple, commercial monitoring systems can be used to record these signals and to assess the extent of the sleep fragmentation and whether upper airways obstruction is the likely cause. Recent work suggests that these simpler measures can predict sleepiness in obstructive sleep apnoea, and its response to treatment, at least as well as EEG-based approaches, which are clearly not the gold standard they were once thought to be. The attention paid to each signal, and perhaps the exact sleep study system used, will depend to some extent on the condition under investigation. Figure 5 shows data provided by the system in routine use in our laboratory, designed primarily to identify obstructive sleep apnoea and its variants, but it will also identify central sleep apnoea (see below) and non-respiratory problems such as periodic movements of the legs during sleep.

Because of the imprecise relationship between the number of abnormalities on a sleep study and the severity of symptoms, trying to count them exactly is pointless, particularly given that there can be considerable night to night variation. Hence, the reporting of sleep studies tends to be more qualitative than previously, with divisions simply into mild, moderate, and severe. The reporter is essentially trying to see if there is an adequate and understandable explanation for the patient's symptoms.

Treatment

Once it is established that the patient's symptoms are likely to be due to sleep disruption from sleep-induced upper airway obstruction, then therapy has to be tailored to symptom severity.

Mild symptoms may resolve with simple treatments and advice (Table 4). Weight loss is undoubtedly effective, but often very difficult to achieve. If sleep disruption only occurs whilst supine, when upper airway obstruction tends to be worst, then learning to lie on one's side may be helpful. Stopping sedatives and evening alcohol can help. Initial enthusiasm for the tricyclic antidepressants has waned, although they may slightly improve mild cases. They are believed to work through REM sleep suppression and by improving upper airway tone. No other drug has shown any consistent effect.

If symptoms are severe, there is one effective therapy—nasal continuous positive airway pressure (NCPAP). This treatment involves wearing a small mask (Fig. 7) over the nose whilst asleep, kept above atmospheric pressure by a pump. Pressures in the region of 10 cmH $_2$O are enough to splint open the pharynx and resist collapse, allowing unobstructed breathing and undisturbed sleep (Fig. 8). The response is dramatic, in terms of both physiology and daytime symptoms. These resolve rapidly, even after one night of treatment. There are several randomized, placebo-controlled trials proving beyond doubt the symptomatic benefit. The unpleasantness and unaesthetic appearance of this treatment initially repels patients, but once the benefits have been experienced, acceptance is high. Off-the-shelf systems, with comfortable soft masks, are now available for home use at about £300 each. Such equipment will last for years and represents extraordinary value for money given the enormous improvement in patient functioning that they produce. NCPAP machines have recently been introduced that hunt automatically the pressure required by the patient to overcome their obstructive sleep apnoea. These are mainly used in the sleep laboratory as there is no evidence to support their more general use.

Fig. 7 A soft silicone nasal mask and its head gear, used in the treatment of obstructive sleep apnoea.

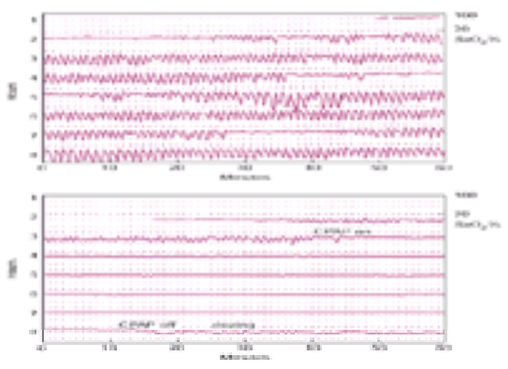

Fig. 8 Two all-night oximetry tracings from a patient with obstructive sleep apnoea, before treatment and during his first night on nasal CPAP. Each tracing starts top left and finishes bottom right. Each tracing is continuous for 8 h with the vertical axis for each individual line scaled 70–100 per cent SaO_2.

Much effort is put into helping patients to become established on NCPAP, through attentive education and comfort-improving measures such as humidification. Once established on NCPAP, a patient with obstructive sleep apnoea is likely to require it for life unless he can lose a significant amount of weight. This may only be achieved through gastric surgery, such as silastic ring gastroplasty to reduce food consumption. Another surgical treatment, whose popularity is decreasing, is uvulopalatopharyngoplasty, which consists of removing part of the soft palate and any residual tonsils, and 'tightening up' the sidewalls of the pharynx. Although it can reduce snoring, its success rate at treating obstructive sleep apnoea is not good: approximately 50 per cent of patients experience a 50 per cent improvement in the number of apnoeas per hour. Attempts to improve the selection of patients have had very limited success, although thin patients with large soft palates, residual tonsils, and milder disease do the best. The operation may have a more significant role in the treatment of snoring-induced arousals than full apnoeas, although this is not established.

Other operative techniques involving advancement of the mandible (and sometimes the maxilla) may be appropriate in highly selected cases. Tracheostomy was the first therapy ever tried and was (of course) very effective. It may still be appropriate in occasional patients.

A newer approach has been the use of mandibular advancement devices, worn in the mouth at night (Fig. 9). These hold the lower jaw closed and forward, thus increasing the space behind the tongue and hence pharyngeal volume. They are undergoing extensive trials in a variety of situations, but the situation is complicated by the plethora of such devices available. The current conclusion is that they do work, but less so as the severity of the obstructive sleep apnoea (and usually therefore the obesity) increase. Their main use currently is in the control of unacceptable snoring.

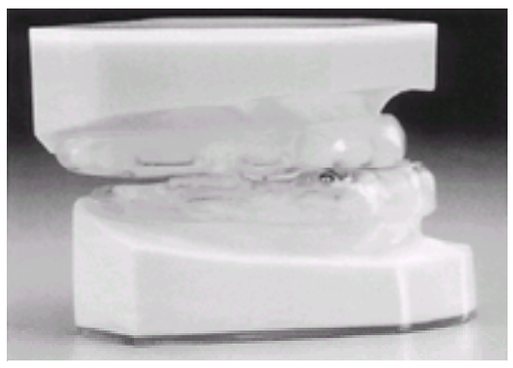

Fig. 9 Example of a mandibular advancement device, worn in the mouth at night. These hold the lower jaw forward and closed, thus increasing pharyngeal dimensions. They are used extensively for the control of snoring but they are not very effective in any more than mild obstructive sleep apnoea.

Epidemiology

Given difficulties over definition, the prevalence of symptomatic obstructive sleep apnoea is hard to establish, and will depend on where an arbitrary cut-off is drawn. In an early study, about 0.3 per cent of men aged 35 to 65 years clearly had severe, symptomatic obstructive sleep apnoea, requiring nasal continuous positive airway pressure (NCPAP) therapy and responsive to such treatment. However, about 5 per cent had more than five dips of more than 4 per cent SaO_2 per hour, one suggested threshold for normality. However, most of these subjects were not obviously symptomatic and would not have wanted a treatment such a NCPAP. Overall in this study, sleepiness correlated with snoring, and more sleepiness seemed to be due to snoring than classical sleep apnoea. Other studies in Israel, the United States, and Italy have found prevalences of 'significant' sleep apnoea in the 0.5 to 2 per cent range.

Predictors of sleep apnoea in these prevalence studies have been obesity, snoring, age, self-reported sleepiness, and alcohol consumption. Snoring is more common in men than women, and obstructive sleep apnoea syndrome itself is about five to ten times more common in men. The prevalence in women probably increases after the menopause with redistribution of body fat to a more male-like, upper body, distribution.

If these prevalence studies are correct, then obstructive sleep apnoea is more common than sarcoidosis and fibrosing alveolitis combined; and every chest physician in the United Kingdom should have about 100 patients on NCPAP.

Prognosis and long-term complications

Although the main reason for treating obstructive sleep apnoea is to relieve the daytime symptoms, mainly sleepiness, there is also limited evidence that these patients have an increased cardiovascular mortality (Fig. 10). Two studies have looked at the long-term survival of patients with treated obstructive sleep apnoea and compared them with some form of untreated control patients. Both found an increased mortality due to cardiovascular events such as myocardial infarction and stroke in the untreated group. The cause of this is not clear and a variety of hypotheses have been advanced. The main problem is that, because a variety of cardiovascular risk factors also contribute to the production of obstructive sleep apnoea (central obesity, smoking, alcohol), it is difficult to identify the real contribution made independently by obstructive sleep apnoea to cardiovascular deaths. Possibilities include sustained hypertension, intermittent nocturnal hypertension, increased catecholamine release, hypoxia-induced cardiac arrhythmias, insulin resistance, hyperlipidaemia, and left ventricular hypertrophy. As yet, most of these risk factors have not been shown to differ between patients with obstructive sleep apnoea and well-matched controls. The exception are the episodic blood pressure rises that occur with each apnoea, a very small increase in diastolic blood pressure (about 2 mmHg), and a small carryover of the raised systolic nocturnal blood pressure into the first few hours of wakefulness (Fig. 4 and Fig. 11). These nocturnal surges in blood pressure may be harmful to the vascular system, and absence of the usual nocturnal fall in blood pressure is said to be a marker of cardiovascular morbidity. Treatment with nasal NCPAP prevents these surges in blood pressure by abolishing the apnoea-induced arousals, and this may be the mechanism by which NCPAP improves mortality in obstructive sleep apnoea, although there are many other possible explanations.

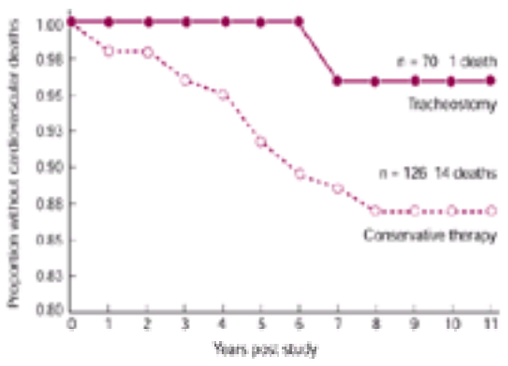

Fig. 10 Long-term survival without a cardiovascular death of 196 patients with obstructive sleep apnoea. The two groups consisted of one with 70 patients who accepted the definitive treatment then available of tracheostomy, and one with 126 patients who declined such therapy and merely attempted weight loss. (Reproduced with permission from C. Guilleminault and M. Partinen (1990). *Obstructive sleep apnoea syndrome*. Raven Press, New York.)

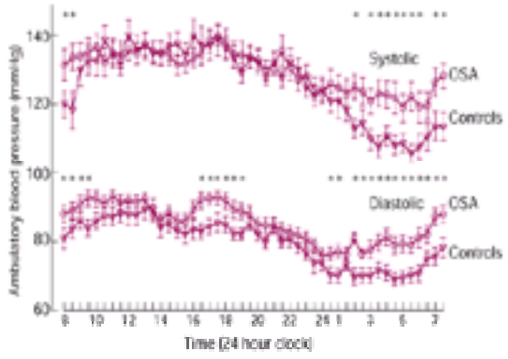

Fig. 11 Twenty-four-h blood pressure records (half hourly) from a group of patients with obstructive sleep apnoea, and individually matched control subjects. The asterisks indicate the times when there were significant differences between the groups. Note the higher systolic and diastolic pressures at night in the patients with obstructive sleep apnoea, with persistence of this into the early morning after awakening, and a generally higher diastolic for much of the day. (Reproduced from Davies C, *et al* (2000). *Thorax*, **55**, 736–40, with permission.)

Nasal CPAP produces a demonstrable improvement in vigilance and sleepiness, which in a randomized, controlled trial was reflected in improved performance on a driving simulator. There is also some evidence that the higher car accident rate in patients falls after treatment. In the 3 years prior to diagnosis, health care utilization by patients with obstructive sleep apnoea, compared to matched controls, is about twice as high. Recent data suggest that this too falls following treatment.

Conclusions

There has been a move away from considering obstructive sleep apnoea to be a condition that one either has or does not have. There is a continuum from light intermittent snoring through to severe, all-night, obstructive sleep apnoea. An analogy can be drawn with hypertension, where there is also a continuum of severity, variability in the measurement, only moderate correlation between the measured abnormality and the physiological consequences, uncertainty over the most relevant way to measure it (one-off versus 24-h blood pressure monitoring), and the target organ damage (e.g. atheroma) may not just be due to the blood pressure. In addition, benefits of treatment have to be weighed against any side-effects. Sleep-induced upper airway obstruction should really be viewed in a similar way.

Sleep-induced hypoventilation and central sleep apnoea

So far we have discussed the sleep-related disorders of breathing that are due to sleep-induced narrowing of the upper airway. Breathing during sleep may also decrease, not because of upper airway obstruction, but because of a reduction in central output to the respiratory muscles—so-called central, rather than obstructive, apnoea (Fig. 3). There are many causes for central sleep apnoea or hypoventilation, and Table 5 shows one way of classifying them. Some of the central apnoeas disturb sleep and present with daytime sleepiness, whereas others tend to present with symptoms of respiratory failure, such as morning headaches with cyanosis and confusion, ankle oedema, and shortness of breath on exertion.

Absent ventilatory drive

Brainstem abnormalities may damage the areas responsible for automatic chemical control of ventilation. Whilst awake, the wakefulness-related ventilatory drive may be adequate to maintain PaO_2 and $PaCO_2$ levels, but on falling asleep, drive falls or even disappears with marked hypoventilation (or apnoea) and hypoxaemia. Arousal is then necessary to restore the blood gases. This failure of brainstem automatic control (known as Ondine's curse) can be congenital, or may be acquired as the result of a stroke, infection, surgical damage, multiple sclerosis, or compression by a tumour or syrinx.

Reduction of chemical drive can occur as a secondary problem when ventilation is reduced by mechanical problems such as chronic airways obstruction or weak respiratory muscles. It appears that chronic underventilation can lead to blunting of ventilatory drive, perhaps through alteration in acid–base buffering in the brainstem, and can also lead to marked falls in ventilation during sleep.

Unstable ventilatory drive

The wakefulness-related ventilatory drive stabilizes ventilation and prevents it from falling below a certain level. If reasons for ventilatory instability exist then, by removing this stabilizing effect, sleep will allow periodic respiration to develop. The usual provoker of instability is an increased drive to breathe. Control theory shows that increasing the gain in any feedback system causes instability through overshoot and undershoot. A good example of this is the hypoxaemia of altitude, which accentuates the response to change in arterial carbon dioxide tension, promoting instability. When sleep occurs there is the usual fall in ventilation. This leads to increased hypoxaemia and a rise in carbon dioxide tension, causing increased ventilation, both provoking arousal. This provides extra awake ventilatory drive, which restores the blood gases, and the cycle repeats. Thus, periodic breathing with recurrent arousals is very common at altitude, with the expected daytime consequence of sleepiness and complaints of insomnia. Acetazolamide produces a metabolic acidosis and increases ventilation overall, the hypoxaemia is relieved, and thus the ventilatory response to carbon dioxide becomes less steep. Both these factors restore stability and reduce periodic respiration.

In left ventricular failure there is also extra ventilatory drive, mainly due to stimulation of interstitial lung receptors by the raised pulmonary venous pressure. In conjunction with the longer circulation time seen in heart failure, this also provokes instability with waxing and waning of the ventilation. This periodic breathing, or Cheyne–Stokes respiration, is quite common in heart failure, and through sleep disruption produces daytime sleepiness and complaints of nocturnal dyspnoea (Fig. 12). The patient is usually aware that on arousal the dyspnoea disappears within a few seconds, unlike the paroxysmal nocturnal dyspnoea of pulmonary oedema which usually takes at least 15 min or so to abate following getting out of bed. Treatment with either overnight oxygen or acetazolamide can sometimes reduce the periodicity and improve both sleep quality and symptoms. There has been some recent interest in NCPAP as a treatment for left ventricular failure and periodic breathing. The mechanism of any action is unknown, and at present the evidence does not justify its general use.

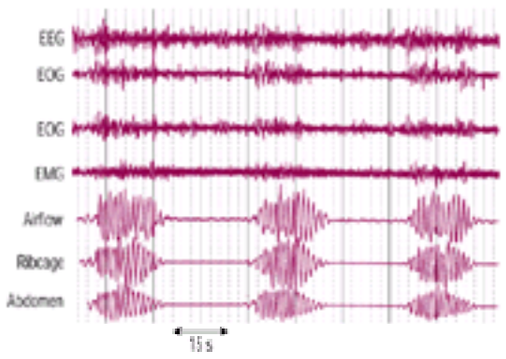

Fig. 12 Tracing of Cheyne–Stokes respiration from a patient with poor left ventricular function but no radiological or clinical evidence of current pulmonary oedema. With each return of respiration there is arousal from sleep (not clearly visible with this compressed EEG tracing).

Instability of respiratory control can also occur in normal subjects in the early stages of sleep or if sleep is disturbed for other reasons. This is because sleep depth is oscillating back and forth between drowsy wakefulness and light sleep, with the ventilatory drive oscillating as well.

REM sleep apnoeas

During normal REM sleep the phasic bursts of eye movements are associated with transient falls in ventilation, and even the occurrence of apnoeas. The rib-cage muscles are affected most, but diaphragmatic excursion can also fall. Such periodicities are entirely normal.

As discussed earlier, the REM sleep inhibition of most muscles (apart from the diaphragm) can greatly reduce overall ventilation when the accessory muscles of respiration are needed for breathing. Thus on entering REM sleep there can be profound falls in ventilation and $Sa\mathrm{O}_2$ in patients with neuromuscular diseases, chest-wall abnormalities, and chronic airways obstruction.

Generalized neuromuscular diseases tend to involve the respiratory muscles in concert with other muscles. However, in some disorders the respiratory muscles, particularly the diaphragm, may be involved very early on, at a time when other muscles are virtually normal. A particular example of this is adult-type acid maltase deficiency, where patients may present in respiratory failure whilst still able to walk normally. REM-sleep-related hypoxaemia may be the first sign that there are problems, and it is not known whether this actually accelerates the onset of eventual diurnal respiratory failure, or is merely a marker that respiratory failure will soon follow. Sometimes there may be associated upper airway obstruction during REM sleep, leading to even larger falls in $Sa\mathrm{O}_2$. Overnight oximetry studies will indicate the degree of hypoxaemia but will not establish if there is additional upper airway obstruction.

There has been great interest in the REM-sleep-related hypoxaemia seen in chronic airways obstruction. It was thought possible that these hypoxic episodes might be the reason why some patients developed respiratory failure but others did not. However, it appears that REM sleep hypoventilation and a fall in $Pa\mathrm{O}_2$ is fairly universal in this group of patients. If the patient is initially well oxygenated and on the flat part of the haemoglobin dissociation curve, the fall in $Sa\mathrm{O}_2$ (which is usually what is monitored) is not particularly dramatic; however, if the patient is initially poorly oxygenated and on the steep part of the curve, similar hypoventilation will produce dramatic falls in $Sa\mathrm{O}_2$. As yet there is no evidence that these REM sleep falls in $Sa\mathrm{O}_2$ contribute to the morbidity and mortality of patients with chronic airflow obstruction, although some centres have shown that overnight oxygen therapy reduces arousals, thus improving sleep quality. The main problem is that the falls in $Sa\mathrm{O}_2$ can look superficially like obstructive sleep apnoea leading to an erroneous diagnosis and the inappropriate use of NCPAP.

Reflex apnoea

Central respiratory output can be modified by a number of reflexes from receptors in the upper airway. There is a reflex from the pharynx that inhibits inspiratory flow when the pharynx is being sucked in and collapsed. This makes teleological sense, as a slowing of inspiratory flow would reduce the tendency to collapse. There are some patients with pharyngeal collapse who, instead of struggling to inspire against the blocked airway, simply stop breathing until they finally arouse, presumably due to the fall in $Pa\mathrm{O}_2$ and rise in $Pa\mathrm{CO}_2$. This then appears as a central apnoea, despite the aetiology being pharyngeal collapse. This tends to happen when the patient is supine, with snoring or ordinary obstructive apnoeas when decubitus. If the pharynx is anaesthetized experimentally, then inspiratory attempts return, suggesting that superficial receptors are responsible. These patients usually present with typical histories of obstructive sleep apnoea, respond to NCPAP, and can be managed in the same way.

Apparent central apnoea

The diagnosis of central apnoea depends on demonstrating the absence of respiratory effort when airflow at the nose and mouth stops. Surface measurements of rib-cage and abdominal movement are usually employed as evidence of continuing respiratory effort. However, in two circumstances, marked obesity and muscular weakness, the surface transducers may fail to register that inspiratory efforts are still being made (although more sensitive measures of inspiratory effort, such as oesophageal pressure tracings, will usually do so). Obesity lessens the sensitivity of surface transducers, and with muscle weakness the inspiratory muscles may not be able to move the chest wall detectably against a closed upper airway.

Overnight ventilation for central sleep apnoea or hypoventilation

The chronic ventilatory failure associated with some neurological disorders (e.g. acid maltase deficiency, postpoliomyelitis syndrome, motor neurone disease, Duchenne dystrophy) usually progress rapidly to death, even when quality of life is otherwise very good. The same is true of chest-wall restrictive disorders such as scoliosis, as well as the ventilatory failure that can develop many years after extensive thoracoplasty. However, supporting breathing overnight can fully reverse ventilatory failure, and the response to treatment can be dramatic, with resolution of all symptoms and restoration of normal blood gases, even when off the ventilator during the day. The mechanism by which supporting breathing at night corrects ventilatory failure is not clear and there are various possibilities. Firstly, it may simply be that the respiratory muscles are rested so that they can respond better to the demands of the respiratory centre during the day. Secondly, it may be that improving the blood gases at night, and preventing the marked REM sleep deteriorations, leads to resetting of the respiratory centre back towards normal, that reverses an acquired blunting of drive. Tricyclic antidepressants such as protriptyline can virtually abolish REM sleep periods and their associated hypoxaemia and have been shown to improve daytime blood gases temporarily, suggesting that simply abolishing these periods of particular hypoxia can help. Thirdly, by increasing chest-wall and lung excursion (tidal volumes in excess of the voluntary vital capacity are sometimes obtained) overall respiratory compliance may improve, allowing the muscles to work more efficiently. Whatever the explanation, there is no doubt that this is a life-saving therapy that in certain conditions can add decades of active life.

Most of the original techniques to support ventilation overnight evolved from the iron lung that was developed to support poliomyelitis victims. Evacuating the air from around the chest expands the lungs, recreating the normal way of breathing. A range of devices involving airtight jackets and shells over the chest were developed, but required much attention to detail and often individual, tailor-made systems. Unfortunately a specific complication, resulted from the abolition of spontaneous ventilatory drive to the diaphragm and pharyngeal muscles, upper airway collapse can occur during the mechanical inspiratory phase and greatly limit efficacy. The recent development of comfortable nasal and face masks has revolutionized the overnight ventilation of these patients. Positive pressure ventilation can be used via a face mask, or more comfortably via the nasal masks used for NCPAP (Fig. 7). Although there are still many problems to be overcome when establishing patients on such equipment (particularly mask comfort and air leaks through the mouth when using nasal masks), the systems can be bought off the shelf ready to use (current cost approximately £3500). Most units now use positive pressure ventilation in preference to the negative pressure systems.

Electrical pacing of the diaphragm is occasionally used for supporting ventilation in conditions where the phrenic nerve and diaphragm are intact and the problem is central. This involves the implantation of bilateral phrenic electrodes and induction coils under the skin that are activated by external induction coils.

Conclusions

Ventilatory failure should not be viewed as a diagnosis. It is a finding that requires a specific explanation. Many of the causes become most obvious at night, when they produce dramatic falls in the level of SaO_2 due to further failure of ventilatory drive mechanisms. Specific treatments aimed at reversing these sleep-related failures of ventilatory drive can produce rapid and satisfactory resolution of symptoms, as well as prolongation of life.

Further reading

Gastaut H, Tassinari CA, Duron B (1966). Polygraphic study of the episodic diurnal and nocturnal (hypnic and respiratory) manifestations of the Pickwick syndrome. *Brain Research* **2**, 167–86.

Guilleminault C, Stoohs R, Duncan S (1991). Snoring (1). Daytime sleepiness in regular heavy snorers. *Chest* **99**, 40–8.

Jenkinson C, Davies RJO, Mullins R, Stradling JR (1999). Randomised prospective parallel trial of therapeutic nasal continuous positive airway pressure (NCPAP) against sub-therapeutic NCPAP for obstructive sleep apnoea. *Lancet* **353**, 2100–5.

Remmers JE, de Broot WJ, Sauerland EK, Anch AM (1978). Pathogenesis of upper airway occlusion during sleep. *Journal of Applied Physiology* **44**, 931–8.

Stradling JR (1995). Obstructive sleep apnoea: definitions, epidemiology and natural history. *Thorax* **50**, 683–9.

Stradling JR (1997). Practical approach to sleep disordered breathing. In: Farthing M, ed. *Horizons in medicine*. Royal College of Physicians, London.

Sullivan CE, Issa FG, Berthon-Jones MCHR, Eves L (1981). Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares. *Lancet* **i**, 862–5.

Weiss JW, Remsburg S, Garpestad E, Ringler J, Sparrow D, Parker JA (1996). Hemodynamic consequences of obstructive sleep apnea. *Sleep* **19**, 388–97.

17.9 Bronchiectasis

D. Bilton

Introduction

The definition of bronchiectasis is based on morbid anatomy first described by Laennec, who in 1819 found abnormal chronic dilatation of the bronchi in an infant who died following whooping cough. The word itself is from the Greek *bronchion* (wind pipe or tube) and *ektasis* (stretched out or extension). By 1891 it was recognized in a textbook of medicine that bronchiectasis was 'not a separate disease' but 'a result of various affectations of the bronchi', hence bronchiectasis is not a final diagnosis but a final pathology of a number of causes which may require their own specific treatment.

Prevalence

Up to 1953, estimates of the prevalence of bronchiectasis in the United Kingdom varied from 0.77 to 1.3 per 1000 population, but it seems that the prevalence has since fallen, at least of severe disease, as judged by a reduction in hospital admissions and deaths. This follows the introduction of antibiotic therapy for pulmonary infection, the control of tuberculosis, and effective vaccination for whooping cough and measles. However, the figures quoted may be an underestimate of the true prevalence of bronchiectasis: the diagnosis depends on demonstrating the cardinal feature of abnormal chronic dilatation of one or more bronchi, and it is likely that many people with chronic sputum production are mislabelled as 'bronchitic'. Indeed, recent CT scanning studies of patients with so-called 'chronic bronchitis' suggest that this is the case, and only the application of non-invasive imaging in large community surveys would tell us the true prevalence of the disorder. Bronchiectasis is regarded as a common problem in less developed countries where antibiotics are less readily available, socio-economic conditions are poor, and the prevalence of both tuberculosis and HIV are high.

Pathology

Macroscopic inspection of bronchiectatic lung reveals permanent dilatation of subsegmental airways that are inflamed, tortuous, and often partially or totally obstructed with secretions. The process also includes bronchioles, and at endstage there may be marked fibrosis of small airways. In allergic bronchopulmonary aspergillosis the changes are predominantly in proximal airways. Bronchiectasis caused by cystic fibrosis is likely to be more marked in the upper lobes. There is a spectrum of disease that ranges from cylindrical, where there is uniform dilatation, to saccular, where there may be gross terminal dilatation of the bronchi (saccules or cyst). An intermediate form is termed varicose bronchiectasis.

Microscopic features

The overall appearance is of chronic inflammation in the bronchial wall with inflammatory cells and mucus in the lumen. There is destruction of the elastin layer of the bronchial wall with a variable amount of fibrosis. Neutrophils are the main cell population in the bronchial lumen, whereas the commonest cells in the bronchial wall are mononuclear. The label follicular is applied when there is lymphoid follicle formation as part of extensive mural inflammation, which in subepithelial sites may cause finger projections blocking the bronchial lumen.

Aetiology and pathogenesis

There is a broad spectrum of causes and underlying conditions associated with bronchiectasis. These are summarized in Table 1.

The pathogenesis of bronchiectasis requires the combination of an infective insult with impaired clearance mechanisms that may result from local obstruction, impaired local structural defences, or defective immune defences. This is supported by work in animal models, which also show that the infection must be active, with damage to the airway wall occurring as a result of direct microbial insult or the secondary effects of the host inflammatory response.

The term 'vicious cycle' has been proposed to explain the development of bronchiectasis in a predisposed individual following a trigger insult, as shown in Fig. 1, which also demonstrates the key role played by neutrophil elastase. Neutrophils are recruited as part of the natural defences, but the inflammation is not self-limiting and in patients with bronchiectasis neutrophils persist in the airway secretions, with free neutrophil elastase activity usually present. Elastase, a neutrophil-derived serine proteinase, is known to inhibit ciliary beating, damage epithelia, act as a mucus secretagogue, and inhibit opsonophagocytosis via cleavage of immunoglobulins. All these actions contribute to persistence of bacteria in the respiratory tract and long-term tissue damage. Figure 1 illustrates that whenever a patient enters the pathway, for example in primary ciliary dyskinesia (which inhibits mucociliary clearance), or with immunoglobulin deficiency (which favours persistence of microbes in the bronchial tree), the vicious cycle becomes self-perpetuating with the final outcome of airway damage.

Fig. 1 The vicious cycle of infection and inflammation leading to progressive tissue damage in bronchiectasis.

Developmental defects

The congenital forms of bronchiectasis frequently show deficiency of elements of the bronchial wall that are necessary to prevent collapse and hence 'obstruction' of the airway. In Williams–Campbell syndrome there is deficiency of the bronchial cartilage. The Mounier–Kuhn syndrome or tracheobronchomegaly is the 'adult equivalent' of congenital deficiency of bronchial cartilage. Pulmonary sequestration predisposes to bronchiectasis because of decreased pulmonary clearance of the affected segment.

Mechanical obstruction

Bronchiectasis confined to a single lobe may be the result of local mechanical obstruction, either in the lumen (intrinsic), for example by a tumour or foreign body, or originating outside the lumen (extrinsic), for example by lymph node enlargement from tuberculosis or a tumour.

Disorders of mucociliary clearance

The disease cystic fibrosis provides the archetypal model of a genetic predisposition for the development of bronchiectasis. In this disorder (described in Chapter 17.10) there is dysfunction of the cystic fibrosis transmembrane regulator, a transmembrane chloride channel and ion transport regulatory protein. The resulting abnormal salt and water transport across respiratory epithelia predisposes to respiratory infection and the effects of the vicious cycle are clearly demonstrated as a structurally normal lung suffers progressive airway damage and the development of bronchiectasis.

Primary ciliary dyskinesia describes a group of inherited disorders in which mutation of several different genes may give rise to non-functional cilia. It is generally considered to be an autosomal recessive disorder with variable penetrance. The diagnosis is made by demonstrating abnormality of the cilia on electron microscopy. In the largest subgroup of this syndrome, and the form first described, the cilia lack dynein arms, which are the structures responsible for movement of cilia or spermatozoa. Subsequently it has been appreciated that a variety of components of the cilia are affected.

The intriguing observation that about 50 per cent of all subjects with immotile cilia syndrome have situs invervus is true for most subgroups, apart from those who have absent cilia or those whose main characteristic is lack of the two central microtubules. When ciliary dyskinesia is associated with abnormal situs the condition is labelled Kartagener's syndrome after the paediatrician who described four patients with the association of dextrocardia, sinusitis, and bronchiectasis in 1933.

Young's syndrome seems to represent an acquired defect of mucociliary clearance in which obstructive azoospermia is associated with sinusitis and bronchiectasis. The condition may occur after a man has successfully fathered a child, and may be associated with mercury poisoning from 'tooth powders' used in infancy (Pink's disease).

Secondary ciliary dyskinesia refers to the situation in which cilia are intrinsically normal but ciliary beating is reduced because of toxic damage from neutrophil or bacterial products. Tobacco smoke and other environmental pollutants have also been implicated in reducing ciliary beat frequency.

Postinfective bronchiectasis

The true incidence of postinfective bronchiectasis is difficult to establish because studies have been retrospective, relying on histories obtained 'second hand' from parents. The micro-organisms known to cause infection likely to progress to bronchiectasis are *Bordetella pertussis*, measles virus, adenoviruses, *Trypanosoma cruzi*, and *Mycobacterium tuberculosis*.

As mentioned above, the pattern of bronchiectasis has changed since the introduction of vaccinations and widespread availability of antibiotics in the developed world. A population with cylindrical bronchiectasis has superseded the gross saccular bronchiectasis associated with severe repeated childhood respiratory infections in the pre-antibiotic era. This may be associated with a childhood history of a chesty cough, with a long period of remission of symptoms through the teens and twenties, followed by the onset of symptoms of productive cough of purulent sputum and/or sinusitis in the third or fourth decade of life. Some of these patients may report having whooping cough or measles in childhood, but it is not appropriate to label them as postinfective unless symptoms have been persistent, without remission since childhood.

Immune deficiency

Immune deficiency is an important cause of bronchiectasis because treatment with intravenous immunoglobulin (where appropriate) will correct the defect and should prevent progression of disease.

Bronchiectasis presenting in childhood should trigger an extensive assessment of phagocytic and cellular immune defences. The rare disorder, X-linked hypogammaglobulinaemia, presents early in life and bronchiectasis is a frequent complication if untreated.

Adult-onset acquired panhypogammaglobulinaemia frequently presents with recurrent respiratory infection and is complicated by bronchiectasis if untreated. Selective deficiencies of IgG and IgM, and of IgG subclasses, are also treatable causes of bronchiectasis. Moreover, functional antibody deficiencies and failure to mount and maintain adequate responses to antigen challenge (for example pneumococcal vaccine) may be present despite normal total immunoglobulins, hence subtle humoral deficiency, a treatable cause of bronchiectasis, can be easily missed.

Immune defects may be secondary to malignancy or be related to treatment with immunosuppressive agents. Bronchiectasis is now a recognized complication of HIV disease.

Excessive immune response

Figure 1 illustrates the mechanism by which damage may occur as a result of the host response to chronic airway infection. Allergic bronchopulmonary aspergillosis is a condition in which the excessive reaction to a 'non-infecting' organism seems to be the major factor in producing the associated characteristic proximal upper lobe bronchiectasis. The appearance of obliterative bronchiolitis and subsequent bronchiectasis in lung transplant rejection further highlights the role of a damaging immune response in the development of bronchiectasis.

Toxic insult

In some patients there is a clear history of an inhalation accident or exposure to hot gases, for example in a fire victim. Aspiration of gastric contents is another

important cause of bronchiectasis insomuch as treatment to prevent aspiration will prevent further airway damage.

Associated conditions

The association of rheumatoid arthritis with bronchiectasis is well recognized. Treatment of bronchiectasis in patients with rheumatoid arthritis can be difficult, reflecting the need to achieve the right balance of immunosuppression, which helps the underlying inflammatory disease process but can simultaneously impair antimicrobial defences. The association of inflammatory bowel disease with bronchiectasis also highlights the issue of immunosuppression: some patients with both conditions report an improvement in chest symptoms when they take systemic corticosteroids for flares of inflammatory bowel symptoms.

Rhinosinusitis

About 80 per cent of patients with bronchiectasis have upper respiratory tract symptoms, with postnasal drip being the most common problem. Some 30 per cent of patients have chronic sinus sepsis, with fewer having recurrent ear infections. In ciliary dyskinesia, however, ear infections are almost invariably present.

Idiopathic bronchiectasis

The underlying cause of bronchiectasis remains unknown in 40 to 60 per cent of patients, even in specialist bronchiectasis clinics.

Clinical features

History

Bronchiectasis should be suspected when there is a history of persistent cough productive of mucopurulent or purulent sputum throughout the year. Patients have frequently been treated for recurrent chest infections and labelled as 'bronchitic', often despite the absence of a history of smoking.

Patients may produce mucoid sputum early in their disease, developing purulent sputum when they suffer an exacerbation associated with viral upper respiratory tract infection. Such exacerbations may be associated with pleuritic chest pain, haemoptysis, fever, and sometimes wheeze. Those presenting as adults often recall a 'chesty cough' or 'wheezy bronchitis' associated with upper respiratory tract infections in childhood, followed by complete resolution of symptoms in the teens and early adult life before these return after a viral trigger. Upper respiratory tract symptoms such as postnasal drip are common, and in about 30 per cent of cases there is a history of chronic sinusitis.

Patients with bronchiectasis also suffer from undue tiredness, which many find more troublesome than the productive cough.

Examination

Severe 'classic' cases of bronchiectasis seen in the pre-antibiotic era or in less developed countries are associated with obvious clinical signs including finger clubbing and widespread course crackles. Nowadays it is much more likely for clinical examination to be normal. The absence of clubbing or lung crackles does not exclude bronchiectasis.

Investigation

Radiological imaging

The chest radiograph may be normal in at least 50 per cent of patients with CT or bronchographic evidence of bronchiectasis. If it is abnormal the findings relate to thickened and dilated bronchi, which produce tram-line opacities and ring shadows. Retained mucus may be seen as tubular opacities, and there may be associated volume loss of the affected lobe.

The gold standard for the diagnosis of bronchiectasis is thin section, high-resolution CT (**HRCT**) of the chest, which has replaced the more invasive investigation of bronchography. The diagnostic criteria for bronchiectasis on HRCT depend on finding both dilatation and thickening of the affected bronchi, dilatation being present if the internal diameter of the bronchus is greater than the diameter of its accompanying pulmonary artery. The classic appearance of a cross-section of a thick-walled dilated bronchus next to the accompanying pulmonary artery is the 'signet ring' sign, as shown in Fig. 2. Bronchial dilatation is also recognized when airways are seen in longitudinal section on CT and there is a failure of tapering as the bronchus courses towards the periphery.

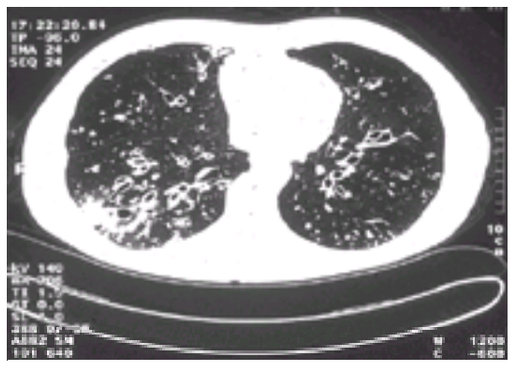

Fig. 2 CT scan of patient with bronchiectasis showing characteristic signet-ring sign.

There is a morphological spectrum of bronchiectasis, with cylindrical bronchiectasis forming one group, cystic or saccular bronchiectasis at the other end of the spectrum, and an intermediate group termed 'varicose' bronchiectasis. The CT appearances of this spectrum are well described. In cylindrical bronchiectasis there is uniform dilatation of the bronchi as they extend towards the periphery. By contrast, varicose bronchiectasis produces a beaded appearance, best shown when bronchi are imaged in the plane of the scan. Cystic bronchiectasis is recognized by rings representing markedly dilated bronchi, which may be clustered together and may contain air–fluid levels.

In addition to defining the extent of bronchiectasis, CT scanning may suggest the need for a bronchoscopy; for example if there is a bronchial obstruction, perhaps due to a foreign body or tumour.

Other tests

Once HRCT has proved the presence of bronchiectasis, investigations are directed towards defining the status of the disease and at attempting to establish an underlying cause. Table 2 highlights the minimum required to assess the current state of disease. Examination of a sputum specimen is crucial, it being important to document its character—mucoid or purulent—and to determine the colonizing organism, typical organisms being non-typeable *Haemophilus influenzae*, *Moraxella iatarrhalis*, *Streptococcus pneumoniae*, and *Pseudomonas aeruginosa. H. influenzae* is the most common (40 to 60 per cent). *P. aeruginosa* is usually associated with worsening symptoms and more severe lung disease. (See Chapter 17.3.3 for further information on the microbiological investigation of patients with lung disease.)

As the sputum microbiology may alter over time it is helpful to obtain repeated samples to ensure that an appropriate antibiotic management plan is in place. Measurement of inflammatory markers allows an assessment of the patient's current 'inflammatory burden'. Many have come to accept persistent purulent sputum over a period of time and may not complain of being particularly unwell. A raised erythrocyte sedimentation rate and/or C-reactive protein would weight the argument

in favour of early antibiotic intervention.

Table 3 outlines the investigations required to tie down a cause of bronchiectasis, some of which will then require specific treatment, for example immunodeficiency. Allergic bronchopulmonary aspergillosis is an important treatable cause: corticosteroid therapy produces substantial improvements in symptoms and well being, restores lung function, and prevents the development of further bronchiectasis. Similarly, the appreciation that chronic aspiration is the precipitant of lung damage can lead to appropriate therapeutic manoeuvres aimed at prevention of further damage.

Cystic fibrosis/bronchiectasis overlap

The diagnosis of cystic fibrosis should be considered in any patient with unexplained bronchiectasis, but particularly in the presence of upper lobe bronchiectasis, colonization with *Staphylococcus aureus* and *Pseudomonas aeruginosa*, or male infertility. A normal sweat test no longer excludes a diagnosis of cystic fibrosis as mutations occur which produce mild disease and a normal sweat test. If doubt exists, then the patient should be referred to specialist centre for further investigation (see Chapter 17.10).

Management

The principles of management of bronchiectasis are outlined in Table 4. The medical approach is two-pronged, with close attention given to treatment of any underlying cause whilst also treating the established bronchiectasis.

Sputum clearance

Since mucociliary clearance is reduced in bronchiectasis it seems sensible to aid sputum clearance by employing physiotherapy. Physiotherapy does not simply prevent mucus retention, but also allows a patient to expectorate sputum at a chosen convenient time rather than coughing throughout the day or night. There are no controlled trials to prove or disprove its usefulness in terms of disease modification or survival.

The use of mucolytics in bronchiectasis is controversial. The success of recombinant human DNAase in cystic fibrosis was not repeated in bronchiectasis that was not due to cystic fibrosis, when patients did not derive benefit in terms of lung function. A recent Cochrane review concluded that there is insufficient evidence to evaluate the routine use of mucolytics for bronchiectasis.

Antimicrobial therapy

There are two approaches to the use of antimicrobial therapy in bronchiectasis. The first involves the treatment of acute exacerbations. The second is based on the vicious cycle hypothesis, suggesting that chronic targeted antimicrobial therapy should reduce bacterial numbers, thereby reducing the host response and hence the potential for further lung damage. Whilst the latter approach has theoretical merits it has not been proved to be better than the former in randomized controlled trials.

In practice, the modern approach to antimicrobial treatment in bronchiectasis has been derived from regimens used in cystic fibrosis, which have yielded impressive improvements in survival (see Chapter 17.10). This depends on knowledge of a patient's colonizing organism, but there are some issues that apply regardless of the bacterial species. First, high doses of antibiotics are often required. These are necessary to penetrate scarred, thickened bronchial walls and the tenacious secretions that act as a physical barrier to antibiotic penetration to the microbes, which may also be harbouring drug-inactivating enzymes such as b-lactamases. Second, to avoid a high oral dose of an antibiotic that may result in unacceptable side-effects, the nebulized or parenteral route may be employed to achieve high levels of drug in the bronchial wall and secretions. Third, to determine the best treatment regimen for a patient it is worth assessing their initial response to an agent appropriate for the colonizing organism and then assessing the rapidity of return of purulent sputum. If purulent sputum becomes mucoid after a 14-day course of oral antibiotics and remains mucoid until the next viral trigger, then one is likely to recommend 'exacerbation only' treatment. However, if sputum returns to purulent within a few days of treatment finishing, then it is likely that chronic suppressive therapy will be required.

Figure 3 suggests an approach to the treatment patients with bronchiectasis depending on the characteristics of their sputum and the colonizing organism, but it must be pointed out that there has not been a systematic study of the benefits of this approach in the management of this condition.

Fig. 3 Guide to therapy for patients with bronchiectasis.

Bronchodilator therapy

Patients with bronchiectasis can have a restrictive or an obstructive picture. Some may have significant reversibility, and it is therefore worth assessing each individual for their response to b_2-agonists and anticholinergic agents.

Anti-inflammatory therapy

The vicious cycle hypothesis would suggest that the addition of anti-inflammatory therapy to antibiotics should be of benefit in patients with bronchiectasis. In cystic fibrosis, trials of oral corticosteroids have shown significant benefit in terms of lung function. Short-term trials of inhaled corticosteroids have been carried out in bronchiectasis, but evidence supporting long-term use (summarized in a Cochrane review) is limited and further trials are required. However, a trial of oral steroids is warranted whenever there is reversible airflow obstruction: if there is a documented improvement in lung function after a 2-week course, then introduction of inhaled steroids is justified.

Monitoring response to treatment

As each patient requires a tailored management plan it is critical that both the patient and physician agree defined criteria for assessing response. Measurement of lung function clearly produces an objective measure of response to corticosteroids, whereas the introduction of antibiotics may not alter lung function to a great degree but does improve sputum colour, volume, and consistency, and may also produce improvement in general well being. Studies have confirmed the validity of grading sputum colour as a marker of the microbial and inflammatory load in these patients, and diary cards documenting these parameters have proved helpful. This approach also facilitates patient education and self-management plans.

Surgery

Surgery represents the only 'curative' treatment for a select group of patients and should be carefully considered. Resection of bronchiectatic areas of lung was common in the pre-antibiotic era and provided a successful treatment at the time. Physicians' judgment regarding surgery may be unduly coloured by the bias of

patients returning to chest clinics with a recurrence of symptoms some years after resection, and the finding of bronchiectasis in other areas of the lung on CT scanning. These patients highlight the need for full assessment of the extent of bronchiectasis and a careful search for an underlying cause. If bronchiectasis is isolated to a single lobe and is the result of a localized obstruction, then surgery provides a cure and removes the need for lifelong treatment. However, surgery is unlikely to effect a cure if bronchiectasis is present in several lobes, and lobar resection is only indicated in two instances. First, if there is uncontrolled bleeding unresponsive to bronchial artery embolization. Second, if it is felt—after failure of aggressive antimicrobial therapy—that a particular lobe is acting as a 'sump' of infection which prevents good control of symptoms with medical therapy.

Lung transplantation

Lung transplantation provides an effective treatment for endstage bronchiectasis providing an underlying cause has been carefully assessed, treated, and is unlikely to jeopardize the transplanted organs. Patients with immunoglobulin deficiencies are not discounted from transplant assessment providing they are receiving adequate immunoglobulin replacement therapy. (See Chapter 17.16 for further discussion.)

Complications

Infective exacerbations are the most common complications to precipitate hospital admissions in patients with bronchiectasis. It is not common for patients to experience chest pain localized over an area of bronchiectasis, which may become pleuritic in nature during an infective exacerbation. Massive haemoptysis is rare nowadays: it is managed by embolization or, if that fails, by resection of the affected lobe. Minor haemoptysis is a common occurrence associated with infective exacerbations. Metastatic spread of infection rarely occurs in the developed world with good control of pulmonary infection with antibiotics. For similar reasons empyema is now very rare. Amyloidosis is often quoted as a classic complication of bronchiectasis, but is now extremely rare in the United Kingdom. Arthropathy is seen as a complication of bronchiectasis: this seems to flare in association with the chest disease, when active antimicrobial treatment will often result in remission of joint pain. Some patients may suffer vasculitic skin lesions in association with flares of bronchiectasis.

Prognosis

It was reported in 1940 that 70 per cent of 400 patients with bronchiectasis were dead before the age of 40. The situation is clearly different now: in the developed world we do not see the florid postinfective saccular type of bronchiectasis, but more commonly see patients presenting in their fourth and fifth decade of life with symptoms developing after a trigger illness and CT findings of cylindrical bronchiectasis. In 1981 a study following 116 patients for 14 years revealed that 20 per cent of patients treated medically and 17 per cent of surgically treated patients died at a mean age of 53 years. A Finnish study published in 1997 used the national hospital discharge register from 1982 to 1986 to identify newly diagnosed cases of bronchiectasis: 842 such patients were age and sex matched with individuals with chronic obstructive pulmonary disease or asthma discharged at the same time. Over a 10-year follow-up the prognosis for those with bronchiectasis was better than for those patients with chronic obstructive pulmonary disease, but poorer than that for patients with asthma. Bronchiectasis was the main cause of death in 13 per cent of the patients with this condition.

Further research

Further studies are required to identify the major factors that affect prognosis. For example, chronic colonization with *Pseudomonas* sp. may be a bad prognostic factor, but this may be negated by aggressive antimicrobial therapy. Study of homogeneous groups of patients (with respect to aetiology and colonizing organisms) should help assess various management regimens with regard to their effect on decline in lung function and survival. It is likely that a careful search for genetic factors that affect lung defences will yield new causes of bronchiectasis and allow the current so-called idiopathic group to be assigned a cause. We may then be able to define an at-risk population and aim to prevent development of bronchiectasis.

Further reading

Afzelius BA (1998). Immotile cilia syndrome: past, present, and prospects for the future. *Thorax* **53**, 894–7. [An overview of cilial disorders.]

Crockett AJ *et al*. (2000). Mucolytics for bronchiectasis. *Cochrane Database System Reviews (England)* (2) pCD001289. [Review of therapy for bronchiectasis.]

Jeffrey J, Swigris DO, Stoller JK (2000). A review of bronchiectasis. *Clinics in Pulmonary Medicine* **7**, 223–30. [This is a comprehensive review of bronchiectasis.]

Jones AP, Rowe BH (2000). Bronchopulmonary hygiene physical therapy for chronic obstructive pulmonary disease and bronchiectasis. *Cochrane Database System Reviews (England)* (2) pCD00045. [Review of therapy for bronchiectasis.]

Keistinen T *et al.* (1997). Bronchiectasis: an orphan disease with a poorly-understood prognosis. *European Respiratory Journal* **10**, 2784–7. [A recent paper describing the prognosis of bronchiectasis compared with asthma and chronic obstructive pulmonary disease.]

Kolbe J, Wells A, Ram FS (2000). Inhaled steroids for bronchiectasis. *Cochrane Database System Reviews (England)* (2) pCD000996. [Review of therapy for bronchiectasis.]

Pasteur MC *et al.* (2000). An investigation into causative factors in patients with bronchiectasis. *American Journal of Respiratory and Critical Care Medicine* **162**, 1277–84. [A paper covering large series of patients demonstrating the variety of aetiology of bronchiectasis.]

Smith IE, Flower CDR (1996). Review article: Imaging in bronchiectasis. *British Journal of Radiology* **69**, 589–93. [A description of HRCT scanning as the gold standard for diagnosis of bronchiectasis with a useful discussion of association between clinical features and radiological appearances.]

Stockley RA (1987). Bronchiectasis—new therapeutic approaches based on pathogenesis. *Clinics in Chest Medicine* **8**, 481–94. [A review of the approach to therapy based on the vicious cycle hypothesis.]

17.10 Cystic fibrosis

Duncan Geddes and Andy Bush

Definition

Cystic fibrosis is a recessively inherited disease caused by mutations in the cystic fibrosis gene located on the long arm of chromosome 7. The classical clinical picture is a combination of pancreatic insufficiency, suppurative lung disease, and high sweat sodium concentration, presenting in childhood and progressing to early death from respiratory failure. However, genetic analysis has identified many patients with less severe disease and cystic fibrosis mutations are also associated with male infertility and idiopathic pancreatitis, so expanding the clinical spectrum of cystic fibrosis lung disease. In general, carriers are healthy.

The genetic defect

The cystic fibrosis gene codes for a 168-kDa membrane protein named the cystic fibrosis transmembrane regulator protein (CFTR). CFTR is primarily an ATP responsive chloride channel but it also influences other cellular functions such as sodium transport across the respiratory epithelium, cell surface glycoprotein composition, and normal antibacterial defences. The protein is expressed in organs involved in cystic fibrosis disease—lungs, pancreas, sweat gland, etc.—but also in some places that do not seem to be affected clinically, such as the choroid plexus, heart, and renal tubules.

More than 900 disease-related mutations of the cystic fibrosis gene have been described. Their distribution in relation to the cystic fibrosis gene and its related protein are shown in Fig. 1. The mutations have been classified according to their impact at a cellular level—type 1: no protein; type 2: disordered trafficking; type 3: defective regulation; type 4: defective channel function; type 5: reduced protein synthesis. The understanding of these abnormalities is valuable as a basis for the design of new potentially corrective treatments, as illustrated in Fig. 2.

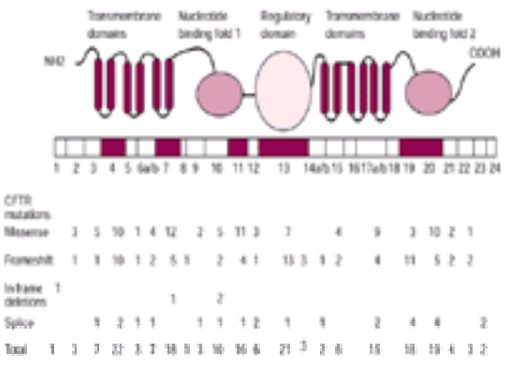

Fig. 1 Distribution of mutations within different regions of the CFTR gene. (Reproduced from Santis G. Basic molecular genetics. In: *Cystic fibrosis* (ed. Hodson ME,

Geddes DM), p. 25. Chapman and Hall, London, with permission.)

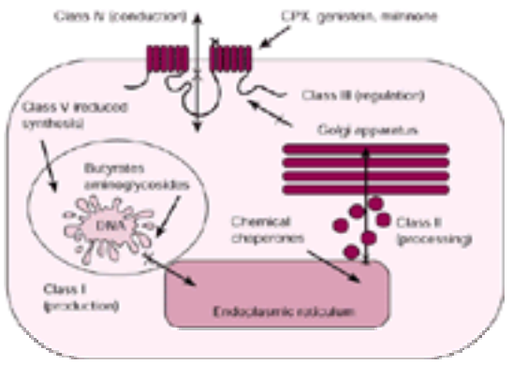

Fig. 2 Five categories of cystic fibrosis mutation with possible corrective treatments. (Reproduced from Rosenstein BJ, Zeitlin PL (1998). Seminar on cystic fibrosis. *Lancet*, **351**, 278, with permission.)

Most mutations are very rare; the commonest in European populations is †F508 which is found on 70 per cent of affected chromosomes. Most genetic laboratories restrict routine testing to the commonest six to eight mutations which account for over 90 per cent within a given population. Genotype–phenotype correlations have shown linkage of so called severe mutations, such as †F508, to pancreatic insufficiency and a tendency to more severe lung disease, while mild mutations go with pancreatic sufficiency and a tendency to less severe lung disease. However, in general, the correlation between genotype and the severity of lung disease is poor. All disease-associated mutations are linked with congenital absence of the vas deferens, resulting in male infertility, while rarer mutations are linked with isolated male infertility with no other evidence of cystic fibrosis disease.

Pathogenesis

Sweat duct

The primary secretion of the sweat duct is normal in volume and electrolyte concentration. However, as this secretion passes along the sweat duct mutant CFTR fails to absorb chloride ions, which therefore remain in the lumen and secondarily impair sodium absorption. The resultant sweat has high concentrations of both sodium and chloride which is useful for diagnosis but can lead to salt depletion in hot weather.

Pancreas

The synthesis and secretion of pancreatic enzymes in the acinus is normal, but disordered ion transport—primarily chloride and secondarily bicarbonate—results in relative dehydration of pancreatic secretions. This in turn leads to low flow and stagnation of secretions in the pancreatic ducts with subsequent autodigestion. The clinical consequences are that low volumes of bicarbonate-depleted pancreatic secretions reach the duodenum, with consequent malabsorption and progressive destruction of the pancreas with cyst formation. Although the islet cells are relatively unaffected at first, they too are progressively destroyed leading to insulin deficiency.

Biliary tract

Intrahepatic biliary secretions are probably normal in cystic fibrosis, but disordered electrolyte transport across the bile duct results in reduced water movement into the lumen. The bile is therefore concentrated, and the volume depleted, leading to plugging and chronic local damage. This eventually causes biliary cirrhosis and associated extrahepatic biliary stenoses. There are secondary changes in bile acids.

Gut

Gastric secretions have decreased volume with increased viscosity and sodium levels. The chloride transport defect similarly leads to altered fluid movement across large and small intestine. These changes are worsened by the addition of dehydrated biliary and pancreatic secretions, as well as by alterations in the osmotic load in the lumen secondary to pancreatic exocrine failure. The resulting deficiency of intraluminal water contributes to meconium ileus in neonates and the distal intestinal obstruction syndrome in adults.

Respiratory tract

The epithelium in the nose, paranasal sinuses, and intrapulmonary conducting airways is disordered in cystic fibrosis, while alveolar function is normal. Defective chloride transport is associated with increased sodium absorption from the lumen. This has two important consequences. Firstly, net surface electrical charge is altered from a normal of –20 mV to cystic fibrosis levels of –40 mV, which can be used for diagnosis. Secondly, increased sodium absorption takes water with it and dehydrates the airway surface liquid, reducing mucociliary clearance and favouring bacterial colonization. In addition, local antibacterial defences—including lactoferrin, lysozyme, and the cationic antibacterial peptides such as the b-defensins—may be impaired by local changes in salt concentration, and bacterial adherence to epithelial cells is increased by changes in cell surface glycoproteins. The net effect is to promote bacterial colonization and to reduce bacterial clearance, with subsequent inflammatory lung damage.

One consequence of bacterial colonization of the lower respiratory tract is an exuberant neutrophilic inflammatory response involving especially interleukin-8 (IL-8) and neutrophil elastase. The combination of elastase and other inflammatory mediators, while initially providing a useful antibacterial defence, is thought to contribute to lung damage and speed the progression of bronchiectasis and small airway narrowing.

Heterozygote advantage

The high frequency of the carrier rate in European populations (1:25) has led to a number of suggested advantages for the carrier, none of which are proven. These range from reduced susceptibility to infections such as cholera (reduced gut chloride secretion) and typhoid (reduced ingestion of bacteria by gut epithelium) to increased fertility among cystic fibrosis carriers.

Epidemiology

Genotype

The prevalence and distribution of the 900 disease-related mutations in the cystic fibrosis gene vary with ethnic origin. †F508 is commonest in northern European populations, accounting for 82 per cent of cystic fibrosis chromosomes in Denmark but only 32 per cent in Turkey. The W1282X mutation is common in Ashkanazi Jews (48 per cent of cystic fibrosis chromosomes) but rare in other populations. All disease-associated mutations are rare in African and almost unknown in Chinese populations.

Phenotype

Birth incidence varies with country of origin from 1:2000 to 1:100 000 as listed in Table 1. Prevalence figures are few and less reliable. In the third world, cystic

fibrosis is likely to be under-diagnosed since early childhood malnutrition, diarrhoea, and chest infections are so common. There are at least 6000 people in the United Kingdom and 30 000 in the United States with cystic fibrosis and numbers are increasing along with life expectancy.

Survival

From 1938 to 1960, most children with cystic fibrosis died before the age of 10. Since 1968, the first year mortality (chiefly from meconium ileus) has fallen from 18 per cent to 4 per cent and survival curves are linear thereafter, showing progressive improvement over succeeding decades. In 1986, the median survival was 25 years and in 1999 about 30 years. Cohort survival analysis shows continuing improvement and estimated survival for a child born with cystic fibrosis in the late 1990s is 40 to 50 years. Age specific mortality rates for females are a little worse then males, although this difference is narrowing and the world record for both sexes is now over 70 years.

Microbiology

People with cystic fibrosis have no detectable immune deficiency and, except for the respiratory tract, have no increased susceptibility to infection. The lungs show evidence of inflammation very early in childhood and thereafter become chronically colonized, characteristically by *Staphylococcus aureus* and *Haemophilus influenzae*, followed some years later by *Pseudomonas aeruginosa*. Many other organisms have been implicated, especially in advanced disease, including *Burkholderia cepacia* and *Stenotrophomonas maltophilia*. *Aspergillus fumigatus* is frequently isolated, but is associated with allergic rather than invasive disease, and atypical mycobateria are occasionally found. Viral, chlamydial, pneumococcal, and other respiratory infections are not more common or severe in cystic fibrosis, but the consequences of these infections may be more important in the damaged and permanently colonized cystic fibrosis lung. The microbiology of the nose and sinuses is the same as for the lung, but the clinical consequences are usually less important.

Staphylococcus aureus

This is the commonest colonizing organism in childhood, with a prevalence of over 50 per cent in children aged 0 to 9. The predeliction of *Staph. aureus* for cystic fibrosis lungs has been ascribed to high electrolyte content of airway surface liquid or enhanced retention in the airways. No phage type predominates and the organism usually remains sensitive to flucloxacillin in spite of prolonged antibiotic treatment. Resistance to tetracycline or erythromycin is relatively common but multiple antibiotic resistance is rare. The prevalence falls in adult life when *Pseudomonas aeruginosa* colonization predominates.

Pseudomonas aeruginosa

This is the commonest colonizing organism after the age of 10 years, with reported prevalence varying between 40 and 80 per cent. Enhanced adherence to cystic fibrosis airways promotes colonization but prior antibiotic treatment may play a part. No particular phage type predominates, but siblings with cystic fibrosis often carry the same type, and environmental sources have been identified in cystic fibrosis centres, dentistry equipment, hydrotherapy pools, and nebulizers. After some months or years of colonization, *Pseudomonas aeruginosa* produces mucoid alginate as a protective biofilm and the organisms live in mucoid microcolonies. This mucoid variant is associated with a worse prognosis and greater antibiotic resistance. Most colonizing *Pseudomonas aeruginosa* are sensitive to antibiotics at first but over the years and in associated with antibiotic treatment multiple resistance to most antibiotics (except colomycin) develops.

Haemophilus influenzae

Non-capsulated *H. influenzae* is a relatively frequent colonizing organism with prevalence of up to 30 per cent, although it may not be isolated due to over-growth of *Staph*. or *Pseudomonas*. Antibiotic resistance is seldom a problem.

Burkholderia cepacia

The overall prevalence of this organism is low, at 3 to 5 per cent, but it poses a particular problem due to occasional cross-infection with a virulent epidemic form that can cause rapid deterioration in patients previously only mildly affected. More usual is chronic asymptomatic carriage or progressive deterioration in the late stage of lung disease. Multiple antibiotic resistance is characteristic.

Diagnosis

Introduction

A consensus document on diagnostic criteria for cystic fibrosis has recently been published. The vast majority of patients with cystic fibrosis can be diagnosed by a sweat test (more than 98 per cent of 19 992 in the United States Cystic Fibrosis Foundation Registry). The occasional patient, particularly with a mutation giving rise to a mild or atypical clinical phenotype, may require more sophisticated testing. However, the major difficulty is usually not in confirming the diagnosis but in thinking of it in an appropriate context (below, and Table 2). Conversely, false positive diagnoses are not rare, and a new referral of a cystic fibrosis patient to the adult clinic should prompt a full review of the diagnosis.

Presenting features

The various age-related problems that can lead to a diagnosis of cystic fibrosis are given in Table 2. Paediatric presentations are relevant to adult life in that if an adult with atypical respiratory disease turns out to have a family history of a child with cystic fibrosis or an illness shown in Table 2, then cystic fibrosis should be considered in the adult. The new diagnosis of cystic fibrosis in a younger relative will also prompt cascade screening (below), usually aiming to discover carriers, but occasionally someone with a clinically mild cystic fibrosis phenotype is discovered. The United States Cystic Fibrosis Foundation Registry data show that as many as 10 per cent of cystic fibrosis patients are not diagnosed until adult life.

Less than 5 per cent pancreatic function is necessary for normal digestive function, and those presenting with cystic fibrosis in adult life are clinically pancreatic sufficient. The main presentation is with respiratory problems, usually recurrent lower respiratory infections with chronic sputum production. Some patients have prior a diagnosis of bronchiectasis, atypical asthma, nasal polyposis, or allergic bronchopulmonary aspergillosis. A new cystic fibrosis diagnosis has been described even in adults in their seventh decade. Depletion of sodium, chloride, and potassium due to excessive sweating, and secondary renal chloride retention, may result in presentation with dehydration and heat exhaustion in an otherwise apparently completely fit adult.

Another important mode of presentation is male infertility due to azoospermia because of congenital bilateral absence of the vas deferens (CABVD). CABVD exists in different forms: firstly, in association with congenital malformations of the upper urinary tract, in which case there is no increased incidence of cystic fibrosis mutations; secondly, as part of classical cystic fibrosis; and thirdly, as a truly isolated forme fruste of cystic fibrosis, with only a single cystic fibrosis mutation and ion transport abnormalities overlapping with, but different from, true cystic fibrosis. Portal hypertension secondary to macronodular cirrhosis in adult life may also be the first presentation of cystic fibrosis.

There is considerable debate as to the status of adults with single organ manifestations characteristic of, but not confined to, cystic fibrosis—for example pancreatitis or allergic bronchopulmonary aspergillosis. Some series report a higher than expected incidence of cystic fibrosis mutations, and the occasional unsuspected cystic fibrosis compound heterozygote. In practice, although cystic fibrosis should be excluded as far as possible by appropriate investigations in the patient with a possible single organ disease, most will not have the traditional clinical cystic fibrosis disease as it is currently defined.

Sweat testing

The test must be performed by someone who is experienced. Techniques include the classical pilocarpine iontopheresis of Gibson and Cooke, and more recently the macroduct collection. For the diagnosis to be established, tests should be performed in duplicate. The normal concentrations of sweat sodium and chloride increase with age. To diagnose cystic fibrosis in a child, the sweat chloride concentration should be greater than 60 mmol/l, and the sweat sodium concentration less than that of chloride. A sweat chloride of less than 40 mmol is normal in older children and adults, and intermediate concentrations are equivocal. However, there are undoubted cases of cystic fibrosis with normal sweat electrolytes, and the sweat test should always be interpreted in the light of the whole clinical picture. If the sweat

test is equivocal, consider repeating it the day after giving fludrocortisone 3 mg/m ^{2} for 2 days. In cystic fibrosis patients, sweat electrolyte concentrations fail to suppress into the normal range. There are a few rare conditions which also cause elevation in sweat electrolyte concentration, but these are rarely a serious diagnostic consideration in practice (Table 3).

Nasal electrical potential difference

The abnormal potential difference across mucosal surfaces can be measured by passing a soft catheter under the inferior turbinate, referencing it to an electrode placed on the abraded skin of the forearm. Normal values are –10 to –30 mV, the cystic fibrosis range –34 to –60 mV. The test is unreliable if the patient has an upper respiratory tract infection. The diagnosis can be further refined by perfusing the nose with solutions of amiloride to block sodium transport, and isoprenaline/low chloride to stimulate CFTR. Nasal potentials require extensive experience if results are to be accurate.

Ion transport can be measured directly from intestinal biopsies in an Ussing chamber, but this remains a research technique only.

Cystic fibrosis genotype

More than 900 different mutations causing cystic fibrosis have been reported. Testing for all is not currently practical. Thus DNA analysis can confirm the diagnosis if two mutations are found, but not exclude it. Linkage analysis can be used for antenatal diagnosis if a couple have already had an affected child, even if the actual mutations are not known.

Other investigations

In doubtful cases, evidence of subclinical organ dysfunction may be sought. Pancreatic dysfunction may be manifest by elevation in 3 day faecal fat excretion, low stool elastase, or abnormal results of pancreatic stimulation tests. CT scan of the chest or bronchoscopy may be used to discover minor bronchiectatic changes or infection with typical cystic fibrosis organisms. Azoospermia is strongly supportive of the diagnosis of cystic fibrosis. But note that it is important not to place too much diagnostic weight on clinically minor changes.

Conclusions

The diagnosis of cystic fibrosis is usually easy to confirm with a properly performed sweat test. There remain a few atypical cases which defy a firm diagnosis. In that event, clinical organ dysfunction should be treated appropriately, and the patient followed up very carefully: often time will clarify the diagnosis.

Screening

Introduction

Screening tests can be used to make an early diagnosis of cystic fibrosis in populations in order for early treatment to be instituted, and to detect cystic fibrosis carriers to allow antenatal diagnosis and the option of termination of affected pregnancies. In both areas there is controversy as to the indications and methods to be used. Currently neither is routinely available.

Methods

In the past, crude tests on meconium have been used, but these lacked accuracy and have been superseded by tests carried out on the routine heel prick blood sample collected from all babies in the first few days of life. These include immunoreactive trypsin, often combined with PCR for one or more common abnormal genes, or pancreatitis-related protein. If routine neonatal screening is to be instituted, it will probably be with immunoreactive trypsin and PCR. Pancreatitis-related protein screening may perform equally well, and obviates the need for genetic testing, which may be an advantage in some cultures.

Carrier screening is by PCR for several of the common cystic fibrosis genes on a blood or mouthwash sample. In principle, this sort of screening may be offered to relatives of known cystic fibrosis patients (cascade screening), by written invitation to the general population, or opportunistically at routine antenatal clinic visits or the GP surgery. It is generally considered that carrier testing at birth will not be useful because of the time lag between obtaining and utilizing the information.

Results

The evidence for the value of screening for the disease has come from a number of retrospective trials, all showing benefit, but with the disadvantage of using historical controls. There has been one prospective, randomized trial of neonatal screening from Wisconsin, United States in which 650 341 babies were screened. Of those in whom the diagnosis of cystic fibrosis was made, in 56 the diagnosis was communicated to the parents, and in 40 the diagnosis was suppressed until it emerged on clinical grounds. There were small but clear-cut nutritional benefits in the group in which the screening diagnosis was communicated, persisting to 10 years of age. The benefits were clearest early in life, at the time when growth is at its most rapid.

In general, carrier screening is poorly taken up when done by invitation, and at antenatal clinics it may be difficult to obtain a sample from the putative father. Cascade screening is generally better utilized, and should be offered at the time of making a new diagnosis.

Conclusions

Any screening test has false positives, which engender unnecessary anxiety, and false negatives, which may result in complacency. The balance of evidence is clearly in favour of neonatal screening so that early treatment can be given, and antenatal diagnosis offered for future pregnancies. The anxiety about false positives seems transient and deemed by the parents to be an acceptable price for subsequent reassurance. Carrier screening other than by cascade is more difficult, and, unless combined with wider public education, is unlikely to have a major impact.

Respiratory management

Introduction

Most of the morbidity and mortality of cystic fibrosis is due to respiratory disease. Much of the treatment is therefore devoted to preventing chronic infection and inflammation, which lead to bronchiectasis, progressive airflow obstruction, cor pulmonale, and ultimately death.

Typical physical findings are finger clubbing, cough with purulent sputum, together with crackles and occasional wheezes, chiefly in the upper lobes. Clinical scoring systems include the comprehensive Schwachman and simpler Taussig scores. The chest radiograph shows thickened bronchial walls and small areas of consolidation which start in the upper lobes and may progress to involve the whole lung (Fig. 3(a)). A variety of radiographic scoring systems have been proposed, for example Crispin–Norman or Brasfield score. Lung function tests show obstruction with relatively well preserved gas transfer. The forced expiratory volume in 1 s (FEV1) is conventionally used to assess the extent and progression of lung disease. Exercise tolerance and arterial blood gases are well maintained until there is extensive lung damage, when hypoxaemic respiratory failure supervenes. Carbon dioxide retention occurs late.

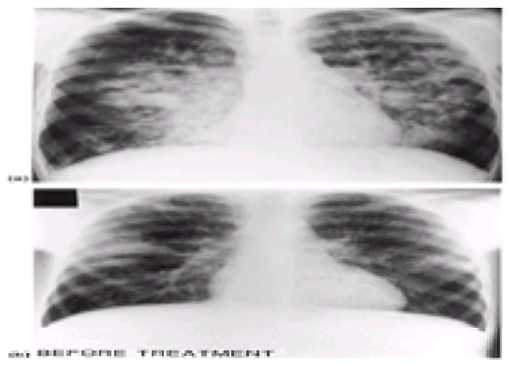

Fig. 3 (a) The typical chest radiograph appearances of advanced cystic fibrosis lung disease. There is also a right pneumothorax. (b) The chest radiograph in an adolescent child with cystic fibrosis complicated by allergic bronchopulmonary aspergillosis. Note the large wedge-shaped shadow.

Infection

Oral antibiotics

The use of prophylactic antistaphylococcal antibiotics is controversial: most would use continuous twice daily oral flucloxacillin if there is evidence of chronic colonization. Minor exacerbations of respiratory symptoms in the patient not colonized with *Pseudomonas* should be treated with a 1-month course of a high-dose antibiotic which will cover *Staph. aureus* and *Haemophilus influenza*.

Ciprofloxacin is used at the time of first isolation of *Pseudomonas aeruginosa*, combined with nebulized antibiotics (below) to try to prevent chronic colonization: the duration of therapy is controversial. Ciprofloxacin is also used to cover exacerbations of symptoms in the *Pseudomonas* colonized patient; however, ciprofloxacin resistance soon becomes common.

Nebulized antibiotics

Nebulized colomycin combined with oral ciprofloxacin is indicated at the time of first isolation of *Pseudomonas aeruginosa*. This approach has been shown in a randomized trial to delay chronic colonization. Once *Pseudomonas aeruginosa* colonization is established, randomized controlled trials have shown benefit from long-term nebulized antibiotics. In Europe, colomycin is the drug most often used. In the United States, nebulized tobramycin is preferred. No comparision of the two has been reported. Occasional patients bronchoconstrict with nebulized antibiotics: a test dose should therefore be given, and if necessary pretreatment with a bronchodilator prescribed.

Intravenous antibiotics

Infective exacerbations not responding to oral antibiotics, particularly those of *Pseudomonas aeruginosa*, are usually treated with a combination of an intravenous aminoglycoside and a semisynthetic antipseudomonal penicillin or cephalosporin. These are frequently given at home. Unlike the circumstance of the febrile patient with neutropenia, there is no evidence to support the use of single daily doses of aminoglycoside in the patient with cystic fibrosis. Drug metabolism in the cystic fibrosis patient is very different from normals and other patient groups. Some centres recommend 3-monthly courses of intravenous antibiotics, irrespective of symptoms, for all cystic fibrosis patients colonized with *Pseudomonas aeruginosa*. A randomized trial in the United Kingdom, although underpowered, did not support this approach.

Cross-infection issues

Fear of nosocomial acquisition of resistant organisms is widespread in the cystic fibrosis community. The apparent increase in prevalence of *Pseudomonas aeruginosa* in specialized clinics probably reflects more assiduous bacterial culture techniques. However, some centres advocate separate clinics for cystic fibrosis patients with and without *Pseudomonas aeruginosa*. Strict cohorting has also been advocated with regard to more resistant organisms such as *Burkholderia cepacia*, but it has subsequently been realized that not all of these organisms are of equal virulence, and cohorting has actually resulted in cystic fibrosis patients acquiring a virulent organism.

Sensible guidelines should be applied to all cystic fibrosis patients: these include diligent handwashing, no sharing of physiotherapy equipment, and the use of single cubicles for inpatients with difficult organisms. Communal physiotherapy and keep fit sessions should be discouraged, and there is no doubt that conferences for cystic fibrosis patients can result in transmission of infection. Careful microbiological surveillance is essential, and special measures may be needed if there is a true epidemic strain within a particular clinic.

Importance of viral infections

Viral infections, trivial in themselves, have been implicated in causing transient reduction in airway defences and an increased risk of *Pseudomonas aeruginosa* acquisition. Most physicians would at least give oral antibiotics (above) to cover viral exacerbations. Annual influenza immunization is advisable.

Atypical mycobacteria

These organisms are often harmless commensals: unlike *M. tuberculosis*, evidence of tissue invasion is generally held to be required to diagnose infection. This evidence cannot often be sought in cystic fibrosis, and decisions as to whether to treat are difficult. Evidence from autopsy studies suggests that atypical mycobacteria should they be treated only if they are repeatedly found in sputum.

Airway clearance

Chest physiotherapy should be performed twice daily as a routine, increasing at times of infective exacerbation. Different groups advocate different techniques (for example active cycle of breathing, autogenic drainage; and mechanical devices, such as the positive expiratory pressure (PEP) mask, flutter, and external oscillation jacket). There are no good comparisons between these approaches, and none has emerged as best. Physical exercise such as swimming supplements, but should not replace formal airway clearance sessions.

Reduction of mucus viscosity

Mucolytics have their advocates but, in general, are not useful. Human recombinant DNase *in vitro* reduces sputum viscosity. In the largest, randomized, controlled trial in the literature, once daily nebulized human recombinant DNase resulted in small but sustained improvement in lung function and reduction in infective exacerbations. Individual responses are very variable, and the treatment is expensive. A carefully monitored n=1 trial is recommended before long-term therapy.

Oxygen and other respiratory support

By analogy with the Medical Research Council and Nocturnal Oxygen Treatment Trial (NOTT) trials of oxygen in COPD, one would anticipate that long-term oxygen would be beneficial to the chronically hypoxic cystic fibrosis patient. The only trial of this approach was underpowered and thus inconclusive. Oxygen is usually prescribed for symptoms only. Nasal ventilation may be a useful short-term expedient while transplantation is awaited.

Bronchodilatation

Bronchial hyper-reactivity is common. Troublesome wheeze may need treatment with short-acting bronchodilators. However, b $_2$-agonists may cause paradoxical bronchoconstriction, and should be used cautiously. Long-acting b $_2$-agonists should only be given if there is clear-cut evidence of benefit. Persistent recurrent wheeze, particularly in the atopic cystic fibrosis patient, may be treated with inhaled or oral corticosteroids.

Aspergillus, including allergic bronchopulmonary aspergillosis

Evidence of exposure to *Aspergillus fumigatus* is common in cystic fibrosis (e.g. positive skin prick test, RAST, IgG precipitins, and sputum culture) but clinical disease is relatively rare. The prevalence of allergic bronchopulmonary aspergillosis is disputed, but is probably around 10 per cent. The major diagnostic criteria for this condition are also common features of cystic fibrosis. Sophisticated immunological testing has been used to try to refine the diagnosis, but an abrupt four-fold rise in total IgE, often in association with IgG precipitins to aspergillus, is the simplest and most reliable investigation. By contrast to typical infective exacerbations of cystic fibrosis, large fleeting radiographic shadows are typical (Fig. 3(b)). Treatment is with oral corticosteroids; the role of itraconazole is controversial.

Anti-inflammatory therapy

The pathogenesis of cystic fibrosis lung disease includes an exuberant IL-8 driven, neutrophil mediated, inflammatory response (see above), which, via the release of neutrophil elastase, may cause much of the tissue damage in the airways. This has lead to the seemingly paradoxical proposal that patients with chronic bronchopulmonary sepsis should be iatrogenically immunosuppressed. Various approaches have been tried, although none are in wide clinical use.

Oral corticosteroids

Usage in severe airway obstruction and allergic bronchopulmonary aspergillosis are discussed above. Long-term routine use was assessed in a multicentre, double-blind trial comparing prednisolone 2 mg/kg on alternate days, 1 mg/kg on alternate days, and placebo. This showed: (a) no benefit, except in patients colonized with *Ps. aeruginosa*; (b) sustained improvement in lung function in colonized patients; (c) unacceptable side-effects (growth failure, cataract, glucose intolerance), necessitating stopping the higher dose after 2 years and the lower dose after 4 years. Although in some patients, regular alternate-day steroids may be considered for up to 2 years, their routine use cannot be justified.

Inhaled corticosteroids

Since oral steroids are beneficial, but at the cost of unacceptable side-effects, it would seem logical to use long-term inhaled corticosteroids. Unfortunately there is no satisfactory trial confirming benefit: only small, relatively short-term studies have been done. Currently, inhaled corticosteroids can only be recommended for persistent wheeze, particularly in the atopic cystic fibrosis patient.

Ibuprofen

A multicentre, double-blind, placebo-controlled trial of ibuprofen showed a slowing of the rate of decline of lung function, particularly in young patients. However, ibuprofen is not widely used. This may be because: (a) not all age groups benefited; (b) there are theoretical reasons for believing that lower doses may actually be harmful, meaning that ibuprofen levels need to be measured and a high dose given; and (c) if intravenous aminoglycosides have to be administered for an acute exacerbation of chest disease, there is a significant risk of nephrotoxicity.

Other anti-inflammatory approaches

Although anti-inflammatory defences are normal in cystic fibrosis, they are overwhelmed by the burden of neutrophil elastase. Boosting the natural defences (a_1-antitrypsin, secretory leukoprotease inhibitor) by nebulizer has been the subject of small and inconclusive trials. Further safety and efficacy data are awaited.

Cytotoxics

There are anecdotal reports of the successful use of methotrexate, cyclosporin, and intravenous immunoglobulin in cystic fibrosis, particularly in those with severe, non-bronchiectatic airflow obstruction. There are no large trials of these approaches.

Haemoptysis

Blood streaking of sputum is common in cystic fibrosis and requires no special treatment. Massive haemoptysis is variously defined, usually as the expectoration of more than 250 ml of blood in 24 h, and is a frightening emergency which does require active management. It is usually a complication of quite severe lung disease, and the source is from hypertrophied bronchial arteries. The patient should be admitted, given antipseudomonal intravenous antibiotics, and any clotting abnormalities corrected. Careful chest physiotherapy should be continued. Trasylol and vasopressin are sometimes used to try to control haemorrhage. If bleeding does not settle, or recurs, then bronchial artery embolism should be considered. All sizeable bronchial arteries should be occluded. Preoperative bronchoscopy does not influence management, and often fails to define the side of bleeding in any case. The major risk of embolization is inadvertent occlusion of a major spinal artery, resulting in paraplegia. Lobectomy is rarely necessary, and carries a high risk in these patients, who are often very compromised.

Pneumothorax

This is usually a complication of late-stage lung disease. Shallow pneumothoraces require no special measures; more severe air leaks are initially treated with tube drainage. Careful physiotherapy must be continued, and intravenous antibiotics given. If there is a continued air leak, pleurodesis should be undertaken. However, it is important to consult with the local transplant service before doing this, because aggressive pleurectomy is seen by some to be a contraindication to subsequent transplantation.

Upper airway disease

Nasal polyps are seen in up to 50 per cent of adults with cystic fibrosis. Treatment is with nasal steroids in the first instance. If medical management fails, surgical polypectomy is indicated, but 50 per cent will require a second procedure within 2 years. Abnormal sinus radiographs are universal, but symptomatic sinusitis relatively rare. If present, sinusitis should be treated medically with prolonged antibiotics, nasal steroids, and possibly decongestants in the first instance; surgery is rarely needed. Rarely, surgery is needed for mucocele of the frontal sinuses.

Gastrointestinal management

Pancreatic insufficiency needs to be treated in 85 per cent of cases; meconium ileus or distal intestinal obstruction syndrome affects up to 30 per cent; symptomatic liver disease occurs in about 5 per cent, but in general the gastrointestinal manifestations of cystic fibrosis are less important than the lung disease. For a few patients, however, they are the dominant problem.

Pancreatic insufficiency

This is usually present from birth with low levels of bicarbonate and lipolytic and proteolytic enzymes in pancreatic secretions. Those with clinical pancreatic sufficiency secrete low but adequate levels of enzymes. Some develop pancreatic insufficiency later in life. The usual presentations are neonatal meconium ileus or failure to thrive with associated steatorrhoea and malnutrition. Consequences can include anaemia, vitamin deficiency, and occasionally oedema; complications include rectal prolapse, intussusception, volvulus, and distant intestinal obstruction.

The diagnosis is confirmed by estimation of stool elastase, demonstration of unsplit fat globules in the stool, or increased faecal fat on a 2 or 3-day stool collection. Formal testing of pancreatic function is seldom required.

Treatment with pancreatic enzyme and vitamin supplementation is usually straightforward and successful. Enteric coated enzyme preparations are taken before meals and the quantity adjusted to achieve normal stools. Most adults need four to eight capsules with main meals and two to four with snacks and learn to adjust the dose according to the fat content of the meal. The commonest cause of failure is poor compliance, although occasionally lactose intolerance, inflammatory bowel disease, coeliac disease, or bowel infection/infestation may coexist. A few patients need to take H_2 blockers, proton pump inhibitors, or antacids to achieve complete control of symptoms. Large bowel strictures have developed in some patients (usually children) taking high-strength enzyme preparations, probably as a toxic effect of the coating rather than the enzymes themselves.

Nutrition

Vitamin supplementation should be given to all patients to cover fat soluble vitamin deficiency. Multivitamin tablets contain vitamins A and D, but vitamin E needs to be given separately to maintain adequate intake. The diet should otherwise be normal, with a high calorie intake, usually 130 per cent of recommended daily allowance. Patients unable to maintain weight in spite of optimal dietary advice can be helped by enteral feeding, which is better tolerated by gastrostomy than by a nasogastric tube in the long term.

Distal intestine obstruction syndrome

Constipation and a loaded colon are relatively common in cystic fibrosis and usually respond to modification of the diet, pancreatic supplements, and a high fluid and roughage intake; occasionally lactulose or cisapride are helpful. Severe constipation merges into the distal intestine obstruction syndrome with pain, palpable faecal masses, and complete obstruction with faecal material in the distal ileum or ascending colon. The cause is multifactorial with imbalance of pancreatic enzymes and diet, disturbed fluid and electrolyte transport, faecal dehydration, and abnormal intestinal mobility all playing a part.

Patients present with chronic intermittent pain or episodes of complete obstruction. Although the differential diagnosis is wide and includes common conditions such as appendicitis, most patients improve with medical treatment and surgery should be avoided unless there is clear evidence of another diagnosis. Treatment with a balanced intestinal lavage solution, 500 to 1000 ml/h by nasogastric tube, usually moves the faecal blockage within 4 to 6 h. Alternatives are gastrografin by mouth or enema, or oral n-acetylcysteine. Occasionally, removal of inspissated faeces at colonoscopy is needed.

Other gastrointestinal complications

Pancreatitis is rare but should be excluded in cases of abdominal pain. It usually affects those who are clinically pancreatic sufficient. Treatment is conventional, with special attention to pulmonary infections, because the pain of pancreatitis may interfere with physiotherapy. Gastro-oesophageal reflux is common, sometimes with overt vomiting, and may be associated with coughing, physiotherapy, and bronchodilators which may relax the oesophageal sphincter. Aspiration of stomach contents is seldom a clinical problem. Although peptic ulcer disease might be expected in view of the low pancreatic bicarbonate secretion, there is only one report of an increased frequency of ulceration. *Helicobacter pylori* infection is uncommon, perhaps because of antibiotic treatment. Lactose intolerance, coeliac disease, and inflammatory bowel disease occur with the expected or slightly increased frequency in the cystic fibrosis population, but symptoms may be misattributed to cystic fibrosis and diagnosis therefore delayed. Both giardiasis and *Clostridium difficile* gut infection have been reported as being more frequent in cystic fibrosis but are not common clinical problems.

Liver disease

Liver disease causes problems in 5 per cent and death in 2 per cent of people with cystic fibrosis, but abnormal liver function tests are very common and up to 50 per cent have biliary cirrhosis demonstrable at post mortem. With increasing survival, liver disease may become more important.

Although liver enlargement and jaundice occasionally occur in early childhood, liver disease is usually signalled by hepatosplenomegaly or abnormal liver function on routine testing. Decompensation with jaundice, ascites, or encephalopathy are rare and occur late. Variceal bleeding only occurs in a minority of those with established chronic liver disease. Minor or modest elevations of aminotransferase, gamma glutamyl transpepddidase, or alkaline phosphatase levels are very common but do not correlate with established liver disease unless the enzyme levels are greater than four times normal. Routine ultrasound detects fatty change or multilobular cirrhosis: the finding of portal vein dilatation, splenomegaly, or colateral vessels indicating portal hypertension. Cholangiography is occasionally needed for treatment of gall stones: this may reveal irregularities of the intrahepatic ducts, suggesting chronic liver disease, and significant strictures of the common bile duct may also be seen. Liver biopsy is seldom needed.

No treatment has been shown to modify the course of chronic liver disease in cystic fibrosis, although clinical and biochemical improvements have been shown following treatment with ursodeoxycholic acid. This bile acid stimulates bile flow, may protect the hepatocyte from toxicity of bile acids, and is helpful in primary biliary cirrhosis. Many hepatologists therefore recommend its use in cystic fibrosis.

Jaundice must be investigated to exclude drug hepatotoxicity or treatable obstructive cause, but is otherwise a late event with poor prognosis. Variceal bleeding is treated with injection sclerotherapy or banding ligation, and in the short-term balloon tamponade or vasoconstrictor drugs may buy a little time. Surgical treatment is hazardous due to lung disease and in a few patients the insertion of a transjugular intrahepatic portal systemic shunt may be an alternative. Prophylactic treatment of varices has not been shown to help and may be detrimental. Ascites and encephalopathy are rare and are usually preterminal events to be managed conventionally.

In most cases of complicated chronic liver disease, management is made more difficult by the presence of lung infection that must be aggressively treated. Respiratory failure may develop concurrently. When this occurs, intubation and ventilation are seldom successful.

Diabetes

Glucose intolerance in cystic fibrosis increases with age, being rare under 10 years, affecting 14 per cent by 15 years, and over 65 per cent at 25 years, by which age 32 per cent are frankly diabetic. Even when glucose tolerance is normal, reduced insulin secretion is frequent. This is caused by gradual and progressive loss of beta cell mass in line with pancreatic fibrosis. Peripheral insulin sensitivity is normal and autoimmune factors are not involved.

Diagnosis is based on conventional WHO recommendations. Some recommend annual oral glucose tolerance tests, but screening for diabetes in a cystic fibrosis clinic is usually done by measurement of HBA1c, together with random or fasting blood sugar levels. Diabetes is usually diagnosed at such screening, but a few patients present with weight loss and increased frequency and severity of chest infections, although polyuria and polydipsia occasionally develop first. It has been suggested that the onset of diabetes is a marker of general deterioration, but many patients return to their previous level of health when diabetes is controlled. Oral hypoglycaemic agents provide control in a minority of patients for a limited time: insulin replacement is usually necessary. Control of blood sugar is relatively simple, with slow release preparations given twice daily. Ketoacidosis and insulin resistance are almost unknown.

The dietary management of diabetes in cystic fibrosis differs from that of other forms of diabetes: high dietary intake is maintained and insulin adjusted to fit the diet rather than the other way round. The usual recommendations are an energy intake of 150 per cent of normal with frequent balanced meals.

Early microangiopathy has been shown in cystic fibrosis patients with diabetes, but retinopathy, neuropathy, and nephropathy are very rare. This is due in part to the mildness of the diabetes and in part to short survival. Nevertheless, cystic fibrosis patients with diabetes tend to have excess morbidity and slightly increased rate of decline in weight and lung function.

Other organ systems

Reproduction

Almost all cystic fibrosis males have obstructive azoospermia with otherwise normal sexual function. This is due to absence of the vas deferens, and although there are no sperm in the ejaculate there is normal spermatogenesis and Leydig cell function. Counselling about infertility should be done by the time of puberty, ideally before permanent relationships develop. Most men opt to confirm the azoospermia by a sperm count. *In vitro* fertilization using aspirated sperm has been successful and there are now many cystic fibrosis fathers.

Early reports of reduced fertility in cystic fibrosis women have not been confirmed and most can conceive normally. The child must carry one mutation from the mother: the risk of cystic fibrosis is therefore 1 in 50 in a Caucasian populations with a carrier frequency of 1 in 25. Counselling and paternal genotyping allows reassurance for the majority of cystic fibrosis pregnancies and identifies a 1 in 2 risk when the father is a carrier. Successful pregnancies have been completed by many hundreds of cystic fibrosis women, but women with severe lung disease may not be able to complete a pregnancy safely, the risks rising with impaired lung function and especially when the 1 s FEV1 is less than 30 per cent predicted. Children born have been healthy, without an increased frequency of birth defects despite the mothers' extensive drug treatment. Lactation is normal.

Vaginal candidiasis secondary to antibiotic treatment is relatively common in cystic fibrosis, but otherwise there are no specific gynaecological problems. Sexual behaviour in both genders may be inhibited by low weight, delayed puberty, cough, sputum, haemoptysis, breathlessness, and indwelling catheters, but most people adapt well and persistent problems are few.

H4>Skin and joints

Clubbing is almost universal in those with significant lung disease and regresses after successful lung transplantation. Hypertropic osteoarthropathy is rare. Episodic arthritis, predominantly affecting the large joints, is quite common and is associated with chest infections. Erosive arthritis is rare. Pain responds to non-steroidal anti-inflammatory drugs and steroids or immunosuppression are seldom needed. Systemic vasculitis has occasionally been reported but is surprisingly rare considering the extent of immune activation, the frequency of circulating immune complexes, and the number of drugs taken.

Kidneys

Glomerulonephritis has been reported but is probably no more frequent than in the normal population. Drug-induced renal damage is rare and is usually associated with higher than recommended aminoglycoside levels. Very large numbers of aminoglycoside treatments appear to be safe when serum levels are well controlled. Renal stones are commoner in cystic fibrosis, probably due to excess oxylate absorption secondary to altered bowel bacterial flora. Systemic amyloidosis has occasionally been reported secondary to prolonged pulmonary infection.

Central nervous system

Ototoxicity occasionally results from aminoglycoside treatment but is not seen when serum levels are well controlled. Cerebral abscess rarely complicates lung sepsis. Vitamin E deficiency leads to a cerebellar syndrome combined with peripheral neuropathy.

Osteoporosis

Reduced bone mineral density is common in cystic fibrosis, with a prevalence among adults of up to 60 per cent. This is partly due to general malnutrition as well as vitamin D malabsorption, but relative immobility is sometimes a factor. An increased rate of fractures has been reported and rib fractures from coughing can interfere with adequate physiotherapy. Vertebral compression fractures are fortunately rare. Regular bone mineral density measurements are recommended with extra vitamin D and calcium supplementation when low. Bisphosphonates can cause bone pain and have not yet been fully evaluated.

Management of respiratory failure

Recurrent and persistent chest infection leads to progressive decline in lung function with eventual respiratory failure in the vast majority of patients; in about 2 per cent the liver fails first. At this stage palliation of symptoms should replace aggressive treatment unless there is a realistic prospect of a lung or liver transplant. If this is feasible, then preoperative work-up, counselling, surgical assessment, and placement on the waiting list should take place 2 years before the predicted date of death.

Lung transplantation (see also Chapter 17.16)

Selection criteria are listed in Table 4. The timing of assessment is judged on the level and rate of decline of lung function, arterial blood gases, and the frequency and severity of chest infections. Patients on the waiting list must be managed optimally to maintain lung function and nutrition, usually with gastrostomy feeding. Non-invasive ventilatory support can provide a bridge to transplantation for patients with progressive respiratory failure but intubation and conventional ventilation are not recommended. Donor organs are scarce and at least 50 per cent of listed cystic fibrosis patients never receive a transplant. The results for lung transplantation are the same as for other lung diseases with a survival of 70 per cent at 1 year and 50 per cent at 3 years.

Liver transplantation is appropriate for the occasional patient dying of liver failure with relatively good lung function: survival at 1 year is 40 per cent. For patients with respiratory failure and severe liver disease combined lung and liver transplantation is a possibility but with poor survival and with limited organ availability the operation is difficult to justify.

Terminal care

The timing of the decision to switch to palliative care is difficult and should be made in conjunction with the patient and relatives. The most distressing symptoms are cough, sputum retention, breathlessness, and exhaustion. Small doses of morphine are usually well tolerated and only seldom worsen respiratory failure.

The cystic fibrosis team

As with many chronic diseases, the purely medical care of cystic fibrosis is relatively straightforward. Proper holistic care requires a team approach, and without such a team, care will be second rate. Typically, the core of the team is formed by a specialist nurse, a physiotherapist, a dietician, a psychologist, and a social worker, together with a specialist doctor. It is unrealistic to expect every hospital to provide this, and so close contact with a tertiary centre is advisable. Many of the physical issues (airway clearance, nutritional management) have been discussed above. Equally important are many of the psychological problems springing from the presence of a chronic disease.

The normal tasks of adolescence include rebelling and breaking free of parental care. In those with cystic fibrosis this may never have been achieved, because the parents have wanted to keep control of treatment regimens, and have been reluctant to allow independence. Although the paediatric clinics should have established a pattern of the adolescent coming into the consulting room alone, frequently this does not happen, and the adult physician is confronted with parents who resent the idea that their now grown-up child should be seen on their own. Conversely, the consequences of a full-blown adolescent revolt (no treatment done, abuse of cigarettes, alcohol, and soft and hard drugs) may be particularly catastrophic in the cystic fibrosis patient. The authors know of no easy answer to adolescence and its aftermath.

Knowledge of fertility issues is notoriously poor amongst adult men with cystic fibrosis: these may need to be tackled tactfully. The issues surrounding pregnancy in the cystic fibrosis girl, who may herself be severely breathless, but desperately wishing for a child, also require sensitive handling. Further education and employment are also difficult issues in the setting of chronic physical disability, but skilled help may allow the cystic fibrosis patient to maximize their potential. A fuller account of the many and complex psychosocial issues surrounding care can be found elsewhere: appreciation of these issues is just as important as knowing the correct management of the physical problems of cystic fibrosis.

Future prospects

The growth in basic scientific understanding of cystic fibrosis will lead to a number of new treatments directed at the mutant CFTR gene or protein. These include gene therapy, which has already reached preliminary clinical trials, protein replacement therapy, and drug therapy to correct the molecular defect (as illustrated in Fig. 2). Research into correction of the disordered electrophysiology with sodium channel blockers, such as amiloride, or promoters of chloride transport, such as UTP, is already well advanced. There is, therefore, a real prospect of new fundamental treatments to prevent the development of cystic fibrosis disease and lead to improved health, prolonged survival, and reduction in life-long supportive therapy.

Further reading

Anguiano A, Oates RD, Amos JA, *et al.* (1992). Congenital bilateral absence of the vas deferens. A primary genital form of cystic fibrosis. *Journal of the American Medical Association* **267**, 1794–7. [Report setting out a new form of CF mutation associated disease.]

Armstrong DS, Grimwood K, Carlin JB, *et al*. (1997). Lower airway inflammation in infants and young children with cystic fibrosis. *American Journal of Respiratory and Critical Care Medicine* **156**, 1197–204. [The other side of the controversy as to whether infection is a necessary prerequisite for inflammation in CF.]

Cantin A (1995). Cystic fibrosis lung inflammation: early, sustained and severe. *American Journal of Respiratory and Critical Care Medicine* **151**, 939–41. [Brief review of lung inflammation.]

Caplen NJ, Geddes DM, Alton EWFW (1998). Gene therapy for respiratory disease. *Clinics in Pulmonary Medicine* **5**, 250–9. Overview relevant to the development of CF gene therapy.

Cleghorn GJ, Stringer DA, Forstner GG, Durie PR (1986). Treatment of distal intestinal obstruction in cystic fibrosis with balanced intestinal lavage solution. *Lancet* **1**, 8–11. [Establishment of modern management of obstruction in CF.]

Cohn JA, Friedman KJ, Noone PG, Knowles MR, Silverman LM, Jowell PS (1998). Relations between mutations of the cystic fibrosis gene and idiopathic pancreatitis. *New England Journal of Medicine* **339**, 653–8. [A paper illustrating the expanding spectrum of cystic fibrosis and the related diseases in which there may be a higher than normal prevalence of cystic fibrosis mutations.]

Cox KL, Ward RE, Furguiele TL, Cannon RA, Sanders KD, Kurland G (1987). Orthoptic liver transplantation in patients with cystic fibrosis. *Pediatrics* **80**, 571–4. [Key paper establishing the success of liver transplantation in CF.]

Cystic Fibrosis Foundation, Patient Registry 1996 (1997). *Annual data report*. Bethesda, Maryland. [Consensus group on diagnosis of atypical cases of cystic fibrosis in particular. Valuable source of epidemiological data from the United States.]

Cystic Fibrosis Genotype-Phenotype Consortium (1993). Correlation between genotype and phenotype in patients with cystic fibrosis. *New England Journal of Medicine* **329**, 1308–13.

Davidson TM, Murphy C, Mitchell M, Smith C, Light M (1995). Management of chronic sinusitis in cystic fibrosis. *Laryngoscope* **105**, 354–8. [Practical paper on clinical management.]

Davis PB, Drumm M, Konstan W (1996). Cystic fibrosis. *American Journal of Respiratory and Critical Care Medicine* **154**, 1229–56. [Comprehensive review article with emphasis on basic science and pathogenesis.]

Eigen H, Rosenstein BJ, Fitzsimmons S, *et al*. (1995). A multicenter study of alternate-day prednisone therapy in patients with cystic fibrosis. *Journal of Pediatrics* **126**, 515–23. [Disappointing full stop to the prednisolone story—an excellent study which failed to confirm previous results.]

Farrell PM, Kosorok MR, Laxova A, *et al*. (1997). Nutritional benefits of neonatal screening for cystic fibrosis. *New England Journal of Medicine* **337**, 963–9. [The only large randomized controlled trial of screening in CF.]

FitzSimmons SC (1993). The changing face epidemiology of cystic fibrosis. *Journal of Paediatrics* **122**, 1–9. [Survey of improvements from the United States registry.]

Frederiksen B, Koch C, Hoiby N (1997). Antibiotic treatment of initial colonization with *Pseudomonas aeruginosa* postpones chronic infection and prevents deterioration of pulmonary function in cystic fibrosis. *Pediatric Pulmonology* **23**, 330–5. [Review of the Danish clinic infection and antibiotic policies.]

Fuchs HJ, Borowitz DS, Christiansen DH, *et al*. (1994). Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. *New England Journal of Medicine* **331**, 637–42. [The largest randomized trial ever performed in CF establishing pulmonary function changes with DNase treatment.]

Hodson ME (1992). Vasculitis and arthropathy in cystic fibrosis. *Journal of the Royal Society of Medicine* **85** (Suppl. 19), 38–40. [Report of systemic manifestations of inflammation in CF.]

Hodson ME, Geddes DM, eds (1995). *Cystic fibrosis*. Chapman and Hall, London. [Comprehensive textbook with a clinical slant.]

Hodson ME, Madden BP, Steven MH, Tsang VT, Yacoub MH (1991). Non-invasive mechanical ventilation for cystic fibrosis patients: a potential bridge to transplantation. *European Respiratory Journal* **4**, 524–7. [A practical contribution to pretransplant care.]

Khan TZ, Wagener JS, Boat T, Martinez J, Accurso FJ, Riches DWH (1995). Early pulmonary inflammation in infants with cystic fibrosis. *American Journal of Respiratory and Critical Care Medicine* **151**, 1075–82. [Important study showing early onset of bronchial inflammation even in infants diagnosed by screening.]

Konstan MW, Byard PJ, Hoppel CL, Davis PB (1995). Effect of high-dose ibuprofen in patients with cystic fibrosis. *New England Journal of Medicine* **332**, 848–54. [Large multicentre study of non-steroidal anti-inflammatory medication in CF.]

Marchant JL, Warner JO, Bush A (1994). Rise in total IgE as an indicator of allergic broncho-pulmonary aspergillosis in cystic fibrosis. *Thorax* **49**, 1002–5. [References criteria for diagnosis of ABPA, and a simple laboratory test for diagnosis and following treatment.]

Middleton PG, Geddes DM, Alton EWFW (1994). Protocols for in vivo measurement of the ion transport defects in cystic fibrosis nasal epithelium. *European Respiratory Journal* **7**, 2050–6. [Methods for using nasal potentials in diagnosis which is also applicable to monitoring new treatments.]

Mukhopadhyay S, Singh M, Cater JI, Ogston S, Franklin M, Olver RE (1996). Nebulised antipseudomonal antibiotic therapy in cystic fibrosis: a meta-analysis of benefits and risks. *Thorax* **51**, 364–8. [Meta-analysis of randomised controlled clinical trials of nebulized antibiotics.]

Ramsey BW, Pepe MS, Quan JM, *et al*. (1999). Intermittent administration of inhaled tobramycin in patients with cystic fibrosis. *New England Journal of Medicine* **340**, 23–30. [Large, randomized trial of nebulized tobramycin in CF.]

Rosenstein BJ, Cutting GR, for the Cystic Fibrosis Foundation Consensus Panel (1998). The diagnosis of cystic fibrosis: a consensus statement. *Journal of Pediatrics* **132**, 589–95.

Rosenstein B, Zeitlin PL (1998). Cystic fibrosis. *Lancet* **351**, 277–82. [Brief but balanced review with selective list for further reading.]

Smyth RL, Van Velzen D, Smyth AR, Lloyd DA, Heaf DP (1994). Strictures of the ascending colon in cystic fibrosis and high strength pancreatic enzymes. *Lancet* **343**, 35–6. [Important report of a new side-effect.]

Tomashefski JF, Stern RC, Demko CA, Doershuk CF (1990). Non-tuberculous mycobacteria in cystic fibrosis: an autopsy study. *American Journal of Respiratory and Critical Care Medicine* **142**, 940–53. [Good autopsy study of atypical mycobacteria in CF.]

Tsang V, Hodson ME, Yacoub MH (1992). Lung transplantation for cystic fibrosis. *British Medical Bulletin* **48**, 949–71. [Details report of early and successful experience with lung transplantation for CF.]

Tsui LC (1995). The cystic fibrosis transmembrane conductance regulator gene. *American Journal of Respiratory and Critical Care Medicine* **151**, S47–53.

Valerius NH, Koch C, Hoiby NM (1991). Prevention of chronic *Pseudomonas aeruginosa* colonisation in cystic fibrosis by early treatment. *Lancet* **338**, 725–6. [Important trial of early and aggressive therapy to eradicate *Pseudomonas aeruginosa* and the results of treatment.]

Wallis C, Leung T, Cubitt D, Reynolds A (1997). Stool elastase as a diagnostic test for pancreatic function in children with cystic fibrosis. *Lancet* **350**, 1001. [Stool elastase has recently been described as a sensitive and specific test for pancreatic insufficiency, requiring only a spot sample; clinically extremely valuable.]

Weaver LT, Green MR, Nicholson K, *et al*. (1994). Prognosis in cystic fibrosis treated with continuous flucloxacillin from the neonatal period. *Archives of Disease in Childhood* **70**, 84–9. [Randomized trial of prophylactic flucloxacillin in CF.]

Webb AK, Govan J (1998). *Burkholderia cepacia*: another twist and a further threat. *Thorax* **53**, 333–4. [Important review of the biology of this increasingly important pathogen.]

Welsh MJ, Smith AE (1993). Molecular mechanisms of CFTR chloride channel dysfunction in cystic fibrosis. *Cell* **73**, 1251–4. [Summary of the molecular pathology of CF.]

17.11.1 Diffuse parenchymal lung disease: an introduction

R. M. du Bois

Definition

The definition of diffuse parenchymal lung disease has become confused. This is due to a combination of muddled nomenclature, the overuse of synonyms, and a lack of precision in defining the individual diseases that come under this 'umbrella' term.

Diffuse parenchymal lung disease used to be known as interstitial lung disease. The change of terminology recognized that it was not just the parenchyma but also the airspace components of the acini that are involved in the diffuse parenchymal lung diseases. Infective pneumonias and some malignancies involve the acinar regions of the lung but are excluded from the classification by convention, although they must be considered as part of a differential diagnosis when diffuse parenchymal lung diseases are being considered.

Each specific disease will be considered in subsequent chapters and this introduction will focus on the approach to the classification of the diffuse lung diseases and their diagnosis and management.

Classification

Diffuse parenchymal lung diseases can be subdivided into five major groupings:

- associated with systemic diseases including rheumatological disease;
- diseases caused by environmental triggers or drug-ingestion;
- granulomatous diseases;
- idiopathic interstitial pneumonias;
- other diffuse lung diseases.

The majority of the environmentally and drug-induced lung diseases and granulomatous lung diseases are of known cause. Diffuse parenchymal lung diseases occurring in the context of systemic disease have known associations but are generally of unknown cause. The majority of the heterogeneous group of 'other' diffuse parenchymal lung diseases is of unknown cause as are, by definition, the idiopathic interstitial pneumonias. Table 1 and Fig. 1 illustrate the diseases of known and unknown cause that fall within each of the above broad headings, and Table 2 illustrates disorders that present more acutely, an important distinguishing feature.

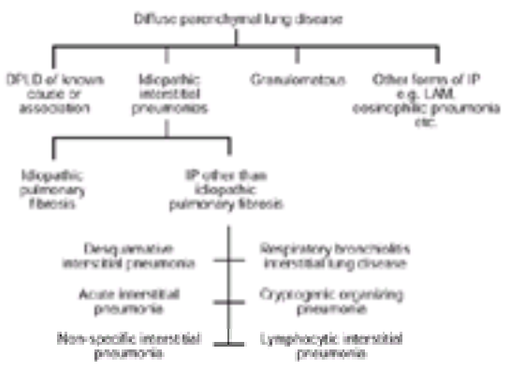

Fig. 1 Classification of diffuse parenchymal lung disease. DPLD, diffuse parenchymal lung disease; IP, interstitial pneumonia; LAM, lymphangioleiomyomatosis.

Idiopathic interstitial pneumonias

The group of diseases known together as the 'idiopathic interstitial pneumonias' are those that have produced most confusion in terms of nomenclature and understanding. This is largely because pathological pattern descriptions have been used interchangeably with disease 'labels' without consistency. Over the last five decades, two parallel processes—clinical and histopathological—were used to define diseases that are now included within the idiopathic interstitial pneumonias.

Clinical

In 1944, Hamman and Rich described four patients who died of a rapidly progressive process; the histopathological appearances showed interstitial pneumonia and fibrosis. Subsequently, similar disease patterns occurring over a more chronic time frame were identified. All of these were characterized by the presence of progressive breathlessness, crackles heard on auscultation of the chest, chest radiography which showed reticulonodular patterns of abnormality in the periphery and the bases of the lung fields, and a restrictive ventilatory defect on lung function testing. Similar but less acute disorders were subsequently described. Together all of these disorders have been loosely labelled as 'cryptogenic fibrosing alveolitis' or 'idiopathic pulmonary fibrosis' in the United States and elsewhere. Bronchoalveolar lavage was later introduced as an investigative tool and the presence of excess neutrophils and/or eosinophils helped to confirm this diagnosis. More recently, high resolution computed tomography has been used to define patterns of disease, helping to identify a number of quite distinct patterns that had previously been included under the single diagnostic 'label' of 'cryptogenic fibrosing alveolitis'.

Histopathological

In 1975, Liebow described five interstitial pneumonias that could be associated with clinical disease that mimics 'cryptogenic fibrosing alveolitis':

- usual interstitial pneumonia (UIP);
- desquamative interstitial pneumonia (DIP);
- bronchiolitis obliterans with usual interstitial pneumonia (BIP);
- lymphoid interstitial pneumonia;

- giant cell interstitial pneumonia.

Over subsequent years it became clear that not all of these interstitial pneumonias were idiopathic. Lymphocytic interstitial pneumonia was generally due to lymphoproliferative disorders, rheumatological disease, or AIDS-related disease. Giant cell interstitial pneumonia was found to be due to the exposure to the alloy hard metal (cobalt, tungsten carbide, titanium salts).

This resulted in a revision of the classification such that the interstitial pneumonias, defined histopathologically, of known cause, were removed. The idiopathic group now included UIP and DIP from the original classification. To these were added respiratory bronchiolitis–interstitial lung disease (RB-ILD), diffuse alveolar damage (DAD), and non-specific interstitial pneumonia (NSIP). BIP was renamed 'organizing pneumonia.'

Usual interstitial pneumonia

Usual interstitial pneumonia is by definition, the pattern seen in cryptogenic fibrosing alveolitis. The hallmark of usual interstitial pneumonia is a patchy distribution of interstitial fibrosis, chronic inflammatory cells, cystic airspaces (honeycomb lung), and areas of relatively normal lung. This pattern of pathology has a non-uniform appearance, best visible under low microscopic power, suggesting that the pathological process is at different stages of evolution throughout the biopsy (i.e. temporally heterogeneous), a feature that distinguishes usual interstitial pneumonia from other interstitial pneumonias. Another key feature is abundant fibroblastic foci.

Desquamative interstitial pneumonia

The characteristic feature of desquamative interstitial pneumonia is the diffuse accumulation of alveolar macrophages within the airspace in a rather monotonous, uniform pattern. There may be some interstitial inflammation and fibrosis but this is minor by comparison with the intra-alveolar inflammation. This pattern of disease is found almost exclusively in cigarette smokers.

Respiratory bronchiolitis–interstitial lung disease

The histopathological features of respiratory bronchiolitis–interstitial lung disease are very similar to those of desquamative interstitial pneumonia. They are differentiated by being less profuse, with pigmented macrophages accumulating in the airspaces around the bronchioles, which are themselves inflamed. The similarity of alveolar macrophage accumulation between desquamative interstitial pneumonia and respiratory bronchiolitis–interstitial lung disease (which also occurs almost exclusively in cigarette smokers) has led some to believe that these histopathological patterns are both part of the same spectrum of disease, but this view is not held unequivocally.

Diffuse alveolar damage

Diffuse alveolar damage is uncommon and the histopathological changes suggest an acute insult. This is the pattern of disease seen in both the adult respiratory distress syndrome and acute interstitial pneumonia (effectively, the adult respiratory distress syndrome of unknown cause). It is characterized by the presence of hyaline membranes lining damaged alveoli and, in more subacute states, the presence of buds of organization in the alveoli of those acini that have been damaged and are undergoing the healing process.

Non-specific interstitial pneumonia

This is arguably the least satisfactory histopathological entity and 'label'. Despite its name, there are specific features that define this histopathological pattern. There are varying degrees of interstitial inflammation and fibrosis within the interstitium but, unlike usual interstitial pneumonia, with which it is most likely to be confused (see Table 3), the appearances are uniform with none of the temporal heterogeneity of usual interstitial pneumonia and significantly less of the fibroblastic foci that are hallmarks of usual interstitial pneumonia.

American Thoracic Society and European Respiratory Society consensus statements on the nomenclature

Over the last few years these parallel clinical and histopathological approaches to the differentiation of the idiopathic interstitial pneumonia subgroup of the diffuse parenchymal lung diseases have become more integrated. The American Thoracic Society and European Respiratory Society have met to provide a consensus statement on the nomenclature of these diseases, which is summarized in Table 4. This classification attempts to provide a disease nomenclature that incorporates clinical, radiological, and histopathological patterns and to separate this from a pure histopathological classification that describes only the pattern of disease down the microscope without taking into account the clinical pathway that the disease has taken to reach that pathology. This is an important point because different insults could result in similar histopathological patterns. For example asbestos exposure can produce a usual interstitial pneumonia pattern of disease but, because it is of known cause, it is not included within the idiopathic interstitial pneumonias. Similarly, exposure to environmental agents to produce extrinsic allergic alveolitis can result in a non-specific interstitial pneumonia pattern, but this is clearly a distinct disease from that of a patient who presents with features similar to cryptogenic fibrosing alveolitis, who may have a similar histopathological pattern on lung biopsy.

The key issue that has emerged from the idiopathic interstitial pneumonias nomenclature debate, therefore, is that all clinical, radiological, and histopathological data must be taken into account before a diagnostic 'label' is applied to an individual patient. This new approach now allows a more precise definition of those diseases that are included within the idiopathic interstitial pneumonias.

Cryptogenic fibrosing alveolitis

The American Thoracic Society and European Respiratory Society consensus statement on this disease now recommends that cryptogenic fibrosing alveolitis (synonymous with idiopathic pulmonary fibrosis in the United States and elsewhere) must have the usual interstitial pneumonia pattern of pathology, or the appropriate clinical features, high resolution computed tomography pattern, and a bronchoalveolar lavage or transbronchial biopsy that excludes other disease (Table 5).

This is arguably the most contentious of the disease definitions in that there is a distinct subgroup of individuals who have all of the clinical, radiological, and bronchoalveolar lavage features of cryptogenic fibrosing alveolitis but who are found to have the non-specific interstitial pneumonia pattern of histopathology. At present it would be best to define these as the non-specific interstitial pneumonia variant of cryptogenic fibrosing alveolitis.

Desquamative interstitial pneumonia and respiratory bronchiolitis–interstitial lung disease

These diseases bear the same name as the histopathological classification because the idiopathic variants appear to be fairly uniform in their clinical and radiological features. Both are rare and it is possible that idiopathic variants may emerge.

Acute interstitial pneumonia

This disease resembles the adult respiratory distress syndrome in mode of onset, radiological and histopathological features, sharing the diffuse alveolar damage pattern of histopathology. Acute interstitial pneumonia is distinguished from adult respiratory distress syndrome by the absence of a known trigger. This nomenclature should not be used in the context of the rheumatological diseases in which a similar pattern of pathology can be seen.

Non-specific interstitial pneumonia

There are a number of clinical and radiological pathways that lead to a similar pattern of histopathology. However, once known causes and associated diseases have been excluded there appear at present to be no disease entities that have this histopathological variant other than the cryptogenic fibrosing alveolitis variant described above. Future studies are needed to validate this conclusion.

Cryptogenic organizing pneumonia

This disease was first defined in 1983 and was then redesignated 'bronchiolitis obliterans organizing pneumonia' in the United States in 1985. This nomenclature was so similar to 'bronchiolitis obliterans' that confusion ensued. Bronchiolitis obliterans organizing pneumonia is a pneumonic-like illness whereas bronchiolitis obliterans is an airway disease with distinct clinical, radiological, physiological, and histopathological appearances. It has therefore been decided to rename this disease cryptogenic organizing pneumonia to make the distinction between bronchiolitis obliterans quite clear and to define this disease as having the histopathological pattern of organizing pneumonia.

Lymphocytic interstitial pneumonia

This is included for completeness because this histopathological pattern can be seen in patients with rheumatological disease that can mimic the idiopathic interstitial pneumonias. It is likely that this will be included in the future within the non-specific interstitial pneumonia classification because the majority of diseases that have this pattern on biopsy are either lymphoproliferative or AIDS-associated.

Diagnostic approach

There are more than 200 entities included within the diffuse parenchymal lung disease group. A logical approach to diagnosis and management helps to avoid the major pitfalls that can be encountered in making a firm diagnosis. This approach can be considered in two phases:

- Phase 1
  - clinical history
  - clinical examination
  - chest radiography
  - pulmonary function tests
  - selective blood tests
- Phase 2
  - high resolution computed tomography
  - bronchoalveolar lavage
  - lung biopsy

This two-phase approach to diagnosis is summarized in Fig. 2.

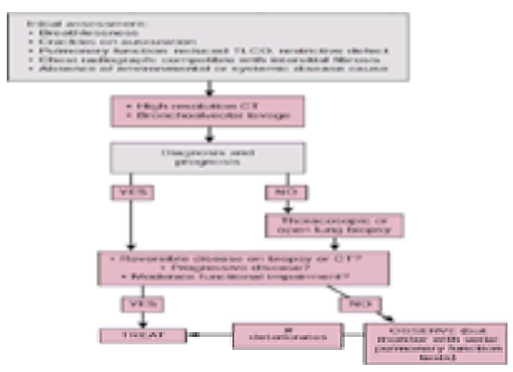

Fig. 2 Algorithm for the diagnosis, assessment, and management of suspected idiopathic interstitial pneumonia.

Clinical history

Most patients present with slowly progressive breathlessness, with or without a cough that is usually dry and non-productive. Duration and speed of onset are important: disease presenting acutely narrows down the differential diagnosis considerably (see Table 2).

The presence of wheeze is discriminatory as this implies that the diffuse lung disease process has an airway component that narrows the differential diagnosis. Examples of this include lymphangioleiomyomatosis or Langerhans cell histiocytosis. Other respiratory symptoms are uncommon and when present may also help to focus diagnosis. Pleurisy may occur in the rheumatological diseases and drug-induced disease, but never in cryptogenic fibrosing alveolitis or extrinsic allergic alveolitis. Haemoptysis may suggest alveolar haemorrhage that can be due to a variety of causes. A history of pneumothorax suggests peripheral lung cysts, which occur most commonly in Langerhans cell histiocytosis and lymphangioleiomyomatosis.

The past medical history can provide crucial information, particularly if there is a history of rheumatological disease, other systemic disease such as systemic vasculitis, or previous diseases that have required long-term drug therapy.

A complete occupational and domestic environmental history is necessary. The occupational history should include all occupations from school-leaving because many diseases caused by occupational environmental exposure occur after a lag period. The environmental dusts that are of relevance are those that are respirable, that is can penetrate the acinar regions of the lungs. The environmental conditions in which such dusts are found involve most commonly the sawing, grinding, drilling, or other working of materials that produce dusts of the appropriate size. A comprehensive occupational history therefore requires knowledge of the materials that are processed in particular job definitions and also of materials that can produce chronic lung inflammation. Organic dusts that provoke disease are most commonly those that occur in the context of fungal contamination of materials such as hay (as in farmers lung) or avian proteins found on the bloom and in the excreta of domestic birds such as budgerigars, pigeons, and hens (producing extrinsic allergic alveolitis). Table 1 includes diseases that need to be considered.

A history of foreign travel is also important because this may suggest the possibility of parasitic infection that can produce pulmonary eosinophilia.

A history of cigarette smoking will help to define predisposition to Langerhans cell histiocytosis, desquamative interstitial pneumonia, respiratory bronchiolitis–interstitial lung disease, and can provoke acute exacerbations of lung vasculitis as in Goodpasture's syndrome. By contrast, smoking appears to 'protect' patients from sarcoidosis and extrinsic allergic alveolitis.

A drug ingestion history also needs to be comprehensive. The list of drugs that can produce a diffuse parenchymal lung disease is increasing rapidly. A good source of drug-induced pulmonary adverse effects can be found on the web site listed at the end of this chapter.

Clinical examination

This is often unrewarding in the respiratory tract. Digital clubbing is common in cryptogenic fibrosing alveolitis but much less so in the other diffuse parenchymal lung diseases. Showers of fine end inspiratory crackles are typical of fibrosing alveolitis but more sporadic crackles can be found in many diffuse parenchymal lung diseases. Expiratory wheeze helps to define associated airway disease, which can be discriminatory.

More helpful is the examination of the rest of the systems to identify ocular disease (suggesting sarcoidosis or vasculitis), skin disease (that might support a diagnosis of sarcoidosis or rheumatological disease), musculoskeletal signs suggesting rheumatological disease, neurological disease (such as mononeuritis multiplex in

vasculitis), or a variety of central and peripheral neuropathies (in sarcoidosis).

Chest radiography

Chest radiography is one of the key components in the process of diagnosis of diffuse parenchymal lung disease. Five features should be noted:

- lung size;
- distribution of abnormalities;
- size and nature of nodular and/or reticular abnormalities;
- presence of confluent shadows;
- presence of pleural disease or lymphadenopathy.

Lung size

Patients with fibrosing lung diseases will generally have small lungs on chest radiography. Pitfalls include patients who have failed to take a full inspiration and patients with neuromuscular and other extrathoracic lesions that preclude full inspiration. This can mimic a fibrosing lung disease and needs to be considered if the small lungs appear to have normal parenchyma.

If other clinical features have suggested a diagnosis of fibrosing alveolitis, normal or large-size lungs indicate the coexistence of emphysema and fibrosing alveolitis: this admixture of disease processes is not uncommon as both are associated with cigarette smoking.

Other causes of large or normal sized lungs on chest radiography occurring together with nodular or reticular shadowing include Langerhans cell histiocytosis, lymphangioleiomyomatosis (a disorder involving smooth muscle proliferation that occurs only in women of childbearing age), tuberous sclerosis, of which lymphangioleiomyomatosis is believed by some to be a *forme fruste*. Chronic sarcoidosis can also feature large lungs, but this is usually associated with severe upper zone fibrosis causing a retraction of the hilar shadows towards the apices. Idiopathic bronchiectasis or cystic fibrosis can also be mistaken for diffuse parenchymal lung disease but the history of regular purulent sputum production will discriminate.

Distribution of abnormalities

The distribution of the abnormalities is always helpful. Fibrosing alveolitis occurring alone, in association with rheumatological diseases, or due to asbestos exposure produces reticular or reticulonodular abnormality predominantly in the basal zones, but also visible in the periphery of the lung, obscuring the diaphragm and the right and left heart borders. Predominantly upper zone disease, particularly involving loss of lung volume with an upward shift of the hilar shadows denoting upper zone fibrosis, occurs in chronic sarcoidosis and tuberculosis, extrinsic allergic alveolitis in its chronic stage, bronchopulmonary aspergillosis (almost always in the presence of asthma), and occasionally in Langerhans cell histiocytosis.

Predominant mid-zone abnormalities occur in sarcoidosis, extrinsic allergic alveolitis in its acute and chronic forms, and Langerhans cell histiocytosis.

Size and nature of nodular and reticular abnormalities

The size and shape of the abnormalities are helpful pointers to diagnosis. Very small 'granular' nodules less than 1 mm in size are seen in conditions such as idiopathic pulmonary hemosiderosis, miliary tuberculosis, and alveolar microlithiasis. Nodules up to 5 mm in diameter are seen in sarcoidosis, extrinsic allergic alveolitis, and silicosis. Shadows greater than 5 mm in size are present in Wegener's granulomatosis, rheumatoid arthritis, lymphoma, and other malignancies. Nodules of differing size and shape are highly suggestive of metastatic malignancy, and nodules that cavitate raise suspicion of Wegener's granulomatosis, other necrotizing granulomas, and rheumatoid nodules. Necrotizing squamous cell carcinomas and multiple staphylococcal abscesses need to be excluded. Lower zone reticulonodular shadowing is the pattern in fibrosing alveolitis of whatever cause.

Confluent shadowing

Confluent shadowing denotes airspace opacification. The presence of air bronchograms is very helpful in defining consolidation. All conditions that have predominant alveolar-filling histopathology will produce this pattern of radiographic change. These disorders include pulmonary alveolar proteinosis, diffuse alveolar haemorrhage, pulmonary eosinophilia (in which the confluent shadowing is often peripheral, as it frequently is in cryptogenic organizing pneumonia). Infection, particularly opportunistic infection in the immunosuppressed patient, alveolar cell carcinoma, and lymphoma must be differentiated, particularly as all of these can be complications of diffuse lung disease or its treatment.

Pleural disease and lymphadenopathy

The presence of pleural disease (with or without effusion) helps to exclude certain diagnoses such as extrinsic allergic alveolitis and cryptogenic fibrosing alveolitis. It is a common feature of the rheumatological diseases, particularly rheumatoid arthritis and systemic lupus erythematosus. It is also a feature of drug-induced lung disease, may occur in sarcoidosis, and is relatively common in Churg–Strauss granulomatosis and Wegener's granulomatosis.

Symmetrical hilar lymphadenopathy is usually due to sarcoidosis. Tuberculosis and lymphoma (and other malignancies) must always be considered, and these are more likely if the changes are unilateral. Lymphadenopathy is rarely observed on chest radiography in other diffuse lung diseases, with the notable exception of silicosis. Hilar calcification occurs in sarcoidosis and silicosis in addition to tuberculosis.

Radionuclide imaging

Gallium scanning has now fallen out of favour as a diagnostic tool for diffuse lung disease. Ventilation perfusion scanning as a means of identifying thromboembolic disease produces frequent false positives in the presence of diffuse lung disease and should not be used in this circumstance, spiral computed tomography being preferred. ^{99m}Tc-DTPA clearance from the lungs has been used in some centres as a tool for early detection of diffuse lung disease and as a predictor of outcome. In patients with fibrosing alveolitis occurring alone or in the context of rheumatological disease, sarcoidosis and other diffuse lung diseases, rapid clearance predicts disease that is more likely to be progressive.

Pulmonary function testing

In the majority of patients with diffuse parenchymal lung disease, lung function tests reveal a restrictive pattern of ventilatory defect with reduced gas transfer (DL_{CO}). Arterial oxygen tensions (PaO_2) may be normal or low and $PaCO_2$ may also be normal or low. In more subtle disease with normal gas transfer at rest, exercise tests can unmask abnormality: PaO_2falls and the alveolar–arterial oxygen gradient (A–a gradient) widens, indicating abnormalities of gas exchange usually due to ventilation–perfusion mismatch, but with an increased diffusion component on exercise. The anatomical dead space to tidal volume ratio (V_D/V_T) falls on exercise in the normal individual but remains the same or increases in restrictive lung disease.These investigations of pulmonary function can confirm the presence of disease but cannot discriminate between the different causes.

In disorders in which an airway component is associated with the diffuse parenchymal lung disease, a mixed obstructive–restrictive ventilatory defect is observed. This occurs in Langerhans' cell histiocytosis and more advanced sarcoidosis, both of which are bronchocentric disease processes. Evidence of subtle airway disease can sometimes be seen in less chronic sarcoidosis and also in early extrinsic allergic alveolitis. Combined pathologies may coexist; in patients with fibrosing alveolitis who have been cigarette smokers, a mixed obstructive–restrictive process is present due to a combination of emphysema and fibrosing alveolitis.

Blood tests

Routine haematology and biochemical tests are not of any discriminatory value in the diffuse lung diseases. Peripheral blood eosinophilia (above 1.5×10^{9}/l) is a

prerequisite for diagnosis of pulmonary eosinophilia. This is particularly helpful in the chronic forms of pulmonary eosinophilia, when total IgE levels are low, as this can discriminate between the allergic forms of disease that can affect the lung in which IgE levels match eosinophil levels.

Angiotensin converting enzyme concentrations are helpful in sarcoidosis but are not diagnostic. Their main value is in monitoring the burden of disease if found to be elevated at presentation. Routine immunoglobulin estimates are of no diagnostic value.

By contrast, autoantibody testing is important. The finding of a positive antinuclear antibody, with particular specific extractable nuclear antigen profiles, or of rheumatoid factor, may indicate that lung disease is the first manifestation of a systemic rheumatological condition. This form of presentation is reported increasingly and has important implications in terms of the precise diagnosis of the diffuse parenchymal lung disease and prognosis. Good examples of this include: the anti-DNA topoisomerase autoantibody, which is associated with fibrosing alveolitis in systemic sclerosis; the anticentromere antibody, which is associated with pulmonary vascular disease in systemic sclerosis; the anti-t-RNA synthetase autoantibodies, which occur when polymyositis is found in association with diffuse parenchymal lung disease; anti-Sm in systemic lupus erythematosus; SS-A and SS-B in Sjögrens syndrome; and the anti-RNP autoantibody in mixed connective tissue disease.

Antineutrophil cytoplasmic antibody (ANCA) is particularly helpful if the pattern is cytoplasmic, suggesting that the diffuse parenchymal lung disease is a manifestation of Wegener's granulomatosis or microscopic polyangiitis. The perinuclear (pANCA) pattern is much less discriminatory and is found in elevated titre in a wide variety of disorders, including rheumatological disease.

Bronchoalveolar lavage

When this technique was first employed, roughly 20 years ago, it was hoped that it could replace surgical biopsy in providing precise diagnostic information. This has proved not to be the case. It was also hoped that serial bronchoalveolar lavage might be a better monitoring tool than other indices of change. This has also been disproved. Nonetheless, bronchoalveolar lavage does still have a role in the diagnosis of diffuse lung diseases. It provides very helpful confirmatory evidence of diffuse lung disease in patients who cannot (for reasons of comorbidity or disease severity) or will not consent to a surgical biopsy following initial investigations. It is also helpful in excluding infection or malignancy.

The pattern of inflammatory cell infiltrate will differentiate the fibrosing lung conditions (characterized by neutrophils and/or eosinophils) from the granulomatous or drug-induced lug diseases (characterized by an excess of lymphocytes with or without granulocytes). Furthermore, the patterns of bronchoalveolar lavage are beginning to become helpful in differentiating the different idiopathic interstitial pneumonias: granulocytes in cryptogenic fibrosing alveolitis; granulocytes and lymphocytes together with 'smoker's inclusions' in macrophages in desquamative interstitial pneumonia; neutrophils, often in very high numbers, in acute interstitial pneumonia; granulocytes and lymphocytes in non-specific interstitial pneumonia and organizing pneumonia; and lymphocytes in lymphocytic interstitial pneumonia.

Bronchoalveolar lavage can provide diagnostic material in some of the rarer lung disorders, such as: pulmonary alveolar proteinosis (milky effluent; PAS-positive material; phospholipid, membrane-like structures under electronmicroscopy; biochemistry); Langerhans cell histiocytosis (increased numbers of Langerhans cells identified by CD1a staining); mineral dust exposure (energy dispersive analysis by X-rays, asbestos bodies); iron-laden macrophages (alveolar haemorrhage). Other very helpful appearances include the bizarre multinuclear giant cells obtained from patients exposed to the alloy hard metal and the proliferative response of T-cells obtained from the lungs of patients with beryllium exposure, confirming a diagnosis of chronic berylliosis.

High resolution computed tomography

High resolution computed tomography provides a three dimensional anatomical reconstruction of the whole of both lungs. This provides a number of significant advantages over plain chest radiography in the diagnosis and discrimination of the different parenchymal lung diseases, and some patterns are pathognomonic (Table 6).

The major advances that computed tomography has provided is in the earlier detection of suspected lung disease, when chest radiography is often normal; the differentiation of one diffuse lung disease from another; an estimate of the likely reversibility of the disease process—in general a 'ground glass' appearance favours reversibility in response to therapy, or sometimes spontaneously, whereas a coarse reticular pattern of abnormality is generally irreversible and indicates fixed fibrosis. Other irreversible features include the thin cystic structures seen in Langerhans cell histiocytosis and lymphangioleiomyomatosis. Less definitive patterns include variably sized nodules of varying density, linear opacification, and more subtle reticular change. Thickening of the interlobular septa or around the bronchovascular bundles imply a lymphatic component of the disease process. This degree of anatomic precision is the reason why high resolution computed tomography increases sensitivity, specificity, predictive value and, most importantly, confidence of diagnosis in the diffuse lung diseases. An assessment of the degree of reversibility also allows a better estimate of prognosis.

In patchy disease, the location of the best site for a surgical biopsy can be identified using high resolution computed tomography and more subtle pleural disease and airway disease can be identified. These additional features, in the context of the features of diffuse parenchymal lung disease, also help in the differential diagnosis. Good examples include the combination of airway and diffuse disease in rheumatological disorders such as rheumatoid arthritis and the subtle combination of patchy parenchymal disease together with pleural tags in drug-induced disease.

Computed tomography provides a better correlation with functional abnormalities than any other index, including surgical biopsy. In this regard, the extent of disease on computed tomography best matches measures of gas exchange and survival. The availability of high resolution imaging has therefore reduced the need for surgical biopsy in patients with diffuse lung disease considerably, but not removed it entirely.

Lung biopsy

It is imperative that a precise diagnosis is made from the large number of potential causes of diffuse parenchymal lung disease. The advantage of having a firm diagnosis is that the appropriate treatment can be instituted, treatment changed if that which has been started is inappropriate, a balance of cost–benefit to the patient can be better established, and, most importantly, the patient can be better informed about outcome, both with and without treatment. The oft used empirical approach of a trial of treatment—and if that fails then biopsy—is flawed. The major flaws being firstly that treatment will modify the disease process, making diagnosis more difficult with a subsequent biopsy, and secondly, the patient may have deteriorated during this waiting period, making biopsy more risky, and side-effects consequent upon treatment may also complicate the process. It is therefore recommended that, if there is any doubt about diagnosis after using all investigative tools other than biopsy, biopsy should be advised.

There are two types of biopsy. Transbronchial biopsy is used where the disease process is centred around the small airways, allowing access to the transbronchial approach, and when the diagnosis can be made on small biopsies, meaning that the pathological appearances must be so characteristic on small biopsies that there is no doubt. The best examples of this are the granulomatous diseases such as sarcoidosis or malignant diseases such as lymphangitis carcinomatosa.

For the more difficult diffuse lung diseases, particularly the idiopathic interstitial pneumonias, a surgical biopsy is necessary so that the overall pattern of disease can be appreciated. This does require a larger sample. The surgeon is also able to take samples from more than one site, thereby increasing the likelihood of obtaining representative tissue. Two approaches have been used: the limited thoracotomy approach or the more recently introduced video-assisted thoracoscopic surgical technique. The latter is a less invasive approach that is now preferred, providing equivalent sized samples to the more open technique but with some evidence of lessened morbidity. For choice of biopsy procedure in particular diseases see Table 7.

Further reading

Classification

Bjoraker JA, Ryu JH, Edwin MK, *et al* (1998). Prognostic significance of histopathologic subsets in idiopathic pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine* **157**, 199–203.

Bouros D, Nicholson AC, Polychronopoulos V, du Bois RM (2000). Acute interstitial pneumonia. *European Respiratory Journal* **15**, 412–8.

Daniil ZD, Gilchrist FC, Nicholson AG, *et al.* (1999). A histologic pattern of nonspecific interstitial pneumonia is associated with a better prognosis than usual interstitial pneumonia in patients with

cryptogenic fibrosing alveolitis. *American Journal of Respiratory and Critical Care Medicine* **160**, 899–905.

Diffuse Parenchymal Lung Disease Group (1999). The diagnosis, assessment and treatment of diffuse parenchymal lung disease in adults. *Thorax* **54** (Suppl. 1).

Katzenstein AL, Myers JL (1998). Idiopathic pulmonary fibrosis: clinical relevance of pathologic classification. [Review] [66 refs]. *American Journal of Respiratory and Critical Care Medicine* **157**, 1301–15.

Liebow AA (1975). Definition and classification of interstitial pneumonias in human pathology. In: Basset F, Georges R, eds. *Progress in respiration research*, pp. 1–33. Karger, New York.

Joint American Thoracic Society and European Respiratory Society Group (2000). Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. *American Journal of Respiratory and Critical Care Medicine* **161**, 646–64.

Drug-induced disease

http://www.pneumotox.com/

Imaging

Hansell DM. High resolution computed tomography and diffuse lung disease (1999). *Royal College of Physicians of London* **33**, 525–31.

Mathieson JR, Mayo JR, Staples CA, Muller NL (1989). Chronic diffuse infiltrative lung disease: comparison of diagnostic accuracy of CT and chest radiography. *Radiology* **171**, 111–6.

Muller NL, Colby TV (1997). Idiopathic interstitial pneumonias: high-resolution CT and histologic findings. *Radiographics* **17**, 1016–22.

Wells AU, Rubens MB, du Bois RM, Hansell DM (1997). Functional impairment in fibrosing alveolitis: relationship to reversible disease on thin section computed tomography. *European Respiratory Journal* **10**, 280–5.

Bronchoalveolar lavage

BAL Co-operative Group Steering Committee (1990). Bronchoalveolar lavage constituents in healthy individuals, idiopathic pulmonary fibrosis, and selected comparison groups. *American Review of Respiratory Diseases* **141**, S169–S202.

Drent M, Mulder PG, Wagenaar SS, Hoogsteden HC, van Velzen-Blad H, van den Bosch JM (1993). Differences in BAL fluid variables in interstitial lung diseases evaluated by discriminant analysis. *European Respiratory Journal* **6**, 803–10.

17.11.2 Cryptogenic fibrosing alveolitis

R. M. du Bois

Introduction

The first description of what we currently recognize as fibrosing alveolitis was in 1907. Since then nomenclature has been confused; multiple synonyms have been used, including Hamman–Rich syndrome. With the advent of a new histopathological classification and the recognition of particular patterns of disease on high resolution computed tomography, it is now possible to define cryptogenic fibrosing alveolitis very specifically

Definition

The International Consensus Statement on cryptogenic fibrosing alveolitis (idiopathic pulmonary fibrosis in the United States) has defined the disease as a specific form of chronic fibrosing interstitial pneumonia, requiring, in the presence of a surgical biopsy showing the usual interstitial pneumonia pattern of pathology (see Chapter 17.11.1):

- the exclusion of other known causes of interstitial lung disease such as drug toxicities, environmental exposures, and rheumatological disease;
- abnormal pulmonary function studies that include evidence of restriction (reduced vital capacity (VC) often with an increased FEV_1/FVC ratio) and/or impaired gas exchange (increased alveolar–arterial oxygen gradient ($P(A–a)O_2$) at rest or on exercise or decreased carbon monoxide transfer factor ($DLCO$));
- typical features on chest radiography or high resolution computed tomography scans.

In the absence of a surgical lung biopsy, the diagnosis of cryptogenic fibrosing alveolitis is less certain. However, in the immunocompetent adult, the presence of all of the following major diagnostic criteria as well as at least three of the four minor criteria increases the likelihood of a correct clinical diagnosis of cryptogenic fibrosing alveolitis:

- Major criteria
  - exclusion of other known causes of diffuse lung disease such as certain drug toxicities, environmental exposures, and rheumatological diseases;
  - abnormal pulmonary function studies that include evidence of restriction (reduced VC often with an increased FEV_1/FVC ratio) and impaired gas exchange (increased $P(A–a)O_2$at rest or on exercise or decreased $DLCO$);
  - bibasilar reticular abnormalities with honeycombing and minimal or no ground glass opacities on high resolution computed tomography scans;
  - transbronchial lung biopsy or bronchoalveolar lavage showing no features to support an alternate diagnosis, such as granulomas on biopsy or an excess of lymphocytes on bronchoalveolar lavage.
- Minor criteria
  - age more than 50 years;
  - insidious onset of otherwise unexplained dyspnoea on exertion;
  - duration of illness more than 3 to 6 months;
  - bibasilar, inspiratory crackles on chest auscultation.

Aetiology

Epidemiology

Cryptogenic fibrosing alveolitis may occur in any decade of life but is most commonly seen between the ages 50 to 60 years; children do not get the disease, although they can develop a diffuse lung disease that mimics the condition but does not have the usual interstitial pneumonia pathology. Cryptogenic fibrosing alveolitis occurs in males slightly more frequently than females. There is no geographic variation. Prevalence and incidence rates (approximately 5 per 100 000) are only estimates in the United Kingdom and elsewhere, but are increasing based on evidence from mortality statistics for England and Wales. A registry-based study from the United States estimated a prevalence of 20.2 cases per 100 000 for males and 13.2 cases per 100 000 for females and an incidence of 10.7 cases per 100 000 per year for males and 7.4 cases per 100 000 per year for females.

Possible trigger factors

By definition there is no known aetiology. Multiple factors are likely to be involved. The disease is more common in patients who have a history of cigarette smoking. Viruses, particularly Epstein–Barr virus, have been implicated as trigger factors but this is not proven. More recently, an increased occupational exposure to metal dusts, wood fires, and antidepressant medication has been associated with excess cryptogenic fibrosing alveolitis by comparison with control populations.

In most patients, there is no family history of fibrosing alveolitis but, in a small subgroup a familial pattern is observed. The inheritance pattern is unpredictable and it is likely therefore that transmission involves variable penetrance. The clinical features of the familial and sporadic variants are identical. There are, as yet, no known

predisposing immunogenetic factors, unlike other diffuse lung diseases that can produce lung fibrosis such as the Hermansky–Pudlak syndrome (a condition characterized by oculocutaneous albinism and abnormal platelets), the diffuse lung disease of systemic sclerosis (associated with the anti-DNA topoisomerase I autoantibody, in turn highly associated with an excess of the HLA DRB1*II and DPB1*1301 class II MHC alleles), and sarcoidosis.

Pathogenesis

The pathogenetic processes that give rise to cryptogenic fibrosing alveolitis involve lung injury, an immunological and inflammatory response, and fibrogenesis. All four components appear to be occurring in parallel and this is reflected in the heterogeneity of the histopathological appearances. The various trigger agents implicated in the disease are likely to be the causes of injury, complemented by an adverse imbalance in the oxidant–antioxidant profile within the lung, together with the release of tissue-damaging enzymes from macrophages and granulocytes.

The inflammatory response is variable and its relationship to fibrogenesis is unclear. In brief, all of the cellular and molecular biological mechanisms that determine immune and inflammatory cell recruitment and activation have been shown to operate in the lungs of patients with cryptogenic fibrosing alveolitis. Key components appear to be the up-regulation of tumour necrosis factor-a, chemokines, notably interleukin 8, and growth factors, most notably transforming growth factor-b and connective tissue growth factor. Whether up-regulation of the mechanisms that result in inflammatory cell traffic is pathological or a physiological response to injury is not clear.

The evidence that the fibrogenetic response has become autonomous (i.e. not a simple tissue repair mechanism) is strong. In gene over-expression animal models, the transient over-expression of transforming growth factor-b resulted in a progressive fibrosing lung disease that mimicked cryptogenic fibrosing alveolitis, with a histopathological appearance that was indistinguishable from usual interstitial pneumonia.

Pathology

Usual interstitial pneumonia is the pattern seen in patients with cryptogenic fibrosing alveolitis. The hallmark of usual interstitial pneumonia is a patchy distribution of interstitial fibrosis, chronic inflammatory cell infiltrate, enlarged cystic air spaces (honeycomb lung), and normal lung (Fig. 1 and Plate 1). Inflammation is often mild and the fibrosis characterized by acellular collagen bundles with foci of proliferating fibroblasts. This non-uniform appearance, which is often visible under low microscopic power, suggests the pathological processes are at different stages of development (temporal heterogeneity), a feature that distinguishes usual interstitial pneumonia from other interstitial pneumonias.

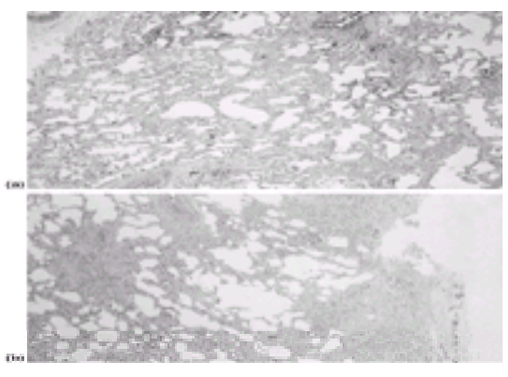

Fig. 1 Histopathological appearance of cryptogenic fibrosing alveolitis and the non-specific interstitial pneumonia 'mimic'. (a) Usual interstitial pneumonia, the histopathological pattern seen in cryptogenic fibrosing alveolitis. Note the pale, fibroblast foci that are the hallmark of usual interstitial pneumonia. (b) The non-specific interstitial pneumonia 'mimic' of cryptogenic fibrosing alveolitis. This is much less common than usual interstitial pneumonia. Note the uniformity of the pathology throughout the section. (See also Plate 1.)

Clinicopathological correlations

Correlations between clinical and physiological indices of disease and lung biopsy appearances have, in general, proved to be disappointing. This is probably a reflection of the fact that biopsy samples a small area of the peripheral (and most involved) part of the lung, whereas other indices reflect the function of the whole of both lungs. Despite this, it has been shown that the degree of lung involvement can be correlated with lung function indices. More recent studies have utilized high resolution computed tomography (that 'samples' the anatomy of the whole of both lungs) instead of biopsy and have shown better correlations with lung function indices, especially gas transfer, and have concluded that high resolution computed tomography provides a more accurate prediction of outcome.

Clinical features

History

A history of progressive breathlessness on exertion in the absence of wheeze is typical. A dry cough may be present, but sputum production is unusual until the later stages of the disease. Haemoptysis is uncommon and should suggest the development of lung malignancy that occurs with a 7 to 14-fold relative risk in cryptogenic fibrosing alveolitis. Chest pain is uncommon. Constitutional symptoms such as weight loss and lethargy are recognized.

A full occupational and domestic environmental exposure (from school leaving), and drug ingestion history, are necessary to identify diseases that can mimic cryptogenic fibrosing alveolitis (see Table 1). A history of other diseases, particularly the rheumatological disorders, is important because the diffuse lung disease in this context can mimic, but is quite different from, cryptogenic fibrosing alveolitis.

Examination

Digital clubbing is present in 70 to 80 per cent of patients. On auscultation, very fine crackles are heard at the lung bases and in the midaxillary line, occurring at the end of inspiration in early cases but becoming paninspiratory in more advanced disease. In the presence of more subtle disease, the crackles may disappear as the patient leans forward, but usually they persist in the midaxillary line.

At more advanced stages of disease, central cyanosis may be evident, also signs of pulmonary hypertension and right ventricular failure. In addition, general examination may reveal non-pulmonary features that would suggest alternative, systemic diseases, such as arthropathy, vasculitis, skin disorders, and peripheral lymphadenopathy.

Investigations

Imaging

Chest radiography

A typical chest radiograph of a patient with cryptogenic fibrosing alveolitis is characterized by small lung fields and reticulonodular shadowing, particularly at the periphery of the lung and at the bases, obscuring the right and left heart borders and making the diaphragmatic surfaces irregular (Fig. 2). Even in more subtle examples of fibrosing alveolitis, this distribution of radiographic abnormality should suggest the diagnosis. In more advanced cases, all lung zones are involved, at which point evidence of honeycomb shadowing may be present.

Fig. 2 Chest radiograph of a patient with cryptogenic fibrosing alveolitis. This shows the typical features of the disease with peripheral, predominantly basal, reticulonodular changes obscuring the heart borders and diaphragms.

Lymphadenopathy is rarely observed on chest radiography and the presence of pleural disease should suggest an alternative diagnosis. Cardiomegaly and prominent pulmonary arteries indicate secondary pulmonary hypertension.

High resolution computed tomography

The use of high resolution computed tomography over the last 10 years has revolutionized the approach to diffuse lung disease. The pattern of abnormality may be characteristic in a number of diffuse lung diseases (see Chapter 17.11.1) and is virtually pathognomonic in cryptogenic fibrosing alveolitis. Typical early changes are of a peripheral rim of reticular change at the bases, posteriorly (Fig. 3). As disease becomes more extensive, these changes are observed in the other lung zones and more centrally. Computed tomography confirms that pleural disease is not present in cryptogenic fibrosing alveolitis but, by contrast to the observation on plain chest radiography, mediastinal lymphadenopathy is commonly present. In more subtle cases, it is important to perform prone as well as supine scans to exclude the contribution of gravity to the radiographic appearances due to vascular and interstitial pooling in the dependent areas.

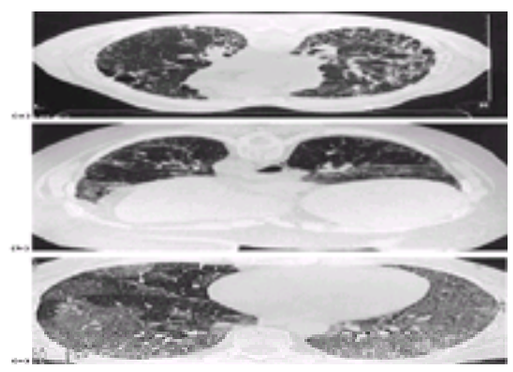

Fig. 3 High resolution computed tomography of patients with diffuse lung diseases that can mimic cryptogenic fibrosing alveolitis. (a) The predominantly peripheral, coarse honeycombing is characteristic of cryptogenic fibrosing alveolitis. (b) More fine fibrotic change is seen in non-specific interstitial pneumonia. Note the absence of honeycombing and the less clear-cut peripheral distribution of disease. (c) Desquamative interstitial pneumonia showing widespread, ground glass opacification with maintenance of the fine architecture of the lung and no evidence of fibrosis.

Radionuclide imaging

Ventilation–perfusion scans

Ventilation–perfusion scans show mismatching of perfusion to ventilation in cryptogenic fibrosing alveolitis, making the test unreliable in excluding thromboembolic disease in this situation. This is important because there is an increased incidence of pulmonary embolus in this disease and the diagnosis of pulmonary embolus will therefore rely on the clinical features, identification of the venous source of emboli, and, in some instances, spiral computed tomography to identify vascular occlusion in the more proximal arteries.

Gallium scanning

Gallium scanning is a highly sensitive but non-specific test, being positive in a wide range of diffuse processes that involve macrophage activation. It has no role in the diagnosis or management of cryptogenic fibrosing alveolitis.

^{99m}Tc-diethylinetriamine pentacetate (DTPA) clearance

Clearance from the lung of inhaled ^{99m}Tc-diethylinetriamine pentacetate (DTPA) is of value in identifying early disease and is a helpful test in prognosis; patients with a persistently normal clearance run a more stable, non-progressive course. Cigarette smoking will produce increased clearance rates and the test is therefore only of value in non-smokers or those who have given up smoking for at least 1 month prior to assessment. The role of DTPA clearance studies in routine clinical management is not yet clear.

Lung function tests

Fibrosing alveolitis is characterized by a restrictive ventilatory defect of mechanical function resulting in reduced pulmonary compliance, vital capacity, and total lung capacity. Residual volume is usually decreased unless there is coincident airflow obstruction due to cigarette smoking and lung recoil pressure is increased.

Carbon monoxide transfer factor (*DL*CO, a measure of diffusion capacity) is reduced and may be the only abnormality in early disease. In most patients the gas transfer measurement adjusted for alveolar volume (*K*CO) is also reduced, but less than *DL*CO, indicating that the capacity to exchange gas is impaired in lung that has not been destroyed. If there is significant coexisting emphysema, lung volumes will be well preserved in the face of a disproportionately depressed gas transfer measurement in both *DL*CO and *K*CO. Gas transfer is reduced by both the emphysematous and the fibrosing processes, whereas lung volumes will tend to be increased by emphysema but reduced by fibrosis and these two opposing influences result in relatively normal-sized lungs radiographically and physiologically.

Typical blood gas measurements will reveal a reduced PaO_2 value with a normal or low $PaCO_2$ measurement. In more advanced cases, the $PaCO_2$ will be reduced because of the increase in ventilatory drive in a patient with more severe lung stiffening due to fibrosis, but in terminal stages CO_2 may rise. The low PaO_2 is largely attributable to ventilation–perfusion mismatching. On exercise, hypoxaemia is exacerbated and a widening of the alveolar–arterial (A–a) gradient is observed. Infrequently, these abnormalities on exercise testing are the only physiological abnormalities, but usually by the time the patient seeks advice there is already some abnormality in the gas transfer measurement at rest.

Lung function measurements (to include gas transfer) should be made sequentially to assess the progression of the disease process. Spirometry alone or, worse, peak flow measurements, are inadequate. It is sensible to plot out serial lung function studies to visualize more gradual change that may be missed if results are

compared only with the previous set of measurements.

Blood tests

Blood tests are of little value in the diagnosis of cryptogenic fibrosing alveolitis. In severe cases, secondary polycythaemia may be observed and a high neutrophil count may indicate superadded infection. Corticosteroid therapy will elevate total white count to around 13 to 14 × 10^{9}/l, sometimes even higher.

Elevations in one or more classes of immunoglobulins, particularly IgG and IgM, may be seen and rheumatoid factor or antinuclear antibody may be present in abnormal titres in approximately 45 per cent of patients, but none of these immunological assessments is specific. The titres of autoantibodies do not approach those seen in the rheumatological diseases. In cases where there is a history of significant antigen exposure, or when there is doubt about relevant exposures, it is helpful to perform precipitin tests against the fungal antigens which produce diseases such as farmer's lung or the avian antigens which can provoke budgerigar or pigeon fancier's lung.

Bronchoalveolar lavage

Bronchoalveolar lavage is valuable in excluding infection and as confirmation that the histopathological pattern is likely to be the usual interstitial pneumonia pattern in patients who cannot or will not tolerate surgical biopsy. In a typical patient with cryptogenic fibrosing alveolitis, bronchoalveolar lavage would produce an increase in total cell returns of three to six fold (up to 6 × 10^{5}/ml of fluid return), and of these up to 20 per cent may be neutrophils or eosinophils. A large excess of lymphocytes suggests an alternative diagnosis such as granulomatous or drug-induced causes of lung disease. Serial bronchoalveolar lavage is of no value in monitoring disease.

Lung biopsy

The only lung biopsy technique that provides useful information is surgical lung biopsy, either through minithoracotomy or video-assisted thoracoscopic biopsy (the most common procedure). The surgeon will biopsy at least two sites to ensure representative sampling and the biopsy can be divided into parts that will be stored, if necessary, for immunohistochemical, molecular, and electron microscopical analysis in addition to the more routine histopathological evaluation.

Transbronchial biopsy cannot be used for diagnosis of cryptogenic fibrosing alveolitis and is only helpful in excluding other conditions such as granulomatous disease. The specimens produced are small, and do not allow an assessment to be made of the individual patterns that are important in differentiating the different idiopathic interstitial pneumonias.

Treatment

Established drug regimens

There have been no prospective, placebo-controlled, randomized trials of treatment in this condition. Most information on drug efficacy is derived from individual comparative series. Many of these have been retrospective. Assessment of efficacy is further complicated by recent knowledge that historic series probably included patients whose disease had the histopathological appearances of interstitial pneumonias other than usual interstitial pneumonia. These qualifications mean that the level of evidence base on which recommendations are made is low. However, world-wide experience, coupled with the recently published American Thoracic Society/ European Respiratory Society statement on cryptogenic fibrosing alveolitis, suggests that a combination of low-dose prednisolone (dosage varies between centres internationally but 20 mg on alternate days is the author's preference) together with azathioprine at 2.5 mg/kg per day up to a maximum of 150 mg/day should be the first-line treatment (Table 2). Azathioprine treatment should be commenced at 50 mg/day for 1 month with weekly full blood count monitoring to ensure that the individual does not have the methyltransferase deficiency that predisposes to enhanced toxicity. Provided blood counts are stable at 4 weeks, dosage can be increased to the maximum. Full blood count and liver function tests need to be performed 6 to 8-weekly.

An alternative to azathioprine is cyclophosphamide at 2 mg/kg per day up to a maximum of 150 mg per day. In addition to the potential bone marrow toxicity, haemorrhagic cystitis and bladder neoplasm are complications. Regular dipstick analysis for blood is recommended by some, ideally at weekly intervals. There is no place for high-dose corticosteroid treatment in the routine management of cryptogenic fibrosing alveolitis.

Other treatment regimens

Other drugs that have been used to treat this disease are colchicine (which has a similar efficacy to corticosteroids alone but with less side-effects), cyclosporin (very little data and no justification for use of this agent), penicillamine (no evidence of efficacy), pirfenidone (possible disease stabilization during corticosteroid reduction in a small subset of patients, but not commercially available), and interferon-g (small study with reservations that the subgroup studied was not representative of cryptogenic fibrosing alveolitis as defined above). None of the approaches other than colchicine are recommended for the reasons indicated.

Acute exacerbations or accelerated disease

Sudden deterioration can occur. Supervening infection, heart failure, and thromboembolism must be excluded. If accelerated disease is believed to be the cause of the deterioration, intravenous corticosteroids (1 g/day methylprednisolone for 3 days) together with intravenous cyclophosphamide (600 mg/m^{2} as a single dose, repeated at roughly 2-week intervals if blood counts are satisfactory) should be considered.

Transplantation

Single lung transplantation is now the organ replacement therapy of choice for end-stage fibrosing alveolitis. The procedure has not been used for a sufficiently long period to provide information about long-term survival, but approximately 60 per cent are alive at 3 years.

Supportive therapy

When all treatment options have failed, supportive therapy is necessary. Supplemental oxygen may be required and this can be provided in the home through oxygen concentrators. Diuretics may be necessary and infection should be treated promptly. In the terminal phases, small dosages of opiates have been shown to suppress the sensation of extreme breathlessness that occurs as the lungs become much less compliant.

This chronic, often relentlessly progressive, disease has a disabling affect on the patient and their close family. Full support by medical and non-medical health-care professionals should be involved in the patient's care; medical social workers, physiotherapists, occupational therapists, and rehabilitation programmes for patients form an important part of supportive management.

Monitoring of treatment

Immunosuppression can take 3 to 6 months, and sometimes longer, to produce a maximal effect. Patient monitoring with full lung function tests at 3-monthly intervals during the first year is recommended. If the disease becomes stable or improvement is seen, treatment should be continued until the disease has stabilised for a total of 1 year. A tapering, with a view to complete withdrawal of immunosuppression, should be tried at this point. If function deteriorates, alternative drug therapies need to be tried. Once lung function is less than 30 per cent predicted, consideration for transplantation is required. The age limit imposed by some transplantation centres needs to be taken into account when planning management.

Prognosis

Historically, survival was thought to be roughly 50 per cent at 4 years. More recent studies comparing survival of incident cases as opposed to prevalent cases, and with a more rigid definition of cryptogenic fibrosing alveolitis as outlined above, have shown the prognosis is even more desperate, with median survival being less than 3 years and 10-year survival of 5 to 10 per cent. These appalling survival figures highlight the need for early detection, early introduction of treatment, and early

consideration of transplantation in patients who fail to respond to therapy.

Areas of uncertainty

What is the cause?

It is unlikely that a single aetiology will be identified and, even if it were, removal of that cause is unlikely to be curative. Multiple trigger factors are likely to operate in conjunction with some form of genetic predisposition resulting in disease. Patients often expect clinicians to be able to identify a cause that results in a cure and they need to be told that this will not happen.

Should all patients undergo surgical biopsy?

Decisions about biopsy need to take into account the degree of certainty of diagnosis using all of the investigations outlined above, particularly high resolution computed tomography and bronchoalveolar lavage. If there is any doubt about diagnosis, then surgical biopsy is needed to confirm the diagnosis. These decisions must always take into account the individual patient and need to be made in the context of other potentially complicating, comorbid conditions and whether there is a likelihood of an increased complication rate. Chronological age should not be a factor but clearly biological age is.

When should treatment be commenced?

It is a mistake to wait until a patient has limiting breathlessness before commencing treatment. At this stage at least 50 per cent of lung function is likely to have been lost and the patient will have little reserve. Many studies have shown that treatment is more likely to be successful if commenced early. The key issue is the likelihood of improving or stabilizing disease against the likelihood of side-effects. This has to be related to the individual. Factors that predict likely response to therapy are ground glass appearances (indicating cellularity) on high resolution computed tomography and comorbid conditions such as hypertension and diabetes mellitus that will increase the risk of corticosteroid-induced side-effect.

It is also important to remember that functional deficit caused by lung damage and scarring cannot be reversed; it is only the more recent, inflammatory pathological response that is responsive to currently used treatment. Improvement in lung function is seen in only 5 to 10 per cent of cases. Therapeutic response, if it occurs, more probably results in disease stability. The patient, and some clinical colleagues, need to be told this. This emphasizes the importance of attempting to stabilize disease at a level of function that is acceptable to the patient.

How long should treatment continue?

Optimal duration of therapy is not known. An approach to the management of immunosuppressant drugs is described above under 'Treatment'. The issue with regard to corticosteroids is less clear-cut. There is anecdotal evidence that complete corticosteroid withdrawal can result in a rebound of disease activity, which cannot then be controlled. It is therefore reasonable to attempt to reduce corticosteroids to approximately 10 mg every other day and, if disease remains stable at this level, then this dosage of drug should be continued indefinitely. The likelihood of significant side-effects on such a small dosage is low set against the possibility of disease flare-up on complete withdrawal of treatment.

Areas for future research

The major areas of future research involve a clarification of epidemiological issues and a resolution of the issue of possible immunogenetic predisposition and the relationship between environmental exposures and immunogenetic predisposition. Differences in the evolution of fibrogenesis in different patterns of idiopathic interstitial pneumonia need to be studied and this will probably involve a better appreciation of different functional phenotypes of fibroblasts, with particular regard to their connective tissue matrix products. Prospective, properly controlled, double-blind studies of treatment efficacy, using cohorts of patients with the usual interstitial pneumonia histopathological pattern, are of paramount importance to identify optimal treatment. This will require an international collaborative effort.

Further reading

Clinical

Joint Authors Group (2000). Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. *American Journal of Respiratory and Critical Care Medicine* **161**, 646–64.

Carrington CB, Gaensler EA, Coutu RE (1978). Natural history and treated course of usual and desquamative interstitial pneumonia. *New England Journal of Medicine* **298**, 801–9.

Diffuse Parenchymal Lung Disease Group (1999). The diagnosis, assessment and treatment of diffuse parenchymal lung disease in adults. *Thorax* **54** (Suppl. 1).

Liebow AA (1975). Definition and classification of interstitial pneumonias in human pathology. In: Basset F, Georges R, eds. *Progress in respiratory research, 8. Alveolar interstitium of the lung,* pp. 1–33. Karger, New York.

Scadding JG, Hinson KFW (1967). Diffuse fibrosing alveolitis (diffuse interstitial fibrosis of the lungs). *Thorax* **22**, 291–304.

Pathology

Katzenstein AL, Myers JL (1998). Idiopathic pulmonary fibrosis: clinical relevance of pathologic classification. *American Journal of Respiratory and Critical Care Medicine* **157**, 1301–15. Review with 66 references.

Pathogenesis

Agostini C, Semenzato G (1996). Immunology of idiopathic pulmonary fibrosis. *Current Opinions in Pulmonary Medicine* **2**, 364–9.

Gauldie J, Sime PJ, Xing Z, Marr B, Tremblay GM (1999). Transforming growth factor-beta gene transfer to the lung induces myofibroblast presence and pulmonary fibrosis. *Current Topics in Pathology* **93**, 35–45.

Sime PJ, Xing Z, Graham FL, Csaky KG, Gauldie J (1997). Adenovector-mediated gene transfer of active transforming growth factor- beta1 induces prolonged severe fibrosis in rat lung. *Journal of Clinical Investigation* **100**, 768–76.

Imaging

Howling SJ, Hansell DM (2000). Spiral computed tomography for pulmonary embolism. *Hospital Medicine*, **61**, 41–5.

Wells A (1998). Clinical usefulness of high resolution computed tomography in cryptogenic fibrosing alveolitis. *Thorax* **53**, 1080–7. Review with 67 references.

Bronchoalveolar lavage

Haslam PL, Turton CWG, Lukoszek A, *et al*. (1980). Bronchoalveolar lavage fluid cell counts and cryptogenic fibrosing alveolitis and their relation to therapy. *Thorax* **35**, 328–9.

Weinberger SE, Kelman JA, Elson NA (1978). Bronchoalveolar lavage in interstitial lung disease. *Annals of Internal Medicine* **89**, 459–66.

Treatment

Johnson MA, Kwan S, Snell NJC, *et al*. (1989). Randomised controlled trial comparing prednisolone alone with cyclophosphamide and low dose prednisolone in combination in cryptogenic fibrosing alveolitis. *Thorax* **44**, 280–8.

Lynch JP, McCune WJ (1997). Immunosuppressive and cytotoxic pharmacotherapy for pulmonary disorders. *American Journal of Respiratory and Critical Care Medicine* **155**, 395–420. Review with 353 references.

Mason RJ, Schwarz MI, Hunninghake GW, Musson RA (1999). NHLBI workshop summary. Pharmacological therapy for idiopathic pulmonary fibrosis. Past, present, and future. *American Journal of*

Respiratory and Critical Care Medicine **160**, 1771–7.

Raghu G, DePaso WJ, Cain K, *et al*. (1991). Azathioprine combined with prednisolone in the treatment of idiopathic pulmonary fibrosis: A prospective double blind randomised placebo controlled clinical trial. *American Review of Respiratory Diseases* **144**, 291–6.

Selman M, Carrillo G, Salas J, *et al.* (1998). Colchicine, D-penicillamine, and prednisone in the treatment of idiopathic pulmonary fibrosis: a controlled clinical trial. *Chest* **114**, 507–12.

Ziesche R, Hofbauer E, Wittmann K, Petkov V, Block LH (1999). A preliminary study of long-term treatment with interferon gamma-1b and low-dose prednisolone in patients with idiopathic pulmonary fibrosis [see comments]. *New England Journal of Medicine* **341**, 1264–9.

Prognosis

du Bois RM (1997). Management of idiopathic pulmonary fibrosis: prognostic indicators. *Monaldi Archives for Chest Disease* **52**, 547–51. Review with 61 references.

Turner-Warwick M, Burrows B, Johnson A (1980). Cryptogenic fibrosing alveolitis: response to corticosteroid treatment and its effect on survival. *Thorax* **35**, 593–9.

Watters LC, King TE, Schwartz MI (1986). A clinical, radiographic and physiologic scoring system for the longitudinal assessment of patients with idiopathic pulmonary fibrosis. *American Review of Respiratory Diseases* **133**, 97–103.

17.11.3 Bronchiolitis obliterans and organizing pneumonia

R. M. du Bois

Introduction

The bronchioles are defined as airways that have no cartilaginous support and comprise the terminal bronchioles and the respiratory bronchioles leading to the alveolar ducts. In many diffuse parenchymal lung disorders, the small bronchioles are involved in the histopathological process and bronchiolar disorders are often included in the classification of diffuse parenchymal lung diseases. The nomenclature of bronchiolitis has become confused. This is because pathological and clinical descriptions have been used interchangeably. There are three main histopathological patterns (Fig. 1 and Table 1):

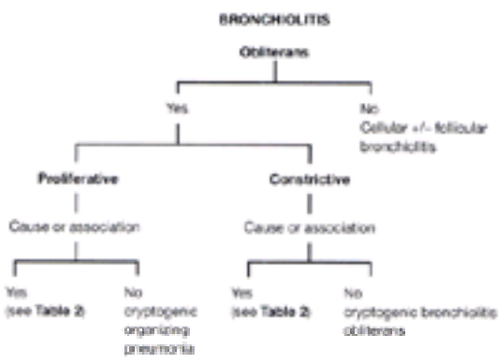

Fig. 1 Algorithm for the classification of bronchiolitis.

- proliferative bronchiolitis;
- constrictive bronchiolitis;
- cellular/ follicular bronchiolitis.

Proliferative and constrictive bronchiolitis patterns both cause obliteration of the bronchioles—'bronchiolitis obliterans'—but are associated with distinct clinical patterns of disease. Both may be the result of a known or unknown cause or association (see Table 2).

Nomenclature

Bronchiolitis obliterans is a term that has been used for many years to denote a relentless fibrosis of the bronchioles that results in severe, generally irreversible, airflow obstruction, and is characterized by obliteration of the terminal bronchioles by fixed fibrosis. In 1983, Davison *et al*. described a group of patients who presented with pneumonic-like features, but without evidence of infection, and in whom histopathological appearances were those of intra-alveolar organization spreading proximally to the terminal bronchiole. This disorder also obliterated the bronchioles, but with intraluminal loose, fibrous tissue, quite different from that seen in bronchiolitis obliterans. No cause was found for this condition and it was termed cryptogenic organizing pneumonitis. In 1985, Epler *et al*. rediscovered the entity, reporting the results of analysis of a large number of histopathological specimens, and gave the disease the name bronchiolitis obliterans organizing pneumonia. This confusion of nomenclature resulted in bronchiolitis obliterans and organizing pneumonia being regarded as synonyms.

An American Thoracic Society/ European Respiratory Society nomenclature committee has now redefined the diffuse lung diseases, including organizing pneumonia. It recommended that the idiopathic (or cryptogenic) variant of organizing pneumonia should be known as cryptogenic organizing pneumonia, with the histopathological features of organizing pneumonia. It must be noted that this disease and these pathological features are synonymous with idiopathic bronchiolitis obliterans organizing pneumonia, a term that will undoubtedly still appear in the literature. Bronchiolitis obliterans should be the term for the irreversible airway disease.

Bronchiolitis obliterans

Introduction

As with every form of bronchiolitis, attempts should be made to identify a potential trigger (see Table 2). Once all known triggers and associations have been excluded, the disease can be considered cryptogenic. Occasionally, the initial presentation of rheumatological disease is with the pulmonary manifestation, and it is therefore important to exclude this as far as possible through a full clinical history and examination, and also by testing for autoantibodies.

This disease is the result of progressive obliteration of the terminal bronchioles with connective tissue matrix. It is a relentless condition that is usually non-responsive to therapy. The pathogenesis is unknown.

Histopathology

The terminal bronchioles are predominantly involved with some extension into the proximal respiratory bronchioles. The lumens are significantly narrowed and often obliterated by a mixture of cellular inflammation and fibrosis–constrictive bronchiolitis (Fig. 2 and Plate 1). In advanced stages, the airways are occluded by dense, relatively acellular connective tissue matrix.

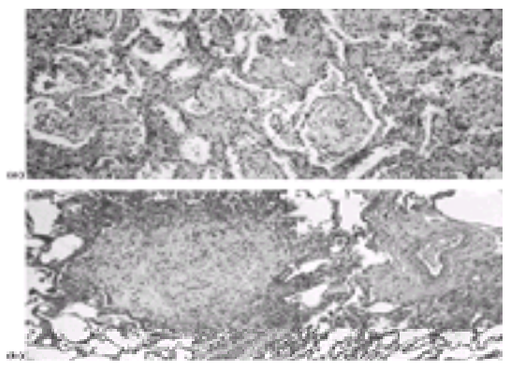

Fig. 2 Histopathology. (a) Proliferative bronchiolitis. (b) Constrictive bronchiolitis. Note the loosely packed granulation tissue in (a) in contrast to the more established scarring in (b). (See also Plate 1.)

Clinical features

The most striking clinical feature is progressive breathlessness, often with little wheeze despite this being predominantly an airway disease. Cough may occur, but is rarely productive, and haemoptysis is not a feature. Chest pain is uncommon but chest tightness may be described. Clinical examination of the respiratory system is unremarkable. Occasionally an inspiratory 'squeak' is heard, which is very characteristic of small airway disease. Examination of other systems may identify subtle rheumatological disease.

Investigations

Imaging

Chest radiography shows large lung fields with some loss of vascular markings but no evidence of infiltration.

High resolution computed tomography is more informative (Fig. 3). There is a patchiness of attenuation, described as 'mosaicism'. This represents variable degrees of lung perfusion caused by patchy vasoconstriction in affected areas, this being due to reflex hypoxic vasospasm in response to airway obliteration. The mosaic pattern is enhanced on expiratory CT because gas trapping enhances the differences between perfused and non-perfused lung.

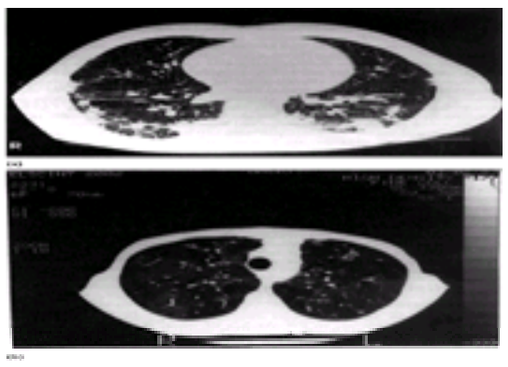

Fig. 3 High resolution computed tomography. (a) Organizing pneumonia. Peripheral consolidation is the classic feature. Histopathology is proliferative. (b) Bronchiolitis obliterans. The attenuation of the lung fields varies from grey (normal perfusion) to black (reduced perfusion due to hypoxic vasoconstriction resulting from airway obliteration). Histopathology is constrictive.

Lung function tests

Lung function tests show fixed airflow obstruction with an increase in residual volume and total lung capacity measured by plethysmography. Total gas transfer for carbon monoxide (*DL*CO) is reduced, but the gas transfer index (total gas transfer corrected for alveolar volume) is preserved. This is an important distinguishing feature from airway diseases that cause obliteration of the vascular bed, such as emphysema. The arterial partial pressure of oxygen (PaO_2) is relatively well preserved until disease is severe.

Other investigations

Blood tests are unhelpful, with the exception of assays for autoantibodies that might suggest rheumatological disease. Bronchoalveolar lavage is not normally indicated and may be dangerous if disease is severe. However, when performed, an excess of neutrophils is characteristic.

Differential diagnosis

Cryptogenic bronchiolitis obliterans should be distinguished from other forms of chronic airflow obstruction, such as asthma, chronic obstructive pulmonary disease (COPD), and known causes or associations that result in bronchiolitis obliterans. The relatively well-preserved gas transfer index is an important feature distinguishing this condition from emphysema. The appearances on high resolution computed tomography (particularly the expiratory scan) are characteristic of obliterated small airways. In exceptional circumstances, surgical biopsy is needed to confirm the diagnosis. If this is undertaken, the characteristic airway abnormalities can sometimes be elusive. Multiple sectioning throughout the whole block is required to obtain regions that would allow a correct diagnosis to be made.

Treatment

Treatment response is generally poor. It is important to give a trial of corticosteroids to exclude reversibility. A reasonable approach would be to give 40 mg of prednisolone per day for 4 weeks, measuring pulmonary function before and after, paying particular attention to changes in residual volume and gas transfer. If there is significant objective response, steroid dosage should be tapered and maintenance inhaler therapy given. This should consist of inhaled corticosteroids and, possibly, a long-acting b_2-adrenoceptor agonist such as salmeterol or eformoterol. Immunosuppressants, typically azathioprine, have been tried but with limited success.

In the younger patient, lung transplantation should be considered. However, the most common long-term problem encountered with lung transplantation is an obliterative bronchiolitis. Counselling of the patient is particularly important in this situation.

Organizing pneumonia

Introduction

Known causes of, and associations with, organizing pneumonia should be identified (see Table 2). In the absence of these, the term cryptogenic organizing

pneumonia is used. Clinical features are very similar for the cryptogenic and non-cryptogenic disease.

Histopathology

The most striking abnormality is the filling of the airways distal to the terminal bronchiole and the alveoli with granulation tissue in a peribronchiolar distribution, comprising buds of loose collagen and connective tissue matrix cells with a uniform appearance (Fig. 2). There may be some surrounding chronic inflammation. The intra-alveolar granulation tissue may spread from one alveolus to another thorough the pores of Kohn.

Clinical features

The disease presents subacutely with a history of, usually, 2 to 3 months' non-productive cough, shortness of breath, and the systemic features of fever, weight loss, and malaise. It affects men and women equally and occurs predominantly in the fifth and sixth decades. Digital clubbing is rare. A few crackles may be heard on auscultation, but this is not invariable. Features of systemic disease may suggest that the lung problems are secondary.

Investigations

Imaging

Chest radiography shows, most commonly, a bilateral, predominantly peripheral, and often basal, pattern of consolidation. Very occasionally there is an interstitial pattern reminiscent of fibrosing alveolitis, and even more rarely the pattern is of a single nodule. Cavitation can occur but is uncommon. Pleural disease is rare.

High resolution computed tomography defines the pattern more clearly and often shows the disease as more extensive than would be suspected from plain chest radiography (Fig. 3). The predominantly peripheral pattern of confluent disease with air bronchograms may be confirmed. A rarer variant, with a very marked peribronchovascular bundle distribution, can be associated with changes suggestive of fibrosis. This more interstitial variant is less likely to resolve.

Lung function tests

Lung function tests show a restrictive ventilatory defect with reduced gas transfer. $Pa{O_2}$ is well maintained unless disease is severe.

Other tests

Blood tests are non-specific, suggesting an inflammatory process with a leucocytosis, elevated erythrocyte sedimentation rate, and C-reactive protein. A seasonal variant of disease (occurring between February and May) has been reported in one series and this was associated with abnormal liver function tests.

Bronchoalveolar lavage

Typical features are in increase in lymphocytes together with neutrophils and/or eosinophils. The CD4:CD8 ratio is low. Foamy macrophages are characteristic. Occasionally, mast cells and plasma cells may be seen.

Differential diagnosis

All other causes of consolidation visible on imaging need to be considered. These include infection, eosinophilic pneumonia, alveolar haemorrhage, alveolar proteinosis, alveolar cell carcinoma, vasculitis, and lymphoma. Patients who have atypical features or who have not responded to treatment as completely as would be expected should have the alternative diagnoses considered.

Diagnosis is most cases will require a surgical biopsy. In exceptional circumstances, the diagnosis may be made in the context of the classical (consolidation) form of disease on the basis of typical bronchoalveolar lavage findings, together with a transbronchial biopsy that is consistent. It is noteworthy, however, that areas of organization are seen on histopathological examination of biopsies from a variety of disorders such as vasculitis, eosinophilic pneumonia, and malignancy, but in these circumstances they are minor features and not the dominant pathology. Unfortunately, a transbronchial biopsy may only sample these minor features, thereby resulting in misdiagnosis.

Treatment

Corticosteroids usually work extremely well. It is usual to commence with a dosage of 0.75 mg/kg per day orally for a month and then gradually taper whilst monitoring symptomatic, radiological, and physiological improvement. It is usually possible to discontinue corticosteroids after 6 to 12 months. Relapses can occur at any time but usually respond to a reintroduction of corticosteroids. Rarely, immunosuppressants, such as azathioprine or cyclophosphamide, need to be used as steroid-sparing agents or in more resistant disease.

Occasionally patients present with respiratory failure due to very extensive disease and in this situation pulsed intravenous methylprednisolone at 500 to 1000 mg daily for 3 days can be life saving. It is very unusual to have to use mechanical ventilation.

Prognosis

The prognosis is generally very good. Indices of a poor prognosis include organizing pneumonia that is secondary to rheumatological or other systemic disease, an imaging pattern that shows an interstitial or mixed pattern of disease, and a bronchoalveolar lavage that has no lymphocytes. Estimates of 5-year survival are 73 per cent for primary disease compared to 44 per cent for disease that occurs secondary to other disorders. Deaths due to organizing pneumonia are uncommon and in various series range from 3 to 13 per cent of cases.

Further reading

Wells AU, du Bois RM (1993). Bronchiolitis in association with connective tissue disorders. In: King TE, ed. *Clinics in chest medicine*, pp. 655–66. WB Saunders, Philadelphia.

Cordier J-F (2000). Organising pneumonia. Thorax rare disease series 8. *Thorax* **55**, 318–28.

King TE Jr (2000). Bronchiolitis. In: du Bois R M and Olivieri D, eds. *European Respiratory Monograph* **14**, 244–66.

Lynch JP, Belperio J, Flint A, Martinez F (1999). Bronchiolar complications of connective tissue disorders. *Seminars in Respiratory and Critical Care Medicine* **20**, 149–67.

Wright JL, Cagle T, Churg A, Colby TV, Myers J (1992). Diseases of the small airways. *American Review of Respiratory Diseases* **146**, 240–62.

17.11..4The lungs and rheumatological diseases

R. M. du Bois and A. K. Wells

Introduction

Lung involvement can occur in all rheumatological diseases, with different patterns of respiratory pathology predominant in different diseases. The frequency and best management of these problems is often uncertain because, although common, they have been less well documented than the more obvious rheumatological features. Respiratory involvement may be subclinical, symptoms often being masked by exercise limitation due to musculoskeletal factors.

In some patients, diffuse lung disease is the presenting feature of the rheumatological disease, when typical computed tomography appearances together with autoantibody studies will point to the correct diagnosis.

Imaging

Chest radiography will identify the predominantly reticulonodular pattern of the diseases that mimic fibrosing alveolitis (see Chapter 17.11.1 and Chapter 17.11.2), the consolidation of organizing pneumonia and alveolar haemorrhage, the ground glass pattern of acute pneumonitis, the hyperinflation of bronchiolitis obliterans, and the presence of pleural disease.

High resolution computed tomography will enhance the sensitivity, precision, and accuracy of diagnosis. It will also reveal combinations of patterns that suggest specific rheumatological diseases. In fibrosing alveolitis, there is a reticular pattern that is initially basal and peripheral with honeycombing. More extensive disease progresses upwards and anteriorly. Ground glass changes are found in more cellular disease; mosaicism (alternating regions of increased and decreased attenuation) is typical of small airway disease. Bronchiectasis and pleural changes are seen more clearly than on chest radiography. Prediction of reversibility is much more precise than with chest radiography.

Lung function

Lung function testing does not discriminate in terms of precise pulmonary diagnosis but will differentiate obstructive ventilatory defects, characteristic of airways disease, from the restrictive pattern with reduction in gas transfer that is seen in all forms of diffuse lung disease. Lung function testing is the gold standard measure of extent of disease. Exercise tests will unmask more subtle disease, often subclinical, but are not routine requirements. A measure of gas transfer, however, is always needed.

Clinical features of lung disease in particular autoimmune rheumatic disorders

Systemic sclerosis

The criteria for classification of systemic sclerosis are described in Chapter 18.11.3. Pulmonary involvement has emerged as the major cause of excess morbidity and mortality in this disorder. The patterns of lung disease with which systemic sclerosis may present are variable and are shown in Table 1.

Fibrosing alveolitis

Clinical features

Lung disease may be the first manifestation of systemic sclerosis, and dyspnoea occurs in roughly 55 per cent of patients with this condition. Cough is a less frequently reported symptom, but when it occurs is dry and non-productive; haemoptysis indicates complicating carcinoma or bronchial telangiectasia. Pleuritic chest pain is uncommon. A history of Raynaud's phenomenon is often present. Digital clubbing is rare because of the poor vasculature of the digits. Fine crackles are heard at the bases and are of 'velcro' character. Capillaroscopy, digital thermography, and a positive antinuclear antibody can be helpful in suggesting the correct diagnosis. Lung fibrosis occurs more commonly in patients with the Scl 70, anti-DNA topoisomerase autoantibody. Chest radiographic series have identified diffuse lung disease in up to 67 per cent of patients. Oesophageal dilatation may be seen on chest radiography and is almost universally present on high resolution computed tomography. ^{99m}Tc-diethylinetriamine pentacetate clearance has been used in some centres to identify early disease and to predict prognosis—persistently normal clearance predicts lung function stability. Lung function abnormalities are found in up to 90 per cent of patients; scleroderma of the chest wall may rarely cause extrathoracic restriction. Bronchoalveolar lavage is of value in excluding complicating infection and malignancy and in suggesting the most likely histopathological pattern, but it is of no value in monitoring disease. In non-specific interstitial pneumonia (NSIP, see Chapter 17.11.1), the most common histopathological pattern, granulocytes and lymphocytes may be present in excess; usual interstitial pneumonia is less common, associated with excess of neutrophils with or without eosinophils on lavage.

Prognosis

The prognosis of fibrosing alveolitis in systemic sclerosis is dependent upon the extent of disease at presentation, but it is much better than in cryptogenic fibrosing alveolitis. Significant improvements can be achieved in response to therapy, despite the prevalent view held by many (Fig. 1). Worse lung function, particularly vital capacity and gas transfer (*DL*CO), indicates a poorer outcome, and similar observations have been made using extent of disease on computed tomography as the criterion. Crude mortality rates are 3.9 per cent per year for males and 2.6 per cent per year for females with systemic sclerosis, with lung disease the commonest cause of death. There is also an increased risk of lung cancer.

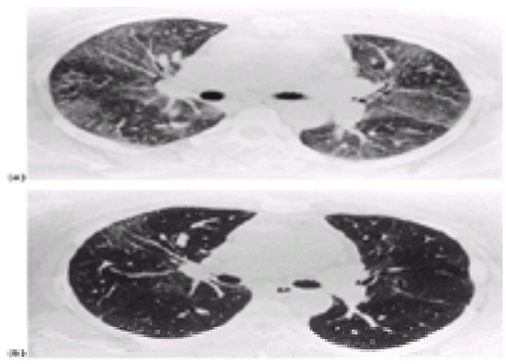

Fig. 1 CT scans of a patient with the fibrosing alveolitis of systemic sclerosis, demonstrating a non-specific interstitial pneumonia pattern of disease. (a) Before treatment. Note the widespread ground glass pattern of disease, with some areas in which the airways have been distracted demonstrating fine fibrosis. (b) After treatment. The residual, more fibrotic areas remain, but much of the abnormality has cleared.

Pulmonary vascular disease in systemic sclerosis

Unlike the other rheumatological diseases, vascular involvement in systemic sclerosis is caused not by vasculitis but by concentric fibrosis replacing the normal intima and media of small arterioles. Isolated pulmonary vascular disease occurs mainly in the limited form of systemic sclerosis (including the CREST syndrome—calcinosis, Raynaud's phenomenon, oesophageal dysmotility, sclerodactyly, telangiectasiae) and in individuals with the anticentromere autoantibody. Chest radiography, high resolution computed tomography, and bronchoalveolar lavage are all normal. Lung function studies show an isolated fall in DL_{CO} and gas transfer index (K_{CO}). When damage to the pulmonary vascular bed is extensive (gas transfer <50 per cent predicted) the risk of pulmonary hypertension increases. Mortality rates are increased with increasing pulmonary hypertension, for which Doppler echocardiography is a good screening test.

Other pulmonary complications

Aspiration pneumonia is uncommon, particularly considering the prevalence of oesophageal dysfunction in systemic sclerosis. Pleural disease is also uncommon. Rarely, the extent of skin tightness over the chest wall produces an extrinsic restriction of ventilation. Occasionally, the first manifestation of pulmonary parenchymal disease is organizing pneumonia

Polymyositis/dermatomyositis

The defining criteria for polymyositis and dermatomyositis are described in Chapter 18.11.7.

Pulmonary complications occur in up to 64 per cent of patients in combined series and are the most frequent cause of death. The main manifestations of lung disease are summarized in Table 1.

Diffuse lung disease

Diffuse lung disease is the commonest problem encountered in the context of dermatomyositis/ polymyositis. The pattern of diffuse lung disease often mimics that of fibrosing alveolitis, but a rapidly progressive form of acute pneumonitis can occur. Organizing pneumonia may also be the presenting feature. A recent report has described an acute presentation of pulmonary capillaritis with alveolar haemorrhage in association with polymyositis.

Clinical features depend upon the nature of the lung process. Breathlessness on exertion without wheeze is a common presenting symptom and, if the myopathy is severe, orthopnoea can be striking. Haemoptysis may occur if there is capillaritis. Pleural disease is uncommon. Computed tomography can be particularly helpful; a combination of patchy consolidation with a peripheral reticular pattern being highly characteristic. In recent alveolar haemorrhage or marked myopathy, there may be a disproportionate preservation of gas transfer index (K_{CO}). It must be remembered, however, that acute haemorrhage can occur in the context of previous chronic disease and so the K_{CO} may be normal or subnormal, but elevated from the baseline level. Autoantibodies to aminoacyl-tRNA synthetases are often found when inflammatory myopathies coexist with diffuse lung disease—Jo-1 (antihistidyl tRNA synthetase) is the most common. This occurs in 20 to 30 per cent of patients with inflammatory myopathy, but in 50 to 100 per cent of cases of inflammatory myopathy and diffuse lung disease, in contrast to less than 5 per cent of patients without diffuse lung disease.

Other pulmonary manifestations

Aspiration pneumonia needs to be considered if there is upper airway/ pharyngeal muscle weakness. The predilection for the posterior lung segments may suggest this diagnosis. Respiratory muscle weakness may be present in the absence of parenchymal lung disease and requires studies of muscle function for confirmation.

Rheumatoid arthritis

The American Rheumatism Association revised criteria for the classification of rheumatoid arthritis (rheumatoid arthritis) are described in Chapter 18.8. Pleuropulmonary manifestations are summarized in Table 1.

Fibrosing alveolitis

The fibrosing alveolitis of rheumatoid arthritis has a male predominance (male:female 3:1), is associated with HLA-B8 and HLA-Dw3 positivity, and is histologically similar to cryptogenic fibrosing alveolitis, but some patients may have non-specific interstitial pneumonia rather than usual interstitial pneumonia. Minor pulmonary fibrosis is common; in one open lung biopsy series pulmonary fibrosis was seen in 60 per cent of patients. Smoking is a risk factor for the development of overt pulmonary fibrosis and has also been associated with subclinical disease. High titres of rheumatoid factor and the presence of rheumatoid nodules are also associated with an increased prevalence of diffuse lung disease.

The most frequent symptom is exertional dyspnoea, although this may be masked by a general loss of mobility due to systemic disease. The clinical picture is usually identical to cryptogenic fibrosing alveolitis with bilateral, predominantly basal crackles and tachypnoea. Digital clubbing is more prevalent than in other rheumatological diseases. Reductions in carbon monoxide diffusing capacity are found in up to 40 per cent of unselected rheumatoid arthritis patients, but radiologically overt pulmonary fibrosis is found in only 1 to 5 per cent of cases (based on three large chest radiographic series).

Organizing pneumonia

The clinical presentation of organizing pneumonia is similar to infective pneumonia with systemic features of fever and weight loss, multifocal consolidation on chest radiograph and computed tomography, a restrictive functional defect, and generally a good response to corticosteroids. Histopathology shows acini filled with loose connective tissue and a variable inflammatory infiltrate. Organizing pneumonia is more common in rheumatoid arthritis than in other rheumatological diseases, with the possible exception of polymyositis.

Bronchiolitis obliterans

The clinical presentation is of progressive breathlessness. Clinical signs reveal hyperinflation of the chest, often with an inspiratory 'squeak'. Chest radiography confirms hyperinflation with no infiltration. High resolution computed tomography shows a 'mosaic' pattern of variable perfusion as a result of the hypoxic vasoconstriction that results from the airway occlusion. Lung function is often a mixed obstructive/restrictive pattern, with obstruction predominating; gas transfer index

(K_{CO}) is preserved. Bronchiolitis obliterans is characterized histologically by destruction of the terminal bronchiolar wall by granulation tissue, effacement of the lumen, and eventual replacement of the bronchiole by fibrous tissue. There is great heterogeneity in the speed of progression, with some patients having indolent disease.

The association of the use of penicillamine with obliterative bronchiolitis was first reported in the late 1970s. Following a number of case reports and small series, obliterative bronchiolitis was identified in a large cohort study in 3/133 rheumatoid arthritis patients using penicillamine, compared to 0/469 who were not using this drug. Obliterative bronchiolitis has, however, been reported in many rheumatoid arthritis patients not using penicillamine, and the relationship of the drug to this airways obliteration is unclear.

Bronchiectasis

The prevalence of bronchiectasis is higher in rheumatoid arthritis than in other rheumatological diseases. A recent literature review identified 289 patients with bronchiectasis associated with rheumatoid arthritis reported since 1928. Although associated with long-standing rheumatoid arthritis in one study, bronchiectasis was found on computed tomography in 30 per cent of 50 rheumatoid arthritis patients with normal chest radiographs on prospective evaluation. Bronchiectatic changes (often subtle) are easily visualized on computed tomography, and a combination of diffuse lung disease and airway disease on computed tomography should raise the suspicion of rheumatological disease, especially rheumatoid arthritis or Sjögren's syndrome, in patients whose lung disease may be the first manifestation of their rheumatological disorder (Fig. 2).

Fig. 2 High resolution computed tomography of a patient with rheumatoid arthritis. There are areas showing reticular changes consistent with fibrosing alveolitis. In addition, in areas where there is no fibrosis, there is evidence of bronchial wall thickening and bronchiectasis. This combination of computed tomographic features is highly characteristic of rheumatoid arthritis.

Pulmonary vasculitis

Pulmonary vasculitis is a potential but uncommon complication of rheumatoid arthritis, given the relatively high prevalence of systemic vasculitis in the disease. Similarly, diffuse alveolar haemorrhage is a rare complication of rheumatoid arthritis.

Pulmonary rheumatoid nodules

These may be single (when distinguishing from malignancy can be impossible, especially in the cigarette smoker) or multiple, and are found on chest radiography in less than 1 per cent of patients with rheumatoid arthritis, usually in association with rheumatoid nodules elsewhere in the body. They may vary in size according to underlying rheumatoid activity, and can be as much as 7 cm in diameter. Nodule cavitation occasionally causes haemoptysis, and pneumothorax can result from the rupture of subpleural nodules, but generally these are found as asymptomatic abnormalities on chest radiography. Caplan's syndrome is the association of single or multiple nodules with coal-miner's pneumoconiosis.

Pleural disease

Pleural disease is seen at autopsy in approximately 50 per cent of patients with rheumatoid arthritis and 20 per cent give a history of pleuritic chest pain. Pleural effusions are found in less than 5 per cent, usually in males, and are frequently asymptomatic, often being identified on routine chest radiography. In a minority, pleuritic pain and fever are prominent and the exclusion of empyema (which may be more prevalent in rheumatoid arthritis) is required. Occasionally, effusions may develop acutely in association with pericarditis or exacerbations of arthritis; more typically, radiographic abnormalities are chronic, often remaining unchanged for years. The fluid is an exudate, with a low glucose level (correlating poorly with serum glucose), a low pH, and, usually, a predominant lymphocytosis (although a neutrophilia is occasionally found).

Other pulmonary manifestations

Other pulmonary complications of rheumatoid arthritis are rare. Lower respiratory tract infection is increased in frequency, bronchopneumonia is a common terminal event, accounting for 15 to 20 per cent of deaths.

Sjögren's syndrome

The defining features of Sjögren's syndrome are described in Chapter 18.11.6. Pulmonary involvement is common, with objective evidence of pulmonary abnormalities in approximately a quarter of cases. This usually consists of lymphocytic infiltration producing diffuse lung disease and tracheobronchial disease.

Diffuse lung disease

Although often asymptomatic, diffuse lung disease in Sjögren's syndrome may present with cough, dyspnoea, crackles on auscultation, reticular or reticulonodular abnormalities on chest radiography, and a restrictive pattern of functional impairment. Interstitial involvement can be classified as fibrosing alveolitis or lymphocytic interstitial pneumonia in up to 10 per cent of patients with primary Sjögren's syndrome. Lymphoma may be a complicating problem and can mimic organizing pneumonia**.**Organizing pneumonia (bronchiolitis obliterans organizing pneumonia, BOOP) has been reported in Sjögren's syndrome, but occurs less frequently than in rheumatoid arthritis or polymyositis.

Tracheobronchial disease

Tracheobronchial disease may take the form of loss of mucus secretion in the trachea (xerotrachea), chronic bronchitis, or small airways disease. Xerotrachea occurs in up to 25 per cent of patients with primary Sjögren's syndrome and consists of atrophy of tracheobronchial mucous glands in association with a lymphoplasmocytic infiltrate, manifesting clinically as a relentless dry cough and endobronchial inflammation at bronchoscopy. It is likely that similar histological abnormalities in bronchi and bronchioles account for an increased prevalence of bronchial hyper-responsiveness, reported in 40 to 60 per cent of patients with primary and secondary Sjögren's syndrome. The evaluation of airflow at low lung volumes in unselected patients with primary and secondary Sjögren's syndrome has demonstrated a high prevalence of small disease.

Systemic lupus erythematosus (SLE)

The 1982 revised ACR criteria for the diagnosis of systemic lupus erythematosus are described in Chapter 18.11.2. Pleuropulmonary manifestations of this condition

are shown in Table 1.

Diffuse lung disease

The prevalence of diffuse lung disease on lung biopsy or at autopsy is highly variable (4–70 per cent). Only 3 per cent of systemic lupus erythematosus patients have clinical evidence of diffuse lung disease at the onset of systemic disease, and a disease resembling fibrosing alveolitis develops during follow-up in less than 5 per cent. The clinical presentation (dyspnoea, cough, predominantly basal crackles, a restrictive lung function defect or isolated reduction in carbon monoxide diffusing capacity, basal infiltrates on chest radiography) is typical of rheumatological fibrosing alveolitis. However, features not typical of cryptogenic fibrosing alveolitis include variably associated pleuritic pain, a paucity of patients with morphologically extensive or functionally severe lung fibrosis, and the frequent presence of enlarged peribronchiolar lymphoid follicles at lung biopsy (although other histological findings are indistinguishable from cryptogenic fibrosing alveolitis).

Acute lupus pneumonitis, seen in less than 2 per cent of systemic lupus erythematosus patients, is life-threatening with a mortality rate despite treatment of over 50 per cent once respiratory failure has developed. It must be distinguished from organizing pneumonia.

Extrapulmonary restriction

Extrapulmonary restriction is a well-recognized complication of systemic lupus erythematosus that results in exertional dyspnoea, a restrictive functional defect, and marked elevation of K_{CO}. The 'shrinking lung syndrome' was first described in patients with severe restriction and a marked reduction in lung volume on chest radiography, and is generally ascribed to respiratory muscle weakness. There is no treatment of proven efficacy, although some patients have improved with corticosteroid or immunosuppressive therapy.

Diffuse alveolar haemorrhage

Although seen more frequently than in other rheumatological disorders, diffuse alveolar haemorrhage is rare in systemic lupus erythematosus. The typical presentation of acute dyspnoea and extensive pulmonary infiltrates on chest radiography may mimic acute lupus pneumonitis, especially in the absence of haemoptysis. Diffuse alveolar haemorrhage is often life-threatening, with a mortality similar to acute lupus pneumonitis.

Pulmonary hypertension

Pulmonary hypertension was once considered rare in systemic lupus erythematosus, but is now reported with increasing frequency, and has a 2-year survival of less than 50 per cent in severe disease. Abnormalities indicative of subclinical pulmonary hypertension are found on echocardiography in 10 per cent of patients, usually in association with Raynaud's phenomenon, and thus it is likely that pulmonary hypertension results from vasoconstriction, rather than pulmonary vasculitis (which is seldom identified in systemic lupus erythematosus even at autopsy). An important alternative mechanism for pulmonary hypertension is thromboembolism, which has a high prevalence in systemic lupus erythematosus, especially in patients with antiphospholipid antibodies.

Pleural disease

Pleural disease is the most common pulmonary manifestation of systemic lupus erythematosus. Clinical or radiographic evidence of pleural involvement is seen in 20 per cent of patients at the onset of systemic disease and occurs in at least 50 per cent at some time (with pleural abnormalities at autopsy in 50–100 per cent). Pleural disease is often asymptomatic, but pleuritic pain may be recurrent or intractable. Pleural fluid is usually serosanguinous (but occasionally haemorrhagic) and exudative, with a neutrophilia in patients with pleurisy but a predominant lymphocytosis in chronic effusions.

Relapsing polychondritis

Relapsing polychondritis is described in Chapter 17.8.1. Respiratory involvement probably accounts for around 10 per cent of deaths in this condition. Pulmonary parenchymal disease is rare, but vasculitis may occur. Destruction and obstruction of the glottis, trachea, and bronchi lead to airway stricture, collapse, and distal infection. Lung function testing shows diminution in maximal inspiratory (large, extrathoracic airways) and expiratory flow (smaller, intrathoracic airways) rates, suggesting airway collapse while static recoil pressures are preserved. Chest radiography may suggest bronchiectasis, with airway thickening and dilatation being confirmed on high resolution computed tomography. Dynamic computed tomography scanning showing collapse of the larger airways on inspiratory manoeuvres can help to localize disease. Treatment is by mechanical stenting. Immunosuppression may be helpful.

Ankylosing spondylitis

Ankylosing spondylitis is described in Chapter 18.7. Fibrotic lung disease on chest radiography, largely or entirely confined to the upper zones, is the main pulmonary complication of this condition; diffuse reticulonodular infiltrates in the upper zones are usually symmetrical and are seldom extensive, except in patients with severe spinal disease or a long history of the disorder. This is uncommon, seen in only 1 to 2 per cent of a series of 2080 patients with ankylosing spondylitis. However, on high resolution computed tomography, limited interstitial lung disease, bronchiectasis, or paraseptal emphysema are present in the majority of patients, even when chest radiographic appearances are normal. There is no proven treatment to prevent the development of apical fibrosis; resistance to corticosteroid therapy being the rule. Cavities may develop within distorted fibrotic apical tissue and are sometimes colonized by mycobacteria or fungi, especially *Aspergillus fumigatus* that are isolated in up to 60 per cent of patients with apical cavitation. Life-threatening haemoptysis is an occasional complication of mycetoma formation within cavities; this may be controllable by bronchial artery embolization, the resection of a mycetoma being a treatment of last resort, due to the high prevalence of postoperative bronchopleural fistula or empyema.

Extrapulmonary restriction due to immobilization of the chest wall (costovertebral ankylosis) is an occasional complication of ankylosing spondylitis, associated with surprisingly little impairment in pulmonary function, perhaps because the diaphragm is able to make a major contribution in the presence of a high resting volume.

Mixed connective tissue disease

This syndrome is defined by the presence of features of systemic lupus erythematosus, systemic sclerosis, and polymyositis (Sjögren's syndrome may also be seen) in association with high titres (>1:1600) of autoantibody directed against the extractable nuclear antigen U1-RNP (see Chapter 18.11.2 for discussion).
Pleuropulmonary complications occur in 20 to 85 per cent of patients, most commonly diffuse lung disease. Pulmonary involvement, investigations, and treatment are as for the individual rheumatological disease.

Treatment of lung disease in particular autoimmune rheumatic disorders

Treatment is directed at the type of lung disease, irrespective of the rheumatological disease in which it is found (Table 2). Opportunistic infection and pulmonary side-effects caused by treatment of the rheumatological disease can mimic pulmonary complications of the rheumatological disease and must be excluded. Similarly, carcinoma and lymphoma are complications of diffuse lung disease in rheumatological lung disease that must be differentiated.

Further reading

General reading

du Bois RM, Stirling RG (1999). Connective tissue disorders. In: Albert R, Spiro S, Jett J, eds. *Principles of respiratory medicine*, pp. 53.2–53.14. Mosby International, Barcelona.

du Bois RM, Wells AU (2000). Pulmonary involvement of connective tissue disease. In: Murray JF and Nadel JA, eds. *Respiratory medicine*, pp. 1691–715. WB Saunders, Philadelphia.

Specific reading

Cervera R, Khamashta MA, Font J, *et al*. (1993). Systemic lupus erythematosus: clinical and immunologic patterns of disease expression in a cohort of 1000 patients. The European working party on

systemic lupus erythematosus. *Medicine* **72**, 113–24.

Friedman AW, Targoff IN, Arnett FC (1996). Interstitial lung disease with autoantibodies against aminoacyl-tRNA synthetases in the absence of clinically apparent myositis. *Seminars in Arthritis and Rheumatism* **26**, 459–67.

Haupt HM, Moore GW, Hutchins G (1981). The lung in systemic lupus erythematosus. Analysis of the pathologic changes in 120 patients. *American Journal of Medicine* **71**, 791–8.

Hyland RH, Gordon DA, Broder I, *et al*. (1983). A systematic controlled study of pulmonary abnormalities in rheumatoid arthritis. *Journal of Rheumatology* **10**, 395–405.

Marie I, Hatron PY, Hachulla E, Wallaert B, Michon-Pasturel U, Devulder B (1998). Pulmonary involvement in polymyositis and dermatomyositis. *Journal of Rheumatology* **25**, 1336–43.

Papiris SA, Maniati M, Constantopoulos SH, Roussos C, Moutsopoulos HM, Skopouli FN (1999). Lung involvement in primary Sjögren's syndrome is mainly related to the small airways disease. *Annals of the Rheumatic Diseases* **58**, 61–4.

Tanoue LT (1998). Pulmonary manifestations of rheumatoid arthritis. *Clinics in Chest Medicine* **19**, 667–85.

Wells AU, Cullinan P, Hansell DM, *et al*. (1994). Fibrosing alveolitis associated with systemic sclerosis has a better prognosis than lone cryptogenic fibrosing alveolitis. *American Journal of Respiratory and Critical Care Medicine* **149**, 1583–90.

17.11.5 The lung in vasculitis

R. M. du Bois

Introduction

The nomenclature of vasculitis has been confused. This is because there is overlap between clinical and histopathological features in a group of disorders of unknown aetiology. It is useful to subdivide pulmonary vasculitides into primary systemic or secondary, and to differentiate them from non-vasculitic disorders that can affect the pulmonary circulation (Table 1). The secondary and non-vasculitic diseases are discussed in other chapters: Table 2 summarizes the primary vasculitides, indicating those in which the lung is involved.

Clinical manifestations of pulmonary vasculitis

Lung involvement in vasculitic disease can manifest as:

1. diffuse alveolar haemorrhage or
2. isolated gas transfer for carbon monoxide (*DL*CO) deficit with or without pulmonary hypertension.

Other features of the underlying or associated disease may be present, and the pulmonary disorder may present as part of a pulmonary–renal syndrome, of which Goodpasture's disease is the best-known example.

Diffuse alveolar haemorrhage

The presenting features of diffuse alveolar haemorrhage often mimic infective pneumonia. The patient may give a history of fever, weight loss, and other systemic symptoms pointing towards the underlying diagnosis, but cough, breathlessness, and clinical signs suggestive of pneumonia are the main respiratory features. Indeed, if pneumonia is suspected on clinical and radiological grounds, but no organism is found, then other explanations need to be considered, including alveolar haemorrhage. A history of previous haemoptysis is helpful but not invariable: the first manifestation of diffuse alveolar haemorrhage can be acute.

Chest radiography shows consolidation, typically resolving within a matter of days, quite unlike the pattern seen in infective pneumonia, a point that can be helpful retrospectively if alveolar haemorrhage is not suspected at presentation. High-resolution computed tomography may reveal more subtle ground-glass partial alveolar filling. An acute fall in haemoglobin can occur; chronic iron-deficient anaemia suggests chronicity.

In the absence of a history of haemoptysis, confirmation of alveolar haemorrhage can be obtained by bronchoalveolar lavage that will reveal frank blood-staining in sequential lavage in the acute presentation, or the presence of numerous macrophages containing iron, identified by Perl's stain, in more chronic disease. The gas transfer corrected for alveolar volume (*K*CO) is elevated in acute haemorrhage, but where this has occurred on a background of small, previously undetected, haemorrhage, the interstitial fibrosis consequent upon this chronic haemorrhage may have reduced gas transfer such that an elevation above normal is not observed. If diagnosis remains elusive, surgical biopsy may be necessary to make the diagnosis and reveal the aetiology if vasculitis is responsible.

If a diagnosis of alveolar haemorrhage is suspected, investigations shown in Table 3 should be considered. Diffuse pulmonary haemorrhage occurring without identifiable cause or association is known as idiopathic pulmonary haemosiderosis (see Chapter 17.11.8). Pulmonary vasculitis can be primary or secondary.

Treatment is for the underlying cause. Prognosis can be poor in certain situations, most notably alveolar haemorrhage in rheumatological disease.

Isolated gas transfer deficit with or without pulmonary hypertension

A patient with an abnormality of the pulmonary vasculature, but in the absence of alveolar haemorrhage, will present with breathlessness on exertion. Clinical examination of the respiratory system will be normal, as will routine lung imaging. Lung function tests will show relatively well-preserved lung volumes, but with a reduced gas transfer factor (*DL*CO and *K*CO). Exercise testing will demonstrate a fall in PaO_2, a drop in oxygen saturation, and widening of the alveolar–arterial (A–a) gradient, with a high ventilatory reserve. In severe pulmonary vascular disease, clinical features of pulmonary hypertension may be observed. This is most easily confirmed by echocardiography, particularly if tricuspid regurgitation is present, when a Doppler estimate of pulmonary artery pressures can be obtained. Right heart catheterization may be required to confirm the diagnosis. Other causes of vascular compartment abnormality such as systemic sclerosis, primary pulmonary hypertension, or coagulation abnormalities must be excluded.

Specific disorders

Churg–Strauss syndrome

This disease, first described by Churg and Strauss in 1951, is defined by the presence of numerous eosinophils and granulomatous inflammation in the respiratory tract, together with a necrotizing vasculitis affecting small to medium-sized vessels, and associated with asthma and eosinophilia. It is a rare disorder, mainly affecting adults around the age of 40 years, but has been reported in individuals from ages 7 to 74 years. In combined series there is roughly a 2:1 prevalence for males to females. There is little information about geographical variation.

Aetiology and pathogenesis

This is not known, but thought to be an eosinophilic granulomatous response to a foreign antigen, akin to the eosinophilic granulomatosis seen in schistosomiasis. In this regard, immunological stimuli in the form of vaccination or immunotherapy have been reported as a trigger for the disease, but the pauci-immune nature of the histopathology raises doubts about this trigger mechanism.

More recently the introduction of antileukotriene therapy for asthma has resulted in an apparent increased incidence of Churg–Strauss syndrome. Two possible reasons have been advanced for this finding. First, that the drug has been a trigger for Churg–Strauss syndrome in predisposed individuals. Second, that the withdrawal or reduction of corticosteroids that has been possible with the introduction of non-steroid therapy has 'unmasked' the underlying condition. However, some

individuals who have never been on corticosteroids have developed Churg–Strauss syndrome with the introduction of one of these drugs.

The relationship of antineutrophil cytoplasmic antibodies (**ANCA**) to disease pathogenesis is unclear. There have been studies demonstrating an up-regulation of the receptor for ANCA on the surface of neutrophils at disease sites, and ANCA can also interact with endothelial cells, causing injury and coagulation.

Clinical

Two sets of criteria have been used to define the disease clinically: Lanham's criteria and the criteria of the American College of Rheumatology. In addition to systemic features such as fever and weight loss, Lanham defined the disease as requiring the presence of the following:

1. asthma;
2. eosinophilia greater than 1.5×10^9/l in the peripheral blood; and
3. evidence of systemic vasculitis in two or more non-lung organs.

The American College of Rheumatology defined Churg–Strauss syndrome as requiring four of the following six criteria:

1. the presence of asthma;
2. eosinophilia greater than 10 per cent in the peripheral blood;
3. evidence of a neuropathy in a vasculitic pattern (e.g. mononeuritis multiplex);
4. transient pulmonary infiltrates;
5. a history of sinus disease; and
6. evidence of extravascular eosinophilia on biopsy.

The disease is considered to evolve in three phases, the first (prodromal) consisting of a long history of rhinitis with nasal polyps that slowly progresses (often over years) to late-onset asthma, which is frequently difficult to treat. The second phase is of increasing peripheral blood and tissue eosinophilia that can wax and wane. The third and final phase is the manifestation of systemic vasculitis. This pattern of disease evolution has been shown in studies to be greater than 95 per cent specific and sensitive for Churg–Strauss syndrome. Table 4 illustrates the major pulmonary manifestations.

Other organ involvement

Skin lesions

These are seen in about 60 per cent of patients, generally manifesting as palpable purpura or subcutaneous nodules. Skin infarcts may also been seen.

Cardiac involvement

The heart may be involved diffusely, producing congestive cardiac failure or restrictive cardiomyopathy. Coronary artery inflammation may also be present, as can pericardial effusion. Cardiac disease is the most common cause of death in Churg–Strauss syndrome.

Renal disease

This is much less common than in Wegener's granulomatosis or microscopic polyangiitis, but the histopathology is very similar, consisting of a focal segmental necrotizing glomerulonephritis. Renal disease is generally mild, but endstage renal failure is reported.

Central nervous system

Mononeuritis multiplex is the most common manifestation, occurring in up to 75 per cent of patients. Cranial nerve involvement is less common, but cerebrovascular disease may occur.

Gastrointestinal involvement

Vasculitis of the mesenteric vessels may produce bowel abnormality, including perforation. Less commonly, eosinophilic infiltration may cause obstruction.

Musculoskeletal system

Arthritis is relatively common, as are myalgias.

Pathology

Lung biopsy shows a necrotizing angiitis, granulomas, and tissue eosinophilia. Giant cells and fibrinoid necrosis are present. It is not unusual to find only some of these features in a single biopsy and there can be overlap with the histopathological appearances of Wegener's granulomatosis and microscopic polyangiitis.

Investigations

Chest radiography shows areas of infiltration in the lung in up to 77 per cent of patients; pleural disease is seen in up to 50 per cent, and pericardial disease may occur with effusions sufficiently severe to cause tamponade. Computed tomography adds greater resolution to the imaging: nodules are uncommon and the main abnormalities are areas of ground-glass infiltrate, particularly if there is alveolar haemorrhage, which may also produce areas of consolidation.

There is a peripheral blood eosinophilia, matched by a marked eosinophilia on bronchoalveolar lavage. Antineutrophil cytoplasmic antibodies are found in roughly two-thirds of cases and are usually of the p-ANCA pattern.

Treatment

Initial treatment depends upon severity of presentation and the organs involved. In isolated pulmonary disease, the first-line treatment of choice is prednisolone at 1 mg/kg per day (up to a total of 60 mg/day) orally or, in more urgent situations such as alveolar haemorrhage, up to 1 g of methylprednisolone intravenously on each of three successive days. Response is usually good.

In more severe disease, particularly life-threatening situations when mechanical organ support is required, cyclophosphamide is added. This is given either orally at 2 mg/kg per day up to a maximum usually of 150 mg/day, or intravenously as a bolus of 600 mg/m^2 as frequently as weekly in more severe disease. It is thought that intravenous administration of cyclophosphamide is less likely to provoke the major complica-tion of cyclophosphamide therapy—haemorrhagic cystitis and subsequent malignancy—that are seen with prolonged oral use of this drug. Other immunosuppressive approaches have been tried, including azathioprine, methotrexate, and mycophenolate mofetil: the evidence in support of these treatments is less firm.

Because of the long-term side-effect profile of cyclophosphamide, especially bladder tumours, arguably the best approach to therapy would be to induce remission with either prednisolone alone or, in more severe presentations, prednisolone and pulsed intravenous cyclophosphamide, before maintaining that remission with prednisolone and one of the other immunosuppressant drugs. There is strong evidence that plasma exchange has no place in this or any other pauci-immune pulmonary vasculitis. Prophylactic co-trimoxazole (trimethoprim 160 mg/sulphamethoxazole 800 mg) three times a week is often used to reduce the risk of *Pneumocystis carinii* opportunistic infection.

Prognosis

Prognosis is generally good for those with isolated intrathoracic disease, but worsens with two or more extrapulmonary complications, particularly proteinuria greater than 1 g/day, renal insufficiency (creatinine greater than 140 µmol/l), cardiomyopathy, gastrointestinal disease, or central nervous system involvement. Five-year mortality for two or more of these complicating factors is 46 per cent, compared with 26 per cent with one and 12 per cent with none of the extrapulmonary features: recognition of this is useful in helping to determine initial therapy. Main causes of death are cardiac disease followed by renal failure, cerebrovascular involvement, and gastrointestinal disease. Lung disease accounts for 10 per cent of deaths.

Wegener's granulomatosis

The main description of Wegener's granulomatosis occurs elsewhere (see Chapter 20.10.3). It is a granulomatous inflammation due to necrotizing vasculitis affecting small to medium-sized vessels, typically involving the respiratory tract (both upper and lower) and often causing necrotizing glomerulonephritis.

Incidence

Wegener's granulomatosis is the third most common systemic vasculitis after giant-cell arteritis and rheumatoid arthritis. In 90 per cent of cases involvement of the upper and/or lower respiratory tract is the first manifestation, and in various series up to 85 per cent of patients have lung involvement at some stage of the disease. Males and females are affected equally.

Pulmonary presentation

In 34 per cent of cases lung involvement is asymptomatic. The main manifestations in the lung are (see Table 4):

1. one or more nodules which can cavitate;
2. localized or diffuse infiltrates (pleural effusions may be seen);
3. alveolar haemorrhage that may be part of a pulmonary–renal syndrome; and
4. large and small airway disease.

There is often a history of rhinosinusitis, often pre-dating other manifestations but seen at some stage of the disease in up to 92 per cent of cases. Presentations of lower respiratory tract involvement include the non-specific symptom of cough, with or without purulent sputum and haemoptysis. There is often fever, weight loss, and the systemic features of non-respiratory disease.

Chest radiography will show the presence of nodules or consolidation, the latter possibly due to inflammatory infiltrate or alveolar haemorrhage. Pleural effusion may be present. High-resolution computed tomography will show these features with greater resolution and is the better modality to identify the number of nodules and determine whether they are cavitating. It is also very helpful in identifying tracheal and bronchial abnormality, including frank bronchiectasis. Characteristic airway involvement includes subglottic stenosis and stenosis of the main airways. Subtle parenchymal fibrosis may be seen as a reticular subpleural process mimicking fibrosing alveolitis.

Fibre-optic bronchoscopy may confirm airway disease, showing evidence of tracheobronchitis including ulceration and 'cobble-stoning' of the mucosa with, in more chronic situations, cicatricial narrowing of the airways with scar tissue. Bronchoalveolar lavage returns an excess of neutrophils and usually of eosinophils (with diffuse infiltrates) or lymphocytes (more interstitial disease), but the most important use of this procedure is to exclude an infective complication of the disease or its treatment, or alveolar haemorrhage. Transbronchial biopsy can be hazardous, resulting in major haemorrhage, and should be avoided. If lung biopsy is necessary to confirm the diagnosis, a surgical approach should be used, but usually it is possible to obtain a histological diagnosis from other tissue, particularly skin or kidney.

Haematological and biochemical investigations reflect the inflammatory process. The presence of c-ANCA is highly specific for this disease (up to 95 per cent in some series) and can be diagnostic in the right clinical context.

Pathology

The histopathological appearances of Wegener's granulomatosis include a necrotizing vasculitis, granuloma formation, necrosis, and surrounding inflammation with a combination of acute and chronic inflammatory cells. This and other non-pulmonary features are discussed in Chapter 20.10.3.

Treatment

First-line treatment is a combination of prednisolone at 1 mg/kg per day (usually up to a maximum of 60 mg), together with cyclophosphamide given either orally up to 150 mg/day or intravenously at 600 mg/m^{2}at intervals dependent on disease severity. The hazards of cyclophosphamide given long-term and the alternative treatments are as for Churg–Strauss syndrome (see above). Co-trimoxazole therapy has been efficacious in this disease for localized upper respiratory tract or minor lower respiratory tract disease, but is not recommended for more systemic disease, although it appears to have a role in the maintenance of remission.

Prognosis

Since the introduction of combination immunosuppression the mortality from this disease has improved from a mean survival of 5 months to a 75 per cent complete remission. However, relapse occurs in up to 50 per cent of cases: long-term follow-up is needed.

Microscopic polyangiitis

The main description of microscopic polyangiitis occurs elsewhere (see Chapter 20.10.3). It is a necrotizing vasculitis that affects small vessels, with few or no immune complex deposits. Lung disease is said to occur in between 34 and 55 per cent of cases.

Pulmonary presentation (see Table 4)

The major presentation in the lung is diffuse alveolar haemorrhage, which can have a poor prognosis. Pulmonary capillaritis may be associated with evidence of disease outside the lung, particularly necrotizing glomerulonephritis, mononeuritis multiplex, and skin lesions. It may be difficult to distinguish microscopic polyangiitis from Wegener's granulomatosis without a biopsy, but if this is performed then the key issue is whether or not granulomas are present. Granulomas are not found in microscopic polyangiitis, whilst they are characteristic of Wegener's granulomatosis. Renal biopsies can be identical in the two conditions. Microscopic polyangiitis also needs to be distinguished from polyarteritis nodosa that, by definition, only affects arteries, rarely arterioles, and never small vessels. Renal vasculitis with microaneurysm formation occurs in polyarteritis nodosa but not microscopic polyangiitis, and diffuse alveolar haemorrhage does not occur in polyarteritis nodosa.

Other diseases

Other primary systemic vasculitides may occasionally present with respiratory features.

Takayasu's arteritis

This arteritis affects predominantly the aorta and its main branches but can involve the pulmonary arteries in up to 50 per cent of patients, presenting with features of pulmonary vascular occlusion.

Giant-cell arteritis

There is rarely objective evidence of lung involvement, but 25 per cent of patients with giant-cell arteritis have associated cough, hoarseness, and sore throat at presentation.

The other systemic vasculitides that feature in the Chapel Hill International consensus nomenclature, but which rarely if ever present with lung disease, are Henoch–Schönlein purpura and essential cryoglobulinaemia.

Behçet's disease

This occurs predominantly in Mediterranean countries and can produce pulmonary vascular inflammation affecting all sizes of vessels and resulting in pulmonary arterial aneurysms, arterial and venous thrombosis, pulmonary infarcts, and pulmonary haemorrhage. It is crucial to differentiate haemorrhage from thrombosis because of the treatment implications.

Pulmonary veno-occlusive disease

This is a disorder of unknown cause that manifests with progressive occlusion of the post-capillary venules, resulting in features similar to those of pulmonary oedema. There is no known effective treatment. Differentiation from cardiogenic causes of raised pulmonary venous pressure must be made.

Lymphomatoid granulomatosis

This has been included historically within the category of pulmonary vasculitis but is now believed to be a lymphoproliferative disease.

Further reading

Conron M, Beynon HLC (2000). Churg–Strauss syndrome. In: du Bois RM, Tattersfield A, eds. *Thorax Rare Disease Series. Thorax*, **55**, 870–7.

Fauci AS *et al.* (1983). Wegener's granulomatosis: prospective clinical and therapeutic experience with 85 patients for 21 years. *Annals of Internal Medicine* **98**, 76–85.

Guillevin L *et al*. (1996). Prognostic factors in polyarteritis nodosa and Churg–Strauss syndrome. A prospective study in 342 patients. *Medicine* **75**, 17–28.

Jennette JC *et al*. (1994). Nomenclature of systemic vasculitides. Proposal of an International consensus conference. *Arthritis and Rheumatism* **37**, 187–92.

Langford CA, Hoffman GS (1999). Wegener's granulomatosis. In: du Bois RM, Tattersfield A, eds. *Thorax Rare Disease Series. Thorax* **54**, 629–37.

Lanham JG *et al*. (1984). Systemic vasculitis with asthma and eosinophilia: the clinical approach to the Churg–Strauss syndrome. *Medicine (Baltimore)* **63**, 65–81.

Lhote F, Guillevin L (1998). Polyarteritis nodosa, microscopic polyangiitis and Churg–Strauss syndrome. *Seminars in Respiratory and Critical Care Medicine* **19**, 27–46.

Specks U (1998). Pulmonary vasculitis. In: Schwarz MI, King TE Jr, eds. *Interstitial lung disease*, pp 507–34. BC Dekker, Hamilton, Canada.

17.11.6

Sarcoidosis

Robert P. Baughman and Elyse E. Lower

The aetiology of sarcoidosis is unknown. It is characterized by the presence of non-caseating granulomas in at least two organs. Patients may experience a variable course ranging from spontaneous remission to severe chronic disease and occasionally death.

Historical introduction

The disease was first recognized in 1869 by Jonathan Hutchinson, who treated a man with skin lesions that appeared unrelated to tuberculosis or any other process that he had encountered. Initially, most reports described patients with skin lesions. Since the disease is often self-limiting, pathological information was scarce. Schaumann in Sweden was one of the first to recognize the multiorgan nature of the disease, with the common pathological feature of granulomas. His original thesis was written in 1914, but not published until 1936.

After the Second World War, the use of routine screening radiography identified asymptomatic patients with abnormal chest radiographs. Lofgren described a group of patients with erythema nodosum, uveitis, and hilar adenopathy. Others began to appreciate the unique aspects of sarcoidosis compared with tuberculosis. Interestingly, as tuberculosis becomes less frequent in a country, sarcoidosis becomes more obvious. The observation that sarcoidosis is a disease of industrial nations and temperate climates may reflect this phenomenon. However, several groups have reported series of patients with sarcoidosis in India, Thailand, and China.

Pulmonary sarcoidosis can be evaluated by chest radiography. Scadding in Scotland and Wurm in Germany described a staging system based on the appearances seen on the chest radiograph, and this became a useful method of describing the extent of lung involvement. It also has prognostic significance and has been used for 40 years as the method of choice for characterizing lung involvement with sarcoidosis.

Newer radiological techniques have been evaluated in sarcoidosis. The chest CT scan provides more detailed information regarding adenopathy, but has not replaced the prognostic information available from the chest radiograph. A gallium scan will reveal increased uptake in areas of inflammation, such as lung and mediastinum. Magnetic resonance imaging (**MRI**) and positron emission tomography (PET) have brought new methods for evaluating extrapulmonary disease.

Bronchoalveolar lavage has provided a new method for sampling lower airway secretions, and it soon became apparent that the lavage findings from patients with sarcoidosis are distinctly different from normal subjects, providing insights into the true inflammatory response of the lung.

The definition of sarcoidosis had been the subject of much debate. An international sarcoidosis group began meeting in 1958. Currently this group, the World Association of Sarcoidosis and Other Granulomatous Diseases (WASOG), is chaired by D. Geraint James and tries to provide order from what was chaos. By consensus, the definition of sarcoidosis has been established and further refined over the years. The group now meets every 2 years and stresses the international aspects of the disease.

Aetiology

The cause of sarcoidosis remains obscure. One hypothesis is that sarcoidosis is an inflammatory response to an environmental agent (including infectious) which occurs in a susceptible host. Susceptibility is determined on the basis of genetic predisposition.

Several potential infectious agents have been proposed as causes of sarcoidosis. The granulomatous reaction reminds many of tuberculosis, and much effort has been expended trying to identify a mycobacterial cause. Several studies using polymerase chain reaction (PCR) and similar molecular biological techniques have been employed, but there is still no convincing evidence that *Mycobacterium tuberculosis* causes most cases of sarcoidosis. It may lead to an occasional case of sarcoid-like reaction. Other mycobacteria have been identified in some cases. Cell wall-deficient mycobacteria have been grown from the blood of patients with sarcoidosis. However, a recently completed control trial failed to demonstrate a difference in the incidence of cell wall-deficient mycobacteria between those with sarcoidosis and controls.

Epidemiology

Sarcoidosis is a worldwide disease. It has been reported to have a higher prevalence in Scandinavian countries as well as in Ireland. Table 1 summarizes the relative frequency of sarcoidosis per 100 000 people around the world. In the United States, a higher incidence of sarcoidosis has been reported in African-American people.

The disease presentation varies in different parts of the world. Table 1 also lists some of the more frequent patterns seen with various ethnic groups. For example, lupus pernio is common among African-American and West Indian people who have migrated to Great Britain, while erythema nodosum is common among Scandinavians. Cardiac disease has been reported at a higher frequency in Japanese patients with sarcoidosis than for other groups. HLA studies have suggested that there may be certain genetic patterns associated with particular manifestations of the disease.

Occupational exposures have been studied as a possible cause of sarcoidosis. Beryllium is a metal used in certain industries (ceramics, nuclear processing) which can cause a reaction in the lung and skin indistinguishable from sarcoidosis. Besides clinical history, the distinguishing feature about berylliosis is that lymphocytes are stimulated to replicate when exposed to beryllium salts. The lymphocyte stimulation test of blood, or the more sensitive bronchoalveolar lavage, is a reliable way of detecting which patients are reacting to beryllium.

There are other occupations associated with sarcoidosis. These include health-care workers, who have been found to be at increased risk. In one study, the odds ratio of a health-care worker acquiring sarcoidosis was 64 times the rate of the general population. Such a high rate could be due to work-related exposures (medications, disinfectants), or be due to an infectious agent. Clusters of sarcoidosis have been reported among firemen and aircraft-carrier personnel.

Obviously, none of these occupational exposures encompass all cases of sarcoidosis. One possible cause of sarcoidosis is that it is the common reaction to several

possible agents.

Pathogenesis

Sarcoidosis is defined by its immunological reaction, the granuloma. Original immunological studies stressed a lack of systemic immune response by the patient with sarcoidosis. This includes anergy, which is a common feature of active sarcoidosis. A reduction in circulating leucocytes, especially lymphocytes, is an important feature of the disease.

In the 1970s, new techniques helped us understand sarcoidosis better. The most important tool introduced at the time was bronchoalveolar lavage, which provided a sample of the inflammatory cells in the lower respiratory tract. In normal lavage fluid, alveolar macrophages are the usual resident inflammatory cell retrieved; lymphocytes and neutrophils are found much less frequently. In lavage fluid from patients with active sarcoidosis, the preponderance of T lymphocytes is usually increased. These lymphocytes are often T-helper/inducer cells (CD4+), and the ratio of CD4 to CD8 lymphocytes is increased from that normally found in the blood, often to greater than 3.5.

The CD4 lymphocyte is a crucial cell in cell-mediated immunity. The CD4 lymphocytes are activated and release several cytokines, including interleukin 2 (**IL-2**) and g-interferon. The T lymphocyte can mount either a T_{H1} or T_{H2} response. The T_{H1} response is associated with granuloma formation, while T_{H2} is associated with an eosinophilic response and fibrosis. The initial response of sarcoidosis follows a T_{H1}pattern. The lymphocytes release IL-2 spontaneously, and g-interferon is released by both lymphocytes and macrophages. An increase in IL-12 and lower levels of IL-10 have also recently been described, consistent with a T_{H1} response.

The alveolar macrophage is also activated in sarcoidosis, and increased levels of IL-1, tumour necrosis factor, and oxygen free radicals are released by macrophages retrieved by bronchoalveolar lavage. During the evolution of the disease, the macrophages and other resident cells may release factors associated with fibrosis. For example, IL-8 is found in the bronchoalveolar lavage fluid of some patients with sarcoidosis, and this increase in IL-8 is associated with the finding of fibrosis in the lung.

The resolution of sarcoidosis has also been studied with serial lavages. The T lymphocytes remain elevated for some time, but the proportion of CD4 to CD8 decreases to the ratio in blood (0.8 to 2.2). The amount of cytokines released also decreases. This return to normal of the inflammatory response has been shown to occur during treatment of sarcoidosis with corticosteroids or methotrexate.

Clinical features

Patients with sarcoidosis may have a variety of presentations. Commonly affected organs include the lung, skin, and eyes. Less commonly, the liver, heart, and brain are affected by the disease. Individual organ involvement by sarcoidosis can be proved by a biopsy showing non-caseating granuloma. Presumed organ involvement is assumed if certain criteria are met. Table 2 lists some of the criteria suggested for definite or probable organ involvement for some of the more commonly affected organs in sarcoidosis.

Lung

Respiratory involvement has been described in more than 90 per cent of patients. The lung involvement includes both the lymph nodes and the lung parenchyma. Scadding and Wurm independently described four stages of the chest radiograph: stage 1 is hilar adenopathy alone (Fig. 1), stage 2 is adenopathy and parenchymal disease, stage 3 is parenchymal disease alone (Fig. 2), and stage 4 is fibrosis. The interstitial disease usually has a diffuse reticulonodular appearance, but confluent patches of disease (alveolar sarcoidosis) have been described. Fibrotic changes due to sarcoidosis are usually in the upper lobe, with retraction. The staging system has proved useful in standardizing reports of pulmonary level of involvement. It has also proved a useful prognostic measure. Patients with stage 1 disease have a 90 per cent rate of resolution within 2 to 3 years, while patients with stage 3 disease possess only a 30 per cent chance of resolution. However, it does not predict the degree of extrapulmonary disease. The choice of the term 'stage' is therefore unfortunate. However, it is so standard that it will not be easily replaced.

Fig. 1 Chest radiograph showing stage 1 sarcoid.

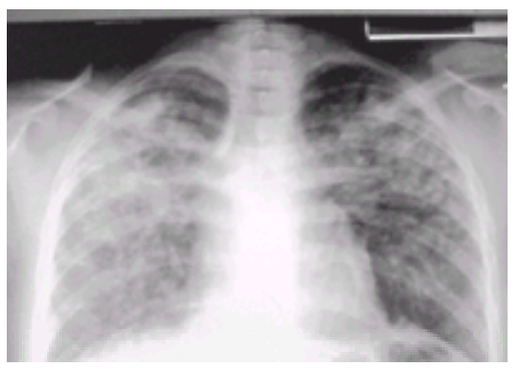

Fig. 2 Chest radiograph showing stage 3 sarcoid.

Table 3 lists the other diseases to be considered in the differential diagnosis based on the chest radiograph pattern. The presence of mediastinal adenopathy alone (stage 1 disease) is certainly consistent with lymphoma or metastatic cancer. It has been pointed out that symmetrical bilateral adenopathy with right paratracheal adenopathy in an asymptomatic individual is almost always sarcoidosis. Asymmetrical adenopathy raises the question of lymphoma, and a tissue diagnosis is usually required. For patients with diffuse infiltrates, adenopathy points one toward sarcoidosis. However, several other conditions may have some adenopathy, including hypersensitivity pneumonitis and idiopathic pulmonary fibrosis. The larger the adenopathy, the more likely is sarcoidosis.

The use of the CT scan has changed our evaluation of many interstitial lung diseases. In sarcoidosis, peribronchial thickening is often seen in the upper lobe. Adenopathy is usually seen in sarcoidosis, making the staging system only applicable for plain radiographs. The CT scan may identify adenopathy in a patient with possible extrapulmonary sarcoidosis. This may help in deciding where to proceed with a tissue diagnosis (brain biopsy or mediastinoscopy).

Pulmonary function studies in patients with sarcoidosis classically demonstrate a restrictive pattern, with reduction of lung volumes. The transfer factor is usually reduced out of proportion to the loss of lung volume, as one would expect in an interstitial lung disease. In advanced cases, the oxygen level will be reduced,

especially during exercise. Obstructive disease can also occur in sarcoidosis. This can be due to airway involvement by the sarcoidosis or associated with cough, a common complaint in the condition.

Skin

The skin is the second most commonly affected organ in sarcoidosis. There are six major manifestations. Hyperpigmentation, hypopigmentation, and keloid reaction may demonstrate granulomas on biopsy. However, their appearance is not always specific. Waxy, maculopapular lesions, which occur on the extremities, back, and face, are usually raised, with the majority less than 2 cm in diameter. When these occur on the face, especially on the cheeks and nose, they are called lupus pernio. Erythema nodosum—red nodular lesions on the extremities—usually involves the legs. The constellation of erythema nodosum, arthritis (in the ankles), hilar adenopathy, and uveitis is referred to as Lofgren's syndrome and is a diagnostic manifestation of sarcoidosis. It is associated with a good prognosis. Interestingly, the skin lesions from erythema nodosum do not contain granulomas, but are felt to be due to circulating immune complexes from the disease.

Eye

The eye can be affected in more than 20 per cent of patients with sarcoidosis. The most common findings are uveitis and lacrimal gland involvement. Anterior uveitis is often self-limiting, and can be treated topically; however, posterior uveitis is a more chronic form of the disease and may require injections of corticosteroids or systemic therapy. Sicca (dry eyes) and glaucoma are long-term complications which are encountered in patients often years after other sarcoidosis symptoms have resolved. They are consequences of the fibrotic changes in the lacrimal glands and eye. They do not respond to anti-inflammatory therapy. Optic nerve involvement can be seen with sarcoidosis, with idiopathic disease and multiple sclerosis being the other major causes of this sight-threatening complication. Retinal disease has also been reported. Fortunately, blindness from sarcoidosis is rare, and usually a consequence of untreated uveitis, retinitis, or optic neuritis.

Neurological

Neurological disease from sarcoidosis includes cranial nerve, central nervous system, and peripheral nerve involvement. Bell's palsy (seventh cranial nerve) is a common complaint in neurosarcoidosis. Central nervous system lesions can lead to a lymphocytic meningitis. Hypothalamic involvement is a characteristic finding, with diabetes insipidus as a resulting complaint. The use of contrast-enhanced magnetic resonance imaging is the most sensitive method for detecting central nervous system disease. The lumbar puncture is complementary, with increased protein and lymphocytes often seen in active disease. Detection of angiotensin-converting enzyme in the spinal fluid is suggestive but not diagnostic of neurosarcoidosis.

Other manifestations

Liver and spleen involvement may be found in over half of patients with sarcoidosis. However, symptomatic disease occurs in less than 10 per cent of patients. Often, elevated liver function tests (especially the alkaline phosphatase and g-glutamyl transferase) are seen, suggesting an obstructive pattern. Hyperbilirubinaemia is relatively rare, but implies extensive disease and is usually an indication for therapy. Massive splenomegaly can occur, and occasionally splenectomy is performed to avoid rupture.

Hypercalcaemia and hypercalcinuria are seen with sarcoidosis. The mechanism is related to the effect of the granuloma on vitamin D_3. The granuloma itself converts the vitamin D_3 to the biologically active form 1,25-D_3. This form of the vitamin has immunological activity as well as enhancing calcium absorption from the gastrointestinal tract. In some patients with sarcoidosis, the 1,25-D_3 can leak into the bloodstream and produce a systemic effect. Increased sunlight exposure also increases the levels of 1,25-D_3. In America, hypercalcaemia is far more common in Caucasian than African-American individuals. Because of the effect of increased calcium absorption, urolithiasis may also be seen in patients with sarcoidosis. Recently, it has been appreciated as a marker for chronic disease.

A less common, but serious complication of sarcoidosis is cardiac involvement. Direct involvement of the heart can lead to arrhythmias such as heart block and ventricular ectopy. This can lead to sudden death. If the problem is recognized, the use of an implanted defibrillator may reduce this risk. Cardiomyopathy is also seen, and cardiac sarcoidosis should be considered in a young patient who presents with unexpected heart failure. Endomyocardial biopsy rarely makes a diagnosis, since the granulomas are patchy. The technetium scan showing non-segmental fixed defects is the most sensitive test. Gallium uptake of the heart is more specific than a thallium scan.

Sarcoidosis granulomas can involve virtually any organ of the body. Rare manifestations include bone cysts, usually in the distal portion of the fingers, sinus invasion, pleural disease, breast disease, and ovarian or testicular masses.

The multiorgan involvement of sarcoidosis distinguishes it from other diseases. Lymphoma and tuberculosis are two diseases often considered in the differential diagnosis of patients with possible sarcoidosis. Table 4 summarizes the common features in all three of these diseases and points out those features that can be used to separate them.

Pathology

The non-caseating granuloma is the characteristic pathological feature of sarcoidosis (Fig. 3). The centre of the granuloma includes macrophages and giant cells which are of the Langerhans type and can contain over 10 nuclei. This core of cells is surrounded by two rings of lymphocytes, the larger inner component of CD4 helper cells, while the outer ring can be CD8 suppressor cells. The granulomas tend to be well formed, and in lung biopsies are often well demarcated from normal tissue. The central area will occasionally contain a Schaumann body, formed of crystallized material (calcium phosphate). This is different in appearance from foreign bodies or caseation, which can be seen in other granulomatous diseases. Occasionally the granuloma will have a necrotic area, but the majority of the granulomas do not.

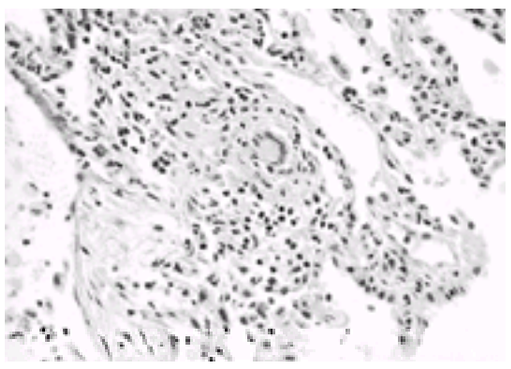

Fig. 3 Transbronchial biopsy demonstrating a non-caseating granuloma in the centre of the field. A multinuclear giant cell is seen within the granuloma. (Haematoxylin and eosin stain, original magnification ×40.)

Because the cause of sarcoidosis is unknown, one can never be absolutely confident of the diagnosis, which is always one of exclusion. However, the finding of non-caseating granulomas in two or more organs is considered diagnostic. Cultures and special stains for tuberculosis and deep-seated fungal infections should be taken to rule out infection as the cause of granulomas. Close examination should also be made for foreign bodies and malignancy, both of which can lead to a granulomatous reaction.

Laboratory diagnosis

Serum angiotensin-converting enzyme (ACE) levels

In 1976, Lieberman reported that ACE level was elevated in the blood of some patients with sarcoidosis. Mild elevations have also been reported in diabetes mellitus and osteoarthritis. High levels have been detected in patients with infectious granulomatous diseases such as tuberculosis, histoplasmosis, and coccidioidomycosis; but also in Gaucher's disease, leprosy, and hyperthyroidism. Because ACE is measured using a biological assay, patients on ACE inhibitors may have low levels.

In those with granulomatous disease, the source of the ACE is not the lung (the usual source), but the granulomas themselves. It appears that ACE has immunoregulatory properties.

Determining the significance of the ACE level in sarcoidosis can be difficult for a variety of reasons. Sixty per cent of patients with acute disease will have elevated values, whilst only 10 per cent of patients with disease for more than 2 years will continue to have an elevated level. The ACE level will decrease in response to treatment or disease resolution, and it has therefore been proposed as a marker for disease activity. However, corticosteroids independently suppress ACE levels, and reducing the dose of corticosteroids may lead to a rise in ACE level without a clinical worsening of disease. Furthermore, since ACE levels are elevated in only a small proportion of those with chronic disease, decreases observed with serial measurement may simply reflect the long-standing nature of the disease.

There is a genetic polymorphism for ACE, with an insertion (I) or deletion (D) of a genomic DNA fragment. There appears to be no difference in the distribution of the alleles in patients with sarcoidosis compared with the general population. Interestingly, ACE levels are higher in DD patients, and this needs to be considered when one is measuring the serum ACE level. There is some evidence that the DD allele is associated with a worse prognosis in patients with sarcoidosis, and it is interesting to note that DD alleles have also been associated with a worse outcome in patients with coronary artery disease.

The serum lysozyme can also be elevated in sarcoidosis in the same way as ACE. However, it is elevated in a smaller proportion of patients and not routinely used in clinical practice.

Tests of the lung

Bronchoalveolar lavage findings can be characteristic in sarcoidosis. The finding of increased lymphocytes, especially an increased CD4 to CD8 ratio, has been interpreted by some groups as enough to make the diagnosis of sarcoidosis, and in a patient with a compatible clinical history and no evidence for infection or malignancy, the lavage findings may be considered sufficient. A more definitive answer from bronchoscopy includes a transbronchial biopsy showing non-caseating granulomas. In over 60 per cent of patients with a stage 1 chest radiograph the biopsy should be positive, rising to 80 per cent in patients with stage 2 or 3 disease. Transbronchial needle aspiration has recently been used to sample hilar lymph nodes, but raises the problem of incomplete sampling in patients with a malignancy and granulomatous response to the lesion. Mediastinoscopy and video-assisted thoracoscopy provide a minimally invasive method to obtain more tissue.

The gallium scan can reveal increased activity in patients with sarcoidosis. The activated macrophage has increased levels of transferrin receptors on its surface and this results in an increase in gallium uptake. Unfortunately, interpreting the uptake in the lung may be difficult as it is non-specific and occurs with other inflammatory lung diseases. It also rapidly returns to normal with corticosteroid therapy. On the other hand, the uptake in the parotid and conjunctiva (the 'panda' sign) and the uptake in the hilar nodes (the 'lambda' sign) are fairly characteristic for sarcoidosis and are useful confirmation in difficult cases.

Other tests

The Kveim–Siltzbach agent is a suspension of spleen tissue from a patient with confirmed sarcoidosis. Six weeks after an intradermal injection of the agent, the site is inspected for a reaction, which will occur in over 60 per cent of patients with acute sarcoidosis. On biopsy, the reaction will show non-caseating granulomas, consistent with sarcoidosis. Properly prepared Kveim–Siltzbach agent has a less than 1 in 500 chance of causing a false positive. However, because of the difficulties in preparing the agent and concerns regarding the risk of transmission of an infectious agent, the test is rarely used except in those centres with a well established reagent.

Other laboratory tests may support the diagnosis of sarcoidosis or be useful in monitoring the level of disease activity. The erythrocyte sedimentation rate can be elevated in sarcoidosis and fall with remission of the disease, but one-third of patients will have a normal sedimentation rate, so it is not specific or sensitive.

Serum calcium is elevated in 10 per cent of cases and is supportive of the diagnosis, but hypercalcaemia can be seen in other conditions which mimic sarcoidosis, such as malignancy. Hypercalcaemia due to sarcoidosis should be associated with a normal to low serum phosphorus. In patients with significant hypercalcaemia, renal failure may occur, reversible in many cases with lowering of the serum calcium.

Hypergammaglobulinaemia is also a feature of sarcoidosis. Activated T lymphocytes in the lung are capable of stimulating circulating peripheral blood B cells, leading to the polyclonal gammaglobulin response. As a result of this non-specific reaction, serological markers for some diseases may be falsely elevated. This includes antifungal antibodies and antinuclear antibodies. Hypergammaglobulinaemia is more common in African-American than Caucasian individuals.

Liver involvement occurs in over half of patients with sarcoidosis. In some cases the liver blood tests are entirely normal, but the majority of patients with liver involvement have elevated serum enzymes. Usually the pattern is obstructive, with a rise in the serum alkaline phosphatase, but in some an elevation of the transaminases is seen. Elevation of the serum bilirubin is less frequent and associated with more extensive liver involvement. Rarely, lymphadenopathy at the porta hepatis can lead to biliary obstruction.

Haematological abnormalities are common in sarcoidosis. Lymphopenia is frequently seen, and is probably due to sequestration of lymphocytes into the area of inflammation, such as the lung. Anaemia has been reported in about 20 per cent of patients. The mechanism is multifactorial, including a high proportion of patients with iron deficiency. Other causes include direct bone marrow invasion by granulomas. Cytokines such as IL-2 may also result in suppression of the bone marrow.

Treatment

The natural course of sarcoidosis is unclear, since corticosteroids are normally used to treat symptomatic patients. For the patient with no symptoms on presentation, the prognosis is often good. Spontaneous resolution commonly occurs within a year or two of diagnosis, but the disease can also take a chronic form, with symptoms for many years. The concept of acute disease, which lasts for less than 2 years, as opposed to chronic disease has been a useful method for considering patients, especially in terms of therapy. Table 5 lists several factors associated with resolution within 2 to 5 years as well as those predicting chronic disease. Acute disease is associated with erythema nodosum, hilar adenopathy, anterior uveitis, and Bell's palsy. Whereas chronic disease includes such manifestations as lupus pernio, stage 4 chest radiograph, posterior uveitis, and bone cysts. Most chronic disease is controllable by therapy, but there is a refractory form of the disease which often involves the cardiac and central nervous systems. Mortality from sarcoidosis occurs, but is less than 5 per cent in most series.

The major indication for therapy in sarcoidosis is symptoms. Hypercalcaemia should be treated, even if the patient is asymptomatic. An eye examination should be performed in all patients with sarcoidosis. Uveitis may be misdiagnosed as sicca (dry eyes). The former will require anti-inflammatory agents, while the latter will only need a wetting agent.

If possible, treatment should be topical. Corticosteroid creams and eye drops are effective if inflammation is superficial. The effectiveness of inhaled steroids is less clear-cut. The higher-potency steroids such as budesonide appear to have a role in reducing the dosage of systemic corticosteroids. Several randomized trials have indicated a role for this drug as maintenance therapy for a patient who has received systemic therapy for 3 months to induce remission.

It is not clear whether corticosteroids change the natural course of the disease. Early randomized trials found no difference in the long-term outcome of patients who received corticosteroids compared with controls. A British Thoracic Society randomized study did demonstrate a small benefit for corticosteroids over placebo for patients with persistent, but not severe, disease. One of the difficulties in assessing this and other trials is that the more severely affected patients were treated with corticosteroids and excluded from the study.

Figure 4 summarizes the use of systemic corticosteroid therapy at three large centres over a period of several years. The genetic background of each group varies,

with the Iowa group being mostly of Scandinavian descent, Philadelphia being African-American, and Milan being Italian. In general, 30 to 60 per cent of patients never required therapy. However, once therapy was instituted, in 18 to 53 per cent of the patients it could not be withdrawn. In patients who were tapered off corticosteroids, one group found that 80 per cent eventually relapsed and required reinstitution of therapy. The differences in rate of continued therapy and relapse between the centres could be due to either the genetic background of the patients or the bias of the treating physicians.

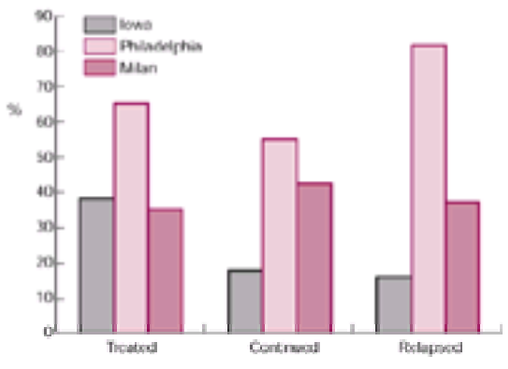

Fig. 4 The results of corticosteroid therapy from three medical centres. The first set of columns demonstrates the percentage of patients treated. The second set of columns represents the percentage of patients who required therapy to be continued beyond 2 years. The last set of columns reports the subgroup of people initially treated with corticosteroids and tapered off treatment who subsequently required reinstitution of therapy.

Once systemic steroids are initiated, one has to remember that a prolonged course is usually necessary. In the beginning, doses may be changed every 1 to 3 months, but after a patient tolerates a lower dose (equivalent to less than 10 mg of prednisolone per day), tapering may be prolonged. The use of alternate-day corticosteroids is strongly advocated by some, but others are less enthusiastic. Rarely will alternate-day therapy be sufficient for initial control of the disease.

The toxicity of corticosteroids is well known. These include weight gain, diabetes mellitus, hypertension, and mood swings. Avascular necrosis and osteopenia are significant problems with prolonged use. Some patients with sarcoidosis will have lost weight as part of their disease. However, the weight gain with treatment often surpasses the amount of weight loss. The longer a patient is on corticosteroids, the greater the risk of problems. Unfortunately, most patients will require more than a year of treatment.

Several alternatives to systemic corticosteroids have been proposed over the years. These are summarized and compared with corticosteroids in Table 6. We include the usual doses, commonly encountered toxicities, an estimate of response rate, and the usual indications for use.

The commonly prescribed antimalarial agents chloroquine and hydroxychloroquine possess anti-inflammatory activity with their major toxicities being eye and gastrointestinal. Because hydroxychloroquine, especially at 400 mg a day or less, is unlikely to cause eye toxicity, it is more frequently prescribed. However, some experts feel chloroquine is a more effective agent. These drugs concentrate in the skin and are most efficacious for skin disease and hypercalcaemia. They are less successful in the treatment of pulmonary disease. A recent report described some utility in patients with neurological disease.

Methotrexate is an antimetabolite chemotherapy used for various solid tumours. Over 30 years ago it was first noted as an immunosuppressant for the treatment of rheumatoid arthritis. In that disease, it prevents joint destruction and is often used to avoid steroid toxicity. In sarcoidosis, methotrexate has been most studied as a treatment for chronic disease. This probably reflects the fact that it may require 6 months for the drug to become effective. The usual dose is 10 to 15 mg orally each week, adjusted if this proves toxic. Acute toxicity including mucositis and nausea can be minimized with supplements of folic acid at 1 mg/day. Leucopenia can also occur, but is usually insignificant unless the patient is already leucopenic from sarcoidosis or the patient has renal insufficiency. We have successfully treated patients with doses as small as 2.5 mg of methotrexate a week. The long-term toxicity of methotrexate can include hypersensitivity pneumonitis and cirrhosis. The latter is a concern, because 50 per cent of patients with chronic disease will have sarcoid granulomas in a liver biopsy, and we recommend liver biopsies every 2 years for patients requiring the drug long term. Methotrexate is teratogenic, but current data suggest it is not carcinogenic.

The response rate to methotrexate in chronic sarcoidosis is 60 to 80 per cent. Most patients who respond can be treated with methotrexate alone. Approximately 20 per cent of patients will require additional low-dose corticosteroids. In most patients skin lesions can be easily controlled with methotrexate, but studies have also reported benefit for disease in the lungs, eyes, and nervous system.

Azathioprine has been used for many years as an immunosuppressant for patients receiving solid-organ transplants and those with idiopathic pulmonary fibrosis. However, its use in sarcoidosis has been more sporadic, usually reserved for chronic cases. Its major side-effects are gastrointestinal and haematological.

Other drugs have been used for refractory sarcoidosis. Cyclophosphamide is used in the treatment of many vasculitic diseases and has been reported as very useful in neurological and cardiac sarcoidosis, but it has more gastrointestinal, haematological, and bladder toxicity than methotrexate or azathioprine.

Case reports suggest that thalidomide may be useful in treating sarcoidosis. It has severe teratogenic effects and may cause peripheral neuropathy and drowsiness. Cyclosporin has been used with limited success in some neurological cases. A recent randomized trial failed to show additional benefit over corticosteroids alone in patients with pulmonary sarcoidosis. The drug is relatively expensive, causes hypertension and renal failure, and requires blood levels to be monitored. Pentoxyfylline has been shown by one centre to provide some benefit in acute sarcoidosis. It is associated with significant gastrointestinal toxicity, which is dose dependent.

There is no single treatment for all patients with sarcoidosis. Figure 5 summarizes our clinical approach. The first step is to determine whether the patient requires treatment. The decision to treat is usually based on the patient's symptoms. The clinician needs to determine the extent of the symptomatic disease and whether the disease is acute or chronic. Asymptomatic or minimally symptomatic patients with hypercalcaemia, cardiac, or central nervous system disease may require therapy to prevent life-threatening complications. The use of systemic therapy usually means corticosteroids first. However, over time, the patient and the physician may need to seek alternatives.

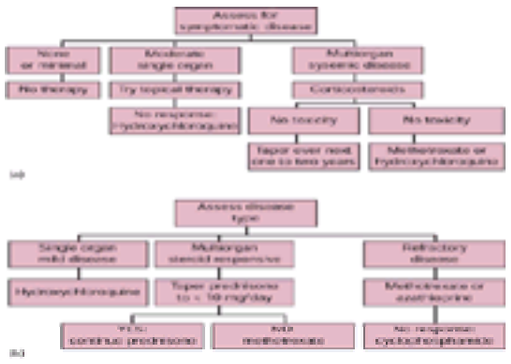

Fig. 5 The approach to treatment of patients with sarcoidosis at the Interstitial Lung Disease Clinic at the University of Cincinnati. Patients are initially classified as having either acute (a) or chronic (b) disease. Further evaluation depends on their symptoms and response to therapy.

Prognosis

Sarcoidosis will resolve within 2 to 5 years in the majority of cases. Approximately 25 per cent of patients will develop residual fibrosis. In a minority of patients, the disease will become chronic and persist for more than 5 years.

For the patient with chronic disease, treatment can usually palliate the symptoms. However, organ failure—including eye, liver, cardiac, or respiratory—can occur as a result of disease.

Most series from referral centres report a 5 per cent mortality from sarcoidosis, most commonly due to respiratory failure, but with cardiac, neurological, and hepatic failure as other causes. Respiratory failure leading to death can be predicted from pulmonary function tests. For example, one study found that of those with a vital capacity persistently less than 1 litre, a third died of respiratory failure, whereas no patient with a vital capacity of over 1.5 litres did so. The 5 per cent mortality reported from referral centres is higher than that encountered in a non-referral setting, where the mortality approximates 1 per cent. Organ transplantation has been performed successfully in patients with sarcoidosis. Although sarcoidosis lesions can occur in the new organ, organ failure due to recurrent sarcoidosis is unlikely.

Special problems

Endstage lung disease is the most common problem for patients with severe sarcoidosis. The fibrotic disease leads to cor pulmonale and respiratory distress. In addition, cavitary lesions can lead to bronchiectasis or become colonized with aspergillus. Aspergillomas can cause fatal haemoptysis. Their treatment can be difficult because most patients are not good surgical candidates. Embolization has been used for life-threatening bleeding.

Steroid-induced osteopenia is a significant problem with long-term corticosteroid therapy. Because of the risk of hypercalcaemia, patients are often not given calcium supplements, but this should be considered if a patient requires long-term systemic steroids. Monitoring serum calcium during therapy is usually sufficient to avoid complications. The use of nasal calcitonin or oral bisphosphonates should also be considered.

Areas of uncertainty/controversy

Some clinicians have proposed the use of bronchoalveolar lavage as the exclusive diagnostic test for sarcoidosis. This is based on the rationale that in the appropriate clinical setting, bronchoalveolar lavage findings of increased lymphocytes and a CD4 to CD8 ratio of greater than 3.5 represents a granulomatous process. In patients with cultures negative for tuberculosis and fungal infection, sarcoidosis is most likely and no further diagnostic testing may be needed. The percentage of patients with increased lymphocytes and CD4 to CD8 ratio varies from centre to centre. In our institution, at least 50 per cent of patients with sarcoidosis will meet these criteria. However, the use of bronchoalveolar lavage does not provide an absolute diagnosis of sarcoidosis. As previously noted, transbronchial needle aspirate may also be useful in making a diagnosis; however, the finding of a granuloma does not guarantee the diagnosis. One way to interpret the bronchoalveolar lavage or tissue results is to consider each as complementary in making the diagnosis of sarcoidosis.

The use of corticosteroids for the treatment of sarcoidosis remains controversial. In the patient with minimal symptoms, treatment can be withheld or topical agents used. If the disease spontaneously resolves, no therapy is indicated. However, if the patient becomes symptomatic, corticosteroids will probably be useful. The treatment of the patient with persistent, mild disease is unclear. The British Thoracic study suggests that these patients should receive corticosteroids, but others argue that the benefits are small and do not justify the use of these agents.

Areas needing further research

The cause of sarcoidosis remains unknown. The newer techniques of molecular biology may provide additional insight into a causative agent or the underlying genetic predisposition for the disease.

Patients with chronic disease represent a disproportionate number of cases with increased morbidity and need for medical services. The use of corticosteroids alone is not adequate for many of these patients. Research into whether other agents are truly steroid sparing and associated with a good clinical outcome are still necessary.

Further reading

Agbogu BN *et al*. (1995). Therapeutic considerations in patients with refractory neurosarcoidosis. *Archives of Neurology* **52**, 875–9.

Agostini C, Semenzato G (1998). Cytokines in sarcoidosis. *Seminars in Respiratory Infections* **13**, 184–96.

Bardelli AM *et al*. (1993). Eye involvement in sarcoidosis: survey of 197 patients. *Sarcoidosis* **10**, 158–9.

Baughman RP, Lower EE (1997). Steroid-sparing alternative treatments for sarcoidosis. *Clinics in Chest Medicine* **18**, 853–64.

Baughman RP *et al*. (1997). Predicting respiratory failure in sarcoidosis patients. *Sarcoidosis Vasculitis and Diffuse Lung Diseases* **14**, 154–8.

Crystal RG *et al*. (1981). Pulmonary sarcoidosis: a disease characterized and perpetuated by activated lung T-lymphocytes. *Annals of Internal Medicine* **94**, 73–94.

Gibson GJ *et al*. (1996). British Thoracic Society Sarcoidosis study: effects of long term corticosteroid treatment. *Thorax* **51**, 238–47.

Gottlieb JE *et al*. (1997). Outcome in sarcoidosis. The relationship of relapse to corticosteroid therapy. *Chest* **111**, 623–31.

Hance AJ (1998). The role of mycobacteria in the pathogenesis of sarcoidosis. *Seminars in Respiratory Infection* **13**, 197–205.

Hirose Y *et al*. (1994). Myocardial involvement in patients with sarcoidosis. An analysis of 75 patients. *Clinics in Nuclear Medicine* **19**, 522–6.

Hunninghake GW *et al*. (1994). Outcome of the treatment for sarcoidosis. *American Journal of Respiratory and Critical Care Medicine* **149**, 893–8.

Izumi T (1988). Sarcoidosis in Kyoto (1963–1986). *Sarcoidosis* **5**, 142–6.

James DG, ed. (1994). *Sarcoidosis and other granulomatous disorders*. Marcel Decker, New York.

Johns CJ, Zachary JB, Ball WC (1974). A ten year study of corticosteroid treatment of pulmonary sarcoidosis. *The Johns Hopkins Medical Journal* **134**, 271–83.

Judson MA *et al*. (1999). Defining organ involvement in sarcoidosis: the ACCESS proposed instrument. *Sarcoidosis Vasculitis and Diffuse Lung Diseases* **16**, 75–86.

Lower EE, Baughman RP (1995). Prolonged use of methotrexate for sarcoidosis. *Archives of Internal Medicine* **155**, 846–51.

Lower EE *et al*. (1997). Diagnosis and management of neurologic sarcoidosis. *Archives of Internal Medicine* **157**, 1864–8.

Lower EE *et al*. (1988). The anemia of sarcoidosis. *Sarcoidosis* **5**, 51–5.

Lynch JPI, McCune WJ (1997). Immunosuppressive and cytotoxic pharmacotherapy for pulmonary disorders. *American Journal of Respiratory and Critical Care Medicine* **155**, 395–420.

Lynch JP, Kazerooni EA, Gay SE (1997). Pulmonary sarcoidosis. *Clinics in Chest Medicine* **18**, 755–85.

Lynch JP, Sharma OP, Baughman RP (1998). Extrapulmonary sarcoidosis. *Seminars in Respiratory Infection* **13**, 229–54.

Maliarik MJ *et al*. (1998). Angiotensin-converting enzyme gene polymorphism and risk of sarcoidosis. *American Journal of Respiratory and Critical Care Medicine* **158**, 1566–70.

Neville E, Walker AN, James DG (1983). Prognostic factors predicting the outcome of sarcoidosis: an analysis of 818 patients. *Quarterly Journal of Medicine* **208**, 525–33.

Newman LS, Rose CS, Maier LA (1997). Sarcoidosis. *New England Journal of Medicine* **336**, 1224–34.

Parkes SA *et al*. (1987). Epidemiology of sarcoidosis in the Isle of Man—1: A case controlled study. *Thorax* **42**, 420–6.

Pietinalho A *et al*. (1999). Oral prednisolone followed by inhaled budesonide in newly diagnosed pulmonary sarcoidosis: a double-blind, placebo-controlled, multicenter study. *Chest* **116**, 424–31.

Rizzato G, Montemurro L (1993). Reversibility of exogenous corticosteroid-induced bone loss. *European Respiratory Journal* **6**, 116–19.

Rizzato G, Montemurro L, Colombo P (1998). The late follow-up of chronic sarcoid patients previously treated with corticosteroids. *Sarcoidosis Vasculitis and Diffuse Lung Diseases* **15**, 52–8.

Selroos O (1969). The frequency, clinical picture and prognosis of pulmonary sarcoidosis in Finland. *Acta Medica Scandinavica Supplementum* **503**, 3–73.

Selroos O, Sellergren TL (1979). Corticosteroid therapy of pulmonary sarcoidosis. *Scandinavian Journal of Respiratory Diseases* **60**, 215.

The Committee on Sarcoidosis of the American Thoracic Society (1999). Statement on sarcoidosis. *American Journal of Respiratory and Critical Care Medicine* **160**, 736–55.

Thomas PD, Hunninghake GW (1987). Current concepts of the pathogenesis of sarcoidosis. *American Review of Respiratory Disease* **135**, 747–60.

17.11.7

Pneumoconioses

A. Seaton

Most lung diseases are caused or provoked at least in part by the inhalation of harmful material. A wide range of lung conditions, including cancer (exposure to asbestos, radon daughters in mines, polycyclic aromatic hydrocarbons, nickel refining, chloromethyl ethers), pneumonia (legionnaire's disease in hospitals), asthma (flour, isocyanates, epoxy resins), allergic alveolitis (farmer's lung, maltworker's lung), and toxic pneumonitis (silo-filler's disease, chlorine poisoning, cadmium poisoning), may occur as a result of workplace exposure. When exposure to mineral dust in the workplace results in a diffuse, usually fibrotic, reaction in the lung acinus, the condition is called a pneumoconiosis.

The distinction between tuberculosis and a specific effect of dust in the causation of respiratory disease was made in the mid-nineteenth century. By this time silicosis, often complicated by tuberculosis, was widespread amongst metal miners, tunnellers, potters, and cutlers. The Industrial Revolution stimulated the need for coal, and the production of this fuel resulted in increasing numbers of sufferers from coal-worker's pneumoconiosis. This in turn was not distinguished from silicosis until the late 1940s, and in some countries the two conditions are still referred to by the one name.

In the United Kingdom, and generally in the West, dust control in mines and decline of traditional industries have resulted in a reduction in the numbers of workers suffering from these two diseases. By contrast, industrialization of developing countries has stimulated the need for indigenous coal and minerals, and in China, South America, and India several million workers are employed in mining, often in conditions which ensure a high incidence of pneumoconiosis. At the same time, the rise of the asbestos and chemical industries has added new problems for society in weighing the benefits of the product against the cost in terms of human morbidity. Fortunately, these problems are potentially soluble by application of preventive measures, as emphasized in the following sections.

Coal-worker's pneumoconiosis

Coal-worker's pneumoconiosis is caused by inhalation of coal-mine dust, a complex mixture of coal, kaolin, mica, silica, and other clay minerals. It is now uncommon in the United Kingdom, and good dust control together with reductions in the workforce imply that the disease may disappear in the next few years. Nevertheless, the strategic importance of coal as a long-term source of fuel supply and as a chemical feedstock means that it will continue to be needed, and any relaxation of dust control in mines will be followed by the reappearance of pneumoconiosis. There has also been a reduction in the incidence of coal-worker's pneumoconiosis in other Western countries, but in China the disease is widespread and in India it afflicts about 1 to 2 per cent of the current workforce of 800 000.

Aetiology and pathology

The pathogenicity of coal dust differs in different areas. If lung damage is to occur, the dust must be inhalable to acinar level within the lung. Thus the particles must have aerodynamic characteristics that make them equivalent to a sphere of unit density between 0.5 and 7 µm in diameter. Once inhaled, the particles must be able to overcome the lung's defences. Some, containing a high proportion of quartz (crystalline silicon dioxide), are toxic to macrophages and cause their disruption after phagocytosis. Such particles are cleared predominantly to the lymph nodes where they remain and set up a fibrotic reaction that ultimately destroys the node. However, some remain in the peribronchiolar and perivascular parts of the acinus, where whorled fibrosis occurs leading to the typical silicotic nodule. The mechanisms of quartz-induced fibrosis are discussed further in the silicosis section. However, since most coal dust contains relatively little quartz and is not particularly toxic to macrophages *in vitro*, some other explanation for its harmfulness must be sought. *In vivo* studies in rats have shown that inhalation of relatively low concentrations of coal dust, comparable with those occurring in United Kingdom mines in the recent past, cause inhibition of macrophage migration and provoke an inflammatory response, mediated *inter alia* by interleukin 1 (**IL-1**) and tumour necrosis factor, resulting in the release of elastase and the degradation of fibronectin. It seems likely that these toxic effects *in vivo* on macrophages are fundamental to the pathological processes in coal-worker's pneumoconiosis, including the concomitant centriacinar emphysema.

The total amount of dust inhaled is also a critical factor in the development of pneumoconiosis. Epidemiological studies have shown clear relationships between cumulative dust exposure and radiological evidence of disease. However, this is not straightforward, as some coal dusts are clearly more toxic than others, and it is not always possible to characterize this toxicity by the relative mineralogical composition of the coal dust. There is evidence that the different minerals in coal dust interact and that some clays may reduce its overall toxicity, perhaps by blocking the surface activity of the toxic fraction. As a general rule, the higher the combustibility (rank) of the coal, the more likely is its dust to cause pneumoconiosis (Fig. 1).

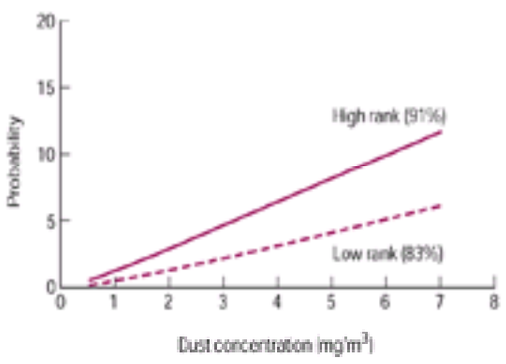

Fig. 1 Relationship between risk of category 2 or 3 radiological simple pneumoconiosis and daily exposure over a working lifetime to different concentrations of coal dust. The greater risk in association with exposure to dust from coals of higher combustibility (rank) should be noted.

Pathologically, coal-worker's pneumoconiosis is characterized by the presence of multiple centriacinar and interlobular foci of dust, inflammatory cells, macrophages, and reticulin or collagen—the coal macule (Fig. 2). In miners exposed to relatively high proportions of quartz, the lesions have a greater resemblance to the silicotic nodule. The presence of small discrete nodules is known as simple pneumoconiosis, and when sufficient numbers of these lesions are present they become visible on a radiograph. Complicated pneumoconiosis, or progressive massive fibrosis, is present by definition when one or more of these lesions is greater than 1 cm in diameter (Fig. 3). This occurs either by aggregation of several, usually collagenous, smaller nodules or by a more diffuse accumulation of dust associated with dead cells and ischaemic necrosis of lung tissue. The former, less common, mechanism occurs particularly in relation to relatively high quartz exposures, while the latter seems more frequent with exposure to high carbon dusts. With either type, and with intermediate types, there is a tendency for the lesions to grow and to be associated with surrounding bullous emphysema, ultimately being responsible for destruction of large volumes of the lung.

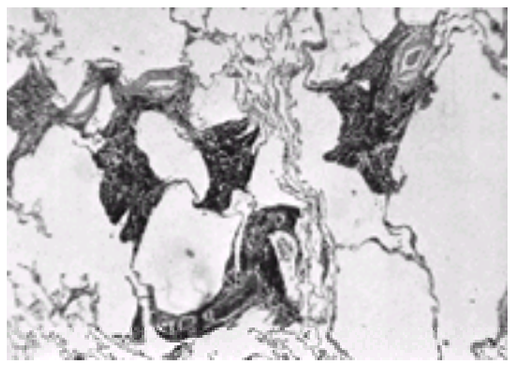

Fig. 2 Simple coal macules, showing accumulations of dust and cells around centre of lobule with associated emphysema.

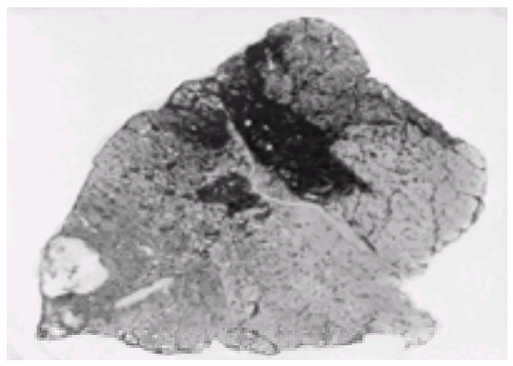

Fig. 3 Whole-lung section of a coal-miner's lung showing progressive massive fibrosis.

The aetiology of progressive massive fibrosis is not completely understood. It is more common in the upper lung zones and in taller men, suggesting a relationship to failure of lung clearance. High carbon or high quartz dusts are particularly liable to cause progressive massive fibrosis, and the higher the dust exposure, the greater is the risk (Fig. 4). Tuberculous infection is no longer an important factor in the developed world, although it may have been in the past. The rheumatoid diathesis is responsible for initiating a rare type of progressive massive fibrosis (Caplan's syndrome), but this is not an important factor overall. For further discussion of the aetiology of progressive massive fibrosis, see the section on silicosis below.

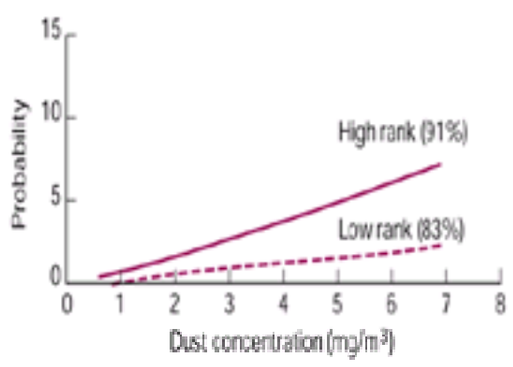

Fig. 4 Relationship between risk of progressive massive fibrosis and exposure to dust over a working lifetime. Again, greater risk in association with exposure to dust from coals of higher combustibility (rank) should be noted.

Clinical features

The people most at risk of coal-worker's pneumoconiosis are those working in the dustiest areas, such as face-workers cutting coal, drilling for shot-firing, developing headings, and drilling bolts into the roof to prevent it falling. Open-cast miners rarely work in such dusty circumstances, except in hot dry countries such as India where loading operations may be extremely dusty. Simple coal-worker's pneumoconiosis causes no symptoms or physical signs, nor any important physiological abnormality. This fact is of importance, as symptoms of respiratory disease in a miner with this condition are due to some other cause, such as bronchitis, heart failure, or asthma, which may be treatable. Radiological progression or regression of simple pneumoconiosis occurs only very rarely after dust exposure ceases, apparent regression sometimes being associated with the development of emphysema.

The danger associated with simple pneumoconiosis is that it predisposes to progressive massive fibrosis, a risk directly related to the profusion of simple pneumoconiosis on the radiograph. Progressive massive fibrosis may occur during working life or appear for the first time after (sometimes many years after) dust exposure ceases, even when there is no apparent simple pneumoconiosis on the radiograph. Progressive massive fibrosis usually progresses and causes a mixture of restriction of lung volumes and, owing to associated emphysema, airflow obstruction. Ultimately it may lead to cor pulmonale and death. However, the rate of progression is very variable. In general, the earlier progressive massive fibrosis develops in a person's life, the more rapidly progressive and thus the greater a threat to health it is.

The patient with progressive massive fibrosis may complain of shortness of breath and symptoms of cor pulmonale. An unusual, but pathognomonic, symptom is melanoptysis—the expectoration of the black contents of a cavitated lesion. Haemoptysis and finger clubbing suggest lung cancer and should not be attributed to pneumoconiosis. Abnormal signs in the chest, if present, relate to the presence of bullae, although sometimes lobar collapse can occur.

Coal-worker's pneumoconiosis is not associated with an increased risk of tuberculosis or lung cancer, although obviously these diseases can occur in coal miners and should be suspected if unusual progression of radiological changes occurs. The association between pneumoconiosis and emphysema has been controversial, but there is now clear evidence of a parallel association between dust exposure and two effects—pneumo-coniosis and airflow obstruction. The more dust that a miner

has been exposed to, the greater are his risks of pneumoconiosis on the one hand, and productive cough, reduction in forced expiratory volume in 1 s (FEV_1), and presence of centriacinar emphysema on the other. Of course, the latter risks are also related to cigarette smoking, and the effect of dust exposure is additive.

The radiological lesions in simple pneumoconiosis are predominantly rounded opacities between 1 and 5 mm in diameter, although small irregular and linear opacities and Kerley B lines are frequently present also. The round opacities tend to be more profuse in the upper and middle zones, whereas the irregular lesions predominate in the lower zone (Fig. 5). Progressive massive fibrosis almost always starts in an upper zone, gradually increasing in size until it may occupy up to a third of the lung. Such lesions are frequently multiple. They are often shaped like short fat sausages, with their outer border curved with the chest wall and separated from the pleura by bullous emphysema (Fig. 6). Calcification is not a feature, but cavitation of progressive massive fibrosis may occur. Caplan's syndrome is the name given to the combination of rheumatoid disease and several round nodules (usually 1 to 5 cm in diameter) in the lungs of a coal miner. The lesions have a rheumatoid histology and rarely cause any serious pulmonary impairment; they often cavitate and disappear. The radiological features of all pneumoconioses are properly described in terms of a set of standard radiographs produced by the International Labour Organization, and use of these standards is mandatory for epidemiological studies.

Fig. 5 Radiograph of a coal miner showing small round lesions of simple pneumoconiosis. Some irregular shadows are also present in the lower zones.

Fig. 6 CT scan of miner, showing central progressive massive fibrosis and surrounding bullous emphysema.

Prevention and management

Epidemiology has shown an exposure–response relationship between the total mass of respirable coal dust to which miners have been exposed and their risks of developing simple pneumoconiosis. This has allowed standards to be set for coal-mine dust levels, which have resulted in falls in the prevalence of the disease in coal mines in the West. Their success depends on regular monitoring of the respirable dust by gravimetric sampler, constant attention to dust suppression by ventilation and the use of water at points of dust production, and regular radiography of the workforce to detect early signs of dust retention. The incidence of progressive massive fibrosis is largely controlled by preventing miners from contracting simple pneumoconiosis, and working conditions in British mines are currently such that this disease is now very rare indeed. The present British standard is 7 mg/m^3, measured in the air returning from the coalface.

If a miner develops simple pneumoconiosis late in his career, no action normally needs to be taken, apart from (in the United Kingdom) advising him to apply to the Respiratory Diseases Board via the Benefits Agency for assessment of disablement and possible benefit payments. A younger man, with several years of further dust exposure ahead, should be advised to work in an area of approved low dust conditions. This advice should be given in the United Kingdom by the employer's occupational health service. Men with more than the earliest stages of radiological change are entitled to disablement benefits from the Benefits Agency, the value of these depending on the extent of disability. Since simple pneumoconiosis *per se* does not disable, these benefits are often small. Payment of benefits for airflow obstruction as an associated effect of coal dust exposure are also made in the United Kingdom if the miner has worked underground for a minimum of 20 years and his FEV_1 is a litre below that predicted. The presence of associated radiological change is no longer necessary.

Silicosis

Silicosis is a fibrotic disease of the lungs due to inhalation of crystalline silicon dioxide, usually in the form of quartz. Such a disease has occurred in metal miners and masons since ancient times, but assumed particular importance in the cutlery and pottery trades in the nineteenth century. Silicosis may affect anyone involved in quarrying, carving, mining, tunnelling, grinding, or sandblasting, if the dust generated contains quartz. In the United Kingdom the traditional trades that caused the disease (pottery, cutlery, flint knapping, sandblasting, tin and iron mining, and slate quarrying) have either introduced safe substitute materials or have declined, so that true silicosis is now quite rare. Between 50 and 60 cases are diagnosed in the United Kingdom each year, usually in the production of slate or granite, among miners cutting through rock, and in fettlers in foundries. However, the author has seen a series of severe cases in British workers who had been employed in circumstances where the risks had been forgotten or were being ignored.

Aetiology and pathology

Crystalline silica is present in the earth's crust usually as quartz, although other forms such as crystobalite and tridymite occur occasionally. All are extremely toxic to macrophages. quartz seems to be most toxic when freshly fractured, suggesting that surface properties are important in toxicity. This concept is supported by experimental evidence that various clay minerals and other chemicals which occlude the surface reduce the toxicity of inhaled quartz when inhaled simultaneously in mixtures of dust. The quartz content of dust from different types of stone may vary considerably from some sandstones which are 100 per cent quartz to shales and slates which may contain less than 10 per cent.

Inhaled particles of quartz small enough (generally less than 7 µm aerodynamic diameter) to reach the acinus are engulfed by macrophages and cause disruption of the phagosome, probably by peroxidation of membrane lipids. Before macrophage death, other reactions occur leading to release of inflammatory mediators, including IL-1, various growth factors, tumour necrosis factor, and fibronectin, largely from interstitial rather than alveolar macrophages. Silica is probably transported across the alveolar epithelium by migrating macrophages and by endocytosis by type 1 alveolar cells, and it is clear from the distribution of pathological lesions that quartz is transported widely in the lung via lymphatics, much of it ultimately being deposited in hilar nodes, which it destroys. This destruction of the nodes is very likely to be responsible for blockage of the exit route for further inhaled dust, and therefore for its retention in the lung and the developmnent of progressive massive fibrosis or, rarely, accelerated or even acute silicosis.

Macroscopic inspection of silicotic lungs shows fibrous pleural adhesions, enlarged lymph nodes that contain fibrotic, often calcified, nodules, and grey nodules throughout the lung. These nodules vary from a few millimetres to several centimetres in diameter and are more profuse in the upper zones (Fig. 7). They may be calcified, and they have a typical whorled appearance when cut across (Fig. 8). The largest lesions consist of many such nodules that have become confluent, and, as

in coal-worker's pneumoconiosis, this progressive massive fibrosis may undergo ischaemic necrosis and cavitate. Under the microscope the silicotic nodule consists of concentric layers of collagen surrounded by a zone of doubly refractile silica particles, macrophages, and fibroblasts. The nodule may contain the remnants of the respiratory bronchiole and arteriole, destroyed by fibrosis. The mechanisms responsible are destruction of macrophages leading to inflammation and laying down of collagen, release of the quartz, further macrophage attraction, and repetition of the cycle. This presumably occurs first in nodes on the drainage pathway, and as these become progressively blocked the process is repeated in the lung. As the quartz never gets removed thereafter, the process continues indefinitely and severity of disease depends on the mass inhaled and retained.

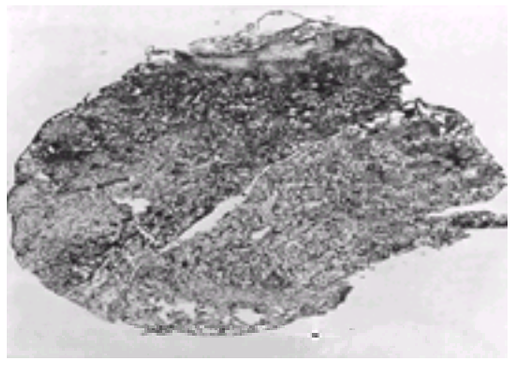

Fig. 7 Whole-lung section from a coal miner whose work had been predominantly in hard rock, showing silicotic nodules in upper parts of upper and lower lobes.

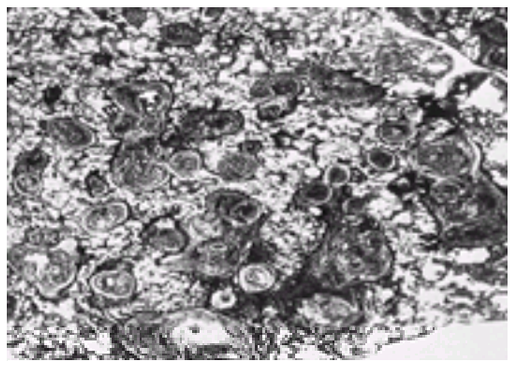

Fig. 8 Silicotic nodules, showing the typical whorled appearance.

Macroscopically, acute silicosis appears like pulmonary oedema. Under the microscope, the alveoli are filled with eosinophilic fluid and the alveolar walls contain plasma cells, lymphocytes, fibroblasts, and silica. In the author's experience, this condition first requires hilar node destruction by the inhaled quartz.

Clinical features

Silicosis presents a spectrum of clinical appearances depending on the circumstances in which it is contracted. The most severe, acute silicosis, may be acquired after very heavy exposure over only months, such as may occur in sandblasting without respiratory protection. Such patients become intensely breathless and die within months. The radiograph shows appearances resembling pulmonary oedema. Less heavy exposure causes progressively less dramatic symptoms, ranging from a progressive upper lobe fibrosis with slowly increasing exertional dyspnoea over several years (accelerated silicosis) to radiographic nodular change similar to coal-worker's pneumoconiosis unassociated with any symptoms or physical signs. The latter type of silicosis is the most common, and is usually associated with exposure to dust containing 10 to 30 per cent silica over a prolonged period. Simple nodular silicosis differs from coal-worker's pneumoconiosis in that the lesions tend to be larger (3 to 5 mm) and that it is progressive even after dust exposure ceases. Lesions increase in size and become more profuse. Moreover, extensive simple silicosis may be associated with some restriction of lung volumes. Simple silicosis rarely seems to be associated with emphysema, unlike coal-worker's pneumoconiosis, but silicotic progressive massive fibrosis is commonly associated with bullous disease. Curiously, it has only recently been recognized that acute enlargement of hilar nodes mimicking sarcoidosis may be an early feature of silicosis. Accelerated silicosis and progressive massive fibrosis cause lung restriction and lead to cor pulmonale and cardiorespiratory failure.

Apart from evidence of cardiac failure or distortion of lung architecture by extreme degrees of massive fibrosis, physical signs are not prominent. Clubbing and crackles are not seen. Diagnosis depends on a history of exposure and the radiographic appearances. The most characteristic of these are nodules between 3 and 5 mm in diameter, predominantly in the upper zones, and eggshell calcification in the hilar nodes (Fig. 9). The latter is virtually a pathognomonic feature, only occurring otherwise, very rarely, in sarcoidosis. All forms of silicosis are liable to be complicated by tuberculosis, usually due to reactivation of a quiescent lesion.

Fig. 9 Radiograph of a hard rock miner, showing massive fibrosis in right mid-zone and eggshell calcification of hilar nodes.

Other mycobacterial diseases (*Mycobacterium kansasii* and *Mycobacterium avium-intracellulare*) also occur more frequently than would be expected in those with silicosis. There is now evidence of a weak association between silicosis and lung cancer, even when exposures to cigarette smoke and other occupational carcinogens have been accounted for. The evidence is sufficient for lung cancer to have been recognized as an occupational disease in patients with silicosis in the United Kingdom. Pneumothorax is an occasional complication of silicosis, as it is of any disease associated with lung fibrosis.

Subjects with silicosis, particularly of the accelerated type, seem to be at increased risk of the development of autoantibodies and of rheumatoid disease, scleroderma, and systemic lupus erythematosus; these conditions have been described in about 10 per cent of some series of patients with silicosis. Focal glomerulonephritis has also been described in silicosis, but the cause of this is unknown.

Prevention and management

The epidemiological evidence suggests that workers exposed to levels of respirable silica in excess of 1 mg/m ^{3} have a high risk of silicosis, and that a risk may still

exist even at levels of around 0.1 mg/m^3. The United Kingdom maximum exposure limit is 0.3 mg/m^3, and industry is obliged to keep exposures of workers below this level as far as practicable, by appropriate ventilation, extraction, and other dust suppression measures. For historic reasons, quartz exposures in coal mining are controlled by total dust levels rather than the silica component of the dust. If higher levels are inevitable, the worker should wear appropriate respiratory protection, although this must be regarded as a second-best and potentially risky procedure. Once a worker has developed the disease, he should be prevented from working with silica again. The only medical management necessary is regular sputum examination for tubercle bacilli, as tuberculosis accelerates the lung damage but responds normally to modern chemotherapy. Acute silicosis would nowadays be an indication for consideration of transplantation. In the United Kingdom, workers with silicosis (whether or not complicated by lung cancer) should apply to the Respiratory Diseases Board of the Benefits Agency for industrial injuries benefits.

Asbestosis

Asbestosis is pulmonary fibrosis caused by exposure to fibres of asbestos. It was originally described in the 1900s and its importance as an occupational disease was recognized by epidemiological studies in the 1930s. However, in the first century AD, Pliny recorded that the weavers of wicks for the lamps of the vestal virgins wore masks for respiratory protection, and so some recognition of its hazards goes back to antiquity.

Asbestos is mined principally in Canada, South Africa, and the former Soviet Union. It is a generic term for a group of fibrous silicates, the most important being chrysotile (white), crocidolite (blue), and amosite (brown). Chrysotile has a serpentine configuration and breaks up into microfibrils, while the other types (amphiboles) are straight and less liable to longitudinal fracture (Fig. 10). All types are resistant to physical and chemical destruction, which gives them their commercial value in fireproofing, insulation, reinforcement of cement, weaving into cloth, bonding in brake linings and plastics, and so on. The asbestos is obtained by crushing the rock to release the fibres, which are then carded and transported in non-porous bags to the user industry.

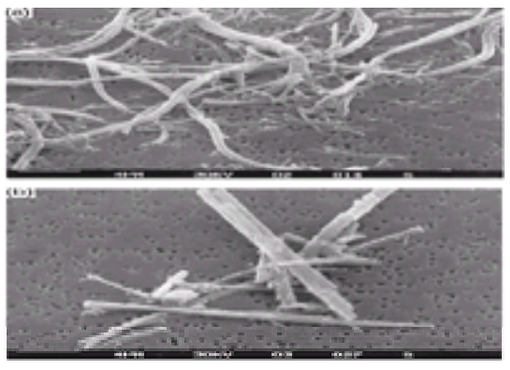

Fig. 10 Scanning electron micrographs of (a) chrysotile and (b) amosite on Millipore filters. The curly configuration and microfibrils of chrysotile should be noted. Scale bar, 4 µm.

Asbestos causes several separate pleuropulmonary lesions. The commonest are benign pleural plaques, but acute effusion and diffuse fibrosis also occur. Mesothelioma, discussed in Chapter 17.14.1), is the most important. It now occurs in over 1000 people in Britain annually and the incidence is predicted to rise further in relation to exposures some 30 years previously. It is most frequent in people who have worked in construction, ship repair, and such trades as electrician, plumber, and insulator, where regular direct or indirect exposure to asbestos has occurred. The pulmonary disease asbestosis occurs in about 100 people annually in the United Kingdom; all have worked regularly with asbestos for many years. Risk of lung carcinoma is also related to asbestos exposure, interacting multiplicatively with smoking. All these risks appear to have been greater with exposure to amphiboles than with exposure to chrysotile, but most workers (except in specific mining/production industries) have usually been exposed to a mixture of the different types.

Aetiology and pathology

The harmful asbestos fibres are those less than 3 µm in transverse diameter and greater than 10 µm in length, that is, sufficiently narrow to be inhaled to the acinus, yet too long to be removed by macrophages. Their toxicity depends on their dimensions and their persistence in lung tissue once inhaled. All types of asbestos of these dimensions can cause fibrosis and carcinoma when inhaled by rats. Moreover, injection of any asbestos type (and indeed many non-asbestos fibres) into the peritoneum of rats causes mesothelioma in a dose-related manner. The lower risk of mesothelioma in association with pure chrysotile exposure in humans is related to this fibre's curly configuration, which reduces the number penetrating the acinus, and its propensity to break up into minute short fibrils that can eventually be removed from the lung by the action of macrophages. As with coal and silica, the fibrogenicity of asbestos is probably related to damage to macrophages which are unable to cope with fibres much longer than themselves and the liberation of substances that activate fibroblasts to produce collagen. Among the substances shown to result from experimental challenge of rats with asbestos are tumour necrosis factor and macrophage- and platelet-derived growth factors.

The macroscopic appearance of an asbestotic lung is of grey fibrosis more marked peripherally and in the lower zones. In severe cases the fibrosis appears like a honeycomb. Yellow shiny parietal pleural plaques are also usually seen in the thoracic cavity, although these frequently also occur in the absence of pulmonary fibrosis. Microscopically there is diffuse alveolar wall fibrosis with minimal cellular infiltrate or desquamation of type 2 pneumocytes, initially around the centre of the acinus and later spreading to destroy the acinar structure, leading to the appearance of honeycombing. Larger asbestos fibres may be seen coated with a protein–ferritin complex (the asbestos or ferruginous bodies), while smaller fibres remain uncoated but may still just be visible with the light microscope (Fig. 11). However, for every fibre visible by light microscopy, several hundred uncoated fine fibres can always be found on electron microscopy. Pleural plaques have the appearance of basket-weave collagen, and fibres are almost never seen within them.

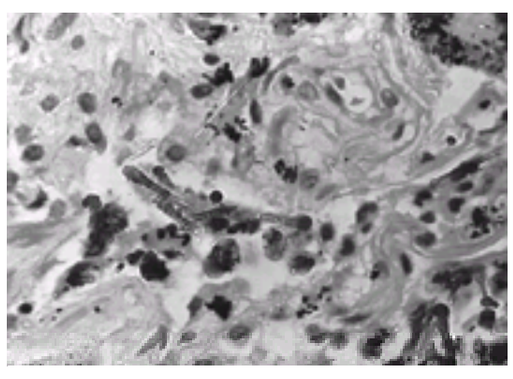

Fig. 11 Histological appearance of asbestosis, with interstitial fibrosis, asbestos bodies, and several uncoated fibres.

Clinical features

Asbestosis occurs in people exposed regularly over years to airborne asbestos as a result of the material being used or removed, and not as a result of occasional exposure. It is more likely to be seen in trades involving the application or removal of asbestos in lagging and insulation than in asbestos mining, preparation, or weaving, where control of fibre levels is more careful. The disease may first become apparent and progress after exposure has ceased.

The symptoms are shortness of breath, initially on exertion, and dry cough. Repetitive end-inspiratory basal crackles commonly precede symptoms, and finger clubbing may occur later. The disease is usually progressive, the speed of progression being related to the dose of asbestos to which the lungs have been subjected, and results in increasing disability and death from cardiorespiratory failure. Forty to fifty per cent of smokers with asbestosis die of bronchial carcinoma, but there is

no increased risk of tuberculosis.

The radiological appearance of asbestosis is identical to that of cryptogenic pulmonary fibrosis —predominantly basal and peripheral irregular linear shadowing progressing to honeycombing (Fig. 12). The presence of pleural plaques, which frequently calcify, is an indication of asbestos exposure and may help in the differential diagnosis (Fig. 13). In advanced asbestosis the fibrosis obscures the cardiac border, giving a shaggy appearance. The radiological appearances are best described by comparison with the International Labour Organization standard radiographs. CT scans are useful in differentiating asbestosis from diffuse pleural fibrosis, which may mimic it clinically and radiologically (Fig. 14).

Fig. 12 Radiograph of a lagger with asbestosis. The irregular basal and middle zone fibrosis should be noted.

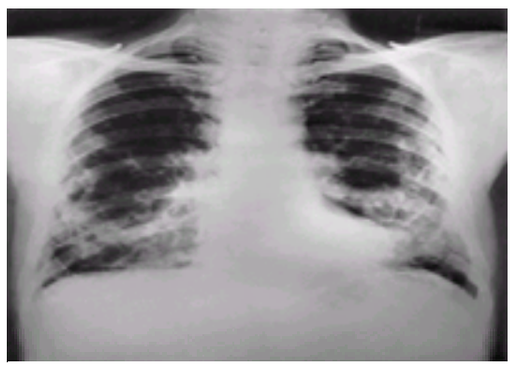

Fig. 13 Radiograph of a lagger, showing extensive calcified pleural plaques.

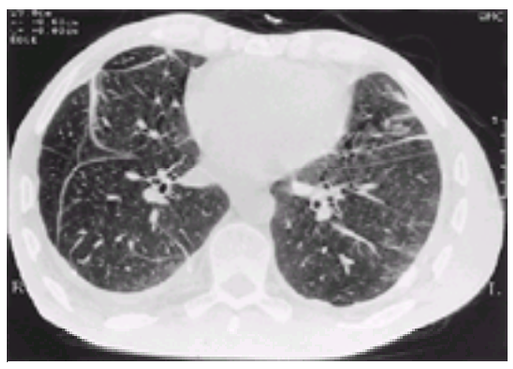

Fig. 14 CT scan of a thermal insulator, showing bilateral pleural fibrosis with areas of calcification and fibrous strands extending into peripheral lung. This does not imply the same functional effects or prognosis as asbestosis.

Asbestosis causes a restrictive pattern of lung function, with reduced volumes and transfer factor. These measurements are the most suitable for screening for the disease and following its progress. Pulmonary compliance is reduced in relation to the extent of the fibrosis, and arterial oxygen desaturation occurs in the later stages.

Pleural plaques cause no symptoms and are usually a coincidental finding on chest radiography. Diffuse bilateral pleural thickening, often calcified, which can cause breathlessness and restricted lung volumes, occurs infrequently. Inspiratory crackles may be heard over this in the absence of significant asbestosis. Very uncommonly, benign pleural effusion may occur. This develops within the first two decades after exposure as a transient haemorrhagic effusion and is diagnosed by the exclusion of infective and malignant causes. There is no evidence that any of these benign disorders predisposes to pleural mesothelioma, the risk of which relates to the prior extent of asbestos exposure.

Prevention and management

The prevention of asbestosis, as of other pneumoconioses, depends on reducing the exposure of individuals to fibre levels that have been shown to be insufficient to cause the disease in a lifetime of exposure. Unfortunately, the difficulties of making valid measurements of airborne fibres and the uncertainties attached to the early diagnosis of asbestosis have prevented the formulation of really reliable evidence on which to base a standard. The present British standard for chrysotile of 0.5 respirable fibres/ml has been based on work that suggests such levels would, when breathed over a working lifetime, result in asbestosis in fewer than 1 per cent of those exposed. The corresponding standard for amphiboles is 0.2 fibres/ml. Many industries have now introduced other fibrous or crystalline minerals in place of asbestos. The potential of such new materials to cause similar diseases depends on their fibre dimensions, solubility in tissue, and the concentrations achieved in the workplace air. It is important that they should be handled with appropriate care by industry.

Regular medical and radiological examination of asbestos workers is essential for the early detection of asbestosis, and there is some evidence that removal of the worker from exposure at this stage is associated with slower progression. Workers should also be advised not to smoke in view of the interaction between cigarettes and asbestos in causing lung cancer. Once asbestosis is suspected, the British worker should apply to the Benefits Agency for assessment for industrial injuries benefit. Diffuse pleural fibrosis also attracts benefits, as does lung cancer in the presence of asbestosis or pleural fibrosis, but pleural plaques do not.

Risks of asbestos-related disease in the non-occupationally exposed population

Much anxiety has been engendered amongst the general public by media interest in asbestos, and doctors may find themselves being asked about, for example, the risks to children of asbestos wall panelling in houses or asbestos inserts in ironing boards. In general it can be stated that asbestosis only occurs in people working regularly with asbestos for years. However, this has included, at least in the past, wives washing the dusty clothes of asbestos workers and people who have lived or worked near polluting asbestos factories. Occasional or incidental exposure to asbestos can be dismissed as a significant cause of asbestosis. Similarly, lung cancer risks seem to be significantly increased only with the doses of asbestos that lead to asbestosis, and individuals who do not smoke and who only have asbestos fittings in their houses can be reassured that their risks of this disease are negligible. Mesothelioma however, while also dose-related, occurs after smaller exposures and it

is well established that a sufficient dose of crocidolite or amosite can be inhaled in a period of intense exposure of a few months. Of the 1000 cases occurring in the United Kingdom each year, almost all individuals give a history of having worked in a trade known to have been associated with asbestos use and have large numbers of fibres in their lungs, suggesting that employment rather than incidental exposure has been responsible. Small and occasional exposures to asbestos are highly unlikely to entail an important risk, but if regular exposures are thought to be occurring in the domestic or general environment, steps should be taken to eliminate them.

Other silicate pneumoconioses

Several silicates apart from asbestos are of commercial importance, and some of these have been shown to cause pneumoconiosis. Talc (hydrated magnesium silicate) is mined as soapstone in the United States, China, and the Pyrenees. It is milled and has many uses including in cosmetics, the rubber industry, paints, ceramics, and pharmaceuticals. Kaolin (hydrated aluminium silicate) is quarried in south-west England, Georgia in the United States, Japan, Egypt, Germany, and former Czechoslovakia. It is used mainly in the manufacture of ceramics, paper and paint, and in pharmaceuticals. Fuller's earth (calcium montmorillonite) is an absorbent clay quarried in England, the United States, and Germany. It was originally used in fulling or removing grease from wool, and is now used in oil refining and bonding foundry moulds. Mica is a complex aluminium silicate occurring in two forms—muscovite and phlogopite. The former is mined in the United States and India and used in fire-resistant windows and the manufacture of paper and paint. Phlogopite, mined in Canada, is used in the electrical industry because of its resistance to heat and electricity.

Two widely used silicate materials—cement and vitreous fibres—are not established as causes of pulmonary disease. Although cement exposure has occasionally been reported to be associated with pneumoconiosis, the evidence for this is flimsy. It is often mixed with asbestos, and asbestosis may occur in its production. Artificial vitreous fibres (glass wool and rock wool) have not so far been shown to cause pulmonary fibrosis or neoplasia in humans exposed to them, although mesothelioma has been produced by intraperitoneal injection in rats.

Other pneumoconioses

Talc pneumoconiosis

Talc is commonly contaminated with tremolite, a non-commercially exploited amphibole asbestos, and with silica. It has been difficult to disentangle the effects of these components. The disease appears clinically to resemble asbestosis, with finger clubbing and basal crackles, although radiological descriptions emphasize lesions predominantly in the middle zones with nodular as well as reticular components. Progressive massive fibrosis has been described.

Talc has also been shown to be associated with pulmonary disease in a number of other circumstances. Bronchoconstriction may occur in children exposed to high concentrations and drug users may have granulomatous reactions in the lungs as a result of either intravenous injection or inhalation of ground-up tablets. Fortunately, the widespread use of talc for producing pleurodesis has not been shown to be associated with the later development of mesothelioma, probably because the grades of talc used have not been contaminated with tremolite.

Kaolin pneumoconiosis

Kaolin causes a pneumoconiosis similar to coal-worker's pneumoconiosis with small discrete nodular lesions initially and a tendency to produce massive fibrosis. It has been described in workers involved in the drying and milling processes in the production of china clay. Kaolin may also have been the component of the dust responsible for pneumoconiosis in the now defunct Scottish shale oil industry. There is no evidence linking kaolin pneumoconiosis with carcinoma or tuberculosis.

Fuller's earth pneumoconiosis

This condition has been described in workers extracting this clay mineral. It is a benign nodular pneumoconiosis similar in pathological and radiological appearance to simple coal-worker's pneumoconiosis; progressive massive fibrosis has not been described.

Mica pneumoconiosis

A few reports of radiological change in those exposed to ground mica have been recorded, but there is no recent publication describing pathological or clinical features.

Fibrous erionite

Exposure to this fibrous hydrated aluminium silicate occurs in certain areas of Turkey and probably elsewhere in the Middle East. The populations of several villages have been exposed for many generations as they use local erionite rock as stucco and whitewash in their homes. Pleural plaques, pulmonary fibrosis, and both lung cancer and mesothelioma are endemic in these villages. Fibrous erionite has no general commercial use, but this episode illustrates the potential dangers of inhaling fine fibrous material, whether asbestos or some other mineral.

Berylliosis

Beryllium is a metal that is used in alloys for the nuclear industry and in the production of X-ray tubes. It was used in ceramics, metallic alloys, and fluorescent lights until its toxicity was recognized and it was replaced by other materials. It is mined as an ore mostly in South America and extracted by chemical processes.

Beryllium is highly toxic when inhaled, and may also cause granulomatous ulcers on contact with the skin. Inhalation of high concentrations causes an acute pneumonitis and tracheobronchitis, which can be fatal. Chronic berylliosis, which may occur as a sequel to acute exposure, usually follows more prolonged exposure to lower levels. It is not common in the United Kingdom, where no more than about 50 cases have been diagnosed, but it has been recorded much more frequently in the United States. Reported cases have occurred in beryllium workers, in wives exposed to dust from their husbands' clothes, and in people living near the factories.

The patient with chronic berylliosis presents with cough and shortness of breath. The features mimic those of sarcoidosis: bilateral pulmonary mottling with upper lobe fibrosis is the usual radiographic feature initially, with bilateral hilar lymphadenopathy being less common. The disease typically progresses to diffuse fibrosis (Fig. 15), but the rate of progression is very variable. The functional lesion is a restrictive pattern with a low transfer factor. The progress of the disease can be controlled with corticosteroid therapy, but this needs to be continued indefinitely in most cases.

Fig. 15 Radiograph of a beryllium refiner worker, showing the diffuse fibrosis of berylliosis.

The pathological lesion is identical with that of sarcoidosis, with non-caseating granulomas and varying amounts of interstitial fibrosis. The diagnosis is made on the basis of a history of exposure, compatible clinical and histological features, and a negative Kveim test. A skin-patch test is inadvisable as it can cause sensitization.

Berylliosis is prevented by keeping exposures below the threshold limit value (2 ng/m ^{3}), although as it is a hypersensitivity disease even this will not prevent all cases. Efficient respiratory protection should also be provided.

Less common pneumoconioses

Many other pneumoconioses have been described, although most are of very limited prevalence and are relatively benign. Haematite lung, occurring in iron ore miners, used to be seen in Cumbria in the United Kingdom; it is a fibrotic reaction to a mixed dust containing silica and iron. Radiographically it resembles silicosis and pathologically only differs from it in that the lungs are coloured red. There was an increased risk of lung cancer, probably due to radiation in the mines. Closely related to haematite lung is siderosis, a benign iron oxide pneumoconiosis occurring in welders and other workers in iron foundries. The radiological lesions often regress after exposure ceases. Barium processing and tin refining may be associated with the development of dramatic radiological nodular shadowing—baritosis and stannosis, respectively. These are also benign conditions, the radiological appearances reflecting radio-opaque dust in macrophages. Pneumoconiosis associated with diffuse lung fibrosis has been described in work with aluminium oxide (Shaver's disease) and tungsten carbide (hard metal disease). This latter condition, which is probably a hypersensitivity reaction to cobalt in cooling liquids, may also present with features of asthma or allergic alveolitis. A pneumoconiosis resembling that in coal miners has been described in workers with graphite and other forms of carbon, and in shale miners. A benign pneumoconiosis, consisting of simple accumulations of dust and macrophages with minimal nodular radiological shadowing, has also been described in workers producing polyvinyl chloride.

Further reading

Henderson VL, Enterline PE (1979). Asbestos exposure: factors associated with excess cancer and respiratory disease mortality. *Annals of the New York Academy of Sciences* **330**, 117–26.

Hurley JF *et al.* (1982). Coalworkers' simple pneumoconiosis and exposure to dust at 10 British coalmines. *British Journal of Industrial Medicine* **39**, 120–7.

International Labour Organization (1980). *Guidelines for the use of ILO International classification of radiographs of pneumoconioses*. Occupational Safety and Health Series No. 22 (rev. 87), International Labour Organization, Geneva.

Marine WM, Gurr D, Jacobsen M (1988). Clinically important respiratory effects of dust exposure and smoking in British coal miners. *American Review of Respiratory Disease* **137**, 106–12.

Morgan WKC, Seaton A (1995). *Occupational lung diseases*, 3rd edn. WB Saunders, Philadelphia.

Mossman BT *et al.* (1990). Asbestos: scientific developments and implications for public policy. *Science* **247**, 294–301.

Peto J *et al.* (1995). Continuing increase in mesothelioma mortality in Britain. *Lancet* **345**, 535–9.

Seaton A (1990). Coalmining, emphysema and compensation. *British Journal of Industrial Medicine* **47**, 433–5.

Seaton A (1998). The new prescription: industrial injuries benefit for smokers? *Thorax* **53**, 335–6.

Seaton A, Cherrie JW (1998). Quartz exposure and severe silicosis: a role for the hilar nodes. *Occupational and Environmental Medicine* **55**, 383–6.

Seaton A *et al.* (1991). Accelerated silicosis in Scottish stonemasons. *Lancet* **337**, 341–4.

17.11.8 Pulmonary haemorrhagic disorders

*D. J. Hendrick and G. P. Spicketi**

Bleeding within the lung and subsequent haemoptysis is common in clinical practice and may be the consequence of many unrelated disorders. What then is the value of the term 'pulmonary haemorrhagic disorder'? The answer lies with its use in special circumstances only—bleeding arising diffusely from pulmonary alveolar capillaries. A preferable and more explicit diagnostic term is therefore pulmonary capillary (or alveolar) haemorrhage. This is not a disease entity of itself, merely a feature of several diseases, but most notable in two conditions, Goodpasture's syndrome and idiopathic pulmonary haemosiderosis.

While the lung can accommodate only small quantities of blood in the major airways without threatening life from asphyxiation, it can sequester surprisingly large amounts (litres) at alveolar level. This leads to a curious characteristic, unique among diffuse parenchymal diseases of the lung and of considerable diagnostic value; the carbon monoxide gas transfer (*TL*CO) is raised significantly above normal. Not only are physiologically useful red cells within the alveolar capillaries able to absorb the inhaled carbon monoxide, but so too are those lost from the circulation into the alveolar spaces.

Pulmonary capillary haemorrhage is thus characterized by haemoptysis, breathlessness, diffuse air space shadowing on the chest radiograph (Fig. 1), anaemia (normochromic normocytic if acute, iron deficient with chronicity), and an elevated *TL*CO (see Chapter 17.3.2). The extravasated red cells are not readily expectorated, although enough generally escape to cause haemoptysis, and so haemosiderin accumulates within alveolar macrophages as the red cells and their debris are engulfed. When haemosiderin-laden macrophages are identified in sputum, the diagnosis of pulmonary capillary haemorrhage is largely confirmed, but if sputum is not expectorated or haemoptysis is absent, minimal, or otherwise explained, then bronchoalveolar lavage and/or lung biopsy are often necessary to establish the diagnosis. An alternative approach is CT and MRI, which may alone provide convincing evidence of blood sited diffusely within the alveoli.

Fig. 1 Radiograph showing gross alveolar shadowing following major pulmonary haemorrhage in a 60-year-old man with systemic vasculitis.

While diffuse alveolar capillary haemorrhage may characterize or complicate a wide variety of specific diseases or disease settings (each associated with its particular range of clinical, diagnostic, and mechanistic features, and management options), the direct effects of the haemorrhage itself are not influenced by the cause, nor are the means by which it can be recognized. There may, nevertheless, be substantial differences at presentation from case to case according to severity and chronicity.

Goodpasture's syndrome

Goodpasture described a man with renal failure, glomerulonephritis, and pulmonary haemorrhage. A number of conditions can cause such a 'pulmonary–renal syndrome', the best characterized of which (although almost certainly not the illness suffered by the patient in the original report) is now termed Goodpasture's disease, which consists of diffuse pulmonary haemorrhage and glomerulonephritis with linear deposition of antibody (90 per cent of which are directed against the a-3 chain of type IV collagen) along the glomerular basement membrane. Goodpasture's disease is described in Chapter 20.7.9 and other causes of pulmonary–renal syndrome in Chapter 20.10.3.

Clinical features

In practice, glomerulonephritis proves to be a much commoner threat to survival than lung haemorrhage, and the diagnosis of Goodpasture's disease is reached more conveniently by serological testing (for anti-GBM antibodies) and from kidney rather than lung biopsy. In some cases, however, lung disease dominates the clinical picture, when the majority of patients are male smokers and some have recent exposure to volatile hydrocarbons; case reports have additionally identified recent exposure to chlorine and smoked cocaine. This suggests that when there is susceptibility, inhaled toxic agents enhance pulmonary endothelial damage and thus allow the initiation of autoimmunity or the ready access of existing autoantibody to basement membrane. Respiratory presentation is with cough, breathlessness, and haemoptysis, which is intermittent and ranges from occasional streaks to massive fatal bleeding. Systemic symptoms of fever, joint pains, or weight loss are unusual. The chest radiograph shows patchy or diffuse shadowing due to intra-alveolar blood, usually resolving over the course of 2 weeks unless there is further bleeding. At the time of bleeding there may be arterial hypoxaemia and reduced lung volumes. Serial measurement of *TL*CO can be used to monitor progression, and prolonged bleeding may lead to iron-deficiency anaemia. Renal function may be normal initially and then deteriorate over days to weeks. Steroids, other immunosuppressant drugs (cyclophosphamide in particular), and plasmapheresis are all used (in some circumstances) to control renal disease, and are additionally helpful in treating pulmonary haemorrhage. Patients should not smoke and should avoid hydrocarbon exposure.

Idiopathic pulmonary haemosiderosis

This is a rare disorder of children and young adults in which there is recurrent alveolar bleeding in the absence of kidney disease. Anti-basement membrane antibody has not been detected, and nor have any other substantial immunological clues to the causal mechanism, although serum IgA is commonly elevated. The electron microscopic appearance of the basement membrane shows no consistent abnormality. The alveolar blood may provoke a fibrogenic stimulus and the development of diffuse pulmonary fibrosis, as may recurrent alveolar bleeding from mitral stenosis and chronic severe left ventricular failure.

Although termed idiopathic pulmonary haemosiderosis, the condition is associated with premature birth and an increasing number of environmental exposures. One such that has incited particular interest is to the mould, *Stachybotrys*, which may contaminate wet or damp accommodation, and which releases a particularly potent toxin with haemorrhagic properties. This is now thought to have aetiological significance in some childhood cases, perhaps in synergy with environmental tobacco smoke. Associations with cow's milk allergy, rheumatoid arthritis, and coeliac disease are also recognized, but the latter might be a consequence of cow's milk allergy also rather than gluten intolerance. A number of other environmental causes have been suggested; but the stronger the evidence for their causal roles, the less appropriate is the diagnostic rubric, idiopathic pulmonary haemosiderosis. They are consequently identified below under the heading 'other causes'.

Clinical features

Recurrent alveolar haemorrhage is generally manifested by cough with haemoptysis and breathlessness, but haemoptysis is not invariably present, and in children a failure to thrive may be prominent, together with the effects of severe chronic iron-deficiency anaemia. Acute bleeds are more common in childhood and may be life threatening. Physical examination is unhelpful. The chest radiograph and CT scan show the non-specific appearances of intra-alveolar blood, which usually clear spontaneously over 1 to 3 weeks. With chronicity, the appearances of diffuse pulmonary fibrosis with honeycombing may supervene. Lung function tests then show a progressive loss of volume and reduction of gas transfer; an obstructive pulmonary defect occurs occasionally, which is unexplained.

Treatment and prognosis

Supportive treatment is required during acute bleeding, and artificial ventilation is occasionally necessary. There are case reports recording responses to the avoidance of milk and gluten, and to the use of immunosuppressive agents including corticosteroids and cyclophosphamide. Some patients recover spontaneously with or without residual pulmonary damage.

Other causes of diffuse alveolar haemorrhage

Although diffuse alveolar haemorrhage is not a principal pulmonary feature of disorders other than Goodpasture's syndrome and idiopathic pulmonary haemosiderosis, it may occur with or complicate a wide variety of disorders with immunological, vasculitic, vascular, haemostatic, toxic, or unknown origins. In many of the cases that have been reported, several different disorders, their complications, and their various treatments could all have played a contributory role.

Vasculitic disorders can occasionally cause prominent diffuse alveolar haemorrhage, particularly Wegener's granulomatosis. This may simulate Goodpasture's disease, since it commonly causes acute necrotizing glomerulonephritis, but is distinguished from it clinically by the common involvement of upper respiratory tract structures (and because diffuse alveolar haemorrhage is an uncommon respiratory manifestation); histologically by the appearances of a granulomatous vasculitis; and immunologically by the presence of circulating anti-neutrophil cytoplasmic antibodies (**ANCA**), directed against proteinase-3, in about 90 per cent of cases. Other vasculitic disorders involving the lung are very rare causes of diffuse alveolar haemorrhage; they include Henoch–Schönlein purpura and Churg–Strauss syndrome. The latter may be associated with ANCA directed against myeloperoxidase and against eosinophil granule enzymes.

Diffuse alveolar haemorrhage may also arise as an unusual respiratory feature of several non-vasculitic immunological disorders. Most prominent is systemic lupus erythematosus (in which lupus anticoagulant, thrombcytopenia, and active nephritis may all play a role), but there are reports also of diffuse alveolar haemorrhage complicating antiphospholipid antibody syndrome, IgA nephropathy, idiopathic membranous nephropathy, scleroderma, renal and bone marrow transplantation, and chronic active hepatitis.

Other reports have implicated hymenopteran stings, moulds other than *Stachybotrys* and their mycotoxins, infections (group A streptococcal, leptospirosis, strongyloidiasis, *Stenotrophomonas*, dengue fever, cytomegalovirus, AIDS, varicella), occupational exposure to tri- and pyro-mellitic anhydride, lymphangiography contrast media, and several medications (valproate, nitrofurantoin, mitomycin C, azathioprine, D-penicillamine, surfactant therapy, anaesthetic agents). The list is completed by causes of chronic pulmonary venous congestion (mitral stenosis, chronic left ventricular failure, pulmonary veno-occlusive disease), malignant hypertension, and disorders that disrupt bleeding and coagulation mechanisms (thrombocytopenia, leukaemia, thrombinolytic therapy, platelet glycoprotein IIb/IIIa inhibitor, anticoagulant poisoning, factor V deficiency). Combinations of such factors are commonly found in individual cases, and it may be that important interactions occur, without which the probability of diffuse haemorrhage is remote. Although capillary stress from high pressure gradients is thought to be a major factor underlying diffuse pulmonary haemorrhage in exercizing horses (and camels), it appears a rare or unheard cause in most other species. Nevertheless, the use of negative pressure ventilation in humans has been reported to have a similar effect.

---

*Dr D. J. Lane wrote on this subject in the third edition of the *Oxford Textbook of Medicine*. Much of his text has been retained in this revision and we acknowledge his contribution with grateful thanks.

Further reading

Anonymous (2000). From the Centers for Disease Control and Prevention. Update: pulmonary hemorrhage/hemosiderosis among infants—Cleveland, Ohio, 1993–1996. *Journal of the American Medical Association* **283**, 1951–3.

Bhandari V *et al.*(1999). Pulmonary hemorrhage in neonates of early and late gestation. *Journal of Perinatal Medicine* **7**, 369–75.

Colombo JL, Stolz SM (1992). Treatment of life threatening pulmonary hemosiderosis with cyclophosphamide. *Chest* **102**, 959–60.

Leatherman JW, Davies SF, Hoidal JR (1984). Alveolar haemorrhage syndromes; diffuse and microvascular lung haemorrhage in immune and idiopathic disorders. *Medicine, Baltimore* **63**, 343–61.

Pacheco A *et al.* (1991). Long term follow-up of adult idiopathic pulmonary hemosiderosis. *Chest* **99**, 1525–6.

Peters DK *et al.* (1982). Treatment and prognosis of anti-basement membrane antibody mediated nephritis. *Transplant Proceedings* **14**, 513–21.

Ryan JJ *et al*. (1998). Recombinant alpha-chains of type IV collagen demonstrate that the amino terminal of the Goodpasture autoantigen is crucial for antibody recognition. *Clinical and Experimental Immunology* **113**, 17–27.

Vats KR *et al* (1999). Henoch–Schönlein purpura and pulmonary hemorrhage: a report and literature review. *Pediatric Nephrology* **13**, 530–4.

Weishaupt D *et al*. (1999). Pulmonary hemorrhage: imaging with a new magnetic resonance blood pool agent in conjunction with breathheld three-dimensional magnetic resonance angiography. *Cardiovascular and Interventional Radiology* **22**, 321–5.

17.11.9 Eosinophilic pneumonia

*D. J. Hendrick and G. P. Spicketi**

When alveolar spaces are consolidated because of eosinophil inflammation/infiltration, there is said to be eosinophilic pneumonia. This is not meant to imply that there is microbial infection, and most commonly there is not. There is characteristically an accompanying eosinophilia of peripheral blood, hence the alternative terms, pulmonary eosinophilia and pulmonary infiltrates with eosinophilia (PIE syndrome). Eosinophilic pneumonia is the preferred term, however, since eosinophilia of peripheral blood may be present coincidently when eosinophils are not relevant to a pulmonary infiltrate, and conversely true eosinophilic pneumonia is occasionally not associated with blood eosinophilia. To avoid further confusion, it should be noted that eosinophilic granuloma is a distinct disease unrelated to eosinophilic pneumonia: it is a disorder of Langerhans (dendritic) cells and not characterized by eosinophil infiltration of the alveolar spaces.

The plethora of diagnostic terms is exceeded by the multitude of causes, and evenly matched by the systems of classification that have been suggested. In essence they reflect the following points concerning eosinophilic pneumonia.

1. It may arise acutely and resolve quickly (acute eosinophilic pneumonia, Löffler's syndrome, simple pulmonary eosinophilia).
2. It may arise gradually and persist for many months, leading sometimes to pulmonary fibrosis (chronic eosinophilic pneumonia, prolonged pulmonary eosinophilia).
3. It may be a consequence of allergy, particularly to bloodborne parasites (tropical eosinophilia), moulds (allergic bronchopulmonary mycosis), or other common environmental allergens.
4. It is often associated with asthma (asthmatic eosinophilia).
5. It is often due to drugs (whether through allergy, idiosyncrasy, or toxicity).
6. It may be associated with pulmonary vasculitis (Churg–Strauss syndrome, polyarteritis nodosa, Wegener's granulomatosis).
7. It may be a component of the hypereosinophilic syndrome.
8. It may be associated with a variety of other distinct disease entities (rheumatoid disease, sarcoidosis, T-cell lymphoma, Hodgkin's disease, shock, and the adult respiratory distress syndrome).
9. It may seem to be idiopathic.

Since there is often overlap—to give a recent example, the affected subject may have asthma, be taking a relevant drug (a leukotriene receptor antagonist), have prolonged manifestations, and have a pulmonary vasculitis (Churg–Strauss syndrome)—there is limited benefit from using any classification. The important issue is to identify any potentially remediable cause.

Diagnosis

In practice the finding of a blood eosinophilia in association with a radiographic pulmonary infiltrate provides a valuable clue that pneumonia of infectious origin may not be the explanation. Since such a disorder is not likely to respond to conventional antibiotic medication, the need to confirm or exclude the possibility of eosinophilic pneumonia will soon arise. Equally, if an apparent pneumonia fails to respond to antibiotics, a blood eosinophil count should be obtained. Once suspected, eosinophilic pneumonia is most conveniently confirmed by demonstrating an excess of eosinophils in bronchoalveolar lavage fluid in the absence of pathogenic micro-organisms. Sometimes sputum alone is sufficient, whether expectorated spontaneously or induced. Alternatively, an excess of alveolar eosinophils is revealed in lung biopsy tissue. The use of CT scanning has shown that episodes of recurrent pulmonary infiltration occur, not surprisingly, more frequently than can be detected from plain chest radiographs in subjects with confirmed eosinophilic pneumonia.

Once eosinophilic pneumonia is confirmed, a variety of possible causes should be considered before it is assumed to be idiopathic in origin and before empirical treatment with corticosteroids is administered.

1. Is there parasitic infestation?
2. Have any drugs been administered?
3. Is there asthma?
4. Is there evidence of allergy to parasites or drugs?
5. Is there evidence of allergic bronchopulmonary mycosis (particularly aspergillosis)?
6. Is there evidence of vasculitis?
7. Is there evidence of the hypereosinophilic syndrome?
8. Is there evidence of other disorders known to be associated with eosinophilic pneumonia?

Treatment

Eosinophilic pneumonia itself often responds well to corticosteroid medication, though treatment may need to be prolonged (6 months or more) in the chronic forms of the disorder. The importance of identifying whether it is associated with the causal factors listed above lies with the additional need to treat these also. Otherwise eosinophilic pneumonia may not respond adequately to steroid therapy and the associated diseases may produce other manifestations.

Particular varieties of eosinophilic pneumonia

Löffler's syndrome (acute eosinophilic pneumonia, simple pulmonary eosinophilia)

The essential features of the syndrome are transitory migratory pulmonary shadows with associated modest peripheral eosinophilia in patients with a mild self-limiting illness. Some cases are asymptomatic and discovered incidentally. Most patients present with cough, sometimes with oddly yellowish sputum containing an abundance of eosinophils, and a few have general malaise and a mild fever. The pulmonary shadows reflect fan-shaped areas of consolidation, often peripheral and sometimes rather nodular, which last a few days only and appear haphazardly in various lobes, seldom following a truly segmental pattern. In some cases they are single and in others they are multiple. The peripheral eosinophilia is obvious but rarely gross; a differential of more than 20 per cent in a modestly raised total white cell count is unusual and more often the absolute eosinophil count ranges between 1×10^{9} and 2×10^{9}/l (normal: $< 0.4 \times 10^{9}$/l).

Patients who develop Löffler's syndrome are often atopic and may have other manifestations of an atopic diathesis such as asthma, urticaria, and angio-oedema. Allergy has been shown since Löffler's original description to play an important role, and cases can be seen to fall into two broad aetiological groups with a third miscellaneous group of unexplained aetiology.

In the first group eosinophilic pneumonia represents an allergic reaction to bloodborne parasites migrating through the lung, particularly larvae of *Ascaris lumbricoides* and occasionally *A. suum. Ancylostoma*, *Trichuris*, *Trichinella*, *Taenia*, and *Strongyloides* species provide further examples.

Drugs form the second major aetiological group. Löffler's syndrome is well described after administration of *p*-amino salicylic acid, aspirin, sulphonamides (including

the antimalarial combination sulphadiazine and pyrimethamine or Fansidar), penicillin, and imipramine. It may also occur with nitrofurantoin (although this can also give a diffuse reticulonodular radiological picture and is a cause of the more chronic type of eosinophilic pneumonia), toxic smoke, and lymphangiography contrast medium.

Successful management requires the irradication of any parasites or the cessation of relevant medication, as well as the administration (if necessary) of oral corticosteroids.

Tropical eosinophilia

Eosinophilic pneumonia in tropical climates is often a consequence of migrating larvae of the filarial worms *Wucheria bancrofti* and *Brugia malayi*. The effects are fundamentally similar to those of Löffler's syndrome, but tend to be more persistent and more serious, are more often associated with asthma, and may be associated with systemic symptoms of weight loss, persistent fever, and lymphadenopathy. Also the peripheral eosinophil count tends to be greater than in Löffler's syndrome (> 3×10^9/l), and the total serum IgE level is markedly elevated. With chronicity, pulmonary fibrosis is characteristic. A cure is to be expected with antifilariasis medication (diethylcarbamazine).

Chronic eosinophilic pneumonia (prolonged pulmonary eosinophilia)

Eosinophilic pneumonia persisting for more than a month is distinguished from the more transitory Löffler's syndrome, although its clinical characteristics are fundamentally similar. As with eosinophilic pneumonia associated with tropical filariasis, it tends to be more persistent and more serious than Löffler's syndrome (it is sometimes life threatening), and to be associated with systemic symptoms (particularly fever), progressive pulmonary fibrosis, and fixed airway obstruction. It may last for several months and be associated additionally with eosinophilic pleural effusion, focal skin lesions, atopic manifestations such as rhinitis, sinusitis, and angio-oedema, hepatosplenomegaly, and even hepatic necrosis. The pulmonary disease is often extensive, causing dyspnoea and hypoxia, and is characterized radiologically by a curious peripheral distribution dubbed a 'negative photographic image of pulmonary oedema'. The radiological abnormalities tend to recur and last for weeks or months, and like the shadows of Löffler's syndrome may vary in site during the course of the illness.

Chronic eosinophilic pneumonia is more commonly idiopathic (cryptogenic) than Löffler's syndrome, but may also be a consequence of parasite infestation (e.g. tropical filariasis) or drug hypersensitivity. Case reports over recent years have identified aminoglutethimide, bicalutamide, chlorpropamide, clomipramine, dapsone, diflunisal, ethambutol, mesalazine, minocycline, nitrofurantoin, sertraline, sotalol, sulphonamides, and venlafaxine as possible causes. Peripheral blood eosinophilia is less consistent with chronic compared with acute forms of eosinophilic pneumonia, although is often of greater level (> 1×10^9/l). When a definitive cause is identified, appropriate specific management should follow, but often no cause is evident and oral corticosteroid therapy should be given. Responses are often dramatic, but recurrences are common if treatment is discontinued within 6 to 12 months. There may be a persistent mixed obstructive and restrictive loss of ventilatory function, and radiographic evidence of persistent pulmonary fibrosis.

Eosinophilic pneumonia with asthma

Eosinophilic pneumonia is commonly associated with asthma, even in the absence of parasite infestation or drug hypersensitivity. Two particular associations are noteworthy.

Allergic bronchopulmonary mycosis

When fungal hypersensitivity develops in atopic subjects with asthma, additional manifestations may occur in the lung: these include eosinophilic pneumonia, mucoid impaction, bronchiectasis, and pulmonary fibrosis. The ensuing syndrome of allergic bronchopulmonary mycosis occurs most commonly with *Aspergillus fumigatus*, though has been reported with other *Aspergillus*, *Candida*, *Curvularia*, and *Helminthosporium* species. It accounts for most cases of eosinophilic pneumonia with asthma in the United Kingdom and is best considered a complication of atopic asthma, appearing to result from airway colonization by the relevant mould. The mechanism, however, is clearly one of hypersensitivity, not infection/invasion, and both IgE and IgG antibodies are necessary to support its diagnosis.

In acute phases, there is patchy obstruction of bronchi with inspissated mucus that, if expectorated, appears as brown rubbery lumps in the sputum (plugs). Fungal hyphae may be recovered from them, indicating fungal growth has occurred within the airway. This impaction of mucus in one or more bronchi leads to patchy atelectasis within, or of, segments (even lobes) and is often associated with eosinophilic pneumonia. The radiographic appearances are of fleeting pulmonary infiltrates (Fig. 1). It usually responds well to corticosteroids, a useful diagnostic feature being the expectoration of plugs during this period of resolution. In the medium term the involved bronchi (generally proximal) may become bronchiectatic, leading in turn to the characteristic features of bronchiectasis (productive cough, intermittent haemoptysis). In the longer term, pulmonary fibrosis may ensue, particularly in the upper lobes and apices, so that the radiographic appearances resemble tuberculosis, and if mucoid impaction and/or eosinophilic pneumonia become superimposed, the radiographic appearances may simulate active tuberculosis very closely. Suspicion of tuberculosis in an individual with atopic asthma should always prompt consideration of allergic bronchopulmonary mycosis.

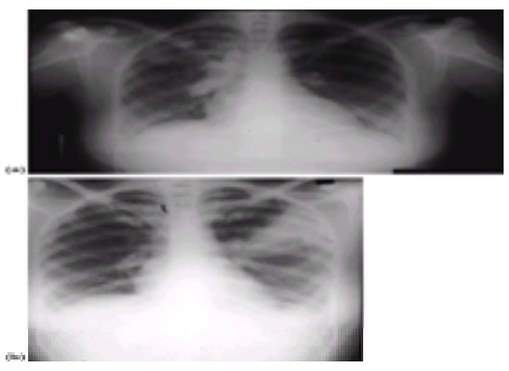

Fig. 1 Allergic bronchopulmonary aspergillosis: two radiographs taken 6 months apart from an East African woman with asthma, peripheral eosinophilia, and high titres of IgE and precipitating IgG antibodies to *Aspergillus fumigatus*.

Churg–Strauss syndrome

A much rarer association of eosinophilic pneumonia with asthma is that involving Churg–Strauss syndrome, a vasculitic and granulomatous disorder that commonly involves lungs, gut, peripheral nerves, skin, and kidneys. It is characterized typically by asthma, eosinophilic pneumonia, and very high numbers of circulating eosinophils (> 5×10^9/l), but the pulmonary manifestations may additionally include localized haemorrhage and haemoptysis. Serological investigation may also demonstrate raised serum levels of IgE and eosinophil cationic protein, P-ANCA (peripheral antineutrophil cytoplasmic antibodies) with myeloperoxidase activity (in 60 per cent of cases), and C-ANCA with proteinase-3 specificity (in 10 per cent). Other autoantibodies against eosinophil granule enzymes have also been described. Pathologically there is vasculitis of small arteries and veins with necrotizing extravascular granulomas. Biopsy tissue is needed to confirm the diagnosis.

The clinical syndrome and the histological features resemble those of polyarteritis nodosa, but the differences are sufficiently clear to establish Churg–Strauss syndrome as a separate entity. Although idiopathic in most cases, a minority appear to be a consequence of drug hypersensitivity, a recent example of particular interest being allergic granulomatosis and angiitis (the nomenclature of Churg and Strauss) complicating the use for asthma of newly introduced oral leukotriene receptor antagonists. It has been suggested, however, that the drugs themselves do not cause the disease, but merely lead to it being uncovered as the beneficial effect of leukotriene receptor antagonism allows a reduction (or withdrawal) of chronic steroid therapy. The disease may become life threatening if there is extensive vasculitic involvement of several organs, though generally responds satisfactorily to immunosuppressive therapy with corticosteroids. Other immunosuppressive or steroid sparing agents, such as azathioprine and cyclophosphamide, are usually required in addition.

Hypereosinophilic syndrome

The hypereosinophilic syndrome completes what might be described as a spectrum of overlapping disorders in which eosinophilic pneumonia is a prominent feature. The eosinophils appear mature, and infiltrate a number of organs by increasing degrees to cause progressive dysfunction, even death. The bone marrow is particularly densely infiltrated, raising the possibility that the disorder is primarily leukaemic in nature. Although the clinical picture does resemble that of eosinophilic leukaemia, the apparent maturity of the cells argues against this; current wisdom favours the hypereosinophilic syndrome as a distinct disorder, though one of unknown cause.

The clinical manifestations vary according to the organ(s) of principal involvement, and the extent of eosinophil infiltration. At the benign end of the spectrum the respiratory effects may be confined to an irritant cough, mild asthma, and minor episodes of eosinophilic pneumonia. When the disorder becomes life threatening, eosinophilic pneumonia may be extensive and associated with pleural effusion, but the chief threat to survival comes from myocardial and central nervous system infiltration. In most cases corticosteroids control progression, but when this fails, there may be progressive cardiac failure, or progressive functional impairment of central and peripheral nervous systems. This is often accompanied by weight loss, muscle weakness, enlargement of spleen and lymph nodes, gut and renal dysfunction, and venous thromboembolism. When corticosteroid therapy is ineffective, the use of antileukaemic agents may provide benefit.

---

*Dr D. J. Lane wrote on this subject in the third edition of the *Oxford Textbook of Medicine*. Much of his text has been retained in this revision and we acknowledge his contribution with grateful thanks.

Further reading

Churg J, Strauss (1951). Allergic granulomatosis, allergic angiitis and periarteritis nodosa. *American Journal of Pathology* **27**, 277–301.

Franco J, Artes MJ (1999). Pulmonary eosinophilia associated with montelukast. *Thorax* **54**, 558–60.

Kim Y *et al.* (1997).The spectrum of eosinophilic lung disease: radiologic findings. *Journal of Computer Assisted Tomography* **21**, 920–30.

Marchand E *et al*. (1998). Idiopathic chronic eosinophilic pneumonia. A clinical and follow-up study of 62 cases. *Medicine* **77**, 299–312.

Middleton WG *et al.* (1977). Asthmatic pulmonary eosinophilia. A review of 65 cases. *British Journal of Diseases of the Chest* **71**, 115–22.

Ong RK, Doyle RL (1998). Tropical pulmonary eosinophilia. *Chest* **113**, 1673–9.

Pearson DJ, Rosenow EC, III (1978). Chronic eosinophilic pneumonia (Carrington's): a follow up study. *Mayo Clinic Proceedings* **53**, 73.

17.11.10Lymphocytic infiltrations of the lung

D. J. Hendrick

Introduction

A number of disorders are characterized by lymphocytic infiltration of the lung. Several are rare and poorly understood, while others are relatively common and possess additional distinctive characteristics. For the latter group, classification poses few problems and individual diseases are readily distinguishable. These include, for example, Sjögren's syndrome, sarcoidosis, Wegener's granulomatosis, extrinsic allergic alveolitis, cryptogenic organizing pneumonia, bronchiolitis obliterans with organizing pneumonia, and cryptogenic fibrosing alveolitis, all of which are described separately in other chapters. However, for the first group, which constitutes the subject of this chapter, classification poses a continuing challenge: precise mechanisms and full natural histories are yet to be defined.

Disorders of lymphocytic infiltration are often considered as a spectrum of overlapping conditions, ranging from relatively benign infiltration of apparently normal lymphocytes without involvement of other cellular lines, through vasculitic and granulomatous inflammation, to frank malignancy. Apparent progression from disorder to disorder within the spectrum is not uncommon, but it is not always clear whether individuals affected in this way truly progress from one disease to another, or have a single disease whose early manifestations are similar to (and mistaken for) those of less serious neighbours in the disease spectrum. This has given rise to an alternative view that one end of the spectrum comprises a group of inflammatory disorders whose vasculitic and granulomatous features link more appropriately with diseases such as Wegener's granulomatosis and sarcoidosis; while the other end comprises the various malignant lymphomas.

Dominant lymphocytic infiltration is, nevertheless, a convenient definitive feature from which to consider the small group of uncommon pulmonary diseases that are described in this section. There is often paraprotein production, implying that a lymphocyte clone is involved. Depending on severity, these disorders are characterized clinically by cough (usually dry) and progressive undue exertional breathlessness; though systemic features of fever, malaise, and weight loss may also be prominent. Clubbing is not common, but there are frequently inspiratory crackles at the lung bases. The chest radiograph shows a diffuse interstitial pattern or patchy 'pneumonic' (i.e. air space) infiltrates with the more benign disorders, but nodular shadows at the more malignant end of the disease spectrum. Lung function tests show a non-specific pattern of ventilatory restriction with impaired parenchymal function.

Lymphocytic (and plasma cell) interstitial pneumonitis (pneumonia)

At the most benign end of the spectrum of lymphocytic infiltrations, lymphocytic (or lymphoid) interstitial pneumonitis is characterized by diffuse infiltration of the lung interstitium and alveolar walls with small mature lymphocytes, immunoblasts (activated lymphocytes), and plasma cells. Occasionally plasma cells dominate the lymphoid cell infiltrate, and in these circumstances the term plasma cell interstitial pneumonitis is preferred.

Lymphocytic pulmonary infiltration may occur in isolation without obvious cause; it may also be a non-specific feature of underlying pulmonary or systemic disease, such as HIV or Epstein–Barr virus (**EBV**) infection, drug hypersensitivity (sometimes toxicity), Castleman's disease (giant follicular lymph node hyperplasia), and (like Castleman's disease) a variety of autoimmune disorders, of which rheumatoid disease, Sjögren's syndrome, and systemic lupus erythematosus are most prominent. It may also be a consequence of a graft-versus-host reaction, and of common variable immunodeficiency—a primary antibody deficiency syndrome. When it occurs in children with AIDS it is thought to be largely a consequence of EBV infection. It may progress to (or be complicated by) the development of lymphoma, or be a feature of it. This too is particularly associated with Sjögren's syndrome.

The infiltrating lymphocytes show various levels of activation, and excess circulating immunoglobulins, whether monoclonal or polyclonal, are commonly observed. Occasionally there is hypo- rather than hypergammaglobulinaemia. When plasma cells rather than lymphocytes are dominant, the immunoglobulins are much less likely to be of the IgM class, though later complications may still include Waldenstrom's macroglobulinaemia or multiple myeloma. Bronchoalveolar lavage shows an excess of CD8+ T cells when lymphocytic interstitial pneumonitis is associated with HIV, whilst in Sjögren's syndrome the recovered lymphocytes are of the CD4+ phenotype.

The radiographic features, best seen with high resolution CT scans, are those of diffuse interstitial shadowing similar to cryptogenic fibrosing alveolitis, or (less commonly) air space filling. Cysts are often present, the mechanism being uncertain, and occasionally there are large nodules (> 10 mm in diameter). Effusion is not characteristic. The overall appearances are not specific, and since the disorder is rare, open biopsy is generally required for definitive diagnosis. It is seen in both sexes, usually in middle age, though children are not uncommonly represented. Slow progression is characteristic, though lymphocytic interstitial pneumonitis is rather more responsive to corticosteroid or other immunosuppressive therapy than is cryptogenic fibrosing alveolitis, and it sometimes remits spontaneously. There may, however, be complicating (even fatal) sepsis. The ultimate prognosis depends most on that of any underlying disease.

Benign lymphocytic angiitis

Lymphocytic infiltration in this condition is centred in small arteries and arterioles, necrosis is characteristically absent, and not infrequently there is granuloma formation. It therefore has both vasculitic and granulomatous features. It is rare, relatively benign, and usually affects the lungs or the skin. Most often there is no obvious provoking cause, but there have been reports of it emerging as a consequence of drug administration (streptokinase), HIV infection, or intrathoracic malignancy (thymoma).

Pulmonary lesions are usually single and most commonly present as asymptomatic nodules on a chance chest radiograph, the diagnosis being made following biopsy or resection. There may be systemic symptoms, however, and treatment with corticosteroids or cytotoxic agents may be necessary. The disease may progress to produce the more characteristic features of lymphomatoid granulomatosis, but more typically there is spontaneous remission. This suggests that benign lymphocytic angiitis is primarily a benign reactive vasculitis rather than a malignant lymphoma. It may be that similar histological features sometimes occur early in the course of lymphomatoid granulomatosis.

Immunoblastic (angio-immunoblastic) lymphadenopathy

This is a systemic and often febrile disorder characterized by widespread reactive lymphadenopathy and the infiltration of various organs by activated lymphoid cells, usually but not uniformly T lymphocytes. CD8+ cells, often clonal, are observed more commonly than CD4+ cells in affected organs, but in peripheral blood active disease is characterized by decreased numbers of T cells and an increase in B lymphocytes. The latter are possibly released from T-cell control and this might explain the frequency of paraprotein production. Infiltration of blood vessels may be prominent, hence the original term, angio-immunoblastic lymphadenopathy.

It occurs most commonly in the elderly and frequently in isolation, but it is often a consequence of infection (prominently HIV infection in recent years in younger subjects) or drug administration (often antibiotics), and it may be associated with autoimmunity.

Respiratory involvement is not common, usually comprising mediastinal or hilar lymphadenopathy, though diffuse interstitial infiltration and pleural effusion may occur. The involvement and enlargement of other lymphoid organs, particularly lymph nodes, liver, and spleen, usually offers a ready biopsy site for definitive diagnosis.

Management requires treatment of any provoking cause and, if necessary, the use of corticosteroids or other immunosuppressive agents. Occasionally there is spontaneous remission, but more commonly a T-cell lymphoma evolves. Indeed, the view is strengthening that most cases are due to peripheral (i.e. post-thymic)

T-cell lymphoma.

Lymphomatoid granulomatosis (angiocentric lymphoma)

Lymphomatoid granulomatosis is now widely considered to be a low-grade lymphoma, though the typical histological appearances of prominent infiltration of blood vessel walls and granuloma formation have, until recently, suggested a disease of more benign nature. The infiltrating cells comprise a mixture of lymphocytes, plasma cells, histiocytes, and atypical (usually malignant) lymphoid cells. The latter are derived from B lymphocytes, in some cases probably because of infection with EBV, but activated T cells are also prominent in focal lesions. Proliferation and vascular infiltration were initially assumed to cause luminal obstruction followed by ischaemic necrosis, but there are now doubts whether true vasculitis and true granuloma formation actually occur. As a consequence it has been proposed that angiocentric lymphoma is a more appropriate descriptive term.

The disease is uncommon in childhood but occurs throughout adult life with a slight predilection for males. It may arise on a background of an immunocompromised state. The lungs are almost invariably affected, but skin, central nervous system, and renal involvement is frequently seen, and there is often peripheral neuropathy. The disease is typically multifocal, affecting several organs, and may simulate disseminated carcinoma. Pulmonary lesions are usually discrete and nodular, whether single or multiple, but may vary in size from less than a centimetre in diameter to several centimetres across. Occasionally outlines are irregular and indistinct, suggesting patchy consolidation. Cavitation may occur, when an inflammatory cause may consequently be suspected, the radiographic appearances simulating those of Wegener's granulomatosis.

Symptoms are commonly dominated by systemic upset (fever, malaise, and weight loss), but respiratory involvement is likely to cause cough (sometimes with haemoptysis) or undue breathlessness. The involvement of other organs may provide valuable diagnostic insight, but biopsy is necessary for definitive diagnosis. Temporary improvement sometimes follows treatment with corticosteroids alone, but a realistic chance of cure requires cytotoxic therapy for lymphoma.

Lymphoma

Unquestionably at the malignant end of the disease spectrum lies lymphomatous infiltration of the lung. All lymphoma types may present with intrathoracic disease, and all may involve the thorax later if they present elsewhere. Lymph nodes, lymphatics, and lung parenchyma may all become infiltrated, but the pattern may vary between tumour types.

A comprehensive review of lymphomas is beyond the scope of this section, and the reader is referred to Section 22 for full details. To the chest physician the distinction of Hodgkin's disease and low-grade non-Hodgkin's disease from high-grade non-Hodgkin's disease is of particular value, since the former are now generally curable—or at least highly responsive to treatment.

Hodgkin's disease is generally a disease of adults and adolescents. When it involves the thorax it usually does so by infiltrating hilar or mediastinal lymph nodes, though patchy or even diffuse parenchymal infiltration does rarely occur, with or without lymphadenopathy. Asymmetrical nodal enlargement favours lymphoma over sarcoidosis, the other disorder (apart from tuberculosis in endemic areas) that commonly produces hilar and mediastinal adenopathy. Hodgkin's disease frequently involves other nodal sites at presentation, but may be confined to the thorax. Pleural effusion is not uncommon, and occasionally there is infiltration of the chest wall. Biopsy and staging are essential to diagnosis and management, though the advent of computed tomographic scanning has greatly simplified the latter by eliminating the need for laparotomy. Radiotherapy is normally curative for localized nodal disease and is invaluable as an adjunct in reducing local 'bulk' when the disease is disseminated. It carries much less risk than chemotherapy, which is required for parenchymal or disseminated disease, and for the small proportion of patients with localized nodal disease who show features of poor prognostic significance (e.g. high 'bulk', systemic symptoms, and anaemia).

Non-Hodgkin's lymphoma is more common than Hodgkin's disease and tends to affect a rather older population. Its thoracic manifestations are similar to those of Hodgkin's disease, but it is the more likely malignant 'complication' of the other lymphocytic pulmonary infiltrations discussed in this section. It also has a greater tendency to be disseminated at presentation, to have more 'high grade' features histologically and clinically, and (not surprisingly) to be less responsive to therapy. Nevertheless, localized and low-grade tumours are often curable, and useful palliation is generally achieved for all but the most high-grade tumours.

High resolution CT scanning shows that both types of lymphoma are most commonly characterized by an air space filling pattern rather than interstitial shadowing, and by large nodules and pleural effusions.

Chemotherapeutic regimens for the treatment of lymphoma continue to develop rapidly and have properly become the responsibility of specialist haemato-oncologists. Chemotherapy is, of course, attended by the familiar risks of bone marrow suppression and an immunocompromised state, but with regard to those with lung disease it is noteworthy that many of the chemotherapeutic agents can themselves cause interstitial lung disease—including lymphocytic infiltration. The supervising physician may consequently face a classic diagnostic dilemma when, following an initial satisfactory remission, the patient's radiographs show pulmonary shadows consistent with infection, drug hypersensitivity/toxicity, or recurrent lymphomatous infiltration. A prompt and accurate diagnosis is essential since each possibility requires fundamentally different management. Expectorated secretions may provide adequate evidence of infection to justify a trial of antibiotic therapy, but if immediate progress is unsatisfactory, fibreoptic bronchoscopy with lavage and/or transbronchial biopsy is generally needed. It may be that with increasing use of the polymerase chain reaction to amplify fragments of genetically specific microbial material, sputum or even oropharyngeal secretions will prove adequate to identify the infecting organisms. However, the cause of pulmonary shadows in this situation is often complex, and any combination of these three diagnostic groups may develop (perhaps with more than one infecting micro-organism). A multidisciplinary approach to management has consequently become essential, involving chest physicians, radiologists, microbiologists, and histopathologists under the expertise of supervising oncologists or haematologists.

Further reading

Calabrese LH *et al.* (1989). Systemic vasculitis in association with human immunodeficiency virus infection. *Arthritis and Rheumatism* **32**, 569–76.

Churg A (1983). Pulmonary angiitis and granulomatosis revisited. *Human Pathology* **14**, 868–83.

Donnelly TJ, Tuder RM, Vendegna TR (1998). A 48-year-old woman with peripheral neuropathy, hypercalcaemia, and pulmonary infiltrates. *Chest* **114**, 1205–9.

Frizzera G, Moran EM, Rappaport H (1974). Angio-immunoblastic lymphadenopathy with dysproteinaemia. *Lancet* **ii**, 1070–3.

Glickstein M *et al.*(1986). Non lymphomatous lymphoid disorders of the lung. *American Journal of Roentgenology* **147**, 227–37.

Guinee DR *et al.* (1998). Proliferation and cellular phenotype in lymphomatoid granulomatosis: implications of a higher proliferation index in B cells. *American Journal of Surgical Pathology* **22**, 1093–100.

Haque AK *et al*. (1998). Pulmonary lymphomatoid granulomatosis in acquired immunodeficiency syndrome: lesions with Epstein–Barr virus. Modern Pathology **11**, 347–56.

Honda O *et al.* (1999). Differential diagnosis of lymphocytic interstitial pneumonia and malignant lymphoma on high-resolution CT. *American Journal of Roentgenology* **173**, 71–4.

Katzenstein A-LA, Carrington CB, Liebow AA (1979). Lymphomatoid granulomatosis. A clinicopathologic study of 152 cases. *Cancer* **43**, 360–73.

Liebow AA (1973). Pulmonary angiitis and granulomatosis. *American Review of Respiratory Disease* **108**, 1–18.

O'Connor NT *et al.* (1986). Evidence for monoclonal T lymphocyte proliferation in angioimmunoblastic lymphadenopathy. *Journal of Clinical Pathology* **39**, 1229–32.

Watanabe S *et al*. (1986). Immunoblastic lymphadenopathy, angioimmunoblastic lymphadenopathy, and IBL-like T-cell lymphoma. A spectrum of T-cell neoplasia. *Cancer* **58**, 2224–32.

Weiss LM *et al*. (1986). Clonal T-cell populations in angioimmunoblastic lymphadenopathy and angioimmunoblastic lymphadenopathy-like lymphoma. *American Journal of Pathology* **122**, 392–7.

17.11.11 Extrinsic allergic alveolitis

D. J. Hendrick and G. P. Spickett

Historical background

Farmer's lung is often regarded as the prototype of the alveolar and bronchiolar disorders that result from hypersensitivity to inhaled organic dusts. These occur worldwide and are known collectively by the term extrinsic allergic alveolitis, although it is recognized that the underlying inflammatory response occurs diffusely throughout the gas exchanging tissues and is not confined to the alveoli. For this reason many prefer the term hypersensitivity pneumonitis. These alveolar disorders were not clearly distinguished from asthma until 1932 when Campbell published his celebrated report describing three affected English farm workers, the appellation 'farmer's lung' being suggested in 1944. However, the disease had been recognized in Iceland in the nineteenth century, and probably contributed to the occupational ailments of grain workers so graphically described by Ramazzini in the eighteenth century.

Part of the eminence of farmer's lung itself stems from its industrial importance, and part from its historical role in the understanding of extrinsic allergic alveolitis. Its relation to the inhalation of dust from mouldy hay, straw, or grain had been recognized from the outset, but it was not until 1961 when Pepys and colleagues demonstrated the presence of precipitins to antigens of mouldy hay in patients suffering from the disease that the idea of an allergic aetiology gained general acceptance. These and other investigators showed that the main sources of antigen were contaminating thermophilic actinomycetes, particularly *Micropolyspora faeni* (now known as *Saccharopolyspora rectivirgula*) and *Thermoactinomyces vulgaris*. These thermophilic microbes (which are actually bacteria not fungi) colonize fermenting damp vegetable produce as it heats up. When it eventually dries, a respirable dust laden with antigenic microbial spores is left. Symptoms are consequently most common during winters following wet summer harvests, when hay or grain is used for feeding stock, and astonishing numbers of spores (thousands of millions per cubic metre) are released into the air.

For deposition of the dust to occur predominantly in the gas exchanging tissues, particle size must be largely confined to the range 0.5 to 5 µm. This encompasses the diameters of many antigenic bacterial and fungal spores, and a large number of microbial species are now recognized as causes of extrinsic allergic alveolitis. In addition, the disease has been described following respiratory exposure to a variety of antigens derived from animal, vegetable, and even chemical sources, both in the workplace and in the home. It may also occur because of allergy to ingested agents, chiefly medications, but only inhalant causes will be addressed in this chapter. Drug-induced examples of the disease are discussed in Chapter 17.11.19.

Causative agents

Table 1 lists the various agents, principally organic proteins, reported to cause extrinsic allergic alveolitis. Most are encountered in working environments and so the disease is usually occupational, but some are encountered in the home or in recreational environments. Most are micro-organisms that are found contaminating a variety of vegetable products, but some are derived directly from animal or vegetable sources, and a few are reactive chemicals. The latter are thought to act as haptens, combining with body proteins to produce a larger and now antigenic molecule. Although the micro-organisms associated with the more celebrated disorders—farmer's lung, mushroom worker's lung, and bagassosis—are usually thermophilic, the majority causing extrinsic allergic alveolitis are not. Even with mouldy hay and farmer's lung there is evidence that non-thermophilic organisms (e.g. *Aspergillus* spp.) may occasionally be involved.

Some microbial contamination may occur during growth of the vegetable host, but most of the antigenic load is usually acquired after harvest. Prolonged storage under damp conditions increases the risk of extrinsic allergic alveolitis substantially, whilst drying to reduce the water content below 30 per cent greatly lessens the risks. Farmer's lung and bagassosis are not therefore primary disorders of hay, grain, or sugar cane harvest. They usually arise months or even years later when the stored product is used or moved. In the interim, moulding is likely to have involved a series of different micro-organisms that colonize the forage material sequentially. As the exothermic process increases the ambient temperature, so thermophilic microbes come to dominate.

Inevitably there are situations where contamination arises with a number of different microbes, and affected subjects show antibodies to several of them. Unless time-consuming inhalation challenge tests are carried out with extracts of the individual microbial species, it is not possible to identify a single responsible agent in a given case or cases, and it is conceivable that several could be relevant in these circumstances. This is a characteristic feature of contaminated water reservoirs in humidifiers and air conditioners, and a great variety of agents have been suggested as possible causes of humidifier lung, including bacteria, mycobacteria, fungi, protozoa (amoebae), and metazoa (nematode debris). Some authors prefer to distinguish extrinsic allergic alveolitis attributable in such circumstances to micro-organisms growing in cool or cold water (humidifier lung) from that arising from heated water (ventilation pneumonitis). Additional sources of causal organisms include hot tubs and saunas, containing both thermophilic and non-thermophilic organisms (including non-tuberculous mycobacteria), and water-based metal working fluids. The latter, often contaminated with oil, are recycled during use to lubricate and cool rotating or cutting equipment in the metal working industry, and may therefore be dispersed as respirable aerosols. The chief microbial contaminants are generally environmental non-tuberculous mycobacteria or fungi, but a variety of other organisms may be involved. Since granulomatous responses might be expected from mycobacterial infection, the mechanism of diffuse pneumonitis when mycobacteria are involved may not be one of hypersensitivity.

Curiously, contamination with multiple microbial species does not seem to be a feature of Japanese summer-type pneumonitis, which arises seasonally in the hot and humid regions in the south and west of Japan and neighbouring countries. This is the result of the excessive growth of *Trichosporon* spp. in unsanitary and poorly ventilated homes.

Epidemiology

Extrinsic allergic alveolitis is an uncommon but not rare disease. Its comparative scarcity limits epidemiological knowledge, as does the use of different methods of investigation. For every case there may be 100 cases of 'extrinsic allergic' asthma, but there is even greater geographical variation than with asthma reflecting the

much larger dependence of extrinsic allergic alveolitis on occupational causes and climate. As a consequence, its incidence and its principal causes vary considerably from country to country, and from region to region.

Incidence

Experience over 3 years with the **SWORD** project (Surveillance of Work-related and Occupational Respiratory Disease) indicated that extrinsic allergic alveolitis of occupational origin accounted for 2 per cent of occupational lung diseases in the United Kingdom. Asthma, the most common, accounted for 26 per cent. This does, of course, ignore extrinsic allergic alveolitis of non-occupational origin, which is much less easily assessed. It also disguises the absolute risk since few workers encounter relevant occupational exposures. Almost 50 per cent of reported cases involved farmers or farm workers, followed by 15 per cent affecting workers in material, metal, or electrical processing trades. Among the farmers, the average incidence was 41 per million per year, though this approached 100 in some regions, and has been estimated at 3000 in Quebec, Canada. However, the estimated incidences are crude and must vary considerably according to the prevailing weather. They may be compared with 200 to 700 per million per year among working groups at greatest risk of developing occupational asthma in the United Kingdom. Contaminating micro-organisms underlie over 50 per cent of the cases of extrinsic allergic alveolitis reported to SWORD, followed in order of importance by animal antigens in 6 per cent and chemicals in 5 per cent. In 27 per cent of reports a suspected agent was not specified.

Prevalence

Figures for prevalence (the proportion affected among a given population at a given point of time) are more readily available than those for incidence, and demonstrate quite marked national differences. In developed countries, humidifier lung is being recognized with increasing frequency in both the workplace and the home, and remarkable prevalences of 15 to 70 per cent have been suggested in populations from contaminated offices in North America. Bird fancier's lung may be more prevalent at present over the whole of the United Kingdom, simply because of the great popularity of keeping budgerigars and pigeons. Budgerigars are kept in some 12 per cent of British homes, and it has been estimated that 0.5 to 7.5 per cent of the population involved are likely to have extrinsic allergic alveolitis as a consequence, albeit mildly in most cases. Pigeon keeping is 40 times less common, and the measured prevalence of pigeon fancier's lung among pigeon keepers has been a good deal more varied (0 to 21 per cent). This may reflect both true differences between groups of pigeon breeders as exposure levels vary, for instance, according to number of birds, duration of exposure, loft ventilation, and cleaning habits, and artefactual differences arising as a result of selection bias and the notorious lack of compliance shown by pigeon fanciers in epidemiological studies. The avian antigen responsible and its precise source have yet to be identified, but bloom from the feathers containing oil, saliva, and secretory IgA is currently favoured over dust emanating from dried droppings.

In areas of high rainfall where 'traditional' farming methods are used, the prevalence of farmer's lung may reach 10 per cent. This is likely to be the commonest cause of extrinsic allergic alveolitis in developing countries. In developed countries, where modern farming methods are used, prevalences rarely exceed 2 to 3 per cent and are usually a good deal less. Furthermore, the farming population at risk represents a mere 1 to 2 per cent of the population at large, although there are marked regional variations. Even smaller populations are employed making whisky from germinating barley (maltings), raising mushrooms on a variety of antigenic composts, or handling bagasse (the fibrous stem that remains when sugar is extracted from sugar cane), but within some of these populations extrinsic allergic alveolitis was a common problem until excessive exposure levels were controlled. Extrinsic allergic alveolitis associated with animals other than birds is extremely uncommon, as is the case with chemical-induced extrinsic allergic alveolitis.

In Japan, the seasonal summer growth of *T. cutaneum* in the home is by far the commonest cause of extrinsic allergic alveolitis where the remarkable 'summer-type hypersensitivity pneumonitis' accounts for about 75 per cent of all cases of extrinsic allergic alveolitis. It is approximately 10 times as common as farmer's lung and 20 times as common as bird fancier's lung. Occasionally other subspecies of *Tricosporon* are responsible.

Pathogenic mechanisms

Histology

There has been little opportunity to characterize the acute form of extrinsic allergic alveolitis histologically because biopsies are very rarely taken within 24 to 48 h of a provoking exposure, and because death leading to autopsy is even less common. Initially there is a non-specific diffuse pneumonitis with inflammatory cellular infiltration of the bronchioles, alveoli, and interstitium, accompanied by oedema and luminal exudation. With ongoing exposure, whether continuous or intermittent, the more familiar appearances of the subacute forms of extrinsic allergic alveolitis evolve. The most characteristic feature is the formation of epithelioid non-caseating granulomas. These are generally less well formed than in sarcoidosis, less profuse, and often evanescent. They can be recognized within 3 weeks of the initiating exposure, and generally resolve within 6 to 12 months. In parallel, fibrosis evolves alongside cellular infiltration of the interstitium with histiocytes, lymphocytes, and plasma cells. Macrophages with foamy cytoplasm may be prominent in the alveolar spaces, and organization of the inflammatory exudate may lead to intra-alveolar fibrosis. Obstruction or obliteration of bronchioles is common. Foreign body giant cells may reflect the dependence of extrinsic allergic alveolitis on antigens derived from inhaled foreign material, as does a peribronchial predominance of the inflammatory response. Vasculitis is notable by its absence. The typical histological appearance of subacute extrinsic allergic alveolitis is illustrated in Fig. 1.

Fig. 1 Histological appearance: subacute disease (by courtesy of Dr T. Ashcroft). Haematoxylin and eosin stain. Medium magnification. There is bronchocentric interstitial fibrosis and chronic inflammation, with poorly formed interstitial granulomas including giant cells.

With continued exposure, progressive, widespread, and irreversible fibrosis may occur, leading to disruption of the normal architecture of the lung. In advanced cases honeycombing may develop. Granulomas are no longer characteristic, and the overall appearances may differ little from other causes of progressive interstitial pulmonary fibrosis. With extrinsic allergic alveolitis, however, there may be disproportionate fibrosis of the upper lobes.

Immune mechanisms

An outline of the possible immunopathology of extrinsic allergic alveolitis through acute, subacute, and chronic phases is illustrated in Fig. 2 and Fig. 3. The presumption that complexes of antigen and complement-activating antibodies are primarily responsible for extrinsic allergic alveolitis is now largely discarded. The evidence for deposition of immune complexes is not convincing, and neither IgG nor IgM antibodies are uniformly demonstrated in the sera of affected subjects unless sensitive detection techniques such as the enzyme-linked immunosorbent assay or radioimmunoassays are used. More importantly, these antibodies are frequently found in subjects who are similarly exposed but clinically unaffected, irrespective of the method of detection. A closer association of disease with the IgG4 antibody subclass has been suggested, but the significance of this is not yet apparent. It is clear, however, that vasculitis—a cardinal feature of the experimental Arthus reaction—is not a characteristic; the inflammatory reaction is dominantly lymphocytic or mononuclear rather than polymorphonuclear. However, a transitory polymorphonuclear leucocyte response is typical immediately following exposure. Lung tissue is most commonly examined during subacute phases of the disease, at which time a non-caseating granulomatous response suggesting cell-mediated hypersensitivity is the usual finding.

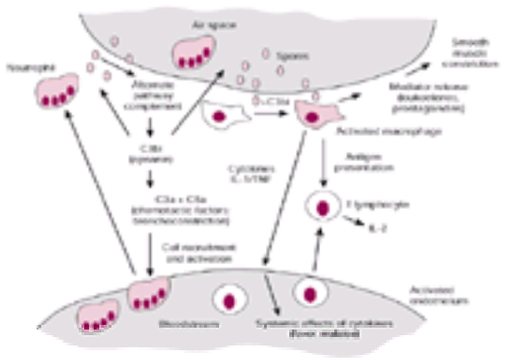

Fig. 2 Possible immunopathogenesis: acute phase.

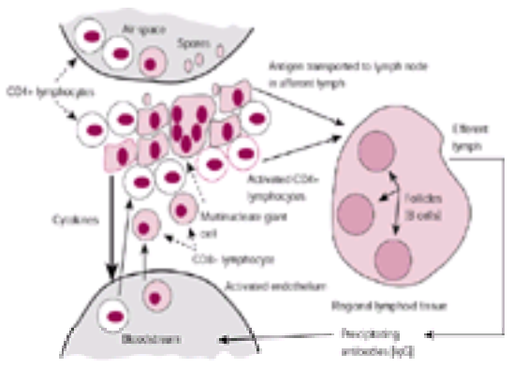

Fig. 3 Possible immunopathogenesis: subacute/chronic phase.

It could be argued that these histological appearances merely represent a healing reaction, but the consistent finding of an acute T-lymphocyte response in fluid obtained at bronchoalveolar lavage supports the current consensus that cell-mediated hypersensitivity plays the dominant pathogenic role in extrinsic allergic alveolitis. The results from animal models of the disease are consistent with this, disease being transferred from animal to animal only with sensitized T lymphocytes. This is not to say that other mechanisms play no role, nor that all inflammatory diseases of the gas exchanging tissues induced by organic dusts share a common mechanism. Indeed, the onset of symptoms within a few hours of exposure, coupled with polymorphonuclear leucocytosis in bronchoalveolar lavage fluid and peripheral blood, favours the participation of an additional (perhaps priming) immunological or toxic process, and B-lymphocyte aggregates have been noted in transbronchial biopsies obtained during the acute phase. Components of a number of organic dusts associated with extrinsic allergic alveolitis are known to activate complement by the alternative pathway and this, with or without humoral hypersensitivity, might also prove to be relevant.

In fact, bronchoalveolar lavage in similarly exposed subjects has shown excess numbers of T lymphocytes whether they were clinically affected or not, though the proportions of T-cell subpopulations has varied according to disease activity and the circumstances of exposure. Most investigators have detected a relative excess of CD8+ T cells in exposed but asymptomatic subjects, thereby 'inverting' the normal CD4+ to CD8+ ratio. The balance appears to be less disturbed in those with disease, and in one sequential study the ratio changed from 0.43 to 1.47 with disease progression. In an intriguing study of an animal model of extrinsic allergic alveolitis, monkeys that developed characteristic reactions to inhalation challenge showed a helper CD4+ cell lymphocytosis in bronchoalveolar fluid and a relative deficiency of suppressor CD8+ cells, compared with the monkeys giving no clinical reaction who showed responses with both CD4+ and CD8+ cells. When the non-reactors were challenged again after low doses of whole-body irradiation had impaired suppressor more than helper cell function, characteristic reactions were noted. These observations suggest that a relative impairment of suppressor cell function, or of its activation following antigenic exposure, is fundamental to the development of extrinsic allergic alveolitis—a situation that has interesting parallels with sarcoidosis. It is also interesting that lymphopenia in peripheral blood is a typical feature of acute exacerbations of the disease, the T lymphocytes migrating from blood to lungs within hours of the provoking exposure. It is small wonder that studies of systemic and local immune responses have given discordant results, and clear that continuing research should address both aspects of the immune response.

It is known that different antigenic determinants from a given inducing microbial source may lead to different immunological responses, and it seems likely that cytotoxic activity and released cytokines (e.g. interleukin 6 and tumour necrosis factor-a) play some role, possibly by activating the vascular endothelium and thereby recruiting and activating further macrophages and inflammatory cells. In experimental models interferon-g has been shown to play a major role (an excess of interferon-g-producing T cells is present in the lungs), and interleukin 10 ameliorates the disease. This indicates that extrinsic allergic alveolitis is a T_{H1}-type disease. Interleukin 8, a chemotactic factor for neutrophils, and monocyte chemotactic protein-1 are both elevated in some types of the disease, perhaps accounting for the increase in macrophage and neutrophil recruitment and activation. Serum levels of interleukin 6 and intercell adhesion molecule 1 (ICAM-1) are also raised. Bronchoalveolar lavage has shown that natural killer cells (CD57+) and mast cells may be prominent additional players in pathogenesis, and there is evidence that the capacity of macrophages to present antigen is enhanced by viral infection. It is diminished by cigarette smoking, which is known to decrease the severity of clinical symptoms as well as the immunological abnormalities.

Cytokines, possibly together with anaphylotoxins from the degradation of complement components (C4, C3, C5), are likely to be responsible for the systemic influenza-like symptoms that are so characteristic of the acute form of extrinsic allergic alveolitis. These symptoms are indistinguishable from those of grain fever in grain workers, 'Monday fever' in cotton workers, humidifier fever in subjects exposed to microbially contaminated humidifiers, and metal fume fever in welders. In these situations, the febrile disorder is not characteristically associated with clinical alveolitis, raising the possibility that its occurrence with the acute form of extrinsic allergic alveolitis is an independent phenomenon, not an integral part of extrinsic allergic alveolitis itself. In favour of this hypothesis has been the finding of high levels of endotoxin from Gram-negative bacteria (which are known to provoke these symptoms) in grain dust, cotton dust, contaminated humidifiers, and many of the 'mouldy' vegetable dusts that cause extrinsic allergic alveolitis. However, neither metal fume nor several other causative agents of extrinsic allergic alveolitis are likely to be contaminated with endotoxin, and so endotoxin-induced release of inflammatory mediators is not an entirely satisfactory explanation. For example, inhalation provocation tests with uncontaminated bird serum in subjects with bird fancier's lung reproduce both alveolar and influenza-like responses. Evidently the influenza-like response is an integral feature of the acute form of extrinsic allergic alveolitis, but it is relatively non-specific and can occur in many other situations.

Extrinsic allergic alveolitis occurs in families only sporadically, and few associations with HLA phenotypes have been demonstrated. However, a number of recent studies have suggested associations between HLA D alleles and pigeon fancier's lung and Japanese summer-type hypersensitivity pneumonitis. Such alleles may exert effects on immune suppression, and offer one mechanism by which a genetic predisposition could play a role in the development of extrinsic allergic alveolitis. It has also been suggested that an acute inflammatory episode (from viral infection or the inhalation of microbial toxins or chemicals) may be necessary to disrupt the normal defence equilibrium of surface membrane and local immune responses, and thereby permit antigen to be presented in a fashion that leads to hypersensitivity. Undue 'leakiness' of the alveolar membrane can be demonstrated by an increased clearance of inhaled $^{99}Tc^{m}$-DTPA, and this has been reported in both the early and continuing phases of extrinsic allergic alveolitis.

Relation to smoking

The disruptive effect of smoking on the alveolar membrane does not appear to augment the risk for extrinsic allergic alveolitis or to increase its severity. Rather, the reverse is true. Although smoking enhances acute-phase reactions and IgE production, it diminishes IgA, IgG, and IgM antibody responses, increases circulating CD8+ T-lymphocyte numbers, and probably reduces the incidence and severity of extrinsic allergic alveolitis. However, the smoker without IgG antibodies is particularly liable to find his respiratory symptoms attributed to other diseases, and so this negative association between extrinsic allergic alveolitis and smoking may have been exaggerated. That it is real is supported by evidence that smoking may also reduce the risk for other T-cell-mediated immunological disorders such as sarcoidosis, ulcerative colitis, and some types of occupational asthma (generally those associated with low molecular weight chemicals). The key cell in a complex series of interactions is probably the alveolar macrophage, which is critical in presenting antigen to CD4+ T lymphocytes and so to activating cellular immune

mechanisms. Although smoking increases macrophage numbers and their metabolic activity, the activated cells show impairment of both the expression of surface major histocompatibility class 2 antigens and the production or release of interleukin 1 and inflammatory mediators derived from arachidonic acid metabolism (leukotriene B4, prostaglandin E_2, thromboxane B_2). It is also argued that the increased macrophage numbers down-regulate pulmonary immune responses in a purely non-specific fashion by impairing antigen access to more effective blood monocytes.

Relation to coeliac disease

Reports that cryptogenic fibrosing alveolitis and extrinsic allergic alveolitis (and particularly bird fancier's lung) might be associated with coeliac disease led to the interesting hypothesis that in some cases absorbed food antigens from the disrupted bowel mucosa might play a role in the pathogenesis of the lung disorder; that is, the lung disorder might be a 'metastatic' complication of the bowel disease. Alternatively, systemic hypersensitivity to a common inhaled and ingested avian antigen might give rise to similar immune reactions and diseases in the relevant target organs. The avian IgG antibody response seen in coeliac disease is, however, distinct from that associated with bird fancier's lung and seems to be a response to dietary egg. It is not related to environmental exposure to birds but does correlate with the activity of the bowel disorder. Subsequent experience suggests a much less strong association between these bowel and lung disorders that, if real, is probably a consequence of their dependence on similar immunological mechanisms and host susceptibility to them.

Clinical features

Acute form

The acute form of extrinsic allergic alveolitis is the most easily recognized, because symptoms are often distressing and incapacitating, and have a high degree of specificity. Following a sensitizing period of exposure, which may vary from weeks to years, the affected subject experiences repeated episodes of an influenza-like illness accompanied by cough and undue breathlessness some hours (usually 3 to 9) after commencing exposure to the relevant organic dust. The systemic influenza-like symptoms generally dominate those that are respiratory, and the affected subject complains most of malaise, fever, chills, widespread aches and pains (particularly headache), anorexia, and tiredness. He is unlikely to feel like exercising and may well put himself to bed, therefore remaining unaware of undue shortness of breath, though he is likely to develop a dry cough without wheeze and to have some difficulty taking deep, satisfying breaths. Occasionally there is an asthmatic or bronchitic response and wheezing or productive cough becomes an additional feature.

Despite the delay in onset after exposure begins, affected subjects soon learn to associate symptoms with the causative environment, especially if they follow a period away from work or the causal exposure. Recognition is particularly easy for groups such as farmers and pigeon fanciers for whom these risks are well known. However, in some cases there may be a tendency to deny such a relationship for fear of compromising the ability to pursue livelihood or hobby, and the clinical history may appear much less convincing than it should.

The severity and duration of symptoms depend critically on exposure dose and individual susceptibility. With low levels of acute exposure, symptoms are mild and persist for a few hours only. When occupation is responsible, the affected worker may feel unwell only at home during the following evening or night, and be fully recovered by the next morning, hence obscuring the relevance of the workplace. When severe responses follow particularly heavy exposures the relation of the one to the other will be more obvious, and complete remission may require several days or even weeks.

In exceptionally severe cases, life-threatening respiratory failure may develop and emergency admission to hospital becomes necessary. Death is not unknown. Respiratory distress at rest with fever and gravity-dependent crackles comprise the major physical signs, with breathing being fast but shallow. Clubbing is very rarely seen. Hypoxaemia is typically accompanied by hypocapnia, and the chest radiograph shows a diffuse alveolar filling and interstitial pattern. Supplemental oxygen may be required, and in rare cases there may be a brief need for mechanical ventilatory support. Spontaneous recovery can be expected to begin within 12 to 24 h, and can be accelerated with corticosteroids.

Most subjects recover fully from each acute exacerbation, and if the cause is recognized and further exposure avoided, there is little risk of persisting pulmonary dysfunction. However, it is not always realistic to expect affected individuals to avoid further exposure, particularly among farming communities, and there is some risk that continuing exposure and repeated acute exacerbations will eventually lead to permanent impairment of lung function.

Chronic form

In some subjects extrinsic allergic alveolitis presents in a much less dramatic but potentially more serious way. Exercise tolerance is gradually lost due to shortness of breath, but without systemic upset aside from (in some cases only) prominent loss of weight. This is the result of diffuse pulmonary fibrosis, which has often been progressing for years before the affected subject seeks advice: the slower the progression, the longer the delay, and the greater the likely degree of permanent fibrotic damage. Eventually hypoxaemia and pulmonary hypertension may supervene, and the right heart fails. There are no acute exacerbations, and each day and each month are much like any other. The clinical features are similar to those of other varieties of pulmonary fibrosis, although clubbing is uncommon, and it may prove extremely difficult to distinguish this form of extrinsic allergic alveolitis from cryptogenic fibrosing alveolitis, sarcoidosis, or other slowly progressive forms of pulmonary fibrosis. There may also be asthmatic or bronchitic symptoms, but these are best regarded as independent airway manifestations of hypersensitivity to the causal agent.

The chronic form of extrinsic allergic alveolitis is typically seen in the subject who keeps a single budgerigar (known as a parakeet in North America) in the home. The level of antigenic exposure to avian dust is comparatively trivial (compared with the farm worker forking bales of heavily contaminated hay in a poorly ventilated barn), but it is encountered almost continuously, particularly if the affected subject is someone confined to the home. Differing exposure patterns are largely responsible for these distinct forms of extrinsic allergic alveolitis, although differences in host responsiveness must exert an important additional influence. There may consequently be considerable variability in clinical features among individuals affected by the same source of antigenic exposure.

Intermediate forms

The fact that the acute form of extrinsic allergic alveolitis can be produced by inhalation provocation tests in subjects with the chronic form of the disease emphasizes the major role that dose exerts in determining the clinical nature of the response that occurs. Depending on exposure dose and host responsiveness a variety of intermediate forms of extrinsic allergic alveolitis are recognized, and some subjects will experience different patterns of response at different times. Hence it is possible for acute exacerbations to occur in subjects manifesting predominantly the chronic form of the disease, and for a limited degree of recovery to follow cessation of exposure. In general, however, the individual affected by the chronic form of extrinsic allergic alveolitis should be satisfied if no further progression occurs following cessation of exposure, because in some cases fibrotic damage continues regardless.

Investigation

Establishing a diagnosis of extrinsic allergic alveolitis involves three areas of investigation: the lungs, the exposure, and the evidence for hypersensitivity.

Pulmonary

In many cases extrinsic allergic alveolitis is first suspected after the presence of diffuse alveolitis or progressive pulmonary fibrosis is established. With the acute form of the disease the chest radiograph commonly shows no abnormality unless symptoms are moderately severe. Normal radiographic appearances are particularly common with humidifier lung, possibly because antigen is largely presented in soluble rather than particulate form. When the radiograph is abnormal, there is a widespread ground-glass appearance or an alveolar filling pattern, particularly in the lower and mid-zones. This may resolve within a mere 24 to 48 h once exposure has ceased. In more subacute forms small reticular opacities, simulating asbestosis, are seen within the same distribution: these may persist for several weeks despite cessation of exposure and, if exposure continues, honeycombing may develop. Occasionally a more nodular pattern occurs. In contrast to the distribution of acute and subacute radiological abnormalities, the upper zones are predominantly affected by the irreversible fibrotic process that characterizes the chronic form of disease. This may simulate sarcoidosis or even tuberculosis, and may lead to considerable shrinkage and distortion. In practice, the radiographic appearances vary considerably from patient to patient, and correlate poorly with the clinical severity of the disease.

Computed tomographic (**CT**) scans provide a much clearer picture of the type of radiographic abnormality and of its extent, particularly when thin-section

high-resolution techniques are used, but they have shown that no single feature or pattern is pathognomonic (Fig. 4). Again, investigation within hours of exposure has been limited and experience is largely confined to patients with subacute and chronic disease. Increased ground-glass density of the lung parenchyma is the most prominent finding in the subacute form, followed almost equally by reticular or nodular infiltration. At end expiration a mosaic pattern is characteristic, reflecting patchy bronchiolar involvement and the different degrees by which residual gas can be expelled from distal lobules. The attenuated areas may then be normal, the translucent areas indicating gas retention. Neither lymph node enlargement nor pleural involvement are characteristic. The CT scan is appreciably more sensitive than the plain chest radiograph, and shows a more uniform involvement of the lung fields in subacute disease than is obvious from plain radiographs. With chronic forms, the CT scan shows a similar pattern of fibrosis and disruption to the plain radiograph, but again is considerably more sensitive.

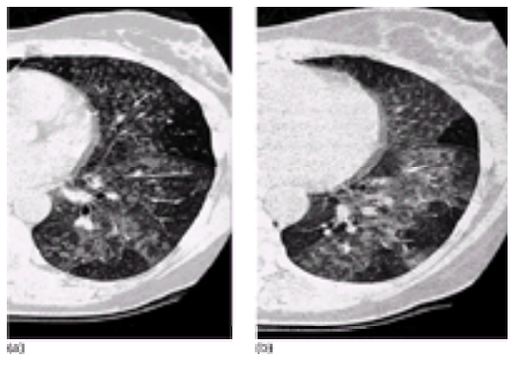

Fig. 4 (a) CT scan of a never-smoking woman aged 44 years whose lung biopsy showed the typical appearances of subacute EAA. She kept two budgerigars in her home and had serum precipitins to avian antigens. The scan shows marked groundglass attenuation of the lung parenchyma, which is nodular in some areas due to characteristic peribronchiolar (and centrilobular) foci. In other areas there is increased translucency because of bronchiolar obstruction and air trapping. Both the groundglass attenuation and the increases in translucency are exaggerated in the expiratory film (b), giving a 'mosaic' pattern. She recovered fully after the birds left her home..

Lung function studies vary according to severity and recent activity. As with asthma, they may be unremarkable in the acute form of the disease when there has been little recent exposure. When lung function is impaired, the pattern suggests parenchymal and interstitial disease but is otherwise non-specific. There is impaired carbon monoxide gas transfer (diminished TL_{CO} and K_{CO}) with restricted ventilation (i.e. forced vital capacity, FVC, is diminished as much as forced expiratory volume, FEV1, or more so, with respect to predicted values), decreased compliance, and (in more severe cases) arterial hypoxaemia and hypocapnia, particularly on exercise. Although total lung capacity is reduced, residual volume is often increased, suggesting air trapping as a result of bronchiolar involvement. Occasionally there is also obstruction of the large airways, but this implies a coincidental asthmatic or bronchitic effect. Serial measurements of lung function may be particularly useful in demonstrating that impairment is closely related to the relevant exposure.

Bronchoscopy and bronchoalveolar lavage are useful in demonstrating that there is no macroscopic abnormality, apart from the occasional presence of mucosal inflammation, and that no microbial growth occurs on culture. If lavage fluid is obtained within a matter of hours of exposure, a polymorphonuclear leucocyte response may dominate, simulating cryptogenic fibrosing alveolitis, but this is followed by an accumulation of lymphocytes over the following 24 to 48 h. In the subacute and chronic forms of the disease, T lymphocytes represent 10 to 20 per cent or more of recovered cells, though the absolute numbers of macrophages are generally increased also. This characteristic cellular picture is not specific for extrinsic allergic alveolitis, but it strongly supports the diagnosis if other suggestive features are present. Other granulomatous disorders, such as sarcoidosis and tuberculosis, hypersensitivity reactions to drugs, and a number of rare lymphoid infiltrative disorders are also associated with a lymphocytosis in lavage fluid, but in practice sarcoidosis is generally the most plausible alternative diagnosis.

In sarcoidosis, B-lymphocyte numbers are decreased and the excess T lymphocytes are typically CD4+ helper cells. The CD4+ to CD8+ ratio normally exceeds 1, and so is exaggerated. By contrast, the ratio is typically reversed in extrinsic allergic alveolitis, CD8+ cells outnumbering CD4+ cells, and B-lymphocyte numbers are not decreased. Lymphocyte markers may therefore help distinguish sarcoidosis from extrinsic allergic alveolitis. Unfortunately, the pattern favouring extrinsic allergic alveolitis does not distinguish so readily between subjects with exposure and symptoms and those with exposure but no symptoms. Both T-cell types show increased numbers if there is exposure, and the number of CD8+ cells tends to show a relatively greater, not lesser, increase in asymptomatic subjects, as described above. The absolute value of the CD4+ to CD8+ ratio therefore provides limited diagnostic benefit in identifying active disease, but this is rarely a relevant issue outside a research setting. A number of cytokines can also be recovered from the lavage fluid, but these are of research rather than diagnostic interest.

Transbronchial or open lung biopsy may occasionally be indicated when other diagnostic procedures do not distinguish extrinsic allergic alveolitis from cryptogenic fibrosing alveolitis or other diffuse infiltrative or fibrotic disorders of the lung. Biopsy is more likely to be needed in the subacute or chronic forms of the disease when hypersensitivity is less obvious, or acutely when there has been an unduly heavy exposure to microbial spores and there is suspicion of microbial invasion.

Environmental exposure

In many cases the history alone provides the evidence of relevant exposure, but this is not always reliable and an independent account of the exposures involved can be invaluable. Ideally, industrial hygiene measurements are made (particularly from personal samplers) so that respirable agents can be recognized and quantified, and microbiological techniques are used to identify specific microbial contaminants. These are sophisticated investigations and usually indicated only when extrinsic allergic alveolitis is first suspected in an environment not previously associated with the disease, particularly in industries where many individuals may be at risk and where modification of the plant and its respirable environment may be a costly matter.

Hypersensitivity

The demonstration of a serum IgG antibody response to the inducing organic dust is the most widely used method of 'confirming' hypersensitivity (saliva may be used more conveniently in children), but this has proved to be unsatisfactory. Although affected subjects tend to have higher antibody levels than those who are exposed but unaffected, the antibody response tends to correlate more closely with exposure than with disease. If the more sensitive enzyme-linked immunosorbent assay (ELISA) is used, rather than the traditional Ouchterlony double-gel diffusion test, even higher rates of false-positive results are obtained. However, this produces greater specificity and in practice the absence of an IgG precipitin response is extremely uncommon in subjects eventually proved to have extrinsic allergic alveolitis, providing they are non-smokers. This is of considerable value in that a negative test generally excludes the diagnosis. The limited value of a positive test is to be expected in view of the current belief that cellular, not humoral, immunity provides the principal mechanism underlying the disease. It is unfortunate that practicable tests for cellular hypersensitivity are not readily available.

When the diagnosis remains in doubt, some form of inhalation challenge test may be necessary. The simplest method involves comparison of experimental periods spent away from the suspected causative environment with similar periods of continuing exposure. The acute form of the disease is likely to be recognized in this way, though the procedure can be time consuming and there may be practical problems of compliance. When a definitive diagnosis is particularly important, laboratory-based inhalation challenge tests can be used. These employ a variety of techniques, ranging from nebulizing soluble extracts to recreating natural environmental exposures in an exposure chamber. The influenza-like component of positive reactions is often uncomfortable, and if excessive doses are administered these tests can be hazardous. What is more, objective evidence for positive reactions may be difficult to obtain from conventional lung function tests. Tests of this nature should therefore be restricted to centres with special expertise. Personal experience of evaluating objective changes in body temperature, circulating neutrophil and lymphocyte numbers, forced vital capacity, and exercise tests from 144 tests is summarized in Table 2. Together they provide high specificity and high sensitivity. Auscultation, chest radiography, measurements of gas transfer, and arterial blood gas analyses are often too insensitive to provide useful diagnostic information.

Differential diagnosis

Acute extrinsic allergic alveolitis is not the only disorder characterized by systemic influenza-like symptoms and respiratory distress to follow an unusually heavy exposure to microbially contaminated vegetable produce. In 1986 an international symposium considered a further disorder that occurs within hours of heavy respiratory exposure to dusts containing fungal toxins, especially those released on decapping silos. It is the result of direct toxicity rather than hypersensitivity and

the term organic dust toxic syndrome was recommended to describe it. Its effects are usually mild and self-limiting, but severe respiratory embarrassment can occur and there is a small risk of ongoing, and potentially fatal, fungal invasion of the lungs. This risk could be enhanced if corticosteroid treatment is given, and death has occurred in subjects who appear to have been fully immunocompetent. Not only does organic dust toxic syndrome occur in circumstances which favour the occurrence of extrinsic allergic alveolitis (particularly silos and swine/poultry confinement buildings), but its clinical features have much in common with extrinsic allergic alveolitis, and to a lesser extent with nitrogen dioxide toxicity, which may also affect silo workers. Indeed, there is so much overlap that it can be very difficult indeed to distinguish one disorder from the others (Table 3).

The acute form of extrinsic allergic alveolitis can only be the result of an acute and recent (a matter of hours) exposure to the relevant causal antigen. This limits the opportunity for diagnostic error, though the circumstances of an unusually heavy exposure may be subtle. For example, a pigeon fancier might spend rather less time than usual with his birds, but much more time than usual in the more hazardous dusty car he uses regularly to transport racing birds for training exercises.

Just as acute and heavy exposures to organic dusts may cause disorders other than extrinsic allergic alveolitis, they may also be quite irrelevant and purely coincidental to the acute respiratory disorder with which the patient presents. Consequently the differential diagnosis should include consideration of other acute disorders of the lung parenchyma and interstitium, such as infections, other immunological disorders, drug reactions, and even paraquat poisoning, which sometimes occurs accidentally in farm workers. In bird keepers the diagnosis of viral, mycoplasmal, and chlamydial infection may itself be confounded by false-positive microbial antibody tests. This is the result of pre-existing avian antibodies cross-reacting with egg protein in the microbial cultures used to provide the test agents.

When subacute or chronic forms of extrinsic allergic alveolitis are encountered, the differential diagnosis lies with other diffuse infiltrative and fibrotic disorders of the lung. Those most frequently resembling extrinsic allergic alveolitis include cryptogenic fibrosing alveolitis, sarcoidosis, pneumoconiosis, tuberculosis, and metastatic cancer, although a huge variety of less common disorders may also need to be considered.

Management

Management of the individual

Management centres on reducing any further exposure to a minimum, though first demands that the diagnosis is secure. There is no place for desensitization. Ideally, the affected individual changes the relevant working, domestic, or recreational environment completely, but this may mean a profound loss in income or great expense, and is often unrealistic. Nor is it fully justified on purely medical grounds since continued exposure does not lead inevitably to progressive disease.

The affected individual who continues to work in the occupation responsible for his disease can often reduce his exposure substantially by changing the pattern of his particular duties. An alternative is the use of industrial respirators, which filter out 98 to 99 per cent of respirable dust from the ambient air. They are especially valuable when exposures are intermittent and short, but may be uncomfortably hot when worn for long periods or when there is heavy work, and so compliance with their use may be poor.

Whatever course is followed, continuing exposure should be accompanied by regular medical surveillance. If there is no progression, it is reasonable for some exposure to continue. When there is progressive disease, exposure should cease. This may involve a loss of earnings, and may entitle the affected worker to compensation. Rarely, the individual with progressive disease will refuse to change his occupation or hobby, and the physician must weigh the possible advantages of long-term corticosteroid therapy against the risks.

Management of the environment

Once extrinsic allergic alveolitis is recognized in one individual, the environment concerned should be assessed for the risk it poses to others. In many circumstances this will be well known already, and exposure levels will be within the range considered acceptable. In others, neither the risk nor the precise causative agent (nor its level of exposure) will be known, and in such unfamiliar circumstances there may be a need to survey the exposed population at risk. Questionnaires and serological tests are most convenient for this, at least as a screening procedure. When large populations are involved, comprehensive investigation is sensible before major modifications are considered to the working environment.

Modifications can always be made to the environment to lessen the level of exposure, but their extent will be limited by expense and should be justified by need. Dry storage and adequate ventilation are the two most important factors when vegetable produce is involved, and in some farming areas there is benefit in drying produce artificially after harvest. An alternative is some form of 'pickling', so that the produce is preserved chemically. With silage, for example, newly cut grass is kept under impervious covering in relatively sealed conditions. Initial enzymic and moulding processes use up available oxygen, and produce aldehydes and other preservative chemicals. These create nearly anaerobic conditions which protect the produce until it is used. Similarly, hay may be sealed in plastic bags, or grain or bagasse may be treated with propionic acid.

When ventilation and humidification systems are themselves responsible for extrinsic allergic alveolitis, major mechanical alterations may be necessary and the methods of humidification and temperature control may need to be changed. The crucial need is to reduce the ease with which normal airborne microbial contaminants are able to proliferate in stagnant collections of water. For this there may be a role for 'biocide' sterilizing agents, but these are also likely to become airborne and respirable and so must have low intrinsic toxicity and sensitizing potency. For one such biocide (isothiazolinone) there have been reports of occupational dermatitis and asthma, though not of extrinsic allergic alveolitis.

The need for rapid air changes coupled with close control of humidity and temperature poses formidable problems. The use of recirculated filtered air is the most economical, but effective filters are expensive and can become contaminated themselves, increasing rather than decreasing the load of respirable microbial antigens. The use of heat exchangers minimizes the cost of temperature control if contaminated exhaust air is not recirculated but does not conserve water.

Outcome

No further exposure

As with occupational asthma, the risk of continuing symptoms following cessation of exposure increases with the duration of exposure. With the acute form of extrinsic allergic alveolitis the exposure period is generally short and the disorder usually resolves without sequelae once the diagnosis is made and exposure ceases. However, one study using bronchoalveolar lavage and DTPA clearance has indicated continuing inflammation and membrane leakiness after a follow up period of 2 to 15 years. The significance of this is unclear, since all subjects were asymptomatic and gave normal results on radiographic and lung function studies.

Continuing exposure

There is greater concern when exposure continues. This may lead not only to recurrent acute attacks but to progressive and permanent fibrotic damage—that is, to the chronic form of the disease. While concern for the risk of progressive fibrosis is undoubtedly justified, this happens in only a minority of affected subjects. A 2- to 40-year follow-up survey of 92 farm workers presenting with the acute form of farmer's lung showed that while most continued to live on farms, only some developed radiographic evidence of pulmonary fibrosis (39 per cent) or impairment of carbon monoxide gas transfer (30 per cent). As many as 28 per cent gave histories of chronic productive cough and 25 per cent had airway obstruction. A similar 10-year outcome has been reported in pigeon fanciers with acute extrinsic allergic alveolitis; again the majority elected to continue their antigenic exposures despite medical advice to the contrary.

Therefore, in some cases—perhaps the majority—important protective mechanisms emerge that lead to tolerance of the effects of further acute exposures, or at least prevent the development of damaging fibrosis. A history of similar increasing tolerance is occasionally noted with occupational asthma, and tolerance not progressive disease is the rule rather than the exception in most animal models of extrinsic allergic alveolitis. However, with both asthma and extrinsic allergic alveolitis some affected subjects give clear accounts of increased responsiveness to a given level of exposure months or years after initial antigen exclusion, which suggests that protective mechanisms may be down-graded more quickly than the causal mechanisms.

As with sarcoidosis, there is debate as to whether the use of corticosteroids for acute episodes confers any long-term benefit. The answer is not clear, but one recent investigation failed to demonstrate any long-term functional differences between groups treated randomly with corticosteroids or placebo for the initial acute episode

of farmer's lung. Whilst the corticosteroid group recovered more quickly from the acute episode, there was the suspicion, already voiced by other investigators, that early steroid therapy carries a greater risk of long-term recurrence. It is possible that the initial response to steroids encouraged less care over subsequent exposures. Alternatively steroid therapy may have induced a different equilibrium between immunological responses, perhaps interfering disproportionately with the development of protective mechanisms.

Compensation of industrial causes

In the United Kingdom, industrial injuries legislation provides compensation from central government for disability in employees (but not employers) from extrinsic allergic alveolitis of occupational origin. The level of disability, and hence compensation, is assessed following examination by a 'Medical Board'. If disability arose before 1991, the affected worker may also be entitled to a 'reduced earnings allowance' if ongoing employment (or lack of it) has resulted in a loss of earned income. Both benefits are limited to a joint maximum figure, which is adjusted from time to time according to inflation. Reduced earnings allowance for new cases was discontinued in 1991.

Acceptance of state compensation in the United Kingdom no longer debars the recipient from seeking redress additionally in the civil courts, which is the primary mechanism of compensation in many countries.

Further reading

Banaszak EF, Thiede WH, Fink JN (1970). Hypersensitivity pneumonia due to contamination of an air conditioner. *New England Journal of Medicine* **283**, 271–6.

Braun SR *et al*. (1979). Farmer's lung disease: long-term clinical and physiologic outcome. *American Review of Respiratory Disease* **119**, 185–91.

Grammar LC (1999). Occupational allergic alveolitis. *Annals of Allergy, Asthma, and Immunology* **83**, 602–6.

Hansell DM, Moskovic E (1991). High-resolution computed tomography in extrinsic allergic alveolitis. *Clinical Radiology* **43**, 8–12.

Hendrick DJ *et al*. (1980). Positive 'alveolar' responses to antigen inhalation provocation tests: their validity and recognition. *Thorax* **35**, 415–27.

Hendrick DJ, Faux JA, Marshall R (1978). Budgerigar fancier's lung: the commonest variety of allergic alveolitis in Britain. *British Medical Journal* **ii**: 81–4.

Kokkarinen JI, Tukiainen HO, Terho EO (1992). Effect of corticosteroid treatment on the recovery of pulmonary function in farmer's lung. *American Review of Respiratory Disease* **145**, 3–5.

Kreiss K, Cox-Ganser J (1997). Metalworking fluid-associated hypersensitivity pneumonitis: a workshop summary. *American Journal of Industrial Medicine* **32**, 423–32.

Leatherman HP *et al*. (1984). Lung T cells in hypersensitivity pneumonitis. *Annals of Internal Medicine* **100**, 390–2.

Meredith SK, Taylor VM, McDonald JC (1991). Occupational respiratory disease in the United Kingdom 1989: a report to the British Thoracic Society and the Society of Occupational Medicine by the SWORD project group. *British Journal of Industrial Medicine* **48**, 292–8.

Morgan DC *et al*. (1973). Chest symptoms and farmer's lung: a community survey. *British Journal of Industrial Medicine* **30**, 259–65.

Ohtani Y *et al*. (1999). Sequential changes in bronchoalveolar lavage cells and cytokines in a patient progressing from acute to chronic bird fancier's lung disease. *Internal Medicine* **38**, 896–9.

Pepys J *et al*. (1963). Farmer's lung. Thermophilic actinomycetes as a source of 'farmer's lung hay' antigens. *Lancet* **ii**: 607–11.

Peterson LB *et al*. (1977). An animal model of hypersensitivity pneumonitis in the rabbit. Induction of cellular hypersensitivity to inhaled antigens using Carageenan and BCG. *American Review of Respiratory Disease* **116**, 1007–12.

Ramirez-Venegas A *et al*. (1998). Utility of a provocation test for diagnosis of chronic pigeon breeder's disease. *American Journal of Respiratory and Critical Care Medicine* **158**, 862–9.

Von Essen S, Donham K (1999). Illness and injury in animal confinement workers. *Occupational Medicine* **14**, 337–50.

Yoshizawa Y *et al*. (1999). Chronic hypersensitivity pneumonitis in Japan: a nationwide epidemiologic survey. *Journal of Allergy and Clinical Immunology* **103**, 315–20.

17.11.12Eosinophilic granuloma of the lung and pulmonary lymphangiomyomatosis

*D. J. Hendrick**

Pulmonary histiocytosis X

Eosinophilic granuloma is a disorder of dendritic (Langerhans) cells that arise from CD34+ progenitors and have special function in presenting antigen to T cells. In this context, however, an abnormal proliferation of these cells is associated with eosinophils (though not peripheral eosinophilia) and granulomatous damage of the lung. Lesions characteristically show cavitation and fibrotic healing, which results in foci of 'stellate' scars, alveolar wall deformation, honeycombing, and the formation of small cysts.

Langerhans cell granulomatosis is the currently favoured generic term rather than histiocytosis X, which was used previously to describe eosinophilic granuloma, Letterer–Siwe disease, and Hand–Schüller–Christian disease. The latter term is used to describe Langerhans granulomas in organs other than the lung (principally bone and the posterior pituitary), and so can be considered a multifocal form of Langerhans cell granulomatosis (eosinophilic granuloma). It is not clear whether Letterer–Siwe disease is truly related to the other two, or whether it is even primarily a disease of Langerhans cells. It is best considered separately. For further discussion, see Section 22.

Eosinophilic granuloma generally affects men and women between the ages of 20 and 40 who are almost exclusively smokers. The main clinical manifestations reflect its chief pathological features—diffuse interstitial fibrosis and cyst formation. The cysts may rupture causing spontaneous pneumothorax, and the fibrosis may progress causing exertional dyspnoea. Cough is common also, and cyst rupture may be associated with pleuritic pain. Because cysts occur bilaterally there is the potential for devastating and immediately life-threatening pneumothoraces to occur on both sides concurrently. One report indicates that occasionally a single pulmonary granuloma may occur, closely simulating carcinoma.

The chest radiograph typically shows diffuse micronodular and reticular changes in the early stages (Fig. 1), which progress with the development of honeycombing. Appearances on high-resolution CT scanning are characteristic (Fig. 1). Pulmonary physiological tests reveal a mixed obstructive and restrictive pattern (decreased vital capacity and total lung capacity, increased residual volume, and decreased carbon monoxide transfer). Bronchoalveolar lavage may reveal Langerhans cells or 'atypical histiocytes', which are most readily identified by immunohistochemical staining of the S100 protein. The major histological characteristics are not readily identified from transbronchial biopsy, and so open biopsy or video-assisted thoracoscopic biopsy is generally required. The major identifying feature of the Langerhans cell is an X body or Birbeck granule, which appears as a pentalaminar rod-like structure in the cytoplasm under the electron microscope.

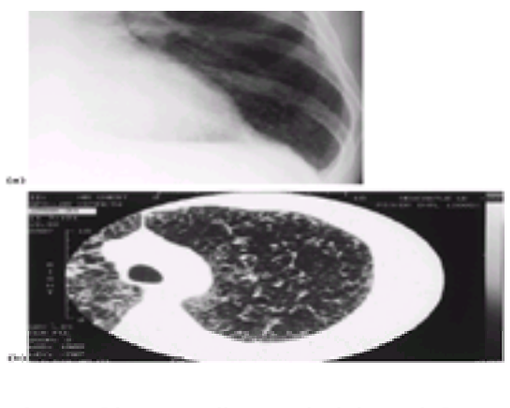

Fig. 1 Chest radiograph (a) and CT scan (b) of a 45-year-old smoking man with biopsy proven (S100 positive) eosinophilic granuloma. The plain film shows a mixed micronodular and reticular pattern which is due principally to multiple small cysts that make the surrounding interstitium more prominent. This is more readily evident from the CT image.

Eosinophilic granuloma of the lung usually occurs in isolation, but may be associated with wider manifestations of multifocal Langerhans cell granulomatosis, particularly skin rashes, bone pain, and diabetes insipidus from granulomas in skin, bone, and pituitary. The clinical course is variable, and treatment in general is supportive rather than specific. In many cases the disease ceases to progress at some point, leaving residual impairment of variable degree. In one case where lung transplantation became necessary, the disease recurred in the transplanted lung, though responded to cyclosphosphamide therapy.

Lymphangiomyomatosis

Lymphangio(leio)myomatosis of the lung is a rare disorder characterized clinically by progressive interstitial disease causing breathlessness and cough, and cyst formation causing spontaneous pneumothorax. Further features are chylous pleural effusion and haemoptysis. It occurs almost exclusively in women of reproductive age, and appears to be a 'forme fruste' of tuberous sclerosis. However, only a minority of patients with tuberous sclerosis show evidence of lymphangiomyomatosis in the lung, and most patients presenting with pulmonary lymphangiomyomatosis do not show other manifestations of tuberous sclerosis. Nor do they appear to pass on the disorder to the next generation. The pulmonary disorder is characterized pathologically by hamartomatous proliferation of smooth-muscle cells in pulmonary lymphatics, venules, and airways. It is often associated with angiomyolipomas of the kidney (and bleeding into the renal tract), and this has been linked with a shared cell surface marker, premelansosome glycoprotein (HMB-45), that can be recognized by immunohistochemical staining.

The typical clinical picture is of progressive loss of exercise capacity associated with a dominantly obstructive loss of ventilatory function and reduced gas transfer. This is interrupted by intermittent episodes of chylous pleural effusion, spontaneous pneumothorax, and haemoptysis. The early chest radiographic appearances are of reticular shadowing and Kerley B lines because of obstructed lymphatics, small nodules due to hamartomatous smooth-muscle aggregates, and pleural effusions. Honeycombing and cystic dilation occur later when pneumothorax becomes increasingly common. These abnormalities are most readily seen by CT scanning (Fig. 2). The diagnosis generally follows lung biopsy, but can been made from cytological examination of aspirated pleural fluid if this contains clusters of immature smooth muscle cells.

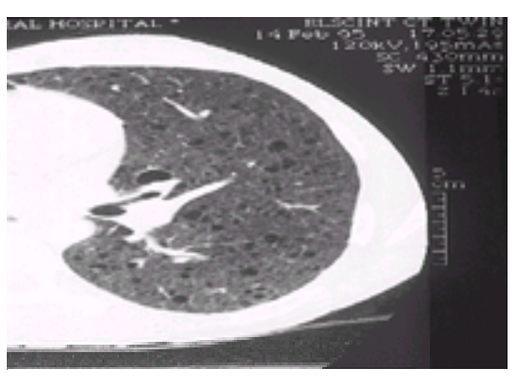

Fig. 2 CT scan of a 37-year-old woman with tuberous sclerosis and pulmonary lymphangiomyomatosis. She had experienced sequential spontaneous

pneumothoraces affecting each side. The scan shows multiple thin-walled cysts throughout the lung. (By courtesy of Dr S. J. Bourke.)

Progesterone therapy or bilateral oophorectomy is sometimes effective in limiting disease progression, and may also help to control pleural effusions. If untreated, most patients die within 10 years, and so lung transplantation becomes the principal option if anti-oestrogen therapy is ineffective. In one case this was associated with recurrent disease in the allograft.

---

*Dr R. J. Shaw wrote on this subject in the third edition of the *Oxford Textbook of Medicine*. Much of his text has been retained in this revision and we acknowledge his contribution with grateful thanks.

Further reading

Anonymous (1994). Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Case 5–1994. A 34-year-old woman with mild exertional dyspnea and interstitial pulmonary lesions [clinical conference]. *New England Journal of Medicine* **330**, 347–53.

Anonymous (1994). Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Case 1–1994. A 37-year-old woman with interstitial lung disease, renal masses, and a previous spontaneous pneumothorax [clinical conference]. *New England Journal of Medicine* **330**, 1300–6.

Bonelli FS *et al*. (1998). Accuracy of high-resolution CT in diagnosing lung diseases. *American Journal of Roentgenology* **170**, 1507–12.

Brauner MW *et al*. (1997). Pulmonary Langerhans cell histiocytosis: evolution of lesions on CT scans. *Radiology* **204**, 497–502.

Chan JK *et al*. (1993). Lymphangiomyomatosis and angiomyolipoma: closely related entities characterized by hamartomatous proliferation of HMB-45-positive smooth muscle. *Histopathology* **22**, 445–55.

Costello LC, Hartman TE, Ryu JH (2000). High frequency of pulmonary lymphangioleiomyomatosis in women with tuberous sclerosis complex. *Mayo Clinic Proceedings* **75**, 591–4.

Gabbay E *et al*. (1998). Recurrence of Langerhans' cell granulomatosis following lung transplantation. *Thorax* **53**, 326–7.

Lieberman PH *et al*. (1996). Langerhans cell (eosinophilic) granulomatosis. A clinicopathologic study encompassing 50 years. *American Journal of Surgical Pathology* **20**, 519–52.

Taylor JR *et al*. (1990). Lymphangioleiomyomatosis: clinical course in 32 patients. *New England Journal of Medicine* **323**, 1254–60.

17.11.13 Pulmonary alveolar proteinosis

D. J. Hendrick

First described in 1958, pulmonary alveolar proteinosis is a rare but interesting disorder that exerts its primary effects in the alveolar spaces. Over a period ranging from months to years, these become filled with an amorphous, largely cell-free, lipoproteinaceous material that is not readily expectorated. Inflammation and fibrosis are conspicuously absent and there are two major consequences. First, depending on the number of alveoli involved, the lungs become stiff, ventilatory function becomes restricted, and shunting occurs at the alveolar capillary level, causing hypoxaemia. The outcome is breathlessness, reduced exercise tolerance, and in some cases death from respiratory failure. Second, and a not uncommon cause of death, is secondary infection. The responsible organisms are generally those that are associated with intracellular infection and impaired T-lymphocyte function, nocardia being particularly prominent. In many cases, however, extensive involvement does not occur, there being little or no progression, or even spontaneous remission. Epidemiological data are scarce but one estimate suggests an annual incidence of the order 2 to 5 per million worldwide.

Pathogenesis

Men appear to be affected more commonly than women: all age groups may be involved, and smoking may enhance the risk. The cause in most cases is unknown, but an apparently identical (though relentlessly progressive) disorder can arise within months of massive exposure to respirable mineral dust, especially silica—both in the unfortunate worker exposed negligently without adequate respiratory protection and in experimental animal models. This has been called acute silicoproteinosis or silicolipoproteinosis. Less commonly aluminium dust may be responsible, and there have been reports implicating titanium and insecticides. A few reports describe affected siblings, implying a possible hereditary factor, and some associate pulmonary alveolar proteinosis with haematological disorders (usually malignant and often after the use of cytotoxic agents) or immunodeficiency disorders.

The secreted material is rich in protein and phospholipid, and stains strongly with periodic acid–Schiff (**PAS**) and eosin. It also contains structures resembling tubular myelin, which are derived from lamellar bodies of surfactant-producing type II pneumocytes. The secretions themselves are chiefly the product of these cells, and the chief phospholipid is dipalmitoyl phosphatidylcholine—the dominant phospholipid of normal surfactant. It is unclear, however, whether the accumulation of these secretions results from an abnormality of the type II pneumocytes (excessive or abnormal production), or from impaired resorption by alveolar macrophages. Recent research suggests that both mechanisms may be relevant, also that the pulmonary alveolar proteinosis phenotype can arise through a number of different processes. In most cases the PAS stain is taken up uniformly, as is peroxidase-labelled immunoglobulin raised against the apolipoproteins of surfactant. In others, particularly those associated with haematological or immunological disorders, uptake is heterogeneous, and it has been suggested that fundamentally different processes underlie this 'secondary' form of pulmonary alveolar proteinosis.

Mice deficient in granulocyte-macrophage colony-stimulating factor (**GM-CSF**), or with a disrupted GM-CSF receptor, develop alveolar accumulations of surfactant material, simulating pulmonary alveolar proteinosis. Benefit occurs when GM-CSF is administered locally, when normal bone marrow is transplanted, or when the faulty gene is replaced by some other method. This suggests that pulmonary alveolar proteinosis could arise because of impaired macrophage differentiation and function as a consequence of an inherited genetic defect, or an acquired defective bone marrow clone. The vulnerability to infection with opportunistic organisms and the *in vitro* demonstration of a number of abnormalities of macrophage function additionally incriminate the macrophage more than the pneumocyte. This is also consistent with pulmonary alveolar proteinosis arising after macrophage function has been disrupted by massive exposure to silica, and with pulmonary alveolar proteinosis being associated with impaired T-cell immunity (and hence diminished macrophage activation).

It has been shown, however, that ingestion of material produced in pulmonary alveolar proteinosis may itself cause impairment of phagocytic function in macrophages harvested from normal controls, and so in some cases the primary abnormality could still lie with the type II pneumocyte and its alveolar secretions. This would be consistent with the belief that chemotherapeutic agents associated with the development of pulmonary alveolar proteinosis (e.g. bleomycin) are more likely to damage pneumocytes than macrophages. Strong support for this view comes from the recent demonstration of a GM-CSF inhibitory factor within the bronchoalveolar secretions of pulmonary alveolar proteinosis, to which GM-CSF binds more avidly than it does to the cells it should stimulate. Its source of origin, however, has not yet been shown to be the type II pneumocyte.

Clinical features

The patient usually presents with progressive shortness of breath due to the disease itself or with a pneumonic illness due to superimposed infection. Occasionally, the disease is without symptoms and is first recognized from the appearances of an incidental chest radiograph, when it may be mistaken for sarcoidosis. Cough is common and may be productive, particularly if there is infection. Low-grade fever, haemoptysis, and pleuritic pain occur infrequently, though some authors report an initial febrile incident. There may be crackles and clubbing in advanced stages, and fever becomes typical when infection supervenes. When nocardia is not responsible for this, aspergillus, candida, cryptococcus, cytomegalovirus, histoplasma, HIV, mucor, mycobacteria, pneumocystis, and viruses are the most common culprits.

Diagnosis

The chest radiograph characteristically shows an alveolar filling pattern, which radiates from the hila and simulates pulmonary oedema. There is no associated evidence of heart failure, however, and the appearances may be somewhat patchy and asymmetrical. Diffuse pulmonary fibrosis is very rare, unless provoked by complicating infection. A micronodular infiltration is occasionally seen, particularly in children, but lymphadenopathy is usually absent. CT scanning, particularly with high resolution, shows the non-specific features of air space filling (ground-glass attenuation), and commonly a patchiness which distinguishes affected from unaffected lobules. There may also be septal thickening and hence the 'crazy paving' appearance typical of combined alveolar and interstitial disease.

Pneumonia or aspiration is often suspected initially, but the cough produces little or no sputum and no organisms are isolated if the disease remains uncomplicated. Occasionally, white gelatinous material is expectorated, and bronchoalveolar lavage fluid is typically milky in colour. Gallium scanning may be useful in showing negligible pulmonary uptake in contrast to the findings in pneumonia. In established pulmonary alveolar proteinosis, a positive gallium scan (or a CT scan) may be invaluable in suggesting the development of superimposed infection.

The key to the diagnosis of uncomplicated pulmonary alveolar proteinosis rests with the demonstration that the alveolar secretions are strongly PAS-positive but contain no organisms and no excessive cellular response. Indeed, the macrophages appear to be deficient in numbers as well as function. Biochemical and immunochemical tests may be used to show that phospholipids and specific surfactant proteins are present in excess. Occasionally the sputum provides diagnostic material, identification of lamellar bodies or their debris by electron microscopy being particularly useful. These may be found within macrophages or pneumocytes, or may lie free within the secretions. More commonly, bronchoalveolar lavage or transbronchial lung biopsy is required, though the former should suffice since PAS-positive amorphous globules demonstrated by cytological smears have high diagnostic specificity. Ultrastructural examination shows that these also generally contain multilaminated structures derived from lamellar bodies. The characteristic histological features are shown in Fig. 1.

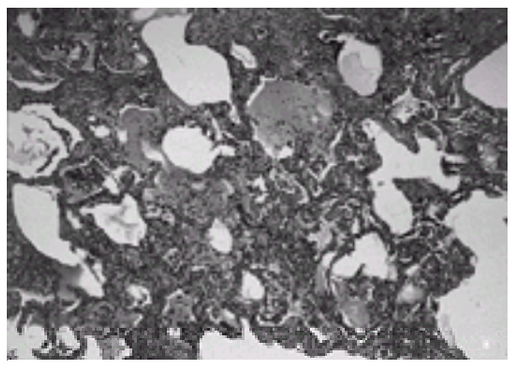

Fig. 1 Pulmonary alveolar proteinosis arising acutely following massive exposure to silica (by courtesy of Dr D.E. Banks). Some alveoli are filled with a non-inflammatory proteinaceous exudate, characteristic of pulmonary alveolar proteinosis. The lung interstitium shows fibrosis and inflammation which can be attributed to acute silicosis (haematoxylin and eosin, medium magnification).

Raised serum levels of surfactant proteins and lactic dehydrogenase are to be expected, but these occur in a number of other diffuse disorders of the lung and so are of no value in diagnosis. Serum levels of the mucin-like glycoprotein KL-6, which is released from type II pneumocytes, are also raised in pulmonary alveolar proteinosis. Although not specific, it has been suggested that these provide a good biochemical correlate for disease activity, and may therefore be a useful marker in assessing when treatment becomes necessary.

Management

In a third to a half of cases, no appreciable disability develops and the disease remits spontaneously or fails to progress. The choice of treatment, when necessary, is strictly limited. Corticosteroids are of no value and may increase the risk of infection. Prolonged periods of inhalation therapy with expectorants (potassium iodide) or proteolytic enzymes (trypsin) have been claimed to offer benefit, but frequently cause irritative responses in the airways. Furthermore, trypsin does not digest *in vitro* the material produced in pulmonary alveolar proteinosis. Neither form of treatment is currently recommended, although the addition of trypsin to therapeutic bronchalveolar lavage fluid has been reported to be both effective and well tolerated.

The most effective measure is physical removal of secretions by bronchoalveolar lavage, usually performed under general anaesthesia using a double-lumen endotracheal tube. The treatment is repeatedly carried out on one lung with a total of 20 to 50 litres of warm sterile buffered saline while the other is mechanically ventilated. The procedure is then reversed so that the other lung is treated. The practice of adding heparin and acetyl cysteine to the lavage fluid has not been shown to be beneficial, but chest percussion during the procedure does seem to enhance the yield. When severe respiratory failure has already supervened despite ventilatory support, cardiopulmonary bypass has been used successfully to maintain gas exchange during the lavage procedure. An alternative is sequential lobar lavage using a fibreoptic bronchoscope and a cuffed catheter. Further lavage is usually necessary every few weeks or months, but the activity of the disease may lessen and the need for frequent treatment may diminish.

Sometimes there is fatal progression with marked loss of weight despite repeated lavage. It may be that current experimental use of GM-CSF might lead to an improved outlook for such patients in the future. In one prominent case treatment involved bilateral lung transplantation, but this was followed by recurrent pulmonary alveolar proteinosis, raising the question of whether bone marrow transplantation might be more appropriate.

The risk of premature death in most series has been low (mostly less than 10 per cent), but a considerable threat to life is associated with complicating infection. This should be recognized and treated promptly. An accelerated clinical course together with the development of fever, increased (and productive) cough, malaise, evidence of systemic illness, and the radiographic demonstration of cavitation or pleural effusion all provide pointers to its development. Blood cultures together with smear and culture studies of sputum may identify the organism or organisms responsible, but often bronchoscopy with brushings and diagnostic lavage is needed. Sometimes a biopsy procedure is considered necessary, particularly when the underlying presence of alveolar proteinosis is not clearly established. When opportunistic organisms are involved, the eradication of infection may prove to be unduly difficult, perhaps reflecting an underlying impairment of macrophage (or GM-CSF) function. It has therefore been argued that regular bronchoalveolar lavage, even in the absence of impaired exercise tolerance, may limit the degree of immunosuppression and provide valuable prophylaxis against such life-theatening infections. A recent study did indeed demonstrate improved macrophage function following lavage, which slowly diminished over 18 months as clinical relapse occurred. If the argument is followed fully, lavage may also play a role in eradicating the acute infection.

Further reading

Burkhalter A *et al.* (1996). Bronchoalveolar lavage cytology in pulmonary alveolar proteinosis. *American Journal of Clinical Pathology* **106**, 504–10.

Claypool WD, Rogers RM, Matuschak GM (1984). Update on the clinical diagnosis, management, and pathogenesis of pulmonary alveolar proteinosis (phospholipidosis). *Chest* **85**, 550–8.

Gaine SP, O'Marcaigh AS (1998). Pulmonary alveolar proteinosis: lung transplant or bone marrow transplant. *Chest* **113**, 563–4.

Goldstein LS *et al.* (1998). Pulmonary alveolar proteinosis: clinical features and outcomes. *Chest* **114**, 1357–62.

Huffman JA *et al*. (1996). Pulmonary epithelial cell expression of GM-CSF corrects the alveolar proteinosis in GM-CSF-deficient mice. *Journal of Clinical Investigation* **97**, 649–55.

Mikami T *et al.* (1997). Pulmonary alveolar proteinosis: diagnosis using routinely processed smears of bronchoalveolar lavage fluid. *Journal of Clinical Pathology* **50**, 981–4.

Reed JA *et al.* (1999). Aerosolized GM-CSF ameliorates pulmonary alveolar proteinosis in GM-CSF-deficient mice. *American Journal of Physiology* **276**, L556–63.

Selecky PA *et al.* (1977). The clinical and physiological effect of whole lung lavage pulmonary alveolar proteinosis: A 10 year experience. *Annals of Thoracic Surgery* **24**, 451–61.

Seymour JF *et al*. (1996). Efficacy of granulocyte-macrophage colony-stimulating factor in acquired alveolar proteinosis. *New England Journal of Medicine* **335**, 1924–5.

Tanaka N *et al.* (1999). Lungs of patients with idiopathic pulmonary alveolar proteinosis express a factor which neutralizes granulocyte-macrophage stimulating factor. *FEBS Letters* **442**, 246–50.

Wang BM *et al.* (1997). Diagnosing pulmonary alveolar proteinosis: a review and an update. *Chest* **111**, 460–6.

17.11.14 Pulmonary amyloidosis

D. J. Hendrick

Clinically important involvement of the lungs in amyloidosis is extremely uncommon, only a few dozen cases having been reported over recent decades, but lung infiltration is said to occur in the majority of cases of primary (**AL**) amyloid, less commonly in secondary (**AA**) amyloid. The proteinaceous material which is responsible for amyloid infiltration is composed of a fibrillar polypeptide and a non-fibrillar glycoprotein. They produce a unique b-pleated structure which may be deposited progressively and widely, eventually interfering with organ function. The glycoprotein (amyloid P or AP protein) comprises a mere 10 per cent of amyloid tissue. It is derived from a parent serum protein (SAP) made in the liver and is common to all types of amyloid tissue. The fibrillar polypeptide on the other hand is of two distinct types (Fig. 1).

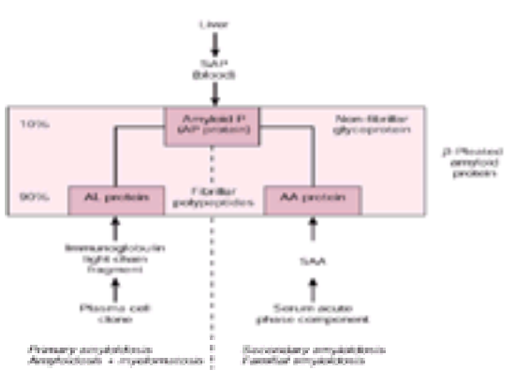

Fig. 1 Pathogenic pathways underlying primary and secondary amyloidosis.

The type seen with 'primary' amyloidosis and amyloidosis associated with myeloma is derived from the variable region of immunoglobulin light chains and is known as AL protein. It is the product of a plasma cell clone, whether benign or malignant. Monoclonal immunoglobulin (M-compo-nent) may sometimes be detected in the serum by electrophoresis, and free light chains (Bence Jones protein) may be detected in the urine. The type of fibrillar polypeptide (AA protein) seen with 'secondary' amyloidosis and most of the familial forms of amyloidosis (such as Mediterranean fever) is derived from an acute phase serum component (SAA). The chronic inflammatory diseases associated with persistently raised levels of SAA (and C reactive protein and interleukin-6) are those that are most commonly associated with secondary amyloidosis, but only a small minority of subjects affected by these chronic inflammatory disorders show evidence of complicating amyloidosis.

Clinical features

Those affected are usually middle aged or elderly, and the sexes are equally represented. Hilar or mediastinal lymphadenopathy is rarely demonstrated on plain chest radiographs, but mild nodal enlargement is commonly seen on CT scanning. The pleura are rarely involved, but recurrent pleural effusion may occur.

Symptomatic pulmonary involvement is generally a consequence of localized disease affecting the upper and central airways, or systemic disease at alveolar–interstitial level. Localized disease is usually due to deposition of amyloid containing the light chain derived AL protein, and is essentially benign in nature. Its effects depend on the site of deposition. Diffuse alveolar–interstitial disease is usually associated with systemic disease and may be the consequence of either AL or AA protein deposition. It carries a poor prognosis, though death is usually due to the involvement of organs other than the lung.

The following varieties of pulmonary amyloidosis, in descending order of epidemiological importance, are the most clearly recognized.

1. Localized laryngotracheobronchial: discrete, and often multiple, masses of amyloid protein enlarge in the walls of the airways or the peribronchial tissues causing cough, obstruction, and sometimes bleeding. Obstruction of airways may lead to wheeze, stridor, breathlessness, atelectasis, and infection, and may eventually give rise to bronchiectasis. When a single lesion is involved it may simulate the effects of a bronchial adenoma, appearing as a polypoid mass on endoscopic inspection.
2. Localized parenchymal nodule(s): discrete nodules or masses, which may be single or multiple and may occasionally reach the size of a tennis ball, are seen within the lung parenchyma on the chest radiograph. They rarely cause symptoms or disrupt lung function and may eventually calcify, cavitate, or even ossify. They are likely to simulate bronchial neoplasms if single and so be resected; if multiple they often lead to biopsy. Biopsy in one unusual case (that of an HIV-positive intravenous drug abuser) showed no evidence of AL protein, but there was focal birefringency and a foreign body giant cell reaction reflecting deposition of the carrier material of the illicit drug. This implies that focal inflammation within the lung may occasionally lead to localized 'secondary' amyloid (i.e. AA protein) deposition. In the future it may be that CT and MRI will offer a useful means of distinguishing amyloid tissue from tumour, so obviating the need for biopsy.
3. Diffuse alveolar–interstitial: amyloid tissue is deposited diffusely throughout the alveolar walls and interstitium of the lung, and is usually a feature of systemic amyloidosis (Fig. 2 and Fig. 3). There is progressive breathlessness and dry cough. Scattered crackles are characteristic and there may be pleural effusions. Eventually respiratory failure may supervene as ventilation becomes increasingly restricted and gas transfer impaired, though death more commonly results from cardiac or renal involvement.

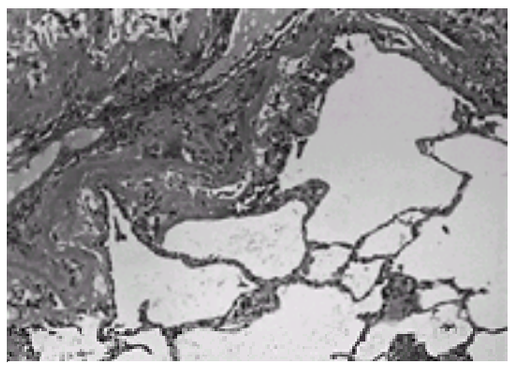

Fig. 2 Amyloidosis of the lung: alveolar-interstitial type [i] (by courtesy of Dr T. Ashcroft). There are interstital deposits of hyaline eosinophilic material with a foreign body type giant cell response in adjacent tissue. This is an almost unique feature of amyloidosis affecting the lung (haematoxylin and eosin, medium magnification).

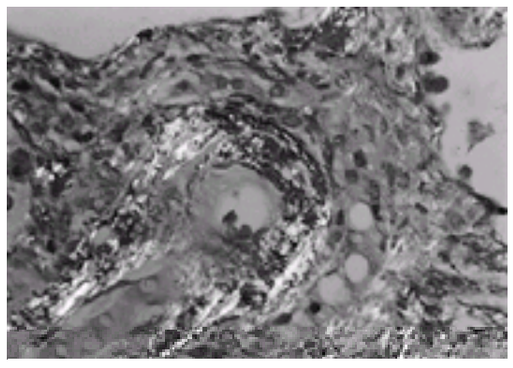

Fig. 3 Amyloidosis of the lung: alveolar-interstitial type [ii] (by courtesy of Dr T. Ashcroft).Amyloid gives a characteristic dichroic birefringence (Congo red stain under polarized light. High magnification).

Histological examination may also show evidence of amyloid infiltration of the pulmonary vasculature. This is usually of no clinical consequence, but has been reported to cause pulmonary hypertension and undue bleeding after biopsy, although other reports suggest that biopsy, particularly transbronchial biopsy, is generally both safe and effective. Another rare effect of amyloidosis on respiratory function is enlargement of the tongue, which can cause or exacerbate obstructive sleep apnoea.

Diagnosis

The diagnosis rests on the demonstration of amyloid tissue in an affected organ. When the protein is derived from plasma cells or lymphocytes, it may be possible to demonstrate light chains in the urine or M-component in the serum, and a plasma cell or lymphocyte dyscrasia may be clinically evident. When systemic reactive amyloidosis is the diagnosis, a provoking chronic inflammatory disease should be obvious, and high levels of SAA will be present in the serum. Measurement of SAA is rarely routinely available, but C-reactive protein is a useful surrogate. Immunohistochemical studies should, in any event, identify the specific biochemical nature of the protein sampled at biopsy, as should the ultrastructural appearances at electron microscopy.

Management

For discussion of the treatment of systemic amyloidosis, which is often unrewarding, see Section 11.12. Ultimately, organ transplantation may become the only hope of survival, and when renal failure or cardiac failure is the only immediate threat to life, this is often carried out. A need for lung transplantation has not yet been reported.

With the local forms of the disease of whatever aetiology, the outlook is a good deal brighter. Progression may be slow, and the disease may become quiescent. The laryngotracheobronchial deposits can sometimes be resected or depleted piecemeal endoscopically, perhaps using laser therapy, but there is some risk of serious bleeding from this. Corticosteroids have been reported to have a beneficial effect when there is critical airway stenosis. Parenchymal nodules in the lung rarely need to be removed, providing their histological nature is not in doubt.

Further reading

Ihling C *et al.* (1996). Amyloid tumors of the lung—an immunocytoma? *Pathology, Research and Practice* **192**, 446–52.

Kavuru MS *et al.* (1990). Amyloidosis and pleural disease. *Chest* **98**, 21–3.

Matsumoto K *et al.* (1997). Primary solitary amyloidoma of the lung: finding on CT and MRI. *European Radiology* **7**, 586–8.

Miyamoto T *et al.* (1999). Monoclonality of infiltrating plasma cells in primary pulmonary nodular amyloidosis: detection with polymerase chain reaction. *Journal of Clinical Pathology* **52**, 464–7.

Pickford HA, Swensen SJ, Utz JP (1997). Thoracic cross-sectional imaging of amyloidosis. *American Journal of Roentgenology* **168**, 353–5.

Rubinow A *et al.* (1978). Localized amyloidosis of the lower respiratory tract. *American Review of Respiratory Disease* **118**, 603–11.

Shah SP *et al*. (1998). Nodular amyloidosis of the lung from intravenous drug abuse: an uncommon cause of multiple pulmonary nodules. *Southern Medical Journal* **91**, 402–4.

Shiue ST, McNally DP (1988). Pulmonary hypertension from prominent vascular involvement in diffuse amyloidosis. *Archives of Internal Medicine* **148**, 687–9.

Utz JP, Swensen SJ, Gertz MA. Pulmonary amyloidosis. The Mayo Clinic experience from 1980 to 1993. *Annals of Internal Medicine* **124**, 407–13.

17.11.15 Lipoid (lipid) pneumonia

D. J. Hendrick

Exogenous

When mineral or vegetable lipids are deposited in the lung, they usually prove to be relatively inert but difficult to remove. Lung lipases have little effect, and the macrophages are slow to transport the free or emulsified material into the lymphatics. The result is often a chronic low-grade inflammatory response that may lead to secondary infection and/or local fibrosis. It is known as lipoid, or lipid, pneumonia. It should be suspected whenever a 'pneumonic' illness is slow to resolve or is recurrent, especially if there is the possibility of impaired swallowing and recurrent aspiration. Some animal lipids are more readily degraded by lung lipases, thus releasing irritating fatty acids. In these circumstances a brisk pneumonitis may occur.

Pathogenesis

Aspiration of vegetable or mineral oil is not common in the population at large, but is seen not infrequently within certain subgroups—particularly those with impaired swallowing mechanisms. Most affected are the very young and the elderly, and the regular nasal instillation of vegetable oils (e.g. olive oil) or paraffin to relieve congestion, or their ingestion to relieve constipation, are often responsible. A portion of any nasal dose is likely to enter the trachea, as may part of an ingested dose if the subject then reclines in bed or has any disturbance of swallowing. The critical point is that paraffin and many vegetable oils are not irritating to the tracheal mucosa, and so coughing is rarely excited and aspiration occurs without immediate sequelae.

The reluctant child forced to swallow cod liver oil is said to have encountered similar risks during the 1940s and 1950s, though infants are likely to face much greater hazard by virtue of less well developed deglutition. Patients fed by nasogastric tube are particularly vulnerable, as are those fed regularly with high lipid diets (for example with milk fat, ghee). A recent case report indicates a rather less obvious risk in an infant—that of lipid embolism from repeated mineral oil enemas. Lipoid pneumonia from embolism may also occur because of intravenous infusion, whether accidental or wilful.

Adults with unimpaired swallowing are affected only sporadically. Shipwrecked sailors have occasionally aspirated floating oil, and lipoid pneumonia has been recognized in workers exposed to oil mists and burning fats. In an exceptional case associated with recurrent tuberculosis, the spontaneous rupture of a longstanding oleothorax (oil plombage) led to a widespread iatrogenic lipoid pneumonia. Aspirated petroleum products such as kerosene may be absorbed from the lung and give rise to toxic responses in other organs (particularly the heart), and this may prove to be life threatening. The potential risk from oil aerosols has been highlighted recently in a diver breathing unfiltered air from an oil-contaminated surface compressor. Less unwilling inhalers of mineral oil and vaseline have been the blackfat tobacco smokers of Guyana, who obtain more satisfaction when these additives are mixed to native tobacco leaf. A distinctive geographical picture of progressive and often fatal pulmonary fibrosis complicates this habit in some 20 per cent of blackfat users, but has not been observed among non-smoking residents.

Clinical features

If there is little or no pulmonary response, there may be no symptoms, and the affected subject presents by chance with an abnormal chest radiograph. In about 50 per cent of cases there is a chronic 'pneumonic' illness with productive cough, low-grade fever, and (occasionally) haemoptysis. Often there is a cyclical course with intermittent symptoms. Repeated aspiration may lead to fibrotic shrinkage of the affected segment or segments (usually in the lower lobes or the middle lobe), bronchiectasis, or persistent consolidation. The radiographic appearances may closely simulate bronchial carcinoma, and many resections have been carried for this reason, sometimes revealing a characteristic granulomatous mass (paraffinoma). When more substantial quantities are aspirated the radiographic abnormalities are necessarily more diffuse, and when the lipid material is more reactive an acute 'pneumonic' illness occurs.

Diagnosis

The key to diagnosis is the demonstration of lipid material within pulmonary secretions or alveolar macrophages, whether obtained from sputum or bronchoalveolar lavage. If lung tissue is resected or a biopsy is taken, there may be fibrosis, evidence of chronic inflammation, and foreign body granulomas/giant cells in addition to lipid material retained within alveoli and macrophages (Fig. 1). An innovative use of computed tomography has recently identified excess deposits of lipid from its X-ray absorption characteristics, a technique that could offer a valuable alternative to biopsy or bronchoalveolar lavage in the diagnosis of atypical pneumonias. More generally, computed tomography shows patchy areas of ground-glass attenuation and interstitial thickening, thereby producing a 'crazy paving' pattern. This may be seen more readily with high resolution scans, which may additionally show interspersed poorly defined small nodules. Nuclear magnetic resonance scanning appears less effective, though a loss of signal intensity has been reported to have diagnostic specificity.

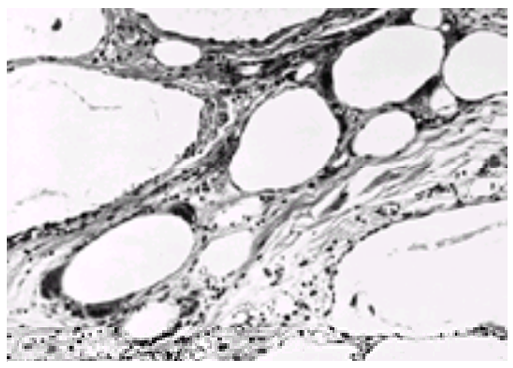

Fig. 1 Lipoid pneumonia (by courtesy of Dr T. Ashcroft). Exogenous lipoid pneumonia due to aspirated paraffin. There is interstitial fibrosis containing oil vacuoles which are enclosed within multinucleated giant cells (haematoxylin and eosin stain, medium magnification).

Management

Prophylactic management centres on minimizing any tendency to aspiration associated with impaired swallowing, and in persuading the misuser (or abuser) of vegetable and mineral oils to adopt alternative habits. Once aspiration has occurred there may be a role for therapeutic bronchoalveolar lavage, since this may remove substantial quantities of lipid from the alveoli. During episodes of secondary bacterial infection, there is an obvious role for antibiotics, and when there is acute inflammation corticosteroids are sometimes used.

Endogenous

The body may itself produce and retain lipid (mainly cholesterol) within the lungs, though this is not a common phenomenon. It occurs chiefly at sites of chronic inflammation, obstruction, or tissue necrosis, and is derived from the necrotic cells. This lipid will also be ingested by macrophages and may be recovered in the sputum. Sputum macrophages laden with lipid are not therefore pathognomonic of aspiration from an exogenous source, though chemical tests can distinguish the two varieties and histological examination of affected lung does not show a granulomatous response to endogenous lipid. Endogenous lipid is most commonly deposited when chronic inflammation accompanies bronchiectasis, bronchial carcinoma, or some other cause of persisting localized bronchial obstruction, and appears to depend on cigarette smoking. The radiological appearances are of a persisting pneumonia, which may also stimulate resection for fear a carcinoma is present. Very recently endogenous lipoid pneumonia in a more diffuse form has been associated with the use of amiodarone in patients dying with adult respiratory distress syndrome.

Further reading

Annobil SH *et al.* (1991). Lipoid pneumonia on children following aspiration of animal fat (ghee). *Annals of Tropical Paediatrics* **11**, 87–94.

Corrin B, Soliman SS (1978). Cholesterol in the lungs of heavy smokers. *Thorax* **33**, 565–68.

Cox JE, Choplin RH, Chiles C (1996). Case report. Chemical-shift MRI of exogenous lipoid pneumonia. *Journal of Computer Assisted Tomography* **20**, 465–7.

Donaldson L *et al.* (1999). Acute amiodarone-induced lung toxicity. *Intensive Care Medicine* **25**, 242–3.

Gondouin A *et al*. (1996). Exogenous lipid pneumonia: a retrospective multicentre study. *European Respiratory Journal* **9**, 1463–9.

Laurent J *et al.* (1999). Exogenous lipoid pneumonia: HRCT, MR, and pathologic findings. *European Radiology* **9**, 1190–6.

Lee JS *et al*. (1999). Exogenous lipoid pneumonia: high-resolution CT findings. *European Radiology* **9**, 287–91.

Miller GJ *et al.* (1971). The lipoid pneumonia of blackfat tobacco smokers in Guyana. *Quarterly Journal of Medicine* **40**, 457–70.

Oldenburger D *et al*. (1972). Inhalation lipid pneumonia from burning fats. *Journal of the American Medical Association* **222**, 1288–9.

Silverman JF *et al.* (1989). Bronchoalveolar lavage in the diagnosis of lipid pneumonia. *Diagnostic Cytopathology* **5**, 3–8.

17.11.16 Pulmonary alveolar microlithiasis

D. J. Hendrick

This is a very rare disorder, with only 200 to 300 cases reported since its initial description in 1918. It is remarkable for a number of unusual if not unique features. Tiny 0.2- to 5-mm calcified concretions, which may be concentrically laminated, form progressively in the alveolar spaces, usually with some degree of interstitial fibrosis. As their profusion slowly increases in the lung, they produce a striking 'white-out' appearance on the chest radiograph as the border of one intensely radio-opaque microlith overlaps that of another, even if the two are not immediately adjacent. There are few clues to the cause of this curious disorder and, apart from transplantation, there is no effective means of therapy.

Pathogenesis

No abnormality of calcium metabolism has been demonstrated, but a disproportionate number of cases (almost half) occurs in siblings and in countries bordering the eastern Mediterranean. This points to a genetic rather than environmental basis, perhaps autosomal recessive inheritance. One hypothesis suggests that the calcific response is directed to an alveolar exudate (whether induced by inhaled particles or microbes) that cannot be reabsorbed adequately. The primary explanation could thus lie with an unusual inhalant stimulus, a deranged alveolar exudative response, or a failure of the usual mechanisms of reabsorption.

That similar concretions are not generally noted in other sites of membrane reabsorption possibly diminishes the probability of the last of these, though microliths have been noted in seminal vesicles in some cases, and azoospermia in others. Analytical studies, including X-ray energy spectroscopy and microscopic infrared spectroscopy, have showed no evidence of mineral dust deposition, and no microbial cause has been confirmed. The possibility of a deranged alveolar response has consequently attracted particular interest, perhaps an abnormality of surfactant function following oxidant stress. The calcified microliths (salts of calcium and phosphorus, and calcium carboxyapatite) are formed in the alveolar spaces, and some can be flushed out by bronchoalveolar lavage or even expectorated in sputum.

Clinical features

The disease usually presents in middle age, but the whole age spectrum may be involved and both sexes are equally represented. Almost invariably the patient is symptom free when an initial film is taken for incidental reasons, and there may be wonder that this can be possible when the radiograph is grossly abnormal. This is a consequence of there being no associated cellular, exudative, fibrotic, or vascular disruption of normal physiological processes in the early stages of the disease. Physical signs are conspicuous by their absence for most of its long course, though crackles, clubbing (even hypertrophic pulmonary osteoarthropathy), and signs of respiratory failure may be observed ultimately as the alveolar spaces are progressively filled with microliths and fibrosis of the interstitium advances. Shunting occurs at alveolar-capillary level causing hypoxaemia, and the bronchial circulation contributes increasingly to pulmonary venous return. In some cases subpleural cysts give rise to spontaneous pneumothoraces, and pleural adhesions may become prominent.

Although death supervened rapidly in the reported cases of two newborn infants, survival of 10 to 20 years is characteristic, and may be much longer. In most cases there is slowly progressive undue breathlessness with dry cough. Haemoptysis occurs occasionally. The lungs stiffen, ventilation becomes restricted, and gas transfer is impaired. Eventually respiratory failure and cor pulmonale supervene. At death, extensive areas of the chest radiograph show a dense 'white-out' appearance due to the considerable accumulation of calcium, the lungs are difficult to cut, and they sink in water.

Diagnosis

The radiographic appearances of profuse small calcified nodules are almost specific, particularly in moderately advanced cases when the dense 'white-out' picture is seen but symptoms are still absent or unimpressive. One case of sarcoidosis has been reported with similar appearances. With less advanced disease, biopsy, bronchoalveolar lavage, or expectorated sputum should provide diagnostic material, but with transbronchial biopsy it may prove difficult to close the forceps and extract them through the fibreoptic bronchoscope. Initially the chest radiograph shows a mere haziness of the lower zones, and computed tomography may be invaluable in demonstrating the nodular shadows and their calcific nature. It may also confirm an early predominance for the basal and posterior segments. High resolution images may also demonstrate the presence of interstitial fibrosis. As profusion and size of the calcified concretions increase, the lung fields become diffusely and densely opaque. Measurement of lung function during the asymptomatic stage reveals little or no abnormality, the affected subject remaining well for many years, but eventually ventilatory restriction, impairment of gas transfer, and arterial hypoxaemia are to be expected.

Management

Corticosteroids, calcium chelating agents, bisphosphonates, and bronchoalveolar lavage have not proved to be effective therapies, and in the absence of lung transplantation, treatment is merely supportive. A detailed description of a 37-year-old man presenting in respiratory failure, recorded severe hypoxia and pulmonary hypertension. Considerable intrapulmonary shunting was demonstrated, which was greatly improved by nasal continuous positive airway pressure, but not by conventional supplemental oxygen therapy. This presumably reflects the dominant effect of the disease in restricting alveolar ventilation over impairing gas diffusion. Although lung transplantation experience is necessarily limited in such a rare disorder, it clearly provides the most optimistic outlook for advanced disease. The dilemma is that the natural history is characteristically one of very slow progression, implying that duration of survival may not be greatly improved unless transplantation is left until the terminal phase. At such a time the probability of success is much reduced.

Further reading

Edelman JD *et al*. (1997). Bilateral sequential lung transplantation for pulmonary alveolar microlithiasis. *Chest* **112**, 1140–4.

Freiberg DB *et al*. (1992). Improvement in gas exchange with nasal continuous positive airway pressure in pulmonary alveolar microlithiasis. *American Review of Respiratory Disease* **145**, 1215–6.

Helbich TH *et al*. (1997). Pulmonary alveolar microlithiasis in children: radiographic and high-resolution CT findings. *American Journal of Roentgenology* **168**, 63–5.

Mariotta S *et al*. (1997). Pulmonary alveolar microlithiasis: review of Italian reports. *European Journal of Epidemiology* **13**, 587–90.

Moran CA *et al*. (1997). Pulmonary alveolar microlithiasis. A clinicopathologic and chemical analysis of seven cases. *Archives of Pathology and Laboratory Medicine* **121**, 607–11.

Nouh MS (1989). Is the desert lung syndrome (non-occupational dust pneumoconiosis) a variant of pulmonary alveolar microlithiasis? Report of four cases with review of the literature. *Respiration* **55**, 122–6.

Volle E, Kaufmann HJ (1987). Pulmonary alveolar microlithiasis in pediatric patients. A review of the world literature and two new observations. *Pediatric Radiology* **17**, 439–42.

Weinstein DS (1999). Pulmonary sarcoidosis: calcified micronodular pattern simulating pulmonary alveolar microlithiasis. *Journal of Thoracic Imaging* **14**, 218–20.

17.11.17Toxic gases and fumes

*D. J. Hendrick**

Noxious substances may be delivered airborne to the respiratory tract in molecular (gases and vapours) or particulate form. A vapour is the gaseous form of a substance that is liquid (occasionally solid) at ambient temperature and pressure. The effects of noxious gases and vapours are determined mainly by their solubility in water; those with high solubility are largely dissolved in the secretions lining the upper respiratory tract, those with low solubilities penetrate to the gas-exchanging tissues and exert their dominant effects there. However, with overwhelming exposures adverse effects will occur at all levels of the respiratory tract, and dose becomes a more important determinant of outcome than solubility.

Particulates that are dispersed in air (aerosols) may be solid (dusts) or liquid (mists), and may carry toxic chemicals through surface adsorption or solution, even if the carrier agent itself is harmless. If the particles are large (diameter > 10 μm), they become trapped in the nose, throat, or major airways. If they are small (diameter < 5 μm), they are deemed 'respirable' and may readily penetrate deeply to become retained in the gas-exchanging tissue and (through macrophage transport) the lung interstitium. Fume is a dispersion of fine (readily respirable) particles that form as vaporized metal condenses at ambient temperature and oxidizes to produce the metal oxide.

Many adverse effects may follow the inhalation of irritant or toxic gases and aerosols. Most are manifested in the lung itself, but some are manifested in other organs after the lung provides a route for absorption (e.g. poisoning from carbon monoxide or hydrogen cyanide). Not only do the respiratory effects occur at different levels, but they may appear at different times. It is useful, therefore, to consider acute and chronic (and sometimes subacute) effects separately, and to recognize that some are dominantly airway effects while others are dominantly parenchymal effects. The chronic effects, such as chronic bronchitis, emphysema, pneumoconiosis, pleural thickening, and lung cancer, generally require months or years of exposure, and arise only after a latency of 10 to 20 years or more. Although 'toxic' or 'irritant' in nature, rather than a consequence of allergy or infection, they are usually considered separately from the disorders attributable to 'toxic gases and fumes', and are described in other chapters.

In general, the acute effects of toxic gases and aerosols are the result of industrial or farming accidents, since the potential for toxic exposure is rare outside occupational environments. However, train or tanker crashes have occasionally caused the rupture of chemical containers and the release of toxic gases into non-occupational environments, and the tragic events at Bhopal illustrate the alarming potential for an industrial accident to have profound effects well beyond the work place.

Site of respiratory injury

Acute upper airway toxicity

If gases or vapours of high solubility (e.g. ammonia, hydrogen chloride, or sulphur dioxide) are involved, or aerosols comprising particles of large average diameter, the adverse effects will dominate in the upper respiratory tract and large airways. Laryngeal oedema, severe enough to cause airflow obstruction and require intubation, is the most important effect, but oedema is to be expected also in the conjunctivae, nose, mouth, and throat, together with inflammatory secretions, even bleeding. One breath is usually sufficient to provoke an immediate withdrawal from further exposure, if this is possible, and so protects against further damage.

Acute tracheobronchitis

If withdrawal from exposure is not possible, or less soluble gases are involved at less pungent levels of exposure (e.g. chlorine), there will be greater penetration beyond the larynx and an acute tracheobronchitis results. This too may become life threatening, and may predispose to secondary infection. If there is survival, full recovery is the rule, but a minority of patients are left with asthma that persists for weeks, months, or even indefinitely. Such an outcome has been called the reactive airways dysfunction syndrome.

Acute pneumonitis

Gases of low solubility (e.g. oxides of nitrogen, ozone, or phosgene) penetrate readily to the gas-exchanging tissues. In the absence of immediate toxicity to the upper respiratory tract they may be encountered in an increasing cumulative and hence dangerous dose. The outcome is an acute pneumonitis and pulmonary oedema some hours later, and is exemplified by nitrogen dioxide toxicity. This gas may be encountered in hazardous concentrations in unventilated silos, particularly when they are decapped (silo filler's disease), when welding is carried out in poorly ventilated sites, and with the combustion of nitrogen-containing substances, such as nitrocellulose. One notorious episode involved a fire of stored radiographs. When grain is stored in silos (or silage is preserved under impervious coverings) microbial contamination leads to the release of nitrogen dioxide along with other toxic gases (principally aldehydes) and asphyxiants (carbon dioxide, methane) and the removal of oxygen. Such processes have the beneficial effect of 'pickling' and preserving the vegetable produce, but they create a dangerous environment for the unwary farm worker. Moulding vegetable produce can similarly provoke a toxic pneumonitis through releasing microbial toxins, and it may be difficult to distinguish this (pulmonary mycotoxicosis or organic dust toxic syndrome, see Chapter 17.11.11) from nitrogen dioxide toxicity.

A curious observation with nitrogen dioxide toxicity has been a recurrent episode of pulmonary oedema 1 to 3 weeks after the initial exposure. The explanation is not clear, though possibly represents the well recognized complication of adult respiratory distress syndrome (**ARDS**) that may follow any cause of toxic pneumonitis. The prophylactic use of oral steroid is said to reduce this risk. Once ARDS occurs, additional risks of pneumothorax and secondary infection arise, sometimes with fatal result. Otherwise recovery is usually full, though rarely bronchiolitis obliterans (or bronchiolitis obliterans with organizing pneumonia) complicates the picture (Chapter 17.11.3).

Illustrative causal agents

Fire smoke

Smoke from fires is a complex mixture of gases and particulates released during combustion and pyrolysis. Its nature can vary greatly with the severity of the fire, the availability of oxygen, and the nature of the burning materials. It may contain toxic concentrations of carbon monoxide, hydrogen cyanide, ammonia, sulphur dioxide, chlorine, phosgene, nitrogen dioxide, aldehydes, and other gases, together with particulates derived from the burning material and surface absorbed gases. Thus, its effects may be diverse, and include suffocation or metabolic poisoning as well as direct toxic injury throughout the respiratory tract.

Metal fume fever

Metal fume fever is an acute and self-limiting febrile illness that characteristically occurs after unusually heavy exposures to metal fume, and recurs on re-exposure after a brief absence from work. It closely simulates other occupational fevers, such as Monday fever in cotton workers (see byssinosis, Chapter 17.4.1.5), polymer fever in chemical workers, and the fevers associated with humidifier lung and allergic alveolitis (see Chapter 17.11.11). It can occur on the first day of exposure. It

results from alveolar deposition of very fine particulate metal oxides (fumes) produced in processes such as welding, burning (oxyacetylene cutting), and smelting of metal. It particularly, but not exclusively, involves zinc, copper, and magnesium. Within some 6 h of exposure, there is thirst, a metallic taste in the mouth, cough, tightness in the chest, and chills, with fever, headache, myalgia, and leucocytosis. Resolution follows within 24 h without permanent sequelae. This benign course is dramatically distinguished from that associated with heavy exposure to fume released from heating cadmium. Cadmium is an anticorrosive metal used in electroplating and the production of alloys. Cadmium fumes may be encountered during extraction, soldering, burning, and welding in poorly ventilated conditions, and may lead to an acute toxic reaction in both lungs and kidneys. It is associated with high mortality.

Simple asphyxiants

Some inhaled gases have no toxic effects, but may severely threaten life through asphyxiation by displacing oxygen from inhaled air. Most common are carbon dioxide and methane, which replace oxygen when vegetable produce decomposes through microbial contamination. A less common source is the slow combustion of coal in disused mines or cellars. Oxygen-deficient air in working mines (blackdamp) has been long recognized as a cause of asphyxiation in miners, but careful monitoring and high levels of ventilation provide effective prevention. Occasionally, however, disused mines accumulate blackdamp during periods of high barometric pressure, only to release the asphyxiant gas when the barometric pressure falls. Most escapes harmlessly to the atmosphere from mine shafts but some may be trapped in the ground under impervious layers of rock or clay. This may find an escape route through faults in the strata and so be emitted at high flow rates into surface buildings. Cellars that breach a layer of clay in coal mining areas may provide a particularly dangerous environment for the unsuspecting. Decaying vegetable matter in the soil may also provide the mechanism for carbon dioxide to replace oxygen, and entry of this oxygen-deficient air into wells during periods of low barometric pressure has also led to asphyxiation.

Management

The management of toxic and asphyxiant insults is essentially supportive. Prompt removal from the source of exposure is followed by attention to the airway, and the administration of specific antidotes when, rarely, this is indicated (for example with cyanide poisoning or methaemoglobinaemia). Because of the risk of laryngeal obstruction or pulmonary oedema following toxic insults, a minimum period of 24 h of hospital care is needed for subjects presenting with hoarseness, stridor, wheeze, or hypoxaemia, and those with a history indicative of heavy exposure to a poorly soluble toxic gas. Humidified air, oxygen supplementation, and bronchodilators may be required. Bronchoscopy may be needed to remove excessive secretions and clear the airway. Laryngeal obstruction demands intubation, and tracheostomy may be necessary if there is extensive upper airway inflammation. Severe pulmonary oedema should be managed as for the adult respiratory distress syndrome (see section 16.04.02). The role of corticosteroids in limiting inflammation is unclear; these drugs add to the risk of secondary infection but have been claimed to prevent the development of late pulmonary oedema after nitrogen dioxide exposure.

---

*Professor J. M. Hopkin wrote on this subject in the third edition of the *Oxford Textbook of Medicine*. Much of his text has been retained in this revision and we acknowledge his contribution with grateful thanks.

Further reading

Ainslie G (1993). Inhalational injuries produced by smoke and nitrogen dioxide. *Respiratory Medicine* **87**, 169–74.

Brooks SM, Weiss MA, Bernstein IL (1985). Reactive airways dysfunction syndrome (RADS). Persistent asthma syndrome after high level irritant exposures. *Chest* **88**, 376–84.

Charan NB *et al.* (1979). Pulmonary injuries associated with acute sulfur dioxide inhalation. *American Review of Respiratory Disease* **119**, 555–60.

Hjortso E *et al*. (1988). ARDS after accidental inhalation of zinc chloride smoke. *Intensive Care Medicine* **14**, 17–24.

Horvarth EP, do Pico GA, Barbee RA (1978). Nitrogen dioxide-induced pulmonary disease. *Journal of Occupational Medicine* **20**, 103–10.

Kanluen S, Gottlieb CA (1991). A clinical pathologic study of four adult cases of acute mercury inhalation toxicity. *Archives of Pathology and Laboratory Medicine* **115**, 56–60.

Schwartz DA. (2002). Toxic tracheitis, bronchitis, and bronchiolitis. In: Hendrick DJ *et al*. eds. *Occupational disorders of the lung: their recognition, management, and prevention* pp.93–103. WB Saunders, London.

Schwartz D, Smith D, Lakshminarayan S (1990). The pulmonary sequelae associated with accidental inhalation of chlorine gas. *Chest* **97**, 820–5.

Smith TJ, Petty TL, Ridding JC (1976). Pulmonary effects of exposure to airborne cadmium. *American Review of Respiratory Disease* **114**, 161.

17.11.18 Radiation pneumonitis

*D. J. Hendrick**

Local therapeutic irradiation of malignancies of the breast, oesophagus, mediastinum (including lymphoma), and lung may damage normal pulmonary tissue. Normal lung is also damaged by the total body irradiation used in preparation for bone marrow transplantation, and by any pulmonary shunting of therapeutic radioactive agents administered via the arterial route to other organs (e.g. yttrium-90 microspheres to the liver). Effects are compounded by the use of chemotherapeutic agents such as methotrexate. However, the risks appear to differ between the various chemotherapeutic agents, for instance that for adriamycin exceeds that for cisplatin. Furthermore, segments of irradiated lung may prove to be unduly susceptible to drug toxicity if chemotherapeutic agents are administered subsequently. Knowledge of the radiation port is therefore important in distinguishing localized toxic pneumonitis from infection.

The scale of pulmonary damage is strongly dependent on the volume of lung exposed, and the dose and fractionation of irradiation. Clinical manifestations range from asymptomatic radiographic opacification to fatal respiratory failure. The latter is fortunately vanishingly rare as radiotherapy techniques are better refined. The dose administered by each fraction is possibly the most important determinant, with values exceeding 2.67 Gy reported to carry most risk. If the same total dose is administered by two fractions during the same day, the risk appears to diminish. Pre-existing fibrotic damage and coincident infection additionally augment the risk.

Radiation releases toxic and mutagenic free radicals within tissue. The resultant DNA damage causes mitotic cell death as tissue cells pass through the first two or three cell divisions after irradiation. The principal cells injured in the lung (the capillary endothelium and type 2 alveolar pneumocyte) have turnover times ranging from 2 to 6 weeks under different circumstances, which explains why the maximum pulmonary effect occurs at about 2 months after injury. The capillary endothelium leaks protein and fluid, and the interalveolar septae become thickened, thereby leading to air space filling (consolidation) and interstitial fibrosis.

The pathogenic pathways appear to involve T lymphocytes. If thymectomy is performed before whole-body irradiation for bone marrow transplantation, the risk of radiation pneumonitis is reduced. When unilateral pneumonitis complicates the irradiation of only one lung in the treatment of breast cancer, a marked but equal increase in lymphocyte numbers from bronchoalveolar lavage in both lungs 4 to 6 weeks later is correlated with a decrease in vital capacity. Animal models suggest that the release of nitric oxide from alveolar macrophages and alveolar epithelial cells also plays a role, as does the recruitment of neutrophils.

The pathological changes can be categorized as (i) acute (up to 3 months), when there is vascular damage with thrombosis and packing of alveoli with surfactant (released from type 2 pneumocytes) and protein-rich fluid; (ii) subacute (2 to 6 months), when there is type 2 pneumocyte renewal and proliferation with macrophage and fibroblast infiltration into the alveoli and interstitium; and (iii) chronic (up to 24 months), when there is alveolar and interstitial fibrosis with capillary sclerosis.

Clinical features

Symptoms begin within a few weeks and may persist for weeks or months. They occur in 10 to 30 per cent of patients following radiotherapy for lung cancer. A rise of the plasma concentrations of transforming growth factor-b_1 and soluble intercellular adhesion molecule-1 have been shown to be markers for those at higher than average risk, thereby providing insights to possible mechanisms, and pointers to identifying subjects who are most suited to escalating radiotherapy treatment.

The severity of symptoms is dependent upon the extent of lung damage, and minor degrees may only be detected incidentally from routine chest radiographs. Cough, which can be severe and may produce thick sputum, and breathlessness are the principal symptoms, but may be accompanied by fever of variable degree. On examination there may be tachypnoea, cyanosis in severe disease, and local crepitations. Telangiectases, the result of cutaneous radiation damage, are often observed in the overlying skin.

The most characteristic radiographic feature is an area of hazy consolidation demarcated by a sharp margin (crossing anatomical pulmonary planes) that corresponds to the limits of the irradiation field, though additional effects are usually detectable beyond these boundaries. Radioisotope scanning shows marked perfusion impairment within the affected portion of lung. In extensive disease, the clinical and radiographic features may be typical of adult respiratory distress syndrome (Chapter 16.04.02). Computed tomography provides the best means of early identification, ground-glass attenuation and interalveolar septal thickening being the early characteristic features.

Up to a year or two after the radiation insult, dense local fibrosis may develop, and magnetic resonance imaging may be required to allow differentiation from tumour recurrence. This can also arise in the apparent absence of earlier pneumonitis, and it may be complicated by pleural effusion, pneumothorax, or fungal colonization (e.g. *Aspergillus* sp.). Fractures resulting from irradiation-induced bone necrosis may occur in nearby ribs.

Treatment

In cases where symptoms are slight, no specific treatment is needed. In more severe disease, corticosteroids produce relief during the acute phase in most patients. Any response to corticosteroids occurs within 3 to 4 days, with clinical and radiographic improvement, and treatment should be continued for 3 to 4 weeks before tapering and stopping. Corticosteroids do not, however, influence the extent of subsequent pulmonary fibrosis. Symptomatic relief of cough and hypoxaemia by an opioid antitussive and oxygen supplementation may also be needed. Prevention offers the best means of control, and further development of methods to detect undue susceptibility and early disease may prove to be valuable in developing safer fractioning protocols.

In laboratory animals, interferon-g has reduced neutrophil recruitment and protein leakage in the early phase of radiation pneumonitis, angiotensin-converting enzyme (**ACE**) inhibitors have prevented or limited increases in central venous and pulmonary artery pressure (and diminished exudation and oedema), and nitric oxide synthase inhibitors have reduced disease progression. Such observations may provide direction for additional measures of prevention and control in humans, though one study has already noted no significant benefit in those who by chance had used ACE inhibitor medication. A further experimental approach has involved the prophylactic use of selenium-enriched spiruline, which appeared to reduce the development of fibrosis by impairing the synthesis of hydroxyproline and type III collagen.

---

*Professor J. M. Hopkin wrote on this subject in the third edition of the *Oxford Textbook of Medicine*. Much of his text has been retained in this revision and we acknowledge his contribution with grateful thanks.

Further reading

Anscher MS *et al*. (1998). Plasma transforming growth factor-b_1 as a predictor of radiation pneumonitis. *International Journal of Radiation Oncology, Biology, Physics* **41**, 1029–35.

Ishii Y, Kitamura S (1999). Soluble intercellular adhesion molecule-1 as an early detection marker for radiation pneumonitis. *European Respiration Journal* **13**, 733–8.

Kwa SL *et al*. (1998). Radiation pneumonitis as a function of mean lung dose: an analysis of pooled data of 540 patients. *International Journal of Radiation Oncology, Biology, Physics* **42**, 1–9.

McBride WH, Vegesna V (2000). The role of T-cells in radiation pneumonitis after bone marrow transplantation. *International Journal of Radiation Biology* **76**, 517–21.

Roach M, III *et al*. (1995). Radiation pneumonitis following combined modality therapy for lung cancer: analysis of prognostic factors. *Journal of Clinical Oncology* **13**, 2606–12.

Roberts CM *et al*. (1993). Radiation pneumonitis: a possible lymphocyte-mediated hypersensitivity reaction. *Annals of Internal Medicine* **118**, 696–700.

Rosiello RA *et al*. (1993). Radiation pneumonitis. Bronchoalveolar lavage assessment and modulation by a recombinant cytokine. *American Review of Respiratory Disease* **148**, 1671–6.

Salinas FV, Winterbauer RH (1995). Radiation pneumonitis: a mimic of infectious pneumonitis. *Seminars in Respiratory Infections* **10**, 143–53.

Sigmund G, Slanina J, Hinkelbein W (1993). Diagnosis of radiation-pneumonitis. *Recent Results in Cancer Research* **130**, 123–31.

17.11.19 Drug-induced lung disease

*D. J. Hendrick and G. P. Spicketi**

Adverse effects of drugs on the lungs frequently present diagnostic problems. This chapter is centred on direct effects of drugs in usual therapeutic doses on the airways, alveoli and interstitium, pulmonary vasculature, pleura, and mediastinal structures. Respiratory disorders arising through occupational exposure (manufacture, transport, dispensing, administration) are considered only briefly. Excluded are indirect effects, such as the predisposition to opportunistic lung infection resulting from cytotoxic agents, the worsening of respiratory failure after sedatives, and the consequences of overdosage or inadequate control of dosage (e.g. pulmonary haemorrhage with anticoagulants).

Asthma

The underlying pathophysiological basis of asthma is an unusually high level of airway responsiveness—the tendency of the airways to constrict following a variety of stimuli, whether specific (e.g. allergens) or non-specific (e.g cold dry air). Airway responsiveness is distributed unimodally in the population at large, individuals with asthma being those in the tail with the highest levels. In theory, drugs (like other environmental agents) may produce asthmatic symptoms either by elevating the pre-existing level of airway responsiveness into the asthma range, or by acting as a specific or non-specific stimulus when airway responsiveness lies already within the asthmatic range. By the first mechanism, the drug acts as a cause of asthma (an asthma inducer), by the second it acts as a cause of asthmatic reactions (an asthma trigger). Some drugs doubtless act through both mechanisms. The mechanism is of limited consequence in the drug setting (though not in the occupational setting), since treatment cessation and future avoidance is the way forward in both circumstances. If, nevertheless, a given drug is known to be a potential trigger but not an inducer, concern over its use need arise only for individuals who are already asthmatic.

In practice, airway obstruction provoked by drugs usually presents as an exacerbation of pre-existing asthma, and the drug acts as a trigger not inducer. In some cases asthma has not previously been recognized until it is exacerbated by the adverse effect of a drug and thus 'uncovered'. In such instances clues to pre-existing asthma are usually elicited when the appropriate history is taken. Drugs that exacerbate symptoms in subjects with pre-existing asthma may conveniently be classified as those that produce a more or less predictable effect, related to their pharmacological properties, and those which produce bronchoconstriction due to an idiosyncratic effect (Table 1). Less commonly, asthma develops *de novo*, probably because immunological hypersensitivity has developed.

Asthmatic symptoms can also be a consequence of the particular formulation of a drug or its method of delivery. For example, nebulized solutions of low osmolality can trigger asthmatic reactions if there is a high level of airway responsiveness. This appears to have been the main mechanism of bronchoconstriction induced paradoxically by nebulized ipratropium bromide. Since the drug was reformulated in isotonic solution the problem has largely disappeared. A further cause of bronchconstriction from nebulized drugs has been the presence of certain preservatives or stabilizers (e.g. benzalkonium chloride, edetate disodium) in the excipient solution. If the administered drug is used for asthma, the effect is particularly unexpected. A further paradox is associated with the evolving use of hydrofluoroalkane propellants (rather than ozone-depleting chlorofluorocarbons) in pressurized aerosol inhalers used to treat asthma. These may have a minor non-specific irritant effect on hyperresponsive airways, which is masked if they are used as the vehicle to administer short-acting b-agonist bronchodilators. However, with the long-acting b-agonist salmeterol (though not eformoterol), the bronchodilator effect is less speedy and hydrofluoroalkane may trigger brief symptoms. The epidemiological importance of this adverse effect is at present unclear.

Pharmacological effects

Cholinergic drugs, such as carbachol, given systemically occasionally produce bronchoconstriction, and in very sensitive asthmatic patients exacerbations have occurred after use of pilocarpine as eye drops. An inhaled anticholinergic agent would seem a logical approach to this problem and has been shown to be effective in reversing occasional untoward effects of cholinesterase inhibitors in asthmatic patients with myasthenia gravis.

The bronchoconstrictor prostaglandin F_{2a}, if used to induce abortion, may be hazardous in asthmatic patients. The occurrence of bronchoconstriction after thiopentone, opiates, and muscle relaxants (tubocurarine, suxamethonium, and pancuronium) is probably due to their capacity to release histamine.

A more common problem is worsening of airway obstruction by b-adrenergic antagonist agents. Although these have been increasingly refined to select agents with the least b_2-antagonism, thus minimizing effects on the airways, none is completely specific for b_1-receptors. The degree of selectivity varies, with propranolol the least and practolol probably the most selective agents used so far. Unfortunately, practolol causes its own distinctive side-effects (see below) and is no longer available. Of the b-blockers currently available, atenolol and metoprolol seem to have the least adverse effects on airway function, but many patients with asthma will show a reduction in forced expiratory volume in 1 s (FEV_1) or peak flow on therapeutic doses of these agents and considerable caution is necessary. The problem of b-blockers in patients with clear-cut asthma is relatively straightforward, but the situation with chronic airway obstruction is less clear. Adverse reactions in such patients are less common and usually less severe, which possibly reflects the coincidental presence of mild asthma rather than a true adverse effect on chronic obstructive pulmonary disease attributable to emphysema or obstructive bronchiolitis. Many patients who develop symptoms and worsening airway obstruction after use of b-blockers are subsequently thought to have had 'latent' asthma.

Although the adverse effects of oral or systemic b-blockers are well recognized, those of ophthalmic preparations are easily overlooked. Timolol, which is used commonly in eye drops for the treatment of glaucoma, is a potent non-selective b-blocker. Its use has frequently been associated with worsening asthma. The ophthalmic formulation of the newer b-blocker betaxolol appears to be less dangerous, but should be avoided in patients with asthma unless no suitable alternative is available.

Effects from sensitization and idiosyncrasy

The mechanism by which drugs lead to asthmatic symptoms when there is no obvious pharmacological effect is often unclear, though immunological sensitization and idiosyncrasy are likely to provide the major pathways.

The most dramatic presentation of drug-related asthma is as part of an acute anaphylactic reaction, and penicillin and intravenously administered iron–dextran are particularly noteworthy among the causal agents. Other drug hypersensitivity reactions that include asthma among the manifestations are often associated with blood eosinophilia and/or eosinophilic pneumonia, and are discussed more fully in Chapter 17.11.9.

Immunological hypersensitivity is presumed to underlie most causes of occupational asthma, some of which involve pharmaceutical agents. Most prominent are certain antibiotics (e.g. cephalosporins, isoniazid, penicillins, piperazine, spiramycin, tetracycline,), the H_2-receptor antagonist cimetidine, the laxative psyllium (ispaghula), pancreatic enzymes, and certain hormones (adrenocorticotrophic hormone, gonadotrophin, pituitary snuff). If a sensitized worker subsequently uses the relevant drug therapeutically, the potential arises for an asthmatic reaction (Fig. 1). The history, when symptoms suggest asthma, should always include details of occupation and medication, and if the patient has ever worked in the pharmaceutical industry the possibility of occupationally induced hypersensitivity to a current medication should be considered.

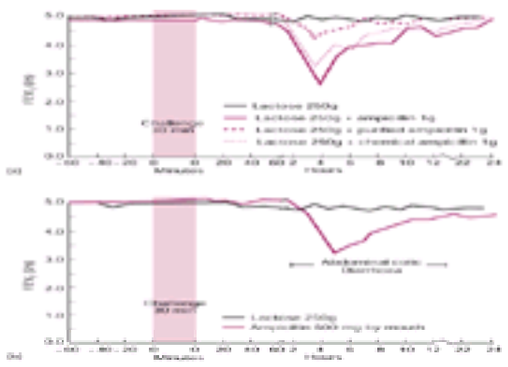

Fig. 1 Results of inhalation and ingestion challenge tests with ampicillin. The inhalation test confirmed that the patient had become sensitized to ampicillin as a consequence of respiratory exposure at work, and the ingestion test showed that asthmatic reactions would be provoked also by oral ingestion at therapeutic dose levels. (Data taken from Davies *et al*.(1974).)

Idiosyncrasy probably underlies many asthmatic symptoms related to medication, and is the likely explanation for exacerbations following use of intravenous *N*-acetylcysteine in severe paracetamol poisoning. Its use in asthmatic patients requires caution. Idiosyncrasy more obviously underlies asthmatic reactions to aspirin and other non-steroidal anti-inflammatory drugs (**NSAIDs**). Exacerbation of asthma after ingestion of aspirin was described as long ago as 1910, but its precise mechanism remains elusive. Most patients who are sensitive to aspirin also react to other NSAIDs, their widely differing chemical structures making an immunological hypersensitivity reaction unlikely. As with cholinergic drugs and b-blockers, asthmatic reactions to NSAIDs may rarely follow ocular administration, and so eye drops deserve careful attention when asthma worsens unexpectedly.

NSAIDs are inhibitors of prostaglandin synthesis via the cyclo-oxygenase pathway, and it is presumed that their adverse effects are mediated in this way. It is possible that metabolism of arachidonic acid is diverted to the production of bronchoconstrictor leukotrienes, but why only a proportion of patients with asthma should be affected is not clear (hence the idiosyncrasy). Deaths have been reported with both aspirin and indomethacin. Of the commonly used analgesic agents, paracetamol is the least likely to provoke a significant response, although occasional adverse reactions are well documented. A further interesting feature is that aspirin-sensitive individuals can be made tolerant to further aspirin by ingesting graded doses over a couple of days. This state of tolerance can then be maintained by daily treatment with aspirin, but sensitivity returns within a few days of discontinuing regular treatment. Any attempt at inducing tolerance in this way requires very careful supervision.

Many patients with analgesic-induced asthma are also sensitive to the azo dye tartrazine (and often to alcoholic beverages). Tartrazine has hitherto been a commonly used colouring agent in medications (particularly those coloured orange or red) and foodstuffs, and since it is an approved food and drug additive, its presence is not always declared and the extent of the problems it may cause is not clear. In the past tartrazine was present, ironically, in some medications used to treat asthma, but most pharmaceutical companies no longer use it in their formulations. Other dyes may, however, have similar adverse effects and some of these still occur in drug formulations. Patients with aspirin and tartrazine sensitivity may also develop troublesome nasal polyposis, as well as asthma. Such patients may benefit from a low salicylate and azo-dye free diet, in addition to strict avoidance of NSAIDs.

The potential exacerbation of asthma by drugs used to treat it presents an acute dilemma, as a drug effect may be difficult to dissociate from spontaneous deterioration. There are well documented reports of worsening asthma after both intravenous aminophylline and hydrocortisone. Sensitivity to hydrocortisone is a particular problem in asthmatic patients who also show adverse reactions to aspirin and NSAIDs. The sensitivity to hydrocortisone of these individuals does not extend to other steroids; it appears to be related to the succinate moiety of the hydrocortisone sodium succinate molecule as it is not seen with the alternative phosphate salt.

The frequent use of nebulized pentamidine for treatment or prophylaxis of pneumocystis infection in patients with HIV infection has been associated with bronchoconstriction in many individuals. The mechanism is unclear. Although patients with asthma show larger responses, others with no previous evidence of asthma may also be affected. The adverse effect is inhibited by prior use of a nebulized bronchodilator, an approach that has become standard in many centres.

Cough

Cough in the absence of asthma is a well-recognized side-effect of treatment with inhibitors of angiotensin-converting enzyme. It develops in 10 to 20 per cent of individuals so treated and is an effect of the class of drug rather than of specific agents. The cough is non-productive. There appears to be a weak relation to dose such that dose reduction may result in some improvement, but in many individuals the symptom is sufficiently troublesome to necessitate drug withdrawal. Deterioration of pre-existing asthma has also been reported occasionally, but in most individuals with cough related to angiotensin-converting enzyme inhibition, features of asthma are not present. The mechanism is unclear; angiotensin-converting enzyme catalyses not only the conversion of angiotensin I to angiotension II, but also the breakdown of bradykinin and substance P. Since these agents are cough stimulants, their accumulation offers a possible mechanism for this unusual adverse effect. The cough disappears on withdrawal of the drug.

Alveolar reactions

There is no generally accepted classification of alveolar reactions to drugs. They range from acute non-cardiogenic pulmonary oedema (e.g. from cremaphor, the agent used to provide soluble cyclosporin A for intravenous use) or the adult respiratory distress syndrome at one extreme to insidiously developing pulmonary fibrosis at the other. The reactions are conveniently considered under three main headings: alveolar capillary leakage, alveolar and interstitial inflammation/fibrosis, and eosinophilia (Table 2). Some overlap is inevitable: inflammatory reactions (whether toxic or allergic) may cause capillary leakage and hence radiographic air space filling; allergic reactions may or may not be characterized by inflammation and eosinophil infiltration.

Of the drugs that may produce the picture of adult respiratory distress syndrome, hydrochlorothiazide and salicylates are the commonest. The reaction to hydrochlorothiazide is idiosyncratic and is not shared by other thiazide drugs. In the case of salicylates there is a clearer relation to dose, with reactions usually occurring with frank overdose (as also occurs with opiates) or, occasionally, with chronic high-level ingestion. Infused b_2-adrenergic agonists are sometimes used as uterine relaxants (tocolytics) to inhibit premature labour. Several, in particular isoxsuprine, ritodrine, and terbutaline, have been associated with florid pulmonary oedema. This reaction is occasionally life threatening and caution is required over the rate of infusion.

Several drugs produce widespread alveolar damage ('pneumonitis' or 'alveolitis'), which may or may not be followed by fibrosis (Table 2). Patients can present acutely with cough, fever, shortness of breath, and occasionally systemic upset. Alternatively, there is slowly progressive fibrosis with gradually worsening dyspnoea and widespread shadowing on the chest radiograph. The mechanism(s) of such reactions are uncertain, but may include toxicity, hypersensitivity, and possibly idiosyncrasy. In some cases, including bleomycin, carmustine, amiodarone, and nitrofurantoin, there is evidence of a relation to dose or duration of treatment. Recent evidence in the cases of nitrofurantoin and bleomycin suggests a role for the production of toxic oxygen radicals in the lungs, perhaps providing a link with the known pulmonary toxicity of oxygen itself and with the synergistic adverse effects of high oxygen concentrations and some cytotoxic agents.

Much recent interest has centred on the cardiac antiarrhythmic drug, amiodarone. It has been estimated that approximately 6 per cent of patients taking 400 mg or more per day for 2 months or more will develop overt pulmonary toxicity. There have also been several well-documented cases involving smaller doses. The mechanism may include both immunologically mediated and direct toxic effects. Histologically the lung shows features of chronic inflammation together with interstitial and intra-alveolar fibrosis (Fig. 2). Characteristic 'foamy' macrophages are seen, but they are not specific for serious toxic reactions as they are also demonstrable in the majority of patients taking the drug without adverse clinical effects. Occasionally, the histological picture is of bronchiolitis obliterans organizing pneumonia (**BOOP**), which is also known as cryptogenic organizing pneumonia. Symptoms include progressive dyspnoea, a troublesome cough, and (occasionally) pleuritic chest pain. Radiographic appearances are varied: most frequently there is a diffuse nodular or alveolar filling pattern, sometimes with upper lobe predominance (Fig. 3); occasionally a pleural effusion is present.

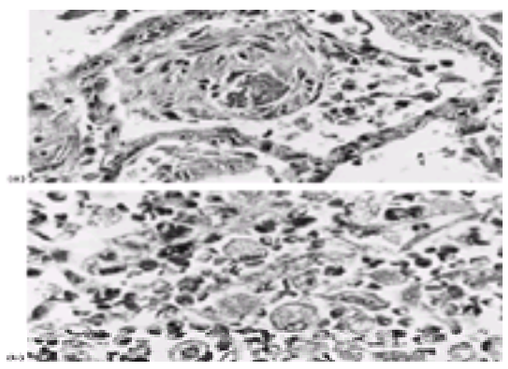

Fig. 2 Histological specimen of the lung of a patient who died from amiodarone pulmonary toxicity: (a) alveolar wall thickening and organizing intra-alveolar exudates; (b) higher power view of alveolar exudate showing characteristic 'foamy' macrophages. (Reproduced from Adams *et al.* (1986). *Quarterly Journal of Medicine* **59**, 449–71, with permission.)

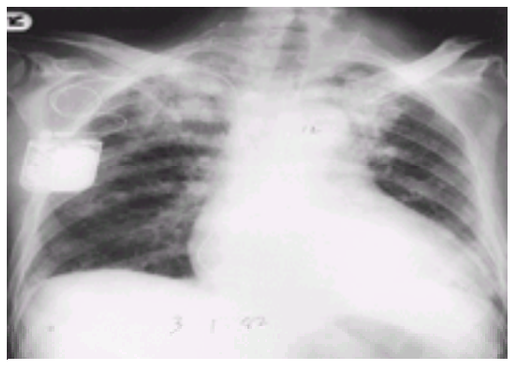

Fig. 3 Chest radiograph of a second patient with amiodarone pulmonary toxicity showing confluent alveolar shadowing in both upper lobes. (Reproduced from Adams *et al.* (1986). *Quarterly Journal of Medicine* **59**, 449–71, with permission.)

Differential diagnoses in the population of patients likely to be taking this drug include left ventricular failure and pneumonia. Further investigation, including measurement of pulmonary wedge pressure and lung biopsy, is often necessary. Bronchoalveolar lavage in some (but not all) patients shows a lymphocytic pattern. This investigation is also of value for the exclusion of infection, but the finding of 'foamy' macrophages in lavage fluid is, for the reasons discussed above, insufficient to confirm the diagnosis. If amiodarone lung toxicity is suspected, cessation of treatment is desirable, but the very long half-life of drug metabolites (many weeks) means that elimination will be very slow. Corticosteroids probably suppress the reaction and sometimes allow treatment to be continued or recommenced in cases of 'malignant' dysrhythmias unresponsive to other agents.

BOOP is a rare manifestation of drug-induced lung disease, but is increasingly recognized in complex settings where drug therapy may have played a dominant or contributory role. In addition to amiodarone, associations have been reported with carbamazepine, nitrofurantoin, phenytoin, sotalol, tacrolimus, ticlopidine, and a number of herbal medications. One of particular interest is a presumed (but unproven) anorectic agent derived from the leaf of an Asian shrub of the *Euphoriaceae* family, *Sauropus androgynus*. In a remarkable period of a few months, more than 60 people who had ingested juice containing uncooked leaf extract of *S. androgynus* presented to hospital in Taiwan with progressive undue breathlessness. In 23 of these patients whose breathlessness was severe, responses to corticosteroid therapy were poor. Plain radiographs were essentially normal, but CT scanning and biopsies demonstrated BOOP with (in a few cases) bronchiectasis in the segmental and subsegmental bronchi.

Cytotoxic and immunosuppressive drugs pose an increasing problem for the lung parenchyma, with the majority reported to cause pulmonary complications. Bleomycin causes problems most frequently, followed by busulphan and methotrexate. Cyclophosphamide and azathioprine are the most widely used drugs in this group, because of their roles in non-malignant disease, but produce adverse pulmonary reactions only occasionally. In most cases it is not clear whether the effects are due to direct toxicity or to hypersensitivity. With bleomycin, however, there is evidence of a dose–response relationship: cumulative doses of less than 150 mg are less likely to cause serious reactions, whereas death due to respiratory failure consequent upon severe fibrosis has occurred in about 10 per cent of patients receiving more than 500 mg.

The recorded frequency of adverse reactions varies with the means by which they are detected; for example, on clinical and functional criteria, fibrosis occurs in 5 to 10 per cent of patients treated with busulphan, but pathological and cytological evidence suggest lung toxicity in much higher proportions. Similarly, the increasing use of CT scanning shows an appreciably higher prevalence than found in surveys that employ plain chest radiography. The frequency of overt lung involvement may also be related to length of survival, as determined by the primary disease. With busulphan, the average interval between starting treatment and the appearance of toxic effects can be as long as 4 years, and in some cases the lung changes appear to progress after the drug has been discontinued. With carmustine (BCNU) it has been shown that pulmonary fibrosis may first be recognized several years after treatment has finished. Other factors that may increase the toxicity of a given drug include advanced patient age and synergism with other drugs, lung radiation, or the subsequent inhalation of high concentrations of oxygen.

Histologically, most cytotoxic drugs produce evidence of diffuse alveolar damage with destruction of lining cells, formation of hyaline membranes, and variable degrees of inflammatory infiltration and fibrosis. Fibrosis is particularly common with busulphan and bleomycin, but rare with methotrexate. With methotrexate and procarbazine (and very occasionally with bleomycin) there may be blood and tissue eosinophilia and correspondingly a good therapeutic response to steroids.

Eosinophilic reactions in the lung include conditions that would be classified as Löffler's syndrome, simple or prolonged pulmonary eosinophilia, and eosinophilic pneumonia (see Chapter 17.11.9). Tissue eosinophilia is a more consistent feature than peripheral blood eosinophilia. Historically, sulphonamides have been the drugs most frequently reported as causes of pulmonary eosinophilia; such reactions have even occurred to a vaginal cream containing sulphonamide. Sulphonamide sensitivity may also explain some of the reactions to sulphasalazine and to chlorpropamide, which is chemically related. The pulmonary eosinophilia recorded with aspirin appears to be distinct from aspirin-induced asthma. Nitrofurantoin may produce an acute eosinophilic reaction in addition to more insidious fibrosis.

The roles of gold salts and penicillamine in eosinophilic reactions have been a matter of some debate, but the evidence suggests that both are involved. It seems unlikely, however, that drugs are responsible for many of the cases of fibrosing alveolitis associated with rheumatoid arthritis. Penicillamine has been incriminated in two other types of adverse pulmonary reaction: first, Goodpasture's syndrome with pulmonary haemorrhage when used in high doses in treatment of Wilson's disease, and second, obliterative bronchiolitis, an unusual form of airway obstruction which is seen occasionally in patients with rheumatoid arthritis. The evidence against penicillamine in the latter is not conclusive.

The clinical severity of eosinophilic reactions is very variable, ranging from a transient and asymptomatic radiographic opacity to a severe illness with dyspnoea, cough, fever, and hypoxaemia due to widespread eosinophilic pneumonia. Concomitant asthma is not uncommon. The chest radiograph shows fluffy opacities, frequently with peripheral or predominantly upper-lobe distribution. The prognosis is usually good: the changes often subside spontaneously on withdrawal of the drug, and in more severely ill patients there is usually a dramatic improvement on instituting treatment with corticosteroids. Although repeated exposure to the offending agents continues to produce reactions, the severity of these may progressively decrease.

Pulmonary vascular reactions

Pulmonary thromboembolism related to use of the contraceptive pill is well established; its frequency correlates with the oestrogen content and has been reduced

since the introduction of low oestrogen preparations. (See Chapter 13.19 and Chapter 13.20 for further discussion.)

The statistical association between pulmonary hypertension and the use of the anorectic agent aminorex in Switzerland, Germany, and Austria in the 1960s was of great theoretical interest. When the drug was withdrawn, the epidemic of pulmonary hypertension subsided and no similar rise was seen in countries that did not introduce this agent. Occasional cases of pulmonary hypertension have been reported also in patients taking various amphetamine-like drugs, but the evidence is not conclusive. (See Section 15.15.2 for further discussion.)

Analgesics given during labour have been implicated in the development of pulmonary hypertension in the newborn; drugs such as aspirin, indomethacin, and naproxen delay premature labour but, by their inhibitory effects on prostaglandin synthesis, may also cause constriction of the ductus arteriosus leading to pulmonary hypertension *in utero* that persists into the postpartum period and causes respiratory distress.

Pleura and mediastinum

Hilar and mediastinal adenopathy are occasionally seen as part of the generalized lymphadenopathy produced by the anticonvulsant phenytoin, and mediastinal lipomatosis has been reported in patients receiving large doses of corticosteroids.

Drugs that have been associated with pleural reactions (effusion or thickening) are shown in Table 3. Several have been reported to produce a syndrome like systemic lupus: the anti-arrhythmic procainamide is most often implicated, but other agents include gold, hydralazine, isoniazid, penicillamine, and sulphonamides. The main respiratory target of this syndrome is the pleura, but (as with pleural disease induced by methysergide and bromocriptine) there is often some fibrosis of underlying lung.

The now obsolete selective b-sympathetic antagonist, practolol, produced a characteristic 'oculomucocutaneous' syndrome. This syndrome differed from systemic lupus erythematosus in that autoantibodies to histones were not usually present, and ocular symptoms are not usually a feature of drug-induced systemic lupus erythematosus. Pleural effusions and subsequent pleural thickening occurred in association with characteristic corneal ulceration, discoid rash, and fibrinous peritonitis. Affected patients sometimes developed effusions months or years after discontinuing the drug. In some the chronic changes led to significant respiratory disability. Minor degrees of pulmonary involvement were reported in some patients, but the predominant abnormality was related to the pleural surface. Other b-sympathetic antagonists, in particular acebutolol, have been reported occasionally to cause an alveolar or pleural reaction, but it seems unlikely that other b-blockers are associated with the full-blown and severe 'oculomucocutaneous' syndrome.

Methysergide, which is used in treatment of the carcinoid syndrome and occasionally for migraine, may induce mediastinal or pleural fibrosis with or without retroperitoneal fibrosis. Improvement follows early withdrawal of the drug. Bromocriptine has some structural similarities to methysergide and can also produce chronic pleural effusions and thickening. The pleural fluid characteristically contains a high proportion of lymphocytes. The frequency of this reaction is uncertain, but it may be relatively common. Methotrexate has been associated with pleurisy, independent of its alveolar effects. The smooth-muscle relaxant, dantrolene, which is used for relief of spasticity, has been reported to produce an unusual type of pleurisy with effusion in which fluid and blood eosinophilia are prominent. There is no evidence of any parenchymal abnormality, and although the changes gradually resolve on withdrawing the drug, some residual pleural fibrosis may remain.

Complications of radiographic and other procedures

Lipoid pneumonia may follow bronchography with oily media. There is an oleogranulomatous reaction that can progress to fibrosis and may sometimes produce a localized mass simulating a neoplasm. Similar reactions can follow aspiration of oily medicines (e.g. laxatives) into the lungs. (See Chapter 17.11.15 for further discussion.)

Lymphangiographic media that drain through the thoracic duct, and so into the venous circulation, enter and can impact in the pulmonary circulation. This is often symptomless, but may cause dyspnoea and cough with the expectoration of fat globules or haemoptysis. Occasional deaths have been recorded. The chest radiograph characteristically shows a fine stippling.

Pleural effusion and, less commonly, mediastinitis occur following endoscopic sclerotherapy of oesophageal varices. The symptoms usually subside within a few days.

---

*Professor G. J. Gibson wrote on this subject in the third edition of the *Oxford Textbook of Medicine*. Much of his text has been retained in this revision and we acknowledge his contribution with grateful thanks.

Further reading

Beasley R *et al*. (1998). Preservatives in nebulizer solutions: risks without benefit. *Pharmacotherapy* **18**, 130–9.

Camus PH, Gibson GJ (1995). Adverse pulmonary effects of drugs and radiation. In: Brewis RAL *et al*., eds. *Respiratory medicine*, 2nd edn, pp 630–57. WB Saunders, London.

Cooper JAD (1990). Drug-induced pulmonary disease. *Clinics in Chest Medicine* **11**, 1–194.

Davies RJ, Hendrick DJ, Pepys J (1974). Asthma due to inhaled chemical agents: ampicillin, benzyl penicillin, 6-amino-penicillanic acid and related substances. *Clinical Allergy* **4**, 227–47.

Lai R-S *et al*. (1996). Outbreak of bronchiolitis obliterans associated with consumption of *Sauropus androgynus*. *Lancet* **348**, 83–5.

Rosenow EC *et al*. (1992). Drug-induced pulmonary disease. An update. *Chest* **102**, 239–50.

Ryrfeldt A (2000). Drug-induced inflammatory responses to the lung. *Toxicology Letters* **112–13**, 171–6.

17.12 Pleural disease

M. K. Benson

Introduction

The pleural surfaces form the interface between the lung parenchyma and chest wall. The parietal pleura is applied to the chest wall and the surfaces of the ribs, with a thin layer of connective tissue separating it from the periosteum. At the hilum, the pleura form a sleeve-like structure encompassing the major vessels and bronchi. The visceral pleura covers the surface of the lungs and extends into the major fissures which separate the lobes of the lung. The pleura is a membranous structure, the surface of which is covered with a single layer of mesothelial cells. These cells have microvilli over their surface that facilitate the absorption of pleural fluid.

The pleura is not essential for adequate functioning of the lungs, although the smooth surfaces do permit movement of the lungs within the thorax with minimal energy loss. Obliteration of the pleural space following surgery or as a result of inflammatory disease does not result in significant respiratory impairment. Between the two layers of pleura there is a potential space, the surfaces of which are lubricated by a thin layer of fluid. The pressures within the pleural cavity are generated by the difference between the elastic forces of the lungs and the chest wall. At functional residual capacity the outward recoil of the chest wall is equal to the inward recoil of the lung parenchyma (see Chapter 17.1.2).

A number of pathological processes can affect the pleura. Inflammation results in characteristic pleuritic pain, aggravated by deep inspiration, coughing or sneezing, and often accompanied by a pleural rub. The accumulation of fluid in the pleural space results in a pleural effusion. Air can also enter the pleural space resulting in a pneumothorax. Primary tumours of the pleura are relatively uncommon. Involvement by metastatic malignant disease or by lymphoma is much more frequent.

Pleural fluid formation

Normal lubrication of the pleura is provided by a thin layer of fluid, an ultrafiltrate of plasma, although surfactant may also be present and plays a role. Although the turnover of pleural fluid is probably in the order of 1 litre per day, the volume of fluid present at any one time is only 20 to 30 ml. Under normal circumstances two factors operate to prevent the accumulation of fluid in the pleural space: the pleura itself acts as a semipermeable membrane, and the flux of fluid across the pleural space is accounted for by the forces involved in Starling's law of transcapillary exchange. The hydrostatic gradient from the capillaries of the parietal pleura favours fluid efflux into the pleural space. Pressure in the capillaries in the visceral pleura is close to that of the pulmonary capillaries, and this lower pressure favours reabsorption of fluid. The lymphatic system provides a second important method of preventing excess fluid accumulation. In addition, it enables proteins to be recovered from the pleural space and return to the circulating plasma. Factors likely to result in excess fluid accumulation in the pleural space can be identified, including the following:

1. imbalance between the hydrostatic and oncotic forces as defined in Starling's equation, such fluid is usually a transudate;
2. alteration in the permeability of pleural capillaries;
3. impaired lymphatic drainage;
4. abnormal sites of entry (e.g. transdiaphragmatic passage of fluid in patients with ascites).

Pleural effusion

A pleural effusion is an abnormal accumulation of fluid in the pleural space. It is traditional to divide effusions into transudates and exudates, although blood, pus, or chyle can also form collections in the pleural space. The main causes are listed in Table .

Clinical features

The clinical history and examination play an important part in diagnosing the presence of pleural fluid and may yield significant clues as to the pathogenesis. Symptoms related to pleural disease are pain and breathlessness. The extent to which these occur is likely to vary and clinical presentation will, at least in part, be determined by the underlying pathogenesis. Pleuritic pain is relatively easy to recognize and causes severe discomfort on deep inspiration or coughing and is most often associated with an inflamed pleural surface or the presence of a pneumothorax. There can, however, be significant collections of pleural fluid without pain. The other major symptom is breathlessness, which only becomes apparent if there is a large effusion or in patients who already have impaired respiratory reserve.

There may be no abnormal physical signs if the effusion is relatively small, but these are often diagnostic if the effusion is large. Chest wall movement may be normal, although it will tend to be limited, particularly if there is pain, and there can also be a lag of chest wall motion on the affected side. The percussion note is very dull and breath sounds are diminished or absent, as are vocal resonance and tactile vocal fremitus. Compression of the lung above the effusion can result in signs of consolidation with bronchial breathing and increased vocal resonance. The position of the mediastinum, as judged by the trachea and apex beat, will help in distinguishing between a large effusion and a collapsed lung. In the former, the mediastinum is central or displaced away from the side of the effusion, whereas in the latter deviation is towards the affected side.

Given the diversity of pathologies that may result in pleural fluid, systemic symptoms and signs often yield important diagnostic information.

Investigation of pleural effusion

The presence of a pleural effusion should be suspected on clinical examination and can be confirmed by using radiographic imaging or ultrasound. Whilst clinical features play an important part in identifying the pathogenesis, examination of the pleural fluid or pleural biopsy material is most likely to lead to a definitive diagnosis.

Radiographic techniques

Radiographic techniques are helpful in identifying the presence of an effusion but are of limited value in determining the pathogenesis. A conventional posteroanterior chest radiograph is usually adequate to confirm the presence of a clinically significant effusion. Fluid tends to accumulate in dependent parts of the thorax and small effusions in the order of 200 ml will result in blunting of the costophrenic angle. Larger effusions produced increased opacification and mediastinal shift (Fig. 1).

Variations from the normal appearance will result if the fluid is loculated, a situation more likely to occur with an empyema or if there are pleural adhesions (Fig. 2).

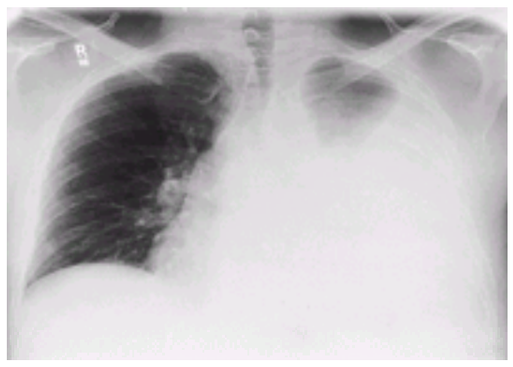

Fig. 1 Chest radiograph showing opacification of the left hemithorax and mediastinal shift indicating a large pleural effusion.

Fig. 2 CT scan demonstrating a loculated effusion due to an empyema.

Ultrasound can be helpful in confirming the presence and site of an effusion. Pleural fluid is identified as an echo-free space between chest wall and lung. The presence of echoes within the fluid may indicate an empyema or haemothorax, and ultrasound can also demonstrate the presence of septation and loculi (Fig. 3).

Fig. 3 Chest ultrasound showing pleural effusion with septation.

CT scan can complement ultrasound examination in demonstrating the site of a pleural collection and has the additional advantage of imaging the underlying lung. The appearance of the pleural surface can be useful in helping to differentiate between benign and malignant pleural disease. In addition, CT-guided percutaneous biopsy techniques can increase the diagnostic yield if taken from areas of gross pleural thickening, visualized on the CT scan.

Examination of pleural fluid

Thoracentesis, whereby pleural fluid is aspirated percutaneously, is a relatively simple procedure that can be undertaken for diagnostic purposes and, in the case of larger effusions, can relieve breathlessness. It is usually performed with the patient upright in a comfortable position with the arms and head supported on a pillow. Unless the fluid is loculated, a conventional site for aspiration is posteriorly about 10 cm lateral to the spine and one intercostal space below the upper level of the fluid as detected by percussion. A common error is to attempt aspiration as low as possible, but this often yields a dry tap since it is impossible on clinical grounds to determine the level of the diaphragm. The procedure is performed with strict aseptic technique. The skin and underlying tissues are infiltrated with local anaesthetic, taking care to avoid the intercostal nerves and vessels that run immediately beneath the rib. For diagnostic purposes it is usually adequate to remove 50 to 100 ml of fluid. If therapeutic aspiration of large amounts of fluid is being undertaken, it is best to introduce a small plastic cannula into the pleural space to minimize the risk of damage to the underlying lung.

Failure to obtain fluid can arise for a number of reasons, including misdiagnosis of the presence of fluid, incorrect site of aspiration, and the presence of viscid fluid. Ultrasound examination can help to identify the reason for a failed tap and guide further attempts if fluid is present. Biochemical, cytological, and microbiological examination of pleural fluid can help to establish a diagnosis if this is not apparent on clinical grounds (Table 2).

Macroscopic appearance

The appearance of the pleural fluid and its odour may provide diagnostic information. Transudates are clear, straw-coloured fluids that do not clot on standing. Many exudates have a similar appearance but they can be turbid due to the presence of cells. Blood-tinged fluid is of little diagnostic significance, but a uniformly bloody effusion is likely to be associated with malignancy. Pus can be very viscid and difficult to aspirate. It is turbid in appearance, yellow in colour, and often foul smelling. Chyle is odourless and milky in appearance.

Biochemistry

Exudates will generally have a higher protein content than transudates, but although a level of 30 g/l has traditionally been used to differentiate between the two, there is significant overlap and values should be interpreted with caution. Better differentiation may be obtained by comparing concentrations of protein and lactic dehydrogenase (LDH) in the pleural fluid with those in blood. The criteria that can prove helpful in identifying an exudate are as follows:

1. fluid to serum ratio of total protein above 0.5;
2. fluid to serum ratio of LDH above 0.6;
3. fluid LDH concentration above 200 international units.

The concentration of glucose in pleural fluid is normally equal to that in serum, but in effusions associated with rheumatoid arthritis the glucose concentration in the pleural fluid is rarely above 1.6 mmol/l. Reduced concentrations are also found in association with tuberculosis, empyema, malignancy, and lupus. Measurement of pleural fluid amylase may be diagnostically useful if the pleural effusion is associated with acute pancreatitis, a pancreatic pseudocyst, or oesophageal rupture. Pleural fluid with a pH less than 7.3 in the presence of a normal blood pH occurs in a number of conditions including empyema and parapneumonic effusions, malignancy, tuberculosis, and collagen vascular diseases. Thus as a diagnostic investigation its use is limited. It may have some prognostic significance in patients with malignant disease in that a low pH is associated with more extensive disease. When associated with infection, a pH of less than 7.2 is one of the criteria indicating the need for tube drainage.

Microscopic and cytological examination

Most transudates have cell counts of less than 1000/mm^3, with the cells being a mixture of lymphocytes, polymorphs, and mesothelial cells. Exudates tend to have higher white counts, although this in itself is of little diagnostic value. A polymorphonuclear leucocytosis is indicative of a bacterial infection but can also be seen in association with a pulmonary infarct or pancreatitis. A predominance of lymphocytes raises the possibility of tuberculosis but can also occur in association with malignancies, including lymphoma, and also is seen after coronary artery by-pass surgery. The presence of excess eosinophils is not in itself diagnostic but tends to be associated with benign inflammation.

Cytological examination of pleural fluid for suspected malignancy is a rapid and efficient diagnostic procedure. Fifty ml of fluid should be sent for immediate examination. The finding of malignant cells is likely to be diagnostic, although actively dividing mesothelial cells can mimic an adenocarcinoma. A positive diagnosis is made in approximately 60 per cent of malignant effusions. The diagnosis of a malignant mesothelioma presents particular difficulties but the use of monoclonal antibodies CEA, B72.3, and Leu-M1 can help distinguish an adenocarcinoma from mesothelioma.

Microbiology

Gram stain and culture of the pleural fluid are of diagnostic value if an infective aetiology is suspected. Identification of an organism confirms the diagnosis and sensitivity testing will assist in the appropriate choice of antibiotics. Tuberculous pleurisy is difficult to diagnose and acid fast smears are positive in only about 10 per cent of cases. Cultures are more likely to be positive if a reasonable volume of fluid is concentrated and then examined.

Pleural biopsy

Pleural biopsy may be indicated if initial analysis of pleural fluid fails to establish a diagnosis. It is most likely to give diagnostic information if there is an underlying malignancy or tuberculosis. The diagnostic yield is likely to be greatest when used in conjunction with CT scanning to identify areas of particular thickening or nodularity. Blind percutaneous biopsies are usually performed using an Abraham's or Cope's needle. Both of these are large blunt-tipped needles with a hook to catch a sample of parietal pleura. The technique is similar to that used for pleural aspiration except that a small incision is made in the skin and subcutaneous tissue to enable ease of insertion of the needle. The Abraham's needle consists of an outer trocar with a side hole and an inner cannula with a cutting edge. Once in the pleural space, confirmed by aspiration of fluid, the side hole is opened by rotating and slightly withdrawing the inner cannula. The needle is then withdrawn at an angle to the chest wall such that the side hole gently catches on to the parietal pleura, and at this point the inner cannula is advanced to obtain a biopsy. Several samples can be taken using this technique, but care is needed to avoid damage to the intercostal nerves and vessels. Samples for histological examination should be placed in 10 per cent formaldehyde and those for culture for mycobacteria should be put into saline.

Thoracoscopy

Direct visualization of the pleura is possible using a thoracoscope, and this should be considered as a diagnostic procedure when less invasive procedures have failed to yield a definitive diagnosis. It is particularly useful in suspected malignancy or tuberculosis. For further details regarding this technique, see Chapter 17.3.4.

The relative diagnostic yields of thoracocentesis, pleural biopsy, and thoracoscopy with respect to tuberculosis and malignancy are given in Table 3. Pleural aspiration is easy to perform, requires limited expertise, and has only minor risks. Pleural biopsy requires more expertise and complications include pain, pneumothorax, haemothorax, and vasovagal reactions. Thoracoscopy is only available in specialist centres and, in addition to the risks associated with pleural biopsy, may result in subcutaneous emphysema. The diagnostic yield and choice of technique will depend on the medical indications and local expertise.

Chest drain insertion

Chest drainage is used therapeutically in a number of clinical situations, including the management of malignant pleural effusions, empyemas, pneumothoraces, postoperatively, and when there has been bleeding into the pleural cavity. Doctors working in a variety of specialties should therefore be capable of inserting an intercostal drain. It is potentially a painful procedure and the patient needs reassurance, comfortable positioning, and adequate analgesia and local anaesthesia. Ideally, narcotic medication should be given prior to the procedure unless medically contraindicated. The position of the patient will depend on the site of insertion. For the axilla, the patient is usually lying at 45° with the arm behind their head such that the axillary area is exposed.

The site of insertion will depend on the clinical and radiological findings. In patients with a large pleural effusion or a pneumothorax, the most usual site is in the axilla, in a triangle bounded by the anterior axillary line, the lateral margin of the pectoralis major, and a horizontal line at the level of the nipple. An alternative site for an apical pneumothorax is in the second intercostal space in the mid clavicular line. Where there are loculated collections of air or fluid, real-time ultrasonography can ensure optimum placement of the catheter.

The appropriate drain size remains a subject of debate. Large bore tubes (28–30 French) are used postoperatively and in the management of haemothorax. Whilst some authorities advocate their use in other clinical situations, such as management of pneumothoraces and empyemas, small bore catheters (10–14 French) are as effective, easier to insert, and better tolerated by patients.

Insertion of a chest drain with aseptic technique is important to minimize the risk of infection. The site for tube insertion is infiltrated with local anaesthetic (1 per cent lignocaine—maximum dose 3 mg/kg) to achieve satisfactory anaesthesia of skin, subcutaneous tissues, and pleura, care being taken to avoid intercostal vessels that run beneath each rib. The infiltrating needle should also be used to aspirate air or fluid from the pleural space once the pleura has been punctured, and a chest tube should not be inserted until this has been achieved. An incision is made in the skin parallel to the ribs and slightly larger than the catheter to be inserted. Blunt dissection of subcutaneous tissue and muscle is achieved by opening and closing a curved clamp: this separates the muscle fibres and can also be used to penetrate the pleura. If a large tube is to be inserted, the track can be explored with a gloved finger to ensure that there are no organs that might be penetrated. However for smaller catheters, an inappropriately large track can result in leakage around the catheter. The central trocar can be used to support the catheter, but forcible insertion should be avoided because of potential for damage to lung and mediastinal structures. Alternative techniques include the use of pigtail catheters after insertion of a guide-wire and dilatation of the track.

Although traditionally the catheter is inserted towards the apex for pneumothoraces and towards the base for pleural effusions, precise positioning is relatively unimportant. There should, however, be sufficient catheter within the thorax to minimize the risk of displacement. A strong non-absorbable suture, such as 1–0 silk, should be used to secure the drain, which should also be supported with a small square of gauze and a bioclusive dressing over the insertion site. Tape can be wrapped around the drain at the point of ligature to help prevent the suture from slipping on the tube. Once the tube has been inserted and anchored, it is attached to an underwater-seal bottle, allowing one-way flow of air or fluid. This has the disadvantage of restricting mobility, but enables observations to be made about tube patency—judged by inspiratory pressure swings—and allows the continued drainage of air or fluid to be monitored. Good nursing care with regular 4-hourly observations are necessary to ensure that the tube has not become kinked or blocked and that it has not become displaced. Uni-jet directional flutter valves permit greater mobility and can be used in ambulant patients for the management of pneumothoraces, but they are inappropriate for use when there are fluid collections.

Specific pleural fluid collections

Transudates

A transudate is characterized by low concentration of protein and other large molecules. Excess fluid forms when there is an increase in capillary hydrostatic pressure or reduction in colloid osmotic pressure. The former occurs predominantly in congestive cardiac failure, and the latter when there is hypoalbuminaemia associated with nephrotic syndrome or hepatic disease.

Cardiac failure

Small effusions are common in congestive cardiac failure or constrictive pericarditis. Right-sided failure results in increased pressures in the systemic capillaries and thus an increased efflux of fluid from the parietal pleura. Elevated left heart pressures will be reflected in the pulmonary circulation with a consequent diminution in fluid reabsorption from the visceral pleura. The clinical features of cardiac failure are usually sufficient to make a diagnosis. Thus cardiomegaly, elevated jugular venous pressure, and third or fourth heart sounds may all be present. The effusions are frequently bilateral. Unilateral effusions are more common on the right side and may cause diagnostic uncertainty. Pleural aspiration can help to confirm the diagnosis by demonstrating the presence of a transudate. Resolution with treatment of heart failure offers further confirmation of the diagnosis.

Hepatic cirrhosis

Hypoalbuminaemia, which may occur in patients with chronic liver disease, is a major contributory factor to the development of generalized oedema. Ascites and pleural effusions are both common, with effusions more often seen on the right than on the left. In some patients, ascitic fluid seems to pass directly into the pleural space, either through a defect in the diaphragm or via lymphatics.

Exudates

Neoplastic pleural effusions

Malignant involvement of the pleura is the commonest cause of a large pleural effusion. Lung cancer may spread directly or via lymphatics and is the commonest metastatic cancer (40 per cent). Breast cancer also spreads via the lymphatic system and is the commonest cause of a malignant effusion in women. Metastatic spread from gastrointestinal or genitourinary tumours is less common. Lymphomas can occur at any age and account for approximately 10 per cent of malignant effusions. Extensive investigation for an asymptomatic primary is of limited value, although it may be appropriate to exclude disease originating in breast or ovary because of the potential response to hormonal treatment or chemotherapy.

Symptoms directly attributable to the effusion are most commonly breathlessness or chest discomfort. The degree of breathlessness depends on the size of the effusion and the presence of pre-existing lung disease. Specific symptoms attributable to the primary tumour are often absent. Non-specific symptoms include malaise, anorexia, weight loss, and sweats.

Appropriate investigations have already been outlined above. They include imaging with a posteroanterior chest radiograph and chest CT scan. Aspirated fluid is usually an exudate and is blood stained in approximately 50 per cent of cases. Malignant cells can be identified in approximately 60 per cent of cases. CT-guided pleural biopsy may be justified where aspiration has proved diagnostically unhelpful. If the diagnosis remains in doubt, the options are either to await events, since the diagnosis may become obvious with the passage of time, or to obtain further biopsy material at thoracoscopy.

Once cancer has metastasized to the pleura, treatment is essentially palliative, although chemotherapy or hormonal treatment may be appropriate depending on the primary, the cell type, and the functional status of the patients. Removal of the pleural fluid and measures to prevent reoccurrence are only necessary if the size of the effusion results in significant breathlessness. If the patient is comfortable, no action may be necessary. Percutaneous needle aspiration of 1 to 2 l of fluid is a simple outpatient procedure and often results in considerable symptomatic benefit. The fluid is likely to recur, but repeated aspiration may be an appropriate therapeutic option. Intercostal tube drainage can be used to remove the bulk of the fluid, but this is also likely to be of temporary benefit unless combined with pleurodesis. If the effusion is large, it should be drained gradually over 24 h to reduce the risk of re-expansion pulmonary oedema. A number of sclerosing agents have been used with varying degrees of success. Unfortunately there is a marked paucity of comparative data. Sclerosing agents include antibiotics (tetracycline, doxycycline, minocycline), antineoplastic agents (bleomycin), and non-specific irritants (sterile talc, *Corynebacterium parvum*, mepacrine). Until recently, tetracycline (1 g/50 ml saline) has been the most widely used sclerosant, but production of the intravenous preparation has been discontinued. Of the remaining sclerosants, talc slurry (2–5 g in 100 ml) achieves success rates of about 90 per cent and is likely to be the most effective. The pleural space is drained to dryness and after adequate local and systemic analgesia, the sclerosant is injected and the tube clamped for 2 to 4 h. Any residual fluid is then drained and the tube removed after 24 h. An alternative approach is to insufflate iodized talc into the pleural space at thoracoscopy. Surgical pleurectomy or pleural abrasion is also very effective at preventing recurrence, although rarely regarded as an appropriate option because of its invasive nature and unacceptably high morbidity and mortality.

Meigs' syndrome

This rare syndrome describes an association between pleural effusions, ascites, and a benign ovarian tumour. Surgical removal of the tumour results in disappearance of the pleural and peritoneal fluid. The mechanism of pleural effusion is uncertain, but it is generally assumed that ascitic fluid reaches the pleura through diaphragmatic channels or lymphatics. There is no evidence of spread of the tumour, and the syndrome should not be confused with effusions that can result from metastatic spread of ovarian cancer.

Endometriosis of the pleura

This rare condition is one in which endometrial tissue is implanted on visceral or parietal pleura. Catamenial pleural chest pain or a pneumothorax can be presenting features. More commonly, there is an associated effusion that on aspiration reveals blood or chocolate brown fluid. Thoracotomy will reveal multiple cystic structures, but surgical ablation is unsuccessful because of the nature of the disease. Treatment is directed at suppressing ovulation using progesterones or androgens.

Infection

Inflammation of the pleura associated with infection is usually due to pneumonia or lung abscess. Other possible sources of pleural infection include subdiaphragmatic pathology, mediastinitis, oesophageal perforation, or direct contamination following penetrating trauma or surgery. Inflammation can result in pleurisy without significant fluid production, a non-infective exudate (a parapneumonic effusion), or infected fluid (an empyema). Distinction between a parapneumonic effusion and an empyema is somewhat arbitrary since there can be a transition from one to the other. A parapneumonic effusion may be slightly turbid, contains an excess of polymorphs, but has no organisms: by contrast, an empyema contains increased numbers of polymorphs and is frankly turbid. pH measurement of the pleural fluid can assist in management: a pH below 7.2 is generally regarded as an indication for pleural drainage. Organisms are likely to be present when there is frank pus, although isolation and identification may be difficult if antibiotics have already been administered. The spectrum of organisms most frequently encountered in the United Kingdom is listed in Table 4.

The presentation will vary depending on the pathogenesis. In a patient with pneumonia, an empyema should be suspected if there is persisting fever and elevated white count despite appropriate use of antibiotics. Pleuritic chest pain may be present but is not invariable. Classical signs of effusion can be difficult to detect, particularly if the pleural collection is loculated. Anaerobic organisms are commonly encountered; these may be secondary to aspiration from the oropharynx or upper respiratory tract.

A chest radiograph that shows apparent loculation of pleural fluid should alert the clinician to the possibility of an empyema. Gas may be present, revealed by a fluid level. Ultrasound examination or CT scan can be useful for identifying the most appropriate approach for attempted aspiration, and also demonstrate the presence of loculi.

Treatment is discussed elsewhere.

Tuberculous pleurisy

Pleural involvement with tuberculosis is a common manifestation of primary infection with direct extension from a subpleural focus. Gross parenchymal disease is rare and the primary site often cannot be identified clinically or radiologically. It is more common in younger patients and in those of Asian origin.

The presenting features are usually acute or subacute with fever, pleuritic pain, and breathlessness, but some patients may give a longer history of malaise, sweats, and weight loss. The effusion is often large (in excess of 2 l) and tends to recur after initial aspiration. The fluid is a serous exudates, often with an excess of lymphocytes whose presence should alert the clinician to the possibility of tuberculosis. Bacilli are rarely identified on pleural aspirate. Culture is more likely to be positive, but even so the diagnostic yield is low, and pleural biopsy is more likely to give a diagnostic result, showing granulomatous inflammation in approximately two-thirds of patients. Thoracoscopic biopsy is most likely to give a definitive diagnosis, but this is not universally available and often it may be appropriate to commence treatment on clinical grounds alone.

A tuberculous empyema with pus in the pleural space is rare, but occasionally complicates cavitating parenchymal disease. A bronchopleural fistula can result, and in advanced disease the empyema can present with a draining sinus through the chest wall. Other bacterial pathogens may be present in the pleural fluid.

Treatment involves the use of standard antituberculous chemotherapy together with adequate drainage if there is frank pus. Steroids have been shown to promote rapid resolution of the effusion and improve symptoms in the short term, but may have no long-term benefits. Rarely, surgical closure of a bronchopleural fistula or decortication is required.

Subdiaphragmatic infection

Inflammation or infection below the diaphragm should always been considered if there is an unexplained effusion with features suggesting infection. A subphrenic abscess is frequently associated with an effusion, usually on the right side. This may follow abdominal surgery, but can also be caused by perforated peptic ulcer, appendicitis, diverticular disease, or cholecystitis. Hepatic abscesses can also cause a right-sided effusion. Even without infection, upper abdominal surgery can result in pleural effusion, but such effusions are usually small and transient.

If there is evidence of sepsis, the source needs to be identified, and if pus is present it must be drained. Ultrasound examination or CT scanning are both effective ways of diagnosing a subphrenic abscess and guiding percutaneous aspiration. The pleural fluid is usually an exudate and although turbid with a polymorphonuclear leucocytosis, it rarely becomes infected.

Pancreatitis is associated with a pleural effusion in approximately 20 per cent of patients. In the majority, the effusion is on the left side and results from inflammation caused by enzyme-rich pancreatic fluid. Whilst the classical symptoms of abdominal pain, nausea, and vomiting usually predominate, pleurisy and breathlessness can occasionally be presenting features. The pleural fluid is often blood stained and contains abnormally high levels of amylase.

Pulmonary emboli

Pleurisy, often associated with a pleural effusion, is a common presenting feature of pulmonary emboli, particularly if there is associated pulmonary infarction. The effusion is usually small and in itself does not require specific treatment. The fluid is often blood stained but the cellular content is variable and there are no specific diagnostic features. The diagnosis of pulmonary emboli is based on clinical features supplemented by appropriate radiographic or isotopic imaging techniques.

Rheumatoid arthritis

Pleural effusions are the commonest pulmonary manifestation of rheumatoid arthritis. They occur in approximately 3 per cent of patients with active rheumatoid disease and are more common in men than in women. The development of an effusion can antedate the onset of joint symptoms in a small proportion of patients. There is no relationship to the severity of the arthritis, but effusions are more likely to occur in patients with subcutaneous nodules and those who have high titres of rheumatoid factor.

The effusions are usually small but can enlarge to a size that results in breathlessness. Although usually unilateral they may be bilateral in about 20 per cent of patients. The fluid is an exudate and may appear turbid due to cholesterol crystals. The cellular content is not diagnostic, but polymorphonuclear cells usually predominate. Although not specific to rheumatoid effusions, a diagnostic clue is the presence of a low glucose concentration (usually below 1.5 mmol/l). The pH is also low and the lactic dehydrogenase concentration elevated. Whilst these findings may also be present in infective and malignant effusions, the associated clinical features rarely lead to diagnostic uncertainty. Pleural biopsy is non-specific although it can reveal the epithelioid cells and multinucleate giant cells found in rheumatoid nodules.

Symptomatic treatment with anti-inflammatory analgesics is indicated if pleuritic pain is a feature. Systemic steroids can speed resolution of the pleural fluid although they are rarely necessary. The majority of effusions resolve spontaneously within a few months but there may be some residual fibrosis.

Systemic lupus erythematosus

Pleural involvement is common in patients with this condition. Approximately 50 per cent of patients will have pleurisy at some stage and the majority of these will have an associated effusion, usually small. Aspiration of the fluid is rarely necessary for either diagnostic or therapeutic purposes. The fluid is an exudate and has high concentrations of antinuclear antibodies. Lupus erythematosus cells can also be identified. There is a good therapeutic response to oral corticosteroids.

Haemothorax

A haemothorax is the result of bleeding into the pleural space and is arbitrarily diagnosed on the basis of having a haematocrit more than half that of peripheral blood. This distinguishes it from a blood stained effusion, which can be associated with a number of different pathological processes. The vast majority of haemothoraces are associated with penetrating or non-penetrating trauma, including iatrogenic procedures such as central venous catheterization. Bleeding usually results from parenchymal laceration or damage to intercostal vessels. A pneumothorax is present in a high proportion of patients.

The treatment of choice is to insert a large intercostal drain (28–32 French), allowing evacuation of blood and reducing the incidence of subsequent fibrothorax. If this reveals continued bleeding, thoracotomy may be required. Surgery is not indicated simply to remove any residual blood clots since in a majority of patients there is spontaneous lysis with no residual damage.

Spontaneous bleeding into the pleural space can occur in association with a pneumothorax (a haemopneumothorax) and presumably results from the tearing of pleural adhesions. Other rare causes of a haemothorax include bleeding disorders, or excess anticoagulants and rupture of the thoracic aorta.

Chylothorax

A chylothorax results from leakage of chylous fluid from the thoracic duct. Absorbed fat is transported as chylomicrons in the intestinal lymphatics and, together with lymph originating in the lower limbs and abdomen, reaches the blood stream via the thoracic duct. The flow of lymph in the thoracic duct is approximately 100 ml/h under basal conditions but can increase five-fold after a fatty meal.

Congenital absence of the thoracic duct is a rare cause of a chylothorax, but the majority of cases are acquired either as a result of trauma or neoplastic invasion. Surgery is the commonest cause of traumatic damage, particularly those operations that involve mobilization of the aortic arch or oesophageal resection. Penetrating trauma occasionally results in damage to the thoracic duct but rupture can also occur from non-penetrating injuries. The commonest single cause of rupture of the thoracic duct is damage caused by neoplastic infiltration, including lymphomas and carcinomas. Other rare associations include pulmonary lymphangioliomyomatosis, the yellow nail syndrome, and filiariasis.

There are no specific clinical features and the diagnosis of a chylothorax is usually made after pleural aspiration. The fat is typically milky and opalescent due to the presence of fat globules, and needs to be distinguished from an empyema or from pseudochyle. In empyema, any discoloration is due to a cellular deposit and after

centrifugation the supernatant is clear. Pseudochyle is due to high lipid levels, usually cholesterol crystals, which occur in chronic effusions, particularly following tuberculosis. Cholesterol crystals can be recognized on smears of the sediment, and the addition of ethyl alcohol to the fluid results in clearing if high concentrations of cholesterol are responsible for the opalescence.

Spontaneous resolution can occur if the chyle is removed by thoracocentesis and the subsequent flow of chyle reduced by the use of medium chain triglyceride diets or parenteral nutrition. If there is known malignancy, mediastinal irradiation may also assist resolution. Malnutrition and lymphopenia are likely to occur if large volumes of chyle continue to be drained, and under such circumstances surgical ligation of the thoracic duct above the diaphragm can be combined with pleurodesis.

Pneumothorax

A pneumothorax results from gas entering the potential space between visceral and parietal pleura. A spontaneous pneumothorax is a consequence of rupture of a bulla or cyst on the surface of the lung, allowing air to escape from the alveoli into the pleural space. Following penetrating trauma, atmospheric air may enter the pleural space through the wound or the visceral pleura may be punctured allowing entry of alveolar gas. An iatrogenic pneumothorax can occur as a result of damage inflicted during catheterization of a subclavian vein or following percutaneous or transbronchial lung biopsy.

Pathophysiology

At functional residual capacity the inward elastic recoil of the lung and the outward recoil of the chest wall results in a negative pressure in the potential space between visceral and parietal pleura. Pressures with respect to atmosphere become more negative during inspiration and only become positive during forced expiration. Because of the elastic recoil of the lung, pleural pressure is always less than alveolar pressure. Thus, if there is a breach of the visceral pleura due to rupture of a surface bulla, gas moves from lung to pleural space. As the lung collapses down, the pressures equilibrate and net flow of gas ceases.

The functional effect of a pneumothorax is to reduce the vital capacity and the total lung capacity as the lung collapses. Ventilation of the affected side is reduced, although perfusion may also fall such that the anticipated alveolar–arterial oxygen gradient and consequent hypoxia are less than might be anticipated. Ventilatory failure with a rise in arterial PCO_2 is rare, except in patients with pre-existing lung disease.

Once the original leak has sealed, reabsorption of pleural gas occurs and re-expansion of the lung takes place at approximately 1.25 per cent of the volume of the hemithorax per day. Pleural gas is absorbed because the total gas pressure, which is similar to that of arterial gas, is greater than that of venous blood.

Tension pneumothorax

Occasionally, and with devastating consequences, the site of air leak acts as a valve, allowing air to enter the pleural space during inspiration but preventing return flow during expiration. If this happens, then pleural pressure rises and a tension pneumothorax results, with compromise of the circulation, mediastinal shift, and impaired function of the opposite lung.

Clinical syndromes

A spontaneous pneumothorax usually occurs without any warning or obvious precipitating factor. A primary pneumothorax occurs in individuals with apparently normal lungs. A secondary pneumothorax is a consequence of pre-existing lung disease.

Primary pneumothorax

A primary pneumothorax is a relatively common condition with an annual incidence of about 9/100 000. It is particularly common in young men, with a male to female ratio of approximately 4 to 1. It is commoner in smokers. Patients are often tall with a marfanoid appearance. The cause of the pneumothorax is assumed to be rupture of a surface bulla or cyst, often near the lung apex. Only rarely can these be visualized radiologically. Approximately 50 per cent of patients suffer from a recurrence within 4 years. Whilst this is usually on the same side as the initial event, there is also an increased chance of contralateral pneumothorax.

Secondary pneumothorax

Older patients presenting with a spontaneous pneumothorax are likely to have underlying lung disease as a predisposing factor, most commonly chronic obstructive pulmonary disease. Rarely, acute exacerbations of asthma may be complicated by a spontaneous pneumothorax, presumably due to high alveolar pressures associated with gas trapping. Some pulmonary infections can result in rupture of necrotic lung with subsequent air leak into the pleura. Staphylococcal pneumonia, anaerobic lung abscesses, and tuberculosis are among the most likely infecting organisms. A secondary pneumothorax can also be caused by lung malignancy. There are also a number of parenchymal and connective tissue disorders in which pneumothorax is a recognized complication; these include cystic fibrosis, lymphangiomyomatosis, pulmonary neurofibromatosis, Langerhan's cell histiocytosis, Marfan's syndrome, and Ehlers–Danlos syndrome.

Iatrogenic pneumothorax

A number of diagnostic and therapeutic procedures can cause pneumothorax. Percutaneous needle aspiration or biopsy of the lung carries the greatest risk, with estimates ranging from 5 to 50 per cent. The risk is related to the presence of underlying lung disease, the size of the needle, and the depth of penetration. Bronchoscopy rarely causes problems, but a transbronchial biopsy carries a small risk, particularly if undertaken in the absence of screening. Intermittent positive pressure ventilation, especially when used with positive end-expiratory pressures, can result in pneumothorax, which can present under tension. Attempted catheterization of subclavian veins can result in puncture of the lung, particularly when carried out by inexperienced personnel.

Clinical features

Symptomatically, a spontaneous pneumothorax will present with chest pain and breathlessness. The pain is usually of sudden onset and typically pleuritic, being localized to the affected side. Inspiration is often painful and breathing is shallow to minimize discomfort. Dyspnoea is partly engendered by the difficulty in taking a deep breath, but is also dependent on the size of the pneumothorax and the presence of underlying lung disease. The initial sensation of breathlessness can improve rapidly before the resolution of the pneumothorax and may be due to reflex changes from receptors in the lung and airways. A significant proportion of patients do not tend to seek medical advice for several days.

Abnormal physical signs may be difficult to detect if the pneumothorax is small or the underlying lung is emphysematous. The most consistent finding is a reduction in breath sounds on the affected side. Movement of the chest wall may be reduced, particularly if there is pain. The percussion note will be resonant, and although hyper-resonance is a recorded feature, it may be difficult to detect any difference from the non-affected side. Vocal fremitus and tactile vocal resonance are diminished. A left sided pneumothorax is occasionally associated with a clicking noise synchronous with the heart beat (Hamman's sign), probably due to contact and separation of the pleural surfaces in time with the heart beat.

Tension pneumothorax

Pneumothorax rarely causes severe respiratory distress and hypoxia unless associated with pre-existing lung disease, or the pneumothorax is under tension. Tension pneumothorax is rare, but a medical emergency. The patient looks as though they are about to die, with severe breathlessness, cyanosis, tachycardia, hypotension, grossly elevated jugular venous pulse, and evidence of mediastinal shift (trachea and apex displaced away from the side of the chest under tension, which may appear unduly prominent).

Associated conditions

Pneumomediastinum

A pneumomediastinum can present in isolation or together with a pneumothorax. Air tracks along the bronchovascular sheath to the hilum and mediastinum. It can be

associated with sudden rises in alveolar pressure during sneezing or straining, and be found in patients undergoing intermittent positive pressure ventilation. Precordial chest discomfort may be a presenting symptom and subcutaneous emphysema can usually be detected in the neck and supraclavicular fossae. No specific treatment is indicated since the condition is benign and self-limiting.

Haemothorax

The presence of blood and air in the pleural space is most commonly the result of trauma. A spontaneous pneumothorax can occasionally have associated bleeding into the pleural space, presumably due to tearing of pleural adhesions.

Pyopneumothorax

A pyopneumothorax usually results from rupture of necrotic lung, but can also be due to oesophageal perforation. The clinical picture is one of a combined empyema and pneumothorax.

Investigations

Confirmation of a pneumothorax is best made by chest radiograph. An erect posteroanterior film is adequate. Although a film during expiration increases the radiodensity of the lung and enhances the contrast between lung and pleural gas, it is rarely necessary. The cardinal radiological features are illustrated in Fig. 4. The outer margin of the lung can be seen as a thin line with the space between it and the chest wall devoid of any lung markings. Pleural adhesions can result in part of the lung being tethered to the chest wall with some distortion of the normal radiographic appearance. A large emphysematous bulla can sometimes be mistaken for a pneumothorax on both clinical and radiological grounds, although the inner margins are usually concave. A CT scan may help to resolve the diagnostic uncertainty.

Fig. 4 Chest radiograph demonstrating a tension pneumothorax.

Lung function tests are inappropriate in the presence of a suspected pneumothorax, although oximetry or arterial blood gases may provide information that will influence decisions regarding treatment.

Management

The diversity of therapeutic options listed in Table 5 is a manifestation of uncertainty as to what is best. Two principal therapeutic objectives are to achieve rapid resolution of the pneumothorax, particularly if there is evidence of respiratory distress, and to reduce the likelihood of recurrence.

Natural resolution

A small pneumothorax with a radiological rim of air of less than 2 cm in an otherwise healthy patient may require no treatment other than reassurance and relief of pain. Non-steroidal anti-inflammatory drugs are usually effective in this respect. Admission to hospital is unnecessary provided the patient has ready access to medical care and is advised to return if symptoms worsen.

Simple aspiration

Simple aspiration is the treatment of choice for a patient with a large primary pneumothorax and is also the initial emergency treatment in patients with a tension pneumothorax. If successful, it will not only speed resolution but also relieve associated breathlessness or chest discomfort. It is simple to perform and has negligible morbidity, even in relatively inexperienced hands. The site for aspiration is usually the second intercostal space in the mid clavicular line. After infiltrating with local anaesthetic, a 16 to 18 gauge intravenous cannula attached to a syringe containing a small amount of sterile water is inserted through the chest wall gently aspirating until bubbling seen in the syringe confirms that the pleural space has been entered. The internal needle is then removed (in tension pneumothorax allowing immediate release of pleural pressure) and a 50-ml syringe with three-way tap is connected to the cannula to allow aspiration of air, which should stop if resistance is encountered or if the patient experiences undue discomfort or coughing. If more than 2 to 3 l of air has been evacuated, it is likely that there is a persisting air leak, and aspiration should be abandoned and a decision made as to whether an intercostal drain should be inserted. This is likely to be necessary if the patient is breathless, but conservative management with repeated aspiration after an interval of a day or two may be more appropriate for those who are relatively asymptomatic. Simple aspiration has been shown to be less painful and require a shorter duration of hospital stay than treatment with an intercostal drain. There are no significant differences in recurrence rate.

Intercostal tube drain

This approach may be indicated if simple aspiration has failed. It is more likely to be needed in patients with underlying lung disease, when even a small pneumothorax can result in severe respiratory failure. The technique for insertion has already been described. Bubbling will cease once the air leak has ceased and the lung fully expanded. A check radiograph should be undertaken before the catheter is removed since drainage of air will also cease if the tube is blocked or has become dislodged. Clamping of the tube prior to removal is unnecessary. The value of additional suction is unproven and may serve to maintain the patency of the original air leak. It can be tried if the lung fails to re-expand, with the aim of evacuating the pleural air and allowing apposition of the pleural surfaces.

Medical pleurodesis

This is undertaken in an attempt to obliterate the pleural space and reduce the likelihood of recurrent pneumothorax. There is a marked paucity of information as to who might benefit and as to the best approach (see previous discussion). It should be considered in those with recurrent pneumothoraces whose underlying condition makes a surgical operation unduly hazardous.

Surgical intervention

Referral rates for surgery vary considerably. The main indications include persisting air leak after prolonged intubation (usually 1–2 weeks) and recurrent pneumothoraces. In the latter group, referral is most commonly made after the second or third ipsilateral recurrence, or if there have been bilateral pneumothoraces. The preferred surgical options include over-sewing or excision of any large bullae, combined with apicolateral pleurectomy or pleural abrasion. Surgical morbidity in otherwise healthy patients is very low. Risks are greater in those with underlying lung disease, but this must be balanced against the life-threatening potential of further pneumothorax.

Video-assisted thoracoscopic surgery (VATS) offers a less invasive approach to the surgical management of pneumothoraces, but experience is limited. The morbidity

does not seem to be significantly different from that of thoracotomy, and there is a significant chance of recurrence (5–10 per cent). Its role may be limited to those who would be regarded as otherwise unfit for an open procedure.

Further reading

Pleural effusions—diagnosis

Ansai T (1998). Management of undiagnosed persistent pleural effusions. *Clinics in Chest Medicine* **19**, 407–17. Useful clinical review with respect to diagnostic approach to the diagnosis of pleural effusions.

Heffner JE (1997). Diagnostic value of tests that discriminate between exudative and transudate pleural effusions. *Chest* **111**, 970–80. Systematic review of tests used to discriminate between transudates and exudates.

Hsu C (1987). Cytological detection of malignancy in pleural effusions: a review of 5,255 samples from 3,811 patients. *Diagnostic Cytopathology* **3**, 8–12. Review of the diagnostic yield from cytological examination of pleural fluid.

Light RW (1995). *Pleural diseases*, 3rd edn. Williams and Wilkins, Baltimore. Authoritative textbook of pleural disease.

Page R (1989). Thoracoscopy: a review of 121 consecutive surgical procedures. *Annals of Thoracic Surgery* **48**, 66–8. Retrospective review of the value of diagnostic thoracoscopy.

Prakash UBS, Reiman HM (1985). Comparison of needle biopsy with cytological analysis for the evaluation of pleural effusion: analysis of 414 cases. *Mayo Clinic Proceedings* **60**, 158–63. Retrospective study comparing results of thoracentesis and needle biopsy in malignant and non-malignant pleural disease.

Sahn S (1988). State of the art. The pleura. *American Review of Respiratory Diseases* **138**, 184–234. Comprehensive review of normal anatomy and physiology, pathology, and clinical features of pleural disease.

Thoracostomy—techniques

American College of Surgeons Committee on Trauma (1993). Thoracic trauma. In: *Advanced trauma life support program for physicians: instructor manual.* Chicago. Practical guidelines regarding indications for and technique of tube thoracostomy.

Hyde J, Sykes T, Graham T (1997). Reducing morbidity from chest drains. *British Medical Journal* **314**, 914–15. Editorial emphasizing importance of not using a trocar to penetrate chest wall.

Kline JS, Schultz S, Heffner JE (1995). Interventional radiology of the chest; image-guided percutaneous drainage of pleural effusions, lung abscess and pneumothorax. *American Journal of Radiology* **164**, 581–8. Descriptive review of image-guided drainage techniques.

Tomlinson MA, Treasure T (1997). Insertion of a chest drain; how to do it. *British Journal of Hospital Medicine* **58**, 248–52. Step by step guide to insertion of chest drain concentrating specifically on large bore tubes.

Malignant effusions

Lynch J Jnr (1993). Management of malignant pleural effusions. *Chest* **103**(Suppl.), 385S–9S. Review of different therapeutic approaches to management of malignant pleural effusions.

Patz ER *et al*. (1998). Sclerotherapy for malignant pleural effusions; a prospective randomised trail of bleomycin vs. doxycycline with small-bore catheter drainage. *Chest* **113**, 1305–11. Prospective study in 106 patients comparing bleomycin and doxycycline sclerotherapy.

Zimmer PW *et al*. (1997). Prospective randomised trial of talc slurry vs. bleomycin in pleurodesis for symptomatic malignant pleural effusions. *Chest* **112**, 430–4. Comparative study in 29 patients with 90 per cent success rate of pleurodesis.

Empyema

Alfageme I *et al*. (1993). Empyema of the thorax in adults. Etiology, microbiologic findings and management. *Chest* **103**, 839–43. Retrospective clinical review of 80 patients with empyema.

Ferguson AD *et al*. Empyema Subcommittee of the Research Committee of the British Thoracic Society (1996). The clinical course and management of thoracic empyema. *Quarterly Journal of Medicine* **89**, 285–9. A multicentre review of clinical features and management of thoracic empyema.

Tuberculosis

Berger HW, Magier E (1973). Tuberculous pleurisy. *Chest* **63**, 88–92. Critical review of clinical presentation and diagnosis of tuberculous pleurisy.

Epstein DM *et al*. (1967). Tuberculous pleural effusions. *Chest* **91**, 107–9. Retrospective review of clinical presentation and treatment.

Miscellaneous

Bynum LJ, Wilson JE (1976). Characteristics of pleural effusions associated with pulmonary embolism. *Archives of Internal Medicine* **136**, 159–62. Features of pleural fluid associated with pulmonary embolism.

Emerson PA (1966). Yellow nails, lymphoedema and pleural effusions. *Thorax* **21**, 247–50. Classical description of yellow nail syndrome.

Eppler GR, McLeod TC, Gaensler EA (1982). Prevalence and incidence of benign asbestos pleural effusion in a working population. *Journal of the American Medical Association* **247**, 617–22. Review of asbestos-related pleural effusions.

Fairfax AJ, McNabb WR, Spiro SG (1986). Chylothorax: A review of 18 cases. *Thorax* **41**, 880–5. A retrospective review of patients presenting with a chylothorax.

Hunninghake GW, Fauci AS (1979). Pulmonary involvement in the collagen vascular diseases. *American Review of Respiratory Disease* **119**, 471–503. State of the art review.

Kay MD (1968). Pleural pulmonary complications of pancreatitis. *Thorax* **23**, 297–306. Description of pleural complications associated with pancreatitis.

Meigs IV (1954). Meigs' syndrome. *American Journal of Obstetrics and Gynaecology* **67**, 962–6. Meigs' original description of the eponymous syndrome.

Pneumothorax

Allmind M, Lange P, Viscum K (1989). Spontaneous pneumothorax: comparison of simple drainage, talc pleurodesis and tetracycline pleurodesis. *Thorax* **44**, 627–30. Randomized study comparing three treatment options with subsequent rates of relapse.

Harvey JE (1993). Comparison of a simple aspiration with intercostal drainage in the management of spontaneous pneumothorax. *Thorax* **48**, 430–1. Randomized study of aspiration versus intercostal tube drainage. No subsequent difference in rate of recurrence.

Massard G, Thomas P, Wihlm J-M (1998). Minimally invasive management for first and recurrent pneumothorax. *Annals of Thoracic Surgery* **66**, 592–9. Review of treatment options in the management of pneumothoraces.

Miller AC, Harvey JE (for Standards of Care Committee, British Thoracic Society) (1993). Guidelines for the management of spontaneous pneumothorax. *British Medical Journal* **307**, 114–6. British consensus-based guidelines for the management of pneumothoraces.

Parry GN, Juniper ME, Dussek JE (1992). Surgical intervention in spontaneous pneumothorax. *Respiratory Medicine* **86**, 1–2. Editorial review of role of thoracoscopic surgery for pneumothorax.

So SY, Yu DYC (1982). Catheter drainage of spontaneous pneumothorax, suction or no suction, early or late removal?. *Thorax* **37**, 46–8. Randomized study on use of suction or no suction drainage in the management of pneumothoraces.

17.13 Disorders of the thoracic cage and diaphragm

J. M. Shneerson

Introduction

Skeletal disorders of the thorax are an important group of conditions that frequently impair ventilation. They are often associated with respiratory muscle weakness due to neuromuscular disorders, which are described elsewhere. Most of these conditions restrict the development and/or the expansion of the lungs so that alveolar ventilation rather than intrapulmonary gas exchange is primarily impaired.

Disorders of the spine

Scoliosis

Scoliosis is defined as a lateral curvature of the spine, but it is invariably also associated with rotation of the vertebral bodies. This results in an unstable lordosis rather than a kyphosis, and hence the frequently used term kyphoscoliosis is inaccurate. A mild degree of scoliosis is very common. Angles of curvature of 5° or 10° have been used to define when it becomes pathological, but these are arbitrary figures. Postural scoliosis can be distinguished from a structural scoliosis by its temporary nature and because it disappears on bending forward.

The age of onset and natural history of scoliosis vary according to its cause (Table 1). When it is due to a neuromuscular disorder ('paralytic' scoliosis) it usually arises during childhood or adolescence, or in poliomyelitis within about 2 years of the acute infection. Typically, the curve has a long C shape and may be severe. The scoliosis is due to asymmetrical weakness of the axial muscles so that the spine rotates and moves to one side. Weakness of chest wall muscles is almost invariable, occurs in a pattern which is characteristic of each disorder, and may have a profound influence on the clinical features.

When the scoliosis is due to a congenital abnormality, such as a hemivertebra or a segmentation defect, it usually becomes apparent early in childhood. The scoliosis of neurofibromatosis and Marfan's syndrome is probably due to an abnormality of connective tissue. Scoliosis due to pleural or pulmonary disease is less common than in the past, now that chronic infections are less frequent and more successfully treated.

The commonest type is adolescent idiopathic scoliosis, where the spinal deformity develops at the time of the pubertal growth spurt. It is around four times as common in girls as in boys, and the convexity of the deformity is on the right in 80 per cent of cases. The scoliosis may continue to worsen slightly even after growth of the spine stops. An infantile form of idiopathic scoliosis is less common, and, although it often resolves spontaneously, it can progress to a severe deformity.

Pathophysiology

The most important organic consequence of scoliosis is the respiratory abnormality. A direct result is that the compliance of the chest wall is reduced. This is more marked in older subjects, possibly owing to degenerative changes in the costovertebral joints. The compliance of the lungs is also reduced, largely because of their small volume. In addition, the distortion of the rib cage puts the inspiratory muscles at a mechanical disadvantage; those on the side of the convexity of the scoliosis are shortened and those on the side of the concavity lengthened. The vital capacity falls when changing from the sitting to the supine position, implying that diaphragmatic function is impaired. A restrictive defect and reduction of the maximum inspiratory and expiratory pressures develops even in the absence of any muscle weakness, but is more marked if this is present.

In adults with severe scoliosis, exercise capacity is linked to the degree of reduction of the vital capacity and the forced expiratory volume in 1 s (FEV_1). The tidal volume increases initially and then remains constant, whilst respiratory rate rises as exercise becomes more intense. Ventilation at any given oxygen uptake is greater than normal, and maximal exercise ventilation, which limits exercise capability, is often severely curtailed. The cardiac output may increase normally during exercise, but pulmonary artery pressure rises rapidly, and its rate of increase is linearly related to oxygen uptake and inversely related to the vital capacity.

In mild scoliosis, the arterial blood gases are often normal, but the first abnormality is a fall in the partial pressure of oxygen (PO_2). This is due to suboptimal ventilation and perfusion matching, particularly at the bases of the lungs. Even when the anatomical distortion of the two lungs is gross, there is usually rather less difference in function between the two lungs than might be expected. Acute ventilatory failure may be precipitated by, for instance, a chest infection or asthma, but chronic hypoventilation initially occurs during sleep. Sleep is associated with loss of the voluntary respiratory drive and a reduction in the reflex drive in response to hypoxia, hypercapnia, and other stimuli. Muscle activity is reduced, and whereas in non-rapid eye movement sleep (NREM) this affects all the respiratory muscles to an equal extent, in rapid eye movement sleep (REM) diaphragmatic activity is selectively retained. Relaxation of the other respiratory muscles is more intense than during NREM and loss of activity in the upper airway dilator muscles increases the upper airway resistance and the work of the chest wall muscles.

These changes during sleep are particularly important in scoliosis, where the diaphragm is attached to an asymmetrical rib cage and where the respiratory pump often has little reserve. The effects of sleep are accentuated when the scoliosis is the result of neuromuscular disorders, because the presence of muscle weakness in addition to the skeletal deformity reduces tidal volume and increases respiratory frequency, leading to alveolar hypoventilation. Arousals from sleep initially occur in REM, which becomes fragmented, and at a later stage in NREM, with loss particularly of stages 3 and 4. Sleep fragmentation itself reduces the respiratory drive and impairs the strength and probably the endurance of the respiratory muscles, promoting a vicious circle in which there are progressively more respiratory-induced arousals and a deterioration in respiratory drive and muscle function. Central apnoeas and hypopnoeas develop; hypercapnia then appears during wakefulness as well as in sleep.

Chronic hypercapnia during the day is uncommon in childhood and is determined by the following:

1. Age of onset: if the scoliosis appears before the age of about 8 years it may prevent normal alveolar multiplication so that the lungs fail to develop fully. The capillary surface area is reduced and there is an increased risk of developing respiratory and right heart failure later in life. The later onset of adolescent idiopathic scoliosis is probably the major reason why these complications only rarely occur in this condition.
2. Level of the scoliosis: in general, the higher the curve in the thoracic spine, the more marked are the cardiac and respiratory problems. Thoracolumbar or lumbar scoliosis has virtually no effect on respiration.
3. Severity of scoliosis: the angle of scoliosis is closely related to the reduction in lung volume. This association is seen with the residual volume, total lung capacity, and functional residual capacity, as well as with vital capacity, except in patients with neuromuscular disorders, where the changes in lung volumes are

due to the weakness of the respiratory muscles as well as the degree of deformity. The changes in lung volumes become significant when the angle of scoliosis is greater than about 100°.

4. Presence of muscle weakness: the functioning of the respiratory muscles is impaired in scoliosis, and any further loss of strength or endurance due to neuromuscular disorders may precipitate respiratory failure. Conversely, respiratory function often worsens in neuromuscular disorders when scoliosis develops as the strength of the axial muscle becomes asymmetrical.
5. Small lung volumes: respiratory failure usually occurs when lung volumes have been reduced to a degree such that vital capacity is less than 1.0 to 1.5 1.

Hypoxia causes pulmonary vasoconstriction which increases the pulmonary vascular resistance and leads to pulmonary hypertension. If this is prolonged, right ventricular and atrial hypertrophy develop. Significant pulmonary hypertension is rarely seen unless the arterial PO_2 is less than around 8 kPa. Pulmonary hypertension by itself rarely causes right heart failure, and the exact mechanisms underlying this are uncertain. The increase in sympathetic activity and circulating catecholamines associated with hypoxia cause renal vasoconstriction and a reduction in renal blood flow. This activates the renin–angiotensin–aldosterone system leading to sodium and water retention. Hypercapnia is associated with an increase in renal tubular hydrogen ion excretion with sodium reabsorption in exchange for hydrogen. This leads to fluid retention, which is accentuated by an increase in antidiuretic hormone secretion. Hypercapnia probably also increases capillary permeability, which contributes to the appearance of oedema.

Polycythaemia occasionally occurs as a result of erythropoietin release from the kidneys in response to hypoxia. This adaptive mechanism increases the oxygen content of the blood, but also increases blood viscosity, which increases the work of the right and left ventricles and predisposes to arterial and venous thrombosis.

Symptoms and physical signs

The earliest symptom of scoliosis is usually a change in the appearance of the patient, such as asymmetry of the shoulders or prominence of the posterior rib hump. Backache is a late and uncommon symptom. With mild curvatures there may be no respiratory symptoms, but mild shortness of breath on exertion is common. A change in this often signifies the development of complications such as respiratory failure. Orthopnoea suggests that diaphragmatic function is impaired. When respiratory failure develops, fatigue, ankle swelling, and even syncope may indicate that pulmonary hypertension and right heart failure are present. Frequent awakenings during sleep, associated with excessive daytime somnolence, indicate sleep fragmentation due to apnoeas and hypopnoeas, and are important symptoms that warn of impending respiratory failure.

Physical examination may reveal the cause of the scoliosis, such as Marfan's syndrome or neurofibromatosis, and other congenital abnormalities. Any associated muscle weakness or congenital heart disease may be apparent. Rib cage expansion may be predominantly lateral or anterior, or achieved by extension of the spine. In some subjects, chest expansion is mainly oblique because of the rotation of the spine, and some areas of the chest wall may move paradoxically. Accessory muscle action is usually prominent. The presence of central cyanosis indicates that the arterial oxygen saturation is below around 80 per cent. Signs of hypercapnia may also be present. These include tachycardia, large volume pulse, peripheral venous dilatation, papilloedema, a flapping tremor, reduction in tendon reflexes, small pupils and, if severe, confusion and coma (CO_2 'narcosis').

Investigations

The severity of scoliosis can be demonstrated radiologically, but chest radiography is often unhelpful in thoracic scoliosis because rotation of the spine obscures much of the lung fields. This can be overcome by obtaining an oblique view of the chest which, by aligning the spine behind the heart, simulates a posteroanterior view. Lung function testing reveals a restrictive defect with reduction in all lung volumes, although the change in residual volume is least marked and, therefore, the ratio of residual volume to total lung capacity is increased. KCO is raised, as in other chest wall disorders that cause a restrictive defect and in which the lung tissue is normal. Maximum inspiratory and expiratory pressures and trans-diaphragmatic pressure are reduced. Chest wall and lung compliance are less than normal, and exercise tolerance is impaired. Arterial blood gas analysis reveals a slightly low PCO_2 in mildly affected subjects, but later in the course of disease a rise in PCO_2 and a proportional fall in PO_2 develop. Sleep studies show a variable degree of hypoxia and hypercapnia which are usually most marked in REM sleep. Electrocardiography and echocardiography may be required to establish whether congenital heart disease is present and, if so, to identify the abnormality.

Prognosis

The prognosis in adolescent idiopathic scoliosis is virtually normal, but life expectancy is reduced in many of the other forms of scoliosis. This is particularly so in scoliosis of early onset, when it is both severe and high in the thorax and associated with respiratory muscle weakness, low vital capacity, and abnormal blood gases.

In most subjects the cause of death is either cardiac or respiratory. Pneumonia and respiratory failure are particularly common in neuromuscular disorders, but hypoxic dysrhythmias during sleep are probably responsible for some deaths. Congenital heart defects, which have an increased prevalence in subjects with scoliosis, particularly when this is due to a congenital abnormality or of the idiopathic type, also contribute to mortality.

Treatment

Mild scoliosis does not need any specific treatment. The prognosis is normal and there is minimal respiratory deficit. However, as the scoliosis becomes more severe, spinal fusion or a costectomy, in which the parts of the ribs comprising the posterior hump are removed, may be of cosmetic value. Spinal fusion may also be required to prevent progression of the scoliosis, to stabilize the spine, particularly in neuromuscular disorders, and in selected cases to try to improve cardiac or respiratory function or to prevent its deterioration.

The value of spinal fusion to prevent cardiorespiratory deterioration in adolescent idiopathic scoliosis is still under debate. A large number of studies of respiratory function before and after surgery have shown remarkably little change in lung volumes, blood gases, or exercise ability. However, in some patients with muscle weakness, particularly Duchenne's muscular dystrophy, the rate of fall of the vital capacity can be slowed considerably, and it can even be improved in patients who have had poliomyelitis. Despite these short-term improvements, there have been no studies which indicate whether or not spinal fusion performed in childhood or adolescence prevents respiratory failure from appearing later in life. If respiratory failure does develop, any acute illness which may have precipitated it should be actively treated. This is most commonly an infection or bronchial asthma. Non-invasive ventilation or endotracheal intubation and ventilation may be required during the acute illness. If the latter is needed, the patient is then weaned from this either completely or onto a non-invasive method of long-term respiratory support.

Chronic ventilatory failure usually responds to long-term mechanical respiratory support. Administration of oxygen at night and/or during the day may be dangerous because of the risk of hypercapnia. Nasal or face mask positive pressure ventilation is the treatment of choice, but a negative pressure system, such as a cuirass or jacket, is an alternative. Non-invasive ventilation is usually only required during sleep, but some patients benefit from 1 or 2 h treatment during the day as well. A tracheostomy is rarely required to provide ventilatory support, but in complex neuromuscular disorders it may be indicated to bypass upper airway obstruction, for instance due to vocal cord adduction, or to gain access to the tracheobronchial tree to aspirate secretions, or to protect the airway from aspiration of material from the pharynx.

Non-invasive ventilatory support can considerably improve the quality of life, arterial blood gases, maximum inspiratory and expiratory pressures, and the quality of sleep, as well as reducing the number of visits required by general practitioners and the quantity of drugs prescribed. Survival once treatment has been instituted is around 75 per cent at 5 years and 60 per cent at 10 years (Fig. 1).

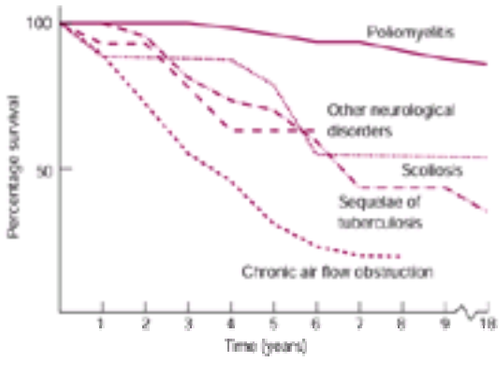

Fig. 1 Actuarial survival during treatment with ventilatory assistance for respiratory failure. (Reproduced with permission from Shneerson JM (1988). *Disorders of ventilation*, Blackwell, Oxford.)

Kyphosis

Exaggeration of the normal thoracic kyphosis is most commonly due to osteoporosis and is not usually associated with any significant changes in respiratory function. The exception to this is when a very sharp kyphosis (gibbus) develops. This is usually caused by tuberculous osteomyelitis of the spine (Pott's disease), although other conditions such as radiotherapy can cause a similar picture.

The spine becomes rigid in the region of the gibbus, and when tuberculosis is the cause the costovertebral joints also become ankylosed and limit the expansion of the rib cage. A restrictive defect in which the total lung capacity is reduced more than the residual volume is characteristic, but respiratory problems are uncommon unless the gibbus is high in the thoracic spine and develops in early childhood. This is probably because the thoracic deformity prevents the normal development of the lungs in a similar way to early-onset scoliosis. Hypoxia and hypercapnia appear during sleep before they become apparent during wakefulness, but may be severe. Pulmonary hypertension and right heart failure frequently develop once chronic hypercapnia has become established.

Slight breathlessness on exertion is common in the presence of a gibbus, but is rare in other types of kyphosis. Physical examination reveals the spinal deformity and limitation of rib cage expansion.

The posteroanterior chest radiograph shows superimposition of the spinal deformity on the lung fields and heart, which makes it difficult to interpret. The extent and severity of the kyphosis is usually well seen on a lateral projection. The typical changes in lung volumes have been described above. The arterial P_{O_2} and P_{CO_2} are usually normal, and, as in scoliosis, the earliest abnormalities are revealed by sleep studies.

Treatment of the acute tuberculous infection with chemotherapy often prevents a gibbus from developing. Once it has been established and respiratory failure has developed, the only effective treatment is long-term respiratory support. This is best provided non-invasively by a nasal positive pressure ventilator, rather than a negative pressure system, since the sharp kyphosis makes it difficult to lie in the supine position required for negative pressure ventilation.

Straight-back syndrome

In this disorder the normal thoracic kyphosis is absent or greatly reduced. This may result in a mild restrictive ventilatory defect, but cardiac problems are more prominent. The heart and great vessels may be compressed between the anterior rib cage and the spine with results similar to those seen in pectus excavatum. The right ventricular outflow tract or pulmonary artery may be narrowed, causing a systolic murmur, and occasionally right ventricular filling is impaired.

Ankylosing spondylitis

The initial manifestation of ankylosing spondylitis is usually painful inflammation of the sacroiliac joints, but this may spread to affect almost any joint including the intervertebral, costovertebral, manubriosternal, costochondral, and chondrosternal joints. When the inflammatory phase of the disease subsides, the joints become ankylosed and the spinal ligaments calcify.

The effect of ankylosing spondylitis on the thorax is that the rib cage becomes rigid. There is little spinal mobility, and a pronounced kyphosis often develops. The changes in lung volumes are characteristic in that, unlike all other skeletal disorders affecting the thorax, functional residual capacity increases. This is because the rib cage becomes fixed at its own relaxation volume. This is greater than the normal functional residual capacity which is influenced by the inward pull of the elastic recoil of the lungs. Total lung capacity and vital capacity are slightly reduced, and residual volume often increases.

The immobility of the rib cage leads to atrophy of the intercostal muscles and both maximal inspiratory and expiratory pressures are reduced. However, there is no impairment of diaphragmatic function and this largely compensates for the restriction of rib cage expansion. The ventilatory responses to exercise are virtually normal and exercise is usually limited by circulatory rather than respiratory factors. Respiratory failure is extremely uncommon in ankylosing spondylitis, probably as a result of the normal diaphragmatic function, unless another complication develops, which may be one of the following:

1. Air flow obstruction: cricoarytenoid arthritis is a feature of ankylosing spondylitis and may present with stridor, hoarseness of the voice, breathlessness, obstructive sleep apnoeas, or respiratory failure.
2. Pleural thickening and effusion: these rare complications of ankylosing spondylitis may precipitate respiratory failure.
3. Aspiration pneumonia: oesophageal motility is often impaired in ankylosing spondylitis and aspiration pneumonia may develop.
4. Bullae: apical fibrobullous lung disease is a feature of ankylosing spondylitis and may be complicated by opportunist infections such as *Aspergillus fumigatus* or saprophytic Mycobacteria, and occasionally pulmonary tuberculosis.
5. Abdominal surgery: this restricts diaphragmatic function on which adequate respiration depends. Conversely, thoracic surgery has relatively little effect on respiration because of the small contribution that rib cage expansion plays.

Chest pain during sudden movements such as coughing and laughing is common if the active phase of inflammation affects the thorax. These symptoms, which originate in either the joints or the muscles, become less prominent as the disease advances. Breathlessness and other respiratory symptoms are uncommon. Occasionally, cricoarytenoid arthritis may present with hoarseness, stridor, or breathlessness, and extensive fibrobullous disease may also cause breathlessness.

The most obvious physical sign is restriction of rib cage movement associated with prominent accessory muscle activity and abdominal respiratory movements.

Chest radiography may show calcification of the paraspinal ligaments (bamboo spine) and reveal evidence of complications of ankylosing spondylitis such as pleural thickening, aspiration pneumonia, and apical fibrobullous disease. The changes in lung volumes have been described above. Chest wall compliance is reduced but lung compliance is normal. The K_{CO} is increased and arterial blood gases are normal during both rest and exertion.

Physiotherapy and non-steroidal anti-inflammatory drugs may improve vital capacity and chest expansion, particularly in the early phase of the disease or during acute exacerbations.

Disorders of the sternum and ribs

Congenital abnormalities

Congenital abnormalities of the ribs and sternum rarely cause any important respiratory problems. Occasionally, multiple congenital rib abnormalities may lead to paradoxical movement of the chest wall or impair diaphragmatic function if they occur in the region of the insertion of this muscle. Severe congenital defects of the sternum, such as agenesis or a bifid sternum, are rare, but may require surgery in the neonatal period in order to stabilize the anterior chest wall.

Pectus excavatum

Pectus excavatum is a depression deformity of the sternum which is often present at birth but may worsen during the adolescent growth spurt. It is occasionally familial and may be associated with other abnormalities such as the straight-back syndrome or scoliosis. It appears to result either from an increased inward pull on the sternum by the sternal diaphragmatic fibres or from an abnormally compliant chest wall.

Transient paradoxical movement of the sternum during respiration is seen in eonates, particularly in the presence of upper airway obstruction or pneumonia. The

sternal depression may become permanent even if the cause, such as enlarged tonsils, resolves completely.

In adults pectus excavatum rarely causes any symptoms. Lung volumes are normal or only slightly diminished, and chest wall mobility appears to be normal. Arterial blood gases are normal both at rest and during exercise. Right ventricular filling can be impaired if the heart is compressed between the depressed sternum and the spine, and compression of the pulmonary outflow tract may cause a systolic murmur. These problems are most marked in the erect position and during exercise. Occasionally atrial arrhythmias develop. Surgery is sometimes indicated for cosmetic reasons, although the result can be disappointing. It has little or no effect on the mild restrictive defect or exercise ability, except in the rare situation when right ventricular filling is impaired or atrial arrhythmias have developed.

Pectus carinatum

Pectus carinatum is a protrusion deformity of the sternum in which the chest is often narrowed transversely as well. It becomes most marked during the pubertal growth spurt, although it may be present from birth and is occasionally associated with severe childhood asthma or ventricular septal defects. It is probably the result of excessive growth of the ribs or costal cartilages, and if this is asymmetrical the sternum becomes oblique.

The respiratory consequences of pectus carinatum have hardly been investigated. Chest pain may arise at the insertions of the intercostal muscles anteriorly, or in the costal cartilages and anterior ribs. Lung volumes appear to be normal, and surgery is indicated only for cosmetic reasons and not in order to improve respiratory function or exercise ability.

Asphyxiating thoracic dystrophy (Jeune's disease)

Asphyxiating thoracic dystrophy is a generalized disorder of cartilage in which the radiological changes are most prominent in the pelvis, phalanges, and other limb bones. Like the other long bones, the ribs are shortened so that the rib cage becomes narrowed. Lung development may be impaired as a result, and respiratory failure often appears during infancy or childhood. Surgical reconstruction of the rib cage with splitting of the sternum to enable lung growth to occur has been largely unsuccessful.

Acquired abnormalities

Flail chest

A flail chest is one in which multiple rib fractures cause paradoxical movement of the chest wall during respiration. It may be associated with other injuries, such as rupture of the aortic arch or spleen, and with fractures of the skull and long bones. It is frequently associated with pulmonary contusion, pneumothorax, or haemothorax.

Surgical stabilization of the chest wall is rarely required. In milder cases, sufficient analgesia to enable the patient to cough adequately may be all that is required, as long as the paradoxical movement does not impair alveolar ventilation. If it is more severe, positive pressure ventilation achieves 'pneumatic splinting' of the flail segment. The effectiveness of this has not been definitely established, but it appears that positive end expiratory pressure or continuous positive airway pressure is beneficial by preventing any negative pressure swings within the pleura.

Thoracoplasty

The operation of thoracoplasty was developed for the treatment of pulmonary tuberculosis. Varying lengths of up to 11 ribs were removed in order to collapse the chest on the affected side. It has been superseded by antituberculous chemotherapy, but is still occasionally required to treat chronic infections, particularly when there is a problem in obliterating the pleural space after pulmonary resection. It is estimated that as many as 30 000 operations were carried out in the United Kingdom between 1951 and 1960, and many of these patients still survive. Increasing numbers of this important cohort are being seen by chest physicians because of the late complications of the surgical procedure.

The consequences of thoracoplasty on respiratory function have been hard to elucidate because they are often combined with the effects of the underlying lung disease for which the surgery was carried out, and those of other treatments such as lung resection. However, the removal of the ribs has the direct result of flattening the chest and reducing the volume of the thorax. The normal movements of the rib cage may be impaired and paradoxical movement at the site of the thoracoplasty is common. The compliance of the chest wall is reduced, and it may fall further because the small range of movements of the costovertebral joints after surgery probably induces soft tissue changes which further limit the mobility of these joints. Chest wall compliance is also reduced by the almost invariable development of a thoracic scoliosis. This is convex to the side of the thoracoplasty and may progress for several years after the surgery. The severity of the scoliosis correlates with the number of ribs removed, but also depends on the details of the surgical technique.

Respiratory muscle function is impaired by a thoracoplasty. The intercostal and shoulder girdle muscles are directly damaged by the surgery, and distortion of the rib cage and the development of a scoliosis put the inspiratory muscles at an additional mechanical disadvantage. Diaphragmatic excursion is reduced, particularly on the side of the thoracoplasty, but also occasionally contralaterally.

The combination of reduced chest wall compliance and impaired respiratory muscle function accounts for the restrictive defect. All lung volumes are reduced, and in general the severity of the restrictive defect is proportional to the number of ribs that have been resected. A rapid respiratory rate with a small tidal volume is the characteristic respiratory pattern, particularly during exertion. Exercise is limited by ventilatory factors rather than by the cardiovascular system. In some patients, chronic air flow obstruction, which may be due to either tuberculous endobronchitis or the effects of tobacco smoking, may be significant, resulting in a progressive fall in exercise ability and contributing to the development of respiratory failure.

Ventilation and perfusion of the lung on the side of the thoracoplasty are usually reduced equally, so that the arterial $P\mathrm{O}_2$ often remains virtually normal. The function of the contralateral lung is much more important in determining the blood gases. Hypoxaemia usually first appears during sleep, as in scoliosis, and may be associated with hypercapnia. The presence of daytime hypercapnia correlates with the reduction in maximal inspiratory and transdiaphragmatic pressures.

The symptoms of patients with a thoracoplasty are similar to those with a scoliosis, but if a productive cough develops, recurrence of pulmonary tuberculosis should be suspected and investigated. Right heart failure often develops insidiously, either when respiratory failure appears or subsequently. It may be manifested by progressively worsening ankle swelling and fatigue. Physical examination reveals a thoracotomy scar and a flattened area of chest in the region of the thoracoplasty which may move paradoxically. Accessory muscle activity, particularly on the side of the thoracoplasty, is often marked.

The chest radiograph shows the extent of the thoracoplasty, other features which indicate the site and extent of previous tuberculous infection, and the sequelae of treatment such as a previous phrenic nerve crush or an artificial pneumothorax. This often causes extensive, calcified pleural thickening. The characteristic physiological defect is restrictive, but airflow obstruction may also be significant. Maximum inspiratory and expiratory pressures and transdiaphragmatic pressures are reduced. Most patients are mildly hypoxic, but later in the clinical course the arterial $P\mathrm{CO}_2$ may rise, particularly during sleep.

Life expectancy is reduced after a thoracoplasty for pulmonary tuberculosis: death occurs particularly from respiratory but also from cardiac causes. These complications are related to the extent of the tuberculosis and to whether or not an artificial pneumothorax was induced on the contralateral side to the thoracoplasty, since this often leads to pleural thickening and may indicate extensive tuberculous damage of the underlying lung. Respiratory failure can develop quite suddenly after a long period of stability, even when an acute illness such as a chest infection is not responsible.

Conventional treatment of airflow obstruction with, for instance, bronchodilators may be effective and right heart failure may respond to diuretics and angiotensin-converting enzyme inhibitors

Chronic ventilatory failure usually responds well to nocturnal, non-invasive respiratory support. Some patients can be managed adequately with oxygen during the day and/or at night as long as the $P\mathrm{CO}_2$remains normal or only slightly raised. When respiratory support is required, a nasal positive pressure ventilation or a negative pressure system such as a cuirass or jacket used at night is usually adequate initially. Some patients gradually require more intensive support, so that treatment is needed during the day as well as at night. This deterioration may be due to progressive worsening of small airway obstruction or respiratory muscle function, or to a

fall in oxygen delivery to the tissues caused by a deteriorating cardiac output associated with advancing pulmonary hypertension.

Disorders of the diaphragm

Aetiology

Diaphragmatic paralysis or paresis may be due to lesions affecting either the diaphragm itself or the phrenic nerve, its nucleus, or higher control centres or pathways. The most common causes of diaphragmatic weakness are shown in Table 2. Often no cause is found in unilateral weakness, and it is presumed to be due to a cryptogenic phrenic neuropathy, either as part of a widespread peripheral neuropathy or isolated to the phrenic nerves.

Pathophysiology

Unilateral weakness of the diaphragm causes it to move upwards (paradoxically) into the thorax during inspiration, instead of descending. This decreases the tidal volume and the mechanical efficiency of breathing. It is worse in the supine position when the weight of the abdominal contents pushes the paralysed diaphragm further into the thorax and decreases the functional residual capacity. The diaphragm is splinted in an expiratory position so that it moves relatively little even though it is paralysed. When the subject lies on one side, the lower half of the diaphragm behaves in this way if it is paralysed, but if the upper half is paralysed it moves paradoxically.

The loss of inspiratory muscle strength is partially compensated by recruitment of intercostal and accessory muscles, but the maximum inspiratory and transdiaphragmatic pressures are reduced. The vital capacity in the upright position is approximately 20 to 25 per cent less than normal and a further fall of about 15 per cent occurs when lying supine. Similar changes in the total lung capacity and functional residual capacity occur. The residual volume is unchanged and expiratory muscle strength is largely preserved.

The distribution of ventilation and perfusion is affected by unilateral diaphragm weakness. Ventilation is slightly diminished, particularly at the base on the side of the diaphragmatic paralysis in the sitting position, but this is more marked when supine. Similar changes occur with perfusion on a regional basis, but ventilation–perfusion matching is impaired and hypoxia results. Hypercapnia does not occur during wakefulness or sleep.

The physiological abnormalities seen with bilateral diaphragmatic weakness in adults are much more marked than in unilateral diaphragmatic disorders. The diaphragm moves paradoxically during inspiration and expiration, and intrapleural pressure changes are transmitted across it so that abdominal pressure falls during inspiration and the anterior abdominal wall moves paradoxically. The maximum transdiaphragmatic pressure falls in proportion to the degree of diaphragm weakness, and since the diaphragm is the main inspiratory muscle, the maximum inspiratory pressure is correspondingly reduced. The vital capacity in the sitting position is about 50 per cent of that predicted and may fall by a further 50 per cent when supine. The influence of the supine position is greater than with unilateral diaphragmatic weakness because the weight of the abdominal contents pushes both halves of the diaphragm into the thorax. Ventilation is particularly reduced at the bases in the supine position, with less change in perfusion so that the arterial P_{O_2}falls. This postural change is partly responsible for the hypoxia that has been observed during sleep, but the rapid respiratory rate, small tidal volume, and short inspiratory time contribute to this and to hypercapnia.

Symptoms and physical signs

Unilateral diaphragmatic paralysis in adults rarely causes symptoms unless there is coexisting pulmonary disease or weakness of other respiratory muscles. In contrast, bilateral weakness can cause severe breathlessness. This may occur during exertion, but a specific feature is orthopnoea. This occurs within a few seconds of lying flat and is relieved promptly by sitting up, in contrast to left ventricular failure and nocturnal asthma with which it is frequently confused. Breathlessness may also occur when standing in water since the passive inspiratory descent of the diaphragm due to gravity is prevented by the raised extra abdominal pressure.

The physical signs of unilateral diaphragm weakness can be subtle. Dullness to percussion over the lower part of the thorax may be present and the level of dullness may rise paradoxically on the paralysed side during inspiration. The normal inspiratory outward movement of the abdomen may be reduced or absent on the side of diaphragmatic paralysis, and expansion of the lower chest may lag behind the normal expansion of the other side.

The physical signs of bilateral diaphragmatic paralysis are much more clear cut. Orthopnoea is usually readily apparent and the abdomen moves paradoxically inwards as the diaphragm ascends during inspiration. A maximum transdiaphragmatic pressure of less than 30 cmH_20 is necessary for this sign to be detected. The accessory muscles are active, particularly in the supine position. The quality of sleep is often poor and as a result excessive daytime somnolence may be a problem. Bilateral, basal dullness due to the high diaphragms is characteristic, but can be mimicked by bilateral pleural effusions.

Investigations

The chest radiograph in unilateral diaphragmatic paralysis shows whether the affected diaphragm is elevated and usually reveals any adjacent mass that may be responsible. If there is bilateral paralysis, both the diaphragms are raised. There is often some basal linear shadowing due to subsegmental lung collapse. Diaphragmatic screening or ultrasound examination reveals that the diaphragm moves paradoxically, particularly during sniffing. This test should be carried out in the supine position with a weight on the abdomen. These precautions prevent abdominal muscle contraction during expiration from mimicking diaphragmatic activity by reducing the end expiratory volume below functional residual capacity, so that inspiration then occurs through the elastic recoil of the lungs and chest wall. In the upright position, the effect of gravity on the abdominal contents can lead to inspiration without any diaphragmatic activity.

A low vital capacity, which falls further in the supine position, is the hallmark of diaphragmatic weakness, particularly when this is severe and bilateral. All lung volumes are reduced except for the residual volume since expiratory muscle strength is largely preserved. Maximum inspiratory pressure is also reduced, but diaphragmatic weakness can be more specifically diagnosed by estimating the transdiaphragmatic pressure. This can be carried out by asking the patient to sniff or to take a maximum inspiratory effort, or by magnetic or percutaneous electrical stimulation of the phrenic nerve in the neck. Care is required to carry out these investigations using a standardized method in order to obtain repeatable results. The function of the phrenic nerve can be estimated by measuring its conduction time. This is normally less than about 9.5 ms, but is prolonged if the nerves are diseased.

The arterial $P_{O_2}$is characteristically slightly reduced, with a normal P_{CO_2}during the daytime and in sleep as long as pulmonary function is normal and there is no other muscle weakness. If either of these are present, however, bilateral diaphragmatic weakness can cause hypercapnia with profound hypoxia during sleep.

Treatment

Plication for hemidiaphragmatic paralysis is rarely required in adults unless coexistent pulmonary disease is severe enough to cause breathlessness. Bilateral plication is not effective and mechanical respiratory support is often required if there is bilateral weakness. Treatment with nasal positive pressure ventilation or a negative pressure ventilator, such as a cuirass or jacket is usually required, although rocking beds have also been found to be effective in the past. Ventilatory support is usually needed only at night and until the function of the diaphragm or phrenic nerve improves. Phrenic nerve pacemakers are only indicated when diaphragmatic weakness is due to lesions above the phrenic nerve nucleus in C3 to 5 or 6. The commonest cause of this is a high cervical spinal cord injury. Breathlessness on exertion often remains a problem, but may lessen as other inspiratory muscles partially compensate for diaphragmatic weakness.

Further reading

Bredin CP (1989). Pulmonary function in long-term survivors of thoracoplasty. *Chest* **95**, 18–20.

Dolmage TE, Avendano MA, Goldstein RS (1992). Respiratory function during wakefulness and sleep among survivors of respiratory and non-respiratory poliomyelitis. *European Respiratory Journal* **5**, 864–70.

Elliott CG, Hill TR, Adams TE, Crapo R, Nietrzeba RM, Gardner RM (1985). Exercise performance of subjects with ankylosing spondylitis and limited chest expansion. *Bulletin Eurropeen de Physiopathologie Respiratoire* **21**, 363–8.

Franssen MJAM, van Herwaarden CLA, van de Putte LBA, Gribnau FWJ (1986). Lung function in patients with ankylosing spondylitis. A study of the influence of disease activity and treatment with

non-steroidal antiinflammatory drugs. *Journal of Rheumatology* **13**, 936–40.

Gibson GJ (1989). Diaphragmatic paresis: pathophysiology, clinical features, and investigations. *Thorax* **44**, 960–70.

Haller JA Jr, Colombani PM, Humphries CT, *et al*. (1996). Chest wall constriction after too extensive and too early operations for pectus excavatum. *Annals of Thoracic Surgery* **61**, 1618–25.

Kafer ER (1975). Idiopathic scoliosis. Mechanical properties of the respiratory system and the ventilatory response to carbon dioxide. *Journal of Clinical Investigation* **55**, 1153–63.

Kinnear WJM, Hockley S, Harvey J, Shneerson JM (1988). The effects of one year of nocturnal cuirass-assisted ventilation in chest wall disease. *European Respiratory Journal* **1**, 204–6.

Laroche CM, Moxham J, Green M (1989). Respiratory muscle weakness and fatigue. *Quarterly Journal of Medicine, New Series* **71** (265), 373–97.

Lindahl T (1954). Spirometric and bronchospirometric studies in five-rib thoracoplasties. *Thorax* **9**, 285–90.

Midgren B, Petersson K, Hansson L, Eriksson L, Airikkala P, Elmqvist D (1988). Nocturnal hypoxaemia in severe scoliosis. *British Journal of Diseases of the Chest* **82**, 226–36.

Mier-Jedrzejowicz A, Brophy C, Moxham J, Green M (1988). Assessment of diaphragm weakness. *American Review of Respiratory Disease* **137**, 877–83.

Newsom-Davis J, Goldman M, Loh L, Casson M (1976). Diaphragm function and alveolar hypoventilation. *Quarterly Journal of Medicine* **145**, 87–100.

O'Brien JW, Johnson SH, Van Steyn SJ, *et al.* (1991). Effects of internal mammary artery dissection on phrenic nerve perfusion and function. *Annals of Thoracic Surgery* **52**, 182–8.

Pehrsson K, Bake B, Larsson S, Nachemson A (1991). Lung function in adult idiopathic scoliosis: a 20 year follow up. *Thorax* **46**, 474–8.

Phillips MS, Kinnear WJM, Shneerson JM (1987). Late sequelae of pulmonary tuberculosis treated by thoracoplasty. *Thorax* **42**, 445–51.

Phillips MS, Kinnear WJM, Shaw D, Shneerson JM (1989). Exercise responses in patients treated for pulmonary tuberculosis by thoracoplasty. *Thorax* **44**, 268–74.

Ras GJ, van Staden M, Schultz C, *et al*. (1994). Respiratory manifestations of rigid spine syndrome. *American Journal of Respiratory and Critical Care Medicine* **150**, 540–6.

Robert D, Gerard M, Leger P, *et al*. (1983). La ventilation mechanique a domicile definitive par tracheostomie de l'insuffisant respiratoire chronique. *Revue Francaise des Maladies Respiratoires* **11**, 923–6.

Sawicka EH, Branthwaite MA, Spencer GT (1983). Respiratory failure after thoracoplasty: treatment by intermittent negative-pressure ventilation. *Thorax* **28**, 433–5.

Shneerson JM (1978). The cardiorespiratory response to exercise in thoracic scoliosis. *Thorax* **33**, 457–63.

Shneerson J (1998). Sleep in neuromuscular thoracic cage disorders. *European Respiratory Monograph* **10**, 324–44.

Smith IE, Laroche CM, Jamieson SA, Shneerson JM (1996). Kyphosis secondary to tuberculous osteomyelitis as a cause of ventilatory failure: Clinical features, mechanisms and management. *Chest* **110**, 1105–10.

Tzelepis GE, McCool FD, Hoppin FG Jr (1989). Chest wall distortion in patients with flail chest. *American Review of Respiratory Disease* **140**, 31–7.

17.14.1 Lung cancer

S. G. Spiro

General epidemiology

Lung cancer is the most common malignant disease in the Western world. It has shown the greatest relative and absolute rise in mortality of any tumour this century in England and Wales, and particularly in Scotland. It causes 38 000 deaths per year in England and Wales, with 80 per cent of these occurring in men. In the European Community, there were 1.35 million deaths per annum in men (the highest death rate from any tumour) and 229 000 deaths per annum in women during the period 1978 to 1982. In the United States, it has been increasing in incidence by up to 10 per cent per year since the 1930s, but over the last decade this trend has levelled off, particularly in men. Nevertheless, approximately 120 000 American men die of lung cancer each year; the figure for women being 34 000, similar to that for breast cancer.

Age-standardized mortality rates for cancer for 1985 to 1989 show that in the European Community lung cancer in men was by far the commonest cause of death. Belgium has the highest mortality (77.16 deaths per 100 000 population) with Scotland (75.9) second, and England and Wales (60.9) fifth. The figures for Central and Eastern Europe are worse in that the death rates for lung cancer are rising exponentially, particularly in men—75.8/100 000 in the Czeck Republic, 74.0 in Hungary, 69.4 in Poland, and 68.7 in Slovakia. For females, Scotland has the highest incidence (27.2, equal to the rate of breast cancer in Scottish women), with England and Wales (20.4) third. Age-adjusted lung cancer death rates in Eastern Europe are still considerably less than in the Western countries, ranging from 14.4 in Hungary to 6.8 in Slovakia. Perhaps the worst epidemic is in China where 0.8 million men will die in the year 2000 from smoking-related diseases. The relative risk of dying from lung cancer is about 3.0 in male smokers compared to non-smokers, and 2.0 for female smokers. Of all deaths attributed to tobacco in China, 15 per cent were due to lung cancer.

Aetiological factors

Tobacco

In every country, the increase in mortality from lung cancer has appeared to coincide with an increase in tobacco usage, particularly cigarette smoking, after what seemed to be an appropriate latent interval. Early retrospective studies showed that, amongst patients with carcinoma of the bronchus, there were many fewer non-smokers and many more heavy smokers than among the controls, and that there was an association between the amount smoked and the risk of lung cancer. Prospective studies, amongst which the long-term study of British doctors was particularly informative, confirmed the increased risk of death from lung cancer from any tobacco use, but most specifically that of cigarettes. There was a strong dose–response relationship with the number of cigarettes smoked, illustrated in Table 1. The most important variable in smoking intensity is the number of cigarettes smoked, but other variables include the depth of inhalation, number of puffs, butt length, use of a filter, and the type of tobacco smoked. Further evidence that the relationship was causal came from a study which documented reduction in mortality after stopping smoking: 15 years after cessation the risk of death fell from 15.8 times to twice that in non-smokers, equivalent to 11 per cent of that pertaining in those who continued to smoke.

Globally, there has been a huge change in cigarette consumption. While there has been a drop of 25 per cent and 9 per cent in consumption in the United Kingdom and the United States, respectively, between 1970 and 1985, the overall world consumption has risen by 7 per cent. This is due to huge increases in Asia (22 per cent), Latin America (24 per cent), and Africa (42 per cent). The current epidemic of smoking in China lags behind Western society by 20 years. Thus in China in 1996, the average number of cigarettes smoked per adult male was 11 per day, a figure that that peaked in the West at 10 a day in 1980.

Wide differences in smoking habits in the United Kingdom are seen between social classes, with 57 per cent of unskilled manual workers smoking compared with only 21 per cent of professional workers. During the last 5 years the number of adult men smoking in England and Wales has fallen from 64 to 36 per cent, but has remained at 35 per cent for adult women. The effect of the lower-tar cigarettes has not yet had time to become established.

Passive smoking

Evidence that passive smoking predisposes to lung cancer is far from certain. Approximately 15 per cent of lung cancers occur in non-smokers, and 5 per cent of these have been attributed to passive smoking.

Occupation

A number of different factors have now been identified as associated with lung cancer; subjects who develop this disease as a result of their occupation represent a small but important group. The association of asbestos with lung cancer is now firmly established; various studies have identified that those exposed are at 4.9 to 7.3 times greater risk than those who are not. This risk is much enhanced if the asbestos industry worker smokes cigarettes; one study estimating this at 93 times that for non-smokers not exposed to asbestos. In Norway, it is illegal to employ a smoker in an asbestos-related job. Exposure to radioactive isotopes, mainly radon

daughters, is associated with a higher risk of lung cancer and occurs among various groups of miners, particularly those involved in extraction of pitchblende and uranium. Polycyclic aromatic hydrocarbons are believed to be responsible for the increased risk in workers in gas and coke ovens and in foundry workers. Nickel refining, chromate manufacture, and arsenical industrial workers are also exposed to a higher risk of lung cancer. The amount of lung cancer caused by occupational exposure may well have been underestimated in the past, and a summary of the importance industrial products and processes involved appears in Table 2.

Air pollution

The decline in male mortality is occurring earlier than would be expected from changes in smoking habits. The high mortality figures in the United Kingdom and Germany compared with France and Italy, for example, seem likely to be due in part to heavy industry and coal burning. Analysis by county in the United States shows an association between lung cancer deaths and counties with chemical, petroleum, ship-building, and paper industries. Legislation for cleaner air has caused both environmental and occupational pollution to fall dramatically in the past 30 years, and this has preceded changes in smoking habits.

Pathology

A detailed understanding of the natural history, pathology, and pathogenesis of bronchial carcinoma is becoming increasingly important as the assessment, management, and prognosis of the disease depends largely upon the cell type and the presence or absence of metastases at the time of presentation. It has been estimated that about seven-eighths of a tumour's life will have passed when it is diagnosed and that the vast majority will be disseminated at the time of diagnosis.

Bronchogenic carcinomas seem to arise most commonly in segmental and subsegmental bronchi in response to repetitive carcinogenic stimuli or inflammation and irritation. The mucosal lining is most susceptible to injury at the bifurcation of bronchial structures. Dysplasia is followed by carcinoma *in situ* when the entire thickness of the mucosa may be replaced by proliferating neoplastic cells. These changes may be strictly localized or multicentric. Tumour infiltration follows loss of the basal membrane. The precise origins of small-cell carcinomas remain an enigma, and those of adenocarcinomas are not precisely defined. The latter may arise from the mucosal lining or from the submucosal bronchial mucous glands. A significant number of lung tumours arise in the periphery of the lung, perhaps three-quarters of adenocarcinomas and large-cell anaplastic malignancies, one-third of squamous (or epidermoid) carcinomas, and one-fifth of small-cell carcinomas.

The WHO classification of lung cancer according to cell type and the approximate distribution of each type as a percentage of all lung cancers is shown in Table 3. The squamous-cell tumour has a relatively slow growth rate (volume doubling time, 90 days) and the lowest incidence of distant haematogenous metastasis. Small-cell tumours grow rapidly (volume doubling time, 30 days) and there is very early dissemination by both the haematogenous and the lymphatic routes, with metastasis being present in more than 90 per cent of patients at the time of diagnosis. Adenocarcinomas and anaplastic large-cell tumours occupy an intermediate position. It is now recognized that significant heterogeneity of cell morphology can be visualized within individual tumours. Squamous-cell tumours, adenocarcinomas, and large-cell tumours are often collectively called non-small-cell lung cancers, and the approach to their management differs from that for small-cell lung cancer.

Squamous (epidermoid) carcinoma

These tumours are composed predominantly of flattened to polygonal neoplastic cells that tend to stratify, form intercellular bridges, and elaborate keratin. About 60 per cent present as obstructive lesions in lobar and main-stem bronchi. The tumours tend to be bulky and to produce intraluminar granular or polypoid masses. As a result, distal pneumonia and abscess formation is common, and cavitation is seen in about 10 per cent. The cells are usually well differentiated, but in some cases differentiation is poor and the appearances are those of predominantly anaplastic cells, frequently arranged in the classical pattern of stratifying sheets.

Small-cell (oat-cell) anaplastic carcinoma

This is now recognized as a pathologically and clinically distinct form of lung cancer. Small-cell lung cancer may originate from the amine precursor uptake and decarboxylation (APUD) series of cells. The tumour is composed of neoplastic cells with dark oval to round spindled nuclei and scanty, indistinct cytoplasm arranged in ribbons, nests, and sheets. The cells tend to crush easily on biopsy, and extensive areas may be necrotic. This type of tumour presents as a proximal lesion in 75 per cent of cases and may arise anywhere in the tracheobronchial tree and rapidly invade vessels and lymph nodes, disseminating widely even before symptoms arise from the primary tumour. Extensive, advanced disease exists in more than half the patients at presentation. The cells secrete peptides which cause clinical syndromes in 10 per cent of cases.

Adenocarcinoma

This tumour forms acinar or granular structures, having prominent papillary processes, and may be mucin-provoking. About 70 per cent appear to originate peripherally in the lung and are frequently fairly circumscribed; in about 10 per cent, the initial presentation is a pleural effusion. If related to bronchi, they tend to cuff and stenose the lumen. They occasionally arise in old tuberculous scars.

Large-cell carcinoma

These tumours, which have been described as an unclassified category, include all tumours that show no evidence of maturation or differentiation. They are composed of pleomorphic cells with variable enlarged nuclei, prominent nucleoli, and nuclear inclusions, and abundant cytoplasm; they are mucin-producing in many instances. The tumours tend to be bulky and are often necrotic; they are frequently peripheral, they invade locally, and disseminate widely, with about half the patients having disseminated disease on presentation. Although they are highly malignant and undifferentiated, the cure rate after surgery is surprisingly high, but radiotherapy is ineffective in controlling the disease. Large-cell carcinoma is a smoking-related disease in over 90 per cent of patients.

Bronchioloalveolar carcinoma

There has been considerable controversy as to whether this tumour, which has the least association with tobacco smoking, arises from alveolar or bronchial epithelium, but derivation from the alveolar type II cell has been suggested. The tumour tends to spread as cuboidal or columnar 'epithelium' along the lining of the alveoli, with single or multiple rows of cells and often papillary formation (Fig. 1). There is production of a large amount of mucus in 20 per cent of cases and it is believed that malignant cells shed into the mucus may carry over into the contralateral lung. The tumour can spread within a lobe and occupy it fully. Sometimes, however, the tumour is multicentric in origin, and diffuse nodular lesions are to be found on radiographic examination. Invasion of neighbouring tissue and lymph nodes is common, but extrathoracic spread is unusual. There is some resemblance to metastases from adenocarcinomas emanating from other organs, and this sometimes leads to confusion. The tumour tends to grow along alveolar septae as a framework, and it may be difficult to distinguish from metastatic tumours from colon, breast, or pancreas.

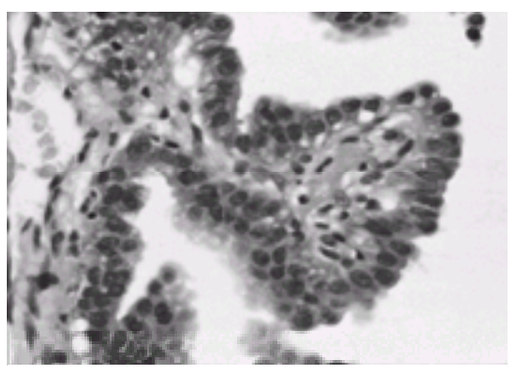

Fig. 1 Bronchoalveolar cell carcinoma: malignant cuboidal epithelium spreads along alveolar walls.

Carcinoid tumours

Carcinoid tumours are described in Chapter 14.8.

Carcinoma *in situ*

Many investigators have suggested that cells undergoing malignant change do not necessarily invade the lungs at the onset of this biological mutation, but continue to exist at a particular location (cancer *in situ*). Exfoliated cancer cells sloughed from such a location may be seen fortuitously by the cytologist; even more rarely, such a site may be biopsied at bronchoscopy.

Genetics and biology

Genetic influences may play a role in the development of lung cancers, particularly in patients under 50. In one study, lung cancers were attributable to a mendelian dominant inheritance pattern in 27 per cent of patients under 50, but only 9 per cent of those over 70.

Both small-cell and non-small-cell lung cancer can be cultured as cell lines which can be transplanted into nude mice. The resultant tumours are morphologically and histologically similar to the original human tumour. Two major categories of cell lines are established for small-cell lung cancer—classic and variant. Classic cell lines account for 70 per cent of the total and are characterized by high expression of neuroendocrine markers such as L-dopa decarboxylase and bombesin/gastrin-releasing peptide, neurone-specific enolase, and creatine kinase-BB. The variant cell lines have selective loss of some of these neuroendocrine markers, and many have substituted amplification of the c-*myc* oncogene. These neuroendocrine and proto-oncogenic properties are thought to be of prognostic significance, playing an important role in regulation of tumourigenesis. In non-small cell lung cancer, approximately 8 per cent of tumours and 20 per cent of tumour cell lines have myc family amplification, and overexpression of these proteins is associated with poorer prognosis.

The ras family of oncogenes (H, K, and N) was the first to be described in association with lung cancer. Ras gene mutations occur in 20 to 40 per cent of non-small cell lung cancer, especially adenocarcinomas, and the presence of K- *ras* mutations is linked with significantly shortened survival.

Lung cancer cells not only show mutations that activate dominant cellular proto-oncogenes, but also genetic mechanisms that inactivate recessive tumour suppressors. The commonest abnormality is a deletion in the short arm of chromosome 3, which is found in over 90 per cent of small-cell lung cancer and 50 per cent of non-small-cell lung cancer patients. Other sites of loss of heterozygosity include 11p, 13q, and 17p. Tumour suppressor genes have been identified in inherited cancers, mainly in studies of familial retinoblastoma. Mutations in P53 occur in 75 per cent of small cell lung cancer and 50 per cent of non-small cell lung cancer. The gene is located on the short arm of chromosome 13q14, and it is thought that it may normally protect cells against accumulation of mutations. Depletions and mutations of P53 are linked with metastatic disease. Alterations of P53 protein have been found in early bronchial neoplasia, and may be a useful marker for the early detection of lung cancer. Other markers, including heterogenous nuclear ribonuclear protein A2/B1 overexpression in sputum, may allow earlier detection of tumours.

Several monoclonal antibodies have been generated against lung-cancer-associated antigens. Thirty-six monoclonal antibodies raised against small-cell lung cancer have been grouped into eight clusters. No antigen is specific for small-cell lung cancer. Antibodies belonging to the major cluster (cluster 1) are directed against the neural-cell adhesion molecule (NCAM), whilst the nature of the other antigens remains unclear. Studies of both small-cell and non-small-cell lung cancer cell lines show that NCAM expression is associated with a neuroendocrine phenotype irrespective of the histological type of lung cancer. Monoclonal antibodies may have a therapeutic value when coupled with a radionuclide or a toxin. Radiolabelled antibodies can be used to detect minimal disease in bone marrow aspirates or biopsy specimens.

The growth factors bombesin/gastrin-releasing peptide, insulin-like growth factor 1, and transferrin stimulate can all stimulate tumour growth. There is much interest in attempts to retard or disrupt these processes.

Clinical features

The clinical features of lung cancer are very variable. In about 5 per cent of patients the presentation is a radiographic abnormality found on routine examination and unassociated with symptoms, but patients may present with extremely advanced disease and die rapidly.

Clinical manifestations may be due to: local presence of the tumour in the lung, including bronchial obstruction or invasion of contiguous structures in the thorax and mediastinum; metastasis through blood or lymph vessels; and endocrine, metabolic, and neurological syndromes.

Cough is the most common initial presenting symptom (Table 4), but because it is a symptom of so many respiratory disorders, the possibility of tumour may be overlooked and cough may be attributed to some other cause, particularly in smokers who have had chronic bronchitis for many years. Patients who have a persistent cough should have a chest radiograph, particularly if they are smokers over 40 years of age. A change in the cough habit is significant and also requires investigation. If the trachea or main bronchi are involved, the cough may be brassy in character and may be accompanied by wheezing or stridor. If cough is manifestly ineffective, involvement of the recurrent laryngeal nerve should be suspected.

Expectoration of sputum may be due to spread of the tumour itself or to infection occurring distal to partial bronchial obstruction. In the early stages of the disease the sputum is often grey and viscid; it is usually purulent in the presence of infection distal to a tumour and in cavitated tumours. The value of sputum cytology in diagnosis is described below.

Haemoptysis, which occurs as a sole presenting symptom in about 5 per cent of cases and at some stage in the disease in 50 per cent of patients, is a symptom not easily ignored by patient or physician. The degree varies from streaking of the sputum with blood to larger amounts, but massive haemoptysis (>200 ml) is rare, except as a terminal event. The most significant description given by patients is that of coughing up blood every morning for several days in succession.

Wheeze may be observed in a few patients. Localized persistent wheeze even after coughing is a significant observation indicating obstruction of a larger or central airway.

Stridor is a feature which is poorly recognized and is often confused with wheeze. It is due to narrowing of the glottis, trachea, or major bronchi, and is best heard after the patient coughs and then breathes in deeply with the mouth open.

Dyspnoea is a presenting symptom in only a small number of patients. As the disease progresses dyspnoea is inevitable, being proportional to the amount of lung involved, either directly by tumour replacement or indirectly by endobronchial disease causing airway narrowing or obstruction. Progressive breathlessness is also a salient feature of malignant pleural and, rarely, pericardial effusion, superior vena caval obstruction, and lymphangitis carcinomatosis.

Chest discomfort is a common symptom, occurring in up to 40 per cent of patients at diagnosis. The discomfort is often of an ill-defined nature and may be described in terms of intermittent aching somewhere in the chest. Definite pleural pain may occur in the presence of infection, but invasion of the pleura by tumour is often painless. However, invasion of the ribs or vertebrae causes continuous, gnawing, localized pain. A tumour in the superior pulmonary sulcus (Pancoast tumour) can cause progressive constant pain in the shoulder, upper anterior chest, or interscapular region, soon spreading to the arm once the brachial plexus is invaded. Other symptoms of this type of tumour include weakness and atrophy of the muscles of the hand, Horner's syndrome, hoarseness, and spinal cord compression at levels D1 and D2.

Lack of energy and, more particularly, loss of interest in normal pursuits are symptoms of great importance; a sensation of vague ill health commonly occurs.

Fever, chills, and night sweats may occur due to chest infection, but fever may very rarely be present in rapidly progressive tumours without evidence of infection, particularly if there are hepatic metastases.

Invasion of adjacent intrathoracic structures gives rise to certain specific clinical features. Involvement of the last cervical and first thoracic segment of the sympathetic

trunk by cancer produces Horner's syndrome. Malignant infiltration of the recurrent laryngeal nerve—almost always the left branch because of its course adjacent to the left hilum—gives rise to vocal chord paralysis. The right recurrent laryngeal nerve is occasionally affected in the base of the neck. Recurrent aspiration pneumonias may follow vocal chord paralysis. Extension of the tumour with invasion or compression of the superior vena cava or by paratracheal lymphadenopathy results in the characteristic features of superior vena caval obstruction—awareness of tightness of the collar, fullness of the head, and suffusion of the face, particularly after bending down, blackouts, breathlessness, and engorgement of veins with a downward venous flow in the neck, the upper half of the thorax, and arms, often accompanied by oedema of the face.

Dysphagia is due to compression of the oesophagus from without by tumour masses and only rarely to direct invasion. Cardiac metastases usually occur late in the disease and are manifested clinically by tachycardia, arrhythmias, pericardial effusion, breathlessness, and cardiac failure. Invasion of the phrenic nerve results in elevation and paralysis of the hemidiaphragm.

The clinical features associated with involvement of the ribs, spine, and pleura are described elsewhere. Very rarely bronchogenic carcinoma causes spontaneous pneumothorax. It must not be forgotten that spread of tumour to the other lung may occur or that synchronous primaries may coexist.

Metastatic lesions from lung cancer may occur in any organ of the body and produce symptoms which form the presenting complaint. Metastases to nodes are frequent and should be sought with great care, particularly those in the scalene area, which are usually the first to be involved and can be palpated. The best position for examination is from behind with the patient seated relaxed in a chair. The side affected usually corresponds to the side of the lung lesion, the exception being that tumours from the left lower lobe may metastasize to the nodes in the right scalene area. Involvement of the nodes in the floor of the supraclavicular fossa is equally common.

Bony metastases are common, particularly in small-cell tumours, and occur predominantly in the ribs, vertebrae, humeri, and femora. Early involvement may be detected by a rise in alkaline phosphatase of bony origin, isotope scanning, or biopsy. Conventional skeletal surveys are often unhelpful and misleading. Liver secondaries are common and may be silent, although a rise in liver enzymes, particularly alkaline phosphatase of liver origin, may be an early sign. Isotope liver scans and ultrasound may detect involvement in a liver which is not clinically enlarged, but as the metastases develop the liver becomes grossly enlarged with an irregular outline. Friction rubs may sometimes be heard over a grossly involved liver. Metastases to brain may account for the presenting symptom in lung cancer in 4 per cent of patients and may be encountered at some time in the illness in 30 per cent. The symptoms simulate those of any expanding brain tumour. The adrenal glands are involved in 15 to 20 per cent of patients, rarely producing symptoms. The skin should be examined for the presence of the typical, slightly bluish, umbilicated lesions of tumour spread. Subcutaneous metastases may be found at almost any site.

Endocrine and metabolic manifestations

It is becoming more apparent that many of the hitherto unexplained and often unusual manifestations of malignant disease are the result of endocrine and metabolic manifestations of the cancer itself. Cancer cells appear to be able to synthesize polypeptides that mimic virtually all the hormones produced by conventional endocrine organs—hence the term 'ectopic hormones'. From time to time the clinical features resulting from ectopic hormone secretion precede those of the pulmonary tumour, emphasizing the importance of a high index of suspicion in such circumstances. Ectopic hormone measurement cannot, however, be used for screening purposes.

Syndrome of inappropriate secretion of antidiuretic hormone (SIADH)

The continued secretion of vasopressin (ADH) in an amount in excess of the body's needs leads to overhydration in both the intracellular and extracellular compartments. The cerebral oedema resulting from water intoxication causes drowsiness, lethargy, irritability, mental confusion, and disorientation, with fits and coma being the most profound features. Peripheral oedema is remarkably rare. The patient is usually asymptomatic until the sodium falls below 120 mmol/l and the hyponatraemia is dilutional in type with a low serum osmolality. Urine osmolality usually exceeds 300 mosmol/kg. The commonest cancer causing this syndrome is small-cell cancer, where it is clinically obvious in 10 per cent of cases, with subclinical involvement detectable by a water-loading test in more than 50 per cent. Restriction of fluid to a daily intake of 700 to 1000 ml may redress the hyponatraemia, but demethylchlortetracycline (demeclocycline) 600 to 1200 mg daily is often highly effective, making water restriction unnecessary. Azotaemia may occur as a result of increased urea production and a mild drug-induced nephrotoxicity so that adjustment of dosage may be necessary. Infusion of hypertonic saline is hazardous, often precipitating cardiac failure or cerebral oedema.

Ectopic ACTH syndrome

Secretion of an adrenocorticotrophic substance by a small-cell carcinoma or bronchial carcinoid leads to bilateral adrenal hyperplasia and to secretion of large amounts of cortisol. The onset of symptoms may be so acute that death may occur within a few weeks, and the typical features of Cushing's syndrome do not have time to develop. Chief clinical features are thirst and polyuria, oedema, pigmentation, and hypokalaemia. Hypertension and profound myopathy may also be present. Serum cortisol is often grossly elevated, with loss of the normal diurnal rhythm; the level is not suppressed by dexamethasone; and hypokalaemic alkalosis can be severe, with plasma potassium frequently below 3.0 mmol/l and HCO_3 above 30 mmol/l. Drugs which block adrenocortical steroid biosynthesis may produce partial and reversible medical adrenalectomy, and metyrapone in doses from 250 mg thrice daily to 1 g four times daily may cause temporary relief of symptoms. Removal of the tumour, if practicable, will cause remission.

Hypercalcaemia (see also Chapter 12.6)

Hypercalcaemia may be associated with ectopic secretion of parathormone by squamous-cell cancers but is more commonly due directly to the presence of multiple bone metastases. The primary tumour may also produce a cyclic-AMP-stimulating factor or a prostaglandin causing hypercalcaemia. Hypercalcaemia is unlikely to cause symptoms unless the serum calcium exceeds 2.8 mmol/l, and levels much higher than this are sometimes encountered. The main clinical features are nausea, vomiting, abdominal pain and constipation, polyuria, thirst and dehydration, muscular weakness, psychosis, drowsiness, and eventually coma. Immediate treatment is to relieve fluid depletion, and large volumes of intravenous saline (up to 5 litres in 24 h) may be required. Corticosteroids (400 mg hydrocortisone or 40 mg prednisolone in 24 h initially) are effective in about half of the cases. However, intravenous diphosphonates followed by oral maintenance therapy is now the treatment of choice. Other treatments which are sometimes effective are calcitonin 200 to 400 units every 8 h, mithramycin 10 to 15 µg/kg by infusion over 4 h every 21 days, aspirin 2 to 4 g/day, and indomethacin 50 to 100 mg/day. The calcium level drops dramatically within 48 h if the tumour is removed.

Gynaecomastia

Swelling of the breasts, which may be painful, occurs mainly in the subareolar area, and there may be atrophy of the testes. The association is chiefly with large-cell carcinomas. Increased gonadatrophin production is the cause.

Other endocrine manifestations

Hyperthyroidism occurs rarely, but neither goitre nor eye signs are prominent features. Spontaneous hypoglycaemia, the masculinizing syndrome in young women, and hyperglycaemia are very rarely encountered. Pigmentation associated with a- and b-melanocyte-stimulating hormone may occur. The carcinoid syndromes are described elsewhere.

Neuromyopathies

The term carcinomatous neuropathy is used to describe those abnormalities of the central nervous system, the peripheral nerves, the muscles, and the autonomic nervous system that occur in association with malignancy. These disorders can be subdivided as follows: myopathies (polymyositis, myasthenia, and dermatomyositis), neuropathies (sensory and mixed sensorimotor, encephalopathy, and myelopathy). Toxic, infective, nutritional, and autoimmune causes have been suggested, but none has been fully substantiated. Neuromyopathies respond variably following treatment of the primary tumour by surgery, radiotherapy, or chemotherapy.

Most neuromyopathies are not tumour-cell-type specific, except for the Lambert–Eaton syndrome seen occasionally in small-cell lung cancer patients. This often precedes the appearance of the tumour by up to 15 months, and is characterized by proximal muscle weakness, depressed tendon reflexes (which often return following repetitive exercise), autonomic features, and difficulty with swallowing. There appears to be an association with HLA-B8 and the IgG heavy-chain allotype

GM2. Prednisolone and 3–4 amidopyridine 10 to 20 mg four times daily are used for treatment.

Finger clubbing and hypertrophic pulmonary osteoarthropathy

Finger clubbing accompanies a variety of intrathoracic disorders. Gross clubbing is readily recognizable; its early presence may best be demonstrated by the ability to rock the nail on its abnormally spongy bed. Clubbing of the toes is usually present also. Its incidence in lung cancer has variously been reported as being between 10 and 30 per cent. Clubbing may disappear after resection of tumour.

Hypertrophic pulmonary osteoarthropathy, which may be preceded by finger clubbing alone, consists of periostitis, arthropathy, and usually gross finger clubbing. It is most commonly associated with lung tumours but is also very common in pleural tumours, preceding the diagnosis of tumour in about one-third of patients. It is much commoner in peripheral lesions and squamous tumours.

The long bones of the extremities are affected by a periosteal reaction resembling elm bark; the changes are symmetrical and affect mainly the distal ends of the bones of the forearms and shins, with the knees and elbows being involved less commonly. Synovial thickening and joint effusions are rare. The typical radiographic appearances are shown in Fig. 2. The affected areas are hot and painful and sometimes oedematous, making walking difficult. Removal of the tumour is followed by immediate regression, but symptoms recur if the tumour recurs. Vagotomy alone is sometimes effective, supporting the theory of a vagal mediation of increased blood flow as an aetiological factor in hypertrophic pulmonary osteoarthropathy.

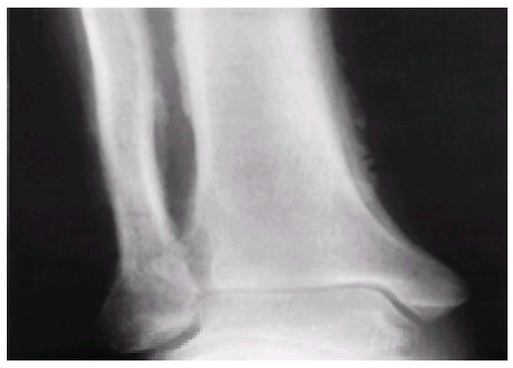

Fig. 2 Hypertrophic pulmonary osteoarthropathy showing persistent new bone formation.

Miscellaneous

The haematological effects of lung cancer are normally non-specific. Normocytic normochromic anaemia is the most common finding. Leucoerythroblastic anaemia denotes bone marrow infiltration and is particularly likely in small-cell lung cancer. Venous thrombosis and thrombophlebitis due to hypercoagulability are common complications of malignancy and may precede the detection of the underlying cancer; recurrent migratory phlebitis resistant to anticoagulation is an ominous feature. Marantic endocarditis is extremely rare, as are skin lesions such as acanthosis nigricans, dermatomyositis, hypertrichosis languinosa, and erythema gyratum repens. Rarely, the nephrotic syndrome due to membranous glomerulonephritis is encountered.

Staging and investigations

The investigations used to make the diagnosis and assess the stage of lung cancer will vary according to the presentation, the cell type, and the age and general condition of the patient.

The very rapid doubling time of small-cell lung cancer causes it to disseminate rapidly and widely, and at diagnosis is very rarely considered operable. However, the slower doubling times for squamous-cell cancers and adenocarcinomas, together with the relatively lesser tendency for the former to disseminate, makes surgery the best option whenever possible for the non-small-cell lung cancers. A precise anatomical staging classification was only applied to lung cancer in 1973 and immediately demonstrated that the prognosis of non-small-cell lung cancer depended strongly on the extent (or stage) of the disease. The introduction of the TNM staging system (T describing the primary tumour, N the extent of regional lymph node involvement, and M the absence or presence of metastases) encouraged an ordered assessment of investigations and selection of cases for surgery. On the basis of this experience, the system was modified in 1997 (Table 5).

Investigations

The following investigations form the basis for the diagnosis and staging of patients with lung cancer.

Radiological assessment

The value of the chest radiograph in the diagnosis and management of pulmonary neoplasm needs no emphasis. No initial examination is complete without a lateral film. Coned views of the ribs may help where rib invasion is suspected clinically.

The finding of a normal radiograph of the chest does not exclude bronchial carcinoma as patients presenting with haemoptysis and a normal chest radiograph are sometimes found to have a central tumour on bronchoscopy. The rounded or ovoid shadow of a peripheral tumour is described in greater detail below; these are sometimes cavitated (Fig. 3). The common appearance of a tumour arising from the main central airways (70 per cent of all cases) is enlargement of one or other hilum (Fig. 4). Even experienced observers sometimes have difficulty in deciding whether or not a hilar shadow is enlarged, and if there is any suspicion, investigation by bronchoscopy and/or CT should be pursued. Consolidation and collapse distal to the tumour may have occurred by the time that the patient presents, with the tumour itself often being obscured in the process. Collapse of the left lower lobe is often hard to identify (Fig. 5), as is a tumour situated behind the heart (Fig. 6). Apically located masses or superior sulcus tumours (Pancoast tumours) may be misdiagnosed as pleural caps, and often have a long history of pain in the distribution of the brachial nerve roots. Loss of the head of the first, second, or third rib is not unusual (Fig. 7).

Fig. 3 Cavitating peripheral squamous-cell carcinoma.

Fig. 4 Enlarged right hilum. Bronchoscopy revealed a tumour in the right intermediate bronchus.

Fig. 5 Collapsed left lower lobe showing loss of the medial third of the left diaphragm.

Fig. 6 Squamous-cell carcinoma lying behind the heart in the left lower lobe.

Fig. 7 Huge apical tumour with destruction of posterior parts of the second and third ribs.

The mediastinum may be widened by enlarged nodes. Involvement of the phrenic nerve may lead to paralysis and elevation of the hemidiaphragm, which then moves paradoxically on sniffing. Tumour spreading to the pleura causes effusion, but such an abnormality may be secondary to infection beyond obstruction caused by a central tumour. The ribs and spine should be carefully examined for the presence of metastasis. Spread of tumour from mediastinal nodes peripherally along the lymphatics gives the appearance characteristic of lymphangitis carcinomatosa—bilateral hilar enlargement with streaky shadows fanning out into the lung fields on either side. Rarely, localized obstructive emphysema may be observed.

Sputum cytology

Cytological examination of sputum is a very useful, non-invasive test for the diagnosis of malignant pulmonary disease. The yield increases according to the number of specimens examined, and three consecutive morning specimens should be submitted in the first instance. The yield increases with the number of specimens examined, rising to 85 per cent after four samples in a study of those in whom a diagnosis of lung cancer was made. The positive incidence is lower with tumours less than 2 cm in diameter (40 per cent) and higher with larger masses (60 per cent). Central tumour yields a higher proportion of positive results (60 per cent) than peripheral lesions (48 per cent).

Bronchoscopy

Bronchoscopy, which is described in detail in Chapter xxx, is frequently the definitive diagnostic method in lung cancer. About 70 per cent of all lung cancers arise in a main bronchus, lobar, first, or second generation subsequent division, and will be visible and within biopsy or cytology brush range. Bronchoscopy also yields valuable information regarding suitability for surgical resection. Attempts to resect are ill-advised if the main carina is obviously involved, or unequivocally broad with splaying of the main bronchi and immobility on respiration, or where there is involvement of the trachea, unless confined to the right lateral wall. Histological confirmation is now obtainable in 85 to 90 per cent of bronchoscopically visible lesions.

Transbronchial biopsy

Transbronchial biopsy via the fibreoptic bronchoscope is rarely used for peripheral tumours. It remains useful for more diffuse lesions such as may be seen in adenocarcinoma, bronchoalveolar-cell carcinoma, and lymphangitis carcinomatosis. However, transthoracic needle biopsy under imaging guidance has largely

replaced this technique.

Percutaneous needle biopsy

Percutaneous needle biopsy may be carried out using a variety of cutting needles to obtain a core of tissue for both histology and cytology. The procedure should be performed under fluoroscopic, CT, or ultrasound control, and is best avoided in patients with poor respiratory function or with bleeding diatheses. Positive yields as high as 90 per cent have been reported. Cytological samples remain the least satisfactory for cell type specificity. It is a useful diagnostic method in patients for whom exploratory thoracotomy may be hazardous, or in attempts to determine whether a solid mass is a primary, secondary, or benign tumour. Pneumothorax occurs following about 25 per cent of procedures, with some 5 per cent requiring a chest drain. Small haemoptyses are a common complication.

Thoracoscopy

Visualization of the parietal and visceral pleura plays an important part in the diagnosis of effusions and pleural tumours. Biopsy of lesions can be carried out under direct vision, and absence of pleural tumour is important in decisions about resectability of a lung tumour. Thoracoscopy is inadvisable in the absence of effusion or pneumothorax, and is unsatisfactory in the presence of empyema or gross haemothorax. However, in otherwise operable tumours with a pleural effusion that is not bloodstained and without positive cytology or pleural biopsy, thoracoscopy may be a useful next step in determining operability. Video-assisted thoracoscopy (VATS) has extended this technique and will also permit inspection and sampling of suspicious mediastinal lymph nodes.

Computed axial tomography (CT)

Thoracic CT scanning is important in the staging of lung cancer. It can identify the site, size, and extension of the primary tumour far more clearly than conventional radiology. It also frequently identifies mediastinal lymphadenopathy when posteroanterior and lateral chest radiographs fail to show any abnormality. Mediastinal lymphadenopathy on CT is arbitrarily taken to be pathological by most centres if the glands are greater than 1.0 cm in transverse diameter. However, previous infective conditions such as tuberculosis or an associated distal pneumonia can cause appearances identical with that of malignant enlargement. Thus positive CT scans of the mediastinum must be confirmed preoperatively by mediastinal lymph node biopsy (mediastinoscopy) to confirm tumour involvement.

Another potential advantage of CT is its ability to detect tumour invasion of the surrounding pleura and chest wall, in addition to the mediastinum itself. However, not all tumours with CT evidence of invasion prove unresectable, and if possible invasion of the mediastinum or chest wall appears the only contraindication to resection, then thoracotomy should be performed.

The predictive value of a negative CT is of the order of 90 to 95 per cent, and in such cases a mediastinoscopy can be omitted before thoracotomy. However, microscopic invasion of normal-sized mediastinal nodes is increasingly reported in patients with adenocarcinoma of the lung. Perhaps patients with this cell type should undergo mediastinoscopy routinely, irrespective of whether the mediastinal lymph nodes appear normal in size on a staging CT scan.

Positron emission tomography (PET) scanning

PET scanning shows promise, particularly in identifying the nature of a solitary pulmonary nodule, showing uptake in mediastinal nodes involved by tumour, and in assessing extrathoracic spread. When used in addition to CT, unsuspected metastases have been identified in up to 30 per cent of cases, changing management in up to 40 per cent of cases. The procedure is expensive and has limited availability but might become the final staging test where CT scans fail to show evidence of metastatic disease.

Lung function testing

The ability to climb one flight of stairs without breathlessness has been claimed to be a very good indication of fitness for resection, but formal evaluation of lung function is essential in all patients for whom surgery is being considered. Simple spirometry is usually adequate, but it may be necessary to evaluate exercise capability in a more sophisticated manner. Pneumonectomy should probably not be undertaken if the patient cannot sustain a forced expiratory volume in 1 s (FEV1) of more than 1.2 litres, bearing in mind that patients who have coexistent chronic obstructive airflow disease may not sustain their usual value during exacerbations. In borderline cases, the risk of resection is a matter requiring careful judgement based on estimation of maximum tolerable resection and assessment of the functional integrity of the non-tumour-bearing lung. A combination of pulmonary function tests, including the 6-min walking test and regional function studies using xenon-133, may be used in patients with borderline function. However, as lung cancer is such a serious disease, consideration may sometimes have to be given to carrying out resection in patients whose physical performance defies the results of pulmonary function tests.

Other investigations

In general, the ability to identify small metastatic deposits is as unsatisfactory for lung carcinomas as for other solid tumours. The available techniques are relatively crude, and this partially explains the high extrathoracic relapse rate following so-called 'curative' resections for non-small-cell lung cancer. In patients with no symptoms other than those caused by their primary tumour, imaging scans of brain, liver, and bones are unhelpful if there is no clinical evidence of neurological, hepatic, or bony disease and normal biochemistry. CT brain scans have a high accuracy in detailing cerebral metastases in patients with neurological symptoms. In patients with a palpable liver and/or abnormal liver function tests, a liver CT scan or ultrasound should be performed. CT scan of the upper abdomen identifies abnormalities of one or both adrenal glands in up to 10 per cent of patients considered for surgery. Fine-needle aspiration of the adrenal gland should be performed if this remains the only contraindication to pulmonary resection. Bone scans have a high false-positive rate due to Paget's disease, active arthritis, healing fractures, renal disease, and hyperparathyroidism. However, a bone scan should be ordered in patients with bone pain, local tenderness, or non-specific symptoms of weight loss or malaise.

Biopsy or cytology aspiration of enlarged lymph nodes and skin metastases should be carried out whenever indicated. If an isolated hepatic or bony lesion identified with isotope or CT scanning appears to be the only contraindication to surgery, this should be biopsied under radiological control.

The staging investigations for non-small-cell lung cancer are summarized in Fig. 8. The final procedure before thoracotomy is assessment of the mediastinum, since this may be involved in up to 50 per cent of patients with a peripheral, poorly differentiated tumour and in a much greater percentage of centrally occurring lesions. If CT scanning is normal, the surgeon can proceed directly to thoracotomy. If the CT scan is abnormal or is not available, mediastinal exploration should be performed first. Cervical mediastinoscopy in patients who appear radiologically operable on conventional films will demonstrate mediastinal lymph node involvement in 10 to 15 per cent of cases considered for surgery. Left anterior mediastinotomy can provide further information. This involves an approach through the bed of the second left costal cartilage to palpate glands draining tumours from the left upper lobe.

Fig. 8 Preoperative staging of non-small-cell lung cancer: +ve, positive; –ve negative; PTNM staging, postsurgical pathological staging. (Reproduced with permission from Spiro and Goldstraw, 1984.)

Treatment and prognosis of non-small cell lung cancer

Surgery

Surgery remains the single modality most likely to be curative in non-small-cell lung cancer. Prior to surgery the patient should have been carefully staged (Fig. 8), and the chances of long-term survival will be greatly influenced by this. All patients with stage IIIB disease (Table 5) should be rejected for thoracotomy, but those with stage I, II, and some with IIIA disease can be resected. In general, patients with squamous-cell carcinomas have higher 5 and 10-year survival rates than those with adenocarcinoma and large-cell carcinomas, and the more differentiated the tumour the better is the prognosis. Table 6 summarizes survival data at 5 years for preoperatively staged non-small-cell lung cancer. Clearly, small peripheral lesions with no nodal disease fare best (up to 70 per cent survival at 5 years), but the survival rate decreases with both size of tumour and increasing involvement of hilar nodes.

In all, approximately 20 per cent of patients who present with non-small-cell lung cancer eventually come to thoracotomy. Most of the others are excluded almost immediately because of clinically evident metastatic disease, radiological or bronchoscopic evidence of inoperability, too advanced an age to withstand surgery, significant associated other illnesses, or inadequate lung function. Of those having a 'curative' resection, the overall survival rate at 5 years is approximately 25 per cent and at 10 years is 16 to 18 per cent. Death from local or distant recurrence of the tumour is equally probable, highlighting the inadequacies of current staging techniques. However, the careful application of the TNM system and the advent of more sophisticated scanning equipment may lead to improvement.

Only very rarely is there an indication for palliative surgery, and resection should not be considered in the presence of intrathoracic or distant metastasis. There are diametrically opposed views as to whether surgery should be undertaken in Pancoast tumours.

Advanced age is not a contraindication to surgery. Patients over 70 years of age appear to tolerate lobectomy as well as younger patients, although the mortality for pneumonectomy (8–10 per cent) is double that of those under 70. There is no evidence that tumours grow more slowly in the elderly, and therefore the disease is as likely to be the terminal event in the aged as in younger patients. Hence resection should be encouraged in fit patients. Smokers should be persuaded to stop smoking before thoracotomy; continued smoking increases perioperative complications.

Thoracoscopic resection of peripheral masses is currently reserved for those with inadequate lung function for lobectomy, as hilar and mediastinal node evaluation and dissection is not always possible. The cure rates for segmentectomy by video-assisted thoracic surgery is less good than by open thoracotomy and lobectomy with lymph node dissection.

Radiotherapy

Patients who are excluded from surgery because of adverse prognostic factors, advanced stage of tumour, or other coincidental disease constitute the largest group treated with radiotherapy. Although the usual aim of radiotherapy will be palliative, there will be a small group of patients in whom more aggressive therapy will be used in the hope of cure, or at least long-term survival, particularly in those who have refused surgery. Radiotherapy for lung cancer is limited by the comparative radiosensitivity of three critical normal tissues likely to be included in the radiation beam: normal lung, spinal cord, and heart, each of which has a critical tolerance dose. Increased radiation dose leads to greater killing of tumour cells but may produce unwanted damage to normal cells. Radiation dose must be expressed not only in terms of total dose but also numbers of fractions and overall time. There is no clear evidence for an optimum radiation dose, but doses of 5000 to 6000 rad (50–60 Gy) in 5 to 6 weeks are appropriate; higher doses will be associated with unacceptable morbidity.

The role of radiotherapy

Alternative to surgery

In some patients with a technically resectable tumour, there may be medical contraindications for resection or the patient may refuse surgery. In general, the results of radical radiotherapy in these patients are inferior to the 5-year survival following surgery. The best result for radiotherapy was a 5-year survival rate of 22 per cent for peripheral squamous-cell cancers, but other series post a 5-year survival rate of 6 per cent.

Preoperative radiotherapy

Preoperative radiotherapy has been attempted in a few uncontrolled studies, but there is no evidence that this approach improves survival.

Postoperative radiotherapy

A recent meta-analysis has shown no benefit from postoperative radiotherapy for stage I and II disease. Any value in stage IIIA disease with nodal involvement is not as clear, but benefit is likely to be small.

Radical radiotherapy for locally inoperable disease

In otherwise fit patients with small-volume intrathoracic disease which is not resectable, usually because of mediastinal involvement, it is common practice to attempt to cure with radiotherapy. Results with daily single fractions are disappointing, even with doses of up to 60 Gy, with 5-year survival rates ranging from 5 to 17 per cent.

Recently, continuous hyperfractionated accelerated radiotherapy (CHART), with three fractions a day for 12 consecutive days to a total of 54 Gy have been compared to conventional radiotherapy in non-small cell lung cancer. CHART gave an absolute improvement in 2-year survival from 20 per cent to 29 per cent, with the greatest benefit (14 per cent absolute improvement) in squamous cell cancers. This appears a real advance in the provision of radiotherapy for locally advanced, inoperable tumours.

Palliation

The value of radiotherapy in palliating certain symptoms is beyond dispute. Haemoptysis and cough, two of the most distressing symptoms, can be controlled by radiotherapy in up to 80 per cent of cases. Administration of single fractions (each of 8.5 Gy, 1 week apart) appears adequate. Dyspnoea from bronchial obstruction and dysphagia are relieved in the majority of cases. The syndrome of superior vena caval obstruction is relieved in about 80 per cent of sufferers, but usually requires a more conventional course of five to ten fractions of radiotherapy. Pain from bone secondaries can be relieved in more than 50 per cent of sufferers by a single fraction of 8 Gy, often given at the same time as a clinic visit. Brain metastases generally respond poorly to radiotherapy. A 48-h trial of dexamethasone, 4 mg orally four times daily, is recommended as initial management. If a worthwhile response follows the resolution of the oedema surrounding the metastases, then radiotherapy will consolidate this gain. The steroids should then be rapidly withdrawn on completion of radiotherapy. Spinal cord compression is a relatively common occurrence associated with vertebral body metastatic disease. Pain and bony tenderness often precede it and may be helpful in localizing the lesion. Responses to radiotherapy are usually incomplete and disappointing, often because of interruption of the vascular supply to the spinal cord by the tumour.

Chemotherapy

Several cytotoxic agents show activity against non-small cell lung cancer, but much less frequently than with small-cell tumours (Table 7). However, combination chemotherapy can achieve impressive response rates; partial responses in 50 per cent of patients with locally advanced disease and in 35 per cent of those with advanced extrathoracic disease have been reported. The most active regimens include: cisplatin, ifosfamide, and mitomycin; or cisplatin, mitomycin, and vindesine. A meta-analysis of all controlled studies randomizing patients to receive or not receive chemotherapy in addition to surgery, radiotherapy, or to best supportive care was published in 1995. This suggested a 5 per cent advantage for the addition of chemotherapy to surgery (confidence intervals –1 to 7 per cent), a smaller non-significant advantage for the addition of chemotherapy to radiotherapy, and a 10 per cent improvement in survival at 1 year for the addition of chemotherapy to best supportive care. Many institutions are attempting to confirm these encouraging data. In advanced disease, chemotherapy only confers a survival advantage of 6 to 8 weeks compared to best supportive care alone, making evaluation of effects on quality of life important. Thus chemotherapy in non-small cell lung cancer is still recommended to be given within clinical trials. The place of the newer agents such as carboplatin, vinorelbine, gemcitabine, the taxols, and campotethecins are being

evaluated.

Treatment and prognosis of small-cell lung cancer

Small cell lung cancer is separated from the other types of lung cancer because of its very different biological and clinical features. It has an explosive growth pattern so that the TNM staging classification makes no impact on prognosis or survival, almost certainly because careful staging puts most patients into the inoperable category and because small metastases remain undetected for a few months. However, simple staging has some prognostic impact and those with limited disease (tumour confined to one hemithorax and the ipsilateral supraclavicular fossa) fare better than those with extensive disease (involvement of any site outside the hemithorax). The life expectancy of those with untreated small-cell lung cancer is about 3.5 months for limited disease and 6 weeks for extensive disease.

Prognostic factors

Multivariate analyses of large patient populations show that routine biochemical values such as serum sodium, albumin, and alkaline phosphatase allow separation of prognostic subgroups. In addition, performance status and extent of disease are important influences. For instance a good performance status and normal biochemical values (i.e. a good prognostic category) has a 2-year survival rate of 20 per cent, yet a correspondingly low performance status with one or more abnormal biochemical parameters (poor prognosis) has virtually no 2 year survivors (Fig. 9). Women tend to do better than men and those under 60 better than those over 60 years of age. These factors are helpful both for stratification within clinical studies and for identifying those patients likely to do well with chemotherapy and those for whom intensive toxic chemotherapy would appear inappropriate. Survival beyond 5 years (cure) is achieved in 4 to 12 per cent of patients with limited disease and in hardly anyone with extensive disease at diagnosis. Most studies of long-term survival report late deaths due to other cancers, including non-small-cell lung cancers in up to 30 per cent of these long-term survivors.

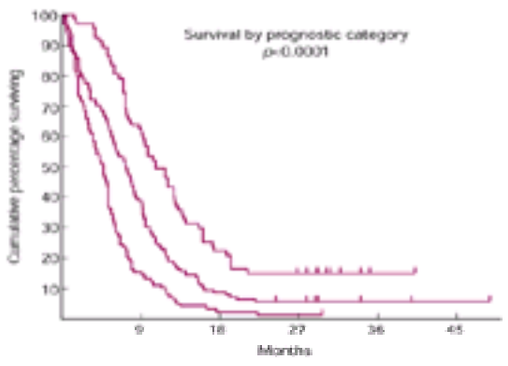

Fig. 9 The effect of prognostic grouping in survival in small-cell lung cancer. The upper curve represents good prognosis patients; the lower two curves are intermediate and poor prognosis groups respectively.

Surgery

In small-cell lung cancer the very occasional patient can be surgically cured, usually those presenting with a peripheral tumour and no evidence of local spread or metastasis despite extensive staging investigations. These patients are rare, but nevertheless have a 5-year survival rate in the region of 30 to 40 per cent.

Radiotherapy

Radiotherapy has an important role in palliation of symptoms that may develop after relapse following chemotherapy. However, in approximately 15 per cent of patients with limited disease, there is a small but definite benefit when thoracic irradiation is added to combination chemotherapy. Chest irradiation also significantly decreases the rate of recurrence at the primary tumour site and in the mediastinum. A total dose of 40 to 50 Gy is usually given. The optimal timing of radiotherapy in relation to chemotherapy is not clear, but the trend is in favour of early radiotherapy given concurrently with chemotherapy. However, acute and late toxicity are both increased when radiotherapy and chemotherapy are given in combination.

Cranial irradiation

Cranial metastases are common. Ten per cent of patients in remission develop brain metastases as their first site of relapse. Prophylactic cranial irradiation given at the end of chemotherapy will delay the presentation of cerebral metastases and also reduce their overall incidence. However, there is no evidence of prolonged survival, and those who receive prophylactic cranial irradiation are at greater risk of late neurological complications—particularly psychometric and psychological impairment. However, the morbidity of cerebral metastases is so great that it seems helpful to attempt to prevent this socially disastrous form of relapse.

Chemotherapy

Small-cell lung cancer is much more sensitive to cytotoxic chemotherapy than the non-small-cell lung cancer tumours, with a much higher response rate for several cytotoxic drugs (Table 7). In the late 1970s, there was a very rapid improvement in median survival using combinations of three and four drugs, but responses have subsequently reached a plateau. Nevertheless, with modern combination cytotoxic treatment, which is usually given on an outpatient basis every 3 weeks, the median survival has been extended to 14 to 18 months for limited disease and to 9 to 12 months for extensive disease. Most combinations include etoposide, cisplatin, cyclophosphamide, doxorubicin, and vincristine. Modern regimens would be expected to achieve a complete response rate (i.e. disappearance of all measurable disease) in 40 to 50 per cent of cases and a partial response rate (greater than 50 per cent reduction in tumour bulk) in a further 40 per cent, giving a total response rate of 80 to 85 per cent. All these regimens have side-effects. Most patients will experience some nausea and vomiting, and alopecia is practically universal. Life-threatening septicaemia occurs in 1 to 4 per cent, but treatment-related deaths are uncommon.

Much effort has been applied during the last 5 years to improve the median and long-term survival of patients with small-cell lung cancer. In general, those patients in whom further progress is to be made are those who present with limited disease and a good performance status. Patients with extensive disease tend to have a universally bad prognosis and very few survive beyond 2 years. However, it seems that some metastatic sites (bone and bone marrow) are not as sinister as others (brain or liver) and the occasional patient with extensive disease does well with chemotherapy, but in general treatment is offered in this circumstance for palliation and not in the hope of cure. Studies assessing the quality of life in patients presenting with small-cell lung cancer have shown that over 70 per cent have important symptoms such as weight loss, malaise, bone pain, dyspnoea, and haemoptysis. The majority of these patients have extensive disease, but after 3 months of chemotherapy symptoms can be relieved in 60 to 70 per cent of sufferers, making chemotherapy worthwhile, with symptomatic benefits far outweighing the potential side-effects. Ten per cent of small-cell lung cancer patients present with superior vena caval obstruction: this responds as well as any presentation to chemotherapy.

Intensity of treatment

Intensifying the dosage or the frequency of administration of cytotoxic agents has been thoroughly explored without real benefit on median survival. Small advantages are occasionally seen, but these have to be balanced by the increased toxicity resulting from a more aggressive approach. Attempts to overcome or delay the emergence of cell resistance to chemotherapy have involved alternating combinations of drugs, but these more complicated regimens have not been rewarding either.

Duration of treatment

Toxicity of chemotherapy increases with the number of courses given. It is now apparent that most of the tumour response to chemotherapy occurs within the first two or three cycles. Studies attempting to minimize the duration of chemotherapy without adversely affecting survival have shown that six courses of combination

chemotherapy, that is a course every 3 weeks, is optimal, with no benefit from maintenance regimens.

General management of patients with lung cancer

There are certain complications which require specific measures to alleviate symptoms.

Patients who seem likely to survive for 6 months or more and who have vocal chord paralysis find considerable help from an injection of Teflon into the affected chord which restores voice production in a high percentage of cases and reduces the risk of aspiration. Occurrence of upper airway obstruction causing stridor, or obstruction of the lower major airways, in non-small-cell lung cancer patients is usually initially treated with radiotherapy. Should this complication recur or be unsuitable for radiotherapy, it can sometimes be treated by laser photocoagulation administered either via a fibreoptic bronchoscope or under general anaesthetic via a rigid instrument. Laser therapy for carcinoma of the bronchus is most suitable as a palliative treatment in central tumours occluding large airways. There are technical limitations to its application via the flexible bronchoscope, but removal of considerable quantities of tumour can be achieved in a single treatment session with the rigid instrument. Laser therapy is used predominantly for recurrence of tumour in the central airways, usually after radiotherapy has failed. Trials are in progress assessing the additional benefits of endobronchial radiotherapy (brachytherapy) using iridium or caesium wires delivered via the fibre optic bronchoscope. This procedure irradiates endobronchial tumour to a circumferential depth of about 1 cm, and will often produce a further remission. It is used where further external beam radiotherapy cannot be given because of the risk of exceeding normal tissue tolerance.

Infection distal to tumour requires antibiotic therapy and, where appropriate, oxygen therapy and bronchodilators. Severe, recurrent haemoptysis may be controlled by radiotherapy or laser.

Malignant pleural effusion recurs after aspiration unless the pleural space is obliterated. Chemical pleurodesis can be induced by intrapleural instillation of a number of agents or by the more invasive procedure of talc pleurodesis. Intrapleural tetracycline is most commonly used, giving successful pleurodesis in 50 to 70 per cent of patients. However, the increasing availability of thoracoscopy makes a talc pleurodesis preferable in all reasonably fit patients who can undergo a general anaesthetic.

Dexamethasone, 4 to 16 mg orally daily, may control the symptoms of brain metastasis and, if so, this should be consolidated with radiotherapy to prevent severe steroid-induced myopathy. Prednisolone, 20 mg orally daily, is often used to improve the sense of well-being, as are blood transfusion or hyperalimentation.

Terminal care is described in Section 30, but the importance of the combined support to the patient and the family given by the family doctor, palliative care medical and nursing staff, and hospice organizations, and the hospital team cannot be overemphasized.

Prevention

Lung cancer is an almost totally preventable disease and is very largely due to smoking, particularly of cigarettes. The strategy of any preventive measures must be based on the following observations. Firstly, that lung cancer is extremely rare in non-smokers. Secondly, that there is no threshold limit below which no effect is produced, although the risk increases proportionately to the amount smoked. Thirdly, that the benefit from stopping smoking is evident within 5 years. Fourthly, that the risk for an exsmoker at any given time after stopping is determined by the length of time he or she had smoked before stopping. Thus strenuous efforts must be made to persuade people not to start smoking, to establish more effective methods of enabling people to stop smoking, and to promote further research into effective methods of health education. The promotion of cigarettes with low tar, nicotine, and carbon monoxide contents may have made a small contribution to prevention, but low-tar cigarettes are not a substitute for giving up smoking. Penal taxation by governments may also help.

The identification of occupational hazards and implementation of appropriate measures to safeguard the health of employees are clearly important preventive measures, even although the number at risk is very small.

Prospective lung cancer screening programmes in males aged 45 years and above who smoke at least 20 cigarettes per day have been carried out using both chest radiography and pooled 3-day sputum analysis every 4 months. They are unlikely to form the basis of standard practice as there is no evidence that early detection is translated into increased cure rate.

Carcinoid tumours

The slow-growing intrabronchial lesions previously grouped under the heading of bronchial adenoma have now been reclassified into bronchial carcinoids, adenoid cystic tumours, and mucoepidermoid tumours. They are not related to cigarette smoking, and tend to be diagnosed at a younger age than carcinoma of the bronchus. True bronchial adenomas derived from bronchial glands are rare. These tumours were once thought to be benign, but they are potentially and often frankly malignant, being capable not only of destructive local growth but also of metastasis to regional lymph nodes in about one-third of patients and to distant organs, particularly liver and brain, in about 10 per cent. They are occasionally located in the trachea.

The most common symptoms are cough, haemoptysis, and recurrent pneumonia, although not infrequently the lesion is discovered on routine radiographic examination before symptoms develop. In the few cases that have extensive liver secondaries, there may be the classical symptom pattern of intermittent cyanotic flushings, intestinal cramps and diarrhoea, bronchoconstriction, and cardiovascular lesions. The radiographic appearances are those of a solitary nodule, pulmonary collapse, or obstructive hyperinflation. As the majority of the tumours occur in main stem or proximal portions of lobar bronchi, bronchoscopy is usually the definitive diagnostic measure. The tumour appears as a white or pink polypoid or lobulated mass, with the bronchial mucosa appearing to be intact. Biopsy may be followed by brisk haemoptysis.

Surgical resection is the treatment of choice. In the absence of regional spread or distant metastases 5-year survival prospects are excellent, but if there is involvement of regional nodes, survival rates fall to 70 per cent. Some aggressive carcinoid tumours carry a much worse prognosis. The mechanism and management of the general symptoms of the carcinoid syndrome are described in Chapter 14.8.

Further reading

Ahrendt SA, Chow JT, Xu LH, Yang SC, Eisenberger CF, *et al*. (1999). Molecular detection of tumor cells in bronchoalveolar lavage fluid from patients with early stage lung cancer. *Journal of the National Cancer Institute* **91,**332–9.

Bruske-Hohlfeld I *et al*. (2000). Occupational lung cancer risk for men in Germany: results from a proband case-control study. *American Journal of Epidemiology* **151,**384–95.

Carney DN (1992). Biology of small-cell lung cancer. *Lancet* **339,** 843–6.

Coggon D, Acheson ED (1983). Trends in lung cancer mortality. *Thorax* **38,** 721–3.

Goldstraw P (1992). The practice of cardiothoracic surgeons in the peri-operative staging of lung cancer. *Thorax* **47,** 1–2.

Hansen HH (1992). Management of small-cell lung cancer. *Lancet* **339,** 846–9.

Izbicki JR *et al.* (1992). Accuracy of computed tomographic scan and surgical assessment for staging of bronchial carcinoma. A prospective study. *Journal of Thoracic and Cardiovascular Surgery* **104,** 413–20.

Kaplan DK (1992). Mediastinal lymph node metastases in lung cancer: is size a valid criterion. *Thorax* **47,** 332–3.

Landreneau RJ *et al.* (1992). Thoracoscopic resection of 85 pulmonary lesions. *Annals of Thoracic Surgery* **54,** 415–19.

Mountain CF (1997). Revisions in the international system for staging lung cancer. *Chest* **111,** 1710–17.

Muers MF, Round CE (1993). Palliation of symptoms in non-small cell lung cancer. *Thorax* **48,** 339–43.

Pless-Mulloli T *et al*. (1998). Lung cancer, proximity to industry, and poverty in northeast England. *Environmental Health Perspectives* **106,** 189–96.

Saunders M *et al*. on behalf of the CHART Steering Committee (1997). Continuous hyperfractionated accelerated radiotherapy (CHART) versus conventional radiotherapy in non-small-cell lung cancer: a randomised multicentre trial. *Lancet* **350**, 161–5.

Scagliotti GV (1995). Symptoms and signs and staging of lung cancer. In: SG Spiro, ed. *Carcinoma of the lung*, European Respiratory Monograph, Vol. 1, No.1, pp. 91–137. European Respiratory Society Journals, Sheffield, UK.

Spiro SG, Goldstraw P (1984). The staging of lung cancer. *Thorax* **39,** 401–7.

Thatcher N, Ranson M, Lee SM, Niven R, Anderson A (1995). Chemotherapy in non-small cell lung cancer. *Annals of Oncology* **6** (Suppl. 1), S83–95.

Tockman MS *et al*. (1997). Prospective detection of preclinical lung cancer: results from two studies of heterogeneous nuclear riboprotein A2/B1 over-expression. *Clinical Cancer Research* **3,** 2237–46.

Wells FC, Kendall SWH (1992). Thoracoscopy: the dawn of a new age. *Respiratory Medicine* **86,** 365–6.

17.14.2 Pulmonary metastases

S. G. Spiro

Further reading

Malignant metastasis to the lung may present as a solitary enlarging nodule, as multiple nodules, or with diffuse lymphatic involvement.

Solitary metastasis represents some 10 per cent of round lesions in general, but some 70 per cent of round lesions in patients with a known malignancy. Colorectal cancer is reported to be the commonest tumour of origin. Diagnosis can usually be secured by percutaneous CT-guided biopsy. In rare cases, surgical excision may prolong survival or result in cure, depending on the state of the primary tumour and the likelihood of other occult metastases. In general, the longer the interval between resection of the primary tumour and the appearance of the metastases the better the prognosis.

Multiple metastases range enormously in size and number from 'cannon balls' to miliary shadowing, and may be accompanied by hilar lymphadenopathy or pleural effusion. Breast, colon, renal, and lung primaries are probably the commonest underlying tumours, but other tumours amenable to chemotherapy, such as testicular cancer and choriocarcinoma, and also sarcomas, occur. Diagnosis may be achieved by cytology or histology on various samples from the pleura or lung and can occasionally be made from cytology on expectoration or induced sputum. Tumours that are suitable for chemotherapy (e.g. choriocarcinoma) or endocrine manipulation (e.g. breast) need to be recognized. Solitary or multiple Kaposi's sarcoma is a feature of AIDS, and can involve the bronchi and pleura as well as lung tissue.

Lymphangitis carcinomatosa is most commonly due to breast and primary lung tumours (usually adenocarcinomas). Patients can be asymptomatic when the disease is first suspected on the basis of a radiograph showing diffusely increased interstitial markings accompanied by Kerley B lines, hilar lymphadenopathy, or pleural effusion. Diagnosis may be established by cytology from sputum or pleural fluid, but often requires bronchoscopic or transbronchial lung biopsy. Later, progressive and severe breathlessness with hypoxaemia often develops, and requires vigorous palliative relief with opiate and oxygen administration.

Occasionally metastasis, presenting as haemoptysis, may be confined to a bronchus and not visible on a plain chest radiograph. Renal carcinoma and malignant melanoma are recorded causes. Diagnosis requires bronchoscopy, and radiotherapy is usually effective in controlling the haemoptysis.

Further reading

Gephardt GN (1981). Malignant melanoma of the bronchus. *Human Pathology* **12**, 671–3.

Ishida T *et al*. (1992). Metastatic lung tumours and extended indications for surgery. *International Surgery* **77**, 173–7.

Lower EE, Baughman RP (1992). Pulmonary lymphangitis metastasis from breast cancer. Lymphocytic alveolitis is associated with favourable prognosis. *Chest* **103**, 1113–17.

Ognibene FP, Masur H, Rogers P (1985). Kaposi's carcinoma causing pulmonary infiltrates and respiratory failure in AIDS. *Annals of Internal Medicine* **102**, 471–5.

Stewart JR *et al*. (1992). Twenty years' experience with pulmonary metastasectomy. *American Surgeon* **58**, 100–3.

17.14.3Pleural tumours

M. K. Benson

Primary pleural tumours are relatively rare, although malignant mesothelioma has received much attention because of its increasing incidence and association with asbestos exposure. Pleural plaques are also associated with asbestos exposure but should not be regarded as tumours since they simply represent local areas of fibrocollagenous thickening. The classical benign tumour of the pleura is a fibrous mesothelioma (pleural fibroma).

By contrast, pleural involvement by metastatic disease is very common. It can occur with most carcinomas, but is particularly associated with primary tumours arising in the lung, breast, colon, and ovary. Malignant lymphomas can also present with pleural involvement. Tumours arising in adjacent structures, such as diaphragm and chest wall, may also involve the pleura, and both benign and malignant tumours can originate from muscle, adipose tissue, nerves, blood vessels, and bony thorax. All are rare, and the diversity of sites and types of tumour results in a variety of clinical presentations. Radiographic techniques can help to demonstrate the site and nature of the tumour, but diagnosis is usually established on biopsy.

Benign tumours

Benign fibrous mesothelioma

These tumours are rare but can occur in virtually any age group. They bear no relationship to the development of malignant mesothelioma and are not associated with exposure to asbestos or other industrial pollutant. They originate from a pedicle, usually from the visceral pleura. Macroscopically they are firm, lobulated, and well encapsulated. The cut surface is white or grey and can have a whorled appearance. They vary in size, but can on occasions be very large, weighing up to 2 or 3 kg.

The tumours are often discovered on routine chest radiology in otherwise asymptomatic individuals (Fig. 1). Large tumours can cause chest discomfort and breathlessness, presumably due to compression of adjacent lung. Spontaneous hypoglycaemia is an associated feature in a small proportion of patients.

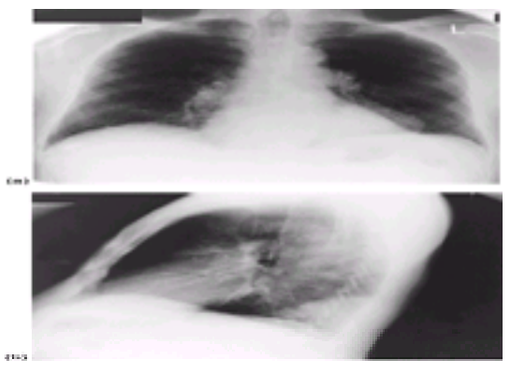

Fig. 1 Chest radiographs, posteroanterior (a) and lateral (b), showing a pleural-based nodule at the left base. The patient had no respiratory symptoms and resection confirmed this to be a benign fibrous mesothelioma.

Radiologically, it can be difficult to decide whether a nodule arises from the pleura or within adjacent lung. The differential diagnosis includes a localized area of pleural thickening, a whorled nodule, although this generally has a less clearly defined outline. The diagnosis is usually established only after surgical excision. Although a fibrous mesothelioma is benign, with no potential for metastatic spread, there is a possibility of local recurrence if the pedicle has not been completely excised.

Pleural plaques

These are areas of fibrocollagenous thickening. They produce no clinical symptoms and are usually detected on routine chest radiographs. They can be single or multiple and are best seen in oblique projection or using tomography. Although associated with asbestos exposure, they are entirely benign and should not be regarded as precursors to the development of a malignant mesothelioma.

Malignant tumours

Malignant mesothelioma

Malignant mesothelioma derives from mesothelial cells, most commonly in the pleura, but also in the peritoneum or (rarely) pericardium. Malignant mesothelioma arising from the pleura was first recognized in the 1950s, and during the 1960s much evidence accumulated indicating a strong link between the condition and exposure to asbestos. Asbestos is a collective term given to a group of silicate minerals, commercially useful because of their heat resistant properties, and widely used in industry for the past 100 years. In addition to asbestos exposure, there has been recent interest in the possible role of Simian virus 40 (SV40) in the aetiology of mesothelioma. This oncogenic virus was a contaminant of polio vaccines used in the 1950s, and SV40-like DNA sequences have been found in pleural mesotheliomas, although it is not clear what this means.

Epidemiology

There is overwhelming evidence that mesotheliomas are caused by asbestos, and the higher incidence in men indicates that most exposure is occupational rather than environmental. The risk is a function of the concentration of fibres and duration of exposure. Fibre type is also of relevance, since although most asbestos workers have been exposed to a mixture of fibres, there is good evidence that crocidolite (blue asbestos) is more hazardous than chrysotile (white asbestos). Exposure is greatest in those involved in mining or quarrying the material and those who handle the raw fibres. Significant exposure has also occurred in individuals employed in the manufacture and use of asbestos-containing products. Many workers engaged in the ship-building industry in the 1940s and 50s were exposed to asbestos, and more recently there has been widespread use in the building industry.

The increasing incidence of mesothelioma is a reflection of the long latent interval between exposure and the development of disease. It is rare for mesotheliomas to develop within 20 years of exposure and most patients were initially exposed 30 or more years before clinical presentation. In the United Kingdom, although the risk was first recognized in the 1960s, it was not until the mid-1970s that there was a significant reduction in exposure to asbestos. Current mesothelioma rates are a quantitive indication of previous population exposure. The increasing rates in the United Kingdom are predicted to continue for the next 20 years, with annual deaths increasing from the current levels of approximately 1100 to 2000 per annum. Men born between 1945 and 50 are those at greatest risk. Similar trends are also being

seen in much of Western Europe, although in North America mortality rates may have already peaked because of earlier measures to limit asbestos exposure.

Low levels of environmental contamination have been shown to result in a slightly increased risk of mesothelioma, although it has been estimated that the annual incidence in subjects with no clear history of asbestos exposure is about 1 per million. Endemic pleural mesothelioma has been reported from certain areas of central Turkey, Cyprus, and Greece. Locally mined zeolite and other environmental asbestos minerals are regarded as responsible.

Pathology

A mesothelioma can arise from either the visceral or parietal pleura, initially as a local mass, often associated with pleural effusion. As it progresses, there is gradual encasement of the lung and extension into adjacent structures including the chest wall, mediastinum, and pericardium. Macroscopically, the tumour is usually white and fibrous in texture, sometimes with areas of necrosis. Metastatic disease is relatively uncommon, but involvement of the contralateral lung and pleura, liver, and bone are recognized sites for secondary spread.

The histological diagnosis can be difficult as there are a variety of appearances, ranging from well-differentiated epithelial or sarcomatous patterns to undifferentiated forms. Even after biopsy there can be difficulty in distinguishing between a malignant mesothelioma and benign pleural disease on purely morphological grounds. A second problem is differentiation between mesothelioma and secondary adenocarcinoma. Histochemistry, immunohistochemistry, and electron microscopy can also provide useful information when there is diagnostic uncertainty.

Clinical presentation

The age of presentation is usually between 50 and 70, although incidence is increasing in older patients. There is a male predominance, reflecting the greater likelihood of previous occupational exposure, which should be sought with a careful lifetime occupational history. Symptoms due to local disease are mainly those of pain and breathlessness. Pain may be pleuritic in nature, but is often a dull ache due to direct involvement of the chest wall. Shortness of breath is usually associated with a pleural effusion, although as the tumour progresses is gradually encases the lung. Systemic symptoms include tiredness, anorexia, weight loss, fever, and occasionally drenching sweats. Finger clubbing has been recorded but is rare. Physical findings in the chest are those of a pleural effusion, but with advanced disease there is progressive reduction in chest wall movement. Direct extension through the chest wall can result in a palpable mass, and this may develop at the site of previous biopsy.

Investigations

A chest radiograph often demonstrates a pleural effusion, with tumour suspected if there is pleural thickening with a lobulated outline (Fig. 2). This can be more easily identified on CT scanning, which can also be used as a staging procedure. Magnetic resonance imaging gives similar information. The presence of benign pleural plaques offers evidence of previous asbestos exposure, although they are not in themselves precursors of malignant change.

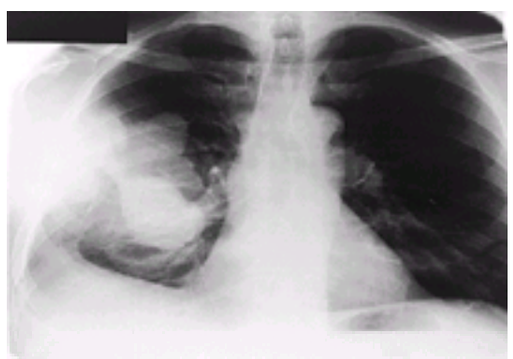

Fig. 2 Chest radiograph showing lobulated pleural thickening due to mesothelioma.

Pleural aspiration with cytological examination of pleural fluid yields a definitive diagnosis in up to a third of patients. There is further modest diagnostic yield from percutaneous needle biopsy. Samples obtained under direct vision, either at thoracoscopy or thoracotomy, have the highest diagnostic yield, with thoracoscopy being the investigation of choice. In some instances, it may be appropriate to make a diagnosis solely on the basis of clinical and radiological features, since even with histological confirmation there is unlikely to be any alteration in management. However, it is important to note that malignant mesothelioma is an industrially notifiable disease for which the patient and family can receive financial compensation.

Treatment and prognosis

A variety of treatments have been used for patients with mesothelioma, but rates of cure are uniformly disappointing. Surgical resection employing extrapleural pneumonectomy has its advocates, but reports are of highly selected patients with no satisfactory control group. There is significant perioperative morbidity and mortality, with median survival figures in the range of 10 to 20 months. Response to radiotherapy is also disappointing, although occasionally palliative relief of pain can be achieved when there has been direct extension into the chest wall. It may also prevent seeding along a biopsy tract. A variety of cytotoxic agents have been tried, either as single agents or in combination, but so far without any convincing success.

Relief of pain usually requires regular opiates, although nerve blocks may also be helpful for localized pain. Pleural aspiration and pleurodesis is of benefit for the relief of breathlessness due to recurrent pleural effusions.

The prognosis is poor with the median survival of approximately 12 months. A few patients seem to have fairly indolent disease and may survive for periods of up to 5 years.

Further reading

Aisner J (1995). Current approach to malignant mesothelioma of the pleura. *Chest* **107**, 332S–44S. Review of treatment options for malignant mesothelioma.

Anthony VB *et al*. (1992). NHLBI workshop summaries, pleural cell biology in health and disease. *American Review of Respiratory Disease* **145**, 1236–9. Workshop overview of pleural cell biology and responses to injury and repair.

Curran D *et al*. (1998). Prognostic factors in patients with pleural mesothelioma; the European Organisation for Research and Treatment of Cancer experience. *Journal of Clinical Oncology* **16**, 145–52. Prospective study of 204 patients with mesothelioma identifying factors influencing prognosis.

Hubbard R (1997). The aetiology of mesothelioma: are risk factors other than asbestos exposure important? *Thorax* **52**, 496–7. Brief review of possible factors in non-asbestos-related mesotheliomas.

Ong ST, Vogelzang NJ (1996). Chemotherapy in malignant pleural mesothelioma: a review. *Journal of Clinical Oncology* **14**, 1007–17. Review of 55 phase two clinical studies using both single agent and combination chemotherapy for the treatment of malignant mesothelioma.

Peto J *et al*. (1999). The European mesothelioma epidemic. *British Journal of Cancer* **79**, 666–72. Elegant analysis linking current and future trends in mesothelioma mortality to prior asbestos exposure.

Rebak J, Selikoff IJ (1992). Survival of asbestos insulation workers with mesothelioma. *British Journal of Industrial Medicine* **49**, 732–5. Epidemiological study in individuals working in the asbestos industry.

17.14.4 Mediastinal tumours and cysts

M. K. Benson

The mediastinum encompasses those structures within the thorax, excluding the lungs. The superior boundary is the thoracic inlet represented by a plane at the level of the first rib. The inferior boundary is the diaphragm. Traditionally, the mediastinum is subdivided into a number of compartments: superior and inferior, with the latter being subdivided into anterior, middle, and posterior divisions. However, there are no true anatomical boundaries, and structures in the superior mediastinum are in general contiguous with those inferiorly. Thus a more logical subdivision is simply into anterior, middle, and posterior compartments (Fig. 1 and Fig. 2). Such a division can help to compartmentalize complex anatomical arrangements and give some guide as to the most likely pathology occurring in any particular area.

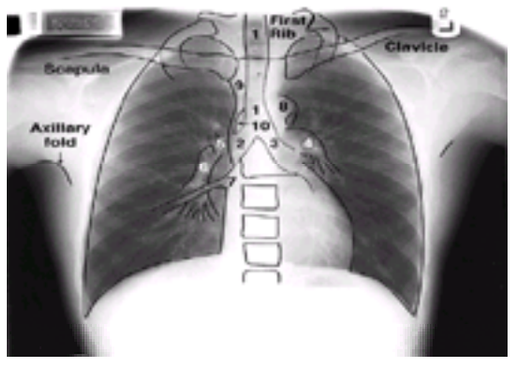

Fig. 1 Posteroanterior chest radiograph with diagrammatic overlay to illustrate normal mediastinal structures: (1) trachea, (2) right main bronchus, (3) left main bronchus, (4) left main pulmonary artery, (5) right upper lobe pulmonary vein, (6) right interlobar artery, (7) right lower and middle lobe vein, (8) aortic knuckle, (9) superior vena cava, (10) azygos vein.

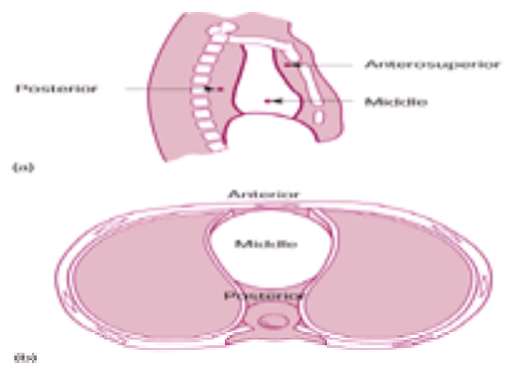

Fig. 2 A schematic representation of the mediastinal compartments: (a) lateral projection showing division into anterior (or anterosuperior), middle, and posterior compartments, (b) cross-sectional depiction.

The anterior mediastinum is bounded anteriorly by the sternum and posteriorly by the pericardium, aorta, and brachiocephalic vessels. It contains the remnant of the thymus gland, branches of the internal mammary artery, veins, and associated lymph nodes. The middle mediastinum contains the pericardium, ascending aorta and aortic arch, the vena cavae, the brachiocephalic vessels, and the pulmonary arteries and veins. It also encompasses the trachea and major bronchi with their associated lymph nodes, the phrenic nerves, and the vagus nerve. The posterior mediastinum is bounded anteriorly by the pericardium, laterally by the mediastinal pleura, and posteriorly by the vertebral bodies, including structures in the paravertebral gutter. It contains the descending thoracic aorta, oesophagus, azygos veins, thoracic duct, lymph nodes, and autonomic nerves.

Lymph nodes are common to all three compartments and knowledge of their anatomical relationships, together with sites of drainage, is helpful in interpreting an abnormal chest radiograph with mediastinal enlargement. The most important group of visceral nodes lie in the middle mediastinum and are predominantly subcarinal and paratracheal. Bronchopulmonary or hilar nodes are numerous but are not visible radiographically unless pathologically enlarged.

Diagnostic approach

The finding of a mediastinal abnormality on chest radiograph, whether or not accompanied by specific clinical features, is usually an indication for further investigation. Computed tomography (CT) provides accurate localization of mediastinal masses. It can define their relationship to and displacement of normal structures, and may be able to define lines of demarcation, particularly if there is adjacent fatty tissue. It is not ideal for determining the composition of any particular mass, although it can demonstrate heterogeneity or the presence of calcification. Contrast enhancement can be used to identify vascularity. Magnetic resonance imaging has relatively little to offer over and above CT scanning, the one exception being in the assessment of spinal tumours.

Fine-needle aspiration biopsy is valuable in the investigation of pulmonary masses, but is of more limited use in assessing those in the mediastinum. The presence of a cyst can be confirmed by aspiration of clear fluid. Anterior mediastinal masses can easily be approached percutaneously, although cytological examination alone may be insufficient for diagnosis. Anterior mediastinotomy allows open biopsy of such lesions. This is performed through an incision in the neck and allows inspection of structures surrounding the superior vena cava and trachea as far as the carina. It is particularly useful for obtaining lymph node biopsies when surgery for lung cancer is contemplated. Bronchoscopy is of limited value in the evaluation of mediastinal masses, except when there is a suspicion of a bronchial neoplasm or possible lymphadenopathy due to sarcoidosis. Neural tumours arising in the posterior mediastinum usually require surgical resection and there is little to be gained by preceding this with fine-needle aspiration.

Clinical features

General considerations

It is not surprising that the diversity of anatomical structures in the mediastinum is reflected in an equally diverse range of neoplastic, developmental, and inflammatory masses (Table 1). Whilst clinical symptoms and signs may give diagnostic clues, a significant proportion of mediastinal masses, particularly those which

are benign, tend to be asymptomatic and are usually detected on routine chest radiography.

Mediastinal masses in children are more likely to be malignant than those in adults. There have been a large number of studies documenting the relative frequency of different causes of primary mediastinal tumours and cysts: neurogenic and thymic tumours are the commonest (approximately 20 per cent each), followed by lymphoma, reduplication cysts, germ cell tumours, and thyroid masses.

Non-specific, constitutional symptoms such as fever or weight loss are more likely to occur with malignant tumours such as lymphomas or thymomas. The commonest symptoms are cough and chest pain, arising as a consequence of distortion of normal mediastinal anatomy. Compression of vital structures can also result in specific symptoms. Thus, tracheal or bronchial compression leads to breathlessness with stridor or wheeze; oesophageal narrowing results in dysphagia; and superior vena caval compression produces the characteristic features of facial and periorbital oedema, chemosis, and distended veins. Involvement of the recurrent laryngeal nerve results in hoarseness and a bovine cough; whilst this usually results from a malignant tumour, it can also occur with benign lesions such as aneurysm of the aortic arch. Involvement of the sympathetic chain as it emerges in the upper mediastinum is also likely to be due to malignant infiltration and results in Horner's syndrome with enophthalmos, miosis, ptosis, and unilateral facial anhidrosis. Compression of intercostal nerves can produce neuralgia, and intraspinal extension of tumours lead to long tract signs.

Anterior mediastinal masses

Thymus

The normal thymus is located in the superior portion of the anterior mediastinum. Its main function is the production of T lymphocytes. Radiographically, the normal thymus can only be seen in infancy and regression occurs during adolescence. Enlargement of the thymus is the commonest single cause of an anterior mediastinal mass. It can be due to the development of a thymoma, thymic hyperplasia, or a thymic cyst. In addition, the thymus can be a site of involvement by lymphoma, particularly Hodgkin's disease.

Thymomas are neoplastic proliferations of the thymus gland, with peak incidence in middle age. They are often benign, but can behave in a malignant fashion, with invasion of adjacent structures and distant metastases, in which case localized symptoms of chest pain and cough are more common. Systemic symptoms that can occur in association with thymic tumours are of particular interest. Myasthenia gravis is the commonest, occurring in some 30 per cent of patients. Other rare associations include red cell aplasia, hypogamma–globulinaemia, systemic lupus erythematosus, and polymyositis.

The majority of thymomas are slowly growing, lobulated masses that are well encapsulated. In these patients surgical resection can be expected to be curative. Local invasion is not common, but often precludes complete resection and recurrence is the rule. Adjuvant radiotherapy or chemotherapy has been used in such patients with debatable benefit.

Thymic cysts are uncommon. They can be unilocular or multilocular and usually contain straw-coloured fluid. The majority of patients are asymptomatic, but since cystic change can occur in some thymomas and in Hodgkin's disease, thorough cytological examination of the cyst's contents and wall must be carried out to exclude malignant disease.

Thymic lymphoma is fairly frequent, particularly in patients with Hodgkin's disease. The histological picture is usually of the nodular sclerosing variety, and the presence of mediastinal or hilar nodes should alert the clinician to the possibility of a lymphoma.

Germ cell tumours

This group of neoplasms includes tumours that are identical to certain testicular and ovarian neoplasms, and thought to be derived from primitive germ cells which have migrated to the mediastinum during oncogenesis.

Benign teratomas (dermoid cysts) consist of a disorganized mixture of ectodermal, mesodermal, and endodermal tissues, which can include skin, hair, cartilage, bone, epithelium, and neural tissue. They often contain cystic areas and CT appearances give a strong clue to the diagnosis. Unless there is a substantial contraindication to surgery, they should be excised to prevent further expansion and to exclude malignant change.

Malignant germ cell tumours are classically divided into seminomas and teratomas, although biopsy often reveals a spectrum of malignant tissue. Non-seminomatous germ cell tumours (malignant teratoma) can range from well-differentiated to trophoblastic. They are associated with elevated serum levels of b-human chorionic gonadotropin and a-fetoprotein, which can be used both diagnostically and to monitor response to treatment. Seminomas tend to be non-secretory. Both types of tumour are very malignant and invade adjacent mediastinal structures. They are not curable by surgery, but both are responsive to chemotherapy using cisplatin-based regimes, although results are disappointing when compared with their testicular counterparts, with response rates ranging from 30 to 70 per cent.

Thyroid masses

Retrosternal extension of an enlarged thyroid represents one the commoner causes of a mass in the superior mediastinum. The majority are multinodular benign goitres, arising in the neck and extending into the mediastinum through the thoracic inlet. They may contain cystic areas, sometimes with haemorrhage and areas of calcification. Radiographically, they have a sharply defined and often lobulated outline. Whilst they rarely cause symptoms, compression of the trachea at the thoracic inlet can result in respiratory distress and is an indication for surgical resection. Thyroid cancer can also involve the mediastinum, either by direct extension or by metastases to intrathoracic nodes.

Middle mediastinal masses

Lymphadenopathy

Enlarged nodes are not confined the middle mediastinum although this represents the commonest site of intrathoracic lymphadenopathy. Reactive changes occur in association with many pulmonary infections, although nodes are not grossly enlarged and are often undetected on plain chest radiograph. Gross lymphadenopathy is a feature of tuberculosis and also occurs in histoplasmosis. Other common causes of significant lymph node enlargement include metastatic carcinoma, lymphomas, and sarcoidosis.

Giant follicular lymph node hyperplasia (Castleman's disease) is rare but merits specific mention (Fig. 3). Its aetiology is unknown and it is not clear whether it represents a focus of lymphoid hyperplasia or has an infectious origin. The lesion consists of a vascular tumour with satellite lymphadenopathy. Two histological subgroups are described, (1) a more common hyaline vascular picture with lymphoid follicles and penetrating capillaries, and (2) a plasma cell type characterized by sheets of plasma cells between germinal centres. Both types can result in symptoms from local pressure, but the plasma cell type also causes systemic symptoms with fever, anaemia, and weight loss. There are no diagnostic radiographic features. The picture is simply one of a solitary mass and the diagnosis is usually made after surgical resection or biopsy. The condition is regarded as benign, but a small group of patients with multicentric disease have progressive hyperplasia, recurrent infections, and subsequently develop a frank lymphoma.

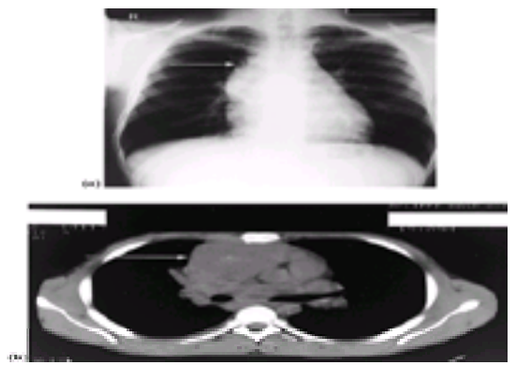

Fig. 3 Chest radiograph and CT scan showing large anterior mediastinal mass which on histology showed features of Castleman's disease.

Mediastinal cysts

Cysts within the mediastinum are a relatively common cause of a mediastinal mass. They can arise in association with the pericardium, bronchi, gut, or thoracic duct. The majority of patients are asymptomatic.

Pericardial cysts develop embryologically in relationship to the pericardium, although direct communication with the pericardial sac is rare. Radiographically they appear as smooth, clear, demarcated densities that can be mistaken for a pericardial fat pad or a hernia through the foramen of Morgagni. Aspiration reveals clear fluid. Surgical excision is not recommended.

Bronchogenic cysts arise in association with the major airways and are lined by respiratory epithelium. They may contain inspissated mucus. Local pressure on the trachea or bronchi can result in cough or wheezing. Occasionally the cysts communicate with the trachea, and when this is the case there is an increased tendency to recurrent infections. Surgical excision is recommended, particularly if there are associated symptoms (Fig. 4).

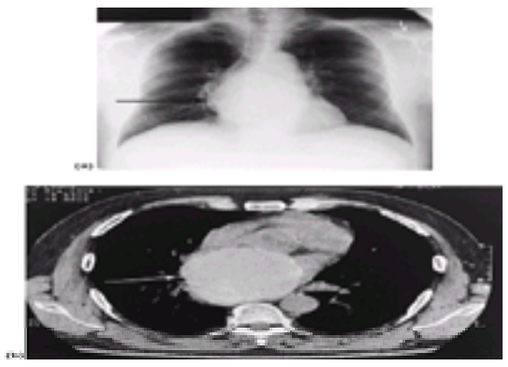

Fig. 4 Chest radiograph and CT scan showing a large mass in the mediastinum. This is a large bronchogenic cyst, present for 20 years, and finally removed when compression of the oesophagus resulted in dysphagia.

Reduplication cysts may also be associated with the oesophagus and can be lined by gastric or oesophageal mucosa.

Posterior mediastinal masses

Oesophageal lesions and aneurysms of the descending thoracic aorta can both result in abnormal shadows in the posterior mediastinum. Tumours, particularly those found in the paravertebral gutters, are likely to be neural in origin. Benign tumours tend to be asymptomatic, whilst malignant tumours cause pressure effects. Occasionally, spinal cord compression results from direct extension into the intravertebral foramen. Tumours arising from peripheral nerve cell sheaths include neurilemmoma (Schwannoma) and neurofibroma together with their malignant counterparts. Tumours of the autonomic chain include ganglioneuroma and neuroblastoma.

A neurilemmoma is the commonest neural tumour arising in the mediastinum. These are more common in middle age and can extend into the intravertebral foramen producing a dumb-bell appearance. Radiographically they can erode adjacent bone and CT scanning or MRI should be undertaken prior to surgical excision. Neurofibromata are also common. They may be solitary and the clinical and radiological features are very similar to those of a neurilemmoma. A significant proportion of patients will have neurofibromatosis. Surgical resection is recommended, in part because of the small risk of developing a malignant neurosarcoma **.** These tumours have a poor prognosis.

Ganglioneuroma arise from the autonomic plexus and are usually found close to the spine. Associated endocrine symptoms include hypertension, flushing, sweating, and diarrhoea. These tumours are often very large before they become clinically apparent. Prognosis is good after surgical resection. Ganglioneuroblastoma and neuroblastoma represent the malignant end of the spectrum and are predominantly tumours of infants and children. Neuroblastoma in particular is very invasive, with metastatic spread often established by the time of presentation.

Further reading

Adkins RB, Maples MD, Ainsworth J (1994). Primary malignant mediastinal tumours. *Annals of Thoracic Surgery* **38**, 648–59. Authors' experience of 38 patients presenting with mediastinal tumours together with a review of the literature.

Bower RJ, Kiesewetter WB (1977). Mediastinal masses in infants and children. *Archives of Surgery* **112**, 1003–9. Experience based on review of 93 infants and children presenting with mediastinal masses.

Davies RD Jr, Oldham HM Jr, Sabiston DC Jr (1987). Primary cysts and neoplasms of the mediastinum: recent changes in clinical presentation, methods of diagnosis, management and results. *Annals of Thoracic Surgery* **44**, 229–35. Report on series of 400 patients presenting with mediastinal tumours and cysts.

Hejna M, Haberl I, Raderer M (1999). Non surgical management of malignant thymoma. *Cancer* **85**, 1871–84. Systematic review of radiotherapy and chemotherapy in treatment of malignant thymoma.

Morrissey B *et al*. (1993). Percutaneous needle biopsy of the mediastinum: review of 94 procedures. *Thorax* **48**, 632–7. Sensitivity and specificity of percutaneous biopsy techniques in the diagnosis of mediastinal tumours.

Shields TW, Reynolds M (1988). Neurogenic tumours of the thorax. *Surgical Clinics of North America* **68**, 645–68. Systematic review of clinical presentation, treatment, and prognosis in patients with intrathoracic neurogenic tumours.

Thomas CR Jr, Wright CD, Loehrer PJ Sr (1999). Thymoma: state of the art. *Journal of Clinical Oncology* **17**, 2280–9. Review article on clinical presentation, pathology, and treatment of thymoma.

17.15 The genetics of lung diseases

J. M. Hopkin

Genetic factors are important in lung biology and disease, ranging from monogenic disorders, for example cystic fibrosis, to multifactorial effects, for example those underlying the syndrome of asthma, in which there are crucial interactions between heterogeneous genetic effects and environmental factors. Subtle genetic factors are being steadily uncovered, for example those that modulate the risk of respiratory infection such as tuberculosis.

The bronchial tree

Cystic fibrosis

Cystic fibrosis is the most common fatal, autosomal recessive disease in Caucasians, with an incidence of 1/1500 and carrier frequency of 1/22. This condition is discussed in detail in Chapter 17.10.

Immotile ciliary syndrome

The immotile ciliary syndrome, or primary ciliary dyskinesia, is a rare (incidence 1/30 000), genetically heterogeneous disorder in which inheritance is usually autosomal recessive. The ciliary abnormality, especially lack of dynein arms, can normally be observed on electron microscopy of a ciliated mucosal biopsy or in spermatozoa. Impaired ciliary function results in impaired mucus clearance and leads to bronchiectasis, sinusitis (often with absence of the frontal sinuses and ear disease), and infertility in males. Kartegener's syndrome is one clinical subgroup of the immotile cilia syndrome where there is also dextracardia or situs inversus as the result of impaired organ rotation in development. Multiple linkages have been found. Loss of function mutations in dyneia chain is one cause

Atopy and asthma

Atopy is an immune disorder in which there is allergy or hypersensitivity to common antigens, such as the house dust mite and pollens, and which results in muocsal inflammation. It is one leading cause of the asthma syndrome and is characterized by exuberant Th-2 immune mechanisms and excessive production of IgE to such common antigens (see Chapter 17.4.1).

Twin studies demonstrate the shared input of genetic and environmental factors. For example the concordance in monozygotic twins for allergy to any common inhaled antigen is 71 per cent, compared with 36 per cent in dizygotic twins. Genetic searches for linkage to either high IgE levels or clinical asthma have identified a whole set of linkages on many chromosomes, thus confirming the heterogeneous genetic input. The first genetic variant shown to have both epidemiological and functional links to atopy and asthma is within the cellular receptor for the Th-2 cytokine IL-4 (IL4Ra). Substitution of isoleucine for valine in the extracellular domain of the receptor (position 50, Ile50Val) associates strongly with atopic asthma in Japanese children. In transfection experiments, Ile50Val up-regulated cellular response to IL-4 challenge and increased secretion of IgE. An amino acid charge-changing variant of another key Th-2 signalling cytokine, IL-13, also predicts asthma and atopy in different populations. One possible benefit of unravelling the heterogeneous genetics of atopy and asthma is the tailoring of novel therapies to patients with genetic variants of specific proteins or pathways.

The pulmonary parenchyma

a_1-Antitrypsin deficiency

Emphysema is mainly the consequence of cigarette smoking, but a_1-antitrypsin deficiency is a substantial risk factor for the development of accelerated emphysema in cigarette smokers. This leads to the proteolysis hypothesis for emphysema in smokers, which proposes that proteases such as elastase are released from pulmonary neutrophils and macrophages 'excited' by cigarette smoke. The actions of elastase on supportive tissue in the lung are opposed by the serine protease inhibitor (serpin) a_1-antitrypsin. When a_1-antitrypsin is deficient, or when smoking is very heavy, then the balance falls in favour of tissue destruction and the development of emphysema. This condition is discussed elsewhere.

In deficient individuals, the most important practical point in management is absolute avoidance of cigarette smoking. Replacement of a_1-antitrypsin protein, now available as a molecular engineered product, can increase lung levels when administered by infusion or by aerosol inhalation—but trials showing major clinical impact are still awaited. In a family with an affected individual, the risk of a_1-antitrypsin deficiency in a further sibling is one-quarter: molecular prenatal diagnosis is possible. The risk to the children of carrier siblings is very small since the incidence of severe a_1-antitrypsin deficiency alleles in the population is low.

Miscellaneous rare disorders

Pulmonary alveolar proteinosis is a rare heterogeneous disorder, characterized by the accumulation of PAS positive proteinaceous material in the alveoli. It can be secondary to malignancy or infection, for example histoplasmosis. The infant form is usually fatal and in some cases is probably a genetic disease. In one family with two affected siblings there was absence of surfactant protein B and its mRNA in the alveoli but markedly increased amounts of surfactant protein C.

Rare, but striking, reports of familial aggregation of early-onset diffuse alveolitis and pulmonary fibrosis are recorded, but their molecular genetic origins are obscure. In pulmonary alveolar microlithiasis, there are multiple, minute calcifications in the alveoli producing a typical radiographic appearance. Affected sib-pairs are well described and the observation of consanguinity suggests a very rare autosomal recessive disorder of obscure molecular genetic origin.

General genetic syndromes with parenchymal lung involvement

A number of well-described genetic syndromes may display a pulmonary component.

Diffuse interstitial pulmonary infiltrates on chest radiograph are recorded in a number of 'in-born errors of metabolism'—such as Farber's lipogranulomatosis, Niemann–Pick's disease type A and Gaucher's disease types I and III. Lethal pulmonary involvement may occur in lysinuric protein intolerance. Niemann–Pick's disease is caused by mutations in the gene for acid sphingomyelinase and results in the development of abnormal 'Niemann–Pick cells' that are histocytes whose

cytoplasm is filled with lipid droplets or particles. Particularly severe pulmonary involvement in Niemann–Pick type B is accompanied by infiltration of these abnormal cells into the substance of the lung and also its lymphatics and vessels. In Gaucher's disease, a lysosomal glycolipid storage disorder characterized by the accumulation of glucosyl ceramide (glucocerebroside), there are mutations in the gene on chromosome 1 encoding the enzyme b-glucosidase. Although pulmonary involvement is rare, it is severe and progressive when present; pathology shows infiltration by Gaucher monocytes/macrophage cells with their characteristic tubular cytoplasmic inclusions.

In the phakomatoses, for example neurofibromatosis or tuberose sclerosis, pulmonary involvement may be observed with pulmonary fibrosis, bulla formation, or lyomyomatosis.

In Marfan's syndrome there is an increased risk of spontaneous pneumothorax. In some subjects, apical bullae are present; more rarely, congenital cystic lung disease may be found. There is also a clear clinical impression that many young individuals presenting with spontaneous pneumothorax are rather tall and thin, whilst not having Marfan's syndrome. They may ultimately prove to have some genetically determined, mild connective tissue disorder.

The immune system

The lung is constantly exposed to the risk of infection and there are well-recognized genetic syndromes that predispose to chest infection. Many genetic variants with subtle effects on the risk of infection, for example tuberculosis, are being discovered.

Antibody deficiency may be secondary (for example to protein loss in renal or bowel disorder) or primary. X-linked and autosomal recessive forms of the latter are described and make infection with encapsulated bacteria and mycoplasmas common and severe. Variable combinations of IgA and IgG subclass and IgM deficiency are recognized. The clinical picture may be of repeated pneumonias, with typical systemic upset, or of bronchiectasis with chronic sepsis.

Severe combined immunodeficiency is a genetically heterogeneous syndrome with profound functional deficiency of both cellular (T cell) and humeral (B cell) immunity. In the X-linked form of disease, most of the mutations are in the gamma chain of the cellular receptor for the cytokine interleukin-2. Recessive disease in most cases is due to mutations in the purine catabolic enzyme, adenosine deaminase, and genotherapy may become possible in this condition. In severe combined immunodeficiency, symptoms start in infancy with failure to thrive, diarrhoea due to parasitic and viral infections, and pneumonia due to *Pneumocystis carinii*, an organism that typically requires effective T-lymphocyte function for its control.

Loss of function mutations in the receptors for Il-12 and IFN-g signalling leads to impaired Th-1 immunity and predispose to disseminated BCG infection, and disease due to environmental mycobacteria of low pathogenicity.

Ataxia telangiectasia is a complex autosomal recessive disorder that includes variable immune deficiency, cerebellar ataxia, and a propensity to develop leukaemias; it results from mutations in ATM kinase. The Wiskcott–Aldrich syndrome includes combined immune deficiency, haemorrhage due to thrombocytopenia, autoimmune disorder, and a tendency to malignancy. The locus X-chromosome encodes for a protein (WASP) that is important in the actin-based cytosketon.

Chronic granulomatous disease is associated with recurrent pyogenic infection of the respiratory tract, skin, and lymphoid tissue by catalase-positive bacteria and fungi. It is genetically heterogeneous. Oxygen-dependent microbial killing is crucial in phagocytes and the process is mediated by a multicomponent NADPH oxidase; the four major oxidase components being encoded at different chromosome locations—1q, 7q, 16p, and Xp. The most frequent form is X-linked and due to mutation in the gene for the large subunit of cytochrome b. Another important property of the phagocyte is accumulation at the site of infection, which is dependent upon the cell's adhesive properties and mobility. In leucocyte adhesion deficiency, an autosomal recessive condition, there are mutations in the CD18 gene, which codes for the b2 integrin subunit.

The complement system plays an essential role in the propagation of inflammation and host defence. Deficiencies have been described for many of its components. Deficiency of C3, an autosomal recessive, increases susceptibility to encapsulated bacteria because of the deficiency of C3b-dependent opsonization.

The vascular system

Pulmonary embolism

Pulmonary embolism, from venous thrombosis, is a common and important pulmonary syndrome. Genetic deficiencies, causing thrombophilia underlie the disorder in some individuals, particularly those with early-onset of disease, unusual sites of venous thrombosis, and recurrent disease.

Miscellaneous rare syndromes

Pulmonary arterial venous fistulas can be single or multiple and may be sporadic or genetic disorders. They can occur as part of the Osler–Rendu–Weber (hereditary haemorrhagic telangiectasia) syndrome. This is a genetically heterogeneous autosomal dominant in which the pulmonary lesions are usually multiple and asymptomatic but can bleed and cause haemoptysis, breathlessness, and chest pain. There are a number of reports of primary pulmonary hypertension clustering in families. This is rare and its mechanism, if distinct from a thrombotic tendency, is unknown.

Neuromuscular disorders affecting pulmonary function

Many genetic syndromes of the neuromuscular system can cause secondary respiratory insufficiency or failure, with a tendency to recurring chest infection because of impaired coughing. The most notable are the muscular dystrophies (for example the X-linked recessive Duchenne muscular dystrophy), dominantly inherited myotonic dystrophy, and autosomal recessive acid maltase deficiency (type II glycogenosis). In acid maltase deficiency, diaphragmatic involvement is common and patients sometimes present with respiratory failure.

Pharmacogenetics

Idiosyncratic reactions to various therapeutic drugs are well recognized and cause a great range of pulmonary disorders. The response of patients to therapeutic agents administered for lung disease also varies in terms of both therapeutic benefit and toxicity. A number of factors underlie such behaviour—including age, state of nutrition, hepatic function, and renal function. In a significant proportion, there may be discrete underlying genetic causes, but relatively few have been well characterized.

The cytochromes P450 are a large family of haemoproteins that metabolize foreign chemicals, including therapeutic drugs and some carcinogens, as well as some endogenous compounds such as steroids. Genetic polymorphisms influence their function and hence the response to drugs; debrisoquine and phenytoin are well described. Genetic polymorphism is also recognized at the locus for NAT2 encoding *N*-acetyltransferase. These polymorphisms or variants of NAT2 influence the rate of acetylation and therefore detoxification of isoniazid; slow acetylators are at risk of drug-related neuropathy (due to accumulated isoniazid) or hepatitis (thought to be due to production of a toxic isoniazid metabolite through an alternative biochemical pathway). Slow acetylators therefore need lower doses of medication to achieve a therapeutic effect and to avoid toxicity. The genetic variants of NAT2 can now be recognized by direct and simple PCR based assays; they are most commonly found in Asian populations.

Tumour genetics

Many potent carcinogens are also potent mutagens, suggesting that malignant transformation is based on heritable mutations in somatic cells. Cigarette smoke is a particularly powerful mutagen, containing a great range of chemical mutagens including aromatic hydrocarbons, nitrosamines, and pyrrholized amino acids.

There are varying sites and types of somatic mutation underlying malignancy, including visible chromosomal rearrangements and discrete mutations at 'oncogene' loci encoding cellular growth factors and their receptors, and growth regulators. For example in both small-cell and non small-cell carcinomas of the bronchus somatic

mutations in the gene for P53, an important growth regulator, are well described. Mutations in the cellular oncogene K *-ras* are found in adenocarcinomas of the lung.

Germline mutations are also important in influencing risk of tumour development in a number of ways—for example through pre-existent mutations in an oncogene or through variant metabolism of carcinogens. Thus polymorphisms of P4501A1 (which metabolizes polycyclic aromatic hydrocarbons) and of P4502E1 (which metabolizes nitrosamines) are associated with increased risk of lung cancer in smokers.

Microbial genetics

Respiratory infection is of massive, world-wide importance. The molecular genetics of the diverse organisms are being defined and the essential genetic foundations for pathogenicity and response to antimicrobial drugs are being clarified.

In antibiotic resistance, a number of genetic mechanisms have been identified. Antibiotic inactivating enzymes are an important mechanism and involve b-lactams, macrolides, and chloramphenicol, for example b-lactamase production by *Haemophilus influenzae*. This kind of resistance can be transferred between bacteria by gene transfer on plasmids.

Change of target is another mechanism, for example mutation in bacterial cell wall peptidoglycans inhibits penicillin binding in *Streptococcus pneumoniae*. In one form of resistance of *M. tuberculosis* to isoniazid, there are mutations in the *InhA* gene (whose protein mediates mycolic acid metabolism and hence cell wall structure), resulting in impaired binding of isoniazid to the target enzyme.

A third mechanism involves bacterial mutations that limit antibiotic penetration through the cell wall; this is thought to occur in the lipoproteins of the cell wall of pseudomonas species and causes reduced permeability and hence resistance to b-lactams and aminoglycosides.

Detailing the molecular genetics of respiratory pathogens offers important practical gains. For example organism-specific DNA sequences can be used as powerful diagnostic tools whose specificity and sensitivity are impressive when linked to *in vitro* DNA amplification. Successful PCR diagnostic assays have been developed for a number of pulmonary pathogens, including *Pneumocystis carinii*. Rapid DNA diagnostics can be used also to identify specific mutations that predict antibiotic resistance. DNA vaccines may become very effective immunizers for intracellular pathogens such as the influenza virus or *M. tuberculosis*, since the microbial DNA is taken up by cells and, following incorporation and expression, presented 'more naturally' to the immune system.

Further reading

Bartsch H *et al*. (2000). Genetic polymorphism of CYP genes, alone or in combination risk modifier of tobacco-related cancers. *Cancer Epidemiology, Biomarkers and Prevention* **9**, 3–28.

Ciba Foundation (1997). *Antibiotic resistance: origins, evolution, selection and spread*. John Wiley and Sons, Chichester.

Shirakawa T *et al*. (2000). Atopy and asthma: genetic variants of Il-4 and Il-13 signalling. *Immunology Today* **21**, 60–4.

Stevenson FK, Rosenberg W (2001). DNA vaccination: a potential weapon against infection and cancer. *Vox Sanguis* **80**, 12–18.

Toloza EM, Roth JA, Swisher SG (2000). Molecular events in bronchogenic carcinoma and their implications for therapy. *Seminars in Surgery and Oncology* **18**, 91–9.

Winf C (2001). Do delta F508 have a selective advantage? *Genetic Research* **78**, 41–7.

17.16 Lung and heart–lung transplantation

K. McNeil

Introduction

The first successful heart–lung transplant in 1981 heralded lung transplantation as a viable therapeutic option for endstage cardiopulmonary and pulmonary lung disease. The successful introduction of single and bilateral lung transplantation in the late 1980s led to greater availability of donor organs, with a consequent expansion in the number of potential recipients. Over 12 000 lung and heart–lung procedures have been performed over the past 18 years, with steadily improving results. Patients can now realistically expect a survival of 5 or more years and there is every prospect that these results will continue to improve.

The lung allograft is unique in that it remains in contact with the external environment, exposing it directly to numerous potential infections and allergens, which predispose to many of the problems encountered immediately post-transplant and in the longer term.

The major obstacles faced by lung transplant programmes are the shortage of suitable donor organs, the need to find more selective (less toxic) methods of immunosuppression, and the means to either prevent or reduce the impact of chronic allograft dysfunction.

The transplant process

The five factors which constitute the transplant process are recipient selection, donor selection, donor/recipient matching and the surgical procedure, immediate post-operative care, and long term follow-up.

Recipient selection

From the 'pulmonary' point of view, recipient suitability is determined by two factors—the underlying disease indication and the severity of that disease. Virtually any endstage pulmonary disease is amenable to transplantation. Most suitable diseases are confined to the thorax, although there are a number of systemic diseases with pulmonary manifestations where carefully selected patients can be transplanted successfully.

The diseases fall into four main categories: septic lung diseases (cystic fibrosis, bronchiectasis), obstructive lung diseases (emphysema/chronic obstructive pulmonary disease, asthma, obliterative bronchiolitis), fibrotic lung diseases (cryptogenic fibrosing alveolitis, sarcoidosis, fibrosis related to drug reactions, occupational exposures, acute lung injury, etc.), and pulmonary vascular diseases (primary pulmonary hypertension, Eisenmenger's syndrome). In addition, a number of other conditions such as lymphangioleiomyomatosis, histiocytosis, and even bronchoalveolar cell carcinoma have been treated successfully with transplantation.

Patients are usually listed for transplantation when their survival is estimated to be less than 2 years and there are no further medical or alternative therapies available. For patients with cystic fibrosis, primary pulmonary hypertension, and cryptogenic fibrosing alveolitis, prognostic indices are available to guide appropriate timing for referral and listing for transplantation. By contrast, it has become apparent that the survival of patients with Eisenmenger's syndrome and emphysema is the same on the waiting list as following transplantation. In this setting, transplantation is performed primarily for quality of life issues. Table 1 summarizes referral recommendations based on the guidelines used at Papworth Hospital, United Kingdom.

Contraindications are well defined. Most are relative, and considered in the context of the patient's overall status. Older patients have a poorer outcome, particularly with the more extensive surgical procedures. Thus, for heart–lung transplantation the upper age limit is usually 55 years, whereas for single and bilateral transplants the limit is usually set at 60 years. Recipients well above these limits have, however, been transplanted successfully.

Well-controlled diabetes and low-dose steroid therapy are no longer considered exclusion criteria. Significant coexisting kidney or liver disease precludes isolated thoracic organ transplantation, but selected patients can be considered for combined thoracic and abdominal organ transplants. Pleural disease (previous pleurectomy, infection, etc.) must be considered carefully because of the increased risk of bleeding. Patients with antibiotic-resistant organisms such as *Burkholderia cepacia* are at increased risk of perioperative sepsis and death, and different units will have individual policies on their suitability for transplantation.

Active malignancy (excluding localized squamous and basal cell carcinomas of the skin), major psychosis, extrapulmonary infection, severe malnutrition, and significant extrathoracic organ dysfunction are considered absolute exclusion criteria. Mechanical ventilation is an absolute contraindication in some units, but this does not generally include non-invasive or ambulatory support.

Donor selection

As the shortage of donor organs increases, criteria for acceptance of donor organs has liberalized. All donors are brainstem dead and should be free of systemic infection or disease. Lung allograft donors are generally less than 55 years of age, although there is an increasing trend to accept organs from older donors with acceptable results. Lung allograft suitability is based on function (gas exchange and compliance) and appearance (macroscopic, bronchoscopic, and radiographic). Haemodynamic performance and the macroscopic appearance of the coronary arteries determine the cardiac status of the heart–lung donor. Echocardiography and angiography are sometimes used but are not routinely available at every donor hospital.

Following acceptance of the allograft, matching of the donor organ with a suitable recipient is based on ABO blood group and size. Size matching is determined by comparing predicted donor total lung capacity (based on height, age, and sex) with measured and/or predicted recipient total lung capacity. Perfect size matching is rarely achieved but significant (more than 10 per cent) oversizing should be avoided because of the resultant lung compression and atelectasis. Conversely, undersizing of 10 to 20 per cent is easily accommodated by the natural compliance of the lung(s).

Surgery

There are three basic options when replacing diseased lung tissue—single-lung transplantation, bilateral sequential single (double) lung transplantation, and heart–lung transplantation. The choice of procedure is determined by the recipient's underlying disease. Diseases involving both the heart and lungs such as Eisenmenger's syndrome and endstage primary pulmonary hypertension are usually treated with heart–lung transplantation. Septic lung diseases such as cystic fibrosis and bronchiectasis require replacement of both lungs (bilateral lung transplantation or heart–lung transplantation). Most other diseases can be treated with

single-lung transplantation.

Single-lung transplantation involves a pneumonectomy followed by implantation of the allograft with anastomoses of the main bronchus, pulmonary vein, and pulmonary artery. Bilateral lung transplantation is performed as two sequential single-lung transplants. These are performed via either a sternotomy or bilateral thoracotomy, with or without cardiopulmonary bypass. Heart–lung transplantation mandates cardiopulmonary bypass and is performed via a sternotomy. Implantation is performed with tracheal, aortic, and atrial anastomoses.

The surgery must be performed meticulously. Careful dissection avoids damage to important structures such as the phrenic and recurrent laryngeal nerves and minimizes the chance of bleeding. Careful implantation reduces the risks of anastomotic complications. Ideally, implantation and reperfusion should be completed within 6 h.

Postoperative care

The first 24 to 48 h are critical to the long-term outcome of the transplant as well as for the survival of the recipient, and a principal aim of immediate postoperative care is to reduce allograft injury during this critical period.

The allograft inevitably sustains endothelial injury because of ischaemia and preservation, causing a breakdown of the capillary–endothelial barrier, resulting in leakage of fluid into alveoli. Increasing damage leads to a progressive impairment in gas exchange. This may necessitate prolonged mechanical ventilatory support, with an increased risk of infection and barotrauma often resulting in irreversible damage to the allograft.

The main principles of postoperative care of the lung transplant recipient are detailed below.

Early extubation

In the majority of patients, extubation is possible within 12 h of the procedure (in many cases much earlier than this). Extubation permits active coughing and clearance of secretions, institution of enteral nutrition, and early commencement of rehabilitation.

Fluid (crystalloid) restriction and diuresis

During the first 24 to 48 h, restriction of crystalloid intake and promotion of diuresis minimize the development of the pulmonary oedema characteristic of ischaemia-reperfusion injury. Crystalloid intake is limited to 1500 ml/day during this time. Colloid solutions are used for haemodynamic requirements.

Early mobilization

Patients with endstage lung disease are usually debilitated and in poor condition. It is therefore vitally important they are mobilized as early as possible. This improves appetite and sleep, and prevents complications such as basal atelectasis and deep venous thrombosis. Most patients are able to sit out of bed within 24 h and can participate in a gymnasium program by day 3. Adequate analgaesia is imperative for effective rehabilitation at this early stage.

Nutrition

The early institution of adequate calorie intake is necessary to overcome the severe catabolism stimulated by surgery. This is particularly important in patients with a poor pretransplant nutritional state. In most cases, enteral feeding can be started within 24 h (orally or via a nasogastric or percutaneous gastrostomy tube). If the gut is not functioning, parenteral nutrition should be used.

Prevention of infection

Bacterial infection remains the most significant early problem and is responsible for most deaths during the perioperative period. In most cases the organism is recipient derived. Antibiotic prophylaxis is administered until the patient is mobile, all drains have been removed, and respiratory secretions are clear. The underlying disease and/or pretransplant microbiology results dictate the choice of antibiotic. In cystic fibrosis and other septic lung diseases the antibiotics used cover *Pseudomonas aeruginosa* and *Staphylococcus aureus*. In other patients, community acquired respiratory pathogens (pneumococcus, haemophilus, etc.) and *Staphylococcus aureus* are targeted.

Fungal infection in the form of oropharyngeal candidiasis is common post-transplant and prophylaxis with topical nystatin or amphotericin is effective. Systemic prophylaxis against candida is not generally necessary. Aspergillus is the commonest cause of invasive fungal disease, and in single and bilateral lung transplantation, nebulized amphotericin (5 mg three times a day) given for the first month post-transplant is effective in reducing aspergillus-related problems. Routine use of itraconazole prophylaxis is dependent on local policy and experience. It is very uncommon for heart–lung transplantation recipients to have problems with aspergillus in this period as the tracheal anastomosis is not ischaemic, and all diseased lung tissue is removed. Routine prophylactic strategies are therefore not required.

Viral infections (specifically herpesviruses) tend to occur later in the recovery period but prophylaxis must be administered from the early stages to be effective. Ganciclovir is very effective in reducing both the incidence and severity of cytomegalovirus and other herpesvirus related illness, but there is no consensus on the optimal prophylaxis regimen. Most units opt for a combination of intravenous and oral therapy for 1 to 3 months, given to any recipient who is serologically cytomegalovirus-positive or who receives an allograft from a cytomegalovirus-positive donor. Herpes simplex virus commonly causes mucocutaneous infection. In the occasional cytomegalovirus-negative donor/recipient match (where ganciclovir is unnecessary) aciclovir is used for herpes simplex prophylaxis.

Cotrimoxazole prophylaxis is used to prevent both pneumocystis infection and toxoplasma reactivation. Standard therapy is 480 mg daily or 960 mg three times a week. Therapy is given for a minimum of 12 months or until corticosteroid therapy has been reduced to physiological replacement doses. If cotrimoxazole is not tolerated, nebulized pentamidine (300 mg per month) is an effective alternative.

Prevention of rejection

In lung transplantation, there are three basic phases of immuno-suppression—induction, consolidation, and maintenance. The details of the exact combinations and doses of agents used vary from unit to unit. Table 2 shows a typical regimen.

Induction phase

Typically, antithymocyte globulin, azathioprine, and corticosteroids are used. Therapy commences immediately pretransplant. The use and length of induction with antithymocyte globulin is not standardized and varies with local unit policy.

Consolidation phase

This phase covers the period when stabilization of the immunosuppressive regimen is achieved. The aim is to achieve optimal immunosuppression (prevention of acute rejection) with tolerable side-effects. It is important to remember that the primary goal is to minimize acute rejection. Inevitably side-effects do occur, but these should not determine immunosuppression levels at the expense of acute rejection. Cyclosporin is introduced and dosing modified to achieve whole blood trough levels above 350 ng/litre. This is usually achieved with doses around 10 mg/kg/day (given in two divided doses). Drug absorption in the perioperative period can be very variable, especially in cystic fibrosis patients where dosing three times a day is usually necessary to achieve satisfactory levels. Azathioprine is continued at a dose of 1 to 2 mg/kg/day, aiming for a white blood cell count of 4×10^9 to 6×10^9/litre. Prednisolone doses are gradually reduced to a maintenance level of 0.2 mg/kg/day.

Maintenance phase

This phase continues for the life of the recipient and is discussed below.

Follow–up

The thrust of management in the longer term is to maintain allograft function, and to minimize the side-effects of immunosuppression. Best lung function is usually established by 6 to 9 months post-transplant. The incidences of acute rejection and infection are highest in the first 3 months. Immunosuppression can be slowly reduced, aiming to stop prednisolone after 12 months and achieving target cyclosporin levels at this time of 150 to 250 ng/litre. Immunosuppression must, however, be tailored to individual requirements and prevention of acute rejection remains the primary goal.

Monitoring of symptoms, chest radiography, and spirometry are the basis of allograft surveillance. Small handheld spirometers enable daily home monitoring of lung function. A 10 per cent or greater fall in the forced expiratory volume in 1 s (FEV_1) prompts review and investigation of the cause.

When allograft dysfunction occurs, transbronchial biopsies are performed to diagnose the cause. Acute rejection and infection cannot be distinguished clinically, and may occur simultaneously. Histopathological diagnosis of the cause of dysfunction is therefore mandatory. Some units perform regular surveillance transbronchial biopsies, but there is no evidence that there is any advantage to this approach in terms of outcomes.

Specific complications

Lung transplantation is rarely an uncomplicated endeavour. All recipients will suffer at some stage either a complication of the process or a side-effect of medication. Surprisingly, however, most patients do not experience life-threatening or major complications until the inevitable onset of chronic allograft dysfunction.

Many of the complications experienced by lung transplant recipients are common to all forms of solid organ transplantation and relate to drug side-effects (hypertension, renal dysfunction, osteoporosis, hypercholesterolaemia, etc.). The following problems relate specifically to lung transplantation.

Rejection

Rejection occurs in three distinct forms, defined by the underlying immunological events and the histological changes. Acute and chronic rejection are not defined by the timing of occurrence after transplant. Acute rejection can occur at any time, and 'chronic rejection' can occur after only 3 months. The two processes can coexist and probably have distinct immunological aetiologies.

Hyperacute rejection

This occurs within 24 to 48 h of transplantation and is caused by the presence of preformed antibodies in the recipient directed against donor HLA. Complement activation and widespread endothelial damage with vascular thrombosis result in rapid and severe allograft dysfunction, which is usually fatal. Removal of the antibodies with plasma exchange, and institution of an anti-B-cell therapy such as cyclophosphamide is the usual treatment.

Acute rejection

Almost all patients will experience at least one episode of acute rejection defined by the strict histological criteria listed in Table 3. The essential findings are a perivascular lymphocytic (predominantly T cell) infiltrate. Clinically, the patient often complains of 'flu-like' symptoms with dyspnoea and low-grade fever. In the early stages the chest radiograph may be normal or show pleural effusion and/or subtle alveolar shadowing. Left untreated, widespread alveolar shadowing and hypoxaemia occur. The diagnosis should be confirmed histologically (via transbronchial biopsy), as it is common for infection to occur simultaneously. These problems cannot be distinguished clinically.

Acute rejection is treated with intravenous methylprednisolone (0.5–1 g daily for 3 days). This is effective in the majority of cases. Steroid-resistant rejection is usually treated with either a polyclonal (ATG) or monoclonal (OKT3) antilymphocytic agent. In addition, it is also now usual to change the background immunosuppression by substituting either tacrolimus for cyclosporin or mycophenolate for azathioprine (or both).

Chronic rejection (obliterative bronchiolitis)

In lung transplantation, the term 'chronic rejection' denotes the presence of obliterative bronchiolitis. It is defined histologically (airway fibrosis and/or vascular sclerosis), not by the time of occurrence after transplant. The term 'chronic rejection' is something of a misnomer in this setting, as there are many processes (both alloimmune and non-alloimmune) that result in fibrotic obliteration of the airway lumen. The term 'chronic allograft dysfunction' is preferred: this implies the presence of obliterative bronchiolitis but does not imply an exclusively alloimmune aetiolgy.

Pathology

Obliterative bronchiolitis is a fibroproliferative scarring process, resulting in either total or subtotal obliteration of the airway lumen. In some cases, there is accompanying vascular sclerosis. Rarely, vascular sclerosis exists without airway changes

Physiology

The airway pathology translates into fixed airflow obstruction. This has enabled a non-invasive marker of the presence and severity of obliterative bronchiolitis to be developed, namely the bronchiolitis obliterans syndrome. Bronchiolitis obliterans syndrome is functionally defined by a fall in the FEV_1, as measured from a baseline average of the two best FEV_1 measurements achieved post-transplant taken at least 1 month apart. No reversible cause of the fall in lung function should be present. Table 4 summarizes the bronchiolitis obliterans syndrome grading system. It has been confirmed in a number of large series that bronchiolitis obliterans syndrome accurately reflects the presence and severity of obliterative bronchiolitis, and is widely used in clinical practice for this purpose.

Aetiology

Obliterative bronchiolitis is undoubtedly multifactorial in aetiology. Any discrete insult resulting in significant damage (e.g. aspiration, viral pneumonitis) will inevitably lead to scarring. The precise alloimmune mechanisms involved remain unknown.

There are two risk factors identified as being strongly predictive of the development of obliterative bronchiolitis—the frequency of early acute rejection and cytomegalovirus serological status. Recipients with more than two episodes of acute rejection in the first 3 months post-transplant have a threefold or higher risk of developing obliterative bronchiolitis than those with no acute rejection. Any cytomegalovirus serological positivity confers an increased risk of developing obliterative bronchiolitis, with cytomegalovirus-negative recipients who receive organs from a cytomegalovirus-positive donor being at greatest risk.

Natural history

Obliterative bronchiolitis confers an increased risk of death, with the hazard ratio increasing with worsening bronchiolitis obliterans syndrome status. Patients in bronchiolitis obliterans syndrome grade 3 have a risk of dying six times higher than those in bronchiolitis obliterans syndrome 0. Once acquired, bronchiolitis obliterans syndrome usually progresses quickly, with a mean time for progression to the next stage or death of only 150 days. This is more pronounced in patients with frequent early acute rejection who are at risk of an even greater acceleration of their disease. The main cause of death in patients with obliterative bronchiolitis is infection. Inevitably, bronchiectasis develops in association with the obliterative bronchiolitis, with organisms such as pseudomonas and aspergillus becoming

problematic.

Treatment

There are no controlled trials to guide treatment of this condition. In some cases, the disease arrests spontaneously. Augmented immunosuppression increases the risk of infection and is not usually effective. Most units now change immunosuppression early in the disease, substituting tacrolimus for cyclosporin, or mycophenolate for azathioprine. In advanced disease, numerous treatments such as methotrexate, phototherapy, and total lymphoid irradiation have been tried, but no consensus exists regarding the best course of action. If the disease progresses despite the above changes, immunosuppression is reduced in an attempt to minimize the impact of infections.

Specific infections

Cytomegalovirus

Before the introduction of ganciclovir, cytomegalovirus pneumonitis had a mortality rate of 50 per cent in cytomegalovirus-naive lung transplant recipients. This led to the matching of cytomegalovirus-negative recipients with cytomegalovirus-negative donors to prevent primary (donor acquired) infection. Ganciclovir therapy (both treatment and prophylaxis) has proven very effective in reducing both the morbidity and mortality associated with this infection.

Prophylaxis is effective in both delaying the onset of clinical infection and reducing the severity of subsequent infective episodes. In lung transplant recipients, cytomegalovirus most commonly causes pneumonitis. Cytomegalovirus syndrome describes an illness characterized by fever, malaise, and leucopenia, and may accompany target organ infections such as pneumonitis or hepatitis. Treatment is with intravenous ganciclovir (10 mg/kg/day). In cases of ganciclovir intolerance or resistance, foscarnet is used. Many patients will have cyclosporin related renal dysfunction and appropriate dose adjustments are necessary. Ganciclovir causes bone marrow suppression and severe leucopenia can occur in association with the low white cell count caused by cytomegalovirus. Antiviral therapy should be continued, and in this setting the addition of granulocyte colony stimulating factor is usually effective. Cytomegalovirus retinitis is uncommon in lung transplant recipients and requires specialist ophthalmology review and follow-up.

Aspergillus

Aspergillus infection occurs in pulmonary and extrapulmonary forms. Extrapulmonary disease is always invasive. There are several forms of pulmonary infection, and significant differences in the timing and pattern of aspergillus infection related to the type of transplant. In single and bilateral lung transplants where the bronchial anastomosis is relatively ischaemic, aspergillus related problems tend to occur early (most within the first 3 months). In addition, there is a very high risk of developing an airway problem if aspergillus is present in this setting. By contrast, in heart–lung transplantation, the tracheal anastomosis is not ischaemic (due to a collateral blood supply from the coronary arteries), and most aspergillus related problems occur later with the onset of obliterative bronchiolitis.

Aspergillus infection of any sort must be diagnosed and treated early. Airway manifestations and early invasive parenchymal disease have a good prognosis. Late-stage or disseminated disease is usually fatal. Treatment with systemic amphotericin in high doses is used. Liposomal amphotericin B is the preferred therapy because of its much improved side-effect profile compared with the conventional preparation (CAB). This is very evident in the setting of cyclosporin induced renal dysfunction which severely limits therapy with CAB. Treatment starts at a single dose of 5 mg/kg/day. It is well tolerated and can be given via a peripheral line.

The role of azoles (primarily itraconazole) as primary treatment of aspergillus infection in this population is not established. Itraconazole is generally reserved for prolonged treatment following a course of liposomal amphotericin B, or where long-term prophylaxis is required. The latter situation usually arises when aspergillus is repeatedly isolated in the setting of an airway complication, or in the presence of obliterative bronchiolitis.

Pneumocystis

Pneumocystis infection is now an uncommon occurrence following the introduction of specific (cotrimoxazole) prophylaxis. Pneumocystis pneumonia in lung transplant recipients presents in a very different manner from pneumocystis pneumonia in AIDS patients, and if not recognized can be fatal. The onset is often insidious, with a low-grade febrile illness and mild to moderate dyspnoea, often only on exertion. Chest radiograph may be normal or show subtle alveolar shadowing, usually in the upper lobes. High-resolution computed tomography scanning shows a ground-glass pattern. Severe hypoxaemia and gross radiographic changes are uncommon. Transbronchial biopsies typically show a granulomatous pneumonitis. On occasions, the histology mimics acute rejection and if special staining is not performed the organisms will be missed. In addition, in distinct contrast to pneumocystis pneumonia in AIDS, the organism load in lung transplant recipients is very low. Pneumocysts are not usually recovered from bronchoalveolar lavage fluid, and can be very difficult indeed to find in biopsy tissue. Repeated transbronchial biopsies may be necessary to diagnose the condition. Treatment is the same as that employed in HIV infected patients, although very high-dose therapy is usually not required with the low organism loads.

Mycobacterial infection

Tuberculous mycobacterial and non-tuberculous mycobacterial disease occur at an increased frequency in lung transplant recipients when compared with the general population. Occult and old mycobacterial disease is sometimes found unexpectedly in explanted organs. In single lung transplantation, disease may remain in the native lung.

With previous inadequately or untreated tuberculous mycobacterial infection, prophylaxis with isoniazid should be considered. The use of rifampicin as a chemoprophylactic agent is not advisable because of the profound interaction with cyclosporin. If full treatment is required, standard robust antituberculous regimens are indicated, and since rifampicin has the profound effect of inducing hepatic metabolism, cyclosporin doses must be adjusted accordingly (dose increases of four- to fivefold are usually required).

Non-tuberculous mycobacterial infections tend to occur later in the post-transplant course than tuberculous mycobacterial disease. Treatment is based on ethambutol and rifampicin with the addition of a macrolide (clarithromycin or azithromycin) and ciprofloxacin. There are no data to guide the choice or length of treatment, but a minimum of 12 months' therapy is usually required, depending on the organism. Intolerance of medication is a significant problem that often limits treatment options in the longer term.

Gastrointestinal complications

Gastrointestinal problems are common after lung transplantation. They often relate to the side-effects of medication and are usually limited to upper gastrointestinal symptoms and/or diarrhoea. All patients receive antibiotics post-transplant and *Clostridium difficile* should be considered in the differential diagnosis of any case of diarrhoea or large bowel problems. Cytomegalovirus (discussed above) is the only other common infective gastrointestinal complication.

Upper gastrointestinal motility problems are common in recipients with cystic fibrosis and are exacerbated by vagus nerve damage during the procedure. All patients are prone to constipation because of immobility and drug therapy such as narcotic analgaesia. Regular laxative therapy should be used throughout the perioperative period. Cystic fibrosis patients are particularly prone to this complication, manifesting as distal intestinal obstruction syndrome. This life-threatening complication is managed aggressively with oral gastrograffin.

Bowel perforations are a catastrophe requiring urgent laparotomy for diagnosis and treatment. The signs of acute abdomen may be modified or absent in patients on steroids, and a high index of suspicion is needed to ensure rapid diagnosis and appropriate intervention. Gastrointestinal lymphoma often presents with small bowel perforation.

Pancreatitis affects a small number of patients and, unless related to gallstones, is usually fatal. Drugs such as azathioprine and prednisolone, and cardiopulmonary bypass are recognized risk factors. a_1-antitrypsin-deficient and cystic fibrosis recipients are also at increased risk. Treatment is along standard lines but there is usually rapid progression to multiorgan failure and death.

Neurological complications

Most of the neurological problems encountered post-transplant are related to cyclosporin toxicity and respond to dose reduction. Tremor and headache are very common, peripheral neuropathy less so. Diffuse cerebral oedema has occurred in a number of patients and been attributed to an idiosyncratic reaction to cyclosporin. Recipients with cystic fibrosis are particularly prone to seizures, particularly in the setting of high cyclosporin levels and/or hypertension (also a side-effect of cyclosporin). Infection of the central nervous system and lymphoma should always be considered in the differential diagnosis of central neurological symptoms.

Malignancy

Solid organ and lymphoid malignancies occur at an increased frequency, affecting up to 4 per cent of recipients. Lymphoproliferative disorders are related to the intensity of immunosuppression, and are driven by replication of Epstein–Barr virus. Most cases are focused in the allograft and most occur in the first 12 to 18 months post-transplant. Classification of these disorders is problematic. Some are clearly a polyclonal lymphocyte proliferation that usually respond to a reduction in immunosuppression and aciclovir therapy. Other cases present as lymphoma (non-Hodgkin's) and are treated with standard chemotherapy regimens. There are many cases where the disorder falls between these two extremes and treatment decisions are often very difficult. In these cases, histology is not a reliable guide to clinical behaviour or response to treatment. Patients are usually given a 1 to 2 month trial of aciclovir and reduced immunosuppression, with either non-response or progression indicating the need for chemotherapy. Reduction of immunosuppression involves decreasing cyclosporin levels by 30 to 50 per cent, stopping azathioprine, and reducing prednisolone to 10 mg or less per day. There are, however, no evidence-based data to support these recommendations, which are based on clinical experience. The prognosis of these disorders is surprisingly good, especially if confined to a single organ system. Most respond to reducing immunosuppression. In those requiring chemotherapy, complications of the treatment are largely responsible for the morbidity and mortality seen. Patients diagnosed with advanced disease invariably have a poor outcome.

The commonest solid organ tumours seen are cutaneous malignancies (squamous or basal cell carcinoma). With early diagnosis and treatment they carry a very good prognosis. By contrast, other solid organ malignancies (lung, gastrointestinal tract, etc.) carry a very poor prognosis, usually resulting in death within 3 to 6 months of diagnosis.

Airway complications

The bronchial anastomosis is prone to problems, with most units experiencing an airway complication rate of around 10 per cent. The bronchial anastomosis is devoid of its normal bronchial arterial supply, relies on retrograde perfusion from the pulmonary arteries, and is therefore relatively ischaemic. Bronchial artery revascularization procedures are time-consuming, technically demanding, and not widely performed. Complications range from minor narrowing of the bronchus, to severe narrowing requiring surgical intervention, to dehiscence and death. Bronchial stenoses are treated with dilatation and/or stenting.

The situation in heart–lung transplantation is completely different. The tracheal anastomosis has a collateral blood supply derived from the coronary arteries. The airway is well vascularized and serious airway/anastomotic problems are rare.

Outcome

Many studies have shown that lung transplantation confers significant survival and quality of life benefits to the majority of recipients. In experienced units, survival figures of better than 80 per cent at 1 year and 50 per cent at 5 years are now achieved for all types of transplant and underlying disease category. For those surviving the first year, the adjusted 5-year survival figure is 60 to 70 per cent.

The main cause of death in the first 12 months is infection (predominantly bacterial). Acute rejection rarely causes death directly. Its main impact is as a risk for the development of obliterative bronchiolitis, which is the main factor determining long-term survival in the majority of lung transplant recipients.

Survival is usually associated with markedly improved lung function and this translates into an improved functional capacity. As an example, in our own cystic fibrosis recipient group, FEV_1 improved from a mean of only 21 per cent predicted pretransplant, to 88 per cent predicted at 1 year post-transplant. Many patients are able to return to work and live a near normal life. Several female lung transplant recipients have undergone normal pregnancies without specific transplant related complications.

Conclusion

Lung and heart–lung transplantation offer the only therapeutic option for many patients with a variety of endstage pulmonary and cardiopulmonary diseases. With increasing experience and the development of more effective immunosuppression, survival figures continue to improve. The limiting factor in providing transplants is the critical shortage of donor organs.

Further reading

Barr ML *et al*. (1998). Recipient and donor outcomes in living related and unrelated lobar transplantation. *Transplantation Proceedings* **30**, 915–22.

Dennis CM *et al*. (1993). Heart–lung transplantation for end-stage respiratory disease in patients with cystic fibrosis at Papworth Hospital. *Journal of Heart and Lung Transplantation* **12**, 893–902. Series highlighting the outcome in terms of both survival and quality of life of cystic fibrosis patients receiving heart–lung transplants.

Dennis CM *et al*. (1996). Heart–lung–liver transplantation. *Journal of Heart and Lung Transplantation* **15,** 536–8. Report of the outcome of a unique series of patients receiving combined abdominal and thoracic transplants.

Gross CR *et al*. (1995). Long-term health status and quality of life outcomes of lung transplant recipients. *Chest* **108**, 1587–93.

Heng D *et al*. (1998). Bronchiolitis obliterans syndrome: incidence, natural history, prognosis, and risk factors. *Journal of Heart and Lung Transplantation* **17**, 1255–63. Largest series dealing with the clinical behaviour of bronchiolitis obliterans syndrome.

Higgins R *et al*. (1994). Airway stenosis after lung transplantation: management with expanding metal stents. *Journal of Heart and Lung Transplantation* **13**, 774–8. Large series detailing the management of airway complications after lung transplantation.

Hosenpud JD *et al*. (1999). The Registry of the International Society for Heart and Lung Transplantation: Sixteenth Official Report—1999. *Journal of Heart and Lung Transplantation* **18,**611–26. Summary of the worldwide experience and outcome of all forms of pulmonary transplantation.

Joint Statement of the American Society for Transplant Physicians/American Thoracic Society/European Respiratory Society/International Society for Heart and Lung Transplantation (1998). International guidelines for the selection of lung transplant candidates. *American Journal of Respiratory and Critical Care Medicine* **158**, 335–9. Outline of the key criteria used for judging the timing of referral and listing for lung transplantation.

Jonas M, Oduro A (1997). Management of the multi-organ donor. In: Higgins RSD *et al*., eds. *The multi-organ donor. Selection and management*, pp 123–9. Blackwell Scientific Publications, Oxford. Detailed account of optimal donor resuscitation and management.

McNeil K, Dennis CM (1998). Heart–lung transplantation: intensive care. In: Klinck JR, Lindop MJ, eds. *Anaesthesia and intensive care for organ transplantation*, pp 115–20. Chapman and Hall, London. A detailed account of the principles involved in the intensive care management of these patients.

McNeil K, Wallwork J (1997). Principles of lung allocation. In: Collins GM *et al*., eds. *Procurement, preservation and allocation of vascularised organs*, pp 223–6. Kluwer Academic Publishers, Dordrecht.

Meester JD *et al*. (1999). Lung transplant waiting list: differential outcome of type of end-stage lung disease, one year after registration. *Journal of Heart and Lung Transplantation* **18**, 563–71. Highlights the importance of the underlying disease in determining death on the waiting list.

Yeatman M *et al*. (1996). Lung transplantation in patients with systemic diseases: an eleven year experience at Papworth Hospital. *Journal of Heart and Lung Transplantation* **15**, 144–9.

Yousem SA *et al*. (1996). Revision of the 1990 Working Formulation for the Classification of Pulmonary Allograft Rejection: Lung Rejection Study Group. *Journal of Heart and Lung Transplantation* **15**,

1–15. Describes the pathological changes and grading of rejection in lung transplants.

18.1 Joints and connective tissues: introduction

Jonathan C. W. Edwards

Modern medicine has moved away from diseases with Latin names towards concepts of disordered physiology. Discussion is increasingly of airflow obstruction, insulin resistance, or reduced ejection fraction. Rheumatology has been behind in this move. Terms such as rheumatoid arthritis and osteoarthritis remain popular and may do so for a few years more. A more physiological approach must come, but old habits die hard.

Five major components of joints are involved in disease: cartilage, bone, tensile tissues (ligament and tendon), and, in diarthrodeal joints, synovium and synovial fluid. Building an understanding of the structure and function of these tissues not only lays down a scientific basis for joint disease, but is also directly relevant to the clinic, providing a framework for explanation and reassurance for patients, which, arguably, is the rheumatologist's main function.

An emerging theme in joint physiology is that biophysics, neurophysiology, and immunoregulation are simply facets of a seamless whole of tissue homeostasis. Disease often arises when interactions between these elements of homeostasis break down.

Cartilage

The primary function of hyaline cartilage is the generation and maintenance of skeletal shape. Most hyaline cartilage ossifies during growth but small amounts are retained in the nose, ribs, and joints. Loss of articular hyaline cartilage reveals two local functions. Spongy bone collapses once cartilage is lost, indicating a force distributing function. The bone margin also remodels progressively in response to changing mechanical stimulation, demonstrating that articular cartilage retains a 'shape memory' function, reflecting its relative resistance to remodelling in response to stress.

Although hyaline cartilage is often seen as adapted to low friction, many human joints (and most avian) function perfectly with fibrocartilaginous surfaces. Fibrocartilage replaces hyaline cartilage following loss, so hyaline cartilage might be considered redundant in this respect.

Cartilage loss

Loss of hyaline articular cartilage is the commonest major joint problem. Loss may occur either by fragmentation, with formation of fissures and the release of debris into the joint, or by resorption. If any central concept survives the debate about what osteoarthritis means, it is cartilage fragmentation under load. Cartilage wear occurs at points of loadbearing. Resorption of cartilage occurs in inflammatory disease and involves replacement of cartilage at the articular margin by fibrovascular 'pannus'. The two processes often coexist.

A major problem with the concept of 'osteoarthritis' is the confusion of events leading up to cartilage fragmentation and the events put in train once fragmentation has started. The events leading to fragmentation are many and affect different joints to differing extents. Dysplasia, heritable biochemical defects, metabolic changes, heritable tendencies to new bone formation, non-physiological usage, obesity, and prior damage from inflammatory disease all contribute. However, it is not clear if all of these factors belong to the concept of osteoarthritis, or whether some represent 'primary' and others 'secondary' disease. In most cases the factors inducing cartilage fragmentation are not understood. The idea that changes in the subchondral bone lead to altered loading on cartilage and subsequent failure is currently popular. This is almost certainly true in Paget's disease, but perhaps more subtle bony changes in middle age underlie many 'primary' cases. It would certainly be the best explanation for the sequential appearance of Heberden's nodes over a period of months in the hands of women who use them rather little.

The events which follow as cartilage fragments are well documented in animal models of non-physiological use, and now in models of excessive bone formation. Changes in glycosaminoglycan composition occur early, with subsequent failure of the collagen framework and disintegration. Disintegration usually starts at the cartilage surface, but there are also examples of cartilage fracturing away from subchondral bone. A major unknown is the role of chondrocyte death in this sequence. Chondrocyte death precedes collagen disruption in at least some cases and could be the critical irreversible event in joints such as the hip.

Resorption of hyaline cartilage is prominent in both septic and aseptic arthritis. Again, several pathways contribute, most involving enzymes. However, chondrocyte death may also play a role here, since most published micrographs of cartilage being resorbed show significant numbers of dead chondrocytes. Toxic factors such as reactive oxygen species and depletion of nutrients may damage cells. Reactive oxygen species and proteolytic enzymes may also attack extracellular matrix. Enzymes may act locally at the surface of cells in pannus activated by cytokines, or indiscriminately, if released into synovial fluid, as may apply to neutrophil elastase.

Chondrocytes also show evidence of degrading their own matrix, both in inflammatory and mechanical disease. Metalloproteinases induced by cytokines such as interleukin 1 and tumour necrosis factor-a are likely mediators. Disruption of collagen by collagenases or gelatinases is likely to be the critical irreversible event. However, depletion of glycosaminoglycan by stromelysin or aggrecanase may make the collagen susceptible to mechanical damage through loss of swelling pressure. Mobilization of aggrecan may also occur through changes in non-covalent interactions involving hyaluronan and binding proteins such as TSG-6 (the product of tumour necrosis factor-senstive gene 6).

Bone

New bone formation commonly occurs in reaction to cartilage fragmentation. However, bone overgrowth has a wider significance. In many joints, subchondral bone remodels gradually over decades, with flaring at the margin and changes in articular congruity. This occurs in the absence of any prior failure of cartilage. In the ninth decade osteophytes are commonplace. One of the major obstacles to effective management of 'hard tissue problems' is the lumping together of new bone formation with loss of (cartilage) joint space as 'degenerative change' in radiographic assessment. The physician cannot reassure the patient with a clear explanation if he or she is not in possession of the facts.

New bone formation

The commonest articular problem attributable to new bone formation is probably pain from impingement of the under surface of the acromion on the rotator cuff tendon of the shoulder. Irritation of the medial collateral ligament of the knee is also common. Compression of other structures occurs, with carpal tunnel syndrome and sciatica being obvious examples.

Very little is known about the factors responsible for the rate of growth of periarticular bone. There is a documented familial element for some sites. A degree of physical stress is probably necessary, in that flaccid limbs may retain a lifelong adolescent bony outline. There is, however, no clear relationship to the degree of use.

A different form of new bone formation occurs at ligament insertion sites. Generalized forms are termed diffuse idiopathic skeletal hyperostosis or Forestier's disease.

These patterns of new bone formation may restrict movement, yet may protect against pain.

Tensile tissues—ligament, tendon, and enthesis

Apart from trauma and tenosynovitis (see synovium), tensile tissue problems are focused at entheses—the attachments of ligament, tendon, or aponeurosis to bone. Entheses are variable, containing fibrocartilaginous and hyaline cartilage elements, and undergo structural change at the cessation of growth, which may influence disease onset. Entheses provide routes for vascularization of the tensile tissue.

Enthesopathy is the central lesion of the seronegative spondarthropathies. Two patterns can be separated—axial and peripheral. Axial enthesopathy affects spinal ligaments and the sacroiliac joint. Peripheral enthesopathy affects ligaments around peripheral joints, the plantar ligament origin, and the Achilles tendon insertion.

Axial enthesopathy is strongly associated with the B27 major histocompatability complex class I allotype. This may reflect an unusual immunological microenvironment in tissues under tension. The same sites which overstretch in Marfan's syndrome become inflamed in ankylosing spondylitis: entheses, the lung apices, the aortic root, the ciliary body, and the sacral root sheaths. Several of these sites are favoured by intracellular infections such as tuberculosis and brucellosis, suggesting that major histocompatability complex class I mediated events, such as cytotoxicity, are inhibited there, perhaps to avoid inflammatory resorption of tensile matrix. Inhibition may be mediated by local production of factors such as transforming growth factor-b. Structural differences between B27 and other allotypes are emerging which raise the possibility that local inhibition of class I associated events at sites such as entheses may be defective in the case of B27.

Synovium

Synovium comprises a superficial intimal cell layer and a subintima, which can consist of any type of connective tissue. The intima, unlike epithelium, is an incomplete and loosely arranged layer of modified macrophages and fibroblasts.

Why such a seemingly trivial tissue should be the major target for several autoimmune and inflammatory disorders may appear puzzling. However, tracing the evolution of synovium reveals some clues. Prior to the development of the jaw in cartilaginous fishes the endoskeleton was a semirigid rod of no immunological interest. Primitive immunity and leucocyte activity is likely to have focused on coelomic (serosal) cavities. Splanchnopleura remains the site of lymphocyte origin in mammals. Cartilaginous fish developed cavities within the skeleton, lined by a new tissue—synovium. Synovial intima and serosae share the expression of a complement regulatory protein, decay-accelerating factor and the immunoglobulin IgG receptor FcgRIIIa, suggesting a shared pattern of immunoregulation. Bone marrow and endoskeletal leucocytopoiesis developed later in teleosts, with bone marrow stroma derived from the same perichondrial stock as synovium. Bone and synovial stromal cells therefore share many features, including readiness to express the adhesion molecule VCAM-1. Thus, the cells of the intimal layer show marked immunological specialization. Intimal macrophages may be adapted to an 'early warning' response to immune complexes, in terms of FcgRIIIa expression, and the fibroblasts have an enhanced capacity to support lymphocyte survival via VCAM-1 and decay-accelerating factor. This combination may make the tissue particularly susceptible to autoimmune disease.

Just as tensile stress may regulate the enthesial immunological microenvironment, other stresses may contribute to the synovial environment. Macrophage FcgRIIIa expression appears to be induced at sites of shearing, not only in synovium but also in dermis over bony prominences, matching the distribution of rheumatoid nodules. Hence biophysics and immunology are inseparable.

Functions of synovium

Functions of synovium are not easily defined because removal of synovium causes few problems. It regenerates rapidly. Misconceptions about synovium, such as the existence of 'Haversian glands' have flourished. Largely as a result of the work of J. R. Levick, it has become clear that the dynamics of synovial fluid are unique and quite different from glandular secretion. Other functions have also been reappraised with the general conclusion that synovium has to be understood on its own terms, not by analogy with other tissues.

Functions of synovium are summarized below.

Disconnection

Unlike other connective tissues, synovium maintains a plane of disconnection which allows movement between rather than within solid tissues. This involves the maintenance of a non-adherent tissue surface. The best example of failure of this function is adhesive capsulitis of the shoulder, in which synovium becomes adherent to itself.

Low friction

Synovium contributes to low friction between cartilage surfaces and between synovial surfaces. Three molecules are important: water, hyaluronan, and a glycoprotein, lubricin. Water is a good lubricant, but cannot alone maintain a film between surfaces under load. Hyaluronan is not a lubricant but a film-maintaining agent. It is a carbohydrate polymer which adds viscosity and elasticity to water making it almost impossible to squeeze a film from between two surfaces. Lubricin is a true lubricant, reducing friction to an extremely low level.

Although changes in synovial fluid in disease probably alter lubricating efficiency, this may have little clinical relevance. In the short term joints appear to function well despite dilution of synovial fluid with exudate. The value of the sophisticated lubricating function of normal synovial fluid is that it reduces long-term articular surface wear to zero. The same collagen molecules are present on the surface 50 years after they were put there. Persistent presence of exudate in joints might facilitate wear, but preserved cartilage in reactive arthritis argues against even this. Attempts to modify the properties of synovial fluid are of doubtful value.

Deformable packing

The articular surfaces of most joints are incongruent. The intervening space is packed with synovium. Synovial fluid fills the tiniest crevices, with a film thickness of 40 µm in many places. With joint movement synovium has to deform with minimal resistance. How this is done remains totally unknown. Loss of this deformability is at least as important a source of disability as bone or cartilage damage in inflammatory arthritis, and can be devastating in children. It deserves serious study.

Chondrocyte nutrition

Cartilage is avascular, and chondrocytes must derive nutrition via the subchondral plate or from synovium. The synovial route is most accessible, although there is no proof that cartilage cannot survive without synovium. Synovectomy is not followed by cartilage death. There is also no evidence that synovium is structurally adapted to the nutrition of cartilage since its vasculature is similar when lining tendons, which have their own vascular supply.

It is commonly suggested that synovial fluid is important in transferring nutrients from synovium to cartilage. In fact, synovial fluid impairs nutrient transfer by acting as a diffusion gap. Without it much of the cartilage would be in contact with vascular synovium. During movement synovial fluid may facilitate nutrition of surfaces which do not come in to contact with synovium, but the fluid capacity is small and routes through solid tissue may be effective. The only likely clinical relevance of synovial fluid in nutrition is that large effusions, particularly those containing fibrin, may impair diffusion from synovium to cartilage leading to ischaemia.

Control of synovial fluid volume and composition

The combined action of muscles and lymphatic 'hearts' normally returns free tissue fluid to the circulation. Since synovial surfaces are permeable to water and there is no active pumping of water into synovial cavities, joints might be expected to be dry. The constant presence of a small volume of synovial fluid appears to be due to hyaluronan. Water from plasma transudate can enter the synovial cavity freely, but once in the cavity is mixed with hyaluronan secreted by intimal fibroblasts. If water molecules leave the joint without hyaluronan molecules, the concentration of hyaluronan at the tissue surface rises until it is so great that it obstructs the further flow of water. This curious process, known as solute polarization, means that a given amount of hyaluronan traps a certain amount of water in the synovial space. The volume can increase with inflammatory exudate, but cannot decrease unless the synovium ruptures (well recognized in disease). Chronic leakage through fistulae

between synovium and lymphatics is probably also common in rheumatoid joints, and may explain why they often show synovial thickening but are dry on aspiration.

Synovitis and synovial lymphoid metaplasia

Synovial tissue responds to acute stimuli in a similar way to other tissues. Vasodilatation, oedema, hyperalgesia, and granulocyte accumulation occur in response to pyogenic infection or crystals. Synovium may be hyper-responsive: accumulations of crystals cause little inflammation at other sites. Synovial cavities are also reputed to be good sites for induction of an immune response to injected foreign antigen.

A peculiar feature of acute synovitis is that cell migration is polarized and segregated in relation to the cavity. Granulocytes pass rapidly into the fluid compartment and are sparse within the tissue. Mononuclear cells remain largely in the tissue and macrophages accumulate at the tissue surface. This leads to a thickening of the intima, previously misnamed 'hyperplasia'. The differential distribution of cells is partly explained by intimal fibroblast expression of VCAM-1, the ligand for which, a4b1 integrin, is present on mononuclear but not polymorphonuclear leucocytes. There is also a difference in the behaviour of macrophages and lymphocytes. Macrophages accumulate in the intima but lymphocytes are confined to the subintima. The reasons for this remain unclear.

Prolonged stimulation of synovium leads to a pattern of infiltration peculiar to the tissue. In addition to the features above, both T and B lymphocytes accumulate in the subintima and generate follicles, and the stroma becomes colonized by plasma cells. The tissue effectively becomes a hybrid between bone marrow and lymph node, contributing to the rubbery or 'boggy' feel characteristic of chronic synovitis. The likely reason for this is, again, the readiness with which synovial fibroblasts express VCAM-1. Synovial fibroblasts as a whole show enhanced expression of VCAM-1, and also decay-accelerating factor and complement receptor 2, in response to tumour necrosis factor-a *in vitro*. This suggests that the high-level expression of VCAM-1 and decay-accelerating factor by intimal fibroblasts in normal tissue reflects a combination of a general responsiveness of synovial cells and an unidentified local intimal stimulus. VCAM-1, decay-accelerating factor, and complement receptor 2 are utilized by lymphoid stromal cells to support lymphocyte survival.

It is likely that the formation of lymphoid tissue in synovium is a stereotyped response to the generation of proinflammatory cytokines such as tumour necrosis factor-a. The source of cytokine is likely to be different in different clinical syndromes. In what we call rheumatoid arthritis the characteristic feature of early synovitis is an increase in size, number, and activation of intimal macrophages, suggesting preferential activation of these cells. This is consistent with the initiating stimulus being small immune complexes capable of crossing endothelium and interacting with FcgRIIIa-expressing macrophages. Extra-articular features can be explained in the same way. The most consistent and specific immunological abnormality in rheumatoid arthritis remains the presence of IgG rheumatoid factors, probably formed from IgG oligomers. In systemic lupus other small immune complex species may have a similar effect. However, tissue damage is more limited, perhaps because IgG rheumatoid factor-secreting plasma cells can generate both antibody and antigen locally in synovium, whereas plasma cells secreting the autoantibodies of lupus only produce antibody.

In the seronegative spondarthropathies increasing evidence that the primary lesion is at the enthesis suggests that peripheral synovitis is secondary to cytokines generated at contiguous entheses. This would explain the relatively mild degree of intimal macrophage activation seen. The net result of lymphoid metaplasia may be very similar, although the precise lymphocyte populations involved may be biased by the (unknown) nature of the underlying immune response.

Pain

A structural approach to joints risks overlooking the essence of rheumatic disease—pain. Synovium, tensile tissues, and bone are innervated by pain fibres and there is growing knowledge of their physiology. However, the relationship between pain and events measurable in the terms described above remains about as ineffable as that between the music of Pablo Casals and the motion of a string of catgut. Prostanoids, central nervous system sensitization, depression, hopes and fears, cultural patterns, holidays, and personality interactions are all essential to an effect which can be simultaneously intractable and responsive in an instant to a word or facial expression. A cyclo-oxygenase inhibitor at night may be essential to face the next day, but the best analgesics may remain explanation and trust. Diagnostic terms such as fibromyalgia are often devised to fudge difficult concepts and tend rapidly to lose value. Appropriately, pain is becoming a discipline in its own right.

Applications to therapy

The growing understanding of the biology of joints has contributed to rational therapy in two ways. It has provided rationales for empirically derived current practices, and has generated new therapeutic avenues.

Articular cartilage has little useful capacity for repair. The best treatment for damaged cartilage, as for teeth, may be what we use now—replacement by non-living material. Attempts to induce cartilage regeneration are probably futile and have the disadvantage of needing to maintain a living material. The main remaining problem is osseo-integration of non-living materials. Dentists have achieved a good solution with titanium but optimum materials for joint prostheses are still under review.

Development of rational physical therapies has been slow, owing to our lack of understanding of the relationship between movement and homeostasis in soft connective tissues. Many studies have demonstrated that exercise reduces pain and improves outcome in arthritis, but the reasons are unclear. Physiotherapeutic techniques tend to remain based on pseudophysiological concepts and are rarely validated by well designed trials. The same applies to techniques of 'joint protection' in which specific actions are discouraged to reduce deformity. Pragmatic management of diseased joints by experienced therapists has a major part to play for patient groups whose psychomotor development is limited—especially children and those with central nervous system disease or trauma. For the average adult with arthritis it is less clear that there is justification for more than sympathetic encouragement to exercise.

The discovery of cyclo-oxygenase inhibition confirmed that aspirin and indomethacin were logical agents to use to reduce pain and stiffness due to inflammatory oedema. It explained why these agents have no long-term effect on inflammation, since cyclo-oxygenase products mediate vasodilatation, pain, and oedema rather than cell influx. Cyclo-oxygenase 2 inhibitors follow this path but their place has yet to be evaluated.

Logical treatment of chronic inflammatory disease depends on whether it is the inevitable result of a genetically determined low inflammatory threshold or an acquired immune response. The former is probably important in ankylosing spondylitis and some forms of juvenile arthritis. These conditions may best be treated by long-term modulation of balances in cytokine and growth factor levels but, as yet, we do not know which molecules should be the best targets. In the meantime, suppression of inflammatory cell function by methotrexate remains the main practical, if not very sophisticated, option. This suppressive approach is also currently the mainstay of therapy for disorders such as rheumatoid arthritis which appear to represent an acquired adaptive immune response. Until recently, synovitis has been suppressed by drugs with unknown and probably disparate modes of action. Intramuscular gold and penicillamine appear only to be effective in rheumatoid arthritis. Sulphasalazine is of benefit in inflammatory bowel disease and seronegative spondarthropathies affecting peripheral joints. Methotrexate and azathioprine have a broad spectrum of action. However, none of these agents has a good risk–benefit profile. Rational suppression of synovitis has come with agents which neutralize tumour necrosis factor-a. Antitumour necrosis factor-a antibodies and soluble receptor fusion proteins have been used to mop up tumour necrosis factor-a with very good results. A similar approach has been used for interleukin 1. Anticytokine therapies appear to be broadly safe but are costly, and there is rapid relapse after withdrawal.

There is an increasing view that disorders such as rheumatoid arthritis, based on an acquired immune response, should be amenable to long-term cure if the immune system can be 'reprogrammed' to 'forget' the autoimmune response. Gold is probably the only agent in current use which occasionally induces complete remission that persists indefinitely after drug withdrawal. Rational 'reprogramming' has been attempted with high-dose chemotherapy, followed by stem cell rescue, but the mortality rate for this procedure is significant and long-term results are awaited. Hopes that anti-CD4 therapy might induce a state of tolerance which would abolish the autoimmune response have not been realized, and long-term T-cell depletion has been a problem. B-lymphocyte depletion is currently under study and has shown some promising results, but further information is needed. A combination approach may well be necessary. At least results so far give grounds for optimism that safe definitive therapy may not be a decade away.

Further reading

Archer CW *et al.*, eds (1998). *The biology of the synovial joint*. Harwood Academic Publishers, Reading, MA.

Bird HA, Snaith ML, eds (1999). *Challenges in rheumatoid arthritis*. Blackwell Scientific, Oxford.

Brandt K, Lohmander S, Doherty M, eds (1998). *Osteoarthritis*. Oxford University Press, Oxford.

Edwards JCW, Morris V (1998). Joint physiology: relevant to the rheumatologist? *British Journal of Rheumatology* **37**, 121–5.

Isenberg DA *et al.*, eds (1998). *Oxford textbook of rheumatology*, 2nd edn. Oxford University Press, Oxford.

Isenberg DA, Miller JJ, eds (1998). *Adolescent rheumatology*. Martin Dunitz, London.

Levick JR (1996). Synovial matrix-synovial fluid system of joints. In: Comper WD, ed. *Extracellular matrix*, vol. 1, pp328–77. Harwood Academic Publishers, Amsterdam.

McGonagle D, Gibbon W, Emery P (1998). Classification of inflammatory arthritis by enthesitis. *Lancet*, **352**, 1137–40.

Sokoloff L, ed (1978, 1980). *The joints and synovial fluid*. Academic Press, New York.

18.2 Clinical presentation and diagnosis of rheumatic disease

Anthony S. Russell and Robert Ferrari

'Medicine is a first-rate profession for a second-rate intellect.' Whilst we may not fully agree with that statement from a first-class iconclast (George Bernard Shaw), we do agree that with experience it is possible to discipline the mind to follow routine pathways to arrive at a correct diagnosis and therefore a valid treatment plan. In modern medicine, rheumatologists are almost unique in relying heavily on the patient's history and physical examination before applying a relatively restricted number of valid tests to clarify the diagnosis.

Pitfalls occur at every stage of the diagnostic process. The first trap is the referral note or phone call from the primary care physician, conveying their impression of the case, plus laboratory results which may or may not be of relevance. This information is obviously important but must never be blindly accepted as fact. After all, the patient has been referred for a second opinion, which should be truly unbiased and not merely an automatic repetition of the views of the referring doctor. For example: 'This man has refractory gout which is not responding to treatment and his uric acid level is high. Allopurinol has not helped him and after 5 days he still cannot walk because of pain and swelling in his right foot.' The referring physician's clinical assumption that the patient has gout may be incorrect, and the plasma uric acid level could well be irrelevant. Furthermore, if the clinical assumption is correct, then the treatment is inappropriate. Therefore, whilst at no time hinting to the patient that aspersions are being cast on the referring physician's assumptions, the consultant must start from the very beginning, taking a careful history. This should include any previous or family history of such attacks, a general history for systemic disease, or any cause for secondary gout. There should be a comprehensive physical examination, commencing with inspection of the affected part. The correct diagnostic process can be laid out as in the algorithm (a term which we use for expediency) for monarticular arthritis, but note that algorithms can only be initiated at some distance down the diagnostic pathway (Fig. 1).

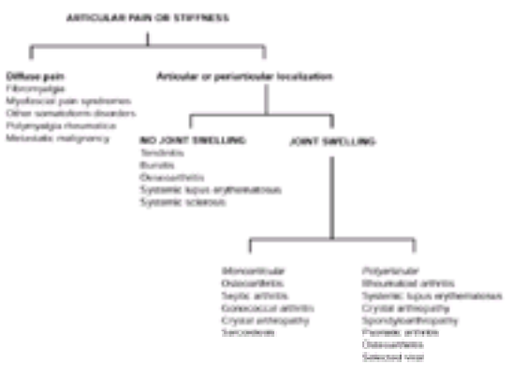

Fig. 1 An algorithm for rheumatology diagnosis.

General approach

The diagnostic process thus begins with the referral consultation note, followed by observation of the patient as they come in, their gait, demeanour, attire, and whether they are alone or accompanied. These, together with the presenting complaint and the initial features of the history, allow the development of an intuitive approach where the physician attempts to define aspects of the disorder and to arrive at a diagnosis (Table 1). The subsequent interview and examination are used to provide feedback and to support or refute those intuitive diagnoses; some observations necessitating a complete rethink of the process. Sometimes this rethink may relate to results of investigation or the development of new clinical features. Central to this diagnostic process is the classification into systemic rheumatological problems, localized—usually structural—problems, or functional somatic syndromes—perhaps best thought of as biopsychosocial disorders, an ungainly term, but one which emphasizes that biological, psychiatric, and social factors are all important. While there is obvious overlap between these categories, we believe it provides a useful framework for proceeding, partly because it allows for a positive diagnostic and therapeutic approach to this latter group, rather than simply regarding them as a group of patients where other diagnoses need to be excluded.

Some physicians feel insecure in a rheumatological diagnosis because of a lack of confidence in their rheumatological examination. This is unfortunate, because although the examination will provide important information, we believe that the principal diagnostic pointers come from a good history. The initial interview also provides a useful way of getting to know the patient and their environment. It is important, by open-ended questioning, to find out not only the details of the problem(s), but what the patient's fears and perceptions are, and especially, in a chronic disorder, what led them or their physician to seek a consultation at this time. This format also allows the patient to become at ease, and make them confident that you are truly listening. Specific directed questions are also of importance, both to elicit less frequent complaints, for example photosensitivity, xerophthalmia, recurrent miscarriage, and to elaborate on misleading terms. Thus, patients often describe 'hip pain', meaning pain in the buttocks, rarely due to hip disease; 'weakness' may reflect true neuromuscular disease, but more commonly reflects pain, for example in the shoulders or hip, or simply exhaustion or fatigue. These distinctions are crucial, and will be reviewed below (Table 2).

One of the key questions in rheumatology is 'where is the pain?', but it must be followed up with, 'and do you have pain anywhere else?'. Sometimes the answer to this, eventually, is 'all over', which by itself is very suggestive of a chronic pain syndrome. Instruments such as a pain diagram where the patient is asked to shade in areas that are painful, and also to indicate their intensity, may present a vivid pictorial representation of this. We find these useful, particularly for demonstrating the diagnostic value of this response.

When confronted with the patient who 'hurts all over', the next set of questions will often readily yield the diagnosis. The patient with diffuse pain should be asked to list all their other symptoms: a lengthy list indicates fibromyalgia or another functional somatic syndrome, and hearing extreme descriptions of individual symptoms is the next best clue. Indeed, after this, examination confirming lack of joint swelling and the presence of tender points provides ready and simple confirmation of the diagnosis, avoiding unnecessary consideration of other conditions in most cases. By contrast, when the patient fails to give a lengthy list of other symptoms, the less common causes of diffuse (poorly localized) pain should be sought. Be wary of those who complain of weight loss: this is rarely one of the long list of symptoms of fibromyalgia patients, and its presence in someone suspected of having fibromyalgia mandates a complete history and emphasis on the physical examination to look for another process. Conditions to be carefully considered in this context include polymyalgia rheumatica and metastatic bone disease (see below).

Systemic disorders

These include rheumatoid arthritis, other polyarthritides, systemic lupus erythematosus, polymyalgia, vasculitides, etc. Pointers to this type of disorder are malaise, anorexia, weight loss, rashes, fever, and multifocal symptoms including a description of actual joint swelling or persistent sensory motor deficits. Rheumatoid arthritis may begin as a monarticular problem, and although there are ways, even here, to approach a probable diagnosis, it may require more prolonged observation to establish this with certainty, obviously coupled with empirical management of symptoms. Even though the systemic disorders are typically widespread they are multifocal—that is a widespread focal problem in different joints. By contrast, pain 'all over' or 'from my head to my toes' suggests fibromyalgia, as indicated previously.

Unfortunately, patients' observations and descriptions of joint swelling are often unreliable and are particularly frequent, for example, in fibromyalgia where, by

definition, it does not occur (unless there is a second disease process going on). There is little point in going through all the diagnostic questions about rheumatoid arthritis and risk factors for gout in a patient who has never had documented swelling. However, if swelling is present on examination, then a further list of questions is aimed at making a specific diagnosis. This means that in practice it is not uncommon for rheumatologists to immediately examine the hands if the patient says that they hurt, and on seeing swelling, return to diagnostic questions about the polyarthritides to expand on a presumptive diagnosis, for example of rheumatoid arthritis.

Inflammatory arthritides are usually associated with morning stiffness of over 30 min, and the patterns of joint involvement (Table 3) and acuteness of presentation may help indicate the likely diagnosis (see Fig. 1). Specific questions directed to associated disorders are important, for example bowel disturbance, rectal bleeding, urethritis, mucosal lesions, conjunctivitis, psoriasis, etc.

Ensure that 'sun sensitivity' is not fatigue, headache, or cholinergic urticaria, but a true photosensitivity. Vague circulatory changes and cold hands are so common in the background population that this is of no help. A diagnosis of Raynaud's requires an extension of this to include at least pallor, usually followed by reactive hyperemia. Even this may occur in 5 per cent of the population without other disease. Always remember to question the validity of previous diagnoses: who made them, and on what grounds?

An important diagnosis in the over-55 age group is polymyalgia rheumatica. Patients may report fatigue and sometimes weight loss. The erythrocyte sedimentation rate is usually very high. Here the problems are located especially around the limb girdles and are associated with marked morning stiffness and sometimes systemic features. The erythrocyte sedimentation rate is virtually always substantially elevated, although this is very non-specific. Commonly the examination may be normal, and muscle tenderness is uncommon. In younger individuals with similar symptoms a somatoform disorder is more likely. Myositis itself is not usually painful, and weakness is the predominant complaint, as it is for myopathies.

Metatastic bone disease is less common. Patients with this condition usually have weight loss and fatigue as well as nocturnal 'bone pain', symptoms which should lead to enquiry about any previous malignancies and risk factors for malignancy.

Although in one sense a focal problem, a patient with a single, hot, swollen joint is best considered as having a systemic disorder. The critical issue here is to decide whether or not the joint is infected. Because of the risk of infection the same initial decision process is involved in a patient with known rheumatoid arthritis who has an acute monarticular 'flare'. If examination confirms an acute synovitis, i.e. not merely tenderness or a periarticular lesion such as cellulitis or erythema nodosum, then joint aspiration and fluid analysis and culture are the most important investigative procedures to be undertaken. Elements of the history are helpful, for example the development 2 to 5 days postoperatively of pain and swelling in the hallux points to gout, and in the knee or wrist to pseudogout. If gonococcus is a possibility then, as culture of joint fluid may be negative, cultures from other sites are equally important. However, the most important point is to remember that even in seemingly classical situations, aspiration remains advisable to achieve a definitive diagnosis.

So-called 'diagnostic criteria' are generally designed not for diagnosis of the individual patient but for classification of groups of patients, for example for studies or reports. They are, however, useful in providing an *aide-mémoire* to direct questions regarding specific features. Thus: the symmetric arthritis of rheumatoid arthritis, the photosensitivity and serositis of systemic lupus erythematosus, the widespread pain above and below the waist of fibromyalgia, the lack of important pain in myositis, the good response of spondylitis to therapy with non-steroidal anti-inflammatory drugs, etc. are all reinforced as important points to record.

Focal disorders

Here the patient presents with pain or other symptoms in one area, although there may be some radiation or spread. For these it is important to know the relevant anatomy and patterns of referral. Some of these are listed in Table 4 together with diagnostic pointers. To recognize meralgia paraesthetica, for example, one has to know of the existence and supply of the lateral cutaneous nerve of the thigh. A diagnosis of tendonitis should not be made unless it is associated with the name of the specific tendon or, if diffuse, with a systemic disease such as rheumatoid arthritis that can induce this. Too often it is an inappropriate attempt at diagnostic specificity in the presence of vague symptoms; diffuse pain and tenderness are often better considered under the functional section below.

Focal problems can be divided into truly articular and periarticular disorders. It must be remembered that they can be early manifestations of a systemic disease— see Table 5—and this illustrates why an inflammatory lesion is best regarded *ab initio* as a systemic problem.

Diffuse muscle pains are common. In the elderly, where radiological changes of osteoarthritis, particularly of the spine, are frequent, the pains may inappropriately be attributed to 'widespread osteoarthritis'. In general this is a diagnosis to be avoided. It does occur, for example in haemochromatosis, epiphyseal dysplasias, etc., but should be confirmed by clear-cut joint tenderness and decreased range of movement. However, the elderly may accumulate a number of focal disorders, for example unilateral osteoarthritis of the knee, a frozen shoulder on the right, postural cervical pain, and an osteoporotic fracture of the dorsal spine, the combination of which may simulate a systemic disease.

Functional somatic syndromes

Functional somatic syndromes are common in rheumatological practice, and are often badly managed because physicians tend to focus on organic disease. We are much more likely to be chagrined at missing the rare secondary deposit as a cause of thigh pain than by initially failing to recognize a patient whose somatic symptoms reflect depression or other emotional distress. We are subject to WHIMS (the 'what have I missed syndrome') that encourages repeated and fruitless investigation in this group to eventually arrive at a diagnosis by exclusion.

It is possible and beneficial to make the diagnosis after a good history and examination. Common symptoms are fatigue, weakness, sleep difficulties, headache, muscle aches, joint pains (plus a description of swelling), paraesthesiae, problems with memory and concentration, gastrointestinal symptoms including nausea, and alternating constipation and diarrhoea, and even irritable bladder. Such symptoms have been termed 'idioms of distress' and may be presented with a characteristic hyperbole. Thus, the pain is 'excrutiating' like 'red hot pokers in the back—you know' (as if this were an everyday experience for physicians). Apart from this, excrutiating pain is seen with fractures, septic/crystal arthritis, or nerve involvement. Patients with rheumatoid arthritis or osteoarthritis, however severe, don't normally use this terminology.

Patients with a functional somatic syndrome may also have arrived at a diagnostic label for their illness: repetitive strain injury, chronic whiplash, side-effects of silicone breast implants, candida hypersensitivity, and Gulf War syndrome to name but a few. We would also include fibromyalgia, chronic fatigue, irritable bowel syndrome, and others. It is possible, as has been suggested, that fibromyalgia (for example) may represent a central disorder of pain perception, perhaps associated with altered levels of substance P or nerve growth factor. We are unconvinced, but in any event this could be equally true of individuals with depression, with dysfunctional personalities, etc., and does not affect the overall approach to these disorders, amongst which there is considerable symptomatic overlap. All of these symptoms are common in the healthy population, and it may be more fruitful to ask oneself why the patient has presented to a physician, and at this time, rather than why they have headaches or fatigue in the first place.

Examination

The ability to detect joint swelling is important, but we are referring to obvious changes—if they are merely 'possible' or subtle, then rely more on the history for diagnostic pointers. A distinction between bony swelling and soft tissue/effusion is important and will often be of diagnostic significance.

Contrary to common belief, evident warmth of a joint (or redness) is unusual and would point to infection or crystal synovitis. The knee is normally somewhat cooler than the thigh or foreleg, and a lack of this coolness may actually be a sign that an observed knee swelling is inflammatory.

Careful palpation should allow one to distinguish between joint line tenderness, seen in arthritis, tenderness in between joints, as in an acute flexor tendonitis, periarticular tenderness, for example in lateral epicondylitis, and diffuse muscle tenderness, seen in some patients with local or generalized fibromyalgia (and very, very rarely in myositis).

The tender points found in fibromyalgia and many other somatoform diagnoses reflect a lowered pain threshold, and it has been suggested that they can be thought of as a 'sed(imentation) rate for emotional distress'. Nevertheless, they may provide diagnostic reassurance to the physician as other aspects of the examination are negative. In particular, joint swelling does not occur—although it is frequently referred to and described by the patient. The physician's observations are important

here, because if swollen joints are found, then some disease process is going on that may also need assessment, perhaps in addition to fibromyalgia.

With some exceptions, physical signs in rheumatology have not yet been subjected to assessments of their validity or positive predictive value. Thus, the stress tests for sacroiliac inflammation, while often described, are of no value. Tinel and Phalen's (sustained palmar flexion of the wrist for 60 s may induce finger paraesthesiae) signs have become modified and integrated into an approach to improve their use in the diagnosis of carpal tunnel compression. Adson's manoeuvre is also of little value. Palpation of tender muscle bands is subject to great intra- and interobserver error, and the relevance of tender trapezius or gluteal muscles in the diagnosis of postural/mechanical neck and back pain, although clinically probable, remains unproven. Even the classic 'limitation of straight leg raising' has a relatively poor sensitivity and specificity. Crossed straight leg raising appears quite specific, but is relatively insensitive. A great deal still needs to be done here.

Investigations

As physicians we are trained to order tests to help throw light on, and perhaps confirm, a diagnosis. We very commonly use them inappropriately. Thus, the idea of a 'rheumatology screen', so popular with some physicians, is entirely inappropriate. There are far, far more healthy people in the population who have a positive test for rheumatoid factor, antinuclear antibodies, or HLA-B27 than there are those with significant disease. Thus, for any test to be useful diagnostically, a Bayesian approach considering the pretest probability of diagnosis is critical, or to put it simply, the result must be taken in context. If the outcome, positive or negative, cannot affect the diagnostic probability, then the test should not normally be ordered. Otherwise, we subject the patient to unnecessary tests and often, when the results are positive, unreasonable anxiety that may take months and a specialty consultation to assuage. This does not include tests done for reasons other than diagnosis, for example to establish a baseline prior to treatment, or to obtain prognostic information, etc.

Diagnostic tests are sometimes ordered for the false reassurance a negative result provides in the presence of an insecure history and/or physical examination. Unfortunately, a false positive may occur and can be disastrous. Particular caution is therefore advised in ordering tests, or further consultations, to reassure 'the patient', especially those with a somatoform disorder. Negative findings generally fail to reassure, and indeed may heighten anxieties: a negative test is interpreted as puzzling and means that the problem is not yet solved. Similarly, if treatment, such as rest, does not improve the situation, the implication is that the disorder is too bad, not that the treatment was inappropriate. In fibromyalgia, for example, acknowledging and legitimizing the patient's distress and complaints is important. But although the patient may want a diagnostic label, and whilst this seems reasonable, labelling has been shown to increase disability and labels should not be applied that can be used to validate the 'sick role', i.e. they must come with reassurance and explanation. This reassurance will only be perceived as helpful rather than dismissive if the physician has initially taken care to legitimize and accept the validity of the complaints. The goal of treatment becomes the recognition and management of factors increasing symptoms and the focus on coping and improving functional status rather than curing 'the disease'.

Treatment

For many rheumatic diseases therapy has advanced enormously in the past 20 years, but the patient will not benefit if the correct diagnosis is not made. Thus, with allopurinol gout should rarely be an active problem, but patients are frequently still admitted to hospital for antibiotics because the correct diagnostic approach of synovial fluid aspiration has not been performed, or because the fluid has been allowed to clot so that crystals are not seen, or because crystals were not looked for, etc. The therapies for rheumatoid arthritis and ankylosing spondylitis, for example, have all progressed, but not to the stage of a cure. Thus, rheumatologists spend a lot of time informing and 'educating' patients. Many of these educational endeavours, when put to the test, have been shown not merely to convey retained information, but to actually alter behaviour and outcomes. It is always rewarding if we can 'fix' a problem, for example by prescribing antimalarials for palindromic arthritis, but all too often the additional role of the rheumatologist is supportive—to inform, to reassure where possible, and to provide continued advice and encouragement.

Further reading

Barsky AJ, Borus JF (1999). Functional somatic syndromes. *Annals of Internal Medicine* **130**, 910–21.

Deyo RA, Rainville J, Kent DL (1992). What can the history and physical examination tell us about low back pain? *Journal of the American Medical Association* **268**, 760–5.

Goodman SN (1999). Toward evidence based medical statistics. 2: the Bayes factor. *Annals of Internal Medicine* **130**, 1005–13.

Straus SE (1999). Bridging the Gulf War syndrome. *The Lancet* **353**, 162–3.

18.3 Clinical investigation

Michael Doherty and Peter Lanyon

Introduction

Disease 'markers' are pathological or physiological characteristics of an individual that assist in determining the diagnosis, the current activity of disease, or the expected prognosis of the condition in that individual (Fig. 1). Some markers relate to just one of these elements; others may relate to two or occasionally all three.

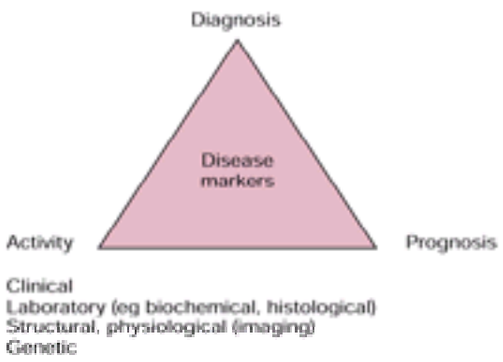

Fig. 1 Markers may be used for diagnosis, assessment of disease activity, or prognosis.

Clinical markers are derived from enquiry and examination of the patient. For many common rheumatic disorders clinical assessment alone gives sufficient information for patient diagnosis and management. In some situations, however, particularly with inflammatory, metabolic, or multisystem disease, a search for additional investigational markers may be warranted. It is important to emphasize that the requirement for and selection of investigations, as well as their subsequent interpretation, is principally determined by the clinical assessment. Investigations are an adjunct, never a substitute, for competent clinical assessment. There is no place for a battery of 'screening tests'. Investigational markers may include:

- Laboratory markers (biochemical, haematological, microbiological, histological) sought through investigation of body fluids and tissues.
- Structural and physiological markers, mainly assessed by imaging (radiography, scintigraphy, magnetic resonance imaging, ultrasound).
- Genetic disease susceptibility and prognostic markers—these hold promise for the future but at present only have clinical application to rare monogenic disorders.

When considering any investigation the following deliberations are pertinent:

- 'Is this the most appropriate investigation to answer the clinical question?' This may depend on various factors, for example the sensitivity and specificity of the marker being sought, its predictive value (which takes into account disease prevalence as well as the sensitivity and specificity), the cost and availability of the investigation, the pros and cons of invasive versus non-invasive tests.
- 'Will the result of this test alter the diagnosis or clinical management of the patient?' It is easy to initiate more investigations than are really required.
- 'Will I be able to interpret and act on the results of this test?' Tests should only be ordered if the implications of either a normal or abnormal result are understood.

In common rheumatological practice the investigations that are of most use in diagnosis are synovial fluid analysis and the plain radiograph. Confirmation of clinically assessed inflammatory disease activity and its response to treatment is mainly by the full blood count and either direct or indirect measures of the acute phase response. These investigations are therefore given special prominence in this chapter. The usefulness of other investigations will be discussed in the context of specific clinical scenarios.

Synovial fluid analysis

This is the key investigation to confirm the diagnosis of the two curable rheumatic diseases—septic arthritis and gout. Other crystal-associated arthropathies and intra-articular bleeding are also diagnosed in this way. Synovial fluid analysis is thus the pivotal investigation for an acute monoarthritis, especially with overlying erythema.

Synovial fluid can be obtained from almost any peripheral joint and only a small volume is required for diagnostic purposes. Aspiration of large joints should be no more uncomfortable than venepuncture. The patient should be informed of the purpose and nature of the procedure and positioned on a couch in a comfortable and relaxed position with full exposure of the relevant joint. The risk of introducing sepsis is negligible as long as sterile equipment and the same sensible precautions used for venepuncture are employed.

Macroscopic appearance

Normal synovial fluid is present in small volume, contains very few cells, is clear, colourless, to pale yellow, and has high viscosity due to macromolecular hyaluronate (Fig. 2 and Plate 1). In general, with increasing joint inflammation the volume increases, the total cell count and proportion of neutrophils rises (causing turbidity), and the viscosity lowers (due to degradation of hyaluronate by protease). However, there is such overlap between arthropathies that these features are of little diagnostic value. Frank pus or 'pyarthrosis', due to very high neutrophil counts, should always lead to exclusion of sepsis but can occur with any florid synovitis such as acute crystal synovitis or rheumatoid. High concentrations of urate or cholesterol crystals may result in white synovial fluid—joint 'milk'.

Fig. 2 Different macroscopic appearances of synovial fluids: (a) on the left, clear straw-coloured fluid from an osteoarthritic knee (easy to read writing behind it); (b) less viscous, turbid (high cell count) 'inflammatory' fluid from a rheumatoid knee; and (c) uniform bloodstaining (haemarthrosis) due to acute pseudogout. (See also Plate 1.)

Non-uniform bloodstaining of synovial fluid is common and reflects inconsequential needle trauma to synovial vessels. Uniform bloodstaining (haemarthrosis) most commonly occurs in association with florid synovitis but may also result from a bleeding diathesis, trauma, or pigmented villonodular synovitis. A lipid layer floating above bloodstained fluid is diagnostic of intra-articular fracture.

Gram stain and culture

If sepsis is suspected synovial fluid should be sent for urgent Gram stain and culture. Placement in blood culture bottles in addition to a sterile universal container increases positive yields, especially of anaerobes. If gonococcal sepsis or uncommon organisms are suspected, especially in immunocompromised patients, it is advisable to discuss this with the microbiologist so that the optimal cultures can be established and molecular techniques of antigen detection used if appropriate. Although a positive result on Gram staining is found in over 50 per cent of cases of adult septic arthritis (predominantly *Staphylococcus aureus*), a negative result does not exclude infection. If there is a strong clinical suspicion of sepsis the patient should be given intravenous antibiotics pending the results of the synovial fluid, blood, and other culture results.

Crystal identification

Accurate identification of common synovial fluid crystals requires a compensated polarized light microscope and an experienced observer. Monosodium urate and calcium pyrophosphate crystals may be seen by ordinary light microscopy but confident identification resides in their light characteristics as well as their morphology. Analysis is best performed on fresh unrefrigerated synovial fluid taken into a plain container to avoid problems of crystal dissolution, postaspiration crystallization, and artefacts from tube additives. If only a few drops are obtained these should be placed straight onto a clean microscope slide and a second slide or coverslip placed on top. Even with an apparently 'dry tap' it is worth expelling the contents of the needle onto a slide as a very small amount of fluid is sometimes obtained and may be diagnostic. Urate crystals are long and needle-shaped and show a strong intensity with negative birefringence (Fig. 3 and Plate 2). Pyrophosphate crystals are smaller, rhomboid in shape, usually less numerous than urate, and have weak intensity and positive birefringence (Fig. 4 and Plate 3).

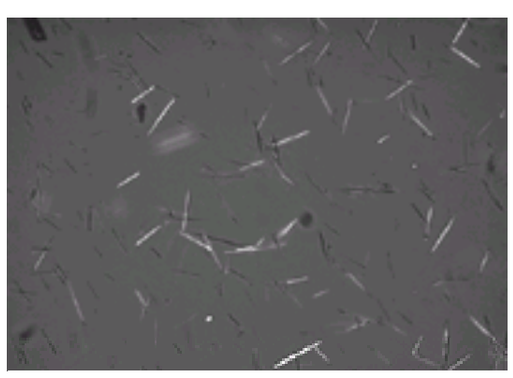

Fig. 3 Monosodium urate crystals viewed by compensated polarized light microscopy (×400) showing bright birefringence (negative sign) and needle-shaped morphology. (See also Plate 2.)

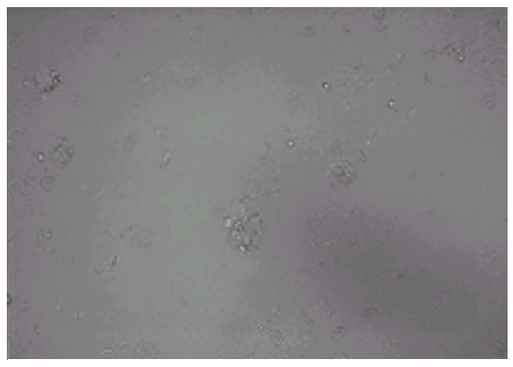

Fig. 4 Calcium pyrophosphate crystals viewed by polarized light microscopy (×400) showing weak birefringence (positive sign), scant numbers, and a predominantly rhomboid morphology. These are clearly more difficult to detect than urate crystals. (See also Plate 3.)

Although usually identified in the setting of acute synovitis, crystals are also often present in fluid aspirated from the joint after the attack has settled. Aspiration of an asymptomatic first metatarsophalangeal joint (gout) or knee (gout, pseudogout) may therefore permit confirmation of a suspected diagnosis. This is particularly important in gout because of the possible implications of life-long hypouricaemic therapy. The diagnosis can also be made by analysis of a tophus aspirate.

Plain radiography

In conjunction with a full history and examination this remains the single most useful imaging technique for assessment of rheumatic disease. Although a radiograph is a static record of predominantly past events, it can demonstrate visually alterations that reflect the underlying pathological processes of rheumatic disease (for example cartilage and bone erosion, bone remodelling, calcification). The abnormalities that may be seen on a plain film include:

- soft tissue swelling—seen as altered skin contours and displaced fat planes and intracapsular fat pads (fat appears dark on a radiograph)
- decreased or increased bone density (localized or generalized; Table 1)
- joint erosion (non-proliferative or proliferative marginal erosion, central erosion)
- joint-space narrowing (osteoarthritis—focal; inflammatory arthritis—generalized)
- new bone formation (osteophyte, enthesophyte, syndesmophyte)
- periosteal reaction (Table 2)
- calcification (cartilage—chondrocalcinosis; synovium, capsule, ligament, tendon, muscle, fat, vascular, skin)
- bone cysts and radiolucent lesions (Table 3)
- intra-articular osteochondral bodies
- deformity.

Although most of these abnormalities taken individually have low specificity, various combinations of some of these features, together with their targeting of certain joint sites (Fig. 5), result in characteristic patterns of abnormality and distribution that have high diagnostic specificity. The distribution of joint involvement, of course, is usually apparent following clinical assessment of the patient, and joints to be investigated by radiography will usually be selected on this basis. An important exception, however, is seronegative spondarthropathy where sacroiliac involvement is often asymptomatic and is difficult to detect clinically. For suspected seronegative spondarthropthy an anteroposterior view of the pelvis and a lateral thoracolumbar spine view (i.e. two films) are usually sufficient to show sacroiliitis and syndesmophytes if these are present.

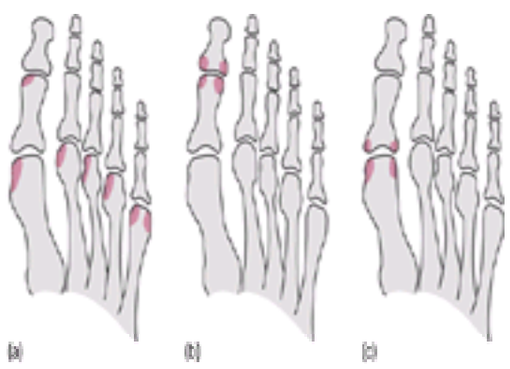

Fig. 5 Diagram to show different target sites of involvement in the forefoot for (a) rheumatoid arthritis, (b) psoriatic arthritis, and (c) osteoarthritis.

Radiographs should be selected to answer specific questions. For example, to address the question of whether a patient with chronic inflammatory polyarthritis affecting hands, elbows, neck, knees, and ankles has erosive disease typical of rheumatoid, posteroanterior views of hands and feet (i.e. two films), but not radiographs of all symptomatic joints, are appropriate. This is because rheumatoid erosions appear first in wrists and the small joints of hands and feet, and may first affect the metatarsophalangeal joints, even if they are relatively asymptomatic. However, if the degree of structural damage in one large joint is a principal cause for concern then a radiograph of that particular joint should obviously be taken. For most joints a single (two-dimensional) view is sufficient (for example anteroposterior view of pelvis, posteroanterior view of both hands, posteroanterior view of both feet), although two views are required for some (for example posteroanterior standing view of both knees plus individual lateral or bilateral skyline patellofemoral view). Thus selection of radiographs will often differ for purposes of diagnosis or disease assessment.

Erosions

An important hallmark of inflammatory arthropathies is cartilage and bone erosion. Intracapsular bone erosion first occurs at the 'bare areas' of the joint margin ('marginal erosion') where bone is exposed directly to inflammatory synovium without the protection of overlying cartilage. Loss of the sharp cortical line, the 'dot-dash' appearance, is the first radiographic sign that precedes more definite scalloping of the bony contour (Fig. 6). Cartilage erosion also commences at the joint margin and slowly works centrally, resulting in relatively late loss of interosseous distance or 'joint space'. Both rheumatoid disease and the seronegative spondarthropathies (especially psoriatic and chronic reactive arthritis) cause marginal erosions. In rheumatoid disease, however, the aggressive synovitis overwhelms any reparative response, presenting a very atrophic appearance ('non-proliferative erosions'; Fig. 6 and Fig. 7) with no new bone or periosteal reaction and only juxta-articular osteopenia (a sign of inflammation) and soft tissue swelling as accompanying early radiographic features. By contrast the seronegative spondarthropathies are characterized by a degree of low-grade inflammation that permits some reparative response. Such inflammation results in a tendency to fibrosis, calcification, and ossification. Marginal erosions in these arthropathies are therefore commonly accompanied by fluffy new bone formation ('proliferative erosions'; Fig. 6 and Fig. 8) with normal or increased periosteal and bone density rather than osteopenia. The fact that different joints are targeted in these conditions, and the common accompanying involvement of entheses—fibrous insertions of tendons, ligaments, or capsule into bone—further assists differentiation in most cases.

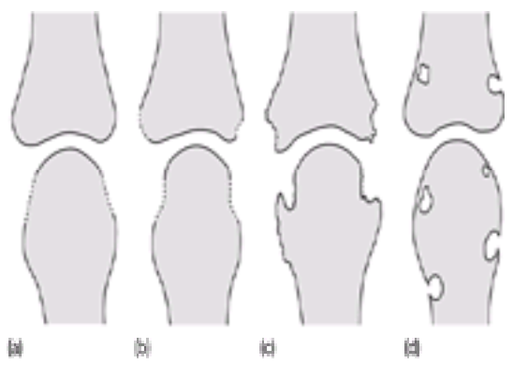

Fig. 6 Diagram of metacarpophalangeal joint showing (a) early dot-dash erosion, (b) the later definite non-proliferative erosion of rheumatoid arthritis, (c) the proliferative erosion of psoriatic arthritis, and (d) the intra- and extracapsular 'pressure erosions' of gout.

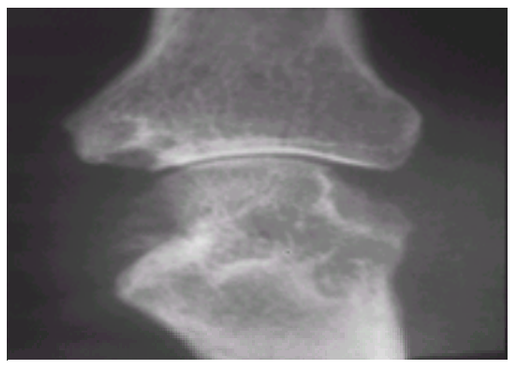

Fig. 7 Radiograph of metacarpophalangeal joint showing late non-proliferative marginal erosions of rheumatoid arthritis, more obvious proximally than distally (reflecting the more proximal than distal distribution of synovium in small finger joints) and eventual global loss of cartilage.

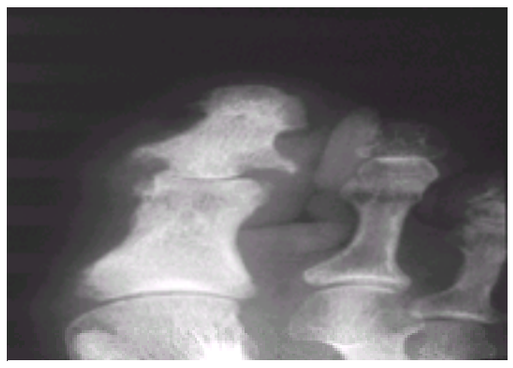

Fig. 8 Radiograph of the hallux showing proliferative erosions and cartilage loss of the interphalangeal joint, and associated increased bone density ('ivory phalanx') typical of psoriatic arthropathy.

In early septic arthritis the radiograph is often normal, apart from osteopenia and soft tissue swelling, for 1 to 2 weeks. However, erosion proceeds rapidly and results in generalized loss of joint space with loss of cortical integrity centrally (central erosion) as well as marginally. In chronic gout bony defects develop slowly as massive crystal concretions ('tophi') causing pressure necrosis to surrounding bone; such 'pressure erosions' (Fig. 6) occur at extracapsular as well as intracapsular sites and are unaccompanied by osteopenia.

Osteoarthritis

The features of osteoarthritis, by far the most common joint disease, are highly characteristic and contrast with those of inflammatory arthropathy. The two cardinal features are narrowing and osteophytes. By contrast to inflammatory arthropathies, joint space narrowing is focal rather than widespread within the joint, mainly targeting the maximum load-bearing region (Fig. 9). Bony osteophyte is most noticeable at the margins of the joint but also occurs centrally and as periosteal osteophyte ('buttressing') at sites such as the femoral neck. Subchondral sclerosis, or increased density of bone is also common, principally below the site of maximal narrowing. Additional features include subchondral 'cysts', osteochondral ('loose') bodies within the synovium, and an increased association with chondrocalcinosis. In contrast to inflammatory arthritis, the bone density is normal or increased and marginal erosions are not a feature.

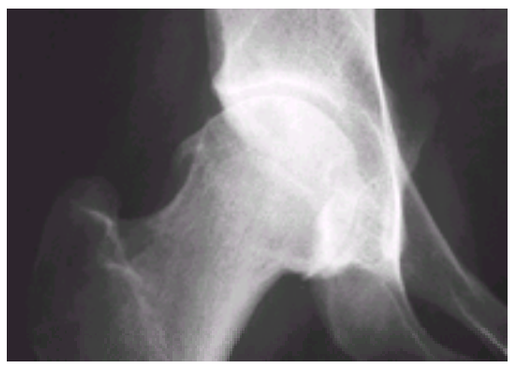

Fig. 9 Radiograph of the hip to show changes of osteoarthritis, specifically superior joint space narrowing, subchondral sclerosis, marginal osteophyte, and cysts.

Calcification

Calcification can affect any locomotor tissue. Calcification of fibro- and hyaline cartilage (chondrocalcinosis) is most commonly due to calcium yrophosphate crystals, less commonly to apatite or other basic calcium phosphates. This can occur as an isolated phenomenon (mainly age-associated, rarely as a result of metabolic or familial disease predisposition) or in association with structural changes of osteoarthritis (chronic 'pyrophosphate arthropathy'). Less commonly pyrophosphate crystals also cause calcification of the synovium and capsule, and linear tendon calcification (mainly hip adductors, Achilles, triceps).

Periarticular calcification is usually apatite. Isolated periarticular calcification mainly affects central sites such as the shoulder (supraspinatus tendons) or hip (abductor tendons), appearing as single dense concretions with rounded contours, as opposed to the linear calcification of pyrophosphate. Shedding of these crystal deposits can result in severe, self-limiting inflammation (acute calcific periarthritis) with reduction or loss of the radiographic calcification.

Spotty, multiple calcification of soft tissues (calcinosis) mainly targets peripheral and intermediate sites such as the finger pulps, wrists, and forearms and is a feature of connective tissue disease, most commonly CREST syndrome (**C**alcinosis, **R**aynaud's, o**E**sophageal dysmotility, **S**clerodactyly, **T**elangiectasia). Calcinosis requires distinction from small blood vessel calcification (increased in diabetes and chronic renal failure) which has a thin, meandering tramline appearance, sesamoids, and solitary dense calcified phleboliths. Myositis ossificans is rare and appears as dense sheets of calcification mainly at proximal sites such as the hip. Fine reticular or linear calcification of subcutaneous fat and muscle may follow young onset derma-tomyositis.

Other imaging

Arthrography

Injection of positive (iodinated) or negative (air) contrast, or a combination of both, can help delineate the soft tissue outline of a joint or other tissue space (for example bursa). The main use of plain film arthrography is at the knee to demonstrate a ruptured popliteal ('Baker's') cyst as a cause of calf pain and swelling. It is also commonly used with either computed tomography (**CT**) or magnetic resonance imaging (**MRI**) to provide better anatomical assessment.

Scintigraphy

Scintigraphy is a cheap, readily available technique that delivers only a very small amount of radiation. It involves gamma camera imaging following an intravenous injection of radioisotope, usually $^{99}Tc^{m}$ diphosphonate. Early 'flow' images obtained immediately postinjection, or a little later when the isotope is in the soft tissues ('blood pool' phase), reflect vascularity and will show, for example, the increased perfusion of inflamed synovium, Pagetic bone, or hypervascular primary or secondary bone tumour (Fig. 10). 'Delayed' images, taken a few hours after injection, indicate bone remodelling due to localization of the diphosphonate to sites of active bone turnover. Although non-specific and lacking high spatial resolution, the major advantage of scintigraphy is its high sensitivity for detecting important bone and joint pathology that may not be apparent on plain radiographs. It is particularly useful, following a normal or inconclusive plain radiograph of the presenting painful region, as the second imaging investigation to detect the following:

- bone metastases (at the presenting site and at clinically occult sites)
- bone or joint sepsis (at the presenting site and at clinically occult sites)
- early osteonecrosis (at the presenting site and at clinically occult sites)
- stress fracture
- reflex sympathetic dystrophy (algodystrophy)
- hypertrophic osteoarthropathy.

Scintigraphy is also useful in delineating the extent and current activity of Paget's disease of bone.

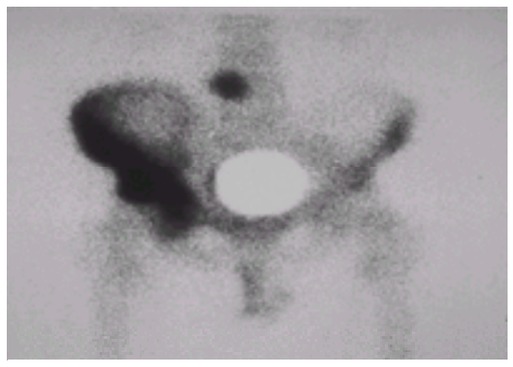

Fig. 10 Bone scan demonstrating secondary deposits of prostate cancer. The presenting painful lesion was in the right hemipelvis and the plain pelvic radiograph was normal. The spinal lesion (and two others not shown on this photograph) were asymptomatic.

Computed tomography

Computerized reconstruction of multiple radiographic scan sections can give detailed information on anatomy, especially of bone, allowing three-dimensional visualization of structures such as the spinal canal and facet joints. Its principal use is therefore in assessing areas of complex anatomy such as the spine or pelvis where plain radiographs may be inadequate (for example to investigate stenosis of the spinal canal). Drawbacks, however, include limited soft tissue resolution and exposure to a considerable radiation dose; in many situations it is has now been superseded by MRI.

Magnetic resonance imaging

The ability of MRI to image the anatomy and biochemistry of soft tissue as well as bone means that it provides detailed information not only on structure but also on the pathophysiology of all locomotor tissues. Further advantages include its capacity for multiplanar imaging (for example coronal, axial, sagittal, oblique) and its safety, without radiation exposure. The physics of MRI is complex. When a patient is placed in the magnetic field of the scanner the protons in the body align along the central axis of the field. Application of a radiofrequency pulse or 'sequence' causes the protons to spin in phase with each other. When the pulse is stopped the protons return to random spinning and 'dephase'. As they do so they emit a signal that is converted to an image by computer manipulation. In general, T_1-weighted short sequences are useful for defining anatomy, and T_2-weighted long sequences are useful for assessing pathology. Other sequences are selected for special purposes, for example the short tau inversion recovery sequence (**STIR**) is used to image marrow since it suppresses fat and makes the marrow appear dark. MRI, with or without enhancement with gadolinium, is particularly useful in detecting and assessing the following:

- early osteonecrosis (at the presenting site and the contralateral clinically occult site)
- intervertebral disc disease, root entrapment, and spinal cord compression
- osteoarticular and soft tissue sepsis
- osteoarticular and soft tissue malignancy
- internal mechanical derangement of joints (particularly the knee)
- assessment of soft tissue and periarticular pathology (for example early synovitis, rotator cuff tears, bursitis, tenosynovitis).

The choice between three-phase scintigraphy and MRI for detection of conditions such as early osteonecrosis, where both have excellent sensitivity (scintigraphy 90 per cent, MRI 100 per cent), will depend on practical issues such as ease of access, musculoskeletal reporting expertise, and local cost.

Ultrasonography

This is a safe and accessible technique for confirming soft tissue changes such as a hip joint effusion, popliteal cyst, or thickened Achilles tendon. Limited resolution, however, makes it inferior to CT or MRI for defining anatomical abnormality.

Blood tests for inflammation and systemic disease

The full blood count, erythrocyte sedimentation rate, and C-reactive protein may show changes that indicate the presence of inflammation somewhere in the body. These changes are very sensitive but are non-specific. They are mainly used as a semiquantitative measure to complement the clinical assessment of inflammatory disease and its response to treatment.

The systemic response to injury that results in these changes is summarized in Fig. 11. At any site of injury or inflammation macrophages and monocytes release soluble intercellular signalling polypeptides (cytokines) including interleukin 1, interleukin 6, and tumour necrosis factor-a. Some of these cytokines enter the systemic circulation and exert effects on the hypothalamus, bone marrow, and liver. These combined systemic effects are called the acute phase response, even though they accompany chronic as well as acute inflammation. Interleukin 6 is the main cytokine to influence the liver, causing increased production of certain acute phase proteins (including fibrinogen and C-reactive protein) but decreased production of other negative acute phase reactants (such as albumin and transferrin).

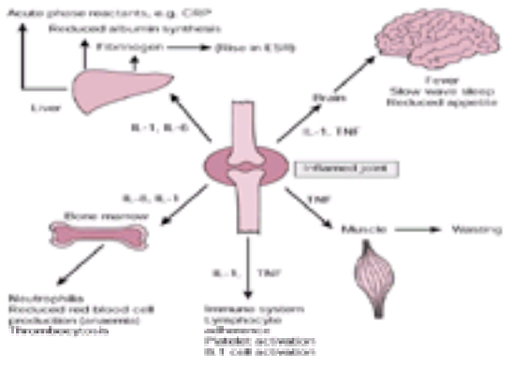

Fig. 11 Diagram to show the important elements of the acute phase response.

Much of the acute phase is beneficial for body defence and adaptation to injury, especially for dealing with the two major complications of injury that threaten life—haemorrhage and sepsis. For example, the thrombocytosis and increased serum levels of clotting factors facilitate haemostasis; neutrophilia and the increased serum levels of complement, immunoglobulin, and C-reactive protein (an opsonin) combat infection; and the anaemia and low serum transferrin levels result in diminished delivery of iron to bacteria and parasites.

Of all the acute phase proteins C-reactive protein shows the greatest shift from very low to very high levels, often representing a several hundred-fold increase in concentration. In addition C-reactive protein closely mirrors the current degree of inflammation, rising rapidly at its onset and falling as inflammation subsides, such that it is therefore the single most useful direct measure of the acute phase response. Interestingly, some rheumatic diseases—specifically lupus, systemic sclerosis, and dermatomyositis—associate with only modest or no elevation of C-reactive protein despite unequivocal pathological evidence of inflammation and tissue damage. The reason for this remains unclear, but patients with such disease are capable of mounting a typical acute phase response, for example in response to infection. In a patient with lupus or scleroderma gross elevation of C-reactive protein should therefore suggest an incidental cause such as sepsis. Some clinical features of active systemic lupus and infection overlap, and in this situation the C-reactive protein can prove a useful test.

The erythrocyte sedimentation rate is an old established indirect measure of the acute phase response. It mainly reflects the degree of rouleaux formation. Normally our circulating erythrocytes do not clump together because of the net balance of three electrical forces (Fig. 12):

- weak attractant van der Waal's forces resulting from red cells being bodies
- a strong repellent net negative surface charge, or zeta potential, due mainly to membrane sialic acid residues, and
- an attractant dielectric constant resulting from the charge characteristics of the plasma constituents.

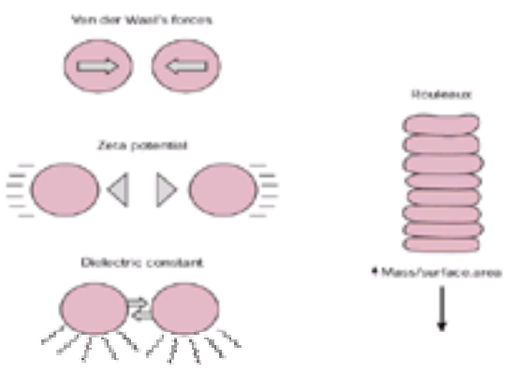

Fig. 12 Diagram showing the balance of three electrical forces that influence clumping of erythrocytes. Rouleaux sediment faster than individual erythrocytes.

In health the zeta potential far exceeds the sum of the two attractant forces, so that erythrocytes electrostatically repel each other and remain single. However, during the acute phase response the change in plasma protein concentrations leads to an increase in dielectric constant. Fibrinogen is particularly important in this respect. Although its increase in concentration is relatively modest, fibrinogen is a very asymmetric molecule that exerts a major electrical charge effect. The resulting increase in dielectric constant is sufficient to overcome the zeta potential so that rouleaux form more readily. Rouleaux have a higher ratio of mass per surface area so sediment faster than single red cells. This property is measured in the erythrocyte sedimentation rate. In the Westergren test system a 200 mm capillary tube is filled with the patient's blood. After 1 h the clearance of red cells from the top is measured. If there is little rouleaux formation the discrete red cells sediment only slowly and the clearance is small (less than 5 to 10 mm). However, if there is significant rouleaux formation the clearance is greater and the erythrocyte sedimentation rate in the first hour is elevated. Therefore in a patient with an acute phase response the erythrocyte sedimentation rate and C-reactive protein are both elevated, the erythrocyte sedimentation rate lagging behind the C-reactive protein in terms of speed of change.

The erythrocyte sedimentation rate, however, may be elevated for reasons other than the acute phase response. Immunoglobulins are very symmetrical molecules and their modest increase in concentration during the acute phase response has relatively little effect, compared with fibrinogen, on the dielectric constant. However, large increases in immunoglobulin concentration (for example in multiple myeloma or associated with autoimmune diseases such as Sjögren's syndrome) will increase the dielectric constant and lead to rouleaux formation. In this situation the patient may have a high erythrocyte sedimentation rate but normal or relatively low C-reactive protein. Such discordance between the erythrocyte sedi-mentation rate and C-reactive protein should lead to consideration of hypergammaglobulinaemia and myeloproliferative disease and to direct measurement of serum immunoglobulins.

In addition to the changes reflecting an acute phase response, the full blood count may show other alterations that are non-specific in themselves but which, taken in the context of the clinical features, may be characteristic of certain rheumatic diseases or their complications (Fig. 13). For example, neutrophilia may be seen in systemic vasculitis, and neutropenia in lupus. Furthermore, many of the slow-acting drugs used to control chronic inflammation have toxicity on the bone marrow so that the full blood count is often included in the routine monitoring of such treatment.

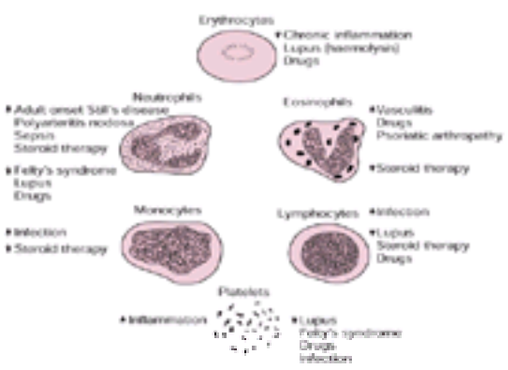

Fig. 13 Diagram showing some of the non-specific changes that may occur in individual elements of the full blood count in patients with systemic rheumatic disease.

Taken together, therefore, the full blood count, erythrocyte sedimentation rate, and C-reactive protein can be a useful complement of tests in the major rheumatic diseases to:

- assess inflammatory disease activity
- assess response to disease-suppressing treatment
- detect certain disease complications, and
- screen for drug toxicity.

However, it is important to appreciate that although an elevated acute phase response is consistent with inflammatory rheumatic disease, it is non-specific; also that the degree of elevation is often proportional to the amount or 'burden' of inflammatory tissue, for example isolated small joint synovitis in rheumatoid arthritis may not be sufficient to cause a detectable acute phase response, but this does not mean that inflammatory disease is not present.

Immunological tests

There are an increasing number of autoantibodies that can be detected from a serum sample by clinical laboratory services. Production of some of these is a common, age-related phenomenon that may be exaggerated by the presence of chronic inflammation. Their mere presence, therefore, often has low diagnostic specificity and little clinical relevance. If present in high concentrations, however, their disease specificity usually increases, so that it is always important to know how much antibody is present (the titre or concentration in units) rather than just whether it is detectable. For some autoantibodies (for example c-ANCA (antineutrophil cytoplasmic antibody)) their titre, if initially high, may be used to monitor the activity of the associated disease; only a few antibodies (for example anti-dsDNA) have high diagnostic specificity. Again, the correct choice and interpretation of these tests will depend on detailed knowledge of the patient. Different detection and assay systems exist for many of these autoantibodies and close liaison with the local immunology service is required.

Rheumatoid factor

The definition of a rheumatoid factor is an antibody directed against a specific region of the Fc (crystallizable) fragment of human IgG. The antibody itself may be of any immunoglobulin class, although IgM anti-IgG is the rheumatoid factor that is most commonly measured in the first instance. One of the traditional methods used to detect IgM rheumatoid factor is to coat latex beads with human IgG. If the patient's serum is then added to the test system the pentameric IgM antibody binds to the IgG causing the latex particles to flocculate producing a positive 'latex fixation test'. The amount the patient's serum must be diluted before this flocculation is lost is then determined; the higher this 'titre' the higher is the concentration of antibody present. Although 'rheumatoid factor' was so named because it was first detected in the sera of patients with rheumatoid arthritis, it also occurs in association with a variety of other conditions as well as in some normal adults (Table 4). It therefore has low diagnostic specificity, particularly in the elderly, and is not a 'test for rheumatoid arthritis'. In terms of sensitivity, it is present in the majority of patients with erosive rheumatoid disease but may only appear after many months or years of disease, once the diagnosis is beyond dispute. It is therefore of little value in making a diagnosis of rheumatoid arthritis, being neither sufficient nor necessary: rheumatoid arthritis is predominately a clinical diagnosis, based on detecting the presence of synovitis (capsular swelling, joint line tenderness, stress pain). However, if present in high titre at the onset of rheumatoid arthritis it associates with a poorer prognosis. IgG rheumatoid factor has greater specificity for major rheumatic disease, but the above caveats still remain. One situation where a negative rheumatoid factor is of diagnostic significance is in a patient with arthritis and nodules. As a general rule all patients with nodular rheumatoid are seropositive so that in this situation other causes of 'arthritis plus nodules' must be considered, for example tophaceous gout or hypercholesterolaemia.

Antinuclear antibody

An antinuclear antibody is any autoantibody directed against one or more components of the nucleus. Immunofluorescence microscopy after serum has been applied to a nucleated tissue substrate (for example rodent organs) or human cell lines (e.g. Hep 2) is the standard method of detection, and four main patterns of staining are reported. As with rheumatoid factor, the higher the titre of antinuclear antibody the greater its significance, but a high titre does not necessarily imply more severe disease. The specificity and sensitivity also vary according to the antigen preparation used in the test system and whether the antinuclear antibody measured is IgG or IgM. However, the tests are not universally standardized, and again, liaison with the laboratory is important in order to determine which cut-off titre is considered to be 'abnormal'. The many causes of a positive antinuclear antibody are outlined in Table 5. The commonest reason to undertake an antinuclear antibody test is in a patient with suspected lupus. For lupus, the antinuclear antibody has high sensitivity (97 to 100 per cent), but because the specificity is very low (10 to 40 per cent) a positive result does not make the diagnosis; by contrast, a negative antinuclear antibody virtually excludes it.

If a screening serum antinuclear antibody test is positive most laboratories will then attempt to determine the specific antigenic determinants. Some of these determinants are soluble and can be extracted from the nucleus, hence 'extractable nuclear antigens' (ENA), although many of the antigen–antibody specificities in human disease remain to be discovered. Compared with the antinuclear antibody, antibodies against specific nuclear antigens may have higher specificity for certain diagnoses or for certain patterns of system involvement within the same disease. For example, antinuclear antibody directed against double-stranded DNA is highly specific for lupus. Unfortunately, it is present in only a minority of patients and those in whom it is positive often have classic severe lupus (for example with renal involvement) and a clear clinical diagnosis.

Antibodies to 'Sm' antigen occur almost exclusively in systemic lupus erythematosus and may imply a poorer disease prognosis. Antitopoisomerase 1 and anticentromere antibodies are found exclusively in diffuse and limited scleroderma respectively. Antibodies to 'Ro' antigen occur predominantly in Sjögren's syndrome and systemic lupus erythematosus and associate with a high frequency of photosensitive rashes and a risk of neonatal heart block. Antibodies to RNP (ribonucleoprotein) are found in systemic lupus erythematosus, but also in a variety of other conditions, including scleroderma, myositis, mixed connective tissue disease, and rheumatoid arthritis. Although these antibodies may associate with disease subsets, there is little evidence that they are involved in disease pathogenesis.

The antiphospholipid syndrome, defined by the occurrence of arterial and venous thromboses, recurrent fetal losses, and thrombocytopenia in the presence of antiphospholipid antibodies, occurs in lupus and other autoimmune diseases and also in subjects with no other underlying disease. Antiphospholipid antibodies can be detected in assays for anticardiolipin antibodies (predominantly directed against a2 glycoprotein 1) and in phospholipid-dependent coagulation studies to detect lupus anticoagulants (prolonged Activated Partial Thromboplastin Time (APTT) which fails to correct with the addition of normal serum). Antiphospholipid antibodies also occur in a wide variety of rheumatic, infectious (bacterial, viral, protozoal) and malignant conditions, although in these situations they are not usually associated with thromboses.

Further information on these tests can be found in Chapter 13.14 and Chapter 18.10.2.

Tests for specific clinical situations

Chronic inflammatory disease at a single site

Patients with unexplained inflammatory disease at a single locomotor site (monoarthritis, bursitis, tenosynovitis, osteitis) should be considered for biopsy. The timing for this will vary according to how florid the lesion appears, but in general this should be undertaken for any undiagnosed lesion that has persisted for 6 months. The reason is to determine or exclude specific disease that can only be diagnosed by this means (Table 6). Although these conditions are uncommon or rare they require a specific treatment approach, rather than a continuing empirical symptomatic one. The commonest site for unexplained inflammatory monoarthritis is the knee, followed by the wrist and small hand joints. Arthroscopic biopsy is ideally used for larger joints (for additional information from direct visualization and guided biopsy), open biopsy for smaller joints and periarticular lesions. Tissue should be examined histologically and sent for culture, including mycobacteria. Apart from the specific conditions in Table 6, histopathology has no role in the diagnosis or management of most common rheumatic disease.

Investigation of suspected muscle disease

There are three principal investigations for the diagnosis and monitoring of muscle disease: serum creatine kinase, electromyography, and muscle histology. None are 100 per cent sensitive so that each may be normal despite abnormality detected by one or both of the others. Although creatine kinase is the most indirect measure it is readily available and commonly measured in the first instance. It is important to realize that elevation of the creatine kinase may result from a variety of causes (Table 7) and certain racial groups (e.g. Afro-Caribbean) have higher 'normal range' values. Electromyography or muscle biopsy will usually be undertaken next, the

choice depending on local availability and expertise. How much information on diagnosis and disease activity has been gained from the first two tests will then often determine whether the third is also undertaken. The one used for subsequent monitoring of disease activity will be that which was most helpful in confirming the diagnosis.

Electromyography measures the action potentials produced at rest and during voluntary contraction. Normal muscle is electrically silent at rest. On slight contraction motor-unit potentials of 500 to 1000 μV in amplitude and 4 to 8 ms in duration are recorded. On maximal contraction, as many motor units as possible are recruited and an interference pattern develops. With inflammatory polymyositis the electromyography may show a diagnostic triad of:

- spontaneous fibrillation
- short-duration action potentials in a polyphasic disorganized outline, and
- repetitive bouts of high-voltage oscillations produced by contact of diseased muscle with the needle.

Muscle histology can readily be obtained from a needle muscle biopsy sample. This is a relatively simple procedure requiring a local anaesthetic, small skin incision (no stitches required), an appropriate muscle biopsy needle, and no subsequent limitation of activity: it can easily be repeated serially for subsequent monitoring of response to treatment. The quadriceps is usually chosen, although the deltoid or other muscles can also be biopsied this way. The two or more small cores of tissue obtained by the needle need to be transported rapidly to the laboratory to be correctly orientated prior to freezing and sectioning. Immunohistochemical staining in conjunction with plain histology gives considerable information concerning primary and secondary muscle and neuromuscular disease. Although open biopsy will yield more tissue than needle biopsy, only a small amount of muscle is actually required and serial open biopsy is clearly problematic.

For further discussion of the investigation of muscle disease, see Chapter 24.22.4.

Investigation of suspected vasculitis

The clinical features of vasculitis often relate to specific organ involvement (ear, nose, and throat, neurological, renal, respiratory) but may be non-specific (malaise, weight loss, night sweats). In view of the potential toxicity of appropriate treatment, further investigation is always required. In terms of laboratory tests, simple measures such as a urine dipstick test and microscopy should not be overlooked, as the prognosis of many of these diseases is dictated by renal involvement. Antineutrophil cytoplasmic antibodies (**ANCA**) were initially detected in patients with glomerulonephritis. These antibodies are directed against enzymes present in neutrophil granules. Two main patterns of immunofluorescence are distinguished, cytoplasmic and perinuclear. The majority of c-ANCAs and p-ANCAs are specific for the enzymes proteinase 3 and myeloperoxidase respectively. These two patterns have been correlated with particular disease manifestations, for example c-ANCA with Wegener's granulomatosis and p-ANCA with microscopic polyangiitis. However, positive ANCAs occur in a variety of other settings, including malignancy and infections (bacterial and HIV infection) as well as other autoimmune diseases (inflammatory bowel disease, rheumatoid arthritis, lupus, pulmonary fibrosis). Therefore, the diagnosis of these conditions cannot be made or refuted on the ANCA test alone. Other evidence should be obtained by biopsy of an appropriate organ (nose, kidney, muscle, skin) or angiography.

For further information see Chapter 20.10.3.

Investigation of multiple regional pain

In most patients who present with widespread musculoskeletal pain the diagnosis is made from clinical examination alone, for example widespread rheumatoid disease. In some cases, however, there may be little to detect on clinical examination to explain the widespread pain. In most cases the diagnosis will be fibromyalgia. This is confirmed clinically by the appropriate symptoms (for example widespread pain, non-restorative sleep, marked fatigue, 'tension' headache, 'irritable bowel' symptoms, anxiety and depression, poor memory and concentration, urinary frequency) and the presence of widespread hyperalgesic tender sites and negative control sites (see Chapter 18.4). However, a number of other conditions may present similarly with multiple regional symptoms and few, if any, physical findings. In this situation, a limited screen (Table 8) is justified to detect conditions that have a specific treatment approach.

Further reading

Brower A (1997). *Arthritis in black and white*, 2nd edn. WB Saunders.

Gabay C, Kushner L (1999). Acute-phase proteins and other systemic responses to inflammation. *New England Journal of Medicine* **340**, 448–54.

Hoffman GS, Specks U (1998). Antineutrophil cytoplasmic antibodies. *Arthritis and Rheumatism* **41**, 1521–37.

McCarty DJ (1997). Synovial fluid. In: *Arthritis and allied conditions* (Coopman WJ, ed.), 13th edn, pp. 81–102. Williams and Wilkins, Maryland.

18.4 Back pain and regional disorders

Simon Carette

Low back pain

Introduction

Low back pain is one of the commonest symptoms and was the fifth leading reason for all visits to doctors' surgeries in the United States in 1990. Between 60 and 80 per cent of adults will suffer from at least one episode of back pain during their lifetime. Acute back pain is usually self-limiting, and the majority of subjects do not seek medical advice. Of those who do, more than 90 per cent will be back to work within 2 months, independent of the treatment received, including those in whom the acute episode results from a work-related injury for which compensation might be available. The 5 to 10 per cent of patients who remain disabled after this time represent a difficult therapeutic challenge due to the influence of psychological and social factors on the continuation of pain. This small percentage of patients is responsible for more than 75 per cent of the total costs of low back pain to our society, estimated to be between 1 and 2 per cent of the gross national product in most industrialized countries.

Significant risk factors for the occurrence of back pain include older age, heavy labour (in particular jobs requiring lifting in an awkward position), lower education and income, smoking, and obesity. Long-distance driving and whole-body vibration such as experienced by lorry drivers is a well-known risk factor for disc herniation. Prior episodes of back pain are strong predictors of recurrence. A number of psychosocial risk factors, or so-called 'yellow flags', predict poor outcomes. These include beliefs that back pain is harmful or potentially severely disabling, resulting in fear/avoidance behaviour and reduced activity levels, excessive reliance on aids and appliances, depressed mood, withdrawal from social interaction, and job dissatisfaction.

Many structures of the back, including the muscles, ligaments, discs, bones, and zygoapophyseal and sacroiliac joints are innervated and can therefore be a source of pain. However, in more than 90 per cent of patients presenting with low back pain, it is extremely difficult—if not impossible—to identify precisely the anatomical source of the pain on the basis of history and physical examination. These patients should be diagnosed as suffering from 'non-specific low back pain'. A host of clinical entities such as muscle strain, degenerative disc disease, facet syndrome, myofascial pain syndrome, segmental instability, minor intervertebral displacement, iliolumbar syndrome, piriformis syndrome, etc. have been described within this broad category based on the localization of pain and tenderness, reproduction of symptoms by specific manoeuvres, radiological features, or pathophysiological hypotheses. Unfortunately, the signs and manoeuvres described for each of these clinical syndromes lack sensitivity and specificity and are not reproducible even by experienced clinicians. Moreover, the claim that any of these entities is responsible for the pain in a given patient can very rarely be validated. For example, it is hazardous to ascribe pain to degenerative disc disease or apophyseal joint osteoarthritis when it has been shown that individuals with similar radiological changes can be completely asymptomatic. The only way to determine if the discs, or zygoapophyseal or sacroiliac joints are the source of pain in a given patient is through injection studies done under stringent, controlled conditions (see below).

Clinical approach o the diagnosis of low back pain

In evaluating a patient presenting with low back pain, the physician should not try to differentiate between the various elusive entities responsible for non-specific back pain but rather should focus on determining if the patient needs emergency surgery, has sciatica with signs of nerve root compression, or has an underlying medical cause of back pain (infectious, inflammatory, metabolic, tumoural, or visceral) (Table 1).

Is this a surgical emergency?

Cauda equina syndrome and an expanding vascular aneurysm are two extremely rare but important conditions to recognize since both are surgical emergencies. In the first instance, the patient will usually present with low back and/or buttock pain, associated with bilateral sciatica, neurological symptoms in the lower extremities, and urinary and/or bowel incontinence. Physical examination may show bilateral weakness, sensory losses, saddle anaesthesia, decreased reflexes in the legs, and decreased rectal tone. Diagnostic procedures (magnetic resonance imaging (**MRI**), computed tomography (**CT**) scan, or myelogram) should be performed on an emergency basis if bowel and bladder control are to be preserved. Central disc herniation is the most common cause of the syndrome, followed by tumours and epidural abscesses.

An aortic aneurysm can be responsible for a dull, gnawing back pain due to direct compression of the aneurysm on the lumbar vertebrae. They are typically seen in elderly patients, especially white men, and physical examination may reveal a pulsating abdominal mass and decreased pulses in the lower extremities. Diagnosis is most important because rupture or dissection of the aneurysm can be fatal in 15 to 70 per cent of cases in various series. In this instance, the patient presents with a sudden, excruciating tearing abdominal or back pain radiating to the groin, buttocks, or thighs with haemodynamic compromise (hypotension, tachycardia, and shock). Up to 30 per cent of ruptured aneurysms are initially misdiagnosed. Preventive surgery (before rupture or dissection) is the optimal treatment.

Does the patient have sciatica and/or neurological signs?

Sciatica can be defined as pain radiating below the knee. It is a rare symptom, being reported by only 1 per cent of patients with back pain, but its presence is usually associated with an identifiable aetiology. Typically, sciatica results from compression of the spinal nerve originating between L4 and L5 (L5 nerve root) and/or L5 and S1 (S1 nerve root) by a herniated disc, bone, or a combination of the two (spinal stenosis). Tumours, infections, or epidural haemorrhage can very rarely produce similar symptoms and signs. The pain in a patient with a herniated disc tends to be aggravated by prolonged sitting as well as any manoeuvre that increases intrathecal pressure such as sneezing, coughing, or defaecation. It is often associated with paraesthesiae and weakness in the distribution of the involved nerve.

Patients with spinal stenosis are usually older and typically complain of pain and/or paraesthesiae in one or both buttocks, thighs, and/or legs that develop on standing or walking and are relieved by 15 to 20 min rest ('neurological claudication'). These patients often walk with the trunk flexed since extension aggravates their symptoms by worsening nerve impingement. The neurological examination is most often normal or shows non-specific abnormalities, such as reduced or absent ankle reflexes. Differentiating neurological from vascular claudication can be difficult since both problems occur in the same age category. Typically, pain from vascular claudication is relieved faster with rest than that of neurological claudication.

Does the patient have an underlying medical cause for their back pain?

The history is by far the most important diagnostic step in the search of potential medical aetiologies of low back pain. A number of clues or 'red flags' should be looked for systematically. These include the presence of fever, chills, night sweats, weight loss, and nocturnal pain that should direct the clinician towards the possibility of neoplasia or infection. An insidious onset of back pain accompanied by significant early morning stiffness in a young individual suggests a spondylarthropathy and should prompt the clinician to enquire about the family history and undertake a detailed review of the ocular (conjunctivitis, iritis), cutaneous (psoriasis, mouth ulcers, balanitis, keratoderma blennorrhagica), gastrointestinal (diarrhoea, haematochezia, abdominal pain), genitourinary (urethritis), and musculoskeletal (peripheral arthritis, dactylitis, enthesitis, heel pain) systems. Risk factors for neoplasia (previous or current history of malignancy), infection (history of tuberculosis, AIDS, intravenous drug abuse, or recent genitourinary procedures), and metabolic bone diseases (previous fractures, menopause, corticosteroid

intake, history of anorexia nervosa) should also be looked for in patients suspected of having a medical problem underlying their back pain.

What are the key signs to look for in the physical examination?

A good examination of the lumbar spine and relevant nerves can be accomplished in less than 5 min if it is done systematically (Table 2). A full physical examination must be completed in patients suspected of having a medical cause for their back pain. The diagnostic utility of the many physical manoeuvres described to identify zygapophyseal and sacroiliac joint pain has been refuted when validated against diagnostic blocks with local anaesthetic. Waddell has described a number of non-organic physical signs (Table 3). When a patient has three or more of these signs, this suggests that psychological factors or secondary gains may be involved.

The classification of non-specific low back pain

In 1987, the Quebec Task Force on Spinal Disorders proposed a classification of activity-related spinal disorders in 11 mutually exclusive categories based on pain localization, neurological examination, paraclinical examinations, and response to treatment. The first four categories are divided according to duration of symptoms and work status as both of these factors can influence management (Table 4).

Who should be investigated and how?

There are now clinical guidelines available from the United States, the United Kingdom, Australia, The Netherlands, Israel, and New Zealand to help physicians manage patients with acute back pain. No such guidelines exist for chronic back pain. There is a general agreement that the initial assessment should focus on the detection of 'red flags' suggestive of a medical aetiology and that the vast majority of patients with back pain do not need any investigations. Recommendations for ordering a plain radiograph in a patient presenting with back pain include the following: age over 50, fever, weight loss, significant trauma, previous history of neoplasia, use of corticosteroids, drug or alcohol abuse, neurological symptoms and signs, particularly if widespread, night pain, morning stiffness (in which case a pelvic rather than a lumbar radiograph is recommended to detect sacroiliitis), and the persistence of pain after 1 month of conservative therapy.

All other tests should be restricted to patients in whom a medical aetiology is suspected from the history and physical examination, and patients with abnormalities on neurological examination who do not improve with conservative management. Ordering blood tests and imaging in any other situation can hardly be justified since not only are these tests unhelpful but they contribute significantly to medical costs. In addition, as many as 25 to 50 per cent of asymptomatic individuals have been shown to have abnormalities such as disc herniation on CT scans and MRI.

The sedimentation rate is the most useful blood test in patients suspected of having spinal infection since it is elevated in up to 80 per cent of cases. Neutrophilia and anaemia are also commonly seen in patients with neoplasia and infection. Laboratory evaluation of patients with osteoporosis and/or pathological fractures should include serum calcium, phosphorus, alkaline phosphatase as well as serum and urine immunoelectrophoresis (to detect myeloma), particularly if the sedimentation rate is elevated.

Many radiologists consider MRI to be the imaging modality of choice for the diagnosis of lumbar disorders. It provides a unique non-invasive means of studying the spine and is unsurpassed for imaging soft tissues. It is particularly helpful in the evaluation of spinal cord tumours, as well as infections of the spine, including discitis, epidural, and paraspinal abscesses. Computed tomography is superior to MRI for the evaluation of bony structures and therefore is the modality of choice for spinal stenosis, particularly when combined with myelography. Plain myelography is rarely used today except in patients who have contraindications to MRI or CT (claustrophobia in particular). The diagnostic accuracy of MRI, plain CT, and CT myelography is comparable for the assessment of nerve root compression due to disc herniation. While MRI is non-invasive and involves no radiation to the patient, the much lower cost of plain CT makes it an excellent choice in this context. CT-guided percutaneous biopsy is commonly used to obtain histological material from patients with tumour mass or infection.

As mentioned previously, injection studies done under fluoroscopic guidance are the only means of diagnosing back pain of discal, zygoapophyseal, or sacroiliac joint origin. When normal discs are injected with contrast material, the individual does not experience pain. A provocative discography should be considered positive only if the injection reproduces the patient's pain and no pain is experienced during the injection of adjacent discs. In a recent report, 40 per cent of subjects with chronic low back pain attending a large specialist spinal centre satisfied this strict definition and demonstrated a radial fissure on CT. Similarly, between 10 and 15 per cent of subjects report a significant improvement in their pain when their zygoapophyseal joints or their sacroiliac joints are injected with a local anaesthetic but not with isotonic saline. When taken together, these figures suggest that the anatomical source of pain can be established in as many as 70 per cent of patients with non-specific back pain by using these invasive techniques. However, the impact of this approach on patient management is unclear, since no treatment has yet been demonstrated to be effective for these specific entities.

Radionuclide bone scintigraphy with technetium-99m is helpful in conditions characterized by increased bone turnover. These include bone metastases, fracture, Paget's disease, and infections. Gallium-67 binds to polymorphonuclear leucocytes and can he helpful in the evaluation of vertebral osteomyelitis and sacroiliac septic arthritis. Typically, bone scans are negative in patients with multiple myeloma which is characterized by lytic lesions.

Neurophysiological studies are rarely indicated except in patients in whom it is difficult to distinguish between a neuropathy, radiculopathy, or plexopathy. Fibrillations in the paraspinous muscles are the most common and earliest findings seen in radiculopathy. Their presence indicates a lesion proximal to the vertebral foramen and excludes a plexopathy.

How best to manage patients with low back pain?

Surgical emergencies

As mentioned earlier, cauda equina syndrome and a ruptured vascular aneurysm are the only two conditions that must be managed surgically on an emergency basis.

Sciatica and neurological deficits

About 90 per cent of patients with a herniated lumbar disc will improve significantly with limited rest, analgesics, and anti-inflammatory drugs. The role of epidural steroids remains unclear. They may afford short-term improvement in leg pain but they do not reduce the need for surgery. Indications for surgery include persistent disabling buttock and/or leg pain despite 2 to 3 months of conservative management, and/or severe or progressive worsening neurological deficit whilst on treatment. Surgery may also be indicated in patients with neurological claudication due to spinal stenosis, but only after all attempts with conservative management have failed. Patients with spinal stenosis who are more incapacitated by back pain than by neurological claudication probably should not be operated on, since surgery is rarely effective and may even worsen back pain.

Medical back pain

Primary and secondary tumours of the spine can be treated by surgery, radiotherapy, or chemotherapy, while antibiotics with or without surgical drainage are the treatment for discitis and osteomyelitis. Postural exercises and non-steroidal anti-inflammatory drugs remain the cornerstone of treatment for patients with spondylarthropathies. While the efficacy of most non-steroidal anti-inflammatory drugs has been demonstrated, indomethacin and phenylbutazone are the two most effective in resistant cases (because of the risk of agranulocytosis and aplastic anaemia, in the United Kingdom phenylbutazone is only available for the treatment of ankylosing spondylitis when other therapy has failed or is unsuitable). Sulfasalazine and methotrexate are helpful for the peripheral arthritis associated with spondylarthropathies but they have no role in the treatment of the spinal disease. The treatment of metabolic bone diseases is beyond the scope of this chapter.

Non-specific low back pain

A number of systematic reviews of randomized controlled trials of the most common interventions have recently been published and form the basis of the recommendations found in the many guidelines published in the past 10 years. For acute back pain, patients should be advised to stay as active as possible. There is strong evidence for the effectiveness of analgesics, non-steroidal anti-inflammatory drugs, and muscle relaxants. Exercise therapy in the acute phase is ineffective. A very important objective at this stage is to reduce the likelihood of patients progressing to chronicity, not least because there are only a few modalities, including manipulation, back schools (programmes using cognitive, physical, and motivational methods to educate patients on how to manage their back problem), and exercise

therapy that have been show to be beneficial in patients with chronic back pain. The early identification of psychosocial risk factors or 'yellow flags' is essential. Screening questionnaires have been developed to help clinicians with this task. Cognitive and behavioural approaches must be used in high-risk patients in an attempt to influence positively some of these factors. Patients with persistent back pain after 6 months represent a very difficult therapeutic challenge, particularly if they have not returned to work. At this stage, their chance of going back to their previous job is only 50 per cent, while after 1 year of absenteeism it decreases to 25 per cent. Health professionals have a major role to play in preventing this unfortunate outcome.

Neck pain

Neck pain is a very common symptom. In a recent large epidemiological survey from Norway, 34.4 per cent of adult respondents reported troublesome neck pain in the previous year, with 13.8 per cent reporting pain lasting more than 6 months. As for low back pain, neck pain can rarely be attributed to a specific anatomical source and the vast majority of patients presenting with this symptom should be diagnosed as suffering from 'non-specific neck pain' or 'cervical spinal pain of unknown origin', rather than applying non-validated diagnostic labels. Trauma, in particular acceleration–deceleration (whiplash) injuries, increasing age, lower education, and psychosocial factors are the most common risk factors associated with the development of neck pain.

The clinical approach to the patient with neck pain should follow the same principles as described for low back pain. Signs of nerve root and/or spinal cord compression should always be looked for, particularly in patients complaining of associated pain, numbness, or weakness in their upper and/or lower extremities. Older patients with cervical spinal stenosis due to severe osteoarthritis may present with wasting and lower motor neurone weakness in the arms or hands and spastic weakness and sensory disturbance in the legs.

A number of diseases from the pharynx (pharyngitis, retropharyngeal abscess), larynx (laryngitis), trachea (tracheitis), thyroid (acute thyroiditis), lymph nodes (lymphadenitis), carotids (carotidynia), lungs (Pancoast tumor), heart (myocardial infarction), pericardium (pericarditis), aorta (dissecting aneurysm), and diaphragm (subphrenic abscess) can refer pain to the neck and should be considered. These conditions will usually have other clinical manifestations to alert the physician to the proper diagnosis. The neoplastic, infectious, inflammatory, and metabolic conditions enumerated in Table 1 can also affect the cervical spine. In addition, rheumatoid arthritis and diffuse idiopathic skeletal hyperostosis should be considered in the differential diagnosis as both can involve the cervical spine and cause spinal cord compression.

A special task force recently proposed a classification of cervical disorders associated with whiplash injuries which takes into account both the severity and duration of symptoms (Table 5). Although the classification was designed to address problems related to whiplash injuries, it can be very useful in classifying and guiding management of patients presenting with non-specific neck pain unrelated to trauma.

Investigation of patients with neck pain—who and how

Guidelines are only available for patients presenting with whiplash injuries. Patients with grade I whiplash-associated disorder do not usually require radiographic evaluation. Those with grade II to IV whiplash-associated disorder need a baseline radiological examination consisting of plain films with anteroposterior, lateral, and open-mouth views. Radiographs are usually unhelpful in patients with non-specific neck pain. Degenerative changes in the discs and zygoapophyseal joints increase with age and do not correlate with symptoms of neck pain. CT is helpful for evaluating the bony structures of the neck but it must be combined with myelography to adequately visualize the neural tissues. Therefore MRI is usually preferred in most cases with spinal cord or nerve root compromise. Fifty per cent of patients with chronic neck pain after motor vehicle accidents respond to diagnostic zygoapophyseal joint injection, suggesting that these joints are responsible for their pain.

Management of patients with neck pain

The majority of treatments recommended for the management of patients with neck pain have not been evaluated in a scientifically rigorous manner. Those that have been have shown very little, if any, evidence of efficacy. These include soft cervical collars, zygoapohyseal joint injections, and acupuncture. Patients with acute neck pain should be encouraged to maintain their usual level of activity. There is evidence that non-narcotic analgesics, non-steroidal anti-inflammatory drugs, mobilization, and manipulation are effective, while the promotion of rest and soft collars tends to prolong disability. Surgery is only indicated for patients with severe radiculopathy not responsive to 2 to 3 months of conservative management. There is no consensus as to how to best manage patients with chronic neck pain.

Regional pain disorders

Regional musculoskeletal pain disorders, defined as painful conditions in a specific region of the body, are extremely common occurrences. A number of clinical entities have been described for the shoulder, elbow, wrist and hand, hip, knee, ankle, and foot regions (Table 6). Most of the regional pain disorders can usually be identified through a careful history and directed physical examination, although recent research indicates that interobserver diagnostic agreement is only moderate for the conditions related to the shoulder region, particularly in patients complaining of severe or chronic pain, and those with bilateral involvement. Paraclinical investigations are not usually required for the diagnosis of most regional pain disorders.

In a patient presenting with regional pain, one should aim to determine whether the pain has its origin in the bones and joints, periarticular soft tissues (tendons, bursa, and fascia), nerve roots and peripheral nerves, or blood vessels or if it is referred from distant musculoskeletal or visceral structures. Lesions of the periarticular soft tissues account for most causes of regional pain disorders. Plain radiographs are helpful in delineating soft tissue calcifications which may or may not be related to the pain presented by the patient. Ultrasonography and MRI are of equal value in confirming a diagnosis of tendon rupture in the shoulder, knee, or ankle regions.

The principles of management include temporary rest, analgesics or non-steroidal anti-inflammatory drugs, local corticosteroid injections, thermal modalities, orthotics, and graded flexibility and strengthening exercises.

Diffuse musculoskeletal pain

Between 8 and 10 per cent of the adult population report suffering from chronic diffuse musculoskeletal pain and about half of these satisfy the classification criteria for fibromyalgia. The aetiology of fibromyalgia is unknown. Patients who seek medical help suffer from more psychological distress than those who don't. Although the pain is felt primarily in the muscles, the muscles show no histological or metabolic abnormalities other than those associated with physical reconditioning. There is evidence that substance P levels are increased in the cerebrospinal fluid of patients with fibromyalgia, thus supporting the hypothesis that the pain may be of central origin. Management which includes patient education, cognitive-behavioural approaches, regular aerobic training, and low-dose tricyclic agents is generally unsatisfactory.

Further reading

Agency for Health Care Policy and Research (1994). Acute low-back pain problems in adults. *Clinical Practice Guideline Number 14*. United States Government Printing Office, Washington, DC.

Boos N, Hodler J (1998). What help and what conclusion can imaging provide? *Ballière's Clinical Rheumatology* **2**, 115–39.

Burton AK, Waddell G (1998). Clinical guidelines in the management of low back pain. *Ballière's Clinical Rheumatology* **12**, 17–35.

Carette S *et al*. (1997) Epidural corticosteroid injections for sciatica due to herniated nucleus pulposus. *New England Journal of Medicine***336**, 1634–40.

Linton SJ, Halldén K (1998). Can we screen for problematic back pain? A screening questionnaire for predicting outcome in acute and subacute back pain. *Clinical Journal of Pain* **14**, 209–15.

Loney PL, Stratford PW (1999). The prevalence of low back pain in adults: a methodological review of the literature. *Physical Therapy* **79**, 384–96.

Schwarzer AC *et al*. (1994) The relative contribution and clinical features of internal disk disruption in patients with chronic low back pain. *Spine* **19**, 801–6.

Schwarzer AC *et al*. (1995). The prevalence and clinical features of internal disk disruption in patients with chronic low back pain. *Spine* **20**,1878–83.

Schwarzer AC, Aprill CN, Bogduk N. (1995) The sacroiliac joint in chronic low back pain. *Spine* **20**, 31–7.

Spitzer WO (1987). Scientific approach to the assessment and management of activity-related disorders: a monograph for clinicians. Report of the Quebec Task Force on Spinal Disorders. *Spine* **12** (Suppl. 7), 1–59.

Spitzer WO *et al*. (1995). Scientific monograph of the Quebec Task Force on Whiplash-associated Disorders: redefining 'whiplash' and its management. *Spine* **20** (Suppl. 8), S1–73.

Van Tulder MW, Koes BW, Bouter LM (1997). Conservative treatment of acute and chronic nonspecific low back pain. *Spine* **22**, 2128–56.

Wolfe F *et al*. (1990). The American College of Rheumatology 1990 criteria for the classification of fibromyalgia. *Arthritis and Rheumatism* **33**, 160–72.

18.5 Rheumatoid arthritis

R. N. Maini

Historical background

The first clinical description of rheumatoid arthritis in the medical literature is generally accorded to Landry-Beavais (1800), although Garrod was the first to use the term in his book published in 1859. Whether rheumatoid arthritis existed in Western Europe in olden times is debated by scholars of medical history: descriptions of chronic deforming arthritis suggestive of rheumatoid arthritis in classical writings of Galen and others have, for example, been ascribed to chronic polyarticular gout. The suggestion has been made that rheumatoid arthritis was imported to Europe after the discovery of the New World in the fifteenth century, where it pre-existed as evidenced by examination of archaic Amerindian skeletal remains. The possibility that rheumatoid arthritis spread from the New World in recent times is not only of historical interest but has also led to speculation suggesting the importance of environmental factors in its causation.

The concept of rheumatoid arthritis as a disease entity continues to evolve with advances in knowledge of the multiple causes of chronic inflammatory joint diseases. Thus improved microbiological, immunological, and epidemiological methods have led to a reclassification of certain forms of chronic arthritis, which in the past may have been labelled as rheumatoid arthritis. These include arthritis caused by infections, such as rubella, parvovirus, and borrelia (Lyme disease), or due to biological responses to non-viable products of micro-organisms ('reactive' arthritis) such as yersinia, salmonella, and chlamydia. Diseases of uncertain aetiology, for example the spondyloarthropathies, sarcoidosis, and chronic arthritis associated with systemic lupus erythematosus, primary Sjögren's syndrome, and other connective tissue diseases, have all been recognized as distinct from rheumatoid arthritis in the relatively recent past.

Definition

Rheumatoid arthritis has been defined as a chronic systemic inflammatory disorder characterized by deforming symmetrical polyarthritis of varying extent and severity, associated with synovitis of joint and tendon sheaths, articular cartilage loss, erosion of juxta-articular bone, and in most patients, the presence of IgM rheumatoid factor in the blood. In some patients systemic and extra-articular features may be observed during the course of the disease, and rarely prior to joint disease. These include anaemia, weight loss, vasculitis, serositis, mononeuritis multiplex, interstitial inflammation in lungs and exocrine salivary and lacrimal glands, as well as nodules in subcutaneous, pulmonary, and scleral tissues.

The American College of Rheumatology (**ACR**) has developed and revised criteria for the classification of rheumatoid arthritis based on a hospital population of patients with established active disease (Table 1). These combine a constellation of clinical, serological, and radiological features and have become widely accepted for epidemiological and clinical studies. These criteria distinguish active rheumatoid arthritis from other forms of inflammatory arthritis with a diagnostic sensitivity and specificity of about 90 per cent. However, they are of less value in prevalence studies, which should ideally include patients with inactive rheumatoid arthritis.

The classification criteria are too restrictive to diagnose rheumatoid arthritis reliably early in its presentation, since not all the required features may be present at this stage of evolution. Moreover, a minority of patients presenting with polyarthritis who satisfy the classification criteria for rheumatoid arthritis may later differentiate into other disease types or follow a self-limiting course.

Epidemiology

Criteria and methods for diagnosis of rheumatoid arthritis have varied in different epidemiological studies: some have been based on retrospective analysis of hospital records and others on prospective observation of patients attending hospitals where clinical examination, rheumatoid factor tests, and radiography have been employed. Questionnaires and clinical examination, with or without tests for rheumatoid factor and radiography, have also been used in population studies. However, in recent years the more widespread use of ACR criteria, including a version with a modified format for use in population studies, has introduced a measure of standardization. Based on studies carried out in various parts of the world, some generalizations can be made about the occurrence of rheumatoid arthritis amongst different ethnic populations.

Given the inherent variability in the methodology employed it is not surprising that estimates of the incidence of rheumatoid arthritis in the United States and Europe vary. In a recent study the incidence was 54 per 100 000 in women and 24.5 per 100 000 in men. The incidence increased sharply to a maximum in women over the age of 45 and in men continued to rise into the seventh decade (Fig. 1). A declining trend in the incidence of rheumatoid arthritis amongst females has been observed in recent years.

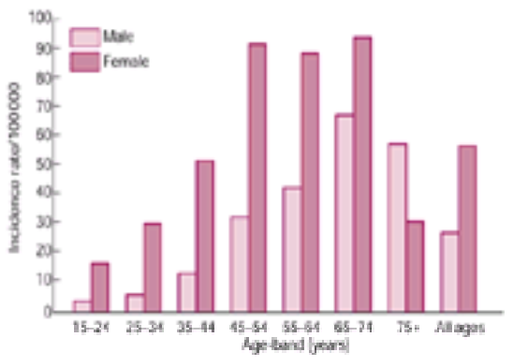

Fig. 1 Age-specific incidence applying modified ACR criteria for rheumatoid arthritis in a United Kingdom population registered in 1990 in whom multiple assessments were made over a 5-year period. Any one of seven criteria may be positive only once during this period. Data from: Wiles N *et al*. (1999). Estimating the incidence of rheumatoid arthritis: Trying to hit a moving target? *Arthritis and Rheumatism* **42**, 1339–46.

The prevalence of rheumatoid arthritis has been consistently assessed as being between 0.8 and 1.1 per cent of the adult population in cross-sectional studies in United States and Western Europe, and translate into higher prevalence rates in the elderly female population. Lower rates of 0.2 to 0.3 per cent have been reported in China and Japan. The prevalence of rheumatoid arthritis amongst the black population is low in rural South Africa (approximately 0.2 per cent) and virtually non-existent in parts of Nigeria. By contrast, prevalence rates of almost 1 per cent have been observed amongst black populations in urban South Africa and in the United States. A strikingly high prevalence rate of over 5 per cent has been noted amongst certain American Indian tribes in the United States, for example the Pima and Chippewa Indians. Differences in both genetic and environmental factors are likely to impact on these variations in incidence and prevalence rates, as are differing access to medical facilities, population age structures, and mortality. Such data are therefore of limited value in providing direct insight into aetiology, but are invaluable in directing research questions and allocating health resources.

Aetiology

Genetic factors

The initiating cause of rheumatoid arthritis remains unknown. A prevalence of 12 to 15 per cent in genetically identical (monozygotic) twins observed in Finland and Great Britain, compared with 4 per cent in non-identical (dizygotic) twins, and between 0.5 to 1 per cent in the general population, strongly favours multigenic influences. It also argues for an environmental trigger.

Advances in molecular genetics have permitted genotyping to confirm an association between the occurrence of rheumatoid arthritis and allelic polymorphisms of genes on the short arm of chromosome 6 that code for a hypervariable region of the b chain of HLA-DR molecules. The critical expressed pentapeptide sequence (glutamine–arginine or lysine–arginine–alanine–alanine) of amino acid residues 70 to 74 has been located to the helical wall of the antigen-binding cleft of the HLA-DRb chain by molecular structural studies (Fig. 2). This pentapeptide region is also referred to as the 'shared epitope' because of its detection by a specific monoclonal antibody. The sequence is present, for example, on HLA-DR4 subtypes Dw4 and Dw14, and HLA-DR1 subtype Dw1, coded by DRB1*0401,*0404, and 0101 genes, respectively. The shared sequence and corresponding allelic genes have been detected in a frequency of up to 90 per cent in patients with rheumatoid arthritis of Western European descent. Their association with rheumatoid arthritis supports the hypothesis that these particular HLA-DR molecules present antigens to T-cell receptors and activate pathogenic reactions.

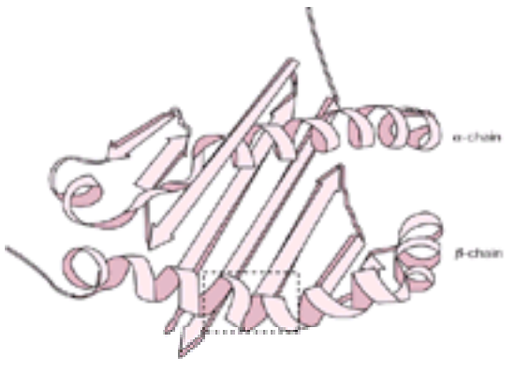

Fig. 2 Ribbon diagram of HLA class II molecule, demonstrating the antigen-binding cleft. The floor consists of a b pleated sheet and the walls are helical structures. The rectangle delineates the hypervariable region of the b chain containing the shared epitope (amino acid residues 70 to 74).

By contrast, HLA-DR4 subtypes, of which Dw10 and Dw13 are examples (coded by DRB1*0402 and *0403 genes, respectively), are negatively associated with rheumatoid arthritis. These subtypes are characterized by a substitution of the basic amino acids glutamine and arginine in positions 70 and 71 by acidic amino acids aspartic and glutamic acid. These alterations are sufficient to alter the specificity of binding such that a different set of antigens binds to the HLA cleft. Another possibility is that specific shared epitope sequences influence T-cell receptor interactions with the HLA-DRb alpha helix independently of peptide . It is proposed that signals delivered to T cells are therefore different and lead to the activation of regulatory pathways that serve a protective function.

It is estimated that major histocompatibility genes confer 30 to 50 per cent of the genetic component of susceptibility to rheumatoid arthritis. However, the presence of DRB1*04 susceptibility genes also correlates with seropositive, erosive, and extra-articular disease. Homozygosity for DRB1*0401 or DRB1*0404, or when they are combined with each other or with DRB1*0101 (as compound homozygotes), appears to correlate with more severe disease, increasing the absolute risk ratio up to 1 in 7 (relative risk of 49). These data have been interpreted as indicating that the shared epitope encoding genes may be more useful as markers of disease severity in established rheumatoid arthritis than as markers of disease susceptibility.

It is intriguing to note that the HLA-DR B1 allele *0405—coding the shared epitope in a different HLA-DR4, Dw15 subtype—is increased in Japanese patients, whilst the DRB1*1402 gene coding HLA-DR6, Dw16 is increased in Yakima American Indians. However, other population studies, for example in black American individuals with rheumatoid arthritis, show no increase in frequency of the gene coding the shared epitope, thus casting some doubt on the hypothesis that it is an essential aetiological factor.

Recent studies have sought positive or negative correlation between disease severity and gene polymorphisms detected by nucleotide sequencing or microsatellite mapping. Using such techniques, associations with polymorphic alleles of candidate genes coding molecules involved in the pathogenesis of rheumatoid arthritis, such as tumour necrosis factor-a (**TNF-a**), a pro-inflammatory cytokine, and interleukin 10 (**IL-10**), an anti-inflammatory cytokine, have been described.

Environmental factors

The similarity of clinical features of rheumatoid arthritis and polyarthritis caused by infectious agents such as rubella, parvovirus B19, and Epstein–Barr virus (**EBV**), and reports of immune hyperactivity to their antigens in rheumatoid arthritis, continues to fuel interest in a potential role for such organisms in initiating rheumatoid disease. In one recent study, for example, the B19 antigen VP-1 was specifically expressed in active lesions in synovium with rheumatoid arthritis but not in osteoarthritis or controls. In other studies, EBV-specific or rubella-specific lymphocytes have been detected in joints with rheumatoid arthritis. However, the arthritis caused by such known infections is almost always sporadic and self-limiting, and the clustering of new cases that one might expect if rheumatoid arthritis was an infectious disease has not been reported. Moreover, corroboration of claims by independent studies is still lacking, hence these theories of causation remain

speculative.

In attempts to define a role for environmental factors in the aetiology of rheumatoid arthritis, epidemiological studies have sought differences in the incidence or prevalence of the disease in genetically similar populations exposed to urbanization, different socio-economic conditions, lifestyles, and known industrial noxious agents. In South Africa, one study found a higher point prevalence of rheumatoid arthritis in Bantu people residing in an urban township area—similar to that recorded for white people in Western countries—compared with Bantu in a rural community. However, in another study, black people living in urban Manchester, United Kingdom, had a lower prevalence of rheumatoid arthritis than white people and the low prevalence of rheumatoid arthritis amongst Chinese people living in urban Hong Kong was no higher than that observed in rural areas. These apparently contradictory data reflect the problems inherent in assessing the relative importance of genetic and environmental factors in heterogeneous populations in a disease with variable expression. The data on correlation with social class are also conflicting. Cigarette smoking, on the other hand, was associated with an increased risk of rheumatoid arthritis in two prospective population studies and in a twin study. Claims of an increased risk have also been made in people exposed to silica dust, organic solvents, and mineral oils.

It is possible that a decline in the incidence of rheumatoid arthritis amongst women noted in Rochester, United States, in the period 1950 to 1975 and in a general practice register in the United Kingdom in the decade following 1976 are indicative of a change in environmental pressures or in lifestyle. Epidemiological studies demonstrating a relationship between the birth weight of babies and future development of cardiovascular disease and premature death have drawn attention to the possibility that the environment of the growing fetus may be as important as environmental factors to which an adult might be exposed.

Host factors

A number of observations implicate sex hormones and prolactin in susceptibility to, or protection from, rheumatoid arthritis. Thus females have a higher incidence of rheumatoid arthritis, which is especially marked before the menopause. Exposure to the oral contraceptive pill confers a level of protection and postpones the onset of rheumatoid arthritis. Pregnancy is associated with suppression of disease, and the incidence of rheumatoid arthritis is increased following parturition and during lactation. Testosterone levels are reported to be low in males with rheumatoid arthritis, and the incidence of disease increases with advancing age when levels of male sex hormones are on the decline. Interconnections between the hypothalmic–pituitary axis, hormones, and cytokines have been described, suggesting possible mechanisms whereby these may influence the evolution of rheumatoid arthritis.

Pathology and pathogenesis

Pathology

Joints

The rheumatoid disease process in the joints is characterized by synovitis, an inflammatory effusion and cellular exudate into the joint space, and by damage to tendons, ligaments, cartilage, and bone in and around articulating surfaces of the joint. Long tendons whose sheaths are lined by synovial membrane, such as in the palms, wrists, ankles, and feet, may also be involved by the inflammatory process and cause malfunction due to damage, rupture, and fibrosis.

In health the synovial membrane (the intima) is a film of one or two cells lining the capsule and its circumferential attachment to the periosteum at the cartilage–bone junction of the joint (Fig. 3). The normal synovial membrane consists of type A and B cells, without a basement membrane and lying on a bed of loose connective tissue and a network of small blood vessels (the subintima). Type A cells have morphological and phenotypic features of macrophages. Type B cells are of mesenchymal origin and share many, but not all, of the phenotypic features of typical tissue fibroblasts and are hence referred to as fibroblast-like synoviocytes.

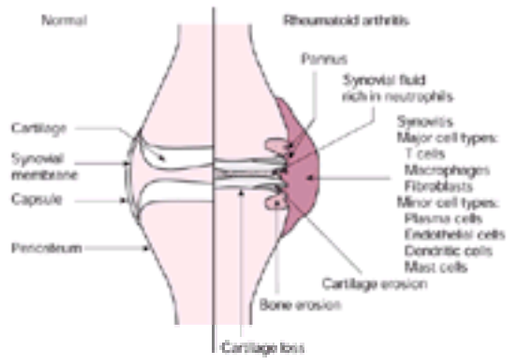

Fig. 3 The pathology of rheumatoid arthritis, normal joint (left) and rheumatoid joint (right).

In established rheumatoid arthritis the synovial membrane typically becomes enormously thickened and assumes a villous appearance. The diseased tissue now consists of an intima that is several (2 to 10) cell layers deep and coated by a film of fibrin. Type A cells predominate over type B cells and tend to lie in the more superficial part of the intima. The sublining layer (subintima) is also greatly expanded by newly formed blood vessels and infiltrating mononuclear cells, including T lymphocytes, lymphoblasts, B cells, plasma cells, monocytes, macrophages, dendritic cells, and synoviocytes (Fig. 3). The cellular infiltrate usually has a recognizable architecture, comprising perivascular aggregates of CD4+ T cells (Fig. 4 and Plate 1). Interaggregate areas show a mixed inflammatory cell population, including dendritic cells and macrophages expressing HLA class II, CD8+ T cells, activated B cells, and plasma cells. The aggregates may be organized into prominent lymphoid follicles, some of which display germinal centre formation.

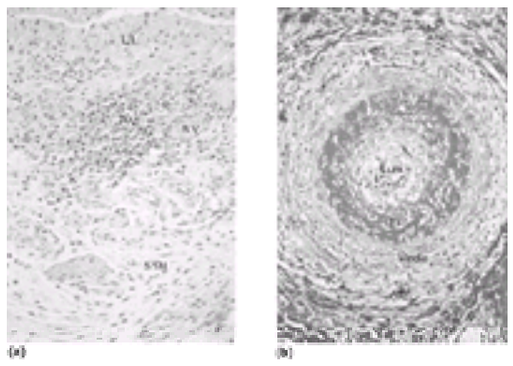

Fig. 4 Histology of rheumatoid arthritis. (a) Rheumatoid arthritis synovitis. L.L., lining layer; P.V., perivascular aggregate of lymphocytes and macrophages; B.V., blood vessel; SYN, synoviocytes; (haematoxylin and eosin staining). (b) Small vessel arteritis. Lum., lumen; Int., Intima; P.V., perivascular inflammation; Adv., adventitial tissue. Arterial wall shows a thrombosed vessel with intimal hyperplasia, destruction of internal elastic lamina, and mononuclear cell infiltration of media and perivascular tissue (methylene blue and safranine staining) (See also Plate 1).

The surface of the thickened synovial membrane is bathed in an inflammatory synovial fluid containing a predominance of polymorphonuclear cells, but also CD4+ and CD8+ lymphocytes, dendritic cells, macrophages, and synoviocytes. The synovial fluid is rich in pro-inflammatory cytokines and immune complexes containing rheumatoid factor. It is a site of local complement consumption, resulting in low haemolytic complement activity, low C3 and C4, and increased complement breakdown products.

The destructive lesion in the joint typically occurs at the circumferential attachment of the joint capsule, just below and adjacent to the articular cartilage and subchondral bone. Here the intima of the adjacent hypertrophic synovial membrane creeps over the cartilage, and tissue rich in blood vessels, macrophages, and synoviocytes (termed 'pannus') invades and destroys variable parts of articular cartilage and subchondral bone. The cells at the cartilage–pannus junction consist of synoviocytes and macrophages, whereas the pannus invading subchondral bone is enriched in osteoclasts. The connective tissue matrix of cartilage adjacent to pannus tissue becomes depleted of proteoglycans and collagen type II as a result of enzymatic degradation and lack of regeneration. A number of enzymes responsible for degradation of cartilage matrix have been demonstrated in the joint with rheumatoid arthritis, including the collagen-degrading matrix metalloproteinases I, III, and XIII, neutrophil-derived cathepsins L and D, and collagenase, as well as aggrecanase that degrades proteoglycans. The matrix in which chondrocytes are embedded becomes depleted, with loss of chondrocyte numbers, suggesting that matrix degeneration is secondary to degradative effects mediated by both pannus and chondrocyte activity or cell death. There may also be a reparative response in the later stages of disease, as suggested by the presence of fibrous tissue replacing areas of destroyed cartilage and bone in some joints removed at surgery.

Extra-articular disease

Extra-articular features associated with rheumatoid arthritis comprise essentially two types of lesion: the first involves arterial walls and the second leads to extravascular lymphocyte–macrophage granuloma formation.

Of those lesions involving arterial walls, two types of pathology are described. The first type is a bland fibro-intimal hyperplasia, without obvious inflammatory changes, resulting in vascular occlusion. This lesion is typically observed in digital vessels of patients with long-standing disease and is associated with collateral blood vessel formation. It correlates with a history of benign, intermittent nail-fold infarcts that develop in winter months. By contrast, the second type of lesion has a polyarteritic pathology and is observed in patients with rheumatoid systemic vasculitis and a poor prognosis. Medium- and small-sized arteries of the limbs, peripheral nerves, and organs are involved, but renal vessels are spared. Histopathological examination of involved vessels reveals lymphocytic, histiocytic, and inflammatory cell infiltration of the medial and perivascular area, disruption of the internal elastic lamina by fibrinoid necrosis, and proliferation of the vessel wall intima with intravascular thrombosis and occlusion (Fig. 4(b)).

Extravascular nodule formation in areas subject to pressure or friction is the characteristic granulomatous lesion of rheumatoid arthritis. Nodules consist of a central core of fibrinoid eosinophilic material surrounded by a palisade of histiocytes, occasional giant cells, and an outer layer of lymphocytes, fibroblasts, and fibrous tissue. Extravascular granulomatous inflammation, with or without nodule formation, has been documented on the surface of the pleura, pericardium, and endocardial valves. As is the case with systemic vasculitis and Felty's syndrome, the occurrence of nodules correlates with seropositive disease and the carriage of HLA-DRb1*04 alleles.

Pathogenesis

Although the initiating cause of rheumatoid arthritis remains uncertain, there has been considerable progress in understanding the cellular and molecular mechanisms involved in chronic inflammation and tissue damage (Fig. 5).

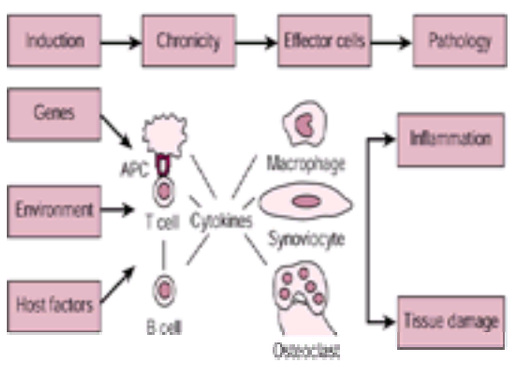

Fig. 5 Aetiopathogenesis of rheumatoid arthritis. In this simplified diagram the four steps in the aetiopathogenesis of rheumatoid arthritis are shown: (a) the induction phase involving interactions among genetic, environmental, and host factors; (b) the chronicity phase dependent upon immunological reactions; and (c) the effector phase of pathology mediated by macrophages, synoviocytes, osteoclasts, and cytokines, which leads to (d) pathology, namely inflammation and tissue damage.

The discovery of rheumatoid factor in the blood of patients with rheumatism over half a century ago led to the immunological hypothesis of disease pathogenesis. Since rheumatoid factor is an autoantibody directed against epitopes on the constant domains of the Fc portion of IgG1, the concept that rheumatoid arthritis is an autoimmune disease gained credibility. However, IgM rheumatoid factor occurs in a variety of other diseases in the absence of joint pathology. In the case of rheumatoid arthritis, rheumatoid factor complexes are present in synovial fluids, and IgG-producing B cells, whose rearranged immunoglobulin gene sequences implicate antigen stimulation, are present in inflamed synovium. B cells in rheumatoid joints also synthesize antibodies to some cartilage components such as collagen type II, although these are not disease specific. By contrast, recent research has shown a highly disease-specific antibody directed against citrullinated peptides in the serum of patients. It seems possible that autoantibodies could interact with complement and Fc receptors expressed on cells in the rheumatoid joint and so contribute to inflammation. Since IgG antibodies are implicated, T-cell help might be required.

The predominance of CD4+ T cells in proximity to antigen-presenting cells in the rheumatoid joint suggests their involvement in perpetuating the immune response. This is also supported by the association of rheumatoid arthritis with HLA class II genes and the beneficial response to T-cell-depleting therapies (lymphophoresis, cyclosporin, and some anti-CD4 monoclonal antibodies). For example, it has been proposed that as a result of molecular mimicry between epitopes on infectious agents and endogenous antigens an immune response is initiated which, by a phenomenon known as 'epitope spreading', overcomes tolerance mechanisms and results in autoimmunity. Alternatively, macromolecules in cells undergoing apoptopic cell death may become modified by oxidative or enzymatic damage and provide neoantigens to which autologous T cells are not tolerant. Candidate autoantigens under current investigation include proteins derived from cartilage, citrullinated peptides, and peptides derived from HLA class II molecules. However, an autoantigen that drives the immune response in rheumatoid arthritis has not yet been identified.

CD4+ T cells in joints appear to be of T_{H1} type, bearing memory and activation markers such as CD45 RO+, CD45B dim+, VLA-4+, CD69+, and HLA class II+. However, whether these T cells are antigen activated is debated since they do not appear to show significant T-cell receptor oligoclonality and do not display the full functional characteristics of such cells, for example the production of IL-2 and interferon-g. Moreover, T cells in joints do not proliferate but increase in number by recruitment and accumulation. It seems possible that T cells are conditioned and activated by cytokines such as IL-15, TNF-a, IL-6, and IL-10 produced by rheumatoid tissues. Cell membrane contact between macrophages and cytokine-activated T cells may also be a key event in driving production of the pivotal cytokine TNF-a (see below).

Cytokines are protein messenger molecules that transmit signals from one cell to another by binding to their specific receptors on the surface of cell membranes. Their activity is usually restricted to adjacent cells in the local milieu. Cytokines are normally produced and exported as soluble molecules into the fluid phase, although some cytokines, such as TNF-a, are also active as molecules displayed on the surface of the producer cells. Expression of mRNA and protein of a large number of cytokines is reported in rheumatoid synovial tissue. These molecules regulate a diverse range of functions relevant to an understanding of the pathogenesis and clinical features of rheumatoid disease. Both pro- and anti-inflammatory cytokines, chemokines, and mitogenic factors are produced, but pro-inflammatory mediators predominate during active phases of disease.

Of the pro-inflammatory cytokines, IL-1 and TNF-a are of key importance in the pathogenesis of rheumatoid arthritis (Fig. 6). They are intimately involved in activation of the cytokine network, leucocyte recruitment and activation, the local immune response, angiogenesis, and fibroblast proliferation. IL-1 and TNF-a also regulate production of a number of mediators of connective tissue damage by synoviocytes, including matrix metalloproteinases and prostaglandins. Furthermore, these cytokines activate osteoclasts that are implicated in bone damage. TNF-a is produced mainly by type A cells of macrophage lineage in the intima, subintima, and the cartilage–pannus junction. The p55- and p75-TNF receptors are coexpressed by cells in the vicinity. The hypothesis that TNF-a is a dominant pro-inflammatory

mediator in the cytokine dysequilibrium observed in the rheumatoid synovium has gained considerable support. In particular, TNF-a regulates production of IL-1 and together these two cytokines orchestrate rheumatoid inflammation and damage. The identification of TNF-a as a molecular target for therapy has been validated by clinical trials of biological inhibitors of TNF-a and their recent application in rheumatological practice.

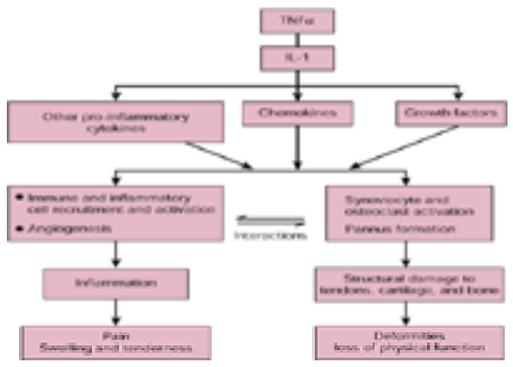

Fig. 6 The role of tumour necrosis factor-a in the pathogenesis of inflammation and structural damage to tendons and bone in rheumatoid arthritis.

Examples of the importance of TNF-a as a mediator of rheumatoid disease include the following observations after treatment with infliximab, a monoclonal anti-TNF-a antibody.

1. There is a reduction in the cellularity of the rheumatoid synovial membrane associated with a reduction in the number of tissue macrophages and lymphocytes.
2. There is evidence that the reduced cellularity is associated with a reduction of cytokine-induced vascular adhesion molecules—intracellular adhesion molecule-1 (ICAM-1), E-selectin, and vascular cell adhesion molecule-1 (VCAM-1)—as well as chemokines such as IL-8 and monocyte-chemoattractant protein-1 (MCP-1).
3. There is reduction in synovial vascular density and neovascularization (avb3 staining of blood vessels), associated with a fall in the concentration of circulating vascular-endothelial growth factor, a major cytokine implicated in new vessel formation.
4. There is direct evidence of reduced retention in joints of autologous ^{111}indium-labelled polymorphonuclear cells (Fig. 7).
5. There is a reduction of serum IL-6 concentrations within 6 h of a single infusion of the anti-TNF-a antibody, followed within 24 h by a reduction in markers of acute-phase response including C-reactive protein, serum amyloid A protein, and erythrocyte sedimentation rate.
6. There is a reduction in the circulating concentrations of pro-matrix metalloproteinase-1 (pro-MMP-1) and pro-MMP-3 in blood. Matrix metalloproteinases are implicated in cartilage and bone degradation and there is clinical trial evidence that anti-TNF-a markedly retards radiographic signs of cartilage loss and bone erosions.

Many other cytokines and chemokines are potential therapeutic targets. Clinical trials have been conducted with anti-TNF-a, anti-IL-1, anti-IL-6, and human recombinant IL-10 and IL-11. So far, anti-TNF-a has been approved for the treatment of rheumatoid arthritis, and in the United States, the IL-1 antagonist, human recombinant IL-1 receptor antagonist (IL-1ra) has been approved.

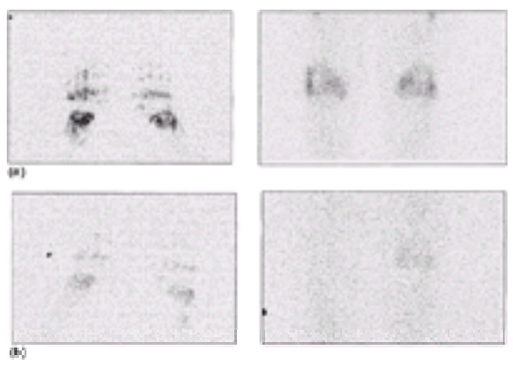

Fig. 7 Gamma camera images of the hands and knees of a patient with rheumatoid arthritis. Images were taken 22 h after a bolus injection of autologous radiolabelled (^{111}indium) granulocytes (a) before and (b) after a single 10 mg/kg intravenous bolus of anti-TNF-a antibody (infliximab). There was a reduction in signal after treatment. (Images kindly provided by P. C. Taylor.)

Clinical features

Presentation

The onset of rheumatoid arthritis is frequently insidious and the principal symptoms are pain and stiffness, mainly of peripheral joints, with associated swelling. Prolonged stiffness of joints on waking and following inactivity is usual and may last an hour or more. There is progressive decline in physical function and ability to perform daily activities. Fatigue and lethargy are common and there may also be low-grade fever and weight loss. Symptoms are persistent in affected joints, although there may be some day to day variation in severity. As the disease evolves, further joints may become involved and some may remit, but ultimately the distribution of arthritis becomes permanently established.

Other patterns of disease presentation are also recognized. Up to one-third of patients present with an explosive or subacute onset of arthritis, leading to severe immobility. In a minority of patients a migratory polyarthritis flitting from joint to joint is observed. This is referred to as 'palindromic rheumatism' and may be a recurring pattern over months before chronic polyarthritis becomes established. About 10 per cent of patients present with features of the syndrome of polymyalgia rheumatica, characterized by prominent limb-girdle pain, stiffness, and painful movement of the neck, shoulders, and hips. Persistent inflammatory arthritis of a single joint such as the knee, wrist, ankle, shoulder, or hip may be the only rheumatological symptom and can antedate the onset of polyarthritis by months or years.

In some patients bilateral diffuse swelling of the fingers and hands may be a presenting complaint, often associated with symptoms of carpal tunnel syndrome. Synovitis of tendon sheaths of the dorsal extensors of the wrist and of flexor tendons in the palm and wrist may be present with concurrent joint signs, but may also occur as a prominent clinical feature in the absence of polyarthritis. Swelling of the ankles with pitting oedema is commonly seen in active rheumatoid arthritis. Lymphoedema of the forearm or lower limb is observed less frequently.

Rarely, the initial manifestations of rheumatoid arthritis are confined to extra-articular disease. Examples include subcutaneous nodules, one or more nodules in the thorax presenting as pulmonary lesions on a chest radiograph, pleurisy with pleural effusion, pericarditis, episcleritis, and vasculitis.

Joint distribution

The expression of rheumatoid arthritis shows interindividual variation with respect to the anatomical sites and numbers of involved joints. For example, some patients have mainly small joints affected, whilst others show simultaneous involvement of small and large joints. The hip and shoulder joints may be spared in some, whilst in others they bear the brunt of the disease. The actual numbers of diseased joints can vary from three or four to over 50. Diseased neck joints may be asymptomatic until, in the late stages, neurological complications alert the physician to subluxation of the cervical spine or the atlantoaxial joint.

In over 80 to 90 per cent of patients, one or more of the metacarpophalangeal and proximal interphalangeal joints of the hand and the metatarsophalangeal joints are involved. Other frequently involved sites include the wrists, glenohumeral joints of the shoulders, knees, and the elbow joints, followed by the mid-tarsal, acromioclavicular, interfacetal, and atlantoaxial joints of the cervical spine and hip joints. The temporomandibular, sternoclavicular, and cricoarytenoid joints are involved in about a third of patients.

Symmetrical involvement of the joints is usual, but joint damage and deformity may be asymmetrical and related to overuse or traumatic injury. Conversely, neurological paralysis of a limb results in joint protection.

In addition to involvement of diarthrodial joints, the rheumatoid process frequently involves tendon sheaths of hands, wrists, shoulders, and ankles.

Features of joint disease

Hands and wrists

In active rheumatoid disease, soft tissue swelling and tenderness of metacarpophalangeal and proximal interphalangeal joints is observed (Fig. 8 and Plate 2). Thickening and nodularity of flexor tendons in the palms may be palpable and tenosynovitis can be a cause of 'triggering' of the fingers. Wasting of the interossei is prominent and fist closure restricted. Flexor tendonitis and wrist synovitis may be associated with signs and symptoms of median nerve compression (carpal tunnel syndrome).

Fig. 8 The hands of a person suffering from rheumatoid arthritis. Features to note include symmetrical soft tissue swelling of the second and third metacarpophalageal joints, early swan-neck deformity of the left ring finger, ulnar deviation at the metacarpophalageal joints, and wasting of the small muscles of the hand. In addition, several small rheumatoid nodules are present. (See also Plate 2.)

Ulnar deviation and volar subluxation of the digits and wrists may develop later. Other recognized deformities include Boutonnière (button hole) flexion deformity of the proximal interphalangeal joint and 'swan-neck' deformities of fingers due to hyperextension of the proximal interphalangeal joint and flexion at the distal interphalangeal joint.

Diffuse synovial swelling may be pronounced at the dorsal aspect of the wrist and the ulnar styloid may become dorsally subluxed. The carpus may drift in a volar direction such that supination of the hand is restricted. In this late stage, the extensor tendons appear stretched across a shrunken carpus ('the bowstring' sign). Extensor tendons may occasionally rupture, most commonly affecting the little or ring fingers.

Nail-fold and finger-tip infarcts and splinter haemorrhages indicate digital vascular occlusive disease. Palmar erythema is common but not specific for rheumatoid arthritis.

Elbows and shoulders

Physical signs in early stages include swelling, limitation of movement, and inability to flex or extend the elbow. Later, pronation and supination are restricted, and the head of the proximal radioulnar joint may dislocate. Olecranon bursitis and subcutaneous nodules around the elbow are common. In the shoulder, aside from glenohumeral joint synovitis, there may be accompanying subacromial bursitis and rotator and biceps tendon involvement.

The neck

Rheumatoid involvement of the apophyseal joints of the neck can cause pain, stiffness, and restricted movement. Loss of stability in the spine may occur at several levels and be associated with symptoms and signs of radicular or cord compression. Subluxation of the atlantoaxial joint diagnosed by plain radiography or magnetic resonance imaging occurs in 6 per cent of the rheumatoid population and up to 30 per cent of patients who are admitted to hospital. It may be asymptomatic, but when severe tends to occur in patients who also suffer from severe generalized disease and advanced disability, and is a recognized cause of quadriplegia and sudden death.

The knees

Involvement of the knees is common, and chronically active synovitis is associated with irreversible destruction and rapid deterioration in functional capacity. In early stages especially, high pressure in the knee joint on active flexion, for example during squatting, can lead to joint rupture and leakage of inflammatory fluid into the calf. This complication simulates signs and symptoms of a calf deep vein thrombosis: it can be diagnosed by arthrography using contrast medium, or by ultrasonography of the knee. A chronic effusion in the knee joint may also be associated with a posterior popliteal (Baker's) cyst and occasionally this extends into the medial aspect of the calf.

Ankles and feet

Inflammation of the metatarsophalangeal joints is common and results in subluxation of the metatarsal heads and, ultimately, claw- or hammer-toe deformities. The soft tissue pad that is normally positioned underneath the metatarsal heads becomes displaced such that the heads of the metatarsal bones become painful to walk on. Patients may describe this as feeling as if they were walking on marbles or stones. Involvement of the tarsal and subtalar joints may result in flattening of the arches of the foot and valgus deformity of the hindfoot. These deformities cause difficulties with footwear, and where shoes rub the feet there is a tendency for callosities to form.

Hips

The hips are less often involved, but there may be erosions in severe cases with remodelling of the acetabulum (protrusio acetabuli). There may also be secondary degenerative disease at the hip. Total hip replacement is generally a highly successful treatment for endstage hip disease.

Extra-articular disease

Nodules

Nodules occur in 25 to 30 per cent of patients with rheumatoid arthritis and are associated with seropositive disease. Common sites for subcutaneous nodules include

the elbow, ischial tuberosity, heel, and dorsum of fingers. Multiple, small, rapidly evolving nodules can occur in those on methotrexate treatment (Fig. 8). Nodules in the pleura may present as single or multiple round shadows on a routine chest radiograph.

Systemic vasculitis

Rheumatoid vasculitis occurs in patients with seropositive and nodular disease. It presents with a severe systemic illness characterized by fever and weight loss. Associated clinical features are consequent upon occlusion of medium- to small-sized arteries. These include Raynaud's phenomenon, nail-fold and digital infarcts, and gangrene, skin ulceration, mononeuritis multiplex, scleromalacia perforans, and occlusion of arteries to visceral organs. The latter include coronary, pulmonary, coeliac axis, and cerebral vessels. In some patients vasculitis may present as a skin rash associated with necrotizing polyangiitis of small cutaneous blood vessels.

Fibrosing alveolitis and obliterative bronchiolitis

Physiological abnormalities in lung function tests indicative of airways and interstitial disease may be present without symptoms. In a proportion of patients with rheumatoid arthritis, more frequently male than female, dyspnoea of insidious onset, physical signs, characteristic lung function abnormalities, a chest radiograph, and high-resolution computed tomography may reveal characteristic features of chronic fibrosing alveolitis. More rarely, acute pneumonitis may be the presenting feature with rapid deterioration and development of respiratory failure. Patients with fibrosing alveolitis are usually seropositive, have a high frequency of antinuclear antibodies, and may also exhibit evidence of multisystem disease, including vasculitis.

Obliterative bronchiolitis can be associated with rheumatoid arthritis. It is usually rapidly progressive, but some patients follow a chronic protracted course that may respond to corticosteroid and immunosuppressive therapy.

Serositis

Past evidence of pericardial and pleural inflammation is common at autopsy and may be discovered by imaging techniques in asymptomatic patients. Both may present with clinical symptoms, generally following a benign course with resolution associated with disease-modifying antirheumatoid drugs (**DMARDs**) or corticosteroid therapy. Rare cases of constrictive pericarditis have been reported. Typically, pleural effusions are exudates with a high protein content and cellular exudate enriched in lymphocytes, but also containing polymorphonuclear cells and macrophages. A low level of complement activity relative to blood concentrations and a low glucose concentration (usually less than 1.4 mmol/l) is of diagnostic value.

Eye complications

Scleritis, episcleritis, scleromalacia perforans, corneal melt, and keratoconjunctivitis sicca have all been described and need evaluation and treatment by a specialist.

Amyloidosis

Secondary amyloidosis due to deposition of amyloid AA fibrils in blood vessels and parenchyma of kidneys, liver, spleen, and gastrointestinal tract has been described in the tissues of 10 to 15 per cent of patients examined at autopsy, or in the blood vessels in the submucosa of rectal and gingival biopsies. Proteinuria, nephrotic syndrome, or renal failure are less common and have a poor prognosis unless detected and treated before irreversible renal failure has occurred. Effective treatment of rheumatoid arthritis with suppression of the acute-phase response with DMARDs and corticosteroids prevents progression and may reverse the disease. In patients with a continuing acute-phase response despite the standard DMARD therapy, treatment with chlorambucil is reported to be of some benefit. Imaging of radionuclide-labelled serum amyloid P protein in the spleen and kidneys may be used to monitor treatment.

Osteoporosis

Juxta-articular osteoporosis is a common feature of radiographs of affected joints and is related to local disease activity. However, decreased bone mineral density of the spine and pelvis has been described in patients with active severe rheumatoid arthritis. This is likely to reflect the response of bone metabolism to prostaglandins and catabolic cytokines such as IL-6, IL-11, and the receptor for activition of NF_b (RANK)-ligand, which increase osteoclast activity. This is distinct from immobility-associated or corticosteroid-induced osteoporosis, although these factors may be additive in individual patients. It has been suggested that increased mobility following low-dose prednisolone may be beneficial and reverse, rather than aggravate, corticosteroid-induced osteopenia.

Felty's syndrome

Felty's syndrome is characterized by a combination of seropositive rheumatoid arthritis, neutropenia, and splenomegaly. Lymphadenopathy, leg ulcers, and nodular hyperplasia of the liver have been described. Patients with severe neutropenia are liable to bacterial infections. Some patients also develop anaemia and thrombocytopenia. In a variant of Felty's syndrome, an expansion of large granular lymphocytes is found in the blood: these are cytotoxic CD8+ lymphocytes and may present as clonally expanded cell populations.

Myocardial disease

Myocardial disease due to diffuse fibrosis or granulomatous lesions is recognized in rheumatoid arthritis, although the more frequently recognized association is with coronary artery disease. Systemic vasculitis may also involve coronary vessels. Aortic incompetence due to valvular thickening and nodule formation or dilation of the ascending aorta have been described.

Neurological complications

A number of compression neuropathies may occur in rheumatoid arthritis. These include compression of the median nerve at the wrist, the ulnar nerve and posterior interosseous branch of the radial nerve at the elbow, and posterior tibial nerve at the level of the knee or ankle. It is important to recognize and confirm these neuropathies by nerve conduction studies since surgical decompression usually cures symptoms.

A mild, symmetrical, sensory peripheral neuropathy involving the hands and legs in a 'glove and stocking' distribution also occurs in rheumatoid arthritis. This is distinct from the rarer and more severe sensorimotor mononeuritis multiplex associated with wrist and foot drop and usually due to vasculitis of vasa nervora, when other features of a systemic vasculitis and extra-articular disease may be present. In some patients, however, no vascular pathology is demonstrable and the cause of axonal degeneration is not understood.

Rheumatoid involvement of the transverse ligament and odontoid process of the atlantoaxial joint may lead to posterior subluxation or upward movement of the odontoid and cause cervical cord compression. Cord compression may also occur due to rheumatoid damage at lower levels of the cervical spine. Compression is a recognized cause of tetraparesis and sudden death. Surgical stabilization of the neck can be successful but cannot always be recommended in patients with associated severe disability and poor health status.

Infections

Patients with rheumatoid arthritis are susceptible to local and systemic bacterial and opportunistic infections. Infections of joints and respiratory and urinary tracts, skin ulcers, and septicaemia are all described, and infections are one of the causes of increased mortality in rheumatoid arthritis. Endogenous disease-related immunosuppressive mechanisms are thought to play an important part. In Felty's syndrome, neutropenia compromises host defence. Drugs for treating rheumatoid arthritis such as cytotoxic and immunosuppressive agents may also be contributory factors.

Clinical course, progression, and outcome

Clinical course

Disease activity

The course of the disease activity fluctuates over time, partly due to the endogenous mechanisms of disease and partly as a result of effective therapy. Recurring periods of weeks or months of exacerbation of symptoms, described as 'flares', alternate with periods of relative quiescence of disease. In about 10 to 20 per cent of patients, the disease continues unabated throughout.

The key clinical features of disease activity in rheumatoid arthritis are pain, fatigue, stiffness of joints on waking, swelling, tenderness of joints on palpation, restriction of joint motion, and loss of physical functional capacity. Joint deformities become apparent as the disease progresses. Symptoms are assessed by taking a history in descriptive terms, but also by attempting to quantify their severity. These measurements have been incorporated into various criteria for assessment of disease activity and response to therapy, developed and validated, for example, by the American College of Rheumatology and the European League Against Rheumatism.

Swelling of joints due to synovial thickening may be detected by palpation as a 'spongy' or 'boggy' feel. Concomitant effusion can be demonstrated by fluctuation. In later stages of disease, subluxed surfaces of bones (such as the heads of metacarpals in the hands, the styloid of the ulna, and distal radius at the wrist) can give the appearance of bony swelling. Tenderness is elicited by digital pressure or squeezing of a joint. The classic signs of inflammation, such as redness and increased temperature overlying joints, are not usually prominent, although readily demonstrable by thermography. Active and passive movement of joints through their anatomical range of motion elicits restriction of movement associated with pain.

Functional capacity can be assessed by testing grip strengths using an inflatable bag attached to a sphygmomanometer, walking time over a standard distance, and by standard health assessment questionnaires (such as the Stanford questionnaire). The degree, quantity, and severity of pain is recorded as experienced by the patient, graded on a visual analogue scale of 1 to 10. The duration of morning stiffness is recorded in minutes. A 'global assessment' of disease activity on a visual analogue scale of 1 to 10 as judged by the patient and physician may also be used as a quantitative measurement of disease activity over time.

Structural damage

The rheumatoid disease process leads to structural damage to the cartilage, bone, and associated joint structures. This is cumulative and irreversible and appears to be related to the severity of inflammatory activity over time. In later stages of disease, loss of normal joint architecture and mechanical derangement also contribute to the perpetuation of symptoms and secondary inflammation. Serial radiographs of the hands and feet are employed to assess structural damage to joints.

Prognosis

The longer-term health status of patients presenting to hospital clinics with recent-onset rheumatoid arthritis has been documented in a number of studies. Functional deterioration occurs rapidly. In one study, half the patient population was moderately disabled in 2 years and severely disabled by 10 years. The most marked deterioration occurs in those patients with the most compromised functional capacity in early disease. These data are compatible with the proportion of patients at work whose disease-associated disability prevents continuing employment: around 20 per cent of patients stop work in the first 2 years, increasing to 30 per cent within 5 years, 50 per cent in 10 years, and 90 per cent prior to retirement age. A low level of manual work, job flexibility, and higher educational and psychosocial status are amongst the determinants that correlate best with the ability to continue work.

Patients with rheumatoid arthritis have a higher than expected prevalence of other serious illnesses and an increased mortality compared with the general population. In one study approximately 20 per cent reported concurrent disorders, and iatrogenic disease is not uncommon. In a 35-year follow-up study on 3501 patients, mortality was twice that of a control population, resulting in a shortening of life by 7 to 10 years. Rheumatoid arthritis itself may contribute to premature death in up to 20 per cent of patients as a result of complications such as fibrosing alveolitis, vasculitis, secondary amyloidosis, cardiac disease, or transection of the cord due to cervical spinal subluxation. More frequently, death is the consequence of comorbid conditions or complications of therapy. These include infection, gastrointestinal haemorrhage or perforation, cardiovascular and cerebrovascular disease, renal failure, and lymphoproliferative diseases.

Survival rates of about 50 per cent at 5 years have been recorded in a subset of patients with polyarticular disease, poor functional status, or extra-articular disease (Fig. 9).

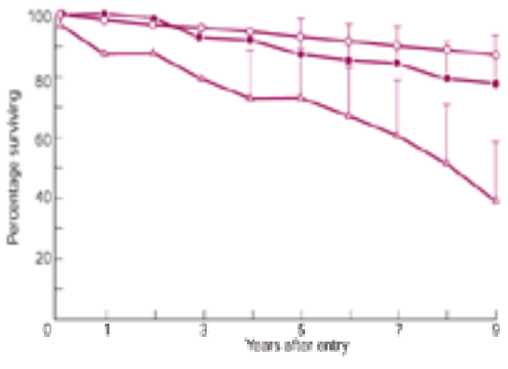

Fig. 9 Survival of female patients with rheumatoid arthritis attending hospital clinics compared with age-adjusted population data for England and Wales. —, survival of whole population; •—• survival of female patients with rheumatoid arthritis and joint disease only (*n* = 38); △–△ survival of patients with rheumatoid arthritis and extra-articular disease (*n* = 33). Bars represent 95 per cent confidence limits. (Reproduced from Erhardt CC *et al*. (1989). Factors predicting a poor life prognosis in rheumatoid arthritis: an eight year prospective study. *Annals of the Rheumatic Diseases* **48**, 7 –13, with permission.)•, survival of female patients with rheumatoid arthritis and joint disease only (*n* = 38); ▲—▲, survival of patients with rheumatoid arthritis and extra-articular disease (*n* = 33). Bars represent 95 per cent confidence limits. (Reproduced from Erhardt CC *et al*. (1989). Factors predicting a poor life prognosis in rheumatoid arthritis: an eight year prospective study. *Annals of the Rheumatic Diseases* **48**, 7–13, with permission.)

Prognostic factors

A number of prognostic factors have been identified in cohorts of patients with rheumatoid arthritis that herald rapid functional deterioration and premature death. These include more than 30 affected joints, a persistently raised level of acute-phase proteins, lower socio-economic status, early development of functional incapacity, a positive rheumatoid factor, cryoglobulinaemia, and in Northern European patients, the presence of the HLA-DRb*04 genes. However, on an individual patient basis, none of these factors are reliably predictive, either singly or in combination.

Remission

Remission has been defined by the American College of Rheumatology as an absence of joint symptoms and signs, duration of morning stiffness less than 15 min, freedom from fatigue, and an erythrocyte sedimentation rate less than 30 mm/h for females and 20 mm/h for males, for 2 consecutive months. Remission of disease is seen in a small proportion of patients, especially in the initial stages of disease. More usually, remission follows in approximately 20 per cent of patients treated with a disease- modifying antirheumatoid drug.

Diagnosis and stages of disease

Recent-onset and established disease

A diagnosis of rheumatoid arthritis is likely if three or more symmetrically distributed joints are found to be swollen and tender for more than 6 weeks in a patient with a

positive rheumatoid factor test and elevated erythrocyte sedimentation rate or serum concentration of C-reactive protein. However, not all these features are necessarily present at the early stages of disease, at which point arthritis is best termed 'undifferentiated'. The broad spectrum described in the 'presentation' section may cause considerable difficulty in making a definitive diagnosis

With the passage of time, the emergence of other features such as subcutaneous nodules, radiographic evidence of joint space narrowing, juxta-articular osteopenia, and bony erosions add further certainty to the diagnosis.

The ACR criteria (Table 1) may not be fulfilled for 6 to 12 months, by which time the pattern of joint involvement and a chronic disease course are usually evident. Prognostic factors declare themselves and the patient is regarded as having reached the stage of established disease.

Differential diagnosis

In patients with recent onset of symptoms of arthritis, the following disorders should be considered in the differential diagnosis.

Polyarthritis associated with connective tissue disease

Systemic lupus erythematosus may present with chronic non-deforming polyarthritis, but features such as Raynaud's phenomenon, photosensitivity, rashes, alopecia, haemolytic anaemia, leucopenia, thrombocytopenia, and renal or neurological involvement are detectable sooner or later, and diagnostic antinuclear antibodies (anti-ds-DNA, anti-Sm and others) are present. Other connective tissue diseases such as systemic sclerosis, polymyositis, mixed connective tissue disease, 'overlap' syndromes, and primary Sjögren's syndrome may also present with marked polyarthralgia or polyarthritis which mimics rheumatoid arthritis. In many such patients the presence of rheumatoid factor can further confuse the diagnosis. Careful clinical examination and measurement of marker autoantibodies directed against nuclear and cytoplasmic antigens will usually permit recognition of the underlying disorder. In some cases the diagnosis may only unfold after a period of weeks or months of a disorder best labelled 'undifferentiated connective tissue disease'.

Infection-related polyarthritis

The polyarthritis of rubella or other microbial agents, such as parvovirus B19 and *Borrelia burghdorferi*, and reactive arthritis associated with genitourinary or gastrointestinal infections can all cause diagnostic difficulty. A positive diagnosis is made by microbiological tests on relevant body fluids and serological tests for the detection of IgM antibodies or a rising titre of IgG antibodies to the suspected micro-organism in sequential serum samples taken over 2 weeks.

Spondyloarthropathies

Peripheral joint disease can be seen in conjunction with ankylosing spondylitis, psoriasis, and inflammatory bowel disease. Clinical examination of the spine, skin, and nails, radiological examination of the bowel using double-contrast enema or small bowel enema, endoscopy, and biopsy may reveal the underlying diagnosis. A high proportion of patients are HLA B27 positive.

Osteoarthritis

Osteoarthritis may present with inflammatory symptoms and signs but is readily distinguished by its different joint distribution (proximal and distal interphalangeal joints, carpometacarpal joints of the thumb) and radiographs that show joint space narrowing, subchondral new bone formation, osteophytes, and subchondrial cysts. Where there is pre-existing osteoarthritis, the superimposition of rheumatoid arthritis can be difficult to distinguish.

Other conditions

In late middle-aged and elderly patients the clinical presentation of polyarticular chronic pyrophosphate arthropathy may be difficult to distinguish from rheumatoid arthritis. The former diagnosis may be suspected where there is an atypical distribution of synovitis together with periarticular complications. Chronic pyrophosphate arthropathy may be associated with a modest acute-phase response and low-titre rheumatoid factor, but can usually be distinguished from rheumatoid arthritis on the basis of typical radiographic appearances and the finding of calcium pyrophosphate dihydrate (CPPD) crystals in synovial fluid aspirates. Rarely, rheumatoid arthritis and chronic pyrophosphate arthropathy may coexist.

Other diagnoses to be considered include hypermobility syndrome, polyarticular gout, psoriatic arthritis, haemochromatosis, sarcoidosis, sickle-cell disease, primary amyloidosis, and paraneoplastic disease.

Laboratory tests

Laboratory studies are an integral part of the management of patients with rheumatoid arthritis and are employed for diagnosis, evaluation of prognosis, assessment of disease activity, response to therapy, and monitoring toxic effects of drugs. Only routinely used tests are considered here.

The measurement of rheumatoid factors is useful in the early stages of assessment of a patient with suspected rheumatoid arthritis. Tests using sensitized sheep erythrocytes (Rose–Waaler) in an agglutination test give better diagnostic specificity than agglutination of human IgG-coated latex particles. Automated tests using nephelometry or enzyme-linked immunosorbent assays are being increasingly used. All assays have to be standardized against a reference standard and may be expressed in international units. A positive result is one that exceeds concentrations (or titres) observed in less than 5 per cent of normal controls, or the value set by an international reference standard, and is observed in about 70 per cent of patients at some point in their disease course. In a patient with recent-onset polyarthritis, a positive rheumatoid factor is moderately specific for rheumatoid arthritis but can also be observed in patients with other connective tissue diseases such as systemic lupus erythematosus and primary Sjögren's syndrome. A repeat test may be positive after an initial test is negative and is therefore necessary before a patient can be categorized as having seronegative rheumatoid arthritis. A significant titre of rheumatoid factor is associated with a poor prognosis and extra-articular disease.

Measurement of erythrocyte sedimentation rate (Westergren method) and serum C-reactive protein are extensively used. High values correlate with disease severity and a reduction is one criterion of response to therapy. Persistently elevated C-reactive protein concentrations correlate with deforming erosive disease.

Patients with active rheumatoid arthritis show haematological abnormalities as a consequence of disease-related mechanisms, such as the overproduction of cytokines suppressing the bone marrow, immune complexes increasing the clearance of polymorphonuclear cells in Felty's syndrome, and hypersplenism reducing platelet counts. A high level of disease activity is associated with a normocytic normochromic anaemia, polymorphonuclear leucocytosis, and thrombocytosis. These abnormal values tend to return to normal as the inflammatory component of disease responds to therapy.

Active disease may also be associated with a raised serum alkaline phosphatase and a low serum albumin. Serum chemistry is otherwise normal. Serum immunoglobulin and complement C3 and C4 levels may be elevated.

Non-steroidal anti-inflammatory drug (**NSAID**) therapy may cause microcytic iron-deficiency anaemia from blood loss from the gastrointestinal tract. A low serum ferritin level suggests iron deficiency, but this is not a reliable guide in cases of rheumatoid arthritis as serum concentrations may be elevated as part of an acute-phase response. Corticosteroids may be responsible for increased polymorphonuclear cell counts and decreased lymphocyte counts. Many DMARDs show dose-related bone marrow toxicity, and sulphasalazine, D-penicillamine, azathioprine, and gold can cause unexpected agranulocytosis due to hypersensitivity, unrelated to the dose administered.

Of the commonly used drugs, methotrexate, sulphasalazine, azathioprine, and leflunomide are hepatoxic and can cause elevation of liver enzymes and alkaline phosphatase. Repeated monitoring is advisable: persistent or highly raised values should prompt further investigation or discontinuation.

Imaging

Radiographs of hands and feet can be used to assess the presence and progression of cartilage loss and bone erosions (Fig. 10). Standardized measurements (the

Larsen or Sharp scoring methods) have been devised to quantify these measures. Changes seen in the hands and feet correlate with radiological changes in other affected joints, showing a linear progression after the initial 1 to 2 years. The erosion count correlates with physical function. Radiographs of affected joints are used for the assessment of integrity and damage. Flexion views of the cervical spine are suitable for demonstration of atlantoaxial subluxation and cervical instability. Arrest or retardation of radiographic change is considered to be a marker of good control of disease.

Fig. 10 Hand radiographs taken soon after symptom onset (left panel) and 12 years into established disease (right panel), showing extensive structural damage especially in the metacarpophalangeal, interphalangeal, carpal, and wrist joints.

Magnetic resonance imaging (**MRI**) and computed tomography (CT) are valuable in assessing neck pathology and pressure on the cervical cord. MRI and high-frequency ultrasound examination are sensitive methods to evaluate synovitis and early change in cartilage and bone, but their place in routine management is not yet established. Dual emission x-ray absorptiometry (DEXA) scanning is in routine use for the assessment of bone mineral density.

Management

Aims of treatment

These are:

1. to relieve symptoms and signs of disease;
2. maintain physical function;
3. prevent structural damage to joints and associated structures;
4. restore and maintain quality of life that permits the pursuit of normal work, domestic, and social life;
5. reduce the comorbidity and increased mortality associated with the disease and therapies; and
6. correct abnormal laboratory-based values of haemotopoietic function, acute-phase proteins, and other markers of disease process.

Achievable goals of current therapy

Considerable progress has been made in developing effective therapies for the relief of symptoms and signs of disease. However, despite the best therapies in current use, the goals of halting structural damage and maintaining a normal quality of life have not yet been realized, although significant progress has been made. The realistic aims, therefore, are to maximize gains whilst minimizing toxicity of drugs (an optimum risk:benefit ratio) and to operate within the pharmacoeconomic constraints (cost–benefit and cost–utility) that apply in the setting in which the patient is being treated.

The costs of treatment of rheumatoid arthritis over the lifetime of a patient are considerable. They include direct and indirect costs that are cumulative and incremental as the disease progresses. Direct costs include those in the primary and hospital sectors of medical and allied health professionals, hospital admissions, drugs, surgery, aids, and appliances. Indirect costs include those arising from loss of economic productivity and earnings, unemployment and disability benefits, the cost of maintaining mobility, and domestic help and daily care for the severely disabled. Patients with rheumatoid arthritis become debilitated by pain and fatigue and may experience psychological depression, anxiety, and loss of self-esteem, which require additional medical treatment and psychological support. Iatrogenic diseases caused by anti-inflammatory and other drugs add further socio-economic burdens over the long term. These considerations have been used as an argument for the aggressive use, within safe limits, of drugs and new therapies that offer rapid relief of symptoms and signs, retardation of structural damage, and maintenance of functional capacity.

The physician has to be able to evaluate these factors and, because of the variability of disease expression and progression, individualize and agree the treatment plan with each patient. It is prudent to re-evaluate and revise the goals from time to time as the prognosis and response to therapy unfolds.

General principles

The heterogeneous and variable course of chronic rheumatoid arthritis presents a complex management problem, with each patient requiring an individualized approach for an optimum outcome. Nevertheless, the general principles that may aid the physician's task are summarized below.

Ascertainment of the severity of disease determines the appropriate choice of drugs. Evaluation of the extent of joint and extra-articular involvement, the level of disease activity, and its progression are judged by clinical examination, laboratory tests, and the rate of progression of radiographically determined damage to joints. The full extent may only unfold over months or years of follow-up. Patients with severe and active disease will require more aggressive medical treatment than those with mild disease.

The main aim is to control disease activity as rapidly as possible. Response to therapy should be monitored to ensure efficacy, using quantifiable clinical and laboratory indices of inflammatory activity and impact on the progression of damage to joints. A lack of response to initial therapy should trigger a change in the management plan and consideration of alternative strategies at intervals of 3 to 6 months. A thorough knowledge of the scope and limitations of treatment modalities is essential in the art of management.

In early rheumatoid arthritis the goal should be to achieve disease remission. With aggressive and continuing use of available therapeutic agents there is evidence that it is possible to achieve this in 20 to 40 per cent of patients over a period of 1 to 2 years. Since low or absent disease activity correlates with retardation of joint damage, the benefit of treatment is likely to be most marked in the early phase of disease.

Remission is rare in established rheumatoid arthritis of more than 2 or 3 years duration. Nevertheless, minimizing disease activity by attention to a measurable response to therapy remains at the core of the management plan. Controlled clinical trials support the concept that optimum use of drugs can ameliorate symptoms and signs and retard progression of joint damage and disability even in later stages of disease.

At all stages of disease, irrespective of its severity, drug therapy constitutes only one part of the whole management plan. Other essential elements include measures such as patient education, psychological and employment counselling, setting appropriate levels of rest and exercise, coping with tasks of daily living, and maintaining mobility, access to splints, aids, and appliances for the disabled, and access to social and financial benefits. In addition, successful pharmacological intervention requires the patient's informed consent in instituting therapeutic decisions and involving the patient and other carers in monitoring of drug toxicity. The provision of holistic care thus requires team work and co-ordination between the treating physician and other medical and health-care professionals, including specialist nurses, physiotherapists, occupational therapists, and social workers.

Surgical treatment plays an important role in relieving intractable symptoms and restoring loss of physical function and mobility due to damage to joints, tendons, and

associated soft tissues. It is also indicated in the treatment of secondary complications such as entrapment of peripheral nerves at the wrist and elbow and cervical cord compression due to instability of the cervical spine.

Evidence base and profile of drugs used in the treatment of rheumatoid arthritis

Non-steroidal anti-inflammatory drugs (NSAIDs)

NSAIDs are widely used for treating symptoms of rheumatoid arthritis. They act by inhibiting the enzymes cyclo-oxygenase I and/or II, which act on lipid substrates in cells, converting them to prostanoids. Tissues such as the gastric and duodenal mucosa, blood vessels, and platelets constitutively express cyclo-oxygenase I, which in the gastroduodenal mucosa regulates the production of prostaglandins, including prostaglandin E_2, that exert a protective effect on its integrity by reducing acid secretion and increasing the secretion of mucus and bicarbonate. Cyclo-oxygenase I-induced prostaglandin E_2 promotes platelet aggregation and its prothrombotic effects. By contrast, cyclo-oxygenase II is mainly induced in macrophages and polymorphonuclear cells at sites of inflammation by pro-inflammatory cytokines such as IL-1 and TNF-a. NSAIDs that inhibit both cyclo-oxygenase I and II activity therefore compromise the gastroprotective effect of cyclo-oxygenase I, whilst simultaneously exerting a therapeutic effect by inhibiting the production of inflammatory prostanoids. Selective or specific cyclo-oxygenase II inhibitors should in theory block inflammation without gastropathic effects.

The conventional and currently widely used NSAIDs are inhibitors of both cyclo-oxygenase I and II. In the United States they are responsible for admission to hospital of over 1 per cent of patients with rheumatoid arthritis per year for complications such as peptic ulceration, gastric haemorrhage, and perforation, and account for a twofold increase in death over expected rates. It is claimed that the least gastrotoxic are ibuprofen and nabumetone, with naproxen and diclofenac carrying intermediate risk, followed by drugs with a high risk such as fenoprofen, ketoprofen, indomethacin and piroxicam, and azapropazone (see Table 2 for dose ranges).

For patients who develop dyspepsia and/or NSAID-induced gastropathy, or elderly patients who have a high risk of gastroduodenal side–effects, concomitant administration of prostaglandin analogues (such as misoprostol) or proton-pump inhibitors (such as omeprazole or lansoprazole) is recommended, and NSAIDs are best avoided for patients with a history of peptic ulcers. Eradication of *Helicobacter pylori* infection results in long-term healing of pre-existing gastric and duodenal ulcers, but whether it decreases dyspepsia or ulceration caused by NSAIDs is uncertain.

Drugs such as meloxicam and etodolac that act by selective or specific inhibition of cyclo-oxygenase II and thus spare cyclo-oxygenase I have fewer gastropathic effects. Two highly cyclo-oxygenase II selective inhibitors, celecoxib and rofecoxib, possessing no significant cyclo-oxygenase I inhibitory activity at anti-inflammatory therapeutic doses, have been recently introduced. Clinical trials have demonstrated their improved safety profile in respect to endoscopically detectable gastroduodenal ulcers, upper gastrointestinal haemorrhage, and perforation when compared with conventional NSAIDs. Although the inhibition of platelet function by NSAIDs that inhibit cyclo-oxygenase I is associated with serious gastrointestinal haemorrhage in susceptible individuals, certain of these drugs, such as low-dose aspirin, appear to be beneficial in the prevention of strokes and coronary thrombosis, a property not shared by selective cyclo-oxygenase II inhibitors.

NSAIDs are valuable in controlling joint pain and stiffness but have insignificant effect on factors mediating joint damage. There is little difference in the efficacy of available NSAIDs at optimal doses. Preferred NSAIDs have the most favourable risk:benefit ratio at low cost and are administered once or twice daily at doses that achieve 8- to 12-h activity to ensure compliance, alleviation of symptoms during nocturnal sleep, and on waking in the morning.

All NSAIDs can cause fluid retention and oedema by a renin–angiotensin-dependent mechanism that may also aggravate congestive cardiac failure and systemic hypertension. Patients with impaired renal function, cirrhosis of the liver, and decreased plasma volume from any cause are at risk from developing NSAID-induced renal toxicity. It is claimed that sulindac may be safer than other NSAIDs in patients with renal failure.

NSAIDs, especially indomethacin, may cause side-effects involving the central nervous system such as headache, dizziness, anxiety, disorientation, and drowsiness. Rarely, use of NSAIDs may be associated with aseptic meningitis. NSAIDs may aggravate asthma and cause hypersensitivity reactions. Blood dycrasias and an increase in serum concentration of liver enzymes and alkaline phosphatase are described. Drug interactions may decrease the efficacy of some concomitantly prescribed therapies, for instance antihypertensives and lithium, and potentiate the effects of others, for instance anticoagulants, anti-epileptics, and oral hypoglycaemics. NSAIDs decrease the excretion of methotrexate but do not appear to increase its toxicity in the dose range used to treat rheumatoid arthritis. They also increase plasma concentrations of cyclosporin and FK-506 and hence may increase the risk of renal toxicity.

Disease-modifying antirheumatoid drugs (DMARDs)

DMARDs, also classified as slow-acting antirheumatoid drugs (SAARDs) because of the lag period of some weeks before their anti-inflammatory effect becomes apparent, are the current cornerstone of drug therapy for rheumatoid arthritis (Table 3). Drugs in this category include: the antimalarials, hydroxychloroquine or chloroquine sulphate; sulphasalazine; weekly low-dose oral or parenterally administered methotrexate; weekly injections of gold aurothiomalate or gold aurothioglucose; leflunomide; cyclosporin; azathioprine; and D-penicillamine. Drugs such as gold, antimalarials, and methotrexate were introduced for use in rheumatoid arthritis by serendipity. Others, such as azathioprine, cyclosporin, and lefluomide, were developed as immunosuppressive agents for preventing transplant rejections and subsequently used to curb the aberrant immunological response in rheumatoid arthritis. The mechanism of action of these drugs in rheumatoid arthritis is complex and still incompletely understood. Inhibitory effects on inflammatory pathways, immune responses, and cell activation have been described in experimental systems and clinical studies.

Clinical trials have demonstrated superior efficacy of all these drugs over placebo in controlling symptoms and signs in patients previously treated with only NSAIDs in early and established rheumatoid arthritis. In addition, compared with placebo, sulphasalazine, methotrexate, and leflunomide appear to retard progression of structural damage as assessed by serial radiographs of the hands and feet in controlled trials lasting 6 to 12 months.

A meta-analysis of clinical trials of commonly used DMARDs has been analysed for efficacy and toxicity relative to each other and to placebo treatment. Methotrexate, sulphasalazine, injectable gold, and D-penicillamine have the best and equal efficacy in the short term compared with placebo. The antimalarials (hydroxychloroquine and chloroquine) and azathioprine appear to be less efficacious in this analysis. The toxicity profile shows a different rank order. Antimalarials are least toxic, followed by methotrexate and sulphasalazine in an intermediate range, and injectable gold, azathioprine, and D-penicillamine at the most toxic end of the spectrum. Methotrexate and sulphasalazine emerge with the best balance between efficacy and toxicity. Since leflunomide was introduced recently it was not included in this meta-analysis, but its efficacy and toxicity profile is similar to methotrexate and sulphasalazine.

Remission of rheumatoid arthritis on DMARD therapy has been described in approximately 20 per cent of those with early disease treated with methotrexate or sulphasalazine as single agents. However, in one example of a follow-up study fewer than 1 in 10 of 18 per cent that achieved remission (i.e. around 2 per cent of the original cohort) sustained it for longer than 3 years. Remissions are rare in patients who have progressed to a stage of physical disability and whose radiographs show bony erosions.

Conclusions from short-term randomized clinical trials do not reflect the effectiveness of DMARDs in controlling disease activity in the longer term. Incomplete responses, relapses, and adverse events are common and account for discontinuation of antimalarials, gold salts, D-penicillamine, sulphasalazine, and azathioprine in the majority of patients in 1 to 3 years. By contrast, responses to methotrexate appear to be more durable in follow-up studies of large cohorts of patients with rheumatoid arthritis, with about 50 per cent continuing therapy at 5 years. Data on long-term effectiveness, tolerability, and toxicity of leflunomide are not yet available.

Combinations of DMARDs have been used in the expectation that their different modes of action might provide added efficacy. However, a meta-analysis showed that there was marginal benefit at the doses and combinations used prior to 1994, especially in the reduction of number of tender joints, with increased toxicity when compared with single agents. Several subsequent randomized controlled trials have demonstrated significantly improved efficacy of combination therapy, without increased toxicity: some examples are given below.

In one trial lasting 2 years, a combination of oral methotrexate (7.5 to 17.5 mg/week), sulphasalazine (0.5 g twice daily), and hydroxychloroquine (200 mg twice daily) showed superior control of symptoms and signs compared with methotrexate alone or a combination of sulphasalazine and hydroxychloroquine. Patients enrolled in this trial with advanced disease had already failed to respond to DMARD monotherapy.

In a further study on patients with disease of less than 2 years duration, the introduction of a combination of methotrexate, sulphasalazine, and hydroxychloquine not

only controlled symptoms better but, at the end of 2 years, had induced remission in 37 per cent compared with 21 per cent of patients on sulphasalazine or methotrexate alone.

In another trial lasting 24 weeks, patients with active disease despite methotrexate (mean dose 12.5 mg/week) showed improvement in their signs and symptoms when cyclosporin at 2.5 to 5 mg/kg daily was added, compared with the addition of placebo. Similarly, in an open-label study, the addition of leflunomide at 20 mg daily enhanced the efficacy of ongoing methotrexate treatment in patients whose disease activity was not well controlled, but was associated with greater toxicity.

Corticosteroids

Corticosteroids are potent anti-inflammatory agents and are most efficacious in treating symptoms and signs of rheumatoid arthritis and for amelioration of systemic features, but their use is limited by toxicity related to dose and duration of exposure. The circumstances in which use of corticosteroids has been established and those in which it is debated are described below.

In patients in whom loss of function and disease activity is restricted to a few joints, local corticosteroid therapy can be most effective. This indication may arise in those whose rheumatoid disease is limited to a few joints, or in patients with an incomplete response to NSAID and DMARD therapy. Several alternative corticosteroid preparations are available, the dose being dependent on the size of the joint. Depot methyl prednisolone (dose range 4 to 40 mg) or triamcinolone acetonide (dose range 2.5 to 40 mg depending on size of joint) are suitable alternatives. Repeat injections may be necessary, but more than three per joint per year should be avoided.

Corticosteroid administered orally in courses lasting a few weeks to months (such as prednisolone at 7.5 to 10 mg daily), or in the form of 'pulse therapy' (such as depot methyl prednisolone at 80 to 120 mg by intramuscular injection), is a suitable adjunctive therapy in patients in whom the benefit of DMARDs is not yet established. Longer-term, more or less indefinite, treatment with low-dose prednisolone is necessary in patients with moderate to severe disease, especially associated with refractory anaemia that is not controlled with currently used antirheumatoid drugs. Long-term low-dose prednisolone retards the progression of rheumatoid bone erosions in radiographs of hands and feet and, hence it is claimed, deterioration of physical function. Whether this benefit is outweighed by the side-effects and morbidity of corticosteroid therapy is debatable. Higher doses of corticosteroids are indicated in the treatment of severe extra-articular disease.

Prevention of corticosteroid-induced osteoporosis and reduction in risk of fractures requires adequate prophylaxis with calcium and vitamin D intake (for example, daily intake of 1000 mg of calcium and 800 IU of vitamin D). In susceptible patients, or those on doses exceeding the equivalent of 7.5 mg of prednisolone daily, measurement of bone mineral density is used to identify and monitor management. Bisphosphonates may be required in addition to calcium and vitamin D, and hormone replacement therapy is recommended in perimenopausal women.

Biological therapy

The identification of TNF-a as a key mediator of rheumatoid inflammation has led to the development of anti-TNF agents that have been introduced recently into clinical practice. Other targeted therapies are in advanced clinical trials, including recombinant IL-1 receptor antagonist (rIL-ra) that inhibits the activity of IL-1, an important pro-inflammatory cytokine. In clinical trials treatment with rIL-ra has shown significant improvement of signs and symptoms and retardation of radiographic progression of joint damage.

Non-pharmacological measures and support

Specialist physicians co-ordinate the management of patients with rheumatoid arthritis using a team of health professionals. The support provided improves the patient's ability to cope with pain, disability, daily activities, and the prospect of continuing work and retaining independence.

Education and counselling is helpful in preparing patients for the likely consequences of their disease, and its development over time. It also allows the patient to participate fully in making informed decisions about taking and monitoring drugs and retaining control. Studies have demonstrated the benefit of this approach in minimizing costs of medical care and improved outcomes.

Bed rest and the use of resting splints is helpful during the very acute stages of joint disease, but should always be accompanied by daily passive joint movements and appropriate isometric exercises to avoid contractures, muscle atrophy, and osteoporosis, and to retain joint function. Exercise initiated under supervision and maintained by patients on a regular basis does not accelerate joint damage, diminishing pain and promoting a sense of well being in those in whom fatigue is a major feature of active disease.

For patients with disability, aids and appliances can be helpful in undertaking daily tasks and leisure activities such as dressing, turning keys and taps, cooking, lifting, domestic tasks, and gardening. Adjustments in the home are helpful, such as use of cushions and chairs with high seats in the bathroom and toilet. For the very disabled, learning techniques for transfers from bed to chair, chair to the toilet, and the installation of chairlifts and use of wheelchairs need expert help and advice.

Maintenance of mobility requires attention to foot care, podiatry, comfortable shoes, a walking stick or elbow crutches, and specially adapted motor vehicles to get to work and for social purposes.

Disabled people have certain privileges in employment and may qualify for disability allowances. Some may benefit from retraining for suitable work. The health-care team needs to recognize that chronic illness and disability places increased pressure on spouses and family, who generally end up as carers of the patient with rheumatoid arthritis: support and counselling should therefore extend to them.

Dietary manipulation, such as exclusion of certain foods and beverages, has enjoyed popularity and in some patients appears to be beneficial, but there is little evidence that most such diets are of value. Fasting followed by a vegetarian diet was shown to be of benefit in a Norwegian study, but the durability of effect is unknown. Diets rich in fish oils and omega fatty acids appear to be of some benefit. As excessive weight accelerates joint damage and increases the risk of complications when undergoing essential surgery, obese patients should be encouraged to lose weight.

Management strategies

Mild disease

Definition

Mild disease may be defined as rheumatoid arthritis with limited joint involvement, low disease activity, and without markers of poor prognosis. Such patients will typically show most of the following features: involvement of less than six or seven individual joints and sparing of weight-bearing joints; pain readily controlled with NSAIDs; less than 15 min of joint stiffness on waking or following inactivity; lack of extra-articular disease; minimally elevated erythrocyte sedimentation rate or concentration of C-reactive protein; negative rheumatoid factor test; a normal haematological profile; little or no impairment of physical function; and ability to undertake activities of daily living, maintaining employment and enjoying non-strenuous social and leisure activities. Radiographs of hands and feet show a lack of significant osteopenia, joint space narrowing, and bony erosions at baseline and annual follow-up. The disease course may be punctuated by self-limiting exacerbations of symptoms and signs. Patients with mild disease constitute a small proportion of patients referred to specialist clinics but are more numerous in the community and in the primary care setting.

Drug treatment

This consists of a judicious use of non-steroidal anti-inflammatory drugs. Corticosteroid injections into individual affected joints, tendon sheaths, and bursas for persistent swelling, tenderness, or loss of normal range of movement can be very effective. Follow-up assessment is necessary to ensure that the disease has not altered to a more severe pattern. DMARDs are indicated in those with recurrent or persistent symptoms and signs, deformities, or radiographic evidence of structural damage. Hydroxychloroquine or sulphasalazine are used initially. If the decision to embark on the use of DMARDs is made, the aims and management strategy are

the same as for patients with moderate or severe disease.

Moderate and severe disease

Definition

This is defined as rheumatoid disease that has evolved into an unremitting pattern of polyarthritis with evidence of significant functional impairment and joint damage. With increasing severity most of the following features are present: 10 to 30 swollen and tender joints; frequent involvement of proximal joints of the upper and lower limbs; moderate to severe pain; inactivity and morning stiffness exceeding 1 h in duration; prominent fatigue; elevated erythrocyte sedimentation rate and/or C-reactive protein concentrations; low haemoglobin concentration; polymorphonuclear leucocytosis and thrombocytosis; and positive rheumatoid factor test. Deformities of joints are apparent early in the course of disease and radiographs of hands, feet, and affected joints already show loss of joint space and subchondral erosions within 2 years of presentation. Such patients show a significant impairment in daily activities and restricted ability to perform domestic and work-related tasks and to enjoy social and leisure activities.

Drug treatment

The aim of drug treatment is to achieve rapid control of disease activity and, if possible, remission of disease. This requires simultaneous or sequential use of drugs belonging to different classes, for instance NSAIDs, DMARDs, corticosteroids, and biological therapies as discussed below.

NSAIDs are used at optimal doses for control of pain and stiffness (Table 2), those most commonly given in practice being naproxen, diclofenac, and indomethacin. Many physicians prefer to administer these drugs in slow-release preparations in the morning and before retiring to bed at night. In the elderly, cyclo-oxygenase II-selective NSAIDs may be preferable, or else the simultaneous use of a gastroprotective agent, most commonly proton pump inhibitors. In addition, simple analgesics such as 0.5 to 1 g of paracetamol every 6 h may be required for relief of pain.

DMARDs should be used in all patients (Table 3), the two most commonly employed being sulphasalazine and methotrexate, provided there are no contraindications. These are given as single drugs in incremental doses over 3 to 4 months to the maximum recommended or tolerated dose. If a clear-cut reduction in disease activity (or remission) is not observed with one of these drugs, then monotherapy with leflunomide, azathioprine, or injectable gold may be attempted. Alternatively, other DMARDs are added at this stage. Commonly used DMARD combinations include: methotrexate and hydroxychloroquine; methotrexate, sulphasalazine, and hydroxychloroquine; and methotrexate and cyclosporin. The choice of therapy is ultimately determined by evaluation of risks of toxicity, efficacy, durability, and direct and indirect costs of treatment. There is no consensus on the most effective combination regimen. Meticulous monitoring of toxic effects is necessary.

In practice over 50 per cent of patients with moderate or severe disease require corticosteroid therapy. If continuing long-term use appears necessary, the aim should be to reduce the dose to the equivalent of 5 to 7.5 mg of prednisolone daily by more aggressive use of DMARDs, or consider anti-TNF therapy.

Despite good initial responses to currently available DMARD treatments, a proportion—probably 10 to 15 per cent of hospital patients—show continuing disease activity and progressive disability. Randomized, placebo-controlled trials of two anti-TNF biological agents have shown these to be efficacious in such cases, and they became available in 2000, although their high cost is likely to restrict widespread use. The two anti-TNF-a drugs licensed for use in rheumatoid arthritis are infliximab (a chimeric monoclonal anti-TNF-a monoclonal antibody) given in combination with methotrexate, and etanercept (a soluble dimeric molecule consisting of a TNF receptor linked to the constant domains of Fc-IgG). Infliximab is given intravenously at a dose of 3 mg/kg over 1 h every 8 weeks to patients already receiving methotrexate therapy once a week. Etanercept given as 25 mg subcutaneously twice weekly is efficacious as monotherapy or when added to methotrexate. Symptoms and signs are rapidly alleviated in approximately 60 to 70 per cent of patients (Fig. 11) in clinical trials. Durable responses are being reported for up to 2 years, with a small increase in upper respiratory infections but without an increase in serious adverse events. Continuing therapy is needed and relapse of disease follows withdrawal.

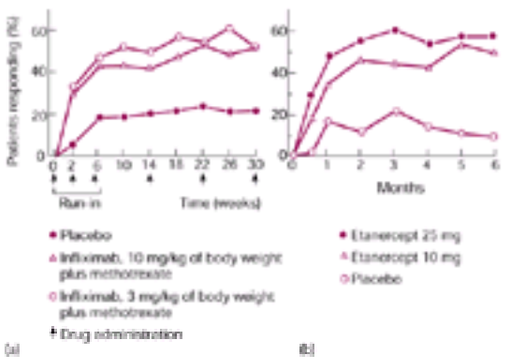

Fig. 11 Anti-TNF therapy. (a) Efficacy of combination of infliximab and methotrexate compared with methotrexate and placebo. Percentage of patients achieving a clinical response of a 20 per cent change from baseline as defined by the American College of Rheumatology 20 (ACR20) criteria. Patients were treated with methotrexate (10 to 35 mg/week) and either placebo, 3, or 10 mg/kg infliximab administered intravenously at time points indicated, in a patient group unresponsive to DMARDs with active disease despite methotrexate therapy (Maini *et al*. 1999). (b) Efficacy of etanercept compared with placebo. ACR 20 results in patients treated with two doses of etanercept or placebo injections administered subcutaneously twice weekly over a 6-month period in a population unresponsive to DMARDs. (Reproduced from Moreland *et al*. (1999), with permission.)

The combination of infliximab and methotrexate has also been reported to inhibit or even reverse significantly progression of joint damage at the end of 1 year in most patients as assessed by serial radiographs. By contrast, damage continues in the control group of patients with an incomplete response to methotrexate. In another study in rheumatoid arthritis, etanercept was found to be more effective than methotrexate in controlling progression of bone erosions, assessed by radiographs of the hands and feet at baseline and the end of 1 year. These data imply that anti-TNF therapy could preserve physical function and quality of life in the long term and hence prove to be cost-effective.

The efficacy of anti-TNF agents represents an important advance in therapy, although their efficacy and safety beyond 2 years of continuous therapy under controlled trial conditions is not yet established. Post-marketing surveillance of adverse events to infliximab and etanercept following exposure of approximately 300 000 patients to date have drawn attention to concerns arising from rare, but significant numbers of cases of sepsis, tuberculosis, fungal, and opportunistic infections. These infections are compatible with the consequences of blockade of the postulated role of TNF in host defence mechanisms. Based on these reports, regulatory authorities in the United States and Europe have advised that anti-TNF therapy is contraindicated in the presence of active serious infections and, in the case of infliximab, latent untreated tuberculosis. Other rare adverse events have included demyelinating syndromes (hence it is advisable not to treat patients with a history of multiple sclerosis), lupus syndrome, and bone marrow depression. Based on reports of an unexpected number of deaths in a phase II clinical trial of infliximab in the treatment of severe congestive cardiac failure, use of infliximab is not advisable for the treatment of patients with rheumatoid arthritis in moderate or severe congestive cardiac failure. Provided suitable screening and monitoring practices are in place, the favourable risk to benefit profile of anti-TNF therapy does not alter the indication for its use in the treatment of moderate to severe rheumatoid arthritis with persistent disease activity despite best available, but also potentially toxic and immunosuppressive, standard therapy. The high cost of anti-TNF drugs has, however, limited access to this treatment in some countries.

Extra-articular disease

Effective treatment of rheumatoid arthritis generally reduces the risk of developing severe extra-articular disease. Systemic rheumatoid vasculitis is potentially a life-threatening complication and may be aggravated by coincidental infection, such as cutaneous ulcers. After due attention to confirming the diagnosis and excluding and treating infections with appropriate antimicrobial drugs, therapy with high-dose corticosteroids and cyclophosphamide is favoured by many specialists, although no randomized placebo-controlled trial data are available. One regimen recommends intravenous methylprednisolone at 1 g daily for 3 days, simultaneously with an initial single pulse of intravenous cyclophosphamide (10 to 15 mg/kg) in a fully hydrated patient to prevent bladder toxicity. Cyclophosphamide is repeated every 3 to

4 weeks, subject to a satisfactory clinical response or lack of toxicity, up to a total dose of 10 to 12 g in a cycle of treatment. Alternatively, oral cyclophosphamide at 2 mg/kg (maximum dose 150 mg daily) may be used. Oral high-dose prednisolone is continued until clinical response is observed or toxic effects occur, when it is rapidly tapered to a maintenance dose, generally about 15 mg daily. Similarly, cyclophosphamide is substituted by the less toxic azathioprine at 1.5 to 2 mg/kg daily or methotrexate at 15 mg/week.

Similar regimens have been used for severe fibrosing alveolitis and for severe scleritis and corneal melt in conjunction with local therapy. Occasional patients with Felty's syndrome and hypersplenism that do not respond to DMARDs benefit from splenectomy, and their neutropenia may respond to recombinant human granulocyte colony-stimulating factor. Keratoconjunctivitis sicca and dry mouth due to secondary Sjögren's syndrome respond to local measures including artificial tears, dental hygiene, and saliva substitute.

Further reading

Arend WP, Dayer JM (1990) Cytokines and cytokine inhibitors or antagonists in rheumatoid arthritis. *Arthritis and Rheumatism* **33**, 305–15.

Arnett FC *et al*. (1988). The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis and Rheumatism* **31**, 315–34.

British Society for Rheumatology (2000). *British Society for Rheumatology drug monitoring guidelines*. BSR Headquarters, 41 Eagle Street, London, WC1R 4AR.

Emery P *et al*. (1999). Celecoxib versus diclofenac in long-term management of rheumatoid arthritis: randomised double-blind comparison. *Lancet* **354**, 2106–11.

Erhardt CC *et al*. (1989). Factors predicting a poor life prognosis in rheumatoid arthritis: an eight year prospective study. *Annals of the Rheumatic Diseases* **48**, 7–13.

Feldmann M, Brennan FM, Maini RN (1996). Role of cytokines in rheumatoid arthritis. *Annual Review of Immunology* **14**, 397–440.

Felson DT, Anderson JJ, Meenan RF (1994). The efficacy and toxicity of combination therapy in rheumatoid arthritis: a metaanalysis. *Arthritis and Rheumatism* **37**, 487–91.

Felson DT *et al*. (1995). American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. *Arthritis and Rheumatism* **38**, 727–35.

Fries JF, Spitz PW, Young DY (1982). The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. *Journal of Rheumatology* **9**, 789–93.

Furst DE (2000). Aggressive strategies for treating aggressive rheumatoid arthritis: has the case been proven? *Lancet* **356**, 183–4.

Gardner DL (1992). Rheumatoid arthritis: cell and tissue pathology. In: Gardner DL, ed. *Pathological basis of the connective tissue diseases*, pp 444–526. Edward Arnold, London.

Gregersen PK, Silver J, Winchester RJ (1987). The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis and Rheumatism* **30**, 1205–13.

Griffiths ID (1998). Extra-articular features of rheumatic diseases. In: Maddison PJ, Isenberg DA, Woo P, Glass DN, eds. *Oxford textbook of rheumatology*, 2nd edn, pp 169–79. Oxford University Press.

Gotzsche PC (2000). Non-steroidal anti-inflammatory drugs. *British Medical Journal* **320**, 1058–61.

Kirwan JR (1995). The effect of glucocorticoids on joint destruction in rheumatoid arthritis. The Arthritis and Rheumatism Council Low-Dose Glucocorticoid Study Group. *New England Journal of Medicine* **333**, 142–6.

Lawrence JS (1970). Rheumatoid arthritis: nature or nurture? *Annals of the Rheumatic Diseases* **29**, 357–69.

Maini RN (1998). The Lumleian lecture: Milestones in the development of anti-tumour necrosis factor a therapy (TNFa) therapy. In: Pusey C, ed. *Horizons in medicine No 11*, pp 131–45. Royal College of Physicians, London.

Maini RN, Feldmann M (1998). Immunopathogenesis of rheumatoid arthritis. In: Maddison PJ, Isenberg DA, Woo P, Glass DN, eds. *Oxford textbook of rheumatology*, 2nd edn, pp 983–1004. Oxford University Press.

Maini RN, Taylor PC (2000). Anti-cytokine therapy for rheumatoid arthritis. *Annual Review of Medicine* **51**, 207–29.

Maini RN *et al*. (1999). Randomised phase III trial of infliximab (Chimeric anti-TNFa monoclonal antibody) versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate. *Lancet* **354**,1932–9.

Mangge H, Hermann J, Schauenstein K (1999). Diet and rheumatoid arthritis—a review. *Scandinavian Journal of Rheumatology* **28**, 201–9.

Moreland LE *et al*. (1999). Etanercept therapy in rheumatoid arthritis. *Annals of Internal Medicine* **130**, 478–86.

O'Dell JR *et al*. (1996). Treatment of rheumatoid arthritis with methotrexate alone, sulfasalazine and hydroxychloroquine, or a combination of all three medications. *New England Journal of Medicine* **334**, 1287–91.

Pinals RS *et al*. (1981). Preliminary criteria for clinical remission in rheumatoid arthritis. *Arthritis and Rheumatism* **24**, 1305–15.

Pincus T (1988). Rheumatoid arthritis: disappointing long-term outcomes despite successful short-term clinical trials. *Journal of Clinical Epidemiology* **41**, 1037–41.

Pincus T, Callahan LF (1993). What is the natural history of rheumatoid arthritis? *Rheumatic Disease Clinics of North America* **19**, 123–51.

Sharp JT *et al*. (2000). Treatment with leflunomide slows radiographic progression of rheumatoid arthritis: results from three randomized controlled trials of leflunomide in patients with active rheumatoid arthritis. Leflunomide Rheumatoid Arthritis Investigators Group. *Arthritis and Rheumatism* **43**, 495–505.

Short CL (1974). The antiquity of rheumatoid arthritis. *Arthritis and Rheumatism* **17**, 193–205.

Silman AJ, Hochberg MC (1993). Rheumatoid arthritis. In: *Epidemiology of the rheumatic diseases*, pp 7–68. Oxford University Press.

Tugwell P *et al*. (1995). Combination therapy with cyclosporine and methotrexate in severe rheumatoid arthritis. The Methotrexate–Cyclosporine Combination Study Group. *New England Journal of Medicine* **333**, 137–41.

Van der Heijde DM *et al.* (1993). Development of a disease activity score based on judgment in clinical practice by rheumatologists. *Journal of Rheumatology* **20**, 579–81.

van Riel PL, Haagsma CJ, Furst DE (1999). Pharmacotherapeutic combination strategies with disease-modifying antirheumatic drugs in established rheumatoid arthritis. *Bailliere's Best Practice and Research: Clinical Rheumatology* **13**, 689–700.

Wiles N *et al*. (1999). Estimating the incidence of rheumatoid arthritis: Trying to hit a moving target? *Arthritis and Rheumatism* **42**, 1339–46.

Wolfe F *et al*. (1994). The mortality of rheumatoid arthritis. *Arthritis and Rheumatism* **37**, 481–94.

Young A *et al*. (2000). How does functional disability in early rheumatoid arthritis (RA) affect patients and their lives? Results of 5 years of follow-up in 732 patients from the early RA study (ERAS). *Rheumatology* **39**, 603–11.

18.6 Spondyloarthropathies and related arthritides

J. Braun and J. Sieper

Introduction and definitions

The spondyloarthropathies are a heterogenous group of inflammatory rheumatic diseases with predominant involvement of axial and peripheral joints and entheses. In addition to these, the various spondyloarthropathies share other characteristic clinical features, for example anterior uveitis and Crohn-like gut lesions. Symptoms in subsets of spondyloarthropathies can overlap, for example psoriatic skin lesions in Reiter's syndrome, and patients can move from one subset to another, for example from reactive arthritis to ankylosing spondylitis.

The various names which have been and are still used for the spondyloarthropathies include seronegative spondarthropathies, spondarthritis, spondylarthropathy, spondyloarthropathy, and spondyloarthritis. There is no substantial difference between them. The prefix seronegative, referring to the general absence of rheumatoid factors in the spondyloarthropathies, is historical and redundant. The term spondyloarthropathy is preferred in this chapter. The spondyloarthropathies are not modern diseases, with ankylosing spondylitis first having been described in 1649 (Table 1).

Epidemiology

The mean age at onset is 20 to 40 years, with a slight preponderance of males in most subsets of spondyloarthropathy. Next to rheumatoid arthritis, the spondyloarthropathies are the most frequent inflammatory rheumatic diseases (Table 2), with ankylosing spondylitis and undifferentiated spondyloarthropathy being the most common subsets. The overall prevalence of spondyloarthropathies in patients presenting with back pain to general practitioners' surgeries in the United Kingdom has been estimated at 5 per cent.

The spondyloarthropathies are associated with the major histocompatability complex class I antigen HLA B27, and the prevalence of spondyloarthropathies in any population correlates with that of HLA B27. The magnitude of association differs between the subsets (Table 3) and has been mainly shown for ankylosing spondylitis, but in Inuit populations Reiter's syndrome is more frequent.

Pathogenesis

In all forms of spondyloarthropathy there is a strong genetic association with the major histocompatability complex class I antigen HLA B27, as shown in Table 3. The overall influence of genes in the pathogenesis of ankylosing spondylitis has been estimated to be 95 per cent, leaving only 5 per cent to other causative factors such as environmental influences. HLA B27 is responsible for about one-third of the total genetic load: 25 subtypes are now recognized by polymerase chain reaction technology, three of which are not associated with ankylosing spondylitis, or are associated less strongly. There is weaker association of spondyloarthropathies with

HLA B60 and HLA DR1, and possibly also tumour necrosis factor-a polymorphisms. Further genes have not yet been identified.

The relevance of HLA B27 to disease pathogenesis is not known: several models have been proposed to explain tissue tropism, the aberrant immune response to certain bacteria, and the HLA B27 association of the spondyloarthropathies (Table 4).

The classical arthritogenic peptide model is backed by the demonstration of HLA B27-restricted CD8+ T-cell clones in the synovial fluid of patients with reactive arthritis. Immunodominant peptide motifs and peptides have been described, but their pathogenetic relevance is not yet clear. Lipopolysaccharide and RNA of bacteria associated with reactive arthritis and a CD4+ T-cell response directed against bacterial antigens have been detected in reactive arthritis, but it is not clear whether this immune response is beneficial or arthritogenic. At the humoral and the cellular level, molecular mimicry (partial sequence homologies at the protein and DNA level) between bacterial antigens and self structures has been described, mainly of the HLA B27 molecule. It also seems possible that patients with HLA B27+ spondyloarthropathies have deficient immune reactivity, for example diminished ability to secrete tumour necrosis factor-a, or a synovial T_{H2} response (secretion of too little interferon-g, too much IL-4, IL-10) making elimination of bacteria difficult. Presentation of HLA B27-derived peptides themselves by HLA class II molecules, or even by HLA class I molecules, has been proposed as an explanation of the association of HLA B27 with disease.

Clinical features

The characteristic clinical features of the spondyloarthropathies are listed in Table 5.

Diagnosis

Five subsets of spondyloarthropathies can be distinguished on clinical grounds: ankylosing spondylitis, reactive arthritis/Reiter's syndrome, psoriatic arthritis, arthritis associated with inflammatory bowel diseases, and undifferentiated spondyloarthropathy.

Diagnostic criteria for spondyloarthropathies are shown in Table 6. Inflammatory back pain is one of the main clinical criteria used. To diagnose this requires four of the following five to be present: insidious onset, onset before the age of 45 years, duration of more than 3 months, morning stiffness, and relief by exercise but not by rest. Other features of possible relevance include waking up at night, alternating buttock pain, initially deep localization, response to non-steroidal anti-inflammatory drugs (**NSAIDs**), other clinical signs of spondyloarthropathies (enthesitis, arthritis, anterior uveitis, family history), elevated acute phase reactants (C-reactive protein, erythrocyte sedimentation rate), and the presence of HLA B27. Note, however, that HLA B27 can never make a diagnosis but increases the probability of an underlying spondyloarthropathy by about tenfold.

Differential diagnosis

The leading clinical symptom of inflammatory back pain may be pain in the lower back radiating to the thighs. Hence an important initial differential diagnosis is sciatica, particularly if symptoms are not insidious but begin abruptly. In inflammatory back pain radiation is more often bilateral than unilateral, rarely extends below the knees, almost never into the foot, and is not associated with paraesthesia. Cough impulse pain may be present. Diagnostic procedures for the detection of disc herniation by magnetic resonance imaging (**MRI**) or computed tomography (**CT**) can be misleading since disc prolapses are found in as many as 30 per cent of normal individuals.

Diffuse idiopathic skeletal hyperostosis or Forestier's disease, a severe radiographic spondylosis, can be difficult to distinguish from spondyloarthropathy. Scoliosis is not usually a marked feature of anklylosing spondylitis. Sacroiliitis occurs in a number of other rheumatic and infectious diseases, as shown in Table 7. The differential diagnosis of peripheral arthritis of the lower limbs includes Lyme arthritis, sarcoidosis (Löfgren's syndrome), gout, and undifferentiated oligoarthritis. The differential diagnosis of enthesitis includes epicondylitis and fibromyalgia, and that of dactylitis is erysipela and infection.

Prognosis

There are no good studies, but the following seem to be poor prognostic factors: hip arthritis, limitation of lumbar spine movements, dactylitis, oligoarthritis, young age at onset (less than 16 years), poor efficacy of NSAIDs, and an erythrocyte sedimentation rate of more than 30 mm in the first hour.

Ankylosing spondylitis

Ankylosing spondylitis is a chronic inflammatory rheumatic disease that mainly affects the axial skeleton, starting in the sacroiliac joints and often progressing to the spine, but peripheral joints, enthesial structures, the anterior uvea, and the aorta can also become affected.

The diagnosis is made on the basis of significant radiological changes in the sacroiliac joints, the typical clinical history of inflammatory back pain and stiffness, and evidence of limited spinal movement and/or chest expansion on physical examination.

Epidemiology

The age of onset is commonly in the twenties, but ankylosing spondylitis can begin in childhood, or considerably later (over the age of 50). The male:female ratio is about 2:1 to 3:1. Approximately 90 per cent of Caucasian ankylosing spondylitis patients are HLA B27-positive. The risk of developing ankylosing spondylitis is increased tenfold in HLA B27-positive individuals, rising to 25 to 30 per cent if a first-degree relative or dizygotic twin is affected, and to 50 to 60 per cent in monozygotic twins. Reactive arthritis, psoriasis, and inflammatory bowel disease are additional, partly independent risk factors.

Immunopathology and pathogenesis

The leading features of ankylosing spondylitis are spinal inflammation and ankylosis, but their cause is unknown. The association of ankylosing spondylitis with bacterial infections is less clear than in reactive arthritis. Antibodies to *Klebsiella pneumoniae* are more frequently detected in patients with ankylosing spondylitis than in healthy controls, but similarly often in patients with Crohn's disease and first-degree relatives of those with ankylosing spondylitis. This finding is probably explained by increased gut permeability, and its predominant clinical association is with peripheral (not axial) arthritis.

The sacroiliac joint is the structure most frequently involved in the initial phase of disease. If biopsy is performed, T cells and macrophages are seen to be the predominant infiltrating cells, with CD4+ and CD8+ T cells both present. The reason for this tropism is unclear. The fact that sacroiliac and spinal joints are affected in diseases caused by mycobacteria and other microbes may argue for a pathogen-triggered pathogenesis in ankylosing spondylitis. However, bacteria associated with reactive arthritis have not been detected in the sacroiliac joints.

Clinical features

The most common initial symptom is inflammatory back pain, commonly in the lower back and the buttocks. Early in the course of disease there may be no limitation of spinal movement or chest expansion. As it progresses, there is restriction of lateral flexion, forward flexion, and extension. There is often a flattening of the lumbar lordosis, or an inability to reverse this on forward flexion. With more advanced disease a thoracic kyphosis develops, with concomitant restriction of thoracic rotation and chest expansion due to inflammation and ankylosis of the costovertebral and costotransverse joints. In severe cases movements of the cervical spine are also restricted in all planes, with dramatic limitation of lateral flexion. The combination of cervical stiffness and severe thoracic kyphosis can lead to difficulties with forward vision. An example of a young patient with severe progressive disease is shown in Fig. 4 and Plate 3. Severe spinal disease is more frequent in men than in women. There is no evidence that pregnancy has a significant impact on the course of the disease.

Fig. 4 30-year-old man with rapidly progressive ankylosing spondylitis (disease of 5 years duration). (See also Plate 3.)

Peripheral joint involvement occurs in 30 to 50 per cent of cases at some time. About 20 to 30 per cent of patients have acute peripheral arthritis of the lower limbs, often with joint effusions as the first symptom, this being especially marked in children. This situation is difficult to differentiate from reactive arthritis. Joint involvement is usually oligoarticular and often asymmetrical. The joints most often involved are the knees, ankles, hips, shoulders, wrists, temporomandibular joints, sternoclavicular joints, manubriosternal joints, costovertebral joints, zygapophyseal joints, and symphysis pubis. Small joints are rarely affected.

Enthesitis occurs at the heel at the insertion of the Achilles tendon (Fig. 1) and the plantar fascia (Fig. 2), and at the iliac crests, the ischial tuberosities, the greater trochanters, and other sites. The diagnosis is often difficult if no swelling is apparent, in which case ultrasound can be revealing. Dactylitis of fingers and toes is uncommon in ankylosing spondylitis, being seen most often in psoriatic arthritis.

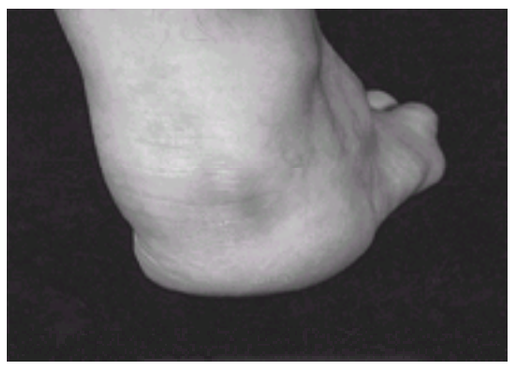

Fig. 1 Enthesitis at the insertion of the Achilles tendon in a patient with reactive arthritis. (See also Plate 1.)

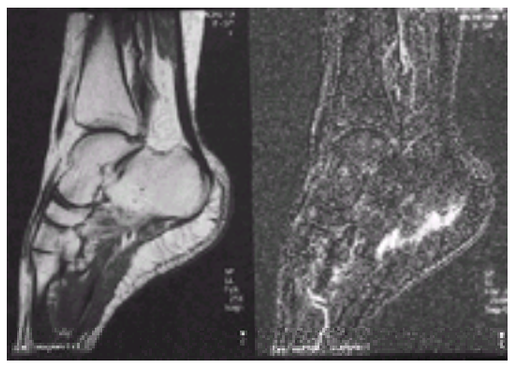

Fig. 2 Magnetic resonance image showing inflammation of the plantar fascia in a patient with undifferentiated spondyloarthropathy.

Fig. 3 Dactylitis of the third finger of the right hand in a patient with undifferentiated spondyloarthropathy. (See also Plate 2.)

Physical examination of the spine and thoracic cage

The physical examination is important in the evaluation of patients with ankylosing spondylitis, in particular to quantitate flexibility of the spine and thoracic cage. The following measurements are useful, but it should be stressed that the values expected of normal individuals are dependent on age and physical training:

1. Schober test (modified):
   - Ventral: with the patient standing upright, a line is drawn across the lumbar spine connecting the two posterior superior iliac spines. Marks are made in the midline over the spine 10 cm cranial and 5 cm caudal to this horizontal line. The patient then bends with legs straight and the distance is measured again. It normally increases by more than 3 cm.
   - Lateral: the distance between the longest finger tip and the floor is measured in the upright position. This is repeated when the patient tries to flex laterally towards the ground as far as possible, normally moving by more than 10 cm.
2. Thoracic excursion. The circumference of the thorax is measured in the fourth intercostal space after maximal inspiration and expiration. It normally alters by more than 3 cm.
3. Occiput/wall distance. In the upright position the patient leans backwards against a wall, and should normally be able to touch the wall with their occiput.
4. Chin/sternum distance. The chin is maximally bent towards the sternum, and should normally be able to touch it.
5. Cervical rotation. The head is rotated to the left and right sides with the angles of rotation measured (normally more than 50°).

6. Intermalleolar distance. The patient tries to stand with their feet together: the malleoli should normally touch.

Physical examination for extra-articular organ involvement

Acute anterior uveitis can occur at any time in the course of disease and is seen in 20 to 30 per cent of patients. It is typically unilateral, but either eye may be affected in separate episodes. Recurrent attacks are common. Aortic regurgitation secondary to aortitis occurs in about 1 per cent of ankylosing spondylitis patients, most frequently in advanced disease, and may be associated with atrioventricular block. Probably on the basis of a restrictive pulmonary defect due to limited chest expansion, apical pulmonary fibrosis occurs in no more than 1 per cent of the patients, especially those with advanced disease. A cauda equina syndrome may complicate severe longstanding disease, with resultant disturbance of the bladder and bowel function. Lumbar diverticulae are seen in myelographic examinations.

Diagnosis

The 1984 modified New York criteria for ankylosing spondylitis are shown in Table 8.

There is a significant diagnostic delay in women (8 years) and in men (5 years). The most probable reason is that back pain is a very frequent complaint, and that primary care and general physicians are often not trained to distinguish inflammatory back pain from other causes of back pain.

Laboratory features

The erythrocyte sedimentation rate and the C-reactive protein are raised in 30 to 50 per cent of patients, with moderate correlation to overall disease activity. Less commonly, serum IgA levels are raised. Mild to severe normochromic normocytic anaemia occurs.

Radiological features

Sacroiliac radiography

Dependent on stage, severity, and duration of disease, there are sacroiliac joint abnormalities in almost all patients. The radiological changes are graded from 0 (normal), through I (minimal changes), II (sclerosis, some erosions), III (severe erosions, pseudodilatation of joint space, limited ankylosis), to IV (ankylosis) (Fig. 5). They are critical for the diagnosis of ankylosing spondylitis and for the differentiation from undifferentiated spondyloarthropathy, but it must be noted that significant inter- and intraobserver variability has been reported—particularly concerning grades I and II—which creates diagnostic problems and confusion. Sclerosis, joint space narrowing, and even synchondrosis occur in healthy elderly individuals. Oblique and other special views are generally not significantly better than normal anteroposterior pelvic radiographs, but can be helpful in a few cases.

Fig. 5 Radiographic sacroiliitis, stage IV in both joints, in a 28-year-old man with ankylosing spondylitis.

Sacroiliac MRI and CT

In early ankylosing spondylitis sacroiliac radiographs may be normal. In clinically suspicious cases dynamic MRI of the sacroiliac joints can be helpful in providing objective evidence of sacroiliitis (see undifferentiated spondyloarthropathy). Active inflammation can be demonstrated by enhancement after application of a contrast agent (gadolinium DTPA) or by special magnetic resonance sequences such as short tau inversion recovery (STIR) or other fat saturation techniques which optimize the visualization of oedematous areas (Fig. 6). Computed tomography of the sacroiliac joints is superior to normal radiographs for documenting bony changes such as erosions and ankylosis. The sacroiliac joint is accessible to biopsy under CT guidance.

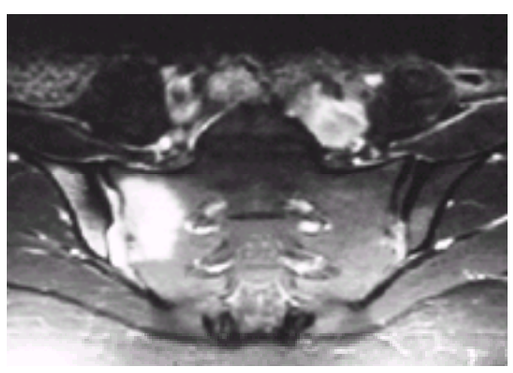

Fig. 6 Dynamic MRI showing right-sided acute sacroiliitis.

Spinal radiography

The characteristic spinal lesion in advanced disease is the syndesmophyte—a bony proliferation originating from an inflammatory area at the ligamentous/discal attachment to the vertebral edge. This early ankylotic structure predominantly grows cranially to fuse with the next vertebral body and has to be distinguished from the spondylophyte, which mainly grows laterally and typically indicates degenerative vertebral disease.

In ankylosing spondylitis the earliest spinal lesions are frequently in the lower thoracic and upper lumbar spine, sometimes preceded by squaring of the vertebrae seen on lateral films. The zygapophyseal joints are frequently involved at all stages. Anterior spondylitis is indicated by lateral spinal radiographs showing hypersclerotic corners (Romanus lesion, Fig. 7). Spondylodiscitis (Anderson lesion) is revealed by erosion of the disc and vertebra with a hypersclerotic lining. In later stages calcification of the anterior and the posterior ligaments occurs, eventually leading to the characteristic 'bamboo spine' (Fig. 8).

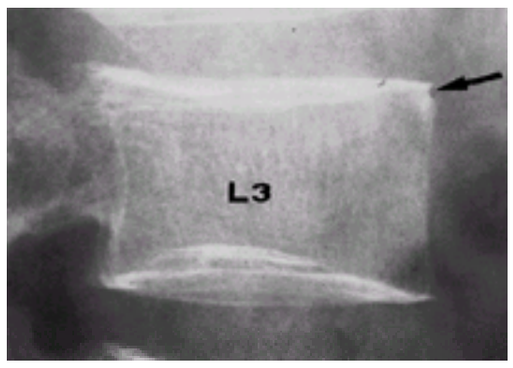

Fig. 8 Radiographic anterior spondylitis (arrow) in a 42-year-old man with ankylosing spondylitis.

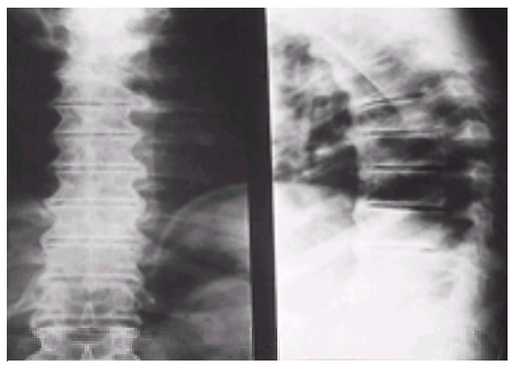

Fig. 8 Spinal radiograph showing classical bamboo spine.

Spinal MRI

Early spinal inflammation (spondylitis, spondylodiscitis) can be detected by dynamic MRI, which can be useful for localizing inflammation in the spine in the early stages when plain radiographs are normal.

Treatment

Although there is no cure for ankylosing spondylitis, several treatments are available. The main therapeutic options are as follows:

1. Acute anti-inflammatory therapy: NSAIDs, local corticosteroids, systemic corticosteroids.
2. Disease-modifying therapy: sulphasalazine, methotrexate(?), gold(?), hydroxychloroquine(?).
3. Anti-resorptive therapy: bisphosphonates (pamidronate)

Non-steroidal anti-inflammatory drugs are better than analgesics and can be used in combination with them. Diclofenac (50–150 mg, and higher in extreme cases), meloxicam (7.5–15 mg), and indomethacin (50–150 mg) are commonly given. Thalidomide and phenylbutazone are reserved for severe cases when other agents have failed. The novel Cox-2 specific agents (rofecoxib 12.5–25 mg, celecoxib 100–200 mg) may be useful. The main risk of NSAIDs is gastrointestinal side-effects: 25 per cent of patients suffer these, ranging from dyspepsia to peptic ulceration and (rarely) bleeding, perforation, and death. It is important to provide patients with proper information about possible symptoms, and prophylactic therapy with proton pump inhibitors or misoprostol is indicated in those at particularly high risk (older age, history of ulcer, disability, comorbidity).

Most patients with ankylosing spondylitis do not respond to small doses of corticosteroids. Transient high-dose steroid treatment has been tried in extreme cases with additional symptoms of inflammatory bowel disease. In our experience, mainly female and HLA B27-negative ankylosing spondylitis patients respond to low-dose corticosteroid therapy.

Sulphasalazine is given in a dosage of 2 to 3 g/day, when effects may be seen after 2 to 4 months. The influence on peripheral joint disease is more significant than for axial symptoms, but this may be due to the preferential study of patients with longstanding disease. Sulphasalazine should mainly be given to patients with early, active disease.

The antitumour necrosis factor-a antibody infliximab in a dosage of 5 mg/kg, which has also been found to be effective in Crohn's disease and rheumatoid arthritis, has been used with significant success in open pilot studies in severe ankylosing spondylitis and recently also in a randomized placebo controlled trial.

The aim of physiotherapy is to maintain and enhance function by improving mobility and muscle strength. Patients affected by spinal stiffness should have physiotherapy on a regular daily basis. Hip replacement is indicated for those with severe hip involvement, and osteotomy can be indicated in cases where visual problems are due to severe kyphosis.

Prognosis

The established myth is that 'patients with ankylosing spondylitis generally do well'. However, one-third are severely disabled and experience intense pain and impairment of health to a comparable degree as those with rheumatoid arthritis. Ankylosing spondylitis does not burn out: disease activity and pain are independent of the duration of the disease. Since the disease usually starts in the second or third decade of life, patients with ankylosing spondylitis typically suffer its effects for many years. The mortality of patients with ankylosing spondylitis may be slightly increased: possible causes of premature death are amyloidosis, NSAID gastropathy (ulcers, bleeding), vertebral fractures, and cardiac or respiratory complications.

Reactive arthritis/Reiter's syndrome

For further information see Chapter 18.7.2.

Undifferentiated spondyloarthropathy

Definition

The term undifferentiated spondyloarthropathy (uspondyloarthropathy) was introduced and defined by the European Spondylarthropathy Study Group in 1991. Terms such as incomplete Reiter's syndrome, syndrome of enthesopathy and arthritis, HLA B27-positive oligoarthritis, and others had been used previously. Patients with uspondyloarthropathies have the typical clinical features of spondyloarthropathies but do not fit into any of the other defined categories. The fact that patients with a clinical picture of peripheral oligoarthritis but without spinal symptoms are also classified/diagnosed with uspondyloarthropathies by the European Spondylarthropathy Study Group criteria can be confusing in some cases. It is possible that uspondyloarthropathies may represent an early form of another spondyloarthropathy subset, or be a genuine spondyloarthropathy subset of their own.

Epidemiology

The prevalence of uspondyloarthropathies is not known precisely, but the frequency is not much less than that of ankylosing spondylitis. They are commoner in men. About 70 per cent of patients are HLA B27-positive. In some contrast to other spondyloarthropathies, late onset disease has been reported.

Clinical features

The main clinical features are inflammatory back pain, asymmetric peripheral arthritis, predominantly of the lower limbs, enthesitis, dactylitis, and anterior uveitis.

Diagnosis

The diagnosis of a uspondyloarthropathy requires inflammatory back pain and/or peripheral arthritis of the lower limbs and at least one other characteristic feature in addition—enthesitis, a positive family history for spondyloarthropathy, psoriasis, or inflammatory bowel disease. Dactylitis, anterior uveitis, and HLA B27 are not part of the European Spondylarthropathy Study Group criteria, but are part of Amor's criteria, and may be taken to indicate uspondyloarthropathy in single cases. Radiographs are not essential for a diagnosis, but in clinically suspicious cases MRI of the sacroiliac joints can be helpful in providing objective evidence of sacroiliitis.

The differential diagnosis of asymmetric peripheral arthritis of the lower limbs in spondyloarthropathies comprises Lyme arthritis, sarcoidosis, gout, osteoarthritis, atypical rheumatoid arthritis, and connective tissue diseases and other rarer conditions.

Treatment

Non-specific therapy with NSAIDs, intra-articular steroid injections, transient immobilization, ice packs, and physiotherapy is similar to that of other arthritides. Sulphasalazine may be effective, but no therapeutic trials with disease-modifying agents has been performed in uspondyloarthropathies.

Prognosis

Knowledge of the long-term prognosis in uspondyloarthropathies is limited. In about 30 to 50 per cent of cases a transition to ankylosing spondylitis has been reported over many years.

Psoriatic arthropathy

Definition

All kinds of arthritis occurring in association with psoriasis can be regarded as psoriatic arthritis, but it is clear that there can be considerable variability in arthritic manifestation. Many patients can be classified as having a spondyloarthropathy, but some are affected in a manner more closely resembling rheumatoid arthritis, and there are other unique forms such as arthritis mutilans. Different forms of psoriasis are associated with different forms of arthritis.

Epidemiology

Psoriasis is common, with a prevalence between 1 and 3 per cent of the population. Arthritic symptoms occur in 20 to 40 per cent of these patients, with the axial skeleton affected in 15 to 25 per cent, such that the overall prevalence of psoriatic arthritis is somewhere around 0.1 to 0.3 per cent. The peak age of onset of psoriatic arthritis is between 20 and 40 years: juvenile disease is rare. Both sexes are equally affected, but women more frequently get polyarthritis and men more often have spinal involvement.

Pathogenesis

Familial aggregation and high concordance rates in monozygotic (70 per cent) compared with dizygotic twins (20 per cent) suggest that there is a clear genetic factor in psoriasis and psoriatic arthritis. About 30 per cent of patients give a clear history of affected first-degree relatives. The genetic impact is thought to be multifactorial. Psoriasis is associated with HLA B13, B17, B37, and HLA DR7. The strongest association is with Cw6 (RR = 24). HLA associations of psoriatic arthritis are with HLA B38 and B39 (peripheral arthritis), with HLA DR4 (symmetric polyarthritis) and HLA B27 (spondylitis).

The importance of genetic linkage may lie in determination of the immunological response to particular antigens, and there has been much interest in the possible role of streptococcal infection. A proliferative response of skin and synovial T cells to streptococcal antigens has been detected in psoriatic arthritis, but also in rheumatoid arthritis, and the (immuno)histology is similar in the two conditions, although some differences have been described.

Koebner's phenomenon is described in psoriasis, when plaques arise at sites of skin injury, scratches, and scars, but the role of trauma in psoriatic arthritis is not clear. Drugs can exacerbate and trigger psoriasis: most well known are b-adrenergic blocking agents, antimalarials, and lithium; and withdrawal of corticosteroids can induce a skin flare; but the relevance of these factors to psoriatic arthritis is uncertain.

Clinical features

Psoriatic arthritis has been divided into five subgroups: distal interphalangeal (overlapping, most common), asymmetrical (spondyloarthropathy-like), symmetrical (rheumatoid arthritis-like), mutilans (unique, rare, Fig. 9 and Plate 4), and spinal (ankylosing spondylitis-like). It must be stressed, however, that these subgroups are not clearcut. In a recent study, when patients were evaluated over a period of 8 years, the initial classification pattern changed significantly over time, and finally only two categories remained: peripheral disease without axial involvement (70 per cent) and axial involvement with or without peripheral arthritis (30 per cent). The latter was correlated with duration of the disease and magnitude of joint involvement. Erosions were found in 70 per cent of the patients.

Fig. 9 Severe psoriatic arthritis (arthritis mutilans). (see also Plate 4.)

Psoriasis precedes joint disease in the majority of cases (70–80 per cent), both occur simulataneously in 15 per cent, and in about 10 per cent arthritis comes first. There is poor correlation between onset, severity, and activity of psoriatic skin lesions and arthritis. More than 80 per cent of patients with psoriatic arthritis have nail dystrophy, while this is the case in only 20 per cent of those with uncomplicated skin disease. Nail dystrophy, ranging from some to many nail pits and horizontal (not longitudinal) ridging to onycholysis, occurs most often in those with distal interphalangeal involvement. In some patients the involvement of interphalangeal joints and

nails is closely correlated, with both appearing on the same finger(s). Acute anterior uveitis occurs mainly in those with radiological sacroiliitis and ankylosing spondylitis.

Different types of psoriatic skin involvement lead to different types of arthropathy. Most frequent is the common psoriasis vulgaris, but a type of skin disease that frequently affects the palms of the hands and soles of the feet with many psoriatic plaques is also seen: pustolosis palmaris et plantaris. This type is associated with the SAPHO syndrome (see below), which is related to the spondyloarthropathies but has unique features that justify the designation as a separate subset of these disorders.

A severe form of psoriatic arthritis can occur in HIV-infected patients, although is not clear whether HIV increases the overall prevalence of psoriatic arthritis. Severe peripheral enthesitis (predominantly of the heel) and dactylitis are characteristic. Knee arthritis can be rapidly destructive. Axial inflammation is less frequent.

There is a classical overlap between psoriatic arthritis and reactive arthritis in the form of keratoderma blenorrhagicum—a desquamating psoriasis-like lesion mostly occurring on the soles of the feet in patients with Reiter's syndrome.

Psoriatic arthritis often improves during pregnancy. There is no adverse effect of the disease on mother or child.

Diagnosis

Scaling erythematous papules and plaques on the scalp and extensor aspects of the extremities, often surmounted by a silvery white micaceous scale that is easily removed, are suggestive of psoriasis. Elbows and knees are often affected. The diagnosis of psoriatic arthritis is based on the presence of these characteristic skin lesions, which are not always obvious. Less accessible areas such as the navel, perineum, and scalp need to be examined carefully. The patient should be asked whether they have a family history of psoriasis or psoriatic arthritis.

Since psoriasis is a frequent disease, it must be remembered that a patient with psoriasis can have an attack of gout or another form of arthritis. The diagnosis of psoriatic arthritis should be considered in those without skin lesions if there is distal interphalangeal joint involvement, dactylitis, the involvement of a whole finger or toe, tendon sheaths and bone of an affected limb, and/or typical radiographic changes.

Laboratory and radiological features

Acute phase reactants are often raised. HLA determinations including HLA B27 do not provide diagnostic help in those with psoriatic arthritis, but in HLA B27-negative patients who appear to have ankylosing spondylitis, psoriasis should always be searched for. The presence of rheumatoid factor does not formally exclude a diagnosis of psoriatic arthritis, there being a background prevalence of rheumatoid factor positivity, but a positive result should always make the physician consider the diagnosis carefully.

The distribution of radiological changes reflects clinical involvement, with the interphalangeal joints involved earlier than larger joints. A characteristic lesion in advanced cases is the so-called pencil-in-cup deformity (Fig. 10), which evolves by resorption of the distal end of a phalanx or metacarpal with uniform deep erosion of the end of the corresponding distal phalanx. In some cases the joints can be completely destroyed and invisible on the radiograph.

Fig. 10 Radiograph of the hands showing destructive psoriatic arthritis.

Radiological grounds for thinking the diagnosis more likely to be psoriatic arthritis than rheumatoid arthritis are distal interphalangeal joint involvement, asymmetric joint involvement, marginal erosions with adjacent bone proliferation (whiskering), osteolysis, periostitis, proliferative new bone formation, and ankylosis. Radiological sacroiliitis is a finding in 20 to 40 per cent of patients. The axial disease in psoriatic arthritis can be indistinguishable from that in primary ankylosing spondylitis, but in psoriatic arthritis the following are more likely: asymmetrical sacroiliitis, less zygapophyseal joint involvement, fewer, coarser, and asymmetric syndesmophytes, and bony bridging that is more often asymmetrical. Psoriatic arthritis syndesmophytes can be indistinguishable from spondylophytes typical of diffuse idiopathic skeletal hyperostosis (Forestier's disease).

When scintigraphy is used to detect the extent and localization of arthritis, an increased uptake of the isotope ^{99}Tc can frequently be detected in the sternoclavicular and manubriosternal joints—this is not necessarily associated with clinical symptoms.

Treatment

Many patients improve with the use of NSAIDs and intra-articular steroids, especially in the case of large joint involvement or flexor tenosynovitis. However, 20 to 40 per cent of patients will not improve and need to be treated with disease-modifying antirheumatic drugs. Sulphasalazine 2 to 3 g daily is often effective against arthritis. Methotrexate 7.5 to 40 mg daily is also good for arthritis, and even better for the skin. Intramuscular gold and azathioprine can be tried. Antimalarials and penicillamine are not used; the former may exacerbate psoriasis. Cyclosporin A is given in severe cases. There is limited information on the use of combination therapy. Systemic corticosteroids are limited to extreme cases of arthritis: psoriasis usually flares when they are withdrawn.

Local skin therapy has no effect on joint symptoms. Etretinate is not clearly beneficial for arthritis and may cause arthralgias and many other adverse reactions. The role of physiotherapy is similar to that in other spondyloarthropathies, and there are no special considerations for surgical intervention in psoriatic arthritis, apart from the fact that the presence of florid skin lesions close to a joint is a relative contraindication to surgery.

Prognosis

Severe psoriasis can lead to significant disability. There are only limited data from long-term studies in psoriatic arthritis. In cross-sectional studies 10 to 20 per cent of patients with psoriatic arthritis were in a poor functional class; the HLA antigens HLA B27, HLA B39, and DQw3 have been associated with such an outcome.

Arthritis associated with inflammatory bowel disease

Definition

An arthropathy with various clinical symptoms occurring in association with Crohn's disease and ulcerative colitis is termed arthritis associated with inflammatory bowel disease. Other forms of arthropathy occurring in association with enteropathy are Morbus, Whipple's disease, and arthritis after intestinal bypass surgery.

History

A relationship between gut and joint disease was postulated in 1922 when Smith treated arthritis patients with segmental bowel surgery. Bargen and Hench in 1929 and 1935 described arthritis in association with ulcerative colitis and Crohn's disease. Moll and Wright included arthritis associated with inflammatory bowel disease in the concept of spondyloarthropathies in 1973. Mielants and Veys described Crohn-like gut lesions in all subsets of spondyloarthropathy in 1984.

Epidemiology

The prevalence of Crohn's disease and ulcerative colitis is between 0.05 and 0.1 per cent of the population, generally higher in Whites and Jews. The peak occurrence of both diseases is between 15 and 35 years, but it may appear in every decade of life; both sexes are equally involved. Arthritis associated with inflammatory bowel disease occurs in 10 to 30 per cent of patients with inflammatory bowel disease; in general more frequently in Crohn's disease than in ulcerative colitis, and more often in patients with colonic involvement and in those with extended bowel disease.

There is a genetic predisposition for inflammatory bowel disease with documented familial aggregation for both Crohn's disease and ulcerative colitis. The association with HLA B16, HLA B18, and HLA B62 is not strong. The peripheral arthritis of inflammatory bowel disease is not associated with HLA B27, but axial inflammation is (50 per cent). The patient with inflammatory bowel disease who is, by chance, HLA B27-positive, is at high risk of developing spondylitis. The relative frequency of sacroiliitis and ankylosing spondylitis in inflammatory bowel diseases varies between 2 and 20 per cent or more, partly depending on the sensitivity of the diagnostic imaging procedure. Four per cent of patients with ankylosing spondylitis develop overt inflammatory bowel disease, while 60 per cent have microscopically detectable Crohn-like gut lesions.

Pathogenesis

The pathogenesis of inflammatory bowel disease and arthritis associated with inflammatory bowel disease is not known. One hypothesis is of an aberrant immune response to gut bacteria, with gut inflammation leading to increased permeability, allowing bacteria to cross the mucosal border and get access to joints. There is some evidence from the HLA B27 transgenic rat model that gut and joints are closely linked: susceptible rats get both colitis and arthritis once they have left a germfree environment.

Clinical features

Patients with ulcerative colitis and Crohn's disease typically present with bloody diarrhoea and abdominal pain, and in severe cases with fever, weight loss, and fatigue. For further details of gastrointestinal and other non-rheumatological presentations, and criteria for diagnosis, see Section 14.

Similar to the other spondyloarthropathies the arthritis is mostly asymmetric and predominantly affects the lower limbs. The arthritis is migratory, often transient, but tends to recur. It does not frequently become chronic but may be associated with erosive disease in some patients. Flaring of gut symptoms is often associated with arthritis, especially in ulcerative colitis. In Crohn's disease, patients experience significantly fewer joint symptoms after colectomy.

Two types of arthropathy were distinguished in a recent study of almost 1500 patients with inflammatory bowel disease, essentially on the basis of the number of joints involved and importantly without knowledge of spinal radiographs. Pauciarticular disease (type I, fewer than five joints involved) affected 3.6 per cent of patients with ulcerative colitis and 6 per cent of those with Crohn's disease and was acute and self-limiting, with episodes lasting 4 to 5 weeks, in 83 and 79 per cent of the cases. Polyarticular disease (type II, five or more joints) affected 2.5 per cent of patients with ulcerative colitis and 4 per cent of those with Crohn's disease and was associated with persistent symptoms in 87 and 89 per cent of the cases.

The onset of peripheral arthritis is associated with exacerbations of colitis, but there is no link between enteric and spinal symptoms. Acute anterior uveitis occurs in 10 per cent of patients with inflammatory bowel disease. It is associated with axial involvement and with HLA B27. Compared with other spondyloarthropathies, the type of uveitis is somewhat different in inflammatory bowel diseases: posterior uveitis and scleritis may occur. The most common skin lesion in arthritis associated with inflammatory bowel disease is erythema nodosum, occurring in association with exacerbation of enteritis.

Diagnosis

Most arthritic symptoms occurring in patients with inflammatory bowel disease can generally be attributed to spondyloarthropathies. However, as in psoriasis, patients can have more than one disease (osteoarthritis, etc.). As many as 50 to 60 per cent of all patients with ankylosing spondylitis have gut lesions resembling those in Crohn's disease, but the majority are asymptomatic. Clinically apparent ankylosing spondylitis often precedes Crohn-like symptoms. This spectrum of diseases clearly and typically belongs to the spectrum of spondyloarthropathies. The differentiation (if needed) will rarely cause problems since one disease is usually predominant. Along with psoriasis, inflammatory bowel disease should always be looked for in HLA B27-negative patients who appear to have ankylosing spondylitis.

Treatment

Treatment of inflammatory bowel diseases is always the first consideration and will probably influence the peripheral arthritis. Treatment with NSAIDs may be effective for arthritis and spondylitis but can exacerbate the bowel disease. There are few data on the use of disease-modifying antirheumatic drugs. Sulphasalazine is effective in ulcerative colitis and other spondyloarthropathies and may, accordingly, be used in arthritis associated with inflammatory bowel disease. Azathioprine is effective in Crohn's disease and can be tried to treat severe and chronic joint disease. Corticosteroids are the therapy of choice in acute inflammatory bowel disease and will generally help arthritis, but they should not be used for mild and transient joint symptoms.

Prognosis

Patients with inflammatory bowel disease have increased mortality due to peritonitis and sepsis. By contrast, the prognosis of arthritis associated with inflammatory bowel disease is generally good. Joint destruction is a rare event. Patients may have ankylosing spondylitis at presentation of inflammatory bowel disease, or develop this later.

SAPHO syndrome

Definition

The acronym SAPHO stands for **S**ynovitis, **A**cne, **P**ustolosis palmaris et plantaris, **H**yperostosis, and **O**steitis. French workers proposed SAPHO as a unifying diagnosis for several idiopathic bone and skin diseases, thereby combining over 50 different terms published in the literature (including pustulotic arthro-osteitis, chronic multifocal osteomyelitis, Tietze syndrome (German), and acquired hyperostosis syndrome). Their description of the common symptoms and overlapping features of this heterogenous group of rheumatic joint, bone, and skin diseases has led to better recognition of the relatively rare condition.

There is an argument that SAPHO simply represents a subset of psoriatic arthropathy; also that it might not really belong to the spondyloarthropathies at all. Only 43 per cent of patients with SAPHO fulfilled the European Spondylarthropathy Study Group criteria for spondyloarthropathies, and only 1 in 19 was HLA B27-positive in one follow-up study.

Pathogenesis

The pathogenesis is unclear. Some authors think that it is similar to that of reactive arthritis. *Propionibacterium acnes*, which can induce arthritis in animals, has been detected in acne lesions and grown from osteitic lesions in some cases. However, cultures are negative in the vast majority of cases, and antibiotics are ineffective.

Diagnosis

There are no evaluated diagnostic criteria for SAPHO. Most convincing clinically is the combination of a classical skin symptom—such as pustolosis or significant acne (acne conglobata and acne fulminans or hidradenitis suppurativa)—with a characteristic joint or bone lesion such as arthritis of the sternoclavicular joint, osteitis, or hyperostosis in the anterior chest wall.

Diagnosis is important to avoid unnecessary biopsy procedures, but can be very difficult, especially in those without typical skin lesions. The most important differential diagnoses are bacterial osteomyelitis and malignancy. The pattern of joints affected differs from other rheumatic diseases: the sternoclavicular joint (Fig. 11 and Fig. 12 and Plate 5), the clavicle, the ribs, and the mandible are frequently involved by arthritis, osteitis, and/or hyperostosis. Sacroiliitis, mostly unilateral, occurs in one-third of patients.

Fig. 11 Arthritis/hyperostosis of the left sternoclavicular joint in a 52-year-old man with SAPHO syndrome. (See also Plate 5.)

Fig. 12 Radiograph showing severe osteitis of the left sternoclavicular joint in a 35-year-old woman with SAPHO syndrome.

Treatment

Analgesics, NSAIDs, and intra-articular steroids are usually effective. In severe cases systemic corticosteroids should be considered. Immunosuppressive agents can be added if the steroid dose cannot be tapered to less than 10 mg/day. Sulphasalzine, azathioprine, and methotrexate have been tried successfully in some cases. Radiation therapy can also be effective in refractory cases. No controlled studies have been performed.

Prognosis

The course of disease is very variable. Initially, occurrence of several flares per year is common. Further progress is usually favourable, but complications such as axillary vein and C8 compression can occur. Some patients may develop ankylosis and few progress into ankylosing spondylitis.

Other enteric arthropathies

Whipple's disease

Whipple's disease is a rare systemic disease which usually involves the small intestine (see Chapter 14.9.6). The associated arthritis is often symmetric and polyarticular, and may antedate the intestinal complaints by years. It is not usually destructive. Axial involvement occurs but is not typical.

Arthritis associated with coeliac disease

For a description of coeliac disease (gluten-sensitive enteropathy) see Chapter 14.9.3. The joint manifestations show a striking response to a gluten-free diet, which strongly suggests a causal relationship. The pattern of arthritis is very variable and overt bowel symptoms are absent in half of cases, making diagnosis difficult. The lumbar spine, hips, knees, shoulders, elbows, wrists, and ankles are most frequently affected, often symmetrically. The arthritis is not destructive. HLA B8 and DR3 are frequently found in the patients. The pathogenesis is unclear.

Arthropathies associated with collagenous colitis

Collagenous colitis is a chronic diarrhoeal disease characterized by a normal or near-normal mucosa endoscopically and a thick subepithelial collagen layer. More than half of patients with this disorder have some form of arthritis and use NSAIDs regularly.

Arthropathies associated with intestinal bypass surgery

Arthritis has been reported in 5 to 50 per cent of patients in the first 3 years after jejunoileal bypass surgery. A symmetric peripheral polyarthritis involves the knees, wrists, metacarpophalangeal and metatarsophalangeal joints, elbows, proximal interphalangeal joints, and ankles and is usually non-destructive. Almost half of those affected also have vesicopustular skin lesions. No specific HLA association has been found, but two previosly healthy HLA B27-positive patients developed spondylitis. Bacterial overgrowth of the blind loop is critical for pathogenesis.

Further reading

Introduction

Amor B *et al.* (1994). Predictive factors for the longterm outcome of spondyloarthropathies. *Journal of Rheumatology* **21**, 1883–7.

Braun J *et al.* (1998). Prevalence of spondylarthropathies in HLA-B27 positive and negative blood donors. *Arthritis and Rheumatism* **41**, 58–67.

Braun J, Bollow M, Sieper J (1998). Radiologic diagnosis and pathology of the spondyloarthropathies. *Rheumatic Disease Clinics of North America* **24**, 697–735.

Calin A *et al.* (1977). Clinical history as a screening test for ankylosing spondylitis. *Journal of the American Medical Association* **237**, 2613–14.

Dougados M *et al*. (1991). The European Spondylarthropathy Study Group preliminary criteria for the classification of spondylarthropathy. *Arthritis and Rheumatism* **34**, 1218–27.

Khan MA, van der Linden SM (1990). A wider spectrum of spondyloarthropathies. *Seminars in Arthritis and Rheumatism* **20**, 107–13.

Moll JM *et al.* (1974). Associations between ankylosing spondylitis, psoriatic arthritis, Reiter's disease, the intestinal arthropathies, and Behcet's syndrome. *Medicine (Baltimore)* **53**, 343–64.

Sieper J, Braun J (1995). Pathogenesis of spondylarthropathies. Persistent bacterial antigen, autoimmunity, or both? *Arthritis and Rheumatism* **38**, 1547–54.

Ankylosing spondylitis

Bollow M *et al.*(2000). Quantitative analysis of sacroiliac biopsies in spondyloarthopathies: T cells and macrophages predominate in early and active sacroiliitis–cellularity correlates with the degree of enhancement detected by magnetic resonance imaging. *Annals of Rheumatic Disease* **59**, 135–40.

Braun J *et al*. (1995). Use of immunohistologic and *in situ* hybridization techniques in the examination of sacroiliac joint biopsy specimens from patients with ankylosing spondylitis. *Arthritis and Rheumatism* **38**, 499–505.

Gran JT, Skomsvoll JF (1997). The outcome of ankylosing spondylitis: a study of 100 patients. *British Journal of Rheumatology* **36**, 766–71.

Kennedy LG, Edmunds L, Calin A (1993). The natural history of ankylosing spondylitis. Does it burn out? *Journal of Rheumatology* **20**, 688–92.

Mau W *et al.* (1988). Clinical features and prognosis of patients with possible ankylosing spondylitis. Results of a 10-year followup. *Journal of Rheumatology* **15**, 1109–14.

McGonagle D *et al.* (1998). Characteristic magnetic resonance imaging entheseal changes of knee synovitis in spondylarthropathy. *Arthritis and Rheumatism* **41**, 694–700.

van der Linden S, Valkenburg HA, Cats A (1984). Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis and Rheumatism* **27**, 361–8.

Zink A *et al.*(2000). Disability and handicap in rheumatoid arthritis and ankylosing spondylitis—results from the German rheumatological database. *Journal of Rheumatology* **27**, 613–22.

Uspondyloarthropathies

Olivieri I *et al.* (1995). Late onset undifferentiated seronegative spondyloarthropathy. *Journal of Rheumatology* **22**, 899–903.

Zeidler H, Mau W, Khan A (1992). Undifferentiated spondyloarthropathies. *Rheumatic Disease Clinics of North America* **18**, 187–202.

Psoriatic arthritis

Gladman DD (1998). Psoriatic arthritis. *Rheumatic Disease Clinics of North America* **24**, 829–44.

Helliwell P *et al.* (1991). A re-evaluation of the osteoarticular manifestations of psoriasis. *British Journal of Rheumatology* **30**, 339–45.

Marsal S *et al*. (1999). Clinical, radiographic and HLA associations as markers for different patterns of psoriatic arthritis. *Rheumatology* **38**, 332–7.

Reece RJ *et al*. (1999). Distinct vascular patterns of early synovitis in psoriatic, reactive and rheumatoid arthritis. *Arthritis and Rheumatism* **42**, 1481–4.

Richter Cohen M *et al*. (1999). Baseline relationships between psoriasis and psoriatic arthritis: analysis of 221 patients with active psoriatic arthritis. *Journal of Rheumatology* **26**, 1752–6.

Salvarani C *et al*. (1995). Prevalence of psoriatic arthritis in Italian psoriatic patients. *Journal of Rheumatology* **22**, 1499–503.

Arthritis associated with inflammatory bowel disease

Leirisalo-Repo M *et al.* (1994). High frequency of silent inflammatory bowel disease in spondyloarthropathy. *Arthritis and Rheumatism* **37**, 23–35.

Mielants H *et al.* (1996). Course of gut inflammation in spondylarthropathies and therapeutic consequences. *Baillière's Clinical Rheumatology* **10**, 147–64.

Orchard TR, Jewell DP (1999). The importance of ileocaecal integrity in the arthritic complications of Crohn's disease. *Inflammatory Bowel Disease* **5**, 92–7.

Orchard TR, Wordsworth BP, Jewell DP (1998). Peripheral arthropathies in inflammatory bowel disease: their articular distribution and natural history. *Gut* **42**, 387–91.

Taurog J *et al*. (1994). The germfree state prevents the development of gut and joint inflammatory disease in HLA B27 transgenic rats. *Journal of Experimental Medicine* **180**, 2359–64.

SAPHO syndrome

Boutin RD, Resnick D (1998). The SAPHO syndrome: an evolving concept for unifying several idiopathic disorders of bone and skin. *American Journal of Rheumatology* **170**, 585–91.

Kahn MF, Khan MA (1994). The SAPHO syndrome. *Baillière's Clinical Rheumatology* **8**, 333–62.

Koehler H *et al.* (1975). Sterno-kosto-klavikuläre Hyperostose. *Deutsche Medizinische Wochenschrift* **100**, 1519–23.

Maugars Y *et al.* (1995). SAPHO syndrome: a followup study of 19 cases with special emphasis on enthesis involvement. *Journal of Rheumatology* **22**, 2135–41.

Sonozaki H *et al*. (1981). Clinical features of 39 patients with pustolotic arthroosteitis. *Annals of Rheumatic Diseases* **40**, 547–53.

Other enteric arthropathies

Fleming JL, Wiesner RH, Shorter RG (1988). Whipple's disease: clinical, biochemical and histopathological features and assessment of treatment in 29 patients. *Mayo Clinic Proceedings* **63**, 539–51.

Goff JS *et al.* (1997). Collagenous colitis: histopathology and clinical course. *American Journal of Gastroenterology* **92**, 57–60.

Pinals RS (1986). Arthritis associated with gluten-sensitive arthropathy. *Journal of Rheumatology* **13**, 201–4.

Stein HE *et al.* (1981). The intestinal bypass arthritis-dermatitis syndrome. *Arthritis and Rheumatism* **24**, 684–90.

18.7.1 Pyogenic arthritis

Anthony R. Berendt

Introduction

Pyogenic arthritis, which may be acute or chronic, describes infection and resulting inflammation in a joint, native or prosthetic. It should not be confused with postinfectious (reactive) arthritis (discussed in Section 18.8). As with other musculoskeletal infections, failings in diagnosis or management may have long-term functional consequences, and clinicians should therefore know when to consider the diagnosis and obtain expert help.

Aetiology

Acute pyogenic arthritis may be primary (by haematogenous spread), or secondary (to trauma, surgery, or arthrocentesis). Organisms that cause primary septic arthritis are usually aggressive pathogens capable of causing a bacteraemia, seeding the joint, and multiplying within it, hence they are also common causes of septicaemia, with *Staphylococcus aureus* dominating in most circumstances. The causes of secondary and chronic septic arthritis are more diverse because they include skin and environmental flora, together with lower-grade pathogens, as well as all the causes of acute infection.

Table 1 in Chapter 19.2 shows the common pathogens involved in both pyogenic arthritis and osteomyelitis.

Epidemiology

The incidence of pyogenic arthritis has been estimated at 7 in 100 000: this is highest in children and the elderly, and more common in males. The increased incidence in the elderly probably reflects a higher prevalence of potential sources of bacteraemia such as urinary tract infection, skin ulceration, pneumonia, and hospitalization with intravenous and/or urinary catheterization. Infection complicates some 0.5 to 2 per cent of total joint replacements, but the true prevalence of prosthetic joint infection is unknown.

Pathogenesis and pathophysiology

In primary septic arthritis, organisms must exit the bloodstream and access the joint. In the case of *S. aureus*, invasion of endothelial cells can occur through interactions between bacterial fibronectin-binding proteins and cell surface-associated fibronectin. This triggers integrin-dependent uptake of bacteria and may be a key first step in bacteraemic seeding.

S. aureus releases a number of toxins and proteases thought to affect host defences. It also expresses numerous cell wall-associated adhesins that mediate attachment to the matrix proteins associated with cell surfaces, cartilage, and bone. Animal models demonstrate that T-cell dependent inflammation plays a central role in damage to articular cartilage following an acute inflammatory response. In these models immunomodulation (for example with corticosteroids) can substantially reduce arthritis, but at the expense of host survival if antibiotic therapy is not also given. Thus the host response appears to reduce the risk of bacteraemia and death, but at the cost of joint damage. If not fatal through septicaemia, untreated septic arthritis generally causes joint destruction or fusion, sometimes with sinus formation and persistent infection (see Fig. 1(a) and (b)).

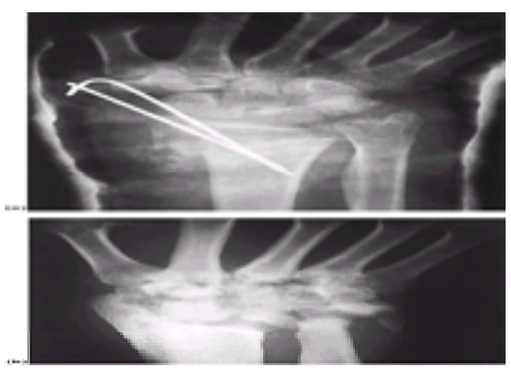

Fig. 1 (a) A Colles fracture fixed with percutaneous K wires. (b) A few months later, after *S. aureus* infection associated with the K wires has led to septic arthritis in the wrist, the wrist joint is completely destroyed.

In a prosthetic joint, the presence of foreign material impairs local antibacterial defences. Bacteria adhere to plastic or metal (in some cases via fibronectin) and in this state become relatively resistant to the action of antibiotics and to phagocytosis. Ineffective but chronic inflammation causes pain and triggers bone loss, with subsequent mechanical loosening.

Clinical features

Patients typically present with fever and an acutely painful joint. The joints involving the long bones are most commonly affected (knee, hip, shoulder, elbow, wrist, and ankle). There may be bacteraemia (in some series up to 70 per cent), giving prostration, vomiting, or hypotension. Infants localize pain poorly and commonly present refusing to use the affected limb. Adults may be unable to localize pain if a sternoclavicular, acromioclavicular, sternocostal, or manubriosternal joint is involved. Infections in these locations often present as chest wall pain. Sacroiliac joint infection presents as buttock or low back pain and may mimic hip or spine pathology.

Clinical examination reveals a joint that is swollen, warm to the touch, tender on palpation, and painful, frequently exquisitely so, on active or passive movement. To minimize pain, the patient will often nurse the joint in a neutral position. A joint effusion is usually present and this may be accompanied by synovitis, depending on the duration of the history. Erythema, not usually prominent, may signal the presence of bursitis.

Septic arthritis must be distinguished from other acute monoarthropathies, notably gout, pyrophosphate arthropathy, and haemarthrosis. Rheumatoid and reactive arthritis can initially present with involvement of only a single joint. Infection is most commonly monoarticular, but multiple joints can be involved. In the patient with

known inflammatory arthritis, polyarticular infection may be mistaken for a flare in the underlying disease.

Prosthetic joint infection may present as an acute wound infection, a periarticular abscess, an acute arthritis or with loosening of the implant (as pain). The differential is from superficial wound infection, haemarthrosis, periprosthetic fracture or dislocation, and aseptic loosening. A sinus discharging in or near the operative scar represents infection of the prosthesis until proved otherwise.

Pathology

The synovial fluid contains polymorphs and the synovium shows an acute inflammatory response with a fibrinous exudate on its surface, which may be ulcerated. A chronic synovitis may develop, with a lymphocytic and mononuclear infiltrate. Late changes include chondrolysis and the development of subarticular osteomyelitis. Tissue from infected prosthetic joints shows a polymorph infiltrate accompanied by chronic inflammatory changes representing a reaction to foreign materials.

Laboratory diagnosis

The criterion for diagnosis of pyogenic arthritis is the isolation of a recognized pathogen from samples of synovium or synovial fluid obtained through biopsy or aspiration. Infected synovial fluid is generally turbid or purulent. It should be sent for Gram stain, semiquantitative or quantitative white cell count, examination under polarized light for pyrophosphate or uric acid crystals, and culture. If tuberculosis, brucellosis, or fungi are suspected, the laboratory should be advised so that the sample can be appropriately processed. For suspected *Neisseria gonorrhoeae*, urethral, endocervical, throat, and rectal swabs should also be obtained for microscopy and culture.

Blood cultures should always be obtained. The white cell count, C-reactive protein, and erythrocyte sedimentation rate are usually raised, but can also be elevated during flares of inflammatory arthritis or acute crystal arthropathy. Their value is probably greatest in following the response to treatment. Measurement of serum uric acid may be elevated in gout, but cannot be used to establish or refute this diagnosis. Serological tests may be of retrospective value in diagnosing *Borrelia burgdorferi* (Lyme disease) and *Brucella* spp. Detection of bacterial nucleic acid in joints remains a research technique.

Plain radiographs can show fracture, effusion, chondrocalcinosis, bone destruction, or loss of joint space. An effusion may be seen acutely, but bony changes appear slowly. If seen at presentation they either indicate chronic infection or represent underlying arthritis. A radiograph at presentation provides a useful baseline for subsequent comparisons. The role of CT scan or magnetic resonance imaging is to demonstrate or exclude surgical disease in the joint or neighbouring bone (Fig. 2). Ultrasound scanning may assist in this and in distinguishing between effusion and synovitis, allowing diagnostic samples to be obtained more reliably.

Fig. 2 Marked synovitis, but no osteomyelitis, in a 10-year-old with group A streptococcal infection of the knee.

Treatment

Acute septic arthritis poses a threat to the joint and is an orthopaedic or rheumatological emergency. Treatment should generally be in an inpatient hospital setting. Patients with suspected chronic septic arthritis may initially be managed as outpatients, but most eventually require surgical intervention.

After obtaining blood cultures and (when possible) synovial fluid, acute pyogenic arthritis should be treated promptly with intravenous antibiotics active against aerobic Gram-positive cocci and Gram-negative organisms. Appropriate regimens would be cefuroxime (or another antistaphylococcal cephalosporin) or a high-dose semisynthetic antistaphylococcal penicillin (flucloxacillin, dicloxacillin, or nafcillin), with or without an aminoglycoside. Patients allergic to b-lactams, or with risk factors for methicillin-resistant *S. aureus*, should receive vancomycin, usually with an aminoglycoside, until culture results are obtained. Treatment can then be modified.

Urgent consultation with an orthopaedic surgeon is advised. Arthroscopic washout has largely replaced arthrotomy, reducing morbidity. Surgery can sometimes be avoided altogether by aspiration once or twice daily until clinical response is evident, but may still be needed if there is clinical deterioration or failure to settle within 5 days. This type of treatment can be applied to children or adults, particularly when anaesthesia is thought to carry high risks. Whether delaying surgery in this way gives worse outcomes than immediate washout is unknown, but there is consensus on the need for prompt surgery on the hip and shoulder joints. The reflection of the capsular vessels up the necks of the humerus and femur makes them vulnerable to thrombosis and subsequent avascular necrosis of the femoral or humeral head. If this occurs, joint destruction, with or without chronic infection of the dead bone, is inevitable.

The optimal duration and mode of administration of antibiotics is unknown. In uncomplicated infection 2 to 3 weeks is adequate, depending on the pathogen (shorter for streptococci, longer for *S. aureus* and aerobic Gram-negative rods). In children it is possible to convert to oral therapy within 48 to 72 h of defervescence, provided that there has been a rapid clinical response. This strategy requires the organism to be sensitive to a reliably bioavailable oral antibiotic, the parents or carers to understand clearly the importance of adhering to the antibiotic regimen, and the clinician to monitor clinical progress carefully. Some authorities treat adults with an oral regimen provided that similar criteria are met, but intravenous antibiotics may be preferred when there has been accompanying bacteraemia or where there are concerns about bony involvement, absorption of antibiotic, or adherence. Many patients are suitable for intravenous therapy at home, provided this is properly organized and supervised.

Chronic septic arthritis has generally led to joint destruction by the time the patient presents. Surgical debridement, arthrodesis, or joint replacement may be necessary, the latter after a considerable interval free of infection. Surgery is commonly required in prosthetic joint infection, although such problems can occasionally be treated successfully with retention of the implant. Antibiotic treatment is usually prolonged in chronic or prosthetic joint infection.

Prognosis

If diagnosed and treated promptly, the prognosis of acute native joint infection is good, with many patients making a complete recovery. Some joint damage is likely when the diagnosis is made late. Infection in young children may lead to disturbance of the growth plate around the infected joint, causing deformity. Mortality is low in uncomplicated septic arthritis, higher when it is complicated by *S. aureus* bacteraemia (up to 20 per cent) and highest in multijoint disease (50 per cent). Recurrence is uncommon and generally indicates a persisting surgical focus in relation to the joint.

Outcomes are less favourable in prosthetic joints, which can be salvaged in 30 to 70 per cent of cases where the prosthesis is retained. Infection can be eradicated in up to 90 per cent of cases with revision surgery, but with very much poorer results when revision surgery for infection is itself complicated by further infection. This and the need for expert surgical and microbiological input, as well as the considerable comorbidity such patients often have, makes the management of infected prosthetic joints a formidable challenge.

Prevention and control

There are no proven means of preventing primary pyogenic arthritis. Secondary cases can be prevented by meticulous attention to infection control measures whenever a joint is aspirated or operated on, and by thorough cleaning and debridement when a joint is contaminated through trauma.

Occupational, quality of life, and psychosocial aspects

The pain of acute pyogenic arthritis generally resolves within the first 1 to 2 weeks of successful treatment, but stiffness and swelling usually persist for very much longer. Chronic infections have significant effects on quality of life through pain and poor function. This is most prolonged and severe in prosthetic joint infection.

Areas of uncertainty needing further research

Further understanding of pathogenesis could potentially offer new targets for therapy or prevention, while diagnostic sensitivity could be improved by robust, well-evaluated molecular methods. Numerous aspects of treatment, including the role of surgery and the duration and route of administration of antibiotics, await clarification in prospective studies, both for native and prosthetic joint infections.

Further reading

Berendt AR (1999). Infections of prosthetic joints and related problems. In: Armstrong D, Cohen J, eds. *Infectious diseases*, pp 2.44.1–2.44.6 Mosby, London. [Review of prosthetic joint infections.]

Dubost J-J *et al*. (1993). Polyarticular septic arthritis. *Medicine* **72**, 296–310. [An extensive review of this challenging condition.]

Girdlestone GR (1943). Acute pyogenic arthritis of the hip: an operation giving free access and effective drainage. *Lancet* **1**, 419. (Reprinted in *Clinical Orthopaedics* **170**, 3–7 (1982).) [Girdlestone's classic description, in the preantibiotic era, of septic arthritis of the hip and the operation that bears his name.]

Howard JB, Highgenboten CL, Nelson JD (1976). Residual effects of septic arthritis in infancy and childhood. *Journal of the American Medical Association* **236**, 932–5. [Outcomes review after infant septic arthritis, emphasizing need for prompt diagnosis and treatment.]

Kaandorp C. *et al*. (1997). The outcome of bacterial arthritis: a prospective community-based study. *Arthritis and Rheumatism* **40**, 884–92.

Le Dantec L *et al*. (1996). Peripheral pyogenic arthritis. A study of one hundred seventy-nine cases. *Reviews in Rheumatology* **63**, 103–10. [Large case series.]

Lowy FD (1998). *Staphylococcus aureus* infections. *New England Journal of Medicine* **339**, 520–9. [Helpful review of this troublesome pathogen.]

Seviour PW, Dieppe PA (1984). Sternoclavicular joint infection as a cause of chest pain. *British Medical Journal* **288**, 133–4. [A catch for the physician.]

Syrogiannopoulos GA, Nelson JD (1988). Duration of antimicrobial therapy for acute suppurative osteoarticular infections. *Lancet* **ii**, 37–40. [Large series that defined oral short course regimens for uncomplicated acute bone and joint infection.]

18.7.2 Reactive arthritis

J. S. H. Gaston

Introduction and historical perspective

The term 'reactive arthritis' was introduced in 1969 by Aho in Finland, where the combination of a high prevalence of HLA B27 and gastrointestinal infection by *Yersinia* afforded opportunities for studying the disease. However, the condition was first recognized in the eighteenth and nineteenth centuries as an arthritis which followed dysentery or venereal disease, and there were descriptions by Hans Reiter and other contemporaries of the disease amongst troops affected by dysentery in the trenches of the First World War. The term 'Reiter's disease' has been used extensively since that time, but is now less favoured for several reasons: Reiter was not the first to describe the disease; he erroneously attributed it to spirochaetal infection; and the triad which makes up Reiter's disease—arthritis, conjunctivitis, and urethritis/cervicitis— is not a clinically meaningful subgroup within reactive arthritis.

Definition

The term reactive arthritis is sometimes used rather loosely to cover any form of arthritis which follows infection, and would then include postviral arthritides, rheumatic fever, Lyme disease, and other forms of arthritis which do not generally have clinical features in common. This usage is not helpful and the term 'postinfectious arthritis' is to be preferred, with this all-embracing term then subdivided into different clinical syndromes, one of which is reactive arthritis (Table 1). Viewed in this way, reactive arthritis is seen as one of the seronegative spondyloarthropathies (ankylosing spondylitis, psoriatic arthritis, arthritis associated with inflammatory bowel disease), sharing clinical and immunogenetic features with those diseases. Other postinfectious arthropathies lack these common features.

In the absence of agreed and validated diagnostic or classification criteria for reactive arthritis, Table 2 presents a useful working classification of those patients who could reasonably be considered to have reactive arthritis. This takes as its starting point the classical pattern of arthritis and typical extra-articular features (Table 3) which are commonly seen after infection by five organisms—*Salmonella*, *Yersinia*, *Campylobacter*, *Shigella flexneri*, and *Chlamydia trachomatis*. The same clinical syndrome (i.e. arthritis and extra-articular signs) is also seen, but more rarely, following many infections, especially of the gastrointestinal tract, for example by *Clostridium difficile*, but genitourinary infections with *Ureaplasma* and respiratory infection with *Chlamydia pneumoniae* can probably also act as triggers of reactive arthritis. The other large group of patients are those who have asymmetric oligoarthritis without extra-articular features, but with definite laboratory evidence of preceding infection by one of the five major reactive arthritis-associated bacteria. Laboratory diagnosis of infection is given priority over symptoms as a classification criterion because infection may be clinically silent. *Chlamydia* is notorious for this, particularly in women, whilst in *Yersinia* infection, arthritis is inversely correlated with the severity of gastrointestinal symptoms. Positively identifying the triggering infection often poses practical problems. Patients in whom arthritis (with or without extra-articular signs) develops after symptomatic episodes of gastrointestinal or genitourinary infection are therefore usually regarded as having reactive arthritis, even when no triggering organism can be identified, although the diagnosis is inevitably less secure in these cases. Improvement in methods for diagnosing preceding infection should decrease the size of this group, and may well show reactive arthritis to be the commonest cause of inflammatory oligo- or monoarthritis in young adults.

Epidemiology

This has been best studied in outbreaks of food poisoning where the proportion of infected patients developing arthritis can be accurately assessed. However, in such studies the proportion going on to develop reactive arthritis varies widely (0–21 per cent). The incidence of clinically significant arthritis is generally low in community surveys of infected patients and high in patients whose infection is severe enough to require hospital referral, but careful population studies of *Campylobacter* infection have shown a high incidence (7–16 per cent) of musculoskeletal symptoms not severe enough to need rheumatological attention. As in other forms of spondyloarthropathy, the influence of HLA B27 on incidence is important: 60 to 80 per cent of patients with reactive arthritis presenting to rheumatology clinics will be B27 positive, but amongst those with mild disease the figure drops to 30 per cent, compared with the population prevalence of 7 to 10 per cent. Thus B27 is associated mainly with the severity and persistence of arthritis, rather than its incidence.

Pathogenesis

There is mounting evidence that, following infection of the gut or genitourinary tract, organisms reach the joints. They may arrive intact, when they can be detected by using the polymerase chain reaction, or as antigenic material (proteins, lipopolysaccharide) which can be demonstrated by immunofluorescence and immunoblotting techniques in synovial macrophages and polymorphs. This process can continue for months or even years, suggesting that some of the infections, for example *Yersinia*, persist as continuing sources of antigens/organisms. Elevated and persistent titres of IgA antibody to these organisms in reactive arthritis compared with uncomplicated infection also favour the idea of persistence. These findings emphasize that the distinction between septic and postinfectious arthritis has become blurred, since viable organisms can be detected in the joint in various forms of postinfectious arthritis, including Lyme disease and reactive arthritis, although they may be difficult or impossible to culture from synovial fluid or synovium.

Within the joint, cellular immune responses to the bacteria responsible for triggering the reactive arthritis are readily detected, particularly by CD4+ helper T lymphocytes, but also CD8+ T cells. Interestingly, although the association with HLA B27 is often taken to imply that CD8+ T cells are the principal effector cells in the disease, observations on reactive arthritis in HIV-positive patients show that they present with arthritis at stage I infection, when numbers of CD4+ T cells are less depressed. By contrast, arthritis can be relatively quiescent in full-blown AIDS. Both CD4+ and CD8+ lymphocytes produce proinflammatory cytokines such as interferon-g which could potentially drive joint inflammation by secondary effects on synoviocytes.

There are several hypotheses about how HLA B27 could influence the course of reactive arthritis, particularly its severity and persistence. For instance, infection may generate a B27-restricted response by CD8+ T cells to a bacterial peptide which crossreacts with a component of the joint, i.e. infection triggers autoimmunity by 'molecular mimicry'. No such autoimmune response has yet been demonstrated. Alternatively, B27 may adversely affect the efficiency with which the immune system clears the triggering organism. In this case disease does not require autoimmunity, but is primarily driven by persistent bacterial antigens. Lastly, B27 might affect the immune response to the triggering organism qualitatively, for example by allowing hyper-responsiveness to particular antigens, or biasing the immune response in favour of the production of proinflammatory cytokines.

Clinical features

Preceding illness

A history of urethritis (dysuria or discharge) and diarrhoea must be sought specifically, there being no reason for patients to automatically link these occurrences with their arthritis. The interval between infection and arthritis is variable but is not usually more than 3 weeks. However, by the time a rheumatologist is consulted, many weeks may have passed and the triggering illness be forgotten, particularly if symptoms were mild. Note that urethritis may be triggered by gastrointestinal infection: if this possibility is forgotten minimally symptomatic gastrointestinal infection and prominent urethritis may cause diagnostic confusion.

Arthritis

The clinical picture in reactive arthritis is usually an asymmetric oligoarthritis (generally fewer than six joints), predominantly affecting the lower limbs (Table 4). However, any joint can be affected, and a proportion of patients have monoarthritis only. Affected joints are often hot and markedly swollen, with septic arthritis and crystal-induced arthritis being the most likely differential diagnoses. Dactylitis, similar to that seen in psoriatic arthritis, also occurs. Many patients complain of low back or buttock pain, suggesting involvement of the sacroiliac joint. Arthritis is usually at its worst early in the course of the disease, but new sites can be affected after several months and relapses are not uncommon, even in those in whom disease eventually settles completely. The presence of enthesitis (inflammation of ligamentous and tendinous insertions) in addition to arthritis is helpful diagnostically, with plantar fasciitis and involvement of the Achilles tendon insertion the commonest sites.

Extra-articular features

In acute severe disease patients have constitutional symptoms of malaise, fatigue, and fever. More useful diagnostically are the specific extra-articular signs listed in Table 3 and illustrated in Fig. 1, Fig. 2, and Fig. 3. The fact that these extra-articular features are common to other forms of spondyloarthropathy greatly strengthens the case for including reactive arthritis in this disease family, and for delineating reactive arthritis as a distinct syndrome within the postinfectious arthritides. Extra-articular features are more common in those with severe joint involvement. Conjunctivitis is often transient and no longer present by the time the patient presents. More persistent eye inflammation or painful eyes should raise the question of an acute anterior uveitis rather than a simple conjunctivitis and prompt full ophthalmological assessment. Circinate balanitis is usually asymptomatic and needs to be looked for specifically in uncircumcized males. Oral ulceration is usually asymptomatic. Keratoderma blennorrhagica is histologically identical to psoriasis; it is most commonly seen on the soles of the feet, but can also involve the hands or trunk. Erythema nodosum is associated with *Yersinia* infection, but is otherwise uncommon in reactive arthritis. Aortitis and cardiac conduction disorders have been described but are rare.

Fig. 1 Ulceration of the tongue in reactive arthritis. (By courtesy of Dr C. J. Eastmond.)

Fig. 2 Circinate balanitis in reactive arthritis. (By courtesy of Dr C. J. Eastmond.)

Fig. 3 Keratoderma blennorrhagica in reactive arthritis. (By courtesy of Dr C. J. Eastmond.)

Differential diagnosis

The differential diagnosis of reactive arthritis is summarized in Table 5. The principal concerns in acute disease are septic arthritis, crystal arthropathies, and other forms of postinfectious arthritis such as Lyme disease, poststreptococcal arthritis, or gonococcal arthritis. In chronic disease it may be difficult to distinguish reactive arthritis from other forms of spondyloarthropathy, especially in those with inflammatory bowel disease, and many patients in whom no infectious trigger can be implicated are classified as having an 'undifferentiated' spondyloarthropathy.

Laboratory features

General

The principal aims of investigation are to exclude important differential diagnoses and to identify the triggering organism. Abnormalities in the early stages when

arthritis is most active are those of a pronounced acute inflammatory response with elevation of erythrocyte sedimentation rate and C-reactive protein, the latter often very marked (more than 100 mg/litre). Rheumatoid factor and antinuclear antibodies are absent. Positive antineutrophil antiplasmic antibodies have been described, but the antibodies are not directed against proteinase-3 or myeloperoxidase and the test is not useful diagnostically. Septic arthritis and crystal arthropathies are best excluded by aspiration of synovial fluid followed by culture and microscopy. Blood cultures should be performed and serum urate checked. A chest radiograph may reveal hilar lymphadenopathy, suggesting the diagnosis of sarcoidosis, although *Yersinia* can cause both reactive arthritis and a sarcoid-like illness. Throat swab and antibodies to streptococcal antigens may produce evidence of poststreptoccocal arthritis, which does not generally share extra-articular features with reactive arthritis.

Microbiological

Stool should be cultured for pathogens associated with reactive arthritis, although cultures are often negative after gastrointestinal symptoms have settled—despite the recent evidence that persistent infection contributes to pathogenesis. *Chlamydia* infection must be sought in sexually active patients, particularly when there is no clear history of gastroenteritis. Formal referral to a department of genitourinary medicine is often helpful. Patients are not infrequently infected with both *Chlamydia* and *Gonococcus*. This can cause confusion, but gonococcal arthritis differs from reactive arthritis with its characteristic rash and absence of the classical extra-articular features. *Chlamydia* can be cultured from urethral or cervical swabs or from urine, and *Chlamydia* antigens can be demonstrated by enzyme-linked immunosorbent assay techniques or by direct immunofluorescence tests. The latter are highly sensitive, in principle able to detect one organism per smear, but they require highly skilled technicians. Polymerase chain reaction techniques achieve similar sensitivity, and can be used on urine specimens, addressing the natural reluctance of many patients with reactive arthritis (or other conditions) to undergo routine urethral and vaginal instrumentation.

The possibility of spondyloarthropathy in the context of HIV infection also needs to be considered, although this appears rare in developed countries. By contrast, large numbers of reactive arthritis cases, often related to dysentery, have recently emerged amongst the HIV-infected population in Africa, where the disease was previously unknown.. Nevertheless, HIV testing should be considered, particularly in patients with unusually severe disease and with relevant risk factors.

Immunological

When the triggering organism cannot be demonstrated directly, infection can be inferred on the basis of immune responses. However, this evidence needs to be interpreted cautiously, since in many cases the findings simply imply immunological memory for the organism in question and do not demonstrate a clear relationship between infection and arthritis. For enteric pathogens, specific IgM antibodies may be demonstrated and these, along with IgG, form the basis of the agglutination tests which are widely used. Rising IgG titres may also be helpful. However, when patients present several months into their illness, IgM may no longer be evident and IgG titres stable. In these circumstances high and persistent IgA titres to organisms such as *Salmonella* and *Yersinia* may be helpful. The possibility of using antibody responses to particular bacterial antigens diagnostically is currently under investigation. Serological diagnosis of *Chlamydia* infection is particularly difficult because of high levels of infection with *Chlamydia pneumoniae* in the community, an organism which shares several highly conserved antigens with *Chlamydia trachomatis*.

Lastly, cellular immune responses to triggering organisms can be demonstrated, particularly in the synovial fluid. Again, these only demonstrate T-cell memory for the organism and do not demonstrate causality—patients with, for instance, rheumatoid arthritis and incidental *Salmonella* infection, will have *Salmonella*-specific T cells in their synovial fluid. Currently such tests are used in research rather than diagnostically.

Radiology

In the acute stages of disease, radiology is not diagnostically helpful, with soft tissue swelling and occasionally periarticular osteoporosis at affected joints being the only abnormalities. Radionuclide scintigraphy can be useful for demonstrating acute sacroiliitis and may show the full extent of acute synovitis and enthesitis, but is not usually required for clinical management. Radiological changes are confined to the minority of patients with persistent disease (more than 1 year's duration). The principal features are erosion of affected joints, including the sacroiliac, and new bone formation manifested as periostitis of metatarsal and metacarpal bones and 'enthesophytes', such as plantar spurs. In the spine paravertebral ossification can be seen in the lumbar region: this is asymmetric and differs from the classical changes of ankylosing spondylitis. Erosive changes are also seen at sites of enthesitis such as the calaneum.

Treatment

Evidence-based therapies for reactive arthritis are lacking, so that consensus opinion is the current guide. In the acute phase affected joints should be rested until they improve substantially. This often requires emphasizing to young, active patients involved in sports, and alternative forms of exercise should be considered. Synovial effusions should be aspirated and when septic arthritis has been excluded will respond well to injection with long-acting corticosteroids. If *Chlamydia* infection is established or thought likely, patients require conventional treatment with short-term antibiotics, but there is no evidence that this has any effect on the progress of reactive arthritis. Enteric infections do not require antibiotics in their own right. Uveitis requires formal ophthalmological assessment and treatment with local steroids. Physiotherapy and advice on exercise is helpful, with quadriceps function needing particular attention in view of the frequent involvement of the knees.

There are two major unresolved treatment issues in reactive arthritis. Firstly, the place of disease-modifying drugs. Sulphasalazine and methotrexate are useful in spondyloarthropathies generally, and on this basis have been used in reactive arthritis, but without controlled trials confined to this condition alone. The frequency with which the disorder is self-limiting makes controlled trials of second-line agents difficult. The second issue is whether long-term antibiotics confer any benefit. In a controlled trial, a subset of patients with evidence of *Chlamydia* infection benefited from prolonged tetracycline, but subsequent trials using ciprofloxacin, mainly in *Yersinia*-associated reactive arthritis, have been negative. Nevertheless, since persistent infection is implicated in our current understanding of pathogenesis, further work on antibiotics is being undertaken, particularly of agents which might be able to treat quiescent or slow-growing bacteria which are refractory to agents that target bacterial cell division.

Psychological and quality of life issues

Reactive arthritis commonly affects young, fit adults who have not previously experienced any form of prolonged illness or disability. The danger is that the rheumatologist, all too used to the gloomy prognosis of rheumatoid arthritis, may treat reactive arthritis, where there is a high likelihood of complete resolution of disease, too lightly. Patients need to be given a realistic prognosis, i.e. that symptoms are likely to persist at some level for 6 to 12 months, although in the latter stages these are usually very mild compared with those experienced in the first 4 to 8 weeks. Exacerbations during this time are not uncommon and do not imply that the disease will not eventually resolve. The chances of the patient developing chronic arthritis are less than 10 per cent. Patients benefit from continuing psychological and clinical support throughout the course of their illness, with rapid access to joint aspiration and intra-articular steroid injection when there is recurrent joint swelling.

Current areas of uncertainty

Current uncertainties concern classification criteria and management strategies. Both may be resolved by developing more secure diagnostic techniques for identifying the triggering infection. Improved treatment is likely to come from either additional evidence about the importance of persistent infection and how to eliminate it, or from discovery of the immune responses responsible for maintaining joint inflammation, whether these are directed against a bacterial antigen or an autoantigen. If a target antigen can be identified, specific immunomodulation strategies will then be relevant.

Reactive arthritis differs from other forms of human inflammatory arthritis in having a clearly defined onset and being triggered by known infectious agents. Genetic influences that result in a minority of infected individuals developing arthritis are being investigated. These include HLA B27, but it is likely that other genes are also involved. Genome screening now being applied to ankylosing spondylitis may throw up likely candidates.

Further reading

Reviews of pathogenesis

Gaston JSH (1995). Symposium: reactive arthritis. *Rheumatology. in Europe* **24**, 5–22.

Sieper J, Braun J (1995). Pathogenesis of spondylarthropathies: persistent bacterial antigen, autoimmunity, or both? *Arthritis and Rheumatism* **38**, 1547–54.

Incidence following salmonella infection

Mattila L *et al.* (1994). Reactive arthritis following an outbreak of salmonella infection in Finland. *British Journal of Rheumatology* **33**, 1136–41.

Inman RD *et al.* (1988). Postdysenteric reactive arthritis: a clinical and immunologic study following an outbreak of salmonellosis. *Arthritis and Rheumatism* **31**, 1377–83.

Evidence that bacteria or bacterial antigens reach the joint in reactive arthritis

Gaston JSH, Cox C, Granfors K (1999). Clinical and experimental evidence for persistent *Yersinia* infection in reactive arthritis. *Arthritis and Rheumatism* **42**, 2239–42.

Gerard HC *et al.* (1998). Synovial *Chlamydia trachomatis* in patients with reactive arthritis/Reiter's syndrome are viable but show aberrant gene expression. *Journal of Rheumatology* **25**, 734–42.

Granfors K *et al*. (1989). Yersinia antigens in synovial fluid cells from patients with reactive arthritis. *New England Journal of Medicine* **320**, 216–21.

Granfors K *et al*. (1990). Salmonella lipopolysaccharide in synovial cells from patients with reactive arthritis. *The Lancet* **335**, 685–8.

Immune responses in reactive arthritis

Gaston JSH *et al*. (1989). Synovial T lymphocyte recognition of organisms that trigger reactive arthritis. *Clinical and Experimental Immunology* **76**, 348–53.

Granfors K, Toivanen A (1986). IgA-anti-yersinia antibodies in yersinia-triggered reactive arthritis. *Annals of the Rheumatic Diseases* **45**, 561–5.

Hermann E *et al.* (1993). HLA-B27-restricted CD8 T-cells derived from synovial fluids of patients with reactive arthritis and ankylosing spondylitis. *The Lancet* **342**, 646–50.

Reactive arthritis and other spondyloarthropathy in HIV infection

Njobvu P *et al.* (1998). Spondyloarthropathy and human immunodeficiency virus infection in Zambia. *Journal of Rheumatology* **25**, 1553–9.

Treatment in reactive arthritis

Dougados M *et al*. (1995). Sulfasalazine in the treatment of spondylarthropathy: a randomized, multicenter, double-blind, placebo-controlled study. *Arthritis and Rheumatism* **38**, 618–27.

Lauhio A *et al.* (1991). Double-blind, placebo-controlled study of three-month treatment with lymecycline in reactive arthritis with special reference to chlamydia arthritis. *Arthritis and Rheumatism* **34**, 6–14.

Sieper J *et al*. (1999). No benefit of long-term ciprofloxacin treatment in patients with reactive arthritis and undifferentiated oligoarthritis—a three-month, multicenter, double-blind, randomized, placebo-controlled study. *Arthritis and Rheumatism* **42**, 1386–96.

18.8 Osteoarthritis

Paul H. Brion and Kenneth C. Kalunian

Introduction

Osteoarthritis is the commonest form of arthritis, detectable radiographically in 80 per cent of patients over the age of 55. Symptomatic osteoarthritis of the knee (pain with radiographic abnormalities) was noted in 6.1 per cent of adults aged 30 and over in the Framingham Study, and the frequency is comparable in the United Kingdom. Approximately 20.7 million people in the United States have physician-diagnosed osteoarthritis. Whilst some may be asymptomatic, many have significant pain and disability, with one study finding that osteoarthritis accounts for 12.3 per cent of all those with limitation of activity. The prevalence of osteoarthritis is likely to increase, paralleling the increase in the absolute and relative number of people who are over 65 years of age. The social impact of this disease is enormous, accounting for more dependency in walking and stair climbing than any other disease. The estimated annual cost associated with osteoarthritis in the United States is $15.5 billion in 1994 dollars, which approaches 1 per cent of the gross national product, with more than 50 per cent of the costs due to work loss.

Definition

Osteoarthritis was described by Solomon as

> A chronic disorder characterized by softening and disintegration of articular cartilage, with reactive phenomena such as vascular congestion and osteoblastic activity in the subarticular bone, new growth of cartilage and bone (osteophytes) at the joint margins, and capsular fibrosis. Osteoarthritis is not accompanied by any systemic illness, and although there are sometimes signs of inflammation, it is not primarily an inflammatory disorder.

Alternative definitions exist, including those based on symptoms, physical findings, and radiographic and arthroscopic findings. The presence of joint symptoms plus evidence of structural change generally defines clinical osteoarthritis, whereas many studies use radiographic assessment alone as the primary means of identifying the condition. Most clinical investigators use the Kellgren and Lawrence scale for grading osteoarthritis of the knee, which defines osteoarthritis on the basis of osteophytes, the presence of which relates well with the presence of knee symptoms. The American College of Rheumatology has developed classification criteria for the presence of osteoarthritis based on the joint involved (Table 1, Table 2, and Table 3): these are based on clinical criteria alone, clinical and laboratory criteria, and clinical plus radiographic criteria. Initially, only the clinical and radiographic criteria were validated; more recently, Klashman and colleagues have validated the other methods. However, the clinical and radiographic criteria have the best specificity (86 per cent) compared with the clinical and laboratory criteria (75 per cent) and clinical criteria alone (69 per cent). Kawasaki and colleagues have shown that the non-radiographic criteria have adequate interobserver reproducibility for outpatients complaining of knee pain with a wide spectrum of rheumatic diseases; however, sensitivity and specificity are lower than those reported by Klashman and colleagues.

Risk factors and epidemiology

A multitude of risk factors exist for the development of osteoarthritis of the knee. These include being female, increasing age, obesity, family history, increased bone density, trauma, and certain occupational exposures (Table 4).

Age is considered the strongest associated risk factor for the development of osteoarthritis in many studies. The National Health and Nutrition Examination Survey found a prevalence of osteoarthritis of only 0.1 per cent in people aged 25 to 34 years, compared with over 80 per cent in those aged 55 to 64 years. This increased incidence occurs in osteoarthritis of the hands, back, hip, and knees.

Gender differences in osteoarthritis are complicated. There is an overall higher prevalence in women, in whom the disease more often involves multiple joints. However, before the age of 50 years there is a higher prevalence and incidence in men, whereas after 50 years the reverse applies, with increasing female predominance as age increases. There is a plateau or decline in both genders by the age of 80 years. The gender- and age-related differences in prevalence parallel the effect of postmenopausal oestrogen deficiency in increasing the risk of osteoarthritis. Other probable factors that help explain the increase in the incidence and prevalence of osteoarthritis with age include a decreased responsiveness of chondrocytes to growth factors that stimulate repair, an increase in the laxity of ligamentous structures, and a decrease in proprioceptive responses.

Although racial differences in osteoarthritis of the hip are conflicting, the higher relative weight of African-American women may predispose them to higher rates of osteoarthritis of the knee. There are few data available for other racial differences in osteoarthritis of the knee among the population of the United States. Several studies have confirmed that inheritance is a risk factor for osteoarthritis: individuals are at higher risk of developing the condition if their parents had it, especially if the parental disease was polyarticular or had its onset in middle age or earlier. The role of inheritance may be more important among women than men. Numerous extended families with high rates of early onset severe osteoarthritis have been characterized in which the condition has been linked to an autosomal dominant mutation in type 11 procollagen. Although the majority of cases of osteoarthritis of the hand are inherited, the percentage for osteoarthritis of the knee is smaller, perhaps because osteoarthritis of the knee often develops more as a result of repeated mechanical insults.

The incidence of osteoarthritis is lower in the setting of osteoporosis: bone density in osteoarthritis patients is greater than in age-matched controls, even at sites distant from the affected joints. Most studies linking osteoarthritis with high bone density are cross-sectional. Although osteoarthritis and high bone density are both linked to obesity, the association of osteoarthritis with high bone density is independent of body mass index. It has been suggested that osteophyte formation rather than cartilage loss is linked to high bone density, which suggests the presence of a circulating bone growth factor in those with osteophytes; possibilities include insulin-like growth factor type 1, platelet-derived growth factor, fibroblast growth factor, transforming growth factor-b, and colony-stimulating factor.

Oestrogen deficiency has been implicated as a risk factor for the development of osteoarthritis as evidenced by the high incidence of osteoarthritis after the menopause. Several studies suggest that oestrogen replacement therapy reduces the risk of osteoarthritis of the hip and knee. Both the Study of Osteoporotic Fractures and the Framingham Study have reported a strong inverse relationship between oestrogen replacement therapy and osteoarthritis among those undergoing long-term oestrogen replacement therapy.

Reactive oxygen species have been implicated in the development of osteoarthritis, and antioxidants may prevent or delay the onset of osteoarthritis. In the Framingham Study, those in the lowest tertile of vitamin C intake had a threefold greater risk of progression of osteoarthritis of the knee, joint space loss, and onset of knee pain compared with subjects with a higher intake. However, the effects of b-carotene and vitamin E against disease progression were inconsistent. No effect of serum 25-hydroxy-vitamin D was seen on incident osteoarthritis; however, among subjects with radiographic osteoarthritis at baseline, those who were in the lowest tertile of serum 25-hydroxyvitamin D had a higher rate of radiographic progression compared with those in the highest tertile.

Local biomechanical factors such as trauma or repetitive joint use are risk factors for osteoarthritis. In animal models, a change in biomechanics that occurs after injury leads to increased shear stress on local areas of cartilage, possibly causing osteoarthritis. In humans, traumatic injury to joints is a common cause of osteoarthritis, and Kellgren and Lawrence found that a history of previous trauma could be elicited in approximately 40 per cent of men and approximately 20 per cent of women aged 55 to 64 years with osteoarthritis of the knee. In the Framingham Study, men with a history of major trauma to the knee had a five-fold increased risk

of osteoarthritis of the knee, whilst women with a similar history had a greater than three-fold increased risk. Trauma that causes damage to a cruciate ligament and/or a meniscus has been associated with subsequent development of osteoarthritis of the knee, perhaps through concurrent damage to articular cartilage. With regard to repetitive use, occupations that require kneeling and squatting are associated with a higher prevalence of osteoarthritis of the knee, but heavy physical work is less consistently associated. The level of physical activity increases the risk of developing osteoarthritis. In the Framingham Study, physical activity (generally consisting of walking and gardening in this population) was found to correlate directly with the risk of developing radiographic osteoarthritis of the knee in elderly subjects followed for 8 years. Those with high levels of these activities had a threefold increase in the risk of osteoarthritis compared with sedentary controls. Elite athletes have higher rates of incidence of osteoarthritis of weight-bearing joints compared with controls, probably because athletic activities often involve both increased risk of injury and repetitive use.

Several longitudinal studies suggest that increased weight is a risk factor for the development of osteoarthritis of the knee, and that overweight patients with established osteoarthritis of the knee are at greater risk of developing progressive disease compared with those who are not overweight. The associations between obesity and osteoarthritis of the knee are significantly greater for women than men and are not affected by adjustments for concurrent diseases. Data from the Chingford Study showed that patients in the highest weight tertile had an odds ratio of 6.17 for radiographic osteoarthritis of the knee compared with the lowest weight tertile. For every two-unit increase in body mass index (approximately 5 kg), the odds ratio for radiographic osteoarthritis of the knee increased by 1.36. A follow-up study of incident osteoarthritis of the knee in women with unilateral disease found the tertile with the highest body mass index had a relative risk of 4.69 for developing osteoarthritis in the contralateral knee compared with patients in the lowest body mass index tertile. Similar findings exist for osteoarthritis of the hand and hip, but the association is less robust. The importance of obesity cannot be understated as this may be a modifiable risk factor. A possible mechanism for the effect of obesity on osteoarthritis of the knee is increased force across the weight-bearing joint, which induces cartilage breakdown by altered walking mechanics. However, obesity may also have effects through metabolic intermediaries.

Quadriceps weakness has been associated with radiographic osteoarthritis of the knee. Muscular strength may be required to stabilize the knee, distribute force, or lessen the effect of an impact load, and maintenance of muscular strength may be important in decreasing the incidence of osteoarthritis of the knee and its progression and disability due to established disease. Proprioceptive sensation, which declines with age, is impaired in elderly patients with osteoarthritis of the knee, suggesting that poor proprioception may contribute to functional impairment in these patients.

Pathogenesis and pathological features

The pathogenesis of osteoarthritis remains controversial. Once thought of as a normal consequence of aging, the complex nature of this disease is only now being understood. Current theories suggest that osteoarthritis results from an imbalance in catabolic and anabolic processes that lead to progressive cartilage damage and destruction. Increased catabolism may be the result of acute injuries such as an acute meniscal tear or of chronic microtraumatic events. Initially, anabolic processes such as proteoglycan synthesis maintain balance with catabolic processes and damage to cartilage is repaired. However, with time and age, anabolic processes decline and progressive cartilage damage ensues.

Histological changes in osteoarthritis are complex. Early stages are characterized by increased water content and cartilage swelling. This swollen cartilage is believed to be more susceptible to injury and may lead to fragmentation of the articular surface. Fragmented cartilage is less able to withstand biomechanical insults, resulting in further deterioration. Chondrocytes become activated and proinflammatory cytokines such as interleukin 1 and tumour necrosis factor-a are synthesized. These cytokines increase the synthesis of degradative proteases such as collagenase, gelatinase, and stomelsin. As cartilage destruction progresses, proteoglycan content becomes reduced. Cartilage becomes thinned, fragmented, and proteoglycan depleted.

Reparative processes may initially lead to joint stabilization, but ultimately contribute to progression of the disease. Fibrocartilage may be synthesized in response to loss of the more durable hyaline cartilage. Fibrocartilage may be denuded bone, improving joint mechanics and protecting the subchondral bone. However, fibrocartilage is less able to withstand mechanical loading. The subchondral bone is exposed to increased force relative to that of hyaline cartilage. The synthesis of fibrocartilage, while a temporary improvement, is ultimately less efficient than hyaline cartilage. Subchondral bony changes such as sclerosis and osteophyte formation develop.

The gross pathological findings of osteoarthritis include cartilage loss and reactive bone formation. Cartilage loss occurs primarily in areas of joint loading and may be related to repetitive mechanical insults. Cartilage loss may be best visualized arthroscopially when findings include cartilage softening, fibrillation, and thinning. Areas of complete cartilage loss may be seen. More commonly, the clinician will recognize these findings as radiographic joint space narrowing. Similarly, bony changes may be seen on pathological specimens or arthroscopically. Arthroscopic findings include osteophyte projections and subchondral bone visualized through denuded cartilage. The classic radiographic bony changes include osteophyte formation and subchondral bony sclerosis and cysts.

Clinical features

Precise definition of clinical osteoarthritis has remained elusive since radiographic findings and symptoms may diverge. In addition, osteoarthritis may be categorized by the joint area involved or as being idiopathic or secondary to other disorders. An essential element to diagnosis of osteoarthritis is the correct attribution of symptoms to the affected joint. Initial evaluation of soft tissue abnormalities such as bursitis, tendonitis, and ligamentous strain should be performed. In addition, consideration of neurological or underlying bone abnormalities should be entertained when appropriate.

A thorough history and physical examination should be performed to evaluate for secondary forms of osteoarthritis. These include developmental, mechanical, or biochemical abnormalities known to increase the risk for osteoarthritis. These forms tend to present earlier in life (for example congenital hip abnormality), in atypical joints (for example calcium hydroxyapatite) or as more inflammatory in nature (for example calcium pyrophosphate deposition disease).

Idiopathic osteoarthritis may occur localized to one body area or as a more generalized disease. Common areas of involvement include the hands, hips, knees, and spine. Less commonly, osteoarthritis involves the shoulders, wrists, ankles, feet, and jaw.

Pain is the predominant symptom of osteoarthritis, usually mild to moderate in nature and increasing with joint use and at the end of the day. Pain is generally improved with rest and moderation of activity. Severe disease may cause pain at rest or at night. The source of pain may be the underlying bone, the joint capsule, or surrounding structures. Cartilage is avascular and without nerves and not itself a source of pain.

Stiffness may occur but is generally limited to less than 30 min in duration (gelling phenomenon). It is typical in the morning or after any prolonged rest (theatre sign). Effusions may occur, but warmth and soft tissue swelling is rare and suggests another diagnosis.

Physical examination reveals tenderness to palpation, bony thickening (osteophyte formation), small effusions, and crepitus. Specific joint findings also occur. Typical in the hand are bony enlargement of the proximal interphalangeal joints (Bouchard's nodes) and the distal interphalangeal joints (Heberden's nodes). The first carpometacarpal joint may be involved causing a 'squared appearance' of the lateral aspect of the hand (in anatomical position) (Fig. 1). Involvement of the foot yields bunions, and of the knee pronounced valgus and varus deformities, Baker's cyst or locking suggesting meniscal damage. Early hip findings include limited internal and external rotation. Back findings include pain: true osteoarthritis occurs at the apophyseal joints; degenerative disc disease and diffuse idiopathic skeletal hyperostosis are distinct entities.

Fig. 1 Osteoarthritis of the hand. Note squaring of first carpometacarpal (CMC) joint and evidence of a Heberden's node on the third distal interphalangeal joint.

In an effort to standardize the diagnosis of osteoarthritis, the American College of Rheumatology formed a subcommittee to define osteoarthritis of the knee, hip, and hand. Clinical, laboratory, and radiographic findings were evaluated by an expert panel and statistical analysis, to yield classification criteria with acceptable sensitivity and specificity values. These instruments should be used with caution in individual patients but provide a framework for analysis (see Table 1, Table 2, and Table 3).

Common clinical mimics of osteoarthritis include rheumatoid arthritis, calcium pyrophosphate deposition disease, and infectious monoarticular arthritis.

Hand osteoarthritis may be confused with rheumatoid arthritis as both cause pain and visible swelling. Less commonly hip or knee arthritis may present as diagnostic challenges. Hand osteoarthritis typically involves the proximal interphalangeal and the distal interphalangeal joints; the 'swelling' is not true swelling but hard, bony thickening due to osteophyte formation; and stiffness is limited. By contrast, rheumatoid arthritis typically involves the proximal interphalangeal, metacarpophalangeal, and carpal joints, sparing the distal interphalangeals. True swelling occurs and is soft with palpation. Multiple joints are involved, symptoms of systemic inflammation occur, and rheumatoid factor is usually positive. In rheumatoid arthritis radiographs demonstrate symmetric joint space narrowing, bony erosions, minimal sclerosis, and minimal osteophyte formation.

Calcium pyrophosphate deposition disease is difficult to differentiate from idiopathic osteoarthritis because the two may coexist. Typical distributions for this disease include the knees, wrist, shoulder, and metacarpophalangeals. Patients may have more prolonged stiffness and pain and swelling may occur. Radiographs with evidence of chondrocalcinosis strongly suggest calcium pyrophosphate deposition disease, but the presence of crystals on arthrocentesis is the gold standard for diagnosis.

Infectious monoarthritis can occasionally mimic osteoarthritis. The distinction is more difficult with subacute infections, such as fungal or mycobacterial. If there is clinical suspicion, radiographs and arthrocentesis should be performed.

Investigation

Laboratory tests, if performed, reveal normal sedimentation rates and non-inflammatory synovial fluid. Radiographic findings include asymmetry, joint space narrowing, subchondral sclerosis, subchondral cysts, and the hallmark osteophyte (Fig. 2, Fig. 3, Fig. 4, and Fig. 5).

Fig. 2 Osteoarthritis of the hand. Note changes in the distal interphalangeal joints and proximal interphalangeal joints as well as the base of the thumb (carpometacarpal). These changes are typical for osteoarthritis of the hand. Note the loss of joint space, bony sclerosis, and the presence of osteophytes. The bony changes seen in the distal interphalangeal and proximal interphalangeal joints would manifest as Bouchard's and Heberden's nodes on clinical examination.

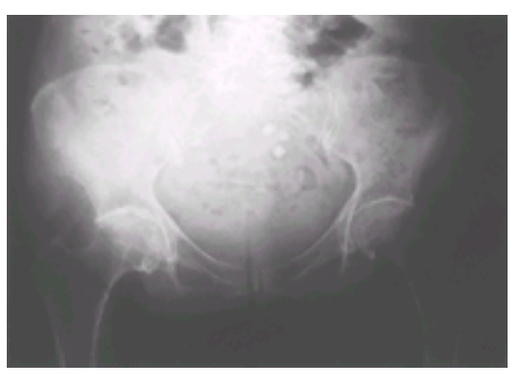

Fig. 3 Hip radiograph demonstrating osteoarthritis. Note joint space narrowing and sclerosis.

Fig. 4 Bilateral knee osteoarthritis. Note the asymmetric joint space narrowing, bony sclerosis, and the presence of osteophytes.

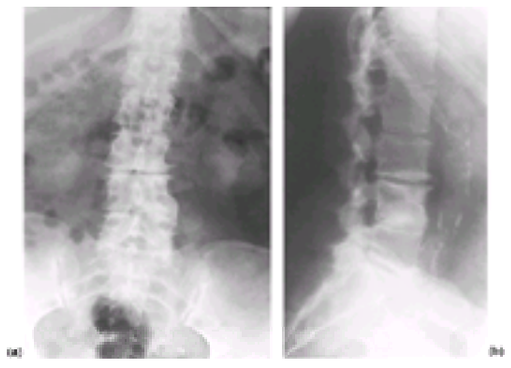

Fig. 5 (a) and (b) Lumbar spine arthritis. Note the changes of degenerative disc disease as narrowing and large osteophytes. Although often called osteoarthritis these changes are not true osteoarthritis. True osteoarthritis occurs at the facet joints. Sclerosis is seen in the inferior facet joints.

Treatment

Traditional treatments

Treatment modalities for all forms of osteoarthritis, listed in Table 5, remain limited. Traditional therapies include analgesics, non-steroidal anti-inflammatory drugs (**NSAIDs**), intra-articular corticosteroid injections, intra-articular hyaluronic acid injections, topical agents, tidal lavage, arthroscopic irrigation, and total joint replacement. With the exception of joint replacement, none of these therapies address the underlying problem of cartilage damage. Newer therapies include weight loss and exercise, both of which have been difficult to maintain over long periods of time. Emerging therapies such as tetracycline, cytokine modulators, and inhibitors of metalloproteinases may potentially alter the progression of osteoarthritis. Nutritional supplements such as glucosamine, chondroitin sulphate, soybean, and avocado products have been reported to provide better long-term analgesia than NSAIDs, and some of these agents may alter the progression of osteoarthritis and repair cartilage damage. However, the efficacy of these emerging therapies and nutritional supplements has not been studied adequately in controlled trials.

Weight loss

Weight loss, while effective, is difficult to achieve and maintain. Evidence suggests that a weight loss of 4.5 kg (10 lb) over 10 years may decrease the risk of developing contralateral knee osteoarthritis by 50 per cent. Studies demonstrating improvement in disease outcome are more controversial. However, given potential benefits in osteoarthritis as well as the additional health benefits of a normal body mass index, obese patients should be encouraged to lose weight.

Exercise and psychosocial support

Physical therapy and exercise are advocated in osteoarthritis for a variety of reasons. Improvements in flexibility and muscle strengthening may decrease joint loading, preventing further damage. They have been demonstrated to improve functional outcome and pain scores in clinical trials. In addition, they provide a sense of self-determination, an adjunct for weight loss, improve depressive symptoms, and decrease patient disability. Obstacles include expense and the lack of motivation to continue exercising after a programme has been completed.

The role of psychosocial support may be significant. Telephone calls providing contact and education have been demonstrated to improve pain and functional status. Education and support improve feelings of frustration, minimize dependency, and improve coping mechanisms.

Simple analgesics

Analgesics such as acetaminophen and paracetamol provide analgesic relief comparable with that of NSAIDs. The lower relative risk of complications has favoured their use over that of NSAIDs, especially in older populations. Typical daily doses of acetaminophen are 4 g (3 g in elderly patients). These agents may be associated with liver problems and interactions with other drugs such as warfarin. Narcotic analgesics are generally avoided because of potential complications including constipation, sedation, addiction, and impairment of balance.

Non-steroidal anti-inflammatory agents

NSAIDs remain a cornerstone of osteoarthritis treatment. Availability, dosing schedule, cost and individual side-effect profile influence the choice of a particular agent. The associations of NSAIDs with peptic ulcer disease and renal insufficiency are well established. Less common side-effects include rash, hepatic dysfunction, platelet dysfunction, and central nervous system effects. NSAIDs with specific cyclo-oxygenase-2 (Cox-2) inhibiting activity have become popular because of their decreased dosing frequency, decreased platelet effects, and—most importantly—decreased incidence of gastrointestinal ulceration compared with traditional NSAIDs.

Corticosteroids

Intra-articular corticosteroids may be effective in decreasing joint pain associated with osteoarthritis. Dosage varies depending on patient body size, comorbidity, and the joint involved. They have multiple side-effects, including risk of infection, bleeding, and (possibly) cartilage damage. To minimize the risk of complications, injections should be limited to three to four per year in any given joint.

Hyaluronic acid

The use of hyaluronic acid preparations (Hyalgan™, Synvisc™) has become popular in recent years. These agents are reported to increase viscosity by replacing depleted hyaluronic acid, which occurs in osteoarthritis. Multiple studies demonstrate efficacy similar to NSAIDs, and the risk profile for side-effects is better than that of NSAIDs.

Joint lavage

Irrigation of osteoarthritic joints has been proposed as a method of relieving joint pain by removing debris or inflammatory cytokines, but remains controversial. Livesley compared arthroscopic irrigation with physical therapy and found that the arthroscopic group experienced significant improvement in pain that was sustained over 12 months. Ike and colleagues compared medical management plus joint lavage without arthroscopy with medical management alone in a multicentre, randomized prospective study: significant improvements in pain and stiffness occurred in the group receiving irrigation. Ravaud and colleagues evaluated the efficacy of joint lavage and intra-articular steroid injection in osteoarthritis of the knee. Patients who underwent joint lavage had improved significantly at 6 months; those only given corticosteroids had early improvement but no long-term benefit. Kalunian and colleagues studied the effectiveness of visually guided arthroscopic irrigation in early osteoarthritis of the knee unresponsive to conservative management. Patients received 3 litre or minimal (< 250 ml) arthroscopic irrigation, the former having an effect on pain as measured on two rating scales.

Surgery

Surgical intervention is generally reserved for patients who have failed conservative management including analgesics, physical therapy, and intra-articular injection. Prescribed treatments include synovectomy, repair of meniscal tears, realignment osteotomy, and total joint replacement. Total joint replacement removes the affected structure and is the only known 'cure' for osteoarthritis to date, providing marked pain relief and functional improvement.

Aids and appliances

A joint that is unstable and painful can be made more stable and less painful by appropriate aids. Wheelchairs and other appliances may make it possible for a patient to maintain their independence. Walking sticks can be very effective, and for a painful hip or knee should be held in the contralateral hand to transfer weight from the affected joint. If the main problem is instablility, the stick should be held in the hand that inspires most confidence. Splinting to correct instability, correction of valgus or varus deformity at the knee or ankle, use of a rocker sole to ease hallux rigidus pain, or a heel raise if the legs are of unequal length, can all allow significant reduction of symptoms, as can the simple recommendation of shoes with good shock-absorbing soles. These simple and apparently mundane issues should not be ignored by the physician.

Other therapies

Glucosamine and chondroitin sulphate

Many patients with osteoarthritis feel that glucosamine salts and chondroitin sulphate improve symptoms and there are abundant data to support these claims, but studies generally involve small numbers of subjects. There are, however, few data to suggest that these supplements repair cartilage damage.

Glucosamine, an aminomonosaccharide, is present in almost all human tissues, but particularly in articular cartilage where it is an intermediate substrate in the synthesis of glycosaminoglycan and proteoglycans. Exogenous glucosamine salts significantly enhance chondrocyte synthesis of glycosaminoglycans, collagen, and DNA. Both glucosamine hydrochloride and glucosamine sulphate are rapidly absorbed after oral administration and are not toxic, even at high oral doses.

Several double-blind studies have compared glucosamine to placebo for periods ranging from 5 to 38 weeks in patients with osteoarthritis of the knee, showing that glucosamine produces significant improvements in pain. Vaz found that glucosamine was as effective as ibuprofen in osteoarthritis of the knee: those receiving ibuprofen had greater initial improvement in pain than those randomized to receive glucosamine, but by 4 and 8 weeks the situation had reversed. Qie and colleagues compared glucosamine with ibuprofen in 178 subjects with osteoarthritis of the knee: both significantly reduced knee symptoms, but glucosamine was better tolerated. In a randomized, controlled, double-blind trial involving 329 patients with medial femerotibial osteoarthritis, Rovati and colleagues found that 3 months of glucosamine was more effective than standard NSAID therapy and placebo, and that the safety of glucosamine did not differ from placebo but was significantly better than NSAIDs.

Reginster and colleagues recently presented data regarding the chondroprotective effects of glucosamine sulphate. This randomized, double-blind, placebo-controlled study included 212 patients with osteoarthritis of the knee, who received either oral glucosamine sulphate 1500 mg/day or placebo. Weight-bearing anteroposterior radiographs of each knee were taken at enrolment, year 1, and year 3 and analysed for joint space narrowing: this progressed by 0.24 mm in the placebo-treated group, while the glucosamine-treated group had preservation of the joint space (–0.12 mm progression). The authors concluded that glucosamine sulphate may be a disease-modifying agent. Criticisms of the study include the use of traditional radiographs rather than fluoroscopically guided semiflexed views, which have better reliability in progression studies.

Chondroitin sulphate is a long-chain polymer of a repeating disaccharide. It is the predominant glycosaminoglycan found in articular cartilage and differs from glucosamine in that it stimulates glycosaminoglycan and proteoglycan synthesis by both extracellular and intracellular mechanisms, whereas glucosamine utilizes only intracellular mechanisms. By virtue of its long chains, chondroitin sulphate competitively inhibits enzymes that degrade proteoglycans, and this may be its mechanism of action, with increased availability of substrates for formation of articular matrix another possibility. It is 70 per cent absorbed after oral ingestion, with affinity for synovial fluid and articular cartilage.

In a randomized controlled trial comparing 3 months of treatment with chondroitin sulphate with placebo, the chondroitin sulphate group demonstrated significant reductions in clinical symptoms. Chondroitin sulphate has also been compared with NSAIDs in patients with osteoarthritis of the knee. In one study, patients treated with chondroitin sulphate had statistically significant decreases in pain by 3 months of therapy, with an overall 72 per cent decrease in concomitant usage of NSAIDs. In another study, patients treated with an NSAID showed prompt reduction in clinical symptoms that reappeared after discontinuation of the NSAID, whereas those given chondroitin sulphate had a therapeutic response that appeared later but lasted for up to 3 months after treatment was discontinued.

The combined use of glucosamine and chondroitin sulphate in the treatment of osteoarthritis has become popular: they have been reported to work synergistically in forming glycosaminoglycans, inhibiting degradative enzymes, and stimulating cartilage metabolism and matrix production. In a randomized double-blind placebo-controlled trial of 93 patients with osteoarthritis of the knee, subjects were randomized to receive glucosamine hydrochloride 1000 mg and chondroitin sulphate 800 mg twice daily orally or placebo. There was significantly greater improvement in the glucosamine and chondroitin sulphate treatment group compared with controls, with a significant drop in requirement for pain medication. A National Institutes of Health sponsored multicentre randomized study is under way comparing glucosamine, chondroitin sulphate, the combination of glucosamine and chondroitin sulphate, and placebo in patients with osteoarthritis of the knee: this will involve more than 1000 patients treated for 4 months.

Tetracyclines, interleukin 1 antagonists, and collagenase inhibitors

Tetracyclines have been demonstrated to inactivate matrix metalloproteinases such as collagenase, stromolysin, and gelatinase. Dog models using doxycycline reduce the incidence of osteoarthritis. A National Institutes of Health sponsored multicentre randomized, placebo-controlled trial of doxycycline is under way.

Other agents demonstrating efficacy in osteoarthritis include use of diacerein and avocado/soybean extracts. Diacerin is an oral agent with analgesic properties, hypothesized to have an effect in osteoarthritis by inhibiting synthesis and activity of interleukin 1 and demonstrating cartilage preservation in an animal model. Human studies have demonstrated improvements in pain and function in hip osteoarthritis, as well as an NSAID-sparing effect. Similarly, avocado and soybean unsaponifiables are believed to exert their effects through interleukin 1. Clinical studies have demonstrated an NSAID-sparing effect and improvement in functional index and pain.

Ro 32–335 (Trocade) is an orally active collagenase inhibitor that has demonstrated chondroprotection by radiographic criteria in a mouse osteoarthritis model. Bay 12–9566 is a stromelysin-1 (MMP-3) inhibitor which demonstrated efficacy both dog and guinea pig meniscetomy models. Further studies of these compounds are needed. Future strategies for chondroprotection include manipulation of tumour necrosis factor, nitrous oxide, and insulin-like growth factor.

Further reading

Altman R *et al*. (1986). Development of criteria in the classification and reporting of osteoarthritis: classification of the knee. *Arthritis and Rheumatism* **29**, 1039–49.

Anderson J, Felson DT (1988). Factors associated with osteoarthritis of the knee in the First National Health and Nutrition Examination Survey (HANES 1). *American Journal of Epidemiology* **128**, 179–89.

Davis MA *et al.* (1988). Sex differences in osteoarthritis of the knee: the role of obesity. *American Journal of Epidemiology* **127**, 1019–30.

Drovanti A, Bignamini AA, Rovati AL (1980). Therapeutic activity of oral glucosamine sulfate in osteoarthritis: a placebo-controlled double-blind investigation. *Clinical Therapy* **3**, 260–72.

Ettinger WH Jr. *et al*. (1997). A randomized trial comparing aerobic exercise and resistance exercise with a health education program in older adults with knee osteoarthritis: the Fitness Arthritis and Seniors Trial (FAST). *Journal of the American Medical Association* **277**, 25–31.

Felson DT *et al*. (1991). Occupational physical demands, knee bending and knee osteoarthritis: results from the Framingham study. *Journal of Rheumatology* **18**, 1587–92.

Felson DT, Zhang Y (1998). An update on the epidemiology of knee and hip osteoarthritis with a view to prevention. *Arthritis and Rheumatism* **41**, 1343–55.

Felson DT *et al*. (1997). Risk factors for incident radiographic knee osteoarthritis in the elderly. *Arthritis and Rheumatism* **40**, 728–33.

Hannan MT *et al*. (1991). Occupational physical demands, knee bending and knee osteoarthritis: results from the Framingham Study. *Journal of Rheumatology* **18**, 1587–92.

Hart DJ, Spector TD (1993). The relationship of obesity, fat distribution and osteoarthritis in women in the general population. The Chingford Study. *Journal of Rheumatology* **20**, 331–5.

Kellgren JH, Lawrence JS (1957). Radiographic assessment of osteoarthritis. *Annals of Rheumatic Disease* **16**, 494–502.

Kelsey JL, Hochberg MC (1988). Epidemiology of chronic musculoskeletal disorders. *Annual Review of Public Health* **9**, 379–401.

Klashman D *et al.* (1996). Validation of nonradiographic ACR knee osteoarthritis classification criteria using arthroscopy damage index as the comparison standard (abstract). *Arthritis and Rheumatism* **39**, S172.

Kujala UM *et al*. (1995). Knee osteoarthritis in former runners, soccer players, weight lifters and shooters. *Arthritis and Rheumatism* **38**, 539–46.

LaPlante MP (1988). *Data on disability from the National Health Interview Survey (1983-5). An InfoUse Report*. National Institute of Disability and Rehabilitation, Washington, DC.

Morreale P *et al*. (1996). Comparison of the antiinflammatory efficacy of chondroitin sulfate and diclofenac sodium in patients with knee osteoarthritis. *Journal of Rheumatology* **23**, 1385–91.

Oliveria SA *et al.* (1995). Incidence of symptomatic hand, hip and knee osteoarthritis among patients in a health maintenance organization, *Arthritis and Rheumatism* **38**, 1134–41.

Setnikar I, Pacinic MA, Revel L (1991). Antiarthritic effects of glucosamine sulfate studied on animal models. *Arzeimitte-Forshung* **41**, 542–5.

18.9 Crystal-related arthropathies

S. C. O'Reilly and M. Doherty

Introduction

Diversity and terminology

A large number of crystals have been associated with acute synovitis, chronic arthropathy, or periarticular syndromes (Table 1). In practice only monosodium urate monohydrate, calcium pyrophosphate dihydrate, and basic calcium phosphates (mainly hydroxyapatite) are commonly encountered.

The taxonomy of these conditions is not universally agreed. Difficulties arise from our poor understanding of pathogenesis, historical extrapolation from gout to other crystal-related conditions, and multiple terms for the same clinical syndrome. Possible relationships between crystals and disease are outlined in Fig. 1. A 'crystal deposition disease' is defined as a pathological condition associated with mineral deposits which contribute directly to the pathology. This is probably the situation for all manifestations of gout, for acute syndromes associated with calcium pyrophosphate dihydrate, and for acute apatite periarthritis. However, the role of non-urate crystals in chronic arthropathy is unclear and confounded by the following observations:

1. Most crystals lack disease specificity and occur in a variety of clinical settings, often unaccompanied by symptoms or other abnormality.
2. Crystal deposition may coexist with other rheumatic disease, most commonly osteoarthritis, and often follows rather than precedes articular damage.
3. Combined deposition of several crystal species is common ('mixed crystal deposition').

For descriptive purposes, confusion may be avoided by specifying the crystal, the site of involvement, and the clinical syndrome (for example, chronic urate olecranon bursitis, acute pyrophosphate arthritis of the knee).

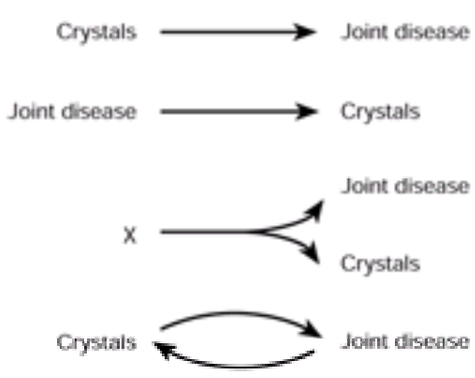

Fig. 1 Possible relationships between crystals and joint disease.

Crystal deposition and clearance

Many factors determine crystal formation and dissolution (Fig. 2). High solute concentrations alone are often insufficient to initiate crystal formation, and the presence of nucleating factors that aid initial particle formation and the balance of growth-promoting and inhibitory factors are probably more important. Little is known of such tissue factors, although they may in part explain:

1. the characteristic, limited distribution of different crystals;
2. the frequency of mixed crystal deposition (via epitaxial nucleation and growth of one crystal on another); and
3. non-specific predisposition to crystal formation in osteoarthritic tissues (via accompanying alterations in proteoglycan, collagen, and lipid).

Formation of crystals *in vivo* is a dynamic process, although usually slow. At any time the crystal load will depend on the rate of formation, the rate of dissolution, and trafficking of crystals away from their site of formation (via 'shedding' from preformed deposits with secondary uptake by synovial and other cells).

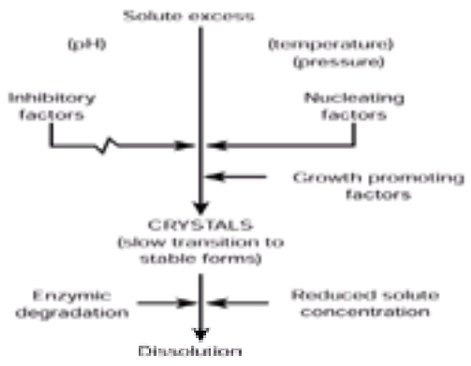

Fig. 2 Factors affecting crystal formation.

Crystal-induced inflammation and tissue damage

Crystals implicated in joint disease are stable, hard particles that exert biological effects via surface-active and mechanical properties. With respect to acute inflammation, they are all markedly phlogistic agents in a wide range of *in vitro* and *in vivo* systems. Surface-active interaction has been demonstrated with:

1. humoral mediators: for example complement activation via classical and alternative pathways; activation of Hageman factor;
2. cell-derived mediators: for example superoxide production and release of lysozymes, chemotactic factor, and lipoxygenase-derived products of arachidonic acid by neutrophils; release of interleukin I, interleukin 6, and tumour necrosis factor by monocytes and synoviocytes;
3. cell membranes: for example membranolysis of lysosomes, erythrocytes, and neutrophils; non-lytic platelet and neutrophil secretory responses.

In general, monosodium urate monohydrate is the most inflammatory, followed by calcium pyrophosphate dihydrate, then apatite and the less common crystals. In general, smaller particle size, marked surface irregularity, and high negative surface charge correlate with inflammatory potential. Some surface effects result from direct crystal contact but others are mediated via adsorbed protein, particularly immunoglobulin. Although adsorbed IgG may enhance inflammation, most other protein binding is inhibitory.

Less is known of chronic crystal-induced tissue damage. Postulated effects include persistent synovial inflammation, altered cell metabolism, and deleterious mechanical effects from large deposits. Evidence for activation of inflammatory mediators in chronic crystal-associated synovitis is lacking, although a chronic 'granulomatous' reaction often occurs around large accretions. The physicochemical effects of hard highly charged crystals embedded within cartilage, or occurring as wear particles at the surface, are largely unknown.

Gout (see also Chapter 11.4)

Monosodium urate monohydrate crystals are undoubted causal agents in gout, usually depositing in previously normal tissues and eliciting acute inflammation and eventual tissue damage. Their effective removal halts progression and results in 'cure'. In these respects gout is a true 'crystal deposition disease'.

The incidence of gout varies in populations from 0.2 to 0.35 per 1000, with an overall prevalence of 2.0 to 2.6 per 1000. Prevalence rises with age and increasing serum urate concentration. There is strong predominance in men (about 10:1), particularly under 65 years of age. Untreated gout evolves slowly through four clinical phases: asymptomatic hyperuricaemia, acute gout, intercritical gout, and chronic tophaceous gout.

Asymptomatic hyperuricaemia

Monosodium urate monohydrate crystals preferentially deposit in peripheral connective tissues in and around synovial joints, favouring lower rather than upper limbs. Deposits occur first in articular cartilage, most commonly the first metatarsophalangeal and small joints of the feet. Deposits later develop in synovium, capsule, and periarticular soft tissues, with progressive involvement of more proximal sites. Monosodium urate monohydrate crystals probably take months if not years to grow *in vivo* to detectable size, implying a long asymptomatic phase. Absence of inflammation during this period may relate to low crystal yield, positioning within hypovascular tissues, or inhibitory protein coating. Around 95 per cent of hyperuricaemic subjects remain asymptomatic throughout life, although how many have occult monosodium urate monohydrate deposits is unknown. Of those who develop gout, one in five will have suffered renal colic due to uric acid urolithiasis, sometimes more than a decade earlier. The first presentation of gout is usually acute synovitis, although an insidious onset of chronic arthropathy or nodular deposits ('tophi') occasionally occurs without preceding attacks.

Acute attacks

The classical attack

In almost all initial episodes a single peripheral joint is involved. This is the first metatarsophalangeal joint ('podagra') in 50 per cent of first attacks and 70 per cent of all attacks. Other common sites are the knee, ankle, midtarsal joints, small hand joints, wrist, and elbow. The axial skeleton and large central joints are rarely involved and never as the first site.

Attacks often wake the patient in the early morning with localized irritation and aching. Within a few hours the joint and surrounding tissues are swollen, hot, red, shiny, and extremely painful. The patient cannot bear even bedclothes to touch the joint and it is often described as 'the worst pain ever experienced'. Inflammation is maximal within 24 h and is often associated with pyrexia and malaise. Examination reveals florid synovitis and swelling, extreme tenderness, and overlying erythema. If left untreated, the attack resolves spontaneously over 5 to 15 days, often with pruritus and desquamation of overlying skin.

Although many attacks occur spontaneously, certain situations encourage shedding of preformed monosodium urate monohydrate crystals and triggering of acute attacks. Suggested mechanisms include mechanical loosening (local trauma), partial dissolution and reduction of crystal size (initiation of hypouricaemic treatment), and local increase in cytokines which encourage inflammatory responses to crystals and facilitate crystal escape via alterations in cartilage matrix (intercurrent illness, surgery). Although some triggers (alcohol, dietary excess, diuretics) increase local urate levels, acute crystallization is considered unlikely.

Atypical attacks

Acute attacks may manifest as tenosynovitis, bursitis, or cellulitis. Many patients describe mild episodes of discomfort without swelling lasting a day or so. Ten per cent of all typical attacks involve more than one joint. Sometimes acute gout, by triggering the acute response, provokes migratory attacks in other joints over subsequent days ('cluster attacks'). Polyarticular attacks are rare, usually occurring after a long history of recurrent attacks: marked systemic upset, fever, and confusion may dominate the clinical picture.

Intercritical periods

These are asymptomatic intervals between attacks. Some patients never have a second attack, in others the next episode occurs after many years; in most, however, a second attack occurs within 1 year. Subsequently the frequency of attacks and number of sites involved gradually increase with time. Later attacks are more often pauciarticular or polyarticular and more severe. Eventually, recurrent attacks and continuing monosodium urate monohydrate deposition cause joint damage and chronic pain. The interval between the first attack and development of chronic symptoms is variable, but averages about 10 years. The principal determinant is the serum uric acid—the higher it is, the earlier and more extensive the development of joint damage and tophaceous deposits.

Chronic tophaceous gout

Large crystal deposits ('tophi') produce irregular firm nodules, principally around extensor surfaces of fingers, hands, the ulnar surface of the forearms, olecranon bursae, Achilles tendons, first metatarsophalangeal joints, and the cartilaginous helix of the ear. Marked asymmetry, both locally and between sides, is particularly characteristic (Fig. 3). Monosodium urate monohydrate crystals beneath the skin may give a 'chalky' appearance (Fig. 4). If untreated, tophi can enlarge into gross knobbly swellings that may ulcerate, discharging material which is white and gritty and causes local inflammation (erythema, pus) even in the absence of secondary infection. If extensive, tophi may rarely involve the eyelids, tongue, larynx, or heart (causing conduction defects and valvular dysfunction).

Fig. 3 Chronic tophaceous gout affecting the hands. Note the eccentric nature of the tophi and the asymmetry between sides.

Fig. 4 Diuretic-induced gout in an elderly women, showing tophaceous deposition on pre-existing nodal osteoarthritis; the white monosodium urate monohydrate crystals are clearly visible beneath the skin.

Joints most commonly involved with signs of damage (restricted movement, crepitus, deformity) and varying degrees of synovitis are the first metatarsophalangeal joints, midfoot, small finger joints, and wrists. As with tophi, joint involvement is usually asymmetrical. Occasionally gross destruction may occur in feet and hands, and less commonly other sites. Acute attacks may become less of a feature as chronic symptoms become established. If untreated, the combination of extensive joint destruction and large tophi may cause grotesque deformities, particularly of hands and feet. Ankylosis is a rare late event. Although axial involvement is rare even in late stages, gouty involvement of hips, shoulders, spine and sacroiliac joints, and spinal cord compression by tophi, are all reported.

Classification

Traditional classification into primary or secondary gout depends on defining predisposing factors for hyperuricaemia. Most gout is primary, strongly predominating in men, with initial onset between 30 and 60 years of age (Table 2). Presentation is with acute attacks, and untreated disease progresses to chronic tophaceous gout. Such patients often give a family history of gout and are 'undersecretors' of uric acid.

Secondary gout usually results from chronic diuretic therapy and presents in older subjects (over 65 years). This increasingly common form affects women and men and is often associated with osteoarthritis. Upper and lower limb joints are affected equally. Acute attacks are less prominent and presentation is often with painful, sometimes discharging, tophaceous deposits in Heberden's and Bouchard's nodes (Fig. 4).

Associations

Hyperuricaemia

Mechanisms resulting in decreased excretion or increased production of uric acid are fully discussed in Section 11.4. Though hyperuricaemia and gout are strongly linked, they are not synonymous. Most hyperuricaemic subjects never develop gout, emphasizing the importance of local tissue factors in crystal nucleation/growth, and gouty patients may not be hyperuricaemic at presentation. Associations also differ: for example, impaired glucose tolerance, ischaemic heart disease, and hypertension are common in men with gout but do not associate with hyperuricaemia *per se*. The majority (75–90 per cent) of patients with primary gout are 'undersecretors' of uric acid, having inherited an isolated renal lesion that reduces fractional urate clearance; fewer than 10 per cent are 'overproducers' of uric acid. The cause usually remains unclear, although a very few have an inherited purine enzyme defect (see Chapter 11.4). Some patients are both undersecretors and overproducers.

Clinical associations differ according to gender. In men important associations are obesity, excessive alcohol intake, type IV hyperlipoproteinaemia, impaired glucose tolerance, and ischaemic heart disease (Table 3). Obesity and lifestyle, rather than hereditary factors, appear to be central factors linking these 'associations of plenty'. Excessive beer drinking, with its high calorie and alcohol intake, is a common form of alcohol abuse associated with gout. The nineteenth century association with port is partly explained by addition of lead to sweeten the port: lead inhibits uric acid excretion and also promotes nucleation of monosodium urate monohydrate, predisposing to 'the gout, the colic and the palsy' (all features of lead poisoning). However, 'saturnine' gout still occurs in individuals who drink alcohol distilled or stored in lead-contaminated containers ('moonshine'). Alcohol is a less common association in women, usually occurring in young to middle-aged, thin, spirit drinkers. In women the main associations are diuretic therapy, chronic renal impairment, and pre-existing osteoarthritis. Other drugs may predispose to gout, such as aspirin in low doses and cyclosporin.

A strong negative association exists between gout and rheumatoid arthritis. This remains unexplained, but probably reflects impaired nucleation/growth of monosodium urate monohydrate crystals rather than masking of monosodium urate monohydrate crystal-induced inflammation (for example, by crystal coating with rheumatoid factors).

Renal disease

Urolithiasis (see also Chapter 20.13)

Uric acid stones account for 5 to 10 per cent of all stones in the United Kingdom and United States, and up to 40 per cent in Israel. A history of renal colic is seen in 10 to 25 per cent of patients with gout, the important aetiological factors being low urinary pH, low urinary volume, and high urinary uric acid concentration. In particular, high urinary concentrations occur in overproducers of uric acid, if renal urate clearance is increased (uricosuric drugs, defects in tubular reabsorption), and

in situations of dehydration with lowering of urinary pH (diarrhoea, ileostomy). Gouty subjects also have an increased incidence of calcium-containing stones, particularly calcium oxalate, with no detectable uric acid nidus.

Acute uric acid nephropathy describes rapid precipitation of uric acid crystals in renal collecting ducts with secondary acute obstructive renal failure. This event correlates with the amount of uric acid excreted rather than the level of hyperuricaemia. Strongly acid urine, which reduces uric acid solubility, potentiates the problem. The condition is most likely in ill, dehydrated patients with lymphoma or malignancy subjected to aggressive chemotherapy without adequate prophylactic treatment with allopurinol. It also occurs in gouty patients with markedly accelerated purine synthesis, for example following excessive exercise or after epileptic seizures. The condition is largely avoidable by appropriate hydration, urine alkalinization, and allopurinol prophylaxis.

Chronic urate nephropathy (see also Section 20.10)

Widespread monosodium urate monohydrate deposition in the interstitium of the medulla and pyramids results in crystal-induced inflammation with surrounding giant-cell reaction and fibrosis, affecting in particular the tubular epithelium of the loop of Henle and juxtaposed interstitial tissues. Subsequent changes include glomerular hyalinization and hypertrophy of the intima and media of arterioles. Hypertensive damage, tubular obstruction, and secondary pyelonephritis may all complicate this picture. Albuminuria and inability to concentrate the urine maximally are early clinical manifestations. Progressive renal disease is an important complication of untreated chronic tophaceous gout, end stage renal failure occurring in up to 25 per cent of cases.

In advanced renal disease of any cause, calcium oxalate or phosphate crystals may deposit in renal parenchyma but are predominantly cortical in location (compare the medullary site of monosodium urate monohydrate).

The association between parenchymal disease and less severe gout remains controversial, being confounded in males by frequent accompanying obesity, hypertension, and drug therapy. The minor progression of renal insufficiency that occurs in most gouty patients, however, is probably largely age-related and life expectancy is not reduced.

Investigations and diagnosis

The history and signs of classical acute or chronic tophaceous gout are highly characteristic, and with a raised serum uric acid a strong presumptive diagnosis is readily made. However, definitive confirmation requires demonstration of monosodium urate monohydrate crystals by compensated polarized light microscopy of fluid from a gouty joint, bursa, or tophus. Synovial fluid in acute attacks is typically turbid with diminished viscosity and greatly elevated cell count (more than 90 per cent neutrophils). Chronic gouty fluid is more variable, but occasionally appears white due to the high crystal load. Only a few drops collected directly on to a slide are required for crystal identification (Chapter 18.3). Monosodium urate monohydrate crystals are seen readily as strongly birefringent (negative sign), needle-shaped crystals, 5 to 20 μm in length, within cells or occurring freely in fluid. In tophaceous material they occur as dense, tightly packed sheets. During intercritical periods, aspiration of an asymptomatic first metatarsophalangeal joint or knee often permits confirmation of the diagnosis by revealing monosodium urate monohydrate crystals.

Hyperuricaemia is confirmed if two or more fasting serum uric acid levels exceed the normal range for the patient's age and sex. Uric acid levels may be lowered during an acute attack of gout and should be remeasured during an intercritical period. In primary gout in a young patient, determination of undersecretion or overproduction of uric acid is best undertaken by measuring total urinary excretion on a low-purine diet, but a quick guide is given by the uric acid/creatinine ratio estimated on a single urine sample (normally less than 0.5). In young overproducers, a purine enzyme defect becomes more likely and should be sought (Chapter 11.4). Assessment of renal function (creatinine, urea, electrolytes, urine testing) should always be undertaken (Table 3). Measurement of fasting lipoprotein concentrations should be made in all patients with primary gout. An intercritical full blood count and measurement of erythrocyte sedimentation rate/viscosity should detect any underlying myeloproliferative disease. During acute attacks a marked acute phase response (high erythrocyte sedimentation rate, neutrophil leucocytosis, thrombocytosis, elevated C-reactive protein) is usual; modest elevations of erythrocyte sedimentation rate also accompany chronic gout.

Radiographs supplement the clinical assessment of structural damage but can also aid diagnosis. In early disease they are usually normal. During acute gout, non-specific soft tissue swelling (rarely juxta-articular osteopenia) may be evident. After repeated attacks, and in chronic disease, joint space narrowing, sclerosis, cysts, and osteophytes (that is, the changes of osteoarthritis) become more frequent in feet and hands. Gouty 'erosions' are a less common but more specific abnormality, occurring as para-articular 'punched-out' bone defects with well-demarcated sclerotic margins, overhanging hooks of bone, and retained bone density (Fig. 5). They are typically asymmetric, eccentric lesions positioned away from the 'bare area' of the joint, contrasting with more symmetrical, ill-defined marginal erosions (with osteopenia) of rheumatoid arthritis. Tophi appear as eccentric soft tissue swellings, occasionally with patchy calcification due to epitaxial growth of apatite. In late disease, severe destructive change with osteopenia may occur and distinction from rheumatoid arthritis or other conditions becomes more difficult.

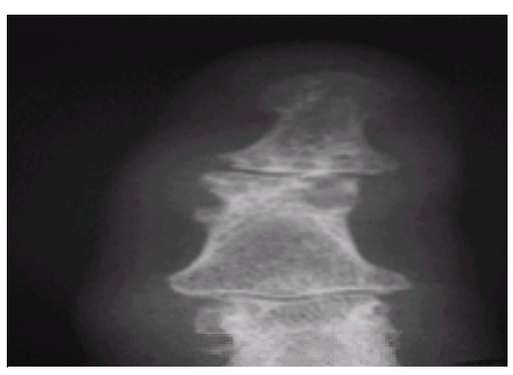

Fig. 5 Characteristic radiographic changes of established gout in a finger: joint space loss and cystic change at the distal interphalangeal joint, 'pressure erosions' with overhanging bony 'hooks' at both interphalangeal joints, and eccentric soft tissue swelling at the proximal joint.

Differential diagnosis

Acute attacks

Sepsis and other crystal-associated synovitis are the main considerations. Gout and sepsis may coexist, as may monosodium urate monohydrate and calcium pyrophosphate dihydrate deposition (particularly in elderly subjects). Examination of aspirated fluid for both crystals and sepsis (Gram stain, culture) is the only sure way of obtaining the correct diagnosis. A wider search for sepsis may be indicated (for example blood and urine cultures), particularly in those who are ill. With less classic attacks, other conditions that may be considered include psoriatic and acute Reiter's arthropathy, acute sarcoid arthropathy, traumatic arthritis, palindromic rheumatism, and exacerbation of osteoarthritis.

Chronic tophaceous gout

Other causes of arthritis and periarticular swellings/nodules that require differentiation are rheumatoid arthritis, generalized nodal osteoarthritis, xanthomatosis with arthropathy, and multicentric reticulohistiocytosis. Gout is usually less symmetrical in distribution than these conditions and, except for xanthomatosis, acute attacks are not a feature. Nodal osteoarthritis, of course, may coexist with gout. Aspiration (joint fluid, nodules) and plain radiographs readily facilitate correct diagnosis.

Treatment

Acute gout

The treatment aim is pain relief by reducing inflammation and intra-articular hypertension. Alteration of uric acid levels is avoided until the attack has resolved, since

initiation of hypouricaemic drugs may prolong the attack.

Rapid symptom relief may be obtained with a quick-acting non-steroidal anti-inflammatory drug (**NSAID**), given in full dosage. Although indomethacin has a long tradition in this context, it is preferably avoided in the elderly due to its frequent renal, gut, and nervous system side-effects.

Oral colchicine is rapidly effective within a few hours (1 mg immediately, followed by 0.5 mg every 6 h until symptoms abate). Unfortunately, at the doses necessary, diarrhoea, nausea, and abdominal cramps are common, causing the patient 'to run before he can walk'. Colchicine, however, is a useful alternative if NSAIDs are contraindicated. Intravenous colchicine, however, is particularly toxic and should never be used. Although previously used as a 'diagnostic test' the efficacy of colchicine is not specific to gout: it also ameliorates other crystal-associated syndromes.

Joint aspiration often provides immediate relief by reducing intra-articular hypertension. Intra-articular steroid is useful for large joints such as the knee, or when NSAIDs or colchicine are contraindicated or unsuccessful. In difficult cases, joint lavage may terminate an attack, and for troublesome polyarticular attacks there is support for the use of parenteral steroid.

Long-term management

Once any acute attack has resolved, long-term strategies need consideration. Gout is potentially curable. Treatment may involve:

1. considering and eliminating modifiable factors that cause hyperuricaemia; and
2. utilizing hypouricaemic drugs.

Management of gout may require alteration in lifestyle and chronic medication: patient compliance and motivation, which depend on appropriate education and counselling, are essential for success.

Modification of provoking factors

In early primary gout, gradual weight loss, reduction in alcohol consumption, and avoidance of toxins (low-dose aspirin, lead) may alone be sufficient. Similarly, in diuretic-induced gout, stopping the diuretic (plus substitution of alternative therapies) may prove possible and be all that is required.

Hypouricaemic drug therapy

Indications for drug therapy are:

1. recurrent, troublesome acute attacks;
2. presence of tophi;
3. bone or cartilage damage;
4. coexistent renal disease, uric acid urolithiasis;
5. very high uric acid levels (particularly with overproduction and hyperexcretion).

The logical approach would be allopurinol for overproducers and uricosurics for undersecretors. In practice, however, allopurinol is the usual drug of choice, permitting flexible tailoring of dose to reduce urate levels below the solubility limit. Allopurinol inhibits xanthine oxidase and often also depresses *de novo* purine synthesis. The starting dose is 100 to 300 mg daily, which is then adjusted within the range 100 to 900 mg daily according to the serum uric acid level (initially checked monthly). In patients with renal insufficiency, particularly the elderly, excretion of the active metabolite oxypurinol is delayed: the starting dose should therefore be 100 mg daily and adjustments made cautiously.

The uricosurics probenecid (0.5–1.0 g twice a day) and sulphinpyrazone (100 mg three or four times daily), which prevent proximal tubular reabsorption of urate, are rarely used. Benzbromarone, a newer uricosuric, is now increasingly used in parts of Europe. Uricosurics are alternatives to allopurinol in patients with normal renal function but are contraindicated in those with renal impairment, urolithiasis, or gross overproduction of uric acid. The therapeutic aim of hypouricaemic therapy is to maintain the serum uric acid well within the normal range (preferably the lower half). Treatment should be lifelong.

Acute attacks may be provoked during the first few months of hypouricaemic treatment. Prophylactic colchicine (0.5 mg twice a day) or a standard dose of NSAID given for the first 2 to 3 months of treatment largely avoids 'breakthrough' attacks. With any uricosuric, high fluid intake and urine alkalinization in the early weeks of treatment are recommended to avoid deposition of uric acid within the kidney.

Serious side-effects are unusual with any hypouricaemic drugs. Rare problems include toxic epidermal necrolysis, interstitial nephritis and vasculitis (allopurinol), nephrotic syndrome (probenecid), and hepatitis and marrow suppression (both drugs). Important interactions with allopurinol occur with coumarin anticoagulants (due to hepatic microsomal enzyme inhibition) and purine analogues (such as azathioprine) which are inactivated by xanthine oxidase. Associated hypertension should be treated, but preferably not with diuretics which elevate serum urate and may provoke acute attacks.

Pyrophosphate arthropathy

Deposition of calcium pyrophosphate dihydrate crystals ($Ca_2P_2O_7.2H_2O$) in articular cartilage is a common age-related phenomenon. Calcium pyrophosphate dihydrate crystals preferentially deposit within fibrocartilage and are the most common cause of cartilage calcification (chondrocalcinosis).

Calcium pyrophosphate dihydrate deposition may occur in otherwise normal cartilage or associate with structural change and clinical arthropathy—'pyrophosphate arthropathy'. A causal role for calcium pyrophosphate dihydrate crystals in acute inflammation is accepted, but their role in chronic arthropathy is unclear. The strong association/overlap with osteoarthritis has led some to consider pyrophosphate arthropathy not as a crystal deposition disease but as a 'subset' of osteoarthritis, with calcium pyrophosphate dihydrate a 'process' marker associating with a hypertrophic articular response.

Radiographic chondrocalcinosis is rare under 50 years of age, but its prevalence rises from 10 to 15 per cent in those aged 65 to 75, to 30 to 60 per cent in those over 85, showing female preponderance (relative risk 1.33) and association with osteoarthritis (relative risk 1.52 at the knee). No epidemiological data exist for pyrophosphate arthropathy, but in patient series the mean age of presentation is around 65 to 75 with female preponderance (about 2:1 to 3:1), particularly in older patients.

Clinical features

Common presentations are acute synovitis, chronic arthritis, or as an incidental finding. Other presentations are rare.

Acute synovitis ('pseudogout')

This is the most common cause of acute monoarthritis in the elderly. Attacks may occur as isolated events or be superimposed upon a background of chronic symptomatic arthropathy. Most attacks occur spontaneously, but provoking factors include intercurrent illness, surgery, and local trauma. Although any joint may be involved, the knee is by far the most common site, followed by the wrist, shoulder, and ankle. Concurrent attacks in several joints are uncommon (fewer than 10 per cent of cases) and polyarticular attacks rare.

The typical attack develops rapidly with severe pain, stiffness, and swelling, becoming maximal within 6 to 24 h of onset. Examination reveals a very tender joint with signs of florid synovitis (increased warmth, tense effusion, restricted movement with stress pain) and often overlying erythema. Fever is common and elderly patients may appear unwell or mildly confused. Attacks are self-limiting, usually resolving within 1 to 3 weeks.

Chronic pyrophosphate arthropathy

This common condition affects mainly elderly women and targets the same large and medium-sized joints as pseudogout. Knees are the usual and most severely affected joint. Presentation is with chronic pain, stiffness, and functional impairment (plus superimposed acute attacks). Symptoms usually relate to just a few joints, although examination often reveals more widespread abnormalities. Affected joints show signs of osteoarthritis (crepitus, bony swelling, restricted movement) with varying degrees of synovitis (often most marked at the knee, radiocarpal, or glenohumeral joint). Knees typically show abnormality of two or three compartments; valgus or varus deformity may occur.

Although symptoms and signs are those of osteoarthritis, chronic pyrophosphate arthropathy may often be distinguished from uncomplicated osteoarthritis by:

1. the joint distribution: in osteoarthritis wrist, glenohumeral, ankle, elbow, and midtarsal involvement are uncommon;
2. the often marked inflammatory component; and
3. superimposition of acute attacks.

The outcome for chronic pyrophosphate arthropathy is generally good, most patients running a relatively benign course, particularly with respect to small and medium-sized joints. If progression occurs, it is usually slow and related to knees, hips, or shoulders. Occasionally severe, rapidly progressive, destructive arthropathy develops at these sites. This is virtually confined to very elderly women and is associated with severe pain, recurrent haemarthrosis (shoulder, knee), and occasional joint leakage.

Incidental finding

As with osteoarthritis, clinical or radiographic evidence of pyrophosphate arthropathy and chondrocalcinosis are not uncommon incidental findings in the elderly, and may confound the cause of regional pain if a through history and examination are not undertaken.

Uncommon presentations

Acute tendinitis (triceps, Achilles), tenosynovitis (hand flexors, extensors), and bursitis (olecranon, infrapatellar, retrocalcaneal) occur uncommonly, usually in patients with widespread calcium pyrophosphate dihydrate crystals. Median and ulnar nerve compression at the wrist may accompany flexor tenosynovitis. Rare tophaceous deposits of calcium pyrophosphate dihydrate usually present as solitary lesions in areas of chondroid metaplasia.

Classification and associations

Calcium pyrophosphate dihydrate deposition is traditionally classified as:

1. hereditary;
2. associated with metabolic disease; or
3. sporadic/idiopathic (by far the commonest, associated with osteoarthritis).

Familial predisposition

This is reported from many countries and different ethnic groups. Two clinical phenotypes occur: early onset (third to fourth decade) florid polyarticular chondrocalcinosis with variable severity of accompanying arthropathy; and late onset (sixth to seventh decade) oligoarticular chondrocalcinosis (mainly knee) with arthritis resembling sporadic disease. The pattern of inheritance varies, although autosomal dominance is usual. Studies on several families have suggested the short arm of chromosome 5 as a potential site for a candidate gene. The mechanism of familial predisposition remains unclear and may differ between families. A primary cartilage abnormality that promotes calcium pyrophosphate dihydrate crystal nucleation and growth (Swedish and Japanese families), and a generalized abnormality of pyrophosphate metabolism resulting in a local increase in cartilage levels (French and American kindreds) have both been reported.

Metabolic disease associations

Inorganic pyrophosphate is a byproduct of many biosynthetic reactions, with a turnover of several kilograms per day. Much extracellular inorganic pyrophosphate derives from ATP via the action of NTP pyrophosphatase, and is rapidly converted to orthophosphate by pyrophosphatases (particularly alkaline phosphatase) (Fig. 6). A number of metabolic diseases associate with deposition of calcium pyrophosphate dihydrate (Table 4), their association being rationalized through putative effects on metabolism of inorganic pyrophosphate. Suggested mechanisms include:

1. Reduced breakdown of inorganic pyrophosphate by alkaline phosphatase, due to (i) reduced levels, (ii) inhibitory ions (calcium, iron, copper), or (iii) impaired complexing with magnesium.
2. Enhanced nucleation by iron or copper.
3. Increased calcium concentration.
4. Increased production of pyrophosphate through stimulation of adenylate cyclase by parathyroid hormone.

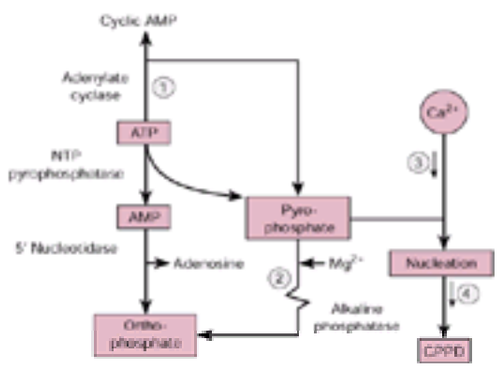

Fig. 6 Simplified scheme of extracellular pyrophosphate metabolism, showing putative sites of interaction by metabolic diseases. Hyperparathyroidism, 1, 2, 3; haemochromatosis, 2, 4; hypophosphatasia, 2; Wilson's disease, 2, 4; and hypomagnesaemia, 2. CPPD, calcium pyrophosphate dihydrate; NTP, nucleotide triphosphate

Osteoarthritis and joint insult

Several observations support a relationship between osteoarthntis and deposition of calcium pyrophosphate dihydrate, the latter often following rather than preceding joint damage. However, a negative association exists between deposition of calcium pyrophosphate dihydrate and rheumatoid arthritis, with atypical radiographic features in coexistent disease (retained bone density; marked osteophyte, cyst, and bone remodelling) suggesting that the primary association of calcium pyrophosphate dihydrate is with hypertrophic tissue response/osteoarthritis and not joint damage *per se*. The explanation for this association is unknown. Levels of inorganic pyrophosphate in the synovial fluid are increased in pyrophosphate arthropathy and osteoarthritis and are low in rheumatoid arthritis, but the order of change is unlikely to influence formation of calcium pyrophosphate dihydrate significantly. These crystals form in pericellular sites and associate with lipid, proteoglycan depletion, and adjacent hypertrophic chondrocytes containing lipid granules. It is therefore possible that reduction of inhibitors (such as proteoglycan)

and increase in promotors (such as lipid) may combine to copromote calcium pyrophosphate dihydrate formation in metabolically active osteoarthritic tissue.

Investigations and diagnosis

Critical investigations are synovial fluid analysis and plain radiographs. In pseudogout aspirated fluid is often turbid or bloodstained with an elevated cell count (more than 90 per cent neutrophils). Compensated polarized microscopy reveals calcium pyrophosphate dihydrate crystals as weakly birefringent (positive sign) rhomboids or rods, about 2 to 10 μm long. Calcium pyrophosphate dihydrate crystals are less readily identified and often less numerous than those of monosodium urate monohydrate; examination of a spun deposit may increase detection.

Radiographic aspects relate both to calcification and arthropathy. Chondrocalcinosis signifies extensive deposition and is not always evident: it mainly affects fibrocartilage (particularly knee menisci, wrist triangular cartilage, symphysis pubis), and less commonly hyaline cartilage (Fig. 7). Although occasionally monoarticular, it usually affects several sites. Calcification of capsule, synovium, and tendons is less common. Chondrocalcinosis and calcification may increase or decrease with time, diminishing chondrocalcinosis often accompanying crystal 'shedding' or cartilage loss.

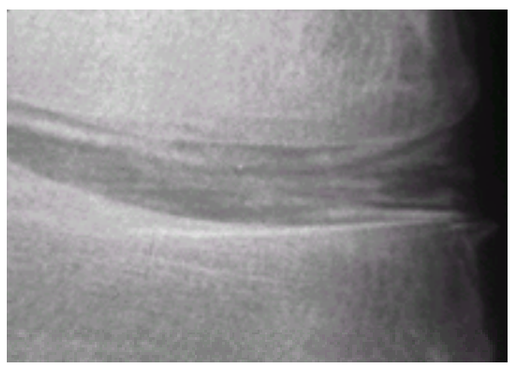

Fig. 7 Radiographic chondrocalcinosis of the knee, affecting meniscal fibrocartilage (central, triangular) and hyaline cartilage (linear, parallel to bone).

Changes of arthropathy are those of osteoarthritis: cartilage loss, sclerosis, cysts, and osteophyte. However, characteristics which suggest pyrophosphate include:

1. distribution between and within joints that is atypical of osteoarthritis (for example glenohumeral disease; isolated or predominant patellofemoral or radiocarpal involvement);
2. prominence of osteophytes and cysts; and
3. prominent osteochondral bodies.

Such combined features may present a distinctive 'hypertrophic' appearance even in the absence of chondrocalcinosis (Fig. 8). In destructive arthropathy, marked cartilage and bone attrition with fragmentation and loose osseous bodies may resemble a Charcot joint.

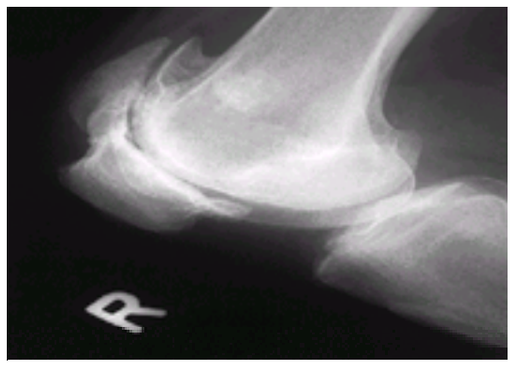

Fig. 8 Lateral knee radiograph showing predominant patellofemoral involvement by 'hypertrophic' osteoarthritis characteristic of pyrophosphate arthropathy.

Metabolic predisposition is rare and routine screening of all patients is unrewarding. Nevertheless, arthritis associated with calcium pyrophosphate dihydrate crystals may be the presenting feature of metabolic disease, and a search is warranted in the following circumstances:

1. early onset arthritis (under 55 years)
2. florid polyarticular chondrocalcinosis; or
3. presence of additional clinical or radiographic clues.

A reasonable screen would include serum calcium, alkaline phosphatase, magnesium, ferritin, and liver function.

Differential diagnosis

The principal differential diagnosis for pseudogout is sepsis or gout, both of which may coexist with calcium pyrophosphate dihydrate deposition. Gram stain and culture of joint fluid should be undertaken even when calcium pyrophosphate dihydrate (and monosodium urate monohydrate) crystals are identified. Marked bloodstaining may lead to consideration of other causes of haemarthrosis, especially a bleeding disorder or subchondral fracture.

Chronic pyrophosphate arthropathy is usually readily distinguished from rheumatoid arthritis by the synovial fluid and radiographic findings, the infrequency of severe systemic upset, absence of extra-articular features, and an acute phase response which is only modest. Proximal stiffness due to glenohumeral involvement may suggest polymyalgia rheumatica, although clinical examination and near normal erythrocyte sedimentation rate should exclude the diagnosis. Destructive pyrophosphate arthropathy may simulate a neuropathic joint, although such joints are severely symptomatic and neurological abnormality is absent.

Treatment

Pseudogout

Since pseudogout usually affects only one or a few joints in elderly patients, local therapy is preferred. Aspiration alone often relieves symptoms, but may be combined with intra-articular steroid in florid cases. Simple analgesics and NSAIDs give additional benefit but should be used cautiously in the elderly. Joint lavage is reserved for troublesome steroid-resistant cases. Colchicine is effective but rarely warranted. Triggering illness (for example chest infection) will require appropriate treatment. Rapid mobilization should be instituted once the synovitis is settling.

Chronic pyrophosphate arthropathy

Unlike gout there is no specific therapy, and treatment of any underlying metabolic disease does not influence outcome. Aims are to reduce symptoms and maintain or

improve function. This may include education of the patient in appropriate use of the affected joints, reduction in obesity, improvement of muscle strength, use of a stick or other walking aid, and surgery for severe disease. Chronic synovitis may be improved by intermittent steroid injection or intra-articular radiocolloid (yttrium-90). As with pseudogout, symptomatic drugs are to be used with caution in older patients; simple analgesics are generally preferable to NSAIDs.

Other crystal-related disorders

Apatite-associated syndromes

Hydroxyapatite is the principal bone mineral. Apatites or basic calcium phosphates (partially carbonate-substituted hydroxyapatite, octacalcium phosphate, rarely tricalcium phosphate) are also the usual mineral to deposit in extraskeletal tissues (for example tuberculous lesions, arteries).

The [calcium × phosphate] product must be kept high to maintain skeletal integrity. Specific cellular mechanisms activate calcification where appropriate (for example matrix vesicles in growing cartilage), whilst other mechanisms (such as pyrophosphate and aggregated proteoglycan) inhibit calcification elsewhere. In general, abnormal calcification results from:

1. elevation of the [calcium × phosphate] product, causing widespread 'metastatic' calcification, or
2. alteration in the balance between inhibitory and promoting tissue factors, resulting in local 'dystrophic' calcification.

In rheumatic diseases abnormal deposition of basic calcium phosphates may occur in:

1. periarticular tissues (particularly tendon);
2. hyaline cartilage, in association with osteoarthritis; or
3. subcutaneous tissues and muscle, principally in connective tissue diseases.

Apatite crystals are too small (5–500 nm) to be seen by light microscopy. Particles may aggregate, however, to form spherulites visible with the light microscope. Confirmation of basic calcium phosphates requires sophisticated analytical techniques and most clinical diagnoses are presumptive, based on radiographic calcification or non-specific staining of joint fluid or histological material.

Acute calcific periarthritis

Apatite deposition in the supraspinatus tendon (Fig. 9) is a relatively common incidental finding (about 7 per cent of adults). It occasionally results in severe acute inflammation of the subacromial bursa, periarticular tissues, or joint itself. Periarticular sites around the greater hip trochanter, the foot, or the hand are less commonly affected.

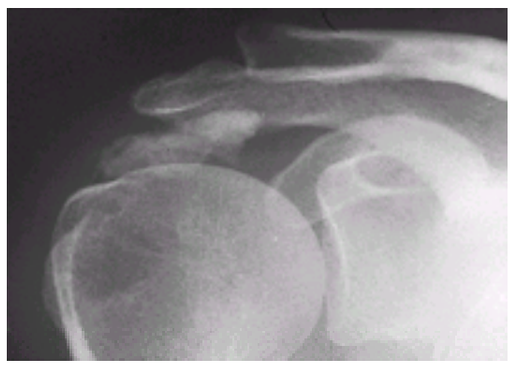

Fig. 9 Shoulder radiograph showing florid supraspinatus tendon calcification (calcific periarthritis).

Acute episodes may follow local trauma or occur spontaneously. Within a few hours pain and tenderness are often extreme and the area appears swollen, hot, and red. Modest systemic upset and fever are common. Sepsis is usually considered first, but the diagnosis is made following demonstration of radiographic calcification. If the lesion is aspirated, thick white fluid containing many apatite aggregates may be obtained. The condition usually resolves spontaneously over 1 to 3 weeks, often accompanied by radiographic dispersal of modestly sized calcifications (crystal 'shedding'). NSAIDs ameliorate symptoms and the attack can be abbreviated by aspiration and injection of steroid. Large deposits may require surgical removal. Calcific periarthritis rarely results from metabolic abnormality (renal failure, hyperparathyroidism, hypophosphatasia) and measurements of serum calcium, alkaline phosphatase, and creatinine are usually normal. Rare families are predisposed to calcific periarthritis at multiple sites with no evidence of altered calcium phosphate product.

Osteoarthritis and apatite-associated destructive arthritis

Modest amounts of basic calcium phosphates are commonly found in synovial fluid from osteoarthritic joints, in isolation or with calcium pyrophosphate dihydrate ('mixed crystal deposition'). Whether apatite plays any part in inflammatory exacerbations or associates with severity or progression of osteoarthritis remains uncertain.

The uncommon condition 'apatite-associated destructive arthritis' is often considered a 'subset' of osteoarthritis. It is virtually confined to elderly women and affects the hip, shoulder ('Milwaukee shoulder'), or knee. It has the general appearance of severe large joint osteoarthritis but is particularly characterized by:

1. rapid progression, often leading to severe pain and disability within a few months of onset;
2. development of marked instability;
3. large, cool effusions;
4. an atrophic radiographic appearance with marked cartilage and bone attrition and little osteophyte or bone remodelling.

Aspirated fluid has normal viscosity and a low cell count but contains large amounts of apatite aggregates, seen readily on light microscopy following non-specific calcium staining (alizarin red, acidic pH). The differential diagnosis may include sepsis (excluded by synovial fluid culture), late avascular necrosis, or neuropathic joint. The pathogenesis of this condition remains unclear. Although apatite particles could contribute to tissue damage by stimulating release of collagenase and other proteolytic enzymes from synovial cells, it is most likely that the apatite is non-contributory and principally reflects the severity of subchondral bone attrition. The outcome is poor and inevitably requires surgical intervention.

Other apatite syndromes

Deposition of tophaceous periarticular apatite may occur in patients with chronic renal failure managed by dialysis. Apatite has also been incriminated in the occasional erosive interphalangeal arthropathy seen in such patients.

Other crystals

Cholesterol

Cholesterol crystals may induce acute synovitis, acute tenosynovitis, and chronic xanthomatous tendinitis in hypercholesterolaemic subjects. Cholesterol and other lipid crystals may also occur as a non-specific finding in chronic synovitis, most commonly due to rheumatoid arthritis. In this situation the lipid probably derives from

cellular debris and its pathogenic significance is uncertain.

Oxalate

Oxalate crystals have been incriminated in acute and chronic articular and periarticular syndromes occurring in association with either primary familial oxalosis (types I and II) or secondary oxalosis (Chapter 11.10). Chronic renal failure managed with dialysis is the commonest cause of secondary oxalosis, particularly if ascorbic acid supplementation has been given. Acute symmetrical interphalangeal and metacarpophalangeal arthritis, with or without tenosynovitis, and digital calcific deposits are the usual manifestation. Large joint involvement, chondrocalcinosis, and tophaceous periarticular masses are less common. Calcium oxalate crystals may also cause life-threatening organ involvement, with peripheral vascular insufficiency and digital necrosis, cardiomyopathy, peripheral neuropathy, and aplastic anaemia. There is no effective treatment.

Extrinsic crystals

These are a rare cause of locomotor problems. Acute flares following intra-articular injection of corticosteroids are uncommon but may represent iatrogenic crystal-induced inflammation. Penetrating injuries involving plant thorns and sea-urchin spines may cause acute and chronic inflammatory synovitis, periostitis, or periarticular lesions which only resolve following surgical removal of the crystalline material.

Further reading

Doherty Mand Dieppe PA (1986). Crystal deposition disease in the elderly. *Clinics in Rheumatic Diseases* **12**, 97–116.

Emmerson BT (1996). The management of gout. *New England Journal of Medicine* **334**, 445–51.

McCarty DJ, ed (1988). Crystalline deposition diseases. *Rheumatic Disease Clinics of North America* **14**, 2.

Reginato A, Kurnik B (1989). Calcium oxalate and other crystals associated with kidney diseases and arthritis. *Seminars in Arthritis and Rheumatism* **18**, 198–224.

Rosenthal AK (1998). Calcium crystal-associated arthritides. *Current Opinion in Rheumatology* **10**, 273–7.

18.10.1 Autoimmune rheumatic disorders and vasculitis

I. P. Giles and D. A. Isenberg

Definition and epidemiology

The autoimmune rheumatic diseases are a heterogeneous group of disorders characterized by clinical involvement of the joints, connective tissues, muscles, internal organs, Raynaud's phenomenon, and cutaneous manifestations. Hence the autoimmune rheumatic diseases include a broad clinical spectrum of disease, including systemic lupus erythematosus, rheumatoid arthritis, Sjögren's syndrome, scleroderma, dermatomyositis, polymyositis, antiphospholipid syndrome, and the vasculitides. This latter group of diseases all share inflammation and necrosis of blood vessels as cardinal features, and may be divided into primary (e.g. giant cell arteritis, Wegener's granulomatosis, polyarteritis nodosa, etc.), occurring in the absence of a recognized precipitating cause, or secondary to established disease (e.g. systemic lupus erythematosus or rheumatoid arthritis) or infection (e.g. hepatitis B, C, or HIV) (see Table 1). On the whole these diseases have a predilection for young women and share defects in immune regulation leading to the production of autoantibodies, activation of the complement system, and generation and deposition of immune complex.

Some autoimmune rheumatic diseases are rare, for example systemic sclerosis; others are common, rheumatoid arthritis affecting approximately 1 per cent of the population (see Table 2). Taken as a whole, however, these autoimmune disorders affect as many as 1 in 20 people. Some are severely debilitating or life-threatening illnesses, others produce minor symptoms that require little, if any, medical intervention.

The clinical spectrum

Each of the autoimmune rheumatic diseases is a distinct entity and can be clearly defined clinically, serologically, and in terms of treatment and prognosis. However, many patients with these diseases have non-specific features of malaise, fever, and arthralgia, and there is also much overlap in terms of multisystem involvement, as shown in Table 3. Organ-specific features, for example lung fibrosis, pericarditis, and less frequently glomerulonephritis, can all occur in several of the autoimmune rheumatic diseases and the presence of such a feature is not pathognomonic of an individual disease.

The clinical features of each patient must be considered together with the laboratory investigations, which should include an autoantibody profile. A preliminary 'autoimmune screen' includes a rheumatoid factor and antinuclear antibody test as a bare minimum, the results of which then guide the need for further autoantibody testing. Immunologically rheumatoid factor (especially if the titre is greater than 1 in 320) remains the most important guide to establishing the diagnosis of rheumatoid arthritis, although the American College of Rheumatology classification criteria for rheumatoid arthritis may still be fulfilled in the absence of rheumatoid factor. The antibody is of no value, however, in the monitoring of the disease.

The presence and pattern of staining of antinuclear antibody is a very useful guide to the presence of disease, as shown in Table 4. In the case of the vasculitides the antineutrophil cytoplasmic antibody (**ANCA**) fulfils this role. An important proviso to the antinuclear antibody test is that it is present in a low titre (up to 1 in 80) in about 1 to 2 per cent of the normal population, and more frequently (up to 10 per cent) in healthy people over the age of 75 years. Hence, its presence alone at low titres does not in itself justify the diagnosis of an autoimmune rheumatic disease: the whole clinical picture must be considered. Further confusion may arise because some autoantibodies may be found in more than one disease, such as anti-U1RNP (in systemic lupus erythematosus and undifferentiated autoimmune rheumatic disease), whilst others may be found in other diseases 'beyond' the autoimmune rheumatic diseases, such as perinuclear staining ANCA (p-ANCA) which is well recognized in patients with inflammatory bowel disease, some chronic infections, and malignancies.

Immunopathogenesis

Autoimmune rheumatic disorders

The precise aetiologies of the autoimmune rheumatic diseases remain unknown, but are undoubtedly complex. Inciting agents, such as infection, are involved, as are genetic susceptibility, hormonal factors, and both cellular and immune dysregulation.

Common to all of the autoimmune rheumatic diseases is the phenomenon of production of autoantibodies by activated B cells. Many of the pathogenic autoantibodies are of the IgG class and have undergone somatic mutation in their hypervariable regions leading to a gradual increase in specificity and binding affinity of an antibody produced by a particular clone of cells. This latter finding is particularly true of anti-dsDNA antibodies in systemic lupus erythematosus and antiphospholipid antibodies in the antiphospholipid syndrome.

The origins of autoantibody production remain an enigma. Mechanisms that have been invoked include antigen-driven T helper cell responses, failure of efficient clearance of nuclear antigens which become surface expressed following cellular apoptosis, and epitope spreading. These might act alone, in combination with each other, or together with other factors. Each has been proposed to lead to increased B-cell activation. Impaired tolerance appears to be the central defect and once this has occurred abnormal immunoregulation leads to persistence of the inappropriate self-directed immune response.

Cellular mechanisms also play a role in the development of autoimmunity in the autoimmune rheumatic diseases: T-cell dysfunction, impaired macrophage and natural killer cell cytotoxicity, decreased clearance of immune complexes by the mononuclear phagocytic system, increase in the number of activated B cells, cytokine dysregulation, and upregulation of adhesion molecules have all been reported.

Genetic factors are important, especially in the case of systemic lupus erythematosus, where there is a higher rate of concordance in monozygotic twins (25 per cent) than dizygotic (3 per cent). The best described of the genetic contributions to autoimmune rheumatic disease is the increased risk associated with particular HLA class II molecules. The HLA DR4 (the Dw4 and Dw14 subtypes, notably the DRb1*0404 allele) and HLA DR1 (Dw1) are particularly associated with rheumatoid arthritis. These subtypes share a similarity of the amino acid sequence in the third hypervariable region of the DRb1 chain, the shared epitope that has been proposed as the underlying unit of susceptibility to rheumatoid arthritis. There are, however, conflicting data proposing that this epitope is better related to the severity of disease. In systemic lupus erythematosus, among Caucasians, the haplotype A1 B8 DR3 is associated with an approximately tenfold increase in risk, although the primary link may be with the complement C4 null allele with which there is linkage disequilibrium.

HLA associations are not only seen with autoimmune rheumatic disease, but also with certain autoantibodies. Anti-Ro and La are strongly correlated with HLA DR3 and DQ, an association that is stronger than that seen with the disease in which these autoantibodies are most frequently encountered (systemic lupus erythematosus and Sjögren's syndrome).

Vasculitides

HLA class I and class II associations are seen throughout the primary vasculitides, whilst infectious agents and circulating immune complexes are pathogenic in the secondary vasculitides. In the primary vasculitides a pathogenic role has been proposed for antiendothelial cell antibodies and sensitized T cells, but undoubtedly the most important role is that of ANCA. Immunofluoresence studies have localized the antigen to the cytoplasm of granulocytes in the azurophilic granules, and two patterns of staining are seen: cytoplasmic ANCA (c-ANCA), of which 90 per cent of sera recognize proteinase 3; and perinuclear staining ANCA (p-ANCA) which is directed against myeloperoxidase in 70 per cent of p-ANCA vasculitis patients. A positive c-ANCA is strongly associated with Wegener's granulomatosis, although 10

per cent of these patients may be p-ANCA positive, whilst antimyeloperoxidase antibodies occur in necrotizing glomerulonephritis (65 per cent), Churg–Strauss syndrome (60 per cent), and microscopic polyangiitis (45 per cent).

Clinical features

As mentioned previously, the presentation of an autoimmune rheumatic disease may be diffuse and non-specific, with fatigue and arthralgia frequently the major features. In this instance, systemic review should enquire for the presence of alopecia, mouth ulcers, Raynaud's phenomenon, rash, sicca symptoms, and lymphadenopathy. The presence of these would lend an autoimmune flavour to the illness, but not necessarily help to make a precise diagnosis. The history should also seek a possible trigger such as a preceding infection, drugs (for example hydralazine, isoniazid, procainamide in drug-induced lupus), or environmental exposure to chemicals, as may be seen in scleroderma-like illnesses. A family history must pay particular attention to the presence not only of other autoimmune rheumatic diseases but also other autoimmune diseases such as diabetes, pernicious anaemia, and thyroid disease, which are often found in association with the autoimmune rheumatic diseases.

The protean clinical manifestations mean that an autoimmune rheumatic disease may present not only to a rheumatologist but to many other specialists, including those in nephrology, dermatology, and less commonly neurology, cardiology, haematology, or even obstetrics, in the case of recurrent miscarriages in the antiphospholipid syndrome.

In many cases it is not possible to make a precise diagnosis on the first encounter with a patient. In those with mild disease, symptomatic relief can be obtained with a non-steroidal anti-inflammatory drug, whilst the results of baseline investigations and an 'immunological screen' of antinuclear antibody and rheumatoid factor are awaited. It is worth noting, however, that there is increasing emphasis on trying to make the diagnosis of rheumatoid arthritis promptly, so that a disease-modifying drug can be used as early as possible, rather than waiting for the development of erosive, destructive joint disease.

Since the autoimmune rheumatic diseases are systemic disorders, it is always important to search for evidence of involvement of any of the major organ systems. Baseline investigations must therefore include urinalysis, a full blood count, simple blood tests of renal and liver function, measurement of serum inflammatory markers, an ECG, and a chest radiograph. The simple bedside test of urinalysis is particularly important: the finding of proteinuria and haematuria immediately identifies those who require further, often urgent, renal investigation and whose prognosis may be chiefly determined by the extent of renal involvement.

Damage to major organ systems can be part of the presenting illness in a patient with an autoimmune rheumatic disease, but may also occur in a previously diagnosed patient with 'stable' disease. Myocardial infarction can occur as the result of a vasculitic illness, or accelerated atherosclerosis in systemic lupus erythematosus. Pericarditis can lead to tamponade (for example in systemic lupus erythematosus or rheumatoid arthritis), whilst myocarditis may induce complex arrhythmias or even heart failure (for example in systemic lupus erythematosus or polymyositis). Seizures or a disturbed level of consciousness can occur due to cerebral infarction or meningoencephalitis (for example in systemic lupus erythematosus, antiphospholipid syndrome, Wegener's granulomatosis). Rapidly progressive glomerulonephritis (systemic lupus erythematosus, Wegener's granulomatosis, microscopic polyangiitis) may be associated with pulmonary haemorrhage, whilst hypertension requires urgent treatment in scleroderma renal crisis. Pneumonitis or myositis due to systemic lupus erythematosus may be life threatening if not recognized and treated appropriately with adequate immunosuppression. Venous or arterial thromboses are likely to complicate the antiphospholipid syndrome, which in its primary form may be catastrophic and characterized by widespread microvascular disease with adult respiratory distress syndrome, profound thrombocytopenia, and acute renal failure.

Physicians treating patients with autoimmune rheumatic diseases need to be constantly aware of the possibility of organ involvement: prompt diagnosis and treatment being necessary to prevent irreversible end organ damage. The immunosuppressive therapy used will be similar, regardless of the particular diagnosis.

Precise identification of an autoimmune rheumatic disease is reliant upon clinical and laboratory features, of which the presence of antinuclear antibody (and its pattern of staining), antibodies to extractable nuclear antigens, disease-specific antibodies, or ANCA are crucial. There are many instances where the disease may not be precisely labelled, and up to 20 per cent of patients have features of several autoimmune rheumatic diseases, most commonly systemic lupus erythematosus/scleroderma and systemic lupus erythematosus/rheumatoid arthritis, or those who would be considered to have an undifferentiated autoimmune rheumatic disease. In the case of these latter diseases, treatment is guided according to disease features and the pattern of organ/system involvement.

Further reading

Kallenberg CGM, Heeringa P (1998). Pathogenesis of vasculitis. *Lupus* **7**, 280–4.

Mason LJ, Isenberg DA (1998). Immunopathogenesis of SLE. *Baillière's Clinical Rheumatology* **12**, 385–403.

Menon S, Isenberg DA (1996). Small vessel vasculitides. In: Tooke JE, Lowe GDD, eds. *The textbook of vascular medicine*, pp 295–313. Arnold, London.

Morrow J *et al.* (1999). *Autoimmune rheumatic disease*, 2nd edn.Oxford University Press, Oxford.

Scott DGI, Watts RA (1994). Classification and epidemiology of systemic vasculitis. *British Journal of Rheumatology* **33**, 897–900.

18.10.2 Systemic lupus erythematosus and related disorders

Anisur Rahman and David Isenberg

Introduction

Systemic lupus erythematosus is an autoimmune rheumatic disorder that can present with symptoms in almost any organ or system of the body. Classification criteria have been published by the American College of Rheumatology which should universally be used to make the diagnosis. These are shown in Table 1 and demonstrate the wide variety of clinical and serological features that are associated with this condition. They provide a useful guide to the clinical features that should place the suspicion of systemic lupus erythematosus in the mind of a clinician. It is important not to be too dogmatic in searching for 'pathognomonic' features of the disease. For example, although the characteristic butterfly rash over the face is perhaps the best known sign of systemic lupus erythematosus, many patients will never develop such a rash.

Aetiology

The aetiology is multifactorial, incorporating genetic, hormonal, and environmental elements. The best established genetic link is with the presence of null alleles of genes encoding early components of the complement cascade (C1q, C2, and C4). Over 90 per cent of patients homozygous for C1q deficiency and 75 per cent of those with C4 deficiency develop a lupus-like disease (similar clinical features but a relative paucity of antibodies). Major histocompatibility complex genes, particularly HLA A1, B8, and DR3, have also been associated with the presence of lupus in family studies, although part of this association may be due to linkage disequilibrium with the C4 and C2 genes also present in that region of chromosome 6.

Hormones are likely to play a role in pathogenesis, since systemic lupus erythematosus is far commoner in women than men (see below). There is a relatively high incidence of the condition in Klinefelter's syndrome (males with the XXY karyotype) and this is associated with abnormalities in oestrogen metabolism.

Viruses may be important in triggering the autoimmune dysfunction that leads to the production of pathogenic autoantibodies in systemic lupus erythematosus. Reactivation of BK polyomavirus infection, in particular, has been associated with the presence of antibodies to double-stranded DNA (anti-dsDNA) in Norwegian studies. This association has not yet been confirmed in large populations.

Certain drugs induce a form of systemic lupus erythematosus which is generally characterized by the presence of antihistone rather than anti-dsDNA antibodies, a milder course of disease, and total remission when the causative drug is withdrawn. The most common drugs involved are isoniazid, procainamide, hydralazine, minocycline, penicillamine, and anticonvulsants.

Epidemiology

The incidence of systemic lupus erythematosus in the United Kingdom is about four cases per 100 000 people per year. The prevalence varies between the sexes and between different ethnic groups. Systemic lupus erythematosus occurs between 10 and 20 times more frequently in women than in men and is commoner in some ethnic groups. A recent study in Birmingham, United Kingdom gave the prevalence of systemic lupus erythematosus in women as 206 per 100 000 in Afro-Caribbeans, 91 per 100 000 in Asians, and 36 per 100 000 in Caucasians. These gender and racial differences are broadly consistent with those reported from studies in the United States and the Caribbean, although the reported prevalence of systemic lupus erythematosus in Africa is much lower.

Mortality from systemic lupus erythematosus has fallen significantly over the last half century. Whereas systemic lupus erythematosus was reported to have a 50 per cent 5-year survival in the 1950s, 10-year survival rates rose to between 80 and 90 per cent by the 1970s. Since then, survival rates have improved a little, but deaths from renal failure have become less common, whilst those from infection have increased. The latter are generally associated with immunosuppressive therapy, highlighting the need for better and more accurately targeted methods of treating the underlying immunological abnormalities in this disease.

Pathogenesis and immune dysfunction

No single abnormality of the immune system can be considered to be the sole cause of systemic lupus erythematosus. The pathogenesis of the disease depends on the interplay of a number of different factors, the relative importance of which may differ from one patient to another. These include autoantibodies, T lymphocytes, cytokines, the complement system, and apoptosis. Research to unravel this complex system of interrelated factors has been carried out by studying properties of cells and tissue components derived from patients with systemic lupus erythematosus and by studying mouse models of the condition.

B lymphocytes and autoantibodies

Autoantibodies are those which bind to antigens present within the tissues of the body itself. A wide variety of different autoantibodies have been described in systemic lupus erythematosus. Those most frequently reported are listed in Table 2.

Antibodies to double-stranded DNA (anti-dsDNA) have been cited widely as possible causative agents in systemic lupus erythematosus, particularly in lupus glomerulonephritis. Raised titres of anti-dsDNA antibodies are found in 50 to 70 per cent of patients with systemic lupus erythematosus but hardly ever in healthy

people or those with other diseases. Levels of these antibodies rise and fall with disease activity in systemic lupus erythematosus, and deposits of anti-dsDNA occur in the glomeruli of patients with lupus nephritis. In experimental murine models of systemic lupus erythematosus, monoclonal anti-dsDNA antibodies can also be shown to deposit in the glomeruli with associated proteinuria.

The titre of anti-dsDNA antibodies present in the bloodstream of patients with systemic lupus erythematosus can be a useful indicator of disease activity. It is increasingly clear, however, that not all anti-dsDNA antibodies are equally likely to be associated with tissue damage. Antibodies of IgG isotype, which show specific, high-affinity binding to dsDNA generally show the closest association with disease activity in patients and the greatest ability to cause renal damage in experimental models.

Why are such antibodies produced in patients with systemic lupus erythematosus? Studies of monoclonal anti-dsDNA antibodies derived from patients or mice indicate that those which show the isotype and binding properties described above often show sequence characteristics suggestive of antigen-driven somatic mutation. This is the process whereby mutations accumulate in the expressed immunoglobulin gene sequences of a B lymphocyte under the influence of a particular antigen. The mutations are accumulated non-randomly, such that the end result is an increase in specificity and affinity of binding. This process is dependent on help from T lymphocytes and on the presence of an appropriate antigen. Naked mammalian DNA, however, is a poor immunogen in experimental animals, and the concentration of free DNA in the bloodstream is low even in patients with systemic lupus erythematosus. It is therefore believed that the antigen which stimulates production of high-affinity anti-dsDNA antibodies is probably a complex of DNA and protein. Nucleosomes derived from cell apoptosis may be the most important antigens involved, although a role for viral DNA binding proteins has also been suggested.

How do the autoantibodies exert their pathogenic effects? Deposition of IgG and complement in inflamed tissues such as kidney and skin is a consistent feature of active systemic lupus erythematosus. The pathogenic potential of autoantibodies in systemic lupus erythematosus (particularly IgG anti-dsDNA) may therefore rest upon their ability to deposit in these tissues and to activate complement. However, the role of complement activation has recently been called into question by work showing that knockout mice deficient in both the classical and alternative pathways of the complement cascade still develop a form of glomerulonephritis similar to that seen in systemic lupus erythematosus, but since these mice do not typically produce anti-dsDNA antibodies they may not be an appropriate model of most patients with systemic lupus erythematosus.

Why are anti-dsDNA antibodies deposited in target tissues? Much of the work designed to answer this question has concentrated on autoantibodies in lupus nephritis. Originally, it was felt that DNA–anti-DNA immune complexes would form in the bloodstream and accumulate in glomeruli as the blood was filtered there. It has not been possible to demonstrate large quantities of such complexes in the blood of patients with systemic lupus erythematosus, though their clearance may well be abnormal due to complement deficiency. Anti-dsDNA antibodies may be targeted to the kidney due to cross-reaction with cell surface proteins there, or may deposit due to an interaction with histones and heparan sulphate. According to this latter model, anti-dsDNA antibodies bind to DNA in nucleosomes, and the positively charged histones in these nucleosomes bind to negatively charged heparan sulphate in the renal basement membrane.

Antiphospholipid antibodies

Between 20 and 30 per cent of patients with systemic lupus erythematosus possess serum antiphospholipid antibodies. The origin of these antibodies may be similar to that of anti-dsDNA antibodies since monoclonal antiphospholipid antibodies from patients with systemic lupus erythematosus also show antigen-driven accumulations of somatic mutations. The antigen in this case may be phosphatidylserine on the outer surfaces of blebs derived from apoptotic cells.

The pathological effects of antiphospholipid antibodies are not due to deposition and complement activation but to promotion of thrombus formation. This leads to arterial and venous thromboses that may be particularly harmful in the cerebral and renal circulation. The mechanism by which thrombosis is altered is not fully understood, but it has become clear that antiphospholipid antibodies found in systemic lupus erythematosus and the primary antiphospholipid antibody syndrome recognize a complex of negatively charged phospholipid with the plasma protein b2-glycoprotein 1. Antiphospholipid antibodies found in infectious diseases such as syphilis bind to phospholipids in the absence of this cofactor, and are not associated with increased thrombosis or adverse clinical effects.

T lymphocytes

Since the process of antigen-driven selection of mutations in B lymphocytes is dependent on help from helper T lymphocytes, it would be reasonable to suppose that antigen-specific T cells might also contribute to the pathogenesis of the disease. The isolation of T-cell clones reactive with DNA and/or DNA binding proteins such as histones has been demonstrated from both patients with systemic lupus erythematosus and murine models of the disease. The clones frequently show specificity for histone epitopes that are cryptic (i.e. not exposed) in normal chromatin. These results reinforce the idea that the antigenic stimulus for production of pathogenic T cells and autoantibodies in systemic lupus erythematosus may be a DNA/histone complex rather than DNA alone.

Patients with systemic lupus erythematosus have decreased levels of the subset of T cells carrying the CD4 and CD45Ro surface markers. This population may be involved in stimulation of suppressor T lymphocytes, so that suppression in these patients is insufficient to prevent the production and survival of autoreactive B-lymphocyte and helper T-lymphocyte clones.

Apoptosis and complement

MRL *lpr/lpr* mice are deficient in apoptosis because they lack the Fas protein which plays a major role in promoting this process. These mice develop a disease very similar to systemic lupus erythematosus with death resulting from glomerulonephritis. One possible reason for this might be the failure of the immune system to delete by apoptosis autoreactive clones of T or B lymphocytes which are then able to cause autoimmune disease. By contrast, humans with the equivalent genetic lesion to MRL *lpr/lpr* mice do not develop systemic lupus erythematosus, and other strains of mice show an accumulation of apoptotic debris within nephritic kidneys which resemble those of systemic lupus erythematosus. A simple deficiency in apoptosis is therefore unlikely to be the underlying mechanism in systemic lupus erythematosus.

Apoptosis leads to the production of surface blebs of cellular material. These blebs include a number of the antigens to which autoantibodies develop in systemic lupus erythematosus, notably DNA and associated nuclear proteins and negatively charged phospholipids. A deficiency in the clearance of products of apoptosis has been demonstrated which might allow the production of a wide spectrum of autoantibodies, as found in systemic lupus erythematosus. Removal of immune complexes containing such potentially antigenic material may be compromised in patients with systemic lupus erythematosus. Monocytes derived from such patients show reduced phagocytosis of cell debris *in vitro*. This process may be complement dependent. Humans with homozygous C2 deficiency process immune complexes very differently from normal controls. Administration of fresh frozen plasma to these patients as a source of complement is successful in ameliorating the symptoms of systemic lupus erythematosus and in normalizing (albeit transiently) the processing of immune complexes.

C1q knockout mice develop a form of glomerulonephritis similar to that seen in systemic lupus erythematosus, and their kidneys are characterized by accumulations of apoptotic debris. Similarly, as noted earlier, humans homozygous for C1q deficiency develop a form of systemic lupus erythematosus with the frequent occurrence of nephritis.

The role of cytokines

Cytokines enhance the ability of cells to interact and are therefore critically important in abnormalities in both T- and B-cell functions seen in patients with lupus. Table 3 summarizes the major differences between the different subsets of T helper (T_H) cells in terms of their cytokine profiles and functions. The balance between cytokines from the T_{H1} and T_{H2} cells is essential in determining the outcome of the immune response. Lupus might be expected to be a disease in which T_{H2} cells predominate, resulting in excessive help for B cells and overproduction of antibodies. In support of this notion, increased levels of IL-10 have been found in patients with lupus. This cytokine suppresses T_{H1} cells and thus impairs cell-mediated immunity, a characteristic feature of the disease. Both macrophage and natural killer cell-mediated cytotoxicity are frequently impaired in patients with lupus. g-interferon-induced enhancement of both types of cytotoxicity is also impaired, despite normal levels of g-interferon production by lupus T_{H1} cells.

Accessory cells in lupus seem to produce insufficient amounts of IL-1 to provide the necessary activation signals for T cells. Both CD4+ and CD8+ T cells have been described as producing either normal or decreased amounts of IL-2 in response to exogenous antigens. Such a reduction is likely to have a profound effect on T-cell

responses.

Clinical features of systemic lupus erythematosus

Systemic lupus erythematosus is a chronic condition in which a low level of baseline activity is punctuated by flares of higher activity. The overall severity of the disease in a particular patient depends on the nature and frequency of these flares.

The diverse clinical features of systemic lupus erythematosus mean that the disease may present to any of a number of different specialists, including rheumatologists, dermatologists, nephrologists, and general physicians. It is important to be aware of systemic lupus erythematosus as a possible diagnosis in any patient, especially a woman aged between 15 and 50, in whom a number of different organs are inflamed either simultaneously or sequentially. The frequency of occurrence of symptoms in various organs is shown in Table 4.

According to the diagnostic guidelines published by the American College of Rheumatology (Table 1) systemic lupus erythematosus may be diagnosed where a patient meets at least four of the 11 criteria specified (though not necessarily at a single time). In everyday practice, however, these requirements may be too stringent, and systemic lupus erythematosus is often diagnosed on the basis of typical clinical findings in one organ or tissue combined with the presence of appropriate autoantibodies.

Constitutional symptoms

Patients with systemic lupus erythematosus find fatigue to be the most troublesome feature of the disease. Excessive tiredness is both very common and difficult to treat. Hypothyroidism coexists in 5 to 10 per cent of patients with systemic lupus erythematosus and so thyroid function tests should be performed in the fatigued patient. There is an ongoing debate as to whether fibromyalgia is a significant comorbid condition.

Weight loss and low-grade fever may both be indicative of disease activity. Lymphadenopathy is also recognized. The nodes may be markedly enlarged but show no diagnostic features on biopsy, which may nevertheless be necessary to exclude other conditions such as lymphoma.

Musculoskeletal involvement

Arthralgia is the commonest symptom in systemic lupus erythematosus, occurring in 90 per cent of patients. This may be severe but is rarely associated with frank synovitis. Effusions may occur but the fluid shows no diagnostic features.

Erosive arthritis is uncommon, though up to 5 per cent of patients may have an overlap syndrome with features of rheumatoid arthritis as well as systemic lupus erythematosus. These patients tend to have both serum rheumatoid factor and erosions.

When progressive deformity of the hands does occur in systemic lupus erythematosus, it is usually due to an aggressive tenosynovitis and tendon dysfunction rather than to joint damage. This leads to reversible subluxation of the joints, often known as Jaccoud's arthropathy (Fig. 1 and Plate 1).

Fig. 1 Deforming Jaccoud's arthropathy. (See also Plate 1.)

Development of hip pain in patients who have been treated with corticosteroids should raise the suspicion of avascular necrosis of the femoral head, which may be diagnosed on plain radiograph or, in earlier stages, by magnetic resonance imaging. Corticosteroids also promote osteoporosis, which can be diagnosed in the presymptomatic phase by bone density scanning, but which may present with the acute pain of a vertebral fracture.

Myalgia is common and a true myositis may occur in 5 per cent of cases. Corticosteroid-induced proximal myopathy may also be a problem where these drugs have been used for long periods.

Cutaneous and mucosal involvement

Photosensitivity is very common, particularly in white female patients. Patients should be advised to avoid strong sunlight and to wear protective clothing and/or a high-factor sunblock.

The butterfly rash over the malar area of the face occurs in up to one-third of patients (Fig. 2 and Plate 2). A number of other forms of cutaneous involvement can occur, although these are less specific for systemic lupus erythematosus. These include maculopapular rash, discoid lesions, alopecia, and nailfold infarcts. Scarring alopecia may be particularly distressing and difficult to treat (Fig. 3 and Plate 3).

Fig. 2 Malar 'butterfly' rash. (See also Plate 2.)

Fig. 3 Severe scarring alopecia. (See also Plate 3.)

A variant of systemic lupus erythematosus in which cutaneous manifestations dominate is known as subacute cutaneous lupus. This condition is often associated with anti-Ro antibodies and may be exacerbated by smoking cigarettes.

Antiphospholipid antibodies are associated with a non-raised lattice-like rash concentrated particularly over the thighs and arms. This rash is termed livedo reticularis (Fig. 4 and Plate 4).

Fig. 4 Livedo reticularis. (See also Plate 4.)

Painless oral ulcers are common enough to be recognized as one of the diagnostic criteria for systemic lupus erythematosus. However, they are rarely troublesome for the patient. Approximately 20 per cent of patients develop secondary Sjögren's syndrome. The dry eyes and mouth in this condition may respond to artificial tears and saliva.

Renal involvement

Glomerulonephritis is the most serious and potentially lethal manifestation of systemic lupus erythematosus. Its presence may be detected by the finding of haematuria and/or proteinuria on routine stick testing of the urine. It may present as the nephrotic syndrome, or less commonly as a florid nephritis with haematuria, proteinuria, hypertension, and acute renal failure with red cell casts in the urine. The diagnosis and management of glomerulonephritis in systemic lupus erythematosus are more fully discussed in Chapter 20.10.4.

It is important to be aware of the possibility of glomerulonephritis in any patient with systemic lupus erythematosus (Fig. 5 and Plate 5). Measurement of blood pressure and analysis of urine should be carried out at each consultation. Early diagnosis and treatment are invaluable in avoiding deterioration of renal function to the extent that dialysis or renal transplantation become necessary.

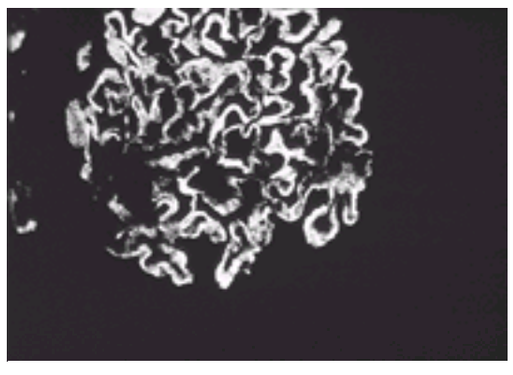

Fig. 5 Immunofluorescence microscopy showing deposition of IgG in the glomerulus of a patient with lupus nephritis. (See also Plate 5.)

Patients with antiphospholipid antibody syndrome may develop a different type of renal lesion characterized by thrombi in small renal vessels rather than by glomerulonephritis. These patients develop hypertension and impairment of renal excretory function, detected as a fall in creatinine clearance, rather than proteinuria and are best managed by anticoagulation rather than immunosuppression.

Respiratory involvement

The commonest form of respiratory involvement in systemic lupus erythematosus is pleuritis, manifesting either as pleuritic chest pain or breathlessness caused by pleural effusion. The lung parenchyma is more rarely involved, but fibrosis can occur.

A patient with systemic lupus erythematosus may present with shortness of breath or chest pain for a number of reasons. Pulmonary emboli must be suspected in those with antiphospholipid antibodies. Infections are common in immunosuppressed patients and rib fractures may occur, particularly in those rendered osteoporotic by treatment with corticosteroids.

The shrinking lung syndrome is characterized by reduced lung volumes and poor respiratory reserve in the face of a normal appearance of the lung parenchyma on computed tomography scanning. It is believed to arise from basal atelectasis in association with diaphragmatic dysfunction.

Cardiovascular involvement

The commonest cardiac manifestation of systemic lupus erythematosus is pericarditis, which occurs in approximately 15 per cent of patients. This generally presents

with chest pain or an asymptomatic friction rub. Pericardial effusions may occur but are rarely large enough to cause haemodynamic compromise.

Myocarditis and endocarditis are less common, though post-mortem and echocardiographic studies suggest that both may occur without symptoms in a significant proportion of patients with systemic lupus erythematosus. For example, the classic endocarditis described by Libman and Sacks is characterized by small vegetations that often do not cause murmurs or cardiac compromise, but which have been identified in up to 50 per cent of patients with systemic lupus erythematosus at autopsy.

It is increasingly recognized that atherosclerosis and the attendant deficiencies of cerebral and cardiac circulation are commoner in patients with systemic lupus erythematosus than in the general population. This may be partially due to the use of corticosteroids, which raise serum cholesterol and can promote hypertension. Patients possessing antiphospholipid antibodies are also at a higher risk of stroke or arterial thrombosis.

Raynaud's phenomenon occurs in about one-third of patients with systemic lupus erythematosus, though it is not usually as severe as that seen in systemic sclerosis. Vasculitis presents with a skin rash or ulcers that may be very difficult to heal, but rarely affects the internal organs.

Gastrointestinal involvement

Like pleuritis, peritonitis may occur in patients with systemic lupus erythematosus and must be considered in the event of abdominal pain.

Involvement of the liver and pancreas is recognized but uncommon. The term 'lupoid hepatitis' was previously used for a form of autoimmune hepatitis characterized by the presence of autoantibodies. These patients, however, do not generally have any form of systemic lupus erythematosus and the term is misleading. Minor enlargements of the liver and/or spleen occur in 10 to 25 per cent of cases but these are usually asymptomatic and require no treatment.

H4>Neuropsychiatric involvement

Systemic lupus erythematosus can affect the central nervous system in many ways, so that the true incidence of this type of involvement is difficult to quantify. Symptoms such as poor memory, change of personality, and depression or anxiety occur in many patients. It is difficult, however, to be sure whether these are caused by cerebral systemic lupus erythematosus or represent a reaction to the diagnosis and treatment of the disease.

More florid presentations such as psychotic episodes and convulsions are well recognized. By contrast to the milder symptoms noted above, these manifestations generally call for immunosuppression.

Migraine occurs in up to 40 per cent of patients with systemic lupus erythematosus, particularly in the presence of antiphospholipid antibodies. Peripheral neuropathy can occur, and is usually sensory rather than motor. Cranial nerve palsies are less common, as is transverse myelitis (another feature linked to antiphospholipid antibodies).

Ocular involvement can include episcleritis, conjunctivitis, and the presence of cytoid bodies (white patches on the retina). Patients treated with high-dose steroids may develop cataracts.

Haematological involvement

A normochromic normocytic anaemia is frequently seen in systemic lupus erythematosus, particularly during periods of high disease activity. Microcytic iron-deficiency anaemia may be present due to blood loss from gastritis and ulcers in patients treated with non-steroidal anti-inflammatory drugs. Anaemia may also result from chronic renal failure in lupus nephritis.

A positive Coombs' test, signifying the presence of antibodies to red blood cells, is present in 10 to 15 per cent of patients with systemic lupus erythematosus but does not always indicate haemolytic anaemia.

The presence of lymphopenia (less than 1.5×10^9 per litre) is a common feature, occurring in up to 80 per cent of patients. Neutropenia may occur secondary to the use of cytotoxic drugs such as azathioprine or cyclophosphamide.

Three different types of thrombocytopenia occur in systemic lupus erythematosus. The mildest form is characterized by stable platelet levels of between 50 and 100×10^9 per litre, is rarely symptomatic, and usually requires no treatment. Other patients develop an acute autoimmune thrombocytopenia with levels dropping rapidly below 10×10^9 per litre, but rising when treated with oral steroids. A third group of patients present with thrombocytopenia alone, are treated with steroids, intravenous immunoglobulins, or splenectomy and some years later develop full-blown systemic lupus erythematosus. Thrombocytopenia is also one of the cardinal features of the antiphospholipid antibody syndrome and may be severe enough to necessitate splenectomy in that condition.

Other complicating disorders

Approximately 30 per cent of lupus patients have another autoimmune condition. Sjögren's syndrome is the commonest of these, being present in some 15 to 20 per cent of patients with systemic lupus erythematosus. In the past 15 years it has been recognized that clinical features of the antiphospholipid antibody syndrome, notably venous and arterial thromboses, recurrent miscarriages, thrombocytopenia, and livedo reticularis, may complicate 10 to 15 per cent of patients with systemic lupus erythematosus. When these features, together with the presence of antiphospholipid antibodies (usually cardiolipin, anti-b2-glycoprotein 1 or the lupus anticoagulant) occur in the presence of other more classical lupus features, the condition is known as secondary antiphospholipid syndrome, but they can occur on their own, in which case the patient is said to have primary antiphospholipid syndrome as described elsewhere. Autoimmune thyroid disease (hyper- or hypothyroidism) occurs in 5 to 10 per cent of the patients, and rather less frequently, rheumatoid arthritis, myasthenia gravis, coeliac disease, diabetes, and pernicious anaemia may be found.

Investigations and pathology

Autoantibodies

The most commonly requested test to screen for systemic lupus erythematosus is the antinuclear antibody assay. A positive antinuclear antibody simply indicates that the patient's blood contains antibodies which will bind to the nuclei of a sample of cells used in the test. The test is a sensitive one since over 95 per cent of patients with systemic lupus erythematosus are antinuclear antibody positive. Although a small group of patients do seem to have persistently antinuclear antibody negative systemic lupus erythematosus, the absence of antinuclear antibody in a patient with suspected lupus raises serious doubt about the diagnosis.

The specificity of the antinuclear antibody test for systemic lupus erythematosus is not high. The titre of antibody represents the highest dilution of the patient's serum at which the test is still positive. Low-titre antinuclear antibody (1 in 10) is of little significance and may occur in healthy people. Higher titres (1 in 160 or more) are more worrying and are found in most patients with systemic lupus erythematosus and in a few patients with other autoimmune conditions including rheumatoid arthritis, systemic sclerosis, and Sjögren's syndrome. However, some people with high-titre antinuclear antibody may be followed in rheumatology clinics for years without developing a frank autoimmune disease.

The finding of a positive antinuclear antibody in a patient with symptoms suggestive of systemic lupus erythematosus should lead to a series of other autoantibody tests. These are listed in Table 2 together with the identity of the target antigen and the approximate prevalence of the antibodies.

Anti-dsDNA antibody levels are particularly useful. This test is virtually specific for systemic lupus erythematosus (as is the anti-Sm antibody), especially if the immunoglobulins are of the IgG isotype. The anti-dsDNA result is usually quantified and this value is a measure of the activity of the disease. Indeed, in one study, trial patients were treated with high-dose corticosteroids on the basis of anti-dsDNA levels alone. In comparison with a control group treated only when symptoms or signs also suggested disease activity, the trial group had less disease activity overall and fewer flares. However, frequent large doses of corticosteroids resulted in significant side-effects and a number of subjects dropped out of this arm of the trial. The current evidence therefore suggests that anti-dsDNA should be used only as

an adjunct to the clinical impression of disease activity when deciding on a treatment regimen.

Anti-Ro and anti-La antibodies are linked to concurrent Sjögren's syndrome. Mothers who have these antibodies have a higher incidence of neonatal lupus (see below) and should be advised about this before embarking upon a pregnancy. Anti-Ro antibodies are also associated with photosensitivity.

There are no good antibody markers for the presence of disease of the central nervous system. Antibodies to ribosomal protein P were previously thought to have some value in the diagnosis of central nervous system lupus, but this has not been borne out by later results and the test is not available routinely in most laboratories.

Antiphospholipid antibodies can be recognized by one of two assays. The enzyme-linked immunosorbent assay for binding to cardiolipin distinguishes IgM and IgG isotypes. This is helpful because the level of IgG antiphospholipid antibodies is a better predictor of clinical sequelae than that of IgM. Antiphospholipid antibodies can also be diagnosed by testing the clotting properties of the blood *in vitro* in the Russell's viper venom test. An abnormal result in this assay is reported as showing the presence of a lupus anticoagulant. It is quite possible for the anticardiolipin test to be positive while the lupus anticoagulant assay is negative or vice versa. If either is positive, the patient may be at risk of manifestations of the antiphospholipid antibody syndrome. Antibodies to the phospholipid cofactor, known as b2-glycoprotein 1, are now becoming available commercially, offering a third way of detecting antiphospholipid antibodies.

Coombs' test and assays for antithyroid antibodies are often requested in patients with systemic lupus erythematosus, particularly those with coexisting anaemia or hypothyroidism.

Measures of disease activity and end-organ damage

Blood and urine tests

The three most reliable measures of highly active disease are high erythrocyte sedimentation rate, depletion of complement, and high anti-dsDNA levels. The erythrocyte sedimentation rate increases much more than the level of C-reactive protein in active systemic lupus erythematosus. The combination of high erythrocyte sedimentation rate and normal C-reactive protein in a patient with a multisystem disorder should raise the suspicion of systemic lupus erythematosus, leading to appropriate autoantibody tests as described above. The C-reactive protein may, however, be raised in the presence of infection, serositis, or arthritis.

Complement components C3 and C4 are the most commonly measured, and both tend to fall in active systemic lupus erythematosus. A persistently very low level of either C3 or C4 (or a high level of their degradation products C3d or C4d), regardless of immunosuppressive therapy, may signify the presence of a homozygous complement deficiency disorder. Though such disorders are very rare, it is important to diagnose them because they respond better to infusions of fresh frozen plasma than to immunosuppression.

It is important to measure creatinine and electrolyte values regularly and to check the urine for proteinuria and/or haematuria. These measures ensure that renal involvement is diagnosed early. It must be remembered that substantial deterioration in renal function may occur before serum creatinine rises beyond the normal range. It is therefore prudent to note even relatively small rises in creatinine if these are persistent. Renal function may be measured more accurately by creatinine clearance using a 24-h collection of urine or by measuring the clearance of a radio-isotope to obtain the glomerular filtration rate. Persistent proteinuria on bedside testing can be investigated further by measuring the total protein in a 24-h urine sample or the albumin to creatinine ratio in a spot sample.

Liver function tests are not usually abnormal in systemic lupus erythematosus (abnormal in less than 10 per cent of patients), but a baseline value should be measured, particularly in cases where potentially hepatotoxic drugs such as azathioprine may be used. Thyroid function abnormalities, particularly hypothyroidism, are well recognized to coexist with systemic lupus erythematosus.

A full blood count should be measured regularly. Falling haemoglobin, white cell count, and platelet counts may all occur (see under haematological involvement above). Anaemia in the presence of a positive Coombs' test may indicate haemolysis which can be confirmed by requesting a blood film and serum haptoglobins.

Infections occur commonly in patients with systemic lupus erythematosus, particularly in those on high-dose immunosuppressants. Infection may not always be accompanied by high fever or leucocytosis, although C-reactive protein is usually raised. It is wise to carry out blood and urine cultures whenever even mild pyrexia is accompanied by a deterioration in health.

Imaging

Plain radiographs are rarely useful in systemic lupus erythematosus. There is no characteristic appearance in the joints and chest radiographs are unlikely to show abnormalities except in the presence of infection or effusion.

Requests for more specialized imaging studies should be directed by the clinical findings. For example, the presence of dyspnoea and abnormal respiratory function tests often necessitates a computed tomography scan of the thorax, which is the investigation of choice for diagnosis of pulmonary fibrosis. Echocardiography is useful if pericardial effusion, myocarditis, or endocarditis are suspected clinically. Bone density scanning is becoming increasingly important, since patients with systemic lupus erythematosus are often at risk of osteoporosis due to use of corticosteroids and reduced capacity for physical exercise during young adult life.

Histopathology

The two tissues most often subjected to biopsy in systemic lupus erythematosus are the skin and kidneys.

Skin biopsies are chiefly carried out to facilitate the diagnosis of an atypical rash. If systemic lupus erythematosus is suspected, it is important to take a sample of apparently normal skin as well as skin from the rash. Both should show deposition of IgG and complement at the dermoepidermal junction (Fig. 6 and Plate 6).

Fig. 6 Immunofluorescence microscopy showing deposition of IgG at the dermoepidermal junction in the skin of a patient with systemic lupus erythematosus (sometimes called the lupus band test). (See also Plate 6.)

There are two main indications for renal biopsy in the patient who has, or might have, systemic lupus erythematosus. Firstly, to establish the diagnosis when this is not certain, for example in a patient with poorly characterized multisystem disease with renal involvement. Secondly, to help determine prognosis and decide upon treatment in the patient known to have systemic lupus erythematosus, but with deterioration in renal function, for example development of nephrotic syndrome and/or declining renal excretory function. The activity of glomerular inflammation is graded on a scale of I to V according to World Health Organization criteria. The biopsy can also be used to grade the degree of chronicity of glomerular disease, i.e. how much irreversible damage such as fibrosis and atrophy has occurred. The activity

and chronicity scores can both be used to determine appropriate treatment and the risk of a progressive decline in renal function, although the predictive value of such data remains controversial. The subject of renal pathology in systemic lupus erythematosus is considered further in Chapter 20.10.4.

Treatment and prognosis

Systemic lupus erythematosus is a disease that still has the potential to kill young people. In many cases, however, the condition runs a fairly indolent course in which an initial flare is followed by many years of low-grade activity. General measures of value in the treatment of systemic lupus erythematosus are shown in Table 5.

In the pharmacological management of a patient with systemic lupus erythematosus, the clinician will typically seek to answer four questions:

1. Can the patient be managed without immunosuppression?
2. If immunosuppression is needed, how should it best be started?
3. If immunosuppression is being used, is the current level of immunosuppression inadequate or excessive? How should it be increased or reduced?
4. Is the patient suffering side-effects from the drugs?

Is immunosuppression required?

Patients whose disease activity is confined to arthralgia, tiredness, and/or mild rash do not usually have greatly raised erythrocyte sedimentation rate or anti-dsDNA antibodies or reduced complement. These patients can often be treated symptomatically, for example with agents such as paracetamol and diclofenac to control joint pain.

Where such symptoms are more severe, the antimalarial agent hydroxychloroquine at a starting dose of 400 mg per day may be useful. This drug has less potential for retinal toxicity than the closely related chloroquine and is therefore preferred in systemic lupus erythematosus. It is often possible to reduce the dose to 200 mg per day after 3 months and gradually withdraw the drug thereafter. Regular blood tests are not required to monitor the effects of hydroxychloroquine, but there is a very small risk of retinopathy such that review by an ophthalmologist every 6 to 12 months is considered advisable in many units.

Where the main symptoms in a patient with systemic lupus erythematosus are those of the antiphospholipid antibody syndrome immunosuppression is rarely useful. Aspirin at a dose of 150 to 300 mg daily is recommended for those with mild symptoms of the disease or who have other risk factors for thrombosis. Patients who have suffered recurrent thromboses or cerebral infarcts and who have serum antiphospholipid antibodies should usually be treated with lifelong anticoagulation. This is a major commitment for a young patient and raises particular problems in pregnancy (discussed below).

Some patients require a low maintenance dose of oral steroids to control their symptoms even though laboratory indices do not indicate high activity of disease. A dose of 5 to 7.5 mg daily is typically used in such cases. Topical steroids may be useful where lupus activity is confined to the skin.

Is the current level of immunosuppression inadequate or excessive?

Corticosteroids and cytotoxic agents are used to treat flares of disease. A mild flare of arthralgia, myalgia, and general fatigue may be alleviated by a single intramuscular dose of a corticosteroid preparation such as prednisolone acetate (usually 50 to 125 mg are given).

More severe flares of arthritis, pleuritis, or pericarditis require oral prednisolone at a dose of 20 to 40 mg daily. This usually leads to a rapid improvement in symptoms and the dose of prednisolone can be reduced by 5 mg every 1 to 2 weeks until it reaches 5 mg per day. It may not be possible to withdraw the drug completely for several months.

Alternatively, a shorter course of corticosteroids can be given intravenously. A typical course would consist of 750 mg to 1 g of methylprednisolone given over 3 to 4 h on each of three successive days. This requires admission to hospital, making it less convenient than oral therapy, and it is generally reserved for those patients who are not responding to oral prednisolone or cannot tolerate that drug in high doses.

Autoimmune haemolytic anaemia requires higher doses (60–80 mg/day) of oral prednisolone, with the dose reduced in 5 to 10 mg increments according to the clinical response. Azathioprine may be required as a steroid-sparing agent and is used at a dose of 2.5 to 3 mg/kg/day.

Renal flares of systemic lupus erythematosus require the most aggressive treatment, generally involving both corticosteroids and cyclophosphamide. A number of regimes have been used and a debate continues as to which is optimal. One of the most common regimes is that recommended by the United States National Institutes of Health, in which the patient is given oral prednisolone at a dose of between 30 and 80 mg/day depending on the severity of the disease. Intravenous boluses of 750 mg to 1 g cyclophosphamide are given at monthly intervals for 6 months, then every 3 months for 2 years. Cyclophosphamide pulses should be accompanied by adequate intravenous hydration and the use of mesna (mercaptoethane sulphonate) to reduce bladder toxicity. It has been suggested that the use of intravenous pulse therapy is preferable to treatment with oral cyclophosphamide on the grounds of improved compliance and less gonadal dysfunction. The latter claim has not been proved beyond doubt and some groups use oral cyclophosphamide (2–4 mg/kg/day) or oral azathioprine (2–3 mg/kg/day) per day as an alternative to intravenous pulses.

In renal systemic lupus erythematosus it is critically important to control the patient's blood pressure. ACE inhibitors, alpha-adrenergic antagonists such as doxazosin, and calcium channel blockers such as nifedipine are the agents most commonly used.

The treatment of central nervous system lupus varies depending on the manifestation of cerebral dysfunction. Mild cases may respond to relatively small doses of oral steroids (up to 30 mg per day). More florid manifestations such as convulsions or major psychosis require treatment with appropriate anticonvulsants or antipsychotic drugs, higher-dose oral steroids (60–80 mg per day), and sometimes azathioprine or intravenous pulses of cyclophosphamide in similar doses to those used in renal systemic lupus erythematosus.

Is the patient suffering side-effects from the drugs?

The side-effects of corticosteroids are well known. The most common early problems are weight gain, hirsutism, easy bruising, and insomnia. It is difficult to prevent them, except by using the lowest dose of steroid that is effective and reducing it as rapidly as possible whilst maintaining control of the disease.

Longer-term sequelae of corticosteroid use include increased susceptibility to infection, osteoporosis, avascular necrosis, and diabetes mellitus. The most rapid loss of bone in steroid-induced osteoporosis occurs within the first year of treatment, though doses of 7.5 mg/day or less of prednisolone are thought to have little effect on bone. At higher doses, it may be advisable to carry out a bone density scan and to give either calcium and vitamin D tablets or a bisphosphonate (either etidronate or alendronate are commonly used) as prophylaxis.

Cyclophosphamide causes alopecia, nausea, bladder toxicity, and gonadal dysfunction that may lead to infertility. The problem of infertility becomes more likely with increasing age. Women over 30 given cyclophosphamide are at particular risk. Again, the best way to prevent such problems is to use as small a cumulative dose of the drug as is feasible. Bone marrow suppression may occur. During a programme of cyclophosphamide pulses, the white blood cell count falls to a nadir 10 days after each pulse and should be measured at that time to decide whether the next pulse can be given safely. Nausea and vomiting during pulses may be so severe that antiemetics such as metoclopramide or granisetron are necessary.

Azathioprine also causes bone marrow suppression and can cause abnormalities of liver enzymes which resolve once the drug is withdrawn.

Systemic lupus erythematosus in pregnancy

Systemic lupus erythematosus itself does not usually reduce the ability to conceive, although as described the drugs used to treat it, notably cyclophosphamide, may induce infertility due to gonadal failure. There is an increased risk of spontaneous abortion, particularly in the presence of high-titre antiphospholipid antibodies. Pregnant mothers with a high antiphospholipid antibody level and a history of previous miscarriage should be considered for anticoagulation until the birth of the baby.

Since warfarin is potentially teratogenic, heparin may be used from the second trimester until parturition.

Mothers often ask whether their children are likely to inherit systemic lupus erythematosus. Inheritance of the adult form of the disease is very rare (approximately 1 per cent of all cases), though a transient illness termed neonatal lupus can occur. The characteristics of this condition are rash, hepatitis, anaemia, and thrombocytopenia which usually resolve by 8 months after birth, and inflammation of the cardiac conducting tissues which may lead to heart block in the fetus. The cardiac problem may be diagnosed by ultrasound scans of the fetal heart between 16 and 24 weeks' gestation. Treatment of the mother with 4 mg oral dexamethasone per day may prevent progression from incomplete to complete fetal heart block. If complete heart block occurs, the neonate may require a cardiac pacemaker. Interestingly, children born with neonatal lupus sometimes develop heart block later in life. In one reported case this problem occurred at the age of 35.

The presence of maternal anti-Ro and anti-La antibodies predicts a higher risk of neonatal lupus. Where both are present the risk is approximately 5 per cent. It is believed that the antibodies cross the placenta and bind to some component of the fetal cardiac tissue (laminin is a particular 'suspect'). The mechanism whereby this leads to heart block is mysterious, especially as the mother's heart is never affected.

Although overall the risk of a flare during pregnancy is probably no greater than at other times, systemic lupus erythematosus may exacerbate during the pregnancy. Corticosteroids may be used in moderate doses without affecting the fetus, but higher doses (over 30 mg) given for long periods can potentially cause fetal adrenal suppression. If lupus activity is such that these doses are required, the risk to the fetus of not treating the disease adequately should outweigh any risk from the drug.

Cyclophosphamide, methotrexate, and azathioprine are contraindicated in pregnancy, although there have been many successful pregnancies in transplant patients taking azathioprine without obvious increased risk of adverse effect. Use of hydroxychloroquine is not recommended by the manufacturers, though there is little evidence that it has adverse effects.

It may be difficult to distinguish pre-eclampsia from a flare of renal lupus. Both can cause hypertension and proteinuria. In pre-eclampsia, unlike systemic lupus erythematosus, there are rarely urinary casts and levels of anti-dsDNA and complement are normal. These tests are therefore useful in making the diagnosis.

For further discussion of autoimmune rheumatic disorders in pregnancy see Chapter 13.14.

Occupational and psychological aspects of systemic lupus erythematosus

Systemic lupus erythematosus typically presents in young people, especially women. The onset of a chronic, essentially incurable condition at a time of life when the patient is otherwise healthy and has many plans and responsibilities is an unexpected and unwelcome burden. Many concerns arise, in particular the outlook for fertility and the ability to care for children are major worries. In those cases where the use of high-dose corticosteroids and immunosuppressive agents are essential, detailed explanations of the benefits and risks of these treatments in both the short and long term are necessary. Although a 10-year survival rate of 90 per cent may appear reassuring, it is probably less so to a 25-year-old who recognizes a 10 per cent chance of dying by the age of 35.

In making the diagnosis of systemic lupus erythematosus, therefore, the doctor must consider the effect of this condition on the overall life of the patient as well as his or her individual organs. A sympathetic understanding of the anxieties associated with the diagnosis is vital.

Controversial areas and future prospects

It is clear that we do not yet possess a cure for systemic lupus erythematosus or even a method of controlling the disease without the risk of major side-effects. The main sources of controversy concern attempts to develop new forms of treatment and to establish indices of disease activity that can be used to measure the effects of these treatments.

Plasma exchange and intravenous immunoglobulin therapy have been tried in systemic lupus erythematosus, particularly in renal crises. Overall, the results do not suggest that either form of treatment should be used routinely. New drugs such as tacrolimus and mycophenolate mofetil have been administered to small numbers of patients. Some encouraging results have been reported, but it is too early to decide on the place of these agents in the management of the disease.

There are now many different murine models of systemic lupus erythematosus. These differ in their clinical and serological characteristics and each represents at best a partial approximation to the human disease. This is important because it is now possible to administer agents such as monoclonal anticytokine antibodies to these mice and to assess the effect on the disease process. In deciding which of these agents might be effective in humans it is important to know how far one can extrapolate from the results in mice.

If new drugs or monoclonal antibodies are to be used in human systemic lupus erythematosus, it is necessary, given that mortality is now (thankfully) a rare end point, to have a recognized index by which to judge the response to treatment. Such an index must include a disease activity index, a damage index, a patient health perception index, a record of toxicity, and cost. Several global score disease activity indices, for example the systemic lupus erythematosus disease activity index, SLAM (systemic lupus activity measure), and ECLAM (European Community lupus activity measure) have been developed and provide a 'rough and ready' guide to activity. A more sophisticated approach based on the 'physician's intention to treat principle', has been derived by the British Isles Lupus Activity Group, providing an 'at a glance' review of activity in eight different organs or systems. These indices have all been compared favourably in both paper and real patient exercises. By contrast, a single damage index (the SLICC/ACR damage index) has been developed by a group of investigators and records a wide variety of potential permanent changes (for example avascular necrosis, myocardial infarction) that can occur in patients with lupus as part of the development of the disease. These principally clinical features have to be present for at least 6 months before they count. The medical outcome survey, short form 36 (SF-36), provides a useful health perception index for patients with lupus. Although not designed specifically for this condition, it has been widely used in a number of ongoing drug trials.

It is likely that the treatment of systemic lupus erythematosus in 10 years' time will be different from that given now. Basic science research is starting to identify the various strands of immune dysfunction at the core of this disease. At the same time, drug development is providing agents that are capable of selectively targeting single cell types or cytokines within the immune system. At least some of these agents are likely to be relevant to the dysfunctional mechanisms in systemic lupus erythematosus. In addition, clinicians are becoming more aware that conditions such as atherosclerosis and osteoporosis are common in patients with systemic lupus erythematosus. By increasing efforts to detect and control these associated conditions, as well as seeking to attack the underlying autoimmune disease, it should be possible to improve the lives of patients with systemic lupus erythematosus, even if a cure for the disease remains a distant prospect.

Further reading

Boumpas DT *et al*. (1992). Controlled trial of methyl prednisolone versus two regimens of pulse cyclophosphamide in severe lupus nephritis. *The Lancet* **340**, 741–5.

Casciola-Rosen LA, Anhalt G, Rosen A (1994). Autoantigens targeted in systemic lupus erythematosus are clustered in two populations of surface structures on apoptotic keratinocytes. *Journal of Experimental Medicine* **179**, 1317–30.

Cervera R *et al*. (1993). Systemic lupus erythematosus—clinical and immunologic patterns of disease expression in a cohort of 1000 patients. *Medicine (Baltimore)* **72**, 113–21.

Isenberg DA *et al.* (1997). The role of antibodies to DNA in systemic lupus erythematosus. *Lupus* **6**, 290–304.

Johnson AE *et al.* (1995). The prevalence and incidence of systemic lupus erythematosus in Birmingham, England; relationship to ethnicity and country of birth. *Arthritis and Rheumatism* **38**, 551–8.

Khamashta MA *et al*. (1995). The management of thrombosis in the antiphospholipid antibody syndrome. *New England Journal of Medicine* **332**, 993–7.

Koffler D, Schur PH, Kunkel HG (1967). Immunological studies concerning the nephritis of systemic lupus erythematosus. *Journal of Experimental Medicine* **126**, 607–24.

Morrow J *et al.* (1999) *Autoimmune rheumatic disease*, 2nd edn. Oxford University Press, Oxford.

Okamura M *et al*. (1993). Significance of enzyme linked immunosorbent assay (ELISA) for antibodies to double stranded and single stranded DNA in patients with lupus nephritis: correlation with severity of renal histology. *Annals of the Rheumatic Diseases* **52**, 14–20.

Tan EM *et al*. (1982). The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis and Rheumatism* **25**, 1271–7.

Walport MJ (1993). Inherited complement deficiency—clues to the physiological activity of complement *in vivo*. *Quarterly Journal of Medicine* **86**, 355–8.

18.10.3 Systemic sclerosis

Carol M. Black and Christopher P. Denton

Introduction

The scleroderma-spectrum of disorders includes a number of diseases with similar clinical and pathological features, and which have Raynaud's phenomenon or skin sclerosis in common. They are clinically important because the systemic forms have the highest case-related mortality of any of the rheumatic diseases, and because of the particular difficulties encountered in their management. In the United Kingdom there are approximately 300 new cases of systemic sclerosis per year and the population prevalence has been estimated to be 100 per million. Both these figures are significantly lower than estimates of disease frequency in the United States. Recent epidemiological survival analyses of patients with systemic sclerosis suggest a reduction in mortality compared with earlier studies, but this may partly be accounted for by the greater awareness of the milder forms of the disease. The disease most often develops in the fifth decade of life, and affects women approximately four times as often as men, with this ratio increasing during the childbearing years.

Those disorders included within the scleroderma spectrum are described in Table 1. The term 'prescleroderma' can be applied to the subgroup of patients with autoimmune Raynaud's phenomenon who manifest an abnormal microcirculation and scleroderma-hallmark autoantibodies (anticentromere antibodies, antitopoisomerase, or anti-RNA polymerase I).

Clinical features of localized scleroderma conditions are summarized in Table 2. The importance of distinguishing between these conditions and their subsets lies in the different clinical features, natural history, and patterns of visceral involvement that are characteristic of each subgroup.

There have been important developments in understanding the pathogenesis, clinical diversity, and management of the scleroderma-spectrum disorders over the last few years. This progress has occurred in parallel with improvements in the management of many of the organ-based complications of the condition.

Clinical features

Although thorough baseline and longitudinal investigation of patients with scleroderma-spectrum disorders is central to their management, the diagnosis of scleroderma is essentially clinical. A number of other causes of skin sclerosis or poor peripheral circulation must be considered in the differential diagnosis (summarized in Table 3). Marked differences between the major subsets of systemic sclerosis in the pattern and time-course of clinical features allow most patients to be characterized into the appropriate subset.

Patients with diffuse, cutaneous systemic sclerosis typically present over 1 to 3 years with widespread changes in skin texture, puffy oedematous extremities, generalized pruritis, and profound constitutional and inflammatory symptoms. Vasospastic symptoms are not usually prominent during the early stages, although within 18 months of their onset most patients will describe definite Raynaud's phenomenon. By contrast, the cutaneous and vasospastic symptoms of limited cutaneous systemic sclerosis are very different. The onset of skin changes is more gradual, often preceded by several years of Raynaud's phenomenon, often becoming progressively more severe, with skin sclerosis limited to the face, neck, and hands distal to the wrists. The main differences between the subsets of systemic sclerosis are summarized in Table 4.

Raynaud's phenomenon

Raynaud's phenomenon is characterized by pallor, cyanosis, suffusion, and/or pain of the fingers in response to cold or stress. The same process can also affect the toes, ears, nose, or jaw. It is present in up to 15 per cent of otherwise healthy individuals. About 1 per cent of those showing the phenomenon develop a connective tissue disease. Other conditions associated with Raynaud's phenomenon include cervical rib, vibration white finger, hypothyroidism, and uraemia. Cold-induced peripheral vasospasm is present in the majority of patients with systemic sclerosis, although generally not in those with localized forms of scleroderma.

It is important to distinguish between the primary and secondary forms of Raynaud's phenomenon. This is most reliably achieved by combining nailfold capillaroscopic assessment with autoantibody testing. Primary Raynaud's phenomenon cases are often familial, typically with onset in the late teens or early adulthood, have normal or only minimally disrupted capillaroscopic architecture, and negative antinuclear antibody tests. Those who present with isolated Raynaud's phenomenon but later develop a connective tissue disease invariably have abnormal autoimmune serology and nailfold capillary studies before they develop associated clinical features. Patients who demonstrate antibodies, including hallmark specificities for lupus or systemic sclerosis, may be designated as having autoimmune Raynaud's phenomenon.

Limited cutaneous systemic sclerosis

This was formerly termed '**CREST**' (calcinosis circumscripta, Raynaud's, (o)esophagus, sclerodactyly, and telangiectasia) and is the most common form of systemic sclerosis, accounting for over 60 per cent of cases. Patients are usually women, between 30- and 50-years old, with longstanding Raynaud's phenomenon.

Early in the disease there is non-pitting oedema of the fingers (sausage-shaped fingers), which, after several weeks or months, is gradually replaced by thickened and shiny skin. This is not usually so closely adherent to underlying structures that mobility is severely impaired, which is in sharp contrast to the findings in those with diffuse disease. Skin involvement does not spread proximally on to the trunk, but the face should be examined carefully for thin, tightly pursed lips, with furrowing and puckering of the surrounding skin, and microstomia. The most striking cutaneous finding is digital and facial telangiectasia caused by dilated capillary loops and venules (Fig. 1). Other evidence of structural vascular change is to be seen in the fingertips, where small areas of ischaemic necrosis or ulceration are common, often leaving pitting scars and pulp atrophy. Loss of the tufts of the terminal phalanges, confirmed on radiography, is also presumed to be due to ischaemia. Patients with limited cutaneous disease often develop intracutaneous and subcutaneous calcification. These deposits frequently occur in the fingers, particularly the digital pads, and in periarticular tissues such as the prepatellar area and olecranon bursa. The calcinotic masses vary in size and are often complicated by ulceration of the overlying skin, extrusion of calcific material, and secondary bacterial infection. Patients may complain of dyspepsia from reflux oesophagitis: this and other visceral

complications are discussed in detail below.

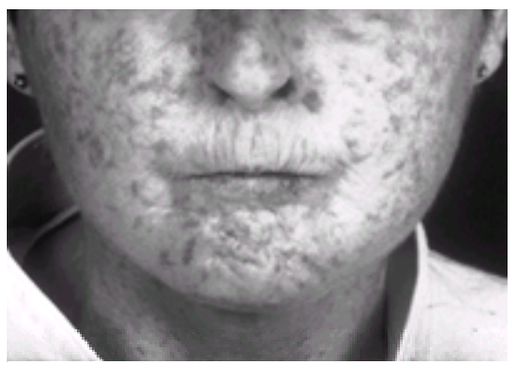

Fig. 1 This patient shows the typical facial features of limited cutaneous systemic sclerosis—microstomia, furrowing, and puckering of the skin around the mouth, beaking of the nose, and telangiectasia on the lips and face.

Diffuse cutaneous systemic sclerosis

By contrast to limited cutaneous systemic sclerosis, the onset of diffuse disease is often abrupt. It may present with widespread, symmetrical, sometimes itchy, painful swelling of the fingers, arms, feet, legs, and face. Rapid weight loss and constitutional symptoms of fatigue or weakness are frequent. If a patient with early disease presents with headache, blurring of vision, and significant hypertension, then this is a medical emergency, portending hypertensive renal crisis and requiring immediate action (see below).

The clinical findings in diffuse scleroderma depend on the stage of the disease. At onset, examination of the skin will usually reveal cold, painful, swollen hands, with swelling and stiffness already extending to the arms, feet, lower legs, face, and trunk. This oedematous phase is usually replaced within a few months by one of induration, when the skin becomes tight, shiny, and bound to underlying structures. Pigmentary changes (hyperpigmentation or hypopigmentation) accompany skin thickening in many patients. Skin involvement in diffuse scleroderma is quite different from that in the limited form of the disease, and can be mapped semiquantitatively by measuring the degree and extent of cutaneous thickening at multiple sites, from which is derived a skin score. In diffuse scleroderma this score increases rapidly at first, often peaking after 1 to 3 years, and is accompanied by impaired mobility of tendons, joints, and muscles that is clinically all too apparent. Contractures and stretching of the skin over bony points often lead to painful ulcers that are slow to heal, particularly over the proximal interphalangeal joints, elbows, and ankle malleoli.

In its earliest stages, diffuse scleroderma can be confused with an acute inflammatory arthropathy, particularly if Raynaud's phenomenon is absent. The oedematous puffy skin is often accompanied by symmetrically stiff, painful joints (hands, feet, knees, ankles, and wrists), but the classic synovitis of rheumatoid arthritis is usually absent. The clinical sign of tendon friction rubs should carefully be sought in this group of patients: these have a distinctive leathery crepitus and can be elicited during joint movement over elbows, knees, fingers, wrists, and ankles. They frequently antedate a rapid increase in cutaneous involvement, or the onset of visceral disease. Signs of carpal tunnel syndrome may be present, due to flexor tenosynovitis at the wrist.

Mild muscle disease is common and can be detected on examination, but is not usually accompanied by an increase in plasma creatine kinase or inflammatory changes on muscle biopsy. It is generally non-progressive. The few patients with florid changes of polymyositis are usually classified as having an overlap syndrome. As with limited disease, evidence of structural vascular damage—sometimes extensive—may be found in the nailfold capillaries and the digital pads.

Scleroderma sine scleroderma

These patients constitute less than 2 per cent of those with systemic sclerosis, but they are the most difficult group to recognize. They may or may not have Raynaud's phenomenon, but by definition they never have the skin changes of scleroderma: common presenting problems include oesophagitis, malabsorption, pseudo-obstruction, renal failure, cardiac arrhythmias, and interstitial lung disease.

Overlap syndromes

There are patients whose disease is not easy to define, having features overlapping with those of other connective tissue diseases. A variety of terms such as 'mixed connective tissue disease', 'undifferentiated connective tissue syndrome', and 'overlap syndromes' have emerged to describe such patients.

Whether or not mixed connective tissue disease is a true entity is controversial. Sharp and colleagues used the term in the 1970s to describe patients with some features of polymyositis, lupus, and scleroderma who ran a benign course with no pulmonary, cerebral, or renal involvement, and no vasculitis. They supposedly responded well to low-dose steroids, and could be identified by the presence of a high-titre antibody with specificity for a nuclear U1 ribonucleoprotein (**RNP**) antigen. However, the clinical features, laboratory tests, and the response to therapy have all proven not to be specific, and these patients do not fulfil the definition of and diagnostic criteria for a single disease. Neither can they be sensibly described as having an 'overlap syndrome', assuming that this definition means the coexistence of two separate diseases. Nevertheless, over time, many do develop major internal organ involvement and evolve into a defined connective tissue disease.

Ribonucleoprotein antibodies can be found in patients with scleroderma or systemic lupus erythematosus. The typical patient with the overlap syndrome presents with Raynaud's phenomenon, puffy hands, arthralgia, myositis, abnormal oesophageal motility, and lymphadenopathy. Over a period of a few years, the skin may become thickened, telangiectasia and calcinosis may appear, signs and symptoms of interstitial lung disease emerge, and the patient has developed scleroderma. Another patient with similar initial findings may develop alopecia, photosensitivity, mouth ulcers, renal disease, antibodies to double-stranded DNA and has developed systemic lupus erythematosus. Other patients may develop a prominent destructive arthropathy reminiscent of rheumatoid disease.

Autoimmune serology in systemic sclerosis

The majority of patients with scleroderma carry a hallmark autoantibody, and almost all have antinuclear reactivity, often with an antinucleolar pattern on Hep2 cells. Thus either anticentromere, and antitopoisomerase or anti-RNA polymerase III (generally also with anti-RNA polymerase I) are present in most patients and are generally (although not always) mutually exclusive reactivities. Rarer specificities include U3-RNP, PM-Scl, and anti-Th. Each serologically defined group shows somewhat different clinical features, which is of some value in risk stratification for management. There are also well-established class II MHC associations with the various autoantibodies, although some differences in association occur in different racial groups. The reactivities and reported clinical associations of the autoantibodies associated with scleroderma are summarized in Table 5.

Studies using an immunoblotting technique, which is more sensitive than immunofluorescence, have demonstrated that the anticentromere antibody (the antigen actually resides in the kinetochore region of the chromosome) is predictive for the development of limited cutaneous disease (sensitivity 60 per cent, specificity 98 per cent) and Scl-70 (an antibody known to recognize the nuclear enzyme DNA topoisomerase I) for the diffuse subset (sensitivity 38 per cent, specificity 100 per cent). Other serum autoantibodies, notably those to nucleolar antigens, are also relatively specific for scleroderma, and the proportion of patients having one or more antibody is over 80 per cent of the total. Some of these antibodies have been shown to have correlations with class II MHC haplotypes.

Less specific serological abnormalities are also found in scleroderma and include hypergammaglobulinaemia, the presence of immune complexes, low concentrations of complement components, and a weakly positive rheumatoid factor. Antibodies to SSA/Ro and SSB/La are found in 50 per cent of patients with scleroderma who also have Sjögren's syndrome, and are nearly always found in those with glandular lymphocyte infiltration rather than fibrosis.

Organ-based complications of systemic sclerosis

Despite the usefulness of an accurate subset classification of patients with systemic sclerosis, management requires that an organ-based approach be taken once the subset has been assigned. This ensures that important complications, which occur with different frequencies in the different subsets, are not missed. The overall prevalence of the different complications is summarized in Table 4.

Vascular manifestations

Raynaud's phenomenon

Episodic acral vasospasm, precipitated by cold or emotional stress (Raynaud's phenomenon), is almost universally present in patients with systemic sclerosis, although its prominence varies considerably between cases. The pathogenetic mechanism is uncertain, but probably represents an imbalance between vasoconstrictor and vasodilator mechanisms in small blood vessels, or an exaggerated release of vasoconstrictor mediators in response to physiological levels of stimulation by cold or emotion.

Raynaud's phenomenon is common in otherwise healthy individuals, with some series estimating its prevalence to be up to 15 per cent in women, with a much lower frequency in men. It may precede the onset of systemic sclerosis, especially the limited cutaneous subset, by many years, whereas in diffuse cutaneous systemic sclerosis it generally first becomes manifest around the time of the onset of other features of the disorder, or afterwards. Patients who have Raynaud's phenomenon in association with one of the hallmark autoantibodies of systemic sclerosis, such as anticentromere or antitopoisomerase-1, will often develop other features of systemic sclerosis, typically within 3 to 5 years, and so may represent a prescleroderma state—but they can also develop features of other autoimmune rheumatic disorders. Current approaches to the management of patients with Raynaud's phenomenon are summarized in Table 6.

Macrovascular disease

There have been several reports that macrovascular disease is increased in patients with systemic sclerosis. This is plausible, given the number of common aetiopathogenic mechanisms between the processes of atherosclerosis and systemic sclerosis, including endothelial-cell perturbation, activation and damage, and subsequent fibroproliferation. Large-vessel disease has important implications for the organ-based complications of systemic sclerosis such as renal disease, peripheral ischaemia, and bowel involvement. Some non-invasive studies have suggested the presence of flow abnormalities in large vessels in the cerebral and renal circulations in systemic sclerosis. Extrapolating from the results of studies investigating cardiac and pulmonary blood flow variations attributable to vasomotor instability, it is certainly possible that episodic vasospasm is not restricted to the extremities in this disease.

Skin manifestations

Scleroderma means 'hard skin' and is the hallmark of the scleroderma-spectrum disorders. The skin lesions of scleroderma differ between diffuse and limited cutaneous subsets, not only by their extent and distribution, but also by the greater tendency for there to be induration and oedema of affected tissues in diffuse cutaneous systemic sclerosis. This may reflect a local release of cytokines or altered endothelial permeability in the diffuse form. The inflammatory phase evolves into established fibrosis, sometimes leading to sheets of thickened skin or a hide-bound texture. The skin sclerosis score (skin-score) is a validated method for assessing the extent of skin involvement, and has been shown to predict survival and to correlate with some other disease features, for example a rapidly increasing skin-score is associated with an increased occurrence of scleroderma renal crisis.

Another common vascular manifestation of scleroderma is the development of local dilated loops of small blood vessels in the skin, termed 'telangiectasias'. These are often distressing for patients and may also cause problems from haemorrhage if they are at sites prone to trauma. Haemorrhage from mucosal telangiectasias is increasingly recognized as a clinical problem that may require local therapy if recurrent: gastrointestinal haemorrhage and epistaxis have both been reported. Men with facial telangiectasia may experience difficulties when shaving. Cosmetic camouflage techniques can be very effective in masking facial telangiectasia, and appropriate advice should be offered to all who might benefit. Recently the pulsed dye laser has also been used with some success.

Pulmonary disease

The most frequent cause of death related to systemic sclerosis is pulmonary disease, which can take the form of interstitial fibrosis or pulmonary vascular disease. These two processes can coexist in patients with secondary pulmonary hypertension, a subgroup that should probably be considered separately from those with isolated pulmonary hypertension.

There has recently been substantial progress in the assessment of pulmonary disease in those patients with systemic sclerosis. This has refined diagnosis and classification, will almost certainly result in different treatment strategies for particular subsets of patients, and illustrates a general theme that the subsetting of organ-based complications is becoming as important as the correct classification of major disease subsets. The investigation and treatment of the pulmonary manifestations of systemic sclerosis are summarized in Table 7.

Fibrosing alveolitis

The initial events in the development of lung disease in patients with systemic sclerosis appear to involve alveolar inflammation and subsequent epithelial and endothelial perturbation. Although all patients should be screened for this complication, it appears to affect only around 25 per cent of those with limited cutaneous disease and up to 40 per cent of those with the diffuse cutaneous form. It is strongly predicted by the presence of antitopoisomerase-1 autoantibodies and also by the *HLA-Dr52a* genotype. Increased frequency also occurs in patients with anti-RNA polymerase III antibodies, but the presence of anticentromere antibodies (**ACA**) is associated with a reduced risk. These tests are therefore of clinical value in planning the frequency and intensity of lung screening tests.

Technical developments have made the early detection of interstitial lung disease possible. Chest radiography is not sufficiently sensitive. Serial lung function tests including CO diffusing capacity are probably the most sensitive screening tool. A significant reduction from predicted values at baseline with a restrictive pattern, or a worsening of serial tests, warrants further testing with a high-resolution, thin-section (3 mm), computed tomography (**CT**) scan of the lungs and bronchoalveolar lavage. The combined use of these techniques and diethylenetriaminepentaacetic acid (**DTPA**) scans can provide much earlier diagnosis and/or indices of progression.

The earliest detectable abnormality on CT is usually a narrow, often ill-defined, subpleural crescent of increased attenuation in the posterior segment of the lower lobe. Other early CT changes include an amorphous ground-glass pattern of parenchymal opacification, or a more reticular appearance (Fig. 2). The relative extent of each pattern is important because there is good correlation between these appearances and histological findings at open-lung biopsy; an inflammatory biopsy equating to an amorphous pattern and fibrosis to a reticular one. Such information may reduce the need for an invasive biopsy.

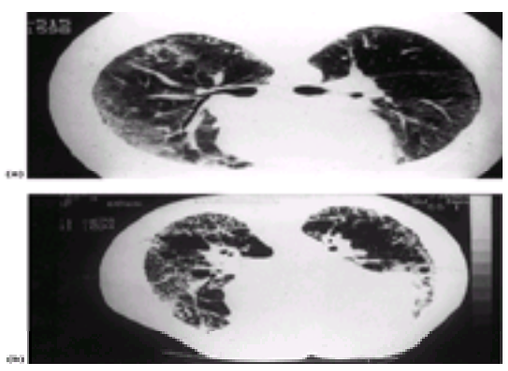

Fig. 2 (a) Thin-section CT scan image illustrating the ground-glass appearance of early pulmonary involvement posteriorly. A chest radiograph taken at the same time was normal. (b) Thin-section CT scan image illustrating extensive honeycomb shadowing and cystic air spaces involving both lower lobes. Chest radiographic appearances at the same time were of advanced interstitial lung disease (bibasilar reticulonodular shadowing). (Both images with grateful acknowledgement to Drs A.

Wells, R. du Bois, and B. Strickland, Departments of Respiratory Medicine and Radiology, Royal Brompton National Heart and Lung Hospitals.)

Bronchoalveolar lavage often identifies patients with alveolitis before the onset of symptoms or abnormalities on chest radiography or in pulmonary function tests. DTPA scans, particularly serial studies, may become useful predictors of progression or improvement: a persistently abnormal DTPA scan is associated with a higher rate of decline in pulmonary function tests subsequently, whereas a reversion to normal clearance is associated with sustained improvement in pulmonary function.

These tests have obvious value for assessing the progress of the disease and are critically important in evaluating new treatments. Lung biopsy is still the 'gold standard' for establishing the diagnosis of fibrosing alveolitis, although it is required less often than previously. However, it is now recognized that histological and CT scan appearances allow further classification of systemic sclerosis-associated fibrosing alveolitis into 'usual-pattern' (**UIP**) or 'non-specific interstitial pneumonia' (**NSIP**). These have different prognoses and probably require different treatment approaches. In general, the outcome for patients with systemic sclerosis-associated fibrosing alveolitis is better than that for cryptogenic fibrosing alveolitis of equivalent extent. Most patients with active alveolitis receive immunomodulatory therapy and many studies have suggested that this is effective. However, formal prospective controlled trials are still needed to confirm this and to refine therapeutic regimens. The management of interstitial lung disease can be further complicated by the development of secondary pulmonary hypertension.

Pulmonary vascular disease

Pulmonary vascular disease is of substantial clinical importance in patients with scleroderma. The overall prevalence is estimated at between 10 and 20 per cent. This is higher than was previously reported, reflecting the use of non-invasive methods of detection, such as echocardiography with Doppler estimation of the peak pulmonary arterial systolic pressure (from the velocity of retrograde flow into the right atrium in the presence of tricuspid regurgitation).

Isolated pulmonary hypertension (**PHT**) in scleroderma, occurring without other pulmonary pathology, is characteristic of limited cutaneous systemic sclerosis, especially in the classical CREST form of this subset with florid cutaneous telangiectasias. Some cases of isolated PHT have been seen in the diffuse form of the disease, associated with antibodies to U3-RNP, but pulmonary hypertension in this condition usually occurs in the context of established pulmonary fibrosis and constitutes secondary PHT.

There are considerable similarities between the histological features of isolated PHT in scleroderma and primary pulmonary hypertension. There is evidence of subintimal cell proliferation, endothelial hyperplasia, and the obliteration of small intrapulmonary vessels. It has been suggested that initial proliferative changes lead to the characteristic plexiform pathology of rapidly progressive pulmonary hypertension, with remodelling leading to the concentric obliterative lesions that predominate at necropsy in patients succumbing to severe systemic sclerosis-associated PHT.

Survival analysis suggests that patients with systemic sclerosis-associated PHT may have a better prognosis that those with primary pulmonary hypertension, especially if they have a mild to moderate elevation in pulmonary arterial pressure. Although patients with this condition are being identified more frequently than previously, and probably earlier, treatment remains difficult. The prognosis for those with systemic sclerosis and PHT is still to be fully defined, and haemodynamic predictors of outcome are being sought. Treatment, up to and including continuous ambulatory prostacyclin infusion or heart–lung transplantation, is probably appropriate for the most severe cases.

Cardiac involvement

Autopsy studies have identified at least three patterns of myocardial involvement in systemic sclerosis, when up to 50 per cent of patients show features of myocardial fibrosis. Other histological patterns of cardiac disease include contraction-band necrosis and, less frequently, inflammatory cardiomyopathy, the latter probably occurring most often in those with an inflammatory skeletal myopathy.

Non-invasive imaging techniques such as magnetic resonance imaging (**MRI**) or spiral CT scanning may allow myocardial fibrosis to be detected. Indirect clues of cardiac involvement may be deduced from ECG or echocardiographic studies. The investigation and management of the cardiac manifestations of systemic sclerosis are summarized in Table 8.

Pericarditis is well recognized as a complication of systemic sclerosis. It is seen particularly in the context of severe diffuse cutaneous disease and seems to be most frequently encountered in patients with established or imminent scleroderma renal crisis. Echocardiographic studies often reveal small haemodynamically insignificant effusions in patients with scleroderma: around 17 per cent of those with diffuse cutaneous and 4 per cent of those with limited cutaneous disease. Therapeutic pericardiocentesis is only occasionally required, but pericardial effusion is associated with active progressive diffuse cutaneous disease.

Electrophysiological cardiac abnormalities are commonly seen in patients with scleroderma. Conduction defects are frequent, especially Q–Tc prolongation on 12-lead ECG. Later, conduction tissue fibrosis may lead to varying degrees of heart block, including first- or second-degree block or complete heart block necessitating pacemaker implantation. Bundle-branch blocks may reflect abnormalities in the conducting tissues or be complications of ventricular strain. Thus right bundle-branch block may be seen in association with PHT, and left bundle-branch block may occur when there is left ventricular strain from hypertension or cardiac muscle disease. Paroxysmal arrhythmias are much more difficult to detect than conduction abnormalities, and in those with occult cardiac disease are probably an important cause of unexplained death in patients with systemic sclerosis.

Renal disease

Several patterns of renal pathology are recognized in patients with scleroderma: all involve vascular abnormalities. The most clearly defined is the scleroderma renal crisis, which describes the occurrence of acute renal failure in a patient with scleroderma, usually associated with accelerated hypertension (further compounding the renal pathology), and in whom no other cause for nephropathy is present. Scleroderma renal crisis was almost always fatal prior to the routine use of angiotensin-converting enzyme (**ACE**) inhibitors and the improvement in outcome over the last 20 years represents a considerable therapeutic triumph. However, in addition to scleroderma renal crisis, many patients demonstrate less severe renal complications, probably associated with reduced renal blood flow and the consequent reduction in glomerular filtration rate. The mechanism of this slowly progressive form of chronic renal disease is unclear. A small number of patients develop significant glomerulonephritis.

Acute renal crisis

This generally occurs in patients with diffuse cutaneous systemic sclerosis within 5 years of disease onset. It is often associated with rapidly advancing skin disease and the presence of tendon friction rubs. Patients often carry the antitopoisomerase-1 or anti-RNA polymerase-III autoantibody, and presence of the latter should prompt extra vigilance in monitoring for renal involvement, particularly in those with limited cutaneous disease who otherwise very rarely develop scleroderma renal crisis. The overall incidence of scleroderma renal crisis varies between different systemic sclerosis subsets and disease stages: in high-risk patients it may be as great as 20 per cent, but overall is probably less than 10 per cent.

The following diagnostic criteria for scleroderma renal crisis have been proposed: abrupt onset of arterial hypertension greater than 160/90 mmHg; hypertensive retinopathy of at least grade III severity; rapid deterioration of renal function; and elevated plasma renin activity. Other typical features include hypertensive encephalopathy and the presence of a microangiopathic haemolytic blood film: the presence of fragmented erythrocytes on a blood film is a simple and inexpensive method of identifying early scleroderma renal crisis.

Symptoms of scleroderma renal crisis usually present abruptly and the condition should be regarded as a life-threatening medical emergency requiring prompt intervention. The pulse rate is increased and patients develop headaches, visual phenomena, and convulsions due to accelerated hypertension. Symptoms and signs of left ventricular failure may follow rapidly. Oliguria or anuria lead to a rising serum creatinine level, and death from renal failure can occur within a short time in untreated patients. Proteinuria is almost universal, and although this may be present long before the renal crisis develops, it often increases with the crisis, though not to nephrotic levels. Microscopic haematuria and granular or red cell urinary casts are typically present.

Although renal crisis usually occurs in patients with established systemic sclerosis, it can occasionally be the presenting feature of the disease. For this reason the

hands and face of any patient presenting with unexplained severe or accelerated hypertension should be examined for clues that might suggest an underlying connective tissue disorder such as systemic sclerosis. Clinical suspicions should be followed up with appropriate investigations, particularly autoimmune serology and nailfold capillaroscopy.

Hypertension should be treated using ACE inhibitors. It has been suggested that quinapril may be preferable to other agents. Historically, most patients have received either captopril or enalapril together with calcium-channel blockers, aimed at reducing both diastolic and systolic pressure by 10 to 15 mmHg per day until baseline levels of diastolic pressure at 80 to 90 mmHg are achieved. Sublingual nifedipine or subcutaneous hydralazine can be used if the patient is vomiting. Intravenous prostacyclin, which may directly benefit the microvascular lesion, is often administered from diagnosis.

It is generally useful to perform a renal biopsy when hypertension has been adequately controlled. This provides prognostic information, allows histological confirmation of the diagnosis, and permits exclusion of other causes for abrupt-onset renal failure such as glomerulonephritis. Histological examination usually shows fibrinoid necrosis; mucoid or fibromucoid proliferative intimal lesions (when extensive, termed 'onion-skinning') in renal arteries, particularly the arcuate and interlobular vessels; glomerular thrombi; and ultimately glomerulosclerosis. The extent of the glomerular lesion can sometimes be useful in predicting the likely degree of ultimate functional recovery. Occasionally a similar pattern of renal dysfunction occurs without hypertension (normotensive renal crisis), suggesting that the pathological features are not simply the end-organ consequences of raised arterial pressure.

Renal function should be monitored closely with daily measurement of serum creatinine. Regular full blood counts, clotting screens, and fibrin-degradation product estimations are important to detect and monitor microangiopathic haemolytic anaemia, which often reflects activity of the disease process. Short-term haemodialysis should be given if necessary, and peritoneal dialysis often works well if long-term renal replacement therapy is needed. It has been observed that skin sclerosis and other features of systemic sclerosis can improve after a renal crisis, particularly if the patient is undergoing maintenance dialysis. The basis for this is uncertain: it may result from the removal or inactivation of circulating mediators, or simply reflect the natural history of the disease. Considerable recovery in renal function often occurs after an acute crisis, sometimes allowing dialysis to be discontinued, and improvement can continue for up to 2 years. Decisions regarding renal transplantation should not be made before this time.

Chronic nephropathy

Patients who survive scleroderma renal crisis may develop similar but less florid proliferative changes in the interlobular and arcuate arteries. Even those who have never had a renal crisis may show reduplication of elastic fibres, sclerosed glomeruli, tubular atrophy, and interstitial fibrosis, presumably reflecting the chronic changes of scleroderma.

Glomerulonephritis

There are a small number of case reports of glomerulonephritis occurring in systemic sclerosis, including a progressive crescentic glomerulonephritis in association with positive antimyeloperoxidase autoantibodies. More commonly, biopsy reveals coincident pathologies such as drug-induced injury or overlap syndromes with features of other connective tissue disorders such as systemic lupus erythematosus.

Gastrointestinal complications

The majority of patients with systemic sclerosis exhibit at least one gastrointestinal manifestation. Most frequent is oesophageal dysmotility and associated reflux oesophagitis. These symptoms often respond dramatically to treatment with proton-pump inhibitors. Involvement can also occur at other sites: these are described in Table 9, together with current management approaches for each complication.

Gastric involvement typically leads to slow gastric emptying and symptoms of postprandial fullness. This, together with sicca symptoms and difficulty swallowing, encourages poor nutritional intake and is a significant contributor to the weight loss observed in patients with this disease. The earliest feature of small bowel involvement is also dysmotility, leading to increased intestinal transit time, which together with a propensity to form wide-mouthed jejunal diverticulae leads to stagnation of the luminal contents and small intestinal bacterial overgrowth. This may in turn lead to bloating, flatulence, malabsorption, and chronic diarrhoea. Endstage involvement of the small bowel leads to profound malabsorption and malnutrition and is a significant cause of miserable scleroderma-associated death. Large-bowel manifestations include constipation and anorectal incontinence. Alternating constipation and diarrhoea is common and complicates management, which is generally empirical.

Musculoskeletal complications

Musculoskeletal features are almost universal in established systemic sclerosis, although often relatively well tolerated. Most patients with diffuse disease experience muscle weakness, although prominent myositis is unusual. Flexion contractures of the interphalangeal joints are common and can be very debilitating. Surgical intervention can be valuable but should focus on functional rather than cosmetic gain. Arthralgia and stiffness are the most frequent symptoms. Frank arthritis is uncommon and points towards an overlap syndrome. Other musculoskeletal manifestations include carpal tunnel syndrome, tendonitis (with friction rubs—most often in diffuse cutaneous disease), and the consequences of contractures—especially affecting the hands, but also more proximal joints in diffuse disease.

Other organ involvement

Neurological involvement is uncommon, but in the late stages of limited cutaneous disease a small but significant proportion of patients develop unilateral or bilateral trigeminal neuralgia. Impotence is a problem for men, usually occurring 1 to 2 years after disease onset, thought to have an neurovascular cause, and is refractory to treatment. Dryness of the mucous membranes is common, leading to dyspareunia. Hypothyroidism occurs in as many as 50 per cent of patients with systemic sclerosis and is frequently missed. Some patients have antithyroid antibodies, but lymphocytic infiltration in the gland is uncommon, fibrosis being the more typical finding.

Disease course

Since the clinical course of particular subsets of systemic sclerosis can to a great extent be predicted, appropriate classification within the scleroderma-spectrum is valuable in planning disease management.

Patients with limited disease have an 'early phase' that lasts about 10 years, when the picture is usually dominated by vascular problems such as Raynaud's phenomenon, pitting scars, digital ulcers, and telangiectasias. Later there may be worsening of the vascular disease, both cutaneously and in the pulmonary circulation. Pulmonary interstitial disease, usually more indolent than that seen in the diffuse form, can also occur as a late complication. Gut involvement may worsen with time, and oesophageal strictures, malabsorption, pseudo-obstruction, and anal incontinence are all possible late and troublesome events in this subset.

During the early phase of diffuse disease (the first 5 years), the patient is fatigued and loses weight. Hypertensive renal crisis is a real risk, and rapid progression of pulmonary and cardiac disease may occur. Arthritis, myositis, and tendon involvement can be most marked at this time. After 5 years, considered to be the late stage of diffuse disease, the constitutional symptoms settle down, the skin and musculoskeletal problems have usually reached a plateau, and there is progression of existing visceral disease but a reduced risk of new organ involvement. These differences in the pattern and natural history influence evaluation and therapy.

Pathogenesis

Systemic sclerosis involves immunological, vascular, and connective tissue abnormalities. Models of pathogenesis focus upon the importance of an initiating stimulus in a susceptible individual and subsequent amplification of pathogenic processes, leading to one of the different subsets of the disease. There is a complex interplay between a number of factors: some of the mechanisms implicated in pathogenesis of systemic sclerosis are indicated in Table 10.

Current models suggest that initiating events eventually lead to the establishment of a fibrogenic population of interstitial fibroblasts that produce increased amounts of extracellular matrix. Disruption of normal tissue architecture and secondary mechanisms such as ischaemia produce the pathological and clinical features. Determination of the discrete patterns of organ involvement within and between disease subsets is not understood, although associations with class II MHC

haplotypes and with particular autoantibodies suggest that genetic or immunological mechanisms may be important. Despite the extreme rarity of familial scleroderma, it seems likely that there is a substantial, if complex, genetic component to the pathogenesis of systemic sclerosis, which is likely to involve both severity and susceptibility loci. Twin data have failed to confirm a substantial inherited component, but some studies—for example, of the Choctaw Native American tribe—have shown a very high incidence of diffuse systemic sclerosis in some populations. A number of candidate genes, including *Fbn-1* (Fibrillin-1), are currently under investigation.

Survival

There have been a number of studies of survival in the past 50 years, and the 5-year cumulative survival rate ranges from 34 to 73 per cent. Even prolonged survival does not protect against an increased mortality risk, which continues for at least 15 years. Factors that adversely affect outcome are increasing age, being male, extent of skin involvement, and heart, lung, and renal disease. Most recent studies point to a substantial improvement in survival over the last 20 years: this is likely to be attributable to the treatment of renal crisis—previously almost invariably fatal—and perhaps to better detection and treatment of other major complications.

Management of systemic sclerosis

There have been significant recent advances in the management of patients with scleroderma, mainly related to improved treatment of organ-based complications and to the appreciation that successful management depends upon accurate diagnosis, subsetting, staging within subset, and screening for specific complications. Many advances have occurred in parallel with the improved treatment of other medical conditions such as hypertension and gastro-oesophageal reflux disease.

One of the most improved areas of clinical understanding of the scleroderma-spectrum of disorders has been the concept of risk stratification. This enables precious resources to be appropriately focused, so that patients at highest risk of particular complications are thoroughly investigated. Subsetting and staging within subsets is the starting point. Associations with particular autoantibodies are helpful, as summarized in Table 5. Additional predictive power is provided by genetic markers, and this is likely to increase considerably over the next few years. In particular, functionally relevant single nucleotide polymorphisms, such as those in cytokine or growth-factor receptors, or other polymorphic markers, including immunogenetic ones, are likely to increase predictive power.

Unfortunately there has been relatively little progress in developing disease-modifying therapies for the most aggressive subset of patients, those with diffuse cutaneous systemic sclerosis, which is probably, in part, a reflection of our relatively limited understanding of the pathogenesis of systemic sclerosis. Effective therapies are likely to be directed against key processes or mediators, and may be different depending upon the subset and stage of disease. In general, it is believed that immunomodulatory strategies are most appropriate in the earlier stages of diffuse disease (1 to 3 years from onset), whereas antifibrotic approaches may be more appropriate in established cases. An induction–maintenance approach has been used in some centres. Current approaches to disease-modifying treatment are summarized in Table 11. Although curative treatments are lacking, scleroderma-spectrum disorders, and especially systemic sclerosis, should be considered treatable. Strategies are now available to treat all the diverse manifestations, and an algorithm summarizing current approaches is shown in Fig. 3.

Fig. 3 Overview of the management of organ-based complications of systemic sclerosis.

Further reading

Black CM (1995). Measurement of skin involvement in scleroderma. *Journal of Rheumatology* **22**, 1217–19.

Black CM, Denton CP (1998). Scleroderma in adults and children. In: PJ Maddison, DA Isenberg, P Woo, DN Glass, eds. *Oxford textbook of rheumatology*, 2nd edn, pp. 1217–48. Oxford University Press, Oxford.

Bunn CC, Black CM (1999). Systemic sclerosis: an autoantibody mosaic. *Clinical and Experimental Immunology* **117**, 207–8.

Denton CP, *et al*. (1996). Systemic sclerosis: current pathogenetic concepts and future prospects for targeted therapy. *Lancet* **347**, 1453–8.

Denton CP, Black CM (2000). Scleroderma and related disorders: therapeutic aspects. *Ballières Clinical Rheumatology*.

Ho M, Belch JJ (1998). Raynaud's phenomenon: state of the art 1998. *Scandinavian Journal of Rheumatology* **27**, 319–22.

Medsger TA Jr, *et al*. (1999). A disease severity scale for systemic sclerosis: development and testing. *Journal of Rheumatology* **26**, 2159–67.

Silman AJ (1997). Scleroderma—demographics and survival. *Journal of Rheumatology* (Suppl.) **48**, 58–61.

Steen VD, Medsger TA Jr (1997). The palpable tendon friction rub: an important physical examination finding in patients with systemic sclerosis. *Arthritis and Rheumatism* **40**, 1146–51.

White B, *et al*. (1995). Guidelines for clinical trials in systemic sclerosis (scleroderma). I. Disease-modifying interventions. The American College of Rheumatology Committee on Design and Outcomes in Clinical Trials in Systemic Sclerosis. *Arthritis and Rheumatism* **38**, 351–60.

18.10.4 Polymyalgia rheumatica and giant-cell arteritis

Alastair G. Mowat

Polymyalgia rheumatica and giant-cell arteritis are common debilitating conditions that may represent opposite ends of a disease spectrum, but since they appear to present with different clinical symptoms and signs and demand different treatment it is best to describe them separately.

Polymyalgia rheumatica

Polymyalgia rheumatica occurs predominantly in patients over the age of 60 years. There is marked pain and stiffness in the shoulder and pelvic girdles associated with variable systemic symptoms and elevated C-reactive protein and erythrocyte sedimentation rate. While an incidence of some 50 per 100 000 of the population aged 50 years and over has been accepted for hospital referrals on both sides of the Atlantic, careful study of defined elderly populations has shown an incidence of 1.5 per cent which exceeds that of any other inflammatory rheumatic disease in the elderly.

Disease characteristics

Although the most common age group involved is that between 60 and 70 years, a third of patients are under 60 years old. Initial symptoms are seldom seen before 45 years or after 80 years. The male to female ratio is 1:2. The onset is often dramatic, with some patients giving the precise date of their first symptoms, and in most cases it is fully developed within a month. Pain and stiffness are usually localized to muscles, although tenderness is not as severe as in myositis. There may be additional tenderness involving periarticular structures. The onset is most common in the shoulder girdle, spreading to involve both shoulders, the pelvic girdle, and proximal muscles with striking symmetry. Involvement of distal muscles is unusual. Immobility is most severe on waking; a characteristic complaint is a need to roll out of bed, often with the aid of a spouse. Such morning stiffness may persist for hours. Most patients look unwell and complain of general malaise, fatigue, and depression. Anorexia and weight loss can be striking, often suggesting neoplasia, while night sweats and fever are frequent and are occasionally the presenting feature.

The spectrum of musculoskeletal features is broader than is often recognized; the prominent proximal myalgia overshadowing asymmetric peripheral synovitis, particularly of the knee and wrist, while in others there are periarticular features and tenosynovitis. The latter may be the principal cause of carpal tunnel syndrome, found in 10 per cent of cases. Peripheral joint features occur in up to 50 per cent of patients. Radiographs show age-related degenerative change but distinctive erosions may occur in some central joints, for example the sternoclavicular, and provide a basis for the pattern of referred pain. Arthroscopic and isotopic studies and recently magnetic resonance imaging have all supported the existence and importance of a wide pattern of synovitis.

Laboratory findings

An acute phase response (raised erythrocyte sedimentation rate, C-reactive protein) is typical. The elevation in erythrocyte sedimentation rate, often to more than 100 mm/h should not be overinterpreted since polymyalgia rheumatica accounts for only 2 per cent of such high values. Although a normal erythrocyte sedimentation rate may occur in 10 per cent of patients and be associated with the same disease course, 40 mm/h has good diagnostic value. A mild hypochromic normocytic anaemia is common. Rheumatoid factor shows a low incidence of positivity consistent with the patient's age. Serum values of liver enzymes, alkaline phosphatase, and g-glutamyl transferase are elevated in most patients, and can be correlated with the erythrocyte sedimentation rate and disease severity. Liver biopsy shows only a mild cellular infiltrate and minor changes in the bile canaliculi.

Despite the prominent muscle symptoms, electromyographic studies and serum muscle enzyme values are normal while changes on muscle biopsy, including recent studies of mitochondrial function are non-specific. While several studies have shown changes in the absolute number and percentage of activated CD8 T cells and interleukin receptors these have not provided diagnostic or therapeutic correlation.

Differential diagnosis

Polymyalgia rheumatica remains a clinical diagnosis. It is critical that a careful history is taken and a full examination carried out. This sometimes to leads to claimed disease association, for example connective tissue disease, malignancy, and thyroid disease, but statistically this has not been substantiated. Several diagnostic criteria have been validated over the years but the first remains practical: the seven best discriminatory features are shown in Table 1 of which three or more criteria are required.

Differential diagnosis includes a wide range of conditions (Table 2). Infection may be viral or bacterial, with miliary tuberculosis and infective endocarditis causing confusion. Bone diseases may be difficult to separate as they are common in this elderly group and alkaline phosphatase is raised in polymyalgia rheumatica. This is also the case with neoplastic disease, which may be associated with myalgia even in the absence of secondary spread to bone. Primary muscle disease can be distinguished by electromyography, biopsy, and enzyme values, but joint disease, particularly osteoarthritis, rheumatoid arthritis, and other connective tissue diseases, all of which may start with a polymyalgic pattern lasting for some months in older patients, cause confusion, although appropriate serological tests should help.

Aetiology and pathogenesis

No distinctive pathophysiological mechanisms have been found. Polymyalgia rheumatica may include several different conditions. The distinctive arteritis may occur in up to 15 per cent of cases on biopsy in North American series but much less frequently in Europe, perhaps reflecting case selection. However, newer techniques such as positron emission tomography again suggest that arteritis may be under-recognized. A prodromal malaise and a possible summer/winter peak incidence has promoted a generally unrewarding search for infective causes, although there may be an arteritic subset associated with parainfluenza virus type 1. The evidence for a central arthritis affecting clavicular, shoulder, and sacroiliac joints comes from a study which reproduced the usual pain patterns by injecting hypertonic saline into these joints. In those with proven arteritis, which need not be confined to the temporal and other central vessels but can be found in larger arteries all over the body, a similar pattern of referred pain can be implicated. An immune destruction of the internal elastic lamina is supported by finding circulating immune complexes, together with immunoglobulins, complement deposition, and mononuclear cell infiltrate adjacent to the lamina (Fig. 1).

Fig. 1 Photomicrograph of a temporal artery biopsy showing giant cells, mononuclear infiltrate, and disruption of the internal elastic lamina.

Although polymyalgia rheumatica is found worldwide, it is more common in Caucasians, particularly those of Scandinavian extraction. The infrequency of the disease in spouses argues against environmental factors, while familial aggregation and an association with HLA DR4 suggests both genetic and immunological mechanisms, possibly similar to seronegative rheumatoid arthritis. Both polymyalgia rheumatica and giant-cell arteritis are negatively associated with pregnancy, suggesting a vascular protection by the hyperoestrogenic state. No benefit from hormone replacement therapy is documented.

Treatment

A decade ago, the average delay to diagnosis and treatment was 6 months. Now, greater awareness and the fear of the link with giant-cell arteritis and possible blindness has led to overdiagnosis and overtreatment, and the commonness of middle-aged muscle ache has been forgotten. It is a myth that a prompt or dramatic response to corticosteroids confirm a diagnosis; many of the listed differential diagnoses show a similar response. Nevertheless corticosteroids are the drugs of choice.

As the average duration of symptoms is 2 to 3 years, with some persisting for 7 or more years, side-effects of corticosteroids are commonly recorded. Fluid retention, weight gain, diabetes, or osteoporosis are shown by 35 to 75 per cent of patients with polymyalgia rheumatica. Many clinicians now advocate starting therapy appropriate to the patient's age and wishes, to prevent osteoporosis at the outset of corticosteroid therapy. Higher initial erythrocyte sedimentation rates tend to be associated with a longer duration of disease and a greater risk of osteoporosis. Suggested corticosteroid regimes for the first 2 months are shown in Table 3. An intramuscular schedule of 120 mg methylprednisolone every 3 weeks, reducing by 20 mg every 12 weeks appears to be associated with fewer side-effects.

The gradual reduction of corticosteroid treatment over 2 years minimizes the risk of relapse; most clinical problems and associated diagnostic doubts appear to be caused by fluctuating corticosteroid dose. Erythrocyte sedimentation rate and C-reactive protein do not show a sufficiently reliable responce to help in adjusting dosage or in predicting relapse. A few patients can be managed on non-steroidal anti-inflammatory drugs, but the value of poorer and slower symptom control and the risk from different side-effects is debatable. Azathioprine 50 mg twice daily or methotrexate 10 mg per week may be helpful in steroid sparing although the evidence from clinical trials is not striking.

Relationship of polymyalgia rheumatica to giant-cell arteritis

William Bruce, a physician practising in Strathpeffer Spa, Scotland, described polymyalgia rheumatica in 1888 using the term senile rheumatic gout, and Jonathan Hutchison described giant-cell arteritis in 1890. The current names were applied much later. For more than 20 years a common cause has been suggested, emphasized by the term polymyalgia arteritica, and based upon the similar clinical and laboratory features. This has led some to search hard for arterial biopsy evidence lest the patient suffer visual features of giant-cell arteritis. Because the arteritis may be patchy and hence missed, attempts have been made to increase the sensitivity of selection of the biopsy site—angiography, sonography, isotope studies, MRI—without success. However, prospective studies emphasize the difference between the conditions, even if part of a spectrum, and the lack of need for biopsy studies in clear polymyalgia rheumatica. In giant-cell arteritis biopsy proof underpins the need for higher and sustained steroid therapy.

Giant-cell arteritis

Giant-cell (cranial, senile, or temporal) arteritis, which is rare before the age of 50 years, chiefly affects those between 65 and 75 years with a male to female ratio of 1:2. An annual incidence of biopsy-proven disease amongst those aged 50 years or more of 18 per 100 000 (25 per 100 000 for women and 10 per 100 000 for men) has been recorded in the United States and Scandinavia, with the rate for women appearing to rise.

Disease characteristics

The features of giant-cell arteritis are protean, but typical ones are shown in Table 4. The diagnosis depends upon clinical suspicion in less typical cases. As with polymyalgia rheumatica, the onset may be dramatic and the condition always becomes fully developed over a few weeks, although the delay in diagnosis may be months. The malaise, fever, and anaemia are similar to those in polymyalgia rheumatica; the differences are in the vascular symptoms. The majority have temporal features with headache, scalp sensitivity, and tender thickened arteries; the classical nodular red streaks are unusual. Overwhelming generalized headache and the feared complication of irreversible loss of vision are more readily recognized. The clinical features listed emphasize developing arteritis. A wide range of cranial manifestations reflects the involvement of larger arteries with an internal elastic lamina in the face, neck, and brain base but not in the cerebral vessels. They include headache, scalp tenderness, skin necrosis, jaw claudication while talking or chewing, tongue pain and claudication, and face and neck pain with nerve damage. The visual manifestations, which include blurred vision, amaurosis fugax, transient and permanent blindness, diplopia, and visual hallucinations, are due to ischaemic changes in the ciliary arteries causing optic neuritis or infarction, with a smaller number of cases being due to thrombosis of the central retinal artery. Fifteen per cent have evidence of arteritis elsewhere, with intermittent claudication, peripheral neuropathy, widespread vessel tenderness with bruits, myocardial ischaemia and damage, and occasionally an aortic syndrome with valve disease. Stroke due to vascular disease in the brainstem is uncommon, accounting for only 1 to 2 per cent of such cases. In contradistinction to other vasculitides, renal involvement is rare.

Laboratory features

These are the same as in polymyalgia rheumatica. Temporal artery biopsy is the definitive diagnostic test. A 2 cm segment of a tender artery will provide positive histology in 70 per cent of cases. The rate may be enhanced by taking longer segments or by the biopsy of other tender scalp vessels. While biopsy confirmation of the diagnosis is important, it should not be a reason for withholding steroids, since characteristic pathological features persist for at least 2 weeks after treatment has begun, and some argue that scar change never clears.

Differential diagnosis

Since the diagnosis of giant-cell arteritis depends upon a positive biopsy, the differential diagnosis does not include other causes of headache, neck pain, anaemia, and weight loss. The vasculitis of rheumatoid arthritis or systemic lupus erythematosus affects arterioles and is associated with other disease features, particularly arthritis and characteristic immunological tests. Polyarteritis affects small arteries with cutaneous, abdominal, and renal rather than cranial features and the histology is distinctive. Although cranial and central nervous system features occur in Wegener's granulomatosis, involvement usually includes characteristic lesions of the respiratory tract. Takayasu's arteritis, in which the pathological lesions mimic those of giant-cell arteritis, is confined to the aortic arch and its major branches and occurs chiefly in young oriental women.

Treatment

Corticosteroids are mandatory; immunosuppressive therapy has no direct effect and the modest steroid sparing rarely warrants the additional hazard. While doses of prednisolone up to 100 mg per day are often advocated, careful sequential studies indicate that lower doses are quite satisfactory (Table 3). Ophthalmologists, who

are likely to see patients with established visual effects or threatening features in the second eye, may use higher doses or methylprednisolone infusions. Dosage reduction must be gradual and should be judged solely on clinical features as acute phase responses are no guide. Most should have achieved a maintenance dose of 10 mg per day after 1 year. Subsequently the known persistence of disease in a significant proportion for 4 years or more and the possible recurrence of symptoms, including blindness, even a year after corticosteroid withdrawal argues for very gradual reduction of dosage. Unfortunately, there are no predictors of these risks. Accordingly, the hazards of therapy are even greater than in polymyalgia rheumatica and require preventive treatment for osteoporosis. Despite all the problems, giant-cell arteritis does not reduce life expectancy.

Further reading

Achkar AA *et al*. (1994). How does previous corticosteroid therapy affect the biopsy findings in giant-cell arteritis? *Annals of Internal Medicine* **120**, 987–92.

Bird HA *et al*. (1979). An evaluation of criteria for polymyalgia rheumatica. *Annals of the Rheumatic Diseases* **38**, 434–9.

Blockmans D *et al*. (1999). New argments for a vasculitic nature of polymyalgia rheumatica using positron emission tomography. *Rheumatology* **38**, 444–7.

Dasgupta B *et al.* (1998). An initially double-blind controlled 96 week trial of depot methylprednisolone against oral prednisolone in the treatment of polymyalgia rheumatica. *British Journal of Rheumatology* **37**, 189–95.

Duhaut P *et al*. (1999). Giant-cell arteritis, polymyalgia rheumatica and viral hypotheses: A multicentre, prospective case-control study. *Journal of Rheumatology* **26**, 361–9.

Duhaut P *et al*. (1999). Giant-cell arteritis and polymyalgia rheumatica: are pregnancies a protective factor ? A prospective, multicentre case-control study. *Rheumatology* **38**, 118–23.

Gonzaley-Gay MA, Garcia-Porrua C, Vazquez-Caruncho M (1998). Polymyalgia rheumatica in biopsy proven giant-cell arteritis does not constitute a different subset but differs from isolated polymyalgia rheumatica. *Journal of Rheumatology* **25**, 1750–5.

Miro O *et al*. (1999). Skeletal muscle mitochondrial function in polymyalgia rheumatica and in giant-cell arteritis. *Rheumatology* **38**, 568–71.

Myklebust G, Gran JT (1996). A prospective stsudy of 287 patients with polymyalgia rheumatica and temporal arteritis: clinical and laboratory manifestations at onset of disease and at time of diagnosis. *British Journal of Rheumatology* **35**, 1161–8.

Pearce G *et al*. (1998). The deleterious effects of low-dose corticosteroids on bone density in patients with polymyalgia rheumatica. *British Journal of Rheumatology* **37**, 292–9.

Proven A *et al.* (1999). Polymyalgia rheumatica with low erythrocyte sedimentation rate at diagnosis. *Journal of Rheumatology* **26**, 1333–7.

Salvarini C *et al.* (1999). Polymyalgia rheumatica: a disorder of extra-articular synovial structures. *Journal of Rheumatology* **26**, 517–21.

18.10.5 Behçet's disease

T. Lehner

Introduction

Behçet's disease is a recurrent, multifocal disorder that persists over many years. It was first described by Hippocrates in ancient Greece and later by Behçet, a Turkish dermatologist. Initial description of the disease comprised oral and genital ulcers and uveitis, but later a number of other clinical features were added, notably skin, joint, neurological, and vascular manifestations. This creates considerable difficulty in diagnosis and a multidisciplinary approach is often required.

An international study group has proposed a set of diagnostic criteria based on data from 914 patients with the disease from 12 centres and seven countries, requiring the presence of recurrent oral ulcers and any two of the following: genital ulcers, defined eye lesions, defined skin lesions, or a positive skin pathergy test. These criteria (Table 1) show better discrimination in sensitivity, specificity, and relative value than the previous criteria. A large number of important clinical manifestations of Behçet's disease have not been included (Table 2) because their lower frequency does not contribute to the accuracy of diagnosis. The same group has proposed that the term 'Behçet's syndrome' be replaced by 'Behçet's disease'.

Epidemiology

A striking feature of the disease is the relatively high prevalence in Japan (1 in 10 000), where in 1977 there were an estimated 11 000 patients. The prevalence is also high in countries bordering the Mediterranean: Italy, Greece, Turkey, Israel, Egypt, Lebanon, Syria, Jordan, Saudi Arabia, as well as Algeria, Tunisia, and Morocco. An epidemiological study in the United Kingdom has shown a prevalence of 1 in 170 000, which compares with 1 in 800 000 in a study in the United States.

Although the disease may develop at any age, onset is most commonly in the third decade. However, it can start in childhood with orogenital ulcers, followed by the other manifestations years or decades later. Male predominance is found in most reported series, but this may vary from 2:1 in Japan to 9:1 in the Middle East. Increased familial prevalence of the syndrome has frequently been recorded.

Aetiology

The cause of Behçet's disease is unknown but an immunogenetic basis has been established. HLA B51 is significantly associated with the disease. As with other HLA disease associations, there are at least two interpretations of these findings:

1. The HLA antigen might function as a specific receptor for viruses (or other pathogens).
2. The antigenic determinants of some pathogens might mimic the HLA antigens.

Recently, the *MIC*A6 allele (*MIC* is the major histocompatability complex class I chain-related gene, thought to be a cell stress response gene), which is in linkage disequilibrium with HLA B51, has been shown to be significantly associated with Behçet's disease.

A viral aetiology for Behçet's disease has often been claimed, but attempts to isolate viruses from patients have failed. Indirect evidence supporting a viral aetiology includes the following. Herpes simplex virus failed to replicate in inactivated cultures of mononuclear cells from patients with Behçet's disease, the interference with viral growth being interpreted as consistent with a viral aetiology of the disease. A more direct approach, using herpes simplex virus DNA probes for complementary DNA obtained from mononuclear cells of patients with Behçet's disease, showed a significant increase in hybridization, suggesting that at least part of the herpes simplex virus genome is transcribed in the circulating mononuclear cells of these patients. However, the role of the virus in the immunopathogenesis has not been elucidated; it may induce some defect in immunoregulation or invoke an autoimmune response.

A variety of streptococci (*Streptococcus sanguis*, *S. pyogenes*, *S. faecalis*, and *S. salivarius*) have been implicated in the aetiology of Behçet's disease, one hypothesis being that heat-shock protein might be a common and perhaps causative agent. Indeed, a significant increase in serum IgA antibodies to the mycobacterial 65 kDa heat-shock protein has been found in patients with Behçet's disease. Earlier reports of autoimmune responses to oral epithelial antigens have been reinvestigated, and a 65 kDa band has been identified with anti-65 kDa heat-shock protein antibodies and mucosal homogenates, as well as streptococci. This evidence, that the 65 kDa heat-shock protein might be involved in the disease is consistent with the finding of a significant increase in circulating T cells with the g† T-cell receptor. Furthermore, four peptides derived from the sequence of the mycobacterial 65 kDa heat-shock protein and the corresponding four homologous human heat-shock protein peptides specifically stimulate T cells from patients with Behçet's disease. The potential pathogenicity of some of these peptides has now been established in rats that developed anterior uveitis when the peptides were injected with adjuvant by the subcutaneous route or given orally. Overall, the evidence is growing that Behçet's disease may be closely associated with heat-shock protein peptides of microbial and crossreactive human origin.

Immunopathology

An early lymphomonocytic infiltration is usually found at the onset of ulceration in the lamina propria, the adjacent epithelium, and around small blood vessels. The latter may show endothelial cell proliferation and some obliteration of the lumen. Although the early stages are suggestive of the type IV cell-mediated immune reaction, this is followed by polymorphonuclear infiltration and fibrinoid necrosis in the blood vessels, consistent with a type III Arthus reaction. The keratinocytes of oral epithelial cells adjacent to an ulcer express HLA class II antigen.

Cell-mediated immune responses can be induced *in vitro* by homogenates of oral mucosa; these elicit lymphoproliferative responses, inhibition of leucocyte migration, and cytotoxicity. The proportion of CD4 cells may be decreased, but that of CD8 cells remains within the normal range.

Circulating immune complexes have been detected in 40 to 60 per cent of patients with Behçet's disease and are associated with disease activity. Although the concentrations of serum C3 and C4 are normal, careful sequential studies have revealed that C3, C4, and C2 are significantly reduced before an attack of uveitis, suggesting consumption of complement by the classical pathway. Electron microscopical examination of centrifuged pellets of serum reveal the presence of small membrane fragments, some of which show complement-dependent holes, suggesting that the soluble immune complex may generate C5b–9 complexes that may bind to the surface of cells and result in lysis.

Acute-phase proteins are increased in Behçet's disease, especially serum C-reactive protein and C9, which is a good marker of disease activity. An increased serum chemotactic activity is found with polymorphonuclear leucocytes and this might be due to IgG complexes releasing chemotactic factors. Serum IgA is often increased in Behçet's disease, but IgG and IgM are variable.

Unlike most autoimmune diseases, nuclear, thyroid, and gastric autoantibodies are not found in greater proportion in Behçet's disease than in the normal population. Rheumatoid factor is also negative, even in patients with joint involvement.

Clinical features

Many patients appear to be generally well and complain only of the localized lesions. However, others present with acute exacerbation of malaise, fever, dysphagia, and loss of weight. Other manifestations are listed in Table 2.

Recurrent oral ulcers

Oral ulcers are the presenting feature in most but not all patients with Behçet's disease. The ulcers can be of the minor or major aphthous or herpetiform type. However, since these ulcers are common in the general population and usually give rise only to local discomfort, they may be missed in the patient's history. Minor aphthous ulcers are found in 67 per cent of the neurological and 76 per cent of the ocular types of Behçet's disease, whereas the more severe major aphthous ulcers are found in 40 and 64 per cent, respectively, of the mucocutaneous and arthritic types. Herpetiform ulcers are found mostly in the mucocutaneous type (45 per cent). An essential feature in relation to the diagnosis of Behçet's disease is that the ulcers recur frequently, at intervals of weeks or months, but this varies from one patient to another. The long-cherished view that oral ulcers in Behçet's disease are rather severe and associated with scarring is no longer tenable. The clinical manifestations can be readily recognized and differentiated from those of similar disorders. The pharynx can also be the site of aphthous ulcers that tend to be rather large, shallow, and covered with a fibrinopurulent exudate (Fig. 1).

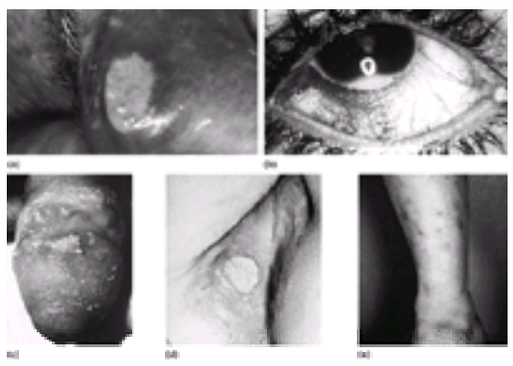

Fig. 1 Behçet's disease: (a) oral ulcer, (b) hypopyon in the eye, (c) ulceration of the head of the penis, (d) vulval ulcers, (e) multiple erythema nodosum lesions of the leg.

Genital ulcers

These are found in most, but not all, patients and can be of the three types described for oral ulcers. They affect women more commonly than men, and scars may follow healing in either sex. Females develop recurrent ulcers of their labia or vagina and they suffer from dysuria and dyspareunia. Males develop recurrent ulcers on the penis or scrotum, again with dysuria and pain on sexual intercourse; occasionally they may develop epididymo-orchitis.

Skin lesions

These vary, but diffuse pustular lesions on the face and (particularly) the back are most common. Erythema nodosum may affect the limbs or other parts of the body. Occasionally, erythema multiforme is found. Both men and women may develop perianal ulcers and, curiously, these may present in the young, well before genital ulcers have appeared.

Ocular lesions

These are the most serious developments in Behçet's disease. Relapsing uveitis, with or without hypopyon, iridocyclitis, retinal vascular lesions, and optic atrophy are common findings. Other manifestations are relapsing conjunctivitis, keratitis, and choroiditis. Gross retinal vascular changes affect both arteries and veins, and fluorescein angiography is particularly helpful in such cases. Both eyes tend to be involved: within 2 years of onset of symptoms in one eye, 90 per cent have involvement of the other eye. There is a painless, bilateral decrease in visual activity, and about 25 per cent of patients with ocular lesions become blind.

Neurological features

These are found in 10 to 25 per cent of patients with Behçet's disease. Patients most commonly develop a transient or persistent brainstem syndrome, resembling a minor stroke, but focal cerebral or spinal cord dysfunction can also occur. Others may present with meningomyelitis or meningoencephalitis, and some with organic confusional syndromes. Multiple sclerosis-like features have also been described. The cerebrospinal fluid sometimes shows pleocytosis, and raised protein and IgG concentrations but more often is normal. Computed tomography scanning does not often reveal abnormalities but the electroencephalogram can show slowing of basic rhythm. Magnetic resonance imaging is the most sensitive and reliable examination, since most patients with neurological involvement may manifest:

1. atrophy of the cerebral cortex, cerebellum, or brainstem;
2. the sinuses may be enlarged;
3. high-intensity focal lesions are found in the brainstem, basal ganglia, or the midbrain;
4. demyelinating processes may be found in the pons and medulla.

Magnetic resonance imaging can help to differentiate Behçet's disease from multiple sclerosis and other neurological diseases, as well as being useful in assessing the response to treatment.

The prognosis of Behçet's disease with neurological features used to be poor, with mortality of about 40 per cent being recorded in the literature before 1970. However, the prognosis has since been improved with reduced mortality, although whether this can be attributed to steroid and/or cytotoxic agents remains uncertain.

Arthritis or arthralgia

In about half of patients with Behçet's disease the joints are affected, typically at irregular intervals and usually more than one joint. The knees, ankles, and elbows are most commonly involved; less frequently the joints of the hands, feet, shoulders, and hips. Effusions, especially in the knees, cause considerable disability. Radiography of the joints does not usually demonstrate erosive or destructive changes, but a number of exceptions have now been recorded with erosive change in the hips, wrists, and elbows. The test for rheumatoid factor is negative.

Vascular lesions

Recurrent thrombophlebitis of leg veins is a significant feature of Behçet's disease. This has been ascribed to decreased plasma fibrinolytic activity. Less frequently,

thrombosis of the superior or inferior vena cava may develop. Arterial aneurysms have also been reported.

Gastrointestinal manifestations

These are ill-defined. The Japanese literature records diarrhoea, distension, nausea, and anorexia in more than half of patients. The ileocaecal region is the most common part of the gut to be affected. However, a British series failed to identify consistent gastrointestinal manifestations, although various transient symptoms were noted in 13 of 70 patients; two of these had rectal ulcers and one each an anal ulcer, a small intestinal ulcer, and perianal fistula. It should be noted that patients with inflammatory bowel disease are excluded from the diagnosis of Behçet's disease by the Mayo Clinic, although they may fulfil current criteria for that diagnosis.

Renal involvement

This has not been established in Behçet's disease. A small number of patients have been reported with Behçet's disease and amyloidosis affecting the kidneys, and a few also with glomerulonephritis. It is doubtful if these renal changes can be considered as primary manifestations of the disease, and they may well be coincidental. Asymptomatic proteinuria and haematuria without evidence either of amyloidosis or nephritis have also been reported in a small number of patients. In a prospective British study, two out of 38 patients with Behçet's disease showed evidence of renal disease: one of these, with biopsy-proven focal proliferative glomerulonephritis, has had no clinical symptoms, and in a 5-year follow-up period the glomerular filtration rate remained normal.

Pulmonary manifestations

These have been reported occasionally, usually with haemoptysis. In some of these patients, pulmonary tuberculosis has been suspected.

Diagnosis

A set of diagnostic criteria is presented in the Introduction (Table 1) that requires recurrent oral ulcers and any two of the following: genital, eye or skin lesions, or a positive pathergy test. However, it is recognized that the wide spectrum of clinical manifestations may not fit the above criteria and the terms incomplete and complete Behçet's disease are often used. The spectrum of the disorder can be divided into four types:

1. Mucocutaneous disease—involving oral and genital ulcers, with or without skin manifestations.
2. Arthritic type—when joint involvement is combined with some or all of the mucocutaneous manifestations.
3. Neurological type—involving the central nervous system and some or all of the features in (1) and (2).
4. Ocular type—affecting the eyes with some or all the features described in (1), (2), and (3).

Thrombosis of blood vessels can be found in any of the types of Behçet's disease, as can some of the other clinical features.

HLA B51 is significantly associated with, but not diagnostic of, the disease. The pathergy test, whereby a sterile subcutaneous puncture (without injection of any material) elicits a pustular reaction within 24 to 48 h, has been used as a diagnostic test in the Middle Eastern countries and in Japan. The presence of immune complexes is consistent with Behçet's disease and so are the raised levels of acute-phase-reacting proteins; C9 is particularly useful in monitoring the course of the disorder.

Patients with rheumatoid arthritis, osteoarthritis, or Reiter's syndrome are excluded from the diagnosis of Behçet's disease, as are patients with a firm diagnosis of ulcerative colitis or Crohn's disease. Stevens–Johnson syndrome may mimic Behçet's disease, but the recurrences are less frequent and tend to be seasonal, the ulcers are large and shallow, the lips are often covered with haemorrhagic crusts, and the skin may show typical lesions of erythema multiforme. Sarcoidosis and viral retinitis should be excluded.

Treatment

The management of patients with Behçet's disease can be difficult, as it requires close liaison between different specialties. Whenever possible, topical treatment of local lesions should be attempted before embarking on systemic anti-inflammatory or immunosuppressive therapy.

Oral and genital ulcers often respond to topical application of steroids or tetracycline or both. Uveitis is initially treated with mydriatic agents and local steroids. However, at some stage systemic prednisolone is usually administered, with a starting dose of 30 to 60 mg/day, which is rapidly brought down to a minimum effective maintenance dose of about 10 mg. There is usually a prompt response, although a small core of patients are resistant to steroid therapy. Azathioprine is often used with prednisolone (2–3 mg/kg body weight daily) and, quite apart from its steroid-sparing function, it may have additional beneficial effects. Colchicine has been advocated by Japanese and Turkish physicians, especially for the treatment of the mucocutaneous type of the disease, with a recommended dose of 0.5 mg twice a day. The rationale is that this drug inhibits the motility of polymorphonuclear leucocytes that is increased in Behçet's disease. There is a general consensus that cyclosporin (2.5–5 mg/kg body weight) should be used in patients with unresponsive uveitis, and is helpful in most patients, though its effect may gradually decline. However, cyclosporin is contraindicated in patients with neurological manifestations, because it may cause or enhance these features. Chlorambucil has also been applied successfully in the treatment of uveitis, but side-effects have limited its application. Recently, interferon-a has been found to be effective in the treatment of ocular manifestations of Behçet's disease. Thalidomide appears to be surprisingly effective in the treatment of orogenital ulcers and pustular lesions, but its teratogenic effect prevents its use in those who are or could become pregnant. Anticoagulant treatment is indicated in deep vein thrombosis.

Further reading

Akman-Demir G *et al.* (1996). Seven-year follow-up of neurologic involvement in Behçet syndrome. *Archives of Neurology* **53**, 691–4.

Benamour S, Zeroual B, Alaoui FZ (1998). Joint manifestations in Behçet's disease: a review of 340 cases. *Review of Rheumatology* **65**, 299–307.

Ehrlich GE (1997). Vasculitis in Behçet's disease. *International Review of Immunology* **14**, 81–8.

Hamuryudan V *et al.* (1997). Azathioprine in Behçet's syndrome: effects on long-term prognosis. *Arthritis and Rheumatology* **40**, 769–74.

Hamuryudan V *et al.* (1998). Thalidomide in the treatment of the mucocutaneous lesions of the Behçet's syndrome: a randomized, double-blind, placebo-controlled trial. *Annales de Medecine Interne* **128**, 443–50.

Hasan A *et al.* (1996). Role of g† T cells in pathogenesis and diagnosis of Behçet's disease. *Lancet* **347**, 789–94.

International Study Group for Behçet's Disease (1990). Criteria for diagnosis of Behçet's disease. *Lancet* **335,** 1078.

Kaneko S *et al.* (1997). Characterization of T cells specific for an epitope of human 60-kD heat shock protein (hsp) in patients with Behçet's disease (BD) in Japan. *Clinical and Experimental Immunology* **108**, 204–12.

Kotake S *et al.* (1999). Central nervous system symptoms in patients with Behçet's disease receiving cyclosporine therapy. *Ophthalmology* **106**, 586–9.

Lehner T (1999). Immunopathogenesis of Behçet's disease. *Annales de Medecine Interne* **150**, 483–87.

Masuda K *et al.* (1989). Double-masked trial of cyclosporin versus colchicine and long-term open study of cyclosporin in Behçet's disease. *The Lancet* **1**, 1093–5.

Mizuki N *et al.* (1997). Triplet repeat polymorphism in the transmembrane region of MICA gene: A strong association of six GCT repitions with Behçet's disease. *Proceedings of the National Academy of Sciences of the USA* **94**, 1298–303.

Nussenblatt RB (1997). Uveitis in Behçet's disease. *International Review of Immunology* **14,** 67–79.

O'Duffy JD *et al.* (1998). Interferon-alpha treatment of Behçet's disease. *Journal of Rheumatology* **25**, 1938–44.

Pervin K *et al.* (1993). T cell epitope expression of mycobacterial and homologous human 65-kilodalton heat shock protein peptides in short term cell lines from patients with Behçet's disease. *Journal of Immunology* **151,** 2273–82.

Sakane T *et al.* (1999). Behçet's disease. *New England Journal of Medicine* **341**, 1284–91.

Serdaroglu P (1998). Behçet's disease and the nervous system. *Journal of Neurology* **245**, 197–205.

Stanford MR *et al.* (1994). Heat shock protein peptides reactive in patients with Behçet's disease are uveitogenic in Lewis rats. *Clinical and Experimental Immunology* **97**, 226–31.

Yazici H (1981). A controlled trial of azathioprine in Behçet's syndrome. *New England Journal of Medicine* **322**, 281–5.

18.10.6 Sjögren's syndrome

Patrick J. W. Venables

Introduction

Sjögren's syndrome is characterized by inflammation and destruction of exocrine glands. The salivary and lachrymal glands are principally involved, giving rise to dry eyes and mouth. It was originally described by Sjögren in 1933 as the triad of dry eyes, dry mouth, and rheumatoid arthritis. It is now classified as primary Sjögren's syndrome where the disease exists on its own, and secondary Sjögren's syndrome where it is associated with other diseases. Well-recognized secondary associations are rheumatoid arthritis, systemic lupus erythematosus, scleroderma, polymyositis, and primary biliary cirrhosis. The recent descriptions of a Sjögren's syndrome-like illness in patients infected with HTLV-I, HIV-1, and hepatitis C virus infection have drawn attention to the importance of considering these viruses in differential diagnosis, as well intensifying the search for a virus underlying idiopathic disease.

Aetiology and pathology

The aetiology of Sjögren's syndrome is unknown but is often considered to be an interaction between constitutional and environmental factors leading to autoimmunity. Primary Sjögren's syndrome is strongly associated with HLA DR3, and the linked genes *B8*, *DQ2*, and the *C4A* null gene. Aetiological candidates for triggering autoimmunity in Sjögren's syndrome are viruses that infect the salivary gland. Sialotropic herpesviruses including Epstein–Barr virus, cytomegalovirus, and human herpesvirus-6 have been examined with conflicting reports of abnormal responses to infection. Using DNA hybridization techniques, Epstein–Barr virus has been detected in the parotid gland and in labial biopsies, but it remains controversial whether the virus is simply persistent in the glands or whether it is triggering inflammation in Sjögren's syndrome. Retroviruses have also attracted interest recently as they infect and persist in cells of the immune system such as T cells and macrophages, and infect salivary gland epithelium. In spite of increasing circumstantial evidence for their involvement in Sjögren's syndrome, the demonstration of a pathogenic role remains elusive.

The cardinal pathological features of Sjögren's syndrome are inflammation and destruction of salivary gland tissue. The inflammatory infiltrates consist of focal aggregates of lymphocytes, mainly CD4-positive T cells, localized around ducts and acini. Scattered interstitial plasma cells are commonly found, although these are not disease specific and are also found in glands from healthy individuals. The destructive changes are predominantly duct dilation, acinal atrophy, and interstitial fibrosis. These latter findings have also been described in biopsies from people without Sjögren's syndrome, particularly in the elderly, and cannot be regarded as diagnostically specific.

The most striking feature of the systemic autoimmune response in Sjögren's syndrome is the marked activation of B cells which can lead to immunoglobulin levels of over three times the upper limit of the normal range. Rheumatoid factors of all isotypes are observed in blood in about 70 per cent of patients and their detection can lead to some patients with Sjögren's syndrome being misdiagnosed as having rheumatoid arthritis. The typical autoantibodies are those against the cellular ribonucleoprotein antigens Ro and La, named after the patients in whom the antibodies were originally described. Anti-Ro antibodies are more frequently detectable (50–90 per cent of cases) than anti-La antibodies (30–50 per cent of cases), but the latter are more diagnostically specific for primary Sjögren's syndrome. More recently two further potential autoantigens have been described: fodrin, which is a cellular protein involved in apoptosis, and the muscarinic acetylcholine receptor, which is important in mediating parasympathetic stimulation of exocrine glands. An astonishing degree of diagnostic sensitivity and specificity has been claimed for serum antibodies to both antigens in Sjögren's syndrome, but these findings have yet to be confirmed by independent laboratories and must be interpreted with caution.

Clinical features

Sjögren's syndrome is nine times more common in women than men and can develop at any age from 15 to 65. Patients rarely complain of dry eyes, but rather a gritty sensation, soreness, photosensitivity, or intolerance of contact lenses. In early disease excessive watering or deposits of dried mucus in the corner of the eye and recurrent attacks of conjunctivitis may occur. The dry mouth is often manifest as the 'cream cracker' sign, inability to swallow dry food without fluid, or waking up in the night to take sips of water. About half of the patients complain of intermittent parotid swelling, sometimes misdiagnosed as recurrent mumps. When the swelling is excessively painful it is often due to secondary bacterial infection. On examination, xerostomia can be detected as a diminished salivary pool, a dried fissured tongue, often complicated by angular stomatitis, and chronic oral candidiasis. The eyes may be reddened and roughened due to shallow erosions in the conjunctivae. Occasionally the front of the eye is eroded to reveal strands of underlying collagen leading to the appearance of filamentary keratitis.

Other exocrine glands may be affected. Dry nasal passages and upper airways may lead to recurrent bouts of sinusitis, a dry cough, and, possibly, a higher than expected frequency of chest infections. Dry skin and dry hair are symptoms frequently elicited on direct questioning. About 30 per cent of women with Sjögren's syndrome have diminished vaginal secretions and may present with dyspaerunia. Involvement of the gastrointestinal tract leads to reflux oesophagitis or gastritis due to lack of protective mucus secretion, and some patients complain of constipation, which may be attributed to defective mucus in the colon and rectum. Rarely, pancreatic failure leading to malabsorption syndromes may occur.

Recent studies have highlighted yet another complication, namely interstitial cystitis. It has been suggested that this is due to an autoimmune immune response to the muscarinic acetylcholine receptor that is extensively expressed in the bladder wall. There is no doubt about the clinical association, although the serological link awaits confirmation.

There is a higher than expected frequency of thyroid autoimmunity in those with Sjögren's syndrome: whether this is part of the same pathological process is debatable, but it is important to check thyroid function from time to time in patients with this condition.

Systemic manifestations

True Sjögren's syndrome is a systemic disease. Two-thirds of patients complain of fatigue, which, according to a recent epidemiological study, is the single most important cause of disability. Occasionally weight loss and fever mimicking an occult malignancy may be the presenting symptoms, particularly in the elderly. Other features include an arthritis that resembles the Jaccoud-like arthritis of systemic lupus erythematosus (Fig. 1). Raynaud's phenomenon occurs in about 50 per cent of patients, though a true vasculitis is less common. Waldenström's benign hypergammaglobulinaemic purpura affecting the lower legs is found in patients with very high IgG levels. Patients with Sjögren's syndrome may also present with polymyalgia rheumatica or, much less frequently, polymyositis. Pleurisy occurs in about 40 per cent of patients and a high prevalence of abnormalities of pulmonary function has been described, although these are rarely clinically significant. A wide range of neurological diseases has been described: peripheral neuropathies are relatively common, particularly mononeuritis multiplex mediated by vasculitis, and a condition resembling multiple sclerosis has been reported. Interstitial nephritis leading to renal tubular acidosis or nephrogenic diabetes insipidus occurs in about 30 per cent of patients: these are usually subclinical but may lead to hypokalaemia causing muscular weakness or, occasionally, nephrocalcinosis. Lymphoma, almost always of B-cell lineage, is a characteristic but unusual feature. This occurs in about 5 per cent of patients referred to specialist centres and is particularly likely in patients with high levels of immunoglobulins, autoantibodies, and cryoglobulins. As the lymphoma develops, the immunoglobulin levels often fall and the autoantibodies become negative. Women of childbearing age are at increased risk of giving birth to babies with congenital heart block. Although rare, about 1 in 20 000 births, this complication is of great immunopathogenic interest as it is thought to be mediated by transplacental transfer of anti-Ro and anti-La antibodies.

Fig. 1 Hands of a patient with long-standing primary Sjögren's syndrome showing correctable swan-necking deformities similar to the Jaccoud-like arthritis seen in systemic lupus erythematosus.

In secondary Sjögren's syndrome the sicca symptoms are less severe than in primary disease. In rheumatoid arthritis with Sjögren's syndrome the patient tends to have more frequent extra-articular disease manifested as digital infarcts and subcutaneous ulcers. In systemic lupus erythematosus, those with Sjögren's syndrome have a lower frequency of renal disease and a relatively good prognosis.

Diagnosis

Keratoconjunctivitis sicca can be detected by Schirmer's test, tear break-up time, and Rose Bengal staining and xerostomia by a reduced parotid salivary flow rate and by reduced uptake and clearance on isotope scans. It is important to remember that both salivary and lachrymal function decline with age and may be impaired in conditions other than Sjögren's syndrome. One cause of diagnostic confusion arises from treatment with drugs with anticholinergic side-effects, the most frequent being the tricyclic antidepressants.

Biopsy and histology of the labial glands from behind the lower lip provides the most definitive diagnostic test. The area is anaesthetized with lidocaine (lignocaine) containing adrenaline and an incision 1.5 cm long allows access to five to ten glands 2 to 4 mm in diameter that are removed by simple blunt dissection. A diagnosis of Sjögren's syndrome depends on finding foci of periductular infiltrates of at least 50 lymphocytes and/or plasma cells at a density of more than one focus/4 mm^2 (Fig. 2).

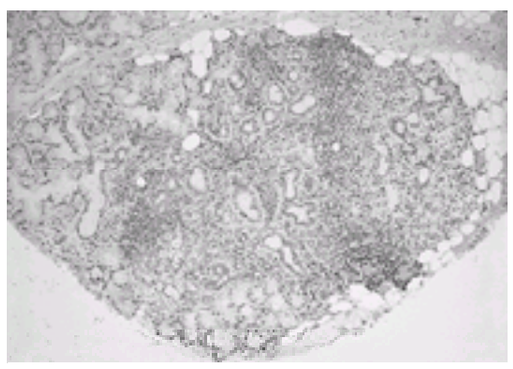

Fig. 2 Biopsy showing a lobule of minor salivary gland from a patient with Sjögren's syndrome. There is a focal inflammatory infiltrate surrounding blood vessels and ducts with the acini being relatively spared.

The majority of patients have a raised erythrocyte sedimentation rate and a mild normocytic anaemia with leucopenia in about 50 per cent of cases. One of the most remarkable features of primary Sjögren's syndrome is a high level of IgG, which can be up to 50 g/litre. Complement levels are usually normal, although C4 levels can sometimes be reduced because of the link between Sjögren's syndrome and the *C4A* null gene. Anti-La antibodies, although of relatively low sensitivity, are of great diagnostic help when present.

Rheumatoid factors, as measured by routine assays, occur in all forms of Sjögren's syndrome and their detection in primary disease is a common reason for misdiagnosing such patients as having rheumatoid arthritis. Similarly, antinuclear antibodies can occur. Both rheumatoid factors and antinuclear antibodies, although not diagnostically specific, can help in distinguishing Sjögren's syndrome from non-autoimmune causes of sicca symptoms. Primary Sjögren's syndrome can be mimicked very closely by infection with HTLV-I, HIV-1, and hepatitis C virus. All three diseases cause dry eyes and mouth, swelling of salivary glands, and biopsy changes very similar to that of primary Sjögren's syndrome. All are associated with hypergammaglobulinaemia, a raised erythrocyte sedimentation rate, and autoantibodies, although anti-Ro and anti-La are unusual. The only way to differentiate them with certainty is by specific serological testing. The Sjögren's-like syndrome associated with HIV infection has been termed diffuse infiltrative lymphocytosis syndrome and occurs in approximately 5 per cent of HIV-positive individuals. Chronic fatigue syndrome is frequently mistaken for Sjögren's syndrome and vice versa (less frequently); a salivary gland biopsy usually clarifies the situation.

Diagnostic criteria

Diagnostic criteria are essential for the standardization of any research involving patient groups, particularly with a disease, or group of diseases, as heterogeneous as Sjögren's syndrome. Currently used criteria depend on the demonstration of keratoconjunctivitis sicca, xerostomia, and a positive labial gland biopsy. The 'European' criteria based on the results of a multicentre European study are probably the most thoroughly evaluated and the simplest to apply. They are based on a short questionnaire (Table 1) about ocular and oral symptoms. Other essential criteria are ocular signs (by Schirmer's test or Rose Bengal staining), lymphocytic infiltrates on lip biopsy, salivary gland involvement (scintigraphy, sialography, or decreased salivary flow rate), and demonstration of serum autoantibodies (rheumatoid factors, antinuclear antibodies, and/or Ro or La antibodies).

Treatment

Most treatment in Sjögren's syndrome is topical and symptomatic. Simple measures can help preserve the integrity of the cornea as well as the gums and teeth and are worth pursuing with enthusiasm rather than with the negative attitude that some patients find in their physicians. Tear substitutes, such as hypromellose eye drops, are the mainstay of treatment for dry eyes, and it is generally worth trying several different types before settling on the most suitable preparation. Where thick mucus strands are a particular problem topical acetylcysteine may help. Eye ointments, particularly at night, can help lubricate sticky eyes. Bacterial infection should be treated immediately with chloramphenicol ointment or drops. Some benefit can be achieved by preventing evaporation of tears by fitting side panels to spectacles. Temporary or permanent occlusion of the canaliculi or, rarely, tarsorraphy may help to retain tears within the conjunctival sac.

The dry mouth may be treated with saliva substitutes that are now available as convenient sprays. Pilocarpine tablets have shown promising results in recent controlled trials but patients often seem to stop taking them after a few weeks or months because of cholinergic side-effects such as palpitations, sweating, and abdominal cramps. Candidal infections are extremely common in Sjögren's syndrome and are often missed. They are best treated with prolonged courses of anticandidal drugs such as fluconozole 50 mg daily for ten days. Attention to dental hygiene may help to prevent the premature caries that is a common problem in Sjögren's syndrome.

Attempts to treat the underlying disease with steroids or cytotoxic drugs are generally thought ill-advised unless there are systemic complications. Fever, weight loss,

parotid swelling, and interstitial cystitis often respond well to a low dose of steroids. Serious systemic complications such as polymyositis, mononeuritis multiplex, or fibrosing alveolitis are treated with steroids and cytotoxic drugs as in other connective tissue diseases. The arthritis of primary Sjögren's syndrome may be treated with anti-inflammatory drugs, although it also responds to hydroxychloroquine. There is no convincing evidence that methotrexate has a role in Sjögren's syndrome. It is generally agreed that other second-line agents for rheumatoid arthritis such as gold or sulphasalazine are associated with a high frequency of side-effects and this is one of the most important reasons for distinguishing between the arthritis of primary Sjögren's syndrome and rheumatoid arthritis.

There is accumulating evidence that hydroxychloroquine may have a beneficial disease-modifying effect in Sjögren's syndrome. Certainly the drug helps with arthralgia, lowers the erythrocyte sedimentation rate and immunoglobulin levels, and may also prevent bouts of purpura. More importantly there is anecdotal evidence that it may help with fatigue. Properly controlled clinical trials are needed to address this point, but it could be that hydroxychloroquine will emerge as the first disease-modifying drug which can be used in patients with uncomplicated disease.

Further reading

Alexander EL *et al.* (1986). Primary Sjögren's syndrome with central nervous system dysfunction mimicking multiple sclerosis. *Annals of Internal Medicine* **104**, 323–30.

Bacman S *et al.* (1996). Circulating antibodies against rat parotid gland M3 muscarinic receptors in primary Sjögren's syndrome. *Clinical and Experimental Immunology* **104**, 454–9.

Fox RI *et al.* (1986). Sjögren's syndrome: proposed criteria for classification. *Arthritis and Rheumatism* **29**, 577–85.

Fox RI *et al.* (1988). Treatment of primary Sjögren's syndrome with hydroxychloroquine. *American Journal of Medicine* **85**, 62–7.

Fox RI, Tornwall J, Michelson P (1999). Current issues in the diagnosis and treatment of Sjögren's syndrome. *Current Opinion in Rheumatology* **11**, 364–71.

Flescher E, Talal N (1991). Do viruses contribute to the development of Sjögren's syndrome? *American Journal of Medicine* **90**, 283–5.

Haneji N *et al.* (1997). Identification of alpha-fodrin as a candidate autoantigen in primary Sjögren syndrome. *Science* **276**, 604–7.

Harley JB *et al.* (1986). Gene interaction at HLA-DQ enhances autoantibody production in primary Sjögren's syndrome. *Science* **232**, 1145–7.

Price EJ and Venables PJW (1995). Aetiopathogenesis of Sjögren's syndrome. *Seminars in Arthritis and Rheumatism* **25**, 117–33.

Thomas E *et al.* (1998). Sjögren's syndrome: a community-based study of prevalence and impact. *British Journal of Rheumatology* **37**, 1069–76.

Vitali C *et al*. (1993). Diagnostic criteria for Sjögren's syndrome: results of a European prospective multicentre study. *Arthritis and Rheumatism* **36**, 340–7.

18.10.7 Polymyositis and dermatomyositis

John H. Stone and David B. Hellmann

Introduction

Polymyositis and dermatomyositis are the two major types of inflammatory muscle disease. Despite numerous shared features, immunopathological evidence now confirms that they are separate disorders. Polymyositis and dermatomyositis are considered to be autoimmune diseases because of their inflammatory nature and the frequent occurrence of autoantibodies (both antinuclear antibodies and myositis-specific antibodies), but the precise causes of these disorders remain unknown. Both are associated with proximal muscle weakness, and both may affect organs other than skeletal muscle, such as the lungs and heart. In adults, dermatomyositis is so frequently a paraneoplastic disorder that its diagnosis should prompt a search for an underlying malignancy.

Clinical features

Classification and epidemiology

The traditional classification of polymyositis and dermatomyositis distinguishes between five subgroups of patients:

1. primary idiopathic polymyositis;
2. primary idiopathic dermatomyositis;
3. either disorder occurring in association with a malignancy;
4. childhood dermatomyositis (or, more rarely, polymyositis); and
5. overlap syndromes, in which polymyositis or dermatomyositis occur along with features of other systemic autoimmune conditions.

This chapter focuses on primary idiopathic polymyositis and dermatomyositis, referring to the other subgroups when appropriate. The principal features of polymyositis and dermatomyositis are displayed in Table 1.

In general, polymyositis and dermatomyositis may afflict individuals of any age and either gender, but female cases outnumber males by a 2:1 ratio. Cases associated with malignancy are clustered among older patients with dermatomyositis, and seldom if ever occur among children.

Polymyositis

Polymyositis is characterized by symmetrical proximal muscle weakness that develops slowly, usually over weeks to months. Routine tasks that require proximal muscle strength, for example rising from a chair or climbing stairs, become increasingly challenging for the patient. In addition to weakness of the extremities, skeletal muscles at many sites, including the upper one-third of the oesophagus, the muscles of neck flexion, the intercostal muscles, and the diaphragm, are also susceptible to muscular inflammation. Dysphagia and nasopharyngeal regurgitation of food may result from oesophageal involvement. Patients with severe neck flexor weakness secondary to polymyositis may be unable to lift their heads from the pillow. Hypercapnoeic respiratory failure sometimes results from weakness of the chest wall muscles and diaphragm. By contrast, polymyositis usually spares the muscles that mediate facial expression and extraocular movements, even in patients with profound weakness elsewhere. Similarly, handgrip strength and the ability to perform fine motor tasks usually remain preserved until advanced stages of the disease.

Prominent muscle pain and tenderness are atypical, and rarely constitute the chief complaint. Deep tendon reflexes and muscle bulk are preserved except in severe, advanced disease. Fasciculations, a manifestation of denervation rather than myopathic injury, are absent in polymyositis. Similarly, sensory function remains normal even as muscle weakness progresses. Endomysial inflammation leads to the release of muscle enzymes into the blood. Thus, polymyositis is characterized by striking elevations of serum creatine kinase, aldolase, aspartate aminotransferase, alanine aminotransferase, and lactate dehydrogenase.

Dermatomyositis

The pattern of muscle involvement in dermatomyositis is clinically indistinguishable from that of polymyositis. In addition to inflammatory muscle disease, however, dermatomyositis has an array of characteristic cutaneous manifestations. Gottron's sign, an erythematous, scaly eruption confined to skin overlying the knuckles, is pathognomonic of this disease (Fig. 1 and Plate 1). Identical lesions known as Gottron's papules also occur over the extensor surfaces of many other joints, particularly the elbows and knees. The heliotrope rash consists of a lilac discoloration of skin over the eyelids, often accompanied by eyelid oedema (Fig. 2 and Plate 2). Cutaneous erythema may involve several sites in dermatomyositis, including the upper back and shoulders (the 'shawl sign'), the upper chest (in a 'V' distribution), and the face and hands. 'Mechanic's hands' (see below) often occur in association with certain types of myositis-specific antibodies.

Fig. 1 Gottron's sign. Roughened, violaceous papules over the dorsal surfaces of several metacarpophalangeal and proximal interphalangeal joints. Note also the erythema at the bases of the fingernail, caused by capillary loop dilatation. (See also Plate 1.)

Fig. 2 Heliotrope rash. An erythematous (often lilac-coloured) rash over the eyelids in a patient with dermatomyositis (reproduced from Mousari HC, Wigley FM (2000). *Journal of Rheumatology* **27**,1542-5 with permission). (See also Plate 2.)

Patients with dermatomyositis may mimic systemic lupus erythematosus in demonstrating photosensitivity and malar rash. Skin biopsies in these two diseases share the histopathological features of 'interface dermatitis', i.e. immune complex deposition at the dermal–epidermal junction. Acral regions are also affected by characteristic features. Both periungual erythema and capillary loop dilatation underscore the vascular nature of this disorder. Sometimes the classic skin features are apparent for months before muscle weakness becomes evident, in which case the condition is termed amyopathic dermatomyositis (or 'dermatomyositis sine myositis').

Extramuscular features

Lung

Weakness of the intercostal muscles and diaphragm occasionally leads to ventilatory failure in polymyositis or dermatomyositis. More commonly, however, patients suffer pulmonary involvement in the form of interstitial lung disease, a complication that occurs in up to 30 per cent of cases. The pattern of pulmonary involvement in the inflammatory myopathies is typical of that which occurs in connective tissue disorders, namely interstitial fibrosis, predominantly at the lung bases. High-resolution computed tomography is very sensitive for detecting this type of pulmonary change, which in the early stages corresponds to an inflammatory alveolitis. The severity of interstitial lung disease in polymyositis and dermatomyositis ranges from asymptomatic radiological findings to a refractory process indistinguishable from idiopathic pulmonary fibrosis. Lung involvement is often, but not always, associated with antisynthetase antibodies (see below). Restrictive findings on pulmonary function testing are the rule. Another common pulmonary complication of the inflammatory myopathies is aspiration pneumonia, caused by weakness of the hypopharynx and upper oesophagus.

Cardiac

Cardiac involvement in polymyositis or dermatomyositis is usually subclinical, and its prevalence is not known with certainty. When measured, creatine kinase-MB isoenzyme levels are frequently elevated in the absence of overt cardiac symptoms. Electrocardiograms usually demonstrate non-specific ST-T wave changes. Even so, the occurrence of clinically evident myocarditis in these diseases may lead to cardiac failure or intractable, life-threatening arrythmias.

Gastrointestinal

Involvement of the gastrointestinal tract beyond the pharynx and oesophagus is particularly common in juvenile dermatomyositis (the childhood form). This complication is mediated by vasculitis, and may result in intestinal haemorrhage or perforation.

Malignancy

A significant proportion of adults with polymyositis or dermatomyositis have underlying malignancies, usually carcinomas. The association with malignancy is stronger for adults with dermatomyositis, who die from cancer significantly more often than age-matched controls. The weaker association of polymyositis with malignancy may be attributable to ascertainment bias resulting from more frequent visits to the doctor and cancer surveillance.

Many types of malignancy have been reported in association with dermatomyositis, including lung, oesophageal, breast, colon, and ovarian tumours. Among women with dermatomyositis, the risk of ovarian cancer may be 20 times greater than that of the general population. Most patients diagnosed with primary polymyositis or dermatomyositis (children excepted) should undergo some surveillance for a disease-associated malignancy. This screening should be based upon careful histories, physical examinations, and the performance of a limited number of routine tests (for example chest radiography), in addition to age-appropriate cancer screening such as mammography and flexible sigmoidoscopy. Costly, undirected 'fishing expeditions' rarely benefit the patient.

Differential diagnosis

Polymyositis and dermatomyositis must be distinguished from numerous disorders that cause subacute weakness. These are shown in Table 2.

Pathology

Polymyositis is characterized by an endomysial infiltrate containing large numbers of CD8+ T cells, combined with foci of cytotoxic T cells and macrophages. The inflammatory infiltrate surrounds and invades non-necrotic muscle fibres. In polymyositis, both uninvolved and involved fibres express increased amounts of HLA class I antigen (normal muscle fibres, by contrast, express neither class I nor class II antigens). Thus, the pathological findings in polymyositis suggest an HLA class I-restricted immune response mediated by cytotoxic T cells. Polymyositis must be distinguished from inclusion body myositis, a more indolent form of inflammatory myopathy associated with asymmetric and distal motor weakness. Inclusion body myositis has distinctive pathological findings: 'rimmed vacuoles' distributed around the myocyte's edge, basophilic 'inclusion bodies' within these vacuoles, and filamentous inclusions within the cytoplasm. Cases of 'refractory polymyositis' are often misdiagnoses of inclusion body myositis.

Muscle biopsies from patients with dermatomyositis contain increased numbers of CD4+ T cells and B lymphocytes. The inflammatory infiltrate in dermatomyositis is localized to perivascular regions. Capillary obliteration, fibrin thrombi, and endothelial cell damage are all hallmarks of dermatomyositis. Evidence of the membrane attack complex, comprising complement components C5 to C9, is present early in dermatomyositis, consistent with the humorally mediated destruction of muscle-associated microvasculature. Focal capillary depletion is one of the earliest pathological changes. In addition to its vascular orientation, the inflammatory infiltrate in dermatomyositis centres on the interfascicular septae and around, rather than within, muscle fascicles. Even in the absence of inflammation, perifascicular atrophy is diagnostic of dermatomyositis.

Myositis-specific antibodies

Approximately 30 per cent of patients with polymyositis/dermatomyositis have myositis-specific antibodies, immune globulins directed against a variety of nuclear or cytoplasmic antigens. Three major types of myositis-specific antibodies have been identified: the antisynthetases; antisignal recognition particle antibodies (**anti-SRP**); and anti-Mi-2 antibodies. Individual patients generally develop only one type of myositis-specific antibody. The antisynthetase antibodies are directed against aminoacyl-tRNA synthetase enzymes, which catalyse the attachment of specific amino acids to their cognate tRNAs. In the case of Jo-1 (the most common type of myositis-specific antibody), the antibody is formed against antihistidyl-tRNA synthetase. Anti-Jo-1 antibodies inhibit the function of their target antigens *in vitro*. In addition to anti-Jo-1 antibodies, antibodies to several other aminoacyl-tRNA synthetase enzymes have been described, including anti-OJ, anti-PL-12, and anti-KJ.

Patients with antisynthetase antibodies often manifest a unique disease phenotype known as 'the antisynthetase syndrome'. This syndrome occurs in 30 per cent of

patients with either polymyositis or dermatomyositis. It is characterized by relatively acute disease onset, the presence of constitutional symptoms (for example fever), interstitial lung disease, Raynaud's phenomenon, arthritis, and 'mechanic's hands'. 'Mechanic's hands' consist of roughened, cracked skin on the lateral and palmar surfaces of the fingers and hands, with irregular, dirty-appearing lines, resembling the hands of a manual labourer. The presence of antisynthetase antibodies in a patient denotes a disease phenotype that usually responds to corticosteroid treatment but which is likely to persist. Thus, patients with antisynthetase antibodies may be candidates for early use of immunosuppressive agents in addition to corticosteroids.

Anti-SRP antibodies occur exclusively in polymyositis. They react with the signal recognition particle, a complex of RNA and protein involved in translocating newly synthesized proteins into the endoplasmic reticulum. Anti-SRP antibodies, which occur in approximately 5 per cent of all adult patients with polymyositis, are associated with muscle inflammation of acute onset, severe degree, and refractoriness to therapy.

Finally, anti-Mi-2 antibodies almost always occur in patients with dermatomyositis, often in patients whose cutaneous involvement is prominent. The target of Mi-2 is a complex of nuclear proteins whose function remains unknown. In comparison with patients with antisynthetases and anti-SRP antibodies, those with antibodies to Mi-2 usually have better treatment outcomes.

Diagnosis and treatment

The unequivocal presence of Gottron's sign in association with proximal muscle weakness and elevation of muscle enzymes obviates the need for muscle biopsy because it is pathognomonic. In virtually all other cases of possible polymyositis or dermatomyositis, however, confirmation of the diagnosis by muscle biopsy is essential. Other studies, such electromyography, nerve conduction studies, and magnetic resonance imaging, are adjuncts to diagnosis but do not supplant tissue biopsy.

Corticosteroids, usually beginning with 1 mg/kg/day of prednisone, remain the cornerstone of all initial treatment regimens for polymyositis and dermatomyositis. Decline of the creatine kinase level within 2 weeks of starting treatment may portend a good outcome, but improvement in muscle strength frequently lags and is sometimes not evident for up to 3 months. Patients should be treated with 1 mg/kg/day of prednisone until the creatine kinase is normal (or nearly so), and then undergo a slow taper that does not exceed 10 mg/month. Once the steroid taper has begun, creatine kinase levels are useful in gauging disease activity, but mild creatine kinase elevations do not justify escalations in treatment, particularly if the patient's muscle strength continues to improve. Conversely, once treatment has begun, low creatine kinase levels do not guarantee inactive muscle disease. Dermatomyositis is particularly notorious for the finding of low or normal creatine kinase levels despite active muscle inflammation. Steroid myopathy is a common complication of treatment, and may be difficult to distinguish from active disease.

Many patients with polymyositis or dermatomyositis require additional immunosuppressive agents during their course.

Methotrexate (up to 25 mg/week) or azathioprine (2 mg/kg/day) are the initial second-line agents of choice. Cyclophosphamide may be preferred for patients who have severe interstitial lung disease at presentation or for the rare patient presenting with overt features of necrotizing vasculitis. Intravenous immune globulin is useful in refractory cases of dermatomyositis, but its expense precludes its use in all patients as an initial therapy. Finally, in addition to pharmacological treatments, physical therapy and rehabilitative medicine play important roles in patient recovery.

Prognosis

Prompter diagnoses, a broader range of therapies, and improved general medical care have improved the 5-year survival rate of patients with polymyositis or dermatomyositis to greater than 80 per cent. However, morbidity from both the diseases themselves and their treatments is high, and few patients emerge from treatment cured and unscathed. Several variables may contribute to worse outcomes or suboptimal therapeutic responses, including delay in diagnosis, the presence of severe myositis, dysphagia, pulmonary or cardiac involvement, the diagnosis of inclusion body myositis, association with malignancy, and the presence of certain myositis-specific antibodies.

Further reading

Bohan A *et al*. (1977). A computer-assisted analysis of 153 patients with polymyositis and dermatomyositis. *Medicine* **56**, 255–86.

Cherin P *et al*. (1993). Dermatomyositis and ovarian cancer: a report of 7 cases and literature review. *Journal of Rheumatology* **20**, 1897–99.

Dalakas MC (1991). Polymyositis, dermatomyositis, and inclusion-body myositis. *New England Journal of Medicine* **325**, 1487–96.

Dalakas MC *et al*. (1993). A controlled trial of high-dose intravenous immune globulin infusions as treatment for dermatomyositis. *New England Journal of Medicine* **329**, 1993–2000.

Kissel JT, Mendell JR, Rammohan KW (1986). Microvascular deposition of complement membrane attack complex in dermatomyositis. *New England Journal of Medicine* **314**, 329–34.

Lie JT (1995). Cardiac manifestations in polymyositis/dermatomyositis: how to get to the heart of the matter. *Journal of Rheumatology* **22**, 809–11.

Plotz PH *et al*. (1995). Myositis: immunologic contributions to understanding cause, pathogenesis, and therapy. *Annals of Internal Medicine* **122**, 715–24.

Schwarz MI (1998). The lung in polymyositis. *Clinical Chest Medicine* **19**, 701–12.

18.10.8 Kawasaki syndrome

Tomisaku Kawasaki

Introduction

Kawasaki disease, first described by Kawasaki in 1967, is an acute febrile, multisystem vasculitic illness of unknown aetiology most commonly affecting children younger than 5 years of age. Originally, the prognosis was believed to be favourable. However, as more studies were carried out, the mortality rate was found to be about 0.3 to 0.5 per cent, but this has subsequently declined to around 0.05 to 0.1 per cent. Autopsy findings revealed exceptional pathological features such as coronary artery aneurysms with thrombosis in many cases. However, since the disease cannot be differentiated histopathologically from infantile periarteritis nodosa (infantile polyarteritis), previously the subject of a few reports in the American and European literature, it is still unclear whether Kawasaki disease is a new entity or a condition that had previously been overlooked. This problem will remain unsolved until the pathogenesis of both diseases is determined.

Although in most cases Kawasaki disease is self-limiting, approximately 25 per cent of untreated patients manifest coronary artery changes such as dilatation and/or aneurysms on echocardiogram. Since 1984, treatment with high-dose intravenous immunoglobulin (**IVIG**) has been used to produce a rapid resolution of the fever and other inflammatory manifestations of Kawasaki disease, in addition to reducing the frequency and severity of coronary artery abnormalities.

Since more than 30 years has elapsed from the first description of Kawasaki disease, some of the earlier patients are now adults. The coronary artery features of the condition may be an important cause of ischaemic heart disease in young adults. Physicians should therefore ask patients with ischaemic heart disease, particularly those under 40 years of age, whether they have a history of childhood Kawasaki disease.

Clinical manifestations

Kawasaki disease is a clear-cut clinical entity that can be diagnosed after the recognition and analysis of six main symptoms (Table 1). The clinical features can be classified into two categories: principal and subsidiary. At least five of the six main features are required for diagnosis. However, patients with four features can also be diagnosed as having the condition, provided that coronary aneurysms are identified by echocardiography or coronary angiography.

Principal features

Fever of unknown aetiology lasting 5 days or more

In general, the onset of Kawasaki disease is with abrupt high fever but without prodromal symptoms such as coughing, sneezing, or rhinorrhoea. Cervical lymphadenopathy is sometimes felt, particularly if the patient complains of neck pain, and this can precede the fever by a day. Usually the fever is remittent or continuous, ranging from 38 °C to 40 °C, for 1 to 2 weeks. High fever lasting more than 2 weeks is seen in 14 to 20 per cent of untreated cases, but it rarely lasts for more than 30 days. The longer the high fever continues, the greater the possibility of coronary artery aneurysm. Nothing apart from IVIG appears to reduce the fever, which resolves significantly faster when IVIG is administered with aspirin, as compared to therapy with aspirin alone. However, in about 10 per cent of cases, IVIG is not effective in reducing fever.

Bilateral congestion of ocular conjunctiva

Conjunctival infection develops 2 to 4 days after the onset of fever. Each capillary vessel is dilated. There is no purulent discharge, so the term 'conjunctivitis' is inappropriate. In most cases redness of the eyes is obvious, but in some cases it can only be seen upon very close examination. Pseudomembrane formation, iris adhesion, or visual disturbance has not been reported. Anterior uveitis can be observed in 66 per cent of cases upon careful slit-lamp examination early in the course of the disease.

Changes of lips and oral cavity

Dryness, redness, and fissuring of the lips occur 3 to 5 days after the onset of fever. The membranes of the oral cavity and pharyngeal mucosa are diffusely red. There is no vesicle, aphtha, or pseudomembrane formation. Frequently there is prominence of the tongue papillae, referred to as a strawberry tongue and similar to that seen in scarlet fever (Fig. 1 and Plate 1).

Fig. 1 Typical appearance of a patient with Kawasaki disease, note the red eyes and red lips (picture of a 5-year-old boy, taken on the fourth day of illness). (See also Plate 1).

Acute non-purulent swelling of cervical lymph nodes

From the day before the onset of fever, or together with fever, there is swelling of the cervical lymph nodes. The patient complains of pain and often suffers a wry neck. In some cases the swelling occurs several days after the onset of fever. The nodes range from 1.5 to 5 cm in size and form a firm, non-fluctuant mass. Sometimes there is bilateral swelling leading to a misdiagnosis of mumps.

Polymorphous exanthema

From the first to the fifth day after the onset of fever, a polymorphic rash appears on the trunk or extremities. It is variously morbiliform, scarlatiniform, urticariform, or

erythema multiform-like. In each case the rash is a different combination of these forms. They are not accompanied by vesicles or crusts, but sometimes there are small aseptic pustules on the knees, buttocks, or other sites. The eruptions usually disappear in less than a week.

Changes in the extremities

Approximately 2 to 5 days after the onset of the disease, when the rash on the trunk has appeared, there is reddening of the palms and soles. Simultaneously, there is an indurative oedema in the hands and feet. From 10 to 15 days after the onset of the illness, desquamation begins from the tips of the fingers and membranous desquamation spreads over the palm up to the wrist. From 45 to 60 days after onset, transverse furrows frequently appear in the nails of both the fingers and toes.

Subsidiary clinical manifestations and complications

Cardiovascular complications

Cardiovascular manifestations can be remarkable in the acute phase of Kawasaki disease and are the leading cause of long-term morbidity and mortality. In this phase, the pericardium, myocardium, endocardium, and coronary arteries may all be involved. Clinically recognizable myocarditis is common, with tachycardia, gallop rhythm, and signs of cardiac failure.

The electrocardiogram is abnormal in one-third of patients, showing low-voltage, ST-segment depression, and T-wave flattening or inversion. Coronary arterial abnormalities develop in approximately 25 per cent of untreated patients (Fig. 2). Aneurysms have been detected within 7 days of illness, but more commonly they occur between 10 days and 3 weeks after the onset of symptoms. The appearance of aneurysms more than 4 weeks after the onset of illness is uncommon. Patients with giant aneurysms (internal diameter of at least 8 mm) have the worst prognosis and are at the greatest risk of developing coronary thrombosis, stenosis, or myocardial infarction. Angiography is sometimes used for diagnosis, particularly in patients with suspected or definite echocardiographic changes or ischaemia. Because of considerable normal variations in the coronary arteries in childhood, only an experienced paediatric cardiologist can properly interpret the angiograms.

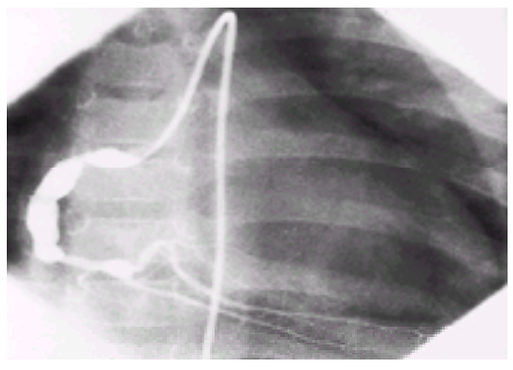

Fig. 2 Right coronary artery aneurysms, seen 1 month after the onset of Kawasaki disease in a boy aged 5 years and 7 months. (By courtesy of Dr T Sonobe of the Japanese Red Cross Medical Center, Tokyo.)

Myocardial infarction is the principal cause of death in patients with Kawasaki disease. It may occur within a year; even later in patients who have giant aneurysms. Children with giant aneurysms more likely have other arterial involvement including that of the renal, brachial, and iliac arteries. Valvular involvement, primarily mitral regurgitation, has been described in about 1 per cent of children with Kawasaki disease.

Gastrointestinal tract

Diarrhoea occurs in approximately 35 per cent of patients. Patients with gall bladder involvement (acute acalculous distension: 'hydrops') often suffer severe abdominal pain, especially in the upper right quadrant. Mild jaundice occurs in approximately 5 per cent of cases. The total serum bilirubin level is almost always lower than 10 mg/dl. In the acute phase, serum transaminase levels are often increased. Serum glutamate-oxaloacetate transaminase (**GOT**) and glutamate-pyruvate transaminase (**GPT**) levels increase from 60 to 200 IU, while the lactate dehydrogenase (**LDH**) level increases from 600 to 900 IU. Paralytic ileus has been reported.

Blood

In almost all cases there is leucocytosis with a shift to the left, an increased erythrocyte sedimentation rate (**ESR**), elevated C-reactive protein, and an increased a2-globulin level. The platelet count increases from the second week and may reach 1000 to 1500 × 10^{9} per litre. Hypoalbuminaemia and slight anaemia are common.

Urinary tract

Albuminuria is frequently seen in the acute phase, with aseptic microscopic pyuria. These findings disappear in the convalescent phase.

Respiratory system

Preceding or concurrent respiratory symptoms such as cough and rhinorrhoea are occasionally seen. Abnormal infiltrates on the chest radiograph are occasionally observed.

Joints

Arthritis or arthralgia can occur in the initial phase of the illness and are usually polyarticular, involving the knees, ankles, and hands. A pauciarticular arthritis involving the knees, ankles, or hips commonly appears during the second or third week of illness. These symptoms disappear within 30 days after their onset in most cases.

Nervous system

Infants with Kawasaki disease are often more irritable than infants with other febrile illnesses. Signs and symptoms suggestive of aseptic meningitis may be present in some patients, and this is found in 20 to 50 per cent of those with Kawasaki disease who undergo lumbar puncture. Other neurological complications such as facial palsy, hemiplegia, and encephalopathy have been reported.

Other systems

Auditory abnormalities, testicular swelling, and peripheral gangrene have also been reported.

Pathological findings

Kawasaki disease is an acute inflammatory disease with systemic angiitis which is distinguishable from classic periarteritis nodosa of the Kussmaul–Maier type. Coronary aneurysms are usually present at autopsy. The angiitis is characterized by acute inflammation with or without mild fibrinoid necrosis. Middle- or large-sized arteries (such as the main coronary, iliac, axillary, or renal arteries and aorta) are commonly involved. The course of the angiitis can be classified into four stages

according to the duration of the illness:

- *Stage 1* (1–2 weeks from onset) shows perivasculitis and vasculitis of the microvessels, small arteries, and veins. There is inflammation of the intima, externa, and perivascular areas in medium- and large-sized arteries. Oedema and infiltration with leucocytes and lymphocytes are also present.
- *Stage 2* (2–4 weeks from onset) shows less inflammation in the vessels than in Stage 1. This stage is characterized by panvasculitis of the main coronary arteries and aneurysm with thrombus in the stems. Myocarditis, coagulation necrosis of heart muscle, lesions of the conduction system, pericarditis and endocarditis with valvulitis are also present.
- *Stage 3* (4–7 weeks from onset) shows subsidence of inflammation in the vessels. Granulation may occur in the medium-sized arteries.
- *Stage 4* (more than 7 weeks from onset) reveals scar formation and intimal thickening with aneurysms, thrombus, and stenosis in the medium-sized arteries.

Other lesions include myocarditis, pericarditis, and inflammation of almost all organs. All these lesions are frequently seen in Stage 1 and 2, but rarely in Stage 4. Ischaemic heart disease usually occurs in Stages 2 to 4. The major cause of death in Stage 1 is myocarditis, including inflammation of conduction systems. In Stage 2 and 3, the causes are ischaemic heart disease, rupture of an aneurysm (rare), and myocarditis. In Stage 4, there may be ischaemic heart disease, and, in rare cases, heart failure due to mitral insufficiency.

Epidemiology

The first nationwide survey was conducted in 1970 by the Japanese Kawasaki disease Research Committee. Since then 15 nationwide surveys have been carried out at 2-year intervals up to December 1998. A total of 153 803 cases (89 272 males and 64 531 females M:F ratio of 1.38:1) has been reported, including 426 (0.28 per cent) deaths. The number of cases reported has been steadily increasing since 1971. There were outbreaks in 1979, 1982, and 1986, when a high incidence of the disease was reported in the early or late spring throughout Japan. A shift of the epidemic wave from warm to cool geographical areas was observed in 1979, but not in 1982 and 1986.

Since 1974, a number of cases have been reported from outside of Japan. Kawasaki disease is now known to have a worldwide distribution, having been observed in all continents and in all ethnic groups. It currently ranks as the leading cause of acquired heart disease in the paediatric population of the United States and Japan.

Aetiology

It is not proven, but the clinical and epidemiological features of Kawasaki disease strongly suggest that it is caused by an infectious agent; one to which the great majority of people become immunized in early life by subclinical infection. The spacing between waves of the disease is determined by the build-up of a new group of susceptible individuals.

The rarity of the illness in the first few months of life and in older children and adults, the low incidence of disease in siblings, and the absence of person-to-person transmission, are all compatible with infection by a ubiquitous agent, to which virtually all adults are immune and from which very young children are protected by passive maternal antibody. The majority of infected individuals probably experience an asymptomatic infection, whilst a select few develop the recognizable clinical features of Kawasaki disease.

Treatment and management

Therapy with the combination of intravenous immunoglobulin (IVIG) and aspirin during the acute phase of Kawasaki disease produces a more marked anti-inflammatory effect and reduction in coronary artery abnormalities than does aspirin alone. It is recommended that patients with acute disease be treated with a single 2 g/kg infusion of IVIG and aspirin (30–50 mg/kg per day) within the first 10 days from onset, and that the aspirin dose be reduced to 3–5 mg/kg per day given as a single daily dose after defervescence. Aspirin is discontinued if no coronary abnormalities have been detected by echocardiography by 6 to 8 weeks after the onset of illness, but continued if coronary artery abnormalities are present. Aspirin should be discontinued if the patient develops an illness suspected to be varicella or influenza, this is to reduce the risk of Reye's syndrome, when the use of an alternative antiplatelet agent should be considered.

Approximately 10 per cent of patients with Kawasaki disease are resistant to IVIG therapy. These patients are at greatest risk for the development of coronary artery aneurysms and long-term sequelae of the disease. As in other vasculitides, blood vessel damage appears to result from an aberrant immune response leading to endothelial cell injury and vessel wall damage. Steroids such as methylprednisolone are the treatment of choice in other forms of vasculitis, yet they have been considered to be unsafe in patients with Kawasaki disease. However, it has been reported that at least some patients with severe disease, resistant to IVIG therapy, may be safely treated with intravenous pulse-steroid therapy and benefit from this treatment.

The management of patients with severe obstructive coronary artery disease, who may develop symptomatic ischaemic heart disease, is an important issue. Patients must be immediately admitted to hospital when myocardial infarction occurs. Management strategies employed are those well defined in the context of atheromatous coronary artery disease. Massive thrombus formation can be visualized by serial echocardiographic studies and is a clear indication for anticoagulation. Recurrence of infarction, which is associated with high mortality, occurs in approximately 20 per cent of patients, and emphasizes the need for the careful management of children with myocardial infarction, even if this is silent.

Patients with cardiac complications or sequelae, such as ventricular dysfunction, heart failure, severe arrhythmias, or postinfarction angina, are managed by conventional medical and/or surgical techniques. As a prelude to surgical treatment, detailed coronary angiography is essential, and viability of the myocardium should be evaluated by thallium scintigraphy. Long-term results and prognosis after surgery remain uncertain. Heart transplantation of patients with Kawasaki disease has been performed in 15 cases.

Further reading

Burns JC, *et al*. (1985). Anterior uveitis associated with Kawasaki syndrome. *Pediatric Infectious Disease* **4**, 258–61.

Checchia P, *et al*. (1995). The worldwide experience with cardiac transplantation for Kawasaki disease. In: Kato H, ed. *Kawasaki disease, Excerpta Medica International Congress Series 1093*, pp 522–6. BV Elsevier Science. Amsterdam.

Dajani AS, *et al*. (1993). Diagnosis and therapy of Kawasaki disease in children. *Circulation* **87**, 1776–80.

Fujiwara H, Hamashima Y (1978). Pathology of the heart in Kawasaki disease. *Pediatrics* **61**, 100–7.

Furusho K, *et al*. (1984). High-dose intravenous gammaglobulin for Kawasaki disease. *Lancet* **2**, 1055–8.

Kato H, Akagi T (1997). Ischemic heart disease in Kawasaki disease. *Progress in Pediatric Cardiology* **6**, 219–26.

Kawasaki T (1967). Acute febrile mucocutaneous syndrome with lymphoid involvement with specific desquamation of the fingers and toes in children. *Japanese Journal of Allergology* **16**, 178–222. [In Japanese.]

Kawasaki T, *et al*. (1974). A new infantile acute febrile mucocutaneous lymph node syndrome (MLNS) prevailing in Japan. *Pediatrics* **54**, 271–6.

Kitamura S, *et al*. (1994). Long-term outcome of myocardial revascularization in patients with Kawasaki coronary artery disease, a multicenter cooperative study. *Journal of Thoracic and Cardiovascular Surgery* **107**, 663–74.

Landing BH, Larson EJ (1987). Pathological features of Kawasaki disease (mucocutaneous lymph node syndrome). *American Journal of Cardiovascular Pathology* **1**, 215–29.

Newburger JW, *et al*. (1991). A single intravenous infusion of gamma globulin as compared with four infusions in the treatment of acute Kawasaki syndrome. *New England Journal of Medicine* **324**, 1633–9.

Shulman ST, Rowley AH (1997). Etiology and pathogenesis of Kawasaki disease. *Progress in Pediatric Cardiology* **6**, 187–92.

Sundel RP, Newburger JW (1997). Management of acute Kawasaki disease. *Progress in Pediatric Cardiology* **6**, 203–9.

Suzuki A, *et al*. (1997). Natural history of coronary artery lesions in Kawasaki disease. *Progress in Pediatric Cardiology* **6**, 211–18.

Tanaka N, Sekimoto K, Naoe S (1976). Kawasaki disease: relationship with infantile periarteritis nodosa. *Archives of Pathology and Laboratory Medicine* **100**, 81–6.

Taubert KA (1997). Epidemiology of Kawasaki disease in the United States and worldwide. *Progress in Pediatric Cardiology* **6**, 181–5.

Yanagawa H, *et al*. (1995). Results of 12 nationwide epidemiological incidence surveys of Kawasaki disease in Japan. *Archives of Pediatrics and Adolescent Medicine*, **149**, 779–83.

Yanagawa H, *et al*. (2000). Results of the 15th nationwide survey on Kawasaki disease in Japan. *Shonika Sinryo* **63**, 121–32. [In Japanese.]

18.11 Miscellaneous conditions presenting to the rheumatologist

D. O'Gradaigh and B. Hazleman

Musculoskeletal symptoms can occur in a wide range of diseases, or as a paraneoplastic manifestation or drug side-effect. Careful assessment of the history, physical signs, and investigation results are required to identify significant underlying conditions that may first present to the rheumatologist. A number of uncommon conditions can present with non-specific musculoskeletal manifestations and are also discussed in this chapter.

Adult Still's disease

In 1971, Bywaters described a series of 14 adults with an illness very similar to the systemic onset-type of juvenile idiopathic arthritis described by Still in 1897. Adult-onset Still's disease is found worldwide with an incidence of 1-3 per million, most commonly in the age range 16 to 35 years and affecting males and females equally in most populations. There is no consistent HLA association.

Features common to both childhood and adult-onset forms are the high, spiking pyrexia, arthralgia or arthritis, and a characteristic rash. The fever typically appears in the evening, and a patient with pyrexia of unknown origin should always be assessed at least once at the end of the day. Spikes in excess of 39°C are typical (and required in diagnostic criteria), though a return to a normal temperature does not occur in 20 per cent. Arthralgia is almost universal and may intensify during the febrile episodes. Distal interphalangeal joint involvement, seen in one in five patients, is useful to distinguish from other inflammatory arthropathies. The classical 'Still's rash' is a maculopapular, salmon-pink rash on the trunk, thighs, and arms or axillae that appears during the temperature spike (termed 'evanescent'). The rash may also appear on the face, palms and soles, and at sites of skin trauma (Koebner phenomenon) in a third of adults. A (culture-negative) severe sore throat is relatively common in adults (though not a feature of the juvenile form).

Other common manifestations are hepatosplenomegaly with or without generalized lymphadenopathy, and polyserositis, of which pericarditis (in a third) and pleuritis are the most common. Rare features include sicca symptoms (dry eyes, mouth), myocarditis, restrictive lung disease, liver or renal failure, panophthalmitis or inflammatory orbital pseudotumour, epilepsy, intravascular coagulopathy or haemophagocytic syndrome, and amyloidosis.

Diagnosis is primarily clinical, it being important to remember that the classical features may only emerge over a period of time, and the possibility of Still's disease may need to be reconsidered as symptoms progress. The differential diagnosis is wide, and while diagnostic criteria have been proposed they have poor sensitivity and specificity until infection (particularly infectious mononucleosis), neoplasia (lymphomas), and connective tissue diseases (such as polyarteritis nodasa and systemic rheumatoid vasculitis) have been excluded.

There are no specific laboratory features, but typical findings include elevated ESR and CRP, thrombocytosis, neutrophil leucocytosis (total leucocytes in excess of 15×10^9/l) and a normochromic normocytic anaemia. Elevated liver enzymes can be found, and may rise further during non-steroidal anti-inflammatory treatment. Both rheumatoid factor and antinuclear antibodies are negative in most cases. A highly elevated (>5 times upper limit) serum ferritin is a useful marker to discriminate from other arthropathies, but is not sufficiently specific to exclude the differentials (especially neoplasia) mentioned above.

Indomethacin (or another non-steroidal anti-inflammatory agent) has largely replaced high-dose (100 mg/kg/day) salicylate as the first line treatment for fever and systemic features. If these agents fail individually, they may be given together, but adequate prophylaxis against peptic ulceration is essential. Corticosteroid is required in two-thirds of cases, and should be initiated without delay in cases of myocarditis, pericardial tamponade, or other severe organ involvement. Doses of prednisolone in the range 0.5–1 mg/kg/day are usually given, and should be continued for 2 to 3 months after remission before gradually tapering the dose.

The role of disease modifying antirheumatic drugs or cytotoxic agents is not established. In refractory cases with systemic features, or as steroid-sparing therapy, methotrexate is particularly useful. Salazopyrin, azothioprine, and intravenous immunoglobulin have also been used. Intramuscular gold is appropriate when arthritis dominates. Most recently, the antitumour necrosis factor-a therapies have been used with some success.

Prognosis is variable. A chronic progressive arthritis is predicted by early arthritis (rather than arthralgia), particularly of the hip and shoulder, and occurs in 30 to 50 per cent of cases, with ankylosis of the carpus and tarsus and involvement of the cervical spine and hips. Equal proportions of the remainder experience either a self-limiting course (lasting up to 1 year), or a polycyclic, relapsing, and remitting course. The rash, fever, and serositis are typically less severe in subsequent relapses, and complete remissions up to 10 years after first presentation have been recorded.

Acne arthralgia

Patients may complain of myalgia, arthralgia, or swelling, typically involving the large joints. Most patients are male adolescents with aggressive acne. *Propionibacterium acnes* has been isolated from joint aspirates; however, effusions are typically sterile and the arthritis is believed to be reactive rather than septic. Hydradenitis suppurativa, producing large abscesses in the axilla and groin, is also associated with a reactive type of large-joint oligoarthropathy. In both conditions, symptoms usually improve with treatment of the skin lesion. A seronegative spondyloarthropathy syndrome of acne, palmoplantar pustulosis, hyperostosis (especially

of the clavicles or sternum), and (sterile) osteomyelitis (**SAPHO**) is associated with enthesitis and an inflammatory polyarthritis that often includes the sacroiliac joints.

Neutrophilic dermatoses

The neutrophilic dermatoses include pyoderma gangrenosum and Sweet's syndrome (acute febrile neutrophilic dermatosis). Erythema nodosum is now considered part of this spectrum.

Pyoderma gangrenosum is a reactive neutrophilic dermatosis associated with ulcerative colitis, rheumatoid arthritis, and monoclonal gammopathies or other haematological malignancies, which produces painful ulcerative skin lesions (Fig. 1 and Plate 1). Approximately 30 per cent of patients describe arthralgia or a seronegative, progressive, erosive polyarthritis. Treatment usually requires corticosteroid therapy in addition to that for the underlying disorder.

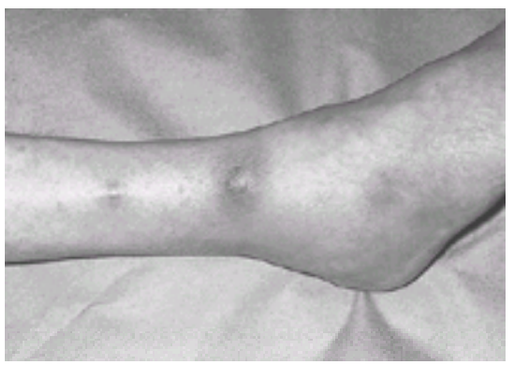

Fig. 1 Pyoderma gangrenosum. (See also Plate 1.)

Sweet's syndrome presents with tender red or purple raised nodules associated with fever and generalized myalgia and/or arthralgia. Joint effusions may occur, and aspirates reveal high neutrophil counts. Sterile osteomyelitic foci have rarely been described. Skin biopsy is diagnostic. Symptoms typically resolve over 2 to 3 months, requiring symptomatic treatment with a non-steroidal anti-inflammatory drug (**NSAID**) or intra-articular steroid. An association with acute myeloid (particularly premyelocytic) leukaemia is noted in about 15 per cent of cases, and recombinant granulocyte colony-stimulating factor (**rG-CSF**) has also been implicated in a number of cases.

Erythema nodosum presents with discrete nodules on the extensor aspect of the lower leg and less commonly on the upper limbs (Fig. 2 and Plate 2). Joint manifestations (arthralgia in two-thirds) occur in 75 per cent of cases. Arthritis with synovial thickening and joint effusions usually affects the knee and ankle symmetrically. The small joints of the hands, wrists, elbows, and shoulders are less commonly affected. Various underlying conditions, infections, and drugs are associated with erythema nodosum. Whilst their recognition is important, the arthritis is usually self-limiting, responding as it does to treatment with NSAIDs, although corticosteroids are occasionally required, and resolving without sequelae.

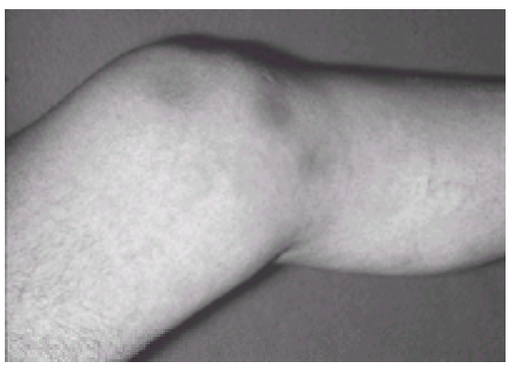

Fig. 2 Erythema nodosum. (See also Plate 2.)

Panniculitis

Also called lupus erythematosus profundus, this is an unusual variation of cutaneous lupus characterized by recurrent inflammation of subcutaneous tissue leading to fibrosis. Asymptomatic, firm, sharply defined subcutaneous nodules or plaques appear on the proximal upper and lower limbs, buttocks, face, and scalp. Histology reveals a non-specific lobular panniculitis with necrobiosis of adipose tissue and fibrotic deposits. Some one in eight patients have systemic lupus erythematosus (**SLE**) at presentation, particularly generalized arthralgia and fatigue. A further 10 to 15 per cent will develop SLE up to 10 years later. Skin and joint features are treated with hydroxychloroquine, though a steroid and dapsone are occasionally required for more florid panniculitis.

Multicentric reticulohistiocytosis

This is a rare systemic disease of unknown aetiology, with infiltration of lipid-laden histiocytes and multinucleate giant cells into various organs. Skin nodules and a rapidly progressive deforming arthritis are the most frequently recognized features. Light copper or red–brown nodules appear on the face and hands, but can appear anywhere, and may number from a few to several hundred. The disease affects middle-aged women who typically present with an insidious onset of polyarthritis in the interphalangeal joints. The spine and other joints may be involved. There is no satisfactory treatment. Underlying malignancy is reported in 20 to 30 per cent of cases.

Sarcoidosis

One-third of patients with sarcoid will have musculoskeletal features, of which an acute symmetrical polyarthritis is the most common. Lofgren's syndrome associates this pattern of arthritis with erythema nodosum and bilateral hilar lymphadenopathy. The ankles and feet are the most commonly affected, followed by the hands, wrists, and elbows. Symptoms develop rapidly and usually respond well to NSAIDs, although corticosteroids are occasionally necessary. Remission without joint destruction occurs after 2 to 4 months. A chronic arthropathy is uncommon, occurring mostly in those with multiorgan involvement and those requiring steroid treatment during the acute phase. Dactylitis, joint space narrowing, and osseous involvement are then the most common features, with superimposed episodes of acute arthritis. Radiographic findings appear late, and include acro-osteolysis, cystic lesions, or a reticulated coarse trabecular pattern in the phalanges. Corticosteroid in doses of 30 to 60 mg are the mainstay of treatment at this stage, but is frequently disappointing, with recurrence of symptoms on cessation of therapy.

Amyloidosis

Musculoskeletal features occur in three main settings. Dialysis-related amyloidosis is due to the accumulation of b $_2$-microglobulin. Synovitis usually involves large joints such as the hip and shoulder. Magnetic resonance imaging (**MRI**) may show characteristic features, and joint fluid aspiration may identify amyloid deposits, particularly using the more sensitive combination of Congo Red staining and immunocytochemistry. Symptomatic treatment with an NSAID is usually sufficient, considerations of the effect of NSAIDs on renal function only being relevant in those dialysis patients with substantial urine output, but the condition can be disabling and refractory. Significant improvement often follows transplantation. Cystic (lytic) bone lesions are typically painless, and present difficulties in the differential

diagnosis. They may be complicated by pathological fracture. Soft tissue amyloid deposits usually present with entrapment neuropathies such as carpal tunnel syndrome.

In primary (AL) amyloidosis, a symmetrical polyarthritis with synovitis and morning stiffness involves large and small joints. Radiographic changes include osteoporosis and, less commonly, joint erosions. Diagnostic confusion can arise, as frequently the erythrocyte sedimentation rate is not significantly elevated. The synovitis is often described as 'pasty', and flexion contractures occur relatively early. The early appearance of carpal tunnel syndrome should also raise the suspicion of underlying amyloidosis. In addition to treatment addressing the underlying paraproteinaemia, joint symptoms may require therapy with a corticosteroid.

Familial amyloidosis and secondary (AA) amyloid due to persistent inflammation are not usually associated with rheumatological symptoms. Exceptions include the Muckle–Wells syndrome of urticaria, deafness, arthritis, and amyloid nephropathy.

Familial Mediterranean fever (FMF)

This is an autosomal recessive disorder appearing in people of Armenian, Arab, and Sephardic Jewish descent. A number of genetic defects have been localized to the *marenostrin* or *pyrin* genes on chromosome 16p. The function of this protein is not known, but the *M694V* mutation is particularly associated with joint involvement. Presenting in childhood with episodes of fever and abdominal pain, synovitis occurs in 75 per cent of cases. Monoarticular involvement of a knee or ankle, or symmetrical involvement of these joints, are the most common of the six patterns of joint involvement described. A symmetrical polyarthritis indistinguishable from juvenile idiopathic arthritis often causes diagnostic confusion, particularly as fever and abdominal pain are not uncommon in this condition. The pattern tends to be similar in subsequent episodes, and despite frequent florid synovitis, residual damage rarely occurs. Episodes typically last for less than 1 week, though more protracted attacks may persist for months. Treatment with colchicine has almost eliminated amyloidosis as a complication of this condition and it also reduces the frequency of symptom relapse. However, it is ineffective once an episode has started, and an NSAID and rest are then the most effective measures. Interferon-a has been used to good effect in resistant cases.

Another two periodic fever syndromes—hyperimmunoglobulin D syndrome (**HIDS**) and familial Hibernian fever (also called autosomal dominant recurrent fever)—occur in The Netherlands and Northern France, and in a few Irish and Scottish families, respectively. To date, neither has been associated with the joint symptoms of FMF. Frequent mouth ulcers resembling Behçet's syndrome appear during attacks of HIDS. However, the markedly elevated levels of IgD and IgA in the latter are diagnostic. The place of colchicine is not yet established in the treatment of these disorders.

Haematological disorders

Leukaemia, lymphoma, and uncommon lymphoproliferative disorders

Between 13 and 60 per cent of patients with acute leukaemia will develop arthralgia or less commonly a frank arthritis. Monoarthritis, symmetrical polyarthritis, and a large joint oligoarthropathy are described. Diagnostic clues include a disproportionate amount of pain, fever, and weight loss, though in children the latter may be mistaken for Still's disease. Arthralgia is an uncommon feature of lymphomas; however, 7 to 25 per cent of patients with non-Hodgkin's lymphoma experience polyarthralgia, secondary gout, or hypertrophic pulmonary osteoarthropathy (see below) during the course of their disease.

Large granular lymphocyte syndrome is a monoclonal expansion of T cells associated with a variety of conditions including rheumatoid arthritis (in one-third of cases). Both neutropenia and splenomegaly can occur, mimicking Felty's syndrome. Some consider it to be indistinguishable in every respect, including its management.

Human T-cell lymphotrophic virus-1 (**HTLV-1**) is associated with the development of leukaemia or lymphoma, and may independently produce a symmetrical polyarthritis closely resembling rheumatoid arthritis.

Haemophilia

Prophylactic, factor replacement between 2 and 18 years of age is cost-effective in preventing disabling joint complications. Without this, acute haemarthroses begin from around 5 years of age, causing recurring episodes of very painful and tender joint swelling, particularly in the hinge joints such as the knee, ankle, and elbow (presumably because these joints are less tolerant of angular or rotational strain). Pain is increased by the additional irritant effect of blood on the synovium. Ultrasonography is useful in the differentiation of haemarthrosis from soft tissue or subperiosteal haemorrhage. Repeated episodes, without appropriate treatment, result in persistent synovitis and joint contracture. Early coagulation factor replacement, ice, joint immobilization, and elevation all reduce further bleeding. Joint aspiration may also be required (after adequate factor replacement). Rehabilitation is required to prevent contraction. Synovectomy by an intra-articular injection of radioactive isotope is a useful treatment in cases of chronic synovitis, but joint replacement continues to be needed where disabling secondary degenerative arthritis has occurred. Acute haemarthrosis due to disseminated intravascular coagulation should be similarly managed.

Cryoglobulinaemia

Cryoglobulins are immune complexes that precipitate spontaneously at low temperatures. Type I (25 per cent) comprises a monoclonal immunoglobulin and is associated with lymphoproliferative disorders including myeloma and Waldenström's macroglobulinaemia. Type II (25 per cent) complexes a monoclonal immunoglobulin, usually of IgM class, with a polyclonal anti-immunoglobulin typically of IgG type (that is, a rheumatoid factor). Previously called mixed essential cryoglobulinaemia, it is now recognized that over 90 per cent of these individuals have serological evidence of hepatitis C virus (HCV) infection. Type III accounts for 50 per cent of cases, and is a complex of two polyclonal immunoglobulins, usually occurring as a paraneoplastic phenomenon.

Precipitation of cryoglobulin leads to complement activation and vasculitis in small vessels. Complete vascular occlusion is less common. A classical triad of a palpable purpuric rash on the extremities, arthralgia, and muscle weakness is described. Joints are involved in 70 per cent of patients in a relapsing and remitting pattern, affecting, in order of frequency, the hands, knees, ankles, and elbows. Inflammatory arthritis is uncommon, and radiological changes do not occur. Other skin presentations include petechiae, urticaria, and acrocyanosis. Other organ involvement is frequently seen in addition to this triad, particularly glomerulonephritis.

Diagnosis requires meticulous attention to phlebotomy and laboratory techniques. A positive rheumatoid factor and raised ESR are supportive features, and urinalysis and microscopy, looking for an 'active sediment' (proteinuria, haematuria, and red cell casts), should always be carried out in patients presenting with purpura and arthralgia. A thorough search is required for underlying malignancy and for associated HCV infection.

Treatment is directed at any underlying cause. NSAIDs, corticosteroids, and steroid-sparing drugs, particularly azathioprine, are used to relieve arthralgia and to prevent the progression of purpura to ulceration. Neurological or renal involvement requires more aggressive therapy, for which cyclophosphamide (oral or intravenous pulses) or chlorambucil are used. Plasmapharesis is sometimes considered for those with rapidly progressive glomerulonephritis, but requires particular care to avoid blood cooling in the extracorporeal circuit.

POEMS

This is an uncommon disorder that may present to any specialty, depending on the dominant feature in the spectrum of **p**olyneuropathy, **o**rganomegaly, **e**ndocrinopathy, **M**-protein (i.e. a monoclonal paraproteinaemia), and **s**kin abnormalities. Skin changes may resemble scleroderma. Radiographs show single or multiple osteosclerotic lesions with unusual patterns of proliferative change, both of which are unexpected in myeloma. The diagnosis of POEMS should therefore be considered in those presenting with osteosclerotic lesions accompanied by paraproteinaemia, particularly when associated with peripheral neuropathy. Treatment must be directed at the principal presenting features—bone lesions are rarely symptomatic unless they result in bone swelling or fracture.

Hypogammaglobulinaemia

Primary hypogammaglobulinaemia is associated in 10 to 30 per cent of patients with a non-erosive polyarthritis resembling rheumatoid arthritis. Features include morning stiffness, pain, and tender swelling in the peripheral joints. Subcutaneous nodules may appear. However, rheumatoid factor is negative, histology reveals the absence of plasma cells, and permanent joint damage is rare. Synovitis may be transient or it may persist for many years, requiring symptomatic treatment. Intra-articular corticosteroid treatment is used, though these patients are somewhat more at risk of septic arthritis. In the absence of any intra-articular procedure, the

rate of septic arthritis is approximately 20 per cent over 20 years.

Sickle-cell disease (SCD)

Sickle-cell crises commonly include bone pain. The cause is believed to be intramedullary hypertension due to vascular occlusion resulting in bone ischaemia. Vasodilator drugs have been used with varied results. Avascular necrosis is less commonly associated with other haemoglobinopathies. Synovitis, frequently complicated by haemarthrosis, usually occurs during crises, and is due to synovial infarction. The effusion is non-inflammatory. Osteomyelitis may complicate avascular necrosis due to SCD, *Salmonella* spp. being particularly common. However, septic arthritis is unusual. Hyperuricaemia and gout occur in 40 per cent of adults with SCD, and is treated in the standard way. Less commonly, a hand-and-foot syndrome affects infants aged between 6 months and 2 years, dactylitis and periostitis producing symmetrical tender, diffuse swelling and stiffness lasting several weeks.

Gastroenterological and metabolic conditions

Hepatitis

The common viral hepatitides, hepatitis-A, -B, and -C viruses (**HAV**, **HBV**, and **HCV**, respectively) are associated with a serum-sickness during their prodromal phase. Early morning stiffness and mild arthralgia, or, less commonly, inflammatory arthritis, affect the small joints of the hands, and, in decreasing order of frequency, the knees, ankles, shoulders, wrists, and feet. The spine and hips are not usually affected. Symptoms typically resolve as hepatitis evolves. Less common features include a leucocytoclastic ('hypersensitivity') vasculitis in HAV and an association with polyarteritis nodosa in HBV. HCV is associated with cryoglobulinaemia (in 50 per cent) and with antiphospholipid antibodies and thrombosis.

Enteropathies

Coeliac disease may result in osteoporosis or osteomalacia with bone pain and pathological fracture. Arthritis is uncommon, but can precede overt bowel symptoms by up to 3 years. Symmetrical involvement with swelling and stiffness can affect the lumbar spine, hips, knees, and shoulders. Dermatitis herpetiformis is more common among those with joint involvement. The joint manifestations resolve on changing to a gluten-free diet, and do not reappear on rechallenge with gluten.

Whipple's disease presents with fever and abdominal pain. Acute or subacute migratory polyarthritis may precede bowel symptoms by years, and typically involves the ankles, knees, shoulders, and elbows. Lymphadenopathy is a prominent feature. Duodenal biopsy shows Periodic acid–Schiff (**PAS**) -staining macrophages, and the polymerase chain reaction (**PCR**) detects the causative organism *Tropheryma whippelii*. Its presence within cells implies a mechanism similar to reactive arthritis with a T_H2-dominant response and inability of the T_H1-cellular immune response to clear the microbe from macrophages, which therefore perpetuate the inflammatory reaction.

Surgical procedures that bypass a section of (proximal) small bowel are associated with so-called 'bypass' arthritis. This ranges from a mono- or oligoarthropathy to a diffuse polyarthritis involving large and small joints. Tenosynovitis of the wrist is a particularly common feature. Treatment is symptomatic, though sulfasalazine is occasionally used as a disease-modifying therapy.

Haemochromatosis

An autosomal recessive inherited disorder of iron transportation and storage, this condition may present up to 10 years before the underlying condition is recognized, usually in men between 50 and 60 years of age, with a painful inflammatory synovitis principally affecting the second and third metacarpophalangeal joints. About 80 per cent of those with genetically identified haemochromatosis will develop arthritis at some point. Acute exacerbations may occur due to calcium pyrophosphate dihydrate deposition, particularly in the knee and wrist. The reason for this is unclear, though iron is known to inhibit the clearance of pyrophosphate from the synovial lining layer. Crystals may be identified on polarized light-microscopy of a joint aspirate. Their presence is implied by radiographic evidence of chondrocalcinosis. Radiographs of the hands may also show cyst formation in the affected joints, with erosive changes and characteristically hook-shaped osteophytes. Treatment is symptomatic, intra-articular corticosteroid therapy being of value during acute episodes. Arthritis persists in the majority of cases, and can be relentlessly progressive, despite regular venesection.

Wilson's disease

A disorder of copper metabolism, this condition typically presents in childhood with neurological problems. Some two-thirds of patients with Wilson's disease will develop musculoskeletal manifestations, half of whom being symptomatic by 15 years of age. Features include arthritis (primary, attributed to copper deposition in synovium; or secondary, due to chondrocalcinosis), rhabdomyolysis, hypermobility (due to effects on collagen synthesis), and osteopenia. Radiographic appearances are generally non-specific, with joint space narrowing, sclerosis, and cyst formation. A fluffy periostitis at the greater trochanter and inferior aspect of the calcaneus, and corticated ossicles near affected joints (particularly the wrist) are characteristic but rare. Diagnosis requires measuring urinary 24-hour copper excretion: caeruloplasmin may be elevated as part of an acute-phase response and therefore is of no diagnostic value in presentations with acute arthritis. Penicillamine is the mainstay of treatment for this condition and alleviates joint symptoms.

Ochronosis

Deficiency of the enzyme homogentisic acid oxidase results in an accumulation of this organic acid. Though a congenital disorder, symptoms rarely appear until the fourth decade. The classical clinical features of pigmentation of the ear and sclera, and urine darkening on standing (giving the alternative name alkaptonuria) allow easy diagnosis. Deposition also occurs in the synovium and may appear in joint fluid aspirate. Pain and swelling affect the large joints, and the thoracolumbar spine is also affected producing pain and stiffness, but the lumbosacral spine is spared. Radiographs show chondrocalcinosis of the intervertebral discs with spondylosis that may progress to ankylosis. In the peripheral joints, radiographic changes of degeneration appear, though osteophytes are often less marked than in other degenerative arthritides. Erosion may occur. Recent treatment efforts are concentrated on early genetic diagnosis, dietary advice, and the possibility in the future of gene therapy.

Hyperlipidaemia

Articular symptoms can occur in a number of the hyperlipidaemias, particularly types II and IV. Joint manifestations typically precede diagnosis of the lipoprotein disorder. Xanthomas of tendons are a useful clue, but the clinical picture is otherwise non-specific. Morning stiffness, pain, and tenderness are noted, but overt joint inflammation is uncommon. A migratory polyarthritis is occasionally described in type II hyperlipoproteinaemia, but oligoarthritis and tendonitis are more common. Tendon xanthomas may result in periarticular bone cyst formation. In two-thirds of patients, symptoms resolve with treatment of the lipid disorder, the remainder requiring symptomatic therapy.

Musculoskeletal manifestations of HIV/AIDS

Rheumatological manifestations include serum-sickness at seroconversion, pyomyositis, and osteomyelitis (particularly in the setting of intravenous drug abuse) and a spectrum of presentations with acute arthropathy. Antiretroviral therapy, particularly zidovudine, may produce a polymyositis with ragged red fibres on muscle biopsy. Very rare manifestations include a vasculitis that appears to be directly induced by the virus, and hypertrophic osteoarthropathy (see below) secondary to *Pneumocystis carinii* pneumonia. Arthritis or arthralgia occur in 1 to 25 per cent of cases. Spondyloarthropathy with dactylitis and enthesitis is the most common. A severe, but self-limiting, large joint oligoarthritis has a predilection for the knees and ankles, resolving over 2 to 6 weeks and responding well to NSAIDs. A generalized articular syndrome is very short-lived though intensely painful, usually lasting only 24 h. Acute symmetrical polyarthritis is relatively uncommon, and septic arthritis is rare.

Reflex sympathetic dystrophy

This is one of a number of terms (including algodystrophy and Sudeck's atrophy) for a condition that has recently been renamed 'chronic regional pain syndrome' (**CRPS**). The dominant feature is pain, with allodynia (pain in response to an innocuous stimuli), hyperalgesia (increased pain perception), and hyperpathia (an

exaggerated delayed reaction). Pain usually involves a single limb or body region, typically distal to the site of some traumatic event, but a definite precipitant is recognized in only 50 per cent of patients. Marked joint stiffness and pain on movement cause considerable disability. CRPS may also follow myocardial infarction, stroke, pregnancy, or deep venous thrombosis. There is an association with HLA-DR2. Other cardinal features relate to excessive activity of the sympathetic nervous system, with localized swelling, sweating, and piloerection in the early stages, the skin often appearing stretched and shiny. Hyperaemia is believed to be responsible for osteopenia in the affected part.

Diagnosis is largely clinical, though diffuse osteopenia on plain radiography (comparing the symptomatic and normal limbs on the same film) and the absence of an acute-phase response are supportive. Bone scintigraphy offers the most reliable confirmation of the clinical impression. A three-phase scan is required, comparing the symptomatic and normal sides in the early blood phase (demonstrating hyperaemia in the affected part), the bone pool phase (increased bone turnover), and delayed phase. Physiotherapy is the key element of treatment and must be quite intensive initially. However, pain may limit patient co-operation. Sympathetic nerve blocks (stellate ganglion or lumbar sympathetic chain) with long-acting anaesthetic and/or guanethidine are specialized techniques that are often quite effective. Other pain-relieving modalities include corticosteroid injection to the involved joint, subcutaneous or intranasal calcitonin, intravenous pamidronate, and (oral) gabapentin.

Charcot's arthropathy

This is a disorder of joint destruction associated with neurological injury or damage. Diabetes mellitus, tabes dorsalis, and syringomyelia are the most commonly associated diseases. There are two leading pathogenesis theories. The more widely accepted neurotraumatic theory suggests that loss of sensation allows repeated subclinical trauma, culminating in a destructive arthropathy. The neurovascular theory is proposed to explain how the arthropathy can appear very early in the absence of use of the limb, and the observation, particularly in the tarsus of diabetic patients, that the arthropathy can be quite painful. Here, it is believed that damage to 'trophic centres' results in altered vascular supply and, hence, impaired bone and cartilage nutrition underlying the subsequent joint damage. Radiologically, a gross proliferative osteoarthrosis is most commonly seen, but significant resorption of bone can also feature, and stress fractures occur in up to one-third of patients. Painfree joints rarely require treatment; moreover, orthopaedic procedures are associated with a high failure rate. Management of painful neuropathic joints is very difficult. Orthoses help to prevent stressing of related soft-tissue structures, and a broad range of analgesics, including amitriptyline, should be considered.

Tietze's syndrome/chostochondritis

Both conditions are of unknown aetiology, though a viral trigger has been proposed in Tietze's syndrome (chondropathia tuberosa). A single chostochondral joint (usually the second or third) is involved in 80 per cent of patients. Coughing or deep breathing exacerbates paracentral chest pain. Tietze's syndrome is also associated with firm, tender lumps at the affected sites. Onset may be acute or more gradual, and the subsequent course is similarly variable, ranging from spontaneous remission to prolonged symptoms lasting for years. As these conditions typically affect middle-aged women, a visceral origin for the symptoms must not be overlooked. Local injection with lidocaine (lignocaine) or a corticosteroid may provide symptomatic relief when necessary.

Miscellaneous disorders of synovium, bone, cartilage, and calcification

Synovium

Pigmented villonodular synovitis (PVNS)

This is a benign synovial hyperplasia of unknown aetiology, possibly reactive or neoplastic. To date, three types of PVNS have been described. Giant-cell tumour of the tendon sheath occurs most commonly in extensor tendons of the hand; although painless, large nodules may restrict movement. Treatment is by surgical excision, which allows histological confirmation of the diagnosis. Recurrence is rare. Isolated nodular and true diffuse PVNS are intra-articular lesions occurring most commonly in the knee of adult males aged between 20 and 50 years. Pain, swelling, and a gradual reduction in the range of movement can continue for some years before the diagnosis is made. Aspiration of erosanguinous fluid in the absence of trauma should raise the suspicion. Intra-articular steroid administration gives effective but short-lived relief, and surgical excision is the treatment of choice as it also allows a definitive diagnosis. In the event of a recurrence (uncommon except in the diffuse form), radioisotope synovectomy or radiotherapy may be used. In late stages, haemosiderin deposition and chronic inflammation can lead to destructive changes requiring arthroplasty.

Synovial (osteo-)chondromatosis (Reichel syndrome)

A benign synovial proliferation, this is probably caused by reactive metaplasia secondary to osteoarthrosis, osteochondrosis, or other joint pathology. Most patients are men in their third to fifth decade. Typically monoarticular, usually in the knee, symptoms include joint swelling, locking, and giving way, suggestive of intra-articular loose bodies. Multiple (up to 200) calcified periarticular bodies of hyaline cartilage, 1 mm to 3 cm in size but usually uniform, fill the joint. Surgery is required to remove loose bodies. Rarely, malignant transformation to chondrosarcoma may occur.

Synovial haemangioma

A synovial haemangioma is a benign lesion comprising vascular and non-vascular tissue in an asymptomatic and well-localized intra-articular mass, most commonly in the knee (60 per cent) and elbow (30 per cent). Surgical excision is curative.

Lipomas are most commonly found in the thenar and hypothenar eminences producing compressive symptoms. They may calcify or undergo fibrosis and infarction. Lipoma arborescens occurs particularly in the suprapatellar bursa, producing painless swelling. MRI changes are diagnostic and surgery is curative.

Some two-thirds of 'synovial sarcomas' arise in the thigh. The tissue of origin is mesenchymal, with differentiation to synovium. Prognosis is poor despite surgical excision and radiotherapy.

Bone and cartilage

Bone cysts may be symptomatic or arise as incidental findings, thereby causing diagnostic difficulty. Cysts may be aneurysmal (primary or secondary) or simple (also called unicameral) and can appear in children or adults. Simple cysts are rarely symptomatic or complicated by fracture, and management is expectant. Aneurysmal bone cysts are rare (1 per million), non-neoplastic expansile lesions occurring principally in the metaphysis of long bones (50 per cent), the posterior part of vertebrae (30 per cent), or in the flat bones, particularly the pelvis. Most present with pain, swelling, or pathological fracture at a mean of 13 years of age. Radiological features that suggest the diagnosis include an eccentric location of a cyst containing fluid-fluid levels and trabeculae which remain distinct within it. Management has evolved from the mainstay of curettage with bone grafting or implanting autologous marrow (rich in osteoblasts) to intralesional corticosteroid injection. However, recurrence rates are high (20 to 50 per cent) and other options include embolization and radiotherapy. Secondary aneurysmal cysts complicate giant-cell tumours, chondroblastomas, and osteosarcomas, or they may develop from simple unicameral cysts.

Diffuse idiopathic skeletal hyperostosis (DISH; Forestier's disease)

Presenting in middle age, and more commonly in men (2:1, male:female ratio), this condition of unknown aetiology affects about 10 per cent of men aged 65 years or over, and up to 58 per cent of men with gout. Usually a radiographic diagnosis, the criteria include the presence of new bone forming bridging osteophytes spanning at least four adjacent thoracic vertebrae in the absence of degenerative disc disease or sacroileitis. (The cortex is preserved, unlike the erosive process seen in the Romanus lesion of ankylosing spondylitis.) New bone formation can occur at any site, though entheseal sites are especially common. Phalangeal tufting, and an increase in the cortical thickness of the tubular bones of the hand and in the size of sesamoid bones are recognized. Symptoms include restriction in range of movement, diffuse limb pain, and symptoms of nerve entrapment or myelopathy. Canal stenosis can occur in the lumbar spine. Fracture through bridging osteophytes may also produce pain. Hyperinsulinaemia is frequently associated and related features such as hypertension, type II diabetes, obesity, and hyperlipidaemia are more commonly seen in this group. There is no medical treatment of proven value for established DISH. In the early stages, physical therapy may preserve the range of movement, and weight reduction is of value, both directly and in reducing hyperinsulinaemia. If oral hypoglycaemic agents are required, those that increase serum insulin levels should be avoided. Efforts to reduce heterotopic bone formation at sites of joint replacement have included radiotherapy and perioperative NSAIDs with

mixed results. Corticosteroid, given into joints or at entheseal sites may also offer symptomatic relief.

Myositis ossificans (MO)

Calcification of muscle complicates an intramuscular haematoma following direct impact, occurring in 17 to 20 per cent of such injuries. The anterior thigh and upper arm are the most common sites. Predictive signs at onset include local swelling, tenderness, and (particularly) reduced range of stretch in the involved muscle. A sympathetic knee effusion is described in up to half of those with MO in the thigh. Diagnosis may be confirmed radiologically after 3 weeks. Magnetic resonance imaging will detect a haematoma very early, but to date has not identified specific features predictive of MO. Therefore, the classic 'rest, ice, compression, elevation' is appropriate in the acute setting, with NSAIDs where pain and swelling are particularly marked. Physical training should not resume until a full range of passive stretching is restored. Surgical debridement of ectopic calcification should only be undertaken if it interferes with limb function, and then only where bone is matured, as assessed by bone scintigraphy.

Fibrodysplasia (myositis) ossificans progressiva, by contrast, is a rare inherited disorder. It is characterized by abnormally short halluces and ectopic calcification of striated muscle leading to disability as the neck, shoulders, spine, hips, and knees become progressively and relentlessly fixed. Additional variable features include fusion of the lateral masses in the lumbar spine, broad femoral necks, and widened metaphyses, as well as episodes of myositis, principally in the neck and upper paraspinal areas, preceding ossification. Histological misdiagnoses include sarcoma or rhabdomyosarcoma and juvenile fibromatosis. The disease appears to be due to a spontaneous genetic mutation in most cases, and prognosis is extremely variable. It has been difficult to evaluate therapeutic options for this reason, and no single measure is clearly of benefit, though there are theoretical grounds for the use of corticosteroids during episodes of myositis, bisphosphonates, and surgical debridement.

Ectopic calcification in renal disease

This is one aspect of renal osteodystrophy, where painful calcification of soft tissue, particularly at sites of repeated trauma, occurs as a result of serum levels of calcium and phosphate exceeding their combined solubility. Careful monitoring of phosphate and calcium levels, particularly when vitamin D analogues are used, and early treatment of hyperparathyroidism are all important, because established calcification may be intractable. Reversal following renal transplantation has been noted.

Melorheostosis

This is a rare disorder of linear hyperostosis associated with fibrosis of the skin and soft tissue. Thickening of cortical bone appears in a linear fashion (akin to spilling wax on the side of a candle), usually involving one or several bones in the same (more commonly the lower) limb. Many cases are associated with skin changes in the dermatome corresponding to the origin (sclerotome) of the affected bone, resulting in joint contracture. Symptoms include joint pain, intermittent swelling, deformity, and nerve entrapment, usually presenting in the second decade of life. Surgical intervention is most successful.

Paraneoplastic presentations

Rheumatological presentations associated with malignancy include gout, poly- and dermatomyositis, necrotizing vasculitis and cryoglobulinaemia, systemic sclerosis, and the presentations of lymphoproliferative disorders mentioned above. There are two specific conditions: hypertrophic pulmonary osteoarthropathy (**HPOA**) and remitting seronegative symmetric synovitis with pitting (o)edema (**RS3PE**). A seronegative polyarthritis without oedema and otherwise indistinguishable from rheumatoid arthritis may also occur.

HPOA is almost always associated with finger clubbing. Patients complain of pain and stiffness of the wrist and ankles, or of a more diffuse polyarthritis. Radiologically, a proliferative periostitis is noted, particularly at the diaphysis of wrists, ankles, and, less commonly, of knees and elbows. Over 90 per cent of cases have an intrathoracic malignancy, though infections or inflammatory conditions in pulmonary, cardiovascular, or gastrointestinal systems are seen. Primary HPOA (pachydermoperiostitis) also occurs. The cause of the condition is unknown: it has been suggested that cytokines such as platelet-derived growth factor might reach the periphery through pulmonary shunts, thereby producing the clinical proliferative features, but, on the basis of this hypothesis, it is difficult to explain the observation that vagotomy can relieve symptoms and signs in some cases. The arthritis is typically coincident with the malignancy, and will resolve with treatment of the underlying disease. Radiotherapy (to the periostitis sites) and infusion of pamidronate have also been successful in the treatment of resistant cases.

RS3PE was first described in 1985. Mostly affecting older men (mean age 71 years), a symmetrical polyarthritis involves the metacarpophalangeal and interphalangeal joints, wrists, and, less commonly, the elbows and shoulders. Tendon sheath involvement is quite common, and diffuse pitting oedema on the dorsum of the hands is characteristic. This condition has diverse clinical associations, but malignancy is detected in only 10 per cent of cases. Resistance to corticosteroid treatment in this otherwise very responsive condition raises the possibility of malignancy, though in most cases the underlying disease is detected within weeks. In paraneoplastic presentations, symptoms mirror treatment and relapse of the tumour.

Drugs producing rheumatological presentations

Myalgia may occur on withdrawal of steroids, especially in those patients taking 10 mg prednisolone for at least 30 days. This is best managed by reintroducing the steroid with a more gradual reduction in dose (for example, 1-mg steps every few days or weeks, depending on severity). Arthralgia and even arthritis are described as rare adverse effects of steroid therapy. Muscle cramps or aching may also complicate therapy with digoxin, penicillamine, clofibrate, and, more recently, with the statins. Myositis and rhabdomyolysis are also recognized in patients prescribed this latter group of drugs. The oral contraceptive is associated with a syndrome of persisting arthralgia, myalgia, morning stiffness, and even synovitis. Myopathy complicates statin and corticosteroid therapy, and chloroquine may cause neuromyopathy, particularly affecting the lower limbs. Myasthenic weakness is an uncommon complication of penicillamine.

Hypersensitivity reaction is associated with penicillamine, sulphonamides, thiouracils, and allopurinol, to name but a few. Presentations vary, but typically include a small vessel vasculitis and generalized arthralgia or arthritis.

Drug-induced systemic lupus erythematosus is well recognized, though ten times less common than classical SLE. It is characterized by the presence of antihistone antibodies, in distinction to the anti-DNA antibodies of classical SLE. Positive antinuclear antigen (**ANA**) antibodies are considerably more common than any clinical evidence of lupus. Other important distinctions from idiopathic SLE include resolution on withdrawal of the drug—renal and CNS involvement being rare, and rash uncommon—older age of onset (50–60 years compared with a mean age of onset of 29 years in idiopathic SLE). Drug-induced SLE is uncommon among the Black population, though this group accounts for 30 per cent of idiopathic cases. The drugs associated with SLE include hydralazine and procainamide, with minocycline an important recent addition. Most rheumatologists believe that these agents can be used safely by patients with idiopathic SLE, but oestrogen-containing contraceptives are generally regarded as being contraindicated. If a patient develops SLE, any concurrent medication should be withdrawn and the patient observed for a period. However, corticosteroids may be required where there is severe involvement, particularly of the renal, CNS, or cardiorespiratory systems. Antibodies may persist after satisfactory clinical resolution and are not of themselves an indication for continued treatment.

Isoniazid and phenobarbital have been associated with a shoulder–hand syndrome (discussed above as reflex sympathetic dystrophy). The mechanism of this association is unclear, though alteration in serotonin metabolism has been implicated.

Quinolone antibiotics can cause a tendinopathy. This may lead to rupture, most commonly of the Achilles tendon in elderly patients who are also taking corticosteroids.

Retinoids have been associated in recent years with a hyperostosis otherwise indistinguishable from DISH discussed above.

This discussion of the associations between drugs and rheumatological presentations is far from complete, and the physician should always consider drug therapy as a potential cause of new symptoms or signs.

Further reading

Ben-Chetri E, Levy M (1998). Familial Mediterranean fever. *Lancet* **351**, 659–64.

Berman A, *et al*. (1999). Human immunodeficiency virus infection associated arthritis; clinical characteristics. *Journal of Rheumatology* **26**, 1158–62.

Braun J, Sieper J (1999). Rheumatologic manifestations of gastrointestinal disorders. *Current Opinion in Rheumatology* **11**, 68–74.

Brower AC, Allman RM (1981). Pathogenesis of the neuropathic joint: neurotraumatic *vs.* neurovascular. *Radiology* **139**, 349–54.

Bywaters EG (1971). Still's disease in the adult. *Annals of Rheumatic Diseases*, **30**: 121–33.

Cush JJ, *et al*. (1987). Adult-onset Still's disease. Clinical course and outcome. *Arthritis and Rheumatism*, **30**: 186–94.

Cuthbert JA (1998). Wilson's disease. *Gastroenterology Clinics of North America* **27**, 655–81.

Ehrenfeld M, Gur H, Shoenfeld Y (1999). Rheumatologic features of haematologic disorders. *Current Opinion in Rheumatology* **11**, 62–7.

Hamdi N, Cvoke TD, Hassan B (1999). Ochronotic arthropathy: case report and review of the literature. *International Orthopaedics* **23**, 122–5.

Hermaszewski RA, Webster ADB (1993). Primary hypogammaglobulinaemia: a survey of clinical manifestations and complications. *Quarterly Journal of Medicine* **86**, 31–42.

Jones SM, Bhalla AK (1997). Algodystrophy. *Osteoporosis Review* **5**, 1–4.

Kaplan G, Haettich B (1991). Rheumatological symptoms due to retinoids. *Baillière's Clinical Rheumatology* **5**, 77–97.

King JB (1998). Post-traumatic ectopic calcification in the muscle of athletes: a review. *British Journal of Sports Medicine* **32**, 287–90.

Klemp P, *et al.* (1993). Musculoskeletal manifestations of hyperlipidaemia: a controlled study. *Annals of the Rheumatic Diseases* **52**, 44–8.

Kraus A, Alarcon-Segovia D (1991). Fever in adult onset Still's disease. Response to methotrexate. *Journal of Rheumatology*, **18**: 918–20.

Lear JT, Atherton MT, Byrne JP (1997). Neutrophilic dermatoses; pyoderma gangrenosum and Sweet's syndrome. *Postgraduate Medical Journal* **73**, 65–8.

Leclet H, Adamsbaum C (1998). Intraosseous cyst injection. *Radiology Clinics of North America* **36**, 581–7.

Mok CC, *et al*. (1998). Clinical characteristics, treatment, and outcome of adult onset Still's disease in southern Chinese. *Journal of Rheumatology*, **25**: 2345–51

Penrod BJ, Resnik CS (1997). Amyloid arthropathy. *Arthritis and Rheumatism* **40**, 1903–5.

Rodriguez-Merchan EC (1999). Common orthopaedic problems in haemophilia. *Haemophilia* **5**(Suppl. 1), 53–60.

Rydholm U (1998). Pigmented villonodular synovitis. *Acta Orthopaedica Scandinavica* **62**, 203–10.

Sibilia J, *et al*. (1999). Remitting seronegative symmetrical synovitis with pitting oedema (RS3PE): a form of paraneoplastic polyarthritis? *Journal of Rheumatology* **26**, 115–20.

Smith K, Fort JG (1998). Phalangeal osseous sarcoidosis. *Arthritis and Rheumatism* **41**, 176–9.

Smith P, Athanasou NA, Vipond SE (1996). Fibrodysplasia (myositis) ossificans progressiva; clinicopathological features and natural history. *Quarterly Journal of Medicine* **89**, 445–56.

Smythe M, Littlejohn G (1998). Diffuse idiopathic skeletal hyperostosis. In: Klippel JH, Dieppe PA, eds. *Rheumatology*, pp. 8.10.1–8.10.6. Mosby, London.

Trendelenberg M, Schifferli JA (1998). Cryoglobulins are not essential. *Annals of Rheumatic Diseases* **57**, 3–5.

Wong K, *et al*. (1999). Monoarticular synovial lesions; radiologic pictorial essay with pathological illustration. *Clinical Radiology* **54**, 273–84.

19.1 Disorders of the skeleton

R. Smith

Introduction

Bone is the only tissue, apart from teeth, that is mineralized to allow it to perform its normal function. The presence of mineral should not encourage the belief that bone is inert or that it is metabolically inactive. Many disorders affect the skeleton but only some can be considered here. Fractures, infections, and tumours that are more often dealt with by orthopaedic surgeons, are excluded from this chapter. The descriptions that follow may be divided into:

1. those disorders generally considered to be metabolic, such as osteoporosis, osteomalacia, Paget's disease of bone, and parathyroid bone disease;
2. those arising primarily from synthetic defects in the major components of the organic bone matrix and connective tissue, including osteogenesis imperfecta, the skeletal dysplasias, and Marfan's syndrome;
3. skeletal disorders that are clearly the result of enzyme defects, such as hypophosphatasia, homocystinuria, alkaptonuria, and the storage diseases;
4. those that appear to be intrinsic disorders of bone cells, such as osteopetrosis, fibrous dysplasia, and inherited ectopic ossification; and
5. various bone disorders that result from excessive minerals, vitamins, and metallic poisons.

To understand how these disorders arise and how to recognize them, a brief account of relevant aspects of bone physiology and clinical features is given here. More detail can be found in specialized texts (see Further reading list).

Physiology of bone

There is an increasing interest in the cells of bone, their control, activities, and communications, and in the non-collagen as well as the collagen components of the organic bone matrix. Advances in understanding of bone diseases such as osteoporosis, osteopetrosis, osteogenesis imperfecta, and Paget's disease reflect this. The causes of many rare skeletal disorders have been discovered (Table 1). Examples are Marfan's syndrome (mutations in the fibrillin gene); vitamin D-dependent rickets type II (mutations in the 1,25-dihydroxycholecalciferol receptor gene); pseudohypoparathyroidism and fibrous dysplasia (abnormalities in the G-protein signalling system); osteogenesis imperfecta (mutations in the type I collagen genes) and skeletal dysplasias (some with similar mutations in the type II collagen gene). Outstanding recent advances in bone physiology include the identification and elucidation of the functions of the parathyroid hormone-related peptide (**PTHrP**) and the bone morphogenetic proteins, known as **BMPs**. The discovery of the calcium-sensing receptor in the parathyroid and other tissues explains many rare disorders of mineral metabolism (Table 1). Further advances have been made in our understanding of the development of the osteoblast from the stromal-cell precursor and the ways in which the osteoblast controls osteoclast development and function (see below).

The mammalian skeleton serves two main functions, the demands of which often conflict. The first is to provide a rigid structure, the second is to act as an accessible mineral store. Both depend on the activities of specialized bone cells, controlled by genetic, mechanical, nutritional, and hormonal influences, and by a host of short-acting messengers produced by cells, collectively known as cytokines.

Structure

Bone tissue consists of cells and an extracellular mineralized matrix (35 per cent organic and 65 per cent inorganic). Some 90 per cent of the organic component is type I collagen. The remainder includes many non-collagen products of the osteoblast, such as osteocalcin, osteonectin, and proteoglycans. The mineral is present mainly as a complex mixture of calcium and phosphate in the form of hydroxyapatite.

Two anatomical types of bone may be defined, trabecular (cancellous) and cortical. The proportion of these differs from one bone to another; for example, vertebral bodies are predominantly trabecular, and the shafts of the long bones cortical. Such a distribution is related both to the functions of the bones and to the development of disorders of them, such as osteoporosis. Trabecular bone contains more metabolically active surfaces in a given volume than cortical bone. Cellular activities take place on the surfaces of trabecular bone and through resorbing channels (cutting cones) in cortical bone. The fine structure of bone is dealt with in anatomical texts.

Bone is often assumed to be inert because of its structural rigidity and persistence after death, and to be composed entirely of chalk because it contains 99 per cent of the body's calcium. These assumptions are superficially reasonable: neither is correct.

Bone cells

Conventional histological sections of bone demonstrate three types of bone cells which are clearly different (Fig. 1): osteoblasts, which may be plump and apparently active, or flat and apparently inactive—otherwise called bone-lining cells; multinuclear osteoclasts, which most often occupy areas of resorption; and osteocytes within their lacunae in the mineralized bone, apparently in contact with other osteocytes and bone cells through their extensions in the canaliculi. All these cells are in close contact with the bone marrow, which contains their precursors and brings them into close relationship with the immune system.

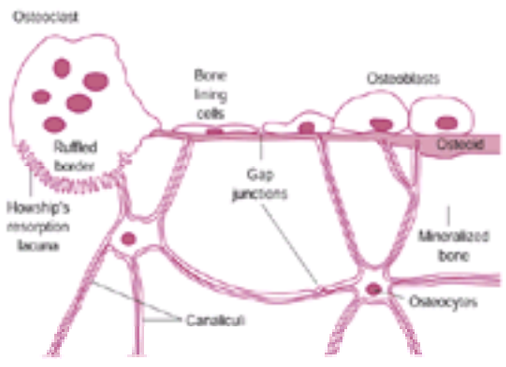

Fig. 1 A diagram showing the structure of bone and the relationship of the different cell types. (From the *Oxford textbook of rheumatology*, with permission.)

Bone cells are at the centre of an information system of astonishing complexity; and it is this complexity of bone that provides both the challenge and the fascination for those interested in its disorders. Histological techniques have been developed to study sequential cellular events in bone tissue; and the techniques of cell biology are used to study the origin and functions of different types of cells and the communications between them. All bone cells communicate with each other to control bone modelling during growth and remodelling throughout life. The constant processes of osteoclastic bone resorption and osteoblastic bone formation which achieve this are closely linked and take place in bone multicellular units (**BMUs**). The cellular cycle of such a unit begins with the activation of multinucleate osteoclasts from their macrophage-like mononuclear precursors, which produce resorption (Howship's) lacunae on the surface of trabecular bone, or cutting cones in cortical bone. These are identical processes; in cancellous (trabecular) bone the BMU may be looked upon as a sagittal section of a cortical BMU. Resorption is followed by a reversal phase, during which a cement line is deposited, and the formation by osteoblasts of new bone matrix which is subsequently mineralized. In the young adult, when the bone mass is constant and there may be several million resorbing sites in the skeleton at any one time, the amount of newly formed bone equals that resorbed. In childhood, more bone is formed than is resorbed; and in later years there is an imbalance between the two processes in favour of resorption, leading to osteoporosis.

The estimated time scale of the remodelling cycle is approximate. In the adult, the replacement of old bone with new occurs at an annual turnover rate of 25 per cent in cancellous bone, and 2 to 3 per cent in cortical bone. In the BMU resorption takes 1 to 2 weeks and new bone formation about 7 weeks. A complete BMU cycle, including reversal and mineralization, takes several months. The turnover of bone at a given site is determined by the frequency with which BMUs are activated and the rates of function of individual cells. Bone loss and gain depend on both factors; and the mechanism of bone loss is different in different disorders. Although the existence of the BMU system is widely accepted, it is far from understood. For instance, what factors lead to activation of the osteoclasts to initiate the resorbing cycle; how do cells talk to each other; and what links osteoblast and osteoclast activity?

It is clear that osteoblasts occupy a central position in bone physiology (Fig. 2). They are derived from the mesenchymal stromal-cell system within the bone marrow.

This system is multipotential and the stromal cells can give rise to osteoblasts, fibroblasts, chondrocytes, myocytes, and adipocytes. Under the influence of the differentiation factor identified as CBFA-1 (core-binding factor-a1) the stromal cells develop into osteoblasts. Osteoblasts respond to hormonal factors, both systemic and local (cytokines), and to mechanical stress. They synthesize the organic bone matrix, mainly collagen, and non-collagen proteins, and they control bone mineralization. Importantly, they also appear to direct the activity of other cell types, particularly the osteoclasts. In this respect they may also activate the bone-resorbing cycle. The osteoclast differentiation factor (**ODF**) has now been identified as osteoprotegerin ligand (**OPGL**, also known as **RANKL** and **TRANCE**), a soluble product of the osteoblast, which together with other factors controls the formation and activity of osteoclasts. It is possible that these many functions are divided between different osteoblasts. The bone-lining cells—resting osteoblasts—may not be as inactive as they appear, since they may provide a cellular barrier separating the so-called bone fluid from the general extracellular compartment. The separate existence of bone fluid has yet to be established.

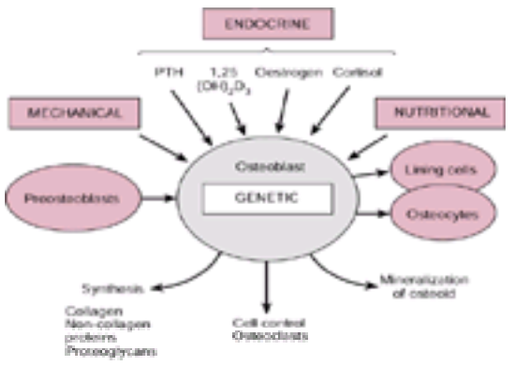

Fig. 2 To show the central position of the osteoblast in bone physiology. (From the *Oxford textbook of rheumatology*, with permission.)

Osteocytes, also derived from osteoblasts, occupy lacunae within the mineralized bone, and communicate with each other through gap junctions via their processes within the canaliculi. They probably have an important function in the detection of mechanical forces and the resultant response of bone.

Osteoclasts have a different origin from osteoblasts, since the former are multinucleated cells derived from the haemopoietic system. The osteoclasts resorb bone by attaching themselves to its surface via integrins and forming a seal to isolate their area of activity. Within this sealed zone they produce a very acid environment, with the aid of a proton pump linked to the enzyme carbonic anhydrase II, to enable digestion of whole bone by lysosomal enzymes. The absence of carbonic anhydrase II is linked to a rare form of osteopetrosis (see below). Osteoclasts have receptors to calcitonin which, when occupied, directly suppress their activity; the existence of any other hormone receptors is controversial. However, they are activated by prostaglandins. The osteoclastic resorptive effects of parathyroid hormone and of 1,25-dihydroxycholecalciferol are probably mediated through the osteoblast.

Bone formation

The factors that control bone formation are complex and not fully understood, but must work largely through the osteoblast. The stromal precursors of osteoblasts are found in the periosteum and the endosteal surfaces close to the bone marrow. The local remodelling stimulus for new bone formation appears to come from some product, or products, of bone resorption, which could, for instance, be a group of polypeptide growth factors or morphogenic proteins liberated from resorbed bone. Such substances are included in the category of cytokines. A cytokine may be defined as 'a peptide produced by a cell which acts as an autocrine, paracrine, or endocrine mediator'. This definition includes a large number of substances with effects on the metabolism of bone and cartilage. Such effects have largely been shown in experimental (and artificial) situations and their physiological role is unknown. Many cytokines have alternative names and multiple actions, featuring both synergism and antagonism. They include interleukins (1 and 6), tumour necrosis factor, g-interferon, platelet-derived growth factor, fibroblast growth factors, insulin-like growth factors, transforming growth factor-b, and bone morphogenic proteins.

Since bone cells contain, synthesize, and respond to many cytokines, they are part of a complex network. As an example, transforming growth factor-b (**TGF-b**) appears to belong to a family of multifunctional regulatory peptides, and bone is probably its most abundant source. Not only do osteoblasts synthesize TGF-b, but they also have high-affinity receptors for it, and are mitogenically stimulated by it. In addition, most of the bone morphogenic proteins belong to the TGF-b family.

Bone resorption

Osteoclasts are controlled by systemic and local hormones but there is no direct evidence that they are influenced by mechanical stress. Calcitonin directly inhibits the osteoclast, temporarily abolishes the active ruffled border, and suppresses the generation of new osteoclasts. Bone resorption is increased by parathyroid hormone and 1,25-dihydroxycholecalciferol. Since the osteoclast contains no receptors to either of these hormones it is proposed that their resorbing effect is mediated via the osteoblast. It is now realized that the interaction between osteoprotegerin (**OPG**) and its ligand (OPGL) is central to osteoblast/osteoclast interaction. The number and activity of the osteoclasts are also increased by a variety of cytokines produced by lymphocytes and monocytes (lymphokines and monokines, respectively), and by peptide growth factors such as epidermal growth factor. In myeloma the malignant plasma cells release interleukin-1 and -6 and tumour necrosis factor, all of which stimulate osteoclastic destruction of bone.

Bone mass (see also osteoporosis)

The development of the skeleton and its eventual size and density are influenced by important genetic factors modified by mechanical stress, nutrition, the systemic effects of endocrines, and by local factors produced by the bone cells themselves. These determine the balance between resorption and formation, and their relative contribution varies with age.

Recent work re-emphasizes the importance of the genetic contribution to bone mass. Apart from the difference in bone mass between races, this work has confirmed the hereditability of bone mass at all sites, which is greater in monozygotic than dizygotic twins. Clearly, mutations in the structural collagen genes will have a considerable effect on bone mass, as in osteogenesis imperfecta (see below). The contribution of vitamin D receptor-gene polymorphisms and genetic changes in the promoter region of type I collagen has been widely discussed (see osteoporosis).

The main function of the skeleton is mechanical and it has long been known that bone is laid down along its lines of stress. Although the way in which this occurs is obscure, experiments show that osteoblasts *in vitro* may respond to mechanical stress by an increase in levels of cyclic adenosine monophosphate (cAMP) and phosphonositol, partly mediated by prostaglandins. It also seems common sense that the size and density of the skeleton should be related to nutritional intake, particularly of calcium, protein, and energy. This has been difficult to prove, but recent co-twin studies in growing children have demonstrated a significantly greater density of bone (which may be temporary) in those taking calcium supplements, and that the starvation associated with anorexia nervosa reduces bone mineral content. This may also be due to oestrogen deficiency and emphasizes the important effect of reproductive hormones on the skeleton. The sex hormones, testosterone, and oestrogen, encourage new bone formation. It has recently been shown that oestrogen-deficient men have osteoporosis, and thus it is clear that the skeleton depends on a full complement of sex steroids for its integrity. Growth hormone is an important anabolic skeletal agent during the early years of life, partly through the local production of somatomedins (insulin-like growth factors). Several hormones that influence bone resorption may also have anabolic actions mediated by osteoblasts. One is parathyroid hormone, which under certain circumstances increases the proliferation of osteoblast precursors.

Collagen

Collagen is the principal extracellular protein in the body, more than half of which is contained within the skeleton, and is the main product of the osteoblast. There are many different molecular types, with different functions, each encoded by distinct genes (Table 2). Collagen in bone is type I. This heteropolymer is composed of two a-1 chains and one a-2 chain. The general structure of the a-1 chain is $(Gly–X–Y)_{338}$. The a-chains are synthesized as precursors within the osteoblasts and undergo a number of synthetic steps, including post-translational hydroxylation of proline and lysine residues; certain hydroxylysine residues are further modified into aldehydes and are also glycosylated (Fig. 3).

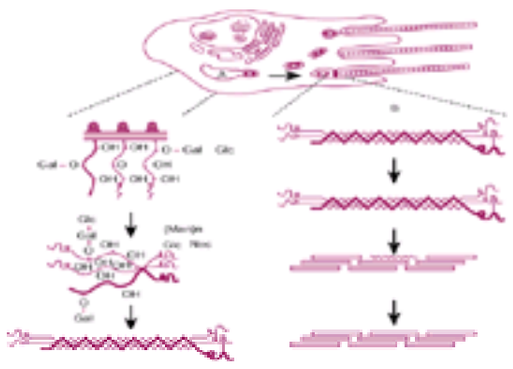

Fig. 3 The synthesis and assembly of collagen molecules. Within the fibroblast (A) the individual pro-a-chains are modified, assembled, and folded into the triple helix. In (B) these chains are exported, shortened, and self-assemble. (From the *Oxford textbook of Rheumatology*, with permission.)

After removal of their extensions, the triple-helical molecules form an exact structure with a quarter-stagger overlap that is subsequently crosslinked. The so-called 'hole zones' within this structure provide a template for early mineralization. Mutations in the collagen genes and defects in post-translational modification cause inherited disorders of connective tissue, of which osteogenesis imperfecta (type I collagen) and type IV Ehlers–Danlos syndrome (type III collagen) are examples (Table 1). Renal excretion of hydroxyproline peptides is an indicator of bone collagen turnover, and excretion of pyridinium compounds is a measure of bone resorption (see below).

Non-collagen proteins

Many such proteins may be extracted from bone, although their abundance differs according to the starting material and the methods used. They include osteocalcin (Gla protein), sialoproteins, various phosphoproteins, such as osteonectin and osteopontin, the bone morphogenetic proteins, and bone-specific proteoglycans.

The nature of non-collagen substances sequestered in bone matrix is complex and most are synthesized by the osteoblasts. Few, if any, are unique to bone, since they can be expressed transiently in other tissues—to date, no unambiguous function has been determined for any of these proteins. Osteonectin is the most abundant non-collagen protein produced by human osteoblasts. It binds strongly to calcium ions, hydroxyapatite, and native collagen, but is not limited to mineralizing tissue, being also found in human platelets. Although osteonectin mRNA is widely distributed in developing tissues, osteonectin is most abundant in bone. Two bone sialoproteins (**BSP**) are now recognized (BSP1 and BSP2). Their relative abundance varies with the species studied: for instance, BSP1 is a minor component of human bone, but a major contributor to total sialoprotein in rat bone. The protein contains an **RGD** (Arg–Gly–Asp) cell-attachment sequence and is therefore called osteopontin. The major human sialoprotein is BSP2.

There are two bone Gla-containing proteins: osteocalcin—bone Gla protein (**BGP**)—and matrix Gla protein (**MGP**). The term Gla refers to the g-carboxylated glutamic acid residues, formed by the vitamin K-modulated, post-translational carboxylation of peptide-bound glutamic acid. These proteins have some sequence homology but are products of different genes. MGP is also a cartilage protein and is found at an earlier developmental stage than BGP. The function of BGP is unknown. BGP biosynthesis is regulated by 1,25-dihydroxycholecalciferol (1,25(OH) $_2D_3$) (and no other hormone), which enhances its nuclear transcription and eventual secretion from bone cells. Plasma BGP has been linked to the rate of bone formation or, less specifically, bone turnover.

Proteoglycans are proteins with one or more attached glycosaminoglycan chains. They vary widely in form and function. Those of bone, which include decorin and biglycan, have been studied less extensively than those of cartilage, and differ from them in their small overall size and relatively larger amounts of protein. Such small proteoglycans are thought to interact with growing collagen fibrils in a precise manner and to regulate their growth, maturation, and interactions. Type IX collagen, closely associated with type II collagen, bridges the gap between the collagens and proteoglycans since it contains a chondroitin sulphate glycosaminoglycan chain.

It has been known for many years that demineralized bone matrix contains substances capable of inducing ectopic bone formation. Because they are present in such small amounts their extraction and isolation have presented great difficulties, but these bone morphogenic proteins have now been isolated and their genes localized and cloned. Interestingly, most belong to the TGF-b supergene family. Some evidence suggests their overexpression in fibrodysplasia ossificans progressiva (**FOP**).

Bone mineral and mineralization

Mineralization occurs on bone matrix collagen. The way in which it occurs has been long debated, but there is now good evidence that, in most mineralized tissues, calcifying vesicles derived from chondrocytes or osteoblasts provide a focus for mineralization. These vesicles are easily demonstrable in cartilage, but their function in the organized matrix of bone is controversial. The precipitation of calcium within these vesicles may be controlled by the action of a pyrophosphatase that locally destroys pyrophosphate, itself an inhibitor of mineralization. Alkaline phosphatase is one such pyrophosphatase which is readily demonstrable both in osteoblasts and in mineralizing vesicles. It is possible, for the purpose of clarity, to consider two types of mineralization: namely (1) homogeneous nucleation, which occurs in the lumen of the matrix vesicles, from amorphous calcium phosphate to form crystalline hydroxyapatite; and (2) heterogeneous nucleation, which is collagen-mediated and may partly rely on adsorbed non-collagen proteins as nucleators. After this first phase (mediated either by vesicles or collagen) there is a second phase of rapid spread of mineralization initially in the hole zones and later the overlap regions of the collagen matrix.

Calcium and phosphorus balance (see also Section 12)

Much has been written about calcium balance and the main hormones that control it. Phosphate balance is less well understood. The circulating level of plasma calcium is determined by the amount of calcium that is absorbed by the intestine, the amount that is excreted by the kidney, and the exchange of mineral with the skeleton. The relative importance of these exchanges differs during growth and in different disorders. Total plasma calcium concentration is closely maintained between 2.25 and 2.60 mmol/l, of which nearly half is in the ionized form (47 per cent ionized, 46 per cent protein bound, and the remainder complexed). The skeleton contains approximately 1 kg (25 000 mmol) of calcium. The main fluxes of calcium in the young adult are shown in Fig. 4.

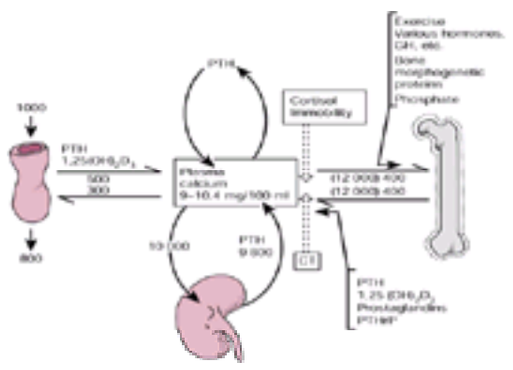

Fig. 4 Factors that control calcium balance. Units are in mg/day (to convert to mmol divide by 40) and refer to an adult. The figures in parentheses are an estimate of exchange through the cellular barrier of bone. CT, calcitonin; GH, growth hormone; PTH, parathyroid hormone; PTHrP, parathyroid hormone-related peptide. (From *Oxford textbook of rheumatology*, with permission.)

Parathyroid hormone (see also Chapter 12.6)

The gene for parathyroid hormone (PTH) is on chromosome 11. PTH is synthesized as a large precursor, in the way of proteins packaged for export, and its secretion is stimulated by a reduction in the plasma ionized-calcium concentration. Changes in plasma calcium are detected by a sensitive calcium-sensing receptor. Mutations in the gene for this receptor can cause hypocalcaemic and hypercalcaemic syndromes (Table 1). Increase in PTH secretion leads to an increase in calcium absorption through the gut, an increase in calcium reabsorption through the kidney, and an increase in bone resorption. Intestinal calcium absorption is mediated by 1,25-dihydroxycholecalciferol, and the 1a-hydroxylation of 25-hydroxy-cholecalciferol is stimulated by parathyroid hormone, so that the effect of parathyroid hormone in increasing intestinal calcium absorption is indirect. In contrast, the renal effect of parathyroid hormone on calcium reabsorption is direct. The cellular effects of parathyroid hormone on kidney and bone appear to involve two cellular systems, namely cAMP and phosphoinositol. Parathyroid hormone encourages osteoclastic bone resorption by its effects on the osteoblast (as previously described). Peripheral resistance to the effect of PTH due to inherited loss-of-function mutations in the G-protein signalling system occurs in pseudohypoparathyroidism (see below and Chapter 12.6).

Vitamin D

Vitamin D is synthesized either as vitamin D_3 (cholecalciferol) within the skin from its precursor 7-dehydrocholesterol under the influence of ultraviolet light (usually as sunlight), or taken in with food, either as vitamin D_3 or D_2 (ergocalciferol) (Fig. 5). It is transported to the liver by a binding protein where it undergoes 25-hydroxylation. 25-hydroxy-vitamin D is then hydroxylated in the 1a-position by the renal 1a-hydroxylase. 1,25(OH) $_2$D is the active metabolite of vitamin D and has widespread effects, the extent of which is only just being appreciated. These are mediated through a widely distributed vitamin D receptor which has DNA- and hormone-binding components. In addition to its classic effect on intestinal calcium transport, vitamin D is linked with the immune system and the growth and differentiation of a wide variety of cells. Measurement of the plasma 25-hydroxy-vitamin D concentration has proved to be a useful indicator of vitamin D status, and work on 1,25(OH) $_2$D and its receptors has illuminated the cause of the rarer forms of inherited rickets (see below). The kidney is the main source of 1,25(OH) $_2$D but it is now clear that this metabolite can be synthesized by macrophages in a variety of granulomas, providing an explanation for the hypercalcaemia of sarcoidosis, disseminated tuberculosis, and (occasionally) lymphomas.

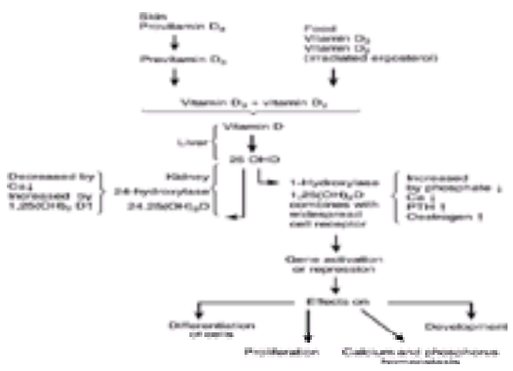

Fig. 5 The synthetic pathways and molecular and cellular effects of 1,25(OH) $_2$D. (From *Oxford textbook of rheumatology*, with permission.)

Calcitonin

The main effect of administered calcitonin is to reduce bone resorption by the direct and reversible suppression of osteoclasts and by inhibition of their production from precursors. The role of calcitonin is uncertain, although it is thought to protect the skeleton during physiological stresses such as growth and pregnancy. It is produced by alternative splicing of the primary gene transcript also responsible for the production of calcitonin gene-related peptide. Recent work has shown that its receptor is widely distributed.

Parathyroid hormone-related protein (PTHrP)

This hormone was discovered through studies on patients with non-metastatic hypercalcaemia of malignancy. PTHrP has close sequence homology to PTH at the amino-terminal end of the molecule and has very similar effects. Its gene is located on the short arm of chromosome 12, and is thought to have arisen by a duplication of chromosome 11, which carries the human *PTH* gene. It has been detected in a number of tumours, particularly of the lung. There is also evidence that it may have a role in fetal physiology, controlling the calcium gradient across the placenta to maintain the relatively higher concentrations in the fetal circulation. PTH and PTHrP have the same receptor. Mutations of this receptor can cause Jansen's metaphyseal dysplasia and it has become clear that PTHrP is involved in early development of the skeleton.

Other hormones

Apart from the recognized calciotrophic hormones, the skeleton is influenced by corticosteroids, the sex hormones, thyroxine, and growth hormone. The main effect of excess corticosteroids (either therapeutic or in Cushing's syndrome) is to suppress osteoblastic new bone formation, although there is also an element of secondary hyperparathyroidism. Androgens and oestrogens promote and maintain skeletal mass. Osteoblasts have receptors for oestrogens, although they are not abundant. Thyroxine increases bone turnover and increases resorption in excess of formation; thyrotoxicosis thus leads to bone loss. Excess growth hormone leads to gigantism and acromegaly (according to the age of onset) with enlargement of the bones. Absence of growth hormone will lead to proportional short stature; where there is general pituitary failure the reduction in gonadotrophins will also induce bone loss.

Biochemical measures of bone turnover

Knowledge of bone physiology allows one to interpret biochemical measures of bone turnover. These include plasma bone-derived alkaline phosphatase and osteocalcin (BGP), and the urinary total hydroxyproline and crosslinked collagen-derived peptides. The first two are produced by osteoblasts and indicate bone formation, the second two, bone resorption. Since formation and resorption are closely coupled, such measurements are usually closely related to each other and to overall bone turnover.

Total plasma alkaline phosphatase (largely derived from osteoblasts) provides a crude but readily accessible index of bone formation, being increased during periods of rapid growth and particularly when bone turnover is greatly increased, as in Paget's disease. Early measurements of serum BGP gave widely variable results and depended on the origin, sensitivity, and stability of the antibodies used. Total urinary hydroxyproline excretion is influenced by dietary collagen (gelatin) and reflects both resorption and new collagen synthesis. The recent development of methods for the measurement of urinary collagen-derived pyridinium crosslinks promises to give a reliable indication of bone resorption rate, unrelated to new collagen formation, and uninfluenced by diet. There are two forms of crosslinked peptide—pyridinoline and deoxypyridinoline, depending on whether they originate from oxidized hydroxylysine or lysine residues. Early assays were dependent on high-pressure liquid chromatography (**HPLC**) of urinary peptides after hydrolysis with acid. Simple and more direct immunoassays have now been developed.

Correct interpretation of collagen-derived fragments depends on knowledge of collagen's metabolic pathway (Fig. 3). Soon after export from the cell, the amino- and carboxy-propeptide extensions are cleaved from the mainly helical central part of the collagen chain. Measurement of these fragments in the plasma indicates the collagen formation rate. Once the collagen chains are crosslinked, measurement of different crosslinked fragments in the urine indicate (mainly bone) collagen resorption.

The diagnosis of bone disease

The diagnosis of bone disorders increasingly depends on investigation, with the result that important clinical points tend to be forgotten.

History

Deformity, pain, and fracture are common features. To these may be added proximal myopathy (in osteomalacia and rickets) and the symptoms of any underlying disease. The family history is always relevant.

Deformity

Deformity suggests previous skeletal disorder, especially if there is a disturbance of growth. Short stature and disproportion are more frequent than excessive height. In children, a knowledge of growth is essential; in the normal adult, height and span are approximately equal and the crown to pubis measurement is equal to the pubis to heel. Those with short stature can be divided into proportionate and disproportionate, of which the most frequent cause is short limbs. Proportionate short stature may occur in children who appear to be otherwise normal, whereas subjects with disproportionate short stature usually appear abnormal from birth. Some causes of short stature are given in Table 3. Skeletal chondrodysplasias are dealt with further below.

Kyphosis, with loss of trunk height, as in osteoporosis and osteomalacia, is the commonest acquired deformity of adult life. It is often noticed because clothes no longer fit. During childhood vertebral collapse will slow the growth rate. Other deformities are characteristic of the underlying disease; for instance, active childhood rickets produces knock knees, bowed legs, enlarged epiphyses, and bossing of the skull; Paget's disease produces thick limb bones and an enlarged skull vault; and severe osteogenesis imperfecta, very short limbs.

Bone pain and fracture

The cause of bone pain is not well understood. In osteomalacia it may be generalized and associated with tenderness on pressure. It may be due to excessive vascularity, with stretching of the periosteum; certainly it can be rapidly relieved by appropriate treatment, such as calcitonin for Paget's disease, or parathyroidectomy for parathyroid bone disease. Fractures of different sorts occur, examples being: the partial, multiple, and painful microfractures ('fissure' fractures) on the convexity of pagetic bone; the Looser's zones on medial borders of osteomalacic bones; and the multiple vertebral compression fractures of osteoporosis.

Myopathy

The cause of the proximal muscle weakness in osteomalacia and rickets remains unknown. The symptoms include a waddling gait, and inability to rise from a chair, to lift objects off high shelves, or to climb stairs. Limbs may be described as stiff rather than weak. Myopathy does not occur in subjects with inherited hypophosphataemia.

Underlying disease

It is necessary to be alert for the symptoms of the underlying disease, such as renal failure, steatorrhoea, or myeloma, and to enquire particularly about previous abdominal operations, including hysterectomy and oophorectomy.

Physical signs

It is important to see the patient out of bed so that an abnormal gait or stature is not missed. The appearance may give vital clues; for instance, the large vault of Paget's disease; the coarse features, large nose, big lower jaw, and widely spaced teeth of acromegaly; and the round face, simplicity, and cataracts of pseudohypoparathyroidism. Endocrine disorders affecting the skeleton, such as hypogonadism and hypopituitarism, are readily recognizable. Special facial features should receive attention; these include the eyes for such signs as corneal calcification, arcus juvenilis, and lens dislocation shown by the shimmering of the unsupported iris, iridodinesis. Further examples are corneal clouding (some mucopolysaccharidoses) and cystine crystals (cystinosis). In dentinogenesis imperfecta, often found with osteogenesis imperfecta, the teeth are abnormal in shape, tend to be transparent, and vary in colour from yellow to grey. Enamel defects occur in hypoparathyroidism, teeth are lost early in hypophosphatasia; and dental abscesses are common in hypophosphataemic rickets.

Hands and feet need particular attention. The fingers may be abnormally long and thin, as in Marfan's syndrome, or excessively short and mobile, as in pseudoachondroplasia; alternatively, they may be short, wide, and stiff in some mucopolysaccharidoses; or the hands may have short metacarpals, as in pseudohypoparathyroidism, or additional digits, as in the Ellis–van Creveld syndrome. The monophalangic big toe (and less often short thumbs) is characteristic of fibrodysplasia ossificans progressiva. Abnormal body proportions are common; the spine is relatively short after vertebral collapse. Scoliosis often dates from adolescence; occasionally it may be a clue to an inherited connective tissue disorder. A thoracolumbar gibbus is a particular (though not exclusive) feature of the mucopolysaccharidoses. Spinal deformity produces secondary changes; thus a young patient with severe osteoporosis will develop a prominent sternum with ribs that touch the iliac crest and a transverse crease across the front of the abdomen. Spontaneous tetany is a rare symptom, but there are two recognized bedside tests for latent tetany; of these Chvostek's sign is more convenient, but that of Trousseau more reliable. The first involves tapping the branches of the facial nerves as they spread out from within the parotid gland; a positive sign is twitching of the appropriate facial muscle. In the second the forearm is made ischaemic with a sphygmomanometer cuff for up to 3 min; if positive, carpal spasm will occur.

Investigations

Biochemistry

Many generalized disorders of the skeleton, such as postmenopausal osteoporosis, achondroplasia, osteogenesis imperfecta, and the epiphyseal dysplasias, have normal routine biochemical values; in others changes are diagnostic (Table 3). In normal persons the fasting plasma calcium concentration remains virtually constant through life, the plasma phosphate declines in adolescence to adult values and the plasma alkaline phosphatase level increases temporarily during rapid adolescent growth. Since total plasma calcium includes a protein-bound fraction, it is usual to relate it to the plasma albumin level and, if necessary, correct it to a plasma albumin of 4 g per 100 ml. Acceptable corrections include: corrected calcium (mg per 100 ml) = measured calcium – albumin (g per 100 ml) + 4; or for SI units: 0.02 mmol/l for every 1 g/l change of albumin from 40 g/l. The fasting plasma calcium is normal in osteoporosis and also in Paget's disease unless the patient is immobilized. It is increased in primary hyperparathyroidism, various neoplasms (including humoral hypercalcaemia of malignancy), in sarcoidosis, in vitamin-D overdosage, and in a number of other states, such as acromegaly and thyrotoxicosis (Table 4). It is often low in osteomalacia, but may be restored towards normal by secondary hyperparathyroidism, and is low in parathyroid insufficiency. Normal values are to be expected in inherited hypophosphataemia and in other forms of renal tubular rickets.

Since the main determinant of the fasting plasma phosphate concentration is its renal tubular reabsorption, hypophosphataemia occurs in primary hyperparathyroidism, in the humoral hypercalcaemia of malignancy, and it is also low in inherited hypophosphataemic rickets. Both oral aluminium hydroxide and prolonged intravenous nutrition also lower plasma phosphate levels. Hyperphosphataemia occurs in hypoparathyroidism, in renal glomerular failure, and in the rare, recessively inherited form of tumoral calcinosis.

Total plasma alkaline phosphatase and bone-derived alkaline phosphatase is normally increased in adolescence and in osteomalacia, particularly in the young, but it may be near-normal in renal tubular osteomalacia. Increases occur in primary hyperparathyroidism, but only where there is demonstrable bone disease. The highest values for plasma alkaline phosphatase are found in young patients with active Paget's disease, and in idiopathic hyperphosphatasia; and the lowest in hypophosphatasia.

Other plasma measurements, which have application in particular circumstances and in research, include: tartrate-resistant acid phosphatase (**TRAP**), a product of the osteoclast and therefore an indication of bone resorption; osteocalcin (bone Gla protein), a product of the osteoblast and therefore sometimes useful as an indicator of bone formation; and the N- and C-propeptide extensions of collagen, again an indicator of bone formation rate.

Glucose in the urine of a patient with inherited rickets suggests multiple renal tubular defects, and proteinuria is an important clue to myelomatosis.

The amount of calcium excreted in the urine is related both to the plasma levels and to the percentage of the filtered load reabsorbed through the renal tubules, itself altered by parathyroid hormone. Hypocalcaemia therefore causes hypocalciuria, particularly in osteomalacia and rickets; and hypercalcaemia leads to hypercalciuria, especially when this is due to rapid bone loss as in neoplastic disease of the skeleton, leukaemia, myeloma, and immobilization. Since parathyroid hormone increases the renal tubular reabsorption of calcium, the normal relationship between plasma and urine calcium is disturbed in parathyroid disease; however, most hypercalcaemic hyperparathyroid patients excrete more calcium than normal. Total hydroxyproline in the urine (after acid hydrolysis of the peptides) is a good indicator of bone breakdown and collagen turnover, provided the patient is ingesting a low-gelatin diet. The physiological changes in hydroxyproline excretion are striking, with a particularly sharp peak in adolescence coinciding with the maximum height velocity. The highest values are seen in active Paget's disease, where the excretion may be up to 50-fold the normal value. Hydroxyproline excretion correlates well with plasma alkaline phosphatase, and is therefore increased in some forms of osteomalacia and in hyperparathyroidism with bone disease. Since thyroxine increases collagen turnover, urinary hydroxyproline is also abnormally high in thyrotoxicosis and abnormally low in myxoedema (either primary or secondary).

Hydroxyproline excretion can be most usefully expressed as the amount in a 24-h urine sample in a patient on a gelatin-free diet, or in a fasting urine sample in relation to creatinine. However, hydroxyproline peptide excretion is related both to newly formed and mature collagen, and is not, therefore, a direct measure of bone resorption. The urinary excretion of pyridinium compounds (see above) from the lysyl- and hydroxylysyl-derived crosslinks of mature collagen is a direct measure of bone resorption, irrespective of dietary collagen.

Radiology

The diagnosis of bone disease often depends on the radiographic appearances, especially where there are no demonstrable biochemical changes. A particular example is in the differential diagnosis of perinatal lethal dwarfism. Conventional radiographs demonstrate well structural changes such as fractures, deformity, areas of resorption, and alteration in size, but are unreliable for the assessment of bone density. As radiographic techniques develop, increasing use is made of isotope bone scans, computed tomographic (**CT**), and magnetic resonance imaging (**MRI**) scans. Bisphosphonate-labelled scanning agents are selectively taken up in areas of increased vascularity or turnover. They are very useful in demonstrating the skeletal extent of Paget's disease of bone, the presence of bony metastases, the pathological fractures of osteoporosis, and Looser's zones in osteomalacia. An isotope scan is preferable to multiple radiographs to assess the distribution (but not the structure) of abnormal bone.

CT scanning can also be very useful in bone disease. Examples include the delineation of ectopic ossification, of spinal cord compression, and of bone tumours. Although magnetic resonance scanning (MRI) finds its most important application in soft-tissue pathology, it is also useful in giving an idea of the composition as well as the structure of bone.

Methods for measuring bone mass are considered under osteoporosis (see below).

Bone biopsy

Direct examination of bone is a valuable but under-used investigation. Bone can be taken by a transiliac trephine (using a local anaesthetic) and sections should be examined with and without decalcification. Ideally the bone should be labelled with tetracycline to allow an estimate of formation rates. In the various metabolic bone diseases the appearances are characteristic, with the: excess osteoid of osteomalacia; the disorganized mosaic pattern, excessive cellular activity, and fibrosis of Paget's disease; and osteitis fibrosa cystica in hyperparathyroid bone disease. In mild osteogenesis imperfecta there is typically an increase in the number of osteocytes, and, in the more severe form, a considerable increase in the amount of woven bone. A normal biopsy will exclude these diseases except where the pathological changes are patchy. Where possible, histological examination should now include transmission and scanning electron microscopy, and the report should include quantitative histomorphometry. More details are given in larger texts (see the Further reading list).

Further investigations

Measurement of the external calcium and phosphorus balance is a classic way of investigating generalized bone disease and the effects of treatment upon it, but it is also tedious. The use of isotopes to measure calcium absorption and apparent bone formation and resorption rates is less direct and also depends on a number of assumptions. This leaves a large number of measurements available for specific problems. Important examples (in the plasma) are intact PTH assays (to investigate hyper- and hypocalcaemia), PTHrP (mainly in research), 25-hydroxy-vitamin D, and 1,25-dihydroxy-vitamin D (for the investigation of rickets and osteomalacia). In inherited disorders, analysis of DNA extracted from whole blood and of collagen synthesized from fibroblast cultures derived from skin samples are used increasingly.

Diagnosis

The diagnosis of a skeletal disorder is not difficult where there are clear biochemical disturbances (Table 4), although, as in osteomalacia, the causes may be many. An exact diagnosis may be impossible when the standard biochemical results are normal, and this is particularly so in some of the rare heritable disorders. Guidance based on the age of the patient and frequency of the disorder is given in Table 5.

Osteomalacia and rickets

Osteomalacia results from a lack of vitamin D or a disturbance of its metabolism; in the growing skeleton it is referred to as rickets, and the terms are often used interchangeably. Very rarely, severe calcium deficiency can lead to rickets. Inherited hypophosphataemia and several other renal tubular disorders may also cause rickets without clear evidence of abnormal vitamin D metabolism. The causal mutations in inherited hypophosphataemia have now been identified.

The main histological feature of osteomalacia is defective mineralization of bone matrix (Fig. 6). Our present understanding of osteomalacia relies on advances in knowledge of vitamin D metabolism (Fig. 7). For clinical purposes two aspects of the physiology of vitamin D require emphasis. The first is the quantitative importance of vitamin D synthesis in the skin in comparison with that in the diet, and the second concerns the relative role of different vitamin D metabolites. The measurement of circulating concentrations of 25-hydroxy-vitamin D (**25(OH)D**) as an index of vitamin D status has identified those groups (Asian immigrants and the elderly) most at risk from vitamin D deficiency; importantly it has also shown the large amounts of vitamin D that can be synthesized in the human skin when exposed to ultraviolet light. The causes of osteomalacia can now be partly understood in terms of its metabolites, and the major importance of **1,25(OH)2D** (1,25-dihydroxy-vitamin D) is established. The effects of giving vitamin D can probably not be ascribed to the actions of $1,25(OH)_2D$ alone, and probably include other biologically active derivatives such as 25(OH)D and possibly 24,25-dihydroxy-vitamin D (**24,25(OH)2D**).

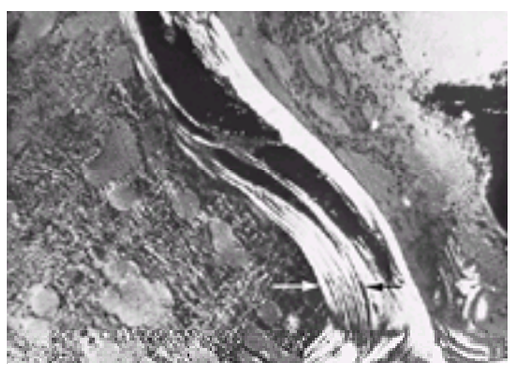

Fig. 6 Bone from a patient with osteomalacia. The birefringent osteoid is abnormally thick (up to 12 lamellae, arrows) and covers all bone surfaces. The bone preparation is undecalcified and viewed under polarized light (von Kossa stain; magnification 300 ×).

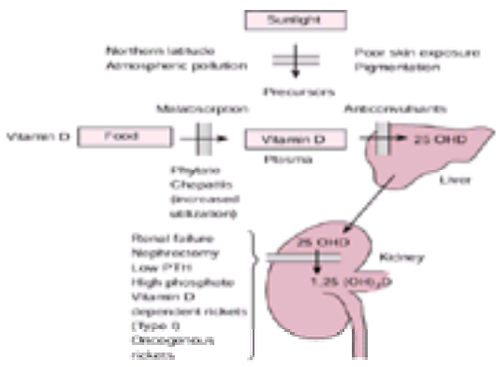

Fig. 7 The causes of rickets and osteomalacia related to the sources and metabolism of vitamin D (From *Oxford textbook of rheumatology*, with permission.)

Pathophysiology

The features of osteomalacia can be predicted largely from the known calciotropic effects of vitamin D. Examination of undecalcified bone shows wide osteoid seams with many birefringent lamellae of collagen (Fig. 6) covering more of the bone surface than normal, and absence of the 'calcification front'. The absence of this front is important since excessive osteoid may also be found in conditions other than osteomalacia, such as hypophosphatasia, Paget's disease of bone, and thyrotoxicosis, where the calcification front is normal; in these disorders the increase tends to be in the amount of bone surface covered rather than in the thickness of osteoid. Excess osteoid also occurs when bisphosphonates, such as etidronate, or aluminium accumulate in the skeleton. In rickets the main change is disorganization of the growth plate.

Since there is intestinal malabsorption of calcium in vitamin D deficiency, both the plasma and urine calcium levels are lower than normal; absorption of phosphorus is also defective, with resultant hypophosphataemia. As hypocalcaemia stimulates the secretion of parathyroid hormone, this will correct the low plasma calcium level and exaggerate hypophosphataemia. In osteomalacia, osteoblastic activity is increased and the plasma alkaline phosphatase is therefore also increased. There appears to be no difficulty in laying down bone matrix collagen, but it cannot be properly mineralized. One should recall that the effects of vitamin D are not confined to the skeleton, although they are clinically most obvious in this tissue—thus vitamin D has important effects on cellular differentiation and on the immune system.

Causes

There are many causes of osteomalacia (and rickets), some of which are very rare. They may conveniently be divided into three main groups: nutritional, malabsorptive, and renal (Table 6). Most can be understood in terms of vitamin D metabolism (see Fig. 7). In the elderly and immigrant populations the food intake of vitamin D is often deficient and the requirements may be increased; the absorption of vitamin D is poor in coeliac disease, after partial gastrectomy, intestinal resection or bypass, and in biliary disease. The intestinal absorption of calcium is reduced by phytate and chapatti ingestion, which may also increase vitamin D requirements (see below). Endogenous synthesis of vitamin D in the skin is reduced, especially in towns and city communities in the Northern hemisphere; it is further reduced by skin pigmentation. The 25-hydroxylation of calciferol may be impaired in some chronic liver diseases, and anticonvulsants may induce hepatic enzymes which degrade vitamin D. The 1a-hydroxylation of 25(OH)D is reduced or absent in renal failure, after nephrectomy, in hyperphosphataemia, parathyroid insufficiency, in type-I vitamin D-dependent rickets, and probably in some bone tumours. Many patients have more than one cause for their osteomalacia; in the elderly person vitamin D intake is poor, exposure to sunlight is limited, and renal glomerular failure progressive. Reduced exposure to sunlight is a frequent consequence of physical immobility and may contribute to osteomalacia in rheumatoid arthritis and other chronic diseases.

The effects of renal glomerular failure on the skeleton are complex (Chapter 20.8). Two main events occur: one is an increase in the plasma phosphate level, which leads to a fall in plasma calcium and to secondary hyperparathyroidism with excessive bone resorption; the other is the reduced formation of $1,25(OH)_2D$, with defective intestinal absorption of calcium and defective bone mineralization. The combination of these events rapidly produces severe deformity, especially in the growing skeleton. In patients receiving dialysis, renal osteodystrophy may be complicated by aluminium intoxication.

Clinical features

The main symptoms of osteomalacia are bone pain and tenderness, skeletal deformity, and proximal muscle weakness, often accompanied by the features of the underlying disorder and by those of hypocalcaemia. In severe osteomalacia all the bones are painful and tender, sometimes sufficiently so to disturb sleep. The tenderness can be particularly marked in the lower ribs and may also be accentuated over Looser's zones. Deformity is most often seen in rickets when the effects of vitamin D deficiency are superimposed on a growing skeleton. The linear growth rate is reduced, there is bowing of the long bones, enlargement of the costochondral junctions (rickety rosary), and bossing of the frontal and parietal bones. Later, osteomalacia may produce a triradiate pelvis, a gross kyphosis, and corresponding deformities of the chest.

Proximal muscle weakness is an important symptom. Its cause is unknown (although myoblasts require $1,25(OH)_2D$ *in vitro*, and the development of myofibrils in animals without the vitamin D receptor may be abnormal). It is more marked in some forms of osteomalacia than in others. Most commonly there is a waddling gait, a difficulty in getting up and down stairs, out of low chairs, and in and out of small cars. In the elderly, weakness may make walking impossible thereby suggesting paraplegia. In younger subjects even muscular dystrophy may be simulated.

Features of the underlying disorder include anaemia; tiredness and steatorrhoea in coeliac disease; pigmentation, thirst, and nocturia in renal failure. Occasionally hypocalcaemia may cause spontaneous tetany; in children the manifestations of carpopedal spasm, stridor, and fits are more dramatic than in the adult.

Examination of the patient with osteomalacia or rickets confirms the main symptoms. Measurement of the body proportions is useful. Thus patients with inherited hypophosphataemia and rickets have relatively short limbs, whereas those with late-onset osteomalacia will have a relatively short trunk. It is important to look for clues as to the cause of the osteomalacia, such as the scars of previous gastric or intestinal surgery.

Investigations

Biochemistry

Since there are many causes of osteomalacia, the detailed biochemical changes differ from one to another. In vitamin D deficiency or malabsorption there are low plasma calcium and phosphate, a low urine calcium, and an increase in the plasma alkaline phosphatase level. However, these may vary with the stage of the disease. Initially, hypocalcaemia may be the only abnormality. Later, with secondary hyperparathyroidism, the plasma calcium level returns towards normal, the plasma phosphate level falls, and the alkaline phosphatase level increases. In inherited hypophosphataemia (vitamin D-resistant rickets) plasma phosphate is low, but the plasma calcium is normal and the alkaline phosphatase may also be normal. Renal glomerular failure causes an increase in plasma phosphate, urea, and creatinine, and hypocalcaemia, and in the rare renal tubular syndromes there may be a marked systemic acidosis. In patients with osteomalacia the urine should always be examined for the presence of glucose and protein. If these are present, it is important to check for the aminoaciduria characteristic of renal tubular disorders.

The measurement of vitamin D metabolites is becoming routine, and a low plasma 25(OH)D level is a good indication of vitamin D deficiency. Estimation of plasma $1,25(OH)_2D$ is important to elucidate the very rare causes of rickets, and particularly to distinguish between type I (low $1,25(OH)_2D$) and type II (high $1,25(OH)_2D$) vitamin D-dependent rickets.

Radiology

The radiological appearances differ according to whether growth has ceased or not. In rickets the main abnormalities are at the ends of the long bones, where the width of the growth plate is increased, and the metaphysis is widened, cupped, and ragged (Fig. 8). Osteomalacia may show the deformities previously described, but the radiological hallmark of active osteomalacia is the Looser's zone (Fig. 9). This is a ribbon-like area of defective mineralization, which may be found in almost any bone but is seen particularly in the long bones, the pelvis, and the ribs, and also around the scapulas. Looser's zones may be bilateral and symmetrical; in bones such as the femur they occur on the medial border of the shaft or neck and are usually single, in contrast to the multiple fissure fractures on the lateral convexity of the bone in Paget's disease. In osteomalacia the vertebral bodies are often uniformly biconcave, to produce an appearance likened to a fish spine. Additionally, in renal glomerular osteodystrophy, the endplates may become relatively more dense than the rest of the vertebral body, to produce the so-called 'rugger jersey' spine. In the adult with inherited hypophosphataemia the bones may also become deformed, buttressed, and dense; in this disorder calcification of the tendons and ligaments at their insertions (enthesiopathy) and of the vertebral ligaments can produce an appearance similar to that of ankylosing spondylitis. Ossification of the ligamenta flava narrows the spinal canal and compresses the spinal cord and its roots. This is well shown on CT scans. In patients with osteomalacia and hypocalcaemia the radiological features of secondary hyperparathyroidism appear with subperiosteal bone resorption that affects the phalanges, the pubic symphysis, and the outer ends of the clavicles. In rickets, periostitis of the distal ends of the long bones, such as the radius and ulna, often occur.

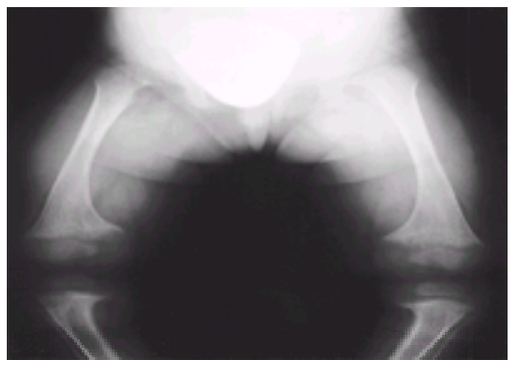

Fig. 8 The radiological appearance of rickets in a child with inherited hypophosphataemia. The growth plates are widened and the metaphyses cupped and ragged.

Fig. 9 To demonstrate the bilateral Looser's zones on the medial border of the femora in a woman with osteomalacia due to adult Fanconi's syndrome.

The most extreme effects of parathyroid overactivity are seen in the skeleton of the child with renal osteodystrophy, where the region of the growth plate and metaphyses may fracture (an appearance likened to a 'rotting stump'). A bone scintigram may be very useful in cases of osteomalacia, demonstrating multiple pathological fractures often not seen on the plain films. The appearance is similar to that of bony metastases.

Bone biopsy

The diagnosis of osteomalacia is often clear without examining the bone. Where doubt exists, a transiliac biopsy examined before and after decalcification will demonstrate the failure of mineralization and the wide osteoid seams. It is important to take all surgical opportunities to examine bone, particularly during operations on fractured femurs in the elderly.

Other investigations

Further investigation is not usually needed to diagnose osteomalacia, but may be necessary to identify its cause. Thus patients with vitamin D-deficient rickets and osteomalacia will have a low plasma 25(OH)D, but not all subjects with such low levels have osteomalacia. In the very rare condition of vitamin D-dependent rickets, measurement of circulating 1,25(OH)$_2$D will be necessary to distinguish the absence of 1a-hydroxylase from resistance to 1,25(OH)$_2$D. Further, CT scanning may help to identify the presence of a mesenchymal tumour causing hypophosphataemic osteomalacia.

Diagnosis

Osteomalacia is not difficult to diagnose once it is thought of. It should be distinguished from other forms of metabolic bone disease (Table 4), from other causes of proximal muscle weakness, and from other disorders causing bone pain. In patients with proximal muscle weakness, polymyalgia rheumatica, thyrotoxic myopathy, muscular dystrophy, neoplastic neuropathy, dermatomyositis, and polymyositis all need to be considered. Multiple myeloma and leukaemia may need to be excluded as causes of pain. Provided that the plasma calcium, phosphorus, and alkaline phosphatase levels are always measured in patients with these symptoms, those with osteomalacia should be easily identified. Patients with psychological illness may have an abnormal gait and complain of pain and weakness in their limbs, but in such patients the biochemistry will be normal. In practice, symptoms of pain and stiffness often first lead the patient with osteomalacia to a rheumatologist.

Treatment

Rickets and osteomalacia should respond rapidly to vitamin D (or to its metabolites) in an appropriate dose, and the response may be a useful way of confirming the diagnosis. Increased mobility with an increase in muscle strength may be the first clinical response, despite a temporary increase in bone pain. Biochemically, plasma phosphate and urine hydroxyproline levels are the first to increase. The alkaline phosphatase level may show a temporary rise and then fall slowly to normal levels. As the plasma calcium and 25(OH)D concentrations increase towards normal, the parathyroid hormone concentration falls.

The effective dose and the particular vitamin D preparation depends on the cause of the osteomalacia. That due to vitamin D deficiency will respond to microgram doses, but it is often useful to give considerably more than this, such as calciferol 1.25 mg daily for 1 to 2 weeks only. Where there is doubt about compliance, vitamin D may be injected intramuscularly in one large dose (up to 15 mg, 600 000 units). Lack of a response to microgram doses suggest that the osteomalacia is not due to simple vitamin D deficiency but, for instance, to malabsorption or renal failure. It is particularly in the last group that the 1a-hydroxylated metabolites of vitamin D are effective (see Chapter 20.8). Clearly, underlying disorders must be treated at the same time: for example, patients with coeliac disease will need a gluten-free diet.

Particular forms of osteomalacia and rickets

Nutritional osteomalacia

In the United Kingdom and other Northern European countries, so-called nutritional osteomalacia occurs particularly amongst the elderly and in Asian immigrants of all ages. In the elderly, the high incidence of osteomalacia is mainly due to their poor exposure to sunlight and to a low intake of vitamin D; and may be contributed to by the effects of drugs such as anticonvulsants and by increasing renal glomerular failure. Since the elderly are often housebound, they may develop osteomalacia despite a sunny climate. Certainly, the prevalence of osteomalacia in the elderly population is significant. The frequency of osteomalacia in patients with fractures of the femoral neck is also higher than previously suspected, but figures of up to 30 per cent, which continue to be reported (according to the histological definitions used), are probably overestimates. Osteomalacia should always be excluded in elderly people with bone disease, and particularly in those with femoral neck fractures. Where possible this should be done by histological examination of bone taken at operation or by biopsy. When this is not appropriate, a therapeutic trial with vitamin D is often useful. Since it is often difficult to define osteomalacia accurately in elderly people it is important to consider the use of such empirical treatment. In the geriatric population the mean concentration of 25(OH)D is much lower than in non-elderly patients; it shows the usual seasonal variation, with lowest values in the winter and early spring and highest in late summer.

Asian immigrants to the Northern hemisphere develop osteomalacia and rickets more often than the indigenous populations. There are probably several reasons for this. They tend to live in northern cities away from sunlight and, especially in women, do not expose their skin to the limited ultraviolet light. Where dermal synthesis of vitamin D is limited, dietary factors become more important, and it is particularly those on a meat-free diet containing chapattis who develop osteomalacia. The role of chapattis and the phytate they contain is not yet fully understood. Phytates bind to calcium so preventing its absorption, and it can be shown, at least experimentally, that reduced calcium absorption increases the vitamin D requirement by increasing its parathyroid-mediated breakdown. It has been suggested that such a mechanism of reduced calcium absorption may also contribute to the osteomalacia of malabsorptive syndromes, such as that following partial gastrectomy.

Pigmentation of the skin can be shown experimentally, using a standardized dose of ultraviolet light, to reduce vitamin D synthesis, but in practice this is of little significance. Since North European immigrants of Afro-Caribbean descent have a lower incidence of rickets than Asians in the same environment, it is clear that factors other than skin colour are important.

As in the elderly, 25(OH)D levels can be very low, especially in Asian immigrants. They increase in the summer, when there may be spontaneous healing of rickets. Important work in Glasgow has shown that Asian rickets can be prevented by fortifying food such as chapatti flour with vitamin D, although the incidence of osteomalacia in Asian adults remains unaffected. Other local lifestyle changes will also influence the diet of children.

Osteomalacia and malabsorption

Coeliac disease (gluten-sensitive enteropathy) (Chapter 14.9.3) is a relatively common cause of osteomalacia. It should be suspected at any age, and confirmed by the presence of circulating endomysial antibody and, if necessary, by a small intestinal biopsy showing an atrophic mucosa. Other causes of malabsorption vary in their frequency according to surgical practice. Thus it is well established that osteomalacia follows classic partial gastrectomy, but the actual incidence is debated and its cause is probably multifactorial. Postgastrectomy subjects tend to take little vitamin D in their diet and there is defective calcium absorption. Available evidence suggests that clinical osteomalacia is rare after vagotomy and pyloroplasty. Osteomalacia can also follow the removal of long segments of small intestine for conditions such as Crohn's disease, and complicates some intestinal bypass operations used for extreme obesity.

Osteomalacia and liver disease

Osteomalacia is uncommon in those with liver disease; in theory it may be due to a number of factors such as malabsorption of vitamin D and its defective 25-hydroxylation. Most research has concerned the osteomalacia of biliary cirrhosis, and osteomalacia in chronic liver disease appears to be a complication related to prolonged cholestasis.

Osteomalacia and renal disease

It is important to distinguish the osteomalacia and rickets of renal glomerular failure from that attributable to renal tubular disorders. Bone disease in renal glomerular failure (renal glomerular osteodystrophy) is dealt with elsewhere (see Chapter 20.8); this includes bone disease in the dialysed patient and the effects of aluminium. Renal glomerular osteodystrophy is a complex disease with excessive bone resorption, defective bone mineralization, and in some cases osteoporosis. Previously it was treated with large doses of native vitamin D; current therapy now includes 1a-hydroxy-cholecalciferol or 1,25(OH) $_2$D.

Many renal tubular disorders lead to osteomalacia (Chapter 20.8) (Table 7). Of these, the most common is inherited hypophosphataemia, so-called vitamin D-resistant rickets, which is normally inherited as an X-linked dominant characteristic; here, the main abnormality is hypophosphataemia due to a reduction in the maximum renal tubular reabsorption rate of phosphate. Some patients in a family will have hypophosphataemia alone, whereas others will have hypophosphataemia with accompanying bone disease. It is now known that inherited hypophosphataemia is caused by mutations in the *PEX* or *PHEX* gene, the cognate protein of which has the features of an endopeptidase. Endopeptidases degrade or activate peptide hormones. It is not yet known how the mutations produce the defect in phosphate homeostasis. Since the 1,25$(OH)_2$D levels are normal where the plasma phosphate is low, it is proposed that the sensitivity of the 1a-hydroxylase enzyme is reduced. Children with hypophosphataemic rickets or osteomalacia are unlike patients with other forms of rickets. They present with deformity but are otherwise well, without muscle weakness; however, growth is defective and their eventual height is usually less than 150 cm. Apart from hypophosphataemia there may be no other abnormality in the biochemical values routinely available, and the plasma alkaline phosphatase level can be normal for age. Radiographs show severe rickets, and later the bones are often dense with buttressing and exostoses. The enthesiopathy with ossification of the ligamenta flava can lead to para-plegia. Ligamentous calcification may also contribute to deafness. Finally, abnormal teeth in this disorder cause periapical translucencies and may lead to abscesses.

The treatment of inherited hypophosphataemia is controversial. For many years its mainstay was large doses of vitamin D; this posed a continuous danger of vitamin D poisoning and did not correct the eventual short stature. There is an improvement in growth rate when oral phosphate is given in addition to vitamin D, but the condition does not appear to respond to phosphate alone. More recently, it has been shown that combined oral phosphate and 1,25(OH) $_2$D produces healing of epiphyseal and trabecular bone and this is now the recommended treatment. This combination produces bone healing and increases eventual stature. However, it is still unusual for affected patients to have an eventual height of more than 1.5 m (5 ft). Accounts of the effects of medical treatment on deformity and height differ; the necessity for corrective osteotomy on the lower limbs is less than previously, but discussion with an orthopaedic surgeon is important.

It is also important that the parents should know the genetics of this condition. Because the defect in phosphate transport is inherited as a dominant on the X chromosome, an affected mother transmits the condition to 50 per cent of her children regardless of their gender. All the daughters of an affected father will have the disease, but none of his sons. In general, affected sons have a more severe disease while some affected daughters may be asymptomatic. Diagnosis can be made from birth, but this demands accurate knowledge of the normal plasma phosphate level at that age. Now that more is known about the exact gene location, prenatal diagnosis may be possible in future. Recently cloned human X-chromosome sequences that reveal restriction fragment length polymorphisms have been used in linkage studies of affected families to map the hypophosphataemic rickets gene. Flanking markers are potentially useful in the identification of mutant gene carriers and in presymptomatic diagnosis, but the distance between these markers and the hypophosphataemic gene is still large, at approximately 10 million base pairs. Hypophosphataemic animal models continue to help in furthering understanding of this disorder. A recently described murine *gy* mutation, in which hypophosphataemia is associated with gyratory activity, has no clear human equivalent. Rare human variants include an autosomal dominant form of hypophosphataemia.

Other renal tubular osteomalacic syndromes include hypophosphataemic osteomalacia presenting in adult life, which may be due to a tumour (see below), inherited and acquired forms of renal tubular acidosis, and rickets associated with multiple renal tubular defects and generalized aminoaciduria (Fanconi's syndrome). Renal tubular acidosis may be proximal or distal, with an inability to reabsorb bicarbonate or to acidify the urine. The osteomalacia may be cured by giving bicarbonate alone or with vitamin D. A persistent acidosis with resultant osteomalacia may also result from ureterosigmoid anastomosis. The commonest cause of Fanconi's syndrome in childhood is nephropathic cystinosis or cystine-storage disease, where there is a widespread deposition of cystine crystals throughout the tissues, and in which thirst, polyuria, dehydration, photophobia, and loss of weight begin at about the age of 1 year. The rickets will heal with the correction of the acidosis, and the administration of phosphate and 1a(OH)D; renal transplantation corrects the renal failure and prolongs survival, but does not prevent non-renal complications.

Other rare causes of renal tubular rickets and osteomalacia with generalized aminoaciduria are inherited, such as Wilson's disease and the X-linked oculocerebral

renal syndrome, or acquired, such as multiple myeloma, cadmium poisoning, and the toxic effects of ifosfamide used in the treatment of childhood malignant disease.

Anticonvulsant osteomalacia

In patients treated with anticonvulsants the incidence of rickets and osteomalacia is higher than normal. This has been attributed to the induction by the anticonvulsants of hepatic enzymes which metabolize vitamin D to biologically inactive derivatives. However, epileptic patients in institutions are often vitamin D-deficient because they are deprived of sunlight, and osteomalacia in such patients probably has several causes.

Tumour rickets

An unusual form of hypophosphataemic rickets or osteomalacia occurs in patients who have mesenchymal tumours, often of a particular histological type, namely sclerosing haemangioperiocytomas or non-ossifying fibromas. A tumour should be considered in any adult who develops hypophosphataemic osteomalacia, particularly with prominent myopathy. The disorder is improved by oral phosphate and cured by removal of the tumour. The way in which the tumour induces hypophosphataemia and subsequent osteomalacia is unknown, but current evidence suggests that it interferes with the renal 1a-hydroxylation of 25(OH)D, since the circulating levels of $1,25(OH)_2D$ are abnormally low but rapidly return to normal when the tumour is removed. Rarely, hypophosphataemic osteomalacia may become apparent in adults with neurofibromatosis and polyostotic fibrous dysplasia.

Osteogenic osteomalacia has also been described in cases of prostatic and small-cell carcinoma of the lung.

Vitamin D-dependent rickets

Patients with these very rare, recessively inherited forms of rickets show the features of severe rickets without vitamin D deficiency. There are at least two types of vitamin D-dependent rickets. In type I the activity of the renal 1a-hydroxylase is reduced so that the concentration of $1,25(OH)_2D$ is abnormally low. However, it can be increased by large doses of the native vitamin, which shows that the enzyme block is not complete. In type II there is an end-organ resistance to $1,25(OH)_2D$, which is present in high concentrations. In both forms there is severe rickets and myopathy from infancy; in type II, lifelong total alopecia is a striking feature. Vitamin D-dependent rickets type I responds to very large doses of vitamin D or physiological doses of $1,25(OH)_2D$. Type II may also respond to large doses of vitamin D or its metabolites, or to prolonged intravenous calcium, but some recorded cases suggest that recovery occurs spontaneously with age.

Recent work on type-II, vitamin D-dependent rickets (otherwise known as hereditary $1,25(OH)_2D$-resistant) shows that the $1,25(OH)_2D$-receptor defects, which are responsible for the end-organ resistance in this disease, are due to a variety of point mutations in the gene for the 1,25 receptor, either at its steroid- or DNA-binding domains.

Phosphate-deficiency rickets

If patients ingest large amounts of phosphate-binding drugs, such as aluminium hydroxide, a form of hypophosphataemic osteomalacia may develop. This differs clinically from inherited hypophosphataemic osteomalacia by the presence of severe muscle weakness. Other biochemical features include increased calcium absorption with hypercalcuria, associated with an increase above normal in the concentration of $1,25(OH)_2D$.

Paget's disease of bone

Paget's disease of bone, osteitis deformans, was described more than a century ago, but existed for many years before. It is the most common of the so-called metabolic bone diseases after osteoporosis. Its hallmark is excessive and disorganized resorption and formation of bone. Its cause is unknown, but recent studies on pagetic bone cells, particularly osteoclasts, have provided clues. The new generation of bisphosphonate drugs now provide effective treatment.

Pathophysiology

The natural history of Paget's disease is similar to that of a multicentric neoplasm or a slow virus disease that begins in young adult life. Electron microscopy shows virus-like inclusion bodies in the osteoclasts of patients with Paget's disease. Immunofluorescence studies suggest that these could represent the measles or respiratory syncytial virus. Another candidate has been the canine distemper virus, but the results of polymerase chain reaction amplification of reverse transcribed DNA from Paget's tissue to identify the putative virus remain controversial.

Histology shows multinucleate osteoclasts which appear to be resorbing bone, and busy osteoblasts which appear to be replacing it; these activities are closely linked. There is also excess fibrosis in the marrow. The bone matrix is laid down in all directions and partially loses its birefringence and strength. Mineralization may be defective, probably because of the excessive rate at which the organic bone matrix is laid down. The cement lines and the mosaic appearances of the bone result from the tidemarks of resorption followed by formation. Osteosarcoma which occurs in Paget's disease is presumably the result of the excessive and prolonged activity of the bone cells. Pagetic bone is large, vascular, and deformed. Its physical characteristics depend on the stage of the disorder and it may be hard or soft. In any event, it fractures more readily than normal.

Incidence

Paget's disease occurs in about 3 to 4 per cent of subjects over 40 years of age, is more common in men than in women and its frequency increases with age. It is not unknown in younger people. In Britain, about 750 000 people may have Paget's disease, of whom fewer than 5 per cent have symptoms. It appears to be a peculiarly Anglo-Saxon affliction, being very rare in countries such as Scandinavia and Japan. Within England, early radiological surveys in the 1970s showed that it occurs most often in Lancashire towns and in northern industrial regions (Table 8). It is also more frequent in recent British immigrants to Western Australia than in the Western Australian population, but less frequent than in those relatives who remained in Britain. Such studies do not distinguish between the effect of environment and heredity. In a disorder as common as Paget's disease many striking examples of 'familial' Paget's disease occur by chance. Recent studies suggest a reduction in the prevalence of Paget's disease.

Clinical features

Pain, deformity, fracture

In Paget's disease the bone itself may be painful, or pain may be due to arthritis of a nearby joint, to an associated fracture, or to the development of sarcoma. It has been suggested that there is a specific type of hip joint disorder associated with Paget's disease. Bone pain could be due to stretching of the periosteum, since this part of the bone (and the vessels within bone) contain nerves sensitive to pain. Clinically, the affected bones are enlarged, deformed, and warm. The enlargement is clearly seen in bones such as the tibia and the skull; in the former the bone is typically bowed forwards; the latter shows a characteristic enlargement of the vault that is said to look like a soft beret, or 'tam-o'-shanter', which appears to descend over the ears. Other long bones may become bent and a kyphosis may develop. Although any of the bones can be affected, including the maxilla and the phalanges, the most common sites for Paget's disease are the pelvis and the spine. Fracture may be the first symptom of undiagnosed Paget's disease, for instance at the junction of a resorbing front with normal bone (Fig. 10), or across a fissure fracture (see Fig. 11).

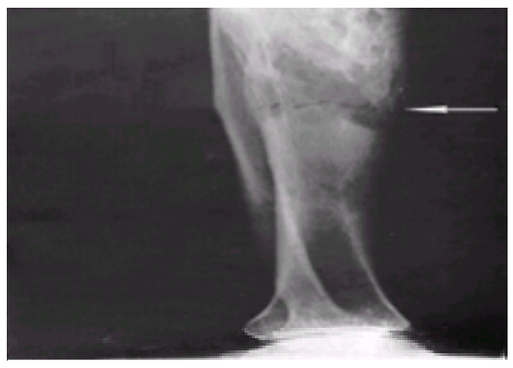

Fig. 10 A fracture in the region of a resorbing front in a pagetic bone (arrowed). Proximal to the area of bone resorption the cortex is thickened and the bone widened by disorganized formation of new bone.

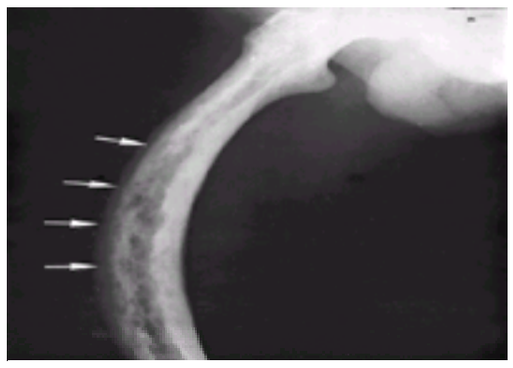

Fig. 11 Multiple microfractures ('fissure fractures' arrowed) on the convex surface of a pagetic femur.

Deafness and nerve compression

Deafness in Paget's disease is one of its most disabling symptoms and responds little to treatment. It has many causes, of which nerve compression is only one.

Most nerves can be compressed by enlarging pagetic bone. The spinal cord is particularly at risk, due to the combined effects of increased bone mass, vertebral collapse, and excessive vascularity. Paraplegia or cauda equina lesions may occur. Alterations in the shape of the skull may produce multiple cranial nerve palsies and brainstem lesions, with dysphagia, dysarthria, and ataxia. Basilar invagination with obstruction of cerebrospinal fluid drainage can lead to internal hydrocephalus, raised intracranial pressure, and confusion.

Heart failure

In severe Paget's disease, cardiac output may be increased by the excessive vascularity of the affected bones, but there is no convincing evidence of large arteriovenous shunts within the skeleton. The heart failure which results may be of the high-output variety, but this is excessively rare. Since heart failure and Paget's disease of bone are common in the elderly, their occurrence together is almost always coincidental.

Sarcoma

The incidence of sarcoma in Paget's disease has sometimes been overestimated in the past; it probably occurs in 1 per cent or less of those with symptoms. Paget's sarcoma often occurs in the humerus, although Paget's disease itself is most common in the pelvis and spine. Sarcoma should be considered in a patient known to have Paget's disease if pain has developed for the first time, or worsened, or if deformity has altered. Radiologically, the appearance of the pagetic bone alters, with evidence of bone destruction (Fig. 12); the tumours occur most often in the medulla. A recent review of 85 bone sarcomas associated with Paget's disease confirmed the humerus as a high-risk site. Rapidly worsening pain was the main symptom; lytic lesions were more common than sclerotic; periosteal reaction was uncommon; and radionuclide bisphosphonate scintigraphy usually showed areas of decreased uptake (contrasting with the underlying pagetic bone).

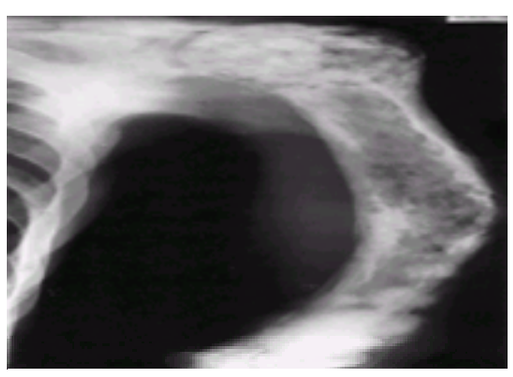

Fig. 12 A sarcoma in the upper end of the left humerus in a 70-year-old man with Paget's disease. The destructive lesion in the proximal humerus has been treated with radiotherapy; there are secondary deposits around the distal end of the bone.

Associated disorders

Paget's disease is said to be associated with other disorders such as osteoarthritis, gout, vascular calcification, and articular chondrocalcinosis. Since all these occur more often in the elderly the associations have little significance.

Investigations

Biochemistry

There is a marked increase in the level of plasma alkaline phosphatase, derived from the overactive osteoblasts, which is roughly related to the extent of clinical and radiological involvement with Paget's disease. In contrast, the acid phosphatase (derived partly from osteoclasts) level is only slightly increased. The rapid turnover of bone matrix collagen increases urinary hydroxyproline (and hydroxylysine), in proportion to the increase in alkaline phosphatase and also the urinary excretion of crosslinked collagen-derived peptides. Plasma calcium and phosphate levels are normal; hypercalcaemia suggests coexistent hyperparathyroidism, malignant disease, or immobility.

Radiology

The radiological appearances of Paget's disease are legion. The most characteristic is an increase in size of the affected bone. Resorption predominates early in the disease and in the young patient. A resorbing front may be seen in a long bone (as a flame-shaped area) (see Fig. 10) or in the skull (as 'osteoporosis circumscripta'). Excessive resorption is inevitably followed by disordered formation, and at this stage the bone becomes thick and deformed. In elderly subjects the affected bone may be very osteoporotic and liable to fracture. Multiple partial fractures (microfractures, fissure fractures) are common on the deformed convex surface of long bones (see Fig. 11), particularly the femur and tibia.

The use of bone-scintigraphic agents (such as $^{99}Tc^{m}$-labelled disodium etidronate (**EHDP**, [disodium] ethane-1-hydroxy-1,1-diphosphonate) has been particularly informative in Paget's disease. Affected bones take up the isotope avidly, which demonstrates both the extent of the bone lesions and the effects of treatment. In one study, 180 patients with Paget's disease underwent whole-body scintigraphy and 826 lesions were identified—one-third of the patients had only one lesion, and only 10 patients had no symptoms. The increase in plasma alkaline phosphatase and urinary total hydroxyproline was proportional to scintigraphic involvement, and patients with skull involvement had the highest values. Apart from the number of sites involved, any distinction between monostotic and polyostotic disease appeared to be artificial.

Diagnosis

The diagnosis of Paget's disease is usually obvious. Bone biopsy is useful to exclude other generalized bone diseases, such as osteomalacia, as well as to confirm Paget's disease. Paget's disease may initially be confused with osteomalacia because of the high plasma alkaline phosphatase level; rarely, an elevated plasma calcium should suggest additional hyperparathyroidism or malignant disease. In prostatic carcinoma with osteoblastic bone secondaries, the dense bones are not enlarged (as they are in Paget's disease) and the acid phosphatase level is considerably and disproportionately increased in relation to that of alkaline phosphatase. Of many other conditions with similar radiological appearance, fibrous dysplasia (see below), in which the alkaline phosphatase may also be slightly increased, may be difficult to distinguish; in the generalized form the unilateral bone lesions, pigmentation, and sexual precocity (in the female) are characteristic. Another very rare disorder usually mistaken for Paget's disease is fibrogenesis imperfecta ossium (see below), where the bone trabeculae are thickened without bony enlargement and there are multiple abnormal fractures.

Treatment

Many patients with Paget's disease require no treatment, but it may be required for symptoms, to suppress the activity of the disease, and to prevent its further progress. Indications include bone pain, nerve compression, and the suppression of vascularity before elective orthopaedic surgery. Since medical treatment is now so effective, these indications may be widened especially in young people.

Medical treatment

Patients with painful Paget's disease should first be treated with a simple analgesic. Where possible it should be determined whether the pain is directly due to the bone disease or to associated arthritis. Specific treatment aimed at the pagetic bone should be considered for those who have pain due to bone disease despite analgesia, or have the complications of deformity, nerve compression, deafness, or very rarely heart failure. This treatment should also be considered in the young person with Paget's disease to prevent further progression. There is no evidence that the rapid course of pagetic sarcoma is altered by any treatment. Of the many agents previously tried in Paget's disease, such as aspirin, fluoride, and corticosteroids, only three are currently in use, mithramycin, bisphosphonates, and calcitonin. Mithramycin is now rarely used. It is an antimitotic agent given intravenously which is hepatotoxic in high doses. It may rapidly abolish pain in Paget's disease, and rapidly reduce the plasma alkaline phosphatase level, but the effect is usually temporary. Mithramycin has been used on its own and in combination with bisphosphonates or calcitonin.

The bisphosphonates (once called diphosphonates) are a series of compounds with a P–C–P structure resistant to the naturally occurring phosphonates and pyrophosphatases. They are effective both orally and parenterally and reduce excessive bone turnover in Paget's disease. EHDP (Didronel®), one of the early bisphosphonates, also interferes with mineralization if given in high doses (20 mg/kg body weight); subsequent derivatives such as dichloromethylene diphosphonate (**Cl2 MDP**) and 3-amino-hydroxypropylidine-l,l-bisphosphonate (**APD**, pamidronate) do not appear to do this. According to their dose, the bisphosphonates may take up to 6 months to produce their effect on symptoms, histology, and biochemistry. The recommended dose for EHDP is 5 mg/kg per day for up to 6 months. It also has been used in combination with calcitonin, and together these agents suppress Paget's disease more effectively than when given alone. The biochemical effects of EHDP appear to last for a long time (possibly several years) after the drug is stopped. Many new bisphosphonates have now been developed based on side-chain substitutions in the basic P–C–P structure. The aminobisphosphonates are particularly effective. The new bisphosphonates are many times more potent than etidronate. They include pamidronate, tiludronate, alendronate, and risedronate. They may produce almost complete and permanent suppression of Paget's disease without significant side-effects. The dose regimes and expected responses are dealt with in larger reviews (see the Further reading list).

The calcitonins are less widely used for the treatment of Paget's disease, and salmon calcitonin is the most effective commercially available form. Various dose regimens are used, for which 100 IU given three times a week is average. Injected calcitonin may produce nausea and vomiting; if side-effects are troublesome, it is best given in the evening together with an antiemetic. Its main effects are seen during the first 3 months of treatment, and continued treatment is ineffective, especially when the alkaline phosphatase level has ceased to decline.

Antibodies to calcitonin do develop but are not necessarily related to calcitonin 'resistance'. Indications for the bisphosphonates and calcitonins are different. Calcitonin is preferred to treat bone pain, for osteolytic Paget's disease, and for preoperative treatment. Some evidence suggests that calcitonin may halt the progression of deafness. Spinal cord compression is also alleviated. Thus treatment of eight patients with paraparesis due to pagetic vertebrae with either calcitonin or bisphosphonate produced marked clinical improvement, at least comparable to the results of surgical decompression. Calcitonin can also be given preoperatively to reduce excessive bleeding when operations such as total hip replacement have to be performed on pagetic bone. Calcitonin can now be given by the nasal route, which is more acceptable to the patient but less effective.

Surgical treatment

Fractures through pagetic bone require the usual surgical treatment, although union may be delayed. Where fracture occurs through a deformed bone, this deformity should be corrected. In addition, elective osteotomy and intramedullary nailing may be considered for a severe long-bone deformity. Spinal cord compression not responding to medical treatment requires surgery. Rarely, hydrocephalus may require a ventriculojugular shunt. Whatever form of surgery is undertaken, it is important that the period of immobility is as short as possible, to avoid the development of hypercalciuria and hypercalcaemia.

Parathyroids and bone disease

Knowledge of the biochemistry of parathyroid hormone has expanded so rapidly that it now occupies a large and deserved part of any clinical description of parathyroid disorders (see Chapter 12.6). The close relationship between these endocrine glands and the skeleton becomes less obvious with increasing recognition of the many ways in which parathyroid disease presents. However, primary hyperparathyroidism was first identified because of its effects on bone, and only later was it realized that it might more often present with renal stones, with pancreatitis, and with the signs and symptoms of hypercalcaemia—or be a chance discovery as a result of multichannel biochemical analysis.

The subject is discussed further in Chapter 12.6.

Molecular advances

With the discovery of the calcium-sensing receptor (**CaR**) and extensive work on the cause of the multiple endocrine neoplasia syndromes our understanding of the rarer causes of abnormal plasma calcium levels has considerably increased. Thus missense mutations of the *CaR* gene cause both familial benign hypercalcaemia and neonatal hyperparathyroidism, whereas gain-of-function mutations in this receptor can cause familial hypoparathyroidism. Multiple endocrine neoplasia (**MEN**) syndromes have traditionally been divided into two types—type 1 with hyperparathyroidism, pituitary adenomas, insulin- and gastrin-secreting tumours of the pancreas, and gastric hyperacidity (Zollinger–Ellison syndrome); and type 2, also known as Sipple's syndrome, with hyperparathyroidism, medullary carcinoma of the

thyroid, and phaeochromocytoma. The molecular elucidation of these differences has identified subgroups. In MEN1 the principal genetic abnormality involves mutations in the *MEN1* gene together with loss of alleles on chromosome 11; in MEN2 (both A and B subgroups) there are mutations in the *RET* proto-oncogene on chromosome 10.

Hypercalcaemia

Of the known causes of hypercalcaemia in hospital inpatients, neoplasm is the most important (Table 9). It should always be considered and excluded clinically. The relative frequency of the causes of hypercalcaemia varies according to the population studied. In apparently healthy outpatients primary hyperparathyroidism is the most frequent cause. In those patients with primary hyperparathyroidism, with hypercalcaemia, hypophosphataemia, hyperphosphatasia, and radiological evidence of osteitis fibrosa, and without clinical evidence of neoplasm, little further investigation is needed. Since only a few patients with hyperparathyroidism have clinical bone disease, further differentiation from other causes of hypercalcaemia is usually necessary. In practice, this means the exclusion of neoplasm, sarcoidosis, thyrotoxicosis, vitamin D overdosage, treatment with thiazide diuretics, or the 'milk alkali' syndrome. The subject is addressed further in Chapter 20.8.

Secondary (and tertiary) hyperparathyroidism

Where hypocalcaemia is prolonged, as in renal glomerular failure or gluten-sensitive enteropathy, the parathyroid glands increase both their size and activity in an attempt to restore the plasma calcium level to normal. This increases bone resorption and is a particular feature of renal glomerular osteodystrophy. Occasionally hypercalcaemia develops and persists in such patients. It has been proposed that one of the hyperplastic parathyroid glands becomes autonomous and thus the label 'tertiary hyperparathyroidism' has been given. Hypercalcaemia may also occur after renal transplantation (see Chapter 20.8).

Hypoparathyroidism (see also Chapter 20.8)

Parathyroid insufficiency may occur after surgical removal of the parathyroids, in idiopathic hypoparathyroidism, and in a familial form of hypoparathyroidism which is often associated with manifestations of autoimmune disease, including moniliasis, malabsorption, thyroid and adrenal failure, and pernicious anaemia. In such patients the levels of immunoreactive parathyroid hormone (**PTH**) are undetectably low but the cAMP response to exogenous PTH is maintained. This distinguishes parathyroid insufficiency from pseudohypoparathyroidism, in which the biochemical features of hypoparathyroidism are associated with characteristic skeletal abnormalities (Albright's hereditary osteodystrophy). Pseudohypoparathyroidism is inherited as an autosomal dominant. In the most common form, the cAMP response to exogenous PTH is defective, and the circulating level of immunoreactive PTH is high. Variations of pseudohypoparathyroidism appear to exist, and disorders are described in which the cAMP response is present but there is still end-organ resistance, and also where the cAMP response is restored by giving vitamin D. Patients who have the skeletal manifestations of pseudohypoparathyroidism but with normal biochemistry may be found in families with pseudohypoparathyroidism, and to them the term 'pseudopseudohypoparathyroidism' is applied. Investigation has shown that the loss-of-function end-organ resistance is due to point mutations in the genes controlling one component of the G-protein signalling system.

So far as the skeleton is concerned, the most striking changes are found in pseudohypoparathyroidism. Clinical features include mental simplicity, short stature, round face, short neck, and abnormal metacarpals (or metatarsals), of which the most common change is shortness of the fourth and fifth. The bones may be excessively dense, and widespread ectopic calcification and ossification may also occur, in the basal ganglia and the subcutaneous tissues respectively. Treatment of the hypocalcaemia is the same as for idiopathic hypoparathyroidism, with 1a-hydroxycholecalciferol.

Osteogenesis imperfecta: the brittle bone syndrome

This disorder, which has emerged from the status of an obscure osteopathy to a metabolic bone disease, provides remarkable lessons concerning the effects of mutations in the collagen genes. The correlation between genotype and phenotype is by no means exact and leaves interesting problems.

Osteogenesis imperfecta is said to occur in about 1 in 20 000 births; since the milder forms may never be diagnosed, this could be an underestimate. It is a leading cause of lethal short-limbed dwarfism and crippling skeletal dysplasia. There is no convincing evidence of different racial frequency. Many patients with osteogenesis imperfecta do not fit easily into the Sillence classification (Table 10), and in some cases hypermobility and features of the Ehlers–Danlos syndrome (see below) are dominant.

Pathophysiology

Osteogenesis imperfecta involves those tissues that contain the main fibrillar collagen, type I. These include particularly bone and dentine, but also the sclerae, joints, tendons, heart valves, and skin. The pathology in bone varies with the type and severity of the disease, with age, previous fracture, and surgery. The skeletal effects of osteogenesis imperfecta are most severe in the lethal forms (type II) and at the region of the growth plate. There is faulty conversion of apparently normal mineralized cartilage to defective bone matrix. The collagen fibres are thin but show the normal striated pattern. The endoplasmic reticulum of the osteoblasts is dilated by retained mutant collagen. The bone structure is completely disorganized and structurally useless. In type III osteogenesis imperfecta, which is less severe, there are variable amounts of woven immature bone, with disorganized trabeculae and an apparent excess of osteocytes—as in other forms of the disorder. At the growth plate there are multiple islands of cartilage in the epiphyses and metaphyses. Accounts of the bone pathology in type IV are sparse. Defective mineralization is described in rare forms of osteogenesis imperfecta.

In mild, type I osteogenesis imperfecta there is a reduction in the amount of bone (and hence in measured bone mineral density) and defective bone formation at the cellular level, such that the osteoblasts each make approximately half as much bone collagen as normal. The result is an osteoporotic bone with an apparent excess of osteoblasts and osteocytes. This appearance of 'hyperosteocytosis' suggests (to some) an increase in bone turnover rate. The overall bone structure is otherwise normal, apart from occasional woven bone. In affected dentine, the odontoblasts produce short, branched dentinal tubules and fill in the dental pulp. In the ear, the auditory ossicles may be imperfect or fractured.

The reduction in collagen is repeated in non-skeletal tissues. Thus, the sclerae are thin (leading to their blueness since the pigmented coat of the choroid becomes visible), the tendons are gracile and weak, the thin heart valves may become incompetent, and the aortic root dilated.

Clinical features

Type I is the most frequent and least serious form, and accounts for 60 per cent of all patients with the disorder. Fractures can occur in the perinatal period or even be delayed until the early perimenopause. After the menopause the overall fracture rate has been recorded at seven times more than in the normal population, and the vertebral bone mineral content in adults with type I osteogenesis imperfecta has been found to be 70 per cent of normal.

Childhood fractures in type I osteogenesis imperfecta may be numerous but rarely lead to deformity unless treated inappropriately. Any type of fracture can occur; they become less frequent with age. Overall, fractures are more frequent in the lower limbs. Significant scoliosis is rare. The skull shows interesting changes; in addition to multiple wormian bones (Fig. 13) (which can occur in other disorders, such as pyknodysostosis, cleidocranial dysostosis, Menkes' syndrome, Prader–Willi syndrome, progeria, and, rarely, in normal subjects), the vault may overhang the base, leading to basilar impression requiring surgical correction.

Fig. 13 The innumerable centres of ossification in the occipital region, wormian bones, in an infant with severe (type III) osteogenesis imperfecta.

Clinical dentinogenesis imperfecta occurs in only some patients; the appearance varies widely and affects some teeth more than others; the teeth are discoloured and the enamel (which is normal) fractures easily from the dentine, leading to rapid erosion of both the first and second dentition. Blueness of the sclerae is a particularly important physical sign of osteogenesis imperfecta. The cause of the frequently early (juvenile) arcus is unknown: limited investigation excludes hypercholesterolaemia. The cardiac manifestations of osteogenesis imperfecta are also important, not only because of their effects but because tissue fragility makes surgery dangerous. Aortic incompetence, aortic root widening, and mitral valve prolapse all occur. Patients with osteogenesis imperfecta often show hypermobility of joints, with resultant flat feet, hyperextensible large joints, and dislocation.

Type II osteogenesis imperfecta is nearly always lethal, but the severity does differ: some children may be born dismembered, whereas others may (rarely) survive the perinatal period to later merge into the type III form. Not all infants with multiple fractures at birth succumb immediately. It is possible to give a prognosis from the extent of ossification of the skull, the shape of the long bones and ribs, and the number of fractures. In the most frequent form of lethal osteogenesis imperfecta (IIA) the infant is short with disproportionately short and deformed limbs, the skull is deformed and soft, the sclerae are often deep grey-blue. Whole-body radiographs, which distinguish osteogenesis imperfecta from other forms of lethal short-limbed dwarfism, show grossly defective mineralization of the skull, short broad limbs with multiple fractures, and broad ribs with innumerable fractures (Fig. 14). In type IIB, the ribs have some structure; in IIC, the long bones are narrow and beaded at the site of fractures and show some evidence of modelling. Perinatal death results from the mechanical uselessness of the skeleton, which leads to respiratory failure or intracranial haemorrhage.

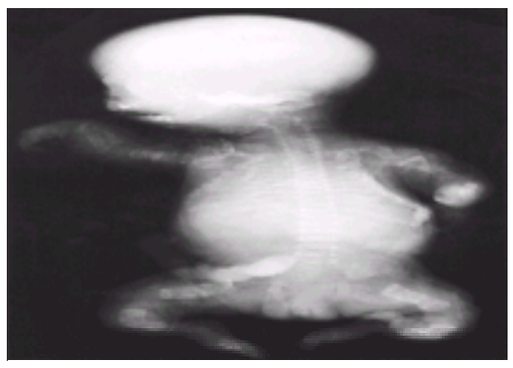

Fig. 14 Whole-body perinatal radiograph (babygram) of lethal (type II) osteogenesis imperfecta. The vault of the skull is not calcified, the ribs and long bones show multiple fractures. There was no family history.

Type III osteogenesis imperfecta causes most clinical difficulty, since the disability is severe and progressive. During the early years of life, progressive deformity affects the skull, the long bones, the spine, chest, and pelvis; the deformity is associated with fractures but can probably occur without them. The radiological appearance of the bones changes rapidly with age. The face appears triangular, with a large vault, prominent eyes, and small jaw. The sclerae may be blue in infancy but take on a normal colour in childhood. Eventual disability and deformity is considerable. Such patients rarely walk, even after multiple operations, and have a very short stature (–4 to –6 standard deviations (**SD**) below the mean). The changes in the long bones are often bizarre, with long, thin diaphyses and comparatively wide metaphyses. Cartilaginous islands often develop at the end of the long bones in the epiphyses and the metaphyses, spreading into the diaphysis, giving the appearance of 'popcorn' bone. Early death may occur from respiratory infections superimposed on the restrictive reduction in vital capacity associated with severe kyphoscoliosis (Fig. 15). Progressive deformity requires specialized orthopaedic care.

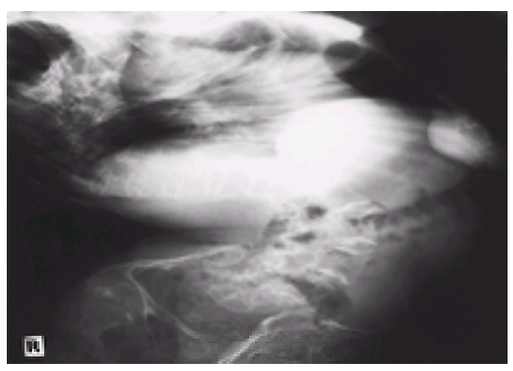

Fig. 15 Severe kyphoscoliosis in type III osteogenesis imperfecta.

Type IV osteogenesis imperfecta is clinically intermediate between type I and type III and is inherited as a dominant trait. The sclerae are of normal colour after infancy. Overall stature is reduced and disability is variable. The rare complication of hyperplastic callus occurs most often in this form (Fig. 16). This begins with a swollen, painful, and vascular swelling, most often over the long bones, an increase in plasma alkaline phosphatase, and sometimes a systemic illness. Recent investigations of osteogenesis imperfecta-affected families with hyperplastic callus have failed to find collagen mutations in affected children. Some classify this form as type V osteogenesis imperfecta.

Fig. 16 The appearance of hyperplastic callus in a patient with osteogenesis imperfecta.

Diagnosis

In the perinatal period, the concern is with alternative causes of lethal, short-limbed dwarfism. These include severe hypophosphatasia (see below), achondrogenesis (see below), thanatophoric dwarfism, and the asphyxiating thoracic dystrophies. A perinatal whole-body radiograph is essential.

In the first few years of life non-accidental injury, 'the battered baby syndrome', is the main differential diagnosis. This is suggested by multiple fractures at different sites and of different ages, especially if associated with clinical signs of neglect. Some fractures, such as metaphyseal 'corner' fractures and posterior rib fractures are more often seen in non-accidental injury, but any type of fracture can occur in osteogenesis imperfecta. The distinction between osteogenesis imperfecta and non-accidental injury is legally important and can be difficult.

Idiopathic juvenile osteoporosis needs to be distinguished during late childhood and adolescence. This begins during growth, with fractures of the long bones, reduction in growth rate (due to vertebral collapse), and metaphyseal compression fractures. In adult life, mild osteogenesis imperfecta may go unrecognized.

Biochemistry

It is impossible to generalize about the clinical effect of a collagen gene mutation but some patterns are emerging. In type I osteogenesis imperfecta there appears to be a null allele for collagen I, so that only 50 per cent of collagen is produced but this is of normal composition. Lethal osteogenesis imperfecta (type II) may result from large gene deletions but more commonly from single base changes in *COL1A1* or *COL1A2*. Such changes convert a glycine codon to one for another amino acid with a side chain. The effect on the triple helix of incorporating such a mutant chain appears to be most marked when the substitution occurs near the carboxy-terminal end of the chain (the helix winds up from this end), when the substituting amino acid is large; and when it occurs in the a-1 rather than the a-2 chain. Such mutations delay helix formation and render collagen mechanically unsound, and also lead to overhydroxylation and overmodification of the lysine residues, detectable by slowing and widening of the a-chains run on conventional polyacrylamide gels. Such abnormalities are common in type II osteogenesis imperfecta and less well defined in type III, which may rarely result from a failure to synthesize a-2 chains. Type IV osteogenesis imperfecta is most often due to changes in the a-2 chain.

Since the genes for a-1 and a-2 collagen have now been mapped and polymorphic sites identified, the mutant locus for osteogenesis imperfecta can be followed through large, dominantly inherited families. Such information can provide the basis for accurate prenatal diagnosis using fetal DNA derived from a chorionic villus biopsy. Methods are also now becoming available which make it possible to identify the mutation directly in the fetal DNA.

Genetic advice

Parents who have already had an infant with osteogenesis imperfecta need accurate advice about further pregnancies. This can be difficult, because the facts are not clear. Where the mutant gene is dominant (type I and IV) and where one parent is affected, the likelihood of affected children is 50 per cent. Difficulties arise where neither parent is clinically affected, and in the lethal and progressive deforming varieties of the brittle bone syndrome. It is impossible to give a statistically accurate prediction of the likelihood of another affected child, particularly since the strict application of Mendelian principles may be inappropriate because of germline and somatic-cell mosaicism. However, there are some guidelines. Where one offspring of clinically unaffected parents has a form of osteogenesis imperfecta which fits into type I or type IV, this is likely to be a new dominant mutation (50 per cent of whose offspring will be affected) and risk of a further affected sibling is probably no more than normal. It used to be considered that infants with the severe lethal form of osteogenesis imperfect (type II) had inherited a mutant gene from both clinically normal parents and were, therefore, homozygous recessives, so that the risk of a further affected infant was 25 per cent. The evidence is now that the great majority (if not all) result from a new dominant (and lethal) mutation. To allow for the possibility of some recessives, the likelihood of phenotypically normal parents having a second baby with lethal osteogenesis imperfecta is put at approximately 7 per cent (more than normal, but significantly less than 25 per cent).

The recurrence risk in progressively deforming osteogenesis imperfecta (type III) is unknown. If recessive inheritance is included in the definition, it is 25 per cent; if not, it is considerably less.

It is now recognized that germline and somatic-cell mosaicism are important factors in the inheritance and expression of osteogenesis imperfecta, and probably in many other disorders. In brief, germline mosaicism means that the sperm (or ova) of an apparently normal person may contain a proportion of mutant genes for lethal (or other forms) of osteogenesis imperfecta. This accounts for those pedigrees where a phenotypically normal man has two or more babies with lethal osteogenesis imperfecta by separate partners. Somatic mosaicism, with variable proportions of mutant cells in different tissues, likewise provides one (but not the only) explanation for phenotypic variability and differing tissue expression. The many factors that control the regulated expression of the vertebrate collagen genes in different tissues are only partly understood.

Prenatal diagnosis

This may be done from the second trimester by ultrasound and appropriate radiographs, and in the first trimester by analysis of fetal DNA from a chorionic villus biopsy. The appropriateness of such an investigation depends on the information previously available. In a dominantly inherited form of osteogenesis imperfecta, analysis of DNA from affected and unaffected family members can establish linkage to a particular collagen gene polymorphism. In such a situation, analysis of chorionic villus DNA is the most direct approach. Alternatively, the cells from such a biopsy may be cultured and the synthesized collagen examined for abnormalities. Where the collagen mutation is known in a previously affected family member, this method may directly confirm the presence of the mutation in the fetus. Except in well-organized laboratories, culture of cells and analysis of collagen will introduce unacceptable delays. Further, it is usually not possible to exclude an affected fetus merely on the grounds of apparently normal collagen. The rapid direct detection of the mutation in DNA from the chorionic villus is an eventual aim.

Amniocentesis also provides amniocytes for DNA linkage analysis and mutation detection. Amniotic fluid cells tend to produce an a-1(I) homotrimer and are not, therefore, appropriate for collagen analysis.

Diagnosis by ultrasound is possible only in the more severe forms of osteogenesis imperfecta (types II and III). Since the severe forms of osteogenesis imperfecta are sporadic and therefore unsuspected, it is important to be able to detect them early and rapidly by routine scanning. Ultrasound features suggestive of osteogenesis imperfecta are shortness and deformity of the limbs, an abnormal skull shape with lack of mineralization, which makes the intracranial structures abnormally visible, and deformity of the ribs leading to a 'champagne cork' appearance on the anteroposterior projection.

Prognosis and management

For an infant born with manifest osteogenesis imperfecta, important questions are asked: how long will he or she survive; what will be the likelihood of further offspring being affected? The immediate prognosis may already be answered by perinatal death, so that it remains to deal with the prognosis of survivors. Not all born with multiple fractures succumb immediately and radiographic appearances can give a good guide to outcome.

It is in those severely affected survivors classified as type III that management will be a lifelong and specialized problem. Such individuals are of normal intelligence and prolonged admission to hospital, either for repeated surgery or for investigation, should not necessarily take precedence over education. Intramedullary rodding and osteoclasis to correct deformity and improve mobility should be very selective since the bones are often so abnormal as to take no advantage from such procedures. An organized programme of rehabilitation is important. Analysis of life expectancy and cause of death in osteogenesis imperfecta show that survival is normal in type I osteogenesis imperfecta and near-normal in type IV. It is those with type III who have the most disability, of which basilar impression with neurological complications is a newly recognized problem, and the shortest lifespan. There is no convincing evidence that fluoride or calcitonin is beneficial, but cyclic intravenous pamidronate (APD) may alleviate symptoms and increase bone density. In severe osteogenesis imperfecta, attempts to transplant normal stromal cells from bone marrow into severely affected infants with osteogenesis imperfecta have been reported.

The Marfan syndrome (see Chapter 19.2)

The Marfan syndrome (Marfan's syndrome) is most often regarded as an inherited disorder of connective tissue rather than as a metabolic bone disease. Where connective tissue disorders significantly affect the skeleton, this distinction is blurred. For many years, it was thought that the defect underlying the Marfan syndrome involved collagen but recent research excludes this.

Pathophysiology

It is now recognized that Marfan's syndrome is caused by mutations in the epidermal growth factor-like regions of the fibrillin gene on chromosome 15. Fibrillin is the major constituent of the microfibrillar system and of the suspensory ligament of the lens; and it is also associated with elastin-containing tissues such as the aorta. This explains the association between dislocation of the lens and dissection of the aorta. The aorta dilates at its proximal part at the sinuses of Valsalva, and returns to normal diameter below the innominate artery, unless a dissection is present. The cusps of the aortic valve do not close efficiently. Dissection is most often above the aortic valves in the area of greatest dilatation. The dissection may progress forwards or backwards. Retrograde dissection may tear the attachment of the coronary arteries and rupture into the pericardial sac. Histopathology shows a reduction in elastic fibres which are swollen and fragmented. The valve cusps are usually diaphanous and redundant. In the eye, the suspensory ligament of the lens is disorganized.

Clinical features

Marfan's syndrome is dominantly inherited. Its main effects are on the skeleton, cardiovascular, and ocular systems. There is considerable phenotypic variation. In the typical patient with Marfan's syndrome, overall height is increased (relative to unaffected siblings or a matched population) and the limbs are long relative to the trunk (so that the crown to pubis measurement is less than pubis to heel). Long, thin fingers (arachnodactyly) are common. Together with hypermobility, this disproportion forms the basis of clinical signs of variable utility. However, not all patients with Marfan's syndrome are long and thin. The skeletal phenotype differs from one family to another and within families. Asymmetrical anterior chest deformity is associated with either depression or prominence of the sternum. Scoliosis is common, may be severe, and worsens during preadolescent growth as in the idiopathic form. The hard palate is often narrow and high-arched (gothic).

Dislocation of the lens is the main ocular feature of Marfan's syndrome. Typically, this occurs upwards or sideways (in contrast to the downward dislocation in homocystinuria), and this may be present at birth or occur later. Dislocation causes the unsupported iris to wobble on movement (iridodinesis). Less important ocular features are myopia and retinal detachment. The axial length of the globe is increased and the cornea tends to be flattened.

The most severe complication of Marfan's syndrome is dilatation of the ascending aorta leading to aortic incompetence and dissection. Progressive widening of the aorta can be readily measured by serial echocardiography. Less well-known manifestations of Marfan's syndrome include cutaneous striae, hernias, spontaneous pneumothorax, and dural ectasia. The mean life expectancy in those with Marfan's syndrome is reduced by nearly 50 per cent, predominantly due to cardiovascular catastrophe.

Diagnosis

At present, there is no certain biochemical way of excluding or confirming Marfan's syndrome, although this is likely to change. In those with few clinical features and no family history, the diagnosis of Marfan's syndrome can be difficult.

The requirements for the diagnosis of Marfan's syndrome have been revised. Where the family history is not helpful, it is necessary to have major criteria in at least two different organ systems and involvement of a third organ system. Where there is an unequivocally affected relative or a mutation known to cause Marfan's syndrome has been detected, one major criterion in an organ system with involvement of a second organ system is necessary for the diagnosis. Homocystinuria (see below), which has a recessive mode of inheritance, should be excluded. Other important alternative diagnoses include congenital contractural arachnodactyly, familial tall stature, isolated mitral valve prolapse, familial or isolated annuloaortic ectasia, and Stickler's syndrome. The latter is a dominantly inherited connective tissue disorder that affects the eyes, ears, and skeleton with severe myopia in childhood, sensorineural hearing loss from adolescence, and degenerative arthritis from early adult life. The diagnosis can be made at birth if cleft palate and micrognathia are present. There is considerable phenotypic variation. In some families the disorder is linked to the type II collagen gene.

Contractures can occur in Marfan's syndrome but are of a late onset. In congenital contractural arachnodactyly, which is inherited as an autosomal dominant trait, contractures involving the hands, feet, and larger joints are present from birth and tend to improve. Abnormal ears are described. Limited studies suggest that this disorder involves mutations of an additional fibrillin gene on chromosome 5.

Treatment

There is no specific treatment for the underlying defect, but many of the clinical manifestations require attention. Scoliosis may be progressive and severe, particularly in adolescence. Bracing is largely ineffective and operative stabilization may be necessary. Excessive height in girls may be prevented by giving oestrogen together with progestogen in the prepubertal years. Marked sternal deformity may need correction for cosmetic or cardiopulmonary reasons, but opinions on the value of surgery vary widely. In the eyes, it is rarely necessary to remove dislocated lenses unless they prolapse into the anterior chamber, but myopia should be corrected. The main decisions concern the management of the cardiovascular problems: when and if to operate on the dilated ascending aorta or to replace incompetent valves; and whether aortic dilatation can be prevented by reducing the intermittent force on its walls due to left ventricular systole. As far as the second point is concerned, giving a b-blocker such as propranolol probably reduces the rate of aortic dilatation. As regards surgery on the aorta, it is clear that progressive aortic widening (measured regularly by echocardiography), together with progressive aortic incompetence and left ventricular strain, provide strong indications for replacement of the proximal aorta by a prosthesis. Mitral valve replacement may also be necessary.

Since both aortic and mitral valves are susceptible to endocarditis, prophylactic antimicrobials must be given at the time of dentistry.

Genetic advice

Genetic advice is at present based on clinical observations and the knowledge that inheritance is of the autosomal dominant pattern. Numerous mutations in the fibrillin genes have now been described. There is no clear relationship between genotype and phenotype.

Ehlers–Danlos syndrome (see Chapter 19.2)

This syndrome initially included only those conditions with the common clinical features of abnormal velvety hyperelastic skin which healed poorly, hyperextensible joints, and lax ligaments. However, the disorders included in this syndrome have now been increased and have brought with them additional specific features, amongst which is vascular rupture, especially in type IV Ehlers–Danlos syndrome, associated with various mutations in type III collagen. In the currently expanded Ehlers–Danlos syndrome the skeleton is particularly affected in types VI and VII (Table 11).

In type VI (ocular scoliotic) Ehlers–Danlos syndrome) the first disorder in which an inborn error of collagen metabolism was identified, the clinical features are due to lysyl hydroxylase deficiency. Since hydroxylation of peptide-bound lysine is an essential post-translational step in collagen synthesis and a necessary precursor to crosslink formation, this defect weakens collagen structure. The main clinical features are severe scoliosis, microcornea, and ocular fragility.

In type VII (arthrochalasia) there is excessive mobility, perinatal joint dislocations (especially of the hips), and short stature. There is persistence in the tissues of collagen molecules with a retained amino-terminal propeptide which leads to defective fibrillogenesis.

Homocystinuria (see also Chapter 11.3)

Homocystinuria is phenotypically similar to Marfan's syndrome but with a different cause and additional important complications. It is due to a deficiency of cystathionine b-synthase, an enzyme whose gene is located on chromosome 21, and firmly bound pyridoxal phosphate (vitamin B_6) is a feature. Homocystinuria is inherited as an autosomal recessive condition. The amount of residual cystathionine synthase varies from 0 to 10 per cent in patients, and in obligate heterozygotes it is less than 50 per cent of normal.

Pathophysiology

Homocysteine lies at the cross-roads of two metabolic pathways and is converted to cystathionine by the addition of serine. This reaction is controlled by

cystathionine b-synthase. The alternative fate of homocysteine is methylation to methionine. Cystathionine b-synthase activity is controlled by pyridoxine, but not all patients with cystathionine-deficient homocystinuria are pyridoxine-sensitive, although this sensitivity or dependency is constant in sibships. In homocystinuria, there is an increase in both homocysteine and homocystine, which accumulate proximal to the metabolic block. Cystathionine, normally present in the brain, is no longer detectable and cysteine (normally made from methionine) becomes an essential amino acid.

The pathological findings include fraying and disruption of the zonular fibres of the lens, defective bone formation, and multiple central nervous system infarcts. It is not known how the biochemical changes lead to the clinical features. The increased thrombotic tendency is not fully explained by changes in platelet function, cellular endothelium, or soluble factors, although abnormalities have been described in all of them. The neurological abnormalities and mental backwardness have not been proven to be due to the biochemical consequences of cystathionine b-synthase deficiency or to repeated vascular thromboses. Homocyst(e)ine may increase the solubility of collagen and interfere with its synthesis; for some, this explains the dislocation of the lens due to failure of the ciliary zonule. Since it is now known that this structure is composed largely of fibrillin, a further explanation is required. There is current interest in the possibility that young adults with premature vascular disease may be heterozygotes for a mutant cystathionine synthase gene. Elevated plasma homocysteine levels are a risk factor for coronary heart disease.

Clinical features

The clinical features of cystathionine b-synthase deficiency involve four systems and develop some time after birth; they are ocular, skeletal, central nervous, and vascular. The main ocular manifestation is downward dislocation of the lens. Myopia, glaucoma, retinal degeneration, and detachment also occur, and cataracts, optic atrophy, and corneal abnormalities are described. Some skeletal features suggest Marfan's syndrome. They include a long, thin habitus, pectus excavatum, scoliosis, and genu valgum. There is often radiological osteoporosis and abnormal modelling of the long bones with epimetaphyseal widening. Many subjects with homocystinuria are mentally backward and may also have seizures and strokes. It is unknown how closely these follow the increased tendency to thrombosis or the biochemical changes, especially a lack of cystathionine. Thromboembolism may occur in any vessel and at any age.

Any patient who has the phenotypic features of Marfan's syndrome associated with thrombosis, mental simplicity, and affected siblings should have a cyanide–nitroprusside test performed on their urine, together with an amino acid analysis of the urine and plasma.

The outlook for patients whose biochemical abnormalities are corrected by large amounts of pyridoxine (that is, those with pyridoxine-sensitive homocystinuria) is usually better than those who are pyridoxine-resistant. The main cause of death is thromboembolism.

The management of patients with homocystinuria differs according to the time of diagnosis and whether or not the patient responds to pyridoxine. In pyridoxine-responsive patients diagnosed after the newborn period, giving pyridoxine, in doses that vary from 250 to 1200 mg a day, appears to prevent thromboembolism.

When homocystinuria is detected in the newborn infant (most are discovered by screening and are pyridoxine non-responsive), a diet low in methionine appears to reduce the incidence of low intelligence. After the newborn period, in those who are unresponsive to pyridoxine, methionine restriction and the administration of betaine (as a methyl donor) are also possibly useful lines of approach.

Alkaptonuria (see also Chapter 11.2)

In this rare autosomal recessive disorder, decreased activity of homogentisate oxidase leads to accumulation of homogentistic acid in the urine and increased pigmentation (ochronosis) in cartilage and connective tissues. Darkening of the urine, alkaptonuria, is due to the presence of 2,5-dehydroxyphenylacetic acid derived from the oxidation and polymerization of homogentisic acid. Polymerization increases in alkaline urine and is slowed down by antioxidants such as vitamin C. The structure of the pigment which causes ochronosis is not known. It is granular or homogeneous and may occur within or outside the cell. It is said to be associated with a reduction in lysyl hydroxylase in the tissue concerned, and impairment of the crosslinks of collagen.

Alkaptonuria is more frequent in the former Czechoslovakia and in Germany than elsewhere and occurs equally in the sexes. It is recessively inherited. The mutant gene has now been identified (*HGO*, chromosome 3q2). Abnormal pigmentation is found in the cartilage of the ear (which may be calcified), the nasal cartilage, and the sclerae. The most important effects of this disease are on the spine (Fig. 17) and later on the larger joints. The intervertebral discs lose height and later calcify; they may also herniate acutely. The spine becomes rigid and short and the lumbar lordosis is lost. In the large joints, such as the knees, shoulders, and hips, there are effusions and loose bodies. The symphysis pubis may be affected but not the sacroiliac joints. Ochronotic 'arthritis' is described with episodes of acute inflammation which resemble those of rheumatoid arthritis. Calcification of the aorta is an additional feature.

Fig. 17 The appearance of the spine in a man with alkaptonuria. There is universal calcification of the intervertebral discs.

The diagnosis of alkaptonuria—often made late—should be suspected where there is a premature disc degeneration, even if there is no excessive darkening of the urine. Early degenerative arthritis suggests the disease, confirmed by finding deeply pigmented articular cartilage at the time of operation. In those patients with a lifelong discoloured urine, the differential diagnosis is from other rare causes of urinary pigmentation. The urine of a patient with alkaptonuria contains reducing substances and will therefore give a positive result suggesting glycosuria except where glucose oxidase is used. An increase in homogentisic acid in the urine and plasma confirms the diagnosis.

In theory, it should be possible to reduce the amount of homogentisic acid, and presumably the side-effects, by cutting down the protein intake to 30 or 40 g/day, thereby reducing tyrosine intake. There is no evidence that such a procedure alleviates the symptoms of alkaptonuria.

Hypophosphatasia (see also Chapter 12.6.2)

This rare disorder has similarities with rickets and osteomalacia. It is due to a reduction in the tissue non-specific alkaline phosphatase (**TNSAP**), which leads to defective mineralization and a triad of biochemical disturbances: increased urinary phosphoethanolamine, plasma pyrophosphate, and plasma pyridoxal phosphate.

Studies on members of the Mennonite sect in Manitoba, in whom the incidence of hypophosphatasia is high, have linked the defective gene to chromosome 1. Numerous mutations have now been described in the *TNSAP* gene. Although TNASP is widely distributed, its absence leads to lesions only in the bone and teeth.

Pathophysiology

The characteristic biochemical changes result directly from the alkaline phosphatase deficiency. Increased urinary pyrophosphate excretion is more reliable than urinary phosphoethanolamine as a marker for carriers of the hypophosphatasia gene. Often, there is also hypercalcaemia and hypercalciuria in childhood; and up to half of affected children and adults have increased plasma phosphate levels. Hyperphosphataemia is described in carriers of the hypophosphatasia gene. The

recorded plasma alkaline phosphatase level must be compared with age-matched control values.

Histological examination of bone shows an excess of osteoid with abnormal tetracycline labelling without evidence of secondary hyperparathyroidism. Matrix vesicles do not contain alkaline phosphatase or hydroxyapatite crystals. The primary dental defect is in the cementum; additionally, the predentine is widened and the dentinal tubules are enlarged and few.

Clinical features

Hypophosphatasia occurs in all races. Since the severe forms are inherited as autosomal recessive traits they are more frequent where there is consanguinity. It has been estimated that hypophosphatasia occurs in 1 in 100 000 live births in Toronto. The four clinical types provide a continuous spectrum, from a lethal perinatal disorder to an asymptomatic disease in adults.

The first is an important cause of lethal, short-limbed dwarfism (see above). Some newborn infants survive for a few days, but fever, failure to thrive, anaemia, seizures, and intracranial haemorrhages occur. Radiographs show grossly defective mineralization, especially in the skull, where only the base may be mineralized, and in diaphyses of the long bones which, rarely, may have bony spurs.

In the infantile form (within the first 6 months), hypotonia, failure to thrive, hypercalcaemia, and hypercalciuria occur. Clinical rickets is noticed and the fontanelle appears wide, but there is a functional synostosis. Craniostenosis can produce optic atrophy, exophthalmos, and raised intracranial pressure requiring surgery.

The most variable expression occurs in childhood. Early loss of deciduous teeth, due to defective cementum, may be the only feature (ondontohypophosphatasia). The pulp chambers are enlarged, the root canals short (shell teeth). If bone disease is present, walking is delayed and deformities occur; for instance, bow legs, knock knees, short stature, and enlargement of the epiphyses at the wrist, knees, and ankles.

In adults, progressive stiffness, pain in the bones, and apparent 'stress' fractures can occur. Approximately 50 per cent of such patients have a childhood history of bone disease resembling rickets, or premature loss of deciduous teeth, or both. There may also be premature shedding of adult teeth, short stature, and abnormal skull shape. Recurrent poorly healing metatarsal fractures occur. Partial fractures of the long bones characteristically occur on the convex outer surface (in contrast to the concave inner position of the Looser's zones in osteomalacia), most often in the upper one-third of the femoral shaft, and are often bilateral; other sites include the ribs, tibias, and ulnas. They may be unaltered for years or increase in size and eventually fracture. Secondary hyperparathyroidism is not seen. Chondrocalcinosis is common and in a proportion is associated with clinical pyrophosphate gout (pseudogout).

Management

In the management of hypophosphatasia, premature synostosis leading to raised intracranial pressure requires surgical relief. Hypercalcaemia may be dealt with by reducing dietary calcium and by giving prednisone. Replacing the defective enzyme by the transfusion of alkaline phosphatase-rich plasma does not produce consistent results. Intramedullary rods may prevent and treat fractures of the long bones. Dental abnormalities, which can occur in biochemically normal members of hypophosphatasia families, may require treatment.

Prenatal diagnosis of a severely affected child can be made by ultrasound. There is also reduced alkaline phosphatase activity in the amniotic fluid cells.

Lysosomal storage diseases (see alsoChapter 11.8)

This large group of diseases is due to various inborn errors that affect the function of specific lysosomal enzymes normally responsible for the breakdown of a variety of complex molecules. As a result, these molecules, or their partially degraded derivatives, accumulate in the lysosomes and the tissues that contain them. The effect of this accumulation varies from one tissue to another according to the particular disorder, and the skeleton is significantly involved in only a proportion of them. They include some mucopolysaccharidoses and Gaucher's disease.

Mucopolysaccharidoses

Failure of the normal lysosomal breakdown of complex carbohydrates leads to their accumulation in the tissues, and produces many clinical abnormalities. The disorders may be divided into two main groups according to the chemistry of the accumulated substance, namely the mucopolysaccharidoses and the mucolipidoses. Specific biochemical defects are described elsewhere in this book (see Section 11). Since some of these disorders have a prominent effect on the skeleton, they are briefly mentioned here; they are the Hurler syndrome (mucopolysaccharidosis type IH; **MPS IH**), the Hunter syndrome (MPS II), and the Morquio syndrome (MPS IV). With certain exceptions the bone changes themselves do not permit precise diagnosis of the type of dysplasia present, or distinction from the mucolipidoses.

The Hurler syndrome (MPS IH)

This is the most severe type of mucopolysaccharidosis and causes death at an early age. The enzyme defect is recessively inherited and all patients have the same appearance, to which the term 'gargoylism' was previously applied. Affected infants appear to develop normally in the first few months of life, but then deteriorate mentally and physically. Death often occurs in late childhood, commonly due to pneumonia or to coronary artery disease associated with mucopolysaccharide deposits.

The physical features include proportionate short stature (Table 3), a typical facial appearance, a short neck with a lumbar gibbus and chest deformity, and a protuberant abdomen. The facial features are coarse and ugly, with flattening of the nasal bridge, with large open mouth and tongue, and often with hypertrophied gums over enlarged alveolar ridges. The eyes are prominent with corneal clouding. There is noisy breathing and variable deafness. The vault of the skull may show scaphocephaly or acrocephaly. Other striking features include the stiff, broad trident hands and the large abdomen with hepatosplenomegaly. Radiographs show the abnormal shape of the skull, the slipper-shaped sella turcica, the beaking of the vertebrae with the thoracolumbar kyphosis, and the bullet-shaped phalanges. Similar but less severe features are seen in the Hunter syndrome, which is inherited as an X-linked recessive.

The Morquio syndrome (MPS IV)

In this disorder the orthopaedic manifestations are striking, but intelligence is normal. Although the disorder is probably heterogeneous and only a proportion of cases excrete an excess of keratan sulphate in the urine, the skeletal changes are uniform. In the first years of life the child becomes progressively more deformed and dwarfed. Characteristically the neck is short, the sternum is protuberant, and there may be a flexed stance with knock knees. There is a striking loss of muscle tone in comparison to the stiffness of MPS IH; hypermobility and a loose skin are features. Radiographs in infancy show a spine similar to that seen in those with Hurler syndrome, but later flattening of the vertebrae with anterior beaking lead to relative shortening of the trunk. The small bones of the hands are very different from those of MPS IH and the metacarpals show diaphyseal constriction (Fig. 18).

Fig. 18 The appearance of the hands in MPS IV (Morquio syndrome).

Importantly, the odontoid may be hypoplastic, leading to atlantoaxial instability, compression of the long spinal tracts, and paraplegia.

Gaucher's disease (see also Chapter 11.8)

This is a rare lysosomal storage disorder in which glucocerebroside-containing macrophages accumulate within the bone marrow, spleen, liver, and other organs. It is recessively inherited and over-represented in Ashkenazi Jews, where the incidence of the adult form (type I) is about 1 in 2500 births. The skeletal manifestations are often severe and disabling. They vary from a characteristic but clinically insignificant failure of remodelling in the lower femora (Erlenmeyer-flask appearance) to diffuse and localized bone loss and osteosclerotic and osteonecrotic lesions, which cause pain and pathological fracture, often requiring precocious joint replacement surgery.

Skeletal dysplasias

The term 'skeletal dysplasia' has traditionally been used to cover a wide range of generalized disorders of the skeleton, often of unknown cause, affecting both cartilage and bone. With increasing knowledge one can distinguish the chondrodysplasias, which are primarily due to mutations affecting cartilage, from such disorders as diaphyseal dysplasia and assorted dense bone diseases, where the causes are less well known. Since osteopetrosis is a well-defined disorder of osteoclast function, it is dealt with separately below.

The mutations in many of the skeletal dysplasias have been described (Table 1) and the skeletal dysplasias can be classified into biochemical families according to their cause (Table 12). The supposition that many of them could be due to mutations in specific collagens has been partially confirmed with mutations found in type I, IX, X, and XI collagens. Achondroplasia is a striking example of a skeletal dysplasia caused by a non-collagen mutation, that is a mutation in fibroblast growth factor (**FGF**)-receptor 3. Further details can be found in reviews (see the Further reading list) and in Table 12.

Clinical features

The physician confronted by a patient with a skeletal dysplasia is unlikely to make the correct diagnosis without much additional help unless it is clearly one of the most frequent, for instance achondroplasia. However, accurate classification of the dysplasias is important and will make clinical and biochemical advance possible. The most convenient simple classification is a clinical one (Table 13). Most patients with skeletal dysplasias have restricted growth, and most are short-limbed. The bodily proportions of people with skeletal dysplasias will provide a clue about whether the limbs are mainly affected, or the spine, or both. In the short-limbed group, achondroplasia and achondroplasia-like dwarfs are the most typical. Those disorders without conspicuous dwarfing include various inherited epiphyseal dysplasias, diaphyseal dysplasias, and some, but not all, metaphyseal dysplasias. An alternative classification, not based on height, groups the dysplasias according to whether they are predominantly epiphyseal or metaphyseal, whether the spine is predominantly involved, and whether single limbs or segments are involved. Radiographs, taken as soon as possible and, where possible, consecutively, are essential to determine whether the metaphyses of the long bones or the epiphyses are primarily affected.

For the purpose of this Section, osteopetrosis (marble bones disease) is dealt with separately as a disorder of bone-cell biology. Other sclerosing disorders of bone, in some of which biochemical abnormalities have been described (Engelmann's disease, van Buchem's disease), receive brief mention.

Achondroplasia

This is the prototype of short-limbed, short stature. It is inherited as an autosomal dominant, with a high mutation rate and the incidence increases with paternal age. It is due to a specific mutation in the gene encoding the FGF-receptor 3. The way in which this produces the skeletal changes is largely unknown. Until recently, any undiagnosed patient with excessively short limbs was given the label of achondroplasia. This explains the apparent high frequency of achondroplasia and its high mortality, since different forms of lethal, short-limbed dwarfism were then included.

As the clinical definition of achondroplasia has not always been exact, its true incidence and natural history are not well defined. There is a failure of the epiphyseal growth cartilage, and bulbous masses of cartilage appear at the ends of the long bones. In contrast, periosteal and membrane bone formation and bone repair are normal. This selective effect on growth cartilage accounts for the skeletal deformity.

Achondroplasia can be diagnosed at birth or within the first year of life, when the disparity between the large skull and short limbs becomes obvious. There is a striking disproportion between the normal length trunk and the short arms and legs. Thus the fingertips may only come down to the iliac crest. The shortness of the limbs particularly affects the proximal segment. The limbs themselves look very broad, with abnormally deep creases, and the hands are trident-like. In contrast to the short limbs is the enlarged bulging vault of the skull, the small face, and flat nasal bridge or 'scooped out' glabella. There is a marked lumbar lordosis and also sometimes some wedging of the upper lumbar vertebrae, which may later lead to a thoracolumbar kyphosis. Radiological features include metaphyseal irregularity and flaring in the long bones, irregular and late-appearing epiphyses, a narrow pelvis in its anteroposterior diameter, with short iliac wings and deep sacroiliac notches, and a spine that shows progressive narrowing of the interpedicular distance from above downwards, which is the reverse of normal.

Children with achondroplasia are of normal intelligence, and the complications of this disease arise particularly from the skeletal disproportion. This may lead to early osteoarthritis, to obstetric difficulties and the need for caesarean section, to hydrocephalus, and to paraplegia. Eventual height can vary between about 80 and 150 cm. Recent reviews emphasize how often narrowing of the spinal canal produces symptoms of spinal stenosis.

Homozygous achondroplasia (the offspring of two affected parents) is severe and lethal. In the condition of hypochondroplasia, which is included in the same *FGFR3* molecular family, the skeletal disproportion and the spinal abnormalities are less and the skull is unaffected.

Achondroplasia-like dwarfism

For details of these and other causes of short-limbed dwarfism the reader should consult more specialized texts). Those that most closely resemble achondroplasia at birth are thanatophoric dwarfism, achondrogenesis, severe hypophosphatasia, and type II osteogenesis imperfecta. All can be distinguished radiologically.

Spondyloepiphyseal dysplasias

This is a heterogeneous group of disorders in which the spine is predominantly affected and the short stature is partly due to shortness of the trunk. The most severe type is spondyloepiphyseal dysplasia (**SED**) congenita; milder forms are referred to as SED tarda. There are various forms of inheritance. Some forms are due to mutations in type II collagen.

SED tarda often has an X-linked mode of inheritance, so that only males are affected and females are carriers. In affected males the disproportionately short trunk becomes obvious at adolescence. Failure of ossification in the anterior part of the so-called ring epiphyses leads to central and posterior humps on the upper and lower parts of the flattened bodies. The condition needs to be distinguished from multiple epiphyseal dysplasia, which involves other major joints more than the spine.

SED congenita can be diagnosed at birth because of the short stature associated with a short trunk. There may be a close resemblance to Morquio's disease (MPS IV, see above). The severe form may be distinguished from the age of about 4 years. The appearance of the capital femoral epiphyses is delayed (in some patients it may never be seen, except by arthrography). Marked lumbar lordosis, waddling gait, back pain, and progressive disproportion may occur. The odontoid is hypoplastic, kyphoscoliosis may develop, and the interpedicular distances of the vertebrae do not increase in the lumbar region. Paraplegia may occur as a result of all these changes. In this disorder there is often myopia and retinal detachment.

There is a form of SED, pseudoachondroplasia, which resembles achondroplasia because of the short limbs, but here the facial appearances are normal. The short stature becomes obvious from about 2 years of age. Lumbar lordosis and scoliosis may develop. The tubular bones are short with irregular metaphyses and small, deformed epiphyses. Hypermobility is marked and early osteoarthritis occurs. The causal mutation is in the gene for the cartilage oligomeric protein (*COMP*, chromosome 19p16).

Proportionate dwarfism

Although it is clinically important to classify short stature into proportionate and disproportionate, there are many conditions in which this distinction is difficult to make. Hypophosphataemic rickets, mucopoly-saccharidoses, vitamin D-dependent rickets, and osteogenesis imperfecta may come into both categories.

Bone dysplasias without conspicuous short stature

The height of patients with multiple epiphyseal dysplasia may be only slightly reduced. Although many epiphyses are affected, the spine is virtually normal. There are also variable forms of inheritance. Some are due to mutations in collagen type IX; others to mutations in cartilage oligomeric protein.

In patients with multiple hereditary exostoses (often referred to as diaphyseal aclasis) there is a juxtaepiphyseal disorder of bone growth, limited to bones developed in cartilage, which gives rise to cartilage-capped exostoses that point away from the joint. Inheritance is autosomal dominant and stature is normal. It is likely that there are causal mutations in putative tumour suppressor genes.

The metaphyseal disorders are rare; some, such as the Jansen type of metaphyseal dysostosis (associated with a mutation in the gene for the PTH/PTHrP receptor) do cause severe dwarfing. In others with less severe growth disturbance, such as Type Schmid (due to a mutation in the type X collagen gene), rickets is simulated, and confusion with inherited hypophosphataemia is possible. In progressive diaphyseal dysplasia (see below) the limbs are disproportionately long.

Sclerosing disorders of bone

Apart from marble bones disease (see below) the experience of most physicians of the osteoscleroses is limited by their extreme rarity.

Engelmann's disease (progressive diaphyseal dysplasia: Camurati–Engelmann disease)

This rare condition is autosomal dominantly inherited. It affects endocrine and muscular systems in addition to the skeleton, where the main feature is a variable but progressive endosteal and periosteal thickening of the diaphyses of the long bones. In severely affected subjects the spine, skull, and axial skeleton are all affected. The cause is unknown.

There is a waddling broad-based gait, muscle wasting and weakness, loss of subcutaneous tissues, and pain in the legs during childhood. The appearance is characteristic; the head is large with a prominent forehead and proptosis, the muscle mass is reduced, and the bones are palpably thickened. Cranial nerve palsies, deafness, and blindness with raised intracranial pressure can occur. Puberty is delayed. Bone pain resistant to analgesia is often a presenting and troublesome feature.

Radiographic appearances vary, from limited thickening of the diaphyses (often in the lower extremities) to widespread new bone formation, affecting all bones, including the skull, demonstrated by scintigraphy.

The increased bone turnover causes a moderate increase in plasma alkaline phosphatase and urinary hydroxyproline levels. There may be a markedly positive calcium balance, associated with hypocalcaemia and hypocalciuria. Hyperphosphataemia has been recorded.

Pathological examination confirms gross thickening of the bone with disorganization of internal structure and external shape. The peripheral subperiosteal new bone is woven. The muscles show non-specific, type-II fibre atrophy.

In the differential diagnosis the proximal myopathy and abnormal gait simulate muscular dystrophy. The radiographic appearances are diagnostic, although idiopathic hyperphosphatasia may present some difficulties.

The course of this disorder is unpredictable and remission of symptoms may occur during adolescence or adult life, so it is difficult to assess treatment. Bone pain may respond to corticosteroids in small, alternate-day doses. Etidronate (20 mg/kg daily) has produced hypocalcaemic tetany, but intermittent administration is reported to reduce pain. Limb pain may be relieved by surgical removal of a cortical window in the diaphysis.

Pyknodystosis

In contrast to Engelmann's disease, pyknodystosis has an autosomal recessive mode of inheritance, with parental consanguinity in some 30 per cent of subjects. It has some similarities to osteopetrosis. Since the disease is caused by mutations that lead to deficiency of cathepsin K, an enzyme necessary for osteoclast function, this is not unexpected. Marked reduction in stature with short limbs is a particular clinical feature.

The vault of the skull is large, the face and chin small, the palate high-arched, and the teeth crowded, with retained deciduous teeth. The anterior fontanelle (and other cranial sutures) remain unfused. The painter Toulouse-Lautrec is regarded as a typical example of this disease. The fingers may appear to be clubbed because of associated acro-osteolysis. The chest is deformed with kyphoscoliosis and pectus excavatum. Recurrent fractures of long bones occur, and occasionally rickets. Radiologically, there are similarities to osteopetrosis with generalized osteosclerosis and fractures. However, the osteosclerosis is uniform; there are no defects of modelling and no endobones. In addition to delayed closure of the cranial sutures there are also wormian bones; the bony fragility, wormian bones, and blue sclerae simulate osteogenesis imperfecta.

Idiopathic hyperphosphatasia

This very rare condition is also labelled 'juvenile Paget's disease'. It has autosomal recessive inheritance. The long bones are abnormal, thickened, and bowed from the first year of life, and the skull may be enlarged. Muscular weakness is common and the plasma alkaline phosphatase level is continuously very high.

Sclerosteosis

This condition is due to an autosomal recessive trait. There is progressive overgrowth and sclerosis of the skeleton, including the skull and the mandible. There are similarities to van Buchem's disease (endosteal hyperostosis), but the skeletal problems are more severe and there is often syndactyly. Prophylactic craniectomy may be necessary to reduce the increased intracranial pressure.

van Buchem's disease

In this rare hyperostosis, endosteal thickening of the shafts of the long bones is associated with generalized hyperostosis, including the base of the skull and the mandible. Bilateral facial nerve weakness, deafness, and optic atrophy may ensue. Severe recessive and mild dominant forms are described.

Cleidocranial dysplasia

In this rare condition the clavicles are hypoplastic or absent, the fontanelles remain open, and there are supernumerary teeth. The heterozygous mutation causes a loss of CBFA1, the osteoblast transcription factor (see above).

Osteopetrosis (marble bones disease)

Among those disorders with increased bone density, marble bones disease or osteopetrosis (Albers–Schönberg disease) is the best known. It is a heterogeneous disorder with a widespread increase in bone density. In most cases, the basic defect lies in the osteoclasts which, for various reasons, are unable to resorb mineralized bone. Many animal models of osteopetrosis exist.

Until recently two main forms were distinguished: recessively inherited severe osteopetrosis causing death in childhood; and the dominantly inherited mild form, in which the diagnosis can be made on radiological grounds alone. This distinction is not absolute—two distinct dominantly inherited forms exist, as well as intermediate forms. Deficiency of carbonic anhydrase II can also cause osteopetrosis associated with cerebral calcification, renal tubular acidosis, growth failure, and mental simplicity.

Severe osteopetrosis

In severe recessively inherited osteopetrosis there is widespread increased density of the bones without modelling or remodelling. This produces the Erlenmeyer-flask deformity of the metaphyses. The increase in bone density is often intermittent, producing alternating bands of sclerosis. The failure of resorption leads to a reduction in bone marrow space with a leucoerythroblastic anaemia and hepatosplenomegaly. It also produces nerve compression, with blindness and often deafness. Other clinical features in this severe form can include hydrocephalus, delayed tooth eruption, and osteomyelitis. Fracture of the dense bones is common. The affected infant is short with an apparently large head with frontal bossing, hepatosplenomegaly, and knock knees. The plasma calcium level appears to alter with the dietary intake and may be sufficiently low to contribute to rickets. The acid phosphatase concentration (derived from the defective osteoclasts) is increased. Secondary hyperparathyroidism leads to an increase in calcitriol levels. Apart from transplantation of bone marrow, as a source of normal osteoclasts, from an appropriate donor, other forms of medical treatment deal only with complications; these include surgery for fractures, blood transfusions for anaemia, and antibiotics for frequent infections.

Mild osteopetrosis

The mild forms vary from subjects with an increased number of fractures affecting both the long bones and the small bones of the hands and feet, to those in which the disorder is so mild that the diagnosis is made by radiology alone (accounting for apparently unaffected generations with the dominant form of the disease). There are more severe forms of dominantly inherited osteopetrosis with nerve compression, deafness and blindness, and anaemia at times of increased physiological requirement, such as pregnancy. Other established features include osteomyelitis and facial nerve palsy.

Recent studies of Danish families define two dominantly inherited forms: one with uniformly dense bones with sclerosis of the cranial vault and the spine and no increase in the plasma acid phosphatase level, and another with variable bone density (giving rise to an endobone appearance, Fig. 19) and lack of modelling, with a significant increase in the plasma acid phosphatase level.

Fig. 19 The appearance of the bones in a boy with dominantly inherited osteopetrosis and a raised plasma acid phosphatase level. There are variations in bone density ('endobones') with recent and old pathological fractures.

Carbonic anhydrase II deficiency

The association of carbonic anhydrase II deficiency with osteopetrosis, renal tubular acidosis, cerebral calcification, some degree of mental retardation, growth failure, and dental malocclusion is of considerable interest because of the clues it provides to the normal function of carbonic anhydrase II in bone resorption. Carbonic anhydrase II is part of the carbonic anhydrase gene family and is widely distributed. It is found in the kidney, brain, red cells, and elsewhere, and its gene is on chromosome 22. Deficiency of carbonic anhydrase II is autosomal recessively inherited, and apparently normal parents of affected offspring have 50 per cent of normal carbonic anhydrase II levels within their red cells. The bone disease is not distinguishable from other forms of osteopetrosis, and fractures occur until adulthood. There is always growth retardation, and height may be more than four standard deviations below the mean. The bone age is also delayed. Radiographic appearances improve in adult life.

The renal tubular acidosis is mixed, both proximal and distal. Cerebral calcification affects the basal ganglia within the first decade. It increases during childhood to include the cortical grey matter and is similar to that occurring in idiopathic or pseudohypoparathyroidism. Bone histology shows unresorbed calcified cartilage and osteoclasts without a ruffled border.

The diagnosis of carbonic anhydrase II deficiency should be considered in any neonate with renal tubular acidosis. Genetic counselling is possible since adult heterozygotes have reduced levels of the enzyme in their red cells. However, the concentration of carbonic anhydrase II is normally very low at birth and cannot be used as a reliable neonatal test for the affected homozygote.

The treatment of carbonic anhydrase II deficiency is symptomatic; it is possible that correction of the renal tubular acidosis temporarily increases the rate of growth.

In the differential diagnosis of osteopetrosis there are many disorders with an excessive amount of bone in various parts of the skeleton; these include other skeletal dysplasias, Caffey's disease (infantile cortical hyperostosis) which causes a temporary increase in bone density from birth, and myelofibrosis, renal glomerular osteodystrophy, inherited hypophosphataemia, and fluorosis in adult life.

Fibrous dysplasia

Fibrous dysplasia of bone is a condition in which areas of immature fibrous tissue, either single or multiple, are found within the skeleton. Recent research has shown a widespread postzygotic activating mutation in the gene for a subunit of the G-protein signalling system. The extent to which this activating mutation affects the bone and other tissues depends on the degree of mosaicism, in other words the proportion and distribution of cells that carry the mutation. It is proposed that such a mutation in the germline would be lethal; certainly the condition is not inherited.

Monostotic fibrous dysplasia

This disorder is relatively common in orthopaedic practice. Although the lesions may occur in any bones, and particularly in the facial bones and ribs, the most frequent presenting symptom at any age is a fracture, often of the upper end of the femur (Fig. 20). The biochemistry is usually normal, and the diagnosis is made from the radiographic and pathological appearances. There is a smooth-walled translucent area within the bone, often with thinning of the cortex and sometimes with associated deformity. Pathologically, areas of disorganized fibrous tissue are are found, associated with woven bone and wide osteoid seams. This represents mosaic

tissue with some normal mesenchymal cells and some carrying the mutation. The differential diagnosis is from other causes of bone cysts, from Paget's disease, and from hyperparathyroidism with osteitis fibrosa cystica. In the monostotic form treatment is largely orthopaedic. However, the large size of some of the defects in the shafts of the long bones may make conventional stabilization of fractures very difficult. Improvement with intravenous pamidronate (APD) has been reported.

Fig. 20 Polyostotic fibrous dysplasia in a 23-year-old woman. A large cyst in the upper femur led to a spontaneous fracture which subsequently united with conservative treatment. Two ribs on the same side of the body show similar abnormalities. Puberty was precocious but pigmentation absent.

Polyostotic fibrous dysplasia (see Chapter 12.8.6)

Interest in this condition, in which the bone lesions are multiple, arises from its frequent association with pigmentation and sexual precocity, especially in females (McCune–Albright syndrome). The bone lesions and the brown pigmentation are typically associated in position (but not in extent), and may be restricted to one side of the body. Sexual precocity is present in about 50 per cent of females with polyostotic disease, and is then the presenting complaint. It may occur at a very early age, with menstruation and the appearance of secondary sexual characteristics from infancy. Where sexual precocity is not a feature, deformity and fracture are often the first symptoms. Gross deformity of the upper femur and femoral neck produces the 'shepherd's crook' appearance. Asymmetry of the long bones and of the skull are also seen; and in about half of the cases the base of the skull is thickened. The macular pigmentation tends to have smooth borders (in contrast to those of neurofibromatosis) and often does not cross the midline. In a recent long-term follow-up of 15 patients with two or more features of the McCune–Albright triad, the bone lesions tended to increase in size and number, but less rapidly after growth had ceased. Skin lesions were generally bilateral and did not correlate with the site of the bone lesions. There are a number of other features which, like the sexual precocity, are explained by the activating mutation. These include thyrotoxicosis, acromegaly, and Cushing's syndrome. The skeletal lesions may cause complications such as spinal cord compression, and may be associated with hypophosphataemic osteomalacia. Sarcoma formation has been reported, but only after irradiation.

In the polyostotic disease both the plasma alkaline phosphatase and the urinary hydroxyproline levels may be slightly increased and that of plasma phosphate slightly reduced. The pathology is similar to the monostotic form, but it is said that cartilage- and fluid-filled cysts are more common. Microscopically, there is an abundance of woven bone and an increase in osteoblasts and osteoclasts. The cortex and marrow may be virtually replaced by fibrous tissue, so that the bones are fragile. Healing is rapid with abundant callus formation. Radiologically, the bones are deformed, the cortex may be difficult to detect, and the medullary bone takes on a 'ground glass' or 'smoky' appearance.

In polyostotic fibrous dysplasia the main differential diagnosis is from osseous neurofibromatosis; in the former condition there is also pigmentation, bone deformity, and sometimes hypophosphataemic osteomalacia. The borders of the pigmentation are less smooth than in fibrous dysplasia, and there are other cutaneous features of neurofibromatosis; the bone deformity in neurofibromatosis can be quite bizarre, with overgrowth or undergrowth of isolated bones. In neurofibromatosis the characteristic spinal change is a very sharp upper thoracic kyphoscoliosis. Finally, neurofibromatosis often shows clear evidence of dominant inheritance pattern.

The medical treatment of the McCune–Albright syndrome is complex. As for the monostotic from, polyostotic fibrous dysplasia may be improved by pamidronate (APD).

Ectopic mineralization

Deposition of calcium in the soft tissues (ectopic calcification) and on ectopic bone matrix (ossification) has many causes (Table 13). These are nearly always pathological, but often the cause is unknown. In the elderly, calcification in the tissues such as the arteries is so common that it may be regarded as a feature of ageing, in the same way as age-related bone loss. There are some disorders in which calcification and/or ossification are associated with biochemical abnormalities.

Ectopic calcification without bone formation

Calcification can result from previous damage in soft tissues (dystrophic calcification) or from an increase in the circulating concentration of calcium or phosphate (metastatic calcification as, for instance, in advanced renal osteodystrophy).

Dystrophic calcification

This occurs in inherited and acquired disorders involving connective tissue, such as alkaptonuria (intervertebral discs), pseudoxanthoma elasticum (blood vessels), systemic sclerosis, and dermatomyositis, and also after infection, tumours, and trauma. In systemic sclerosis, subcutaneous calcification, often around the phalanges (calcinosis circumscripta), may be part of a syndrome with Raynaud's phenomenon and telangiectases (**CRST**; calcinosis, Raynaud's phenomenon, sclerodactyly, telangiectasia) syndrome; see also Chapter 18.10.3). The calcific deposits can be sufficiently extensive to break through the skin as toothpaste-like material. In dermatomyositis, sheets of subcutaneous calcification can be deposited some time after the initial inflammatory episode characterized by a systemic illness and painful weak muscles; the calcification can be very extensive (calcinosis universalis), but can also disappear rapidly, sometimes in adolescence. Rarely this is associated with hypercalcaemia.

Metastatic calcification

The distribution of the calcification varies inexplicably with its cause; for example, in hypoparathyroidism there is subcutaneous and basal ganglia calcification and in hyperparathyroidism vascular calcification, suggesting that metastatic calcification is not only related to the Ca:P product. Calcification and ossification may also coexist.

Calcification and hypocalcaemia

This occurs in idiopathic and postsurgical hypoparathyroidism, as well as in pseudohypoparathyroidism. There may be extensive ectopic calcification, calcification within the basal ganglia (and outside it), and cataract formation. Pseudohypoparathyroidism is inherited as an autosomal dominant disorder with variable expression; additional clinical features include mental simplicity, round face, short stature, and short third and fourth metacarpals. An important feature is subcutaneous endochondral ossification. End-organ resistance to parathyroid hormone may be due to mutations in the gene responsible for one component ($G_s a$) of the G-protein signalling system (Table 1).

Calcification in hyperphosphataemia

Idiopathic hyperphosphataemia is a rare autosomal recessive disorder, with an increase in the maximal tubular reabsorption of phosphate and an inappropriate increase in the plasma $1,25(OH)_2D$ concentration. Masses of ectopic mineral, which form around the joints from childhood (tumoral calcinosis), may discharge through the skin. Treatment with large oral doses of aluminium hydroxide or other phosphate-binding agents can reduce the plasma phosphate level and the size of the

deposits.

Calcification in inherited hypophosphataemia

A particular feature of X-linked inherited hypophosphataemia is the widespread calcification and ossification of ligaments and tendons at their insertions into the periosteum (so-called Sharpey fibres). This is termed an enthesiopathy. Calcification and new bone formation in the ligamenta flava may produce spinal cord compression.

Idiopathic soft-tissue calcification

This includes calcific tendinitis and so-called calcinosis circumscripta.

Ectopic ossification

Acquired ectopic ossification may occur at the site of injury, such as after hip replacement or at a distance from it, for instance, following paraplegia; or in tumours and in a variety of other disorders. Fibrodysplasia (myositis) ossificans progressiva is a very rare disorder inherited as an autosomal dominant (see below).

Acquired ectopic ossification

Post-traumatic ossification

Local ossification can occur after total hip replacement. The quoted incidence varies widely, depending on the method used to detect it. It is said to occur more often in men than in women and in certain individuals; for instance, where ossification follows hip replacement on one side, it is likely to occur if the contralateral hip is also replaced. The reason for this is unknown. The bone mainly forms in the hip abductors and ossification is classified according to its severity. Disodium etidronate may delay mineralization, but only while it is being given, and non-steroidal anti-inflammatory drugs are also useful. A small dose of radiotherapy may also delay ectopic ossification after total hip replacement.

Ossification after neurological injury

Extensive myositis ossificans can also occur 1 to 4 months after injuries to the head or spinal cord, in muscles distant from the injury such as the major muscles of the thigh. Affected muscles become swollen, red, and warm, and, unless the cord lesion is complete, pain and tenderness also occur. At this time the differential diagnosis may include cellulitis, arthritis, and thrombophlebitis. Radiological calcification is initially absent (appearing at about 6 weeks or more after the injury), but an isotope bone scan will show increased uptake before that. Later there is progressive mineralization, with the eventual appearance of organized bone. Because the bone affects the major periarticular muscles, it leads to joint fixation, particularly of the hips. The plasma alkaline phosphatase level may be increased in the early stages.

Attempted surgical removal of ectopic bone is technically difficult and produces little increase in movement. The ectopic bone recurs, especially if it is removed too early. Oral disodium etidronate at full dose (20 mg/kg body weight per day) may delay the onset of mineralization, but only while it is being given. Likewise, the prevention of further ectopic bone formation after its removal may be delayed by non-steroidal anti-inflammatory drugs or radiotherapy, which should be commenced as soon as possible.

Myositis ossificans can also occur after other neurological diseases, such as poliomyelitis and meningitis, and also after prolonged coma. The reason why ectopic ossification occurs after head injury is unknown; interestingly head injury is associated with an increased rate of fracture healing and excessive callus formation. In such patients the serum contains increased mitogenic activity for osteoblast-like cells; the source of this activity is unknown, but there could be an increase in bone morphogenic proteins.

Ossification can coexist with calcification, and, for instance, extensive ossification of the spinal ligaments in hypoparathyroidism can lead to progressive stiffness. The enthesiopathy in inherited hypophosphataemia (vitamin D-resistant rickets) is a form of ectopic ossification. Ossification of the posterior longitudinal ligament and sternoclavicular hyperostosis is particularly described in Japan. Ligamentous ossification has been noted in patients treated with vitamin A analogues, such as etretinate, for dermatological disorders. The term 'osteoma cutis' covers a number of rare conditions of uncertain cause. Finally, ectopic bone may complicate varicose veins, chronic venous insufficiency, and surgical incisions.

Inherited ectopic ossification

The main inherited cause of ectopic ossification is myositis ossificans progressiva. This disorder is currently classified as a 'heritable disorder of connective tissue'.

Histology suggests (to some) that it is the connective tissue within muscles that is primarily involved and therefore the alternative term 'fibrodysplasia ossificans progressiva' is widely used.

Fibrodysplasia ossificans progressiva is rare, with an incidence of between 1 and 2 per million, which increases with paternal age. Since patients rarely reproduce, most instances represent new mutations. The few family histories demonstrate that the mutant gene is inherited as a dominant with full penetrance but variable expression. Diagnosis depends on the combination of progressive myositis, leading to ossification in the major skeletal muscles, and characteristic bony skeletal abnormalities.

Pathophysiology

Initially there is oedema and cellular infiltration throughout the muscle, with myofibrillar breakdown. Later endochondral ossification leads to mature bone, within which is haemopoietic marrow. Information on the earliest histological appearances is scanty because biopsies are often taken after the acute phase of myositis; for this reason there is still doubt about the primary lesion. Ectopic ossification occurs when mesenchymal or stromal cells take on the behaviour of osteoblasts. This form of cell differentiation could result from an increase in bone-inducing substances or (for unknown reasons) a change in stromal-cell expression. Although the timing of myositis differs widely from one affected patient to another, there is a specific order in which they are affected, from the upper paraspinal to the lower, and from the centre to the periphery. Recent work suggests overexpression of the bone morphogenetic proteins (**BMP**-4). Localization of the mutant gene is hampered by the lack of affected families.

Clinical features

Episodes of myositis are the non-skeletal hallmark of this disease. Typically, the affected muscle becomes swollen and hard, sometimes following injury; after a week or two these features subside, but the apparent improvement is followed in a month or so by ossification within the muscle and progressive joint fixation. Myositis usually begins in the upper paraspinal muscles. By late childhood or adolescence ossification will have occurred within the muscles around the shoulders, hips, and knees, to fix these joints and to complete the disability. The large, striated muscles are affected; ossification does not involve the small muscles of the hands and feet, the diaphragm, the cardiac, or the smooth muscles. Ossification in the muscles around the jaw may fix it almost completely. Although the overall sequence of ossification is characteristic from large upper paraspinal to lower limb muscles, it varies considerably in its rate. For instance, neonates may have sufficient ossification to produce torticollis while, in contrast, late and slow ossification producing stiffness may delay the correct diagnosis until adolescence. Likewise, there may be long symptom-free periods.

The diagnostic skeletal abnormalities affect the big toes (Fig. 21) (and to a lesser extent the thumbs), the cervical spine (Fig. 22), and the metaphyses. The big toes are always abnormal; in the infant, bony changes produce bilateral hallux valgus and, in the adult, fusion produces a short fixed monophalangic big toe. In the cervical spine the vertebral bodies are small and the laminae large. Both are variably fused; and this fusion is independent of nearby ossification of the cervical muscles. Finally, the femoral necks are short and wide and there are exostoses from the metaphyses.

Fig. 22 Fusion of the cervical spine in a young woman with fibrodysplasia ossificans progressiva.

Rare clinical features include early onset baldness, difficulty in hearing, and mental retardation.

Differential diagnosis

Bilateral hallux valgus in the neonate should suggest the possibility of fibrodysplasia ossificans progressiva. In childhood, myositis may be mistaken for soft-tissue sarcoma; and a biopsy showing oedema and increased cellularity may support this or suggest an aggressive fibromatosis. Painful swelling of the masticatory muscles simulates mumps; and progressive stiffness with a fixed abnormal neck suggests the Klippel–Feil syndrome or childhood rheumatoid arthritis.

Management

Once the diagnosis has been made, and this is often delayed, there are four main questions: can the myositis be prevented; if myositis does occur, can subsequent ossification be prevented; what will be the eventual disability; and should ectopic bone be removed?

Since the onset of myositis is quite unpredictable, it is almost impossible to assess the effect of any form of therapy. Corticosteroids have been used, sometimes associated with symptom-free periods. Myositis often follows injury, which should be avoided where possible. It seems likely, but difficult to prove, that myositis is normally followed by ossification. It is to prevent or slow down this ossification that the bisphosphonate EHDP (disodium etidronate) can be given in full doses (20 mg/kg body weight daily by mouth), but there is little evidence that this is effective. In children, continued high-dose etidronate interferes with mineralization, disorganizes the growth plates, and delays fracture healing so that it is not an acceptable long-term treatment. Surgical removal of ectopic bone is technically difficult and recurrence at the site of surgery worsens the disability.

The eventual disability produced by fibrodysplasia ossificans progressiva is severe (Fig. 23). The body moves as in one piece with the legs usually fixed in partial extension. All major joints become completely fixed. The help of a specialized rehabilitation centre is essential.

Fig. 23 Extensive ectopic ossification in the paraspinal muscles fixing the shoulders in a patient with fibrodysplasia ossificans progressiva.

Familial osteoma cutis

This has been reported as a dominantly inherited disorder in a New Zealand family. The proposita had extensive subcutaneous ossification in one leg; relatives had insignificant multifocal subcutaneous ossification in childhood. Similar patients have been described under the name of progressive osseous heteroplasia.

Miscellaneous bone disorders

The skeleton is affected in many systemic diseases (for example, scurvy and the haemoglobinopathies), by the methods used to treat them (for example, parenteral nutrition), and by excessive ingestion of minerals, vitamins, and metals (for example, fluorosis, overdose of vitamins A and D, and metal poisoning). In some, the skeletal changes are clinically important; in others they are a minor aspect of the general illness. This Section ends with a brief description of the obscure disorder fibrogenesis imperfecta ossium.

Scurvy

Vitamin C (ascorbic acid) is necessary for intracellular hydroxylation of peptide-bound proline. In its absence, formation of the collagen molecule is defective, structurally incompetent precursors accumulate within the cell, and collagen-containing tissues are weak. Scurvy is very rare, occurring most often in neglected infants who do not receive fruit juice or ascorbic acid for several months. Extensive subperiosteal haemorrhage leads to pain and immobility; the legs are held in a 'frog-like'

position. In the adult, there is perifollicular haemorrhage, purpura, and bleeding gums. Radiographs in infancy show a widened zone of provisional calcification in the metaphyses, with a proximal disordered area representing the destroyed primary spongiosa and failure of new bone formation. The edges of the metaphyses may show small spurs, and epiphyseolysis may occur. With healing the subperiosteal haematoma calcifies.

The clinical picture of scurvy may suggest non-accidental injury, but scurvy is far less common. Similar radiographic appearances have been described in cases of copper deficiency.

The haemoglobinopathies

In the inherited disorders of haemoglobin (see Section 22) the skeleton is often abnormal. This may result from a hyperplastic bone marrow and overactivity of the osteoblasts, so that the skull, facial bones, and long bones are thickened. Additional features include collapse of the weight-bearing bones and disorganization of the joints following bone infarction. This is especially seen in sickle-cell disease, haemoglobin C disease, and haemoglobin SC compound-heterozygotes. In b-thalassaemia an increase in osteoid thickness has been described which resembles that of osteomalacia.

Parenteral nutrition

Prolonged parenteral nutrition can produce a form of bone disease with similarities to osteomalacia. The main symptom is periarticular bone pain, particularly in the ankles. Histology shows impaired mineralization of bone, and biochemistry an increase in plasma alkaline phosphatase, in urinary calcium, and sometimes in plasma calcium levels. The radiographic appearances suggest osteoporosis. Since patients on total parenteral nutrition are invariably ill to begin with, and many have malabsorption, there are several probable causes for this disorder; aluminium intoxication may contribute.

Fluorosis

Deposition of excess fluoride in the skeleton can result from an excess in the diet (endemic fluorosis), from industrial exposure (during the manufacture of aluminium, steel, and glass, and from exposure to the dust of fluoride-containing rock), and from the administration of sodium fluoride in treatment. The most severe effects are seen in endemic fluorosis, well described from the Punjab.

There is considerable disability, with spinal rigidity, restricted movements of the joints, and flexion deformities of the hips and knees. There is a generalized increase in bone density (with loss of the normal corticomedullary junction), and the tendons, ligaments, and sometimes muscles may be mineralized. This can produce compression of the spinal cord and its roots, with progressive neurological disability. Mineralization of tendon insertions may be seen in other situations, such as inherited hypophosphataemia (see above), retinoid treatment, and fibrogenesis imperfecta ossium (see below).

Increased levels of fluoride can affect the enamel of developing teeth, producing chalky-white patches, yellow-brown discoloration, and other defects.

The diagnosis of fluorosis depends on the radiographic changes and an increased urinary excretion of fluoride (which is an index of current exposure). When a bone biopsy is performed (most often to exclude other causes of increased bone density), histology shows an increase in new bone formation with an increase in the width of osteoid borders. There is also an increase in fibrous tissue and bone resorption. When the biopsy includes an area of tendinous insertion, this may be mineralized.

Sodium fluoride has been given widely to treat osteoporosis and produces increased vertebral density. Current evidence does not suggest that this increases vertebral bone strength and suggests that there is an increase in appendicular bone fracture. The main effect of the fluoride ion is to stimulate new bone formation, while fluoroapatite may also reduce resorption. The new bone appears to be mainly woven in character and imperfectly mineralized. Although there is no doubt about the anabolic effects of fluoride on bone, current controversies about its clinical usefulness depend, in part, on the dose used.

Vitamin A

Retinoic acid and its derivatives have profound effects on osteoblast function . Vitamin A poisoning produces characteristic periostitis in the young skeleton, and therapeutic retinoids (such as etretinate) causes widespread ligamentous calcification. Acute and chronic forms of vitamin A overdosage are described. In infants, it is uncommon under the age of 1 year. There is anorexia and failure to thrive; other features include pruritus, hepatosplenomegaly, jaundice, alopecia, dry skin, and fissures around the lips. Hard, tender masses appear in the limbs, and radiographs show periosteal new bone formation, especially in the diaphyses of the tibias, which later blends into the cortex. A number of other radiological features include shortening of the shafts of the long bones, splaying of the metaphyses, enlargement and premature fusion of the ossification centres, and flexion deformities of the legs.

The prolonged use of retinoids for the treatment of skin disease, such as psoriasis and ichthyosis, leads particularly to calcification of the spinal ligaments, causing stiffness and reduced mobility. There is a resemblance to Forestier's disease (diffuse idiopathic skeletal hyperostosis).

Vitamin D

Vitamin D poisoning can result from inappropriate therapeutic overdosage or accidental overconsumption. This leads to the features of hypercalcaemia (see Chapter 10.3 and Chapter 12.6.2) without detectable effects on the skeleton. The main opportunities for overdose exist when potent preparations of vitamin D are used inappropriately (as, for instance, to treat skin conditions and tuberculosis). Chronic vitamin D overdosage leads to soft-tissue calcification, especially in the arteries and kidneys. After several years, progressive stiffness in the spine, major joints, and feet lead to difficulty in walking. Radiographs show ligamentous calcification. Another cause of hypercalcaemia is an excess of $1,25(OH)_2D$ produced by granulomas, especially those of sarcoidosis following exposure of the skin to sunlight. The biochemical effects are the same as those of vitamin D poisoning. The $1,25(OH)_2D$ concentration increases with that of 25(OH)D, and the hypercalcaemia of sarcoidosis often occurs during the spring in persons with outdoor jobs (such as farmers, window cleaners) or after foreign holidays in the sun. Treatment with corticosteroids and removal from sunlight rapidly reduces the hypercalcaemia and the elevated $1,25(OH)_2D$ levels.

Idiopathic hypercalcaemia can occur in infancy. Now named the Williams syndrome, it is associated with an unusual 'elfin face', mental simplicity, and congenital heart disease. Radiographs of the long bones show increased density of the metaphyses. The cause is not fully understood, but the concentration of 25(OH)D may be increased when the patients are hypercalcaemic. Deletion of the elastin locus on chromosome 7 has also been described and may explain the cardiac abnormalities.

Lead (see also Chapter 8.1)

Lead has unique effects on the skeleton, which may be combined with the other manifestations of lead intoxication. Lead deposition in the growing skeleton produces a radiologically dense line near the growth plate. When exposure to lead has been intermittent, or the condition has been treated, this may be a single, relatively narrow line, which is superseded by apparently normal bone. If exposure to lead recurs, a further line will appear.

Lead poisoning due to industrial pollution is thought to be widespread, but the skeleton is affected only when lead exposure has been considerable. In children, one recognized source is lead-containing paint. Other sources include contaminated water from old lead pipes, eye blackener used by Asian women (which contains up to 88 per cent lead sulphide), inhalation of lead fumes from burning old battery cases, and alcoholic drinks stored in vessels of lead glass or coated by lead enamel glaze.

Clinical features involve: the gastrointestinal tract (abdominal pain, colic with constipation, and a blue pigmentation of the gingival margin); the neuromuscular system, with weakness; and encephalopathy, with restlessness, irritability, and lethargy. Renal manifestations are described in Section 20.

Characteristic radiological features include widened skull sutures (due to raised intracranial pressure in infants), dense deposits in the gastrointestinal tract indicating heavy-metal ingestion, and dense lines in the metaphyses (lead lines). Such lines are an important clue to lead poisoning in infants and children up to about the age of 6 years. They occur most commonly around the knees, wrists, and ankles, and appear after about a month of chronic poisoning. The diagnosis of lead poisoning is confirmed by an increase in plasma and urinary lead levels. There are other causes of radiologically dense metaphyses. These include: other heavy metals—bismuth, mercury, or phosphorus; vitamin D intoxication and idiopathic hypercalcaemia of infancy; cretinism; and healing rickets. In practice, there is often difficulty in deciding

on the significance of dense metaphyses due to excessive calcium in an otherwise well child, since this appearance can occur in the normal growing skeleton.

Aluminium (See also Chapter 8.1)

Aluminium in water is not significantly absorbed through the intestine, but this barrier was effectively removed in the early days of haemodialysis treatment for endstage renal failure. The resultant accumulation of aluminium in the skeleton in patients in some units where the aluminium content of tap water was high led to the occurrence and recognition of 'dialysis bone disease'. There was a close clinical association with dialysis dementia, also related to aluminium poisoning. The clinical features of this bone disease were proximal myopathy, multiple painful spontaneous fractures with radiographic evidence of osteopenia (osteoporosis), histological evidence of excess osteoid with aluminium deposition near the calcification front, and an absence of response to vitamin D metabolites.

In renal glomerular failure, aluminium may also accumulate in patients given oral aluminium hydroxide to reduce plasma phosphate (in order to lessen hypocalcaemia and subsequent secondary hyperparathyroidism). Aluminium bone disease can also occur in patients on prolonged parenteral nutrition.

The pathology of aluminium bone disease is not fully understood, but it seems likely that in some instances aluminium reduces osteoblast activity. Two different forms are described: in the first, there is excessive osteoid (with an appearance like that of osteomalacia); and in the second, there is little increase in osteoid with reduced osteoblastic activity. It is likely that the different histological features are related to the amount of aluminium in the bones.

Cadmium (See also Chapter 8.1)

Contamination of drinking water by cadmium and its accumulation in the body causes renal tubular damage with multiple biochemical defects. Cadmium intoxication is one of the acquired causes of the Fanconi syndrome which leads to rickets or osteomalacia. Industrial exposure to cadmium fumes can produce hypophosphataemic osteomalacia.

Exceptionally, chronic lead poisoning can also produce osteomalacia by the same mechanism as cadmium; and copper 'poisoning' causes the Fanconi syndrome and bone disease of Wilson's disease.

Fibrogenesis imperfecta ossium

This is a very rare, apparently acquired disorder, characterized by excessive bony fragility due to the replacement of normal bone with a fibre-deficient, poorly mineralized matrix. The cause is unknown. In recorded cases, the main clinical feature has been pathological fractures first presenting in adult life. In most patients, progressive disability has followed, with more fractures that fail to unite. Radiologically, the trabeculae throughout the skeleton appear to be thickened. There is also ectopic mineralization around large joints and tendon insertions. Biochemically, plasma calcium and phosphate levels are normal, but the alkaline phosphatase level is moderately raised. In the urine, monoclonal light chains may be present. The diagnosis is confirmed by the examination of undecalcified bone. This shows defective mineralization and wide osteoid seams suggesting severe osteomalacia, but the osteoid is not birefringent under polarized light and the normal structure of bone collagen under electron microscopy is absent. The differential diagnosis is from those disorders that produce widespread coarse trabeculation throughout the skeleton, and those that produce the histological changes of osteomalacia. In the first category, Paget's disease of bone, renal glomerular osteodystrophy, and fluorosis should be excluded, and in the second, axial osteomalacia has some similarities; in this very rare osteosclerotic disorder, both histology and radiographs suggest that the osteomalacia is limited to the spine, pelvis, and ribs.

Since the cause of fibrogenesis imperfecta ossium is unknown, treatment to date has been largely empirical. The occasional finding of an excess of plasma cells in the bone marrow, or a monoclonal gammopathy, or light-chain proteinuria, has led to apparently successful treatment with melphalan and prednisolone. Where surgery is indicated for fractures, particularly of the femoral neck, this is difficult because of the extreme fragility of the bones.

Although it seems likely that the defect may be related to an acquired disorder of bone collagen, no consistent abnormality has been detected.

Sudeck's atrophy

This is one of many synonyms of what is more often referred to as 'algodystrophy' (painful dystrophy). Its features are pain, swelling, and tenderness, most often of a limb, which is persistent and recurrent. Early oedema and erythema may be replaced by a dystrophic phase that may last for months, with pallor or cyanosis. The skin and subcutaneous tissues may atrophy and there is increased sweating and often worsening of the pain. The main known precipitating cause of algodystrophy is trauma, such as forearm (Colles) fracture. In at least one-quarter of patients there is no identifiable cause. The recorded prevalence of algodystrophy varies widely, depending on its definition and how closely it is looked for; in a specific study made to identify algodystrophy it was found in 25 per cent of patients 9 weeks after Colles fracture, a frequency far higher than normally recorded.

The treatment of algodystrophy is difficult and requires consideration of the whole patient. Adequate pain relief, reassurance, and explanation are essential. Numerous other measures have been proposed; these include courses of corticosteroids and calcitonin, regional sympathetic block, and surgical sympathectomy, none of which is consistently useful. Bisphosphonates, such as pamidronate, may prove to benefit symptoms and outcome, but no controlled trial data are currently available to substantiate their general use where bone loss occurs (see below).

Algodystrophy is often associated with localized bone loss, and some include so-called regional and transient migratory osteoporosis within the algodystrophy syndrome, although the reasons for doing so are tenuous. Characteristically, there is severe localized osteoporosis associated with pain which recurs in different limbs. The increased bone resorptive activity associated with bone loss may be identified by the use of bisphosphonate skeletal scintigraphy, which may assist diagnosis in the early phases of this condition. Further investigation may show that the osteoporosis is more widespread than suspected with, for instance, vertebral compression fractures. Rarely, osteoporosis occurs in pregnancy (see above) predominantly affecting the spine, but also other peripheral bones, leading, for instance, to femoral neck fractures.

Further reading

Bone physiology

Avioli LV, Krane SM (1998). *Metabolic bone disease*, 3rd edn. Academic Press, San Diego.

Bilezikian JP, Raisz LG, Rodan GA (2002). *Principles of bone biology*, 2nd edn. Academic Press, San Diego.

Byers PH (1995). Disorders of collagen biosynthesis and structure. In: Scriver CR, *et al.*, eds. *The metabolic basis of inherited disease*, 7th edn, Volume III, pp 4029–77. McGraw Hill, New York.

Jilka RL (1998). Cytokines, bone modelling and oestrogen deficiency: a 1998 update. *Bone* **23**, 75–81.

Manolagos SC, Jilka RL (1995). Bone marrow, cytokines, and bone remodelling. *New England Journal of Medicine* **332**, 305–11.

Royce PM, Steinmann B (1992). *Connective tissue and its heritable disorders. Molecular, genetic and medical aspects*, 1st edn. Wiley–Liss, New York.

Russell RGG (1997). The assessment of bone metabolism *in vivo* using biochemical approaches. *Hormone and Metabolic Research* **29**, 138–44.

Suda T, *et al*. (1999). Modulation of osteoclast differentiation and function by the new members of the tumour necrosis factor receptor and ligand families. *Endocrine Reviews* **20**, 345–57.

Diagnosis of bone disease

Blumsohn A, Eastell R (1997). The performance and utility of biochemical markers of bone turnover: do we know enough to use them in clinical practice? *Annals of Clinical Biochemistry* **34**, 449–59.

Favus MJ (1999). *Primer on the metabolic bone diseases and disorders of mineral metabolism*, 4th edn. Lippincott, Williams and Wilkins, Philadelphia.

Osteomalacia and rickets

Francis RM, Selby PL (1997). Osteomalacia. *Baillieres Clinical Endocrinology and Metabolism* **II**, 145–63.

O'Riordan JLH (1997). Rickets, from history to molecular biology, from monkeys to YACS. *Journal of Endocrinology* **154**, S3–S13.

Parfitt AM (1998). Osteomalacia and related disorders. In: Avioli LV, Krane SM, eds. *Metabolic bone disease*, 3rd edn, pp 327–86. Academic Press, San Diego.

Paget's disease

Barker DJP *et al.* (1977). Paget's disease of bone in 14 British towns. *British Medical Journal* **1**, 1181–3.

Cooper C, *et al.* (1999). Epidemiology of Paget's disease of bone. *Bone* **24**, 35–55 (Suppl.).

Delmas PD, Meunier PJ (1997). The management of Paget's disease of bone. *New England Journal of Medicine* **336**, 558–66.

Kanis JA (1998). *Pathophysiology and treatment of Paget's disease of bone*, 2nd edn. Martin Dunitz, London.

Parathyroids and bone disease

Bassett JHD, Thakker RV (1995). Molecular genetics of disorders of calcium homeostasis. *Baillieres Clinical Endocrinology and Metabolism* **9**, 581–608.

Osteogenesis imperfecta: the brittle bone syndrome

Glorieux FH, *et al*. (1998). Cyclic administration of pamidronate in children with severe osteogenesis imperfecta. *New England Journal of Medicine* **339**, 947–52.

Pope FM (1998). Molecular abnormalities of collagen and connective tissue. In: Maddison PJ, *et al.*, eds. *Oxford textbook of rheumatology*, 2nd edn, pp 353–404. Oxford University Press, Oxford.

Smith R (1995). Idiopathic juvenile osteoporosis: experience of twenty one patients. *British Journal of Rheumatology* **34**, 68–77.

Smith R (1999). Osteogenesis imperfecta; the brittle syndrome. An update. *Current Orthopaedics* **13**, 218–22.

Marfan's syndrome

De Paepe A, *et al.* (1996). Revised diagnostic criteria for the Marfan syndrome. *American Journal of Medical Genetics* **62**, 417–26.

Pyeritz RE (1993). The Marfan syndrome. In: Royce PM, Steinmann B, eds. *Connective tissue and its heritable disorders*, pp 437–68. Wiley–Liss, New York.

Shores J, *et al.* (1994). Progression of aortic dilatation and the benefit of longterm b adrenergic blockade in Marfan's syndrome. *New England Journal of Medicine* **330**, 1335–41.

Ehlers–Danlos syndrome

Pope FM (1998). Molecular abnormalities of collagen and connective tissue. In: Maddison PJ, *et al.*, eds. *Oxford textbook of rheumatology*, 2nd edn, pp 353–404. Oxford University Press, Oxford.

Steinmann B, Royce PM, Superti-Furga A (1993). The Ehlers–Danlos syndrome. In: Royce PM, Steinmann B, eds. *Connective tissue and its heritable disorders*, pp 351–401. Wiley–Liss, New York.

Homocystinuria

Isherwood DM (1996). Homocystinuria. Early diagnosis and intervention reduces risk of visual impairment and thromboembolism. *British Medical Journal* **313**, 1025–6 [Editorial].

Nygard O, *et al*. (1997). Plasma homocysteine levels and mortality in patients with coronary artery disease. *New England Journal of Medicine* **337**, 230–6.

Skovby F (1993). The homocystinurias. In: Royce PM, Steinmann B, eds. *Connective tissue and its heritable disorders*, pp 469–86. Wiley–Liss, New York.

Alkaptonuria

Hazleman BL, Adebajo AO (1993). Alcaptonuria. In: Royce PM, Steinmann B, eds. *Connective tissue and its heritable disorders*, pp 591–602. Wiley–Liss, New York.

Hypophosphatasia

Whyte MP (1993). Osteopetrosis and the heritable forms of rickets. In Royce PM, Steinmann B, eds. *Connective tissue and its heritable disorders*, pp 563–9. Wiley–Liss, New York.

Whyte MP (1999). Hypophosphatasia. In: Favus MJ, ed. *Primer on the metabolic bone diseases and disorders of mineral metabolism*, 4th edn, pp 337–9. Lippincott, Williams and Wilkins, Baltimore, MD.

Lysosomal storage diseases

Leroy JG, Weismann U (1993). Disorders of lysosomal enzymes. In: Royce PM, Steinmann B, eds. *Connective tissue and its heritable disorders*, pp 613–39. Wiley–Liss, New York.

Mankin HJ, *et al.* (1990). Metabolic bone disease in patients with Gaucher's disease. In: Avioli LV, Krane SM, eds. *Metabolic bone disease*, 2nd edn, pp 730–52. WB Saunders, Philadelphia.

Skeletal dysplasias

Francomano CA, McIntosh I, Wilkins DJ (1996). Bone dysplasias in man: molecular insights. *Current Opinion in Genetics and Development* **6**, 301–8.

Gelb BD, *et al.* (1996). Pycnodysostosis, a lysosomal disease caused by cathepsin K deficiency. *Science* **273**, 1236–8.

Horton WA (1996). Molecular genetic basis of the human chondrodysplasias. *Endocrinology and Metabolism Clinics of North America* **25**, 683–97.

Horton WA, Hecht JT (1993). The chondrodysplasias. In: Royce PM, Steinmann B, eds. *Connective tissue and its heritable disorders*, pp 641–75. Wiley–Liss, New York.

Schipani E, *et al*. (1996). Constitutively activated receptors for parathyroid hormone and parathyroid hormone-related peptide in Jansen's metaphyseal chondrodysplasia. *New England Journal of Medicine* **335**, 708–14.

Osteopetrosis (marble bones disease)

Key LL, Ries WL (1996). Osteopetrosis. In: Bilezikian JP, Raisz LG, Rodan GA, eds. *Principles of bone biology*, 1st edn, pp 941–50 Academic Press, San Diego, CA.

Fibrous dysplasia

Chapurlat RD, *et al.* (1997). Long term effects of intravenous pamidronate on fibrous dysplasia of bone. *Journal of Bone and Mineral Research* **12**, 1746–52.

Ectopic ossification

Connor JM, Evans DAP (1982). Fibrodysplasia ossificans progressiva. The clinical features and natural history of 34 patients. *Journal of Bone and Joint Surgery* **64B**, 76–83.

Kaplan FS (1998). Fibrodysplasia ossificans progressiva. *Clinical orthopaedics and related research* **346**, 1–140 (Symposium).

Smith R, Athanasou N, Vipond SE (1996). Fibrodysplasia (myositis) ossificans progressiva; clinicopathological features and natural history. *Quarterly Journal of Medicine* **89**, 445–56.

Miscellaneous bone disorders

Carr AJ, *et al.* (1995). Fibrogenesis imperfecta ossium. *Journal of Bone and Joint Surgery* **77B**, 820–9.

Smith R (1993). Heritable bone diseases, chondrodysplasias and skeletal poisons. In Nordin BEC, Need AG, Morris HA, eds. *Metabolic bone and stone disease*, 3rd edn, pp 213–48. Churchill

Livingstone, Edinburgh.

Sudeck's atrophy

Littlejohn GO (1998). Algodystrophy (reflex sympathetic dystrophy). In: Maddison PJ, *et al.*, eds. *Oxford textbook of rheumatology*, 2nd edn, pp 1679–89. Oxford University Press, Oxford.

19.2 Inherited defects of connective tissue: Ehlers-Danlos syndrome, Marfan's syndrome, and pseudoxanthoma elasticum

F. M. Pope

Introduction

Ehlers–Danlos syndrome, pseudoxanthoma elasticum, and the Marfan syndrome are characterized by the fragility of connective tissues and thus cause diverse clinical disease. These inherited defects disrupt the integrity of structural proteins found in the skin, ligaments, cartilage, and vasculature and share common clinical features (Table 1). Connective tissue is also affected in the pleura, peritoneum, heart valves, gastrointestinal system, muscles, and other tissues with similar scaffolding components. Moreover, defects of connective tissue also disrupt basement membranes present in the eyes, skin, and kidneys.

A notable feature of connective tissues is the existence of molecular interactions between structural proteins and extracellular matrix. These interactions are important early in development and involve early embryonic patterning co-ordinated by the expression of homeobox proteins, including those of the hedgehog family. Inherited defects of protein constituents in connective tissues may thus disturb many tissues during development and organogenesis.

For all these reasons, inherited defects of connective tissue impinge widely on the practice of medicine. In the ageing population, increased fragility of the skin, rupture of blood vessels, and laxity of ligaments, as well as defects of cartilage and bone, overlap to form a series of disorders that declare themselves in adult life. Such 'degenerative disorders' include osteoporosis, osteoarthritis, and arterial aneurysms. In the light of spectacular advances in the understanding of the molecular structure and genetics of connective tissue components, it seems likely that many aspects of medicine hitherto ascribed to age-related degeneration, will ultimately prove to have strong genetic components. A valid, molecular understanding of these processes may well emerge. It also appears likely that discrete clinical conditions now recognized as the Marfan syndrome, Ehlers–Danlos syndrome, and pseudoxanthoma elasticum will prove to have diverse and genetically determined counterparts that are responsible for the so-called degenerative disorders in the population at large.

Ehlers–Danlos syndrome (EDS)

Many clinical subtypes of Ehlers–Danlos syndrome have been recognized and increasingly these variants are being associated with distinct abnormalities of collagen structure. Abnormal collagen structure in EDS leads to multisystem disease (Table 2). Common features of this syndrome include fragile skin and laxity of the joints and ligaments (Fig. 1). In certain subtypes there is a particular fragility of tissues including arteries, the wall of the intestine, spinal ligaments, or even teeth; these propensities have led to discrete syndromic recognition (Table 3). The classical form of Ehlers–Danlos syndrome is characterized by excess cutaneous extensibility, bruising, and molluscoid pseudotumours. Careful examination of the skin of patients with EDS may show laxity, pendulousness, and fragility with easy splitting at different stages. In childhood, the skin tends to be hyperelastic but with advancing age, laxity combined with pendulousness, becomes obvious. Other forms of EDS may show hyperelastic and droopy skin in the same or different sites from the outset. The EDS V subtype is now obsolete.

Fig. 1 Ehlers–Danlos syndrome. (a) Hyperelasticity of the skin (EDS II); (b) atrophic and pigmented papyraceous scars (EDS II); (c) joint hypermobility (EDS III); (d) severe pes planus (EDS VI); (e) dentinogenesis imperfecta (EDS VII), this patient had a deletion of exons 3–6 of the *COL1A1* gene; (f) premature periodontitis (EDS VIII).

After the original description, it was realized that some patients with EDS were susceptible to spontaneous arterial rupture with its associated lethal consequences. Affected women suffered fetal prematurity, and examination of their skin showed depletion of collagen and an increase in elastin fibres. This particular condition, which is associated with early fatality from vascular rupture, is correlated with defects in a type III collagen, as shown below. There are obstetric, rheumatological, orthopaedic, and abdominal complications of this vascular form of Ehlers–Danlos syndrome (type IV)—as well as an appreciable incidence of bladder-neck obstruction and ureteric reflux. Representative clinical features of EDS subtypes are illustrated in Fig. 1 and Fig. 2 (and see also Plate 1).

Fig. 2 EDS type IV (vascular type). (a) Acrogeria—a specific clinical feature of EDS IV. Note the large eyes and thin nose (Madonna facies) with perioral wrinkling. (b) Premature wrinkling of the skin on the dorsum of the hands; note also the joint contractures superficially resembling rheumatoid arthritis. (c) Pretibial bruising and haemosiderosis. (See also Plate 1.)

Clinical genetics

Ehlers–Danlos syndrome type VI is caused by a recessively inherited deficiency of lysyl hydroxylase leading to underhydroxylation of collagen molecules (see Fig. 4(b)). Ehlers–Danlos type IV is due to defects and a deficiency of type III collagen in blood vessels, and is inherited as an autosomal dominant disorder. The biochemical defect in autosomal dominant Ehlers–Danlos syndrome type VII results from a failure to remove the N-terminal procollagen extensions of collagen. Defects in type III collagen are now implicated in the cutaneous ligaments and arterial fragility of the three major EDS variants. These defects compromise the assembly of collagen fibrils and provide an important biological model for other EDS subtypes that have now been identified, thus allowing diagnostic criteria to be married to an understanding of the molecular defect in collagen and connective tissue matrix.

Ehlers–Danlos syndrome may show classical X-linked, autosomal recessive or autosomal dominant transmission. These patterns reflect the complex assembly of the helical collagen molecule and its susceptibility to protein suicide. Because it is a complex structure, mutations in a single glycine located in the collagen helices disrupt up to seven-eighths of the assembled homocollagen trimers and 75 per cent of collagen heterotrimers, depending on the particular stoichiometry. For these reasons, collagen defects of this type behave as autosomal dominant traits, while the enzymatic deficiencies of collagen formation segregate as autosomal recessive traits with little or no expression in heterozygotes. Convincing autosomal recessive inheritance is common in EDS type VI associated with pronounced cutis laxa, due to gross distortions of the assembly of type I collagen that lead to a hieroglyphic appearance to these fibres under electron microscopy (Fig. 3(a)). In contrast, mutations leading to EDS type IV result from dominant mutations of type III collagen.

Ehlers–Danlos syndrome types I and II

EDS I and II are associated with the classical features outlined in Table 3. Easy skin-splitting, especially over bony prominences on the forehead, elbows, knees, and chin, shows itself in childhood. Notable features of the condition include epicanthic folds, blue sclerae, and fibrous nodules over the knees and ankles. Mitral valve prolapse is common but does not usually result in dilatational rupture of the valve. Unlike Ehlers–Danlos syndrome type IV, fragile blood vessels do not occur, but venous varicosities, premature bilateral hallux valgus, and distortion of the cornea leading to astigmatism, as well as premature osteoarthritis, are common. Most EDS type I or II families show linkage to the collagen V a-1 or V a-2 genes; which are located, respectively, on human chromosomes 9 or 2q34. Mutation analysis usually reveals substitutions of glycines and exon-skipping events. In some instances, null alleles lead to the deficiency of type V collagen domains. It seems likely that defects in the interactive properties of the N-terminus of type V collagen, which normally protrudes from the surface compound fibres comprising types I, III, and V collagen, impair normal interactions with other matrix components. Misdirection of the collagen fibrils leads to the generation of the so-called 'cauliflower' fibrils of EDS types I and II (Fig. 3(b) and (c)). The clinical consequences being of fragile skin, ligaments, tendons, and corneas, as well as defective articular surfaces.

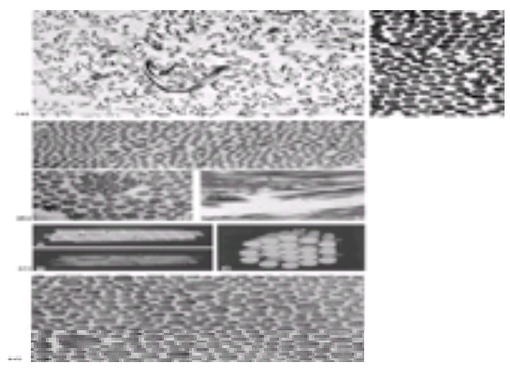

Fig. 3 Ultrastructural abnormalities of collagen in EDS. (a) 'Hieroglyphic' collagen fibres in EDS VII, indicating very severe disruption of fibril packing—compare with healthy collagen shown on the right of the figure. (b) Misassembled 'cauliflower' fibrils of skin and ligaments in EDS II. The left panel shows transversely fused fibres, which in longitudinal sections appear to splay distally. They resemble transversely sectioned cauliflower heads (from Nicholls *et al*. (1996) with permission). (c) Diagrammatic representation (in a) of compound collagen I and III fibres, composed of quarter-staggered individual triple helices. The dark collagen type V molecules (b) regulate fibril diameter; their protruding N-termini (shown in (c)) can interact with other matrix components (from Birk *et al*. (1990) with permission). (d) Dual distribution of collagen fibre size in EDS type IV.

Ehlers–Danlos syndrome type III/benign hypermobile syndrome

This is the most common and least differentiated variant of Ehlers–Danlos syndrome and, for this reason, its diagnosis is frequently overlooked. The disorder is associated with tall stature, ready bruising, blue sclerae, and osteoporosis. Occasionally, more severe defects including osteogenesis imperfecta and Marfan's syndrome occur. Occasionally pseudoxanthoma and other Ehlers–Danlos syndrome variants occur as part of this disorder. As yet uncharacterized defects in collagen, elastin, fibrillin, proteoglycan, and connective tissue-modifying enzymes may all be responsible for the hypermobility of joints and other manifestations in this category of Ehlers–Danlos syndrome (Fig. 1(c)). Patients with Marfan's syndrome may show extensible skin and osteoporosis that complicate Ehlers–Danlos syndrome types III, VI, and VII; Marfan-like stature may be observed in patients with osteogenesis imperfecta.

Ehlers–Danlos syndrome type IV—'vascular form'

This autosomal dominant vascular form of Ehlers–Danlos syndrome is typically accompanied by pretibial ecchymoses over the knees and shins as well as acrogeria (Fig. 2). Acrogeria refers to prematurely aged extremities with thinning of the skin on the dorsum of the hands, feet, and shins. The features are combined with the so-called 'Madonna' facial appearance of large eyes, nasal thinning, and small earlobes. Some patients may have a Marfanic appearance. Rarely there is acro-osteolysis with unexplained androgenic alopecia in females as well as congenital talipes, hip dislocations, and tendon contractures; displacement of the metacarpophalangeal joints in the hands may superficially resemble the changes of rheumatoid arthritis. Fragility of pleuroperitoneal membranes or the colonic wall is a common feature in this vascular subtype but occasionally it may complicate other types of EDS, including types I, II, and III.

The main concern with type IV Ehlers–Danlos syndrome is of spontaneous rupture of medium or large arteries, which may be lethal at any point from mid-adolescence to late-adult life. Aneurysms of small, medium, and large arteries, including the aorta, are common. Angiographic studies reveal a dilatated and tortuous arterial tree, including the carotid bifurcation, and major aortic or iliac disease. Dissections are common especially after ill-advised angiography or surgery; these present with unexplained abdominal pain that proves to be due to intestinal ischaemia. Histological examination of the skin reveals dermal thinning with depletion of dermal

collagen and an overproliferation of elastic fibres. Examination of the skin by electron microscopy usually reveals marked variability in collagen fibril size (Fig. 3(d)).

In a recent review by Pepin and colleagues of the medical and surgical complications in 220 index patients and 199 of their affected relatives with biochemically confirmed EDS type IV, the underlying COL 3A1 mutation was identified in 135 propositi. One-quarter of the index patients had had a first complication by the age of 20 years and more than 80 per cent had at least one complication by the age of 40. Most deaths resulted from arterial rupture but bowel rupture, usually of the sigmoid colon was frequent. It was noteworthy that in 81 women with EDS IV who had become pregnant, 12 died as a result of disease complications during pregnancy. Overall the median lifespan of the whole group was reduced to 48 years.

Molecular pathology

Collagen III is the dominant collagen in skin, blood vessels, tendons, ligaments, gastrointestinal tract, and pleuroperitoneal cavity linings. This explains the diverse multisystem phenotype of type IV EDS. Disturbed assembly, as well as haploinsufficiency of type III collagen explains the wide-ranging severity of EDS type IV, although some affected patients suffer a mild clinical phenotype resembling EDS type III (Fig. 4(a)). Numerous mutations in the type III collagen gene have been found in EDS type IV; most of these mutations are private, although several are associated with 'hot spots' in the complex collagen gene structure that are located in exons 7, 16, and 24. The complications of EDS type IV cannot be predicted from the nature of the specific mutations in COL 3A1.

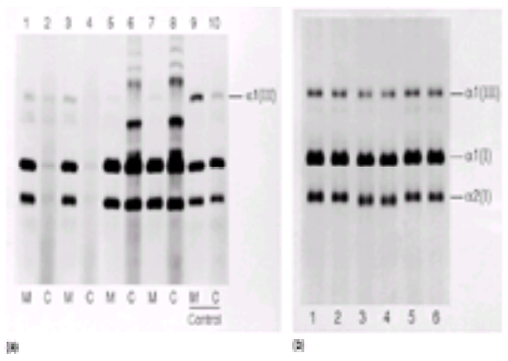

Fig. 4 Molecular analysis of collagen in EDS. (a) Typical collagen type III electrophoretic profile in fibroblasts after biosynthetic labelling in culture. There is virtually complete deficiency (tracks 5–8) or haploinsufficiency (tracks 1–4) compared with the normal pattern (9–10). M, collagen in culture medium; C, collagen recovered from cells. (b) Electrophoresis of radiolabelled collagen proteins in fibroblasts obtained from a patient with severe pes planus due to EDS VI, showing accelerated migration (tracks 3–4) of underhydroxylated, compared with normal, collagen molecules (tracks 1–2; 5–6).

Ehlers–Danlos syndrome types III, IV, and VI, with overlap to pseudoxanthoma elasticum

Much overlap is seen between these syndromes and others that affect connective tissue, including Stickler syndrome.

A characteristic feature of Ehlers–Danlos syndrome type III is the absence of scarring after skin injury; none the less, the skin is doughy and extensible but without fragility. The benign hypermobile syndrome is associated with joint hypermobility but without extensible skin. While this syndrome may be adapted to extreme sporting skills and gymnastics, it may also indicate heterozygosity for other collagen defects such as Ehlers–Danlos syndrome type VI, or even pseudoxanthoma elasticum.

The syndrome is associated with persistent arthralgia without evidence of inflammatory joint disease and is difficult to treat. Treatment includes physiotherapy, rest, and graded exercise combined with conventional pain relief. Later, joint-stabilizing exercises or supports combine with proprioceptive enhancement, and cognitive therapy may be beneficial. It is unknown whether these patients have defective pain receptors or whether they have acquired or inherited disturbed pain reception as a result of their long-standing collagen abnormality.

Marfan's syndrome

The Marfan syndrome is an archetypal defect of connective tissue with a strong hereditary basis causing characteristic skeletal, cardiovascular, and ocular disease. Patients with Marfan's syndrome are disproportionately tall and thin with abnormally long extremities and, often, a cadaverous physique (Fig. 5). Abraham Lincoln was possibly affected. Marfan's syndrome is caused by mutations in the human fibrillin gene and is an autosomal dominat trait.

Fig. 5 Marfan's syndrome. Early illustration of a family with skeletal and ophthalmic features transmitted from the affected father to his daughter and two sons.

Marfan's syndrome is not rare; it affects both sexes and occurs with a frequency of about 1 in 10 000. Marfan's syndrome has been reported in nearly all ethnic groups.

Diagnostic criteria

The Marfan syndrome overlaps with other inherited connective tissue disorders including Ehlers–Danlos syndrome type III, benign hypermobility syndrome, pseudoxanthoma elasticum, osteogenesis imperfecta, and homocystinuria. Typically, there is joint hypermobility, hyperextensibility of the skin with striae, blue sclerae, and tall stature. Combinations of one major and two minor criteria are sufficient to diagnose Marfan's syndrome. Marfan's syndrome also overlaps with ectopia lentis and Beals' syndrome (congenital contractural arachnodactyly)—both of which are unassociated with aortic dilatation and rupture but are caused, respectively, by mutations in the genes encoding fibrillin I or II.

Clinical features

It seems likely, in retrospect, that the patient originally described by Marfan may have suffered from Beals' syndrome with congenital contractural arachnodactyly, elongated feet, and crumpled ears. Classical Marfan's syndrome arises from mutations in fibrillin I. Typically, there are long, elegant (though spidery) fingers and toes with dislocated lenses caused by a rupture of the ciliary zonules early in life. The syndrome is associated with mitral valve prolapse and aortic dilatation. Aortic disease is associated with dissection and rupture; rarely, dissection and rupture of the pulmonary artery occurs in Marfan's syndrome.

Detailed manifestations diagnostic of Marfan's syndrome (Table 4)

- *Skeletal*—anterior chest deformity, including asymmetrical pectus excavatum or carinatum (Fig. 6); dolichostenomelia, arachnodactyly, scoliosis lordosis, tall stature, narrow arched palate and dental crowding, protrusio acetabuli, congenital flexion contractures, hypermobility, ocular ectopia lentis (Fig. 7), flat cornea, elongated globe, retinal detachment, myopia;
- *Cardiovascular*—dilatation of the ascending aorta (Fig. 8), aortic dissection, aortic regurgitation, mitral regurgitation with prolapse, calcification of mitral annulus, abdominal aortic aneurysm, arrhythmia, endocarditis;
- *Pulmonary*—spontaneous pneumothorax, atypical bullae;
- *Skin*—abdominal striae and abdominal hernias, including diaphragmatic and umbilical hernias;
- *Central nervous system*—dural ectasia, meningocele, dilatation of cisterna magna, learning disability and hyperactivity.

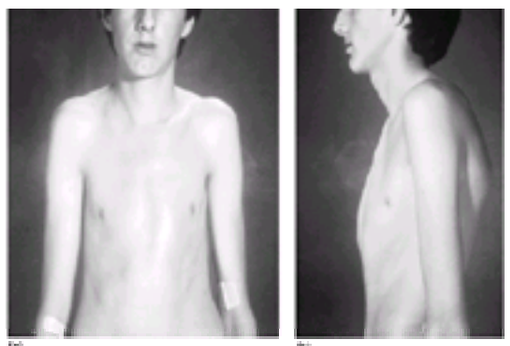

Fig. 6 Marfan's syndrome. (a) Frontal and (b) lateral views showing pectus deformity and mild kyphosis; the abnormal sternum and ribs are laterally compressed.

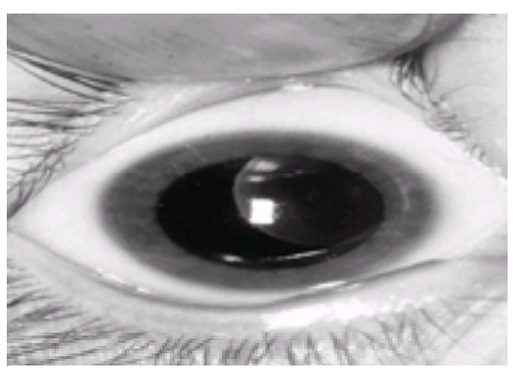

Fig. 7 Ectopia lentis in Marfan's syndrome. The lens is displaced upwards and medially; typically strong concave spectacle (aphakic) lenses are required to correct the high myopia.

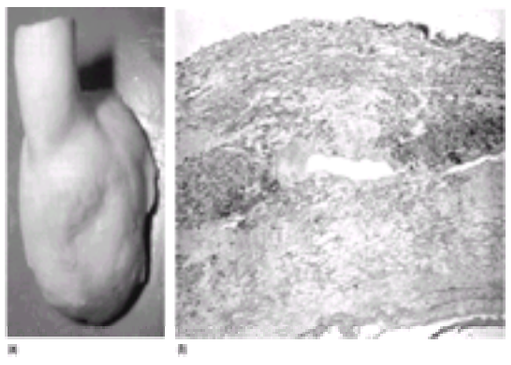

Fig. 8 Aortic disease in Marfan's syndrome. (a) Excised dilatated aortic root; (b) histological section of the aorta showing elastic degeneration of the aortic media.

The diagnosis of Marfan's syndrome may be confirmed by the involvement of at least two systems with at least one major manifestation as listed in Table 4; urinalysis (in patients not receiving vitamin B_6 supplements) should confirm the absence of homocystinuria.

The differential diagnosis includes: Stickler syndrome (vitreoretinal changes); osteoarthroses; mid-facial hypoplasia; Shprintzen–Goldberg syndrome (craniocarniosynostosis and retarded neurodevelopment, with Marfanoid features); and homocysteinuria. Since homocystinuria and Marfan's syndrome are distinct disorders—and because many patients with homocystinuria respond to specific therapies, for example pyridoxine supplements—clear distinction is necessary. Confusion between these two conditions is particularly likely in tall young patients with ectopia lentis and, except in unequivocal cases, patients with suspected Marfan's syndrome should always undergo appropriate biochemical testing for homocystinuria due to cystathionine b-synthetase deficiency or other causes.

Genetics

Marfan's syndrome is typically inherited as an autosomal dominant trait (see Fig. 5) and belongs to that group of genetic diseases in which a strong paternal-age effect occurs. The mean age of fathers of individuals who appear to harbour 'new' mutations is 5 to 10 years greater than average. Approximately one-third of all patients with Marfan's syndrome appear to be sporadic cases.

The gene responsible for Marfan's syndrome maps to chromosome 15. Mutations in the gene encoding fibrillin, a novel elastin-matrix protein, are responsible for Marfan's syndrome (Fig. 9). A second fibrillin gene maps to chromosome 5, mutations in which are responsible for Beals' syndrome (congenital contractural arachnodactyly that is not associated with defects in the ciliary zonules).

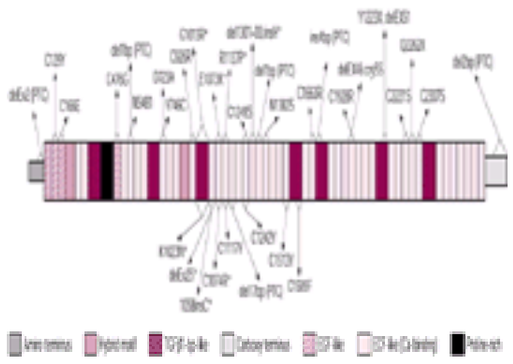

Fig. 9 Organization of the human fibrillin I gene (see text). Mutations associated with Marfan's syndrome are depicted.

The fibrillin I gene has a complex multi-exon organization and encodes calcium-binding, epidermal growth-factor-like and non-calcium binding, epidermal growth-factor regions. The gene encodes 65 exons encoding several conserved cysteines, common mutations of which are responsible for Marfan's syndrome; other fibrillin mutations have also been identified. Mutations in certain cysteines between exons 59 and 65 appear to be responsible for mild Marfan's syndrome that lacks aortic dilatation and is only associated with ectopia lentis. However, substitutions of cysteine codons elsewhere in the fibrillin gene may be responsible for more severe clinical phenotypes. Despite these findings, however, phenotype–genotype correlations in Marfan's syndrome are at present very loose. It is clear, however, that fibrillin mutations responsible for Marfan's syndrome are associated with disruption of the assembly of the fibrillin gene product and that dominant-negative mechanisms operate, as in the collagen defects. It seems likely that mutations of regions encoding epidermal growth-factor domains will disturb important homologous as well as heterologous, protein-matrix interactions of fibrillin.

The fibrillins are elastin-associated microfibrils, which assemble autonomously to form beaded microfilaments with ordered quasi-crystalline structures that can be studied by electron microscopy and other methods. Mutations that lead to the premature termination of the fibrillin I gene also appear to result in a milder phenotype than those mutations that cause small in-frame deletions or insertions. Exon-skipping events lead to the formation of in-frame mutant fibrillins lacking small protein sections which tend to interact normally with their wild-type congeners. In contrast, truncated mutants of fibrillin I typically disrupt the formation of the fibrillar structure. Most mutations in fibrillin I are private, and the presence of 65 exons in this large and complex genetic structure greatly impedes the molecular analysis of the fibrillin gene in patients with suspected Marfan's syndrome.

Treatment

The main causes of death in patients with Marfan's syndrome result from cardiovascular disease and complications elsewhere in the vascular system. Vigorous and regular surveillance is recommended with careful monitoring of aortic-root width and of the function of aortic and mitral valves by transthoracic echocardiography and periodic electrocardiography.

In patients with evidence of progressive aortic disease, including dilatation of the ascending aorta and valve ring, a Dacron graft, with or without an artificial or reconstituted aortic valve (the Bentall procedure), may be considered. There is also a strong case for joint management with experienced cardiac surgical colleagues in special clinics dedicated to the treatment of patients with Marfan's syndrome. Insertion of a Dacron graft to the aortic valvular ring, after excision of a terminally dilated aorta, requires reimplantation of the coronary arteries; in the best hands, the mortality rate of this procedure is less than 5 per cent, more than three-quarters of patients surviving 5 years. Scrupulous joint cardiac monitoring including echocardiography and magnetic resonance and CT imaging is recommended. A recent report by Gott and colleagues from Johns Hopkins Hospital, in the United States describes the results of aortic root replacement in 271 patients with Marfan's syndrome over the period 1976 to 2000. Most (> 85 per cent) patients underwent the Bentall procedure involving composite graft replacement of the aortic root. There were no 30-day deaths and more than 80 per cent of the operated patients were alive at the time this report was submitted for publication.

Patients with Marfan's syndrome benefit from the early introduction of adrenergic b-blockers to reduce the mean arterial pressure and pulse rate. Serial studies show that the mortality rate for aortic rupture in Marfan's syndrome is significantly reduced by the introduction of prophylactic b-blocking therapy; those so treated have a slower rate of aortic root dilatation. b-Adrenergic blockade should probably be introduced at an early age.

More than half of the patients with Marfan's syndrome suffer from dislocation of the optic lens—often in early childhood. Typically, the lens dislocates upwards. Thus surgical removal is indicated only if cataract or secondary glaucoma, due to the unusual lens position, intervenes. Upward dislocation may also lead to a greatly diminished visual acuity that cannot be corrected with spectacles; extraction of the lens is recommended under these circumstances.

Other complications of Marfan's syndrome including unstable joints, dislocation of the patella, progressive kyphoscoliosis, and recurrent pneumothoraces require surgical intervention. Clearly, many patients with Marfan's syndrome will require support from the psychological aspect and in the light of their diminished life expectancy. Women with Marfan's syndrome require counselling, not only about the genetic risk to their offspring but also because of the intrinsic risks of carrying a pregnancy to term. Re-evaluation of the heart during pregnancy may be recommended as a means of providing reassurance; the introduction of propanolol may be required during the course of labour and delivery to reduce cardiovascular stress.

Patients with Marfan's syndrome are at risk if they participate in competitive athletics. Indeed, the presence of Marfan's syndrome is a common cause of fatal aortic dissection and rupture in young adults. Those individuals with a clear diagnosis of the condition should be advised on an appropriate lifestyle and referred for genetic, as well as cardiological review, upon diagnosis. Several charities exist to support patients with Marfan's syndrome: the National Marfan Foundation in the United States as well as other lay groups in Europe and other countries provide critical information as well as psychological support for these patients and their families. Marfan's syndrome is a complex condition but, as a subject of intensive genetic and clinical study, benefits from active intervention with specific therapy (for instance, b-blockers) as well as rigorous monitoring, counselling, and supportive therapy.

Prognosis

Patients with Marfan's syndrome have a reduced life expectancy, principally as a result of the cardiovascular complications. Indeed, about 80 per cent of all deaths are due to aortic dilatation and its complications; the mean age of death in a series of 257 patients published in 1972 was 32 years. However, with the introduction of b-blocking agents and improvements in vascular and cardiac surgery, the prognosis has improved greatly and the early cohort studies were almost certainly subject to bias, since outcome was better in patients ascertained on the basis of family studies compared with those with sporadic disease.

Men with Marfan's syndrome have a lower survival rate than affected women—a conclusion based on several cohort studies. In a study of patients with Marfan's syndrome in Wales and Scotland between 1970 and 1990, the median survival for men was 53 years and for women, 72 years; the median age at death was 45 ± 16.5 years.

In a trial of 70 adolescent and adult patients with classical Marfan's syndrome, Shores and colleagues reported the results of treatment with propanolol. The treatment group received individualized doses of propanolol sufficient to induce negative inotrophy. When compared with the control group, those treated showed a reduced rate of aortic dilatation and improved survival—fewer of the treated patients reached clinical endpoints as determined by death, congestive cardiac failure, aortic regurgitation, aortic dissection, or the need for cardiac surgery.

Recently, several reports in the surgical literature describe the outcome of aortic root replacement in many hundreds of patients with Marfan disease: about one-third have had aortic dissection and in approximately one-half, aortic dilatation, 6.5 cm or less, was documented. Late operative death occurred in less than one-fifth of patients but subsequent series suggest that the prophylactic repair of smaller aortic aneurysms is preferable.

Pseudoxanthoma elasticum

Pseudoxanthoma elasticum (**PXE**) is an inherited defect of connective tissue with specific histological characteristics that reflect abnormally fragmented and calcified

elastic tissue. Fragmented elastin fibres occur principally in the skin, arteries, and retina where they are responsible for the main clinical effects of the disorder. Electron microscopy shows abnormalities of elastic fibres as well as associated collagen fibres. Pseudoxanthoma is now defined principally as a result of examination of the mutational spectrum of the responsible *ABCC6* gene and rigid clinical criteria. The systemic manifestations result principally from premature arterial stiffening and calcification; hypertension, thrombosis, and haemorrhage thus occur with widespread effects. Gastrointestinal bleeding, retinal disease causing visual loss, and stroke phenomena are among the most frequent complications of PXE.

Clinical genetics

The mode of transmission of PXE has been puzzling and controversial. This is reflected in the assignment of several Mendelian inheritance in man (**MIM**) catalogue numbers for autosomal dominant and recessive disease, respectively (MIM 177850, 177860; and 264800). In most (about 85 per cent) affected patients, the disease appears to occur sporadically without an overt family history. A minority of patients with PXE belong to pedigrees in which several siblings are affected in one generation; their parents (and other heterozygotes) often show generalized laxity of the joints, similar to Ehlers–Danlos syndrome type III. Transmission of authentic PXE through two or more generations consistent with an autosomal dominant pattern of inheritance occurs in no more than 5 per cent of cases.

Puzzling segregation patterns are frequent and obligate heterozygotes may show premature atherosclerosis. It is probable that homozygous or doubly heterozygous PXE is relatively common and that the carrier frequency may be as high as 0.5 to 1 per cent. Thus most families exhibiting anomalous inheritance of PXE may prove to be pseudodominant with matings between homozygotes or double heterozygotes and randomly distributed heterozygotes. In practical terms, the sibling recurrence risk for sporadic PXE lies between 1 in 800 and 1 in 4000; for authentic known autosomal recessive disease, this is 1 in 4. The risks of recurrent disease in the offspring of affected or potentially heterozygous children of known heterozygote parents are, providing no matings occur with consanguineous relatives, twice or the same as the odds for the sporadic cases—that is to say, no greater than 1 in 400. In due course, uncertainties about the transmission of PXE and genetic counselling will be clarified by mutation testing of the *ABCC6* gene responsible for this disease (see below).

Clinical features

Until recently there has been a lack of connection between the cutaneous and ophthalmological manifestations of PXE.

Although the typical orange–yellow xanthomatous appearance of flexual skin with underlying fragmentation of elastic had been long known, the association between this syndrome and the development of retinal angioid streaks due to elastic fragmentation of Bruch's membrane was not initially recognized. Pseudoxanthoma elasticum is associated with disease of both large and small arteries, as shown in the retina. Indeed, widespread involvement of the arterial media with attendant thrombotic complications is referable to many organs. Pseudoxanthoma is associated with bleeding from the gastrointestinal tract as well as stroke and, occasionally, lung disease.

Cutaneous

Typical changes of PXE in the skin develop in the face, neck, and flexual regions, including the axillas, antecubital fossas, inguinal folds, as well as the umbilicus. Similar changes may be observed around the mouth and nasolabial creases (Fig. 10 and Plate 2). The first changes are of thickening with a raised yellow discoloration between grooves, leading to plaque formation and an appearance resembling the skin of a plucked chicken, ' *peau d'orange*'. With time, the skin becomes inelastic and lax, thus causing considerable cosmetic alarm. In full-blown cases, an unattractive hound-dog appearance to the face and skin folds around the neck and groins develops, 'cutis laxa'. These changes may be aggravated by sun exposure and smoking.

Fig. 10 Skin lesions in pseudoxanthoma elasticum (PXE). (a) Typical flexural skin lesions of PXE of the lateral neck. (b) Widespread cutis laxa in PXE. (c) Mucosal infiltration of the lower lip in PXE. (d) Elastic van Gieson's stain of skin section showing mid-dermal elastic fragmentation and degeneration. (See also Plate 2).

Examination of the palate may show similar changes. Endoscopy of the stomach may reveal nodular submucosal lesions comparable to those present in peripheral skin.

Ophthalmic

The characteristic features of pseudoxanthoma elasticum in the eye are revealed by ophthalmoscopy (Fig. 11 and Plate 3). There may be only minor pigmentary changes, extending to disruption of the fibres in the preretinal membrane of Bruch causing angioid streaks. These retinal streaks vary in colour from dark red or maroon to black. Pressure on the eye, however, may discolour the streaks that underlie the retinal vessels and radiate from the optic disc.

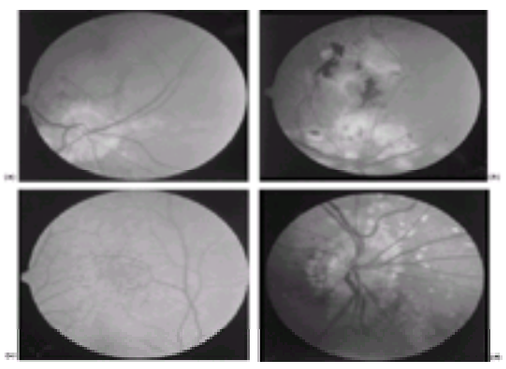

Fig. 11 Retinal changes in PXE. (a) Angioid streaks caused by fracture of the retroretinal Bruch's membrane—an early feature. (b) Macular haemorrhage with consequential choroideretinitis. (c) Specked *peau d'orange* mottling. (d) Salmon spotting and drusen. (See also Plate).

Angioid streaks occur in most patients beyond the age of 50 years but only in about one-third of patients with pseudoxanthoma ascertained in childhood.

In a cross-sectional study of 186 British patients with pseudoxanthoma elasticum conducted by the author in 1973, normal visual acuity was present in two-thirds: central visual loss occurred in 6 per cent; moderate visual impairment in 15 per cent; and mild impairment in 10 per cent. In contrast, fundoscopy was normal in less than 10 per cent of the patients: about one-third had angioid streaks and 13.5 per cent had degenerative maculopathy—a few patients had signs of haemorrhage. Myopia occurred in more than one-third of the patients and appeared to be at least three times more prevalent than in the general population. This implies that PXE

may predispose to myopia, perhaps by effects on corneal curvature, lens power, optical length, or vitreous composition.

Cardiovascular

In the cardiovascular system, the most striking abnormality of PXE is hypertension associated with an absence or weakness of peripheral arterial pulses. The author's cross-sectional survey of 186 British patients with PXE showed that 40 per cent had systemic hypertension: one in five of the patients experienced angina pectoris and one in ten had intermittent claudication; slightly more than 5 per cent had suffered transient ischaemic attacks; and 5 per cent had residual hemipareses following stroke episodes.

Occasionally, ischaemic features develop in the hands associated with resorption of digital tufts; these abnormalities are associated with a diminished Doppler pulse-wave velocity, reflecting a reduced amplitude of the systolic pulse wave. Other features include mitral valve prolapse and episodic and often torrential gastrointestinal haemorrhage usually from the stomach with or without a coincidental hiatal hernia or peptic ulcer. Bleeding may occur at other points including the renal, retinal, uterine, bladder, or subarachnoid spaces.

Many patients with PXE are hypertensive. This appears to increase the risk of bleeding, which may also be associated with premature arterial calcification in peripheral arteries as well as coronary vessels. Women with pseudoxanthoma elasticum often develop severe hypertension during pregnancy and complain of rapid progression of skin changes. Joint manifestations are not a feature of pseudoxanthoma elasticum.

Diagnosis

Several clinical criteria for the diagnosis of PXE have been defined by different authorities. The major and minor criteria preferred by this author are set out in Table 5; any two major, or one major and two minor, criteria are sufficient to diagnose the disease. The most important criteria for establishing the diagnosis are the presence of the appropriate skin eruption with histological evidence of elastic degeneration and calcification as determined by tissue biopsy.

Differential diagnosis

Angioid streaks may be seen in Paget's disease and sickle cell anaemia but are rarely as florid as those occurring in pedigrees affected by pseudoxanthoma elasticum. Rarely, diabetic retinopathy may be associated with angioid streaks. Angioid streaks have also been reported in patients with neurofibromatosis and tuberous sclerosis. Cutaneous manifestations of pseudoxanthoma elasticum may resemble those of extreme solar injury to skin associated with ageing. Characteristically, long-term penicillamine therapy leads to a syndrome that is a close phenocopy of pseudoxanthoma elasticum: elastosis serpingiosa perforans is frequently associated with pseudoxanthoma elasticum-like features. Pseudoxanthoma elasticum may also be considered in patients with Ehlers–Danlos syndrome in which cutis laxa is the sole manifestation.

Pathology

Pseudoxanthoma elasticum is diagnosed principally because of the occurrence of the constellation of clinical features, the family history, and a skin biopsy that reveals a characteristic fragmentation and disruption as well as calcification of the elastic fibres of the middle and deep zones of the corium. The use of von Kossa's stain, which identifies carbonate and phosphate complexes of calcium, together with Van Gieson's stain for elastic fibres is usually diagnostic; electron microscopy usually reveals electron-dense deposits throughout elastin fibres in the skin with a central core of mineral.

Molecular genetics

The gene responsible for pseudoxanthoma elasticum has been mapped in some affected pedigrees to a single locus on the short arm of chromosome 16 (16p31.1) (Fig. 12). Unexpectedly, the gene proved to be a multidrug-resistance and membrane ion transporter rather than an integral structural matrix protein. It spans 31 exons and encodes at the ATP-binding cassette and is a member of the cystic fibrosis-transmembrane regulator, *CFTR*, gene family.

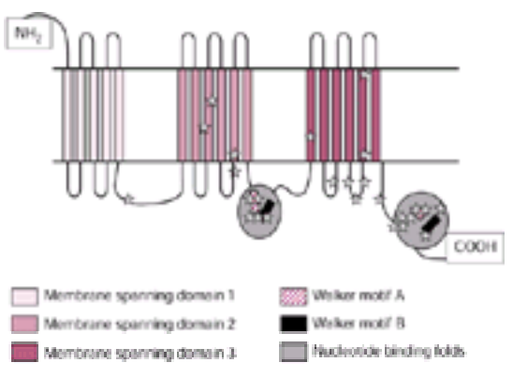

Fig. 12 Organization of the human PXE gene, *ABCC6*, a member of the ABC transmembrane ion transporter family. There are three membrane-spanning domains, each of which contains nucleotide-binding folds.

As discussed above, many patients with pseudoxanthoma elasticum belong to healthy families and are the only member with the condition, thus having no overt family history of the disease. Of the remaining 15 per cent of PXE patients, the most common family pattern is of several affected siblings in one generation showing diverse clinical expression.

Some families appear to show either autosomal recessive or dominant patterns of transmission. The relationship between these defects and mutations in the *ABCC6* gene implicated in pseudoxanthoma is unclear. Most patients appear to be homozygous for a defect in the *PXE* gene and some shared mutations appear to segregate with homozygosity for stop codons. At present, it is unclear what proportion of mutations in the *ABCC6* gene account for pseudoxanthoma elasticum and what proportion of patients are homozygous or compound heterozygotes at this locus. Despite the ambiguities, however, molecular analysis of the *ABCC6* gene provides an opportunity for systematic testing and clarification of the risk of transmission of the disease in affected pedigrees.

Disease frequency

At present, the population frequencies of pseudoxanthoma elasticum and mutations in the *ABCC6* gene are unknown, but in the United Kingdom at least 400 families are members of a PXE self-help patient association. Systematic analysis will be needed to provide additional information for diagnosis and genetic counselling in the population affected by PXE at large. So far, no clear correlation between the genotype and phenotype of the complex pseudoxanthoma syndromes has been possible.

Treatment and management

With time, patients with pseudoxanthoma elasticum are prone to premature-ageing phenomena in their skin and cosmetic embarrassment as a result. Patients with PXE are likely to suffer the results of vascular disease with stroke complicating hypertension and subretinal haemorrhages with visual loss; they are at increasing risk from severe gastrointestinal bleeding. Women with pseudoxanthoma elasticum should be advised to limit their family size. Excess exposure to ultraviolet light should be avoided as far as possible and sunscreen lotions used when this is not possible. Regular light exercise and avoidance of cigarette smoking are simple measures that also likely to be beneficial.

Although the skin and vascular lesions of pseudoxanthoma are associated with calcification, there is no evidence that calcium restriction influences the development of the disease. Nonetheless, some authorities recommend restricting calcium intake without evidence that this impedes the progression of the disorder. Because of

their severe systemic arterial disease, patients with pseudoxanthoma elasticum are advised to undergo regular monitoring of their vascular integrity and blood pressure. The rapid onset of severe systemic hypertension that is refractory to treatment may be due to unilateral renal artery stenosis—a well-described abnormality in PXE.

Contact sports, including boxing, and arduous exercise such as cross-country running should be avoided. Regular monitoring by a cardiologist and ophthalmologist may be beneficial, in that the occurrence of new vessel formation in relation to angioid streaks can be arrested by ocular laser therapy to prevent or diminish the risk of retinal haemorrhage. Similarly, regular blood pressure monitoring with the prompt use of b-blockers for hypertension where possible may delay the onset of peripheral vascular insufficiency and coronary heart disease.

Prompt treatment of systemic hyperlipidaemia, which may independently complicate the arteriopathy of PXE, is indicated to arrest arterial narrowing and prevent thrombosis. Antiplatelet drugs such as aspirin are contraindicated because of the increased risk of visual loss due to retinal bleeding and of gastrointestinal haemorrhage. Coronary bypass surgery is as successful and no riskier than for the general population; there is little evidence to judge the outcome of vascular surgical procedures that may be indicated for stenoses of carotid or other major peripheral arteries.

Patients with pseudoxanthoma elasticum may benefit from plastic surgery to remove redundant skin around the neck and groins, abdomen and breasts. This is particularly applicable to women who become embarrassed by rapid cutaneous changes after pregnancy or the menopause. Keloid formation may complicate such cosmetic surgery and it is advisable that those who operate are apprised of this risk in PXE.

Special problems in pregnancy

Pregnancy usually proceeds with only minimal difficulties; the theoretical risks of recurrent gastrointestinal haemorrhage and perineal tearing at delivery result only rarely in adverse outcomes in women with pseudoxanthoma elasticum. Maternal PXE does not appear to increase the risk of fetal abnormalities. However, scrupulous monitoring of systemic arterial blood pressure is recommended in pregnant patients with this disorder and during the peripartum period.

Prognosis

In some patients with pseudoxanthoma elasticum, premature death results from vascular disease which may cause critical occlusion of the arterial supply to essential organs or fatal bleeding. Death from a recurrent massive gastrointestinal haemorrhage was recorded in a 13 year-old patient and severe bleeding due to PXE has been reported in many younger children. McKusick has shown that many patients may live beyond 70 years and die of conditions unrelated to their connective tissue disorder; his study of 52 patients with PXE from the case records of the Johns Hopkins Hospital in the early 1970s showed that the median survival of this selected cohort was about 46 years. All the patients in this early group were dead by the age of 76 years.

Further reading

Barabas AP (1967). Heterogeneity of the Ehlers–Danlos syndrome: description of three clinical types and hypothesis to explain the basic defects. *British Medical Journal* **2**, 612–13.

Beighton P (1968). Lethal complications of the Ehlers–Danlos syndrome. *British Medical Journal* **2**, 656–60.

Beighton P (1993). The Ehlers–Danlos syndromes. In: Beighton P, ed. *McKusick's heritable disorders of connective tissue*, 5th edn, pp 189–252. Mosby Year-Book, St. Louis.

Beighton P, *et al*. (1992). Molecular nosology of heritable disorders of connective tissue. *American Journal of Medical Genetics* **42**, 431–48.

Beighton P, *et al*. (1998). International nosology of heritable disorders of connective tissue. *American Journal of Medical Genetics* **29**, 581–94.

Beighton P, *et al.* (1999). Ehlers–Danlos syndrome: revised nosology, Villefrache, 1997. *American Journal of Medical Genetics* **77**, 31–7.

Bergen AAB, *et al*. (2000). Mutations in a gene encoding an ABC transporter cause pseudoxanthoma elasticum. *Nature Genetics* **25**, 223–7.

Birk DE, *et al*. (1990). Collagen fibrillogenesis *in vitro*. Interaction of types I and V collagen regulates fibril diameter. *Journal of Cell Science* **95**, 649–57.

Buntinx IM, *et al*. (1991). Neonatal Marfan syndrome with congenital arachnodactyly flexion contractures and severe cardiac valve insufficiency. *Journal of Medical Genetics* **28**, 267–73.

Burrows NP, *et al.* (1996). The gene encoding collagen alpha 1 type V (COL5A1) to type II Ehlers–Danlos type I/II. *Journal of Investigative Dermatology* **106**, 1273–6.

Byers PH, *et al*. (1979). Clinical and ultrastructural integrity of type IV Ehlers–Danlos syndrome. *Human Genetics* **47**, 141–50.

Carlborg U, *et al*. (1959). Vascular studies in pseudoxanthoma elasticum. *Acta Medica Scandinavica* **350**, 1–17.

Collod-Beraut G, *et al*. (1998). Marfan database (third edition): new mutations and new routines for the software. *Nucleic Acids Research* **26**, 229–33.

De Paepe A, *et al*. (1996). Revised diagnostic criteria for the Marfan syndrome. *American Journal of Medical Genetics* **62**, 417–26.

Elejalde BR, *et al*. (1984). Manifestations of pseudoxanthoma elasticum during pregnancy: a case report and review of the literature. *American Journal of Medical Genetics* **18**, 755–62.

Fattori R, *et al*. (1999). Importance of dural ectasia in phenotypic assessment of Marfan's syndrome. *Lancet* **354**, 910–13.

Godfrey M (1993). The Marfan syndrome. In: Beighton P, ed. *McKusick's heritable disorders of connective tissue*, 5th edn, pp 51–135. Mosby Year Book, St Louis.

Gott VL (2002). Aortic root replacement in 271 Marfan patients: a 24-year experience. *Annals Thoracic Surgery* **73**, 438–43.

Grahame R (2000). Heritable disorders of connective tissue. *Bailliere's Clinical Rheumatology* **14**, 345–61.

Gray JR, *et al*. (1998). Life expectancy in British Marfan syndrome populations. *Clinical Genetics* **54**, 124–8.

James AE, *et al.* (1969). Roentgen findings in pseudoxanthoma elasticum (PXE). *American Journal of Radiology* **106**, 632–47.

Kielty CM and Shuttleworth AC (1994). Abnormal fibril assembly by dermal fibroblasts from two patients with the Marfan syndrome. *Journal of Cell Biology* **124**, 997–1004.

Lee B, Godfrey M, Vitale E (1991). Linkage of the Marfan syndrome and a phenotypically related disorder to two different fibrillin genes. *Nature* **352**, 330–4.

Le Saux O, *et al*. (2001). A spectrum of ABCC6 mutations is responsible for pseudoxanthoma elasticum. *American Journal of Human Genetics* **69**, 749–64.

McKusick VA (1956) and (1972). *Heritable disorders of connective tissue*, 1st and 4th editions. Charles Thomas and Mosby Year Book, St Louis.

Neidner KH (1988). Pseudoxanthoma elasticum. *Clinics in Dermatology* **6**, 1–157.

Nicholls AC, *et al*. (1996). An exon-skipping mutation of the type V collagen gene (COL 5A1) in Ehlers–Danlos syndrome. *Journal of Medical Genetics* **33**, 940–6.

Palz M, *et al*. (2000). Clustering of mutations associated with mild Marfan-like phenotypes in the 3-prime region of FBN1 suggests a potential genotype-phenotype correlation. *American Journal of Medical Genetics* **91**, 212–21.

Pepin M, *et al.* (2000). Clinical and genetic features of Ehlers–Danlos syndrome type IV, the vascular type. *New England Journal of Medicine* **342**, 673–80.

Pope FM (1975). Historical evidence for the genetic heterogeneity of pseudoxanthoma elasticum. Brit. Journal Dermatol. **92**: 493–509.

Pope FM (1997). Molecular abnormalities of collagen. In: Maddison PJ, *et al.*, eds. *Oxford textbook of rheumatology*, 2nd edn, pp. 353–404. Oxford University Press, Oxford.

Pope FM (1998). Components of the dermis in anatomy and organization of skin. In: Champion RH, *et al*., eds. *Textbook of dermatology*, 6th edn, pp 59–92. Blackwell Science, Oxford.

Pope FM and Burrows NP (1997). Ehlers–Danlos syndrome has varied molecular mechanisms. *Journal of Medical Genetics* **34**, 400–10.

Pope FM, *et al*. (1975). Patients with Ehlers–Danlos syndrome type IV lack type III collagen. *Proceedings National Academy of Sciences, USA* **72**, 1314–16.

Pope FM, *et al*. (1996). CoL3A1 mutations cause variable clinical phenotypes, including acrogeria and vascular rupture. *British Journal of Dermatology* **135**, 1617–20.

Reeve EB, *et al*. (1979). Development and calcification of skin lesions in thirty-nine patients with pseudoxanthoma elasticum. *Clinical and Experimental Dermatology* **4**, 291–301.

Scheie HG and Hogan TT (1957). Angioid streaks and generalized arterial disease. *Archives of Ophthalmology* **56**, 855–68.

Shores J, *et al.* (1994). Progression of aortic dilatation and the benefit of long-term beta-adrenergic blockade in Marfan's syndrome. *New England Journal of Medicine* **330**, 1335–41.

Steinmann B, *et al*. (1979). Evidence for a structural mutation of procollagen type I in a patient with Ehlers–Danlos syndrome type VII. *European Journal of Paediatrics* **130**, 203–5.

Struch B, *et al*. (1997). Mapping of both autosomal recessive and dominant variants of pseudoxanthomata elasticum to chromosome 16p13.1. *Human Molecular Genetics* **6**, 1823–8.

Viljoen DL (1993). Pseudoxanthoma elasticum. In: Beighton P, ed. *McKusick's heritable disorders of connective tissue*, 5th edn, pp 335–65. Mosby Year Book, St Louis.

Viljoen DL, Beatty S, Beighton P (1987). The obstetric and gynaecological implications of pseudoxanthoma elasticum. *British Journal of Obstetrics and Gynaecology* **94**, 884–8.

19.3 Osteomyelitis

Anthony R. Berendt and Martin McNally

Introduction

Osteomyelitis is an ancient disease with a formidable reputation for persistence and relapse. It has been diagnosed in human fossil remains from the late Neolithic and was described by many classical writers including Hippocrates. The term indicates infection of the marrow (the suffix 'myelitis'), but will be used here to indicate any infection of bone, even if confined to the cortex (sometimes called 'osteitis').

Aetiology

The pathogens causing osteomyelitis are dominated by *Staphylococcus aureus*, but there are many other known causes as shown in Fig. 1.

Fig. 1 Microbiological causes and contexts in pyogenic arthritis and osteomyelitis.

Epidemiology

Classical acute haematogenous osteomyelitis has its peak incidence in childhood. Males are more commonly affected than females. In children, there appears to be a greater incidence in the Southern hemisphere and among certain racial groups (for example, aboriginal Australians), with rates varying from 10 to 100:100 000/year. Socioeconomic factors may contribute to this. Chronic osteomyelitis is such a diverse disease that an overall incidence and prevalence rate is not available, but incidence rises with age due to numerous causes including diabetes, peripheral vascular disease, infirmity, and ulceration.

Pathogenesis and pathophysiology

The critical step in pathogenesis is the access of bacteria to the bone. This may occur from a contiguous focus such as chronic ulceration, surgery, trauma, or soft tissue infection. Alternatively, the route may be haematogenous, with bacteria reaching bone via the circulation. The exact mechanism by which this occurs is uncertain. It is believed that the tortuous capillary loops in the metaphysis of the long bones, a favoured site for haematogenous osteomyelitis, are particularly vulnerable to thrombosis, leading to bacterial seeding. This is supported by a history of recent blunt trauma to the affected part in some 30 per cent of cases, and by observations that in most animal models it is necessary to injure bone to infect it. Even minor bone and soft tissue trauma exposes components of blood clot, extracellular matrix, and bone matrix to the bloodstream. Many pathogens, notably *S. aureus*, can adhere to such host proteins through specific receptors, and hence to tissues and cells, including endothelial cells and osteocytes.

An acute inflammatory response is elicited once bacteria gain access to bone and begin to multiply. This causes oedema within bone and soft tissue, and the procoagulant effect of inflammation may also cause thrombosis in vessels. The result can be bone infarction, possibly contributed to by bacterial toxins.

As infection progresses, it propagates within the bone marrow, and through the cortical bone via the Haversian canals. Pus may form within cancellous bone and beneath the periosteum (see Fig. 2 for a schematic diagram). It may break into the soft tissues and even extend to the surface as a sinus tract. Subperiosteal pus under pressure will strip off the overlying periosteum, tracking along the length of the bone and around its circumference. The vascular consequences of this are critical to the evolution of the disease, since the outer aspect of the cortical bone is vascularized by the periosteum, the inner by the endosteal circulation. If the endosteal blood supply is already compromised, periosteal stripping causes bone death. Thus, large pieces of bone, segments, or even whole long bones can die.

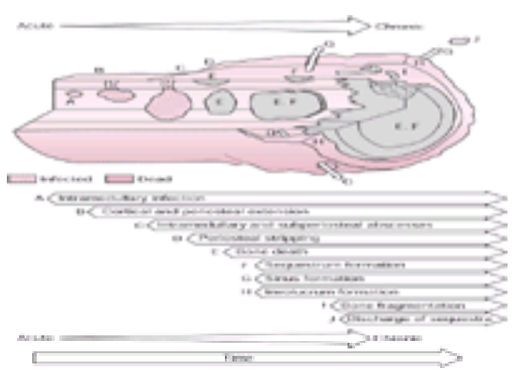

Fig. 2 Schematic diagram showing the evolution from acute to chronic osteomyelitis, with progressive necrosis, sequestration, and sinus formation.

Dead bone can potentially be revascularized and remodelled, but only if it remains in physical continuity with living bone. However, the action of bone-resorbing cells,

recruited and activated by inflammation and some bacterial products, is frequently to separate dead from healthy bone. This produces a detached piece of dead bone called a sequestrum. Small sequestra can be extruded from sinuses or wounds and the episode of osteomyelitis may arrest spontaneously; larger sequestra result in continuing infection and inflammation. Over time more bone tends to be involved, sometimes resulting in new sinuses, with extension into soft tissues and contiguous joints. As bone is resorbed and killed, the loss of structure may lead to pathological fracture.

Chronicity and relapse result both from this host response and from features of bacterial physiology. The body cannot mount effective inflammatory responses in dead tissue or chronic abscesses. Bacteria adhere to the inanimate surfaces of dead bone and, as in implant-related infections, form complex structures in which they are enmeshed in an antiphagocytic polysaccharide matrix (the whole known as a biofilm). Their growth state alters within this, rendering them phenotypically resistant to almost all antibiotics. They may even be able to persist as metabolically crippled forms called small-colony variants: these can exist within cells and are also resistant to many antibiotics that would otherwise kill wild-type organisms.

If periosteum has been stripped and remains viable, it produces new bone called the involucrum. This may develop circumferentially, producing a shell of living bone around the dead segment, thus preserving mechanical strength. Defects in the involucrum, through which sinuses communicate with sequestra, are called cloacae.

Variations on this theme occur when flat bones or those of the spine are involved in haematogenous infection. In discitis and vertebral osteomyelitis, infection of the disc space is rapidly followed by involvement of the two adjacent vertebral bodies. The infection may arrest as disc material is replaced by granulation tissue, eventually leading to fusion of the two involved vertebral bodies. In flat bones such as the pelvis or the skull, infection can spread very rapidly in the cancellous bone between the two tables, before exciting a periosteal reaction.

The 'inside-to-out' nature of haematogenous osteomyelitis is in distinction to the 'outside-to-in' nature of contiguous focus osteomyelitis. In this case, periosteum is destroyed as part of the same process that has destroyed the overlying soft tissues. Cortical bone is killed and infection can enter the medullary cavity, thereafter extending as for haematogenous disease. Sequestra may separate and be discharged, but the adverse biological factors that led to the initial soft tissue loss may impair subsequent healing and permit further bone infection to occur.

Clinical features

Acute osteomyelitis presents as a rapid onset of pain and loss of function in the affected limb, usually accompanied by high fever and malaise. It predominantly affects the metaphyses adjacent to the large weight-bearing joints. Prostration, sweating, rigors, and vomiting from accompanying bacteraemia (in 50 per cent of cases) may also be present. In neonates and infants, extension from the medullary cavity of the metaphysis through the cortex leads into the joint space, since the joint capsule extends beyond the growth plate. Thus in this age group an acute septic arthritis can be an early complication, or a presenting feature, of an acute osteomyelitis (see Chapter 18.7.1). In older children, the joint capsule is much tougher and inserts at the growth plate, the cartilage of which forms a barrier to the passage of infection from the metaphysis to the epiphysis and the joint.

Chronic osteomyelitis presents more variably. Pain is the rule, unless there is underlying neuropathy, and there may be severe disability in the context of an ununited fracture or when the spine is involved. Wound or sinus tract drainage is usually present when osteomyelitis complicates ulceration, instrumentation, or other surgery. Bone may be visible, palpable with a gloved finger or located with a sterile metal probe in the base of an ulcer or sinus. There may be evidence of soft tissue swelling or induration, and bony tenderness on palpation or percussion. Some patients experience repeated flares of fever and acute illness due to inadequate drainage of deep pus or rapid extension into previously uninvolved soft tissue or bone. Minor ill health is common, manifesting as weight or appetite loss, general malaise, or poor glycaemic control in diabetics. This is often only noticeable in retrospect, once the infection has been treated.

Patients with vertebral osteomyelitis may present with bacteraemia and acute back pain (raising the possibility of spinal epidural abscess and the need for urgent diagnosis and treatment), but more often they present with chronic back pain and non-specific illness. Differential diagnoses of degenerative back pain, osteoporotic fracture, metastatic disease, and myeloma should be considered. The presence of severe back pain at rest, often of a deep and unremitting character that patients can distinguish from previous back pains, and of night pain, should prompt consideration of the diagnosis. Spinal tenderness is an unreliable sign. Deformity and the development of neurological signs are late features.

Special forms of osteomyelitis include chronic multifocal osteomyelitis (this presents with pain but, despite radiological and histological features of osteomyelitis, is culture-negative), unifocal osteomyelitis with a similar behaviour, and Brodie's abscess (a well-defined chronic abscess in bone with a very indolent presentation). The interested reader is referred to specialist texts for details.

Laboratory diagnosis

The white-cell count, erythrocyte sedimentation rate (**ESR**), and C-reactive protein (**CRP**), though generally elevated in acute infection and flares of chronic disease, are non-specific and occasionally normal in chronic disease. It is helpful to see elevated inflammatory markers fall after treatment, but this may take several weeks. The alkaline phosphatase level is of no value, being neither sensitive nor specific for bone infection. Blood cultures are essential in acute infection, when they may be the only means of obtaining a microbiological diagnosis. Serological tests are useful for the diagnosis of syphilis, yaws, brucellosis, and occasionally bartonellosis.

Plain radiography of chronic osteomyelitis typically shows patchy osteopenia or frank bone destruction, loss of definition of the cortex, areas of sclerosis, or periosteal reaction with new bone formation. These changes take many weeks to develop fully. In acute infection, the earliest changes visible on plain radiography are soft tissue swelling (minimum 2–3 days), followed by periosteal reaction (7 days), and last, bone destruction (10 days). If radiographs are abnormal, the changes need to be distinguished from those of a tumour, trauma, or degenerative bone disease. Repeat imaging at an interval of 2 to 4 weeks can sometimes help as osteomyelitis is usually an aggressive process. For more rapid clarification of diagnosis, however, specialized imaging is needed.

Ultrasound can identify subperiosteal collections and soft tissue abscesses, and demonstrate sinuses. Computed tomography (**CT**) scanning may be able to identify cortical erosion that has been missed on plain films and can demonstrate sequestra within bone. Reformatted images make it possible to produce saggital or coronal images (for example, to view vertebral body endplates) and three-dimensional images for surgical planning. Soft tissue collections are easily identified. Other than a lack of sensitivity early in the disease, the principal pitfalls of CT scanning are the radiation dose, its lack of ability to determine the extent or activity of infection, and its sensitivity to image degradation from orthopaedic metalware.

Isotope scanning is widely used, but there is a lack of consensus on the utility of various tests. Conventional, three-phase, technetium bone scans are sensitive but non-specific. Specificity may be increased by the addition of indium-labelled leucocyte scanning. Other reagents include labelled immunoglobulins, anti-leucocyte monoclonal antibodies, and even radiolabelled antibiotics, but the performance of these tests has not yet been rigorously evaluated.

Magnetic resonance imaging (**MRI**) is the standard and best method for diagnostic imaging of osteomyelitis (Fig. 3 and Fig. 4). It can detect intra- and extraosseous oedema, abscesses, dead bone, and sinus tracts. It can distinguish active from inactive infection. Other than cost (rapidly falling) and availability (rising), the main problem with MRI is its extreme sensitivity to physiological changes that may persist long after surgery or treatment and to metal artefact from orthopaedic implants (and even to microscopic metallosis when they have been removed).

Fig. 3 Acute osteomyelitis of the femur in a child. (a) The plain radiograph, after one day of illness, is normal. *S. aureus* was isolated from blood cultures. (b) MRI

scan (STIR sequence) of the same patient at day 2. There is marked soft tissue and intraosseous oedema (high signal). Subperiosteal abscesses can clearly be seen as linear areas of high signal just outside the cortex, tracking proximally up the femur from the metaphysis.

Fig. 4 (a) A plain radiograph showing chronic osteomyelitis of the proximal tibia in an adult. There is patchy sclerosis and lysis. (b) MRI of the same patient (*T*1 sequence) showing oedema (low signal) in the area corresponding to plain film changes, but also an additional, distal, intramedullary satellite lesion.

The microbiological standard for the diagnosis of osteomyelitis is the growth of bacteria from samples of bone, taken with precautions to prevent contamination from superficial flora. Pus or soft tissue associated with infected bone may be acceptable, but sinus tract or wound swab cultures are not. The bacteria isolated from wounds are poorly predictive of the deep flora because of asymptomatic colonization. Cultures of this kind should be reserved for detecting multi-resistant organisms (such as methicillin-resistant *S. aureus* (**MRSA**)) for infection control purposes. Fluid for microscopy and culture can be aspirated from periosteal or subperiosteal abscesses. In infants, needle aspiration of bone itself is safe and well tolerated if performed by someone experienced in the technique. Bone biopsy can be performed surgically or percutaneously (by needle biopsy). In neuropathic ulcers, bone can be obtained by curettage. The laboratory must be made aware of the importance and nature of any specimen sent so that it can be appropriately processed and interpreted.

Bone histology is also an important diagnostic test: the presence of inflammatory cells, dead bone, and active bone remodelling are hallmarks of infection. They may provide the only confirmation of infection in cases where the culture results are unhelpful.

Treatment

Acute osteomyelitis

Acute osteomyelitis may respond to antibiotics alone, and with better outcomes, if treated before the onset of bone death or abscess formation. It is therefore an orthopaedic emergency. Treatment should be initiated on the basis of the clinical diagnosis, with investigations used to confirm the diagnosis once treatment has begun. Following blood cultures, high-dose intravenous antibiotics effective against *S. aureus*, b-haemolytic streptococci, and aerobic Gram-negative rods should be given. Appropriate regimens include a cephalosporin or the combination of an antistaphylococcal penicillin and gentamicin. Vancomycin and gentamicin will be necessary if the patient has risk factors for infection with MRSA. Antibiotics can be modified based on culture results. For patient comfort, the limb should be splinted and elevated, and analgesia given.

Surgery is indicated if abscesses are present: these must be opened, and the bone is usually drilled to allow free drainage of contained pus. The soft tissues are protected to avoid further devascularization and consequent bone death.

The necessary duration of antibiotic therapy is unclear. Treatment for less than 4 weeks is associated with higher rates of relapse.

In children, oral therapy can be considered when:

- the patient is afebrile after an initial 48 to 72 h of intravenous treatment;
- there is no evidence of abscess formation, metastatic infection, or bacteraemia;
- there is no suspicion from the history or imaging that infection has been prolonged or is associated with dead bone;
- the organism is sensitive to reliably bioavailable oral antibiotics; and
- compliance with therapy can be assured.

Less information is available for adults. The greatly lower rates of bone blood flow and turnover make revascularization and absorption of necrotic bone, and delivery of antibiotics and white cells, less certain. For these reasons it is common, but not universal, to treat adult acute osteomyelitis with intravenous therapy for periods of at least 4 weeks. Certain drugs, notably clindamycin and ciprofloxacin, are highly bioavailable and have proved useful in the treatment of osteomyelitis. There are no randomized studies to inform decisions about the requisite total duration or duration of intravenous therapy.

Chronic osteomyelitis

To achieve long-term arrest of infection, the management of chronic osteomyelitis usually requires multiple, co-ordinated inputs. The aims of treatment are to:

- remove dead bone and soft tissue;
- drain abscesses;
- eliminate cavities (which act as surgical 'dead spaces');
- ensure skeletal stability;
- restore soft tissue cover (if necessary using plastic surgery);
- define pathogens from high-quality specimens and administer appropriate antibiotics;
- correct adverse local and systemic host factors;
- support the patient physically and psychologically;
- reconstruct the skeleton if need be; and
- rehabilitate the patient.

Surgery

Detailed consideration of surgical methods is beyond the scope of this book, but major surgical advances include the use of free-tissue transfer and bone transport techniques to close very large bony and soft tissue defects. These permit much more radical approaches to the resection of diseased and dead tissues. In this way, surgery can potentially convert chronic infected wounds with dead bone and soft tissue into contaminated wounds of living bone with healthy soft tissue cover.

Antibiotics

These play an important role in increasing success after surgery, though the 'added value' they confer is uncertain and may depend on the extent of surgical resection. Some conditions often respond well without surgery including:

- discitis and vertebral osteomyelitis, with surgery reserved for abscess formation, progressive pain or deformity, instability, spinal cord compression, or persistent

sepsis;
- tuberculous osteomyelitis, reserving surgery for mechanical complications, pain, or persistent infection;
- osteomyelitis of small bones such as the phalanges. In the treatment of a diabetic patient with foot osteomyelitis, accompanied only by limited podiatric debridement of bone, some authorities quote that chronic osteomyelitis can be arrested in 70 to 80 per cent of cases, but recurrences are common.

Antibiotics may also help when the patient refuses surgery, when there is no clearly definable surgical 'target', or when the risks and consequences of surgical resection would be worse than the disease itself.

The choice of antibiotics should be guided by the culture results. Intravenous therapy may need to be prolonged (for up to 6 weeks) where there is thought to be a risk of unreliable compliance, absorption, or efficacy with oral therapy. If properly supervised, many patients can be discharged from hospital while remaining on intravenous therapy. Periods of total antibiotic treatment vary from weeks to many months, but there is a growing trend to shorten the duration of treatment when an expert surgeon has achieved a radical surgical clearance, provided that local and systemic host factors are favourable. Antibiotics can also be delivered locally, by implanting antibiotic-loaded bone cement or collagen fleece at the time of surgery. The relative efficacies of intravenous, oral, or local antibiotics have received little attention and treatment protocols vary widely.

Adjunctive treatment

It is important to correct other host factors that may affect wound and bone healing. These include ischaemia (which may need intervention), anaemia, diabetes, hypoxia from respiratory or cardiac failure, peripheral oedema, poor nutrition, and smoking. Where neuropathy has contributed to ulceration, appropriate pressure relief is essential for healing and for secondary prevention. This must be continued indefinitely through the provision of specialist footwear, cushions, or beds. The patient must be taught about neuropathy and trained in methods to prevent further ulceration. Hyperbaric oxygen therapy has been widely employed with anecdotal success, but its efficacy and precise role are unclear and definitive randomized trials are awaited.

Prognosis

Chronic osteomyelitis can be arrested in 80 to 90 per cent of cases, usually when surgery has been combined with antibiotic treatment. Though most common within the first year, delayed recurrence is well recognized, and relapse can occur up to 50 years after an initial infection has apparently been treated successfully. This poses major difficulties for the design of trials on new treatment, as extended follow-up is needed to make definitive statements about success or failure. Chronic multifocal osteomyelitis has a good prognosis, usually self-arresting, albeit after some years. Longstanding active disease may be associated with the eventual development of squamous metaplasia or carcinoma in a sinus, and rarely with the deposition of amyloid.

Prevention and control

There are no proven means of preventing haematogenous osteomyelitis, but prompt treatment can prevent chronicity. Contiguous osteomyelitis can be prevented by the appropriate management of open fractures, of infective foci close to bone, and of chronic wounds whenever these are close to a bone or joint. Pressure-area care in immobile patients and of a diabetic patient's foot can prevent ulceration and subsequent osteomyelitis.

Occupational, quality of life, and psychosocial aspects

Pain, chronic sepsis, and physical disability have a significant impact on quality of life. Psychological well being is further affected by issues common to all chronic diseases, together with anxiety and depression over:

- risks of death, paralysis (e.g. in spinal infection), and limb loss;
- stigmatizing effects of chronic discharging wounds; and
- feelings of anger or failure where infection has resulted from an accident or surgery.

Areas of uncertainty needing further research

There is a need to define the true health, psychosocial, and economic burden of osteomyelitis. More information is needed on pathogenetic mechanisms and on molecular diagnosis. The optimal duration and route of administration of antibiotics and their place alongside differing forms of surgery needs clarification, as does the role of adjunctive medical treatments such as hyperbaric oxygen.

Further reading

Carr AJ, *et al*. (1993). Chronic multifocal osteomyelitis. *Journal of Bone and Joint Surgery* **75B,** 582–91. [Series of this rare, fascinating, and poorly understood condition.]

Case records of the Massachusetts General Hospital (1993). *New England Journal of Medicine* **328,** 422–8. [Case of late relapsing chronic osteomyelitis with much interesting historical and pathological discussion.]

Cierny, III G, Mader JT (1984). Adult chronic osteomyelitis. *Orthopaedics* **7,** 1557–64. [Classification scheme for chronic osteomyelitis, now widely accepted and used.]

Cremieux A-C, Carbon C (1997). Experimental models of bone and prosthetic joint infection. *Clinical Infectious Diseases* **25,** 1295–302. [Useful review of animal models.]

Gristina A, *et al*. (1985). Adherent bacterial colonisation in the pathogenesis of osteomyelitis. *Science* **228,** 990–3. [Important exposition of the role of bacterial adhesion in disease.]

Jacobs RF, McCarthy RE, Elser JM (1989). Pseudomonas osteocondritis complicating puncture wounds of the foot in children: a 10-year evaluation. *Journal of Infectious Diseases* **160,** 657–61. [Appropriate surgery allowing short-course antibiotic therapy for 'tennis shoe osteomyelitis'.]

Lew DP, Waldvogel RA (1997). Osteomyelitis. *New England Journal of Medicine* **336**, 999–1007. [Review.]

Lipsky BA (1997). Osteomyelitis of the foot in diabetic patients. *Clinical Infectious Diseases* **25,** 1318–26. [Excellent comprehensive review of a common and difficult form of osteomyelitis.]

McNally MA, *et al*. (1993). Two-stage management of chronic osteomyelitis of the long bones. *Journal of Bone and Joint Surgery* **75B,** 375–80. [Combined orthopaedic and plastic surgical management of complex osteomyelitis.]

Mader JT, *et al*. (1990). Hyperbaric oxygen as adjunctive therapy for osteomyelitis. *Infectious Diseases Clinics of North America* **4,** 433–40. [An expert's review of this mode of treatment.]

Norden CW (1996). Bone and joint infection. *Current Opinion in Infectious Diseases* **9,** 109–14. [A still-useful review of advances up to 1996 from an *eminence grise* of bone infection.]

Rissing JP (1997). Antimicrobial therapy for chronic osteomyelitis in adults: role of the quinolones. *Clinical Infectious Diseases* **25,** 1327–33. [Useful review of intravenous and oral therapies.]

Swiontkowski MF, *et al.* (1999). A comparison of short and long course i.v. antibiotic therapy in the post-operative management of adult osteomyelitis. *Journal of Bone and Joint Surgery* **81B**, 1046–50. [Useful paper.]

Syrogiannopoulos GA, Nelson JD (1988). Duration of antimicrobial therapy for acute suppurative osteoarticular infections. *Lancet* **2,** 37–40. [Large series that defined oral short course regimens for uncomplicated acute bone and joint infection.]

Tice AD (1991). Once-daily ceftriaxone outpatient therapy in adults with infections. *Chemotherapy* **37** (Suppl. 3)**,** 7–10. [Description of outpatient intravenous antibiotic therapy.]

Wininger DA, Fass RJ (1996). Antibiotic-impregnated cement and beads for orthopaedic infections. *Antimicrobial Agents and Chemotherapy* **40,** 2675–9. [A review of this mode of treatment.]

19.4 Osteoporosis

Juliet Compston

Introduction

Osteoporosis is characterized by a reduction in bone mass and disruption of bone architecture, resulting in increased bone fragility and an increase in fracture risk. These fractures are widely recognized as a major health problem in the elderly population, resulting in an estimated annual cost to British health services of £1.5 billion. One in three women and one in five men surviving to the age of 80 years will suffer a hip fracture due to osteoporosis; demographic changes over the next 50 years are predicted to lead to at least a doubling in the number of these fractures, largely as a result of increased longevity.

Epidemiology

Osteoporotic fractures are termed fragility fractures (defined as occurring after a fall from standing height or less). They may occur at a number of skeletal sites but fractures of the distal radius (Colles' fracture), spine, and proximal femur are most characteristic. The incidence of osteoporotic fractures increases markedly with age; in women, the median age for Colles' fractures is 65 years and for hip fracture, 80 years. The age at which vertebral fracture incidence reaches a peak has been less well defined but is thought in women to be between 65 and 80 years. In men, no age-related increase in forearm fractures is seen but hip fracture incidence rises exponentially after the age of 75 years. The prevalence of vertebral fractures rises with age in men, although less steeply than in women.

The remaining lifetime risk of osteoporotic fracture in 50-year-old British white women has been estimated at 14 per cent for the hip, 11 per cent for the spine, and 13 per cent for the radius; for any osteoporotic fracture, this risk approaches 40 per cent in women and 13 per cent in men. For women this risk is similar to that of cardiovascular disease and is approximately six times higher than that of breast cancer. In the United Kingdom it is estimated that approximately 60 000 hip fractures and 50 000 Colles' fractures occur annually; for vertebral fractures, the figure of 40 000 reflects only those which are clinically diagnosed and probably represents only about one-third of all fractures. There are marked geographical variations in the incidence of osteoporotic fractures, the reasons for which remain only partially defined. The condition is most common in Asian and Caucasian populations but rare in African and American black populations.

Clinical features

Colles' fractures typically occur after a fall forwards on to the outstretched hand. They cause considerable inconvenience, usually requiring 4 to 6 weeks in plaster and long-term adverse sequelae are seen in up to one-third of patients. These include pain, sympathetic algodystrophy, deformity, and functional impairment.

Spinal fractures are characterized by varying degrees of vertebral deformity (Fig. 1) and may occur spontaneously or as a result of normal activities such as lifting, bending, and coughing. A minority of vertebral fractures (possibly around one-third) present with acute and severe pain at the site of the fracture, often radiating around the thorax or abdomen. The natural history of this pain is variable; in general, there is a tendency for improvement with time but resolution is often incomplete. Multiple vertebral deformities result in spinal deformity (kyphosis), height loss, and corresponding alterations in body shape with protuberance of the abdomen and loss of normal body contours. These changes are commonly associated with loss of self-confidence and self-esteem, difficulty with daily activities, and increased social isolation. The clinical impact of spinal fractures is thus substantial, although often underestimated.

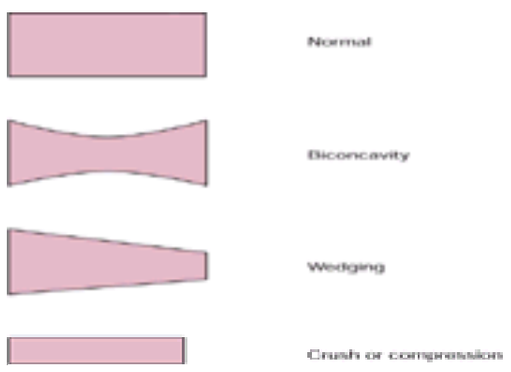

Fig. 1 Vertebral deformities associated with osteoporosis. It should be noted that some degree of biconcavity can be a normal variant and wedge and crush deformities are most commonly associated with symptoms.

Of all the osteoporotic fractures, hip fractures cause the greatest morbidity and mortality. They almost always follow a fall, either backwards or to the side, and require admission to hospital and surgical treatment. Because hip fractures characteristically affect frail elderly people, postoperative morbidity and mortality are high; at 6 months after fracture, mortality rates of 12 to 20 per cent have been reported. Only a minority of patients regain their former level of independence following a hip fracture and up to one-third require institutionalized care.

Pathogenesis

Lifetime changes in bone mass are shown in Fig. 2. Peak bone mass is attained in the third decade of life and age-related bone loss is believed to start in both men and women around the beginning of the fifth decade; thereafter bone loss continues throughout life. In women, there is an acceleration of the rate of bone loss around the time of the menopause, the duration of which is poorly characterized but may be 5 to 10 years.

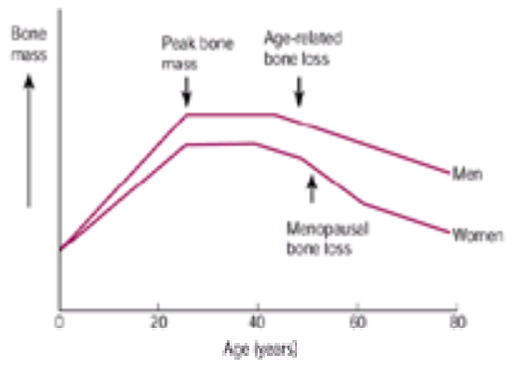

Fig. 2 Schematic representation of lifetime changes in bone mass in men and women. (Reprinted with permission from Compston JE, 1995, Osteoporosis, corticosteroids and inflammatory bowel disease. *Alimentary Pharmacology and Therapeutics* **9**, 237–50.)

Bone mass in later life thus depends both on the peak bone mass achieved in early adulthood and on the rate of age-related bone loss. Genetic factors strongly influence peak bone mass, accounting for up to 70 to 80 per cent of its variance. A number of genes are likely to be involved; a polymorphism of the collagen type IA1 gene has been shown to be associated both with low bone mineral density and fracture and it is likely that other genes will be identified in the near future. Sex hormone status, nutrition, and physical activity also influence peak bone mass.

In women, oestrogen deficiency is a major pathogenetic factor in menopausal bone loss. In men, the relationship between age-related bone loss and declining testosterone levels is less well documented. In the elderly, vitamin D insufficiency and secondary hyperparathyroidism are common and contribute to age-related bone loss. Other potential pathogenetic factors include declining levels of physical activity and intestinal calcium malabsorption.

A number of endogenous and exogenous risk factors for osteoporosis have been identified. In the former category, advancing age, female gender, Caucasian or Asian race, and a family history of osteoporosis (particularly a maternal history of hip fracture) are the most important. Strong exogenous risk factors include hypogonadism in either sex, glucocorticoid therapy, low body weight, and previous or prevalent fragility fracture. Several diseases, including hyperthyroidism, hyperparathyroidism, inflammatory bowel disease, and chronic renal or liver dysfunction are also associated with increased risk of osteoporosis. Finally, low dietary calcium intake, vitamin D deficiency, immobilization, cigarette smoking, and excessive alcohol intake have adverse effects on bone mass.

Risk factors for falling are major determinants of fracture risk, particularly for hip fracture in the elderly. Their recognition is important, since many are modifiable. They include poor visual acuity, neuromuscular weakness or incoordination, reduced mobility, cognitive impairment, and the use of sedatives, tranquillizers, and alcohol. There are also many environmental hazards that increase the risk of falling, such as uneven paving stones, poor lighting, and loose carpets and wires.

Pathophysiology

The mechanical competence of the skeleton is maintained by the process of bone remodelling, in which a quantum of bone is removed by osteoclasts followed by the formation, in the cavity so created, of new bone by osteoblasts. Under normal circumstances resorption always occurs before formation and the amounts of bone resorbed and formed within each bone remodelling unit are similar.

In menopausal bone loss, there is an increase in the number of bone remodelling units on the bone surface (increased bone turnover), resulting in a higher number than normal of remodelling units undergoing resorption at any one time. In addition, within each of these units less bone is formed than resorbed, leading to a negative remodelling imbalance. It is believed that one of the early, and probably transient effects, of oestrogen deficiency is to increase the activity of osteoclasts, probably by suppressing apoptosis. Increased osteoclastic activity causes an increase in the depth of erosion of bone by these cells, contributing to the trabecular penetration and disruption of bone architecture that characterizes postmenopausal osteoporosis.

The pathophysiology of other forms of osteoporosis remains to be fully defined. In glucocorticoid-induced osteoporosis, reduced bone formation and low bone turnover predominate in those treated long term, but there is evidence that in the early stages of treatment there is an increase in bone turnover and osteoclast activity. The alterations in bone remodelling responsible for osteoporosis in men have not been established, but the lesser degree of structural disruption of cancellous bone during ageing suggests that reduced bone formation plays a greater role in age-related bone loss in men than women. Whether this applies to men with osteoporosis, however, is uncertain.

Diagnosis

Several techniques are available for the assessment of bone mass; these include single and dual energy X-ray absorptiometry, quantitative computed tomography, and ultrasound. At present, dual energy X-ray absorptiometry is regarded as the optimal approach in clinical practice because of its ability to measure bone mineral density at axial and appendicular sites, its good reproducibility, and the very low doses of radiation required. However, ultrasonography is currently being evaluated and may become more widely adopted in the future.

The rationale for the use of bone densitometry in clinical practice is the demonstration, in prospective studies, of a continuous and inverse relationship between bone mineral density and fracture risk. For every standard deviation decrease in bone mineral density, there is a two- to threefold increase in fracture risk; this is quantitatively similar to the relationship between blood pressure and stroke and greater than that observed for serum cholesterol and coronary heart disease. Bone mineral density at a wide range of sites, including spine, hip, radius, and os calcis, is able to predict fracture risk, although the best predictive value, at least in the case of hip fracture, is provided by measurement at the potential fracture site.

The relationship between bone mineral density and fracture risk forms the basis for the densitometric classification of osteoporosis. This is based on T scores—standard deviation scores above or below normal peak bone mass—and defines osteoporosis as a bone mineral density T score at the hip and/or spine below –2.5. Osteopenia is defined as a T score between –1 and –2.5 and established osteoporosis as a T score below –2.5 in the presence of one or more fragility fractures. Although these are diagnostic rather than interventional criteria, a T score below –2.5 would generally be considered as an indication for intervention because of the high associated fracture risk.

Since population-based screening cannot be justified at present, patients are selected for bone densitometry in clinical practice on the basis of risk factors and/or symptoms or signs suggestive of osteoporosis. Indications for bone densitometry are shown in Table 1. Bone densitometry should only be performed where the result will influence clinical management; it is not indicated in patients with clear evidence of established osteoporosis and in the very elderly, in whom low bone mass is almost universal, bone densitometry is rarely useful in directing clinical management. In addition, the presence of osteophytes, extraskeletal calcification, and vertebral and/or spinal deformity in the elderly significantly reduces the accuracy of spinal measurements.

Secondary causes of osteoporosis should be excluded where appropriate. A full blood count, liver function tests, serum calcium and phosphate levels, thyroid function tests, plasma immunoelectrophoresis, and Bence-Jones protein determination should be performed in the first instance with further investigation if indicated. In men, in whom secondary causes are more common, serum testosterone, gonadotrophins and prolactin, and 24-h urinary cortisol should also be performed.

Biochemical markers of bone resorption (such as urinary deoxypyridinoline, pyridinoline, N-terminal and C-terminal cross-linked telopeptides of type I collagen) and formation (such as osteocalcin, bone-specific alkaline phosphatase, C-terminal propeptide of type I procollagen) have been shown to be useful in the prediction of fracture risk, particularly when combined with bone mineral density measurements, and in the monitoring of response to treatment. However, their role in clinical practice has not been firmly established.

Current therapeutic options for osteoporosis

A number of options are now available for the prevention of osteoporotic fractures in postmenopausal women (Table 2). For historical reasons the level of evidence on which the registration of these interventions is based varies widely; thus adequately powered, randomized, controlled trials with fracture as the primary end-point exist only for alendronate, raloxifene, and combined calcium and vitamin D, whereas in the case of hormone replacement therapy, evidence for antifracture efficacy is based almost solely on observational data which are subject to bias and likely to overestimate beneficial effects.

Hormone replacement therapy

Hormone replacement at the menopause, whether unopposed or combined with progestogens, prevents bone loss and there is evidence, mainly from observational studies, for protection against fracture at the hip, spine, and wrist. At least in terms of its effects on bone mineral density, oestrogen replacement is effective both when given at the menopause and if started some years later; however, it is generally less well tolerated in more elderly women and compliance is correspondingly lower.

There is growing evidence for attenuation of the beneficial effects of hormone replacement therapy on the skeleton following cessation of therapy, both in terms of its effects on bone mineral density and protection against fracture. The implications of these findings are first, that lifelong treatment after the menopause is likely to be required to maintain optimal fracture protection, and second, that the most cost-effective treatment strategies will be those which target high-risk women for treatment.

Short-term side-effects of hormone replacement therapy include breast tenderness and withdrawal bleeding. Although continuous combined 'no-bleed' preparations are now available, some vaginal bleeding is experienced by up to 30 per cent of women during the first few months of such therapy. The principal concern with long-term use is an increase in the risk of breast cancer; after 5 to 10 years of use, the relative risk increases by around 30 per cent, which is significant in terms of absolute risk for a disease which affects 1 in 12 postmenopausal women. Other adverse extraskeletal side-effects include a two- to threefold increase in risk of venous thromboembolism and there may also be a small increase in the risk of endometrial cancer.

Hormone replacement therapy also has important short-term and long-term benefits. It is effective in alleviating vasomotor and other menopausal symptoms and may also have beneficial effects on postural stability. Observational data indicate a significant reduction in morbidity and mortality attributable to coronary heart disease, although this remains to be confirmed in prospective studies. However, a recent study indicates that combined hormone replacement therapy in older women may not be effective in the secondary prevention of coronary heart disease; in that study, more deaths due to heart disease occurred during the first year in treated women than in controls. Other potential but as yet unproven benefits of long-term hormone replacement therapy include improved cognitive function, protection against Alzheimer's disease, and a reduction in risk of colon cancer.

Accurate evaluation of the risk/benefit ratio of hormone replacement therapy cannot at present be performed because of the lack of prospective data which are required to demonstrate the magnitude of potential risks and benefits. Nevertheless, many women are reluctant to take indefinite hormone replacement therapy because of the increase in breast cancer risk and treatment is thus often limited to a finite period of between 5 and 10 years.

Bisphosphonates

The bisphosphonates are synthetic analogues of the naturally occurring compound pyrophosphate. They inhibit bone resorption by complex and only partially understood mechanisms and may also inhibit mineralization. Three bisphosphonates are currently licensed for use in osteoporosis. Etidronate is administered cyclically and intermittently, the 3-month cycle consisting of 400 mg daily of etidronate for 2 weeks followed by calcium only for 76 days. In contrast, alendronate is given as a single daily dose of 10 mg and calcium is not included in the formulation; once weekly doses of 70 mg of alendronate have an equivalent therapeutic effect. Thirdly, risedronate is given as a single daily dose of 5 mg.

These bisphosphonates have been shown to prevent bone loss in the spine and hip, both in healthy perimenopausal women and in more elderly women with osteoporosis. The magnitude of this effect is similar for both agents; however, unlike the majority of antiresorptive agents, bisphosphonate therapy (at least in the case of alendronate) is associated with a sustained although small increase in bone mineral density, which is believed to be due to hypermineralization of bone. This may occur as a result of the suppression of bone turnover and its consequences for bone strength have not been established.

In a large, randomized controlled trial, treatment with alendronate for nearly 3 years was associated with a reduction of approximately 50 per cent in vertebral and non-vertebral fractures in postmenopausal women with osteoporosis. In the case of cyclic etidronate therapy, the clinical trials did not demonstrate statistically significant fracture reduction, but favourable trends for vertebral fracture were observed after 3 years of treatment and observational data also indicate protective effects against hip and other non-vertebral fractures. Finally, risedronate has been demonstrated to reduce vertebral and non-vertebral fractures, including hip fractures, in postmenopausal women with osteoporosis. A significant reduction in vertebral fractures was observed in the first year of treatment.

Bisphosphonates are generally well tolerated. Gastrointestinal side-effects may occur, especially with aminobisphosphonates, and a small number of cases of erosive oesophagitis have been reported with alendronate. It is therefore important that patients take the drug according to the instructions, namely in the morning with a full glass of water, 30 min before food, drink, or other medications, and remaining upright for 30 min after the dose. Alendronate is contraindicated in patients with oesophageal abnormalities or disease and should be withdrawn immediately if dyspepsia or dysphagia develop during therapy.

The optimum duration of bisphosphonate therapy is unknown. Despite their high skeletal retention, preliminary indications are that bone loss resumes soon after treatment is discontinued, although further studies are required in this area. There are theoretical concerns that prolonged suppression of bone turnover may have adverse effects on bone strength and, at present, treatment is usually given for a period of 5 to 10 years.

Raloxifene

Raloxifene is a selective oestrogen receptor modulator which has oestrogenic effects in the skeleton without the unwanted effects of oestrogen in the breast and endometrium. In randomized controlled trials, raloxifene has been shown to prevent menopausal bone loss in healthy early postmenopausal women; in addition, a 30 per cent reduction in vertebral fracture rate was seen after 3 years of treatment with 60 mg daily in postmenopausal women with osteoporosis. Beneficial effects on bone mineral density are seen both in the spine and proximal femur but no effect of raloxifene therapy on non-vertebral fracture has been demonstrated.

Raloxifene is taken orally as a single daily dose. Minor adverse effects include leg oedema, leg cramps, and hot flushes. As with hormone replacement therapy there is a two- to threefold increase in the relative risk of venous thromboembolism. Raloxifene does not alleviate, and may exacerbate, menopausal vasomotor symptoms; it should therefore be avoided in perimenopausal women with such symptoms.

The extraskeletal effects of raloxifene are of considerable interest. In particular, a highly significant protective effect against breast cancer which is oestrogen receptor positive has emerged in the clinical trials; overall, after 4 years of treatment there was a 75 per cent reduction in new cases of breast cancer, this figure rising to 90 per cent when only cases that were oestrogen receptor positive were considered. Raloxifene use is not associated with vaginal bleeding and does not increase the incidence of endometrial hyperplasia or carcinoma. The effects of raloxifene on cognitive function and cardiovascular disease risk have not been established; in the context of the latter, effects on serum lipid profile similar but not identical to those observed with oestrogen have been reported.

Calcitonin

Calcitonin may be administered parenterally or intranasally; both forms of treatment have been shown to prevent spinal bone loss in postmenopausal women, but treatment benefits at other sites such as the proximal femur and radius have not been clearly demonstrated. The effects of calcitonin on fracture rate are controversial although some randomized controlled trial data indicate beneficial effects on vertebral fracture risk. Adverse effects with intranasal calcitonin are rare. Nausea and flushing may occur shortly after parenteral administration of calcitonin and vomiting and diarrhoea also sometimes occur. These symptoms are usually transient but may persist for some hours after injection.

Vitamin D and calcium

There is increasing evidence that vitamin D and calcium supplementation protects against non-vertebral fractures in elderly subjects. Thus in a randomized controlled

trial of vitamin D and calcium in daily doses of 800 IU and 1.2 g, respectively, a significant reduction in hip and other non-vertebral fractures was seen after 12 to 18 months of treatment in a cohort of very elderly women (mean age 84 years) who were living in sheltered accommodation. Subsequently, a significant reduction in non-vertebral fractures was reported in community-dwelling men and women aged over 65 years in a randomized controlled trial of 700 IU of vitamin D and 500 mg of calcium daily. It is not possible from these studies to deduce the relative contribution of vitamin D and calcium to the observed benefits; vitamin D without calcium has been shown in some studies to reduce non-vertebral fracture rate in the elderly, but this finding has not been universal. The important question of whether vitamin D alone reduces hip fracture thus remains unanswered at present.

Calcitriol

Calcitriol (1,25-dihydroxyvitamin D, the active metabolite of vitamin D) preserves bone mineral density in women with postmenopausal osteoporosis and there is evidence that it also reduces vertebral fracture rate, although the latter finding has not been universal. Effects on non-vertebral fracture have not been documented. Calcitriol is given orally in a dose of 0.5 to 1.0 µg daily; hypercalciuria and hypercalcaemia may occur and serum calcium levels should be monitored at regular intervals.

Calcium

Beneficial effects of calcium on bone mineral density have been documented in children and adults, particularly at appendicular skeletal sites. In lumbar spinal bone, these effects are generally less evident and may be transient; the benefits of calcium are also less marked in perimenopausal women, presumably because of the dominant effects of oestrogen deficiency. Although several small studies have reported a reduction in vertebral fracture rate in calcium-supplemented individuals, evidence from adequately powered studies is not available and calcium should be regarded as an adjunct to treatment rather than as definitive therapy.

Non-pharmacological interventions

Hip protectors have been shown to protect against hip fracture in randomized controlled trials in the elderly and should be considered in all those at high risk, including those who have already sustained a hip fracture. Physiotherapy has an important role to play in the management of pain and restricted mobility and measures such as hydrotherapy and TENS (transcutaneous electrical nerve stimulation) are often effective. In elderly patients, occupational therapy is also often helpful and assessment of the risk of falling should be performed with advice on reducing risk where appropriate.

Weight-bearing exercise can produce modest, site-specific increases in bone mineral density in younger adults, but its skeletal effects in postmenopausal women are less certain and it should not be regarded as a definitive treatment. In the elderly, exercise may reduce the risk of falling and, if a fall should occur, improve the neuromuscular protective responses; however, the efficacy of this approach in reducing fracture risk has yet to be proved in randomized controlled trials.

Other aspects of treatment

Pain associated with acute vertebral fracture is often underestimated and can be difficult to manage. Very strong analgesics should be avoided where possible since these may increase the risk of falling and bed rest should be restricted to a minimum to avoid further bone loss associated with immobilization. Calcitonin is often effective in the treatment of pain associated with vertebral fractures; salcatonin is usually given subcutaneously in a dose of 100 IU daily or on alternate days for a period for 3 to 6 weeks.

Treatment of glucocorticoid-induced osteoporosis and osteoporosis in men

Cyclical etidronate, alendronate, and risedronate therapy have all been shown to be effective in the prevention of glucocorticoid-induced osteoporosis and are licensed in the United Kingdom for this indication. In patients receiving high doses of glucocorticoids (such as 15 mg of prednisolone daily or equivalent) for 3 months or more or who have strong risk factors for osteoporosis, bisphosphonate therapy should be started immediately; in those receiving lower doses (more than 7.5 mg) for 6 months or more, bone densitometry should be performed and prophylaxis instituted if the T score is below –1.5.

Alendronate is the only treatment currently licensed for the prevention of osteoporotic fractures in men. Beneficial effects on bone mineral density have been reported after treatment with testosterone or cyclical etidronate, but effects on fracture risk have not been reported for either agent.

Further reading

Adachi JD *et al*. (1997). Intermittent etidronate therapy to prevent corticosteroid-induced osteoporosis. *New England Journal of Medicine* **337**, 382–7.

Barrett-Connor E (1998). Hormone replacement therapy. *British Medical Journal* **317**, 457–61.

Black DM *et al*. (1996). Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. *Lancet* **348**, 1535–41.

Chapuy MC *et al*. (1992). Vitamin D_3 and calcium to prevent hip fracture in elderly women. *New England Journal of Medicine* **327**, 1637–42.

Cohen S *et al.* (1999). Risedronate therapy prevents corticosteroid-induced bone loss. *Arthritis and Rheumatism*, **42**, 2309–18.

Compston JE (1997). Prevention and management of osteoporosis: current trends and future prospects. *Drugs* **53**, 727–35.

Compston JE, Cooper C, Kanis JA (1995). Bone densitometry in clinical practice. *British Medical Journal* **310**, 1507–10.

Dawson-Hughes B *et al*. (1997). Effect of calcium and vitamin D supplementation on bone density in men and women 65 years of age and older. *New England Journal of Medicine* **337**, 670–6.

Eastell R (1995). Management of corticosteroid-induced osteoporosis. *Journal of Internal Medicine* **237**, 439–47.

Eastell R *et al*. (1998). Management of male osteoporosis: report of the UK Consensus Group. *Quarterly Journal of Medicine* **91**, 71–92.

Ettinger B *et al*. (1999). Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: results from a 3-year randomized clinical trial. *Journal of the American Medical Association* **282**, 637–45.

Gluer C-C (1997). Quantitative ultrasound techniques for the assessment of osteoporosis: expert agreement on current status. *Journal of Bone and Mineral Research* **12**, 1280–8.

Harris ST *et al*. (1999). Effects of risedronate treatment on vertebral and non-vertebral fractures in women with postmenpausal osteoporosis. A randomized controlled trial. *Journal of the American Medical Association*, **282**, 1344–52.

Hulley S *et al*. (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Journal of the American Medical Association* **280**, 605–13.

Kanis JA, McCloskey EV (1999). Effect of calcitonin on vertebral and other fractures. *Quarterly Journal of Medicine* **92**, 143–9.

Lauritzen JB, Petersen MM, Lund B (1993). Effect of external hip protectors on hip fractures. *Lancet* **341**, 11–13.

Marshall D, Johnell O, Wedel H (1996). Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. *Lancet* **312**, 1254–9.

Melton III LJ (1995). How many women have osteoporosis now? *Journal of Bone and Mineral Research* **10**, 175–7.

Michaëlsson K *et al*. (1998). Hormone replacement therapy and risk of hip fracture: population based case–control study. *British Medical Journal* **316**, 1858–63.

Orwell E *et al*. (2000). Alendronate for the treatment of osteoporosis in men. *New England Journal of Medicine*, **343**, 604–10.

Ralston SH (1997). Osteoporosis. *British Medical Journal* **315**, 469–72.

Seibel MJ, Woitge HW (1999). Basic principles and clinical applications of biochemical markers of bone metabolism: biochemical and technical aspects. *Journal of Clinical Densitometry* **2**, 299–322.

Soag KG *et al*. (1998) Alendronate for the prevention and treatment of glucocorticoid-induced osteoporosis. *New England Journal of Medicine*, **339**, 292–9.

Storm T *et al*. (1990). Effect of intermittent cyclical etidronate therapy on bone mass and fracture rate in women with postmenopausal osteoporosis. *New England Journal of Medicine* **322**, 1265–71.

Tilyard MW *et al*. (1992). Treatment of postmenopausal osteoporosis with calcitriol or calcium. *New England Journal of Medicine* **326**, 357–62.

WHO Study Group (1994). Assessment of fracture risk and its application to postmenopausal osteoporosis. *World Health Organization Technical Report Series* **843**.

19.5 Avascular necrosis and related topics

D. O'Gradaigh, C. A. Speed, and A. J. Crisp

Introduction

Osteonecrosis is a non-specific term for the death of bone, which was attributed to suppuration or trauma; Pasteur's study of bacteria in 1860 and the advent of clinical radiography led to recognition that many cases were aseptic. Phemister and others showed the importance of ischaemia in many cases and the collective term avascular necrosis has since been used. Associated disorders are the osteochondroses and osteochondritis dissecans. Although these conditions are separate entities, they are linked by the involvement of ischaemia (Table 1).

Avascular necrosis

There are 15 000 new cases of adult avascular necrosis annually in the United States. Males are more commonly affected (8:1), the majority under 50 years of age, with the exception of knee avascular necrosis, which particularly affects women over the age of 50. Any bone can be affected, the femoral head and condyles, the head of humerus, and the talus being the most commonly involved. he small cuboidal bones of the wrist and foot are less frequently affected. The unifying feature is a relatively poor vascular supply to subchondral bone through end arterioles with a limited collateral network.

Bone comprises osseous tissue and cartilage, with myeloid tissue, fatty marrow, and a sinusoidal network of vessels packing the inexpandable bone compartment. Several local mechanisms may be involved alone or in combination in vascular compromise. These include disruption of the vessel wall, raised intramedullary pressure through intraosseous venous congestion, and intravascular occlusion by atherosclerosis, thrombosis, or embolization. Death and lysis of haemopoietic cells and lymphocytes occurs with a subsequent macrophage and fibroblast response. Revascularization follows with deposition on the dead trabeculae of fibrous tissue (in traumatic cases) or of lamellar bone (in non-traumatic cases). Simultaneous formation and resorption occur, with progressive loss of cartilage. Should adequate revascularization and bone deposition not occur, articular collapse ensues.

Numerous conditions have been associated with avascular necrosis (Table 2). Alcoholism and corticosteroids are the most common, where fat embolization precipitating thrombosis is implicated. Higher doses and longer duration of steroid treatment increase the risk of avascular necrosis, though there is considerable variation. Interactions between steroids and the conditions for which they are prescribed may contribute, as some conditions (systemic lupus erythematosus, renal transplantation) appear to confer a particular risk of avascular necrosis.

Avascular necrosis should be considered in anyone presenting with bone pain, particularly those with associated risks. Since dead bone is painless and biomechanically sound, clinical symptoms of activity-related pain and joint dysfunction usually develop insidiously with the onset of repair processes or following articular collapse. Local tenderness, warmth, oedema, synovitis, and reduced range of movement may be evident. Infection must always be excluded.

Low-grade symptoms may precede radiological changes by months or years. In the early stages of avascular necrosis, radiographs are normal. Areas of mottled increase in radiodensity may be seen initially. A radiolucent crescent in subchondral bone (suggestive of fracture), cyst formation, and flattening and fragmentation of the articular surface may be noted in intermediate stages. Joint space narrowing and other degenerative signs occur in advanced stages. Bone scintigraphy is more sensitive in early, potentially reversible, avascular necrosis. A 'cold' area with a surrounding area of increased uptake is characteristic but rare, and non-specific increased uptake is more commonly seen. Magnetic resonance imaging (**MRI**) has become the first choice of investigation for early diagnosis. In addition to its high sensitivity, it is valuable for assessing the extent of the lesion.

Primary prevention is the ideal in 'at risk' groups, where a high index of suspicion must be maintained. Judicious use of corticosteroids, early diagnosis, and treatment of sickle-cell and metabolic syndromes and of alcoholism may reduce risk to bone. Guidelines are available for decompression in divers to reduce dysbaric complications.

Management of avascular necrosis depends on stage classified according to the imaging oulined above. Three main groups can be identified. Early stages (normal radiographs, positive scintigraphy or MRI) may resolve with conservative management. In some centres, raised intramedullary pressure is believed to be a critical factor, and core decompression biopsies (also useful for diagnosis) are therefore advocated; however, fracture is a significant complication. The intermediate group (crescent formation, flattening, etc.) often require surgery, including attempts at reperfusion (with vascularized pedicle grafts or an implanted artery), osteotomy, or arthroplasty. Arthroplasty is recommended in those with advanced degenerative changes. Rare complications of avascular necrosis include osteomyelitis (especially in sickle-cell disease) and malignancy.

Osteochondroses

Osteochondrosis is due to disturbance of endochondral ossification at a previously normal site of growth, involving chondrogenesis and osteogenesis. This can occur through mechanical (macro- or microtrauma) and/or vascular mechanisms which may vary according to the site involved. For example, Legg–Calvé–Perthes disease is considered to be of vascular aetiology. Osteochondrosis can occur at any epiphysis and involve the articular surface, the epiphyseal plate, or apophysis (secondary ossification centre or site of ligament or musculotendinous attachment). These areas are significantly weaker than surrounding soft tissue structures and are particularly vulnerable in growth spurts, where musculotendinous tightness (resulting in poor flexibility) contributes to apophyseal disorders. Growing children and adolescents are most commonly affected, males three times more frequently than females, the onset of symptoms occuring earlier in girls. Osteochondroses are classified into articular, non-articular, and physeal disorders (Table 1). Although they do not continue past the attainment of skeletal maturity, complications may come to light in adulthood.

Specific stages have been identified in the pathogenesis. Arrest of ossification at the affected site occurs, followed by revascularization and bone resorption. Later reossification may result in alteration in shape. It is often difficult to differentiate osteochondroses from normal ossification centres on radiography. Initial reduction in size and fissuring of the ossification centre may appear, followed by sclerosis and alteration in shape. Radiographs of the contralateral side should be obtained for comparison. In early lesions, scintigraphy may show increased uptake, and MRI may show cartilage disruption. While most cases are self-limiting, a poorer prognosis with development of osteoarthrosis is associated with larger lesions and older age at presentation. In apophysitis, radiography is not usually indicated, but may show loose ossicles or bony enlargement at the enthesis. Management is symptomatic, with reassurance, modification of activities, local application of ice, and non-steroidal anti-inflammatory drugs as required combined with a stretching regime.

Osteochondritis dissecans

Osteochondritis dissecans is a distinct form of osteochondral injury through the articular cartilage in a diarthrodial joint. It can affect all ages, but usually presents in teenage males, typically at the distal femur and particularly the lateral aspect of the medial femoral condyle. Other commonly affected sites are the patella, talus, and capitellum of the humerus. Although trauma has been identified in 50 per cent of cases, this is an unlikely aetiological factor in those under 15 years of age. A familial pattern is noted in 10 per cent, and lesions may be multiple and occur at several sites suggestive of multiple epiphyseal dysplasia.

The osteochondral fragment is susceptible to avascular necrosis. It may remain *in situ* or become partially or completely detached, which may precipitate effusion and mechanical symptoms of locking, catching, and giving way. The disorder usually presents with progressive activity-related pain. Local swelling may be evident if trauma has occurred. External tibial rotation when walking is characteristic in medial femoral involvement.

Plain radiographs may be normal and specialized views (such as the notch view of the knee) may be required, showing a typical subchondral crescent sign or loose bodies. Scintigraphic findings resemble those of the osteochondroses, while computed tomography is useful to determine the site and size of the lesion; the overlying cartilage can be evaluated by MRI.

Treatment aims to achieve union of the fragment and restoration of joint surface integrity. In young patients with open epiphyses, healing may be achieved with conservative management, as outlined for osteochondrosis. Joint immobilization may be necessary. Surgery should be considered in skeletally mature patients, those who fail conservative management, or those with detached fragments. This involves debridement and internal fixation of the fragment with drilling or vascular grafting of the base. The prognosis depends on the patient's age, and the stability and location of the fragment. Degenerative joint disease in later life is a major complication.

Further reading

Bohndorf K (1998). Osteochondritis (osteochondrosis) dissecans: a review and new MRI classification. *European Radiology* **8**, 103–12.

Chang CC, Greenspan A, Gershwin ME (1993). Osteonecrosis: current perspectives on pathogenesis and treatment. *Seminars in Arthritis and Rheumatism* **23**, 47–69.

Mitchell DG, Rao VM, Dalinka MK (1987). Femoral head avascular necrosis: correlation of MR imaging, radiographic staging, radionuclide imaging and clinical findings. *Radiology* **162**, 709–15.

Mont MA, Carbone JJ, Fairbank AC (1996). Core decompression versus non-operative management for osteonecrosis of the hip. *Clinical Orthopaedics* **324**, 169–78.

Williams JS Jr, Bush Joseph CA, Bach BR Jr (1998). Osteochondritis dissecans of the knee. *American Journal of Knee Surgery* **11**, 221–32.

20.1 Structure and function of the kidney

J. D. Williams and A. Phillips

> The organs of the human body were created to perform ten functions, among which is the function of the kidney to furnish the human being with thought.
>
> Leviticus Rabba 3, Talmud Berochoth 61B

Introduction

The human kidney is formed by the fusion of a number of lobes. The structure of a single lobe is best understood by examining the unilobar kidney of a small mammal: in the rat, for example, the medulla is enfolded by the cortex on all sides other than its pelvic aspect, where it projects as a papilla into the renal pelvis. The renal cortex, containing all the glomeruli and the proximal and distal convoluted tubules appears in coronal section to be distinct from the pyramidal-shaped medulla, which contains the loops of Henle. The medulla is divided into an outer and inner medulla, the outer medulla being subdivided into an outer and inner stripe.

The nephron

The functional unit of the kidney is the nephron, which begins at the glomerulus (Fig. 1). The urinary space (the cavity between the glomerulus and its surrounding Bowman's capsule) leads into the proximal tubule, which itself can be subdivided into a convoluted segment and a straight segment. The straight segment of the proximal tubule descends into the medulla and changes abruptly into the descending limb of Henle's loop. This loop penetrates for varying distances into the medulla before returning to the cortex. The longer loops pass all the way into the inner medulla, whilst the short loops only reach the outer medulla. Generally speaking, long loops belong to nephrons of glomeruli lying adjacent to the medullary region, while the shorter loops belong to the more superficial glomeruli. The descending limb of Henle bends sharply back at its lowest point to form the ascending limb, which at another abrupt transition forms the medullary part of the thick ascending limb. This leads up into the cortex where it becomes convoluted and comes into close contact with the vascular pole of its own glomerulus, forming the juxtaglomerular apparatus. Further along the nephron the thick ascending limb becomes the distal convoluted tubule and then the connecting tubule, which joins the cortical collecting duct. Each collecting duct receives connecting tubules from about 12 nephrons and then opens onto the surface of a papilla.

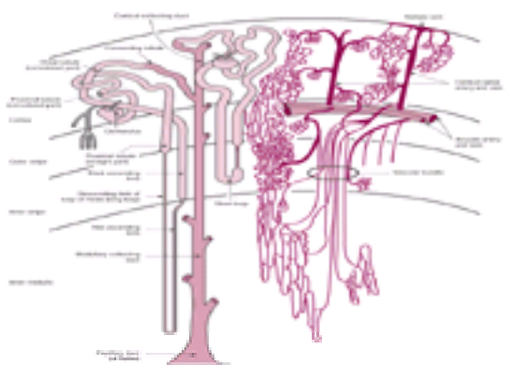

Fig. 1 The nephron and its blood supply. (Reproduced from Williams JD, *et al. Clinical atlas of the kidney*. Gower Publishing, London, with permission.)

The renal blood supply

Structure

The renal artery divides into the interlobar arteries and enters the renal substance at the columns of Bertin (the area between adjacent lobes). At the junction of the cortex and medulla the arteries divide again and form the arcuate arteries (Fig. 1). Each arcuate artery gives rise to cortical radial arteries that ascend through the cortex: there is no direct arterial supply to the medulla. The afferent glomerular arteries arise from the cortical radial arteries and directly supply the glomeruli. Efferent glomerular arteries drain the glomeruli and then supply the peritubular capillaries of the cortex and medulla, a unique arrangement meaning that the peritubular capillary supply is exclusively postglomerular. Efferent glomerular arteries can be divided into two types: those from the superficial and midcortical glomeruli supply the capillary plexus of the cortex; those from juxtamedullary glomeruli form the blood supply to the renal medulla. Within the outer stripe they divide into the descending vasa recta, which penetrate the inner stripe in vascular bundles. The renal medulla is drained by the ascending vasa recta, which traverse the inner stripe within the vascular bundle and then join the cortical radial veins. The vascular bundles of the medulla represent the vascular component of the countercurrent exchange mechanism between the blood entering and leaving the medulla. Interestingly, the vascular bundles are organized such that the perfusion of the inner medulla is kept totally separate from the perfusion of the outer medulla. The cortical radial veins join the arcuate veins to eventually form the interlobular veins, which run alongside corresponding arteries.

Function

Renal blood flow is influenced by intrarenal and extrarenal factors. Autoregulation within the kidney maintains a relatively stable blood flow to the glomerulus over a range of arterial pressure. This phenomenon seems to be mediated by events intrinsic to the kidney since it has been demonstrated in both denervated and isolated kidney preparations.

The glomerulus

Structure

On entering the glomerulus (Fig. 2(a)) the afferent arteriole divides into primary capillary branches, each of which gives rise to an anastomosing capillary network that

forms a glomerular lobule. These capillaries then coalesce into the efferent arteriole within the tuft. The structural organization of the capillaries is unlike that found in any other part of the body. The capillary basement membrane (glomerular basement membrane) forms the barrier across which filtrate is generated. Embryologically, the glomerulus is the interface between the ureteric bud (or hollow nephrogenic vesicle) and the metanephrogenic cap, which develops into the capillary plexus. The result of this is a basement membrane formed by the fusion of the basement membrane of the capillaries and the basement membrane of the nephrogenic vesicle. This glomerular basement membrane (**GBM**) forms the skeletal framework of the glomerular tuft.

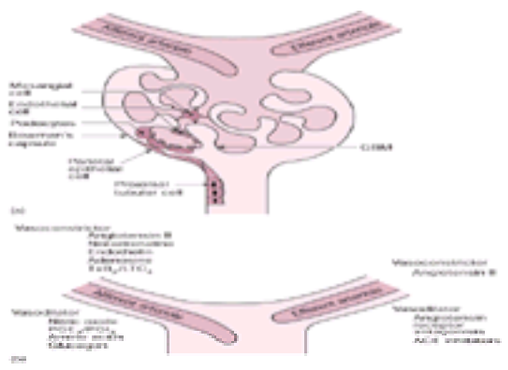

Fig. 2 The glomerulus: (a) structure; (b) regulation of glomerular blood flow by vasoactive agents.

Although on electron microscopy the GBM appears as a three-layer structure with a central lamina densa and outer lamina rara interna and externa, this is probably an artefact. Freeze–fracture studies have suggested uniformity in the basement membrane from its outer to inner aspects. The major components of the membrane include a framework of type IV collagen linked by heparan sulphate proteoglycans (**HSPG**) and laminin, the basement membrane charge being provided by the heparan sulphate component (subtypes of which include perlican and agrican). Type IV collagen consists of a triple helix of fibres with a large non-collagenous globular domain at the C-terminal end (called NC1). This NC1 domain of the collagen molecule is the target for Goodpasture's disease, and mutations of the collagen chains are responsible for Alport's syndrome.

The endothelial cells, the basement membrane, and the podocytes form the filtration barrier. The endothelial cells are fenestrated (60 and 100 nm in diameter) and the lack of a diaphragm across the fenestrations exposes the basement membrane directly to the glomerular capillary contents. The luminal surface of the endothelial cells is negatively charged by polyanionic glycoproteins, but these are not present on the fenestrae. The capillary loops are incomplete (a tube of fenestrated endothelial cells surrounded only on its epithelial aspect by a basement membrane) and are held together on their inner aspect by the mesangial cells. Thus the basement membrane has an opening on its mesangial aspect so that the endothelial cells are in direct contact with the mesangium. At the vascular pole of the glomerulus the capillary basement membrane is reflected to form the parietal epithelium of Bowman's capsule.

The outer aspect of the filtration barrier is provided by the epithelial cells (podocytes), which interdigitate on the surface of the glomerular lobules. The foot processes of adjacent podocytes are separated by the filtration slits, which are bridged by the slit diaphragms and are the sites through which the glomerular filtrate passes. The pores have a central proteinaceous core with side arms linking to each adjacent cell, forming a structure with a zipper-like appearance and a width of about 40 nm. The luminal surface of the podocyte and the slit diaphragm are rich in negative charge, being covered in glycoproteins. The podocyte surface adjacent to the basement membrane expresses a number of adhesion proteins that ensure firm anchorage to the membrane.

The mesangium forms the pillar to which the GBM scaffold is attached. The interaction between the mesangial cells and the basement membrane provides the mechanism for the contractility of the glomerular tuft, and the means whereby the surface area of the tuft can be varied. The spaces between the mesangial cells are filled by the mesangial matrix and consist of a number of different collagens, as well as glycoproteins, fibronectin, and proteoglycans. This mesangial matrix provides a channel for the migration of a variety of molecules from the glomerular capillaries, with trafficking centrally towards the vascular pole of the glomerulus.

Function

The glomerular filtration barrier, consisting of the endothelial pores, the glomerular basement membrane, and slit diaphragms, will exclude molecules on the basis of size, shape, and charge. Size selectivity is imparted by the matrix of the GBM itself, as well as by the integrity of the podocytes. The matrix, formed by the type IV collagen molecules, consists of a series of interlinking pores, the narrowest of which determines the size of molecules that can pass through. Thus any pathological change to the structure of the matrix is likely to result in a greater permeability of the GBM. The resistance to the movement of water and small molecules is provided by the endothelial pores, the basement membrane, and by the available surface area of the slit diaphragms, the last of these probably being the effective barrier. The charge barrier, whose efficiency is disputed, is provided by the negative charge of the basement membrane and to a lesser extent by the surface of the endothelial and epithelial cells.

The effectiveness of the filtration barrier is dependent not only on the integrity of the basement membrane but also on the function of the epithelial cells. Most studies have demonstrated that changes in barrier function are closely correlated with significant alteration to the podocytes. These include changes in the surface area of the slit diaphragm (slit diaphragm frequency) as well as detachment of podocytes from the basement membrane. There is now a large body of evidence to suggest that HSPGs are involved in both the charge- and size-selective properties of the GBM and, furthermore, that alterations in GBM HSPGs may be important in the development of proteinuria.

Production of glomerular filtrate

The number of functioning glomeruli and the filtration rate at each single glomerulus determines the GFR. There are approximately a million glomeruli per human kidney, of which 90 per cent are in the outer two-thirds of the cortex and are fairly homogenous in terms of structure and function. The remaining 10 per cent, which are located in the juxtamedullary region, are larger with a higher single-nephron GFR compared with cortical glomeruli.

Single-nephron GFR is determined by a number of factors. First, the pressure of blood in the glomerular capillary and the hydrostatic pressure of the fluid in Bowman's space determine the pressure difference that drives the movement of fluid across the glomerular capillary wall, the transglomerular hydrostatic pressure difference or ‡*P*. Second, the gradient in colloid osmotic pressure (‡*p*) across the filtration barrier: this is equal to the colloid osmotic pressure within the glomerular capillary less the colloid osmotic pressure in Bowman's space (which, in effect, is zero). The difference between ‡ *P* and ‡*p* is the net ultrafiltration pressure. Other determining factors are water permeability (*K*) and *f* (the area available for filtration, namely the surface area of the slit pores between podocytes), these two combined forming the glomerular filtration coefficient or *Kf*. Hence:

$$\text{single-nephron GFR} = (‡P - ‡p) \times Kf.$$

SNGFR can be regulated by alterations in the ultrafiltration coefficient *Kf*, the net ultrafiltration pressure, or both. A change in the net ultrafiltration pressure may arise due to a change in the hydraulic pressure ‡*P*, the capillary plasma oncotic pressure (‡*p*), and/or alterations in the initial glomerular capillary plasma flow rate (the latter dictates changes in protein concentration with distance along a capillary network and hence affects colloid osmotic pressure).

Glomeruli contain receptors for a number of hormones that are capable of modifying the filtration rate (see Fig. 2(b)). These include vasoconstrictors such as adenosine, angiotensin II, and endothelin as well as vasodilators including dopamine, bradykinin, prostacyclin, and nitric oxide. Some of these vasoactive molecules are produced within the kidney, whilst others are delivered by the systemic circulation, and many studies have examined the effects of hormones on glomerular ultrafiltration. It is clear from the preceding discussion that, in addition to ‡ *P* (that is, change in hydraulic pressure), the glomerular filtration rate (**GFR**) is dependent on the capillary plasma flow rate and the ultrafiltration coefficient (*Kf*), all of which may be altered by hormones. Vasoconstrictor substances such as angiotensin II and norepinephrine are capable of producing substantial reductions in renal plasma flow, generally with little change in GFR. Angiotensin II, for example, causes constriction of both afferent and efferent arterioles with a resultant decrease in capillary plasma flow, a reduction in *Kf*, but little change in the single-nephron GFR (**SNGFR**) due to an increase in ‡*P*. Increased afferent arterial tone caused by endothelin or adenosine will decrease renal blood flow, decrease ‡ *P*, and therefore

decrease GFR. By contrast, dilatation of the afferent arteriole by nitric oxide or prostaglandins will also cause an increase in ‡ *P*, but with an increase in renal blood flow and hence an increase in GFR.

In the normal adult human, water is filtered by the glomerulus at a rate of 80 to 200 ml/min. The glomerular filtration rate is critically related to all functions of the kidney and is closely regulated by mechanisms that maintain a constant high value for GFR. In practical clinical terms, the estimation of GFR is achieved by measuring the renal clearance of a substance that is freely filtered at the glomerulus and not absorbed or secreted by the renal tubules. For discussion of the methods of measuring GFR in clinical practice, see Chapter 20.3.1.

The proximal convoluted tubule

Structure

The main function of the proximal tubule is to reabsorb the bulk of filtered water and solutes, and its structure shows numerous adaptations for this purpose. Proximal tubular epithelial cells are tall and columnar with a well-developed brush border, resulting in a 40-fold increase in the apical surface area of the cells (Fig. 3). In addition they possess extensive basolateral interdigitation, increasing the basolateral cell surface area. The apices of the cells are held together by junctional complexes: these are called 'tight junctions' (zona occludens), of the leaky variety, with a low electrical resistance that allows some transepithelial transport. The bases of the cells rest on the tubular basement membrane, which separates them from the peritubular capillaries. Another characteristic feature is the presence of large numbers of mitochondria, intimately associated with the basolateral cell membranes where the Na^+/K^+-ATPase is located, and whose function is to provide the energy source for fluid and electrolyte reabsorption.

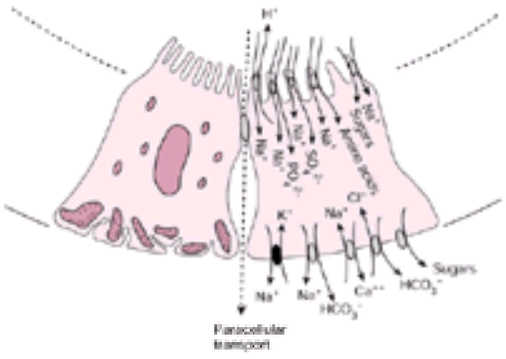

Fig. 3 Proximal tubular cell function. Principal transport processes of the proximal tubular cell.

Function

Sodium and water reabsorption

About seven-eighths of the volume of the glomerular filtrate is reabsorbed in the proximal tubule. Sodium enters the proximal tubular cells passively from the tubular fluid down an electrochemical gradient that is the driving force for fluid and electrolyte reabsorption. This gradient is produced by the action of the Na^+/K^+-ATPase on the basal surface, which transports sodium out of the cell in excess of the potassium transported into the cell, thereby generating a transmembrane potential of –70 mV. Chloride ions follow the same route by cotransport with Na^+, and the resulting increase in osmolality in the intercellular spaces results in the absorption of water by osmosis, such that the volume of the filtrate in the renal tubule is substantially reduced by the time it reaches the beginning of the loop of Henle, although its net osmolality does not change. In addition to this transcellular route for the transport of salt and water, there is also a paracellular route through the 'leaky' tight junctions.

Reabsorption of other substances

The proximal tubule is also responsible for the reabsorption of other substances such as glucose, phosphate, amino acids, and organic anions, including citrate and lactate. These enter the proximal tubular cells across the apical membrane by a series of cotransport systems, each of which binds one or more sodium ions and its specific substrate, and carries them across the cell membrane (Fig. 3). Thus, the rate of sodium entry into the cell is linked by cotransport systems to the reabsorption of these substances.

The energy for secondary active transport or cotransport of substances (glucose, phosphate, etc.) against their concentration gradient is therefore provided indirectly by Na^+/K^+-ATPase, which is responsible for the concentration gradient for sodium across the cell membrane. This is illustrated by the reabsorption of glucose, which involves brush-border, Na^+-coupled glucose transporters, termed **SGLT**, and basolateral facilitated glucose transporters (**GLUT**). In the human, the major site for glucose reabsorption is the early S1 segment of the proximal tubule, where 90 per cent of the filtered glucose is reabsorbed, such that only a small fraction of the filtered load reaches the S2 and S3 segments. Glucose reabsorption in the S1 proximal tubular segment is mediated by the low-affinity, high-capacity, Na^+/glucose cotransporter, SLGT2, whilst reabsorption in the later segments is mediated by the high-affinity, low-capacity SGLT1. Similarly, the high rate of glucose efflux characteristic of the early proximal tubular segment is mediated by the low-affinity, facultative glucose transporter GLUT2 and high-affinity GLUT1, whereas only GLUT1 is expressed in the late proximal tubule where a minor portion of the filtered glucose load is reabsorbed.

The kidneys are also involved in maintenance of the acid–base balance of the body by regulating the serum bicarbonate concentration to approximately 24 mmol/l. The proximal tubule reabsorbs between 80 and 90 per cent of the filtered bicarbonate, largely by the following mechanism. H^+ is secreted by the Na^+/H^+-exchanger on the luminal membrane. It then reacts with the filtered HCO_3^- to form H_2CO_3, which is converted to CO_2 and H_2O catalysed by carbonic anhydrase present on the luminal brush-border membrane. CO_2 diffuses passively into the cell where it is split to yield the H^+ that is secreted and OH^-, the hydroxyl ion then reacts with CO_2 (catalysed by carbonic anhydrase) to yield HCO_3^-, which exits the cell via a Na/HCO_3 synporter thus restoring filtered HCO_3^- to the plasma.

Handling of protein

In addition to its role in fluid and electrolyte balance, almost all the protein that is filtered at the glomerulus is reabsorbed by the proximal tubule via a process of endocytosis. To date, four major routes of tubular handling of peptides have been identified: (1) reabsorption of filtered protein/peptides by endocytosis and intracellular lysosomal degradation; (2) luminal hydrolysis and reabsorption of free amino acids; (3) carrier-mediated reabsorption of small intact peptides; and (4) peritubular uptake of peptides. The most important of hese is probably the endocytotic route.

In recent years there has been considerable interest in the role of proteinuria in the progression of renal disease. Amongst the hypotheses currently under investigation are those that focus on the effect of excess protein trafficking on the generation of profibrotic factors by proximal tubular cells and the subsequent initiation of interstitial fibrosis. These theories suggest that cells of the proximal tubule play a role in maintaining the normal architecture of the renal interstitium. In support are numerous studies demonstrating that tubular cells are a rich source of many components of the extracellular matrix, which may modify matrix turnover by alterations in the synthesis of both matrix-degrading enzymes and their inhibitors, as well as through the production of cytokines. More recent studies have suggested that cells of the proximal tubule may migrate into the interstitium and transdifferentiate into the cortical fibroblasts during conditions of inflammation.

Interplay between the regulation of GFR and proximal tubular function

One aspect of the control of renal function is the correlation between the volume of filtrate produced by the glomerulus and the reabsorptive capacity of the renal tubule. The movement of sodium and water from the proximal tubular lumen into the capillary network depends on the hydrostatic pressure of the blood in the peritubular capillary complex, as well as the osmotic pressure of the blood within those capillaries. An increased hydrostatic pressure will reduce reabsorption, but an increased oncotic pressure will enhance reabsorption. Thus, increased systemic blood pressure will increase the interstitial pressure within the interstitium and result

in the movement of sodium from the interstitial fluid into the lumen (pressure natriuresis).

The loop of Henle

The loop of Henle begins where the straight (S3) part of the proximal tubule changes abruptly in diameter to become the descending thin limb. Long loops pass into the inner medulla, before performing a hairpin bend and returning as the thin ascending limb, when an abrupt transition at the inner stripe of the outer medulla marks the beginning of the thick ascending limb, which is structurally distinct from its thin counterpart. In the case of the short loops the transition to ascending thick limb takes place before the bend, so that the thick part of the tubule forms the loop.

Although there are only minor structural differences between the thin segments of descending and ascending limbs, there are major differences in their permeability properties. The thin descending limb, like the proximal tubule, is highly permeable to water as a result of the presence of aquaporin 1, whereas the thin ascending limb is impermeable to water. By contrast, the descending limb is impermeable to sodium, whilst significant sodium and urea reabsorption occurs in the thin ascending limb. This allows an osmotic gradient to be established in the medulla, which is the basis of the countercurrent multiplier mechanism (Fig. 4).

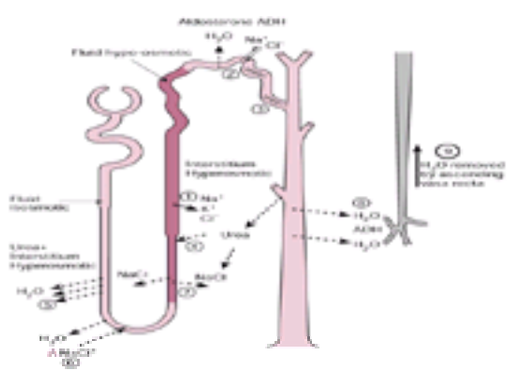

Fig. 4 Diagram to illustrate the mechanism of concentration of the urine. The darkened part of the nephron is impermeable to water. 1, Active transport of Na^+ and Cl^- into the insterstitium. 2, Reabsorption of Na^+ and Cl^-. Passive absorption of water under ADH control. 3, Increased concentration of urea in tubule following reabsorption of water. 4, Urea passes into the interstitium, thereby increasing osmolality. 5, The increased interstitial osmolality results in more water being extracted. 6, This leads to an increased salt concentration in the loop of Henle. 7, In the ascending limb, salt diffuses into the interstitium, further increasing its osmolality. 8, In the presence of ADH the permeability of the distal nephron and collecting ducts is increased and water is reabsorbed. 9, Water is removed from the interstitium by vasa recta. (Reproduced from Williams JD, *et al. Clinical atlas of the kidney*. Gower Publishing, London, with permission.)

The juxtaglomerular apparatus

The juxtaglomerular apparatus comprises the macula densa, the extraglomerular mesangium, the terminal portion of the afferent arteriole with its renin-producing granular cells, and the early portions of the efferent arteriole.

The thick ascending limb of the loop of Henle returns to its own glomerulus, where the cells that lie nearest to the glomerulus become taller to form the macula densa, the most obvious structural feature being that these cells are tightly packed and have large nuclei. The basal aspect of the macula densa is firmly attached to the extraglomerular mesangium.

The granular cells (also termed the juxtaglomerular cells) are assembled in clusters within the terminal portion of the afferent arteriole. These are modified smooth muscle cells containing cytoplasmic granules in which renin is stored. This enzyme is responsible for controlling the synthesis of angiotensin II by converting angiotensinogen to angiotensin I. This in turn is converted to angiotensin II by the action of the angiotensin-converting enzyme.

Granular cells appose the extraglomerular mesangial cells, adjacent smooth muscle cells, and endothelial cells, and are densely innervated by sympathetic nerve terminals. The secretion of renin by the granular cells is controlled by signals generated intrarenally (such as perfusion pressure and tubular fluid composition) and extrarenally, due to changes in sympathetic output and by stimuli that decrease the extracellular fluid (**ECF**) volume and blood pressure. Many factors may therefore be involved in the control of renin release, a particularly important one of these being an intrarenal baroreceptor mechanism that causes renin secretion to increase when the intrarenal arteriolar pressure at the granular cells is decreased. A major level of control also lies in the macula densa, where renin secretion is proportionate to the concentration of Cl^- or Na^+ in the tubular fluid. Decreased delivery of Na^+ and Cl^- to the macula densa is associated with increased renin secretion. Angiotensin II, by contrast, inhibits renin secretion by its direct action on the granular cells; it is also a major stimulant of aldosterone secretion, thereby stimulating sodium retention (see below), which closes the renin–angiotensin–aldosterone negative-feedback loop. In addition to these factors, increased activity of the sympathetic nervous system increases renin secretion, both by increased circulating catecholamines and by way of the renal sympathetic nerves.

It has been postulated that the intrarenal renin–angiotensin mechanism is the prime hormonal mediator of the tubuloglomerular feedback system, whereby a stimulus perceived at the macula densa, presumably related to luminal flow or ion concentration, influences filtration rate (Fig. 5). Evidence for this is inconclusive and it is almost certainly not the sole mediator of this feedback mechanism.

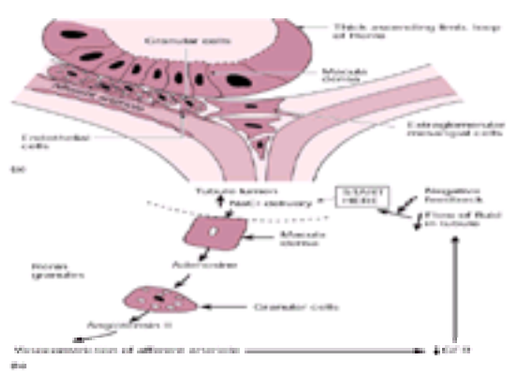

Fig. 5 Tubuloglomerular feedback: (a) anatomical basis; (b) putative mechanism.

The distal tubule and collecting duct

The bulk of sodium and water reabsorption occurs primarily in the proximal tubule, but fine regulation is necessary to maintain a precise sodium and water balance. The distal tubule and the collecting duct are responsible for the necessary final adjustments, which ultimately determine the rate of urinary water and sodium excretion, a mechanism substantially influenced by antidiuretic hormone (vasopressin, **ADH**) and aldosterone, respectively.

Structure

The distal convoluted tubule begins just beyond the macula densa and ends at the cortical collecting duct. Its structure is similar to that of the main part of the thick ascending limb of the loop of Henle. The collecting duct system includes the connecting tubule and the cortical and medullary collecting ducts. The connecting tubule

and the collecting ducts, unlike the distal tubule, are lined by two cell types: principal cells, with small basal infoldings, some mitochondria, and small microvilli; and intercalated cells with darkly staining cytoplasm that contains mitochondria, smooth endoplasmic reticulum, and prominent Golgi apparatus. There are at least two types of intercalated cells, distinguished on the basis of immunocytochemical and functional characteristics: type A cells express H^+-ATPase at their luminal membrane and secrete protons, whilst type B cells express H^+-ATPase at their basolateral membrane and secrete bicarbonate ions.

Function

Cells of both the connecting tubule and the collecting duct share sensitivity to ADH, but only those of the collecting duct are sensitive to mineralocorticoids. The renal concentrating and diluting processes are ultimately dependent on the ability of ADH to modulate the water permeability of collecting ducts. Regulation of ADH is dependent on osmoreceptors in the hypothalamus, which recognize changes in ECF osmolality, but this can also occur in the absence of changes in plasma osmolality, for example intravascular volume depletion, pain, nausea. Once released from the posterior pituitary, vasopressin exerts its biological action on water excretion by binding to receptors in the basolateral membrane of the collecting duct (Fig. 6). This results in increased adenylate cyclase activity, increased cAMP formation, and ultimately causes the apical (luminal) cell membrane to become more permeable to water through insertion of aquaporin 2 channels.

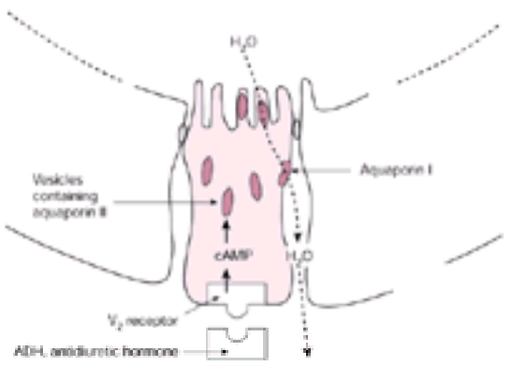

Fig. 6 Action of antidiuretic hormone.

The principal cells of the collecting duct are responsible for the modulation of sodium reabsorption. Entry of sodium into these cells occurs down a concentration gradient through specific sodium ion channels in the luminal membrane. This creates a negative potential difference in the lumen, which promotes either the secretion of potassium or the reabsorption of chloride via the paracellular route. These processes, which are the final regulators of sodium balance, are under the control of aldosterone, which increases the number of open sodium-ion channels in the luminal membrane (Fig. 7). As previously discussed, angiotensin II is a major stimulant of aldosterone secretion. Hence during periods of volume depletion, activation of the renin–angiotensin system leads to increased aldosterone production and sodium retention; whereas when volume-replete, the system is suppressed and renin release and aldosterone secretion are reduced, resulting in natriuresis. Although the acute production of aldosterone is linked to the renin–angiotensin system, other mechanisms (including that of sodium or potassium balance) can also affect the ability of the adrenal glands to produce aldosterone.

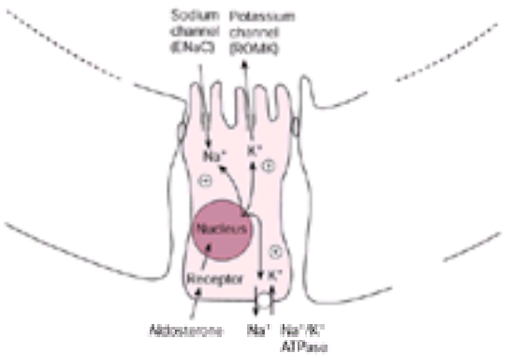

Fig. 7 The action of aldosterone on the collecting duct. Aldosterone stimulates an increase in the numbers and activity of apical ENaC and ROMK, and of basolateral Na^+/K^+-ATPase by direct and indirect effects.

The intercalated cells of the collecting duct are involved in maintenance of the acid–base balance. The method by which they excrete acid by generating ammonium ions is discussed in Chapter 11.11.

The interstitium

The renal interstitium is the space that is not occupied by the glomeruli or nephrons, and the vasculature of the kidney can be thought of as lying within it. The interstitium amounts to some 5 to 7 per cent of the volume of the cortex, 3 to 4 per cent of the outer stripe, 10 per cent of the inner stripe, and up to 30 per cent of the inner medulla. It is involved in virtually all functions of the healthy kidney, as well as in many pathological events. The transit of molecules from the tubules to the blood necessitates a crossing of the interstitial space, and vice versa. Thus, changes to the interstitium have a profound effect on the function of the tubules and indeed of the nephron itself.

The cells of the interstitium are not a homogeneous population but comprise different cell types that vary anatomically within the kidney and between health and disease. The major cellular component is the fibroblast; however, there is evidence to suggest that there is a significant difference between the phenotype of the cortical fibroblast and that of the inner medullary fibroblast. Fibroblasts are important for the integrity of the interstitial matrix and are considered to be the source of matrix component production, as well as being responsible for their turnover. The renal fibroblasts also have endocrine functions: those of the cortex are the source of erythropoietin; the inner medullary fibroblast produces significant amounts of prostaglandins, primarily PGE_2, and has a function in modifying electrolyte transport. Renal fibroblasts may have a pivotal role in renal interstitial fibrosis; it is now well established that the progression of renal disease is intimately linked to the degree of renal interstitial fibrosis, and it is likely that the key cell involved in this may well be the cortical fibroblast.

Dendritic cells are present in small numbers throughout the interstitium and express MHC class II receptors. In addition, there are a few macrophages as well as some large lymphocytes. Dendritic cells, macrophages, and lymphocyte-like cells mostly have immunological and defence-like functions.

Further reading

Davison AM *et al*., eds (1998). *Oxford textbook of clinical nephrology*, 2nd edn. Oxford University Press, Oxford.

Johnson RJ, Feehally J (2000). *Comprehensive clinical nephrology*. Harcourt Publishers Ltd.

Windhage E, ed. (1992). *Handbook of physiology*, section 8 (renal physiology). Published for the American Physiological Society by Oxford University Press, New York.

20.2.1 Water and sodium homeostasis and their disorders

Peter H. Baylis

Introduction

Total body water accounts for about 60 per cent of the body weight of a healthy adult: two-thirds of this is intracellular and one-third extracellular. The extracellular fluid compartment is divided into the vascular (blood volume) and the interstitial fluid compartments in the ratio 1:2. For a 75 kg adult, the total volume of body water is approximately 45 litres, with intracellular and extracellular volumes of 30 and 15 litres respectively; the latter comprising the blood (5 litres) and interstitial (10 litres) compartments. Sodium is the main extracellular cation, which with its anion, chloride, contributes 95 per cent of the extracellular solute. By contrast, the major intracellular cation is potassium. Many cell membranes are freely permeable to water, but not to most electrolytes, which results in the same total solute but very different electrolyte concentrations in the extracellular and intracellular compartments (Fig. 1).

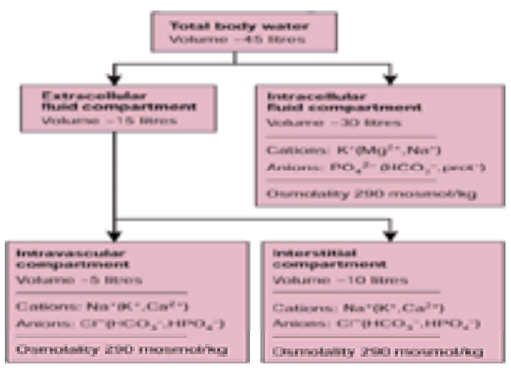

Fig. 1 Composition of body compartments. Body water is distributed uniformly throughout all compartments. Major and some minor (in parentheses) anions and cations are indicated. Osmolality remains the same inside and outside the cell.

The maintenance of stable volume and solute concentrations is essential to all complex animals, including humans. In the extracellular fluid compartment, particularly the vascular component, the control of water and sodium balance is inextricably linked. Water is distributed uniformly throughout all compartments, and is clearly a determinant of volume, but is also essential in establishing the concentration of sodium (and other solutes). Sodium, however, contributes ot only to its own concentration, but its total quantity in the extracellular fluid is the main factor determining the volume of that compartment. A variety of integrated mechanisms ensure that minimal fluctuations, probably less than 1 per cent, in blood volume and in sodium concentration occur in healthy adults.

Body water and therefore solute concentration is regulated mainly by vasopressin (antidiuretic hormone) mediated alteration of renal water excretion, but also to some extent by thirst as a motivation for drinking. The secretion of vasopressin and thirst are influenced principally by changes in circulating concentration of sodium, but also in part by significant falls in blood volume or pressure, and there are instances where these stimuli can be pulling vasopressin secretion in opposite directions (see later).

The volume of the extracellular compartment is determined by its total sodium content, which is regulated by numerous mechanisms. Sodium intake is poorly controlled in humans, although some animals do demonstrate a specific sodium appetite. The kidney is the major effector organ influencing sodium homeostasis. Complex intrarenal mechanisms contribute to the maintenance of sodium homeostasis, in addition to which are endocrine factors that either tend to reduce excretion of sodium by the kidney (for example the renin–angiotensin–aldosterone system) or which produce a natriuresis (for example atrial natriuretic peptide, ouabain-like substances). The situation is very complex. Blood volume and pressure are also influenced by a variety of vasoactive substances that act locally or systemically (for example catecholamines, prostaglandins, nitric oxide, endothelins), as well as by changes in sympathetic nerve activity. Any change in blood volume and/or pressure will, in turn, have an effect on vasopressin secretion. It can therefore be appreciated that there is an intricate network of homeostatic mechanisms controlling both sodium and water balance.

In clinical practice, the precise measurement of circulating concentrations of electrolytes, specific non-electrolytic solutes (for example glucose), and total solute is relatively simple, and approximates closely to their concentrations in interstitial fluid. Sodium is measured in molar terms (mmol/litre) using flame photometry or an ion-selective electrode. Total solute concentration is assessed by determining the depression of the freezing point of the sample plasma using an osmometer, and is expressed as the number of osmoles of solute per kilogram (osmolality). Thus, a solution of glucose at 1 mmol/litre will provide an osmolality of 1 mosmol/kg but a 1 mmol/litre solution of a salt (for example NaCl), which dissociates completely in the solvent into sodium and chloride ions, will have an osmolality of 2 mosmol/kg. The clinical assessment of volume—whether it be intravascular, interstitial, or extracellular—is extremely important, but difficult and inaccurate.

The physiology of water homeostasis

The maintenance of normal water balance is achieved through the combined action of three main factors: vasopressin, the kidney, and thirst. There needs to be secretion of adequate quantities of osmotically stimulated vasopressin, which must be able to bind to the renal tubule to modulate the flow of solute-free water and produce antidiuresis. Most healthy adults excrete 1 to 2 litres of urine per 24 h, but the normal kidney is capable of wide variation in urine output, ranging from 0.5 to (in very extreme cases) 25 litres per 24 h. Osmotically stimulated thirst must be able to promote drinking and is particularly important when the kidney is concentrating urine maximally but there is still persistent water loss from, for example, excessive sweating or copious watery diarrhoea. Under these circumstances water homeostasis cannot be maintained without adequate fluid intake.

Fine control of water balance ensures that the concentration of solutes, particularly extracellular sodium, remains stable. The extraordinary sensitivity of the function of the three homeostatic mechanisms detailed above allows plasma osmolality to be maintained within the narrow range 285 to 295 mosmol/kg (equivalent to serum sodium, 137 to 142 mmol/litre) in healthy adults.

Thirst and water intake

Drinking behaviour of humans can be divided into two types. The first, primary drinking, occurs as a result of physiological stimulation of thirst. This initiates drinking behaviour to allow ingestion of sufficient fluid to lower blood osmolality. Secondary drinking, which is far more common in our culture, occurs for social reasons (the endless cups of coffee throughout the day, or the ritual visit to the 'pub'), habit, or the need to drink with food. For the majority of adults living in temperate climates, secondary drinking ensures that they remain in a state of mild water excess, and water balance is maintained by regulating renal water excretion.

The mechanism of primary drinking is believed to be as follows: as the body loses water, blood osmolality starts to rise and stimulates thirst osmoreceptors. Studies in animals indicate that these are situated in the anterior hypothalamic structures, probably in the organum vasculosum of the lamina terminalis or the subfornical organ, where there is a defect in the blood–brain barrier. Isolated lesions in this area can occur in humans following haemorrhage from an aneurysm in an anterior communicating artery and result in loss of thirst appreciation, suggesting that human thirst osmoreceptors are located in a similar area to those in animals. The precise mechanism by which blood hyperosmolality stimulates thirst is not known, but it is believed that increase in extracellular osmolality draws water from within the osmoreceptor cells of the organum vasculosum of the lamina terminalis and subfornical organ, resulting in cellular hypovolaemia which is translated into neuronal impulses that migrate to the cortex and allow conscious appreciation of thirst.

With the use of visual analogue scales it is possible to obtain an estimate of the degree of thirst sensation. There is a simple relationship between increasing blood osmolality and the intensity of thirst (Fig. 2(b)). Furthermore, there is also a direct relationship between intensity of thirst and the amount of fluid drunk to quench thirst. The act of drinking quickly reduces thirst, usually within a few minutes and certainly before there are substantial falls in blood osmolality. Drinking is therefore able to override or inhibit osmotically stimulated thirst, probably through an oropharyngeal reflux.

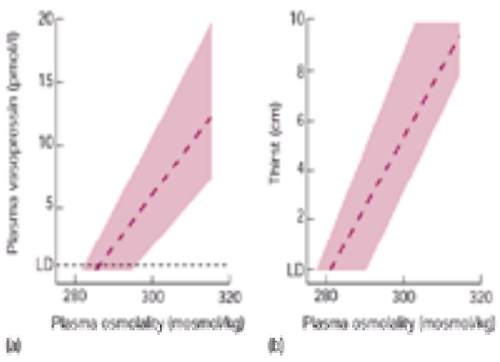

Fig. 2 (a) The relationship between plasma osmolality and plasma vasopressin during infusion of hypertonic (850 mmol/litre) saline in healthy adults. There is a linear relationship between the two variables, represented by the dashed line, termed the mean osmoregulatory line for vasopressin secretion. The abscissal intercept of the line represents the threshold for vasopressin secretion, approximately 283 mosmol/kg. LD is the limit of detection of the assay, and the shaded area is the extent of the normal response. (b) The relationship between plasma osmolality and thirst intensity assessed on a 10 cm visual analogue scale during the same hypertonic infusion. The dashed line is the mean osmoregulatory line for thirst, which has an abscissal intercept (thirst threshold) of 281 mosmol/kg. (Adapted from Thompson CJ *et al*. (1986). The osmotic thresholds for thirst and vasopressin release are similar in healthy man. *Clinical Science* **71**, 651–6, with permission.)

Thirst is not only stimulated by increases in blood osmolality, but also by acute substantial falls in blood volume and/or pressure. A sudden decrease in volume in excess of 15 per cent is required before thirst is influenced. Low-pressure baroreceptors located in the atria of the heart and great veins of the chest mediate the response. In addition, significant extracellular volume depletion is a potent stimulus to the release of renin from the juxtaglomerular apparatus of the nephron: this generates increasing concentrations of circulating angiotensin II, known to be a profound dipsogen in animals. Systemic angiotensin II therefore augments the baroregulatory influence on thirst in acute hypovolaemia. Animal studies have also revealed that intrahypothalamic angiotensin II is the most potent neurotransmitter involved in the generation of the sensation of thirst.

As humans age, so their thirst appreciation becomes blunted, and primary drinking is reduced such that individuals tend to become mildly hyperosmolar and hypernatraemic. Fortunately, most elderly people continue secondary drinking and rely on mechanisms to control renal water excretion, providing protection from significant hypernatraemia.

During human pregnancy there is a fall in plasma osmolality of the order of 10 mosmol/kg, with an appropriate fall in serum sodium. This is due to alteration in the osmoregulatory systems for both thirst and vasopressin secretion. The thirst osmoregulatory line (Fig. 2) is displaced to the left of the normal, non-pregnant position, which runs parallel—but the abscissal intercept, known as the osmotic threshold for thirst, is reset to about 275 mosmol/kg. Similar changes occur with the osmoregulatory line for vasopressin secretion (see below). The precise mechanisms for this 'resetting of the osmostat' are unknown, but they have important implications for the ability of pregnant women to handle a water load.

Vasopressin and renal water excretion

The antidiuretic hormone of humans is arginine vasopressin (in contrast to lysine vasopressin which is specific to the pig family), a nonapeptide, the gene for which is located on chromosome 20. Arginine vasopressin is synthesized from a large precursor molecule in the supraoptic and paraventricular nuclei of the hypothalamus, transported in neurosecretory granules to the posterior pituitary, median eminence of the hypothalamus and to a lesser extent to other areas of the brain and brainstem. It is secreted from the posterior pituitary into the systemic circulation to influence renal function, and into the hypothalamopituitary portal circulation to enhance pituitary ACTH secretion.

Control of vasopressin release

Secretion of vasopressin from the posterior pituitary is regulated mainly by changes in blood osmolality. The vasopressin osmoreceptors, distinct from the thirst osmoreceptors, are located in the same anterior hypothalamic area, i.e. the circumventricular structures, the organum vasculosum of the lamina terminalis, and possibly the subfornical organ. Rising blood osmolality is believed to cause water to flow out of the osmoreceptor cells, with cellular hypovolaemia then initiating a neuronal signal that passes principally to the supraoptic nucleus and stimulates the process of vasopressin synthesis and secretion. There is an exquisitely sensitive linear relationship between blood osmolality and vasopressin secretion (Fig. 2(a)), the slope of the vasopressin osmoregulatory line being a measure of the sensitivity of the system and the abscissal intercept representing the threshold for vasopressin release. At plasma osmolality values below 285 mosmol/kg, on average, vasopressin secretion is inhibited to allow a maximum water diuresis (15–25 litres/24 h) with urine osmolality of 50 to 70 mosmol/kg (Fig. 3). Increase in blood osmolality above this threshold induces progressive vasopressin release, thus increasing urine concentration, so that at plasma vasopressin values of 2 to 4 pmol/litre, maximum antidiuresis occurs. Drinking inhibits osmoregulated vasopressin secretion.

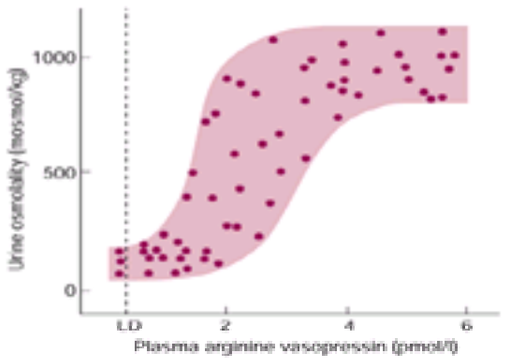

Fig. 3 The effect of vasopressin on urinary concentration during varying states of hydration in humans. Each closed circle represents a single value, and the stippled area the normal range. Values of plasma vasopressin greater than 4 pmol/litre fail to increase urinary concentration further. LD is the limit of the assay.

Each individual has a unique threshold and sensitivity for both thirst and vasopressin release. Circulating solutes have varying abilities to stimulate the osmoregulatory system, with sodium chloride being among the most potent and glucose having little or no effect. By contrast to osmoregulated thirst, there is no blunting of the vasopressin response to osmotic stimulation with ageing. Pregnancy is associated with a lowering of the vasopressin threshold similar to the thirst threshold, allowing osmoregulation to occur about a lower set-point of 275 rather than 285 mosmol/kg.

Non-osmotic release of vasopressin is stimulated by a number of factors: acute substantial reductions in blood volume or pressure, of the order of 10 to 15 per cent or more; nausea and/or emesis; hypoglycaemia; and a variety of circulating substances (for example angiotensin II). Low-pressure receptors in the great veins of the chest and cardiac atria mediate the effect of hypovolaemia, while receptors in the arch of the aorta and carotid vessels sense reductions in arterial pressure. The sensory information is carried via the vagus and glossopharyngeal nerves to the brainstem vasomotor centres and then transmitted to the hypothalamus, principally the paraventricular nucleus. There is an exponential relationship between the fall in blood volume/pressure and vasopressin release, such that large reductions (~40 per cent of normal) raise plasma vasopressin to huge concentrations (100 to 500 pmol/litre) that have vasoconstrictor effects. Similarly high vasopressin concentrations can be achieved with nausea/emesis.

Actions of vasopressin

The major physiological action of vasopressin is to increase urinary concentration (Fig. 3). Circulating arginine vasopressin binds to a specific renal tubular receptor, designated the V_2 receptor, of the collecting ducts. Adenyl cyclase is stimulated, via the coupled G protein, to produce cyclic 5'AMP, which activates intracellular protein kinases and accelerates the expression and trafficking of aquaporin 2, the vasopressin-sensitive water channel protein. Aquaporin 2 is organized into a tetramer and inserted into the luminal cellular membrane of the distal tubule, allowing water to flow from the tubular lumen into the cellular compartment. Two other aquaporins (aquaporins 3 and 4) are located on the contraluminal cell membrane: these are not vasopressin responsive but facilitate the flow of water across the distal tubule under the influence of the osmotic gradient between the hypotonic urine within the tubular lumen and the hypertonic renal interstitium. Thus, urine volume is decreased and urine is concentrated (Fig. 4).

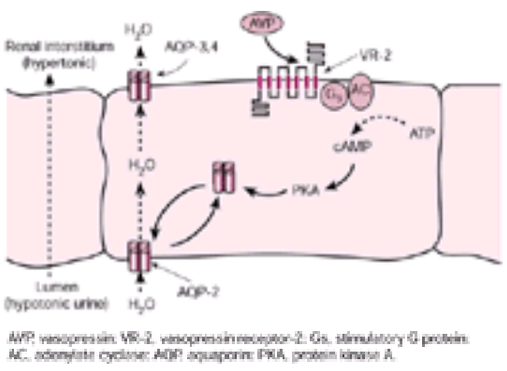

AVP, vasopressin; VR-2, vasopressin receptor-2; G_s, stimulatory G protein; AC, adenylate cyclase; AQP, aquaporin; PKA, protein kinase A.

Fig. 4 A schematic diagram of the distal renal tubule indicating the effect of arginine vasopressin initiating a cascade of intracellular events after binding to the V_2 receptor. Protein kinase A is activated, which mobilizes the arginine vasopressin-sensitive water channel protein, aquaporin 2, for insertion into the luminal tubular membrane. Non-vasopressin-sensitive water channel proteins, aquaporins 3 and 4, are positioned in the contralateral membrane, which allows water to flow through the cell along the osmotic gradient. Loss of arginine vasopressin binding to its receptor promotes re-entry of the luminal aquaporin 2 into the cell, resulting in a shuttling of aquaporin 2.

Animal studies have indicated that vasopressin also stimulates the transport of urea across the collecting tubule and of sodium chloride across the medullary thick ascending limb of the loop of Henle, both of which enhance the osmotic gradient. As renal prostaglandins reduce the generation of cyclic 5'AMP, they blunt the effect of vasopressin, and therefore prostaglandin synthetase inhibitors augment the antidiuretic action of arginine vasopressin.

The first action attributed to vasopressin was the elevation of systemic blood pressure by peripheral vasoconstriction. Arginine vasopressin binds to vascular smooth muscle receptors (V_1 receptors) that activate phosphatidyl–inositol pathways, increase intracellular calcium concentration, and cause the contraction of vascular muscles. High circulating concentrations of vasopressin are necessary to achieve this pressor effect: at physiological levels it probably plays little (if any) role in maintaining blood pressure, but it is involved in the pressor response to hypovolaemia or hypotension.

Vasopressin released from the hypothalamic median eminence binds to a modified V_1 receptor on the pituitary corticotroph and acts to enhance the release of ACTH stimulated by corticotrophin-releasing factor. At high concentrations vasopressin increases circulating concentrations of the clotting factors, plasma factor VIII and the von Willebrand factor, by releasing them from vascular endothelium via a V_2 receptor. Hepatic glycogenolysis is also promoted by high concentrations of vasopressin via a V_1 hepatic receptor.

The physiology of sodium homeostasis

The volume of the extracellular compartment is determined by its sodium content, changes in which result in alterations in blood and interstitial volumes (Fig. 1) with little influence on sodium concentration. The reason is that any rise in sodium concentration causes transient stimulation of thirst and increased water intake, as well as increased vasopressin secretion and reduced renal water excretion, both leading to an increase in body water and a return of extracellular sodium concentration to normal. It therefore appears that extracellular osmolality is conserved at the expense of volume in healthy adults, this integration illustrating the close links between sodium and water homeostatic mechanisms.

Sodium intake

There is little regulation of sodium intake in humans, although some animals demonstrate a specific sodium appetite and have sodium receptors in the hypothalamus. Sodium balance is maintained largely by the kidney, which is normally capable of controlling sodium excretion over a very wide range, 1 to 5000 mmol/24 h. In Western countries, including Britain, the usual intake of sodium is grossly in excess of body needs, being about 100 to 200 mmol/24 h. There is little sodium loss from the healthy bowel and in temperate climates sweating is minimal (the sodium concentration in sweat is 40 to 50 mmol/litre). Most people are therefore at continuous

risk of sodium excess, which is prevented by the kidney.

Control of renal sodium excretion

The glomerular filtration rate of normal kidneys is 170 litres per 24 h, the filtrate containing 140 mmol of sodium per litre. Most of the filtered sodium (60–70 per cent) is reabsorbed iso-osmotically by the proximal tubule. Much of the remainder is reabsorbed in the medullary thick ascending limb of the loop of Henle and the distal nephron, such that only a small fraction of the load of sodium filtered at the glomerulus is excreted in the urine (0.1 to a few per cent).

At the single nephron level the glomerulotubular feedback mechanism operates to maintain a balance between the amount of sodium and fluid filtered by the glomerulus and the reabsorptive function of the corresponding nephron. The mechanisms responsible for glomerulotubular feedback are not known precisely, but the macula densa cells of the thick ascending limb detect changes in composition of tubular fluid entering the terminal portion of the thick ascending limb and transmit signals (possibly intrarenal angiotensin II) to modulate glomerular vascular resistance and glomerular pressure. Thus, acute changes in glomerular filtration determine appropriate changes in sodium reabsorption in the proximal tubule, such that if glomerular filtration rate increases, then reabsorption increases proportionally, and vice versa.

Most regulation of sodium balance occurs in the distal nephron, where 'fine tuning' of sodium excretion occurs under the control of a variety of mechanisms.

Renin–angiotensin–aldosterone system

In addition to their role in the glomerulotubular feedback mechanism, the macula densa cells also influence the juxtaglomerular cells of the renal afferent arterioles to synthesize and secrete renin into the systemic circulation if sodium delivery to the distal nephron drops. In addition, reductions of renal perfusion pressure appear to directly increase renin secretion, while baroreceptors within the great veins of the chest influence renin secretion via the sympathetic nerves. Renin catalyses the conversion of angiotensinogen to angiotensin I (Fig. 5), which is then converted into the highly active octapeptide angiotensin II. The activity of angiotensin III is only 30 per cent that of angiotensin II. These peptides influence body sodium content and extracellular fluid volume in a number of ways. Angiotensin II is a potent vasoconstrictor which readily increases systemic blood pressure, stimulates the secretion of aldosterone from the zona glomerulosa of the adrenal cortex to enhance sodium reabsorption in the distal nephron, increases cardiac contractility, stimulates thirst, and is involved in the glomerulotubular feedback mechanism.

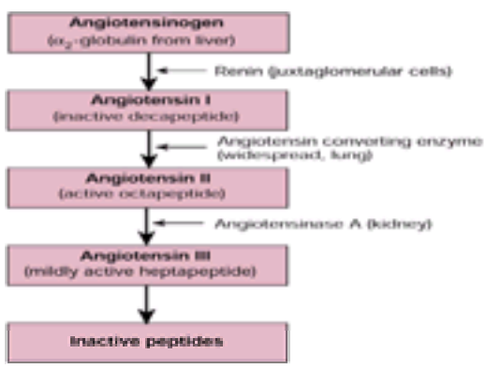

Fig. 5 The renin–angiotensin–aldosterone system. The enzyme, renin, secreted by juxtaglomerular cells of the renal afferent arterioles converts angiotensinogen to angiotensin I (inactive). The active peptides, angiotensin II and III, are potent vasoconstrictors and stimulate aldosterone secretion from the adrenal cortex to expand blood volume and raise blood pressure.

Atrial natriuretic peptide

Following the observation that expansion of blood volume caused a rise in renal sodium excretion that could not be accounted for by inhibition of the renin–angiotensin–aldosterone system, specific humoral natriuretic factors were proposed. One that has been well characterized is human alpha atrial natriuretic peptide, which is synthesized and secreted primarily by cardiac atrial myocytes. Moderate increases in blood volume and postural changes cause atrial distension, directly stimulating atrial natriuretic peptide. Its most important actions are on the kidney, where at physiological concentrations it causes a modest natriuresis, minimal diuresis, and reduces plasma renin activity, and plasma aldosterone concentration. In addition, atrial natriuretic peptide acts as a vasodilator agent, reducing systolic blood pressure and cardiac contractility, and is a potent inhibitor of aldosterone synthesis and release. Indeed, many of the effects of atrial natriuretic peptide could be explained by antagonism of the renin–angiotensin–aldosterone system, but this is not its mechanism of action: specific receptors for atrial natriuretic peptide have been identified in the kidney—on cortical glomeruli, the inner medulla, the vasa rectae in the outer medulla, and the collecting duct. Brain natriuretic peptide has similar actions.

Atrial natriuretic peptide and brain natriuretic peptide play minor roles in regulating extracellular sodium content and volume, but do counterbalance to some extent the actions of the renin–angiotensin–aldosterone system. It is likely that there are other natriuretic factors: ouabain-like substances that inhibit renal sodium–potassium ATPase activity have been described, but they have not been characterized fully and their possible significance for sodium homeostasis remains unknown. Furthermore, there are numerous other intrarenal factors that influence renal sodium excretion, including dopamine, prostaglandins (particularly prostaglandin E_2), and the kallikrein–kinin system.

Summary

In health, the control of renal sodium excretion and extracellular volume is multifactorial, complex, and not completely understood: so also in disease. It is not therefore surprising that it can sometimes be difficult to determine precisely why a particular patient has developed a particular disorder of extracellular volume or serum sodium concentration at a particular time. The physician with a good grasp of the underlying pathophysiological mechanisms stands the best chance of making the correct diagnosis.

Disorders of water and salt homeostasis

The polyuric states

Polyuria describes excessive urinary volume, normal being less than 2.5 to 3.0 litres per 24 h. This can occur due to solute diuresis, the commonest cause being hyperglycaemia in poorly controlled diabetes mellitus, but when the urine is hypotonic three basic pathogenetic mechanisms can account for polyuria. First, lack of osmoregulated vasopressin secretion, termed cranial, central, hypothalamic, or neurogenic diabetes insipidus. Second, reduction in responsiveness of the renal tubules to adequate vasopressin, called nephrogenic or vasopressin-resistant diabetes insipidus. Third, persistent excessive intake of fluid due to inappropriate thirst or drinking behaviour, known as primary polydipsia or dipsogenic diabetes insipidus.

Cranial diabetes insipidus

Cranial diabetes insipidus is a disorder of urinary concentration that is due to decreased secretion of osmoregulated vasopressin. At least 80 per cent of vasopressin-synthesizing neurones must be destroyed before overt clinical features become manifest. Cranial diabetes insipidus is rare, with an estimated prevalence of 1 in 25 000 and equal gender distribution.

Aetiology

The causes of cranial diabetes insipidus are given in Table 1.

Familial varieties account for 5 per cent of cases. Autosomal dominant familial cranial diabetes insipidus is caused by mutations of the arginine vasopressin gene located on chromosome 20. Typically the onset is in early childhood (2 to 7 years) and not infancy. A variety of different mis-sense and non-sense mutations and deletions have been identified in numerous kindreds. Mutant arginine vasopressin precursors accumulate in the magnocellular neurones where they are neurotoxic.

Approximately 30 per cent of acquired cases of cranial diabetes insipidus are idiopathic. One-third of these have circulating antibodies to the hypothalamic neurones that produce vasopressin, suggesting an autoimmune aetiology, supported by lymphocytic infiltration of the neurohypophysis that leads to thickening of the pituitary stalk. Trauma to the hypothalamus or pituitary stalk is a frequent cause of cranial diabetes insipidus, but the trans-sphenoidal surgical approach to the pituitary is less traumatic than the transfrontal approach, and rarely causes permanent cranial diabetes insipidus. Head injury may cause cranial diabetes insipidus, with some patients following a triple-phase response to trauma characterized by initial polyuria for a few hours or days, followed by antidiuresis for a variable period, then progressing to permanent polyuria. Primary pituitary tumours rarely cause cranial diabetes insipidus. Germinoma is a common cause of cranial diabetes insipidus in childhood.

Clinical features

The main clinical manifestations of cranial diabetes insipidus are polyuria, nocturia, and excessive thirst and drinking. Children may present with enuresis. Most patients have partial deficiency of vasopressin. Urine volumes range between 3 and 25 litres per 24 h, with random urine osmolalities of 50 to less than 300 mosmol/kg and plasma osmolality within the normal reference range.

Patients with cranial diabetes insipidus maintain normal values of plasma osmolality and serum sodium because they have an intact osmoregulated thirst mechanism. Defective thirst or restricted access to water leads to hypernatraemia and hyperosmolality. With severe polyuria the slightest obstruction to outflow from the urinary tract can lead to hydronephrosis and hydroureter.

Cranial diabetes insipidus may be masked by deficiency of glucocorticoid hormone due either to hypopituitarism or primary adrenal failure because cortisol is necessary for the maximal dilution function of the distal nephron and for normal secretion of arginine vasopressin. The symptoms of partial cranial diabetes insipidus are often worse in pregnancy due to the increase in metabolic clearance of arginine vasopressin caused by cysteine aminopeptidase (vasopressinase), a circulating enzyme of placental origin.

Nephrogenic diabetes insipidus

In nephrogenic diabetes insipidus the renal tubules are partially (most often) or totally resistant to the action of vasopressin.

Aetiology

Table 1 lists the causes of nephrogenic diabetes insipidus.

The X-linked form is rare. Infant males have profound polyuria, dehydration, vomiting, fever, irritability, and fail to thrive. Females, when tested, have slightly impaired urinary concentration. Molecular studies of kindreds with X-linked nephrogenic diabetes insipidus have identified mutations or deletions of the gene that encodes for the V_2 receptor located on Xq28. The V_2 receptor is a classic seven-domain transmembrane protein: genetic abnormalities have been demonstrated in external and internal segments of the receptor as well as the transmembrane portions.

Approximately 10 per cent of cases of familial nephrogenic diabetes insipidus are due to genetic defects of aquaporin 2, the water-channel protein that is encoded on chromosome 12q13. Mutant aquaporin 2 is misrouted within the distal tubule and fails to be inserted into the cellular membrane. Inheritance is autosomal recessive.

Hypercalcaemia-induced nephrogenic diabetes insipidus is thought to be due to a combination of factors: reduced medullary hyperosmolality and adenyl cyclase activity, dysfunction of aquaporin 2, and calcium deposition with scarring of the kidney. The effect of sustained hypokalaemia on renal function is complex: it inhibits sodium–potassium cotransport in the thick ascending limb, reduces adenyl cyclase activity, increases intrarenal prostaglandin synthesis (which blunts the antidiuretic effect of vasopressin) and may reduce intracellular protein kinase function. Aquaporin 2 trafficking is reduced. Reversal of these metabolic derangements often returns renal function to normal, but does not always do so.

A third of patients taking long-term lithium carbonate develop nephrogenic diabetes insipidus. Lithium blunts the generation and action of cyclic 5'AMP in the distal nephron and may reduce osmoregulated vasopressin secretion and/or stimulate thirst. Demeclocycline also inhibits the generation and function of cyclic 5'AMP.

Clinical features

Adults with nephrogenic diabetes insipidus usually have partial nephrogenic diabetes insipidus with mild symptoms. Similar to patients with cranial diabetes insipidus, these individuals have serum sodium and plasma osmolality within the normal range, but low urine osmolality (< 300 mosmol/kg). The familial forms present soon after birth with profound polyuria, dehydration, and fever (see above). Infants may become hyper-natraemic due to inadequate fluid intake.

Primary polydipsia

Some patients drink copious quantities of fluid, well in excess of body requirements, for reasons that are ill understood. This condition is termed primary polydipsia, or dipsogenic diabetes insipidus.

Aetiology

Many patients have a psychological disturbance leading to compulsive drinking, some of whom have a lowered osmotic thirst threshold but a normal threshold for vasopressin release. Up to 20 per cent of patients with chronic schizophrenia have primary polydipsia. Very rarely a structural hypothalamic lesion (for example sarcoidosis) is believed to be the cause of primary polydipsia (Table 1). Some drugs cause a dry mouth, thus stimulating thirst.

Clinical features

Although the clinical manifestations of primary polydipsia are similar to cranial diabetes insipidus and nephrogenic diabetes insipidus, nocturia is less of a feature: patients with primary polydipsia tend to sleep through the night. Individuals with primary polydipsia lower plasma osmolality sufficiently to suppress vasopressin secretion to allow polyuria. Their serum sodium therefore tends to be lower than that of patients with cranial diabetes insipidus and nephrogenic diabetes insipidus, but usually remains within the reference range. The fact that many patients with primary polydipsia can drink up to 20 litres in 24 h and still remain normonatraemic is testament to the remarkable effectiveness of homeostatic mechanisms.

Diagnostic evaluation of the polyuric patient

Before embarking on expensive and time-consuming tests, it is always wise to establish that the urine volume is in excess of 3 litres per 24 h. Urine output less than this with a normal serum sodium and plasma osmolality excludes significant disturbance of water balance. Routine biochemical investigation of glucose, calcium, and potassium may point towards some causes of polyuria (Table 1). Three types of specialized diagnostic tests are available: dehydration tests, measurement of plasma vasopressin after dehydration or osmotic stimulation, and therapeutic trial of desmopressin.

Dehydration tests

These can aid the diagnosis of severe forms of cranial diabetes insipidus and nephrogenic diabetes insipidus. Many protocols have been described: all are based on observing the responses in urine and blood to a period of fluid deprivation, followed by noting the ability to concentrate urine after exposure to exogenous vasopressin (for example desmopressin). A typical commonly used test is (briefly) as follows. The patient is encouraged to drink as usual during the night before the test which is to start in the morning. Basal measurements of urinary volume and osmolality and of plasma osmolality are made, and the patient weighed. All fluid is then withheld for 8 h, with the patient weighed and urine and blood samples taken every 1 to 2 h. The test must be stopped if the patient loses in excess of 5 per cent of their initial body weight. Thereafter 2 µg of desmopressin is injected intramuscularly, the patient being allowed to drink cautiously and eat. Urine samples are collected over the following 16 h.

A guide to interpretation of the results of the water deprivation test is given in Table 2. Substantial difficulty arises in the differentiation of partial diabetes insipidus disorders from each other, and from primary polydipsia. The reason for this is that prolonged polyuria, irrespective of its cause, leads to a reduction in the maximal concentrating ability of the kidney by removing renal medullary interstitial solute and altering aquaporin 2 function, such that exogenously administered vasopressin cannot elicit its maximal renal effect. Direct measurement of plasma vasopressin aids diagnosis of partial nephrogenic diabetes insipidus in these circumstances.

Response of plasma vasopressin to osmotic stimulation and dehydration

Measurement of plasma vasopressin and osmolality during a 2-h hypertonic (850 mmol/litre) saline infusion at a rate of 0.06 ml/kg/min will diagnose partial or complete cranial diabetes insipidus, as the vasopressin response to osmotic stimulation is subnormal (Fig. 6(a)). Patients with nephrogenic diabetes insipidus or primary polydipsia have results that fall within the normal range.

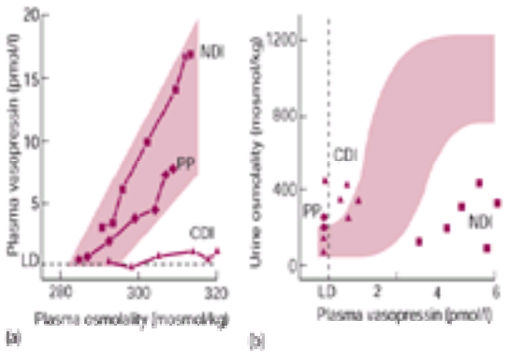

Fig. 6 (a) Relationship between plasma vasopressin and plasma osmolality during hypertonic saline infusion in typical patients with (i) cranial diabetes insipidus (CDI), (ii) nephrogenic diabetes insipidus (NDI), and (iii) primary polydipsia (PP). The shaded area represents the normal response. (b) Relationship between urine osmolality and plasma vasopressin in patients with cranial diabetes insipidus (triangles), nephrogenic diabetes insipidus (squares) and primary polydipsia (circles) after a period of dehydration. The shaded area is the normal relationship under various degrees of hydration. LD represents the limit of detection of the plasma vasopressin assay.

After a period of water deprivation, the measurement of urine osmolality and plasma vasopressin will define nephrogenic diabetes insipidus (Fig. 6(b)), as vasopressin will be inappropriately elevated with respect to the low urine osmolality.

Therapeutic trial of desmopressin

If the water deprivation–desmopressin test gives equivocal results and facilities to measure vasopressin are not available, then a formal therapeutic trial of low-dose desmopressin should be instituted to differentiate the cause of polyuria. The trial must be supervised closely, preferably in hospital, because of the potential hazard of severe water intoxication in those with primary polydipsia. After a basal period of 3 to 4 days, desmopressin (1 µg intramuscularly daily for 10 days) is administered to patients who are weighed and have urine and plasma osmolalities or serum sodium and urine volume measured daily. Patients with cranial diabetes insipidus will be identified by a reduction of thirst, little or no weight gain, a reduction in urine flow, and normal plasma osmolality. Nephrogenic diabetes insipidus is characterized by a lack of response. Primary polydipsia patients remain thirsty, continue to drink, gain weight, and become progressively hyponatraemic.

Having established the pathogenetic mechanism causing polyuria, it is important to search for a specific underlying cause (Table 1). Magnetic resonance imaging or high-resolution computed tomography scans of the pituitary and surrounding structures are invaluable in those with cranial diabetes insipidus. Patients with cranial diabetes insipidus frequently lose the posterior pituitary bright spot on T_1-weighted MRI.

Treatment

Cranial diabetes insipidus

Mild forms of cranial diabetes insipidus (urine output less than 4 litres/24 h) may not require any specific therapy other than advice to drink sufficient quantities to quench thirst. Such patients can, however, get into difficulty if they are unable to get and retain an adequate fluid intake for any reason. In more severe forms, the drug of choice is desmopressin, a synthetic vasopressin V_2-receptor agonist analogue possessing potent antidiuretic, but no pressor, activity, and with a prolonged duration of action. Desmopressin is administered orally, intranasally by spray or tube, or parenterally. There are wide individual variations in the dose required to control symptoms. Requirements for oral desmopressin range from 50 µg to 1200 µg daily; intranasal from 2.5 µg to 120 µg daily; and parenteral up to 2 µg intramuscularly daily. Dilutional hyponatraemia is a potential hazard if desmopressin is given in excess for a prolonged period: this can be avoided by instructing the patient to forgo the drug for 1 day each week. Other side-effects are minimal. Desmopressin is a safe drug in pregnancy, and is resistant to the circulating placental enzyme, vasopressinase.

Lysine vasopressin, given intranasally, is a shorter-acting alternative but as it possesses pressor activity it can cause intestinal and/or renal colic, increase blood pressure and induce coronary artery vasospasm. Pitressin is rarely used nowadays because of similar pressor side-effects.

Chlorpropamide, clofibrate, carbamazepine, and thiazide diuretics have been used either singly or in combination to reduce urine volume by up to 50 per cent, but they are rarely prescribed these days because of their side-effects and the efficacy of desmopressin.

Nephrogenic diabetes insipidus

Correction of the underlying cause of acquired nephrogenic diabetes insipidus (for example removal of drug or correction of hypercalcaemia) may allow recovery of renal concentrating ability. If matters do resolve, this typically takes a number of weeks.

Severe polyuria of the familial form of nephrogenic diabetes insipidus can be reduced by about 50 per cent using a combination of salt restriction, thiazide and/or amiloride diuretics, and a prostaglandin synthetase inhibitor (indomethacin, 1.5–3.0 mg/kg/day). A promising new therapeutic approach is the combination of a thiazide, indomethacin, and desmopressin, which may reduce urine output by up to 80 per cent.

Primary polydipsia

There is no efficacious drug treatment available for primary polydipsia although propranolol in doses up to 120 mg daily has been recommended to reduce thirst. Therapy directed towards underlying psychiatric problems may prove helpful. Clozapine has reduced polydipsia associated with hyponatraemia in those with chronic

schizophrenia.

Hyponatraemic states

Hyponatraemia, defined as a serum sodium less than 130 mmol/litre, is a common electrolyte disturbance, affecting up to 5 per cent of hospital patients. Severe hyponatraemia (serum sodium < 115 mmol/litre) is rare (< 0.5 per cent).

Pseudohyponatraemia

Spuriously low measurements of serum sodium can occur in patients with very high circulating concentrations of lipids or proteins because the volume of these substances contributes substantially to serum volume. The concentration of sodium in the water phase of blood remains normal, hence plasma osmolality is normal and its measurement proves an easy diagnostic test. Pseudohyponatraemia also arises with severe hyperglycaemia, although the mechanism is different: high blood glucose concentration draws intracellular water into the extracellular space, resulting in hyponatraemia. Plasma osmolality will be elevated due to the hyperglycaemia.

Classification and causes of hyponatraemia

In all hyponatraemic states there is an excess of extracellular water relative to the total sodium content of the extracellular compartment. The sodium content, however, can vary markedly, such that patients can be divided into three groups, forming the basis of a classification of hyponatraemia. Total extracellular sodium quantity can be:

- lower than normal, resulting in extracellular hypovolaemia,
- normal, with slightly expanded extracellular volume (not clinically evident), or
- higher than normal causing extracellular hypervolaemia.

Significant extracellular volume changes can be detected clinically:

- hypovolaemia leads to thirst, reduced skin turgor, tachycardia, low jugular venous pressure, and postural hypotension (or supine hypotension in severe cases), and
- hypervolaemia leads to dependent oedema, also possibly to elevation of jugular venous pressure, pulmonary oedema, and ascites.

Minor volume changes are difficult to assess clinically and there are no simple quick diagnostic tests to aid classification.

Table 3 presents the classification and pathogenesis of hyponatraemia. Measurement of urinary sodium helps diagnosis. The majority of hyponatraemic patients have urine osmolalities in excess of 300 mosmol/kg and, of course, plasma is hypo-osmolar (< 280 mosmol/kg).

Large sodium losses and hypovolaemic hyponatraemia commonly occur with persistent vomiting and/or diarrhoea, extensive skin burns, and excessive prolonged sweating. The healthy kidney will conserve sodium and urinary concentration will be less than 10 mmol/litre. Renal sodium loss leading to hyponatraemia can be due to renal diseases, typically those affecting the renal medulla (analgesic nephropathy, chronic pyelonephritis, polycystic kidneys, recovery from acute tubular necrosis, or post bilateral ureteric obstruction), mineralocorticoid deficiency (Addison's disease, hyporeninaemic hypoaldosteronism), or diuretic excess.

Normovolaemic hyponatraemia is usually due to the syndrome of inappropriate antidiuresis (see below) or inappropriate administration of intravenous fluid (for example 5 per cent dextrose solutions) in the postoperative period. Rarely it may be caused by isolated glucocorticoid deficiency (for example partial hypopituitarism) or severe prolonged hypothyroidism. Beer drinker's potomania occurs in some individuals who drink excessive volumes of beer over short periods, for example 10 litres in 6 h, which overwhelms the kidney's capacity to excrete water.

Hypervolaemic hyponatraemia is commonly observed in severe heart failure, decompensated cirrhosis, and nephrotic syndrome. In these disorders glomerular filtration is reduced and proximal tubular sodium reabsorption increased. Renal afferent arteriole perfusion falls leading to increased circulating angiotensin II concentrations, contributing to stimulation of thirst. Furthermore, non-osmotic release of vasopressin also contributes to water retention.

Clinical features

In addition to the features associated with extracellular (and therefore blood) volume reduction or expansion described above, there are clinical manifestations due to hyponatraemia *per se* (Table 4). The severity of hyponatraemic symptoms depends upon both the absolute serum sodium concentration and its rate of fall. Chronic mild hyponatraemia (serum sodium 120–130 mmol/litre) is often totally asymptomatic; but a sudden fall to only 125 mmol/litre from normal values (usually iatrogenic) can cause convulsions.

General principles of management

Specific therapy for mild hyponatraemia is often not necessary: treatment should be reserved for symptomatic or severe life-threatening hyponatraemia. Treatment of the underlying cause is obviously essential and will frequently correct the serum sodium concentration.

For hypovolaemic hyponatraemia, volume replacement is mandatory. Infusion of isotonic saline is usually sufficient, but occasionally intravascular volume expanders are required to raise blood pressure, particularly in an acute situation. Immediate hydrocortisone and subsequently fludrocortisone are essential to treat Addison's disease.

Treatment of normovolaemic hyponatraemia is described in the section on the syndrome of inappropriate antidiuresis.

Hyponatraemia associated with hypervolaemic disorders responds to single or combined therapy with potent diuretic drugs to remove extracellular sodium, inhibitors of angiotensin-converting enzyme to lower angiotensin II values, and water restriction to less than 1 litre per 24 h.

Irrespective of the cause, chronic severe hyponatraemia (defined as serum sodium < 120 mmol/litre lasting more than 3 days) should be corrected slowly (i.e. at a rate of less than 0.5 mmol/litre/h). Infusion of hypertonic saline should be avoided, but if used the rate of infusion should increase serum sodium by no more than 0.5 mmol/litre/h (10 mmol/litre/24 h) and infusion stopped when serum sodium reaches 120 mmol/litre. A convenient formula determines the likely rate of change of serum sodium concentration following infusion of hypertonic saline and should always be used in patient care:

$$\text{rate of infusion of 3 per cent NaCl solution (ml/h)} = \text{body weight (kg)} \times \text{desired rate of correction of serum sodium (mmol/litre/h)}.$$

Acute symptomatic severe hyponatraemia (developed in less than 3 days; with drowsiness, convulsions, or coma) can be corrected more quickly, with a rate of serum sodium increase of up to 2 mmol/litre/h, but again infusion should stop at a serum sodium concentration of 120 mmol/litre. Regrettably, this condition is almost always iatrogenic, most commonly arising when young women recovering from surgery are ill-advisedly given large quantities of 5 per cent dextrose by intravenous infusion.

Central pontine myelinolysis

The first case of central pontine myelinolysis was described in a young alcoholic. Following many similar cases, it was initially assumed that the condition was due to some form of nutritional deficiency. Association with hyponatraemia was then reported, and the view that this (or its subsequent management) might cause central pontine myelinolysis (also known as osmotic demyelination syndrome) was supported by the finding that dogs made hyponatraemic by repeated injections of vasopressin and intraperitoneal infusions of water, then given hypertonic saline, became quadriparetic with brainstem lesions indistinguishable from those seen in the human condition.

There is undoubtedly a high morbidity and mortality in those with serum sodium concentrations of less than 110 mmol/litre, but debate continues as to whether this is caused by the hyponatraemia itself, or by overzealous treatment of that hyponatraemia. The most feared outcome remains the neurological sequelae of cerebral demyelination, thought to result from large shifts of intracellular water, which occur outside the brainstem as well as in the pons. In most cases reported (if not all), serum sodium has rapidly been corrected to normal levels, hence the recommendation given above that if hypertonic saline is used to correct hyponatraemia, then the infusion must be stopped when serum sodium rises to 120 mmol/litre, allowing more gradual correction from that point onward.

Neurological signs usually develop 2 to 4 days after rapid correction of hyponatraemia. Typical features are quadriplegia and pseudobulbar palsy: these can take the form of a 'locked in' syndrome of mutism with paralysis.

Syndrome of inappropriate antidiuresis

This syndrome is due to inappropriate secretion of vasopressin and is the commonest cause of normovolaemic hyponatraemia. The diagnosis of syndrome of inappropriate antidiuresis is established by ensuring that all the syndrome's criteria are fulfilled (Table 5). This is important: too often the diagnosis is incorrectly claimed on the basis that the first two criteria are met. If they are, this establishes that the level of vasopressin is inappropriate to plasma osmolality, but for the term to be properly applied the level must also be inappropriate to the intravascular volume (it is appropriate to have a high vasopressin level in the context of volume depletion/hypotension). Measurement of urinary sodium is essential: this is persistently elevated in the range 50 to 70 mmol/litre; if it is low, then this suggests that the kidney is attempting to conserve sodium, either due to volume depletion or a sodium-retaining state. The final two criteria specifically exclude those with hypo- and hypervolaemic states, and make the point that the diagnosis can only be confidently made if renal and adrenal function are normal.

Plasma vasopressin estimations are unhelpful in differentiating the syndrome of inappropriate antidiuresis from other causes of hyponatraemia, because the majority of all hyponatraemic states (> 95 per cent) have detectable or elevated values, due to non-osmotic release of the hormone. Persistent circulating vasopressin causes the relative excess of water in all types of hyponatraemia.

Pathophysiology and causes of the syndrome of inappropriate antidiuresis

A very large number of disorders have been associated with the syndrome of inappropriate antidiuresis, some of which are listed in Table 6. In brief, they include a variety of neoplastic conditions, the commonest being small cell carcinoma of the bronchus, non-malignant chest diseases including infections, neurological disorders (infective and vascular), drugs (cytotoxic agents, chlorpropamide, carbamazepine, antidepressants, oxytocin, and thiazide diuretics), recreational agents ('Ecstasy'), and a miscellaneous group (porphyria, cortisol deficiency, idiopathic).

The persistent natriuresis, central to the diagnosis of syndrome of inappropriate antidiuresis, can be explained, in part, by the expanded total body water which is not clinically detectable, causing a reduction in aldosterone production, an increase in circulating natriuretic factors, and a decrease in proximal sodium reabsorption.

Treatment of the syndrome of inappropriate antidiuresis

Identification and successful treatment of the underlying cause of the syndrome of inappropriate antidiuresis will usually correct hyponatraemia. If chronic symptomatic or life-threatening hyponatraemia remains, specific measures to remove the excess total body water are required.

Fluid restriction to 500 ml per 24 h to increase serum sodium to about 130 mmol/litre remains the therapy to be tried first. If this approach is unsatisfactory, additional methods to remove water are justified, the most successful of which is the induction of partial nephrogenic diabetes insipidus with demeclocycline (600 to 1200 mg daily in divided doses), but the maximal effect may take 2 weeks to achieve. It is preferable to lithium carbonate, which although inducing nephrogenic diabetes insipidus is more nephrotoxic. An alternative approach is the administration of furosemide (frusemide) (40–80 mg daily) in combination with oral sodium chloride supplementation (3 g daily). Phenytoin has occasionally proved helpful by suppressing inappropriate neurohypophyseal vasopressin secretion. Infusion of isotonic or hypertonic solutions of saline are not advised because of the real danger that rapid increase in serum sodium concentration might cause the osmotic demyelination syndrome (see above).

The most logical therapy for the syndrome of inappropriate antidiuresis would be a V_2-receptor antagonist. A recently synthesized, linear non-peptide V_2-receptor antagonist, OPC-31260, increases solute-free water excretion. Clinical trials of the treatment of syndrome of inappropriate antidiuresis with this agent are encouraging. This new class of drugs, called aquaretics, should improve management of the syndrome of inappropriate antidiuresis, and other hyponatraemic states associated with excess arginine vasopressin.

Sick cell concept

This concept was formulated following the observation that hyponatraemia developed quickly in severe trauma or overwhelming infection in humans or animals, and in malnourished very ill patients. It is classified as a cause of normovolaemic hyponatraemia. There is a shift of intracellular water into the extracellular compartment due to reduction of intracellular solute either by leakage across a damaged cell membrane, enhanced intracellular catabolism, or possibly due to movement of sodium into the cell. There is no specific therapy other than treatment of the underlying cause.

Hypernatraemic states and thirst deficiency

Hypernatraemia, defined as a serum sodium concentration greater than 150 mmol/litre, is less common than hyponatraemia.

Aetiology and pathophysiology

Hypernatraemia can be classified into two categories dependent on extracellular volume states (Table 7).

Hypervolaemic hypernatraemia is caused by extracellular sodium excess, usually as a result of accidental iatrogenic overdoses of sodium-containing preparations.

Acute hypovolaemic hypernatraemia occurs when patients lose large quantities of hypotonic fluid (for example gastrointestinal). Chronic hypovolaemic hypernatraemia is the result of prolonged water deficit, usually the result of impaired or absent thirst, hypodipsia, or adipsia, which implies a lesion of the thirst osmoreceptor. It is sometimes associated with abnormal osmoregulated vasopressin secretion. Four patterns of osmoregulatory dysfunction have been described (Fig. 7):

- Type A, shows elevation of osmotic thresholds for both thirst and vasopressin. It has been termed 'essential' hypernatraemia; patients continue to dilute and concentrate urine normally, but do so around a higher serum sodium concentration.
- Type B is characterized by decreased sensitivity (slope) of the vasopressin and thirst osmoregulatory lines.
- Type C, the most serious defect, termed adipsic hypernatraemia, is due to complete destruction of both thirst and arginine vasopressin osmoreceptors. Patients never experience the desire to drink, even when serum sodium reaches values as high as 190 mmol/litre.
- Type D is very rare, with selective complete loss of thirst but vasopressin osmoregulation remaining normal.

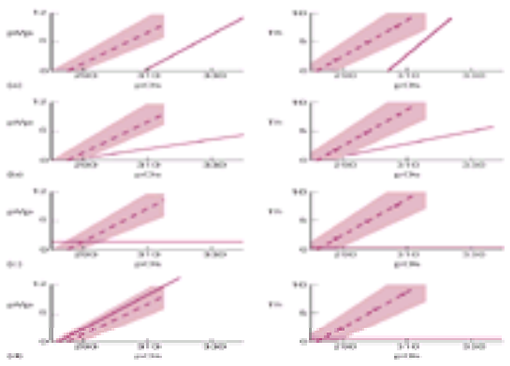

Fig. 7 Patterns of osmoregulated thirst (Th) and plasma vasopressin concentration (pVp) in hypodipsic or adipsic hypernatraemia. The units of pVp are pmol/l and of thirst are 0 to 10 on a visual analogue scale. Stippled areas are the normal responses to increases in plasma osmolality (pOs) and the dashed lines represent mean osmoregulatory lines: (a) reset thirst and vasopressin thresholds or 'essential' hypernatraemia; (b) decreased sensitivity of thirst and vasopressin release; (c) complete destruction of both thirst and vasopressin osmoreceptors; (d) absent thirst osmoregulation with normal osmoregulated vasopressin release. See text for further explanation. (Reproduced from Baylis PH and Thompson CJ (1988). Osmoregulation of vasopressin secretion and thirst in health and disease. *Clinical Endocrinology* **29**, 549–76, with permission.)

Clinical features

Hypervolaemic hypernatraemia causes severe thirst, irritability, and hypotonia and may lead to convulsions, seizures, and death. Clinical manifestations of hypovolaemic hypernatraemia relate to extracellular and intracellular fluid loss, the striking feature being lack of thirst. The slow development of hypernatraemia is often associated with minimal symptoms of confusion or drowsiness.

Treatment of hypernatraemia

Patients require water to lower serum sodium concentration slowly. The safest route of administration is oral, but unconscious patients will require infusion of 5 per cent dextrose solutions. Care must be taken to avoid rapid falls in serum sodium, with the rate of decline no greater than 10 mmol per 24 h.

'Essential' hypernatraemia (Fig. 7, type A) requires little specific therapy as patients are protected from extremes of hypernatraemia. Patients with total loss of thirst and vasopressin osmoregulation pose major management problems. They should be instructed to drink a daily volume of about 2 litres, which should be adjusted according to changes in daily body weight. Desmopressin may be required. Regular checks of serum sodium concentration are essential to avoid wide fluctuations. Constant vigilance is necessary to maintain water balance in chronic hypodipsic or adipsic patients.

Further reading

Anderson RJ *et al*. (1985). Hyponatremia: a prospective analysis of its epidemiology and the pathogenetic role of vasopressin. *Annals of Internal Medicine* **102**,164–8.

Arieff AL, Guisado R (1976). Effects on the central nervous system of hypernatremic and hyponatremic states. *Kidney International* **10**, 104–16.

Ball SG, Vaidja B, Baylis PH (1997). Hypothalamic adipsic syndrome: diagnosis and management. *Clinical Endocrinology* **47**, 405–9.

Bartter FC, Schwartz WB (1967). The syndrome of inappropriate secretion of antidiuretic hormone. *American Journal of Medicine* **42**, 790–806.

Baylis PH, Cheetham T (1998). Diabetes insipidus. *Archives of Disease in Childhood* **79**, 84–9.

Baylis PH, Robertson GL (1980). Plasma vasopressin response to hypertonic saline to assess posterior pituitary function. *Journal of the Royal Society of Medicine* **73**, 255–60.

Baylis PH, Thompson CJ (1988). Osmoregulation of vasopressin secretion and thirst in health and disease. *Clinical Endocrinology* **29**, 549–76.

Berl T *et al*. (1976). Clinical disorders of water metabolism. *Kidney International* **10**, 117–32.

Bibi D *et al*. (1999). Treatment of central pontine myelinolysis with therapeutic plasmaphoresis. *The Lancet* **353**, 1155.

Bichet DG *et al*. (1994). Nature and recurrence of AVPR2 mutations in X-linked nephrogenic diabetes insipidus. *American Journal of Human Genetics* **55**, 278–86.

Charmondari E, Brook CGD (1999). 20 years of experience of idiopathic central diabetes insipidus. *The Lancet* **353**, 2212–13.

Davison JM *et al*. (1988). Serial evaluation of vasopressin release and thirst in human pregnancy. *Journal of Clinical Investigation* **81**, 798–806.

De Bellis A *et al*. (1999). Longitudinal study of vasopressin cell antibodies, posterior pituitary function and magnetic resonance imaging evaluations in subclinical autoimmune central diabetes insipidus. *Journal of Clinical Endocrinology and Metabolism* **84**, 3047–3051.

Deen PMT *et al*. (1994). Requirement of human renal water channel aquaporin-2 for vasopressin dependent concentration of urine. *Science* **264**, 92–5.

De Zeeuw D, Janssen WMT, de Jong PE (1992). Atrial natriuretic factor: its (patho)physiological significance in humans. *Kidney International* **41**, 1115–33.

Flear CTG, Gill GV, Burn J (1981). Hyponatraemia: mechanisms and management. *The Lancet* **i**, 26–31.

Kenyon CJ, Jardine AG (1989) Atrial natriuretic peptide: water and electrolyte homeostasis. *Baillière's Clinics in Endocrinology and Metabolism* **3**, 431–50.

Martin P-Y, Schrier RW (1998). Role of aquaporin-2 water channels in urinary concentration and dilution defects. *Kidney International* **53** (Supplement 65), 557–62.

McKenna K, Thompson C (1998). Osmoregulation in clinical disorders of thirst and thirst appreciation. *Clinical Endocrinology* **49**, 139–52.

Miller WL (1993). Molecular genetics of familial central diabetes insipidus. *Journal of Clinical Endocrinology and Metabolism* **77**, 592–5.

Mitchell KD, Navar LG (1989). The renin-angiotensin-aldosterone system in volume control. *Baillière's Clinics in Endocrinology and Metabolism* **3**, 393–430.

Robertson GL, Shelton RL, Athar S (1976). The osmoregulation of vasopressin. *Kidney International* **10**, 25–37.

Robertson GL (1995). Diabetes insipidus. *Endocrinology and Metabolism Clinics of North America* **24**, 549–72.

Saito T *et al*. (1997). Acute aquaresis by the non-peptide arginine vasopressin (AVP) antagonist OPC-31260 improves hyponatraemia in patients with the syndrome of inappropriate secretion of antidiuretic hormone. *Journal of Clinical Endocrinology and Metabolism* **82**, 1054–7.

Siggaard C *et al*. (1999). Clinical and molecular evidence of abnormal processing and trafficking of the vasopressin preprohormone in a large kindred with familial neurohypophyseal diabetes insipidus due to a signal peptide mutation. *Journal of Clinical Endocrinology and Metabolism* **84**, 2933–41.

Sterns RH, Riggo J, Schochet SS (1986). Osmotic demyelination syndrome following correction of hyponatremia. *New England Journal of Medicine* **314**, 1535–42.

Strom TM *et al*. (1998). Diabetes insipidus, diabetes mellitus, optic atrophy and deafness (DIDMOAD) caused by mutations in a novel gene (wolframin) coding for a predicted transmembrane protein. *Human Molecular Genetics* **7**, 2021–8.

Thompson CJ *et al*. (1986). The osmotic thresholds for thirst and vasopressin release are similar in healthy man. *Clinical Science* **71**, 651–6.

Thrasher TN, Keil LC, Ramsay DJ (1982). Lesions of the organum vasculosum of the lamina terminalis (OVLT) attenuate osmotically induced drinking and vasopressin secretion in dogs. *Endocrinology* **110**, 1837–41.

Verbalis JG (1989). Hyponatraemia. *Baillière's Clinics in Endocrinology and Metabolism* **3**, 499–530.

Verbalis JG (1998). Adaptation to acute and chronic hyponatraemia: implications for symptomatology, diagnosis and treatment. *Seminars in Nephrology* **18**, 3–19.

Vokes TJ, Robertson GL (1988). Disorders of antidiuretic hormone. *Endocrinology and Metabolism Clinics of North America* **17**, 281–99.

Walker LA, Valtin H (1982). Biological importance of nephron heterogeneity. *Annual Reviews in Physiology* **44**, 203–19.

Yasui M *et al*. (1997). Adenylate cyclase-coupled vasopressin receptor activates AQP-2 promoter via a dual effect on CRE and AP1 elements. *American Journal of Physiology, Renal Fluid and Electrolyte Physiology* **272**, F443–F450.

20.2.2 Disorders of potassium homeostasis

J. Firth

Potassium homeostasis

Introduction

Potassium is the most abundant cation in the body. Total body potassium ranges between 37 and 52 mmol/kg body weight, and of this 98 per cent is found within cells, where its concentration is in the range 150 to 160 mmol/l. By contrast, the normal range of potassium concentration in serum is from 3.5 to 5.0 mmol/l. The ratio of intracellular to extracellular potassium concentration is a critical determinant of cellular resting membrane potential and thereby of the function of excitable tissues, particularly the nerves and muscles. Potassium tends to leak out of cells through a variety of ion-selective potassium channels found in all cell membranes. The maintenance of the intracellular to extracellular gradient is largely dependent on the ubiquitous Na^{+},K^{+}-ATPase enzyme, which pumps two potassium ions into the cell for every three sodium ions extruded.

The mechanisms of potassium homeostasis can be considered in terms of internal balance (the relationship between intracellular and extracellular potassium concentration) and external balance (which determines total body potassium).

Internal balance

A wide variety of factors modulate the distribution of potassium between the intracellular and extracellular fluid compartments. These factors either alter the function of the Na^{+},K^{+}-ATPase or the rate of efflux of potassium from cells, which together dictate intracellular potassium concentration. In view of the importance of the ratio of internal to external potassium concentration for critical neuromuscular functions, some of these mechanisms serve as essential acute defence mechanisms to counteract life-threatening hyperkalaemia. Factors modulating internal potassium balance are shown in Table 1.

External balance

Dietary potassium intake in Western society typically varies between 50 and 150 mmol/day, but balance can be attained with intake of up to 500 mmol/day if homeostatic mechanisms are intact. In normal circumstances potassium excretion in the stool is not regulated, but it amounts to only 5 to 15 mmol/day. When renal function is compromised, the absolute magnitude as well as the proportion of potassium in the faeces is increased, but variation in renal excretion of potassium is usually the only means by which the body achieves external potassium balance by ensuring that excretion equals intake.

With normal intake of potassium, 10 to 20 per cent of the load filtered at the glomerulus is excreted, but fractional excretion of potassium can vary from 1 per cent when intake is restricted to over 100 per cent when intake is excessive. Micropuncture studies have shown that the amount of potassium reaching the distal convoluted tubule does not vary in these circumstances, indicating that modulation of renal potassium excretion is normally a property of the distal nephron. Factors that modify potassium excretion by the distal nephron are shown in Table 2. These factors are clearly interrelated: it is rare that one is modified in isolation and the overall effect on potassium excretion is almost invariably the aggregate result of several complementary or competing stimuli.

Hypokalaemia

A low serum potassium concentration (3.5 mmol/l or less) is the commonest electrolyte abnormality seen in clinical practice, found in up to 20 per cent of patients in hospital. Most have mild hypokalaemia, with serum potassium in the range 3.0 to 3.5 mmol/l, but 5 per cent have a level lower than 3.0 mmol/l, and 0.03 per cent (more in some series) have very severe hypokalaemia with serum potassium concentration less than 2.5 mmol/l.

Clinical features of hypokalaemia

Patients with mild hypokalaemia often have no symptoms attributable to their low serum potassium concentration. A variety of non-specific symptoms develop with more severe hypokalaemia, including lassitude, generalized weakness, and constipation. At a serum potassium level of less than 2.5 mmol/l serious neuromuscular problems sometimes arise. Rhabdomyolysis (see Chapter 20.4) can occur, and increases in serum creatine phosphokinase activity indicative of muscle injury are frequently detectable in those with a serum potassium concentration below 3.0 mmol/l. Hypokalaemia can cause intestinal ileus, and is particularly likely to do so in the postoperative period when other factors also conspire to prevent normal gut motility. Paralysis of skeletal muscle has been reported, most dramatically in cases of hypokalaemic quadraparesis, which appears to be more common in India than elsewhere. Paraesthesias and tetany have rarely been described.

Hypokalaemia can cause polyuria and polydipsia, also a metabolic alkalosis. Severe prolonged potassium depletion is associated with chronic interstitial nephritis, the presence of renal cysts, and with the development of chronic renal failure. It is not always clear, however, whether hypokalaemia is the cause or effect of this condition.

Hypokalaemia may be suspected from the clinical context (for example the patient taking diuretics or vomiting copiously), but there are no specific physical signs. Alterations induced in the ECG include flattening of the T wave, depression of the S–T segment, and the development of prominent U waves, which can give the impression of a prolonged Q–T interval. These changes, typically observed with a serum potassium concentration lower than 3.0 mmol/l, provide a diagnostic clue to the presence of hypokalaemia, but do not have any serious clinical implications in a patient with a normal heart. However, hypokalaemia can cause problems in those whose heart is abnormal. There is a correlation between hypokalaemia and the development of ventricular tachycardia or ibrillation during the acute phase of myocardial infarction; hypokalaemia can provoke life-threatening arrhythmias in those receiving digoxin; and there is controversy as to whether the mild hypokalaemia often produced by diuretic therapy constitutes a risk factor for sudden cardiac death.

Treatment of hypokalaemia

In emergency

Emergency treatment of hypokalaemia is rarely required. In the rare circumstances of life-threatening cardiac arrhythmia or muscular paralysis, intravenous infusion of potassium (usually potassium chloride) should be given immediately. This must be administered into a central vein (internal jugular, subclavian, or femoral) since solutions containing the necessary high concentration of potassium cause pain and phlebitis if given peripherally, and can cause chemical burns if they extravasate. There is no good evidence on which to base a recommendation regarding dose and rate, but the maximum rate of infusion usually employed is 1 mmol/min, which

should be controlled with a volumetric pump. The main danger of giving potassium with such rapidity is the development of hyperkalaemia, hence the patient and their ECG should be observed continuously, the serum potassium should be checked frequently, and infusion slowed as soon as the life-threatening problem has resolved (arrhythmia settled, muscular power improved). In one study, administration of 40 mmol of potassium over 1 h was found to increase serum potassium concentration by an average of 1.1 mmol/l in hypokalaemic patients with both normal and impaired renal function.

In cases that are not emergencies

In most circumstances the management of a patient with hypokalaemia requires a methodical approach to establishing the diagnosis, which is often readily apparent (but not always so), rectification (if possible) of the underlying cause, and administration of potassium at a less hurried rate than that described above. In most cases of hypokalaemia, the fall in the serum potassium concentration represents the tip of an iceberg, a reduction of 0.3 mmol/l typically reflecting a 100 mmol deficit in body stores. This relationship is variable, but it is important to remember that patients with even modest hypokalaemia may have a very considerable deficit of total body potassium that needs to be replaced.

Potassium can be given orally or intravenously. Foods with high potassium content are listed in Table 7, but it should be noted that the potassium which they contain is almost entirely coupled with phosphate. In the absence of adequate chloride intake they are therefore ineffective in replenishing body potassium in the many and common causes of hypokalaemia associated with chloride depletion (such as diuretics or vomiting). Potassium chloride can be given in either liquid or tablet form, typically 2 to 4 g (approximately 25 to 50 mmol) daily in divided doses. Both are well absorbed, but the liquid preparations are unpalatable to many patients and slow-release tablets have been associated with gastrointestinal ulceration, bleeding, and stricture, such that they must be taken with fluid whilst sitting or standing and not just before retiring to bed for the night. If intravenous administration of potassium is required, infusions containing a concentration of 20 mmol/l can usually be tolerated through a good peripheral line. If a higher concentration than this is required, central venous access will be necessary. Care must always be taken to monitor serum levels closely.

Common causes of hypokalaemia

There are a very large number of possible causes of hypokalaemia (Table 3), but in most instances the diagnosis is immediately apparent. Whenever this is not so, it is wise to remember that common things are the most likely. In the case of patients with hypokalaemia it is also important to recognize that concealment of the diagnosis is not infrequent, with diuretic abuse or covert vomiting more likely than the more exotic and rare causes of this condition. Hypokalaemia is not a prominent feature of many of the disorders listed in Table 3: discussion in this chapter will be limited to those conditions that are common, or where hypokalaemia is an important manifestation.

A pragmatic approach is first to consider the most frequent causes of hypokalaemia—diuretic ingestion and gastrointestinal fluid loss—and then proceed to a systematic analysis if these are not evidently the cause of the problem.

Treatment with diuretics

The most common cause of hypokalaemia is diuretic therapy. All diuretics other than those acting directly on the collecting duct (amiloride, triamterene, spironolactone) block some form of chloride-associated sodium transport. As a result they increase the delivery of sodium to the collecting duct, where its reabsorption creates a favourable electrochemical gradient for and obligates potassium secretion. Hypokalaemia frequently occurs together with metabolic alkalosis (serum bicarbonate concentration 28 to 36 mmol/l). In general, the hypokalaemia is mild, with serum potassium in the range 3 to 3.5 mmol/l; the average fall after inititation of the usual doses of loop diuretics (frusemide, bumetanide, torasemide) being about 0.3 mmol/l, somewhat more with the usual doses of thiazides (bendrofluazide, chlorothiazide, chlorthalidone) at about 0.6 mmol/l. In one analysis of publications on hypokalaemia and diuretics it was found that the fall in serum potassium was little influenced by the reason for prescription (hypertension or heart failure), or by the dose or duration of treatment.

The question of whether or not patients receiving diuretics prone to induce hypokalaemia should be prescribed potassium supplements or potassium-retaining diuretics has been much debated. There is no strong evidence on which to base recommendations. It seems common sense to monitor for and intervene to prevent hypokalaemia in those considered at particular risk of hypokalaemic complications, including those with a history of cardiac arrhythmia, those on digoxin, and those with liver disease in whom electrolyte imbalance might precipitate encephalopathy. Most patients do not fall into any of these categories, and here the balance is between an attempt to prevent a hypothetical but unproven hazard and the requirement for medication that is unpalatable to many and in rare cases can have significant side-effects. As in many other aspects of medicine, the behaviour of the physician will say as much about them as about the condition that they are dealing with. Those that like all test results to be in the 'normal range' will prescribe, but short of stopping diuretic therapy, correcting diuretic-induced hypokalaemia is not easy. In one study that monitored adverse drug reactions in 5047 consecutive inpatients, 2439 were taking potassium-losing diuretics, in whom serum potassium was less than 3.5 mmol/l in 21 per cent, and below 3.0 mmol/l in 3.8 per cent. If the group taking potassium-losing diuretics was broken down into those taking them without any attempt to prevent hypokalaemia, those taking them in conjunction with potassium supplements, and those taking them together with a potassium-sparing diuretic, then serum potassium below 3.5 mmol/l was found in 24.9, 19.7, and 15.2 per cent, respectively.

Loss of gastrointestinal fluid

In one study of severe hypokalaemia (serum potassium less than 2.5 mmol/l), gastrointestinal fluid loss was the main cause in 22 per cent of cases.

Vomiting

The concentration of potassium in gastric and upper intestinal secretions is between 3 and 12 mmol/l. Reduced intake and direct loss of potassium in vomit are not, therefore, the main causes of hypokalaemia, which arises due to increased renal excretion of potassium. Why does this happen? Circumstances can arise in which the renal response to one pathophysiological abnormality takes precedence over another, and where the attempt to correct one imbalance actually has the effect of worsening another. This is the situation when the kidney responds to prolonged vomiting. Aside from modest quantities of potassium, gastric juices contain sodium ions (30 to 90 mmol/l), protons (90 mmol/l), and chloride (50 to 125 mmol/l). Loss of gastric acid (HCl) pulls the buffer equation $H_2CO_3 + Na^+ + Cl^-$ in equilibrium with $Na^+ + HCO_3^- + H^+ + Cl^-$ to the right, hence the main effect is metabolic alkalosis. Depletion of extracellular fluid volume also occurs, activating the renin–angiotensin–aldosterone system. As the bicarbonate concentration in the blood rises, more is filtered at the glomerulus and some is excreted in the urine, partly in conjunction with potassium, whose distal excretion is stimulated by high levels of aldosterone. Considerations of acid–base balance have taken precedence over those of potassium homeostasis, and hypokalaemia results.

An important point to note is that the combination of direct chloride loss in vomit and contraction of extracellular fluid volume lead to a situation where the kidney avidly retains chloride and the urinary concentration of chloride falls to a very low level (less than 10 mmol/l, sometimes as low as 1 to 2 mmol/l, when the normal range is 30 to 120 mmol/l). This has critical clinical significance in two circumstances. First, since reabsorption of filtered sodium and potassium ions by the renal tubule can only be achieved in combination with an anion, usually chloride, then if urinary chloride concentration is already close to zero there is no way in which sodium and potassium can be reabsorbed efficiently. Hence, sodium and potassium that are administered can only be retained if provided in conjunction with chloride, and not if given as other salts. Second, measurement of urinary chloride concentration can be helpful in making the diagnosis of surreptitious vomiting (see later).

Resuscitation of the patient with hypokalaemia due to vomiting requires the intravenous infusion of 0.9 per cent sodium chloride, together with potassium supplementation as described above. In severe cases the total body deficit of fluid may be in excess of 5 litres, and of potassium of many hundreds of millimoles.

Diarrhoea

The concentration of potassium in stool is 80 to 90 mmol/l. Hence, given normal stool weight of 100 to 200 g/day, faecal loss of potassium is usually in the range 5 to 15 mmol/day. The potassium concentration in the stool decreases as stool volume increases, but volume can increase massively, such that substantial potassium loss and profound hypokalaemia can complicate any severe diarrhoeal illness.

Potassium loss in diarrhoeal states is usually associated with loss of bicarbonate, resulting in a coexisting metabolic acidosis, such that serum levels of potassium may not reflect the true body deficit. In this circumstance the renal excretion of potassium is broadly appropriate, and potassium deficiency is not due to a renal leak. However, in some situations potassium is lost in conjunction with chloride, resulting in a metabolic alkalosis and a picture similar to that seen with vomiting (see

above).

A villous adenoma of the colon or rectum can rarely result in profound hypokalaemia. The mechanism seems to involve secretion of cyclic AMP and prostaglandin E_2 by the tumour, leading to disturbance of ion transport in the normal colonic mucosa. Treatment with non-steroidal anti-inflammatory agents can significantly reduce stool volume and help to correct both volume depletion and hypokalaemia. Similar disturbances probably underlie the hypokalaemia of patients with the watery diarrhoea, hypokalaemia, and achlorhydria (WDHA) syndrome, caused by excess vasoactive polypeptide (VIP) secreted by certain tumours. In addition to treatment directed at the tumour itself, somatostatin or somatostatin analogues are effective in controlling symptoms.

Ureteric diversion

Diversion of the ureters into the colon (ureterosigmoidostomy) is most commonly performed in children for the treatement of bladder exstrophy, but occasionally for other reasons. If urine remains in contact with the colonic mucosa for a long time there is a tendency for the colon to reabsorb urinary ammonium and secrete bicarbonate, leading to hyperchloraemic acidosis, and also for stimulation of colonic potassium secretion, resulting in hypokalaemia. These can have serious consequences: profound acidosis can occur with concurrent illness, and chronic renal failure can develop. Close metabolic monitoring of patients with ureterosigmoidostomies is essential, and substantial metabolic disturbance is an indication for revision of the procedure, which is performed less frequently following improvement in surgical techniques for ileal conduits and alternative urinary diversions.

Diagnosing the cause of hypokalaemia in difficult cases

The diagnosis of the cause of hypokalaemia is usually straightforward and explained by diuretic therapy or gastrointestinal fluid loss, as described above. In other patients the abnormality is mild, with the occasional serum potassium concentration measured at just below the lower limit of the normal range, such that extensive investigation is almost certainly inappropriate (and likely to be fruitless if pursued). However, some patients present with unexplained severe hypokalaemia, and these represent a considerable challenge for both diagnosis and management. The differential diagnosis in these cases usually lies between concealed ingestion of diuretics, concealed vomiting and/or usage of purgatives, and various abnormalities of tubular potassium transport.

It is important to ask directly for a history of vomiting or diarrhoea, and about present or past use of any medications, particularly diuretics or purgatives. It is also worthwhile to ask about consumption of liquorice or chewing tobacco (see below). Examination is likely to be unremarkable in cases of unexplained hypokalaemia, but pay particular attention to body weight/height/body mass index, and to any other features that might support the diagnosis of an eating disorder such as anorexia nervosa or bulimia nervosa (see Chapter 26.5.5).

One study reported the findings of extensive investigation of 27 adult patients (17 women) who presented with chronic hypokalaemia (serum potassium concentration less than 3.4 mmol/l) that was sustained for over 5 years and which had previously eluded diagnosis. The following diagnoses were established: diuretic abuse (in five patients), surreptitious vomiting (eight), laxative abuse (one), renal tubular acidosis (one), and Gitelman's syndrome (12). Medical work-up that had sought to make the diagnoses by measurement of plasma renin activity, plasma aldosterone concentration, and urinary potassium concentration failed to discriminate between these conditions. Investigations that were diagnostically helpful are given in Table 4, the most useful being the plasma pH and chloride concentration, urinary chloride concentration and screen for diuretics, and (in one case) stool weight.

The finding of a low plasma chloride concentration with the virtual absence of chloride from the urine supports the diagnosis of surreptitious vomiting. Screening the urine for diuretics is appropriate if the urinary chloride concentration is above 20 mmol/l, and if no diuretics are found in samples with a chloride concentration of above 50 mmol/l then Gitelman's syndrome is likely. Vomiting, diuretics, and Gitelman's syndrome all cause alkalosis, whereas laxative abuse is associated with acidosis, as is renal tubular acidosis. The diagnosis of renal tubular acidosis can be established by demonstrating an inability to produce acid urine in the presence of systemic acidosis (see Section 20.8 for further discussion).

The management of cases of surreptitious vomiting, or diuretic or purgative abuse is difficult. Many patients will fulfil diagnostic criteria for anorexia nervosa or bulimia nervosa, and issues other than those simply and directly related to potassium homeostasis will clearly need to be considered. The physician may well need to seek expert psychiatric help. See Chapter 26.5.3 for further discussion.

Rare causes of hypokalaemia

Altered external potassium balance

Mineralocorticoid excess

Hypokalaemia can be caused by a large number of causes of mineralocorticoid excess, as shown in Table 3. Primary aldosteronism is discussed in Chapter 12.7.1 and Chapter 15.16.2.3, congenital adrenal hyperplasia in Chapter 12.7.2, and glucocorticoid-remediable aldosteronism in Chapter 15.16.1.2. Hypokalaemia is rarely a prominent feature of the other conditions of mineralocorticoid excess listed, which are discussed elsewhere in this book.

Apparent mineralocorticoid excess

Activating mutations in the b- or g-subunits of the epithelial sodium channel in the collecting duct causes Liddle's syndrome. Disabling mutations in the type 2 11b-hydroxysteroid dehydrogenase gene cause a deficiency of the enzyme, allowing cortisol access to the mineralocorticoid receptor and the syndrome of apparent mineralocorticoid excess. Acquired inhibition of the action of 11b-hydroxysteroid dehydrogenase can be caused by liquorice, carbenoxolone, and chewing tobacco. Hypokalaemia with low plasma concentrations of renin and aldosterone are features of all of these conditions, which are discussed in Chapter 15.16.1.2.

Renal transport abnormalities

Patients with renal tubular acidosis type I are prone to hypokalaemia, as discussed in Chapter 20.8.

In 1962 Bartter described 'hyperplasia of the juxtaglomerular complex with hyperaldosteronism and hypokalemic alkalosis: a new syndrome'. Well over a hundred papers were subsequently written to describe features of what was believed to be the same eponymously named condition. The picture became immensely confused, but since 1995 has been clarified by the recognition of distinct phenotypes within the group of patients previously thought to have 'Bartter's syndrome' and the application of powerful molecular genetic methods to their study. These have revealed that most patients previously thought to have Bartter's syndrome do not have this condition, but have Gitelman's syndrome instead.

Bartter's syndrome

This is now classified into three types: each is an autosomal recessive disorder caused by mutation of an ion transporter or ion channel that is present in cells of the thick ascending limb of the nephron (Fig. 1).

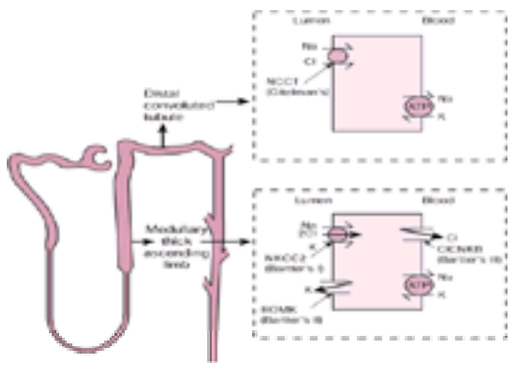

Fig. 1 Some genetic disorders of the renal tubule that cause hypokalaemia. NCCT, Na-Cl cotransporter; NKCC2, Na-K-2Cl cotransporter; ROMK, ATP-regulated potassium channel; CLCNKB, kidney-specific chloride channel.

Bartter's syndrome type I: *NKCC2* mutations

This was originally described in six consanguineous families. Affected individuals were born prematurely after pregnancies complicated by polyhydramnios and developed severe dehydration in the first few days of life. All affected individuals had severe hypercalciuria in addition to hypokalaemic alkalosis, and most had nephrocalcinosis. In all cases disease was associated with destructive mutations in the gene encoding the Na-K-2Cl cotransporter (NKCC2), which is localized to chromosome 15. Other clinical manifestations can include short stature, mental retardation, rickets, generalized weakness, and muscle cramps. Other reported abnormalities on investigation can include hyperreninism, hyperaldosteronism, increased renal prostaglandin production, erythrocytosis, a platelet aggregation defect, impaired vascular responses to angiotensin II, and hypertrophy or hyperplasia of the juxtaglomerular apparatus. Management is supportive: dehydration must be avoided and potassium supplementation is needed; non-steroidal anti-inflammatory agents may be helpful (although they tend to be more useful in type II disease). There is a high mortality rate before diagnosis, with infants often dying due to volume depletion caused by intercurrent illness, but the prognosis in cases where the diagnosis is made and where care is taken to avoid volume depletion is not known.

Bartter's syndrome type II: *ROMK* mutations

Abnormality of the *NKCC2* gene product was excluded in five families where individuals also presented with severe neonatal dehydration, hypercalciuria, and nephrocalcinosis. The only clinical distinction from Bartter's syndrome type I was that patients often had a transient initial hyperkalaemia, with serum potassium falling rapidly into the hypokalaemic range as soon as they were rehydrated. An obvious explanation was mutation in another gene or genes whose product interacted with NKCC2 in some way. In 1993 and 1994 a renal potassium channel had been cloned from rat and humans, and this apical ATP-sensitive potassium channel (ROMK) that recycles potassium back into the lumen and is critical for continued activity of the NKCC2 cotransporter was an obvious candidate. Functionally significant mutations of *ROMK* (also known as KCNJ1) were identified in all affected individuals. Management and prognosis is as for Bartter's syndrome type I.

Bartter's syndrome type III: *CLCNKB* mutations

To determine whether mutation of other genes could account for the Bartter's phenotype a large number of patients with inherited hypokalaemic alkalosis, normomagnesaemia, and normocalciuria or hypercalciuria were studied. Most (those in 44 of 66 families) did not have mutations in *NKCC2* or *ROMK*. In 1994 two highly homologous renal chloride channels were cloned, *CLCNKA* and *CLCNKB*, and the latter was shown to be the cause of the Bartter's syndrome in a number of the families. The clinical picture was more varied than that for types I or II Bartter's syndrome, ranging in severity from near fatal volume depletion with hypokalaemic alkalosis and respiratory arrest to mild disease presenting in a teenager with polyuria and weakness. None of the patients had nephrocalcinosis, distinguishing them phenotypically from those with *NKCC2* or *ROMK* mutations. Management is with potassium supplementation and care to avoid dehydration. Long-term prognosis is uncertain.

Gitelman's syndrome

Gitelman's syndrome is the commonest genetic cause of hypokalaemia. If it presents clinically, it typically does so in early adulthood with hypotension, alkalosis, and salt wasting, along with hypomagnesaemia, hypocalciuria, and hypermagnesuria (see Table 4). The marked similarity between this picture and that induced by thiazide diuretics, which are potent inhibitors of the Na-Cl cotransporter (NCCT) in the distal convoluted tubule of the nephron, led to a candidate gene approach to the condition as soon as the thiazide-sensitive NCCT gene had been cloned. Mutations in the *NCCT* gene are responsible for Gitelman's syndrome (Fig. 1), which is an autosomal recessive condition.

Since there are no dramatic clinical symptoms or signs, suspicion of the diagnosis of Gitelman's syndrome often arises only when hypokalaemia is found (or in screening of family members of a known case). However, in one recent study it was clearly shown that patients with Gitelman's syndrome are significantly more symptomatic than controls, reporting salt craving, musculoskeletal symptoms (cramps, muscle weakness, and aches), constitutional symptoms (fatigue, generalized weakness, and dizziness), nocturia, and polydipsia. Forty-five per cent of patients considered their symptoms to be a moderate problem or worse.

Management is with potassium and magnesium supplements, it being important to recognize that in the face of magnesium depletion the kidney cannot retain potassium, but these are often poorly tolerated. Diuretics that block sodium reabsorption in the collecting duct (spironolactone, triamterene, and amiloride) can reduce urinary potassium excretion and raise the serum potassium concentration, but they often need to be accompanied by salt-loading to prevent volume depletion and hypotension. Non-steroidal anti-inflammatory agents can sometimes be helpful, but their mechanism of action is uncertain. The long-term prognosis of patients with Gitelman's syndrome is not known.

Heterozygote carriers of Bartter's or Gitelman's mutations

The phenotypes that might be associated with heterozygote carriage of the Bartter's and Gitelman's mutations have not been well characterized. Heterozygote carriers of Gitelman's mutations have increased urinary sodium excretion (due to a self-selected higher salt intake), modestly lowered blood pressure (in childhood if not in adulthood), a serum potassium concentration towards the lower limit of the normal range, and increased susceptibility to hypokalaemia induced by diuretics. Similar features would be anticipated in Bartter's heterozygotes. Carriers of Gitelman's mutations have increased bone density; carriers of Bartter's may have a predisposition to osteoporosis or nephrolithiasis caused by hypercalciuria.

Abnormal internal potassium balance

Although there are many causes of hypokalaemia (Table 3), there are relatively few causes of hypokalaemia associated with extreme weakness, the commonest explanation for this rare presentation being hypokalaemic periodic paralysis. In Western countries most cases of hypokalaemic periodic paralysis are familial, termed familial periodic paralysis, whereas in Asian populations the commonest cause is thyrotoxic periodic paralysis. In all forms of hypokalaemic periodic paralysis the hypokalaemia and paralysis result from an acute shift of potassium into cells, the mechanism for which is unknown, although there is speculation that it is due to a transient hyperadrenergic state.

One study reviewed the medical records of 97 patients who presented over a 10-year period to hospital in Taiwan with severe hypokalaemia (plasma potassium less than 3.0 mmol/l, mean 2.2 mmol/l) and acute loss of muscle strength with inability to walk. The final diagnoses established are shown in Table 5.

Treatment of acute attacks of hypokalaemic periodic paralysis traditionally involves the administration of intravenous potassium, some patients recovering with as little as 20 mmol, but others requiring over 200 mmol. In all types of this condition a paradoxical fall in serum potassium concentration can occur at the start of treatment, and rebound hyperkalaemia is also seen.

Thyrotoxic periodic paralysis

The diagnosis of thyrotoxic periodic paralysis is established if hyperthyroidism is present when hypokalaemic paralysis occurs. About 50 per cent of patients give a history of thyrotoxic symptoms, but there is no family history of paralysis. Attacks are often provoked by a large carbohydrate meal (perhaps via the mechanism of an exaggerated response to insulin) or adrenergic stress. Physical findings during an attack include tachycardia (a useful diagnostic discriminator from sporadic periodic paralysis) and high blood pressure; signs of hyperthyroidism are absent in 20 to 40 per cent of cases. In 39 patients reported from Taiwan the mean serum T_3 concentration was 4.5 nmol/l (range 2.3 to 8.4, upper limit of normal being 3.0), the mean serum T_4 concentration was 201 nmol/l (range 154 to 299, upper limit of normal 154), and the mean thyroid-stimulating hormone was less than 0.06 mU/l (range less than 0.06 to 0.32; normal 0.5 to 5.0). Hypophosphataemia and hypomagnesaemia are also found, the latter also being low in patients with Gitelman's syndrome.

Although treatment of thyrotoxic periodic paralysis conventionally involves administration of potassium, recent experience suggests that patients with this condition respond rapidly to the b-blocker propanolol (3 mg/kg, given orally). This, rather than potassium, is now the preferred first-line treatment, with the expectation that serum potassium concentration will return to normal and paralysis will resolve within 2 h.

Sporadic periodic paralysis

The cause of sporadic periodic paralysis is not known: patients do not have a family history of hypokalaemic periodic paralysis and do not have hyperthyroidism. There are no obvious precipitating factors. Heart rate at presentation is lower than for those with thyrotoxic periodic paralysis (mean 76 compared with 105 beats/min). Emergency treatment is with intravenous potassium. Propanolol is ineffective. Oral potassium chloride supplements or acetazolamide are used to prevent recurrent attacks. The mechanism of action of acetazolamide is uncertain, but it has been reported in an animal model that it causes activation of the sarcolemmal calcium-activated potassium channel.

Familial hypokalaemic periodic paralysis

The diagnosis is established by finding a family history of attacks of flaccid weakness and hypokalaemia. These can be precipitated by administration of insulin or glucose and aborted by exercise, which induces an exaggerated rise in serum potassium concentration, or by administration of potassium.

Familial hypokalaemic periodic paralysis can be caused by mutations in three genes. First, the *CACNL1A3* gene, which encodes a dihydropyridine receptor that functions as a voltage-gated calcium channel and is also critical for excitation–contraction coupling in a voltage-sensitive and calcium-independent manner. Second, the *SCN4A* gene that encodes for a sodium channel and is also the site of mutations causing hyperkalaemic periodic paralysis. Third, the *KCNE3* gene, which encodes a potassium channel. The conditions are autosomal dominant, with 100 per cent penetrance in males, but much less in females.

Treatment is as for sporadic periodic paralysis. Acetazolamide usually prevents recurrent attacks, but one family has been reported where this made the condition worse (a beneficial response to triamterene was observed).

Sudden unexplained death syndrome

Between 1982 and 1990 there were a total of 235 cases of sudden unexplained death syndrome (**SUDS**) in apparently healthy male Thai migrant workers in Singapore. SUDS, known locally as *laitai*, is a leading cause of death in young men in rural north-eastern Thailand, where one study reported an annual incidence of 38 per 100 000 men aged 20 to 49 years. It is also reported elsewhere in Asia. Women are rarely, if ever, affected. Death occurs at rest and is nocturnal in most (84 per cent) cases. There is a family history of SUDS more often than would be expected by chance. In cases that are observed, witnesses often report that death is preceded by a few minutes of groaning, choking, coughing, and muscular spasticity or paralysis.

The cause of SUDS is not known; hypotheses include stress, genetic factors, dietary deficiency (perhaps of thiamine), potassium deficiency, melioidosis, and sleep disorders. With regard to potassium, survivors of SUDS-like attacks and relatives of victims of SUDS have been reported to have significantly lower activity of erythrocyte Na^+,K^+-ATPase and lower plasma potassium concentration than controls, but the reason for and significance of these findings is not certain.

Hyperkalaemia

Clinical features and treatment of hyperkalaemia

Hyperkalaemia is the most serious of all electrolyte disorders, despite being relatively infrequent, because it typically produces no recognizable symptoms before causing cardiac arrest. Some patients report muscular symptoms such as weakness, stiffness, or simply a 'funny feeling', but the significance of these is rarely appreciated. A high serum potassium concentration leads to membrane depolarization in excitable tissues, making the initation of an action potential more likely, and to increased membrane potassium conductance, which impairs recovery after an action potential. The effect is to cause electrical instability with the risk of life-threatening arrhythmia. The likelihood of such an event increases as the serum potassium concentration rises, but some patients are more resistant to the cardiac effects of hyperkalaemia than others: for instance, those with endstage renal failure on long-term dialysis may be habitually hyperkalaemic (although this is not to be encouraged) and tolerate a plasma potassium concentration that would kill a normal person if imposed acutely.

The best guide to the significance of hyperkalaemia in any particular individual is the impact that it is having on the ECG, and an ECG should be obtained immediately in any patient in whom the question of hyperkalaemia arises. The earliest change is tenting of the T wave, progressing as the plasma potassium concentration rises to P-wave flattening, prolongation of the P–R internal, widening of the QRS complex, and eventually a 'sine wave' pattern as a prelude to ventricular fibrillation and death. All involved in the care of acutely ill patients must be able to recognize this pattern of ECG changes and give effective emergency treatment for severe hyperkalaemia, as described in Chapter 20.5.

Causes of hyperkalaemia

There are many causes of a high serum potassium concentration (Table 6), but a survey of over 400 cases found that renal failure was present in 43 per cent and potassium supplements or potassium-sparing diuretics had been taken by 37 per cent. Life-threatening hyperkalaemia is almost exclusively seen in those with renal failure, often in conjunction with another exacerbating cause. Common scenarios would be the patient with acute renal failure who is hypercatabolic or has extensive tissue destruction, as in rhabdomyolysis, or the patient with endstage renal failure who has missed a dialysis treatment, not adhered to a low potassium diet (see Table 7), or suffered an upper gastrointestinal haemorrhage, thereby inadvertently consuming a high potassium meal.

Hyperkalaemia is not a prominent feature of many of the conditions listed in Table 6: further discussion in this chapter will be limited to disorders other than renal failure that are not discussed elsewhere in this textbook and in which hyperkalaemia is a common or important manifestation.

Pseudohyperkalaemia

Haemolysed samples show hyperkalaemia, which also occurs when there is considerable delay between venepuncture and separation of red cells and plasma or serum in the laboratory, allowing potassium to leak out of red cells after venesection. However, aside from these common and banal explanations, there are other reasons for pseudohyperkalaemia.

Potassium is released from white blood cells and platelets as blood coagulates, causing the serum potassium concentration to exceed, by a few tenths of a millimole per litre, that of plasma estimated in a parallel sample. This process is greatly exaggerated when gross leucocytosis or thrombocytosis is present, such that the serum potassium concentration can be over 2 mmol/l higher than that in plasma. The plasma and not the serum potassium concentration should obviously be measured in this circumstance.

There is also the rare syndrome of familial pseudohyperkalaemia, first described in 16 members of three generations of a kindred from Edinburgh who had elevated plasma potassium if the red cells were not separated promptly. Several other families have been described, in each of which there appears to be one of a variety of abnormalities in the temperature sensitivity of the ouabain-plus-bumetanide-resistant potassium flux, which reflects the passive leak. The blood film may show a few target cells, red cell survival is shortened, but there is no frank haemolysis. Other phenotypic abnormalities have been described in some families.

Abnormal external potassium balance

Mineralocorticoid deficiency

Hyporeninaemic hypoaldosteronism

It is not uncommon to find patients with chronic renal failure who have hyperkalaemia despite a glomerular filtration rate that should be sufficient to maintain

normokalaemia. Two-thirds of these will have the syndrome of hyporeninaemic hypoaldosteronism, which should be suspected in any patient with hyperkalaemia without other obvious explanation. Tubulointerstitial forms of renal disease predominate in this population and diabetes mellitus is common. Hyperkalaemia is usually asymptomatic, but presentation with cardiac arrhythmia and/or muscle weakness has been described.

Characteristics of the syndrome include low levels of plasma renin activity, which are unresponsive to sodium restriction or frusemide, low plasma and urinary aldosterone, hyperkalaemia, and hyperchloraemic metabolic acidosis. Fractional potassium excretion is low for the glomerular filtration rate, and the response to kaliuretic stimuli is blunted. Glucocorticoid metabolism is normal.

The cause of both hyporeninism and hypoaldosteronism is not known. Decreased renin secretion may be the result of pathological involvement of the juxtaglomerular apparatus, but this is not obvious histologically in cases where renal biopsies have been performed. Other hypotheses include defective prostacyclin production and disordered conversion of inactive (prorenin) to active renin, which is a well-documented observation in diabetes mellitus. Chronic expansion of extracellular fluid volume has also been blamed since plasma renin activity may be increased by prolonged sodium restriction or diuretic therapy. By contrast, acute salt restriction can worsen hyperkalaemia in these patients by diminishing the distal delivery of sodium without a concurrent rise in aldosterone secretion, and illnesses causing volume depletion can precipitate presentation with dangerous hyperkalaemia. Hypoaldosteronism is probably related to the low level of plasma renin activity, but the situation is more complicated than this since most patients secrete subnormal amounts of aldosterone in response to infusion of both angiotensin II and ACTH, suggesting a defect in the function of the adrenal gland.

Criteria for establishing the diagnosis of hyporeninaemic hypoaldosteronism are not well defined, and it is uncommon for patients to be intensively investigated in routine clinical practice since treatment is usually straightforward. However, establishing the diagnosis with certainty depends on demonstrating deficient responses of renin and aldosterone to sodium depletion. One study reported plasma renin and aldosterone concentrations in subjects in an upright posture after administration of 60 mg of intravenous frusemide: the renin concentration in those with hyporeninaemic hypoaldosteronism was 6 ± 2 ng/ml per minute (compared with 34 ± 6 in controls matched for degree of renal failure) and aldosterone concentration was 7 ±2 ng/dl (compared with 28 ±2 8 in controls).

Therapy for hyporeninaemic hypoaldosteronism includes dietary potassium restriction and avoidance of drugs that can cause hyperkalaemia. Measures to increase urinary excretion of potassium, such as the use of thiazide or loop diuretics, can be useful. Cation exchange resins can be used to increase elimination of potassium from the gut, but compliance with long-term use of these medications is difficult to achieve. Although mineralocorticoid replacement (fludrocortisone, 0.2 mg/day) effectively treats the hyperkalaemia, sodium retention and worsening hypertension are often unacceptable side-effects.

Effects of drugs on the renin–angiotensin–aldosterone system

Non-steroidal anti-inflammatory agents

Prostaglandin synthetase inhibitors produce hyporeninaemic hypoaldosteronism by interfering with prostacyclin-mediated renin secretion, with reduction in glomerular filtration rate and distal sodium delivery as potential contributory factors. These effects, as might be expected, become more important in the context of renal impairment: in one study approximately a quarter of patients with chronic renal failure developed hyperkalaemia after treatment with indomethacin.

Angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers

Angiotensin-converting enzyme (**ACE**) inhibitors and angiotensin II receptor blockers produce hyperkalaemia by impairing angiotensin II-mediated secretion of aldosterone. In one study hyperkalaemia was found in 46 of 119 (39 per cent) patients taking ACE inhibitors who were attending a renal clinic. The higher the serum creatinine concentration, the greater the chance of hyperkalaemia. Those with diabetes were also at particular risk. The treatment had to be stopped in 15 patients (13 per cent).

ACE inhibitors and spironolactone have been found to improve prognosis in heart failure, but care is needed when prescribing for those who might be prone to hyperkalaemia. One study reported life-threatening hyperkalaemia (mean serum potassium 7.7 mmol/l) in 25 patients who had received this combination of medications.

Calcineurin inhibitors

Hyperkalaemia is a well-documented complication of the immunosuppressive drugs cyclosporin and tacrolimus (FK506). Two mechanisms are possible, both of which may be exacerbated by reduction in glomerular filtration rate caused by nephrotoxicity. First, drug-induced hyporeninaemic hypoaldosteronism, which is well documented with tacrolimus. Second, in association with a distal tubular acidification defect that is caused (mechanism unknown) by both cyclosporin and tacrolimus.

Heparin

Hyperkalaemia occurs in about 7 per cent of patients given heparin, which is a potent inhibitor of aldosterone production. It can arise with doses as low as 10 000 units/day, but—as with most other hyperkalaemic stimuli—clinically important elevations in the plasma potassium concentration are found only when more than one homeostatic mechanism for potassium is deranged. Patients with endstage renal failure who receive unfractionated heparin to provide anticoagulation during haemodialysis treatments have a higher predialysis plasma potassium than those given low-molecular-weight heparin.

The most important mechanism of aldosterone inhibition appears to involve reduction in both numbers and affinity of angiotensin II receptors in the zona glomerulosa, which is reduced in width by prolonged use of heparin. Direct inhibition of the enzyme 18-hydroxylase has also been postulated. Production of other corticosteroids is not affected.

Renal transport abnormalities

Tubulointerstitial renal disease

A few hyperkalaemic patients with chronic renal failure but a glomerular filtration rate that should be adequate for potassium homeostasis have normal levels of aldosterone and plasma renin activity and seem to have a primary defect in the ability of the distal nephron to excrete potassium. They typically have tubulointerstitial types of renal diseases, the abnormality being documented in patients with obstructive uropathy, renal transplants, sickle-cell disease, systemic lupus erythematosus, amyloidosis, and medullary sponge kidney—all of which can also be associated with hyporeninaemic hypoaldosteronism. In contrast to patients with hyporeninaemic hypoaldosteronism, their hyperkalaemia is unresponsive to mineralocorticoid replacement therapy.

Type IV renal tubular acidosis

Hyperkalaemia due to impaired renal excretion of potassium may be a feature of type IV or voltage-dependent renal tubular acidosis. The lumen negative potential difference along the distal nephron normally facilitates the excretion of potassium and hydrogen ions, and hyperkalaemia and metabolic acidosis occur when this is reduced. This condition is discussed in Chapter 20.13.

Drugs

Potassium-sparing diuretics are obviously likely to cause hyperkalaemia in those with any predisposition to this condition. They should not be used in those with renal failure, and the serum potassium concentration must be monitored closely in patients taking these agents who become acutely unwell.

Trimethoprim–sulphamethoxazole

A review of 80 patients treated with standard-dose trimethoprim (up to 320 mg/day) and sulphamethoxazole (up to 1600 mg/day) showed that this increased the serum potassium concentration by an average of 1.2 mmol/l, whereas there was no change in a control group receiving other antibiotics. Some studies have shown a lesser effect than this, but even larger increases in serum potassium concentration have been reported in those receiving high-dose trimethoprim–sulphamethoxazole to

treat pneumocystis, and hyperkalaemia is also reported with use of pentamidine. Both trimethoprim and pentamidine block the apical sodium channel in the distal nephron in a manner similar to amiloride.

Pseudohypoaldosteronism

Pseudohypoaldosteronism type 1

Autosomal recessive pseudohypoaldosteronism type 1 is caused by mutations in the a-, b-, or g-subunits of the epithelial sodium channel (ENaC). Mutations of the mineralocorticoid receptor gene can cause an autosomal dominant form of this condition.

The recessive form typically presents in infancy with vomiting and feeding difficulty. There are signs of volume depletion and laboratory findings of hyponatraemia, hyperkalaemia, and acidaemia. The plasma renin concentration is usually increased and plasma aldosterone concentration is markedly elevated. The sodium concentration in urine, sweat, saliva, and stool is high. Treatment is with salt supplements that must usually be continued into adulthood. By contrast, the autosomal dominant form has a milder phenotype, with symptoms that remit with age.

Pseudohypoaldosteronism type 2 (Gordon's syndrome)

Pseudohypoaldosteronism type 2 (Gordon's syndrome) is a rare autosomal dominant condition in which there is hyperkalaemia despite normal glomerular filtration rate, hypertension, and correction of physiological abnormalities with thiazide diuretics, which may provide effective treatment.

The condition is usually asymptomatic, detected fortuitously if serum potassium concentration is measured for any reason, or in the course of family studies, but can rarely present in late childhood or adulthood with hyperkalaemic periodic paralysis. There is a hyperchloraemic acidosis, a low level of plasma renin activity, and normal or slightly low plasma aldosterone concentration. Giving exogenous aldosterone does not increase urinary potassium excretion or reduce hyperkalaemia.

Loci have been mapped to 1q (PHA2A), 12, and 17q21 (PHA2B), but the gene(s) responsible have not been identified. Because a kaliuresis can be provoked by infusion of sodium sulphate or sodium bicarbonate, but not sodium chloride, it has been suggested that enhanced reabsorption of chloride at a distal nephron site may underlie the abnormality in potassium secretion.

Abnormal internal potassium balance

Exercise

Exercise-related rises in the plasma potassium concentration are a normal phenomenon and usually modest, but increases to 7.0 mmol/l occur during acute, maximal, physical performance and levels as high as 10.0 mmol/l have been reported with prolonged exhaustive exercise such as in marathons. Exercise-induced hyperkalaemia is accentuated by b-adrenergic blockade, a-adrenergic agonists, and in patients with chronic renal failure.

Acidosis

Acidosis diminishes potassium uptake by cells (Table 1) and causes hyperkalaemia. The increase in the plasma potassium concentration is greater with metabolic than respiratory acidosis, and occurs with hyperchloraemic but not with organic acid-induced forms of metabolic acidosis. Stimulation of insulin release by organic acids appears to account for this divergent response, explaining the pathophysiology of disturbed potassium homeostasis in diabetic ketoacidosis. At presentation, when insulin is deficient, potassium is redistributed in a fashion comparable with mineral acid-induced metabolic acidosis and patients are hyperkalaemic. However, the preceding kaliuresis (caused by polyuria) has rendered the body enormously deficient in potassium, and the plasma potassium concentration falls rapidly as soon as insulin is provided, allowing potassium to return to the cells. Indeed, dangerous hypokalaemia can develop if adequate potassium is not given during treatment.

Drugs

Several drugs can produce hyperkalaemia by altering the transcellular distribution of potassium. Digitalis preparations diminish cellular potassium uptake by inhibiting the Na^+,K^+ pump and substantial hyperkalaemia can accompany digitalis intoxication. Succinylcholine and other depolarizing muscle relaxants increase the potassium permeability of muscle: the plasma potassium concentration typically increases by 0.5 to 1.0 mmol/l, but in patients with burns or neuromuscular diseases, hyperkalaemia can be more severe. Infusion of 30 g of the cationic amino acid arginine HCl increases plasma potassium concentration by 0.5 to 1.0 mmol/l and can produce life-threatening hyperkalaemia in individuals with deranged potassium metabolism. Fluoride intoxication appears to increase the plasma potassium concentration by provoking leakage from the intracellular compartment: associated hypocalcaemia enhances the cardiac risks of fluoride-induced hyperkalaemia.

Although b_2-adrenergic stimulants cause hypokalaemia and can be used to treat hyperkalaemia (see section 20.5), the administration of b-blockers typically increases the plasma potassium concentration only modestly (by 0.1 to 0.2 mmol/l). However, the hyperkalaemic effect can be much more prominent when other potassium homeostatic mechanisms are deranged, for example in patients receiving intermittent haemodialysis, the predialysis plasma potassium concentration is increased on average by 1.0 mmol/l.

Hyperkalaemic periodic paralysis

Hyperkalaemic periodic paralysis is a rare autosomal dominant condition in which mutations in the sodium channel gene *SCN4A* are associated with episodes of flaccid generalized weakness (rather than paralysis) and elevation of the serum potassium concentration, typically into the range from 6.0 to 8.0 mmol/l. Mutations in the same gene can cause hypokalaemic periodic paralysis and paramyotonia congenita. Attacks of weakness last from minutes to hours, occurring without any obvious precipitant but sometimes following exercise or administration of potassium.

Treatment with kaliuretic diuretics (not potassium sparing) is used to prevent attacks. b_2-Agonists can be used both to prevent and abort paralytic attacks.

Myotonia of the ocular muscles and tongue is sometimes observed both between and during attacks, the former demonstrable as slow opening of the lids after forced active closure of the eyes, or as myotonic lid lag lasting 15 to 20 s after elevation of the eyes. There can be generalized muscle wasting and progressive myopathy. In some families cardiac arrhythmia, cardiac sudden death, short stature, microcephaly, and clinodactyly (typically a bent little finger) are reported.

Further reading

Brater DC (1998). Diuretic therapy. *New England Journal of Medicine* **339**, 387–95.

Cruz DN *et al*. (2001). Mutations in the Na-Cl cotransporter reduce blood pressure in humans. *Hypertension* **37**, 1458–64.

Cruz DN *et al.* (2001). Gitelman's syndrome revisited: an evaluation of symptoms and health-related quality of life. *Kidney International* **59**, 710–17.

Gennari FJ (1998). Hypokalemia. *New England Journal of Medicine* **339**, 451–8.

Gladziwa U *et al*. (1995). Chronic hypokalaemia of adults: Gitelman's syndrome is frequent but classical Bartter's syndrome is rare. *Nephrology, Dialysis, Transplantation* **10**, 1607–13.

Grier JF (1995). WDHA (watery diarrhea, hypokalemia, achlorhydria) syndrome: clinical features, diagnosis, and treatment. *Southern Medical Journal* **88**, 22–4.

Halevy J *et al*. (1988). Life-threatening hypokalemia in hospitalized patients. *Mineral and Electrolyte Metabolism* **14**, 163–6.

Hamill RJ *et al*. (1991). Efficacy and safety of potassium infusion therapy in hypokalemic critically ill patients. *Critical Care Medicine* **19**, 694–9.

Lin SH, Lin YF (2001). Propranolol rapidly reverses paralysis, hypokalemia, and hypophosphatemia in thyrotoxic periodic paralysis. *American Journal of Kidney Diseases* **37**, 620–3.

Lin SH *et al*. (2001). Hypokalaemia and paralysis. *Quarterly Journal of Medicine* **94**, 133–9.

Morgan DB, Davidson C (1980). Hypokalaemia and diuretics: an analysis of publications. *British Medical Journal* **280**, 905–8.

Nadler JL *et al*. (1986). Evidence of prostacyclin deficiency in the syndrome of hyporeninemic hypoaldosteronism. *New England Journal of Medicine* **314**, 1015–20.

Older J *et al*. (1999). Secretory villous adenomas that cause depletion syndrome. *Archives of Internal Medicine* **159**, 879–80.

Oster JR *et al*. (1995). Heparin-induced aldosterone suppression and hyperkalemia. *American Journal of Medicine* **98**, 575–86.

Paice BJ *et al*. (1986). Record linkage study of hypokalaemia in hospitalized patients. *Postgraduate Medical Journal* **62**, 187–91.

Preston RA *et al*. (1998). University of Miami Division of Clinical Pharmacology therapeutic rounds: drug-induced hyperkalemia. *American Journal of Therapeutics* **5**, 125–32.

Scheinman SJ *et al*. (1999). Genetic disorders of renal electrolyte transport. *New England Journal of Medicine* **340**, 1177–87.

Schepkens H *et al*. (2001). Life-threatening hyperkalemia during combined therapy with angiotensin-converting enzyme inhibitors and spironolactone: an analysis of 25 cases. *American Journal of Medicine* **110**, 438–41.

Simon DB, Lifton RP (1998). Ion transporter mutations in Gitelman's and Bartter's syndromes. *Current Opinion in Nephrology and Hypertension* **7**, 43–7.

Tosukhowong P *et al*. (1996). Hypokalemia, high erythrocyte Na+ and low erythrocyte Na^+,K^+-ATPase in relatives of patients dying from sudden unexplained death syndrome in north-east Thailand and in survivors from near-fatal attacks. *American Journal of Nephrology* **16**, 369–74.

Widmer P *et al*. (1995). Diuretic-related hypokalaemia: the role of diuretics, potassium supplements, glucocorticoids and b $_2$-adrenoceptor agonists. Results from the comprehensive hospital drug monitoring programme, Berne (CHDM). *European Journal of Clinical Pharmacology* **49**, 31–6.

Wong ML *et al*. (1992). Sudden unexplained death syndrome. A review and update. *Tropical and Geographical Medicine* **44**, S1–19.

20.3.1 The clinical presentation of renal disease

Alex M. Davison

Introduction

Effective diagnosis has four essential requisites: an awareness of the patterns of clinical presentation; obtaining a complete history; undertaking a structured clinical examination; and formulating an appropriate investigative plan. In many instances, the clinical symptoms of renal disease are non-specific, the underlying condition may not be suspected from the history alone, and the physical examination may be surprisingly unrevealing. It is often routine investigations, such as urine analysis or estimation of renal function, that suggest the presence of renal pathology. Clinicians therefore need to be aware of the symptoms and signs that give a clue to an underlying renal disease, and which act as a prompt for appropriate initial diagnostic investigations.

The presence of renal disease in a patient may be detected because of:

1. presentation with a symptom or clinical sign that indicates an underlying renal disorder;
2. the presence of a systemic disease known to involve the kidneys;
3. a family history of inherited renal disease;
4. the finding of asymptomatic urinary abnormalities or disordered renal function tests.

Many patients remain asymptomatic, even with advanced renal disease, hence the importance of urine analysis and the estimation of blood urea and serum creatinine in anyone suspected of having renal disease. Unlike most other organ systems, patients with renal failure may remain asymptomatic despite the loss of up to 80 per cent of excretory function. Not surprisingly they may therefore be unaware of the presence of advanced renal disease, and as a consequence find it difficult to come to terms with the severity and seriousness of their illness.

Clinical syndromes

Asymptomatic urinary abnormalities

Asymptomatic proteinuria

Urinary protein excretion can amount to 150 mg daily in normal persons, consisting of albumin, Tamm–Horsfall protein, and secretory IgA. An accurate 24-h urine collection is difficult to obtain, particularly in outpatients, and therefore it is frequently more convenient to estimate the urinary protein/creatinine ratio on a mid-morning sample of urine, a normal value would be less than 130 (as this is a ratio it is without units). Approximately half consists of low molecular weight proteins or protein fragments, with the rest being albumin.

The most common method of detecting proteinuria is by using dipstix. These paper strips are impregnated with tetrabromophenol blue which changes colour from yellow-green to blue-green in the presence of protein. This test is very observer-dependent, and it should be remembered that Bence-Jones protein will not be detected and that false-positive results can occur both in alkaline urine and in urine contaminated with antiseptics (see Section 20.4 for further discussion).

Urinary protein excretion can increase during pyrexial illnesses, with strenuous exercise, congestive cardiac failure, and hypertension. In such patients the proteinuria is commonly mild (generally less than 1.5 g daily) and resolves with remission of the underlying cause. If proteinuria is detected in these circumstances the test should be repeated once the potential cause has resolved. If persistent proteinuria is detected then further investigation to determine the nature of the underlying disease is indicated.

Microalbuminuria

'Microalbuminuria' is the term used for urinary protein excretion greater than normal but still less than that detectable by dipstix testing. The excretion of more than 30 µg/min of albumin in an overnight collection or 70 µg/min in a 24-h collection in a patient with diabetes mellitus is indicative of early diabetic nephropathy. It is, however, not specific for diabetes: microalbuminuria may also be present in hypertension, obesity, systemic lupus erythematosus, and following exercise. Specifically designed stix tests are now available for screening purposes, but these remain only semiquantitative.

Postural (orthostatic) proteinuria

In some patients it has been noted that the proteinuria is present in samples obtained during the day, but absent from samples obtained first thing in the morning after overnight recumbency. This has been termed postural or orthostatic proteinuria. It is most common in children and young adults, and, although the pathogenesis is uncertain, it most likely represents an exaggerated intraglomerular haemodynamic response to a change in posture and/or entrapment of renal veins. Urinary protein excretion in such patients rarely exceeds 1.0 g/24 h. A renal biopsy may be normal or reveal only minor abnormalities.

The long-term prognosis is excellent, hence in patients with minor proteinuria it is important to obtain an early morning sample for testing. If this is found to be negative on stix testing then a diagnosis of postural proteinuria can be made, no biopsy is necessary, and the patient can be reassured that they do not have a sinister renal problem. As the majority of patients with this condition are young adult males this diagnosis may be of value in obtaining employment or insurance.

Investigation of asymptomatic proteinuria

The Scottish Intercollegiate Guidelines Network (**SIGN**) have produced a reasonable guideline for the investigation in primary care and by non-nephrologists of adult patients with asymptomatic proteinuria (Fig. 1). Different nephrologists will take different views regarding further investigation. If there are no symptoms, blood pressure is normal, the glomerular filtration rate (**GFR**) (usually estimated from serum creatinine by the Cockcroft and Gault formula, or from a 24-h creatinine clearance) is within the normal range, and there is no clinical suspicion of a multisystem disorder, then most will adopt a 'watching brief' (reviewing in 6 months and annually thereafter if all parameters remain stable) and perform a renal biopsy only in the event of a change in one or more of these parameters. The justification for this approach is that the chances of a renal biopsy revealing a diagnosis that alters management are extremely small in this context, and that the benefit does not outweigh the risk.

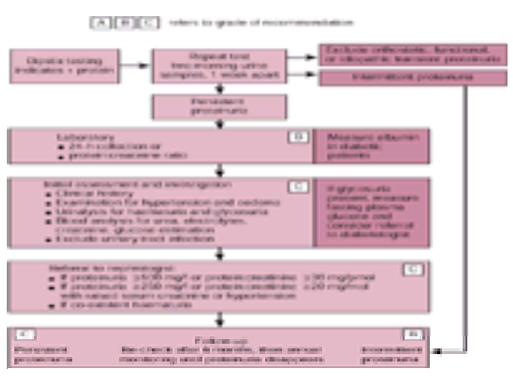

Fig. 1 Summary of the Scottish Intercollegiate Guidelines Network (SIGN) recommendations for the investigation of asymptomatic proteinuria in adults. Evidence required: A, at least one randomized controlled trial as part of the body of literature of overall good quality and consistency, addressing specific recommendations. B, availability of well-conducted clinical studies but no randomized clinical trials on the topic of recommendation. C, expert committee reports or opinions and/or clinical experiences of respected authorities. Note that these recommendations for investigation and referral are made solely on the basis that significant pathology may be detected. It is presently unproven whether there is any effect on patient outcomes. (Reproduced with kind permission from the SIGN Secretariat. Copies of the full text can be obtained from the SIGN Secretariat, Royal College of Physicians, 9 Queen Street, Edinburgh EH2 1JQ, Scotland.)

Microscopic haematuria

There is no agreed definition of microscopic haematuria as all urine samples contain some red blood cells. The Scottish Intercollegiate Guideline Network (SIGN) has suggested the presence of a positive result on dipstix testing and/or the presence of more than five red blood cells per high-power field on urine microscopy. Asymptomatic microscopic haematuria is occult haematuria, however detected, and excludes haematuria visible to the naked eye or associated with urinary tract pain, infection, or other symptom.

Asymptomatic microscopic haematuria is most commonly detected during routine medical examinations, such as for employment or insurance purposes, in well-person clinics, and when visiting general practitioner or hospital clinics where urine samples are examined by dipstix testing. The dipstix detects the peroxidase activity of haem from its reaction with orthotoluidine, which shows the presence of haemoglobin, whether in red cells or as 'free' haemoglobin, and myoglobin. False-positive results may be obtained from oxidizing contaminants (for example, from bacteria and from iodine or hypochlorite in antiseptic solutions). In obtaining urine samples for analysis it is important that the labia or prepuce are not first washed with any antiseptic solution. False-negative results may occur in the presence of excess ascorbic acid or the presence of rifampicin or phenolphthalein: if such substances are suspected to be present then a repeat sample should be obtained after appropriate withdrawal of these agents (see Chapter 20.4 for further discussion).

If a positive result is obtained it is important to repeat the test on a different occasion to confirm the result. Tests that are positive in women who are menstruating should be repeated between menses. In many patients stix testing for blood may be positive after exercise, particularly strenuous exercise, and a repeat sample should be obtained after 2 days without exercise. A positive result in the latter situation is probably due to the presence of myoglobinuria—urine microscopy is likely to be negative.

The further investigation of a patient found to have asymptomatic haematuria depends on a number of factors. As haematuria can arise from any part of the urinary tract from the glomerulus to the urethra, the first issue is to determine whether further investigations should be urological or nephrological. In men, particularly those over the age of 50 years, it is most likely that urological investigations will be required because of the increasing incidence of prostatic problems and urothelial malignancy, and in such patients clinical examination must include a rectal examination to determine whether any prostatic abnormality can be detected. By contrast, the presence of significant proteinuria, clinical evidence of renal disease, and/or impaired renal function indicate the need for nephrological investigation.

Urine microscopy can be of value if performed on a fresh sample. Red cell casts are diagnostic of glomerular bleeding and do not arise from bleeding anywhere else in the renal tract. Consideration of the morphology of red cells in the urine can also be useful (Fig. 2): dysmorphic red cells, in particular those appearing as a ring form with bubbles (Fig. 2(d)), are probably the consequence of glomerular bleeding, whereas red cells of normal appearance are more likely to arise from a site in the lower urinary tract. However, discrimination is not always straightforward (Fig. 2(a)), considerable interobserver variability is reported, and the technique is not robust enough to be routinely applied in most centres.

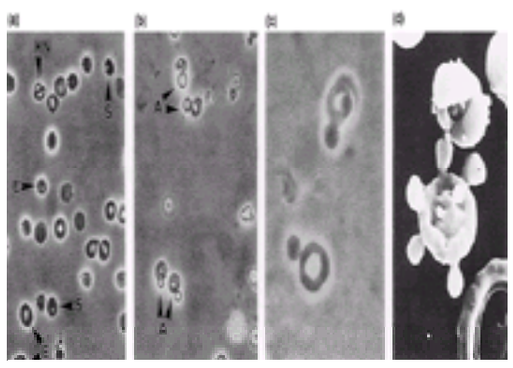

Fig. 2 Appearances of red cells in the urine. (a) phase-contrast microscopy (400 ×) showing a variety of non-specific 'dysmorphic' erythrocytes in acute interstitial nephritis. (b) phase-contrast microscopy (400 ×), (c) phase-contrast microscopy (1000 ×), and (d) scanning electron microscopy (5000 ×), all showing acanthocyturia (ring form and bubbles) in glomerulonephritis. A, acanthocyte; E, echinocyte; KN, knizocyte; S, stomatocyte. (Reproduced with permission from Kohler H, Wandel E (1993). *Nephrology, Dialysis, Transplantation* **8**, 879.)

Screening for and investigation of asymptomatic microscopic haematuria

Background

Should patients or the population in general be screened for microscopic haematuria, and what should be done if they test positive and this is confirmed on repeat testing? There is no consensus.

Studies of apparently healthy populations indicate that asymptomatic microscopic haematuria has an incidence of between 2.5 and 13 per cent, the incidence increasing with age, as does the chance of finding an underlying cause. In men over the age of 50 years conditions such as transitional-cell tumours, stones, outflow obstruction, and infections are the most common explanation. However, the chance of a patient with asymptomatic microscopic haematuria detected by population screening having a serious and curable condition is small. A retrospective study using dipstix to screen 10 050 men for asymptomatic haematuria in the United Kingdom found that 2.5 per cent tested positive. Questionnaires were subsequently sent to the general practitioners of all those who were registered, asking what further investigations had been performed. In 39 per cent (59 of 152 respondents) no further investigations had been undertaken. Abnormalities of some sort were found in 28 per cent of those who had undergone some investigation. This rose to 50 per cent in the few patients (24) who were 'fully investigated' by examination of an midstream urine (**MSU**) specimen, intravenous urography, and cystoscopy: two of whom were found to have bladder cancer. Long-term follow-up of all participants in this study would clearly be of great interest, but is not available. In another study, similar screening was performed in 20 571 men (35 years or older) and women (55 years or older) who were members of a prepaid United States health plan: 867 (4.2 per cent) tested positive, of whom 278 were known to have urological disease to account for this, leaving 589 (2.9 per cent) with newly discovered asymptomatic microscopic haematuria. Over the next 3 years two of these individuals developed prostatic cancer and one bladder cancer. However, the likelihood of developing urological cancer was the same in those whose urine had tested negative for blood on screening. The sensitivity of microhaematuria detected on a single dipstix analysis for urological cancer within 3 years was 2.9 per cent, specificity was 96.7 per cent, and positive predictive value was 0.5 per cent. Multivariate analysis that adjusted for age, gender, and race showed that the relative risk of 2.1 (95 per cent confidence intervals, 0.7 to 6.6) for urological cancer was not significantly increased among patients with asymptomatic microhaematuria compared with patients who had negative test results.

The picture derived from hospital-based studies is rather different. A recent study described the outcomes in 1930 patients referred to a urological clinic because of haematuria, including 982 with microscopic haematuria. Evaluation in all cases consisted of basic demographics, history and examination, routine blood tests, urinalysis and cytology, plain abdominal radiography, renal ultrasound, intravenous urography (**IVU**), and flexible cystoscopy. Of the 982 cases of microscopic haematuria, 53 (5.4 per cent) had cancer, including three with renal and one with urothelial carcinoma.

Recommendations

The report of the American Urological Association Best Practice Policy Panel on Asymptomatic Microhematuria in Adults suggested: 'that the patient's history and physical examination should help the physician decide whether testing is appropriate', and that 'patients with asymptomatic hematuria who are at risk for urological disease or primary renal disease should undergo an appropriate evaluation. In patients at low risk for disease, some components of the evaluation may be deferred'. Risk factors for significant disease are specified as smoking history, occupational exposure to chemicals or dyes, history of gross haematuria, age >40 years, history of urological disorder or disease, history of irritative voiding symptoms, history of urinary tract infection, analgesic abuse, and a history of pelvic irradiation. For patients with any of these risk factors 'complete evaluation' is recommended, comprising upper tract imaging, urinary cytology, and cystoscopy. However, in those with none of these risk factors and thereby at low risk of urological malignancy, the report does not specify clearly which components of the evaluation may be deferred, or for how long.

The Scottish Intercollegiate Guideline Network has also produced recommendations for the investigation of adults detected as having asymptomatic microscopic haematuria (Fig. 3). These do not, however, describe usual practice in many centres, and the implications of employing them would be considerable since they advocate radiological imaging (plain film and ultrasonography of the urinary tract, or intravenous urography) and cystourethroscopy for all those with genuine microscopic haematuria (that is to say, false and known benign positive causes excluded) who do not have clinical suspicion of renal disease. This would mean that (on the figures given above) up to 13 per cent of the population would require cystourethroscopy, which is scarcely sensible and certainly not practicable in most healthcare systems. Given the epidemiology of urothelial malignancy, all would recommend such investigation in patients over the age of 50 years with asymptomatic microscopic haematuria. Some would advocate cystourethroscopy in those over the age of 45, or sometimes 40 years, but many renal physicians and urologists would not do so in their routine practice in those under this age. However, all have seen unfortunate cases of urothelial malignancy in young patients, and prudent instruction to all with microscopic haematuria is that they should report any new symptoms, in particular macroscopic haematuria, immediately. It is also worth noting that knowledge of the length of time for which microscopic haematuria is or has been present is very helpful: malignancies declare themselves eventually, and the patient who is documented retrospectively or prospectively to have microscopic haematuria for more than a few years can almost invariably be reassured with confidence that they do not have urinary tract malignancy.

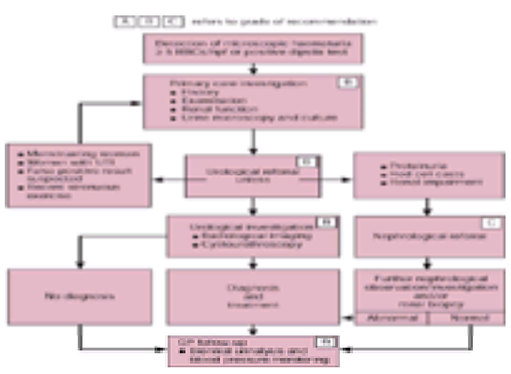

Fig. 3 Summary of the Scottish Intercollegiate Guidelines Network (SIGN) recommendations for the investigation of asymptomatic haematuria in adults. See text for further discussion and caption to Fig. 1 for explanation of grades. (Reproduced with kind permission from the SIGN Secretariat. Copies of the full text can be obtained from the SIGN Secretariat, Royal College of Physicians, 9 Queen Street, Edinburgh EH2 1JQ, Scotland.)

Symptomatic presentations

Nephrotic syndrome

The nephrotic syndrome is a common mode of presentation of glomerular disease. It is not a diagnosis but rather a term used to describe a clinical syndrome that arises when the urinary protein excretion is sufficient to produce hypoproteinaemic oedema. The correlation between urinary protein excretion, plasma albumin concentration, and the presence of oedema is poor. In adults it is uncommon to develop oedema unless the plasma albumin is less than 30 g/l, but many patients will remain oedema-free with a plasma protein of less than 25 g/l. Similarly, the degree of proteinuria needed to cause hypoprotinaemia is variable: some quote a urinary protein excretion of 3.5 g/day as being 'nephrotic', but it is best to avoid a particular value as some patients develop oedema with an excretion of less than 3.5 g/day, whereas others will excrete greater amounts and remain oedema-free.

Clinically, the presentation is with peripheral oedema. In adults this is commonly of the lower limbs and tends to be progressive throughout the day, in children facial oedema is more common. Ascites and pleural effusions may be present in severe cases. Some patients may mention that they have noticed their urine becoming frothy. Commonly, patients complain of an inexplicable tiredness and lethargy: the cause for which is unknown. Other clinical features include anorexia, muscle wasting, susceptibility to infections, and nail changes (with prolonged severe nephrotic syndrome the nails become white). These latter symptoms and signs are uncommon: they are a feature of prolonged protein loss and the majority of patients present for investigation before they develop. Hypertension and impairment of renal function, if present, will, to a large extent, depend on the underlying cause of the syndrome.

Pathophysiology

Mechanisms of proteinuria

The mechanisms whereby the filtration of protein by the glomerular capillary wall is restricted is poorly understood. The barrier consists of an inner endothelial cell, presenting little in the way of restriction to filtration, a basement membrane consisting of matrix proteins arranged in a complex three-dimensional pattern that allows the passage of small molecules by convection and of larger molecules by diffusion, and an outer cellular layer of epithelial cells producing a complex barrier to filtration. There is also a charge on the surface of the basement membrane due to the presence of negatively charged heparan sulphate that allows the penetration of cationic molecules to the basement membrane but repels anionic (negatively charged) molecules. In addition, the podocytes of the epithelial cells play an important role in controlling glomerular filtration: these adhere to the outer surface of the basement membrane, there being a complex interaction between the podocyte and the basement membrane in maintaining glomerular wall integrity. In patients with proteinuria there is frequently flattening and detachment of the podocyte of the epithelial cell from the basement membrane. The pathogenesis of this is unknown. Proteinuria most likely results from a combination of disruption of the charge barrier, an alteration in the spatial configuration of the basement membrane, and disruption of the normal podocyte–basement membrane interaction. A greater understanding of this may lead to improved methods of controlling proteinuria. (See Chapter 20.1 for further discussion.)

Oedema formation

The mechanisms of oedema formation in the nephrotic syndrome are complex. Initially the loss of protein in the urine will stimulate an increased production of proteins by the liver, but if the proteinuria is sufficient the loss will exceed the capacity to replace them and so the plasma protein concentration will decline. The diminution in plasma protein concentration reduces the plasma oncotic pressure and, as a result, there is accumulation of fluid in the extravascular space, resulting in oedema. Traditional teaching proposes that this is accompanied by a reduction in the intravascular volume and, as a consequence, there is reduced renal perfusion. This results in enhanced renin secretion, and thus through the renin–angiotensin–aldosterone system there is increased sodium retention by the distal tubule. The reduced intravascular volume also stimulates ADH secretion, the net result of these effects being an avid retention of salt and water, which because of the continuing low oncotic pressure increases oedema formation. This proposed mechanism is attractive but does not adequately explain the clinical findings. In the majority of patients with the nephrotic syndrome the plasma volume is not diminished, and in some it is even expanded. In addition, in many patients with the nephrotic syndrome there does not appear to be increased plasma renin, and, furthermore, blocking of the renin–angiotensin system with angiotensin-converting enzyme (**ACE**) inhibitors is not accompanied by a diuresis. It is possible that the retention of sodium and water is due to an intrarenal mechanism involving proximal and distal tubular function, in addition to physical factors such as peritubular oncotic pressure. Hence, oedema formation does not appear to have a uniform mechanism and it is more than likely that it results from a complex of effects within the kidney, the intravascular volume, and possibly the peripheral capillary integrity.

Investigation of patients with the nephrotic syndrome

The presence of the nephrotic syndrome indicates that the patient has at least one of a wide range of glomerular pathologies, which may reflect primary or secondary renal disease. The different conditions have very variable prognoses and responses to treatment, hence making a precise diagnosis is essential to guide management. Aside from the estimation of GFR (usually estimated from serum creatinine by the Cockcroft and Gault formula, or from a 24-h creatinine clearance), quantitation of proteinuria, and measurement of serum albumin, the routine investigation of all patients with the nephrotic syndrome should include a full blood count, tests for systemic lupus erythematosus (**SLE**) (antinuclear antibody tests, anti-double-stranded DNA (- **dsDNA**), estimation of serum complement, hepatitis B and hepatitis C serology, serum immunoglobulins, and protein electrophoresis. Further serological tests may be indicated in some cases. Given the association of membranous glomerulonephritis with malignancy (see Chapter 20.7.9), older patients with the nephrotic syndrome should undergo chest radiography, and there should be a low threshold for the investigation of gastrointestinal symptoms. However, a precise diagnosis of most cases of the nephrotic syndrome can only be made histologically, and renal biopsy should be performed after checking platelets and a coagulation screen, and imaging—usually by ultrasonography—to confirm the presence of two anatomically normal kidneys. Such extensive investigation is not required when a confident diagnosis of the cause of the nephrotic syndrome can be made on clinical grounds, for example in the patient with diabetes and long-standing proteinuria who becomes nephrotic.

Treatment of patients with the nephrotic syndrome

Specific causes of the nephrotic syndrome may require specific treatments: these are discussed in the relevant subsections of Section 20.7, but some general measures are applicable to all nephrotic patients.

Patients with the nephrotic syndrome may feel generally 'washed out' and exhausted, and they may suffer psychologically from uncertainties and fears surrounding their diagnosis and prognosis, but the main concern in all cases is likely to be oedema. This can be massive: over 10 litres of excess fluid is not infrequent, and some have over 20 litres, when the patient is bed-bound with massively swollen and weeping legs, distressing genital oedema, and pitting of the abdominal and sometimes the chest wall.

Patients with the nephrotic syndrome are unable to excrete salt or water normally, hence it is prudent to recommend moderation in the consumption of both. Strict limitation of salt intake renders the diet unpalatable, but patients should be advised not to add salt to food at the table and to avoid foods that are rich in salt. It seems reasonable to suggest a total fluid intake of no more than 1.5 litres per day for those who are very oedematous, and perhaps no more than 1 litre per day in the most severe cases. The mouth can be kept moist using swabs or by sucking boiled sweets, and the daily fluid ration will go much further if the patient is given an ice cube to suck rather than a jug of water if they feel thirsty.

Diuretics are the mainstay of oedema removal, with loop diuretics often the only effective agents. The aim should be to reduce the patient's weight by 0.5 to 1 kg daily. Oral frusemide (furosemide), 40 to 80 mg daily (or other loop diuretic), would be the usual starting dose, but some patients require much higher doses, with up to 250 mg twice daily not uncommon. If oral frusemide proves ineffective, then the addition of oral spironolactone (usually 100 to 200 mg daily, with particularly close monitoring of the serum potassium level) or oral metolazone (usually 2.5 to 20 mg daily, with close monitoring to ensure that it is stopped promptly in the event of massive diuresis) can be helpful. If these fail then admission to hospital for bed rest and intravenous diuretic (usually frusemide 250 to 500 mg) is required, and some would also give daily intravenous infusions of concentrated albumin. There is little evidence that the latter is effective in adult practice, but there is theoretical justification in giving it to those who appear to have intravascular volume depletion (postural hypotension, low jugular venous pressure), and it is reasonable to give it (for example, human albumin solution 20 per cent, 100 ml daily) as a therapeutic trial for a few days to all with severe refractory nephrotic oedema. However, many studies would suggest that the likely effect is simply an increased proteinuria, and if this is indeed the case and no benefit in terms of diuresis/weight loss is seen, then albumin infusion should be abandoned. It is interesting to note that when nephrotic oedema has been removed the patient is often able to sustain a new steady state with a much reduced burden of oedema, with no change in serum albumin or proteinuria.

A number of agents, including ACE inhibitors, non-steroidal anti-inflammatory agents, and ciclosporin, can reduce proteinuria at the expense of some reduction in the glomerular filtration rate. These drugs, most commonly ACE inhibitors, are sometimes used for this purpose (after deliberately inducing volume depletion with diuretics) in managing patients with severe nephrotic syndrome when the difficult judgement has been made that the benefits of reducing proteinuria (and hopefully thereby oedema) more than outweigh the disadvantage of reducing the GFR. This is a step that should never be taken lightly: the reduction in GFR may induce endstage renal failure (so-called 'medical nephrectomy'), and whilst renal replacement therapy may be preferable to intractable massive oedema, this possibility clearly needs to be thoroughly discussed with the patient beforehand.

There is considerable debate about the protein intake that should be recommended for those with the nephrotic syndrome: protein restriction diminishes proteinuria and a high protein diet increases it, but the impact on oedema or long-term prognosis of either manoeuvre is unknown. It is probably reasonable to suggest that patients with the nephrotic syndrome should consume about 1 g/kg per day of mainly first-class protein.

The management of specific complications of the nephrotic syndrome such as infection, thromboembolism, and hyperlipidaemia are discussed below.

Complications of the nephrotic syndrome

A number of complications are recognized in patients with the nephrotic syndrome: these result from the metabolic consequences of prolonged protein loss.

Infections

The incidence of infections, particularly bacterial, is increased in the nephrotic syndrome. Peritonitis is well recognized in children and is an important cause of

mortality. Cellulitis, particularly streptococcal, is common and may spread rapidly. The cause of the increased susceptibility to infections is a combination of physical factors such as the accumulation of fluid in the interstitial space, the peritoneal and/or pleural space, and the impairment of defence mechanisms, such as the reduction in immunoglobulin concentration and impaired white cell function. Impairment of the alternative pathway of complement activation through loss in the urine of Factor B, which has a molecular weight of only 55 kDa, is crucial in causing impairment of the phagocytosis of encapsulated organisms. This leads in childhood to the particular vulnerability to *Streptococcus pneumoniae* described above: most adults are protected by virtue of having acquired antibodies against a variety of pneumococcal capsular antigens.

Prompt induction of remission of oedema and proteinuria is the best method of preventing infection. When this cannot be achieved, good skin care is important, and there is a good case for using prophylactic penicillin to prevent pneumococcal infection in oedematous children. Antipneumococcal vaccines against capsular antigens induce an adequate response when given to patients in remission. Any suspicion of infection should be treated aggressively in those with the nephrotic syndrome, with a low threshold for starting a parenteral antibiotic regimen, including benzylpenicillin in children, as soon as necessary cultures have been taken.

Thromboembolism

There is an increase in both arterial and venous thromboses in patients with the nephrotic syndrome, which may be aggravated by the use of excessive diuretic therapy and corticosteroids. There is an increase in the plasma concentration of a number of factors involved in the coagulation cascade, notably fibrinogen, and in addition plasminogen concentration is commonly reduced, thereby increasing the tendency to thrombus formation. Other factors include increased platelet aggregation and alterations in endothelial-cell function.

The increased thrombotic tendency is manifest by an increased risk of deep leg vein thromboses and pulmonary embolism and an increase in arterial thrombotic episodes. Deep leg vein thrombosis is clinically evident in about 6 per cent of nephrotic adults and can be detected in about 25 per cent if Doppler ultrasonography is used. Pulmonary embolism is also clinically evident in around 6 per cent of cases, but if ventilation/perfusion scans are used as a screening test the figure rises to 12 per cent, or even higher in some series. Nevertheless, mortality from thromboembolism appears to be low (only one death in 2100 years of patient follow-up in a series reported by Cameron) and very few nephrologists would routinely anticoagulate all patients with the nephrotic syndrome, although most would have a low threshold for doing so (for instance, during periods of immobility in hospital). Likewise, after any thromboembolic event they would advise anticoagulation for as long as the nephrotic state persisted.

Renal vein thrombosis may occur, presenting acutely with loin pain, haematuria, deterioration in renal function, and with swelling of the kidney detectable on imaging (usually by ultrasonography). It can also present more insidiously with a significant increase in urinary protein excretion and a gradual reduction in renal function. Some one-third of cases are associated with pulmonary embolism. Renal vein thrombosis can be associated with all causes of the nephrotic syndrome but, for reasons that are unknown, is most common in membranous glomerulonephritis, when it is clinically apparent in 6 to 8 per cent of cases and detectable on imaging in 10 to 45 per cent. It is not routine practice to investigate patients with the nephrotic syndrome, even that caused by membranous glomerulonephritis, for renal vein thrombosis: when there is clinical suspicion the diagnosis can be made by magnetic resonance imaging (**MRI**), computed tomographic (**CT**) scanning or renal arteriography (looking at the venous phase). Treatment of symptomatic renal vein thrombosis is by anticoagulation with full-dose therapeutic intravenous heparin or low molecular weight heparin (although there is little experience of using the latter in this condition) followed by warfarinization. Thrombolysis has also been used, although the indications for this are not well defined. The prognosis is usually benign, with recanalization of the veins and recovery of renal function.

Alterations in lipid metabolism

Alterations in lipid metabolism are well recognized in patients with nephrotic syndrome. There is concern that these changes might lead to atheromatous vascular disease and also be an adverse factor with respect to the development of progressive renal function impairment. Moreover, there is some evidence that patients with the nephrotic syndrome have an increased incidence and prevalence of vascular disease, particularly coronary vascular disease leading to ischaemia and infarction, although not all studies have confirmed this.

Patients with the nephrotic syndrome have an increased concentration of both free cholesterol and cholesterol esters, which have an inverse correlation with plasma albumin concentration. There is an increase in phospholipids, although this is not marked, and in severe cases fasting triglycerides are elevated. Plasma free fatty acid concentrations are reduced. The changes are not determined by the nature of the underlying glomerular disease but are closely linked to the degree of hypoproteinaemia. There are alterations in lipoproteins: very low-density lipoproteins (**VLDL**) and low-density lipoproteins (**LDL**) are increased, whilst high-density lipoprotein (**HDL**) concentrations are less predictably affected.

The causes of the alterations in the lipid profile are incompletely understood. There is an increase in the hepatic production of VLDL and LDL, the stimulus for which is unknown but may be related to plasma oncotic pressure. In addition there is evidence for the diminished removal of LDL, possibly due to a reduction in the activity of lipoprotein lipase. The changes in HDL concentrations are probably related to the reduction in lecithin cholesterol acyl transferase (**LCAT**) activity that occurs in nephrotic patients.

Should patients with nephrotic hyperlipidaemia be treated for this? There is no good evidence. It seems reasonable to give dietary advice, although there have been few studies to show how effective (or not) this is in reducing lipid levels, and none looking at cardiovascular outcome. Statins are effective at reducing serum cholesterol, which is virtually always elevated in the nephrotic syndrome, sometimes massively so, but most would not prescribe these as a routine for the following reasons:

1. Given that the aetiology of hyperlipidaemia in the nephrotic syndrome is different from that in the primary hyperlipidaemias, it is not certain that the biological impact is the same, for example it might not be as deleterious from a cardiovascular point of view.
2. Few patients remain nephrotic for years: they either go into remission (spontaneously or after treatment) or develop endstage renal failure.
3. Many patients with the nephrotic syndrome are young and, aside from their hyperlipidaemia, at low absolute risk of vascular events.
4. Many patients will require complex drug therapy, such that there is a reluctance to add further medication, particularly since there is a suspicion that those with the nephrotic syndrome may be more likely to experience side-effects from lipid-lowering medications.

Proteinuria and the progression of renal disease

It is well recognized that prolonged profuse proteinuria is associated with a poor prognosis in patients with glomerulonephritis. In experimental studies proteinuria is associated with interstitial damage, tubular injury, and glomerulosclerosis. Heavy proteinuria is associated with progressive glomerular scarring, and intervention to reduce proteinuria is accompanied by reduction in sclerosis. It is possible that the prolonged and heavy proteinuria has an adverse effect on mesangial cells leading to mesangial sclerosis and eventually global sclerosis, but it is likely that proteinuria is only one of many factors responsible for progressive renal disease.

Acute nephritic syndrome (haematoproteinuria syndrome)

Acute nephritis is a clinical syndrome characterized by the acute onset of haematuria, proteinuria, hypertension, and oliguria. The urine typically appears 'smoky' due to the presence of red blood cell casts, and rarely it will appear frankly red. Proteinuria is variable in amount, but is rarely sufficient to produce a nephrotic syndrome. Hypertension is variable and oliguria depends to a large extent on the degree of glomerular involvement. Not all four clinical features may be present simultaneously. In some patients there is oedema due to salt and water retention in the oliguric phase. Encephalopathy, particularly in children, may occur due to hypertension or electrolyte disorders such as hyponatraemia.

The 'classical' cause of acute nephritis is poststreptococcal glomerulonephritis: this is described in detail in Chapter 20.7.5 and other infective causes in Chapter 20.7.8. However, these diseases are becoming less common, particularly in developed countries, and it is more usual to see patients who have proteinuria and haematuria accompanied by variable hypertension and renal functional impairment in whom no identifiable preceding infection can be identified. The presence of blood and protein in the urine is a sign of glomerular inflammation and is not indicative of any particular glomerular pathology. On investigation such patients have a wide variety of glomerular appearances (Table 1), hence renal biopsy is essential for precise diagnosis. For further discussion of the conditions listed in Table 1 see the relevant subsections of this section.

Recurrent haematuria

'Recurrent haematuria' is the term used to describe patients who have episodic macroscopic haematuria. This most commonly is due to IgA nephropathy, described in Section 20.8.2, where episodes are immediately preceded by mucosal inflammation, usually of the upper respiratory tract but sometimes of the gastrointestinal tract. Some patients with Alport's syndrome may present with episodes of recurrent haematuria but these are usually unrelated to mucosal inflammation. It must also be remembered that other causes of repeated episodes of macroscopic haematuria include polycystic renal disease, renal stone disease, sickle-cell disease, and tumours of the renal tract. In such patients there are frequently other clinical indications to allow confident differentiation from IgA nephropathy.

Loin pain haematuria syndrome

This is a relatively uncommon but well-recognized syndrome, which can only be diagnosed by excluding other conditions that can be associated with pain in the loins and microscopic or intermittent macroscopic haematuria (Table 2).

Patients present with intermittent or persistent loin pain accompanied by persistent microscopic haematuria, although on very rare occasions macroscopic haematuria can occur, even to the extent of causing clot colic. Most are young women, many of whom will have undergone numerous investigations. Symptoms are frequently unilateral but usually become bilateral, although one side may predominate. Clinical examination is unremarkable, although there may be some loin tenderness, and blood pressure is normal. The urinary red cells are dysmorphic, suggesting a glomerular origin, but red cell casts are not present. There is no evidence of urinary tract infection, normal urinary protein excretion or only minor proteinuria (occasionally up to 1 g daily), normal GFR (usually estimated from serum creatinine values by the Cockcroft and Gault formula, or from a 24-h creatinine clearance), and no clear structural abnormality on thorough imaging (cystoscopy and at least one of intravenous urography/ultrasound/CT scanning/MRI). Indices of inflammation such as a raised erythrocyte sedimentation rate (**ESR**) or C-reactive protein (**CRP**), or elevated plasma viscosity are notable by their absence.

A radiological abnormality consisting of focal or generalized tortuosity, beading, and occlusion of intrarenal medium-sized arteries leading to cortical infarcts has been described, but these findings are not universal. On renal biopsy the glomeruli appear normal or show only minor changes: a number of reports describe complement deposition (C3 and C4) in arterioles, but the significance of this is uncertain and it may be a non-specific finding.

The syndrome runs a relapsing/remitting course over a number of years, with symptoms gradually subsiding with time. The loin pain can vary greatly in severity, ranging from a mild ache to disabling colic requiring opiate analgesia. There is no effective specific treatment and, not surprisingly, some patients can be difficult to manage as it is difficult for them to accept that there is no 'cure' for their symptoms. Renal denervation and renal autotransplantation have been tried, but pain typically returns over 6 to 12 months, and even after nephrectomy pain can develop in the remaining kidney. These procedures should not be performed, and other methods of pain control (transcutaneous nerve stimulation, amitriptyline, carbamazepine, and similar agents) are usually ineffective. In most cases pain control can be obtained using regular opiate analgesia, after which the dosage can usually be reduced and the patient eventually weaned off medication. Although this can take a long time, it is rare for pain to persist into the patient's forties or fifties.

Disorders of micturition

Frequency

'Frequency' is the term applied when the bladder is emptied more often than normal, hence in obtaining a history it is therefore necessary to determine how often the patient passes urine. This may be associated with a normal or increased 24-hour urine volume. It is important to distinguish between these two situations as frequency in the presence of a normal output indicates a bladder (lower urinary tract) problem, whereas an increase in output is indicative of a disorder of urinary concentration or excessive fluid intake (see 'polyuria').

Frequency in the presence of a normal urine volume is most commonly due to bladder inflammation from a bacterial infection (cystitis), when dysuria is a common accompanying symptom. It can also be produced by chemical irritation (for example, as sometimes occurs during treatment with cyclophosphamide), or from a calculus or tumour involving the bladder wall. A reduction in bladder capacity is uncommon but may result from radiation-induced fibrosis following treatment to a pelvic malignancy. In males, prostatic hypertrophy, benign or malignant, is associated with frequency and a diminution in urinary stream, together with hesitancy (a difficulty in initiating micturition) and dribbling (a difficulty in terminating micturition).

Nocturia

Nocturia may arise from the many conditions that cause frequency. On lying down there is an increase in renal perfusion resulting in increased urine flow, but ADH is secreted during sleep, thereby increasing urinary concentration and meaning that urine volume diminishes during sleep. In patients with sleep disturbance there is less ADH production and thus urine concentration is reduced, with increased urine volume such that nocturia may occur. Enquiry should be made regarding sleep patterns in patients presenting with nocturia, in addition to considering those conditions that cause polyuria and frequency.

Dysuria

Dysuria is pain or discomfort on micturition and one of the most frequent symptoms, accounting for about 2 per cent of consultations in primary care. It is more common in women, and is usually described as a burning, scalding, or tingling sensation in the urethra or at the urethral meatus occurring during or immediately after micturition. Most commonly it is due to urinary infection, but it may also be caused by chemical irritation such as rarely occurs with cyclophosphamide. If associated with frequency and urgency of micturition it indicates bladder irritation such as cystitis. In young women this is usually associated with sexual activity, but in older persons it may indicate a lesion in the bladder or prostate. Prostatic inflammation usually gives rise to perineal or rectal pain. Very young children will be unable to complain of dysuria but urethral irritation may be inferred if the child cries during micturition. (See Chapter 20.12 for further discussion.)

Polyuria

Polyuria is an increase in the daily volume of urine and may arise from a number of different conditions. The normal daily urine volume varies considerably depending on fluid intake and insensible loss, but is normally in the range of 1 to 2 litres. Most patients have no idea of their urine volume and so it is necessary to obtain a 24-h collection to verify urine output. Excessive fluid intake, as occurs in compulsive water drinking, results in an increased volume. An increase in solute load, most commonly due to hyperglycaemia, reduces tubular reabsorption and increases urine production. Inadequate ADH secretion, such as following a head injury or associated with tumours or infection, result in an impaired urinary concentration and increased output (central diabetes insipidus). Conditions that impair the tubular response to ADH, such as potassium depletion, lithium toxicity, and some rare inherited diseases, also increase urine volume (nephrogenic diabetes insipidus), as do renal disorders that impair medullary concentration, such as analgesic nephropathy, papillary necrosis, medullary cystic disease, and nephrocalcinosis.

Oliguria and anuria

Oliguria is a reduction in urine volume to such an extent that there is inability to excrete the residues of normal daily metabolic functions. This normally means to a volume of less than 400 ml daily in an adult, usually indicating acute renal failure of whatever cause (see Chapter 20.4). Anuria is the lack of any urine output and is indicative of obstruction, although it may occur in some forms of severe acute renal failure. If anuria is present it is essential to perform a rectal examination to determine if there is any pelvic malignancy, such as a rectal or cervical carcinoma, to account for the obstruction.

Pain

Renal pain

Stretching of the capsule of the kidney causes renal pain that is felt in the loin ('renal angle'). It can be produced by any condition that distends the kidney, such as inflammation, mass lesions, or an obstruction. The last is the most common cause, particularly obstruction of the pelviureteric junction, when the patient may give a history that anything that causes an acute increase in urine volume (for example, drinking a large quantity of water, beer, or lager or taking a diuretic) precipitates the pain. Inflammatory pain, such as in pyelonephritis and (uncommonly) in glomerulonephritis, develops gradually, is usually constant in nature, and is variable in severity. A perirenal abscess, which may not always be associated with fever or tenderness, can give rise to symptoms and signs of diaphragmatic irritation and/or

psoas irritation. In the latter case, the patient usually prefers to rest with the hips flexed, and reports that extension of the hips is accompanied by an increase in pain.

It can be difficult to distinguish renal pain from musculoskeletal pain, hence the history should enquire specifically about the relationship of pain to movement or position, neither of which greatly affects renal pain. Clinical examination of the back and spine should determine any limitation of movement or localized point tenderness, which would suggest a musculoskeletal problem.

Some patients with polycystic renal disease complain of a constant dull loin ache. They may also suffer from the sudden onset of renal pain if there is bleeding into a cyst, or from pain of a more gradual onset if there is cyst infection.

Ureteric colic

Pain arising from an acute obstruction is frequently sudden in onset, severe, colicky, and may radiate to the groin, scrotum, labia, or upper thigh. Many describe it as 'the worst pain that they have ever had', and the patient with ureteric colic typically thrashes about, unable to find comfort, looks pale and sweaty, and often vomits, which can lead to diagnostic confusion. The pain is due to acute distention of the pelvis of the kidney and the upper ureter and the associated increased peristalsis. If the obstruction is ureteric the pain resolves rapidly once the cause is extruded into the bladder, although when in the bladder it may result in bladder irritation with strangury or further obstruction if it becomes impacted at the urethral orifice. The most common differential diagnoses of right-sided renal colic are biliary colic and appendicitis: diagnostic difficulty is less likely on the left side, although colonic pain requires consideration. (See Chapter 20.14 for further discussion.)

Chronic obstruction may be surprisingly asymptomatic. Retroperitoneal fibrosis is accompanied by a dull-aching back discomfort but is not associated with colic in spite of an obstruction.

Disorders of renal function

Acute renal failure

An acute deterioration in renal function may arise in patients with normal renal function or in those with known renal insufficiency, the latter being known as acute on chronic renal failure. The clinical features depend to a large extent on the underlying cause. Patients may be seriously ill with profound hypotension from such causes as multiple trauma or severe sepsis, or they may appear remarkably well, such as with rapidly progressive glomerulonephritis. The diagnosis may be suspected or proven in patients with the following clinical features:

1. *Diminished urine volume.* This is not invariable and some patients have a normal volume but a reduction in urinary concentration to such an extent that there is retention of urea, creatinine, and other substances that are normally excreted. It should be remembered that the quality of urine is as important as the volume.
2. *Increasing blood concentrations of urea and creatinine.* This usually indicates impaired excretion and is the usual way in which the diagnosis of acute renal failure is established. Creatinine is a more accurate reflection of renal function than urea, as its concentration is influenced to a lesser extent by protein catabolism and the state of hydration. An important clinical indicator as to whether a patient has acute or chronic renal failure can be the state of consciousness at a given serum creatinine concentration. If renal function declines rapidly, the patient is often obtunded once the creatinine is in excess of 800 µmol/l, whereas if the creatinine concentration has increased slowly with time the patient is unlikely to have any impairment of consciousness with a creatinine of 1000 µmol/l.
3. *Electrolyte disturbance.* Hyperkalaemia may be the first indication that a patient is developing acute renal failure. In the presence of normal renal function it is difficult to produce hyperkalaemia.
4. *Acidosis.* This is most commonly detected by measurement of a declining serum bicarbonate concentration, reflecting the development of a metabolic acidosis and clinically manifested by tachypnoea.
5. *Pulmonary oedema.* Most patients with acute renal failure are prone to this as their ability to excrete fluid is limited, but it is most commonly iatrogenic, when an oliguric patient is ill-advisedly given inappropriate intravenous fluids.
6. *Appropriate clinical circumstances*. In some instances there is an obvious risk for the development of acute renal failure, such as in septicaemic shock, severe multiple trauma, prolonged hypotension, and indeed any clinical circumstance that impairs normal kidney function.
7. *Consumption of certain medications.* Some medications have the ability to reduce renal function. The most common mechanism is by reduction of effective renal perfusion, such as can be caused by non-steroidal anti-inflammatory agents (**NSAIDs**), ACE inhibitors, and angiotensin II-receptor blockers. Less common mechanisms include the induction of acute interstitial nephritis (for example, caused by NSAIDs, penicillins) or acute toxic tubular dysfunction (for example, aminoglycosides, paracetamol overdose). Rare mechanisms include the precipitation of acute vasculitis (for instance, hydralazine) or stimulation of retroperitoneal fibrosis (for example, methysergide). Commonly it is a combination of medications such as analgesics, antibiotics, and non-steroidal drugs together with radiocontrast agents that result in acute renal failure, because during an illness one agent after another is added without the recognition that renal function is declining.

It must be remembered that to maintain normal renal excretory function there is a need for effective cardiac output and renal perfusion, glomerular filtration, tubular function (reabsorption and excretion), and the drainage of urine from the renal pelvis via the ureters, bladder, and urethra. It is thus possible to acutely disrupt this function in many different ways, and accurate diagnosis is dependent upon obtaining a good history, paying particular attention to detail, followed by a thorough clinical examination supported by a structured investigative protocol. See Section 20.5 for further discussion.

Chronic renal failure

Patients with a slow progressive deterioration in renal function may remain remarkably free from symptoms until renal function is seriously impaired. When symptoms of renal failure do develop, they are non-specific, such that many patients attribute them to increasing age or being less physically active than previously. This can make diagnosis difficult and renal impairment may be unsuspected until the results of some screening investigations, such as haemoglobin and blood urea, are obtained. (See Section 20.5 for further discussion.)

History

Presenting complaint

Obtaining a clear and concise history is about one of the most difficult things to achieve in clinical medicine. It is important to elicit carefully the nature and chronological sequence of all the symptoms experienced by the patient, who should be encouraged to describe each symptom in detail and—if medical terminology is used—the physician needs to determine what exactly the patient understands by the particular word or phrase used. It is not uncommon for a patient to describe their symptoms using words that their friends or previous medical practitioners have used, which may not be appropriate. For each symptom it is necessary to make specific enquiry regarding the mode of onset, whether sudden or gradual, and the nature of any associated features such as precipitating and relieving factors. If the patient describes pain then additional enquiry needs to be made regarding its position, nature, and radiation.

In patients with proteinuria and/or haematuria it is essential to ask about the results of any previous urine testing, even from years previous. It is standard practice to test urine during pregnancy and during employment and insurance medical examinations. It is often profitable to review all available medical notes: frequently urine will have been tested and the results recorded, even if no notice has been taken of an abnormal result.

In most instances the pattern of the symptoms will follow a recognized syndrome; in some an apparently unconnected number of symptoms will suggest the presence of a systemic disease; while in others there will be a condition known to affect the kidneys. In any event, it is important to document carefully all symptomatology and subsequently to record all new or changing symptoms.

Past history

The past medical history of a patient can provide much useful information to aid diagnosis. It is important to obtain as full a history as possible detailing childhood illnesses, all major illnesses, and hospital admissions. A history of unexplained febrile illnesses in childhood or prolonged enuresis suggests the possibility of recurrent urinary infections due to structural urinary abnormalities such as vesicoureteric reflux.

A number of chronic conditions may be associated with renal involvement, either directly as in systemic diseases, or indirectly as in the development of amyloidosis in prolonged inflammatory diseases such as chronic osteomyelitis, bronchiectasis, or rheumatoid arthritis. In addition, certain drugs used to treat chronic conditions may have adverse renal effects such as gold, penicillamine, ciclosporin, NSAIDs and analgesics used in the management of rheumatoid arthritis and similar conditions.

Systemic vasculitides frequently have multisystem manifestations with variable renal involvement. As a general rule the prognosis of any systemic disease is significantly and adversely influenced once there is renal involvement. Diabetes mellitus, either insulin or non-insulin-dependent, may result in glomerulopathy. In the patient with insulin-dependent diabetes the first manifestation of glomerular involvement typically occurs about 10 years after onset. In the non-insulin-dependent patient, however, renal involvement may be detected at or soon after diagnosis, not because of more aggressive glomerular involvement but rather because the condition may well have been present in an undiagnosed form for many years. (See Chapter 20.10.1 for further discussion.)

It is important to obtain details regarding a woman's menstrual, contraception, and pregnancy histories. A delayed menarche may indicate the presence of renal failure, as may the unexplained development of amenorrhoea. Hypertension is more likely in patients taking a combined oestrogen–progesterone oral contraceptive if there is an underlying renal disease. Similarly, hypertension during pregnancy is much more common in patients with renal disease, particularly in the third trimester. Patients with asymptomatic proteinuria will have an increase in proteinuria during pregnancy, which may come to medical attention when the urine is tested for the first time or because it becomes sufficient to produce the nephrotic syndrome. In the latter circumstance it is sometimes difficult to differentiate between a simple nephrotic syndrome and the development of pre-eclampsia. The proteinuria of pre-eclampsia resolves following delivery, usually with 3 months, whereas in patients with glomerular disease, although the proteinuria will diminish, it does not completely resolve (see Chapter 13.5). Recurrent fetal loss may be associated with antiphospholipid antibodies and so prompts appropriate investigations.

It is not uncommon to find that patients are unable to remember significant events in their past medical history. It is important, therefore, to obtain any previous medical notes to obtain the maximum information. It is surprisingly common to find that urinary abnormalities such as proteinuria or haematuria have been documented previously, perhaps during medicals for work or insurance purposes, but not investigated. This can be a reassuring finding, because if it is known that haematuria has been present for many years then it is a safe bet that it is not due to a serious condition (assuming that renal function is normal).

Drug history

It is essential to obtain a detailed history of all medications recently consumed, whether obtained by prescription, over the counter, or from health shops: all have the potential for precipitating adverse renal effects. Many patients do not consider that medications they can buy from their local chemist, such as analgesics, are drugs. In some patients it may be necessary to obtain information from the family practitioner. There may be covert drug use, such as with laxatives and diuretics, which is particularly difficult to detect. The presence of unexplained hypokalaemia may provide the only clue that the patient is abusing diuretics or taking excessive amounts of liquorice (see Chapter 20.2.2). Frequently the patient has some connection to the medical profession and thus access to medications. In addition, some who take diuretics unnecessarily experience oedema on withdrawal of the drug, reinforcing their belief that such medication is necessary and resulting in continued ingestion.

Factors that increase the risk of adverse drug effects include age, impaired renal function, and multiple drug therapy. The elderly are at particular risk as renal function declines with age, and with a diminishing muscle mass this may not be obvious from an estimation of the serum creatinine alone; they may also have multiple pathologies resulting in an increased chance that they are taking many drugs. The pharmacokinetics of drugs are altered in elderly patients and so caution is required when prescribing.

All compartments of the kidney can be involved in adverse drug reactions. In the situation where there is a constraint on renal blood flow there is enhanced secretion of renin from the juxtaglomerular apparatus, activating angiotensin and thereby causing vasoconstriction of the efferent arteriole, resulting in an increase in intraglomerular pressure to maintain filtration. If there is vascular stenosis, whether of the main renal artery or of intrarenal arteries, the introduction of an ACE inhibitor or angiotensin II-receptor blocker may result in a significant reduction in glomerular filtration as a consequence of the abolition of this constrictive effect of angiotensin II. A number of patients, particularly the elderly and those with renovascular disease, are particularly susceptible to this adverse effect of these medications. When the renal circulation is compromised glomerular blood flow is also supported by the action of vasodilator prostaglandins, and in these circumstances NSAIDs can cause substantial decrement in the filtration rate. Much more rarely, the blood vessels may exhibit vasculitis following treatment with hydralazine or propylthiouracil.

Glomerular changes may be induced by gold or D-penicillamine, resulting in proteinuria and, in certain cases, the nephrotic syndrome. The glomerular changes usually revert once the drug is withdrawn.

Tubular function can also be adversely influenced. Lithium is associated with a nephrogenic diabetes insipidus-like syndrome due to inhibition of the action of ADH on the cells of the distal tubule and collecting duct. This may be accompanied by an incomplete distal renal tubular acidosis and, in some patients, with a chronic interstitial nephritis. Acute renal failure due to an acute reversible reduction in glomerular filtration occurs in lithium toxicity.

Interstitial nephritis may occur as an acute allergic reaction occurring shortly after the introduction of a drug, or in a more chronic form after several months of ingestion. In the acute form there may be other manifestations of an allergic reaction such as rash, eosinophilia, and the detection of eosinophils in the urine. In the more chronic form there may be no such indication of an allergic reaction. A wide variety of medications, including antibiotics and analgesics, may give rise to interstitial disease, as can certain herbal remedies, as evidenced by the nephropathy that has been associated with Chinese herbs used as a slimming aid. Analgesic abuse may give rise to papillary necrosis, which may present as polyuria due to impairment of renal concentration or from obstruction if a sloughed papilla occludes the ureter.

Some medications have the ability to elevate blood pressure, and so a drug-induced cause should be considered in any patient presenting with hypertension: this is well recognized in patients receiving oestrogen/progesterone preparations, corticosteroids, ciclosporin, and erythropoietin.

In patients with renal disease there may be a greater propensity for adverse reactions. In the nephrotic syndrome the diminished plasma albumin concentration will result in a lessening of protein binding and thus an increased availability of the 'free' drug. If there is impaired renal function there may be diminished excretion, thus prolonging the half-life of the drug and increasing the risk of toxicity if the dose is not adjusted appropriately. In addition, some treatments given for some renal diseases may have adverse interactions with other drugs, for example high-dose corticosteroids have the ability to displace protein-bound drugs, thereby increasing the concentration of 'free' drug.

Dietary history

A carefully obtained dietary history can, in a few cases, be helpful in reaching a diagnosis. Excessive sodium intake may be associated with hypertension or apparent resistance to antihypertensive medication. Idiopathic, stone-forming patients seem to have a greater than average protein intake, resulting in increased excretion of calcium, oxalate, and uric acid, all of which are risk factors for stone formation. Some patients have a preference for acid-tasting foods that they may consume in excessive amounts: fruit juices and rhubarb, which are high in oxalate, may be the cause of oxalate deposition in the kidney if taken to excess. Other patients consume an excessive amount of liquorice that interferes with the inactivation of cortisol by blocking the enzyme 11b-hydroxysteroid dehydrogenase, and, as a consequence, the cortisol binds to mineralocorticoid receptors mimicking hyperaldosteronism.

There are differences in the composition of some ethnic diets compared to a 'standard' Western diet: Japanese food is high in sodium, whereas Indian food is high in potassium. In addition, some foods may aggravate the effects of renal failure. Unleavened bread, chapattis, with a high phytate content bind intestinal calcium reducing absorption and thereby stimulating parathyroid hormone secretion which, in addition to other effects, increases the metabolism of vitamin D and as a consequence aggravates the osteomalacic or rachitic component of renal osteodystrophy.

Family history

It is important to obtain as full a family history as possible as this can significantly aid diagnosis. If an inherited renal disease is identified, the diagnosis is of value to other members of the family, allowing the identification of affected members and appropriate clinical review to control any associated complications such as hypertension and anaemia. In any patient with suspected renal disease it is advisable to enquire about familial renal disease, deafness (Alport's syndrome), and whether the parents were related (all recessive conditions are much more common with inbreeding). It may be necessary to examine medical notes and obtain death

certificates. Sometimes it is possible to infer inherited renal conditions that were unsuspected: a patient with polycystic renal disease may give a history that some relatives have died suddenly from a 'stroke', indicating the possibility of subarachnoid haemorrhage from a ruptured cerebral aneurysm. Similarly, most, if not all, cases of vesicoureteric reflux are familial, when a history of troublesome urinary infections in family members can be informative.

Not all familial conditions seem to have a clear inherited pattern. There is a familial predisposition to systemic lupus erythematosus, other autoimmune diseases, and IgA nephropathy. In some families there is a greater than expected incidence of hypertension or diabetes mellitus, indicating genetic influences as yet undetermined. In some conditions there is considerable genetic heterogeneity: Alport's syndrome is classically an X-linked dominant condition, but an autosomal recessive form has been described, as has a form associated with macrothrombocytopenia (Epstein's syndrome). Furthermore, not all diseases that have similar features are the same condition, for example nerve deafness and urinary abnormalities may be due to Alport's syndrome or to the Muckle–Wells syndrome (heredofamilial amyloidosis), Refsum's syndrome, a rare form of Charcot–Marie–Tooth syndrome, or Cockayne syndrome.

Social history

Socioeconomic factors are important in patients with renal disease. Increasing affluence is associated with an increasing incidence of renal stone disease, possibly related to an increase in protein intake. Low socioeconomic status is associated with an increased incidence of bacteriuria during pregnancy.

Smoking is a risk factor for the development of atherosclerosis, renovascular hypertension, and accelerated hypertension. It is also associated with anti-glomerular basement membrane (**anti-GBM**) nephritis and the risk of developing nephropathy in patients with diabetes mellitus.

Intravenous drug abuse is a risk factor for septicaemia that may lead to bacterial endocarditis with associated glomerulonephritis. In addition, such patients may develop acute renal failure due to septicaemia, rhabdomyolysis or (rarely) vasculitis, or chronic renal failure due to amyloidosis. They are also at risk from needle-transmitted infections such as hepatitis B and C and human immunodeficiency virus (**HIV**) that may lead to vascular and glomerular disease. Many patients will not admit to drug abuse and so obtaining an accurate history may be difficult: on clinical examination particular attention needs to be taken of any skin indication of intravenous needling.

Occupational history

Some renal diseases may be work related, hence specific enquiry should be made with respect to the working environment and, in particular, to exposure to any chemical substances: for example, the use of chemicals, pesticides, exposure to fumes, and the need to wear protective clothing. A full occupational history, proceeding in chronological order from the time of leaving school to the present day, needs to be taken. Aniline dye workers have an increased risk of developing urothelial tumours; exposure to solvents may be causally associated with anti-GBM glomerulonephritis; vaporized lead fumes, as occur in the welding of lead pipes, may cause lead nephropathy and chronic renal failure, which is being associated with exposure to an increasing number of toxic substances. As information becomes available, appropriate health and safety regulations are introduced to lessen the risk to workers, but some workers ignore advice, and in some countries—particularly in the developing world—preventive measures are not enforced or are deliberately ignored.

Acute renal failure may arise from leptospirosis in miners, sewage workers, farm workers, and those recreationally exposed, or from hantavirus infection in laboratory technicians and farmers in endemic areas.

Ethnic and geographical factors

Geographical factors are important in a number of renal diseases. Hantaviruses infect various animal species worldwide, but the incidence of clinical infection is variable. In Asia it is due mainly to the Haantan and Seoul viruses, in China it is known as epidemic haemorrhagic fever, and in Korea as Korean haemorrhagic fever. In Europe infections are mainly caused by the Puumala serotype and result in acute renal failure, whereas in North America, although two serotypes have been identified, no cases of renal disease have been reported. Other infections, such as malaria and schistosomiasis, also have particular geographical prevalence. It is important to obtain a full history of travel abroad, as this will alert the clinician to the potential of diseases not frequently seen in everyday practice but which may have an infective basis. (See Chapter 20.7.10 for further discussion of aspects of renal disease particular to developing countries.)

Patients of African or Asian origin are likely to have been exposed to tuberculosis at some time, which may give rise to clinical disease if they are immunosuppressed for treatment of glomerular disease or following transplantation. It is common practice to treat with prophylactic therapy in such circumstances, generally using isoniazid (with pyridoxine).

Some renal diseases have an ethnic association. The amyloidosis that complicates familial Mediterranean fever occurs in Arabs from the Mediterranean area, Turks and Sephardic Jews, whereas Sephardic Jews from Baghdad, southern Russia, the Balkans, and Ashkenazic Jews are rarely affected. IgA nephropathy is more common in Caucasians and people from some Asian countries (Japan, Singapore, and China) and is apparently less common in Afro-American and African peoples. Systemic lupus erythematosus is more common in those from the Middle East and the Orient than from Europe. Diabetes mellitus has an increased prevalence in Asian communities and in some of the Indian communities of North America.

Tubulointerstitial diseases such as Balkan nephropathy have a particular geographical distribution. Analgesic nephropathy, associated with the regular consumption of compound analgesics, particularly but not exclusively those containing phenacetin (before the use of this drug was restricted or banned), was particularly common in Australia where renal effects may be aggravated by minor dehydration due to increased insensible fluid loss. Less easy to explain is the increased prevalence in Switzerland and certain towns in Belgium.

Further reading

Birch D, *et al*. (1983). Urinary erythrocyte morphology in the diagnosis of glomerular hematuria. *Clinical Nephrology* **20**, 78–84.

Britton JP, *et al*. (1992). A community study of bladder cancer screening by the detection of occult urinary bleeding. *Journal of Urology* **148**, 289–92.

Burden RP, *et al*. (1979). The loin pain/haematuria syndrome. *Lancet* **i**, 897–900.

Castenfors J, Mossfeldt F, Piscator M (1967). Effect of heavy prolonged exercise on renal function and urinary protein excretion. *Acta Physiologica Scandinavica* **70**, 194–206.

Devarajan P (1993). Mechanisms of orthostatic proteinuria: lessons from a transplant donor. *Journal of the American Society of Nephrology* **4**, 36–9.

Dorhout Mees EJ, Geers AB, Koomans HA (1984). Blood volume and sodium retention in the nephrotic syndrome: a controversial pathophysiological concept. *Nephron* **36**, 201–11.

Ginsberg JM, *et al*. (1983). Use of single voided urine samples to estimate quantitative proteinuria. *New England Journal of Medicine* **309**, 1543–6.

Gorensek MJ, Lebel MH, Nelson JD (1988). Peritonitis in children with nephrotic syndrome. *Pediatrics* **8**, 849–56.

Grossfeld GD, *et al.* (2001). Asymptomatic microscopic hematuria in adults: summary of the AUA best practice policy recommendations. *American Family Physician* **63**, 1145–54.

Grossman E, Messerli FH (1995). A side effect of drugs, poisons, and food. *Archives of Internal Medicine* **155**, 450–60.

Hiatt RA, Ordonez JD (1994). Dipstick urinalysis screening, asymptomatic microhematuria, and subsequent urological cancers in a population-based sample. *Cancer Epidemiology, Biomarkers and Prevention* **3**, 439–43.

Khadra MH, *et al.* (2000). A prospective analysis of 1,930 patients with hematuria to evaluate current diagnostic practice. *Journal of Urology* **163**, 524–7.

Little PJ, Sloper JS, deWardener HE (1967). A syndrome of loin pain and haematuria associated with disease of peripheral renal arteries. *Quarterly Journal of Medicine* **36**, 253–9.

MacGregor GA, De Wardener HE (1988). Idiopathic oedema. In: Schrier RW, Gottshalk CW, eds. *Diseases of the kidney*, 4th edn, pp 2743–53. Little, Brown and Co., Boston, MA.

Mallick NP, Short CD (1981). The nephrotic syndrome and ischaemic heart disease. *Nephron* **27**, 54–7.

Messing EM, *et al*. (1992). Home screening for haematuria: results of a multi-clinic study. *Journal of Urology* **148**, 289–92.

Mohr DN, *et al*. (1986). Asymptomatic microscopic haematuria and urologic disease: a population based study. *Journal of the American Medical Association* **256**, 224–9.

Naish PF, Aber GM, Boyd WN (1975). C3 deposition in renal arterioles in the loin pain and haematuria syndrome. *British Medical Journal* **3**, 746.

Nuyts GD, *et al*. (1995). New occupational risk factors for chronic renal failure. *Lancet* **346**, 7–11.

Ordoñez JD, *et al*. (1993). The increased risk of coronary heart disease associated with nephrotic syndrome. *Kidney International* **44**, 638–42.

Polenakovic MH, Stefanovic V (1998). Balkan nephropathy. In: Davison AM, *et al.*, eds. *Oxford textbook of clinical nephrology*, pp 1203–10. Oxford University Press, Oxford.

Rabelink TJ, *et al*. (1994). Thrombosis and hemostasis in renal disease. *Kidney International* **46**, 287–96.

Reuben DB, *et al*. (1982). Transient proteinuria in emergency medical admissions. *New England Journal of Medicine* **306**, 1031–3.

Ritchie CD, Bevan EA, Collier StJ (1986). Importance of occult haematuria found at screening. *British Medical Journal* **292**, 681–3.

Robinson RR (1980). Isolated proteinuria in asymptomatic patients. *Kidney International* **18**, 395–406.

Springberg PD, *et al*. (1982). Fixed and reproducible orthostatic proteinuria: results of a 20 year follow-up study. *Annals of Internal Medicine* **97**, 516–19.

van Ypersele de Strihou C (1998). Hantavirus infection. In: Davison AM, *et al.*, eds. *Oxford textbook of clinical nephrology*, 2nd edn, pp 1688–92. Oxford University Press, Oxford.

Wass VJ, Cameron JS (1981). Cardiovascular disease and the nephrotic syndrome. The other side of the coin. *Nephron* **27**, 58–61.

20.3.2 Clinical investigation of renal disease

A. Davenport

Introduction

The key to making any correct diagnosis depends upon a careful history and thorough examination. In patients with renal failure the history and examination should attempt to differentiate acute from chronic renal disease, single-organ system involvement from multisystem disease, and obstruction from intrinsic or prerenal disease. Renal disease may be associated with preceding infections and the ingestion of drugs or herbal remedies. An accurate history and careful examination will determine the sequence and spectrum of clinical investigations required to make a diagnosis.

Examination of the urine

Urine collection

To minimize contamination, standard investigation is of a midstream urine (**MSU**) sample. Voiding from a full bladder containing at least 200 ml of urine should remove urethral organisms before the MSU is collected. Even so, in women, vaginal leucocytes and bacteria may contaminate the urine, and men should retract the foreskin to minimize contamination. Suprapubic aspiration is the technique of choice in babies and infants, and occasionally in adults who can not co-operate to provide an MSU. The second urine of the morning is the best for microscopy as it is still acidic and concentrated, but without the overnight stay in the bladder that results in the degeneration of casts and cells. Cell lysis can occur in both hypotonic and alkaline urine. Only the first 10 ml of the stream are collected in cases of suspected urethritis.

Macroscopic appearance

Fresh urine usually has a yellow colour due to the presence of urochromes. Occasionally urine will have a milky appearance due to pus, spermatozoa, insoluble phosphates in alkaline urine (sometimes seen following heavy meals), or occasionally in cases of chyluria, or urate crystals in acid urine. Foamy or frothy urine is typical of heavy proteinuria.

Certain agents and conditions can discolour urine:

•*Pink to red coloration*—haematuria may result in a range of colours from smoky pink through to port-wine red in cases of frank macroscopic haematuria. Other causes of a pink or red urine include eating sweets containing aniline dyes, beetroot or other foodstuffs containing anthocyanins, haemoglobin, myoglobin, some drugs such as phenindione and phenolphthalein, and (if the urine is left to stand) porphyrins in cases of acute intermittent porphyria.

•*Blue or green coloration*—can be caused by pseudomonas urinary sepsis, methylene blue, biliverdin, triamterene, amitriptyline, chlorophyll-containing breath mints (Clorets®), excessive use of mouthwash and deodorants, magnesium salicylate (Doan's pills®), phenyl salicylate, guanicol (in cough remedies), thymol (in volatile oils and horesemint), iodochlorhydroxyquin, tolonium, Evans blue, methocarbamol, Diagnex blue, indigo blue, resorcinol, azuresin, bromoforium, and occasionally propofol. Phenol and lysol can result in a green or black discoloration.

•*Orange coloration*—can be caused by anthraquinone-containing laxatives, rifampicin, and excess urobilinogen.

•*Yellow urine*—may be found in patients prescribed mepacrine, phenacetin, and those taking excessive amounts of riboflavin, as well as icteric patients with conjugated hyperbilirubinaemia.

•*Black or brown urine*—alkaptonuria results in black or brown urine, whereas myoglobin and melanin only lead to black urine on standing. Other causes of a brown urine include bilirubin, L-dopa, niridazole, furazolidone, and phenazopyridine, and, following standing, haemoglobin and myoglobin. As mentioned above, phenol and lysol can result in a black or green discoloration.

Stick testing

The upper limit of normal for protein excretion in the urine is 128 mg/24 h. Although albumin is the largest single component, more than half of the protein content comprises low molecular weight proteins and protein fragments. Commercial sticks such as Albustix™ are very sensitive, detecting protein in urine starting at concentrations around 100 mg/l. Since these sticks detect protein on a concentration basis, using bromocresol green as an indicator dye, the results they give are affected by urine flow rate and urine dilution or concentration. The sticks are treated with a buffer to keep their pH constant. An elevated urinary protein concentration can erroneously be recorded if the buffer is washed off by leaving the stick in the urine for too long, and with very alkaline urine. Some antiseptics used to clean the skin, including cetrimide and chlorhexidine, may also react and cause a false-positive result.

pH

Normal urine is slightly acidic, but can vary between pH 4.5 and 8.0. If an early morning urine specimen is under pH 5.3, then there is unlikely to be a significant defect in urinary acidification. Alkaline pH is often found in urine infected with urea-splitting bacteria. In some cases of renal calculus disease, particularly in cystinuria and urate nephropathy, crystal solubility is greater in alkaline urine, and patients should regularly check their urine pH. Haemoglobin and myoglobin are also more

soluble in alkaline urine. Thus maintaining a forced alkaline diuresis is important in the management of patients following tumour lysis and those with rhabdomyolysis or haemoglobinuria.

Glycosuria

The stick reaction is based on glucose oxidase, which releases hydrogen peroxide from glucose, so producing a graded colour change by oxidizing an indicator. This reaction is specific for glucose, and does not detect other sugars. The reaction can be blocked by large doses of ascorbic acid. A positive stick test for glucose must be interpreted in light of the plasma glucose level, as glycosuria may reflect a defect in renal tubular glucose absorption.

Specific gravity

Specific gravity is a measure of the number of particles dissolved in a litre, whereas osmolality is the number of particles per kilogram. Protein and glucose increase the specific gravity more than the osmolality as they are dense particles. In normal patients the early morning, or concentrated, urine sample should have a specific gravity of 1.024 or more.

Nitrite stick test

Nitrite sticks contain an aromatic amine which reacts with nitrites, produced by bacterial reduction of nitrate, to form a pink-coloured diazonium complex. More than 90 per cent of the common urinary pathogens are nitrite-forming bacteria. However, *Pseudomonas* spp., *Staphylococcus albus*, *Staphylococcus saprophyticus*, and *Streptococcus faecalis* may have minimal or no nitrite producing capacity. Other false-negative results may be obtained in alkaline urine, in patients taking large doses of vitamin C, and with frequent voiding of dilute urine when the urinary nitrite concentration is too low.

Leucocyte esterase stick test

This stick test is based on the presence of a leucocyte esterase, and is very specific for the presence of urinary leucocytes, both intact and lysed. This test may be more accurate than microscopy when the urine is alkaline or hypotonic. However, the test can be inhibited by high concentrations of glucose (³30 g/l), ketones, and antibiotics including cefalexin, cephalothin, nitrofurantoin, tetracycline, and tobramycin. The sensitivity of this test is also reduced when the specific gravity of the urine is high, for instance in the presence of a heavy proteinuria.

Urine microscopy

To obtain reproducible results urine should be processed in a standard manner and examined under the microscope as soon as possible. In our own institution a few drops of acetic acid (10 per cent v/v) is added to ensure a pH of 6.0 or less; then 10 ml of urine is centrifuged for 5 min at 1500 r.p.m. (750 *g*); following which, 9.5 ml of supernatant is removed and the deposit resuspended. One drop (50 µl) is placed on a microscope slide and covered with a standard coverslip (24 × 32 mm). Although phase-contrast microscopy is an advantage in identifying red cells and casts, a standard microscope will suffice. A semiquantitative assessment of casts is made at low power (160 ×) and other elements at high power (400 ×), expressing the counts as numbers per field. Normal urine contains 1 or 2 leucocytes per high-power field (**HPF**), 1 erythrocyte per 2 or 3 HPF, 1 tubular cell per 10 HPF, and both hyaline casts (1 per low-power field, **LPF**) and granular casts (1 per LPF). Physical exercise can result in haematuria and cylinduria for several hours. Stains such as modified Sternheimer's stain (Sedi-stain™) can be used to help differentiate renal tubular cells from leucocytes. To improve the detection of casts, urine can be filtered through a 5-µm Millipore™ filter, and the retained casts stained with Papanicolaou's stain.

Cellular elements

The morphology of the erythrocytes in the urine can give valuable information as to the source of bleeding. Erythrocytes which have passed through the glomerulus and then along the renal tubule can become distorted or dysmorphic. Those originating from other sources within the urinary tract, such as the bladder, typically show much less signs of damage so that they more closely resemble erythrocytes in the peripheral blood, these are termed isomorphic. To establish a diagnosis of glomerular haematuria there should be a minimum of three different forms of dysmorphic erythrocytes present. One particular type of dysmorphic erythrocyte, the acanthocyte, is reported to have 52 per cent specificity and 98 per cent sensitivity for glomerular haematuria when the acanthocyte count is 5 per cent or more. However, not all workers have found erythrocyte morphology to be useful in discriminating glomerular from non-glomerular bleeding, and the physician who only occasionally examines urine under the microscope is unlikely to obtain clear, reproducible, and useful discrimination between dysmorphic and isomorphic cells.

Some centres use automated haematological cell counters (Coulter Counter™) to assess red cell morphology in both urine and peripheral blood. The red cell size-distribution pattern for lower urinary tract haematuria is similar to that of the peripheral blood, with a relatively narrow size range and a high-frequency distribution curve. Whereas the typical pattern for dysmorphic haematuria is one of a broader range of red cell sizes, with a lower frequency distribution. To have any reliability, urine samples must be processed rapidly by those who do it regularly.

Microscopy may also reveal renal tubular epithelial cells. These cells are shed into the urine in acute tubular necrosis; in response to certain drugs, both nephrotoxic and ischaemic; and also in acute renal allograft rejection. In patients with nephrotic syndrome, these cells are seen as oval fat bodies, laden with lipid droplets. Squamous epithelial cells from the urethra and vagina and transitional cells from the ureter and bladder may also be present in normal urine.

During infection, the urine may contain large numbers of leucocytes and bacteria. When large numbers of leucocytes are present in the absence of bacteria, so called sterile pyuria, then a variety of conditions should be considered: renal calculus disease, analgesic nephropathy, interstitial nephropathy, proliferative glomerulonephritis (rarely), renal tuberculosis, schistosomiasis, and partially treated bacterial urinary tract infection. Phase-contrast microscopy can distinguish lymphocytes from neutrophils, but eosinophils can only be identified with specific stains (Hansel's stain). Classically, urinary eosinophilia occurs in cases of acute interstitial nephritis, typically due to drugs, and also in cholesterol atheroembolic disease.

Urinary casts

Casts form from the transformation of Tamm–Horsfall glycoprotein, secreted by the distal tubular cells, into a gel matrix. They typically assume a tubular structure. Hyaline casts only contain Tamm–Horsfall glycoprotein, and are found in a variable amount in the urine of normal subjects (Fig. 1 and Plate 1). Fever, cardiac failure, strenuous exercise, and some drugs, such as furosemide and ethacrynic acid, increase hyaline cast excretion. During passage through the distal tubule and collecting duct a variety of proteins, pigments, and cells adhere to the Tamm–Horsfall protein, producing a wide variety of casts. Granular casts have deposits of either fine or coarse protein granules (Fig. 2 and Plate 2). Although they may occur in normal subjects, or after exercise, they are typically found in cases of parenchymal renal disease. In patients with proteinuria, the protein deposited comes from the glomerulus, whereas in acute tubular necrosis the protein comes from degenerate tubular cells. Broad waxy casts are much larger than normal casts and have clear-cut edges: they are formed in dilated hypertrophied tubules, as found in patients with chronic renal failure. Casts containing erythrocytes (red cell casts) indicate renal bleeding and are typically found when there is acute glomerular inflammation caused by glomerulonephritis or vasculitis (Fig. 3 and Plate 3). White cell casts (containing leucocytes) can be found in proliferative glomerulonephritis, acute interstitial nephritis, and acute pyelonephritis.

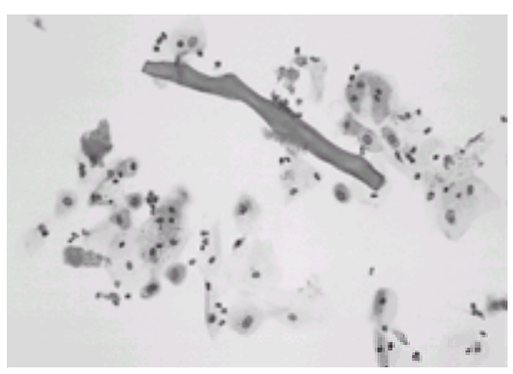

Fig. 1 Papanicolaou-stained urine showing a hyaline cast with both normal transitional and squamous cells and renal tubular cells. (By courtesy of Dr Deery.) (See also Plate 1.)

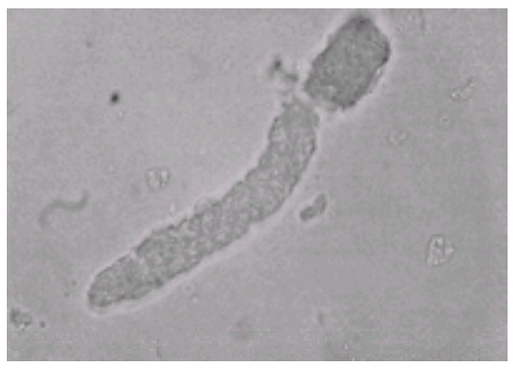

Fig. 2 Unstained urine specimen showing a granular cast. (See also Plate 2.)

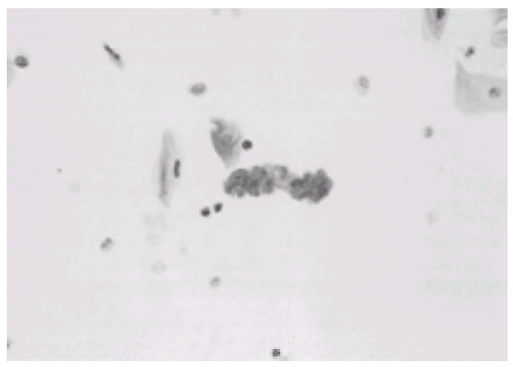

Fig. 3 Papanicolaou-stained urine deposit showing a red cell cast. (See also Plate 3.)

Measurement of proteinuria

Quantification of proteinuria is important as the risk for progression of underlying renal disease to renal failure is related to the amount of protein in the urine. Traditionally, proteinuria has been measured using 24-h urine collections and expressed as grams per day (g/day). This has the advantage that it averages-out protein excretion, and is not therefore affected by the normal diurnal variation in protein excretion (less overnight and first thing in the morning) or urine concentration. Several different methods are used to measure the protein content of 24-h urine collections, ranging from the Biuret method, which uses a copper-based method to precipitate proteins, to dye-binding methods using Coomassie Brilliant Blue as the indicator. These are more accurate than the turbidimetric methods, which use trichloroacetic or sulphosalicylic acid and measure turbidity with a densitometer. Radiocontrast media, and some drugs (including penicillin, sulphonamides, and tolbutamide) may give false-positive results for proteinuria with the sulphosalicylic acid method. The Biuret method measures total proteins, the turbidimetric method provides different readings for albumin and globulins, and the dye-binding methods may do so also.

Testing spot urine samples for protein has been introduced to overcome the inherent problems of patient accuracy and reliability with 24-h urine collections. The urinary albumin concentration is measured by radioimmunoassay. Under resting conditions, urinary creatinine excretion is relatively constant throughout the day. Thus to overcome the problems of timing urinary collections, proteinuria in spot urine samples is expressed as an albumin:creatinine ratio (normal <2.0 mg/mmol creatinine in a daytime urine or 24-h collection, and <1.5 mg/mmol for an overnight or early morning sample). An albumin:creatinine ratio of 100 mg/mmol approximately corresponds to 1.5 g/day, and 350 mg/mmol to nephrotic-range proteinuria. Aside from their use to replace 24-h urine collections, spot urine collections are particularly useful in the diagnosis of orthostatic proteinuria, in other words where the patient has a normal urinary protein excretion when recumbent, or overnight, but has marginally increased proteinuria in the ambulant or daytime sample.

Microalbuminuria

Radioimmunoassays for albumin can detect an increased urinary albumin excretion in patients with normal levels of proteinuria. Normoalbuminuria is defined as an excretion rate of 20 µg/min or less. Proteinuria is usually detectable on dip-stick testing at rates of 200 µg/min or more, and thus microalbuminuria is defined as an excretion rate between 20 and 200 µg/min. The albumin excretion rate (**AER**) is some 25 per cent higher during the day than the night. There is a good correlation between the morning AER and the albumin:creatinine ratio in the first urine sample of the morning. The advantage of spot urines is that all patients can provide a sample when they attend the clinic. Provided the urine samples are taken at the same time, and the patient's dietary intake is relatively constant, then these samples are very useful in assessing patients over time. The advantage of measuring the albumin: creatinine ratio is that it eliminates the timing of urinary samples, which is important in calculating the AER. The albumin:creatinine ratio can also be used to assess the progress of patients with proteinuria, especially if patients fail to collect 24-h urine samples properly.

Microalbuminuria is not only an adverse factor for the progression of diabetic renal disease, but is also predictive of cardiovascular events in both the diabetic and non-diabetic population. In addition to diabetic subjects, microalbuminuria may be found in those with hypertension, cardiac failure, and following a pyrexial or viral illness. Similarly, microalbuminuria may be present in healthy subjects after exercise and during normal pregnancy.

Selectivity of proteinuria

Patients with glomerular disease typically have a non-selective proteinuria, with a similar clearance of both high and low molecular weight plasma proteins. However, those with minimal-change disease may have selective proteinuria, with clearance of predominantly small molecular weight proteins. The demonstration of selective proteinuria is useful in paediatric practice, where patients are often treated with steroids without a renal biopsy.

Most laboratories compare the clearance of IgG, as the large molecular weight protein (mol. wt 150 kDa), to that of albumin (or transferrin, mol. wt 88 kDa) as the low molecular weight protein. Both plasma and spot urine samples are required. Protein concentrations are measured either by laser nephelometry or radial immunodiffusion. Non-selective proteinuria is taken as a [IgG]U/[IgG]P × [transferrin]P/[transferrin]U ratio of 0.20 or more, whereas selective proteinuria is taken as a ratio of 0.10 or less (U is the protein concentration in urine, P the protein concentration in plasma).

Spill-over proteinuria

Patients with myeloma, some types of amyloidosis, and those with reticuloendothelial disorders may have a spill-over proteinuria, due to glomerular filtration of complete and incomplete kappa (_) and lambda (l) chains and immunoglobulin light chains. These small molecular weight proteins are not detected by simple urine stick-testing, or by standard biochemical methods to determine urine protein concentration. Thus, when clinically appropriate, urine should specifically be sent for immunoelectrophoresis to exclude myeloma. However, light chains in particular may still not be detected, hence further investigation with specific antisera may be

required if their presence is suspected.

Renal tubular proteinuria

Interstitial renal disease can result in proteinuria, usually less than 2 g/day. Proximal tubular injury leads to increased low molecular weight proteinuria, characterized by an excess of intestinal alkaline phos-phatase, *n*-acetylglucosaminidase, retinol binding protein, tissue-specific alkaline phosphatase, a-glutathione *S*-transferase, a_1-macroglobulin, and b_2-microglobulin. By contrast, Tamm–Horsfall glycoprotein and a-gluta-thione *S*-transferase are increased in distal tubular injury.

b_2-Microglobulin is freely filtered at the glomerulus and then reabsorbed in the proximal tubule, such that less than 1 per cent of the filtered load is excreted in the urine of normal subjects (normal <370 µg/24 h). Thus urinary b_2-microglobulin excretion has been used as a marker of proximal tubular damage. However, b_2-microglobulin is unstable in urine, and its excretion can be affected both by an increased production rate (found in cases of myeloproliferative disease, chronic inflammatory states, and acute liver disease) and by saturation of b_2-microglobulin tubular uptake due to an excess of dibasic amino acids.

More reliable markers of tubular proteinuria are now available. These include a-glutathione *S*-transferase, a_1-macroglobulin, and retinol binding protein. Turbimetric or enzyme assays are now available. Results are expressed as either excretion rates (for example, normal a-glutathione *S*-transferase, <12.5 ng/min or <11.5 µg/l) or as a ratio to urinary creatinine (for example, normal reference range for retinol binding protein:creatinine, <0.019 mg/mmol).

These tests of renal tubular proteinuria are helpful in investigating patients with suspected Chinese herbal nephropathy, Asian subcontinent nephropathy, and Balkan nephropathy. Industrial workers exposed to heavy metals and organic chemicals, such as those used in the dry-cleaning industry, may develop interstitial renal disease characterized by increased urinary low molecular weight proteinuria.

Estimation of renal function

Biochemical tests

Measurement of plasma creatinine is the standard biochemical test used to assess renal function. Unfortunately the plasma creatinine concentration is not linearly related to the glomerular filtration rate. Thus some 30 per cent of patients with significantly impaired renal function still have a plasma creatinine value within the normal range (<120 µmol/l).

Creatinine

Creatine, which is endogenously synthesized in the liver or exogenously supplied by meat in the diet, is transported to muscle and converted to creatinine by non-enzymatic dehydration. Muscle mass represents some 98 per cent of the total body creatine pool. Thus gender, racial and age-related differences in body composition, physical training and exercise, muscle-wasting diseases, paralysis, and intercurrent illnesses will all affect the production rate of creatinine, and therefore both the plasma creatinine concentration and urinary creatinine excretion. Hence, in young children there is a steady increase in the plasma creatinine level as their muscle mass increases. Dietary influences will affect plasma creatinine levels, with a reduction in strict vegans and increased values in those with a high meat intake (particularly stewed meat: cooking leads to the conversion of creatine to creatinine) or those taking creatine supplements. For any individual, the plasma creatinine level is relatively constant throughout the day, although there is a tendency for it to increase slightly in the afternoon.

Creatinine is not only freely filtered by the glomerulus, but is also secreted into the renal tubule. Creatinine reabsorption may occur at low urinary flow rates, such as in congestive cardiac failure. The relative proportion of renal tubular creatinine secretion to that filtered increases as renal function declines. In addition, in oedematous states such as nephrotic syndrome, calculated creatinine clearance exceeds inulin clearance, suggesting increased tubular creatinine secretion. Several drugs are known to block the tubular secretion of creatinine, and thus cause an increase in the serum creatinine level: these include the diuretics amiloride, spironolactone, and triamterene; and also cimetidine, aspirin, probenecid, and trimethoprim.

Most laboratories measure plasma creatinine using standard automated analysers, which assess the chromagenic product of creatinine and alkaline picrate (Jaffé reaction). Table 1 lists some substances which in high concentration can act directly or indirectly as chromogens, or affect the background control blanks, and so result in a spurious increase in the plasma creatinine level. In clinical practice these may lead to an overestimation of creatinine in poorly controlled diabetics, and an underestimation in deeply jaundiced patients, such as those with primary biliary cirrhosis. Under these circumstances a more accurate method is to determine the plasma creatinine level enzymatically.

Reciprocal creatinine or logarithm of creatinine values

As the plasma creatinine level roughly doubles for every 50 per cent reduction in glomerular filtration rate (**GFR**), expressing (transforming) the results as the reciprocal or logarithm is useful in assessing serial plasma values—this changes the graph from an exponential to a straight-line plot. The advantage of using a straight-line plot of plasma creatinine is that it allows the rate of renal decline to be calculated, which can then be used to predict the onset of endstage renal failure and the requirement for dialysis treatment in many patients. The reciprocal creatinine plot assumes a constant rate of loss, whereas the logarithm a constant fractional loss of renal function.

Patients with diabetic nephropathy tend to have a faster rate of decline in renal function than those with glomerular disease, who, in turn, have a faster rate than those with tubulointerstitial renal disease. In addition, it is easier to assess the effect of treatment interventions on the progression of renal disease by analysing transformed data, and also to recognize when there has been a sudden and unexpected deterioration in function that requires urgent investigation.

Prediction of creatinine clearance from the plasma creatinine level

Despite the potential inaccuracies in the determination of plasma creatinine, variations in endogenous creatinine production rates, and the relative increase in renal tubular and intestinal creatinine secretion with deteriorating renal function, formulas based on the plasma creatinine level are used in clinical practice to estimate creatinine clearance. The most common equation, validated in adults, is the formula of Cockcroft and Gault, later modified by Gault:

$$\text{GFR ml/min} = 1.2 \times [140 - \text{age (years)}\ ] \times \text{weight (kg)}]/[\text{plasma creatinine concentration}]\ (\mu\text{mol/l}).$$

In the original formula, there was a different equation for women, with a factor of 0.85 (instead of 1.2) to allow for the lower rate of creatinine production in women due to differences in their body composition. Although these formulas may be helpful in clinical practice to provide an estimation of renal function, they are not always accurate, particularly in diabetic subjects and Afro-Americans (due to differences in body composition).

Creatinine clearance

In clinical practice, creatinine clearance remains the most commonly used parameter for assessing the GFR. However, this depends upon patient compliance to provide an accurate 24-h urine collection. Even when patients are in a steady state, urinary creatinine excretion varies from day to day, and reliability can be increased by performing consecutive daily clearances.

Creatinine clearance is calculated thus:

$$\text{creatinine clearance (ml/min)} = [\text{urine volume (ml/24 h)} \times \text{urine creatinine concentration (µmol/l)}/\text{plasma creatinine concentration (µmol/l)}\ ] \times 24 \times 60.$$

As regards the use of the creatinine clearance measurement as an estimate of GFR, two errors tend to balance each other out. The chromagenic assay tends to overestimate the plasma, but not urinary, creatinine concentration, leading to an underestimation of GFR. By contrast, creatinine is not only excreted by glomerular filtration: some is secreted by the renal tubules, leading to an overestimation of the GFR. However, in patients with impaired renal function these contrasting effects are not balanced, and the relative increase in tubular creatinine secretion results in creatinine clearance exceeding GFR. This problem can be overcome by the

administration of 400 mg of cimetidine to block renal tubular creatinine secretion, but this manoeuvre is rarely (if ever) performed in clinical practice solely for this purpose. By convention, creatinine clearance values are commonly corrected for body surface area to adjust for differences in muscle mass, assuming a fixed mathematical relationship between body surface area and the relative proportions of fat to muscle. However, body composition is not only age- and gender-dependent, but also varies from race to race, and other inaccuracies occur in oedematous states.

Cystatin C

Cystatin C is a low molecular weight basic protein (13.26 kDa) produced by all nucleated cells. The cystatin gene is a housekeeping gene and a member of the cystatin superfamily of cysteine proteinase inhibitors. Cystatin C is produced at a constant rate, and is not affected by acute inflammation, nutrition, gender, race, or changes in body mass. The production rate is stable over a wide age range, from infants older than 1 year through to the elderly, although there is a slight increase in the latter age group. As cystatin C is freely filtered by the glomerulus and is unaffected by renal tubular degradation or tubular secretion, it can be used as a marker of GFR. Rapid and fully automated accurate assays are now available, which have a superior analytical specificity and precision to that of serum creatinine.

Carbamylation

Urea accumulates with deteriorating renal function. In plasma, urea can spontaneously dissociate to form a reactive cyanate species which can react with the terminal valine of haemoglobin a and b chains (and also similar valine molecules in other proteins). This reaction is termed 'carbamylation' and the product 'carbamylated haemoglobin' (or other protein). Whereas glycosylated haemoglobin has proved useful in clinical practice for assessing time-averaged diabetic control, carbamylated haemoglobin or carbamyl-lysine adducts have not been shown to be superior to simple serum creatinine measurements in determining stable renal function. However, they are useful in helping to differentiate acute from chronic renal failure, because of the time course of the carbamylation reaction, and also in the assessment of time-averaged urea levels in the dialysis patient with endstage renal failure. However, until the relevant assays are commercially available, their use will remain experimental.

Isotopic methods

The glomerular filtration rate can be determined by the clearance of a compound which is freely filtered by the glomerulus and then passes through the nephron without tubular reabsorption or secretion. Traditionally, inulin—a naturally occurring polyfructose—was given as a constant infusion to achieve a constant plasma concentration, and then clearance determined from timed urinary collections. This was a considerable laborious technique. Furthermore, the biochemical estimation of inulin was initially tedious and difficult, with significant interassay variation, and accurate timed urine collections are unreliable in patients with urinary tract anomalies. To overcome these and other difficulties, compounds other than inulin are generally used to estimate GFR, and methods other than constant infusion.

Following a single bolus injection, depending on the compound used, the fall in plasma concentration follows either a single- or two-compartment model related to renal clearance. Chromium-labelled ethylenediaminetetraacetic acid ([^{51}Cr]EDTA) is the most commonly used isotope. After the single injection, three timed plasma samples are taken to calculate the plasma decay rate, and thereby the GFR. More recently it has been showed that for a GFR over 30 ml/min, only a single blood sample at 4 h is required. At GFRs above 30 ml/min there is a very good correlation between inulin and [^{51}Cr]EDTA clearance, but below 30 ml/min the accuracy of the isotope techniques is reduced, there being some renal tubular reabsorption. Accuracy can be improved in this situation by taking a delayed (24 h) plasma sample.

Other isotopes that have been used to estimate GFR include [^{125}I]iothalamate, which when given as a subcutaneous injection results in a constant plasma concentration equivalent to the infusion technique, and Tc ^{99m}-diethylenetriaminepentaacetic acid (**DTPA**), which is less accurate due to its short half-life (6 h) and dissociation of DTPA from the radionuclide.

With all the isotopic methods, it is conventional for the GFR to be corrected for the size of the patient. This correction assumes a fixed relationship between the weight and height of an individual: hence serial estimations to detect a change in renal function are more likely to be accurate than single estimations.

Radiological methods

Iohexol is a non-ionic, low-osmolality, radiocontrast dye. It can be used to estimate glomerular filtration rate following a single bolus injection of between 2 and 5 ml. In patients with a clearance of over 30 ml/min, a single plasma sample taken 3 h after injection provides an accurate estimation, whereas additional later samples are required to improve the accuracy in those with severely impaired renal function.

Summary

When the plasma creatinine concentration is below 150 µmol/l, it cannot be used as an accurate assessment of renal function. When appropriate, an isotopic assessment of GFR is the most accurate method of determining GFR. Otherwise, two 24-h urine collections with corresponding plasma samples should be used to calculate the GFR, although in some centres cystatin C has replaced creatinine for the assessment of renal function. To examine changes in renal function, creatinine concentrations should be transformed to either the reciprocal or the logarithm to assess trends in serial results.

Renal blood flow

Renal blood flow can be estimated non-invasively using Doppler flow probes, provided there is a single renal artery and adequate imaging is possible. This is technically easier for the transplanted kidney than the native kidney. The recent development of contrast agents for ultrasound may increase the reliability of these estimations. Alternatively, renal blood flow can be estimated from the measurement of the renal plasma flow and the haematocrit. However, the haematocrit of peripheral venous blood may not be the same as that entering the renal artery.

Renal plasma flow

Ideally any compound used to assess renal plasma flow should have 100 per cent uptake by the kidney. Thus that fraction not filtered by the glomerulus must be extracted by the tubules and secreted. *p*-Aminohippurate is the most commonly used compound, but is only 85 per cent extracted during a single passage through the kidney, and thus at best only provides an estimate of renal plasma flow. Continuous infusion of *p*-aminohippurate provides a more accurate estimation of renal plasma flow than single injection techniques.

Renal blood flow varies in normal subjects with pain, stress, physical exercise, normal pregnancy, and following a high protein meal. In patients with impaired renal function, the decline in renal plasma flow generally corresponds to the decrease in GFR. However, in some conditions where there may be renal tubular hypoxia or toxicity, such as in patients with severe heart disease or those with ciclosporin nephrotoxicity, the reduction in estimated renal plasma flow is greater than that expected for the change in GFR, due to a reduction in the renal tubular uptake of *p*-aminohippurate. Similarly, *p*-aminohippurate uptake is reduced in small children. [^{125}I]*o*-Iodohippurate has also been used to estimate renal plasma flow, but this has a lower extraction than *p*-aminohippurate (75 per cent), and is less reliable.

Investigation of tubular function

In a normal subject, some 180 litres of glomerular filtrate is produced each day and less 3 per cent of this is excreted, due to reabsorption by the tubules. The proximal and distal tubules have different functions, and traditionally each is considered separately.

Proximal tubular function

Defects in proximal tubular function may be isolated or generalized, as in the Fanconi syndrome. Glucose, amino acids, phosphate, and organic ions are reabsorbed by the apical border of proximal renal tubular cells by sodium-dependent cotransporters, and then cross out from the basolateral membrane by different, sodium-independent, cotransporters.

Glucose

There is a maximum reabsorption rate for glucose (T_MG) in the proximal tubule of 15.1 ± 2.5 mmol/l (T_MG/GFR), above which glycosuria will be present. To determine TMG/GFR, a 20 per cent glucose infusion is administered at increasing rates to produce a slow rise in the plasma glucose up to a maximum of 30 mmol/l, which is maintained for a minimum of 1 h. Plasma and urine samples are collected every 30 min. Renal function is determined by [^{51}Cr]EDTA-GFR. The glucose absorption rate is calculated as the difference between the filtered load in urine (urine volume × $[glucose]_{urine}$) and the filtered load in plasma (GFR × $[glucose]_{plasma}$). Patients with type A renal glycosuria typically have a reduced threshold of around 5 mmol/l.

Phosphate

Phosphate is normally filtered at the glomerulus and reabsorbed in the proximal tubule, with only 10 to 20 per cent of the filtered load being excreted. The normal tubular reabsorption of phosphate (**TRP**) is above 85 per cent and can be calculated from:

$$\%TRP = \{1 – (\text{phosphate clearance/GFR or creatinine clearance})\} \times 100.$$

If renal function is normal, then this can be simplified by collecting an early morning specimen of urine, and:

$$\%TRP = \{1 – ([phosphate]_{urine} \times [creatinine]_{plasma}/[creatinine]_{urine} \times [phosphate]_{plasma})\} \times 100.$$

Alternatively, the theoretical maximum tubular threshold of phosphate (T_MP) can be estimated from:

$$T_MP/GFR = [phosphate]_{plasma} – ([phosphate]_{urine} \times [creatinine]_{plasma}/[creatinine]_{urine}),$$

or measured directly as for T_MG, following an infusion of phosphate (1.0 litre of 0.1 M sodium phosphate at pH 7.4) with a corresponding [^{51}Cr]EDTA-GFR.

Excessive urine phosphate losses occur in proximal tubular disorders such as the Fanconi syndrome, primary and secondary hyperpara-thyroidism. In the various forms of hypophosphataemic rickets, phosphaturia occurs with a characteristically reduced T_MP/GFR of less than 0.56 mmol/l.

Amino acids

Apart from the reabsoption of histidine (90–95 per cent), that of other amino acids is almost complete (97–99 per cent). Although amidoaciduria can occur as a result of overflow when the plasma concentration exceeds the tubular transport maximum, this is very rarely the cause of aminoaciduria in adults. In general, five types of renal aminoaciduria are distinguished: dibasic amino acids, neutral (monoaminomonocarboxylic acids) amino acids, glycine and imino acids, dicarboxylic amino acids, and generalized amino aciduria in the case of the Fanconi syndrome. Generalized and specific amino acidurias can be detected and quantified by thin-layer chromatography. In the Fanconi syndrome amino acids from all four groups are present, whereas there is only excess glycine in glycinuria. Classic cases of cystinuria have increased urinary arginine, ornithine, lysine, and cystine; and patients with Hartnup disease have an excess of neutral amino acids.

For more detailed discussion of other aspects of proximal tubular function and their diseases, see Chapter 20.8.

Distal tubular function

Patients with primary or secondary nephrogenic and/or cranial diabetes insipidus and those with primary polydipsia may present with polyuria. A water-deprivation test can help to differentiate between these conditions, and should be performed as follows. The patient should be admitted to a metabolic ward on the evening prior to the test, be weighed, and have samples taken for baseline plasma osmolality, chemistries, and arginine vasopressin measurement (**AVP**). An osmolality above 295 mosmol/kg and a sodium concentration above 143 mmol/l, excludes a diagnosis of primary polydipsia. After midnight, no oral fluids are allowed until completion of the test. The early morning urine osmolality is measured, and if it is above 800 mosmol/kg (normal response) the test is abandoned. Thereafter the weight, plasma and urine osmolality, and plasma AVP concentration should be recorded regularly. If weight loss exceeds 5 per cent, then the test should be abandoned to prevent dangerous dehydration. Once urine osmolality reaches a plateau (an hourly increase of less than 30 mosmol/kg for 3 consecutive hours), then 5 units of aqueous vasopressin is administered subcutaneously and urine and plasma osmolality measured after a further 30 min, and then at hourly intervals.

Comparison of the last urine osmolality reading prior to the administration of vasopressin with the maximum osmolality following vasopressin helps to categorize patients. Those with nephrogenic diabetes insipidus will produce a urine osmolality under 300 mosmol/kg with no response to exogenous vasopressin and have high AVP levels. Those with severe cranial diabetes insipidus will have dilute urine, again less than 300 mosmol/kg, but they will respond to exogenous vasopressin by increasing urine osmolality by 50 per cent or more, accompanied by low endogenous AVP levels. Both cranial and nephrogenic diabetes insipidus can occur as partial forms, which show some response to dehydration, but they can be discriminated by analysing the relative changes in endogenous AVP and the urinary and plasma osmolalities. Patients with primary polydipsia do not show pituitary suppression, and have little or no response to exogenous vasopressin.

Renally induced electrolyte imbalances

Sodium and water

Hyponatraemia may occur both in patients with a reduced effective circulating plasma volume and those with the syndrome of inappropriate ADH secretion (**SIADH**). Patients with reduced renal perfusion, such as those with cardiac failure, chronic liver disease, nephrotic syndrome, and prerenal acute renal failure, will have a reduced fractional excretion of sodium (FE_{Na}) of less than 1 per cent (normal 1–2 per cent), where:

$$\%FE_{Na} = ([Na]_{urine}/[Na]_{plasma} \times [Cr]_{plasma}/[Cr]_{urine}) \times 100.$$

Those with SIADH preferentially retain water and have a normal FE_{Na}. However, when interpreting measurements of FE_{Na} it must be remembered that this is increased by diuretic administration and in chronic renal failure.

Both those with a reduced effective circulating plasma volume and those with SIADH have impaired free-water excretion, which can be tested by giving the patient 20 ml of water/kg body weight to drink after voiding. More than 75 per cent of the water load should be excreted within 3 h, and the urine osmolality should fall to under 100 mosmol/kg (specific gravity <1.003). This test can be affected by gastrointestinal disease, smoking, and emotional factors. The free-water clearance (C_{H2O}) can be quantitated from:

$$C_{H2O} = \text{urine volume (ml/min)} – [\text{osmolality}_{urine}/\text{osmolality}_{plasma} \times \text{urine volume (ml/min)}].$$

A positive free-water clearance occurs when the urine is more dilute than plasma, and a negative free-water clearance when the urine is more concentrated.

For further discussion of these issues, and the clinical approach to disorders of sodium and water homeostasis, see Chapter 20.2.1.

Potassium

To determine whether there is a renal tubular cause for potassium disturbances, the transtubular potassium gradient (**TTKG**) can be calculated. This attempts to estimate the potassium concentration in the cortical collecting duct.

Using $TTKG = [potassium]_{urine} \times$ osmolality plasma/urine, a TTKG under 2 suggests a non-renal cause of hypokalaemia, whereas a high TTKG (>10) is associated with mineralocorticoid excess, Liddle's syndrome, or drugs such as acetazolamide, fludrocortisone, and amphotericin. A TTKG above 10 implies a non-renal cause of hyperkalaemia and a low TTKG (<2) would be found in cases of potassium-sparing diuretics, hypoaldosteronism, and pseudohypoaldosteronism. Whilst having

theoretical attraction, it is doubtful whether such analysis helps greatly in the diagnosis or management of patients with hypokalaemia or hyperkalaemia. For further discussion of these issues, and the clinical approach to disorders of potassium homeostasis, see Chapter 20.2.2.

For more detailed discussion of other aspects of distal tubular function and their diseases, see Chapter 20.8 and Chapter 20.13.

Imaging of the patient with renal disease

Plain radiography

Plain abdominal radiographs may demonstrate opaque renal stones, nephrocalcinosis, and the renal outlines. Ultrafast, non-contrast CT scanning with three-dimensional reconstruction has generally replaced nephrotomograms for detecting low-opacity renal stones.

Chest radiography may be helpful in the diagnosis of pulmonary oedema, and also in demonstrating the cardiac silhouette and lung pathology sometimes associated with renal disease, such as pulmonary haemorrhage and cavitation. Multiple rib fractures may suggest multiple myeloma.

Intravenous urography

Although intravenous urography (**IVU**) is no longer the standard investigation in nephrology, it still has an important place in the investigation of patients with suspected obstruction of the urinary tract, as it does provide imaging of the entire urinary tract. As with all radiographic procedures, potential fetal irradiation should be avoided. Bowel preparation is no longer standard, due to the risks of dehydration in the elderly and of gaseous distension of the bowel obscuring the urinary tract. Even the newer non-ionic contrast media can cause nephrotoxicity in some patients, and care should be taken to ensure that those at risk (elderly and those with diabetes, myeloma, or a pre-existing renal impairment) are adequately hydrated. Normal renal length is between 3 and 4 lumbar vertebrae, with a width approximately half that of the length.

The IVU may provide valuable information about renal size and possible intrarenal masses. It remains the best method for investigating the patient with acute renal colic, and for assessing the level of any obstruction. Other techniques, such as ultrasound and ultrafast computed tomography (**CT**) scanning, can also be used to investigate renal colic—the main advantage of CT scanning being that it can detect other pathologies which mimic this condition.

The calyces and papillae are well demonstrated on the IVU, which may be diagnostic in cases of medullary sponge kidney, papillary necrosis, and sloughed papillae. Similarly, intraluminal radiolucent foreign bodies may be demonstrated surrounded by contrast, typified by radiolucent stones, blood clots, fungal ball, tumour, or sloughed papillae.

Abnormalities of the ureteric wall such as localized thickening are found in cases of transitional-cell carcinoma, oedema, tuberculosis, and parasitic granuloma. The IVU may also demonstrate external compression: this can be due to aberrant blood vessels in the upper tract, retroperitoneal fibrosis affecting the middle ureter, or prostatic pathology in the lower tract.

Other conventional uroradiological techniques

Further information about the site and nature of any obstruction can be obtained by ureteropyelography. This may be performed by an antegrade or a retrograde approach. An antegrade study involves percutaneous puncture of the renal pelvis, with immediate relief of the obstruction by nephrostomy, and allows demonstration of the site of obstruction following an injection of contrast media (antegrade ureteropyelography). A retrograde study requires cystoscopy, allowing direct visualization of the distal ureter, the possibility of removing an obstructing stone, and the passage of double JJ stents from below to relieve the obstruction. Injection of contrast media from below demonstrates the site of any obstruction (retrograde ureteropyelography). Antegrade techniques are usually more successful in relieving obstruction, particularly in those with pelvic malignancy or obstruction of a renal transplant. In cases when renal obstruction is considered, but investigation inconclusive, then a pragmatic trial of antegrade stent insertion should be undertaken. Improvement of renal function confirms obstruction.

Retrograde urethrocystography is performed in female patients to detect lower urinary tract abnormalities, such as fistulas or urethral diverticulae. Sequential films taken during micturition may detect active reflux. In males, urethrocystography can be complicated by trauma and infection to the lower urinary tract, and therefore suprapubic bladder puncture is recommended.

Renal ultrasonography

The normal kidney and chronic renal disease

The normal adult kidney is between 10 and 12 cm long, with a thin, bright capsule surrounded by highly reflective perinephric fat. The healthy cortex returns mid-level grey echoes, the pyramids are darker, and the renal sinus, containing fat and the major vascular pedicle, is bright with high reflectivity. Colour, flow Doppler can be used to visualize the flow of urine from the native ureters into the bladder. In most causes of chronic renal disease the kidneys become smaller, with reduced cortical thickness and increased reflectivity. Diastolic blood flow is reduced on the Doppler scan. The renal ultrasound appearances are characteristic in some conditions: these include focal segmental glomerular sclerosis secondary to HIV infection in which the kidney is reported to be large and the cortex uniformly of a high reflectivity, greater than that of the renal sinus. Scars, either vascular or infective, may often be too small to be detected by ultrasound examination, especially in the neonate.

Renal masses

Ultrasound is useful in the assessment of renal masses. Benign cysts have a smooth outline with well-demarcated borders and an echo-free centre, whereas renal tumours are usually irregular with heterogeneous echo reflectivity. Most tumours are vascular, with high flow during both systole and diastole on colour-Doppler scanning, and adenocarcinomas in particular may be seen to extend into the renal vein. Renal transitional-cell carcinomas are not readily detected unless large, as ultrasound does not visualize individual calyces well. Angiolipomas may have a characteristic appearance due to their fat content which has high reflectivity, but confirmatory CT scanning is required.

In adult polycystic kidney disease, the kidneys are typically enlarged with multiple bilateral cysts. Middle-aged women may also have hepatic cysts. It is important to remember that, if patients are scanned in their teenage years or before, then cysts may not have developed or they may be below the level of resolution for ultrasound detection. Haemorrhage, infection, or malignant change all result in complex echoes within cysts, which cannot be differentiated by ultrasound scanning. Autosomal recessive, polycystic renal disease can be detected *in utero* with antenatal scanning.

There is an increased incidence of cystic change in the kidneys of patients with endstage renal disease, and occasionally these cysts may become malignant. It has been recommended that dialysis patients should be screened by ultrasound every 3 years, and then annually if cystic changes develop.

Urinary obstruction

In most centres, ultrasonography of the urinary tract is the first investigation performed when an obstruction is suspected. When urinary obstruction has been present for some time, the high reflectivity of the central renal sinus becomes replaced by echo-free urine, with distention of the calyces. However, it is important to recognize that in acute obstruction, and in cases where the kidney and ureter are encased (usually the result of tumour), the standard ultrasound examination may appear normal. In these circumstances, a colour-Doppler scan may show reduced diastolic blood flow due to increased intrarenal pressure; also absence of the normal pulsatile jets of urine from the ureter into the bladder on the side with the acute obstruction. Ultrasound is not usually diagnostic of the cause of obstruction, but it may detect para-aortic nodes, a bladder mass, prostatic enlargement, or a ureterocele. Further investigation with transvaginal, transrectal, or transurethral ultrasound may confirm the cause of obstruction, transrectal ultrasound being particularly useful in the detection of local invasion from prostatic carcinoma.

Urinary tract stones

Renal stones appear on ultrasound as a bright echogenic focus with a distal acoustic shadow. Ultrasound can be used to follow up patients with renal calculus

disease by assessing the number and size of stones. Nephrocalcinosis may result in an increase in medullary echoreflectivity due to calcium deposition, which usually affects the whole medulla, whereas calcification from papillary necrosis has an appearance more like that of a renal stone.

Renovascular disease

Colour Doppler can be used to investigate renal arterial and venous disease. Thrombosis of major vessels produces absent flow or changes to the intrarenal blood flow pattern. More recently, colour-Doppler scanning has being used as a screening test for renovascular disease, with changes at the site of stenosis characterized by an increase in the peak systolic frequency followed by a diastolic spectral broadening. The sensitivity and specificity of this test has not been determined, and it remains a 'research' rather than a 'standard' clinical investigation.

Renal transplantation

Ultrasound examination is an important investigation in the management of the renal transplant recipient. Early graft dysfunction mandates investigation to exclude a technical problem with either the renal artery or vein, or a urinary leak. Colour-Doppler scanning provides valuable information about the vascular supply of the graft (Fig. 4). Fluid collections (commonly lymphoceles) appear as echo-free or echo-poor areas, and perinephric collections can be drained under ultrasound guidance for diagnostic purposes or to relieve obstruction. As with the native kidney, percutaneous nephrostomy is the emergency treatment of choice for obstruction of the renal transplant. Colour-Doppler scanning can detect the presence of arteriovenous fistulas, not uncommon following transplant biopsy.

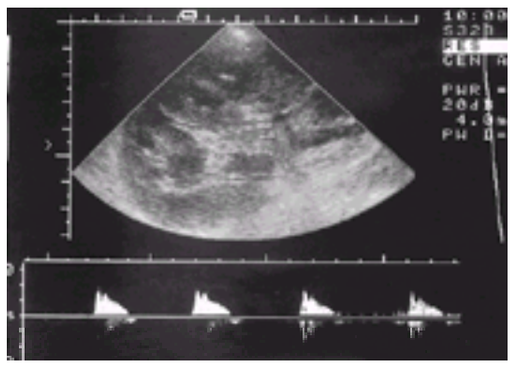

Fig. 4 Doppler ultrasound of a renal transplant showing normal systolic and diastolic wave forms.

Contrast agents for ultrasound

Colour-Doppler ultrasound can detect bubbles present in injected contrast medium. Application of this technique can change the use of ultrasound from simple anatomical visualization of the kidneys to dynamic testing. This will allow ultrasound to determine relative renal function, and improve investigation for obstruction and renovascular disease.

Computed tomography (CT) scanning

Computed tomography has advantages over conventional intravenous urography by imaging the perirenal and retroperitoneal spaces, and differentiating soft tissues within the kidney. Ultrafast CT scanning, without contrast, is being used more frequently to image ureteric renal stones and define the site of ureteric obstruction. In addition, CT provides vital information regarding the cause of ureteric obstruction by imaging the ureter, retroperitoneal space, and pelvis. Spiral, or helical CT scanning allows a three-dimensional reconstruction of the images, overcomes respiratory artefacts, and is useful in the investigation of congenital and anatomical abnormalities of the renal tract, such as renal agenesis. High-resolution CT scanning may detect early nephrocalcinosis, before calcification can be detected on plain films. These imaging techniques can be enhanced by contrast to give additional information: for example, simple renal cysts do not change in density following contrast, but occlusion of vessels may be demonstrated (Fig. 5).

Fig. 5 Contrast-enhanced CT scan showing thrombosed aorta and renal arteries.

Apart from the investigation of cystic renal disease, CT scanning is used to investigate renal masses. Renal-cell carcinomas vary in appearance: some show calcification both within and surrounding the tumour on non-enhanced scans, some are solid, and others are cystic or have necrotic centres. The majority of tumours are vascular and readily enhance with contrast, but those with heavy calcification may not. CT scanning is important in tumour staging, in determining the extent of perirenal spread, renal vein involvement, and enlargement of local lymph nodes. Occasionally, secondary deposits due to metastatic spread and secondary involvement in lymphomas and leukaemia, can be found on contrast-enhanced scans. These are usually small multiple intrarenal masses, often bilateral, typically homogenous, and solid in lymphomas. Although ultrasound is used to screen and assess Wilms' tumours in children, CT scanning is important in excluding pulmonary metastases.

Angiolipomas can be recognized with ultrasound, but should be confirmed on CT scanning as some renal-cell carcinomas may contain small amounts of fat. In tuberous sclerosis, angiolipomas may be associated with renal cysts. Although angiolipomas are benign mesenchymal tumours, they can rarely rupture, especially those with intrarenal haematomas and aneurysms. Early detection by CT scanning allows prophylactic embolization of these vascular lesions.

Renal oncocytomas, are another benign renal tumour, and on CT scanning may have a central lucent area due to fibrosis. However, a proportion of oncocytomas may become malignant. Thus any small renal lesion which is not a simple cyst or angiolipoma must be regarded as potentially malignant and therefore surveillance with repeat CT scanning (or ultrasound) should be recommended.

Renal tract imaging in patients with acute pyelonephritis is usually requested to exclude the presence of an obstruction, or when there has been an inadequate response to treatment. CT scanning defines the extent of disease better than ultrasound, detects abscesses, and can also exclude obstruction. Whereas focal acute bacterial pyelonephritis should respond to antibiotics, renal abscesses may require drainage. CT scanning may also detect gas bubbles within the renal parenchyma or perirenal space, characteristic of emphysematous pyelonephritis, typically found in diabetics. Similarly, CT scanning may establish a diagnosis of xanthogranulomatous pyelonephritis, with an enlarged kidney containing areas of scarring, focal loss of renal parenchyma, and multiple low-density masses, often following recurrent infections in patients with staghorn calculi.

In cases where renal trauma is suspected, contrast-enhanced CT scanning provides information not only about renal anatomy and function, but also perirenal collections, differentiating blood from urine. In addition, CT scanning provides valuable information about trauma to other intra-abdominal structures.

Magnetic resonance imaging (MRI)

CT scanning and ultrasound are good reliable techniques for detecting and evaluating renal masses. MRI is an alternative in patients who are allergic to conventional iodine-based radiocontrast media or those at risk of contrast nephropathy. The gadolinium contrast used in MRI is taken up by the proximal tubule, in a similar manner to aluminium, but has not been shown to cause nephropathy. MRI is expensive, but does have some advantages over conventional CT. Tissues surrounded by fat, such as enlarged lymph nodes, or tumour extension into the renal vein, are better demonstrated on MRI than CT. Thus MRI is useful in staging renal-cell carcinoma, and by being able to distinguish blood from tissue can help to differentiate simple cysts complicated by haemorrhage from those that are malignant.

The whole of the urinary tract can be visualized, in a manner similar to an IVU, by using a heavily weighted T_2 fast spin–echo sequence. This rapid acquisition and relaxation enhancement scan can be used to assess potential live donors for renal transplantation, by demonstrating the renal vasculature, renal anatomy, and urinary drainage with one investigation.

The quality of image provided by MRI can be very high (Fig. 6 and Fig. 7).

Fig. 6 Gadolinium-enhanced MRI showing left-sided pyelonephritic scarring, with a reduction in cortical thickness and scarring.

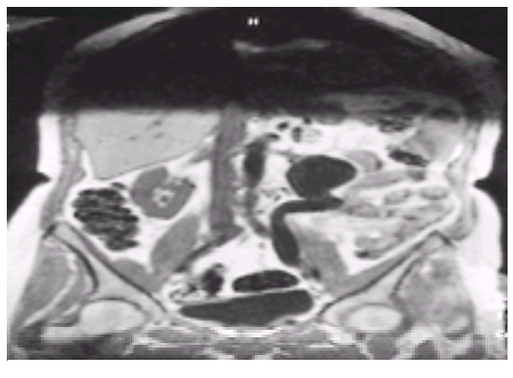

Fig. 7 Gadolinium-enhanced MRI showing a hydronephrotic left kidney and dilated upper two-thirds of the ureter following gynaecological surgery.

Angiography and digital subtraction angiography (DSA)

Formal renal angiography remains the 'gold standard' technique for assessing renovascular disease (Fig. 8). With the advent of expandable renal artery stents, it is important to determine precisely the anatomy of any stenosis, so that appropriate intervention can be planned. Direct pressure measurements can be made either side of any stenosis, so that the degree of stenosis can be assessed both anatomically and functionally. However, renal angiography is not without hazard: it involves an arterial puncture, the use of potentially nephrotoxic contrast agents, and carries the risk of dislodging aortic and renal artery plaques, which can result in intrarenal, intra-abdominal, and peripheral cholesterol embolization.

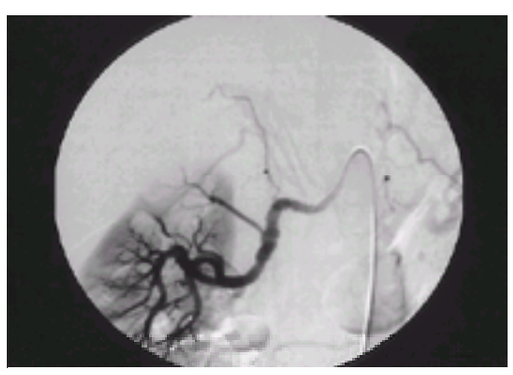

Fig. 8 Renal arteriogram showing fibromuscular hyperplasia of the renal artery.

Aside from the investigation of suspected chronic renovascular disease, renal arteriography can be indicated in the investigation of sudden renal ischaemia due to renal artery thrombosis or dissection, aortic dissection with extension into the renal arteries, or trauma to the renal artery. In addition, renal and coeliac arteriography can establish a diagnosis of classical macroscopic polyangiitis nodosa. Occasionally renal angiography is helpful in assessing renal tumour vascularity, and in determining whether partial nephrectomy can be performed. In some cases of persistent non-glomerular haematuria, formal renal angiography reveals a vascular abnormality as the underlying cause.

Digital subtraction angiography (DSA) uses a venous injection of contrast and computer-derived images to view the major renal arteries and intrarenal vessels. High doses of contrast media may be required, even so insufficient anatomical definition is obtained in between 5 and 20 per cent of cases. Thus, with appropriate indications, DSA is a good screening test but may need to be followed by formal angiography.

Interventional renal arteriography

Interventional renal arteriography should only be undertaken by experienced interventional radiologists with the support of vascular surgeons, as renal artery dissection or rupture may occur. Embolization with gel foam or metal coils can be used to selectively control renal haemorrhage, which is particularly useful when this

follows renal biopsy, and also in cases of arteriovenous malformation or tumour. Occasionally a whole kidney is embolized. Some renovascular stenotic lesions can be usefully treated by transluminal angioplasty or stenting.

Spiral CT angiogram

Spiral CT, by taking pictures which are then reconstructed by computer to provide a three-dimensional picture, can be used to investigate renal vascular disease. Radiocontrast is required, which can be administered by arterial or peripheral venous injection. To reduce the risk of contrast-induced nephropathy, or the possibility of 'flash pulmonary oedema' due to an intravenous volume load, some centres use carbon dioxide gas as the contrast agent. Compared to standard renal angiography, spiral CT renal angiography tends to overestimate any stenosis. The overestimation is greater with carbon dioxide than conventional contrast. However, this is a useful technique for excluding significant renovascular and intrarenal vascular disease.

Magnetic resonance angiography

Magnetic resonance angiography (**MRA**) using gadolinium chelates is indicated when there is a clinical risk of the patient developing contrast-induced nephropathy with standard renal or spiral CT angiography. MRA overemphasizes any stenotic area or other vascular abnormality. Thus MRA is useful in confirming normality, and is used in the preoperative assessment of living related kidney donors. A normal MRA of the renal arteries excludes renal artery stenosis, intrarenal vascular disease, and polyarteritis nodosa.

Magnetic resonance venography

As with MRA, magnetic resonance venography (**MRV**) using gadolinium contrast can be used to assess renal venous patency. Patients with nephrotic syndrome, and those with renal adenocarcinoma, may develop renal venous thrombosis, which can be difficult to positively diagnose with other imaging techniques.

Renal venography

Selective renal venous catheterization for blood sampling is still useful in patients with severe renovascular disease. The relative renal vein renin concentrations may aid the decision-making process in deciding whether to perform a surgical or medical nephrectomy in a patient with a small poorly functioning kidney due to severe renal artery stenosis.

Nuclear medicine

Static imaging

Radiolabelled dimercaptosuccinic acid (DMSA)

Technetium-labelled dimercaptosuccinic acid binds to renal proximal tubular cells, and after an intravenous injection some 70 per cent of the dose is taken up by viable tubules within 3 to 4 h. This can be detected by a gamma camera. DMSA scans provide information about the relative function of each kidney, and show areas of scarring due to renal stone disease, infection, and vascular disease. In children with urinary tract sepsis, suspected of reflux nephropathy, then serial DMSA scans are used to assess progressive cortical scarring. During acute pyelonephritis the DMSA scan may appear to show scars. These photopenic areas are due to inflammation and increased intrarenal pressure and can return to normal following resolution of infection. DMSA scans are also used to confirm the congenital absence of a kidney, to detect ectopic kidneys and other congenital malformations such as horseshoe kidney, and to confirm absence of renal function.

More recently, the introduction of single-photon emission computed tomography (**SPECT**) DMSA scans has improved resolution. These have shown that renal scars occur more frequently than previously thought, both in patients with acute pyelonephritis and also following lower urinary tract infection in renal transplant recipients.

Dynamic imaging

Radiolabelled diethylenetriaminepentaacetic acid and hippuran

Technetium-labelled diethylenetriaminepentaacetic acid (**DTPA**) or ^{131}I-labelled hippuran are both filtered by the glomerulus and then rapidly excreted by the kidney. These renograms have three phases: vascular, accumulation within the kidney, and excretion. Renal artery stenosis and acute tubular necrosis can reduce uptake, flattening the second and third phases of the renogram. Similarly, intrinsic renal disease flattens the second phase, and makes interpretation difficult when renal function is impaired.

Radiolabelled DTPA and hippuran scans are used to assess urological obstruction (Fig. 9). Occasionally patients with polycystic kidney disease present with severe pain due to the obstruction of a cyst, and DTPA scanning provides a dynamic test to confirm obstruction. In patients with dilated collecting systems, it is important to a differentiate congenital megaureter from an obstructed system. Excretion may be slow due to pooling in a dilated system, but obstruction is unlikely if there is a brisk wash-out following the administration of intravenous furosemide. Patients with impaired renal function may have a reduced response to furosemide, making interpretation of the renogram less reliable. Thus, in cases with impaired renal function, direct pressure measurement within the renal pelvis following percutaneous puncture may be required to exclude partial obstruction (Stamey test).

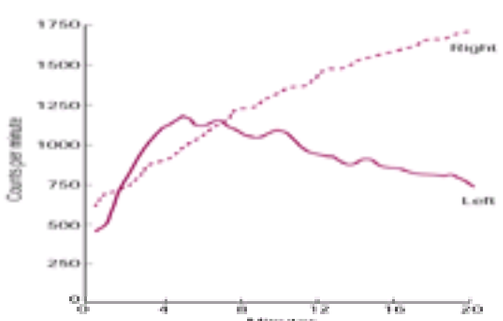

Fig. 9 DTPA renogram showing increasing uptake by the right kidney in a case of right-sided ureteric obstruction.

DTPA scans are also used to detect reflux in children, as reflux may be demonstrated during the 'emptying' phase of the renogram. If not, then an indirect micturating cystogram can be performed using the radioactivity which has passed into the child's bladder.

Following renal transplantation, DTPA and hippuran isotope scans can be used to monitor graft function. In cases of major arterial or venous thrombosis, and hyperacute rejection, the graft appears to have no perfusion. Acute tubular necrosis, rejection, and immunophylin toxicity may all have similar appearances. Serial scans can help to differentiate these conditions. DTPA scans may also reveal perirenal and urinary leaks before they are clinically manifest. Later isotope scans may detect obstruction due to ureteric stenosis.

Radiolabelled mercaptoacetyltriglycine (MAG3)

Technetium-labelled mercaptoacetyltriglycine ([^{99}Tc]MAG3) is protein-bound, and renal excretion is both by glomerular filtration and renal proximal tubular secretion. The excretion pattern is similar to that of hippuran and DTPA. The advantage of MAG3 is that it provides better image definition than DTPA and hippuran, especially

in those with impaired renal function. Thus MAG3 can be used to provide both anatomical (as DMSA) and functional (as DTPA and hippuran) information (Fig. 10).

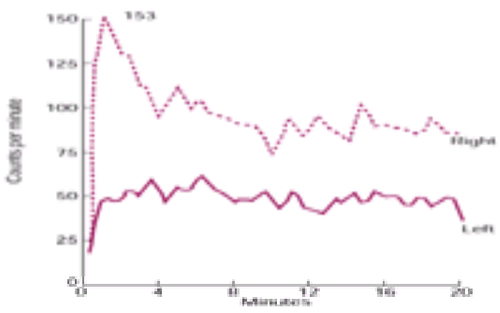

Fig. 10 MAG3 renogram demonstrating reduced uptake by the left kidney in a case of left-sided renal artery stenosis.

Patients with renal artery stenosis may have a delay in uptake time (the time taken from injection to peak activity) and an increased intensity and duration of MAG3 accumulation (due to increased tubular salt and water reabsorption, not seen in the case depicted in Fig. 10). If there is major stenosis of a major branch artery, then perfusion to one pole may be delayed. To improve the sensitivity and specificity of the MAG3 renogram in the detection of renal artery stenosis, some centres employ a method whereby two scans are performed—one with, and one without, prior administration of captopril. The captopril–MAG3 renogram can also be used as a screening test to determine whether the use of angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers might be detrimental to renal function in patients with an increased risk of atheromatous renovascular disease, including those with severe cardiac failure or diabetes and elderly hypertensive patients.

Other isotopes

Methyldiphosphonate (**MDP**) is filtered by the glomerulus, providing an immediate dynamic renogram. It is later taken up by inflamed muscles (found in patients with myositis and rhabdomyolysis) and the skeleton (detecting single or multiple bone metastases, and also metabolic bone disease in patients with endstage renal failure).

Renal biopsy

Indications

A renal biopsy should be considered in any patient with disease affecting the kidney when the clinical information and other laboratory investigations have failed to establish a definitive diagnosis or prognosis, or when there is doubt as to the optimal therapy. All renal biopsies have the potential to result in morbidity and (on rare occasions) mortality. The risk of biopsy must therefore be outweighed by the potential advantages of the result to the individual patient. Biopsies which would be 'of interest', but 'not in the patient's interest', should not be performed. Indications for renal biopsy should therefore be considered on an individual basis. Table 2 sets out the clinical presentations that warrant native renal biopsy.

Diabetic patients with proteinuria would not normally be biopsied, unless they had other conditions suggesting there might be an alternative or additional diagnosis to diabetic nephropathy. Most paediatricians would treat small children presenting with nephrotic syndrome with steroids, and only consider renal biopsy if they did not respond to treatment. Some conditions, in particular lupus nephritis and membranous glomerulonephritis, may change histological grading, so requiring repeat biopsy.

Renal biopsy is an important investigation in the management of the renal transplanted patient. Postoperative oliguria requires urgent investigation to differentiate acute ischaemic tubular necrosis from immunophylin (ciclosporin or tacrolimus) or other drug toxicity, acute rejection (vascular and/or cellular), or even frank infarction. Further biopsies may be required to monitor the response to antirejection therapy, and at a later stage to examine for recurrence of the original renal disease, or *de novo* glomerulonephritis in the graft.

Contraindications

Percutaneous renal biopsy should not be undertaken in patients with polycystic kidney disease. Similarly, patients with renal masses, such as tumours or cysts, should only be biopsied under direct vision, either by real-time ultrasound or CT scanning, or by formal open surgical biopsy. Patients with a solitary (or solitary functioning) native kidney are normally considered for open surgical biopsy.

Haemorrhage is more likely to occur in patients with uncontrolled hypertension, hereditary or acquired coagulation disorders, and those taking anticoagulants or antiplatelet agents. Blood pressure should be controlled and coagulation abnormalities treated before biopsy. Patients with renal amyloid also have an increased risk of haemorrhage, as may those with classic polyarteritis nodosa.

Patients with chronic renal failure and bilaterally small kidneys should not undergo biopsy. This would be technically difficult (the kidneys are small and hard) and the biopsy appearances of endstage renal failure are exceedingly unlikely to provide any information that might alter the clinical course or management. Percutaneous renal biopsy should not be performed in patients with untreated acute pyelonephritis due to the risk of developing a perinephric abscess.

Technique

'Blind' biopsy of the native kidney, meaning biopsy without imaging for localization, should not be performed unless there are truly exceptional circumstances. It is possible to visualize the kidney and biopsy under fluoroscopic control after injection of radiocontrast medium as for an IVU, but the most commonly used method for directing biopsy is ultrasound guidance. This can either be used to record the depth of the lower pole from the skin and mark the surface position vertically above it on inspiration, or to provide real-time guidance. The latter technique is described below.

Percutaneous renal biopsy should be carried out using sedation and local anaesthesia. Children may require general anaesthesia. The patient should be placed prone on top of pillows or folded sheets to compress the upper abdomen and lower ribs and fix (to some degree) the position of the kidneys. Under real-time ultrasound the kidneys are visualized, the patient asked to take and hold a deep breath in inspiration, and the kidney which is thought to be technically the most easy to biopsy is targeted. To avoid the major vessels, the aim should be for the lateral border of the lower pole. Either 14- or 18-gauge, trucut-type needles are commonly used, some centres now use an automated spring-loaded biopsy gun. Under direct vision the needle tip is advanced to the renal capsule, and with the kidney fixed in inspiration, biopsy is performed. The advent of colour Doppler means that the operator can deliberately avoid the major intrarenal vessels.

Transjugular biopsy can be performed in patients who have an increased likelihood of bleeding complications. Technical developments have now allowed biopsy needles to be passed reliably from the renal vein into the renal cortex, such that in our own institution all such biopsies in the last 3 years have been diagnostic. Occasionally, open surgical biopsy is required, with the biopsy taken under direct vision and local bleeding controlled.

Renal transplants, usually placed in one or other iliac fossa, are biopsied in the supine position. Pillows can be placed under the side with the transplant to help move bowel and fat pad away from the transplant. Biopsies are taken from the lateral border of the upper pole, avoiding the major vessels and ureter.

The obvious risk of renal biopsy is haemorrhage. All patients should be placed on strict bed rest for at least 6 h after the procedure, and pulse and blood pressure should be checked frequently during this period. Hypotension, tachycardia, abdominal/back pain, and macroscopic haematuria are indications for urgent medical review.

Complications

Postbiopsy scanning has shown that the vast majority of patients develop a perirenal haematoma, which is usually asymptomatic. Arteriovenous fistulas may also develop acutely following biopsy. The majority disappear spontaneously with time, and only the occasional one requires treatment by the interventional radiologist. Macroscopic haematuria occurs in fewer than 10 per cent of patients, and bleeding sufficient to warrant blood transfusion in around 1 per cent. Rarely, severe haemorrhage may require treatment with the insertion of coils or gel foam embolization. Exceptionally, death may occur, usually due to the failure to detect haemorrhage and provide appropriate resuscitation.

Complication rates are increased in patients with both acute and chronic renal failure. Uraemia prolongs the bleeding time, even when the conventional coagulation screening is normal (prothrombin time, activated partial thromboplastin time, and peripheral platelet count). The risk of uraemic haemorrhage can be at least partially reversed prior to biopsy by good dialysis to improve platelet function, correction of the haematocrit and any underlying coagulation defect, and by giving an infusion of deamino-D-arginine vasopressin (**DDAVP**; desmopressin) immediately prior to the procedure (0.3 µg/kg over 30 min).

Further reading

Urine microscopy

Birch DF, *et al*. (1994). *A color atlas of urine microscopy*, 1st edn. Chapman and Hall, London.

Fogazzi GB, *et al*. (1993). *The urinary sediment. An integrated view*. Masson, Milan.

Renal function

Davison AM, *et al*. (1997). *Oxford textbook of nephrology*, 2nd edn. Oxford University Press, Oxford.

Randers E, *et al*. (1998). Serum cystatin C as a marker of the renal function. *Scandinavian Journal of Clinical and Laboratory Investigation* **58**, 585–92.

Seldin DW, Giebisch G (1992). *The kidney physiology and pathology*, 2nd edn. Raven Press, New York.

Valtin H, Schafer JA (1994). *Renal function*, 3rd edn. Little Brown, Boston, MA.

Renal imaging

Allan PL, Dubbins P, Pozniak MA (1997). *Clinical Doppler ultrasound*. Churchill Livingstone, Edinburgh.

Ghantous VE, *et al*. (1999). Evaluating patients with renal failure for renal artery stenosis with gadolinium enhanced magnetic resonance angiography. *American Journal of Kidney Diseases* **33**, 36–42.

Helenon O, *et al*. (1997). Renovascular disease: Doppler ultrasound. *Seminars in Ultrasound* **18**, 136–42.

Testa HJ, Prescott MC (1996). *Nephrourology, British Nuclear Medicine Society*, 1st edn. BPC Wheatons, Exeter.

20.4 Acute renal failure

J. Firth

The clinical approach to the patient with acute renal failure

Introduction

Acute renal failure is defined as a significant decline in renal excretory function occurring over hours or days. This is usually detected clinically by a rise in the plasma concentration of urea or creatinine. Oliguria, defined (arbitrarily) as a urinary volume of less than 400 ml/day, is usually present, but not always. Acute renal failure may arise as an isolated problem, but much more commonly occurs in the setting of circulatory disturbance associated with severe illness, trauma, or surgery; transient renal dysfunction complicates some 5 per cent of medical and surgical admissions. A community-based study conducted during 1993 reported that the incidence of severe acute renal failure in adults (serum creatinine >500 µmol/l) was 172 per million, rising from 17 per million in those under 50 years of age to 949 per million in those aged between 80 and 89 years. A recent (year 2000) study from renal units and intensive care units (ICUs) in a defined geographical area of Scotland found that 131 patients per million per year required renal replacement therapy for acute renal failure. There are many possible causes (Table 1, Table 2, and Table 3), but in any given clinical context few of these are likely to require consideration.

Diagnosis of the presence of acute renal failure

A high index of clinical suspicion is required to diagnose acute renal failure at an early stage of its development. This is because symptoms and signs attributable to the accumulation of fluid, electrolytes, acid or uraemic wastes within the body may not be apparent until the condition is far advanced. Furthermore, the symptoms and signs that may arise are not specific: unsuspected hyperkalaemia is the greatest danger, since this may produce no symptoms whatsoever before causing cardiac arrest.

All patients admitted to hospital with acute illness should be considered at risk of developing acute renal failure. Those who have some pre-existing chronic impairment of renal function are particularly susceptible to acute exacerbations. This group includes all elderly patients, in whom a combination of low muscle mass and low dietary meat consumption may conspire to maintain an apparently 'normal' plasma creatinine level, despite a reduction in glomerular filtration rate to as little as 25 per cent of that expected in a healthy young adult.

To recognize impairment of renal function early, the basic care of all acutely ill patients should include careful monitoring of fluid input and output, daily weighing, lying and standing (or sitting) blood pressure, and regular estimation of plasma creatinine, urea, and electrolytes. Although it might seem to the physician to be a simple matter to monitor fluid input and output, this simplicity is often only present in theory, excepting in patients who are restricted to parenteral fluids and who have a urethral catheter. Drinks may be spilt, extra drinks may be acquired from a variety of sources, urine may be spilt, and vomit and diarrhoea are often found in places where they are difficult to quantitate. These considerations mean that the most likely explanation for fluid balance charts being difficult to interpret is the erroneous recording of input or output. Daily weighing on accurate scales provides a much more reliable picture of net overall fluid balance. Patients who are acutely ill invariably lose flesh weight, commonly at a rate of up to a few hundred grams per day. If weight appears to fall at a rate faster than this, then negative fluid balance is likely: the occurrence of greatly increased 'insensible' losses through the skin and lungs during fever being a common explanation. Aside from weight loss, the development of a postural drop in blood pressure is a reliable sign that a patient has become significantly volume-depleted. If weight rises at any time, then this must be due to positive fluid balance, whatever the input/output charts may suggest. It may not be obvious from clinical examination where the fluid has gone: the possibilities of sequestration in the peritoneal cavity or in the tissue interstitium should be recognized.

Plasma urea, creatinine, and electrolytes should be measured on admission in all acutely ill patients, and repeated daily or on alternate days in those who remain so. These measurements will ensure that advanced acute renal failure does not seem to have occurred 'suddenly' in patients already in hospital. However, many patients will be found to have significant renal impairment on admission, and many more will develop some degree of renal impairment whilst on the ward. In all cases the physician must try to make a precise diagnosis of the cause.

Diagnosis of the cause of acute renal failure

In the initial assessment of a patient who appears to have acute renal failure three questions should be asked.

Question 1: is the renal failure really acute?

The only basis for excluding the possibility of pre-existing chronic renal impairment with absolute confidence is the knowledge of a previous normal measurement of renal function. In cases where there is uncertainty, a diligent search for previous notes and biochemical information may save the patient and the doctor the inconvenience (and occasionally hazard) of unnecessary investigation. The finding of two small kidneys on ultrasound examination indicates the presence of chronic renal disease. Other clinical features are poor discriminators between acute and chronic renal impairment. A history of vague ill health of some months' duration, of nocturia, of pruritus, or the findings of skin pigmentation or anaemia would all suggest chronicity (see Section 20.5). However, anaemia is not invariable in chronic renal failure (for example, in polycystic kidney disease the haemoglobin concentration may be normal), and anaemia can develop over a few days in acute renal failure, as may hypocalcaemia and hyperphosphataemia. Radiological evidence of renal osteodystrophy is only found in patients with obviously long-standing renal failure and never aids the clinical distinction between acute and chronic renal failure.

Question 2: is urinary obstruction a possibility?

One of the merits of the traditional division of the causes of acute renal failure into prerenal, renal, and postrenal is that it encourages consideration of the possibility of urinary obstruction, which in community studies accounts for about 25 per cent of severe acute renal failure cases, mostly due to prostatic obstruction.

It is extremely important that obstruction should not be missed, since most cases are readily treatable and delayed diagnosis may lead to permanent renal damage. Obstruction is particularly likely to cause acute renal failure in those with a single functioning kidney, in those with a history of renal stones or of prostatism, and after pelvic or retroperitoneal surgery, but the possibility of obstruction should be seriously considered in all cases where another positive diagnosis cannot be made. The presence of anuria, or of alternating polyuria and oligoanuria, are helpful clues. However, it is not widely appreciated that a patient may pass normal or elevated volumes of urine despite significant obstruction, although this is extremely rare. The mechanism is poorly understood, but three factors present in obstruction tend to

impair urinary concentrating ability, thereby leading to the preservation of urinary volume despite obstructive depression of the filtration rate. These factors are structural damage to the inner medulla and papilla, functional changes in the distal nephron resulting from increased intraluminal or interstitial pressure, and loss of medullary hypertonicity at low filtration rates.

Ultrasound examination of the kidneys and bladder is the usual first method of investigation for the presence of obstruction. However, it is important to remember that the quality of the image obtained by renal ultrasonography is highly variable, depending on the patient, the equipment, and the operator. Furthermore, ultrasound detects calyceal dilatation, not obstruction, and the test may be 'negative' (because the calyces fail to dilate, or do so only minimally) in about 5 per cent of cases of acute obstructive renal failure. If doubt as to the diagnosis persists in the clinician's mind, then the examination should be repeated, and other investigations pursued if uncertainty still remains. If renal function is adequate (creatinine concentration less than about 250 µmol/l) then **DTPA** (diethylenetriaminepentaacetic acid) or **MAG3** (mercaptoacetyl triglycine) renography with furosemide (frusemide) injection may be helpful, showing delayed excretion and clearance of radionucleide from the obstructed kidney(s). If renal function is severely impaired then imaging modalities that depend upon renal excretion (including intravenous pyelography) are not useful, and percutaneous antegrade nephrostomy/pyelography or cystoscopy with retrograde ureteric catheterization and pyelography should be undertaken. (See Chapter 20.14 for further discussion.)

Obstruction, once diagnosed, must be relieved urgently by bladder catheterization, percutaneous nephrostomy, or cystoscopic insertion of ureteral stents, as a prelude to definitive treatment (where possible) of the underlying obstructive lesion. The most important causes of urinary obstruction are renal calculi, retroperitoneal fibrosis, and malignant diseases of the uterine cervix, prostate, bladder, and rectum (see Chapter 20.14).

Question 3: are glomerulonephritis, interstitial nephritis, vasculitis, or other rarities possible?

To make these diagnoses, which although rare have critically important management implications, stick testing of the urine and microscopy of the urinary sediment is an essential part of the assessment of any patient with unexplained acute renal failure. If stick testing indicates more than + of protein or more than a trace of blood, then a sample of urine should be examined under the microscope. This should be done by centrifuging 10 to 15 ml of urine at 1500 to 2500 r.p.m. (approximately 400 to 1120 g) for 5 min, carefully discarding all but 1 ml of the supernatant, and then resuspending the pellet. Examination should be made under high power, preferably after staining, which makes the cellular elements of casts more obvious. Red cell casts (Fig. 1) are present in acute glomerulonephritis, renal vasculitis, accelerated-phase hypertension, and (sometimes) in interstitial nephritis, but not in other conditions. Their presence indicates the need for urgent specialist renal referral.

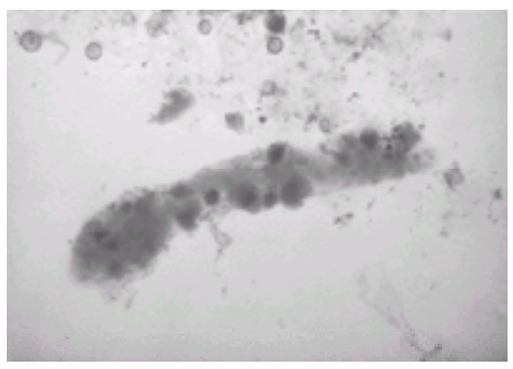

Fig. 1 A red cell cast. The red cells incorporated in the cast, which has a typical cylindrical appearance, look dark in this tone image due to staining with a blue dye.

Clinical features of acute renal failure

In the early stages of acute renal failure there are few warning symptoms. The patient may notice a reduction in urinary volume, but non-oliguric renal failure comprises as many as 50 per cent of cases in some series, and most patients who are unwell do not drink as much as usual and therefore are not concerned if they pass little urine. The clinical picture is likely to be dominated by the primary condition, of which acute renal failure is a complication, and by the effects of intravascular volume depletion, with dizziness caused by postural hypotension a common reason for patients being brought to medical attention.

In the later stages of acute renal failure there are manifestations of uraemia with anorexia, nausea, vomiting (or occasionally diarrhoea), muscular cramps, and signs of encephalopathy—including a 'metabolic' flapping tremor (asterixis), progressing in extreme cases to depressed consciousness and *grand mal* convulsions. Skin bruising and gastrointestinal bleeding may occur. Uraemic haemorrhagic pericarditis is another potentially fatal complication, but this occurs much less frequently in acute renal failure than in (neglected) chronic renal failure.

Biochemical changes

The clinical diagnosis of renal failure, acute or chronic, is made when the plasma urea and creatinine concentrations rise. Other important biochemical changes include the development of hyperkalaemia, metabolic acidosis, hypocalcaemia, and hyperphosphataemia. Hyperkalaemia is due not only to reduced urinary excretion, but also to potassium release from cells—either as a consequence of cell death or as a result of metabolic acidosis. Particularly rapid rises are to be expected when there is extensive tissue damage or hypercatabolism, as in rhabdomyolysis, burns, and sepsis. Transfusion of stored blood is sometimes said to cause dangerous rises in plasma potassium concentration in oliguric patients. However, the transfused blood may not really be to blame, but the circumstances that demand transfusion. Loss of blood into the gastrointestinal tract or body tissues is followed by red cell lysis and the absorption of a considerable potassium load.

Protein catabolism produces sulphuric and phosphoric acids. These are normally buffered by bicarbonate and excreted by the kidney. In acute renal failure these systems fail, leading to the development of acidosis. This is usually modest in degree (plasma pH 7.2–7.35), but can be more severe, manifesting as sighing Kussmaul respiration and/or with circulatory compromise. Acidosis is sometimes the metabolic abnormality most obviously necessitating urgent institution of renal replacement therapy, but overzealous administration of bicarbonate should be avoided (see below).

Calcium malabsorption occurs early in acute renal failure and is probably secondary to disordered vitamin D metabolism. Hypocalcaemia can develop with surprising rapidity. It is usually asymptomatic, but tetany and fits may be provoked by injudicious over-rapid correction of acidosis with resultant depression of ionized calcium. Profound hypocalcaemia and marked hyperphosphataemia, together with hyperuricaemia, is to be expected in rhabdomyolysis. Transient hypercalcaemia is frequently seen during the recovery phase from acute renal failure, and this is particularly common after rhabdomyolysis, probably being caused by secondary hyperparathyroidism related to preceding hypocalcaemia. The hypercalcaemic phase may be prolonged and accompanied by metastatic calcification in patients in whom there has been extensive muscle injury.

The plasma sodium concentration is usually normal in cases of acute renal failure: any deficit of sodium is usually matched by that of water, thus leading to reduction of the extracellular fluid volume but with an unchanged plasma sodium concentration. However, on occasion the intake of water, either drunk in response to thirst or inflicted iatrogenically, may exceed the rate of excretion such that hyponatraemia results.

The retention of uric acid, sulphate, and magnesium occurs in acute renal failure, but these biochemical abnormalities are rarely clinically significant, with the exception of the grossly elevated levels of uric acid that can be seen in rhabdomyolysis and following tumour lysis.

General aspects of medical management

The immediate management of the patient with renal impairment is directed towards three goals. The first is the treatment of any life-threatening complications of acute renal failure. The second is prompt diagnosis and treatment of hypovolaemia. The third is specific treatment of the underlying condition: if this persists untreated then renal function will not improve.

Life-threatening complications

Hyperkalaemia (see also Chapter 20.2.2)

Hyperkalaemia is most commonly dangerous in the context of acute renal failure, and is important because it can cause cardiac arrest. Patients may occasionally notice muscle weakness or paralysis, but the significance of these symptoms is rarely appreciated, and usually there are no symptoms whatsoever. All doctors who work with acutely ill patients should be able to recognize the characteristic electrocardiogram (**ECG**) appearances, which are a better indicator of cardiac toxicity than the serum potassium level. As serum potassium rises, the following changes progressively occur (Fig. 2):

1. 'tenting' of the T wave;
2. reduction in size of P waves, increase in the PR interval, widening of the QRS complex;
3. disappearance of the P wave, further widening of the QRS complex;
4. irregular 'sinusoidal' ECG;
5. asystole.

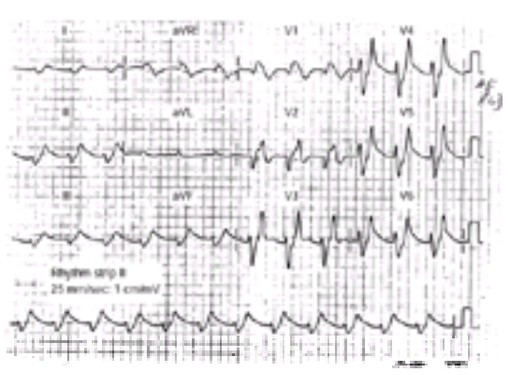

Fig. 2 An electrocardiogram showing severe hyperkalaemic changes in a patient with a serum potassium level of 8.6 mmol/l.

Treatment of hyperkalaemia is described in Table 4.

Pulmonary oedema

The most serious complication of salt and water overload in acute renal failure (usually iatrogenic) is the development of pulmonary oedema. Severe cases are dramatic. The patient is terrified, restless, and confused. Examination reveals cyanosis, tachypnoea, tachycardia, widespread wheeze or crepitations in the chest, and a gallop rhythm (if the heart can be heard). Investigation demonstrates arterial hypoxaemia and widespread interstitial shadowing on the chest radiograph. (See Chapter 15.15.2.2 and Chapter 16.1 for further discussion.)

The patient should be sat up and supported, and given oxygen by face-mask in as high a concentration as possible using a reservoir bag. Furosemide (frusemide) may work as a venodilator but is unlikely to provoke a substantial diuresis in a patient with renal failure. Morphine can relieve symptoms rapidly and should be given in small (2.5 to 5 mg) doses, repeated if necessary and if tolerated, and with the opioid antagonist naloxone to hand in the event of deterioration due to toxicity. An intravenous infusion of a venodilator such as isosorbide dinitrate may be helpful.

The definitive treatment for pulmonary oedema caused by renal failure is the removal of fluid by haemodialysis or haemofiltration. Acute peritoneal dialysis is much less effective in this capacity and should only be considered in circumstances where haemodialysis and haemofiltration are not available. The immediate beneficial effects of venesection of 200 to 400 ml of blood from the patient *in extremis* should not be forgotten.

Recognition and treatment of volume depletion

A key part of the immediate assessment and management of any patient who is very ill, which will include many of those with acute renal failure, is to make a correct assessment of the intravascular volume status and to resuscitate rapidly and effectively. (See Chapter 16.1 for further details.)

Fluid and electrolyte requirements in established acute renal failure

Fluid

Many patients with acute renal failure are volume-depleted at the time of presentation. An urgent priority is to correct such depletion rapidly. Once this has been achieved—as judged by an improvement in peripheral perfusion, a fall in pulse rate, loss of postural drop in blood pressure, and a rise in jugular venous pressure—the perspective changes. In the absence of normal renal function the greatest care must be taken to regulate the intake of fluids and electrolytes to match losses in the urine, from the gastrointestinal tract, and from other 'insensible' sources. As a working rule, fluid intake is limited to the volume of the previous day's urine output and gastrointestinal losses, plus 500 ml, but this allocation may need to be substantially increased in the presence of fever or in hot environments, when insensible losses may be much increased. However, as discussed above, fluid-balance charts are frequently inaccurate and unthinking adherence to the 'output plus 500 ml' rule can lead to grief. There is no substitute for careful, twice-daily clinical examination for signs of intravascular volume depletion or excess, supplemented by accurate daily weighing to gauge the overall net fluid balance, and an intelligent flexible response to the findings.

Sodium

In the patient who is not being dialysed, the intake of sodium must also be matched to output. Requirements are usually very small in those who are oliguric, perhaps only 15 to 30 mmol/day, but if the patient is polyuric the requirements can be considerable, with a danger of volume depletion if these are not met. The urine of a patient with polyuric renal failure will usually contain sodium at a concentration of 50 to 70 mmol/l, hence if urine output is 3 litres/day then over 200 mmol of sodium may be required. On occasion, the urine output in polyuric acute renal failure can be massive (even up to 1 litre/hour)—if the response is to administer an even greater quantity of fluid (output plus insensible losses), then it is possible to contrive a vicious cycle whereby an ever-increasing urinary output is rewarded by ever-increasing fluid infusion. To avoid this situation in the patient with polyuria it is best to limit input to urinary output alone, thus allowing other fluid losses to establish a mild overall negative balance, only increasing fluid input if the patient develops significant postural hypotension, which should be checked for twice daily. For unknown reasons, an excess of sodium and water in patients with tubular necrosis leads to peripheral or pulmonary oedema, whereas in those with glomerulonephritis it tends to produce hypertension.

Potassium

Because hyperkalaemia is one of the most important problems in the management of acute renal failure, it is essential to check plasma potassium levels at least daily, and in those with hypercatabolism or gastrointestinal bleeding, or who require surgery, more frequent estimations are advisable. In oliguric cases, dietary consumption should be limited to the minimum compatible with an adequate intake of protein and amino acids (20–30 mmol/day).

Diuretics that work on the distal tubule (for example, spironolactone, amiloride, and triamterene) promote potassium retention: they should never be used in renal failure, and it is important when reviewing the drug chart to remember that these agents are frequent constituents of tablets containing a combination of diuretic/antihypertensive compounds. Intravenous preparations of antimicrobial agents that contain large amounts of potassium should also be avoided whenever

possible.

Excretion of potassium can sometimes be enhanced in those who are oliguric by the use of high doses of furosemide (0.5–1 g daily). Oral potassium-exchange resins (e.g. Calcium Resonium), prescribed concurrently with a laxative, can be useful in controlling serum potassium for a few days or weeks, but they are not effective treatments for acute severe hyperkalaemia (see Table 4) and are usually found to be unpalatable for long-term use. By contrast, in polyuric acute renal failure substantial losses of potassium can occur and need to be replaced. Measurement of the urinary potassium concentration can be helpful in estimating how much potassium is required.

Renal replacement therapy

Mandatory indications for immediate instigation of renal replacement therapy are:

1. refractory hyperkalaemia;
2. intractable fluid overload;
3. acidosis producing circulatory compromise;
4. overt uraemia manifesting as encephalopathy, pericarditis, or uraemic bleeding.

These indications will be present in some patients on their admission to hospital. However, in most cases renal function will be seen to decline over a period of days or a few weeks despite optimal medical therapy. In this situation here is no hard and fast rule as to when renal replacement therapy should be initiated. There is no level of nitrogenous waste at which the patient suddenly becomes susceptible to overt uraemic sequelae. Nevertheless, it is clearly not sensible to wait until an obvious uraemic complication (which might be fatal) arises. Modern practice is (whenever possible) to begin renal replacement therapy when the blood urea reaches 25 to 30 mmol/l and the serum creatinine 500 to 700 µmol/l, unless there is clear evidence that spontaneous recovery is occurring. There are three basic options for renal replacement therapy: peritoneal dialysis, haemodialysis, and haemofiltration.

Peritoneal dialysis

Peritoneal dialysis is technically the simplest form of renal replacement therapy and is commonly used worldwide, although remarkably little has been published recently about its use in those with acute renal failure. The principle is the same as that described for the long-term treatment of patients with chronic renal failure (see Section 20.5), the major differences being: (1) that catheters are used which can be inserted percutaneously using a metal stylet (although some use the same type of catheter as that used for continuous ambulatory treatment); and (2) that smaller volume exchanges with shorter dwell-times are the norm. The technique requires an intact peritoneum and is therefore precluded in the many patients whose renal failure is associated with abdominal surgery. Other problems include difficulties in maintaining dialysate flow, leakage, peritoneal infection, protein losses, and restricted ability to clear fluid and uraemic wastes. These limitations mean that, particularly in the hypercatabolic patient, peritoneal dialysis is frequently unable to provide good dialysis of the patient with acute renal failure as judged by modern standards. It is fair to say that peritoneal dialysis is virtually never the first choice modality for renal replacement therapy in an adult with acute renal failure in those centres that have a range of techniques at their disposal.

Haemodialysis and haemofiltration

Traditional haemodialysis, which is usually performed on alternate days but may be associated with better outcome when applied daily, can provide good control of uraemia in patients with acute renal failure who do not have severe haemodynamic compromise. The major disadvantage and limitation of the technique (apart from cost) arises from the fact that it is intermittent: in each 4-h treatment at least 2 to 3 litres of fluid must typically be removed to make 'space' either for the infusion of drugs/parenteral fluids or for oral fluid intake during the 24- to 48-h period before the next dialysis. This imposes a substantial haemodynamic stress, which often cannot be tolerated by those who are cardiovascularly unstable, and is the main reason why continuous haemofiltration techniques have largely replaced haemodialysis in intensive care units.

The standard haemofiltration technique works as follows: a mechanical pump (but sometimes the patient's own arterial pressure) drives blood through a haemofilter of high hydraulic conductivity. An ultrafiltrate of plasma is removed, usually at a rate of between 1 and 2 litres per hour. This is replaced, minus the volume of other fluid inputs and the amount of 'negative balance' required, using (most commonly) a lactate/acetate-based substitution fluid. The process is tolerated well, even by patients who are very ill, and the continuous nature of the technique permits continuous fine tuning of the intravascular volume. A large number of technical variations are possible—for example, combination of filtration and dialysis elements (haemodiafiltration), use of differing replacement fluids—but there is nothing to suggest that any one of these is better than another, excepting in those who are unable to metabolize lactate, when bicarbonate-based substitution fluid is essential.

In the same way that there is no evidence on which to make firm recommendations as to when to start renal replacement therapy in those with acute renal failure whose chemistry is gradually 'going off', there is also little information on which to base targets for the clearance of metabolic wastes that should be achieved by treatment. One recent study compared the outcome of patients treated with different doses of venovenous haemofiltration: those randomly assigned to ultrafiltration at a rate of 20 ml/h per kg did less well than those receiving 35 ml/h per kg or 45 ml/h per kg, there being no significant difference between the latter two groups.

Other issues in the management of patients with acute renal failure

Indications for renal biopsy

Most cases of acute renal failure are due to prerenal failure or to the clinical syndrome of acute tubular necrosis. They occur in an appropriate clinical setting and follow a typical time course, with recovery of renal function over a few weeks. In such instances renal biopsy should not be performed, since the information gained is exceedingly unlikely to influence management, and the risks of the procedure are therefore not warranted. There are, however, circumstances in which renal biopsy is essential to establish a correct diagnosis, with important implications for both management and prognosis. Biopsy should be considered when:

1. the history, examination, or laboratory tests suggest a systemic disorder that could cause acute renal failure and could be diagnosed by renal biopsy;
2. the urine sediment contains red cell casts;
3. the case history is atypical; and
4. renal failure is unusually prolonged (say beyond 6 weeks), although in this context cortical necrosis (see below) is better diagnosed by computed tomography (**CT**) scanning or angiography.

Nutrition

Patients with acute renal failure are invariably catabolic and derive a larger fraction of their energy expenditure from protein breakdown than normal. Insulin resistance, metabolic acidosis, the release of proteinases into the circulation, and changes in the metabolism of branched-chain amino acids have all been suggested as possible reasons. If nutrition is neglected, patients with acute renal failure lose weight very rapidly, and those that lose most have the highest mortality. However, it has not been proven in controlled trials that any form of nutritional support can generate a positive nitrogen balance, improve nutritional status, or alter the mortality rate in patients with acute renal failure. Nevertheless, there is a consensus that early institution of nutritional support probably improves prognosis. Despite this, and almost certainly to the patient's detriment, action is frequently delayed or not taken at all, particularly if it is thought that the extra fluid load required will mandate the institution of dialysis or the need for additional dialysis sessions in an already busy unit.

Typical recommended daily adult requirements are total energy 35 kcal/kg body weight, protein 1 g/kg but and nitrogen 0.16 g/kg but there is no good evidence on which to base stipulations and some would advocate more calories and more protein for those who are catabolic. If patients with acute renal failure are oliguric, the nutritional support should be given in a restricted fluid volume, with reduced amounts of sodium, potassium, and phosphate. For practical purposes it is sensible to have enteral and parenteral fluids that satisfy these needs available routinely (a variety of commercial preparations are available): extra water and electrolytes can always be added when required. In the many patients who are too unwell to take adequate food by mouth, commonly those who need it most, tube feeding or parenteral nutrition should be started early. Protein restriction, aimed at moderating the rise of plasma urea, is not appropriate management for the patient with acute renal failure.

Bleeding

In uraemia the bleeding time is prolonged, and in acute renal failure this summates with any abnormality of haemostasis that might be simultaneously induced by the precipitating condition. Better control of uraemia and the routine use of H_2-receptor antagonists have been associated with a greatly reduced risk of upper gastrointestinal bleeding, a previously frequent and grave occurrence. Impairment of haemostasis is not a cause of great clinical concern in most patients, but there are some who bleed—from anywhere and everywhere. Guidelines for the management of such cases are given in Table 5.

Sepsis

Overwhelming septicaemia is a common cause of acute renal failure, and in such instances the diagnosis is often straightforward. However, in many more cases the role of sepsis is insidious and difficult to diagnose with certainty. There is often strong clinical feeling, but little in the way of hard proof, that sepsis underlies the slide towards worsening renal and multiorgan failure in patients who have been apparently successfully resuscitated from major trauma or surgery. Septicaemia is the commonest cause of death in those with acute renal failure. The index of clinical suspicion must therefore be very high: if a patient with acute renal failure appears to be deteriorating in any way, the question must be asked 'is this sepsis?'. Unused intravenous lines and urinary catheters should be removed, and those that are necessary but in any way 'suspicious' should be replaced. The patient should be examined regularly for signs of a septic focus. There should be a low threshold for repeated, thorough microbiological investigation. Proven infection should be treated promptly with appropriate antimicrobial agents (dose modified as required). In many cases, however, it will be necessary to start treatment 'blind', having taken specimens for culture and having made an educated guess as to the likely pathogen, with the possibility of Gram-negative septicaemia high on the list.

In the patient who appears 'obviously septic' or to be 'going off', but in whom no cause can be found, attention should be directed towards the abdomen, this being the most likely place for hidden mischief, either infective or ischaemic. Radiological investigations, in particular CT scanning, can be very useful in searching for abdominal sepsis or dead bowel, but should not be relied upon too faithfully. However, surgical exploration may be required, both to diagnose and to treat, especially in patients whose renal failure follows previous abdominal surgical procedures.

Prescription of drugs

Many drugs are excreted by glomerular filtration or tubular secretion and must be given in reduced dosage or at longer intervals than normal in patients with renal failure (see Chapter 20.16). For patients with acute renal failure the following should not be given without very good reason: non-steroidal anti-inflammatory drugs, angiotensin-converting enzyme inhibitors, angiotensin-II receptor antagonists (all of which have adverse effects on renal perfusion and glomerular filtration), and aminoglycoside antibiotics (these are discussed later in this chapter). A note about two other drugs that may be given to patients with acute renal failure is also appropriate here: both aciclovir and penicillins can cause encephalopathy if given in the doses used to treat severe infection in patients with normal renal function. The dose of aciclovir needs to be reduced from between 5 and 10 mg/kg every 8 h to between 2.5 and 5 mg/kg every 24 h in those receiving renal replacement therapy, and physicians should restrain themselves from prescribing the maximum recommended doses of penicillins. If in doubt, consult the manufacturer's data sheet before prescribing any drug to a patient with acute renal failure.

Prognosis

Acute renal failure of sufficient severity to require renal replacement therapy carries a high mortality. In a series of over 1300 cases, the actuarial 1-year survival of all medical and surgical cases rose from 39 per cent to 58 per cent between 1956 and 1988, despite an increase in the median patient age from 41 to 61 years over this period. The prognosis varies according to the cause of acute renal failure: mortality is between 40 and 60 per cent in patients with renal failure as part of the multiple organ failure syndrome, but less than 10 per cent in those who have renal failure alone. Death should rarely be attributable to a primary sequel of renal failure, for example uraemia or hyperkalaemia, and the incidence of life-threatening gastrointestinal haemorrhage is much reduced: sepsis is the major killer. Patients die *with* but not directly *of* renal failure.

Specific causes of acute renal failure

Prerenal failure and acute tubular necrosis

Introduction

Between 80 and 90 per cent of the cases of acute renal failure seen by physicians will fall into the categories of prerenal failure and acute tubular necrosis (those due to prostatic obstruction usually being managed by others). The term 'prerenal failure' is used when renal dysfunction is entirely attributable to hypoperfusion, and where restoration of renal perfusion leads to rapid recovery. The term 'acute tubular necrosis' does not find favour with all; although necrosis of tubular cells can usually be found by diligent examination, the lesion may be inconspicuous and the pathophysiological implications of such necrosis as might be seen remain uncertain. The glomeruli and vessels are usually normal. In common usage (retained here), the term 'acute tubular necrosis' describes a clinical entity comprising acute renal failure with three main characteristics:

1. it is seen in specific clinical contexts, frequently involving circulatory compromise and/or nephrotoxins;
2. urinary abnormalities usually suggest tubular dysfunction; and
3. essentially complete recovery of renal function is expected within days or weeks in most cases if the patient survives the precipitating insult, with a period of polyuria commonly following oliguria (but see the later section on prognosis).

The syndrome can be seen after virtually any episode of severe circulatory compromise, but not all causes of circulatory derangement are equally devastating to renal function. Primary impairment of cardiac performance, for example following myocardial infarction, may cause plasma creatinine to rise somewhat, but rarely causes renal failure of sufficient severity to require renal replacement therapy. By contrast, an apparently similar haemodynamic upset caused by sepsis frequently does, as demonstrated in Table 2 and Table 3. Multiple insults are the rule rather than the exception. Circumstances associated with a particularly high risk of acute renal failure include repair of a ruptured aortic aneurysm (20 per cent, as opposed to 3 per cent for elective repair), hepatobiliary surgery (10 per cent), pancreatitis (10 per cent), and burns (2 to 38 per cent, depending on the series).

Pathophysiology

The perfusion of the kidney seems to suffer more than that of any other organ when the circulation is compromised. In the face of modest underperfusion, the glomerular filtration rate is relatively preserved by a compensatory increase in the filtration fraction. This increase has repercussions on tubular function which, along with other factors, leads to the increased tubular reabsorption of sodium, water, and urea—a situation rapidly reversed by restoration of renal perfusion. However, following prolonged circulatory shock, renal function frequently deteriorates in a manner that is not immediately reversible, and it is not at all obvious why this should be so. Lack of a clear pathophysiological understanding has bedevilled all attempts at the development of rational therapy. Under normal conditions the kidney enjoys high blood flow, exceeded on a volume/weight basis only by the carotid body, and oxygen tension in the renal venous effluent is high, suggesting that oxygen supply greatly exceeds demand. Such a situation might be expected to confer protection from the effects of circulatory compromise, but no such benefit is observed: indeed the kidney appears to be more susceptible to damage than other organs. Acute renal failure resembling acute tubular necrosis can be produced in animal models by ischaemia, and the condition often arises clinically in the setting of profound haemodynamic disturbance, leading to the supposition that—despite apparently generous blood flow normally—renal ischaemia is the cause of renal failure in such circumstances. Two main hypotheses, not necessarily mutually exclusive, have been proposed to explain this. The first stresses that arteriovenous shunting of oxygen, resulting from the specialized anatomical arrangement of the vasa rectae that is essential for the countercurrent mechanism involved in urinary concentration and dilution, leads to the presence of areas of profound hypoxia within the normal kidney. These areas might therefore be operating on the verge of anoxia in the normal organ and hence be susceptible to ischaemic damage in response to a modest compromise of whole-organ blood flow. The second hypothesis is based on clinical and experimental evidence of intense constriction of renal vessels during shock, and suggests that very severe reduction in renal blood flow (perhaps only transient) may be responsible for the initiation of ischaemic damage. The justification for many of the interventions proposed in the management of patients at risk of acute renal failure, or with established acute renal failure, is that they might preserve renal blood flow and/or reduce renal oxygen consumption, thus rendering the development of ischaemic injury less likely.

Once damage to the kidney has been sustained, a variety of factors may be responsible for the persistence of excretory failure that is characteristic of the clinical

syndrome of acute tubular necrosis. There is evidence from experimental models and in humans that backleak of filtrate can occur from damaged tubules, but reduced renal blood flow and the prevention of fluid flow through tubules by internal blockage or external compression may also contribute to filtration failure. Even in experimental models it is very hard to determine what is happening at any time, and impossible to do so in clinical practice. However, the processes involved are beginning to be dissected, but matters are ferociously complicated and progress is slow. Many of the abnormalities have a structural as well as a functional basis, hence rapid reversal cannot be expected, there being good evidence that recovery from acute tubular necrosis depends upon cellular regeneration.

Diagnosis

The diagnosis of acute tubular necrosis is based on the clinical context, which often involves circulatory compromise, and the exclusion of obstruction or renal inflammatory conditions, usually by ultrasound examination of the kidneys and testing of the urine for blood and protein, respectively.

In prerenal failure the biochemical composition of the urine reflects the response of normal tubules to impaired renal perfusion. There is avid retention of sodium and water, leading to low urinary sodium and high urinary urea and creatinine concentrations, together with a high urinary osmolarity. Restoration of renal perfusion leads to rapid improvement in renal function. By contrast, conventional wisdom holds that in acute tubular necrosis the urinary sodium concentration is elevated and the urinary urea and creatinine concentrations and urinary osmolarity are relatively low, but this is not always so. Biochemical analysis of the urine is rarely useful in clinical practice, as explained in Table 6. From a practical point of view, treatment is begun on exactly the same lines whether the expected diagnosis is of prerenal failure or of acute tubular necrosis. The response to resuscitation retrospectively defines the diagnosis and determines further management.

Circumstances predisposing to prerenal failure are almost invariably associated with raised plasma levels of ADH. This acts on the collecting duct to increase the tubular reabsorption of both water and urea, hence the plasma concentration of urea rises out of proportion to that of creatinine in prerenal failure. Plasma urea may also appear to be disproportionately raised in the presence of sepsis, steroids, tetracycline (catabolic effect), and gastrointestinal haemorrhage (protein meal).

Avoidance

One of the main aims of the basic nursing and medical care provided to all acutely ill patients is to minimize the chances of the development of renal impairment. This can arise despite exemplary treatment, but poor care increases the likelihood. As described above, regular measurement of serum creatinine will permit early recognition of declining renal function, but is not of itself therapeutic. The best way to prevent the development of prerenal failure or acute tubular necrosis is to maintain an optimal intravascular volume (as described above, with further information given in Chapter 16.1), and to avoid or reduce exposure to nephrotoxic agents.

One common clinical situation worthy of specific note is the patient about to undergo a major elective surgical procedure such as repair of an abdominal aortic aneurysm. In the past the risk of acute renal failure following such an operation was substantial, but this has been considerably reduced by recognition of the importance of careful attention to fluid management—with the aim of avoiding episodes of hypovolaemia—both before, during, and after the procedure. It is good practice to maintain a diuresis, which can often be accomplished simply by infusion of crystalloid at moderate rate, since this appears to render the kidney less susceptible to insult. Although the routine use of diuretic agents is advocated by some, they would appear to have no specific advantages over a simple saline diuresis in protecting the kidney. Modest doses of diuretics (furosemide (frusemide) 40–80 mg, mannitol 25 g) given intravenously to a volume-replete patient undergoing a procedure that might compromise renal blood flow (for example, bile duct surgery, resection of aortic aneurysm, cardiac bypass) will increase urinary volume and may afford protection from acute renal failure. This is not proven, but the treatment should do no harm provided that the patient is not volume-depleted. However, the tendency of some to administer very large doses of diuretic agents should be restrained, since these can provoke massive diuresis (urine output >500 ml/h) and thereby lead to considerable difficulties in the control of electrolytes, especially potassium. A multicentre, randomized, double-blind, placebo-controlled trial of the use of dopamine in critically ill patients with evidence of early renal impairment did not show that this treatment was of any benefit.

For high-risk cases the insertion of a central venous pressure line preoperatively is a sensible precaution: the positioning of the patient for surgery and the presence of drapes may prevent proper intraoperative clinical assessment of cardiovascular status, and the risks of elective insertion of a central venous pressure line in the relative calm of the anaesthetic room are considerably less than those incurred if the attempt is made with the patient 'going off' on the operating table.

Clinical findings

There are no specific clinical features of prerenal failure or acute tubular necrosis. There may be symptoms of acute renal failure, as described previously, but these are also not specific and are rarely prominent, hence the clinical picture at presentation is likely to be dominated by signs of volume depletion and those of the precipitating condition.

If the patient does not die of acute renal failure, either because the degree of uraemia is modest or renal replacement therapy is provided, then renal recovery occurs in the vast majority of those who survive the precipitating insult. This may begin at any time from a few days to a few months (median 10–14 days) after the onset of acute renal failure, with a progressive increase in urinary volume typically preceding improvement in the plasma levels of creatinine and urea. Due to a relatively persistent defect in renal tubular sodium reabsorption and concentrating ability, a period of polyuria may ensue, placing the patient at risk of sodium and water depletion. Young patients can be expected to recover clinically normal renal function, but in those over 70 years of age recovery may be delayed, incomplete, and sometimes does not occur at all—leading to lifelong dependence on renal replacement therapy.

Specific treatment

The importance of effective treatment of the underlying condition and of rapid correction of hypovolaemia are above clinical dispute, although neither has been subject to controlled trial as regards the outcome of prerenal failure and acute tubular necrosis. Diuretic agents, in particular loop diuretics such as furosemide (frusemide), and/or 'renal dose' dopamine are often given to the patient who is thought to have acute tubular necrosis and whose urine output is inadequate. Decent trials are very thin on the ground, but there is no compelling evidence that any of these 'specific' remedies are helpful. In established acute tubular necrosis large doses of furosemide (0.5–2 g/day) may substantially increase urinary volume, and this can ease the management of fluid balance and reduce the degree of hyperkalaemia. However, such treatment is most unlikely to lead to improvement in the renal clearance of metabolic wastes, almost certainly does not alter the requirement for renal replacement therapy, and does not affect mortality. The evidence in favour of 'renal dose' dopamine (1–3 μg/kg per min) is generally weak and, although dopamine receptors certainly exist in the renal vasculature and on the renal tubules, it may well be that the effects that are sometimes observed clinically relate to an improvement in cardiac output rather than to any direct effect on the kidneys. However, many nephrologists have seen cases where the administration of a loop diuretic and dopamine has appeared to a have beneficial effect, hence it is not unreasonable to try such treatment, but not to prolong it if it is ineffective in a particular instance. Practical recommendations are given in Fig. 3.

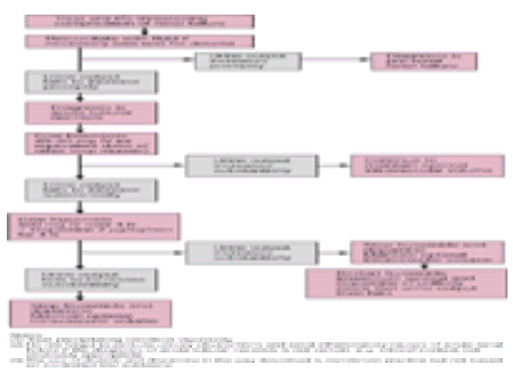

Fig. 3 An algorithm for the practical management of the patient with prerenal failure or acute tubular necrosis.

All other medical treatments should be regarded as experimental and not given except in the context of controlled trials. In experimental models the use of growth factors has been shown to speed renal recovery from acute tubular necrosis, and the possibility that such an approach might be applied clinically has generated the

greatest recent excitement. Unfortunately, the only substantial clinical trial published so far showed no evidence of benefit.

Prognosis

Complete recovery of renal function can be anticipated in those with acute tubular necrosis who survive the precipitating insult, excepting in the elderly (over 70 years) in whom there is a substantial chance (10–20 per cent) that dependence on dialysis will be lifelong.

Nephrotoxic causes of acute renal failure

Exogenous nephrotoxins

A wide variety of exogenous agents, including therapeutically prescribed drugs, can cause acute renal failure. Some of these agents are listed in Table 7. The following are worthy of particular note.

Aminoglycosides

Gentamicin, amikacin, kanamycin, and streptomycin are all potentially nephrotoxic, as are tobramycin and netilmicin to a lesser degree. These drugs are usually prescribed for patients thought to be suffering from potentially fatal infections, hence in clinical practice it is frequently impossible to separate with certainty the harmful effects of aminoglycosides from those of the underlying condition, or of other drugs used in treatment. However, evidence from animal models supports the view that these agents are genuinely nephrotoxic, rather than that their prescription is simply a marker for severe infection, which is itself a potent cause of acute renal failure.

The risk of nephrotoxicity is increased by old age, pre-existing renal insufficiency, high dosage, prolonged treatment, combined treatment with other nephrotoxic drugs, renal ischaemia, and volume depletion. It has been stated that acute renal failure complicates up to 25 per cent of therapeutic courses of gentamicin, even when monitoring optimally controls drug levels. Parenteral administration is not required for the development of toxicity: acute renal failure can occur as a result of systemic absorption when aminoglycosides are used in irrigating or bowel-sterilizing solutions. The typical clinical picture is of relatively mild non-oliguric renal failure coming on 1 to 2 weeks after starting treatment. Tubular proteinuria and impaired ability to concentrate the urine precede a loss of glomerular filtration rate. Proximal tubular damage involves the brush border, reflected by increased urinary excretion of g-glutamyl transferase, alanine aminopeptidase, and of lysosomal enzymes. Recovery may be slow, delayed, or incomplete.

The nephrotoxicity of particular aminoglycosides is related to the strength of their positive charge. They bind to negatively charged membrane phospholipids, particularly in the kidney to parts S1 and S2 of the proximal tubule, where they are delivered to megalin (the Heymann nephritis autoantigen, a member of the low-density lipoprotein (**LDL**) receptor family) in coated pits. The complex is endocytosed and trafficked to the endosome, where gentamicin inhibits fusion *in vivo* and *in vitro*. Polyaspartic acid polymers normalize fusion and ameliorate nephrotoxicity, suggesting that binding of other ligands to megalin may be useful in limiting aminoglycoside uptake and nephrotoxicity, but this possibility has not yet been explored clinically.

Aminoglycosides should only be used in the relatively uncommon circumstance that there is no suitable alternative antibiotic that is not nephrotoxic, and careful monitoring of levels is mandatory to avoid toxicity if gentamicin or similar agents must be used.

Radiographic contrast media

The incidence of acute renal failure associated with the use of radiographic contrast media has been reported to vary between 0 and 50 per cent. This extraordinary variability reflects differences in other risk factors in the populations under examination and in the definition of renal failure used. Recent prospective studies, using non-ionic contrast media and in which careful attention has been paid to the maintenance of adequate hydration, have shown a very low incidence of significant renal impairment—even in groups reported to be at high risk (diabetes, myeloma). When renal impairment does occur it is usually mild.

Endogenous nephrotoxins

Myoglobin

Myoglobinuric acute renal failure, the mechanism of which remains uncertain, is typically associated with crush injury to muscle, but there are a large number of causes of non-traumatic rhabdomyolysis (Table 8). A high index of suspicion is required to diagnose cases that are not obviously associated with muscle injury, since muscular pain, swelling, and tenderness may not be prominent features and can even be absent. The key to making the diagnosis is to detect myoglobin in the urine, or a very high level of enzymes released from muscle in the plasma. The former is recognized by the combination of dark-brown ('coca cola') urine that tests positive for 'blood' on a reagent strip, but which does not contain red cells on microscopy. The muscle enzyme usually measured in plasma is creatine kinase: the normal range of this is up to just below 200 U/l; in rhabdomyolysis values above 10 000 U/l are commonly seen, a value of only 1–2000 U/l not being enough to establish the diagnosis of rhabdomyolytic acute renal failure in the absence of other supporting evidence. Extremely high levels of plasma myoglobin, aldolase, and lactic dehydrogenase are also seen, all being released from damaged muscle.

Rhabdomyolysis can be associated with very high plasma levels of urate (>750 µmol/l), phosphate (>2.5 mmol/l), aspartate and alanine transaminase (**AST** in the many hundreds of U/l, exceptionally in the thousands; **ALT** in the few hundreds of U/l; respectively), and with an unusually low plasma calcium concentration (<1.5 mmol/l). Any of these findings should lead to serious consideration of rhabdomyolysis in any patient with unexplained acute renal failure.

If the diagnosis of rhabdomyolysis is made, then the question of whether to initiate an alkaline diuresis arises, since on theoretical grounds it would be anticipated that alkalinization of the urine would lead to enhanced excretion of the putative toxin and protect against acute renal failure. Victims of crush injury have been treated with infusion of very large volumes of fluid (12 litres/day) and high doses of mannitol (160 g/day) and bicarbonate (240 mmol/day). In comparison with historical (almost certainly volume-depleted) controls the incidence of renal failure has been impressively reduced, but the difficulties of controlling potassium balance in the face of such a massive diuresis should not be underestimated. It may well be that avoidance of hypovolaemia using a less aggressive and more easily managed fluid regimen would be equally efficacious.

Haemoglobin

In several situations, acute renal failure is seen in association with massive haemolysis: malaria, glucose-6-phosphate dehydrogenase deficiency, mismatched blood transfusion, arsine poisoning, copper sulphate poisoning, burns, and as a complication of bladder irrigation with hypotonic solutions. In each circumstance it is possible, but not proven, that the development of acute renal failure might be attributable to, or exacerbated by, the presence of large amounts of free haemoglobin within the circulation.

Urate (see also Chapter 20.10.5)

The tumour lysis syndrome is associated with a rapid rise in plasma uric acid concentration (and almost certainly liberation of other nephrotoxins) as a complication of the treatment of lymphoma, leukaemia, myeloma, or other 'high-turnover' tumours. This can result in the deposition of urate crystals in the distal tubule, which can both cause physical obstruction and initiate an inflammatory response, leading to acute renal failure in which freshly voided urine is heavily laden with urate crystals. Hyperuricaemia and renal failure have been described on rare occasions after recurrent epileptic seizures.

Hyperuricaemic acute renal failure is predictable and hence potentially avoidable in the context of the treatment of malignancy. The most important issue is that dehydration should be avoided at all costs, and a brisk saline diuresis should be initiated at least 24 h before the initiation of chemotherapy if possible. The use of an alkaline diuresis has been advocated, since uric acid is undoubtedly more soluble in alkaline urine, but this may encourage the precipitation of phosphate within the renal tubules and should not be employed if the serum phosphate is high. It is common practice in many centres to give allopurinol, even at a dosage above the usual 300 mg/day, but others would not do so on the grounds that this may encourage xanthine nephrotoxicity, although the risks of this seem to have been overstated.

If hyperuricaemic acute renal failure does develop, then it is unlikely that any of the treatments described above, or diuretics, will reverse the condition. Prompt improvement usually follows reduction of the plasma uric acid concentration, which is best accomplished by haemodialysis. On very rare occasions the ureters can become obstructed by urate crystals—indicated by colic, pelvicalyceal distension, or persistent oliguria—and ureteral catheterization and washout may be required.

Other endogenous nephrotoxins

More uncommon even than intratubular obstruction by urate crystals is similar obstruction by phosphate, also seen in the context of massive cell destruction in the treatment of malignant disease. Urinary alkalinization should be avoided because it may promote intratubular phosphate precipitation.

The possible role of immunoglobulin light chains in causing acute renal failure related to myeloma is discussed in Chapter 20.10.5.

Vascular causes of acute renal failure

Acute cortical necrosis

Acute cortical necrosis is an uncommon cause of acute renal failure, accounting for around 1 per cent of cases in the developed world, but more (3.8 per cent) in the experience of one large centre in the developing world (North India). However, these figures may be an underestimate, given that investigation is not pursued in many patients who fail to recover from what was presumed to be acute tubular necrosis, on the grounds that test results do not reliably predict prognosis or affect management, which is supportive.

Acute cortical necrosis presents in the same context as acute tubular necrosis, which is almost always the diagnosis made initially on clinical grounds. Suspicion should arise immediately if a patient without obstruction is anuric, as was found in 79 per cent of 113 patients in the largest study reported, but cortical necrosis is often considered only when renal function fails to improve.

Most cases of acute cortical necrosis are the result of obstetric disasters, particularly postpartum haemorrhage, abruptio placentae, eclampsia, or septic abortion. Snake bite, haemolytic uraemic syndrome, acute gastroenteritis, pancreatitis, septicaemia (often with disseminated intravascular coagulation), trauma, and drug-induced intravascular haemolysis are risk factors in the non-obstetric population.

The pathological findings are of microvascular thrombosis, mainly affecting interlobular arteries, arterioles, and glomeruli, with complete infarction of affected areas of cortex. The medulla and a rim of juxtamedullary tissue are spared.

The best investigations to establish the diagnosis of acute cortical necrosis are renal angiography and contrast-enhanced CT scanning. The former reveals attenuation of interlobular arteries, an increase in the subcapsular vessels, and a negative outer cortical nephrogram. The latter shows enhancement of the renal medulla, but no enhancement of the renal cortex and no excretion of contrast. Biopsy necessarily samples only a very small piece of tissue and may mislead because of the patchy nature of renal damage. Radiopharmaceutical investigations that depend upon renal excretion (for example, **DMSA** (dimercaptosuccinic acid) scans) are unhelpful in patients with very poor renal function.

In the months or years after an episode of acute cortical necrosis, the kidneys tend to contract: cortical calcification, producing an eggshell or tramline appearance on the abdominal radiograph, is a characteristic sequel, but this is not useful in making the diagnosis acutely.

Return of renal function in cases of acute cortical necrosis occurs very slowly, if at all, and is attributable to the survival of islands of intact cortical tissue. About 50 per cent of patients recover sufficiently to come off dialysis, but the glomerular filtration rate rarely exceeds 10 to 20 ml/min. Hypertension (including accelerated phase) may be a major problem, and a subsequent decline in renal function with the necessity for a return to dialysis/transplantation is not uncommon.

Large-vessel obstruction

Arterial obstruction

Occlusion of the main renal arteries—or of the artery supplying a solitary functioning kidney—by trauma, dissection, thrombosis, or embolism may rarely be the reason for acute renal failure. Loin pain sometimes occurs, and there is usually a low-grade fever, such that the clinical picture may mimic acute pyelonephritis, but symptoms can be notable by their absence. Proteinuria and haematuria may occur.

Diagnosis is important because thrombolysis and/or renovascular surgery can be surprisingly effective in restoring function, even when undertaken a considerable time after arterial occlusion (up to many weeks), in those with atherosclerotic renovascular disease in whom (prior to occlusion) a collateral blood supply to the renal parenchyma has developed. Suspicion should be aroused by complete, sudden anuria in the absence of urinary obstruction, especially if the clinical setting is appropriate, for example atrial fibrillation in an arteriopath. A useful pointer to the diagnosis is the finding of a urinary sodium concentration similar to that of plasma (see Table 6), but DTPA renography and renal angiography are the appropriate diagnostic tests if the diagnosis of renal artery occlusion is suspected. CT scanning may reveal wedge-shaped infarcts when occlusion is incomplete.

Venous obstruction

Renal vein thrombosis can cause acute renal failure, most commonly in adults as a complication of the nephrotic syndrome, but in infants and children as a result of abdominal sepsis or severe dehydration. Renal pain is common, as is increasing proteinuria and haematuria (which can be macroscopic), but there may be no symptoms. If there is clinical suspicion of the diagnosis, for example unexplained deterioration of renal function in a nephrotic patient, then appropriate investigation includes ultrasound/Doppler examination of the renal veins and inferior vena cava, CT/MRI scanning, or renal arteriography with late films taken specifically to look for filling of the renal veins. Treatment by anticoagulation is the usual practice. (See Section 20.3 for further discussion.)

Small-vessel obstruction

Accelerated-phase hypertension (see also Chapter 15.16.3)

'Accelerated-phase' hypertension (a term preferred to 'malignant' hypertension because the implication of malignancy is terrifying for patients) occurs when the blood pressure is elevated sufficiently to cause fibrinoid necrosis of blood vessels, leading to the development of haemorrhages and exudates in the ocular fundi. It may develop as a consequence of pre-existing renal disease, but does not always do so, and is itself a potent cause of renal damage. Acute renal failure is a common complication in those with previously normal renal function, and is associated with proteinuria, haematuria, and the presence of urinary red cell casts. The higher the creatinine at presentation, the poorer the prognosis for both patient survival and renal outcome: in one study only 9 per cent of those with an initial plasma creatinine below 300 µmol/l progressed to need renal replacement therapy, compared with two-thirds of those with a plasma creatinine above this level. The ability of the kidney to autoregulate perfusion is disturbed in accelerated-phase hypertension, hence the therapeutic lowering of arterial pressure may be associated with reduced renal perfusion and an abrupt decline in renal function. Accelerated-phase hypertension is one of the conditions in which renal function sometimes recovers after a lengthy period on dialysis. Renal failure was the cause of two-thirds of the deaths in patients with accelerated-phase hypertension in the days before dialysis was available.

Systemic sclerosis (see also Chapter 18.10.3)

This disease does not usually involve the kidney, but a syndrome resembling accelerated-phase hypertension and termed 'scleroderma renal crisis' is well recognized in patients with diffuse cutaneous systemic sclerosis. It usually occurs within the first 5 years of the disease, may be the presenting feature, and often appears during the winter months. Rapid worsening of skin manifestations may precede the crisis, but frequently there is no warning. The patient may develop headaches, visual disturbance, and convulsions. Arterial pressure is usually grossly elevated, but the renal syndrome can occur without a rise in arterial pressure. Haemorrhages and exudates are often seen in the ocular fundi. Renal failure, with proteinuria and haematuria, develops rapidly. A microangiopathic haemolytic anaemia may complicate the situation. Plasma levels of renin are grossly elevated. There have been a number of case reports of arrest or reversal of the syndrome after treatment with

angiotensin-converting enzyme inhibitors or nifedipine. These agents should be tried, but more in hope than expectation that they will prevent relentless progression to endstage renal failure.

Glomerulonephritic and vasculitic causes of acute renal failure

A large number of glomerulonephritic and vasculitic diseases can cause acute renal failure, sometimes in association with pulmonary haemorrhage (see Table 1 of Chapter 20.10.3). These are discussed in detail in the relevant subsections of Section 20.7, and in Chapter 20.10.3 and Chapter 20.10.4. Together they form only 5 to 10 per cent of cases of acute renal failure, but making the correct diagnosis is of extreme importance because of the management implications. Regrettably, most nephrologists have seen cases where the diagnosis has been much delayed because renal impairment has incorrectly been attributed to acute tubular necrosis, and infiltrates on the chest radiograph to oedema or infection. This error, which can be catastrophic, should be avoided in patients in whom the cause of acute renal failure is not obvious, by:

1. A history and examination specifically directed towards determining whether one of the conditions listed in Table 1 of Chapter 20.10.3 might be present.
2. Microscopy of the urine to look for the presence of red cells and red cell casts.
3. The following blood tests:
   a. measurement of antiglomerular basement membrane (**anti-GBM**) antibodies—positive in Goodpasture's disease (see Chapter 20.7.9);
   b. measurement of antineutrophil cytoplasmic antigen antibodies (**ANCA**)(screening by indirect immunofluorescence test, specific tests for antiproteinase-3 and antimyeloperoxidase antibodies)—positive in microscopic polyangiitis and Wegener's granulomatosis (see Chapter 20.10.3);
   c. estimation of serum complement levels (C3 is depressed in postinfectious glomerulonephritis, mesangiocapillary glomerulonephritis, systemic lupus erythematosus) (see Chapter 20.7.7, Chapter 20.7.8, and Chapter 20.10.4);
   d. measurement of anti-streptolysin O titre (**ASOT**—elevated in poststreptococcal glomerulonephritis) (see Chapter 20.7.7);
   e. serological tests for systemic lupus erythematosus (see Chapter 20.10.4);
   f. cryoglobulins (see Chapter 20.10.5) (tests of serum immunoglobulins and for urinary light chains should also be performed—see below).
4. Considering the possibility that pulmonary infiltrates in a patient with acute renal failure might be due to haemorrhage. The chances of this are increased if there is a history of haemoptysis (associated with several forms of rapidly progressive glomerulonephritis), nasal discharge, or bleeding (associated with Wegener's granulomatosis), or if anaemia is unusually profound and otherwise unexplained. Lung function tests demonstrating an increase in carbon monoxide transfer factor can establish the diagnosis.
5. Performing an urgent renal biopsy. In any patient with acute renal failure and an active urinary sediment, renal biopsy should be performed unless the diagnosis is clear (for example, a classical history of poststreptococcal nephritis, obvious infective endocarditis/shunt nephritis) or there is a strong contraindication, for example a single kidney or serious bleeding disorder.

The possibility of the presence of a rapidly progressive glomerulonephritis/vasculitis constitutes a medical emergency. Anti-GBM disease responds well to immunosuppression with plasma exchange, steroids and cyclophosphamide, but only if treatment is begun before dialysis is required. Immunosuppressive treatment should be given as early as possible in the course of acute renal failure complicating microscopic polyangiitis/idiopathic rapidly progressive (crescentic) glomerulonephritis, Wegener's granulomatosis, and systemic lupus erythematosus. The urgency is such that it may well be appropriate to start these treatments while the results of blood tests and renal biopsy are awaited, and to stop them if the findings do not corroborate the initial clinical diagnosis. The management of these patients is complex and patients benefit from the judgement and expertise of specialists.

Interstitial nephritis as a cause of acute renal failure (see also Chapter 20.9.1)

Drugs

Drugs are the commonest cause of acute interstitial nephritis, the usual culprits being penicillins, non-steroidal anti-inflammatory drugs (**NSAIDs**), and diuretics, but many others have been implicated (see Table 1 of Chapter 20.9.1). The classical clinical picture is that a few days or weeks after taking a drug the patient develops flank pain (sometimes), fever, a skin rash, arthralgias, haematuria, blood eosinophilia and elevated IgE, disturbed liver function (sometimes), interstitial pneumonia (rarely), and renal impairment, but renal failure may be the only manifestation. The urine contains protein and blood, with white and red cell casts. Proteinuria may be in the nephrotic range, particularly in association with NSAIDs. The diagnosis can only be established with certainty by renal biopsy, where typical histological findings are of an interstitial infiltrate of lymphocytes and monocytes/macrophages, together with some eosinophils. Epithelioid granulomas may be seen, strongly supporting the diagnosis of a drug-induced interstitial nephritis, but they are not pathognomonic. When large numbers of cells are present in the renal interstitium the diagnosis is not contentious: more difficult are those cases (not too infrequent) in which the infiltrate is modest—how many cells turn 'acute tubular necrosis' into 'interstitial nephritis'? The importance of making the distinction lies in the belief that, apart from withdrawal of the offending drug, treatment of drug-induced interstitial nephritis with steroids is beneficial. There is some evidence that those given prednisolone (typical dosage 20 to 60 mg/day) have an earlier and more complete recovery of renal function than those left untreated.

Leptospirosis (see also Chapter 7.11.31)

Acute renal failure due to an interstitial nephritis may appear within a few days of the onset of disease, but more commonly in the second week. It occurs in about 10 per cent of cases of leptospirosis and is frequently mild, but may be severe, with the plasma urea level rising rapidly due to hypercatabolism. The diagnosis of leptospirosis should be considered in any patient with unexplained acute renal failure who has myalgias/muscle tenderness, conjunctival infection, and/or haemorrhage or jaundice. Direct enquiry must be made as to whether any such patient has been exposed to rats.

Aside from renal impairment, blood tests commonly reveal a dramatic conjugated hyperbilirubinaemia (often >250 µmol/l) and thrombocytopenia (seen in 40 per cent of cases). There may also be elevation of serum creatine kinase and a slight increase in serum AST. Anaemia may be severe due to intravascular haemolysis. By contrast to most other causes of acute renal failure, serum potassium is often normal or low in cases of leptospirosis. Mild abnormalities of blood clotting tests can be seen, but disseminated intravascular coagulation is not a feature, which is an important point in its distinction from bacterial septicaemia.

The diagnosis is established by culture of *Leptospira* spp. (from blood during the first phase or urine afterwards) or positive serology. Doxycycline prophylaxis is effective at preventing leptospirosis, but antibiotics are not of proven benefit in treating disease. Mild cases are self-limiting; most physicians treat patients who are symptomatic with a 7-day course of oral doxycycline or intravenous benzylpenicillin, on the grounds that this appears to shorten the duration of fever and leptospiruria.

Hantavirus disease (see also Chapter 7.10.15)

In Europe

In Europe the Puumula serotype of hantavirus produces an illness that can have many similarities to that produced by leptospirosis, although serological studies indicate that many patients must have a subclinical infection. In those that are symptomatic, high fever is typically followed within a couple of days by loin/abdominal pain and often by nausea and vomiting; photophobia and signs of meningeal irritation can also occur. Acute renal failure follows when these symptoms have settled and is associated with conjunctival haemorrhage (20 per cent), proteinuria (almost 100 per cent of cases), microscopic haematuria (70 per cent), thrombocytopenia (50 per cent), and a transient mild rise in serum liver enzymes. There may be a small increase in serum bilirubin (maximum 40 µmol/l). Mild abnormalities of blood clotting tests are seen, but disseminated intravascular coagulation is rare.

Renal biopsy, performed for the indication of unexplained acute renal impairment, shows interstitial nephritis. This has no pathognomonic features, leading in this clinical context to the differential diagnosis of leptospirosis and sometimes (depending on exposure) disease induced by NSAIDs. Leptospirosis is much more likely if the serum bilirubin is markedly elevated. NSAID-induced disease does not cause conjunctival haemorrhages or thrombocytopenia. The diagnosis of Puumula hantavirus infection is made on the basis of serological evidence. Prognosis is good: no deaths have been reported and renal function returns to normal.

In some areas of Eastern and Central Europe there is a more severe form of hantavirus infection, which is similar to that seen in Asia.

In Asia

The Hantaan and Seoul viruses cause hantavirus disease in Asia: the former causes more severe illness, but both are considerably more dangerous that the Puumula hantavirus seen in Europe. A total of five phases of disease are recognized, comprising:

1. high fever and myalgias, followed by headache and severe abdominal/loin pain, often with an erythematous rash that may become petechial, also conjunctival haemorrhages;
2. severe hypotension;
3. gradual recovery of blood pressure, but associated with oliguria and renal failure with proteinuria and microscopic haematuria—one-third of patients in this stage have major problems with bleeding: gastrointestinal, intracerebral or massive purpura (hence the terms epidemic or Korean haemorrhagic fever);
4. presence of polyuria;
5. convalescence.

Differential diagnosis is from severe leptospirosis and other causes of haemorrhagic fever found in Asia, including dengue and murine typhus. The diagnosis is made serologically. Treatment is supportive. Mortality is between 3 and 7 per cent; survivors recover completely.

'Haematological' causes of acute renal failure

Haemolytic uraemic syndrome and idiopathic postpartum renal failure (see also Chapter 13.5 and Chapter 20.10.6)

The haemolytic uraemic syndrome (**HUS**) is a condition, or group of conditions, in which acute renal failure, characterized on biopsy by thrombosis and necrosis of intrarenal vessels, occurs together with thrombocytopenia, haemolytic anaemia, and red cell fragmentation. (See Chapter 20.10.6 for further information.) A similar picture developing immediately (or up to several weeks) after an entirely uneventful pregnancy and delivery is termed 'idiopathic postpartum acute renal failure'.

Myeloma (see also Chapter 20.10.5)

Acute renal failure complicates about 7 per cent of cases of myeloma, often being the presenting feature, and subacute progressive renal failure is even commoner, affecting 14 to 61 per cent of cases. The cause of renal failure is often multifactorial, with varying contribution from the reversible factors of dehydration, infection, hypercalcaemia, and hyperuricaemia, and with renal damage caused by free immunoglobulin light chains. The reason why some patients with myeloma develop renal failure and others do not remains a mystery. There has been much speculation as to whether variation in the isoelectric point of light chains, and hence their capacity for reabsorption by the renal tubules, might be responsible. However, individual patients with light chains of very similar physicochemical properties can present totally different clinical pictures, varying from no perceptible renal involvement to irreversible renal failure.

In a patient with acute renal failure, a history of bone pain, the findings of clumping of erythrocytes on the blood film, or of gross and unexpected elevation of the erythrocyte sedimentation rate, are clues that myeloma might be the underlying diagnosis. Such clues may be absent when excess production of monoclonal light chains is the predominant problem, hence all patients with unexplained acute or subacute renal failure should undergo investigation both of serum for a monoclonal immunoglobulin component (with immunoparesis) and of their urine (if available) for free _ or l light chains. The renal biopsy appearances are of tubulointerstitial nephritis, with fractured casts in the tubular lumina, tubular atrophy, interstitial oedema/fibrosis, and an interstitial infiltrate that may contain multinucleate giant cells. However, the definitive test for myeloma is a bone marrow biopsy for immunochemical analysis of the plasma-cell population, and this should be performed whenever myeloma is a likely or possible cause of acute renal failure.

The first priority in management is to deal promptly with those factors that can be reversed— dehydration, infection, hypercalcaemia, and hyperuricaemia. Volume resuscitation should be given as described in Chapter 16.1, along with broad-spectrum antimicrobials (after appropriate cultures have been taken) if there is any suspicion of infection. After the intravascular volume has been restored, then (assuming adequate urine output) hypercalcaemia can be treated rapidly and effectively using a two-pronged approach: a diuresis provoked by infusion of 0.9 per cent saline (1 litre every 4–6 h) and furosemide (40 mg as necessary), and intravenous bisphosphonate (for example, disodium pamidronate, 15–60 mg as a single dose, maximum of 90 mg over 2–4 days). It has been suggested that alkalinization of the urine using intravenous sodium bicarbonate may be advantageous in promoting light-chain excretion, but it is unclear whether this is better than adequate rehydration with saline alone.

If there is a clear precipitant for the decline in renal function, then the prospects for renal recovery in patients with myeloma are good; if not, then the renal outlook is less favourable. Although some report that aggressive treatment with cytotoxic agents and/or plasmapheresis can restore renal function in such cases, this is not everyone's experience, and renal recovery seems to be the exception rather than the rule.

The prognosis for patients with myeloma and established renal failure requiring dialysis is poor: 50 per cent 1-year survival, 30 per cent at 2 years. However, many patients will have few symptoms from their myeloma, excepting renal failure, and these patients should certainly be offered the opportunity of renal replacement therapy. In those with considerable extrarenal manifestations the situation is much more difficult, and it may not be appropriate or kind in such circumstances to offer aggressive haematological regimens, producing considerable side-effects, and/or dialysis. The decisions to be made are rarely straightforward: they will substantially depend on an assessment of the overall burden to the patient of their disease and a realistic appraisal of what benefits treatment might produce.

Hepatorenal syndrome

The hepatorenal syndrome consists of the association of severe and usually progressive liver disease with acute renal failure. The renal failure is characterized by:

1. no evidence of renal parenchymal damage (when kidneys from patients with the hepatorenal syndrome have been transplanted, they function normally in the recipient);
2. characteristic 'prerenal' urine biochemistry, in particular a very low urinary sodium concentration (<10 mmol/l) (Table 6);
3. no sustained response to volume expansion; and
4. exclusion of other causes of acute renal failure.

The mechanism of renal failure is uncertain, but is associated with markedly reduced renal perfusion that may be due to excessive action of the vasoconstrictor endothelin.

One of the aims of the general management of patients with liver disease is prevention of the hepatorenal syndrome, the most important consideration being avoidance of known precipitants (drugs, excessive diuresis, delay in the treatment of sepsis). Nevertheless, the syndrome develops in up to 20 per cent of patients with cirrhosis admitted to hospital. There is no specific treatment and the prognosis is extremely poor. In the presence of potentially reversible liver disease, or with the prospect of liver transplantation, intensive therapy and renal replacement therapy are justified. If these criteria are not met, then aggressive support is almost certainly inappropriate.

Tropical

Acute renal failure in the developing world, as elsewhere, is usually a consequence of acute tubular necrosis. The causes of hospital-acquired acute renal failure are the same as in the developed world, with nephrotoxic drugs, major surgery, and hospital-acquired sepsis the dominant factors. By contrast, the causes of community-acquired acute renal failure are very different.

The Chandigarh study showed that 30 years ago diarrhoeal disease and obstetric complications each accounted for about 25 per cent of cases of acute renal failure in North India, but more recently each has accounted for about 10 per cent. Infections that often cause acute renal failure in the developing world include falciparum malaria, leptospirosis, melioidosis, cholera, salmonellosis, and shigellosis. Intravascular haemolysis is a common feature of many cases, being found in over 20 per cent of 325 patients receiving dialysis for acute renal failure in Chandigarh. This was most frequently seen in those with glucose-6-phosphate dehydrogenase deficiency, with copper sulphate poisoning and snake bite the next commonest causes. Poisoning by deliberate (occasionally accidental) ingestion of paraquat (herbicide) is not uncommon in agricultural communities: aside from renal failure this can lead to inexorably progressive respiratory failure with an extremely high mortality. Treatment with corticosteroids and cyclophosphamide has been used in an attempt to prevent pulmonary fibrosis and may be of benefit. Heatstroke can

cause acute renal failure.

Snake bite

Acute renal failure develops in 5 to 30 per cent of the victims of severe viper poisoning and is the cause of between 2 and 3 per cent of cases of acute renal failure, but a very much higher proportion in some centres at some times of the year. It develops from a few hours to 72 h following the bite, and is non-oliguric in 50 per cent of cases. Hyperkalaemia may be prominent in bites associated with myonecrosis, such as those of sea-snakes. The usual renal pathology is acute tubular necrosis, but acute cortical necrosis can occur (see above). Renal management is supportive. (See Chapter 8.2 for further discussion.)

Copper sulphate poisoning

Copper sulphate is extensively used in the leather industry and is a relatively common (although decreasing) cause of poisoning in India. Symptoms include nausea, vomiting, diarrhoea, epigastric pain, gastrointestinal bleeding, and coma. Signs include jaundice, hypotension, and shock. Investigation reveals intravascular haemolysis, methaemoglobinaemia, haemoglobinuria, haematuria, and renal failure, the latter complicating 11 of 29 cases in one series.

Histological examination of the kidneys shows acute tubular necrosis with luminal haemoglobin casts, rupture of tubular basement membranes, and copper in degenerated tubules.

Aside from gastric lavage, management usually involves the administration (urine output permitting) of intravenous 0.9 per cent saline, mannitol, and/or diuretics to encourage copper elimination, intramuscular dimercaprol (but with extreme caution if renal failure has developed), and dialysis.

Further reading

Abassi ZA, *et al*. (1998). Acute renal failure complicating muscle crush injury. *Seminars in Nephrology* **18**, 558–65.

Ash SR (2001). Peritoneal dialysis in acute renal failure of adults: the safe, effective, and low-cost modality. *Contributions to Nephrology* **132**, 210–21.

Barton IK, *et al*. (1993). Acute renal failure treated by haemofiltration: factors affecting outcome. *Quarterly Journal of Medicine* **86**, 81–90.

Bellomo R, *et al*. (2000). Low-dose dopamine in patients with early renal dysfunction: a placebo-controlled randomised trial. Australian and New Zealand Intensive Care Society (ANZICS) Clinical Trials Group. *Lancet* **356**, 2139–43.

Better OS, Stein JH (1990). Early management of shock and prophylaxis of acute renal failure in traumatic rhabdomyolysis. *New England Journal of Medicine* **322**, 825–9.

Bhandari S, Turney JH (1996). Survivors of acute renal failure who do not recover renal function. *Quarterly Journal of Medicine* **89**, 415–21.

Brosius FC, Lau K (1986). Low fractional excretion of sodium in acute renal failure: role of timing of the test and ischemia. *American Journal of Nephrology* **6**, 450–7.

Chugh KS (1989). Snake-bite-induced acute renal failure in India. *Kidney International* **35**, 891–907.

Chugh KS, *et al*. (1977). Acute renal failure following copper sulphate intoxication. *Postgraduate Medical Journal* **53**, 18–23.

Chugh KS, *et al*. (1977). Acute renal failure due to intravascular hemolysis in the North Indian patients. *American Journal of the Medical Sciences* **274**, 139–46.

Chugh KS, *et al*. (1989). Changing trends in acute renal failure in third-world countries—Chandigarh study. *Quarterly Journal of Medicine* **73**, 1117–23.

Chugh KS, *et al*. (1994). Acute renal cortical necrosis—a study of 113 patients. *Renal Failure* **16**, 37–47.

Cramer BC, *et al*. (1985). Renal function following infusion of radiologic contrast material. A prospective controlled study. *Archives of Internal Medicine* **145**, 87–9.

Denton MD, *et al*. (1996). 'Renal-dose' dopamine for the treatment of acute renal failure: scientific rationale, experimental studies and clinical trials. *Kidney International* **50**, 4–14.

Feest TG, *et al*. (1993). Incidence of severe acute renal failure in adults: results of a community based study. *British Medical Journal* **306**, 481–3.

Firth JD (1996). Acute irreversible renal failure. *Quarterly Journal of Medicine* **89**, 397–9.

Gines P, Arroyo V (1999). Hepatorenal syndrome. *Journal of the American Society of Nephrology* **10**, 1833–9.

Hirschberg R, *et al*. (1999). Multicenter clinical trial of recombinant human insulin-like growth factor I in patients with acute renal failure. *Kidney International* **55**, 2423–32.

Holt SG, Moore KP (2001). Pathogenesis and treatment of renal dysfunction in rhabdomyolysis. *Intensive Care Medicine* **27**, 803–11.

Hou SH, *et al*. (1983). Hospital-acquired renal insufficiency: a prospective study. *American Journal of Medicine* **74**, 243–8.

Jha V, *et al*. (1992). Spectrum of hospital-acquired acute renal failure in the developing countries—Chandigarh study. *Quarterly Journal of Medicine* **83**, 497–505.

Kleinknecht D, *et al*. (1973). Diagnostic procedures and long-term prognosis in bilateral renal cortical necrosis. *Kidney International* **4**, 390–400.

Leverve X, Barnoud D (1998). Stress metabolism and nutritional support in acute renal failure. *Kidney International Supplement* **66**, S62–6.

Levy M (1993). Hepatorenal syndrome. *Kidney International* **43**, 737–53.

Liano F, *et al*. (1994). Use of urinary parameters in the diagnosis of total acute renal artery occlusion. *Nephron* **66**, 170–5.

Lieberthal W, Nigam SK (1998). Acute renal failure. I. Relative importance of proximal vs. distal tubular injury. *American Journal of Physiology* **275**, F623–31.

Lieberthal W, Nigam SK (2000). Acute renal failure. II. Experimental models of acute renal failure: imperfect but indispensable. *American Journal of Physiology* **278**, F1–F12.

Lindner A (1983). Synergism of dopamine and furosemide in diuretic-resistant, oliguric acute renal failure. *Nephron* **33**, 121–6.

Maillet PJ, *et al*. (1986). Nondilated obstructive acute renal failure: diagnostic procedures and therapeutic management. *Radiology* **160**, 659–62.

Milligan SL, *et al*. (1978). Intra-abdominal infection and acute renal failure. *Archives of Surgery* **113**, 467–72.

Molitoris BA (1997). Cell biology of aminoglycoside nephrotoxicity: newer aspects. *Current Opinion in Nephrology and Hypertension* **6**, 384–8.

Murphy SW, *et al*. (2000). Contrast nephropathy. *Journal of the American Society of Nephrology* **11**, 177–82.

Nigame S, Lieberthal W (2000). Acute renal failure. III. The role of growth factors in the process of renal regeneration and repair. *American Journal of Physiology* **279**, F3–F11.

Parfrey PS, *et al*. (1989). Contrast material-induced renal failure in patients with diabetes mellitus, renal insufficiency, or both. A prospective controlled study. *New England Journal of Medicine* **320**, 143–9.

Pilmore HL, *et al*. (1995). Acute bilateral renal artery occlusion: successful revascularization with streptokinase. *American Journal of Nephrology* **15**, 90–1.

Ramsay AG, *et al*. (1983). Renal functional recovery 47 days after renal artery occlusion. *American Journal of Nephrology* **3**, 325–8.

Rasmussen HH, Ibels LS (1982). Acute renal failure. Multivariate analysis of causes and risk factors. *American Journal of Medicine* **73**, 211–18.

Remuzzi G (1988). Bleeding in renal failure. *Lancet* **1**, 1205–8.

Ronco C, *et al*. (2000). Effects of different doses in continuous veno-venous haemofiltration on outcomes of acute renal failure: a prospective randomised trial. *Lancet* **356**, 26–30.

Schiffer H, *et al*.(2002). Daily hemodialysis and the outcome of acute renal failure. *New England Journal of Medicine* **346**, 305–10.

Shilliday IR, *et al*. (1997). Loop diuretics in the management of acute renal failure: a prospective, double-blind, placebo-controlled, randomized study. *Nephrology, Dialysis, Transplantation* **12**, 2592–6.

Solez K, Racusen LC (2001). Role of the renal biopsy in acute renal failure. *Contributions to Nephrology* **132**, 68–75.

Solez K, *et al*. (1979). The morphology of 'acute tubular necrosis' in man: analysis of 57 renal biopsies and a comparison with the glycerol model. *Medicine (Baltimore)* **58**, 362–76.

Spital A, *et al*. (1988). Nondilated obstructive uropathy. *Urology* **31**, 478–82.

Sponsel H, Conger JD (1995). Is parenteral nutrition therapy of value in acute renal failure patients? *American Journal of Kidney Diseases* **25**, 96–102.

van Ypersele de Strihou C (1998). Hantavirus infection. In: Davison AM, *et al.*, eds. *Oxford textbook of clinical nephrology*, pp 1688–92. Oxford University Press, Oxford.

van Ypersele de Strihou C, Mery JP (1989). Hantavirus-related acute interstitial nephritis in western Europe. Expansion of a world-wide zoonosis. *Quarterly Journal of Medicine* **73**, 941–50.

Winearls CG, *et al*. (1984). Acute renal failure due to leptospirosis: clinical features and outcome in six cases. *Quarterly Journal of Medicine* **53**, 487–95.

20.5.1 Chronic renal failure

C. G. Winearls

Introduction

Chronic renal failure is the clinical syndrome of the metabolic and systemic consequences of a gradual, substantial, and irreversible reduction in the excretory and homeostatic functions of the kidneys. It can be difficult to recognize because the symptoms and clinical manifestations are non-specific. However, if suspected, it is easily diagnosed by simple biochemical measurements. Early and specific diagnosis is worthwhile because this allows the application of effective treatments of both cause (in some cases) and consequences (in all cases), and substitution treatments are readily available for complete kidney failure. An understanding of chronic renal failure is important for doctors in both general and specialty practice: they will have to accommodate its consequences in their own work, for example in surgery, obstetrics, and in prescribing, and they will need to refer patients appropriately for specialist investigation and supervision.

For the patients who suffer it, chronic renal failure is an ever-increasing burden that they carry for the rest of their lives. Eventually, when end-stage renal failure (**ESRF**) is reached, they embark on a career of substitution treatments (dialysis and renal transplantation) referred to as renal replacement therapy. The illness and these treatments intrude on every aspect of their lives—physical, social, vocational, and emotional.

Definition

Chronic renal failure is defined as the state resulting from a permanent (and usually progressive) reduction in renal function, sufficient to have adverse consequences on other systems. The threshold at which these develop is at around 40 per cent of normal excretory capacity. A reduction in renal function below the third centile for age and gender (for example, by removing one of a pair of normal kidneys or limited damage to one or both), does not amount to chronic renal failure: although there is a loss of renal reserve, there are no clinical consequences.

The severity of chronic renal failure is graded according to the fraction of kidney function remaining (Table 1). These are useful descriptions because they give an indication of the likely symptoms and complications that should be anticipated, and provide cues to the key steps in both immediate and future management. Accurate measurement of glomerular filtration rate is not necessary to categorize the degree of renal failure: knowledge of the patient's age, gender, and body-weight applied to the Cockcroft–Gault formula:

$$\text{glomerular filtration rate} = [\,(140 - \text{age in years}) \times \text{weight (kg)}\,] / \text{plasma creatinine } (\mu\text{mol/l}) \times 0.82 \text{ (subtract 15 per cent for females)}$$

will give an acceptable estimate (see Chapter 20.4 for further discussion). However, this grading of the severity of renal failure has limitations: as ESRF approaches there is a poor correlation between the actual glomerular filtration rate and symptoms which are caused by downstream consequences of that reduction: for example, breathlessness by pulmonary oedema or acidosis, fatigue by anaemia, muscle weakness by abnormalities in calcium and phosphate. The decision when to start dialysis should be made after integrating knowledge of the estimated glomerular filtration rate, symptoms, and recognition of complications. Decisions should not be based on estimates of plasma creatinine or urea.

Incidence and prevalence

The only accurate data on the incidence of chronic renal failure is for ESRF that is treated with renal replacement therapy, that is to say the number of patients starting such treatment. This is available from national and other databases, which also provide data on the prevalence of patients receiving renal replacement treatment (Table 2).

The incidence or prevalence of ESRF that is not treated with renal replacement therapy is unknown, nor is such information available for lesser degrees of chronic renal failure (mild, moderate, and severe): many patients are undiagnosed or not yet referred. This is illustrated by the fact that as many as 30 per cent of new patients starting renal replacement therapy meet a nephrologist for the first time less than 3 months before dialysis is begun. Crude estimates suggest there are about

500 to 1000 patients per million population with significant chronic renal impairment. Many of these patients will never require renal replacement therapy and will die with renal failure and not of it, examples include patients with malignant urinary tract obstruction, diabetes mellitus, renovascular disease with widespread cerebro- and cardiovascular disease.

From a workload point of view there are about 500 to 600 per million patients on renal replacement therapy in the United Kingdom and Australia, and over 1000 per million in the United States and Japan. This means that an average British general practitioner with 2000 registered patients will have only one dialysis or transplant patient on their list, and will see one new ESRF patient every 5 years. They will, however, be involved in the care of one or two patients with chronic renal failure. Renal failure is, when compared to chronic cardiovascular and respiratory diseases, a relatively small part of the general practitioner's workload.

The epidemiological study of renal failure has revealed some stark facts. It is a disease of the elderly: the incidence in a population over 75 years of age is 10 times higher at 400 per million population (**pmp**) than it is in those under 40 years of age; 50 per cent of patients now starting on renal replacement therapy are over 65 years of age. The incidence is higher in males (1.3:1), in areas of social deprivation, and in particular ethnic groups. In the United Kingdom it is 3.5 times higher in citizens of Asian or Afro-Caribbean backgrounds. In 1997 in Australia the incidence in Aboriginals was 435 pmp, which was six to seven times higher than in Caucasoids at 68 patients pmp. In New Zealand the incidence in Maoris is three to four times higher than in Caucasoids. These ethnic variations (which are related to the higher prevalence of diabetes and hypertension) can account only for part of the huge difference in the incidence of ESRF between Europe and the United States. There is still no easy explanation for the higher number of ESRF patients in Caucasian Americans than in the same age groups in Europe. Acceptance criteria for renal replacement therapy are much more stringent in the United Kingdom than in the United States, but although patients with terminal illness, dementia, or severe comorbid conditions are not usually started on dialysis, few patients who would benefit (that is, those who would survive independently for longer than 12 months) are excluded.

The prevalence of treated ESRF, in other words the number of patients receiving dialysis or a life-sustaining renal transplant, varies according to the capacity and availability of renal replacement programmes. In countries such as Germany, the United States, and Japan, where there are no constraints on the acceptance of patients for treatment, the prevalence is higher than elsewhere in the world. There are important health economic implications: renal replacement treatment has proved so successful that it is now accepted as a right in most developed countries. Improvements in patient survival and increases in acceptance rates (these have trebled in the United Kingdom in the last decade) mean that the total numbers of patients, and therefore the aggregate cost of maintenance treatment (already 2 per cent of the National Health Service (**NHS**) budget) is still increasing. A steady state has not yet been reached, and until it does the resources allocated will have to be increased too.

Causes of chronic renal failure

End-stage renal failure databases are the usual sources for descriptions of the causes of chronic renal failure. There are flaws in these because the meaning of terms such as 'pyelonephritis' may vary; diagnoses are allocated as best guesses by clinicians; glomerulonephritis is diagnosed without renal biopsy; and hypertension cited when it may be no more than a consequence of whatever caused the renal failure. The most rigorous database is that of the Australia and New Zealand Data (ANZDATA) Registry. The data in Table 3 should be interpreted with these caveats in mind.

Glomerulonephritis

Primary glomerulonephritis and secondary inflammatory glomerular disease (see Section 20.7, and Chapter 20.10.3 and Chapter 20.10.4).

Glomerulonephritis remains the most common cause of chronic renal failure outside the United States, accounting for 34 per cent of new cases in Australia. The patients usually suffer the common chronic glomerulonephritides, especially IgA disease, but including focal sclerosis, membranous nephropathy, and mesangiocapillary nephritis. These patients, more often males, are identified as marked for ESRF by heavy proteinuria, hypertension, interstitial changes in their renal biopsy specimens, and early and progressive renal dysfunction, in other words they have severe disease. Others have glomerular lesions secondary to systemic diseases such as systemic lupus erythematosis, Henoch–Schönlein purpura, systemic vasculitis, and, rarely, antiglomerular basement membrane antibody disease (**anti-GBM** disease). Progression to ESRF can be rapid or gradual after a severe acute nephritic illness has been halted but leaving substantial residual injury.

Diabetes (see Chapter 20.10.1)

This is now the commonest cause of ESRF in the United States (40 per cent of new patients) but is still behind that for glomerulonephritis in the United Kingdom and Australia (21 per cent). The diagnosis of diabetic nephropathy is usually assumed because of proteinuria, usually with concomitant retinopathy, in patients with a history of diabetes for 10 or more years. About half of the patients have type 2 diabetes, meaning that the onset was in middle life and not immediately requiring insulin.

By 40 years from the onset of diabetes, some 30 to 40 per cent of patients with type 1 diabetes have developed nephropathy and some, but not all, of these will survive long enough to develop ESRF. Between 5 and 10 per cent of type 2 diabetics already have nephropathy at the time their diabetes is diagnosed, and by 20 years from diagnosis 25 per cent of these will have overt nephropathy. It is not possible to predict with certainty which diabetics will develop nephropathy, but it is more likely in Blacks, Asians, and males, and in those with a family history of hypertension. There is now persuasive evidence that good glycaemic control reduces the risk of diabetic nephropathy, and that improving control will reduce the risk of progression of patients with microalbuminuria to overt nephropathy. It has not been shown that glycaemic control affects the prognosis of established diabetic nephropathy. Once overt nephropathy (proteinuria above 0.5 g/day) develops, the median time to ESRF is about 7 years for those with type 1 diabetes but it is more variable in type 2.

Young people with type 1 diabetes present the greatest challenges—they often have many additional complications such as blindness from retinopathy and vitreous haemorrhage, peripheral and autonomic neuropathy, precocious cardiovascular disease, and, unfortunately, some scepticism of conventional medical advice. The older diabetics are usually obese and often have severe peripheral vascular and coronary disease. Both these groups need a multidisciplinary team to care for them—nephrologists, diabetologists, ophthalmologists, vascular surgeons, and podiatrists.

When assessing diabetics with renal problems it should be remembered that they are also susceptible to non-diabetic renal disease such as glomerulonephritis and (particularly) renovascular disease. Their renal failure can be exacerbated by papillary necrosis, usually associated with pyelonephritis. They are more susceptible to renal tuberculosis and fungal infections, especially if autonomic neuropathy has affected bladder function.

There is a general and probably appropriate tendency to start dialysis earlier in those with diabetes. In the younger patient one aims for renal transplantation, ideally with whole pancreas transplantation, as early as possible. Islet-cell transplantation may, in the future, prove to be the best option. Dialysis and diabetes seem to have a synergistically adverse effect on the vasculature. Before embarking on transplantation, a vigorous search for silent coronary artery disease is essential. The prognosis for the survival of all diabetics is much worse than for any other cause of renal failure except malignancy.

Hypertension and renal vascular disease (see Chapter 15.16.2.2 and Chapter 20.10.2)

Hypertension is cited as the primary renal disease causing ESRF in 12 per cent of Australian patients and in 25 per cent in the United States. Chronic renal failure is a rare complication of primary essential hypertension, but the large number of patients with hypertension means that it is a relatively common cause. The risk of renal failure is higher in Blacks. Accelerated-phase hypertension was once a frequent cause of ESRF but is now relatively rare, except again in Blacks. Renovascular disease, which is not separately classified, is an increasingly common cause, especially in the elderly. The patients have a history of other vascular diseases—coronary, cerebral, or peripheral—and present with renal impairment, hypertension, and occasionally 'flash' pulmonary oedema. If the arterial stenoses or occlusions are bilateral or the stenosis is in the artery supplying the functionally dominant kidney, angiotensin-converting enzyme (**ACE**) inhibitors or a2-receptor blockers will cause a sharp but usually reversible rise in plasma creatinine. Angioplasty and stenting seldom rescue much renal function but may slow progression or prevent acute occlusion. In atheromatous renal arterial disease, stenting makes relatively little difference to blood pressure control.

Adult polycystic kidney disease (see Chapter 20.11)

The development of renal failure is seldom a surprise in patients with adult polycystic kidney disease whose condition has usually been diagnosed for other reasons (for instance, a family history, hypertension, haematuria, or a loin mass). Once renal impairment is diagnosed, the loss of function is predictable at about 5 to 6 ml/min per year, tending to be more rapid in males than females. Adult polycystic kidney disease accounts for about 6 per cent of those receiving renal replacement therapy

in the United Kingdom and Australia, the median age at which ESRF (which is not inevitable) is reached being 55 years, but with a wide range from 25 to 75 years. ESRF occurs 10 to 15 years later in those with type 2 adult polycystic kidney disease than in those with the commoner type 1 disease.

Reflux nephropathy (see Chapter 20.12)

Reflux nephropathy (a congenital and often inherited abnormality of the vesicoureteric junction) and congenital structural abnormalities of the urinary tract are important causes of chronic renal failure in the young, the mean age of entering renal replacement therapy programmes being 30 years. Once a threshold fraction of nephron mass has been lost, perhaps as a consequence of infection-related scarring, the remaining glomeruli develop segmental hyalinosis and eventually sclerosis, manifesting as proteinuria and hypertension, both heralding a steady decline towards end-stage disease. Antireflux procedures make no difference to the prognosis, and management is medical.

Miscellaneous (Table 4)

Many renal conditions, some primary and others secondary to systemic diseases, can cause chronic renal failure.

Drugs

Analgesic nephropathy is declining in incidence but is still common in Australia and parts of Europe (5 per cent). Ciclosporin toxicity can cause chronic renal failure in patients who have received cardiac, liver, and lung allografts. A few patients on long-term lithium medication for bipolar affective disorder develop chronic renal failure. Long-term, non-steroidal anti-inflammatory drugs (**NSAIDs**) use is also associated with chronic renal failure.

Obstructive uropathy (see Chapter 20.14)

Neglected or unrecognized obstruction, associated sometimes with calculi, infection, or malignancy, accounts for a small but significant number of patients requiring dialysis. Any manoeuvre (for example, ureteric stenting) which relieves obstruction, is worthwhile for the preservation of renal function.

Dysproteinaemias (see Chapter 20.10.5)

Primary amyloid, myeloma kidney, and the other immunoglobulin deposition diseases are relatively rare. The patients, usually elderly, generally have a poor prognosis, having to cope with the consequences and treatment not only of the underlying disease, but also the hazards and inconvenience of dialysis. Survival on dialysis is seldom longer than 2 years.

Pregnancy (see Chapter 13.5)

Irreversible postpartum renal failure is now rare in developed countries but is still a problem in the developing world.

Unknown

In a significant proportion of patients no confident diagnosis of the cause of chronic renal failure can be made. There are no clues in the history, although a renal condition may have been suspected because of long-standing minor urinary abnormalities (such as asymptomatic proteinuria or haematuria). Imaging reveals small echogenic kidneys, which cannot safely be biopsied, and even if tissue does become available it seldom reveals a specific diagnosis. The glomeruli are sclerosed and there is widespread interstitial fibrosis and vascular changes, which are probably secondary. It may be that there are unrecognized renal diseases caused by environmental toxins, but this is speculation. The recent description of the development of renal failure after the consumption of Chinese herbal remedies containing aristocholic acid shows the need to be agnostic on causality. Some patients do probably have unrecognized glomerulonephritis (as shown by the development of IgA nephropathy in subsequent kidney allografts). Others may have silent cholesterol emboli, analgesic nephropathy, or 'burnt out' tuberculosis.

The pattern in developing countries of Asia, Africa, and Latin America is quite different. Chronic glomerulonephritis (especially infection-associated disease, see Chapter 20.7.10) and hypertension are the dominant causes in Africa. In Asia, diabetes mellitus—usually type 2—is almost as common a cause as glomerulonephritis. Obstructive uropathy is a more common cause than in Europe because of the higher incidence of tuberculosis, schistosomiasis, urethral strictures, and renal stone disease. Other diseases more common in these countries, which may cause chronic renal failure, include systemic lupus erythematosus (especially in Asian women), sickle-cell disease, and human immunodeficiency virus (**HIV**) infection.

Pathophysiology of chronic renal failure

In chronic renal failure, compensatory and adaptive mechanisms maintain acceptable health until the glomerular filtration rate is about 10 to 15 ml/min, and life-sustaining renal excretory and homeostatic functions continue until the glomerular filtration rate (**GFR**) is less than 5 ml/min. The favoured explanation for the pathophysiology of this condition centres on the 'intact nephron hypothesis' first proposed by Bricker, which states that despite distortion of renal architecture and a widened range of single nephron GFR in diseased or damaged kidneys, glomerular and tubular function remains closely integrated in all individual nephrons, both normal and damaged. As the GFR of the whole kidney falls, still-functioning nephrons produce an increased volume of filtrate ('hyperfiltration') and their tubules respond appropriately for overall homeostasis by excreting fluid and solutes in amounts that maintain external balance. For sodium and potassium, compensation can occur down to a glomerular filtration rate as low as 5 ml/min and plasma values are commonly normal. For phosphate and urate, adaptation is less precise and plasma concentrations are increased in many patients at a glomerular filtration rate of 20 ml/min and in almost all at 5 to 10 ml/min.

The functional adaptations of individual nephrons that allow homeostasis to be maintained in the face of a substantially reduced GFR do not come without a price, and the 'trade off' hypothesis needs to be considered alongside the intact nephron hypothesis. This describes the concept that adaptations arising in chronic renal failure may control one abnormality, but only in such a way as to produce other changes characteristic of the uraemic syndrome. The best example of 'trade off' is the increase in parathormone secretion essential for the increased fractional excretion of phosphate: as the glomerular filtration rate falls plasma phosphate rises, parathormone secretion increases, and plasma phosphate is lowered by decreased tubular reabsorption. The cost of normal plasma phosphate is then secondary hyperparathyroidism, sometimes leading to metastatic calcification (see Chapter 20.6.1).

Electrolytes and water

Inability to concentrate urine in the presence of dehydration is often the first symptom of chronic renal failure, resulting in polyuria, nocturia, and thirst when GFR is about 30 ml/min, although diseases that predominantly affect the medulla, such as pyelonephritis, interstitial nephritis, and medullary cystic disease, may present with a concentration defect at an earlier stage. Defective urine concentration is due to an increased solute load in surviving nephrons, with minor contributions from decreased tubular function and increased GFR per nephron. Thirst accompanies polyuria and water balance is maintained provided there is free access to fluid. As obligatory water loss is increased, careful attention needs to be paid to fluid balance in the presence of anorexia, fever, surgery, and other sources of extrarenal loss if dehydration, hypotension, and further impairment of renal function are to be avoided.

Diluting capacity is preserved until renal failure is advanced, the asymmetrical narrowing of the range of urinary osmolality eventually producing the fixed (300 mosmol/kg) urinary osmolality of chronic renal failure with its obligatory polyuria (Fig. 1). It should be noted, however, that although urinary dilution is maintained until late in chronic renal failure, large water loads are excreted more slowly than in normal subjects and excessive intake (by drinking or an ill-advised iatrogenic infusion of dextrose-containing solutions) can result in hyponatraemia, mental disturbances, and convulsions.

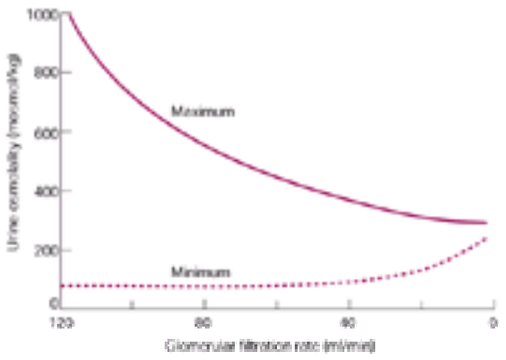

Fig. 1 Progressive loss of flexibility in water handling as renal failure worsens. Concentrating ability is impaired earlier than the ability to excrete a dilute urine.

Sodium excess and hypertension

As renal function decreases, hormonal mechanisms increase the fraction of filtered sodium excreted so that the sodium balance and extracellular fluid volume are maintained until the GFR is less than 10 ml/min. The extent of this adaptation is such that the 1 per cent or less of filtered sodium excreted by normal subjects increases to 30 per cent in those with late chronic renal failure. However, adaptive mechanisms are not unlimited, and in late renal failure increased total body sodium, with water to maintain osmotic equilibrium, presents as fluid overload and hypertension. Initially, excess extracellular fluid does not cause oedema, but in late renal failure an elevated jugular venous pressure, functional incompetence of the mitral valve, and pulmonary and peripheral oedema are often seen. Another major consequence of sodium and fluid excess is hypertension, present in 80 per cent of patients in late chronic renal failure and occasionally presenting in the accelerated phase, although precisely how sodium retention and increased extracellular fluid volume lead to high blood pressure remains uncertain.

Sodium depletion

Patients with chronic renal failure can also be vulnerable to sodium depletion. In the presence of dietary sodium restriction or loss of sodium by various routes, functioning nephrons cannot restrict sodium excretion promptly so that the extracellular fluid, plasma volume, and GFR all decrease. Although this sodium and fluid loss has been attributed to an osmotic diuresis, other mechanisms are involved and may dominate; thus, if sodium restriction is induced slowly over months, patients can reduce their urinary sodium concentration to less than 10 mmol/l without significant reduction in GFR. A few patients with early chronic renal failure, usually with diseases affecting the renal medulla (for example, obstructive uropathy and medullary cystic disease), present with a urinary sodium leak and sodium depletion on a normal sodium diet. Blood pressure in these patients is normal or low, often with a postural drop of arterial pressure. Sodium supplements may be needed.

Potassium

Most patients maintain a normal external potassium balance until their GFR is less than 5 ml/min, but their capacity to excrete potassium is limited and severe hyperkalaemia may follow a sudden reduction in residual GFR, excess dietary intake (chocolate, nuts, instant coffee, some fruits and their juices, wine), potassium-sparing diuretics (spironolactone, amiloride, triamterene), medication with a high potassium content, surgery, and hypercatabolic states. Acidosis raises serum potassium by ion transfer out of cells and interference with renal excretion. Hypoxia causes hyperkalaemia by impaired uptake of potassium from extracellular fluid. In some patients, particularly those with diabetes mellitus and/or interstitial nephritis, and sometimes in early chronic renal failure, hyperkalaemia may be due to selective aldosterone deficiency (hyporeninaemic hypoaldosteronism) or the use of angiotensin-converting enzyme (ACE) inhibitors. Tubular resistance to aldosterone is another rare cause of hyperkalaemia. Complications occur at plasma potassium concentrations above 7.0 mmol/l; a weakness in pelvic and shoulder girdle muscles may be the presenting symptom, but in most patients serious electrocardiographic abnormalities and cardiac arrhythmias are the first sign of hyperkalaemia (see Chapter 20.2.2 and Chapter 20.5.2 for further discussion).

Calcium and phosphate and Vitamin D

The role of the kidney in regulating calcium and phosphate in body fluids and tissues is described in Chapter 20.6.1. Magnesium concentrations are usually high and care must be exercised with the use of magnesium-containing drugs.

Acid–base

The kidney is an essential organ for maintenance of the acid–base balance by reabsorption of filtered bicarbonate, acidification of urinary buffers. and excretion of ammonia. As renal failure progresses, at least until GFR is less than 20 ml/min, intact nephrons increase their excretion of hydrogen ions to prevent acidosis. Increasing acidosis, variable between patients, occurs at a GFR of less than 10 ml/min—when normal net acid production exceeds the excretory capacity of remaining nephrons and diminished tubular function impairs ammonia synthesis and bicarbonate regeneration. Renal diseases that principally affect tubules and interstitial tissues are associated with acidosis quite early in the course of chronic renal failure. Acidosis seldom requires treatment unless the bicarbonate concentration is less than 15 mmol/l and the pH less than 7.30, except in children in whom prevention of severe acidosis with bicarbonate supplements may have a beneficial effect on renal osteodystrophy and growth retardation. Delayed excretion of excess base is also a feature of late chronic renal failure so that metabolic alkalosis may occur more easily and resolve more slowly after, for instance, prolonged gastric aspiration.

Endocrine dysfunction

The endocrine functions of the kidney can be disturbed in kidney disease, for example reduced production of 1,25-dihydroxy vitamin D (see Chapter 20.6.1) and erythropoietin (Fig. 2). There are also diverse abnormalities in the production, control, protein binding, catabolism, and tissue effect of extrarenal hormones in renal failure. Hormone concentrations may be elevated as a result of reduced degradation (insulin) or increased secretion in appropriate response to metabolic alterations (parathormone, **PTH**). Hormone concentrations may be reduced owing to impaired production (oestrogen, testosterone). There may also be disturbances of activation through altered prohormones. Finally, reductions in hormone-binding proteins are most commonly a consequence of protein loss in nephrotic patients or in those on continuous ambulatory peritoneal dialysis.

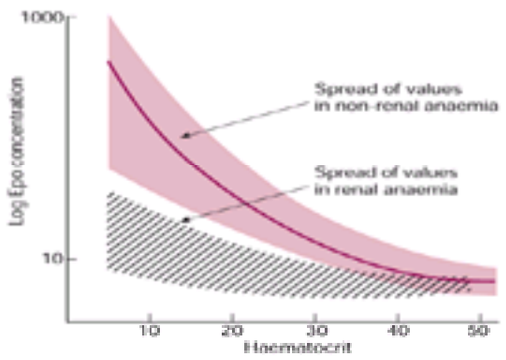

Fig. 2 In renal anaemia the erythropoietin concentration rises in response to anaemia, but to a much lower concentration than in non-renal anaemia. This is a consequence of defective oxygen sensing, reduced synthesis, or both.

Thyroid hormones

Total thyroxine (T_4) may be low with increased reverse tri-iodothyronine (T_3) as a result of impaired conversion of T_4 to T_3. Loss of thyroid-binding globulin (**TBG**) may further lower total circulating T_4 concentrations. However, patients are not clinically hypothyroid and measurements of thyroid-stimulating hormone (**TSH**) remain a reliable diagnostic test for hypothyroidism in patients with renal failure.

Growth hormone

Plasma growth-hormone levels are abnormally high in patients in renal failure because of delayed clearance and alterations in the hypothalamic–pituitary control of growth-hormone release. In adults the clinical implications of these changes are not clear. In children with renal failure and growth retardation, production of insulin-like growth factor-1 (**IGF-1**) in response to growth hormone is impaired; this can be overcome by treatment with exogenous recombinant growth hormone in supraphysiological dosage.

Insulin

Decreased clearance of insulin seems to be balanced by increased peripheral resistance to the effects of insulin, hence there are usually no clinical effects and patients are not prone to hypoglycaemia or diabetes, but there is a reduced requirement for insulin in diabetics as renal function declines.

Sex hormones

Males

Prolactin levels are high in renal failure and may contribute to gynaecomastia and sexual dysfunction in men. Testosterone levels are often low to normal in males, but gonadotrophins are raised, implying testicular failure as the cause.

Females

Raised prolactin levels contribute to infertility. In severe renal failure the pituitary–ovary axis is disturbed: luteinizing hormone is raised, but the normal pulsatile release and preovulation surge are absent, hence cycles are often anovulatory, causing oestrogen deficiency.

Middle molecules and the uraemic syndrome

Although many of the manifestations of the uraemic syndrome are attributable to the derangement in electrolyte concentrations, fluid imbalance, and endocrine deficiencies, there are others that are explained by the actions of substances and metabolites retained because of excretory failure. Examples include encephalopathy, glucose intolerance, platelet dysfunction, anaemia, and leucocyte dysfunction. Knowledge of the nature, mode of action, and contribution of those substances referred to as 'uraemic toxins' or 'middle molecules' to the syndrome is incomplete. The best example of a middle molecule is b $_2$-microglobulin, which is normally excreted by the kidneys, reaches a concentration of 30 times higher than normal in dialysis patients, and accumulates as b $_2$-microglobulin amyloid in joints and bone. It meets fully the criteria of a 'uraemic toxin'. However, because a substance is retained in uraemia and removed by dialysis does not mean that it can be indicted for a particular element of the uraemic syndrome. To be sure that a substance is relevant to the uraemic syndrome, its concentration should be raised in renal failure, should relate to the severity of the particular effect, and the effect should be reproduced by the substance alone and ameliorated by reducing the concentration. The subject of uraemic toxins has been approached by considering: (1) their existence predicted by abnormalities detected in *in vitro* systems; (2) known chemical effects of urea retention; and (3) identifiable substances.

Urea itself has rather modest effects, but it is used as a marker of accumulation of other metabolites. It inhibits cell-membrane electrolyte transport *in vitro* and is thought to have a direct effect on appetite and protein anabolism. It forms isocyanic acid, which reacts with amino groups on amino acids, carbamoylating them or the proteins of which they are constituents. Similarly, modification of lipoproteins may decrease their binding to receptors, thereby delaying their metabolism.

In a manner analogous to reverse genetics, the existence of certain uraemic toxins can be predicted. Examples include insulin resistance, which in uraemia is improved by dialysis, and a number of factors in uraemic plasma that inhibit glucose metabolism have been found but not identified. Similarly, inhibitors of calcitriol binding to its receptor have been found in ultrafiltrates of uraemic plasma.

Some compounds that accumulate in uraemia are thought to have a direct effect, including:

1. homocysteine—raised concentrations of which are thought to be atherogenic by an oxidation effect on lipoproteins;
2. methylguanidine—a neurotoxin that may explain the uraemic peripheral neuropathy;
3. the aminoguanidine **ADMA** (asymmetric dimethylarginine)— is a potent inhibitor of nitric acid synthesis that may have a bearing on hypertension of renal failure;
4. b$_2$-microglobulin—accumulation of which causes b$_2$-microglobulin amyloid;
5. drugs such as morphine—which are metabolized to excretable glucuronides that accumulate in renal failure and cause or exacerbate encephalopathy.

Although incomplete, knowledge of the presence of uraemic toxins is what underpins the perceived benefit of early and adequate dialysis for uraemia.

One of the clearest manifestations of the uraemic syndrome is anaemia. The physiological mechanisms controlling red cell mass are shown in Fig. 3. The maintenance of a normal red cell mass requires a rate of red cell production by a healthy bone marrow, with no substrate limitations and under the influence of an adequate amount of erythropoietin, to balance red cell loss and destruction. In uraemia all components are disturbed. Red cell lifespan is shortened by accelerated destruction, possibly caused by substances within uraemic plasma that alter the red cell membrane. To compensate for a shorter lifespan, a higher than normal production of red cells is required, which is dependent on an increase in the erythropoietin secretion rate. Concentrations do indeed rise, but not enough to set erythropoiesis at a sufficient level (Fig. 2). The classic experiment of providing exogenous erythropoietin to treat anaemia shows that: (1) red cell survival is not altered by erythropoietin; (2) red cell mass is restored; (3) marrow activity has to be maintained at a higher than normal level to achieve this. It is a much-argued point whether the marrow is itself normal. *In vitro* experiments reveal that constituents of uraemic plasma can suppress erythropoiesis. However, the fact that the effects of these constituents can obviously be overridden by exogenous erythropoietin does not exclude the possibility of resistance. It is not possible to say whether the doses of erythropoietin required to reverse anaemia are physiological because the modes of delivery—subcutaneous and intravenous injection—are not.

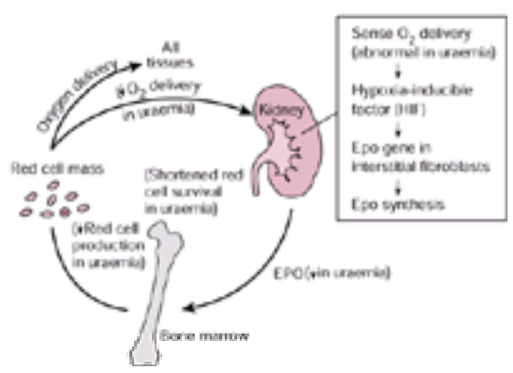

Fig. 3 The relationship between red cell mass, oxygen delivery, anaemia, erythropoietin (Epo) synthesis, and red cell production in the bone marrow. Erythropoietin production is reduced in uraemia, either because of defective sensing or reduced synthesis.

Progression of chronic renal failure

The clinical course of most nephropathies is a progressive decline in renal function. However, the rate of progression varies considerably between patients and diseases, generally being faster in chronic glomerulonephritides than in tubulointerstitial nephritides.

Factors influencing the rate of progression

Proteinuria

The degree of proteinuria correlates with the rate of progression of the underlying nephropathy and is the most reliable prognostic factor in chronic renal failure. In chronic glomerulonephritis, persistent heavy proteinuria (greater than 3 g/24 h) predicts a poor outcome. Conversely, the absence of significant proteinuria or its partial or complete remission indicates a favourable prognosis. In patients with diabetes or tubulointerstitial nephropathy, such as reflux nephropathy, the onset of significant proteinuria (greater than 1 g/24 h) usually predicts a decline in renal function. Proteinuria may be a marker of severe renal disease, but it is possible that the filtration and overloading of the tubules with protein may itself damage the nephron.

Hypertension

The most important factor influencing the rate of progression is systemic hypertension, which appears early in the course of renal diseases and long precedes the onset of ESRF. As with proteinuria, hypertension is a marker of more severe renal disease, but there is good evidence that the raised arterial pressure is itself pathogenetic.

Other factors

Ethnicity and genes

Certain major histocompatibility antigens have been associated with a poor outcome in some forms of glomerulonephritis, for example in membranous nephropathy. In adult polycystic kidney disease, patients with type 1 (abnormal gene on chromosome 16) have an earlier onset and a faster rate of decline compared to those with type 2 (whose abnormal gene is on chromosome 4) (see Chapter 20.11). Afro- and, to a lesser extent, Hispanic Americans suffer a faster rate of progression when compared to Caucasians. Diabetic nephropathy progresses more rapidly in Afro- and native Americans.

Gender

Renal function deteriorates faster in males with adult polycystic kidney disease, mesangial IgA disease, and membranous nephropathy. In Western societies, males tend to have a higher blood pressure than age-matched females, which may explain this difference.

Mechanisms of progression

The loss of filtration rate in chronic renal disorders is a consequence of progressive glomerulosclerosis, tubulointerstitial fibrosis, and vascular sclerosis. Glomerulosclerosis has been attributed to immunological (glomerulonephritis), haemodynamic (hypertension), or metabolic (diabetes mellitus) insults leading to glomerular endothelial injury. In surviving ('remnant') glomeruli, a compensatory increase in intraglomerular capillary pressure (glomerular hypertension) results from a disproportionate afferent arteriolar vasodilatation and the loss of autoregulation, exposing them to systemic hypertension that in turn is associated with endothelial damage. Injury to the glomerular endothelium favours platelet adhesion, aggregation, and the formation of glomerular microthrombi, allowing the transudation of macromolecules, including lipids and growth factors, into the glomerular mesangium. These stimulate mesangial proliferation and the increased synthesis of extracellular collagenous matrix.

Tubulointerstitial scarring

There is a correlation between the severity of tubulointerstitial scarring and GFR. Tubulointerstitial inflammation and widespread interstitial fibrosis are markers of a worse outcome in renal disease: these are characterized by inadequate healing with excessive collagen deposition and involve interactions between renal tubular cells, inflammatory cells, and resident fibroblasts through the release of cytokines and growth promoters.

Vascular sclerosis

The extent and severity of renal vascular changes (arterial and arteriolar) is also relevant to outcome. Although hyalinosis of smaller renal vessels is common in patients of all ages with chronic renal disease, severe arteriolar hyalinosis is often seen in the kidney tissue of patients with chronic nephropathies in the absence of significant systemic hypertension. Moreover, the everity of these vascular changes is greater than that seen in patients with essential hypertension. This arteriolar hyalinosis further jeopardizes the glomerular and tubular blood supply, causing ischaemic injury and further scarring.

Clinical presentation

The presentation of chronic renal failure will depend on the degree of renal dysfunction at the time medical help is sought.

- *Asymptomatic*—At one extreme are asymptomatic patients in whom an abnormal creatinine is noticed on a 'routine' biochemical screen. Such patients may be shocked when it is explained that they have lost what might be a substantial amount of their renal function, and counselling them and persuading them to comply with follow up and medication is sometimes difficult. Patients with illnesses known to cause renal failure, such as adult polycystic kidney disease, are easier to manage because they usually understand the progressive nature of renal failure and the need to introduce various measures in steps.
- *Associated disease*—Much renal failure is picked up in general medical, hypertension, diabetic, cardiac, and urology clinics because clinicians are aware of the effect of other diseases on renal function.
- *Symptomatic presentation*—Relatively few patients are diagnosed because they present with the non-specific symptoms of chronic renal failure, such as lethargy, dyspnoea, and anorexia. Those that are will be relieved that their symptoms have an explanation. At the extreme end of this category are the patients who present with an acute uraemic emergency requiring urgent dialysis, constituting about 5 per cent of patients entering renal replacement treatment programmes. Another 25 per cent are close to end-stage renal failure when they are first seen by a nephrologist and need dialysis within 3 months of the first consultation.

It may at first be difficult to distinguish acute from chronic renal failure, but a systematic history, examination, and appropriate investigations should soon distinguish the two (Table 5). The presentation of ESRF as a uraemic emergency is often the result of missed diagnostic opportunities, but may be the presentation of a rapidly progressive illness such as rapidly progressive nephritis, myeloma, or renal vascular disease.

Assessment

All patients with chronic renal failure should be referred for specialist opinion, although their care can often best be shared with the primary care physician or other specialist, for instance a diabetic physician.

The patient with ESRF

The 5 per cent of patients with chronic renal failure who present as uraemic emergencies may be comatose, may have fitted, and may have asterixis. The skin shows excoriation from pruritus, purpura, and bruising on a sallow yellow-brown background. The blood pressure is raised and examination of the fundi may reveal haemorrhages and exudates. The apex beat is displaced laterally and there is often a pericardial friction rub. There are basal lung crepitations and oedema of the

face, sacrum, and ankles. Blood investigations show a urea concentration above 50 mmol/l, a creatinine concentration above 1000 µmol/l, hypocalcaemia, hyperphosphataemia, hyperkalaemia, and a partially compensated metabolic acidosis. There is a normochromic normocytic anaemia, a normal white blood count, and a platelet count in the low normal range.

The patient may be too ill to give a history, but family and/or friends report a general deterioration in health over the preceding 6 months with dyspnoea, anorexia, pruritus, and nocturia. Such a patient is easy to diagnose, indeed the ammoniacal smell of the breath often alerts the family practitioner. This medical emergency is now infrequently encountered in societies with developed medical services, but the gradual nature of the deterioration is such that it may take some dramatic event like a fit to provoke referral. The morbidity in such patients is high and it is obvious that the opportunity for halting the underlying pathology or slowing progression will have been lost. Most patients present with a much milder combination of symptoms and signs and are often irritated that their non-specific symptoms had not earlier been attributed to chronic renal failure. They should be told that renal failure is rare and that, in the absence of obvious clues, no doctor should be criticized for missing the diagnosis in the early stages.

Many patients will present to hospitals without dialysis facilities. If so, it needs to be established whether there are any immediately life-threatening complications that mandate urgent transfer to a hospital where dialysis can be provided. These include hyperkalaemia, refractory pulmonary oedema, severe hypertension, metabolic acidosis, and encephalopathy. The specific management of these is described in Section 20.5. After dealing with these issues, the next point is to determine whether there are factors which have caused or are causing an acute reduction in chronically impaired renal function. If so, can they be reversed? (Table 6). Examples include:

- *Hypoperfusion*—This can result from dehydration caused by diarrhoea, vomiting, iatrogenic deprivation of fluid (e.g. following surgery), or the overzealous use of diuretics. An occasional cause is the renal loss of salt and water in conditions such as medullary cystic disease. Significant dehydration is associated with a reduction in weight and postural hypotension. Worsening renal arterial stenosis and cholesterol emboli should be sought in arteriopaths.
- *Drugs*—Many drugs, particularly non-steroidal anti-inflammatory drugs (**NSAIDs**), aminoglycosides, and antihypertensive agents, can cause a reduction in GFR, and many others cause acute interstitial nephritis. Tetracyclines cause nausea and vomiting. Clofibrate causes rhabdomyolysis and myoglobinuria. Contrast media in the dehydrated patient are another cause of sudden deterioration in function.
- *Infection*—Systemic infection such as pneumonia can reduce the GFR, and renal parenchymal infections in patients with diabetes, analgesic nephropathy, or adult polycystic kidney disease can damage the remaining functioning renal tissue.
- *Obstruction*—Renal function may worsen substantially in a patient with chronic renal failure if one kidney is obstructed by calculi or papillary necrosis, for example. Sloughed papillae should be sought in those with analgesic nephropathy, diabetes, an obstruction, or sickle-cell disease. Retroperitoneal fibrosis may be occult and is not always detected by ultrasound examination (see Chapter 20.14).
- *Relapse* of the underlying disease—Patients with diseases such as systemic lupus erythematosus, IgA nephropathy, or systemic vasculitic syndromes will deteriorate when the underlying disease relapses, causing further damage to glomeruli. Diagnosis can be difficult because the kidneys are too small to biopsy. Serology and examination of the urine deposit can be helpful. Occasionally, membranous nephropathy and membranoproliferative glomerulonephritis can change in character with the development of extracapillary proliferation (crescent formation), and this is associated with a rapid decline in function. Renal vein thrombosis also causes deterioration in function and should be considered in those with chronic nephrotic syndrome, particularly with underlying membranous nephropathy or focal segmental glomerulosclerosis.
- *Hypertension*—The development of accelerated-phase hypertension may cause a sharp and irreversible reduction in residual renal function. This is most likely in patients with glomerulonephritis.
- *Congestive heart failure*—Independent of the salt and water retention of uraemia, congestive heart failure itself can lead to a reduction in GFR. This can be a result of hypertension, myocardial infarction, or arrhythmias.
- *Hypercalcaemia*—The use of vitamin D analogues such as alfa-calcidol (1a-hydroxycholecalciferol) to prevent hyperparathyroidism often leads to hypercalcaemia. When marked (plasma [Ca^{2+}] >3 mmol/l), this causes a reduction in GFR, usually by causing dehydration.
- *Pregnancy*—Early in pregnancy the plasma creatinine concentration tends to fall, but the course of diseases such as reflux nephropathy or glomerulonephritis may accelerate (see Chapter 13.5).

Once the pressure of the emergency situation is resolved, or if the patient has been referred to the outpatient department with apparently stable chronic renal impairment, the clinician will need to make a thorough assessment of the cause and degree of renal failure and its complications, and institute the appropriate treatments, counsel the patient about the prognosis, and (if appropriate) plan for renal replacement treatment.

History and examination

Questions will be directed towards possible causes, duration of illness. and complications (Table 3, Table 4, and Table 5). Patients with advanced chronic renal failure usually admit to a gradual deterioration in health during the previous 6 months, but some are remarkably uncomplaining, claiming to 'feel fine' despite very abnormal blood tests. This is usually an indication of a very gradual decline towards chronic renal failure. Clues as to the cause may come from all past interactions with doctors: for instance, proteinuria during a medical examination for employment and insurance purposes or during pregnancy, or a urological assessment for microscopic haematuria.

The severity of renal failure will be gauged from the uraemic symptoms of anorexia, vomiting, lassitude, breathlessness, and ankle swelling. The clinician will need to know all about the patient's family as well as their social and employment background to allow advice on the future and to plan treatment.

The examination often provides little extra diagnostic help, but it is important to look for evidence of multisystem disorder, generalized vascular disease (which might indicate a renovascular cause for renal failure), and urinary obstruction. The latter cannot be excluded on physical examination, but if the cause of chronic renal failure is not apparent it is essential to palpate carefully for an enlarged bladder and to perform a rectal examination. Examination also allows an assessment of the consequences of chronic renal failure on blood pressure, left ventricular hypertrophy, and salt and water balance.

Investigations

The work-up of a patient newly diagnosed with chronic renal failure (Table 7) is partly diagnostic and partly for staging and preparation for dialysis and/or transplantation.

Radiological

A minimum set of investigations includes: ultrasonography to determine renal size, echogenicity, and calyceal appearance and to ensure complete emptying of the bladder; and a chest radiograph for heart size (cardiothoracic ratio— **CTR**) and lung fields. Some nephrologists would routinely also take radiographs of the hands and pelvis, looking for renal osteodystrophy and vascular calcification.

Biochemical

Creatinine clearance can be estimated using the Cockcroft–Gault formula (see above). The calcium, phosphate, and parathormone levels will give a pointer to the presence of renal osteodystrophy, and hence to the need for dietary advice and to the scope for prescription of vitamin D analogues and calcium-containing phosphate binders. Measurement of cholesterol and triglyceride concentrations will dictate the use of HMG-CoA (3-hydroxy-3-methylglutaryl coenzyme A) reductase inhibitors (statins). There is now a low threshold for statin prescription to patients with hypertensive renal failure because of their very high risk of cardiovascular events.

Haematological

Most patients will have a normochromic normocytic anaemia; although this will be due mainly to erythropoietin deficiency, it can be exacerbated or caused in some by iron deficiency. The iron status (plasma ferritin, percentage hypochromic red cells) should always be measured so that a trial of intravenous iron can be given before the more expensive option of erythropoietin treatment is initiated, the usual threshold for which is a haemoglobin concentration of less than 11 g/dl. Vitamin B_{12} and folate concentrations are rarely abnormal and should not be part of a routine work-up.

Immunological

These tests are mainly of diagnostic use and are performed only when clinically indicated, that is to say when there is diagnostic uncertainty (Table 7).

Cardiac work-up

An ECG is a routine but insensitive marker of left ventricular hypertrophy or ischaemia, except for previous myocardial infarction. Echocardiography is now the preferred investigation to detect left ventricular hypertrophy. In patients being considered for renal transplantation, exercise testing, radionuclide scanning, or angiography are required if there is any suspicion of asymptomatic ischaemic heart disease, for example in juvenile-onset diabetics or a past history of ischaemic events.

Virological

The presence of the hepatitis B surface antigen (**HepBsAg**) and hepatitis C (**Hep C**) antibodies or RNA will need to be established to decide whether a patient needs to be dialysed in isolation (Hep B), or using a dedicated machine (Hep C). HIV testing in many centres is only performed in high-risk groups, or if renal transplantation is contemplated, but in other centres it is done (after appropriate counselling) as a routine. Knowledge of the patient's immune status for cytomegalovirus (**CMV**), Epstein–Barr virus (**EBV**), and herpes zoster virus (**HZV**) will be useful baseline information if the patient is likely to undergo renal transplantation.

Clinical complications of chronic renal failure

The clinical complications of chronic renal failure are widespread (Fig. 4).

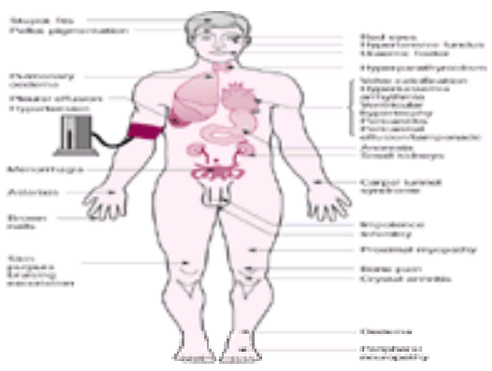

Fig. 4 Symptoms and signs of uraemia.

The cardiovascular system

The single most important complication of chronic renal failure is raised arterial blood pressure, which accelerates atherosclerosis and is the main cause of the left ventricular hypertrophy that is found in 75 per cent of patients. Left ventricular dilatation, coronary atherosclerosis, ventricular dysfunction, and cardiac failure are common, explaining why cardiac disease is the leading cause of death in patients with ESRF, the relative risk being highest in the young. There is a high risk of acute myocardial infarction, but sudden arrhythmic death is the more common fatal cardiac event. Patients are prone to develop pulmonary oedema with relatively small increases in extracellular fluid and tolerate dialytic removal of fluid poorly.

The factors giving rise to this dangerous combination of left ventricular hypertrophy and coronary atherosclerosis operate early in chronic renal failure and include hypertension, dyslipidaemia, anaemia, hyperparathyroidism, and hyperhomocysteinaemia. The combination of these factors has a synergistic effect on the risk of cardiovascular disease.

Pericarditis is a dreaded complication of chronic renal failure because it may lead to tamponade and death. It was more common in the days when dialysis was delayed until the patient was extremely uraemic, but it still occurs in underdialysed, chronically fluid overloaded, and infected patients. Progression to constriction is rare.

Calcific aortic stenosis and mitral valve calcification leading to incompetence occurs in about one-third of patients in dialysis. Endocarditis is not uncommon in haemodialysis patients, but could still be considered as surprisingly rare considering the frequency with which access to the circulation is made. It is usually caused by *Staphylococcus aureus* and leads to destructive valve disease often needing surgery.

Musculoskeletal system

Chronic renal failure causes major problems in the skeleton. Osteitis fibrosa, osteomalacia, and reduced bone turnover are a consequence of hyperparathyroidism, phosphate retention, and deranged vitamin D metabolism. These manifest as bone pain, deformity, pathological fractures, soft tissue and especially vascular calcification, and proximal myopathy. (See Chapter 20.5.2 for further discussion.)

Patients are also prone to develop crystal arthropathy, either from urate or pyrophosphate. Uric acid concentrations are high because of reduced excretion and the effect of diuretics. Long-term dialysis patients develop a specific b $_2$-microglobulin amyloid, deposits of which cause a large joint and spinal arthropathy, and the carpal tunnel syndrome. Large-joint haemarthroses are seen in anticoagulated renal failure patients, probably because of synergy between the use of heparin and uraemia-associated effects on haemostasis. Gout can be prevented with xanthine oxidase inhibitors such as allopurinol (but note the need for a reduced dosage in patients with renal impairment), or uricosuric agents such as probenecid. Initiation of these agents should be covered by colchicine for they can provoke acute attacks. NSAIDs can be used during acute attacks in chronic renal failure for short periods, provided that the effect on salt retention and GFR is acknowledged. Colchicine or corticosteroids are alternatives. The cyclo-oxygenase-2 inhibitors do not have an advantage over conventional agents.

Gastrointestinal system

Anorexia and nausea are almost universal symptoms in uraemia and are accompanied by a blunting of taste. Both lead to decreased caloric intake and malnutrition. If there is poor oral hygiene, mouth bacteria will break down urea in saliva to ammonia, giving an unpleasant taste in the mouth and uriniferous smell to the breath. Patients often suffer early morning vomiting in the late stages of renal failure. These upper gastrointestinal symptoms are aggravated or even caused by opioid analgesics, the metabolites of which accumulate in renal failure. Diverticular disease is a problem in dialysis patients who may become constipated because of a reduction in the fluid and bulk of their diet. Patients on dialysis with primary amyloid are, for the same reasons, more at risk of perforation of the colon.

Clostridium difficile is endemic in renal units. Elderly patients in particular often develop pseudomembranous colitis after treatment with broad-spectrum antibiotics, especially cephalosporins. Treatment is with oral metronidazole or vancomycin, but if a toxic megacolon develops then colectomy is essential to preserve life.

Gastrin levels are higher in patients with chronic renal failure than in controls, but peptic ulceration is not obviously more common than in the general population. However, gastrointestinal haemorrhage, both acute and chronic, is believed to be more common in renal failure and is attributed to angiodysplasia or non-specific gastric ulceration aggravated by the platelet dysfunction of uraemia. The chronic blood loss is more noticeable because the erythroid bone marrow is already near the limit of compensation for shortened red cell survival.

Pancreatitis is only more common in uraemia because it can be provoked by hypercalcaemia, which is a hazard of vitamin D-analogue treatment. Chronic dialysis patients do have a fibrotic pancreas, but this does not seem to have clinically relevant effects on exocrine secretion.

Hepatitis B infection, if contracted in the presence of renal failure, is likely to become chronic. Patients fail to clear the virus because of their depressed cell-mediated immunity, but they seldom develop severe hepatitis or chronic liver disease. Hepatitis C infections are only more common in renal failure because of exposure to blood transfusions and its transmission in haemodialysis units. The natural history does not seem different in patients in renal failure.

The nervous system

Obvious encephalopathy is a very late manifestation of uraemia, typically leading to confusion, myoclonic twitching of distal muscle groups, and impaired consciousness. Seizures are rare unless there is also accelerated-phase hypertension. Before this preterminal state is reached, higher mental function is impaired and patients will complain of difficulty concentrating and of lethargy. It is important to exclude synergic sedation from drugs such as codeine, dextropropoxyphene, carbamazepine, or benzodiazepines. Electroencephalography (**EEG**), although an unnecessary investigation in these circumstances, shows slowing of the background rhythm. Brain computed tomography (**CT**) scans are unhelpful and magnetic resonance imaging (**MRI**) can be frankly misleading. Treatment is with dialysis—frequent, short, and gentle. The myoclonic jerks can be suppressed by benzodiazepines such as clonazepam, 500 to 2500 µg per day.

A specific encephalopathic ('dialysis disequilibrium') syndrome can occur during or after the institution of dialysis in uraemic individuals. From a normal mental state, the patient develops a headache, confusion, involuntary movements, and seizures, all suggesting the development of cerebral oedema. This is attributed to rapid urea removal leading to changes in the water content of brain cells. It is prevented by slow dialysis, avoiding rapid shifts in urea concentration.

Aluminium-induced encephalopathy—dialysis dementia—has disappeared as a clinical problem since dialysis water is properly purified to exclude aluminium, and aluminium-containing phosphate binders are not given for very long periods. Such patients exhibited a gradual deterioration in intellectual performance, progressing to dementia with involuntary movements.

Sensorimotor peripheral polyneuropathy is a late complication of chronic renal failure. This presents as dysaesthesiae, restless legs, and eventually weakness with foot drop, also loss of power in the small muscles of the hand. Nerve conduction studies do not show specific diagnostic features, there being a delay in the conduction velocity and a reduction in the amplitude of the action potential. The neuropathy is thought to be a result of the effect of an unidentified toxic 'middle' molecule. Dialysis results in a slow improvement, but patients are often left with motor disability. Autonomic neuropathy manifests largely as abnormal cardiovascular reflexes, especially during dialysis.

A specific mononeuropathy of renal failure is the carpal tunnel entrapment syndrome caused by b $_2$-microglobulin-derived amyloid deposition. Almost all dialysis patients develop this by 10 years of treatment unrelieved by renal transplantation.

One final comment: although renal failure can lead to many neurological problems, as detailed above, it is important that such problems are not automatically attributed to uraemia—drug accumulation and vascular disorders are common differential diagnoses.

The skin

The effects of uraemia on the skin are obvious and cause much distress to the patients. Yellow-brown pigmentation prominent in sun-exposed areas is a feature of prolonged chronic renal failure. It is attributed to an effect of retained melanocyte-stimulating hormone (**MSH**), retention of vegetable-derived lipochrome and carotenoids, and iron.

Pruritus is the most exasperating symptom for both patient and nephrologist because it is so difficult to treat. It is associated with xerosis (dry skin) and is worse when the skin is warm. A number of explanations are advanced including sensitivity to histamines, a raised calcium phosphate product, and uraemia itself. The itch–scratch cycle can lead to infection and nodular prurigo. Treatment includes starting or increasing dialysis, applying skin emollients, controlling the plasma phosphate level, keeping cool, and the prescription of antihistamines, for example chlorphenamine (chlorpheniramine) 4 mg at night (which is also slightly sedative). Naltrexone, an opioid antagonist, and ultraviolet phototherapy are effective in the short-term.

Cutaneous and subcutaneous calcification (calciphylaxis), a result of small-vessel calcification and occlusion, leads to painful livedo reticulosis and ischaemic ulceration. Once established, it is difficult to treat.

Bullous eruptions in sun-exposed areas, mimicking those of porphyria, are seen in dialysis patients. They are attributed to retained uroporphyrins or other photosensitizing chemicals that inhibit porphyrin breakdown. Iron excess is implicated in the pathogenesis, so phlebotomy and erythropoietin are part of the treatment.

Sexual function

Males

A combination of loss of libido and erectile impotence are experienced by about half the men on dialysis. Loss of libido will be a consequence of ill health and depression and is improved when well being is restored. Impotence has many causes, including pelvic vascular disease, venous leakage because of neuropathy, and drugs (for example, thiazides or a-blockers). Treatment of impotence with vacuum devices, intracavernosal injection of phentolamine, or oral sildanefil may be effective. Low sperm counts and motility that are not improved by hormonal treatment account for the lower fertility of uraemic men. *In vitro* fertilization is an option. Priapism is a rare complication of haemodialysis treatment. Gynaecomastia is common.

Females

Most women with severe renal failure develop amenorrhoea or irregular menses due to hypothalamopituitary dysregulation. Oestrogen levels are low, accounting for atrophic vaginitis and contributing to osteoporosis. Women with severe uraemia are usually infertile and the rare pregnancies almost always end in miscarriage. Patients with mild to moderate renal failure do become pregnant, when the risk of an accelerating decline in function and chance of a successful outcome are related to the severity of preconception renal dysfunction, proteinuria, and blood pressure (see Chapter 13.5).

Haematological effects and host defence

Renal anaemia, which is normochromic and normocytic, accounts for many of the symptoms that previously were attributed to uraemia. These include lethargy, cold intolerance, and loss of stamina. Anaemia increases the cardiac output and therefore contributes to the development of left ventricular hypertrophy and dilatation. The low haematocrit itself has an effect on platelet function as measured by prolonged bleeding times.

Platelet numbers are usually normal but function is impaired at the level of endothelial contact. The coagulation system is not affected. The uraemic bleeding diathesis manifests as occult gastrointestinal blood loss, oozing from any injuries or surgical incision, menorrhagia, and epistaxes. It is aggravated by the consumption of aspirin, a regular component of drug regimens in patients with cardiovascular disease.

T-cell immunity is impaired in renal failure, but the mechanism has not been explained. Evidence for this defect is the higher risk of reactivation of tuberculosis and herpes zoster, a failure to clear hepatitis B, and a poor response to immunization with hepatitis B vaccines. Neutrophil function is abnormal in a number of *in vitro* tests and may explain the high incidence and severity of bacterial infections, especially those associated with vascular access. The defect is attributed to the effects of iron, increased cytosolic calcium, and 'granulocyte inhibitory proteins'.

Metabolic effects

Renal failure causes a degree of glucose intolerance explained by resistance to insulin-mediated glucose uptake in skeletal muscle. There are complex effects on

lipids resulting in an increased concentration of very low-density lipoproteins (**VLDL**) and an increase in high-density lipoproteins (**HDL**). Protein catabolism is enhanced in renal failure, perhaps as a consequence of metabolic acidosis. This has significant effects during periods of malnutrition and infection.

Psychological manifestations

The psychological problems of patients with chronic renal failure, usually anxiety and depression, are the predictable and understandable consequences of loss of health, control, and pleasure. They are most obvious in those with the most to lose—the young and ambitious—and may be relatively minor in the elderly who are grateful that they have a treatable illness and not an immediately lethal one.

The best treatment is good sympathetic symptomatic care from physicians, nurses, and other staff with whom they can build a relationship. In particular, one should try to eliminate fear of the unknown (caused by ignorance and often by gossip in the clinic waiting room) and encourage an optimistic approach. Psychiatrists usually have little to offer unless there is a specific mental illness, but psychotherapists may be able to help with phobias, guilt, and anger. Antidepressants should be used sparingly, but gentle night sedation is frequently helpful.

Medical treatment of chronic renal failure

Patients with chronic renal failure should be managed by nephrologists (or at least physicians with an interest in renal disease) in 'low clearance clinics' or outpatient departments geared to the three issues—conservation of renal function, compensation for the effects of chronic renal failure, and preparation for eventual renal replacement therapy.

Conservation of function and prevention of progression

Measures to conserve renal function may be specific for the cause of renal disease or general, i.e. applicable to all patients. There are a few treatable causes of chronic renal failure diseases for which the pathology can be modified if not actually arrested.

Specific measures

Urinary obstruction

Relief of obstruction is very rewarding, allowing the patient many years of survival without the need for dialysis because 'natural' progression seems to be relatively slower than for other parenchymal diseases. Proof of a persistent obstruction can be difficult in an already dilated renal tract. Isotope renography with furosemide (frusemide) is unreliable when renal function is poor, so if there is doubt, attempts to improve drainage should be made. This may involve an indwelling bladder catheter for high-pressure bladder outflow obstruction, stenting of ureteric strictures, or even antegrade nephrostomy drainage.

Drug-induced renal disease

If analgesic abusers stop taking the drugs, renal function can stabilize. The same may apply to ciclosporin, NSAIDs, and lithium.

Glomerulonephritis—primary and secondary

The progression of most of the common primary glomerulonephritides is relentless. There are no proven treatments of IgA nephropathy, membranoproliferative glomerulonephritis, or focal segmental glomerulosclerosis. Aggressive membranous nephropathy does respond, but not permanently, to regimens of alkylating agents (chlorambucil or cyclophosphamide) with corticosteroids.

Unsuppressed systemic lupus erythematosus and systemic vasculitis can both accelerate the decline in renal function. Proving 'activity' can be difficult and will require knowledge of serological tests, urinary sediment, and other non-specific markers, such as proteinuria, anaemia, and the concentration of acute-phase reactants. Renal biopsy may help, but is not an option when the kidneys are already small and scarred. A trial of increased immunosuppression is often worthwhile.

Ischaemic renal disease

Angioplasty with or without stenting of atheromatous renal artery stenosis seldom reverses renal failure, but it does seem to stabilize or slow progression. It is now a safe undertaking with a low risk of acute occlusion of the renal arteries.

Myeloma

Chemotherapy, which reduces the paraprotein load, improves renal function in patients with myeloma provided the renal failure is not advanced.

Amyloidosis

Treatment of the underlying cause of amyloid A (AA) disease (for example, familial Mediterranean fever (**FMF**) and Still's disease) improves the renal consequences of the process. Chemotherapy of primary amyloidosis (AL amyloid) has been disappointing. Although prednisone and melphalan regimens improve survival, they do not delay the progression of renal failure. It remains to be seen whether marrow ablation and stem-cell rescue will be an effective and applicable treatment for more than a select minority of patients.

Urinary tract infection

Treatment of renal tuberculosis or infection of polycystic kidneys is effective in reducing renal destruction. In reflux nephropathy, suppressing lower urinary tract infections makes relatively little difference, but true pyelonephritis must be treated with the appropriate antibiotics. Antireflux procedures do not alter the natural history.

Stone disease

Measures to reduce stone formation in cystinuria, hypercalciuria, and hyperuricaemia are effective (see Chapter 20.13). Allopurinol is effective in stabilizing renal function in patients with inherited forms of hyperuricaemia.

Diabetes

Control of blood pressure halts or retards diabetic nephropathy, although there is no evidence that better treatment of diabetes (glycaemic control) has a substantial effect (see Chapter 20.10.1).

General (non-specific) measures

Once there has been a loss of more than 50 per cent of renal function the residual nephrons become vulnerable to injury from glomerular hypertension, quite independent of the primary pathology. The clearest example of this phenomenon is the patient who has 1.5 kidneys removed because of cancer and is left with one-quarter of their normal renal mass. Despite the fact that the remaining kidney tissue has no intrinsic pathology, the patient develops hypertension, proteinuria, and progressive glomerulosclerosis leading to renal failure, the mechanisms involved being described above (see ' mechanisms of progression'). Countering the factors that are thought to favour the vicious cycle of nephron loss and autologous injury could stop or delay the process.

Hypertension

There is good evidence relating the presence of systemic hypertension to progression of chronic renal failure and height of the blood pressure to the rate of decline of renal function. Lowering blood pressure is effective in delaying the rate of progression of renal failure, best exemplified in diabetic nephropathy. The optimum target blood pressure is uncertain, but most nephrologists operate on a 'lower the better' principle. Excepting in diabetic nephropathy, it is still unclear whether ACE inhibitors or a2-receptor blockers are superior to other blood pressure-lowering drugs, but the prejudice of most nephrologists is that they should be the first-line agents. The additional benefit of ACE inhibitors is attributed to their effects on glomerular hypertension and the prevention of angiotensin-induced vascular injury.

Dietary protein

Dietary protein restriction slows the progress of glomerulosclerosis in residual nephrons in animal experimental models of chronic renal failure. Trials in humans have evoked much controversy, but the consensus is that there is a modest effect of a 0.2 g/kg per day reduction in protein consumption (from that consumed in a normal Western diet) in patients with GFRs of 13 to 24 ml/min. In practice, nephrologists do no more than advise against high-protein diets. Promotion of low-protein diets has been abandoned in favour of maintaining good nutrition and starting dialysis earlier.

Other

There is no evidence that antiplatelet drugs, anticoagulants, lipid-lowering measures, and low-phosphate diets delay the progression of chronic renal failure.

Compensation for the effects of chronic renal failure

Although there is usually a remarkable adaptation to the loss of as much as 90 per cent of kidney function, a figure which allows the survival of patients with severe chronic renal failure, there is much that should be done to improve health and prevent complications.

Water and electrolyte balance

Only those with oliguric end-stage renal failure need to restrict their fluid intake precisely, when the usual (but seldom complied with) recommendation is that the patient's daily intake should be 500 ml (for insensible losses) plus a volume equivalent to their daily urine output. Patients with chronic renal failure pass normal volumes of urine. However, they do need to be counselled against binge drinking or ignoring extra fluid losses in hot weather and during episodes of diarrhoea or vomiting, because the free-water clearance is blunted and concentration is impaired in renal failure.

Dietary restriction to 60 mmol/day each of sodium and potassium will not exceed the capacity of the failing kidney to maintain balance. Tolerant and flexible advice to provide a safe and tasty diet is more likely to be adhered to than one with absolute and complete exclusions. Sodium balance and blood pressure will be improved by diuretics, usually of the 'loop' type, and in resistant cases in combination with a thiazide such as metolazone.

If the potassium level rises above 7 mmol/l, haemodialysis should be initiated unless there is an otherwise remediable cause. Occasionally this is an isolated finding in an otherwise stable patient, when the ECG should be checked (with emergency treatment as described in Section 20.5 if there are ominous changes) and the measurement repeated. Causes of hyperkalaemia to be considered in those with chronic renal failure—other than a fruit, chocolate, or coffee binge—include gastrointestinal haemorrhage, acidosis, and tissue necrosis, such as a gut infarction or gangrene. Chronic disproportionate hyperkalaemia (for example, when the GFR is still above 10 ml/min) is encountered in diabetics with hyporeninaemic hypoaldosteronism, hypoadrenalism, and as a response to ACE inhibitors.

Calcium, phosphate, and vitamin D

Secondary hyperparathyroidism—so difficult to suppress or reverse when established—starts early in chronic renal failure, when the GFR falls below 40 ml/min. Prevention requires countering the three key stimuli: hyperphosphataemia by diet and phosphate binders; provision of 1,25-dihydroxycholecalciferol, either as calcitriol or 1a-hydroxy-cholecalciferol; and maintaining a normal ionized calcium level. To control phosphate, milk products and fish will be limited; the favoured phosphate binder is calcium acetate taken three times a day with meals. A vitamin D analogue should be started in low dose (e.g. 0.25 µg of alfacalcidol three times/week) as soon as the parathormone is found to be above the normal range. It can be quite difficult to persuade patients to adhere to phosphate restriction or to remember to take binders because the immediate benefit is not obvious. (See Chapter 20.5.1 for further discussion.)

Control of blood pressure

A major focus of the follow-up of patients with chronic renal failure is to achieve and maintain a satisfactory blood pressure: the aim is for less than 140/90 mmHg, but in practice this can be difficult to achieve. It is essential that the patients be engaged as partners in the endeavour: it will help if they understand that good blood pressure control will delay the need for dialysis, prevent left ventricular hypertrophy, which will have an adverse effect on survival, and reduce the risk of cardiovascular events. Patients should be encouraged to measure and record their own blood pressures because readings in hurried clinics and general practitioner surgeries are often unreliable. When there is a disparity between clinic and home readings a 24-h ambulatory recording may provide reassurance that increased doses or extra drugs are not needed. Equally, they will justify such a change for a reluctant patient. The choice of drugs will depend on clinician preference and patient tolerance. There is a move towards the use of ACE inhibitors or angiotensin-receptor blockers as first choice because of the potential added benefits, but most patients will require two to four-drug regimens.

Nutrition

Chronic renal failure causes anorexia, acidosis, and insulin resistance. All three contribute to the subtle malnutrition that may develop in the months during which the decision to start dialysis is procrastinated. The dietician's role here is as much to ensure the prevention of malnutrition as to monitor the excess consumption of the problem items. If anorexia causes a reduction in calorie and protein intake, supplements should be prescribed only as a bridge to the starting of dialysis.

Metabolic acidosis

This frequently goes unnoticed as many laboratories do not routinely report plasma bicarbonate levels and it is unusual to take an arterial blood sample in a low-clearance clinic. Acidosis is more common in patients with interstitial renal disease who have an acquired renal tubular acidosis. The usual symptom is effort dyspnoea not explained by pulmonary oedema or anaemia. A chronic acidosis will aggravate hyperkalaemia, inhibit protein anabolism, and accelerate calcium loss from bone where the excess hydrogen ions are buffered. Sodium bicarbonate 1.2 to 1.8 g thrice daily can be prescribed to patients who can bear this sodium load (for example, those with obstructive uropathy who are acidotic salt wasters).

Anaemia

There is no absolute haemoglobin concentration at which the symptoms of anaemia become manifest, so the decision to treat is a matter of judgement: generally the aim is to maintain the haemoglobin level at or above 11 g/dl. Because chronic anaemia leads to left ventricular hypertrophy, which has adverse effects on patient survival and cardiac function, there is a move towards earlier (and even preventive treatment) of anaemia in patients with chronic renal failure. Whether this will improve survival and reduce cardiac complications is a subject of clinical trials.

For patients not yet on dialysis who have a haemoglobin level under 11 g/dl, one can either start erythropoietin, subcutaneously, at 50 U/kg per week (rounded up to reach 1000 units) in two divided doses, or give a trial of intravenous iron first. This can either be a single dose of iron dextran (1 g), taking precautions against the occurrence of anaphylaxis, or intravenous iron saccharate 200 mg weekly for 5 weeks.. A similar regimen is used for dialysis patients. Patients should learn to inject erythropoietin themselves using the prefilled syringe and pens that are available. Longer acting analogues of erythropoietin are now available and will allow once-weekly administration. Doses can be titrated up, the usual maintenance dose being between 25 and 150 U/kg per week. If patients require higher doses or respond poorly they should be investigated for iron deficiency, sepsis, severe hyperparathyroidism (which causes marrow fibrosis), chronic blood loss, or non-compliance. If no cause is found then a bone marrow examination may be helpful. Recently red cell aplasia caused by autoantibodies to erythropoietin has been

described. Patients on haemodialysis can receive the erythropoietin intravenously, but the dose required will be about 30 per cent higher and they are especially liable to develop iron deficiency, which can be prevented by the intravenous administration of 100 mg iron saccharate every 2 weeks. The ferritin concentration should be monitored and the iron temporarily stopped if it rises above 500 µg/l. A checklist for the management of renal anaemia is shown in Table 8.

The reversal of anaemia increases blood pressure in about 30 per cent of patients. High/normal haematocrits are associated with an increasing risk of vascular access thrombosis and do not appear to alter cardiac prognosis.

Drugs

It is essential that the appropriateness of the prescription of drugs to patients with renal failure be checked and the doses adjusted according to the estimated GFR (see Chapter 20.16).

Preparation for dialysis and transplantation

Once end-stage renal failure is inevitable, the patient must be prepared physically and psychologically for renal replacement treatment. In many patients it is possible to predict approximately when the end-stage will be reached (Fig. 5). This information is useful for the patient and provides a guide for the timing of the creation of vascular access, placement of peritoneal dialysis catheters, or activating the patient on to a transplant waiting list. One should avoid the temptation to delay starting dialysis for as long as possible, for the quality of life and health of a well-dialysed patient is superior to that of a non-dialysed, uraemic malnourished one.

Fig. 5 A reciprocal creatinine plot showing the progressive decline in renal function in a patient with glomerulonephritis. The timing of the need to start dialysis could be predicted sufficiently well to allow planning of treatment.

The absolute indications for dialysis are the development of complications that cannot be contained by conservative and pharmacological means. These are hyperkalaemia, fluid overload, severe hypertension, pericarditis, encephalopathy, and neuropathy. To wait for these is bad practice. Nephrologists generally wait until the patient has some uraemic symptoms such as anorexia, lassitude and pruritus, if only because their relief reinforces the need to adjust to regular dialysis. Apart from potassium concentrations and the degree of acidosis, blood tests such as urea and creatinine do not provide a safe guide to when to start. Nevertheless, it is advisable to start dialysis, in the absence of symptoms, at creatinine clearances of less than 10 ml/min. In small patients with little muscle bulk the urea concentration is often between 30 and 40 mmol/l and the creatinine concentration between 650 and 800 µmol/l; in larger subjects the blood urea concentration is typically 45 to 50 mmol/l and that of creatinine above 1000 µmol/l. Initiation of dialysis at lower blood levels of urea and creatinine is recommended in diabetic patients.

The choice of modality—haemodialysis, continuous ambulatory peritoneal dialysis, or renal transplantation—depends on many factors, not least their availability and the patient's preference (see Section 20.6 for further discussion). If transplantation is appropriate, there is no reason not to perform it before dialysis is mandatory. If haemodialysis is chosen, vascular access should be created 4 to 6 months before it is needed. If continuous ambulatory peritoneal dialysis is to be used, the Tenckhoff catheter should be placed 2 to 3 weeks before dialysis needs to be started to allow it to seal.

Management of terminal uraemia

There will be patients for whom dialysis is inappropriate or who either choose not to start or to discontinue treatment. Because, intuitively, one would predict that instituting dialysis in a patient with renal failure and other comorbid conditions should result in some improvement by ameliorating at least one element of their clinical condition, it is very hard not to start. There are those who argue that there is no harm done by starting because treatment can always be stopped or the patient will die despite dialysis. However, withdrawing dialysis or dying while on treatment are traumatic for both the patient's family and staff. If possible, one should discuss the option of not starting before treatment is actually needed. The patient will need to know what the treatment can achieve and at what cost—access, travel to dialysis, restrictions, and complications. If one takes the view that dialysis is a treatment offered to allow the patient to continue living with a reasonable quality of life as opposed to delaying death in the short term, dialysis will not be offered to patients with other life-limiting conditions. Certainly one could argue that it should not be started when survival beyond 3 months outside of hospital is unlikely, indeed at least 10 per cent of deaths in dialysis programmes follow withdrawal of treatment. The ethical and legal issues are complex and require that the patient makes the decision not to start or to discontinue when fully informed and able to do so.

Properly managed death from uraemia is peaceful and free of suffering. It is important to ensure that the patient has peace of mind, in that they are comfortable with the decision, and that their family members are understanding and supportive. They will be comforted to know that their doctor respects their decision. Several distressing symptoms may need to be controlled. The first is breathlessness from pulmonary oedema and acidosis, best controlled with a morphine infusion. The second is nausea and anorexia, which can be helped with regular chlorpromazine 25 mg four times daily: ondansetron 8 mg twice daily can also be effective. Food and fluid should be offered in small palatable helpings and no pressure to eat exerted on the patient. The mouth can become dry and crusted from mouth breathing and will smell foul from the uraemic saliva. Regular mouth washes and gum care will help. Pruritus is managed by keeping the skin cool, and soft with emollients. The patient may not be aware of myoclonic jerks but these will distress the family, so benzodiazepines, such as clonazepam, should be prescribed.

Further reading

Baigent C, Burbury K, Wheeler D (2000). Premature cardiovascular disease in chronic renal failure. *Lancet* **356**, 147–52.

El-Nahas AM, Tamimi N (1999). The progression of chronic renal failure: a harmful quartet. *Quarterly Journal of Medicine* **92**, 421–4.

Hörl W (1999). Neutrophil function and infections in uraemia. *American Journal of Kidney Disease* **33**, xlv–xlviii.

Locatelli F, *et al.* (2000). The management of chronic renal insufficiency in the conservative phase. *Nephrology, Dialysis, Transplantation* **15**, 1529–34.

Luke RG (1999). Hypertensive nephrosclerosis: Pathogenesis and prevalence. *Nephrology, Dialysis, Transplantation* **14**, 2271–8.

McLaughlin K, Jardine AG, Moss JG (2000). Renal artery stenosis. *British Medical Journal* **320**, 1124–7.

Maschio G, *et al*. (1996). Effect of angiotensin-converting enzyme inhibitor benazepril on the progression of chronic renal insufficiency. *New England Journal of Medicine* **334**, 939.

Phillips AO (2000). Diabetic nephropathy—where next? *Quarterly Journal of Medicine* **93**, 643–6.

Ruggenenti P, *et al*. for the GISEN Group (2000). Pretreatment blood pressure reliably predicts progression of chronic nephropathies. *Kidney International* **58**, 2093–101.

Schaefer F, Wiecek A, Ritz E (1998). Endocrine disorders in chronic renal failure. In: Davison AM, *et al.*, eds. *Oxford textbook of clinical nephrology*, 2nd edn, pp 1854–66. Oxford University Press, Oxford.

United States Renal Data System (2000). 2000 Annual Data Report. Atlas of end-stage renal disease in the United States. *American Journal of Kidney Disease* **36**, Suppl. 2.

Working Party (1999). European best practice guidelines for the management of anaemia in patients with chronic renal failure. *Nephrology, Dialysis, Transplantation* 14, Suppl. 5.

20.5.2 Bone disease in chronic renal failure

Michael Schömig and Eberhard Ritz

Introduction

Renal bone disease is a major cause of disability in patients with terminal renal failure. It is mainly, but not exclusively, due to secondary hyperparathyroidism. Previously, aluminium-induced bone disease, secondary to high aluminium concentrations in the dialysate or ingestion of aluminium-containing phosphate binders, played an important role, but this iatrogenic complication has virtually been eliminated. With more efficient prevention and treatment of secondary hyperparathyroidism, patients with uraemia with low bone turnover are encountered with increasing frequency, but whether this condition (so-called adynamic bone disease) has any clinical consequences, other than the propensity to hypercalcaemia, remains unresolved.

It is important to recognize that abnormal calcium/phosphate metabolism impacts not only on parathyroid glands and bone, but also on cardiovascular function; for example, it increases the risk of cardiac death, calcific aortic stenosis, and coronary plaque calcification. This adds a new dimension to the importance of returning calcium/phosphate metabolism to normal in patients with renal failure.

Pathogenesis of renal bone disease

The role of phosphate excess

In early renal failure, plasma phosphate concentration is normal or low, but renal phosphate excretion (more precisely the fractional clearance of phosphate) is increased. Hyperphosphataemia develops when the glomerular filtration rate is approximately 30 ml/min. This causes and aggravates secondary hyperparathyroidism due to indirect mechanisms, such as inhibition of the synthesis of the active vitamin D metabolite 1,25-(OH) $_2$vitamin D$_3$ (calcitriol) in tubular epithelial cells, and possibly also by inducing a tendency for hypocalcaemia. More recently, it has been shown that phosphate directly stimulates parathyroid hormone (**PTH**) synthesis and secretion as well as causing parathyroid cell proliferation independent of low 1,25-(OH) $_2$D$_3$ and hypocalcaemia.

The role of 1,25-(OH)$_2$vitamin D$_3$ (calcitriol) deficiency

The hepatic vitamin D metabolite 25-(OH)D$_3$ is transformed in tubular epithelial cells to the active vitamin D metabolite 1,25-(OH) $_2$vitamin D$_3$. Synthesis is stimulated by PTH and inhibited by hyperphosphataemia. Even in early renal failure there is a tendency for 1,25-(OH) $_2$D$_3$ concentration to decrease, although this is very often compensated by increased PTH concentrations. The average concentration of 1,25-(OH) $_2$D$_3$ falls as renal failure progresses (Fig. 1). One specific problem is that vitamin D metabolites bound to plasma-binding protein (DBP) may be lost in the urine in patients with nephrotic-range proteinuria, so that deficiency of active vitamin D may ensue. The renal 1-a-hydroxylase reaction is normally substrate-independent, but becomes dependent on the availability of the substrate 25-(OH)D $_3$ in some patients with renal failure, hence vitamin D deficiency aggravates the deficit in the synthesis of 1,25-(OH) $_2$D$_3$.

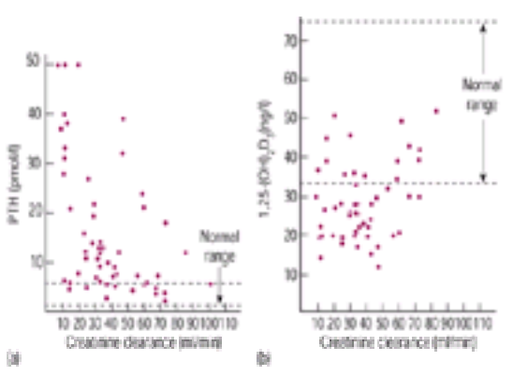

Fig. 1 Intact plasma parathyroid hormone values and serum 1,25-(OH) $_2$D$_3$ levels as a function of the glomerular filtration rate (GFR) in patients with chronic renal failure.

The role of hypocalcaemia

Plasma (total and ionized) calcium concentrations are maintained until the patient reaches pre-endstage renal failure. Nevertheless, the tendency to hypocalcaemia, which is due to a combination of (1) reduced active calcium resorption in the intestine as a result of insufficient active vitamin D, and (2) resistance of the skeleton to release bone mineral and calcium as a result f (partial) resistance to PTH and active vitamin D, may play a more important role in the genesis of secondary hyperparathyroidism than previously thought. The parathyroid gland senses the calcium concentration in the extracellular space via the calcium receptor (CaR) and there are some observations that argue for abnormal sensing of Ca^{2+} even in early renal failure, which may possibly be reversed by agents that improve calcium sensing (calcimimetics).

Clinical manifestations

Pattern of skeletal involvement

The following bone lesions are found in the skeleton of patients with renal failure, in isolation or in combination (Table 1).

Osteitis fibrosa

This is increased osteoclastic bone resorption and increased osteoblastic bone aposition with (1) consecutive intense remodelling of bone trabeculae in the spongiosa, and (2) rarefaction and tunnelization of cortical bone with or without deposition of fibrous tissue (endosteal fibrosis) (Fig. 2).

Osteomalacia

Osteomalacia is a disparity between the rate of bone matrix synthesis and bone matrix mineralization, leading to widening of the seam of unmineralized bone matrix (osteoid), usually associated with signs of diminished numbers and activities of cells at the bone surface. Pure osteomalacia is rarely seen nowadays. In the past it was mainly due to aluminium toxicity and vitamin D (cholecalciferol) deficiency.

Mixed lesions

In many patients with renal failure a combination of osteitis fibrosa and osteomalacia are present.

Adynamic bone disease

In patients with a low serum PTH concentration the number and activity of cells on the bone surface is strikingly reduced and bone turnover is reduced, as evaluated by isotope- or tetracycline-labelling techniques. This condition is relatively frequent in patients with renal failure treated with active vitamin D. It predisposes to hypercalcaemia because the capacity of the skeleton to sequester calcium is reduced, but whether it has more far reaching clinical implications is currently unknown.

Osteopenia or osteoporosis

The problem of diminished bone mass, superimposed upon uraemia-specific bony abnormalities, is very common in patients with renal failure. The most common causes are a history of treatment with steroids and (premature) menopause. It is currently unresolved whether the risk is aggravated by smoking and low calcium diets and whether it can be prevented by substitution of oestrogens/gestagens or selective oestrogen receptor modulators.

Other pathologies

There are several pathologies unrelated to calcium metabolism that have to be taken into consideration in patients with renal failure (Table 1). A dialysis-specific type of amyloidosis with preferential osteoarticular involvement—b $_2$-microglobulin-related amyloidosis—must also be considered in the differential diagnosis of bone pain or bone destruction (see Chapter 20.6.1).

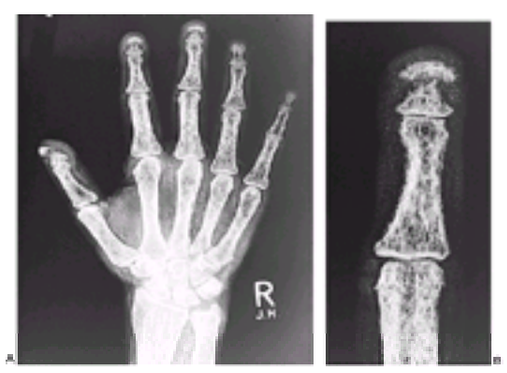

Fig. 2 The radiograph of the hand (a) shows reduced mineral density as well as fluffy and mottled texture of the bones. Note (i) subperiosteal resorption zones at the radial site of the middle phalanges (see also (b)—erosion cavities with overlying areas of calcification (periosteal neostosis), (ii) longitudinal striation of cortical bone (corresponding to enlarged Haversian channels), (iii) thinning of cortical bone by endosteal bone resorption, and (iv) loss of the terminal lamella of the terminal phalanx. The terminal phalanx of the second digit had collapsed, such that the patient presented with 'pseudo-clubbing'. Vascular calcifications are seen above the first digit and along the exterior side of the radius. A detail of the index finger is shown in (b).

The salient differences between osteitis fibrosa and aluminium-related bone disease are summarized in Table 2.

Signs and symptoms

While patients with renal failure left untreated usually have hypocalcaemia and hyperphosphataemia, patients with advanced secondary hyperparathyroidism are characterized by hypercalcaemia and hyperphosphataemia associated with an increase in alkaline phosphatase and its bone isoenzyme.

In the patient with hypercalcaemia it is important to consider causes other than secondary hyperparathyroidism which necessitate specific treatment (Table 3). Similarly, bone pain is not common even in advanced osteitis fibrosa, but bones subjected to mechanical stress (spine, calcaneus, foot) may be painful. Whilst fractures are uncommon, skeletal deformity, leontiasis faciei, and avulsion of the patella may occur. By contrast, osteomalacia, particularly that secondary to aluminium intoxication, may be very painful, especially when Looser zones—fatigue fractures—occur. Again it is important to exclude alternative causes of bone pain (Table 4).

Severe extra-osseous calcifications—periarticular, bursal, or visceral calcifications (see Fig. 2, Fig. 3, and most dramatically, Fig. 4)—are usually the consequence of severe hyperphosphataemia with or without elevated serum PTH concentrations. Tumoral tissue calcification is often triggered by trauma, for instance haematoma. It is favoured by low bone turnover, a situation in which the capacity of the skeleton to sequester calcium phosphate is diminished. Calciphylaxis is a medical emergency where ischaemic skin eschars form secondary to calcification of cutaneous arterial vessels: it usually responds to parathyroidectomy, which may need to be performed as an emergency.

Fig. 3 Calcification of the popliteal artery in a patient with diabetes and severe hyperparathyroidism.

Fig. 4 Tumorous calcification around the left shoulder in a patient on dialysis with aluminium intoxication.

Prophylaxis of secondary hyperparathyroidism

Rationale

Secondary hyperparathyroidism is the combined result of failing excretory function of the kidney (leading to phosphate excess) and failing endocrine function of the kidney (leading to calcitriol deficiency). Consequently, appropriate management requires that both abnormalities must be treated.

Phosphate control

It is usually recommended that phosphate-lowering interventions should begin once plasma phosphate concentrations exceed the upper limit of the normal range—1.45 mmol/l. This is usually the case when the creatinine clearance is approximately 30 ml/min, but the plasma phosphate concentration depends not only on renal clearance, but also on dietary phosphate intake, protein catabolism, and other confounding factors. The problem of phosphate retention persists when patients are on dialysis: the normal dietary intake of phosphate is 50 to 100 mmol/day, of which 50 to 70 per cent is absorbed in the intestine. This exceeds the amount of phosphate that is eliminated by conventional thrice-weekly haemodialysis (33 mmol per session, i.e. 100 mmol/week), such that an average daily positive phosphate balance of 30 mmol ensues.

The risk of precipitation of calcium phosphate is particularly high if hyperphosphataemia is accompanied by hypercalcaemia, reflected by the calcium × phosphate product (desirable range below 5.6 $mmol^2/l^2$, although the practical value of such calculation is limited).

Phosphate is present in virtually all foods, hence reduction of dietary intake is difficult without incurring the risk of malnutrition. Patients should be advised, however, to avoid items with very high phosphate content, for example dairy products and those to which phosphate is added, such as sausages and phosphate-rich soft drinks. A protein-restricted diet is often recommended to patients with renal insufficiency (although there is controversy regarding this, see Section 20.6) and one desirable consequence is that this diet reduces the dietary intake of phosphate.

However, since dietary restriction of phosphate is usually not feasible or sufficient, patients with uraemia remain in positive phosphate balance unless oral phosphate binders are administered. The agents most commonly used are calcium carbonate and calcium acetate: aluminium-containing substances have been widely used in the past, but because of the risk of aluminium intoxication (encephalopathy, osteopathy, anaemia, etc.) they should generally be avoided. These substances trap phosphate in the intestinal lumen by forming insoluble calcium phosphate complexes, hence it follows that they must be taken together with meals because phosphate in the food can only be precipitated within the intestinal lumen when phosphate binders are present. Furthermore, ingestion of calcium-containing phosphate binders without meals increases the risk of hypercalcaemia. If aluminium-containing phosphate binders are used (in cases where hypercalcaemia develops with calcium-containing phosphate binders), then plasma aluminium concentrations must be monitored at regular intervals (relatively safe range: below 60 µg/l). Phosphate binders without calcium or aluminium are currently under investigation.

If hyperphosphataemia does not respond to intervention, one should consider non-compliance, increased phosphate release from the skeleton (e.g. in marked osteitis fibrosa), or insufficient efficacy of dialysis.

Reversal of cholecalciferol deficiency

Deficiency of the parent compound cholecalciferol (vitamin D_3) is common among patients with renal failure as a result of altered lifestyle with insufficient sun exposure, hyperpigmentation of the skin, and loss of protein-bound vitamin D (metabolites) into proteinuric urine or peritoneal dialysis fluid. Vitamin D deficiency can be diagnosed when plasma 25-(OH)D_3 concentrations are low (< 50 nmol/l). In renal failure the synthesis of 1,25-(OH) $_2D_3$ depends on the concentration of the precursor substance 25-(OH)D_3, which explains why administration of 1000 U vitamin D per day (which is two to three times the average daily intake) leads to an increase of calcitriol and decrease of intact PTH (**iPTH**) in many patients with renal failure. Note, however, that treatment with pharmacological doses of native vitamin D is never appropriate in renal failure: these are much less effective than hydroxylated metabolites (see below) and, if they do raise the serum calcium concentration, carry a substantial risk of inducing prolonged hypercalcaemia.

Administration of active vitamin D

Although there is not complete consensus, most authorities advise that prophylactic administration of active vitamin D should be considered when 1,84-iPTH concentrations are elevated in early renal failure, or two- to threefold above the normal range in advanced renal failure despite measures to correct plasma phosphate and plasma calcium concentrations. In advanced chronic renal failure, or when patients are on dialysis, complete return of iPTH concentrations to normal is not desirable, because in patients with renal failure a normal bone turnover is found only if PTH concentrations are slightly above the normal range. It is currently unknown whether this reflects PTH resistance of the skeleton or insufficient specificity of the PTH assay, which also measures some inactive fragments of PTH.

It has emerged that relatively low doses of calcitriol or alternative active vitamin D preparations (e.g. 1-a-calcidol) are necessary to prevent the progressive increase of iPTH in patients with renal failure, for instance 0.125 or 0.25 µg 1,25-(OH) $_2D_3$ per day. The rationale for administration of 1,25-(OH) $_2D_3$ is not only the acute reversal of oversecretion of PTH, but also prevention of parathyroid hyperplasia. This is important because hyperplasia is at least partially irreversible. Administration of active vitamin D preparation is fraught with the risks of hypercalciuria, hypercalcaemia, and accelerated loss of renal function. However, monitoring urinary and plasma calcium can prevent this, and the risk is negligible with very low doses, that is, 0.125 µg/day of 1,25-(OH) $_2D_3$. 1-a-Hydroxy-cholecalciferol is a prodrug, which is hydroxylated in the liver *in vivo* to 1,25-$(OH)_2$vitamin D_3. Biotransformation may be abnormal if hepatic disease is present, but otherwise the two compounds yield comparable therapeutic results. Table 5 provides an algorithm for the prophylaxis of secondary hyperparathyroidism.

Selection of dialysate calcium concentration

Active intestinal calcium transport is impaired in uraemia, hence patients with renal failure without additional calcium intake are in negative calcium balance. On dialysis, calcium may be lost into the dialysate, indeed convective calcium loss is obligatory with ultrafiltration and may amount to up to 200 to 400 mg/week. Loss of calcium into the dialysate also occurs by diffusion if the plasma concentration of diffusible calcium is higher than the dialysate calcium concentration. In the past, a high dialysate calcium concentration of 7 mg/100 ml (1.75 mmol/l) was recommended, so that net uptake of calcium occurs during the dialysis session to compensate for convective loss of calcium via ultrafiltration during, and negative intestinal calcium between, dialysis sessions. If calcium-containing phosphate binders or active vitamin D preparations are administered, intestinal uptake of calcium is high and the patients may develop hypercalcaemia. Lowering of dialysate calcium concentration to 6 mg/100 ml (1.5 mmol/l) (temporarily even to 5 mg/100 ml or 1.25 mmol/l) counteracts this tendency. It is important, however, to verify that patients take their medication when low dialysis calcium concentrations have been selected. If calcium carbonate and/or active vitamin D preparations are not taken, there is a

definite risk that the calcium balance becomes negative and that secondary hyperparathyroidism is exacerbated.

Treatment of advanced hyperparathyroidism

Administration of active vitamin D

In the patient with advanced hyperparathyroidism (i.e. 1,84-iPTH above approximately 50 pmol/l or eightfold above the normal range), higher doses of active vitamin D are required. However, it is important to stress that active vitamin D must only be administered if hyperphosphataemia and hypercalcaemia are not present (or have been reversed) to prevent extra-osseous calcifications and further stimulation of the parathyroid gland in response to hyperphosphataemia, which is aggravated by administration of active vitamin D. Treatment should start with relatively modest doses, for instance 0.5 µg calcitriol per day. If this dose is tolerated without provoking hyperphosphataemia or hypercalcaemia, then it can be gradually increased until plasma 1,84-iPTH concentrations begin to fall. Several schedules of administration of active vitamin D are currently under discussion, but a complete consensus has not yet emerged.

Continuous compared with pulse administration

In experimental studies, continuous (daily) administration is less effective than intermittent (pulse) administration in lowering PTH concentration and preventing parathyroid hyperplasia. So far, there is no good clinical evidence that this effect is sufficiently marked to be of clinical importance.

Oral compared with intravenous administration

It can be shown that intravenous administration causes rapid lowering of 1,84-iPTH concentrations, but head-on comparisons of intravenous and oral administration have failed to show any superiority of the intravenous route.

Alternative vitamin D analogues

The major side-effects of treatment with active vitamin D are hypercalcaemia and hyperphosphataemia. There has therefore been an intense search for vitamin D analogues that suppress the parathyroid gland while having less hypercalcaemic and hyperphosphataemic potential. Several analogues are available (19-nor-1,25-dihydroxyvitamin D_2, namely Paricalcitol; 19-nor-22-oxa-1a,25-dihydroxyvitamin D_3, namely 22-oxacalcitriol), but so far there is no evidence that they are clinically better.

Calcimimetics

Calcimimetic substances—those that stimulate the calcium sensor—cause substantial acute and sustained decrease in the elevated PTH concentration of patients with moderate and advanced hyperparathyroidism. In experimental studies they also prevent further parathyroid hyperplasia, a finding of great importance because advanced parathyroid hyperplasia is irreversible (see below). There is currently only limited clinical experience, and concerns have been raised because the calcium receptor is expressed on numerous tissues other than the parathyroid gland. Anecdotal observations of long-term administration without side-effects in patients with parathyroid carcinoma raise the hope that these compounds will become an important ingredient in the management of the patient with renal failure.

Parathyroidectomy

It has recently been recognized that marked parathyroid hyperplasia is a process that bears many similarities to tumour growth. In patients whose estimated parathyroid mass exceeds 1 to 1.5 g, nodular hyperplasia is usually found. The nodules frequently exhibit monoclonal growth, with microsatellite analysis showing loss of heterozygosity for many alleles, including putative tumour suppressor genes. These nodules also express few vitamin D and calcium receptors, explaining the frequent lack of response to medical management. It appears that continuous stimulation of the parathyroid gland selectively favours cells with higher proliferative potential, so that the gland progressively escapes from growth inhibitory control mechanisms. This is illustrated by the fact that regrowth, including locally invasive regrowth, occurs in a high proportion (approximately one-third, or even more in studies with longer follow-up) of patients after subtotal parathyroidectomy or autotransplantation of parathyroid tissue.

There has recently been a tendency to consider parathyroidectomy early on if patients with marked elevation of 1,84-iPTH (above approximately 50 pmol/l) fail to respond to medical treatment within 4 to 8 weeks by decreasing their PTH concentration and have massive parathyroid enlargement on imaging procedures, with an estimated mass greater than 1 to 1.5 g.

An absolute indication for parathyroidectomy is calciphylaxis—ischaemic skin necrosis secondary to calcification of skin arteries; a relative indication is intractable pruritus associated with high PTH, or biomechanical problems that require urgent stabilization (e.g. rupture of the patella or epiphyseolysis in children with uraemia).

There is a long-standing debate as to whether total parathyroidectomy or subtotal parathyroidectomy (with a remnant left *in situ* or autotransplanted into the subcutaneous abdominal fat or forearm musculature) is preferable. Leaving parathyroid tissue behind is associated with a relatively high risk of recurrence, presumably because of the higher growth potential of the parathyroid. The risk can be reduced if only non-nodular parts of the gland are autotransplanted. As an alternative to surgery, alcohol injection into the enlarged parathyroids under ultrasonographic guidance has been tried successfully, but this procedure is not completely devoid of risk (paresis of the recurrent nerve).

Table 6 summarizes the approach to the management of patients with advanced renal secondary hyperparathyroidism, and Table 7 gives guidelines on how to interpret the common laboratory values used to diagnose abnormal calcium metabolism or follow therapeutic intervention.

Further reading

Arnold A *et al.* (1995). Monoclonality of parathyroid tumors in chronic renal failure and in primary parathyroid hyperplasia. *Journal of Clinical Investigation* **95**, 2047–53. [The first study to document that monoclonal growth occurs in the parathyroid nodules in patients with uraemia with nodular hyperplasia of the parathyroids.]

Couttenye MM *et al.* (1999). Low bone turnover in patients with renal failure. *Kidney International* **56** (Suppl 73), S70–6. [A review summarizing the current information concerning diminished bone turnover in patients with renal disease—so-called adynamic bone disease.]

Drueke TB (1998). Primary and secondary uraemic hyperparathyroidism: from initial clinical observations to recent findings. *Nephrology, Dialysis, Transplantation* **13**, 1384–7. [An up-to-date review of the pathomechanisms of secondary hyperparathyroidism, emphasizing molecular aspects.]

Felsenfeld AJ (1997). Considerations for the treatment of secondary hyperparathyroidism in renal failure. *Journal of the American Society of Nephrology* **8**, 993–1004. [Very complete update on the pathogenesis of secondary hyperparathyroidism and the rationale for therapeutic interventions.]

Fukuda N *et al.* (1993). Decreased 1,25-dihydroxyvitamin D3 receptor density is associated with a more severe form of parathyroid hyperplasia in chronic uremic patients. *Journal of Clinical Investigation* **92**, 1436–43. [A study documenting deficient vitamin D receptor expression in parathyroid glands with nodular hyperplasia. This explains, at least in part, the resistance of advanced hyperparathyroidism to active vitamin D.]

Gagne ER *et al.* (1992). Short- and long-term efficacy of total parathyroidectomy with immediate autografting compared with subtotal parathyroidectomy in hemodialysis patients. *Journal of the American Society of Nephrology* **3**, 1008–17. [A study documenting a high rate of hypoparathyroidism and relapse of hyperparathyroidism in patients with total parathyroidectomy and parathyroid autografts and subtotal parathyroidectomy, respectively.]

Hamdy NA *et al.* (1995). Effect of alfacalcidol on natural course of renal bone disease in mild to moderate renal failure. *British Medical Journal* **310**, 358–63. [A study documenting that early intervention with active vitamin D causes less increase in PTH and interferes with the development of bony lesions.]

Hergesell O, Ritz E (1999). Phosphate binders on iron basis: a new perspective? *Kidney International* **56** (Suppl 73), S42–5. [A review summarizing the rationale for the use of phosphate binders and discussing novel developments in this field.]

Hutchison AJ *et al.* (1993). Correlation of bone histology with parathyroid hormone, vitamin D_3, and radiology in end-stage renal disease. *Kidney International* **44**, 1071–7. [A study documenting

marked skeletal abnormalities in patients with renal disease before they are taken into renal replacement therapy programmes.]

Naveh-Many T *et al.* (1995). Parathyroid cell proliferation in normal and chronic renal failure rats. The effects of calcium, phosphate, and vitamin D. *Journal of Clinical Investigation* **96**, 1786–93. [A study showing that phosphate is an important modulator of parathyroid function and involved in the genesis of parathyroid hyperplasia in renal failure.]

Quarles LD *et al.* (1994). Prospective trial of pulse oral versus intravenous calcitriol treatment of hyperparathyroidism in ESRD. *Kidney International* **45**, 1710–21. [A controlled prospective trial documenting the lack of superiority of intermittent intravenous high-dose active vitamin D therapy over oral active vitamin D therapy.]

Ritz E (1994). Early parathyroidectomy should be considered as the first choice. *Nephrology, Dialysis, Transplantation* **9**, 1819–21. [Summarizes the arguments for parathyroidectomy compared with aggressive therapy using active vitamin D in advanced secondary hyperparathyroidism.]

Ritz E *et al.* (1995). Low-dose calcitriol prevents the rise in 1,84 iPTH without affecting serum calcium and phosphate in patients with moderate renal failure (prospective placebo-controlled multicentre trial). *Nephrology, Dialysis, Transplantation* **10**, 2228–34. [A study which documents that very low doses of active vitamin D are effective in preventing the rise of PTH, with no change of serum or urinary calcium, serum phosphate, or creatinine clearance.]

Yalcindag C, Silver J, Naveh-Many T (1999). Mechanism of increased parathyroid hormone mRNA in experimental uremia: roles of protein RNA binding and RNA degradation. *Journal of the American Society of Nephrology* **10**, 2562–8. [A study detailing the molecular mechanism causing increased synthesis of pre-pro-PTH mRNA in experimental uraemia.]

20.6.1

Haemodialysis

Ken Farrington and Roger Greenwood

Introduction

The availability of effective renal replacement therapy has transformed the outlook for patients with chronic renal failure over the past 40 years, replacing certain and imminent death with the prospect of long-term survival. Growing numbers of patients worldwide are now dependent on regular dialysis treatment to sustain life, the escalating proportion of older and frailer patients being direct testimony to the durability and flexibility of the treatment. Inevitably this success needs qualification. First, the functions of the kidney are many and diverse and dialysis effects only partial replacement of a few of these, notably the excretion of nitrogenous waste products, and control of water, electrolyte, and acid–base balance. Second, although there is agreement about the general aims of dialysis treatment, to prolong life and prevent or reduce morbidity from the uraemic syndrome, consensus on the way to achieve these aims is still far off. There is still debate about what constitutes adequate dialysis and even about which parameters are appropriate. Third, although dialysis undoubtedly prolongs life in patients with endstage chronic renal failure, mortality still far exceeds that in the general population. This is mainly due to cardiovascular disease, which is endemic and runs an accelerated course in patients on dialysis, whatever the modality employed. This chapter outlines the scope of current practice in haemodialysis stressing those aspects most relevant to clinical management.

The development of haemodialysis

The pioneers

Although dialysis had been known since the mid-nineteenth century as a means of separating dissolved elements by diffusion through a semipermeable membrane, it was not until over 80 years later that haemodialysis was first used clinically to treat acute uraemia. The development of reliable means of vascular access in the 1960s (the Scribner silastic shunt and the Brescia–Cimino arteriovenous fistula) allowed its extension to maintenance treatment for patients with chronic renal failure and Pandora's box was open.

Expanding services

The initial rigid selection criteria for access to dialysis treatment soon bent under the combined strain of ethical concern and patient expectation, eventually giving way to more liberal policies. There were marked geographical variations, largely economically driven, in the rates of the subsequent expansion of dialysis programmes, as well as in the modalities employed and in the patterns of service provision (Fig. 1 and Fig. 2). Renal transplantation has had a limited impact on this expansion, having hit the ceiling of donor organ availability.

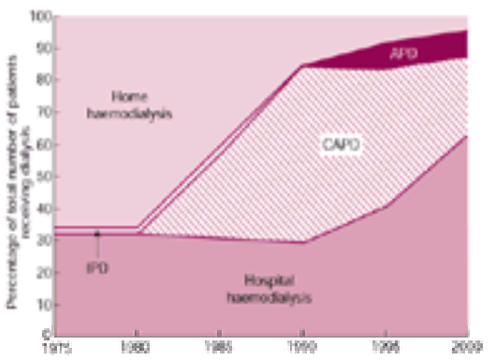

Fig. 1 Dialysis treatment modalities in the United Kingdom. Schematic representation of changes to treatment modality. IPD, intermittent peritoneal dialysis; CAPD, continuous ambulatory peritoneal dialysis; APD, automated peritoneal dialysis.

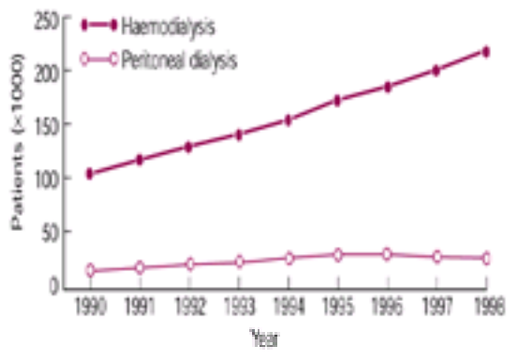

Fig. 2 Dialysis treatment modalities in the United States (United States Renal Data System (USRDS) data).

In the United States and most of Europe, haemodialysis was rapidly decentralized from the pioneering units but remained centre-based in hospitals and free-standing facilities in cities, towns, and rural areas. In the United Kingdom little decentralization occurred, with self-supervised home haemodialysis the chosen means of expansion and selection criteria remaining tight. During the 1980s continuous ambulatory peritoneal dialysis (CAPD) was seized upon as the means to liberalize access to treatment in the United Kingdom. It became the dominant dialysis mode, displacing centre-based haemodialysis to a rescue mode for those in whom CAPD was precluded or had failed. Elsewhere in Europe and in the United States centre-based haemodialysis remained the norm.

The impact of adequacy

Although patients continued to dialyse thrice weekly, there was a general trend to reduce dialysis times below 4 h. This first occurred in the United States as a mechanism to cope with constrained funding. 'Short dialysis' was held responsible for the excess mortality in American patients on haemodialysis during the 1980s, and inadvertently focused attention on the concept of dialysis adequacy, such that most units now prescribe and monitor dialysis dose by urea kinetic methods. Applying the same methods to CAPD, it became apparent that adequacy in this mode was critically dependent on residual renal function. Unless dialysate volumes are increased, adequacy is compromised, sometimes to the point of technique non-viability, when residual renal function has been lost. This has been a major factor in the steady relative decline in the United Kingdom CAPD population during the 1990s (Fig. 1), which has been more than offset by increased centre-based haemodialysis provision. The automated peritoneal dialysis (APD) programme has also grown but remains small. During the last decade, practice in the United Kingdom appears to be converging on United States and European norms.

Changing demographics

The mean age of the dialysis population has increased considerably over the last two decades and is now around 61 years. One-third of all new patients in the United Kingdom are over 70. The elderly, most of whom will not receive transplants, account for most of the increased acceptance and prevalence rates, which now exceed 90 and 300 patients per million population, respectively.

The proportion of patients with non-renal comorbidities, particularly cardiovascular disease, has increased dramatically. Multiple pathologies are common: about 15 per cent of patients on United Kingdom programmes are diabetic, and the United States figure is nearer 50 per cent (see section 20.6). Many such patients have widespread micro- and macrovascular complications at the time of dialysis initiation. These changes have placed increased demands on nephrological and other specialist resources.

Technical aspects

The principles of dialysis

Dialysis is a physicochemical process allowing separation of the components of a complex solution by solute exchange across a semipermeable membrane. Such membranes act as molecular size-selective filters, the size threshold depending on the nature of the membrane. In haemodialysis the membrane is interposed between the patient's bloodstream and a rinsing solution (dialysis fluid). Diffusive and convective mass transfer takes place across the membrane, allowing changes in the composition of body fluid compartments. The rate of diffusive solute transport is dependent on flow rates, concentration gradients, and membrane characteristics. Convection involves the bulk movement of solvent and dissolved solute across the membrane. The driving force is transmembrane hydrostatic pressure, which can be adjusted by application of variable negative pressure to the dialysate side of the membrane. Solute transport (by solvent drag) is independent of diffusion. In general, convection contributes little to the clearance of rapidly diffusible small solutes such as urea (molecular weight 60), but can make a major contribution to the clearance of larger, poorly diffusible molecules such as b $_2$-microglobulin (molecular weight 11 200), provided the membrane is porous to so-called 'middle' molecules. Convective movement of water from blood across the membrane is known as ultrafiltration.

Membranes and dialysers

The original haemodialysis membranes were fashioned from regenerated cellulose, but technology has since proliferated and there are now three classes of membrane, though there is much overlap (Table 1). Membranes are arranged and supported in devices called dialysers to form separate paths for blood and dialysis fluid flow, usually in a hollow-fibre design. Dialysers are classified by design type, membrane composition, surface area, and permeability characteristics defined in terms of dialyser clearance (K_d) for a range of solutes and ultrafiltration coefficient (K_{uf}), which is the water flux per unit of transmembrane pressure. In contrast to cuprophane, high-flux synthetic membranes are highly permeable (high K_{uf} and high K_d for middle molecules), remove b_2-microglobulin and other potentially toxic middle molecules, and tend to be more 'biocompatible', meaning that they cause less activation of inflammatory cells, the complement cascade, and contact pathways, and less cytokine production. High-flux membranes, when employed in countercurrent mode to maximize diffusion, permit 'back-filtration' of dialysis fluid into blood, hence use of ultrapure water to prepare dialysis fluid is mandatory. Many would argue that this improves biocompatibility and is highly desirable anyway. Synthetic membranes are expensive and dialyser reuse is still an economic necessity in most high-flux programmes.

Dialysis water and fluids

Patients on haemodialysis are intimately exposed to huge quantities of water (300 litres in a single week compared with a standard weekly exposure of 15 litres). The potential for poisoning by waterborne impurities is significant. Aluminium and chloramines are examples of proven toxins, which must be removed. Bacterial and endotoxin contamination can produce acute problems. Ultrapurity is crucial in high-flux modes in which dialysis fluid is passively (back-filtration) or actively (on-line haemodiafiltration) infused directly into the patient. A combination of purification techniques are employed that include filtration, activated carbon adsorption, ion-exchange resin perfusion, reverse osmosis, and ultraviolet irradiation. Regular monitoring ensures chemical and microbiological standards are maintained. Acid and bicarbonate concentrates are then mixed with treated water in a single-patient proportionating system to produce dialysis fluid of the desired composition (Table 2). Regulation of dialysis fluid composition is the main tool to achieve a return to normal of electrolyte and mineral content and acid–base balance in body fluid compartments. Although there is great potential for individualization, a programme-standard composition is typically adopted, which may be varied in particular circumstances.

The dialysis machine and the extracorporeal circulation

Dialysis machines control and monitor much of the haemodialysis process and have crucial fail-safe functions. In the standard extracorporeal circuit, arterial blood is withdrawn from the arteriovenous fistula via the 'A' needle by a peristaltic pump (Fig. 3), circulated through the dialyser, through a bubble trap, and returned 'downstream' into the fistula through the 'V' needle. Heparin is infused downstream from the blood pump. The venous pressure monitor (*P*v) protects against blood loss from the circuit to the environment and detects downstream obstruction to flow. The bubble-trap level detector protects against air embolus. The arterial pressure monitor (*P*a) protects the fistula by detecting excessive negative pressure. Fail-safe mode activates the venous clamp and switches off the blood pump.

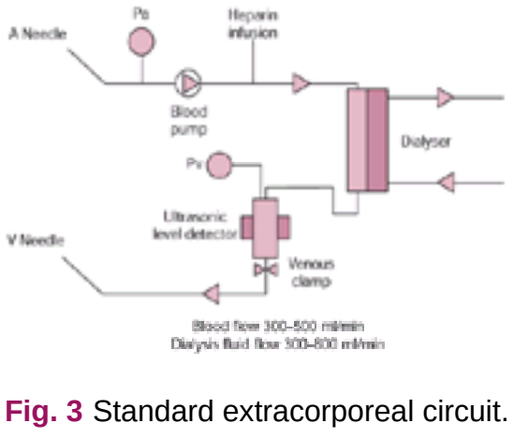

Fig. 3 Standard extracorporeal circuit. *P*a = arterial pressure detector, *P*v = venous pressure detector. A needle, arterial needle; V needle, venous needle.

Control of ultrafiltration

Modern dialysis machines utilize volumetric methods that permit precise control of ultrafiltration. A balancing system regulates dialysis fluid flow rates to and from the dialyser allowing for the removal of the required ultrafiltration volume, which is preset by the operator.

Anticoagulation

Routine anticoagulation with heparin, administered by intravenous bolus and subsequent infusion, is monitored by the whole-blood activated clotting time. Heparin-free dialysis, employing regular saline flushes of the circuit, is possible in high-risk patients. Prostacyclin is an expensive alternative.

Quantification and adequacy of dialysis

Predialysis blood urea and creatinine concentrations are poor indicators of dialysis adequacy. Low levels have been associated with increased mortality, suggesting that they are as (or more) likely to be due to reduced generation resulting from protein malnutrition and muscle wasting than due to increased clearance indicative of adequate dialysis. This was the lesson of the 1980s.

Urea kinetic modelling

The reanalysed data from the National Cooperative Dialysis Study (**NCDS**), still the only completed randomized study of the effect of haemodialysis dose on outcome, defined a new parameter of adequacy—the normalized urea clearance, *Kt*/*V*. This is a dimensionless parameter in which *K* is the urea clearance of the dialyser, *t* is the duration of dialysis in minutes, and *V* is the urea distribution volume, which approximates to total body water volume and is normally estimated from anthropomorphic data. The reanalysis showed that a *Kt*/*V* greater than 0.8 per dialysis was associated with good outcomes provided the patients were adequately nourished as defined by a normalized protein catabolic rate (**NPCR**) greater than 0.8 g/kg/day. NPCR can be calculated from urinary urea excretion in an interdialytic urine collection together with blood urea measurements taken at the start and finish of the collection (see equations in Appendix).

It is important to stress that the NCDS findings defined a minimum adequacy standard. The emerging view is that higher delivered *Kt*/*V*'s produce improved outcomes. It is not yet possible to say whether there is an upper threshold above which no further improvement is obtained. Current guidelines suggest a minimum target *Kt*/*V* of 1.2 to 1.3. The dose of dialysis prescribed can be adjusted to achieve the target *Kt*/*V* by changing the surface area of the membrane, blood flow rate, dialysis fluid flow rate (these influence urea clearance by the dialyser), and dialysis duration. This logic allows adequate dialysis to be delivered in a shorter time using larger dialysers and high flow rates (high-efficiency dialysis). It is important to have a means of monitoring the effectiveness of delivery of the prescribed dose.

Monitoring dialysis delivery using the urea reduction ratio

The simplest measurement of dialysis dose is the urea reduction ratio (**URR**), which is given by:

$$\mathrm{URR} = 100(1 - C_{post}/C_0)$$

where C_0 is the initial blood urea concentration, and C_{post} is the blood urea concentration in a blood sample taken immediately post-dialysis.

URR is a quality assurance tool, and cannot be used to prescribe dialysis dose. It takes no account of urea generation, ultrafiltration, or residual renal function, but does correlate with outcome testifying to its clinical utility. URR can also be converted to *Kt*/*V* (see Appendix).

Monitoring dialysis delivery using urea kinetic modelling

Assuming urea is distributed in a single pool within the body and that the effects of urea generation and ultrafiltration during dialysis are small, the blood urea concentration (C_t) at any time (t) during the dialysis is given by:

$$C_t = C_0 e^{-Kt/V}$$

Hence, the delivered dose of dialysis *Kt*/*V*, can be calculated from the expression:

$$Kt/V = \ln(C_0/C_{post})$$

The expression that corrects for urea generation and ultrafiltration during dialysis is more complex, and the single pool assumption is also an oversimplification. The rapid removal of urea (and other solutes) from the bloodstream during dialysis creates intercompartmental disequilibra. The intracellular concentration of urea exceeds the extracellular, and that in poorly perfused peripheral pools exceeds that in well-perfused body compartments. Urea exchange between these compartments continues after cessation of dialysis and causes a post-dialysis rebound of blood urea concentrations. This rebound can be substantial in high-efficiency treatments and can cause overestimation of *Kt*/*V* delivery by as much as 20 per cent, making the single pool assumption untenable in short high-efficiency treatments. There are a number of ways of dealing with this problem: the most straightforward is to delay the post-dialysis sample until rebound is complete (equilibrated post-dialysis sample), but this can be inconvenient (the patients want to leave the dialysis unit as soon as their treatment is completed). Much more complex is to model the system as two pools, requiring the assumption of a number of physiological parameters and iterative solution by computer. There are a number of less complex approximations, which are usually preferred and have been shown to produce equivalent results. The bottom line is that urea kinetic modelling can be used to prescribe the amount of dialysis necessary to attain the target *Kt*/*V*. The method has the flexibility to take account of residual renal function,

in which case the target *Kt*/*V* (total *Kt*/*V*) has dialysis ($K_d t/V$) and residual renal ($K_R t/V$) components.

Other approaches to adequacy

Urea clearance is the basis of most current methods of assessment of dialysis adequacy, in spite of the fact that urea transfer is not representative of the kinetics of most uraemic toxins. An alternative, the solute removal index (ratio of mass of solute removed by dialysis to the mass present at the start of dialysis) has a theoretical advantage over *Kt*/*V* in that it allows direct comparison of the adequacy of all treatment modalities. Larger molecules such as b $_2$-microglobulin are certainly toxic in dialysed patients but do not figure in our currently accepted notions of adequacy. Broader definitions are required. Urea kinetic modelling needs to be regarded as an essential component of a more global view of adequacy, which includes clinical as well as other laboratory data.

Initiation of dialysis

Blood urea and creatinine concentrations are poor indicators of dialysis adequacy and are subject to the same misinterpretation in the predialysis phase as endstage chronic renal failure is approached. This may lead to delay in dialysis initiation. The National Kidney Foundation Dialysis Outcomes Quality Initiative (**NKF-DOQI**) guidelines for dialysis initiation are thus based on urea kinetic modelling and suggest that dialysis should be started when weekly renal *Kt*/*V* is less than 2.0 unless the patient is asymptomatic, has a stable oedema-free body weight, and a normalized protein equivalent of urea nitrogen appearance (equivalent to NPCR) greater than 0.8. The guidelines have the benefit of aligning the approach to the assessment of severity of uraemia in predialysis and dialysis phases, but still lack a firm evidence base.

Incremental dialysis

This approach recognizes that the target total urea *Kt*/*V* has dialysis ($K_d t/V$) and residual renal ($K_R t/V$) components. As residual renal function declines during the first few years of dialysis, the dialysis component is gradually increased to ensure the target continues to be achieved. This allows a gentler initiation and maximizes the use of scarce resources. It does require regular estimates of residual renal function, and also assumes an equivalence of renal and dialyser clearance which holds for urea, but not necessarily for other solutes, or other renal functions. It is relevant that the viability of CAPD as a renal replacement modality also depends on this assumption.

Dry weight

Regulation of salt and water balance is one of the key functions of the kidney. Renal failure results in salt and water retention, which along with activation of the renin–angiotensin–aldosterone system, contributes to hypertension, left ventricular hypertrophy, and dilatation. These are potent causes of morbidity and mortality. 'Dry weight' is an important concept dating back to the early days of maintenance haemodialysis. It assumes that body weight at any time consists of two components: the dry weight or target weight, at which the patient's fluid compartments are normal in volume, and an excess weight consisting of surplus volume, which expands body fluid compartments and elevates blood pressure. The only way of defining dry weight is trial and error. The protocol requires cessation of antihypertensive agents and weight reduction during successive dialyses during the first few weeks or months after initiation. The dry weight is the point at which the patient is oedema free and below which hypotension occurs on further fluid removal. The implicit assumption is that patients on dialysis have normal cardiovascular responses, which may have been reasonable in the highly selected dialysis population of 1970 but is much less tenable in the older, sicker patients on dialysis today. Applying such principles, most of the early patients on dialysis became normotensive without the need for antihypertensive agents. In most current patients the target weight is likely to be the best achievable weight and hypertension is more likely. Shorter treatment times and less rigorous salt restriction have undoubtedly added to these difficulties.

Vascular access

The creation and maintenance of adequate, dependable, and robust vascular access is of vital importance to the continued well-being of patients on haemodialysis, and has rightly been referred to as their 'lifeline'.

Temporary access

For acute haemodialysis, temporary, non-cuffed, dual-lumen catheters can be inserted into femoral, internal jugular, or subclavian veins. The femoral route is simplest and preferred in the very sick patient, but the infection risk is high if femoral catheters are left *in situ* for more than a few days. Temporary catheters in other sites can remain for weeks, though use of the subclavian route risks stenosis of the vein and potentially compromises future permanent access in the ipsilateral arm.

Permanent access

Fashioning an arteriovenous fistula causes arterialization and expansion of the draining vein allowing its repeated puncturing for haemodialysis. A radiocephalic (Brescia–Cimino) fistula at the wrist is preferred, being less likely to produce distal limb ischaemia than proximal fistulas. Forearm vasculature is extremely vulnerable, especially when the patient is in hospital, and needs protecting in those destined for dialysis. Maturation of distal fistulas is slow, so these should be fashioned 3 to 6 months before planned initiation. It may be necessary to resort to other types of access including, in order of preference, an elbow brachiocephalic fistula, an arteriovenous graft composed of synthetic material (e.g. polytetrafluoroethene—PTFE), and cuffed tunnelled internal jugular venous catheters. Many patients still present late for dialysis and require primary central venous access by default. Tunnelled catheters are also required when other access options have been exhausted. Access failure is a significant cause of morbidity and mortality.

Recirculation

If there is a stenosis in the fistula severe enough to limit fistula blood flow to a level less than that demanded by the blood pump in the extracorporeal circulation, then blood returning from the dialyser to the fistula can be drawn directly from the 'V' needle to the 'A' needle and dialysed again. This is known as access recirculation, an effect that can also be produced by misplacement of fistula needles with the 'A' needle downstream to the 'V' needle. Also during dialysis a proportion of the blood returning through the 'V' needle will pass directly to the 'A' needle after passage through the heart and lungs without traversing a capillary bed to be 'replenished' with solute. This is known as cardiopulmonary recirculation and is an inevitable consequence of having a fistula as the access. The higher the blood pump speed the greater the degree of recirculation in all of these circumstances. Access recirculation is a major cause of underdelivery of prescribed dialysis dose, and unexplained reductions of monitored *Kt*/*V* or urea reduction ratios demand further investigation to exclude this. Significant recirculation (greater than 10 per cent), which can be detected and quantified by a variety of sampling and dilution methods (not described here), may require further investigation by Doppler ultrasonography or fistulography to define the lesion before attempting angioplastic or surgical correction.

Haemodialysis and related techniques

Conventional haemodialysis

'Conventional' refers to the use of low-flux cellulosic dialysers in standard circuits (Fig. 3). A decade ago the definition would also have included the use of acetate as buffer, but bicarbonate dialysis is now the norm.

High-flux haemodialysis

Concerns about the biocompatibility of cuprophane, and its poor clearance of middle molecules, especially b $_2$-microglobulin, have fuelled the increasing use of high-flux membranes, and concomitant investment in the use of ultrapure water for preparation of dialysis fluids.

Haemodiafiltration

Haemofiltration is a purely convective mode of treatment, which involves filtration of uraemic plasma and simultaneous infusion of replacement fluid. This greatly

improves middle molecule clearance. However, small molecule clearance is slow (Fig. 4), so the technique is not suitable for routine treatment of patients with chronic renal failure. It has, however, proved highly successful as a continuous treatment for patients with acute renal failure, particularly in the context of multiple organ failure in the intensive care unit (see Section 16). Adding a greater convective component (haemofiltration) to the diffusive and convective clearances offered by high-flux haemodialysis (a technique referred to as haemodiafiltration) allows the benefits of both these modalities to be maximized. The capacity to utilize dialysis fluid as the infusion fluid, made possible by use of ultrapure water in its preparation, means that the technique is economically viable (on-line haemodiafiltration, Fig. 5).

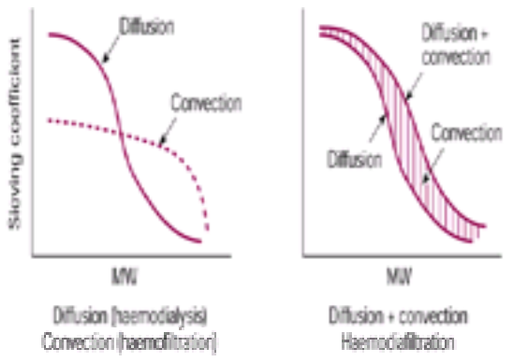

Fig. 4 Comparison of solute removal by diffusion and convection according to molecular weight. Convection has better 'middle molecule' clearance but much poorer clearance of small-molecular-weight solutes than diffusion. Haemodiafiltration combines the strengths of both techniques to broaden the spectrum of solute removal.

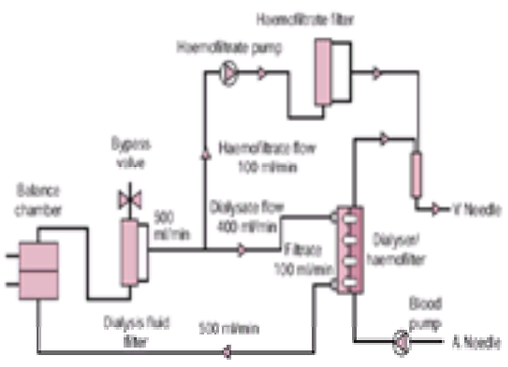

Fig. 5 Circuit for on-line haemodiafiltration. The haemofiltration replacement fluid is haemodialysis fluid pumped through an additional filter before infusion directly into the venous line. HDF, haemodiafiltrate. A needle, arterial needle; V needle, venous needle.

High-efficiency modalities

High-efficiency dialysers have high dialyser urea clearances and large surface areas. Dialysis treatments achieving high urea clearances (usually more than 200 ml/min) are referred to as high-efficiency treatments. Dialysis duration can be shortened by these means, often to below 3 h per session ('short dialysis').

High-efficiency and high-flux relate to different properties of dialysers and dialysis modes, which can be neither, either, or both.

Complications of haemodialysis

Acute complications

Hypotension

Symptomatic hypotension occurs in up to 30 per cent of dialysis sessions. Symptoms include nausea, vomiting, cramps, palpitations, dizziness, and syncope. The major cause is hypovolaemia, resulting from an imbalance between the rate of fluid removal from the circulation by ultrafiltration, and the rate of vascular refilling from the interstitium. Underlying cardiovascular disease, the use of antihypertensive drugs, autonomic dysfunction, and shortened dialysis times, increase the likelihood. The mainstays of management are careful assessment and reassessment of target weight, limited use of antihypertensive agents, reduction of interdialytic weight gain by fluid and sodium restriction, and reduction of ultrafiltration rate. Newer dialysis machines have the capacity to monitor relative blood volume and to profile the sodium concentration of dialysis fluid throughout the dialysis session. These techniques may be useful in particular situations, but neither is yet used routinely. Episodes of hypotension may occasionally have more sinister causes such as primary myocardial events and heparin-induced bleeding.

Disequilibration

The severest forms of disequilibration occur shortly after dialysis initiation (dialysis disequilibrium syndrome). The major predisposing factors are late presentation with severe uraemia and aggressive dialysis initiation with lengthy dialyses and high solute clearance rates. Restlessness, headache, tremors, fits, and coma can result. Dialysis should not be initiated in this way. Cerebral oedema due to fluid shifts induced by intercompartmental differences in urea concentrations and paradoxical cerebrospinal fluid acidosis are among the suggested causes. Post-dialysis headache is a common symptom in patients undergoing regular haemodialysis and may be a minor manifestation of disequilibrium.

Dialyser reactions

Anaphylactic (IgE-mediated) reactions occurring on first use of a dialyser are usually due to ethylene oxide sensitivity, which is used by manufacturers as a sterilizing agent. Reactions with reused dialysers are usually due to disinfectants, such as formaldehyde and peracetic acid, used in reprocessing. Bradykinin-mediated anaphylactoid reactions can occur in patients taking angiotensin-converting enzyme (ACE) inhibitors who are dialysed with polyacrylonitrile synthetic membranes (AN69). All the above occur within the first 20 min of treatment.

Pyrexias

Infected central lines are a potent cause of bacteraemia. Pyrogen reactions due to contaminated water became rare when rebuildable dialysers were replaced by disposable devices in the late 1970s.

Other complications

Use of modern fail-safe dialysis machines and ultrapure water systems has fortunately rendered a number of previously well-described complications exceedingly rare. These include air embolism, severe hypercalcaemia due to dialysis against hard water, and acute haemolysis.

Chronic complications

Hypertension and cardiovascular disease

There are many factors contributing to hypertension in patients on dialysis, including stimulation of the renin–angiotensin–aldosterone system and sympathetic overactivity, but the overriding factor is volume overload. Fluid status varies throughout the dialysis cycle and so does blood pressure. Just what constitutes hypertension in those on dialysis is not well defined, but in most units 60 per cent or more of the patients receive antihypertensive agents, although in other units this is less common. This may be explained by differences in definition and case mix, but other factors are undoubtedly important too, particularly the emphasis placed on the maintenance of optimum sodium balance and adequate ultrafiltration. In some units there may be a tendency to compromise, perhaps too soon, and use drugs, especially in patients with cardiovascular and other comorbidities, making dry weight even more difficult to achieve. Getting this right is crucial since hypertension is an important risk factor for cardiovascular disease, which is the major cause of the excess mortality in the dialysis population.

Volume overload, hypertension, anaemia, hyperparathyroidism, excessive fistula flow rates, and uraemia itself all predispose to left ventricular hypertrophy, which is an independent risk factor for mortality. Correction of anaemia can favourably influence the natural history of left ventricular hypertrophy in patients on dialysis, and early use of erythropoietin in the predialysis period may prevent it. Patients on dialysis also have a variety of lipid abnormalities, hyperhomocysteinaemia, increased oxidative stress, and elevated inflammatory markers, all of which may predispose to cardiovascular disease.

Anaemia

Erythropoietin deficiency is the major cause of anaemia in patients on haemodialysis. The introduction of recombinant erythropoietin in the early 1990s has redefined the uraemic syndrome, in the sense that many debilitating 'uraemic' symptoms can be remedied by successful treatment of anaemia (see Chapter 20.6.2 for further discussion).

A number of additional causes of anaemia may arise from the dialysis process itself. The most common is iron deficiency, which results from the repeated loss of small amounts of blood. Regular iron replacement, preferably given intravenously, is necessary since iron deficiency is a potent cause of resistance to exogenous erythropoietin. Deficiencies of other haematinics can occur, particularly in high-flux treatments and regular supplementation with vitamin B_{12} and folate is recommended. Mechanical and chloramine-induced haemolysis should not occur with modern techniques. There are other causes of erythropoietin resistance, the most potent being infection, often arising from central venous lines in this context, also severe hyperparathyroidism and underlying malignancy.

Bone disease

For discussion of renal bone disease, see Chapter 20.5.2.

Amyloidosis

Dialysis-related amyloidosis is a serious complication of chronic dialysis. The incidence increases with duration of haemodialysis and symptomatic involvement is almost universal after 15 years. Older patients are more susceptible. The syndrome manifests mainly as carpel tunnel syndrome and destructive arthropathy associated with bone cysts, but other organs can be involved. Deposits of amyloid, mainly composed of b_2-microglobulin fibrils, can be found at these and other sites. b_2-Microglobulin is an 11 200-Da protein that is part of the human class 1 major histocompatibility complex. It is 95 per cent eliminated by glomerular filtration, hence levels are elevated in renal failure. Low-flux membranes do not clear b_2-microglobulin but clearance occurs by a combination of convection and adsorption, using high-flux membranes. Elevated plasma levels are the major predisposing factor to amyloid deposition. Other factors may also be important, including modification of b_2-microglobulin by advanced glycation end-products and by oxidative and carbonyl stress. It is also possible that b_2-microglobulin or modified b_2-microglobulin is directly toxic to tissues. Use of high-flux synthetic membranes, especially in haemodiafiltration mode (Fig. 5), reduces plasma levels of b_2-microglobulin, although the level remains about tenfold higher than in those with normal renal function. High-flux dialysis and especially haemodiafiltration may prevent or delay the onset of symptomatic disease. Use of ultrapure water may also be protective. Treatment options are limited for established disease, but renal transplantation may enable slow resorption of deposits.

Patient management

Infection control

Strict adherence to universal precautions is necessary to minimize the risk of cross-infection by bloodborne viruses. Transmission from contaminated external surfaces, rather than through the dialyser membrane, is the major cross-infection threat. Screening of patients about to start dialysis for evidence of prior infection with hepatitis B and C is routine, and should be repeated at least 6 monthly thereafter. Patients negative for hepatitis B surface antigen should be vaccinated: positive patients should be segregated and use a dedicated machine. Patients positive for hepatitis C or HIV should be managed similarly.

Dialysis prescription and monitoring of dialysis delivery

There are a number of elements to the dialysis prescription that should be regularly and systematically reviewed. Target Kt/V is normally 1.2 to 1.3. Patients with intercurrent illness may require more. The value of K is obtained from data sheets from the dialyser manufacturer, taking into account membrane area, and blood and dialyser flow rates. The value of V is obtained empirically from age, sex, weight, and height. If the target Kt/V includes a component for residual renal function then residual urea clearance should be measured monthly and dialysis time readjusted accordingly. Dialysis delivery should be monitored monthly. Prescription and monitoring protocols are now often computer based. Inefficient delivery requires prompt investigation to exclude problems such as access recirculation. Membrane type and flux should be specified. Use of high-flux synthetic membranes is increasingly standard rather than targeted by evidence of amyloid deposition. Setting the dry weight allows the ultrafiltration requirement to be specified for each dialysis. Regular reassessment of dry weight is required, especially during intercurrent illness, when loss of flesh weight predisposes to covert fluid overload. The prescription should also refer to the heparin loading dose and maintenance infusion rate.

Control of hypertension

Predialysis blood pressure measurements may be misleading and cause overdiagnosis and overtreatment of hypertension. Measurements taken 20 min post-dialysis, allowing time for compartmental re-equilibration, are a better reflection of interdialytic ambulatory readings. Target levels have not been defined. We suggest targeting a 20-min post-dialysis reading of 135/85 in patients less than 65 years old. Drug therapy of hypertension is second-line treatment to be deployed after the achievement of optimal fluid status. No class of antihypertensive agent is contraindicated in dialysis patients, although the half-life of renally excreted drugs may be markedly prolonged and care is required with dosing schedules.

Access care and surveillance

Regular clinical examination of fistulas, serial measurement of static and dynamic venous pressures during dialysis, and monitoring of access flow rates by ultrasound dilution may enable access problems to be detected and corrected before the risks of underdialysis supervene.

Diet and nutrition

Malnutrition is common and is usually due to underdialysis. The early signs are subtle and often masked by fluid overload, which can compound the problem. There are no simple, fool-proof laboratory tests. Serum albumin reflects the presence of inflammation as much as it reflects malnutrition. Monitoring of NPCR (normalized protein catabolic rate) may be helpful but cannot replace regular dietetic review. When malnutrition is identified a range of oral nutritional supplements may be deployed. There is a limited role for intradialytic parenteral nutrition. Protein requirements in patients on haemodialysis are not well characterized but have been estimated at 1.2 g/kg ideal body weight/day, which is considerably greater than the non-uraemic requirement. Intakes greatly in excess of this may cause problems unless dialysis dose is correspondingly increased. There is no role for protein restriction. The energy requirement of a moderately active patient on haemodialysis is about 35 kcal/kg body weight/day, which is similar to that of normal subjects.

Attempts should be made to limit interdialytic fluid gains to 1 to 2 litres. This is difficult when residual renal function has been lost. The value of limiting sodium intake (40 to 80 mmol/day) to control thirst is often understated. Potassium restriction (to about 60 mmol/day) is usually required when residual renal function is minimal. The recommended intake of elemental calcium is 1 to 1.5 g/day. Achieving this is seldom difficult given the extensive use of calcium salts as phosphate binders. Phosphate restriction to about 0.8 g/day of elemental phosphorus may reduce the requirement for these agents. There is no consensus on the need to supplement water-soluble vitamins (B and C), but the practice is widespread. Vitamin B_{12} supplements are recommended in high-flux treatments.

Anaemia

Treatment with recombinant erythropoietin has become the mainstay of anaemia management, which is often nurse-led and protocol-driven. Uncertainties remain about the optimal target haemoglobin, recommendations varying between 10 and 12 g/dl. In the dialysis population the use of erythropoietin has unmasked a huge requirement for intravenous iron supplementation. Many patients have a state of absolute or functional iron deficiency, which is difficult to diagnose in those on dialysis. NKF-DOQI recommendations that serum ferritin levels less than 100 ng/ml and transferrin saturation less than 20 per cent are indicative of iron deficiency are reasonably sensitive when taken together, but poorly specific. Other parameters such as the percentage of hypochromic red cells and the reticulocyte haemoglobin content may be useful, but the best test appears to be the pragmatic one of response to intravenous iron. Oral iron is usually ineffective, probably because intestinal iron absorption—already low in endstage chronic renal failure—fails to respond to the stimulus of erythropoiesis, and is further suppressed by high tissue iron stores. Intravenous iron saccharate is used in moderate doses (such as 100 mg on each of 10 successive dialyses) to correct iron deficiency, and as maintenance treatment (such as 50 mg weekly) provided 3-monthly serum ferritin levels remain below about 800 ng/ml. (See Chapter 20.6.2 for further discussion of the use of erythropoietin.)

Bone disease

The goals of treatment are to maintain optimal bone structure and function and prevent metastatic calcification. The short-term surrogate is to maintain biochemical parameters in their target ranges. Serum calcium should be maintained in the normal range. The target range for phosphate is ill-defined and there is confusion about what is desirable and what is achievable. The product of calcium and phosphate (in mmol/l) should be less than 5. Serum parathyroid hormone (**PTH**) levels should be maintained at about two to three times the upper limit of normal. Calcium carbonate and calcium acetate are the phosphate binders in common use. Hypercalcaemia is a real risk. Sevelamer, a new polymeric phosphate binder that does not contan metal ions should be less toxic in this regard. Calcitriol and a-calcidol are used mainly to suppress PTH secretion. When used with calcium salts, the risks of hypercalcaemia are multiplied and careful monitoring is required. Pulsed therapy (usually thrice weekly on dialysis days) may improve the therapeutic ratio, but the benefits of intravenous administration have been overstated. Failure to control hyperparathyroidism still necessitates parathyroidectomy in significant numbers of patients. On the other hand, oversuppression of PTH levels should be avoided, being a risk factor for adynamic bone disease and enhanced metastatic calcification. (See Chapter 20.5.2 for further discussion.)

Other aspects

Selection and preparation of patients on dialysis for transplantation is a crucial aspect of care, which now usually includes rigorous assessment of cardiovascular fitness.

Cardiovascular risk factors such as smoking and lipid disorders should be addressed, and exercise encouraged, both during dialysis sessions and in general. Low-dose aspirin may be beneficial. Folate and vitamin B supplements may reduce elevated homocysteine levels. There is a high incidence of sexual dysfunction, especially in males: some may benefit from androgen replacement, sildenafil may be effective if not contraindicated, skilled counselling may be helpful.

Outcomes

Dialysis undoubtedly prolongs the life of patients with endstage chronic renal failure, but survival remains markedly inferior to that of age-matched peers with normal renal function. Cardiovascular disease is the main cause of death, followed by infection. Comparison of outcome in the different eras of dialysis is fraught with problems, largely because of the dramatic differences in case mix of patients entering programmes. Age, comorbidity, and functional status are independent predictors of morbidity (rate of admission to hospital) and mortality. Late presentation for dialysis has a profound effect on survival; late planned initiation may also have an effect, perhaps mediated through malnutrition; and we have previously alluded to the effects of dialysis adequacy and nutrition on outcome. The relationship between post-dialysis systolic blood pressure and mortality is 'U' shaped, as in the normal population, those with the lowest pressures faring poorly, probably due to coexisting heart failure.

It is difficult to compare the outcome of patients treated with haemodialysis and peritoneal dialysis in any meaningful way. Data from single centres, multicentre studies, and analysis of registry data do not show consistent differences in survival between these modalities. There are a number of confounding factors. Patients initiated on CAPD are younger, have less coexisting non-renal comorbidities, better functional status, and are less likely to have presented late. In addition, technique survival is poor in CAPD, and many patients require transfer to haemodialysis because of peritonitis, inadequate dialysis, or ultrafiltration failure. Haemodialysis can be regarded as the default mode of renal replacement therapy. Quality of life assessments are similar in both groups, but both are inferior to those obtained in patients with successful transplants. It is probably safe to conclude that, in the early years of therapy at least, morbidity and mortality are similar on both modalities if risk-stratified groups are compared.

There are few data to allow comparison of the outcome of conventional haemodialysis and more modern haemodialysis modes. Registry data suggests better survival with more biocompatible membranes, possibly related to enhanced clearance of larger solutes. Evidence is hardening that high-flux modes protect against the development of dialysis-associated amyloidosis.

The future

Haemodialysis is likely to remain centre based for the majority. Technical advances will allow treatments to become more tailored to the specific requirements of the individual. On-line dialysis quantification could guarantee the adequacy of each session. On-line blood volume monitoring coupled with algorithms to control ultrafiltration rate, dialysis fluid temperature, and sodium content, on a minute-to-minute basis, could prevent intradialytic hypotension and allow patients to finish dialysis at their optimal achievable weight. The encouraging results with daily dialysis suggest that this may emerge as a home, self-supervised modality, perhaps for a small proportion of younger, less dependent patients for whom transplantation or retransplantation is not an option. Vascular access will remain the Achilles' heel.

Appendix

$$\mathrm{NPCR} = 148.7(G/V + 0.17),$$

where G is the urea generation rate given by:

$$G/V = [C_{pre2}(V + w_g)/V - C_{post} + (V_u \times U_u)/V]/t_{id}$$

Where C_{pre2} = predialysis blood urea concentration before succeeding dialysis

w_g = interdialytic weight gain

V_u = volume of interdialytic urinary collection

U_u = urinary urea concentration

t_{id} = duration of intradialytic urine collection

$$Kt/V = -\ln(\text{URR} - 0.008t) + (4 - 3.5\text{URR}) \times (W_2 - W_1)/W_2$$

Where t = duration of dialysis

W_1 = predialysis weight

W_2 = post-dialysis weight

Further reading

Bergstrom J (1993). The nutritional requirements of hemodialysis patients. In: Mitch WE, Klahr S, eds. *Nutrition and the kidney*, 2nd edn, pp 263–89. Little, Brown and Company, Boston.

Block GA, Port FK (2000). Re-evaluation of the risks associated with hyperphosphatemia and hyperparathyroidism in dialysis patients: recommendations for a change in management. *American Journal of Kidney Diseases* **35**, 1226–37.

Chandna SM *et al.* (1999). Factors affecting survival and morbidity on chronic dialysis. Is there a rationale for rationing? *British Medical Journal* **318**, 217–23.

Drucker W (1979). Haemodialysis: a historical review. In: Drucker W, Parsons FM, Maher JF, eds. *Replacement of renal function by dialysis*, pp 3–37. Martinus Nijhoff, The Hague.

Gokal R (1993). Quality of life in patients undergoing renal replacement therapy. *Kidney International* **40** (Suppl 8), S23–7.

Hirsch DH (1989). Death from dialysis termination. *Nephrology, Dialysis, Transplantation* **4**, 41–4.

Keshaviah P, Star RA (1994). A new approach to dialysis quantification: an adequacy index based on solute removal. *Seminars in Dialysis* **7(2)**, 85–9.

Khan IH *et al*. (1993). Influence of coexisting disease on survival on renal replacement therapy. *Lancet* **341**, 415–18.

Lameire N, Van Biesen W (1999). The pattern of referral to the nephrologist: a European survey. *Nephrology, Dialysis, Transplantation* **14** (Suppl.6), 16–23.

Leypoldt JK *et al.* (1999). Effect of dialysis membranes and middle molecule removal on chronic hemodialysis patient survival. *American Journal of Kidney Diseases* **33**, 349–55.

Locatelli F *et al.* (1996). The effects of different membranes and dialysis technologies on patient treatment tolerance and nutritional parameters. The Italian Cooperative Dialysis Study. *Kidney International* **50**, 1293–302.

Lowrie EG, Lew NL (1990). Death risk in haemodialysis patients: the predictive value of commonly measured variables and an evaluation of death rate differences between facilities. *American Journal of Kidney Diseases* **15**, 458–82.

Mitra SM, Chandna SM, Farrington K (1999). What is hypertension in chronic haemodialysis? The role of interdialytic blood pressure monitoring. *Nephrology, Dialysis, Transplantation* **14**, 2915–21.

National Kidney Foundation (1997). NKF-DOQI clinical practice guidelines for hemodialysis adequacy. *American Journal of Kidney Diseases* **30** (Suppl 2), S15–66.

National Kidney Foundation (1997). NKF-DOQI clinical practice guidelines for vascular access. *American Journal of Kidney Diseases* **30** (Suppl 3), S150–91.

Owen WF *et al*. (1993). The urea reduction ratio and serum albumin concentration as predictors of mortality in patients undergoing haemodialysis. *New England Journal of Medicine* **329**, 1001–6.

Renal Association (1997). *Treatment of adult patients with renal failure. Recommended standards and audit measures*. Royal College of Physicians, London.

Schiffl H *et al.* (2000). Clinical manifestations of AB-amyloidosis: effects of biocompatibility and flux. *Nephrology, Dialysis, Transplantation* **15**, 840–5.

Tattersall JE, Greenwood RN, Farrington K (1995). Urea kinetics and when to commence dialysis. *American Journal of Nephrology* **15**, 283–9.

Tattersall JE, Farrington K, Greenwood RN (1998). Adequacy of dialysis. In: Davison AM, *et al*., eds. *Oxford textbook of clinical nephrology*, 2nd edn, pp. 2075–87. Oxford University Press, Oxford.

Vonesh EF, Moran J (1999). Mortality in end-stage renal disease. A reassessment of differences between patients treated with hemodialysis and peritoneal dialysis. *Journal of the American Society of Nephrology* **10**, 354–65.

20.6.2 The treatment of endstage renal disease by peritoneal dialysis

Paul F. Williams

Introduction

Peritoneal dialysis has been an established treatment modality for acute and chronic renal failure since the early 1960s, but it was not until 1976 with the description of peritoneal equilibration dialysis (later to become known as 'continuous ambulatory peritoneal dialysis'— **CAPD**) by Popovich and Moncrieff that the technique was popularized. By the end of the twentieth century some 15 to 20 per cent of all chronic dialysis patients worldwide were being treated with either CAPD or automated peritoneal dialysis (**APD**). Peritoneal dialysis plays an important role in the integrated care of the patient with endstage renal disease along with haemodialysis and renal transplantation.

Practical aspects of peritoneal dialysis

The process of peritoneal dialysis involves the instillation of dialysis fluid into the peritoneal cavity via a dialysis catheter, thereby allowing the fluid to come into contact with the uraemic blood in the patient's peritoneal capillaries. Dialysis takes place through the diffusion of uraemic toxins down a concentration gradient from capillary blood to dialysis fluid, with fluid removal from the circulation being achieved by varying the concentrations of osmotic agents (usually glucose) in the dialysis fluid. Once the uraemic toxins in the patient's blood have equilibrated with the dialysis fluid then this can be drained out and replaced manually (CAPD) or by machine (APD).

Peritoneal dialysis solutions

The most commonly used dialysis solutions today contain varying concentrations of glucose as the osmotic agent, along with a balanced electrolyte solution using lactate as a buffer to correct uraemic acidosis. Although these solutions have been in widespread use for many years they are recognized to have a number of disadvantages, including low pH and hyperosmolality, which depends on the glucose concentration. These make the dialysis fluids relatively bioincompatible and may be responsible for long-term membrane damage with mesothelial cell loss and glycation of the membrane. A variety of new peritoneal dialysis solutions have been introduced in recent years in an attempt to improve biocompatibility, membrane viability, fluid removal, and malnutrition. These include: (1) solutions using a bicarbonate/lactate mixture instead of lactate alone as the buffer; (2) a glucose polymer-based fluid that allows reduced membrane exposure to glucose and allows slow prolonged ultrafiltration to take place; and (3) an amino acid-based solution which may help to correct hypoalbuminaemia and malnutrition.

Peritoneal dialysis catheters

Access to the peritoneal cavity can be achieved on a semipermanent basis using a silastic Tenckhoff catheter. A variety of modifications on the basic design exist but none has conclusively been shown to be superior. The intraperitoneal portion of the catheter can be straight or coiled and has side and end holes; the subcutaneous portion provides anchorage and a barrier against infection by means of tissue ingrowth into two Dacron cuffs, one placed preperitoneally in the rectus sheath and one subcutaneously.

Insertion techniques also vary: the catheters may be placed percutaneously via a trochar and cannula, via a laparoscope, or with a formal mini-laparotomy. The success or otherwise of the insertion procedure seems to depend on the skill and experience of the operator rather than any intrinsic advantage for one method of insertion, although a formal surgical procedure is warranted in the presence of significant obesity or suspected adhesions after previous surgical operations. Depending on the urgency of starting dialysis, the insertion technique, and local practice, a variable amount of time is usually allowed to elapse prior to starting peritoneal dialysis to allow satisfactory wound healing to take place. The catheters may remain *in situ* for many years if needed, the main reasons for removal being infection, transfer to haemodialysis, or successful renal transplantation.

Peritoneal membrane function

The transport of small molecular weight uraemic toxins across the peritoneal capillary membrane is governed by its permeability to such solutes. This can be assessed by means of the Peritoneal Equilibration Test (PET) introduced by Twardowski. Instilling a 2.27 per cent glucose-based dialysis solution into the peritoneal cavity and taking samples of blood and dialysate for glucose and creatinine measurements over a 4-h period enables the membrane transport characteristics to be classified as low, low-average, high-average, or high with respect to glucose and creatinine concentrations. A high-transporter status will enable rapid equilibration of urea and creatinine, thus enabling adequate solute removal, but the increased membrane permeability prevents the maintenance of the glucose gradient that is essential for adequate fluid removal. Thus patients will achieve satisfactory small-solute clearance, but may have problems with adequate fluid removal. In a similar fashion, patients with a low-transporter membrane will sustain an adequate glucose gradient and achieve adequate ultrafiltration, but poor equilibration of urea and creatinine will leave them at risk of underdialysis. Knowledge of a patient's membrane function soon after starting dialysis can therefore aid in the rational prescription of peritoneal dialysis and may also have prognostic importance, particularly in large patients with declining residual renal function.

Prescription of peritoneal dialysis

CAPD

In the early years of CAPD many patients received so-called standard peritoneal dialysis prescriptions—4 × 1.5- or 2.0-litre exchanges per day—with little or no regard for body weight, membrane function, or level of residual renal function. With increasing experience worldwide and recognition of the need to attain small-solute clearance targets for maximum patient well being and survival, it has become obligatory to provide an individualized dialysis prescription for each patient.

When patients first begin peritoneal dialysis the great majority will have some degree of residual renal function, which may provide some 25 to 30 per cent of the required small-solute clearance. Under these circumstances most will achieve adequate dialysis with standard CAPD or APD prescriptions. Once residual renal function has been lost after 2 to 3 years on dialysis, then it becomes increasingly important to individualize the patients' prescription to enable adequate dialysis to be delivered. In general, patients who have low-transporter membranes will require high-volume exchanges on CAPD to achieve adequate small-solute clearance. As they still maintain adequate glucose gradients they still have adequate ultrafiltration. Patients with high-transporter membranes will often still achieve adequate

small-solute removal, but they will be unable to obtain adequate ultrafiltration to keep them oedema-free unless short frequent exchanges are used. Under these circumstances, if patients do not achieve adequate dialysis in terms of small-solute clearance and ultrafiltration with the above prescription modification, transfer to haemodialysis may be required.

Table 1 summarizes the problems seen with high body-weight anuric patients.

APD

APD makes use of a machine to drain and fill the abdomen with dialysis fluid overnight, usually leaving the patient relatively free from the need to perform any dialysis-related activities during the day. The modality of APD may therefore be particularly suitable for certain groups of patients at the beginning of dialysis, for example the young, or those still at school or work. The increased quality of life achievable with APD is an added benefit in these patient groups, which is partly the reason why APD is the fastest growing renal replacement modality worldwide.

When a patient starts on APD as an initial dialysis modality, the presence of significant residual function may allow the use of relatively low dialysate volumes. However, as the patient's residual renal function declines then the need to reach small-solute and salt and water-removal targets will require the use of larger fill volumes, with the length of overnight dwells matched to peritoneal membrane function. At the other end of the spectrum, APD can also be used to prolong a patient's time on peritoneal dialysis when it is used as a salvage therapy for those no longer adequately dialysed on CAPD. But in some patients with membrane failure or some other reason for inadequate dialysis, then elective transfer to haemodialysis is the appropriate course of action.

Outpatient monitoring of the patient on peritoneal dialysis

Once the patient and their family have been trained to perform peritoneal dialysis in the community, then outpatient review need not be that frequent—perhaps monthly initially, and as time passes and confidence increases the intervals between outpatient visits can be increased up to once every 2 to 3 months. At each clinic visit the patient should be assessed by the nephrologist, dialysis nurse, and dietician, with the aim of ensuring that they are compliant with the dialysis regime and adhering to dietary and fluid balance recommendations.

Adequacy of dialysis

As mentioned previously, the need for individualized dialysis prescriptions is paramount. The CANUSA study, and others, have documented the links between adequacy of dialysis, in terms of small-solute clearances, and patient survival, Hence, it is important to monitor the urea and creatinine clearances achieved by the combination of dialysis and residual renal function on a regular basis—at least yearly. If the clearances fall below recommended targets then changes in the dialysis prescription may be needed to keep the patient healthy. Table 2 shows the targets recommended for urea clearance (*K*t/v urea) and creatinine clearance (Cr Cl) per week.

Nutrition

Malnutrition is a common problem in dialysis patients and may be seen in up to 30 per cent of those on peritoneal dialysis. This may be due to appetite suppression caused by glucose absorption from the dialysate, the presence of dialysis fluid in the abdomen, uraemia secondary to underdialysis, or peritoneal protein losses. If the patient is unable to maintain an adequate protein (1–1.2 g/kg body weight per day) and calorie (30–35 kCal (125–146 joules)/kg body weight per day) intake then malnutrition and hypoproteinaemia will develop with increased risk of death. Dietary advice from an experienced dietician is invaluable under these circumstances.

Cardiovascular status

Up to 50 per cent of all deaths in patients on peritoneal dialysis will be from cardiovascular causes, and therefore attention to the control of risk factors may be of benefit. It is important to encourage patients to stop smoking, take exercise, and to control fluid overload and hypertension by a combination of fluid restriction, fluid removal by dialysis, and the use of antihypertensive drugs where appropriate. Control of hyperlipidaemia by diet and drug therapy is also indicated.

Anaemia

The anaemia of renal failure is readily treatable with recombinant erythropoietin and, providing iron stores are kept at adequate levels, it should be possible to keep the majority of patients at or above target haemoglobin levels. This will improve the quality of life, improve exercise tolerance, and may also reduce cardiovascular mortality.

Infective complications

Exit site infections

Bacterial infection around the catheter exit site is a frequent occurrence, which, if neglected or inadequately treated, may lead to peritonitis and/or the need for catheter removal. The majority of these infections are caused by Gram-positive organisms such as *Staphylococcus aureus* and *S. epidermidis*, with occasional episodes being caused by Gram-negative organisms, for example *Pseudomonas* spp.

Meticulous exit site care, with attention to catheter immobilization locally and regular cleaning of the exit site with antiseptics, may diminish the frequency of such infections. Local application of the antibacterial agent mupirocin on a regular basis may also help eradicate *S. aureus* carriage and consequent infections. However, once erythema, purulent discharge, and pain develop, then broad-spectrum antibiotic therapy is indicated. If this is not successful in eradicating catheter sepsis then it may be necessary to replace the catheter.

Peritonitis

Peritonitis remains the major infective complication of peritoneal dialysis, and either acute or repeated episodes of peritonitis are a common cause of transfer from peritoneal dialysis to haemodialysis.

Common causative organisms include *S. aureus*, *S. epidermidis*, coliforms including *Pseudomonas* spp., and, rarely, fungal organisms. The diagnosis of peritonitis is usually straightforward, being based on the presence of cloudy dialysate fluid with or without abdominal pain. Treatment should be initiated on clinical suspicion, before laboratory culture results are available, and should cover both Gram-positive and Gram-negative organisms. Recommended regimes include intraperitoneal cefazolin or vancomycin (to give Gram-positive cover) with an intraperitoneal aminoglycoside or oral quinolone such as ciprofloxacin (to give Gram-negative cover), with subsequent modification of therapy depending on the culture results. The use of vancomycin has declined over recent years because of worries concerning the development of resistant organisms.

The frequency of episodes of peritonitis in patients on peritoneal dialysis has been falling with increasing experience and technological advances, including the use of disconnect systems and APD machines. It should be possible to reduce the frequency of peritonitis to one episode every 2-patient years or better. If peritonitis does develop then the cure rate without removing the dialysis catheter should exceed 85 per cent, and the majority of patients should be able to continue with peritoneal dialysis after the episode has resolved, unless it was caused by bowel perforation or a fungal peritonitis. A successful continuation of peritoneal dialysis is unlikely in both circumstances.

Transplantation

Patients who are established on peritoneal dialysis may be transplanted safely, and there is evidence that they have a lower incidence of delayed graft function and early rejection. The peritoneal dialysis catheter may be used after transplantation if dialysis is required, provided that the peritoneum has not been breached. It is

usually removed electively some 2 or 3 months later, when the risk of graft failure has diminished.

Survival data

Studies over the years comparing both technique and survival in patients on haemodialysis and peritoneal dialysis have produced conflicting results. In the absence of a randomized controlled trial the differences in the case mix of patients on the two modalities makes interpretation of the data difficult. It seems likely that providing attention is given to monitoring and adjusting dialysis therapy as residual renal function falls, then the two therapies are equivalent and most patients should be allowed to choose the modality that suits them best. There is some evidence that peritoneal dialysis is the more suitable first modality as residual renal function is better preserved with this technique, but once this advantage is lost after 3 to 5 years then patients are more likely to require haemodialysis as a long-term treatment.

The two techniques should be seen as complementary and both should be available to the patient waiting for renal transplantation. The fact that worldwide peritoneal dialysis varies from less than 5 per cent to more than 60 per cent, with a world average of around 15 per cent, probably reflects factors other than patient selection.

Summary

Peritoneal dialysis has been a well-recognized mode of therapy for endstage renal disease since the mid-1970s. Approximately 15 per cent of the world's dialysis population are currently kept alive and healthy by this technique. Patient survival in the short to medium term with peritoneal dialysis is equivalent to that seen with haemodialysis. In the past, the main complications of peritoneal dialysis have been peritonitis, inadequate dialysis, malnutrition, membrane failure, and patient 'burn-out'. With increasing experience, individualized dialysis prescription, and technological advances including new solutions and improved automated peritoneal dialysis machines, the frequency of these complications is diminishing. Peritoneal dialysis should be seen as part of an integrated care package for patients with endstage renal disease, indeed—there being no significant difference between haemodialysis and peritoneal dialysis—the patient should be encouraged to choose the dialysis technique that fits best with their lifestyle. There may be advantages in starting dialysis with peritoneal dialysis and then, as residual renal function declines, moving to haemodialysis while waiting for a renal transplant to become available.

Further reading

CANUSA Peritoneal Dialysis study group (1996). Adequacy of dialysis and nutrition in continuous peritoneal dialysis: association with clinical outcomes. *Journal of the American Society of Nephrology* **7**, 198–207. [Important trial documenting the link between adequacy of dialysis, nutrition, and patient survival]

Diaz-Buxo JA, Suki WN (1994). Automated peritoneal dialysis. In: Gokal R, Nolph KD, eds. *Textbook of peritoneal dialysis*, pp 399–418. Kluwer Academic, Dordrecht. [Comprehensive review of the history, application, and outcomes of APD]

Fenton SSA, *et al*. (1997). Hemodialysis versus peritoneal dialysis: a comparison of adjusted mortality rates. *American Journal of Kidney Diseases* **30**, 334–42. [Report from Canadian registry showing that peritoneal dialysis and haemodialysis are at least equivalent over the first 5 years of treatment]

Keane WF, *et al*. (1996). Peritoneal dialysis related peritonitis treatment recommendations; 1996 update. *Peritoneal Dialysis International* **16**, 557–73. [Consensus statement by panel of international experts]

Lameire NH (1997). The impact of residual renal function on adequacy of peritoneal dialysis. *Nephron* **77** (1), 13–28. [Important article documenting the importance of residual renal function to adequacy of peritoneal dialysis]

National Kidney Federation (1997). DOQI clinical practice guidelines for peritoneal dialysis adequacy. *American Journal of Kidney Diseases* **30** (Suppl. 2), S67–134. [North American consensus statement on peritoneal dialysis prescription guidelines]

Popovich RP, *et al*. (1978). Continuous ambulatory peritoneal dialysis. *Annals of Internal Medicine* **88**, 449–52. [First description of the technique of CAPD]

Shockley TR, Martis L, Tranaeus AP (1999). New solutions for peritoneal dialysis in adult and paediatric patients. *Peritoneal Dialysis International* **19** (Suppl 2), S429–434. [Review article discussing the role of the new peritoneal dialysis solutions available]

Twardowski ZJ, *et al*. 1987 Peritoneal Equilibration Test. *Peritoneal Dialysis Bulletin* **7**, 128–47. [Important article describing how to perform the PET test and its importance to PD prescription]

Van Biesen W, *et al*. (2000). An evaluation of an integrative care approach for endstage renal disease patients. *Journal of the American Society of Nephrology* **11** (1), 116–25. [Important single-centre report documenting excellent patient survival with integrated PD and HD therapies]

Young G, *et al*. (1991). Nutritional assessment of CAPD: an international study. *American Journal of Kidney Diseases* **17**, 462–71. [International study of the prevalence and causes of malnutrition in CAPD patients]

20.6.3 Renal transplantation

P. Sweny

Introduction

Renal transplantation is the preferred option for the treatment of endstage chronic renal failure in patients for whom there are no major medical contraindications. With improvements in immunosuppression and in the equally important general medical support of the immunocompromised patient, the age ranges and permissible comorbidities continue to be extended. In well-selected recipients, both life expectancy and quality of life are superior to long-term dialysis. The two impediments to the extension of transplantation are the shortage of donor organs and the side-effects of the still crude immunosuppressive agents. Xenotransplantation may remove the first of these hurdles, but is likely to increase our dependence on potent immunosuppressive regimes. In humans, immunological tolerance to the graft with preservation of normal immunoresponsiveness to infections and tumours has not yet been achieved.

Supply, demand, and kidney donation

At the beginning of the year 2000 there were approximately 5000 patients waiting for renal transplantation in the United Kingdom. The annual rate of renal transplantation in the United Kingdom is approximately 1300 per year or 24 per million of the population per year and has not altered greatly over the last 5 years. The dialysis population, however, is increasing by 7 to 10 per cent per annum. In most countries the maximum achievable number of cadaver donors is about 35 per million per year against a need of 50 per million per year. The shortage of cadaver donors has been attributed to three main factors: a decline in deaths from road traffic accidents and cerebral haemorrhage, and inadequate numbers of intensive care unit beds. In the United Kingdom living donation represents only 10 to 15 per cent of all transplants (3 to 5 donors per million of the population), whereas in Scandinavia 10 live donors per million of the population has been achieved.

There are currently three possible sources of donor organs for transplantation: cadaveric, living related, and the living unrelated donor. In most countries the last group is restricted to donors that are closely 'emotionally related', for example spouses and partners. Two factors sustain our continued reliance on living donation as a source of kidneys: the first is the shortfall in available cadaver donors, and the second is the superior survival of a well-matched graft from a living related donor. The continuing need for an adequate supply of cadaver organs for transplantation requires an equally continuing education of both the medical and general population. Acceptance of the brainstem death criteria (see Chapter 16.6.3) in many countries has helped greatly in establishing a secure definition of death for both legal and religious purposes. However, given the continuing shortage of organs, many centres are re-exploring the possibility of the rapid procurement of organs for transplantation from non-heart-beating cadavers.

Living donors

Every care must be taken to protect the interests of the donor. Informed consent is crucial. Potential donors must be aware that giving a kidney carries risks, albeit that the mortality rate is only 0.01 to 0.03 per cent, with most deaths attributable to acute pulmonary embolus. The other risks that are involved in a general anaesthetic and an abdominal operation must also be fully explained. Needless to say, the donor should be in good general physical health and have normal kidney function and surgically acceptable renal anatomy. The assessments required are summarized in Table 1. Apart from exceptional circumstances, donors outside the age limits of 18 to 70 years are not considered. It is usual to wait for a young female potential donor to complete her family.

Most studies have shown an increase in life expectancy of donors when compared with age-matched controls. A small proportion of donors will develop hypertension, but at a risk that is similar to that of the general population. A small number develop proteinuria, but this is usually less than 0.5 g per 24 h and does not affect survival. Renal function usually returns to 75 to 80 per cent of the predonation level.

The superior outcome of living related donor kidney transplantation is partly due to better matching, with donor and recipient sharing one or two extended haplotypes in almost all cases. An additional benefit, shared also by kidneys from living unrelated donors, is the physiological state of the organ when recovered under ideal and planned conditions. Rejection is more likely and more severe with cadaver donors: this may be due to the 'cytokine storm' that accompanies the agonal phase of death and ischaemia reperfusion injury. This is thought to increase the expression of HLA antigens and adhesion molecules in the donor organ, making it more visible to the recipient's immune system.

Cadaver donors

In this situation the prime responsibility is to the potential recipient. The kidney should be in as good a physiological state as possible, and there should be no obvious risk of transfer of infection or malignancy by the donor organ. The major contraindications to organ procurement are listed in Table 2. Marginal donors are increasingly being considered, particularly for older recipients and for those with a limited life expectancy. In some situations it may be appropriate to consider organs from hepatitis C (**HCV**)-positive or hepatitis B (**HBV**)-positive donors for positive recipients. Experience in parts of the world where safe long-term dialysis is not available have shown that an acceptable quality of life can be sustained with substandard kidneys, and—given the shortage of organs—marginal donors should not be discarded out of hand without discussion with the local transplant unit.

Recipient assessment

Patients may be transplanted before the need for dialysis (pre-emptive transplantation) or from an established dialysis programme (haemodialysis or peritoneal dialysis). It is essential that all patients are fully assessed by both a transplant surgeon and transplant physician before being placed on the waiting list or offered a kidney, whether it be from a cadaver or a relative. Patients with chronic renal failure develop a multitude of complications that need assessment prior to surgery. Transplantation carries with it the risks of any major surgical procedure together with the added risks of prolonged immunosuppression. An additional consideration is that given the shortage of organs for transplantation, it is important that the best use is made of all organs. Whilst all would agree with this in principle, making decisions in individual cases can be difficult. In some situations the general health and life expectancy of the potential recipient argue strongly against transplantation. In patients with viral hepatitis or cirrhosis there is increasing evidence, particularly for HBV-related disease, that survival will be longer on dialysis. Patients with congenitally abnormal lower urinary tracts can be difficult to transplant and ideally should be managed in centres with urological transplant expertise, some needing complex bladder augmentation or drainage procedures prior to transplantation.

Allocation of kidneys

Fully matched kidneys (zero A, zero B, and zero DR mismatch—denoted 0–0–0 mismatch) and DR identical kidneys do better than less well-matched organs, hence most countries have local or national kidney sharing schemes so that more recipients can receive the benefits of a well-matched organ. Use is increasingly being made of point scoring systems to allocate kidneys fairly, patients accruing points based on the degree of match as well as the length of time they have been waiting for a transplant.

Surgical technique

The new kidney is placed in one or other iliac fossa, usually in an extraperitoneal position that allows ease of repeated biopsy to detect the cause of graft dysfunction. The renal artery is anastomosed end to side to the common iliac artery or end to end to the internal iliac artery. The renal vein is usually anastomosed to the common iliac vein. The transplant ureter only has a short distance to run before it can be implanted into the bladder, which is usually done through a submucosal tunnel to reduce the chances of reflux of urine from the bladder into the transplant. Some surgeons routinely place a vesicoureteric stent to reduce the risks of urine leakage and to promote healing. A drain is usually placed near the renal hilum. Lymphatics in the perihilar region are tied off. A urethral catheter and/or suprapubic bladder catheter is inserted and left *in situ* for about 5 days. The ureteric stent is removed at cystoscopy after a few weeks. Most units use prophylactic heparin routinely.

Note that in the standard renal transplant operation described above the native kidneys are left *in situ*. In some patients one or both may need to be removed (at a separate operation) before the patient can be listed for transplantation: mandatory indications for this include suspicion of renal tumour (usually in those with cystic disease), chronic renal infection, and massive organomegally, when there is literally no space in which to put a new kidney (in patients with adult polycystic kidney disease). Some would also advocate nephrectomy (Table 3) as a prelude to transplantation in those with gross ureteric reflux, renal stone disease, or analgesic nephropathy.

Retransplantation is increasingly being undertaken as the general medical care of patients with renal failure has improved. Second transplants are now not uncommon, and even third and fourth transplants may be occasionally undertaken. Third and fourth transplants are more surgically demanding as vessels available for anastomosis become limited. Aortic and venocaval anastomoses can be performed.

Ischaemia times

Warm ischaemia is defined as the time between circulatory arrest and renal artery cannulation for ice-cold perfusion, together with the time between the removal of the kidney from ice and release of the vascular clamps at implantation. With the beating heart donor, the first component is zero. The maximum permissible warm ischaemia time before irreversable damage occurs is 60 min.

Cold ischaemia time (preservation time) is defined as the time between ice-cold perfusion of the kidney and removal from the ice at the start of the implantation operation. Cold ischaemia times of up to 96 h have resulted in functioning grafts, but times in excess of 30 h are associated with a less favourable outcome. The permissible cold ischaemia time of 30 h allows for organ sharing and equitable operating times for the surgical team.

Postoperative management

Excepting for transplants between identical twins, immunosuppression is required to allow transplantation. The first dose of this is often given pre- or intraoperatively. Details are discussed below.

Following implantation, the function of the new kidney is assiduously monitored. Most units give low-dose dopamine, mannitol, or a loop diuretic, singly or in combination, to ensure good urine flow rate on return from theatre. Hourly urine volumes are closely monitored for the first few days. Fluid balance is usually maintained by a prescription that requires 100 per cent replacement of urine volumes and drain losses with crystalloid, and central venous pressure is monitored and maintained in the high normal range (+10 cmH_2O) with blood or colloid.

Serum creatinine is measured daily. A failure to fall rapidly, or a 15 per cent rise once it has fallen to a plateau, is evidence of graft dysfunction and requires prompt investigations. A kidney that fails to function initially, despite good perfusion on the table when the vascular clamps were removed, is usually suffering from acute tubular necrosis, which is expected to recover. A sudden cessation of urine flow usually means a surgical problem, for example clot obstruction, urinary leak, or vascular catastrophe. A slow tailing-off of the urinary volumes is more suggestive of rejection, hypovolaemia, or developing drug nephrotoxicity. Two of the major immuno-suppressive agents, cyclosporin A and tacrolimus, are nephrotoxic: doses have to be carefully adjusted to maintain the therapeutic range. Blood pressure should be returned to normal, obstruction excluded, and coagulation checked before a biopsy is undertaken. Close and careful monitoring needs to continue for the first 6 months following transplantation as the risk of rejection is at its greatest during this period.

One of the 'holy grails' of transplant medicine is a method of determining the immunological relationship between the recipient and their transplanted organ, since this would allow tailoring of immunosuppression to immunological need. However, immunological monitoring of transplant recipients is still in its infancy: lympocyte T- and B-cell subsets and activation markers can be of value, particularly when antilymphocyte preparations are being used; serial estimation of post-transplant anti-HLA antibodies can help predict patients at risk of chronic rejection. Much work continues to look for better ways of monitoring patients, such as testing for cytokine gene polymorphisms to predict those at highest risk of rejection, and examination of graft biopsies for expression of adhesion molecules, HLA, cytokines, and enzymes (e.g. granzyme, perforin) to characterize better the rejection process. Protocol biopsies may demonstrate subclinical rejection, and some argue that treatment of these may improve outcome, but most units are not convinced and do not perform 'routine' biopsies.

Complications of renal transplantation

Surgical

Table 4 summarizes the main surgical complications, which include those of any general anaesthetic and laparotomy. Extra risk is added because patients on dialysis are immunosuppressed by uraemia *per se* and transplant patients also require immunosuppressive drugs following surgery. Wound healing is significantly delayed in

the early post-transplant period, particularly by steroids. Some patients on dialysis will have a marked bleeding tendency related to defective platelet–endothelial cell interaction. The combination of uraemia, surgical stress, a bleeding tendency, and high-dose steroids produces an increased risk of bleeding peptic ulceration, which the routine use of H_2-blockers has virtually abolished. Many donor organs have small polar arteries that can be lost during or shortly after surgery, in which case the resulting segment of kidney will atrophy. Occasionally a polar infarct can lead to necrosis of a significant segment of renal cortex causing a calyceal fistula and urinary leak. An area of ischaemia around a polar infarct may drive post-transplant hypertension. Perirenal collections of fluid (whether from inadequately tied-off perihilar lymphatics, haemorrhage, or a urinary leak) can become infected: these are best demonstrated by ultrasound, which can guide aspiration for culture and drainage.

Rejection

Rejection can be classified into four main categories (Table 5). These are not mutually exclusive and there is overlap in the pathological processes.

Hyperacute rejection

In the presence of preformed cytotoxic antibodies the new graft infarcts within minutes of insertion. This can occur if transplantation is attempted across ABO incompatibilities. It is a rare event as the lymphocyte crossmatch usually identifies pre-existing anti-HLA antibodies. Transplantation is not undertaken in the presence of a positive lymphocyte crossmatch, but hyperacute rejection can rarely occur in the presence of non-HLA cytotoxic antibodies. There is no treatment save nephrectomy.

Accelerated rejection

A fierce, predominantly T-cell mediated rejection crisis may occur within the first few days of transplantation. This is thought to be due to sensitization of the recipient by a previous pregnancy, blood transfusion, or a failed transplant. Patients present clinically with fever, an acutely swollen tender graft, and a rapidly rising serum creatinine. Salvage usually requires the combination of high-dose intravenous pulse methylprednisolone (10 to 15 mg/kg per day infused over 30 min on 3 successive days) and an antilymphocyte antibody such as antithymocyte (ATG) or antilymphocyte globulin (ALG). The murine monoclonal antibody OKT3 may also be used. It is unusual to be able to reverse fully this type of severe rejection and long-term graft survival is compromised

Acute cellular rejection

In most centres about 25 per cent of patients will experience an acute cellular rejection, usually occurring between days 7 to 21 but up to 3 months following transplantation. Acute cellular rejection is often clinically silent as the inflammatory component of the rejection is masked by immunosuppression. Fluid retention, increasing hypertension, and a sharp rise in creatinine are typical. Assessment of renal perfusion (Doppler ultrasound or renography studies) may show a dramatic reduction in graft perfusion, but these tests are not sensitive or specific enough for a confident diagnosis of rejection. Most centres routinely take kidney biopsies for all episodes of graft dysfunction once infection, toxic levels of the calcineurin blocking drugs (cyclosporin and tacrolimus), and mechanical factors causing obstruction have been excluded. Obtaining a histological diagnosis is very important since several processes can mimic rejection, including drug nephrotoxicity, bacterial pyelonephritis, recurrence of original disease, and post-transplant lymphoproliferative disorder. The hallmark of acute cellular rejection is tubulitis in which the invading lymphocytes have penetrated the tubular epithelial cell basement membrane and directly engage tubule epithelial cells. Late acute rejection episodes usually imply inadequate immunosuppression, sometimes due to poor compliance. Treatment is very effective and usually involves a bolus of intravenous steroid therapy as described above. Long-term graft survival is severely jeopardized if the rejection episode is not completely reversed. With some of the newer, very potent induction regimes, the incidence of acute rejection episodes can be reduced to 10 per cent. Whether this will translate into a higher rate of infection and neoplasia awaits further follow-up.

Chronic rejection

Chronic rejection is a complex pathological process that is difficult to define. At its simplest, it represents the breakthrough of humoral immunity with an antibody-mediated attack on the graft endothelium. The result is an insidious and obliterative endovasculitis with progressive graft dysfunction from ischaemia. Clinically, the features of chronic rejection include difficult hypertension, proteinuria, and a slowly rising serum creatinine. (See the subsection on chronic allograft nephropathy for further discussion.) Chronic rejection is associated with the presence of anti-HLA antibodies in the serum and the deposition of C4d complement in the peritubular capillaries.

Immunosuppression

There is no clear consensus on the best immunosuppressive regime for renal transplantation, and for commercial reasons the large multicentre trials that the community of transplant physicians and surgeons would most like to see performed are unlikely ever to be funded. The choice of agents available is summarized in Table 6.

Most centres worldwide use what is now called standard triple therapy, comprising cyclosporin A, prednisolone, and azathioprine. However, in North America in particular, there is widespread use of additional serotherapy given as induction therapy for the first 10 days after grafting, and azathioprine has been replaced by mycophenolate mofetil in many centres, and cyclosporin by tacrolimus in some.

Agents of established efficacy for induction therapy include polyclonal antibodies such as antithymocyte globulin (ATG) or antilymphocyte globulin (ALG), and the murine monoclonal antibody, OKT3. Two anti-CD25 antibodies, daclizumab and basiliximab, which bind to the a-chain of the interleukin 2 (**IL-2**) receptor, have recently been introduced. Both are heavily engineered antibodies, comprising a murine antigen-binding site and human immunoglobulin. Both are very effective and have been shown to reduce acute rejection episodes by about 30 per cent.

Many centres are now exploring the possibility of tailoring immunosuppression to the needs of the individual recipient, but as described above we are not good at assessing immunological risk. In practice this involves giving immunosuppressants that are perceived to be more powerful—for example tacrolimus instead of cyclosporin, mycophenolate mofetil instead of azathioprine, or adding antibody treatments—to recipients who are thought to be at greatest immunological risk, such as those who are highly sensitized, patients who have rejected a previous transplant, or those who have suffered an acute rejection episode.

The best long-term therapy is equally in doubt. This is partly due to the nephrotoxicity of two of the main agents employed for the prevention of rejection: both cyclosporin A and tacrolimus can produce a nodular arterioleopathy resulting in ischaemic renal damage. It is also clear that the morbidity and mortality from long-term steroid therapy is significant, such that many centres are now attempting steroid-free immunosuppression, reducing and even withdrawing steroids at 3 to 6 months despite the associated risk of rejection, which has been reported to be as high as 30 per cent. Unfortunately, it is not possible to predict who is going to reject on withdrawal of steroids, hence one of the main aims of those developing new immunosuppressive drugs and regimens is to devise agents or protocols that allow less dependence on steroids without increased rates of rejection or other unacceptable toxicities.

The side-effects of immunosuppression

It is important to remember that all currently available immunosuppressive regimen are non-specific in the sense that they suppress not only the immune response to the allograft, but also the immune response to infections and tumours. All the agents used have significant side-effects and toxicities, and to a very large extent the long-term complications of renal transplantation are those of the immunosuppressive agents used. Some side-effects are more related to the total burden of immunosuppression rather than to any specific single agent, for example infections and cancer.

Specific side-effects of particular agents

Steroids

Steroids are responsible for many of the complications of transplantation (Table 7). In recent years the dose of steroids used has been safely reduced, thanks in part to the introduction of the calcineurin-blocking drugs, but attempts to produce totally steroid-free transplantation are only successful in about half of cases. One of the most significant side-effects of steroids is that they mask the inflammatory response so that symptoms develop late, which is particularly important in cases of

intra-abdominal catastrophy such as a perforated hollow viscus.

Calcineurin-blocking drugs

The major drawback of both cyclosporin A and tacrolimus is nephrotoxicity (Table 8), which adds another level of complexity to the differential diagnosis and management of both acute and chronic graft dysfunction. Most consider tacrolimus to be more potent than cyclosporin A, but perhaps more toxic (diabetes mellitus and neurotoxicity). It does, however, have real cosmetic advantages over cyclosporin A, perhaps mediated by lower levels of transforming growth factor-b.

Azathioprine and mycophenolate mofetil

Both agents block purine synthesis. The main side-effects of azathioprine are hepatotoxicity and marrow suppression (Table 9). Mycophenolate mofetil is more potent and more specific than azathioprine, blocking purine synthesis in lymphocytes with a degree of specifity. Its most troublesome side-effect is that of abdominal colic and diarrhoea: about 10 per cent of patients are so badly affected that they are unable to tolerate the drug. A higher incidence of invasive cytomegalovirus disease has been associated with mycophenolate.

Serotherapy

A range of antibodies to lymphocytes is available for clinical use (Table 6). Side-effects vary with the preparation used: it is important to remember that the consequences of augmenting immunosuppression with serological agents may last many months, even though administration is usually limited to 10 to 14 days.

Polyclonal antilymphocyte preparations can cause a marked first-dose effect in which lymphocytes are activated and secrete cytokines. High fever, rigors, and muscle and back pains are common, and hypotension may occur. With successive doses this reaction subsides. The murine anti-CD3 antibody (OKT3) is particularly prone to produce a first-dose effect that can lead to a widespread capillary leak syndrome with non-cardiogenic pulmonary oedema, hypotension, and shock. It should not be given to patients who are fluid overloaded. Aseptic meningitis and encephalitis are also seen occasionally. By contrast, the humanized and chimeric anti-IL2 receptor antibodies that have recently been introduced do not appear to have any short-term side-effects.

General side-effects of immunosuppression

Infectious complications

Introduction

The calcineurin-blocking drugs used for immunosuppression act to inhibit the T-helper cell (CD4) and prevent the elaboration of IL-2 and other cytokines. In some respects this is akin to the effects of HIV infection and it is therefore not surprising that the renal transplant recipient may develop the same range of opportunistic infections and tumours as is seen in patients with AIDS (see Chapter 7.10.21). Clinical features are often dramatic and rapidly evolving, hence prompt and precise microbiological diagnosis is essential. This requires early recourse to invasive techniques, for example biopsy, node aspiration, excision, bronchoalveolar lavage and even lung biopsy. Neurological symptoms and signs may herald central nervous system infection and require urgent CT scanning or MRI and the examination of cerebrospinal fluid whenever possible. Any pyrexial episode in a transplant recipient should prompt a search for infection. Blood and urine cultures should be undertaken routinely.

Figure 1 summarizes the timetable of infections. In the first month, before immunosuppression is fully established, renal transplant recipients may develop the same sort of infection seen after any general anaesthetic, abdominal operation, or urological procedure. From months 1 to 6, immunosuppression is maximal and the risk of opportunistic infections greatest. Thereafter, the risk of infection declines but remains greater than the general population.

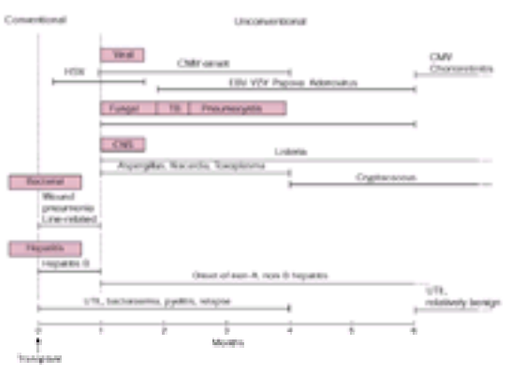

Fig. 1 Timetable of infections (reproduced by permission from Rubin R.H. and Young L.S. (eds) 1994, *Clinical approach to infection in the compromised host*, 3rd edn. Plenum Medical Book Co., New York.)

Viral infections

Not all virus infections prove dangerous to the immunosuppressed renal transplant recipient. Those with particularly important clinical sequelae are summarized in Table 10. The most important group are the DNA viruses of the herpes group: infection with these is immunomodulating in its own right and further immunosuppresses the patient, hence they are not infrequently associated with superinfections, for example *Pneumocystis carinii*, listeria, and bacterial sepsis. Several of the viruses have proven oncogenic potential and are considered later.

Cytomegalovirus

Cytomegalovirus (**CMV**) is the main infectious complication in solid organ transplantation (Table 11), with a primary infection more likely to produce serious disease than either reinfection or reactivation. Viral load, as indicated by quantitative polymerase chain reaction (**PCR**), and the total burden of immunosuppression, are the main determinants of disease. Use of potent serological agents, either for induction or rescue, is strongly associated with CMV disease. As would be expected, the total number of treated rejection episodes is an important risk factor. Diagnosis is usually via PCR for viral DNA or by an antigen assay (pp65) on peripheral blood leucocytes. Monitoring the serological response for diagnostic purposes is obsolete as it is far too insensitive and routine cultures are too slow. A range of effective prophylactic regimes is available: oral valaciclovir is effective and oral valganciclovir is awaited. Another equally valid approach is careful monitoring combined with pre-emptive treatment of infection, where therapy can be given before clinical disease in vulnerable individuals. Two or three weeks of intravenous ganciclovir is usually effective. Foscarnet is a more toxic alternative.

CMV may play a role in triggering or augmenting both acute and chronic rejection. If this is confirmed then more widespread prophylaxis may be indicated.

Epstein–Barr virus (EBV)

EBV-related syndromes (Table 11) are an important cause of morbidity and mortality in renal transplant recipients, the most serious problem being so-called post-transplant lymphoproliferative disorder, which is considered later.

Varicella zoster

Reactivation of latent varicella zoster (**VZV**) produces shingles, which is a common and unpleasant complication of transplantation. Immediate treatment with

intravenous aciclovir can limit spread and reduce post-herpetic pain. Much more dangerous is a primary VZV infection in an immunocompromised individual: this can cause a fulminating disease with hepatitis, pneumonitis, and disseminated intravascular coagulation occurring within a few days. Mortality is high. All patients who are to receive immunosuppression should have their VZV antibody status established. Those who are seronegative should be counselled about exposure to chicken pox and should report any contact immediately. Vaccination is available, but if exposed, susceptible individuals should be given zoster immune globulin (ZIG) and monitored closely. High-dose intravenous aciclovir should be given at the first suggestion of disease.

Herpes simplex

Although the classic herpetic cold sore is common after transplantation, herpes simplex virus (**HSV**) can produce a variety of serious clinical sequelae in the immunocompromised patient (Table 11). Use of prophylactic aciclovir or ganciclovir (primarily for CMV prophylaxis) dramatically reduces the risks of HSV infection. Treatment with aciclovir is very effective.

Human polyomavirus (BK and JC)

Most adult recipients are already seropositive for these viruses, indicating childhood infection that is usually asymptomatic. Primary infection can occur from the allograft. In most cases this is also asymptomatic, but rarely these viruses can cause an acute interstitial nephritis and graft dysfunction. The JC virus has been reported to cause a progressive multifocal leucoencephalopathy in renal transplant recipients, although this is very rare.

Papilloma viruses

Papilloma viruses cause an extensive range of viral warts in renal transplant recipients. Some types have been implicated in the pathogenisis of anogenital carcinomas and squamous cell carcinomas of the skin (see below). The management of viral warts in the immunocompromised patient is difficult when they are very extensive and consideration should be given to reducing immunosuppression. Localized lesions can be treated conventionally with topical agents such as glutaraldehyde or laser therapy, but widespread surgical excision is sometimes required. Local recurrence in scar tissue is common. A combination of oral etidronate (50 mg daily) and topical tretinoin cream (0.05 per cent) can control the lesions in severe cases.

*Human immunodeficiency virus (**HIV**)*

Infection with HIV is considered an absolute contraindication to transplantation. The time to AIDS and death is significantly shortened, particularly if HIV is acquired at or shortly after transplantation. However, the recently introduced intensive antiviral therapy for HIV infection may alter this approach. It is important to remember that HIV infection or behaviour considered to be at risk of contracting HIV or other viruses (lifestyle assessment) excludes such individuals from organ donation.

Bacterial infections

There are a limited number of bacterial infections that are significantly more common and more severe in the transplant population (Table 11). However, there is little doubt that bacteraemias are more common in transplant recipients, usually as a result of urinary tract infections. Metastatic abscesses in joints, skin, muscles, and the brain are also more frequent.

Mycobacterial infections

Reactivation of mycobacterial infection following transplantation is very common in the 'at risk' population, and most United Kingdom units recommend prophylaxis with isoniazid in these groups, although some debate the need for this. Experience in the Indian Subcontinent suggests that pretransplant BCG vaccination is not effective. Mycobacterial infections (both atypical and tuberculosis) can present in many different guises, for example pneumonia, lymphadenopathy, intracranial space-occupying lesions, discharging sinus, pyrexia of unknown origin, and skin ulcers. Tissue biopsy and cultures and smears employing special stains are essential. PCR, particularly of cerebrospinal fluid, is proving helpful. Gallium scanning may identify nodes that can be aspirated under CT guidance. Skin testing is unreliable in the immunocompromised patient.

Treatment is compromised by serious drug interactions between rifampicin and both the calcineurin-blocking drugs and prednisolone. Rifampicin is such a potent inducer of cytochrome P450 that sub-therapeutic levels of the calcineurin-blocking drugs and steroids can develop within weeks. Graft loss from rejection will occur unless doses are increased: that of prednisolone is usually doubled, and the calcineurin blockers may have to be increased still further and given three times daily. Monitoring of drug levels is essential.

In many units a four-drug antituberculous regimen is recommended, comprising rifampicin, ethambutol, isoniazid, and pyrazinamide. When sensitivities become available, this can be reduced. Treatment should be continued for at least a year, particularly in the case of atypical mycobacterial infections. Therapy may be further complicated by hepatotoxicity, for which the differential diagnosis is complex as many other factors can cause deranged liver function tests in renal transplant recipients (for example virus infections—HBV, HCV, CMV, and other drugs)

Nocardia

Nocardia typically produces either a pseudotuberculosis or a pseudostaphylococcal syndrome. Central nervous system infections can occur. Dissemination is common, occurring in 25 to 30 per cent. Diagnosis often requires a biopsy or aspiration, with cultures needing to be pro-longed for at least 3 weeks. Prolonged treatment (at least 6 months) with co-trimoxazole is usually effective, following which long-term co-trimoxazole should continue indefinitely.

Non-typhoid salmonella

Non-typhoid salmonella infections are noteworthy because of their tendency to produce metastatic abscesses following bacteraemia. With control of the acute illness, the continued excretion of the organism may occur in stool or urine. Relapse is common so treatment needs to be prolonged. Suitable antimicrobials include ciprofloxacin, co-trimoxazole, and ampicillin.

Listeria

Listeria has a tendency to localize in the central nervous system following a bacteraemic phase. Neurological syndromes vary from meningitis and meningoencephalitis to space-occupying lesions. It is the commonest cause of post-transplant meningitis. In the absence of evidence of raised intracanial pressure, all patients will require lumbar puncture and examination of cerebrospinal fluid. Delayed or inadequate treatment may result in permanent neurological deficit. Treatment usually includes high-dose ampicillin for at least 6 weeks, combined with gentamicin for the first week. The source of listeria is usually contaminated dairy products, chicken, or uncooked vegetables contaminated by manure.

Fungal infections

Oral candidiasis is a common post-transplant infection. Spread to the oropharynx and lungs may occur. All patients should receive prophylaxis (nystatin mouthwashes or amphotericin lozenges) for at least 6 weeks, but some practitioners would recommend longer courses, or even indefinite treatment in patients with diabetes. Intercurrent courses of antibiotics should be covered with oral prophylaxis against candida.

The spectrum of diseases produced by fungal infections is wide, ranging from mucocutaneous syndromes, severe pneumonias, central nervous system syndromes, to skin or muscle abscesses. This variation in clinical presentation again highlights the need for aggressive invasive investigation. Outbreaks of aspergillus are usually related to hospital building projects and should prompt a search for the source. Deep-seated fungal infections carry a very high mortality. Dissemination is common. Specialist microbiological advice is usually required, but if the fungus is sensitive then liposomal amphotericin is the drug of choice.

Parasitic infections

Some of the parasitic infections listed in Table 10 are geographically restricted and therefore will only be of specific relevance in those areas. Schistosomiasis, for example, can cause ureteric strictures and leaks following transplantation. *Strongyloides stercoralis* is usually found in patients from the West Indies or the Far East: in the immunocompromised it can reactivate, complete its lifecycle in the patient without need for an intermediate host, and produce a hyperinfestation syndrome. A pretransplant eosinophilia is sometimes present. Clinical presentation is with recurrent bouts of Gram-negative septicaemia as the worm penetrates the gut mucosa. Other clinical features include pruritus ani, haemorrhagic enteritis, lava currens, cough, wheeze, and a haemorrhaging bronchopneumonia. Meningitis may also occur. Diagnosis usually requires a duodenal aspirate. Treatment is with thiabendazole, which should be given pretransplant to susceptible patients. Several courses of treatment may be needed to eradicate the infestation.

Scabies may occur in transplant recipients and can produce so-called Norwegian scabies in which there may be many parasitic mites per burrow. In the immunocompromised patient, skin organisms are readily carried into the bloodstream, hence cellulitis and septicaemia are common.

The transplant organ, particularly the heart, can transmit toxoplasmosis. The organism becomes widely disseminated, including the central nervous system. Other clinical features may include low-grade fever, lymphadenopathy, pneumonia, myocarditis, retinopathy, and myositis. It can mimic cytomegalovirus. Treatment is with pyrimethamine and sulphadiazine for at least 4 weeks. Prophylaxis with co-trimoxazole has greatly reduced the incidence of toxoplasmosis following solid organ transplantation.

Pneumocystis carinii

Until the widespread introduction of prophylactic low-dose co-trimoxazole, *Pneumocystis carinii* pneumonia was a dreaded complication of solid organ transplantation. Oral co-trimoxazole or inhaled pentamidine (300 mg monthly) is effective prophylaxis. *Pneumocystis carinii* pneumonia is now most commonly seen in the setting of augmented immunosuppression (additional serotherapy) and in patients who already have developed CMV disease. Presentation is with fever, dry cough, and profound shortness of breath, occurring in the context of few added sounds in the chest and a remarkably clear chest radiograph. By the time the chest radiograph has altered, pulmonary fibrosis is occurring. Successful treatment demands an early diagnosis, such that the renal transplant recipient who complains of shortness of breath on exercise and who desaturates on exercise should be admitted and investigated as a medical emergency. Bronchoalveolar lavage is virtually mandatory under these circumstances. Overall immunsuppression should be reduced in patients with *Pneumocystis carinii* pneumonia, but steroids may need to be increased to cover a stress response (e.g. prednisolone at 20 to 25 mg daily). High-dose intravenous co-trimoxazole is given: 15 to 20 mg of trimethoprim and 75 to 100 mg of sulphamethoxazole per kilogram body weight per day, although these doses may need to be reduced in severe renal failure. Treatment should be continued for at least 2 weeks. It is essential to monitor respiratory effort carefully in the renal transplant recipient with an interstitial pneumonitis and intervene with continuous positive airways pressure or full ventilation if the patient tires or cannot protect his airways. Nutrition should be ensured, using total parenteral nutrition if necessary.

Specific infective problems

Pulmonary disease

Recurrent chest infections are common. Many are viral and will be self-limiting, even in the immunosuppressed transplant recipient. An abrupt clinical onset with fever and a lobar pattern of lung infiltrates is likely to be due to a bacterial infection. A more insidious onset with scattered or diffuse pulmonary infiltrates is more likely to be due to an opportunistic infection. Blood and sputum should be cultured urgently. Sputum samples need careful microscopy and cultures should be set up for mycobacteria, fungi, and legionella. PCR is available for *Mycobacterium tuberculosis*, *Pneumocystis carinii*, and CMV. Antibiotics may be started pending culture results. The regimen that will cover most of the common organisms is penicillin V, clarithromycin (NB drug interactions), and a third-generation cephalosporin.

Failure to respond promptly to therapy or a non-lobar pattern of infiltration is an indication for brochoscopy and bronchoalveolar lavage, the diagnostic accuracy of which is about 80 to 90 per cent. It is essential to examine the fluid thoroughly, which will involve viral and bacterial cultures, special stains, and PCR where available. In clinical practice it is often necessary to start therapy blindly in seriously ill patients. Sometimes this will involve the addition of high-dose co-trimoxazole and ganciclovir to conventional antibiotics. When the results of culture and sensitivity testing become available it may be possible to reduce the antimicrobial regime or change to specific antituberculous or antifungal therapy.

The greatest mimic of a chest infection is pulmonary oedema: measurement of an elevated pulmonary capillary wedge pressure is diagnostic, and a therapeutic test of a potent diuretic sometimes produces a dramatic clearing of the chest radiograph. Other non-infectious causes of acute pulmonary syndromes that may occur in the renal transplant recipient are shown in Table 12.

Urinary tract infection

One-third of renal transplant recipients will develop urinary tract infection. In most this is related to postoperative bladder catheterization and usually resolves with removal of the catheter and a short course of antibiotics. There is an exponential relationship between the incidence of urinary tract infections and the duration of bladder catheterization. Some patients develop numerous recurrent infections, particularly in the first couple of years following transplantation. In some this can be related to a focus of infection in the native kidneys, when bilateral native nephroureterectomy may be indicated if sepsis is severe. A few patients will develop encrustation or even a stone in the bladder as a result of the surgical implantation of the ureter: a plain abdominal radiograph may reveal such calculi, which should be removed cystoscopically.

More worrying is infection ascending into the transplant kidney itself during the intermediate period of post-transplantation immunosuppression when the patient is most immunocompromised. A severe bacterial pyelonephritis can develop in the transplant, presenting as an acute rejection episode with a swollen kidney, low-grade fever, and deteriorating graft function. Such upper tract infections are frequently complicated by septicaemia, and it is always worth remembering that urinary sepsis is the commonest cause of post-transplant bacteraemia. It is essential that episodes of graft dysfunction due to upper tract infection are clearly diagnosed and aggressively treated with appropriate high-dose parental antibiotics. Misdiagnosis resulting in treatment with high-dose intravenous steroids for a presumed rejection episode can be catastrophic.

The advent of technetium-99m labelled DMSA SPECT isotope scanning (single photon emission computed tomography using dimercaptosuccinic acid labelled with technetium-99m) has enabled three-dimensional reconstructions of the grafted kidney to be produced. Progressive scarring can develop in some patients with recurrent infections and reflux to the graft, hence transplant recipients with recurrent urinary tract infections need full investigation and aggressive treatment. Every effort should be made to establish and maintain sterile urine. Long-term prophylactic low-dose antibiotics may be indicated.

Neurological syndromes

The main concerns are those of post-transplant lymphoproliferative disorder or an opportunistic infection producing progressive neurological deterioration due to an increasing space-occupying lesion. Examples of neurological syndromes seen in the renal transplant recipient and their common causes are given in Table 13. The range of infectious micro-organisms that can cause central nervous system lesions is such that a diagnostic aspirate is usually essential. Tuberculosis is common in at-risk patients. Investigation should include a CT scan with contrast or an MRI so that abscesses are not missed. In the absence of evidence of a raised intracranial pressure, cerebrospinal fluid should be examined. As with the processing of bronchoalveolar lavage fluid, close co-operation between clinician and the cytological and microbiological laboratories is essential.

Fits may occur in the early post-transplant period, when the cause is usually multifactorial, including hyponatraemia, hypertension, hypomagnesaemia, hypocalcaemia, and the toxic effects of the calcineurin-blocking drugs. The rejection process itself can cause a rise in intracranial pressure, so-called rejection encephalopathy. Fits occurring after the first month should prompt a search for a serious intracranial space-occupying lesion.

Renal transplantation in the presence of liver dysfunction

The two main liver conditions encountered in patients on transplant waiting lists that give concern in the post-transplant period are hepatitis B (HBV) and hepatitis C (HCV).

Hepatitis B

HBV usually causes persistent infection in patients with chronic renal failure. In many this may be subclinical, but in others a chronic hepatitis and cirrhosis can develop. Serology is of little help in assessing suitability for transplantation, which is contraindicated in the presence of cirrhosis and biopsy evidence of active hepatic inflammation since immunosuppression causes rapid viral replication and progressive liver disease. Death within 5 years of transplantation may occur in up to 50 per cent of patients if wrongly transplanted, usually from extrahepatic sepsis. Long-term therapy with the newer antiviral agents may improve the outlook and allow access to transplantation to those at present denied this.

Hepatitis C

Although HCV is now the most common cause of both pre- and post-transplant liver disease, the effects of immunosuppression on HCV seem much less dramatic than the effects on HBV. Occasional patients do develop a fulminating hepatitis post-transplantation, but overall HCV does not appear to have a major impact on the short- to medium-term outcome after renal transplantation.

Post-transplant liver dysfunction

Abnormal liver function tests following transplantation are common: both drugs and infectious agents may be responsible. Full investigation is required, including imaging of the liver, bile ducts, and gallbladder as well as a liver biopsy. In some instances transient elevation of liver transaminases may herald CMV disease. In other situations raised liver enzymes can represent progressive HCV- or HBV- induced liver disease. It is important to remember that the donor organ can transmit most of the hepatotropic viruses. Treatment clearly depends on the cause. Where possible the offending drug (for instance azathioprine) should be withdrawn and antiviral therapy may be appropriate in the case of HBV and HCV. Interferon therapy is contraindicated as it induces expression of HLA antigens and may provoke acute rejection. In the case of HBV it is often possible to reduce the dosage of immunosuppressive agents significantly without precipitating a rejection episode.

Neoplasia

Post-transplant neoplasia is an important cause of morbidity and mortality. There is some debate as to whether some of the conditions often regarded as neoplastic can truly be classed as cancers, since several are clearly viral related and will regress with reduction of immunosuppression. Table 14 summarizes the tumours seen with increased frequency after transplantation. There is a marked geographical variation: for instance in Japan, renal, thyroid, and uterine cancers as well as lymphoma are common; in Saudi Arabia, Kaposi's sarcoma is the most common; in Australia squamous cell carcinoma of the skin is almost ubiquitous 20 years after transplantation (75 per cent). It is also important to remember than the donor organ can transmit cancer.

Post-transplant lymphoproliferative disorder

Post-transplant lymphoproliferative disorder is driven by Epstein–Barr virus (EBV) present in a latent form (episomal or circular DNA) in B lymphocytes. In non-immunosuppressed individuals a normal T-cytotoxic lymphocyte response terminates infected proliferating B cells. In the presence of effective immunosuppression this does not happen and an unrestricted, increasingly monoclonal B-cell proliferation develops. The more potent the immunosuppressive regime, the earlier post-transplant lymphoproliferative disorder occurs. In most centres, the incidence of this disorder is about 2 per cent. The clinical features are summarized in Table 14. In common with many other infections following transplantation, a primary infection (i.e. the recipient is naive or seronegative for EBV antibodies, while the donor is seropositive) leads more frequently to disease.

Early diagnosis is important as the stepwise reduction of immunsuppression with careful monitoring of graft function can lead to regression of the tumour without graft rejection. Stimulating the patient's immune system with interferon-a or IL-2 may be tried. Conventional cytotoxic therapy should be introduced if post-transplant lymphoproliferative disorder progresses despite the withdrawal of immunosuppression, but this further suppresses the patient's immune system and death from overwhelming infection is all too common. Anti-B-cell antibodies (e.g. rituximab) and infusions of EBV-specific cytotoxic T lymphocytes are promising new avenues of therapy.

It remains to be proved, but seems very likely, that the carefully monitored stepwise reduction of immunosuppression may also be appropriate for other virally induced neoplasms in renal transplant recipients, for example Kaposi's sarcoma and squamous cell carcinoma. A few patients who have lost their grafts in the context of post-transplant lymphoproliferative disorder have been successfully retransplanted.

Kaposi's sarcoma

Kaposi's sarcoma is a vascular tumour composed of proliferating spindle cells (latently infected lymphatic endothelial cells) and thin-walled neovascular formations that is driven by the Kaposi's sarcoma virus (KSV), recently designated HHV8. Aetiological factors are very similar to those of post-transplant lymphoproliferative disorder. Lesions may develop at almost any site on the skin and visceral involvement is common. Prompt diagnosis and early reduction of immunosuppression may result in regression. As with post-transplant lymphoproliferative disorder, the use of cytotoxic agents is associated with a greatly increased risk of death from sepsis. Attempts at retransplantation after regression are associated almost universally with recurrence.

Human papilloma virus related carcinoma

Human papilloma virus (HPV) is responsible for skin, vulval, and anogenital warts, and some types are now clearly associated with carcinoma. Renal transplant recipients should therefore receive full dermatological and gynaecological examinations at regular intervals.

Aetiological factors for skin cancer include exposure to ultraviolet light (which may act by depletion of cutaneous Langerhan's cells as well as direct DNA damage), duration and intensity of immunosuppression, and the HPV virus itself. The prevalence increases progressively with time, such that after 20 years most renal transplant recipients will have cutaneous squamous cell carcinoma. Management should involve cautious dose-reduction of immunosuppressive agents. There is great interest in the role of combined oral and topical retinoids, which are associated with repopulation of the skin with Langerhan's cells and augmentation of natural killer cell activity, and may also act by blocking IL-6 pathways. Some cases of squamous cell carcinoma can metastasize, when reduction of immunosuppression dose, interferon-a therapy, and a willingness to abandon the graft should be considered before recourse to systemic cytotoxic chemotherapy.

Hypertension

The aetiology of post-transplant hypertension is complex (Table 15). Over 75 per cent of renal transplant recipients will need drug therapy for hypertension in addition to lifestyle modification. Most units aim for a systolic blood pressure of less than 145 mmHg and a diastolic pressure of less than 85 mmHg, but there is a lack of clear data regarding the ideal blood pressure. Hypertension plays a crucial role in accelerating vascular disease and chronic allograft nephropathy. Care should be taken with the choice of agents. On theoretical grounds, angiotensin-converting enzyme (**ACE**) inhibitors or angiotensin I (**AT-I**) receptor antagonists appear a rational first choice in view of the multiple proinflammatory, profibrotic, and proproliferative actions of angiotensin II. There are overwhelming data showing reduction of proteinuria and a slowing of progression of renal disease in native kidneys treated with ACE inhibitors, and by implication also by AT-I receptor-blocking drugs. However, the risks of using ACE inhibitors in patients with renal artery stenosis should not be forgotten, transplant renal artery stenosis being most likely to develop between 3 and 12 months after transplantation, sometimes (but not always) associated with a bruit over the kidney. Serum creatinine must be carefully monitored, any substantial rise after ACE inhibition leading to immediate cessation of the drug and consideration of angiography of the transplant renal artery. It is preferable to avoid drugs that exacerbate dyslipidaemia, such as b-blockers and thaizides. Poor blood pressure control with short-acting dihydropyridine calcium-channel blockers may increase proteinuria and cardiovascular mortality. In refractory cases, or those where treatment is problematic, estimation of renin levels in the veins draining the native and transplant kidneys may help in the decision to proceed to a native kidney nephrectomy.

Accelerated atherosclerosis

In common with patients on dialysis, one of the major causes of death following renal transplantation is cardiovascular disease. Indeed, death with a functioning graft is now the major cause of late graft failure. Much of the cardiovascular disease that shortens life expectancy in renal transplant recipients will have developed and be

established pretransplantation. Table 15 summarizes the pre- and post- transplant aetiological factors. Prevention and treatment of established vascular disease is essential. About a third of renal transplant recipients will have hypercholesterolaemia and many will also be hypertensive. Lifestyle modification is important. All renal transplant recipients should be strongly advised not to smoke. Following transplantation, some 10 per cent of transplant recipients become quite grossly obese. It is important to remember that the cardiovascular risk factors multiply rather than summate, hence the long-term management of renal transplant recipients has to address all cardiovascular risk factors.

Electrolyte disorders

Hypophosphataemia

In the presence of inadequately controlled secondary hyperparathyroidism a well functioning transplant will waste phosphate, and in a few cases phosphaturia persists despite resolution of the secondary hyperparathyroidism. In some patients there is steroid-related malabsorption of phosphate. In the first few months following renal transplantation, phosphate wasting can be severe and oral supplements will be required. Untreated chronic hypophosphataemia can lead to bone fractures (hypophosphataemic rickets).

Hyperkalaemia

The calcineurin-blocking drugs cause hyperkalaemia, particularly when levels are toxic. This is thought to be due to type IV renal tubular acidosis in which distal tubular potassium secretion is reduced in response to a fall in renin secretion due to reduced renal prostaglandins. The addition of ACE inhibitors, AT-I receptor-blocking drugs, non-steroidal anti-inflammatory drugs, or potassium-conserving diuretics can produce a brisk rise in serum potassium in renal transplant recipients. Dietary advice and loop diuretics are usually sufficient but a small number of patients may require fludrocortisone (100 to 200 µg daily.)

Hypomagnesaemia

Renal tubular magnesium wasting is a component of the nephrotoxicity of the calcineurin-blocking drugs and can be exacerbated by diuretics and diarrhoea. Hypomagnesaemia may predispose to fits and cardiac arrhythmias in susceptible individuals. Levels should be monitored and oral supplements of magnesium glycerophosphate given if required.

Hypercalcaemia

Hypercalcaemia can develop after grafting if renal osteodystrophy has been poorly controlled and severe secondary hyperparathyroidism is present at the time of transplantation. The transplant kidney produces adequate amounts of 1,25-dihydroxycholecalciferol, which in the presence of high levels of parathyroid hormone will result in hypercalcaemia. Widespread metastatic deposition of calcium can occur if hypercalcaemia is severe. Simple controlling measures include adequate fluids and the use of loop diuretics rather than thiazides. Occasional patients will require regular infusions of pamidronate (15 to 30 mg), which can be combined with intermittent doses of oral a-calcidol to suppress parathyroid hyperplasia. Parathyroidectomy is occasionally required. It may take 12 to 24 months for secondary hyperparathyroidism to resolve following renal transplantation.

Bicarbonate wasting

The transplant kidney may waste bicarbonate as well as phosphate. This may be due to persistent hyperparathyroidism, but can also reflect acute tubule damage from rejection. A chronic metabolic acidosis will contribute to post-transplant osteoporosis and should be treated with bicarbonate supplements.

Hyponatraemia

Hyponatraemia may develop, particularly in the early postoperative period. It is usually due to inappropriate intravenous fluids (excess 5 per cent dextrose or dextrose saline) in the context of deteriorating graft function, and may be an important contributing factor to fits after transplantation.

Musculoskeletal complications

Tendon rupture

Steroids impair collagen synthesis. Tendons and tendon insertions are weakened and avulsions may occur, most commonly in the fingers or Achilles' tendon.

Myopathy

An important complication of steroid therapy is proximal myopathy, which can be incapacitating in some patients. Physiotherapy plus vitamin D supplements and a rapid reduction of steroids (alternate-day prescription or even cessation) can produce improvement. Hypophosphataemia should be corrected. Acute rhabdomyolysis may develop if fibrates or statins are used with the calcineurin-blocking drugs.

Avascular necrosis of bone

Avascular necrosis of bone, particularly of the weight-bearing ends of the long bones, causes an extremely painful joint. When the hips are involved, walking can become impossible and total hip replacement is the only treatment. Prevention may be possible by the careful control of secondary hyperparathyroidism prior to transplantation and the early use of bisphosphonates to minimize post-transplant osteoporosis may be beneficial.

Osteoporosis

Osteoporosis is a common and progressive complication of long-term steroid therapy such that regular bone-density assessment should be part of long-term renal transplant follow-up. The problem is particularly severe in postmenopausal women, in whom hormone replacement therapy is of benefit. Prophylaxis with intravenous pamidronate has been advocated in the first few months after transplantation, but this is not standard practice in most units.

Renal osteodystrophy

In the presence of a poorly functioning graft, control of parathyroid hormone (**PTH**) and the calcium-phosphate product (ideally to be kept at less than 5) is as important as it is in the pretransplant patient with chronic renal failure (see Chapter 20.5.1). PTH levels should be kept at one to two times the upper limit of normal by the careful use of a-calcidol and calcium supplements. Serum phosphate should be kept at 1 to 1.5 mmol/l using calcium carbonate as an oral phosphate binder.

Gout

The calcineurin-blocking drugs impair urate secretion in the proximal tubule, and urate retention is exacerbated by the concomitant use of diuretics, particularly in patients with poorly functioning grafts. Uric acid levels may rise dramatically and be associated with attacks of clinical gout as well as tophi. Management is complicated, both for acute attacks and for prophylaxis. For acute episodes the physician must choose between three treatments, all of which are problematic. Non-steroidal anti-inflammatory drugs are the usual first-line treatment for acute episodes in general medical practice, but in those with renal impairment—including many transplant recipients—they are best avoided, although the recently introduced Cox 2 inhibitors may prove safer. Colchicine can be used for acute attacks, but the transplant recipient tolerates diarrhoea and the attendant hypovolaemia poorly. Oral prednisolone (20 mg daily) can be effective, but will clearly exacerbate steroid side-effects, which are already a problem in many patients. As regards prophylaxis, allopurinol is (relatively) contraindicated if the patient is receiving azathioprine (the dose of azathioprine should be reduced to about 25 per cent of normal) with very careful monitoring for leucopenia. Most uricosuric agents work poorly in the presence of renal impairment, but an important exception to this is benzbromarone (100 to 200 mg daily), which can be safely administered to patients on

azathioprine. In some patients it may be helpful to stop azathioprine altogether and use mycophenolate mofetil in its place so that allopurinol can be used safely.

Haematological complications of renal transplantation

Venous thromboembolism is not uncommon following renal transplantation. The local effects of surgery on the pelvic veins together with immediate postoperative bed rest contribute to the risk. The calcineurin-blocking drugs have an activating and procoagulant effect on endothelial cells and platelets. Nephrotic patients have a profound disturbance of many coagulation factors and represent an extremely high-risk group for perioperative venous thromboembolism. Prophylactic subcutaneous low-molecular-weight heparin (e.g. enoxaparin at 20 mg daily) is standard practice, with higher doses (e.g. enoxaparin at 40 mg daily) in those at highest risk.

A direct endothelial effect of the calcineurin-blocking drugs can result in a *de novo* post-transplant haemolytic uraemic syndrome, and it seems likely that both cyclosporin and tacrolimus may increase the risk of recurrent bouts in patients with this as their primary disease.

Bone marrow suppression may occur as a result of intercurrent viral infection and a variety of drugs. In the context of severe CMV infection it is safe to continue with ganciclovir or aciclovir, but the bone marrow should be stimulated with granulocyte colony-stimulating factor. Profound bone marrow suppression may occur when allopurinol is used with azathioprine if the dose of the latter is not appropriately reduced (see above).

An acute haemolytic anaemia can develop at any time following transplantation. Sometimes this is triggered by an intercurrent infection, but in many cases may be due to antibodies to minor blood antigens. Treatment consists of intravenous immunoglobulin and an increase in steroids.

Patients with a poorly functioning graft will become anaemic just as their predialysis counterparts. Haematinics should be prescribed if patients are deficient, using intravenous iron if iron stores cannot be restored by oral supplements. Erythropoietin may be required.

About 20 per cent of renal transplant recipients develop erythrocytosis. The mechanism is probably multifactorial. The transplant kidney may produce an excess of erythropoietin, occasionally stimulated by renal artery stenosis in the donor kidney. The use of diuretics may produce blood volume contraction. Reduction in erythrocytosis following administration of ACE inhibitors or AT-I receptor-blocking drugs suggests that angiotensin II may contribute to pathogenesis. In most cases the condition is self-limiting, but there is a risk of cerebrovascular occlusion if the haematocrit is grossly elevated and most practitioners recommend regular venesection (or ACE inhibitors) to keep the haemoglobin below 15 or 16 g/dl. A high haematocrit will also exacerbate hypertension.

Cosmetic complications

It is important not to underestimate the psychological importance of the cosmetic disfigurement that can be produced by some of the treatment regimes used in transplantation. They are an important contributing factor to non-compliance, particularly in adolescents, and can even lead to agoraphobia and suicide. With the currently available choices of immunosuppressive agents it should be possible to minimize cosmetic complications when these cause great distress, for example steroid withdrawal, substitution of tacrolimus for cyclosporin A, and use of mycophenolate mofetil to reduce reliance on steroids and calcineurin-blocking drugs. The better cosmetic profile of tacrolimus is thought to be due to lower transforming growth factor-b production, but expense may be a limiting factor in some health-care systems.

Outcome

Graft and patient survival

Figure 2 and Figure 3 summarize the graft survival rates for first cadaver and living related transplants. The highest rate of graft loss is within the first few months. Graft losses due to technical factors should be less than 5 per cent. The major cause of early graft loss continues to be acute rejection. However, it is a matter of concern that the attrition of grafts following the first year has not altered, even with the introduction of newer, more potent immunosuppressive agents. Currently some 4 per cent of grafts fail annually for a variety of causes loosely grouped together as chronic allograft nephropathy (see discussion below). Death with a functioning graft is now the most common cause of late graft failure.

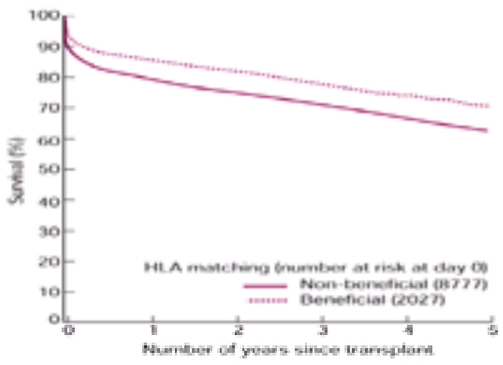

Fig. 2 Graft survival: first cadaver graft (by courtesy of UK Transplant).

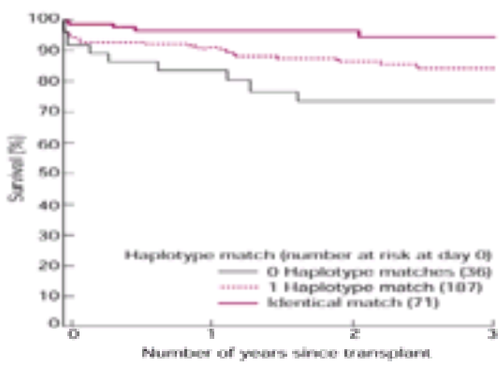

Fig. 3 Graft survival: living related grafts (by courtesy of UK Transplant).

Factors affecting graft survival are summarized in Table 16. Even with potent immunosuppressive regimes, HLA matching remains extremely important, forming the rationale for local and national organ sharing schemes to ensure that the best possible matches can be obtained. Figure 2 indicates that beneficially matched cadaver kidneys (1–0–0 or 0–1–0 mismatch) fare significantly better than non-beneficially matched, and well-matched living related transplants do best of all (Fig. 3).

Early studies indicated that an acute rejection episode had a major impact on long-term graft survival, reducing it by almost a half. If an acute rejection episode is completely reversed the effect on long-term graft survival is markedly reduced. Long-term graft survival can be clearly related to the creatinine level at 1 year. This has led to great emphasis on efforts to reduce the rate of early acute rejection episodes. One crucial observation that predicts those at increased risk is the presence of widely reactive anti-HLA antibodies to potential donors. Those patients who have already rejected a kidney within 6 months of transplantation also do poorly on subsequent transplantation unless immunosuppression is augmented. Increasingly potent induction regimens and combinations of drugs have been introduced, but in the absence of accurate predictors of the risk of rejection this has the effect that a significant number of patients will be grossly over-immunosuppressed, whilst others

remain under-immunosuppressed. Poor long-term graft survival is also related to hypertension, proteinuria, hyperlipidaemia, and a high body-mass index (Table 16).

In the early post-transplant period the major causes of death are related to cardiovascular complications of surgery. However, with good patient selection, mortality in the first year should be very low (less than 1 to 2 per cent), despite the fact that increasingly older patients are being offered transplantation, as are those with significant comorbidities, diabetes mellitus in particular. In the first year the major cause of death is infection (Fig. 4). Later on, death from neoplasia and accelerated vascular disease are more common.

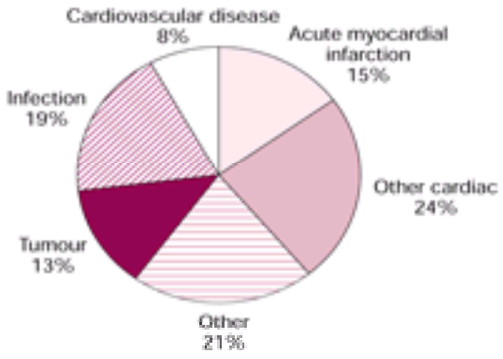

Fig. 4 Causes of death in renal transplant recipients.

Chronic allograft nephropathy

First-year transplant losses from rejection have been dramatically reduced from about 40 per cent in the 1970s to 5 to 10 per cent. Similarly, early rejection rates can be reduced from around 50 per cent to less than 20 per cent. However, as stated above, the rate of chronic graft loss remains at about 4 per cent per year. Some of these graft losses will be due to death with a functioning graft, but many patients present with an insidiously rising creatinine, increasing proteinuria, and worsening hypertension. The causes of this late graft dysfunction are multifactorial and involve many different conditions and several overlapping pathogenic pathways. Histologically, many of these kidneys will show tubular atrophy, interstitial scarring, and an obliterative vasculopathy of the intrarenal arteries and arterioles, but even with a transplant biopsy it may not be possible to define the pathololgical process. Nevertheless, every effort should be made to produce an accurate diagnosis. Obstruction due to ureteric ischaemia must be excluded by renography or ultrasound. A review of past urine cultures and a technetium-99m labelled DMSA SPECT scan may reveal pyelonephritic scarring. Late graft dysfunction due to renal artery stenosis can be demonstrated by angiography. It is possible that the single transplant kidney may have insufficient numbers of nephrons to cope with the work demanded of it, particularly if many nephrons have been lost early from acute rejection, manifest as a raised serum creatinine at 1 year. This situation may also arise if there is a very large size mismatch, with a very large recipient being given a kidney from a smaller donor, leading to the self-perpetuating inherently progressive cycle of hyperperfusion and hyperfiltration of the surviving nephrons (Brenner's hypothesis). Treatment is empirical but must involve control of blood pressure, preferably using an ACE inhibitor or an AT-I receptor-blocking drug to reduce intraglomerular blood pressure.

Some pathogenic aspects of insidious graft loss are akin to atherosclerotic changes, and interventions designed to limit vascular damage may prolong the life of poorly functioning kidneys. The early use of ACE inhibitors (especially if there is proteinuria), aspirin, fish oils, and control of cholesterol may help. Chronic allograft nephropathy may also be due in part to the nephrotoxic effects of the calcineurin-blocking drugs. The newer agent, mycophenolate mofetil, may have a particular role to play in that it has been shown to reduce proliferation of smooth muscle cells, which may ameliorate the obliterative vasculopathy typical of the condition, and permit reduced reliance on long-term calcineurin-blocking agents, but this is yet to be proved.

Recurrence of original disease and *de novo* glomerulonephritis

Most of the primary glomerular diseases can recur in the transplant, but few are associated with graft loss. Overall, histologically demonstrable recurrence occurs in about 60 per cent of patients, but less than 10 per cent will lose their graft as a result. Accurate recurrence rates are difficult to determine, being roughly 2.5 per cent at 2 years, 10 per cent at 5 years, and perhaps as high as 20 per cent by 8 years. Treatment of recurrent glomerulonephritis is not particularly effective but has usually involved intensive plasma exchange or an increase in immunosuppression.

Oxalosis will recur rapidly in the kidney unless a liver transplant is also done to correct the underlying enzyme defect. Transplantation in the presence of circulating antibody to glomerular basement membrane will result in the immediate recurrence of Goodpasture's disease. The dense deposit variety of mesangiocapillary glomerulonephritis, which is usually associated with hypocomplementaemia, predictably recurs and may destroy the graft. In paediatric practice, the nephrotic syndrome associated with focal segmental glomerulosclerosis may recur in the immediate post-transplant period and can be associated with massive proteinuria, hypovolaemia, and thromboembolism. The risk of recurrence should be taken into account when assessing patients for the possibility of a living related transplant.

Membranous glomerulonephritis can develop *de novo* in patients in whom the original disease was demonstrably different (between 2 and 10 per cent).

Other aspects of medical management of transplant recipients

Drug interactions

Care has to be taken when prescribing drugs for renal transplant recipients. Renal function must be considered, as well as the potential for drug interactions between immunosuppressive agents and other pharmaceuticals. Table 17 summarizes the more common interactions. ACE inhibitors, AT-I receptor antagonists, and non-steroidal anti-inflammatory drugs can compromise the perfusion of a single transplanted kidney, particularly if there is a degree of renal artery stenosis. Great care must be taken with potent enzyme inducers such as rifampicin as subtherapeutic levels of steroids and the calcineurin-blocking drugs can occur.

Diet

The help of a renal-trained dietician is essential. Patients may eat voraciously after release from the restrictions of dialysis and, with the euphoric effects of steroids, some gain in excess of 20 kg in the first year. About 5 per cent become grossly obese, which is associated insulin resistance, hyperlipidaemia, sympathetic overactivity, and hypertension. Hypercholesterolaemia is present post-transplant in about 30 per cent of patients and is related to drugs, proteinuria, and diet. Patients should avoid a high intake of saturated fats. Sodium intake—easily gauged from monitoring the 24-h urinary sodium excretion—is often excessive, a desirable intake being less than 100 mmol per day. A high urinary sodium causes urinary calcium wasting and may contribute to post-transplant osteoporosis as well as making hypertension more difficult to control. All patients should spend time with the dietician and be encouraged to adopt healthy eating guidelines. In addition to dietary advice, transplant recipients need education about the risks of contaminated food, for example with listeria, campylobacter, and cryptosporidium.

Additional therapy

The medical complications of renal transplantation are so numerous that many recipients will require many different drugs. The regimen often become intolerable and non-compliance can be a major problem. In the early post-transplant period it is necessary to give prophylaxis with co-trimoxazole, an H_2 antagonist, and possibly a bisphosphonate, as well as an appropriate anti-CMV regime. Patients who are at risk of tuberculosis require isoniazid for the duration of immunosuppression. Hypertension needs aggressive control. In the early post-transplant period there is also the possibility of wasting of magnesium, phosphate, and bicarbonate, each requiring supplements. Uric acid levels may be high and clinical gout may develop requiring either allopurinol or benzbromarone. Long-term management needs to include regular vaccinations (influenza and pneumococcus). To reduce the risks of accelerated vascular disease, aspirins and statins may also be indicated.

In patients with poorly functioning transplants, medical management must include the same measures as would be undertaken in a low clearance clinic for patients

expected to start dialysis. Under these circumstances, treatment may include erythropoietin therapy, iron and vitamin supplementation, a-calcidol, and oral phosphate-binders.

Follow-up

With an uncomplicated transplant operation, patients may only be in hospital for about 7 days. Following discharge, patients will need to be seen two or three times a week for the first month, once or twice a week for the second month, and then weekly for the third month. At each visit blood pressure and graft function is checked. Many units undertake weekly CMV surveillance for at least the first 3 months following transplantation. After 3 months, outpatient visits are gradually reduced with patients eventually being reviewed only every 3 to 4 months. Particular attention has to be paid to cardiovascular risk factors, infections, and neoplasia. Ideally, all patients should have an annual dermatological examination and women should have an annual cervical smear and colposcopy if indicated. Bone density should be monitored regularly. Even in an apparently stable transplant, some units perform a renogram every 1 to 2 years to detect deteriorating renal perfusion or an obstruction from an ischaemic ureteric stenosis. Patients at risk of tuberculosis will require a regular chest radiograph. Vaccinations should be kept up to date. Many centres offer an anniversary clinic when these medical complications can be more fully assessed. Accelerated atherosclerosis will lead to early coronary and peripheral vascular disease. Increasingly, renal transplant recipients are being put forward for coronary revascularization procedures, when it is interesting to note that angioplasty alone is less successful in renal patients and should be combined with stenting.

Pregnancy

A successful renal transplant restores fertility and pregnancy with normal vaginal delivery (unless there are obstetric indications for caesarean section) is possible. Most recommend that pregnancy is not embarked upon in the first year or if the serum creatinine is above 150 µmol/l or proteinuria greater than 2 g/day. Many successful pregnancies have been undertaken with renal function worse than this, but the risks are greater.

There is little evidence that immunosuppression with prednisolone, azathioprine, and cyclosporin A has a significant adverse effect on the fetus. Cyclosporin A may be associated with intrauterine growth retardation and prednisolone may produce neonatal adrenal suppression. Experience with tacrolimus is limited but is probably broadly similar to cyclosporin A. Pregnancy with mycophenolate mofetil is contraindicated. There is an increased risk of hypertension and pre-eclampsia in renal transplant recipients. Care has to be taken with the choice of antihypertensive agent. During delivery, intravenous fluid should be given and great care taken to avoid episodes of hypovolaemia and hypotension. It is usual to give an extra dose of steroid during the delivery, for instance 100 mg of intravenous hydrocortisone, and to increase oral prednisolone for a few days afterwards.

Further reading

General

Morris PJ (1994). *Kidney transplantation: principles and practice*, 4th edn. WB Saunders Co., Philadelphia.

Rubin RH, Young LS, eds (1994). *Clinical approach to infection in the compromised host*, 3rd edn. Plenum Medical Book Co., New York.

Suthanthiran M, Strom TB (1994). Renal transplantation. *New England Journal of Medicine* **331**, 365–76.

Infections

Brennan DC, Garlock KA, Lippmann BA (1997). Control of cytomegalovirus-associated morbidity in renal transplant patients using intensive monitoring and either pre-emptive or deferred therapy. *Journal of the American Society of Nephrology* **8**, 118–25.

Jassal SV, Roscoe JM, Zaltzman JS (1998). Clinical practice guidelines: prevention of cytomegalovirus disease after transplantation. *Journal of the American Society of Nephrology* **9**, 1697–708.

Lowance D *et al*. (1999). Valacyclovir for the prevention of cytomegalovirus disease after renal transplantation. *New England Journal of Medicine* **340**, 1462–70.

Lufft V *et al*. (1996). Incidence of *Pneumocystis carinii* pneumonia after renal transplantation: impact of immunosuppression. *Transplantation* **62**, 421–3.

Paya CV (1993). Fungal infections in solid organ transplantation. *Clinical Infectious Diseases* **16**, 677–88.

Sternberg RI *et al*. (1993). Utility of bronchoalveolar lavage in assessing pneumonia in immunosuppressed renal transplant patients. *American Journal of Medicine* **95**, 358–64.

Tumours

Alloub MI *et al*. (1989). Human papillomavirus infection and cervical intraepithelial neoplasia in women with renal allografts. *British Medical Journal* **298**, 153–6.

Barr BBB *et al*. (1989). Human papilloma virus infection and skin cancer in renal allograft recipients. *Lancet* **i**, 124–8.

Bouwes-Bavinck JN *et al*. (1995). Prevention of skin cancer and reduction of keratotic skin lesions during acitretin therapy in renal transplant recipients: a double-blind placebo-controlled study. *Journal of Clinical Oncology* **13**, 1933–8.

Bouwes-Bavinck JN *et al*. (1996). The risk of skin cancer in renal transplant recipients in Queensland, Australia. *Transplantation* **61**, 715–21.

Gotti E, Remuzzi G (1997). Post-transplant Kaposi's sarcoma. *Journal of the American Society of Nephrology* **8**, 130–7.

Opelz G *et al*. (1995). Analysis of non-Hodgkin's lymphomas in organ transplant recipients. *Transplant Reviews* **9**, 231–40.

Penn I (1986). Cancers of the anogenital region in renal transplant recipients. *Cancer* **58**, 611–16.

Penn I (1994). The problems of cancer in organ transplant recipients: an overview. *Transplantation Science* **4**, 23–31.

Rook AH *et al*. (1995). Beneficial effect of low-dose systemic retinoid in combination with topical tretinoin for the treatment and prophylaxis of pre-malignant and malignant skin lesions in renal transplant recipients. *Transplantation* **59**, 179.

Shah KV (1997). Human papillomavirus and anogenital cancers. *New England Journal of Medicine* **337**, 1386–8.

Cardiovascular complications

Arnadottir M, Berg AL (1997). Treatment of hyperlipidaemia in renal transplant recipients. *Transplantation* **63**, 339–45.

Curtis JJ (1991). Distinguishing the causes of post-transplant hypertension. *Pediatric Nephrology* **5**, 108–11.

Fervenza *et al*. (1999). Renal artery stenosis in kidney transplants. *American Journal of Kidney Disease* **31**, 142–8.

Kaisiske BL *et al*. (1996). Cardiovascular disease after renal transplantation. *Journal of the American Society of Nephrology* **7**, 158–65.

Liver disease

Rao KV, Anderson WR (1992). Liver disease after transplantation. *American Journal of Kidney Disease* **19**, 496–501.

Muscoskeletal complications

Julian BA, Quarles LD, Nieman KMW (1992). Musculoskeletal complications after renal transplantation: pathogenesis and treatment. *American Journal of Kidney Disease* **19**, 99–120.

Torregrosa JV, Campistol JM (1999). Reflex sympathetic dystrophy syndrome in renal transplant patients: a mysterious and misdiagnosed entity. *Nephology, Dialysis, Transplantation* **14**, 1364–5.

Haematological complications

Gaston RS, Julian BA, Curtis JJ (1994). Post-transplant erythrocytosis: an enigma revisited. *American Journal of Kidney Disease* **24**, 1–11.

Grupp C *et al*. (1998). Haemolytic uraemic syndrome (HUS) during treatment with cyclosporin A after renal transplantation—is tacrolimus the answer? *Nephrology, Dialysis, Transplantation* **13**, 1629–31.

Diabetes mellitus

Hariharan S *et al*. (1996). Diabetic nephropathy after renal transplantation. *Transplantation* **62**, 632–5.

Vesco L *et al*. (1996). Diabetes mellitus after renal transplantation. *Transplantation* **61**, 1475–8.

Weir MR, Fink JC (1999). Risks for post-transplant diabetes mellitus with current immunosuppressive medications. *American Journal of Kidney Diseases* **34**, 1–13.

Immunosuppression

Denton MD, Magee CC, Sayegh MH (1999). Immunosuppressive strategies in transplantation. *Lancet* **353**, 1083–91.

First MR (1997). An update on new immunosuppressive drugs undergoing preclinical and clinical trials: potential applications in organ transplantations. *American Journal of Kidney Diseases* **29**, 303–17.

Gummert JF, Ikonen T, Morris RE (1999). Newer immunosuppressive drugs: a review. *Journal of the American Society of Nephrology* **10**, 1366–80.

Koene RAP, Hilbrands LB (1998). Choices of long-term immunosuppression in the renal transplantation: balancing the benefits and risks. *Nephology, Dialysis, Transplantation* **13**, 844–6.

Paul LC, Zaltzman J, Cardiella CJ (1995). Prophylactic anti-lymphocyte antibody therapy in kidney transplantation: quo vadis? *Transplant Review* **9**, 200–6.

Chronic allograft nephropathy

Halloran PF, Melk A, Barth C (1999). Rethinking chronic allograft nephropathy: the concept of accelerated senescence. *Journal of the American Society of Nephrology* **10**, 167–81.

Paul LC (1999). Chronic allograft nephropathy. *Kidney International* **56**, 783–93.

Terasaki PI *et al*. (1994). The hyperfiltration hypothesis in human renal transplantation. *Transplantation* **57**, 1450–4.

Pregnancy

Ehrich JHH *et al*. (1996). Repeated successful pregnancies after kidney transplantation in 102 women (report by the EDTA Registry). *Nephrology, Dialysis, Transplantation* **11**, 1314–17.

Recurrance of original disease

Mathew TH (1988). Recurrence of disease following renal transplantation. *American Journal of Kidney Disease* **12**, 85–96.

Outcome

Fischel RJ *et al*. (1991). Long term outlook for renal transplants recipients with 1-year function. *Transplantation* **51**, 118–22.

Hirata M *et al*. (1996). Patient death after renal transplantation: an analysis of its role in graft outcome. *Transplantation* **61**, 1479–83.

Laupacis A *et al*. (1996). A study of the quality of life and cost-utility of renal transplantation. *Kidney International* **50**, 235–42.

Morris PJ *et al*. (1999). Analysis of factors that affect outcome of primary cadaveric renal transplantation in the UK. *Lancet* **354**, 1147–52.

Pratshke J *et al*. (1999). Brain death and its influence on donor organ quality and outcome after transplantation. *Transplantation* **67**, 343–8.

20.7.1 The glomerulus and glomerular injury

John Savill

Introduction

Patients with glomerular injury consequent upon diverse and often poorly understood stimuli present with remarkably stereotyped clinical features (see Section 20.3). Nevertheless, when investigated by percutaneous renal biopsy, clinically similar patients may exhibit an array of histopathological types of glomerular disease that can bewilder all but the most experienced of clinicians. The patterns of glomerular disease that the skilled histopathologist can recognize will continue to be extremely useful to clinicians. However, recent advances in glomerular cell biology point to a complementary approach toward understanding the pathogenesis and present or future treatment of glomerular disorders. Thus, apparently complex histopathological changes can be viewed as the sum of a small number of pathological alterations in the cell biology of the glomerulus. In this scheme, it is apparent that three potentially reversible cellular processes are active to varying degrees in most forms of glomerulonephritis: (1) leucocyte infiltration; (2) changes in the number, size, or phenotype of resident glomerular cells; and (3) changes in the amount and composition of extracellular matrix, which includes the specialized glomerular basement membrane. Furthermore, although three further processes—(4) crescent formation, (5) glomerular capillary thrombosis, and (6) glomerular sclerosis—may be viewed as irreversible events that are best prevented, there is some hope that resolution of such processes and subsequent repair could be encouraged. Consequently, a basic knowledge of the cellular processes that constitute a threat to long-term function in glomerular disease may not only assist in the understanding of complex glomerular pathologies but will also prepare the contemporary clinician for implementing future therapies targeted at key pathological processes.

Key cellular processes in glomerular disease

Leucocyte infiltration

Accumulation of potentially injurious leucocytes is a hallmark of severe inflammatory glomerular injury (Fig. 1).

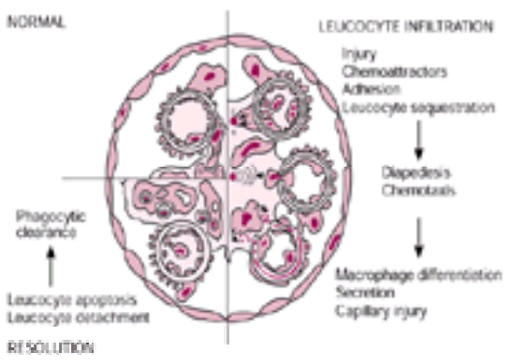

Fig. 1 Leucocyte infiltration in glomerular injury. Generation of chemoattractants and other mediators (consequent upon, for example, immunological injury), causes endothelial cells to express adhesion molecules to which leucocytes bind, these commonly then being sequestered in capillary lumens. However, leucocytes may also migrate towards areas with a high concentration of chemoattractants (chemotaxis), crossing the basement membrane by diapedesis. Monocytes can mature into secretory macrophages, which may further exacerbate capillary injury. Resolution can occur, due to detachment of sequestered leucocytes and apoptosis of leucocytes that have left the bloodstream, the latter leading to anti-inflammatory phagocytic clearance of apoptotic leucocytes.

Granulocytes (usually neutrophils, but sometimes eosinophils) represent the 'rapid response force' of the inflammatory response and are prominent in severe acute inflammatory conditions such as ANCA-positive vasculitis (**ANCA**, antineutrophil cytoplasmic antibody). Granulocytes threaten to exacerbate injury by generating reactive oxygen species and releasing granules yielding toxic cationic proteins, potent degradative enzymes, and monocyte chemoattractants. However, they are also present in self-limited conditions such as poststreptococcal glomerulonephritis, suggesting that their presence does not *per se* direct progressive glomerular injury. Indeed, this shows that granulocytes can be removed safely from inflamed glomeruli. Neutrophils accumulating in the capillary lumen can return to the circulation, while some extravasated cells reaching Bowman's space are flushed out in the urine. Furthermore, neutrophils 'trapped' in the mesangium or in partially occluded glomerular capillaries are cleared away by undergoing constitutive cell death by apoptosis. In turn this leads to anti-inflammatory engulfment of the intact dying cells by phagocytic cells, including macrophages and mesangial cells, with suppression of the phagocyte synthesis of inflammatory mediators by mechanisms involving local release of the anti-inflammatory cytokine transforming growth factor-b1 (**TGF-b1**).

If granulocytes are the 'storm-troopers' of inflammatory responses to tissue injury, then monocyte/macrophages are the 'regimental officers' controlling the attack. Time-course studies in animal models and in human disease emphasize that blood monocytes are recruited within a few hours of neutrophil sequestration in the injured glomerulus. These monocytes mature, by poorly understood mechanisms, into macrophages, which can be very easily overlooked unless detected by specific immunostaining. This confirms abundant macrophage infiltration in many different types of glomerular injury. There is persuasive evidence that macrophages can, depending on circumstances, either exacerbate injury or promote repair. The microenvironment in which macrophages mature, particularly the cytokines present, can irreversibly 'programme' the cells to adopt particular phenotypes. Thus interferon-gamma (**IFN-g**) programmes macrophages to adopt an 'activated/inflammatory' phenotype—which, for example, threatens neighbouring cells with injury or death induced by an enhanced macrophage release of nitric oxide, consequent upon IFN-g-directed induction of the inducible nitric oxide synthase (**iNOS**, also known as **NOS2**). However, if, for example, interleukin-4 (**IL-4**) predominates, macrophages adopt a 'reparative' phenotype in which TGF-b1 secretion will promote extracellular matrix deposition (see below).

T (thymic type) lymphocytes can also infiltrate injured glomeruli, although their presence usually indicates severe damage and a high risk of progression to scarring. T cells have a reciprocal relationship with macrophages—each can tell the other what to do. For example, T cells can summon macrophages to inflamed sites, while macrophage-lineage cells can direct T-cell behaviour by the presentation of antigens on MHC molecules, in the context of other macrophage surface molecules and macrophage-secreted cytokines that help to activate T cells. CD8-positive cytotoxic T cells and their close cousins, the natural killer (**NK**) cells, can injure or kill glomerular cells presenting antigen from infectious or other sources. Helper, CD4-positive T cells are probably important in orchestrating autoimmune glomerular injury, especially where this depends upon autoantibody production by distant B lymphocytes, as in Goodpasture's disease. While T cells may meet their fate by undergoing apoptosis at inflamed sites, these cells are typically 'visitors' circulating through the glomerulus from the blood to the lymphatics, unwanted T cells dying in lymph nodes.

Therapeutic approaches

Glucocorticoids reduce leucocyte infiltration of inflamed tissue by multiple mechanisms, including inhibition of recruitment, promotion of deletion by apoptosis (eosinophils and lymphocytes), and increased phagocyte clearance of dying leucocytes. Immunosuppressive agents such as cyclophosphamide may also reduce leucocyte infiltration, in part by diminishing circulating leucocyte numbers. New therapies in development aim: to inhibit adhesion molecules involved in leucocyte recruitment; to block the action of chemoattractants, such as chemokines or complement fragments; or to inhibit the upregulation of both adhesion molecules and chemoattractants by blocking the action of 'master' proinflammatory cytokines such as tumour necrosis factor-a (**TNF-a**), an approach that has been successful in treating patients with rheumatoid arthritis.

Changes in resident cell number, size, or phenotype

Mesangial cells are smooth muscle-like pericytes of the glomerulus, regulating structure by supporting glomerular capillaries and by modulating the amount or composition of the extracellular matrix (see below). They may also modulate glomerular function, controlling perfusion and filtration by the release of vasoactive mediators and effecting constriction/relaxation. Many glomerular diseases exhibit 'mesangial expansion' or 'mesangial proliferation' in their early stages (Fig. 2), an example being IgA nephropathy. Such pathological features usually reflect an increase in the number of mesangial cells—mesangial hyperplasia, which is due to increased mesangial cell mitosis (true proliferation)—and/or decreased mesangial cell death. Interestingly, however, diabetic nephropathy appears to exhibit a predominant increase in mesangial cell size (hypertrophy) rather than number. Nevertheless, both hyperplasia and hypertrophy involve the activation of intracellular regulatory proteins called cyclin-dependent kinases. In hyperplasia, these cyclin-dependent kinases drive mesangial cells through the cell cycle so that they divide and increase in number, while in hypertrophy such progression does not occur but the cell grows in size. These intracellular controls on mesangial cell number and size are subject to exquisite external controls. Thus, increases in mesangial cell number can be driven by platelet-derived growth factor (**PDGF**) and basic fibroblastic growth factor (**bFGF**). Both mesangial cell hyperplasia and hypertrophy threaten later progression to scarring, this is because these changes in mesangial cell number and size are usually accompanied by the adoption of an abnormal myofibroblast-like phenotype characterized by the expression of a-smooth muscle actin. Such myofibroblast-like mesangial cells are remarkably similar to skin myofibroblasts, which mediate both wound repair and contraction/closure and whose presence is strongly associated with the deposition of excess, abnormal extracellular matrix (see below).

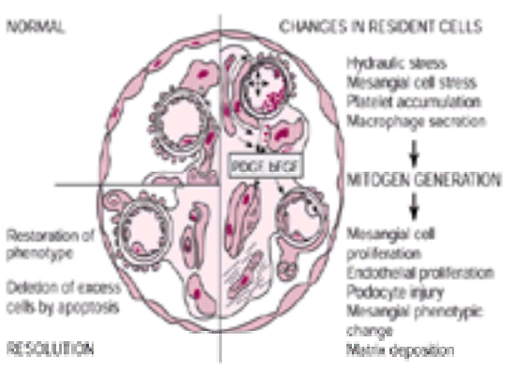

Fig. 2 Resident cell changes in glomerular injury. Hydraulic stress upon mesangial cells (as may occur in glomerular hypertension), secretory macrophages, and recruited blood platelets may all release mitogens such as PDGF (platelet-derived growth factor) or bFGF (basic fibroblastic growth factor). These trigger the proliferation of mesangial and endothelial cells and the phenotypic change seen in these and epithelial cells (podocytes). Such changes are associated with alterations in matrix (see Fig. 3). Resolution involves the deletion of excess resident cells by apoptosis and the restoration of the resident cell phenotype.

Glomerular endothelial cells seem to obey similar 'rules' as mesangial cells. They undergo true proliferation (i.e. hyperplasia) in poststreptococcal and other forms of glomerular injury, although whether this has the ominous implications of a mesangial cell increase is unclear. By contrast, glomerular epithelial cells or podocytes do not readily undergo proliferation in humans after birth: although they can respond to injury by synthesizing DNA and even becoming binucleate, experimental work in rodents emphasizes that adult podocytes rarely divide. Indeed, it is important to emphasize that the number of glomerular resident cells can also decline undesirably, especially when acute glomerular injury progresses to glomerular sclerosis or scarring (see below).

Therapeutic approaches

It has not been established whether any current therapies specifically modulate changes in the resident cell number, size, or phenotype. However, because such changes may be triggered by mechanical cellular stress resulting from an increased glomerular perfusion pressure, some of the beneficial effects of antihypertensive agents, especially angiotensin-converting enzyme (**ACE**) inhibitors and angiotensin II blockers, may prove to be mediated by effects on resident cells. New therapies in development include drugs that specifically inhibit the cyclin-dependent kinases, which may mediate undesirable glomerular cell hyperplasia and/or hypertrophy.

Increased deposition of abnormal extracellular matrix

The normal glomerular extracellular matrix is a network of proteins and proteoglycans (such as heparan sulphate proteoglycan) of critical importance in regulating the survival and properties of glomerular cells, in addition to supporting the glomerular structure. The accumulation of extracellular matrix (**ECM**) is a prominent feature of most glomerular diseases (Fig. 3) and reflects the combined effects of an increased secretion of matrix components and tissue inhibitors of metalloproteinases (**TIMPs**) from glomerular cells. The increased deposition of ECM is thus achieved both by laying down new matrix and preventing the degradation of new or existing ECM. Furthermore, the protein composition of ECM is abnormal in injured glomeruli: there is an accumulation of 'interstitial' type I and type III collagens and of plasma-type fibronectin, alongside an increase in normal constituents such as laminin and type IV collagen. Indeed, there is now a growing body of data to demonstrate that potentially deleterious alterations in glomerular ECM may be dependent upon an excess local secretion (or action) of the cytokine TGF-b1. The propensity of TGF-b1 to promote an undesirable accumulation of ECM is thought to represent the 'dark side' of mechanisms that evolved to promote wound-healing. For example, the accumulation of abnormal ECM is a prominent feature of diabetic nephropathy and may be so exuberant as to form Kimmelstiel–Wilson nodules, constituting a grave threat of progression to scarring. However, in keeping with the dynamic balance between synthesis and degradation that underlies ECM accumulation, abnormal glomerular ECM can be remodelled in self-limited disorders such as some cases of IgA nephropathy.

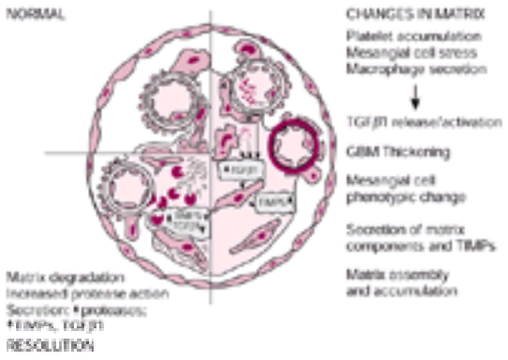

Fig. 3 Extracellular matrix changes in glomerular injury. Mesangial cells (stressed by, for example, hydraulic stimuli), secretory macrophages, and recruited platelets may all release TGF-b1 that is locally activated. In turn, this fibroblastic cytokine causes mesangial cells to adopt a myofibroblastic phenotype, secreting both matrix components and TIMPs, with a net accumulation of abnormal matrix. Similar mechanisms, probably also impinging on endothelial cells, can cause thickening of the glomerular basement membrane (**GBM**). Resolution requires a switch in glomerular cell secretion so that proteases (depicted as 'pacmen') predominate over TIMPs, leading to beneficial matrix degradation.

The general principles of matrix changes in glomerular injury are also evident in the most specialized compartment of the glomerular ECM, the glomerular basement membrane (**GBM**). Thus, thickening of the GBM in cases of diabetic nephropathy largely reflects the accumulation of normal constituents, while grossly similar changes can be caused by the deposition of abnormal proteins, such as amyloid. Examination under polarized light, special stains, immunofluorescence, or electron microscopy may be required to reveal the deposition of abnormal proteins, which can also include immune deposits or fibrils, or intrinsic defects such as those of Alport's syndrome.

Therapeutic approaches

There is no unequivocal evidence that currently available treatments directly and specifically effect changes in the glomerular matrix. New therapies in development include agents to inhibit the profibrotic effects of TGF-b1 (which may be particularly useful in diabetic nephropathy) and drugs that target the control of matrix deposition by antagonizing, for example, TIMPs.

Glomerular crescent formation

Glomerular crescents, abnormal masses of cells filling Bowman's space, are a reflection of severe glomerular injury associated with disorders such as vasculitis, systemic lupus erythematosus (**SLE**), and anti-GBM disease. Indeed, scanning electron microscopy studies of crescentic glomerulonephritis have demonstrated holes in the glomerular basement membrane, consistent with the idea that bleeding into Bowman's space and the deposition of fibrin are crucial pathogenetic factors (Fig. 4). Nevertheless, crescents can be viewed as being a special consequence of the three cellular processes described above. First, monocyte/macrophages and sometimes other leucocytes such as lymphocytes frequently infiltrate them. econd, especially in early crescents, there is prominent evidence of an increase in number and change of phenotype in resident cells, that is to say the 'parietal' epithelial cells of Bowman's capsule, which unlike the 'visceral' glomerular epithelial cells (podocytes) undergo true proliferation in crescentic disease. Third, there is deposition of extracellular matrix, which is obviously abnormal in that it should not be present at all. These cellular processes of glomerular crescent formation are worthy of special consideration because they are widely believed to be incompatible with recovery of function, especially when Bowman's capsule has been breached, potentially allowing ingress of myofibroblasts from the locally injured interstitium.

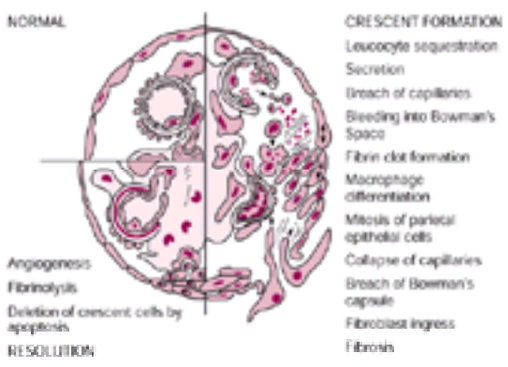

Fig. 4 Crescent formation in severe glomerular injury. Severe glomerular capillary injury, most likely mediated by leucocyte-derived reactive oxygen species and injurious proteins, causes breaches in capillary walls. Bleeding into Bowman's space, with formation of fibrin clot, then ensues. Blood monocytes mature into macrophages and parietal epithelial cells of Bowman's capsule may proliferate. The resultant crescent may compress capillaries, causing their collapse and occlusion, especially when injury is sufficiently severe to allow ingress of fibroblasts from the periglomerular space. It seems that glomerular fibrosis and loss is the rule. However, resolution may occur rarely, for example in poststreptococcal glomerulonephritis. Fibrinolysis, deletion of crescent cells by apoptosis, and repair of damaged capillary networks by angiogenesis may all play a role.

Therapeutic approaches

In some types of crescentic nephritis there is good anecdotal evidence that glucocorticoids/immunosuppressive agents may retard crescent formation by mechanisms likely to include the downregulation of monocyte/macrophage efflux into Bowman's space. Such drugs might also encourage resolution by promoting apoptosis and safe clearance of cells in crescents. Indeed, most nephrologists will have encountered cases of childhood poststreptococcal glomerulonephritis with 100 per cent crescents, which nevertheless exhibit a large degree of resolution, typically in association with immunosuppressive therapy. The fibrin component may be susceptible to anticoagulants/ancrod but new therapies are urgently required. There is active interest in the antagonism of cytokines/chemokines and in the blockade of leucocyte/endothelial adhesion molecules, but new insights into disease mechanisms are needed.

Glomerular capillary thrombosis

Thrombotic occlusion of glomerular capillaries is another manifestation of severe injury, consequent upon disorders such as systemic vasculitis, SLE, malignant hypertension, and radiation nephritis. Formation of platelet thrombi in glomeruli often reflects loss/retraction of glomerular endothelial cells with exposure to blood elements of the glomerular basement membrane (Fig. 5). Indeed, thrombosis often goes hand in hand with segmental necrosis of glomeruli. In such circumstances there would appear to be little prospect of successful repair and scarring of the affected lobule/glomerulus. However, all may not be lost: the astonishing capacity of mesangial cell precursors to repopulate glomeruli completely denuded of mesangial cells by experimental antibody-mediated injury is a testament to the capacity that the glomerulus may have—under suitable circumstances—to 'rebuild' itself. In keeping with such observations, recent data emphasize that angiogenesis, the growth of new blood vessels, can and does occur in the glomerular response to injury, under the direction of cytokines such as vascular endothelial growth factor (**VEGF**). This raises the possibility that such blood vessel growth could replace occluded, destroyed glomerular capillaries.

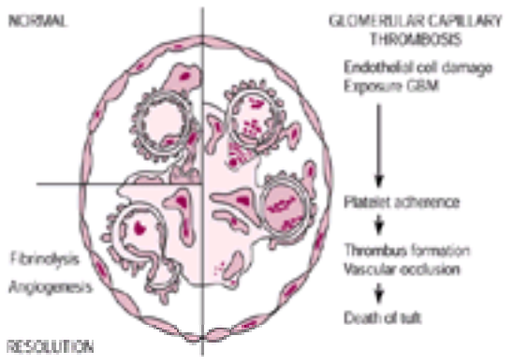

Fig. 5 Glomerular capillary thrombosis. Platelet adherence to the glomerular capillary may reflect mediator generation within the glomerulus, causing endothelial cell activation with adhesion molecule expression, or endothelial cell retraction leading to GBM exposure. Endothelial cells may also be damaged by bloodborne elements or ischaemia–reperfusion injury, as may occur in sickle cell disease. Tufts in which glomerular capillaries are occluded by thrombosis usually die by ischaemic necrosis. However, resolution is possible, since fibrinolysis can result in the recanalization of capillaries, which may be repaired by angiogenesis.

Therapeutic approaches

Clinicians dealing with glomerular capillary thrombosis frequently consider using anticoagulants/antiplatelet drugs, and growing experience of using thrombolytic agents in other vascular territories may lead to their increased use in the hope of recanalizing occluded vessels. New therapies under active consideration are likely

to concentrate on promoting recovery, but selective stimulation of angiogenesis capable of leading to glomerular repair is a distant prospect.

Glomerular scarring/sclerosis

Leucocyte infiltration, increased glomerular resident cell number/size, increased deposition of abnormal extracellular matrix, glomerular crescent formation, and glomerular capillary thrombosis may all be reversible to varying extents. However, all too frequently these cellular pathologies set the scene for glomerular sclerosis (Fig. 6). Nevertheless, although scarred, functionless glomeruli are beyond resurrection. However, recent insights into the pathogenesis of glomerular sclerosis offer some prospect of prevention and are therefore worthy of consideration.

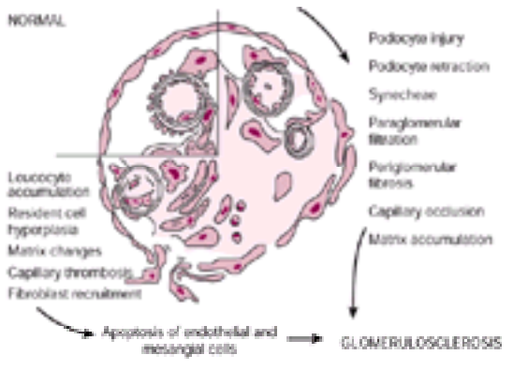

Fig. 6 Two pathways to glomerulosclerosis. At the bottom left, leucocyte accumulation, resident cell hyperplasia, matrix changes, capillary thrombosis, and recruitment of extraglomerular fibroblasts set up a situation in which progressive and unscheduled apoptosis of mesangial and endothelial cells leads to a featureless, non-functioning glomerular scar. This endstage glomerular lesion may also arise because of podocyte injury and retraction (top right), as may occur in HIV infection. 'Naked' GBM then adheres to Bowman's capsule because of local matrix deposition. As the adhesion enlarges, glomerular filtration may occur directly into the paraglomerular space (i.e. 'under' Bowman's capsule) causing gross periglomerular fibrosis and capillary occlusion, with ultimate progression to functionless scarring.

Although an increased number of cells in glomeruli (reflecting leucocyte infiltration and resident cell hyperplasia) characterizes potentially progressive glomerulonephritis, the scarred glomerulus is striking for its lack of cells, being replaced by featureless, acellular matrix. Growing evidence points to the undesirable deletion of resident glomerular cells via unscheduled apoptosis as a final common pathway in glomerular sclerosis. The mechanisms responsible are likely to be complex, since a wide range of possible apoptotic stimuli is likely to occur in injured glomeruli, but the potential consequences of glomerular cell loss are rather easier to appreciate. For example, loss of endothelial cells from glomeruli, as observed in models of progressive glomerular injury, obviously poses a hazard for the disruption of the glomerular blood supply, structure, and function. Such unscheduled loss may be prevented by the antiapoptotic properties of cytokine survival factors. Thus, recent studies suggest that the 165 amino acid isoform of VEGF is of central importance for endothelial cell survival in glomerular injury, while insulin-like growth factor-1 (**IGF-1**) may play a similar role in retarding mesangial cell apoptosis, emphasizing that glomerular cell loss might be preventable.

Recent work, based on elegant morphological studies by Kriz and colleagues, has emphasized the subtle and destabilizing effects of unscheduled glomerular cell loss upon the biomechanics of a delicate multicellular structure, that must contain blood at a much greater pressure than is usual in capillary networks. Loss of supporting mesangial cells is associated with glomerular capillary dilatation, which exerts deleterious mechanical stress upon both glomerular endothelial and epithelial cells, threatening their unscheduled death. Loss of podocytes (see Fig. 6) may be particularly dangerous because of the limited capacity of glomerular epithelial cells for mitosis; rather like precious neurones, if podocytes die they may not be replaced. The consequences of such loss include the formation of tuft adhesions to Bowman's capsule, which may allow abnormal filtration into the interstitium and promote periglomerular fibrosis, the adhesion propagating around an ever-more constricted glomerular tuft. This model may be particularly pertinent to the pathogenesis of focal segmental glomerulosclerosis (**FSGS**). Indeed, the model also predicts the course of the 'collapsing' form of FSGS seen in human immunodeficiency virus (HIV) infection, in which apparent podocyte dedifferentiation seems to deny the glomerulus an essential podocyte-mediated force that counterbalances mesangial cell-mediated constriction of the glomerulus.

Therapeutic approaches

All clinicians employ antihypertensives in the hope of retarding the progression of glomerular scarring, perhaps reducing mechanical stresses that trigger undesirable apoptosis. However, dietary protein restriction has yet to achieve widespread use, despite some encouraging trial evidence and animal data that this reduces potentially harmful glomerular hyperfiltration. New therapies are likely to concentrate on dealing with the cellular pathologies described above in the hope of preventing progression. Nevertheless, the advent of stem-cell therapies and tissue engineering raises the very distant prospect of growing replacement tissue/organs in the laboratory.

Conclusions

When confronted by apparently complex glomerular histopathological changes, it may be helpful to think of the six cellular pathologies described above. Although glomerular histopathology can provide extremely useful prognostic information, considering the cellular pathologies at the root of the problem may help the clinician address the primary concern as to whether the glomerular injury is likely to be amenable to treatment. This will be especially important as new therapies that target pathological cellular processes become available.

Further reading

Kriz W, Lemley KV (1999). The role of the podocyte in glomerulosclerosis. *Current Opinion in Nephrology and Hypertension* **8**, 489–97.

Preisig P (1999). What makes cells grow larger and how do they do it? Renal hypertrophy revisited. *Experimental Nephrology* **7**, 273–83.

Savill J (1999). Regulation of glomerular cell number by apoptosis. *Kidney International* **56**, 1216–22.

Savill J, Rees AJ (1998). Mechanisms of glomerular inflammation. In: Davison AM, *et al.*, eds. *Oxford textbook of clinical nephrology*, 2nd edn, pp 403–40. Oxford University Press, Oxford.

Shankland S, Al'Douahji M (1999). Cell cycle regulatory proteins in glomerular disease. *Experimental Nephrology* **7**, 207–11.

20.7.2Iga nephropathy and Henoch–Schönlein purpura

John Feehally

Introduction and definitions

IgA nephropathy

IgA nephropathy (**IgAN**) is the commonest pattern of glomerulonephritis identified in areas of the world where renal biopsy is frequently performed. It was first described by Berger in 1968 and at one time was known as Berger's disease. It is defined by IgA deposition in the glomerular mesangium, accompanied by a mesangial proliferative glomerulonephritis which may vary greatly in severity. Although recurrent macroscopic haematuria is the hallmark of the disease, the old term 'benign recurrent haematuria' is a discredited misnomer since it is now clear that IgAN is an important cause of endstage renal failure (**ESRF**).

Henoch–Schönlein purpura

Henoch–Schönlein purpura (**HSP**) is a somewhat misleading historical term. The purpuric rash is a cutaneous vasculitis (Plate 1), and HSP is a small-vessel systemic vasculitis characterized by IgA deposition in affected blood vessels. The renal lesion (HSP nephritis) is a mesangial proliferative glomerulonephritis, usually indistinguishable from IgAN.

Aetiology and pathogenesis

Mechanism of mesangial IgA deposition

Mesangial proliferative glomerulonephritis, such as is seen in IgAN and HSP nephritis, may be the consequence of immune complex deposition, either due to trapping of circulating IgA immune complexes or the formation of complexes *in situ* by the reaction of IgA with antigen that has already been deposited. No exogenous antigen has consistently been identified in the mesangial deposits in IgAN, which may indicate that the IgA complexes are a common response to different antigens, or that the initiating antigen has disappeared by the time of the renal biopsy. Alternatively, the IgA may be deposited by some mechanism independent of classical antigen–antibody interactions, such as a physicochemical abnormality of the IgA.

The frequent recurrence of both IgAN and HSP nephritis after renal transplantation strongly suggests that the abnormality resides in the host IgA immune system. The mesangial IgA deposits are polymeric IgA1 (pIgA1). Most pIgA is synthesized in the mucosa, and the clinical association of macroscopic haematuria with mucosal infection originally led to the assumption that an exaggerated mucosal IgA response resulted in mesangial IgA deposition. However, IgA production is downregulated in the mucosal immune system and upregulated in the bone marrow, and exaggerated IgA1 responses to immunization in these patients are marrow rather than mucosally derived.

There is increasing evidence of abnormal glycosylation of both serum and mesangial IgA1 in patients with IgAN and HSP nephritis. The glycosylation abnormality may favour the development of immune complexes or may directly provoke mesangial deposition, but these putative mechanisms have not yet been further defined.

Progression of IgA nephropathy

IgA deposition may occur in many patients with mild disease with little mesangial injury. What decides the prognosis in any individual is the extent to which IgA deposition is followed by mesangial proliferation, inflammation, and scarring. There is nothing to suggest that these subsequent mechanisms of damage and scarring are unique to IgAN, rather they are generic to many forms of glomerulonephritis.

Relationship of IgAN and HSP

There is much indirect evidence to suggest a close relationship between IgAN and HSP. Monozygotic twins have been described, one developing IgAN and the other HSP at the same time. HSP developing on a background of proven IgAN has been described in both adults and children. Many abnormalities of the IgA immune system, including abnormal IgA1 glycosylation, have been described in both conditions. IgAN is increasingly thought of as 'HSP without the rash'. Why some individuals get a renal-limited disease (IgAN) and others a systemic disease (HSP) is not known.

Epidemiology

IgAN is the commonest glomerulonephritis in countries where renal biopsy is widely used, typically found in 30 per cent of biopsies with primary glomerular disease, but the apparent prevalence varies markedly around the world. It is commoner in the Pacific Rim and Mediterranean countries, less so in North America and Northern Europe. At least part of this apparent difference is explained by variations in the use of urine testing in health screening and varying attitudes to the value of renal biopsy in individuals with isolated haematuria or other minor clinical evidence of renal disease. In Japan, for example, where there is routine urine testing of schoolchildren and employed adults, the threshold for renal biopsy is low, and the reported prevalence of IgAN is high.

There are also important racial differences in susceptibility. IgAN is uncommon in Afro-Caribbeans. It is also less common in Polynesians than Caucasians in Australasia—a particularly striking finding given the exaggerated susceptibility of Polynesians to most forms of renal disease. Despite many studies of potential immunogenetic associations, the genetic basis for such variations in susceptibility to IgAN has not yet been identified. IgAN is occasionally familial, one very large kindred having been described in Kentucky, in the United States; but the great majority of cases are sporadic.

Clinical features

IgA nephropathy

Macroscopic haematuria

IgAN can occur at any age, but the peak age of onset is in the second and third decades of life (Fig. 1). IgAN is three times more common in males than females.

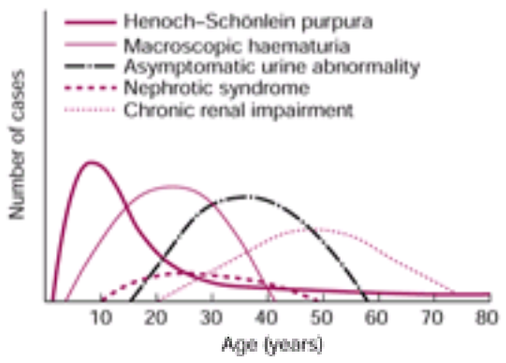

Fig. 1 Clinical presentations of IgA nephropathy (IgAN) and Henoch–Schönlein purpura (HSP) in relation to age at diagnosis. HSP is most common in childhood but may occur at any age. Macroscopic haematuria is very rare in people over the age of 40 years. The importance of an asymptomatic urine abnormality as the presentation of IgAN will depend on attitudes to routine urine testing and renal biopsy. It is uncertain whether those presenting with chronic renal impairment have a disease distinct from that of those presenting at younger ages with macroscopic haematuria. (Reproduced from Johnson RJ, Feehally J. *Comprehensive clinical nephrology*. London: Harcourt Publishers, 1999: 26.3, with permission.)

The characteristic clinical picture of recurrent macroscopic haematuria occurs in about 40 to 50 per cent of cases. A child or young adult develops episodes of painless macroscopic haematuria occurring within a day or so of the onset of an upper respiratory tract infection, or occasionally infections of other mucosal or IgA-secreting surfaces such as the gastrointestinal tract, bladder, or breast. The urine may be frankly bloody, but more often is brown (like 'Coca-Cola' or tea without milk), there are no clots passed and it is usually painless, although there may be dull loin ache. The episodes settle spontaneously after 1 to 5 days and may be recurrent, but rarely for more than a year or two. Serum IgA is moderately elevated in 30 per cent of cases but serum complement C3 and C4 levels are normal. Between episodes there will be persistent microscopic haematuria. This presentation does not occur beyond the age of 40 years (Fig. 1).

Asymptomatic haematuria/proteinuria

Some 30 to 40 per cent of cases of IgAN are identified by urine testing—microscopic haematuria may be combined with proteinuria (usually <2 g/24 h). Since this is glomerular haematuria, dysmorphic red cells may be seen on phase-contrast microscopy, but red cell casts are frequently absent in mild disease.

Nephrotic syndrome

Nephrotic syndrome is the presentation in only 5 per cent of patients with IgAN. Very occasionally in children or young adults this appears to be the consequence of coincidental minimal-change nephrotic syndrome: the proteinuria resolves completely with corticosteroid therapy, but haematuria and IgA deposits persist. More commonly, nephrotic syndrome may develop in cases of IgAN with overt mesangial proliferative glomerulonephritis, or it may be a consequence of glomerular scarring in advanced IgAN.

Acute renal failure

Acute renal failure occurs for two reasons in patients with IgAN. Episodes of macroscopic haematuria may produce acute tubular occlusion by red cells in the face of minor glomerular injury. Alternatively, there can be acute severe necrotizing glomerulonephritis with crescent formation—'crescentic IgA nephropathy', which may be the presenting feature or occur on a background of known milder disease.

Chronic renal failure

Patients with IgAN may also present with hypertension and established renal impairment. This often occurs in older patients (Fig. 1). Too little is yet known about the pathogenesis of IgAN to understand whether this is a distinct disease entity, or simply the same disease presenting much later in the absence of macroscopic haematuria or a urine test to bring it to earlier medical attention.

Clinical associations with IgA nephropathy

The commonest secondary cause of IgAN is chronic liver disease, typically alcoholic liver disease, in which it is probable that IgA deposition is a consequence of impaired IgA clearance from the circulation via the liver. Most hepatic IgAN is asymptomatic, and progression to ESRF is unusual. The other best established associations are with coeliac disease and dermatitis herpetiformis; with rheumatoid arthritis, ankylosing spondylitis and Reiter's disease; and with HIV infection. Many other conditions have been reported occasionally with IgAN, but since IgAN is so common it is difficult to know if these are more than chance associations.

Henoch–Schönlein purpura nephritis

HSP can occur at any age but is commonest in the first decade of life (Fig. 1). There is a slight male preponderance. A palpable purpuric rash caused by cutaneous vasculitis is the presenting feature. It has a characteristic extensor surface distribution, with sparing of the trunk and face. Crops of rash, often provoked by intercurrent infection, may continue for some time, but rarely beyond a year from first presentation. Polyarthralgia is common. Abdominal pain, due to gut vasculitis, is usually mild and transient, but severe pain and bloody diarrhoea may develop due to intussusception.

Apart from intussusception the major sequelae of HSP come from renal involvement. Renal disease in HSP is transient in many cases, asymptomatic haematuria or proteinuria disappearing in a few weeks. Of those with persistent evidence of renal disease, asymptomatic haematuria and proteinuria is the commonest clinical state, but 20 per cent will have nephrotic syndrome. Serum IgA is raised in 50 per cent of patients, but complement C3 and C4 levels are normal. Acute renal failure due to crescentic HSP nephritis usually occurs early and is commoner than crescentic IgAN.

Pathology

Immune deposits

IgAN and HSP nephritis are defined by the presence of mesangial IgA detected by immunofluorescence or immunoperoxidase staining (Plate 2). C3 frequently accompanies IgA in the same mesangial distribution, IgG and IgM are less common. Electron microscopy identifies mesangial electron-dense deposits corresponding to the mesangial IgA (Fig. 2).

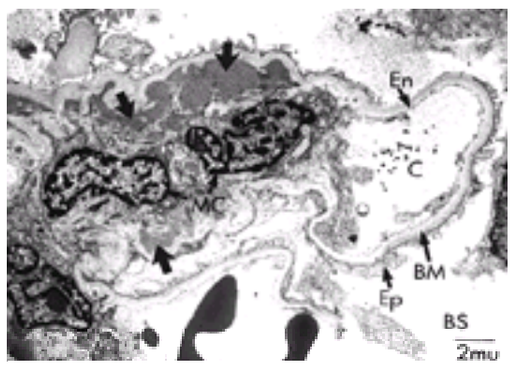

Fig. 2 Electron micrograph of a glomerular capillary loop in IgA nephropathy. Numerous electron-dense deposits representing deposits of IgA (large arrows) are seen within the expanded mesangium (5200 ×). BM, basement membrane; C, capillary lumen; Ep, visceral epithelium; En, fenestrated endothelium; MC, mesangial cell nucleus; BS, Bowman's space.

Light microscopy

Mesangial proliferative glomerulonephritis is the characteristic appearance. Although when haematuria is the only clinical finding, abnormalities seen on light microscopy may be minimal despite florid IgA deposition. Mesangial hypercellularity and matrix expansion are usually global but may be focal and segmental. The hypercellularity is followed by increasing mesangial matrix deposition and eventual sclerosis (Plate 3). In acute renal failure there may be severe glomerular inflammation with crescent formation. In advanced cases there is glomerulosclerosis and corresponding tubular atrophy and interstitial fibrosis, which are entirely non-specific changes of 'endstage kidney'.

Diagnosis and differential diagnosis

IgA nephropathy

By definition, the diagnosis of IgAN requires a renal biopsy: no serological or other laboratory indices provide diagnostic information reliable enough to avoid the need for tissue.

Macroscopic haematuria

Non-glomerular causes of haematuria must always be considered, including renal stones and neoplasia, and excluded where appropriate by urological investigation. While episodic macroscopic haematuria coinciding with an upper respiratory tract infection in children and young adults is the hallmark of IgAN, it is not pathognomonic. Similar episodes can occur with other glomerular diseases, most commonly hereditary nephropathies such as Alport's syndrome and thin membrane nephropathy. The distinction of IgAN from postinfectious (usually post-streptococcal) glomerulonephritis is also important. In post-streptococcal glomerulonephritis there is a 10- to 14-day latency period from the onset of infection and the development of symptomatic renal disease, contrasting with the immediacy of haematuria in IgAN for which the term 'synpharyngitic haematuria' has been coined. The haematuria is usually less heavy in post-streptococcal glomerulonephritis, such that the urine is typically smoky rather than frankly bloody; hypertension, oedema, and other features of the acute nephritic syndrome are usually present. Serological evidence of a recent streptococcal infection (such as antibodies to endostreptosin) and a low C3 level are not found in IgAN.

Nephrotic syndrome

The differential diagnosis when IgAN presents with nephrotic syndrome includes the usual range of glomerular diseases known to cause nephrotic syndrome given the age of the patient.

Chronic renal failure

Advanced IgAN presenting with hypertension, proteinuria, and renal impairment is clinically indistinguishable from many other causes of chronic progressive renal disease. If it is considered important to attempt a precise diagnosis, renal biopsy remains a valuable diagnostic tool since mesangial IgA can often still be identified even when light microscopy shows 'endstage kidney' disease.

Henoch–Schönlein purpura

In children HSP is the commonest form of vasculitis. A clinical diagnosis is often made from the characteristic rash and abdominal pain, but ultimate confirmation requires identification of tissue IgA deposition, which can be found in the vessels of affected skin as well as the kidney. In adults the differential diagnosis is wider, including many other forms of small-vessel vasculitis that must be distinguished on the basis of clinical, serological, and histopathological findings.

Prognosis

IgA nephropathy

Some 30 per cent of children will have a spontaneous clinical remission with complete disappearance of haematuria within 10 years of diagnosis. But IgAN, despite the apparently benign presentation in many cases, is an important cause of endstage renal failure (ESRF). Up to 25 per cent of patients reach ESRF within 20 years of diagnosis. Where a lower risk of ESRF is reported the series will contain larger numbers of patients with mild disease, such as those with isolated microscopic haematuria.

Perhaps unexpectedly, a history of episodic macroscopic haematuria is a favourable prognostic feature. The prognosis for patients who present with microscopic haematuria and minimal proteinuria (<1 g/24 h) is very good, but not perfect. Even in this group up to 5 per cent of patients will develop worsening proteinuria and hypertension during follow-up and are at eventual risk of ESRF. Consequently the long-term follow-up of any patient with biopsy-proven IgAN is mandatory. The risk of progressive renal failure can be predicted by clinical and pathological features at diagnosis (Table 1). These predictive features are not specific to IgAN, but identify the risk of progression in any glomerular disease.

Renal transplantation

Both IgAN and HSP nephritis recur after renal transplantation. Mesangial IgA deposits appear within a few months in 60 per cent of patients with IgAN. Initially this is benign, accompanied by little mesangial injury, but recurrent disease in the long term will contribute to progressive graft loss in a number of patients. However, overall transplant success and graft longevity do not differ in patients with IgAN or HSP from other primary renal disease. The changes in immunosuppressive regimens used to prevent rejection over the last two decades have not altered the recurrence rate or its prognosis.

Treatment

IgA nephropathy

Treatment proposals for IgAN are summarized in Table 2. Only in a small minority of patients with IgAN is there any evidence that drug therapy alters the natural history of the disease. Despite being so common among renal diseases, there is still a dearth of well-conducted, prospective, randomized controlled trials in IgAN on

which to base therapeutic decisions.

Specific treatment for IgAN would either restrict the formation of relevant pathogenic IgA molecules or prevent their deposition in the mesangium. So little is understood about the pathogenesis of the disease that the prospect for such treatment is still remote.

Haematuria

There is no specific treatment for the great majority of patients with IgAN who have isolated haematuria, with or without low-grade proteinuria (<1 g/24 h).

Microscopic haematuria should merely be observed. Recurrent macroscopic haematuria settles without treatment: there is no role for prophylactic antibiotics, and in any case the majority of precipitating infections are viral. Tonsillectomy may reduce the number of episodes of macroscopic haematuria, but there is no evidence that it reduces the risk of progressive renal failure.

Proteinuria

Those with proteinuria above 1 g/24 h in addition to haematuria have a worse prognosis. Immunosuppressive therapies have been tried, although the frequent recurrence of IgAN in transplanted kidneys when patients are receiving immunosuppressive therapy argues against their value. Short-term, randomized controlled trials of corticosteroids have shown no benefit. However, a 6-month controlled trial of treatment with corticosteroids (prednisolone 0.5 mg/kg per day) showed a significant reduction in proteinuria and a reduced risk of developing renal impairment at 5 years' follow-up. This requires further confirmation: corticosteroid treatment is not presently recommended, except in the rare circumstance where the biopsy suggests coincidental minimal-change nephrotic syndrome which may be fully steroid-responsive. All with proteinuria above 1 g/24 h should receive an ACE inhibitor to minimize protein excretion.

Other immune-modulating drugs have been tried in the treatment of IgAN, including cyclophosphamide, azathioprine, ciclosporin, and pooled human intravenous immunoglobulin. However, there are few properly controlled studies, and for none is there consistent evidence of benefit or an acceptable risk–benefit ratio in the great majority of patients who have indolent slowly progressive disease.

Acute renal failure

A renal biopsy is essential when acute renal failure develops in patients with IgAN. If the biopsy shows mild glomerular disease but tubular occlusion with erythrocytes and accompanying acute tubular necrosis, supportive treatment is required while recovery is awaited. If there is crescentic IgAN, a regimen such as that used for renal vasculitis and other forms of crescentic glomerulonephritis should be considered, unless the histological appearances are thought to be advanced and irreversible. Such treatment would typically include oral prednisolone 0.5 mg/kg per day (reducing to a maintenance dose of 5 to 10 mg daily by 3 months), and oral cyclophosphamide 2 to 3 mg/kg per day, the latter being replaced by azathioprine 2 to 3 mg/kg per day after 3 months. There are no randomized controlled trials of these treatments in crescentic IgAN. Although the initial response to treatment is excellent, the medium-term outlook is much less good, and up to 50 per cent of patients may be on long-term dialysis after 12 months.

Progressive renal impairment

Slowly progressive renal impairment due to IgAN requires a management approach common to any form of chronic renal failure. Rigorous control of blood pressure is the one established method of delaying progressive renal failure. Angiotensin-converting enzyme (**ACE**) inhibitors are widely used as first-line therapy for their special role in lessening proteinuria and giving a degree of blood pressure control, although there are no specific prospective studies to prove their additional efficacy in the treatment of IgAN compared to other hypotensive drugs. Fish oil therapy (which provides a supplement of w-3 fatty acids) has effects likely to impact favourably on mechanisms of progressive renal damage and has been used in randomized controlled trials in IgAN, but there is no reason to expect its effects are specific for IgAN, rather than other progressive disease. One such trial has shown a substantial reduction in the risk of progression to ESRF, but other studies have not shown comparable benefit and at present the use of fish oil is not recommended until confirmatory studies are available.

Henoch–Schönlein purpura nephritis

There is very little information to guide the treatment of patients with HSP nephritis. As there are no published randomized controlled trials, and most therapeutic studies in IgAN exclude patients with HSP, it is unclear whether their conclusions can be extrapolated to HSP.

Transient, early nephritis requires no specific treatment. There is no evidence that corticosteroids or other immunosuppressive regimens alter the natural history of the nephrotic syndrome or slowly progressive glomerular damage in Henoch–Schönlein purpura. Crescentic HSP nephritis is more common than crescentic IgAN. Regimens used in the therapy of renal vasculitis have also been applied to crescentic HSP nephritis with apparent benefit, although there are no controlled trials.

Further reading

Clinical

Galla JH (1995). IgA nephropathy. *Kidney International* **47**, 377–87. [Clinical overview of IgAN]

D'Amico G (2000). Natural history of idiopathic IgA nephropathy: role of clinical and histological prognostic factors. *American Journal of Kidney Disease* **36**, 227–37. [Comprehensive review of the natural history of IgAN]

Pouria S, Feehally J (1999) Glomerular IgA deposition in liver disease. *Nephrology, Dialysis, Transplantation* 14, 2279–82. [Review of hepatic IgAN]

White RHR (1994). Henoch–Schönlein nephritis. *Nephron* **68**, 1–9. [A clinical review of HSP nephritis including long-term outcome]

Pathogenesis

Feehally J (1999). Pathogenesis of IgA nephropathy. *Annales Medicine Interne* **150**, 91–8.

van Es LA, de Fijter JW, Daha MR (1997). Pathogenesis of IgA nephropathy. *Nephrology* **3**, 3–12.

Treatment

Dillon JJ (1997). Fish oil therapy for IgA nephropathy. Efficacy and interstudy variability. *Journal of the American Society of Nephrology* **8**, 1739–44. [Meta-analysis of fish oil studies in IgAN]

Donadio JV *et al*. (1999). The long term outcome of patients with IgA nephropathy treated with fish oil in a controlled trial. *Journal of the American Society of Nephrology* **10**, 1772–7. [Evidence of benefit of fish oil in IgAN]

Feehally J (1999). IgA nephropathy and Henoch–Schönlein purpura. In: Pusey CD, ed. *Treatment of glomerulonephritis*, pp 93–112. Kluwer Academic Publishers, Dordrecht. [Review of all treatment evidence in IgAN and HSP nephritis, except new information on corticosteroids and fish oil cited in the other references here]

Pozzi C *et al*. (1999). Corticosteroids in IgA nephropathy: a randomized controlled trial. *Lancet* **353**, 883–7. [A randomized controlled trial of corticosteroids showing benefit in IgAN]

Roccatello D, *et al*. (1995) Report on intensive treatment of extracapillary glomerulonephritis with focus on crescentic IgA nephropathy. *Nephrology, Dialysis, Transplantation* **10**, 2054–9. [Best available evidence on treatment of crescentic IgAN]

20.7.3 Thin membrane nephropathy

John Feehally

Introduction and definition

Thin membrane nephropathy (**TMN**) must always be considered alongside IgA nephropathy (**IgAN**) in the differential diagnosis of glomerular haematuria. TMN is an autosomal dominant condition diagnosed by examination of a renal biopsy by electron microscopy, which shows thin but otherwise morphologically normal glomerular basement membranes (**GBM**). The term 'benign familial haematuria' was used before the GBM abnormality had been identified.

Aetiology and pathogenesis

The genetic basis for TMN has not been defined, although it seems probable that defects in type IV collagen or other GBM proteins will eventually be identified. The gene defects in the a-3 and a-5 chains seen in Alport's syndrome are not found in TMN. A deletion in the gene for the a-4 chain of type IV collagen has been identified in one kindred, but this is not confirmed in other families and it is likely that TMN is genetically heterogeneous.

Pathology

The GBM is diffusely thin but otherwise morphologically normal (Fig. 1). This contrasts with Alport's syndrome in which the GBM is thickened and lamellated and the normal lamina densa of the GBM is disrupted. The normal range for GBM thickness must be determined in each laboratory because of the influence of techniques used for tissue fixation, but typically normal GBM thickness is between 350 and 450 nm. A uniform reduction to less than 250 nm is diagnostic of TMN.

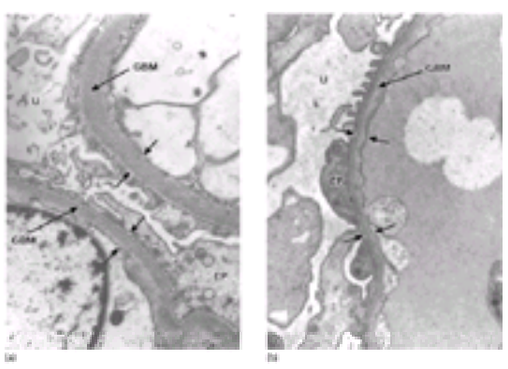

Fig. 1 Thin membrane nephropathy. Electron micrographs contrasting (a) glomerular basement membranes of normal thickness (350–450 nm) with (b) uniform membrane thinning (150–200 nm) in thin membrane nephropathy (both micrographs 20 000 ×). GBM, glomerular basement membrane; Ep, visceral epithelial cells; U, urinary space. (Glutaraldehyde fixation with osmium post-fixation. Ultrathin resin sections stained with uranyl acetate and lead citrate.)

Clinical features

TMN is common and is estimated to be the diagnosis in 20 to 25 per cent of patients presenting to a nephrologist with isolated microscopic haematuria. Autopsy studies suggest it may be present in 5 to 9 per cent of the population. It is an autosomal dominant condition but may also be sporadic. Persistent microscopic haematuria is usually lifelong, and episodic macroscopic haematuria may also occur. Proteinuria is uncommon and progressive renal impairment is rare but has been described in a number of families. Deafness and other extrarenal manifestations seen in Alport's syndrome are absent. There is no specific treatment.

Differential diagnosis

TMN can only be distinguished from IgAN by renal biopsy. Although the coexistence of TMN and IgAN is well recorded, it is a matter of debate whether this merely represents the coincidence of two common glomerular diseases. TMN must be distinguished from Alport's syndrome (hereditary nephritis with deafness), of which the commonest form is X-linked. If there is a clear autosomal dominant pattern of haematuria without renal insufficiency or extrarenal problems a clinical diagnosis of TMN may be established with reasonable confidence, but a renal biopsy in at least one family member is still preferable. Once the diagnosis is established in a kindred, biopsy is not required unless there are unexpected clinical findings. The differentiation from the less common autosomal forms of Alport's syndrome may be less straightforward. Subclinical deafness must be excluded by audiography if necessary. The renal biopsy also requires particularly careful assessment: in TMN there is uniform thinning; early in the course of Alport's syndrome, even if the characteristic structural disruption of the GBM has not yet developed, marked variability in GBM width is typical. Staining of GBM for the a-chains of type IV collagen is highly informative since in Alport's syndrome a-3, a-4, and a-5 are absent, whereas normal a-chain distribution is preserved in TMN.

Prognosis

The prognosis is excellent in the great majority of families with TMN, but there is a small but real risk of developing chronic renal failure, identified by the onset of proteinuria and hypertension. Long-term follow up of those with TMN is therefore required: urinalysis and measurement of blood pressure and renal function is recommended every 1 to 2 years.

Further reading

Dische FE, *et al*. (1990). Incidence of thin membrane nephropathy: morphometric investigation of a population sample. *Journal of Clinical Pathology* **43**, 457–60. [Information on the population incidence of TMN]

Kashtan CE (1998). Alport syndrome and thin glomerular basement membrane disease. *Journal of the American Society of Nephrology* **9**, 1736–50. [Review of molecular basis and diagnosis of TMN]

Nieuwhof CM, *et al.* (1997). Thin GBM nephropathy. Premature glomerular obsolescence is associated with hypertension and late onset renal failure. *Kidney International* **51**, 1596–601. [Evidence that TMN may be associated with progressive renal failure]

Tiebosch AT, *et al*. (1989). Thin-basement-membrane nephropathy in adults with persistent hematuria. *New England Journal of Medicine* **320**, 14–18. [The first prospective study of TMN]

20.7.4 Minimal-change nephropathy, focal segmental glomerulosclerosis, and membranous nephropathy

D. Adu

Classification of glomerulonephritis

The most helpful classification of glomerulonephritis is one based on histology. Careful clinical and pathological studies have established the histological patterns of glomerulonephritis in patients with a nephrotic syndrome inhabiting temperate regions of the world (Table 1). The aetiology and patterns of glomerulonephritis in tropical countries differ considerably and are considered elsewhere (see Chapter 20.7.10): discussion in this chapter refers to disease seen in temperate regions. Idiopathic glomerulonephritis accounts for 90 per cent of all childhood cases of the nephrotic syndrome and for approximately 80 per cent in adult patients. Although these histological changes are usually of unknown aetiology, they may also be secondary to well-defined aetiological factors.

General clinical approach

Children

In the original studies of the International Study of Kidney Diseases in Children (**ISKDC**) the diagnosis of minimal-change nephropathy was based on renal biopsies. From these and other studies it was established that for a child aged between 1 and 6 years with nephrotic syndrome and highly selective proteinuria, and who did not have microscopic haematuria, hypertension, or renal impairment, the likely diagnosis was minimal-change nephropathy. When treated with steroids, such children had a greater than 90 per cent chance of going into remission within 4 weeks. Based on these observations, children of this age with the features summarized above are no longer subjected to renal biopsy, but instead are treated with a trial of steroids. This leads to the term 'steroid-responsive nephrotic syndrome of childhood' and most, but not all, of such children will have minimal-change nephropathy. If the proteinuria does not respond to steroids at 1 month then a renal biopsy should be considered to establish the diagnosis. Children over 8 years of age are more likely to have a steroid non-responsive lesion and probably need a renal biopsy. In neonates and in children under 1 year of age there is a high probability of the congenital nephrotic syndrome or diffuse mesangial sclerosis, and therefore renal biopsy should be considered: neither of these lesions respond to steroids.

Adults

Only 20 per cent of adults with a nephrotic syndrome have minimal-change nephropathy and for that reason a renal biopsy is necessary to establish the type of glomerulonephritis. There have been suggestions that renal biopsy is not essential and that all nephrotic adults should be treated with steroids. However, this approach means unnecessary treatment of a large proportion of patients with a toxic drug, also that no assessment would be available of the type of glomerulonephritis or an estimate of the likelihood of a response to treatment and of the prognosis for long-term renal function. In skilled hands the dangers of renal biopsy are small and outweighed by those of steroid treatment.

General aspects of the management of the nephrotic syndrome

Although steroids and immunosuppressants have been widely used in the treatment of the nephrotic syndrome, general measures remain an important part of the treatment of these disorders. Initial treatment of oedema is with salt restriction and if there is hyponatraemia with fluid restriction. Adults are commonly treated with loop diuretics such as furosemide (frusemide), but this must be used with care, particularly in children, because of the risk of volume depletion and consequent renal impairment. Patients with a nephrotic syndrome are at an increased risk of developing thromboemboli but prophylactic anticoagulation is not normally recommended. There is now good experimental and clinical evidence that angiotensin-converting enzyme (**ACE**) inhibitors reduce proteinuria and slow the progression of renal impairment in patients with glomerulonephritis. These agents are recommended in those who are likely to have a prolonged nephrotic syndrome. Such patients are also likely to benefit from lipid-lowering therapy, although this has not been tested by randomized controlled study.

Minimal-change nephropathy

Aetiology

There is a well-recognized association between Hodgkin's lymphoma and minimal-change nephropathy. Rarely, minimal-change nephropathy has been reported in patients with a carcinoma. There are also case reports of minimal-change nephropathy in individuals, often atopic, exposed to bee stings, poison oak, grass pollen, and cow's milk. Non-steroidal anti-inflammatory drugs can cause an interstitial nephritis, which in some cases is accompanied by a nephrotic syndrome with renal histology showing the changes of minimal-change nephropathy.

Pathogenesis

The responsiveness of the nephrotic syndrome of minimal-change nephropathy to steroids, cyclophosphamide, chlorambucil, and ciclosporin A is strong evidence that this disorder is immune-mediated. The pathogenetic mechanisms remain obscure. The low serum IgG and high IgM levels in these patients appear to be a consequence of the nephrotic syndrome and shed no light on pathogenesis. The hypothesis that proteinuria is caused by a lymphokine produced by an abnormal clone of T lymphocytes has been extensively studied: as yet it has neither been proved nor disproved. In Europe an increased incidence of HLA-DR7 is found in patients with minimal-change nephropathy and in Japan the association is with HLA-DR8, thus suggesting a genetic predisposition.

Pathology

The histological features are similar in both children and adults. On light microscopy the glomeruli appear normal or small (Fig. 1 and Plate 1) and on electron microscopy there is effacement of epithelial-cell foot processes over the outer surface of the glomerular basement membrane. Some authors accept a minor degree of mesangial IgM deposition and mesangial proliferation as being consistent with this disorder.

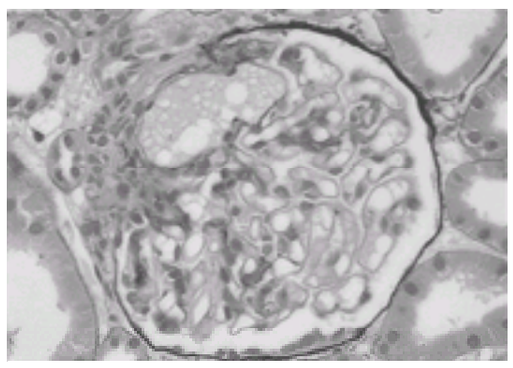

Fig. 1 Minimal-change nephropathy. The glomerulus looks normal on light microscopy. Periodic acid–methenamine silver staining (64 ×). (By courtesy of Dr A. J. Howie.) (See also Plate 1.)

Minimal-change nephropathy in children

Minimal-change nephropathy is found in approximately 76 per cent of children with an idiopathic nephrotic syndrome. Most affected children are under 6 years of age (80 per cent), with a peak age of onset of 2 to 4 years. The condition is responsible for 59 per cent of those aged between 6 and 15 years and about 20 per cent of adults with the nephrotic syndrome. It is more common in boys than in girls, with a male to female childhood ratio of 2:1.

Clinical presentation

The clinical presentation is with a nephrotic syndrome that is characterized by severe hypoalbuminaemia, with a serum albumin level of less than 10 g/l in some 38 per cent of cases. Microscopic haematuria is infrequent (22 per cent), as is hypertension (9 per cent). Renal impairment is infrequent at diagnosis, being found in about 10 per cent of cases, and presentation in acute renal failure is rare. These children are prone to infections, in particular cellulitis and pneumococcal peritonitis.

Diagnosis

The role (or not) of renal biopsy has already been discussed. In 75 per cent of children with minimal-change nephropathy the proteinuria is highly or moderately selective. The concept of protein clearance selectivity was based on the hypothesis that the glomerular basement membrane provided a size-selective barrier to the loss of protein molecules. Although it is now clear that glomerular permselectivity is also based on the charge of molecules, the selectivity test remains useful. A simple version involves dividing the ratio of the urine and plasma concentrations of a large protein (for example, IgG or a $_2$-macroglobulin) by the ratio of the urine and plasma concentrations of a small molecule (for example, albumin or transferrin).

Treatment of minimal-change nephropathy in children

The first-line treatment of minimal-change nephropathy is prednisolone at an initial dose of 60 mg/m ^{2} (maximum dose 80 mg) daily for 4 to 6 weeks, reducing the prednisolone to 40 mg/m^2 (maximum dose 60 mg) on alternate days for a further 4 to 6 weeks. With this treatment 93 per cent of children respond with complete loss of proteinuria within 8 weeks. This duration of treatment is more effective in maintaining a remission than a shorter course of steroids. However, once remission is induced, 66 per cent of children have at least one relapse. The major problem is that between 40 and 55 per cent of children who initially respond to steroids develop multiple relapses when steroids are discontinued, or they become steroid-dependent and relapse when the steroid dosage is reduced. Early, frequent relapses (three or more) in the 6 months following an initial response to steroids predict a frequently relapsing course.

Treatment of relapses

Prednisolone 60 mg/m^2 should be given until the urine is free of protein for 3 days (maximum 4 weeks) then prednisolone 40 mg/m ^{2} on alternate days for 4 weeks. Children's growth should be carefully monitored, using growth curves, during repeated steroid treatment of relapses.

Cyclophosphamide/chlorambucil

Treatment with an immunosuppressant drug should be considered in the following groups of patients: children who are frequent relapsers (two relapses within 6 months of the initial response or four relapses within any 1 year); children who are steroid-dependent (two consecutive relapses occurring during alternate-day treatment for an earlier relapse) or who relapse within 14 days of treatment of an earlier relapse (fast relapse); children in whom two out of four relapses within 6 months were fast relapses; children who are steroid-toxic. There is good evidence that short-term treatment with cyclophosphamide can induce a sustained or even permanent remission in such children. Cyclophosphamide is given in a dose of 2 mg/kg per day (ideally, height for weight) for 8 weeks. Approximately 50 per cent of treated children are in remission at 2 years and 40 per cent at 5 years. One study suggested that the duration of remission was longer with a 12-week course compared with an 8-week course of cyclophosphamide, but this was not subsequently confirmed. Chlorambucil has also been used to treat these patients, but there is no evidence that it is better than cyclophosphamide and it is probably more toxic. Cyclophosphamide has been carefully evaluated in these children and is the drug of choice. Children with the HLA allele HLA-DR7 are less likely to respond to cyclophosphamide.

Toxicity of cyclophosphamide and chlorambucil

The risk of gonadal toxicity is greater in boys than in girls. Gonadal toxicity occurs with chlorambucil at a cumulative dose of 8 to 10 mg/kg. The borderline dose for permanent gonadal toxicity with cyclophosphamide is a cumulative dose of 200 mg/kg. Bone marrow toxicity with both drugs means that the leucocyte count should be regularly measured during treatment. These drugs also increase the long-term risk of developing cancer. Other toxic side-effects of cyclophosphamide include leucopenia, haemorrhagic cystitis, and alopecia. At the doses and duration of treatment outlined above it is relatively safe.

Levamisole

This has been used in children with frequently relapsing or steroid-dependent minimal-change nephropathy and appears to have a steroid-sparing effect. However, the benefits appear marginal and most patients relapse after stopping treatment. Levamisole can cause a reversible neutropenia.

Long-term outcome

The risk of a future relapse is low for those children in whom the nephrotic syndrome goes into remission within 8 weeks of steroid therapy and who do not relapse for 6 months. Early relapse within 6 months is reported to be associated with a risk of relapses for up to 3 years. Some 5.5 per cent of affected children continue to relapse into adult life. All of these children presented with a nephrotic syndrome before the age of 6 years. Children who had persistent proteinuria at 8 weeks had a 21 per cent risk of developing endstage renal failure, and this increased to 35 per cent if they still had proteinuria at 6 months. The long-term mortality rate in children ranges from 2.6 to 7.2 per cent.

Minimal-change nephropathy as part of a spectrum of glomerular disease

In temperate countries, the majority of children with a nephrotic syndrome have minimal-change nephropathy, focal segmental glomerulosclerosis (**FSGS**), or a mesangial proliferative glomerulonephritis (Table 1). Since there is considerable overlap between these conditions, it has been argued that they are all variants of the same disease, termed the 'idiopathic nephrotic syndrome'. In favour of this view is the observation that the histological lesion may evolve with time in a proportion of patients with steroid-responsive nephrotic syndrome. Repeat renal biopsies performed if the character of illness changes—for example, if patients become frequent relapsers, steroid-dependent, or steroid-resistant—sometimes show progression from minimal-change nephropathy or mesangial proliferative glomerulonephritis to focal segmental glomerulonephritis. One study showed that patients with presumed minimal-change nephropathy whose renal biopsies showed large glomeruli were more likely to develop FSGS. In general, those patients with minimal-change nephropathy who develop FSGS but remain steroid-responsive have a good prognosis for renal function, whilst those who are steroid-resistant develop progressive renal failure. The prognosis for renal function is therefore determined by the responsiveness to steroids and not by the histological lesion.

Minimal-change nephropathy in adults

About 20 per cent of adults with a nephrotic syndrome have minimal-change nephropathy. The mean age of onset is 40 years but the condition can occur at any age. The histology is identical to that found in children, with the exception of a higher incidence of globally sclerosed glomeruli that are a feature of ageing.

Clinical presentation

As in children the clinical presentation is with a nephrotic syndrome, although this is not generally as severe. Profound hypoalbuminaemia (serum albumin level under 10 g/l) is rare in adults, being found in only 6 per cent of cases. The disease is slightly more common in men than in women, with a male to female ratio of 1.3:1. More adults than children are hypertensive (30 per cent), have microscopic haematuria (28 per cent), and have renal impairment at diagnosis (60 per cent). These abnormalities are more severe in patients aged over 60 years who are also at particular risk of developing acute renal failure.

Diagnosis

Only 50 per cent of adults with minimal-change nephropathy have highly selective proteinuria, which together with the high incidence of microscopic haematuria and renal impairment makes it impossible to differentiate minimal-change nephropathy from other forms of glomeru-lonephritis on clinical grounds. A renal biopsy is essential to make the diagnosis in adults with a nephrotic syndrome,.

Treatment of minimal-change nephropathy in adults

Treatment is with prednisolone at an initial dose of 60 mg/day: response occurs slightly less often than in children and also more slowly. Some 80 per cent of adults with minimal-change nephropathy do respond, but remission can take up to 16 weeks to occur. Relapse is less frequent in adults (1.7/patient) than in children, and only 21 per cent of adults develop multiple relapses or are steroid-dependent. In adults as in children, cyclophosphamide is effective in inducing a long-lasting remission. In one study 62.5 per cent of patients treated with cyclophosphamide were in remission at 10 years.

Ciclosporin A

There is now good evidence that ciclosporin A is effective in the treatment of minimal-change nephropathy in both adults and children. Patients who are steroid-responsive or multiple relapsers are more likely to respond with complete or partial remissions (70 to 80 per cent) than patients who are resistant to steroids (40 to 50 per cent). The drug should be considered in those patients who develop steroid toxicity because they have multiple relapses or who are steroid-dependent. Ciclosporin A appears to be effective at blood levels of between 100 and 200 ng/ml, and at these levels significant short-term nephrotoxicity and hypertension are uncommon. However, relapses appear to recur with the same frequency after ciclosporin A has been discontinued as before, and for that reason it is still advisable to use cyclophosphamide as the first-choice treatment in patients with a multiple relapsing or steroid-dependent minimal-change nephropathy in the hope of inducing a sustained remission. In this author's view, ciclosporin A can best be viewed as a steroid-sparing agent in patients with minimal-change nephropathy.

Long-term outcome

Some 6 per cent of adult patients are still nephrotic after a mean follow-up of 7.5 years. The survival in patients over 60 years of age has been reported to be 50 per cent at 10 years, and in those aged 15 to 59 it was 90 per cent.

Focal segmental glomerulosclerosis

Focal segmental glomerulosclerosis (FSGS) was first described by Rich in 1957 at autopsy in children who died from a nephrotic syndrome. Fewer terms have generated more disagreement amongst pathologists and nephrologists: it is not a disease entity but a histological lesion that is often of unknown aetiology.

Secondary FSGS

FSGS may be a sequel of glomerular scarring in patients with previous proliferative glomerulonephritis and is seen in biopsies from patients with Alport's syndrome. It is also seen in patients with reflux nephropathy and other conditions leading to a reduced renal mass, and it is likely that the segmental sclerosing lesions in these circumstances are a consequence of glomerular hypertension and hyperfiltration (Table 2). FSGS has also been found late on during the clinical course of patients with a nephrotic syndrome who had an initial renal biopsy showing minimal-change nephropathy.

Pathogenesis

In experimental models, focal segmental sclerosis can develop from different pathogenic mechanisms. These include toxic injury (puromycin nephropathy), immunological injury (anti-GBM nephritis), the lupus-associated nephritis in NZB/NZW F1 mice, and hyperfiltration injury (5/6ths nephrectomy). Some of these models have clinical counterparts, and the diversity of pathogenic mechanisms may explain the variability in the clinical presentation and response to FSGS therapy. As with minimal-change nephropathy there are suggestions that the glomerular injury in FSGS is caused by a lymphokine. The rapid development of heavy proteinuria following renal transplantation in some patients with FSGS indicates that the glomerular injury is caused by a circulating factor.

Primary FSGS

Focal segmental glomerulosclerosis may be apparently idiopathic and found early on during the clinical course of patients with proteinuria or nephrotic syndrome. About 7 per cent of children and 20 per cent of adults with a nephrotic syndrome have FSGS. Even when FSGS is found early on in the course of a nephrotic syndrome there is no evidence to suggest that it represents a homogenous disease.

Pathology

The histological lesions of FSGS comprise segmental areas of glomerular sclerosis with hyalinization of glomerular capillaries, the segmental areas usually being adherent to Bowman's capsule. In childhood FSGS, these lesions predominantly affect juxtamedullary glomeruli. Typically, the areas of segmental sclerosis are randomly distributed within the glomerular tuft with a predilection for the hilar regions, and these patients may be regarded as having classical FSGS (Fig. 2 and Fig. 3 and Plate 2 and Plate 3). In some biopsies the glomerular lesions are located peripherally at the glomerulotubular junction, the so-called glomerular tip lesion. Focal areas of tubular atrophy and interstitial nephritis are prominent. On immunofluorescent microscopy, deposits of IgM and C3 may be seen in the sclerotic areas. Electron microscopy shows diffuse foot-process effacement in apparently unaffected glomeruli.

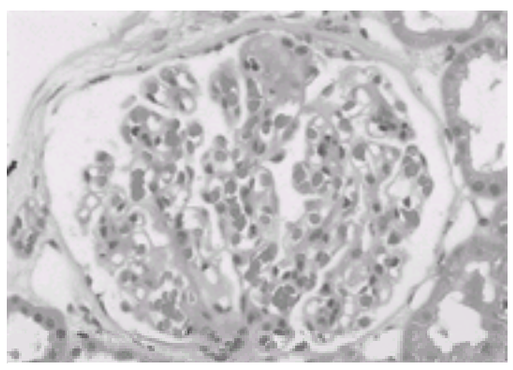

Fig. 2 Classical segmental sclerosing glomerulonephritis at an early stage. The glomerulus shows an erratic increase in mesangium with a segmental area of foamy cells and sclerosis opposite the vascular pole, next to the tubular origin. Haematoxylin and eosin staining (50 ×). (By courtesy of Dr A. J. Howie.) (See also Plate 2.)

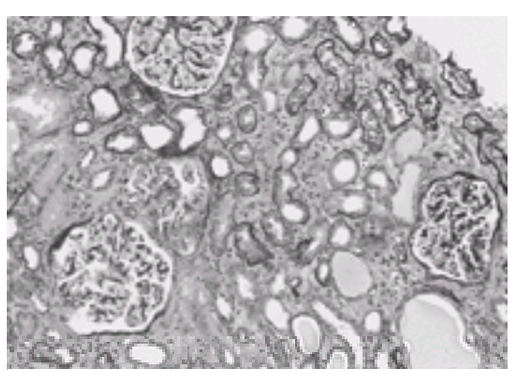

Fig. 3 Classical segmental sclerosing glomerulonephritis at a late stage. Four glomeruli show an erratic increase in mesangium and segmental lesions at various sites. Periodic acid–methenamine silver staining (× 64). (By courtesy of Dr A. J. Howie.) (See also Plate 3.)

Pathogenesis

In approximately 30 per cent of patients with primary FSGS there is a circulating factor that causes an increase in glomerular permeability *in vitro*. This factor appears to be a protein with a molecular weight of between 30 and 50 kDa and is not an immunoglobulin. This factor has also been found in the serum of patients who develop recurrent FSGS postrenal transplantation.

Clinical presentation

Children

Approximately 7 per cent of children presenting with an idiopathic nephrotic syndrome have FSGS. Males and females are equally affected and the peak age at onset is between 6 and 8 years. The majority of patients (75 per cent) present with a nephrotic syndrome, 20 per cent have persistent proteinuria, and 5 per cent haematuria as well as proteinuria. Clinically these patients differ from other children with minimal-change nephropathy, in that two-thirds have microscopic haematuria, half have impaired renal function at diagnosis, and one-third are hypertensive. The proteinuria is usually poorly selective.

Adults

The clinical presentation in adults does not differ in any significant respects from that in children. The mean age at onset is between 20 and 30 years but FSGS has been found in patients aged 70.

Treatment of primary ('classical') FSGS

The prognosis in patients with primary FSGS and proteinuria in the non-nephrotic range is good, and 80 per cent of such patients survive for 10 years without developing endstage renal failure. These patients therefore do not need treatment with either prednisolone or immunosuppressants and should be treated with general measures only.

The main problem is the treatment of patients with FSGS and a nephrotic syndrome. In most studies patients have been given steroids, and approximately 30 per cent of those treated for 8 weeks with prednisolone go into remission. There is now good data from uncontrolled studies that a more prolonged course of steroids, of up to 6 months, is associated with a higher rate of remission. Children are treated with prednisolone at an initial dose of 60 mg/m ^{2} per day and adults with a dose of 60 mg/day. Patients who go into remission have a good prognosis with fewer than 10 per cent developing endstage renal failure. However, the prognosis in patients who do not respond to steroids is poor, with between 30 and 50 per cent developing endstage renal failure over 5 to 10 years. There is no difference in prognosis between adults and children. Adverse prognostic factors include tubulointerstitial fibrosis and a high serum creatinine level.

Other immunosuppressants

The evidence supporting the addition to prednisolone of cyclophosphamide or chlorambucil in the treatment of FSGS is not convincing. Useful remissions have been reported, but in the absence of controlled trials the value of this is difficult to assess.

Several uncontrolled studies have looked at the effects of ciclosporin A: in general, the responsiveness has been poor and paralleled that of steroids. Recent controlled studies in adults suggest that ciclosporin A when added to prednisolone is more effective in inducing remission of the nephrotic syndrome than steroids alone in patients with steroid-resistant FSGS.

Glomerular tip lesion

Some studies suggest that the site of the segmental sclerosing lesions predicted steroid responsiveness. Adult patients with a peripheral segmental sclerosing lesion at the tubular origin, the glomerular tip lesion, have a steroid- or immunosuppressant-responsive nephrotic syndrome and do not progress to endstage renal failure.

Similar observations have been reported in children, although in both children and adults these observations have not been confirmed.

Recurrence after renal transplantation

The nephrotic syndrome recurs in 20 to 40 per cent of patients with primary FSGS, often within days of renal transplantation, and this leads to graft failure in some 50 per cent of cases. After recurrence in a first transplant the rate of recurrence in a subsequent transplant approaches 75 per cent. Plasma exchange and protein immunoadsorption have resulted in a reduction of proteinuria or a remission of the nephrotic syndrome in some patients. These data are not controlled and our experience is that the reduction in proteinuria is transient.

Collapsing glomerulopathy

This is a type of focal segmental sclerosing glomerulonephritis, characterized by segmental or global collapse of glomerular capillaries with basement-membrane wrinkling and crowding of glomerular epithelial cells. These appearances represent a distinct subset of patients with focal segmental glomerulosclerosis, and were initially described in patients with HIV-associated nephropathy in the context of a severe nephrotic syndrome and rapid progression to endstage renal failure. Subsequent reports show that it may also be idiopathic.

Presentation is with a nephrotic syndrome and renal impairment (70 per cent of cases). Treatment with steroids or cytotoxic drugs has been ineffective in inducing remission of the nephrotic syndrome or preventing the development of endstage renal failure. There is a rapid deterioration of renal function and over 70 per cent of patients are in endstage renal failure after a follow up of 5 years.

Membranous nephropathy

Membranous nephropathy accounts for between 20 and 30 per cent of cases of the nephrotic syndrome in adults and about 2 to 5 per cent of those in childhood. Histologically it is defined by the presence of subepithelial immune deposits on the outer surface of the glomerular basement membrane. No cause for this histological lesion is found in most patients living in temperate countries, and it is therefore termed idiopathic membranous nephropathy, but it is unlikely that membranous glomerulonephritis is a homogenous disorder. Its aetiology (where identifiable), genetic basis, frequency as a cause of the nephrotic syndrome, and clinical evolution with or without treatment differ substantially between studies from different countries.

Aetiology

In about 20 to 25 per cent of adults and 35 per cent of children with membranous nephropathy there is an identifiable associated condition (Table 3). The frequency of this varies in different parts of the world. Malignancy, usually a carcinoma and rarely Hodgkin's lymphoma and non-Hodgkin's lymphoma, is found in between 3 and 7 per cent of all cases, rising to 16 per cent in those aged over 60 years. The most common tumours are carcinoma of the bronchus, colon, kidney, breast, stomach, and prostate. Gold and penicillamine are prominent causes of membranous nephropathy and this complication is more common in individuals who carry the *HLA-DR3* gene. There is also some evidence that membranous nephropathy can develop in patients with rheumatoid arthritis who are not taking these drugs. Approximately 3 per cent of all patients with membranous nephropathy have systemic lupus erythematosus; a further 2 per cent of patients have serological features of this disorder or histological changes that are suggestive of it, sometimes predating clinical evidence of the disease by many years. In Northern Europe about 1 per cent of patients with membranous nephropathy have positive hepatitis B serology, but this association is much more common in South-east Asia and in Africa, particularly in children.

Pathogenesis

The immune mechanisms that lead to the development of membranous nephropathy are unknown. In rats, the administration of antibodies against renal tubular epithelial antigen leads to a membranous nephropathy that histologically resembles the human condition. The antibody responsible for this Heymann's nephritis in rats binds to an antigen called gp330, which is found on renal tubular brush border and on glomerular epithelial cells. In glomeruli this leads to the development of subepithelial deposits through the *in situ* formation of immune complexes. Although human renal tubular cells express gp330, this is not found in glomerular epithelial cells, and there is no evidence that a similar mechanism plays a role in human membranous nephropathy. In Europe there is a strong association between membranous nephropathy and the MHC haplotype HLA-A1 B8 DRw3, whilst in Japan the association is with HLA-DR2. By contrast, no such association is seen in the United States.

Pathology

Idiopathic membranous nephropathy is characterized histologically by diffuse thickening of the glomerular basement on light microscopy, usually with argyrophyllic subepithelial spikes (Fig. 4 and Plate 4). On immunofluorescent or immunoperoxidase microscopy this thickening is shown to be due to the presence of immune deposits, usually consisting of IgG and C3, on the subepithelial surface of the glomerular basement membrane (Fig. 5 and Plate 5). The size and extent of incorporation of immune deposits into the glomerular basement membrane on electron microscopy forms the basis of histological classification—stage 1: subepithelial deposits without spikes; stage 2: large subendothelial deposits separated by spikes of basement membrane; stage 3: deposits incorporated into a thickened basement membrane with many spikes; stage 4: a very thick irregular basement membrane with no spikes and resorbed deposits. The presence of mesangial proliferation, mesangial immune deposits, and IgA and C1q on immunofluorescent microscopy raises the possibility that membranous nephropathy is secondary to systemic lupus erythematosus.

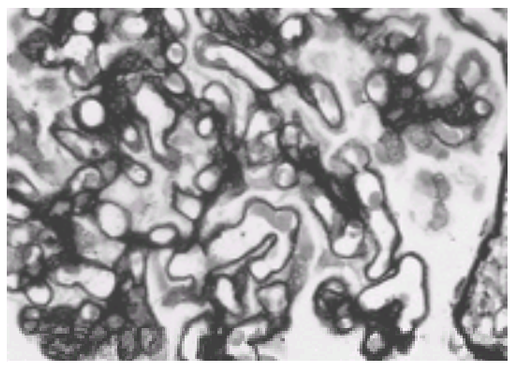

Fig. 4 Membranous nephropathy. There are regular short spikes on the outside of glomerular capillary loops. Periodic acid–methenamine silver staining (80 ×). (See also Plate 4.)

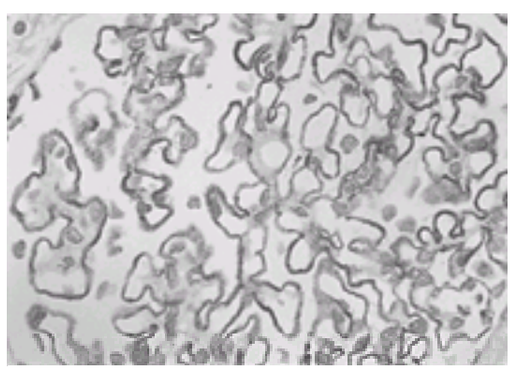

Fig. 5 Membranous nephropathy. Immunoperoxidase staining shows uniform granular deposits of IgG on the epithelial side of glomerular basement membranes (80

×). (By courtesy of Dr A. J. Howie.) (See also Plate 5.)

Clinical presentation

In children, boys are affected three times as often as girls. In adults, most studies report a preponderance of men, with a male to female ratio of 2–3:1. The ajority of patients are aged between 30 and 50, although the condition has been described in patients aged up to 80 years. The clinical presentation is with the nephrotic syndrome in about 75 per cent of cases, with the remainder having proteinuria only. Microscopic haematuria is found in 50 per cent of adults and 90 per cent of children. Macroscopic haematuria is found in about 10 to 20 per cent of children but it is rare in adults. About 25 to 40 per cent of adults and 6 per cent of children are hypertensive at diagnosis, and between 10 per cent and 30 per cent of patients have a raised serum creatinine level.

Renal vein thrombosis

Patients with membranous nephropathy appear to be at particular risk of developing renal vein thrombosis, although this is not as high as originally suggested. Most such patients are asymptomatic, but they may present with pulmonary emboli. Detection is by Doppler ultrasound of the renal veins, computed tomography (**CT**), magnetic resonance (**MRI**) imaging, or using the venous phase of renal arteriography. In practice, a renal vein thrombosis should be looked for if there is a sudden deterioration of renal function in a patient with membranous nephropathy. It is now known that renal vein thrombosis is a consequence of the hypercoagulable state of the nephrotic syndrome and is not a cause of membranous nephropathy.

Membranous nephropathy with a crescentic glomerulonephritis

About 5 per cent of patients with a membranous nephropathy develop a crescentic glomerulonephritis with rapid deterioration of renal function. Most such patients have antibodies to glomerular basement antigen or to neutrophil cytoplasmic antigens. Treatment has been with prednisolone and cyclophosphamide as for other patients with a crescentic glomerulonephritis.

Clinical evolution of untreated membranous nephropathy

In the long-term, untreated membranous nephropathy evolves either to remission or to the development of chronic renal failure. The rate at which either outcome occurs varies in different studies. After a mean follow-up of 4.5 to 6 years, between 9.5 and 22 per cent of patients are in endstage renal failure, 9.5 to 19 per cent have significantly impaired renal function, and 23 to 50 per cent are in remission. The actuarial survival rate shows that about 75 per cent of patients are alive at 10 years and 60 per cent have functioning kidneys. Examination of the control untreated patients in recent treatment trials shows that, of 205 patients followed for between 2 and 5 years, 15 per cent were in complete remission and 9 per cent in endstage renal failure. Any study of treatment in membranous nephropathy must therefore address the difficulty of treating large numbers of patients with toxic drugs who have little risk of developing endstage renal failure.

Treatment

The twin aims of treating membranous nephropathy are first to induce a remission of the nephrotic syndrome and second to prevent the development of endstage renal failure. Despite several careful studies using steroids and immunosuppressants, there is still no agreement that these aims can be achieved.

Steroid treatment

In the 1979 Collaborative study conducted in the United States, 72 adults with membranous nephropathy were randomized to 8-weeks' treatment with either 125 mg prednisolone on alternate days or placebo. The steroid dose was then tapered and stopped over several weeks. Deterioration of renal function, as measured by the glomerular filtration rate (**GFR**), was significantly more rapid in untreated than in treated patients. Further, a significantly lower proportion of treated patients than untreated patients developed renal failure (serum creatinine level over 440 µmol/l).

In the United Kingdom Medical Research Council (**MRC**) study (1990), 107 adult patients with membranous nephropathy were randomized to treatment with either prednisolone 125 mg on alternate days for 8 weeks or placebo. At 36 months there were no significant differences in the plasma creatinine level, creatinine clearance, and 24-h urine protein between treated and untreated patients. In the Canadian study, 158 patients were treated with either prednisolone 45 mg/m ^{2} body surface area per day for 6 months or no specific treatment. No benefits were seen in renal function or proteinuria after a mean follow-up of 48 months. These data indicate that short-term steroids are of no benefit in the treatment of membranous nephropathy.

Steroid and chlorambucil treatment

One study reported that chlorambucil was more effective than azathioprine or placebo in the treatment of membranous nephropathy. This provided the rationale for the Italian multicentre study in which patients were randomized to symptomatic treatment only or treatment with the following alternating regime—month 1: intravenous methylprednisolone, 1 g on each of 3 consecutive days, followed by oral methylprednisolone (0.4 mg/kg per day) or prednisolone (0.5 mg/kg per day) for 27 days; month 2: oral chlorambucil (0.2 mg/kg per day) alone for 1 month, the dose was lowered if the leucocyte count fell below 5×10^{9}/l. Alternating monthly cycles of methylprednisolone and chlorambucil were given for a total of 6 months. After a mean follow-up of 31 to 37 months, significantly more treated than untreated patients were in remission (either total or partial): 23/32 (72 per cent) versus 9/30 (30 per cent). Furthermore, 8 of 30 controls showed a 50 per cent rise in serum creatinine in contrast to none of the treated patients. The side-effects of treatment were minor and consisted of epigastric pain and leucopenia. To answer the question of whether the beneficial effect of this regime was due solely to the steroid component, a further study compared the effect of methylprednisolone alone with methylprednisolone and chlorambucil. Patients treated with the combination were more likely to have an early remission of the nephrotic syndrome, but this benefit was lost after 4 years. There was no difference in the rate of decline of renal function between the two therapies.

Other immunosuppressive regimes

In a recent, randomized controlled study, cyclophosphamide 2.5 mg/kg per day was compared with chlorambucil 0.2 mg/kg per day in the regime described above. The results in terms of remission of the nephrotic syndrome and deterioration of renal function were comparable.

A small, randomized controlled study of 17 patients with a persistent nephrotic syndrome and declining renal function suggested that ciclosporin A slowed the rate of decline of renal function: this requires confirmation in a larger trial.

Meta-analysis of steroid and immunosuppressant treatment of membranous nephropathy

A meta-analysis of four randomized controlled studies comparing treatments of membranous nephropathy showed that regimes comprising chlorambucil or cyclophosphamide, either alone or with steroids, were more effective than symptomatic treatment or treatment with steroids alone in inducing remission of the nephrotic syndrome. These agents increased the relative risk of achieving a complete remission by 4.6 (95 per cent confidence interval of 2.2 to 9.3). The number of patients studied and the design of the clinical trials were such that no conclusion could be drawn on the effects of these treatments on renal function.

Management of the patient with membranous nephropathy and deteriorating renal function

Drug-induced interstitial nephritis, renal vein thrombosis, and crescentic glomerulonephritis should be excluded. In patients with deteriorating renal function due to the progression of membranous nephropathy, several uncontrolled studies have suggested that treatment with intravenous methylprednisolone, or with oral prednisolone and chlorambucil or cyclophosphamide, may reverse the rate of decline in renal function. These studies are difficult to interpret as renal function may stabilize or improve without treatment in some cases.

Prognostic factors

Identifying those patients who at the onset of membranous nephropathy were likely to have a poor outcome for renal function would be helpful in deciding who to treat. Most studies show that adverse risk factors for the development of renal failure include male sex, a nephrotic syndrome, persistent heavy proteinuria, tubulointerstitial fibrosis, renal impairment at diagnosis, and deterioration of renal function in the first 2.5 years after diagnosis. In particular, patients with proteinuria of over 6 g/day for longer than 9 months were found to have a 55 per cent likelihood of progressing to renal failure. Children appear to do better than adults; in one study, 42 per cent of children went into complete remission and only 10 per cent developed endstage renal failure after a mean follow-up of 4 years.

How should membranous nephropathy be treated?

There is still no agreement on how membranous nephropathy should be treated, as up to 40 per cent of patients with this disorder enter spontaneous remission with long-term preservation of renal function. The dilemma is that early treatment of all patients exposes those who were going into remission anyway to the toxicity of drugs, whilst delayed treatment of high-risk patients may be ineffective. The results of current randomized controlled trials into the benefits of treatment with alkylating agents or cyclophosphamide in patients at high risk of progressive renal failure will guide treatment decisions.

Further reading

Minimal-change nephropathy

Arbeitsgemeinschaft fur Padiatrisch Nephrologie (1988). Short versus standard prednisolone for initial treatment of idiopathic nephrotic syndrome in children. *Lancet* **1**, 380–3.

Bargman J (1999). Management of minimal lesion glomerulonephritis: evidence-based recommendations. *Kidney International* **55** (Suppl. 70), 3–16.

British Association for Paediatric Nephrology (1991). Levamisole for corticosteroid dependent nephrotic syndrome in childhood. *Lancet* **337**, 1555–7.

International Study of Kidney Disease in Children (1978). Prediction of histopathology from clinical and laboratory characteristics at time of diagnosis. *Kidney International* **13**, 159–65.

International Study of Kidney Disease in Children (1981). The primary nephrotic syndrome in Children. Identification of patients with minimal change nephrotic syndrome from initial response to prednisolone. *Journal of Pediatrics* **98**, 561–4.

Nolasco F, *et al*. (1986). Adult-onset minimal change nephrotic syndrome: a long term follow-up. *Kidney International* **29**, 1215–23.

Tarshish P *et al*. (1997). Prognostic significance of the early course of minimal changes nephrotic syndrome: report of the International Study of Kidney Disease in Children. *Journal of the American Society of Nephrology* **8**, 769–76.

Ueda N, Kuno K, Ito S (1990). Eight and 12 week courses of cyclophosphamide in nephrotic syndrome. *Archives of Disease in Childhood* **85**, 1147–50.

Focal segmental glomerulosclerosis

Burgess E (1999). Management of focal glomerulosclerosis: evidence based recommendations. *Kidney International* **55** (Suppl. 70), 26–32.

Cattran D, *et al*. (1999). A randomized study of cyclosporine in patients with steroid-resistant focal segmental glomerulosclerosis. *Kidney International* **56**, 2220–6.

D'Agati V (1994). The many masks of focal segmental glomerulosclerosis. *Kidney International* **46**, 1223–41.

Detweiler R, *et al*. (1994). Collapsing glomerulopathy: a clinically and pathologically distinct variant of segmental glomerulosclerosis. *Kidney International* **45**, 1734–46.

Howie A, *et al*. (1993). Different clinicopathological types of segmental sclerosing glomerular lesions in adults. *Nephrology, Dialysis, and Transplantation* **8**, 590–9.

Korbet S, Schwartz M, Lewis E (1994). Primary focal segmental glomerulosclerosis: clinical course and response to therapy. *American Journal of Kidney Disease* **23**, 773–83.

Niaudet P for The French Society of Pediatric Nephrology (1992). Comparison of cyclosporine and chlorambucil in the treatment of idiopathic nephrotic syndrome: a multicenter randomized controlled trial. *Pediatric Nephrology* **6**, 1–3.

Niaudet P for The French Society of Pediatric Nephrology (1994). Treatment of childhood steroid resistant idiopathic nephrosis with a combination of cyclosporine and prednisolone. *Journal of Pediatrics* **125**, 981–6.

Rich A (1957). A hitherto undescribed vulnerability of the juxta-medullary glomeruli in lipoid nephrosis. *Bulletin of John Hopkins Hospital* **100**, 173–86.

Savin V, *et al*. (1996). Circulating factor associated with increased glomerular permeability to albumin in recurrent focal segmental glomerulosclerosis. *New England Journal of Medicine* **334**, 878–83.

Membranous nephropathy

Cameron J, Healy M, Adu D (1990). The Medical Research Council Trial of short-term high-dose alternate day prednisolone in idiopathic membranous nephrotic syndrome in adults. *Quarterly Journal of Medicine* **274**, 133–56.

Cattran D, *et al*. (1989). A randomized controlled trial of prednisolone in patients with idiopathic membranous nephropathy. *New England Journal of Medicine* **320**, 210–15.

Cattran D, *et al*. (1995). A controlled trial of cyclophosphamide in patients with progressive membranous nephropathy: Canadian Glomerulonephritis Study Group. *Kidney International* **47**, 1130–5.

Collaborative Study of the Adult Idiopathic Nephrotic Syndrome (1979). A controlled study of short-term prednisolone treatment in adults with membranous nephropathy. *New England Journal of Medicine* **301**, 1301–6.

Honkanen E, Tornroth T, Gronhagen-Riska C (1992). Natural history, clinical course and morphological evolution of membranous nephropathy. *Nephrology, Dialysis, and Transplantation* **7** (Suppl. 1), 35–41.

Imperiale T, Goldfarb S, Berns J (1995). Are cytotoxic agents beneficial in idiopathic membranous nephropathy? A meta-analysis of the controlled trials. *Journal of the American Society of Nephrology* **5**, 1553–8.

Muirhead N (1999). Management of idiopathic membranous nephropathy: evidence-based recommendations. *Kidney International* **55** (Suppl. 70), S47–55.

Pei Y, Cattran D, Greenwood C (1992). Predicting chronic renal insufficiency in idiopathic membranous nephropathy. *Kidney International* **42**, 960–6.

Ponticelli C, *et al*. (1995). A 10-year follow-up of a randomized study with methylprednisolone and chlorambucil in membranous nephropathy. *Kidney International* **48**, 1600–4.

Ponticelli P, *et al*. (1998). A randomized study comparing methylprednisolone plus chlorambucil versus methylprednisolone plus cyclophosphamide in idiopathic membranous nephropathy. *Journal of the American Society of Nephrology* **9**, 444–50.

Schiepatti A, *et al*. (1993). Prognosis of untreated patients with idiopathic membranous nephropathy. *New England Journal of Medicine* **329**, 85–9.

Books and monographs

Cameron J, Glassock R, eds (1988). *The nephrotic syndrome*. Marcel Dekker, New York.

Cattran DC, ed. (1999). Management of glomerulonephritis. *Kidney International* **55** (Suppl. 70), S1–62.

Kincaid-Smith P, d'Apice A, Atkins R, eds (1978). *Progress in glomerulonephritis*. Wiley, New York.

Pusey CD, ed. (1999). The Treatment of Glomerulonephritis. In: *Developments in nephrology*, Vol. 40. Kluwer Academic, Dordrecht.

20.7.5 Proliferative glomerulonephritis

Peter W. Mathieson

The term proliferative glomerulonephritis covers a variety of conditions (Table 1) where there is increased cellularity of the glomerulus, either due to the proliferation of resident glomerular cells, or infiltration of leucocytes, or both. The proliferative changes may be focal (that is to say, they only affect some glomeruli) and/or segmental (in other words, only affecting parts of each glomerulus). Many of these entities will be considered in other chapters, and only those not covered elsewhere (*****) will be described here.

Mesangial proliferative glomerulonephritis

Patients will typically have haematuria and this may be associated with proteinuria and/or impairment of excretory renal function and/or hypertension. The majority of patients whose renal biopsies show only mesangial proliferation will have IgA nephropathy (see Chapter 20.7.2), but a few will have no IgA deposits and their classification is not straightforward: possibilities include IgM nephropathy and so-called 'idiopathic' mesangial proliferative glomerulonephritis (**GN**).

IgM nephropathy

There is continuing controversy about this diagnostic entity. In patients with nephrotic syndrome, if the only abnormalities on the renal biopsy are in the mesangial region, with proliferation of mesangial cells and deposition of IgM, many authorities would assign a diagnosis of minimal-change nephropathy and advocate treatment with corticosteroids; some would consider that these morphological features are markers for a poorer prognosis and a reduced likelihood of a response to corticosteroids. Others would consider the patient to have a completely different disease entity and give a diagnosis of IgM nephropathy. Some of the confusion may be explained by methodological factors: assessment of the degree of mesangial hypercellularity is subjective, and reagents to detect IgM are notoriously unreliable since they may give high background staining. Mesangial IgM has been found in up to 60 per cent of 'normal' kidneys donated for transplantation; the diagnostic significance of IgM is also cast into doubt by its presence in over 75 per cent of controls as well as in patients with various other forms of glomerulonephritis. The best support for the existence of IgM nephropathy, as an entity distinct from minimal-change nephropathy, comes from the occurrence of a familial form and from the identification of this pattern of glomerular injury in patients who, after lengthy follow-up, have an appreciable risk of developing impaired excretory kidney function.

Idiopathic mesangial proliferative GN

This term may be applied if there is isolated mesangial proliferation without deposition of IgA or IgM. Again there is overlap with minimal-change nephropathy: if the patient presents with nephrotic syndrome, most nephrologists would not allow the presence of mesangial proliferation to deflect them from treating the patient with corticosteroids, although there is evidence that the presence of this histological finding is associated with a poorer response rate. If, however, the patient has haematuria and/or hypertension and/or impaired kidney function, none of which are typical features of minimal-change nephropathy, it is difficult to resist the need for another separate diagnostic category. Unfortunately there are no informative studies to guide treatment or give information on prognosis.

Endocapillary proliferative GN

Patients will often have impaired excretory function, haematuria, proteinuria, and hypertension, sometimes presenting acutely as a 'nephritic syndrome'. On renal biopsy, the glomerular hypercellularity is confined within the glomerular capillary tuft, which is probably due to the combination of a proliferation of intrinsic (endothelial and mesangial) cells together with an infiltration of inflammatory cells. This can occur in systemic lupus erythematosus and as a complication of a variety of infections (see Chapter 20.7.8). Only poststreptococcal GN will be considered here.

Poststreptococcal GN (Plate 1)

Most infection-related GN occurs concurrently with the infection. By contrast, postinfectious GN (of which poststreptococcal GN is the most frequent and best characterized) occurs, as the name implies, after the infection. In poststreptococcal GN the delay between the inciting infection and the onset of the renal complication may be long enough for the infection to have been forgotten, and this may contribute to diagnostic confusion. The typical case follows infection with streptococci of Lancefield group A (b-haemolytic streptococci, *S. pyogenes*), either causing pharyngitis or skin infections such as cellulitis or impetigo. Children are the most common victims. Around 2 weeks later, sometimes longer after skin infections, the patient develops nephritis which may be sufficiently acute and severe to cause a nephritic syndrome with oliguria, hypertension, and oedema. If a renal biopsy is performed, it will show diffuse proliferative GN, with infiltration by neutrophil polymorphs often particularly prominent (Fig. 1). Immunohistology shows deposition of IgG, IgM, and complement in the mesangial and subepithelial areas, and electron microscopy shows large subepithelial deposits ('humps').

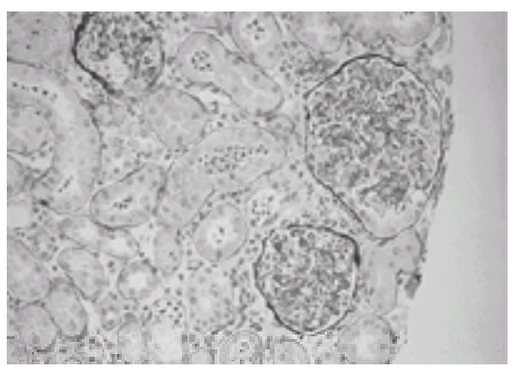

Fig. 1 Poststreptococcal glomerulonephritis.

Serological tests

There are typical serological features which give clues to the pathogenesis: these include antibodies to streptococcal antigens and evidence of activation of the complement cascade. The antibodies are IgG; reactivity with numerous streptococcal antigens has been reported including streptolysin O, deoxyribonuclease B, hyaluronidase, and streptokinase. Anti-streptolysin O is the most useful diagnostic test after pharyngitis, anti-DNAse B is best after skin infections. Hypocomplementaemia (low C3 in the majority of cases, also low C4 in a smaller proportion) reflects activation of both the alternative and the classical pathways (the complement system is discussed in more detail in Chapter 20.7.6. In poststreptococcal GN, the alternative complement pathway may be activated by bacterial antigens and/or by IgG autoantibodies called nephritic factors which resemble those seen in MCGN; the classical pathway may be activated by circulating immune complexes.

Pathogenesis

It is believed that the pathogenesis of poststreptococcal GN can be explained as follows: streptococcal antigens are deposited in glomeruli, by virtue of some aspect of their charge, size, or other physicochemical characteristics, during the early phase of the infection. After the 10 to 14 days necessary for the host to mount an immune response to the bacterial infection, circulating antibody appears and binds to the 'planted' antigens in the glomeruli. Complement is activated, leucocytes are attracted (by complement-activation products C3a and C5a among other chemoattractants) and an inflammatory reaction is provoked, injuring the glomeruli. The precise nature of the streptococcal antigens that act in this nephritogenic manner remains controversial; only certain serological types of streptococci (referred to as M types and serotyped according to cell-wall protein antigens) are capable of inciting GN, but the M proteins themselves are not believed to be nephritogenic. In addition to the planted antigen mechanism, streptococci may lead to GN by their other complex effects on the immune response. These include the direct activation of T cells by a superantigen effect, whereby M proteins can bind to particular Vb regions of the T-cell receptor and activate families of T cells sharing receptors of this 'family'. Antigenic crossreactivity ('molecular mimicry') akin to that thought to be responsible for rheumatic fever may also occur, so that anti-streptococcal antibodies crossreact with, and therefore bind to, renal autoantigens such as laminin and collagen.

Management

Poststreptococcal GN is less common in the developed than in the developing world, possibly influenced by socioeconomic factors. Its general importance lies in the fact that early recognition allows appropriate treatment, with the prognosis often being very good, and also that the immunopathological mechanisms outlined above may be instructive in understanding other forms of GN where the inciting stimulus is not so evident. Treatment of patients with poststreptococcal GN should be directed at eradicating the infection (a 10-day course of penicillin or erythromycin is advised even if the original infection appears to have resolved) and providing symptomatic relief of the consequences of the acute nephritis: aggressive treatment of hypertension, salt, and water restriction with or without diuretics for oedema; and dialysis if necessary (which is uncommon). Recovery is the rule, although haematuria and proteinuria may persist and some authors believe that in the long-term there is a risk of chronic renal failure.

Idiopathic diffuse proliferative GN

A few cases will have no preceding history of infection, no evidence of lupus, and/or atypical features on the renal biopsy: these may be assigned the unsatisfactory 'idiopathic' sobriquet, with the implication that the prognosis and the appropriate treatment are uncertain.

Further reading

Bloom PM, Filo RS, Smith EJ (1976). Immunofluorescent deposits in normal kidneys. *Kidney International* **10**, 539. [Report that IgM is present in glomeruli of 60 per cent of kidneys donated for transplantation]

Ji-Yun Y, *et al*. (1984). No evidence for a specific role of IgM in mesangial proliferation of idiopathic nephrotic syndrome. *Kidney International* **25**, 100–6. [High incidence of glomerular IgM deposits in controls as well as in nephritic kidneys]

Kefalides NA, *et al*. (1986). Antibodies to basement membrane collagen and laminin are present in sera from patients with post-streptococcal glomerulonephritis. *Journal of Experimental Medicine* **163**, 585–602. [Suggests 'molecular mimicry' as a pathogenetic mechanism in poststreptococcal glomerulonephritis]

O'Donoghue DJ, *et al*. (1991). IgM-associated primary diffuse mesangial proliferative glomerulonephritis: natural history and prognostic indicators. *Quarterly Journal of Medicine* **79**, 333–50. [Supports a separate diagnostic entity of IgM nephropathy, with implications for prognosis and treatment]

Oliveira DBG (1997). Poststreptococcal glomerulonephritis: getting to know an old enemy. *Clinical and Experimental Immunology* **107**, 8–10. [Editorial review of pathogenetic mechanisms in poststreptococcal glomerulonephritis]

Scolari F, *et al*. (1990). Familial IgM nephropathy: a morphologic and immunogenetic study of three pedigrees. *American Journal of Nephrology* **10**, 261–8. [Suggests familial form of IgM nephropathy]

Watanabe-Ohnishi R, *et al*. (1994). Characterization of unique human TCR V beta specificities for a family of streptococcal superantigens represented by rheumatogenic serotypes of M protein. *Journal of Immunology* **152**, 2066–73. [Evidence of superantigen effects of streptococcal proteins, which may contribute to pathogenesis of poststreptococcal glomerulonephritis]

20.7.6 Mesangiocapillary glomerulonephritis

Peter W. Mathieson

Mesangiocapillary glomerulonephritis (**MCGN**) is synonymous with membranoproliferative glomerulonephritis (**MPGN**). The term describes a morphological pattern of glomerular injury in which there is diffuse thickening of the glomerular basement membrane (**GBM**) associated with increased cellularity, giving a characteristic lobular appearance to the glomeruli (Fig. 1 and Plate 1). As with other forms of glomerulonephritis, such as membranous nephropathy (see Chapter 20.7.4), the appearances on light microscopy are indistinguishable whether the lesion occurs as a primary 'idiopathic' renal disease or secondary to an extrarenal/systemic disorder. Extra information is obtained with the use of immunohistology and electron microscopy, which allow further subdivision into three patterns. In type I MCGN there is typically IgG, IgM, and complement C3 in mesangial areas as well as along the glomerular capillary loops in a subendothelial or intramembranous location, and electron microscopy shows discrete electron-dense deposits in these regions. In type II MCGN there is typically no immunoglobulin deposited, but C3 is detected in a linear distribution along the capillary loops and often also in tubular and vascular basement membranes. Electron microscopy shows typical thick, linear, electron-dense material along these basement membranes, giving rise to the other term for type II MCGN which is '(linear) dense-deposit disease' (Fig. 2). Type III MCGN is similar to type I, except that there are subepithelial as well as subendothelial deposits and there may be disruption of the GBM with accumulation of new basement membrane material in layers. Most 'secondary' forms of MCGN are of the type I pattern.

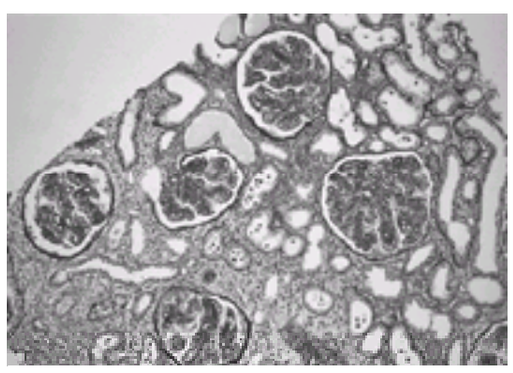

Fig. 1 Mesangiocapillary glomerulonephritis. Note characteristic lobular appearance of expanded glomerulus. (See also Plate 1.)

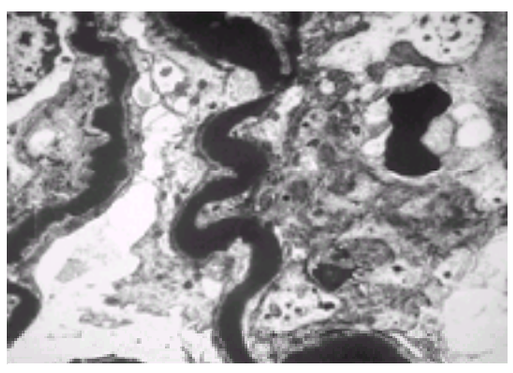

Fig. 2 Electron micrograph of type II MCGN, 'dense deposit disease'. Note linear, electron-dense material along the glomerular basement membrane.

The complement system

The subdivision of MCGN into types I, II, and III is not just of academic importance. MCGN is the type of glomerulonephritis most closely associated with activation of the complement cascade and there is evidence, at least for some types of MCGN, that complement dysregulation may directly cause the renal lesion. The pattern of complement activation differs between the three subtypes of MCGN. Complement activation can occur via two main pathways, the classical and alternative pathways (Fig. 3). A recently described third pathway, the lectin or mannan-binding pathway, yields similar results to classical pathway activation and is of unknown relevance to nephritis. In general, classical pathway activation leads to depletion of plasma C3 and C4, whereas alternative pathway activation leads to a low C3 with normal C4. This is an oversimplification, since both C3 and C4 are acute-phase reactants whose synthesis is upregulated in inflammation. Thus there may be considerable complement activation without depletion of circulating levels, due to increased production. Further complexity is introduced by the fact that genetic deficiencies of C4 are common: there are four C4 genes encoded within the major histocompatibility complex (**MHC**) on chromosome 6, and one or more null alleles are commonly present, which result in no C4 protein production. These result in a reduction of circulating C4 concentrations—it is estimated that only 60 per cent of the normal population has all four normal C4 genes. Thus a single low C4 level must be interpreted with caution unless a previous 'normal' result is available for that individual; serial measurements are helpful since they give an indication of the level of C4 consumption.

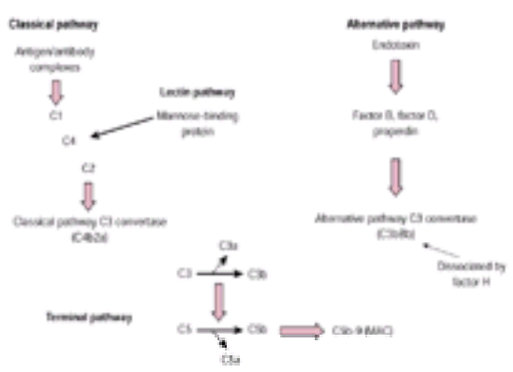

Fig. 3 The complement system.

The classical pathway is activated predominantly by immune complexes: immunoglobulin molecules linked to antigen. The alternative pathway is more concerned with host defence, being activated by bacteria or other foreign surfaces, and existing in a state of constant low-level activity: so-called 'tickover'. This state of constant

activity demands tight regulation to avoid excessive activity, and this is achieved by regulatory proteins factor H and factor I. The classical and alternative pathways converge at the point at which C3 is cleaved. The enzyme formed by classical pathway activation is called the classical pathway C3 convertase, and denoted C4b2a. The enzyme formed by alternative pathway activation is the alternative pathway C3 convertase, denoted C3bBb. Each of these enzymes leads to the cleavage of C3, releasing C3a and leading to the formation of a C5 convertase enzyme which cleaves C5, thereby releasing C5a and leading to the formation of C5b–9, the membrane-attack complex (**MAC**).

Pathogenesis of MCGN

In MCGN, the pattern of complement activation (and therefore possibly the pathogenesis) is different in each of the three forms. In type I, the complement activation predominantly affects the classical pathway (causing low levels of both plasma C3 and C4); in type II it is the alternative pathway which is predominantly affected (low C3 with normal C4); and in type III there is activation of the terminal pathway leading to depletion of C5, sometimes associated with mild depletion of C3 and/or C4. 'Secondary' MCGN may occur in systemic lupus erythematosus (**SLE**); cryoglobulinaemia with or without hepatitis C; infections such as infective endocarditis or other chronic bacteraemic states (for example, 'shunt nephritis', originally described with infected ventriculoatrial shunts); or in association with neoplasms. In each of these situations there is activation of the classical pathway which is presumed to be due to circulating immune complexes, and this is associated with a type I MCGN pattern. In idiopathic type I MCGN, a variety of complement-activating factors have been described: there may be circulating immune complexes, some patients have antibodies to Clq which probably directly activate the classical pathway, and some have other autoantibodies which interfere with the normal regulation of the classical pathway. In type II MCGN, the alternative pathway activation is due to the presence of an IgG autoantibody (known as C3 nephritic factor C3NeF, or more simply as nephritic factor, NeF) which binds to a neoantigen formed when the alternative pathway C3 convertase enzyme, C3bBb, is assembled. The antibody stabilizes this enzyme and protects it from degradation by factor H. Thus the half-life of the enzyme is prolonged, and the normal regulatory mechanism is subverted. This type of nephritic factor has also been described in patients with type I and type III MCGN, but its presence is virtually invariable in type II MCGN. In type III MCGN, the presence of a circulating factor which activates complement slowly in a properdin-dependent manner has been postulated; the reasons for the preferential depletion of terminal pathway components, and whether there is a direct relationship of this activation to the renal injury in type III MCGN, remain unanswered questions.

The best evidence for a causative role of complement activation in MCGN comes in type II disease. As mentioned above, most patients with type II MCGN have the IgG autoantibody known as nephritic factor (**NeF**) which allows unregulated alternative pathway activation. Two other situations in which there is similar overactivity of the alternative pathway, and an associated renal lesion with the appearances of type II MCGN, have recently been characterized. First, genetic deficiency of the regulatory protein known as factor H, which normally serves to degrade the alternative pathway C3 convertase, has been reported in a variety of inbred pigs and also in rare human cases. Second, there is a case report of an individual whose serum contained a monoclonal lambda light chain which interacted with factor H *in vitro* and prevented its action, allowing unregulated alternative pathway activation. Therefore, in these three situations (the presence of NeF, genetic deficiency of factor H, or functional blocking of factor H), there is dysregulated alternative pathway activation, but due to completely different mechanisms. In each case, the renal lesion is type II MCGN, strongly suggesting that it is the complement activation *per se* which leads to the renal injury. Importantly, in the factor H-deficient pigs, replacement of factor H leads to prevention of the excessive alternative pathway activation and an improvement in the MCGN. These observations have clear implications for the therapy of human MCGN (discussed further below).

The NeF autoantibody, and the resultant unregulated alternative pathway activation, have another striking clinical association: with partial lipodystrophy in which there is permanent loss of adipose tissue from the face and neck and sometimes also from the upper trunk (Fig. 4). Such patients may also have type II MCGN, and as with the renal lesion, there is evidence to suggest that the complement activation directly causes the tissue injury: NeF containing IgG can cause complement-mediated lysis of adipocytes *in vitro*. Furthermore, adipocytes are probably susceptible to this injury because they produce complement components: these observations have contributed to the recent appreciation that the complement system plays a previously unsuspected role in the normal physiological regulation of adipose tissue.

Fig. 4 Facial appearance in partial lipodystrophy. This patient has had silicone pads inserted into her cheeks, accounting for the bulges in the regions where adipose tissue has been completely lost.

Clinical presentation

This is with proteinuria, which may be sufficiently severe to cause nephrotic syndrome; and/or haematuria, which, especially in children, may be macroscopic. Hypertension and/or impairment of excretory kidney function may be associated. Acute presentation as a nephritic syndrome is recognized in children. As mentioned above, type II MCGN may be associated with partial lipodystrophy; the loss of adipose tissue can precede the onset of nephritis by many years. Abnormalities in the eye are also recognized in type II MCGN, and may rarely be the presenting feature. Drusen-like deposits and mottled pigmentation are visible in the fundi: retinal neovascularization may occasionally threaten sight and require laser therapy.

Natural history of MCGN

Overall, the renal survival in MCGN at 10 years from diagnosis is about 50 per cent. Children tend to have a more acute presentation and a slower decline in renal function, although with lengthy follow-up the overall renal survival is similar to that in adults. The prognosis differs between the three subtypes of MCGN, with type II carrying the greatest risk of the development of endstage renal failure (**ESRF**): in one recent study the median time to ESRF in types I, II, and III was, respectively, 15.3 years, 8.7 years, and 15.9 years. Since presentation with the nephrotic syndrome carries a substantially increased risk of ESRF compared to other milder clinical syndromes, the adverse prognosis of type II MCGN may simply reflect the greater likelihood of nephrotic presentation with this histological type. As in many other forms of glomerular disease the presence of tubular atrophy and interstitial fibrosis indicate a worse prognosis, as does hypertension at the time of presentation.

Treatment of MCGN

The forms of therapy which have been applied to MCGN are similar to those used in other forms of nephritis: antiplatelet drugs, anticoagulants, corticosteroids, and alkylating agents have been used alone or in various combinations. The studies tend to be small, with varying proportions of children and adults, and of the three subtypes of MCGN. There is a dearth of randomized controlled trials and reviews of the subject have usually concluded that there is no treatment of proven efficacy in MCGN. Nevertheless, there are hints from some of the studies that certain drugs may have useful effects. In particular, high doses of prednisone, usually given on alternate days (especially in children) are favoured by some authors, especially the Cincinnati group who have published most extensively on this subject. However, high corticosteroid dosages are required for prolonged periods and the magnitude of benefit obtained may be too small to justify the risks of such treatment. Possibly by refining the dosage schedule and by applying the treatment only to high-risk groups, such as those with severe nephrotic syndrome, the risk:benefit ratio may be more favourable.

Since complement activation is so prominent in MCGN, therapy aimed at the complement system may be rational: promising anticomplement agents are now becoming available. At present, the best strategy is to try to identify any underlying cause of complement activation and remove it if possible. As in other forms of GN, associated hypertension should be aggressively treated. Patients with proteinuria of any cause have their renal prognosis improved if they are treated with angiotensin-converting enzyme (**ACE**) inhibitors. They are at increased cardiovascular risk, and attention should also be paid to other modifiable risk factors,

especially cigarette smoking and hyperlipidaemia.

Recurrent MCGN in renal transplants

MCGN is one of the types of nephritis that tends to recur in kidney transplants (see Chapter X): type I recurs in around 25 to 30 per cent of grafts and type II recurs even more frequently, possibly in 85 to 90 per cent of cases. Recurrent MCGN only causes graft failure in a minority of cases, presumably because the antirejection immunosuppressive therapy modulates the damage done by the nephritis.

Further reading

Cameron JS (1982). Glomerulonephritis in renal transplants. *Transplantation* **34,** 237–45. [Review of glomerulonephritis (primary and recurrent) in transplanted kidneys]

Donadio JV, Offord KP (1989). Reassessment of treatment results in membranoproliferative glomerulonephritis, with emphasis on life-table analysis. *American Journal of Kidney Diseases* **14,** 445–51. [Cautionary tale about the assessment of treatment effects in MCGN, refuting earlier claims of effectiveness of antiplatelet therapy]

McEnery PT (1990). Membranoproliferative glomerulonephritis: the Cincinnati experience—cumulative renal survival from 1957 to 1989. *Journal of Pediatrics* **116,** S109–14. [Review of the Cincinnati experience with childhood MCGN, especially supporting the role of corticosteroid therapy]

Mathieson PW (1998). Is complement a target for therapy in renal disease? *Kidney International* **54,** 1429–36. [Review of the role of complement in various forms of renal injury, discussion of currently available anticomplement therapies, and speculation on future possibilities]

Mathieson PW (1999). Mesangiocapillary glomerulonephritis. In: CD Pusey, ed. *Treatment of glomerulonephritis*, pp 81–92. Kluwer Academic, Dordrecht, The Netherlands. [Review of literature on the treatment of MCGN, suggestions for future strategies]

Mathieson PW, Peters DK (1997). Lipodystrophy in MCGN type II: the clue to links between the adipocyte and the complement system. *Nephrology, Dialysis, and Transplantation* **12,** 1804–6. [Review of the evidence for a causative role of complement activation in type II MCGN and partial lipodystrophy, discussion of the complement system's role in the biology of adipose tissue]

Ruggenenti P *et al*. (1998). Renal function and requirement for dialysis in chronic nephropathy patients on long-term ramipril. *Lancet* **352**, 1252–6. [Influential recent trial confirming the benefits of ACE inhibition in patients with proteinuria, irrespective of the underlying cause]

Schena FP, Cameron JS (1988). Treatment of proteinuric idiopathic glomerulonephritides in adults: a retrospective survey. *American Journal of Medicine* **85,** 315–26. [Review of literature on treatment of all forms of glomerulonephritis in adults]

Schwertz R, *et al*. (1996). Outcome of idiopathic membranoproliferative glomerulonephritis in children. *Acta Paediatrica Scandinavica* **85,** 308–12. [Review of natural history of MCGN in children, especially trying to analyse any differences between types I, II, and III]

Sissons JGP, *et al*. (1979). The complement abnormalities of lipodystrophy. *New England Journal of Medicine* **294,** 461–5. [Largest series in which patterns of complement activation have been analysed in patients with different types of lipodystrophy]

Varade WS, Forristal J, West CD (1990). Patterns of complement activation in idiopathic membranoproliferative glomerulonephritis, types I, II, III. *American Journal of Kidney Diseases* **16,** 196–206. [Analysis of patterns of complement activation in subtypes of MCGN, review of reported mechanisms]

20.7.7 Antiglomerular basement membrane disease

Jeremy Levy and Charles Pusey

Introduction

Antiglomerular basement membrane disease is an autoimmune disease in which patients develop pathogenic autoantibodies against the glomerular basement membrane (**GBM**). Patients typically present with renal failure and pulmonary haemorrhage, but isolated renal disease is well recognized. The triad of anti-GBM antibodies, rapidly progressive glomerulonephritis (**RPGN**), and pulmonary haemorrhage is referred to as Goodpasture's disease in the United Kingdom, whilst the term Goodpasture's syndrome describes patients with RPGN and pulmonary haemorrhage of various aetiologies.

The term 'Goodpasture's syndrome' was first used in 1958 by Stanton and Tange in their report of nine patients with pulmonary–renal syndrome, which referred to a patient with fulminant pulmonary haemorrhage and proliferative glomerulonephritis described by Goodpasture during the influenza pandemic of 1919. In retrospect, this original patient may have had systemic vasculitis, not anti-GBM disease. In recent years much has been learnt about the immune response in Goodpasture's disease, but despite huge advances in our understanding of aetiopathogenesis, the therapy has changed little in the last 20 years.

Aetiology and pathogenesis

The Goodpasture antigen

All patients with Goodpasture's disease have circulating antibodies that bind a glomerular basement membrane antigen, the a3 chain of type IV collagen (a3(IV)). Type IV collagen is found in all basement membranes; but the a3, a4, and a5 chains are restricted in their distribution primarily to the GBM and alveolar basement membranes. The epitope for autoantibodies in Goodpasture's disease is carried at the amino terminus of the 230 amino acid, non-collagenous, carboxy-terminal domain of the a3 chain (a3(IV)NC1), which is normally hidden within the collagen network. The a3(IV)NC1 is also found in the basement membranes of the choroid plexus, the cochlea, Bruch's membrane in the eye, retinal capillaries, and the thymus.

Anti-GBM antibodies and the T-cell-mediated immune response

Transfer of antibodies from patients into squirrel monkeys initially confirmed the pathogenicity of the autoantibodies. Clinical studies report a correlation between antibody levels at presentation and disease activity, and the disease recurs immediately in renal transplants when the recipient still has circulating antibodies. All patients have antibodies against a3(IV)NC1, either circulating or bound to the GBM. A small number of patients develop antibodies against other GBM components, particularly the a1 (15 per cent of patients) or a4 (4 per cent of patients) chains of type IV collagen. However, anti-GBM antibodies are unlikely to be the only cause of glomerular injury, and a cell-mediated immune response is also important in inducing renal damage.

Alveolar haemorrhage generally requires a second insult, either local to the lungs (for example, cigarette smoking or pulmonary oedema), or systemic with activation of cytokines and inflammatory mediators (for example, sepsis).

Genetic predisposition

Goodpasture's disease has been reported in four sibling pairs and two sets of identical twins; however, discordant twin pairs are also documented. More striking is the association with the HLA serotype HLA-DR2, which is carried by more than 85 per cent of patients with Goodpasture's disease, compared with 30 per cent of controls. Molecular analysis of HLA alleles has confirmed the association with HLA-DR15 (DRB1*1501 and -1502), and a weaker association with HLA-DR4 (DRB1*04). A negative association with HLA DR7 (DRB1*07) has been demonstrated. Thus, specific characteristics of the HLA molecules on antigen-presenting cells determine susceptibility to Goodpasture's disease.

Environmental factors

No specific pathogens or toxins have been identified that can initiate Goodpasture's disease; but many case reports have documented exposure to hydrocarbons prior to the development of clinical manifestations, and cigarette smoking undoubtedly precipitates pulmonary haemorrhage. It seems more likely that organic solvents trigger overt pulmonary damage (and possibly renal injury) in the presence of circulating autoantibodies than that hydrocarbons are involved in the initiation of autoimmunity. Several clusters of cases have been reported, but no clear associations with influenza virus or other infectious agents have been proved.

Disease associations

Anti-GBM disease is only rarely associated with other autoimmune disorders, apart from systemic vasculitides. Increasing numbers of patients (up to 30 per cent) have been shown to have circulating antineutrophil cytoplasmic antibodies (**ANCA**), generally P-ANCA, in addition to anti-GBM antibodies. Conversely, only few patients with ANCA-associated vasculitis also have anti-GBM antibodies (2.5–8 per cent). This is an important distinction since the 'double positive' patients tend to behave more like those with 'pure' Goodpasture's disease than systemic vasculitis. Anti-GBM disease has been reported after lithotripsy and urinary tract obstruction, and in some patients with membranous nephropathy. In all these cases it is possible that disruption of the GBM in susceptible individuals can lead to a breakdown in tolerance to the a3 chain of type IV collagen, with the development of autoantibodies and clinical disease.

Epidemiology

Limited epidemiological studies suggest that Goodpasture's disease has an incidence of 0.5 to 1 new case per million of the population per year. It is found in 1 to 2 per cent of renal biopsies. In comparison, systemic vasculitides have an incidence of 15 to 30 new cases per million of the population per year. The disease is less common in Afro-Caribbean and Asian populations. There is a bimodal age distribution, with peak incidence in the third and sixth decades, and a slight excess of males.

Clinical features

Most patients present with RPGN or lung haemorrhage, or both. Some patients have isolated lung haemorrhage and never develop renal failure (although most of

these have haematuria and proteinuria), and a few have mild isolated nephritis. General malaise, fatigue, weight loss, and anaemia are the commonest systemic features, whilst other signs and symptoms are much rarer than in patients with systemic vasculitis.

Pulmonary features

Pulmonary haemorrhage occurs in two-thirds of patients, more commonly in young men, it usually precedes presentation with acute renal failure, and is strongly associated with cigarette smoking. Patients often complain of breathlessness and cough, and there is a poor relationship between overt haemoptysis and the degree of alveolar haemorrhage. Haemoptysis can be triggered by cigarettes, inhaled toxins, fluid overload, and intercurrent infection, either local (pneumonia) or systemic (sepsis). Clinical signs are often indistinguishable from those of pulmonary oedema or infection. The most sensitive indicator is an elevated K_{CO} (diffusing capacity for carbon monoxide), which identifies the presence of haemoglobin in alveolar spaces by increased binding of inhaled carbon monoxide. Radiographic features are not specific, but alveolar shadowing in the central lung fields is typically seen (Fig. 1).

Fig. 1 Chest radiograph from a patient with Goodpasture's disease showing florid pulmonary haemorrhage.

Renal features

Patients can present with isolated haematuria, chronic renal failure, or mild renal insufficiency, but classically present with severe acute renal failure due to rapidly progressive glomerulonephritis. The clinical features of the nephritis are indistinguishable from any other cause of RPGN, with cellular casts in the urine, haematuria, and mild to moderate proteinuria (nephrotic range proteinuria is rare). Hypertension and oliguria are late features. A small number of patients have relatively normal renal function at presentation, but always have abnormal urine findings and evidence of antibody deposition on renal biopsy.

Differential diagnosis

It is crucial to distinguish anti-GBM disease from other cause of RPGN, and especially ANCA-associated vasculitis. There is only a small window of opportunity in which to rescue renal function in patients with anti-GBM disease, by contrast to systemic vasculitis in which renal failure can be reversed at a later stage. All patients with suspected RPGN, acute renal failure of unknown cause, or lung haemorrhage and urinary abnormalities, should have both anti-GBM antibody and ANCA assays performed urgently. Double-positive patients may respond better to therapy at a late stage compared to those with pure anti-GBM disease. Other differential diagnoses to consider include systemic lupus erythematosus (**SLE**), cryoglobulinaemia, haemolytic uraemia syndrome (**HUS**), and other causes of pulmonary renal syndrome (see Table 1).

Pathology

Immunohistology is characteristic (Fig. 2 and Plate 1), with linear deposition of IgG (sometimes with IgA or IgM) and complement C3 along the glomerular basement membranes. Rare patients have been reported with IgM or IgA alone. Less intense linear staining with IgG can occasionally be seen in diabetes, SLE, myeloma, and transplanted kidneys. The most characteristic morphological finding is severe crescentic glomerulonephritis, with almost all the glomeruli exhibiting cellular crescents, usually at the same stage of evolution. Segmental necrosis and cellular proliferation may occur. Blood vessels are usually normal, but rarely vasculitis has been reported even in the absence of detectable ANCA. There is often a prominent interstitial cellular infiltrate.

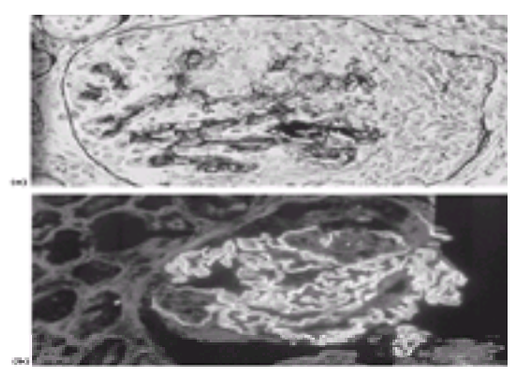

Fig. 2 Renal biopsy from a patient with Goodpasture's disease. (a) Light microscopy showing a single glomerulus with cellular crescent and focal necrosis (silver stain). (b) Immunofluorescence of a single glomerulus with linear deposition of IgG along the GBM. (Figure by courtesy of Dr HT Cook.) (See also Plate 1.)

Histological specimens are rarely obtained from lungs, since transbronchial biopsy does not usually penetrate beyond the bronchial mucosa. Open-lung biopsy can reveal alveoli full of red blood cells, macrophages, and fibrin, interspersed between relatively normal alveoli. Immunofluorescence inconsistently reveals linear deposition of antibody.

Laboratory diagnosis

Serological testing for anti-GBM antibodies and ANCA is crucial for confirming the diagnosis, and a renal biopsy is almost always warranted. Some healthy individuals exposed to inhaled oils, hydrocarbons, or solvents may have borderline raised anti-GBM antibody levels, and anti-GBM antibody has also been detected in HIV-negative patients with pneumocystis pneumonia. Other investigations are detailed in Table 2. Alveolar haemorrhage is an important cause of mortality and must be identified early. All patients should have baseline K_{CO} and chest radiology, repeated as necessary.

Treatment

Untreated anti-GBM disease is usually fatal, and renal function never recovers. In most centres immunosuppressive treatment is given immediately upon diagnosis to those with a serum creatinine concentration below 600 µmol/l at presentation and/or with active pulmonary haemorrhage. Those with a serum creatinine concentration above 600 µmol/l and without active pulmonary haemorrhage need more careful consideration before being treated since they have a small chance of recovery (see below).

Treatment with plasma exchange, cyclophosphamide, and corticosteroids, together with dialysis when required, can allow up to 90 per cent of patients to survive, but

only around 40 per cent of survivors will recover renal function. Daily plasma exchange removes circulating antibodies, whilst cyclophosphamide prevents further antibody synthesis. There has only been one controlled trial of plasma exchange, which utilized a low intensity of exchanges in a small number of patients and showed a non-significant trend towards an improved outcome. However, the dramatic improvement in overall mortality and renal function coincident with the introduction of a treatment regimen of the type described above has led to their widespread use. The regimen we use is shown in Table 3. An alternative to plasma exchange is protein-A immunoadsorption. Ciclosporin has been used in occasional patients unresponsive to other therapies, but is of doubtful benefit. Long-term treatment is unnecessary, and patients can stop taking cyclophosphamide after 2 to 3 months, and withdraw prednisolone over approximately 6 months.

Prognosis

The outcome of patients with Goodpasture's disease in published series is shown in Table 4. Most will now survive the acute illness, but pulmonary haemorrhage and infection remain important causes of death. In those with a serum creatinine concentration below 600 µmol/l at presentation, the creatinine should begin to fall within 1 to 2 weeks of treatment, and the majority will recover renal function. However, patients with a creatinine concentration above 600 µmol/l, or with oligoanuria, less commonly recover renal function. For this reason most centres would not give immunosuppressive agents to this group with the sole intention of trying to restore renal function, although they would for concurrent active pulmonary haemorrhage or if the renal biopsy suggests that tubular necrosis may be contributing to the severity of the renal failure (see above). Crescent scores over 50 per cent, high antibody titres, and a delay in diagnosis are also markers of a poor renal prognosis.

The prognosis in anti-GBM disease is in marked contrast to that of patients with a diagnosis of ANCA-associated RPGN. Renal recovery is to be expected in the latter group with immunosuppression, and around 70 per cent of patients presenting with a creatinine concentration above 600 µmol/l will recover renal function.

Relapses of pulmonary haemorrhage and worsening of renal function can occur early during the course of treatment in the presence of circulating autoantibodies, and can be triggered by smoking, infection, or fluid overload. True late recurrence is very unusual. Transplantation is safe once autoantibodies are no longer detectable, and is best delayed until between 6 and 12 months after the disappearance of anti-GBM antibody.

Anti-GBM disease in Alport's syndrome

Patients with X-linked Alport's syndrome have a mutation in the a5 chain of type IV collagen, but also have undetectable Goodpasture antigen in their kidneys despite a normal a3(IV) gene. Transplantation of a normal kidney into such recipients may allow the development of anti-GBM antibodies as a result of the exposure of the immune system to neoantigens to which tolerance has not developed. These antibodies are usually anti-a5(IV)NC1, but can also be anti-a3(IV)NC1 (classical Goodpasture autoantibodies). Most patients do not develop overt nephritis, but simply deposit antibody along the GBM without recruiting a glomerular inflammatory response. However, a minority develop severe glomerulonephritis. In the absence of lung antigen, pulmonary haemorrhage never occurs.

Further reading

Herody M, *et al.* (1993). Anti-GBM disease: predictive value of clinical, histological and serological data. *Clinical Nephrology* **40**, 249–55. [Detailed description and outcome of French series.]

Johnson JP, *et al.* (1985). Therapy of anti-glomerular basement membrane antibody disease: analysis of prognostic significance of clinical, pathological and treatment factors. *Medicine* **64**, 219–27. [Single controlled trial of plasma exchange in Goodpasture's disease.]

Lerner RA, Glassock RJ, Dixon FJ (1967). The role of anti-glomerular basement membrane antibodies in the pathogenesis of human glomerulonephritis. *Journal of Experimental Medicine* **126**, 989–1004. [Classic paper describing the transfer of disease by anti-GBM antibodies.]

Levy JB, Pusey CD (1997). Anti-glomerular basement membrane disease. In: Wilkinson R, Jamison R, eds. *Nephrology*, pp 599–615. Chapman Hall, London. [Comprehensive review of anti-GBM disease.]

Levy JB *et al*. (2001). Long-term outcome of anti-GBM antibody diseases treated with plasma exchange and immunosuppression. *Annals of Internal Medicine* **134**, 1033–42. [Largest series of patients reported.]

Levy JB, Pusey CD (1999). Plasmapheresis. In: Johnson R, Feehally J, eds. *Comprehensive clinical nephrology*, pp 83.1–8. Mosby, London. [Review of the techniques, use and complications of plasma exchange.]

Lockwood CM, *et al.* (1976). Immunosuppression and plasma exchange in the treatment of Goodpasture's syndrome. *Lancet* **i**, 711–15. [Classic description of treatment of Goodpasture's disease.]

Merkel F, *et al.* (1994). Course and prognosis of anti-basement membrane antibody mediated disease, a report of 35 cases. *Nephrology, Dialysis, Transplantation* **9**, 372–6. [A recent series from Europe documenting prognostic factors.]

Saus J, *et al.* (1988). Identification of the Goodpasture antigen as the a3(IV) chain of collagen IV. *Journal of Biological Chemistry* **263**, 13374–80. [Initial identification of the Goodpasture antigen.]

Savage COS, *et al.* (1986). Anti-GBM antibody mediated disease in the British Isles 1980–1984. *British Medical Journal* **292**, 301–4. [Largest series of patients reported in the United Kingdom.]

Stanton MC, Tange JD (1958). Goodpasture's syndrome, pulmonary haemorrhage associated with glomerulonephritis. *Australasian Annals of Medicine* **7**, 132–44. [Initial description of Goodpasture's syndrome.]

Turner N, *et al.* (1992). Molecular cloning of the human Goodpasture antigen demonstrates it to be the alpha 3 chain of type IV collagen. *Journal of Clinical Investigation* **89**, 592–601. [Molecular characterization of the Goodpasture antigen.]

20.7.8 Infection-associated nephropathies

A. Neil Turner

Introduction

Almost all varieties of renal lesion, particularly glomerular, may be associated with infections. The pathways leading to these are sometimes understood, but sometimes obscure. In the West, infection-associated nephritis was once predominantly recognized during episodes of acute infections in apparently healthy individuals. This is still the pattern in less developed regions; but this could be partly a problem of recognition, as in all populations infections are more common and more severe in malnourished or otherwise debilitated individuals. Improvements in living conditions and healthcare in developed countries have reduced the numbers of healthy people succumbing to complications of infection. By contrast, infections occurring on a background of debilitating illnesses and previous medical interventions have become more common, and are certainly more often diagnosed.

In this chapter, glomerular diseases and interstitial diseases associated with infection are considered in turn. Particular attention is given to those glomerulopathies associated with bacterial endocarditis and other chronic bacterial infections, and to three viral infections of worldwide importance—human immunodeficiency virus (**HIV**), hepatitis B, and hepatitis C.

Pathogenesis

Infection-associated glomerular disease is usually attributed to trapping of circulating antigen–antibody complexes or to immune responses to pathogen-derived antigens that become 'planted' in the glomerulus. The evidence for the deposition of circulating immune complexes is unequivocal for cryoglobulinaemia, and highly plausible for infections occurring within the vascular system such as bacterial endocarditis. In most other infections the evidence is less clear.

Interstitial renal disease is often blamed on direct invasion by micro-organisms, and for some, particularly viruses, there is evidence that this is true. The pathogen may cause injury directly, or indirectly by causing cells to express foreign antigens that generate an immune response. More speculatively, an immune response generated to an organism may crossreact with a remote self-antigen, triggering autoimmunity through molecular mimicry, but there are no unequivocal examples of this.

Infection may also involve the kidney by interfering with the circulation either generally (septic shock) or locally (for instance, by causing thrombotic microangiopathy, as for *Escherichia coli* O157), or *Capnocytophaga canimorsus* (previously DF-2). On occasions, toxins may be released that harm the kidney directly (for example, haemoglobin in malaria). Medically administered toxins include antimicrobial agents that impair renal function by crystallization (aciclovir, indinavir, etc.), or by predictable toxicity (e.g. aminoglycosides and amphotericin), or by idiosyncratic reactions such as acute interstitial nephritis (e.g. penicillins).

Glomerulonephritis associated with chronic and acute bacterial infections

Classic, acute postinfectious glomerulonephritis is considered separately elsewhere. This chapter centres on the more subacute or chronic diseases, although other causes of a 'classical' picture are mentioned.

Shunt nephritis was first recognized in the 1960s and remains the archetype of an immune-complex nephritis. The glomerulonephritis occurring in association with infective endocarditis is very similar. Both are caused by subacute infection within the bloodstream, with constant production and shedding of antigen and the formation of antigen–antibody complexes. Other bacterial infections cause similar pictures, or patterns more similar to acute 'postinfectious' glomerulonephritis.

Shunt nephritis

In shunt nephritis a ventriculoatrial shunt implanted for the treatment of hydrocephalus becomes colonized by bacteria, usually of low pathogenicity. More common modern equivalents of this clinical syndrome are due to infected long-term, indwelling, central vein catheters and other intravascular devices. The syndrome does not occur with ventriculoperitoneal shunts, which are therefore now the preferred neurosurgical option. Although *Staphylococcus epidermidis* has been most commonly implicated, *Propionibacterium acnes* or other organisms are sometimes involved. Typically, the diagnosis is only appreciated after weeks to months of symptoms of mild to moderate pyrexia and malaise associated with haematuria, proteinuria, and progressive renal impairment. Fevers have often been attributed to urinary infection in patients with neurogenic bladders. There may be moderate splenomegaly. Investigations show variable renal impairment, complement consumption, and an acute-phase response with a normochromic normocytic anaemia. The renal lesion is characteristically a type-1 mesangioproliferative glomerulonephritis with deposition of multiple immunoglobulins and complement components beneath the endothelium, the classic appearance of a circulating immune-complex nephritis. Sometimes the picture is more severe, showing a diffuse proliferative lesion, occasionally with crescents. In other cases the histological appearances are less pronounced with focal proliferative changes.

Antibiotic treatment alone is almost never adequate to cure these infections, which require removal of the shunt, followed by its replacement after an interval if drainage is still required. Delayed diagnosis and delayed removal may lead to severe and irreversible renal damage and sometimes to endstage renal failure. Substantial recovery often follows successful treatment.

Infective endocarditis

A similar syndrome occurs in patients with subacute bacterial endocarditis (see Chapter 15.10.2), when minor glomerular involvement is probably extremely common. The majority of signs and symptoms in these circumstances are common to shunt nephritis. Typical streptococcal infections are well represented in case series, but there have been many reports involving 'slow' infections such as Q fever (*Coxiella burnetti*) and more unusual causes including chlamydia and fungi. Infection of prosthetic or native heart valves may be implicated. Right-sided endocarditis occurring in intravenous drug abusers may be particularly likely to present as nephritis, perhaps because the diagnosis is often delayed. Depletion of serum complement is again diagnostically useful, but, as for shunt nephritis, most other serological and

haematological changes are non-specific. Partial treatment with antibiotics makes diagnosis more difficult, as positive blood cultures are usually a key part of proving the diagnosis and planning appropriate therapy.

The pathological lesion is often similar to that of shunt nephritis. A more acute endocarditis (for instance, that associated with *Staphylococcus aureus*) is more likely to cause glomerulonephritis in a diffuse proliferative pattern, sometimes with crescent formation. A third lesion has been increasingly reported in recent literature – focal changes that are indistinguishable from ANCA-associated vasculitis– indeed, in some cases ANCA have been detected. Cutaneous vasculitis may be seen in association with bacterial infection, and this seems particularly likely in endocarditis (Fig. 1 and Plate 1).

Fig. 1 Cutaneous vasculitis in a patient with *Staphylococcus aureus* endocarditis. (See also Plate 1.)

In most cases the outcome depends on the response of the endocarditis to treatment, but renal involvement is a poor prognostic factor for survival, which may be simply because it reflects a long-standing infection. Recovery from dialysis-dependence may occur.

Patients with endocarditis are also prone to two other renal lesions. Interstitial nephritis is frequently due to the prolonged administration of drugs, including high doses of antibiotics. Those with disease on the left side of the heart or with right–left shunts may suffer renal emboli. These are common at autopsies, but glomerulonephritis is probably a more common cause of urinary abnormalities in most patients.

Deep-seated bacterial infections

Amyloidosis is a well-recognized consequence of very chronic bacterial (including mycobacterial) and other infections, and is described in Chapter 11.12.4 and Chapter 20.10.4. As in reactive amyloidosis of other aetiologies, progression of the renal lesion may be prevented or even reversed by treatment of the cause.

Deep-seated infections, particularly abscesses, may also be associated with acute renal pathology. Although the mechanisms involved are presumably similar to those of shunt nephritis and nephritis associated with endocarditis, blood cultures have often been negative in reported cases. *Staphylococcus aureus* is the most frequently implicated organism. A specific type of renal disease in association with methicillin-resistant *Staph. aureus* (**MRSA**) infection has been postulated, but without strong support. A wide variety of renal lesions have been described, usually inflammatory/proliferative and with immunoglobulin deposition. Unsuspected abscesses or other deep-seated infections are occasionally found only after the renal biopsy appearances trigger a search. Such hidden abscesses are more likely to occur in the obese, the elderly, and in those prescribed corticosteroids or who are immunosuppressed by other means or by disease.

Acute glomerulonephritis and other infections

Acute glomerulonephritis resembling poststreptococcal nephritis has been reported in association with a large number of other organisms: including current (as opposed to recent) infection with staphylococci, streptococci, and other bacteria, and with acute viral infections that are usually self-limiting. These include Epstein–Barr virus, cytomegalovirus, coxsackieviruses, and the varicella, measles, and mumps viruses. Some may cause a clinical syndrome that is very similar to poststreptococcal nephritis, while others typically cause a less florid 'nephritic' or mixed 'nephritic/nephrotic' picture.

Diagnostic difficulties in bacterial infection-related glomerulonephritis

Infection-related nephritis may present in a very similar manner to nephritis associated with other systemic diseases, notably microscopic polyangiitis and other small-vessel vasculitides. As both types of disease process may be associated with fever, a systemic illness, and an acute-phase response, it is important to consider the possibility of infection in all patients thought to have systemic vasculitis. Blood cultures should be routine. **ANCA** (antineutrophil cytoplasmic antibody) assays are extremely useful, but it is important to note that ANCA positivity has been recorded in many infections, both by fluorescence and by solid-phase assays: ANCA are not diagnostic of small-vessel vasculitides. Renal biopsy is often the most discriminating investigation: infection-associated glomerulonephritis is usually associated with plentiful immunoglobulin deposition, although the pattern is variable, whereas small-vessel vasculitis is characteristically pauci-immune. Non-glomerular causes of renal impairment (interstitial nephritis, acute tubular necrosis) are also distinguished by renal biopsy.

Interstitial nephritis associated with infections

Bacterial infections

Acute bacterial pyelonephritis is usually a florid and painful disorder associated with symptoms of urinary tract infection, as described in Chapter 20.12. Substantial renal impairment is usual only if a single functioning kidney is affected. Occasionally, however, the diagnosis is masked by immunosuppression (for example, in a transplanted kidney), age, or other factors, and the diagnosis is made by the renal biopsy appearances of neutrophils in the interstitium and in tubules, which are rarely found in any other renal lesions.

Acute interstitial nephritis is a key feature of Weil's disease, a severe form of leptospirosis (see Chapter 7.11.31). Jaundice and renal failure follow a febrile illness caused by infection with *Leptospira interrogans*. The renal lesion comprises interstitial oedema with predominantly mononuclear infiltrates and foci of tubular necrosis. Renal failure is usually oliguric but may be polyuric. Dialysis may be required for days to weeks, and renal recovery may sometimes be incomplete.

Other bacterial infections that may cause a similar pathological picture include Rocky Mountain Spotted Fever (*Rickettsia rickettsii*), in which there may be an interstitial nephritis with foci of haemorrhage, and acute *Yersinia pseudotuberculosis* infection, in which an acute lymphocytic interstitial nephritis has been described in several patients. Legionnaire's disease (*Legionella pneumophila*) has been reported to be associated with renal impairment due to an interstitial nephritis, but in some instances may show a picture of acute tubular necrosis. The same is probably true of other severe pneumonias.

Mycobacteria spp. can cause a chronic granulomatous interstitial nephritis that is discussed below.

Viral infections

Hantaviruses

Hantaviruses are carried by small rodents and have been associated with a range of human syndromes that involve the kidneys with varying severity. 'Haemorrhagic fever with renal syndrome' (**HFRS**) was originally described in Eastern Asia. The usual renal syndrome is of oliguric renal failure, associated histologically with lymphocytic interstitial nephritis that may be haemorrhagic in severe cases, reflecting a systemic bleeding diathesis. Some patients have been reported to have persistent renal impairment after recovery.

HFRS was originally associated with Hantaan strains of hantavirus in Korea, while milder disease, usually with less severe or frequent renal impairment and without haemorrhagic diathesis, was associated with the Seoul strain. The milder disease ('nephropathia epidemica') recognized in Northern Europe, and subsequently more widely, was associated with Puumula strain. However, it has become apparent that there are many more subtypes of hantavirus, and that the association of a serotype with a particular clinical picture is not rigid. Severe disease with shock, variable haemorrhage, and (sometimes) pulmonary impairment has been encountered in patients in the Balkans and Greece. Disease with predominantly pulmonary manifestations and shock has been recognized particularly in North America, although these geographical variations in the clinical picture are no more rigid than the strain variations.

Ribavirin was shown to be effective in treating patients with HFRS in China, but in a smaller trial in North American patients with the pulmonary syndrome it was found to be no better than placebo.

Cytomegalovirus and polyomavirus

Cytomegalovirus (**CMV**) may lie dormant in renal tubular cells, and during new or reactivated infection cause characteristic inclusion bodies. This rarely has significant impact on renal function outside the setting of renal transplantation, where CMV infection commonly occurs concurrently with acute rejection. Although there is evidence that CMV infection may precipitate rejection, it is also clear that the risk of CMV infection is greatly increased by most types of antirejection therapy. CMV may also rarely cause a florid glomerular lesion characterized by gross endothelial-cell damage and swelling, resembling pre-eclampsia. This has again been recognized almost exclusively in renal transplant patients, where some believe that the appearances are due to, or complicated by, vascular rejection.

Human polyomaviruses (BK and JC) were previously believed to be benign passenger viruses that replicated without causing damage during immunosuppression. However, BK virus was recently recognized as a cause of impaired renal transplant function, generally months after transplantation. The histological changes of tubulitis had usually suggested acute cellular rejection, leading to further immunosuppression and favouring further infective damage, but inclusion bodies and immunohistochemical or *in situ* hybridization studies provided evidence for active virus replication—one report described an improvement of renal function in several patients after the reduction of immunosuppressive agents. Similar manifestations have not yet been described in other immunosuppressive settings.

Other viruses

A wide range of other viruses and micro-organisms have been less regularly associated with interstitial lesions. HIV may cause an interstitial nephritis and is considered separately below. Another condition that is likely to be infective in origin, Kawasaki disease (see Chapter 18.10.8), is associated with interstitial nephritis, although glomerular lesions have also been described occasionally.

HIV and renal disease

Renal impairment is commonly encountered at some stage of HIV infection, the largest single cause of serious renal disease being the distinct entity of HIV nephropathy. However, this generalization is misleading since this specific diagnosis is largely restricted to Black patients, and there are many other causes of renal disease in patients with HIV infection.

Focal segmental glomerulosclerosis associated with HIV infection (HIV nephropathy)

HIV nephropathy is characterized by heavy proteinuria and renal impairment. It has become the third most frequent cause of endstage renal failure (**ESRF**) in Black adults of working age in the United States. Although it has often been described as the initial manifestation of HIV infection, detailed analyses suggest that even in these cases the infection is advanced and CD4 counts are usually low. The histological appearances are of focal segmental glomerulosclerosis (**FSGS**) of the 'collapsing' form, with injury and hypertrophy of glomerular epithelial cells accompanied by variable interstitial inflammation with oedema and microtubular dilatation. It typically progresses to ESRF very rapidly, over weeks to months. Perhaps because of its association with low CD4 counts, the medium-term prognosis is usually poor despite renal replacement therapy. Highly active antiretroviral therapy is likely to have improved the outlook, and there are isolated reports of responses of nephropathy to such treatment, but at the time of writing no clear picture has emerged. Renal function in some patients has been reported to improve dramatically in response to treatment with corticosteroids, but this is not predictable and carries significant risk. It has been suggested that any improvement may be due to responses of interstitial nephritis rather than the glomerular lesion, as evidenced by the lack of reduction in proteinuria in one study.

Non-FSGS nephropathy in HIV infection

The large proportion of patients with HIV infection and nephropathy of other causes (a majority in most populations) can rarely be reliably distinguished by clinical criteria. Renal biopsies in this group have shown a very wide range of diagnoses encompassing almost all types of glomerular lesion, interstitial nephritis, cryoglobulinaemia, and thrombotic microangiopathy. Some of these lesions may be related to concurrent infections with other micro-organisms. Others may be related to therapy. Aciclovir and indinivar have replaced sulphonamides as common causes of crystal nephropathy. Idiosyncratic reactions to drugs may be more frequent in patients with HIV infection, but they also receive many drugs with predictable nephrotoxicity. The occurrence of autoimmune phenomena in patients with HIV infection may also be accompanied by an increase in immune-mediated primary glomerular diseases.

Nephropathy associated with hepatitis B virus

Chronic infection with hepatitis B virus (**HBV**) is strongly associated with membranous nephropathy and is an important secondary cause of the lesion, with a frequency that depends on the population. A less clear relationship holds with membranoproliferative nephropathy, while for hepatitis C virus (**HCV**) the converse is true.

Chronic HBV infection is much more common in some regions and racial groups, and the distribution of HBV-related nephropathy closely follows this distribution. The clinical picture may be complicated by the concurrence of HBV infection with infection by HCV or other organisms, or by coincidental significant renal and hepatic disease. HBV membranous nephropathy has a close relationship with virus multiplication, so affected individuals are **HBeAg**- and **HBsAg**-positive (hepatitis B e antigen, hepatitis B surface antigen, respectively) and hepatitis usually coexists, although it may be minor and subclinical. Membranous nephropathy is a more common complication of HBV infection in children, but it is also more benign in this group. The lesion may be static or in some cases (particularly in adults) associated with progressive deterioration to ESRF.

The histopathological appearances are typical of membranous nephropathy (see Section 20.8.6), and HBV antigens may be detectable in glomerular deposits. Whether this is relevant to pathogenesis is debatable. Animal models suggest that antibodies to podocyte surface molecules are a more likely way of producing the membranous lesion than trapping of preformed antigen–antibody complexes.

Seroconversion from HBeAg-positive to HBeAb-positive status is associated with remission of the renal lesion, whether the conversion occurs naturally or is induced by treatment. Spontaneous remission of the renal lesion is more likely in children. Antiviral treatment is the appropriate therapy when required, as immunosuppression may increase the viral burden. Unfortunately this is least likely to be successful in those populations in which the problem is greatest.

HBV infection has been associated with classic polyarteritis nodosa (**PAN**) in some populations, such as in France and North America, but even in these areas HBV–PAN is uncommon and apparently decreasing in incidence. Furthermore, the association of the two diseases is rare in some countries with both low (for example, the United Kingdom) and high (for example, Thailand) rates of HBV carriage. Usually the infection has been acquired within months of the onset of arteritic manifestations. Clinically the disease is typical of PAN, affecting medium and somewhat smaller vessels but not capillaries, and therefore not usually associated with focal necrotizing or crescentic nephritis. ANCA are not usually detected. Treatment is difficult as immunosuppression favours viral replication and exacerbation of liver disease, while remission is associated with seroconversion from HBeAg- to HBeAb-positivity.

Nephropathy associated with hepatitis C virus

Chronic hepatitis C virus (**HCV**) infection is the major cause of mixed essential (type II) cryoglobulinaemia in most populations. The mechanism is unknown, but the cryoglobulinaemia associated with HCV infection is entirely typical (see Chapter 20.10.5). The clinical picture includes cutaneous vasculitis, glomerular pathology

(membranoproliferative glomerulonephritis), and other manifestations. The cryoglobulins contain quantities of HCV antigens and bound antibody, in addition to monoclonal IgM rheumatoid factors.

HCV may also be associated with mesangioproliferative glomerulonephritis in the absence of detectable cryoglobulins. A relationship with membranous nephropathy has been suggested but is not proven.

As for HBV, reduction of viral replication has been associated with disease remission, but this is harder to achieve for HCV than for HBV with current therapies. Immunosuppression with corticosteroids and sometimes other agents may be required to control disease manifestations caused by vasculitis.

Renal sequelae of other chronic infections

Amyloidosis may be a consequence of all sorts of chronic infection, but of the 'tropical' infections is most frequently associated with schistosomiasis, filariasis, or leishmaniasis.

Mycobacteria

Mycobacterial infections cause a chronic granulomatous interstitial nephritis that is characteristically associated with inflammatory and fibrotic abnormalities in the ureters and lower urinary tract. Symptoms often relate to this lower tract involvement; however, the disease may be asymptomatic and in the earliest stages involvement is presumed to be restricted to the kidneys, with subsequent spread to the lower tract. Sterile pyuria is the rule. Impaired renal function is common at presentation. Intravenous urography will show blunting of the calyces, progressing to changes typical of pyelonephritis or papillary necrosis, along with lower tract abnormalities such as ureteric strictures and scarring and contraction of the bladder. Amyloidosis is a well-recognized secondary complication of mycobacterial infections. Idiosyncratic reactions to antituberculous drugs are the other common cause of late renal dysfunction.

Syphilis

Congenital syphilis may cause severe nephrotic syndrome with the histological pattern of membranous nephropathy. This is also the usual pattern in the rare instances when secondary syphilis causes a nephrotic syndrome. Both respond to antispirochaetal treatment.

Malaria

Plasmodium falciparum infections may cause acute renal disease, but chronic lesions are usually associated with chronic *P. malariae* infection, as described in Chapter 20.7.10.

Schistosomiasis

In renal/urological practice, schistosomiasis is best recognized for causing disease of the lower urinary tract, but chronic infections associated with hepatosplenomegaly may be associated many years later with glomerular disease. In *Schistosoma haematobium* infection this is often due to secondary infections with *Salmonella* species rather than directly associated with schistosomal infection. In *S. mansoni* infection the usual relationship is directly causal, producing a mesangiocapillary or mesangioproliferative picture.

Filariasis

Long-standing filariasis may also be associated with glomerular lesions. An acute syndrome with tubulointerstitial nephritis has also been described in association with the presence of microfilariae in renal capillaries.

Further reading

Bonarek H, *et al.* (1999). Reversal of c-ANCA positive mesangiocapillary glomerulonephritis after removal of an infected cysto-atrial shunt. *Nephrology, Dialysis, Transplantation* **14**, 1771–3. [An example of the association of c-ANCA and proteinase-3 antibodies with infection. Other examples are listed in the discussion.]

Conlon PJ, *et al.* (1998). Predictors of prognosis and risk of acute renal failure in bacterial endocarditis. *Clinical Nephrology* **49**, 96–101.

Daghestani L, Pomeroy C (1999). Renal manifestations of hepatitis C infection. *American Journal of Medicine* **106**, 347–54.

Gee WM, (1993). Causes of death in a hospitalized geriatric population: an autopsy study of 3000 patients. *Virchows Archives A* **423**, 343–9. [Pyelonephritis was commonly clinically unsuspected in this large series of unselected autopsies.]

Guillevin L, *et al.* (1995). Polyarteritis nodosa related to hepatitis B virus. A prospective study with long-term observation of 41 patients. *Medicine (Baltimore)* **74**, 238–53.

Haffner D, *et al.* (1997). The clinical spectrum of shunt nephritis. *Nephrology, Dialysis, Transplantation* **12**, 1143–8. [A report of a condition now rarely seen.]

Johnson RJ, Couser WG (1990). Hepatitis B infection and renal disease: clinical, immunopathogenetic and therapeutic considerations. *Kidney International* **37**, 663–76.

Johnston RJ, *et al.* (1994). Renal manifestations of hepatitis C virus infection. *Kidney International* **46**, 1255–63.

Jones JM, Davison AM (1986). Persistent infection as a cause of renal disease in patients submitted to renal biopsy: a report from the Glomerulonephritis Registry of the United Kingdom MRC. *Quarterly Journal of Medicine* **58**, 123–32. [A rare review of biopsy diagnoses from many centres.]

Klotman PE (1999). HIV-associated nephropathy. *Kidney International* **3**, 1161–76. [A very good general review.]

Koo JW, *et al.* (1996). Acute renal failure associated with *Yersinia pseudotuberculosis* infection in children. *Pediatric Nephrology* **10**, 582–6.

Majumdar A, *et al.* (2000). Renal pathological findings in infective endocarditis. *Nephrology, Dialysis, Transplantation* **15**, 1782–7. [Forty-two of the 62 kidneys studied here were autopsy samples, and of these, half showed localized infarction, whereas this was not identified in biopsies from living patients. Focal changes and crescent formation were the most common glomerular lesions in this group with severe renal and other disease.]

Montseny JJ, *et al.* (1995). The current spectrum of infectious glomerulonephritis: experience with 76 patients and review of the literature. *Medicine (Baltimore)* **74**, 63–73. [Little detail of individual cases but a revealing survey of causes and frequencies in a modern hospital environment.]

Neugarten J, Baldwin DS (1984). Glomerulonephritis in bacterial endocarditis. *American Journal of Medicine* **77**, 297–304. [Although some years old, it describes the current situation.]

Peters CJ, Simpson GL, Levy H (1999). Spectrum of hantavirus infection: hemorrhagic fever with renal syndrome and hantavirus pulmonary syndrome. *Annual Review of Medicine* **50**, 531–45.

Randhawa PS, *et al.* (1999). Human polyoma virus-associated interstitial nephritis in the allograft kidney. *Transplantation* **67**, 103–9.

20.7.9 Malignancy-associated renal disease

A. Neil Turner

Malignant disease may affect the kidneys and urinary tract by five broad mechanisms (Table 1).

Direct involvement of the urinary tract

Solitary kidney tumours in adults are usually caused by renal-cell carcinoma (hypernephroma). Bilateral tumours may occur, but multicentric tumours should lead to the suspicion of an inherited disorder—the least rare of these are von Hippel–Lindau syndrome (see Chapter 20.11; cystic and solid lesions, some malignant), or tuberous sclerosis (see Chapter 20.11; benign lesions), both having an autosomal dominant mode of inheritance. Lymphoma and leukaemia may occasionally invade the renal substance on a sufficient scale to cause renal impairment, but it is rare for other umours to do so.

A rare and aggressive renal medullary tumour has recently been described in young Blacks with sickle-cell trait or disease: these are easily confused with tumours of the collecting system; all reported cases have been rapidly fatal.

The collecting system and lower urinary tract may be affected by transitional-cell tumours or by invasive malignancies. Transitional-cell tumours affecting the bladder are common, and sometimes cause urinary obstruction if they are extensive or if they block one or both ureters or bladder outflow. Lesions in the ureters and collecting system are less common. They occur multifocally in association with analgesic nephropathy and Balkan nephropathy (see Chapter 20.9.2). Multifocal premalignant or malignant changes have also been described in patients with 'Chinese herb nephropathy', an epidemic of renal interstitial fibrosis associated with the ingestion of a herbal slimming aid in Europe during the early 1990s.

Metabolic effects

Hypercalcaemia is a feature of many malignancies, both with and without metastasis. Its renal effects are discussed in Chapter 20.10.5. Hypokalaemia may be a consequence of acute leukaemias or rectal tumours, and occasionally may be severe enough to cause renal dysfunction (see Chapter 20.2.2).

Hyperuricaemia and tumour lysis syndrome

Severe hyperuricaemia (>900 µmol/l) is characteristically associated with the occurrence of massive cell death following chemotherapy for haematological or solid tumours (tumour lysis syndrome), when it is usually accompanied by hyperphosphataemia and often hypocalcaemia. High serum lactate dehydrogenase (**LDH**) levels may also be diagnostically useful. Similar gross hyperuricaemia may also be seen following radiotherapy of radiosensitive tumours; sometimes this also occurs in untreated patients who have haematological or other malignancies with a very high rate of cell turnover. Levels this high can lead to precipitation within renal tubules and acute renal failure. Allopurinol should therefore be given prophylactically before commencing any chemotherapeutic regimen where such a response might occur.

Allopurinol therapy is still appropriate once hyperuricaemia has developed, and maintenance of high urinary output, and possibly urinary alkalinization, should theoretically be beneficial. If oliguric renal failure has developed then prolonged haemodialysis treatment may remove enough urate to permit renal recovery. Urate oxidase (uricase) has also been used in these circumstances to convert urate to the more soluble compound allantoin, but the enzyme is not widely available.

Remote effects of malignant tumours on the kidney

Thrombotic microangiopathy

Thrombotic microangiopathy occurring in association with malignant disease (also known as malignancy-associated thrombotic thrombocytopenic purpura, see Section 22). is often attributed to chemotherapy: although this is particularly associated with some agents (for example, bleomycin, mitomycin), isolated reports do implicate other drugs. However, in some instances the classic presentation with thrombocytopenia, microangiopathic haemolytic anaemia, and renal failure occurs in association with primary tumours. This has been particularly reported for malignancies of the stomach, pancreas, and prostate when occasionally it is the presenting sign, but it more often occurs in the setting of a known tumour.

In the absence of specific evidence, tumour-related thrombotic microangiopathy is usually treated in the same way as thrombotic microangiopathy of other types, by plasma exchange with fresh-frozen plasma. Microangiopathy generally subsides if the tumour is responsive to treatment. Renal function may be recoverable if the process is halted rapidly, an outcome most likely to be achieved in patients with prostatic carcinoma.

Other tumour products

The protean effects of the monoclonal overproduction of immunoglobulins, or their component parts, are considered elsewhere in this section. The tubulotoxic effects of light chains may be amplified by hypercalcaemia in patients with myeloma, or by the concurrent administration of other nephrotoxins—notably intravenous contrast media. Although amyloidosis was reported in older series as a consequence of lymphomas, it is now very rarely encountered as a consequence of malignancy.

Immune reactions

Malignant disease is common, so cancer will be associated with nephropathy by chance on occasion. There are therefore many case reports in the literature, but some associations have been reported consistently and are beyond doubt. The best evidence for a linkage between malignancies and intrinsic renal diseases is in minimal-change disease and membranous nephropathy (see Chapter 20.7.4). There is also substantial evidence for an association of malignancy with various types of vasculitis.

Some malignancies are particularly likely to be associated with renal disease. Chronic lymphocytic leukaemia and similar low-grade, B-cell tumours are associated with a variety of types of glomerulopathy. Thymomas have frequently been associated with glomerular lesions, usually causing nephrotic syndrome exhibiting a variety of histological patterns. There is little evidence, by contrast, for a common association of malignancies with primary interstitial renal diseases.

Minimal-change nephrotic syndrome

Lymphomas, usually Hodgkin's disease, are rarely associated with minimal-change nephropathy; this may be the presenting sign of the lymphoma, and it may also herald relapses. More so than with other renal lesions that are putatively associated with malignancy, it has often been possible to show a close temporal relationship between the occurrence of nephrotic syndrome and the presentation of the tumour. However, there is no way of proving the association in an individual patient, or of

suspecting an underlying lymphoma in patients who present with nephrotic syndrome without systemic symptoms. As the association is very rare in comparison to the number of young patients with minimal-change disease, screening other than by clinical examination and simple investigations does not seem justified. The renal lesion is typical in its pathological characteristics, and usually also in response to corticosteroid treatment.

Less commonly, minimal-change disease has been associated with solid tumours, and particularly with malignant and benign thymomas.

Membranous nephropathy

Membranous nephropathy is sometimes associated with malignancy, especially in the elderly. Series have reported rates of malignancy ranging from 5 to 11 per cent, with the risk being greatest in those at the upper end of the age range. However, different inclusion criteria have sometimes been used to assess risk: for example, some series have included tumours recognized long after a diagnosis of renal disease has been made, when the association may be coincidental. Most reported tumours have been of solid organs but haematological malignancies are also implicated. Very often the disease is advanced and obvious by the time that nephrotic syndrome or a heavy proteinuria is recognized. In some cases, effective treatment of the malignancy has led to an improvement in the nephrotic syndrome or proteinuria. The use of alkylating agents or corticosteroids to treat membranous nephropathy is not recommended in this setting, unless it would be appropriate for treatment of the malignancy itself.

In patients presenting with membranous nephropathy, controversy surrounds the value of screening for malignancy when this is not apparent from initial investigations. However, palpation of the breasts, faecal occult blood testing, and rectal examination should not be neglected. Routine haematological and biochemical investigations are appropriate for all patients, as is chest radiography and renal ultrasound. In older patients there should be a low threshold for investigating gastrointestinal or other symptoms/signs: for example, with upper gastrointestinal endoscopy, sigmoidoscopy/colonoscopy, or mammography. However, in clinical practice the number of treatable and otherwise subclinical tumours uncovered in this way is low, hence an exhaustive series of investigations is not indicated in the absence of clear and specific clinical indications.

Systemic vasculitis

Focal necrotizing and crescentic nephritis (rapidly progressive glomerulonephritis, RPGN), with or without evidence of small-vessel vasculitis affecting other organs, may occur in association with malignancy. Small-vessel vasculitis is more common in the elderly, so that some of the associations will be chance associations. However, there are sufficient reports of unusual associations to strongly suggest a causal relationship in some cases.

As well as true vasculitis, cancer-related thrombotic microangiopathy and thrombotic events complicating disseminated intravascular coagulation in association with cancer may resemble systemic vasculitis and lead to diagnostic confusion. Recent evidence suggests that thrombotic thrombocytopenic purpura itself may be an autoimmune condition caused by autoantibodies to von Willebrand factor protease, but it is not yet clear whether this association also applies to cases associated with malignancy.

The commonest type of vasculitis to be associated with malignancy is small-vessel cutaneous vasculitis. In other cases, the presence of a small-to-medium vessel systemic vasculitis, not usually associated with autoantibodies to neutrophil granule proteins (**ANCA**, antineutrophil cytoplasmic antibody), has been reported in the bowel and other organs, including the kidney. More typical ANCA-associated vasculitis has also been associated with malignancy, and there may be a particular relationship between Wegener's granulomatosis and renal-cell carcinoma. Usually the kidney is not involved in cancer-associated systemic vasculitis, but when it is the appearances are indistinguishable from those of small-vessel vasculitides of other aetiologies. Immune deposits are not usually found in the glomeruli (pauci-immune). Atrial myomas have been associated with lesions of larger and smaller vessels, and it appears that embolization is not always the explanation for this.

Effects of treatment

These include the tumour-lysis syndrome (discussed above), as well as idiosyncratic or predictable reactions to therapeutic agents. On occasions, minimal-change disease or other lesions have been associated with interferon therapy.

Further reading

Biava CG, *et al.* (1984). Crescentic glomerulonephritis associated with nonrenal malignancies. *American Journal of Nephrology* **4**, 208–14.

Bursteinß DM, Korbet SM, Schwartz MM (1993). Membranous glomerulonephritis and malignancy. *American Journal of Kidney Diseases* **22**, 5–10.

Cosyns JP, *et al.* (1999). Urothelial lesions in Chinese-herb nephropathy. *American Journal of Kidney Diseases* **33**, 1011–17.

Dabbs DJ, *et al.* (1986). Glomerular lesions in lymphomas and leukemias. *American Journal of Medicine* **80**, 63–70.

Gordon LI, *et al.* (1999). Thrombotic microangiopathy manifesting as thrombotic thrombocytopenic purpura/hemolytic uremic syndrome in the cancer patient. *Seminars in Thrombosis and Hemostasis* **25**, 217–21.

Kaplan BS, Klassen J, Gault MH (1976). Glomerular injury in patients with neoplasia. *Annual Review of Medicine* **27**, 117–25.

Kurzrock R, Cohen PR, Markowitz A (1994). Clinical manifestations of vasculitis in patients with solid tumors. A case report and review of the literature. *Archives of Internal Medicine* **154**, 334–40.

Lesesne JB, *et al.* (1989). Cancer-associated hemolytic-uremic syndrome: analysis of 85 cases from a national registry. *Journal of Clinical Oncology* **7**, 781–9.

Ronco PM (1999). Paraneoplastic glomerulopathies: new insights into an old entity. *Kidney International* **56**, 355–77. [Excellent review covering most glomerulopathies, particularly membranous and minimal-change disease, and those associated with haematological malignancy.]

Tatsis E, *et al.* (1999). Wegener's granulomatosis associated with renal cell carcinoma. *Arthritis and Rheumatism* **42**, 751–6. [A retrospective survey of 477 patients showed a specifically increased risk of renal-cell carcinoma, which in five of the seven patients occurred simultaneously with Wegener's granulomatosis.]

Valli G, *et al*. (1998). Glomerulonephritis associated with myasthenia gravis. *American Journal of Kidney Disease* **31**, 350–5. [Of three patients, two had thymomas. Ten previous case reports, including one previous series of three patients (Scadding 1983), are reviewed.]

Warren KE, Gidvani-Diaz V, Duval-Arnould B (1999). Renal medullary carcinoma in an adolescent with sickle cell trait. *Pediatrics* **103**, E22.

20.7.10 Glomerular disease in the tropics

Kirpal S. Chugh and Vivekanand Jha

Introduction

Glomerulonephritis continues to be the commonest cause of endstage renal failure in tropical countries. Reliable statistics on the various types of glomerular diseases encountered in different geographical regions are not available: the published data is mostly based on individual experiences, and suggests a significant difference in the epidemiology, aetiology, and natural history of glomerulonephritis between populations living in countries with tropical and temperate climates. The most important factor that appears to account for this difference is the high prevalence of infection-related glomerulonephritis in tropical regions. Furthermore, the impact of glomerulonephritis on the individual is often more severe: a lower degree of protein loss leads to more severe peripheral oedema and serous effusions in a malnourished individual, and the condition often remains undiagnosed and untreated for long periods due to the lack of uniform access to healthcare, culminating in a higher morbidity and mortality. Because of enormous variation in the prevalence of various endemic infections in different countries, the pattern of glomerulonephritis is different throughout the tropical region.

The overall prevalence of glomerulonephritis appears to be 60 to 100 times higher in tropical countries than in temperate regions. Hospital-based surveys from South Africa, Zimbabwe, Senegal, Uganda, Nigeria, Yemen, and Papua New Guinea show that nephrotic syndrome accounts for between 0.2 and 4 per cent of all hospital admissions. Primary glomerular diseases account for the majority of cases, but secondary causes are responsible for nephrotic syndrome in 40 to 55 per cent of patients in some countries like Zimbabwe and Jamaica.

Primary glomerular diseases

The relative frequencies of the various primary glomerulonephritides in the tropical countries and the rest of the world are shown in Fig. 1 and Fig. 2. Minimal-change disease is as prevalent in Asian countries as in the developed world, but is seen less frequently in Africa. In a study from South Africa, minimal-change disease was responsible for nephrotic syndrome in 75 per cent of children of Indian ancestry, whereas only 13.5 per cent of Black children showed this lesion. The frequency of membranous nephropathy is high in countries with a high hepatitis B carrier rate. A variant of membranous nephropathy associated with hypocomplementaemia has been described in Senegal.

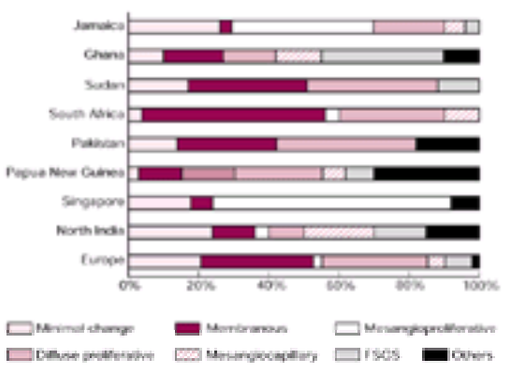

Fig. 1 Prevalence of primary glomerular disease in adults with nephrotic syndrome.

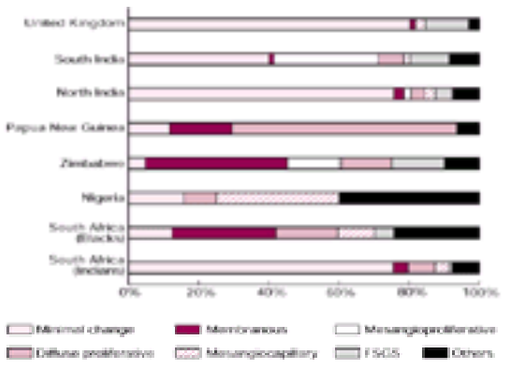

Fig. 2 Prevalence of primary glomerular diseases in nephrotic children.

IgA nephropathy (incorporated under the heading mesangioproliferative nephropathy in Fig. 1 and Fig. 2) has emerged as the commonest primary glomerular disease in many parts of the world, including Far Eastern tropical countries where it accounts for between 30 and 50 per cent of all cases of primary glomerulonephritis. By contrast, its frequency is estimated to be 5 to 15 per cent in countries of the Indian subcontinent and South America, and less than 1 per cent amongst the Black population in Africa. However, the reported differences in prevalence may be partly related to the lack of facilities for performing immunofluorescence studies in many centres in the tropical regions. High dietary fibre intake has been suggested to protect Black people from IgA nephropathy.

Postinfectious glomerulonephritis, due to both streptococcal and non-streptococcal organisms, continues to be encountered in a significant proportion of patients in tropical countries.

Glomerular diseases specific to the tropics

In addition to the well-recognized bacterial and viral infections with worldwide distribution that can cause glomerular disease (see Chapter 20.7.8), several other pathogens have been recognized to be associated with glomerular lesions in the kidneys in tropical countries, leading to a significantly different profile of infection-related glomerular disease. Causal relationships are initially suggested by epidemiological studies and then established by demonstrating the resolution of renal lesions following treatment of the infection. More recently, improved diagnostic techniques and well-designed experimental studies have provided more concrete evidence in favour of a cause-and-effect relationship. Direct confirmatory evidence has been obtained by the demonstration of specific antigens (using monoclonal antibodies), immunoglobulins, and complement within the lesions, pointing to the immunological origin of glomerular disease. The infections that cause glomerular

lesions specifically in the tropics are listed in Table 1.

Malarial nephropathy

Malaria, caused by members of the protozoan *Plasmodium* genus, is endemic in the Indian subcontinent, Middle and Far Eastern Asia, sub-Saharan Africa, and Central America where the hot and humid tropical climate is conducive to multiplication of the disease vector, the Anopheles mosquito. Of the four species that are pathogenic in humans, *P. vivax*, *P. ovale*, *P. malariae*, and *P. falciparum*, only the latter two are associated with clinically significant renal disease. Acute renal failure is the chief complication in falciparum malaria infection. Glomerulonephritis is observed with *P. malariae* infection (quartan malarial nephropathy) and less commonly with *P. falciparum*.

Quartan malaria

Although the first record of quartan malarial nephropathy in the medical literature dates back to 1884, a definite cause-and-effect relationship between *P. malariae* and the nephrotic syndrome was established in 1930 by Giglioli in the surveys carried out in British Guyana. A number of reports subsequently supported his observations, most notable being that by Gilles and Hendrickse who recorded an increased prevalence of *P. malariae* parasitaemia amongst nephrotic children in western Nigeria. The condition has also been reported from Uganda, Kenya, Ivory Coast, Sumatra, New Guinea, and Yemen.

The exact incidence of glomerulonephritis associated with quartan malaria is unknown, but the frequency of nephrotic syndrome in areas endemic for this infection is between 20 and 60 times higher than in non-endemic areas. The prevalence has shown a consistent decline with the eradication of malaria in many areas, most notably in Uganda.

Clinical features

Quartan malarial nephropathy is rare during the first 2 years of life, with the peak incidence being between 5 and 8 years of age. The prevalence declines thereafter, although sporadic cases are reported in adult life. Most patients are poor and malnourished. The most frequent presentation is with a nephrotic syndrome developing several weeks after the onset of quartan fever. Oedema is prominent and accentuated by concomitant protein energy malnutrition. Proteinuria is non-selective in 80 per cent of cases. Gross haematuria is not seen, but microscopic haematuria is noted in some cases. The blood pressure is normal at the onset of disease, but increases once renal failure sets in. Anaemia is a universal feature and enlargement of the liver and/or spleen is noted in over 75 per cent cases. Hypoalbuminaemia is usually profound, with values commonly under 1 g/dl. By contrast with other causes of nephrotic syndrome, the serum cholesterol level tends to be normal or low, reflecting low dietary intake. Serum creatinine is usually normal at presentation. Serum complement (C3) is within the normal range. In the early stages, *P. malariae* parasitaemia is detected in about 75 per cent cases.

Pathology

The predominant light microscopic abnormality is segmental thickening of glomerular capillary walls. This is focal in the early stages, but the number of affected glomeruli increases as the disease progresses. The thickening is due to the subendothelial deposition of Periodic acid–Schiff (**PAS**)- and silver stain-positive fibrils arranged in a plexiform manner. Laying down of new basement membrane material on the opposite side gives rise to the classical 'double contouring'. Eventually the capillary lumina are obliterated and mesangial sclerosis extends to involve all components. Proliferative lesions, including mesangial hypercellularity and small fibroepithelial crescents, have been described in adults.

Immunofluorescence shows deposits of IgG (usually subclass 3), IgM, and C3. Three patterns have been described. The commonest is a coarse, granular deposition along the capillary walls; a minority show diffuse, homogeneously distributed IgG2 deposits; and a mixed pattern is observed in the remaining cases. *P. malariae* antigen is detected in about one-third of patients. Electron microscopy chiefly reveals subendothelial deposits of basement membrane-like material. Intramembranous deposits may also be seen.

Pathogenesis

Demonstration of malarial antigen in the deposits and binding of specific antibody to circulating malarial antigens suggest an immunological basis for the condition. The Rhesus monkey (*Macaca mulatta*) when infected with *P. inui* develops an immune complex glomerulonephritis, and so has been proposed as an experimental model for quartan malarial nephropathy. Subendothelial location of the deposits indicates formation of immune complexes in the circulation rather than *in situ*. The permissive role of environmental factors, such as malnutrition or co-infection with Epstein–Barr virus, has been speculated to explain the development of lesions in some but not all cases of *P. malariae* infection. It has been suggested that the liver may act as a source of continuous antigen supply by harbouring the parasite.

Management

Treatment of quartan malarial nephropathy is highly unsatisfactory. Once established, the disease follows an inexorably progressive course, culminating in renal failure within 2 to 4 years. Antimalarial drugs such as chloroquine and pyrimethamine have been ineffective in controlled trials. Prednisolone is ineffective in inducing remissions and may lead to infections and worsening of hypertension, although some workers have reported a 50 per cent response rate to steroids in those with highly selective proteinuria. Remission of nephrotic syndrome has been reported occasionally with cyclophosphamide in patients with mild histological lesions, but there is no improvement in overall survival. Azathioprine is associated with increased mortality and is contraindicated in this condition.

Falciparum malaria

The incidence of glomerulonephritis associated with falciparum malaria is difficult to estimate as the disease is mild, transient, and overshadowed by other complications. Autopsy studies reveal glomerular lesions in about 18 per cent of cases, but urinary abnormalities including non-selective proteinuria, microhaematuria and casts are noted in 20 to 50 per cent. Full-blown nephrotic and acute nephritic syndromes are seen occasionally. By contrast to quartan malarial nephropathy, glomerulonephritis associated with falciparum malaria resolves within 4 to 6 weeks of eradication of infection.

Pathology

The histological lesions are characterized by mesangial hypercellularity and a modest increase in the mesangial matrix. Basement membrane changes are usually absent. An eosinophilic granular material is present in capillary walls, mesangium, and Bowman's capsule, along with pigment-laden macrophages in the capillary lumina. Immunofluorescence shows finely granular IgG3, IgM, and C3. Malarial antigen can be demonstrated in some cases. Electron microscopy reveals subendothelial and mesangial deposits.

Pathogenesis

Transient glomerular lesions akin to those in falciparum malaria can be induced in BALB/c mice and Sprague–Dawley rats infected with *P. berghei*, also in *P. falciparum*-infected squirrel monkeys, the latter model showing endocapillary proliferation in addition to mesangial hypercellularity. Recent studies have suggested that CD4 cell subpopulations, in particular TH2 cells, may be playing a significant role in the genesis of these glomerular lesions, with renal cytokine expression strongly correlated with the severity of proteinuria in C57BL/6J mice infected with *P. berghei*. The role of concomitant infection, especially with the hepatitis B virus, is under investigation.

Schistosomal nephropathy

Schistosomiasis is a chronic infection caused by trematodes (blood flukes) and affects over 300 million people in Asia, Africa, and South America. Of the seven species pathogenic to man, the most prevalent are *Shistosoma haematobium* (Africa and the Middle East), *S. mansoni* (South America and Africa) and *S. japonicum* (China and the Far East). *S. haematobium* primarily involves the lower urinary tract, whereas *S. mansoni* involves the gastrointestinal tract and portal system, leading to hepatic fibrosis and portal hypertension.

Glomerulonephritis has been described most frequently in association with hepatosplenic schistosomiasis produced by *S. mansoni*. The first reports came from autopsy series in Brazil during 1964. Several clinical observations from endemic areas of Africa, Saudi Arabia, Aden, and Yemen and experimental studies have since confirmed the cause-and-effect relationship of *S. mansoni* infection with glomerulonephritis. *S. japonicum* is known to cause glomerulonephritis only in experimental animals.

In clinical studies, overt proteinuria has been reported in 1 to 20 per cent of patients infected with *S. mansoni* and 2 to 5 per cent with *S. haematobium* infection; microalbuminuria is found in about 22 per cent of patients with hepatosplenic schistosomiasis. Histological studies have documented subclinical glomerular lesions in a much higher proportion of patients, with glomerular abnormalities detected on renal biopsy in 40 per cent of patients who underwent splenectomy in one study, none of whom had clinical evidence of renal disease, hence the true extent of subclinical glomerular involvement remains unknown.

Clinical features

Though described at all ages, glomerulonephritis is seen most commonly in young adults with overt hepatosplenic disease. Males are affected twice as frequently as females. Peripheral oedema and ascites are the hallmarks of clinical glomerular involvement. Hypertension is seen in 50 per cent of cases, appearing late in the disease. Proteinuria is poorly selective and haematuria is uncommon in the absence of lower urinary tract involvement. About 30 per cent of patients exhibit hypergammaglobulinaemia; the serum cholesterol is not elevated in an equal proportion; serum complement (C3) levels are usually low. Non-specific antibody production is demonstrated by false-positive rheumatoid factor or the **VDRL** (Venereal Disease Research Laboratory) test, especially in those with concomitant salmonella infection. Demonstrating viable eggs in the stool or egg-containing granulomas in a rectal or liver biopsy makes the diagnosis of schistosomiasis. It is important to exclude other causes of nephrotic syndrome before attributing the lesions to schistosomiasis. Some patients with schistosomiasis and proteinuria have been shown to have a coexistent salmonella infection.

Pathology

Several patterns of glomerular pathology have been described (Table 2). The Class I lesion is the earliest and most frequent, all three types of mesangioproliferative lesion being seen with equal frequency (see Table 2). It is also the principal lesion in renal allografts with recurrent schistosomal nephropathy. Class II lesions are more frequent in patients with concomitant salmonella infection. The frequency of Class III lesions varies from 20 per cent in asymptomatic patients to over 80 per cent in those with overt renal disease, over 90 per cent of Class III cases belonging to Class III A. Immunofluorescence shows IgG and C3, less commonly IgM and IgA. Schistosomal antigen is detected in a minority of cases. Electron microscopy shows subendothelial and epimembranous deposits in Class IIIA and IIIB, respectively. The Class IV lesion, seen in 15 to 40 per cent of cases, cannot be distinguished from idiopathic focal segmental glomerulosclerosis (**FSGS**) on the basis of light microscopy, but immunofluorescence reveals IgA deposition in most cases. Class III and IV lesions are seen in patients with fibrotic livers. Class V prevalence varies from 15 to 40 per cent, with a higher frequency in African patients. This form is not usually affected by hepatic fibrosis.

Pathogenesis

Schistosomal glomerulopathy is caused by the immunological reaction to specific parasitic antigens. Out of over 100 immunological constituents extracted from adult worms, Cercariae and Schistosomulae, only a few have been identified *in vivo*. The pathogenic antigens originate in the gut of the adult worm, are regurgitated into the host's bloodstream and find their way to the glomeruli. Schistosoma antigens unaccompanied by any immunoglobulin have been demonstrated in the glomeruli of Kenyan baboons infected with *S. mansoni*, suggesting a role for the kidneys in the disposal of circulating antigens. Circulating immune complexes have been documented in humans and experimental animals, with the highest titres in those with hepatosplenic disease. The complexes usually localize in the mesangial region. An additional component of *in situ* immune complex formation is suggested by the extramembranous location of deposits. Interspecies antigenic variation could be responsible for the differential expression of glomerular injury.

Portocaval shunting secondary to hepatic fibrosis is critical in the genesis of an immune reaction to such antigens. Diversion of portal blood carrying the primary load of worm antigen directly into the systemic circulation prevents their normal processing and degradation by hepatic macrophages. This hypothesis is supported by the low frequency of glomerular disease in cases of *S. haematobium* infection that do not show hepatic involvement. Baboons with *S. mansoni* infection develop neither portal fibrosis nor glomerular disease despite heavy infestation. IgA antibodies predominate in the circulation of patients with schistosomal glomerulonephritis, whereas those with hepatosplenic schistosomiasis without glomerular involvement show IgM antibodies. Impaired clearance by the liver and increased production secondary to immunoglobulin isotype 'switching' from IgM- to IgA-producing B cells are postulated to be responsible for this alteration. The number of circulating mononuclear IgA-bearing cells is also increased in these patients.

Elevated immunoglobulin levels and false-positive rheumatoid factor and VDRL tests suggest the presence of an autoimmune disorder. Antinuclear antibodies, specific against the public anti-DNA idiotype 16/6 ID, have been found in the sera and glomerular deposits of humans and experimental animals.

The role of salmonella infection in the genesis of schistosomal nephropathy is unclear. Epidemiological studies have shown that urinary abnormalities disappear following therapy for salmonella alone, suggesting that these abnormalities could be purely due to salmonella infection in some cases.

Management

Treatment of schistosomal glomerulopathy is disappointing. Antischistosomal drugs like oxamniquine, hycanthone, or praziquantel are unsuccessful in altering the clinical course, which is one of inexorable progression to renal failure. Steroids or cytotoxic agents, alone or in combination, are ineffective in inducing remission. Isolated reports of response to these agents have not been evaluated in controlled trials. Salmonella infection should be looked for and treated in all patients.

Filarial nephropathy

Filarial worms are nematodes that dwell in the subcutaneous tissues and lymphatics. These are transmitted to humans through arthropod bites. Clinical manifestations depend upon the location of microfilariae and adult worms in the tissues. Of the eight filarial species that infect humans, *Loa loa*, *Onchocerca volvulus*, *Wuchereria bacrofti*, and *Brugia malayi* are associated with glomerular disease.

Loiasis is prevalent in West and Central Africa and characteristically manifests with localized areas of allergic inflammation and calabar swellings. Onchocerciasis (river blindness) is characterized by subcutaneous nodules, a pruritic skin rash, sclerosing lymphadenitis, and ocular lesions. Bancroftian and brugia infections cause febrile episodes associated with acute lymphangitis and lymphadenitis, leading ultimately to lymphoedema manifesting as hydrocele and elephantiasis. This form of filariasis is endemic in Africa and SE Asia.

The frequency of glomerular involvement in filariasis is difficult to estimate. Urinary abnormalities have been described in 11 to 25 per cent of cases of loiasis and onchocerciasis, with a nephrotic syndrome in 3 to 5 per cent. Nephrotic syndrome is more common in those with polyarthritis and chorioretinitis, and impaired creatinine clearance is more frequent with onchocerciasis than loiasis. Regarding wuchererian and bancroftian filariasis, early studies indicated an increased incidence of proteinuria in patients infected with *B. malayi* compared to controls, but there was no correlation with the severity of filariasis. In a recent survey in an endemic area, proteinuria was detected in over 50 per cent of patients with filariasis, with 25 per cent showing glomerular proteinuria. The frequency of proteinuria, also of microhaematuria and hypertension, is significantly higher in patients with chronic sclerosing filariasis than in those with an acute febrile illness or microfilaraemia. False-positive rheumatoid factor, both of IgG and IgM type, and anti-DNA and antiphospholipid antibodies and autoantibodies against a variety of cytoplasmic proteins may be noted in some cases.

Pathology

Light microscopy reveals a gamut of lesions, including mesangial proliferative, mesangiocapillary, minimal-change disease, chronic sclerosing glomerulonephritis, and the collapsing variant of focal segmental glom-erulosclerosis. A diffuse basement membrane thickening with a mild increase in the number of endocapillary cells is the commonest finding. Mononuclear interstitial infiltration and microinfarcts around blood vessels have been demonstrated in patients with loiasis. Microfilariae may be found in the glomerular capillary lumina, tubules, and interstitium. Electron microscopy shows widely spaced subepithelial, subendothelial, and intramembranous deposits and spikes. *O. volvulus* and *B. malayi* antigens along with IgM, IgG, and C3 have been demonstrated in the deposits. Biopsies from patients with loiasis

exhibit reactivity with hyperimmune anti-Onchocerca serum, raising the possibility of a shared antigen.

Pathogenesis

Filarial glomerulonephritis appears to be immune complex-mediated. Circulating immune complex levels correlate with the adult worm burden. Dogs infected with *Dirofilaria immitis* develop glomerular lesions similar to human filariasis. Immune complexes can also form *in situ*, as suggested by one experimental study that showed the development of glomerular lesions in kidneys after selective catheterization and infusion of *D. immitis* into the renal arteries. The contralateral kidneys either remained uninvolved or showed very minor lesions. Diethylcarbamazine treatment, by killing the parasite, may lead to antigen release into the circulation, thus exacerbating the immune process. A temporal relationship between the administration of this agent and the development of proteinuria has been noted.

Treatment

A good response to antifilarial therapy with diethylcarbamazine is observed in patients with non-nephrotic proteinuria and/or haematuria. The response is inconsistent in those with nephritic syndrome, and deterioration of renal function may continue despite clearance of microfilariae with treatment.

Mycobacterial infections

Leprosy

Leprosy is a chronic granulomatous disorder caused by the acid-fast bacillus *Mycobacterium leprae*. Nephritis in patients with leprosy was recognized by Hansen and Looft in 1894, and continued to be an important cause of death until the 1950s. The two major glomerular lesions encountered in leprosy include glomerulonephritis and secondary amyloidosis.

Glomerulonephritis

The incidence of glomerulonephritis varies from under 2 per cent on clinical evaluation to over 50 per cent on histology. Interpretation of various studies is confounded by bias in patient selection, with specialized centres reporting high figures. Glomerulonephritis is seen in both lepromatous and non-lepromatous forms of leprosy and is more common during episodes of erythema nodosum leprosum.

Clinical features

Most patients present with asymptomatic urinary abnormalities but nephrotic syndrome, acute nephritic syndrome, and rapidly progressive renal failure have all been described in a small number of patients. Hypertension is uncommon. Reduced creatinine clearance is noted in patients with erythema nodosum leprosum, and impaired urinary acidification and concentration may be demonstrated. Hypocomplementaemia is common, and circulating cryoglobulins are present in many cases.

Pathology

The histological picture is varied. The most frequent light microscopic lesions are those of mesangial proliferative and diffuse proliferative glomerulonephritis, although other morphological lesions have been reported. Acid-fast bacilli are seen rarely. Electron microscopy reveals electron-dense deposits in the mesangial and subendothelial regions, focal foot-process widening, glomerular capillary basement membrane reduplication with mesangial interposition, and endothelial cytoplasmic vacuolation. Immunofluorescence reveals granular deposits of IgG and C3, and less frequently IgM, IgA, and fibrin in the mesangium and along capillary walls.

Pathogenesis

The lesions are manifestations of an immune complex process. Circulating immune complexes can be detected in about one-third of those with lepromatous disease and over 75 per cent of patients with active erythema nodosum leprosum. The antigen is thought to be derived from *M. leprae*, but there is also speculation about the role of a non-mycobacterial antigen derived from co-infecting micro-organisms or dapsone:antidapsone antibodies. Alternate-pathway complement activation by cryoprecipitates can exacerbate the glomerular injury.

Management

In general, steroids or antileprosy drugs have no effect on the course of glomerular disease. Prednisolone may hasten the recovery of renal function in patients with renal failure during episodes of erythema nodosum leprosum.

Amyloidosis

The incidence of renal amyloidosis in leprosy ranges from 2 to 55 per cent in different geographical regions. Amyloid was documented in 55 per cent cases in older autopsy and biopsy studies from the United States, but reports from Mexico, Africa, and India found the incidence to be less than 10 per cent. The amyloid is of AA type and is far more frequent in lepromatous compared to non-lepromatous leprosy: erythema nodosum leprosum further increases the risk as each episode is associated with a marked and persistent elevation of serum amyloid A protein. Patients with tuberculoid leprosy who have long-standing and infected trophic ulcers can also develop this complication.

Amyloidosis can be prevented by early and aggressive antileprosy treatment, with particular attention to preventing erythema nodosum leprosum.

Tuberculosis

Renal involvement in patients with tuberculosis takes the form of granuloma formation, interstitial nephritis, and caseous destruction. An association of glomerulonephritis with tuberculosis was postulated in the preantibiotic era, but only stray reports have described immune complex glomerulonephritis and dense-deposit disease in tuberculosis in recent times. The cause-and-effect relationship remains speculative, and a chance association cannot be excluded. A well-known complication, however, is amyloidosis, which is still seen in a significant proportion of patients in poor countries where the disease often remains untreated for long periods. Once established, the course of amyloidosis is unaffected by treatment of the underlying tuberculosis.

Other infections

Variable degrees of glomerular involvement are seen with a variety of infections encountered in the tropical countries. These include: bacterial infections such as typhoid and pneumococcal infection; viral infections including dengue haemorrhagic fever; protozoal infections such as toxoplasmosis, kala-azar, and trypanosomiasis; and parasitic infestations like trichinosis. In most cases, the glomerulonephritis is mild and transient and resolves with treatment of the primary illness. In general, the frequency of glomerular involvement is falling with the reduction in incidence of these infections.

Further reading

Abdurrahman MB, *et al.* (1990). Clinicopathological features of childhood nephrotic syndrome in northern Nigeria. *Quarterly Journal of Medicine* **75**, 563–76.

Aikawa M, *et al.* (1988). Glomerulopathy in squirrel monkeys with acute *Plasmodium falciparum* infection. *American Journal of Tropical Medicine and Hygiene* **38**, 7–14.

Barsoum RS (1993). Schistosomal glomerulopathies. *Kidney International* **44**, 1–12.

Barsoum RS (1999). Tropical parasitic nephropathies. *Nephrology Dialysis Transplantation* **14**, 79–91.

Chugh KS, Jha V (2000). Glomerulonephritis due to other bacterial, viral and parasitic infections. In: Massry SG, Glassock RJ, eds. *Textbook of nephrology*, 4th edn. Williams and Wilkins, Baltimore,

MD.

Chugh KS, Sakhuja V (1990). Glomerular diseases in the tropics. *American Journal of Nephrology* **10**, 437–50.

Chugh KS, Sakhuja V (1998). Glomerular disease in the tropics. In: Davison AM, *et al.*, eds. *Oxford textbook of clinical nephrology*, 2nd edn, pp 703–19. Oxford University Press, Oxford.

Eiam-Ong S, Sitprija V (1998). Falciparum malaria and the kidney: a model of inflammation. *American Journal Kidney Diseases* **32**, 361–75.

Gilles HM, Hendrickse RG (1963). Nephrosis in Nigerian children, role of *Plasmodium malariae*, and effect of anti malarial treatment. *British Medical Journal* **1**, 27–31.

Grauer GF, *et al.* (1989). Experimental *Dirofilaria immitis*-associated glomerulonephritis induced in part by *in situ* formation of immune complexes in the glomerular capillary wall. *Journal of Parasitology* **755**, 585–93.

Pakasa NM, Nseka NM, Nyimi LM (1997). Secondary collapsing glomerulopathy associated with Loa loa filariasis. *American Journal of Kidney Diseases* **30**, 836–9.

Sinniah R, Rui-Mei L, Kara L (1999). Up-regulation of cytokines in glomerulonephritis associated with murine malaria infection. *International Journal of Experimental Pathology* **80**, 87–95.

Sitprija V (1988). Nephropathy in falciparum malaria. *Kidney International* **34**, 867–77.

Weiner ID, Northcutt AD (1989). Leprosy and glomerulonephritis. *American Journal of Kidney Diseases* **13**, 424–9.

WHO (1988). *Renal disease, classification and atlas of infectious and tropical diseases*, Sinniah R, *et al.*, eds. ASCP Press, Chicago, IL.

20.8.1 Renal tubular disorders

J. Cunningham

Renal glycosuria

Definition and pathophysiology

Renal glycosuria occurs when there is failure of tubular mechanisms to reabsorb the entire filtered load of glucose under conditions of normoglycaemia. In health, the volume of the glomerular filtrate is approximately 180 litres/day, containing approximately 800 mmol of glucose. The average daily urinary excretion of glucose is approximately 1 mmol, implying that some 99.9 per cent of the filtered load of glucose is normally reabsorbed, mostly in the proximal tubule. The primary event at this site is the reabsorption of very large amounts of Na+. Glucose reabsorption is a two-step process in which a carrier-mediated Na+-glucose cotransporter moves both sodium and glucose passively across the brush border (apical) membrane and thereby into the proximal tubular cell. The active extrusion of sodium across the basolateral membrane of the cell is subsequently linked with diffusion of glucose facilitated by specific glucose transporters, such that glucose is transported, in a manner tightly linked to sodium transport, from the tubular lumen to the peritubular capillaries. The molecular mechanisms by which cotransported solutes such as glucose are moved across the tubular epithelium are not fully understood, although it is likely that the binding of the cotransported solute (in this case glucose) leads to conformational changes in the transport protein and the opening of a sodium gate. Sodium thus moves down a concentration gradient (from lumen to cell), with secondary active transport linking the movement of glucose to that of Na+. These disturbances can be identified functionally, either as a reduction of the tubular maximum capacity for glucose reabsorption or a reduction of the tubular threshold for glucose.

Clinical features

The isolated form of renal glycosuria is familial with a mixed inheritance pattern, suggesting that the condition results from any one of several mutations affecting the glucose transport processes described above. Isolated renal glycosuria has no clinical sequelae and is easily distinguished from diabetes mellitus by the fact that the patient is normoglycaemic. Isolated renal glycosuria is frequently seen in otherwise normal pregnancies, where it results from the increased glomerular filtration rate (**GFR**) taking the filtered glucose load to a level that exceeds the renal tubular maximum capacity for reabsorption. By contrast with the glycosuria of uncontrolled diabetes mellitus, renal glycosuria is never of sufficient magnitude to drive a clinically significant osmotic diuresis.

Abnormalities of glucose transport may be seen in association with other defects of proximal tubular transport. Collectively these are designated as Fanconi's syndrome, in which renal glycosuria is found as part of a generalized disorder of proximal tubular function with aminoaciduria, renal tubular acidosis, and phosphaturia. It is important to distinguish forms of isolated glycosuria from multiple tubular defects, including the Fanconi syndrome, in which there may be important clinical consequences, albeit not directly in relation to the glycosuria itself.

Many patients with chronic renal insufficiency of mild to moderate degree exhibit renal glycosuria, usually in combination with other disorders of tubular function. These are frequently subtle and of little or no clinical significance. They reflect tubular damage as part of the general renal parenchymal pathology, or the consequences of the high filtered solute load per nephron in patients with reduced numbers and mass of functioning nephrons.

Phosphate-handling disorders

Physiology and pathophysiology

The renal handling of inorganic phosphate is the major determinant of extracellular phosphate concentration. In health, the kidney shows a powerful adaptive capacity that is capable of maintaining a normal phosphate concentration in the face of wide fluctuations in dietary phosphate intake. Between 80 and 95 per cent of the filtered load of phosphate is normally reabsorbed, mostly in the proximal tubule, but up to 20 per cent of phosphate reabsorption occurs at more distal sites, namely the distal convoluted tubule and cortical collecting duct. The initial process is the movement of filtered phosphate into proximal tubular cells via a number of specific Na+–phosphate cotransporters that are located in the luminal membrane and have a 3Na+:1HPO_4^{2-}stoichiometry. This linkage allows the movement of sodium ions down their electrochemical gradient to drive the movement of phosphate up its electrochemical gradient even as the tubular phosphate concentration falls. In the later parts of the proximal tubule other higher affinity phosphate transporters retrieve much of the residual phosphate. The final step, namely the exit of phosphate across the basolateral membrane and into the peritubular capillaries, appears to be passive.

As indicated above, the kidney provides not only the principal means for phosphate excretion, but also acts as the principal regulator of phosphate homeostasis. This regulation takes place over a wide range: the kidneys excrete virtually all the typical dietary phosphate intake of approximately 30 to 40 mmol/day, but phosphate deprivation or hypophosphataemia leads to such effective renal phosphate conservation that renal phosphate excretion virtually ceases. These adjustments are mediated by the cotransporter activity described above.

Parathyroid hormone (**PTH**) is an important hormonal regulator of phosphate excretion, stimulating phosphaturia by acting directly on proximal tubular cells to inhibit sodium-dependent phosphate transport by mechanisms that operate through both the cAMP protein kinase-A and the protein kinase-C phosphoinositide pathways. There are receptors for PTH on both the apical and basolateral membranes of the proximal tubular cells, the functional effect of PTH being to decrease the *V*max of both the more-proximal, high-capacity, low-affinity cotransporter and the more-distal, low-capacity, high-affinity cotransporter systems, both resulting in phosphaturia.

Other hormones influencing proximal phosphate transport include growth hormone, insulin-like growth factor-1 (**IGF-1**), insulin, thyroid hormone, and 1,25-dihydroxyvitamin D, all of which augment phosphate reabsorption. By contrast, in addition to PTH itself, phosphate excretion is augmented by PTH-related peptide, calcitonin, glucocorticoids, and atrial natriuretic peptide (**ANP**).

In addition to the above, it appears that the tubular phosphate transport mechanism can respond to changes in dietary phosphate intake, even when the plasma phosphate level changes little or not at all. The nature of this dietary signal is unclear. In parallel with the antiphosphaturic effect of reduced dietary phosphate, there is an increase in bone resorption leading to mobilization of skeletal phosphate (and calcium). Both the renal and skeletal responses to low dietary phosphate are unimpaired by parathyroidectomy and are therefore not mediated by PTH.

Disorders associated with increased urinary phosphate excretion

There are various types of phosphate transport defect, but all cause hypophosphataemia with inappropriate phosphaturia (Table 1). The fractional excretion of phosphate (the percentage of filtered phosphate that appears in the final urine) is increased and the tubular transport maximum for phosphate (*TmP*/GFR) is decreased. The clinical disturbances that result from such disorders may be very severe, important ones being rickets (in children) and osteomalacia (in adults). The development of these depends on the severity and chronicity of the hypophosphataemia, also on the presence or absence of any associated non-renal abnormalities.

Hereditary hypophosphataemic rickets

The terminology is potentially confusing. Vitamin D-resistant rickets (**VDRR**) originally described a syndrome of hypophosphataemia and metabolic bone disease (rickets or osteomalacia) that in many ways resembled that of vitamin D deficiency but which did not respond to treatment with vitamin D. This condition is now more properly called hereditary hypophosphataemic rickets, a designation more consistent with the phosphate-wasting aetiology. That these patients do not respond to vitamin D is undeniable, but true vitamin D resistance (that is, resistance even to 1,25-dihydroxyvitamin D) appears to exist only in patients with functional defects (usually inherited) of the vitamin D receptor.

X-linked hypophosphataemic rickets

This is the most important type of renal phosphate-handling disorder. Presentation is generally with poor growth and rickets in early childhood. The inheritance pattern is consistently of X-linked dominant type. There is a defect in proximal tubular phosphate transport that results in persistent hypophosphataemia and inappropriate phosphaturia. Females (heterozygotes) are less severely affected than males (hemizygotes). There also appears to be a subtle disturbance of vitamin D metabolism, such that the plasma 1,25-dihydroxyvitamin D concentration does not show the increase during hypophosphataemia that is seen in otherwise normal subjects.

Understanding of the pathogenesis and molecular biology of X-linked hypophosphataemic rickets has been greatly assisted by the existence of a murine model, the *hyp* mouse. It is clear that the defect of phosphate transport has nothing to do with the normal PTH modulatory control system. Cross-circulation and kidney transplant experiments in *hyp* mice have shown that the defect can be transferred from affected to non-affected animals. This, together with the observation that cultured proximal tubular cells from *hyp* mice exhibit normal phosphate transport, points strongly to mediation by an extrarenal humoral factor termed phosphotonin. The *hyp* gene is called PHEX (Phosphate regulating gene Homologous to Endopeptidases on the X chromosome). It is expressed in bone and not in kidney, a distribution compatible with the extrarenal origin of X-linked hypophosphataemic rickets.

The diagnosis is made on the basis of characteristic clinical features coupled with persistent hypophosphataemia and a reduced *TmP*/GFR, indicating an inappropriate reduction of the tubular reasborptive capacity for phosphate.

Because the bone disease in X-linked hypophosphataemic rickets is at least partly a consequence of the hypophosphataemia, treatment attempts to normalize plasma phosphate—a difficult task in practice. The administration of oral phosphate supplements increases phosphaturia, hence large oral doses have to be taken at frequent intervals, thereby presenting a substantial compliance problem in these patients, many of whom are young children. Additionally, the administration of large doses of phosphate reduces the plasma 1,25-dihydroxyvitamin D concentration, slightly lowers the ionized calcium concentration in plasma, and thereby triggers secondary hyperparathyroidism. This in turn may compound the skeletal disease and also further increase phosphaturia. These troublesome compensations can be attenuated by the addition of calcitriol (1,25-dihydroxyvitamin D) therapy to the phosphate supplement, with substantially improved clinical outcomes. However, treatment must be monitored extremely closely, there being a constant risk of calcitriol-induced hypercalciuria, hypercalcaemia and nephrocalcinosis. If this occurs, the degree of tubular calcium phosphate deposition appears to be largely determined by the oral phosphate dose: the presence of large amounts of phosphate in the gut lumen prevents the normal association of luminal calcium with oxalate, thereby increasing oxalate availability and absorption. Thus the stage is set for enteric hyperoxaluria, another potent risk factor for nephrocalcinosis and stone formation (see Chapter 20.13).

Oncogenic rickets/osteomalacia

A disturbance similar to X-linked hypophosphataemic rickets is rarely acquired in association with certain mesenchymal tumours, especially giant-cell tumours of bone, neurofibromas, and cavernous haemangiomata. Removal of the tumour is followed by complete normalization of renal phosphate handling, an observation that strongly favours the involvement of a tumour-generated humoral factor in the pathogenesis of the hypophosphataemia. Extracts of these tumours have been found to inhibit phosphate transport in renal tubular cells and to initiate phosphaturia in experimental animals. The humoral factor has been named 'phosphatonin' and it is likely that 'phosphatonin' is the same in X-linked hypophosphataemia and in oncogenic rickets/osteomalacia. It affects only phosphate transport and has no direct effects on calcium metabolism, PTH, or the PTH receptor.

Other phosphate-wasting disorders

Autosomal recessive and autosomal dominant phosphaturic disorders have rarely been reported. Common to these disorders is impaired proximal tubular phosphate transport (low *TmP*/GFR) with inappropriate phosphate wasting and variable metabolic bone disease. The clinical presentations are variable: some present during adolescence or even adulthood, while others present in early life, with some cases resolving spontaneously at puberty.

Syndromes of hereditary hypophosphataemia with hypercalciuria are described below in the section on calcium-handling disorders.

Disorders associated with reduced urinary phosphate excretion

Excessive tubular phosphate reabsorption (high *T*mP/GFR) with resulting hyperphosphataemia is seen in conditions where PTH is lacking or there is renal resistance to PTH (Table 1). In these patients hyperphosphataemia coexists with hypocalcaemia. The hyperphosphataemia is the result of an inappropriately raised *T*mP/GFR. The hypocalcaemia is largely the result of the failure of adequate 1,25-dihydroxyvitamin D (calcitriol) production by the PTH-deprived kidney, arising because the renal 25-hydroxyvitamin D la-hydroxylase is downregulated in the absence of PTH or its receptor.

Hypoparathyroidism

In hypoparathyroidism the renal tubular response to PTH is normal when tested by the administration of exogenous PTH. The metabolic abnormalities merely reflect the lack of PTH. The diagnosis depends on a low or undetectable PTH concentration in plasma, despite a prevailing hypocalcaemia that would normally trigger secondary hyperparathyroidism. For a detailed discussion of hypoparathyroid disorders, see Chapter 12.4.

Pseudohypoparathyroidism

There are two main types of pseudohypoparathyroidism, renal resistance to PTH being a feature of both. In these disorders the resulting hypocalcaemia evokes an appropriate PTH response. Type 1 pseudohypoparathyroidism is associated with a G-protein defect, with failure of coupling between the PTH receptor itself and adenylate cyclase. As a result, PTH (whether endogenous or exogenous) evokes neither a urinary cAMP nor a phosphaturic response. Despite plasma PTH being elevated, the *T*mP/GFR is high with associated phosphate retention. In the type 2 variant, G-protein activity is normal and PTH induces a cAMP response but no phosphaturia, implying a defect of cAMP-dependent protein kinase C.

Clinically, hypocalcaemia and hyperphosphataemia dominate the metabolic picture. There are also somatic features comprising short stature, short fourth and fifth metacarpals, and a variable degree of mental deficiency.

All types of hypoparathyroidism can be treated effectively using oral calcitriol or alfacalcidol. Pharmacological doses of these agents are needed to bring calcium into the normal range: it is frequently helpful to incorporate a calcium supplement into each meal, principally to reduce the intestinal absorption of dietary phosphate in these hyperphosphataemic individuals.

Calcium-handling disorders

Disorders of urinary calcium output (hypercalciuria and, to a lesser extent, hypocalciuria) are quite common and have important sequelae, particularly in regard to

hypercalciuric renal stone disease (see Chapter 20.13). Whilst it is important to recognize that such abnormalities of urinary calcium excretion may reflect intrinsic abnormalities of calcium handling within the kidney, it is also the case that primary disorders remote from the kidneys can also be responsible for disturbances of the calcium excretion rate. For example, hypercalcaemia—as seen in vitamin D intoxication, excessive dietary calcium intake, and osteolytic metastases—is associated with marked hypercalciuria, albeit in circumstances where the kidneys are responding appropriately to the high plasma calcium concentration. Many factors influence the handling of calcium by the renal tubule (Table 2).

Physiology and pathophysiology

In health, between 200 and 250 mmol of calcium appear in the glomerular filtrate each day, assuming an ultrafiltrable blood calcium concentration of 1.3 mmol/l out of the total blood calcium concentration of 2.5 mmol/l. About 70 per cent of filtered calcium is reabsorbed in the proximal convoluted tubule and the rest in the thick ascending limb of Henle's loop (20 per cent), the distal convoluted tubule (5 to 10 per cent), and the collecting tubule (less than 5 per cent). These reabsorptive processes reclaim nearly all the filtered calcium, such that only about 3 to 5 mmol of calcium appears in the urine each day. Assuming constancy of the total-body calcium content, this urinary calcium loss is equal to the net intestinal calcium absorption.

Calcium reabsorption in the proximal tubule and in the loop of Henle occurs passively down the electrochemical gradient generated by sodium and water reasbsorption at these sites. Regulation of calcium transport occurs at more distal sites where both parathyroid hormone and 1,25-dihydroxyvitamin D augment calcium reabsorption, the former by activation of adenylate cyclase via the PTH receptor and the latter by upregulation of calcium binding proteins – calbindins. Thus, parathyroid hormone itself, and also parathyroid hormone-related peptide (**PTHrp**), acts in an anticalciuric fashion, thereby contributing to the hypercalcaemia of primary hyperparathyroidism and the humoral hypercalcaemia of malignancy, respectively.

The extracellular calcium-sensing receptor (CaR)

It is now clear that the extracellular calcium-sensing receptor plays a central role in regulating the renal handling of calcium. It does this by both indirect and direct mechanisms. Indirectly, the CaR in the parathyroid gland senses extracellular calcium and adjusts the output of parathyroid hormone appropriately. This in turn regulates the renal handling of calcium in the distal nephron, with a fall of plasma calcium concentration triggering PTH secretion and thereby renal calcium retention—an appropriate response. Directly, the CaR in renal epithelial cells (most heavily expressed in the cortical thick ascending limb and also present in the proximal tubule, medullary thick ascending limb of Henle's loop, the distal convoluted tubule, and the collecting duct) regulates the handling of both calcium and water. Binding of calcium to the CaR in Henle's loop is thought to diminish the reabsorption of calcium (and magnesium). Thus, conditions of high calcium delivery to the loop of Henle lead to an appropriate increase in calciuria. In addition, it appears that the CaR in the kidney provides a link between distal tubular calcium delivery and the rate of ADH-stimulated water reabsorption at that site. This mechanism would explain the observed nephrogenic diabetes insipidus that accompanies significant hypercalcaemia and which reduces the likelihood of urinary supersaturation of calcium in circumstances of hypercalciuria.

Action of diuretics

Loop diuretics (furosemide (frusemide) and bumetanide), thiazide diuretics, and amiloride all affect renal tubular calcium transport and, as experimental probes, have been extremely useful in the elucidation of the mechanisms of renal calcium handling, as well as in the therapy of hypercalcaemic and hypercalciuric disorders. Loop diuretics inhibit sodium chloride reabsorption in the thick ascending limb of the loop of Henle and with it the passive reabsorption of various cations, including calcium. The resulting increase in the calcium excretion rate can be beneficial (as in the treatment of hypercalcaemia), or deleterious (increased risk of osteopenia or stone formation in chronic drug-induced hypercalciuria). Conversely, thiazide diuretics substantially reduce the urinary calcium excretion rate. Two mechanisms appear to underlie this effect. First, the mild volume depletion arising from the natriuretic and diuretic actions of the thiazide serves to accelerate proximal sodium and water reabsorption, and with it passive calcium reabsorption in this part of the nephron. Second, thiazides appear to increase distal calcium reabsorption directly, although the mechanism is unclear. These actions of thiazide diuretics are extremely useful in the treatment of hypercalciuric stone disease. In addition, there is evidence that thiazides can reduce the negative calcium balance in elderly people, which may translate to a reduction in the incidence of fractures in this age group. Thiazide diuretics slightly elevate the plasma calcium concentration, although usually to a trivial extent only, but this may become clinically significant in patients who have a tendency to hypercalcaemia, such as those with very mild hyperparathyroidism, Paget's disease, or who are immobilized. Amiloride also exerts a hypocalciuric action, probably at the cortical connecting segment, although the precise mechanism is uncertain. The hypocalciuric effect of thiazides and amiloride are, therefore, exerted at different sites of the renal tubule and in clinical practice are additive. Thus combinations of thiazides and amiloride are useful treatments for patients with hypercalciuric stone disease (see Chapter 20.13).

Disorders associated with increased urinary calcium excretion

Idiopathic hypercalciuria

This is an extremely important metabolic disturbance because of the high associated risk of calcium stone formation. 'Idiopathic' in this context implies hypercalciuria without hypercalcaemia and in the absence of other factors known to accelerate bone resorption (for instance, acromegaly, hyperthyroidism, hyperparathyroidism, osteolytic metastases, immobilization, metabolic acidosis) or reduced tubular calcium reabsorption (such as loop diuretics, chronic metabolic acidosis).

In most cases, the hypercalciuria is driven by calcium hyperabsorption by the intestine, and the hypercalciuria is thus of the 'overspill' type. This usually results from an initial defect of renal phosphate handling (reduction of *T*mP/GFR and consequent renal phosphate leak), which stimulates the production of 1,25-dihydroxyvitamin D. In a minority of cases the primary defect is of renal tubular calcium reabsorption ('renal leak' hypercalciuria), with a secondary increase of parathyroid hormone, calcitriol, and intestinal calcium absorption.

The assessment and management of these patients has benefited greatly from an increased understanding of the underlying defects. For example, 'absorptive' hypercalciuria is logically managed initially by measures to reduce intestinal calcium absorption, namely avoidance of excessive dietary calcium intake and, in those with evidence of 1,25-dihydroxyvitamin D excess driven by hypophosphataemia, oral phosphate therapy as well. Thiazides and amiloride are added if these initial measures are inadequate. Conversely, the management of 'renal leak' hypercalciuria requires an increase in tubular calcium reabsorption by reducing the dietary sodium and protein intake (acid load) and giving thiazide diuretics and amiloride.

Hereditary hypercalciuric nephrolithiasis

Recent studies have identified four rare disorders, all characterized by low molecular weight proteinuria, hypercalciuria, nephrocalcinosis, renal stone formation, and (in many cases) renal failure. In some, there are also defects of proximal tubular function with aminoaciduria, phosphaturia, renal glycosuria, and uricosuria (Fanconi's syndrome), as well as impairment of urinary acidification (renal tubular acidosis). These four disorders are Dent's disease, X-linked recessive nephrolithiasis, X-linked recessive hypophosphataemic rickets, and idiopathic low molecular weight proteinuria in Japanese children. The underlying defect appears to be the result of a mutation of a chloride-channel gene (*CLCN5*), the functional loss of which results in a widespread defect of proximal tubular transport.

Disorders associated with reduced urinary calcium excretion

Most patients exhibiting hypocalciuria do so in association with hypocalcaemia and a reduced filtered load of calcium. This is seen in patients with secondary hyperparathyroidism as a response to an underlying vitamin D and/or calcium deficiency, when a reduced filtered load of calcium is combined with accelerated tubular calcium reabsorption driven by increased levels of PTH. Most patients with advanced renal failure also exhibit hypocalciuria.

Familial hypocalciuric hypercalcaemia

These conditions reflect functional aberrations of the extracellular calcium-sensing receptor (**CaR**). In familial hypocalciuric hypercalcaemia (**FHH**, also known as familial benign hypercalcaemia) there is an inactivating mutation in the calcium-sensing receptor gene. Several different mutations are known, most appearing to result in receptors that are truncated or have an abnormal amino acid sequence. These render the receptor less sensitive to calcium which, at the level of the parathyroid gland, makes the parathyroid attempt to set calcium at a supraphysiological concentration. At the level of the kidney, the defect leads to increased tubular calcium and magnesium reabsorption. The resulting metabolic disturbance is characterized by hypercalcaemia, hypocalciuria, and hypermagnesaemia. The PTH hormone concentration is within the 'normal range', but this is inappropriately elevated with regard to the serum calcium concentration. The inheritance pattern is

autosomal dominant with high penetrance.

Most patients tolerate the hypercalcaemia well and the characteristic symptoms of hypercalcaemia (polyuria, constipation, neuropsychiatric disturbance) are conspicuously absent. The disorder may be distinguished from primary hyperparathyroidism by the presence of a family history, the reduction in urinary calcium excretion rate, and the normal urinary excretion rate of cyclic AMP (increased in conditions of parathyroid hormone excess). It is important to distinguish these patients from those with mild primary hyperparathyroidism: the benign natural history and the poor response to subtotal parathyroidectomy means that parathyroid surgery should not be undertaken in these individuals.

Autosomal dominant hypocalcaemia

By contrast to familial hypocalciuric hypercalcaemia, autosomal dominant hypocalcaemia results from an activating mutation in the calcium-sensing receptor that renders it overly sensitive to extracellular calcium. This leads to a downward setting of the normal parathyroid hormone and calcium relationship with resulting hypocalcaemia. Because the mutation also affects the calcium receptor in the renal tubular cells, urinary calcium excretion is inappropriately high. Principal clinical sequelae are the result of the hypercalciuria that predisposes to stone formation, nephrocalcinosis, and renal insufficiency. The hypocalcaemia is generally well tolerated. Treatment with thiazide–amiloride combinations to reduce hypercalciuria is logical but not of established benefit. Hypocalcaemia is generally asymptomatic and treatment of this with vitamin D or calcitriol is generally inappropriate, serving only to increase the degree of hypercalciuria and the risk of nephrocalcinosis and stone formation.

The Fanconi syndrome and aminoaciduria

Physiology and pathophysiology

Filtered amino acids are subject to very rapid proximal tubular reabsorption, at least 95 per cent having been cleared from the glomerular filtrate by the time it reaches the end of the proximal tubule. The transport of organic solute at this site is a two-step process that is both carrier-mediated and sodium-coupled; amino acids enter the proximal tubular cell via the brush border membrane against an electrochemical gradient, and exit via another transporter at the basolateral membrane. Movement in this fashion is accomplished by secondary active mechanisms, whereby coupling to sodium transport allows the movement of amino acids to be driven indirectly by the basolateral membrane Na+/K+-ATPase.

Based on loose structural similarities, the amino acids segregate into four groups, each with a group-specific carrier system (Table 3). Common to many defects of amino acid transport is a reduction of the electrochemical sodium gradient across the proximal tubular cells, leading to the impaired linked transport of glucose, amino acids, phosphate, and a range of other solutes. Some of the amino acid transporters are also expressed in the intestine; in which case defects may be evident at both sites, and clinical disease can result from the intestinal defect, the renal defect, or both.

The Fanconi syndrome

This syndrome comprises a disturbance of proximal tubular functions with generalized aminoaciduria, phosphate wasting (hypophosphataemic rickets and osteomalacia), renal tubular acidosis type-2 (proximal **RTA**), and renal glycosuria. The terms 'juvenile-' and 'adult Fanconi's syndrome' are widely used, but refer only to the age of onset and serve no additional classification purpose. More helpful is to classify, as far as possible, the many causes of the Fanconi syndrome on the basis of aetiology and pathogenesis (Table 4).

Clinical presentations of the Fanconi syndrome usually depend more on the associated underlying abnormality than on the renal tubular defect *per se*. However, the diagnosis ultimately depends on the demonstration of characteristic multiple tubular defects. These may not all be present in all patients and may even fluctuate in an individual patient, hence it is often best to define the specific defects that are present rather than to use the catch-all Fanconi eponym.

Treatment focuses on two issues. First, the cause of the Fanconi syndrome: for example, fructose avoidance in hereditary fructose intolerance, galactose avoidance in galactosaemia, copper chelation therapy in Wilson's disease. Second, the consequences of the Fanconi syndrome: for example, alkali and potassium for RTA type-2, oral phosphate and calcitriol for phosphate wasting.

Specific aminoacidurias

These are classified according to four principal carrier defects (see Table 3).

Neutral aminoacidurias

Hartnup disease

This rare (1:16 000 births), autosomal recessive disorder comprises three features:

1. intestinal tryptophan malabsorption;
2. a pellagra-like syndrome with photosensitive skin lesions, ataxia, and neuropsychiatric disturbances; and
3. neutral aminoaciduria with increased renal clearance of alanine, asparagine, glutamine, histidine, isoleucine, leucine, phenylalanine, serine, threonine, tyrosine, valine, and tryptophan.

The clinical manifestations of Hartnup disease result from the tryptophan malabsorption that leads to nutritional deficiency, which is exacerbated by the accelerated urinary losses of tryptophan. It presents much like pellagra, although is usually less severe and tends to fluctuate in its course. Analysis of the urine distinguishes the two disorders. Hartnup disease responds well to oral nicotinamide therapy (40–200 mg daily).

Dibasic aminoacidurias

These comprise cystinuria, lysinuric protein intolerance, and lysinuria, of which cystinuria is the most common and the most important.

Cystinuria

The group-specific carrier protein for the dibasic amino acids is located on the brush border membrane of the proximal tubular cells and is thought to be the product of a single pair of allelic genes. So far, three potential mutant alleles have been identified that appear to be capable of causing both homozygous and heterozygous forms of cystinuria. When expressed, this transport defect is found in both the kidney and in intestinal epithelium.

Cystinuria occurs in 1 in 7000 births and has serious clinical manifestations. Inheritance is autosomal recessive. Presentation is usually during childhood or adolescence, the syndrome of nephrolithiasis presenting with pain, infection, and, in some cases, renal impairment and hypertension. The stones are radio-opaque (although less so than calcium-containing stones), smooth, and sometimes staghorn-shaped. The diagnosis is confirmed by a positive nitroprusside test, the presence of typical hexagonal crystals in morning urine specimens, and the quantitative measurement of urinary cystine output.

Treatment requires reduction of the cystine concentration in the urine combined with measures to increase its solubility. Typical regimens comprise a high fluid intake, alkalinization of the urine to over pH 7.5 (usually with large quantities of potassium citrate and sodium bicarbonate), and penicillamine. Cystine solubility changes little across the acidic range of pH but increases rapidly above pH 7. At 37 °C and pH 7, the solubility is only 1.66 mmol/l, but this increases to between about 3.3 and 3.5 mmol/l at pH 7.8. However, pushing the pH to even more alkaline levels may be counterproductive since alkalinization decreases the solubility of calcium phosphate, which may be deposited on the cystine stones. Poor compliance is a frequent and unsurprising practical problem with this demanding regimen, particularly in regard to maintenance of a high fluid intake. To be fully effective this requires oral fluids to be taken at least once during the night. Nevertheless, at least 50 per cent of patients respond to these measures if they are rigorously applied and adhered to, and in some cases the stones regress significantly.

Sulphydryl-containing drugs, such as penicillamine, react with cystine to form penicillamine–cysteine, which is much more soluble. In the past this treatment was often reserved for those who failed on the fluid/alkali regimen. However, penicillamine is now used much earlier and often forms part of the initial therapy. Although potentially toxic (cutaneous reactions, marrow suppression, and glomerulopathy), serious reactions are rare, and penicillamine is currently the most effective therapy known. It is given at doses of 1 to 2 g daily, the aim being to reduce the free-cystine concentration in urine to below 1.66 mmol/l, when stone formation is prevented and existing stones can be dissolved.

In cystinuria other basic amino acids (lysine, arginine, and ornithine) are also present in increased amounts in the urine, but only cystine—by virtue of its low solubility—is of clinical importance.

Lysinuric protein intolerance

This is a rare autosomal recessive disorder that results from widespread defects of dibasic amino acid transport, involving particularly the intestine, proximal renal tubule, and liver. Cystine transport is normal. The renal tubular defect plays no part in the pathogenesis of the disease.

Mental retardation, growth failure, and osteopenia are prominent, and are thought to result from reduced activity of the urea cycle with a low plasma urea concentration and hyperammonaemia after food. Treatment with citrulline is sometimes effective, probably by regenerating the deficient urea cycle intermediates, arginine and ornithine.

Imino acids and glycine

Familial iminoglycinuria

This is a relatively common condition, arising in approximately 1 in 15 000 births. Clinically, the inheritance appears to be autosomal recessive, but there are multiple alleles and gene loci for the transport of imino acids and glycine. Proximal tubular transport of proline, hydroxyproline, and glycine is impaired, accompanied in some, but not all, cases with defects in intestinal transport.

Few if any clinical sequelae result from isolated imminoglycinuria, although it was once thought that the abnormality was associated with mental retardation and seizures.

Acidic aminoaciduria

These disturbances are exceedingly rare, involving the dicarboxylic amino acids (aspartic acid and glutamic acid). They are not well understood. The possibility of accelerated renal production and/or failure to transfer these amino acids into the renal circulation is suggested by the observation of renal clearance in excess of the GFR. There are no clinical sequelae.

Further reading

Baron DN, *et al.* (1956). Hereditary pellagra-like skin rash with temporary cerebellar ataxia, constant renal amino-aciduria and other bizarre biochemical features. *Lancet* **ii**, 421–33.

Bergeron M, *et al.* (2001). The renal Fanconi syndrome. In: Scriver CR, *et al*., eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 5023–38. McGraw-Hill, New York.

Brown EM (2000). Familial hypocalciuric hypercalcemia and other disorders with resistance to extracellular calcium. *Endocrinology and Metabolism Clinics of North America* **29**, 503–22.

Brown ME, *et al*. (1995). Calcium-ion-sensing cell-surface receptors. *New England Journal of Medicine* **333**, 234–40

Calcium-sensing receptor. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 601199.

Chesney RW (2001). Iminoglycinuria. In: Scriver CR, *et al*., eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 4971–82. McGraw-Hill, New York.

Coe FL, Parks JH, Moore ES (1979). Familial idiopathic hypercalciuria. *New England Journal of Medicine* **300**, 337–40.

Cystinuria. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Numbers 220100 (type 1) and 600918 (types 2 and 3).

De Marchi S, *et al*. (1984). Close genetic linkage between HLA and renal glycosuria. *American Journal of Nephrology* **4**, 280–6.

Dent's disease, X-linked recessive nephrolithiasis, X-linked recessive hypophosphataemic rickets, and idiopathic low molecular weight proteinuria of Japanese children. All due to mutations of chloride channel 5; CLCN5. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 300008.

Dicarboxylicaminoaciduria. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 222730.

Familial hypocalciuric hypercalcaemia. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 241530

Familial idiopathic hypercalciuria (hypercalciuria, absorptive, type 1). http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 143870

Fitch N (1982). Albright's hereditary osteodystrophy: a review. *American Journal of Medical Genetics* **11**, 11–29.

Gregory MJ, Schwartz GJ (1998). Diagnosis and treatment of renal tubular disorders. *Seminars in Nephrology* **18**, 317.

Hartnup disease. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 234500.

Iminoglycinuria. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 242600.

Kanai Y, *et al*. (1994). The human kidney low affinity Na(+)/glucose cotransporter SGLT2: delineation of the major renal reabsorptive mechanism for D-glucose. *Journal of Clinical Investigation* **93**, 397–404.

Kumar R (2000). Tumor-induced osteomalacia and the regulation of phosphate homeostasis. *Bone* **27**, 333–8.

Levy HL (2001). Hartnup disorder. In: Scriver CR, *et al*., eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 4957–70. McGraw-Hill, New York.

Lloyd SE, *et al*. (1996). A common molecular basis for three inherited kidney stone diseases. *Nature* **370**, 445–9.

Lysinuric protein intolerance. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 222700.

Marx SJ (2000). Hyperparathyroid and hypoparathyroid disorders. *New England Journal of Medicine* **343**, 1863–75.

Nesbitt T, *et al*. (1992). Crosstransplantation of kidneys in normal and Hyp mice: evidence that the Hyp mouse phenotype is unrelated to an intrinsic renal defect. *Journal of Clinical Investigation* **89**, 1453–9.

Palacin M, *et al*. (2001). Cystinuria. In: Scriver CR, *et al*., eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 4909–32. McGraw-Hill, New York.

Pearce SHS, *et al*. (1996). A familial syndrome of hypocalcemia with hypercalciuria due to mutations in the calcium-sensing receptor. *New England Journal of Medicine* **335**, 1115–22.

Pseudohypoparathyroidism type IA. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 103580.

Pseudohypoparathyroidism type IB. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 603233.

Renal glycosuria. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 233100.

Rosenberg LE, Durant JL, Elsas LJ (1968). Familial iminoglycinuria: an inborn error of renal tubular transport. *New England Journal of Medicine* **278**, 1407–13.

Rosenberg LE, *et al.* (1966). Cystinuria: biochemical evidence for three genetically distinct diseases. *Journal of Clinical Investigation* **45**, 365–71.

Rowe PS (2000). The molecular background to hypophosphataemic rickets. *Archives of Disease in Childhood* **83**, 192–4.

Simell O (2001). Lysinuric protein intolerance and other cationic aminoacidurias. In: Scriver CR, *et al*., eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 4933–56. McGraw-Hill, New York.

Strewler GJ (2000). The parathyroid hormone-related protein. *Endocrinology and Metabolism Clinics of North America* **29**, 629–45.

Tenenhouse HS, Econs MJ (2001). Mendelian hypophosphatemias. In: Scriver CR, *et al*., eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 5039–68. McGraw-Hill, New York.

Thakker RV (2000). Pathogenesis of Dent's disease and related syndromes of X-linked nephrolithiasis. *Kidney International* **57**, 787–93.

Tieder M, *et al*. (1985). Hereditary hypophosphatemic rickets with hypercalciuria. *New England Journal of Medicine* **312**, 611–17.

Verge CF, *et al*. (1991). Effects of therapy in X-linked hypophosphatemic rickets. *New England Journal of Medicine* **325**, 1843–8.

Vitamin D resistant rickets with end-organ unresponsiveness to 1,25-dihydoxycholecalciferol. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 277440.

Wright EM, Martin MG, Turk E (2001). Familial glucose–galactose malabsorption and hereditary renal glycosuria. In: Scriver CR, *et al*., eds. *The metabolic and molecular basis of inherited disease*, 8th edn, pp 4891–908. McGraw-Hill, New York.

X-linked hypophosphataemic rickets. http://www3.ncbi.nlm.nih.gov/omim/ OMIM Number 307800.

20.9.1 Acute interstitial nephritis

Dominique Droz and Dominique Chauveau

Introduction

Acute interstitial nephritis has a clinicopathological definition: acute renal failure with prominent inflammation of the renal interstitium, composed mainly of lymphocytes and more rarely of polymorphonuclear cells or granulomas. Since a variable degree of tubular-cell damage is consistently found, the term 'tubulointerstitial nephritis' is preferred by some authors. Because the clinical presentation of acute renal failure is devoid of specific findings, renal biopsy is required in all cases both to establish the diagnosis of acute interstitial nephritis and to allow appropriate management.

Epidemiology and incidence

Although its precise incidence is difficult to determine, acute interstitial nephritis (**AIN**) is a rare disease, found in 1 to 3 per cent of specimens in unselected series of renal biopsy material. In patients presenting with acute renal failure, the proportion with acute interstitial nephritis varies from 6.5 to 15 per cent. Values at the lower end of this range are found in paediatric series. Higher values are found in studies containing larger numbers of elderly patients, where the increased likelihood of drug ingestion is paralleled by an increase in drug-induced hypersensitivity. Overall, acute interstitial nephritis is the third leading cause of acute drug-induced nephropathy, following haemodynamically mediated and direct tubular injuries.

Pathophysiology of acute interstitial nephritis (AIN)

The pathogenesis of AIN is not completely elucidated, and what we know has been extrapolated from animal models. These include spontaneous T cell-mediated interstitial nephritis in kd/kd mice, interstitial nephritis related to antitubular basement antibodies, interstitial nephritis associated with autoimmune systemic disease, and tubulointerstitial damage associated with glomerular diseases and severe proteinuria. Although in rodents these models induce tubulointerstitial damage, very few reproduce the human picture of AIN.

Experimental evidence implicates both humoral and cell-mediated mechanisms, and both antigen-specific and antigen-nonspecific pathways of T-cell activation, but the relevant antigen responsible for T-cell activation is unknown in the vast majority of cases of human AIN. However, in drug-related acute interstitial nephritis the offending drug or one of its metabolic products could function as a hapten, binding to membrane or cellular proteins and being presented in the context of MHC molecules to T-helper cells. Sensitized activated T-helper cells could then produce cytokines, activate macrophages, induce the differentiation and proliferation of B cells producing specific antibodies, and activate other T cells, either cytotoxic cells or those responsible for the delayed-type hypersensitivity reaction. As a consequence, this cascade of reactions could ultimately result in a variable mixture of specific antibody production, death of the target tubular cells, and development of parenchymal granulomas. This cascade-type response could be modified by a variety of mechanisms such as removal of the antigen (for example, the offending drug) or production of inhibitory cytokines.

Aetiology

The causes of acute interstitial nephritis fall into four main categories: (1) drug hypersensitivity reactions; (2) infections; (3) systemic immune-mediated diseases; and (4) idiopathic (Table 1). A drug hypersensitivity reaction is the most common cause of acute interstitial nephritis, accounting for 40 to 60 per cent of the cases, while infections—of bacterial or viral origin—account for less than 5 per cent of cases today. Table 1 provides a detailed list of drugs and infectious agents repeatedly associated with AIN. Although more than 100 drugs have been implicated in acute interstitial nephritis, only a few are repeatedly incriminated, while most reports remain anecdotal. In the presence of acute interstitial nephritis, this should not distract from considering any drug currently or recently consumed by the patient as a potential culprit.

Clinical and pathological features

Most clinical features are non-specific, but the history may point to AIN when renal failure develops in the context of a systemic infection, typical drug reaction, sarcoidosis, Sjögren's syndrome, or uveitis. Blood pressure remains unchanged. Urinary output is variable with mild or moderate proteinuria (<1–2 g/day). Nephrotic syndrome is only found in cases induced by non-steroidal anti-inflammatory drugs. There is no haematuria except in cases related to b-lactam hypersensitivity. Leucocyte casts are common. Fractional excretion of sodium is often high.

Eosinophilia and eosinophiluria are not always present, but argue in favour of an adverse drug reaction. Eosinophils can be detected in the urine using Wright or Hansel stains, the latter being more sensitive. However, eosinophiluria is not specific for drug-induced acute interstitial nephritis and is therefore of poor predictive value.

Kidney size is typically normal or enlarged. Increased cortical echogenicity shown by ultrasound imaging correlates with the degree of interstitial inflammation. Computed tomography (**CT**) scanning is of limited value in the diagnosis of acute interstitial nephritis and other diffuse renal parenchymal diseases.

Renal biopsy remains the sole means of unequivocally establishing the diagnosis of acute interstitial nephritis. The characteristic lesion is the presence of numerous mononuclear cells in the renal interstitium (Fig. 1 and Plate 1). Tubular changes of focal cell necrosis and tubulitis, together with interstitial oedema, are commonly found. Since a discrete interstitial cell infiltrate may be present in primary acute tubular necrosis (either of ischaemic or toxic origin), the assessment of primary acute interstitial nephritis requires careful evaluation of the abundance of the cell infiltrate in comparison with the degree of tubular damage. In addition, the biopsy must be scrutinized for the presence of glomerular pathology, significant immune complex deposition, and vasculitic lesions. If these coexist with a prominent interstitial-cell infiltrate, then diagnoses such as lupus glomerulonephritis, mixed cryoglobulinaemia, and systemic vasculitis need to be considered. Lymphomas or leukaemias may invade the renal parenchyma and cause acute renal failure. The tumoral (clonal) nature of the infiltrating cells is usually obvious, but needs to be distinguished from acute interstitial nephritis.

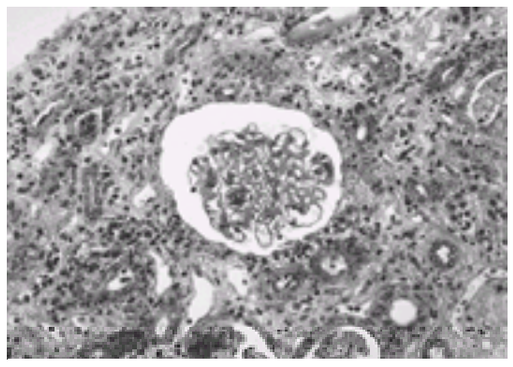

Fig. 1 Acute interstitial nephritis. The renal interstitium is invaded by numerous mononuclear cells. The glomerulus is normal. Mason's trichrome 250 ×. (See also Plate 1.)

In acute interstitial nephritis, the degree of interstitial inflammation is variable and predominates in the cortex. Whatever the cause, T lymphocytes (with an equal proportion of CD4+ and CD8+ cells) and macrophages (CD14+ and CD68+) expressing cell-activation markers, comprise 80 per cent of the infiltrate. Natural killer cells are rare, while polyclonal plasma cells are often observed. Polymorphonuclear cells constitute only a minor part of the cell infiltrate, even in the early phases of the disease, and in cases related to bacterial infection. Eosinophils are rare, even in drug-induced acute interstitial nephritis. Tubular cells express MHC class II antigens. Of note, some interstitial epithelioid and non-caseating granulomas are present in nearly 20 per cent of cases, but they are more common (25 to 45 per cent of cases) in drug-induced acute interstitial nephritis.

In the vast majority of cases (90 per cent), staining for immunofluorescence shows no significant immunoglobulin deposits in the renal parenchyma. However, in rare cases mainly related to systemic lupus erythematosus, granular deposits of immunoglobulin and complement are observed along the tubular basement and in the interstitium. In less than 5 per cent of the cases, linear IgG and C3 deposits are present along the tubular basement membrane, corresponding to the presence of antitubular basement membrane antibodies. Such findings are observed only in cases related to antibiotic use (b-lactams or ciprofloxacin).

Little is known about the long-term histopathological consequences of AIN: most patients recover and serial biopsies have rarely been performed. However, in severe cases, especially those with severe tubule destruction, irreversible lesions with fibrous scars ultimately develop. Late chronic interstitial nephritis is also more frequent when the initial biopsy reveals granulomas, but overall the most reliable histological parameter for prognosis is the extent of interstitial fibrosis on the initial biopsy. The mechanism of fibrogenesis is thought to be that interstitial infiltrating cells and damaged tubular cells release cytokines and growth factors, promoting fibroblast and myofibroblast proliferation and the production of collagen.

Distinctive features of particular causes of acute interstitial nephritis

According to the cause, clinical and laboratory features characterize certain forms of AIN.

Drug-induced acute interstitial nephritis

The most common drugs implicated in acute interstitial nephritis are b-lactam antibiotics and non-steroidal anti-inflammatory drugs (Table 1), but their clinical presentation differs. The *b*-lactams (and especially methicillin, which is no longer used) give the most characteristic picture of drug-related acute interstitial nephritis. The symptoms occur 2 to 60 days after the beginning of treatment and comprise fever, skin rash (usually maculopapular), arthralgias, liver involvement, abundant haematuria (often macroscopic), blood eosinophilia, and a variable degree of renal failure, requiring dialysis in one-third of the cases. Skin tests or *in vitro* evaluation of hypersensitivity reaction are not valuable for implicating a given drug.

Non-steroidal anti-inflammatory drug-associated acute interstitial nephritis usually affects the elderly, is particularly likely to occur when the drug is taken discontinuously, and evolves with a more progressive course. Renal failure develops several months to years after the initiation of the offending therapy; extrarenal signs of drug sensitization are often lacking. Abundant proteinuria and nephrotic syndrome are found in more than 80 per cent of patients (versus 1 per cent in *b*-lactam-induced acute interstitial nephritis). In addition to the interstitial lesions, renal biopsy shows minimal glomerular changes with diffuse podocyte foot-process fusion.

How can the physician recognize a drug as being responsible for the onset of acute interstitial nephritis? Given the variability of clinical course, as exemplified for *b*-lactams and non-steroidal anti-inflammatory drugs, the clinician should first establish an exhaustive list of the drugs used by the patient, including over-the-counter medications, and the date of exposure. This is often a difficult task. In a very few cases, re-challenge with the same drug has mistakenly been performed. When such exposure is elicited and associated with recurrence of the renal and systemic manifestations, then the drug can definitely be regarded as the culprit, but in clinical practice re-challenge should never be recommended. In decreasing order of importance, clues for considering a drug as being responsible for AIN include: (1) appropriate timing, (2) well-documented knowledge of similar nephrotoxicity, and (3) exclusion of other causes (see below).

Infectious acute interstitial nephritis

In children, infections remain the main cause of acute interstitial nephritis. In adults, predisposing factors include old age, diabetes, cytotoxic drug administration, and prolonged corticosteroid therapy. Septicaemia due to various micro-organisms such as *Escherichia coli*, *Proteus* spp., *Staphylococcus* spp., and *Candida albicans* can be associated with direct invasion of the renal parenchyma and acute interstitial nephritis. The clinical picture is that of acute pyelonephritis with a biopsy revealing renal microabscesses.

Haemorrhagic fevers due to Hantaviruses are recognized in Europe with increasing frequency. In these cases, the renal biopsy shows interstitial oedema and medullary haemorrhage, whilst interstitial inflammation remains discrete.

An acute interstitial nephritis with an infiltrate predominantly of CD8+ lymphocytes may be seen in patients with human immunodeficiency viral (HIV) disease, with or without glomerular involvement. Enlargement of the liver and salivary glands with hypergammaglobulinaemia and lymphocytic infiltration of these organs is often found, as well as lymphocytic pneumonitis.

Acute interstitial nephritis in systemic immune-mediated diseases

Lupus erythematosus and Sjögren's syndrome are rare causes of AIN. By contrast, sarcoidosis can present with acute renal failure, with typical sarcoid granulomas found in the renal interstitium. Concomitant hypercalcaemia and/or extrarenal manifestation make the diagnosis of sarcoidosis easy in most cases, but renal involvement can be isolated. In such instances, CT-scanning of the lung may disclose asymptomatic chest involvement. It is conceivable that a localized form of sarcoidosis might be restricted to isolated granulomatous interstitial nephritis, but it is impossible to distinguish between this possibility and that of 'idiopathic' acute interstitial nephritis.

Idiopathic forms of acute interstitial nephritis

Since its description in 1975, an increasing number of cases of acute interstitial nephritis associated with anterior uveitis (or iritis) have been reported. Such an association—referred to as tubulointerstitial nephritis and uveitis, or the **TINU** syndrome—affects mainly young women and adolescent girls. It is characterized by fever, weight loss, blood eosinophilia, hypergammaglobulinaemia, and renal failure. Uveitis may precede the appearance of renal failure by several weeks. An identical clinical picture lacking uveitis has also been described. In the TINU syndrome the biopsy shows diffuse interstitial inflammation, but some granulomas may be present. Whether this association is a manifestation of a limited form of sarcoidosis is still unclear.

Treatment

Treatment of AIN should be first directed against its cause: withdrawal of any drug that might be involved, or prompt treatment of infection. Prednisone improves the renal failure of sarcoidosis and the TINU syndrome, although persistent dysfunction is not uncommon when diagnosis is delayed. Whether or not to use steroids in drug-related AIN remains a matter of debate. Uncontrolled studies suggest that a short course of high-dose prednisone promotes an earlier and more complete decline of serum creatinine toward baseline than in patients left untreated. Some advocate its use when renal failure persists for more than 1 week after withdrawal of the drug or if granulomas are found in the renal biopsy.

In patients with drug-related AIN, the physician should inform the patient that they should not be treated with the presumed culprit or related compounds. More specifically, all b-lactam antibiotics should be avoided in patients who have suffered acute interstitial nephritis attributed to a penicillin compound or a cephalosporin, even though cross-sensitization between the two classes of drugs is not consistent. In those who have recovered from non-steroidal anti-inflammatory-related AIN, acute renal failure may or may not recur after resuming therapy with a non-steroidal anti-inflammatory drug belonging to another family of this class. In some countries, severe drug-induced side-effects, including AIN, should be reported to the health authorities.

Further reading

Buysen JG, *et al*. (1990). Acute interstitial nephritis: a clinical and morphological study in 27 patients. *Nephrology, Dialysis, Transplantation* **5**, 94–9.

Cameron JS (1988). Allergic interstitial nephritis: clinical features and pathogenesis. *Quarterly Journal of Medicine* **66**, 97–115.

Davison AM, Jones CH (1998). Acute interstitial nephritis in the elderly: a report from the UK MRC Glomerulonephritis Register and a review of the literature. *Nephrology, Dialysis, Transplantation*, **13**(Suppl. 7), 12–16.

Dobrin RS, Vernier RL, Fish AL (1975). Acute eosinophilic interstitial nephritis and renal failure with bone marrow-lymph node granulomas and anterior uveitis. A new syndrome. *American Journal of Medicine* **59**, 325–33.

Droz D, Kleinknecht D (1998). Acute interstitial nephritis. In: Davison AM, *et al.* eds. *Oxford textbook of clinical nephrology*, pp 1634–48. Oxford University Press, Oxford.

Ellis D, *et al.* (1981). Acute interstitial nephritis in children: a report of 13 cases and review of the literature. *Pediatrics* **67**, 862–70.

Ivanyi B, *et al.* (1996). Acute tubulointerstitial nephritis: phenotype of infiltrating cells and prognostic impact of tubulitis. *Virchows Archiv* **428**, 5–12.

Kleinknecht D (1995). Interstitial nephritis, the nephrotic syndrome, and chronic renal failure secondary to nonsteroidal anti-inflammatory drugs. *Seminars in Nephrology* **15**, 228–35.

McRae Dell K, Kaplan BS, Meyers CM (1999). Tubulointerstitial nephritis. In: Barratt TM, *et al.* eds. *Pediatric nephrology*, pp 823–4. Lippincott Williams & Wilkins, Baltimore, MD.

Meyers CM, Neilson EG (1995). Immunopathogenesis of tubulointerstitial disease. In: Massry SG, Glassock RJ, eds. *Massry and Glassock's textbook of nephrology*, pp 671–7. Williams & Wilkins, Baltimore, MD.

Michel DM, Kelly CJ (1998). Acute interstitial nephritis. *Journal of the American Society of Nephrology* **9**, 506–15.

Mustonen J, *et al.* (1994). Renal biopsy findings and clinicopathologic correlations in nephropathia epidemica. *Clinical Nephrology* **41**, 121–6.

Nochy D, *et al.* (1993). Renal disease associated with HIV infection: a multicentric study of 60 patients from Paris hospitals. *Nephrology, Dialysis, Transplantation* **8**, 11–19.

Okada H, *et al.* (1993). Steroid-responsive renal insufficiency due to idiopathic granulomatous tubulointerstitial nephritis. *American Journal of Nephrology* **13**, 164–6.

Reddy S, Salant DJ (1998). Treatment of acute interstitial nephritis. *Renal Failure* **20**, 829–38.

Ruffing KA, *et al.* (1994). Eosinophils in urine revisited. *Clinical Nephrology* **41**, 163–6.

20.9.2 Chronic tubulointerstitial nephritis

Marc E. De Broe, Patrick C. D'Haese, and Monique M. Elseviers

Autoimmune

Sarcoidosis

Sarcoidosis is a multisystem disorder of unknown aetiology characterized by the accumulation in many tissues of T lymphocytes, mononuclear phagocytes, and non-caseating granulomas. The pathogenesis and clinical features of the condition are discussed in Chapter 17.11.6.

Clinically important renal involvement is an occasional problem—hypercalciuria and hypercalcaemia are most often responsible, although granulomatous interstitial disease, glomerular disease, obstructive uropathy, and (rarely) endstage renal disease may also occur. The true incidence of renal involvement in sarcoidosis remains unknown, but several small series of renal biopsies suggest that some degree of renal involvement occurs in approximately 35 per cent of patients with sarcoidosis.

Clinical features

Hypercalciuria, hypercalcaemia, nephrolithiasis, granulomatous interstitial nephritis, glomerular disease, and urinary tract disorders can all be observed in patients with sarcoidosis. Macrophages in a sarcoid granulomas contain a 1a-hydroxylase enzyme, but not a 24-hydroxylase enzyme, capable of converting vitamin D to its active form. The resultant increase in the absorption of calcium from the gut, which occurs in up to 50 per cent of those with sarcoidosis, leads to hypercalciuria and, in roughly 2.5 to 20 per cent of cases, to hypercalcaemia. Most patients remain asymptomatic, but nephrolithiasis, nephrocalcinosis, renal insufficiency, and polyuria are potential complications. Nephrolithiasis occurs in approximately 1 to 14 per cent of patients with sarcoidosis and may be the presenting feature. Nephrocalcinosis, observed in over half of those with renal insufficiency, is the most common cause of chronic renal failure in sarcoidosis. The increase in urine output associated with hypercalcaemia and hypercalciuria is due to a reduced responsiveness to antidiuretic hormone.

An interstitial nephritis with granuloma formation is common in sarcoidosis, but the development of clinical disease manifested by renal insufficiency is unusual. A survey of all renal biopsies over a 6-year period at three general hospitals found clinically significant sarcoid granulomatous interstitial nephritis in only four cases. Most affected patients have clear evidence of diffuse active sarcoidosis, although some present with an isolated elevation in the plasma creatinine concentration and no or only minimal renal manifestations. Renal biopsy reveals normal glomeruli, interstitial infiltration mostly with mononuclear cells, non-caseating granulomas in the interstitium, tubular injury, and—with more chronic disease—interstitial fibrosis. Granulomatosis interstitial nephritis is also seen in other diseases, including allergic interstitial nephritis (mainly drug-induced, caused by non-steroidal anti-inflammatory drugs (**NSAIDs**) and 5-aminosalicylic acid), Wegener's granulomatosis, berrylliosis, and tuberculosis. The urinary manifestations of granulomatous interstitial nephritis are relatively non-specific, with urinalysis typical of other chronic tubulointerstitial diseases, being normal or showing only sterile pyuria or mild proteinuria.

Glomerular involvement is rare in sarcoidosis. A variety of different lesions have been described in isolated cases, including membranous nephropathy, a proliferative or crescentic glomerulonephritis, and focal glomerulosclerosis. The presence of heavy proteinuria or red cell casts tends to differentiate these glomerulopathies from interstitial nephritis.

Occasionally, retroperitoneal lymph node involvement, retroperitoneal fibrosis, or renal stones may produce ureteral obstruction.

Diagnosis and treatment

Sarcoid nephropathy should be considered in any patient with unexplained renal failure and hypercalcaemia, nephrocalcinosis, renal tubular defect, or increased immunoglobulins. These patients often have signs and symptoms of pulmonary, ocular, and/or dermal involvement with sarcoidosis. The presence of granulomas on renal biopsy, while not specific to sarcoidosis, should strongly suggest this diagnosis in an appropriate setting. In patients with known sarcoidosis, sarcoid nephropathy should be considered in the presence of renal failure, hypercalcaemia, nephrolithiasis, nephrocalcinosis, or renal tubular defects.

Granulomatous interstitial nephritis can be treated effectively with glucocorticoids, typically prednisolone 1 to 1.5 mg/kg initially, tapered off following signs and symptoms of disease activity. Patients often respond quickly with an improvement in renal function, but this depends greatly on the extent and severity of inflammation and fibrosis before treatment was initiated. There are no controlled trials regarding the dose or length of the treatment.

The hypercalcaemia/hypercalciuric syndrome also responds quickly to corticosteroids: in general the dose needed to treat this complication is significantly lower than that required to treat granulomatous interstitial nephritis, and can be as low as 35 mg of prednisolone daily. Chloroquine, by decreasing the level of 1,25-dihydroxycholecalciferol, is an effective therapy for the hypercalcaemic/hypercalciuric syndrome. Ketoconazole, an inhibitor of steroidogenesis, has been used in a single patient who could not tolerate corticosteroids and was effective in decreasing the level of active vitamin D as well as serum and urinary calcium.

Although uncommon in patients with sarcoidosis, endstage renal failure (**ESRF**) requiring renal replacement therapy is most often due to hypercalcaemic nephropathy rather than granulomatous nephritis. Graft loss due to disease recurrence has not been reported.

Drug-induced nephropathy (Table 1)

Analgesics

Analgesic nephropathy is characterized by renal papillary necrosis and chronic interstitial nephritis caused by the prolonged and excessive consumption of analgesics. It is invariably caused by compound analgesic mixtures containing aspirin or other antipyretic agent in combination with phenacetin, paracetamol, or

salicylamide and caffeine or codeine in popular 'over-the-counter' proprietary medicines.

In the recent past, analgesic nephropathy has been one of the commoner causes of chronic renal failure, particularly in Australia and parts of Europe. Estimates made before phenacetin was removed from over-the-counter analgesics and prior to the enactment of legislation making combined analgesic preparations only available by prescription (in Sweden, Canada, and Australia), suggested that analgesic nephropathy was responsible for 1 to 3 per cent of cases of endstage renal disease in the United States as a whole: up to 10 per cent in areas of North Carolina, and 13 to 20 per cent in Australia and some countries in Europe (such as Belgium and Switzerland). During the 1990s, there was a clear decrease in the prevalence and incidence of the condition among patients undergoing dialysis in several European countries and Australia. Some authors have associated this decrease with the removal of phenacetin from analgesic mixtures. However, it is impossible to draw definitive conclusions from the epidemiological observations since other factors, such as eligibility criteria for dialysis treatment and the availability of analgesic mixtures, may also have had an influence.

Pathogenesis and pathology

The aetiology of analgesic nephropathy remains a controversial issue and the question of which kinds of analgesic are nephrotoxic is still a matter of debate. Since 1955 experimental studies on the nephrotoxicity of analgesics have been performed, mainly using rats fed with large amounts of drugs, sometimes aggravating the renal effects by dehydration or by introducing bacteria into the blood, peritoneum, or bladder. The results have been difficult to interpret, but it could be concluded that renal papillary necrosis was most frequently observed after the administration of aspirin in combination with phenacetin or paracetamol.

In humans, the long-standing excessive use of analgesics observed in patients with analgesic nephropathy is preferentially that of analgesic mixtures rather than single agents, abusers taking these products for their mood-altering effects rather than for the relief of physical complaints. Hence, all these mixtures contain caffeine and/or codeine, substances that can create a psychological dependence. In most of the early analgesic nephropathy reports, nearly all patients had taken large amounts of analgesic mixtures containing phenacetin. There is strong evidence that phenacetin-containing analgesic mixtures showed a high nephrotoxic potency in the past, and several case-control studies, as well as the prospective controlled longitudinal epidemiological study of Dubach *et al*., demonstrated a high increased risk associated with the regular consumption of analgesic mixtures containing phenacetin.

However, the withdrawal of phenacetin from analgesic mixtures in western Europe, Australia, and the United States gives rise to the question of the nephrotoxic potency of different kinds of products. Clinical observations in countries where analgesics without phenacetin have been on the market for more than 20 years (for example, Australia, Belgium, Germany) have shown that identical renal pathology is observed in patients abusing analgesic mixtures that have never contained phenacetin. We observed a cohort of 226 patients with analgesic nephropathy diagnosed according to objective renal imaging criteria: abuse of analgesic mixtures was documented in all except seven cases, and in 46 patients nephrotoxicity was found in the absence of any previous phenacetin consumption. These patients abused the combinations of: aspirin and paracetamol; aspirin and pyrazolones; paracetamol and pyrazolones; and two pyrazolones.

The mechanisms responsible for the renal injury are incompletely understood. Phenacetin is metabolized to acetaminophen and to reactive intermediates that can injure cells, in part by lipid peroxidation. These metabolites tend to accumulate in the medulla along the medullary osmotic gradient created by the countercurrent system. As a result, the highest concentrations are seen at the papillary tip, the site of the initial vascular lesions. The potentiating effect of aspirin with both phenacetin and acetaminophen may be related to two factors. First, acetaminophen undergoes oxidative metabolism by prostaglandin H synthase to reactive quinoneimine that is conjugated to glutathione. If acetaminophen is present alone, there is sufficient glutathione generated in the papillae to detoxify the reactive intermediate. However, if acetaminophen is ingested with aspirin, the aspirin is converted to salicylate, which becomes highly concentrated and depletes glutathione in both the cortex and papillae of the kidney. With the cellular glutathione depleted, the reactive metabolite of acetaminophen then produces lipid peroxides and arylation of tissue proteins, ultimately resulting in necrosis of the papillae (Fig. 1). Second, aspirin (and other NSAIDs) suppress prostaglandin production by inhibiting cyclooxygenase enzymes. Renal blood flow, particularly within the renal medulla that normally exists on the verge of hypoxia, is highly dependent upon the systemic and local production of vasodilatory prostaglandins. The final injury is therefore due to both the haemodynamic and cytotoxic effects of these drugs resulting in papillary necrosis and interstitial fibrosis.

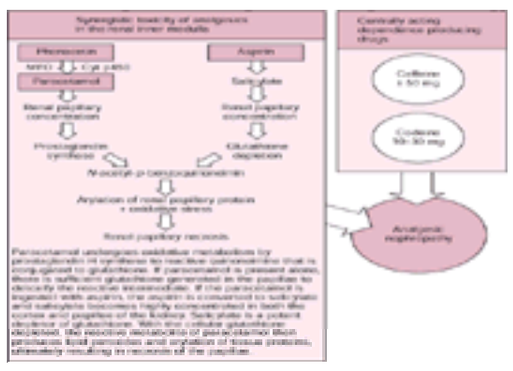

Fig. 1 Synergistic toxicity of analgesics in the renal inner medulla and centrally acting dependence-producing drugs leading to analgesic nephropathy. (Reproduced with permission from Kincaid-Smith P, Nanra RS (1993). In: Schrier RW, Gottschalk CW, eds. *Diseases of the kidney*, pp 1099–129. Little, Brown and Company, Boston, MA, and Duggin G (1996). *American Journal of Kidney Diseases* **28/1** (Suppl. 1), S39–S47.)

The renal damage induced by analgesics is most prominent in the medulla. The earliest changes consist of prominent thickening of the vasa recta capillaries (capillary sclerosis) and patchy areas of tubular necrosis; similar vascular lesions can be found in the renal pelvis and ureter, suggesting that the primary effect is damage to the vascular endothelial cells. Later changes include areas of papillary necrosis and secondary cortical injury with focal and segmental glomerulosclerosis and interstitial infiltration and fibrosis.

Clinical features

The renal manifestations of analgesic nephropathy are usually non-specific: normal renal function or slowly progressive chronic renal failure, and urinalysis that may be normal or may reveal sterile pyuria and mild proteinuria (less than 1.5 g/day). Hypertension and anaemia are commonly seen with moderate to advanced disease; more prominent proteinuria that can exceed 3.5 g/day can also occur at this time, a probable reflection of secondary haemodynamically mediated glomerular injury. Most patients have no symptoms referable to the urinary tract, although flank pain or macroscopic/microscopic haematuria from a sloughed or obstructing papilla may occur or as a result of a transitional-cell carcinoma. Urinary tract infection is also somewhat more common in women with this disorder.

Despite the non-specific nature of the renal presentation, there are frequently other findings that point toward the presence of analgesic nephropathy. Most patients are between the ages of 30 and 70 years, and careful questioning often reveals a history of chronic headaches or low back pain that leads to the analgesic use. Also common are other somatic complaints (such as malaise and weakness), and ulcer-like symptoms or a history of peptic ulcer disease due in part of chronic aspirin ingestion.

The decline in renal function can be expected to progress if analgesics are continued, whereas renal function stabilizes or mildly improves in most patients if analgesic consumption is discontinued. However, if the renal disease is already advanced, then progression may occur in the absence of drug intake, presumably due to secondary haemodynamic and metabolic changes associated with nephron loss. The late course of analgesic nephropathy may also be complicated by two additional problems: malignancy and atherosclerotic disease. Urinary tract malignancy will develop in as many as 8 to 10 per cent of patients with analgesic nephropathy, but in well under 1 per cent of phenacetin-containing analgesic users without kidney disease. In women under the age of 50, for example, analgesic abuse is the most common cause of bladder cancer, an otherwise unusual disorder in young women. The potential magnitude of this problem has also been illustrated by histological examination of nephrectomy specimens obtained prior to renal transplantation, when the incidence of urothelial atypia approaches 50 per cent. The tumours generally become apparent after 15 to 25 years of analgesic abuse, usually but not always in patients with clinically evident analgesic nephropathy. Most patients are still taking the drug at the time of diagnosis, but clinically evident disease can first become apparent several years after cessation of analgesic intake and even after renal transplantation. It is presumed that the induction of malignancy results from the intrarenal accumulation of *N*-hydroxylated phenacetin metabolites that

have potent alkylating action. The highest concentration of these metabolites will be in the renal medulla, ureters, and bladder (as described above), possibly explaining the predisposition to carcinogenesis at these sites. The pathogenetic importance of phenacetin metabolites is suggested indirectly from the observation that the prolonged ingestion of other analgesics that can cause papillary necrosis, but do not form the same metabolites, such as acetaminophen and the NSAIDs, is not associated with tumour formation. The main presenting symptom of urinary tract malignancy in patients with analgesic nephropathy is microscopic or gross haematuria, hence continued monitoring is essential, and new haematuria should be evaluated by urinary cytology, and—if indicated—cystoscopy with retrograde pyelography. The incidence of urothelial carcinoma after renal transplantation in patients with analgesic nephropathy is comparable to the general incidence, of up to 10 per cent, of urothelial carcinomas in ESRF patients with analgesic nephropathy. Removal of the native kidneys prior to renal transplantation has also been suggested, but the efficacy of this regimen has not been proven.

Diagnosis and treatment

The lack of reliable criteria and the high prevalence of analgesic nephropathy during the 1980s in Belgium (17.9 per cent in 1984) led us to perform a series of prospective, multicentre, controlled studies to define and validate the diagnostic criteria for this disease. We could provide strong evidence that specific anatomical changes, best seen by non-contrast computed tomography (**CT** scan), have much greater sensitivity and specificity than other clinical signs and symptoms in the diagnosis of endstage renal disease due to analgesic nephropathy. These changes are: (1) decrease in renal volume; (2) bumpy renal contours; and (3) papillary calcifications. In a more recent study, these observations were validated in a representative sample of patients with analgesic abuse with endstage renal disease and extended to patients with moderate renal failure. In patients with ESRF, decreased renal volume had the greatest sensitivity at 95 per cent, whilst papillary calcification had the highest specificity, and contour or papillary necrosis had a sensitivity and specificity of 90 per cent. In patients with moderate renal failure, papillary calcification was most sensitive at 92 per cent and specific at 100 per cent. The combination of papillary necrosis with either a bumpy renal contour or small kidneys did not improve sensitivity or specificity. In clinical practice, however, it is important to remember that the predictive value of this test, like any other diagnostic test, is very much dependent on the prevalence of the disease in the population under study. This test should therefore be utilized in patients with a reasonable risk for analgesic nephropathy and not as a general screening test.

As indicated above, patients with normal or only mildly/moderately impaired renal function should be strongly encouraged to stop taking analgesics, in the hope that further deterioration in renal function can be avoided. Those with severe or endstage renal failure are unlikely to recover renal function, although there may be other valid medical reasons for recommending that they stop ingesting large quantities of analgesics. The medical management of chronic renal failure is along conventional lines, as is provision of renal replacement therapy.

Non-steroidal anti-inflammatory drugs (NSAIDs)

NSAIDs are popular for treating a wide range of clinical conditions, available both over-the-counter and on prescription. Despite their usefulness, there is substantial evidence from experimental and clinical studies that NSAIDs have a variety of effects on the kidney. The most common renal disorder associated with NSAIDs is acute, largely reversible, insufficiency due to the inhibition of renal vasodilatory prostaglandins in the clinical setting of a stimulated renin–angiotensin system. Older age, hypertension, concomitant use of diuretics or aspirin, pre-existing renal failure, diabetes, and plasma-volume contraction are known risk factors for renal failure after the ingestion of NSAIDs. Rarely, NSAIDs may cause acute interstitial nephritis with proteinuria. These effects appear to be common to all NSAIDs, and are likely to be observed with cyclooxygenase-2 inhibitors as well as cyclooxygenase-1 inhibitors because both have been identified in adult and fetal human kidneys, suggesting a role for both enzymes in normal renal physiology.

By contrast to the well-characterized acute effects of NSAIDs on the kidney, the chronic effects are less well documented. However, a recent report demonstrated that NSAIDs are the most frequent cause of permanent renal insufficiency after acute interstitial nephritis. Risk factors for irreversible failure are pre-existing renal damage, long-standing intake of the causative drug, slow oligosymptomatic disease development, and histological signs of chronicity with those of acute interstitial nephritis. Although renal papillary necrosis and chronic renal failure can occur after the prolonged use of NSAIDs, the actual risk of these serious complications is unknown. Furthermore, the frequency of renal papillary necrosis as a primary or contributing cause of endstage renal disease remains unknown.

5-Aminosalicylic acid

Over the past few years an association between the use of 5-aminosalicylic (**5-ASA**) in patients with chronic inflammatory bowel disease and the development of a particular type of chronic tubulointerstitial nephritis has been suggested.

For many years, sulfasalazine, an azo-compound derived from sulphapyridine and 5-ASA, the latter being the pharmacologically active moiety, was the only valuable non-corticosteroid drug in the treatment of inflammatory bowel disease. Since the therapeutically inactive sulphapyridine moiety was largely responsible for the mainly haematological side-effects of sulfasalazine, this stimulated the development of a number of new 5-ASA formulations (mesalazine, olsalazine, balsalazine) for topical and oral use. In the last decade, these new 5-ASA products replaced sulfasalazine as the first-line therapy for mildly to moderately active inflammatory bowel disease. However, a literature search revealed 17 published cases of renal impairment associated with 5-ASA therapy in patients with inflammatory bowel disease, and in several it was shown that this did not recover completely upon stopping the drug, even after a follow-up period of several years. In a retrospective study, nephrologists reported 40 patients with inflammatory bowel disease showing renal impairment, including 15 cases with interstitial nephritis and previous use of 5-ASA. Stimulated by these findings we started a European prospective registration study aiming to register all patients with inflammatory bowel disease and renal impairment and to control for a possible association with 5-ASA therapy. A cohort of 1449 patients with inflammatory bowel disease seen during 1 year in the outpatient clinics of 28 European gastroenterology departments was investigated: preliminary results showed 30 patients (2 per cent) with decreased renal function, and a possible association with 5-ASA therapy was found in half of them.

However, determining the cause of renal disease in those with inflammatory bowel disease is not straightforward. The most frequent renal complications are oxalate stones and their consequences, such as pyelonephritis, hydronephrosis, and (in the long-term) amyloidosis. Chronic inflammatory bowel disease is also associated with glomerulonephritis: minimal-change glomerulonephritis, membranous, membranoproliferative, focal glomerulosclerosis, and proliferative crescentic glomerulonephritis have all been reported. As for many drugs, reversible, acute interstitial nephritis has been described with the use of 5-ASA compounds. In view of this complexity, the association of 5-ASA and chronic interstitial nephritis in patients with inflammatory bowel disease can be difficult to interpret, since renal involvement may be an extraintestinal manifestation of the underlying disease. However, the particular form of chronic tubulointerstitial nephritis in patients with inflammatory bowel disease treated with 5-ASA is characterized by an important cellular infiltration of the interstitium with macrophages, T cells, and also B cells (Fig. 2).

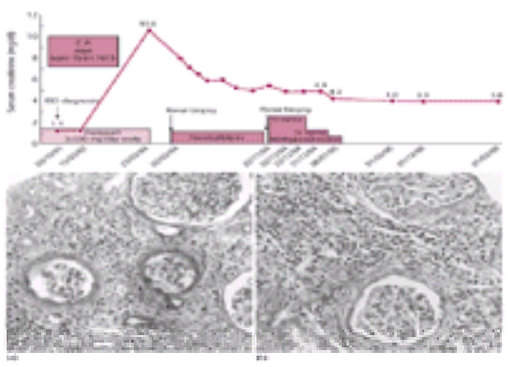

Fig. 2 5-ASA. A 36-year-old male patient suffering from Crohn's disease presented with severe renal failure after 23 months of 5-aminosalicylic acid (Pentasa®) treatment. A first renal biopsy showed widening and massive cellular infiltration of the interstitium, tubular atrophy, and relative spacing of glomeruli. The cellular infiltration was identified using appropriate monoclonal antibodies and consisted not only of T cells and macrophages, but also B cells. A second renal biopsy performed after the drug had been stopped for 8 months, when there was a modest improvement in renal function, again showed a significant cellular infiltration of the interstitium, tubular atrophy, and fibrosis. (a) First renal biopsy, (b) second renal biopsy.

Pathogenesis and pathology

That 5-ASA causes renal disease is supported by the number of case reports appearing in the recent literature of patients with inflammatory bowel disease using 5-ASA as their only medication, the improvement (at least partially) of impaired renal function upon stopping the drug, and a worsening after resuming 5-ASA use. Furthermore, the molecular structure of 5-ASA is very close to that of salicylic acid, phenacetin, and aminophenol, drugs with well-documented nephrotoxic potential (Fig. 1). Calder *et al*. found that necrosis of the proximal convoluted tubules and papillary necrosis developed in rats after a single intravenous injection of 5-ASA at doses of 1.4, 2.8, and 5.7 mmol/kg body weight (high pharmacological doses). The mechanism of renal damage, possibly caused by 5-ASA itself, may be analogous to that of salicylates by inducing hypoxia of renal tissues, either by uncoupling oxidative phosphorylation in renal mitochondria, by inhibiting the synthesis of renal prostaglandins, or by rendering the kidney susceptible to oxidative damage by a reducing renal glutathione concentration after inhibition of the pentose phosphate shunt.

Clinical features

A typical case is shown in Fig. 2. An intriguing aspect of this type of toxic nephropathy is the documented persistence of the renal interstitium inflammation even several months/years after first taking the drug. The disease is more prevalent in men, with a male:female ratio of 15:2. The age of reported cases ranges from 14 to 45 years. By contrast with analgesic nephropathy, where renal lesions are only observed after several years of analgesic abuse, interstitial nephritis associated to 5-ASA was already observed during the first year of treatment in 7 out of 17 reported cases, most of whom had started 5-ASA therapy with documented normal renal function. In several cases, particularly those in which there was a delayed diagnosis of renal damage, recovery of renal function did not occur, and some needed renal replacement therapy.

Diagnosis and treatment

Since this type of chronic tubulointerstitial nephritis produces few if any symptoms, and if diagnosed at a late stage progresses to irreversible chronic endstage renal disease, serum creatinine levels should be measured in any patient with inflammatory bowel disease treated with 5-ASA at the start of the treatment, every 3 months for the remainder of the first year, and annually thereafter. The use of concurrent immunosuppressive therapy, as is the case in severe forms of chronic inflammatory bowel disease, may necessitate extension to the period of intensive renal function monitoring. If serum creatinine increases, a renal biopsy is the only way to demonstrate the cause.

Chinese herbs

In 1992, physicians in Belgium noted an increasing number of women presenting with renal failure, often near end stage, following their exposure to a slimming regimen containing Chinese herbs. An initial survey of seven nephrology centres in Brussels identified 14 women under the age of 50 who had presented with advanced renal failure due to biopsy-proven, chronic tubulointerstitial nephritis over a 3-year period; nine of whom had been exposed to the same slimming regimen. As of early 2000, a total of more than 120 cases had been identified. The epidemiology is unknown, as is the risk for the development of severe renal damage, but the recent publication of case reports from several countries in Europe and Asia would seem to indicate that the incidence of herbal medicine-induced nephrotoxicity is more common than previously thought.

Pathogenesis and pathology

The aetiology of Chinese herbal nephropathy is not fully understood. A plant nephrotoxin, aristolochic acid, was proposed as a possible aetiological agent, but this compound was not part of the herbal preparations used by all the patients. Furthermore, aristolochic acid (0.15 mg/tablet) has been used as an immunomodulatory drug for 20 years in Germany by thousands of patients, sometimes in doses comparable to the Chinese herb slimming regimen; despite this exposure, there is no report relating chronic tubulointerstitial nephritis to aristolochic acid.

In addition to aristolochic acid, patients with Chinese herb nephropathy also received the appetite suppressants fenfluramine and diethylpropion, which have vasoconstrictive properties, and acetazolamide, which alkalinizes the urine, thereby potentially enhancing the nephrotoxic effect of aristolochic acid. Another uncertain factor is why only some patients exposed to the same herbal preparations develop renal disease. Women appear to be at greater risk than men: other factors that might be important include toxin dose, batch-to-batch variability in toxin content, individual differences in toxin metabolism, and a genetically determined predisposition toward nephrotoxicity and/or carcinogenesis.

At one centre in Belgium, 19 native kidneys and ureters were removed in a series of 10 patients during and/or after renal transplantation: multifocal, high-grade, flat, transitional-cell carcinoma (carcinoma *in situ*) was observed in four (40 per cent), whilst all had multifocal moderate atypia. Tissue samples revealed aristolochic acid-related DNA adducts, indicating a possible mechanism underlying the development of malignancy. In another study of 39 patients with Chinese herbal nephropathy and endstage renal disease who underwent prophylactic removal of the native kidneys and ureters, urothelial carcinoma was discovered in 18 and mild-to-moderate urothelial dysplasia in 19. All atypical cells were found to overexpress a p53 protein, suggesting the presence of a mutation in the gene.

The main histological lesion, which is located principally in the cortex, is extensive interstitial fibrosis with atrophy and loss of the tubules (Fig. 3). Cellular infiltration of the interstitium is scarce. Thickening of the walls of the interlobular and afferent arterioles result from endothelial cells swelling. The glomeruli are relatively spared and immune deposits are not observed. These findings suggest that the primary lesions may be centred in the vessel walls, thereby leading to ischaemia and interstitial fibrosis.

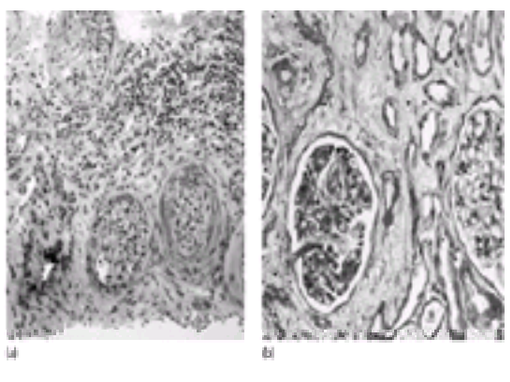

Fig. 3 Renal biopsy showing tubular atrophy, widening of the interstitium, cellular infiltration, and fibrosis, with glomeruli surrounded by a fibrotic ring, in a case of Chinese herb nephropathy. (a) Masson staining, (b) haematoxylin–eosin staining.

Clinical features

Patients present with renal insufficiency and other features indicating a tubulointerstitial disease. Blood pressure is either normal or only mildly elevated, and the urinary sediment reveals only a few red and white cells. The urine contains protein (less than 1.5 g/day), consisting of both albumin and low molecular weight proteins that are normally reabsorbed by the proximal tubules, hence tubular dysfunction—also marked by glycosuria—contributes to the proteinuria. The plasma creatinine concentration at presentation has ranged from 1.4 to 12.7 mg/dl (123 to 1122 µmol/l). Follow-up studies have revealed relatively stable renal function in most patients, with an initial plasma creatinine concentration below 2 mg/dl (176 µmol/l). However, progressive renal failure resulting in eventual dialysis or transplantation may ensue in patients with more severe disease, even if further exposure to Chinese herbs is prevented.

An extremely similar clinical and pathological process has been reported in a group of patients from Taiwan who had ingested a selection of uncontrolled traditional

Chinese herbs that differed from those of the slimming regimen. Despite discontinuation of these remedies, progressive renal failure was common.

Diagnosis and treatment

There are no specific criteria for the diagnosis of this type of renal disease. The condition should be suspected in any patient with unexplained relatively rapidly progressive renal disease who is using/abusing herbal remedies. The presence of tubular proteinuria may be a clue to the diagnosis, particularly in the early stages. The histological appearances are not specific, but renal biopsy is necessary to exclude other conditions in this clinical context.

There is no proven effective therapy for this disorder, which typically presents with marked interstitial fibrosis without prominent inflammation. An uncontrolled study suggested that corticosteroids may slow the rate of loss of renal function. The high incidence of cellular atypia of the genitourinary tract suggests that, as a minimum, these patients should undergo regular surveillance for abnormal urinary cytology. Whether more aggressive management strategies, such as bilateral native nephrouterectomies (particularly in those undergoing renal transplantation), are required is unclear. Findings from a recent report support the more aggressive option. Renal transplantation is an effective modality for those who progress to endstage renal disease, one report noting no recurrence in five patients.

Lithium

Lithium is used extensively in the treatment of patients with manic-depressive psychosis. Different forms of renal effects/injury have been described: most frequently nephrogenic diabetes insipidus, but also renal tubular acidosis, chronic interstitial nephritis, nephrotic syndrome, and focal segmental glomerular sclerosis/global glomerular sclerosis. Hyperparathyroidism is observed in patients treated with lithium.

Pathogenesis and pathology

Lithium is eliminated from the body almost entirely by the kidney, being filtered at the glomerulus and reabsorbed in the proximal tubule, resulting in a clearance of one-third of the normal creatinine clearance. It moves in and out of cells only slowly and accumulates in the kidney, particularly in the collecting tubule, entering these cells through sodium channels in the luminal membrane. Hence, its principal toxicity relates to distal tubular function, where inhibition of adenylate cyclase and generation of cyclic AMP result in downregulation of aquaporin-2, the collecting tubule water channel, and a decrease in ADH receptor density, leading to resistance to antidiuretic hormone. Further effects compound this. A low intracellular level of cyclic AMP leads to the increased cellular levels of glycogen observed in kidney biopsy specimens from patients taking lithium, as does the fact that lithium also directly inhibits enzymes involved in glycogen breakdown. The ensuing increased glycogen storage may interfere with distal tubular function and be responsible for the observation that polyuria and polydipsia in lithium-treated patients is due to nephrogenic diabetes insipidus.

The tubular defect in the distal nephron can also impair the ability to maximally acidify the urine. A lithium-induced decrease in the activity of the H+-ATPase pump in the collecting tubule may be responsible for this defect.

Lithium treatment has been aetiologically related to parathyroid hypertrophy and hyperfunction, the latter seeming to be due to an upward resetting of the level at which the plasma calcium concentration depresses parathyroid hormone (**PTH**) release. Persistent hypercalcaemia (in 5–10 per cent of the patients) may exacerbate both the concentrating defect and the interstitial nephritis seen in lithium-treated patients.

Renal biopsies from patients taking lithium show a specific histological lesion in the distal tubule and collecting duct. On light microscopy there is swelling and vacuolization in cells associated with a considerable accumulation of Periodic acid–Schiff (**PAS**)-positive glycogen. This is present in all renal biopsies from patients taking lithium, it appears within days after the administration of lithium and disappears when lithium ingestion is ceased.

Hestbech *et al*. were the first to suggest that progressive chronic interstitial lesions occurred in the kidneys of patients receiving lithium. However, a controlled study showed no difference between biopsies from patients taking lithium and those from a group of patients who had affective disorders but were not doing so. Specifically, there was no difference in the incidence of glomerular sclerosis, interstitial fibrosis, tubular atrophy, cast formation, or interstitial volume, but there was a significant increase in the number of microcysts in the lithium-treated patients. One reason why it has been difficult to determine the nature of lithium-induced chronic renal damage has been the lack, until recently, of an animal model in which lesions similar to those noted in human biopsies could be demonstrated. However, a recent study on lithium nephrotoxicity carried out in the rabbit showed clear-cut evidence of progressive histological and functional impairment, with the development of significant interstitial fibrosis, tubular atrophy, glomerular sclerosis, and cystic tubular lesions. A recent publication by Markowitz *et al*. revealed a chronic tubulointerstitial nephropathy in 100 per cent of 24 patients having received lithium for several years, associated with cortical and medullary tubular cysts or dilatation. There was also a surprisingly high prevalence of focal segmental glomerulosclerosis and global glomerulosclerosis, sometimes of equivalent severity to the chronic tubulointerstitial disease. Despite discontinuing lithium treatment, seven of nine patients with initial serum creatinine values above 2.5 mg/dl progressed to endstage renal disease. Nevertheless, an answer to the question as to whether or not chronic lithium therapy causes chronic interstitial nephritis still needs more hard data.

Clinical features

Apart from acute lithium intoxication, chronic poisoning can occur in patients whose lithium dosage has been increased or in those with a decreased effective circulating volume, decreased sodium intake, diabetes mellitus, gastroenteritis, and renal failure, thereby resulting in an increase in serum lithium levels. Symptoms associated with poisoning include lethargy, drowsiness, coarse hand tremor, muscle weakness, nausea, vomiting, weight loss, polyuria, and polydipsia. Severe toxicity is associated with increased deep tendon reflexes, seizures, syncope, renal insufficiency, and coma. The commonest manifestation is altered mental status.

Chronic lithium poisoning is frequently associated with electrocardiogram changes, including ST-segment depression and inverted T waves in the lateral precordial leads. Lithium is concentrated within the thyroid and inhibits the synthesis and release of thyroxine, which can lead to hypothyroidism and hypothermia. It may also cause thyrotoxicosis and hyperthermia. Symptoms of hypercalcaemia may also be present, exacerbating the urinary concentrating defect already present in these patients.

In patients with glomerular lesions such as minimal-change or focal glomerular sclerosis, proteinuria generally begins within 1.5 to 10 months after the onset of therapy, completely or partially resolving in most patients within 4 weeks after lithium is discontinued. Reinstitution of lithium has led to recurrent nephrosis in some cases.

The hyperparathyroidism observed in patients receiving lithium treatment is characterized by elevated parathyroid hormone levels, hypercalcaemia, hypocalciuria, and normal serum phosphate levels, by contrast to primary hyperparathyroidism in which hypophosphataemia and hypercalciuria are seen.

Diagnosis and treatment

The severity of chronic lithium intoxication correlates directly with the serum lithium concentration and may be categorized as mild (1.5–2.0 mEq/l), moderate (2.0–2.5 mEq/l), or severe (>2.5 mEq.l).

Polyuria and polydipsia due to nephrogenic diabetes insipidus and other acute manifestations of the effect of lithium on the kidney usually disappear rapidly if lithium is withdrawn. The decision about management, however, usually revolves around the relative benefit of the lithium in controlling and preventing the manifestation of manic-depressive psychosis, and the disadvantage to the patient of the major side-effect of lithium, namely polyuria. In most cases the lithium is so clearly beneficial that the polyuria is accepted as a side-effect and treatment continued. It is likely that the serum concentration of lithium is important, and that renal damage is more likely to occur if the serum concentration is consistently high or if repeated episodes of lithium toxicity occur. The serum lithium concentration should therefore be monitored carefully (at least every 3 months) and maintained at the lowest level that will provide adequate control of the manic depressive psychosis.

Much more difficult to handle is the situation where a patient on long-term lithium therapy is found to have impaired renal function for which there is no obvious alternative cause. As stated above, renal failure may progress even if lithium therapy is withdrawn, and in some patients the discontinuation of lithium can lead to a devastating deterioration in their psychiatric condition. The decision as to whether or not to discontinue lithium should therefore be made after frank and open discussion, admitting all uncertainties, with the patient, psychiatric colleagues, and (if appropriate) relatives/carers.

Endemic Balkan nephropathy

Endemic Balkan nephropathy (**EBN**) is a chronic, familial, non-inflammatory tubulointerstitial disease of the kidneys. A high frequency of urothelial atypia, occasionally culminating in tumours of the renal pelvis and urethra, is associated with this disorder.

As the name suggests, EBN is most commonly seen in South-Eastern Europe, including the areas traditionally considered to comprise the 'Balkans': Serbia, Bosnia and Herzegovina, Croatia, Romania, and Bulgaria. It is most likely to occur among those living along the confluence of the Danube river, a region in which the plains and low hills generally have a high humidity and rainfall (Fig. 4). There is a very high prevalence in endemic areas, with rates ranging between approximately 0.5 and 4.4 per cent, increasing to as high as 20 per cent if the disorder is suspected and carefully screened for among an at-risk population. A striking observation is that nearly all affected patients are farmers.

Fig. 4 Foci of endemic Balkan nephropathy.

Pathogenesis and pathology

Although the aetiology of EBN is unknown, many environmental and genetic factors have been evaluated as possible underlying causes.

Environmental factors

Given that it is endemic to a specific geographic area, toxins and/or environmental exposures that are unique to the Balkans have been investigated. However, no agent and/or general group of compounds or organisms, including trace elements (lead, cadmium, silica, selenium), viruses, fungus, and/or plant toxin, has yet been successfully identified.

One intriguing possibility is that aristolochic acid, a mutagenic and nephrotoxic alkaloid found in the plant *Aristolochia clematis*, may underlie both Chinese herbal nephropathy (see above) and EBN (see Table 1). There are striking pathological and clinical similarities between the progressive interstitial fibrosis observed in young women who have been on a slimming regimen containing Chinese herbs (as well as other agents) and EBN, but this putative association between EBN and aristolochic acid remains speculative.

Genetic factors

Support for a genetic aetiology includes observations that the disease clearly affects particular families, and that some ethnic populations who have lived in endemic areas for generations do not suffer from EBN. The mode of inheritance has not yet been established and possible causative gene(s) have not been identified, but a locus in the region between 3q25 and 3q26 has been incriminated.

By contrast, some observations are inconsistent with a genetic basis. First, EBN is observed in individuals who have immigrated into the 'Balkan' area from regions without the disorder, and in previously unaffected families who have lived for at least 15 years in endemic areas. Second, EBN does not develop in members from previously affected families who have left endemic areas early in life or who spent less than 15 years in these areas.

A unifying hypothesis may be that the disease most likely occurs in genetically predisposed individuals who are chronically exposed to a causative, as yet unidentified agent.

In the early stages of disease, renal histology reveals focal cortical tubular atrophy, interstitial oedema, and peritubuloglomerular sclerosis with limited mononuclear-cell infiltration. Narrowing and endothelial swelling of interstitial capillaries (e.g. capillarosclerosis) is also described. In advanced cases, marked tubular atrophy and interstitial fibrosis develop along with focal segmental glomerular changes and global sclerosis. There is an extremely high incidence of cellular atypia and urothelial carcinoma of the genitourinary tract.

Clinical features

EBN is a slowly progressive tubulointerstitial disease that may culminate in endstage renal disease. Clinical manifestations first appear between 30 and 50 years of age, with findings prior to the age of 20 being extremely rare. One of the first signs is tubular dysfunction, which is characterized by an increased excretion of low molecular weight proteins (such as b2-microglobulin). Early tubular injury can also lead to renal glycosuria, aminoaciduria, and diminished ability to handle an acid (NH_4Cl) load. Over a period of more than 20 years there is a progressive decrease in concentrating ability (resulting in polyuria) and in the glomerular filtration rate (resulting in endstage renal disease). Patients are usually without oedema and are normotensive, hypertension only developing with endstage disease. A normochromic normocytic anaemia occurs with early disease, which becomes increasingly pronounced as the disorder progresses. Urinary tract infection is rarely observed. Kidneys are of normal size early in the course of the disease. A symmetrical reduction of kidney size with a smooth outline and normal pelvicaliceal system is subsequently observed in patients with late-stage disease. Intrarenal calcifications are not observed.

EBN is also associated with the development of transitional-cell carcinoma of the renal pelvis or ureter, with studies noting a wide range in incidence (2 to nearly 50 per cent). These tumours are generally superficial and slow-growing.

Diagnosis and treatment

The diagnosis of EBN is based upon the presence of some combination of the following findings:

- symmetrically shrunken kidneys with absence of intrarenal calcifications;
- farmers living in the endangered villages;
- familial history positive for endemic Balkan nephropathy;
- mild tubular proteinuria, hyposthenuria; and
- normochromic hypochromic anaemia occurring in patients with only slightly impaired renal function.

As with many other chronic tubulointerstitial diseases of unclear origin, there is no specific prevention or treatment. Therapy is therefore supportive, with renal replacement therapy being initiated in patients with endstage renal disease.

The high incidence of cellular atypia in the genitourinary tract suggests that regular surveillance should be performed for abnormal urinary cytologies. Whether

bilateral native nephroureterectomies are required, particularly in those undergoing renal transplantation, is unclear.

Radiation nephropathy

Radiation nephropathy is a renal disorder caused by ionizing radiation. The kidney may be injured by radiation administered to tumours within the kidney or nearby tissues (testis, ovary, retroperitoneum). Clinicians were aware of the potential adverse effects of X-rays on renal function from the beginning of the twentieth century, and between 1940 and 1960 a significant number of cases were reported. In 1953 Luxton established the clinical features of the condition and defined the tolerance of the kidney to irradiation, leading to preventive shielding of the kidneys in patients receiving radiation therapy and to a marked decline in the frequency of radiation nephropathy. In the last decade, however, total-body irradiation preceding bone marrow transplantation has resulted in an increasing incidence of radiation nephropathy, with late chronic renal failure developing in 20 per cent of patients who receive this treatment.

Pathogenesis and pathology

The radiation doses traditionally associated with radiation nephropathy were above 2000 rad (20 Gy) (less in children). By contrast, in patients receiving total-body irradiation preceding bone marrow transplantation, renal impairment was observed after doses of 1000 to 1400 rad (10–14 Gy). Fractionation, time, and effects of cytotoxic chemotherapy can probably explain the differences. In laboratory rodents, fractionation of the total dose into multiple separated doses decreases the risk, probably due to repair of sublethal radiation damage during the time between the fractionated doses. Total-body irradiation before bone marrow transplantation is usually administered over a short period, which does not allow sufficient time for the repair of radiation injury to the kidney. Moreover, the additional cytotoxic chemotherapy given to these patients potentiates the effects of ionizing radiation.

The precise pathogenesis of radiation nephropathy remains to be determined. The initial target of ionizing radiation within the kidney appears to be the endothelial cell. Radiation kills cells by damaging DNA, so that cell death after radiation is delayed until the cell divides. After the initial glomerular endothelial injury, vascular occlusion subsequently develops, leading to tubular atrophy. Because inflammatory cells are not seen in the renal parenchyma, the previously used terminology of 'radiation nephritis' is a misnomer.

The pathological features of radiation nephropathy comprise a continuous spectrum of changes that vary in relation to the dose of irradiation administered and the time elapsed after exposure. Large doses are followed by complete atrophy, thickening of basement membranes, and interstitial fibrosis.

Clinical features

Radiation nephropathy can take several forms. Acute radiation nephropathy occurs between 6 and 12 months after radiation therapy and presents with hypertension, anaemia, and oedema. The severity of hypertension ranges from mild to malignant, and more than half of the patients progress to chronic renal failure. Radiation nephropathy after total-body irradiation before bone marrow transplantation most closely corresponds to this acute form of radiation nephropathy. A more insidious chronic form of radiation nephropathy develops over a period of several years and presents primarily with diminished glomerular filtration rate, hypertension, and (occasionally) proteinuria. Another subset of patients may develop hypertension within a few years of irradiation, evolving in some to malignant hypertension with accelerated loss of renal function. Isolated persistent or intermittent proteinuria may also occur, frequently developing more than a decade after radiation exposure.

Diagnosis and treatment

Radiographic studies may help in the diagnosis of acute radiation nephropathy. CT scans with contrast demonstrate sharply demarcated, dense, persistent nephrograms corresponding to the irradiated areas.

The treatment of radiation nephropathy is supportive. Aggressive treatment of hypertension may slow the progression of renal disease. Hypertension due to unilateral disease may respond to nephrectomy. Additionally, the use of ACE inhibitors may have its classical renoprotective effect independent of antihypertensive action.

Since radiation nephropathy is an irreversible process, preventive measures should be taken during the administration of radiation. This includes selective shielding of the kidneys and the use of fractionated doses. Patients exposed to additional nephrotoxins remain at an increased risk of toxic effects.

Toxins

Lead

Lead toxicity affects many organs, resulting in encephalopathy, anaemia, peripheral neuropathy, gout, and renal failure. It was the epidemic of lead nephropathy in Queensland (Australia) that provided the strongest link between lead and chronic tubulointerstitial nephritis. Henderson noted an excess mortality due to chronic interstitial nephritis in Queensland but not in other parts of Australia, and correlated the incidence of granular contracted kidneys at autopsy with the lead content of the skull in people from Queensland and Sydney, showing that this correlated closely with the incidence of renal failure. Exposure was due to the lead-based paints used between 1890 and 1930, but recently the source of lead is industrial exposure. This type of exposure is often insidious, occurring over a very long period. Two studies have shown an inverse relationship between low-level lead exposure and renal function in the general population. Recent studies have failed to show any effect on renal function 17 to 50 years after an episode of acute childhood plumbism, the difference with Henderson's findings reflecting the greater lead burden in his study compared to the recent ones. Although low-level lead exposure in the general population is associated with mild but significant depression of renal function, its role in the development of endstage renal disease is unclear.

Pathogenesis and pathology

The pathogenesis of renal disease seen in the context of lead exposure may be related to proximal tubule reabsorption of filtered lead, with subsequent accumulation in proximal tubule cells. Aminoaciduria, glycosuria, and phosphaturia representing the Fanconi syndrome are observed after lead exposure, and thought to be related to an effect of lead on mitochondrial respiration and phosphorylation. Since lead is also capable of reducing 1,25-dihydroxyvitamin D synthesis, prolonged hyperphosphaturia and hypophosphataemia caused by lead poisoning in children could result in bone demineralization and rickets. Chronic lead poisoning can affect glomerular function: after an initial period of hyperfiltration the glomerular filtration rate is reduced and nephrosclerosis and chronic renal failure may ensue. Protracted lead exposure also interferes with the distal tubular secretion of urate, leading to hyperuricaemia and gout.

Renal biopsies in patients with subclinical lead nephropathy and a mild to moderate decrease in glomerular filtration rate primarily show focal tubular atrophy and interstitial fibrosis with minimal cellular infiltration. Electron microscopy shows mitochondrial swelling, loss of cristae, loss of basal infoldings, and a lysosomal-like structure containing dense bodies in the proximal tubules. In Australian patients who died as a result of severe lead exposure, their kidneys were fibrotic and shrunken, the interstitium showed variable degree of fibrosis with tubular dilatation, and the vessels had thickened muscular walls with subintimal hyaline deposition in afferent arterioles, but these findings in patients with endstage renal failure were non-specific.

Clinical features

Renal failure becomes apparent years after exposure and is associated with gout in most, if not all, cases. Hypertension is a very common feature of lead nephropathy, and an association between hypertension without renal failure and low-level lead exposure has gained increasing recognition over the past two decades. Although hyperuricaemia is common in renal failure, gout is unusual and its presence should raise the possibility of lead nephropathy.

However, whether chronic lead nephropathy exists as a clinical entity has been questioned. Many studies of occupational lead poisoning have not taken into account the coexposure to other toxins such as cadmium. Additionally, the relationship between early markers of renal tubular dysfunction, such as the urinary excretion of low molecular weight proteins or *N*-acetyl b-D-glucosaminidase, to the subsequent development of renal failure remains to be determined.

Diagnosis and treatment

As the blood lead level only reflects recent lead exposure, and is usually normal in patients with chronic renal failure due to their previously sustained low-level lead

exposure, the diagnosis has to be based on measurement of the body lead burden. The test of choice is the **EDTA** (ethylenediaminetetraacetic acid) mobilization test: this involves the administration of 2 g of EDTA intramuscularly in two divided doses 8 to 12 h apart, and collection of three consecutive 24-hour urine samples. A cumulative excretion of more than 600 µg is suggestive for a high lead body burden. Renal failure in itself does not increase body lead load but it does delay the excretion of lead. The diagnosis of lead nephropathy should be considered in any patient with progressive renal failure, mild to moderate proteinuria, significant hypertension, history of gout, and an appropriate history of exposure.

There is very little experience of the therapeutic use of EDTA in patients with chronic renal failure. Wedeen *et al*. treated eight industrially exposed patients with EDTA injections thrice weekly for 6 to 15 months, all having mild renal failure with GFRs of around 50 ml/min before treatment—four patients improved with a 20 per cent increase in their GFR.

Cadmium

Cadmium is a cumulative environmental pollutant and accumulates in the human body after inhalation or gastrointestinal absorption. Due to its various applications and increased industrial production, this element's release into the environment increased considerably from the 1950s onwards, particularly in Belgium and Japan, which are among the most important cadmium-producing countries worldwide. However, the atmospheric emissions of cadmium from zinc smelters have been reduced since the 1970s. At the present, normal cadmium values are set at 0.1 to 0.8 µg/l (non-smokers) in blood and 0.02 to 0.7 µg/g creatinine in urine.

Cadmium is a highly toxic metal. The kidney is the element's most important target organ and it has long been recognized that high-level exposure to cadmium after inhalation or ingestion can give rise to nephrotoxicity in humans, and that this effect is usually considered to be the earliest and most important feature from the point of view of health. Cadmium induces a tubular proteinuria (of low molecular weight plasma proteins). Hence, when exposed to high levels of cadmium (cadmium in renal cortex >100–400 µg/kg wet weight) in the workplace, workers have developed tubular proteinuria, renal glycosuria, aminoaciduria, hypercalciuria, phosphaturia, and polyuria, and in a few severe cases (long-standing high exposure and urinary excretion >20 µg/g creatinine and b $_2$-microglobulin >1500 µg/g creatinine) renal damage may progress to an irreversible reduction in glomerular filtration. Signs of distal tubular damage such as a cadmium-induced inhibition of ADH-stimulated ion transport have also been reported.

The extent to which chronic low-level environmental exposure to cadmium affects renal function is much less clear. The Cadmibel study, in which a random sample of 1699 subjects was recruited from four areas of Belgium with varying degrees of cadmium pollution, showed that (after standardization for several confounding factors) five markers of renal dysfunction (retinol binding protein, *N*-acetyl-b-glucosaminidase, b $_2$-microglobulin, amino acids, and calcium) were significantly associated with urinary cadmium excretion. There was a 10 per cent probability of these variables being abnormal when urinary cadmium levels exceeded 2 to 4 µg/24 h. However, in a 5-year follow-up of a subcohort from the Cadmibel study, the so-called Pheecad study, in which 593 individuals with the highest urinary cadmium excretion were re-examined on average 5 years later, it was demonstrated that the subclinical tubular effects previously documented were not associated with a deterioration in glomerular function. Hence, in the environmentally cadmium-exposed population, the renal effects due to cadmium appear to be weak, stable, and even reversible. These findings in environmentally exposed subjects may reasonably be extrapolated to the current, moderately exposed, occupational population, where, in various epidemiological studies, increased cadmium levels/exposure have repeatedly been associated with disturbed levels of markers of early renal dysfunction, but without evidence for accelerated progression towards chronic renal failure.

Metabolic disorders

Chronic hypokalaemia

Several renal abnormalities, most of which are reversible with potassium repletion, can be induced by hypokalaemia. Vasopressin-resistant impairment of the ability to concentrate the urine, increased renal ammonia production, enhanced bicarbonate reabsorption, altered sodium reabsorption, and hyperkalaemic nephropathy have all been described.

Persistent hypokalaemia can induce a variety of changes in renal function, impairing tubular transport, and possibly inducing chronic tubulointerstitial disease and cyst formation. Hypokalaemic nephropathy in humans produces characteristic vacuolar lesions in the epithelial cells of the proximal tubule and (occasionally) the distal tubule. This abnormality probably requires about 1 month to develop. More severe changes, including interstitial fibrosis, tubular atrophy, and cyst formation that is most prominent in the renal medulla, occur if prolonged hypokalaemia is maintained. The pathogenesis of these changes is not well understood.

Renal growth accelerates when rats are placed on a potassium-deficient diet, and within 8 days there is a 25 per cent increase in kidney mass. The changes are most prominent in the outer medulla, especially the inner stripe, where hyperplastic, enlarged, collecting-duct cells form cellular outgrowths that project into the lumen causing partial obstruction. If the potassium-deficient state persists, then cellular infiltrates appear in the renal interstitial compartment and tubulointerstitial fibrosis develops. It has been proposed that some of these pathological changes may be initiated by the high levels of ammonia generated in potassium-deficiency states and may be mediated through the activation of the alternate complement pathway. In support of this hypothesis is the finding that bicarbonate supplementation sufficient to suppress renal ammoniagenesis attenuates the renal enlargement and tubulointerstitial disease: against it are reports that increased renal ammoniagenesis induced by acid loading causes renal enlargement without cellular proliferation or interstitial disease. A recent paper provides results consistent with a sustained role for insulin-like growth factor-1 (**IGF-1**) in promoting the marked tubular epithelial-cell hypertrophy and hyperplasia that occurs in the inner stripe of the outer medulla of the kidney with chronic potassium depletion. The same study also showed that potassium depletion causes a selective increase in the renal expression of transforming growth factor-b (**TGF-b**) in the hypertrophied, non-hyperplastic, thick ascending limb, but—unlike IGF-1—it is absent from the hyperplastic collecting-duct cells. This might be responsible for preventing the conversion of the mitogenic stimulus of IGF-1 into a hypertrophic one. It is possible that TGF-b causes the prominent interstitial infiltrate that develops in chronic hypokalaemia, since this 'growth factor' is a well-known chemoattractant for macrophages.

Hyperoxaluria

Hyperoxaluria may be primary or acquired. The primary form is a rare inherited disorder due to an enzymatic abnormality in the metabolism of glyoxylic acid. The acquired forms of hyperoxaluria are more common and result either from the ingestion of oxalate precursors, such as ethylene glycol and ascorbic acid, and exposure to methoxyflurane anaesthesia, or from increased absorption from the intestinal tract in those with inflammatory bowel disease or who have undergone small-bowel resection.

The microcrystallization of calcium oxalate first occurs in the proximal tubules where oxalate secretion occurs. However, the lesions that develop are more severe in the renal medulla, where the increasing concentration of the tubular fluid and its acidification promote the precipitation of calcium oxalate. If the overload is insidious and chronic, inflammatory-cell infiltration, oedema, interstitial fibrosis, tubular atrophy, and dilatation result in a chronic tubulointerstitial nephritis with progressive renal failure.

Hypercalcaemia

Prolonged elevation of urinary and serum calcium levels may result in the deposition of calcium in the kidney (nephrocalcinosis, see Chapter 20.13). This also occurs in some clinical conditions not associated with hypercalcaemia. Increased intestinal absorption of calcium occurs with vitamin D intoxication, sarcoidosis, and the milk alkali syndrome. Skeletal deossification due to neoplasms, hyperparathyroidism, and multiple myeloma can also produce nephrocalcinosis, stones, and functional abnormalities.

Calcium is most concentrated in the medulla, where degeneration and tubular necrosis begins due to intracellular overload with damage to mitochondria and other critical organelles. Reactive inflammatory changes occur in the adjacent interstitium, and necrotic cells may cause intratubular obstruction and tubular atrophy. The final results of these changes are focal areas of tubular atrophy, interstitial fibrosis, and a mononuclear-cell infiltrate.

Hyperuricaemia/hyperuricosuria

There are three different types of renal disease induced by abnormal uric acid metabolism: acute uric acid nephropathy; chronic urate nephropathy; and uric-acid stone disease—the latter being discussed in Chapter 20.13.

The kidneys are the major organs for the excretion of uric acid and a primary target organ affected in disorders of urate metabolism. Renal lesions result from the crystallization of uric acid either in the urinary outflow tract or in the renal parenchyma. The determinants of uric acid solubility are its concentration and the pH of the medium in which it is dissolved. Hence the supersaturation of fluid within the renal tubules as excreted uric acid becomes concentrated in the medulla, and the acidification of the urine in the distal tubule, are both conducive to the precipitation of uric acid. The major sites of urate deposition are the renal medulla, the collecting tubules, and the urinary tract. The p*K*a of uric acid is 5.7, and at the acid pH of the fluid in the distal tubule the bulk of filtered urate will be present in its non-ionized form as uric acid, whereas at the more alkaline pH of the blood and interstitium it is in its ionized form as urate salts.

Acute uric acid nephropathy

Acute uric acid nephropathy is an uncommon condition caused by the precipitation of birefringent uric acid crystals in the collecting tubules, with consequent tubular obstruction, dilatation, and inflammation. This can occur in disorders associated with an increased production of uric acid, for example myeloproliferative or lymphoproliferative disorders, tumour lysis syndrome (see Chapter 20.10.5), chronic haemolytic anaemia, psoriasis, or the Lesch–Nyhan syndrome, or when there is increased renal clearance of uric acid, for example inherited or acquired defects of tubular urate transport, uricosuric drugs.

In those prone to acute uric acid nephropathy, management centres on prophylaxis with a plentiful fluid intake, with or without alkalinization of the urine, and pretreatment with allopurinol, although the latter can increase the risk of xanthine nephropathy. Presentation of acute uric acid nephropathy is with acute renal failure, with urine microscopy revealing plentiful birefringent crystals. Management is supportive. If allopurinol is prescribed, then the dose must be substantially reduced in renal failure.

Chronic urate nephropathy

The principal lesion in chronic hyperuricaemia is the deposition of microtophi of amorphous urate crystals in the interstitium, with a surrounding giant-cell reaction. This results in a secondary chronic inflammatory response similar to that seen with microtophus formation elsewhere in the body, potentially leading to interstitial fibrosis and chronic renal failure.

Evidence linking chronic renal failure to gout is weak, and the long-standing notion that chronic renal disease is common in patients with hyperuricaemia has been questioned in the light of prolonged follow-up studies of renal function in people with this condition. Renal dysfunction could be documented only when the serum urate concentration was more than 10 mg/dl (600 µmol/l) in women and more than 13 mg/dl (780 µmol/l) in men for prolonged periods. Furthermore, the deterioration of renal function in those with hyperuricaemia of a lower magnitude has been attributed to the higher-than-expected occurrence of hypertension, diabetes mellitus, abnormal lipid metabolism, and nephrosclerosis. Nonetheless, it seems reasonable to prescribe allopurinol (in a dose appropriate to the level of renal function) to those very rare patients with biopsy evidence of 'gouty nephropathy', and possibly to patients with chronic renal failure who have a grossly elevated serum urate.

There is an association between severe lead intoxication, chronic renal failure, and gout (saturnine gout) (see above). It has also been suggested that there might be an association between renal disease and hyperuricaemia in those with a past history of exposure to lead and consequent subclinical lead toxicity (saturnine nephropathy). Evidence for this association is not clear-cut, nor is the mechanism whereby lead exposure might aggravate hyperuricaemia and renal failure.

Further reading

Bach PH, Hardy TL (1985). Relevance of animal models to analgesic-associated renal papillary necrosis in humans. *Kidney International* **28**(4), 605–13.

Benabe JE, Martinez-Maldonado M (1978). Hypercalcemic nephropathy. *Archives of Internal Medicine* **138**, 777–9.

Bennett WM, De Broe ME (1989). Analgesic nephropathy—a preventable renal disease. *New England Journal of Medicine* **320**, 1269–71.

Bergstein JM (1998). Radiation. In: Davison AM, *et al*., eds. *Oxford textbook of clinical nephrology*, pp 1190–5. Oxford University Press, Oxford.

Bia MJ, Ansognia K (1991). Treatment of sarcoid-associated hypercalcemia with ketoconazole. *American Journal of Kidney Diseases* **18**, 702–5.

Bjerregaard HF, Faurskov B (1997). Cadmium-induced inhibition of ADH-stimulated ion transport in cultured kidney-derived epithelial cells. *Alternatives to Laboratory Animals* **25**, 271–7.

Blohme I, Johansson S (1981). Renal pelvic neoplasms and atypical urothelium in patients with end-stage analgesic nephropathy. *Kidney International* **20**, 671–5.

Boton R, Gaviria M, Battle CD (1987). Prevalence, pathogenesis, and treatment of renal dysfunction associated with chronic lithium therapy. *American Journal of Kidney Diseases* **10**, 329–45.

Brunner FP, Selwood NH (1994). End-stage renal failure due to analgesic nephropathy, its changing pattern and cardiovascular mortality. *Nephrology, Dialysis, Transplantation* **9**, 1371–6.

Buchet JP, *et al*. (1990). Renal effects of cadmium body burden of the general population. *Lancet* **336**, 699–702.

Calder IC, *et al*. (1972). Nephrotoxic lesions from 5-aminosalicylic acid. *British Medical Journal* **1**, 152–4.

Casella FJ, Allon M (1993). The kidney in sarcoidosis. *Journal of the American Society of Nephrology* **3**, 1555–64.

Ceovic S, Hrabar A, Saric M (1992). Epidemiology of Balkan endemic nephropathy. *Food and Chemical Toxicology* **30**, 183–98.

Cohen EP (2000). Radiation nephropathy after bone marrow transplantation. *Kidney International* **58**, 903–18.

Cosyns JP, *et al*. (1994). Chinese herbs nephropathy: a clue to Balkan endemic nephropathy? *Kidney International* **45**,1680–8.

Cosyns JP, *et al*. (1999). Urothelial lesions in Chinese-herb nephropathy. *American Journal of Kidney Diseases* **33**, 1011–17.

Cremer W, Bock KD (1976). Symptoms and course of chronic hypokalemic nephropathy in man. *Clinical Nephrology* **7**, 112–19.

Dafnis E, Kurtzman NA, Sabatini S (1992). Effects of lithium and amiloride on collecting tubule transport enzymes. *Journal of Pharmacology and Experimental Therapeutics* **261**, 701–6.

De Broe ME (1999). On a nephrotoxic and carcinogenic slimming regimen. *American Journal of Kidney Diseases* **33**, 1171–3. [Editorial]

Depierreux M, *et al*. (1994). Pathologic aspects of a newly described nephropathy related to the prolonged use of Chinese herbs. *American Journal of Kidney Diseases* **24**, 172–80.

Diamond JR, Pallone TL (1994). Acute interstitial nephritis following use of Tung Shueh pills. *American Journal of Kidney Diseases* **24**, 219–21.

Djukanovic L, Velimirovic D, Sindjic M (1998). Balkan nephropathy. In: De Broe ME, *et al*., eds. *Clinical nephrotoxins—renal injury from drugs and chemicals*, pp 425–36. Kluwer Academic, Dordrecht.

Dubach UC, Rosner B, Pfister E (1983). Epidemiologic study of abuse of analgesics containing phenacetin. Renal morbidity and mortality (1968–1979). *New England Journal of Medicine* **308**, 357–62.

Dubach UC, Rosner B, Sturmer T (1991). An epidemiologic study of abuse of analgesic drugs. Effects of phenacetin and salicylate on mortality and cardiovascular morbidity. *New England Journal of Medicine* **324**, 155–60.

Duffy WB, Senekjian HO, Knight TF (1981). Management of asymptomatic hyperuricemia. *Journal of the American Medical Association* **246**, 2215–16.

Duggin GG (1996). Combination analgesic-induced kidney disease: the Australian experience. *American Journal of Kidney Diseases* **28**(1 Suppl 1), S39–S47.

Elseviers MM, De Broe ME (1996). Combination analgesic involvement in the pathogenesis of analgesic nephropathy: the European perspective. *American Journal of Kidney Diseases* **28**(1-Suppl. 1): S48–S55.

Elseviers MM, *et al*. (1992). Diagnostic criteria of analgesic nephropathy in patients with end-stage renal failure—results of the Belgian study. *Nephrology, Dialysis, Transplantation* **7**, 479–86.

Elseviers MM, *et al*. (1995). Evaluation of diagnostic criteria for analgesic nephropathy in patients with end-stage renal failure: results of the ANNE study. *Nephrology, Dialysis, Transplantation* **10**,

808–14.

Elseviers MM, *et al*. (1995). High diagnostic performance of CT scan for analgesic nephropathy in patients with incipient to severe renal failure. *Kidney International* **48**, 1316–23.

Farkas WR, Stanawitz T, Schneider M (1978). Saturnine gout: lead-induced formation of guanine crystals. *Science* **199**, 786–7.

Ganote CE, *et al*. (1975). Acute calcium nephrotoxicity. An electron microscopical and semiquantitative light microscopical study. *Archives of Pathology* **99**, 650–7.

Henderson DA (1958). The etiology of chronic nephritis in Queensland. *Medical Journal of Australia* **25**, 196–202.

Hensen J, Haenelt M, Gross P (1996). Lithium induced polyuria and renal vasopressin receptor density. *Nephrology, Dialysis, Transplantation* **11**, 622–7.

Hestbech J, *et al*. (1977). Chronic renal lesions following long-term treatment with lithium. *Kidney International* **12**, 205–13.

Hodgkinson A, Wilkinson R (1974). Plasma oxalate concentration and renal excretion of oxalate in man. *Clinical Science and Molecular Medicine* **46**, 61–73.

Hotz P, *et al*. (1999). Renal effects of low-level environmental cadmium exposure: 5-year follow-up of a subcohort from the Cadmibel study. *Lancet* **354**, 1508–13.

Hu H (1991). A 50-year follow-up of childhood plumbism. Hypertension renal function, and hemoglobin levels among survivors. *American Journal of Diseases of Children* **145**, 681–7.

Inglis JA, Henderson DA, Emmerson BT (1978). The pathology and pathogenesis of chronic lead nephropathy occurring in Queensland. *Journal of Pathology* **124**, 65–76.

Ivic M (1970). The problem of etiology of endemic nephropathy. *Acta Facultatis Medicae Naissensis* **1**, 29–38.

Jensen OM, *et al*. (1989). The Copenhagen case-control study of renal pelvis and ureter cancer, role of analgesics. *International Journal of Cancer* **44**, 965–8.

Johnson RJ, *et al*. (1999). Reappraisal of the pathogenesis and consequences of hyperuricemia in hypertension, cardiovascular disease, and renal disease. *American Journal of Kidney Diseases* **33**, 225–34.

Kabanda A, *et al*. (1995). Low molecular weight proteinuria in Chinese herbs nephropathy. *Kidney International* **48**, 1571–6.

Kido T, Nordberg G (1998). Cadmium-induced renal effects in the general environment. In: De Broe ME, *et al.*, eds. *Clinical nephrotoxins*, pp 345–61. Kluwer Academic Press, Dordrecht.

Kim R, *et al*. (1996). A longitudinal study of low-level lead exposure and impairment of renal function. The normative age study. *Journal of the American Medical Association* **275**, 1177–81.

Kömhoff M, *et al*. (1997). Localization of cyclooxygenase-1 and –2 in adult and fetal human kidney: implication for renal function. *American Journal of Physiology* **272**, 460–8.

Korzets Z, *et al*. (1985). Acute renal failure due to sarcoid granulomatous infiltration of the renal parenchyma. *American Journal of Kidney Diseases* **6**, 250–3.

Lakkis FG, Campbell OC, Badr KF (1996). Microvascular diseases of the kidney. In: Barry M. Brenner BM, eds. *The kidney*, pp 1721–2. WB Saunders, Philadelphia.

Luxton RW (1961). Radiation nephritis: a long-term study of fifty-four patients. *Lancet* **2**, 1221.

Markowitz GS, *et al*. (2000). Lithium nephrotoxicity: a progressive combined glomerular and tubulointerstitial nephropathy. *Journal of the American Society of Nephrology* **11**, 1439–48.

Marples D, *et al*. (1995). Lithium-induced downregulation of aquaporin-2 water channel expression in rat kidney medulla. *Journal of Clinical Investigation* **95**, 1838–45.

McCredie M, Stewart JH (1988). Does paracetamol alone cause urothelial cancer or renal papillary necrosis? *Nephron* **49**, 296–300.

McCredie M, *et al.* (1982). Analgesics and cancer of the renal pelvis in New South Wales. *Cancer* **49**, 2617–25.

McCredie M, *et al*. (1986). Phenacetin and papillary necrosis: independent risk factors for renal pelvic cancer. *Kidney International* **30**, 81–4.

McLeary TJ, *et al*. (1986). The effect of chloroquine on serum 1,25-dihydroxyvitamin D and calcium metabolism in sarcoidosis. *New England Journal of Medicine* **315**, 727–30.

Messerli FH, *et al*. (1980). Serum uric acid in essential hypertension: an indicator of renal vascular involvement. *Annals of Internal Medicine* **93**, 817–21.

Mihatsch MJ, *et al*. (1983). Capillary sclerosis of the urinary tract and analgesic nephropathy. *Clinical Nephrology* **20**(6), 285–301

Moel DI, Sachs HK (1992). Renal function 17 to 23 years after chelation therapy for childhood plumbism. *Kidney International* **42**, 1226–31.

Morlans M, *et al*. (1990). End-stage renal disease and non-narcotic analgesics. A case-control study. *British Journal of Clinical Pharmacology* **30**, 717–23.

Murray MD, Brater DC (1993). Renal toxicity of the nonsteroidal anti-inflammatory drugs. *Annual Review of Pharmacology and Toxicology* **33**, 435–65.

Murray MD, Henrich WL, Stoff JS (1996). The renal effects of nonsteroidal anti-inflammatory drugs: summary and recommendations. *American Journal of Kidney Diseases* **28**(Suppl. 1), S56–S62.

Murray TG, Goldberg M (1978). Analgesic-associated nephropathy in the USA: epidemiologic, clinical and pathogenetic features. *Kidney International* **13**, 64–71.

Muther RS, McCarron DA, Bennett VM (1981). Renal manifestations of sarcoidosis. *Archives of Internal Medicine* **141**, 643–5.

Nanra RS (1993). Analgesic nephropathy in the 1990s: an Australian perspective. *Kidney International* **42**(Suppl. 44), 86–92.

Nanra RS, Kincaid-Smith P (1993). Experimental evidence for nephrotoxity of analgesics. In: Stewart JH, ed. *Analgesic and NSAID induced kidney disease*, pp 17–31. Oxford University Press, Oxford.

Nanra RS, *et al.* (1978). Analgesic nephropathy: etiology, clinical syndrome, and clinicopathologic correlations in Australia. *Kidney International* **13**, 79–92.

Nortier JL, *et al*. (2000). Urothelial carcinoma associated with the use of a Chinese herb. *New England Journal of Medicine* **342**, 1686–92.

Petronic VJ, *et al*. (1991). Balkan endemic nephropathy and papillary transitional cell tumors of the renal pelvis and ureters. *Kidney International* **34**, S77–S79.

Piper JM, Tonascia J, Matanoski GM (1985). Heavy phenacetin use and bladder cancer in women aged 20 to 49 years. *New England Journal of Medicine* **313**, 292–5.

Polenakovic MH, Stefanovic V (1998). Balkan nephropathy. In: Davison AM, *et al*., eds. *Oxford textbook of clinical nephrology*, 2nd edn, pp 1203–9. Oxford Medical Publications, Oxford.

Pommer W, *et al*. (1989). Regular analgesic intake and the risk of end-stage renal failure. *American Journal of Nephrology*, **9**, 403–12.

Reginster F, Jadoul M, van Ypersele de Strihou C (1997). Chinese herbs nephropathy presentation, natural history and fate after transplantation. *Nephrology, Dialysis, Transplantation* **12**, 81–6.

Roels H, *et al*. (1993). Markers of early renal changes induced by industrial pollutants. III Application to workers exposed to cadmium. *British Journal of Industrial Medicine* **50**, 37–48.

Rose BD (1994). *Clinical physiology of acid–base and electrolyte disorders*, 4th edn, pp 802–5. McGraw-Hill, New York.

Sandler DF, *et al*. (1989). Analgesic use and chronic renal disease. *New England Journal of Medicine* **320**, 1238–43.

Schwarz A, *et al*. (2000). The outcome of acute interstitial nephritis: risk factors for the transition from acute to chronic interstitial nephritis. *Clinical Nephrology* **54**, 179–90.

Smilde TJ, *et al*. (1994). Tubulo-interstitiële nefritis door mesalazine (5-ASA)-preparaten. *Nederlands Tijdschrift voor Geneeskunde* **138**, 2557–61.

Staessen JA, *et al*. (1992). Impairment of renal function with increasing blood lead concentrations in the general population. *New England Journal of Medicine* **327**, 151–6.

Staessen JP, *et al*. on behalf of the Working Groups (1996). Public health implications of environmental exposure to cadmium and lead: an overview of epidemiological studies in Belgium. *Journal of Cardiovascular Risk* **3**, 26–41.

Stefanovic V, Polenakovic MH (1991). Balkan nephropathy. *American Journal of Nephrology* **11**, 1–11.

Thompson CS, Weinman EJ (1984). The significance of oxalate in renal failure. *American Journal of Kidney Diseases* **4**, 97–100.

Timmer RT, Sands JM (1999). Lithium intoxication. *Journal of the American Society of Nephrology* **10**, 666–74.

Tollins JP, Hostetter MK, Hostetter TH (1987). Hypokalemic nephropathy in the rat. *Journal of Clinical Investigation* **79**, 1447–58.

Tonceva D, Dimitrov T, Tzoneva M (1988). Cytogenetic studies in Balkan endemic nephropathy. *Nephron* **48**, 18–21.

Torres VE, *et al*. (1990). Association of hypokalemia, aldosteronism, and renal cysts. *New England Journal of Medicine* **322**, 345–51.

Tsao T, *et al*. (2001). Expression of insulin-like growth factor-1 and transforming growth factor-b in hypokalemic nephropathy in the rat. *Kidney International* **59**, 96–105.

Vanherweghem JL (2000). Nephropathy and herbal medicine. *American Journal of Kidney Diseases* **35**, 330–2.

Vanherweghem JL, *et al*. (1993). Rapidly progressive interstitial fibrosis in young women: association with slimming regimen including Chinese herbs. *Lancet* **341**, 387–91.

Vanherweghem JL, *et al*. (1996). Effects of steroids on the progression of renal failure in chronic interstitial renal fibrosis: a pilot study in Chinese herbs nephropathy. *American Journal of Kidney Diseases* **27**, 209–15.

Viero RM, Cavalla T (1995). Granulomatous interstitial nephritis. *Human Pathology* **26**, 1347–53.

Walker RG, *et al*. (1982). Structural and functional effects of long-term lithium therapy. *Kidney International* **21**(Suppl. 11), S13–S19.

Walker RG, *et al*. (1982). A clinicopathological study of lithium nephrotoxicity. *Journal of Chronic Disease*, **35**, 685–95.

Wedeen RP, Batuman V (1983). Tubulointerstitial nephritis induced by heavy metals and metabolic disturbances. *Contemporary Issues in Nephrology* **10**, 211.

Wedeen RP, Mallik DK, Batuman V (1979). Detection and treatment of occupational lead nephropathy. *Archives of Internal Medicine* **139**, 53–7.

Wolf ME, *et al*. (1997). Lithium therapy, hypercalcemia, and hyperparathyroidism. *American Journal of Therapeutics* **4**, 323–5.

World Health Organization (1992). *Cadmium (environment health criteria 134)*, pp 174–88. World Health Organization, Geneva.

World MJ, *et al*. (1996). Mesalazine-associated interstitial nephritis. *Nephrology, Dialysis, Transplantation* **11**, 614–21.

Yang CS, *et al*. (2000). Rapidly progressive fibrosing interstitial nephritis associated with Chinese herbal drugs. *American Journal of Kidney Diseases* **35**, 313–18.

20.10.1 Diabetes mellitus and the kidney

R. W. Bilous

Introduction

Diabetic nephropathy is the commonest single cause of endstage renal failure (**ESRF**) requiring renal replacement therapy in the United States, and the second most common in Europe and Japan. The incidence is increasing, largely because the incidence of diabetes itself is reaching what some have termed epidemic proportions, this growth being greatest in the developing world.

Definition

Nephropathy is a clinical diagnosis based upon the finding of proteinuria in a patient with diabetes and in whom there is no evidence of urinary infection. Conventionally, the level of proteinuria for a diagnosis of 'clinical nephropathy' or 'overt nephropathy' is 0.5 g/day, which is roughly equivalent to a urinary albumin excretion rate (**UAER**) of 300 mg/day. Patients with a UAER between 30 and 300 mg/day are defined as having 'microalbuminuria' or 'incipient nephropathy'. In this chapter, the terms 'incipient' and 'clinical nephropathy' will be used.

Although timed urine collections remain the 'gold standard' for diagnosis, they are cumbersome to use in routine clinical practice and most definitions of clinical or incipient nephropathy depend upon a 'spot' urine sample and thus a test of albumin concentration. Results in excess of 300 mg/l and more than 50 mg/l define clinical and incipient nephropathy, respectively. Sensitivity and specificity can be improved by using an early morning, first-voided specimen and correcting the albumin level for creatinine concentration (albumin:creatinine ratio (**ACR**)). Defining levels are shown in Table 1.

Pathology

Patients with newly diagnosed type 1 disease have large kidneys. Studies in experimental animals suggest that this enlargement is due to tubular hypertrophy and hyperplasia and an expansion of the tubulointerstitium. These changes are probably in response to the increased filtration of glucose and can be reversed in animals, but not man, by glycaemic correction. Otherwise, glomerular and tubular structure is normal at diagnosis in patients with type 1 diabetes.

The pathological hallmarks of diabetic nephropathy are thickening of the glomerular basement membrane (**GBM**) and mesangial expansion with or without nodule formation. GBM thickening can be detected in nearly all patients with diabetes of more than 10 years' duration, irrespective of the UAER. Those with clinical nephropathy almost invariably have GBM widths two to three times the upper limit of normal (350 nm). Mesangial volume remains in the normal range in patients who have a normal UAER. Nodule formation, although virtually pathognomonic, is not invariable. A combination of mesangial expansion and afferent arteriolar hyalinosis with ischaemia leads to eventual total glomerulosclerosis and subsequent loss of filtration capacity, ultimately leading to ESRF.

Patients with type 2 diabetes have been much less well studied, but the pathological appearances of subjects with clinical nephropathy are very similar to those with type 1. However, the pattern of changes in incipient nephropathy is more heterogeneous and a significant prevalence of non-diabetic pathology (around 10 per cent) has been reported in some biopsy series.

Clinical course

The pathological lesions underpinning nephropathy have been recognized since 1936. However, clinical progression is usually defined in terms of changes in the UAER, glomerular filtration rate (**GFR**), and blood pressure. There are few long-term prospective studies of individual patients, and much of our current understanding is based upon cross-sectional data. Albuminuria is clearly a continuous variable and any separation into stages must be regarded as somewhat artificial. However, the distinction between incipient and clinical nephropathy is a useful one for practical purposes.

Urinary albumin excretion rate (UAER)

UAER may increase at diagnosis of type 1 diabetes and during acute hyperglycaemia, but it rapidly returns to normal with glycaemic correction. Thereafter the majority of patients (>60 per cent) will have a normal UAER throughout their diabetic life. The remainder will develop incipient nephropathy at incidence rates of between 1 and 2 per cent per annum, usually preceded by intermittently positive tests for microalbuminuria. The rate of increase of UAER in patients with incipient nephropathy is around 20 per cent per annum, although these rates will be less in those treated with antihypertensive therapy or intensified insulin regimens (see later).

It is unusual to develop incipient nephropathy within the first 5 years of diabetes onset, but it can develop at any time thereafter, even after 40 years. Most patients with type 1 disease and incipient nephropathy will progress to clinical nephropathy unless treated. The average rate of progression is 20 per cent over 5 years, but those with longer durations of diabetes prior to incipient nephropathy tend to progress more slowly. Once the UAER exceeds 300 mg/day there tends to be a relentless increase, sometimes into the nephrotic range, although the rate of change varies between patients and is very dependent upon systemic blood pressure. Historically, the incidence of clinical nephropathy peaks after diabetes of 15 to 17 years' duration, but these figures were obtained before the availability of effective antihypertensive therapy.

Because the onset of type 2 diabetes is more difficult to define, the precise incidence of incipient nephropathy is difficult to estimate. Up to 7 per cent of patients in the United Kingdom have incipient nephropathy and 1 per cent clinical nephropathy at diagnosis of diabetes. As in type 1 diabetes, the UAER will reduce with glycaemic

correction, but usually only in a minority of patients, therefore implying an established nephropathology. The United Kingdom Prospective Diabetes Study (**UKPDS**) in newly diagnosed patients reports rates of transition from normal to incipient nephropathy of 2 per cent per annum and from incipient to clinical nephropathy of 3 per cent per annum, which are very similar to those seen in patients with type 1 diabetes.

Glomerular filtration rate (GFR)

At diagnosis, the GFR is increased in a proportion of patients with type 1 and type 2 diabetes, a phenomenon termed hyperfiltration. The precise prevalence depends upon the definition of 'normal GFR', which varies with methodology and age, but a raised GFR has been described in up to 40 per cent of patients with untreated, non-ketotic type 1 diabetes and in 45 per cent of those newly diagnosed as type 2.

The GFR returns to normal in most patients with treatment, although a significant minority maintain persistent hyperfiltration. In normal people, the GFR declines by 1 per cent per annum beyond 40 years of age, a rate that is no different in normotensive diabetic patients with a normal UAER. Patients with type 1 and type 2 disease with incipient nephropathy also tend to have a stable GFR. However, as the UAER approaches and exceeds the clinical nephropathy threshold there is a steady decline. The rate of loss of GFR is very dependent on systemic blood pressure, and also varies considerably between individuals. In historical series of hypertensive type 1 and type 2 patients, the average decline was 10 ml/min per year, thus leading to ESRF within 7 to 10 years. In patients with well-controlled blood pressure, the rates are around 4 ml/min per year, effectively delaying ESRF by more than 10 years (Table 2).

Blood pressure

In patients with type 1 diabetes, their blood pressure is virtually always normal at diagnosis. In newly diagnosed type 2 patients, over one-third will have hypertension, as conventionally defined (>160/95 mmHg). Type 1 patients who go on to develop incipient nephropathy have significantly higher blood pressures than those who remain with a normal UAER, although the averages remain within the 'normal' range. Patients with newly developed incipient nephropathy show a steady increase thereafter, such that over 45 per cent have a blood pressure of more than 140/90 mmHg within 4 years. Cross-sectional studies also consistently show higher blood pressures with increasing UAER, most type 1 and type 2 patients with clinical nephropathy will be hypertensive.

Clinical concomitants of nephropathy

Most patients with nephropathy will also have retinopathy and neuropathy (the so-called 'triopathy' of microvascular complications), which will also tend to progress along with nephropathy.

However, the most serious comorbid complication of nephropathy is macrovascular disease. The reported relative mortality for European 40-year-old, type 1 patients with clinical nephropathy in Denmark was between 80 and 100 times that of the non-diabetic population, whereas the World Health Organization (**WHO**) study revealed a three- to fourfold excess for patients with type 2 disease. Most of these deaths are due to stroke or myocardial infarction, and in Finland type 1 patients with nephropathy have a 10-fold relative risk for both compared to non-diabetic controls. Similar increases, but of lower magnitude, are seen in type 1 and 2 patients with incipient nephropathy. The reasons are unclear, but increasing blood pressure is certainly a factor, together with the unfavourable blood lipid profile found in patients with an increased UAER.

Non-nephropathic renal disease in diabetes

It is important to remember that not all renal or urinary tract disease in diabetic patients is due to 'diabetic nephropathy'. Urinary tract infection is more common in diabetic women, many of whom are asymptomatic. Urine culture should always be performed in patients with an isolated positive urinalysis for protein. Papillary necrosis is also more common in women with longstanding type 1 diabetes, and is a recognized complication of hyperosmolar coma in patients with type 1 and type 2 disease. Atheromatous vascular disease is common in diabetics and can cause renal artery stenosis, but the prevalence of functionally significant stenosis is not known.

Epidemiology

The incidence of diabetes is increasing worldwide, most rapidly in developing countries. It is estimated that by 2010 there will be a near-doubling of people with the condition to 221 million (5 million with type 1 and 216 million with type 2 disease).

Some of the wide variation in the reported prevalence of incipient and clinical nephropathy can be explained by different selection of the population under study, and by the use of different defining levels of UAER. However, selecting only population-based cohorts with good patient ascertainment, gives prevalence rates for incipient nephropathy of between 5 and 21 per cent for type 1 and 11 to 42 per cent for type 2 disease. Annual incidence rates are similar at around 2 per cent for both type 1 and type 2 patients.

For clinical nephropathy, the prevalence is 6.4 per cent in type 1 patients in the United Kingdom, with a range from 5 to 33 per cent reported worldwide for type 2. A cumulative incidence of approximately 20 per cent after 20 years' duration was found in the type 1 diabetic cohorts of the Steno Hospital in Denmark and Joslin Clinic in the United States, and similar figures have been reported for patients with type 2 diabetes in the United States (25 per cent) and Germany (27 per cent). Although reported annual incidence rates vary from 0.4 to 3.6 per cent, these are highly dependent upon the duration of diabetes in the population under study.

The Steno and Joslin cohorts have shown a 20 to 30 per cent lower incidence of clinical nephropathy in those type 1 patients diagnosed with diabetes in the 1950s and 1960s, compared to those diagnosed 20 years earlier. A single clinic in Sweden reports no patients with clinical nephropathy after 15 years' disease duration in patients diagnosed between 1976 and 1980, compared to 15 per cent and 5 per cent in patients with diabetes onset from 1961 to 1965 and 1966 to 1975, respectively. This reduction has not been seen in other clinics such as the Steno Hospital, for example. There are fewer data in patients with type 2 diabetes, but a recent study from the United States suggests that proteinuria is less common at diagnosis of diabetes, but that its incidence thereafter has not changed in the last 20 years.

Many countries now have registers of patients entering renal replacement therapy and all have shown a dramatic increase in the numbers with diabetic nephropathy. It is not clear, however, whether this is a true increase in the numbers of diabetic patients developing ESRF or a reflection of a change in acceptance policy. Either way, there is going to be a continuing increase in the number of patients with diabetes presenting for renal replacement, particularly from ethnic minorities (Afro-Caribbean and South African), who will make up around 50 per cent of such patients in the United Kingdom by 2001. The reasons for the excess risk of ECRF in these groups is unclear but may be genetic, related to hypertension, or the result of fetal programming. Although patients with type 2 disease have always been thought to develop ESRF less frequently than those with type 1, this may have been because such patients were not referred, or that they died of cardiovascular disease before entering renal failure. In 1997, 71 per cent of all diabetic patients on dialysis in the United States were classified as having type 2 disease.

Pathogenesis

Glycaemia

Observational studies have shown that sustained poor glycaemic control is associated with a greater risk for the development of nephropathy in both type 1 and type 2 diabetes. There are several potential mechanisms by which hyperglycaemia may cause nephropathy. These are common to all the microvascular complications of diabetes and are reviewed in Section 12.11.

Haemodynamic factors

Studies of experimental diabetes in the rat suggested that hyperfiltration alone could cause glomerulosclerosis. The evidence in humans is conflicting, and not helped by differing definitions of an abnormally high GFR. It appears that the rate of decline of GFR in hyperfiltering type 1 patients with a normal UAER is greater than that seen in age- and duration-matched normal GFR controls, but the numbers that go on to develop incipient nephropathy are similar in both groups. In Pima Indians with type 2 diabetes, their baseline GFR is not linked to the subsequent development of incipient or clinical nephropathy.

Growth factors

In experimental animals, the initial increase in kidney size is preceded by an increase in renal production of insulin-like growth-factor 1, and there are reports of increased circulating and urinary levels in diabetic people. However, there is no conclusive link between the initial kidney size or the renal expression of growth factors and the subsequent development of nephropathy in humans.

Hypertension

Systemic blood pressure is higher in patients with type 1 diabetes who subsequently develop incipient nephropathy. In Japanese people and Pima Indians, a prediabetic mean arterial pressure higher than 97 mmHg (>130/80 mmHg) strongly predicts the development of proteinuria. A family history of hypertension has been found in one study of type 1 patients with nephropathy, but not by another.

The situation in Europid type 2 diabetes may be different. In the UKPDS, hypertension (defined as above 160/90 mmHg or above 150/85 on treatment) was present in over 30 per cent of newly diagnosed patients, only a third of whom had increased albuminuria. Cohorts of normotensive (below 140/90 mmHg) type 2 patients with incipient nephropathy from Israel, Japan, and India showed little change in blood pressure over 7, 4, and 5 years, respectively, despite an increase in their UAER over this time.

It is not clear whether the observed changes in blood pressure initiate the nephropathic process or occur as a result of it. What is certain is that progression of nephropathy is much faster in patients with higher systemic blood pressures.

Genetics

There is a greater than 80 per cent concordance for nephropathy and a more than 73 per cent concordance for normal UAER in siblings of patients with type 1 diabetes. In Pima Indians, the prevalence of nephropathy is 14 per cent in the offspring of parents neither of whom have nephropathy, compared to 46 per cent of offspring when both parents have the condition. These observations have led to many studies looking for a possible genetic cause of nephropathy, most of which have used the candidate-gene approach. The most intensively studied genetic abnormality has been the insertion/deletion polymorphism in the angiotensin-converting enzyme (**ACE**) gene, but the results are inconsistent. Other candidate genes that have been studied include polymorphisms of the angiotensinogen, angiotensin-II type 1 receptor, collagen type IV, and aldose reductase genes: results have been variable.

Alterations in cell-membrane ion transporters have been described in patients with nephropathy, notably the red cell sodium–lithium exchanger and the sodium–hydrogen antiporter. Increased activity of the sodium–lithium exchanger is linked to increased blood pressure in non-diabetic subjects, and the sodium–hydrogen antiporter is closely linked to cell responses to growth factors, both of which have a degree of heritability. There is continuing controversy about the importance of these abnormalities in diabetic nephropathy.

Mechanical and structural factors

In experimental diabetes, intraglomerular capillary pressure is closely linked to the development of glomerulosclerosis. It is not possible to measure intraglomerular pressure directly in humans, but mathematical modelling suggests an increase in early nephropathy.

Glomerular volume is increased at diagnosis of type 1 diabetes and is a feature of clinical nephropathy in both type 1 and type 2 disease. A link between baseline glomerular size and subsequent progression to sclerosis has been described in patients with minimal-change disease, but the connection in diabetes is not proven.

Reductions of heparan sulphate proteoglycan in the intercellular matrix of diabetic patients and the GBM of those with incipient and clinical nephropathy have been reported. Workers at the Steno Hospital in Denmark suggested that this might relate to increased tissue damage in the micro- and macrovasculature, but this interesting hypothesis remains unproven.

Fetal programming

The low birth weight-thrifty phenotype hypothesis proposes intrauterine malnutrition as a possible cause of adult hypertension and diabetes, perhaps mediated via reduced numbers of renal glomeruli or islets of Langerhans, respectively. Studies have failed to find lower glomerular numbers in diabetic patients with nephropathy compared to those without, and no consistent correlation between birth weight and adult glomerular number has been found.

Other factors

Smoking rates are higher in diabetic patients with nephropathy, although a plausible mechanism of effect has yet to be defined. A link between raised blood lipids and the causation and progression of renal disease is still hotly debated. Both cross-sectional and prospective studies have shown an association between plasma total cholesterol and triglyceride levels and the UAER, but not between plasma lipids and a change in GFR.

In experimental diabetes, dietary protein restriction can prevent glomerulosclerosis. The cross-sectional EURODIAB study of patients with type 1 diabetes found that the UAER increased in patients with a protein intake of more than 20 per cent of their total food energy, whilst in the Hoorn Study of type 2 diabetic and normal subjects, a 0.1 g/kg body weight per day increase in protein intake was associated with a greater risk of developing microalbuminuria.

Finally, abnormalities of endothelial function, assessed by increases in plasma von Willebrand factor and homocysteine levels, have been described in diabetic patients with initially normal albuminuria who go on to develop incipient nephropathy, as well as in those with a persistently increased UAER at baseline. The EURODIAB investigators suggest that endothelial dysfunction provides a unifying hypothesis of micro- and macrovascular disease in diabetes.

Investigation

The diagnostic criteria for incipient and clinical nephropathy have already been discussed: the choice of urine sample (either timed or spot) and test (either absolute concentration or corrected for creatinine) depend largely upon local factors and patient acceptability. A diagnostic cascade is shown for a single (spot) urine sample in Fig. 1.

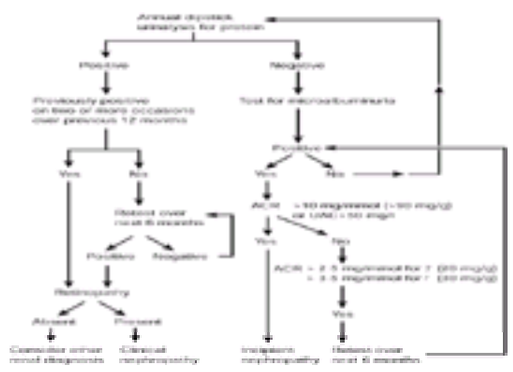

Fig. 1 Flowchart for the diagnosis of incipient and clinical nephropathy. (NB Assumes sterile urine throughout. Exclude infection when proteinuria is first detected and at any time thereafter if there is a history of a urinary tract infection.) ACR, albumin:creatinine ratio; UAC, urine albumin concentration.

Because diabetes is so common, particularly in the elderly, other common nephropathies and uropathies must be excluded if the clinical picture is atypical (for example, renal impairment in the absence of significant proteinuria). In addition, non-diabetic glomerular disease should be considered in proteinuric patients without retinopathy, and in those with an unexpectedly rapid deterioration in renal function, or in whom there are features of other systemic disease. Renovascular disease should be considered whenever an acute increase in plasma creatinine follows initiation of ACE inhibitor therapy, or in those with hypertension that is refractory to treatment.

Treatment

Most studies of treatments for diabetic nephropathy have used surrogate endpoints of efficacy such as a reduction in the UAER. Few have had sufficient statistical power to determine whether the therapy under investigation prevents death, reduces the number of patients entering ESRF, slows the rate of decline of GFR, or slows or reverses the progression of the pathological lesions underpinning nephropathy.

Glycaemic control

Up until 1993, there had been several well-planned, but relatively small, studies of the impact of intensified glycaemic control on the development of nephropathy in type 1 diabetes, with meta-analysis confirming significant benefit. Later that year the definitive Diabetes Control and Complications Trial (**DCCT**) showed that 9 years of intensive (mean Hb A_{1c} 7.2 per cent (normal <6.05 per cent)) compared to conventional insulin therapy (Hb A_{1c} 9.1 per cent) produced a 44 per cent reduction of development of incipient nephropathy (UAER >40 mg/day) in patients with no retinopathy, and 35 per cent reduction in patients with early retinopathy at entry (Table 3). However, the cumulative incidence of incipient nephropathy was still 15 and 27 per cent in the intensively treated patients in the two cohorts.

In patients with type 2 diabetes, there has been one small study from Japan (*n* = 110) and the much larger UKPDS cohort of 3867 newly diagnosed patients of mixed ethnicity. The Japanese study used an almost identical protocol to that of the DCCT, and showed a reduction in incipient nephropathy (UAER >30 mg/day) in the intensively treated group from 28.0 to 7.7 per cent (*p* = 0.032) in those without and 32.0 to 11.5 per cent (*p* = 0.044) in those with retinopathy at entry. In the UKPDS, the percentage of patients developing incipient nephropathy (urinary albumin concentration >50 mg/l) was lower at 19.2 versus 25.4 per cent (*p* <0.001) in the intensively treated cohort after 9 years, representing a risk reduction of 24 per cent (Table 3).

There is continuing controversy as to whether intensive glucose control alone can prevent the progression of incipient to clinical nephropathy. Careful analysis shows that of the 73 patients with a UAER over 40 mg/day at entry into the DCCT, equal numbers developed clinical nephropathy in both the intensive and conventional groups. The UKPDS also found no impact of intensive therapy on the development of a urinary albumin concentration above 300 mg/l (4.4 versus 6.5 per cent at 9 years, relative risk (99 per cent confidence interval (**CI**)) 0.67 (0.42 to 1.07). It therefore seems that once the UAER exceeds 30 to 40 mg/day, other factors (such as blood pressure control) are of more importance for progression.

There are no conclusive data on the impact of improved glycaemic control on the development of endstage renal disease, GFR progression, or death in patients with type 1 diabetes. The UKPDS did show a positive benefit of intensive therapy on the rate of doubling of serum creatinine at 12 years (0.91 versus 3.52 per cent, *p* <0.003) in patients with type 2 diabetes. Pancreas transplantation in type 1 patients has demonstrated that long-term (10 years) glycaemic normalization can reverse established pathological changes in glomeruli. Thus glomerulopathy may take as long to reverse as it does to develop. Notwithstanding this, it is important to remember that good glycaemic control is of proven benefit for retinopathy and therefore still an important goal of treatment in patients with nephropathy.

Blood pressure control

There have been many more studies of antihypertensive therapy than of improved glycaemic control in diabetic nephropathy. For clarity, these will be dealt with under three headings: primary prevention (of incipient nephropathy); secondary prevention (of clinical nephropathy); and tertiary prevention (of endstage renal disease and death).

Primary prevention

The EUCLID study showed that a UAER between 5 and 20 µg/min could be reduced by 2 years' treatment with the ACE inhibitor lisinopril in normotensive (systolic blood pressure <155, diastolic 75–90 mmHg) type 1 patients. There are no comparable studies in patients with type 2 diabetes, but short-term reductions in the UAER have been demonstrated. In the blood pressure control arm of the UKPDS, the percentage of hypertensive patients developing incipient nephropathy at 6 years was 2.3 per cent in the tight (average blood pressure 144/82 mmHg) and 12.5 per cent in the less tight (average blood pressure 154/87 mmHg) control groups (*p* <0.009). This benefit was seen whether the main treatment was with ACE inhibitors or b-blockers.

Secondary prevention

All studies falling into this category have shown a short- to medium-term benefit of all antihypertensive therapies on UAER in the incipient nephropathy range, although, as a general rule, ACE inhibitors seem to be more effective.

In an attempt to explore whether ACE inhibitors are uniquely beneficial, investigators have tried to select normotensive patients and compare active treatment to placebo. In mainly European patients with type 1 diabetes with a mean entry blood pressure of 122/77 mmHg, a combined analysis of one European and one American study (total *n* = 225) showed an adjusted risk reduction of 63 per cent (95 per cent CI, 16–84 per cent; *p* = 0.017) for the development of clinical nephropathy comparing captopril 100 mg/day with placebo. Three smaller studies in normotensive (<140/90 mmHg) type 2 patients have reported a similar reduction in the rate of development of clinical nephropathy. In hypertensive type 2 patients, the angiotensin II receptor blocker irbesartan also reduced the rate of progression from incipient to clinical nephropathy by around two-thirds. The much larger Heart Outcomes Prevention Evaluation (**HOPE**) study demonstrated fewer patients progressing from incipient (albumin:creatinine ratio (ACR) >2 mg/mmol) to clinical nephropathy when treated with 10 mg ramipril. Thus blockade of the renin–angiotensin system by any means appears to confer benefit. Accurate data on GFR are not given in these studies, but in type 1 hypertensive patients, long-term ACE inhibitor therapy appears to stabilize renal function. Interpretation of these studies is difficult: actively treated patients have nearly always had significantly lower blood pressures than the placebo groups, and whilst mathematical corrections for these differences can be applied, the magnitude of the biological consequences of blood pressure reduction cannot be precisely determined.

Tertiary prevention

Studies in the early 1980s established that lowering blood pressure in hypertensive (>160/95 mmHg) type 1 patients with clinical nephropathy resulted in a more than 50 per cent reduction in the UAER and a significant slowing down of the rate of decline of the GFR from 10 to 3 ml/min per year. In those with normal blood pressure, the Collaborative Study Group Trial in type 1 diabetes compared the addition of captopril 100 mg per day to placebo in 409 patients with nephropathy and an entry blood pressure of less than 140/90 mmHg. A significant reduction (35 versus 78 per cent, *p* <0.001) in the numbers of patients doubling their baseline serum creatinine concentration was seen in the captopril-treated patients, although this significance was confined to those with an entry serum creatinine concentration of more than 133 µmol/l (1.5 mg/dl). A similar reduction of 30 versus 70 per cent (*p* = 0.002) in the combined endpoint of death or the need for renal replacement therapy was also seen in the same group.

In patients with type 2 diabetes, the results are less consistent and complicated by the greater frequency of cardiovascular comorbidity. In those studies of more than 2 years' duration, all show sustained reductions in proteinuria but a variable, although usually consistent, slowing of the rate of GFR decline. Two recently reported studies using angiotensin II receptor blocking agents in patients with clinical nephropathy have shown a reduction in the rate of doubling of serum creatinine of between 25 and 33 per cent, less than that seen in type 1 patients but significantly better than that observed in patients randomized to the calcium channel blocker amlodipine. Taken together the studies in type1 and 2 patients support the use of drugs which block the renin–angiotensin system as first-line therapy in both incipient and clinical nephropathy.

Non-renal outcomes

Several large studies of the effect of antihypertensive therapy on cardiovascular mortality and morbidity have been published; many have included sizeable cohorts of

diabetic patients, although rarely specifying their nephropathic status. All have shown that low blood pressure is associated with a reduction in overall mortality and stroke incidence, although the effect on myocardial infarction is inconsistent. Diabetic patients on the whole had greater benefit, with no clear-cut advantage from any specific drug.

Treatment targets

There is uncertainty as to what should be the target blood pressure for diabetic patients, irrespective of their nephropathy status. The recommendation of the British Hypertension Society is 140/85 mmHg, and 130/85 mmHg from the American Diabetes Association. Achieving these targets is difficult, particularly in patients with type 2 disease, and especially for systolic hypertension. In the UKPDS by year 9, 27 per cent of 758 patients in the tight blood pressure control arm were on three or more drugs, their average blood pressure was 144/82 mmHg, and 44 per cent had levels above 150/85. However, as with blood glucose levels, it seems that any reduction confers benefit and the lowest achievable (and tolerated) is probably the best target.

Other treatments

Low-protein diets have been shown by meta-analysis to slow the rate of decline of GFR in diabetic patients. Current dietary recommendations are for an intake of between 0.7 and 0.9 g protein/kg body weight per day.

Aspirin in a dose of 325 mg/day reduced myocardial infarction (relative risk (**RR**) 0.72, 99 per cent CI 0.55–0.95) in 3711 type 1 and 2 patients with retinopathy. Although nephropathy status was not determined in this study, aspirin use is advised for all patients with an increased UAER (unless contraindicated) because of their high risk of cardiovascular disease. Similarly, lipid-lowering therapy should also be considered, but there are no data on the impact of such treatment on the progression of nephropathy.

Experimental therapies include agents that inhibit the formation of advanced glycation endproducts (such as aminoguanidine), and third-generation aldose reductase inhibitors. Antagonists of endothelin and neuroendopeptidases are in advanced phase III studies.

Treatment of endstage renal disease

Diabetic patients do less well on all modalities of renal replacement therapy than their non-diabetic counterparts, with higher cardiovascular mortality. The European survival figures for diabetic patients from 1983 to 1992 are dismal: 23 per cent alive at 5 years, compared to 56 per cent of non-diabetic patients. In the United States, the death rate for never-transplanted, 20- to 44-year-old diabetic patients from 1993–1995 was 160.7 per 1000 patient-years, compared to 83.3 for non-diabetic subjects. Overall survival is best in those with a successful kidney transplant, and recent data from the United States estimates an increased survival of 11 years for diabetic patients receiving an allograft compared to those who remain on dialysis.

Management strategy

Because the outcomes of patients with diabetic nephropathy are so poor, many national guidelines now suggest a multiple risk-factor approach to management. These initiatives have been given impetus by the observation that many patients referred to renal units in Europe have: inadequate blood pressure control; low use of therapies of proven benefit in heart and kidney disease (for example b-blockers, ACE inhibitors, lipid-lowering therapy, low-dose aspirin); and poor assessment of comorbidities such as retinopathy and foot care.

The St Vincent Declaration Taskforce recommends that all patients with nephropathy need to be referred to a nephrologist when their plasma creatinine concentration exceeds 200 µmol/l (2.2 mg/dl). However, before this, many are now proposing multiple cardiovascular risk-factor reduction, aiming for:

- a target blood pressure at least below 140/80 mmHg;
- a target total cholesterol concentration below 5 mmol/l;
- ACE inhibitor use as part of, or in addition to, antihypertensive therapy;
- low-dose aspirin; and
- invasive investigation and correction of coexistent cardio-, cerebro-, and peripheral arterial disease.

Screening for nephropathy

Because of the strong associations between an increased UAER and cardiovascular disease, a case for screening for diabetic nephropathy can be made with some confidence, although the evidence base for beneficial intervention at lower levels of albuminuria is less secure. Current recommendations from national diabetes associations recommends at least annual screening based upon the diagnostic flowchart shown in Fig. 1. Extrapolating the known effects of ACE inhibitors on a reduction of UAER to the prevention of clinical nephropathy and thus endstage renal disease, several authors show a potential cost benefit from the early use of these drugs. However, only long-term prospective studies of primary prevention can conclusively answer this question.

Further reading

Useful reference texts

Alberti KGMM, *et al.*, eds (1997). *International textbook of diabetes mellitus*, 2nd edn. Wiley, Chichester.

Mogensen CE, ed. (2000). *The kidney and hypertension in diabetes mellitus*, 5th edn. Kluwer Academic, Boston, MA.

Ritz E, Rychlik I, eds (1999). *Nephropathy in type 2 diabetes,* Oxford University Press, Oxford.

Guidelines for the management of diabetes and its complications

American Diabetes Association: clinical practice recommendations 2001. *Diabetes Care* **24**(Suppl. 1), S1–S133.

European Diabetes Policy Group (1998). A desktop guide to type 1 (insulin-dependent) diabetes mellitus. *Experimental and Clinical Endocrinology and Diabetes* **106**, 240–69.

European Diabetes Policy Group (1999). A desktop guide to type 2 diabetes mellitus. *Diabetic Medicine* **16**, 716–30.

Ramsay LE, *et al*. (1999). British Hypertension Society guidelines for hypertension management 1999: summary. *British Medical Journal* **319**, 630–5.

Epidemiology

Amos AF, *et al*. (1997). The rise in global burden of diabetes and its complications: estimates and projections to the year 2010. *Diabetic Medicine* **14**(Suppl. 5), S7–S85.

Causes of diabetic nephropathy

Cooper ME (1998). Pathogenesis, prevention and treatment of diabetic nephropathy. *Lancet* **252**, 213–9.

Krolewski AS (1999). Genetics of diabetic nephropathy: evidence for major and minor gene effects. *Kidney International* **55**, 1582–96.

Lee HB, Ho H (1997). Experimental approaches to diabetic nephropathy. *Kidney International* **51**(Suppl. 60), S1–S103.

Mogensen CE (1999). Microalbuminuria, blood pressure and diabetic renal disease: origin and development of ideas. *Diabetologia* **42**, 263–85.

Clinical trials

Brenner B *et al*. (2001). Effects of Losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *New England Journal of Medicine* **345**, 861–9

DCCT (Diabetes Control and Complications Trial) Research Group (1993). The effect of intensive treatment of diabetes on the development and progression of long term complications of insulin dependent diabetes mellitus. *New England Journal of Medicine* **329**, 977–86.

Gaede P *et al*. (1999). Intensified multi-factorial intervention in patients with type 2 diabetes mellitus and microalbuminuria: the Steno type 2 randomised study. *Lancet* **353**, 617–22.

HOPE (Heart Outcomes Prevention Evaluation) study investigators (2000). Effects of ramipril on cardiovascular and microvascular outcomes in people with diabetes mellitus: results of the HOPE Study and micro-HOPE Sub-Study. *Lancet* **355**, 253–9.

Keane WF, Brenner BM, Kurokawa H (1997). Progression of renal disease: clinical patterns, therapeutic options, and lessons from clinical trials. *Kidney International* **52**(Suppl. 63), S32–S53.

Lewis EJ *et al*. (2001). Reno-protective effect of the angiotensin receptor antagonist Irbesartan in patients with nephropathy due to type 2 diabetes. *New England Journal of Medicine* **345**, 851–60.

Parving H-H *et al.* (2001). The effect of Irbesartan of the development of diabetic nephropathy in patients with type 2 diabetes. *New England Journal of Medicine* **345**, 870–8.

The ACE Inhibitors in Diabetic Nephropathy Trialist Group (2001). Should all patients with type 1 diabetes mellitus and microalbuminuria receive angiotensin converting enzyme inhibitors? A meta-analysis of individual patient data. *Annals of Internal Medicine* **134**, 370–9.

UKPDS (UK Prospective Diabetes Study) Group (1998). Intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* **352**, 837–53.

Viberti GC (1994). Outcome variables in the assessment of progression of diabetic kidney disease. *Kidney International* **45**(Suppl. 45), S121–S124.

Wang P-H, Lau J, Chalmers TC (1993). Meta-analysis of effects of intensive blood glucose control on late complications of type 1 diabetes. *Lancet* **34**, 1306–9.

Screening for nephropathy

Bilous RW (1996). Early diagnosis of diabetic nephropathy. *Diabetes and Metabolism Reviews* **12**, 243–53.

20.10.2 Hypertension and the kidney

Lawrence E. Ramsay

This section describes the effects of hypertension on kidney function; renal and renovascular disease causing hypertension are described in Chapter 15.16.2.2. The impact of accelerated hypertension on renal function, and of hypertension and its treatment on the progression of established renal failure, are reasonably clear. There is debate about whether hypertension causes renal failure in the absence of accelerated phase, renovascular disease, or undetected primary renal disease.

Accelerated hypertension and the kidney

Accelerated (or malignant) hypertension is defined by bilateral fundal haemorrhages and exudates. Papilloedema may be present but is not necessary for the diagnosis. The condition is fully described in Chapter 15.16.3. Pathological changes in the kidney include 'onion skin' intimal proliferation in interlobular arteries with narrowing or loss of the lumen, ischaemic atrophy of nephrons, and fibrinoid necrosis of arterioles. These are caused by disruption of the arteriolar wall by severe or rapidly increasing hypertension, allowing insudation of plasma and fibrin deposition. The clinical correlate is moderate to severe renal impairment in one-third of patients with accelerated hypertension, or (uncommonly) oliguric acute renal failure. When blood pressure is controlled there may be slight deterioration of renal function over 1 to 2 days, but renal function then stabilizes and often recovers slightly. Patients with oliguric acute renal failure caused by accelerated hypertension may have excellent recovery of renal function even after prolonged dialysis. However, those who present with serum creatinine of 300 µmol/l or higher (but not oliguric renal failure) often progress to endstage renal disease despite good control of blood pressure and in the absence of a primary renal or renovascular cause. By contrast, in idiopathic accelerated hypertension with serum creatinine less than 300 µmol/l renal function usually remains stable, and deterioration despite good blood pressure control should raise suspicion of a primary renal or renovascular cause.

The 5-year survival of patients with accelerated hypertension has improved dramatically, from 1 per cent without treatment to 75 per cent with modern treatment, but the prognosis is still impaired, with mortality of 25 per cent over 5 years. Renal impairment is a powerful determinant of prognosis, and prevention by urgent treatment of accelerated hypertension is therefore crucial. Bilateral haemorrhages and exudates, even without papilloedema, constitute a medical emergency requiring immediate admission. An underlying cause for hypertension, commonly renal or renovascular disease, is present in about a quarter of patients with accelerated hypertension (see Chapter 15.16.2.1 and Chapter 15.16.3).

Hypertension and progression of chronic renal failure

Hypertension is the rule in advanced renal failure of any cause, but occurs earlier in glomerular than interstitial renal disease. The mechanism is inability to excrete a sodium load plus inappropriately high peripheral vascular resistance, i.e. an imbalance between volume and vasoconstriction. Uncontrolled hypertension accelerates the progression of renal failure.

Diabetic nephropathy

As discussed in Chapter 20.10.1, antihypertensive treatment retards the progression of diabetic nephropathy from its earliest stages, slowing progression of microalbuminuria to overt proteinuria, and the decline of glomerular filtration rate in established diabetic nephropathy. Treatment with angiotensin converting enzyme inhibitors confers additional protection, but control of blood pressure probably outweighs any specific effect of angiotensin converting enzyme inhibition. The evidence is stronger for type 1 than type 2 diabetes, but all patients with diabetic nephropathy should have antihypertensive treatment that includes an angiotensin converting enzyme inhibitor. The blood pressure target is (in most cases) as low as is achievable, ideally less than 130/80 mmHg, or less than 125/75 mmHg when there is proteinuria greater than 1 g/24 h.

Non-diabetic nephropathy

There is a relation between the level of blood pressure and the rate of decline of the glomerular filtraton rate in patients with renal impairment, suggesting that uncontrolled hypertension accelerates progression. The putative mechanism is impaired autoregulation in damaged kidneys so that systemic hypertension translates to glomerular hypertension, glomerulosclerosis, and progression of renal failure. Control of blood pressure slows progression, and blood pressure of 140/90 mmHg or more should be treated in any patient with renal disease.

Target blood pressure in renal failure

The Modification of Diet in Renal Disease Study, in addition to looking at the effects of diet, compared antihypertensive treatment titrated to targets equivalent to 140/90 mmHg or 125/75 mmHg in patients with various non-diabetic renal diseases and renal impairment. The overall analysis showed no significant relation between target blood pressure and decline of glomerular filtration rate. However, subgroup analysis showed a substantially slower decline in renal function in patients with proteinuria of 3 g/24 h or more, a similar trend with proteinuria of 1.0 to 2.9 g/24 h, but no benefit when proteinuria was less than 1 g/24 h, or in patients with polycystic kidneys (Fig. 1).

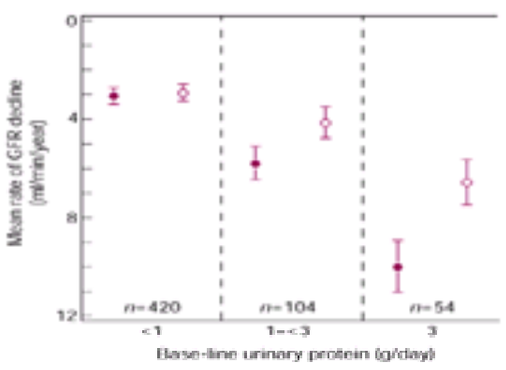

Fig. 1 Decline in glomerular filtration rate related to baseline urinary protein excretion and target blood pressures in patients with non-diabetic chronic renal disease. Solid circles (•) indicate target blood pressure of 140/90 mmHg, open circles () a target blood pressure of 125/75 mmHg. More aggressive blood pressure control

slowed the rate of progression with proteinuria of 3 g/day or more, with similar trends for proteinuria of 1 to 2.9 g/day. There was no apparent effect with proteinuria of less than 1 g/day. (From the Modification of Diet in Renal Disease Study study (Klahr S *et al*. 1994 *New England Journal of Medicine* **330**, 877–84) with permission.)

Angiotensin converting enzyme inhibitors

Treatment that includes an angiotensin converting enzyme inhibitor slows progression of renal failure in selected patients, and the Ramipril Efficacy in Nephropathy (REIN) study suggests specific benefit from angiotensin converting enzyme inhibition over and above reduction of blood pressure. Benefit is large in those with glomerular disease, and small or absent in interstitial disease, polycystic kidneys, or nephrosclerosis. Rapid progression determines the outcome of treatment with angiotensin converting enzyme inhibitors and is predicted by proteinuria and the severity of renal failure. Treatment with angiotensin converting enzyme inhibitors is indicated in any patient with hypertension (³ 140/90 mmHg), renal impairment, and proteinuria of 1 g/24 h or more. Angiotensin converting enzyme inhibitors cause renal failure in critical renovascular disease, and there was concern that patients with renal parenchymal disease could be at similar risk because of 'functional' vascular insufficiency. Trials do not support this concern as withdrawals because of renal failure or hyperkalaemia were no more common with angiotensin converting enzyme inhibitor than with placebo.

Conclusions

Blood pressure of 140/90 mmHg or more should be treated in all patients with renal impairment. When proteinuria exceeds 1 g/24 h the regimen should include an angiotensin converting enzyme inhibitor and the target is less than 125/75 mmHg. Angiotensin converting enzyme inhibitors rarely cause serious adverse effects, but renal function and serum potassium should be monitored.

Idiopathic hypertension and renal failure

Introduction

Renal function is altered in early idiopathic hypertension, with reduced renal blood flow, increased efferent arteriolar tone, but maintained glomerular filtration rate. In experimental animals such changes can cause intraglomerular hypertension, hyperfiltration, and glomerulosclerosis. An unresolved question is whether idiopathic non-accelerated hypertension causes similar renal damage in humans, and, if so, whether antihypertensive treatment prevents this. This question is important. If idiopathic hypertension causes endstage renal disease despite conventional antihypertensive treatment, possible responses are to treat even milder hypertension, to lower blood pressure targets, or to prefer antihypertensive drugs that may be renoprotective. If idiopathic hypertension is an innocent bystander when proteinuria or renal failure develops, the response should be to seek the true cause of the renal abnormality which might be undiagnosed glomerular, interstitial, or vascular renal disease.

Pathology

Nephrosclerosis is characterized by hyaline arteriosclerosis in arterioles and intimal fibroelastic reduplication in interlobular arteries. Some view these these changes as confined to vessels, with no encroachment on lumina, glomerular loss, or other important consequences, and consider them entirely benign and not even specific for hypertension. Others believe that preglomerular arteriosclerosis leads to glomerulosclerosis and loss of renal function through narrowing or occlusion causing ischaemia, or by failure of autoregulation causing intraglomerular hypertension.

Outcome trials

The incidence of endstage renal disease or other renal endpoints has been very low in outcome trials and no effect of treatment on renal function has been observed. These trials have included over 50 000 hypertensives, suggesting that renal failure is at most a rare complication of idiopathic hypertension. However, the trials have been short, averaging 2 to 3 years, and have not studied renal function in detail because of the focus on stroke and coronary complications.

Cohort studies

Uncontrolled studies over 5 to 14 years in treated hypertension have shown renal failure or proteinuria developing in 0 to 18 per cent of patients despite adequate control of blood pressure. These studies were in hospital patients with fairly severe hypertension and had inadequate criteria for excluding intrinsic renal disease causing the hypertension. In the MRFIT study (see below) men with serum creatinine of less than 106 µmol/l and proteinuria of less than 1+ at entry were followed for 16 years. The risk of endstage renal disease almost doubled with increased baseline blood pressure of 15/10 mmHg, but the incidence was still very low in absolute terms. Furthermore, undiagnosed primary renal disease could still have caused the endstage renal disease and explained its relation to blood pressure, even in this cohort, because normal serum creatinine and urine protein do not exclude renal causes for hypertension such as renovascular disease, glomerulonephritis, or renal scarring.

Endstage renal disease

Hypertensive nephrosclerosis is said to account for a quarter of entries to endstage renal disease programmes, but this diagnosis is usually presumptive and may often be incorrect. Presumptive diagnoses will include undiagnosed renovascular disease, undiagnosed parenchymal disease, and undiagnosed or forgotten episodes of accelerated hypertension. In one centre 11 per cent of cases having renal replacement were attributed to hypertension. When examined in detail 45 per cent had documented accelerated phase, 15 per cent had scarring or glomerular disease, 4 per cent renovascular disease, and only 2 per cent had documented hypertensive nephrosclerosis. The remaining 34 per cent of patients had hypertension, endstage kidneys, and no definite diagnosis, but had been labelled as having hypertensive nephrosclerosis.

Epidemiological data

The association of endstage renal disease with seven blood pressure strata in the MRFIT cohort of 332 544 men followed for 16 years is shown in Table 1. There is a clear graded association, but the absolute risk in men with mild to moderate hypertension is very small, for example 1/637 over 10 years with blood pressure of (140 to 159)/(90 to 99) mmHg when compared with blood pressure of less than 120/80 mmHg. As discussed above, this association could be explained by undiagnosed intrinsic renal disease or development of renovascular disease. The significant risk factors for developing endstage renal disease were older age, low income, high cholesterol, smoking, diabetes, and hypertension. Note that these are the major risk factors for atherosclerosis, and development of atherosclerotic renovascular disease could cause endstage renal disease in mild to moderate hypertension.

Conclusions

Renal failure and proteinuria in mild to moderate hypertension usually indicate underlying renal or renovascular disease. When renal damage is present, consider whether the hypertension is, or has been, severe enough to cause accelerated phase. If not, assume that intrinsic renal disease or renovascular disease is present and investigate appropriately. If it has, a presumptive diagnosis of renal damage caused by present or past accelerated hypertension should be confirmed by careful follow-up. Deterioration of renal function despite reasonable blood pressure control should prompt investigation for a cause other than the hypertension.

Further reading

Giatras I, Lau J, Levey AS for the Angiotensin-Converting-Enzyme Inhibition and Progressive Renal Disease Study Group (1997). Effect of angiotensin-converting enzyme inhibitors on the progression of nondiabetic renal disease: a meta-analysis of randomized trials. *Annals of Internal Medicine* **127**, 337–45. [Meta-analysis of trials of angiotensin converting enzyme inhibitors in non-diabetic renal disease.]

Klag MJ *et al*. (1996). Blood pressure and end-stage renal disease in men. *New England Journal of Medicine* **334**, 13–18. [Largest and longest epidemiological study relating endstage renal disease to

blood pressure.]

Klahr S (1989). The kidney in hypertension—villain and victim. *New England Journal of Medicine* **320**, 731–2. [Brief review—does idiopathic hypertension cause endstage renal disease?]

Klahr S *et al*. for the Modification of Diet in Renal Disease Study Group (1994). The effects of dietary protein restriction and blood-pressure control on the progression of chronic renal disease. *New England Journal of Medicine* **330**, 877–84. [Modification of diet in renal disease study of different blood pressure targets in non-diabetic renal failure.]

Madhavan S *et al*. (1995). Renal function during antihypertensive treatment. *Lancet* **345**, 749–51. [Cohort study suggesting idiopathic hypertension does not cause endstage renal disease.]

Rostand SG *et al*. (1989). Renal insufficiency in treated essential hypertension. *New England Journal of Medicine* **320**, 684–8. [Cohort study suggesting that idiopathic hypertension may cause endstage renal disease.]

Ruggenenti P *et al*. (1999). Renoprotective properties of ACE-inhibition in non-diabetic nephropathies with non-nephrotic proteinuria. *Lancet* **354**, 359–64. [Second report from REIN trial suggesting that inhibition of angiotensin converting enzyme has a specific renoprotective effect in non-diabetic renal impairment.]

Tomson CRV, Petersen K, Heagerty AM (1991). Does treated essential hypertension result in renal impairment? A cohort study. *Journal of Human Hypertension* **5**, 189–92. [Cohort study and examination of endstage renal disease registry suggesting idiopathic hypertension does not cause endstage renal disease.]

Whitworth JA (1992). Renal parenchymal disease and hypertension. In: Robertson JIS, ed. *Handbook of hypertension, vol 15, Clinical hypertension*, pp 326–56. Elsevier, Amsterdam. [Excellent review of hypertension in renal disease.]

20.10.3 Vasculitis and the kidney

A. J. Rees

Introduction

This chapter is concerned with a heterogeneous group of disorders commonly referred to as the primary systemic vasculitides. These are defined by the presence of inflammation and necrosis of blood vessels, with the individual clinical syndromes defined by the size and distribution of the vessels involved and whether vasculitis is accompanied by granulomas. The diseases include polyarteritis nodosa, Wegener's granulomatosis, microscopic polyangiitis, and Churg–Strauss syndrome, the renal aspects of which are emphasized in this chapter. The aetiology and pathogenesis of these disorders have not been elucidated, but the strong association between some of them and autoantibodies to neutrophil cytoplasmic antigens (antineutrophil cytoplasmic antibodies: **ANCA**) suggests an autoimmune pathogenesis.

Vasculitis is an important treatable cause of renal failure, most commonly when the glomerular capillaries are involved thus causing focal necrotizing glomerulonephritis. This is common in patients with generalized small vessel vasculitis, such as Wegener's granulomatosis and microscopic polyangiitis, both of which are closely associated with ANCA. Some patients presenting with focal necrotizing glomerulonephritis without evidence of involvement of other organs also have positive assays for ANCA. The appearance on renal biopsy is identical to microscopic polyangiitis, as is the response to treatment. Other organs can be involved later in the course of the disease, and autopsies have demonstrated more widespread disease even at the outset. For these reasons, ANCA-associated focal necrotizing glomerulonephritis in the absence of any evidence of generalized disease is usually considered to be a form of microscopic polyangiitis, and will be discussed as such in this chapter.

Focal necrotizing glomerulonephritis usually presents with rapidly deteriorating renal function (rapidly progressive glomerulonephritis) and is caused most frequently by ANCA-associated vasculitis. However, this is not its only cause: focal necrotizing glomerulonephritis can complicate diseases described in other chapters, including hypersensitivity to drugs, systemic infections such as infective endocarditis, systemic immunological diseases (for example systemic lupus erythematosus, Henoch–Schönlein purpura, and rheumatoid arthritis), and as a paraneoplastic syndrome (Table 1). Whilst there are only subtle differences in the morphological appearances of the different types of focal necrotizing glomerulonephritis, there are marked differences in immunohistology, different types usually being separated on the basis of their pattern of glomerular immunoglobulin deposition (Table 1).

Focal necrotizing glomerulonephritis should be regarded as a medical emergency: because different types respond differently to treatment a precise diagnosis must be made promptly on clinical and serological grounds and confirmed by biopsy.

Historical perspective

Vasculitis was first described in autopsy material in the 1840s by Rokitansky but was not recognized as a specific entity until 20 years later. In 1866, Kussmaul and Maier described the case of a young man who presented with fever and muscle, renal, and gastrointestinal disease and at autopsy had widespread inflammation of medium-sized arteries with aneurysms to which the name periarteritis nodosa was applied. The heterogeneity of the disorder was recognized over the next 100 years and various different clinical syndromes were described. In 1923 Wohlwill distinguished a type of vasculitis that affected arterioles, capillaries, and venules as well as muscular arteries. Davson and his colleagues studied this condition in great detail in the 1940s and called it the microscopic form of polyarteritis (now more accurately called polyangiitis). In the 1930s Klinger, and then Wegener, described patients with microscopic polyarteritis in which many of the arthritic lesions were surrounded by granulomas. Disease in these patients had a predilection for the nose, upper airways, and lungs and constituted a distinct clinical syndrome, now called Wegener's granulomatosis. In 1954 Churg and Strauss described another distinct form of primary vasculitis characterized by granulomatous inflammation of medium-sized vessels associated with eosinophilia and asthma: this disease now bears their names. Despite their obvious differences, it should be remembered that Churg in particular regarded Wegener's granulomatosis, microscopic polyangiitis, and Churg–Strauss syndrome as closely related (with macroscopic polyarteritis regarded as being somewhat separate). The subsequent demonstration that ANCA are commonly present in all three of these conditions (but not in macroscopic polyarteritis) supports this suggestion.

Classification

The classic accounts of vasculitis established strict clinical and pathological criteria for diagnosis based on clinical histories and autopsies of patients with endstage disease. They defined the different clinical disorders and remain invaluable descriptions of their evolution. However, the diagnostic criteria they introduced are now far too restrictive because nowadays patients usually present at an early stage, with relatively mild or limited disease that can be controlled by present treatments before it has evolved into a 'classical' syndrome. The current emphasis on early diagnosis and treatment demands a more flexible approach. The finding that most patients with active small vessel vasculitis have circulating ANCA has proved to be especially useful, even though the role of ANCA in pathogenesis is uncertain. Differences in ANCA specificity have also tended to reinforce the idea that different vasculitic syndromes are distinct entities. Thus, most patients with active small vessel vasculitis have positive assays for ANCA, whereas patients whose disease in confined to the arteries do not.

The current more pragmatic approach to diagnosis is exemplified by the criteria developed by Lanham for Churg–Strauss syndrome, but generally the move towards more clinically based definitions has been a source of considerable controversy. New terms were introduced to describe 'incomplete variants', the distinction between classic (large vessel) and microscopic forms of polyangiitis became blurred, and different terms were used for the same entities, not only in different parts of the world but—worse still—by physicians practising in different specialties. In 1990, the American College of Rheumatologists attempted to clarify the situation by establishing standard diagnostic criteria for each of the main vasculitic syndromes, analogous to those that had proved highly successful for systemic lupus erythematosus. These criteria were based on clinical and pathological data collected from patients diagnosed as having the various types of systemic vasculitis, the data being used to derive mutually exclusive sets of criteria for diagnosing each condition. Unfortunately, they have not proved to be robust, especially for patients with extensive renal disease: for example, they do not accurately discriminate between patients with small vessel vasculitis and focal necrotizing glomerulonephritis and those with classical polyarteritis nodosa, or between cutaneous leucocytoclastic vasculitis and Henoch–Schönlein purpura. However, the American College of Rheumatologists

project was an important attempt to introduce greater uniformity, although it was flawed because differences in the pattern of disease of patients referred to different physicians was underestimated, and also because the drive for early treatment of patients with vasculitis reversed the injury before patients developed signs that enabled individual conditions to be distinguished from each other with confidence.

In an attempt to overcome these difficulties, the Chapel Hill Consensus Conference in 1994 brought together physicians from all the disciplines to which patients with vasculitis are commonly referred and pathologists with extensive experience of these diseases. The purpose was to establish a classification of vasculitis with a common nomenclature based on set working definitions for the different syndromes. It was recognized that it would not always be possible (or even relevant) to try to distinguish between some of the syndromes before treatment was started, and no attempt was made to establish strict diagnostic criteria although clearly the Chapel Hill classification provided a framework for diagnosis. The definitions of individual clinical syndromes built on previous work that classified vasculitis in terms of the size of vessel involved and the presence or absence of granulomas. Clinical definitions of the different types of vasculitis are shown in Table 2 and illustrated in Fig. 1. The essential feature is that the diagnosis depends on the size of the smallest vessel involved. Thus the term microscopic polyangiitis replaced the term microscopic polyarteritis because capillaries and venules are involved as well as arteries, and the diagnosis depends on their involvement regardless of whether or not small or medium-sized muscular arteries are also affected. By contrast, polyarteritis nodosa was defined as a form of vasculitis in which the vasculitis is confined to muscular arteries and the diagnosis excludes patients with evidence of injury to capillaries or venules. This terminology was justified by differences in the natural history and response to treatment of patients with and without small vessel involvement and is supported by the serological findings. Although not used as a primary diagnostic criterion, ANCA are very frequent in patients with small vessel vasculitis but uncommon in those whose disease is confined to larger vessels.

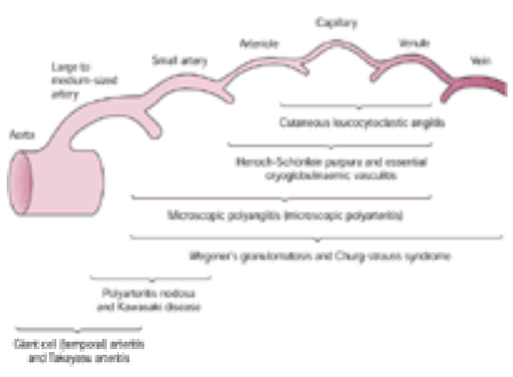

Fig. 1 Summary of the Chapel Hill classification of systemic vasculitis.

The Chapel Hill classification has been increasingly widely used since it was first introduced but should not be regarded as definitive. The very high prevalence of ANCA in patients with small vessel vasculitis has lead to the useful generic term ANCA-associated vasculitis, which includes Wegener's granulomatosis and microscopic polyangiitis as well as patients with isolated focal necrotizing glomerulonephritis without clinical evidence of extrarenal disease.

Antineutrophil cytoplasmic antibodies

ANCA were first identified in patients with small vessel vasculitis thought to be due to Ross River virus and were subsequently reported to be highly specific for Wegener's granulomatosis and microscopic polyangiitis. The antibodies are usually assayed by indirect immunofluorescence using ethanol-fixed normal human neutrophils as the substrate. They are heterogeneous, as reflected in the two distinct patterns of fluorescence that can be seen (Fig. 2 and Plate 1). Some sera display granular staining throughout the cytoplasm (cytoplasmic ANCA or **cANCA**), whereas with others the staining concentrates around the nucleus (perinuclear ANCA or **pANCA**); exceptional patients demonstrate a mixed pattern of staining.

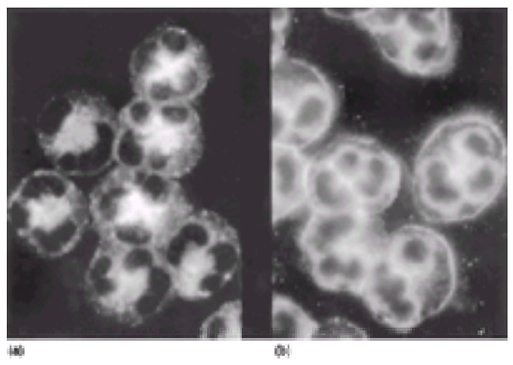

Fig. 2 Indirect immunofluorescence assay for ANCA. (a) Typical staining of cytoplasmic ANCA that is usually due to antibodies to proteinase 3. (b) Typical staining pattern of perinuclear ANCA most often due to antimyeloperoxidase antibodies. (See also Plate 1.)

Most ANCA bind to enzymes found in the neutrophil granules, and the commonest antigens in patients with small vessel vasculitis are proteinase-3, which is the principal target of cANCA, and myeloperoxidase, which is the most common target for pANCA. Antiproteinase-3 antibodies are almost invariably present in untreated patients with active systemic Wegener's granulomatosis, as well as a proportion of those with microscopic angiitis. Antimyeloperoxidase antibodies are usually found in the remaining patients with microscopic polyangiitis and in those with isolated focal necrotizing glomerulonephritis. However, it is important to recognize that ANCA are not unique to small vessel vasculitis and there is a growing list of conditions in which they have also been described, including inflammatory bowel disease, systemic lupus erythematosus, and some chronic infections. In these contexts it is rare for the ANCA detected to be directed against proteinase-3 and uncommon that they recognize myeloperoxidase. ANCA that recognize other antigens are found in a minority of patients with small vessel vasculitis (Table 3), and their targets include elastase, cathepsin G, lactoferrin, bacterial-permeability-increasing peptide, and the lysosomal membrane protein h-lamp-2. This last antigen is interesting in that it is also expressed on glomerular endothelium, and antibodies that recognize glomerular endothelium have previously been described in Wegener's granulomatosis.

A critical issue is whether assays for ANCA are specific and sensitive for small vessel vasculitis. Meta-analyses and large-scale multicentre prospective studies have confirmed the original impression that they are, provided that certain safeguards are met. When used alone, ANCA detected by indirect immununoflurescent assays have a sensitivity of 80 to 85 per cent and a specificity of around 75 per cent compared with disease controls. Specificity improves considerably when indirect immunofluorescence assays are used together with specific enzyme-linked immunoabsorbent assays to exclude patients who have ANCA directed against targets other than proteinase-3 and myeloperoxidase, and those with false positive assays caused by antinuclear antibodies. Combining indirect immunofluorescence with specific immunoassays for antibodies to proteinase-3 and myeloperoxidase increases both the sensitivity and the specificity of the assays for Wegener's granulomatosis and microscopic polyangiitis to over 90 per cent. However, the standard enzyme-linked immunoabsorbent assay for proteinase-3 and myeloperoxidase can occasionally give false negative results when circulating ANCA are bound to their natural inhibitors a $_1$-antitrypsin and caeruloplasmin. There are far fewer data about the use of ANCA for diagnosing Churg–Strauss syndrome: overall about 50 per cent of patients have positive assays for either proteinase-3 or myeloperoxidase, but a sufficiently large cohort of untreated patients has not been studied to establish a true incidence. ANCA are positive in fewer than 10 per cent of patients without evidence of involvement of capillaries or venules.

Incidence and aetiology

Primary systemic vasculitis is relatively uncommon, but the incidence is probably increasing, even allowing for greater awareness and improvements in diagnosis. The overall incidence in southern England is about 20 new patients per million population per year, with a similar incidence reported in Scandinavia. Men are more

commonly affected than women, and the incidence rises progressively with age, amounting to 60 per million per year for those aged 65 to 75. The annual incidence of ANCA-associated crescentic nephritis is 6 to 7 per million **,**accounting for at least 0.5 per cent of patients on dialysis in Europe. However, surveys of the ANCA status of dialysis patients suggest that both figures are underestimates because of underdiagnosis in patients presenting with renal failure and few extrarenal symptoms. The disease appears to be equally common in South Asian caucasoids, but studies from the United States indicate that African Americans are much less susceptible.

Genetic factors

The cause of primary systemic vasculitis is unknown, but heredity undoubtedly influences susceptibility as evidenced by reports of the disease occurring in siblings; however, identical twins discordant for the disease have also been observed. Polymorphisms of the a $_1$-antitrypsin locus influence susceptibility to Wegener's granulomatosis and cANCA positive vasculitis, case control studies showing that the Z allele, which causes relative antitrypsin deficiency, is much more common in patients than in controls. a$_1$-antitrypsin is the natural inhibitor of proteinase-3 (but not of myeloperoxidase) and so loss of this inhibition provides an apparent explanation for the findings, although an unconvincing one given the relatively trivial reduction in a $_1$-antitrypsin activity produced by the Z allele. It should also be noted that nearly 80 per cent of patients have the fully sufficient phenotype. There is a single study describing an association between the complement C3 allele *C3F* and vasculitis, which resulted in a relative risk of 2.6 in heterozygotes and 5.1 in homozygotes. By contrast, studies of the HLA class II complex have produced conflicting results.

Environmental factors

Environmental agents responsible for primary systemic vasculitis have proved even more difficult to identify than genetic factors, but drugs and infections have both been invoked. Hepatitis B carriers have been reported to develop polyarteritis nodosa in some populations (United States and France), but not in others such as the United Kingdom. Infection with parvovirus BIg has also been reported to cause polyateritis nodosa. Hypersensitivity to hydrallazine, rifampicin, and minocycline can cause ANCA-associated vasculitis, and there is convincing evidence that penicillamine causes pANCA-associated crescentic glomerulonephritis with or without pulmonary haemorrhage. Case controlled studies indicate that silica predisposes to ANCA-associated crescentic nephritis and Wegener's granulomatosis. It is difficult to evaluate descriptions of Churg–Strauss syndrome in patients with asthma treated with the leukotriene inhibitor Zafirlukast: asthma could have been the first symptom of the vasculitis, and corticosteroids, which would have been the alternative treatment, would also have been effective treatment for Churg–Strauss.

Pathogenesis

Explanations for the cause of injury in small vessel vasculitis must take account of the close association with ANCA, the extensive infiltration of vessels with neutrophil, endothelial activation, and the near absence of immunoglobulin deposition on immunohistology. Most authorities regard these conditions as autoimmune diseases, but proof is lacking and it is important to remember that 30 years ago they were generally believed to be systemic immune complex diseases.

A direct role for ANCA is suggested by their close association with injury in vasculitis, especially as ANCA titres often, though not invariably, increase before relapses of disease activity. The observation that vasculitis in spontaneous and induced models of autoimmunity in rodents is also associated with ANCA provides further support for the hypothesis. For example, *MRL/lpr* lupus prone mice and Kinjo mice both spontaneously develop vasculitis and focal necrotizing glomerulonephritis in association with ANCA. The real difficulty is to understand how ANCA could cause vasculitis, because neither proteinase-3 nor myeloperoxidase are expressed on endothelium. Three types of explanation have been proposed:

1. They could cause endothelial injury after release from neutrophils and monocytes because they prolong the activity of proteinase-3 and myeloperoxidase by interfering with binding to their natural inhibitors.
2. They activate neutrophils (and possibly endothelium) and so could facilitate endothelial injury.
3. They could bind to endothelium and so facilitate *in situ* immune complex formation between ANCA and their targets.

There has been widespread interest in the effects of ANCA on neutrophil function since the initial reports that they stimulated respiratory burst activity in neutrophils that had been primed with proinflammatory cytokines such as tumour necrosis factor. Proteinase-3 translocates to the plasma membrane when polymorphs are exposed to tumour necrosis factor, and the process is enhanced by the chemokine interleukin 8. *In vivo*, neutrophils from patients with active vasculitis express more proteinase-3 on their surface than neutrophils from controls. Incubation of primed neutrophils with ANCA increases superoxide generation and enzyme release and also interferes with normal control of apoptosis.

Regardless of how ANCA activate neutrophils, there is now a strong body of evidence to show that they facilitate the adherence of neutrophils to endothelium and neutrophil mediated killing of endothelial cells *in vitro*. At one stage it was thought that activated endothelium expressed proteinase-3 but this is now known not to be the case. It seems probable that ANCA aggravate neutrophil-dependent injury once they adhere in the microvasculature.

The question as to whether immune responses to proteinase-3 and myeloperoxidase cause vasculitis and glomerulonephritis directly has also been examined experimentally by infusing neutrophil granule extracts together with hydrogen peroxide (H $_2$O$_2$) into the renal arteries of rats that had previously been immunized with myeloperoxidase. This resulted in the development of severe focal necrotizing glomerulonephritis with very transient deposition of immunoglobulin, hence the injury simulated that seen in ANCA-associated vasculitis. Rats in whom the perfusion with H $_2$O$_2$ was omitted developed less severe injury and more persistent IgG deposits. In most respects this antimyeloperoxidase model is just another example of '*in situ* complex formation nephritis' similar to models used to examine other types of glomerular injury. The main difference is the transience of the IgG deposits, but even this has been recorded previously in models of chronic serum sickness under conditions of marked excess of antigen. These are exactly the conditions that would be expected in inflamed glomeruli in which neutrophils and monocytes continued to release proteinase-3 and myeloperoxidase in the presence of ANCA. Two additional pieces of evidence support the notion that IgG deposition in ANCA-associated crescentic nephritis might be unusually shortlived: first, proteinase-3 degrades IgG including ANCA, even when complexed to it; secondly, it has been known for many years that IgG deposited in vessels in some forms of cutaneous vasculitis rapidly becomes undetectable. Thus it is entirely plausible that IgG is deposited in glomeruli of patients with crescentic nephritis and is then rapidly removed or destroyed.

In summary, there is no decisive proof that autoimmunity to neutrophil antigens cause injury in human vasculitis, but a number of clear statements can be made:

1. Autoantibodies to ANCA are very closely associated with some types of systemic small vessel vasculitis, both clinically and in experimental models.
2. The autoantibodies modify the function of neutrophil enzymes and promote neutrophil mediated endothelial injury *in vitro*.
3. Injection of antimyeloperoxidase antibodies does not in itself cause glomerular injury, except in rats in which myeloperoxidase has been planted in the kidney, when the result is severe glomerulonephritis with very similar morphological and immunohistological characteristics to human ANCA-associated glomerulonephritis.
4. Injection of antirat myeloperoxidase antibodies into rats with nephrotoxic nephritis markedly aggravates injury.

Thus it reasonable to suggest that the cellular or humoral autoimmune response to neutrophil antigens contributes to the pathogenesis of the glomerular injury in vasculitis.

Clinical features of primary vasculitis

Wegener's granulomatosis and microscopic polyangiitis

Wegener's granulomatosis is characterized by a predominantly small vessel vasculitis associated with granulomas and a predilection for involving the upper and lower airways. Microscopic polyangiitis is characterized by a very similar small vessel vasculitis without granuloma formation or a predilection for the upper airways. These two conditions will be considered together because of the many features they share, and because clinically it can be impossible to distinguish between them. Furthermore, the approach to diagnosis and management of both conditions is identical.

Both Wegener's granulomatosis and microscopic polyangiitis present with non-specific symptoms and signs that can be difficult to distinguish from those of viral infection, except that they are more persistent. These consist of flu-like symptoms with muscle pains, night sweats, and weight loss. Many patients also complain of arthralgia, which is symmetrical, relatively mild, and affects the hands and feet; exceptional patients have frank arthritis. These non-specific symptoms either precede

or occur as a background to damage to particular organs, which provides the basis for making a more specific diagnosis.

Features unique to Wegener's granulomatosis

Upper airway involvement

Over 90 per cent of patients with Wegener's granulomatosis have obvious upper airways disease. Nasal discomfort and blockage together with ulceration, crusting, rhinorrhoea, and epistaxis are usually prominent at presentation. Destruction of nasal cartilage with perforation of the nasal septum or the appearance of the characteristic saddle nose deformity rarely occur until much later. Gross destruction of bone is very uncommon and suggests an alternative diagnosis such as malignancy. Disappointingly, nasal biopsies rarely show the characteristic lesions of necrotizing vasculitis with loosely formed granulomas unless special techniques are used to get adequate samples. Nasal disease is frequently accompanied by the involvement of the paranasal sinuses, blockage of the Eustachian tube and otitis media. The larynx is less commonly involved but subglottic stenosis is a characteristic late phenomenon, which may develop insidiously and can be severe.

Pulmonary involvement

Almost all the patients have evidence of granulomatous lung disease at presentation, which is often accompanied by alveolar capillaritis. The bronchi can also be affected and bronchial stenoses occur as late manifestations. Symptoms include cough, dyspnoea, haemoptysis, and chest pain, which can be pleuritic. Signs on chest examination depend on the nature of the pulmonary lesions and include fine crepitations and bronchial breathing or less commonly pleural rubs and signs of pleural effusions. Fixed rhonchi are suggestive of bronchial stenosis and stridor of subglottic stenosis.

Pulmonary granulomas are usually diagnosed from chest radiographs and CT scans. They may appear as single or multiple rounded lesions, which can cavitate (Fig. 3). They are often intermingled with alveolar shadowing to produce composite lesions, which reflect pulmonary capillaritis and haemorrhage as well as granulomas. The radiological findings can be confused with neoplasms, infection, or fluid overload. Results from pulmonary function tests are not specific, but it is important to note that increases in the K_{CO} are a much less sensitive test for pulmonary haemorrhage in vasculitis than in disease mediated by antiglomerular basement membrane antibodies (see Chapter 20.7.9). Bronchoscopy often reveals granulomatous inflammation and the diagnosis can sometimes be made from bronchial biopsies. By contrast, needle biopsies and transbronchial biopsies of pulmonary lesions are rarely adequate and can delay definitive diagnosis by open or video-endoscopic lung biopsy. Histology of lung biopsies reveals pulmonary necrosis, vasculitis affecting alveolar capillaries and bronchial vessels, and loosely formed granulomas together with a mixed inflammatory infiltrate (Fig. 4 and Plate 2).

Fig. 3 Chest radiograph from a patient with Wegener's granulomatosis showing the typical appearance of pulmonary granulomas together with alveolar shadowing caused by capillaritis of lung haemorrhage.

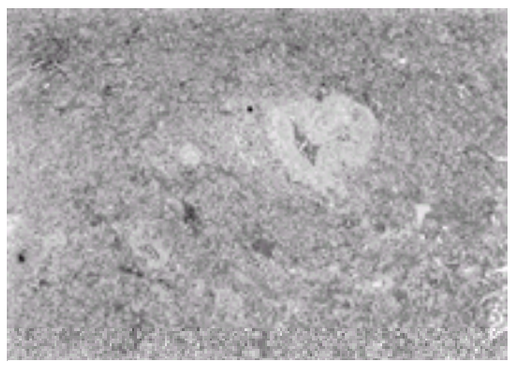

Fig. 4 Morphological appearances of pulmonary granulomas in a specimen obtained by video-endoscopic lung biopsy from a patient with Wegener's granulomatosis. (See also Plate 2.)

Granulomatous inflammation at other sites

Typical Wegener's granulomas have been reported at numerous sites outside the lung, but are much less common in part at least because of the greater difficulty in diagnosing them. Microscopic granulomas are seen in a minority of renal and skin biopsies from patients known to have Wegener's granulomatosis, but macroscopically visible granulomas are rare. However, these have been reported in kidney, simulating carcinoma, in the orbit of the eye, causing exophthalmos, and also in the brain, pituitary, salivary glands, prostate, gingiva, and vertebrae. Masses of inflammatory tissue have also been described in the intestine (simulating Crohn's disease) and breast, as well as in the retroperitoneal space and mediastinum.

Features common to microscopic polyangiitis and Wegener's granulomatosis

Renal disease

Focal necrotizing glomerulonephritis is the characteristic renal lesion of generalized Wegener's granulomatosis and microscopic polyangiitis. Typically it presents with deteriorating renal function that progresses to renal failure within 3 months, i.e. as rapidly progressive glomerulonephritis, although a few patients have more indolent disease. Proteinuria (typically 2 to 3 g/24 h, but occasionally in the nephrotic range) and microscopic haematuria provide clinical evidence of glomerulonephritis, as does urine microscopy that reveals granular and red cell casts. Despite the glomerular inflammation hypertension is uncommon, except in patients with pre-existing hypertension or obvious fluid overload. Renal ultrasound examination and CT scans show normal or enlarged kidneys, isotope renograms reflect the renal function, and renal arteriograms only exceptionally demonstrate aneurysms or other evidence of vascular injury in the renal arteries. The diagnosis is made by renal biopsy, the appearances reflecting the extent of the injury and the speed with which it has developed.

Cutaneous vasculitis

Cutaneous vasculitis occurs with equal frequency in Wegener's granulomatosis and microscopic polyangiitis, being present in about half of patients. Typically there is palpable purpura or splinter haemorrhages. Blistering and urcitarial lesions are less common, as are infarcts. Biopsies reveal leucocytoclastic vasculitis without deposition of immunoglobulins.

Vasculitis of the eye

Involvement of the eye occurs in both diseases but is more common in Wegener's granulomatosis. Orbital granulomas and exophthalmos have already been mentioned as features unique to Wegener's granulomatosis, and other manifestations include episcleritis, uveitis, and retinal vasculitis.

Gastrointestinal vasculitis

Vasculitis of the gut can occur in Wegener's granulomatosis and microscopic polyangiitis. Oral ulceration is relatively common. Intestinal vasculitis usually presents as abdominal pain and bloody diarrhoea. Endoscopy reveals purpuric lesions with or without ulceration, and white cell scans show increased uptake and are a useful way of delineating the extent of disease.

Vasculitis of the nervous system

Small vessel vasculitis can affect both central and peripheral nervous systems. Cranial nerve palsies, mononeuritis multiplex, and symmetrical polyneuropathies are straightforward to diagnose but are relatively uncommon in Wegener's granulomatosis and microscopic polyangiitis. Cerebral vasculitis also occurs: this is difficult to diagnose but can cause strokes and convulsions. Computed tomography and magnetic resonance imaging can demonstrate lesions, but appearances are not specific.

Diagnosis

Diagnosis of Wegener's granulomatosis and microscopic polyangiitis depends on an appropriate clinical history together with clinical evidence of vasculitis or focal necrotizing glomerulonephritis and biopsy evidence that is at least consistent with the diagnosis if not pathognomic of it. A positive assay for ANCA with specificity for proteinase-3 or myeloperoxidase provides strong confirmatory evidence, and negative ANCA assays are grounds for reconsidering the diagnosis. However, it is important to remember that 10 per cent of patients with generalized disease have negative ANCA, as do up to 50 per cent of those with Wegener's granulomatosis apparently confined to the respiratory tract. Early diagnosis of the patients with non-specific 'flu-like' symptoms can be very difficult and the value of urinary dipstick testing to search for asymptomatic proteinuria and haematuria cannot be overestimated.

Haematology and biochemistry

Non-specific haematological abnormalities are the rule. The full blood count shows a normocytic normochromic anaemia and often a neutrophil leucocytosis. Thrombocytosis is almost invariably present (typically in the range of 400 to 800 × 10^{9}/litre) and provides a useful measure of disease activity. Increased values for C-reactive protein and hypoalbuminaemia provide further evidence of the acute phase response. Alkaline phosphatase in often increased, but abnormalities of other liver function tests are much less common. Serum urea and creatinine values reflect the severity of the renal injury.

Serology

The sensitivity and specificity of ANCA for small vessel vasculitis have already been discussed. The results of other serological tests are variable. Up to 50 per cent of patients have detectable antiendothelial cell antibodies, and there is often a polyclonal increase in immunoglobulins. Rheumatoid factors are present in up to a third of patients and positive antinuclear factor without antibodies to double-stranded DNA in at least 10 per cent. Complement concentrations reflect the intensity of the acute phase response and are either normal or increased.

Renal biopsy

Renal biopsies show focal necrotizing glomerulonephritis of varying degrees of severity. Segments of the glomerular tuft are initially infiltrated by leucocytes, individual capillary loops become necrotic, and intravascular contents escape into Bowman's space (Fig. 5(a) and Plate 3(a)). This provokes an intense inflammatory response that is termed extracapillary proliferation, which eventually surrounds the glomerular tuft to form a 'crescent' (Fig. 5(b) and Plate 3(b)). Freshly formed crescents are composed entirely of cells, but gradually these are replaced by fibrosis. Progressively more glomeruli are involved until the disease becomes diffuse, hence it is usual to see crescents in all stages of their evolution, some being entirely cellular whilst others are completely fibrotic. Immunohistology shows that the glomeruli contain scanty deposits of immunoglobulins and complement and for that reason the lesions are often referred to as pauci-immune glomerulonephritis. The glomerular changes are accompanied by interstitial inflammation and progressive tubular atrophy as increasing numbers of nephrons are destroyed. Arteries affected by vasculitis are seen occasionally.

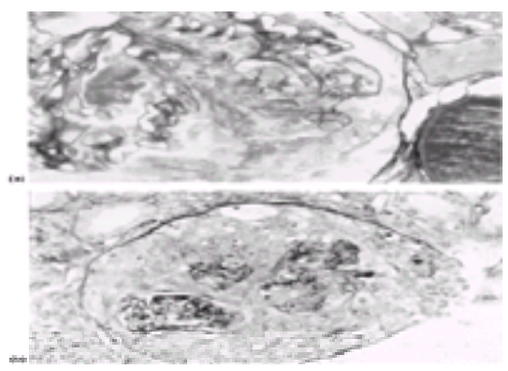

Fig. 5 Morphological appearances on a renal biopsy from a patient with pauci-immune focal necrotizing glomerulonephritis. (a) An early lesion with necrosis of one glomerular segment. (b) A much more florid lesion with the whole glomerular tuft surrounded by a crescent. (See also Plate 3.)

Renal morphology on biopsy is critical for diagnosis but has remarkably little prognostic power. In contrast to the situation with Goodpasture's disease (antiglomerular basement membrane antibody disease), even patients whose biopsies show severe glomerular destruction usually respond to treatment with a marked improvement in renal function. Similarly, patients who appear on biopsy to have indolent disease with marked glomerular scarring and tubular atrophy frequently get worthwhile improvements in function with immunosuppressive therapy. The number of normal glomerular capillary loops provides the best, albeit uncertain, guide to prognosis.

Differential diagnosis

Wegener's granulomatosis and microscopic polyangiitis are distinguished from each other on the basis of evidence for granulomatous inflammation and by whether or not the upper airways are involved. They should also be distinguished from other types of vasculitis. Although they often present with similar symptoms, other types of primary vasculitis are distinguished by clinical (for example asthma and eosinophilia in the case of Churg–Strauss syndrome), serological, radiological, and morphological evidence of the size and distribution of the vessels involved. Recognition of secondary forms of small vessel vasculitis (Table 1) depends on making an alternative diagnosis such as rheumatoid arthritis. In some cases this requires a biopsy: for example in adults Henoch–Schönlein purpura can be difficult to distinguish from microscopic polyangiitis without histological demonstration of prominent granular IgA deposits in glomeruli, or in vessels in vasculitic skin. Mixed essential cryoglobulinaemia, rheumatoid vasculitis, and systemic lupus erythematosus can all be diagnosed on the basis of specific serological tests. Infective endocarditis can cause vasculitis that is sometimes associated with positive ANCA assays (usually with specificity for myeloperoxidase) and should therefore be considered in all patients presenting with a vasculitis.

Other causes of rapidly progressive glomerulonephritis should also be considered in patients with few signs of extrarenal disease. Antiglomerular basement membrane disease and crescentic transformation of other types of chronic glomerulonephritis can be excluded on the basis of serology and renal biopsy. Assays for

antiglomerular basement membrane antibodies are especially important because 10 to 20 per cent of patients with antiglomerular basement membrane disease also have positive ANCA and 1 to 2 per cent of patients with ANCA also have antiglomerular basement membrane antibodies.

Management of small vessel vasculitis

Before any effective treatments were available, Wegener's granulomatosis was a fatal illness with mean survival of less than 6 months. Corticosteroids extended survival, but the benefits were not prolonged and patients still died with progressive disease. The prognosis was transformed in the 1970s by the introduction of regimens that combined cyclophosphamide and steroids. These proved highly effective at suppressing disease activity and are the basis of treatment today. Cyclophosphamide has become the standard immunosuppressive drug, but prolonged treatment imposes considerable morbidity. The goal is to minimize the long-term use of this drug by developing regimens that control disease activity equally well but with less toxicity.

The prognosis of microscopic polyangiitis in the pretreatment era was little better than that of Wegener's granulomatosis, but a greater proportion of patients were treated successfully with corticosteroids alone, with up to 50 per cent survival at 5 years. Again the combined use of steroids and immunosuppressive drugs has improved the prognosis further. This is especially evident in the prognosis for focal necrotizing glomerulonephritis, which previously almost invariably resulted in endstage renal failure, whereas nowadays up to 70 per cent of dialysis-dependent patients with ANCA-associated focal necrotizing glomerulonephritis can be expected to regain independent renal function.

Current approaches to management separate immunosuppressive treatment regimens into induction and maintenance phases (Table 4). Effective regimens have evolved empirically without being assessed by well-designed randomized controlled trials. Formal comparisons are increasingly needed to compare newer and potentially less toxic approaches to treatment with the standard protocols.

Induction treatment

The combination of prednisolone and cyclophosphamide is now established as the standard induction therapy for patients with generalized Wegener's granulomatosis or microscopic polyangiitis. There is consensus on how corticosteroids should be used, but less so for cyclophosphamide. Prednisolone is given in doses of around 1 mg/kg/day initially, after which the dose is reduced rapidly, typically at weekly intervals. Controlled trials show that the addition of pulses of methyl prednisolone is unlikely to confer additional benefit.

Traditionally, patients received daily oral cyclophosphamide (2 mg/kg/day), but latterly intravenous boluses have proved increasingly popular, given in doses of 0.5 to 0.75 g/m^2 body surface area at intervals of 2 weeks (at least for short periods) to 2 months. Pulse therapy has three potential advantages: a lower total dose of cyclophosphamide is used, MESNA (2-mercaptoethane sulfonate sodium) can be given with each dose to minimize bladder toxicity, and compliance is assured. Originally it was believed that pulse therapy might also be more effective, but this has not proved to be the case and indeed the opposite may be true. Results of three randomized prospective controlled trials comparing the two regimens have been reported but are not decisive. They show that both regimens are highly effective for most patients, and that pulse therapy has fewer toxic side-effects, including infections. Importantly, however, they suggest that pulse therapy may be less effective for patients with the most aggressive disease, including those with dialysis-dependent renal failure. Thus the available evidence suggests that the daily oral cyclophosphamide regimen should still be employed for patients with very severe disease. The choice of daily versus pulse therapy is open to personal preference for less severely affected patients.

The duration of induction treatment with cyclophosphamide is also controversial. The approach usually employed in North America and Continental Europe has been to use prolonged courses, albeit with progressively smaller doses. The evident toxicity of this approach has forced a re-evaluation and many now switch to weekly methotrexate after an initial course of cyclophosphamide, but there is a problem in that methotrexate is contraindicated in patients with reduced renal function. Management has been different in the United Kingdom, where it has been routine to use a 3-month induction course of cyclophosphamide and then switch to azathioprine as the immunosuppressive drug. There is now controlled trial evidence that this is as effective as prolonged cyclophosphamide. Thus limiting the induction course of cyclophosphamide to 3 months can be recommended and should be used in all patients with evidence of renal involvement.

Induction treatment for less severely affected patients

Another major issue about induction therapy is whether or not drugs which are less toxic than cyclophosphamide should be used for patients with less severe or more localized disease. In the past, azathioprine was used successfully in this situation, it being clear from uncontrolled studies that it was effective in many patients, but not all, and that its substitution by cyclophosphamide was usually effective in the event of treatment failure. Thus azathioprine did not find a regular place in induction therapy, and weekly pulses of methotrexate at an initial dose of 10 to 15 mg/week, increasing to 20 and then 25 mg/week if necessary, is the current most common alternative to cyclophosphamide. This is an effective treatment but is contraindicated in patients with reduced renal function, which limits its use. Controlled trials are now in progress to compare this approach with standard induction therapy for patients without renal disease. Liver function must be monitored carefully in patients receiving methotrexate because of the risk of hepatotoxicity. Other side-effects of methotrexate include infections and pulmonary fibrosis.

The patient with fulminant disease

The management of patients with fulminant or life-threatening disease presents additional problems. It remains controversial whether they benefit from additional treatment. The two regimens most commonly added in this situation are pulse methylprednisolone (typically 0.5 g/day for three doses) or plasma exchange. A prospective randomized controlled clinical trial of plasma exchange (4 litres daily for 7 to 10 days) demonstrated that this had significant benefit for those already on dialysis, but conferred no benefits over conventional treatment for those with less severe renal disease. A second reported trial showed broadly similar results with no overall benefit but a trend in favour of plasma exchange for those on dialysis. There are no controlled trials of methylprednisolone in dialysis-dependent patients but results of uncontrolled studies of dialysis-dependent patients are similar to those achieved using plasma exchange. A formal controlled comparison of the different regimens is being conducted as part of European Multicentre EUVAS trial group.

Antithymocyte and anti-T-cell antibodies have been used with apparent success in small numbers of patients with severe disease, but this approach should be regarded as experimental and restricted to treatment of patients who have failed to respond to conventional therapy in centres with special experience of these disorders. There is not enough experience of the use of newer immunosuppressive drugs to recommend their use outside the clinical trials.

Maintenance therapy

Systemic vasculitis is a chronic disease and so long-term treatment with steroids and immunosuppressive drugs is required both to maintain remission and to preserve renal function. The key to maintenance therapy is to balance the risks of relapse with those of immunosuppression. For this reason the current approach is to minimize the duration of treatment with cyclophosphamide. The evidence from controlled trials already cited has shown that substituting azathioprine (2 mg/kg) for cyclophosphamide after 3 months is as effective as continuing cyclophosphamide. This is continued for at least a year, together with reducing doses of prednisolone. Thereafter the dose of azathioprine is reduced to 1 mg/kg. An alternative approach is to use weekly pulses of methotrexate (10 to 20 mg/week) instead of azathioprine. Further reductions of treatment can be made during the second year provided that the patient's disease remains in remission, and one can consider stopping treatment altogether in patients who are clinically well provided that their assays for ANCA are negative. Even patients with prolonged clinical remissions are at risk of relapse if immunosuppression is stopped when ANCA are still detectable. There are no data that indicate whether it is best to stop the steroids or cytotoxic drugs first, but there is a strong clinical impression that immunosuppressive drugs are better at maintaining remission than are steroids.

Adjunctive therapy

Intravenous immunoglobulin has been widely used to treat autoimmune disease because of its anti-inflammatory properties and possibly because of its proposed effects on idiotypic networks. It appears to have a modest beneficial effect when used to treat vasculitis, but not sufficient to warrant its regular use. Cotrimoxazole has been widely used as adjunctive therapy both during the induction phase of treatment and to prevent relapses. Evidence from controlled trials demonstrates that it confers significant benefit in reducing the incidence of relapses, especially in those with severe upper airway involvement with Wegener's granulomatosis. It is likely that this effect is mediated through better control of infection. Nasal carriage of *Staphylococcus aureus* is associated with an increased risk of relapse and so long-term use of mupirocin cream may be a useful adjunct in affected patients.

Management of relapses

Relapses are common and occur in between a third and a half of patients with small vessel vasculitis, more frequently in those with Wegener's granulomatosis than with microscopic polyangiitis. They may also be more common in patients with microscopic polyangiitis and antiproteinase-3 antibodies. Relapses can occur at any time, even decades after the initial presentation; they may occur spontaneously or be provoked by a reduction of therapy or by the development of an intercurrent infection. As already stated the risk of relapse is significantly greater in chronic nasal carriers of *S. aureus*.

Most relapses occur in patients who are no longer receiving therapy. Clinical evidence of disease activity is the most important determinant of whether to increase therapy. It is usually unwise to increase therapy in the absence of any clinical signs or symptoms, but it is often wise to do so in patients with minimal non-specific symptoms in whom a change in ANCA assays from negative to positive or an increase in titre provide additional evidence of grumbling activity. Minor relapses can usually be managed by minor adjustments to the maintenance dose of steroids, whereas severe ones require more drastic measures. High-dose corticosteroids may need to be introduced, and depending on baseline therapy the dose of azathioprine may need to be increased or cyclophosphamide reintroduced.

Whenever possible overt relapses should be anticipated by prolonged follow-up of patients. This requires careful monitoring, both of clinical symptoms and signs, and also of investigations including full blood count, biochemical profile, urinary protein and sediment, C-reactive protein, and ANCA. Clinical relapses often affect different organs from those of the presenting illness, and the nature of the inflammation can be different, for example granulomas can develop in patients previously thought to have microscopic polyangiitis. The Birmingham vasculitis activity score has been developed to provide an objective assessment of disease activity. Use of such tools is essential for clinical trials and can also be useful in routine practice for monitoring the activity of patients with difficult disease.

Prognosis

Conventional treatments are very effective at suppressing disease activity, with numerous reports in recent literature indicating that more than 90 per cent of patients achieve remission. Resolution of inflammation often leaves patients with long-standing damage to particular organs, including the kidney. None the less, up to 70 per cent of patients with focal necrotizing glomerulonephritis severe enough to require dialysis at presentation regain independent renal function, which is sustained for many years.

One-year survival is around 80 per cent, with 5-year survival in the region of 55 to 75 per cent. About 10 per cent of patients with small vessel vasculitis die early, usually with uncontrolled disease. Opportunist infections contribute to deaths later in the first year, but have become uncommon with increased experience in the use of immunosuppressive drugs in this group of patients. Some units use cotrimoxazole as prophylaxis against *Pneumocystis carinii*. Some late deaths are due to unrelated causes in this elderly population, while others can be attributed to complications of therapy including haematological malignancies and other tumours. Patients with severe renal disease have a worse mortality, as inevitably do the elderly.

Dialysis and transplantation

Patients who develop endstage renal failure due to vasculitis appear to do no worse than other groups when treated by dialysis. Similarly, successful transplantation has been repeatedly reported and recurrence of small vessel vasculitis in renal transplants is exceptional.

Special problems in pregnancy

Vasculitis is uncommon in women of child-bearing age and so pregnancy is rare in these conditions. However, Wegener's granulomatosis has occasionally been reported to present in pregnancy and the puerperium, and there are also reports of a relapse of disease activity in the third trimester. Women of child-bearing age should be warned of the potential teratogenic effects of drugs that they may be taking to control vasculitis, for example cyclophosphamide and cotrimoxazole. Azathioprine does not pose a threat to the fetus but is a contraindication to breast feeding.

Renal involvement in other vasculitic disorders

Churg–Strauss syndrome

Churg–Strauss syndrome (see Chapter 17.11.5) is a small to medium-sized vasculitis frequently associated with ANCA and sometimes with focal necrotizing glomerulonephritis. Characteristic clinical features are those of asthma and transient pulmonary infiltrates, eosinophilia, and a necrotizing vasculitis affecting particularly the peripheral nervous system (mononeuritis multiplex), the bowel, and sometimes the heart and the skin. The full blood count shows normochromic normocytic anaemia and eosinophilia, which may exceed 1.5×10^9/litre, and a raised platelet count. The chest radiograph reveals pulmonary infiltrates and, as with other forms of vasculitis, severe pulmonary haemorrhage can occur. Renal disease is much less common in Churg–Strauss syndrome than in Wegener's granulomatosis or microscopic polyangiitis. When present, it consists of a focal necrotizing glomerulonephritis or, less commonly, a severe interstitial nephritis with large numbers of eosinophils.

Steroids have been the main treatment for many years, but many would advocate the concurrent use of other immunosuppressive drugs in those with severe disease. This should include those with life-threatening pulmonary haemorrhage, apidly developing mononeuritis multiplex, or focal necrotizing glomerulonephritis. Overall the prognosis of Churg–Strauss syndrome is good with initial clinical remission being achieved in over 90 per cent of patients and a 5-year survival of in excess of 75 per cent. However, it should be emphasized that patients with mononeuritis may be left with considerable disability.

Polyarteritis nodosa

Polyarteritis nodosa is an uncommon condition characterized by necrotizing vasculitis affecting small and medium-sized muscular arteries. Affected arteries often develop aneurysmal swellings (Fig. 6 and Plate 4), which can be palpable as nodules when they occur in subcutaneous tissue. By definition the diagnosis excludes patients who also have involvement of capillaries, arterioles, and venules. Typically it is a disease of the middle aged, with males being affected twice as commonly as females. Patients usually present with fever, weight loss, and night sweats. Other symptoms depend on which vessels are involved. Abdominal pain is common and is due to intestinal ischaemia or pressure from aneurysms; some patients present with symptoms of hypertension caused by renal ischaemia.

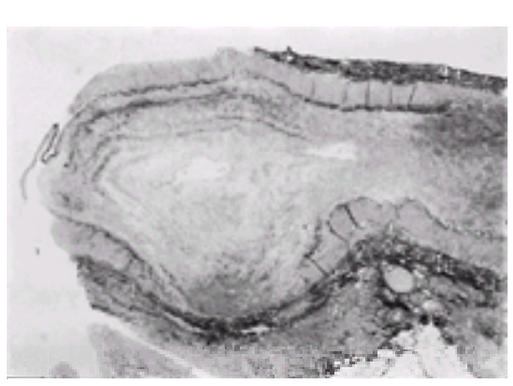

Fig. 6 Morphological appearances of a renal artery from a patient with polyarteritis nodosa. The elastic lamina has been destroyed and the artery has become aneurysmal. (See also Plate 4.)

The aetiology is unknown in most patients, but polyarteritis nodosa has been reported in carriers of hepatitis B virus and also after infection with parvovirus. Further

evidence of infective aetiology comes from reports of polyarteritis nodosa in intravenous drug users.

Diagnosis

Polyarteritis nodosa is associated with the usual non-specific signs of inflammation, but in marked contrast to small vessel vasculitis, ANCA are usually negative. The diagnosis depends on the angiographic demonstration of vasculitis in muscular arteries. All patients should be tested to determine whether they are carriers of hepatitis B or infected with hepatitis C.

Treatment

Polyarteritis nodosa is treated with corticosteroids and cytotoxic drugs using the regimen described for small vessel vasculitis (Table 4). This regimen is effective in the short term in those with hepatitis B, but may be deleterious in the longer term, and success has been reported when these patients are treated with antiviral therapy alone, for example interferon-a or lamivudine.

Conclusion

There has been enormous progress in the management of patients with systemic vasculitis over the past decade, in particular for those with severe renal involvement. Common approaches to classification, diagnosis, and treatment have been developed that for the first time provide a basis for prospective randomized controlled trails of treatment methods. There is no doubt that current treatment regimens are highly effective in the short and medium term, but minimizing long-term toxicity of treatment remains a major issue.

The identification of ANCA in patients with vasculitis has been extremely valuable, both for purely practical reasons as an aid to diagnosis, and more generally as a focus for investigation of these disorders. Nevertheless, it is disappointing that despite nearly two decades of research no clear conclusions can be drawn about why ANCA are so closely associated with systemic vasculitis, or about their roll in pathogenesis.

There is an urgent need for better understanding of pathogenesis to rationalize therapy. Controlled clinical trials designed to optimize the use of currently available drugs are certainly needed, but assessing the potential of the large numbers of new immunosuppressive drugs and recombinant molecules presents impossible difficulties. A better understanding of pathogenesis would help in choosing which other new approaches are most likely to be valuable.

Further reading

General

Heeringa P *et al.* (1998). Animal models of anti-neutrophil cytoplasmic antibody associated vasculitis. *Kidney International* **53**, 253–63.

Hoffman GS (1998). Classification of the systemic vasculitides: antineutrophil cytoplasmic antobodies, consensus and controversy. *Clinical and Experimental Rheumatology* **16**, 111–15.

Jennette JC *et al.* (1994). Nomenclature of systemic vasculitides: proposal of an international consensus conference. *Arthritis and Rheumatism* **37**, 187–92.

Pusey CD *et al.* (1991). Plasma exchange in focal necrotising glomerulonephritis without anti-GBM antibodies. *Kidney International* **40**, 757–63.

Watts RA *et al.* (2000). Epidemiology of systemic vasculitis: a ten-year study in the United Kingdom. *Arthritis and Rheumatism* **43**, 414–19.

Antineutrophil cytoplasmic antibodies

Hagen EC *et al.* (1998). Diagnostic value of standardised assays for anti-neutrophil cytoplasmic antibodies in idiopathic vasculitis. *Kidney International* **53**, 743–53.

Hoffman GS, Specks U (1998). Antineutrophil cytoplasmic antibodies. *Arthritis and Rheumatism* **41**, 1521–37.

McLaren JS *et al.* (2001). The diagnostic value of antineutrophil cytoplasmic antibody testing in a routine clinical setting. *Quarterly Journal of Medicine* **94**, 615–21.

Small vessel vasculitis

Adu D *et al.* (1997). Controlled trial of pulse versus continuous prednisolone and cyclophosphamide in the treatment of systemic vasculitis. *Quarterly Journal of Medicine* **90**, 401–9.

Exley AR, Bacon PA (1996). Clinical disease activity in systemic vasculitis. *Current Opinion in Rheumatology* **8**, 12–18.

Gaskin G, Pusey CD (1998). Systemic vasculitis. In: Davison AM *et al.*, eds. *Oxford textbook of clinical nephrology*, vol. 2, pp 877–910. Oxford University Press, Oxford.

Guillevin L *et al.* (1997). A prospective, multicenter, randomised trial comparing steroids and pulse cyclophosphamide in the treatment of generalised Wegener's granulomatosis. *Arthritis and Rheumatism* **40**, 2187–98.

Haubitz M *et al.* (1998). Intravenous pulse administration of cyclophosphamide versus daily oral treatment in patients with antineutrophil cytoplasmic antibody-associated vasculitis and renal involvement: a prospective, randomised study. *Arthritis and Rheumatism* **41**, 1835–44.

Jennette JC, Falk RJ (1997). Small-vessel vasculitis. *New England Journal of Medicine* **337**, 1512–23.

Savage COS *et al.* (1985). Microscopic polyarteritis: presentation, pathology and prognosis. *Quarterly Journal of Medicine* **56**, 467–83.

Westman KWA *et al.* (1997). Relapse rate, renal survival, and cancer morbidity in patients with Wegener's granulomatosis or microscopis polyangiitis with renal involvement. *Journal of the American Society of Nephrology* **9**, 842–52.

Churg–Strauss Syndrome

Eustace JA, Nadasdy T, Choi M (1999). The Churg–Strauss syndrome. *Journal of the American Society of Nephrology* **10**, 2048–55.

Guillevin L *et al.* (1999). Churg–Strauss syndrome: clinical study and long term follow-up of 96 patients. *Medicine* **78**, 26–37

Lanham JG *et al.* (1984). Systemic vasculitis with asthma and eosinophilia: a clinical approach to the Churg–Strauss syndrome. *Medicine* **63**, 65–81.

Polyarteritis nodosa

Gayraud M *et al*. (2001). Long-term followup of polyarteritis nodosa, microscopic polyangiitis, and Churg–Strauss syndrome: analysis of four prospective trials including 278 patients. *Arthritis and Rheumatism* **44**, 666–75.

Guillevin L (1999). The treatment of classic polyarteritis nodosa in 1999. *Nephrology Dialysis Transplantation* **14**, 2077–9.

20.10.4 The kidney in rheumatological disorders

D. Adu

Lupus nephritis

Systemic lupus erythematosus is a multisystem autoimmune disease that is characterized by the presence of antinuclear antibodies (see Chapter 18.10.2). The overall survival of patients with systemic lupus erythematosus and a nephritis has improved considerably over the last few decades; from less than 50 per cent survival at 5 years in the 1960s to over 80 per cent survival at 10 years in the 1990s. This is due the wider use of corticosteroids and immunosuppressants and the availability of more effective antihypertensive drugs, antibiotics, renal dialysis, and transplantation.

Pathogenesis

The pathogenesis of systemic lupus erythematosus in general, and lupus nephritis in particular, is complex and multifactorial. Immunological dysregulation leads to the production of autoantibodies to nuclear (in particular double-stranded DNA) and other cellular antigens. The renal lesions of lupus nephritis show glomerular and (less often) tubular deposits of immunoglobulins and complement in a granular pattern indicating immune aggregation. It now seems likely that this is due to *in situ* assembly of antigen–antibody complexes rather than the deposition of immune complexes from the circulation.

Clinical presentation

Renal disease may rarely be the presenting feature of systemic lupus erythematosus, although at presentation 10 to 20 per cent of patients with the condition have evidence of renal involvement, and this develops in about 40 to 50 per cent of patients, typically during the first 5 years after diagnosis. Whilst renal disease is a major complication of systemic lupus erythematosus it is always important to recognize that lupus is a systemic disease and that nephritis typically occurs in patients with extrarenal symptoms such as a rash, arthralgia, Raynaud's phenomenon, and pleuropericarditis. Other major organ systems may be involved including the central nervous system, heart, and lungs.

Proteinuria is found in all patients with lupus nephritis and in 50 to 60 per cent of cases is heavy enough to lead to a nephrotic syndrome. Microscopic haematuria accompanies the proteinuria in about 80 per cent of patients; hypertension is found at presentation in 20 to 50 per cent; and some 20 to 30 per cent present with rapidly deteriorating renal function that may occasionally be severe enough to lead to acute renal failure.

Diagnosis

Immunology

A fluorescent antinuclear test is positive in more than 95 per cent of patients with systemic lupus erythematosus although it lacks specificity as it is also found in other connective tissue diseases. More specific, but less sensitive, tests include antidouble-stranded DNA and anti-Sm (Smith) autoantibodies. (For discussion of immunological tests for systemic lupus erythematosus see Chapter 18.10.1 and Chapter 18.10.2.) In general antidouble-stranded DNA antibody levels reflect disease activity, particularly if accompanied by falling complement levels, but as regards the kidney they are less consistently related to features of active glomerulonephritis. Reduced serum concentrations of the complement proteins C1q and C4 as well as C3 indicate activation of the classical pathway of complement. These are useful in the diagnosis of systemic lupus erythematosus, but although more commonly found in lupus nephritis they are not useful in predicting its onset. Patients with lupus nephritis have antibodies to phospholipids in 30 to 50 per cent of cases, resulting in prolongation of the partial thromboplastin time and leading to the term lupus anticoagulant.

Pathology

A renal biopsy is justified when there is evidence of glomerular disease in the form of proteinuria (more than 200 mg/24 h), microscopic haematuria, a urinary sediment indicative of active nephritis (more than 10 dysmorphic red blood cells per high-power field and/or casts of red and white blood cells), or renal insufficiency. Histology allows an assessment of disease activity and provides a basis for therapy and prognosis.

A distinctive feature of lupus nephritis on light microscopy is the variability of the glomerular changes seen in a single biopsy, and sometimes within the same glomerulus. This makes classification of renal histology difficult, but that most widely used is the modified World Health Organization (**WHO**) classification shown in Table 1. Segmental glomerular thrombosis, necrosis, and extracapillary proliferation (crescents) are frequently found in association with the proliferative type lesions (WHO class III and IV). On immunofluorescent microscopy there is often florid deposition of immunoglobulins IgG, IgA, and IgM as well as complement proteins C3, C4, and C1q.

Patients with minimal changes or mesangial glomerulonephritis (WHO class I and II lesions) (Fig. 1 and Plate 1) usually have an inherently low rate of progressive renal failure. Patients with membranous nephropathy (WHO class V) have an intermediate prognosis for renal function. By contrast, patients with focal or diffuse proliferative glomerulonephritis (WHO class III and IV) (Fig. 2 and Plate 2) have a high risk of progressive renal failure.

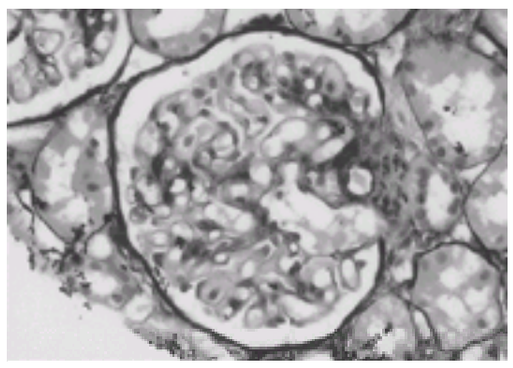

Fig. 1 Lupus nephritis. The glomerulus has mild mesangial increase (WHO class II). Periodic acid-methenamine silver staining (×50). (By courtesy of Dr A. J. Howie.) (See also Plate 1.)

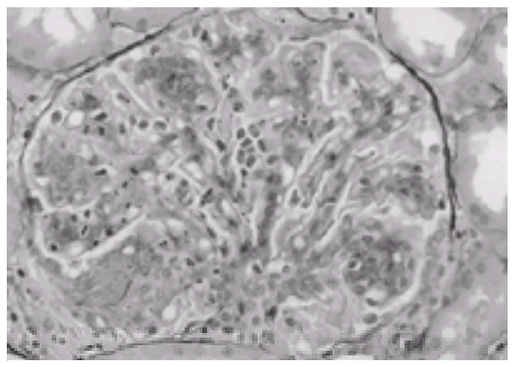

Fig. 2 Lupus nephritis. The glomerulus has marked mesangial increase with wire loops, a few doubled basement membranes and segmental lesions (WHO class IV). Periodic acid-methenamine silver staining (×40). (By courtesy of Dr A. J. Howie. (See also Plate 2.)

Treatment

There are several considerations in the approach to the treatment of patients with lupus nephritis. The first is based on the histological severity of the renal lesion. The second is based on the severity of the clinical presentation. The third consideration is the choice of therapy for inducing remission of acute disease and for maintaining remission and treating relapses. The heterogeneity of the clinical course of lupus nephritis and the relatively few randomized controlled trials means that decision making is difficult and there are still substantial disagreements on the optimum treatment. Steroids and/or immunosuppressants are used—drugs with major toxicities that need to be offset against any benefit.

Treatments for particular classes of lupus nephritis

Mesangial proliferative glomerulonephritis (WHO class II)

Most such patients present with proteinuria and microscopic haematuria, often with little in the way of renal impairment. There are no controlled trials to guide treatment. We treat such patients with corticosteroids in the hope that this will prevent progression to a more severe glomerulonephritis, but this is not certain.

Membranous nephropathy (WHO class V)

In patients with lupus nephritis the frequency of membranous nephropathy is approximately 12 per cent when the definition of the renal histology is confined to pure membranous nephropathy with or without mild mesangial hypercellularity, expansion, and scattered deposits (WHO classes Va and Vb). With the revision of the WHO criteria in 1995, biopsies with focal segmental proliferative or diffuse proliferative glomerulonephritis in addition to membranous changes are now classified as WHO classes III and IV because they behave similarly: this causes some difficulties in interpreting earlier studies where these appearances were classified as Vc and Vd.

The clinical presentation of lupus membranous nephropathy is with proteinuria and in about 50 per cent of cases a nephrotic syndrome. Patients with WHO class Va and Vb lesions have a low rate of progressive renal failure. There are no controlled trials of treatment and thus there is no consensus on treatment. In some studies patients with WHO class Va and Vb disease have been treated with prednisolone, with a smaller proportion also receiving pulses of methylprednisolone or oral cyclophosphamide and azathioprine. By contrast, most patients previously classified as WHO class Vc and Vd have been treated with cyclophosphamide or azathioprine in addition to prednisolone. With these approaches to treatment the 10-year survival free of death and renal failure in WHO class Va and Vb was 72 to 92 per cent and in WHO class Vc and Vd was 35 to 81 per cent.

Most nephrologists treat patients with pure lupus membranous nephropathy with or without minor mesangial proliferation with prednisolone and consider adding in azathioprine as a corticosteroid sparing agent.

Focal and diffuse lupus proliferative glomerulonephritis (WHO class III and IV)

It has been argued that the addition of immunosuppressive drugs to corticosteroids does not improve the prognosis of lupus. Others have concluded that patients treated with prednisolone plus cyclophosphamide or azathioprine have fewer unfavourable outcomes than patients treated with prednisolone alone. Formal meta-analysis has shown that treatment with cyclophosphamide or azathioprine combined with prednisolone reduced the risk of developing endstage renal disease and possibly mortality when compared with prednisolone alone.

A series of clinical trials from the National Institutes of Health provided evidence of the effectiveness of intermittent intravenous cyclophosphamide together with oral prednisolone in preserving renal function in patients with severe lupus nephritis. This regimen is preferable to continuous oral cyclophosphamide as it leads to less bladder toxicity, although the frequency of gonadal toxicity is unaffected and it is not yet known whether pulse cyclophosphamide is less carcinogenic than continuous oral therapy, although this is unlikely. From the National Institutes of Health data, monthly pulse cyclophosphamide (0.5 to 0.75 g/m ^{2}) adjusted for the glomerular filtration rate and leucocyte count at 10 to 14 days is given monthly for the first 6 months, then quarterly for 18 to 24 months, the longer course of cyclophosphamide being associated with fewer relapses than a shorter 6-month course, but at the expense of greater gonadal toxicity. Preliminary data with pulse oral cyclophosphamide have shown encouraging results and, if validated, will minimize the inconvenience associated with intravenous therapy. To reduce the bladder toxicity of intravenous cyclophosphamide patients should be hydrated either with oral or intravenous fluid and mesna given concomitantly. Prednisolone is given in conjunction with the cyclophosphamide at an initial dose of 0.5 to 1 mg/kg/day for 6 to 8 weeks with gradual tapering, preferably to an alternate day regimen to minimize toxicity.

In other uncontrolled but extensive observations, treatment with intravenous methylprednisolone followed by combined prednisolone and azathioprine or oral cyclophosphamide gave long-term results comparable with those of the National Institutes of Health data. It has also been reported that azathioprine may prevent relapse. However, on the basis of the randomized controlled data, we feel that the National Institutes of Health regimen is a reasonable initial treatment for severe lupus nephritis. Much of the toxicity is due to the prolonged maintenance course of pulse cyclophosphamide: whether conversion after 6 months to azathioprine is as effective in maintaining remission as continued pulse cyclophosphamide remains to be established.

A key feature of the care of patients with lupus nephritis is close monitoring of the white cell count and renal function and detailed surveillance and management of

infection, extrarenal lupus, and hypertension. A summary of the treatment strategies for lupus nephritis is shown in Table 2.

Notes on particular treatments for lupus nephritis

Toxicity of cyclophosphamide

Intravenous cyclophosphamide often leads to nausea and vomiting: giving serotonin antagonists such as ondansetron together with dexamethasone can control this. The most common toxic effect is depression of normal haematopoiesis, which is dose dependent and reversible on discontinuing therapy. A further major side-effect is an increased risk of infections, worsened by the concomitant use of corticosteroids. In particular, an increased incidence of herpes zoster is seen with cyclophosphamide.

Cyclophosphamide is metabolized to phosphoramide mustard and acrolein that are excreted by the kidneys. Acrolein can lead to a haemorrhagic cystitis, which is particularly common with oral cyclophosphamide. The use of intravenous cyclophosphamide with vigorous hydration and concomitant administration of 2-mercaptoethane sulfonate sodium has essentially eliminated bladder complications. Prolonged oral cyclophosphamide is associated with an increased risk of malignancy and it is likely that intravenous cyclophosphamide also carries this risk.

Cyclophosphamide causes dose- and age-related gonadal toxicity with oligospermia in men and premature ovarian failure in women. Few data on gonadal toxicity in men with lupus are available. In one study of six men treated with oral cyclophosphamide at a daily dose of 50 to 100 mg, germinal aplasia occurred after a cumulative dose of 9 to 18 g. All studies show that the risk of ovarian toxicity rises substantially with age and is correlated with the duration of treatment and the cumulative dose of cyclophosphamide. In patients aged less than 25 the risk of ovarian failure after 6 months of monthly intravenous cyclophosphamide was nil, whilst a further 24 months of quarterly cyclophosphamide increased this risk to 17 per cent. Comparable figures for women aged over 31 years were 25 per cent and 100 per cent respectively. By contrast, one out of 20 patients (5 per cent) treated with azathioprine developed ovarian failure. Since lupus nephritis chiefly afflicts women of reproductive age, one must balance this risk of premature ovarian failure, which may be permanent, with the benefits of treatment. Cyclophosphamide, unlike azathioprine, is a potent teratogen and must not be used in pregnancy.

Toxicity of azathioprine

Some patients are intolerant of azathioprine and develop nausea, vomiting, and diarrhoea. It also causes marrow suppression, which can be severe in individuals who have a deficiency of thiopurine methyltransferase. Other toxicities include an increased risk of infection and the development of malignancies with prolonged usage.

Pulse methylprednisolone

Pulse methylprednisolone has been used in at least two different ways. In the first, pulse methylprednisolone has been used in three consecutive daily doses of 0.5 to 1 g at the initiation of treatment of severe proliferative lupus nephritis or for the treatment of renal flares. This has been used together with cyclophosphamide or azathioprine and in conjunction with oral prednisolone. The long-term results of this approach are good, although this approach has not been examined in a randomized controlled study. The second way that methylprednisolone has been used is as monthly pulses together with continuous low-dose prednisolone. However, this is less effective in inducing remission and preventing endstage renal failure than treatments that include cyclophosphamide and it cannot be recommended.

Plasma exchange

Several studies have examined the role of plasmapheresis in the treatment of patients with lupus nephritis: although it was well tolerated with few adverse effects the impact on renal function was disappointing. Controlled trials of plasmapheresis in patients with all types of proliferative or membranous glomerulonephritis showed no benefit over treatment with prednisolone and immunosuppressants alone. We currently use plasma-pheresis only in patients with a severe diffuse proliferative glomerulonephritis and pulmonary haemorrhage whose disease does not respond to prednisolone and cyclophosphamide.

Intravenous immunoglobulins

Uncontrolled studies have shown a temporary benefit in patients with systemic lupus erythematosus from the infusion of high doses of intravenous immunoglobulin. Prospective controlled studies are needed to evaluate critically the efficacy of this therapy.

Cyclosporin A

Several studies have examined the effectiveness of cyclosporin A in the treatment of lupus nephritis: none of these were controlled and it is difficult to discern whether cyclosporin was of any benefit. The nephrotoxicity of cyclosporin is a major problem, and pending randomized controlled studies comparing this drug with other immunosuppressive agents we cannot recommend its use in lupus nephritis.

Methotrexate

Methotrexate may be useful as a steroid-sparing agent in lupus with arthritis and serositis, and may have potential benefits in mild nephritis. However, methotrexate is excreted by the kidneys and cannot be used safely in patients with renal impairment.

Mycophenylate mofetil

The active metabolite of mycophenylate mofetil inhibits inosine monophosphate dehydrogenase and thereby the *de novo* pathway of guanosine nucleotide synthesis. Preliminary experimental and clinical studies have indicated a potential role in systemic lupus erythematosus, but this awaits confirmation in controlled trials.

Prognostic factors in lupus nephritis

Patients with proliferative glomerulonephritis (WHO classes III and IV) tend to have a worse outcome for renal function than those with milder lesions, although with treatment this difference is now small. The combination of severe active and chronic histological changes on a renal biopsy adversely affects outcome. Even in the face of active lupus nephritis, patients without chronic histological changes have a lower risk of developing renal failure, 90 per cent or more remaining free of renal failure after 10 years. A number of clinical variables are associated with a greater probability of renal progression in lupus nephritis, including low haematocrit, raised serum creatinine level at diagnosis, 'nephritic flares', hypertension, heavy proteinuria, and poor socio-economic status.

Long-term outcome

Studies reported in the 1990s show a 10-year patient survival in lupus nephritis that ranges from 70 to 90 per cent. Renal failure can now be treated by dialysis and transplantation, the major causes of death now being treatment-related sepsis, which occurs early, and myocardial ischaemia, which occurs late. The other major cause of death is extrarenal lupus.

Between 17 and 30 per cent of patients with lupus nephritis develop endstage renal failure by 10 years. Both haemodialysis and continuous ambulatory peritoneal dialysis are well tolerated and there is a tendency for the activity of lupus disease to diminish after the start of dialysis. We discontinue immunosuppressants in patients on dialysis if there is no overt disease activity, only persisting with a small dose of prednisolone. Overall survival on dialysis is good, being 75 per cent at 10 years. After transplantation, graft survival and function in patients with lupus are comparable to those obtained in patients with other diseases, and recurrence of lupus nephritis is uncommon.

Renal disease in systemic sclerosis

Systemic sclerosis is a systemic disorder characterized by skin thickening due to the deposition of collagen in the dermis (see Chapter 18.10.3). Adverse prognostic features are renal, cardiac, and pulmonary involvement. A major complication is the development of scleroderma renal crisis, which is characterized by the abrupt

onset of severe hypertension, usually with retinopathy, together with the rapid deterioration of renal function and heart failure. Scleroderma renal crisis develops in approximately 8 to 15 per cent of patients with diffuse systemic sclerosis, the most important risk factor being the rapid progression of diffuse skin disease. It usually occurs early, within 3 years of the onset of illness, and develops more commonly in the autumn and winter.

Pathogenesis

The pathogenetic mechanisms leading to renal damage in systemic sclerosis are not known. Whilst plasma renin activity is almost always raised in scleroderma renal crisis, there is no evidence that this occurs before the development of this complication and plasma renin activity does not predict the problem. Patients with systemic sclerosis may show cold-induced reduction in renal perfusion and increased plasma renin activity, but this does not correlate with the presence of renal histological vascular abnormalities. There is evidence of endothelial activation in patients who develop renal damage with raised serum levels of circulating endothelial derived adhesion molecules including s-ELAM, s-VCAM, and s-ICAM, but these are likely to reflect the presence of endothelial injury rather than being of pathophysiological significance.

Antecedent hypertension does not increase the risk of development of scleroderma renal crisis, which is as common in men as in women, although systemic sclerosis is more common in the latter, with a female to male ratio between 3:1 and 4:1. In one retrospective case-controlled study the risk of scleroderma renal crisis was increased by prior treatment with steroids (more than 15 mg prednisolone/day) with an odds ratio of 4.37 (95 per cent confidence interval 2.03–9.42) and reduced by treatment with penicillamine (odds ratio 0.41, 95 per cent confidence interval 0.24–0.69). One report suggests that cyclosporin A may predispose to the development of scleroderma renal crisis.

Pathology

The smaller arcuate and interlobular arteries are predominantly involved in scleroderma renal crisis, showing intimal hyperplasia with concentric mucoid intimal degeneration, but the internal and external elastic laminae remain intact. In addition the adventitia of interlobular arteries show an abnormal degree of fibrosis. There is fibrinoid necrosis of afferent arterioles and glomeruli and also glomerular thrombosis. Ischaemia of the glomerular tuft leads to wrinkling and thickening of the glomerular basement membrane and glomerular sclerosis (Fig. 3 and Plate 3). These lesions resemble those seen in accelerated hypertension or the haemolytic uraemic syndrome, although the vessels involved tend to be larger and adventitial fibrosis is not seen in accelerated hypertension.

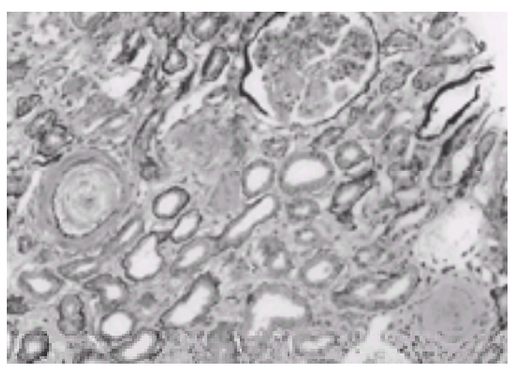

Fig. 3 Scleroderma kidney. A small artery has concentric mucoid intimal thickening, an arteriole has thrombosis and fibrinoid necrosis, and tubules and a glomerulus have ischaemic damage. Periodic acid-methenamine silver staining (×25). (By courtesy of Dr A. J. Howie.) (See also Plate 3.)

Clinical presentation

Hypertensive scleroderma renal crisis

The clinical presentation is typically with the symptoms of malignant hypertension with headaches, blurred vision, fits, and heart failure. Renal function is impaired and deteriorates rapidly. The hypertension is almost always severe with a diastolic blood pressure in excess of 100 mmHg in 90 per cent of patients. There is hypertensive retinopathy in about 85 per cent of cases with exudates and haemorrhages and at times papilloedema.

Normotensive scleroderma renal crisis

Scleroderma renal crisis can also develop in individuals with a normal blood pressure. They are more likely to have a microangiopathic haemolytic anaemia (90 per cent versus 38 per cent), thrombocytopenia (83 per cent versus 21 per cent), and pulmonary haemorrhage than patients with hypertensive scleroderma renal crisis.

Diagnosis

The clinical presentation described above in a patient with the typical diffuse skin thickening of systemic sclerosis is diagnostic. Typically the renal impairment is accompanied by a microangiopathic haemolytic anaemia with thrombocytopenia and fragmented red blood cells (schistocytes or burr cells). Once the blood pressure has been well controlled for at least 7 days then, if there is doubt, the diagnosis can be established by renal histology.

Treatment

Scleroderma renal crisis is a medical emergency. The hypertension should be treated with an angiotensin converting enzyme inhibitor, which can also help with the treatment of the heart failure. The aim should be for a slow and gradual reduction in blood pressure as an abrupt fall can lead to cerebral ischaemia or infarction, as it can in accelerated phase hypertension. Calcium channel blockers may be required in addition to the angiotensin converting enzyme inhibitors. Deterioration of renal function in these patients is often rapid and they can precipitately develop pulmonary oedema, hence they should be treated in a hospital with facilities for dialysis.

Prognosis

Prior to the early 1970s scleroderma renal crisis was almost always a fatal illness with most patients dying within a year. The survival improved slightly with the use of dialysis and better hypotensive agents, but it is only since the introduction of angiotensin converting enzyme inhibitors that prognosis has improved. In one study of 23 patients treated with captopril, 20 responded favourably and in 14 the serum creatinine fell. After a median follow-up of 29 months, six patients had died and four remained dependent on dialysis. Prior to treatment with angiotensin converting enzyme inhibitors, patients with normotensive scleroderma renal crisis had a worse prognosis, with a 1-year survival of 13 per cent as compared with 35 per cent in those who were hypertensive.

Although the renal vascular lesions are acute, recovery of function is unusual once renal failure has developed and most patients require long-term dialysis. However, some patients may recover renal function after a period of dialysis, and there is a tendency for the skin lesions of scleroderma to improve on this treatment.

Renal disease in rheumatoid arthritis

Death certificate and autopsy studies in rheumatoid arthritis show that there is an excess mortality from renal failure, which accounts for between 3 and 20 per cent of deaths. About a half of these are due to amyloid, the remainder being due to nephritis and renal infections. There are no good figures on the prevalence of renal disease during life in rheumatoid arthritis, although this does seem to be lower.

There are three broad categories of renal disease in rheumatoid arthritis (Table 3). The first and the most common is nephrotoxicity from the drugs used in the treatment (see Chapter 18.5). Gold and penicillamine lead to proteinuria and glomerulonephritis in between 10 and 30 per cent of patients, often severe enough to

cause a nephrotic syndrome. Non-steroidal anti-inflammatory drugs are widely used for pain relief and are associated with the development of a variety of renal syndromes ranging from a reversible reduction in glomerular filtration rate to acute renal failure, either due to an acute tubular necrosis or an acute interstitial nephritis. The latter may be complicated by nephrotic range proteinuria. The second major but diminishing cause of renal disease in rheumatoid arthritis is amyloidosis. Thirdly, patients with rheumatoid arthritis may develop a renal vasculitis and also a glomerulonephritis.

Secondary amyloidosis in rheumatoid arthritis

Secondary amyloidosis results from deposition of fibrils containing amyloid A protein that is antigenically related to the acute phase reactant serum amyloid A (see Chapter 11.12.1 and Chapter 11.12.4 for further discussion). Rheumatoid arthritis is the commonest disease producing secondary amyloidosis in developed countries. At autopsy, prevalence rates of 8 to 17 per cent are found, whilst data from biopsy series show a lower prevalence of around 5 to 10 per cent. There is some evidence for a decline in prevalence of amyloid over the last 20 years, and in the last 5 years the incidence appears to have dropped dramatically (unpublished evidence). The reason for this is likely to be much more aggressive therapy, with fewer patients being left with a persistently elevated acute phase response.

Clinical presentation and diagnosis

The presentation of renal amyloid is with proteinuria that is often severe enough to cause a nephrotic syndrome. Renal vein thrombosis is particularly common. Diagnosis is established by renal biopsy (Fig. 4 and Plate 4), where histological Congo red staining, which is birefringent in polarized light, is characteristic of amyloid. This staining is abolished by potassium permanganate in reactive amyloidosis but not in primary amyloidosis. Monoclonal and polyclonal antibodies that specifically bind amyloid A are now available and are of use for histological diagnosis. The diagnosis of amyloid has also been aided by the availability of scans using radiolabelled serum amyloid P (SAP) protein, utilizing the strong calcium dependent affinity of SAP for amyloid fibrils of any protein type.

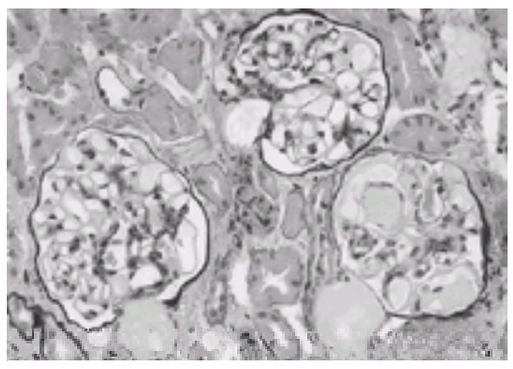

Fig. 4 Amyloidosis in rheumatoid arthritis. Arterioles and glomeruli contain acellular masses of amyloid. Periodic acid-methenamine silver staining (×40). (By courtesy of Dr A. J. Howie.) (See also Plate 4.)

Treatment and prognosis

There is no specific therapy for amyloid A amyloidosis, the general principle being suppression of the underlying chronic inflammation. Uncontrolled evidence suggests that aggressive treatment of rheumatoid arthritis may be effective in delaying the deterioration of renal function in patients with renal amyloid. There are some reports that treatment with prednisolone and cyclophosphamide or methotrexate can induce remission of the nephrotic syndrome due to amyloid in patients with rheumatoid arthritis, but in other studies no benefit was seen. Randomized controlled studies are needed to establish the role of aggressive treatment of renal amyloid in this condition.

Renal amyloid leads to progressive renal failure. After 5 years 50 per cent of patients develop endstage renal failure and this rises to 90 per cent at 10 years. Treatment of endstage renal failure from amyloid is by dialysis and renal transplantation.

Gold and penicillamine nephropathy

Clinical features

The most frequent presenting feature is proteinuria, which occurs in approximately 10 per cent of patients receiving gold and up to 30 per cent of those taking penicillamine. This progresses to the nephrotic syndrome in 30 and 16 per cent respectively. Haematuria is uncommon, although it is seen more frequently with penicillamine, and still requires the exclusion of other causes when occurring in the context of therapy with these drugs. Renal function is usually normal.

Pathology

About 80 per cent of patients who present with D-penicillamine or gold-induced proteinuria will have a membranous glomerulonephritis. Subepithelial spikes and a mild increase in mesangial cells are usually seen, and the diagnosis can be confirmed with immunofluorescence/immunoperoxidase microscopy that shows granular subepithelial deposits of predominantly IgG. On electron microscopy electron-dense subepithelial deposits are seen.

Other renal lesions are less common and include a mesangial glomerulonephritis, minimal change nephropathy, and tubulointerstitial inflammation. Penicillamine may lead to the development of a rapidly progressive glomerulonephritis with crescents and the clinical picture of Goodpasture's syndrome, also to a renal vasculitis.

Treatment and prognosis

In general gold and penicillamine should be discontinued when significant proteinuria develops (more than 0.5 g/24 h). Renal biopsy should be confined to those patients who have deteriorating renal function, or who fail to improve after withdrawal of the drug. Regular monitoring of proteinuria and the glomerular filtration rate are mandatory. No specific immunosuppression is required although supportive measures for the nephrotic syndrome are given as indicated (see Section 20.3).

After cessation of the drug, proteinuria peaks at around a month then gradually disappears; the majority of patients will have clear urine by 1 year and almost all will achieve this by 2 years. Renal function does not deteriorate in uncomplicated cases.

The susceptibility to gold- or penicillamine-induced nephrotoxicity is linked to the major histocompatibility genes: HLA DR3 confers a relative risk of 14.0 to 32.0 for gold-induced nephropathy and 3.2 to 10.0 for D-penicillamine-induced nephropathy. There also appear to be metabolic factors that determine the toxicity of these drugs: individuals with poor sulphoxidation appear to be at increase risk. Given this basis it is not surprising that rechallenge with the same drug at the same dose usually leads to a recurrence of the renal problem, although a lower dose may be tolerated. The dilemma of whether to restart treatment is now less of a problem because of the increasing number of alternative therapies.

Cyclosporin A nephrotoxicity

The renal toxicity of cyclosporin in rheumatoid arthritis is well documented, hence in these patients cyclosporin should be started at a dose of 2.5 mg/kg/day and not exceeding 5 mg/kg/day, with a reduction of cyclosporin if creatinine rises to 130 per cent of baseline. Indeed, so sensitive is the rise in creatinine in patients with rheumatoid arthritis that other measures of renal function are used only to confirm changes.

Non-steroidal anti-inflammatory drugs increase the nephrotoxicity of cyclosporin A, which can lead to chronic irreversible renal failure: this is more common with doses in excess of 5 mg/kg, in patients with pre-existing renal impairment, in elderly patients, and in those treated for more than 6 months. Renal function should be carefully

monitored in patients on cyclosporin therapy.

Non-steroidal anti-inflammatory drugs

Non-steroidal anti-inflammatory drugs are potentially nephrotoxic and in patients with rheumatoid arthritis may lead to a reversible reduction in glomerular filtration rate, acute tubular necrosis, an acute interstitial nephritis often with heavy proteinuria, renal papillary necrosis, and chronic tubulointerstitial nephritis.

Glomerulonephritis

The most commonly described glomerulonephritis in rheumatoid arthritis that is not related to drug use is a mesangial proliferative glomerulonephritis, which in many cases is accompanied by IgA deposits (IgA nephropathy). The other major type of glomerulonephritis reported in rheumatoid arthritis is membranous nephropathy.

Renal vasculitis

The clinical spectrum of rheumatoid arthritis includes a systemic necrotizing vasculitis with involvement of blood vessels ranging in size from capillaries to small and medium-sized arteries. The clinical presentation includes nailfold infarcts, a leucocytoclastic vasculitis, a peripheral neuropathy, pericarditis, gastrointestinal infarcts, and renal vasculitis. Renal abnormalities are found in about 25 per cent of patients with rheumatoid vasculitis, usually microscopic haematuria, proteinuria, and renal impairment. Renal histology shows a large vessel renal arteritis and a segmental necrotizing glomerulonephritis with crescent formation (vasculitic glomerulonephritis) (Fig. 5 and Plate 5). Treatment is with prednisolone and cyclophosphamide, usually leading to improvement of renal function.

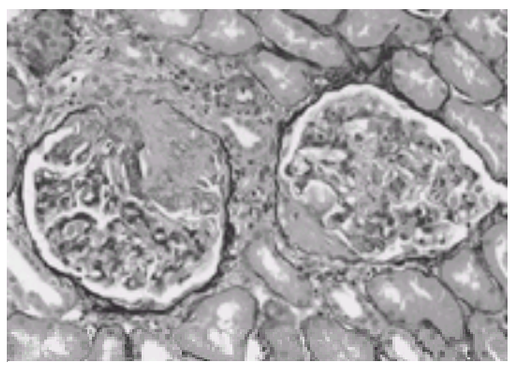

Fig. 5 Vasculitic glomerulonephritis in rheumatoid arthritis. Two glomeruli have sharply defined segmental lesions where there has been disruption of the tuft and partial obliteration of Bowman's space. Periodic acid-methenamine silver staining (×32). (By courtesy of Dr A. J. Howie.) (See also Plate 5.)

Renal disease in juvenile chronic arthritis

Renal failure accounts for 38 per cent of deaths in patients with juvenile chronic arthritis. Proteinuria is found in between 3 and 12 per cent and microscopic haematuria in between 3 and 8 per cent of these patients. Nephrotic range proteinuria is commonly due to renal amyloid, found in between 1.2 and 6.7 per cent of patients with juvenile chronic arthritis, whilst haematuria and proteinuria may be due to amyloid or to gold treatment. Interstitial nephritis is also common and may be due to drug treatment. Amyloid is a major problem, accounting for more than 40 per cent of deaths, the majority due to renal failure.

Renal disease in primary Sjögren's syndrome

Sjögren's syndrome is characterized by a lymphocytic infiltration of exocrine glands leading to a dry mouth (xerostomia) and dry eyes (keratoconjunctivitis sicca) (see Chapter 18.10.4). It may be primary or secondary to a variety of autoimmune disorders including rheumatoid arthritis, systemic lupus erythematosus, systemic sclerosis, and mixed connective tissue disorder.

Clinically significant renal disease has been reported in about 10 to 25 per cent of patients with Sjögren's syndrome. The most common renal disorder is mild and often subclinical distal renal tubular acidosis, impairment of urinary concentration, and rarely hypokalaemia. Clinical manifestations of these renal tubular disorders include the development of renal calculi, polyuria, and rarely hypokalaemic periodic paralysis. Renal biopsy in these patients shows a tubulointerstitial nephritis with interstitial lymphocytic infiltrates. Glomerulonephritis is rare in primary Sjögren's syndrome, and is most commonly membranoproliferative glomerulonephritis or membranous nephropathy.

Renal disease in mixed connective tissue disease

Some patients with a connective tissue disorder do not fit easily into the accepted definitions of a single disease. In patients with mixed connective tissue disease there is the sequential or concurrent development of the clinical features of systemic lupus erythematosus, systemic sclerosis, polymyositis, and less commonly of rheumatoid arthritis (see Chapter 18.10.2). Renal involvement is found in 10 to 47 per cent of patients with mixed connective tissue disease, the clinical presentation being with asymptomatic proteinuria or haematuria and less commonly with a nephrotic syndrome.

Membranous nephropathy and a mesangial proliferative glomerulonephritis are the most common histological changes, found in 34 and 30 per cent of cases respectively. A focal or diffuse proliferative glomerulonephritis is found in 17 per cent, a mixed lesion with membranous nephropathy in 5 per cent, and in 7 per cent renal histology is normal. Immunofluorescent microscopy of glomeruli in patients with mixed connective tissue disease has shown immunoglobulin and complement deposits; dense deposits are found on electron microscopy.

Treatment of renal disease in mixed connective tissue disease is with steroids, initially in high doses, subsequently tapering to a low maintenance dose over weeks. Treatment of patients with a nephrotic syndrome with high-dose steroids leads to a significant reduction of proteinuria in 62 per cent of cases. Whether those with renal disease resistant to steroids would benefit from the addition of immunosuppressant drugs is not known. Some 14 per cent of patients with mixed connective tissue disease and renal disease develop chronic renal failure.

Further reading

Adu D, *et al*., eds. (2001). *Rheumatology and the kidney*. Oxford University Press, Oxford.

Lupus nephritis

Austin HA *et al*. (1986). Therapy of lupus nephritis: controlled trial of prednisolone and cytotoxic drugs. *New England Journal of Medicine* **314**, 614–19.

Balow JE *et al*. (1996). Management of lupus nephritis. *Kidney International* **53**, S88–S92.

Bansal VK, Beto JA (1997). Treatment of lupus nephritis: a meta-analysis of clinical trials. *American Journal of Kidney Diseases* **29**, 193–9.

Berden J (1997). Lupus nephritis (nephrology forum). *Kidney International* **52**, 538–58.

Boumpas DT *et al*. (1992). Controlled trial of pulse methylprednisolone versus two regimes of pulse cyclophosphamide in severe lupus nephritis. *The Lancet* **340**, 741–5.

Cameron J (1999). Lupus nephritis. *Journal of the American Society of Nephrology* **10**, 1–17.

Cameron JS (1994). Lupus nephritis in childhood and adolescence. *Pediatric Nephrology* **8**, 230–49.

Donadio JV, Glasscock RJ (1993). Immunosuppressive drug therapy in lupus nephritis. *American Journal of Kidney Diseases* **21**, 239–50.

Fessel WF (1988). Epidemiology of systemic lupus erythematosus. *Rheumatic Disease Clinics of North America* **14**, 15–23.

Lewis EJ *et al*. (1992). A controlled trial of plasmapheresis therapy in severe lupus nephritis. *New England Journal of Medicine* **326**, 1373–9.

Moroni G *et al*. (1996). 'Nephritic flares' are predictors of bad long-term renal outcome in lupus nephritis. *Kidney International* **50**, 2047–53.

Pasquali S *et al*. (1993). Lupus membranous nephropathy: long-term outcome. *Clinical Nephrology* **39**, 175–82.

Tse WY, Adu D (1999). Treatment of glomerulonephritis in systemic disease. In: Pusey CD, ed. *The treatment of glomerulonephritis*, pp 143–76. Kluwer Academic, Dordrecht.

Walport M (1997). The pathogenesis of systemic lupus erythematosus. In: Davison AM *et al.*, eds. *Oxford textbook of clinical nephrology*, vol. 2, pp 917–35. Oxford University Press, Oxford.

Systemic sclerosis

Helfrich D *et al*. (1989). Normotensive renal failure in systemic sclerosis. *Arthritis and Rheumatism* **32**, 1128–34.

Steen V *et al*. (1984). Factors predicting development of renal involvement in progressive systemic sclerosis. *American Journal of Medicine* **76**, 779–86.

Steen V, Medsger TJ (1998). Case-control study of corticosteroids and other drugs that either precipitate or protect from the development of scleroderma renal crisis. *Arthritis and Rheumatism* **41**, 1613–19.

Thurm R, Alexander J (1984). Captopril in the treatment of scleroderma renal crisis. *Archives of Internal Medicine* **144**, 733–5.

Traub Y *et al*. (1983). Hypertension and renal failure (scleroderma renal crisis) in progressive systemic sclerosis: review of a 25-year experience with 68 cases. *Medicine* **62**, 335–52.

Wasner C, Cooke R, Fries J (1978). Successful medical treatment of scleroderma renal crisis. *New England Journal of Medicine* **299**, 873–5.

Rheumatoid arthritis

Adu D *et al*. (1993). Glomerulonephritis in rheumatoid arthritis. *British Journal of Rheumatology* **32**, 1008–11.

Anttila R (1972). Renal involvement in juvenile rheumatoid arthritis. A clinical and histopathological study. *Acta Paediatrica Scandinavica Supplement*, **227**, 1–73.

Boers M *et al.* (1987). Renal findings in rheumatoid arthritis: clinical aspects of 132 necropsies. *Annals of the Rheumatic Diseases* **46**, 658–63.

Boers M (1990). Renal disorders in rheumatoid arthritis. *Seminars in Arthritis and Rheumatism* **20**, 57–68.

Cohen DJ, Appel GB (1992). Cyclosporine: nephrotoxic effects and guidelines for safe use in patients with rheumatoid arthritis. *Seminars in Arthritis and Rheumatism* **21** (suppl. 3), 43–8.

Hall CL *et al*. (1987). The natural course of gold nephropathy: long term study of 21 patients. *British Medical Journal* **295**, 745–84.

Hall CL *et al.* (1988). Natural course of penicillamine nephropathy: a long term study of 33 patients. *British Medical Journal* **296**, 1085–6.

Harper L *et al*. (1997). Focal segmental necrotizing glomerulonephritis in rheumatoid arthritis. *Quarterly Journal of Medicine* **90**, 125–32.

Helin H *et al*. (1986). Mild mesangial glomerulopathy, a frequent finding in rheumatoid arthritis patients with haematuria or proteinuria. *Nephron* **42**, 224–30.

Honkanen E *et al*. (1987). Membranous glomerulonephritis in rheumatoid arthritis not related to gold or D-penicillamine therapy: a report of four cases and review of the literature. *Clinical Nephrology* **27**, 87–93.

Kuznetsky KA *et al*. (1986). Necrotizing glomerulonephritis in rheumatoid arthritis. *Clinical Nephrology* **26**, 257–64.

Landewe RB *et al*. (1996). Longterm low dose cyclosporine in patients with rheumatoid arthritis: renal function loss without structural nephropathy. *Journal of Rheumatology* **23**, 61–4.

Maezawa A *et al*. (1994). Combined treatment with cyclophosphamide and prednisolone can induce remission in a patient with nephrotic syndrome with amyloidosis associated with rheumatoid arthritis. *Clinical Nephrology* **42**, 30–2.

Rodriguez F *et al*. (1996) Renal biopsy findings and followup of renal function in rheumatoid arthritis patients treated with cyclosporin A. *Arthritis and Rheumatism* **39**, 1491–8.

Sandler DP *et al.* (1989). Analgesic use and chronic renal disease. *New England Journal of Medicine* **320**, 1238–43.

Scott DGI, Bacon PA, Tribe CR (1981). Systemic rheumatoid vasculitis: a clinical and laboratory study of 50 cases. *Medicine* **60**, 288–97.

Sellars L *et al*. (1983). Renal biopsy appearances in rheumatoid disease. *Clinical Nephrology* **20**, 114–20.

Sjögren's syndrome

Moutsopoulos H *et al*. (1978). Immune complex glomerulonephritis in sicca syndrome. *American Journal of Medicine* **64**, 955–60.

Moutsopoulos H *et al*. (1991). Nephrocalcinosis in Sjögren's syndrome. *Journal of Internal Medicine* **230**, 187–91.

Shearn M, Tu WH (1965). Nephrogenic diabetes insipidus and other disorders of renal tubular function in Sjögren's syndrome. *American Journal of Medicine* **39**, 312–18.

Talal N, Zisman E, Schur P (1968). Renal tubular acidosis, glomerulonephritis and immune complex glomerulonephritis in Sjögren's syndrome. *Arthritis and Rheumatism* **11**, 774.

Mixed connective tissue disease

Kitridou RC *et al.* (1986). Renal involvement in mixed connective tissue disease; a longitudinal clinicopathologic study. *Seminars in Arthritis and Rheumatism* **16**, 135–45.

20.10.5 Renal involvement in plasma cell dyscrasias, immunoglobulin-based amyloidoses, and fibrillary glomerulopathies, lymphomas, and leukaemias

P. Ronco

Introduction

Plasma cell dyscrasias are characterized by uncontrolled proliferation of a single clone of B cells, usually with plasma cell differentiation, that is responsible for the secretion in blood of a monoclonal immunoglobulin (**Ig**) or Ig subunit. This monoclonal component can become deposited in tissues, and the recognized spectrum of renal diseases in which there is deposition or precipitation of Ig-related material has expanded dramatically in recent years.

These conditions can be classified into two categories on the basis of their ultrastructural appearances (Table 1). Those with organized deposits include diseases with crystal formation, mainly myeloma cast nephropathy; diseases with fibril formation, mainly light-chain amyloidosis; and diseases with microtubule formation, including cryoglobulinaemia kidney and immunotactoid glomerulonephritis. A second category of diseases is characterized by the presence of non-organized granular electron-dense deposits made of light and/or heavy chains along the basement membranes of many tissues, most importantly the kidney. First described by Randall and associates, they are named monoclonal immunoglobulin deposition diseases. It is now well established that the spectrum of plasma cell dyscrasia-related renal complications is due to intrinsic properties of the monoclonal component.

Except for myeloma cast nephropathy, diagnosis often relies on careful analysis of a biopsy specimen taken from the kidney, which should systematically include immunohistochemical studies with specific antibodies and also electron microscopy in all ambiguous cases. Since most of these patients will develop renal failure, it is essential to identify the underlying plasma cell dyscrasia because appropriate treatments may halt the extension of visceral deposits, and even induce their regression. Except in patients with myeloma cast nephropathy, who usually present with a high-mass myeloma, 'malignancy' more often results from life-threatening visceral deposits than from the Ig-secreting clone itself.

Renal involvement in Ig light-chain amyloidosis

Definition and epidemiology

Amyloidosis is a general term for a family of diseases defined by morphological criteria and characterized by deposition in extracellular spaces of a proteinaceous material that stains with Congo red and is metachromatic. Amyloid deposits are composed of a felt-like array of 10-nm wide rigid, linear, aggregated fibrils of indefinite length with a b-pleated sheet configuration. They occur in a variety of conditions including Alzheimer's disease and other neurodegenerative disorders, tumoural and inflammatory diseases, and plasma cell dyscrasias. The various types of amyloidosis differ essentially by the nature of the precursor protein that yields the main component of fibrils, and are classified accordingly. (See Chapter 11.12.4 for further discussion.)

Light-chain amyloidosis has become the most frequent form of amyloidosis with renal involvement. Amyloid deposits are found in approximately 10 per cent of myeloma patients, their prevalence reaching 20 per cent in those with pure light-chain myeloma. (Fig. 1 and Plate 1)

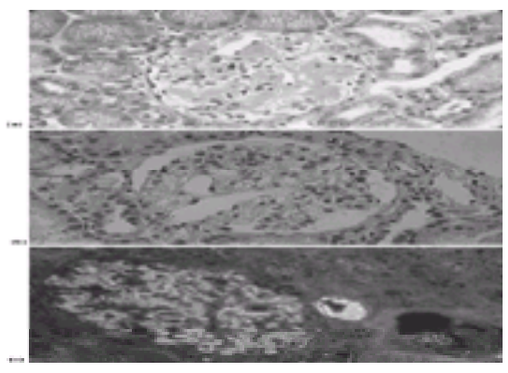

Fig. 1 Light-chain amyloidosis. (a) Amyloid deposits in a renal glomerulus (Masson's trichrome stain, ×312). (b) Congo red stain. Apple-green/yellow dichroism under polarized light (×312). (c) Immunofluorescence with anti-l antibody. Note glomerular and arteriolar deposits (×312). (From Béatrice Mougenot's personal collection.) (See also Plate 1.)

Clinical presentation

The main clinical features of light-chain amyloidosis at presentation are fatigue (62 per cent) and weight loss (52 per cent), followed by purpura (15 per cent), pain (5 per cent), and gross bleeding (3 per cent). Hepatomegaly is found in 24 per cent of patients, and macroglossia in 9 per cent. A palpable spleen and lymphadenopathy can also be found.

Proteinuria is the usual symptom of renal amyloidosis, detected in 55 per cent of patients at presentation and often progressing to a severe nephrotic syndrome, which can be complicated by renal vein thrombosis. Haematuria is uncommon, and when present should prompt examination for a bleeding lesion of the urinary tract. Progressive decline in renal function usually occurs, leading finally to endstage renal failure. In those rare patients in whom renal tubulointerstitial deposits predominate, renal failure may progress without a nephrotic stage, and in some of these cases renal tubular dysfunction may be the presenting problem, including Fanconi syndrome, renal tubular acidosis, or even nephrogenic diabetes insipidus. Hypertension is uncommon but may develop concomitantly with renal failure. The kidneys are generally of normal size or large, even when renal function is impaired.

Light-chain amyloidosis can infiltrate almost any organ and thus be responsible for a wide variety of clinical manifestations. It frequently involves the tongue, gastrointestinal tract, peripheral neural system, carpal tunnel, heart, and skin. Purpuric macules of the superior eyebrow are very typical of light-chain amyloidosis.

Diagnosis

Light-chain amyloidosis should be suspected when the clinical manifestations described above are associated with a monoclonal component. Monoclonal light chains can be detected by immunoelectrophoresis of urine in around 73 per cent of cases, with the l isotype being twice as frequent as the _. With the use of more sensitive techniques (immunofixation), a monoclonal Ig is found in the serum and/or the urine in nearly 90 per cent of patients, but still not in all of them. Light-chain amyloidosis is, however, always the result of the proliferation of a plasma cell clone. Fifty six per cent of patients with light-chain amyloidosis have an increased number of plasma cells in the bone marrow, and 15 per cent of them have a true myeloma.

In all cases, diagnosis of light-chain amyloidosis should be established by taking a biopsy specimen from a superficial organ including skin, salivary glands, gum, or by aspiration biopsy of abdominal fat. These biopsies should be performed before biopsies of rectal mucosa (which should include vessels of the submucosa where amyloid deposits usually start) and/or of kidney because of the risk of bleeding complications due to factor X deficiency or amyloid infiltration of vascular walls. After Congo red staining, amyloid deposits appear faintly red and show the characteristic apple-green birefringence under polarized light. Metachromasia is also observed with crystal violet, which stains the deposits in red. In the kidney, the earliest lesions are located in the mesangium, along the glomerular basement membrane, and in the blood vessels. Because there are specific diagnostic and therapeutic strategies depending on the type of protein deposited within tissues, immunofluorescence with specific antisera including anti-_ and anti-l light chains should be performed routinely.

Treatment

Light-chain amyloidosis is a wasting disease with a poor outcome irrespective of the underlying haematological abnormality. Median survival is 18 months in patients treated with chemotherapy (melphalan and prednisolone) and less than 9 months in those given colchicine only. Among patients with the nephrotic syndrome, a normal serum creatinine and no echocardiographic evidence of heart amyloidosis are associated with a higher response rate (39 per cent) to chemotherapy, as defined by a 50 per cent reduction in proteinuria without an increase in serum creatinine.

The results of chemotherapy in amyloidosis are difficult to document because there is no easy way to measure the amount of amyloid present. Resolution of the nephrotic syndrome does not necessarily reflect the disappearance of amyloid deposits, and the progressive deposition of amyloid can occur in the presence of improved clinical and laboratory findings. Scintigraphy after the injection of ^{123}I-labelled SAP may be helpful for monitoring the extent of systemic amyloidosis, but it is available in a rather limited number of centers (see Chapter 11.12.4). The effects of chemotherapy are better evaluated by the level of circulating monoclonal immunoglobulin and by the daily urinary excretion of Ig light chain.

Amyloid nephropathy requires symptomic management of the nephrotic syndrome and of renal failure. Patients in endstage renal disease are candidates for regular dialysis and/or kidney transplantation. Their prognosis is compromised by the risks of extension of extrarenal deposition, especially to the heart, and by recurrence of amyloidosis in the graft, although the latter is uncommon.

Based on promising results in myeloma patients, high-dose melphalan treatment with autologous bone marrow or blood stem cell transplantation is being attempted in light-chain amyloidosis with interesting results. It should be applied in appropriately designed trials in centres with special expertise.

Renal involvement in myeloma

Definition and epidemiology

Renal failure is one of the major complications of myeloma, found at presentation in 20 per cent of patients and occurring in 50 per cent of patients during the course of the disease. It is mostly due to cast nephropathy, although other forms of renal disease can occur as well, including light-chain amyloidosis (10 per cent of myeloma patients), light-chain deposition disease (5 per cent), fanconi syndrome, infiltration of renal interstitium by plasma cells, calcium precipitation, and renal infection. Myeloma cast nephropathy is due both to alterations in tubule cells induced by massive reabsorption of light chains in proximal tubule cells and to cast formation involving light chains and Tamm–Horsfall protein in the distal tubule. The risk of developing renal failure is twice as high in patients with pure light-chain myeloma, and five to six times greater in patients with light-chain proteinuria of more than 2.0 g/day compared with those with proteinuria of less than 0.05 g/day.

Clinical presentation

Myeloma cast nephropathy usually presents as acute or subacute renal failure, often revealing myeloma with a high tumour burden (found in 70 to 80 per cent of myeloma patients with renal failure). Common triggering factors include hypercalcaemia, dehydration, infection, use of toxic compounds including radiocontrast media, non-steroidal anti-inflammatory drugs, and angiotensin converting enzyme inhibitors, all of which reduce renal perfusion, especially in those who are dehydrated.

Renal failure induced by cast nephropathy is remarkably silent. The clinical and urinary syndrome is characterized by non-specific signs including weakness, weight loss, bone pain, and signs of infection, all due to myeloma, and by urinary excretion of a monoclonal light chain. It must be emphasized that urinary dipsticks do not detect the light chain, which is measured by quantitative tests of proteinuria. Light chain accounts for more than 70 per cent of total proteinuria by urine electrophoresis.

Tubular dysfunction is rarely a presenting symptom. fanconi syndrome due to proximal tubule impairment may result from intratubular crystalline inclusions of _ light chains. This can lead to osteomalacia and may precede the diagnosis of myeloma by several years.

Diagnosis

Diagnosis of myeloma cast nephropathy relies on the detection of a urinary monoclonal light chain in patients with subacute or acute renal failure of apparently unknown origin. In those patients with pure light-chain myeloma, diagnosis can be suspected before urinalysis on the basis of dramatic hypogammaglobulinaemia detected by serum electrophoresis.

A renal biopsy should not be performed routinely in patients with a presumed diagnosis of myeloma cast nephropathy. It can, however, be useful for a number of reasons: firstly, to analyse tubulointerstitial lesions and allow diagnosis and treatment of other potential causes of renal impairment in those with multiple possible precipitating factors (infection, drugs, etc.). Secondly, to establish the diagnosis of Fanconi syndrome. Thirdly, to identify glomerular lesions in patients with albuminuria over 1 g/day and no evidence of amyloid deposits in 'peripheral' biopsies. Myeloma casts have unique characteristics, including a 'fractured' appearance due to crystal formation, polychromatism upon staining with Masson's trichrome, and the presence of multinucleated giant cells. They are consistantly associated with

dramatic epithelial tubular lesions.

Treatment

The first aim of treatment is to prevent or retard renal impairment in all patients with myeloma, most particularly those with light-chain myeloma, by prevention of dehydration, maintenance of a high urinary output and urine alkalinization, avoidance of nephrotoxic drugs, and control of hypercalcaemia (if present). hypercalcaemia requires correction of salt and water deficit, steroids, and/or bisphosphonates, which are potent inhibitors of osteoclast activity. Renal failure of recent onset should be promptly and vigorously managed. Adequate administration of salt and water and forced alkaline diuresis (which may help to prevent intratubular light-chain precipitation) are required when urine output persists. Plasma exchange has been advocated to remove light chains more rapidly, but its value is unproven. In patients with oliguria, dialysis should be provided early.

Most patients with overt myeloma cast nephropathy should be given chemotherapy to reduce the production of monoclonal light chains, which is justified because partial or complete recovery of renal failure occurs in about half of patients. Only patients with refractory haematological disease should be given purely symptomatic treatment. However, median survival in those with progressive renal failure (about 2 years) remains shorter than that of patients without renal failure (3 to 4 years). Two main options should be discussed. Firstly, conventional chemotherapy regimens can induce remissions, but they have not markedly lengthened median survival. The melphalan–prednisone regimen remains the first choice for chemotherapy, but its drawbacks are slow antitumour action and the necessity of reducing melphalan doses because the drug is renally eliminated. Regimens including vincristine and doxorubicin induce earlier remission and are safer in those with renal failure since the drugs are metabolized in the liver. Secondly, in younger patients (those aged less than 65), high-dose chemotherapy with the support of autologous bone marrow or blood stem cell transplantation should systematically be considered because substantially longer survival can be achieved.

In patients with irreversible renal failure and in those whose renal function deteriorates later, regular dialysis may be indicated if the clinical condition and bone lesions allow it. Recombinant human erythropoietin may be helpful to correct anaemia, although very high doses (and therefore great expense) are likely to be needed.

Light-chain and heavy-chain deposition disease

Definition and epidemiology

It has been known since the late 1950s that non-amyloidotic forms of glomerular disease resembling the lesion of diabetic glomerulosclerosis could occur in multiple myeloma. Randall and associates recognized the presence of monoclonal light chains in these lesions in 1976, defining light-chain deposition disease. Monoclonal heavy chains can also be found in association with light chains (defining light- and heavy-chain deposition disease) or occasionally in the absence of light chains (heavy-chain deposition disease). In clinical and pathological terms light-chain deposition disease, light- and heavy-chain deposition disease, and heavy-chain deposition disease are similar and hence are also collectively referred to as monoclonal immunoglobulin deposition disease. They differ from amyloidosis by the lack of affinity for Congo red and fibrillar organization. Monoclonal immunoglobulin deposition disease occurs in a wide range of ages (31 to 79 years) with a slight male preponderance. Myeloma accounts for only 45 per cent of cases, but as in amyloidosis a monoclonal plasma cell proliferation can be found in virtually all patients by immunofluorescence examination of the bone marrow with specific antiheavy- and antilight-chain antisera.

Clinical presentation

Light-chain deposition disease is a systemic disease with deposition of Ig light chains along basement membranes in most tissues. However, deposition in tissues other than the kidney is often (but not always) totally asymptomatic and renal involvement dominates clinical presentation, mainly in the form of proteinuria and renal failure. In 23 to 67 per cent of patients with light-chain deposition disease, albuminuria is associated with the nephrotic syndrome. In 25 per cent, the urinary albumin output is less than 1 g/day, and these patients mainly exhibit a tubulointerstitial syndrome. Haematuria is more frequent (44 per cent) than one would expect for a nephropathy in which cell proliferation is usually modest. Renal failure is remarkable for its high prevalence (89 per cent), early appearance, and severity, irrespective of urinary albumin output. Hypertension occurs in about half of the patients.

Diagnosis

Diagnosis of monoclonal immunoglobulin deposition disease relies on the association of the clinical features described above with the finding of a monoclonal Ig component in the serum and/or the urine. Since this component cannot be detected even by immunofixation in 15 to 30 per cent of patients, the diagnosis of monoclonal immunoglobulin deposition disease is often made by renal biopsy. In virtually all patients with this condition tubular lesions are characterized by the deposition of a periodic acid-schiff positive ribbon-like material along the basement membrane. This is usually associated with a marked interstitial fibrosis and nodular glomerulosclerosis (found in two-thirds of patients with light-chain deposition disease and in all patients with heavy-chain deposition disease reported so far). Nodules are composed of membrane-like material with nuclei at the periphery (Fig. 2 and Plate 2). A key step in the diagnosis of the various forms of monoclonal immunoglobulin deposition disease is immunofluorescence examination of the biopsy specimen, revealing evidence of monotypic light- and/or heavy-chain deposits along glomerular and tubular basement membranes in all cases. By contrast with light-chain amyloidosis, the _ isotype is two to three times more frequent than the l one. In those patients with heavy-chain deposition disease, a deletion of the first constant domain of the heavy chain can invariably be demonstrated by immunofluorescence analysis of the kidney specimen with specific antisera. Finally, non-fibrillar granular electron-dense deposits are visible by electron microscopy along tubular basement membranes and in glomerular lesions.

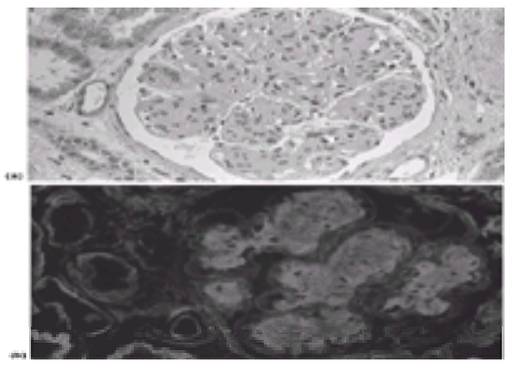

Fig. 2 Monoclonal immunoglobulin deposition disease. (a) Typical nodular glomerulosclerosis. Note the membrane-like material in the centre of the nodules and nuclei at the periphery. Some glomerular capillaries show double contours. Note also thickening of the basement membrane of atrophic tubules (Masson's trichrome stain, ×312). (b) Bright staining of tubular basement membranes and mesangial nodules and, to a lesser extent, of glomerular basement membrane with anti-_ antibody in a case of _ light-chain deposition disease (immunofluorescence, ×312). (See also Plate 2.)

Treatment

The natural history of monoclonal immunoglobulin deposition disease is more uncertain than that of light-chain amyloidosis because extrarenal deposits can be totally asymptomatic or cause severe organ damage, including severe heart failure and occasionally hepatic insufficiency and portal hypertension. The 5-year actuarial rates for patient survival and survival free of endstage renal failure (with chemotherapy) are 70 and 39 per cent respectively.

Patients with monoclonal immunoglobulin deposition disease and myeloma should be treated with conventional chemotherapy if they are over 60 years of age, but intensive therapy with blood stem cell autografting should be discussed in younger patients (see above). As in light-chain amyloidosis, deposited light chains have disappeared in isolated instances after intensive therapy. The correct treatment for those without myeloma is uncertain, the rarity of the disease meaning that there are no controlled trials. A pragmatic approach is to use alkylating agents plus prednisolone in those with moderate but rapidly progressive renal insufficiency in an endeavour to prevent progression to endstage renal failure, but not to treat those with severe renal failure unless there are significant extrarenal complications.

Recurrence of the disease has usually been observed in the few patients who have received renal transplants.

Non-amyloid fibrillary and immunotactoid glomerulopathies

Definition and epidemiology

Fibrillary glomerulonephritis and immunotactoid glomerulopathy are recently described entities characterized, respectively, by fibrillar and microtubular deposits in the mesangium and the glomerular capillary loops. These deposits do not have a b-pleated sheet organization and are readily distinguishable from amyloid by the larger thickness of fibrils and the lack of Congo red staining. Whether they are totally distinct entities remains the subject of considerable debate. However, the distinction between the two diseases may be of great clinical and pathophysiological interest in the context of plasma cell dyscrasias because monotypic deposits are detected in 50 to 80 per cent of immunotactoid glomerulopathies, whilst they are found in fewer than 20 per cent of fibrillary glomerulopathies.

The prevalence of glomerulopathy with non-amyloid deposition of fibrillary or tubular material in a non-transplant adult biopsy population is around 1 per cent, but this is almost certainly an underestimate because insufficient attention is given to atypical reactions with histochemical stains for amyloid and the fact that most specimens are not examined by electron microscopy. The age range extends from 10 to 80 years with a peak incidence between 40 and 60 years.

Clinical presentation

The usual presentation is with the nephrotic syndrome and microscopic haematuria, often in the setting of chronic lymphocytic leukaemia or lymphoma.

Diagnosis

Diagnosis relies entirely on analysis of the renal biopsy specimen by immunofluorescence microscopy with antilight-chain antibodies and by electron microscopy. In immunotactoid glomerulopathy this reveals either membranous glomerulonephritis (often associated with segmental mesangial proliferation) or lobular membranoproliferative glomerulonephritis. Immunofluorescence shows coarse granular deposits of IgG and IgC3 along capillary basement membranes and in mesangial areas. Although monotypic deposits are common, a circulating monoclonal Ig is detected in only a minority of patients. Electron microscopy shows immunotactoid glomerulopathy to be remarkable for the presence of organized deposits of large, thick-walled microtubules, usually greater than 30 nm in diameter, at times arranged in parallel arrays. When immunotactoid glomerulopathy occurs in the setting of chronic lymphocytic leukaemia or related B-cell lymphoma, inclusions showing the same microtubular organization and containing the same IgG subclass and light-chain type as the renal deposits are then often detected in the cytoplasm of leukaemic lymphocytes in the blood (Fig. 4).

Mesangial proliferation and membranoproliferative glomerulonephritis are the commonest lesions observed in fibrillary glomerulonephritis. Immunofluorescence studies mainly show polyclonal IgG deposits (of the g4 isotype in one series). Electron microscopy shows the fibrils to be randomly arranged with a diameter varying between 12 and 22 nm.

Infection wuth hepatitis c virus has recently been reported in patients with non-amyloid fibrillary glomerulonephritis and immunotactoid glomerulopathy.

Treatment

In patients with immunotactoid glomerulopathy and monotypic immunoglobulin deposits, especially in those with chronic lymphocytic leukaemia, corticosteroids and/or chemotherapy are associated with partial or complete remission of the nephrotic syndrome, parallel with improvement of the haemopathy. More variable results are obtained with these treatments in patients with crescentic fibrillary glomerulonephritis. Recurrence of these diseases has been reported in patients receiving a renal allograft.

Renal involvement in cryoglobulinaemia

Definition and epidemiology

Cryoglobulinaemia is a pathological condition in which the blood contains immunoglobulins that precipitate on cooling (4 °C) and resolubilize on warming (37 °C). according to Brouet's classification, there are three types of cryoglobulinaemia defined by their composition. Renal involvement is observed mainly in patients with mixed type II cryoglobulinaemia involving a monoclonal IgM (most often including a _ light chain) with rheumatoid factor activity and a polyclonal IgG. Type II cryoglobulinaemia can be associated with overt lymphoproliferative disorders of the B-cell lineage, although in many cases no underlying haemopathy is found. Therefore, this type of cryoglobulinaemia has long been referred to as essential mixed cryoglobulinaemia (Fig. 3 and Plate 3).

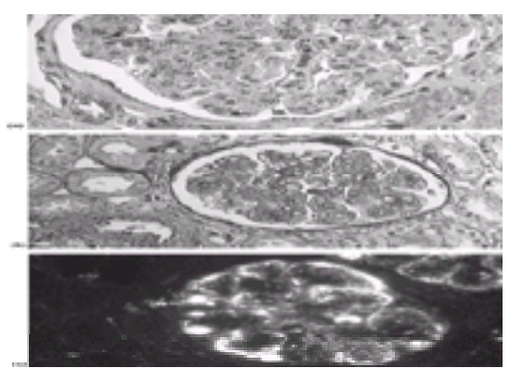

Fig. 3 Cryoglobulinaemic glomerulonephritis. (a) The glomerulus shows a marked endocapillary hypercellularity with massive infiltration of mononuclear leucocytes (Masson's trichrome stain, ×500). (b) Frequent double-contour aspect, and intraluminal thrombi (periodic acid-Schiff stain, ×312). (c) thrombi and segments of glomerular basement membrane are brightly stained with anti-IgM antibody (immunofluorescence, ×312). (From Béatrice Mougenot's personal collection.) (See also Plate 3.)

Viral infections may trigger the formation of cryoglobulin. Whereas hepatitis B and Epstein–Barr virus infections have been implicated in the past, the role of hepatitis C virus infection is now recognized to be an important factor in the pathogenesis of type II cryoglobulinaemia. Antibodies to hepatitis C virus and hepatitis C virus RNA are frequently found in the sera of patients with type II cryoglobulinaemia, probably explaining the uneven geographical distribution of mixed cryoglobulinaemias, which predominate in southern Europe where hepatitis C infection is more prevalent.

The condition is commonest in adults in the fifth and the sixth decades of life, with a slight female predominance.

Clinical presentation

Renal disease most often occurs in patients with a long history of cryoglobulinaemia-related vasculitis symptoms including palpable purpura (70 per cent), arthralgias (50 per cent), fatigue, Raynaud's phenomenon, peripheral neuropathy (22 per cent), and hepatic involvement.

The renal disease may present as an acute nephritic syndrome (in 20 to 30 per cent of cases) with gross haematuria, heavy proteinuria, hypertension, and renal failure of sudden onset, sometimes with oliguria (5 per cent of cases). The pathological finding in these patients is membranoproliferative glomerulonephritis with the presence of numerous intraluminal thrombi and/or necrotic vasculitic lesions. Remission may occur spontaneously or during therapy, with relapses following in up to

20 per cent of cases.

Most patients with mixed cryoglobulinaemia (55 per cent) have an indolent and protracted renal course, presenting with proteinuria, haematuria, and hypertension. The usual renal lesion in this context is membranoproliferative glomerulonephritis, with some of the peculiarities described above.

Nephrotic syndrome affects another 20 per cent of patients. Arterial hypertension is observed in more than 80 per cent of patients at the time of onset of renal disease. Endstage uraemia develops in fewer than 10 per cent of patients.

Diagnosis

Mixed type II cryoglobulinaemia should be suspected in patients with the clinical picture described above, an IgM rheumatoid factor, and a very low serum C4 fraction and total haemolytic activity of complement. In this context a careful search for the presence of cryoglobulin must be made, requiring that a blood sample from a fasting patient should be placed in warm water and taken promptly to the laboratory, which needs to be forewarned that such a sample will arrive.

Cryoglobulinaemia-related membranoproliferative glomerulonephritis usually shows several distinctive histological features, including massive subendothelial deposits filling the capillary lumen and forming so-called thrombi, and dramatic infiltration by leucocytes, mainly monocytes. The thrombi are brightly stained with anti-µ and anti-_ antibodies and present a microtubular crystalline structure similar to that of the cryoprecipitate. These glomerular changes may be associated with acute vasculitis of the small and medium-sized arteries (33 per cent) and lymphocytic infiltrates in interstitium. Crescentic extracapillary proliferation is rare and always limited.

Treatment

The best treatment is not firmly established because the course of the disease is unpredictable and acute exacerbations may remit spontaneously. High-dose steroids, cyclophosphamide, and plasma exchanges are used in the more severe cases, particularly those with signs of systemic vasculitis. The place of antihepatitis C virus therapy including interferon-a and vidarabine has still to be evaluated. These antiviral drugs may be useful for controlling virus replication enhanced by steroids and/or immunosuppressive agents. Hypertension needs to be carefully controlled because cardiovascular complications are the major causes of death.

Renal involvement in Waldenström's macroglobulinaemia

A glomerulonephritis with intracapillary thrombi of IgM is rare, but is almost specific for Waldenström's macroglobulinaemia. It is characterized by periodic acid-Schiff positive, non-congophilic endomembranous deposits in a variable number of capillary loops, which are sometimes so large as to occlude the capillary lumen either partially or completely, thus forming thrombi. There may also be a B-cell interstitial infiltrate. Renal presentation is with proteinuria or renal impairment. Some patients have a cryoglobulin, in others the amount of circulating IgM seems to be higher than that in patients with Waldenström's without obvious renal involvement, or with amyloidosis, leading to the suggestion that hyperviscosity is important in pathogenesis of the renal lesion. The haemopathy is treated on its own merits (see Section 22). In those with acute renal failure there is anecdotal experience that plasma exchange can be effective in restoring renal function at least temporarily, allowing time for other treatments to be applied.

Renal involvement in lymphomas and leukaemias

Renal complications of lymphomas and leukaemias are summarized in Table 2. Patients with unexplained renal failure should undergo ultrasound examination of the kidney, which should be arranged as a matter of urgency, to identify either enlarged kidneys due to tumour infiltration or hydronephrosis. The presence of heavy albuminuria in this setting is suggestive of paraneoplastic glomerulopathy.

Hodgkin's disease and non-Hodgkin's lymphoma

Glomerulonephritis is a rare complication of lymphoma, most often described in patients with Hodgkin's disease, of whom 0.4 per cent have minimal change disease and 0.1 per cent have AA amyloidosis. This low incidence of amyloidosis in patients with Hodgkin's disease is most likely attributable to modern treatment protocols that induce a rapid remission of the haemopathy. Hodgkin's lymphoma-related minimal change disease shows features of a paraneoplastic glomerulopathy: the nephrotic syndrome usually appears early, revealing the haemopathy in about one-half of the cases; it rapidly disappears after effective treatment of Hodgkin's disease and it usually relapses simultaneously with the haemopathy. Cases of crescentic glomerulonephritis with rapidly progressive renal failure due to antiglomerular basement antibodies have also been reported.

There are about 50 reports of glomerulonephritis in patients with non-Hodgkin's lymphoma, including both T- and B-cell proliferations. In these conditions, unlike in Hodgkin's lymphoma, minimal change disease is uncommon, and membranoproliferative glomerulonephritis and necrotizing crescentic glomerulonephritis with or without vasculitis are the most frequent lesions. Some cases are associated with type II cryoglobulinaemia or immunotactoid glomerulopathies with monotypic immune deposits. In other cases the association between non-Hodgkin's lymphoma and renal involvement may be coincidental. Presenting renal symptoms are nephrotic syndrome and/or renal impairment. Full remission of these symptoms can be achieved in some patients by aggressive therapy of the lymphoma.

Chronic lymphocytic leukaemia and low-grade B-cell lymphoma

These haemopathies, particularly chronic lymphocytic leukaemia, have been reported in association with glomerular disease in about 50 cases. Most commonly the nephropathy, usually manifesting as nephrotic syndrome with impaired renal function, and the leukaemia are detected simultaneously. The most frequent glomerular disease is membranoproliferative glomerulonephritis with or without cryoglobulinaemia. In the absence of cryoglobulinaemia, a molecular link can be established between the haemopathy and the glomerulopathy when monotypic Ig deposits are found in the glomerulus, which can occur even in the absence of detectable circulating M component. Some of these patients present with typical immunotactoid glomerulopathy or monoclonal immunoglobulin deposition disease (Fig. 4 and Plate 4). Improvement of the nephropathy after treatment for the leukaemia is well described.

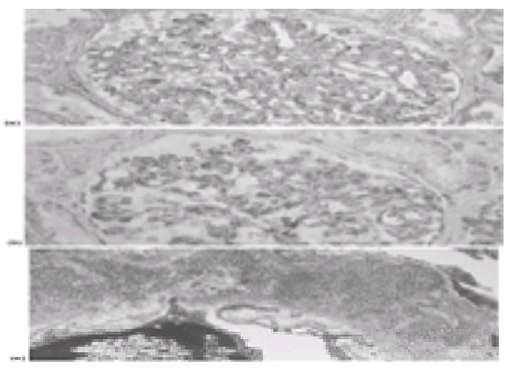

Fig. 4 Immunotactoid glomerulopathy in a patient with chronic lymphocytic leukaemia. Atypical membranous glomerulonephritis showing exclusive staining of the deposits with anti-g (a) and anti-_ (b) antibodies (immunohistochemistry, alkaline phosphatase, ×312). (c) Electron micrograph of glomerular basement membrane, showing the microtubular structure of the subepithelial deposits (uranyl acetate and lead citrate, ×12 000). (From Béatrice Mougenot's personal collection.) (See also Plate 4.)

Acute leukaemias

Disseminated intravascular coagulation has been associated with acute progranulocytic leukaemia. Other renal complications are commonly due to treatment, most

particularly the tumour lysis syndrome (see below).

Tumour lysis syndrome

Tumour lysis syndrome is a life-threatening metabolic emergency. It occurs in patients with haemopathies involving a high cell turnover, mostly at the onset of chemotherapy and/or upon radiation therapy. The ensuing massive cytolysis generates high levels of uric acid, phosphate, potassium, and xanthine (especially in patients treated with allopurinol), with a concomitant decrease in serum calcium concentration. Oliguric or anuric acute renal failure may occur, especially in those who are dehydrated or have pre-existing impairment of kidney function. This acute renal failure is mostly the consequence of acute precipitation of urate crystals in the tubular lumen, but in those with a moderate increase in uric acid concentration, the role of severe hyperphosphataemia causing precipitation of calcium/phosphate complexes in renal interstitium and the tubular system has been assumed.

Prevention is better than cure and intensive monitoring is mandatory to prevent the development and the consequences of this syndrome. Patients at risk of the tumour lysis syndrome should be vigorously hydrated with 0.9 per cent saline (assuming normal or near normal baseline renal function, and with care taken to avoid inducing pulmonary oedema) before receiving chemotherapy or radiotherapy. Some physicians would also pretreat with allopurinol, but this does carry the risk of formation of xanthine nephropathy/stones due to accumulation of xanthine and cannot be generally recommended. Urine alkalinization should be avoided because it increases the risk of phosphate precipitation. In those who develop tumour lysis syndrome but are passing urine, vigorous hydration with 0.9 per cent saline to encourage urine output is required, with close clinical monitoring to prevent iatrogenic fluid overload should the urine output drop. Administration of urate oxidase has been advocated as a treatment for acute hyperuricaemia, but experience is very limited and there is no therapeutic formulation that is widely available. Patients with severe acute renal failure should be treated with haemodialysis, which allows recovery of renal function following the reduction of serum phosphate and serum uric acid concentrations.

Further reading

Renal involvement in Ig light-chain amyloidosis

Comenzo RL *et al.* (1998). Dose-intensive melphalan with blood stem-cell support for the treatment of AL (amyloid light-chain) amyloidosis: survival and responses in 25 patients. *Blood* **91**, 3662–70.

Kyle RA, Gertz MA (1995). Primary systemic amyloidosis: clinical and laboratory features in 474 cases. *Seminars in Hematology* **32**, 45–59.

Kyle RA *et al.* (1997). A trial of three regimens for primary amyloidosis: colchicine alone, melphalan and prednisone, and melphalan, prednisone, and colchicine. *New England Journal of Medicine* **336**, 1202–7.

Ronco PM, Aucouturier P, Moulin B (1999). Renal amyloidosis and plasma cell dyscrasia-related glomerulopathies. In: Feehally J and Johnson R, eds *Comprehensive nephrology*, section 5, chapter 31, pp 1–14. Mosby, London.

Renal involvement in myeloma

Ronco PM, Aucouturier P, Mougenot B (1997). The kidney in plasma cell dyscrasias. In: Davison AM *et al.*, eds. *Oxford textbook of clinical nephrology*, 2nd edn, vol. 2, pp 811–35. Oxford University Press, Oxford.

Winearls, C.G. (1995). Acute myeloma kidney. *Kidney International* **48**: 1347–61.

Light-chain and heavy-chain deposition disease

Heilman RL *et al.* (1992). Long-term follow-up and response to chemotherapy in patients with light-chain deposition disease. *American Journal of Kidney Diseases* **20**, 34–41.

Moulin B *et al.* (1999). Nodular glomerulosclerosis with deposition of monoclonal immunoglobulin heavy chains lacking C_H1. *Journal of the American Society of Nephrology* **10**, 519–28.

Ronco PM, Aucouturier P, Moulin B (1999). Renal amyloidosis and plasma cell dyscrasia-related glomerulopathies. In: Feehally J and Johnson R, eds *Comprehensive nephrology*, section 5, chapter 31, pp 1–14. Mosby, London.

Non-amyloid fibrillary and immunotactoid glomerulopathies

Brady HR (1998). Fibrillary glomerulopathy. *Kidney International* **53**, 1421–9.

Fogo A, Qureshi N, Horn RG (1993). Morphologic and clinical features of fibrillary glomerulonephritis versus immunotactoid glomerulopathy. *American Journal of Kidney Diseases* **22**, 367–77.

Markowitz GS *et al.* (1998). Hepatitis C viral infection is associated with fibrillary glomerulonephritis and immunotactoid glomerulopathy. *Journal of the American Society of Nephrology* **9**, 2244–52.

Touchard G *et al.* (1994). Glomerulonephritis with organized microtubular monoclonal immunoglobulin deposits. *Advances in Nephrology from the Necker Hospital* **23**, 149–75.

Renal involvement in cryoglobulinaemia

Brouet JC *et al.* (1974). Biologic and clinical significance of cryoglobulins. A report of 86 cases. *American Journal of Medicine* **57**, 775–88.

D'Amico G (1998). Renal involvement in hepatitis C infection: cryoglobulinemic glomerulonephritis. *Kidney International* **54**, 650–71.

Johnson RJ *et al*. (1993). Membranoproliferative glomerulonephritis associated with hepatitis C virus infection. *New England Journal of Medicine* **328**, 465–70.

Renal involvement in Waldenström's macroglobulinaemia

Veltman GA *et al.* (1997). Renal disease in Waldenstrom's macroglobulinaemia. *Nephrology, Dialysis and Transplantation* **12**, 1256–9.

Renal involvement in lymphomas and leukaemias

Moulin B *et al.* (1992). Glomerulonephritis in chronic lymphocytic leukemia and related B-cell lymphomas. *Kidney International* **42**, 127–35.

Ronco PM (1999). Paraneoplastic glomerulopathies: new insights into an old entity. *Kidney International* **56**, 355–77.

Tumour lysis syndrome

Haas M *et al.* (1999). The spectrum of acute renal failure in tumour lysis syndrome. *Nephrology, Dialysis and Transplantation* **14**, 776–9.

Wolf G *et al.* (1999). Hyperuricemia and renal insufficiency associated with malignant disease: urate oxidase as an efficient therapy? *American Journal of Kidney Diseases* **34**, E20.

20.10.6 Haemolytic uraemic syndrome

Paul Warwicker and Timothy H. J. Goodship

Introduction

The haemolytic uraemic syndrome (**HUS**) is characterized by the triad of microangiopathic haemolytic anaemia (Coombs' test negative), thrombocytopenia, and acute renal failure. In thrombotic thrombocytopenic purpura (**TTP**) neurological manifestations and fever occur in addition.

Classification (Table 1)

HUS can be divided into diarrhoeal (D+)- and non-diarrhoeal-associated (D–) disease. Historically, some disorders that present with the characteristic triad of features have been considered separately from HUS. This is illogical, and they too are listed as forms of D– HUS. Because the clinical presentation, prognosis, and treatment differ for each form of HUS they are described separately in this chapter.

Pathogenesis of HUS

The anticoagulant state of normal endothelium is maintained by: (1) a lack of the constitutive expression of tissue factor; (2) the endothelial expression of heparin, tissue-plasminogen activator, and thrombomodulin; and (3) the local secretion of vasoactive substances preventing platelet aggregation, including nitric oxide, prostacyclin, and adenosine.

In HUS, several factors can result in endothelial cell activation, these include: antiendothelial antibodies, immune complexes, lipopolysaccharides, toxins, complement, and drugs. Tissue factor and tissue-plasminogen activator inhibitor are expressed. In addition, von Willebrand factor (**vWF**) is synthesized and secreted in increased amounts, promoting platelet aggregation by binding to the $a_{iiib}b_3$ integrins on the platelet surface and to the endothelial matrix. Downregulation of this process is usually achieved by endothelial secretion of a metalloproteinase that cleaves vWF, rendering it inactive.

The effect of endothelial activation, a crucial feature of all the conditions listed in Table 1 as causing HUS, is that five core changes occur: loss of vascular integrity, expression of leucocyte-adhesion molecules, cytokine production, upregulation of HLA molecules, and the change in phenotype from an anticoagulant to a procoagulant state. It is the latter that predisposes to the development of a thrombotic microangiopathy.

There have been many descriptions of phenomena associated with HUS, such as complement activation and the secretion of ultra-large vWF, selectins, and tissue plasminogen activator-I. Causal relationships have been proposed. In retrospect many of these observations merely reflect a state of endothelial activation. However, if the mechanisms responsible for downregulating the sequelae of endothelial activation are impaired then it is possible that a procoagulant state will be maintained. This has recently been shown for both HUS (overactivity of the alternative complement pathway) and TTP (abnormalities of the metalloproteinase that cleaves vWF).

Histopathology (Fig. 1 and Plate 1)

In D+ HUS it is predominantly the glomerular endothelium that is affected (hence termed 'thrombotic glomerulopathy'), whilst in D– HUS it is the endothelium of the preglomerular vessels (hence termed 'arterial thrombotic microangiopathy'). Severe intimal proliferation and luminal stenosis also affects arterioles and interlobular arteries in some forms of D– HUS. The subendothelial spaces are widened and these may contain fibrin deposits.

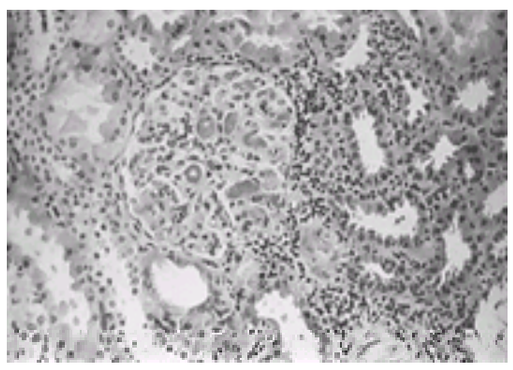

Fig. 1 Typical changes of glomerular thrombotic microangiopathy in a patient with HUS (figure kindly provided by Dr Marie O'Donnell). (See also Plate 1.)

Diagnosis and laboratory features

The diagnosis of HUS is based on demonstrating the aforementioned triad of microangiopathic haemolytic anaemia, thrombocytopenia, and acute renal failure. It should be considered and excluded in all cases of acute renal failure, especially those associated with diarrhoea and/or severe hypertension. Urine output may be reduced, although non-oliguric renal failure is also seen. Urinalysis usually shows microscopic haematuria and proteinuria.

The anaemia of HUS may be severe, with features of haemolysis, including reticulocytosis and increased unconjugated bilirubin, decreased haptoglobin, and raised lactate dehydrogenase (**LDH**) levels, which can be used as a marker of disease. The Coombs' test is negative, and careful examination of the blood film may reveal fragmented and deformed red cells. Thrombocytopenia ranges from severe to mild: 50 per cent of patients will have platelet counts in excess of 100×10^9/l. There may be an associated leucocytosis, the extent of which is correlated with disease severity in D+ HUS. Fibrinogen degradation products may be increased, but clotting tests are characteristically normal. If they are deranged, then septicaemia and disseminated intravascular coagulation should be considered.

Hyponatraemia may complicate D+ HUS, which can then be associated with seizures. Hyperkalaemia is occasionally severe. Complement levels should be measured: C3 is often low in both D+ HUS, where it is associated with a poor prognosis, and D– HUS, where it is associated with recurring or familial types of disease. Human immunodeficiency virus (**HIV**) infection should also be considered in appropriate risk groups.

Diarrhoeal (D+) HUS

Incidence

This is the commonest form of HUS, accounting for over 90 per cent of cases in industrialized countries. It is the commonest cause of renal failure in children and is associated with a diarrhoeal prodrome, hence D+ HUS. The bloody diarrhoea is caused by bacterial infection, predominantly with strains of enterohaemorrhagic *Escherichia coli*, notably *E. coli* O157:H7. This strain has only recently become a significant health hazard—the first descriptions of haemorrhagic colitis and HUS associated with it appeared in 1983. Since then there has been a rapid increase in the number of cases reported. One of the largest and most serious outbreaks was the Central Scotland outbreak of 1996 where 496 individuals were infected, 27 developed HUS, and 18 died. Other diarrhoeal pathogens, particularly *Shigella dysenteriae*-type 1, can also cause D+ HUS.

Clinical features

E. coli is very virulent with as few as 50 organisms causing disease. The bacteria adhere to the large bowel and release a toxin, known as verocytotoxin or Shiga toxin, into the bloodstream. This toxin belongs to the **RIP** (ribosomal inhibitory protein) group of proteins which are amongst the most potent toxins known to man. They include ricin, which gained notoriety in 1978 when an iridium pellet containing trace amounts of the toxin was injected with the aid of an umbrella into the calf of the dissident Bulgarian journalist, Georgi Markov, on Waterloo bridge in London. He subsequently died from multiorgan failure. These toxins consist of five b-subunits and a single a-subunit. The b-subunit binds to Gb3 (a glycolipid found in the membranes of eukaryotic cells), following which the toxin enters the cell by endocytosis and the a-subunit blocks protein synthesis at the ribosome.

The delay between exposure and illness is on average 3 days, typically starting with abdominal cramps and diarrhoea, which becomes bloody over the following 2 days. The majority of patients then recover, although late sequelae such as colonic strictures and chronic pancreatitis are occasionally seen. Between 3 and 20 per cent go on to develop HUS of varying severity. Approximately 50 per cent of patients with D+ HUS require renal replacement therapy, 5 per cent are left with chronic renal failure, and 3 to 5 per cent die of the acute illness. In those who do recover, between 15 and 40 per cent show evidence of persistent renal damage with proteinuria and/or hypertension. Acute neurological complications such as cerebrovascular accident, seizures, and coma develop in approximately 25 per cent of patients with D+ HUS.

Diagnosis

E. coli O157:H7 infection is diagnosed by stool culture and subsequent detection of the O157:H7 antigen. It is also possible to detect antibodies to the O157 lipopolysaccharide in sera from convalescent patients, although this is of limited use in acute illness and is not widely available.

Treatment

Identification of infection and early diagnosis of HUS are important. However, it is unclear whether antibiotics in the early stages of the disease are of benefit, and it is possible that they may increase the risk of developing HUS. Antimotility agents are contraindicated in the early stages as they too may increase the risk of HUS. Treatment of D+ HUS is predominantly supportive, including careful fluid and electrolyte balance, control of hypertension, nutritional support, and renal replacement therapy if necessary. Vigilance should be maintained for complications such as ischaemic colitis, myocardial dysfunction, and pancreatitis. There is limited evidence for a benefit of either plasma exchange or plasma infusion, but plasma exchange continues to be used, particularly in complicated and prolonged cases. Specific treatments such as toxin-binding resins and toxoid vaccines are currently being investigated. Prevention and good public health policy is likely to be of the utmost importance in the future. The recent outbreak in Scotland led to the commissioning of the *Pennington report*, which produced recommendations for disease prevention ranging from the 'farm to the table'.

Non-diarrhoeal (D–) HUS

Different forms of D– HUS

Idiopathic

Idiopathic, also known as sporadic or atypical, HUS accounts for approximately 5 to 10 per cent of cases. It is typically insidious, although it may present following an upper respiratory tract infection. People of all ages can be affected, and there is no seasonal incidence. Severe hypertension is frequent and mortality is higher than with D+ HUS. Renal involvement is usually pronounced with significant proteinuria and uraemia. The disease may recur in both native kidneys and allografts.

As the name suggests, the cause of idiopathic HUS is unknown. Recent research has focused on a dysfunction of complement-pathway modulators; case reports often reveal low levels of C3, indicating overactivity of the alternative complement pathway (see familial HUS below). Other possible aetiological factors implicated include dysfunction of prostacyclin metabolism, abnormalities of von Willebrand factor multimers (see the section on TTP), abnormalities of platelet-activating factor, and tissue plasminogen activator inhibitor-type 1.

Familial HUS

HUS may be inherited as both autosomal recessive and autosomal dominant forms. In autosomal recessive disease the onset is early, with a peak incidence in infancy. Affected children may suffer recurrent episodes. Autosomal dominant disease can affect all ages, and may be precipitated by pregnancy. Phenotypically, familial HUS most closely resembles idiopathic HUS. Severe hypertension is a prominent feature, being found in approximately 80 per cent of autosomal dominant and 40 per cent of autosomal recessive cases. There is usually no diarrhoeal prodrome, prognosis is poor, mortality is high (often over 50 per cent), and recurrence is common. Management includes aggressive control of blood pressure, particularly with angiotensin-converting enzyme (**ACE**) inhibitors, careful fluid balance, and plasma exchange (although it is often unsuccessful in reversing the disease).

Although familial HUS is rare, it affords an opportunity to elucidate underlying mechanisms relevant to the acquired form of D– HUS. Genetic linkage of the familial form of the disease has been established to an area of chromosome 1q32 containing the gene for complement factor H. This is a soluble serum protein that downregulates the spontaneous activity of the alternative complement pathway. Factor H deficiency has now been described in several families with HUS. Mutations in the factor H gene have been identified in both familial and sporadic HUS.

Transplantation

Both *de novo* and recurrent HUS are seen following renal transplantation. HUS is a well-established side-effect of both ciclosporin and tacrolimus (FK506). Other risk factors include acute rejection, cytomegalovirus (**CMV**) infection, and simultaneous pancreas transplantation. Treatment includes discontinuation of the drug, the use of intravenous corticosteroids, and the substitution of other immunosuppressive agents.

HUS is also seen with other forms of transplantation, particularly that of bone marrow, where between 6 and 26 per cent of patients show evidence of a microangiopathy. The aetiology is unclear, but may be related to endothelial damage secondary to total body irradiation, intensive conditioning chemotherapy, ciclosporin treatment, CMV infection, or graft-versus-host disease. Management is supportive, including renal replacement therapy, the aggressive treatment of hypertension with ACE inhibitors, and the withdrawal of ciclosporin. Fresh-frozen plasma replacement may be of benefit, but treatment is usually unsuccessful in the more fulminant forms of the disease and prognosis is poor.

Drug-related

Other drugs besides ciclosporin and tacrolimus are associated with HUS. In particular, several cytotoxic drugs used in chemotherapy are complicated by the syndrome (so-called **C-HUS** or cancer-chemotherapy HUS). They include mitomycin C, 5-fluorouracil, and cisplatin either alone or in combination with daunorubicin, vinblastine, and bleomycin. The disease may be associated with severe hypertension, pulmonary oedema (often after transfusion of blood products), neurological features, and a high mortality (60 to 75 per cent in some series). With mitomycin C the disease is dose-related: it is rarely seen in patients receiving less than 30 mg/m^2, but more frequently at doses above 60 mg/m^2. The incidence is between 2 and 15 per cent, and usually presents weeks or months after the last treatment with mitomycin, often when the patient is in clinical remission. Few treatments have proven effective, although staphylococcal protein-A column perfusion to remove

circulating immune complexes seems to be more effective than plasma exchange and may be of benefit in less severe forms of the disease.

Other drugs implicated in HUS include crack cocaine, ticlopidine, and quinine. Use of the oral contraceptive has also been said to be associated with HUS, although with such a commonly prescribed medication it is difficult to be certain whether the association is coincidental. However, it is wise to advise against its use in survivors of D– HUS and in families with the inherited form of HUS.

Pregnancy-related

The incidence of HUS in pregnancy is approximately 1 in 25 000. Presentation is usually peripartum or within a few weeks after delivery, when it is also known as postpartum renal failure. In women presenting in the third trimester, it may be difficult to differentiate HUS from severe forms of pre-eclampsia, such as the **HELLP** syndrome (haemolysis, elevated liver enzymes, and low platelets). Pre-eclamptic syndromes tend to be associated with less severe haemolytic anaemia, the presence of hepatocellular necrosis, and a rapid improvement postdelivery. Features of pregnancy-related HUS include severe hypertension, neurological symptoms, fever, and renal failure requiring renal replacement therapy. Although plasma exchange increases survival rates, maternal mortality remains between 5 and 20 per cent, and preterm delivery and intrauterine fetal death (approximately 30 per cent) are frequent complications. Long-term follow-up is important because of the later development of renal failure and hypertension. HUS will reoccur in approximately 50 per cent of patients, not only during a further pregnancy but also at other times.

Malignancy-related

HUS is associated with malignancy, particularly adenocarcinoma. Patients with a gastric primary and metastatic disease are at increased risk.

HIV-related

There are several forms of nephropathy associated with HIV infection, and a thrombotic microangiopathy with features resembling both idiopathic HUS and TTP is being increasingly recognized. The incidence of HUS in patients infected with HIV is estimated at approximately 1 per cent; although it usually presents in the later stages of the disease, occasionally it can be the first presentation. HIV infection should be considered in the differential diagnosis of HUS and TTP in high-risk groups and in patients originating from a high-prevalence area. The p24 antigen in endothelial cells may reflect either a direct cytopathic effect of the virus or functional impairment of the endothelium. Concurrent CMV infection has also been associated with HUS. Neurological involvement is common and severe hypertension is a prominent feature. Plasma exchange can lead to renal recovery, but overall the prognosis is poor and few patients survive 1 year from diagnosis. Since ACE inhibitors are used in the therapy of other forms of HIV-associated nephropathy, they would seem a logical choice in the treatment of hypertension.

Other infective causes

Non-diarrhoeal bacterial infections are occasionally associated with HUS. *Pneumococcus* and some *Clostridia* species can produce neuraminidase which strips sialic acid from cell membranes, thereby exposing the cryptic Thomsen–Friedenreich antigen on erythrocytes, platelets, and glomerular cells. An IgM antibody, present in most plasma, causes agglutination, endothelial injury, and HUS. Plasma exchange is therefore contraindicated and treatment consists of washed red cells and antibiotics. Difficulty in red cell typing and a blood film demonstrating both agglutination and red cell fragments may give a clue to diagnosis. Capnocytophagia sepsis from dog bites has recently been associated with cases of HUS.

Immunological disorders

HUS has been reported in association with systemic lupus erythematosus, primary antiphospholipid syndrome, and a variety of glomerulonephritides.

'HUS-like' syndromes

As the renal crisis of systemic sclerosis can be clinically and histologically indistinguishable from idiopathic D– HUS, a diagnosis of systemic sclerosis is often made retrospectively from serological markers or with the development of other features of the disease. Likewise, the thrombotic microangiopathy of accelerated hypertension may be difficult to distinguish from HUS. The most important aspect of treatment in these syndromes is good blood pressure control; ACE inhibitors should be prescribed early, balanced with the requirement to reduce blood pressure gradually. In the renal crisis of systemic sclerosis there may be late recovery of renal function, even when the patient has been established on dialysis for some weeks or months.

Treatment of D– HUS

Few randomized controlled trials have been conducted into the treatment of D– HUS.

Supportive

This consists of careful fluid balance, blood transfusion, and renal replacement therapy. In oliguric patients care should be taken to prevent fluid overload, and central venous monitoring may be required. Platelet transfusions should be avoided, unless there is evidence of bleeding. Patients with deteriorating renal function should always be referred to a renal unit.

Plasma treatment

Previous studies of the efficacy of plasma treatment in HUS are difficult to interpret because of the inclusion of patients with TTP who respond well to plasma exchange. Nevertheless plasma exchange is recommended for most forms of D– HUS.

Corticosteroids

The use of corticosteroids is controversial. Although there is evidence of their efficacy in the treatment of TTP, in small retrospective studies of D– HUS they appear to have no significant effect on survival or the need for renal replacement therapy.

Hypertension

Because of the histological changes in HUS, it is not surprising that hypertension, presumably renin driven, is common. ACE inhibitors are the treatment of choice, with clear parallels existing with their use in the renal crisis of systemic sclerosis. Bilateral nephrectomy has been advocated for severe D– HUS with widespread manifestations unresponsive to plasma exchange. In a series of four patients (three of whom suffered ACE inhibitor-resistant hypertension), bilateral nephrectomy led to complete haematological and clinical remission within 2 weeks.

Renal transplantation

Renal transplantation in patients with renal failure secondary to D+ HUS is safe, with little risk of disease recurrence. In contrast, patients with D– HUS have a 30 to 50 per cent chance of recurrence, usually within the first 2 months, and often within the first 2 weeks. Once recurrent HUS is established, most grafts are lost despite treatment with plasma exchange. This is reflected in the poor 2-year overall graft survival rate of 35 per cent. Graft nephrectomy should not be delayed in this situation. Acute vascular rejection may be difficult to differentiate from recurrent HUS, both clinically and histologically.

Further reading

Reviews of subtypes and treatment of HUS

Kaplan BS, Trompeter RS, Moake JL (1992). *Hemolytic uremic syndrome and thrombotic thrombocytopenic purpura*. Dekker, New York.

Neild GH (1994). Haemolytic-uraemic syndrome in practice. *Lancet* **343**, 398–401.

Neild GH, Barratt TM (1998). Acute renal failure associated with microangiopathy. In: Davison AM, *et al.*, eds. *Oxford textbook of clinical nephrology*, pp 1649–66. Oxford University Press, Oxford.

Remuzzi G, Ruggenenti P (1995). The haemolytic uraemic syndrome. *Kidney International* **48**, 2–19.

Pathophysiology of HUS

Moake JL (1994). Haemolytic-uraemic syndrome: basic science. *Lancet* **343**, 393–7.

Rougier N, *et al.* (1998). Human complement factor H deficiency associated with haemolytic uremic syndrome. *Journal of the American Society of Nephrology* **9**, 2318–26.

Warwicker P, *et al.* (1998). Genetic studies into haemolytic uraemic syndrome. *Kidney International* **53**, 836–44.

Comprehensive review of *E. coli* and D+ HUS

Mead PS, Griffin PM (1998). *Escherichia coli* 0157:H7. *Lancet* **352**, 1207–12.

Editorial summarizing recent studies in TTP and HUS

Moake JL (1998). Moschcowitz, multimers, and metalloprotease. *New England Journal of Medicine* **339**, 1629–31.

Pregnancy-associated HUS

Dashe JS, Ramin SM, Cunningham FG (1998). The long term consequences of thrombotic microangiopathy (thrombotic thrombocytopenic purpura and hemolytic uremic syndrome) in pregnancy. *Obstetrics and Gynecology* **91**, 662–8.

Egerman RS, *et al.* (1996). Thrombotic thrombocytopenic purpura and hemolytic uremic syndrome in pregnancy: review of 11 cases. *American Journal of Obstetrics and Gynecology* **175**, 950–6.

HIV-associated HUS

Badesha PS, Saklayen MG (1996). Hemolytic uremic syndrome as a presenting form of HIV infection. *Nephron* **72**, 472–5.

Sutor GC, Schmidt RE, Albrecht H. (1999). Thrombotic microangiopathies and HIV infection: report of two typical cases, features of HUS and TTP, and review of the literature. *Infection* **27**, 12–15.

Transplantation and HUS

Conlon PJ, *et al.* (1996). Renal transplantation in adults with thrombotic thrombocytopenic purpura/haemolytic uraemic syndrome. *Nephrology, Dialysis and Transplantation* **11**, 1810–14.

Ducloux D, *et al.* (1998). Recurrence of hemolytic-uremic syndrome in renal transplant recipients: a meta-analysis. *Transplantation* **65**, 1405–7.

Verburgh CA, *et al.* (1996). Haemolytic uraemic syndrome following bone marrow transplantation. Case report and review of the literature. *Nephrology, Dialysis and Transplantation* **11**, 1332–7.

20.10.7 Sickle-cell disease and the kidney

G. R. Serjeant

Homozygous sickle-cell (**SS**) disease results in anaemia, a hyperdynamic circulation, less deformable red blood cells, and probably widespread endothelial damage and dysfunction. These processes affect structure and function in the kidney: medullary and glomerular involvement occurs at different ages and with different implications for outcome. Other genotypes of sickle-cell disease such as sickle-cell haemoglobin C (SC) disease, sickle-cell b°-thalassaemia, and sickle-cell b^+-thalassaemia manifest similar but less frequent and less severe changes. Even the sickle-cell trait is associated with some abnormalities of renal function.

Medullary involvement

Vascular damage

The vasa rectae system of the renal medulla with its low oxygen tension, high pH, and hypertonicity is uniquely conducive to sickling, causing disruption of the blood vessels and secondary damage to the renal tubules. Microradioangiographic studies have shown almost complete obliteration of the fine-vessel system of the vasa rectae, the remaining vessels being distorted into spirals, dilatations, and appearing to end blindly. These changes have been observed in SS disease in childhood and occur to a lesser extent in the sickle-cell trait in which haemoglobin S levels are only 20 to 45 per cent of total haemoglobin.

Tubular dysfunction

The functional effect of these vascular changes is an inability to concentrate the urine normally, which becomes worse with age but can be improved by transfusion in children under 2 years. Proximal tubular functional abnormalities include an increased secretion of urate and an increased reabsorption of phosphate and of b_2-microglobulin. Distal tubular functional abnormalities include an inability to excrete an acid load, defective maximal potassium excretion, and occasionally evidence of hyporeninaemic hypoaldosteronism.

Clinically, these changes are reflected in hyposthenuria with larger urinary volumes contributing to nocturia, enuresis, and possibly a tendency to dehydration.

Glomerular involvement

Large hypercellular glomeruli are characteristic of SS disease from the age of 2 years, the size continuing to increase with age even over the age of 40 years. The large glomeruli in childhood are associated with supranormal glomerular filtration rates, effective renal blood flows, and effective renal plasma flows. All these indices fall with age and in many patients aged 30 to 40 years are below normal, with particularly rapid decline occurring in some patients who proceed to chronic renal failure. This functional deterioration is assumed to reflect progressive glomerular loss, the mechanism of which is unclear, but there may be contributions from immune complexes derived from renal tubular epithelial antigen and mechanical damage to the nephron from hyperfiltration. The notion that glomerular capillary hypertension might be involved gained support from the observation that angiotensin-converting enzyme inhibitors significantly reduce proteinuria in some proteinuric patients with SS disease.

Clinical syndromes

Nocturnal enuresis

Enuresis is common in SS disease, and in the Jamaican Sickle Cell Cohort study occurred at least 2 nights weekly in 52 per cent of SS males and 38 per cent of SS females aged 8 years compared with values of 22 and 17 per cent in normal control children. Enuretic children with SS disease have slightly higher urine volumes and lower mean maximal functional bladder capacities than those without enuresis, and the ratio of overnight urine volume to bladder capacity was significantly greater in enuretic subjects. Enuresis alarms may therefore be the most appropriate therapy but require testing in controlled studies.

Haematuria

Haematuria occurs in both sickle-cell disease and sickle-cell trait and is believed to result from ischaemic lesions of the renal papilla, varying from minute ulcerations to renal papillary necrosis. Treatment is conservative, although prolonged haematuria may require blood transfusion or rarely limited surgery. Epsilon aminocaproic acid, which inhibits urokinase, has been effective in some cases, but promotes clots that may obstruct the ureters and its use requires assessment in controlled trials.

Urinary tract infections

The frequency of urinary tract infections is increased in subjects with the sickle-cell trait during pregnancy, and may be increased in SS disease, although no reliable data are available. *Escherichia coli*, *Klebsiella*, and *Enterobacter* spp. are most commonly responsible.

Acute glomerular disease

Poststreptococcal glomerulonephritis occurs in SS disease and may affect patients at a later age than the normal population. An association of proteinuria with leg ulceration raised the possibility that leg ulcers acted as a portal of entry for b-haemolytic streptococci, but further analysis did not support this hypothesis.

Nephrotic syndrome

It is unclear whether patients with SS disease are more prone to nephrotic syndrome, but the histological picture of membranoproliferative glomerulonephritis accounts for over half of adult cases. Nephrotic syndrome has been reported following parvovirus B19 infection, and B19-specific DNA demonstrated within renal biopsy tissue 1 year after the onset of nephrotic syndrome. If nephrotic syndrome is due to acute glomerulonephritis, the prognosis is good, but it is not if the cause is otherwise, with a 50 per cent mortality occurring within 16 months in one study.

Chronic renal failure

This is an important contributor to illness and death among adults with SS disease, especially those over 40 years of age. It is usually insidious in onset, manifested initially by a falling haemoglobin level attributable to lowered erythropoietin levels. The renal function of older patients should therefore be monitored regularly. Serum

creatinine levels tend to be lower than normal in steady-state SS disease and creatinine levels within the accepted normal range should not be interpreted as indicating normal renal function, indeed in SS disease it is likely that levels of 60 to 70 µmol/l reflect significant renal damage.

Since the early symptoms of chronic renal failure in those with SS disease are principally due to a low haemoglobin level, patients may be maintained in tolerable health for years by regular transfusion. The response to subcutaneous erythropoietin is unpredictable, large doses of erythropoietin being required to induce a response, but some patients showing dramatic increases in haemoglobin that may precipitate painful crises. Endstage renal failure may require long-term renal replacement therapy or renal transplantation, but there are conflicting data on outcome. Successful transplantation can be followed by striking increases in haemoglobin levels sufficient to precipitate painful crises, and recurrent sickle nephropathy may affect the transplanted kidney.

Further reading

Assar R *et al*. (1988). Acute poststreptococcal glomerulonephritis and sickle cell disease. *Child Nephrology and Urology* **9**, 176–9.

Bakir AA *et al*. (1987). Prognosis of the nephrotic syndrome in sickle glomerulopathy. *American Journal of Nephrology* **7**, 110–15.

Barber WH *et al*. (1987). Renal transplantation in sickle cell anemia and sickle disease. *Clinical Transplantation* **1**, 169–75.

Bhathena DB, Sondheimer, JH (1991) The glomerulopathy of homozygous sickle hemoglobin (SS) disease: morphology and pathogenesis. *Journal of the American Society of Nephrology* **1**, 1241–52.

Falk RJ *et al*. (1992). Prevalence and pathologic features of sickle cell nephropathy and response to inhibition of angiotensin-converting enzyme. *New England Journal of Medicine* **326**, 910–15.

Morgan AG, Serjeant GR (1981). Renal function in patients over 40 with homozygous sickle-cell disease. *British Medical Journal* **282**, 1181–3.

Readett DRJ, Morris JS, Serjeant GR (1990). Determinants of nocturnal enuresis in homozygous sickle cell disease. *Archives of Diseases in Childhood* **65**, 615–18.

Statius van Eps LW *et al*. (1970). Nature of concentrating defect in sickle-cell nephropathy. Microradioangiographic studies. *Lancet* **i**, 450–2.

Tejani A *et al*. (1985). Renal lesions in sickle cell nephropathy in children. *Nephron* **39**, 352–5.

Wierenga KJJ *et al*. (1995). Glomerulonephritis after human parvovirus infection in homozygous sickle cell disease. *Lancet* **346**, 475–6.

20.11 Renal involvement in genetic disease

J. P. Grünfeld

The spectrum of inherited renal disorders (and of inherited diseases with kidney involvement) is summarized in Table 1. Attention will be focused in this section on the commonest inherited kidney diseases leading to renal failure.

Cystic kidney diseases

Autosomal dominant polycystic kidney disease

Autosomal dominant polycystic kidney disease is by far the most frequent inherited kidney disorder, accounting for approximately 7 per cent of cases of endstage renal failure in Western countries. It is one of the most frequent human inherited monogenic diseases (approximately 1 in 1000 individuals). The spectrum of genetic cystic kidney diseases is summarized in Table 2.

Definition

The disease is characterized by the presence of multiple cysts, arising from various segments of the nephrons and involving both kidneys. The mechanisms underlying cyst formation and progression are poorly understood: cysts develop only in a small percentage of nephrons and only focally, whereas all nephron cells carry the mutated gene. This has been explained by a two-hit phenomenon which postulates that renal tubular (or liver biliary) cells which are at the origin of cysts bear first the germinal *PKD* gene mutation, and then acquire a somatic *PKD* gene mutation involving the other allele, this event occurring at random in a limited number of cells. This explanation does not exclude other mechanisms. The link between the genetic event(s) and cystic fluid accumulation is not known.

The disease is also characterized by its autosomal dominant mode of inheritance, so that the risk of any child of either parent carrying the abnormal gene is one in two. New mutations are very rare. The mutant gene responsible for 85 per cent of cases of autosomal dominant polycystic kidney disease has been located to the short arm of chromosome 16 (*PKD1* gene) by linkage analysis, and then identified. A second gene (*PKD2*) has been located to the long arm of chromosome 4 and cloned. The corresponding gene products have been named polycystin 1 and 2. Their normal function is unknown, but the two proteins interact. There is possibly a third locus, so far unidentified.

The diagnosis of autosomal dominant polycystic kidney disease is therefore based on the two following features:

1. evidence for autosomal dominant inheritance;
2. demonstration of multiple renal cysts in both kidneys, which are often enlarged, by ultrasonography (Fig. 1).

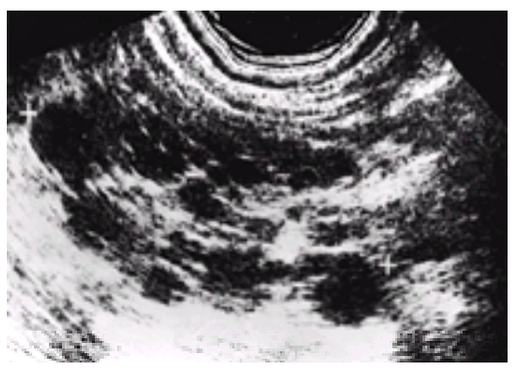

Fig. 1 Typical ultrasonographic findings in a patient with autosomal dominant polycystic kidney disease. The kidney is enlarged and contains multiple cysts of different sizes; the contralateral kidney had similar changes. The concentration of serum creatinine was 120 µmol/l at the time of ultrasonography. (By courtesy of Dr O. Helenon.)

The latter criterion deserves further comment. Renal cysts are initiated in the fetal kidney and develop progressively in life over the course of years. They may be too small to be detected by ultrasound in childhood, and kidney enlargement also progresses with age. Thus the sensitivity of ultrasonography for detecting autosomal dominant polycystic kidney disease is poor before 20 years of age (but the specificity is high since solitary renal cysts, *a fortiori* bilateral, are very rare at this age). In families with the *PKD1* gene mutation, false-negative ultrasonographic diagnosis is very unlikely at ages above 30 years, and rare at ages 20 to 29 years. Routine screening in asymptomatic members of affected families should not be performed before 20 years of age. By contrast, renal cysts, even bilateral, are relatively common in patients aged 50 years or more, hence strict criteria (multiple cysts in both enlarged kidneys and clear-cut inheritance) are required for diagnosing autosomal dominant polycystic kidney disease in older patients.

Symptoms

Renal manifestations

In some patients, autosomal dominant polycystic kidney disease is asymptomatic and discovered during family investigation or by chance on abdominal ultrasonography. In most cases, however, there are symptoms and patients complain of one or more of the following at some time during their life: renal pain due to cyst development, or stone or blood clot migration; bleeding within a cyst, leading to flank pain, the hyperdense cyst fluid then being visualized by computed

tomography; bleeding into the urinary tract, with gross haematuria occurring in approximately 30 per cent of cases; fever due to upper urinary tract infection, which is more frequent in women, or to cyst fluid infection. Renal stones, predominantly uric acid (for unknown reasons), develop in about 20 per cent of the patients.

Hypertension is a common and early finding in autosomal dominant polycystic kidney disease, occurring in about 30 to 50 per cent of patients at a stage when renal function is normal. Subsequently, with the development of renal failure, up to 80 per cent of patients become hypertensive. Why hypertension develops is not known: it has been ascribed to compression and ischaemia of the normal renal parenchyma by cysts.

Renal failure is also a common finding in autosomal dominant polycystic kidney disease. When it occurs, it usually progresses to endstage at between 40 and 60 years of age. However, in 30 per cent of cases it reaches endstage later, and in 5 per cent earlier, including very rare instances when it develops in the first years of life. Recent epidemiological studies have indicated that autosomal dominant polycystic kidney disease may have a much more indolent course in a substantial number of cases: 25 to 50 per cent of affected subjects are not in endstage renal failure by 70 years of age, and some patients may reach 80 or 90 years without the need for renal replacement therapy. This information is crucial for genetic counselling.

Genetic and non-genetic factors determine renal prognosis: the renal disease may progress more slowly in families with *PKD2* disease (mean age at endstage renal disease 55 years in *PKD1* disease compared with 70 years in *PKD2* disease), it progresses more slowly in women than in men, and control of hypertension may reduce the rate of progression.

Extrarenal manifestations

Liver cysts

These develop in 70 per cent of patients, usually later in life than renal cysts. Liver cysts are more frequent and more diffuse in women than in men. They are usually asymptomatic, detected by ultrasonography, but may be clinically palpable. Liver function tests are usually normal. Liver cyst infection may occur, particularly in patients on dialysis or transplant recipients. Massive liver involvement can cause severe discomfort in some cases, mostly in women.

Cardiovascular abnormalities

These include intracranial aneurysms and mitral valve prolapse. Subarachnoid haemorrhage or intracerebral bleeding due to rupture of intracranial aneurysm are among the most severe complications of autosomal dominant polycystic kidney disease and occur in approximately 1 to 2 per cent of patients. Rapid diagnosis and urgent neurosurgical opinion are required. Diagnosis should be suspected early, before complete rupture, in patients with autosomal dominant polycystic kidney disease with recent and severe headache or with any transient focal neurological deficit.

In cross-sectional studies performed using non-invasive screening methods such as high-resolution computed tomography or magnetic resonance angiography, intracranial aneurysms have been found in 7 to 8 per cent of asymptomatic middle-aged patients with autosomal dominant polycystic kidney disease. The prevalence is higher in those with a family history of intracranial aneurysm. The risk of rupture is largely dependent on aneurysm size. Routine screening by non-invasive methods is not indicated for all asymptomatic patients with autosomal dominant polycystic kidney disease, but it seems reasonable in certain subgroups, in particular those with a family history of intracranial aneurysm or subarachnoid haemorrhage, those who have already bled from an aneurysm (since recurrent aneurysm is possible), and possibly those who are to undergo major elective surgery. In high-risk groups, screening should be repeated every 5 to 10 years since the cerebral vascular disease is progressive.

Mitral valve prolapse is discovered in 20 per cent of patients with autosomal dominant polycystic kidney disease by echocardiography, whereas it is found in only 2 to 3 per cent of the general population. Other cardiac valve abnormalities and occasionally artery dissection or aneurysm may also be detected.

Other extrarenal defects are observed in autosomal dominant polycystic kidney disease: pes excavatum, colonic diverticulas, and abdominal hernias are more prevalent than in the general population.

Treatment

High fluid intake and regular follow-up of blood pressure and renal function are indicated in all patients with autosomal dominant polycystic kidney disease. The control of hypertension is an essential part of management, achieved with standard antihypertensive agents. Haematuria requires conservative management, although bleeding may sometimes be prolonged over several days and even weeks.

The relief of pain or abdominal discomfort can be difficult. In addition to symptomatic treatment, surgical renal cyst decompression should be restricted to very selected cases. Surgery is rarely needed in the management of renal stones. Liver cyst aspirations by needle under CT guidance, fenestration, or resection may be needed when massive involvement gives rise to pain; and in very rare cases such patients have come to liver transplantation.

Kidney infection requires administration of antimicrobials currently used in upper urinary tract infection. In some cases control of infection is not obtained, most probably because agents penetrate some infected cyst fluids poorly and do not achieve adequate concentration. Lipophilic drugs such as trimethoprim–sulphamethoxazole and ciprofloxacin have the best penetration into cyst fluid. Liver cyst infection also requires antimicrobials and drainage if infection is not controlled. Ciprofloxacin penetrates well into liver cyst fluid.

Standard medical management of chronic renal failure is indicated, as are renal replacement therapy and kidney transplantation when the patient reaches endstage, the results being similar to those obtained in other renal diseases.

Genetic counselling

The pattern of inheritance of autosomal dominant polycystic kidney disease means that the offspring of an affected subject have a 50 per cent risk of having the disease. The disease has a highly variable clinical course, even within a given family. Prenatal diagnosis by gene linkage studies using material derived from chorionic villous sampling has been performed and can be considered if required and if adequate family information is available. The demand for such prenatal diagnosis has, however, been very low in Western countries. This is explained by the late onset and the variable clinical course of the disease, often relatively benign, which cannot yet be predicted by DNA analysis.

Ultrasonography may occasionally show renal cysts in the fetus, but late in pregnancy. Obviously, due to the slow and late development of macrocysts, negative ultrasonography in the fetus (as well as in a child) does not rule out the disease.

Autosomal recessive polycystic kidney disease

Autosomal recessive polycystic kidney disease is a rare inherited disease (approximately 1 in 40 000 individuals), the first manifestations of which appear early in childhood.

Three features characterize this disease:

1. its recessive nature: both heterozygous parents are unaffected, with normal renal ultrasonography; parental consanguinity is found in some families; the mutant gene has been located on chromosome 6;
2. renal cysts derive from the collecting ducts, accounting for the striations in the dilated collecting system seen on intravenous pyelography; and
3. the renal disease is invariably associated with congenital hepatic fibrosis: this may be responsible for portal hypertension due to presinusoidal block, or for bacterial angiocholitis due to intrahepatic bile duct dilatation.

In children, autosomal recessive polycystic kidney disease should be differentiated from autosomal dominant polycystic kidney disease, which can be detected in childhood, even in neonates. Family history and renal ultrasonography in parents are decisive for correct diagnosis. In very rare families with *PKD1* disease, renal

involvement may be revealed in neonates and may progress to endstage within the first year of life.

The diagnosis of autosomal recessive polycystic kidney disease may be made before birth by antenatal ultrasonography, showing renal enlargement and increased echogenicity (as well as oligohydramnios). However, prenatal diagnosis may be uncertain and, since cystic changes occur in well-developed collecting ducts, these are detected only in the second half of pregnancy. When there is huge renal enlargement, pulmonary hypoplasia and respiratory distress may lead to death within hours after birth. With prolonged survival, liver and renal involvement becomes prominent. Gastrointestinal bleeding due to portal hypertension may be life-threatening and necessitate surgical portocaval shunt. Systemic hypertension is a frequent finding in the first year of life but, surprisingly, it may regress in subsequent years. Urinary tract infection is common. The rate of progression of renal failure is variable. Among patients who survive the neonatal period, approximately 50 per cent reach endstage in childhood, whilst this occurs in adulthood in the remainder.

Other cystic kidney diseases

Renal cysts may be found in von Hippel–Lindau's disease and in tuberous sclerosis (see below). Renal medullary cysts are also found in juvenile nephronophthisis, but not early in the course (see below). By contrast, such cysts—well localized in adults by ultasonography or CT scan—are seen early in autosomal dominant renal medullary cystic disease. This very rare condition progresses to endstage renal failure. Three different genetic loci have so far been localized.

Inherited diseases with glomerular involvement

Alport's syndrome

First described in 1927 by Dr Arthur Cecil Alport, this syndrome is characterized by the association of progressive haematuric hereditary nephritis and bilateral sensorineural hearing loss. Its prevalence is approximately 1 in 5000 individuals. In 85 per cent of kindreds the mode of transmission is compatible with X-linked dominant inheritance. In the remaining families, autosomal dominant or recessive inheritance should be considered.

Symptoms

Renal manifestations

The first clinical manifestation is typically gross haematuria, occurring sometimes in the first year of life, recurring during childhood, and followed by permanent microscopic haematuria. Proteinuria appears later. A nephrotic syndrome, usually moderate, develops in 30 to 40 per cent of patients. In other cases, moderate proteinuria and microscopic haematuria are the presenting symptoms in adulthood. The disease is progressive, leading to renal failure in all affected males, but the rate of progression is heterogeneous from one family to another, although usually homogeneous within a given family. In some endstage is reached at or before 30 years of age, sometimes in childhood; in others renal failure progresses to endstage between the ages of 30 to 60 years.

Carrier females of X-linked Alport's syndrome often have slight or intermittent urinary abnormalities. They may develop mild impairment of renal function late in life but do not usually have progressive renal disease as occurs in males, although this does happen in a few cases. In the autosomal recessive form of Alport's syndrome, renal disease progresses to endstage before 20 to 30 years of age at a similar rate in both affected men and women.

Extrarenal manifestations

The hearing defect may lead to severe perceptive deafness, but it is often moderate or slight, only detected at audiometric testing. The hearing loss labels a given family but is not found in all patients with renal disease. In some kindreds, familial haematuric progressive nephritis without hearing defect is documented: this form belongs to the spectrum of Alport's syndrome.

Other extrarenal manifestations may be found. Eye abnormalities are detected in 30 to 40 per cent of cases. These include bilateral anterior lenticonus detected by slit-lamp examination—a pathognomonic abnormality—and perimacular or macular retinal flecks that are seen by fundoscopic examination and do not alter visual acuity. Recurrent corneal erosions occur in some patients. In some families, macrothrombocytopenia is associated with nephritis and hearing defect. In other rare kindreds the latter features are found in association with diffuse leiomyomatosis, mainly oesophageal, and congenital cataracts.

Pathogenesis

The primary defect in Alport's syndrome involves the glomerular basement membrane. By electron microscopy, this membrane can be abnormally thickened with splitting of the lamina densa, thinned with focal thickening, or diffusely thin in younger children. In some patients, antigenicity of the glomerular basement membrane is abnormal: antiglomerular basement membrane antibodies do not bind linearly along the Alport glomerular basement membrane, whereas they show linear fixation along the glomerular basement membranes of normal and diseased kidneys which contain the corresponding Goodpasture antigen (the Goodpasture syndrome, an autoimmune disorder characterized by the development of antiglomerular basement membrane antibodies directed against this antigen (see Chapter 20.7.7).

In X-linked Alport's syndrome, the molecular defect involves the gene encoding for the a5 chain of the type IV collagen molecule. Type IV collagen is a major component of basement membranes. Six a chains of type IV collagen have been identified so far, with each molecule of type IV collagen being made up of three of these chains, differently associated in various basement membranes. In Alport's syndrome, mutations have been identified in the gene encoding for the a5 chain that maps to the long arm of the X chromosome. The Goodpasture antigen is located in the a3 chain, the gene of which has been mapped on chromosome 2. Absence or severe alteration of the a5 chain possibly prevents normal integration of the a3 chain into the glomerular basement membrane, leading to the defect in antigenicity.

In the autosomal recessive form of Alport's syndrome, the genes encoding for a3 or a4 chains are mutated. Affected subjects are homozygotes in consanguineous families, or compound heterozygotes in other cases. In families with leiomyomatosis, a5 and a6 genes, located contiguously on the X chromosome, are both involved in a large deletion.

Skin biopsy has become valuable for diagnosis of Alport's syndrome. Epidermal basement membrane normally contains a5 but not a3/a4 chains. Thus negative a5 staining by immunofluoresence is highly specific for X-linked Alport's syndrome, but is found in only 75 per cent of cases because a5 chains that are only slightly mutated can be detected. a5 Staining is normal in the autosomal recessive forms of Alport's syndrome.

In the disease with macrothrombocytopenia, mutations involve the *MYH9* gene, encoding the non-muscle myosin heavy chain IIA.

Genetic counselling and treatment

Genetic counselling first requires the correct identification of the mode of inheritance. If X-linked dominant inheritance is documented, affected men will not transmit the disease to their sons, whereas all their daughters will carry the mutant gene; affected women will transmit the mutant gene to 50 per cent of either sons or daughters. DNA analysis may be helpful for genetic counselling in these families.

Treatment of hypertension and supportive management of renal failure are indicated in patients with progressive disease. The results of kidney transplantation are similar to those obtained in other renal diseases. In rare cases, however, antiglomerular basement membrane crescentic glomerulonephritis develops in the graft. It is assumed that this complication is related to alloimmunization to the 'missing antigen' introduced by the transplant.

Benign familial haematuria

This disease is characterized by isolated microhaematuria, without proteinuria and progression to renal failure, in both men and women. Renal biopsy usually shows a thin glomerular basement membrane and immunofluorescence studies are negative. The mode of transmission is compatible with autosomal dominant inheritance. In some families, subjects with microhaematuria are heterozygotes carrying mutations involving the a3 or a4 chain gene.

Congenital nephrotic syndrome of the Finnish type

This disease specifically affects the kidney and is characterized by massive proteinuria, which occurs already *in utero* and then persists in infancy. Intense therapy is needed to afford the children a chance of survival: nutritional support to compensate for protein loss, prevention of infection and thrombosis, bilateral nephrectomy, continuous peritoneal dialysis, and finally kidney transplantation.

It is an autosomal recessive disease. The gene has been located on chromosome 19q and cloned. It encodes for a protein named nephrin, localized at the slit diaphragm between podocyte foot processes, which are both absent in affected subjects. Nephrin probably has a zipper-like structure and plays a key role in the normal glomerular filtration barrier.

Nail–patella syndrome

This syndrome, also known as hereditary osteo-oncho dysplasia, is a rare autosomal dominant disorder, defined by the association of nail hypoplasia or dysplasia, bone abnormalities (including iliac horns), and renal disease. The latter is found in 50 to 60 per cent of cases, progressing to endstage in approximately 15 per cent. The hallmark of renal involvement is the detection by electron microscopy of fibrillar collagen bundles within the glomerular basement membrane. Open angle glaucoma is a feature in rare families.

The mutated gene, *LMX1B* (located on 9q), belongs to a family of transcription factors that are involved in pattern formation during development. *LMX1B* is more specifically involved in the dorsoventral patterning of the limbs, and mice with a deletion in their *lmx1*-homologue exhibit skeletal defects similar to those observed in nail–patella syndrome and abnormal dorsoventral patterning of the extremities of the limbs.

Metabolic diseases with glomerular involvement

Anderson–Fabry disease

This disease is X-linked recessive (prevalence approximately 1 in 40 000 individuals) and due to a-galactosidase A deficiency resulting in glycosphingolipid deposition, mainly in the cardiovascular and renal system. The first manifestations in hemizygotes are painful acroparaesthesias, appearing in childhood, often prevented by continuous administration of carbamazepine or diphenylhydantin. Subsequently, angiokeratomas, anhydrosis, and corneal deposits develop. Ischaemic cerebrovascular complications, cardiac valve abnormalities, myocardial deposition of glycolipids, and coronary accidents are the most severe manifestations, along with renal involvement.

In the kidney, glycolipid deposition involves glomerular epithelial cells, tubular cells, and endothelial and smooth muscle cells of intrarenal arteries. The latter changes are responsible for progressive renal ischaemia. Renal disease is revealed by proteinuria at around 20 years, and then progresses to endstage between 40 and 60 years of age, necessitating regular dialysis and/or kidney transplantation. Glycolipid deposition does not recur in the renal graft that contains normal a-galactosidase activity. Enzyme replacement therapy is available.

Heterozygote female carriers usually have few symptoms. Corneal deposits are found in 70 per cent of them. They can develop cardiac changes and, very rarely, symptomatic renal disease.

Lecithin-cholesterol acyl-transferase (LCAT) deficiency

This is a very rare autosomal recessive disorder. LCAT is a key enzyme in the metabolism of cholesterol, responsible for its esterification. In affected subjects the proportion of cholesteryl ester to total cholesterol is very low. Lipid accumulation occurs in the eyes (causing corneal deposits), erythrocyte membranes (leading to low-grade haemolytic anaemia), arterial walls (contributing to premature atherosclerosis), and kidneys, predominating in glomerular mesangial cells and progressing to endstage renal disease. LCAT is expressed primarily in the liver, hence liver transplantation would theoretically be the treatment of choice, but it has not so far been performed in this disease. Patients have received kidney transplants: lipid deposition recurs slowly in the graft.

Type I glycogen storage disease

Also named von Gierke's disease (see Chapter 11.2), this disease is due to glucose-6-phosphatase deficiency. Affected infants develop hypoglycaemia, growth retardation, and hepatomegaly. Fanconi syndrome may occur as a consequence of glycogen deposition in the proximal tubule. Progressive renal involvement is not due predominantly to glycogen accumulation. It is rather related to the development of focal segmental glomerulosclerosis, the mechanism of which is unclear, usually after 20 years of age. This complication has only been recognized recently since children with severe hypoglycaemia have now survived to adulthood thanks to the progress achieved by paediatricians, dieticians, and families in providing adequate feeding and nutrition.

Familial primary glomerulonephritis

In most types of primary glomerulonephritis, familial cases have been anecdotally reported. The most frequent form, albeit rare, is probably familial IgA nephropathy, either primary (Berger's disease) or associated with Henoch–Schönlein purpura. Familial focal segmental glomerulosclerosis with either autosomal dominant or autosomal recessive inheritance has also been well characterized and several genetic loci have been identified.

Inherited tubulointerstitial disorders

Juvenile nephronophthisis

This complex represents an inherited form of chronic tubulointerstitial disease. Cysts located at the corticomedullary junction or in the medullary region appear late in the course of the disease. Renal pathological examination reveals tubular atrophy and interstitial fibrosis (which are non-specific lesions), and extreme thickening and multilamellation of the tubular basement membrane.

Juvenile nephronophthisis is a major cause of endstage renal disease in children, accounting for 10 to 20 per cent of cases. It is transmitted as an autosomal recessive trait. The gene involved, *NPH1*, has been mapped to the short arm of chromosome 2 and its product has been named nephrocystin. In 80 per cent of cases a homozygous deletion is found. Heterozygotes are asymptomatic.

The clinical manifestations first appear around the age of 4 years and consist of polyuria, secondary enuresis, and polydipsia, reflecting a urinary concentration defect. Renal failure, metabolic acidosis, anaemia, and growth retardation subsequently develop, and endstage renal failure is usually reached at the age of 10 to 13 years. More or less severe renal salt wasting is a common finding. In approximately 10 to 15 per cent of cases renal involvement is associated with retinal changes: tapetoretinal degeneration with or without retinitis pigmentosa, leading to blindness early or later in life (this association is referred to as the Senior–Loken syndrome; in most of these cases, no deletion of the *NPH1* gene has been detected). Other extrarenal features (skeletal changes, cerebellar ataxia, and liver fibrosis) are rare.

Familial nephropathy with juvenile hyperuricaemia and gout (or familial juvenile hyperuricaemic nephropathy)

This disease is characterized by its autosomal dominant inheritance, juvenile onset of gout, hyperuricaemia disproportionate to the age, sex, or degree of renal dysfunction, which is due to low renal fractional excretion of urate, and renal failure often recognized between 20 and 40 years of age. Renal biopsy shows non-specific tubulointerstitial changes. Allopurinol is indicated to prevent gout and perhaps to slow the progression of renal disease.

Genetic disorders with nephrolithiasis

Pertinent clinical data on these disorders are summarized in Table 3. Additional information can be found in Chapter 20.13, Chapter 20.8, Chapter 11.10, and Chapter

11.4.

Other genetic diseases with kidney involvement

Phakomatoses

Two diseases of this group have significant renal involvement: von Hippel–Lindau's disease and tuberous sclerosis.

In von Hippel–Lindau's disease, renal cysts and bilateral multifocal renal cell carcinomas are found in 70 per cent of the patients. Carcinomas are often asymptomatic, should be screened for regularly (Fig. 2), and occur at a mean age of 45 years. Nephron-sparing surgery (tumorectomy) is advocated when technically feasible. The other clinical features of von Hippel–Lindau's disease are described in Chapter 14.8.

Fig. 2 Computed tomography of the kidneys in a patient with von Hippel–Lindau's disease. In the right kidney, a solid tumour is found as well as cystic changes. In the left kidney, a voluminous multilocular tumour is detected with thick walls, corresponding to renal clear cell carcinoma, associated with other cystic lesions.

The most typical renal lesion encountered in tuberous sclerosis is angiomyolipoma, which is a benign tumour, often multiple and bilateral. By ultrasonography, this tumour is hyperechogenic and by CT is characterized by its fat content (Fig. 3). Bleeding is the main complication of renal angiomyolipoma. Multiple angiomyolipomas may severely reduce renal mass and lead to renal failure, but this is rare. The development of segmental glomerulosclerosis may accelerate the progression to endstage. Renal cysts may also be found in *TSC2* forms (see below). The incidence of renal cell carcinoma is slightly higher than in the general population. The other features of tuberous sclerosis involve the skin and the central nervous system (Chapter 24.8).

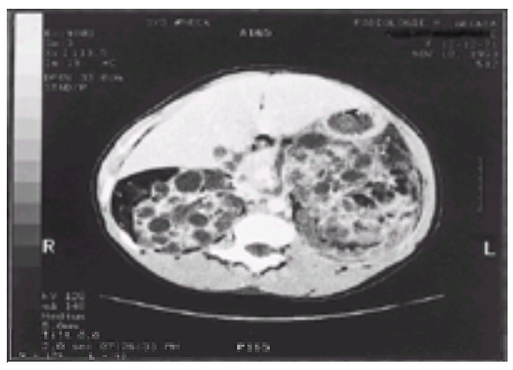

Fig. 3 Multiple bilateral renal angiomyolipomas in a patient with tuberous sclerosis (CT scan). Note the voluminous angiomyolipoma (with high content of fat that is black) at the periphery of the right kidney.

The genes mutated in von Hippel–Lindau's disease and tuberous sclerosis are tumour-suppressor genes. Two mutations ('two-hit' phenomenon) are required to trigger tumour formation: the first one germinal, inherited, and the second one somatic. The *VHL* gene has been cloned and located on 3p. Two somatic mutations of the same gene are involved in sporadic renal cell carcinoma. Two genes are identified in tuberous sclerosis: *TSC1* on chromosome 9q encoding for hamartin, and *TSC2* on chromosome 16p, encoding for tuberin.

Cystinosis

Cystinosis results from defective carrier-mediated transport of cystine through the lysosomal membrane. The disease is transmitted as an autosomal recessive trait with an incidence of about 1 in 200 000 live-born babies. The gene has been mapped on chromosome 17, has been identified, mutations have been characterized, and it encodes for a protein named cystinosine. The diagnosis is based on the findings of cystine crystals in tissues, such as the eyes, and on the elevated cystine content in leucocytes. It should be clear that cystinosis is completely different from cystinuria, which is due to a defective reabsorption of cystine in the proximal tubule.

The clinical manifestations are due to progressive intralysosomal accumulation of cystine. In the infantile form, the first symptoms are related to the clinical consequences of Fanconi syndrome (salt and water depletion, hypokalaemia, acidosis, rickets) appearing before 6 months of age. Renal failure develops later, reaching endstage generally before 12 years. Cystine accumulates in other tissues, before and after kidney transplantation: eyes (photophobia due to corneal deposits, then retinal depigmentation and visual impairment), thyroid gland (hypothyroidism), liver and spleen (portal hypertension), pancreas (diabetes mellitus), muscles, testis, and central nervous system (encephalopathy) (see Chapter 11.3).

In addition to symptomatic management, cysteamine has proved to be effective in cystinosis. It accumulates within lysosomes, promotes cystine outflow, and thus reduces tissue cystine content. Administration of this drug should be started as soon as the diagnosis is made. It may slow the rate of progression of renal failure and prevent most extrarenal complications. Despite recent progress, tolerance of the drug is not good because of its offensive taste and odour: compliance may therefore be poor. Topical cysteamine prevents corneal crystal deposition.

Juvenile cystinosis presents in late childhood or early adult life. Renal involvement occurs. By contrast, in the adult form only corneal crystals are found. Both forms are very rare.

Malformation syndromes with kidney involvement

The most frequent of these rare syndromes is Bardet–Biedl syndrome. This is a heterogeneous autosomal recessive condition for which six different genetic loci have been identified. Clinical features comprise obesity, hypogonadism (in males), polydactyly or dystrophic extremities, retinal dystrophy (leading to blindness), and renal abnormalities. The last have only been recognized recently as a cardinal feature in the syndrome. Renal imaging often shows the following abnormalities: calyceal clubbing and pronounced diverticulas, and lobulated renal outlines of the fetal type. These changes are probably dysplastic in nature, and are characteristic when associated. Renal cortical and medullary cysts have also been found by ultrasonography, but the latter may be difficult to differentiate from calyceal diverticulas. Approximately 25 per cent of patients develop chronic renal failure, progressing to endstage, which is probably the major cause of death. The most important treatment is the provision of specialized education with low-vision aids. Symptomatic management of diabetes mellitus (found in 30 per cent), hypertension, and renal

failure is required.

Renal hypoplasia or unilateral renal agenesis is found in other malformation syndromes, such as the following.

1. Kallmann's syndrome—with hypogonadism and hyposmia or anosmia.
2. Branchio-oto-renal syndrome—where laterocervical fistulas or cysts and otic abnormalities, involving the outer, middle, or inner ear are found, and the *EYA1* gene, on the long arm of chromosome 8, is mutated. This gene is the homologue of a gene present in drosophila, the mutation of which leads to eye absence.
3. Renal–coloboma syndrome—with optic nerve coloboma and sometimes hearing defect, and where the *PAX2* gene, located on 10q, is mutated.
4. Alagille's syndrome—characterized by paucity of intrahepatic bile ducts leading to cholestasis, vertebral abnormalities (butterfly vertebra), and heart defects. The gene has been identified.

All these genes implicated in malformation syndromes are involved normally in the control of kidney development.

Further reading

Cystic kidney diseases

Gabow PA (1993). Autosomal dominant polycystic kidney disease. *New England Journal of Medicine* **329**, 332–42.

Pirson Y, Chauveau D (1996). Intracranial aneurysms in ADPKD. In: Watson ML, Torres VE, eds. *Polycystic kidney disease*, pp. 530–47. Oxford University Press, Oxford.

Pirson Y, Chauveau D, Grünfeld JP (1998). Autosomal-dominant polycystic kidney disease. In: Davison AM *et al.*, eds, pp 2393–415. *Oxford textbook of clinical nephrology*. Oxford University Press, Oxford.

Zerres K, Volpel MC, Weiss H (1984). Cystic kidneys: genetics, pathologic anatomy, clinical picture, and prenatal diagnosis. *Human genetics* **68**, 104–35.

Alport's syndrome

Flinter FA *et al.* (1988). Genetics of classic Alport's syndrome. *Lancet* **ii**, 1005–7.

Grünfeld JP, Knebelmann B (1998). Alport's syndrome. In: Davison AM *et al.*, eds. *Oxford textbook of clinical nephrology*, pp 2427–37. Oxford University Press, Oxford.

Kashtan C, Michael AF (1996). Alport syndrome. *Kidney International* **50**, 1145–63.

Pirson Y (1999). Nephrology Forum: making the diagnosis of Alport's syndrome. *Kidney International* **56**, 760–75.

Inherited diseases with glomerular involvement

Morgan SH, Grünfeld JP (1998). *Inherited disorders of the kidney*. Oxford University Press, Oxford.

Inherited tubulointestinal disorders

Cameron JS *et al.* (1998). Inherited disorders of purine metabolism and transport. In: Davison AM *et al.*, eds. *Oxford textbook of clinical nephrology*, pp 2469–84. Oxford University Press, Oxford.

Hildebrandt F, Jungers P, Grünfeld JP (2001). Medullary cystic and medullary sponge renal disorders. In: Schrier RW, ed. *Diseases of the kidney*, pp 521–46. Little, Brown and Company, Boston.

Genetic diseases with kidney involvement

Parfrey PS (1998). Bardet–Biedl syndrome. In: Morgan SH, Grünfeld JP, eds. *Inherited disorders of the kidney*, pp 321–39. Oxford University Press, Oxford.

20.12 Urinary tract infection

C. Tomson

Introduction and definitions

Infection of the urinary tract is important for different reasons in different age groups. In infants and children, ascending infection is thought to be a preventable cause of renal parenchymal scarring and eventual renal failure, although (as discussed below) it is controversial how frequently this occurs. In adult women, recurrent lower urinary tract infection ('cystitis') is a common cause of misery and time off work. In all age groups, persistent or relapsing infection is an important indicator of abnormal host defences, usually due to abnormal anatomy or function of the urinary tract, and may result in irreversible renal damage unless the underlying cause is dealt with. Urinary tract infections are the cause of over 50 per cent of Gram-negative septicaemic episodes. In the elderly, non-specific symptoms including toxic confusional states are often due to occult urinary tract infection, and asymptomatic bacteriuria is associated with increased mortality.

'Urinary tract infection' refers to bacterial or fungal infection of the kidneys, pelvis, ureters, or bladder (viral infections may involve the urinary tract, as in Hantaan virus infection, but viruria more commonly reflects systemic viral infection). Infections primarily involving the urethra are nearly lways sexually acquired and are dealt with elsewhere (see Section 21). 'Pyelonephritis' refers to infection primarily involving the kidneys and collecting systems. 'Cystitis' refers to infections localized to the urinary bladder. 'Recurrent' urinary tract infections are due to repeated reinfection, whether by similar organisms on each occasion or by different species; 'relapsing' and 'persistent' infections are due to the continued presence of the same organism, suppressed or not suppressed during antibiotic therapy. 'Uncomplicated' urinary tract infection occurs in an anatomically and functionally normal urinary tract; 'complicated' infection refers to all infections occurring in patients either with impaired host defence (e.g. diabetes) or with abnormal urinary tract anatomy (e.g. urinary tract obstruction).

Epidemiology

Symptomatic bacterial urinary tract infection is one of the commonest bacterial infections. Around 1 per cent of boys and 3 per cent of girls will develop a urinary tract infection during childhood, and 50 per cent of women have a history of at least one episode of urinary tract infection, with recurrent infections in a significant minority. Urinary tract infection is rare in men until after the age of 60, when the rising prevalence of prostatic bladder outflow obstruction leads to an increased risk of infection. Asymptomatic bacteriuria is found in about 10 per cent of elderly men and in 20 per cent of elderly women. Each year in the United Kingdom around 60 women per 1000 population visit their general practitioner with urinary tract infection. Urinary tract infection is responsible for over 25 per cent of all community-acquired bacteraemias, more than any other source of infection, and accounts for over 40 per cent of hospital-acquired infections, often as a result of bladder catheterization.

Causative organisms

The commonest causative organisms in bacterial urinary tract infection are Gram-negative gut organisms, particularly *Escherichia coli* (Table 1). This reflects the fact that most infections reach the urinary tract via the urethra from the perineum. However, as discussed below, only some subtypes of *E. coli* and only some of the other species of gut organisms have the necessary virulence characteristics to enable infection of the normal urinary tract.

Pathophysiology

The great majority of urinary tract infections are acquired by ascent of the infecting organism up the urethra; only a very small minority result from haematogenous spread or—even loss commonly—from vesicoenteric fistulas. The pelvis, ureters, bladder, and urethra possess a highly specialized epithelium, which normally maintains complete impermeability to all components of urine, including toxins and water. This is maintained by tight junctions between the surface layers of epithelial cells, with a very high transepithelial electrical resistance. In the bladder, this impermeability has to be maintained despite repeated large changes in surface area as the bladder fills and empties. This is maintained by unfolding and refolding of the large, highly folded 'umbrella' cells that form the uppermost layer of the epithelium, together with insertion and endocytosis of vesicles, ready-lined with uroplakin, a hexagonal transmembrane protein found only on the surface of umbrella cells. In experimental models, infection is associated with a marked reduction in transepithelial resistance and loss of tight junctions, allowing components of urine to stimulate pain fibres and inflammatory cytokine release.

Ascending infection takes place in a series of steps, at each of which defective host defence increases the chance of successful establishment of infection (Fig. 1).

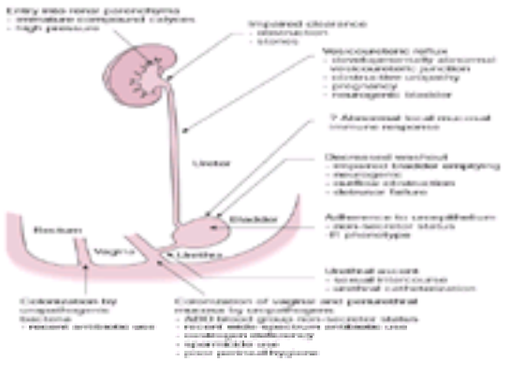

Fig. 1 Mechanisms allowing ascent of infection up the urinary tract.

The ability of a bacterium to colonize the gut and periurethral mucosa, and subsequently to adhere to the uroepithelium, is a major determinant of its ability to cause clinical infection, particularly if other host defences are intact. This ability to adhere is governed by specific interaction between bacterial adhesins, located on the tips of thin filaments ('pili' or 'fimbrias'), with genetically determined glycoproteins on the cell surface of the host cell. Type 1 fimbrias bind to mannose-containing glycoproteins that are present on the surface of uroepithelial cells, but also to Tamm–Horsfall protein, which is present in urine and can competitively inhibit binding of bacteria to cell surface glycoproteins. Type P pili bind the a-galactosyl-1,4-b-galactose disaccharide sequence present in some glycoproteins and glycosphingolipids, including the human P blood-group antigen system and also on the cell surface of uroepithelial cells as well as red cells. Some uropathogens are particularly adapted to colonizing foreign surfaces, particularly those coated by biofilm or mucin; for example, *Proteus* spp. are able to transform into a swarming phenotype with massive flagellas, organize into rafts, and move very rapidly against the flow of urine—they are therefore important causes of infection in patients with indwelling urinary catheters and those with ileal conduits.

Following adherence, fimbriae appear to retract, drawing the organism closer to the surface of the uroepithelial cell. Adherence is followed by apoptosis, exfoliation, and excretion of infected superficial cells and replacement by less differentiated cells, a process that may also contribute to host defence.

Bacterial adherence results in the local production of interleukin 8, which results in neutrophil migration through the uroepithelium into the bladder. Inflammatory cytokine release may also be promoted by soluble bacterial stimuli, such as lipopolysaccharide.

Variability in host defence

Frequency and completeness of bladder emptying

For an ascending infection to become established in the bladder, the number of organisms needs to reach a critical mass. The chance of this happening is reduced by increased urine flow rate, causing dilution of organisms within the bladder, and by frequent voiding, which also flushes the urethra and helps to prevent ascent of organisms into the bladder. This is termed 'hydrokinetic' defence. Habitual infrequent voiding is thought to be a risk factor for recurrent urinary tract infection for this reason. Patients with recurrent urinary tract infections are routinely advised to increase fluid intake and frequency of voiding, and some women report that a high fluid intake alone is enough to clear symptomatic infection. Incomplete voiding, which may be present in both sexes and is not necessarily due to outflow obstruction (Table 2) is an important cause of increased susceptibility to urine infection.

Vesicoureteric reflux

During normal micturition urine is expelled into the urethra, while retrograde flow ('reflux') of urine into the ureters is prevented because muscular contraction of the bladder wall results in closure of the vesicoureteric junctions. Reflux of urine into the ureters can occur if this mechanism is defective, sometimes as far as the renal pelvis, followed by return to the bladder once bladder contraction has finished. The most common cause of reflux is abnormal insertion of the ureters into the bladder: this occurs as a relatively frequent developmental anomaly, the other major cause being abnormally high intravesical pressure, for instance in high-pressure chronic retention of urine due to bladder outflow obstruction, or in neurogenic bladder in patients with partial spinal cord lesions. Whatever the cause, reflux of urine results in failure to expel all bladder urine during micturition, and therefore significantly impairs host defence against infection, as well as being associated with a greatly increased risk of infection ascending to the kidneys and causing acute pyelonephritis. Vesicoureteric reflux is frequently found in children with urinary tract infection. The question of whether ascending infection is a cause of renal damage in children with reflux is discussed later in this chapter.

Foreign bodies, stones, and privileged sites

The presence of a foreign body, such as a urinary catheter or ureteric stent, or of a stone within the urinary tract, creates a protected site where uropathogenic organisms can adhere and multiply, relatively immune to both hydrokinetic and mucosal defence mechanisms. In this situation it is often impossible to eradicate urine infection unless the foreign body or stone is removed completely, and prolonged use of antibiotics often results in the acquisition of resistance by the infecting organism. Urinary infection is nearly inevitable after a few weeks of bladder catheterization, which has led to attempts to develop catheter materials that are less easily colonized by bacteria and might thereby reduce the risk of infection. Meta-analysis shows that silver alloy urinary catheters may be cost-effective in preventing urinary infection in patients catheterized for a short time. Other 'privileged' sites include renal cysts (as in polycystic kidney disease, discussed below) and bladder diverticulas.

Sexual behaviour

Many women first experience acute cystitis shortly after becoming sexually active. Most women have transient bacteriuria after sexual intercourse, which develops into symptomatic cystitis only in a minority. In case–control studies of young women, the risk of urinary tract infection was associated with vaginal intercourse, and increased further by condom use. These findings are explained by the mechanical effect of intercourse encouraging ascent of organisms up the urethra, an effect that may be exacerbated by condom use, particularly without lubricants. The risk of urinary tract infection is also increased by a change in sexual partner, which may reflect male to female transmission of uropathogens. Use of spermicides as an adjunct to barrier contraceptive methods is also associated with an increased rate of periurethral colonization with *E. coli* and other uropathogens and with an increased risk of symptomatic urinary tract infection, probably because the active component in spermicides (nonoxynol-9) is bactericidal against lactobacilli. The protective effect of micturition soon after intercourse, based on the supposition that washout of recently introduced bacteria will prevent establishment of infection, remains unproved.

Vaginal and periurethral flora

Vaginal secretions are normally colonized by lactobacilli that appear to protect against colonization by uropathogenic bacteria such as *E. coli*. The mechanism of this protection is uncertain but may in part be related to the maintenance of an acidic pH, which suppresses growth of some uropathogenic bacteria. Suppression of this normal vaginal colonization by antibiotic treatment or by spermicide use increases the risk of colonization of the periurethral mucosa by uropathogenic bacteria and subsequent ascending urinary tract infection. In addition, atrophic vaginitis caused by oestrogen deficiency is associated with the absence of lactobacillus colonization, which may be part of the reason for the increased risk of urinary tract infection in postmenopausal women. However, attempts to prevent recurrent urinary infection by re-establishing colonization by lactobacilli have so far not yielded convincing results.

Genetic factors

In laboratory studies, adherence of *E. coli* to both vaginal and buccal cells is greater in women with recurrent urinary tract infection than in healthy controls, and women with recurrent urinary tract infection more frequently have gut colonization by uropathogenic strains of *E. coli*, suggesting that they experience more frequent urinary tract infections because they are more susceptible to colonization of the periurethral area by uropathogenic bacteria. It appears that this difference in susceptibility to colonization and infection, especially in patients in whom there is no other defect of host defence (such as vesicoureteric reflux), is due to genetically

determined differences in the extracellular antigens to which bacteria adhere, in particular in the expression of blood group antigens. The density of glycosphingolipids is higher in patients with the P1 blood group than those with the P2 blood group, and the P1 blood group is a risk factor for acute pyelonephritis among girls without vesicoureteric reflux. Expression of the large oligosaccharide A, B, H blood group antigens on the cell surface partially or completely obscures the smaller glycosphingolipids, preventing them from being bound by type P pili, which is why women with the secretor phenotype, in which these antigens are both expressed on the cell surface and secreted, are less prone to most *E. coli* infections than non-secretors. Non-secretors also have an increased inflammatory response (fever and acute-phase response) to urinary infection compared with secretors, and non-secretors are over-represented among patients with urographic evidence of reflux nephropathy. However, some *E. coli* strains only bind to cells from subjects who are secretor-positive blood group A.

Local immunity

Another aspect of host defence is the local production of antimicrobial peptides, secreted by uroepithelial cells into the urine, and the secretion of immunoglobulin A into the urine. However, there is little convincing evidence that impaired local IgA secretion is responsible for increased susceptibility to urinary tract infection. Patients with inherited or systemic defects in systemic immunity, whether cellular or humoral, do not appear to be at greatly increased risk of urinary tract infection; the increased risk of urinary tract infection in homosexual men with the acquired immune deficiency syndrome is associated with anal intercourse.

Pathogenicity

A few species of bacteria, collectively known as 'uropathogenic' bacteria, together account for most urinary tract infections (Table 1); the presence of non-uropathogenic species suggests an abnormality of host defence. Within uropathogenic species there are strains that are capable of causing infection and other strains that are far less likely to do so: uropathogenicity is determined by expression of cell surface molecules determining adhesion to receptors on uroepithelial cells, toxin production, factors conferring resistance to the membrane attack complex, and virulence factors. The factors determining pathogenicity of *E. coli* have been studied extensively, but much less is known about the determinants of pathogenicity of other uropathogenic bacteria, although motility may be an important determinant of the ability of *Proteus* spp. to ascend the urinary tract and cause infection; and *Staphylococcus saprophyticus*, an important cause of urinary tract infection in sexually active young women, is probably better able to cause urinary tract infection than *Staphylococcus aureus* or *Staphylococcus epidermidis* because of its possession of a lactosamine adhesin permitting adherence to uroepithelial cells.

Clinical presentation of uncomplicated urinary tract infection

Frequency and dysuria

The commonest presentation of urinary tract infection is with 'cystitis', a symptom complex associated with lower urinary tract infection in which many of the symptoms are directly attributable to increased bladder irritability caused by local infection. Classic symptoms include:

1. severe dysuria, often described as 'scorching' or 'like peeing barbed wire', worse towards the end or immediately after micturition;
2. increased urinary frequency, including nocturia—which helps to distinguish cystitis from other causes of daytime frequency;
3. urgency—the feeling of having to pass urine straight away to avoid incontinence;
4. urge incontinence—leakage of urine associated with the desire to pass urine;
5. strangury—the feeling of needing to pass urine despite just having done so;
6. offensive-smelling urine, often described as 'strong' or 'fishy';
7. macroscopic haematuria—particularly in women under 50, less commonly in girls or older women;
8. constant lower abdominal aching, not just in the genital area but also in the back, flanks, and lower abdomen; and
9. non-specific malaise, aching all over, nausea, tiredness, irritability, and cold sweats.

Not all of these symptoms are specific for lower urinary tract infection. Dysuria may be due to cystitis, urethritis, or vaginitis, but the latter two conditions are usually not associated with urinary frequency, and may be associated with vaginal discharge or itching and with specific findings on vaginal examination. Symptoms of cystitis may be due to causes other than lower urinary tract infection, for instance drug-induced cystitis.

Genuine inflammation of the bladder wall may or may not be present, and it is important to remember that there may be non-infective causes of increased bladder irritability, such as chemical- or drug-induced cystitis. Most patients with cystitis do not have fever, nor is there evidence of an acute-phase response.

Asymptomatic bacteriuria

By definition, this is an incidental finding in patients whose urine is cultured despite the absence of urinary tract symptoms. It is seldom justified to send a urine sample from an asymptomatic patient for culture, so this diagnosis should only rarely be made in clinical practice. An important exception is during pregnancy. Pregnant women with bacteriuria are at increased risk of acute pyelonephritis (occurring in 20 to 30 per cent of untreated women), of delivering low birth-weight babies, and of premature delivery, and antibiotic treatment significantly reduces these risks. Elderly patients with asymptomatic bacteriuria are also at increased risk of death, but this is probably because bacteriuria is a marker of poorer general health, and antibacterial treatment has not been shown to improve survival in this situation.

Acute pyelonephritis

The term acute pyelonephritis denotes infection within the renal pelvis, with or without active infection within the renal parenchyma. The diagnosis is usually made on the basis of the presence of flank pain (usually unilateral), fever, rigors, raised C-reactive protein (or erythrocyte sedimentation rate or plasma viscosity), neutrophilia, and evidence of urine infection on culture of a mid-stream urine sample. However, rigorous tests to localize the site of infection (discussed below) show that the correlation between the presence or absence of bacteriuria in the upper urinary tract and the presence or absence of flank pain, systemic symptoms, and an acute-phase response is dismally poor; many patients with infection confined to the bladder have flank pain and fever, whereas over 60 per cent of elderly women with asymptomatic bacteriuria have upper tract infection. The symptoms and signs of so-called 'acute pyelonephritis' are therefore in reality those of a marked host response to urinary tract infection, irrespective of whether organisms are multiplying in the renal pelvis or in the bladder.

Diagnosis

Inspection and dipstick testing

In a 'classic' case of established urinary tract infection the urine is cloudy, has an offensive smell, and is positive for blood, protein, leucocyte esterase, and nitrite on dipstick urinalysis. In this situation it is reasonable to make a diagnosis of urinary tract infection without further delay, and to institute empirical treatment. Whether a midstream urine sample should also be sent to the laboratory for confirmation and identification of the causative organism depends on the clinical situation, as discussed below. However, in many situations the diagnosis is not so obvious, and the diagnostic accuracy of inspection and dipstick testing less good.

1. Cloudy urine may be caused by bacteria and pyuria, but may also be caused by amorphous phosphate crystals that form in normal urine as it cools. Low concentrations of bacteria and white cells will not cause sufficient turbidity to be detected on visual inspection.
2. An offensive, fishy smell is highly suggestive of urinary tract infection, but relatively infrequent.
3. Macroscopic haematuria can certainly occur as a result of severe cystitis, but is frequently absent in genuine urinary tract infection and is more often due to glomerular bleeding or urothelial bleeding as a result of tumours or stones. Dipstick detection of haematuria is neither sensitive nor specific for the detection of urinary tract infection.
4. Proteinuria can occur in urinary tract infection as a result of the release of proteins from white cells, but is neither specific nor sensitive.
5. Leucocyte esterase is an enzyme released by white cells and a reliable test for pyuria, which is in most situations a major diagnostic criterion for urinary tract infection, as discussed in the next section. A positive test indicates 10 white cells/ml. Transport of urine samples in containers containing boric acid can result in false-negative leucocyte esterase tests, as the boric acid inhibits the enzyme.
6. Nitrite is produced by most uropathogens, which reduce urinary nitrate to nitrite, but not by Gram-positive organisms. A positive test for nitrite is highly

suggestive of urinary tract infection. False negative tests can be seen in patients with low dietary nitrate and in those taking high-dose ascorbic acid.

A combination of visual inspection and dipstick testing is therefore a reasonable screening test for patients in whom uncomplicated urinary tract infection is suspected on clinical grounds: in this situation, crystal clear urine and negative dipsticks for nitrite and leucocyte esterase make the diagnosis of urinary tract infection very unlikely (Table 3). The worst that is likely to happen if the diagnosis is missed is that the patient will re-present with more obvious abnormalities due to progression of the urinary tract infection to a more severe stage. However, in situations in which it would be important not to miss the opportunity to start treatment early, for example in patients with known abnormalities of host defence, pregnancy, or previous acute pyelonephritis, or in suspected atypical infections, formal microscopy and culture of the urine is required.

Microscopy and culture of urine

At first sight, the diagnosis of bacterial infection in the urinary tract should be straightforward, relying on culture of freshly voided urine. However, urine samples are very easily contaminated during voiding by bacteria from the perineal skin (or, to a lesser extent, the foreskin in males), resulting in false-positive results. The only certain way to circumvent this problem is to take urine directly from the bladder, either by suprapubic needle aspiration of urine from the bladder, which is invasive and seldom performed in clinical practice, or by urethral catheterization, which carries a 1 to 2 per cent risk of introducing infection into the bladder. In men, contamination of the voided urine sample can largely be avoided by retraction of the foreskin prior to voiding. In women, the reliability of urine culture can be improved by instructing women to part the labia with one hand and ensuring collection of a midstream sample, without either the initial portion or the 'afterdrip', but is not improved further by perineal washing or antiseptic use. These precautions only reduce the risk of contamination, rather than abolishing it altogether.

Microscopy of urine samples allows quantification of pyuria—the presence of white blood cells in the urine. However, the methodology used to report pyuria varies enormously: microscopy of urine that has been centrifuged and resuspended, with reporting of the number of cells per high-power field, gives results which bear little relation to leucocyte excretion rate or to counting cells from unspun urine in a counting chamber, when significant pyuria is usually defined as a urinary white cell count of 10 leucocytes/μl or more.

Bacterial urinary tract infection is by far the commonest cause of pyuria, and symptomatic patients with pyuria whose urine cultures are reported as showing no significant pathogens should be suspected either of having 'low-count' bacteriuria due to early infection or infection with a slow-growing organism, chlamydial infection, or one of the causes of sterile pyuria (Table 4). However, vaginal leucorrhoea can also result in 'false-positive' pyuria.

Microscopy also gives information on whether the urine sample is contaminated by cells from the periurethral area. Squamous cells are five to seven times larger than red cells and are easily recognized on microscopy: their presence in a midstream urine sample has conventionally been taken to indicate contamination, but they may originate from the urethra as well as from the epithelium of the vulva and vagina, as well as from areas of squamous metaplasia in the bladder, which is a common finding; and squamous cells are frequently seen in urine obtained by bladder catheterization, showing that their presence is not an absolute indicator of contamination.

Once a urine sample is obtained, the conditions under which it is cultured determine whether any organisms present grow. Standard laboratory culture conditions are designed to encourage the growth of recognized urinary pathogens (if present), but may not be optimal for the growth of atypical organisms or of those not usually recognized as urinary pathogens. Because small numbers of organisms are frequently cultured from urine as a result of contamination, growth of an organism is conventionally reported as a 'significant growth' if it meets several criteria:

1. there is a pure growth, i.e. of a single organism;
2. the organism grown is a 'recognized' urinary pathogen;
3. quantitative urine culture results in greater than 10^5 colony-forming units per millilitre (**cfu/ml**); and
4. there is significant pyuria on urine microscopy, and few if any squamous cells.

However, there are important exceptions to these criteria.

1. Genuine mixed growth of two or more bacteria may occur in complicated urinary tract infection (Table 5), as may the growth of an organism not usually associated with the urinary tract.
2. The spectrum of organisms recognized as capable of causing genuine urinary tract infection is widening. *Staphylococcus saprophyticus* was only fairly recently recognized as a cause of urinary tract infection in sexually active women, and it is possible that other true urinary pathogens are yet to be identified, perhaps accounting for some cases of the so-called urethral syndrome (see below).
3. 'Low-count' bacteriuria may reflect genuine bladder infection, particularly in early urinary tract infections, and may occur in patients who have increased their fluid intake and are 'diluting' their bacterial counts by generating a high urine output; also in patients infected with slow-growing organisms such as *Staph. saprophyticus*. The criterion of 10^5 cfu/ml was originally validated in asymptomatic women, but subsequent studies showed that nearly 50 per cent of women presenting with frequency and dysuria had genuine bladder infection but with counts between 10^2 and 10^5 cfu/ml on culture of a midstream urine sample. If symptomatic women with counts of between 10^2 and 10^4 cfu/ml are left untreated, most have persistent symptoms and counts of more than 10^5 cfu/ml 2 days later.
4. In men, bacterial counts of 10^3 cfu/ml or more are very likely to reflect significant infection, as the potential for significant contamination is lower.
5. The presence of pyuria further increases the likelihood that low counts are significant, although pyuria is not always present in proven bladder infection, particularly if the sample is taken early after the onset. The traditional method of expressing urinary white cell counts as cells per high-power field is very poorly reproducible as the volume in a high-power field is extremely variable. If a counting chamber or equivalent is used, then a criterion of 10 white cells/mm^3 separates patients with genuine bacteriuria from those without.

Localization to upper or lower urinary tract

It is sometimes important to discover whether infection is confined to the bladder or whether it has spread to involve one or both kidneys. The 'gold standard' for diagnosis of upper urinary tract infection is culture of urine obtained from each ureter by direct catheterization during cystoscopy, but such an invasive procedure can only be justified in exceptional circumstances, and even then may be difficult to interpret due to contamination of ureteric samples by bladder urine during passage of the catheters. An alternative, the Fairley test, involves a bladder washout using neomycin and fibrinolytic enzymes. Urine is cultured following completion, to confirm eradication of bladder bacteria, and then at 10, 20, and 30 min after completion of the washout. Bacteriuria returns slowly, if at all, in patients with infection confined to the bladder, but because the washout procedure has no effect on bacteria in the upper urinary tracts, rapid reappearance of bacteriuria indicates upper urinary tract infection. Using this test it has been shown that both upper and lower urinary tract infection are frequently asymptomatic and that flank pain and fever are extremely unreliable indicators of the presence of upper urinary tract infection.

Detection by immunofluorescent staining of immunoglobulin-coated bacteria is suggestive of tissue invasion, and has been advocated as a test to distinguish upper from lower urinary tract infection. However, compared with ureteric catheterization or Fairley tests, tests for antibody-coated bacteria are not reliable. This may be partly because tissue invasion can also occur in severe lower urinary tract infection, such as that complicating urethral catheterization. Antibody-coated bacteria may indicate a higher risk of treatment failure after standard antibiotic courses, but this remains uncertain.

Tubular damage due to ascending infection may be detected by measurement of urinary b_2-microglobulin, although this is also raised in patients with chronic renal disease and is therefore not specific for acute pyelonephritis. Renal excretory function usually remains unchanged during acute pyelonephritis unless obstruction is present, but acute renal failure is occasionally seen, often associated with coincident use of non-steroidal anti-inflammatory drugs. Abnormal appearances on contrast CT scanning and/or dimercaptosuccinyl acid (**DMSA**) scanning have been reported, including generalized renal swelling, focal areas of decreased parenchymal enhancement, and perirenal abscess formation, with the development of cortical scars and calyceal diverticulas if imaging is repeated on follow-up. In general, the more severe the infection is clinically (assessed by acute-phase response, duration of fever, etc.), the more marked the scarring. However, significant loss of renal excretory function following acute pyelonephritis in patients without diabetes, obstruction, or pre-existing reflux nephropathy/dysplasia is remarkably uncommon, and the significance of such scars is therefore uncertain.

Culture-negative syndromes

Occasionally patients may present with symptoms and signs highly suggestive of urinary tract infection, with or without pyuria, but with negative urine cultures. These

patients may have 'false-negative' urine cultures, for instance a low growth of a genuine pathogen; infection with a 'fastidious' organism, the presence of which is not detected by routine laboratory cultures; or may have a non-infectious cause. It is dangerous to label symptoms in such patients as psychogenic without careful thought and investigation; prolonged symptoms combined with numerous unsuccessful trials of antibacterials or with different explanations from different doctors may result in psychological stress, which in turn may amplify symptoms, whereas there is little evidence that psychological disease is the primary problem even in a subgroup.

'Urethral syndrome'

The term 'urethral syndrome' was used in the past as a synonym for the typical symptoms of cystitis, namely frequency, urgency, and dysuria. More recently it has been applied to the subgroup of women with typical symptoms but in whom a recognized urinary pathogen cannot be cultured from the urine. A significant proportion of these patients, particularly those with pyuria, have chlamydial urethritis, which can be diagnosed by urethral swab or by detection of chlamydial antigens in a first-pass urine sample. Chlamydial infection can be treated with tetracyclines, but as the infection may be sexually transmitted it is also important to treat the patient's sexual partner(s), who may be asymptomatic. Other patients have 'low-count' infection with a true bacterial urinary pathogen. Vaginal infection or atrophy should be excluded, as these can cause similar symptoms.

The pathogenesis and optimal management of the remaining patients with frequency and dysuria with no identifiable bacterial infection remains controversial. There is controversy over the role of 'fastidious bacteria' that are difficult to grow in the laboratory, particularly lactobacilli. Empirical antiobiotic treatment is equally successful in eradicating symptoms in women presenting to primary care whether or not urinary pathogens are found on urine culture, suggesting that the syndrome is frequently due to bacterial infection that is not detected by routine laboratory urine culture. However, a few women with persistent symptoms do not respond to antibiotics, and in these women repeated courses of antibiotics are likely to lead to the emergence of antiobiotic-resistant organisms, which may later cause true infection that is difficult to treat. Psychological distress is common in patients with persistent lower urinary tract symptoms, but the prevalence of emotional or psychiatric disorders is no higher in women presenting to general practitioners with dysuria and frequency whose urine cultures are negative than in those with proven cystitis. Urologists often offer such women urethral dilatation on the assumption that the symptoms are due to urethral spasm or stricture, but there is minimal evidence beyond clinical anecdote that this procedure is of any benefit; one randomized, controlled trial showed no difference in outcome between urethral dilatation and cystoscopy alone. Urethral dilatation may itself cause periurethral fibrosis, resulting in the later formation of genuine urethral strictures.

Women with recurrent episodes of frequency and dysuria, with or without pyuria, whose urine cultures remain sterile should be carefully evaluated for the presence of vaginitis (either infective or atrophic) and for sexually acquired urethritis (where relevant). It is justified in this situation to obtain urine direct from the bladder during an episode, preferably by suprapubic aspiration or alternatively by urethral catheterization, and ensure that this is cultured in conditions permitting the identification of fastidious or low-growing organisms. In urine obtained direct from the bladder, any growth of organisms is clinically significant. Any infection so detected should be treated, preferably with a prolonged course of an appropriate antibiotic to ensure complete eradication. If no infection can be detected, cystoscopy is required to exclude non-infective causes of cystitis. Patients should be treated with compassion and their symptoms believed: attributing the symptoms to psychiatric disease is almost certain to be incorrect and likely to alienate the patient.

Interstitial cystitis

Interstitial cystitis causes chronic suprapubic pain, dysuria, and frequency despite, by definition, sterile urine. Urine microscopy shows pyuria. Cystoscopy shows variable inflammation, sometimes with ulceration. Bladder biopsies show a chronic inflammatory infiltrate; mast cell infiltration is common, but is also seen in infective cystitis. The condition may progress to cause contracture of the bladder. Many of the features would be explained by an acquired defect in the barrier function of the uroepithelium, but the cause of such a defect remains unclear. It remains possible that infection by a fastidious organism is responsible for initiating the disease in some patients. Numerous therapies have been tried, including intravesical instillation of glycosoaminoglycans.

Drug-induced cystitis

This presents similarly, although often more acutely and with macroscopic haematuria. It may be caused by acrolein, a metabolite of cyclophosphamide and ifosfamide, and also by non-steroidal anti-inflammatory drugs, particularly tiaprofenic acid, and by danazol.

Investigation of patients with urinary tract infection

Most women with uncomplicated cystitis do not require investigation other than urine culture, and may even be treated empirically, the choice of antibiotic being based on locally prevalent sensitivity patterns of the most common uropathogens, rather than waiting for the results of culture and sensitivity. The yield in such women of investigation with cystoscopy and/or intravenous urography is low. Because minor abnormalities such as duplex collecting systems are common in the general population, these will often be found in women presenting with cystitis, but detection of such abnormalities does not lead to any change in treatment. Investigation of women should therefore be reserved for those with atypical features (Table 6). In men, urinary tract infection is nearly always associated with an underlying abnormality of host defence, and all men with proven urinary tract infection should therefore be offered investigation. If investigation in either sex is thought necessary, the important abnormalities being looked for are:

- diabetes
- urinary tract stones
- anatomical abnormalities of the upper urinary tract (e.g. papillary necrosis, reflux nephropathy)
- urinary tract obstruction (anywhere from the renal pelvis to the tip of the urethra)
- bladder diverticulas
- impaired bladder emptying.

The choice of investigation of the upper urinary tract depends on local facilities, and usually lies between plain abdominal radiography with ultrasound, on the one hand, and intravenous urography on the other. Both may need to be supplemented by cystoscopy and by bladder voiding studies, using urinary flow rate and measurement of pre- and post-void bladder volumes.

Whether adults should be investigated for vesicoureteric reflux is open to doubt, as there is no good evidence that antireflux surgery (e.g. ureteric reimplantation, injection of Teflon around the ureteric orifice) is of benefit in preventing either ascending infection or renal damage.

Treatment of uncomplicated 'cystitis'

Rational treatment of urinary tract infection requires the physician to balance the costs and dangers of treatment (including cost of the drug, risk of unwanted side-effects, and the induction of resistance) with benefit.

Is treatment necessary at all? Many women with recurrent uncomplicated cystitis report that they can clear their own infections by increased fluid intake and frequent voiding. Many buy alkalinizing agents (e.g. potassium citrate) to ameliorate the symptoms, which work by reducing bladder irritability. Placebo-controlled studies have confirmed that infection may clear spontaneously, although this may take several weeks or even months, and a small percentage of women remain infected until given antibiotics. There is therefore no justification in insisting on antibiotic treatment in those who wish to try to do without.

Choice of antibiotic

It is usually impracticable to await the results of culture and sensitivity testing, if these tests are justified at all. The choice of antibiotic is therefore usually empirical, based on the likelihood that the drug will clear the infection (efficacy), cost, side-effect profile, and the risk of selection of resistant organisms, both in the patient being treated and in the community. The efficacy of antibiotics is not fully predictable from *in vitro* sensitivity testing, which is probably part of the reason why trimethoprim (with or without sulphamethoxazole) remains the first-line choice in many areas, despite resistance rates on *in vitro* testing of up to 20 per cent. This is at least in part because many antibiotics are concentrated in the urine to levels far greater than those found in tissues, and at these concentrations may remain active against organisms that are reported to be resistant to the concentrations found in tissues, which are usually used to define resistance *in vitro*. However, increasing resistance *in vitro* to trimethoprim is sure to lead sooner or later to increased clinical failure rates, as has already been observed for b-lactam antibiotics. Some of the properties of the most commonly used antibiotics are reviewed in Table 7. The most recent recommendations of the Infectious Diseases Society of America are summarized in

Table 8.

Duration of treatment

A single high dose of an antibiotic will cure many women and is simple, cheap, and may reduce the risk of side-effects and bacterial resistance. Single-dose treatment is thought to be popular amongst patients, although those paying for prescription medications may feel 'short-changed' by paying a full prescription fee for a single tablet. However, cure rates after single-dose treatment are lower than for longer courses of antibiotics, this difference being more marked for b-lactam antibiotics, which in general need to be given for 7 days, than for trimethoprim (with or without sulphonamide) and for quinolones. Fosfomycin given as a single dose gives higher cure rates than single doses of other antibiotics, but is less effective and more likely to cause adverse effects than 3-day courses of trimethoprim (with or without sulphonamide) or quinolones. The rate of adverse reactions also increases with duration of therapy, particularly for trimethoprim–sulphonamide combinations, which should therefore be given for no more than 3 days. Cure rates of urinary tract infections caused by *Staph. saprophyticus* and in elderly women are low with 3-day regimens, and these infections should be treated with 7-day courses.

Alternatives to antibiotic therapy

Fructose, present in many fruit juices, inhibits binding of type 1 fimbrias of *E. coli* to the uroepithelium. Cranberry juice also contains proanthocyanidin, which inhibits adherence of P-fimbriated *E. coli*. Cranberry juice is popular as an alternative treatment for urinary tract infection, but the evidence for its effectiveness in clinical practice remains uncertain.

Treatment of uncomplicated 'acute pyelonephritis'

Choice of antibiotic

The antibiotic chosen in this situation needs good tissue penetration as well as high urinary excretion, and must be fully active against the infecting organism at typical serum concentrations. It is therefore much more important to identify the infecting organism and its antibiotic sensitivity pattern by sending urine (or blood from patients in hospital) for culture. However, empirical treatment must be started while awaiting culture and sensitivity results, as acute pyelonephritis can evolve rapidly into a life-threatening illness. Oral therapy with a quinolone antibiotic (ciprofloxacin, ofloxacin, norfloxacin) is probably the best choice, although trimethoprim or trimethoprim–sulphamethoxazole are alternatives if local rates of resistance among uropathogens remain low. Treatment with b-lactam antibiotics, even if the infecting organism is fully sensitive *in vitro*, is associated with a high rate of recurrence compared with treatment by other agents. Patients with septicaemia should receive a quinolone (for which oral administration is as effective as intravenous) or a combination of an aminoglycoside with ampicillin plus b-lactamase inhibitor, or an extended-spectrum cephalosporin with or without an aminoglycoside. Once-daily administration of aminoglycosides is as effective as thrice-daily and reduces the risk of toxicity.

Duration of therapy

It is widely recommended that acute pyelonephritis is treated with a significantly longer course of antibiotics than acute cystitis. Since, as discussed above, the clinical distinction between acute pyelonephritis and cystitis relies on the presence of fever, flank pain, and an acute-phase response, and since all of these may be present in acute cystitis with no involvement of the upper urinary tract and entirely absent in acute pyelonephritis, what these recommendations really mean is that those patients demonstrating a more marked host response should be treated more aggressively and for longer. It is therefore reasonable to suggest that in a patient with systemic symptoms (including flank pain), fever, or leucocytosis, or a raised C-reactive protein, plasma viscosity, or erythrocyte sedimentation rate, antibiotic treatment should be continued until these abnormalities have disappeared. A 14-day course is as effective as a 6-week course in uncomplicated acute pyelonephritis, and a 7-day course may be sufficient for patients with mild illness.

Treatment of asymptomatic bacteriuria

The only situation in which treatment of asymptomatic bacteriuria is mandatory is during pregnancy (see Chapter 13.5). Eradication of asymptomatic infection in children with or without proven vesicoureteric reflux is widely practised in the hope that this will prevent ascending infection and renal damage. However, prophylactic treatment for 2 years of covert bacteriuria in schoolgirls without renal scarring has no effect on glomerular filtration rate at age 18, but is associated with lower fractional reabsorption of glucose and with a smaller increment in glomerular filtration rate and greater degrees of glycosuria during subsequent pregnancy. Screening for asymptomatic bacteriuria with the aim of preventing these minor abnormalities is not currently thought justified. Treatment of asymptomatic bacteriuria in patients with anatomically abnormal urinary tracts or with indwelling urinary catheters is unjustified and is likely only to lead to the emergence of antibiotic-resistance urinary infection. Treatment of asymptomatic bacteriuria in the elderly has been shown to be of no benefit.

Prophylaxis of recurrent urinary tract infection

Some women with recurrent cystitis choose to have antibiotic treatment for each infection as it arises, particularly if they are allowed to self-administer treatment as soon as symptoms start. Others may opt for prophylactic treatment. Long-term low-dose antibiotic treatment has been shown to be effective in reducing the rate of infection in such women, but no regimen offers 100 per cent protection. Prophylactic treatment should be considered in women with at least two symptomatic infections per year and probably works by preventing colonization of periurethral tissues by uropathogens. Trimethoprim (100 mg at night) is widely used for prophylaxis because it achieves very high concentrations in vaginal fluid and may therefore remain active against organisms that are resistant to the concentrations used in *in vitro* sensitivity testing. Nitrofurantoin (100 mg at night) has also been widely used, and may be more effective, but can cause rare but serious adverse effects (pulmonary and hepatic toxicity) with long-term therapy, making regular monitoring of liver enzymes and lung function tests necessary. Because both are well absorbed they do not reach high concentrations in the colon, hence emergence of resistant strains in colonic flora is uncommon, whereas this problem does arise with long-term use of b-lactam antibiotics. Long-term use of quinolones is expensive and associated with a significant risk of selection of resistant strains. There is no proven advantage in 'rotating' antibiotic prophylaxis. A number of dosage regimens have been used, including nightly treatment, thrice-weekly treatment, and postcoital treatment, with no convincing evidence of the superiority of one regimen over another. Treatment should be continued for at least 6 months, because, for reasons that are not clear, this results in a lower relapse rate once treatment is stopped than shorter periods of prophylaxis.

Cranberry juice, as discussed above, contains substances that inhibit adherence of uropathogenic bacteria to the uroepithelium, and has become popular as an alternative treatment to prevent recurrent urinary tract infection. However, the evidence that regular use of cranberry juice reduces the risk of recurrent symptomatic infections remains poor, and there is no consensus on what dosage and regimen, if any, is effective.

Complicated urinary tract infection

'Complicated' urinary tract infections are those occurring in a patient with abnormal host defence, and as a result are often more severe.

Urinary tract infection in men

Cystitis

In the first year of life, urinary tract infection is commoner amongst boys than girls; circumcision reduces the risk. Urinary tract infection in men is uncommon, as the length of the urethra and the fact that the penile mucosa is seldom colonized with faecal organisms including uropathogens confer major protection against ascending infection. The occurrence of urinary tract infection in a man therefore suggests an abnormality of host defence, which may predispose to more severe infection and should be investigated unless the cause is immediately obvious (e.g. the presence of a urinary catheter). Risk factors that may be identified by investigation include:

- bacterial prostatitis and prostatic calcification
- lack of circumcision
- impaired bladder emptying (particularly if this has resulted in bladder catheterization or instrumentation)
- anal intercourse
- urinary tract stones

- reflux nephropathy.

The symptoms of urinary tract infection in men are similar to those in women. The risk of bacterial contamination of voided urine is low apart from in elderly men with foreskins, in whom precautions should be taken to minimize contamination by retraction of the foreskin and collection of a midstream sample. Colony counts of 10^3/ml or higher usually indicate significant infection.

Acute prostatitis

Acute bacterial prostatitis causes fever, rigors, backache, and dysuria, and may result in acute urinary retention. Symptoms and signs of epididymitis may also be present. Rectal examination reveals an enlarged, tender prostate. Untreated, acute prostatitis may culminate in prostatic abscess formation. The causative organism can be identified on urine culture: an antibiotic which has good tissue penetration (e.g. trimethoprim, a tetracycline, or a quinolone) should be used and continued for 4 weeks, as it is thought that this reduces the risk of chronic prostatitis.

Chronic bacterial prostatitis

This is an uncommon syndrome caused by the persistence of a uropathogen within the prostate, with repeated episodes of acute infection caused by the same organism on each occasion, and few if any symptoms between episodes. Obtaining bacteriological proof that the infecting organism is 'hiding' in the prostate gland between acute episodes is difficult. The 'textbook' method described by Stamey and Mears involves culture of four specimens obtained during voiding of the bladder. The first 10 ml voided and a midstream sample are collected. The patient then interrupts the flow of urine, bends forward, and digital prostatic massage is performed, resulting (sometimes) in the collection of a few drops of 'expressed prostatic secretions'. Finally, voiding is completed and a fourth sample collected. Prostatitis is diagnosed when bacterial counts are highest in the expressed prostatic secretions and the final voided urine sample; urethritis, by contrast, results in high counts in the first sample. Due to its complexity and the unpleasantness of performing digital prostatic massage *per rectum* during interrupted micturition, this test is very rarely performed in practice, and many patients are treated with a prolonged course of a quinolone antibiotic. a-Blockers have been shown to reduce recurrence rate, possibly by reducing reflux of urine into prostatic ducts during micturition.

Culture-negative pelvic pain in men

Patients may complain of chronic pelvic pain, dysuria, strangury, urinary frequency, and pain during sexual intercourse but have no evidence of bacterial infection on cultures of prostatic secretions, semen, or post-massage urine specimens. Patients with this symptom complex may be further subclassified as having chronic abacterial prostatitis or non-inflammatory pelvic pain syndrome according to the presence or absence of leucocytes in semen. There is no gold standard for diagnosis, no clear understanding of the pathophysiology, no correlation between symptoms and prostatic histology, and no satisfactory treatment for this ill-understood group of conditions. Occasionally, patients are found to have evidence of prostatic inflammation on biopsy, or to have leucocytes in prostatic fluid in the absence of symptoms. As in the urethral syndrome in women, some cases may be caused by persistent infection by fastidious bacteria, such as *Chlamydia* or *Mycoplasma*; a prolonged trial of a tetracycline is therefore often used. Other treatments include regular prostatic massage, non-steroidal anti-inflammatory drugs, a-blockers, and 5-a reductase inhibitors. a-Blockers have been shown to be of some benefit in all types of symptomatic chronic prostatitis in one randomized study.

Urethral catheterization

P>Urinary tract infection occurs after 2 per cent of in/out urethral catheterizations, after 10 to 30 per cent of 5-day indwelling catheterization, and is nearly inevitable in patients with long-term indwelling catheters. It is an important cause of hospital-acquired infection, increasing the risk of Gram-negative septicaemia fivefold and carrying a threefold increase in mortality after adjustment for age, severity and type of underlying illness, duration of catheterization, and renal function. Organisms enter the bladder either by migration between the catheter and the urethral mucosa or by ascent up the column of urine in the lumen after entry into the drainage system following contamination at disconnection or drainage points. Although most infections are probably caused by ascent of the patient's own faecal flora, there is evidence from investigation of clusters of infections by highly antibiotic-resistant organisms that inadequate handwashing by hospital staff may also cause some infections. A sample obtained directly from the catheter (not from the drainage bag) represents bladder urine, when any bacterial growth should be considered as evidence of urinary tract infection; low-count infection (e.g. less than 10^2 cfu/ml) usually progresses within days to higher counts. Mixed growths are common in patients with long-term catheterization and may be associated with mixed-growth bacteraemia.

Risk factors for the acquisition of infection include increasing duration of catheterization, increasing age, female sex, renal impairment, diabetes mellitus, and the nature of the underlying illness. Use of prophylactic antibiotics is associated with a delay in the onset of infection and may be justified in high-risk patients requiring catheterization for 3 to 14 days, whereas in those with long-term catheters, antibiotic use simply increases the risk of emergence of antibiotic-resistant pathogens. Use of silver alloy-coated catheters also reduces the risk of infection and may be justified in high-risk patients. Progress is being made in the development of new catheter materials that may provide further resistance against colonization by micro-organisms.

Urethral catheters should not be inserted unless absolutely necessary (is knowledge of hourly urine output really going to change your management?) and removed as soon as they are no longer needed. Consideration should always be given to methods of urine collection that may carry lower risks, such as condom drainage in men, intermittent catheterization in patients with abnormal bladder empyting, and suprapubic catheterization.

Abnormal bladder emptying

Incomplete bladder emptying, removing the 'washout' part of host defence, greatly increases the risk of urinary tract infection, as in patients with prostatic bladder outflow obstruction and those with neurogenic bladder due to spinal cord injury. Long-term catheterization only increases these risks. Where possible, the cause of incomplete bladder emptying should be treated. However, patients shown on urodynamic study to have underactive detrusor activity will not benefit from prostatectomy or a-blockade and may require long-term intermittent self-catheterization. Bladder dysfunction in patients with neurogenic bladder, for instance due to spina bifida or spinal cord injury, depends on the level of injury. Patients with lesions above T-11 have hyperreflexic bladder activity, often with sphincter dyssynergia (failure of the sphincter to relax during detrusor contraction), resulting in a high-pressure system, often with high-pressure reflux, combined with impaired emptying. In combination with urinary tract infection, this frequently results in progressive renal damage. Those with lesions below L-1 have decreased detrusor activity with large amounts of residual urine, which also increases the risk of urinary tract infection. Diabetic neuropathy may also cause decreased detrusor activity. The aim of treatment in both situations is to achieve a low-pressure bladder with low residual volumes. This may involve teaching patients to utilize reflexes to induce bladder contraction and sphincter relaxation, condom drainage for incontinence, anticholinergics to reduce detrusor overactivity, sphincterotomy, augmentation cystoplasty, and intermittent self-catheterization. Urethral catheterization should be avoided wherever possible. There is no evidence that regular use of antiseptics to wash the perineum and urethral meatus are of benefit. Bladder washouts with saline or boiled water may be of benefit in eliminating mucus in patients with augmentation cystoplasties: antiseptic bladder washouts are of minimal value in prevention, probably due to the fact that uropathogens become embedded in a biofilm adherent to the bladder wall. Methenamine, a drug that releases formaldehyde into acidic urine, may be of some benefit in preventing infection.

Treatment of urinary tract infection in patients with abnormal bladder emptying should be reserved for those with evidence of invasive infection. The diagnosis is obvious in those with cloudy urine combined with fever, rigors, and flank pain, but it is important to remember that symptoms and signs, particularly flank pain, dysuria, urgency, and frequency may be absent in those with neurological dysfunction.

Urinary diversion

Ileal or colonic conduits have been used for many years in patients requiring cystectomy for malignancy, and occasionally (although increasingly less frequently) for non-malignant conditions such as neurogenic bladder. Such conduits are frequently complicated by urine infection as the bowel mucosa and the mucus it produces readily permits adherence of uropathogens. Upper urinary tract dilatation is common, irrespective of whether the ureteric anastomoses are designed to be non-refluxing or not, and there is a high incidence of recurrent 'acute pyelonephritis' with flank pain, fever, and rigors. Diagnosis of urinary tract infection in patients with a conduit requires insertion of a catheter to the far end of the conduit and collection of urine via the catheter, rather than culture of urine collected from the conduit bag. Preventive measures include ensuring that the ileal segment is as short as possible at the time of surgery and ensuring a high fluid intake. The belief that cranberry juice reduces the incidence of urinary tract infection by reducing bacterial adherence is as yet unproven, although it seems likely that treatments designed to interfere with bacterial adherence or with mucin production are more likely than antibiotic treatment to help prevent symptomatic infection in these patients.

Renal tract stones

Renal tract stones are an important cause of persistent or relapsing urinary tract infection, as they provide a 'hiding place' in which organisms are protected from antibiotics. Management of such patients is complicated, as it may be impossible to eradicate infection without aggressive stone management (which may involve extracorporeal shock-wave lithotripsy, percutaneous and ureteroscopic stone removal). Attempts at stone removal may be complicated by septicaemia unless combined with antibiotic treatment, yet prolonged antibiotic therapy may encourage the emergence of resistance in the infecting organism.

Infection stones are caused by chronic infection with urease-producing organisms, usually *Proteus mirabilis*, and account for around 5 per cent of urinary tract stones. These stones are made of 'struvite' ($MgNH_4PO_4.6H_2O$), which forms as a result of the action of the alkaline pH caused by the production of ammonium and hydroxyl ions from the breakdown of urea by urease. Pure struvite stones may result from *de novo* urinary tract infection by a urease-producing organism, and are commoner in women and, probably, in patients with pre-existing anatomical abnormalities of the upper urinary tract such as reflux nephropathy, pelviureteric junction obstruction, or urinary diversion. Sruvite stones may also form as a secondary complication of metabolic stones. Struvite stones often expand to fill the entire renal pelvis, forming 'staghorn' calculi, but such calculi should not be assumed to be due to infection (rather than a metabolic cause) without demonstration of chronic infection by a urease-producing organism and/or biochemical analysis showing that the stone is made of struvite. The usual presentation is with symptomatic 'acute pyelonephritis' and alkaline urine; renal colic is unusual due to the large size of the stones. Treatment is with a combination of antibiotic and stone removal, which is imperative to prevent stone recurrence, and may require a combination of extracorporeal shock-wave lithotripsy and percutaneous nephrolithotomy, aided by dissolution therapy for larger stones. Urease inhibitors (acetohydroxamic acid, propionhydroxamic acid) may reduce stone recurrence but are too toxic for clinical use. See Chapter 20.13 for further discussion.

Encrusted cystitis and pyelitis occur as a result of chronic infection by urease-producing organisms, including *Corynebacterium* spp., in immunosuppressed patients, causing deposition of struvite in the bladder wall.

Autosomal dominant polycystic kidney disease

Cystitis is common in women with polycystic kidney disease, and in 20 per cent it is the presenting clinical finding, but there is no evidence that host defence in the lower urinary tract is abnormal. However, the risk of upper urinary tract infection is increased, and its diagnosis and treatment complicated. Acute parenchymal infection presents as acute pyelonephritis with flank pain, fever, and infected bladder urine, and usually responds to conventional therapy. Infection of cysts is more difficult to diagnose: the urine may be sterile and there may be no pyuria if the infected cyst does not communicate with the urinary space. Presentation is with fever and a discrete area of tenderness in the affected kidney. Blood cultures are the most reliable way of making a bacteriological diagnosis. Imaging studies, looking for cysts with increased fluid density, septations, and thick walls, are seldom conclusive, as similar appearances may occur normally or after previous cyst haemorrhage. The spectrum of causative organisms suggests that ascending infection rather than haematogenous spread is the usual route of infection. Hydrophilic antibiotics, including aminoglycosides and b-lactam antibiotics, penetrate poorly into those cysts which maintain large ionic gradients, whereas quinolones, trimethoprim–sulphamethox-azole, doxycycline, and clindamycin achieve better penetration. Prolonged courses of antibiotics are usually needed to eradicate infection, with surgical resection a last resort.

Renal transplantation

Urinary tract infection is the commonest bacterial infection after renal transplantation. Risk factors include urethral catheterization in the early postoperative period, the use of ureteric stents, pre-existing abnormalities of bladder emptying (such as diabetic autonomic neuropathy, previous bladder outflow obstruction, small contracted bladders in anuric patients on dialysis), anatomical abnormalities in the upper urinary tract (such as reflux nephropathy), contamination of the transplanted organ during retrieval and storage, abnormal drainage of urine from the transplanted kidney, vesicoureteric reflux into the transplant, areas of renal infarction, and immunosuppression. The commonest causative bacteria are those found in the general population with urinary tract infection, but many organisms not usually considered as urinary tract pathogens may also cause significant infection in these patients. Many infections are asymptomatic. Prophylactic antibiotics may reduce the early postoperative risk and many centres use co-trimoxazole as it also reduces the risk of *Pneumocystis* pneumonia. Antibiotic treatment must be chosen with care because of the risk of interactions with immunosuppressive treatment and of nephrotoxicity.

Infections with the polyomaviruses BK virus and JC virus may cause cystitis, ureteric stenoses, and interstitial nephritis (easily mistaken for acute rejection) in renal transplant recipients. The diagnosis may be suggested by recognition of infected transitional uroepithelial cells on urine cytology or by histological recognition of inclusion bodies on renal biopsy. Treatment is by reduction of immunosuppression, but this is often complicated by further rejection.

Pregnancy

Asymptomatic bacteriuria early in pregnancy is associated with the development of acute pyelonephritis in up to 30 per cent of patients if left untreated. It is commoner in women of lower socio-economic status and is associated with an increased incidence of preterm delivery and low birth weight, particularly if the pregnancy is complicated by acute pyelonephritis towards term. The increased risk of pyelonephritis is attributed to ureteric dilatation caused primarily by progesterone-induced smooth muscle relaxation. Antibiotic treatment of infection reduces the risk of acute pyelonephritis and of preterm delivery and low birth weight. Similar benefit is seen from a short course of treatment and from continued antibiotic prophylaxis. (See Chapter 13.5 for further discussion.)

Reflux nephropathy

As discussed above, vesicoureteric reflux (retrograde flow of urine up into the ureters and, in severe cases, as far as the renal pelvis) is often found in children with recurrent urinary tract infection. At the time of first diagnosis of urinary tract infection or subsequently, a small proportion of such children are found to have a characteristic pattern of renal parenchymal scarring at the upper and lower poles, with underlying clubbing and distortion of calyces. This pattern of scarring has become known by a variety of terms including 'reflux nephropathy' and 'chronic pyelonephritis'. Patients with reflux nephropathy have an increased risk of recurrent urinary tract infection, may develop infection stones, and a proportion develop hypertension, proteinuria, and progressive renal impairment with an inexorable progression to endstage renal failure. Under the age of 1 year, when only relatively severe cases come to clinical attention, slightly more boys than girls are affected; in older children the disease is diagnosed up to 5 times more frequently in girls, possibly because the disease is often discovered during investigation of urinary tract infection, which is commoner in females. Reflux nephropathy is commonly familial, best modelled by an autosomal dominant pattern of inheritance with variable penetrance. Linkage has been demonstrated to an area of chromosome 1 in some large pedigrees.

The diagnosis is conventionally made in adults by intravenous urography, which permits the detection both of focal parenchymal scarring and of the underlying calyceal abnormality (Fig. 2). Ultrasound scanning can show focal scarring but does allow visualization of the calyces. DMSA isotope scanning is the most sensitive test for the detection of parenchymal scars, and is widely used in children, as there are few alternative causes of focal scarring in this age group. Lateral displacement of the ureteric orifices can be demonstrated by Doppler ultrasound in most patients with reflux nephropathy. Demonstration of vesicoureteric reflux by direct or isotopic micturating cystography is commonly used to confirm the diagnosis in children, but is rarely justified in adults, as the absence of reflux could be due to spontaneous resolution of reflux with age (it often resolves in childhood), and its presence seldom justifies a change in clinical management. The histological appearances of 'chronic pyelonephritis' are well described and may occasionally be seen in patients with no scarring on urography or even DMSA scanning, probably because the scars are too small in these patients to be detected radiologically.

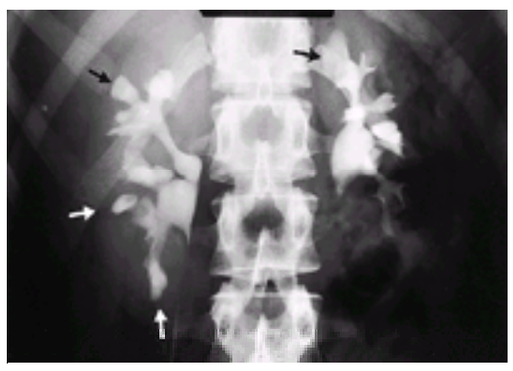

Fig. 2 Reflux nephropathy on intravenous urography, more marked on the right side than the left. Several focal scars (arrowed) involving the full thickness of the renal parenchyma and associated with calyceal clubbing are most obvious in the polar regions (reproduced from Bailey, 1993, with permission).

The conventional view is that reflux nephropathy is 'postinfectious focal renal scarring' and caused by the ascent of infected urine into the renal pelvis and then into the collecting ducts and renal parenchyma via compound papillas (papillas in which more than one collecting duct opens into the pelvis), which are found at the upper and lower poles but not in the middle calyces—explaining the polar distribution of scars. Sequential radiological imaging studies in children with urinary tract infections appear to support this theory, with the emergence of new scars up until the age of around 5 years, after which it is thought that maturation of the papillas prevents entry of infected urine into the renal parenchyma. Experimental infection in pigs causes a pattern of scarring very similar to that seen in human reflux nephropathy.

An alternative hypothesis is that at least a proportion of children with the radiological diagnosis of reflux nephropathy have congenital renal dysplasia, caused by abnormal nephrogenesis *in utero*, and associated with abnormal embryogenesis of the ureterovesical junction leading to vesicoureteric reflux. Vesicoureteric reflux is often found in the rare genetic syndromes that include renal dysplasia, and in non-syndromic renal dysplasia or aplasia, vesicoureteric reflux in the contralateral ureter is commonly seen. This theory would explain the presence of classic reflux nephropathy in neonates and in children with no documented history of urinary tract infection. Even the emergence of new scars during the first 5 years of life could be due to differential growth around areas of renal dysplasia. The rarity with which acute pyelonephritis in adults results in renal impairment, even in the presence of radiological evidence of scar formation, is perhaps further evidence that progressive loss of renal function is more likely to be due to 'remnant nephropathy' in dysplastic kidneys rather than the result of postinfectious scarring alone.

These two theories have different implications for the prevention of reflux nephropathy. Proponents of the 'postinfectious focal renal scarring' theory believe that diagnosis in infancy and treatment to prevent the ascent of infected urine into the renal pelvis until at least the age of 5 years should prevent the emergence of renal scarring and the later sequelae of hypertension, proteinuria, and progressive renal failure; by contrast, such treatment will not prevent these sequelae if reflux nephropathy is a disease of embryogenesis. Of course, the two theories are not mutually exclusive: in an individual patient reflux nephropathy may be due to the interaction of dysplasia and ascending infection during infancy. Antireflux surgery (ureteric reimplantation) and long-term prophylactic antibiotic treatment have been compared in several large randomized trials. Surgery is more effective at preventing episodes of acute pyelonephritis than medical treatment, but no other major differences in outcome were observed, and potential complications of antireflux surgery include ureteric obstruction, itself a potent cause of renal parenchymal damage.

The confusion over the pathogenesis of reflux nephropathy probably explains the lack of hard clinical evidence to guide the management of children with urinary tract infection. Current United Kingdom guidelines are summarized in Table 9.

Whatever the cause of reflux nephropathy, there is little doubt that women with it are more prone to recurrent acute pyelonephritis than those with anatomically normal upper urinary tracts, particularly during pregnancy.

Invasive/destructive renal parenchymal infection

As discussed above, ascending infection may cause the clinical syndrome of 'acute pyelonephritis' but seldom causes significant renal parenchymal damage. However, this is not the case if there is further impairment of host defence against infection, particularly by diabetes or urinary tract obstruction.

Acute papillary necrosis

This is an unusual complication of acute pyelonephritis, but more likely to occur in the elderly and especially those with diabetes. It should be suspected, as should urinary stones, in the patient with symptoms and signs of acute pyelonephritis who also has pain suggesting renal colic. This situation requires immediate imaging, usually with ultrasonography, to exclude urinary obstruction, and if obstruction is present then it must be relieved urgently, most often by antegrade nephrostomy.

The use of non-steroidal anti-inflammatory drugs is associated with an increased incidence of chronic renal papillary necrosis, perhaps because they compromise the renal medullary circulation. It therefore seems reasonable to say that these agents should be discontinued, at least temporarily, in the presence of acute pyelonephritis.

Renal carbuncle

Renal carbuncle is the formation of renal cortical abscesses, often only in one kidney, caused by bloodborne infection, usually associated with untreated *Staph. aureus* septicaemia. It is most commonly seen in intravenous drug abusers and patients with diabetes. There is usually a significant time delay between the initial infection and presentation with renal carbuncle, typically 6 to 8 weeks. Presenting symptoms include fever, malaise, and abdominal or flank pain, and are often non-specific. Because the infection is limited to the renal cortex and does not communicate with the collecting system, the urine is sterile and acellular. Blood cultures are usually negative. Radiological studies show a semisolid, thick-walled mass, percutaneous aspiration of which yields pus.

Pyonephrosis

Pyonephrosis is bacterial infection within a completely obstructed collecting system, for instance due to an obstructing ureteric stone. Patients usually present with fever, rigors, and flank pain, and have a marked neutrophilia and acute-phase response. Radiological differentiation from hydronephrosis relies on the presence of echogenic material and/or septas in the pelvicalyceal system, and confirmation is by percutaneous aspiration; as with other localized urinary tract infections, the voided bladder urine may be sterile. Untreated pyonephrosis rapidly results in complete destruction of the renal parenchyma, followed by death from complications of sepsis if nephrectomy is not performed; correction of obstruction and aggressive intravenous antibiotic therapy may prevent this if instituted soon enough.

Perinephric abscess

Perinephric abscess may complicate renal carbuncle or, more commonly, acute pyelonephritis—particularly if complicated by an anatomical or functional abnormality of the urinary tract. Typical presenting symptoms are those of acute pyelonephritis, with flank pain, fever, and rigors. If the abscess does not communicate with the collecting system, for instance in abscesses caused by haematogenous spread or complicating obstruction or renal cysts, there may be no lower urinary tract symptoms, no pyuria, and the urine may be sterile. Response to antibiotic treatment is much less rapid than in patients with uncomplicated acute pyelonephritis. Diagnosis is by ultrasound, urography, or CT scanning, followed by percutaneous (or occasionally surgical) aspiration, drainage, and culture of the aspirate. Prolonged antibiotic treatment of the organism identified is needed, stopping only when there is evidence that the infection has resolved, based on defervescence, resolution of the acute-phase response, and repeated radiological studies. This may take as long as 8 weeks.

Xanthogranulomatous pyelonephritis

Xanthogranulomatous pyelonephritis is an atypical form of chronic infection of the renal parenchyma in which bacterial infection, usually in the presence of obstruction or staghorn calculi, results in formation of granulomas with the accumulation of lipid-rich foamy macrophages. The process may be multifocal and can be complicated by extension into the perinephric fat, causing perinephric abscess. Patients are typically febrile and ill, with a history of progressive weight loss, anaemia, and malaise, without lower urinary tract symptoms, and have a mass in the flank on examination. Radiologically, the multifocal mass crossing tissue planes may be indistinguishable from a renal cell carcinoma, which may also cause systemic symptoms such as fever, anaemia, and weight loss. Although both require surgical excision, radical surgery can be avoided if the diagnosis is made preoperatively.

Emphysematous pyelonephritis

Emphysematous pyelonephritis is a rare and life-threatening form of acute pyelonephritis in which there is tissue necrosis together with formation of hydrogen and

carbon dioxide, which accumulate in pockets in the renal parenchyma, perinephric space, and collecting systems—'gas gangrene of the kidney' (Fig. 3). The typical patient is an obese, elderly woman with type 2 diabetes; urinary tract obstruction is another important risk factor. Presentation is with fever, vomiting, and abdominal pain. The patient is often extremely ill with hypotension, neutrophilia, and renal impairment. The commonest causative organism is *E. coli*; clostridial infection has not been reported. Even with aggressive medical treatment the mortality is high, and although occasional successes with antibiotics combined with percutaneous drainage have been reported, the standard treatment is nephrectomy.

Fig. 3 Gas-forming infection, seen as the three black holes in the single remaining (right) kidney of a patient with diabetes. The left kidney had been removed 2 years earlier for a similar gas-forming infection. This infection was successfully treated by intravenous antibiotics and percutaneous drainage.

Malakoplakia

Malakoplakia (Greek: 'soft plaque') is a rare disease characterized by destructive tumour-like granulomatous infiltrates in the urinary bladder, kidneys, and occasionally other organs. Bladder involvement usually presents with haematuria, frequency, and dysuria; renal involvement presents with fever, flank pain, and renal enlargement, and may frequently be bilateral. The diagnosis may be suspected at cystoscopy or on renal imaging, but is confirmed histologically by detection of large eosinophilic granular macrophages containing characteristic intracellular lamellated 5- to 10-µm inclusion bodies. It is caused by bacterial urinary tract infection, commonly *E. coli*, together with an ill-understood acquired defect of microtubule assembly within phagocytic cells, resulting in the accumulation within the cytoplasm of bacterial remnants that subsequently calcify. Treatment with bethanechol (to stimulate intracellular cGMP and thus microtubule assembly) and ascorbic acid (to stimulate the intracellular hexose monophosphate shunt, which is involved in phagocytosis) have been recommended on theoretical grounds, but seldom arrest the disease. The best chance of avoiding nephrectomy comes from the use of long-term quinolone antibiotics such as ciprofloxacin, which penetrate macrophages well.

Unusual infections

Tuberculosis

Genitourinary tuberculosis is an uncommon late manifestation of tuberculosis, and is often clinically silent, with few if any systemic symptoms. Most cases of renal tuberculosis probably result from haematogenous spread, although unilateral disease is common. Seeding of infection from above leads to ulceration and distortion of the collecting system, pelvis, and ureter, followed by stricture formation and calcification. Obstruction and parenchymal infection may eventually lead to 'autonephrectomy' (Fig. 4). The disease is usually detected either during investigation of asymptomatic sterile pyuria or during investigation of irritative lower urinary tract symptoms or haematuria due to bladder involvement. Reactivation of disease may result from acquired deficiency of 1,25-dihydroxyvitamin D. Occasionally, renal tuberculosis may present with a cold abscess in the flank. Chronic renal failure due to bilateral diffuse interstitial renal tuberculosis may occur, and may account for some of the excess of chronic renal failure in Asian immigrants in the United Kingdom.

Fig. 4 Calcified 'autonephrectomy' as a result of long-standing tuberculous infection. (Reproduced by permission of Professor P. W. Mathieson.)

Diagnosis of renal tuberculosis is by culture of early morning urine samples. Treatment is with rifampicin, isoniazid, pyrazinamide, and ethambutol for 2 months, followed by rifampicin and isoniazid for a further 4 months, and should be supervised by a physician experienced in the chemotherapy of tuberculosis, and with adjustment of the dose of ethambutol in the presence of renal impairment. Corticosteroids may help to prevent or reverse ureteric obstruction, which may otherwise require stent insertion or surgery to prevent renal destruction. Nephrectomy is seldom necessary.

Schistosomiasis

Schistosoma haematobium infection in the venules of the urinary bladder may cause irritative symptoms and terminal haematuria, starting 2 to 3 months after the initial infection. Eosinophilia may be present. The diagnosis is made by detection of ova in a midday terminal urine specimen or by cystoscopy and biopsy. Treatment is with systemic anthelmintic drugs, currently praziquantel.

Fungal infections

Fungal urinary tract infections typically occur in patients whose host defence is compromised by indwelling urethral catheters or ureteric stents, previous wide-spectrum antibiotic therapy, immunosuppressive drugs, or diabetes. Most infections are caused by *Candida* spp. Many patients with funguria have asymptomatic colonization, but some develop life-threatening ascending disease. Severity of infection does not correlate with pyuria. It is important to differentiate funguria from contamination of voided urine by *Candida* in patients with vaginal candidiasis. Many infections clear spontaneously on removal of the urethral catheter, although this can take many months. Treatment options for patients thought to be at high risk of invasive infection (for instance patients with diabetes with indwelling catheters, renal transplant recipients) include continuous bladder irrigation or antegrade perfusion via a nephrostomy tube with amphotericin B at 50 mg/l, and oral fluconazole. Patients with clinical features of acute pyelonephritis require parenteral antifungal treatment, adjusted to *in vitro* sensitivities.

Fungaemia is often complicated by renal parenchymal infection, possibly because the hypertonic and hypoxic conditions in the renal medulla favour transformation of *Candida* from the yeast to the mycelial phase. Infection starts with multiple cortical abscesses and progresses to invasion of the renal pelvis and ureter, with eventual obstruction by fungus balls.

Prospects for the future

Current methods for prevention and treatment of uncomplicated and complicated urinary tract infection are unsatisfactory, with persisting high morbidity and mortality from complicated infection and increasing rates of antibiotic-resistant organisms. Development of new antibiotics is likely only to remain half a step ahead. We hope to see major advances in the prevention of urinary tract infection, perhaps with the development of substances designed to inhibit bacterial adherence to the uroepithelium, the development of new catheter materials and of alternatives to urethral catheterization, and the possibility of vaccines against the virulence determinants of uropathogenic bacteria.

Further reading

Abrutyn E *et al*. (1993). Does asymptomatic bacteriuria predict mortality and does antimicrobial treatment reduce mortality in elderly ambulatory women? [published erratum appears in *Annals of Internal Medicine* 1994, **121**, 901]. *Annals of Internal Medicine* **120**, 827–33.

Bailey RR (1993). Vesicoureteric reflux and reflux nephropathy. In: Schrier RW, Gottschalk CW, eds. *Diseases of the kidney*, 5th edn, pp 689–727. Little, Brown, Boston.

Cardenas DD, Hooton TM (1995). Urinary tract infection in persons with spinal cord injury. *Archives of Physical Medicine and Rehabilitation* **76**, 272–80.

Cattel WR, ed. (1996). *Infections of the kidney and urinary tract*. Oxford University Press, Oxford.

Chew LD, Fihn SD (1999). Pyelonephritis in non-pregnant women. In: Godlee F *et al*., eds. *Clinical evidence*, pp. 761–75. BMJ Publishing Group, London.

Franz M, Horl WH (1999). Common errors in diagnosis and management of urinary tract infection. I: pathophysiology and diagnostic techniques. *Nephrology, Dialysis, Transplantation* **14**, 2746–53.

Franz M, Horl WH (1999). Common errors in diagnosis and management of urinary tract infection. II: clinical management. *Nephrology, Dialysis, Transplantation* **14**, 2754–62.

Gordon I (1995). Vesico-ureteric reflux, urinary-tract infection, and renal damage in children. *Lancet* **346**, 489–90.

Gorelick MH, Shaw KN (1999). Screening tests for urinary tract infection in children: a meta-analysis. *Pediatrics* **104**(5). URL: http://www.pediatrics.org/cgi/content/full/104/5/e54

Hooton TM *et al*. (1996). A prospective study of risk factors for symptomatic urinary tract infection in young women. *New England Journal of Medicine* **335**, 468–74.

Hunt GM, Oakeshott P, Whitaker RH (1996). Intermittent catheterisation: simple, safe, and effective but underused. *British Medical Journal* **312**, 103–7.

Kunin CM, White LV, Hua TH (1993). A reassessment of the importance of 'low-count' bacteriuria in young women with acute urinary symptoms. *Annals of Internal Medicine* **119**, 454–60.

Lachs MS *et al.* (1992). Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Annals of Internal Medicine* **117**, 135–40.

Leibovici L, Wysenbeek AJ (1991). Single-dose antibiotic treatment for symptomatic urinary tract infections in women: a meta-analysis of randomized trials. *Quarterly Journal of Medicine* **78**, 43–57.

Mabeck CE (1972). Treatment of uncomplicated urinary tract infection in non-pregnant women. *Postgraduate Medical Journal* **48**, 69–75.

Platt R *et al.* (1982). Mortality associated with nosocomial urinary-tract infection. *New England Journal of Medicine* **307**, 637–42.

Raz R, Stamm WE (1993). A controlled trial of intravaginal estriol in postmenopausal women with recurrent urinary tract infections. *New England Journal of Medicine* **329**, 753–6.

Saint S, Lipsky BA (1999). Preventing catheter-related bacteriuria: Should we? Can we? How? *Archives of Internal Medicine* **159**, 800–8.

Schaeffer AJ (1994). Urinary tract infection in men—state of the art. *Infection* **22**, S121.

Smaill F (2002). Antibiotics for asymptomatic bacteriuria in pregnancy (Cochrane Review). In: *The Cochrane Library*, Issue 1. Update Software, Oxford. [A substantive amendment to this systematic review was last made on 28 December 2000.]

Stamm WE, Hooton TM (1993). Management of urinary tract infections in adults. *New England Journal of Medicine* **329**, 1328–34.

Stamm WE *et al.* (1982). Diagnosis of coliform infection in acutely dysuric women. *New England Journal of Medicine* **307**, 463–8.

Stapleton A (1999). Prevention of recurrent urinary-tract infections in women. *Lancet* **353**, 7–8.

Svanborg C (1993). Resistance to urinary tract infection. *New England Journal of Medicine* **329**, 802–3.

Warren JW *et al.* (1999). Guidelines for antimicrobial treatment of uncomplicated acute bacterial cystitis and acute pyelonephritis in women. *Clinical Infectious Diseases* **29**, 745–58.

Wong-Beringer A, Jacobs RA, Guglielmo BJ (1992). Treatment of funguria. *Journal of the American Medical Association* **267**, 2780–5.

20.13 Urinary stones, nephrocalcinosis, and renal tubular acidosis

Robert J. Unwin, William G. Robertson, and Giovambattista Capasso

Nephrocalcinosis and urinary stone disease

> 'No stretch of chemical or physical imagination will permit so heterogeneous a group of compounds (as renal stones) to be ascribed to a common origin, or their disposition in the kidney, ureter or bladder to be uniformly charged to an identical cause'
>
> (Howard Kelly)

Nephrocalcinosis and uro-(nephro-)lithiasis frequently coexist and the terms are often loosely combined when describing patients with urinary stone disease. Whether they are aetiologically distinct is unclear, although it is generally believed that nephrocalcinosis represents one end of the spectrum of urinary stone disease. However, although nephrocalcinosis is often associated with urinary stones, most patients with urinary stones do not have macroscopic nephrocalcinosis.

Urolithiasis

Introduction

Urolithiasis has no geographical, demographic, or genetic boundaries: patterns of stone formation have changed in the past and are continuing to change today. The earliest evidence of the disorder is the stones found in mummies entombed in the predynastic Egyptian era, around 4000 BC. In Western countries before 1900, stones occurred commonly in children, particularly boys, and were formed mainly in the bladder. These stones consisted of ammonium urate and/or calcium oxalate and were associated with poor nutrition. Although this form of stone disorder is still found today in rural areas within the so-called 'endemic stone belt' (which stretches from Jordan, through Iraq, Iran, and the Indian subcontinent to the furthest reaches of South-East Asia), it is gradually disappearing with improving standards of nutrition, as it did in most developed countries 100 years ago.

By contrast to the gradual decrease in the occurrence of bladder stones in children, the incidence of upper urinary tract stones (mainly renal) in adults has steadily increased in most countries over the last century. Kidney stones are more common in the industrially developed nations and less so in countries whose economies are more dependent on agriculture. Overall, upper tract stone disease seems to be a disorder associated with affluence, presumably through effects on diet and lifestyle.

Epidemiological factors in the formation of urinary stones

Although stones generally occur more frequently in men than in women (male:female ratio about 2.5:1), recent studies in the United Kingdom and Portugal have shown that, within the past 25 years, there has been a progressive decrease in the age at onset of stone formation in both men and women, particularly women. Within the population of stone formers as a whole, the male:female ratio is now 1.7:1 among patients who formed their first stone before the age of 20 years; but in patients currently aged less than 20 years, the ratio has fallen to 1.1:1. These changes have been attributed to alterations in diet and lifestyle over the last 25 years.

Epidemiological factors important in the formation of urinary stones are summarized in Table 1. Each has been shown to increase the risk of stone formation through effects on the balance between supersaturation and inhibitors and promoters of crystallization in urine (see later).

Calcium oxalate stones

Most (80 per cent) urinary calculi contain calcium oxalate, often on its own, but frequently mixed with calcium phosphate or, occasionally, uric acid. In about 90 per cent of these cases (the so-called idiopathic or primary stone formers), there is no obvious metabolic cause for stone formation. In the remainder, calcium-containing stones form as a result of some disorder of calcium metabolism, oxalate metabolism, or acid–base balance.

For idiopathic calcium stone formation the main epidemiological factors are age, gender, season, climate, stress, occupation, affluence, diet (including fluid intake), and genetic/metabolic factors. The role of diet, in particular, has been studied in detail and appears to explain much of the changing pattern of stone incidence over the past 100 years. As the diet becomes 'richer' in a given population (with an increased consumption of protein, particularly animal protein, refined sugars, and salt), the incidence of stones increases. This often follows periods of economic expansion, whereas the incidence of stones decreases during periods of recession in parallel with a return to a diet containing more fibre and less energy-rich foods. The recent increase in consumption of soft drinks, especially in the young, is becoming an important 'new' factor in the risk of urinary stone formation. These contain phosphoric acid, providing a small acid load, but one that may become significant if large volumes are drunk. Paradoxically, potential sources of oxalate, such as beer, may be associated with a reduced stone risk, perhaps because a minimum ingestion of oxalate is necessary to bind dietary calcium and limit calciuria. Antacid ingestion (as distinct from 'milk-alkali syndrome') may also reduce stone risk by increasing urine pH, binding oxalate, and providing a source of magnesium.

Infection stones

So-called 'infection stones', composed of magnesium ammonium phosphate, usually in conjunction with calcium phosphate, are more common in women and now constitute between 4 and 15 per cent of stones, depending on the country of origin. They are caused by urinary tract infection with urea-splitting organisms that secrete the enzyme urease. This converts urea to ammonium (NH_4^+) and bicarbonate, making the urine more alkaline (urinary NH_4^+ concentration is normally low in sterile alkaline urine). As a result, phosphate-containing salts, such as calcium phosphate and magnesium ammonium phosphate, precipitate and increase the risk of stone formation. Infection stones can also form secondary to most other types of stone. The relative incidence of infection stones has decreased over the past 25 years in most Western countries, presumably as a result of better clinical diagnosis and earlier treatment of urinary tract infections.

Uric acid stones

Stones consisting of uric acid constitute between 4 and 25 per cent of published series, depending on the relative consumption of animal and vegetable protein in the population studied. There are three factors that promote formation of uric acid stones: (i) low urine volume, (ii) acid urine pH, and (iii) high uric acid excretion. For a given diet, these stones are more common in elderly men, at least in part because of the decline in urine pH with age. Because the pK of uric acid/urate is approximately 5.7, a more acid urine pH favours the less soluble undissociated form of uric acid. 'Pure' uric acid stones are infrequent in most developing countries and are most common in the oil-rich states of the Arabian Gulf, or in countries where there is a cheap local source of meat, fish, or poultry protein. Most uric acid stones are idiopathic; a small number form secondary to some disorder of purine metabolism (such as Lesch–Nyhan syndrome), or to a condition in which there is high tissue turnover (such as tumour necrosis following chemotherapy).

'Rare' stones

In all series of stones analysed, between 1 and 2 per cent consist of a range of 'rare' constituents derived from either some hereditary or congenital inborn error of metabolism, such as cystinuria (not to be confused with cystinosis), xanthinuria, or 2,8-dihydroxyadeninuria, or from a prescribed drug or metabolite, which is relatively insoluble in urine. Examples are silica (from excess ingestion of the antacid magnesium trisilicate, or from the use of pectin and silicium to thicken milk for infant feeding), sulphonamides, indinavir, and triamterene. All stones contain a small percentage by weight of mucoproteinaceous matrix. Some 'stones' consist almost entirely of mucoprotein, and usually result from inflammation of the urinary tract in patients whose urine is not sufficiently supersaturated to mineralize the organic matrix.

Causes of urinary stone formation

Insolubility of mineral components

The overriding factor that is common to all types of stone is the relative insolubility of their respective mineral component(s) in the urine. However, whether or not this is the only factor responsible for the formation of stones is still open to debate. There are two possible models (which are not mutually exclusive) for the initiation of stones. The 'free-particle' model is that stones are initiated when urine becomes so excessively supersaturated with one of the salts or acids occurring in kidney stones that crystals spontaneously precipitate in urine. If this happens frequently, and if the crystals grow or aggregate sufficiently within the transit time of urine through the kidney, then the risk increases that one of these particles will become trapped at some narrow point along the urinary tract and act as a focus around which a stone can form. The other model of stone formation, which is currently favoured, requires chemical 'fixation' of a crystal, or aggregate of crystals, to the renal epithelial cell lining. This fixed particle may result from injury to the cell wall (caused either by the crystals themselves or by viruses or bacteria) and/or from some 'gluing' material—present only in the urine of stone formers—that causes crystals to adhere to these sites and then results in stone formation.

Both models require urine to be supersaturated to some degree with respect to the stone-forming salt or acid concerned, sufficient to cause crystals to be formed by nucleation that is either homogeneous (spontaneous) or heterogeneous (on a pre-existing nucleus of some foreign material). The factors causing urine to become supersaturated with one or more of the various constituents of stones are shown in Table 2.

Modifiers of crystallization

One factor that might affect the kinetics of the processes involved is the presence or absence in urine of so-called modifiers of crystallization, claimed to be of particular importance in the formation of calcium-containing stones. One group of crystallization modifiers is said to retard the rate of growth and/or aggregation of crystals, or the binding of calcium-containing crystals to cell walls. These are known as inhibitors of crystallization and include magnesium, citrate, pyrophosphate, ADP, ATP, at least two phosphopeptides, glycosaminoglycans, Tamm–Horsfall protein, nephrocalcin, calgranulin, fibronectin, various plasma proteins, osteopontin (uropontin), a_1-microglobulin, b_2-microglobulin, urinary prothrombin fragment 1, and inter-a-trypsin inhibitor. Of these, urinary citrate is probably the most important. The second group of modifiers is claimed to promote one or more of the processes involved in crystallization. These are known as promoters of stone formation and include matrix substance A, various uncharacterized urinary proteins and glycoproteins, and the polymerized form of Tamm–Horsfall protein (uromucoid). However, the clinical importance of these compounds in the pathogenesis of stone formation remains unclear.

In the final analysis, stone formation is probably due to an abnormal combination of factors that affect either the thermodynamic (supersaturation driving force) or kinetic (rate-controlling) processes involved in the crystallization of the various stone-forming minerals. For some types of stone formation (cystine, xanthine, 2,8-dihydroxyadenine, uric acid, and probably magnesium phosphate ammonium stones) the thermodynamic factors predominate; in others (calcium oxalate and calcium phosphate) both sets of factors may be involved.

Idiopathic hypercalciuria

This is often familial, accounts for the majority (more than 50 per cent) of patients with renal stones, and can be divided into three types: absorptive, resorptive (or fasting), and renal. The most common is absorptive, due to increased intestinal absorption of calcium, the cause of which remains unknown. Resorptive hypercalciuria is associated with reduced bone mineralization, although primary hyperparathyroidism must be excluded. Renal hypercalciuria is distinct from that seen in tubular disorders such as the Fanconi syndrome (see under renal tubular acidosis). Although the underlying mechanism is not known, there may be a primary defect of renal phosphate reabsorption, suggested in some cases by an increase in plasma calcitriol (1,25-OH vitamin D) levels, which will enhance intestinal absorption of calcium.

Hypocitraturia

This is an important risk factor for renal stone disease. The blood citrate pool is maintained by delivery from bone and the gastrointestinal tract and by removal by hepatic and renal metabolism. High urinary citrate prevents calcium stones by encouraging formation of soluble calcium citrate; it also reduces formation of urate stones by alkalinizing the urine. Hypocitraturia is present in about 40 per cent of calcium stone formers, but in most cases the reason for this is unknown.

Low urinary citrate excretion results from metabolic acidosis in conditions such as chronic diarrhoea, urinary diversion, and distal renal tubular acidosis. The hypocitraturia of distal renal tubular acidosis is due to increased reabsorption of citrate in the proximal tubule as a result of intracellular acidosis (see section on renal tubular acidosis below for more detail). Citrate excretion is also reduced because of acid retention in subjects on a high protein diet. The widely prescribed angiotensin-converting enzyme (**ACE**) inhibitor enalapril has been shown to decrease citrate excretion in rats, although the effect is small in humans. It is not known whether this is true of all ACE inhibitors, or if it is of any significance in patients at risk of renal stones (particularly given the increased prevalence of hypertension in renal stone disease).

Hyperuricosuria

The contribution of increased uric acid excretion to uric acid stone formation occurs mainly in patients on a high protein (purine) diet, which leads to the production of more acid urine and increases the risk of urate precipitation. Hyperuricosuria is less commonly due to a defect of urate metabolism *per se*.

Up to a fifth of patients with gout also have urinary stone disease (urate, calcium oxalate, calcium phosphate, or mixed). Hypertension is commonly associated with both conditions; hence, as already mentioned, it may be important to consider the effect of antihypertensive therapy on the risk of urolithiasis. The angiotensin receptor blocker Losartan lowers plasma urate concentration and increases uric acid excretion by an unknown mechanism, although this does not seem to increase the risk of uric acid stones.

Hyperoxaluria

The excretion of oxalate is often mildly elevated in idiopathic stone formers. Most oxalate is derived from the metabolism of glycine and ascorbic acid, vitamin C intoxication being a rare cause of hyperoxaluria (vitamin C can be non-enzymatically converted to oxalate in urine; this can be prevented by collecting urine in acid or EDTA; up to 4 g/day of vitamin C has no significant effect on urinary oxalate excretion). Intestinal absorption of oxalate is normally low, but rises when dietary calcium content is reduced. It is also increased following small bowel resection and in Crohn's disease (so-called enteric hyperoxaluria), when saponification and the action of bile salts increase the permeability of the large bowel to oxalate. In these conditions diarrhoea and malabsorption can lead to chronic dehydration, low urinary volumes and metabolic acidosis, with the risk of interstitial oxalate deposition causing acute or chronic renal failure.

Recent research suggests that normal bowel colonization with the bacterium *Oxalobacter formigenes* is an important determinant of urinary oxalate excretion, because this organism digests dietary oxalate, thereby reducing its absorption. This might be relevant to the association between renal stones and long-term antibiotic use in patients with cystic fibrosis, although their high protein intake from pancreatic enzyme supplements may also be a factor.

The two genetic types of primary hyperoxaluria are autosomal recessive and cause oxalate overproduction. Type 1 (PH1) is the more severe form, producing widespread tissue deposition of oxalate (oxalosis), early renal failure, and nephrocalcinosis. It is due to a defect of the liver transaminase that converts glyoxylate to glycine, resulting in glyoxylate oxidation to oxalate and reduction to glycollate. The rarer type 2 (PH2) is due to a deficiency of liver D-glycerate dehydrogenase and is

characterized by glyceraturia. Since pyridoxine (vitamin B_6) is a cofactor for the defective enzyme in PH1, high doses of this vitamin can sometimes help to reduce oxalate production; more modest doses are sometimes also effective in patients with mild 'idiopathic' hyperoxaluria.

Cystinuria, xanthinuria, and 2,8-dihydroxyadeninuria

A small number of stones are found in patients with cystinuria, xanthinuria and 2,8-dihydroxyadeninuria arising from inherited or congenital errors of metabolism. Cystine is normally excreted in very low concentrations, well below the limit of solubility of cystine in urine (1 to 1.5 mmol/l). In cystinuria, due to a defect in the tubular reabsorption of cystine, the urinary concentrations are 200-fold higher than normal, leading to precipitation of cystine and stone formation. Only homozygotes form stones; the excretion of cystine in heterozygotes is higher than normal, but insufficient to cause crystal formation in most cases. The basic amino acids ornithine, lysine, and arginine share the same amino acid transport mechanism as cystine and their excretion is also increased.

Xanthinuria results from a deficiency of the enzyme xanthine oxidase, such that xanthine is not converted to uric acid. Radiolucent stones form in acid urine. Secondary xanthinuria producing xanthine stones is an unusual complication of allopurinol therapy.

A deficiency of the enzyme adenine phosphoribosyl transferase, inherited as an autosomal recessive trait, is associated with an increased urinary excretion of 2,8-dihydroxyadenine and this sometimes leads to stone formation.

Clinical features

Calculi can occur at any point in the urinary tract, although they are more often located in the kidney and ureter than in the bladder. Upper urinary tract stones can occur on either side and are often bilateral. Over 60 per cent are small enough to be passed spontaneously. Within the kidney itself, concretions may be found in the calyces, in the renal pelvis, or extending from the calyces into the pelvis. They may be attached to the epithelial surfaces of the pelvicalyceal system, be encapsulated within the renal parenchyma, or lie free within the pelvis or lower pole of the kidney. Occasionally, they may occupy the entire pelvicalyceal space to form a so-called 'staghorn' calculus. The clinical presentation of calculi depends primarily upon their position in the urinary tract, and whether or not they cause obstruction to urinary flow.

Pain from urinary stones

The commonest presentation of urinary calculi is with pain. Stones can become lodged at any point in the ureter, but most commonly do so at the upper and lower ends, where there are constrictions at the pelviureteric and vesicoureteric junctions. A stone in the renal pelvis typically causes dull loin pain with occasional colic. The most problematic differential diagnosis is from musculoskeletal pain arising in the back, lower ribs, or their muscular and ligamentous attachments. Such pain is common, and it can be very difficult to decide whether a stone seen on a radiograph is 'incidental' (see later) or the cause of symptoms. Musculoskeletal pain is more likely to be precipitated by bad posture, exercise, or movement, to come on suddenly, to last for a few seconds or minutes only, and to be associated with a localized 'superficial' point of tenderness. Renal pain is less likely to be brought on by exercise or movement and more likely to be felt 'deeper inside', last for hours, and be associated with diffuse loin tenderness and no comfortable position.

Partial or total obstruction of the pelviureteric junction or ureter gives rise to the agonizing pain of renal colic, described by many sufferers as the 'worst pain they have ever had'. The patient can be in absolute agony, rolling around and crying out as the waves of colic strike them. If the stone is at the pelviureteric junction, the pain is felt in the loin. If it moves down the ureter, the pain radiates into the groin and (sometimes) into the scrotum or labia. The patient sweats profusely with the waves of pain and often vomits. Treatment with non-steroidal anti-inflammatory agents may be helpful, but the physician should not refrain from giving prompt and adequate doses of powerful analgesics (pethidine, morphine) and will certainly gain the undying gratitude of their patient if they do so. The diagnosis of renal colic is often straightforward from the history alone, but in some cases differential diagnoses need to be entertained, which include biliary colic, small bowel obstruction, appendicitis, and diverticulitis. Children and pregnant women are less likely to present with characteristic symptoms and a high index of suspicion is sometimes required to make the diagnosis.

The patient with stones in the upper urinary tract may have macroscopic haematuria and frequently will have microscopic haematuria, together with episodes of frequency and dysuria caused by the passage of 'sand' or 'gravel'.

'Staghorn' calculi are associated with infection, as described previously, but urinary infection is probably more common in any individual with urinary stones, and the combination of urinary obstruction with sepsis can be particularly dangerous. Added to symptoms arising from obstruction are those from infection, with high fever (often 39°C or higher), rigors, and (in severe cases) circulatory collapse. These cases are medical emergencies requiring resuscitation, intravenous antimicrobial therapy, and urgent relief of obstruction.

Another circumstance worthy of note is the patient with a single functioning kidney. Obstruction by a stone, or any other cause, will lead to acute obstructive renal failure, demanding urgent relief of obstruction.

Calculi may be found in the bladder, either lying free or lodged in a diverticulum in the vesical wall. Most probably they originate in the upper urinary tract and continue to grow in their new location, although others may be initiated in the bladder. They occur almost exclusively in elderly men and usually consist of uric acid, or are associated with infection due to chronic prostatic obstruction. Stones in the bladder may be asymptomatic, but can cause discomfort in the suprapubic and perineal regions, and also the dramatic symptom of sudden and painful cessation to urine flow.

Imaging

Diagnosis is usually confirmed by a combination of ultrasound, plain abdominal radiography (kidney, ureters, and bladder), and particularly in the case of radiolucent stones (i.e. those composed of uric acid, 2,8-dihydroxyadenine, or xanthine), an intravenous urogram. The place of the intravenous urogram is still hotly debated, although there is really no better means of defining the anatomy of the renal drainage system and at the same time providing some index of renal function, which an ultrasound examination cannot do. Occasionally, a computed tomographic (CT) scan (especially spiral) can provide useful additional information, particularly if there is associated nephrocalcinosis or lucent stones. Magnetic resonance imaging is poor at showing calcium deposits.

Recurrence

If patients are not provided with proper preventative management, the risk of recurrence is high—40 per cent within 3 years, rising to 74 per cent at 10 years, and 98 per cent at 25 years. The rate of recurrence appears to be higher after the use of minimally invasive techniques for the disintegration and/or removal of stones than it was in the days of open kidney surgery. This is particularly noticeable in patients treated with extracorporeal shock-wave lithotripsy, which by the nature of the technique tends to leave behind fragments of stone that may act as foci for subsequent stone formation. Percutaneous nephrolithotomy is slightly less of a problem in this respect, because it is usually easier to ensure that most of the stone fragments are removed during this procedure: it is usually preferred to lithotripsy for treatment of a stone present in the lower pole of the kidney, or greater than 2 cm in diameter. It is important to remember that sole reliance on surgical procedures for the overall management of stone formers is inadequate: stones tend to recur unless dietary and/or medical treatment is also instituted.

Asymptomatic urinary stones

About 3 per cent of patients undergoing abdominal ultrasound or radiographic investigations for other indications are found to have 'silent' kidney stones (Fig. 1(a)). If the stone is greater than 4 mm in diameter and of uncertain location, an intravenous urogram should be performed to determine this in relation to the renal pelvis and ureter and to assess the risk of future obstruction (Fig. 1(b)). A simple metabolic screen (see later) is also required. If the patient remains asymptomatic, no specific abnormality is found, and no intervention deemed necessary, then a follow-up radiograph or ultrasound at 2 years is probably advisable, depending on the size of the stone.

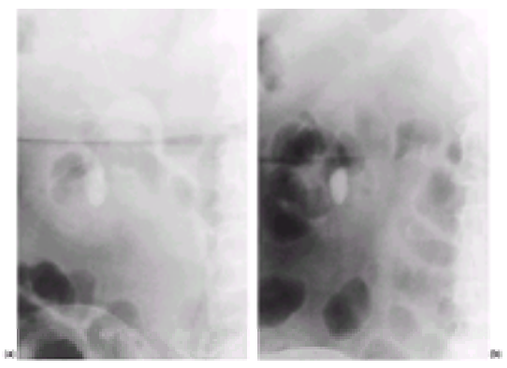

Fig. 1 A plain radiograph (a) and intravenous urogram (b) of an adult woman showing an incidental right kidney stone. The stone is in a lower pole calyx or calyceal diverticulum and measures 8 by 15 mm.

Biochemical screening to determine risk of urinary stones

Once the presence of a stone is confirmed and a decision is reached on the most appropriate urological intervention (if any), the patient should be screened for a biochemical cause. It is usually best to do this when the patient is eating and drinking 'normally', but not when there is haematuria or immediately prior to stone removal or lithotripsy. After such treatment it is sensible to wait for 2 to 3 months, because during this period patients often consume a diet that is very different from their 'normal'one. If they do not form another stone within 3 months of the presenting episode, then they will frequently return to former dietary habits and an increased risk of stone formation.

There are several published biochemical screening procedures for assessing the risk of stone formation, most of which require a stone analysis, a metabolic screen, a 24-h urine screen, and in some instances, a dietary history. These should not be carried out when the patient is in hospital, because the diet is usually very different from that consumed at home. In addition to the routine clinical history from the patient, which should include occupation and family history of stones, the screen includes the following.

1. Analysis of a simultaneous blood and spot urine sample. The blood sample should be analysed for urea, creatinine, calcium, magnesium, sodium, potassium, bicarbonate, phosphate, urate, alkaline phosphatase, albumin, and oxalate (where necessary). The urine sample should be analysed for urea, creatinine, pH, calcium, sodium, potassium, phosphate, and urate.
2. Analysis of two 24-h urine samples collected on consecutive days at home: the first in a container with 50 ml of 2.2 mol hydrochloric acid as a preservative and analysed for volume, creatinine, calcium, magnesium, sodium, potassium, phosphate, oxalate, and citrate; the second in a plain container and analysed for volume, creatinine, pH, protein, urate, and a qualitative test for cystine.
3. Dietary assessment carried out using the diet diary system during the week leading up to, and including, the 2 days of the 24-urine collections. The patient is asked to complete a diet diary of everything that he or she consumes each day. This is analysed for fluid intake, calories, calcium, magnesium, sodium, potassium, phosphate, oxalate, purine, protein (and its various fractions, including animal protein, meat plus fish plus poultry protein, dairy protein, and fruit plus vegetable plus cereal protein), fibre, fat, and refined sugars.
4. Quantitative stone analysis (whenever possible). All patients should be encouraged to retain their stones or stone fragments for quantitative analysis by infrared spectroscopy. This is an important tool, often providing the first clue as to the cause of the stone(s) in a given patient.

From the combined analyses, a number of algorithms can be used to assess the overall biochemical risk of forming stones containing uric acid, calcium oxalate, or calcium phosphate, or various mixtures of these constituents (details of several such algorithms that are currently used in various specialist urinary stone clinics can be found in chapter 2 of Coe *et al.*1996). These are used to give an indication of the risk of stone formation in an individual patient. High risk is rarely caused by a single abnormal urinary constituent (except in the case of primary hyperoxaluria), but usually due to a combination of several lesser abnormalities, depending on the stone type. Indeed, it is possible to be at high risk with every single individual risk factor within its 'normal range', but with several lying towards the upper or lower limits of these ranges. This is an important feature of these models because they allow a risk assessment to be made in the patient who would otherwise be described as 'having normal urine', yet has an abnormal combination of the variables that lead to crystalluria and stones. The models can be used both to assess the patient's probability of stones before treatment, and also to follow progress during preventative management.

Prevention of stone recurrence

The main aim in the prevention of stone recurrence is to decrease the likelihood of crystals forming in the urinary tract by reducing the supersaturation of urine with respect to the particular constituent(s) that occurs in a patient's stone.

A summary of the available dietary and medical treatments for the various types of urinary calculi is shown in Table 3. The injunction to 'drink more' is crucially important for all at risk of urinary stones, and patients should aim to maintain a urinary volume of at least 2.5 litres per day. Those with recurrent stones will certainly benefit from maintaining a higher urinary volume than this, and also from making a point of drinking when they get up at night to micturate. This is undoubtedly an inconvenience, but recommended to ensure that the urine is as dilute as possible throughout the 24 h, and does not become concentrated at night. Patients taking treatment to modify their urinary pH should be given appropriate sticks to measure this, instructed how to use them, and how to modify their treatment to achieve the desired effect.

Some patients will benefit from alteration of their diet, in particular those who eat excessive amounts of animal protein and purine from meat, fish, and poultry; those who consume large quantities of oxalate-containing foods; and those who not only consume high amounts of calcium but also hyperabsorb calcium from the intestine. Too low an intake of calcium, on the other hand, may increase the intestinal absorption and hence urinary excretion of oxalate. It is important, therefore, not to advise patients to cut out all dairy produce to correct their hypercalciuria, as they may end up with a higher risk of forming stones than when they started. Other dietary excesses that may increase the risk of stones include a high intake of salt, which leads to a renal leak of calcium, and a high intake of refined sugars, which increase the intestinal absorption of calcium. When taken together, gross hypercalciuria often results. A list of foods that should be avoided in excess and taken only in moderation is contained in Table 4.

Although most of the treatments listed in Table 3 and Table 4 are effective in reducing the risk of stone recurrence, the main problem in the long-term management of patients with stones is compliance. Stone formers typically feel well for most of the time, except when experiencing an attack of renal colic. It is therefore often difficult to maintain co-operation and motivation to adhere to preventative treatment for long after a first stone episode, particularly since this is socially intrusive, with drinking of large volumes leading to urinary frequency. If patients do not have a recurrent stone within a few months, they will generally return to their original abnormal pattern of urine biochemistry by 3 to 6 months and eventually produce another stone. Once they have had several episodes of renal colic, compliance with treatment is usually better.

Figure 2 illustrates how the important urinary risk factors for stone formation can be represented visually to encourage understanding, compliance, and motivation. The patient can see immediately how changes in diet and urine composition affect his/her stone risk, which can be an incentive to try and reduce the combined risk by moving it from outside (high risk) toward the central (low risk) 'bull's eye'. It is important to review the patient regularly as an outpatient and to repeat the relevant screening tests, preferably annually but at least biennially, to encourage adherence to the recommended treatment.

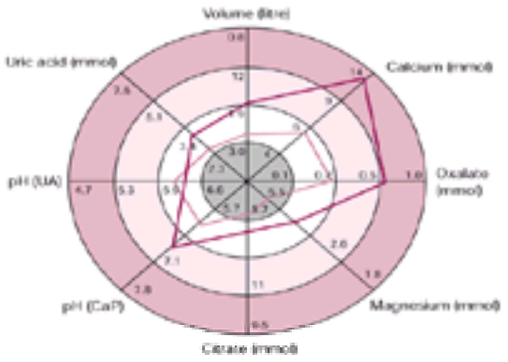

Fig. 2 A radar plot ('target diagram') showing the baseline urinary risk factors for stone formation (purple lines) and the corresponding data during preventative treatment (pink lines) in a patient with 'mixed' calcium oxalate and calcium phosphate stones. The individual urinary risk factors are plotted such that abnormal values fall in the dark pink outer ring and normal values in the grey bull's eye. Intermediate values fall in the white and light pink areas. The objective for the patient is to aim to get his/her risk factors in the 'bull's eye'. Units of axes are urinary volumes or amounts per 24 h. UA, refers to a patient with uric acid stones; CaP, refers to a patient with calcium phosphate-containing stones.

Economics of management of urinary stones

The undoubted success of lithotripsy, less invasive surgery (such as percutaneous nephrolithotomy), and ureteroscopy for the disintegration and removal of calculi has lulled some into the belief that urinary stone disease can be managed solely by these interventions. Although these minimally invasive techniques are often the procedures of choice for the removal of stones, they do not prevent their recurrence. Without biochemical screening and appropriate dietary and/or medical management, the patient will often return for further stone removal in the future, which can be uncomfortable for the patient and is expensive to perform. Failure to provide proper prophylactic treatment and follow-up can also result in missed infections and eventually compromised renal function. Financial analysis has shown that the projected costs of treating patients with stones by only removing their stones as they form is much more costly than removing the initial stone(s) and then screening to identify risk factors to provide appropriate advice and long-term management.

Nephrocalcinosis

Nephrocalcinosis is the deposition of calcium salts (mainly phosphates and oxalates) within the kidney, but the term is usually reserved for those conditions in which there is a generalized increase in kidney calcium content, rather than any localized intrarenal calcification, as may occur in some tumours, cysts, tuberculous granulomas, and areas of renal infarction.

Microscopic, or mild, nephrocalcinosis is a common incidental finding at autopsy, but macroscopic nephrocalcinosis is uncommon. It is diagnosed by radiography or ultrasound, although there is still some debate amongst radiologists as to whether radiography or ultrasound is more sensitive in detecting early disease. Plain radiographs have the advantage that they can more readily be used to judge progress from one year to the next, and they are better at detecting associated urinary stone disease. It is useful to have both tests, except in children and women of childbearing age.

Cortical nephrocalcinosis is seen following extensive acute cortical infarction, and in chronic glomerulonephritis or pyelonephritis. It can occur in the transplanted kidney and in severe oxalosis, conditions that also produce a medullary distribution. An example of the more common (and clinically more important) medullary form of nephrocalcinosis is shown in Fig. 3. There are many causes of medullary nephrocalcinosis: it is usually associated with disordered calcium homeostasis and can, like urolithiasis, be broadly divided into hypercalcaemic, hypercalciuric/non-hypercalcaemic, and non-hypercalciuric forms. However, the presence of nephrocalcinosis is more likely to signify an underlying metabolic disturbance than is isolated urinary stone disease.

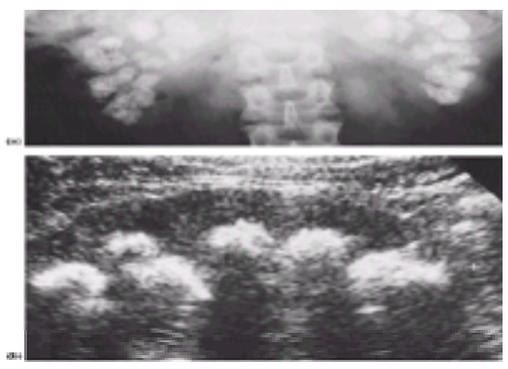

Fig. 3 A plain radiograph (a) and renal ultrasound scan (b) of a man with familial distal renal tubular acidosis and severe medullary nephrocalcinosis. Note the striking medullary distribution of calcification in the radiographic picture and the increased medullary reflections and acoustic shadows 'behind' in the ultrasound picture.

The most important risk factors for nephrocalcinosis are the same as those for urolithiasis:

1. hypercalcaemia (leading to an increased filtered load of calcium), which is linked to diet;
2. enhanced intestinal absorption of calcium, increased parathyroid hormone activity, or vitamin D excess;
3. altered renal tubular handling of calcium resulting in hypercalciuria; and
4. absence of factors in urine (such as citrate) that help to maintain calcium salts in solution.

Nephrocalcinosis associated with hypercalcaemia

Hyperparathyroidism is a common cause, accounting for approximately a third of cases in the large series of Wrong and Feest. Skeletal breakdown due to neoplasia or bone loss from chronic immobilization and severe osteoporosis can also be associated with nephrocalcinosis, although the latter, especially when the result of steroid therapy, usually only causes hypercalciuria. Vitamin D intoxication is often iatrogenic, a consequence of treating combined hypophosphataemic and hypocalcaemic states with vitamin D, calcium, and phosphate supplements. Sarcoid-induced hypercalcaemia is also related to increased activity of vitamin D, due in this circumstance to increased synthesis in granulomas, and sometimes causing significant hypercalciuria without overt hypercalcaemia. Nephrocalcinosis in association with hypothyroidism has been reported. By contrast, nephrocalcinosis is not seen in thyrotoxicosis, which can cause hypercalcaemia, except when there is an autoimmune basis and may therefore be associated with autoimmune distal renal tubular acidosis (see later in this section). A rare cause is the idiopathic hypercalcaemia of infancy (William's disease).

Nephrocalcinosis associated with hypercalciuria without hypercalcaemia

The commonest cause in this category is distal renal tubular acidosis, which is associated with low urinary excretion of citrate. Hypercalciuria is a consistent finding in patients with the complete form of this disease and systemic acidosis (see under renal tubular acidosis). Although hypercalciuria is also a feature of the Fanconi syndrome and proximal renal tubular acidosis, these conditions are less commonly associated with nephrocalcinosis, presumably because of supranormal excretion of citrate in many cases.

The next most common cause of nephrocalcinosis is medullary sponge kidney. The cause of this condition is unknown. The collecting ducts are dilated and become

the site of crystal formation and precipitation, probably related to stasis. The diagnosis can only be made reliably on an intravenous urogram, which shows grape-like clusters of ectatic medullary collecting ducts filled with contrast dye. Medullary sponge kidney can run in families and in some cases is associated with hemihypertrophy affecting the face, an arm, or a leg. Because damage to the collecting ducts can affect acid excretion, medullary sponge kidney can also be associated with a 'secondary' form of distal renal tubular acidosis, as well as mild nephrogenic diabetes insipidus.

Absorptive hypercalciuria is another cause: this may be idiopathic, but is also seen with vitamin D excess and in sarcoidosis. 'Milk-alkali syndrome' as a cause, due to excess calcium ingestion from calcium carbonate, is now rarely seen. Hypercalciuria and nephrocalcinosis can also occur in inherited tubular disorders like Bartter's syndrome and familial magnesium-losing nephropathy, and may follow intensive loop diuretic treatment in premature infants.

Non-hypercalciuric

Hyperoxaluria and oxalosis have been described earlier (see under urolithiasis). Various chronic hypokalaemic states have been associated with nephrocalcinosis. Hypokalaemia causes an intracellular acidosis, which reduces citrate excretion and may cause unspecified tubular damage. Reported examples are the hypertensive syndromes of apparent mineralocorticoid excess, due to a defect in the enzyme 11b-hydroxysteroid dehydrogenase, and Liddle's syndrome, due to an activating mutation of the collecting duct sodium (Na^+) channel (ENaC). Other conditions in which hypokalaemia may be a contributory factor include Bartter's syndrome and loop diuretic use in infancy. These and other causes are listed in Table 5.

Clinical approach to the patient with nephrocalcinosis

In most patients with nephrocalcinosis, a clinical diagnosis can be made and contributory factors identified. However, as with urinary stones, it is still poorly understood why changes in important risk factors (such as urinary calcium or citrate excretion) occur, and why or how calcium is deposited. The impact of nephrocalcinosis on health varies and its presentation can range from incidental, when detected on abdominal radiographs or ultrasounds performed for another reason, to life threatening. It may result in uncontrolled hypertension, renal infection, scarring, renal colic, defects of renal tubular function (impaired urinary concentrating ability and acid excretion—mild diabetes insipidus and secondary distal renal tubular acidosis), and even renal failure, although this is unusual.

Treatment

A metabolic cause must always be sought and treated if found. Treatment and management are otherwise very similar to those for macroscopic renal stone disease, and are mainly symptomatic. Assessment of dietary risk factors may be of some help if the underlying cause, such as distal renal tubular acidosis, cannot be easily corrected. Surgical intervention is only required if significant stones form, are passed frequently, or cause obstruction and infection. There is no place for lithotripsy, which may actually do harm, as the calcium deposition is largely parenchymal.

Renal tubular acidoses

The term 'renal tubular acidosis' is used to describe a group of clinical disorders in which net renal excretion of acid (hydrogen ions, H^+) is impaired as a result of renal tubular dysfunction. Strictly speaking, this definition could also include chronic renal failure, but by convention a reduced glomerular filtration rate as the cause of failure of acid excretion is excluded. Primary abnormalities of urinary acidification, renal tubular acidosis, are responsible for approximately 20 per cent of cases of medullary nephrocalcinosis with stones (see under nephrocalcinosis). To understand the pathogenesis of the renal tubular acidoses, an understanding of the role of the kidney in acid–base balance is required.

Role of the kidney in acid–base balance

Normal H^+ excretion depends on the ability of the renal tubule to reabsorb filtered bicarbonate (HCO_3^-) in the proximal nephron, followed by net H^+ secretion (approximately 50 mmol/day) by excretion of titratable acid and ammonium (NH_4^+) in the distal nephron (distal tubule and collecting duct) (see Fig. 4).

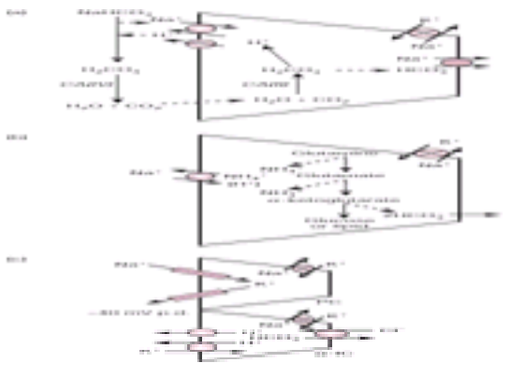

Fig. 4 Diagram showing cellular mechanisms of bicarbonate (HCO_3^-) reabsorption, ammonium (NH_4^+) generation and secretion, and hydrogen ion (H^+) secretion along the nephron. (a) Reclamation of filtered HCO_3^-, also mechanism of titratable acid; (b) generation and secretion of NH_4^+; (c) H^+ secretion. CA, carbonic anhydrase; PC, principal cell; a-IC, a-intercalated cell. See text for further details.

Reclamation of filtered bicarbonate

Normally, almost all filtered HCO_3^- is reabsorbed, the bulk of it in the proximal tubule (about 80 per cent), the remainder in the loop of Henle (about 15 per cent) and distal nephron (about 5 per cent). The mechanism of reabsorption is indirect: H^+ and HCO_3^- are generated in renal tubular cells (with the aid of carbonic anhydrase type II, **CA-II**), and the H^+ are secreted apically into the lumen, while the HCO_3^- enters the plasma via the basolateral cell membrane. In the proximal tubule and loop of Henle, most of the H^+ secretion is via Na^+/H^+ exchange in the apical (luminal) membrane, although there is also a contribution from primary active H^+-ATPase. The secreted H^+ reacts with filtered HCO_3^- to produce carbonic acid (H_2CO_3), which is rapidly converted to CO_2 and H_2O by carbonic anhydrase type IV (**CA-IV**) present on the luminal membrane, and the CO_2 and H_2O diffuse into the cell. The net result is that for every filtered HCO_3^- removed from tubular fluid, another replaces it in plasma.

Addition of 'new' bicarbonate to plasma

The bulk of H^+ generated within tubular cells is involved in reclaiming filtered bicarbonate (more than 4000 mmol/day). However, under normal conditions, because of the excess acid produced from food metabolism (about 50 mmol/day), it is necessary to add 'new' (extra) HCO_3^- to the plasma to replace that which has buffered this acid load. This is achieved by the generation of H^+ (and HCO_3^-) within tubular cells in addition to those needed to effect HCO_3^- reabsorption. The extra HCO_3^- enters the plasma, thus making the bicarbonate content of blood in the renal vein slightly higher than that in the renal artery. What happens to the H^+ produced simultaneously is a little more complicated. It would seem simplest to secrete these H^+ directly and independently into the tubular lumen and excrete them in the urine. However, the excretion of around 50 mmol of free H^+ per day in urine would lower urine pH to approximately 1.5 ($[H^+]$, ? 31 mmol/l), which the tubular epithelium cannot do, because it is only able to sustain a maximum pH gradient between plasma (pH ? 7.4; $[H^+]$? 40 nmol/l) and tubular fluid of 3 pH units (minimum urine pH ? 4.5; $[H^+]$? 31 600 nmol/l). The extra H^+ are excreted in two different ways: as titratable acid and as NH_4^+.

Excretion of titratable acid

Some of the extra H^+ are secreted into the lumen, where they can react with buffer anions in the tubular fluid (principally filtered HPO_4^{2-}); any buffer that is not reabsorbed will therefore excrete acid. The total amount of H^+ lost in the urine in this way can be determined by back-titrating the urine with a strong base, such as NaOH, until the urine pH is raised to 7.4 (that of arterial plasma); hence the term 'titratable acid'. It usually amounts to about 20 mmol/day. Approximately half the titratable acid production occurs in the proximal tubule, where tubular fluid pH falls to around 6.8 (equal to the pK of the $HPO_4^{2-}/H_2PO_4^-$ buffer system). The remainder occurs in the collecting duct, where H^+ can be secreted against a higher concentration gradient and consequently, as indicated above, urine pH can fall to approximately 4.5.

Excretion of ammonium

The proximal tubular cells are capable of taking up the amino acid glutamine and deaminating it to form NH_4^+ and a-ketoglutarate. NH_4^+ is secreted into the tubular lumen (substituting for H^+ on the Na^+/H^+ exchanger—Fig. 4) and eventually excreted, whereas a-ketoglutarate is converted to glucose, through reactions that consume H^+. The secreted H^+ (in NH_4^+) is generated from CO_2 and H_2O, and the HCO_3^- that is formed at the same time is added to plasma. It is important that the NH_4^+ produced from glutamine enters the tubular fluid and not the plasma: if it did so it would be taken up by the liver and combined with bicarbonate to produce urea ($CO(NH_2)_2$) and CO_2, effectively neutralizing the bicarbonate generated in the proximal tubule thus: $2NH_4^+ + 2HCO_3^- \rightarrow CO(NH_2)_2 + CO_2 + 3H_2O$. In severe liver disease, metabolic alkalosis occurs because urea synthesis is impaired and bicarbonate consumption in this way is reduced. The converse of this is that when whole kidney NH_4^+ production and 'new' bicarbonate generation are impaired (as in chronic renal failure), continuing synthesis of urea by the liver will generate unneutralized H^+ (from CO_2), which may contribute to the metabolic acidosis of uraemia. This renal NH_4^+ system for the generation of 'new' bicarbonate is adaptable: the activity of the glutaminase enzyme that deaminates glutamine is enhanced during acidosis (including intracellular acidosis due to chronic hypokalaemia).

The elimination of NH_4^+ in the urine occurs only after a complicated process that involves active NH_4^+ secretion in the proximal tubule, NH_4^+ reabsorption in the thick ascending limb of the loop of Henle (which can be blocked by the loop diuretic frusemide—see under urinary acidification and the diagnosis of renal tubular acidosis), and finally, NH_3 secretion by diffusion into the collecting duct. The reabsorption of NH_4^+ in the thick ascending limb leads to accumulation of NH_4^+ in the medullary interstitium, which is increased further by countercurrent multiplication (similar to generation of the corticomedullary osmotic gradient for urinary concentration). At the pH of interstitial fluid, and with a high medullary concentration of NH_4^+, some dissociates and increases the local level of NH_3. This can then diffuse into the collecting duct, where owing to the lower pH it is converted to NH_4^+ again, trapped in the tubular lumen, and excreted. This conversion maintains a concentration gradient for further NH_3 diffusion into the collecting duct ('diffusion trapping'). Anything that interferes with H^+ secretion in the collecting duct, as in distal renal tubular acidosis, would be expected to reduce diffusion trapping and thereby cause not only a higher urine pH (which depends on the presence of free H^+), but also a reduction in NH_4^+ and titratable acid (net acid) excretion; in fact some authorities advocate defining renal tubular acidosis in terms of reduced NH_4^+ excretion. The converse of this is decreased availability of NH_3 (as in hyperkalaemia, which suppresses NH_3 synthesis), but relatively normal tubular H^+ secretion per nephron. In this situation, urine pH will be low (because of less NH_3 to buffer H^+) and net acid excretion reduced. Like distal renal tubular acidosis, in chronic renal failure (uraemic acidosis), urinary excretion of NH_4^+ and titratable acid (mainly phosphate) is decreased; however in contrast to renal tubular acidosis, urine pH is usually low, and when the amounts of excreted phosphate and NH_4^+ are corrected for the reduced glomerular filtration rate (i.e. excretion per nephron), they are both normal or even increased.

What determines the presence of acidosis?

The following factors are important:

1. H^+ intake and endogenous generation;
2. Net H^+ excretion in urine = urinary [titratable acid] + $[NH_4^+]$ – $[HCO_3^-]$;
3. [titratable acid] in urine is dependent on urine pH and the amount of buffer available;
4. $[NH_4^+]$ in urine is dependent on NH_3/NH_4^+ generation and delivery (proximal tubule secretion, thick ascending loop reabsorption, and diffusion trapping in the collecting duct); and
5. $[HCO_3^-]$ in urine is dependent on urine pH and pCO_2.

Specific transport proteins differentially located on the apical and basolateral membrane of tubular cells mediate the processes described above. They are either responsible for H^+ or HCO_3^- transport. The Na^+/H^+ exchanger and H^+-ATPase, together with a small and uncertain contribution from a renal H^+/K^+-ATPase (similar to that found in the stomach) are responsible for H^+ secretion. Six isoforms of the Na^+/H^+ exchanger have been identified (NHE-1 to NHE-6), but so far none has been linked to a clinical form of renal tubular acidosis. The H^+-ATPase is not a single protein, but composed of at least nine distinct subunits. It is found mainly in the distal nephron, although it is also present in the proximal tubule and loop of Henle. The role of the H^+/K^+-ATPase in normal urinary acidification is still debated, although it seems to be upregulated along the distal nephron in potassium deficiency.

To generalize, it can be said that the main exit pathway for bicarbonate from tubular cells of the proximal tubule and loop of Henle is the Na^+-HCO_3^- cotransporter, while from cells of the distal nephron it is the Cl^-/HCO_3^- exchanger. The latter transporter can also mediate HCO_3^- secretion when present in the luminal cell membrane.

Nomenclature of the renal tubular acidoses

The 'old' numbered classification of renal tubular acidosis is a chronological one and reflects historical description of clinical disease. The numbering is often a source of great confusion to students and doctors alike, since it is not based on any functional understanding of acid excretion by the nephron. It is easier to subdivide renal tubular acidosis into proximal ('old' type II) and distal ('old' type I) nephron types (which may be inherited, or due to a variety of drugs or systemic diseases) and to consider underlying mechanisms. What is now emerging is that we can begin to subclassify these two main types according to the transport defect of acid excretion. The fact that several of the transporters involved in the renal handling of H^+ and HCO_3^- have now been cloned means that our classification of renal tubular acidosis is likely to change in the future, to one based on underlying and specific molecular transport defects.

Tests to diagnose renal tubular acidosis

From the above, it is apparent that in the presence of a normal glomerular filtration rate, renal tubular acidosis can result from (alone or in combination): (1) failure to reclaim filtered bicarbonate in the proximal tubule; (2) failure to generate and excrete NH_4^+; and (3) impaired H^+ secretion along the distal nephron. The finding of a low serum bicarbonate concentration in an appropriate clinical context should raise the possibility of renal tubular acidosis, but to make the diagnosis, specialized tests are usually required. A variety of such tests have been described, some of which we find to be more useful than others, as indicated below.

Measurement of urinary pH and HCO_3^- excretion

Dipstix urine pH values are unreliable: pH should be measured in a freshly passed specimen of urine with a glass pH electrode. The pH of a urine sample that has been left to stand, and has not been covered with a film of oil, will rise as CO_2 is lost; this is usually the reason for a measured urine pH of higher than 8, unless it is infected (high NH_4^+ from urea-splitting organisms).

Whilst a high urinary pH might lead one to diagnose renal tubular acidosis in an acidotic patient with normal glomerular filtration rate and uninfected urine, especially if they had nephrocalcinosis and/or urinary stones, the diagnosis of renal tubular acidosis cannot be based on the measurement of urine pH alone. For example, even in the presence of systemic acidosis, and excluding chronic renal failure, a low urine pH may not rule out a diagnosis of renal tubular acidosis, since renal acid excretion can still be impaired if the generation and delivery of NH_3/NH_4^+ are reduced. In addition, in proximal renal tubular acidosis, when bicarbonate reabsorption is impaired, urine pH can be low (rather than high) if previous loss of bicarbonate has lowered plasma $[HCO_3^-]$ to such a level (less than 18 mmol/l) that the filtered load of bicarbonate is then decreased, resulting in low urinary bicarbonate excretion. In this situation, failure of the proximal tubule to reabsorb bicarbonate can only be demonstrated by showing a high fractional (as a proportion of the amount of filtered bicarbonate) excretion (urinary $[HCO_3^-]$/plasma $[HCO_3^-]$ × plasma

[creatinine]/urinary [creatinine]) × 100 per cent), although to do this plasma [HCO_3^-] must be increased to approximately 24 mmol/l by intravenous infusion of bicarbonate. Fractional bicarbonate excretion is normally less than 5 per cent, but in proximal renal tubular acidosis it exceeds 15 per cent. Ammonium generation can also be reduced in proximal renal tubular acidosis (as part of more generalized proximal tubular dysfunction), which may be another reason for the finding of a low urine pH.

Urinary net negative charge and osmolal gap

The difficulty in interpretation of urinary pH in isolation is why some have argued for the calculation of the urinary net charge, or anion gap ([Na^+] + [K^+] – [Cl^-]), which is usually negative (because Cl^- is balanced by unmeasured NH_4^+), to estimate urinary NH_4^+concentration indirectly. Because a non-renal cause of metabolic acidosis will increase NH_4^+ generation and excretion (making the net charge more negative), reduced NH_4^+ excretion is evident as a positive net charge, which would indicate a renal cause for acidosis, including uraemic acidosis (chronic renal failure). However, an increase in the excretion of ketoacid anions (such as in diabetic ketoacidosis) will also make the net charge positive, despite increased NH_4^+ excretion, because excreted NH_4^+ accompanies unmeasured ketoacid anions (A^-), rather than measured Cl^-.

To get around the problem of unmeasured anions, calculation of the urinary osmolal gap (urine osmolality-(2 × ([Na^+] + [K^+]) + [urea] + [glucose])) has been proposed. Assuming that NH_4^+ and its accompanying anion constitute the main unmeasured osmotically active particles in the urine, then urinary [NH_4^+] is approximately half the osmolal gap. A low osmolal gap suggests low NH_4^+ excretion and thus a defect of renal acidification. However, to determine where the defect might lie, urine pH must still be measured: if it is high (more than 6), there may be reduced distal H^+ excretion or increased distal delivery of HCO_3^-, that is, distal or proximal renal tubular acidosis, respectively; if urine pH is low (less than 5), there may be chronic renal failure (low glomerular filtration rate) or hyperkalaemia (inhibiting NH_4^+ generation); an intermediate urine pH of 5 to 6 is said to indicate renal interstitial disease.

However, despite the sound theory, there are too many confounding variables, and both estimation of urinary net negative charge and osmolal gap fail to distinguish reliably between the acidosis of chronic renal failure and renal tubular acidosis. In our opinion these measurements often seem confusing and are of limited practical value in the diagnosis of renal tubular acidosis.

Urinary pCO_2/blood difference

Since the distal nephron H^+ secretory mechanism is intact in proximal renal tubular acidosis, measurement of the urine–blood pCO_2 (U–B pCO_2) difference has been proposed as a means of distinguishing proximal from distal renal tubular acidosis. In the presence of adequate bicarbonaturia and alkaline urine (pH greater than 7.4; intravenous bicarbonate may be necessary to achieve this) this difference is high (around 30 mmHg (4 kPa) or more), whereas it is low (less than 25 mmHg) when distal H^+ secretion is defective. This test is also said to distinguish between reduced distal H^+ secretion due to a primary secretory ('pump') defect and that due to increased cell membrane permeability and backleak of H^+ (see Table 5). Whilst there are theoretical reasons why this may be so, our incomplete understanding of the factors that determine urinary pCO_2 make interpretation difficult; moreover, in practice the test is not easy to perform and results can be variable. Like the urinary anion and osmolal gaps, we do not find it clinically useful.

Acid loading

The most straightforward and reliable means of making the diagnosis of impaired distal nephron acidification is an acid load test. The easiest and best-established method is the short oral ammonium chloride (0.1 g/kg) test of Wrong and Davies. The criterion for a diagnosis of a distal acidification defect is a failure to lower urine pH to less than 5.5, making measurements as urine is passed for at least 6 h (and up to 8 h) after the ingestion of NH_4Cl.

A simpler test, which compares well with the NH_4Cl test in normal subjects, although it has not been formally validated in patients with renal tubular acidosis, is the frusemide/fludrocortisone test. In this much more palatable test, a single dose of fludrocortisone (1 mg) is given orally, followed 1 h later by oral frusemide (40 mg), and urine pH measured for up to 5 h; again the threshold pH is 5.5. Frusemide works by enhancing H^+ secretion through increased delivery of Na^+ (from the thick ascending limb) to the distal tubule and collecting duct, where increased Na^+ reabsorption facilitates H^+ secretion; fludrocortisone directly stimulates both distal Na^+ reabsorption and H^+ secretion. Frusemide may also promote NH_4^+ excretion by blocking its reabsorption in the thick ascending limb, which could be used to demonstrate that impaired NH_4^+ generation *per se* is not the primary cause of reduced distal acidification. In distal renal tubular acidosis secondary to a defect of H^+ secretion, frusemide should still increase NH_4^+ excretion, although it may be less than normal.

The clinical features of renal tubular acidosis

A hyperchloraemic (normal anion gap) metabolic acidosis is a feature of both proximal and distal forms of renal tubular acidosis.

Proximal renal tubular acidosis

As already mentioned, urine pH in this form of renal tubular acidosis (also known as type 2) is often within the normal range; plasma bicarbonate concentration can range between 10 and 20 mmol/l. This type of renal tubular acidosis is uncommon and an isolated failure to reabsorb bicarbonate is unusual. Most cases are associated with other proximal tubular transport defects, as in the Fanconi syndrome, which may have many causes (see Table 5). Thus, it is likely that other clues to a primary abnormality of proximal tubular function will be present, such as glycosuria, hyperphosphaturia (hypophosphataemia), hyperuricosuria (hypouricaemia), aminoaciduria, and tubular (low molecular weight, e.g. retinol-binding protein) proteinuria. Osteomalacia (rickets and growth retardation in children) is common and due to impaired synthesis of active metabolites of vitamin D and urinary loss of 25-OH vitamin D.

Although hypercalciuria and hyperphosphaturia occur, nephrocalcinosis or urinary stones are rare, probably because of the associated increase in urinary citrate excretion. However, nephrocalcinosis is sometimes seen as a consequence of treatment with vitamin D, calcium, and phosphate supplements, and is often attributable to periods of iatrogenic hypercalcaemia.

Hypokalaemia is common and probably a consequence of osmotic diuresis due to reduced solute reabsorption (mainly bicarbonate) in the proximal tubule, leading to increased flow rate in the distal nephron, the site of potassium secretion. For this reason, bicarbonate supplementation tends to exacerbate hypokalaemia.

The molecular basis of a particular form of inherited non-cystinotic Fanconi syndrome, now known as Dent's disease, has been defined. This condition is X-linked (Xp11.22) and recessive, typically (about 80 per cent) presenting with nephrocalcinosis and urolithiasis, hypercalciuria, and progressive renal failure. A proportion of cases are initially thought to have a proximal or distal type of renal tubular acidosis. The reason for is that the underlying defect is a mutation of an intracellular membrane-associated chloride channel (ClC-5) found in cells of the proximal tubule, thick ascending loop, and collecting duct (a-intercalated cells) (Fig. 5). This channel appears to be involved in normal endosomal function (acidification), which affects both the reabsorption of low-molecular weight-proteins in the proximal tubule and perhaps the normal turnover and recruitment (recycling) to the luminal membrane of transport proteins, like the H^+-ATPase of the a-intercalated cell. Tubular proteinuria is a characteristic feature, distinguishing it from distal renal tubular acidosis proper.

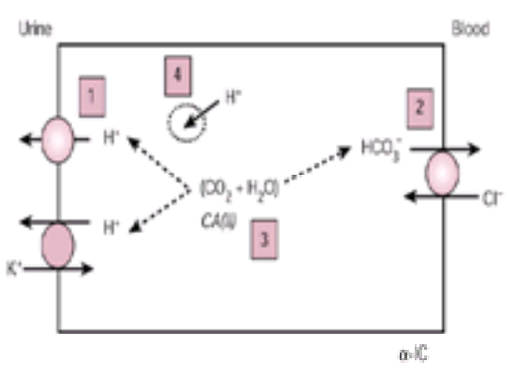

Fig. 5 Diagram of the a-intercalated cell (a-IC) of the collecting duct, showing the sites of ion transport defects that have been linked to distal renal tubular acidosis (dRTA). In boxes: **1**, H^+-ATPase mutations in recessive dRTA with and without deafness; **2**, Cl^-/HCO_3^- exchanger mutation in dominant (and recessive) dRTA; **3**, CA-II (carbonic anhydrase) deficiency in inherited mixed pRTA and dRTA; **4**, Cl^- channel (ClC-5) mutation in Dent's disease (pRTA/Fanconi with some dRTA features).

Because the proximal tubule is the dominant site of bicarbonate reabsorption, and this is ultimately dependent on CA-II activity, lack of this enzyme produces a predominantly proximal form of renal tubular acidosis, although it is widely expressed along the nephron. This type of proximal renal tubular acidosis is very rare and associated with increased bone mineralization (osteopetrosis), deafness, and cerebral calcification. Interestingly, a recent publication describing a *CA-II* 'knockout' mouse with renal tubular acidosis (but without bicarbonaturia) demonstrated that the condition could be partially corrected by gene therapy (retrograde injection of *CA-II* cDNA in liposomes), which restored *CA-II* function to the distal nephron.

Distal renal tubular acidosis

The most common form of renal tubular acidosis is 'classic' (or type 1) distal renal tubular acidosis and is due to an underlying defect of H^+ secretion ('pump defect'). The key feature of this form of renal tubular acidosis is an inability to lower urine pH below 5.5. In the absence of systemic acidosis, a patient who cannot reduce their urine pH to less than 5.5 in the face of an acid load is said have the 'incomplete' form. In patients with 'complete' distal renal tubular acidosis, plasma bicarbonate concentration usually ranges between 15 and 20 mmol/l, but can fall to less than 10 mmol/l in severe cases. In patients with 'incomplete' distal renal tubular acidosis, NH_4^+ excretion is normal (in relation to glomerular filtration rate) and this may be why these patients are not acidotic and their plasma bicarbonate concentration stays within the normal range. The causes of distal renal tubular acidosis are listed in Table 5.

Hypokalaemia can occur, especially in the autoimmune form (such as Sjögren's syndrome) and when there is systemic acidosis, but it is not as consistent a feature as it is in proximal renal tubular acidosis. Nephrocalcinosis is common and present in 70 to 80 per cent of adults, whether or not they have a systemic acidosis. Osteomalacia (rickets in children) is seen only in acidotic patients with 'complete' distal renal tubular acidosis. Hypercalciuria is said to be a feature of this form of renal tubular acidosis, but it is not a consistent finding and may depend on the presence of systemic acidosis and resulting calcium loss from bone.

A characteristic of distal renal tubular acidosis, which can be useful in screening for this condition in children, is a low urinary excretion of citrate, which in alkaline urine is normally increased. The probable explanation for reduced citrate excretion in distal renal tubular acidosis is that it is reabsorbed in the proximal tubule by a mechanism dependent on intracellular pH: intracellular acidosis (due to systemic acidosis or chronic hypokalaemia) increases mitochondrial metabolism of citrate and thus its reabsorption. In addition, trivalent citrate buffers H^+ in systemic acidosis and is converted to the more readily reabsorbed divalent form. Because it inhibits the precipitation of calcium salts (mainly phosphate), reduced citrate delivery to the distal nephron is a key factor in the development of nephrocalcinosis and urinary stone disease in distal renal tubular acidosis.

The main site of H^+ secretion along the distal nephron is the H^+-ATPase in the luminal cell membrane of the a-intercalated cell (Fig. 5). As already described, H^+ and HCO_3^- are generated from intracellular hydration of CO_2, the latter exiting the basolateral cell membrane in exchange for Cl^- by an anion exchanger (AE1). This basolateral exchanger is essential for normal apical H^+ secretion in the a-intercalated cell and it has been established that a mutation of the gene encoding this transport protein is the cause of dominantly inherited distal renal tubular acidosis and some cases of the recessive form without deafness. Recessive distal renal tubular acidosis with deafness is due to a mutation of the B1 subunit of the H^+-ATPase (and recently another H^+-ATPase subunit mutation in some cases without deafness). However, unlike the dominant form, the recessive form appears to be genetically heterogeneous. Expression of both these transport proteins is decreased in the collecting duct a-intercalated cells of patients with autoimmune distal renal tubular acidosis.

A rare drug-related form of distal renal tubular acidosis with hypokalaemia is that due to amphotericin B. This drug accumulates in the kidney and increases the permeability of the tubular cell membrane, resulting in a back-leak of secreted H^+ ('permeability' or 'gradient' defect) and failure to maintain the plasma–urine pH gradient across the tubular epithelium.

Distal renal tubular acidosis with hyperkalaemia is also known as type IV renal tubular acidosis, or voltage-dependent renal tubular acidosis. It occurs in situations in which the lumen negative potential difference along the distal nephron (Fig. 4) is reduced. This normally facilitates potassium and hydrogen ion secretion, depends on sodium reabsorption, and is stimulated by aldosterone. Thus, when sodium reabsorption through the epithelial Na^+ channel is reduced by drugs like amiloride and trimethoprim, or when the Na^+ channel is inactive (autosomal recessive pseudohypoaldosteronism type 1a), there is a decrease in both potassium and hydrogen ion secretion, leading to hyperkalaemia and metabolic acidosis. Because of the stimulatory effect of aldosterone, this form of renal tubular acidosis is also seen in states of hypoaldosteronism.

Another rare and inherited form of hyperkalaemic renal tubular acidosis is autosomal dominant pseudohypoaldosteronism type II, or Gordon's syndrome. The transport defect underlying this condition remains unknown, although it has been proposed that an abnormal increase in collecting duct permeability to Cl^- is responsible, leading to a decrease in the lumen negative potential difference. In contrast to patients with the type 1 variant of pseudohypoaldosteronism, these patients are usually hypertensive, and their blood pressure and hyperkalaemia are particularly responsive to thiazide diuretics.

In all forms of hyperkalaemia, renal NH_3/NH_4^+ generation is reduced, which also impairs acid excretion and exacerbates systemic acidosis.

Treatment

Administration of oral bicarbonate, which can also be given as citrate, is the mainstay of treatment in all forms of renal tubular acidosis, unless the underlying defect can be corrected (for instance drug withdrawal or aldosterone replacement). It can prevent the long-term complications of rickets and growth retardation, and may also limit the progression of nephrocalcinosis and ephrolithiasis in distal renal tubular acidosis by increasing citrate excretion. In both proximal renal tubular acidosis and 'classic' distal renal tubular acidosis, a potassium supplement may also be necessary, since oral sodium bicarbonate tends to increase urinary potassium loss.

Treatment that is more specific to proximal renal tubular acidosis may include vitamin D, calcium, and phosphate supplementation; a thiazide diuretic can also be tried, so as to increase bicarbonate reabsorption (secondary to mild extracellular volume contraction), but this may also exacerbate the tendency to hypokalaemia.

In distal renal tubular acidosis, plasma bicarbonate concentration should be maintained above 20 mmol/l. Morbidity in classic distal renal tubular acidosis is due to calcium phosphate renal stones; endstage renal failure is a rare complication and is usually the result of unrecognized urinary tract obstruction and recurrent infection. In hyperkalaemic distal renal tubular acidosis, a loop diuretic plus fludrocortisone, or an ion exchange resin, can be used to control the hyperkalaemia, and improve NH_3/NH_4^+ generation and thus acid excretion.

A seeming contradiction in the treatment of distal renal tubular acidosis with nephrocalcinosis and urolithiasis is the use of alkali therapy. On the one hand, this treatment will correct acidosis-related hypercalciuria and increase urinary citrate excretion, which should reduce or arrest nephrocalcinosis and stone formation. On the other hand, calcium phosphate precipitation (the main component of nephrocalcinosis and stones related to distal renal tubular acidosis) is favoured by an alkaline urine pH, and if bicarbonate is given as the sodium salt, this will also tend to increase urinary calcium excretion. It is possible that these opposing effects are why amelioration of nephrocalcinosis is often not seen in treated patients followed long-term.

Further reading

Urolithiasis and nephrocalcinosis

Coe FL *et al.*, eds. (1996). *Kidney stones, medical and surgical management*. Lippincott-Raven, Philadelphia.

Lingeman JE *et al.*, eds. (1989). *Urinary calculi*. Lea & Febiger, Philadelphia.

Pak CYC (1998). Kidney stones. *Lancet* **351**, 1797–801.

Robertson WG (1992). Factors involved in stone-formation. In: Cameron S *et al.*, eds. *Oxford textbook of nephrology*, 1st edn, pp 1822–46. Oxford University Press, Oxford.

Robertson WG (1993). Urinary tract calculi. In: Nordin BEC, Need AG, Morris HA, eds. *Metabolic bone and stone disease*, 3rd edn, pp 249–311. Churchill Livingstone, Edinburgh.

Wickham JEA, Buck AC, eds (1990). *Renal tract stone, metabolic basis and practice*. Churchill Livingstone, Edinburgh.

Wrong O (1998). Nephrocalcinsosis. In: Cameron S *et al.*, eds. *Oxford textbook of nephrology*, 2nd edn, pp 1375–96. Oxford University Press, Oxford.

Renal tubular acidosis

Alpern RJ (1990). Cell mechanisms of proximal tubule acidification. *Physiological Reviews* **70**, 79–114.

Bruce LJ *et al.* (1997). Familial distal renal tubular acidosis is associated with mutations in the red cell anion exchanger (band 3, *AE1*) gene. *Journal of Clinical Investigation* **100**, 1693–707.

Capasso G *et al.* (1986). Amphotericin B and amphotericin B methylester: effect on brush border membrane permeability. *Kidney International* **30**, 311–17.

Capasso G *et al.* (1994). Acidification in mammalian cortical distal tubule. *Kidney International* **45**, 1543–54.

Cohen EP *et al*. (1992). Absence of H(+)-ATPase in cortical collecting tubules of a patient with Sjögren's syndrome and distal renal tubular acidosis. *Journal of the American Society of Nephrology* **3**, 264–71.

Halperin ML, Vasuvattakul S, Bayoumi A (1992). A modified classification of metabolic acidosis: a pathophysiologic approach. *Nephron* **60**, 129–33.

Karet FE *et al.* (1999). Mutations in the gene encoding B1 subunit of H+-ATPase cause renal tubular acidosis with sensorineural deafness. *Nature Genetics* **21**, 84–90.

Lai LW *et al.* (1998). Correction of renal tubular acidosis in carbonic anhydrase II-deficient mice with gene therapy. *Journal of Clinical Investigation* **101**, 1320–5.

Lloyd SE *et al.* (1996). A common molecular basis for three inherited kidney stone diseases. *Nature* **379**, 445–9.

Wrong O (1991). Distal renal tubular acidosis: the value of urinary pH, pCO_2 and NH_4^+ measurements. *Pediatric Nephrology* **5**, 249–55.

Wrong O, Davies HEF (1959). The excretion of acid in renal disease. *Quarterly Journal of Medicine* **28**, 259–313.

20.14 Urinary tract obstruction

L. R. I. Baker

Introduction

If the flow of urine is impeded at any point in its course from the renal calices to the exterior, urinary tract obstruction is present. The terms 'obstructive uropathy', 'obstructive nephropathy', and 'hydronephrosis' are frequently used interchangeably and are taken to have the same meaning as the term 'urinary tract obstruction'. A more rigorous approach is preferable: 'obstructive uropathy' should be taken to mean pathological change occurring in the urinary tract and kidney consequent upon urinary tract obstruction. 'Obstructive nephropathy' is present when pathological change in the kidney resulting from urinary tract obstruction is associated with prolongation of the transit time of glomerular filtrate down the nephron. The term 'hydronephrosis' should be taken to denote dilatation of the renal pelvis and calyceal system.

Intraluminal obstruction between the commencement of the proximal tubule and the distal end of the collecting duct, such as occurs in uric acid nephropathy, as a result of sulphonamide crystal deposition, and in multiple myelomatosis, falls outside the definition of urinary tract obstruction employed here and will not be considered further.

Although dilatation of the outflow system proximal to the site of obstruction is a characteristic finding, widening of the ureter and/or pelvicalyceal system does not necessarily indicate the presence of obstruction. Causes of such anatomical abnormality in the absence of obstruction are listed in Table 1.

Obstruction may be partial or complete, unilateral or bilateral. Bilateral obstruction, or obstruction of a single kidney, is a greater threat to the patient than unilateral obstruction. Obstruction associated with infection is a greater threat to kidney function and to life than obstruction in the absence of infection. Since it is common, and often reversible, obstruction of the urinary tract should be considered in every uraemic patient, whether acute or chronic.

Incidence

Urinary tract obstruction occurs most frequently in the young and the old. Hydronephrosis is the most common cause of an abdominal mass in neonates, and obstruction, usually due to congenital abnormalities, is also relatively common in children. Its incidence declines after the age of 10 and is at its lowest in middle age, but begins to rise again after the age of 60, particularly in males, in whom the commonest cause is prostatic enlargement. Although the overall frequency of urinary tract obstruction is the same in both sexes, between 20 and 60 years of age it is more frequent in women, and over the age of 60 years the reverse is true. Urinary tract obstruction has been found in 3.8 per cent of a large series of routine autopsies and 25 per cent of autopsies carried out upon uraemic patients.

Causes

Obstructing lesions may lie within the lumen or the wall of the urinary tract, or may cause obstruction by pressure from outside, the major causes in each group being listed in Table 2.

Calculi and neuromuscular malfunction at the junction of the renal pelvis and ureter are common causes of unilateral obstruction. Prostatic obstruction, stone disease, and bladder tumours account for approximately 75 per cent of cases of bilateral obstruction in developed countries. Wide geographical variations occur in the relative incidence of some causes of obstruction, for example schistosomiasis. To the clinician, the most important questions are whether urinary tract obstruction affects the upper or the lower urinary tract, and whether it is of recent onset (acute obstruction) or is long-standing (chronic obstruction).

Pathophysiology

Acute upper tract obstruction

Urine flows from the kidney to the bladder as a result of ureteric and pelvic peristalsis, the effects of gravity, and the pressure of glomerular filtration. Peristalsis normally generates high pressures within the ureteric lumen, sufficient to propel urine down the ureter without the transmission of the increased pressure to the renal parenchyma. Initially, an upward movement occurs in the ureter: thereafter, proximal contraction of ureteric circular muscle, with eventual complete occlusion of the lumen, forms a bolus of urine. Contraction of longitudinal smooth muscle then propels the bolus along the ureter. Baseline ureteric pressure is similar to that in the renal pelvis, but during this process rises to values between 10 and 25 mmHg. These pressures are not transmitted to the renal pelvis, where pressure seldom rises above 4 mmHg.

Any change in blood flow or glomerular filtration rate resulting from ureteric obstruction would have important effects on tubular pressures and flows. In humans, the time of onset of obstruction is seldom known with any precision, and methods of measurement of renal blood flow or filtration rate using clearance techniques are indirect and depend upon tubular function, which is affected by urinary tract obstruction; hence pathophysiological explanations must depend on animal experiments. In the normal dog, pressure within the ureter more than doubles when the ureteric lumen is occluded during peristalsis, and similar changes occur in ureteric wall tension. Baseline and peak pressure and wall tensions are about twice as high as control values 3 min after acute ureteric obstruction; at 5 to 20 min they

approximate to peak values; and at 1 h there is a threefold increase. At this point, occlusion of the ureter fails to occur and pressures generated by ureteric wall tension are transmitted to the renal pelvis and parenchyma. Any further increase in pressure results in dilatation of the ureter.

The effect of an increase in pressure within the ureter transmitted to the nephron depends upon the degree of obstruction (whether complete or incomplete), whether obstruction is unilateral or bilateral, and the duration of obstruction. In the dog, renal blood flow falls to 50 per cent of control values 3 or 4 days after induction of complete ureteric obstruction and at 4 weeks it is about one-third that of the contralateral unobstructed kidney. Three phases, which are not well understood, are discernible in the relationship between changes in ureteral pressure and renal blood flow with time (Fig. 1). Phase I occurs during the first hour after induction of obstruction. Renal blood flow increases, presumably owing to a reduction in intrarenal vascular resistance, associated with a gradual increase in ureteric pressure. In phase II, which takes place over the next 2 to 5 h, ureteric pressure continues to rise and renal blood flow begins to fall. Thereafter, in phase III, renal blood flow continues to fall and ureteric pressure returns towards or to normal.

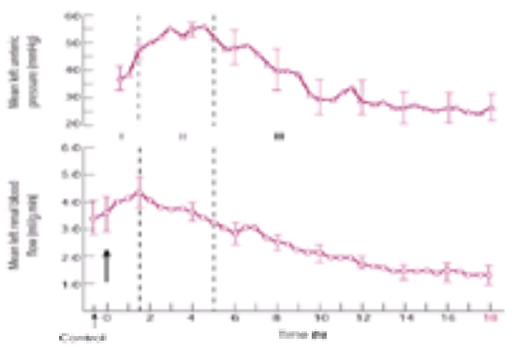

Fig. 1 The relationship between ipsilateral renal blood flow and ureteric pressure during experimental ureteric occlusion in the dog. In phase I, renal blood flow and ureteric pressure rise. In phase II, blood flow declines while ureteric pressure continues to rise. In phase III, both renal blood flow and ureteric pressure decline. The arrow indicates the time of ureteric occlusion. Mean ± standard error; $n = 5$. (Reprinted with permission from Moody *et al.* 1975.)

Chronic upper tract obstruction

Three months after the production of experimental obstruction in dogs, baseline ureteric wall tension is increased and there is no difference between baseline and peak values of wall tension, the latter being measured during ureteric occlusion. By contrast, baseline and peak pressures within the ureteric lumen are not significantly different from control values. This is a consequence of the relationship between pressure and wall tension expressed in Laplace's law, which states that $P = K(T/R)$, where P is the transluminal pressure, K is a constant, T is wall tension, and R is the radius of the ureter. In chronic obstruction, therefore, normal intraluminal pressures are maintained as a consequence of ureteric dilatation.

These experimental findings suggest that the major component of damage to the kidney due to obstruction occurs soon after its onset. In humans the highest measured ureteric pressures have been found during acute obstruction (as high as 50 mmHg during passage of a stone) and there appears to be an inverse relationship between pressure within the renal pelvis and time of measurement in patients with complete obstruction. The notion that chronic obstruction with dilatation of the ureter may be relatively benign is supported by the observation that patients with incomplete urinary tract obstruction due to congenital anomalies lose renal function only slowly.

Acute lower tract obstruction

The mechanical efficiency of smooth muscle fibres is reduced when they become overstretched. As obstruction to bladder outflow increases, a point is reached when acute urinary retention will result. Factors which may precipitate acute retention include a sudden diuresis, such as occurs after diuretic therapy (particularly loop agents) for heart failure, urinary infection, and drugs that have pharmacological effects upon the bladder, provoking retention, such as those with antimuscarinic and calcium channel blocking activity.

Chronic lower tract obstruction

In adults, chronic obstruction of the outflow from the bladder is most commonly due to benign prostatic hypertrophy. Prostatic malignancy and urethral strictures may also be responsible. In children, posterior urethral valves and urethral strictures are most often the cause. Such organic causes are easy to understand, but functional obstruction may occur at the bladder neck and at the level of the distal sphincter owing to a failure of co-ordination between bladder contraction and sphincter relaxation. The bladder wall may become increasingly compliant or non-compliant: this is of significance to the urologist, since patients with poorly compliant bladders fare much better after removal of the obstruction than those with highly compliant bladders. The highly compliant bladder tends not to be associated with upper tract dilatation, whereas the high pressure that exists within a bladder of low compliance may be transmitted to the upper tracts and may be the cause of renal impairment, which on occasion will be severe.

Histopathological changes

Acute obstruction results in increased ureteric pressure and decreased renal blood flow, and may be complicated by bacterial infection. The rise in intraluminal pressure and dilatation of the system proximal to the site of obstruction both result in compression of the renal substance. In the early phase of obstruction, the kidney becomes oedematous and haemorrhagic; tubular dilatation initially affects mainly the collecting duct and distal tubular segments; and Bowman's space may be dilated. The ducts of Bellini are first affected by dilatation of the system proximal to the site of obstruction. Subsequently, other papillary structures are affected, and ultimately compression of renal cortical tissue occurs with thinning of the renal parenchyma. Enlargement of the kidney occurs in association with dilatation of the renal pelvis.

Atrophy of the renal parenchyma with reduction in size of the kidney (obstructive atrophy) is believed to result from the effects of compression of the renal parenchyma and from prolonged renal ischaemia. Slowly progressive increasing resistance to outflow tends to result in gross dilatation of the collecting system, dilated calyces, and the renal pelvis being surrounded by only a thin rim of renal parenchyma. In acute complete obstruction dilatation tends to be less marked.

In patients with long-standing obstruction there is flattening of the renal tubular epithelium, periglomerular fibrosis, interstitial fibrosis, and mononuclear cell infiltration. These changes are thought to result in part from renal ischaemia and in part to reflect the effects of bacterial infection (Fig. 2).

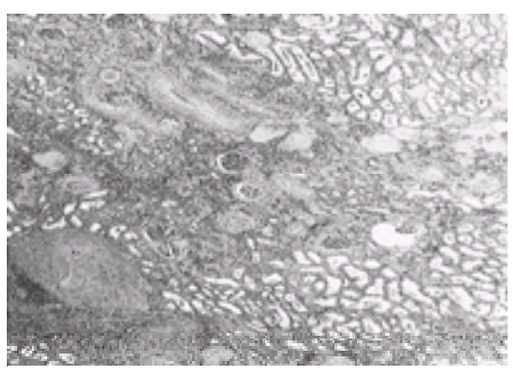

Fig. 2 Histological appearances in long-standing obstruction. Note dilated tubules, interstitial fibrosis, vessel wall thickening, and global sclerosis of some glomeruli.

Effects of obstruction upon renal function

Little detailed information is available about the effects of urinary tract obstruction on the glomerular filtration rate in humans, but it is clear that ureteric obstruction results in a marked fall in glomerular filtration rate, and that incomplete bilateral obstruction causes progressive renal failure. Glomerular filtration must continue to some extent even after development of complete acute obstruction since a nephrogram (albeit a delayed one) can be obtained after injection of intravenous contrast medium during urography.

Distal tubular function is more strikingly disturbed than proximal tubular function in chronically increased resistance to outflow, as would be expected from the histopathological findings. A characteristic feature of patients with such pathology is an impaired ability to concentrate urine. The concentration defect is resistant to administration of antidiuretic hormone and is thus an example of nephrogenic diabetes insipidus. Such patients may present with polyuria, dehydration, and hypernatraemia secondary to a reduction in free water reabsorption. Animal experiments indicate that production of cyclic AMP in response to vasopressin is reduced in chronic partial obstruction and this may, in part, explain the concentration defect. The extent to which the need to excrete an increased solute load in bilateral ureteric obstruction contributes to the concentration defect is unclear. Chronically increased bilateral resistance to outflow in humans may be associated with a salt-losing state, although the frequency with which this occurs has not been defined.

Since ureteric obstruction preferentially affects distal segments of the nephron, an acidification defect is associated with chronically increased resistance to outflow in humans. In many patients with obstructive nephropathy, urinary pH is inappropriately high for any associated degree of metabolic acidosis. This distal renal tubular acidosis is present in both unilateral and bilateral ureteric obstruction, and may be associated with hyperkalaemia.

Renal function after relief of obstruction

The relationship between duration of obstruction and recovery of function has been defined in experimental animals. In dogs subjected to unilateral ureteric ligation the ipsilateral glomerular filtration rate was 25 per cent of the ipsilateral control value and 16 per cent of that concurrently measured in the contralateral kidney when a ligature causing complete obstruction was removed after 1 week. This discrepancy resulted from a compensatory increase in function of the non-obstructed kidney during the week of complete obstruction of its partner. Improvement in the glomerular filtration rate of the previously obstructed kidney continued up to, but not beyond, 2 months after release of the obstruction, but complete recovery never occurred, maximum improvement being to only 50 per cent of the glomerular filtration rate of the non-obstructed kidney.

In humans, renal blood flow increases after relief of obstruction and the glomerular filtration rate either remains the same or increases, but no large study has been performed in which the duration of obstruction has been correlated with the degree of recovery of glomerular filtration rate. However, there is no reason to doubt that the extent of recovery depends upon whether the resistance to outflow is mild or severe, the duration of obstruction, and whether or not obstruction is complicated by bacterial infection.

After relief of prolonged bilateral obstruction or obstruction of a functionally or anatomically single kidney, there may follow a pronounced salt and water diuresis and kaliuresis, requiring appropriate oral or intravenous replacement. This is usually attributed to damage to the renal tubules and collecting ducts induced by the resistance to outflow, resulting in failure of sodium, water, and potassium conservation. However, this cannot be the sole explanation because a salt, water, and potassium diuresis is not seen as a clinical problem in humans following relief of unilateral obstruction. The osmotic diuretic effect of uraemia, hypervolaemia owing to salt and water retention, and perhaps retention in renal failure of natriuretic factors may play a part.

Hormonal changes induced by obstruction

Erythropoietin

Levels of erythropoietin are low in humans with renal failure due to obstructive uropathy, but neither the degree of anaemia nor the degree of depression of erythropoietin concentration is known to differ from that occurring in chronic renal failure of similar severity and different aetiology.

Erythraemia is a recognized association of chronic upper urinary tract obstruction and correction after relief of obstruction has been recorded. Erythropoietin concentrations in such patients have rarely been documented.

Vitamin D metabolism

Anatomical considerations suggest that renal 1a-hydroxylase activity might be particularly severely affected in chronic obstruction, but scant data are available in respect of levels of vitamin D metabolite and vitamin D metabolism in renal failure associated with obstruction, compared with renal failure of similar degree but of different aetiologies. There is an impression that osteomalacia may be more common in patients with chronic renal failure due to obstruction, but the fact that such renal failure is very slowly progressive may account for this. Among patients about to start dialysis, radiological evidence of hyperparathyroid bone disease is most common in those whose renal failure is a consequence of obstruction, even when a correction is applied for duration of renal failure and gender.

Hypertension and the renin–angiotensin system

Hypertension is more common in patients with bilateral urinary tract obstruction than in normal individuals matched for age and sex. The prevalence of hypertension resulting from unilateral obstruction is unknown.

An increase in total exchangeable sodium has been demonstrated in chronic bilateral upper tract obstruction, and blood pressure frequently returns to normal with correction of obstruction. Patients of this sort appear to have a volume-dependent form of hypertension consequent upon salt and water retention. Concentrations of renin in renal and peripheral veins are normal.

Patients with chronic unilateral ureteric obstruction and hypertension have been described in whom renal vein renin concentrations were elevated on the side of obstruction and in whom both blood pressure and renal vein renin concentration returned to normal after the relief of obstruction, but there are no clinical features or preoperative investigations that will predict outcome.

Atrial natriuretic peptide

Release of atrial natriuretic peptide is augmented in patients with bilateral ureteric obstruction and uraemia, probably owing to salt and water overload, and this may contribute to the postobstructive diuresis and natriuresis which occurs after surgical correction of the problem.

Clinical features

Acute upper tract obstruction

Typically, this gives rise to pain in the flank that may radiate to the iliac fossa, inguinal region, testis, or labium. The pain may be dull or sharp, intermittent or persistent, though waxing and waning in intensity. A high fluid intake, alcohol, or diuretics—all measures that increase urinary volume and distend the collecting system—may provoke it, which is particularly noticeable when obstruction occurs at the pelviureteric junction. Loin tenderness may be detected and an enlarged kidney felt. Upper urinary tract infection with malaise, fever, and symptoms and signs of septicaemia may dominate the clinical picture.

Complete anuria is strongly suggestive of complete bilateral obstruction or complete obstruction of a single kidney. The differential diagnosis includes bilateral total renal cortical necrosis, acute anuric glomerulonephritis, and bilateral renal arterial occlusion. Intermittent anuria indicates the presence of intermittent complete

obstruction.

Chronic upper tract obstruction

Patients may present with flank or abdominal pain, renal failure, or both, and the symptoms and signs of urinary tract infection and septicaemia may be superimposed. Rarely, presentation is with erythraemia or hypertension and their complications. Some patients are asymptomatic, obstruction being found during investigation of another condition.

Polyuria often occurs when there is chronic resistance to outflow owing to impairment of the concentrating capacity of the renal tubules. Intermittent anuria and polyuria indicate intermittent complete and partial obstruction.

Acute lower tract obstruction

Acute urinary retention is often preceded by a history of symptoms of obstruction of bladder outflow. It is typically associated with severe suprapubic pain, but this may be absent if acute retention is superimposed on chronic retention or if there is an underlying neuropathy. A potential clinical pitfall is failure to recognize that patients who have had an epidural anaesthetic may develop painless acute retention of urine. Most modalities of bladder sensation are mediated via sacral parasympathetic nerves. The pain from overdistension of the bladder is sympathetically mediated and will be abolished by a high epidural reaching to D10. Obstetricians need to be particularly alert to this problem.

Pre-existing obstruction may have provoked changes in the bladder, such as muscle wall hypertrophy, sacculation, and diverticulum formation; these in turn predispose to persistence of lower urinary tract infection once acquired and occasionally to bladder stones. Epididymo-orchitis may occur.

Chronic lower tract obstruction

Symptoms may be minimal or may be accepted by the patient as within normal limits. Hesitancy, narrowing, and diminished force of the urinary stream, terminal dribbling, and a sense of incomplete bladder emptying are typical features. If a large volume of residual urine remains in the bladder after urination, the frequent passage of small volumes of urine may be a prominent symptom even in the absence of infection. Incontinence of such small volumes of urine is termed overflow incontinence or retention with overflow.

There are no pathognomonic clinical features that differentiate high-pressure and low-pressure chronic retention. In each the bladder may be palpably distended if the volume of residual urine is sufficient. The size and consistency of the prostate is variable.

Acute complete retention of urine, usually with severe suprapubic and perineal pain, may complicate chronic retention and is commonly precipitated by lower urinary tract infection. Frequency, urgency, urge incontinence, dysuria, strangury, suprapubic pain, haematuria, and cloudy, smelly urine may be present. Asymptomatic bacteriuria is common.

Examination of the abdomen and genitalia, and rectal and vaginal examination are essential. It should be noted that the apparent size of the prostate is a poor guide to the presence of prostatic obstruction. Median lobe enlargement of a palpably normal prostate may give rise to severe obstruction, whereas an apparently grossly enlarged gland may cause little or no obstruction.

Investigation

Acute upper urinary tract obstruction

The range and specificity of imaging techniques available for the investigation of possible upper urinary tract obstruction is continually expanding. The decision as to which method of investigation to use will be influenced by local expertise and availability, but new developments seem certain to alter the optimum plan of investigation in the future. At present ultrasonography, excretion urography, and computed tomography (**CT**) scanning compete for the role of first-line investigation in this field, each having its advantages and disadvantages. Magnetic resonance urography is an alternative investigation that is currently very little used but shows considerable promise.

Ultrasonography

Ultrasonography appears to be completely safe and is relatively inexpensive. It is non-invasive, no intravenous injection, exposure to radiation, or exposure to contrast medium being involved. It has become the first-line imaging procedure in many centres, but an ultrasound report saying 'normal kidneys and urinary tract' should not be taken as meaning 'obstruction is excluded'. A relative disadvantage is its dependence on the operator: the ultrasonographer sees much more during the course of an examination than the clinician examining the few images produced. Relatively minor degrees of upper tract dilatation may be missed, indicative in some cases of clinically important obstruction. Ultrasound may visualize a few centimetres of dilated upper ureter and can show a dilated distal ureter posterior to the bladder, but it does not visualize most of the mid ureter. Calculi are easily missed: those in a dilated upper or lower ureter may be detected, but since much of the ureter is not visualized, ultrasonography cannot diagnose many ureteric calculi and hence should be accompanied by a full-length plain abdominal radiograph and, if necessary (usually owing to bowel gas overlying the renal areas), plain tomograms. Ultrasonography is less sensitive for detection of opaque renal calculi than either plain films and plain renal tomography or CT. The site and nature of obstruction often cannot be defined by ultrasonography, and upper tract dilatation on ultrasound (Fig. 3) is not synonymous with urinary tract obstruction since ultrasound cannot differentiate a baggy, low-pressure, unobstructed system from a tense, dilated, obstructed one. Ultrasound has advantages compared with intravenous urography, although not with CT, in patients with severe impairment of renal function (see below). In contrast with urography, but again not with CT, it does not exacerbate pain due to obstruction, caused in excretion urography by the diuretic effect of intravenous contrast medium employed.

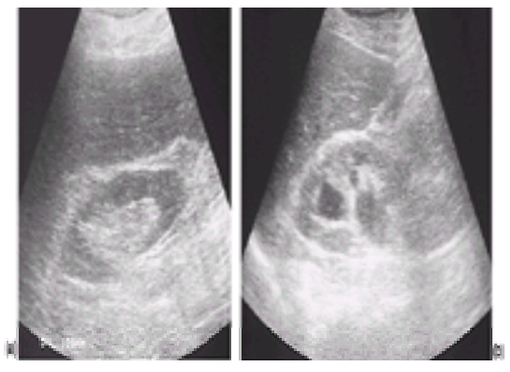

Fig. 3 Ultrasound showing (a) a normal kidney, and (b) a kidney with a dilated pelvicalyceal system in urinary obstruction.

Intravenous urography

Intravenous urography will usually demonstrate the site, cause, and degree of obstruction and is much less operator dependent than ultrasound since the number of images checked by the clinician is equal to the number of images reported by the radiologist. Its major disadvantage is that the technique carries a mortality owing to contrast hypersensitivity reaction of perhaps 1 in 200 000. It also involves an intravenous injection, exposure to radiation (of particular concern in pregnant women and children), worsening of pain due to the diuretic effect of contrast medium when an upper tract (or tracts) is obstructed, and the potential for contrast nephrotoxicity. Patients with impaired renal function, particularly those with diabetes and perhaps patients with myelomatosis, are at particular risk. Such risk is

minimized by employment of low osmolality contrast medium and the avoidance of prior dehydration. As with ultrasound, upper tract dilatation does not necessarily indicate the existence of obstruction, but this is much less often a diagnostic problem in urography than with ultrasonography because of the better ureteric visualization and the ability to assess drainage of the upper tracts. Ultrasonography has replaced high-dose urography as the first-line method for detecting obstruction in renal failure.

The initial sequence of radiographs must include sufficient films to identify calcifications in the urinary tract, a good combination being a full length and a coned renal area plain film. The plain films must be examined carefully for opaque calculi along the line of the ureter—calculi overlying bone are easily missed (Fig. 4(a) and Fig. 4(b)). Some obstructing calculi are very small and only faintly calcified or non-opaque. Ureteric calculi within the bony pelvis are often impossible to distinguish from calcified phleboliths on the plain film.

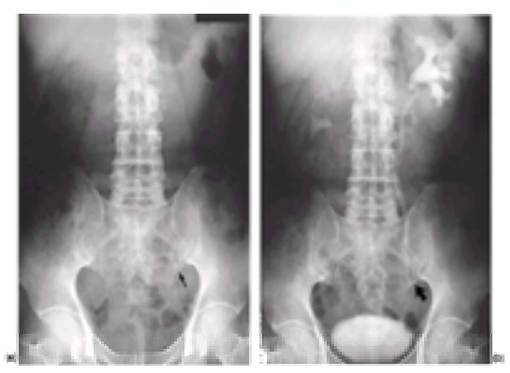

Fig. 4 (a) Plain abdominal radiograph. Opaque calculus (arrowed) medial to the left lower sacroiliac joint is easy to overlook. (b) Same patient as in (a) after contrast radiograph. Note dilatation of collecting system and ureter to the level of the calculus.

A large dose of contrast, preferably of low osmolality, should be given to those who may have acute obstruction to compensate for the lack of preparation of the patient and the probability of a low glomerular filtration rate. After contrast injection, the recently obstructed kidney is typically enlarged and smooth in outline. There is an immediate nephrogram, but the calyces and pelvis fill with contrast later than normal and the nephrogram becomes increasingly dense over time owing to the prolonged nephron transit time, which allows greater than normal concentration of contrast medium within the tubules (Fig. 5). In time, the site of obstruction may become obvious owing to dilatation of the system to the level of the block (Fig. 6). A full length film should be taken 20 min after contrast injection and after the patient has been asked to empty their bladder, since contrast in a full bladder may cause spurious upper tract dilatation and may obscure the lower end of the ureter and make it impossible to confirm that a ureter is dilated down to an opacity seen in the line of the ureter in the bony pelvis.

Fig. 5 Acute left ureteric obstruction. Note the increased density of the nephrogram and the absence of a pyelogram on the left side 15 min after injection of contrast.

Fig. 6 Same patient as in Fig. 5. A later radiograph showing a persistent dense nephrogram on the left. The pelvicalyceal system and ureter, which have now filled, are only slightly dilated due to the fact that obstruction is of very recent onset. The obstructing calculus at the left ureteric orifice is not visible.

Since contrast medium enters the pelvicalyceal system and ureter slowly, opacification of the system and ureter may never be seen in severe acute obstruction. However, in most instances filling of the pelvicalyceal system and ureter to the level of obstruction can be demonstrated on delayed films. In acute ureteric obstruction the pelvicalyceal system and ureter are typically only slightly dilated. Occasionally the only abnormality may be a ureter that remains full throughout its length to the level of the vesicoureteric junction, with this finding persisting on the full length postmicturition film. Acute obstruction is also characterized by increased excretion of contrast medium by the liver, leading to opacification of the gallbladder on delayed films.

When typical obstructive changes are present, with a ureter dilated down to a calcified opacity, diagnosis is simple. If there is an obstructive nephrogram or dilatation of the pelvicalyceal system and/or ureter but no radiodense calculus is seen, diagnosis is more difficult. If the history is of pain of recent onset, the likely diagnostic possibilities are a small low-density stone not detected by urography, recent passage of an opaque stone, a uric acid stone, acute pelviureteric junction obstruction, a blood clot, or sloughed papillae. Ultrasonography can often demonstrate small low-density stones at the vesicoureteric junction not shown by urography. The presence of a uric acid stone may be suggested by a previous history of such stones, a personal or family history of gout, or clinical circumstances associated with uric acid stone formation, such as cytotoxic drug therapy or chronic small bowel disease. Urography shows uric acid stones as lucent filling defects (Fig. 7); similar filling defects may also occur with transitional cell tumours, sloughed papillae, or blood clots. Since most ureteric stones pass spontaneously, investigation of a possible transitional cell tumour or blood clot should be delayed. If a persistent lucency is present, CT scanning may be very helpful (Fig. 8).

Fig. 7 Uric acid stones seen on intravenous urography as lucent filling defects in the collecting system on the left.

Fig. 8 Same patient as in Fig. 7. The CT scan clearly shows uric acid stones as opacities within the collecting system.

Acute idiopathic pelviureteric junction obstruction should be suspected if there is a large, soft tissue density inferomedial to the kidney on the plain film produced by the distended pelvis. This usually fills on delayed films of the urogram, with no filling of the ureter.

Clot colic is always associated with macroscopic haematuria. When it is suspected, the urogram should be repeated after 2 weeks, by which time the clot should have lysed and any underlying lucent filling defect can be seen. Such patients require further investigation to define the cause of bleeding.

Sloughed papillae result from papillary necrosis. Typically, abnormal calyces are seen in both kidneys, but papillary necrosis may occasionally be unilateral, usually as a result of a previous episode of infection associated with unilateral obstruction, especially in diabetics. Occasionally, calcified papillae may mimic stones (Fig. 9).

Fig. 9 Bilateral papillary necrosis with papillary calcification mimicking stones.

CT scanning

CT scanning carried out without the use of oral or intravenous contrast medium offers obvious safety advantages compared with intravenous urography, since no intravenous injection is required. The radiation dose is, however, significantly higher (four to 10 times) than with urography. Unenhanced CT is well established as a second-line method for detecting pelvicalyceal and ureteric dilatation when ultrasonography is non-diagnostic in patients with suspected obstruction. CT also has an established role in demonstrating the cause of obstruction when this is not shown by ultrasonography (for example ureteric calculus, retroperitoneal mass). With the advent of helical (spiral) CT, this method is being increasingly widely used, especially in the United States, in patients with suspected ureteric colic where it has a sensitivity and specificity (in expert hands) of over 95 per cent.

A major advantage of helical CT scanning over both ultrasonography and intravenous urography is that it is capable of diagnosing non-urological conditions presenting with flank pain and not associated with urinary tract obstruction, as well as conditions causing obstruction by extrinsic compression of the ureter. These include appendix abscess, diverticular perforation, torsion of an ovarian mass, a leaking abdominal aortic aneurysm, and pancreatitis.

Other techniques

Magnetic resonance imaging

A number of magnetic resonance urography techniques are available. In suspected obstruction when irradiation and/or contrast exposure are contraindicated (for example pregnancy, contrast allergy, impaired renal function) and ultrasonography is inconclusive, heavily T_2-weighted sequences may be used to generate a magnetic resonance urogram. The dilated pelvicalyceal system and ureter show increased signal and can be delineated without using contrast medium. However, since magnetic resonance does not demonstrate calculi, it may be difficult to be sure that the obstructing lesion is a stone.

Antegrade and retrograde pyelography and ureterography

If the site of obstruction is not demonstrated by intravenous urography or other imaging techniques, antegrade or retrograde examination may be helpful. Both have the advantage that they can be initiated as a method of diagnosis and then extended to provide a therapeutic role by providing drainage.

Radionuclide imaging

There is no role for the use of radionuclides in the investigation of acute urinary tract obstruction.

Which imaging technique to use?

The time taken for each of the examinations described above is not a major consideration. With appropriate equipment, unenhanced helical CT is very quick (approximately 5 min), ultrasonography takes approximately 15 min, and standard urography 30 min. If there is an obstruction, delayed films up to 24 h after contrast injection may be necessary with urography. The cost of ultrasound is approximately half that of intravenous urography but neither is expensive by the standards of the developed world. It has been claimed that the cost of helical CT scanning is equivalent to that of urography, although the initial cost of providing the necessary equipment is high.

The plan and sequence of investigation in suspected upper tract obstruction is dictated by the mode of presentation and the presence or absence of uraemia. If there is suspected chronic urinary tract obstruction, ultrasonography is the method of choice. Uraemia, if present due to urinary tract obstruction, must indicate bilateral obstruction or obstruction of a functionally or anatomically single kidney. Renal ultrasonography (plus plain films to screen for the presence of calculi) is the investigation of first choice. The same is true if a palpably enlarged kidney or kidneys are present in the absence of pain or flank tenderness. Where signs of sepsis are present and pyonephrosis (an infected and obstructed system) is suspected, ultrasound is also the investigation of first choice. However, many patients with acute urinary tract obstruction present with pain and without uraemia or signs of sepsis. In this situation, intravenous urography (or unenhanced spiral CT, if available) remains the investigation of first choice for reasons given above. Magnetic resonance urography may be of value in the minority of patients in whom ultrasonography is inconclusive and irradiation and/or contrast exposure are contraindicated.

Chronic upper urinary tract obstruction

Obstruction must be excluded in all patients with unexplained renal failure. In patients with known renal disease, rapid deterioration in renal function unexplained by the primary renal problem also demands investigation. Relapsing urinary tract infections should also raise the possibility of an associated obstructing lesion. The diagnosis of chronic resistance to outflow should not be discounted simply because the volume of urine is normal or even increased.

The history should include questions relating to analgesic abuse (associated with papillary necrosis, transitional cell tumours, and periureteric fibrosis) and vitamin D consumption (associated with calculus formation). Ingestion of methysergide and other drugs may be associated with retroperitoneal fibrosis. A history or family history of gout, diabetes, or renal stone formation should be sought.

Ultrasonography

In suspected chronic upper tract obstruction, including the initial investigation of patients with unexplained impairment of renal function, ultrasonography is usually the imaging method of choice. Plain films should also be obtained, often with plain renal tomography to check for low-density calculi if the bowel overlies the renal area. Ultrasonography has a high sensitivity for the detection of pelvicalyceal dilatation but cannot distinguish between dilatation caused by obstruction and other types of pelvicalyceal abnormality—caused for example by a distensible system, extrarenal pelvis, vesicoureteric reflux, or dilated calyces due to reflux nephropathy (Table 1). To avoid missing the relatively minor dilatation that may occur with some causes of severe functional obstruction (for example in retroperitoneal fibrosis and tumours), all questionable pelvicalyceal visualization on ultrasonography must be evaluated further. This, together with the difficulty in differentiating dilated obstructed from dilated non-obstructed systems, leads to a significant false positive rate when ultrasonography is used to exclude obstruction. Further evaluation with intravenous urography or CT may be necessary depending on renal function. Intravenous urography is the method of choice in patients with loin pain since it best diagnoses the more common causes—calculous obstruction and idiopathic pelviureteric junction obstruction. Scintigraphy is not recommended as the first investigation in suspected obstruction but is useful in defining whether dilatation shown by other methods is obstructive (see below).

Intravenous urography and CT scanning

Intravenous urography may be helpful when ultrasonography yields equivocal results in a patient with suspected obstruction, especially when renal function is normal.

Unenhanced CT is used as a secondary method of diagnosing obstruction when the results of ultrasonography are equivocal. On CT the dilated collecting system appears as a multiloculate fluid collection with the density of water in the renal sinus. It is possible to distinguish the intrarenal collecting system from the extrarenal portion of the pelvis; this is important since obstruction can only be diagnosed on CT when there is dilatation of the intrarenal collecting system since a prominent extrarenal pelvis may be a normal variant. The whole dilated ureter is well shown on CT.

The main value of CT in the investigation of chronic upper tract obstruction is in defining the cause (Fig. 10). It is particularly helpful for diagnosing retroperitoneal masses causing obstruction, since the retroperitoneum is often obscured by the bowel at ultrasonography. CT also demonstrates calculi causing obstruction, including 'lucent' calculi such as urate stones that appear with a high density on CT.

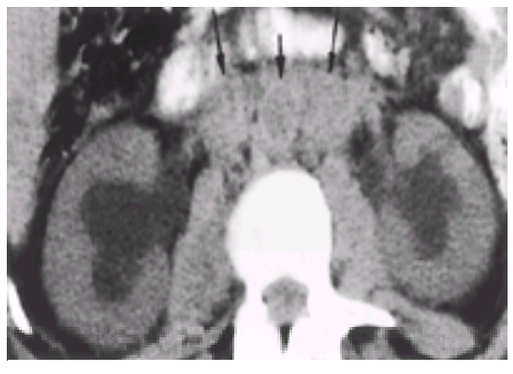

Fig. 10 CT scan in a patient with bilateral urinary tract obstruction due to prostatic cancer. Tumour tissue is clearly delineated on the scan (arrowed).

Scintigraphy

Renal scintigraphy (also called renography) provides functional evidence of obstruction. A radioactive tracer is injected, its passage through the kidney and pelvis is recorded by serial images, and the data are computerized for further analysis (see section 20.4 for further discussion). A rise in resistance to flow in the pelvis or ureter, sufficient to result in impaired renal function, prolongs the parenchymal transit of tracer and there is usually a delay in emptying the pelvis. On whole-kidney renograms, the activity–time curve fails to fall after an initial peak, or continues to rise (Fig. 11). These activity–time curves alone do not enable a distinction to be made between obstructive nephropathy, in which parenchymal transit time is prolonged, and retention of tracer within a large, baggy, low-pressure unobstructed pelvis, where the parenchymal transit time is normal. Parenchymal transit times must therefore be measured through renographic data analysis to make this distinction. Whereas obstructive nephropathy is associated with prolonged transit of tracer through the renal parenchyma, a normal parenchymal transit time with delayed outflow indicates a non-obstructed dilated pelvis.When the possibility of obstruction is suspected, a dynamic renal scintigram is performed with diuresis. Frusemide (furosemide) (0.5 mg/kg, adult dose 40 mg) is given intravenously about 18 to 20 min into the study. Activity–time curves show an immediate fall in the normal kidney after the injection of frusemide (furosemide): in the presence of obstructive uropathy activity persists in the pelvis, the activity–time curve fails to fall, or falls to a lesser extent than its previous rate of increase (an 'inappropriate' response). The half-time of the descending part of the activity–time curve is prolonged, meaning that 'outflow efficiency' is impaired, which compares the amount of activity taken up by the kidney with that excreted within 30 min (the normal value is greater than 78 per cent). However, this test is not infallible: a poor response to frusemide (furosemide) may result from poor renal function or blood volume depletion rather than from obstruction and in the presence of massive pelvic or ureteric dilatation washout may not be observed on the images. Under these circumstances this part of the test is uninterpretable. However, the parenchymal transit time index will indicate whether obstructive nephropathy is or is not present.

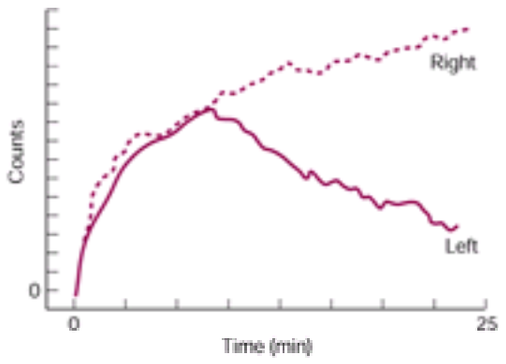

Fig. 11 Dynamic diethylene-triamine penta-aceticacid (DTPA) scintigram. Note the progressive rise of the right kidney curve to a plateau in contrast to the normal left kidney curve.

In conclusion, radionuclide scintigraphy methods are of particular value in differentiating obstructive nephropathy from a baggy, dilated, but unobstructed system and in defining the contribution of each kidney to overall uptake function. A decision as to whether conservative surgery or nephrectomy should be carried out in unilateral obstruction may be much facilitated by the latter assessment.

Antegrade pyelography and ureterography

Percutaneous introduction of contrast medium directly into the renal pelvis or a calyx via a needle, with subsequent radiographic examination of the pelvicalyceal system and ureter (antegrade pyelography and ureterography) is used increasingly to define the site and cause of chronic upper tract obstruction. Diagnostic antegrade examination can be combined with therapeutic drainage of an obstructed system.

Retrograde ureterography

Cystoscopy and catheterization of one or both of the ureters from below, followed by retrograde injection of contrast medium (retrograde ureterography), is indicated if antegrade examination cannot be carried out, or if there is a prospect of dealing with ureteric obstruction from below at the time of retrograde examination. The technique carries the risks of introducing infection into an obstructed urinary tract and of septicaemia, and should be performed only when absolutely necessary. In obstruction due to neuromuscular dysfunction at the pelviureteric junction and in retroperitoneal fibrosis, the collecting system may fill normally from below.

Pressure flow studies

This investigation provides a quantitative assessment of the effect of obstruction on the outflow tract. The technique involves the insertion of a needle transparenchymally into the upper collecting system. Local anaesthetic is sufficient in adults, but general anaesthesia is required in children. The bladder is catheterized and the intravesical pressure measured. The pressure differential between the kidney and the bladder is monitored while the collecting system is perfused with dilute contrast at a rate of 10 ml/min. Perfusion must be maintained for long enough to ensure that the upper urinary tract is filled.

Normal systems can accommodate a flow rate of 10 ml/min without a pressure differential of more than 15 cm of water. If an obstruction is present, there will be a pressure differential of more than 22 cm of water. An equivocal range exists when the differential pressure is between 15 and 22 cm of water, but such a result occurs in only a small proportion of patients.

This diagnostic technique is relatively simple and can readily be extended into a therapeutic one by leaving a catheter *in situ*, to provide drainage of an obstructed system. The disadvantages are that it is an invasive test with a risk (albeit small) of provoking haemorrhage or infection, that the technique investigates the collecting system and gives no information on parenchymal function, and that it is not readily repeatable. Leakage around the needle can invalidate the pressure measurements.

With the advent of more sophisticated renography, pressure flow studies are rarely performed, but they still have a place in the investigation of obstruction in patients with very poor renal function, in whom radioisotope techniques are less reliable, in equivocal obstruction, particularly at the pelviureteric junction, and intraoperatively in patients with retroperitoneal fibrosis undergoing ureterolysis.

Difficult clinical situations

Significant incomplete chronic upper urinary tract obstruction

It may be very difficult to tell whether a given degree of resistance to outflow or intermittent obstruction is impairing, or potentially impairing, renal function or causing symptoms. Symptoms may be present in the absence of deleterious effects upon renal function and the converse may also be true. Different methods of diagnosing obstruction define subtly different pathological features of the condition and a valid correlation between the results of different investigations cannot invariably be made. Incomplete obstruction is clinically important if it causes deterioration in kidney function that can be halted or corrected by intervention, or symptoms that are improved thereby. In patients with one kidney, or in those with bilateral partial obstruction, a decline in serial measurements of glomerular filtration rate attributable to obstruction may define the situation. There may be a similar change in uptake of radionuclides in unilateral obstruction. Other proposed methods of detecting significant incomplete obstruction are given in Table 3. Strict validation of these methods would require them to be carried out in patients with supposed obstruction who would then be allocated randomly to intervention or no intervention. Deterioration in function predicted by each of the methods by comparison with matched controls would provide validation. This study is never likely to be done.

Differential diagnosis of non-obstructive dilatation of the collecting system

A number of non-obstructive conditions may cause dilatation of the collecting system (Table 1).

Extrarenal pelves may mimic pelviureteric junctional obstruction. If intravenous frusemide (furosemide) is administered after the 20-min full length contrast radiograph in this condition, contrast medium will have washed out of the affected side on a full length film 15 min later. In the presence of obstruction of the pelviureteric junction the contrast is retained and the pelvicalyceal system increases in size.

Megacalyces are readily identified on urography. The renal cortex is normal and the calyceal infundibula, pelvis, and ureter are normal with no evidence of obstruction.

Vesicoureteric reflux may be associated with dilatation of the ureters and in severe cases dilatation of the pelvicalyceal system too. The presence of reflux on urography is suggested by the degree of dilatation varying at different times during the examination, by dilatation which is greatest from the vesicoureteric junction upwards, and by a postmicturition film which shows a large volume of residual urine in the bladder composed of urine that has refluxed into the ureters during voiding and drained back thereafter.

A decision as to whether or not significant obstruction is present at the pelviureteric junction and whether operation is indicated may be facilitated by frusemide (furosemide) urography (Fig. 12) or frusemide (furosemide) scintigraphy. In some patients the urographic findings are unremarkable during asymptomatic periods, while emergency intravenous urography during an episode of pain may define the condition.

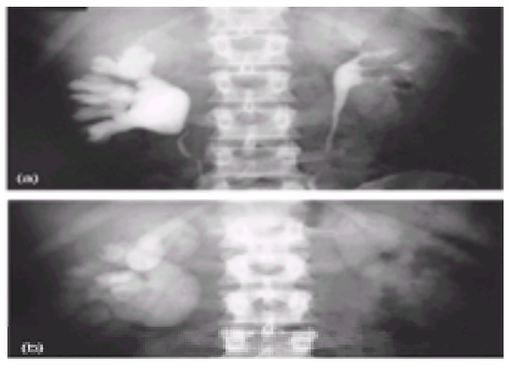

Fig. 12 (a) Right-sided pelviureteric obstruction. (b) Same urogram as in (a) 15 min after intravenous injection of frusemide (furosemide). Note the increase in size of the pelvicalyceal system, indicating significant pelviureteric junction obstruction.

In women who have been pregnant, particularly those who have suffered pregnancy bacteriuria, one or both upper ureters (more often the right) may be dilated to the pelvic brim, but the system is seen to empty on a full length postmicturition film and there is no dilatation of the pelvis and calyces (see Chapter 13.5 for further discussion).

Acute lower urinary tract obstruction

Most patients presenting with acute urinary retention require no investigation before treatment. Suprapubic pain coexisting with a bladder that is palpably or percussibly distended above the level of the symphysis pubis is sufficient evidence for immediate catheterization.

If there is doubt about the diagnosis, an ultrasound examination will confirm or refute the presence of a distended bladder. Transrectal ultrasound of the prostrate can demonstrate both the size of the gland and, to some extent, the benign or malignant nature of any enlargement. Such an investigation is not indicated in the acute situation but is of potential benefit after the relief of obstruction.

An ascending urethrogram may be indicated if an attempt at urethral catheterization proves unsuccessful. This is done as an elective procedure after bladder drainage has been achieved by suprapubic catheterization.

Chronic lower urinary tract obstruction

In one series of patients presenting with acute retention of urine, approximately half had bladders of low compliance and half of high compliance. Investigation is aimed at demonstrating associated pathology such as urinary tract infection, upper tract dilatation, stones, and renal impairment, and also at defining the severity of obstruction of bladder outflow.

Urine culture is essential. In most centres, ultrasonography of the upper and lower urinary tract, together with a plain abdominal radiograph and measurement of urinary flow rate have replaced the intravenous urogram. Full urodynamic investigations, with combined videopressure cystourethrography may be necessary. Serum biochemistry and routine haematology are also required, as is measurement of the level of prostatic-specific antigen in men.

Management

Acute upper tract obstruction

Stones

Most patients presenting with renal and ureteric colic will have a stone in the lower third of the ureter, often in that portion of the ureter lying within the bladder wall. Such patients can be managed conservatively, since the stone has already passed through two areas of relative ureteric narrowing—the pelviureteric junction and the site at which the ureter crosses the bifurcation of the common iliac artery. A conservative policy is likely to prove successful if the stone is 5 mm or less in its maximum diameter. It is unusual for acute episodes of colic to persist for more than 72 h.

Patients with ureteric colic are usually admitted to hospital, although this is unnecessary in many cases since the only medical requirement is the provision of regular analgesia, which can be given parenterally, orally, or rectally. There is a time-honoured recommendation that patients with colic should be encouraged to maintain a very high fluid intake to induce a diuresis. An antimuscarinic drug such as propantheline is also often prescribed. There is no reason to think that these measures are of benefit, and they may even be harmful since both encourage ureteric dilatation and would be expected to reduce forward peristalsis of the ureter, which is the very effect needed to encourage spontaneous passage of the stone. A diuresis will also tend to increase intratubular pressure and may increase the risk of forniceal rupture. It might be argued that forniceal rupture may, by decompressing the system, encourage the return of peristalsis but few would regard this as an appropriate approach to management.

Although it has long been argued that morphine should be avoided as an analgesic as it may provoke prolonged ureteric constriction, there is no evidence that therapeutic doses of morphine have this adverse effect. Pethidine may provoke nausea and vomiting, particularly when administered parenterally, but since nausea and vomiting frequently accompany colic it is difficult to disentangle the effects of such treatment from the effects of colic alone. Very satisfactory pain relief can often be obtained using non-steroidal anti-inflammatory agents administered orally or rectally.

With the advent of new, less invasive methods of surgical management of ureteric stones, there is a temptation to intervene earlier. Stones in the intramural ureter can be treated readily with most lithotripters, whether imaging is by ultrasound or radiology. Since most stones at that site will pass spontaneously, the extent to which lithotripsy will hasten the process is difficult to establish. Lithotripsy may be worthwhile for stones in the upper third of the ureter since, by disintegrating the stone into small fragments, spontaneous passage will be encouraged. However, the availability of lithotripsy for patients with acute colic varies very markedly between countries and the precise role and benefits of the technique have yet to be established. Endoscopic manoeuvres, which are usually performed under general anaesthesia, are reserved for those patients with persistent colic.

Drainage of an obstructed system

If there is clinical evidence of infection above an obstruction, drainage must be established as a matter of urgency. The diagnosis is a clinical one, made on the basis that the patient is pyrexial, often with a very high fever and rigors, and the degree of loin tenderness is greater than when obstruction is not associated with infection. Examination of bladder urine may be unhelpful since ureteric obstruction may prevent red and white blood cells and organisms from reaching the lower urinary tract. Leucocytosis may be present but this is not invariably the case, especially in the elderly.

The choice between antegrade and retrograde intervention will depend on the facilities and expertise available. In most specialist centres there is a clear preference for the percutaneous insertion of an antegrade needle to provide a nephrostomy under local anaesthesia. In a dilated high-pressure system the procedure is usually easy, and such a system may be used to provide drainage for weeks or even months if necessary. If excretion of intravenous contrast has outlined renal anatomy, renal puncture may be guided radiographically; if not, the initial puncture may be better directed under ultrasound control using a fine needle. The collecting system is then outlined with contrast and an accurate transparenchymal calyceal puncture can be placed, usually through a lower calyx. By contrast, a retrograde ureteric catheter can be relied on to provide drainage for only for a matter of days at best. Occasionally a retrograde catheter cannot be passed beyond the obstruction and the diagnostic role of retrograde ureterography cannot then be extended to a therapeutic one.

Other causes of acute obstruction

Aside from urinary stone, the two other most common causes of acute obstruction are sloughed papillae and blood clots. The principles of management vary little from those already outlined for ureteric stones, but greater attention must be paid in the acute phase to the underlying cause. In the patient with papillary necrosis, infection is a more common accompaniment of obstruction, and intervention, usually with a percutaneous needle nephrostomy, is required more often. When colic results from a blood clot, treatment of the underlying cause may be necessary at an early stage. Renal parenchymal tumours and transitional cell tumours of the collecting system may both cause persistent bleeding and colic, and ablative open surgery is usually required. More difficulty is encountered when bleeding occurs from a non-malignant cause. An arteriovenous fistula, whether spontaneous or traumatic, may be embolized with every prospect of success. The most difficult case of recurrent bleeding to manage is that associated with papillary necrosis in sickle cell trait or disease. Antifibrinolytic agents may be of value, but administration of such treatment during active bleeding may produce hard, rubbery clots that fill the collecting system and require surgical removal.

Chronic upper tract obstruction

The aim of management is to relieve symptoms, improve or conserve renal function, and avoid complications such as septicaemia. Important surgical advances in the past decade include the increasing use of ureteric stents to provide short-term, or even long-term, relief of obstruction. Recently a means of avoiding problems associated with endoluminal stenting has been described, in which a stent is used to drain urine from the renal pelvis through a subcutaneous tube directly into the bladder, bypassing the ureter.

Obstruction is the most readily reversible cause of chronic renal failure. Acute obstruction caused by ureteric stones commonly resolves spontaneously; but the longer a stone remains in the same position within the ureter the less likely it is that a conservative policy will be successful.

Probably the second most common cause of chronic obstruction in adults is obstruction of the pelviureteric junction. The Anderson–Hynes pyeloplasty gives very satisfactory results and provides the gold standard against which other open and endoscopic techniques (such as endopyelotomy) must be assessed.

Idiopathic obstruction at the pelviureteric junction may present in childhood, when obstructed megaureter and ureteric obstruction secondary to a ureterocele are also more common. Since all three congenital anomalies cause pelvicalyceal dilatation, it is becoming increasingly common for obstruction to be diagnosed *in utero*. Treatment of the obstruction *in utero* by the insertion of a nephrostomy tube has been reported: it is too early to know whether such early intervention will prove to have long-term benefits, but at the moment it seems, on balance, best to wait until immediately after delivery to investigate and relieve the problem.

Ureteric obstruction can occur in a transplanted kidney, most commonly at the site of the ureteroneocystostomy, but sometimes more proximally in the ureter. Vesicoureteric stenosis is caused by ischaemia of the ureter, but it is never possible to define whether this ischaemia is associated with rejection or is a result of poor vascularization following donor dephrectomy. More proximal ureteric obstruction may be due to mechanical kinking of the ureter or, occasionally, to extrinsic compression by a lymphocele. Irrespective of the site and cause of the obstruction, the diagnosis presents special problems. The possibility of obstruction is raised either because of deteriorating renal function or because ultrasound during routine follow-up demonstrates increasing dilataton of the collecting system. The differential diagnosis includes rejection, cyclosporin nephrotoxicity, and arterial insufficiency. An obstruction may be demonstrated by intravenous urography and/or antegrade pyelography. Retrograde studies are usually difficult and not infrequently impossible, and since, in the case of stenosis at a ureteroneocystostomy, they involve passing a catheter across the segment of ureter under suspicion, the investigation is only indicated if intravenous urography and antegrade pyelography prove unsatisfactory.

Minimally invasive stone surgery

The past decade has seen a revolution in the management of urinary tract stones, due to the adoption of minimally invasive techniques including percutaneous surgery and extracorporeal shock wave lithotripsy. Open operation for renal stones can now be avoided in many cases by creating a nephrostomy track to the calculus, dilating the track, and then either removing the calculus endoscopically via the track or causing it to disintegrate by direct application of an ultrasound probe. It is possible to extract ureteric stones endoscopically with the assistance of a ureteroscope, and bladder calculi may be disintegrated by the endoscopic application of electrohydraulically produced shock waves.

Externally delivered shock waves can be used to shatter calculi into many fragments which are then passed spontaneously, hence extracorporeal shock wave lithotripsy offers a solution to the problem of the presence of a calculus or calculi within the kidney without the need for a surgical operation. This carries the promise of a reduction in morbidity and perhaps mortality, a much shorter hospital stay for the patient, a more rapid return to work, and may also be suitable for those who are unfit for conventional surgery. The technique is unsuitable for hard uric acid and cystine stones, for very large stones (which must be debulked percutaneously before lithotripsy), and for some ureteric stones (although a proportion of these can be manoeuvred into the upper collecting system endoscopically and then dealt with by this method). Other disadvantages include the high capital cost of the necessary equipment and the need for further intervention in 10 to 15 per cent of patients in whom stone fragments do not pass. Such fragments can, in general, be removed endoscopically. Despite these difficulties, non-operative dissolution of calculi is being increasingly used. The recurrence rate of stones in the long term seems to be no higher than after open surgery and to date no unforeseen long-term complications have emerged.

Large staghorn calculi are still usually removed by a cutting procedure. Surface cooling of the kidney at the time of operation allows time for more complete clearance of stones with the renal artery clamped, and protects against the development of ischaemic damage to the kidney.

For details of other aspects of diagnosis and management of urinary stones see Chapter 20.13.

Other causes of urinary obstruction

Pelviureteric junction obstruction

This often appears to result from a functional disturbance in peristalsis of the collecting system in the absence of mechanical obstruction. A percutaneous procedure for managing pelviureteric junction obstruction was first described in 1983 (endopyelotomy): this involves a full thickness incision through the stenosed region with a stent left *in situ*; healing occurs by re-epithelialization from either side of the incision and very little new scar tissue is formed. There is no consensus on the indications: patients with secondary pelviureteric junction stenosis in association with stones, infection, or previous surgery tend to be offered the percutaneous operation, whereas those with primary idiopathic obstruction are usually treated by open pyeloplasty.

Malignant obstruction

A wide variety of tumours may cause ureteric obstruction, either by local spread (cervix, prostate, bladder), or secondary to para-aortic nodal enlargement (lymphoma, testicular tumours). The diagnosis rests upon the same investigations as for any other cause of chronic obstruction, but the treatment will vary widely, depending on the cause. An aggressive or radical approach is almost always indicated in a patient who has received no previous treatment for the underlying malignancy. Unilateral or bilateral ureteric stenting or percutaneous nephrostomies may be necessary to cover the period of time during which chemotherapy or radiotherapy is given with the expectation of controlling the tumour. More difficulty arises when ureteric obstruction is due to recurrent tumour, when the potential benefits of chemotherapy and radiotherapy are significantly less and patients may be facing debilitating treatment for an advancing malignant disease, the prognosis of which is poor. To be confined by nephrostomy drainage for what is left of life significantly diminishes its quality, but may be right in certain circumstances. A percutaneously placed pigtail nephrostomy, which can be inserted under local anaesthetic, has a tendency to fall out or to be pulled out. Open surgery can be avoided by the use of a ring nephrostomy inserted percutaneously under general anaesthesia: this provides secure long-term drainage, for years if necessary.

Obstruction in patients with urinary diversion

There are many reasons for diverting the urine into an isolated loop of ileum or colon. One of the recognized complications is stenosis at the site of anastomosis between the bowel and ureter(s).

The thin muscle wall of the ileum means that it is not possible to fashion an antireflux anastomosis between the bowel and the ureter when diverting the urine into an ileal conduit, hence a loopogram (a radiograph carried out after injection of contrast into an ideal loop) will normally show bilateral ureteric reflux, and the absence of such reflux is strong evidence of a stenosis at the ureteroileal junction.

The operation of ureterosigmoidostomy, in which the ureters are anastomosed to sigmoid colon, has fallen into disfavour owing to the associated complications of infection, metabolic acidosis (caused by reabsorption of hydrogen ions from the gut), and osteomalacia.

Idiopathic retroperitoneal fibrosis (peri-aortitis)

In this condition the ureters become embedded in dense fibrous tissue, with resultant unilateral or bilateral obstruction, usually at the junction between the middle and lower thirds of the ureter. The condition is progressive: initially, the fibrous tissue is fairly cellular, later becoming relatively acellular. The mechanism by which obstruction occurs is unclear, not least because of the frequent observation that contrast medium injected into the lower ureter may pass freely up to the pelvicalyceal system despite the presence of clinical, radiological and isotopic evidence of functional urinary tract obstruction.

Pathogenesis

'Retroperitoneal fibrosis' is an unfortunate term, since there are many causes of fibrosis in the retroperitoneal area, for example malignant disease of the breast, colon, or prostate, and because it is anatomically misleading and says nothing about pathogenesis.

The location, long known to surgeons and pathologists, and further emphasized by the advent of CT scanning (Fig. 13), is around the aorta, hence the term peri-aortitis is preferable to retroperitoneal fibrosis. The histological appearances are of aortic atheroma, medial thinning, splits in the media, and an increase in the adventitia, which contains an inflammatory infiltrate. These findings are present to some extent in the aortas of some patients with advanced atherosclerosis who have not suffered a clinical illness and may reasonably be classified as having 'subclinical peri-aortitis'. The fibrous tissue itself contains macrophages and plasma cells but not polymorphs, and it now seems likely that this is due to an autoimmune response to leakage of material derived from atheromatous plaques in the diseased aorta.

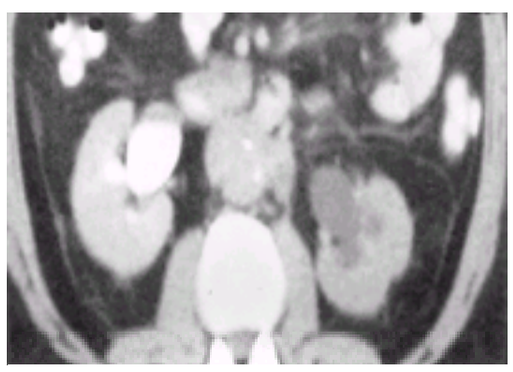

Fig. 13 A CT scan of idiopathic retroperitoneal fibrosis (peri-aortitis) causing urinary tract obstruction. Note the aortic calcification and peri-aortic nature of the mass.

The substance ceroid, an insoluble polymer of oxidized lipid and lipoprotein, which can be synthesized artificially by oxidizing low-density lipoprotein, may be involved in generating the fibrotic reaction. It is found in atheromatous plaques and is identified by staining with oil red O. Examination of sections of aorta containing such plaques incubated with mouse monoclonal antibody to human IgG localizes the antibody to the region of the plaque where ceroid has been identified by oil red O staining. Identical findings are obtained on incubation with polyclonal rabbit antihuman IgG. Moreover, IgG and some IgM, but not IgA or IgE, can be identified in plasma cells in the fibrotic tissue where there are splits in the adjacent media. Circulating antibodies to oxidized low-density lipoprotein and to ceroid extracted from human atheroma are detected in patients with peri-aortitis in much higher concentrations than in normal individuals and in those with ischaemic heart disease. Stored sera obtained from individuals subsequently shown at autopsy to have had subclinical peri-aortitis also show significantly increased antibody titres compared with controls. It thus seems likely that chronic peri-aortitis has an autoimmune aetiology in which the allergen is a component of ceroid, probably oxidized low-density lipoprotein, produced in human atheroma, and that a specific immune response involves T cells and plasma cells, which secrete IgG. Oxidized low-density lipoprotein is known to be highly immunogenic.

This concept clarifies some issues that were previously difficult to explain. For example, the definite, although uncommon, association between mediastinal fibrosis and idiopathic retroperitoneal fibrosis has always been difficult to understand. If one regards at least some cases of mediastinal fibrosis (see below) as a peri-aortitis, occurring in this instance around the thoracic aorta, the association becomes comprehensible. Surgeons operating on aortic aneurysms quite often see fibrosis around the aneurysm. Sometimes the surgeon encounters technical difficulties in adhesions between the aorta and duodenum, and a dense, fibrotic, chronic inflammatory infiltrate is present around the aorta: the term 'inflammatory aneurysm' is used to describe this condition. The unifying hypothesis of an autoimmune peri-aortitis accounts for this finding. Certainly so-called idiopathic retroperitoneal fibrosis, idiopathic mediastinal fibrosis, perianeurysmal fibrosis, and inflammatory aneurysm have much in common. The hypothesis also accounts for the well-known association between aortic disease, including aneurysm and aortic wall calcification, and retroperitoneal fibrosis. Finally, it may be no coincidence that carcinoid tumours and drugs such as methysergide and ergot derivatives which are sometimes responsible for the condition all have well described effects on the vasculature.

Clinical features

The condition is three times as common in men as in women. Patients' ages range from the third to the ninth decade, but peak incidence occurs in the sixth and seventh decades; in one series of 60 patients the mean age of the group was 56 years. The early clinical manifestations are not distinctive. Most commonly there is pain in a girdle-like distribution from the low back to the lower abdomen, occasionally spreading to the buttocks or thighs. Examination is usually unremarkable apart from hypertension, which is found in over 50 per cent of patients. Oedema of the legs, a palpable kidney, or hydrocele is found in fewer than 10 per cent of patients. There is usually a normochromic, normocytic anaemia and a raised erythrocyte sedimentation rate and plasma C-reactive protein, but a significant minority are normal in one or both of these respects. Proteinuria is uncommon and significant bacteriuria rare.

Diagnosis

Peri-aortic fibrosis is clearly more common than hitherto appreciated, particularly if one takes subclinical forms of the condition into account. Diagnostic delay is the rule: in one series 6 to 12 months, or even longer, elapsed from the onset of symptoms to diagnosis. Perhaps for this reason, bilateral rather than unilateral upper tract obstruction was present in the majority of patients.

When taking the history, enquiry should be made regarding consumption of relevant drugs, including methysergide, b-blockers, methyldopa, and bromocriptine, another ergot-like drug.

Ultrasonography, radionuclide methods, and the intravenous urogram will reveal findings typical of urinary tract obstruction, and the last technique may show medial deviation of the ureters, although this may also be present in normal subjects and is an unreliable guide to diagnosis. CT scanning will show the peri-aortic mass (Fig. 13). The differential diagnosis includes lymphoma (in which case splenomegaly and lymphadenopathy may be seen on CT scanning) and various forms of cancer, including particularly those of the bladder, bowel, and cervix.

A histological diagnosis should be obtained if at all possible, and laparotomy is required in order to obtain a sufficiently large sample to exclude lymphoma and cancer with certainty. Conversely, a CT-guided needle biopsy may be sufficient to make a definitive diagnosis of malignancy.

Management

Management of the idiopathic and probably autoimmune syndrome is empirical and controversial since controlled trials of treatment are lacking. Corticosteroid therapy, with or without temporary relief of obstruction by insertion of ureteric stents, ureterolysis alone, and ureterolysis followed by steroid therapy to shrink the peri-aortic mass and maintain remission are all employed. Corticosteroid therapy alone may correct obstruction, but is not invariably effective. Ureterolysis alone may also correct obstruction in the long term but is sometimes associated with recurrence or the development of a further obstruction in a previously unaffected kidney. Surgical relief of obstruction by ureterolysis followed by corticosteroid therapy (initial dose of prednisolone 20 mg daily, begun when sutures are removed) has proved to be a reliable and successful strategy. Corticosteroid dosage is reduced progressively thereafter according to clinical response. When bilateral obstruction is present, bilateral ureterolysis followed by steroid therapy is preferable to unilateral ureterolysis with reliance upon corticosteroid therapy to free the contralateral side, since this is sometimes unsuccessful. Ureterolysis of kidneys shown to be non-functioning on high-dose excretion urography or by appropriate radionuclide techniques is usually unsuccessful in restoring useful renal function. A reasonable policy for management would seem to be to perform unilateral or bilateral ureterolysis, as appropriate, followed by corticosteroid therapy in patients fit for operation and able to take steroids safely. Surgery alone should be employed in those with a particular contraindication to corticosteroid treatment, such as the presence of a peptic ulcer or severe osteoporosis. Steroid therapy alone (methylprednisolone 500 mg intravenously daily for 3 days, followed by prednisolone 20 mg daily), with or without insertion of ureteric stents, should be reserved for patients unfit for ureterolysis. A dramatic response to parenteral steroid treatment sometimes occurs, a marked diuresis being seen within 24 h of commencing treatment.

In the United Kingdom in recent years there has been an increasing tendency to avoid operation, even in patients fit for ureterolysis, whether unilateral or bilateral. Reliance is placed upon the insertion of a ureteric stent or stents plus corticosteroid therapy. This approach has the advantage of avoiding operative mortality and morbidity, of providing a rapid solution to the problem of urinary tract obstruction, and a much reduced hospital stay. Disadvantages include difficulties in stent insertion, incomplete relief of obstruction by stents, the need for periodic (say 6-monthly) change of stents if steroid therapy is unsuccessful, and the potential for urinary tract sepsis in the presence of a stent, which is a foreign body. To these must be added, in 'real life' clinical practice, the potential for patients to be lost to follow-up and the presence of a stent to be forgotten. No completely reliable method has ever been devised to render this last occurrence impossible. By contrast, when the ureter or ureters are displaced surgically, well away from the periaortic mass, the patient and the patient's urinary tract are secure, and a further advantage of open surgery is the potential it offers to obtain adequate tissue to permit a firm histological diagnosis to be made. Patients judged fit for operation should in general undergo ureterolysis unless or until prospective controlled trials are published showing equivalent or better results from alternative approaches.

Peri-aortitis in the absence of ureteric obstruction

The use of CT scanning in the investigation of abdominal pain has revealed an increasing number of patients to have peri-aortitis before the onset of urinary tract obstruction. Management of these cases is controversial. The development of bilateral ureteric obstruction with severe uraemia within 3 months of diagnosis (at which time renal function was normal and the ureters unobstructed) has occurred in at least one patient. Until more is known of the natural history of the disease in such patients, it would seem prudent to obtain a histological diagnosis at open operation and to consider corticosteroid therapy to shrink the mass. Whether an attempt to reduce the risk of ureteric obstruction by insertion of stents or displacement of the ureters from the mass at the time of the operation should be carried out is not known.

Prognosis

The older and the more uraemic the patient at the time of presentation, the worse is the prognosis. Nevertheless, if treated appropriately, most patients do well.

Follow-up

In some patients long-term remission is achieved by surgery alone. In those receiving maintenance prednisolone, the dose can be reduced progressively, and in some patients long-term remission occurs after complete withdrawal of corticosteroid therapy. In one series of 60 patients, 10 relapsed more than 5 years after the time of diagnosis when steroid therapy had been stopped, in that their erythrocyte sedimentation rates rose to an abnormal level, and obstruction and diminished renal function redeveloped. Five patients relapsed as late as 10 years after the onset of the disease. Lifelong follow-up is therefore mandatory, but the best way to monitor such patients is not certain. Clinical assessment, serial measurement of erythrocyte sedimentation rate and C-reactive protein, and assessment of renal function, together with imaging to detect redevelopment of obstruction, is appropriate. Reduction in size of the peri-aortic mass can be detected on serial CT scanning, but residual peri-aortic tissue is seen frequently, even after steroid therapy, and the usefulness of CT in monitoring disease activity is limited.

Associated fibrotic conditions

Mediastinal fibrosis

The pathological process described in connection with retroperitoneal fibrosis can also develop in the upper mediastinum where it tends to be located around the bronchi, the cardiac atria, the pulmonary arteries and veins, the superior vena cava, and the azygos vein; rarely, it also envelops the oesophagus. Symptoms vary according to the structures principally affected. There may be cardiopulmonary manifestations because of scar tissue about the atria, pulmonary vessels, or bronchi, and dysphagia can result from oesophageal constriction. One of the commonest clinical manifestations results from obstruction of the superior vena cava, with distension of veins in the neck and upper extremities. When mediastinal fibrosis appears without discernible cause, it may be associated with retroperitoneal fibrosis and then probably has the autoimmune origin described above. There are, however, other causes to consider. The condition is encountered most commonly in people who reside in places where histoplasmosis is endemic, notably in the central parts of the United States, and most of the case reports have come from clinics located in the Mississippi river valley (see Chapter 7.13.1 for further information). Studies of some of these patients have revealed the existence of large granulomas due to histoplasmosis, with eventual rupture into the superior mediastinum and subsequent growth of the dense masses of scar tissue characteristic of the fibrosing syndromes. A curious anomaly in this context is that tuberculosis rarely, if ever, causes the syndrome.

The diagnosis is suggested by radiographic demonstration of the fibrous tissue in the affected areas. CT scanning may be of great value in this context, and histological verification of the diagnosis can be made by CT-guided needle biopsy or by mediastinoscopy. Surgical treatment of mediastinal fibrosis is much more hazardous than of retroperitoneal fibrosis and is much less likely to be beneficial. Despite this, some experienced thoracic surgeons recommend that attempts be made to remove large granulomatous masses of histoplasmosis when this is the diagnosis. Chemotherapy for histoplasmosis has not been very effective. In view of the tendency of this fibrosing process to burn out eventually, it may be possible to ameliorate the manifestations by steroid therapy and thus gain time for a collateral circulation to develop. However, there is little experience of the effects of steroid treatment or other forms of immunosuppression, but there is obvious potential for this approach if the origin of the disorder is autoimmune.

Other rarer fibrosing syndromes

In association with retroperitoneal fibrosis and mediastinal fibrosis, other fibrotic processes have been reported to involve the thyroid gland (Riedel's thyroiditis), the pancreas, the salivary glands, and orbital tissue. The last mentioned can cause severe proptosis and damage to the optic nerve leading to loss of vision, some cases of which may be due to undiagnosed Wegener's granulomatosis. Peyronie's disease is characterized by the deposition of fibrous plaques in the corpora cavernosa of the penis. These plaques, which can be detected by palpation, may cause discomfort and angulation during penile erection.

Chronic inflammatory bowel disease

Chronic inflammatory bowel disease is associated with chronic and unsuspected urinary tract obstruction in 10 to 15 per cent of patients. The obstruction is nearly always right-sided in patients with Crohn's disease, and a valuable clue to its existence is pain radiating down from the right iliac fossa into the right leg. The ureter is usually involved in an inflammatory mass. By contrast, in patients with ulcerative colitis the problem may occur on either side and nearly always follows colectomy. There should be a low threshold for ultrasound examination of the urinary tract to detect obstruction in patients with chronic inflammatory bowel disease.

Further reading

Baker LRI *et al.* (1988). Idiopathic retroperitoneal fibrosis. A retrospective analysis of 60 cases. *British Journal of Urology* **60**, 497–503.

Baker LRI *et al.* (1992). Rate of development of ureteric obstruction in idiopathic retroperitoneal fibrosis (periaortitis). *British Journal of Urology* **69**, 102–5.

Better OS *et al.* (1973). Studies on renal function after relief of complete unilateral ureteral obstruction of three months' duration in man. *American Journal of Medicine* **54**, 234–40.

Brooks AP (1990). Computed tomography of idiopathic retroperitoneal fibrosis ('periaortitis'): variants, variations, patterns and pitfalls. *Clinical Radiology* **42**, 75–9.

Dines DE *et al.* (1979). Mediastinal granuloma and fibrosing mediastinitis. *Chest* **75**, 320–4.

Früh D, Jaeger W, Küfer O (1975). Orbital involvement in retroperitoneal fibrosis (morbus ormond). *Modern Problems in Ophthalmology* **14**, 651–6.

Ghose RR (1990). Prolonged recovery of renal function after prostatectomy for prostatic outflow obstruction. *British Medical Journal* **300**, 1376–7.

Gillenwater JY (1986). The pathophysiology of urinary obstruction. In: Walsh PC, ed. *Campbell's Urology*, 5th edn, p 554. WB Saunders, Philadelphia.

Graham JR *et al.* (1966). Fibrotic disorders associated with methysergide therapy for headache. *New England Journal of Medicine* **274**, 359–68.

Higgins PM *et al.* (1988). Non-operative management of retroperitoneal fibrosis. *British Journal of Surgery* **75**, 573–7.

Jaworski ZF, Wolan FT (1963). Hydronephrosis and polycythaemia, a case with erythrocytosis relieved by decompression of unilateral hydronephrosis and cured by nephrectomy. *American Journal of Medicine* **34**, 523.

Keuhnelian JG, Bartone F, Marshall VF (1964). Practical considerations from autopsies in uraemic patients. *Journal of Urology* **91**, 467–73.

McDougal WS, Wright FS (1972). Defect in proximal and distal sodium transport in post-obstructive diuresis. *Kidney International* **2**, 304–17.

Mitchinson MJ (1970). The pathology of retroperitoneal fibrosis. *Journal of Clinical Pathology* **23**, 681–9.

Moody TE, Vaughan ED, Gillenwater JY (1975). Relationship between renal blood flow and ureteral pressure during 18 hours of total unilateral occlusion. *Investigative Urology* **13**, 246–51.

Ormond JK (1948). Bilateral ureteral obstruction due to envelopment and compression by an inflammatory process. *Journal of Urology* **59**, 1072–9.

Parums DV, Brown DL, Mitchinson MJ (1990). Serum antibodies to oxidised LDL and ceroid in chronic periaortitis. *Archives of Pathology and Laboratory Medicine* **114**, 383–7.

Parums DV, Chadwick DR, Mitchinson MJ (1986). The localisation of immunoglobulin in chronic periaortitis. *Atherosclerosis* **61**, 117–23.

Pryor JP *et al.* (1983). Do beta adrenoceptor blocking drugs cause retroperitoneal fibrosis? *British Medical Journal* **287**, 639–42.

Roy C, Saussine C, Jacqmin D (2000). Magnetic resonance urography. *British Journal of Urology* **86** (suppl. 1), 42–7.

Sacks SH *et al.* (1989). Late renal failure due to prostatic outflow obstruction: a preventable disease. *British Medical Journal* **298**, 156–9.

Schowengerdt CG, Suyemoto R, Main FB (1969). Granulomatous and fibrous mediastinitis: a review and analysis of 180 cases. *Journal of Thoracic and Cardiovascular Surgery* **57**, 365–79.

Smith RC, Coll DM (2000). Helical computed tomography in the diagnosis of ureteric colic. *British Journal of Urology* **86** (suppl. 1), 33–41.

Smith RC *et al.* (1996). Diagnosis of acute flank pain: value of unenhanced helical CT. *American Journal of Roentgenology* **166**, 97–101.

Webb JAW *et al.* (1984). Can ultrasound and computed tomography replace high-dose urography in patients with impaired renal function? *Quarterly Journal of Medicine* **xx**, 411–25.

Whelan JS *et al.* (1991). Computed tomography (CT) and ultrasound (US) guided core biopsy in the management of non-Hodgkin's lymphoma. *British Journal of Cancer* **63**, 460–2.

Whitaker RH (1990). The diagnosis of upper urinary tract obstruction. *Postgraduate Medical Journal* **66** (suppl. 1), 25–30.

Whitfield HN *et al.* (1979). Frusemide intravenous urography in the diagnosis of pelvi-ureteric junction obstruction. *British Journal of Urology* **51**, 445–8.

Whitfield HN *et al.* (1981). Renal transit time measurements in the diagnosis of ureteric obstruction. *British Journal of Urology* **53**, 500–3.

Whitfield HN *et al.* (1983). Percutaneous pyelolysis: an alternative to pyeloplasty. *British Journal of Urology* **55**, 93–6.

Wickham JEA, Buck AC, eds (1990). *Renal tract stone*. Churchill Livingstone, London.

20.15 Tumours of the urinary tract

P. H. Smith, H. Irving, and P. Harnden

Introduction

Tumours of the kidney, bladder, prostate, and testis are grouped together only because of their association with the genitourinary tract. Each has, over the years, inspired surgeons to prove that ever more complex operations may be undertaken, but patients have not always benefited from this approach and surgeons now rely increasingly upon physicians working in radiotherapy and medical oncology. Recent emphasis has focused on earlier diagnosis and population screening in the hope that overall mortality will fall as more and more localized tumours are detected. Such an approach carries considerable economic consequences.

Transitional cell carcinoma of the bladder and upper urinary tract

Aetiology and incidence

Risk factors include: tobacco smoking; several chemicals related to the dye, rubber, leather, painting, and organic chemical industries; chronic irritation involving *N*-nitrosamine production; and therapeutic pelvic irradiation in women. Genetic factors modulate individual responses to environmental carcinogens, partially explaining variations in incidences between ethnic groups. The slow *N*-acetylation genotype and inherited defects of the glutathione- *S*-transferase M1 gene are susceptibility factors in occupational and smoking-related bladder cancer. Occupational exposure is thought to account for up to 25 per cent of cases in the United States, but less than 5 per cent of patients with bladder cancer are eligible for prescribed disease benefit in the United Kingdom.

The incidence in males and females is 32.5 and 12.9 per 100 000 population, respectively. The changes in industrial practice (and in supervision of workers) ver the last 50 years has reduced the incidence of bladder cancer due to industrial carcinogens and it is now believed that over 40 per cent of all bladder tumours arise primarily as a consequence of cigarette smoking. The recent changes in attitude to smoking, both in the workplace and socially, may well result in a significant fall in the incidence of bladder cancer in the years to come.

The peak age at presentation is 65 to 69 years for men and 75 to 79 years for women, with less than 5 per cent of tumours occurring in patients younger than 60 years.

Clinical features

Though most bladder tumours present because of haematuria, less frequent presentations include bladder irritability, difficulty with micturition, and symptoms of uraemia or backache. Tumours of the renal pelvis and ureter may bleed or may silently obstruct the relevant kidney and ureter.

Investigation and diagnosis

Urinary cytology and high quality ultrasound is usually adequate to make the diagnosis of bladder cancer (Fig. 1). The flexible cystoscope allows the nature and extent of the lesion within the bladder to be seen. Tumours of the upper tract are revealed by intravenous pyelography or by ultrasound.

Fig. 1 Bladder ultrasound scan showing small (1 cm diameter) transitional cell carcinoma arising from the posterior wall.

Patient categories and treatment

The TNM classification of bladder tumours is unique in that it recognizes a category of papillary carcinoma that is non-invasive (invasion through the basement membrane is the hallmark of malignancy in other tumour types, otherwise the lesion is regarded as premalignant or ' *in situ*'). This evolved because of the difficulties in using standard microscopical techniques to distinguish urothelial papillomas (implying a benign lesion with no associated cancer risk) from the 70 per cent of tumours which recur and may progress. Understanding the biological behaviour of urothelial tumours has been further hampered by the tendency until recently to group together non-invasive tumours (pTa) and the tumours invading the lamina propria (pT1) under the umbrella of 'superficial disease' for treatment purposes, despite

evidence that tumours with lamina propria invasion are at higher risk of progressing further.

Tumours not invading the lamina propria (pTa)

The now accepted primary treatment of the superficial lesion is by transurethral resection with a single instillation of an intravesical agent. Once the tumour has been resected, continuing supervision by check cystoscopy is mandatory as 70 per cent of lesions recur at some stage, whilst approximately 10 per cent will subsequently show progression of grade or stage requiring additional therapy.

Tumours involving the lamina propria (pT1)

These tumours have already demonstrated their invasive potential and require more intensive treatment. Intravesical BCG is commonly used.

If there is associated carcinoma *in situ*, or if the tumour is poorly differentiated, approximately half of the patients will develop invasive disease within 5 years. Many urologists believe that early cystectomy is better for this small subgroup, comprising perhaps 5 to 6 per cent of the whole.

Muscle invasive disease of the bladder (pT2 and above)

For the patient with disease invading the bladder muscle, cystectomy or radiotherapy will be necessary. Cystectomy is well established, but is a major procedure that still carries a mortality of between 2 and 4 per cent in different institutions and in different trials. In addition, the patient must accept the need for some form of urinary diversion or bladder replacement. Both are inconvenient to a greater or lesser extent, and the large segments of bowel needed for bladder replacement or a continent diversion bring with them the risks of metabolic problems associated with impaired absorption from the bowel, of diarrhoea, and of electrolyte imbalance. Radiotherapy may be thought to be a more attractive option, especially for the elderly patient. It must be accepted, however, that radical radiotherapy causes troublesome proctitis in 10 per cent of patients and, if survival is prolonged, telangiectases within the bladder that lead to haematuria. In addition, bladder contracture due to ischaemic fibrosis can necessitate secondary cystectomy for intractable symptoms.

As the two forms of radical treatment, cystectomy and radiotherapy, have not been adequately compared in randomized trials, comparisons of outcomes are limited as cystectomy may only be offered to fitter, younger patients. However, both treatment types offer similar overall 5-year survival rates related to the depth of invasion (which can only be accurately assessed in surgical series). This ranges from 60 to 70 per cent for tumours involving superficial muscle only, to 0 to 10 per cent for tumours which have spread beyond the bladder.

Adjuvant chemotherapy appeared to be very promising 10 years ago, but randomized trials have shown no survival advantage for single agent therapy or for combinations including cyclophosphamide, methotrexate, and vinblastine (CMV). The addition to this combination of adriamycin (M-VAC) may have marginally greater benefit but is also more toxic.

In the United Kingdom it is increasingly being suggested that patients with muscle invasive bladder cancer be managed in cancer centres in which there is surgical, radiation, and medical oncological expertise to advise the patient of the options and to supervise effectively the treatment that is chosen.

Transitional cell tumours of the upper urinary tract

Radical nephro-ureterectomy excising a cuff of bladder mucosa remains the treatment of choice in most patients whose other kidney is normal, since accurate staging of tumours of the ureter and pelvis is often impossible because of difficulties in obtaining biopsies of adequate depth endoscopically. Similarly, the possibility of regular assessment of the upper tracts for evidence of recurrence or progression is limited. Unifocal small lesions may be treated endoscopically or by local excision. In the patient in whom overall renal function is impaired or in whom there is only one kidney a conservative approach is desirable. Such lesions may be treated endoscopically or by local excision in certain cases.

Other tumours

Squamous and adenocarcinomas may be found in the bladder and, more rarely, in the upper urinary tract in association with long-standing stone disease, schistosomiasis, or spinal cord injury. The management of these tumours is surgical and the prognosis is poor.

Screening for transitional cell tumours of the bladder and upper urinary tract

Routine urinary cytology in workers in the chemical and dye industries is well established. The situation in relation to the general public is more debatable: screening by dipstick for microscopic haematuria in the elderly yields a positive test in up to 20 per cent, but a tumour incidence of only 1 per cent. Whilst a single positive dipstick test may not merit full urological investigation, if the test is repeatedly positive, formal investigation (urinary cytology, ultrasound of the urinary tract, and perhaps flexible cystoscopy) becomes inevitable, even in patients taking aspirin or anticoagulants, to reassure the physician, the patient, (and his lawyer). Unfortunately there is no evidence that 'screening' picks up invasive bladder cancer more effectively than the superficial tumour, nor that—in patients known to have bladder cancer—screening and early diagnosis will prevent invasion in the 10 per cent of patients in whom this subsequently recurs. For further discussion of the approach to microscopic haematuria, see Section 20.7.

Carcinoma of the prostate

Aetiology and incidence

There is some relationship between the number of sexual partners or extent of sexual activity and prostate cancer. The condition is less common in those who are celibate. The tumour does not occur in those who are castrated before puberty. Dietary factors are also of importance. The incidence is high at 50.7 per 100 000 men.

Clinical features

Patients may present themselves with symptoms of bladder outflow obstruction, bladder irritation, or the effect of some secondary deposit. Increasingly, however, the diagnosis is made as a result of a prostate-specific antigen (**PSA**) test. PSA is a glycoprotein of molecular weight 33 000 containing 7 per cent carbohydrate. It is a serine protease and an esterase with chymotrypsin-like and trypsin-like activity. It is found almost exclusively in the epithelial cells of the prostate.

Investigation and diagnosis

The introduction of the PSA test has led to increasing public awareness of prostate cancer, whilst the introduction of the nerve-sparing technique of radical prostatectomy by Walsh in the early 1980s offered the hope of effective cancer surgery with preservation of sexual function for the patients whose disease proves to be localized. As a consequence, diagnosis is increasingly made before metastases in lymph nodes or bone are evident. Bony metastases are very unlikely unless the PSA exceeds 20 ng/ml and a bone scan at diagnosis is probably an unnecessary luxury in patients whose PSA is below 20. Repeated bone scans during the course of treatment are unnecessary in the absence of clinical indications since repeated PSA tests are as effective and much cheaper.

Large numbers of cancers are now diagnosed following a PSA test at a stage before they are palpable (category T1c) by sextant biopsies under transrectal ultrasound control. The usual indication for biopsy is a PSA exceeding 4 ng/ml, but there is logic in taking a biopsy from all patients whose PSA exceeds 3 ng/ml (see below). As a consequence, large numbers of patients without symptoms and with no clinical findings are discovering that they have prostate cancer and are having to consider the form of therapy that they are prepared to accept. The situation is further complicated by the difficulty in staging the disease when it is not advanced. There is no true prostatic capsule. The prostate merges with the muscle of the bladder above and with the fibres of the levator ani below.

Radical therapy is appropriate for patients whose disease is confined to the prostate—a particularly important point with this disease since the surgeon has little opportunity to excise any normal tissue around the prostate because of its close association with the rectum and the base of the bladder. Surgical studies reveal that

up to 50 per cent of prostatic tumours are understaged clinically, implying the need for even earlier diagnosis if surgical treatment is to be effective.

Treatment

'Localized disease'

In patients whose PSA is less than 20 ng/ml bone metastases are hardly ever seen and many patients have no involvement of their lymph nodes. Such early diagnosis will be of help only if existing therapy is curative or provides improved survival or quality of life. Unfortunately, 30 to 40 per cent of patients, despite early diagnosis, are found to have tumour extending beyond the prostate at the time of surgical intervention, suggesting that recurrence is likely if not inevitable. This is typically discovered as a rising level of PSA during follow-up ('PSA failure'): following radical prostatectomy or radiotherapy the PSA is expected to fall—to zero after operation or to less than 0.5 after radical radiotherapy since the prostate remains *in situ*. PSA failure is more common in those with higher stage and less well differentiated disease. It is seen in half of patients whose PSA at diagnosis is 10 ng/ml or more and in the vast majority of those with lymph node involvement at the time of treatment. None the less, it is possible that the reduction of tumour bulk by operation or radiotherapy may have affected the natural course of events such that the death rate from prostatic cancer will fall significantly, but this will not be clear for at least another 5 to 10 years.

'Advanced disease'

The dramatic change in outlook following the introduction of oestrogen therapy and orchidectomy in the 1940s was followed 20 years later by a search for agents that were more effective and without cardiotoxicity. Unfortunately, neither the steroidal nor non-steroidal anti-androgens, estramustine phosphate, nor the luteinizing hormone releasing hormone (LHRH) analogues used alone or in combination have been shown to prolong life to any clinically significantly extent, and many have concluded that little progress has been made in hormone therapy in the last 50 years. It is true, however, that sexual function may be preserved somewhat longer if the patient takes a non-steroidal anti-androgen and that the modern compounds have a lower rate of cardiovascular complications than stilboestrol.

Screening for prostate cancer

This topic creates great emotion in urological circles. The debate is polarized between those who believe that screening and radical therapy (prostatectomy) will cure the condition and those who are less certain that screening will identify only those patients whose cancers are likely to be life-threatening during the remaining years of the individual's life.

Post-mortem studies of victims of road accidents reveal that histological changes of prostate cancer can be found at increasing frequency with age from 30 per cent of 40-year olds to 50 per cent of 80-year olds. Of 1000 men in their 50s to 70s approximately 400 will have histological changes of prostate cancer, but only 100 will develop symptoms of the disease in the remaining years of their lives, whilst the 'forces of competing mortality' result in only one-quarter of these dying of the condition. Those anxious to prove that radical prostatectomy will cure the condition are naturally in favour of population screening. Epidemiologists are less convinced of its benefits to society as a whole.

A recent report of a working party of the British Association of Urological Surgeons recommends the following.

1. There is a need for scientifically valid controlled trials. These are currently being undertaken in the United States and Europe by the International Prostate Study for Treatment and Evaluation Group (IPSTEG) and by the European Randomized Study for Screening of Prostate Cancer (ERSPC).
2. Diagnostic tests, particularly prostate-specific antigen (PSA) measurements, should not be used in those for whom they are inappropriate, particularly the very elderly and those with diseases severely limiting their life expectation.
3. No asymptomatic man who requests PSA testing should undergo tests for prostate cancer without adequate counselling as to the possible consequences.
4. The role of PSA as part of a routine protocol for investigating men with lower urinary tract symptoms is controversial.
5. Transurethral ultrasound and biopsy must be carried out by trained operators with appropriate equipment.
6. The diagnosis of early prostate cancer will identify those who need or desire treatment for confined disease.
7. Early prostate cancer should be managed in designated clinics where the patient has access to a urologist and an oncologist.

The approach to a patient who has symptoms or has a genuine fear of having prostate cancer must be different from that adopted towards the population as a whole. Even here, however, it must be remembered that approximately half the men subjected to radical prostatectomy whose PSA at diagnosis is between 4 and 10 ng/ml have disease outside the prostate, and that a PSA limit of 3 ng/ml is probably more effective for population screening since:

1. 12 per cent of tumours are found in men whose PSA is between 3 and 4 ng/ml;
2. almost half the patients diagnosed at screening have extraprostatic disease; and
3. 'PSA failure' (a gradually rising PSA following radical treatment) is found in up to 50 per cent of patients within 5 years of operation.

However, it should be noted that the death rate from prostatic cancer in the United States which rose during the 1990s and peaked in 1996 has since fallen for reasons which are not fully understood, but which may relate to early diagnosis and to radical therapy.

Carcinoma of the kidney

Aetiology and incidence

Though tumours may be produced in animals following irradiation, prolonged administration of oestrogens, exposure to nitrosamines, aromatic agents, and certain alkylating agents, the cause of renal tumours in humans is less well understood. Renal cell cancer is four times as common in males as in females, is linked with smoking, and associated with exposure to cadmium. It is more commonly found in areas with urban or industrial pollution than in rural areas. The incidence in males and females is 10.3 and 5.6 per 100 000 population, respectively.

Rarely, renal cancer appears to run in families, when a defect in the short arm of chromosome 3 is found in many cases (88 per cent in one series). Similar abnormalities are common in non-familial renal cancer and are found uniformly in von Hippel–Lindau disease—an inherited syndrome in which cysts or tumours in the kidney, pancreas, adrenal gland, epididymis, cerebellum, and spinal cord may form. Between one-third and one-half of patients with this condition develop renal cell tumours that are often bilateral and multifocal.

Investigation and diagnosis

Though some tumours are still diagnosed with the classic triad of haematuria, flank pain, and a palpable mass, many tumours are now discovered when completely asymptomatic and at a much earlier stage as a consequence of an incidental upper abdominal ultrasound (Fig. 2). It is also well recognized that renal cancer can present with a variety of apparently unrelated paraneoplastic syndromes. These are a consequence of the production of hormones or cytokines, or perhaps arise from an immune response to the tumour. Relatively common presentations include anaemia, hypertension, pyrexia of unknown origin, fatigue, and an increased plasma viscosity or raised erythrocyte sedimentation rate. Less common presentations include hypercalcaemia, polycythaemia, liver dysfunction, enteropathy, and neuromyopathy.

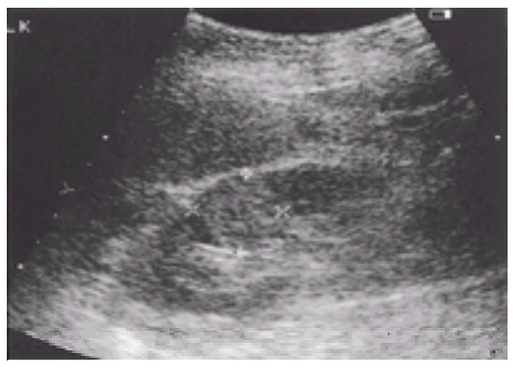

Fig. 2 Renal ultrasound scan showing an incidentally detected, small (2.5 cm diameter) renal cell carcinoma, proved to be stage T1 following nephrectomy. The edge of the tumour is marked by the four crosses.

Until 20 years ago it was necessary to rely upon an intravenous urogram to show evidence of a space-occupying lesion within the kidney and difficult to detect tumours less than 3 cm in diameter. Modern ultrasound and CT can detect and correctly characterize 95 per cent of renal masses greater than 1 cm in diameter. Detection rates for lesions less than 1 cm are 50 per cent.

Survival is crucially dependent upon tumour stage. The free use of upper abdominal ultrasound in the investigation of many patients with upper abdominal symptoms or with symptoms consistent with renal cancer has led to a rapid increase in the diagnosis of early stage tumours that are potentially curable. These small lesions do, however, present diagnostic difficulties to both the pathologist and radiologist whose task it is to separate the renal carcinoma from benign tumours such as oncocytomas, angiomyolipomas, and complex cysts.

The use of ultrasound, CT (Fig. 3), and MRI (Fig. 4), if necessary with image-guided biopsy, can nearly always determine the nature of the tumour and the extent, if any, of lymph node involvement and venous invasion, whilst a bone scan will complete the basic investigations required prior to surgical excision. Tumours diagnosed whilst still small (less than 3 cm) and without extension to lymph nodes or the venous system are likely to have a 10-year survival of 90 per cent if treated effectively.

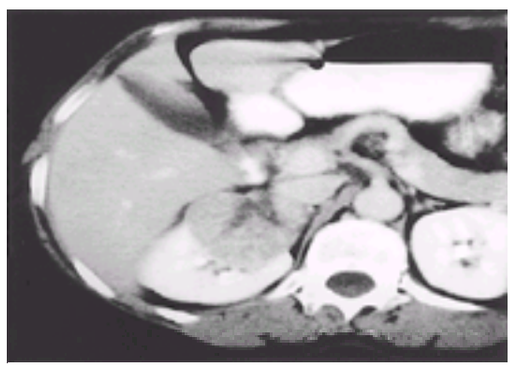

Fig. 3 CT scan showing a renal carcinoma extending into the retrocaval tissues.

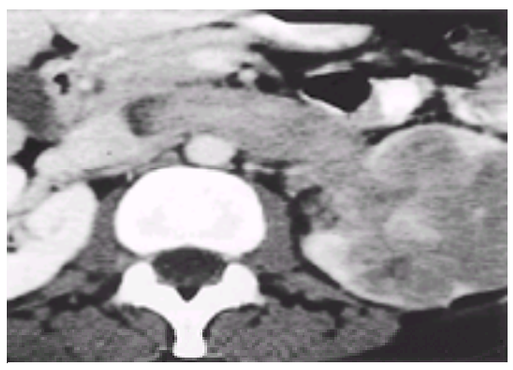

Fig. 4 MR scan showing a renal carcinoma distending and filling the left renal vein and the inferior vena cava.

Treatment

As yet renal tumours can only be cured by surgical excision. The standard treatment is still that of radical nephrectomy in which the affected kidney is removed with its perinephric fat and fascia. It has not yet been established that routine lymph node dissection is of value, but it is recognized that radiotherapy has little role in the management of the primary tumour or local recurrence, although it can be helpful in 'sterilizing' solitary bone metastases.

The use of nephron-sparing surgery is appropriate for smaller tumours, for the patient with a solitary kidney, for those with bilateral tumours, and in patients with von Hippel–Lindau disease. If the partial nephrectomy which such an excision involves leaves half of one functioning kidney, renal function will be adequate. For patients with very small lesions near the cortex of the kidney a simple wedge excision may be sufficient. If local recurrence occurs following partial nephrectomy, the remaining part of the kidney can be removed. It is often impossible for the surgeon to know exactly what sort of operation will be required until 'they get in there'. If there is any likelihood that the patient may be rendered anephric, this prospect should be fully discussed, and the patient introduced to the local renal unit for discussions regarding long-term dialysis before surgery is undertaken.

As yet neither chemo- nor immunotherapy have proved effective, but the use of interferon and interleukin 2 alone or in combination offers a 10 to 40 per cent chance of a partial or complete response in patients with advanced disease following preliminary nephrectomy and in good general condition. Metastases in the lungs are those which respond most frequently. Patients with a complete response have a two-thirds chance of surviving for longer than 1 year. Adoptive immunotherapy and trials of gene therapy remain at the investigational stage.

Advanced disease

For those in whom the diagnosis of renal carcinoma is not made until the local tumour is advanced, or with invasion of the vena cava, nodal involvement, or distant metastases, the outcome remains poor despite therapy. Nephrectomy may be considered for relief of symptoms but is unlikely to be able to offer the prospect of long-term survival. Spontaneous regression of metastases has been reported in fewer than 1 per cent of patients treated by nephrectomy, the responding lesions usually being in the lungs. The surgeon may also be of help to the patient with an apparently solitary skeletal metastasis, since it may be possible to excise it and provide a prosthetic replacement; also to relieve incipient spinal cord compression. In the management of such metastases adjunctive postoperative radiotherapy also plays a role.

Screening for carcinoma of the kidney

This tumour is relatively uncommon, is known to present with a wide variety of symptoms, and 30 per cent of people have metastases at diagnosis. Though no formal screening programme has been suggested, the increasing use of abdominal ultrasound for investigation of problems arising within the upper abdomen has brought to light many small and asymptomatic space-occupying lesions within the kidneys which present a diagnostic problem to the clinician and radiologist. CT scan or MRI with or without needle biopsy can confirm the diagnosis, but the decision as to the type of treatment to be offered is difficult, especially in the elderly patient in whom so many of these lesions are found.

Testicular tumours

Aetiology and incidence

Maldescent increases the risk of testicular cancer approximately fourfold and a range of other abnormalities in urogenital development, such as testicular atrophy or intersex states, have also been associated with an increased risk. This indicates that there are aetiological events associated with embryonic development, but the nature of these events is not known. Excessive oestrogenic exposure may play a role. The incidence has doubled in the last 20 years for reasons that are not understood, but germ cell tumours of the testis are rare, accounting for no more than 1 to 2 per cent of malignancies in males. The incidence (5.1 per 100 000 men) peaks between the ages of 25 and 34 years.

Clinical features, investigation, and diagnosis

From the clinical perspective it is important for the practising physician to remember that whilst most patients notice a lump, a few have symptoms consistent with epididymo-orchitis and a minority present with symptoms of metastases in nodes or other organs. Any disabling symptoms in a male under 35 years of age, such as malaise, loss of weight, backache, and pulmonary symptoms, including haemoptysis, may arise from a previously unrecognized testicular tumour that has metastasized. In a young adolescent or adult male in whom the diagnosis is in doubt or obscure, blood samples for a-fetoprotein and b-human chorionic gonadotrophin should be taken.

Examination of the abdomen in outpatients or on admission must include examination of the scrotal contents. Unless a completely normal testis can be identified, urgent ultrasound of the scrotum is indicated. This is very accurate in determining whether any mass is intratesticular (probably malignant) or extratesticular (probably benign). Samples of blood for a-fetoprotein and b-human chorionic gonadotrophin are mandatory investigations prior to surgery. Any tumour found must be removed without delay. The urologist should deal with such a patient on their next list by removing the testicle with spermatic cord as far as the internal inguinal ring.

Treatment

Following diagnosis by 'excision biopsy', as outlined above, all further management (staging, treatment, and subsequent follow-up) must be in the hands of a specialist cancer centre where expertise should be available to ensure that appropriate additional care is given without delay. Pathological assessment of the orchidectomy specimen is performed to confirm the germ cell origin of the tumour (other diagnoses to be considered include Sertoli or Leydig cell tumours and lymphoma). Germ cell tumours are divided into two main categories, seminoma and teratoma (non-seminomatous germ cell tumours in the American literature), which differ in terms of their relapse rates, patterns of spread, and treatment. Important clinical distinction is also made between patients who have disease clinically localized to the testis at the time of presentation and those who have evidence of metastatic disease.

Localized tumours

Relapse occurs in 16 per cent of patients with seminomas and twice as many with teratomas. Radiotherapy is the treatment of choice in seminoma because it is radiosensitive, and also because it tends to spread in a contiguous manner, with 80 per cent of relapses occurring in the retroperitoneum. Sites of relapse are less predictable in teratoma, and patients considered to be at high risk (generally because of the presence of vascular invasion in the orchidectomy specimen) are given chemotherapy in the United Kingdom. In North America, standard treatment involves a surgical retroperitoneal lymph node dissection and chemotherapy is reserved for those with evidence of established metastatic disease.

Metastatic disease

The dissemination of testicular cancer beyond the testis was a uniformly fatal illness until the 1960s. During the 1970s the introduction of a series of chemotherapy combinations, particularly those involving cisplatin, resulted in improving cure rates for patients with advanced disease. In modern practice, when treatment is delivered carefully by specialized multidisciplinary teams, 90 per cent of patients with testicular cancer, even if disseminated, can expect to be cured.

The International Germ Cell Consensus Classification, based on multivariate analyses of prognostic factors for progression and survival, determined that the most important factors for patients with a testicular primary were the level of serum markers (a-fetoprotein, b-human chorionic gonadotrophin, and lactate dehydrogenase) and the presence or absence of non-pulmonary metastases. The 5-year survival for patients with non-pulmonary visceral metastases was 18 per cent compared with 80 to 92 per cent for patients without such metastases. For the minority of patients for whom conventional therapy is not curative—identifiable by careful consideration of these prognostic factors—therapy including high-dose treatment, new drugs, and combined modality treatment with surgical resection of residual disease can still result in significant benefit.

Screening for tumours of the testis

Testicular self-examination is to be encouraged. Teenagers and men up to the age of 50 years should examine their testicles on a monthly basis, presenting themselves for consultation, diagnosis, and therapy if any abnormality is detected. Any abnormality not otherwise easily explicable should be investigated urgently by testicular ultrasound together with blood samples for a-fetoprotein and b-human chorionic gonadotrophin. Of the urological tumours, this is the only group for which therapy must rightly be given on an emergency basis.

Further reading

General

Vogelzang NJ *et al*., eds (1996). *Comprehensive textbook of genito-urinary oncology*. Williams & Wilkins, Baltimore. [A first class and modern reference work on the topic.]

http://www.nice.org.uk/pdf/urologicalcancerimprovingoutcomes.pdf

Bladder

International collaboration of Trialists on behalf of the Medical Research Council Advanced Bladder Cancer Working Party *et al.* (1999). Neoadjuvant cisplatin, methotrexate and vinblastine chemotherapy for muscle invasive bladder cancer: a randomised controlled trial. *Lancet* **354**, 533–40. [A large international randomized trial to investigate the role of neoadjuvant chemotherapy in patients with invasive bladder cancer.]

Mayfield MP, Whelan P (1998). Bladder tumours detected on screening: results at 7 years. *British Journal of Urology* **82**, 825–8. [A good analysis of a screening study.]

Michaud DS *et al*. (1999). Fluid intake and the risk of bladder cancer in men. *New England Journal of Medicine* **340**, 1390–7. [Simple advice for a complex problem.]

Mills RD, Studer UE (1999). Metabolic consequences of continent urinary diversion. *Journal of Urology* **161**, 1057–66. [A detailed analysis of the consequences and complications of urinary diversion.]

Prostate

Auvinen A *et al*. for the International Prostate Screening Trial Evaluation Group (1996). Prospective evaluation plan for randomised trial of prostate cancer screening. *Journal of Medical Screening* **3**,

97–104. [The outline of the two major randomized screening studies.]

Dearnaley DP *et al*. (1999). Diagnosis and management of early prostate cancer. Report of a British Association of Urological Surgeons Working Party. *British Journal of Urology* **83**, 18–33. [A thoughtful analysis of the present situation.]

Prostate Cancer Trialists Collaborative Group (2000). Maximum androgen blockade in advanced prostate cancer: an overview of the randomised trials. *Lancet* **355**, 1491–8. [A meta-analysis showing only a minimal advantage for combination therapy at the time of first treatment.]

Schroder FH (1995). Detection of prostate cancer. *British Medical Journal* **310**, 140–1.

Schroder FH, Bangma CH (1997). The European Randomised Screening Study for Prostate Cancer ERSPC. *British Journal of Urology* **79**(Suppl 1), 68–71. [An analysis of the results of a pilot study leading to the randomized trial.]

Kidney

Avisrror MU (1998). Renal carcinoma and other tumours. In: Davison AM *et al*., eds. *Oxford textbook of clinical nephrology*, 2nd edn, pp 2573–94. Oxford University Press.

Belldegrun A, deKernion JB (1998). Renal tumours. In: Walsh PC *et al*., eds. *Campbell's urology*, 7th edn, pp 2283–325. Saunders, Philadelphia.

Davidson AJ *et al*. (1997). Radiologic assessment of renal masses: implications for patient care. *Radiology* **202**, 297–305.

Pavone-Macaluso M, Ingargiola GB, La Martina M (1983). Aetiology of kidney tumours. In: Smith PH, ed. *Cancer of the prostate and kidney*, pp 475–88. NATO ASI series, Plenum Press, New York. [An extensive review of possible aetiological factors.]

Vogelzang NJ *et al*., eds (1996). *Comprehensive textbook of genito-urinary oncology*. Williams & Wilkins, Baltimore. [A detailed analysis of what is possible and not possible for the patient with renal cell cancer.]

Testis

Bueton SA (1996). Testicular cancer—to screen or not to screen? *Journal of Medical Screening* **3**, 3–7.

Collette L *et al*. (1999). Impact of the treating institution on survival of patients with 'poor prognosis' metastatic non-seminoma. *Journal of the National Cancer Institute* **91**, 839–46. [Emphasizes the importance of treating patients with advanced disease in large centres.]

Colls BM *et al*. (1992). Results of the surveillance policy of stage I non-seminomatous germ cell testicular tumours. *British Journal of Urology* **70**, 423–8.

Donohue JP, Foster RS (1994). Management of retroperitoneal recurrences: seminoma and non-seminoma. *Urological Clinics of North America* **21**, 761–72.

Scheinfeld J, Bajorin D (1993). Management of the post-chemotherapy residual mass. *Urological Clinics of North America* **20**, 133–43.

Sternberg CN (1993). Role of primary chemotherapy in stage I and low volume stage II non-seminomatous germ cell testis tumours. *Urological Clinics of North America* **20**, 93–109.

20.16 Drugs and the kidney

D. J. S. Carmichael

Introduction

The kidney is the major route of elimination for many drugs and their metabolites. This excretion may be by glomerular filtration, tubular secretion, or in some cases both. In practice a minority of drugs need dose adjustment (dosage and/or interval). The major problems occur in those drugs with a narrow therapeutic range or whose adverse effects are related to the concentration of the drug or its metabolites.

Excretion is affected most by reduction in the glomerular filtration rate, but absorption, distribution (including protein binding), metabolism, and pharmacodynamics may be altered in patients with renal impairment. However, the major determinant of alteration in dosage is the change in drug clearance, which can be estimated by measurement of the glomerular filtration rate, and many handbooks provide guidelines for the adjustment of dosage in renal impairment. Many of these data are derived from measurement or estimation of changes in clearance, half-life ($t_{1/2}$), and volume of distribution (V_d). Renal impairment will often be defined as mild (glomerular filtration rate > 50 ml/min), moderate (glomerular filtration rate > 20 to < 50 ml/min) or severe (glomerular filtration rate < 20 ml/min).

Patients with established chronic renal impairment are often on many medications, either for treatment of the primary disease or its consequences (dialysis, transplantation) or concurrent medical problems. Conversely many patients who are prescribed drugs have impaired renal function (often unrecognized) coincidentally or as a result of other medical problems. Although alteration in pharmacokinetics in renal disease is important, problems are more likely to arise for the following reasons:

i. ignorance of renal impairment before a drug is prescribed;
ii. ignorance of how a drug is cleared from the body;
iii. failure to monitor therapeutic and adverse effects.

Pharmacokinetics

Renal excretion

Renal excretion of drugs depends upon:

i. filtration
ii. active tubular secretion and reabsorption
iii. passive diffusion.

Renal clearance of drugs is a function of the glomerular filtration rate, but tubular reabsorption, secretion, and passive diffusion are contributory factors. If renal clearance is less than the glomerular filtration rate, then tubular reabsorption must be taking place; if it is greater than the glomerular filtration rate, then there must be active tubular secretion.

Compounds with a molecular weight below 60 000 Da are filtered through the glomerulus to a variable extent depending on molecular size, unless they are protein bound when only the unbound portion is filtered. Non-polar (lipid soluble) drugs diffuse readily across tubular cells whereas polar (water soluble) compounds do not. Hence polar drugs generally remain in the tubular fluid and are excreted in the urine, whilst non-polar drugs are reabsorbed by passive diffusion down their concentration gradient into plasma. Some polar drugs are eliminated in the urine as a result of active or facilitated transport mechanisms that transport organic acids or bases (see Table 1). Many drugs are metabolized, primarily in the liver, to produce more polar compounds that cannot be passively reabsorbed and so are eliminated in the urine. In renal failure there may be reduced clearance of these metabolites, which could have therapeutic or adverse effects (see Table 2).

Elimination of organic acids (AH) or bases (B) is affected by the H^+ ion concentration of the tubular fluid, with any change of urinary pH that favours ionization leading to more drug excretion:

$$H^+ + B \; pK_B ?2; \; BH^+$$
$$AH \; ?2; pK_A \; H^+ + A^-$$

The amount of ionized drug at any particular pH is determined by its p K, this being the pH at which 50 per cent of the drug is ionized. If an organic acid has a p K_A of less than 7.5, making the urine alkaline (i.e. increasing its pH) increases the amount of ionized drug (A^-) and therefore its excretion. The converse is true for organic bases with a pK_B of more than 7.5, which are eliminated as the charged (BH^+) form favoured by acid pH. The excretion of salicylates (weak acids) and amphetamines (weak bases) exemplifies these principles.

Although it is an oversimplification to disregard the tubular handling of drugs in renal impairment, both filtration and secretion of drugs appear to fall in parallel and in proportion to the glomerular filtration rate. Hence by far the most important aspect of prescribing in renal disease is awareness of the existence of renal impairment and of changes in renal function: some measure of glomerular filtration rate is needed, serum creatinine measurement usually being sufficient, but more precise measurement is required in some circumstances, for example before the use of known nephrotoxins in chemotherapeutic regimes.

Drugs present in tubular fluid may affect the excretion of other compounds; for example, aspirin and paracetamol reduce methotrexate excretion and probenecid

reduces tubular secretion of penicillins and cephalosporins.

Drug kinetics

Most drugs that are eliminated by the kidney display first-order kinetics, meaning that the rate of removal is proportional to the concentration of the drug. The elimination rate constant k_e is the proportion of the total amount of drug removed per unit time, producing a simple exponential decline (and therefore a straight line on a semilogarithmic plot) in concentration (Fig. 1). The half-life ($t_{1/2}$) of a drug is the time for its plasma concentration to fall by half after absorption and distribution are complete. It is useful in determining dosage interval, drug accumulation (both extent of accumulation and the time taken to reach steady state), and persistence of drug after dosing is stopped. $t_{1/2}$is inversely related to k_e:

$$t_{1/2} = 0.693/k_e$$

(note: 0.693 = ln 2). The clearance of a drug depends upon $t_{1/2}$ (k_e) and the volume of distribution (V_d). The latter does not usually correspond to a real (physiological) volume, although for a drug confined exclusively to the plasma it would approximate to the plasma volume: it represents an apparent volume in which the amount of drug administered would have distributed to produce the measured plasma concentration. The volume of distribution itself may be affected by the protein and tissue binding of drugs, changes in intravascular and extravascular fluid volumes, and lean body mass. Digoxin is one of the few drugs in which a smaller loading dose is needed because of changes in V_d in renal impairment.

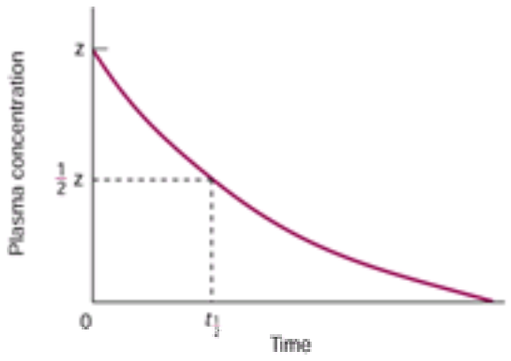

Fig. 1 Plot of log concentration of a drug against time to demonstrate the half-life of a drug.

Clearance can be used to calculate the steady state concentration of a drug (C_{ss}) that can be anticipated in response to any particular dosage regimen. This concentration is proportional to the dose and $t_{1/2}$ of the drug and inversely proportional to the V_d and dosage interval. From these relationships it can be seen that C_{ss} will increase with a longer $t_{1/2}$ and a smaller V_d; it can be reduced either by lowering the dose or by increasing the dose interval.

The relationship between the elimination of a drug in the presence of renal impairment and normal renal function can be estimated and used to gauge the appropriate reduction in dose or increase in dose interval compared with a standard regimen. If non-renal clearance accounts for 50 per cent or more of total drug clearance then no dose adjustment will be needed, provided that non-renal clearance is not affected by renal impairment. If dose reduction is required, then dose, dose interval, or both can be adjusted to reduce dose per unit time. However, in practice the decision whether to reduce dose or prolong dose interval is not always equivalent: for example in the use of aminoglycoside antibiotics, which must achieve a threshold peak concentration in order to kill bacteria effectively, when small but frequent doses may fail to achieve efficacy whereas the same total dose delivered less frequently may achieve the desired therapeutic effect without leading to accumulation and toxicity.

Other pharmacokinetic topics

Gastrointestinal absorption, distribution, metabolism, and renal haemodynamics are other factors that may be altered in the presence of renal impairment. They are less significant than the effects of changes in renal excretion and are dealt with in more specialized pharmacology texts.

Dialysis and haemofiltration

The clearance of drugs by haemodialysis and haemofiltration follows first-order kinetics. Estimates of clearance can be obtained by use of the sieving coefficient, which is the proportion of the drug (or solute) that will cross the membrane and should be constant for a particular drug and membrane (see Table 3). This depends on the drug's molecular weight (and size) and protein binding. Haemofilters have a pore size of 0.01 µm and artificial kidneys for haemodialysis have a pore size of 0.001 µm. In haemofiltration, drugs with a molecular weight below that of inulin (5200 Da) will pass through, whereas in haemodialysis most drugs with a molecular weight below 500 Da (which includes most antibiotics) will be cleared, but drugs such as vancomycin (1800 Da), amphotericin (960 Da), and erythromycin (734 Da) behave differentially with respect to haemofiltration/haemodialysis. Heavily protein bound drugs, even those of low molecular weight (propranolol 259 Da), will not be filtered, but drugs that are displaced from binding sites in the presence of renal impairment will become available for filtration. Water soluble drugs pass through filters more readily than those which are fat soluble. The other differences between the different techniques of haemodialysis, haemodiafiltration, and intermittent and continuous haemofiltration are dealt with in Chapter 20.6.1 and other specialized texts.

Digoxin and antidepressants are examples of drugs with a large volume of distribution that will have low plasma concentrations, hence little of the drug is available for filtration.

Peritoneal dialysis clears drugs very much less efficiently than haemodialysis/filtration, some needing careful adjustment of prescription, for example antibiotics in the treatment of continuous ambulatory peritoneal dialysis peritonitis.

Poisoning

Haemodialysis, peritoneal dialysis, haemofiltration, and haemoperfusion have all been used to eliminate poisons and drugs in overdose from the body. Peritoneal dialysis is very slow but can be used to clear alcohols, salicylate, and lithium if haemodialysis or filtration are unavailable. Similar principles apply to those for the handling of drugs in therapeutic dialysis and filtration. A substance that has a low V_d, low protein binding, and is unipolar will be cleared more efficiently. Note, however, that measurements of the amount of any drug in the body almost invariably report the plasma concentration, and simple estimates of the clearance of any substance will reflect the rate at which the plasma is cleared, ignoring the relationship between the plasma and other compartments. This is a gross oversimplification in some instances, when rebound of plasma concentration occurs as drug redistributes from intracellular compartments.

Haemodialysis

All toxins of size 100 to 2000 Da can be removed using conventional filters, preferably by dialysis with bicarbonate buffering as many toxins will cause acidosis. Should rebound of plasma levels of a substance occur after a period of treatment, as described above, then continuous haemofiltration or repeated haemodialysis may be required; for example thallium poisoning may need prolonged treatment over several days. High-flux membranes used in continuous haemofiltration and haemodiafiltration will remove larger sized molecules (up to 10 000 Da).

Haemoperfusion

Haemoperfusion relies on the affinity of an adsorbent for a toxin and can be used preceding filtration or dialysis in series. Activated charcoal or polystyrene exchange

resins are usually used in a column arranged like an artificial kidney. The technique is particularly useful for toxins with a low V_d: poisons and drugs that can be removed by these means are listed in (Table 4). Further information should be obtained from the National Poisons Centre (Toxbase http://www.spib.axl.co.uk/toxbaseindex.htm).

Prescribing drugs for patients with renal failure

Once the clinician has identified that renal impairment is present, and that a drug to be administered has clinically important renal excretion, the dose can be adjusted in two main ways. Either the size of each dose or the frequency of administration can be reduced. Plasma drug concentrations can be used to confirm that the initial adjustment of dosage is correct in that particular individual. Steady state concentrations of anticonvulsants, digoxin, and theophylline can be measured after the equivalent of five half-lives of the drug. For antibiotics such as gentamicin monitoring is required after the first day's administration and continued if renal function is impaired or changing.

The combination of reduction in dosage and less frequent administration is suitable for most drugs. Dosage reduction alone is more likely to lead to subtherapeutic plasma concentrations. Unfamiliar dosages and administration of drugs at odd times may result in errors, hence adjustments of dose and timing must be kept simple and clear. The physician should use a limited number of drugs and learn about their clearance in renal impairment and changes during renal replacement therapy.

Many of the most complex prescribing problems arise in patients with acute renal failure, particularly if this occurs as part of multisystem failure, perhaps in a patient requiring artificial ventilation and some form of renal replacement therapy.

Acute renal failure on the intensive care unit

Neuromuscular blocking agents

Succinylcholine is rapidly hydrolysed by plasma cholinesterase: no dose adjustment is needed. For more prolonged paralysis atracurium should be used; this is degraded by non-enzymatic Hofmann elimination independent of renal or hepatic function. It is removed by dialysis and haemofiltration: the dose must be titrated to produce a therapeutic effect. Avoid tubocurarine, gallamine, alcuronium, pancuronium, and vercuronium.

Anaesthetic and sedating agents

Propofol, fentanyl, and alfentanyl require no dose adjustment, but the latter two may have prolonged effects if there is concomitant hepatic dysfunction. Metabolites of diazepam accumulate and midazolam is preferable (with dosage reduction) if the glomerular filtration rate falls below 10 ml/min. Phenothiazines, butyrophenones, and chlormethiazole are given in the usual doses.

Treatment of pain or inflammation: analgesics and anti-inflammatories

Narcotic analgesics

Opiates are affected by renal failure, and retention of metabolites can produce adverse effects which may be reduced by intermittent dosing, epidural administration, or low-dose continuous infusions. Diamorphine is metabolized into morphine and then to morphine-3-glucuronide and morphine-6-glucuronide, both of which accumulate to prolong both analgesia and respiratory depression. Morphine and its metabolites are not cleared well by haemofiltration or dialsysis. Pethidine (meperidine) is converted to norpethidine (normeperidine), which accumulates and can cause seizures. Papaveretum is a mixture of alkaloids of opium including morphine, codeine, noscapine, and papaverene: its use is not recommended. Codeine and dihydrocodeine, although weaker analgesics, still have the potential to cause severe respiratory depression in some patients with renal failure, likewise dextropropoxyphene (combined with paracetamol as coproxamol). Its metabolite norproxyphene, which accumulates in renal failure, sometimes causes cardiac toxicity. Buprenorphine is metabolized in the liver and does not appear to have any important toxic metabolites.

If confronted with a patient with renal failure who is inexplicably drowsy or has depressed respiration and who has (or might have) received opiates in the last 72 h, then a trial of intravenous naloxone should be administered: this is sometimes dramatically effective.

Non-narcotic analgesics

Paracetamol (which in overdosage is an important cause of acute renal failure) is excreted in small amounts by glomerular filtration, with some passive tubular reabsorption. Most of the drug is metabolized and the glucuronide and sulphide metabolites, which are subject to active tubular secretion, accumulate in renal impairment, with some regeneration of the parent compound. Despite this, paracetamol is used in the usual doses. Aspirin has the disadvantage of causing gastric damage and increasing the bleeding diathesis of patients with renal failure. Renal elimination of its metabolite salicylate is enhanced in alkaline urine (see above).

Anti-inflammatory agents

Non-steroidal anti-inflammatory drugs including aspirin inhibit the synthesis of prostaglandin by inhibition of cyclo-oxygenase. The principal renal prostaglandins in humans are prostaglandin E_2 and prostaglandin I_2, each of which is vasodilator and natriuretic, having direct effects on both renal blood flow and tubular ion transport. In healthy individuals inhibition of cyclo-oxygenase has no detectable effect on renal function, but in patients with cardiac failure, nephrotic syndrome, liver disease, glomerulonephritis, and other renal disease cyclo-oxygenase inhibitors predictably cause a reversible fall in glomerular filtration rate that can be severe. They can also cause fluid retention and hyperkalaemia. There is evidence that sulindac causes less inhibition of renal cyclo-oxygenase than a dose of ibuprofen that is equieffective on extrarenal tissues and sulindac may cause less renal impairment than other non-steroidal anti-inflammatory drugs. Aspirin may also spare cyclo-oxygenase in the kidney to some extent. The clinical relevance of these observations remains uncertain and caution is needed in severe renal impairment when using any non-steroidal anti-inflammatory drugs.

Indomethacin, azapropazone, and diflunisal have important renal excretion, whereas most other non-steroidal anti-inflammatory drugs are eliminated by metabolism. The non-steroidal anti-inflammatory drugs are highly protein bound and are not removed by dialysis.

Treatment of circulatory disturbance, cardiac disease, or hypertension

Problems can be avoided or minimized by titration of a low starting dose of any drug to produce the required therapeutic effect.

Vasopressors and vasodilators

Adrenaline, dobutamine, dopexamine, dopamine, and noradrenaline should be used in the minimum doses possible to avoid renal vasoconstriction. They all have a short half-life and are not affected by haemo- or haemodiafiltration. Intravenous nitrates are given in the normal dosage. Sodium nitroprusside is metabolized in the liver to sodium thiocyanate, which is eliminated by the kidney and thus may accumulate in renal failure, causing toxicity. It is removed by haemofiltration or haemodiafiltration.

Antiarrhythmics

In patients with abnormal renal function it is advisable to keep treatment simple. Most antiarrhythmic drugs are used without dose modification, for example lidocaine (lignocaine) and verapamil. Digoxin is a notable exception, with a lower loading and maintenance dose than usual. Flecainide and disopyramide require dosage reduction. Amiodarone requires a lower maintenance dose (100 mg daily) when the glomerular filtration rate falls below 20 ml/min.

Diuretics

Spironolactone, triamterene, and amiloride, all potassium sparing diuretics, should be avoided or used with extreme caution in renal impairment because of the danger of hyperkalaemia. The same applies for combination diuretics such as 'Moduretic' (amiloride and hydrochlorothiazide) or 'Dyazide' (triamterene and hydrochlorothiazide). Thiazides, apart from metolazone (a quinolone), become less effective in treatment of both fluid retention and hypertension if the glomerular filtration rate is below 25 ml/min. Higher doses of loop diuretics are needed, but the synergistic effect with metolazone may overcome 'diuretic resistance' in refractory oedema.

Severe sodium and water depletion can occur in diuretic therapy. This may affect renal haemodynamics with secondary effects upon renal function and concomitant drug therapy.

Angiotensin converting enzyme inhibitors

Whether angiotensin converting enzyme inhibitors are being used to treat cardiac failure or hypertension the same precautions apply. Starting doses should be low and increased slowly with careful monitoring of serum creatinine and potassium. Particular caution is necessary if these drugs are used in combination with diuretics or in other high-renin states (for example volume depletion), when marked hypotension ('first-dose effect') may be anticipated. Caution is also necessary when there is (or may be) a possibility of renal artery stenosis, both because of the risk of hypotension and also because of the reduced glomerular filtration rate in the affected kidney(s). Angiotensin converting enzyme inhibitors should not generally be used with potassium sparing diuretics because of the added risk of hyperkalaemia. The kidney eliminates all angiotensin converting enzyme inhibitors, accounting for the reduced dose usually required in the elderly. Exactly the same advice pertains for angiotensin II receptor antagonists.

b-blockers

Atenolol, bisoprolol, pindolol, nadolol, and sotalol are all excreted by the kidney and reduced doses may be needed. The metabolites of acebutolol may accumulate. Other b-blockers are prescribed unchanged.

Vasodilators, calcium channel blockers, and a-blockers

No dose adjustment is needed for any of these drugs. Minoxidil therapy often requires concomitant use of a loop diuretic.

Centrally acting agents

a-Methyldopa, moxonidine, and clonidine are given in the usual doses and titrated for their effect.

Treatment of infection: antimicrobials

Many antimicrobial agents are excreted by the kidney. With the exception of aminoglycosides and vancomycin, most have a wide therapeutic index and little or no dose adjustment is typically made until the glomerular filtration rate is less than 20 ml/min. Antimicrobials that are removed by dialysis should be administered after dialysis, or a supplemental dose given at that time. Adjustments are shown in Table 5.

Penicillins

All penicillins need to be given in reduced dose. Carbenicillin and ticarcillin solutions contain approximately 5 mmol Na^+/g, and caution is needed in the presence of salt and water retention. Mezlocillin (unlike other penicillins) is not removed by dialysis.

Intravenous cephalosporins and other b-lactams

Cephalosporins are excreted in similar fashion to penicillins and need dose reduction in renal impairment (Table 5). The later generation drugs are relatively safe, but caution is still needed with cefuroxime. The carbapenems imipenem and meropenem have a broad spectrum of activity: the former is partially inactivated by a renal dipeptidase and hence administered in combined preparation with cilastatin, an inhibitor of this enzyme. Aztreonam (a monobactam) requires dose adjustment.

Aminoglycosides

Aminoglycosides need dose adjustment in mild renal impairment. Furthermore, they are inherently nephrotoxic and their use may worsen renal impairment as well as causing ototoxicity. Several factors predispose to nephrotoxicity: these include prior or prolonged treatment, hypovolaemia, dehydration, concomitant administration of diuretics, hypokalaemia, and hypomagnesaemia. Obstructive jaundice also increases the risk. The simplest way to prevent aminoglycoside toxicity is to avoid their use altogether in patients with any suspicion of renal impairment and prescribe alternatives.

Nomograms and other guidelines should be used for dose adjustment of aminoglycosides in patients with renal impairment, usually based on conventional multiple dosages. Since the volume of distribution of aminoglycosides is not materially affected by renal impairment, an adequate loading dose is required whatever the method of dose adjustment. Reduction of dose (without alteration in frequency of administration) may lead to an increased likelihood of subtherapeutic peak plasma levels. However, if only the frequency of administration is reduced, then subtherapeutic plasma concentrations are more likely to occur over longer periods. A combination of both methods with frequent peak measurements (taken 1 h after intravenous dosing) and trough measurements (immediately before the next dose) is optimal. An alternative method is to give a single daily dose, there being evidence that gentamicin is less nephrotoxic with an initial dose of 3 to 5 mg/kg body weight, with adjustment made on a daily trough level. Such measurements should be made daily when alterations in renal function are anticipated and two to three times a week under other circumstances. Doses and therapeutic concentrations are shown in (Table 6).

Vancomycin and teicoplanin

Vancomycin is excreted by the kidney and is not dialysed (except at extremely high flow rates in haemofiltration, by haemodiafiltration, or by haemodialysis with high-flux dialysers). In patients with endstage renal failure on dialysis, therapeutic concentrations can be maintained for 5 days or more after a single intravenous dose, the target steady state plasma concentration being approximately 15 mg/l. Following a loading dose of 15 mg/kg, further doses are given on the basis of plasma concentration.

Teicoplanin, a glycopeptide related to vancomycin, behaves rather differently. The half-life is prolonged by approximately threefold in renal failure, and after a loading dose of 400 mg the maintenance dose of 200 mg/day is reduced after 3 days, even in mild renal failure. It is not cleared by dialysis.

Ciprofloxacin

Renal excretion exceeds the glomerular filtration rate, and in patients with normal renal function approximately 60 per cent is cleared by passage through the kidneys. It is recommended that the dose should be reduced in renal impairment, but the proportion that is eliminated by the kidney is reduced in renal failure as a result of an increase in hepatic clearance and of secretion through the wall of the bowel. This becomes more problematic if there is combined renal and hepatic or intestinal failure. Ciprofloxacin is not significantly removed by haemodialysis, but is partially removed by haemofiltration.

Tetracyclines

All the tetracyclines, with the exception of minocycline and doxycycline, are renally excreted. Plasma half-lives are markedly prolonged (up to 100 h) in renal impairment. Tetracyclines are antianabolic and cause a concentration-related increase in blood urea, setting up a vicious cycle leading to deterioration in renal function. Doxycycline or minocycline can be used cautiously in patients with renal impairment, but the other tetracyclines are contraindicated. Demeclocycline, peculiar to itself, has an inhibitory effect on the tubular action of antidiuretic hormone.

Sulphonamides and cotrimoxazole and trimethoprim

Sulphonamides are eliminated by acetylation followed by renal excretion, and acetylated metabolites (which have no antibacterial activity) are a cause of crystalluria and tubular damage. High doses of cotrimoxazole are needed in the treatment of *Pneumocystis carinii*, the risk of adverse effects being balanced against the seriousness of the condition. Such patients often have impaired renal function. The dose is trimethoprim 20 mg and sulphamethoxazole 100 mg/kg body weight/day divided into two or more doses. The plasma concentration should be maintained at approximately 5 to 8 µg/l, measured after five doses. Full dosage should be given initially to those with renal impairment and then reduced if necessary.

Pentamidine

The usual dose is 4 mg/kg, given by slow intravenous infusion over 90 min. There is considerable tissue binding and the drug is excreted in the urine over long periods. The dose should be reduced in patients with renal impairment and, since the drug is nephrotoxic, the dose should be reduced by 30 to 50 per cent if the serum creatinine increases by 88 µmol/l (1 mg/dl).

Antituberculous chemotherapy

Rifampicin and isoniazid (given with pyridoxine) are given in the usual doses. Rifampicin is a potent inducer of the cytochrome P-450 system and will therefore affect cyclosporin metabolism. The dosage of ethambutol and pyrazinamide needs to be reduced. Capreomycin is nephrotoxic and, although streptomycin can be used, careful monitoring after each dose is required.

Antiviral agents

Aciclovir and ganciclovir are both eliminated by the kidney and both are dialysed. Similar dose reductions are needed for each (see Table 5). Antiretroviral drugs such as lamivudine, zalcitabine, and zidovudine all need reduced dosage.

Antifungal agents

Amphotericin is nephrotoxic: it should only be used with great caution in patients who already have renal impairment and discontinued if the plasma creatinine concentration exceeds 260 µmol/l. Liposomal amphotericin avoids this toxicity. Amphotericin is cleared by haemodialysis and should therefore be given after treatment. Both flucytosine and fluconazole are excreted in the urine. Fluconazole should be given at a dose of 200 mg/day after an initial dose of 400 mg. Ketoconazole is less well absorbed in renal failure and interferes with cyclosporin metabolism through action on the cytochrome P-450 system (see Table 7).

Antiprotozoal agents and malaria

Quinine is given in the usual doses unless acute renal failure develops, in which case the dose is reduced after two to three days. The dose of chloroquine is reduced by half if the glomerular filtration rate is less than 50 ml/min and to a quarter if the glomerular filtration rate is less than 10 ml/min; primaquine is given in the usual doses. For prophylaxis chloroquine (usual dose 300 mg/week) can be given to patients with renal impairment. Proguanil (usual dose 200 mg daily) should be given in half the usual dose if the glomerular filtration rate is less than 10 ml/min. Mefloquine has been used in haemodialysis patients in a dose of 250 mg weekly: it is not cleared by dialysis and no adverse effects were reported.

Drugs acting on the central nervous system

Drugs acting on the central nervous system may have a prolonged effect in renal failure, not only because of changes in pharmacokinetics, but also because of increased sensitivity as a consequence of uraemia.

Antidepressants

The data on antidepressant drugs are conflicting, but all should be used with caution. Tricyclics are given in the usual dosage. Fluoxetine, paroxetine, and other selective serotonin reuptake inhibitors have been used widely in patients on dialysis, although dosage reductions are advised. It is best to avoid citalopram and venaflaxine.

Lithium

Lithium is filtered and then reabsorbed, mainly in the proximal tubule. The dose should be reduced in renal impairment with careful monitoring of plasma concentration. In sodium depletion (for example with chronic use of thiazide diuretics) tubular reabsorption of lithium is increased, leading to higher plasma concentrations and toxicity. Lithium is a cause of chronic tubulointerstitial damage.

Major tranquillizers

No dose change is required when phenothiazines or butyrophenones are used in patients with renal impairment. Newer drugs such as clozapine, rispiridone, and sulpiride should be used with caution.

Minor tranquillizers

Benzodiazepines can be prescribed in the usual dosage. Diazepam and chlordiazepoxide have active metabolites that may accumulate in renal failure: drugs without active metabolites such as nitrazepam and temazepam may avoid hangover the morning after use as night sedation.

Anticonvulsants

Phenytoin, carbamazepine, and valproic acid are given in the usual dosage. Protein binding of phenytoin is reduced in renal impairment with a rise in the free (active) fraction such that plasma or serum levels may need to be adjusted downwards. Vigabatrin and gabapentin need dose reduction.

Antihistamines

Terfenadine should be avoided because of prolongation of the QT interval, perhaps made more likely in renal impairment. Prochlorperazine and chlorpheniramine are used in the usual dosage but may cause drowsiness.

Treatment of hyperlipidaemia and diabetes mellitus

Lipid lowering agents

There is now wide experience in the use of HMG-coenzyme A reductase inhibitors. Simvastatin and pravastatin are given in the usual doses, that of fluvastatin is reduced. All may cause myopathy and myositis, particularly in renal impairment or if used with cyclosporin or gemfibrizol. The fibrates (gemfibrizol, bezafibrate) can be used with dose reduction at a glomerular filtration rate of less than 20 ml/min.

Insulin

Insulin requirements fall with declining renal function, probably as a consequence of the reduced metabolism of insulin by the kidney in both acute and chronic renal

failure. In patients on haemodialysis it is often necessary to give supplemental insulin during treatment. The same situation applies in patients on haemofiltration for acute renal failure, particularly if they are being fed parenterally, and in continuous haemodiafiltration when the dialysate is a glucose-based solution. Non-diabetic patients may require insulin temporarily under these circumstances. Patients on continuous ambulatory peritoneal dialysis may need a change in insulin preparation and adjustment in the frequency and route of administration. The intraperitoneal requirement is approximately 50 per cent of that needed intravenously.

Oral hypoglycaemic agents

Glicazide, gliquidone, and glipizide are the safest drugs to use, although dose reduction may be needed if the glomerular filtration rate is below 10 ml/min. Other sulphonylureas, particularly chlopropamide, have a prolonged half-life. The biguanides should not be used if the glomerular filtration rate is below 20 ml/min.

Treatment of asthma

b-Agonists administered by inhalation, oral, or parenteral routes need no adjustment in patients with renal impairment, although tobuterol is an exception. Aminophylline and theophylline can be given in the usual doses but metabolites may accumulate. The leukotrione antagonist zafirlukast (but not montelukast) needs dosage reduction.

Treatment of gastrointestinal disorders

H_2-Antagonists and antiulcer drugs

Cimetidine is cleared by the liver but metabolites accumulate if the glomerular filtration rate is less than 20 ml/min. Ranitidine, which causes less cerebral confusion, is preferable in this situation, but may interfere with creatinine secretion and raises plasma creatinine. It is partly cleared by the kidneys (in common with famotidine) and the dose should be halved when the glomerular filtration rate less than 10 ml/min. It is dialysed, and a supplemental dose is needed after dialysis but not after haemofiltration. Omeprazole and misoprostol are given in the usual doses. Misoprostol may cause reductions in glomerular filtration rate through haemodynamic changes in the kidney.

Antacids

Alginates, magnesium trisilicate mixture (but not magnesium trisilicate powder), and sodium bicarbonate all have high sodium contents. The use of aluminium containing compounds, such as aluminium hydroxide or sulcralfate, in patients with severe renal impairment or those on dialysis is controversial because of the potential risks of aluminium retention with deleterious effects on bone, bone marrow, and the central nervous system. Calcium carbonate should not be used as an antacid but only as a phosphate binder.

Treatment of hyperuricaemia and gout

Allopurinol

Allopurinol is metabolized to oxypurinol, which is retained in renal impairment and may be responsible for some of the adverse effects including rashes, bone marrow depression, and gastrointestinal upset. The dose should be reduced to 100 mg/day when the glomerular filtration rate is less than 20 ml/min. The dose should be given after haemodialysis. Allopurinol interferes with the metabolism of 6-mercaptopurine (an active metabolite of azathioprine) causing accumulation and toxicity (for example leucopenia).

Probenecid

Probenecid inhibits secretion of acids in the proximal tubule and prevents reabsorption of urate from the tubular lumen. It prolongs the effect of penicillins, cephalosporins, naproxen, indomethacin, methotrexate, and sulphonylureas (all of which are weak acids), causing accumulation and the potential for toxicity. It also inhibits tubular secretion (and hence activity) of furosemide (frusemide) and bumetanide.

Colchicine

Colchicine has been largely replaced by non-steroidal anti-inflammatory drugs (see above) for the treatment of acute gout. However, it remains valuable in patients in whom non-steroidal anti-inflammatory drugs are undesirable (for example in those with peptic ulcer disease, cardiac failure, or renal impairment), and can be used without dose adjustment in renal failure.

Treatment or prevention of thrombosis and thromboembolism: anticoagulants and antiplatelet agents

Warfarin is used in the normal dosage and its effect is monitored by measuring prothrombin time in the usual way. It is highly protein bound and there may be slight displacement and consequent reduction in the volume of distribution in uraemia. In nephrotic patients hypoalbuminaemia leads to an increased sensitivity to warfarin, which is not removed by dialysis. Heparin is used in the normal dosage, but dosage reduction is required for tinzaparin. Prophylactic aspirin is given in the usual low dose (75–150 mg/day), but the dosages of clodiprogel and ticlopidine should be reduced.

Treatment of autoimmune rheumatic or vasculitic disorders: corticosteroids and immunosuppressive agents

Prednisone and prednisolone are not eliminated by the kidney. Methylprednisolone is cleared by haemodialysis, and should be given after dialysis. Azathioprine accumulates in renal impairment and the dose should be reduced from a maximum of 3 mg/kg/day to 1 mg/kg/day if the glomerular filtration rate falls below 10 ml/min. The dose of cyclophophamide, if given intravenously for systemic lupus erythematosus or systemic vasculitis, should be at the lower end of the therapeutic range with careful monitoring of the blood count before further doses are given.

Methotrexate is used frequently in rheumatological disorders (rheumatoid arthritis, psoriatic arthropathy), usually as a small (7.5 to 20 mg) weekly dose. It should be noted that the drug is a weak acid and is eliminated by proximal tubular secretion, which can be blocked by salicylates or non-steroidal drugs.

Cyclosporin is a highly lipid soluble drug that is extensively bound to plasma proteins and has a large volume of distribution. It is metabolized in the liver via the cytochrome P-450 system by mono- and dihydroxylation as well as *N*-demethylation. Only minor amounts are excreted as the parent drug or metabolites in the urine. Renal impairment does not affect its metabolism. However, since many other drugs may be prescribed to patients on cyclosporin therapy, several important interactions may occur. These may both increase plasma concentration and therefore increase the risk of nephrotoxicity, or reduce plasma concentrations to increase the risk of transplant organ rejection. Aminoglycosides may have an additive effect upon the nephrotoxicity itself. The common interactions are listed in Table 7.

Miscellaneous drugs

Acetazolamide

Acetazolamide may produce electrolyte disturbance, particularly in renal impairment and in the elderly. Its use should be avoided or carefully monitored.

Bisphosphonates

It is appropriate to use intravenous disodium pamidronate or etidronate in hypercalcaemia caused by malignancy, even if this is causing renal failure. The likely outcome is an improvement in renal function, particularly if other measures such as sodium and water depletion are addressed. Bisphosphonates (with the exception of alendronic acid) can be used to treat postmenopausal or corticosteroid induced osteoporosis and Paget's disease, with dosage reduced in moderate renal

impairment. Etidronate may cause hypercalcaemia if combined with calcium supplementation.

Anticancer drugs

The kidney excretes many anticancer drugs or their metabolites, and doses need to be calculated with accurate measurements of glomerular filtration rate (for example cisplatin). It is important to obtain appropriate information before prescribing.

Summary

- Always check the method of elimination of any drug before prescribing in the presence of known or suspected renal impairment.
- Monitor any changes in renal function.
- Look out for any adverse or side-effects.

Further reading

Barclay ML, Kirkpatrick CMJ, Begg EJ (1999). Once daily aminoglycoside therapy. *Clinical Pharmacokinetics* **36**, 89–98.

Bellisant E, Sebilla V, Vaintand G. (1998). Methodological issues in pharmaco-pharmacodynamic modeling. *Clinical Pharmacokinetics* **35**, 151–66.

Benet LZ, Kroetz DL, Sheiner LB (1996). Pharmacokinetics: the dynamics of drug, absorption, distribution, and elimination. In: Hardman JG and Limberd LE, eds. *Goodman and Gilman's the pharmacological basis of therapeutics*, section I, pp 3–28. McGraw-Hill New York.

Benet LZ, Zia-Amirhossaini P (1995). Basic principles of pharmacokinetics. *Toxicology and Pathology* **23**, 115–23.

Bohler J, Donauer J, Keller F (1999). Pharmacokinetic principles during continuous renal replacement therapy: drugs and dosage. *Kidney International* **72**, S24–S28.

Bonate PL, Reith K, Weir S (1998). Drug interactions at a renal level. *Clinical Pharmackinetics* **34**, 375–404.

Carmichael DJS (1998). Handling of drugs in kidney disease. In: Davison AM *et al.*, eds. *Oxford textbook of clinical nephrology*, pp.2659–78. Oxford University Press, Oxford.

Czock D, Keller F (1999). The area under the effect-time curve as a target for dosage adaptation in renal insufficiency. *Nephrology, Dialysis and Transplantation* **14** (suppl. 4), 4.

Davies G, Kingswood C, Street M (1996). Pharmacokinetics of opiods in renal dysfunction. *Clinical Pharmacokinetics* **31**, 410–22.

Golper TA, Marx MA (1998). Drug dosing adjustments during continuous renal replacement therapies. *Kidney International* **66**, S165–S168.

Hammertstein A, Derendorf H, Lowenthal DT (1998). Pharmacokinetic and pharmacodynamic changes in the elderly. *Clinical Pharmacokinetics* **35**, 49–64.

Joos B, Schmidli M, Keusch G (1996). Pharmacokinetics of antimicrobial agents in anuric patients during continous venovenous haemofiltration. *Nephrology, Dialysis and Transplantation* **11**, 1582–5.

Keller F *et al.* (1999). Individualized drug dosage in patients treated with continuous hemofiltration. *Kidney International* **72**, S29–S31.

Keller F, Czock D (1999) Pharmacodynamic half-life and effect-time in renal impairment. *Nephrology, Dialysis and Transplantation* **14** (suppl. 4), 6–8.

Kramer BK, Schweda F, Riegger GAJ (1999). Diuretic treatment and diuretic resistance in heart failure. *American Journal Of Medicine* **106**, 90–6.

Joest M, Ritz E, Mutschler E (1999). Renal handling of drugs in the healthy elderly. *European Journal of Clinical Pharmacology* **55**, 205–11.

Laville M *et al.* (1989). Restrictions on use of creatinine clearance for measurement of renal functional reserve. *Nephron* **51**, 233–6.

Taylor CA *et al.* (1996). Clinical pharmacokinetics during continuous ambulatory peritoneal dialysis. *Clinical Pharmackinetics* **31**, 293–308.

Toxbase http://www.spib.axl.co.uk/toxbaseindex.htm.

21.1 Epidemiology

M. W. Adler and A. Meheus

Introduction

The sexually transmitted infections (**STIs**) are mainly spread by sexual intercourse (Table 1). However, some, such as genital candidosis, are only rarely spread in this way; others, like scabies and pediculosis pubis, are spread by close bodily contact without penetrative intercourse. The range of diseases spread by sexual activity continues to increase, with familiar bacterial and treponemal infections now being superseded in developed countries by herpes, warts, and human immunodeficiency virus infection (**HIV**). The detrimental effect of STIs on pregnancy and the newborn (for example, miscarriage, prematurity, congenital and neonatal infections, blindness) are more common and severe than had previously been realized. Furthermore, complications such as pelvic inflammatory disease, ectopic pregnancy, infertility, and cervical cancer are major health problems. The World Bank in 1993 estimated that for women aged 15 to 44 years, STIs, excluding HIV infection, were second only to maternal morbidity and mortality as causes of healthy lives lost.

Their incidence, distribution, and risk of complications is strongly influenced by a wide array of determining factors, including behavioural and sociocultural factors, population composition, susceptibility of individuals, changing characteristics of pathogens, and society's efforts at primary prevention and disease control.

Since the advent of HIV/AIDS, increased attention is being given to the other STIs, because they are important causes of morbidity and mortality in their own right and are important markers of behaviour associated with a high risk of HIV transmission. Some 80 per cent of HIV infection is spread sexually. STIs, especially genital ulcer disease, can enhance the acquisition and transmission of HIV by damaging the mucosa or skin, due to increases in HIV-susceptible macrophages and increased viral shedding.

The global burden of STIs is unknown because of the lack of effective control and notification systems in some countries. The World Health Organization (WHO) has estimated a total of 340 million new cases of curable STDs in adults per annum, mainly in South-East Asia (151 million new cases per year) and Sub-Saharan Africa (69 million). In Eastern Europe and Central Asia, the estimate is 22 million, and 17 million in Western Europe. The prevalence and incidence varies regionally: for instance, between Sub-Saharan Africa and Western Europe, 4.6- and 3.3-fold, respectively (Table 2).

The diseases

The accuracy of European data suffers from the reluctance of private doctors, who may see most of the patients, to notify the appropriate authorities. Notification systems include STD clinics (United Kingdom, France, Italy), laboratory/systems (Denmark, Sweden), and sentinel general practitioner (**GP**) systems (Belgium).

In the United Kingdom, a notification system was established in 1916, which also provided a free and confidential service for people with sexually transmitted infections. This service, mainly run outside hospitals, was integrated into the hospital structure following the creation of the National Health Service in 1948. At that time, syphilis and gonorrhoea made up the majority of the 132 000 cases diagnosed in clinics in England and Wales. Since then, the number of people attending clinics has increased exponentially and the disease profile has changed (Fig. 1). Gonorrhoea and syphilis now represent less than 2 per cent of all cases seen, while the new viral diseases, in particular genital warts and herpes, increased by 236 per cent and 160 per cent, respectively, between 1980 and 2000.

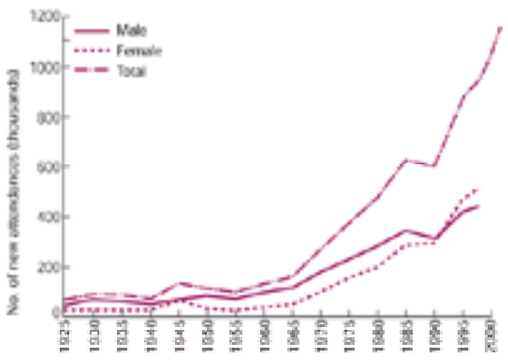

Fig. 1 New attendance at genitourinary medicine (**GUM**) clinics England and Wales 1925–2000.

Gonorrhoea

Although rates of gonorrhoea vary between countries, there was a general upward trend over most of the last 40 or so years, with a flattening out in the mid-1970s and a subsequent decrease in the number of cases reported with recent increases in some countries. It is difficult to interpret differences between countries because of the variation in reporting practices and the provision of facilities.

Data tend to be more complete in developed countries. However, standards vary and countries like the United Kingdom, which has a network of genitourinary medicine/STD clinics and routine notification requirements, produce more accurate figures than some other European and North American countries where most patients are treated by private physicians who do not usually report cases. There was a peak in the number of cases of gonorrhoea during the early to mid-1970s in most European countries. The advent of AIDS/HIV infection in the 1980s led to safer sexual practices and a reduction in the number of cases of gonorrhoea, but this has not been sustained in all countries. Recently, there has been a substantial increase in both male and female cases of gonorrhoea in England and Wales. Between 1995 and 2000 there was a 105 per cent increase in the number of cases in men, from 6759 to 14 231, and in females an 85 per cent increase from 3394 to 6289. In males, this increase was seen in all age groups, particularly in those aged 16 to 19 and 35 to 44 years, but in females, it was mostly in teenagers. The incidence of gonorrhoea has increased since 1993 in homosexual men, particularly in those living in London and its immediate surrounding area.

In Nordic countries, the annual incidence of gonorrhoea declined (Fig. 2) from over 100 per 100 000 population in most countries in the early 1980s to less than 10 per 100 000 population by the late 1990s, but there have been recent slight increases. Although the incidence of gonorrhoea has declined in the United States, there

are considerable differences between ethnic groups (Fig. 3). In 2000 there was a total of 358 995 cases.

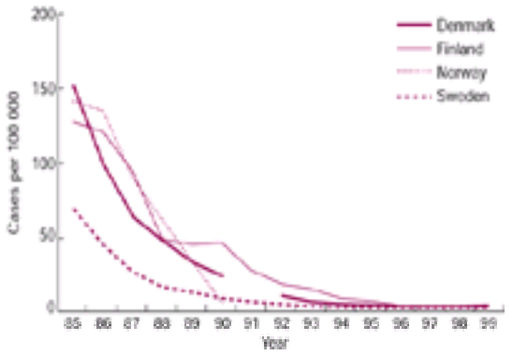

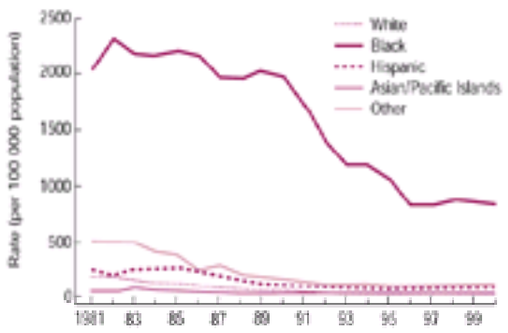

Fig. 2 Annual incidence of gonorrhoea per 100 000 population in Nordic countries.

Fig. 3 Gonorrhoea—rates by race and ethnicity: United States, 1981–2000, and the 'Healthy People year 2000 objective'. 'Other' includes Asian/Pacific Islander and Native American/Alaskan populations. Georgia did not report gonorrhoea statistics in 1994.

There is an epidemic of STIs in Eastern Europe, in the newly independent states of the former Soviet Union. The highest rates of gonorrhoea are in Estonia (166), Russia (139), and Belarus (125) per 100 000, compared to France 18.5, Germany 5, The Netherlands 8, and the United Kingdom 22 per 100 000.

Syphilis

The dramatic impact of penicillin on the incidence of early infectious syphilis throughout the world in the 1950s has not been maintained everywhere. The United States has experienced a continuous increase in the total number of cases of primary and secondary syphilis of seven- to ninefold in males and females, respectively, since 1956. These increases were particularly noticeable in the 1990s and are partly explained by the deployment of resources away from traditional STD control programmes to those for AIDS. However, the control of HIV and AIDS will only come through integrated STD/AIDS control programmes.

In most Western European countries, but particularly in Scandinavia, there has been a decline in incidence to below 5 per 100 000. However, in Eastern Europe there is an epidemic of syphilis in all Newly Independent States (**NIS**) of the former Soviet Union. The 1999 incidence of syphilis in these NIS ranged from 55 to 180 per 100 000 (Fig. 4); increases are particularly evident in older adolescents. Between 1986 and 1996, the incidence of syphilis increased in 18- to 19-year-old Russians, from 6 to 607 per 100 000 in men and from 20 to 1321 per 100 000 in women. The increase in Russia between 1992 and 1996 was 20-fold, Estonia 6-fold, Latvia 12-fold, and Lithuania 14-fold. An HIV/AIDS epidemic can be predicted in Eastern Europe, and already there have been outbreaks of HIV among intravenous drug users, particularly in Belarus, Russia, and Ukraine.

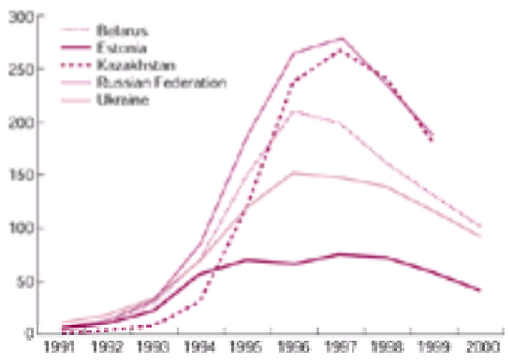

Fig. 4 Annual incidence of syphilis in Belarus, Estonia, Kazakhstan, the Republic of Moldova, the Russian Federation, and Ukraine, 1990–2000 (rate per 100 000 of the population).

There has been an encouraging fall in the incidence of congenital syphilis in developed countries, largely due to the control of early acquired infectious syphilis in women and the screening of all pregnant women for syphilis. However, congenital syphilis is still a major health problem in the countries of the former Soviet Union and in many developing countries.

Chlamydia

The introduction of antigen detection tests such as direct immunofluorescence techniques and enzyme immunoassays in the 1990s, and, more recently, sensitive nucleic acid detection-based tests such as polymerase or ligase chain reactions, has allowed a more widespread screening for chlamydia in most European countries. After the start of wide-scale screening in Sweden during the 1980s, the number of cases declined from 38 000 in 1987 to 14 000 by 1997, with an associated decrease in ectopic pregnancies.

In England and Wales there has been no such decline, and chlamydia remains a major public health problem; long-term sequelae associated with chlamydial infection include pelvic inflammatory disease, ectopic pregnancy, infertility, and abdominal pain. Genital *Chlamydia trachomatis* infection is now the commonest curable bacterial STI in England and Wales. There has been an increase in the number of cases since 1993, with females outnumbering males; in 2000, 63 037 people attended clinics—27 222 males, 35 815 females. It is commonest in young people: the peak age in men is between 20 and 24 years, and between 16 and 19 in women. Screening surveys carried out in antenatal and gynaecological clinics, general practice, family planning units prior to pregnancy termination, and those attending STD clinics have shown median prevalences ranging from 4.5 to 16.4 per cent. Similar prevalence rates have been seen in the United States. The increased availability of Chlamydia testing and more sensitive detection tests will, to some extent, account for the apparent increase in the number of cases seen.

Genital herpes and warts

The greatest increase in STIs in England and Wales during the 1980s was in the number of reported cases of genital herpes and warts. In 1978, 8406 cases of herpes and 24 136 of warts were seen in STD clinics, increasing to 30 199 and 100 124, respectively, by 2000 (Fig. 5). Compared to Chlamydia and gonorrhoea there has been a slowing down in the increase in the last few years.

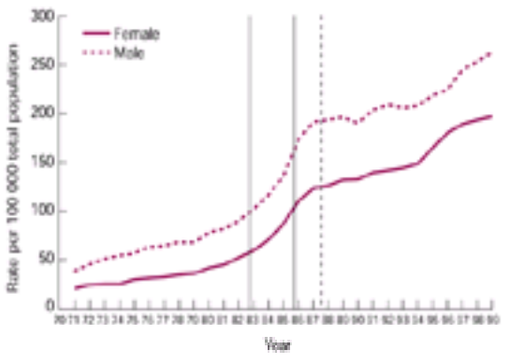

Fig. 5 New cases of genital warts infection seen in GUM clinics, 1970–1999 (England).

Genital herpes

There are between 200 000 and 500 000 new cases of herpes each year in the United States, with a prevalence of approximately 40 million cases. These extrapolations from STD clinic data may not be valid, for most of these patients attending such clinics are of lower socioeconomic status and unrepresentative of the whole population. In a Swedish STD clinic, herpes simplex virus (**HSV**) was isolated from the cervices of 8.4 per cent of all female attenders.

HSV-1 and -2 antibody tests allow prevalence studies to be carried out in different populations. In the United Kingdom, HSV-2 seroprevalence is low in blood donors (3 per cent in men, 12 per cent in women), intermediate in pregnant women (4 per cent in those under 20 years of age, 11 per cent in the 20 to 29 age group, and 16 per cent in those over 29 years), and high in STD clinic attenders (21 per cent in men, 25 per cent in women). In the United States, the pattern is similar but rates are approximately twice as high.

In the United Kingdom, STD clinic attenders showed an increasing proportion of HSV-1 genital infections during the 1980s and early 1990s. HSV-1 now accounts for most first episodes in women. The proportion of HSV-1 in genital specimens was 48 to 68 per cent in women, 25 to 35 per cent in men. In the United States most genital infections are caused by HSV-2.

Genital human papillomavirus (HPV)

Genital warts is now the commonest sexually transmitted disease seen in STD clinics in England and Wales (100 124 cases in 2000). They are difficult and time-consuming to treat. HPV infection is of particular concern because certain types are associated with cervical dysplasia and cervical cancer (see Chapter 7.10.17).

Pelvic inflammatory disease

Pelvic inflammatory disease (**PID**) is the most serious complication following gonococcal, chlamydial, and non-specific infections. Its incidence and prevalence is increasing in most countries. In Western industrialized countries its estimated annual incidence is 10 per 1000 women aged 15 to 39 years, with a peak incidence of 20 per 1000 in the age group 15 to 24 years. Risk factors include STIs, the use of intrauterine devices (**IUDs**), and postabortion and puerperal infections.

STIs cause most cases of PID. In developed countries, 75 per cent of cases in under 25-year-old women are attributable to STIs. In Uppsala, 50 per cent of patients with PID in 1965 had cervical gonorrhoea, but by 1975 this had dropped to approximately 10 to 15 per cent, at which time *Chlamydia trachomatis* was found to be responsible for 60 per cent of cases.

The most serious consequence of PID is infertility (see Chapter 21.4). It is difficult to obtain accurate data on the trends for PID since not all patients are hospitalized or correctly diagnosed.

STD epidemiology in developing countries

The frequency of STIs is much higher in developing countries. They are among the top five causes of consultation at general health services in many African countries. The incidence and prevalence of these infections are very high in specific population groups, such as female prostitutes and their clients. Prostitution is an important factor in the transmission of STIs in developing countries.

Genital ulcers are much more frequent. The so-called 'tropical STIs'—in particular chancroid and, to a lesser degree, lymphogranuloma venereum and granuloma inguinale—are major causes. More genital ulcers in developing countries are caused by syphilis than in industrialized countries; genital herpes accounts for a smaller proportion, but has become the leading cause of genital ulcers in areas of high HIV/AIDS incidence.

The incidence of STI complications and their sequelae is much higher in developing countries due to the lack of resources for adequate diagnosis and treatment. Important STI complications and sequelae include adverse pregnancy outcome for mother and newborn, neonatal and infant infections, infertility in both sexes, ectopic pregnancy, urethral stricture in males, blindness in infants due to gonococcal and chlamydial ophthalmia neonatorum and in adults due to gonococcal keratoconjunctivitis, as well as genital cancers, particularly cancer of the cervix uteri and penis.

The epidemiology of HIV infection and AIDS is very different from that in Western countries: level of sexual activity, not sexual orientation is, apparently, the major risk factor. HIV is predominantly transmitted heterosexually in developing countries. Genital ulceration, and other STIs, facilitate the sexual transmission of HIV. Many governments and international donor agencies have tended to ignore the real magnitude of the problem. It needed a fatal STI to alert decision-makers worldwide and the community to the STI problem and to generate resources for its prevention and control.

Control

Prevention is more effective and cheaper than treatment. The objectives of STI control are to interrupt the transmission of STIs, to prevent their development, complications, and sequelae and to reduce the risk of sexual transmission of HIV. These aims can be achieved by primary and secondary prevention and comprehensive patient care (Table 3).

Primary prevention

Primary prevention is achieved through health promotion (information, education, communication— **IEC**) and counselling activities (see Chapter 3.5). The aim is to educate people about the advantages of discriminative sex and prophylaxis ('safer sex'). Clearly, the best way to avoid STIs is to avoid sexual intercourse. This may not be acceptable to those who are already sexually active, but the dangers of frequent changes of sexual partners and the methods of reducing the risk must be emphasized. It is essential to use a condom to avoid contact with partners who have symptoms or lesions, and to have regular check-ups. This must be taught to

children before they become sexually active. The best place for this is at home or school in the context of growing up, understanding one's body, and being responsible for one's own health. Since the arrival of HIV infection and AIDS, sexuality has become less of a taboo in many countries. Health workers and health educators must, therefore, take advantage of this new openness. Through mass media campaigns targeted at the general public, IECs have become part of STD/AIDS control nearly everywhere.

A special target group for health promotion and individual counselling are patients with STIs; such infections prove them to be at risk of contracting STIs, including HIV infection. Their current infection and the awareness of their vulnerability might make STD patients more inclined to change risky sexual behaviour.

No effective vaccines are as yet available for STIs, apart from that for hepatitis-B virus (**HBV**).

Secondary prevention

P>Here the aim is to detect STIs early through screening ('check-up') and by education about what to do once disease is suspected (healthcare-seeking behaviour). Sexual intercourse should be stopped until medical care has been sought. Education should indicate how and where such advice can be found, and emphasize the importance of adhering to the treatment and advice.

Screening aims to detect asymptomatic or mildly symptomatic infections, and is carried out in specific populations if the prevalence of STIs is high. Good laboratory support is essential. Pregnant women may be screened for syphilis, HIV, or chlamydial infection. Where prostitution is legal or tolerated, prostitutes can be screened for various STIs. Blood donors are screened for HIV, HBV, and syphilis.

Adequate and comprehensive management of STI patients

The management of infected patients is a cornerstone of STI control. The aims of patient care are: to detect or rule out infection; to give treatment if necessary; to educate and counsel on treatment compliance and on STD/HIV prevention and condom use; to ensure that sexual partner(s) are evaluated and managed (contact tracing); and eventually to test for other STIs, including HIV infection. In a developed country, it is not appropriate to attempt the management of STIs without microbiological facilities, since doctors providing care in these countries usually have specialist knowledge of the diseases. However, in developing countries with limited resources, it is more realistic to use a syndromic approach based on STI signs, symptoms, and simple laboratory tests. To implement major intervention strategies, the following support components need to be developed:

- *Training* should be given to health workers and health educators, for instance in the use of flow-charts to simplify the management of STD patients, or to strengthen their health education and counselling skills.
- *Laboratory services* need to be expanded, depending on the level of healthcare provided. A reference laboratory should be developed in each country to allow the quality control and analysis of referred specimens.
- *Research* should be undertaken to include epidemiological and sociobehavioural baseline studies, assessment of antimicrobial sensitivity, as well as operational research to render the programme more cost-effective.
- *Information systems* or surveillance should be implemented to gather epidemiological data for magnitude and trend assessments, and to provide data for programme planning and monitoring. Various surveillance methods can be used—clinician notification, laboratory notification, sentinel site surveillance (either of syndromes or of aetiological diagnoses), prevalence studies in specific population groups, and aetiological surveys in patients.

Conclusion

A successful STI control programme, by reducing both the incidence and prevalence of STIs, will reduce the morbidity, suffering, and economic cost associated with these diseases. By eliminating STIs as a facilitating factor in HIV transmission, and by contributing to behavioural changes towards safer sex, it will play an important role in the prevention and control of HIV/AIDS.

Further reading

Adler MW (1998). *The ABC of sexually transmitted diseases*, 4th edition. British Medical Association, London.

Adler MW (1980). The terrible peril—a historical perspective on the venereal diseases. *British Medical Journal* **281**, 206–11.

Adler MW (1999). Cinderella and the glass slipper: the growth and modernisation of a specialty. *Sexually Transmitted Infections* **75**, 439–44.

Adler MW, *et al*. (1998). Sexual health and healthcare: sexually transmitted infections—guidelines for prevention and treatment. Department for International Development, London.

Arya OP, Hart CA, eds (1998). *Sexually transmitted infections and AIDS in the tropics*. CABI Publishing, Wallingford, Oxon, UK.

De Schryver A, Meheus A (1990). Epidemiology of sexually transmitted diseases: the global picture. *Bulletin of the World Health Organisation* **68**, 639–54.

Meheus A, Piot P (1986). Provision of services for sexually transmitted diseases in developing countries. In: Oriel, JD and Harris, JRW, eds. *Recent advances in STIs*, 3rd edn. Churchill Livingstone.

World Health Organization (1991). *Management of patients with sexually transmitted diseases.* WHO Technical Report Series 810. WHO, Geneva.

World Health Organization (1995). *An overview of selected curable sexually transmitted diseases.*

World Health Organization/United Nations Programme on HIV/AIDS (1997). *Sexually transmitted diseases: policies and principles for prevention and care*, pp 1–47. WHO/UNAIDS, Geneva.

21.2Sexual behaviour

Anne M. Johnson

Most men and women are sexually active for a large part of their adult lives and sexual fulfilment is an important part of the quality of life. Patterns of sexual behaviour in populations are a key determinant of fertility and transmission of sexually transmitted infections (**STIs**).

Discussion of sexual lifestyle and the ability to take a sexual history is relevant to a wide range of clinical consultations. A few common examples include management of genitourinary symptoms, contraceptive advice, sexual dysfunction, and resumption of sexual activity following childbirth, major illnesses, and surgery.

Sexual orientation

Sexual behaviour studies in representative population samples show that the majority of men and women are predominantly attracted to, and have experience with, members of the opposite sex throughout their lives. However, sexual orientation is not a simple dichotomy between 'homosexual' and 'heterosexual', but varies between individuals from exclusively heterosexual experience through various shades of attraction and experience with both genders, to exclusively homosexual experience.

In a large-scale British study of adults aged between 16 and 44, 8.5 per cent of men and 9.7 per cent of women reported having had any sexual experience with someone of the same gender. For some this may be a fleeting experience in adolescence, with subsequent exclusively heterosexual partnerships. A smaller proportion of the British population report homosexual partnerships involving genital contact (5.4 per cent of men and 4.9 per cent of women). Similar findings are reported from surveys in France and the United States. Most of those with same-gender partners have also experienced heterosexual intercourse at some time in their lives. Exclusively homosexual experience throughout life is thus a relatively unusual phenomenon.

Homosexual experience is more common among men in large metropolitan areas. For example, 10.5 per cent of men sampled in Greater London reported ever having a homosexual partner. Capital cities typically provide a more tolerant atmosphere and better social facilities for those with a homosexual lifestyle. This is reflected in the high proportion of homosexually acquired STIs which are reported from clinics in London, and the higher rates of homosexually acquired HIV and AIDS reported in many metropolitan areas in Europe and the United States.

Age of first heterosexual intercourse

The age of first heterosexual intercourse has been gradually declining over recent decades. The proportion of people having sexual intercourse before marriage has rapidly increased so that sex before marriage has become almost universal in Britain. For men born in between 1930 and 1935, the median age of first intercourse was 20 and for women 21. For men and women born between 1965 and 1975, the median age at first intercourse was 17. Similar trends have been observed in France and the United States.

English law gives the age of consent for heterosexual intercourse as 16 and it is illegal for a man to have sex with a women under 16 in England. The proportion of men and women reporting first intercourse before the age of 16 has risen rapidly over recent decades to 30 per cent of men and 26 per cent of women aged between 16 and 19 in 2000 in Britain. This has important implications for the provision of sex education. Those who are embarking on their sexual careers may be most susceptible to the unwanted consequences of unprotected sexual intercourse. The incidence of sexually transmitted diseases and termination of pregnancy is higher among 16- to 24-year-olds than in older men and women.

Heterosexual partners

There is great variability between individuals in the number of reported heterosexual partners. While many people have few partners, a small proportion have many. Among 16- to 44-year-old men in Britain, 51 per cent reported having none or one partner in the last 5 years; 8 per cent reported more than 10; and a small proportion reported hundreds or even thousands of partners in the course of their lives.

The risk of acquiring or transmitting an STI increases with the number of sexual partners. Those with high numbers of partners may account for a relatively high proportion of STI transmission in a society, and for sustaining endemic STI transmission. They are sometimes referred to as a 'core group for STI transmission'. The choice of partner also influences STI transmission in populations. Age, gender, and ethnic mixing are important, as well as the extent to which people choose partners with similar lifestyles to their own (assortative mixing) or different from their own (disassortative); and whether they have serially monogamous or concurrent partnerships.

Prostitutes and their clients remain at high risk of contracting HIV and STI in some parts of the developing world where condoms are rarely used. In some countries, such as Thailand, public health campaigns have recently led to considerable success in increasing the use of condoms during client and prostitute contacts. In many developed countries, although prostitutes are at increased risk of STIs, they use condoms frequently to protect both themselves and their clients.

The proportion of men who use prostitutes varies widely between countries. In the British survey, 4 per cent of men reported having paid money for sex with a woman in the last 5 years but considerably more frequent exposure is reported in other countries.

Multiple heterosexual partnerships are most common among the young, and among those who are neither married nor cohabiting. Close to one in seven men between the ages of 16 and 24 in Britain reported more than 10 partners in the last 5 years, even though in this group a high proportion are not yet sexually active. Age *per se* is not the only influence on sexual behaviour. Whatever their age, those who are separated, divorced, or widowed are more likely than married people of a similar age to have multiple partners, illustrating the effects of the lifecourse on patterns of partnership formation. Since the emergence of the HIV epidemic, public health campaigns have emphasized the need for a change in sexual behaviour. STI incidence fell in the 1980s but rose again in the late 1990s. In Britain since there has been an increase in number of partners as well as an increase in condom use. In some parts of the developing world, such as Uganda and Thailand, there is evidence of a reduction in the incidence of HIV infection attributable to a recent change in behaviour.

Heterosexual practices

There is variability in the repertoire and frequency of sexual practices between individuals. In heterosexual relationships, vaginal intercourse is the most common practice, but most couples include other practices in their repertoire, particularly mutual masturbation and orogenital contact.

The frequency of sexual contact varies with age, lifestage, and the availability of a sexual partner. For married couples, the median frequency of sexual intercourse is in the order of four times per month, but this is highly variable. The frequency of intercourse appears to decline with age in married and cohabiting couples, although this is partly a function of the increasing length of their relationship.

Among 16- to 44-year-old men and women in Britain, close to 80 per cent reported experience of orogenital contact in the last year, with the majority of couples who experience any orogenital contact practising both cunnilingus and fellatio. There is evidence of an increasing practice of orogenital stimulation in recent decades.

Mutual masturbation is also a common practice and has become more frequent in recent decades.

Anal intercourse is a relatively infrequent activity in heterosexual couples. In the British survey, 26 per cent of men and 24 per cent of women had experienced anal intercourse, but only around 12 per cent had experienced it in the last year. Anal intercourse can result in transmission of STIs. The practice of anal intercourse in addition to vaginal intercourse may increase the risk of heterosexual transmission of HIV. Since anal intercourse is a relatively infrequent practice in all parts of the world, most heterosexual HIV transmission worldwide is attributable to vaginal intercourse.

Homosexual behaviour

Male homosexual lifestyles have been rather more intensively studied than female homosexual lifestyles. Research in the 1970s of volunteer samples of homosexual men in the United States identified a particular lifestyle characterized by multiple casual sexual partners, often encountered at gay meeting places such as bars, clubs, and 'bath-houses'. These men were at high risk of contracting STIs and were among the first to suffer high rates of HIV infection. Research in Britain identified a group of homosexual men with similar lifestyles. However, studies of homosexual men recruited from sites other than STD clinics and gay meeting places show lower rates of sexual partner change and a lower prevalence of sexually acquired pathogens.

Men with multiple homosexual partnerships are at increased risk of HIV infection as well as other STIs, including hepatitis B and syphilis. Women with homosexual partnerships tend to be at low risk of STI and HIV as a result of their different sexual lifestyles and sexual practices, such as non-penetrative sex and orogenital contact.

Many male homosexual partnerships do not involve penetrative anal intercourse, but are restricted to mutual masturbation or orogenital contact. Anal intercourse, however, is the most important mode of transmission of sexually acquired organisms between homosexual men. Many men practise both anal receptive and insertive intercourse. Receptive anal intercourse is the highest risk behaviour for HIV transmission. Since the emergence of the HIV epidemic, there is evidence of a reduction in high-risk behaviour among gay men, characterized by increased condom use and reduced exposure to unprotected anal intercourse. However, there have been recent concerns of a resurgence in high-risk behaviour.

Risk-reduction strategies and sexual health

Increasing attention is being paid to promoting sexual health and reducing the adverse consequences of sexual behaviour. Extensive discussion of population strategies is outside the scope of this chapter. However, people can reduce their risk of STI and unwanted pregnancy by reducing the numbers of partners with whom they have unprotected intercourse, by using condoms, by using effective contraception, and by enjoying sexual practices that may reduce transmission risks. Negotiating sexual fulfilment is a more difficult matter, but greater focus on communication between partners, and sexual technique, is also important. Health professionals have an important role to play, not only by being well informed but by including tactful sexual history-taking among their clinical skills and by being concerned about sexual health and health promotion.

Further reading

ACSF investigators (1992). AIDS and sexual behaviour in France. *Nature* **360**, 407–9.

Cleland J, Ferry B (1995). *Sexual behaviour and AIDS in the developing world*. Taylor and Frances, London.

Johnson AM, *et al.* (1994). *Sexual attitudes and lifestyles*. Blackwell Scientific, Oxford.

Johnson AM, *et al.* (2001). Sexual behaviour in Britain: partnerships, practices, and HIV risk behaviours. *Lancet* **358**, 1835–42.

21.3Vaginal discharge

J. Schwebke and S. L. Hillier

Vaginal discharge is an extremely common reason for women to seek medical care, but its causes, treatment, and potential complications are poorly understood by patients and medical personnel. In the past, it has often been regarded as simply a nuisance. Only recently have accurate diagnosis and treatment of vaginal discharge been recognized as means of preventing future costly morbidity.

The healthy vagina

It is important to understand the normal vaginal ecosystem. At puberty the vagina becomes colonized predominantly with lactobacilli (Fig. 1). These Gram-positive facultative bacilli convert glucose to lactic acid, which helps maintain the normal vaginal pH at less than 4.5. This acidic environment helps to stabilize the ecosystem, as many pathogens, with the exception of *Candida* spp., are inhibited at this pH. Vaginal lactobacilli produce other antibacterial compounds, most notably hydrogen peroxide, which can inhibit the growth of bacterial vaginosis-associated pathogens *in vitro*. Women with peroxide-producing lactobacilli in the vagina are less likely to have gonorrhoea, chlamydial infection, trichomoniasis, or bacterial vaginosis.

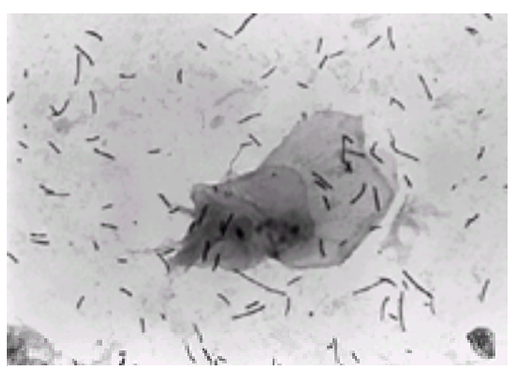

Fig. 1 Gram stain of normal vaginal fluid (copyright Dr Sharon Hillier).

Factors that alter the normal vaginal milieu and predispose the vagina to infection are poorly understood. They probably include hormonal and immunological factors as well as exogenous influences such as antimicrobial therapy, multiple sexual partners, and the use of vaginal douches.

Vaginitis and vaginosis

There are three main types of vaginal infections: trichomoniasis, bacterial vaginosis, and vulvovaginal candidiasis. Trichomoniasis is the only one known to be sexually transmitted, although bacterial vaginosis is most commonly seen in sexually active women and frequently coexists with sexually transmitted infections.

Trichomoniasis (see also Chapter 7.13.13)

Trichomoniasis is one of the few sexually transmitted infections that is more easily diagnosed in the female than in the male. In males, infection is usually asymptomatic and so they may be an important reservoir of infection. Annual worldwide incidence is estimated at 180 million cases. Prevalence varies with the population being studied. It is high among women attending sexually transmitted disease clinics. *Trichomonas* spp. and *Neisseria gonorrhoeae* are frequent coinfections. Symptomatic patients with trichomoniasis most frequently complain of discharge and vaginal pruritus. Intermenstrual or postcoital spotting may occur because the ectocervix is involved. Occasionally the urethra and Skene's glands may be infected, resulting in dysuria. Some 50 per cent of all infected women are probably asymptomatic.

Signs of trichomoniasis are vaginal discharge (42 per cent), odour (50 per cent), oedema or erythema (22–37 per cent). The often copious discharge is described as frothy and yellowish-green, but varies in consistency and colour and is actually frothy in only 8 to 12 per cent of women. Colpitis macularis ('strawberry cervix'), detected by colposcopy, is reported in almost half the patients and is the most specific clinical sign for trichomoniasis, but is rarely seen during routine examination.

Because the clinical signs and symptoms are not diagnostic, the organism should be sought either by direct wet-mount preparation or culture. Trichomonads are motile by means of flagella that can be seen beating even when the organism is at rest; the organism is about the size of a white blood cell (10–20 μm wide). The vaginal fluid contains numerous polymorphonuclear neutrophils because this disease is a true vaginitis, causing a local inflammatory response. The vaginal pH is usually elevated to above 4.5 and can be as high as 6.5 to 7.5. The background bacterial flora is often abnormal as bacterial vaginosis may also be present. The sensitivity of microscopical examination of the vaginal fluid for the diagnosis of trichomoniasis ranges from 40 to 80 per cent. Culture is the current diagnostic 'gold standard' but requires 2 to 5 days before a result can be obtained. As trichomoniasis can occur with other sexually transmitted diseases, screening should be carried out for gonococcal and chlamydial infections.

Treatment is with nitroimidazoles; metronidazole is most frequently used. A single oral dose of 2.0 g is preferred to 250 mg orally three times daily for 7 days because of better compliance and reduced total dosage. The use of nitroimidazoles is often complicated by gastrointestinal side-effects. Some strains of *T. vaginalis* are resistant to metronidazole, but treatment fails most often because the sexual partner has not been treated. For strains of *T. vaginalis* with decreased susceptibility to metronidazole, higher doses are given orally (up to 1 g three times daily), often in combination with intravaginal metronidazole.

New data link trichomoniasis with preterm labour, and no adverse outcomes of pregnancy have been associated with the use of metronidazole, therefore treatment during pregnancy is recommended.

Women in whom *T. vaginalis* is detected, but who are asymptomatic, should be treated as above. If left untreated they may later become symptomatic and, without treatment, they serve as an important reservoir for continued ransmission of the disease. Furthermore, there is data which suggests that trichomoniasis may facilitate the transmission of the human immunodeficiency virus (**HIV**). Because *T. vaginalis* is sexually transmitted it is imperative that the male partners of women with trichomoniasis be treated empirically for *T. vaginalis* and screened for other sexually transmitted diseases.

Bacterial vaginosis

Bacterial vaginosis is the most common diagnosis made in women complaining of abnormal vaginal discharge. The microbiology of bacterial vaginosis is now much better understood, but not its pathogenesis. Recently, complications associated with bacterial vaginosis have been recognized, such as preterm birth, low birth weight,

and infectious complications of pregnancy including postabortive endometritis, intra-amniotic infection, and postpartum endometritis. Among non-pregnant women, bacterial vaginosis has been linked to infection of the vaginal cuff following hysterectomy, and possibly pelvic inflammatory disease.

Despite the fact that bacterial vaginosis is most often diagnosed in sexually active women and is frequently seen together with other sexually transmitted infections, there remains no direct proof that it is exclusively sexually transmitted.

Unlike trichomoniasis and candidiasis, bacterial vaginosis does not appear to be caused by a single organism. Instead there is a change in the entire vaginal flora, resulting in the loss of normal hydrogen peroxide-producing lactobacilli and the appearance of increased numbers of mycoplasmas, *Gardnerella* spp., and anaerobic bacteria. Included among the anaerobes are the black-pigmented *Bacteroides (Prevotella)* spp., *B. (Prevotella) biviens*, *Peptostreptococcus* spp., and *Mobiluncus* spp. There is no inflammation of the vaginal epithelium, unlike in trichomoniasis and candidiasis, hence use of the term 'vaginosis' instead of 'vaginitis'. However, one-third of women with bacterial vaginosis and without other infections have more than 30 white blood cells per high-power field in the vaginal fluid. The predominant cell type is the squamous epithelial cell, many of which are covered by adherent bacteria ('clue cells') (Fig. 2).

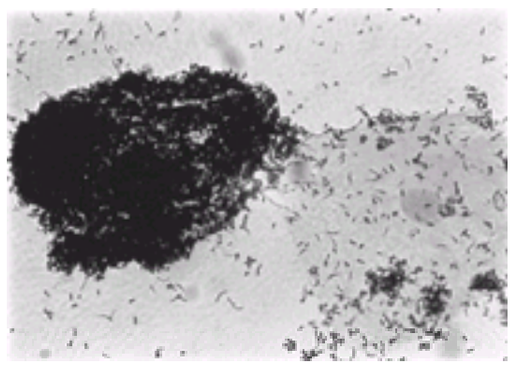

Fig. 2 Gram stain from a patient with bacterial vaginosis (copyright Dr Sharon Hillier).

The symptoms of bacterial vaginosis are vaginal discharge (90 per cent) and an unpleasant odour (90 per cent). The odour is often first noticed after intercourse as the alkaline semen mixes with the vaginal fluid releasing volatile amines. Many women complain of increased odour during menses, at which time vaginal pH increases due to the presence of menstrual blood. As with sexually transmitted infections there is a wide range of symptoms, and many women with bacterial vaginosis are asymptomatic or unaware of their symptoms. Thus, it is of great importance to make a careful evaluation of the following clinical signs:

1. homogeneous-looking, white to grey vaginal discharge;
2. vaginal pH greater than 4.5;
3. positive amine odour when vaginal secretions are mixed with 10 per cent potassium hydroxide ('whiff' test);
4. the presence of 'clue cells'.

It is important to note the type of bacteria present in the vaginal fluid. Careful observation will alert the microscopist to the absence of characteristic lactobacilli and the presence of large numbers of coccobacilli and curved, motile rods (*Mobiluncus* spp.). Culture techniques are not helpful in the diagnosis of this infection, as many of the offending organisms are present in the normal vagina but in low numbers.

Antimicrobial therapy directed at anaerobic organisms is the mainstay of treatment. The most commonly used antibiotic is metronidazole (500 mg orally twice a day for 1 week). Clindamycin is also effective. Intravaginal therapy in the form of clindamycin 2 per cent cream or ovules and metronidazole 0.75 per cent gel is associated with fewer side-effects than oral metronidazole and are of equivalent efficacy.

Although antimicrobial therapy alleviates symptoms in up to 80 per cent of women, recurrences are common. There is no proven regimen for the management of recurrences; however, retreatment with an identical regimen is generally successful in four out of five women. Recurrences may occur because therapy is directed towards eliminating organisms rather than re-establishing the normal vaginal flora. Microbiological studies have shown that the return to normal is slow, often taking several weeks. During this time the vagina is vulnerable to regrowth of the organisms associated with bacterial vaginosis and to clinical relapse. Future directions in treatment may involve some means of reintroducing healthy lactobacilli into the vaginal ecosystem.

Treatment of women with asymptomatic bacterial vaginosis remains controversial. Further studies are needed to define the natural history of bacterial vaginosis in these women. However, because of the association of bacterial vaginosis with infectious complications of gynaecological surgery, treatment is justified in this setting. Although bacterial vaginosis has been associated with complications of pregnancy, routine screening and treatment of pregnant women with asymptomatic bacterial vaginosis is not yet recommended. There is some data to support screening and treatment of women at high risk for preterm delivery, namely those who have had a prior preterm delivery. However, published studies have yielded different results, suggesting that simple treatment of bacterial vaginosis is inadequate to prevent preterm birth. A clearer understanding of the mechanisms by which bacterial vaginosis could cause preterm birth are needed before devising appropriate intervention trials. Because bacterial vaginosis has now been associated with an increased risk for acquisition of HIV, there is interest in promoting widespread screening and treatment in areas with high incidences of HIV. However, the current suboptimal cure rates and high recurrence rates associated with current therapies may make these initiatives untenable.

Candidiasis (see also Chapter 7.12.1)

Vaginal candidiasis or 'yeast' infections are perhaps the best known of vaginal infections to patients and physicians alike. Any type of vaginal discharge is likely to be self-diagnosed as a yeast infection by the patient, yet these represent only 20 to 30 per cent of all vaginal infections. Therefore, many women who have identified themselves as having yeast infections actually have other types of infections, or sometimes no infection at all.

Risk factors for candidiasis include the use of oral contraceptives, containing a high dose of oestrogen, recent use of broad-spectrum antimicrobials, pregnancy, diabetes mellitus, and immunosuppression.

Vaginal candidiasis is not thought to be a sexually transmitted infection but an overgrowth of vaginal yeasts with the development of local symptoms. The most common species causing vaginitis is *Candida albicans*, although other candidal species account for 10 per cent of genital yeast infections.

The symptoms of vaginal candidiasis are a thick, white discharge and pruritis. Frequently, there is extensive inflammation, which may involve the vulva. Examination may reveal a discharge, erythema, and often excoriations. The discharge is classically described as resembling cottage cheese but it can be variable.

Diagnosis is confirmed microscopically by the presence of pseudohyphae and budding yeasts in the vaginal fluid. Addition of 10 per cent potassium hydroxide may help to clarify these appearances by dissolving epithelial cells and bacteria. Typically the pH is below 4.5 and there are many neutrophils present, although candidiasis can occur simultaneously with trichomoniasis or bacterial vaginosis. Culture is not useful because low numbers of yeasts may be present in the vagina without causing disease. Culture may be of use when characteristic signs and symptoms are present, but pseudohyphae are not identified in the wet-mount examination.

Treatment of vaginal candidiasis relies heavily upon intravaginal imidazole preparations such as clotrimazole, terconazole, miconazole, and butoconazole for 3 to 7 days. Fluconazole is an attractive, single-dose, oral alternative to the topical preparations in women with uncomplicated vaginal candidiasis.

Diagnostic approach to the patient with vaginal discharge

As with any problem in medicine, the history is of great importance. This should include a detailed sexual history to help assess the patient's level of risk for sexually transmitted infections. During the examination the pH of the vaginal fluid should be determined and a sample of the fluid placed in small amounts of both saline and 10 per cent potassium hydroxide for microscopy. The presence or absence of an amine odour when the potassium hydroxide preparation is made should be noted ('whiff test'). The presence of blood, semen, or exogenous vaginal preparations (douches, creams) will interfere with the determination of pH and with the 'whiff test'. Microscopical examination of the saline preparation should be done at 400 × to look for pseudohyphae, 'clue cells', motile trichomonads, and polymorphonuclear leucocytes. The predominant type of bacteria should also be noted.

Vaginal Gram stains may also be useful to determine if 'clue cells' and bacterial morphotypes suggestive of bacterial vaginosis are present. Table 1 reviews the bedside diagnosis and treatment of vaginal infections.

Conclusion

The aetiology of vaginal discharge can be easily determined by taking a careful history, by physical examination, and simple laboratory techniques. Timely and appropriate treatment will prevent recurrent illness and costly complications to the patient and, perhaps, to neonates.

Further reading

Eschenbach DA, *et al.* (1988). Diagnosis and clinical manifestations of bacterial vaginosis. *American Journal of Obstetrics and Gynecology* **158,** 819–28.

Holmes KK, *et al.* (1999). *Sexually transmitted diseases*, 3rd edn. McGraw-Hill, New York

Joesoef MR, Schmid GP, Hillier SL (1999). Bacterial vaginosis: review of treatment options and potential clinical indications for therapy. *Clinical Infectious Diseases* **28**, 57–65.

Lossick JG (1990). Treatment of sexually transmitted vaginosis/vaginitis. *Reviews of Infectious Diseases* **12,** S665–81.

Mårdh P (1991). The vaginal ecosystem. *American Journal of Obstetrics and Gynecology* **165**, 1163–8.

Wolner-Hanssen P, *et al.* (1989). Clinical manifestations of vaginal trichomoniasis. *Journal of the American Medical Association* **261**, 571–6.

21.4 Pelvic inflammatory disease

David Eschenbach

Pelvic inflammatory disease (**PID**) comprises a spectrum of female upper genital tract infections that includes any combination of endometritis, salpingitis, tubo-ovarian abscess, and pelvic peritonitis. Salpingitis, or infection of the fallopian tubes, is the most important feature of PID. This is one of the most common and serious infections of the female genital tract because of the long-term effects after PID including infertility, ectopic pregnancy, and pelvic pain.

PID is caused by the canalicular spread of micro-organisms along the mucosal surfaces from the cervix, and to a lesser extent from the vagina, into the upper genital tract. Cervical mucus provides a relative barrier to this spread, but virulent microbes can traverse cervical mucus, which, in any case, is lost during menses. Little is known of local defence mechanisms to prevent this spread of micro-organisms. Certain HLA types appear to be important in chlamydial PID. Factors that appear to influence the ascent of microbes from the cervix into the upper genital tract include surgical procedures such as dilatation and curettage, induced abortion, intrauterine device (**IUD**)-insertion, and hysterosalpingograms. Vaginal douching with medicated products disrupts the vaginal flora and appears to increase the incidence of PID. Furthermore, contraceptive use influences PID rates. Barrier contraception reduces the risk of PID by preventing the acquisition of *Gonorrhoea* and *Chlamydia* spp. Oral contraceptives also appear to decrease the incidence of PID, perhaps by reducing the inflammatory response to chlamydia infection. The IUD appears to increase the risk of PID by allowing the attachment of bacteria.

Most initial episodes of PID are attributable to *Neisseria gonorrhoeae* and *Chlamydia trachomatis*. While one or both bacteria can be isolated from the cervix in up to 75 per cent of patients with PID, there is a wide variation in the prevalence of these bacteria. In populations where gonorrhoea is highly endemic, there is a 50 to 80 per cent prevalence of *N. gonorrhoeae* in women with PID. A study conducted across eight countries of 1900 patients with PID found that the prevalence of *N. gonorrhoeae* varied from 5 to 80 per cent, with a mean of 26 per cent. In the same 1900 patients, *C. trachomatis* was isolated from 5 to 50 per cent of these women (mean prevalence of 29 per cent). In Europe, there was a 30 to 50 per cent prevalence of *C. trachomatis* in women with PID, and a 5 to 15 per cent prevalence of *N. gonorrhoeae*. Between 10 and 20 per cent of patients with PID harbour both bacteria. However, the use of DNA amplification techniques may reveal even higher numbers of infections with these bacteria.

Mycoplasma hominis and *M. genitalium* cause tubal infection in primates, but their role in human PID is unclear. *Ureaplasma urealyticum* has been isolated from the fallopian tubes but appears to play little role in the development of PID.

Facultative and anaerobic bacteria common to the vagina are also isolated from the fallopian tubes of women with PID. In the initial episode of PID, these bacteria appear less frequent than *N. gonorrhoeae* and *C. trachomatis*, but they can become secondary invaders and are important among women with prolonged symptoms. Anaerobic bacteria are virtually always present in pelvic abscesses associated with initial or recurrent episodes of PID. These bacteria are important in recurrent PID, where *N. gonorrhoeae* and *C. trachomatis* are infrequent.

Women with PID have a vast array of clinical symptoms, ranging from virtually no symptoms to ones that are severe. No symptom, clinical sign, or laboratory result is pathognomonic of PID. In women with mild or uncharacteristic manifestations, the diagnosis of PID is usually missed. Perhaps two-thirds of cases of PID go unrecognized. Some patients' symptoms are too mild or are suggestive of common, less serious conditions. The diagnostic threshold of PID must be sufficiently low to include women with mild PID, but as the threshold is lowered, specificity decreases. However, it is better to overdiagnose than to fail to treat mild PID. Most recognized cases of PID present with moderate symptoms and signs such as lower abdominal pain. Symptoms such as abnormal vaginal discharge or bleeding, dysuria, or vomiting do not distinguish women with PID from those with apparently normal fallopian tubes at laparoscopy. Among women with PID diagnosed by laparoscopy, a temperature over 38 °C is present in only 40 per cent of cases, 60 per cent have a leucocytosis, and 75 per cent an elevated erythrocyte sedimentation rate. Women with severe symptoms and signs usually have peritonitis, often from *N. gonorrhoeae* infection, or they have an abscess. These patients can be very ill. Laparoscopy should be considered both for those with florid peritonitis, to exclude other causes such as appendicitis, and for those with abscesses greater than 6 cm in diameter to allow percutaneous drainage.

Perihepatitis occurs in about 10 per cent of women with PID. Often there is moderate to severe pleuritic pain and tenderness, usually in the right upper quadrant. These symptoms are often so severe that lower abdominal pain, suggesting PID, may not be noticed. Perihepatitis must therefore be distinguished from other causes of upper quadrant pain and tenderness by careful pelvic examination. Perihepatitis is associated with *N. gonorrhoeae*, *C. trachomatis*, and other aetiologies.

Cervical samples should be obtained to identify *N. gonorrhoeae* by culture or DNA technology and to identify *C. trachomatis* by DNA technology. Patients with severe manifestations should have peripheral white blood cell counts. Other laboratory tests are usually of little benefit. Ultrasound is helpful for identifying the presence, and particularly the size, of an abscess. Dilated or thickened tubes or fluid within tubes are found in 80 to 90 per cent of women with severe to moderate PID, but in only two-thirds of those with mild PID. Most sonographers have little experience with these findings. Laparoscopy provides an accurate diagnosis and is particularly useful for excluding serious surgical conditions such as ectopic pregnancy, appendicitis, bleeding ruptured ovarian cyst, or a ruptured abscess. Laparoscopy is also useful for difficult cases, such as those unresponsive to antibiotics in which the only objective finding is pelvic tenderness.

Differential diagnosis of PID

Among 814 women who underwent laparoscopy because of a clinical diagnosis of PID, 12 per cent had intra-abdominal conditions other than PID: ectopic pregnancy, appendicitis, ruptured ovarian cysts, and endometriosis. In older women, pyelonephritis, gastroenteritis, and diverticulitis can masquerade as PID. A patient with severe signs of peritonitis should be admitted to hospital for ultrasound examination and/or exploratory laparoscopy. A pregnancy test is needed. If positive, an ectopic pregnancy or other pregnancy complications must be considered. If the pregnancy test is negative and a wet mount of vaginal/cervical secretions reveals no neutrophils or bacterial vaginosis, an ultrasound is needed to diagnose gynaecological diseases other than PID or a gastrointestinal or urinary disorder. If the wet mount shows more neutrophils than vaginal epithelial cells, PID is probable, but other pelvic conditions are not completely excluded. Ultrasound or an endometrial biopsy examined for plasma cells is useful to increase the accuracy of diagnosis.

Treatment is aimed at eradicating *N. gonorrhoeae* and *C. trachomatis*, and, especially for those with moderate to severe disease, anaerobic bacteria (Table 1). Women with PID who are HIV-seropositive appear less likely to be infected with *N. gonorrhoeae* and *C. trachomatis*, but they are more likely to develop abscesses. These patients respond to treatment as promptly as those who are HIV-seronegative, unless they are severely immunosuppressed.

Despite prompt treatment, sequelae are common. Tubal infertility is the most common and disturbing complication. About 10 per cent of women develop tubal infertility after a single episode of PID. Tubal infertility is increased by delaying the treatment of abdominal pain by more than 3 days in chlamydial PID, in women aged over 25 years at the time of PID, and particularly by the number of episodes of PID. Tubal infertility occurs in about 20 per cent of women after two episodes and 40 per cent after three episodes of PID. Tubal infertility occurs in about two-thirds of those with a pelvic abscess or severely damaged tubes observed laparoscopically. There is no correlation between clinical manifestations and the degree of tubal damage observed laparoscopically. Thus, women with mild symptoms but severe tubal damage may become infertile from a single episode of PID. About 7 to 10 per cent of women who become pregnant following PID develop an ectopic pregnancy. Chronic pelvic pain of over 6 months' duration occurs in about 15 per cent of patients following PID.

Attempts should be made to prevent PID. Ideally *N. gonorrhoeae* and *C. trachomatis* infection should be diagnosed and treated before PID can develop. Reduction of *C. trachomatis* infection has lowered the incidence of PID. This should be the aim of primary care providers, especially since *C. trachomatis* can now be diagnosed more readily using new sensitive DNA detection methods.

Further reading

Centers for Disease Control and Prevention (1998). 1998 Guidelines for treatment of sexually transmitted diseases. *Morbidity and Mortality Weekly Report* **47**(No. RR-1), 79–86.

Cohen CR, *et al*. (1998). Effect of human immunodeficiency virus type 1 infection upon acute salpingitis: a laparoscopic study. *Journal of Infectious Disease* **178**, 1352–8.

Eschenbach DA, *et al*. (1975). Polymicrobial etiology of acute pelvic inflammatory disease. *New England Journal Medicine* **293**, 166–71.

Jacobsen L, Westrom L (1969). Objectivized diagnosis of pelvic inflammatory disease. Diagnostic and prognostic value of routine laparoscopy. *American Journal of Obstetrics and Gynecology* **105**, 1088–98.

Scholes D, *et al*. (1996). Prevention of pelvic inflammatory disease by screening for cervical chlamydial infection. *New England Journal of Medicine* **334**, 1362–6.

Spirtos NJ, *et al*. (1982). Sonography in acute pelvic inflammatory disease. *Journal of Reproductive Medicine* **27**, 312–20.

Westrom L (1980). Incidence, prevalence, and trends of acute pelvic inflammatory disease and its consequences in industrialized countries. *American Journal of Obstetrics and Gynecology* **138**, 880–92.

Westrom L, Eschenbach D (1999). Pelvic inflammatory disease. In: Holmes KK, *et al*., eds. *Sexually transmitted diseases*, pp 783–809. McGraw-Hill, New York.

21.5 Infections and other medical problems in homosexual men

A. McMillan

Many homosexual men have adopted safer sexual practices to prevent the acquisition or transmission of the human immunodeficiency virus (**HIV**). However, condom use is often inconsistent and these men may be at risk of contracting sexually transmissible infections (**STIs**), spread by unprotected receptive/insertive genitoanal and orogenital sex, and faecal–oral infections.

An approach to patients with a suspected sexually transmitted infection that has been acquired homosexually

History

Symptomless, sexually transmissible infections in homosexual men are common, particularly when the pharynx or rectum is affected. A history of a sore throat developing within a few days of receptive orogenital contact is common, and often is not associated with a detectable infection. However, pharyngeal gonorrhoea sometimes produces a sore throat. In HIV-infected men, oral discomfort may be a feature of candidiasis or of oral hairy leucoplakia. A rash may be a feature of several STIs, including primary HIV infection. The patient with various viral STIs, including HIV infection, may notice the presence of enlarged lymph nodes. There may be symptoms of viral hepatitis. Constipation, a mucopurulent anal discharge, anal bleeding, perianal discomfort or pruritus ani, and, in severe cases, pain and tenesmus are symptoms of proctitis caused, for example, by *Neisseria gonorrhoeae*. Many men with proctocolitis, such as results from campylobacter infection, have similar symptoms, but diarrhoea with abdominal cramping, bloating, and fever are the principal features in some people. Diarrhoea, epigastric fullness, abdominal cramps, increased flatulence, and nausea may be features of enteritis caused by *Giardia intestinalis*. All patients, but particularly those with a diarrhoeal disease, should be asked about recent travel to tropical or subtropical areas and sexual contacts there. Perianal pain may be a feature of proctitis, but it is also a symptom of localized disease such as traumatic anal fissure and perianal haematoma. A common cause of pruritus ani is threadworm infestation.

Urethral discharge and dysuria are symptoms of urethritis, caused, for example, by *N. gonorrhoeae*.

The proper interpretation of serological tests for syphilis requires information about previous infection. Similarly, a history of vaccination against hepatitis B should be noted.

Since many homosexual men have psychological problems, careful enquiry should be made about, for example, problems with sexual identity.

Physical examination and investigations

The following should be examined:

- *The skin*: for example, a macular rash may indicate early secondary syphilis, acute HIV, or primary Epstein–Barr virus infections. There may be other dermatological features of HIV (see Chapter 7.11.23). Patients with acute hepatitis may be jaundiced.
- *The mouth and pharynx:* tender superficial ulceration may be caused by herpes simplex virus infection but can also occur during acute HIV infection. Other oral manifestations of HIV infection may be seen (Chapter 7.10.21). Painless superficial ulceration may be a manifestation of secondary syphilis.
- *The superficial lymph nodes*: generalized lymphadenopathy (each node being at least 1 cm in diameter) may be associated with some STIs, including HIV infection. Tender enlargement of the inguinal or femoral nodes may be found in, for example, herpes simplex virus infection of the external genitalia or of the perianal region, respectively.
- *The abdomen*: tender hepatic enlargement is a feature of viral hepatitis and splenomegaly may be found in acute viral infections.
- *The external genitalia:* see Chapter 7.10.21.
- *The anal region*: erythema without specific features may be found in patients who have pruritus ani secondary to an anal discharge. Threadworms may be seen. Multiple tender ulcers are most commonly caused by herpes simplex virus (**HSV**) infection. A solitary ulcer at the anal margin may be traumatic in origin, but primary syphilis must always be excluded. Papillomatous lesions of the perianal region are usually caused by human papillomavirus infection, although the condylomata lata of secondary syphilis should always be considered in the differential diagnosis.
- *The rectum*: the distal rectal mucosa should be inspected in those who give a history of receptive anal intercourse or who have anorectal symptoms. In primary perianal and anal HSV infection, however, proctoscopy should be postponed until the lesions have healed. Signs of proctitis are: loss of the normal vascular pattern of the mucosa (although sometimes this may be a normal finding in the distal rectum), mucosal oedema, friability with contact bleeding, and the presence of mucopus in the lumen (Plate 1). Table 1 lists the sexually transmissible causes of proctitis. Ulceration may be noted in HSV infection, and rarely in primary syphilis or lymphogranuloma venereum. As the proctoscope is withdrawn, the anal canal should be examined for ulceration and condylomata (Plate 2).

In patients with anorectal symptoms and in whom microbiological tests yield negative results for the more common pathogens, sigmoidoscopy may define the extent of proctitis and may identify any lesions beyond the reach of the proctoscope. Rectal biopsy is only occasionally helpful in the diagnosis of rectal sexually transmitted diseases (for example, in lymphogranuloma venereum), but may help to exclude other causes of proctocolitis such as Crohn's disease.

Table 2 indicates the routine microbiological investigations that should be undertaken in the management of a symptomless homosexual man who requests a sexual health screen.

Homosexually transmissible infections

The clinical features, diagnosis, and treatment of these conditions are detailed elsewhere and only those aspects that particularly concern homosexual men are discussed here.

Bacterial infections

Gonorrhoea

In homosexual men the urethra is the most frequently infected site, but infection at multiple sites occurs in about 10 per cent of men.

Although pharyngeal gonorrhoea is usually symptomless, occasionally the patient complains of a sore throat, the pain sometimes radiating to the ear. The physical signs are non-specific but include pharyngeal erythema and sometimes tender enlargement of the anterior cervical lymph nodes. Systemic spread of the gonococcus from the site is extremely rare.

At least 40 per cent of homosexual men with rectal gonorrhoea are symptomless and the rectal mucosa appears normal. When present, symptoms and signs are those of a distal proctitis; the histological findings are non-specific. Perianal abscess is an uncommon complication and disseminated infection is rare.

Neisseria meningitidis infection

N. meningitidis is the most common *Neisseria* species to colonize the oropharynx, and is isolated more frequently from this site in homosexual than in heterosexual men. Urethral carriage of the organism occurs in fewer than 1 per cent of homosexuals and rarely causes urethritis. The rectum is colonized in about 2 per cent of homosexually active men, but this organism is a rare cause of proctitis.

Syphilis

The primary lesion may occur on the penis and have the classical features. A chancre may be found at the anal margin or, rarely, in the distal rectum or in the oropharynx. The clinical presentation of extragenital lesions is often atypical. For example, an anal chancre often resembles a traumatic anal fissure—it is often painful, tender, bleeds easily, and often lacks induration; there is usually femoral lymph-node enlargement. The most common symptoms of primary syphilis of the rectum are rectal pain, an alteration of bowel habit, a mucoid anal discharge, and bleeding; the lesion is usually ulcerative but can be polypoidal, resembling a carcinoma. Biopsy can cause profuse haemorrhage.

Chlamydia infection

Although non-gonococcal urethritis is common in homosexual men attending sexually transmitted disease (**STD**) clinics, *C. trachomatis* is isolated much less frequently from homosexual than heterosexual men with non-gonococcal urethritis. The aetiology of non-chlamydial, non-gonococcal urethritis is uncertain. Rectal chlamydial infection is found in about 6 per cent of STD clinic attenders who have had unprotected, receptive anal intercourse. Pharyngeal chlamydial infection, diagnosed either by culture or by the detection of chlamydial DNA, is uncommon. In temperate climates, infection with the lymphogranuloma venereum serovars of *C. trachomatis* is rare.

The clinical features of non-gonococcal urethritis associated with infection by the oculogenital serovars (D–K) of *C. trachomatis* are described in Chapter 7.11.40. Pharyngeal infections are often symptomless but can be associated with pharyngitis lacking specific features. Most men with rectal infection are symptomless with normal proctoscopic findings, but there may be features of a distal proctitis. Histologically, there is a non-specific proctitis consisting of a mild increase in the number of chronic inflammatory cells and polymorphonuclear leucocytes within the lamina propria.

Infection with lymphogranuloma venereum serovars is associated with a more severe proctitis with systemic features. Although the sigmoidoscopic findings are those of a severe proctitis, the inflammatory changes seldom extend more proximally than 12 cm from the dentate line. Occasionally, the inflammation is more localized, with an irregular ulcerated mass that may be polypoidal and extend circumferentially to produce stenosis. Inguinal lymph-node involvement may be a feature of lesions of the anal canal and distal rectum. Untreated, lymphogranuloma venereum can be complicated by perianal abscess formation, strictures, and fistulas in the anus. Histologically, there is a dense infiltration of the lamina propria and submucosa by lymphocytes, plasma cells, histiocytes, and sometimes eosinophils. Occasionally, granulomas with giant cells are found with focal areas of acute inflammation with crypt abscesses.

Chancroid

Chancroid, caused by *Haemophilus ducreyi*, is common in tropical countries. Although there are few reports on the features of perianal chancroid, it is likely that the ulceration is similar to that occurring on the genitalia (Chapter 7.11.13).

Donovanosis (granuloma inguinale)

Klebsiella granulomatis infection is regarded as a sexually transmitted disease. Perianal donovanosis usually occurs in homosexual men, presenting as ulceration. Extensive fibrosis with anal stenosis may occur and, rarely, extensive areas of skin and subcutaneous tissue undergo necrosis. Basal-cell or squamous-cell carcinomas may complicate the infection.

Shigellosis

The sexual transmission of *Shigella* spp. among homosexual men was first recognized in 1974 in San Francisco. Subsequent reports have confirmed the spread of this organism through oroanal sexual contact.

Salmonellosis

Cases of typhoid fever acquired from anilingus with symptomless carriers of *Salmonella typhi* have been described.

Campylobacter infection

In some areas of the United States, *Campylobacter* spp. (particularly *C. jejuni* and, to a lesser extent, *C. fetus*, *C. fennelliae*, and *C. cinaedi*) can be isolated from over 20 per cent of homosexual men with diarrhoea. The source of infection is often uncertain but symptomless carriers are known to exist.

Corynebacterium diphtheriae infection

An increased pharyngeal carriage rate of non-toxigenic *C. diphtheriae* has been reported, but, although these organisms may be associated with pharyngitis, the significance of the finding is uncertain.

Viral infections

Human papillomavirus (HPV)

Condyloma is the most common clinical presentation of HPV infection, a lesion that is almost always associated with HPV type 6/11. In homosexual men, condylomas are found on the genitalia and in the perianal region and within the anal canal, where they may cause pruritus ani and bleeding during defaecation. Rarely, condyloma acuminata may develop in the oropharynx. Although condylomas may be very extensive and persistent in immunocompetent individuals, this is particularly so in immunocompromised patients, including those with HIV infection.

Non-condylomatous HPV lesions of the anal canal may be associated with squamous intraepithelial lesions (SIL). The lesions can be identified through an operating microscope after the application of acetic acid (5 per cent v/v), they are white, well-demarcated from the surrounding mucosa, and have a punctate appearance similar to that seen in HPV infection of the uterine cervix. Definitive diagnosis is by biopsy.

Herpes simplex virus (HSV)

Both HSV-1 and -2 can affect the anogenital region. In Edinburgh, between 1989 and mid-1999, 65 per cent of 65 primary or initial episodes of anogenital disease in homosexual men were associated with HSV-1. These had probably been acquired during orogenital or oroanal sexual contact. Inapparent infection with either type is common.

Primary perianal herpes causes anal pain, constipation, tenesmus, anal discharge, and bleeding on defaecation; systemic symptoms are often prominent. Sacral nerve-root involvement may result in paraesthesia in the affected nerves, urinary hesitancy or acute retention, and impotence. There are often multiple tender ulcers in the perianal region and within the anal canal. Distal proctitis associated with HSV may have no specific features but discrete vesicular or pustular lesions or ulcers may be seen. HSV is an uncommon cause of proctitis. It can occur in the absence of perianal or anal ulceration. Rectal biopsies show a marked infiltration of the lamina propria with neutrophils (sometimes with the formation of crypt abscesses), perivascular infiltration of the submucosal vessels with lymphocytes, and multinucleated cells occasionally with intranuclear inclusions.

HSV-2 is more likely to recur than in HSV-1, but the symptoms and signs are generally much less severe and there are no systemic features.

Cytomegalovirus (CMV)

CMV infection is more prevalent among homosexual than heterosexual men and women. A study in San Francisco during 1981 reported that 94 per cent of sexually active, homosexual men were seropositive for CMV compared with only 54 per cent of heterosexuals. Receptive anal intercourse is the most likely means of acquisition. The virus is present in semen. Restriction enzyme analysis of serial isolates from homosexuals has shown that infection with multiple strains of CMV is common and that multiple strains can be shed simultaneously. CMV is a rare cause of anorectal ulceration.

Hepatitis A virus (HAV)

HAV is transmitted by the faecal–oral route and, although there are conflicting data, homosexual men may be at increased risk of infection. Epidemic outbreaks of homosexually acquired, acute hepatitis A are reported occasionally. Inactivated HAV vaccine may be indicated in sexually active homosexual men to prevent the spread of this infection.

Hepatitis B virus (HBV)

The homosexual transmission of HBV was first recognized more than 20 years ago when it was shown that the prevalence of hepatitis B surface antigen (**HBsAg**) was significantly higher in sera of Caucasian homosexuals than in the sera of Caucasian heterosexual men attending an STD clinic in London. HBV-seropositivity has been related to the duration of regular homosexual activity and to the numbers of different sexual partners. As hepatitis B e antigenaemia is closely associated with infectivity, and as the sera of some 70 per cent of homosexual men who are persistent carriers of HBsAg contain e antigen, their sexual contacts are at particular risk of infection. The means of transmission of HBV between homosexual men is uncertain. In some areas there has been a recent decline in HB prevalence, presumably as a result of the adoption of safer sexual practices to avoid HIV infection and to the more widespread use of hepatitis B vaccine. Vaccination of sexually active homosexual and bisexual men who are antiHBs-negative, is a cost-effective method of preventing spread. However, in HIV-infected homosexual men, the humoral response to vaccination is frequently impaired.

Hepatitis C virus (HCV)

HCV can be transmitted sexually, but the risk of infection to homosexual men who do not inject drugs is low.

Hepatitis D virus (HDV)

In the United States, sera from 7.7 per cent of 298 homosexual men who were HBsAg-positive contained anti-HDV antibodies. There was an association with the number of sexual partners and the occurrence of anorectal trauma in the 2 years before testing. These data, and the finding of HDV markers in the serum and viral RNA in liver tissue from HBV-infected homosexuals who had never injected intravenous drugs, suggests that sexual transmission of this virus occurs within this population.

Hepatitis G virus/GB virus C

The prevalence of serum antibodies against this virus is significantly higher among homosexual than heterosexual men, suggesting that it is sexually transmissible. The pathogenic significance of the virus, however, remains uncertain.

Human immunodeficiency virus (HIV)

The epidemiology, pathogenesis, and manifestations of this viral infection are discussed in Chapter 7.10.21.

Human T-cell leukaemia viruses (HTLV) types 1 and 2

Although HTLV-1 and HTLV-2 can be transmitted sexually, the prevalence of infection among homosexual men is low.

Human herpesvirus 8 (HHV-8)

Kaposi's sarcoma is more common among homosexual men than other groups affected by the HIV. The association with HHV-8 is now well established, but its epidemiology is not fully understood. The prevalence of antibodies against HHV-8 in the sera of homosexual men is significantly higher than that of the general population and correlates with the number of male sexual partners. Orogenital insertive and orogenital receptive sex have been implicated in the transmission of HHV-8 amongst homosexual men.

Protozoal infections

Amoebiasis

Entamoeba histolytica can be transmitted sexually, but the incidence of infection among homosexual men in industrialized countries is low. Previous reports on the apparent high prevalence of infection were erroneous because the organism that was reported as *E. histolytica* was almost invariably *E. dispar*, a non-pathogenic protozoan that is morphologically indistinguishable from *E. histolytica*.

Giardiasis

The prevalence of *Giardia intestinalis* in homosexual men attending STD clinics in temperate climates varies between 2 and 12 per cent. Most infections are symptomless, but a diarrhoeal illness may result.

Cryptosporidiosis

Cryptosporidium parvum can cause diarrhoeal illness. Person-to-person spread has probably been responsible for infection in household contacts. It is a cause of diarrhoea in some homosexual men, but the importance of sexual transmission in the epidemiology of cryptosporidium is uncertain.

Nematode infection

Enterobius vermicularis

Sexual spread by oroanal contact is the most likely route of infection in homosexual men.

Strongyloides stercoralis

Non-infective rhabditaform larvae may develop into infective filariform larvae before leaving the colon. During oroanal contact, these larvae may be transmitted by ingestion of faeces. Penetration of the skin or mucous membrane of the penis by these larvae cause infection during or after anal intercourse.

Other medical conditions in homosexual men

Urinary-tract infection and epididymitis

Bacteriuria may be more prevalent among homosexual men. Acute epididymitis in homosexuals under 35 years of age is more likely to be caused by enterobacteria than *N. gonorrhoeae* or *C. trachomatis*, the most common aetiological agents in heterosexual men under the age of 35 years.

Anorectal trauma

Violent anal intercourse can cause anal fissure or a perianal haematoma. Profuse rectal haemorrhage from mucosal laceration or rupture of the colon at the rectosigmoid junction may result from the insertion of a closed fist into the rectum. Extraperitoneal microperforation of the rectum during 'fisting' may result in pelvic cellulitis: within a few days, lower abdominal and rectal pain develops and the temperature rises. Abdominal examination is usually normal but there may be tenderness in the left iliac fossa. There is marked proctitis and induration of the pararectal tissues. Treatment is with broad-spectrum antimicrobial agents.

Rectal spirochaetosis

In this condition, spirochaetes lie parallel to the microvilli of the epithelial cells of the rectum and superficial portions of the crypts. The condition is indicated by the presence in a haematoxylin and eosin-stained section as a haematoxyphil zone, 3 µm wide, on the luminal surface of the cells. At least one species of spirochaete, *Brachyspira aalborgi*, has been associated with this condition.

Rectal spirochaetosis is found in at least one-third of homosexual men who attend STD clinics but in only 2 to 7 per cent of patients attending general outpatient departments, suggesting sexual transmission. The pathogenicity of the spirochaetes is uncertain.

Carcinoma of the anal canal

In men, receptive anal intercourse may be a risk factor for squamous- and transitional-cell carcinomas of the anal canal. HPV types, particularly HPV-16, have been detected in anal squamous-cell carcinomas. The development of squamous-cell carcinomas in HIV-infected individuals is now well recognized. The immune deficiency associated with HIV infection may permit reactivation of latent HPV resulting in epithelial abnormalities. The situation may be analogous to the development of cancers at other sites in iatrogenically immunosuppressed patients.

Anal squamous intraepithelial lesions (**SIL**) were first described in 1986 and are most commonly found at the junction of the squamous epithelium of the anal canal and the columnar epithelium of the rectum. SILs have been found in tissue removed from the anal canal of homosexual men, and with the operating microscope, after the application of acetic acid, high-grade SILs may be seen as irregular white areas with cobblestoning and mosaicism, and corkscrew vessels.

Low-grade SILs tend to be associated with HPV types 6 and 11 and high-grade SILs with types 16 and 18. Abnormal anal cytology is common in homosexual men with late-stage HIV disease (Centers for Disease Control group IV) and is significantly associated with infection with multiple types of HPV, including those with oncogenic potential. The natural history of SILs, particularly progression to invasive cancer and whether treatment is necessary, is as yet unknown.

Further reading

Doll LS, Ostrow DG (1999). Homosexual behaviour and bisexual behaviour. In: Holmes KK, *et al.*, eds. *Sexually transmitted diseases*, pp 151–62. McGraw-Hill, New York. [Well-referenced review in specialist textbook]

Friedman RC, Downey JI (1994). Homosexuality. *New England Journal of Medicine* **331**, 923–30. [Useful review of psychosocial aspects of homosexuality]

Frisch M *et al.* (1997). Sexually transmitted infection as a cause of anal cancer. *New England Journal of Medicine* **337**, 1350–8.

Katz MH *et al.* (1997). Seroprevalence of and risk factors for hepatitis A infection among young homosexual and bisexual men. *Journal of Infectious Diseases* **175**, 1225–9. [Study that identified risk factors for the acquisition of hepatitis A in young men in the United States]

Martin JN *et al.* (1998). Sexual transmission and the natural history of human herpesvirus 8 infection. *New England Journal of Medicine* **338**, 948–54. [Epidemiological study]

Palefsky JM *et al.* (1998). High incidence of anal high-grade squamous intra-epithelial lesions among HIV-positive and HIV-negative homosexual and bisexual men. *AIDS* **12**, 495–503.

Scallan MF *et al.* (1998). Sexual transmission of GB virus C/hepatitis G. *Journal of Medical Virol*ogy **55**, 203–8. [Epidemiological study]

21.6 Cervical cancer and other cancers caused by sexually transmitted infections

V. Beral

Occurrence

Worldwide, cervical cancer is the second most common cancer in women. It is far more frequent in Third World countries than in the West (Fig. 1). About 1 per cent of women in Britain have invasive cervical cancer diagnosed during their lifetime, and 0.4 per cent die from it. Mortality rates have been falling throughout the twentieth century, except among recent generations of women who became sexually active during the 1960s, a time when exposure to sexually transmitted diseases increased rapidly: 0.2 per cent develop vulval cancer and 0.2 per cent anal cancer.

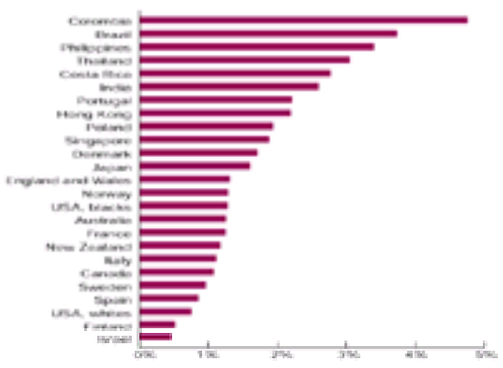

Fig. 1 Percentage of women who develop cervical cancer before the age of 75 years, by country.

The role of human papillomaviruses and other factors

There is now overwhelming evidence that the vast majority of cervical, vulval, anal, and penile cancers are caused by specific types of the human papillomavirus (see Chapter 7.10.2). DNA from human papillomavirus types 16, 18, 31, 33, and 35 (mostly type 16) has been found in as many as 99 per cent of cervical and other anogenital cancers. Fewer than 10 per cent of people without such cancers have detectable evidence of these types of papillomaviruses in their anogenital cells.

The risk of cervical cancer is increased in women who are poor, have little education, were young when they first had sexual intercourse, had many sexual partners and multiple sexually transmitted infections, had many children, especially when they were young, and who smoked cigarettes and used oral contraceptives. Recent evidence suggests that hormonal and reproductive factors and cigarette smoking may independently influence the development of cervical cancer in papillomavirus-infected women, whereas other sexually transmitted infections may be of no direct aetiological significance.

The natural history of infection with the human papillomavirus and associated changes in the cervical epithelium

Some cervical papillomavirus infections cause no obvious epithelial changes—cervical colposcopy, cytology, and biopsy are normal—and the only way infection can be identified is by virological study. Other papillomavirus infections cause 'cervical warts', which are asymptomatic but can be seen as white patches at colposcopy after acetic acid has been applied to the cervix. Cervical smears from women with cervical warts may show various degrees of dysplasia or dyskaryosis, and cervical biopsy may show various grades of cervical intraepithelial neoplasia (**CIN**) (or squamous intraepithelial lesions (**SIL**) according to a new classification known as the 'Bethesda system'). The most extreme change in the epithelium is the development of invasive cervical cancer, which seems to be associated with persistent viral infection and long-standing epithelial abnormalities.

Very little is known about why some lesions progress or regress. Most changes (up to the development of invasive cancer) seem to be reversible. Moreover, lesions of different severity often coexist in the same woman. The more severe lesions tend to be rarer and found in older women: estimates of the proportion of women in Britain likely to develop various types of lesions are given in Table 1, and the age-specific prevalences of some of these lesions are shown in Fig. 2. At least 1 woman in 10 is likely to be infected with human papillomavirus types 16 or 18; 1 in 20 is likely to develop persistent infection or some abnormality of her cervical epithelium; 1 in 50 is diagnosed with *in situ* cervical cancer; and 1 in 100 develops invasive cervical cancer. Papillomavirus infection is most prevalent in women in their twenties, corresponding to the age when they acquire other sexually transmitted infections; the peak prevalence of *in situ* cancer tends to be about 10 years later (in women in their thirties), whereas invasive cancer is rare before 30 years of age.

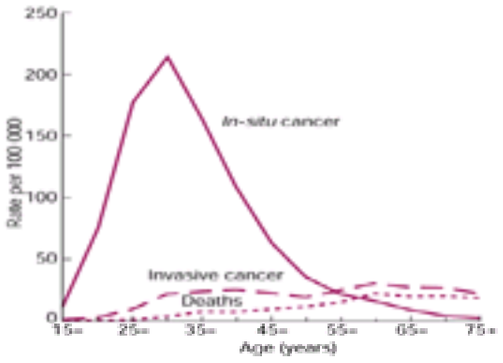

Fig. 2 Age-specific incidence of *in situ* and invasive cervical cancer and of death from cervical cancer in England and Wales.

Clinical implications

Most premalignant cervical changes are caused by papillomaviruses, although most infections resolve spontaneously. Cervical infection with papillomaviruses is especially common in young women, but invasive cervical cancer is rare before the age of 30 years and seems to be associated with persistent infection and long-standing epithelial abnormalities.

We do not know why infection persists in some women and what determines who goes on to develop cervical cancer. Although intensity of infection, age at exposure,

immune response, genital hygiene, reproductive history, and the use of oral contraceptives could be relevant, further research is needed to clarify the role of these possible risk factors. In the meantime, it seems sensible not to resort to active treatment for women in their twenties with papillomavirus infection and related mild cervical lesions, but to reassure them that most lesions will resolve spontaneously and encourage them to have regular cervical smears. Only in young women with severe lesions and older women with persistent epithelial abnormalities is there an appreciable risk of progression to invasive cervical cancer, and active treatment is required.

Prevention

Well-organized screening programmes, based on exfoliative cervical cytology, are known to be effective in reducing the incidence and mortality of cervical cancer. Testing for papillomaviruses should not replace cervical cytology as a first-line approach in screening, since this would result in unnecessary treatment in many young women and because not all cervical cancers are associated with papillomavirus infection. Testing for papillomaviruses may be useful, however, in deciding how to manage the large number of women who have equivocal cervical cytology. Its efficacy must be evaluated before it is adopted on a large scale.

Effective vaccines have already been produced for bovine tumours caused by papillomavirus. Vaccines against the human papillomaviruses have been developed in the last few years and are now being tested in clinical trials.

Further reading

Schiffman MH (1992). Recent progress in defining the epidemiology of human papillomavirus infection and cervical neoplasia. *Journal of the National Cancer Institute* **84**, 394–8.

22.1 Introduction

D. J. Weatherall

The study of blood is one of the most fascinating branches of clinical medicine. Almost all diseases produce changes in the blood at some time during the course of the illness. Furthermore, the primary disorders of the blood and blood-forming tissues can give rise to extremely diverse clinical manifestations, which may involve any of the organ systems. Textbooks often give their readers an unbalanced picture of haematology in the real world. Primary disorders of the blood and blood-forming organs account for only a small percentage of a haematologist's practice. Most patients who are referred with haematological abnormalities have diseases in other systems. Anaemia is a good example. Most anaemias are due to blood loss, infection, renal failure, malignant disease, and malnutrition or parasitic infestation. Anaemia may be the first indication of a chronic urinary tract infection, hypothyroidism, pituitary failure, bacterial endocarditis, polymyalgia rheumatica, or even that endemic disease of 'medical grand rounds', atrial myxoma.

This section emphasizes primary diseases of the blood and blood-forming organs, but the reader should be aware that many of these conditions are relatively uncommon. The summary of the haematological manifestations of non-haematological diseases that appears later in this section should leave the reader with a more balanced view of the scope of this absorbing subject.

An approach to patients with haematological disorders

The diagnosis of blood diseases follows the same process as any other condition; expertise in the laboratory will never make up for an inadequate history and clinical examination. It should be remembered that many patients who are referred to hospital for a specialist opinion on their blood are worried about the possibility of leukaemia, although they will rarely say so. It is important to reassure them as soon as possible if this is not the diagnosis. Where leukaemia is suspected, no time should be lost in determining an accurate diagnosis and a well-worked out plan of management. The situation can then be discussed frankly with the patient and his or her family; knowledge of what they face and precisely what form of treatment is to be instituted often engenders a great sense of relief after weeks or months of fearing the worst.

History

In taking a history from a patient who is suspected of having a haematological disorder, certain factors are of particular importance. The symptoms of anaemia are described in detail later in this section. However, a slowly developing anaemia may be completely asymptomatic, even when the haemoglobin level is extremely low. Individuals who are otherwise healthy should be able to compensate for a relatively mild anaemia; a young individual with a haemoglobin level of 10.5 g/dl who complains of tiredness and an inability to cope with life is more likely to have these symptoms because of chronic anxiety rather than the anaemia. Other general symptoms are of great importance, particularly weight loss, night sweats, bone pain, and pruritis. Moderate nocturnal sweating is common in anxiety states; drenching sweats requiring several changes of nightclothes and sheets are a more ominous symptom, often associated with infection or lymphoproliferative disease. Pruritus occurs in conjunction with many disorders of the blood. When associated with lymphoma it is non-specific, but when it accompanies the myeloproliferative disorders it is often precipitated by warmth such as getting into bed or a hot bath. A detailed drug history is essential; many drugs produce haematological side-effects.

Although a complete systematic history must be taken, gastrointestinal and haemostatic functions are particularly relevant to diseases of the blood. A detailed dietetic history is essential when investigating anaemia, and it is important to ask specifically about symptoms such as a sore tongue, bleeding gums, dysphagia, dyspepsia, disturbance of bowel habit suggestive of malabsorption, and rectal bleeding. Patients are often referred to haematological departments for investigation of easy bruising. Many people, particularly women, bruise easily and the key question is whether the bruising is unusual for them. Is it spontaneous or related to only mild trauma? It is also extremely helpful to enquire into certain key episodes in a patient's life that may provide a clue as to whether there is an inborn bleeding tendency. These include circumcision, dental extraction (was a return to the dentist for stitching or packing ever required?), menstruation, surgical procedures, and so on.

Assessment of menstrual blood loss is an important part of the history in women with iron deficiency, as well as for assessing haemostatic function. It is not enough to ask a woman whether she considers that her periods are normal. If she only uses internal tampons, she probably does not have menorrhagia. However, the use of one or more packets of the more absorbent brands of external pads, or the need to get up at night to change pads or to stay at home during the menstrual period, suggests a heavy loss.

Family histories are particularly important for the diagnosis of blood diseases. It is not only essential to ask for a family history of anaemia or bleeding disorders, the racial origin of the patient's ancestors may also give valuable clues to the cause of anaemia. The long-forgotten, Italian great-grandparent may have been the source of the thalassaemia gene that is responsible for a refractory hypochromic anaemia or the red-cell enzyme deficiency that leads to a haemolytic drug reaction. A detailed personal history is also essential. Cigarette- or cigar-smoking is probably the most common cause of mild polycythaemia. Alcohol can produce remarkably diverse haematological changes. A detailed occupational history may reveal exposure to industrial solvents or other agents responsible for bone marrow depression; unusual hobbies may also result in contact with toxic agents.

Physical examination

The examination of a patient with a haematological disorder follows the same pattern as any physical examination, but there are certain aspects of particular importance. On general inspection it is essential to examine the skin carefully for evidence of bruising, purpura, infiltration, or ulceration. The distribution and pattern of bruising or petechiae may be diagnostic, particularly in disorders such as Henoch–Schönlein purpura, senile purpura, scurvy, purpura due to venous obstruction, and the painful bruising syndrome. Thrombocytopenic purpura is often seen most easily over pressure areas; a few lesions in these regions are easily overlooked. Cutaneous lymphoma may mimic a variety of skin diseases. Chronic leg ulceration is a common finding in sickle-cell anaemia; it occurs occasionally with other genetic haemolytic anaemias. The perianal region and perineum should be carefully inspected. There may be perianal infiltration, particularly in the monocytic leukaemias, and it is very important to recognize perianal infection early in neutropenic patients. Digital examination of the rectum should be avoided in neutropenia for fear of disseminating an infection. Potential sites of infection in compromised patients must be examined daily. They include the skin, intravenous infusion sites, the mouth and throat, and the perineum. The mucous membranes, nailbeds, and palmar creases should be examined carefully for pallor, always remembering that the clinical assessment of anaemia is very inaccurate. Pigmentation of the face is sometimes a feature of folic acid deficiency. Mild jaundice may be a useful indicator of haemolysis, while a greyish pigmentation of the skin is common in patients with iron overload, both primary and secondary to repeated transfusion. There is an association between vitiligo and pernicious anaemia. In patients with polycythaemia there may be suffusion of the conjunctivae, a high colour, and prominence of the vessels over the face, neck, and upper part of the chest. The nails should be examined for unusual fragility; flattened, spoon-shaped nails, koilonychia, which are supposed to be diagnostic of chronic iron deficiency, are now rarely seen.

An assessment of the size of the lymph nodes and an inspection of other lymphatic tissue are a major part of the examination of patients with haematological

disorders. It is most important to develop a systematic approach to lymph-node examination. Each group of nodes in the head and neck, axillae and groins, together with the epitrochlear nodes, must be examined in detail. In the head and neck it is useful to start with the occipital nodes, then move to the preauricular and postauricular nodes, and, finally, to examine systematically the anterior and posterior triangles and supraclavicular regions. In patients with enlarged occipital or posterior cervical nodes, the scalp should be inspected for signs of infestation and secondary infection due to scratching. A simple way of describing enlarged lymph nodes should be used, without the use of too many adjectives. Nodes should be labelled as hard, firm, or soft, and tender or non-tender. Ambiguous terms such as 'rubbery' should be avoided. Soft, tender nodes usually indicate infection. Large, firm nodes are characteristic of lymphoma. Hard nodes occur in secondary carcinoma, although calcified nodes, matted together and attached to skin, are still encountered in patients with tuberculous adenitis. The approximate size of the nodes should be recorded, together with whether they are mobile, attached deep or superficially, and discrete or matted together. It is also very important to examine the tonsils and adenoids, particularly in a patient suspected of having a lymphoproliferative disease.

A detailed examination of the mouth should include the state of the tongue, mucous membranes, gums, teeth, and fauces. Glossitis, as evidenced by a smooth, depapillated tongue, occurs in iron-deficiency and megaloblastic anaemia. Small, black bullae (blood blisters) on the tongue or mucous membranes, which burst and leave superficial ulcers, are characteristic of thrombocytopenic purpura. Gingival hypertrophy is sometimes found in patients with acute leukaemia, particularly the monocytic type, and in some individuals with megaloblastic anaemia due to phenytoin therapy. Ulcers of the mouth and fauces occur in all forms of acute leukaemia. Oral infection, often associated with ulceration, is very common in neutropenic patients. Candidosis may be seen on the fauces, tongue, or mucous membranes. Candidal infection of the throat, associated with dysphagia, should raise the suspicion of oesophageal candidosis (-iasis). The teeth may be badly formed and the bite may be abnormal in patients with severe forms of thalassaemia. Dental abscesses are common in patients with neutropenia; suspect teeth should be gently percussed for evidence of apical infection. Telangiectases may be found on the lips and oral mucous membranes of patients with hereditary telangiectasia.

On abdominal examination the most important questions are the size of the liver, whether there is splenomegaly, and if there are any palpable para-aortic lymph nodes. It is not possible to learn how to examine the spleen from a textbook, but a few hints may be helpful. Large spleens can often be seen to move up and down on respiration if the abdomen is well illuminated and the observer stands at the end of the bed. Very large spleens tend to move downwards and medially towards the right iliac fossa and can be missed if the examiner does not start palpating from this region, moving upwards and medially towards the left subcostal region. A sure way to miss a moderately enlarged spleen is to go digging in with the fingers without eliciting the patient's help. With the left-hand hooked round the region above the left costal margin, and the right hand resting lightly on the abdomen, the patient should be asked to gently breathe in and out through the mouth. The secret of success is to persuade the patient to breathe just deeply enough to move the spleen down without contracting the abdominal muscles. The examiner should wait for the spleen tip to meet their fingers rather than to try to find it by deep palpation. Once defined, the position of the lower border of the spleen should be recorded in centimetres, vertically below the costal margin. Manoeuvres designed to facilitate the palpation of a slightly enlarged spleen, such as turning the patient on their right side, while useful for impressing clinical examiners, are rarely of much help in practice. Be gentle! The author has seen enlarged spleens ruptured by overenthusiastic medical students. If there is pain over the spleen or referred to the left shoulder, don't forget to listen for a rub. Finally, remember that spleens come in all sizes and shapes, and often lie more laterally than expected. Do not be disappointed not to feel the much publicized notch; it happens once or twice in a clinical lifetime! The differential diagnosis of palpable masses in the region of the spleen is considered later in this section.

The eyes are a mine of information in patients with haematological disorders. Periorbital oedema is sometimes seen in infectious mononucleosis. The conjunctivae may show mild icterus not obvious in the skin, and there may be haemorrhages in bleeding disorders. Pingueculae of the conjunctivae are seen in Gaucher's disease. Retinal haemorrhages are common in patients who have had a sudden fall in haemoglobin level. They are less frequent in severely thrombocytopenic patients with normal haemoglobin levels; the combination of anaemia and thrombocytopenia is particularly likely to lead to severe retinal bleeding. Papilloedema occurs commonly in patients with leukaemia involving the central nervous system. Proliferative abnormalities of the retinal vessels are often seen in patients with sickling disorders, particularly haemoglobin SC disease. The hyperviscosity syndrome associated with macroglobulinaemia and some forms of myeloma is characterized by fullness of the retinal veins, which are sometimes broken up into segments like a string of sausages. These changes are often associated with widespread retinal haemorrhages. Optic atrophy may occur in patients with severe vitamin B_{12} deficiency. Unilateral exophthalmos occurs occasionally in patients with myeloma deposits or lymphoma involving the orbit.

Examination of the musculoskeletal system may be particularly rewarding in patients suspected of having genetic blood disorders. In patients with coagulation defects such as haemophilia or Christmas disease, recurrent bleeding into joints may produce a chronic deforming arthritis. Muscle haematomas are also common and are easily missed. For example, bleeding into the psoas sheath may produce a discrete swelling above the inguinal ligament, which may later be associated with nerve compression leading to weakness of the quadriceps and anaesthesia over the anterior aspect of the thigh. If muscle pain is the presenting symptom, it is very important to palpate the muscle groups carefully for the cystic swellings that may occur in haemophiliacs after bleeding into muscles. The joints have other important associations with blood disorders. A mild refractory anaemia is a very common accompaniment of rheumatoid arthritis. Painful arthritis of the large joints may be the presenting symptom of primary haemochromatosis. Gout is a common complication of all the myeloproliferative diseases; the ears should be examined carefully for tophi, in addition to a full assessment of the joints. The value of bone tenderness in the diagnosis of acute leukaemia has been overemphasized. When present it is best elicited by carefully palpating the sternum or tibias, or by rib compression. Be gentle, because sometimes the tenderness is quite exquisite. Bone tenderness or local swelling are also found in patients with myeloma or sickle-cell anaemia. In children with thalassaemia or other hereditary haemolytic anaemias there may be reduced growth, bossing of the skull, and facial deformities. A wide variety of skeletal changes may occur with congenital hypoplastic anaemia.

The use of the laboratory

The diagnosis and management of blood disease requires an examination of the blood and, if appropriate, the bone marrow. Clinicians will obtain the maximum information from their colleagues in the laboratory if they ask the right questions. Scribbling down 'full blood count' on a laboratory request form is useless. It is essential to ask for an examination of the blood film in any patient who is suspected of having a haematological disorder. More can be learned from the help of an experienced morphologist than any other investigation in clinical haematology. Some haematological investigations are underused; others are requested far too often. For example, the often forgotten reticulocyte count is an invaluable guide to the response of the bone marrow to anaemia and for the recognition of bleeding or mild haemolysis. On the other hand, bone marrow examination is an unpleasant investigation and should only be requested with very clear indications. For example, clinicians should stop and think why they are ordering a bone marrow examination in an elderly patient with a peripheral blood lymphocyte count of 80×10^{9}/l. This can only be chronic lymphatic leukaemia; the bone marrow will be infiltrated with lymphocytes. Why put the patient through this traumatic investigation? The result is predictable and will not help in their management.

It cannot be emphasized too strongly that the most useful information is obtained by very close liaison between the laboratory and the ward. Clinicians should visit the haematology laboratory regularly, review films and haematological data with their laboratory colleagues, and be very precise in setting out the reasons for the investigations they order. Much valuable information is lost because of the lack of good liaison between the bedside and laboratory.

Examination of the blood

Constituents of normal blood

Blood consists of several different types of cells suspended in plasma. The classification and morphological analysis of blood cells was made possible by the studies of Ehrlich, who, in 1877, described the use of aniline dyes for staining dried blood films. This approach has been refined over the years. The fine structure of the blood cells has been analysed in greater detail with the electron microscope and, more recently, with the scanning electron microscope (Fig. 1).

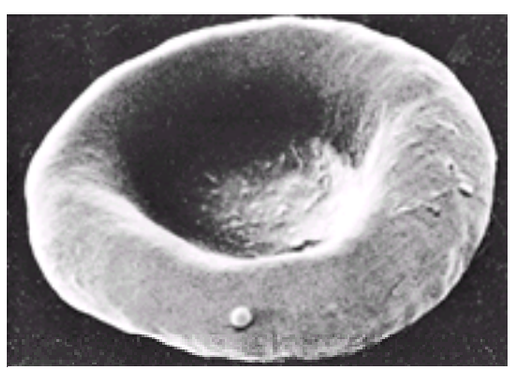

Fig. 1 A human erythrocyte as viewed through the scanning electron microscope. (By kind permission of Dr S. M. Lewis.)

The formed elements of the blood, or blood cells, consist of the red cells, white cells, and platelets. The red cells are biconcave discs approximately 7 to 8 µm in diameter (Fig. 1). They consist of a membrane that contains a concentrated solution of haemoglobin and a variety of other proteins, salts, and vitamins. Normally they are of a uniform shape and size, and contain similar amounts of haemoglobin. On supravital staining, approximately 1 per cent of the red cells show a reticular appearance. These are newly released cells and because of their staining characteristics are called reticulocytes.

The white cells are classified according to their morphological appearances into granulocytes (polymorphonuclear leucocytes (PMNs)), monocytes, and lymphocytes. The granulocytes and monocytes are phagocytic cells, while the lymphocytes are involved in a variety of immune mechanisms. The granulocytes can be further classified according to their maturity. In the newly produced forms, band cells or juvenile polymorphonuclear leucocytes, the nucleus is horseshoe-shaped but single. In a normal blood film the majority of the granulocytes have matured beyond this stage and their nuclei consist of two or more lobes separated by thin, filamentous chromatin strands. These cells are about 12 to 15 µm in diameter. The granulocyte series is further classified according to the staining characteristics of the granules into neutrophils, eosinophils, and basophils. The monocytes are of similar size to the granulocytes but have oval nuclei with a slate-coloured cytoplasm, which may contain some fine granules.

There are two morphologically distinct forms of lymphocyte: a large cell with a diameter of 8 to 16 µm and a smaller one measuring 7 to 9 µm. Both forms are round and have a light blue cytoplasm. In the large lymphocytes the nucleus fills about half of the cell whereas in the small lymphocytes it almost completely fills the cell.

The platelets are disc-shaped cells measuring approximately 2 to 3 µm in diameter. In normal blood they are relatively homogeneous in structure; their fine structure cannot be distinguished by conventional light microscopy.

A more detailed description of the structure and function of these different blood cells and their precursors appears later in this chapter.

Investigation of the blood—the normal blood count

A full blood count can be carried out on a 5-ml anticoagulated blood sample. A stained blood film is prepared for examination of the morphology of the different cells. Using either chemical and physical methods, or the more accurate electronic cell counters, the relative volume of packed red cells and white cells, the haemoglobin level, and the red-cell, white-cell, and platelet counts can be determined. From a series of calculations relating the volume of packed cells, haemoglobin level, and red-cell count, it is possible to derive a series of absolute indices that provide useful information about the size and degree of haemoglobinization of the red cells. Finally, the relative numbers of reticulocytes and the erythrocyte sedimentation rate can be determined.

The stained blood film

An examination of the stained blood film is the most important investigation in haematology. Each of the cell types is studied separately.

The red cells are examined to assess their degree of haemoglobinization and their shape; if both are normal, they are described as normochromic and normocytic. Disorders of the red cell are frequently associated with changes in their morphology or staining properties. These include variation in size or anisocytosis; an increase in size or macrocytosis; a reduction in size or microcytosis; variability in shape or poikilocytosis; pale staining or hypochromia, which suggests underhaemoglobinization; and variation in the degree of staining from cell to cell, which is called anisochromia. In addition to these changes there may be more specific alterations in the morphology of the red cells. Some of these, together with the different clinical disorders with which they are associated, are summarized in Table 1 and illustrated in Fig. 2.

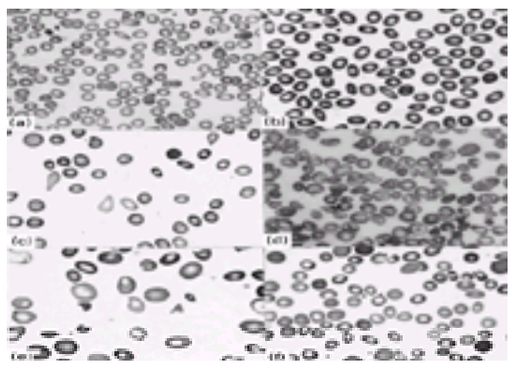

Fig. 2 Morphological changes of the red cells (600–800 ×). (a) Hypochromia and microcytosis. (b) Elliptocytosis. (c) Poikilocytosis (myelosclerosis). (d) Target cells and intracellular crystals (haemoglobin C disease). (e) Macrocytosis and anisocytosis (pernicious anaemia). (f) Dimorphic picture—normochromic and hypochromic (sideroblastic anaemia).

The white cells may be abnormal in number or morphology. An increased white-cell count is called a leucocytosis. If this involves the polymorphonuclear series, it is called a polymorphonuclear leucocytosis or granulocytosis. An elevated eosinophil, basophil, monocyte, or lymphocyte count is called an eosinophilia, basophilia, monocytosis, or lymphocytosis, respectively. A reduced white count is called a neutropenia or lymphopenia, depending on the cell type involved. An absence of granulocytes in the blood is called agranulocytosis. Much can be learned by morphological examination of the white cells. A blood film is said to show a 'shift to the left' if there are relatively more 'young' polymorphonuclear leucocytes present than normal. This is reflected by an increased proportion of band forms and, in more extreme cases, by a variable number of myelocytes or metamyelocytes. In acute bacterial infections, vacuoles may appear in the cytoplasm of polymorphonuclear leucocytes. In addition, the granules may become morphologically abnormal; heavy granulation of this type is called toxic granulation. This change is sometimes associated with the presence of small (1–2 µm) oval bodies called Döhle bodies. A variety of genetic changes of nuclear configuration or of the granules of the polymorphonuclear leucocytes has been described; these are discussed later in this section.

The packed-cell volume, haemoglobin level, and red-cell indices

A great deal can be learned about the character of an anaemia from a few simple haematological tests. The volume of packed red cells (**PCV** or haematocrit) can be estimated either by centrifugation of a blood sample, or by a conductivity method in which it is derived from measurement of the red-cell volume and the number of red cells using an electronic counting system. The haemoglobin concentration is usually determined spectrophotometrically by comparing a test sample with a stable standard, usually of the cyanmethaemoglobin derivative. Red-cell counting has become part of a standard blood count because of the accuracy of electronic cell counters.

Normal values for the PCV, haemoglobin level, and red-cell count are shown in Table 2. It is important to become familiar with the variability of these figures between the sexes and at different stages of development (Table 3). Furthermore, it should be emphasized that the accuracy of these measurements relies very much on the method used for their determination. An electronic cell counter gives extremely reproducible results for all three measurements, whereas a red-cell count made with a counting chamber is of little value. The red-cell indices can be estimated by combining information obtained from these measurements. The mean cell haemoglobin (**MCH**), which is derived from the haemoglobin value and the red-cell count and is expressed in picograms (pg), gives a reliable indication of the amount of haemoglobin per cell. The mean cell haemoglobin concentration (**MCHC**) represents the concentration of haemoglobin in g/dl (g/100 ml) of erythrocytes. The mean cell volume (**MCV**), calculated in femtolitres (fl), gives an indication of the size of the erythrocytes. Hence it is elevated in patients with macrocytic disorders and

reduced in the presence of microcytic red cells. The normal values at different stages of development are summarized in Table 3.

It should be emphasized that the red-cell indices give an indication of the average size and degree of haemoglobinization of the red cells. They are only of value if combined with an examination of a blood film to provide information about the relative uniformity of any changes in size or haemoglobin concentration.

The total and differential leucocyte count

The leucocyte count can be determined either by using a counting chamber or electronically. The differential count is obtained from analysing the different types of white cells in a total of 200 to 300 cells, or more if the total white-cell count is unusually low. It should be remembered that the total white-cell count shows remarkable variability even in the same individual at different times. There are variations during the menstrual cycle and a marked diurnal rhythm with minimum counts in the morning with subjects at rest. Activity may increase the white-cell count slightly, as may emotional stress and eating. Furthermore, the differential white-cell count varies considerably during normal human development. There is a preponderance of lymphocytes during the first few years of life and of polymorphonuclear leucocytes during later development and in adult life. These normal variations are shown in Table 3.

The platelet count

This is most accurately determined with an electronic cell counter, although a rough approximation can be obtained by using a counting chamber. There is marked variation in the normal platelet count and the range in health is approximately 150 to 400×10^{9}/l. A slight drop in the count occurs before menstruation but on the whole it varies less within an individual than the white-cell count.

Blood volume, red-cell mass, and plasma volume

Because the haemoglobin level or PCV may vary due to expansion or contraction of the plasma volume, it is sometimes necessary to measure the red-cell mass and plasma volume directly. This is usually done by radioisotope dilution. The red-cell volume (**RCV**) is measured by labelling the red cells with ^{51}Cr and the plasma volume (**PV**) by the use of isotope-labelled albumin. These measurements are fraught with difficulties because of the variation of vascularity and PCV between different organs, and because fat is a relatively avascular tissue. There is still considerable controversy about how best to express the results. A variety of correction factors has been derived, which attempt to relate the measured RCV or PV to an ideal body weight. In practice it is usual to simply calculate the RCV or PV in ml/kg. The wide range of normal values is summarized in Table 2.

The erythrocyte sedimentation rate (ESR)

The ESR is a measure of the suspension stability of red cells in blood. It is usually expressed in millimetres (mm) and is obtained by measuring the distance from the surface meniscus to the upper limit of the red-cell layer in a column of blood after 60 min. The ESR depends on the difference in specific gravity between the red cells and plasma but is influenced by many other factors, particularly the rate at which the red cells clump or form rouleaux. The increased sedimentation rate of clusters of cells reflects reduced fluid friction resulting from a decreased surface:volume ratio. Rouleaux formation is related to the concentration of fibrinogen and, to a lesser extent, of a_2- and g-globulins in the plasma. Unfortunately, the ESR is also subject to many technical difficulties including the dimensions of the tube, the nature of the anticoagulant used, and any degree of tilt of the tube from the horizontal.

The ESR is still widely used as a non-specific index of organic disease. It is elevated in many acute or chronic infections, neoplastic diseases, collagen diseases, renal insufficiency, and any disorder associated with a significant change in the plasma proteins. Anaemia may cause an increased rate of sedimentation. Although many attempts have been made to develop correction factors to allow for this variable, none is satisfactory. Like all haematological measurements, the ESR changes in certain physiological states, particularly in pregnancy and with increasing age. In men and women over the age of 60 a slightly elevated ESR is often found without an obvious cause (Table 2).

Other haematological investigations

The simple tests that have been outlined in this section form the general screening investigations for all haematological disorders. In later sections we will describe the more specialized investigations that are often required to diagnose specific disorders of the red cells, white cells, and platelets, or of haemostasis and coagulation. Normal values for some of these investigations are given in Table 2.

Examination of the marrow

Bone marrow can be examined by needle aspiration, closed needle biopsy, or open surgical biopsy. In adults the sites most easily available are the sternum and the anterior or posterior iliac crests, although the marrow at the iliac crests tends to become rather fatty in elderly subjects. In under 1-year-old children the anterior surface of the tibia is the site of choice, but in older children the iliac crest or the lumbar vertebral spines are suitable. After aspiration of the marrow, films are made and stained with a Romanowsky stain. Needle or surgical biopsy samples are fixed and sectioned by standard methods.

The marrow films are initially examined under low power to assess the overall cellularity and for the presence of abnormal cells. It is sometimes useful to obtain a differential count and from this to determine the myeloid/erythroid (**M/E**) ratio. This is approximately 3:1 in health, although, if there is increased erythroid activity, it may fall to unity or less. It should be remembered that differential counts may be quite inaccurate because the precursors may not be distributed homogeneously. This is a particular problem in disorders in which there are abnormal cells in the marrow. Having determined the overall cellularity, the morphology of the individual cells is examined. The degree of maturation of the red cells, white cells, and megakaryocyte series is assessed and the marrow examined carefully for the presence of any abnormal cells.

A biopsy specimen is particularly useful for looking at overall cellularity and relating the amount of haemopoiesis to the amount of fatty tissue. It is of particular value if an aspiration yields a 'dry tap' when it may show replacement by fibrous or tumour tissue, which may not aspirate readily. Using appropriate stains it is possible to estimate the amount of iron and reticulin in the marrow.

Assessment of bone marrow activity and distribution

Some indication of marrow function is obtained from its morphological appearances and from the M/E ratio. It is also possible to measure the rates of production and turnover of the red-cell series using radioactive iron. It is sometimes necessary to attempt to estimate the distribution of the haemopoietic marrow, and this is usually done by using isotopes to produce scintograms that show the distribution of erythropoietic or reticuloendothelial marrow throughout the body. Erythropoietic marrow can be visualized using the short-lived, positron-emitting isotope ^{52}Fe with a scintillation camera. In health this shows erythropoietic marrow in the ribs, spine, pelvis, scapula, and clavicle, with a variable amount in the skull. The reticuloendothelial portion of the marrow can be labelled with a radiocolloid with an appropriate particle size; the most effective and commonly used is ^{99}Tc^{m}-sulphur colloid.

Further reading

Beutler E, *et al.*, eds (2001). *Williams hematology*, 6th edn. McGraw-Hill, New York.

Dacie JV, Lewis SM (1994). *Practical haematology*, 8th edn. Churchill Livingstone, Edinburgh.

Hoffman R, *et al*. (2000). *Hematology. Basic principles and practice*, 3rd edn. Churchill Livingstone, New York.

Nathan DG, Orkin SH (1998). *Hematology of infancy and childhood*, 5th edn. WB Saunders, Philadelphia.

22.2.1 Stem cells and haemopoiesis

C. A. Sieff and D. G. Nathan

Introduction

Normal haemopoiesis in the adult depends on the production of blood cells from their recognizable precursors in the bone marrow, their survival in the vasculature, and their demise in the reticuloendothelial system, predominantly in the spleen, liver, lung, and the marrow itself. Though the concentration of cells in the blood varies widely, the values observed in normal individuals are remarkably consistent, particularly considering the vast differences in the lifespans of these cells. For example, the mean lifespan of granulocytes in the peripheral blood may be measured in hours. In contrast, platelets survive for 7 to 10 days. Though platelets are removed from the blood in part by random forces, most of their lifespan is dictated by metabolic changes within them that lead to predetermined death. Normally, red cells are lost by a process of metabolic decay that begins after the erythrocyte has attained an age of approximately 100 days. Lymphocytes have very dramatic differences in lifespan. Some are removed from the circulation in 2 or 3 weeks by a process that is not understood. Others, particularly certain T lymphocytes, appear to survive for the entire lifespan of the individual, carrying within them the programmes embossed upon them by the thymus.

The steady-state concentrations of blood cells vary from one another by three logs or more, but the marrow production rates that maintain them are very similar. Approximately 5×10^4 red cells, 2×10^4 platelets, and 2×10^4 granulocytes are produced per microlitre of blood per day to maintain a normal blood count. Lymphocyte production must be considerably lower because the bulk of lymphocytes in the peripheral blood are long-lived T lymphocytes.

The relatively constant production rates of blood cells are regulated by a highly complex marrow tissue characterized morphologically by recognizable, differentiating precursor cells. These are partially renewed by a variable population of invisible progenitor cells, some of which have the characteristics of stem cells. Precursor cells and their progenitors are packed together into fronds surrounded by endothelial cells that separate marrow cells from the venous sinuses. The completed blood cells find apertures through the endothelial cells and migrate between them to fall into the sinuses, the currents of which carry them into the peripheral blood.

In this chapter, we shall describe critical aspects of the physiology of haemopoiesis in the marrow. To understand this process, we must first review its ontogeny and comparative development.

Phylogeny and ontogeny

In the developing human being, haemopoiesis moves through several overlapping anatomical and functional stages, beginning in the yolk sac, entering the hepatic phase at 6 weeks', and the marrow phase at 20 weeks' gestation. Transfer to the bone marrow phase is generally complete at birth. These anatomical shifts are associated with marked alterations in functional properties, particularly with respect to the pattern of globin synthesis in the red cell. These changes are referred to as the 'fetal switch'. This transition is not a single event involving only the g-chains of fetal haemoglobin, it is instead polygenic involving a series of changes regulated in a programmed fashion. The mechanism of this co-ordinated series of changes is as yet undetermined. It appears to be mediated at the level of the progenitors of haemopoietic cells and is strongly influenced by site-specific regulatory factors.

Marrow anatomy

The relative red (active) marrow space of a child is much greater than that of an adult, presumably because the high requirements for red-cell production during neonatal life demand the resources of the entire production potential of the marrow. During postnatal life the demands for red-cell production ebb. Much of the marrow space is progressively filled with fat (Fig. 1). In certain diseases that are usually associated with anaemia, such as myeloid metaplasia, haemopoiesis may return to its former sites in the liver, spleen, and lymph nodes and may also be found in the adrenals, cartilage, adipose tissue, thoracic paravertebral gutters, and even in the kidneys.

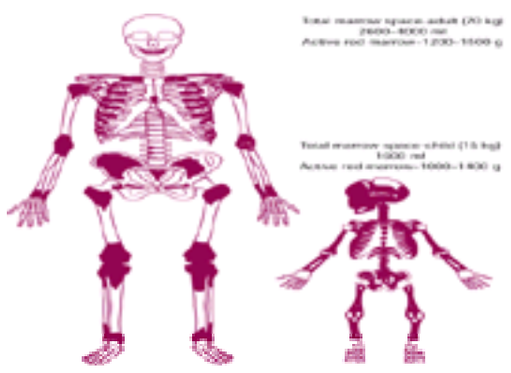

Fig. 1 A comparison of active red marrow-bearing areas in a child and adult. Note the almost identical amount of active red marrow in the child and adult despite a fivefold discrepancy in body weight. (Reproduced from MacFarlane RG and Robb-Smith AHT, eds, 1961. *Functions of the blood*, p 357. Blackwell Scientific, Oxford, with permission.)

The microenvironment of the marrow cavity is a vast network of endothelial cell-lined vascular channels or sinusoids that separate clumps of haemopoietic cells, including fat cells, that reside in the intrasinusoidal spaces. These two compartments are separated by reticular cells (derived from fibroblasts) that form the adventitial surfaces of the sinuses and extend cytoplasmic processes to create a lattice on which blood cells are found. The lattice is demonstrated by reticulin stains of marrow sections (Fig. 2). The conformation of the meshwork of cytoplasmic extensions and the placement of haemopoietic cells in the network of sinuses are best illustrated by scanning electron microscopy. The fibroblast–endothelial cell network provides two major functions: (i) an adhesive framework on to which the developing cells are bound by fibronectin and other integrins, and (ii) the production of haemopoietic growth factors by these cells. Cell–cell adhesion may be mediated by binding of the haemopoietic VLA-4 integrin to stromal fibronectin or VCAM-1.

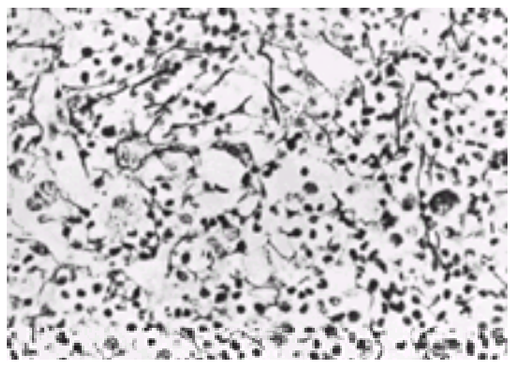

Fig. 2 Bone marrow biopsy of a patient with mild myelofibrosis. A slight increase in the number of reticulin fibres in a delicate discontinuous fibre network is present. Gomori stain, ×350. (Reproduced from Lennert K *et al.*, 1975, *Clinical Haematology* **4**, 335, with permission.)

The central and radial arteries ramify in the cortical capillaries, which in turn join the marrow sinusoids and drain into the central sinus. Cells that egress from the marrow sinusoids then join the venous circulation through concomitant veins. The inner, or luminal, surface of the vascular sinusoids is lined with endothelial cells, the cytoplasmic extensions of which overlap, or interdigitate, with one another. The escape of developing haemopoietic cells into the sinus for transport to the circulation occurs through gaps that develop in this endothelial lining and even through endothelial-cell cytoplasmic pores.

The haemopoietic growth factors comprise a family of small glycoproteins that not only affect immature cells but also influence the survival and function of mature cells. They do so by binding to specific cell-surface receptors. The genes for many of the growth factors and their receptors have been isolated. The cellular origin and the major sites of action of important members of the haemopoietic growth-factor family are shown in Fig. 3. Three of the receptors, c-*kit*, the receptor for Steel factor, Flt-3, the receptor for Flt-3 ligand, and c- *fms*, the monocyte colony-stimulating factor (**M-CSF**) receptor, are members of the transmembrane tyrosine kinase family. In contrast, the receptors for the other haemopoietic growth factors such as interleukin 3(**IL-3**), granulocyte–macrophage colony-stimulating factor (**GM-CSF**), granulocyte-CSF (**G-CSF**), IL-5, IL-6, erythropoietin, and thrombopoietin are members of the haemopoietic growth-factor receptor family. They share several structural features; lacking cytoplasmic tyrosine kinase domains, they activate cells by dimerizing after binding their cognate ligands. This promotes the recruitment and activation of cytoplasmic tyrosine kinases such as members of the Janus kinase (**JAK**) family. The JAK proteins in turn activate members of the signal transducer and activator of transcription (STAT) family and phosphorylate tyrosines of the cytoplasmic domains of the receptor itself. This stimulates recruitment of other signalling or adaptor proteins that activate pathways, such as that involving the RAS protein.

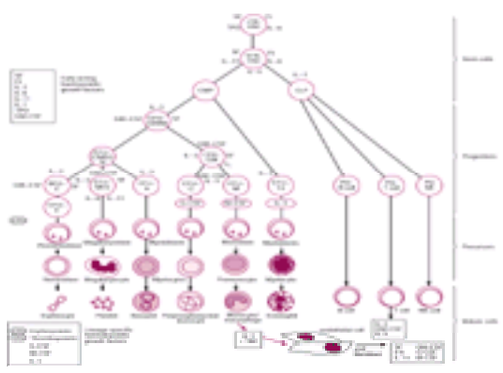

Fig. 3 Maturation of haemopoietic cells and sources and actions of haemopoietic growth factors (HGFs). The predominant actions of several major HGFs are indicated diagrammatically. Steel factor (SF), Flt-3 ligand (FL), IL-11, IL-3, IL-6, and thrombopoietin (TPO) act early during haemopoiesis on haemopoietic stem cells (HSC), while GM-CSF probably act slightly later. The lineage-specific factors erythropoietic (EPO), granulocyte colony-stimulating factor (G-CSF), monocyte-CSF (M-CSF), IL-5, and thrombopoietin act on single lineage-committed progenitors and precursors. Monocytes and macrophages also produce IL-1 and tumour necrosis factor (TNF), potent inducers of HGF production by microenvironmental endothelial and reticular fibroblastoid cells. In addition, macrophages can produce these factors after induction with endotoxin (not shown) and T cells produce IL-3, GM-CSF, and IL-5 in response to IL-1 plus antigenic stimulation.

The location of the different haemopoietic cells is not random. Clumps of megakaryocytes are found adjacent to marrow sinuses. They shed platelets, the fragments of their cytoplasm, directly into the lumen. This reduces the requirement for movement of bulky mature megakaryocytes, a mobility characteristic of the granuloid- and erythroid-differentiated precursors as they approach the point at which they egress from the marrow.

The formed elements of blood in vertebrates, including humans, continuously undergo replacement to maintain a constant number of red cells, white cells, and platelets. The number of cells of each type is maintained in a very narrow range in normal adults—approximately 5000 granulocytes, 5×10^6 red blood cells, and 150 000 to 300 000 platelets per microlitre of whole blood. In the following section we shall review the nature of the signals that affect the proliferation of the stem and progenitor cells, and the normal regulatory mechanisms that maintain balanced production of new blood cells.

Function of stem cells and progenitors

The progenitors of recognizable precursor cells are mononuclear 'blast' cells with large nuclei, prominent nucleoli, and basophilic cytoplasm devoid of granules. These primitive progenitors are present at extremely low frequencies, approximately 1 in 10^4 to 10^5 marrow cells for the stem cell population and 1 in 10^3 for their committed progenitor progeny. A single pluripotent stem cell is capable of giving rise, in a stochastic fashion, to increasingly committed progenitor cells according to the schema outlined in Fig. 3. These committed progenitors are destined to form differentiated recognizable precursors of the specific types of blood cells.

Pluripotent stem cells are defined as cells capable of both self-renewal and multilineage differentiation under the influence of certain non-lineage-specific growth factors such as Flt-3 ligand, Steel factor, IL-6, and thrombopoietin. Their differentiation programme is random and leads to a broad array of more mature lineage-committed progenitors that are themselves responsive to broadly active factors such as IL-3, GM-CSF, IL-11, and subsequently to the more lineage-restricted growth factors, including erythropoietin, thrombopoietin, G-CSF, IL-5, and M-CSF. Some of the lineage-restricted growth factors, particularly erythropoietin, are produced in response to the circulating levels of differentiated blood cells.

Lineage-committed progenitors are characterized by limited proliferative potential that depends upon the presence of specific growth factors that interact with specific receptors on progenitor surfaces. Progenitors are not capable of indefinite self-renewal. In fact, they 'die by differentiation' to mature precursors of the blood cells. The maintenance of their numbers ultimately depends upon the presence of lineage-specific growth factors and on random influx into their pool from the pluripotent stem-cell pool. Therefore, amplification of blood-cell production occurs at the level of the committed progenitor pool, while maintenance of the progenitors depends upon the capacity of members of the pluripotent stem-cell pool to differentiate into the committed progenitor pool.

Haemopoietic differentiation requires an appropriate microenvironment. In normal adults, this is confined to the bone marrow, whereas in the mouse it includes both the spleen and bone marrow. The existence of certain strains of mice that exhibit a deficiency in the haemopoietic microenvironment suggests that the interactions between haemopoietic cells and the bone marrow microenvironment involve very specific molecular mechanisms. Insight into the nature of one of these interactions has come from isolation of the genes that determine the *White Spotting (W)* and *Steel (Sl)* mutations in mice. Animals affected by mutations at both of these loci have a severe macrocytic anaemia associated with defects in skin pigmentation and fertility. The mutations, however, map to different chromosome loci (*W* to chromosome 5, *Sl* to chromosome 10). This is consistent with the results of transplant experiments, which demonstrate that the *W* mutation is one of stem cells whereas the *Sl* defect is one of the bone marrow microenvironment. The *W* gene has now been shown to be allelic with the c-*kit* proto-oncogene, a member of the tyrosine kinase cell-surface receptor family; in contrast, the *Sl* mutation results in defective production of the ligand for this receptor (Steel factor, also known as kit ligand, stem-cell

factor, or mast-cell growth factor). Interestingly, Steel factor is produced in both a secreted and membrane-bound form by fibroblasts and other cells. The latter form may thus provide one molecular explanation for interactions between the stem cells and their microenvironment.

Progenitors can exist outside the marrow. Early haemopoietic cells, including the pluripotent stem cells and certain committed progenitor cells, have been demonstrated in the circulation of normal individuals and experimental animals. The capacity of haemopoietic stem cells to negotiate the circulation is especially significant in relation to stem cell transplantation. While this procedure is still often carried out by infusion of bone marrow from the donor into the circulation of the recipient, mobilized blood stem and progenitors cells are now frequently being used.

The relatively limited production of lymphocyte progenitors has made it difficult to demonstrate that the lymphocyte is derived from the same population of stem cells as the other cellular elements of blood. Recent evidence indicates, however, that both T and B lymphocytes and natural killer cells are derived from a common lymphoid progenitor, while a common myeloid progenitor cell matures to form the committed progenitors of red blood cells, phagocytes, and megakaryocytes (Fig. 3).

The pluripotent stem cell

The concept that sustained haemopoiesis comes from pluripotent stem cells derives from the observation that mice can be protected from the lethal effects of whole-body irradiation by exteriorization and shielding of the spleen. This protective effect was shown to be cell mediated as the injection of spleen cells could initiate recovery and re-establish haemopoiesis in irradiated animals. Till and McCulloch demonstrated that colonies of haemopoietic cells could be observed in the spleen in bone-marrow transplanted, irradiated recipient mice within 10 days after the transplant. These spleen colony-forming units (**CFU-S**) produce colonies that contain precursors of erythrocytes, granulocytes, macrophages, and megakaryocytes. Subsequent experiments using karyotypically marked donor cells confirmed the clonal origin of the differentiated cells. Recent experiments in which foreign genes have been inserted into spleen colony-forming cells have further substantiated this finding. Each colony contains a variable number of stem cells that could again form spleen colonies of differentiated progeny in a second irradiated recipient, indicating the self-renewal property of stem cells. The demonstration of a stem cell that can differentiate to form progenitor cells for erythropoiesis, granulopoiesis, and megakaryopoiesis is completely consistent with subsequent observations in diseases such as chronic myeloid leukaemia and polycythaemia vera in which a clonal origin of abnormal erythroid, granulocytic, and megakaryocytic precursor cells can be demonstrated (see Chapter 22.3.5). In addition, these studies of chronic myeloid leukaemia have demonstrated a pluripotent stem cell that gives rise to B cells as well as to the aforementioned blood cells.

Recent studies provide a model in which the CFU-S is viewed as part of a continuum of cells with a decreasing capacity for self-renewal, increasing likelihood for differentiation, and increasing proliferative activity. The cells progress in a unidirectional fashion in this continuum. CFU-S can be distinguished from a more primitive precursor that has the capacity for long-term haemopoietic reconstitution after bone marrow transplantation (**LTR-HSC**, Fig. 3). In the mouse, as few as 30 cells of a highly purified marrow population that lacks lineage-specific antigens but expresses Ly-6 (Sca-1) and low levels of Thy-1 (that is, lin $^{-}$ Sca-1^{+} Thy-l^{10}) can reconstitute haemopoiesis in 50 per cent of lethally irradiated mice. This fraction appears to comprise virtually 100 per cent CFU-S, but single-cell transfer experiments have shown that it is still heterogeneous. Indeed, cell elutriation studies have shown that most of the CFU-S population is contained within a cell fraction that confers short-term radioprotective capacity (STR-HSC, Fig. 3). It can be separated from a cell fraction with the capacity for long-term haemopoietic reconstitution.

Differences in physical properties and expression of the antigens CD34 and CD33/CD38 have been used to enrich for human stem cells. Most colony-forming cells express all three cell-surface molecules. Cells that give rise to colony-forming cells in long-term bone marrow cultures (that is, long-term culture-initiating cells; LTC-IC) can be separated by their expression of CD34, lack of expression of CD33, CD38, and other lineage-specific markers, and intermediate forward light scattering properties. The importance of CD34^{+} marrow cells is emphasized by *in vivo* simian studies. Like human bone marrow, the CD34 antigen is expressed by a minority of baboon cells. Infusion of these purified cells can reconstitute lymphohaemopoiesis in lethally irradiated baboons. The recent cloning of the murine CD34 cDNA has cast some doubt on expression of CD34 by LTR-HSC. A monoclonal antibody raised to a murine CD34–GST fusion protein was use to separate marrow cells into CD34$^{lo/-}$, CD34^{lo}, and CD34^{+} fractions. Interestingly, long-term multilineage reconstitution was observed after transplantation of the CD34 $^{lo/-}$ cells, whereas the CD34^{+} fraction gave early but unsustained multilineage reconstitution. These data are supported by experiments demonstrating that a tiny subset of murine bone marrow cells that exclude the Hoechst 33342 dye at blue and red wavelengths (called the side population) contains all the LTR-HSC activity, but is CD34 $^{-}$. Recent human studies have also raised the possibility that LTR-HSC do not express CD34. When primitive human lin $^{-}$ cells are separated into CD34^{+} and CD34^{-} fractions, the capacity to reconstitute haemopoiesis in immunodeficient mice (called SCID repopulating cells or SRC) is found in both cell fractions. A resolution to this controversy may come from the recent demonstration that resting murine haemopoietic stem cells are CD34^{-}, while activated haemopoietic stem cells express the CD34 antigen.

Studies with purified populations of stem cells have shown that combinations of specific haemopoietic growth factors such as Steel factor, Flt-3 ligand, IL-6, and surprisingly, thrombopoietin can act at the stem cell level to induce cell cycling and proliferation. IL-3, produced by T cells and natural killer cells, and GM-CSF, a product of both stromal cells and T cells, appear to be factors essential for the survival *in vitro* of a class of stem cells that forms blast colonies in methylcellulose culture. These 'blast' colonies contain multilineage and unilineage progenitors. They are probably at the myeloid stem-cell stage of differentiation (Fig. 3). When isolated from bone marrow, these stem cells are mostly in a non-cycling, quiescent state. The addition of IL-3, GM-CSF, or Steel factor or other stromally produced haemopoietic growth factors such as IL-6, IL-11, or G-CSF shortens the G_0 phase in these cultures, thus hastening the onset of blast colony formation.

The factors that control the fate of stem cells to undergo either self-renewal or commitment to differentiate down a lineage pathway are poorly understood. However, nuclear transcription factors have been shown to play a role in haemopoietic cell proliferation and lineage commitment. The tal-1/SCL, Rbtn2/LMO2, and GATA family of transcription factors are important in this regard. In particular, tal-1/SCL, a basic helix–loop–helix (bHLH) transcription factor, is expressed in biphenotypic (lymphoid/myeloid) and T-cell leukaemias, and in both early haemopoietic progenitors and more mature erythroid, mast, megakaryocyte, and endothelial cells. Targeted disruption of the *tal-1/SCL* gene in murine embryonic stem cells leads to death *in utero* from absence of blood formation; a lack of *in vitro* myeloid colony formation suggests a role for this factor very early during haemopoiesis.

Another transcription factor implicated in T-cell ALL is the LIM domain nuclear protein rhombotin 2 (rbtn2/LMO2). Mice that lack this factor die *in utero* and have the same bloodless phenotype as the tal-1/SCL$^{-/-}$ animals. GATA-2 is expressed in the regions of the *Xenopus* and zebrafish embryos that are fated to become haemopoietic, and is highly expressed in progenitor cells. Overexpression of GATA-2 in chicken erythroid progenitors leads to proliferation at the expense of differentiation. Targeted disruption of the *GATA-2* gene by homologous recombination in embryonic stem cells leads to reduced primitive haemopoiesis in the yolk sac and embryonic death by day 10 to 11. Definitive haemopoiesis in liver and bone marrow is profoundly reduced with loss of virtually all lineages. *In vitro* differentiation data show a marked deficiency of Steel factor-responsive definitive erythroid and mast cell colonies and reduced macrophage colonies, suggesting that GATA-2 serves as a regulator of genes that control haemopoietic growth factor responsiveness or proliferation of stem and/or early progenitor cells. These data contrast with the later time of embryonic death from anaemia (day 15) in mice with targeted disruption of the c-*myb* or retinoblastoma (*Rb*) genes, or with severe forms of *W* and *Sl* mutations. Similarly, loss of function of the AML-1 gene, which encodes one of the a subunits of the heterodimeric core-binding factor (**CBF**), results in fetal death by day 12.5 due to failure of production of all definitive haemopoietic lineages. CBF recognition sequences are present in the IL-3, GM-CSF, M-CSFR, and T-cell antigen receptor promoters. The *AML-1* gene is frequently rearranged in acute myeloid leukaemia (AML) and childhood acute lymphoblastic leukaemia (ALL), and is expressed in myeloid and lymphoid cells.

The survival of a particular stem cell in the marrow requires a 'niche'; thus isogeneic marrow infusions are not successful unless the recipient is irradiated or treated with sufficient doses of cytotoxic drugs to create an adequate number of niches. Therefore, reports of failure of engraftment in aplastic anaemia using identical twin donors do not necessarily implicate an immunological basis for the disease. Equally likely is persistence of non-functional pluripotent progenitors in the aplastic marrow niches. These abnormal cells must be destroyed in order to allow implantation of transfused normal progenitors.

The stem cell model of haemopoiesis has parallels in other organ systems. That rapidly self-renewing epithelial tissues like skin and intestine have stem cells that continually replenish the cells lost by differentiation is well described. It is likely that most epithelial tissues, for example liver and pancreas, also contain stem cells that are brought to bear after organ damage. The demonstration of the existence of neural stem cells in the adult brain has raised the possibility that many organ systems might retain a population of self-renewing stem cells. Muscle satellite cells also appear to fulfil this role.

Much more surprising are recent demonstrations of stem cell plasticity. Transplantation of genetically marked bone marrow showed that the donor cells can migrate into areas of damaged muscle, differentiate into myogenic cells, and participate in tissue regeneration. Similarly, transplanted haemopoietic cells can differentiate into glial cells in adult mouse brain, into endothelial cells, and into hepatocytes in damaged liver. Mesenchymal stem cells, derived from an adherent bone marrow cell population, express neither CD34 nor CD45, markers of primitive haemopoietic cells. Mesenchymal stem cells are capable of marked expansion in culture, and can then be induced to differentiate into osteoblasts and osteocytes, chondrocytes, adipocytes, and myotubules.

That stem cell plasticity may be therapeutically useful is suggested by bone marrow transplant studies in three patients with ostegenesis imperfecta, in which low-level osteoblast engraftment was demonstrated 3 months post-transplantation, with histological changes indicating new dense bone formation. These studies do not address the question of the identity of the engrafting cell, since whole unfractionated marrow was used. Further studies with purified cell populations and longer follow-up will be required. The relationship of the mesenchymal stem cells to the haemopoietic stem cells is not clearly defined, especially in view of the recent demonstration that the highly purified CD34– side population of murine marrow can reconstitute not only haemopoietic activity, but also contribute nuclei to muscle fibres, partially restoring expression of dystrophin in the *mdx* mouse, a model of Duchenne muscular dystrophy. Finally, it is now apparent that the plasticity of primitive cells is not confined to the haemopoietic system, with the demonstration that blood cells can be derived from both neural and myogenic satellite cells.

Erythropoiesis

The rate of erythropoiesis is driven by anaemia or hypoxia. Both stimulate a class of peritubular kidney cells, through a haem-containing oxygen sensor, to transcribe the erythropoietin gene and release the hormone into the blood. The hormone binds to the erythropoietin receptor in erythroblasts and erythroid progenitors to stimulate their division and differentiation. The least mature committed erythroid progenitor is known as an erythroid burst-forming unit (**BFU-E**), because when it differentiates *in vitro* it forms large colonies of erythroblasts and reticulocytes that may contain as many as 50 000 cells. The colonies, derived from single cells, have a burst-like appearance because they may be composed of multiple subcolonies. Thus one BFU-E may first divide in culture to form subcolony-forming cells, which then differentiate into colonies of erythroblasts and reticulocytes. BFU-E progressively mature during their sojourn in the marrow. In doing so they lose their capacity to divide and migrate *in vitro,* but gain in sensitivity to erythropoietin until they reach the stage at which they are known as erythroid colony-forming units (**CFU-E**).

The regulated proliferation and maturation of erythroid progenitors depends on interaction with a number of growth factors. Erythropoietin is essential for the terminal maturation of erythroid cells. Its major effect appears to be at the level of the CFU-E during adult erythropoiesis. Recombinant preparations are as effective as the natural hormone. These very mature progenitors and proerythroblasts do not require 'burst-promoting activity' in the form of IL-3, GM-CSF, or Steel factor. Their dependence on erythropoietin is emphasized by the observation that they will not survive *in vitro* in its absence.

Steel factor has also been shown to have marked synergistic effects on BFU-E cultured in the presence of erythropoietin. Alone, it has no colony-forming ability. The majority of CFU-E are in cycle; their survival in the presence of erythropoietin is probably tightly linked to their proliferation and differentiation to proerythroblasts and mature erythrocytes. Erythropoietin also acts on a subset of presumptive mature BFU-E that require it for survival and terminal differentiation. A second subset of BFU-E, presumably less mature, survive deprivation of erythropoietin if 'burst-promoting activity' is present, either as Steel factor, IL-3, or GM-CSF. Under serum-deprived culture conditions, the combination of erythropoietin and IL-3 or GM-CSF results in more BFU-E-derived colonies than when erythropoietin is added alone.

Factors distinct from the classic colony-stimulating factors may positively regulate erythropoiesis, either directly or indirectly. Limiting dilution studies of highly purified CFU-E in serum-free culture show that insulin and insulin-like growth factor I act directly on these cells. The presence of erythropoietin is also essential. CFU-E and mature BFU-E are highly responsive to the mitogenic effect of erythropoietin as well as to its differentiating role. Therefore, in haemorrhagic or haemolytic anaemias with elevated levels of erythropoietin, the numbers of CFU-E and mature BFU-E may rise remarkably in the marrow. Immature BFU-E are less responsive to the mitogenic effect of erythropoietin, and therefore, the frequency of this subset of BFU-E changes little in anaemia.

Negative regulation of erythropoiesis

Subsets of lymphocytes with an immunological suppressor phenotype isolated from normal subjects can inhibit erythroid activity *in vitro.* Similarly, some patients with anaemia or granulocytopenia have an associated expansion of certain T-lymphocyte populations. In the rare disorder 'T lymphocytosis with cytopenia', *in vitro* suppression of erythropoiesis (or granulo-cytopoiesis) has been correlated with the expansion of a T-lymphocyte population that may be the counterpart of the haemopoietic suppressor cells isolated from normal peripheral blood. The phenotype of these cells has been described in detail. The cell is a large, granular lymphocyte that is both CD2 and CD8 (classic suppressor phenotype) positive. Suppressor T cells may also be involved in some cases of aplastic anaemia or neutropenia without any underlying immunological disorder or an overt T-cell proliferation.

Exactly how suppressor T cells interact with haemopoietic progenitors, and what surface antigens are 'seen' by the suppressors is not known. There is evidence to support the concept that suppression of erythroid colony expression *in vitro* can be regulated by T cells and may be genetically restricted. Certain phenotypes of T cells 'recognize' distinct classes of histocompatibility antigens on immunological cell surfaces. Thus, the observation that haemopoietic progenitors have a unique distribution of class II histocompatibility antigens on their cell surface suggests a role for these antigens in the cell–cell interactions that regulate haemopoietic differentiation.

T cells may also inhibit erythropoiesis in a non-HLA restricted fashion by the production of inhibitory cytokines. Some lymphokines may inhibit erythropoiesis *in vitro* by a complex lymphokine cascade. Activation of T cells by the T-cell antigen receptor CD3 results in cell-surface expression of the IL-2 a-chain (p55) and the acquisition of IL-2 responsiveness. IL-2 inhibits BFU-E in the presence of these IL-2R positive cells, possibly by inducing their release of interferon-g. CD2 can serve as an alternative pathway of T-cell activation, and may do so through binding to its ligand LFA-3 on antigen-presenting cells. Blockade of CD2 with monoclonal antibody leads to abrogation of IL-2/ interferon-g-mediated BFU-E suppression. These data are difficult to reconcile with the observation that IL-2 incubation of activated CD4+ T cells leads to marked expansion of IL-3 and GM-CSF mRNA-positive cells by *in situ* hybridization. Most, but not all, CD4+ T cells express CD28 as well, and there is evidence to suggest that IL-3 production is restricted to CD28+ T cells. It thus appears paradoxical that potent stimulating and inhibitory lymphokines can be produced by activation of T cells through the same pathway.

Tumour necrosis factor also suppresses erythropoiesis *in vitro*. The injection of peritoneal macrophages into animals infected with Friend murine leukaemia virus results in rapid but transient resolution of the massive erythroid hyperplasia associated with this disease. This may be due to elaboration by macrophages of IL-1a, which does not suppress erythropoiesis itself, but acts by the induction of tumour necrosis factor. This effect is reversed by erythropoietin.

Proerythroblasts represent the ultimate stage of differentiation of committed erythroid progenitors. In contrast to the progenitors, which comprise less than 0.1 per cent of the marrow cell population, proerythroblasts are present at 3 to 5 per cent, and their daughters, the recognizable erythroid precursors, comprise 30 per cent of the population.

Estimates of reticulocyte production and erythroblast content of marrows, together with measurements of the rate at which the proerythroblast compartment is renewed from the progenitor pool, suggest that approximately 10 per cent of the daily reticulocyte production is derived from the terminal differentiation of proerythroblasts newly developed from the progenitor department. During anaemic stress the rate at which progenitors differentiate to proerythroblasts may increase 10-fold or more. This increase in the rate of proerythroblast formation from progenitors is associated with an increase in the production of fetal haemoglobin in a large fraction of the erythroid cells derived from them. The basis of this reactivation of fetal haemoglobin synthesis is not understood. The extent to which fetal haemoglobin may be increased in such settings could be genetically controlled. It is an important phenomenon because those with the capacity to develop large increases in fetal haemoglobin who are also homozygous for major b-chain haemoglobinopathies may have a remarkably mild course. Fetal haemoglobin elevation occurs in many forms of accelerated erythropoiesis and is a marker of such a condition.

Phagocytopoiesis

The development of a clonal assay for granulocyte and macrophage progenitors preceded the development of erythroid progenitor assays by nearly a decade, yet a clear understanding of the regulation of myeloid differentiation remains elusive. Figure 3 describes the development and regulation of granulocyte, monocyte, and macrophage production from the pluripotent stem cell. The colony-forming unit-granulocyte-macrophage (**CFU-GM**) is derived from the pluripotent progenitor. It gives rise to separate granulocyte and monocyte progenitors (CFU-G and CFU-M), which, under the influence of unique colony-stimulating factors, differentiate to mature granulocytes and/or monocytes, respectively. Both IL-3 and GM-CSF affect a similar broad spectrum of human myeloid progenitor cells. This includes colonies that contain granulocytes, erythrocytes, monocytes, and megakaryocytes (**CFU-GEMM**), eosinophils (**CFU-Eo**), CFU-GM, CFU-G, and CFU-M. Data from serum-free cultures suggest that in the presence of IL-3 or GM-CSF alone, myeloid colony formation is much reduced. Optimal CFU-G or CFU-M proliferation requires the addition of G-CSF or M-CSF, respectively, to the cultures. Even in serum-replete conditions, IL-3 acts additively or synergistically with G-CSF to induce more granulocyte colony formation than is observed with either factor alone.

Serum-free studies may have important implications for the use of combinations of colony-stimulating factors *in vivo.* The use of such culture conditions should

provide further insight into the *in vitro* activities of the different factors.

Colony-stimulating factors also induce a variety of functional changes in mature cells. GM-CSF inhibits polymorphonuclear neutrophil migration, induces antibody-dependent cellular cytotoxicity (**ADCC**) for human target cells, and increases neutrophil phagocytic activity. Some of these changes may be related to GM-CSF-induced increase in the cell-surface expression of a family of antigens that function as cell adhesion molecules. The increase in antigen expression is rapid and is associated with increased aggregation of neutrophils. Granulocyte–granulocyte adhesion can be inhibited by an antigen-specific monoclonal antibody. GM-CSF also acts as a potent stimulus of eosinophil ADCC, superoxide production, and phagocytosis. G-CSF acts as a potent stimulus of neutrophil superoxide production, ADCC, and phagocytosis, while M-CSF activates mature macrophages and enhances macrophage cytotoxicity.

Monocytes leave the circulation and differentiate further to become fixed tissue macrophages. These tissue macrophages include alveolar macrophages and hepatic Kupffer cells, dermal Langerhans cells, osteoclasts, peritoneal macrophages, pleural macrophages, and possibly brain microglial cells, though the origin of these is still uncertain. The wide variety of cells with diverse functions that must be supplied from the granulocyte–macrophage progenitor requires that this system be highly regulated at many levels of differentiation.

The granulocyte compartment itself is more complex than either the erythroid or megakaryocyte compartments. The circulating half-life of the newly rapidly deployed granulocyte is only 6.5 h. In order to meet sudden demands, an additional non-circulating granulocyte pool exists in the spleen, marginated around blood vessels, and in a readily releasable bone-marrow pool. The rate at which new myeloblasts or monoblasts are produced by progenitors *in vivo* is not known, but exhaustion of progenitors in infection, particularly in the neonatal period, is associated with a fatal outcome due to a failure of granulocyte production.

Suppression of phagocyte production

An elaborate system exists for suppression of granulocyte and macrophage production. It involves T lymphocytes and their products, particularly interferon-g, monocytes, and perhaps acidic isoferritins. In some circumstances, clones of T cells that suppress granulocyte production *in vitro* and *in vivo* have caused profound granulocytopenia. Clearly, a twin regulatory system exists that contributes to the fine control of phagocyte production by close control between progenitors and adventitial cells that secrete inducer and suppressor molecules. It is well established that T lymphocytes capable of the suppression of phagocyte colony formation may be present in human marrow and induce neutropenia.

Megakaryocytopoiesis

The cloning of thrombopoietin has greatly clarified our understanding of the regulation of megakaryocytopoiesis. Prior to the discovery of thrombopoietin, several factors including IL-3, IL-6, IL-11, Steel factor, and even erythropoietin were shown to stimulate megakaryocytopoiesis and thrombopoiesis *in vitro* and *in vivo*. IL-11 has even entered clinical trials. Hence, all of the above mentioned haemopoietic growth factors, except erythropoietin, can contribute collectively to 'megakaryocyte colony-stimulating activity' (**Meg-CSA**). Meg-CSA is therefore a 'soup' of growth factors that transduce three of the four classes of receptors that drive haemopoietic differentiation; these comprise the b common, tyrosine kinase, and gp130 families. All of these receptors, when engaged, drive early progenitor proliferation and partial differentiation to more mature progenitors. The final steps of lineage-committed mature progenitor development into recognizable marrow precursors require a lineage-specific growth factor—G-CSF for the granulocyte, M-CSF for the macrophage, IL-5 for the eosinophil, and erythropoietin for the erythrocyte.

The discovery of thrombopoietin provides the final step of understanding of megakaryocytopoiesis because this factor, and probably none other, actually induces lineage-restricted megakaryocyte progenitor proliferation, differentiation of those committed progenitors to megakaryoblasts, and finally, differentiation of megakaryoblasts to the megakaryocytes that in turn produce platelets. However, this in no way implies that other Meg-CSA components may not be useful in the therapy of hypoplastic thrombocytopenias. Circulating thrombopoietin levels are high in those conditions, just as erythropoietin levels are elevated in the erythroid hypoplasias. Administration of high doses of erythropoietin is usually of little benefit in the latter conditions. Thrombopoietin may be just as unsuccessful in certain megakaryocyte hypoplasias because those conditions are often associated with severe depletion of lineage-specific or multipotent progenitors. One or more of the growth factors that comprise Meg-CSA, such as IL-11, may be more useful in such circumstances. Clinical trials now in progress will decide this issue.

Thrombopoietin

Identification of the proto-oncogene c-*mpl* revealed an orphan haemopoietic growth factor receptor that proved to be crucial for megakaryocytopoiesis. In 1993, Methia and coworkers performed a critically important experiment, when they demonstrated that exposure of CD34+ progenitor cells in culture to oligonucleotides that were antisense to c-mpl inhibited the ability of these cells to form megakaryocyte, but not other haemopoietic colonies. In 1994 several laboratories cloned the all-important ligand for this receptor, the growth factor thrombopoietin, and important physiological studies of thrombopoietin were launched.

The thrombopoietin gene is localized on the long arm of chromosome 3. It contains five exons, the boundaries of which line up precisely with those of the erythropoietin gene. The gene is widely expressed in liver, kidney, smooth muscle, endothelial cells, and fibroblasts. Thus thrombopoietin is produced at the site of stoma supporting haemopoiesis. Though its activity is increased in the blood during episodes of thrombocytopenia, it does not necessarily function as a hormone because it is produced directly at the site of thrombopoiesis. In this sense, it differs from erythropoietin, which is not produced at all in marrow stroma. It is likely that the level of production of thrombopoietin is quite constant in all tissues. The blood levels may increase in thrombocytopenic states merely because circulating platelets and tissue megakaryocytes sop up the growth factor and carry it out of the circulation. This theory has received support from observations in mice with disruption of the murine transcription factor gene called *NF-E2*; although these animals are thrombocytopenic they have an increase in megakaryocyte mass and no increase in serum thrombopoietin levels.

The thrombopoietin molecule is considerably longer than the other haemopoietic growth factor polypeptides. Its 5' half bears 23 per cent sequence homology to erythropoietin, while the 3' half bears no structural homology to any cytokine and may be removed by a proteolytic mechanism. Indeed, removal of this half does not ablate physiological function. The resemblance of the 5' domain of the molecule to erythropoietin may explain the synergy of thrombopoietin and erythropoietin in megakaryocyte colony formation and platelet production. It is well recognized that splenectomized individuals with persistent anaemia usually have significant thrombocytosis and many individuals with red cell aplasia and high erythropoietin levels also have thrombocytosis and megakaryocytosis.

Circulating platelets

The differential diagnosis of thrombocytopenia rests first on evaluation of platelet morphology. In conditions in which megakaryocytopoiesis is accelerated, circulating platelet volume (and usually diameter) is increased. The reasons for this shift in volume are disputed. Some claim that young platelets are larger than old platelets, while others suggest that large megakaryocytes give rise to large platelets. Neither explanation satisfies all experimental and clinical conditions, but in general, thrombocytopenia secondary to increased destruction of platelets is associated with platelets of large volume. Thrombocytopenia related to decreased production of platelets is associated with platelets of normal size.

There are major exceptions to this rule. Patients with hyposplenism tend to have large platelets in their blood, whether thrombopoiesis is increased or not, and patients with primary abnormalities of platelet function, such as Wiskott–Aldrich syndrome or Bernard–Soulier syndrome, have small and large platelets, respectively, that bear no relationship to platelet production. Thrombopoietin increases platelet production by increasing both the number and size of individual megakaryocytes. Though thrombopoietin is probably solely responsible for the later stages of recognizable megakaryocyte differentiation and proliferation of megakaryocyte progenitors, its function depends, at least in part, on the additional stimulation of earlier megakaryocyte progenitors with other growth factors, including IL-3, IL-11, and Steel factor.

Down-regulation of megakaryocytes

There is great uncertainty about possible down-regulation of megakaryocytes. Platelet factor 4 seems to down-regulate colony formation *in vitro*. If active *in vivo*, this would provide an interesting feedback loop. Transforming growth factor-b is also a potent inhibitor *in vitro*. Natural killer cells, which are thought by some to be general suppressors of haemopoiesis *in vitro*, actually enhance megakaryocyte colony formation *in vitro*. In addition, an antibody to natural killer cells, when it is given intraperitoneally in massive doses to mice, abolishes the formation of colonies of mega-karyocytes that can be grown in culture from murine marrow. Natural killer cells may thus actually play a stimulating role *in vivo*.

Megakaryocyte progenitors in disease

A number of attempts have been made to relate diseases associated with elevated or depressed platelet counts to the number or the growth characteristics of megakaryocyte progenitors. Megakaryocyte progenitors in essential thrombocythaemia are similar in their growth characteristics to the expanded numbers of erythroid progenitors in polycythaemia vera. The latter develop into erythroid colonies without additions of erythropoietin to the culture medium. The trace of erythropoietin in the serum is sufficient to drive the sensitive receptor system in these progenitors. In a similar fashion, the numerous CFU-Meg in essential thrombocythaemia develop into megakaryocyte colonies in the absence of stimulation by aplastic anaemia serum. They are 'thrombopoietin independent' and many produce endogenous thrombopoietin.

Clinical studies with haemopoietic growth factors

Several recombinant haemopoietic growth factors are currently in use and under evaluation in a variety of clinical settings. Initial studies focused on erythropoietin in the anaemia of chronic renal failure, and GM-CSF and G-CSF in both transient and long-standing bone marrow-failure syndromes. These three factors are now commercially available for clinical use. More recently, other haemopoietic growth factors such as M-CSF, IL-3, and Steel factor are coming under scrutiny.

Anaemia is a major complication of endstage renal failure, and is due primarily to a reduction in erythropoietin production. Several phase I, II, and III studies have documented that recombinant human erythropoietin can induce a dose-dependent increase in effective erythropoiesis. The extension of this treatment to patients who do not yet require dialysis has met with similar success. Erythropoietin may also be useful in the anaemia of chronic disease and in the anaemia that complicates azidothymidine treatment of patients with acquired immune deficiency disease (**AIDS**).

G-CSF has proven to be useful for shortening the period of neutropenia following myelosuppressive anticancer chemotherapy, and has been approved in the United States and Europe for reduction of infection in patients with non-myeloid malignancies. GM-CSF and G-CSF can accelerate haemopoietic reconstitution after bone marrow transplantation, and GM-CSF has been approved for use in the United States in autologous transplantation. In the context of bone marrow failure, GM-CSF is a useful palliative treatment as it can increase the neutrophil count, particularly in the majority of children with acquired aplastic anaemia. GM-CSF can also increase neutrophils, eosinophils, and monocytes in AIDS. Most patients with Kostmann syndrome, a rare inherited severe failure of neutrophil production, respond dramatically to G-CSF treatment. Patients with other defects of neutrophil production such as cyclic neutropenia and chronic idiopathic neutropenia have also responded to this factor.

Recombinant human thrombopoietin or its polyethylene glycol (**PEG**)-derivatized 163 residue aminoterminus (PEG-MGDF) stimulates megakaryocyte proliferation and endoreduplication *in vitro* and is a potent inducer of megakaryocytopoiesis and platelet production *in vivo* in mice and non-human primates. Both recombinant human thrombopoietin and PEG-MGDF are safe and show no organ toxicity. In normal volunteers a single bolus of 3 µg/kg per day of PEG-MGDF doubles the blood platelet concentration by day 12, with a return to baseline by day 28. A stimulatory effect on platelet production was observed when thrombopoietin or PEG-MGDF was administered after chemotherapy to more than 100 patients with cancer, with a decrease in the time for platelet counts to return to normal and elevated platelet nadirs. Antibodies to thrombopoietin have been reported in one patient with cancer and in volunteers given subcutaneous PEG-MGDF. Further clinical development of this thrombopoietin formulation has been stopped, since transient decreases in platelet count were noted. It is possible that the factor is more antigenic when given by the subcutaneous route.

Summary

Haemopoiesis is the process of terminal differentiation of recognizable immature precursors of the formed elements of the blood. Renewal of the precursor pool is accompanied by the differentiation of committed progenitor cells that are themselves renewed by a process of stochastic maturation of stem cells. A group of haemopoietins derived from T cells, monocytes, and fibroblasts governs the differentiation of committed progenitor cells by mechanisms yet to be defined.

The *mélange* of marrow cells described above exists in delicate fronds thrust into the venous sinuses. Cells are packed in close proximity within the fronds, held together by extensions of fibroblast cytoplasm and fibronectin. Such a delicate anatomy is subject to a myriad of abnormalities that can disturb the orderly progress of cell–cell interactions that govern the system. The multiple symptoms of bone marrow failure are the results of these disturbances.

Further reading

Clark S, Nathan DG, Sieff CA (1997). The anatomy and physiology of hematopoiesis. In: Nathan DG, Orkin SH, eds. *Hematology of infancy and childhood.* W.B. Saunders, Philadelphia. [Comprehensive chapter that discusses haemopoiesis in more detail.]

Cosman D *et al*. (1990). A new cytokine receptor superfamily. *Trends in Biochemical Sciences* **15**, 265–70. [Review of the haemopoietic growth factor receptors.]

Drachman JG (2000). Role of thrombopoietin in hematopoietic stem cell and progenitor regulation. *Current Opinion in Hematology* **7**, 183–90.

Gerson SL (1999). Mesenchymal stem cells: no longer second class marrow citizens. *Nature Medicine* **5**, 262–4.

Goodell MA *et al.* (1997). Dye efflux studies suggest that hematopoietic stem cells expressing low or undetectable levels of CD34 antigen exist in multiple species. *Nature Medicine* **3**, 1337–45. [Demonstration that purified murine and rhesus 'side population' cells do not express CD34 and have stem cell properties.]

Gussoni E *et al.* (1999). Dystrophin expression in the mdx mouse restored by stem cell transplantation. *Nature* **401**, 390–4. [Example of the plasticity of both haemopoietic and muscle stem cells.]

Horwitz EM *et al.* (1999). Transplantability and therapeutic effects of bone marrow-derived mesenchymal cells in children with osteogenesis imperfecta. *Nature Medicine* **5**, 309–13.

Kaushansky K (1995). Thrombopoietin: the primary regulator of megakaryocyte and platelet production. *Thrombosis and Haemostasis* **74**, 521–5.

Metcalf D (1984). *The hemopoietic growth factors*. Elsevier, Amsterdam.

Metcalf D, Moore MAS (1971). *Haematopoietic cells*. North-Holland Publishing Company, Amsterdam.

Miyajima A *et al.* (1992). Cytokine receptors and signal transduction. *Annual Review of Immunology* **10**, 295–331. [Review of haemopoietic growth factor signal transduction pathways.]

Nicola NA (1989). Hemopoietic cell growth factors and their receptors. *Annual Review of Biochemistry* **58**, 45–77.

Osawa M *et al*. (1996). Long-term lymphohematopoietic reconstitution by a single CD34– low/negative hematopoietic stem cell. *Science* **273**, 242–5. [Original paper showing that murine long-term reconstituting stem cells do not express CD34.]

Shivdasani RA, Orkin SH (1996). The transcriptional control of hematopoiesis. *Blood* **87**, 4025–39.

Weissman IL (2000). Translating stem and progenitor cell biology to the clinic: barriers and opportunities. *Science* **287**, 1442–6. [Review of current research and clinical potential of haemopoietic stem cells.]

22.2.2Stem-cell disorders

D. C. Linch

Concept of the haemopoietic stem cell and its disorders

The haemopoietic stem cell is a poorly defined entity with an undifferentiated phenotype, which resides within the haemopoietic tissue. It has the ability to self-renew, and the capacity to generate large numbers of mature progeny of multiple haemopoietic lineages. It was considered that such cells were irreversibly committed to haemopoiesis. More recent data suggest that stem cells have far greater plasticity than previously appreciated. Within the adult marrow, stem cells have been found that can give rise not only to blood cells, but also to other mesodermally derived tissues such as endothelium and muscle and to hepatic and neuronal cells traditionally thought to be derived from the endoderm and ectoderm, respectively. In some instances the extensive repertoire is due to the presence of non-haemopoietic mesenchymal stem cells within the bone marrow, but there is also evidence that stem cells with haemopoietic potential can give rise to other tissue-types. In addition, stem cells within the brain have been shown to be capable of generating cells that are embryologically derived from all three germ layers, and this includes the generation of haemopoietic cells. Either very undifferentiated multipotent stem cells persist in many adult tissues or, under appropriate conditions, some stem cells can undergo dedifferentiation and reprogramming. The most dramatic evidence for the possibility of dedifferentiation comes from 'Dolly the sheep', generated by the transfer of a mature cell nucleus into an enucleated egg. Clearly, the environment within the egg cytoplasm can reprogramme the genetic material within the nucleus. It is conceivable that similar processes could be induced by an appropriate extracellular milieu. It is not yet clear whether these new insights into the potential of the stem cell will alter the way in which we consider the haemopoietic stem cell disorders, but some of the fundamental assumptions of recent decades may be challenged.

Quantitative and qualitative abnormalities of haemopoietic stem cells can be envisaged. The stem-cell population might become depleted and fail to produce adequate mature progeny. Similarly, a normal number of abnormal stem cells could fail to proliferate normally and thus generate the same deficiency of end-cells. Abnormal stem cells could also produce normal numbers of defective end-cells, or he stem cells could undergo malignant transformation. These possibilities are not mutually exclusive and transitional forms can be envisaged.

In practice, the definition of a stem-cell disorder is often extremely difficult. Self-renewal, one of the hallmarks of a stem cell, cannot be considered in the context of malignant disorders, as any malignant clone, arising in any tissue, at any stage of differentiation, must have undergone immortalization and be capable of self-renewal. The term 'stem-cell disorder' is usually used, therefore, to imply that the target cell for the disease process has occurred in a cell with the potential to develop into cells of different lineages. Such a cell could be a relatively 'late' or 'lineage-restricted, stem-cell' capable, for example, of giving rise to phagocytes and erythrocytes, or it could be a very primitive stem cell capable of giving rise to all myeloid and lymphoid lineages. There are, however, a number of difficulties with such a definition. First, malignant change in a very primitive cell does not necessarily lead to the production of mature cells of multiple lineages. It is a feature of the acute leukaemias that there is a block in differentiation; in some cases of acute myeloid leukaemia (**AML**), no mature progeny are produced by the malignant clone. In other cases of AML, neutrophils alone are produced. However, it cannot be assumed that the target cell of the original oncogenic event was not a cell with the potential to form all the myeloid elements, including the red cell series. Second, whereas an immature phenotype of a malignant cells was always considered to be indicative of the transformation of a very early cell, we must now consider the possibility that transformation could arise in a later cell with subsequent dedifferentiation.

The concept of a haemopoietic stem-cell disorder is, therefore, imprecise. Pragmatically a stem-cell disorder is most easily considered as a disease with multilineage involvement.

Detection of multilineage involvement

In the bone marrow hypoplastic states, stem-cell involvement is obvious because of the pancytopenia that occurs. In the myeloproliferative disorders examination of the blood count and blood film also reveals the involvement of multiple lineages; neutrophil leucocytosis may coexist with eosinophilia, basophilia, and thrombocytosis. A number of more sophisticated techniques have also been used to demonstrate that a particular cell lineage is involved in a clonal process (Table 1).

Analysis of non-random cytogenetic abnormalities represents one of the most longstanding techniques for examining cell-lineage involvement in the haematological malignancies. This approach was used in the investigation of acute myeloid leukaemia to show that the large majority of T cells in the peripheral blood were not part of the malignant clone. The combination of this technique with immunophenotyping for lineage-specific markers provides a powerful addition to this approach. Conventional cytogenetic studies suffer the disadvantage that only cells in metaphase can be analysed, but techniques such as fluorescent *in situ* hybridization (**FISH**) (with or without immunophenotyping) allow cells in interphase to be examined.

Somatic mutations, from major chromosomal alterations to point mutations, can also be detected using non-microscopic methods employing recombinant DNA technology. Polymerase chain reaction (**PCR**) methods are particularly sensitive and make it possible to study small samples of cells, but it is difficult to make the techniques fully quantitative. Errors in interpretation can readily arise from minor contaminating cells in supposedly purified cell populations. Analysis of somatic mutations is subject to a further potential problem if the mutation is a secondary event in the disease process. Under these circumstances the mutation observed could have arisen in a subclone and not be present in all the malignant cells. In acute myeloid leukaemia mutations in *N-ras* may be present in all or just a few cells; sometimes a mutation detected at presentation cannot be found in relapse, indicative of the process of clonal evolution and the fact that the *ras* mutation was not an early event in the leukaemogenic process.

The use of polymorphic X-linked markers has been a very useful tool for examining clonality in informative females, and is not subject to the problems of clonal evolution. The original studies used the enzyme glucose-6-phosphate dehydrogenase (**G6PD**). However, there are many structural variants of this enzyme, the most common normally active variant being designated as B type. In populations of African descent a common normal variant exists which is designated as A type. Although this variant only differs by one amino acid it can be readily separated from the B type on starch-gel electrophoresis. Because the gene which codes for G6PD is on the X-chromosome, individual cells of a heterozygous female (AB) express only one enzyme type, with approximately half the cells expressing type A and half type B (in other words, the individual is a mosaic). This restricted pattern of gene expression arises because of the process of random X-inactivation, known as lyonization, which occurs in early embryonic life and is passed on to the progeny of those cells in a stable manner. (Fig. 1). Malignant disorders nearly always arise in a single cell, and thus all the malignant cells in a particular patient will have the same X-chromosome inactivated. In an informative G6PD female all the tumour cells will express either type A or type B enzyme. Analysis of the G6PD levels in blood cells of different lineages will help to determine whether they are involved in a haematological malignancy. This technique is limited by the low frequency of informative polymorphisms in populations other than those of African descent.

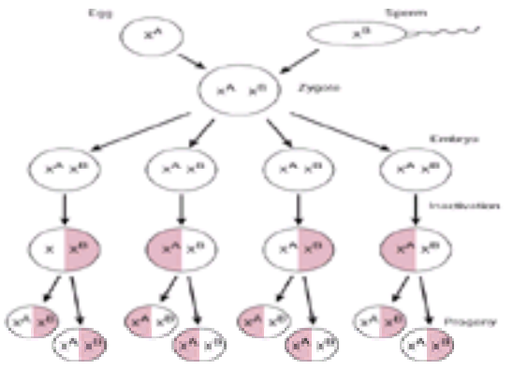

Fig. 1 X-chromosome inactivation.

Clonality can also be investigated using X-linked DNA polymorphisms which do not result in different protein products. This is based on the fact that the active and inactive genes are differentially methylated at specific cytosine residues. DNA samples are first digested with an appropriate restriction endonuclease to distinguish maternal and paternal copies of the gene, and subsequently with a restriction endonuclease sensitive to cytosine methylation in its recognition sequence to distinguish active from inactive copies of the gene. Useful genes to study include the hypoxanthine phosphoribosyl transferase (*HPRT*) gene and the phosphoglycerate kinase (*PGK*) gene (Fig. 2). The X-linked, multiple tandem repeat recognized by the probe M27B is highly informative (approximately 80 per cent of females are heterozygous), but as this is not a gene it is not really correct to talk about active and inactive copies. None the less, it acts as a useful marker of the inactivated X-chromosome, and results with this probe are concordant with those obtained with HPRT or PGK. PCR-based assays have also been developed to examine the methylation status of a number of genes, including the monoamine oxidase A gene (*MOA*), the human androgen-receptor gene (*HUMARA*), and the Fragile X gene. These assays require far fewer cells than are required for techniques based on Southern blotting and hence are now more frequently used. The *HUMARA* gene is the most informative, with heterozygosity rates in Caucasian populations of about 90 per cent. More recently a number of reverse transcriptase, polymerase chain reaction (**RT-PCR**) assays have been introduced that enable direct analysis of the relative expression of the two alleles at the transcript level, which may circumvent the problem of complex DNA methylation patterns. Informative genes must contain polymorphisms in the coding sequence, although these do not necessarily have to lead to changes in the amino acid sequence. It is also essential that the gene to be analysed is expressed by the cell type being investigated.

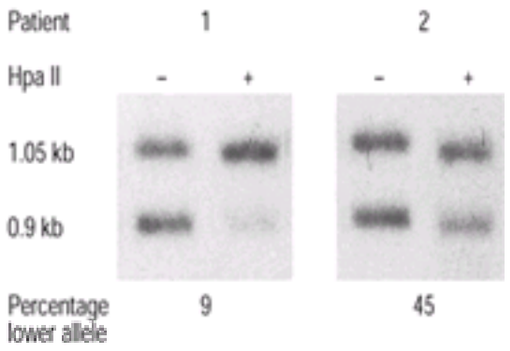

Fig. 2 Clonal analysis of PCK heterozygotes.

The *HUMARA* gene is again the most informative gene as the polymorphic variable number tandem repeat (**VNTR**) is contained within the coding sequence. The two alleles can be readily distinguished on electrophoresis of the RT-PCR product, but unfortunately this gene is not expressed in all haemopoietic tissues. The transcripts most commonly studied in haematological samples include G6PD, iduronate-2-sulfatase (**IDS**), and the palmitoylated membrane protein p55 (p55). Together these three RNA transcripts are informative in about 70 per cent of females. Once the mRNA has been reverse-transcribed into cDNA the different alleles, which differ only by single base changes, are then be detected by allele-specific PCR or allele-specific restriction analysis. Even with these genes expression is not constant between different haemopoietic cell-types. In one study IDS expression was shown to be sixfold higher in T cells than in neutrophils. High sample purity is thus essential when using this methodology.

Clonality studies based on X-chromosome inactivation patterns have three main drawbacks. First, it must be appreciated that lyonization occurs early in embryogenesis when there are few stem cells destined to give rise to the different tissues. As a consequence of this and the random nature of X-inactivation, considerable constitutive skewing away from the expected 50:50 expression of maternal and paternal alleles occurs in some individuals. An ill-defined inherited component may also contribute to this random process. In approximately one-quarter of females more than 75 per cent of the expressed genes derive from the same allele, and in 3 per cent of normal individuals more than 90 per cent. It is therefore essential that X-chromosome inactivation patterns are interpreted with reference to normal tissue. This has frequently been omitted and many of the reports in the literature are thus suspect. Furthermore, in the case of the haematological malignancies it is not always easy to obtain appropriate control samples. Non-haemopoietic tissues can be misleading controls, as X-inactivation patterns can vary between tissues. T cells are probably a good control in most myeloid malignancies as they derive from the same stem-cell pool as the myeloid cells, and it is unlikely that the majority of T lymphocytes will be involved in the malignant clone. Second, it has been clearly demonstrated that skewing of the myeloid lineages is acquired in a significant proportion of elderly females, so clonal analysis of haemopoietic lineages must be limited to females under the 65 years of age. Third, it must be remembered that the study of X-inactivation patterns is an insensitive technique which can not be used to detect a minor clone within a polyclonal population.

Myeloproliferative disorders

The term 'myeloproliferative disorders' was invented by Dameshek and others in the 1950s in an attempt to explain the variability of the haematological findings in polycythaemia rubra vera, chronic myeloid leukaemia, and myelofibrosis, and the existence of intermediate and transitional forms. The myeloproliferative disorders are characterized by the predominant cell type produced by the malignant clone, but they all involve a primitive stem cell (Fig. 3). Acute myeloid leukaemia frequently arises at the stem-cell level and diseases such as chronic myeloid leukaemia and polycythaemia rubra vera not infrequently terminate in 'blastic transformation'. By convention, however, the term 'myeloproliferative disorders' is usually reserved for the chronic malignancies where mature myeloid cells predominate.

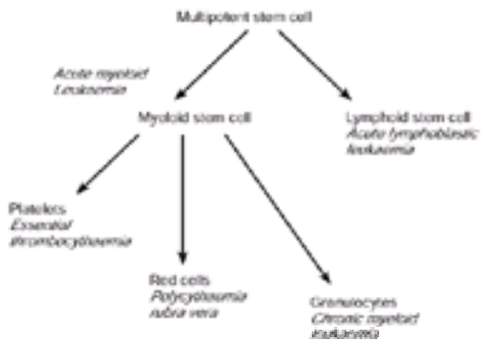

Fig. 3 The myeloproliferative disorders.

Lineage involvement was first studied in chronic myeloid leukaemia because of the presence of the characteristic Philadelphia chromosome [t(9;22)(q34;q11)]. Not only were all cells of the phagocytic- and red-cell series found to be involved, but Epstein–Barr virus (**EBV**)-transformed B lymphocytes were also shown to contain the Ph1 chromosome, thus indicating that the target cell for malignant transformation was a primitive stem cell with both myeloid and lymphoid potential. This is confirmed by the fact that about one-third of blastic transformations are due to the accumulation of primitive B cells.

A similarly primitive stem cell is thought to be the cell of origin of polycythaemia rubra vera (**PRV**), and at least some cases of essential thrombocythaemia (**ET**). Clonality studies have revealed, however, that a significant proportion of cases of ET are polyclonal and not malignant disorders, despite fulfilling the usual diagnostic criteria. The cause of the dysregulated blood cell production in such cases is unknown. An important observation made by several groups is that people with the

polyclonal forms of ET have a lower incidence of thrombosis.

Dameshek had considered that idiopathic myelofibrosis was part of the myeloproliferative disease spectrum, and indeed this is the case. The fibroblasts are not part of the malignant clone, however, but are a reaction to an underlying myeloid malignancy. Where studied, the myeloid cells have been shown to be clonal. Fibrosis is particularly common when cells of the megakaryocytic series predominate, and may be due to the excessive local production of platelet growth factors such as platelet-derived growth factor (**PDGF**) and transforming growth factor-b (**TGF-b**).

Acute leukaemias

In children with acute myeloid leukaemia (**AML**) the red cells and platelets do not appear to be part of the malignant clone, whereas such tri-lineage involvement is frequent in adults. There is no convincing evidence of lymphoid involvement. Considerable attention has been focused on the notion that remission in AML may represent persistence of the malignant clone with full differentiation to give a normal blood count. This view is based on studies that did not pay adequate attention to the skewing of X-inactivation patterns that can occur in normal individuals; true 'clonal remission' is rare.

Myelodysplasia, which frequently precedes AML, often involves all myeloid lineages, as is evident from examination of the blood and bone marrow films. There is considerable controversy within the literature, but the majority of studies do not demonstrate involvement of the lymphoid series.

Acute lymphoid leukaemia (**ALL**) is usually restricted to the B-cell or T-cell lineage. An exception occurs in cases of Ph1-positive ALL where the myeloid lineages are often involved. This entity is akin to chronic myeloid leukaemia (**CML**) presenting in lymphoid blast crisis.

Aplastic anaemia

Aplastic anaemia by definition refers to involvement of multiple myeloid lineages (pancytopenia). In severe cases the lymphocyte count is also reduced, suggesting that the defect is at the level of stem cells with the potential to give rise to both myeloid and lymphoid elements. Although T-cell numbers tend to be relatively well preserved, it must be appreciated that many T cells are long-lived cells and their numbers would not be expected to fall rapidly if their production from stem cells ceased. Furthermore, the basis of immunological memory, and a characteristic difference between myeloid and lymphoid cells, is that the mature progeny of the lymphoid stem cells can undergo amplification-division and self-renewal.

In those patients who respond (at least partially) to immunosuppression, with long-term follow-up there is a very high incidence of the development of clonal disorders such as paroxysmal nocturnal haemoglobinuria (**PNH**), myelodysplasia, and AML. In some patients at presentation, the few remaining granulocytes are clonal, although it is not clear how a clonal disorder can give rise to a hypoplastic bone marrow.

Paroxysmal nocturnal haemoglobinuria

PNH is a clonal disorder, due to a somatic mutation in the haemopoietic stem cell in the X-linked phosphatidylinositol glycan-A (**PIG-A**) responsible for the assembly of glycosyl phosphatidylinositol (**GPI**)-linked proteins on the cell surface of that cell and its progeny. This results in complement hypersensitivity and low-level expression of a number of antigens which are useful for defining lineage involvement. These include CD59 for red cells, platelets and T cells, CD67 for granulocytes, CD14 for monocytes, and CD24 for B cells. Flow cytometric studies have revealed variable lineage involvement: red cells, granulocytes, and monocytes and natural killer (**NK**) cells are involved in most cases; B cells are involved in a proportion of cases, and there is one report of a subpopulation of T cells involved in the PNH clone. It is possible that the variable lineage involvement represents differences in the target cell for the initiating mutation. The immunophenotypic studies have also confirmed that there is variable persistence of normal haemopoiesis, and some patients have more than one PNH clone with different levels of expression of GPI-linked molecules. One of the major unresolved questions in PNH is how does the PNH clone, which is not usually considered to be a malignancy, acquire a growth advantage over the normal haemopoietic tissue.

Further reading

Abrahamson G, *et al*. (1991). Clonality of cell populations in refractory anaemia using combined approach of gene loss and X-linked restriction fragment length polymorphism-methylation analyses. *British Journal of Haematology* **79**, 550–5.

Adamson JW, *et al.* (1976). Polycythaemia vera: stem cell and probable clonal origin of the disease. *New England Journal of Medicine* **295**, 913–16.

Beutler E, Collins Z, Irwin LE (1967). Value of genetic variants of glucose-6-phosphate dehydrogenase in tracing the origin of malignant tumours. *New England Journal of Medicine* **276**, 389–91.

Bjornson CR, *et al.* (1999). Turning brain into blood: a haemopoietic fate adopted by adult neural stem cells *in vivo*. *Science* **283**, 534–7.

Busque L, *et al.* (1996). Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* **88**, 59–65.

Dameshek W (1951). Some speculations on the myeloproliferative syndromes. *Blood* **6**, 392–5.

Fialkow PJ (1972). Use of genetic markers to study cellular origin of development of tumours in human females. *Advances in Cancer Research* **15**, 191–226.

Fialkow PJ, Jacobson RJ, Papayanopoulou T (1977). Chronic myeloid leukaemia: clonal origin in a stem cell common to the granulocytic, erythrocyte, platelet and monocyte/macrophage. *American Journal of Medicine* **63**, 125–31.

Fialkow PJ, *et al*. (1987). Clonal development, stem cell differentiation and clinical remissions in acute non-lymphocytic leukaemia. *New England Journal of Medicine* **317**, 468–73.

Gale RE, Wheadon H, Linch DC (1991). X-chromosome inactivation patterns using HPRT and PGK polymorphisms in haematologically normal and post-chemotherapy females. *British Journal of Haematology* **79**, 193–7.

Gale RE, *et al*. (1993). Frequency of clonal remission in acute myeloid leukaemia. *Lancet* **341**, 138–42.

Gale RE, *et al.* (1994). Tissue specificity of X-chromosome inactivation patterns. *Blood* **83**, 2899–905.

Gale RE, *et al.* (1997). Acquired skewing of X chromosome inactivation patterns in myeloid cells of the elderly suggests stochastic clonal loss with age. *British Journal of Haematology* **98**, 512–19.

Harrison C, *et al.* (1999). A large proportion of patients with a diagnosis of essential thrombocythaemia do not have a clonal disorder and may be at lower risk of thrombotic complications. *Blood* **93**, 417–25.

Hillmen P, Richards SJ (2000). Implications of recent insights into the pathophysiology of paroxysmal nocturnal haemoglobinuria. *British Journal of Haematology* **108**, 470–9.

Kurzrock R, Gutterman JU, Talpaz M (1988). The molecular genetics of Philadelphia chromosome-positive leukaemias. *New England Journal of Medicine* **319**, 990–8.

Lyon MF (1961). Gene action in the X-chromosome of the mouse (*Mus musculus* L). *Nature* **190**, 372–3.

Naumova AK, *et al.* (1996). Heritability of X-chromosome inactivation phenotype in a large family. *American Journal of Human Genetics* **58**, 1111–19.

Nissen C, *et al*. (1986). Acquired aplastic anaemia: a PNH-like disease? *British Journal of Haematology* **64**, 355–62.

Rowley JD (1973). A new consistent chromosomal abnormality in chronic myeloid leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290–3.

Turhan AG, *et al.* (1988). Molecular analysis of clonality and *bcr* rearrangements in Philadelphia chromosome positive acute lymphoblastic leukaemia. *Blood* **71**, 1495–500.

van Kamp H, *et al*. (1991). Clonal haematopoiesis in patients with acquired aplastic anaemia. *Blood* **78**, 3209–14.

Vogelstein B, *et al*. (1987). Clonal analysis using recombinant DNA probes from the X-chromosome. *Cancer Research* **47**, 4806–13.

Weissman I (2000). Translating stem and progenitor cell biology to the clinic: barriers and opportunities. *Science* **287**, 1442.

22.3.1 Cell and molecular biology of human leukaemias

*Thomas Look**

Introduction

The human leukaemias arise from haemopoietic stem and progenitor cells, and exhibit differentiation arrest in any lineage and in any stage of maturation. Attempts to understand the pathobiology of these diseases have focused on clinical presentation, cell morphology, histochemistry, cell immunophenotype, cytogenetics, and, in recent years, molecular genetics. Although useful in diagnosis and risk assessment, clinical and cell-biological findings ultimately fail to reveal underlying mechanisms of leukaemic transformation and, therefore, cannot account for treatment failures among groups of patients with ostensibly similar features. The emerging picture of leucocyte transformation indicates that most cases of leukaemia involve chromosomal translocations that result in aberrantly expressed transcription factors or activated tyrosine kinases, which drive malignant conversion and maintain the leukaemic phenotype. Some of these genetic changes affect cell proliferation or survival, while others exert their primary effects on cell differentiation. Almost always, the critical lesion involves a 'master' transcriptional regulatory gene or tyrosine kinase signalling molecule that stands near the top of a hierarchy of gene control, so that leukaemia is efficiently instigated by a limited number of alterations rather than by multiple changes affecting tens of responder genes in the biochemical cascade.

This chapter describes the human leukaemias in terms of their characteristic molecular genetic and developmental features. A recurring theme is that the genes most frequently altered in these diseases are those with evolutionarily conserved roles in the embryological development of various cell lineages and organ systems, including, but not limited to, genes that control normal blood cell production (haemopoiesis). Whenever possible, I have attempted to link advances in the molecular biology of a particular leukaemia to any unique implications for treatment and prognosis.

Oncogenic transcription factors and activated tyrosine kinases

Transcriptional control genes are common mutational targets in the human leukaemias because their protein products (transcription factors) bind to regulatory elements in DNA, such as promoters and enhancers, and stimulate or inhibit gene expression. These proto-oncogenes are frequently activated by chromosomal translocations, either by fusion of disparate gene fragments or by mobilization of the gene into the vicinity of transcriptionally active T-cell receptor (*TCR*) or immunoglobulin (*IG*) genes. More than 80 per cent of the oncogenic transcription factors identified to date can be classified according to the structural motifs within their DNA- and protein-binding domains: **bHLH** (basic region/helix–loop–helix), **bZIP** (basic region/leucine zipper), **HTH** or homeodomain (helix–turn–helix), A–T hook, Ets-like, Runt homology, zinc finger, and LIM (cysteine-rich). Each of these motifs has functional significance, in that it defines the downstream responder genes that determine the altered patterns of gene expression during normal development and malignant transformation.

Current observations indicate that oncogenic transcription factors act positively to aberrantly upregulate critical target genes that coordinate the production of proteins needed for normal cell proliferation, differentiation, or survival (gain of function), or negatively to interfere with normal regulatory cascades controlling apoptosis and the growth that normally accompanies differentiation (loss of function). Similarly, activated tyrosine kinases act through signal-transduction cascades that ultimately dysregulate the transcriptional control of gene expression. The oncogenic transcription factors or activated tyrosine kinases involved in the human leukaemias have unique transforming properties, which tend to be specific for different types of progenitor cells developing in the lymphoid or myeloid pathways. In most cases of acute lymphoblastic leukaemia and acute promyelocytic leukaemia, the initial lesion appears to affect progenitors at the same stage of differentiation as the predominate phenotype in the malignant clone. However, most types of acute myeloid leukaemia appear to arise in a primitive haemopoietic stem cell rather than a committed myeloid progenitor, with subsequent blockade of differentiation that determines the morphological (FAB) subtypes of myeloid leukaemia apparent at diagnosis. Primitive normal stem cells are also thought to be the targets of leukaemic transformation in chronic myeloid leukaemia and in some cases of acute lymphoblastic leukaemia, at least those expressing the *BCR–ABL* and *MLL* fusion oncogenes.

Acute lymphoblastic leukaemia

Approximately 3000 to 4000 people in the United States, two-thirds of them children, develop acute lymphoblastic leukaemia (**ALL**) each year. Only 20 to 25 per cent of the childhood cases are resistant to modern multiagent therapy, attesting to the remarkable treatment advances that have been made over the past three decades. Unfortunately, the prognosis for adults with ALL remains poor. Fewer than 50 per cent of the patients treated solely with chemotherapy become long-term survivors, and many patients are ineligible for or fail in spite of bone marrow transplantation. The difference in curability between acute leukaemias in children and adults can be attributed partly to the much larger proportion of adult cases with the *BCR–ABL* chimeric tyrosine kinase oncogene, resulting from the Philadelphia chromosome, and to the poor tolerance of intensive chemotherapy by patients over 50 years of age. There is also the possibility that ostensibly similar cases of ALL in children and adults differ in, as yet undefined, genetic ways that influence the outcome of therapy.

B-Lineage ALL

Normal B-lymphoid cell populations undergo diverse, clonal rearrangements of their *IG* genes, followed by highly regulated proliferation of cells that successfully complete the process and produce immunoglobulin. When this developmental process is altered, the developing B lymphocytes may undergo transforming events that eventually lead to overt ALL. In most instances, the pathobiology of transformed lymphoid cells mirrors the altered expression of genes that contribute to the normal functioning of pro-B (immunoglobulin undetectable) or pre-B (cytoplasmic immunoglobulin-positive) lymphocytes or occasionally mature virgin B cells (surface immunoglobulin-positive), although it may involve the aberrant expression of normally quiescent genes. Approximately 80 per cent of patients with ALL have lymphoblasts whose phenotypes correspond to those of B-cell precursors. Only 2 to 3 per cent of these patients have mature B-cell leukaemia, which is thought to represent a disseminated form of Burkitt lymphoma. The prevailing concept is that lymphoid leukaemic cells are classified immunophenotypically according to their 'normal' developmental stage, as this convention provides a firm basis for the study of cell type-specific genetic alterations.

BCR–ABL

The first constitutively activated tyrosine kinase described in ALL results from fusion of 5' sequences of the *BCR* proto-oncogene to 3' sequences of *ABL* due to action of the t(9;22) translocation, leading to creation of the so-called Philadelphia (**Ph**) chromosome. Most investigators agree that the t(9;22) translocation occurs in haemopoietic stem cells possessing both lymphoid and myeloid differentiative potential. Differences between *BCR–ABL*$^{+}$ leukaemias in children and adults are striking. Found in only 3 to 5 per cent of newly diagnosed paediatric ALL cases, they represent at least 25 per cent of adult cases. Also, the *BCR* breakpoints on chromosome 22 in paediatric cases cluster mainly in the *m-bcr* (minor breakpoint cluster region), whereas in adults they can occur in either the *m-bcr* or the *M-bcr* (major breakpoint cluster region). Breaks in the former region lead to fusion genes encoding a 190-kDa protein (p190), while those in the latter generate a 210-kDa protein (p120).

Very little is known about the normal function of ABL, although as a nuclear protein with tyrosine kinase activity, it may stimulate p53-dependent growth arrest, suggesting a role as a cell-cycle checkpoint after damage to the cellular genome. Both the p190 and p210 forms of BCR –ABL are localized in the cytoplasm and have increased tyrosine kinase activity. Their enforced expression can transform haemopoietic cells *in vitro* and can induce a syndrome similar to chronic myeloid leukaemia in mice. The precise mechanism by which BCR–ABL transforms human haemopoietic cells is unclear, but this process almost certainly involves activation

of the *CRKL* gene, whose product stimulates JUN kinase and ultimately JUN through the RAS signalling pathway, as well as STATs, MYC, and cyclin D1.

Patients with *BCR–ABL*$^+$ ALL respond poorly to conventional therapy but are excellent candidates for bone marrow transplantation. A new drug, STI-571, has been identified that selectively inhibits the ABL tyrosine kinase and appears to be highly active in BCR –ABL$^+$ CML (see below). Trials are currently underway to test the activity of this drug in BCR–ABL$^+$ leukaemias that present initially as ALL.

TEL–AML1

One of the exciting recent developments in molecular genetics research of ALL was the identification of *TEL–AML1* (also termed *ETV6–CBFA2*) as the most common genetic alteration in leukaemic pro-B lymphoblasts. This chimeric oncogene is generated by the t(12;21), a translocation that is difficult to detect by standard Giemsa-banding, but which is identified in 25 per cent of cases by molecular assays. The resultant TEL –AML1 oncoprotein consists of the HLH domain of TEL, an ETS family transcription factor, fused to the DNA-binding and transactivation domains of AML1. AML1 is the DNA-binding component of the core-binding factor (**CBF**) transcription-factor complex, which is also the most frequent target of myeloid cell-associated translocations, including the t(8;21), t(3;21), and inv(16), and a critical participant in normal haemopoiesis.

The relative contributions of TEL and AML1 to leukaemia induction are only beginning to be understood. One attractive model suggests that the crucial pathogenic event is interference with AML1-mediated expression of *HOX* or other master control genes, which have pivotal roles in normal lymphopoiesis. Thus, the leukaemogenic role of TEL–AML1 may well depend on the DNA-binding and transactivating domains of AML1, although a definitive explanation will probably not be found until better *in vitro* transformation assays become available.

The prognostic impact of *TEL–AML1* expression is controversial. Some treatment centres claim long-term, event-free survival rates of approximately 90 per cent in patients with this genetic abnormality, while others report a lack of a favourable prognostic significance. Evaluation of the components of the treatment regimens suggest that TEL–AML1$^+$ cases respond better to drug combinations that rely on the intensive use of L-asparaginase and increased dosages of methotrexate with leucovorin (calcium folinate) rescue. Thus, the *TEL–AML1* fusion gene may identify a large, previously unrecognized subset of pro-B ALL cases that can be treated with regimens that rely on antimetabolites and an enzyme that depletes an essential amino acid in lymphoid cells, rather than genotoxic agents (for example, epipodophyllotoxins) that are associated with secondary AML.

E2A fusion genes

The *E2A* gene, which encodes a bHLH transcription factor on chromosome 19, is targeted by two reciprocal translocations in patients with ALL—the t(1;19) in 5 per cent of children and 3 per cent of adults, and the t(17;19) in approximately 0.5 per cent of children. The former generates one of the best-characterized fusion oncogenes in ALL, in which the two transcriptional transactivation domains of *E2A* are linked to the homeodomain of *PBX1*. *PBX1* is an orphan *HOX* gene that shares homology with the *exd* gene of *Drosophila*, which regulates in segment identity through direct interaction of its product with specific HOM proteins of the *Bithorax* and *Antennapedia* complexes.

E2A–PBX1, which binds to DNA in a site-specific manner, is clearly oncogenic in fibroblast transformation assays and in mice, and appears to induce programmed cell death (apoptosis) in lymphoid cells. Like its exd homologue, PBX1 is directed to its consensus DNA-binding site by a subset of interacting HOX proteins, whether or not it is fused to E2A. Surprisingly, these site-specific recognition sequences of PBX1 are not required for the transforming activity of E2A–PBX1. How, then, would the chimeric oncoprotein induce leukaemia, if not by disruption of gene expression normally regulated by HOX proteins? The answer appears to lie in a short peptide sequence that regulates the HOX-specific protein–protein interaction, and which is essential for leukaemogenic activity.

In about 0.5 per cent of childhood ALL cases, the t(17;19) translocation generates the *E2A–HLF* fusion gene, consisting of the amino-terminal transactivation regions of *E2A* and the C-terminal DNA-binding and dimerization domains of *HLF*, a member of the bZIP transcription factor gene family. E2A–HLF mediates leukaemic transformation by inhibiting a normal apoptotic pathway in pro-B lymphocytes, one that closely resembles the pathway responsible for eliminating extraneous neuronal cells in the nematode *Caenorhabditis elegans*. Of seven reported patients with ALL whose blast cells expressed *E2A–HLF*, each has died of leukaemia during or after aggressive treatment, suggesting a profound resistance to chemotherapy, which may arise from E2A–HLF-induced inhibition of drug-induced apoptosis.

MLL fusion genes

Structural abnormalities of chromosome 11, band q23, are found in 80 per cent of infants with ALL, but only 7 per cent of older children and adults with this disease. Recent findings of Greaves and co-workers in the United Kingdom, indicate that translocations affecting the 11q23 region can arise *in utero*, predisposing the child to the very early development of leukaemia. The vast majority of 11q23 translocations affect the *MLL* gene, which encodes a protein of 3968 amino acids with a predicted molecular mass of 431 kDa and with three regions of homology with the *Drosophila* trithorax protein (two central zinc-finger domains and a C-terminal segment of 210 amino acids). The N-terminal region, which is uniformly retained in reciprocal translocations, contains three A –T hook motifs that apparently bind A–T-rich sequences in the minor groove of DNA. The A–T hook motifs are separated from the zinc-finger regions by a 47-amino-acid sequence with homology to the non-catalytic domains of human DNA-methyltransferase, an enzyme that produces fully methylated double-stranded DNA from a hemimethylated substrate. The structural features of MLL indicate that its normal physiological function, as well as its role in leukaemogenesis, is mediated through direct interactions with DNA and other DNA-binding and signal-transduction proteins.

As with PBX1, the MLL protein appears to play a role in *HOX* gene regulation. Homozygous inactivation of the *MII* gene in mice is lethal during embryogenesis, with homeotic transformations of the skeleton reminiscent of the phenotype associated with *trithorax* mutation in flies. Knockout mice heterozygous for *MII* are small at birth, have retarded growth, and develop both anaemia and thrombocytopenia. Thus, loss of function of one *MLL* allele through chromosome breakage and translocation of the resultant fragment to a new site, together with a gain of function due to gene fusion, would be expected to interfere with the normal haemopoietic role of HOX proteins, thereby contributing to leukaemogenesis.

More than 30 discrete chromosomal sites participate in 11q23 translocations, most commonly 4q21, 9p22, and 19p13, resulting in fusion of the *AF4*, *AF9*, and *ENL* genes to *MLL*,. A recurring question has been whether the partner DNA fragments in *MLL* gene fusions are needed to induce leukaemia. In the case of *MLL–ENL*, created by the t(11;19), neither the *MLL* fragment alone nor a fusion gene lacking the *ENL* C-terminal region is sufficient to transform retrovirally transduced haemopoietic cells, suggesting that a gain of function due to the *ENL* C-terminus is required for *MLL–ENL*-induced leukaemia. Similar results have been obtained for *AF9*, in regions of the molecules required for transcriptional activation, which occurs through recruitment of co-activators like **CBP** to the transcriptional machinery. In fact, CBP itself has recently been identified as a fusion partner with MLL in rare leukaemia cases. Taken together, recent findings suggest a model in which CBP or related histone acetylases like p300 contribute in a highly regulated fashion to normal MLL function. MLL fusion proteins apparently result in aberrant gene regulation through the constitutive association of this enzymatic activity with the MLL DNA-binding domain.

The *MLL–AF4*, *MLL–AF9*, and other *MLL* fusion genes predict a dismal outcome in infants and adults treated exclusively with chemotherapy. Trials testing the efficacy of high-dose chemotherapy and radiation with stem cell replacement are currently underway.

MYC

Chromosomal translocations in B-lineage cells can also mobilize proto-oncogenes to sites adjacent to normally active enhancer or promoter elements of *IG* genes. The prototype for this mechanism in B-lineage ALL is the t(8;14), which arises in mature B cells and places the *MYC* proto-oncogene on chromosome 8 under the control of *IG* heavy-chain gene regulatory sequences on chromosome 14. Similar repositioning of *MYC* adjacent to the light-chain regulatory sequences results from the t(2;8) or the t(8;22), although in much smaller percentages of cases. Through one of these translocations, *MYC* expression becomes dysregulated, leading to abnormally increased amounts of the MYC protein, a transcription factor that forms a DNA-binding complex with another cellular protein (MAX), and eventually leads to disruption of gene expression involved in the control of cell proliferation.

Dysregulated *MYC* genes confer an exceptionally poor prognosis in children and adults with mature B-cell ALL who are treated with conventional regimens used for other types of ALL. Recent studies, however, indicate that at least 80 per cent of these patients can be cured with intensive short-term chemotherapy including high-doses of cyclophosphamide, methotrexate, and cytarabine. Thus, B-cell ALL provides the first example of a subtype of ALL requiring tailored therapy with a vastly different drug regimen for effective disease control.

T-Lineage ALL

Cell biology

First recognized as a distinct clinical entity in the early 1970s, T-cell ALL accounts for 10 to 15 per cent of acute lymphoid leukaemias in children and 20 to 25 per cent of cases in adults. The disease can arise in thymocytes at any stage of maturation, defined on the basis of reactivity with monoclonal antibodies (CD4/CD8 double-negative immature thymocytes, cytoplasmic CD3+ CD7+ CD2+ CD5+; CD4/CD8 double-positive common thymocytes, cytoplasmic CD3+ CD1+ CD2+ CD5+ CD7+ CD10+ CD4+ and CD8+; and CD4/CD8 single-positive late thymocytes, cytoplasmic CD3+ CD2+ CD5+ CD7+, CD4+ or CD8+).

Molecular oncogenesis

In contrast to the fusion oncogenes that drive the development of B-cell precursor ALL, oncogene activation in T-cell ALL reflects the mobilization of genes encoding structurally intact T-cell proto-oncogenes into the vicinity of transcriptionally active sites of the beta or alpha/delta loci of the T-cell receptor genes (TCRb or TCRa/†). Among the genes that are aberrantly expressed in thymocytes and cause leukaemic transformation through this mechanism are those representing the bHLH family of transcription factors (TAL1/SCL1, TAL2/SCL2, LYL1, and BHLHB1), the bHLH/ZIP family (MYC), other nuclear regulatory proteins (LMO1 and LMO2), homeotic proteins (HOX11), and a truncated and constitutively activated form of the human homologue of the *Drosophila* Notch 1 receptor (TAN1). The relationship of these genes to the pathogenesis of T-cell ALL has been established by their recurrent involvement in translocations that affect thymocytes or their precursors. Surprisingly, most of the T-cell oncogenes identified to date are not usually expressed in T cells; hence, their ability to induce leukaemia most likely reflects the misexpression of master transcriptional control genes with the disruption of normal T-cell developmental pathways. This is illustrated by *HOX11*, which is not normally expressed in lymphoid cells, but has been shown to be absolutely essential for normal development of the spleen.

Despite intensive cytogenetic research, chromosomal translocations have been identified in only about 25 per cent of T-cell ALL cases, suggesting that additional pathogenetic mechanisms remain to be identified. One intriguing possibility is that such cases harbour mutations (not attributable to translocations) that disrupt key transcriptional control networks in thymocyte development, leading to overt ALL. Support for this hypothesis comes from the existence of mechanisms that can induce misexpression of bHLH transcription factor genes, including *TAL1/SCL*, without a requirement for rearrangement to a site near the *TCRa/†* locus. Thus, *TAL1/SCL* may be abnormally expressed in as many as 60 per cent of T-cell leukaemias, in contrast to the 3 per cent accounted for by chromosomal translocations. Even so, it is unlikely that dysregulation of *TAL1/SCL* by itself would be sufficient to generate a fully malignant phenotype; rather, consistent misexpression of two or more T-cell oncogenes and their cooperative interaction, both among themselves and through a lack of other regulatory proteins encoded by tumour suppressor genes, is probably needed to induce clinically apparent leukaemia.

Therapeutic implications

With the availability of effective multidrug regimens, the prognostic importance of T-lineage leukaemia has disappeared in children. In adults, the addition of cyclophosphamide and cytarabine to first-line treatments boosted complete remission rates from 72 per cent to 85 per cent and improved the probability of continuous complete remission from less than 10 per cent to 46 per cent or more. The best prospect for accelerated therapeutic progress in this disease appears to lie in a fuller understanding of the molecular aspects of T-cell pathogenesis and how they relate to available modes of treatment. Ongoing studies to clarify these interactions should greatly enhance the value of molecular genetics in predicting the clinical responses of patients with T-cell ALL and, ultimately, could provide lucrative targets for novel drug therapies.

Acute myeloid leukaemia

The estimated number of new cases of acute myeloid leukaemia (**AML**) occurring annually in the United States is vastly higher in adults than in children (22 500 versus 500). Prognosis is poor in both age groups, especially in older patients with AML, although recent improvements in allogeneic bone marrow transplantation may boost cure rates above 50 per cent in those patients under 50 years of age with histocompatible donors. AML can develop from transformed cells within the granulocyte/monocyte lineage or from haemopoietic stem cells capable of differentiating into erythrocytes and megakaryocytes, as well as granulocytes and monocytes. Regardless of the cell of origin, AML pathogenesis is a multistep process, involving alterations of both proto-oncogenes and tumour suppressor genes. Although important tumour suppressors are thought to exist in regions of non-random loss of heterozygosity in this disease (for example, monosomy 7, 5q $^{-}$, 20q^{-}), current knowledge focuses on chromosomal translocations which activate transcription factor genes, reminiscent of those giving rise to ALL.

Differentiation stage in AML is best described in the context of the French–American–British (**FAB**) cell-classification system, which distinguishes among eight morphological subtypes of myeloid leukaemia, allowing useful cell type-specific comparisons of molecular genetic findings.

Acute promyelocytic leukaemia

Characterized by the clonal expansion of transformed myeloid cells blocked at the promyelocyte stage of development, the FAB M3 subtype of AML harbours a t(15;17) translocation that generates the *PML–RARa* fusion gene. *RARa* is a member of the nuclear hormone-receptor superfamily, functioning as a ligand-dependent, zinc-finger transcription factor with a critical role in normal myeloid cell differentiation. Much less is known about the normal biological function of PML, although it is a nuclear protein thought to function with p53 in cellular senescence induced by oncogene expression. The fusion protein contains nearly all the key functional domains of each molecule, including the protein–protein interaction motifs of PML and the DNA-binding, dimerization, ligand-binding, and transcriptional activation domains of RARa. The PML–RARa oncoprotein induces leukaemia by inhibiting, in a dominant fashion, the normal biological activities of both RARa and PML. The net effect is a blockade of differentiation with immortality and sustained proliferation among promyelocytes, the hallmark of acute promyelocytic leukaemia (M3 AML).

Treatment of *PML–RARa*+ AML with pharmacological dosages of the RARa ligand all-*trans*-retinoic acid (**ATRA**) results in the release of co-repressor complexes from the PML–RARa fusion protein, reversing the protein's inhibitory activity and enabling the leukaemic promyelocytes to proceed to terminal differentiation. The unique specificity of ATRA for the underlying molecular lesion in M3 AML affords a paradigm for molecularly targeted cancer chemotherapy.

Acute myeloblastic leukaemia

AML1–ETO

The *AML1–ETO* fusion gene results from the t(8;21) translocation in approximately 40 per cent of cases involving myeloblastic leukaemic cells with some evidence of differentiation (FAB M2 subtype). The oncogene product consists of the N-terminal portion of AML1, including its entire Runt homology domain (RHD), and the C-terminal portion of ETO, the mammalian homologue of the *Drosophila* protein, nervy. The AML1–ETO chimeric oncoprotein exerts its leukaemic activity by dominantly repressing the DNA-binding sites normally recognized by the AML1/CBFb core-binding factor transcriptional complex, whose function is essential for the development of all haemopoietic lineages.

Patients with *AML1–ETO*+ leukaemia respond more readily to chemotherapy than most other patients with AML, so that most experts advise reserving stem-cell transplantation for patients in first relapse. Multiple groups have documented the persistence of *AML1–ETO* fusion transcripts in patients with long-term remissions, suggesting that additional mutations within the leukaemic clone are essential for manifestation of the malignant phenotype. Since AML–ETO is a dominant-negative chimeric transcription factor that relies on co-repressor complexes, novel treatments that disrupt the formation, stability, or activity of such complexes, such as histone deacetylase inhibitors, might reverse the leukaemic phenotype, as seen with the use of ATRA in patients with acute promyelocytic leukaemia carrying the *PML–RARa* oncogene.

CBFb–MYH11

The *CBFb–MYH11* fusion product, due to inv(16) or t(16;16), is another frequent genetic lesion in newly diagnosed cases of AML. Once thought to be pathognomonic of cases with dysplastic eosinophilic precursors among myeloblasts and monoblasts (M4Eo subtype), this finding has since been made in acute myeloblastic leukaemia (M1 and M2 subtypes). These genetic rearrangements fuse the N-terminal portion of CBFb to a variable amount of the C-terminal a-helical rod domain of

MYH11, a smooth-muscle, myosin heavy-chain protein that possesses both actin binding and ATPase activity. *CBFb–MYH11* also inhibits the normal activity of the AML1/CBFb transcription factor complex, depriving the cell of requisite developmental signals. There are also suggestions that the oncoprotein generates positive signals resulting in abnormalities of cell growth. As with *AML1–ETO*, the *CBFb–MYH11* oncogene confers a better-than-average probability of achieving a sustained remission with chemotherapy.

Chronic myeloid leukaemia

Several forms of leukaemia are included under the generic term 'chronic myeloid leukaemia' (**CML**). About 90 per cent of patients with CML, both children and adults, harbour the classic t(9;22) translocation in myeloid cells, giving rise to the Ph chromosome and the *BCR–ABL* oncogene. In most patients the resulting disease is biphasic, with an initial (chronic) phase that lasts 3 years on average, and a terminal (blast) phase that is refractory to all treatment and is generally fatal within a median of 2 to 4 months. Children also can develop a juvenile form of CML (**JCML**) that often involves loss of the NF1 tumour suppressor.

Allogeneic bone marrow transplantation is the main curative treatment for CML. Patients with either the adult or juvenile form of the disease, who lack histocompatible sibling donors, may benefit from long-term treatment with interferon-a, which improves karyotypic responses, delays progression of the disease, and prolongs overall survival in randomized clinical trials. An exciting new development is the experimental agent STI-571, which targets the tyrosine kinase activity of BCR –ABL. STI-571 has produced striking improvements in each of 31 patients who were resistant to interferon and who then received STI-571 in cumulative doses of at least 300 mg per day. In three cases, the drug eradicated all cells with the t(9;22) chromosomal marker, and the malignant clone has remained undetectable in these patients for the relatively short duration of this trial.

Clinical summary

Molecular genetic changes at diagnosis are sensitive markers of potential leukaemia aggressiveness and therefore can be used as guides to treatment. Table 1 summarizes the clinical utility of recognized oncogenic transcription factors in the human leukaemias. Thus far, only two specific lesions, the *PML–RARa* fusion gene in acute promyelocytic leukaemia and the BCR–ABL kinase in CML and ALL, have been productive targets for molecular-oriented therapy, but this number will likely increase as we learn more about the genetic mechanisms that transform normal blood cells and maintain leukaemic phenotypes.

---

*I would like to thank Markus Seidel and Adolfo Ferrando for assistance and helpful discussions, and John Gilbert for editorial review and critical comments. Supported in part by NIH grants CA-59571,

Further reading

Bash RO, *et al*. (1995). Does activation of the *TAL1* gene occur in a majority of patients with T-cell acute lymphoblastic leukemia? A pediatric oncology group study. *Blood* **86,** 666–676.

Blackwood EM, Eisenman RN (1991). Max: a helix–loop–helix zipper protein that forms a sequence- specific DNA-binding complex with Myc. *Science* **251,** 1211–1217.

Bonnet D, Dick JE (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Medicine* **3,** 730–737.

Clark SS, *et al*. (1987). Unique forms of the abl tyrosine kinase distinguish Ph[1]-positive CML from Ph[1]-positive ALL. *Science* **235,** 85–88.

de The H, *et al*. (1990). The t(15;17) translocation of acute promyelocytic leukaemia fuses the retinoic acid receptor alpha gene to a novel transcribed locus. *Nature* **347,** 558–561.

Fletcher JA, *et al*. (1991). Translocation (9;22) is associated with extremely poor prognosis in intensively treated children with acute lymphoblastic leukemia. *Blood* **77,** 435–9.

Ford AM, *et al*. (1993). In utero rearrangements in the trithorax-related oncogene in infant leukaemias. *Nature* **363,** 358–60.

Golub TR, *et al*. (1995). Fusion of the TEL gene on 12p13 to the AML1 gene on 21q22 in acute lymphoblastic leukemia. *Proceedings of the National Academy of Science, USA* **92,** 4917–21.

Head DR, Pui CH (1999). Leukemia diagnosis and classification. In: Pui CH, ed. *Childhood leukemias*, pp 19–37. Cambridge University Press, New York.

Hughes TP, Goldman JM (1995). Chronic myeloid leukemia. In: Hoffman R, *et al.*, eds. *Hematology*, pp 1142–59. Churchill Livingstone, New York.

Hunger SP (1996). Chromosomal translocations involving the *E2A* gene in acute lymphoblastic leukemia: clinical features and molecular pathogenesis. *Blood* **87,** 1211–24.

Inaba T, *et al*. (1996). Reversal of apoptosis by the leukaemia-associated E2A-HLF chimaeric transcription factor. *Nature* **382,** 541–4.

Look AT (1997). Oncogenic transcription factors in the human acute leukemias. *Science* **278,** 1059–2064.

Look AT, Kirsch IR (1997). Molecular basis of childhood cancer. In: Pizzo PA, Poplack DG, eds. *Pediatric oncology*, pp 37–74. Lippincott-Raven, Philadelphia.

Meyers S, Downing JR, Hiebert SW (1993). Identification of AML-1 and the (8;21) translocation protein (AML-1/ETO) as sequence specific DNA binding proteins: the runt homology domain is required for DNA binding and protein-protein interactions. *Molecular and Cellular Biology* **13,** 6336–45.

Okuda T (1996). AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell* **84,** 321–30.

Pui CH, Evans WE (1998). Acute lymphoblastic leukemia. *New England Journal of Medicine* **339,** 605–15.

Rabbitts TH (1994). Chromosomal translocations in human cancer. *Nature* **372,** 143–9.

Rubnitz JE, *et al*. (1997). *TEL* gene rearrangement in acute lymphoblastic leukemia: a new genetic marker with prognostic significance. *Journal of Clinical Oncology* **15,** 1150–7.

Shtivelman E, *et al*. (1985). Fused transcript of *abl* and *bcr* genes in chronic myelogenous leukemia. *Nature* **315,** 550–4.

Thirman MJ, *et al*. (1993). Rearrangement of the MLL gene in acute lymphoblastic and acute myeloid leukemias with 11q23 chromosomal translocations. *New England Journal of Medicine* **329,** 909–14.

Wang Q, *et al*. (1996). The CBFb subunit is essential for CBFa2 (AML1) function *in vivo*. *Cell* **87,** 697–708.

Warrell RP, Jr, *et al*. (1991). Differentiation therapy of acute promyelocytic leukemia with tretinoin (all-trans-retinoic acid). *New England Journal of Medicine* **324,** 1385–93.

22.3.2 The classification of leukaemia

D. Catovsky

Introduction

The classification of leukaemia has evolved from a purely morphological approach, based on the appearances of the leukaemic cells in peripheral blood and bone marrow films, through cytochemical techniques and, more recently, with the use of monoclonal antibodies (**MAbs**) against cellular antigens. There has also been a major input from cytogenetic and molecular methods. The new methodologies are introducing greater precision and objectivity to the diagnostic criteria and in the assessment of prognosis in the well-defined disease entities. A new WHO classification on haemopoietic malignancies is a positive step forward and represents the consensus of most pathologists and clinicians in the field.

A broad classification of acute leukaemia includes two large groups, historically designated 'acute' and 'chronic'. Acute leukaemias represent malignancies with little evidence of differentiation; the characteristic cells are immature precursors or blasts. It is for this group where techniques other than morphology are essential for classification. The chronic leukaemias (lymphoid and myeloid) show maturation that is easily recognized morphologically, although in the diseases of lymphocytes the various subtypes can only be accurately defined by means of immunological methodology using panels of MAbs (Chapter 22.3.5).

Acute leukaemia

There are two major groups of acute leukaemia, lymphoblastic (**ALL**) and myeloid (**AML**) (Table 1). Both affect children and adults but with different frequency: 80 per cent of patients with AML are adults (over the age of 15 years) and 20 per cent children, including infants; in contrast, 85 per cent of those with ALL are children (under 15 years) and 15 per cent are adults. There are few differences in most disease features in AML between children and adults; conversely, the biological and clinical differences between childhood ALL and adult ALL are substantial.

For a diagnosis of acute leukaemia it is necessary to identify blasts as the major cellular component. With a few cytochemical reactions, namely myeloperoxidase (**MPO**), Sudan Black B, and a-naphthyl acetate esterase (**ANAE**), it is possible to distinguish the two main forms, ALL and AML. However, because cytochemistry is largely negative in ALL, it is essential to apply a battery of MAbs which can be used as markers for the two ALL cell lineages, B and T, and which can also characterize the AML blasts (Table 2). These MAbs define the immunophenotype of the disease. In immature cells some antigens are first expressed in the cytoplasm then in the cell membrane. This must be taken into account when testing blasts in suspension by flow cytometry as this method needs to be adapted for the detection of cytoplasmic antigens. Some of the markers that are specific for the T, B, and myeloid lineages, CD3, CD22, CD79a, and MPO (Table 2), are localized in the cytoplasm and not in the membrane.

Bone marrow trephine biopsies may be useful for the diagnosis of those rare cases that yield a dry tap or a hypocellular specimen in aspirates. This is the case in two uncommon forms of acute leukaemia: megakaryoblastic, or AML-M7, and hypocellular acute leukaemia, the blasts of which may be myeloid or lymphoid.

In the last decade it has become apparent that some of the acute leukaemias, both AML and ALL, can be defined by distinct chromosome translocations. These translocations often result in a fusion gene that encodes a new chimeric protein. The leukaemias so classified correlate with their response to therapy and their prognosis. For this reason, the new WHO classification has defined several types of AML and ALL by the molecular events crucial in their pathogenesis. It is therefore essential that, whenever possible, cytogenetic analysis and/or molecular methods—such as fluorescent *in-situ* hybridization (**FISH**), polymerase chain reaction (**PCR**), or Southern blots—should be part of the diagnostic procedures performed in all these cases.

Acute myeloid leukaemia (Table 3)

AML with cytogenetic abnormalities

There is now good evidence that four types of AML defined by reciprocal chromosomal translocations represent distinct disease entities: these are listed in Table 3.

The translocation t(8;21)(q22;q22) is seen in 10 per cent of cases of AML, most of which have the M2 morphology (myeloblastic with maturation). This form of AML is relatively more common in children than adults. In rare cases the percentage of blasts in the bone marrow is less than 20 per cent, which is currently the recommended threshold for the diagnosis of AML. The immunophenotype of this type of AML is similar to others (Table 2), except for the frequent expression of the B-lineage antigen CD19, in 70 per cent of cases. AML with t(8;21) is associated with a high complete remission rate and a long-term disease-free survival. In a proportion of cases, mainly children, t(8;21) presents as a chloroma (myeloid sarcoma), sometimes with little bone marrow involvement.

The translocation t(15;17)(q22;q21) is consistently associated with acute promyelocytic leukaemia (FAB M3). The chromosome abnormality is seen in both the typical or hypergranular form and the less common microgranular type or M3V. The translocation involves the *PML* and *RARa* (retinoic acid receptor) genes, resulting in a *PML/RARa* fusion gene. M3 represents 7 to 8 per cent of all AMLs. Typical cases can be recognized morphologically: the blasts are bilobed, heavily granular, and show bundles of Auer rods or 'faggots'. MPO and Sudan Black B are strongly positive, myeloid antigens are expressed but HLA-DR and CD34, a feature of myeloblasts, are negative. AML M3 affects young adults who present with low white blood counts (**WBC**) and a bleeding tendency. M3 variant cases have high WBC, and the blasts are deceptively hypogranular and may resemble monocytes but the nucleus is bilobed rather than reniform. Cytochemistry helps to exclude monocytic leukaemia by a strong MPO and weak or negative ANAE reaction. Patients with the rare variant translocations such as t(11;17)(q23;q21) also involving *RARa* usually lack typical M3 morphology and do not respond to treatment with all-*trans* retinoic acid (**ATRA**).

AML with myelomonocytic features and abnormal eosinophils (FAB M4Eo) is associated with abnormalities of chromosome 16, inv16, del(16) or t(16;16) (Table 3). The cytogenetic changes result in the fusion of the *CBFb* and *MYH11* genes and often require analysis by FISH or PCR as they may be missed by conventional karyotype analysis. The granules of the bone marrow eosinophils are often large and react with chloracetate esterase.

Analysis of more than 4000 AML cases entered into MRC AML trials (MRC AML 10, 11, and 12) showed that patients with t(8;21), t(15;17), or inv(16) represent 23 per cent of all cases, and have a significantly better prognosis and treatment outcome than those with normal karyotypes or other abnormalities. In contrast, a particularly unfavourable prognosis is seen in AML with abnormalities of chromosomes 3, 5, and 7.

Translocations involving the *MLL* gene at 11q23 take place with close to 20 other partner chromosomes. These abnormalities are more common in childhood AML and include the t(9;11)(p21;q23) and t(11;19)(q23;p13). The demonstration of *MLL* rearrangement, which is also affected in childhood ALL, is better shown by Southern blot and FISH than by conventional karyotyping. Abnormalities of the *MLL* gene are seen mainly in infants (<1 year) with AML and therapy-related AML (see below). Morphologically, these cases often have monocytic/monoblastic features and a strong ANAE reaction sensitive to inhibition by sodium fluoride.

AML not otherwise categorized

These represent more than two-thirds of cases of AML not included in the above group. Some have distinct chromosome abnormalities, but because they are uncommon they have not yet been considered as specific entities. The WHO classification has retained the FAB categories (M0 to M7) with some additions for this large group of AML (Table 4).

AML M0 represents the most immature form of myeloblastic leukaemia. The blasts are negative with MPO and ANAE cytochemistry, negative with B- and T-lymphoid markers, as distinct from ALL, and positive with one or more of the AML markers (Table 2), including anti-MPO which is more sensitive than the cytochemical method for MPO. The incidence of M0 is around 3 per cent of all AML. There is evidence that this disease has a poor prognosis with a lower remission rate and shorter survival than other forms of AML.

AML M1 is also poorly differentiated; it differs from M0 in the demonstration of myeloid features by the cytochemical reactions with MPO and Sudan Black B. The majority of cases have more than 25 per cent positive blasts and often show Auer rods. It is advisable, in cases with less than 10 per cent MPO-positive blasts, to confirm the diagnosis of AML with myeloid markers (Table 2; Fig. 1).

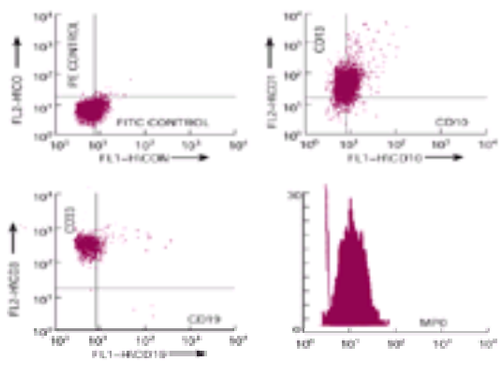

Fig. 1 Flow cytometry analysis of a case of AML using double MAbs conjugated with two different fluorescent dyes (phycoerythrin (**PE**) and fluorescein-isothiocyanate (**FITC**)) shown in the control panel (top left). The blast cells are CD13+ and CD10– (top right), CD33+ and CD19– (bottom left) and anti-MPO+ (bottom right). (See Table 2 and Farahat *et al*. (1994) for technical details).

Some one-third of cases with AML M2 will show t(8;21) and will therefore be included in the group defined by cytogenetic abnormalities. Other cases may have other changes, for example the rare variants of M2 and basophilia involving abnormalities at 12p or the translocation t(6;9)(p23;q34) with the chimeric fusion of the *DEK* and *CAN* genes.

The majority of cases of M3 and M3V recognized morphologically will have the t(15;17) described above. It has been retained here only for cases in which cytogenetic or molecular studies are not possible.

The myelomonocytic (M4) and pure monocytic (M5) leukaemias can be defined by morphology and cytochemistry. The variant M4Eo is associated with inv(16) and is included in the group with cytogenetic abnormalities.

There are two forms of M5 AML identified by morphology. M5a is immature, common in infants, and is recognized by a strong ANAE cytochemical reaction. In M5b the cells are more mature and lysozyme levels are raised. M5 is associated with high WBC, lymphadenopathy, gum hypertrophy, skin deposits, and central nervous system involvement. A rare form of M4 or M5 has heavily granular blasts, prominent haemophagocytosis, and a bleeding diathesis and is characterized by the translocation t(8;16)(p11;p13).

M6 (erythroleukaemia) may show features of trilineage dysplasia. The WHO recommends that the 20 per cent blasts required for the diagnosis of AML could be reached in M6 by excluding erythroblasts in the bone marrow differential count. The blasts in M6 are often myeloblasts. A rare pure form of AML M6 with erythroid precursors could be identified with MAbs against glycophorin A and Hb A.

AML M7 has a fibrotic bone marrow with more than 20 per cent blasts, abnormal megakaryocytes, and circulating blasts shown to be megakaryoblasts by their reactivity with MAbs against platelet glycoproteins, for example CD41/42/61. Bone marrow trephines are essential for diagnosis if there are insufficient blasts to examine with MAbs. Electron microscopy shows that the megakaryoblasts contain peroxidase activity in their nuclear membranes and endoplasmic reticulum. Both M6 and M7 are associated with a poor prognosis.

A subtype of AML seen in children with Down's syndrome has been described as a transient myeloproliferative disorder. If they develop acute leukaemia this is often megakaryoblastic in type.

Other rare forms of AML, acute basophilic leukaemia, acute panmyelosis with myelofibrosis (which may be indistinguishable from AML M7), and myeloid sarcoma (a myeloid tumour developing in an extramedullary site) are now incorporated in the WHO classification (Table 4).

AML with multilineage dysplasia

This form of AML is seen mainly in elderly patients and is defined by the presence of dysplasia in the three bone marrow lineages: granulocytic, erythroid, and megakaryocytic. Chromosome abnormalities, gains or losses, involving chromosomes 7 (–7,7q–) and 5 (–5,5q–), are common. These cases include those evolving from a myelodysplastic syndrome (**MDS**) or arising apparently *de novo*. The key element to distinguish from MDS is the presence of more than 20 per cent blasts in the bone marrow.

Therapy-related AML and MDS

This new category recognized in the WHO classification results from the increased number of patients developing AML and MDS following therapy for other malignancies. Two major types are recognized:

1. *Related to therapy with alkylating agents*. These cases occur late, 5 to 6 years following exposure. A phase of MDS often precedes AML. Morphologically, these cases have trilineage dysplasia and are difficult to classify in the types listed in Table 4. Abnormalities of chromosomes 5 and 7 are also common.
2. *Related to therapy with topoisomerase II inhibitors*. This form of AML is seen in patients treated with drugs targeting topoisomerase II, such as etoposide and teniposide, and also with anthracyclines. It has a short period of latency, usually from 12 to 50 months (median 33 months) and is rarely associated with MDS changes. Chromosome abnormalities often involve the *MLL* gene, e.g. t(9;11) and t(11;19), and also t(8;21), inv(16), and t(6;9).

Acute lymphoblastic leukaemia

ALL represents the clonal proliferation of immature lymphoid precursors. The characteristic cell is the lymphoblast, which has a high nuclear to cytoplasmic ratio and usually lacks cytoplasmic granules. Staining reactions with MPO, Sudan Black B, and ANAE are negative. Evidence of lymphoid differentiation towards a B or T lineage is provided by immunological markers (Table 2), including a positive terminal deoxynucleotidyl transferase (**TdT**) response.

In the morphological groups described by the FAB group, L1 is seen in 80 per cent of childhood cases, L2 in 20 per cent, and L3 (or Burkitt type) in 1 to 3 per cent. However, only the latter—which corresponds to membrane Ig-positive B-lineage ALL and has the same chromosome translocation, t(8;14), as Burkitt lymphoma—remains as a morphological/pathological entity described as Burkitt leukaemia/lymphoma (Table 5). The L1 and L2 types are no longer taken into consideration for classification purposes.

The current classification of ALL takes into account the cell lineage, B or T, which is based on the sequential appearance of B- and T-cell antigens during differentiation (Table 2), and distinct and consistent abnormalities that define discrete forms of the disease (Table 5). The immunophenotype of ALL correlates at the DNA level with the rearrangement of the immunoglobulin heavy-chain genes and T-cell receptor (**TCR**) genes in the B and T lineages, respectively. This type of analysis is not necessary for diagnosis or classification purposes, but it is important for monitoring minimal residual disease. Molecular techniques, such as PCR and FISH analysis, on the other hand, are also becoming more informative for the detection of rearranged genes and chimeric mRNA involved in the specific chromosome abnormalities of ALL.

Precursor B-lineage ALL

This is more frequent in children and the elderly. In infants with precursor B-ALL, there is a strong association with translocations of the *MLL* gene, which confers a poor prognosis.

An important marker of the B lineage is the common-ALL antigen recognized by MAbs of the CD10 cluster. CD10 is expressed in the blasts of 75 per cent of childhood ALL cases. Less mature lymphoblasts (early or pro-B ALL) do not express CD10, and this is seen in 10 per cent of childhood cases and in 30 per cent of adult patients. Pre-B ALL blasts express cytoplasmic μ chains (without light chains).

The various forms of B-lineage ALL have distinct prognostic features that are best characterized by the associated chromosome abnormalities (Table 5). The best two prognostic groups of precursor B-ALL defined by cytogenetic/molecular rearrangements are t(12;21)(q21;q22) and hyperdiploidy (51–65 chromosomes). The t(12;21) translocation is the most common, found in 25 per cent of cases, but it is difficult to detect by conventional cytogenetics and needs FISH analysis or Southern blotting to identify rearrangements of the *TEL* gene. Recent data show that at 5 years' follow-up the event-free survival of children with *TEL* rearrangements is 91 per cent. Similarly, 85 per cent of children with ALL and 51 to 65 chromosomes can be cured with current protocols. One of the possible explanations for this high cure rate is the propensity of lymphoblasts from hyperdiploid cases to undergo apoptosis.

The translocation t(1;19) is associated with the expression of cytoplasmic μ chain and immunologically it corresponds to a pre-B precursor. Although formerly associated with a poor prognosis, this has improved with newer modalities.

Poor prognosis in patients with ALL is associated with two molecular events: *MLL* rearrangement and the *BCR/ABL* fusion gene (Table 5). The *MLL* gene is rearranged in several translocations. The most common in infants is t(4;11)(q21;q23); this event is now known to occur before birth. The t(4;11) translocation is seen in 4 to 8 per cent of adult cases of ALL, 3 to 5 per cent of childhood ALL, and in up to 70 per cent of infants with ALL.

The t(9;22), which results in a short chromosome 22, the Philadelphia (**Ph**) chromosome, and the associated *BCR/ABL* fusion gene, increases in frequency with age. In childhood ALL up to 5 per cent of cases have been reported, but 25 to 30 per cent of adult patients with ALL have the translocation. There are important subtle differences between children and adults. In the latter, 50 per cent of the *BCR/ABL* fusion results in a p210 chimeric protein (as in all cases of Ph-positive CML), whilst in childhood ALL the common product is p190. In addition to the higher incidence of t(9;22) in adult ALL, their poor prognosis is compounded by the rarity of hyperdiploidy and of the t(12;21) translocation.

Precursor T-lineage ALL

T-lymphoblasts are defined by several T-cell markers (Table 2). In about one-third of cases, chromosome translocations involving the *TCRa/d* loci at 14q11 and the *TCRb/g* loci at 7q34. Examples are t(1;14)(p32;q11) involving the *TAL1* gene, t(10;14)(q24;q11) with the *HOX11* gene, and t(11;14)(p13;q11) involving the *RBTN2* gene. In 25 per cent of cases of T-ALL the *TAL1* locus (1p32) is dysregulated by submicroscopic deletion.

Precursor T-ALL comprises 15 per cent of childhood ALL and is more common in adolescents and males. In adult ALL it comprises 25 per cent of cases. Disease characteristics are a high WBC, large mediastinal mass, and less involvement of the bone marrow than precursor B cases.

Burkitt leukaemia/lymphoma

This leukaemia represents the leukaemic presentation of Burkitt lymphoma and is defined by evidence of membrane immunoglobulin (heavy and light chains); it is associated with L3 morphology and the translocation t(8;14)(q24;q32). In adults it may also represent the transformation of a pre-existing follicular lymphoma; in such cases t(8;14) coexists with t(14;18), but this is rare. The lymphoblasts have a deep basophilic cytoplasm and prominent nucleolus. L3 blasts are TdT- and CD34-negative and CD19-, CD20-, and CD22-positive; CD10 may be expressed, although it is more frequently absent.

Patients present with bulky disease and, biochemically, a high lactate dehydrogenase (**LDH**) level. Prognosis has now improved with very intensive protocols of short duration.

Biphenotypic acute leukaemia or acute leukaemia of ambiguous lineage

The wider use of MAbs has brought to light the existence of cases in which markers of different cell lineages (usually B and myeloid) are coexpressed on the same blast cells. Many of the markers used for defining the immunophenotype of acute leukaemia are not always lineage-specific. Hence, it is necessary to use restrictive criteria to define as 'biphenotypic' those cases that coexpress two markers of which at least one is lineage specific, for example: CD22, CD79a, and μ chain for the B lineage; CD3 for the T lineage; and MPO or CD117 for the myeloid lineage.

Some cases of biphenotypic leukaemia have all the features of ALL but coexpress two or more myeloid markers. Others present as typical AML and, in addition to myeloid antigens (Table 2), they express TdT and two or more B (or rarely T) antigens. Biphenotypic leukaemias may represent up to 5 per cent of all acute leukaemias and frequently show rearrangement of immunoglobulin and/or *TCR* genes, even in cases presenting as AML. Because of the distinct biological and molecular changes and current poor prognosis associated with these cases, it is relevant to recognize biphenotypic leukaemia as a distinct type using well-defined criteria. Cytogenetic changes include t(4;11) and other 11q23 abnormalities as well as t(9;22), which may explain the overall poor prognosis.

Chronic leukaemias

These are malignancies in which mature leucocytes are the predominant cells. As in the acute leukaemias, there are two main groups: myeloid and lymphoid, with two representative disorders, chronic myelogenous (**CML**) and chronic lymphocytic (**CLL**) leukaemia. The latter, as well as the less common forms of leukaemias of mature B and T lymphocytes, are described in Chapter 22.3.5. The chronic myeloid leukaemias reflect granulocytic or monocytic differentiation, or a combination of both: myelomonocytic (Table 6).

CML (Ph-positive)

CML (or CGL) has a distinct chromosome marker, the Ph chromosome, resulting from the reciprocal translocation t(9;22), found in 95 per cent of cases. Cases with similar haematological features, but which are cytogenetically Ph-negative, often have the same molecular rearrangement that results from the juxtaposition of the *BCR* and *ABL* genes to form a hybrid *BCR/ABL* gene. Understanding of the molecular biology of CML leads to significant advances in treatment (see Further reading

list).

The diagnosis of Ph-positive CML can be established by the morphological appearance of peripheral blood and bone marrow films, and confirmed by cytogenetic and/or molecular analysis. The leucocyte differential count in CML shows the full spectrum of granulocytic cells but with a predominance of myelocytes (about 30 per cent) and mature neutrophils (about 50 per cent), as well as almost invariably basophilia (approximately 5 per cent) and frequent eosinophilia; the percentage of monocytes is low, usually less than 3 per cent. Blasts represent 1 or 2 per cent of the circulating cells unless the disease is in accelerated phase or in transformation; myelodysplastic changes are minimal. The bone marrow aspirate is hypercellular with granulocytic hyperplasia and numerous megakaryocytes, and is less useful than the peripheral blood for a differential diagnosis between CGL and the other chronic myeloid leukaemias. The myeloid:erythroid ratio is greater than 10:1 with few erythroblasts. The bone marrow trephine is necessary to assess the degree of fibrosis and, occasionally, to distinguish from idiopathic myelofibrosis.

Atypical chronic myeloid leukaemia (Ph-negative)

Patients with high leucocyte counts who are Ph chromosome-negative (and *BCR/ABL*-negative) represent a slightly heterogeneous group. Most cases have atypical morphological features compared with those with Ph-positive CML: namely, slight monocytosis, absence of basophilia and eosinophilia, granulocytic dysplasia (Pelger and hypogranular neutrophils), and 2 to 3 per cent circulating blasts.

Chronic myelomonocytic leukaemia (CMML)

CMML has features of MDS and myeloproliferative disorders. The two main differences with atypical CML and Ph-positive CML are the proportion of monocytes, which ranges from 25 to 50 per cent in CMML, and the lower percentage of immature granulocytes (5–10 per cent) in the bone marrow. The erythroid cells in the bone marrow are also more prominent in CMML, with a lower myeloid:erythroid ratio than in the other conditions. CMML cases are always Ph-negative and *BCR/ABL*-negative, and show moderate to high levels of serum and urinary lysozyme and myelodysplastic changes in the bone marrow. The WHO includes CML in an MDS/myeloproliferative grouping and no longer purely as MDS.

Rare forms of chronic myeloid leukaemia

Chronic neutrophilia

This occurs in adults over 50 years of age. Haemoglobin and platelet counts are normal. The blood film shows predominantly neutrophils without immature forms. The neutrophil alkaline phosphatase score is high, in contrast with the very low levels in classic Ph-positive CML. Most cases were thought to be Ph-negative. Recent evidence suggests that some cases have the t(9;22), but, in contrast to CML, the *BCR/ABL* rearrangement results in a p230, rather than a p210, protein product.

Juvenile CMML

Juvenile chronic myelomonocytic leukaemia represents 2 per cent of the childhood leukaemias. Patients are under 5 years of age and have systemic symptoms. In contrast to CML, there is monocytosis without basophilia or eosinophilia. Characteristically, the levels of fetal haemoglobin are high. Cytogenetic analysis is important to distinguish juvenile CMML from Ph-positive CML in childhood and from the myelodysplastic syndrome with monosomy 7 seen in young children.

Chronic eosinophilic leukaemia (CES)

CES is now included together with the hypereosinophilic syndrome because a clear distinction between these conditions is currently not possible. A comprehensive review of the topic has recently been published (see Further reading list).

Further reading

Bain BJ (2000). Hypereosinophilia. *Current Opinion in Hematology* **7**, 21–5.

Bennett JM, *et al*. (1976). Proposals for the classification of the acute leukaemias. *British Journal of Haematology* **33**, 451–8.

Bennett JM, *et al*. (1985). Criteria for the diagnosis of acute leukemia of megakaryocyte lineage (M7). *Annals of Internal Medicine* **103**, 460–2.

Bennett JM, *et al*. (1991). Proposal for the recognition of minimally differentiated acute myeloid leukaemia (AML-M0). *British Journal of Haematology* **78**, 325–9.

Bennett JM, *et al*. (1994). The chronic myeloid leukaemias: guidelines for distinguishing chronic granulocytic, atypical chronic myeloid, and chronic myelomonocytic leukaemia. Proposals by the French–American–British Cooperative Leukaemia Group. *British Journal of Haematology* **87**, 746–54.

Farahat N, *et al.* (1994). Demonstration of cytoplasmic and nuclear antigens in acute leukaemia using flow cytometry. *Journal of Clinical Pathology* **47**, 843–9.

Harris NL, *et al*. (1999). World Health Organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Committee meeting—Airlie House, Virginia, November 1997. *Journal of Clinical Oncology* **17**, 3835–49.

Jaffe ES, *et al*. (2001). *World Health Organization classification of tumours – pathology and genetics of tumours of haematopoietic and lymphoid tissues*. IARC Press, Dijon.

Löwenberg B, Downing JR, Burnett A (1999). Acute myeloid leukemia. *New England Journal of Medicine* **341**, 1051–62.

Melo JV (1996). The diversity of BCR–ABL fusion proteins and their relationship to leukemia phenotype. *Blood* **88**, 2375–84.

Rowley JD (2000). Molecular genetics in acute leukemia. *Leukemia* **14**, 513–17.

Savage DG, Antman KH (2002). Imatinib mesylate–a new oral targeted therapy. *New England Journal of Medicine*, **346**, 683–93.

Wheatley K, *et al*. (1999). A simple, robust, validated and highly predictive index for the determination of risk-directed therapy in acute myeloid leukaemia derived from the MRC AML 10 trial. United Kingdom Medical Research Council's Adult and Childhood Leukaemia Working Parties. *British Journal of Haematology* **107**, 69–79.

22.3.3 Acute lymphoblastic leukaemia

Philip J. Burke

Acute lymphoblastic leukaemia (ALL)

Some 40 years ago prednisone and vincristine were found to be effective in reducing visible leukaemia in children with ALL. The concepts of remission, consolidation, and maintenance therapies for all leukaemias evolved from trials in this disease, and until now have been the mainstay of tumour control strategies. Although the outcome in children has been excellent, outcomes in adults have so far been discouraging. However, with improved support, thereby permitting more intensive treatment modalities, the survival rate in patients at obvious poor risk has improved. Unfortunately, a significant increase in the cure rate without untoward early and late toxicity is unlikely using currently available drugs.

It is unclear how best to increase the cure rate in older patients. Real improvement may lie in applied molecular biology aimed at specific genetic targets. Subset identification will help to resolve the question of risk versus benefit in individual patients. The next step will be to obtain evidence supporting biological intervention with specific novel agents. The most appropriate candidates for study are patients with minimal leukaemia after initial tumour reduction and bone marrow homeostasis.

Diagnosis

Classically, acute lymphoblastic leukaemia has been categorized by morphology. There are three groups in the French–American–British (**FAB**) carcinoma staging classification, with L1 being defined by small monomorphic cells, L2 by large heterogeneous cells, and L3 by a Burkitt's cell-type with vacuoles. These distinctions are not as definitive as for acute myeloblastic leukaemia (**AML**), but flow cytometry has relieved some of the confusion with identification by phenotype. A prognosis in ALL can now be more accurately assigned with the addition of both immunophenotyping and cytogenetics (Table 1). Some 20 per cent of ALL are T cell, 75 per cent are precursor B cell, and 5 per cent are mature B cell in origin. In adults, 35 per cent of cases express both lymphoid and myeloid antigens. These more immature subtypes, early-pre-B-ALL and pre-T-ALL occur more frequently in adults than in children. Classification by immunological subtype is reviewed in Chapter 22.3.2. In brief, early pre-B-ALL expressing HLA-DR, terminal deoxynucleotidyl transferase (**TdT**), and CD19 occur in 10 per cent of adults and 5 per cent of children. Common ALL, found in 50 per cent of adults, is characterized by CD10 and CD19. Some 15 per cent of ALL are pre-B with expression of a cytoplasmic immunoglobulin absent in common ALL.

The T-cell phenotype is present in 25 per cent of adult ALL, with CD7, CD2, and/or CD1. In most, the T-cell receptor gene is rearranged. Bilineage or biphenotypic hybrid leukaemias which express both lymphoid and myeloid antigens account for 35 per cent of adult cases.

Cytogenetic analyses of lymphoblasts reveal clonal chromosomal aberrations in 50 to 75 per cent of ALL. Most are structural abnormalities with a balanced translocation. Other more random aberrations are classed by chromosomal number—hypo- or hyperploidy. This syndrome identification confers prognosis for outcome with standard therapy.

Pathophysiology

Clinical manifestations

The pathophysiology of ALL is the consequence of bone marrow failure caused by tumour-related suppression of normal haemopoiesis and the clinical expression of the malignant clone. The growth of lymphocytic precursors arrested at an immature stage of differentiation suppresses production of the normal bone marrow elements, resulting in anaemia, infection, and bleeding. Circulating leukaemic cells infiltrate the central nervous system, liver, spleen, testes, ovary, lymph nodes, skin, and gastrointestinal tract. Diffuse lymphadenopathy and hepatosplenomegaly occur in half the cases. Varied degrees of failure of each organ, and the metabolic effects of increased cell turnover, contribute to the overall symptom complex at presentation. Centripetal joint swelling and bone pain with rheumatic symptoms in the adult contrast with the distal bone involvement seen in children. Punched out bony lesions may be present on the radiograph.

In children, the history of a recent severe upper respiratory infection is frequently elicited. Although evidence for a viral aetiology is sketchy, immune stimulation may lead to proliferation of B-cell precursors. There are models in the lymphomas with similar theoretical pathogenesis.

Failure to obtain disease remission late in therapy involves either failure of support or tumour resistance, while early mortality relates to the pathophysiology of the tumour and its immediate management.

Infection

As seen with AML, clinical signs and symptoms are those of bone marrow failure and tumour mass, but because of a lack of normal lymphocytes there is an associated immunosuppression in addition to neutropenia. This loss of immune surveillance places the patient at risk for infections unique to the immunocompromised host. In addition to the infectious agents encountered in the myeloid malignancies with granulocytopenia, nosocomial and parasitic infections, such as *Pneumocystis carinii* and atypical fungal organisms, must be suspected with each new fever. Since therapy is prolonged in ALL, consistent monitoring for pathogenic organisms is necessary even during remission (Chapter 22.3.4).

Lymphoblastosis

The immediate challenge at presentation is the diagnosis of the leukaemia presenting with a high white blood cell (**WBC**) count. Acute causes of an increased WBC count which may be mistaken for lymphoid leukaemia include the FAB classifications M0 and M7 of acute myelocytic leukaemia, chronic myelocytic leukaemia, and chronic lymphocytic leukaemia.

The WBC can range from zero to more than 100×10^9/l, with a count of less than 5×10^9/l in 25 per cent of patients. The need for immediate intervention is less urgent with lymphoblasts because they are relatively small and deformable compared with young myeloid cells, lessening the probability of hyperviscosity. Their

rupture with chemotherapy, however, releases high levels of phosphates and urates which can cause tumour-lysis syndrome unless precautions are taken. An approach to hydration, leukapheresis, and possible early dialysis is reviewed in Chapter 22.3.4. If the diagnosis remains in doubt after slide review, flow cytometry will quickly distinguish between myeloid and lymphoid cells. The untoward effects of inappropriate chemotherapy prior to leukapheresis and fluid balance must be avoided in ALL.

Occasionally (less than 5 per cent of patients) the differential count shows reactive myelopoiesis with pseudo Pelger–Huët cells, but without circulating lymphoblasts. Normal lymphocytes may be absent. A bone marrow aspirate will provide the diagnosis in these cases of subleukaemic leukaemia.

Prognostic factors

Although statistics indicate considerable improvement in the treatment of adult ALL over the past decade, the majority of patients relapse and die of their disease. Certain subsets of patients are predicted to respond poorly with current therapies. For example, those with leukaemia characterized by the Philadelphia chromosome (Ph_1) or its molecular equivalent, the fusion gene *BCR–ABL*, have essentially no chance of cure without allogeneic bone marrow transplantation. Even then, only 40 per cent of patients who achieve remission will be cured.

Other adverse prognostic factors include over 60 years of age, WBC in excess of 30×10^9/l, absence of a mediastinal mass, L3 morphology, myeloid phenotype, and lack of expression of CD10. Combinations of these markers decrease the predicted survival. In one trial, the presence of three high-risk factors in B-lineage ALL decreased the 1-year survival rate from 75 per cent (one risk factor present) to 25 per cent. There were no long-term survivors among patients with four risk factors.

The strongest variable, genetic identification, now provides a reliable basis for assigning risk and selecting therapies (see cytogenetic and phenotypic classification, Chapter 22.3.1 and Chapter 22.3.2). Patients known to do poorly with childhood-type strategies, usually those with the chromosomal anomalies t(9;22), t(4;11), t(8;14), or 14q 11–13 (T-ALL), now respond to intensive initial treatment schedules.

Management

Therapy

The highly successful treatment of ALL in young children combines numerous drugs known to be active against malignant lymphoblasts. Therapy consists of a three-phase treatment induction, CNS prophylaxis, and maintenance. Induction therapy with prednisone, vincristine, L-aspara-ginase, and daunorubicin produces remission rates of 85 per cent. Maintenance therapy with 6-mercaptopurine, methotrexate, cyclophosphamide, and prednisone is given in a cyclical fashion for 2 or more years.

In contrast, complete remission (**CR**) rates in adults with ALL range from 65 per cent to 85 per cent, the time to remission is longer, the relapse rate is higher, and cure rates are between 20 per cent and 40 per cent. Multiple factors related to both the biology of leukaemic cells and ability of the host to tolerate treatment contribute to the inferior prognosis in adults as compared with childhood ALL.

A variety of strategies are designed to decrease the relapse rates in adults. Early intensive therapy to rapidly destroy more leukaemic cells that have not yet developed drug resistance is a logical approach. Increasing induction intensity with sequential high-dose cytotoxic agents results in an improved long-term outcome in poor-risk groups. In recent studies, the addition of early intensification therapy, particularly a timed sequence of daunorubicin and cytarabine, effectively prolongs remissions and prevents relapse. This approach in children results in a CR rate of 95 per cent, with 70 per cent of children remaining in CR beyond 5 years.

Intense therapy derived from AML models, given early in remission (consolidation) using drugs not previously employed and at high doses, effectively destroys residual leukaemic cells. Cyclophosphamide and cytarabine are examples of AML drugs of obvious value which are now better tolerated, and with a greater therapeutic advantage provided by effective support measures. The role of colony-stimulating factors remains unclear.

The present investigative approach focuses on intensification of therapy in all subgroups with ALL. Repeatedly given high-dose drugs rapidly reduce tumour mass, reduce the possibility of drug-resistance, and invade sanctuary sites to eliminate sequestered cells. With these more aggressive therapies, CNS relapse in adults is less than 5 per cent.

However, in contrast to the experience with two short, intensive courses in AML, long-term success in ALL requires prolonged chemotherapy with multiple courses similar to present lymphoma regimens. Long-term maintenance therapy remains standard in most group trials but long-term value is not proven (Table 2).

The model of high-dose chemotherapy is intense induction followed by bone marrow transplantation (**BMT**). The optimal indications for BMT in adult ALL remain to be determined. While a definitive role of allogeneic BMT during first remission in patients with high-risk features has been established, significant controversy exists regarding the indications for BMT in patients with standard-risk adult ALL. Autologous BMT has a failure rate similar to that of high-dose chemotherapy.

High-dose treatment in poor-risk patients with ALL

Below are examples of cytogenetically poor-risk groups of ALL which respond to aggressive therapies:

1. ALL-L3 (B-ALL) treated with intensive therapies with high doses of five drugs (vincristine, doxorubicin, cytarabine, methotrexate, cyclophosphamide) given in short repeated cycles now achieve CR rates of 70 per cent with 50 per cent long-term survival rates in patients with this rare (5 per cent) leukaemia. CNS prophylaxis with high-dose systemic therapy or intrathecal drugs is essential.
2. t(4;11) ALL with a pro-B-cell phenotype occurs in 5 per cent of adults and infants. Intensive therapy with high-dose cytarabine and mitoxantrone has improved the dismal outcome, with a predicted 50 per cent disease-free survival. Bone marrow transplantation is a reasonable alternative therapy.
3. t(9;22) ALL with *BCR–ABL* fusion gene occurs in 30 per cent of adults; 50 per cent of adults with B-lineage surface antigens have this fusion protein p190, in contrast to the p210 subtype of CML. Although remission rates of 75 per cent can be achieved with intensive therapy, even with BMT the relapse rate is between 40 and 80 per cent.
4. t(1;19), 11q23, t(12;21) are specific syndromes which have also demonstrated a significantly improved outcome with aggressive treatment.

Allogeneic bone marrow transplantation in poor-risk patients is a reasonable choice, since the median age in adults is 33 years. Relapse after BMT is 30 per cent, with a 15 to 20 per cent mortality from immediate and late toxicities.

With present support and intensive treatment, most patients can achieve a stable and durable remission. Most will relapse with resistant disease. A continued search for new approaches to the treatment of such patients with these non-random subsets of ALL is critical.

Proto-oncogene leukaemogenesis (seeTable 1)

Most of the lymphoid malignancies appear to be related to oncogene rather than tumour-suppressor malfunction. Oncogenes direct the synthesis of products that contribute to the malignant transformation of a cell. When a genetic alteration activates a particular proto-oncogene, it is implicated in malignant transformation. The mechanisms may also involve either a point mutation or insertional mutagenesis. The most mechanistically defined are those syndromes related to visible translocation of proto-oncogenes to active promoter sites, a coupling that results in production of a unique protein or overproduction of the normal product. The diseases may also be associated with a single-point mutation in cells committed to specific lineage. A number of the leukaemias have been fully characterized. Minimal numbers of leukaemia cells are quantifiable by detection of the residual fusion gene with the polymerase chain reaction.

Measuring cause and change in these specific and cytotoxic drug-responsive leukaemias may determine and remedy the oncogenic mechanism, and establish models for treatment strategies in similar, but less definable, tumours.

Regulation of lymphoid-cell growth

Biological modifiers of lymphoid-cell proliferation are being sought for therapy. These are best understood by reviewing the normal regulation of cell growth and survival. The proliferation and maturation of lymphocytes is ultimately determined by peptide growth factors that bind and activate specific receptors. These activated receptors propagate the signal through a series of protein interactions to the cell nucleus where the signals are converted into predetermined responses. Proto-oncogenes encode growth factors and growth-factor receptors, transcription factors, and cell-cycle proteins that determine the level of gene expression. Abnormalities of these functions are produced by genetic change in haematological malignancies.

The rate of proliferation versus its maturation and normal programmed cell death by growth factors is determined by the cell. In the nucleus, low levels of MYC cause apoptosis (cell death) in response to cytokines. MYC is an intracellular protein ordinarily sequestered in the cytoplasm by the retinoblastoma suppressor protein (**Rb**). MYC protein, functioning in the nucleus as a heterodimer with another molecule, MAX, regulates transcription and therefore gene expression and cell survival. BCL-2 allows MYC to drive the cells to proliferation by blocking the pathway to cell death, determining the ultimate expansion of the cell population. This activity of BCL-2 is balanced by an opposing protein, BAX, whose activity as a dominant regulator of BCL-2 is increased by expression of P53. P53 overcomes the BCL-2 block of cell death and inhibits cell-cycle traverse of mutated or damaged DNA. It blocks cyclin-dependent phosphorylation, forcing DNA repair, or apoptosis if repair is faulty. For example, elimination of excess lymphoid cells activated by antigenic stimuli via the programmed cell-death mechanism is a physiological event involved in normal immune responses.

The initial aim of intervention in the lymphoid tumours is to downregulate the overexpressed gene *BCL2*, and to force apoptosis in patients with leukaemias defined by this *BCL2* abnormality. In those without specific abnormalities in gene translocation and protein production, the transfer of lymphocytes to upregulate cytoplasmic antigens for tumour-specific attack is a testable thesis.

Novel agents

A few of many targets for investigation in patients with senescent leukaemia in remission are listed below. See Table 5 of Chapter 22.3.4 for a wider list of potential agents.

Drug resistance

Although most leukaemias are responsive to initial therapies, most patients relapse and die of drug-resistant disease. The refractory clone, either acquired or selected with chemotherapy, contains the multidrug resistance gene and its encoded 170-kDa P-glycoprotein. This mechanism may also protect normal bone marrow stem cells from cytotoxicity. Drugs that modulate the effect of this gene—the rapid pumping out of drug from the cell—are calcium-channel blocking agents such as verapamil and ciclosporin-A.

Telomerase

Telomeres are structures at the ends of chromosome which in normal cells progressively shorten with each cell division. Without telomerase action terminally short telomeres cause cell-growth arrest. Suppression of telomerase is a tumour-suppressor and ageing mechanism. In contrast, most cancer cells activate telomerase, resulting in stable telomeres and immortalization. Normal stem cells may express telomerase in the adult, allowing perpetuation, while tumours arising from stem cells may have a similar prolongation of survival. Telomerase inhibition is a therapeutic target to force cell senescence. Since telomerase levels reflect the persistent growth of AML, measurable suppression of telomerase activity in patients with minimal residual disease in remission may force tumour senescence and cell death.

Methylation

An imbalance of DNA methylation, involving widespread hypomethylation, regional hypermethylation, and increased cellular capacity for methylation is characteristic of human neoplasia. Beginning in preneoplastic cells, it becomes extensive in subsequent stages of tumour progression. This aberrant methylation, particularly of cytosine, may mark abnormalities of chromatin reorganization and mediate progressive losses of gene expression associated with tumour development.

Abnormal methylation sites have been detected in both myeloid and lymphoid malignancies. Drugs presently in trial to prevent leukaemic transformation or suppress relapse are the butyrates and cytidine analogues.

Angiogenesis

Leukaemia-cell growth requires the aggressive infiltration of capillaries into nests of tumour in excess of that seen in normal bone marrow. Relative selectivity of this neovascularization results in selective tumour regression on administration of antiangiogenic agents. These include angiostatin, endostatin, vasostatin, and other endogenous inhibitors of angiogenesis that block the effects of vascular endothelial cell-growth factor (**VEGF**). Other drugs act to block oncoprotein function by inhibiting signal transduction, for example Ras farnesyltransferase inhibitors. Such factors will prevent blood supply to new tumour growth when given at the time of maximal leukaemia-cell reduction.

Monoclonal antibodies

Cell-surface directed approaches are now in large clinical trials with encouraging responses. Constructs aimed directly at receptors include monoclonal antibodies (MAbs) specific for the cell-surface receptors CD20, CD19, and CD33 linked with a radionuclide, an endotoxin, or an antibiotic. All have achieved early success in relatively resistant leukaemias. Unequal expression of surface receptors in the leukaemic clone may require upregulation for sensitivity.

Monoclonal antibodies have achieved utility in the treatment on lymphoma and hold promise in ALL.

Further reading

Cassileth PA, *et al*. (1992). Adult acute lymphocytic leukemia: the Eastern Cooperative Oncology Group experience. *Leukemia* **6**,178–81.

Chao NJ, *et al*. (1991). Allogeneic bone marrow transplantation for high-risk acute lymphoblastic leukemia during first complete remission. *Blood* **78**,1923–7.

Copelan EA, *et al*. (1995). The biology and treatment of acute lymphoblastic leukemia in adults. *Blood* **85**,1151–68.

Hoelzer D, *et al*. (1988). Prognostic factors in a multicenter study for treatment of acute lymphoblastic leukemia in adults. *Blood* **71**,123–31.

Hoelzer DF (1993). Therapy of the newly diagnosed adult with acute lymphoblastic leukemia. *Hematology/Oncology Clinics of North America* **7**,139–60.

Kantarjian HM (1994). Adult acute lymphocytic leukemia: critical review of current knowledge. *American Journal of Medicine* **97**,176–84.

Larson RA, *et al*. (1995). A five-drug remission induction regimen with intensive consolidation for adults with acute lymphoblastic leukemia: cancer and leukemia group B study 8811. *Blood* **85**, 2025–37.

Linker Ca, *et al*. (1991). Treatment of adult acute lymphoblastic leukemia with intensive cyclical chemotherapy: a follow-up report. *Blood* **78**, 2814–22.

Mandelli F (1992). GIMEMA ALL 0288: a multicentric study on adult acute lymphoblastic leukemia. Preliminary results. *Leukemia* **6**,182–5.

Rivera GK (1993). Treatment of acute lymphoblastic leukemia. 30 years' experience at St. Jude Children's Research Hospital. *New England Journal of Medicine* **329**,1289–95. [See comments.]

Rohatiner AZ, *et al*. (1990). High dose cytosine arabinoside in the initial treatment of adults with acute lymphoblastic leukemia. *British Journal of Cancer* **62**, 454–8.

Woods WG, *et al*. (1996). Timed-sequential induction therapy improves postremission outcome in acute myeloid leukemia: a report from the Children's Cancer Group. *Blood* **87**, 4979–89.

Yeager AM, *et al*. (1986). Autologous bone marrow transplantation in patients with acute nonlymphocytic leukemia, using *ex vivo* marrow treatment with 4-hydroperoxycyclophosphamide. *New England Journal of Medicine* **315**,141–7.

22.3.4 Acute myeloblastic leukaemia

Philip J. Burke

Introduction

Efforts in leukaemia research over the last 30 years or so have been rewarded with much success. Moreover, the principles of clinical research and their application to patient care are now defined.

With the availability of effective extrinsic haemopoietic support, leukaemia-induced bone marrow failure no longer restricts the intensity of therapy necessary to achieve maximal tumour kill. Pharmacologically deter-mined drugs and schedules combined with rational concepts of maximal therapy have improved the survival rates of patients with leukaemia. The identification of tools and specific targets heralds a new era of therapy in patients with small amounts of residual leukaemia.

Many new technologies—DNA hybridization, purified specific probes, monoclonal antibodies marking lineage specificity, genetic array, and rapid measures of hybrid proteins—permit the genetic classification and identification of the origin and activity of a tumour. For some types of leukaemias, a combination of these methods provides an accurate prognosis for tumour eradication with the utilization of enhanced chemotherapy regimens and the application of new biological modalities. Single-site genetic mutations, translocations or deletions on chromosomes, loss of suppression or overexpression of genes, and transcription of an abnormal product that induces a malignant phenotype are all obvious targets for unique approaches to tumour control. Future success depends on understanding genetic multistep pathogenesis and its control, ultimately with specific interdiction of the premalignant clone.

Pathogenesis of myeloid leukaemia

There is evidence for a multistep pathogenesis of solid tumours. Similarly, many of the acute myeloid leukaemias are the final stages of clonal homeopathies initially transformed by exposure to ionizing radiation, certain chemicals, some chemotherapeutic drugs—especially alkylating agents and topoisomerase inhibitors—or by progression (for example, in certain genetic disorders).

Leukaemogenesis proceeds after the deleterious effect of a chemical or environmental agent on chromosomes. In some cases amplification or modification of the gene product then perpetuates the process. These events can be subclinical and of long duration in the prodromal stage. Progression occurs, for instance, when DNA damage is not repaired in proliferating cells which then escape programmed cell death.

The prima-facie evidence of induced genetic instability comes from the leukaemias that developed in people exposed to high-dose radiation from the Hiroshima and Nagasaki atomic bombs. A similar leukaemogenic effect of alkylating agents was manifested in 10 per cent of patients treated with **MOPP** (mechlorethamine, Oncovin (vincristine), procarbazine, prednisone) for Hodgkin's disease. These therapy-related leukaemias (**tAML**; secondary AML, **sAML**) usually have non-random genetic deletions of -5, -7, 5q, and/or 7q, while those related to topoisomerase II inhibitors (etoposides and anthracyclines) have consistent abnormalities of chromosome 11 (11q,23), and FAB 4 to 5 morphology of the French–American–British carcinoma staging classification. Many evolve from trilineage bone marrow failure (myelodysplastic syndrome, **MDS**), suggesting a genetic instability of multistep origin. These syndromes of long duration culminate in an aggressive phase with a clonal cohort that escaped DNA repair. The increasing numbers of patients with MDS reflect the cumulative effects of a series of sporadic incidents having their ultimate expression in an ageing population.

Some 30 per cent of patients have leukaemic syndromes associated with a non-random loss or gain of DNA on chromosomes 5, 7, and 8. These are stem-cell derived, trilineage tumours classified by FAB as either M_6, M_7, transformed MDS, or secondary acute myelocytic leukaemia (**sAML**). As in colon cancer, sequential genetic alterations ultimately result in metastatic malignancy. Secondary leukaemia may take years to evolve (Fig. 1). This class of haematological malignancy—the loss of a tumour-suppressor gene(s)—only transiently responds to current therapeutic modalities. However, remissions of predictable duration are achievable with intensive therapy in 40 per cent of such patients. Once stable, these genetic carcinoma-like tumours are targets for specific novel agents.

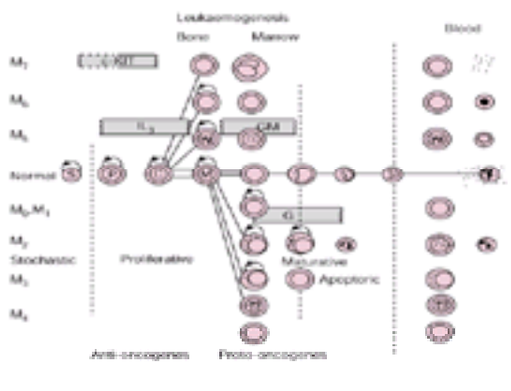

Fig. 1 Myeloid leukaemogenesis. Normal proliferation and maturation of granulopoiesis is contrasted to those leukaemias of proto-oncogene and antioncogene derivation. This is a conceptual model loosely relating FAB with a cytogenetic classification. The activities of growth factors relative to proliferation, maturation, and apoptosis are shown. The figure depicts the progression of leukaemia according to the FAB classification and ligand stimulation. During the stochastic kinetic period the stem cell reproduces itself, with some cells distributed to the committed pool in a random manner. While in this proliferative pool they are influenced by normal regulatory factors. The first legend, c-*kit*, affects the stem cell directly, with the subsequent committed cells filling various compartments. In the M7 through M5 leukaemias, this compartment consists of trilineage-derived cells. The disease reflects the activities of the cell type most prominently involved, i.e. megakaryocyte, erythrocyte, or monocyte. Leukaemias of FAB classification M1 through M3 are specifically derived from cells committed to the granulocytic pathway, while M4, a combination of monocytic and granulocytic phenotypes, is probably derived from one or more earlier cells. Normal growth factors, which modify proliferation and maturation, are c-kit, IL-3 (active early), and GM-CSF (active late) in the proliferation period; and G-CSF specifically affecting the more mature granulocytic precursors (GM-CSF, granulocyte–macrophage colony-stimulating factor; G-CSF, granulocyte colony-stimulating factor). The resultant forms commonly released to the peripheral blood are blasts, some showing signs of maturation. The leukaemias with balanced translocations represent a loss of oncogene control, while those derived from earlier progenitors, more likely to be involved with a loss of DNA, relate to a loss of suppressor gene activity. Patients with balanced translocations or inversions comprise 20 per cent of the total AML population, while 50 per cent have grossly normal cytogenetics. Most of these are found within FAB classifications M0–5, while those antioncogene tumours with DNA loss occur in the FAB M5–7, secondary AMLs, and transformed MDS. This distribution is not absolute, and prognosis cannot be determined by FAB classification alone.

In contrast, myeloid leukaemias which occur spontaneously or with only brief incubation periods are probably clonal progeny with a single proto-oncogene, non-random genetic mutation. In these, chromosomal malfunction at the gene level follows transposition of genetic material to promoter areas and activation. These leukaemias, representing 20 per cent of all cases of AML, have either a chromosomal balanced translocation or an inversion. Some 50 per cent of patients with a normal karyotype have submicroscopic genetic mutations and overexpression. These alter nuclear oncogenes, many yet to be defined, which affect signal transduction, transcription, gene splicing, protein kinase C, and many other cellular activities.

Pretreatment management of acute leukaemia

Initial evaluation

The clinical signs and symptoms of acute leukaemia result from the suppression of normal haemopoiesis and the extramedullary expression of the malignant clone.

Bone marrow failure

Tumour suppression of the trilineage elements of the bone marrow produces pancytopenia. The signs and symptoms of anaemia result from the non-production of red blood cells, and may be associated with overt blood loss and with petechia and purpura at pressure sites in the skin. Variations in the dominance of loss of function of any the trilineage products relate to the specific type of AML (see FAB classification). Neutropenia is not commonly associated with infection in the rapid-onset subsets of AML. With a lymphoid-sparing pathogenesis, immunosuppression and its associated opportunistic infections are not frequently encountered in patients with AML. The incidence of compounding medical diseases is no greater that in the general population. Most signs and symptoms of acute multiorgan impairment disappear with adequate tumour control and homeostasis. The performance factor score is not a reliable indicator for making treatment decisions.

Extramedullary disease

A minority of patients present with a subleukaemia without circulating blast cells and require a bone marrow aspirate for diagnosis. The white blood cell (**WBC**) count may reach levels in excess of 100 000 × 10^9/l in 5 per cent of patients, although in the majority it is between 5 and 20 × 10^9/l. However, even in those patients with low WBC counts at presentation the initial decisions are urgent. Disseminated intravascular coagulation (**DIC**) can occur in promyelocytic leukaemia (M3) in young patients with pancytopenia. Emergency approaches have been designed to manage patients who will suffer the morbidity of leucostasis and tumour-lysis syndrome before and during treatment.

Diagnosis and supportive care

The advent of HLA typing, recognition of the risks of neutropenia, evidence of stem-cell resilience, availability of blood products, the acceptance of new aggressive chemotherapy regimens, recognition of genetic subsets, and identification of special training needs of staff now provide the basis for maximal levels of support for the patient presenting with acute leukaemia. These concepts of oncology care are essential for patients' survival throughout the intensive therapy required to reduce their leukaemic state to a minimum.

Support at presentation

Although the combination of bone marrow failure and tumour invasion determine the presenting clinical features of acute leukaemia, the aspects of clonal expression based on tumour mass, specific leukaemia-cell characteristics, and the rate of cell turnover determine the need for prompt intervention. Paradoxically, rapid cell killing with intensive drug treatment results in severe metabolic imbalances that are sometimes difficult to ameliorate. Survival depends on the prevention of both disease and treatment-related complications. The rational use of supportive measures combined with treatment permits both disease eradication and survival with treatment.

Immediate intervention is required in some cases. Efforts to decrease tumour mass both rapidly and safely entail awareness, rapid intervention, and aggressive treatment. Intense and appropriate critical-care management in emergency situations in patients with hyperleucocytosis, tumour-lysis syndrome, and DIC has now markedly reduced the historical early death rate of 50 per cent.

Hyperleucocytosis

Early mortality is related to the sheer mass of tumour cells. Patients with hyperleucocytic leukaemia (WBC count of more than 75 000 × 10^9/l), can suffer early death with central nervous system (**CNS**) haemorrhage and pulmonary capillary leakage sometimes due to a delay in treatment. These patients are at grave risk of vessel rupture. Clinically, high leukaemia blast-cell counts in AML and chronic myelocytic leukaemia (CML) are associated with fever and evidence of functional impairment of the lungs and brain. The most predictive sign of ongoing vascular rupture is a target purpura with a single, deep central nodule. The rigidity of the myeloblast causes stasis with expanding aggregates, arteriole infarction, and bleeding (Fig. 2 and Plate 1). Increased viscosity is offset by a concurrent anaemia. The combined haematocrit and leucocrit at presentation is usually no more than 45 per cent. Because this critical level is not exceeded unless iatrogenically increased by red blood cell transfusions, the incidence of clinical leucostasis in patients with WBC counts less than 75 × 10^9/l is low.

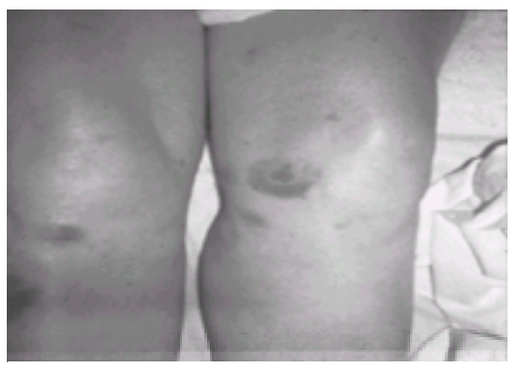

Fig. 2 Target purpura. Classic target appearance of purpura formed by infarction of an arteriole by a dividing cluster of leukaemic myeloblasts. Typically, a deep, firm nodule can be felt in the pale centre of the lesion. (See also Plate 1.)

Not all patients with an elevated WBC count are at risk of leucostasis. Those with acute lymphocytic leukaemia (**ALL**) and chronic lymphocytic leukaemia (**CLL**) are less likely to be affected, as are those with CML. Those in jeopardy are patients with myeloblasts having a high proliferative rate and short cell-cycle time. These patients form thrombi in the slow-flow microcirculation, with haemorrhage and necrosis of lung and brain tissue.

In a high percentage of patients CNS damage can be prevented with the rapid institution of a continuous infusion of high-dose cytarabine. Hydroxycarbamide (hydroxyurea) (6 g/m^2) and cyclophosphamide (3 g/m^2) are also of value in rapidly reducing the WBC count (Table 1).

Tumour lysis

An indwelling, triple-lumen catheter provides access for all necessary parenteral support throughout the treatment and recovery period. Initial platelet support and subsequent prophylactic antibiotics prevent the complications of bleeding and infection at the puncture site. Measures to allay the complications of rapid tumour kill and chemotherapy toxicity include hydration and diuresis. Allopurinol offsets the metabolic consequences of hyperuricaemia, particularly in lymphoid leukaemias, and aluminium hydroxide counters the opposing phosphataemia. The addition of bicarbonate is unnecessary. Furosemide (frusemide) given every 6 h is the critical component affecting the morbidity associated with the tumour-lysis syndrome and fluid overload. Low-dose dopamine helps to maintain renal blood flow and the

essential high urinary output. The toxicity of cytarabine when given by continuous infusion—namely, pulmonary vascular permeability—is prevented by rigorous control of the fluid balance. If major renal insufficiency does occur, then early dialysis effectively maintains the metabolic balance for the short at-risk period.

Disseminated intravascular coagulation (DIC)

An aggressive attack on ongoing or anticipated DIC has changed the dire prognosis of patients presenting with FAB M3, M3V, and occasionally M2 to M4 and M_5 myeloid leukaemias. Prophylactic infusion of heparin at 10 units/kg per hour in patients with M_3 and M_3V disease, together with close monitoring of changes in the prothrombin and partial thromboplastin times, fibrinogen, fibrin degradation products, and the platelet count, will prevent haemorrhage in these ultimately long-term survivors. Ongoing platelet consumption is countered with platelet infusions to maintain counts in excess of 50×10^9/l, and heparin to prevent the fibrinogen level from falling below 100 mg/ml. Some centres depend more heavily on the use of cryoprecipitate infusions.

Bleeding as a result of low-dose anticoagulation is seldom encountered; the risk of fatal haemorrhage is far greater without its use. This syndrome of DIC is driven by tumour-cell destruction, and should be monitored and treated throughout the initial days of chemotherapy. The WBC count will fall to 10 per cent of the presenting count within 72 h. The major cause of morbidity is any delay in the institution of leukaemic treatment.

Fever

At initial presentation, fever is related either to disease progression (B signs) or caused by a sensitive Gram-positive organism. Opportunistic and Gram-negative organisms and parasites are uncommon in patients with AML prior to their initial therapy. However, if a patient has been previously treated, the manifestations and organisms previously encountered must be anticipated.

Prophylactic antibiotics

If the patient is afebrile, prophylaxis is begun with an oral antiviral agent and vancomycin to prevent the first fever, most frequently caused by *Staphylococcus epidermitis* and/or *Corynebacterium* spp. This approach prohibits the use of empirical Gram-negative antibiotics for up to 5 days. The guidelines for antibiotic use at the Johns Hopkins' Oncology Center (**JHOC**) are outlined in (Table 2).

Laboratory monitoring

Laboratory data for tumour lysis, DIC, and chemotherapy-induced organ toxicity are determined daily until after tumour clearance, then twice a week.

Blood products

Once the WBC count is less than 50×10^9/l the haematocrit is maintained above 24 per cent. If evidence of bleeding and platelet deficiency is present on admission, and to anticipate DIC in all patients with AML, the platelet count is maintained at 50×10^9/l until measurements are reviewed and the leukaemia stabilized.

Hyperalimentation

Nausea and vomiting associated with intensive induction therapy limits a patient's oral intake to less than 500 calories/day. Parenteral feeding provides nourishment for normal bone marrow stem-cell regrowth and decreases the mitotic activity of the villus crypt cells of the gut, thereby eliminating them as a target for DNA synthesis-dependent drugs.

Three-step treatment

The immediate goal of therapy is to alter disease progression and stabilize the biology in favour of the patient by eradicating leukaemia and restoring bone marrow homeostasis. Currently, the use of potent chemotherapeutic agents is the first-choice therapy at diagnosis. Early and aggressive therapy, in contrast to the so-called 'standard' regimens, greatly improves survival.

Reviews of large randomized studies confirm that early intensive therapy with the few available drugs used at full dose cures a growing percentage of patients with myeloid malignancies, while providing a period of meaningful survival for many others (Table 3). This approach calls for an initial aggressive reduction of a large tumour mass with cautious management of tumour and therapy-related physiological instabilities, as well as chronic biological infirmities (stroke, myocardial infarction, DU, emphysema, etc.). When followed by greater intensification in remission this two-course sequence of rapid high-dose cytotoxicity results in a 40 per cent long-term survival in definable, selected populations. The key second course eliminates leukaemia in some, or substantially reduces it to small amounts in others, resulting in measurable and anticipated stable periods prior to relapse. These subgroups of patients with predicted outcomes furnishes the model for studies of novel therapeutic agents. Once homeostasis has been re-established and the patient is in remission, further interventions with novel agents are warranted in all groups, since there is a risk of relapse in all such patients. Many varied approaches now exist, or soon will, which are non-cytotoxic innovative modalities based in new biology.

Management of induction therapy—step 1

Drugs and dose

Variables predicting early failure must be taken into account when clinical instability is identified in patients presenting with leukaemia. However, since a delay in intervention is frequently not a valid option, it is important to select a dose schedule that provides stability and some therapeutic advantage based on the anticipated duration of bone marrow failure. The use of a moderately intense induction regimen (compared to the aggressive second course in remission) has reduced the clinical failure rate in successive studies. Examples of some two-step therapies are given in Table 4.

Support in remission

The interval between remission and the second course of therapy should be long enough to allow the patient to fully recover from all untoward consequences of the induction treatment. Those with FAB 1 to 5 disease usually recover without sequelae within a month, whereas those patients with FAB 6 to 7, SAML, and transformed MDS may be slower to recover. Any infections, but most importantly previous fungal infections, require treatment with appropriate antibiotics/antifungal agents prior to the patient receiving further chemotherapy: the use of amphotericin B will prevent the recurrence of fungal infections during the second course of therapy. On admission to hospital, patients should be treated with prophylactic antibiotics, since past infections in the patient will recur (Table 2).

Management of augmentation therapy—step 2

Intensive chemotherapy

Until recently it was assumed that cures could only be achieved in AML with therapy that destroyed both leukaemic and normal host bone marrow tissues. Requiring bone marrow transplanted from a suitable donor, this strategy is the ultimate aggressive single-course therapy. Although a high cure rate is apparent in those patients surviving allogeneic bone marrow transplantation, its application is limited by the age and availability of a suitable donor.

Another approach considers the infinite character of the bone marrow stem cell. With the development of four drugs effective in the treatment of AML, and the exploration of more intensive therapy, evidence accumulates that cures can be attained. Central to the success of these intensive regimens is the relative resistance of the host progenitor haemopoietic cell to tumoricidal doses of drugs. Such aggressive therapies contrast with those based on some of the principles developed in rodent lymphoid models and extrapolated to trials in children with lymphocytic leukaemia. These regimens initially reduce the tumour bulk to levels allowing the recovery of normal haemopoiesis, but they require continued long-term treatment to suppress tumour regrowth. These approaches using low-dose drugs are generally effective for only short periods in patients with AML. Conceptual change, using intensive high-dose therapy to eradicate residual leukaemia after its initial reduction,

has prompted a number of successful innovative studies that have matured sufficiently to allow comparison of the two approaches.

Patients with good-risk prognostic factors—normal genetics or balanced translocations with core-binding transcripts—are candidates for intensive non-ablative chemotherapy in remission. The data strongly support the use of a brief exposure to an intense regimen, high-dose cytarabine (**HDAC**) or timed sequential therapy (**TST**), which, with one or two courses, produce significant improvements over lesser therapies. There is no role for 'maintenance' therapy with similar cytotoxic drugs after the recovery of normal bone marrow function.

Alternately, patients with poor-risk prognostic factors (that is to say, genetic loss of a tumour suppressor) or with negative prognostic factors are candidates for bone marrow transplantation with stem-cell ablation.

Bone marrow transplantation (BMT)

The outcome, availability, and technology for BMT have all improved, and the indications for its use have increased in recent years. The number of bone marrow transplants performed annually has increased 10-fold since 1985, when essentially all transplants were allogeneic; 60 per cent are now autologous and 50 per cent use sources of stem cells other than from bone marrow.

Non-myeloablative BMT regimens are now used to produce mixed donor chimerism instead of full donor engraftment. This approach (referred to as 'minitransplants' or 'transplant-lite') can serve as a foundation for adoptive immunotherapy with donor lymphocyte infusions against leukaemias.

T-cell depletion and intensive immunosuppression can successfully prevent graft-versus-host disease, but these approaches are associated with an increased frequency of Epstein–Barr virus-associated lymphoproliferative disease.

High-dose cytotoxic therapy followed by BMT produces high response rates, prolongation of survival, and cures in some patients. It is indicated in patients with AML who are at risk of a poor response to non-ablative intensive chemotherapy. Autologous transplantation procedures are not as successful as intensive non-ablative chemotherapy.

Limitations of two-step treatment

Currently available drugs allow the most aggressive therapy to be rationally applied, and meaningful survival achieved. These brief intensive treatments aimed at leukaemia cure include doses of active drugs given to host tolerance levels, although some patients will require bone marrow rescue. Initial tumour reduction with a less intense regimen is followed by a more aggressive tumour-ablative treatment in a medically stable patient with normal bone marrow function. In an attempt to further reduce the amount of leukaemia while preserving host recovery capability, this approach, as in the ablative therapy for BMT, produces a large tumour kill while sparing host stem cells.

However, much more needs to be done for treating patients with all subsets of leukaemia. Only 70 per cent of adults with AML achieve complete remission (**CR**) following cytotoxic therapy. Further intensive chemotherapy or BMT during a patient's early first complete remission results in a longer than 5-year disease-free survival (**DFS**) in 35 to 50 per cent of adults treated in this manner. However, 30 per cent of all newly diagnosed AMLs are primarily refractory to induction therapy, in particular those evolving from myelodysplastic syndrome (**MDS**), an antecedent haematological disorder, and those linked to environmental/occupational exposures, including the secondary AMLs. The myelodysplastic syndrome-associated AMLs occur in older adults (over 60 years of age). Current treatment approaches to these AML variants yield CR rates of 40 per cent or less (however, the CR is brief, at less than 12 months), and low cure rates even with BMT.

At present, the cure rate for all adult acute leukaemias is only between 25 and 30 per cent. New approaches are needed to improve the clinical outcome of those adult leukaemias which are refractory to current therapeutic modalities.

Elimination of senescent leukaemia—step 3

New strategies with biological agents will test current approaches in all patients with minimal residual leukaemia. These new agents will probably be most effective after two courses of intensive timed sequential therapy, which reduces the tumour mass to minimal amounts, frequently to levels only detectable using PCR or fluorescent *in situ* hybridization (**FISH**) techniques.

Therapeutic strategies (novel agents)

Genetically engineered, human biological agents having minimal host toxicity and possessing activity at multiple levels of gene function are expanding the scope of applied medicine.

Brief, two-course, tumour-reduction therapy in all patients with AML provides a biologically stable array of genetically determined subgroups with predictable cell mass. These residual cells are vulnerable to designer agents. Coupled with the immunologically intact host, novel agents with a wide therapeutic advantage could annihilate the remaining 100 cells.

Examples of possible novel genetic agents of use in the treatment of varied leukaemias with minimal residual disease are outlined in (Table 5).

The ability to detect probable subtle differences in early outcome induced by the new biological agents is critical to expanding their trials, and hence rapidly moving to alternate treatments. These methods vary with the disease syndrome. In some good-risk patients, such as those with balanced chromosomal translocations, fluctuation of hybrid proteins may reflect and quantify any modifying effects of new agents. A prime example is the signal transduction inhibitor (STI571). Designed by flow-through and configuration technologies, this selected blocker of a tyrosine kinase specific for bcr-able has demonstrated significant antitumour effect and apoptosis in patients with CML in all stages. Persistent bone marrow depression and activity of STI571 against the rare gastrointestinal stromal tumours may relate to interaction with platelet derived growth factor and C-KIT as well as abl. Less effect in CML blast crisis may relate to the second-step genetic changes in that disease, with progressive proto-oncogenes characteristic of carcinoma.

In other groups of patients, the persistence of molecular abnormalities detectable in leukaemic cells (such as *RAS* mutations), kinetic responses to recruiting agents, or maturation effects may all be of specific value, but are as yet not clinically relevant to the study design. The duration of complete remission will be a valuable tool for measuring effect, particularly in that group of patients with leukaemic syndromes mirroring one-step carcinogenesis. Those with a loss or gain of DNA are predicted for relapse within a defined period. These genetic abnormalities promote biological instability and a tumour growth advantage. When minimal residual disease after induction therapy is achieved, any effective intervention that alters the time of relapse may be transferable to the management of patients with cancer, since the genetics of these leukaemias mimic those of carcinoma. Thus, the paradigm for the leukaemic model of cure of carcinoma is in those patients with an anti-oncogene with significant tumour reduction, but an anticipated short remission. Lengthening of this duration, or cure with genetically targeted novel biologies, will be of great significance.

Further reading

Bishop JF, *et al.* (1996). A randomized study of high-dose cytarabine in induction in acute myeloid leukemia. *Blood* **87**, 1710–17.

Bloomfield CD, *et al.* (1998). Frequency of prolonged remission duration after high-dose cytarabine intensification in acute leukemia varies by cytogenetic subtype. *Cancer Research* **58**, 4173–9.

Burke PJ (1993). Leukemia and the new biology. In: Niederbeuber JE, ed. *Current therapy in oncology*, pp 575–91. Mosby, New York.

Cassileth PA, *et al.* (1998). Chemotherapy compared with autologous or allogeneic bone marrow transplantation in the management of acute myeloid leukemia in first remission. *New England Journal of Medicine* **23**, 1649–56.

Fearon ER, *et al.* (1986). Differentiation of leukemia cells to polymorphonuclear leukocytes in patients with acute nonlymphocytic leukemia. *New England Journal of Medicine* **315**, 15–24.

Geller RB, *et al.* (1989). A two-step timed sequential treatment for acute myelocytic leukemia. *Blood* **74**, 1499–506.

Jaffee EM, *et al.* (1995). Use of murine models of cytokine-secreting tumor vaccines to study feasibility and toxicity issues critical to designing clinical trials. *Journal of Immunotherapy with Emphasis on Tumor Immunology* **18**, 1–9.

Mayer R J, *et al.* (1994). Intensive postremission chemotherapy in adults with acute myeloid leukemia. *New England Journal of Medicine* **231**, 896–903.

Phillips GL, *et al.* (1991). High-dose cytarabine and daunorubicin induction and postremission chemotherapy for the treatment of acute myelogenous leukemia in adults. *Blood* **77**, 1429–35.

Woods WG, *et al.* (1996). Timed-sequential induction therapy improves postremission outcome in acute myeloid leukemia: a report from the Children's Cancer Group. *Blood* **87**, 4979–89.

22.3.5 Chronic lymphocytic leukaemia and other leukaemias of mature B and T cells

D. Catovsky

Advances in the last decade have revealed a greater recognition of the heterogeneity in leukaemias arising from immunologically mature B and T cells. These advances resulted mainly from the systematic use of monoclonal antibodies against lymphocyte differentiation antigens, greater attention to morphological detail, and a more consistent evaluation of the patterns of lymphocytic infiltration in the bone marrow, lymph nodes, and spleen. Precise diagnosis is critical because there are new treatment modalities for these disorders and the related low-grade non-Hodgkin's lymphomas.

The principal methods used for diagnosis and classification include: films of peripheral blood and bone marrow aspirates, bone marrow trephine biopsies, monoclonal antibodies against lymphocyte differentiation antigens and antibodies specific to immunoglobulin (**Ig**) heavy and light chains, histology of involved organs such as lymph nodes and spleen, and cytogenetic analysis, chiefly by means of fluorescence *in situ* hybridization (**FISH**), which allows the study of cells in interphase. DNA analysis may help to elucidate pathogenesis and demonstrate clonality in cases of uncertain diagnosis. Other investigations which may provide information are protein electrophoresis, tests for free light chains in the urine, and imaging techniques. Physical examination should include palpation of lymph nodes, liver, and spleen, and detection of any skin infiltration.

There are two broad disease categories to be considered: primary lymphoid leukaemias and leukaemia/lymphoma syndromes, which as a rule correspond to non-Hodgkin's lymphomas manifesting with peripheral blood and/or bone marrow involvement. Both groups can be subdivided, according to their cell derivation, into B- and T-cell types (Table 1).

Chronic lymphocytic leukaemia

This is the most common form of lymphocytic leukaemia, accounting for at least 50 per cent of cases presenting with a lymphocyte count of 5×10^9/l or higher. In Western countries chronic lymphocytic leukaemia accounts for 25 per cent of all cases of leukaemia. It is less common in the Far East, comprising 2 per cent of cases. It is estimated that 1000 new cases are diagnosed in the United Kingdom each year.

Chronic lymphocytic leukaemia affects adults over the age of 50 years, with only 5 per cent of patients aged between 30 and 50, and it is rare below the age of 30. The peak incidence is between 60 and 80 years. The male to female ratio is 2:1. This ratio is greater in younger patients and lower in older patients.

Diagnostic criteria include lymphocytosis greater than 10×10^9/l and more than 30 per cent lymphocytes in the bone marrow aspirates. With the use of membrane markers it is possible to make a diagnosis with lymphocyte counts of at least 5×10^9/l. The lymphocytes in chronic lymphocytic leukaemia have distinct morphology (Plate 1): small size, round nucleus, clumped nuclear chromatin, and scanty cytoplasm; smear cells are common. A minority of cells are larger with a prominent nucleolus and have been designated prolymphocytes. Cases with more than 10 per cent prolymphocytes have a progressive clinical course, a higher proliferation rate than stable cases, and a correspondingly shorter lymphocyte doubling time. This variant form of chronic lymphocytic leukaemia has been described as chronic lymphocytic leukaemia/prolymphocytic.

Clinical features

Almost one-third of patients are diagnosed by chance with lymphocytosis and no specific symptoms or physical signs. Others present with lymphadenopathy or symptoms of anaemia. The lymphadenopathy is usually symmetrical and of moderate size, involving the neck, axillas, and inguinal regions. Other nodal areas can be ascertained by chest radiography and abdominal CT scan. Splenomegaly of variable size is found in 50 per cent of cases; hepatomegaly is less common and more difficult to document as being of clinical relevance.

Systemic symptoms such as weight loss or night sweating are not common but, when present, they correlate with bulky abdominal disease. Fever in chronic lymphocytic leukaemia usually indicates infection, but if the latter is excluded it may suggest transformation (see below).

Membrane markers

Chronic lymphocytic leukaemia lymphocytes are clonal B cells with weak kappa or lambda light-chain expression (**SmIg**) in the membrane. The immunophenotype of chronic lymphocytic leukaemia is unique within the B-cell disorders: CD5 and CD23 positive, FMC7 negative, and weak or negative expression of CD22 and CD79b. These markers, including the weak SmIg expression, represent the typical immunophenotype of chronic lymphocytic leukaemia. In a majority of cases (about 90 per cent) chronic lymphocytic leukaemia lymphocytes will show the expected findings with four or five of the above reagents. The most consistent markers are CD5 and CD23, positive in 92 and 94 per cent of cases, respectively. This contrasts with observations in other B-cell leukaemias and non-Hodgkin's lymphoma in leukaemic phase (Table 2). Chronic lymphocytic leukaemia/prolymphocytic has the same membrane markers as typical chronic lymphocytic leukaemia although in 20 to 30 per cent of cases it departs from the expected phenotype by expressing strong SmIg or CD79b or FMC7. As CD5 is a marker of both B and T cells, it may be necessary to establish in cases with a low white blood cell count that another B-cell antigen, such as CD19, is coexpressed in the CD5+ lymphocytes by simultaneous double labelling.

The expression of CD38 on chronic lymphocytic leukaemia lymphocytes seems to provide a strong new marker of prognosis. To be accurate, the assessment of CD38 needs to be done simultaneously with CD5 and CD19 to identify chronic lymphocytic leukaemia cells. CD38 appears also to be a surrogate marker for two forms of chronic lymphocytic leukaemia: (i) a benign one, CD38 negative, which arises from post-follicular (memory) B cells as shown by somatic mutations of the IgVH genes; and (ii) a more active and progressive form of chronic lymphocytic leukaemia, CD38 positive, which usually requires treatment and is associated with unmutated IgVH

genes (naive B cells). The proportion of cases with one or other form of chronic lymphocytic leukaemia appears to be similar.

Bone marrow findings

A bone marrow trephine biopsy should always be performed in chronic lymphocytic leukaemia to complement the clinical staging (see below), to exclude other B-cell leukaemias and non-Hodgkin's lymphoma, and as a baseline to assess disease progression and/or response to therapy. Immunohistochemistry with CD20 and CD79a is useful to highlight the areas of leukaemic infiltration.

The patterns of bone marrow infiltration in chronic lymphocytic leukaemia are variable and correlate with clinical stages. Early chronic lymphocytic leukaemia shows minimal interstitial or nodular involvement; with disease progression the normal bone marrow fat spaces are gradually replaced by lymphocytes. There is a mixed interstitial and nodular pattern and in advanced disease, the involvement is diffuse or 'packed'. The latter correlates with the presence of anaemia and/or thrombocytopenia. A paratrabecular pattern characteristic of non-Hodgkin's lymphoma, follicular lymphoma in particular, is not seen in chronic lymphocytic leukaemia. A nodular pattern, on the other hand, is common in lymphoplasmacytic, mantle cell, and splenic lymphoma with villous lymphocytes.

Bone marrow aspirates are necessary to confirm lymphocyte morphology and to evaluate infiltration. A minimum of 30 per cent lymphocytes is required for a diagnosis of chronic lymphocytic leukaemia, but this needs to be confirmed by cell marker studies.

Examination of the bone marrow is important in chronic lymphocytic leukaemia for three reasons: (i) to assess the degree of infiltration, which is an independent prognostic factor; (ii) to establish the possible mechanism of anaemia or thrombocytopenia by assessing the normal haemopoietic reserves; and (iii) to distinguish chronic lymphocytic leukaemia from cases of low-grade non-Hodgkin's lymphoma by the pattern of involvement in trephine biopsy sections. In patients with cytopenias and a large spleen, assessment of the bone marrow is critical to decide whether splenectomy may be beneficial.

Staging

The course of chronic lymphocytic leukaemia is very variable. Some patients may never require treatment and others have a progressive course with short survival. A major advance in the management of chronic lymphocytic leukaemia was the development of staging systems which can predict prognosis. The first system, used in the United States, was described by Rai *et al*. (1975). A new, simplified proposal by Binet *et al*. (1981) was subsequently adopted by the International Workshop on Chronic Lymphocytic Leukaemia (1981) whilst retaining some aspects of Rai's staging.

Both systems use simple information: blood counts and physical signs, namely lymphadenopathy and hepatosplenomegaly. Findings by imaging techniques are not taken into account for staging but are useful for assessing accurately disease bulk and measuring response to treatment.

Binet's system is currently used in chronic lymphocytic leukaemia trials in the United Kingdom. Stages A and B have no anaemia (haemoglobin, **Hb** > 10 g/dl) or thrombocytopenia (platelets > 100 × 10^9/l) and have a different degree of organ enlargement. Patients with stage A disease have either no palpable nodes (including liver and spleen as nodal areas) or have one or two involved areas. Patients with stage B disease have three, four, or all five nodal areas involved, which include nodes in cervical, axillary, and inguinal regions, spleen, and liver. Patients with stage C disease have anaemia (Hb < 10 g/dl) and/or thrombocytopenia (platelets < 100 × 10^9/l) and correspond to stages III and IV of the Rai system. The relative distribution of stages at presentation in chronic lymphocytic leukaemia is as follows: stage A, 45 to 50 per cent; stage B, 25 to 30 per cent; and stage C, 20 to 25 per cent. The proportion of patients in stages A, B, and C varies between the sexes. More women are likely to present with stage A and more men with stages B and C. Stage A is also more common over the age of 70 years in both sexes. Stage is the single most important prognostic factor of the disease (see below).

Attempts have been made to identify, within the large stage A group, further prognostic substages. One is to retain Rai stage 0 (lymphocytosis with no physical signs) as stage A(0). The other, proposed by the French Cooperative Group on Chronic Lymphocytic Leukaemia (1990), is to separate stage A into patients with (i) haemoglobin greater than 12 g/dl and a lymphocyte count less than 30 × 10^9/l (stage A') or (ii) haemoglobin less than 12 g/dl and a lymphocyte count greater than 30 × 10^9/l (stage A''). The 5-year survival of the first group was 87 per cent and of the latter 60 per cent. This difference was confirmed in the analysis of the United Kingdom Medical Research Council (**MRC**) Chronic Lymphocytic Leukaemia 3A study, which also confirmed the survival advantage of patients with stage A disease with lymphocyte doubling times greater than 12 months.

Chromosome abnormalities

Progress in this area has resulted from the routine use of fluorescence *in situ* hybridization (FISH), which helps detect the abnormalities in chronic lymphocytic leukaemia with a higher frequency (80 per cent of cases) than conventional cytogenetic methods (30 per cent). The most common abnormality is the interstitial deletion of 13q14 (50 to 60 per cent of cases), followed by trisomy 12 (20 per cent), 11q23 deletion (20 per cent), 17p13 (p53 locus) deletion (10 per cent), and 6q21 deletion (5 per cent of cases). Chromosome translocations have been described in chronic lymphocytic leukaemia but with low frequency. None of the above abnormalities are unique to chronic lymphocytic leukaemia as they may also be seen in other low-grade non-Hodgkin's lymphomas. On the other hand, translocations seen in non-Hodgkin's lymphomas such as t(11;14)(q13;q32), a feature of mantle cell lymphoma, and t(14;18)(q32;q21) are not a feature of chronic lymphocytic leukaemia.

Some of the changes detected by FISH in chronic lymphocytic leukaemia have been associated with distinct disease characteristics, namely: trisomy 12 with high proliferative rate and increased number of prolymphocytes; p53 deletion, which correlates with p53 overexpression and gene mutation, is associated with chronic lymphocytic leukaemia/prolymphocytic and poor response to therapy; 11q23 with younger age, massive lymphadenopathy, and poor prognosis; and 6q21 with high lymphocyte counts and bulky disease. It is likely that most of the abnormalities described in chronic lymphocytic leukaemia are relatively late events in the evolution of the disease. The nature of the early genetic event triggering naive (unmutated IgVH genes) or memory (mutated IgVH genes) B cells to become neoplastic is unknown.

Prognostic factors

The main features of poor prognosis in chronic lymphocytic leukaemia are listed in Table 3. The most important one for predicting survival is clinical stage, assessed by either the Binet or Rai systems, followed by age, sex, and response to therapy. The median survival of patients with stage A disease is more than 10 years, with stage B disease it is 6 years, and stage C disease 4 to 5 years.

For patients with stage A disease, both the French substaging (A', A'') and the lymphocyte doubling time are important independent prognostic variables (see above). A period of close observation with blood counts every 2 or 3 months for the first year is recommended in newly diagnosed patients with stage A disease in order to assess the pace of the disease and calculate the lymphocyte doubling time. The degree of bone marrow infiltration is also an independent prognostic variable, particularly in patients with stage B disease. A packed bone marrow pattern is associated with worse prognosis. In contrast to the anaemia caused by bone marrow infiltration, autoimmune haemolytic anaemia is not necessarily considered to indicate poor prognosis. Because chronic lymphocytic leukaemia affects elderly people, it is essential to investigate thoroughly other causes of anaemia and exclude, as not related to chronic lymphocytic leukaemia, those caused by iron, folate, or vitamin B_{12} deficiency, before deciding that the patient has stage C disease and requires chemotherapy.

The prospective value of new prognostic factors such as CD38 expression, IgVH mutations, and cytogenetic changes as detected by FISH analysis is still being evaluated. However, they are likely to define or to be associated with distinct clinical behaviour and to explain in large part the contrasting evolution of chronic lymphocytic leukaemia seen in patients.

Complications

Infections, particularly of the upper respiratory tract, are the main cause of morbidity in chronic lymphocytic leukaemia. Pneumonia is the main cause of death in 30 per cent of cases, usually in patients with advanced disease. The major predisposing factor for infections is hypogammaglobulinaemia.

Autoimmune phenomena, commonly haemolytic anaemia with a positive direct antiglobulin test due to warm autoantibodies, is a feature in 5 to 7 per cent of cases at presentation. However, the proportion of cases with a positive Coombs' test is higher than the number with frank haemolysis. Not infrequently, the haemolytic anaemia

is precipitated by the initiation of therapy, particularly the nucleoside analogue fludarabine, which causes a drop in CD4+ T cells, or following a viral illness. Immune thrombocytopenia is seen in 2 per cent of cases.

Other malignancies are not uncommon in chronic lymphocytic leukaemia and it is not clear whether this relatively high incidence correlates with age or with a greater predisposition of the disease itself or its associated immunodeficiency. Up to 30 per cent of patients may die of causes unrelated to chronic lymphocytic leukaemia and in half of them the cause is another cancer. In patients with early chronic lymphocytic leukaemia (stage A), half of the causes of death are not due to the disease itself and are often age related, such as cardiovascular events. In contrast, in advanced stages (stages B and C) 80 to 90 per cent of deaths are a direct consequence of chronic lymphocytic leukaemia and its complications.

Treatment

Because of the variable outlook of patients, which relates to stage and other disease features, it is important to consider the treatment separately for patients with early and stable disease and for those with progressive, symptomatic, and/or advanced disease. For this purpose staging is the first criterion to take into account.

The majority of patients with stage A disease have no symptoms and may be observed for a while to determine, by the lymphocyte doubling time or other features, whether the disease has a stable pattern, before deciding whether treatment is necessary. A number of randomized trials have considered whether patients having stage A disease should be treated early with chlorambucil, with or without prednisolone, and these have been summarized in an overview (Chronic Lymphocytic Leukaemia trialists, 1999) which showed conclusively that patients treated early did not fare better and, if anything, show a trend towards shorter survival.

Treatment is indicated for patients with stage B and C disease or stage A with evidence of progression. Disease progression is defined as a downward trend in haemoglobin or platelets, rising lymphocyte counts, development of lymphadenopathy, and systemic symptoms, among other parameters. Most of the treatments listed in Table 4 have been, or still are, subject to clinical trials. It is accepted that the addition of prednisolone to an alkylating agent, chlorambucil or cyclophosphamide as in **COP** (cyclophosphamide/oncovin/prednisolone) does not confer a survival advantage. There is good evidence, on the other hand, that the use of prednisolone alone for the first 4 weeks in patients who present with stage C disease facilitates the subsequent introduction of other drugs and corrects more rapidly the cytopenias. One needs to be aware that the use of corticosteroids results in a significant rise in the lymphocyte count whilst lymph nodes and spleen are reducing in size.

The role of anthracyclines, as in the combination **CHOP** (cyclophosphamide/oncovin/ prednisolone/doxorubicin) or by adding epirubicin to chlorambucil (Table 4), has been tested in several trials, including recently MRC CLL 3, but no survival advantage has been shown in a large meta-analysis of randomized trials. Although the response rates are slightly higher with anthracycline-containing combinations in previously untreated patients, for example 80 per cent for partial plus complete remissions, against 70 per cent with chlorambucil or COP, this has not translated into a survival advantage. The 5-year survival in the MRC CLL 3 trial was 44 per cent with chlorambucil and 45 per cent with chlorambucil plus epirubicin.

One of the difficulties in assessing survival as the only end point in clinical trials of chronic lymphocytic leukaemia is that patients who do not respond to a first-line therapy may respond to another and whenever there is a response (partial or complete) the outlook improves. Only patients who do not respond to any modality, such as those with p53 deletion/mutation or 11q23 deletion, fare really badly with respect to survival.

A new generation of drugs, the nucleoside analogues (Table 4), have shown promise for the treatment of chronic lymphocytic leukaemia and other low-grade lymphoid malignancies. The agent with greater activity in chronic lymphocytic leukaemia is fludarabine. The encouraging findings with fludarabine result from higher complete remission rates, for example 33 per cent in previously untreated patients, which compares favourably with the 15 per cent rate of complete remission that is observed with other agents. Furthermore, there is no evidence for cross-resistance between fludarabine and chlorambucil or anthracyclines. This makes fludarabine the agent of choice for second-line therapy in chronic lymphocytic leukaemia. It is not yet clear whether using fludarabine as first-line treatment provides any survival advantage.

Trials carried out in previously untreated patients by American and French co-operative groups have shown a higher rate of complete remission with fludarabine over chlorambucil or the combination CHOP and a prolonged disease-free interval, but no survival advantage. The reason for the latter, as stated above, may be that crossover design allowed good responses to fludarabine in the non-responders to chlorambucil or CHOP.

Data from the MD Anderson group in Texas suggest that fludarabine inhibits DNA and RNA synthesis as well as DNA repair, thus being potentially beneficial in combination with DNA-damaging agents. The combination of fludarabine with cyclophosphamide is currently being tested as first-line therapy in the Chronic Lymphocytic Leukaemia 4 trial in the United Kingdom. The potential benefit of the monoclonal antibodies, anti-CD20 and CD52 (Table 4), is currently being explored in chronic lymphocytic leukaemia with promising results. High-dose methylprednisolone is an effective salvage therapy for highly resistant patients, including those with p53 abnormalities. Although complete remissions are rare, partial responses, including some reverting to a bone marrow nodular pattern, have been observed. We have used high-dose methylprednisolone alone or in combination with other agents, such as vinca alkaloids, depending on results of an *in vitro* cytotoxicity assay (Bosanquet *et al.* 1999).

Stem cell transplantation

Efforts to improve treatment results in younger individuals with chronic lymphocytic leukaemia led to protocols using allogeneic and autologous stem cell transplantation. Allogeneic transplants may be curative in a minority, but this procedure is limited to those with an HLA-identical donor and has been associated in chronic lymphocytic leukaemia with a high (about 35 per cent) treatment-related mortality. To improve treatment-related mortality, less intensive conditioning regimens ('mini' transplants) have been devised in an attempt to exploit the host versus leukaemia effect. Currently, the use of the patient's own stem cells harvested after a good remission has been achieved is the preferred method of transplant for chronic lymphocytic leukaemia. The treatment-related mortality of autografts worldwide is less than 10 per cent and is more likely to be less than 5 per cent with improvements in supportive care and the ability to mobilize peripheral stem cells using haemopoietic growth factors such as granulocyte colony-stimulating factor (**G-CSF**). Due to its low toxicity, the 3- to 5-year survival after autografts is in the order of 75 to 80 per cent. However, the event-free survival is shorter due to 40 per cent relapses in the first 3 years. Thus, autologous transplants seem to increase survival in chronic lymphocytic leukaemia but may not be considered a curative procedure. The MRC Chronic Lymphocytic Leukaemia Working Group has conducted a pilot study since 1996 in which 102 patients have been entered so far. Remissions were induced with fludarabine and stem cells were harvested after cyclophosphamide priming followed by G-CSF. A harvest was not possible in a minority (about 15 per cent). The projected 5-year survival of the 50 patients who received transplants is 80 per cent, with a high proportion of bone marrow samples after autograft showing no sign of chronic lymphocytic leukaemia by polymerase chain reaction (PCR) for IgH gene rearrangement. A European randomized trial is planned, to assess further the value of autografts in younger patients with chronic lymphocytic leukaemia.

Splenectomy in chronic lymphocytic leukaemia

There are three indications for splenectomy in chronic lymphocytic leukaemia. First, for therapy-resistant disease with significant residual splenomegaly. Second, in patients with evidence of hypersplenism, that is cytopenia(s) and active bone marrow haemopoiesis. Third, for autoimmune complications, haemolytic anaemia, or thrombocytopenia that do not respond to therapy with corticosteroids and immunosuppressive drugs. In our experience, splenectomy is always beneficial in any of the above indications. In patients in whom the spleen is the dominant organ, with little or no lymphadenopathy, splenectomy could revert the clinical staging from C to A with the corresponding improvement in survival.

Because of the poor humoral immunity in chronic lymphocytic leukaemia, the prophylaxis after splenectomy should rely on oral penicillin as well as on antipneumococcal vaccines. The latter should also be used in all patients with chronic lymphocytic leukaemia.

Supportive care

The recurrent infections in patients with chronic lymphocytic leukaemia, particularly with advanced disease, makes supportive care an important component of management. This includes long-term antibiotics and their availability as soon as signs or symptoms of infections appear, and intravenous gammaglobulin replacement therapy to prevent serious upper respiratory tract infections in selected patients. Co-trimoxazole should be used in all patients treated with fludarabine or other nucleoside analogues that cause lymphopenia. Blood products should always be irradiated after fludarabine. Annual influenza vaccinations are strongly recommended. Other measures include blood transfusions and vitamin supplements such as folic acid to correct deficiencies. Anaemia in chronic lymphocytic leukaemia should always be thoroughly investigated and it should not be assumed that it is caused by bone marrow infiltration. The treatment of autoimmune

complications includes corticosteroids, splenectomy, danazol, azathioprine, cyclophosphamide, and cyclosporin A.

Transformation

There are two forms of transformation in chronic lymphocytic leukaemia: a subtle one with increased proportion (usually more than 10 per cent) of prolymphocytes, known as chronic lymphocytic leukaemia/prolymphocytic, and a more dramatic change to a high-grade non-Hodgkin's lymphoma with diffuse large B-cell/immunoblastic histology, known as Richter's syndrome. Chronic lymphocytic leukaemia/prolymphocytic may be seen in 10 per cent of patients and Richter's syndrome in at least 5 per cent. The former has a progressive course and is associated with trisomy 12 and p53 abnormalities in about 50 per cent of cases. The immunophenotype of chronic lymphocytic leukaemia/prolymphocytic is identical to typical chronic lymphocytic leukaemia with occasional cases scoring 3 instead of the usual 4 or 5. Chronic lymphocytic leukaemia/prolymphocytic should not be confused with B-cell prolymphocytic leukaemia (see below), which is a distinct entity and in which the immunophenotype usually scores less than 2 (Table 2).

The large cell transformation may be localized or generalized; very rarely, it may resemble an acute leukaemia with circulating large blasts. Richter's syndrome is associated with deteriorating clinical status and the systemic symptoms of fever, weight loss, and sweating, particularly when large para-aortic nodes are involved. Systemic symptoms or rapidly enlarging asymmetric nodes should always raise the question of transformation and be properly documented. Hypercalcaemia, a rare feature of chronic lymphocytic leukaemia, has been documented in patients developing Richter's syndrome.

One question which has generated interest is whether Richter's transformation represents a new malignancy or a new change within the chronic lymphocytic leukaemia B-cell clone. Studies with anti-light-chain antibodies and DNA analysis with probes for heavy- and light-chain genes seem to indicate that in 50 per cent of cases the transformation occurs within the pre-existing chronic lymphocytic leukaemia cells; in the rest it represents a new B-cell clone. In cases arising from a separate clone, it has been suggested that the new malignancy may be mediated by the Epstein–Barr virus (**EBV**) as shown by the expression of the EBV latent membrane protein (LMP-1) and EBV mRNA in tissue sections. Severe immunosuppression, usually through CD4 lymphopenia such as caused by fludarabine, may be involved in the development of the new malignancy, which, in the strict sense, is not truly a transformation event. Cases which histologically resemble Hodgkin's disease have been documented with an incidence of 0.5 per cent. These may also relate to treatment with nucleoside analogues.

Richter's syndrome has been associated with poor prognosis with a median survival of less than 6 months following presentation. Alkylating agents are no longer effective at this stage and neither is fludarabine. Combinations of the type used in high-grade non-Hodgkin's lymphoma, such as CHOP, may induce remissions in some patients. If complete remission is obtained, the outlook may be favourable. Patients with localized transformation seem to respond better than those with generalized lymphadenopathy.

B-cell prolymphocytic leukaemia

B-cell prolymphocytic leukaemia was originally described by Galton in 1974 as a variant form of chronic lymphocytic leukaemia but is now recognized as a distinct clinicopathological entity. The main disease features are splenomegaly without peripheral lymphadenopathy, anaemia, and thrombocytopenia and a high white blood cell count, usually over 100×10^9/l. The diagnosis is made by examination of peripheral blood films in which the predominant cells are prolymphocytes (Plate 2); small lymphocytes, as in chronic lymphocytic leukaemia, are rarely seen. B-cell prolymphocytic leukaemia is rare, representing 1 per cent of cases of lymphocytic leukaemia. Most patients are over the age of 60, with a median age of 70 years.

The immunophenotype of B-cell prolymphocytic leukaemia (Table 2) is different from that of chronic lymphocytic leukaemia: most cases express strongly SmIg, FMC7, CD22, and CD79b; two-thirds of cases are CD5 negative and CD23 is also often negative. The differential diagnosis should be considered with chronic lymphocytic leukaemia/prolymphocytic, mantle cell non-Hodgkin's lymphoma, and a rare variant form of hairy cell leukaemia. In chronic lymphocytic leukaemia/prolymphocytic the proportion of prolymphocytes is less than 50 per cent, there are many small lymphocytes in the blood films, and the immunophenotype is similar to that of chronic lymphocytic leukaemia. The circulating cells in the leukaemic phase of mantle cell lymphoma have a pleomorphic appearance, the nucleolus is not prominent, and they often have stippled nuclear chromatin and an indented nuclear outline. The membrane phenotype may be similar to B-cell prolymphocytic leukaemia except for CD5 which is positive in most cases. The cells in variant hairy cell leukaemia have a prominent nucleolus resembling prolymphocytes but their cytoplasm is abundant and has distinct 'hairy' projections. Their immunological profile may be similar to that of B-cell prolymphocytic leukaemia.

Many cases of B-cell prolymphocytic leukaemia have been reported with break points at chromosome 14q32 including the translocation t(11;14)(q13;q32) in 20 per cent. Such cases need to be distinguished from blastoid forms of mantle cell lymphoma presenting with leukaemia, but this may not be easy; cases with t(11;14) in both conditions overexpress cyclin D1. Abnormalities of the p53 gene (loss of heterozygosity, overexpression, and mutations) have been reported in more than 50 per cent of cases of B-cell prolymphocytic leukaemia, the highest incidence reported in lymphoid malignancies. Deletions at 11q23 and 13q14 have also been shown by FISH.

Treatment and prognosis

In contrast to chronic lymphocytic leukaemia, the evolution of B-cell prolymphocytic leukaemia is always progressive with a median survival of 3 to 4 years. Several treatment modalities have been used with moderate success: splenic irradiation, the combination CHOP, and splenectomy. Recently the nucleoside analogues fludarabine and cladribine have been shown to induce remissions in 50 per cent of patients. Chlorambucil and other alkylating agents are largely ineffective. The high incidence of p53 abnormalities may underlie the resistance of B-cell prolymphocytic leukaemia to chemotherapy.

Hairy cell leukaemia

Hairy cell leukaemia is a rare disorder comprising 2 per cent of lymphoid leukaemias characterized by pancytopenia and splenomegaly in two-thirds of cases; monocytopenia is a consistent finding. Most patients have circulating hairy cells but leucocyte counts rarely exceed 10×10^9/l. Hairy cells are larger than lymphocytes, their nuclei show a homogeneous loose chromatin pattern without a visible nucleolus (except in variant hairy cell leukaemia) and have an abundant cytoplasm with broad-based projections or villi. The nuclear outline is often kidney shaped.

The bone marrow trephine biopsy is the main diagnostic test and shows a unique pattern of infiltration with characteristic clear zones in between the cells. This infiltration is usually interstitial but may be also be focal. Bone marrow aspirates are, as a rule, unsuccessful (dry tap) due to the heavy deposition of reticulin fibres.

Hairy cells are B cells positive with CD19, CD20, and SmIg with light-chain restriction. When tested with the five markers listed in Table 2, the immunophenotype is different from chronic lymphocytic leukaemia but similar to other B-cell disorders. Four other monoclonal antibodies, CD103, CD11c, CD25, and HC2, which have shown specificity for hairy cells, are positive in most cases. Cases of variant hairy cell leukaemia have high white blood cell counts and no monocytopenia. Histologically they resemble typical hairy cell leukaemia. The cells are CD11c+, CD103+ (in 50 per cent of cases), but always HC2 and CD25 negative.

A cytochemical property of hairy cells is the presence of tartaric acid-resistant acid phosphatase, which is still used for diagnosis. In paraffin-embedded sections of bone marrow, hairy cells are positive with the monoclonal antibodies CD20 and DBA44. These reagents can help monitor residual disease after treatment as recognition of clusters of hairy cells in histological sections may be difficult.

Treatment and prognosis

The prognosis of patients with hairy cell leukaemia has improved dramatically with the advent of three treatment modalities: interferon-a, pento-statin, and cladribine. Splenectomy is reserved for patients presenting with very large spleens that are disproportionate to the degree of bone marrow involvement. Interferon-a improves the blood counts and the bone marrow but does not induce prolonged remissions when treatment is discontinued. Pentostatin induces complete remissions in 85 to 90 per cent of patients with few (less than 5 per cent) non-responders. Once treatment is discontinued the majority of responders remain in remission for more than 5 years; 40 to 50 per cent of patients are still in remission after 10 years. The response to these agents in variant hairy cell leukaemia is usually poor and the survival significantly shorter than in the typical form.

Transformation

A subtle transformation takes place in patients with hairy cell leukaemia in the form of massive abdominal lymphadenopathy with few systemic symptoms. The overall incidence of abdominal nodes in hairy cell leukaemia is about 25 per cent and is documented by performing routine CT scan investigations. The proportion with lymphadenopathy is higher in patients who relapse after previously successful treatments and/or who had long-standing disease. Abdominal lymphadenopathy may be associated with resistance to further therapy and with the presence of large hairy cells in both the bone marrow and the enlarged lymph nodes, supporting the concept of transformation suggested by clinical findings.

B-cell lymphomas in leukaemic phase

Several types of low- or intermediate-grade non-Hodgkin's lymphoma of B-cell type present or evolve with a leukaemic blood picture, for example more than 5×10^9/l circulating lymphoid cells (Table 1). The types of non-Hodgkin's lymphoma that most commonly develop a leukaemic phase are follicular lymphoma (Plate 4), mantle cell lymphoma (Plate 3), and splenic marginal zone lymphoma with villous lymphocytes (**SLVL**) (Plate 5). The main differential diagnosis is with chronic lymphocytic leukaemia, other non-Hodgkin's lymphoma, and in the case of SLVL, with hairy cell leukaemia.

The circulating cells in follicular lymphoma are small, have no visible cytoplasm, the nuclear chromatin has a smooth pattern and shows regularly deep nuclear clefts or indentations and an angular or irregular nuclear shape. Leukaemia in follicular lymphoma is associated with widespread disease, such as hepatosplenomegaly and lymphadenopathy. The membrane phenotype is different from chronic lymphocytic leukaemia (Table 2) and the cells often express CD10. Lymph node biopsy is essential for a definitive diagnosis. Cytogenetic analysis will show the translocation t(14;18)(q32;q21) and rearrangement of the BCL-2 gene by molecular techniques. Cases with leukaemia tend to run a more aggressive course and require a more intensive treatment approach than those without leukaemia.

SLVL is a distinct low-grade non-Hodgkin's lymphoma characterized by splenomegaly, moderate lymphocytosis (10 to 30 × 10^9/l), a small monoclonal band, and/or free light chains in the urine in 50 per cent of cases. The circulating lymphocytes have a small nucleolus and a cytoplasm with conspicuous villous projections that are often seen polarized in one end of the cell (Plate 5). A minority of cells show plasma cell differentiation. The bone marrow is minimally involved early in the disease and the biopsies show a nodular pattern and intrasinusoidal infiltration highlighted in trephine biopsies by CD20 staining. The immunophenotype of SLVL cells can be distinguished from that of chronic lymphocytic leukaemia (Table 2) and hairy cell leukaemia, both diseases with which it can be confused. SLVL lymphocytes do not express HC2 and CD25 as typical hairy cells. The distinction with variant hairy cell leukaemia is difficult unless there is tissue histology.

Splenectomy is a useful treatment modality for SLVL and fludarabine has recently been shown to induce complete remissions. The spleen histology shows predominant white pulp involvement with a prominent marginal zone, which contrasts with the predominantly red pulp infiltration pattern in typical and variant hairy cell leukaemia. Most cases have somatic mutations of the IgVH genes, suggesting a disease derivation from late B cells. Chromosome abnormalities are not consistent but chromosome 7q21–32 deletion has been found in 40 per cent of cases with abnormalities of the CDK6 gene at 7q21. Trisomy 3, a feature of marginal zone lymphoma of MALT type, has been found in 17 per cent of SLVL cases. The disease course is indolent. Transformation to a high-grade diffuse large B-cell lymphoma has been observed in 5 per cent of cases.

Leukaemia is common in mantle cell lymphoma presenting with splenomegaly and lymphocytosis. The circulating cells in mantle cell lymphoma (Plate 3) are of medium to large size with an irregular nuclear outline. The bone marrow biopsy shows nodular or paratrabecular involvement (Plate 6). In addition to histology, mantle cell lymphoma is characterized by the translocation t(11;14)(q13;q32) in 80 per cent of cases, involving the BCL-1/PRAD-1 gene at 11q13. This chromosome translocation, which can be shown in interphase cells by FISH, results in the overexpression of cyclin D1 in the nucleus, which can be demonstrated by immunohistochemistry in tissue sections and by flow cytometry in cell suspensions.

Large granular lymphocytic leukaemia

Most cases with persistent T-cell lymphocytosis greater than 2×10^9/l lasting for more than 6 months without an identifiable cause are likely to represent clonal proliferations of large granular lymphocytes. Clonality can be demonstrated by the rearrangement of T-cell receptor genes and sometimes, also, by chromosome translocations, which are not consistent in every case. Large granular lymphocytes have abundant cytoplasm with prominent azurophil granules and an excentric nucleus without a visible nucleolus. Half of the patients have splenomegaly without lymphadenopathy and are neutropenic or, less frequently, suffer from other cytopenias, in particular red cell hypoplasia. Rheumatoid arthritis and the presence of autoantibodies are a feature of 25 to 30 per cent of cases. The membrane phenotype shows mature T cells which are CD4–, CD8+ and, characteristically, express one or more antigens associated with natural killer cells, such as CD11b, CD16, CD56, and CD57. Bone marrow involvement is variable but is usually present in true large granular lymphocyte leukaemia. The spleen involvement is in the red pulp with reactive normal follicles (white pulp) and frequent granuloma formation. The bone marrow involvement is variable, often with an interstitial pattern, but representing less than 50 per cent of the bone marrow cells. Although many patients do not require active treatment, a significant number (about 60 per cent) present a therapeutic problem related to the associated cytopenia. Treatments that have been effective in some patients are cyclosporin A, prednisolone plus an alkylating agent, and pentostatin.

T-cell prolymphocytic leukaemia

This aggressive form of T-cell leukaemia is characterized by hepatosplenomegaly, lymphadenopathy, and high leucocyte counts, usually rising fairly rapidly above 100×10^9/l, but cases with an initial indolent course have been recognized. There is skin infiltration in the dermis around the blood vessels and appendages in 20 per cent of cases and pleural effusions are often seen. The blood picture (Plate 7) may resemble B-cell prolymphocytic leukaemia but typical T prolymphocytes are smaller than B prolymphocytes and have some distinct features: irregular nuclear outline and a deep basophilic cytoplasm with protrusions or blebs. The nucleolus is often prominent, but it may be hidden in 20 per cent of cases when small cells predominate (small cell variant). In 5 per cent of cases the cells have a cerebriform configuration (Sézary cell variant), but erythroderma is not a feature in these patients. Serology for HTLV-I is always negative.

The membrane phenotype corresponds to that of mature (post-thymic) T lymphocytes, CD2+, CD3+, CD7+, with CD4+, CD8– markers. One-third of cases coexpress CD4 and CD8 or are CD4–, CD8+. The diagnosis is made by examination of peripheral blood and bone marrow films and confirmed by the appropriate markers. Ultrastructural examination has been used to define the morphology in cases with small cells.

Cytogenetic abnormalities

There are consistent non-random chromosome abnormalities in T-cell prolymphocytic leukaemia affecting 90 per cent of cases. The most commonly involved inversion is of chromosome 14 with break points at 14q11, the T-cell receptor (**TCR**) a/b locus, and 14q32.1, locus of the proto-oncogenes TCL1 and TCL1b, which are activated through the translocation. Other karyotypic changes include: idic(8)(p11), t(8;8)(p11–12;q12) and trisomy 8q, and deletions at 12p13. The translocation t(X;14)(q28;q11) is less common but also involves the TCR a/b locus and the MTCP1 gene, which has 70 per cent homology with TCL1. Both TCL1 and MTCP1 can induce a T-cell leukaemia with a CD4–, CD8+ immunophenotype in a transgenic mouse model. Deletions and missense mutations involving the ataxia telangiectasia mutated (ATM) gene at 11q23 have been documented in a high proportion of cases. It is of interest that patients with ataxia telangiectasia have circulating T-cell clones with inv(14)(q11;q32) and that some develop a T-cell leukaemia indistinguishable from T-cell prolymphocytic leukaemia.

Treatment and prognosis

The median survival of T-cell prolymphocytic leukaemia in our historical series has been 7 months. Responses can be obtained in 50 per cent of cases with pentostatin, but the majority of these are partial. Complete responses can now be obtained with the humanized monoclonal antibodies CAMPATH-1H in two-thirds of patients who are resistant and only partially responsive to other agents. Currently CAMPATH-1H should be considered the best first-line therapy for T-cell prolymphocytic leukaemia. Autologous stem cell transplantation has been used in patients achieving a complete remission, with some success.

T-cell lymphomas in leukaemic phase

T-cell non-Hodgkin's lymphomas develop leukaemia with higher frequency that B-cell lymphomas. Two diseases in particular regularly evolve with circulating lymphoma cells in the peripheral blood: adult T-cell leukaemia/lymphoma and Sézary syndrome.

Adult T-cell leukaemia/lymphoma has a distinct geographical distribution affecting mainly the south-west islands of Japan, the Caribbean basin, and some parts of

South America—Brazil and Chile. The serological demonstration of antibodies to HTLV-I, the causative agent of adult T-cell leukaemia/lymphoma, is one of the tests necessary for diagnosis. At genomic level there is evidence of clonal integration of HTLV-I in the malignant T cells. Diagnosis is further suggested by the demonstration of adult T-cell leukaemia/lymphoma cells in peripheral blood films (Plate 8). These cells have an irregular nucleus with polylobed configuration and many atypical forms including large transformed ones. These cells have been described as 'flower' cells. Patients with adult T-cell leukaemia/lymphoma have generalized lymphadenopathy, splenomegaly, and skin rashes. Leucocyte counts are variable but often less than 50×10^{9}/l. Hypercalcaemia is present in two-thirds of patients and tends to be difficult to control.

The pathogenesis of the T-cell malignancy by HTLV-I involves multiple steps. Infection with HTLV-I is essential but only 1 out of 2000 infected patients may develop full-blown adult T-cell leukaemia/lymphoma. Cases with non-specific symptoms and signs and minimal lymphocytosis are considered as smouldering disease. Lymphoma forms without blood or bone marrow involvement have been recognized in 20 per cent of cases.

Adult T-cell leukaemia/lymphoma cells may resemble Sézary cells. The latter have more uniform features and a cerebriform rather than a hyperlobulated nucleus. Lymph node histology in adult T-cell leukaemia/lymphoma shows diffuse infiltration with pleomorphic T cells of small, medium, and large size (Plate 10). The median survival of acute forms of adult T-cell leukaemia/lymphoma is less than 12 months. Patients are treated as those with high-grade non-Hodgkin's lymphoma, but remissions are transient and opportunistic infections are common. Some benefit has been reported using the combination of interferon-a and zidovudine.

Sézary syndrome is a distinct form of cutaneous T-cell lymphoma characterized by erythroderma and circulating Sézary cells, usually of small size (also known as Lutzner cells) (Plate 9). The skin infiltration is epidermotrophic with the formation of Pautrier microabcesses. Sézary cells, as well as adult T-cell leukaemia/lymphoma cells, are mature T cells with a CD4+, CD8– immunophenotype. The main difference is strong expression of the interleukin-2 receptor, demonstrated by the monoclonal antibodies CD25, in adult T-cell leukaemia/lymphoma but not in Sézary cells. Cases of T-cell leukaemia without skin involvement and with cells that resemble Sézary cells morphologically but lack skin involvement are now considered as part of the spectrum of T-cell prolymphocytic leukaemia (see above).

Further reading

Bennett JM *et al*. (1989). Proposals for the classification of chronic (mature) B and T lymphoid leukaemias. *Journal of Clinical Pathology* **42**, 567–84.

Bolam S, Orchard J, Oscier D (1997). Fludarabine is effective in the treatment of splenic lymphoma with villous lymphocytes. *British Journal of Haematology* **99**, 158–61.

Bosanquet AG, Johnson SA, Richards SM (1999). Prognosis for fludarabine therapy of chronic lymphocytic leukaemia based on *ex vivo* drug response by DiSC assay. *British Journal of Haematology* **106**, 71–7.

Brito-Babapulle V *et al*. (1997). The impact of molecular cytogenetics on chronic lymphoid leukaemia. *Acta Haematologica* **98**, 175–86.

Catovsky D, Foa R (1990). *The lymphoid leukaemias*. Butterworths, London. [A description of the clinical and laboratory features of chronic lymphocytic leukaemia and other B- and T-cell leukaemias and lymphomas evolving with leukaemia.]

Catovsky D, Matutes E (1999). Splenic lymphoma with circulating villous lymphocytes/splenic marginal zone lymphoma. *Seminars in Hematology* **36**, 148–54.

Chronic Lymphocytic Leukaemia Trialists' Collaborative Group (1999). Chemotherapeutic options in chronic lymphocytic leukemia: a meta-analysis of the randomized trials. *Journal of the National Cancer Institute* **91**, 861–8.

Damle RN *et al*. (1999). Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* **94**, 1840–7.

Dearden CE *et al*. (1999). Long-term follow-up of patients with hairy cell leukaemia after treatment with pentostatin or cladribine. *British Journal of Haematology* **106**, 515–19.

Dearden CE *et al*. (2001). High remission rate in T-cell prolymphocytic leukaemia. *Blood* **98**, 1271–6.

Döhner H *et al*. (1997). 11q deletions identify a new subset of B-cell chronic lymphocytic leukemia characterized by extensive nodal involvement and inferior prognosis. *Blood* **89**, 2516–22.

Döhner H *et al*. (1999). Chromosome aberrations in B-cell chronic lymphocytic leukemia: reassessment based on molecular cytogenetic analysis. *Journal of Molecular Medicine* **77**, 266–81.

Dreger P *et al*. (1998). Early stem cell transplantation for chronic lymphocytic leukaemia: a chance for cure? *British Journal of Cancer* **7**, 2291–7.

Dyer MJ *et al*. (1997). *In vivo* 'purging' of residual disease in CLL with Campath-1H. *British Journal of Haematology* **97**, 669–72.

Grever M *et al*. (1995). Randomized comparison of pentostatin versus interferon-a_{2a} in previously untreated patients with hairy cell leukemia: an intergroup study. *Journal of Clinical Oncology* **13**, 974–82.

Hamblin TJ *et al*. (1999). Unmutated Ig VH genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* **94**, 1848–54.

Jaffe ES *et al*., eds. (2001). *World Health Organization classification of tumours–pathology and genetics of tumours of haematopoietic and lymphoid tissues*. IARC Press, Lyon.

Keating *et al*. (1998). Long-term follow-up of patients with chronic lymphocytic leukemia (CLL) receiving fludarabine regimens as initial therapy. *Blood* **92**, 1165–71.

Lamy T, Loughran TP Jr (1999). Current concepts: large granular lymphocyte leukemia. *Blood Reviews* **13**, 230–40.

Lens D *et al*. (1997). p53 abnormalities in B-cell prolymphocytic leukemia. *Blood* **89**, 2015–23.

Lens D *et al*. (1997). p53 abnormalities in CLL are associated with excess of prolymphocytes and poor prognosis. *British Journal of Haematology* **99**, 848–57.

Maljaei SH *et al*. (1998). Abnormalities of chromosomes 8, 11, 14, and X in T-prolymphocytic leukemia studied by fluorescence *in situ* hybridization. *Cancer, Genetics and Cytogenetics* **103**, 110–16.

Matutes E (1999). *T-cell lymphoproliferative disorders—classification, clinical and laboratory aspects.* Harwood Academic Publishers, Australia.

Matutes E *et al*. (1991). Clinical and laboratory features of 78 cases of T-prolymphocytic leukemia. *Blood* **78**, 3269–74.

Matutes E *et al*. (1996). Trisomy 12 defines a group of CLL with atypical morphology: correlation between cytogenetic, clinical and laboratory features in 544 patients. *British Journal of Haematology* **92**, 382–8.

Matutes E *et al*. (1999). FISH analysis for BCL-1 rearrangements and trisomy 12 helps the diagnosis of atypical B cell leukaemias. *Leukemia* **13**, 1721–6.

Mercieca J *et al*. (1992). Massive abdominal lymphadenopathy in hairy cell leukaemia: a report of 12 cases. *British Journal of Haematology* **82**, 547–54.

Mercieca J *et al*. (1994). The role of pentostatin in the treatment of T-cell malignancies: analysis of response rate in 145 patients according to disease subtype. *Journal of Clinical Oncology* **12**, 2588–93.

Montserrat E *et al*. (1996). Bone marrow assessment in chronic lymphocytic leukaemia: aspirate or biopsy? A comparative study in 258 patients. *British Journal of Haematology* **93**, 111–16.

Moreau EJ *et al*. (1997). Improvement of the chronic lymphocytic leukemia scoring system with the monoclonal antibody SN8 (CD79b). *American Journal of Clinical Pathology* **108**, 378–82.

Österborg A *et al*. (1996). Humanized CD52 monoclonal antibody Campath-1H as first-line treatment in chronic lymphocytic leukaemia. *British Journal of Haematology* **93**, 151–3.

Pawson R *et al*. (1997). Treatment of T-cell prolymphocytic leukemia with human CD52 antibody. *Journal of Clinical Oncology* **15**, 2667–72.

Pawson R *et al*. (1997). Sézary cell leukemia: a distinct T-cell disorder or a variant form of T prolymphocytic leukaemia? *Leukemia* **11**, 1009–13.

Pekarsky Y *et al*. (1999). Abnormalities at 14q32.1 in T cell malignancies involve two oncogenes. *Proceedings of the National Academy of Sciences (USA)* **96**, 2949–51.

Sainati L *et al*. (1990). A variant form of hairy cell leukemia resistant to a-interferon: clinical and phenotypic characteristics of 17 patients. *Blood* **76**, 157–62.

Sood R *et al*. (1998). Neutropenia associated with T-cell large granular lymphocyte leukemia: long-term response to cyclosporine therapy despite persistence of abnormal cells. *Blood* **91**, 3372–8.

Vorechovský I *et al*. (1997). Clustering of missense mutations in the ataxia-telangiectasia gene in a sporadic T-cell leukaemia. *Nature Genetics* **17**, 96–9.

22.3.6 Chronic myeloid leukaemia

Tariq I. Mughal and John M. Goldman

Introduction

Chronic myeloid leukaemia (**CML**)—a term historically used interchangeably with chronic granulocytic leukaemia, chronic myelogenous leukaemia, and chronic myelocytic leukaemia—is a clonal malignant myeloproliferative disorder believed to originate in a single abnormal haemopoietic stem cell. It involves myeloid, monocytic, erythyroid, megakaryocytic, B-lymphoid, and sometimes T-lymphoid lineages. The first cases were described in the 1840s. A major landmark in the study of CML was the discovery of the Philadelphia (**Ph**) chromosome in 1960. Later it was established that the Ph chromosome was linked to the genetic events that cause CML. CML became the first human cancer in which a specific cytogenetic abnormality could be linked to its pathogenesis.

The past two decades have witnessed an enormous increase in our understanding of the molecular biology of CML. Such knowledge has enabled specialists to define precisely some of the molecular events and relate them to the prognostic factors of the individual patients with CML. During this period the prognosis of patients with CML has evolved from incurable to potentially curable by treatment with allogeneic haemopoietic stem-cell transplantation (allo- **SCT**). However, only a relatively small proportion of all patients with CML are eligible for allo-SCT, which still carries an appreciable risk of mortality and morbidity. For the majority of patients, interferon-alpha (IFN-á) have been found to suppress the CML cells and prolong survival in comparison to hydroxyurea. Adoptive immunotherapy using donor-lymphocyte infusions have proven valuable in treating selected patients with mini-SCT, and for rescuing those who relapse following a conventional allogeneic SCT. A new Abl-specific tyrosine kinase inhibitor, designated STI571 or imatinib mesylate, has recently been introduced and early clinical results are extremely encouraging.

Epidemiology

The annual incidence of CML is about 1 to 1.5 per 100 000 of the population. It accounts for approximately 15 per cent of all leukaemias in adults but less than 5 per cent of all childhood leukaemias. In the United Kingdom there are about 700 new cases each year. The median age of onset is 60 years and there is a slight male excess. With the possible exception of China there appears to be no clear geographical variation in the incidence.

Aetiology

Most cases of CML occur sporadically. No aetiological factor can be incriminated in the great majority of cases. A marginally increased risk of developing CML has been reported following exposure to high doses of irradiation, as occurred in survivors of the Hiroshima and Nagasaki atomic bombs in 1945. No familial predisposition or specific HLA genotypes have been recognized, but a small number of families with a high incidence of the disease have been reported.

Natural history

Characteristically CML is a biphasic or triphasic disease. Most patients present in the initial stable 'chronic' phase which typically lasts for 4 to 7 years. The natural history involves a spontaneous but largely predictable progression to an 'advanced phase', a term that covers the 'accelerated' phase and also 'blast crisis' or 'blastic transformation' (Table 1). About half of all patients in chronic phase transform directly into blast crisis, and the remainder do so following an intervening period of accelerated phase. In blastic transformation the CML cells fail to mature, and the blast cells resemble either the myeloblasts (myeloid blastic transformation) or lymphoblasts (lymphoid blastic transformation) found in patients with *de novo* acute myeloid or acute lymphoblastic leukaemia respectively.

Clinical features

Most patients typically present with lethargy and anorexia or abdominal discomfort due to splenomegaly, but 30 to 40 per cent of patients are asymptomatic and the diagnosis is made following a routine blood test. The principal physical finding is a palpable spleen, which is found in up to three-quarters of patients. Hepatomegaly and lymphadenopathy are uncommon. The clinical features of 430 patients referred to the Hammersmith Hospital in London are shown in Table 2. Occasional patients have 'chloromas' or 'granulocytic sarcomas' with subcutaneous deposits of extramedullary leukaemia.

In contrast to patients in the chronic phase, patients in the advanced phase are often symptomatic with fever, bone pain, bleeding, and/or excessive sweating. Splenic pain due to splenic infarct is not uncommon. During the accelerated phase patients often require increasing doses of hydroxycarbamide (hydroxyurea) to control the neutrophil counts.

Molecular biology

The Ph chromosome is an acquired cytogenetic abnormality present in all CML cells (Fig. 1). It is formed as a result of a reciprocal translocation of genetic material from the long arm of one of the two no. 9 chromosomes with material from the long arm of one of the no. 22 chromosomes, referred to as t(9;22)(q34;q11). This translocation involves transections of the *BCR* (breakpoint cluster region) gene normally on chromosome 22 and the *ABL* (Abelson) gene normally on chromosome 9 and so results in the juxtaposition of 5' sequences from the *BCR* gene with the 3' sequences from the *ABL* gene. The end results is the creation of a chimeric or 'fusion' *BCR–ABL* gene.

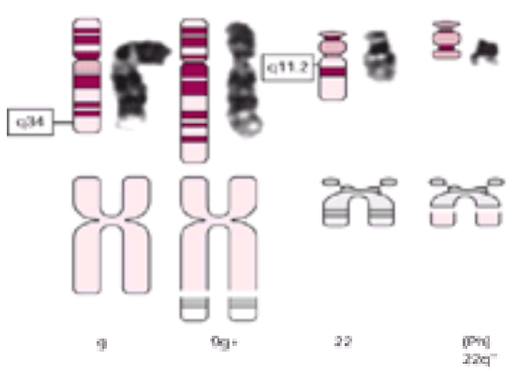

Fig. 1 Partial karyotype showing the Philadelphia chromosome translocation t(9;22)(q34;q11) with the positions of the breakpoints on chromosomes 9 and 22.

The classical Ph chromosome is easily identified in 80 per cent of patients with CML. Variant translocations are seen in a further 10 per cent of patients; these variants may be 'simple' involving chromosome 22 and a chromosome other than chromosome 9, or 'complex', where chromosomes 9, 22, and other additional chromosomes are involved. About 8 per cent of patients with classical clinical and haematological features of CML lack the Ph chromosome and are referred to as cases of Ph-negative CML. About half of such patients have a *BCR–ABL* chimeric gene and are referred to as Ph-negative, BCR-ABL-positive cases; the remainder are BCR-ABL-negative, and some of these have mutations in other genes. It is probable that these latter patients have a more aggressive clinical course than those with Ph-negative, BCR-ABL-positive disease. As the disease progresses patients may acquire additional cytogenetic abnormalities, including duplication of the Ph chromosome, trisomy 8, and isochromosome 17q. Mutations or deletions of tumour-suppressor genes such as *p16* and *p53* may contribute to the disease progression.

The *BCR–ABL* fusion gene transcribes an mRNA which encodes a protein that has a greater tyrosine kinase activity than the normal ABL protein. Depending on the site of the breakpoint in the *BCR* gene, the fusion protein can vary in size from 185 kDa to 230 kDa (Fig. 2). To date, three separate breakpoint locations on the *BCR* gene have been identified. When the break occurs in *major breakpoint cluster region* (**M-BCR**) it is nearly always in the intron between exons e13 and e14 or in the intron between exons e14 and e15 (toward the telomere). By contrast, the position of the breakpoint in the *ABL* gene is highly variable and may occur at almost any position upstream of exon a2 (toward the centromere). Most patients with the classical Ph chromosome have express transcripts with e13a2 or e14a2 junctions which translate as 210-kDa oncoproteins ($p210^{BCR–ABL}$) (Table 3). A break in the first intron of the *BCR* gene, between exons e1 and e2, in an area designated the *minor breakpoint cluster region* (m-bcr) results in the transcription of an e1a2 mRNA which encodes a 190-kDa protein ($p190^{BCR–ABL}$). This is found in about two-thirds of patients with Ph-positive acute lymphoblastic leukaemia. The third position for a break in the *BCR* gene is between exons e19 and e20, in an area designated *micro breakpoint cluster region* (**μ-bcr**). The associated mRNA product, e19a2, encodes a larger protein of 230 kDa ($p230^{BCR–ABL}$), which is found in the very rare cases of chronic neutrophilic leukaemia associated with a Ph chromosome.

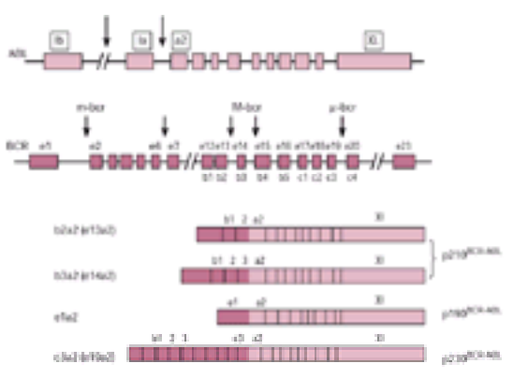

Fig. 2 Schematic representation of the various breakpoints in the *ABL* and *BCR* genes and the encoded proteins in the BCR–ABL-positive leukaemias

The different breakpoints in the M-BCR result in two slightly different chimeric *BCR–ABL* genes. A break occurring in the M-BCR intron between exons e13 and e14 yields an e13a2 (previously known as b2a2) mRNA, whereas a break occurring in the intron between exons e14 and e15) produces an e14a2 mRNA (previously known as b3a2). Most patients have either e13a2 or e14a2 transcripts, although both transcripts are present in about 10 per cent of cases. The type of BCR-ABL transcript has no important prognostic significance, altough patients with the e14a2 transcripts may have higher platelet counts.

It is believed that the various BCR–ABL transcripts play a central role in the pathogenesis of CML, though the precise details are still not fully understood. Various efforts to determine the function of the BCR–ABL proteins have established that, as a consequence of increased tyrosine kinase activity, the BCR–ABL protein can phosphorylate several substrate molecules, such as CRKL, p62Dok, paxillin, CBL, and RIN, thereby activating multiple signal-transduction pathways affecting cell growth and differentiation. The details of the pathways are incomplete, but a popular hypothesis is that the BCR–ABL protein activates the same pathways as are activated by cytokines that control the growth and differentiation of haemopoietic cells, thereby allowing CML cells to circumvent normal cellular growth and differentiation and become malignant. This would not, however, explain the mechanism by which the preferential proliferation and differentiation of myeloid progenitors occur. The molecular basis for transformation of CML from chronic phase to more advanced phases remains poorly understood, although several molecular changes, in particular mutations or deletions of p53, p16, retinoblastoma protein, and mutations or overexpression of Ras and EVI-1, have been identified.

Normal haemopoietic stem cells may be maintained in a resting state (designated G_0) as a result of the proliferation of CML cells. Under certain circumstances, however, these normal cells can be induced to proliferate, and this provides the rationale for autografting as a treatment of CML. There may also be a subpopulation of deeply quiescent Ph-positive CML cells, that might be relatively resistant to eradication by cycle-active cytotoxic drugs even when administered in high doses.

Diagnosis

The diagnosis of CML is commonly based on the characteristic appearances of the peripheral blood film and bone marrow aspirate and trephine biopsy. Cytogenetic analysis for the presence of the Ph chromosome is confirmatory. Molecular studies for the evidence of the BCR–ABL product provide additional confirmation.

The peripheral blood usually shows a leucocytosis which involves cells at all stages of differentiation within the myeloid lineage (Fig. 3 and Plate 1). Basophilia is an important diagnostic feature as its absence suggests other myeloproliferative disorders, particularly if the Ph chromosome and *BCR–ABL* gene are also absent. Eosinophilia may also be present but has no diagnostic relevance. There is a relative monocytopenia, although absolute numbers may be increased corresponding with the leucocytosis. This differentiates CML from chronic myelomonocytic leukaemia. Thrombocytosis with platelet anisocytosis and nucleated red cells are common.

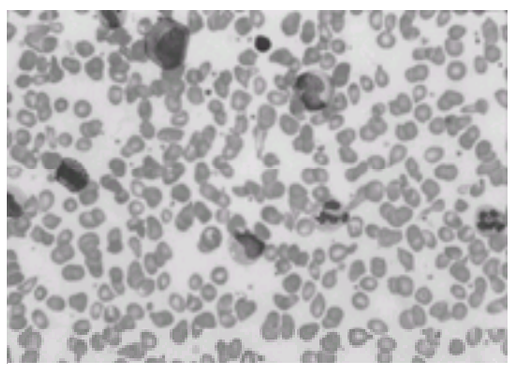

Fig. 3 Peripheral blood film from a patient with CML in chronic phase. (Photograph kindly provided by Professor Barbara Bain, Imperial College London.) (See also Plate 1.)

The bone marrow is markedly hypercellular with an increased myeloid to erythroid ratio due to the predominance of myeloid cells, particularly neutrophils and myelocytes. There may be no features of abnormal maturation in the precursors. Megakaryocytes are increased and may form clusters, which are less striking than those seen in essential thrombocythaemia. Reticulin fibrosis is usually absent or mildly increased at diagnosis.

Blastic transformation is often abrupt and striking with blast cells comprising up to 100 per cent of nucleated cells seen in the blood and marrow; it thus resembles an acute leukaemia arising *de novo*. These blast cells are of myeloid lineage in about 70 per cent of the patients, while in about 20 per cent the express lymphoid surface markers, have rearrangement of immunoglobulin genes and are presumably lymhoid in origin. The remaining patients in blast crisis have a blast population with a mixed or indeterminate morphology and immunophenotype.

Prognostic factors

Various efforts have been made to establish criteria definable at diagnosis that may help to predict survival for individual patients. Historically the most frequently used method was that proposed by Sokal, whereby patients can be divided into various risk categories based on a mathematical formula that takes account of the patient's age, blast-cell count, spleen size, and platelet count at diagnosis. The Sokal index has recently been updated by a retrospective study of patients treated with IFN-a in Germany — the new Euro or Hasford system is similar to the Sokal index but includes consideration of basophil and eosinophil numbers. Currently the best prognostic indicator may be the response to initial treatment with **IFN-a**; patients who achieve a degree of cytogenetic response have the best survival. These approaches help to predict survival with a non-transplant strategy. Other methods, such as the risk assessment proposals by Gratwohl *et al*. and Lee *et al*., are designed to calculate the risk of transplant-related mortality after an allo-SCT. Other possible prognostic factors are the presence or absence of deletions in the derivative 9q+ chromosome and the rate of shortening of telomeres in the leukaemia clone.

Management

The substantial recent developments in treating CML, including the introduction of IFN-a and allogeneic SCT in the 1980s and of STI571 very recently, have made management decisions for individual patients in chronic phase fairly complex. Thus some patients with CML can be cured by an allogeneic SCT, but the risks associated with it need to be carefully assessed. In contrast, IFN-a induces haematological control in a significant proportion of patients and confers some improvement in survival in comparison with hydroxyurea, especially for those patients who achieve a degree of cytogenetic response. It is rare, however, to achieve a molecular remission with IFN-a. STI571 induces haematological remission in almost all patients and is associated with a high incidence of cytogenetic response, but there is not yet information on its ability to confer molecular remission; neither is there any evidence as yet that it prolongs survival in comparison to the other treatments of CML. It is therefore prudent to discuss the relative merits of the various treatments with the patient at the time of diagnosis and to explain some aspects of the disease and the proposed management strategy.

Non-transplant treatment options

IFN-a is a member of a large family of glycoproteins of biological origin with antiviral and antiproliferative properties. It is active in reducing the leucocyte count and reversing the clinical features in 70 to 80 per cent of patients with CML. Five to 15 per cent of patients achieve a major reduction in the proportion of Ph-positive marrow metaphases. A number of prospective studies comparing IFN-a to hydroxycarbamide (hydroxyurea) and busulfan have been reported. The standard treatment for most patients with CML in the 1980s was hydroxycarbamide which had largely replaced busulfan. By the mid-1990s it was suggested that IFN-a was probably superior to both hydroxycarbamide and busulfan, and, more importantly, it appeared that IFN-a treatment resulted in prolongation of survival, in particular for patients achieving substantial cytogenetic responses. A meta-analysis of seven prospective randomized trials has confirmed the superiority of IFN-a over both busulfan and hydroxycarbamide. The meta-analysis involved over 1500 patients and found a 5-year survival of 57 per cent for the IFN-a treated patients compared to 42 per cent for the chemotherapy treated cohort. A major cytogenetic response (>66 per cent Ph-chromo-some negativity in the marrow) was seen in 10 to 38 per cent of all patients and occurred usually within 12 to 18 months of starting IFN-a therapy.

There are several key issues still unresolved with regard to IFN-a. The optimal dose remains a matter of some controversy although the recent study conducted by the UK Medical Research Council found no difference between 'high' and 'low' doses. Moreover, the optimal duration of IFN-a treatment remains undefined. Toxicity is common but is generally mild and reversible. Most patients suffer from influenza-like symptoms on starting treatment. Later they may experience lethargy and weight loss. Less common effects include autoimmune-mediated complications, such as thrombocytopenia and hypothyroidism. In an effort to improve the treatment results, several trials focused on combining IFN-a with cytotoxic drugs. Encouraging cytogenetic results have been obtained with the addition of cytarabine to IFN-a in French trial and a survival advantage in comparison to IFN-a was observed; this has not however been confirmed in a comparable Italian study. In an attempt to reduce the toxicity of IFN-a therapy, a pegylated form of IFN-a has also been investigated and preliminary experience is encouraging.

The precise mode of action of IFN-a remains uncertain. *In vitro* studies, mainly from long-term cultures of CML cells, suggest a prominent antiproliferative effect of IFN-a in CML. Some of this activity against CML cells may be mediated through the dendritic cells. IFN-a may also indirectly influence the survival of CML cells by restoring the defective cytoadhesion of the CML cells or by recruiting accessory cells of the immune surveillance system, by inhibiting other cytokines or by augmenting the action of natural killer cells.

Clinical trials designed to assess the efficacy and safety of STI571 in patients in all phases of CML began in 1998. In patients with CML in chronic phase refractory to or intolerant of IFN-a who receive at least 300mg of STI571 daily the incidence of complete haematological response is 98 per cent of the patients; about 40 to 45 per cent achieve cytogenetic responses. STI571 was administered orally and so far no important side-effects have been noted. The drug also demonstrated significant activity in CML in accelerated phase and in blast crisis. The follow-up is still relatively short, but if the responses are sustainable, this novel drug will become the preferred non-transplant treatment option. The drug is now licensed for these indications on both sides of the Atlantic and studies to compare STI571 alone with STI571 plus IFN-á and STI571 plus cytarabine are being designed. The combination of STI571 with pegylated IFN-á is also being tested.

Stem-cell transplantation

Allogeneic SCT, using blood- or marrow-derived stem cells derived from an HLA-matched sibling donor, performed in the chronic phase can cure a substantial proportion of patients with CML. International Bone Marrow Transplant Registry (**IBMTR**) showed that the leukaemia-free survival (**LFS**) at 5 years is 55 to 60 per cent. The probability of relapse at 5 years was 15 per cent. In contrast, the results of allogeneic SCT performed in the advanced phase of CML were generally poor. The probabilities of LFS at 5 years for those transplanted in the accelerated phase and blast crisis were 28 per cent and 10 per cent, respectively. The clinical results of transplantation using HLA-matched sibling donors appear to be relatively consistent worldwide.

The major determinants for survival, other than the phase of the disease, include the patient's age at transplant and the cytomegalovirus (**CMV**) status of the patient, and the age and sex of the donor. Survival appears to be best for patients who are transplanted within 1 year of diagnosis, are less than 40 years of age, have a young male donor, and both patient and donor are CMV-negative. For such a cohort, the 5-year LFS is probably 70 to 80 per cent; the relapse rate would be 10 to 20 per cent. It is possible that the precise details of the transplant procedure also influence the outcome. The cytoreductive regimens used prior to the transplant and the preventive measures for graft-versus-host disease (**GvHD**) appear especially important. The source of stem cells may also influence the result; following a peripheral-blood SCT patients achieve rapid engraftment, but there is a slight excess of chronic GvHD, perhaps due to the increased T-cell numbers in the peripheral blood compared to the bone marrow. There has also been some concern with regard to the effect of prior IFN-a therapy following an initial report suggesting a possible detrimental effect, further careful monitoring is therefore necessary.

Although allogeneic SCT using an HLA-matched sibling donor appears to cure some patients with CML, this treatment is only available to about 15 to 20 per cent of all patients with CML. This is largely due to a lack of suitable family donors and to age limitations. The rigors associated with SCT mean that most patients over the age of 60 years have a substantial incidence of transplant-related mortality (**TRM**). For these reasons, efforts have been made to identify suitable volunteer unrelated donors (**VUD**) and better conditioning regimens to reduce the toxicity. Historically, the results of VUD-SCT are inferior to those of HLA-matched sibling SCT due to an increased rate of graft failure, GvHD, and TRM. The presence of GvHD greatly increases the risk of TRM and morbidity post-SCT from infections. GvHD prophylaxis with a combination of ciclosporin and methotrexate is superior to ciclosporin alone. T-cell depletion of the allograft effectively abrogates GvHD, in particular in the VUD-SCT setting, but it is accompanied by a relapse rate of over 60 per cent. The observation that removal of donor T cells greatly increased the incidence of relapse was the unequivocal proof of the existence of a graft-versus-leukaemia (**GvL**) effect.

Current results of VUD-SCT from the Seattle group suggest an LFS of 74 per cent at 5 years in CML patients who are under the age of 50 years and are transplanted within a year of diagnosis. These and other similar results have led to further refinements toward the search for suitable VUDs to make SCT more available. Newer molecular techniques for subtyping the HLA class I genes should improve the chances of finding optimally matched or acceptably mismatched donors. Syngeneic SCT has comparable overall survival, but, due to the lack of a GvL effect, there is a higher relapse rate resulting in a lower LFS.

Efforts to minimize the toxicity of conditioning regimens have been benefited by the use of the purine analogues (for example, fludarabine) which are potent immunosuppressive drugs. These regimens are non-myeloablative but ensure engraftment and are designed to exploit maximally the GvL effect. These procedures have been termed reduced intensity conditioning SCTs (also mini-SCTs or non-myeloablative SCTs) and reflect the exciting advances in our understanding of how SCT actually works.

Treatment of relapse of CML post-transplantation

Most patients who relapse after allogeneic SCT (allo-SCT) do so within the first 3 years. This relapse tends to follow an orderly progression, with the patient initially demonstrating evidence of a molecular relapse with increasing positivity of BCR–ABL transcripts assayed by the polymerase chain reaction (**PCR**), followed by a cytogenetic relapse (finding of the Ph chromosome in marrow metaphases), and then by haematological and clinical relapse. Molecular monitoring of allo-SCT recipients is therefore valuable. For patients with molecular relapse, remission can be induced by withdrawing immunosuppression or by the infusion of lymphocytes collected from the original transplant donor (**DLI**) without any other antileukaemic measures. DLI can induce remissions in 60 to 80 per cent of patients with molecular or cytogenetic relapse. These important results lend further support to the concept that a GvL effect plays an important role in the cure of CML after allografting. Patients who fail to enter remission with DLI may be candidates for treatment with STI571 or a second allo-SCT but the risk of TRM is relatively high. The 4-year LFS for a second allo-SCT is around 28 per cent. The potential benefit of using STI571 initially instead of DLI is now being assessed.

The mechanisms by which T lymphocytes exert a GvL effect remain highly speculative. It is possible that they release cytokines, such as interleukin-2, IFN-a, or transforming growth factor-a, that selectively suppress the proliferation of Ph-positive cells. It is possible that T cells or natural killer cells act directly against leukaemia cells. A third possibility is that CML cells express leukaemia-specific antigens, possibly coded by the *BCR–ABL* chimeric gene, that provoke a true leukaemia-specific, T-cell response.

Autologous stem-cell transplantation

Despite the qualified success of allo-SCT, the majority of CML patients are not eligible for this therapy and a substantial number have a marginal survival benefit from IFN-a treatment. Autologous SCT following high-dose chemotherapy has a lower TRM and is available to more patients. Retrospective analyses suggest that autografting with blood- or marrow-derived stem cells can prolong survival. The fact that some Ph-negative stem cells survive at the time of diagnosis in most patients provides the rationale for developing techniques that favour reconstitution with Ph-negative haemopoiesis. The Genoa group has pioneered procedures where Ph-negative stem cells are harvested during the recovery phase after chemotherapy; they demonstrated successful engraftment resulting in Ph-negative haemopoiesis. In most cases, however, the Ph-positive haemopoiesis recurs. This recurrence may be due in some cases to residual Ph-positive cells in the autografted material and this provides the rationale for 'purging' techniques. Various *in vitro* and *in vivo* methods have been developed with variable degrees of success. The new generation of *in vitro* purging studies using tyrosine kinase inhibitors targeted against the *BCR–ABL* oncoprotein, such as STI571, may be informative. Although the clinical feasibility and safety of these differing autografting strategies have been demonstrated, the precise therapeutic role of autografting remains unclear.

Conclusions and a suggested therapeutic algorithm

The choice of therapy for the younger CML patient newly diagnosed in chronic phase now requires the benefits and risks of SCT to be balanced against the predicted results of non-transplant therapy. The decision to offer an allogeneic SCT early after diagnosis to a patient under the age of 40 years, in particular if an HLA-matched sibling donor is available, is probably straightforward. For most other patients it is useful to assess all the treatment options carefully. For the present, patients over the age of 60 years cannot safely be treated by SCT, but this may change in the near future following further experience with non-myeloablative SCT. For patients aged between 40 and 60 years who have a suitable molecularly matched unrelated donor and wish to be transplanted, the risk of TRM is such that a trial of STI571 in the first instance is reasonable. We suggest this form of an integrated approach in the therapeutic algorithm shown in Fig. 4. It is likely that some of the other new treatment strategies currently being investigated, such as immunotherapy to activate a leukaemia-specific immune response, will also be integrated in the treatment plans of patients with CML in the near future.

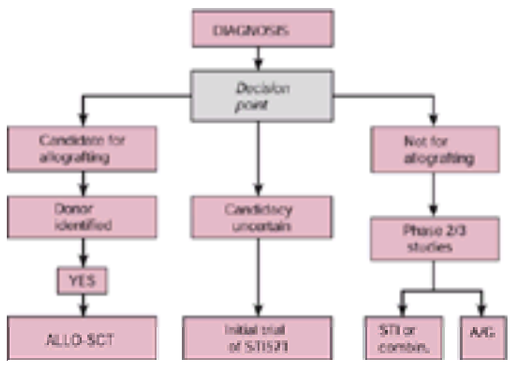

Fig. 4 A suggested therapeutic algorithm for the management of patients with CML.

Further reading

Ahuja H, *et al*. (1991). The spectrum of molecular alterations in the evolution of chronic myeloid leukemia. *Journal of Clinical Investigation* **87**, 2042–6.

Allan NC, Richards SM, Shepherd PCA (1995). UK Medical Research Council randomised multicentre trial of interferon-an1 for chronic myeloid leukaemia: improved survival irrespective of cytogenetic response. *Lancet* **345**, 1392–7.

Barrett AJ, Malkovska V (1996). Graft-versus-leukaemia: understanding and using the allo-immune response to treat haematological malignancies. *British Journal of Haematology* **93**, 754–61.

Barrett AJ, van Rhee F (1997). Graft-versus-leukemia. *Baillière's Clinical Haematology* **10**, 337–56.

Bolin RW, *et al*. (1982). Busulfan versus hydroxyurea in the long-term therapy of chronic myelogenous leukemia. *Cancer* 50, 1683–7.

Carella AM, *et al*. (1997). Mobilization and transplantation of Philadelphia-negative peripheral blood progenitor cells early in chronic myelogenous leukemia. *Journal of Clinical Oncology* **15**, 1575.

Collins RH, *et al*. (1997). Donor leukocyte infusions in 140 patients with relapse malignancy after allogeneic bone marrow transplantation. *Journal of Clinical Oncology* **15**, 433.

Devergie A, *et al*. (1995). Allogeneic bone marrow transplantation for chronic myeloid leukemia in first chronic phase: a randomized trial of busulfan–cytoxan versus cytoxan–total body irradiation as preparative regimen: a report from the French Society of Bone Marrow Graft (SFGM). *Blood* **85**, 2263.

Druker BJ, *et al*. (2001). Efficacy and safety of a specific inhibitor of the Bcr-Abl tyrosine kinase in chronic myeloid leukemia. *New England Journal of Medicine* 344, 1031-1037

Druker BJ, *et al*. (1996). Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of BCR–ABL positive cells. *Nature Medicine* **2**, 561.

Faderl S, *et al*. (1999). The biology of chronic myeloid leukemia. *New England Journal of Medicine* **341**, 164–72.

Fialkow PJ (1981). Evidence for a multistep origin of chronic myeloid leukemia. *Blood* **58**, 158–63.

Goldman JM, *et al*. (1986). Bone marrow transplantation for patients with chronic myeloid leukemia. *New England Journal of Medicine* **314**, 202.

Goldman JM (1998). Cost effectiveness of interferon-a for chronic myeloid leukaemia. *Annals of Oncology* **9**, 351–2.

Goldman JM (1999). Donor lymphocyte infusion for chronic myelogenous leukemia. *Blood* **94** (Suppl. 1), 60.

Goldman JM (2000). Tyrosine-kinase inhibition in treatment of chronic myeloid leukaemia. *Lancet* **355**, 1031–2.

Gratwohl A, *et al*. (1998). Risk assessment for patients with chronic myeloid leukaemia before allogeneic blood or marrow transplantation. *Lancet* **352**, 1078–92.

Gratwohl A, *et al*. (1986). Bone marrow transplantation for chronic myeloid leukemia: long-term results. *Bone Marrow Transplantation* **12**, 509.

Groffen J, Heisterkamp N (1997). The chimeric *BCR–ABL* gene. *Baillière's Clinical Haematology* **10**, 187.

Guilhot F, *et al*. (1997). Interferon alpha-2b combined with cytarabine versus interferon alone in chronic myelogenous leukemia. *New England Journal of Medicine* **337**, 223.

Hansen JA, *et al*. (1998). Bone marrow transplantation from unrelated donors for patients with chronic myeloid leukemia. *New England Journal of Medicine* **338**, 962.

Hasford J, *et al*. (1998). A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. *Journal of the National Cancer Institute* **90**, 850–8.

Holyoake T, *et al*. (1999). Isolation of a highly quiescent subpopulation of primitive leukemic cells in chronic myeloid leukemia. *Blood* **94**, 2056–64.

Horowitz MM, *et al*. (1996). Effect of prior interferon therapy on outcome of HLA-identical sibling bone marrow transplants for chronic myelogenous leukemia (CML) in first chronic phase. *Experimental Hematology* **24**, 1143.

Horowitz MM, Rowlings PA, Passweg JR (1996). Allogeneic bone marrow transplantation for CML: a report from the International Bone Marrow Transplant Registry. *Bone Marrow Transplantation* **17**(Suppl. 3), S5–6.

Huntly BJ, *et al*. (2001) Deletions of the derivative chromosome 9 occur at the time of the Philadelphia translocation and provide a powerful and independent prognostic indicator in chronic myeloid leukemia. *Blood* **98**, 1732–38.

Ichimaru M, Ishimaru T, Belsky JL (1978). Incidence of leukemia in atomic bomb survivors belonging to a fixed cohort in Hiroshima and Nagasaki, 1950–1971: radiation dose, years after exposure, age at exposure, and type of leukaemia. *Journal of Radiation Research* **19**, 262.

Kantarjian HM, *et al*. (1995). Prolonged survival in chronic myelogenous leukemia after cytogenetic response to interferon-a therapy. *Annals of Internal Medicine* **122**, 254–61.

Lee S, *et al*. (1998). Initial therapy for chronic myelogenous leukemia: playing the odds. *Journal of Clinical Oncology* **16**, 2897–903.

Lee SJ, *et al*. (1997). Unrelated donor bone marrow transplantation for chronic myelogenous leukemia: a decision analysis. *Annals of Internal Medicine* **127**, 1080–8.

Lin F, *et al*. (1996). Kinetics of increasing BCR–ABL transcript numbers in chronic myeloid leukemia patients who relapse after allogeneic bone barrow transplantation. *Blood* 87, 4473.

McGlave P, *et al*. (1994). Autologous transplant therapy for chronic myelogenous leukaemia prolongs survival: results from eight transplant groups. *Lancet* **343**, 1486.

Marmont A, *et al*. (1984). Recurrence of Ph'-leukemia in donor cells after marrow transplantation for chronic myelogenous leukemia. *New England Journal of Medicine* **310**, 903.

Melo JV (1996). The diversity of the BCR–ABL fusion proteins and their relationship to leukemia phenotype. *Blood* **88**, 2375.

Mughal TI, Goldman JM (1995). Chronic myeloid leukaemia: a therapeutic challenge. *Annals of Oncology* **6**, 637–44.

Mughal TI, Hoyle C, Goldman JM (1993). Autografting for patients with chronic myeloid leukemia—The Hammersmith experience. *Stem Cells* **11**, 20–2.

Mughal TI, *et al*. (2001). Molecular studies in patients with chronic myeloid leukaemia in remission 5 years after allogeneic stem cell transplant define the risk of subsequent relapse. *British Journal of Haematology* **115**, 569–74.

Nowell PC, Hungerford DA (1960). A minute chromosome in human chronic granulocytic leukemia. *Science* **132**, 1497.

Pane F, *et al*. (1996). Neutrophilic-chronic myeloid leukemia: a distinct disease with a specific molecular marker. *Blood* **88**, 2410–14

Sawyers CL (1999). Chronic myeloid leukemia *New England Journal of Medicine* **340**, 1330–8.

Schofield JR, *et al*. (1994). Low doses of interferon-a are as effective as higher doses in inducing remissions and prolonging survival in chronic myeloid leukemia. *Annals of Internal Medicine* **121**, 736.

Shepherd P, *et al*. (1995). Analysis of molecular breakpoint and mRNA transcripts in a prospective randomized trial of interferon in chronic myeloid leukaemia: no correlation with clinical features, cytogenetic response, duration of chronic phase, or survival. *British Journal of Haematology* **89**, 546–54.

Silver RT, *et al*. (1999). An evidence-based analysis of the effect of busulfan, hydroxyurea, interferon, and allogeneic bone marrow transplantation in treating the chronic phase of chronic myeloid leukemia: Development for the American Society of Hematology. *Blood* **94**, 1517–36

Slavin S, *et al*. (1998). Nonmyeloablative stem cell transplantation with cell therapy as an alternative to conventional bone marrow transplantation with lethal cytoreduction for the treatment of malignant and non-malignant hematologic diseases. *Blood* **91**, 756.

Sokal JE, *et al*. (1984). Prognostic discrimination in 'good-risk' chronic granulocytic leukemia. *Blood* **63**, 789.

Spencer A, *et al*. (1995). Cytotoxic T-lymphocyte precursor frequency analysis in bone marrow transplantation with volunteer unrelated donors: value in donor selection. *Transplantation* **59**, 1303.

Spencer A, *et al*. (1995). Bone marrow transplantation for chronic myeloid leukemia with volunteer unrelated donors using ' *ex vivo*' or '*in vivo*' T-cell depletion: a major prognostic impact of HLA class I identity between donor and recipient. *Blood* **86**, 3590.

Szydlo R, *et al*. (1997). Results of allogeneic bone marrow transplants using donors other than HLA-identical siblings. *Journal of Clinical Oncology* **15**, 1767.

Talpaz M, *et al*. (1998). The MD Anderson Cancer Center experience with interferon-a therapy in chronic myelogenous leukemia. *Baillière's Clinical Haematology* **10**, 291.

The Chronic Myeloid Leukemia Trialists' Collaborative Group (1997). Interferon versus chemotherapy for chronic myeloid leukemia: a meta-analysis of seven randomized trials. *Journal of the National Cancer Institute* **89**, 1616–20.

Tura S (1998). Cytarabine increases karyotypic response in alpha-IFN treated chronic myeloid leukemia patients: results of a national prospective randomized trial. *Blood* **92**, 317a.

van Rhee F, *et al*. (1994). Detection of residual leukemia more than 10 years after allogeneic bone marrow transplantation. *Bone Marrow Transplantation* **14**, 609.

22.3.7 Myelodysplasia

Lawrence B. Gardner and Chi V. Dang

Definition

The myelodysplastic syndromes are a collection of acquired, clonal, haemopoietic disorders characterized by cytopenias and abnormal cellular morphologies. The vast majority of myelodysplastic syndromes (**MDS**) are marked by progressive, multilineage cytopenias, ineffective maturation of cells with dysplastic appearances and chromosomal abnormalities, and a tendency to degenerate into poorly responsive leukaemias. Historically, these syndromes have been referred to as preleukaemias, smouldering leukaemias, and refractory anaemias. In 1976 the French–American–British (**FAB**) Cooperative Group proposed a classification system of these heterogeneous disorders based on the histological appearance of the peripheral blood and bone marrow, with primary emphasis placed on the percentage of immature cells, or myeloblasts, present. This artificial classification has provided a helpful outline for physician communication and research. However, the presentation and prognosis of individuals with MDS vary greatly, depending on the biological impact of the genetic mutation(s) present in an individual's clone. MDS is more common in adults than any acute or chronic leukaemia. While MDS can occur at any age, it is primarily a disease of the old; over 80 per cent of those affected are older than 60 years, and after age 70 there is a prevalence of approximately 33/100 000. The incidence of the disease appears to be increasing, probably in part due to more common screening with complete blood counts, and an increase in secondary, treatment-related MDS. There is currently no effective cure for MDS, except for allogeneic bone marrow transplantation, which is often not an option for the older patient with MDS. Treatment therefore remains supportive, consisting of transfusion and antibiotic treatment for documented infections. Patients often die as a result of their cytopenias (for example, from bleeding or infection) or transformation to leukaemia.

Pathogenesis and pathophysiology

The biology of MDS is difficult to study, since disorders grouped under MDS probably include diagnoses of disparate aetiologies. It is clear from karyotypic analysis, including chromosomal studies, and X-linked inactivation of genetic markers, that MDS is a clonal disorder involving a defect in an early haemopoietic progenitor cell. Thus erythrocytes, platelets, neutrophils, and monocytes may all be affected in MDS. The ability of progenitor cells from patients with MDS to form colonies *in vitro* is markedly diminished. Haemopoietic growth-factor production by lymphocytes from these patients is often decreased, and growth-inhibitory cytokines may be increased. In addition to the abnormal growth characteristics found in MDS progenitor cells in culture, there appears to be a higher rate of programmed cell death, or apoptosis, in such cells, which may contribute to the peripheral cytopenias and ineffective haemopoiesis noted on bone marrow examination.

Chromosomes are often abnormal in MDS. In both therapy-related and *de novo*-acquired MDS, specific chromosomal abnormalities (including deletions of chromosomes 5 and 7, and trisomy 8) are relatively common. A specific chromosomal translocation resulting in an oncogenic fusion protein is often found in a subtype of chronic myelomonocytic leukaemia (**CMMoL**). While the regions of chromosome 5 and 7 often deleted in MDS contain several genes important for haemopoiesis, including granulocyte–macrophage colony-stimulating factor (**GM-CSF**), erythropoietin, interleukin-6 (**IL-6)**, and the receptors for several haemopoietic growth factors, no single gene, or group of genes, has been found to be consistently mutated in MDS. In addition, the pathogenic importance of specific deletions has not been well established. A number of genes important for proliferation and apoptosis have been described to be abnormally expressed in MDS, including the mutated *ras* oncogene which is present in up to 30 per cent of cases, but the importance of these abnormalities in the causation of MDS is unclear. Progression of MDS is often associated with the accumulation of additional chromosomal abnormalities. This stepwise accumulation of mutations, often found in many types of cancers, suggests the dominance of new clones with a proliferation and/or survival advantage.

Primary acquired sideroblastic anaemia is a unique subset of MDS. Other causes of an anaemia with the morphological appearance of ringed sideroblasts include an X-linked inherited form, and a secondary toxic form. Inherited and secondary sideroblastic anaemias are believed to result from a disruption in haem synthesis, producing ineffective haemopoiesis and iron overload. Alcohol, isoniazid, and pyrazinamide are among the common medications that can cause acquired sideroblastic anaemia.

Clinical features

While MDS has been described in children, it is uncommon and other congenital haemopoietic diseases should be strongly considered. And while secondary MDS may occur in younger adults, MDS is primarily a disease of the older adult. The clinical presentation of a patient with MDS depends on the specific cytopenias present, and the extent to which a lineage is depressed. Most commonly, patients present due to a symptomatic anaemia. Because the onset of MDS is gradual and progressive, patients typically present with signs and symptoms of chronic, not acute, anaemia, including fatigue and exertional dyspnoea. If the platelet count is are low, petechiae or other forms of bleeding, typically mucosal or gastrointestinal, may be present. Infections occur with suppressed numbers of white cells, particularly when the absolute neutrophil count is below 500/ml. One form of MDS, CMMoL, shares many characteristics with myeloproliferative diseases. In CMMoL the monocyte count can be quite high, resulting in pleural, pericardial, and peritoneal effusions, as well as splenomegaly. Rheumatological and autoimmune processes have been noted to occur with MDS.

Laboratory diagnosis

Because MDS is a chronic disease, obtaining old laboratory data documenting a progressive, often macrocytic anaemia, thrombocytopenia, and leucopenia, can be invaluable in making the diagnosis. All patients diagnosed with MDS have an anaemia, which is typically either normocytic or macrocytic, although extreme macrocytosis (mean corpuscular volume (**MCV**) >120 fl) is not common. Anaemia and leucopenia in the setting of a normal or elevated platelet count should lead one to consider a subtype of MDS, the 5q-syndrome. The hallmark of MDS is dysplasia, which is often noted on a peripheral blood smear (Fig. 1(a)). Erythrocytes may show anisocytosis (varying sizes), poikilocytosis (abnormal morphology) with bizarre shapes, and basophilic stippling. Polymorphonuclear neutrophils may be hyperlobulated or, more commonly, hypolobulated, sometimes showing the characteristic bilobed appearance of pseudo-Pelger–Huët cells. Hypogranulation may be present, and chromatin may be abnormally clumped. Myeloblasts may be seen in the periphery, and are a poor prognostic sign.

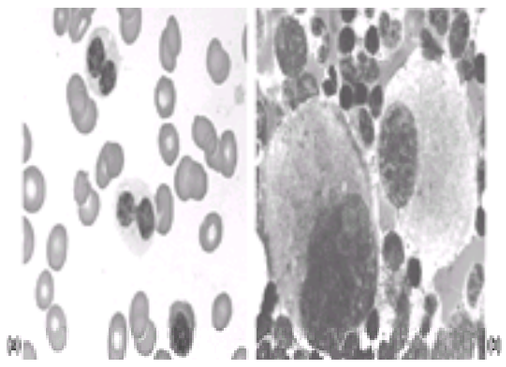

Fig. 1 (a) Peripheral blood smear and (b) bone marrow appearance in MDS. (a) Several dysplastic neutrophils, including a pseudo-Pelger–Huët neutrophil with a bilobed nucleus. The chromatin in the neutrophils is clumped, and the red cells show a range of sizes and appearances. (b) A bone marrow aspirate with a small, monolobed megakaryocyte, typical in MDS.

A bone marrow aspirate and biopsy, along with cytogenetics, are crucial for the diagnosis of MDS (Fig. 1(b)). Although the bone marrow biopsy may reveal a hypocellular bone marrow, this is rare, and should lead one to question the diagnosis and consider aplastic anaemia as the cause of a cytopenia. Typically, the bone marrow biopsy reveals a hypercellular marrow, consistent with the ineffective haemopoiesis common in MDS. The morphology of early progenitor cells shows a lack of maturation, with abnormal forms. Dyserythropoiesis, with multilobed erythroid progenitors, may be present. Megaloblastoid erythropoiesis is common; this is evidenced by an asynchrony of nuclear/cytoplasmic maturation, so that haemoglobin synthesis occurs while the erythroid nucleus is large and young. Megakaryocytes may be diminished, and/or be small and hypolobulated The bone marrow biopsy may show an abnormal localization of immature myeloid precursors (**ALIP**); typically, granulopoiesis occurs in a paratrabecular location, but in MDS there may be a shift to a central intratrabecular site. Depending on the stage of MDS, the number of blasts may be elevated. According to the FAB classification scheme, more than 30 per cent blasts in the marrow defines acute leukaemia, while less than 30 per cent are consistent with MDS. Staining the bone marrow biopsy for CD34+ cells may be helpful in the diagnosis.

Chromosomal abnormalities (deletions, additions, translocations) are very common in MDS. In MDS as a whole, chromosomal abnormalities are found 40 per cent of the time. While MDS may be associated with a normal karyotype, especially in an early stage with few blasts, some chromosomal abnormalities are so typical that their presence strongly suggests the diagnosis. Similarly, multiple complex chromosomal abnormalities leads one to strongly consider a diagnosis of MDS. Typically, the longer a patient has MDS, the more chromosomal abnormalities may occur, leading to a more aggressive clone. The most common abnormalities include monosomy 7, 7q-, monosomy 5, 5q-, trisomy 8, and 20q-. 11q23 is commonly found in secondary MDS due to treatment with topoisomerase II inhibitors, such as the epipodophyllotoxin, VP16 (etoposide). As discussed below, chromosomal abnormalities are important not only in suggesting the diagnosis of MDS, but also in its prognosis. Other diagnostic tests, such as abnormal growth of progenitor colonies in *in vitro* assays, are not widely utilized.

Differential diagnosis

When working up the possible aetiologies of a mild asymptomatic anaemia or pancytopenia in the older patient, where making a diagnosis may not change management, the extent of the work-up should depend on the patient's wishes. However, certain reversible diseases, several of which may be significant to the patient's overall health, must be ruled out. Aplastic anaemia also presents with pancytopenia; however, in contrast to MDS, the bone marrow is hypocellular, CD34+ early progenitor stem cells in the bone marrow are relatively diminished, and there is little evidence of dysplasia. In paroxysmal nocturnal haemoglobinuria the absence of the marker CD59 on the surface of cells, a lack of dysplasia, and low iron stores should differentiate this disease from MDS. Macrocytic, megaloblastoid anaemias as well as pancytopenias are common in patients with vitamin B12 and folate deficiency. These may be ruled out with serum and red cell measurements, respectively. Additionally, these defects should respond rapidly to replacement therapy. The anaemia of chronic disease is primarily a clinical diagnosis, but little dysplasia should be present. Alcoholism and/or hypersplenism can result in a mild pancytopenia, and an abnormal physical examination and normal bone marrow biopsy will rule these out. Bone marrow infiltration by a tumour or fibrosis usually presents with a myelophthisic blood smear consisting of nucleated red cells, teardrop red-cell forms, and a left-shifted myeloid series. Although both MDS and myeloproliferative diseases may present with a hypercellular marrow, it should not be difficult to delineate these two. Myeloproliferative disorders are not marked by dysplasia and bone marrow failure, but by increased proliferation and usually elevated cell counts.

Classification

The 1976 FAB classification divided MDS into five subgroups based on the morphological appearance of the peripheral blood and bone marrow (Table 1). While newer prognostic systems, including a modification of the FAB classification recently proposed by the World Health Organization, are becoming more popular, and aspects of the FAB MDS classification, especially the inclusion of CMMoL, have been criticized, this classification system remains helpful clinically and is the basis of patient stratification in most recent MDS research. In many patients, there is a gradual progression through the subgroups, eventually leading to acute leukaemia.

Refractory anaemia

Refractory anaemia accounts for approximately 25 per cent of all cases of MDS. By definition, less than 5 per cent blasts are present in the bone marrow, and less than 15 per cent ringed sideroblasts are seen with iron staining. Typically, dysplasia in the peripheral blood and bone marrow are minimal, and there are few chromosomal abnormalities. These patients have a relatively low risk of progressing to leukaemia, and may do well for prolonged periods of time.

Refractory anaemia with ringed sideroblasts (RARS)

Ringed sideroblasts are erythroblasts with iron-laden mitochondria encircling more than one-third of the nucleus. More than six Prussian Blue-stained iron granules must be noted, in more than 15 per cent of the cells to make the diagnosis of RARS. In addition, fewer than 5 per cent blasts must be found in the bone marrow. As alluded to above, several drug-induced and hereditary syndromes may also present with ringed sideroblasts: for example, alcohol- and isoniazid-induced and X-linked disease, respectively. It is important to differentiate these states, as their prognosis and treatment may be different than for RARS. For example, inherited RARS may sometimes be successfully treated with pyridoxine, and the most common complication is usually iron overload. RARS, like refractory anaemia, has a relatively low risk of progression to acute leukaemia (approximately 10 per cent), especially when only the erythroid series is suppressed. Median survival is 50 months.

Refractory anaemia with excess blasts (RAEB)

RAEB accounts for approximately 25 per cent of cases of MDS. While ringed sideroblasts may be present, a diagnosis of RAEB is made when the bone marrow contains 5 to 20 per cent blasts. Multiple lineages of cells are usually affected, and chromosomal abnormalities are often found. RAEB has a high rate of progression to acute leukaemia and of bone marrow failure.

Refractory anaemia with excess blasts in transformation (RAEBt)

A diagnosis of RAEBt is made when there are more than 5 per cent blasts in the peripheral blood, or 21 to 30 per cent blasts in the bone marrow. Auer rods (that is, abnormal, oblong lysosomes) may be present in these blasts. These patients have multiple chromosomal abnormalities and a dismal prognosis, with most progressing to acute leukaemia. Obviously the decision that 29 per cent blasts in the marrow confers a diagnosis of MDS, while 31 per cent blasts meet the criteria of leukaemia is artefactual and arbitrary. Even some cases of RAEB, where the bone marrow blasts are less than 20 per cent, clinically behave as—and might be better considered—an evolving acute leukaemia. Thus a proportion of RAEBt cases may be acute leukaemia diagnosed relatively early. The classification of RAEBt is most helpful in those with documented long-term cytopenias and MDS, and characteristic chromosomal abnormalities. Mortality of patients with RAEBt is not dramatically different than those with acute leukaemia, and in fact survivors whose acute leukaemia has evolved from RAEBt have a higher rate of relapse after aggressive chemotherapy.

Chronic myelomonocytic leukaemia (CMMoL)

CMMoL, characterized by fewer than 20 per cent bone marrow blasts and a peripheral monocytosis of more than 1000 monocytes/μl, has many similarities to myeloproliferative diseases such as chronic myelogenous leukaemia. The white blood cell count is typically very elevated, marrow fibrosis can occur, and extramedullary diseases (hepatosplenomegaly, skin) and serositis are common. Splenomegaly is found in approximately 20 per cent of the cases. However, similar to other MDS types, trilineage dysplasia is typically evident. Prognosis best correlates with the percentage of bone marrow blasts, not with the degree of peripheral monocytosis. Approximately 25 per cent of patients progress to acute leukaemia, and death due to cytopenia is common.

Other subtypes of MDS- 5q- and secondary MDS

The 5q- syndrome is a unique subtype of MDS with specific morphological, laboratory, and clinical characteristics. Platelet counts are typically normal, or even elevated. Megakaryocytes are small and hypolobulated. When the only chromosomal abnormality is a deletion of 5, patients have an excellent prognosis with a low risk of transforming to a leukaemic state.

Secondary, or treatment-related MDS, is becoming more prevalent. This is probably due to several factors. Patients with solid malignancies are being treated with more aggressive chemotherapeutic regimens, and they are living longer after these treatments. In addition, it has been postulated that haemopoietic growth-factor support during intensive chemotherapy may be a contributing factor to the development of MDS. Most cases of secondary MDS present within the first decade after treatment. Chromosomal abnormalities are common, occurring more than 90 per cent of the time. Chromosomes 5 or 7, are typically involved, with monosomy 7 occurring in 60 per cent of the cases. Alkylating agents and topoisomerase II inhibitors are the most commonly implicated in causing secondary MDS.

Treatment

With the exception of allogeneic transplant there is no curative treatment for MDS. In addition, once acute leukaemia has evolved from MDS, treatment with aggressive chemotherapy does not usually result in long-term, curative, remission. Although there are several treatments for MDS, there are little data to suggest that drug treatment of MDS prolongs survival over that of standard, supportive care. However, treatment can improve quality of life, and new treatments are being explored.

Because of the limits of therapy, there is no need to treat asymptomatic patients, except when an allogeneic bone marrow transplant is clinically possible and the patient wishes to undergo such a procedure. Thus, mild anaemia and thrombocytopenia without bleeding do not necessitate transfusion. There is no evidence that prophylactic antibiotics are beneficial. However, for symptomatic anaemia, or anaemia in the older patient with cardiovascular disease, and for severely thrombocytopenic patients with episodes of bleeding or who are at high risk for significant bleeds, supportive transfusions are the mainstay of care. Patients may need regular transfusions. It is important to recognize patients who will live for long periods with red cell transfusion; such patients should have their iron status followed, and be initiated on iron chelation therapy to avoid the side-effects of haemosiderosis.

A wide range of myeloablative chemotherapeutic regimens have been explored; for instance, regimens commonly used to treat acute myelogenous leukaemia (**AML**), including aplasia-inducing doses of cytosine arabinoside (**Ara-C**), anthracyclines, cytoxan, and topotecan. A few studies have suggested that selected patients with good-risk characteristics may do as well with such regimens as patients with AML, but the results have usually been disappointing. Complete responses have ranged from 10 to 50 per cent, but these are generally of short duration and accompanied by significant morbidity and mortality. This is probably due to several factors, including the relatively older age of most patients with MDS, the drug resistance of MDS due to the increased expression of multidrug resistance proteins, and limited reserves of normal, healthy marrow for recovery. Low-dose chemotherapy has also been used. While initially explored because these dosing regimens are better tolerated in the older MDS population, many of these agents may have differentiating as well as cytotoxic effects. Low-dose Ara-C, 5-azacytidine, and topotecan have been used. These have resulted in complete responses from 10 to 40 per cent, but again these responses are not durable. Trials with low-dose Ara-C or 5-azacytidine showed no improvement in survival over supportive care. There is still significant scientific enthusiasm for exploring other dosing regimens, other differentiating agents, alone or in combination with growth factors and cytotoxic agents, in the treatment of MDS. For CMMoL, hydroxycarbamide (hydroxyurea) and the control of peripheral monocytosis has been shown to be as effective as aggressive chemotherapy.

Although most patients with MDS are ineligible for an allogeneic stem-cell transplant (either because of their age, comorbid disease, or lack of suitable donor), for some this remains a viable option and hope for cure. Bone marrow transplantation has been successively carried out in highly selected 55- to 66-year-old patients with MDS. The 5-year survival rate for all patients undergoing transplantation ranges from 30 to 70 per cent, but early mortality is common. Patients with refractory anaemia and RARS and younger patients have the best outcomes with transplantation. Normal or good chromosomal abnormalities are also predictive of a better response with a transplant. All high-risk (see next section) young patients should be considered for allogeneic bone marrow transplantation.

Multiple trials have utilized haemopoietic growth factors, such as erythropoietin and granulocyte colony-stimulating factor (**G-CSF**) (or granulocyte–macrophage colony-stimulating factor, **GM-CSF**), sometimes in combination. Short-term (1–2 weeks) and prolonged treatment with erythropoietin and G-CSF or GM-CSF do not appear to increase the progression to acute leukaemia. In a majority of patients, neutrophil counts increase, sometimes with a documented decrease in infection, and there are often improvements in red cell and platelet counts with decreased transfusion requirements. Some patients may actually respond with a decreased platelet count, and some patients may not tolerate some of the side-effects of the injections. Current evidence suggests that overall survival does not appear to be improved. Generally higher doses of erythropoietin (>200 U/kg per day) may be necessary, and patients with lower serum erythropoietin levels (<500 mU/ml) tend to have better responses. Laboratory data has suggested that the combination of erythropoietin and G-CSF may be synergistic in promoting the growth of haemopoietic progenitor colonies. Several clinical studies have suggested that the combination of erythropoietin and G-CSF is more effective than either alone, especially in patients with low serum erythropoietin levels and in those with ringed sideroblasts. Ciclosporin, an immunosuppressant, has also recently been reported as effective in MDS, especially when the marrow is hypocellular. In general, treatment choices should be based on the patient's performance status and age, the prognosis of the disease, and, whenever possible, in the setting of a clinical research protocol.

Prognosis

The median survival of all those with MDS has been reported to be between 12 and 28 months. However, since MDS consists of a variety of diseases, prognosis varies widely. A number of prognostic factors have been studied. Many of the factors which have proven to be predictive in prospective and retrospective studies are intuitive. Because the most common causes of death in patients with MDS are transformation to leukaemia or symptomatic cytopenias, a poorer prognosis is seen in those with increased bone marrow blasts and with more severe cytopenias. Other important prognostic factors include age and specific karyotypic abnormalities.

Because the FAB classification is for the most part dependent on the percentage of bone marrow blasts, the FAB subtypes closely correlate with both overall survival and evolution to acute leukaemia (Fig. 2). Those with refractory anaemia (**RA**) and RARS have a low incidence of leukaemia, both within the first 2 years and overall, and an improved survival rate. Patients with increased blasts, seen in RAEB and RAEBt do worse, and the overall survival in RAEBt is not significantly different than AML. Virtually all patients with RAEBt evolve to acute leukaemia within 30 months. Univariate analysis has also indicated that those with two or three cytopenias do worse than those with none or just one cytopenia, probably because those with more cytopenias have increased blasts, and are more prone to bleeding and infection. Marrow cytogenetics have also been found to be important. A relatively good prognosis is found in deletions of 5q, 20q, -Y in men, or normal cytogenetics. Of note, when these deletions are accompanied with other chromosomal abnormalities, they do not connote a good prognosis. Those with complex chromosomal abnormalities, or deletions of 7 do particularly poorly, with a high progression to acute leukaemia.

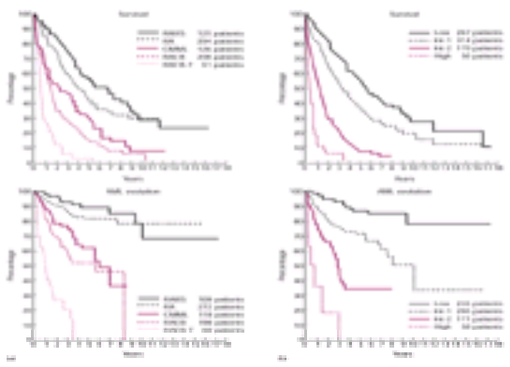

Fig. 2 Survival and evolution to leukaemia in MDS in groups defined by FAB classification (a) and IPSS (b). RA, refractory anaemia; RARS, refractory anaemia with ringed sideroblasts; RAEB, refractory anaemia with excess blasts; RAEBt, refractory anaemia with excess blasts in transformation; CMMoL, chronic myelomonocytic leukaemia. IPSS as defined in Table 2. (Figures taken from Greenberg P, *et al.* (1997). International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood* **89**, 2079–88, with permission from the publisher.)

Several of these prognostic indicators have been combined in various scoring systems to predict the survival of patients with MDS. One of the most accurate, and most widely used, is the International Prognostic Scoring System (**IPSS**) (Table 2). The IPSS assigns points for unfavourable characteristics, such as unbalanced chromosomal translocations, percentage of blasts in the bone marrow, and lineages affected by the MDS. Although the IPSS is somewhat cumbersome to use for clinicians, and no scoring system is perfect for individual patients, the IPSS is useful for investigators and for making general decisions regarding the aggressiveness of treatment.

Further research

Increasing information on normal stem-cell biology and haemopoiesis will clearly lead to increased understanding of the abnormal haemopoietic development seen in MDS. Particular areas of research being explored include the role of apoptosis in MDS, and specific genetic mutations (or groups of mutations) that are necessary for the development of MDS. The importance of individual mutations or chromosomal abnormalities for prognosis is still being explored. A major emphasis in the stem-cell transplantation field is on increasing the availability of transplants. This includes making transplantation less toxic through non-myeloablative induction strategies, capitalizing on the graft-versus-tumour phenomena, and increasing the number of potential donors with international registries and by minimizing the importance of HLA barriers. Clearly these improvements will aid in the treatment of MDS. The most active area of research in specific treatment of MDS, is in the role of non-toxic differentiating agents.

Further reading

Bennett JM, *et al.* (1982). FAB Cooperative Group L Proposal for the classification of the myelodysplastic syndromes. *British Journal of Haematology* **51**, 189–99.

Cheson BD (1998). Standard and low-dose chemotherapy for the treatment of myelodysplastic syndromes. *Leukemia Research* **22**(Suppl 1), S17–21.

Deeg HJ, Appelbaum FR (2000). Hematopoietic stem cell transplantation for myelodysplastic syndrome. *Current Opinion in Oncology* **12**, 116–20.

Deeg HJ *et al*. (2000). Allogeneic and syngeneic marrow transplantation for myelodysplastic syndrome in patients 55–66 years of age. *Blood* **15**, 1188–94.

Eillman CL (1998). Molecular genetic features of myelodysplastic syndromes. *Leukaemia* **12**(Suppl 1), S2–6.

Estey EH (1998). Prognosis and therapy of secondary myelodysplastic syndromes. *Haematologica* **83**, 543–9.

Greenberg P (2000). Myelodysplastic syndrome. In: Hoffman R, *et al.*, eds. *Hematology: basic principles and practice*, pp 1106–29. Churchill Livingstone, New York.

Greenberg P, *et al.* (1997). International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood* **89**, 2079–88.

Hellstrom-Lindberg E, *et al*. (1997). Erythroid responses to treatment with G-CSF plus erythropoietin for the anaemia of patients with myelodysplastic syndromes: proposal for a predictive model. *British Journal of Haematology* **99**, 344–51.

Sole F, *et al.* (2000). Incidence, characterization and prognostic significance of chromosomal abnormalities in 640 patients with primary myelodysplastic syndromes. Grups Cooperativo Espanol de Citogenetica Hematologica. *British Journal of Haematology* **108**, 346–56.

Yoshida Y, Mufti GJ (1999). Apoptosis and its significance in MDS: controversies revisited. *Leukemia Research* **23**, 777–85.

22.3.8 The polycythaemias

David M. Gustin and Ronald Hoffman

Introduction

Polycythaemia or erythrocytosis is associated with an aetiologically distinct group of disorders characterized by an abnormal increase in the numbers of red blood cells, leading to an elevation in the haemoglobin concentration and haematocrit. Absolute polycythaemias (increased red cell mass) can be attributed to either an intrinsic defect of haemopoietic stem cells (primary) or to the stimulation of progenitor cells by excessive levels of circulating growth factors (secondary). A pathophysiological classification of polycythaemia is offered in Table 1. Patients with absolute polycythaemias should be distinguished from individuals in whom a minimally elevated haematocrit is not accompanied by a corresponding absolute increase in the red cell mass (spurious polycythaemia, stress erythrocytosis, Gaisbock's syndrome), but rather by a contraction of plasma volume.

Erythropoiesis

Erythropoietin (EPO), a 34.4-kDa glycoprotein hormone, is the primary humoral regulator of red blood cell production. Erythropoiesis occurs at a low baseline level to replace normal, senescent red blood cells. Decreased tissue oxygen tension increases erythropoietin levels by enhancing expression of the *EPO* gene. The circulating, secreted erythropoietin binds to receptors present on erythroid progenitor cells, promoting the proliferation and differentiation of erythroid precursors, ultimately resulting in an increased red cell mass. This complex process enhances the oxygen-carrying capacity of blood, leading to elevated tissue oxygen tension and down-modulating erythropoietin production. A peritubular interstitial cell is the primary site of erythropoietin production in the kidney. An oxygen-sensing haem protein present in these cells is essential for the control of erythropoietin production. It induces the transcription of EPO mRNA, a phenomenon mediated by hypoxia-inducible factors (**HIF**), a group of nucleoproteins that interact with the enhancer element of the *EPO* gene. The kidney has no preformed stores of erythropoietin. Increases in plasma levels are primarily due to new hormone synthesis. Since anaemia or hypoxia does not seem to influence, to any significant degree, the plasma clearance of erythropoietin, the control appears to occur mainly at the level of gene expression. Erythropoietin production also occurs in the liver to a lesser degree. The severity of hypoxia required to induce its production is much greater than that required in the kidney.

Oxygen transport is a complex process dependent on a number of variables, such as ambient oxygen levels, minute ventilation, lung diffusion capacity, cardiac output, red cell mass, regional blood flow, tissue capillary density, and haemoglobin–oxygen affinity. Acute changes in tissue oxygen demands or in environmental oxygen levels are compensated not only by increased erythropoietin production but also in minute ventilation, cardiac output, blood flow distribution, and haemoglobin–oxygen affinity (thro-ugh the modulation of 2,3-diphosphoglycerate (**2,3-DPG**) production). Sustained hypoxia is required for polycythaemia to occur as a compensatory mechanism.

Relative polycythaemias

A common group of disorders is marked by an elevated haemoglobin or haematocrit as the result of the contraction in plasma volume. The red cell mass remains normal. There are two major groups of patients with relative polycythaemias. The first includes patients with more acute conditions associated with significant degrees of dehydration with a consequent decrease in extracellular volume: for example, gastrointestinal fluid losses, therapeutic diuresis, endocrine disorders such as Addison's disease, and hypercalcaemia. In most cases, the presence of volume contraction is clinically obvious. The aetiology of the increase in haematocrit does not usually present a diagnostic challenge.

The second group is associated with a chronic low-level increase in the haematocrit. These patients are frequently active, hard-working, middle-aged, mildly hypertensive, obese males subjected to considerable stress who present with persistent polycythaemia. Characteristically, they appear plethoric but without any of the other typical features of polycythaemia vera. The cause for the contraction in the plasma volume is poorly understood, but autonomic dysregulation with changes in venous capacitance have been suggested.

The usual range of haemoglobin in these individuals is between 18 and 20 g/dl with haematocrits ranging from 49 to 55 per cent. Most of these patients seek medical evaluation for an unrelated condition, and are incidentally found to have increased haemoglobin and haematocrit values.

Suitable advice regarding weight reduction, control of hypertension, and smoking cessation is usually given to these patients. Evidence associating the mild elevations in viscosity encountered in this condition with episodes of thrombosis is incomplete. It seems sensible to recommend phlebotomy for those patients who have already experienced thrombotic episodes and who have persistent elevations of their haematocrits. It is more difficult to justify phlebotomy in asymptomatic individuals with mild elevations in haematocrit.

Absolute polycythaemia

Absolute polycythaemias may be classified as being primary, due to autonomous cell growth or to an enhanced response to growth factors that promote the proliferation of developing erythroid cells, or secondary, due to excessive production of erythropoietin in response to a variety of stimuli. Primary polycythaemia caused by defects in haemopoietic stem cells, is accompanied by low levels of circulating erythropoietin. Germline mutations that lead to enhanced erythropoiesis cause primary familial congenital polycythaemias. Polycythaemia vera, the most common primary polycythaemia, is caused by an acquired defect in haemopoietic stem cells resulting in an excessive proliferation of myeloid cells. By contrast, secondary polycythaemia is generally associated with elevated erythropoietin production. Raised levels of plasma erythropoietin can accompany systemic hypoxaemia, certain neoplasms, and disorders that impair oxygen delivery to tissues.

Absolute polycythaemias are accompanied by an elevated red cell mass. Documentation of such a mass to confirm the presence of an absolute polycythaemia usually requires a blood volume study with direct quantitation of both the red cell mass and plasma volume. An haematocrit greater than 60 per cent in men and greater than 55 per cent in women, however, are almost always associated with absolute erythrocytosis. In such cases it is frequently unnecessary to perform blood volume studies to be assured that the patient has an absolute polycythaemia.

Secondary polycythaemias associated with appropriate erythropoietin secretion

This group of polycythaemias encompasses a number of conditions that are ultimately the result of tissue hypoxia and subsequent excessive erythropoietin production leading to erythrocytosis. These disorders are collectively regarded as hypoxic erythrocytoses.

Hypobaric hypoxia

At high altitudes the barometric pressure and, consequently, the ambient oxygen tension are reduced, resulting in alveolar and arterial hypoxia. Natives of the Andes mountains who live above 4200 metres have been reported to have haematocrits 30 per cent higher than individuals who live at sea level. Acutely, changes in minute ventilation, heart rate, blood flow, and haemoglobin–oxygen affinity occur as an individual reaches a high altitude. Serum erythropoietin is elevated initially, but

eventually returns to the normal range in the absence of extreme hypoxia. This decline will not prevent the increase in red cell mass, which will be sustained, because early unsustained elevations of erythropoietin promote expansion of the erythroid progenitor pool. Only very small quantities of the hormone are subsequently required to sustain the red cell mass under normal circumstances.

A decrease in the plasma volume frequently accompanies hypoxic erythrocytoses resulting in further elevation of the haematocrit. Some individuals who live at high altitudes for prolonged periods develop chronic mountain sickness. They suffer from headaches, fatigue, impaired exercise tolerance, cyanosis, clubbing, right heart failure, and absolute polycythaemia. These symptoms frequently resolve with therapeutic phlebotomy.

Chronic pulmonary disease

Pulmonary diseases are a common cause of secondary polycythaemias. Defects in gas exchange result in hypoxia, with consequent increases in erythropoietin and red cell mass. Not every patient with hypoxia secondary to respiratory disease develops polycythaemia. The presence of concurrent inflammation or infection may blunt the marrow response to hypoxia. It is important to be aware that smoking itself may also contribute significantly to the polycythaemia associated with chronic respiratory disease. Phlebotomy may be indicated in patients with relatively high haematocrits (55–60 per cent), given the known deleterious effects of increased viscosity on tissue perfusion.

Alveolar hypoventilation

Hypoventilation may lead to hypoxia and an erythropoietin-mediated increase in red cell mass. These disorders include the sleep apnoea syndrome and supine hypoventilation. In these conditions, significant degrees of hypoxia may occur without evident parenchymal pulmonary disease. Decreases in blood oxygen content may occur intermittently, consequently erythropoietin levels and arterial blood gas values may be normal. Diseases affecting the central nervous system may impair respiratory centre function and trigger hypoventilation. These defects have been described in association with encephalitis, cerebrovascular accidents, and drug intoxication (that is, barbiturates). Impaired skeletal muscle function of the chest wall or diaphragm may also sufficiently compromise alveolar ventilation to trigger polycythaemia. In these cases, correction of hypoxia is warranted. The role of phlebotomy is unclear, but not unreasonable in patients with significant elevations in haematocrit and associated cardiovascular or cerebrovascular disease.

Cardiovascular disease

Cyanotic congenital heart diseases with associated right-to-left shunt result in oxygen desaturation and an elevation of erythropoietin, causing secondary polycythaemia. After compensatory erythrocytosis in response to oxygen desaturation occurs, serum erythropoietin levels may return to normal levels. Characteristically, therapeutic phlebotomy is associated with marked increases in erythropoietin levels. Some children with congenital heart disease may develop extreme haematocrit values (80 per cent or higher), which lead to a clear risk of a thrombotic event, especially during periods of dehydration. Phlebotomy may be indicated in some instances: for example, in preparation for elective surgery. Further clinical information is required to establish the precise target haematocrit value for therapeutic phlebotomy in the management of these disorders.

Carbon monoxide intoxication

Chronic carbon monoxide intoxication most commonly occurs as a consequence of smoking. Elevated haematocrits have been reported in 3 per cent of all smokers. Other less common causes include work-related exposures such as those seen in caisson workers or tunnel toll-collectors. Carbon monoxide has a much higher affinity for haemoglobin than oxygen, thereby reducing the amount of oxygen that can be bound and transported by haemoglobin. It also shifts the oxygen–haemoglobin dissociation curve to the left, decreasing the ability of haemoglobin to release oxygen to peripheral tissues. Furthermore, carbon monoxide impairs normal compensatory mechanisms; carboxyhaemoglobin is known to decrease 2,3-DPG production by red cells and to reduce the affinity of haemoglobin for 2,3-DPG. Polycythaemia due to chronic carbon monoxide intoxication may be associated with an increased risk of thromboembolic phenomena. Phlebotomy may be indicated in patients with very high haematocrits (>55–60 per cent).

The decreased oxygen-carrying capacity associated with carbon monoxide intoxication is not detected by standard blood gas measurements, therefore a direct measure of carboxyhaemoglobin levels is required. Morning carboxyhaemoglobin levels ranging from 4 per cent to 20 per cent have been reported. Individuals with chronic carbon monoxide poisoning may experience neuropsychiatric and cardiac abnormalities. The treatment is smoking cessation or removal of the patient from the alternative source of carbon monoxide.

High-affinity haemoglobins

At least 50 haemoglobin variants exhibit increased avidity for oxygen. Oxygen transport by haemoglobin occurs as a function of the oxygen–haemoglobin affinity curve. This function is represented by a sigmoid-shaped curve and is a reflection of the initial binding of oxygen by deoxygenated haemoglobin occurring with significant difficulty. As oxygen molecules are bound to normal haemoglobin, further binding is facilitated by structural changes that occur to the haemoglobin molecule. High-affinity haemoglobin variants arise when mutations alter key amino acid residues in regions of haemoglobin that affect these rearrangements, or at the interface between a- and b-chains. Another group of mutations induces changes in oxygen affinity indirectly, by causing structural changes in haemoglobin regions that are critical for the binding of 2,3-DPG.

Increases in oxygen affinity result in a shift of the oxygen dissociation curve to the left. Consequently haemoglobin binds oxygen more readily and retains more oxygen at lower *P*O2 (partial pressure of O2) levels. This ultimately results in decreased delivery of oxygen to tissues where capillary *P*O2 is low (35–45 mmHg). Mild tissue hypoxia then triggers an increase in the production of erythropoietin with consequent polycythaemia.

Oxygen affinity by a variant haemoglobin is usually measured as the $P_{50}O_2$, which represents the partial oxygen pressure at which 50 per cent of haemoglobin is saturated with oxygen. This analysis is necessary for the identification of patients with high-affinity haemoglobins. High-affinity haemoglobins are associated with lower than normal values of $P_{50}O_2$. Haemoglobin electrophoresis may, on occasion, aid in the recognition of an abnormal haemoglobin, but many high-affinity haemoglobins display normal electrophoretic mobility. Conversely, the presence of an abnormal band *per se* does not provide information regarding oxygen affinity. A study of family members is important, but a negative family history does not negate the diagnosis since there is a high rate of spontaneous mutations.

Most patients with high-affinity haemoglobins have mild polycythaemia and are asymptomatic since the compensatory polycythaemia results in normal oxygen delivery to tissues. On rare occasions, very high haematocrits (>55–60 per cent) associated with elevated blood viscosity may be sufficient to warrant therapeutic phlebotomy.

Methaemoglobinaemias

Hereditary methaemoglobinaemia may be associated with a mild polycythaemia. Methaemoglobin results from the oxidation of ferrous ions (Fe^{2+}) to the ferric state (Fe^{3+}). Oxygen does not bind reversibly to methaemoglobin, so resulting in a left shift of the oxygen dissociation curve, impaired oxygen delivery, and chronic tissue hypoxia.

2,3-Diphosphoglycerate (2,3-DPG) deficiency

This rare familial form of polycythaemia is due to a deficiency of the enzyme diphosphoglyceromutase. Deficiency leads to a decrease in 2,3-DPG, resulting in the increased affinity of oxygen to haemoglobin, peripheral tissue hypoxia, and hypoxic erythrocytosis. This disorder should be suspected in patients with familial polycythaemia with a low $P_{50}O_2$ in the absence of a mutant haemoglobin. Measurements of 2,3-DPG in fresh red cells reveals reduced levels.

Chuvash polycythaemia

Chuvash polycythaemia is a recently recognized form of congenital and familial polycythaemia that is endemic to the Chuvash population of the Russian Federation. The extreme elevations of haemoglobin in this autosomal recessive disorder are accompanied by increased erythropoietin levels. Excessive elaboration of erythropoietin is thought to be due to an abnormality of the oxygen-sensing mechanism. These patients present with isolated erythrocytosis without elevations of white

cells or platelets. Death frequently occurs before the age of 40 due to thrombotic and haemorrhagic complications.

Secondary polycythaemias associated with the inappropriate secretion of erythropoietin

Enhanced erythropoietin levels and secretion occur in the absence of tissue hypoxia in this group of disorders. The erythropoietin response is therefore inappropriate to systemic oxygen requirements.

Polycythaemia of renal disease

As the kidney is the major site of erythropoietin production, it is not surprising that renal disorders may be associated with erythrocytosis or anaemia. Patients with hypertension and renal artery stenosis have a higher incidence of erythrocytosis than similarly hypertensive patients without renal artery disease. Other benign kidney diseases associated with an increase in erythropoietin production and erythrocytosis include polycystic kidney disease (acquired or familial) and renal cysts. Unusual patients with glomerulonephritis may also occasionally present with an elevated haematocrit. An uncommon cause of polycythaemia is Bartter's syndrome, an hereditary tubular disorder characterized by hypokalaemia secondary to renal potassium loss in association with elevated plasma renin activity and aldosterone secretion. Between 5 and 13 per cent of patients have been reported to develop erythrocytosis following renal transplantation. It has been postulated that the excessive erythropoietin response originates from the patient's own kidneys and not from the transplanted one. Angiotensin-converting enzyme inhibitors may prove useful in controlling post-transplantation polycythaemia. Phlebotomy may be required in patients with haematocrit levels over 55 to 60 per cent.

Tumour-associated polycythaemia

A number of tumours are associated with an inappropriately increased production of erythropoietin, including benign and malignant tumours of the kidney, hepatomas, cerebellar haemangioblastomas, and phaeochromocytomas. Polycythaemia occurs in 1 per cent of patients with renal carcinomas, 9 to 20 per cent of patients with cerebellar haemangioblastomas, and 10 per cent of patients with hepatomas. Resection of the tumour, if feasible, may be associated with regression of the polycythaemia. Therapeutic phlebotomy is recommended in patients with extreme increases in the haematocrit.

Endocrine disorders

Phaeochromocytomas and aldosterone-producing adenomas have been associated with increased levels of erythropoietin. Mild forms of polycythaemia have also been observed in some patients with Cushing's syndrome, probably related to marrow stimulation by steroid hormone.

Primary polycythaemia

Primary familial and congenital polycythaemia (PFCP)

PFCP is an inherited form of polycythaemia caused by mutations in the erythropoietin receptor, thereby resulting in the hypersensitivity of erythroid progenitor cells to erythropoietin and low serum erythropoietin levels. In the autosomal dominant form of the disease, family members have plethora, headaches, dizziness, nose bleeds, and exertional dyspnoea. These symptoms resolve with phlebotomy and reduction of the haematocrit. Not all cases of PFPC can be attributed to the mutations of the erythropoietin receptor, suggesting that other genetic defects can lead to a similar phenotype.

Polycythaemia vera

Polycythaemia vera is a malignancy characterized by excessive proliferation of erythroid, myeloid, and megakaryocytic elements in the bone marrow. Its hallmark is an absolute increase in the red cell mass usually associated with leucocytosis, thrombocytosis, and splenomegaly. In contrast to other haematological malignancies, patients suffering from polycythaemia vera may enjoy prolonged survival, provided that the excessive production of red cells and platelets is controlled. This survival is occasionally punctuated by the development of myelofibrosis and/or acute leukaemia.

Epidemiology

Polycythaemia vera is a rare disorder, the estimated yearly incidence in the Western world is between 5 and 17 cases per 1 million of the population. Its true prevalence is unknown but it seems to be more common in Ashkenazi Jews and rarer in Afro-Americans. A very low incidence of 2 cases per year per million population has been reported in Japan. These differences suggest that environmental as well as genetic factors might be important.

Polycythaemia vera is slightly more common in males than in females, with a male to female ratio of 1.2:1. The average age at diagnosis is 60 years, it is very rare in individuals younger than 30 years of age. Only a handful of cases have being reported during childhood.

Biological and molecular aspects

The exaggerated production of red cells, granulocytes, and platelets in polycythaemia vera suggests that the fundamental defect occurs at the level of the pluripotent haemopoietic stem cell. The clonal, and thereby malignant, nature of polycythaemia vera was first established by the cellular analysis of blood cell production, in heterozygous Afro-American women, of X-linked glucose-6-phosphate dehydrogenase (**G6PD**) isoenzymes. These results have recently been confirmed using restriction fragment length polymorphisms (**RFLPs**) of the active X-chromosomes.

In patients with polycythaemia vera, erythropoietin levels often fall below the 95 per cent confidence interval for normal individuals. These low levels persist even after repeated phlebotomies, suggesting that excessive production of erythropoietin is not a critical component in the pathogenesis of this disorder. Using *in vitro* cell-culture systems, polycythaemia-vera bone marrow can form erythroid colonies in the absence of erythropoietin (endogenous colonies). Erythroid progenitor cells derived from normal individuals are incapable of forming such colonies in the absence of exogenous erythropoietin. In addition to the erythroid colonies, mixed-lineage colonies are frequently formed in the absence of exogenous erythropoietin, suggesting the involvement by not just erythroid precursors but also of more primitive haemopoietic progenitor cells. Both malignant and normal cells are present in the bone marrow of patients with polycythaemia vera, but the malignant cells have a growth advantage. Most peripheral blood elements are thus derived from the neoplastic clone. We now know that the clonal assay systems that were first used to produce endogenous erythroid colonies contained trace quantities of erythropoietin. In newer systems that are absolutely devoided of erythropoietin, the progenitor cells in polycythaemia vera require erythropoietin, but clearly demonstrate increased sensitivity to this growth factor.

Polycythaemia vera progenitor cells are also hypersensitive to other cytokines such as steel factor, interleukin-3 (**IL-3**), and granulocyte–macrophage colony-stimulating factor (**GM-CSF**). These responses require the presence of insulin-like growth factor-1 (**IGF-1**) or of its binding protein (**IGFBP-1**). At present, the primary signalling protein defect that underlies the characteristic hypersensitivity to this broad group of cytokines remains unknown.

Pathobiology

Patients with polycythaemia vera have an increased thrombotic tendency resulting from the expansion of the red cell mass. There is a direct relationship between the risk of thrombosis and age, suggesting that the presence of vascular disease might be important. Younger individuals are also at risk for thrombotic episodes, many of them life-threatening. The main rheological abnormality is elevated total blood viscosity. Cerebral blood flow is reduced in patients with polycythaemia vera and a haematocrit of 53 to 62 per cent. Reductions in blood flow are correctable by phlebotomy. Even small reductions in the haematocrit result in significant reduction in blood viscosity and cerebral blood flow. An increased haematocrit may facilitate the transport of platelets to the vessel wall, an event that may lead to thrombus formation. Elevation in blood viscosity also results in greater peripheral vascular resistance and a consequent reduction in organ blood flow, thereby increasing the likelihood of thrombus development. Thrombocytosis and functional platelet abnormalities are frequently present, and may play a role in the development of thrombosis. This relationship is still highly controversial.

Patients with polycythaemia vera are also at an increased risk of developing life-threatening haemorrhagic complications. Abnormalities in platelet function and number have been implicated. Qualitative platelet abnormalities include defective platelet aggregation *in vitro*, acquired storage pool disease, and dysregulated thromboxane A_2 metabolism. Patients with acquired von Willebrand syndromes have been described who have very high platelet counts (>1000 × 10^6/μl), in

association with life-threatening bleeding episodes. Leucocytosis, found in 50 per cent of patients, may impact negatively on the rheology of the microcirculation. No laboratory test has proven useful for the *a priori* identification of patients at an increased risk of developing haemorrhagic or thrombotic events.

The progression to the so-called 'spent phase' is a common cause of morbidity. This stage, also known as postpolycythaemic myeloid metaplasia (**PPMM**), is characterized by cytopenias, myelofibrosis, and extramedullary haemopoiesis. The fibroblastic component represents a reactive event, and may be due to the local release of growth factors, particularly platelet-derived growth factor. The association between the treatment modality and the development of myeloid metaplasia is as yet unclear. Clearer is the association between treatment type (alkylating agents and ^{32}P; see below), and the development of acute leukaemia. It must be emphasized, however, that even those patients treated with phlebotomy alone have a leukaemogenic risk significantly higher than that expected in the general population.

Clinical manifestations

The clinical manifestations of polycythaemia vera are the direct consequence of the excessive proliferation of cellular elements of the various haemopoietic cell lineages.

The routine and widespread use of laboratory screening tests during medical evaluations has led to an increased detection of asymptomatic patients. In contrast, symptomatic patients may present to their physician with a large array of non-specific complaints including headache, weakness, pruritus, dizziness, excessive sweating, visual disturbances, paraesthesias, joint symptoms, and epigastric distress. Some one-third of patients will have lost 10 per cent of their body weight by the time they present, presumably due to the associated hypermetabolism. Joint disease is usually the manifestation of gout, due to the increased production of uric acid. The most important signs on physical examination include ruddy cyanosis, conjunctival plethora, hepatomegaly, splenomegaly, and hypertension.

Patients left without appropriate treatment are at a particularly high risk of developing thrombotic or haemorrhagic events. In fact, thrombosis may be the cause of death in up to 30 to 40 per cent of patients. Thrombosis may occur in the deep venous system of the lower extremities, or present as a pulmonary embolism. Cerebrovascular, coronary, and peripheral vascular occlusions are not rare.

Thromboses at unusual sites are characteristic of polycythaemia vera. They include occlusion of the splenic, portal, hepatic, and mesenteric veins. Cardiac valve abnormalities affecting the aortic or the mitral valves are commonly seen, frequently in the form of leaflet thickening or frank vegetations. These lesions are associated with the occurrence of arterial thromboembolism. Hepatic venous or inferior vena caval thrombosis is known as Budd–Chiari syndrome. It is characterized by hepatosplenomegaly, ascites, oedema of the peripheral extremities, jaundice, abdominal pain, and distension of superficial abdominal veins due to portal hypertension. Some ten per cent of patients who present with Budd–Chiari syndrome have polycythaemia vera. At times, these patients will present with normal haemoglobin and haematocrit levels as a consequence of blood loss and expansion of the plasma volume. This phenomenon is regarded as 'inapparent polycythaemia vera' and requires direct quantification of the red cell mass for confirmation. Iron deficiency may also mask the expected erythrocytosis in some patients with polycythaemia vera. Leucocytosis, thrombocytosis, and splenomegaly are usually present. The definitive diagnosis of polycythaemia vera in this patient population requires the documentation of an elevated red cell mass.

Neurological abnormalities are also common and occur in up to 60 to 80 per cent of patients. They include transient ischaemic attacks, cerebral infarction, cerebral haemorrhage, confusional states, fluctuating dementia, and involuntary movement syndromes. Dizziness, paraesthesias, tinnitus, visual problems, and headaches are common symptoms attributed to the hyperviscosity state. Small infarcts in the basal ganglia region, also known as lacunae, might be related to some of the transient neurological manifestations. Symptoms of carotid or vertebral and basilar artery insufficiency occur frequently. Peripheral vascular insufficiency may be manifested by intense redness or cyanosis of the digits, burning, classical erythromelalgia, digital ischaemia with palpable pulses, or thrombophlebitis. Erythromelalgia consists of a burning pain in the digits of either the lower and/or upper extremities, an objective sensation of increased temperature, and relief by cooling. If left untreated it may evolve into gangrene. Antiplatelet aggregation therapy rapidly reverses the symptoms. Peripheral pulses are usually normal in these patients, as this phenomenon is due to changes in the microcirculation related to arteriolar activation and aggregation of platelets *in vivo*.

Haemorrhagic complications are the cause of death in 2 to 10 per cent of patients with polycythaemia vera: 30 to 40 per cent of patients will experience a haemorrhagic event sometime during the course of their disease. Peptic ulcer disease occurs frequently and contributes to the gastrointestinal tract being the most common source of bleeding. The bleeding diathesis may relate to abnormalities in platelet function, and thus occurs frequently after the ingestion of anti-inflammatory agents. Spontaneous bleeding is rare. Recent data suggests that low-dose aspirin might not increase the frequency of life-threatening haemorrhages.

The risk of postoperative complications is high in patients with polycythaemia vera. Bleeding, thrombosis, or a combination of both can occur. The risk is higher for those patients who undergo surgery with uncontrolled erythrocytosis. Generalized pruritus affects 50 per cent of all patients, but its aetiology is unknown. Increased blood and urine histamine levels have been implicated by some. Pruritus triggered by water contact is characteristic and tolerated very poorly. There is no relationship between the severity of the disease and the intensity of the pruritus. Up to 20 per cent of patients experience persistent pruritus in even after normalization of their counts.

Polycythaemia vera evolves to PPMM in up to 50 per cent of the patients 10 years, on average, after the initial diagnosis. It is characterized by increased splenomegaly, tear-drop red cells, a leucoerythroblastic blood picture, marrow fibrosis, and a normal or decreasing red cell mass. Fatigue, dizziness, weight loss, anorexia, progressive anaemia, and thrombocytopenia associated with bleeding are common. Patients with progressive anaemia should be evaluated for folate and iron deficiency. Occasional patients will respond to iron supplementation with resurgence of erythropoiesis. Severe hyperuricaemia may induce gout or uric acid nephropathy. PPMM portends a very grave prognosis with over two-thirds of patients dying within 3 years.

The evolution to acute leukaemia is probably the natural consequence of the malignant nature of polycythaemia vera, which can be accentuated by the therapeutic interventions commonly used for its treatment, such as alkylating agent use or radioactive phosphorus (^{32}P) administration. A randomized study conducted under the auspices of the Polycythaemia Vera Study Group (**PVSG**) comparing chlorambucil, ^{32}P, and phlebotomy is instructive. After 15 years of follow-up, 17.5 per cent of patients treated with chlorambucil and 10.9 per cent of those who received ^{32}P had developed acute leukaemia. Only 1.5 per cent of patients treated with phlebotomy alone developed acute leukaemia. This still represents a much higher incidence than the one expected in a normal age-matched population. Between 30 and 50 per cent of patients who develop leukaemia have previously entered the spent phase, whereas 50 per cent progress directly from the erythrocytotic phase. A significant number of patients experience a myelodysplastic interval before transforming. Large-cell lymphomas have been observed in roughly 3.5 per cent of patients treated with chlorambucil.

Laboratory evaluation

Polycythaemia vera is a panmyelosis. Some two-thirds of the patients present with leucocytosis and approximately 50 per cent with thrombocytosis. Red cell morphology usually reflects an underlying iron deficiency state: microcytosis, hypochromia, polychromatophilia, poikilocytosis, and anisocytosis are frequently seen. White blood cell morphology is usually normal. Increased numbers of basophils, eosinophils, and immature myeloid cells are observed. Megathrombocytes are often seen in the peripheral blood smear. Platelet counts are usually under $1 \times 10^6/\mu l$, but higher counts may be seen. The PPMM phase is characterized by a leucoerythroblastic response with the presence in the peripheral blood of tear-drop red cells (dacrocytes), immature myeloid cells, and nucleated red blood cells. Bleeding time and platelet aggregation studies are frequently, but not always, abnormal. Prolongation of prothrombin and partial thromboplastin times are frequently encountered, usually reflecting a laboratory artefact due to erythrocytosis (the volume of plasma in the collection tube might be too small relative to the amount of anticoagulant (citrate) present in these tubes).

Acquired von Willebrand factor (**vWF**) defects are frequently characterized by a significant decrease in large vWF multimers. This acquired defect resembles type II vWF disease. It occurs mainly in patients with very high platelet counts ($>1 \times 10^6/\mu l$), implicating the adsorption of larger forms of vWF multimers on to platelet membranes. The defect is corrected by normalization of the thrombocytosis. Elevations in leucocyte alkaline phosphatase (70 per cent), serum vitamin B12 levels (40 per cent), and serum vitamin B12 binding proteins (70 per cent) are common, as are hyperuricaemia and increased histamine levels.

Bone marrow examination reveals a hypercellular marrow with an increased number of megakaryocytes. The cellular elements are morphologically normal. Iron stores are usually absent prior to treatment. Reticulin is often seen, but is not predictive of evolution into the spent phase. At diagnosis, erythropoietin levels are either reduced or within the lower limits of normal. Low levels persist in two-thirds of patients after normalization of their haematocrit. Cytogenetic abnormalities have been described, but none are characteristic. Abnormalities in chromosomes 1, 5, 7, 8, 9, 12, 13, and 20 have been detected. The frequency and complexity of these

chromosomal abnormalities are a function of the disease longevity and duration of treatment.

Diagnostic criteria of polycythaemia vera

Clinical criteria for the diagnosis of polycythaemia vera have been developed. These criteria have proven useful in defining a homogeneous study population for incorporation into clinical studies, but, in our opinion, the previously utilized criteria have a number of shortcomings that are of clinical relevance. In Table 2 we offer what we believe to be a more flexible and useful criteria utilizing modern diagnostic tools for the diagnosis of polycythaemia vera.

Approach to the patient with polycythaemia

It is wise to avoid the temptation of diagnosing polycythaemia on the basis of a single blood count determination unless extremely high levels are identified. A rational diagnostic approach is required to avoid unnecessary emotional distress to the patient as well as expensive and unnecessary evaluations (see Fig. 1).

Fig. 1 Diagnostic approach to the patient with polycythaemia.

Dehydration from any cause can produce a spurious elevation in the blood counts. Heavy smokers with mild polycythaemias should be asked to stop smoking and their counts repeated after a few weeks. Once a genuine elevation of haemoglobin or haematocrit has been established, the next step is to decide whether this represents an absolute increase in total red cell mass, or just a relative phenomenon. A blood volume study with direct quantitation of both red cell mass and plasma volume is helpful in making this distinction. If absolute polycythaemia is confirmed, it is essential to elucidate whether it is the consequence of a primary myeloproliferative disorder such as polycythaemia vera or a secondary condition.

The determination of erythropoietin levels may prove to be of diagnostic utility. An elevated serum erythropoietin level is indicative of the presence of a secondary polycythaemia and a low level supports the diagnosis of polycythaemia vera, but a normal erythropoietin value excludes neither hypoxia nor the autonomous production of erythropoietin as the cause. Normal values may also be encountered in some cases of polycythaemia vera.

The presence of leucocytosis, thrombocytosis, or splenomegaly is suggestive of polycythaemia vera as the cause for the elevated red cell mass. Arterial blood gases and the direct determination of oxygen saturation in arterial blood, if decreased they may aid in the recognition of a chronic pulmonary or congenital cardiovascular abnormality. If blood oxygen saturation is normal, the quantification of haemoglobin's oxygen affinity ($P_{50}O_2$) may indicate the presence of a high-affinity haemoglobin variant. Otherwise, causes for a physiologically inappropriate polycythaemia should be sought.

There is a small but definite group of patients in whom a specific cause for polycythaemia remains elusive, despite appropriate diagnostic testing. Examining close relatives might disclose the presence of a hereditary polycythaemia, a rare condition caused by an abnormality in erythropoietin control. Regular, continued surveillance is recommended for all non-categorized patients, as some of them develop polycythaemia vera in the future.

Management of polycythaemia

The two major tasks in the management of polycythaemia involve the identification of a correctable cause and the reduction of the red cell mass. The untoward effects of an increased red cell mass on tissue blood flow occur independently from the specific cause of the polycythaemia. It is thus reasonable to recommend that all patients with uncorrectable erythrocytosis be offered phlebotomy. A haematocrit under 45 per cent represents a reasonable target.

Polycythaemia vera is an incurable disorder. The main therapeutic goals are the maintenance of well being and the prevention of complications for as long as possible. Several therapeutic strategies have resulted in dramatic increases in the survival of patients. Historical evidence suggests a median survival of approximately 18 months in untreated patients, whereas with appropriate management survival of over 10 years is now common. The main therapeutic objective is the reduction of the haematocrit to a safe level. This is usually accomplished by the implementation of repeated phlebotomies. Every possible effort should be made to discourage patients with polycythaemia vera from smoking. A regimen of phlebotomies should be prescribed as soon as the diagnosis has been clearly established. It is often feasible to remove between 350 and 500 ml of blood every other day until the desired haematocrit level is attained. Haematocrit levels of less than 50 per cent and preferably below 45 per cent are desirable. The removal of smaller aliquots might be necessary in older patients.

Once the target haematocrit level is achieved, a maintenance regimen should be instituted. Venesection is preferred in those younger individuals without critical elevations in their platelet counts. Myelosuppressive therapy should be considered in elderly patients who are intolerant of phlebotomies, and in younger individuals with repeated thrombotic episodes and extremely high platelet counts. There is controversy regarding what represents the optimal myelosuppressive agent. A major concern has been the growing evidence that supports the association between exposure to some of these agents and the development of leukaemia. A number of clinical studies have been conducted in order to clarify the risk/benefit ratios of some of these approaches.

The Polycythemia Vera Study Group conducted a randomized comparison of three interventions: (1) phlebotomy only; (2) chlorambucil supplemented by phlebotomy; and (3) ^{32}P supplemented by phlebotomy. Median survival was shorter for the chlorambucil group (9.1 years) and statistically similar for ^{32}P (10.9 years) and for the phlebotomy-only group (12.6 years). Chlorambucil treatment was clearly associated with an unacceptably large number of acute leukaemias. Thrombosis as a cause of death was much more common in the phlebotomy-only group, whereas the use of ^{32}P led to a higher incidence of leukaemias, lymphoma, and non-haematological malignancies. In a subsequent study, high-dose platelet antiaggregating agents (aspirin and dipyridamole) were added to phlebotomy and compared with ^{32}P. Surprisingly, a higher number of thrombotic and serious bleeding episodes were seen in the former group. In a recent study, however, 40 mg/day of aspirin was administered to patients with polycythaemia vera in order to prevent thrombosis and minimize the bleeding risk. This regimen was shown to be well tolerated, safe, and appeared to reduce the thrombotic risk. Hydroxycarbamide (hydroxyurea) is useful for the management of a number of patients with polycythaemia vera, despite emerging evidence suggesting a mild leukaemogenic potential. At present, hydroxyurea is the chemotherapeutic drug of choice in those patients who can not be treated with phlebotomy alone.

In younger patients, given their potential long-term survival, strong consideration should be given to the use of phlebotomy therapy in combination with low-dose aspirin, as well as with other apparently non-leukae-mogenic interventions such as interferon-a and anagrelide. In patients where uncontrolled thrombocytosis is a problem, anagrelide, an inhibitor of megakaryocytic maturation, has proven effective.

Elective surgery should only be undertaken after adequate and sustained control of the blood count has been achieved. When emergency surgery is required the patient should be phlebotomized rapidly until a normal haematocrit is achieved, and platelets should be available in case excessive operative bleeding occurs. Patients should be mobilized promptly, and the use of prophylactic doses of low molecular weight heparin should be considered unless contraindicated. Dental extractions are associated with an increased bleeding risk and should only be pursued in patients with good haematological control.

In the later spent phases of the disease the management is quite similar to that for primary myelofibrosis. It mainly consists of blood product replacement, folate and

iron replacement, and splenectomy if there is significant hypersplenism. The prognosis of patients with acute leukaemia that has evolved from pre-existing polycythaemia vera is very poor, with very few long-term survivors after aggressive combination chemotherapy.

Pregnant patients with polycythaemia vera experience an increased incidence of fetal wastage, with 30 per cent of pregnancies culminating in spontaneous abortions. Interestingly, pregnancy in patients with polycythaemia vera is frequently associated with a gradual normalization of blood values, and it is not unusual for a woman who has required extensive therapy for control of her disease to no longer require phlebotomies during pregnancy. Delivery appears not to be complicated by excessive haemorrhage or by an increased risk of venous thrombosis.

Prognosis

The outcome of patients with secondary polycythaemia is usually related to the prognosis of the underlying disorder. In polycythaemia vera, the nature and severity of the complications during the clinical course of the disease are the most important determinants of outcome. Disease duration is also important, as long-term survival is strongly associated with progression into the spent phase or acute leukaemia. As was previously emphasized, prompt and appropriate therapy results in dramatic improvements in survival. Young patients should be initially managed with phlebotomy and low doses of aspirin. Supplemental therapy with interferon, anagrelide, or hydroxyurea might be required in patients with serious haemorrhagic or thrombotic episodes. The use of either hydroxyurea or ^{32}P appears warranted in the treatment of elderly patients who, because of their age, have a limited survival.

Further reading

Copelan EA, Balcerzak SP (1995). Secondary polycythemia. In: Wasserman LR, Berk PD, Berlin NI, eds. *Polycythemia vera and the myeloproliferative disorders*, pp 195–221. WB Saunders, Philadelphia.

Fruchtman SM, *et al*. (1997). From efficacy to safety: a Polycythemia Vera Study Group report on hydroxyurea in patients with polycythemia vera. *Seminars in Hematology* **34**, 17–23.

Gruppo Italiano Studio Policitemia (1995). Polycythemia vera: the natural history of 1213 patients followed for 20 years. *Annals of Internal Medicine* **123**, 656–64.

Gruppo Italiano Studio Policitemia (GISP) (1997). Low dose aspirin in polycythaemia vera: a pilot study. *British Journal of Haematology* **77**, 453–6.

Hoffman R (2000). Polycythemia vera. In: Hoffman R, *et al.*, eds. *Hematology basic principles and practice*, pp 1130–55. Churchill Livingstone, Philadelphia, PA.

Messinezy M, *et al.* (1995). Low serum erythropoietin—a strong diagnostic criteria of primary polycythaemia even at normal haemoglobin levels. *Clinical and Laboratory Haematology* **17**, 217–20.

Michiels JJ (1996). The myeloproliferative disorders, a historical appraisal and personal experiences. *Leukemia and Lymphoma* **22**(Suppl. 1), 1–14.

Michiels JJ, *et al.* (1985). Erythromelalgia caused by platelet mediated arteriolar inflammation and thrombosis in thrombocythemia. *Annals of Internal Medicine* **102**, 466–71.

Michiels JJ, *et al.* (1996). Erythromelalgic thrombotic and hemorrhagic manifestations in 50 cases of thrombocythemia. *Leukemia and Lymphoma* **22**(Suppl. 1), 47–56.

Pearson TC, Messinezy M (1996). The diagnostic criteria of polycythemia rubra vera. *Leukemia and Lymphoma* **22**(Suppl. 1), 87–93.

Pearson TC, Wetherley-Mein G (1978). Vascular occlusive episodes and venous haematocrit in primary proliferative polycythemia. *Lancet* **2**, 1219–22.

Petitt RM, Silverstein MK (1997). Anagrelide for control of thrombocythemia in polycythemia and other myeloproliferative disorders. *Seminars in Hematology* **34**, 51–4.

Prchal JF, Prchal JT (1999). Molecular basis for polycythemia. *Current Opinion in Hematology* **6**, 100–9.

Silver RT (1997). Interferon alpha: effects of long term treatment for polycythemia vera. *Seminars in Hematology* **34**, 40–50.

Spivak JL (2000). Erythrocytosis. In: Hoffman R, *et al.*, eds. *Hematology basic principles and practice*, pp 388–96. Churchill Livingstone, Philadelphia, PA.

22.3.9 Idiopathic myelofibrosis

Jerry L. Spivak

Introduction

Idiopathic myelofibrosis (also called myelofibrosis with myeloid metaplasia, agnogenic myeloid metaplasia, primary myelofibrosis, or primary myelosclerosis) is a chronic clonal disorder of unknown aetiology, involving a multipotent haemopoietic progenitor cell that results in abnormalities in red cell, white cell, and platelet production in association with marrow fibrosis and extramedullary haemopoiesis. Although myelofibrosis in association with leucoerythroblastosis and splenomegaly are the clinical hallmarks of idiopathic myelofibrosis, these abnormalities can also be seen in other chronic myeloproliferative disorders such as polycythaemia vera and chronic myelogenous leukaemia as well as in a variety of benign and malignant disorders that involve the bone marrow (Table 1). Since there is no specific clonal marker for idiopathic myelofibrosis and since many of the disorders listed in Table 1 are responsive to specific therapies not effective in idiopathic myelofibrosis, the diagnosis of this disorder is one of exclusion.

Aetiology

The aetiology of idiopathic myelofibrosis is unknown. Analysis of glucose-6-phosphate dehydrogenase (**G6DP**) isoenzyme expression and X-linked gene inactivation patterns in informative women, as well as N-*ras* mutations, have established the clonality of idiopathic myelofibrosis and its origin in a multipotent haemopoietic progenitor cell. In some patients, T lymphocytes express the same clonal marker as B lymphocytes and myeloid cells, suggesting involvement at the level of the pluripotent stem cell. Non-random chromosome abnormalities primarily involving chromosomes 13 (del. 13q), 20 (del. 20q), and 1 (partial trisomy 1q) occur, but are found in fewer than 40 per cent of patients at the time of diagnosis. No abnormalities of the known tumour suppressor genes associated with these chromosomes have yet been identified.

The basis for the myelofibrosis has proved equally problematic. Marrow fibroblasts in idiopathic myelofibrosis are polyclonal, suggesting that the fibrosis is a reactive process initiated by expansion of the monoclonal malignant clone. Marrow collagen is argyrophilic, so that changes in its distribution and content can be analysed histochemically by silver staining. Under normal circumstances, the connective tissue stroma of the bone marrow is composed of collagen types I, III, IV, and V together with non-collagen proteins such as fibronectin, laminin, vitronectin, and the proteoglycans. Collagen types I, III, and V form a delicate and usually non-continuous supporting network for haemopoietic cells, while type IV collagen, laminin, and fibronectin are localized in the basement membranes of arteries in a continuous fashion and along marrow sinusoids in a discontinuous fashion. With increasing marrow cellularity, the collagenous supporting network also increases. In myelofibrosis, however, there is both an increase in the collagen network and a change in its physical characteristics. Condensation of the interstitial fibres results in the formation of thick, continuous and often wavy bundles in association with an increase in reticular or fibroblastic cells. Sinusoidal basement membrane collagen becomes continuous, leading to sinusoidal dilatation and obliteration with an associated capillary neovascularization. The content of basement membrane fibronectin as well as stromal fibronectin and vitronectin also increases. While best studied in idiopathic myelofibrosis, the types of collagen involved in marrow fibrosis in this condition do not appear to differ from those involved in the marrow fibrosis associated with the other disorders listed in Table 1.

Neither the stimulus nor the molecular basis for the increase in marrow collagen and non-collagenous extracellular matrix proteins in idiopathic myelofibrosis is understood. The commonality of the types of collagen involved and the similarity of the histological process, regardless of disease association, implies that marrow fibrosis *per se* represents a final common pathway involved in the response to diverse immunological, metabolic, toxic, or infectious stimuli (Table 1). Megakaryocytic hyperplasia, dysplasia, and clustering is characteristic of idiopathic myelofibrosis. These cells produce proteins such as platelet-derived growth factor (**PDGF**) and transforming growth factor-b (**TGF-b**) that promote fibroblast proliferation, and platelet factor 4, which, like TGF-b, inhibits collagenase. These findings suggest that inappropriate release of these fibrogenic proteins by dysfunctional megakaryocytes is the stimulus for myelofibrosis. In support of this contention, elevated levels of PDGF and TGF-b as well as basic fibroblast growth factor have been observed in platelets and megakaryocytes from patients with idiopathic myelofibrosis. Circulating levels of TGF-b are also increased in idiopathic myelofibrosis, as is the urinary excretion of basic fibroblast growth factor and calmodulin, another potential fibroblast stimulant present in platelets.

However, these findings must be reconciled with the observations that neither overexpression of PDGF nor high levels of circulating TGF-b alone are associated with marrow fibrosis, presumably because PDGF in the circulation is inactivated by a2-macroglobulin and TGF-b usually circulates in an inactive from. Furthermore, in some studies the appropriate control populations were not examined, while in others there was significant overlap in the overexpression of these proteins when they were examined. Importantly, there was no correlation between marrow TGF-b content and megakaryocyte number, nor was there a correlation between marrow fibrosis and platelet number, although such a correlation existed between the extent of fibrosis and granulocyte number. Nevertheless, given what is known about the process of tissue fibrosis in other organs, it is likely that a combination of cytokines and growth factors such as TGF-b, PDGF, basic fibroblast growth factor, interleukin-1, and tumour necrosis factor (**TNF**) acting in synergy are required to initiate and perpetuate collagen deposition. In addition, thrombopoietin may also have a role, since overexpression of this hormone can cause myelofibrosis in animal models. Neither the cells responsible for the elaboration of these cytokines nor their targets, whether reticulum cells or fibroblasts, have been defined but recent evidence suggests that monocytes may be an important source of cytokine production. What has been learned, however, is that myelofibrosis is reversible by chemotherapy or bone marrow transplantation and occasionally spontaneously resolves.

Clinical features

Although considered to be an uncommon disorder with an incidence of approximately 1/100 000 person-years, clinical studies of more than 1000 patients have been reported over the last 40 years. In contrast to the other chronic myeloproliferative disorders, the median age at diagnosis of myelofibrosis, 61 years (range 15–94), is much older. No gender differences exist and familial clustering is sufficiently unusual that another myeloproliferative disorder such as polycythaemia vera should be considered when this occurs. The presenting manifestations depend on the state of the illness but are often bland. Many patients are asymptomatic at the time of discovery. Fatigue is the commonest presenting complaint followed by weight loss, night sweats, fever, dyspnoea, and abdominal discomfort due to splenomegaly. Hearing loss due to otosclerosis is an interesting but often non-elicited symptom. Easy bruising or bleeding and acute gout or renal stones are other presenting manifestations that are reasonably common and directly related to the underlying disease process. Rarely, periostitis may occur.

Splenomegaly is present in virtually every patient with idiopathic myelofibrosis at diagnosis. When absent, one should consider other causes for the clinical abnormalities. The degree of splenomegaly varies but is frequently substantial. Moreover, the rate of splenic enlargement is also variable; spleen size cannot be used as an indication of disease duration. Hepatomegaly, invariably of a lesser extent than the splenomegaly, is present initially in approximately 50 per cent of patients and is usually proportional to the degree of splenomegaly. Lymphadenopathy is uncommon. With substantial splenomegaly, wasting may be prominent.

Laboratory studies

Because of its origin in a multipotent haemopoietic progenitor cell, idiopathic myelofibrosis affects all cell lines but not in a predictable manner. Anaemia, usually mild, is the most consistent abnormality. Indeed, a normal haemoglobin or haematocrit in the presence of substantial splenomegaly should lead to immediate consideration of polycythaemia vera, since the expanded plasma volume associated with splenomegaly can mask a substantial increase of the red cell mass. The leucocyte and platelets counts can be low, normal, or high without reference to spleen size. Inevitably, due to extramedullary haemopoiesis, metamyelocytes, myelocytes, promyelocytes, myeloblasts, and nucleated red cells will be present in the circulation together with the tear drop-shaped red cells characteristic of this situation (Fig. 1). While this so-called leucoerythroblastic reaction is not specific for idiopathic myelofibrosis, its absence should challenge the clinical impression. Abnormalities in liver function tests are not uncommon, usually mild and most often involve a reduction in serum albumin and an elevation of the alkaline phosphatase, an abnormality

that is magnified by splenectomy. The lactate dehydrogenase (**LDH**) level is usually mildly increased and correlates best with the leucocyte count. Hyperuricaemia is not infrequent. The leucocyte alkaline phosphatase concentration can be low, normal, or high and so cannot be recommended as a diagnostic test.

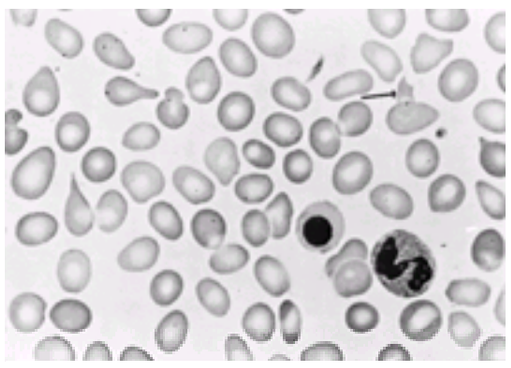

Fig. 1 Haematological changes in myelosclerosis. (a) A peripheral blood film showing tear-drop cells, a nucleated red cell, and a grossly distorted cell marked by the arrow. (From Liebold PF, Weed RI (1975). *Clinical Haematology*, **4**, 353, reproduced with permission.) (b) A scanning electron-microscope study showing characteristic poikilocytes. (From Liebold PF, Weed RI (1975). *Clinical Haematology*, **4**, 353, reproduced with permission.) (c) A leucoerythroblastic reaction in myelosclerosis showing young red-cell precursors. Note, in addition, the abnormal platelet morphology and platelet clumps.

Perhaps the most intriguing laboratory abnormalities in idiopathic myelofibrosis are those linked to autoreactivity, such as circulating immune complexes, complement activation, elevations in antinuclear antibody (**ANA**) and rheumatoid factor titres, and a positive Coombs' test in the absence of an overt connective tissue disorder. Although marrow fibrosis has been documented in patients with systemic lupus erythematosis, the linkage between autoimmune abnormalities and marrow fibrosis is unclear. It does, however, provide another therapeutic option as discussed below.

The presence of marrow fibrosis is essential for a diagnosis of idiopathic myelofibrosis and usually results in a 'dry tap' or the inability to aspirate marrow from a properly placed needle. A prefibrotic phase of idiopathic myelofibrosis has been described retrospectively. However, given the similarity of the histopathology of polycythaemia vera, essential thrombocytosis, and premyelofibrotic myelofibrosis, prospective substantiation of the latter disorder is not possible in the absence of marrow fibrosis and/or a specific clonal marker. Even the presence of myelofibrosis, while mandatory, is not in itself sufficient for diagnosis. This is because polycythaemia vera and chronic myelogenous leukaemia and other disorders such as hairy-cell leukaemia, myelodysplasia, and acute leukaemia can present with myelofibrosis. Thus, it is essential to employ the appropriate diagnostic tests (cytogenetics, *BCR–ABL* polymerase chain reaction (**PCR**), flow cytometry, and immunohistochemistry) to exclude these and the other disorders listed in Table 1 that can cause myelofibrosis.

Marrow cellularity in idiopathic myelofibrosis may be increased with trilineage hyperplasia and erythroblastic and megakaryocytic islands, decreased with scattered areas of hyperplastic marrow embedded in a collagenous matrix, or hypoplastic with intense osteomyelosclerosis and residual megakaryocytic islands (Fig. 2). While there is a correlation between the degree of fibrosis and osteosclerosis, there is no correlation between bone marrow histology and disease duration, platelet count, or splenomegaly; marrow fibrosis does, however, appear to correlate with the leucocyte count. In general, marrow fibrosis and extramedullary haemopoiesis with myeloid metaplasia appear unrelated, and the latter abnormalities cannot be considered as compensation for the former. Increased marrow angiogenesis is a recently recognized feature of idiopathic myelofibrosis which correlates with increased cellularity and extramedullary haematopoiesis independently of marrow fibrosis.

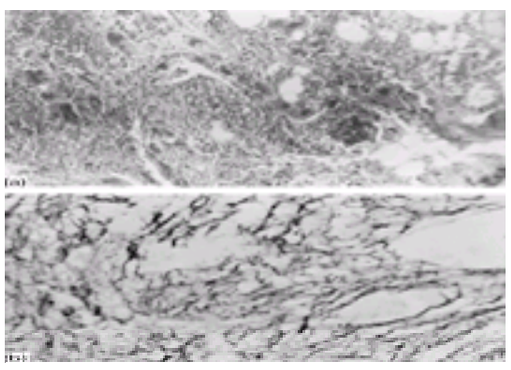

Fig. 2 Bone marrow appearances in myelosclerosis. (a) A biopsy showing a hyperplastic fragment with marked megakaryocytic hyperplasia; (b) silver stain showing the marked increase in reticulin.

In conjunction with the most severe form of marrow fibrosis, osteosclerosis, radiographic abnormalities become apparent. These primarily involve the axial skeleton but can include the skull, with thickening of bony trabeculae and patchy or coalescent sclerosis. With obliteration of axial marrow, extension of the marrow into the long bones occurs. Interestingly, the increase in trabecula bone formation in idiopathic myelofibrosis is not accompanied by an increase in either osteoblastic or osteoclastic activity. This feature distinguishes the osteosclerosis of idiopathic myelofibrosis from that associated with metabolic causes of osteosclerosis.

Course and prognosis

Idiopathic myelofibrosis is a chronic progressive disorder with a median lifespan (5.5 years) that is much shorter than for polycythaemia vera and essential thrombocytosis. However, the heterogeneity characterizing the initial clinical presentation is also evident with respect to survival, which can range from less than a year to more than 30 years. Death is usually a consequence of bone marrow failure (haemorrhage, anaemia, or infection), transformation to acute leukaemia, portal hypertension, heart failure, cachexia, or myeloid metaplasia with organ failure. Retrospective analysis of the adverse prognostic value of presenting manifestations has identified a number of factors that may be useful for both prognostic and therapeutic purposes. These include age at onset (>64 years), anaemia (haemoglobin <10 g/dl), constitutional symptoms, white cell count abnormalities (<4000/μl or >12 000/μl), thrombocytopenia, circulating blast cells (>1 per cent), and cytogenetic abnormalities. A number of scoring systems have been devised for identifying long- and short-term survivors based on the presence of more than one adverse presenting manifestation. Two such scoring systems that are useful in separating patients of any age with myelofibrosis into low- and high-risk groups with respect to survival are shown in Table 2.

Complications

The major complications of idiopathic myelofibrosis are the consequences of bone marrow failure and extramedullary haemopoiesis.

Anaemia may be the result of ineffective erythropoiesis, but haemodilution due to the expanded plasma volume associated with splenomegaly, iron deficiency due to gastrointestinal blood loss, folic acid deficiency due to the increased demands of haemopoiesis, haemolysis due to autoimmune phenomena or hypersplenism, and, rarely, pyridoxine deficiency are also considerations. In some patients, erythropoietin production may be inappropriately low for the degree of anaemia but in this instance haemodilution needs to be excluded. Red cell survival and splenic sequestration studies can be useful in determining the splenic contribution to anaemia. Ferrokinetic studies, although no longer easily obtained, provide a means of assessing effective marrow erythropoiesis.

Hyperuricaemia is a consequence of increased cell turnover and can provoke acute gout or renal stone formation if left untreated.

Splenic enlargement is inevitable and can lead to splenic infarction, malnutrition due to early satiety, plasma volume expansion, hypersplenism, portal hypertension,

extreme discomfort due to its mass, and eventually cachexia (Fig. 3). Hepatomegaly is associated with splenomegaly. Impaired hepatic function is a consequence of extramedullary haemopoiesis, which can lead to hepatic fibrosis and portal hypertension.

Fig. 3 Autopsy showing massive splenomegaly together with a splenic infarct on the medial surface of the spleen in a patient with advanced myelosclerosis.

Although myeloid metaplasia due to exuberant extramedullary haemopoiesis is most common in the spleen and liver, it can occur at any site and compromise organ or tissue function. For example: peritoneal involvement can lead to ascites; epidural involvement to spinal cord compression; retroperitoneal involvement to obstructive uropathy or portal hypertension; and intravascular haemopoiesis to pulmonary thrombosis. The reason why myeloid metaplasia is more aggressive in some patients than in others is unclear.

Approximately 20 per cent of patients with idiopathic myelofibrosis develop acute leukaemia as a terminal event. Although some clinicians do not distinguish acute leukaemia presenting with myelofibrosis (malignant myelosclerosis) from idiopathic myelofibrosis, they are clinically distinct entities. The extent to which therapeutic intervention with mutagenic drugs such as hydroxycarbamide (hydroxyurea), alkylating agents, or irradiation predisposes patients with idiopathic myelofibrosis to progress to acute leukaemia (as it does in patients with polycythaemia vera or essential thrombocytosis) is unknown. Again for unknown reasons, splenectomy also appears to be a predisposing factor for the development of acute leukaemia.

Platelet dysfunction is a common feature of the chronic myeloproliferative disorders and can lead to spontaneous haemorrhage as well as increased bleeding during surgical procedures. Although abnormalities in platelet morphology, prolongation of the bleeding time, and abnormal platelet aggregation are frequently observed in patients with idiopathic myelofibrosis, no consistent biochemical abnormality has been identified and no platelet function test is predictive for the risk of haemorrhage.

Therapy

There is no specific therapy for idiopathic myelofibrosis. Treatment should be individualized based on the patient's risk group and age. Asymptomatic, low-risk patients without hyperuricaemia or a remedial cause of anaemia require no therapy, although the oral administration of folic acid (1 mg per day) and a trial of oral pyridoxine (250 mg per day for 3 months) appears reasonable. Anaemia associated with an inappropriately low endogenous erythropoietin level may respond to recombinant erythropoietin therapy but the hormone can cause an increase in splenomegaly or hepatomegaly. Hyperuricaemia should be treated with allopurinol. Asymptomatic leucocytosis or thrombocytosis requires no therapy. Patients of appropriate age, who are in a high-risk category and who have a matched, related donor should be considered for allogeneic bone marrow transplantation. In the absence of a suitable donor, therapy with recombinant a-interferon should be employed. This can alleviate splenomegaly and reduce myeloid metaplasia but may not reverse marrow fibrosis. Interferon therapy can be limited by the induction of leucopenia or thrombocytopenia and by its side-effects in elderly patients, in whom interferon therapy should be initiated at a low dose.

Given the known sensitivity of patients with polycythaemia vera to chemotherapeutic agents and irradiation, these forms of therapy should be used judiciously in the treatment of idiopathic myelofibrosis. Hydroxycarbamide, while easy to use and with a low incidence of acute toxicity, is leukaemogenic and should not be employed before a trial of interferon. Busulfan is another effective agent that has been demonstrated to reduce organomegaly, reverse marrow fibrosis, and improve blood counts, occasionally in a durable fashion. However, busulfan has significant toxicities, not least of which is prolonged marrow aplasia. Its influence on the development of acute leukaemia in patients with idiopathic myelofibrosis is unknown.

Splenomegaly is the most distressing complication of idiopathic myelofibrosis, leading to mechanical discomfort, inanition, splenic infarction, portal hypertension, and blood cell sequestration. Reduction in splenic size can be achieved with interferon, alkylating agents, hydroxcarbamide, splenectomy, and splenic irradiation. Interferon is the treatment of choice followed by chemotherapy with either busulfan or hydroxycarbamide. Splenic irradiation can be effective at alleviating splenic pain and temporarily reducing spleen size. However, its use should be restricted to inoperable patients since there is an unpredictable risk of severe cytopenias as well as an increased risk of haemorrhage if the irradiation precedes splenectomy. Local irradiation is, of course, appropriate for the management of patients with symptomatic extramedullary haemopoiesis.

Splenectomy in idiopathic myelofibrosis is a prodigious procedure, given the large size of the spleen and its vessels, the inevitable presence of adhesions, the haemorrhagic tendency of patients with idiopathic myelofibrosis, and their often poor nutritional status. Evaluation for portal hypertension should precede surgery and, if necessary, parental hyperalimentation should be employed to avoid postoperative complications. Epsilon-aminocaproic acid should be used if bleeding is a problem.

Leucocytosis, thrombocytosis, and postoperative hepatic enlargement are the usual consequences of splenectomy, as is elevation of the alkaline phosphatase. Postoperative splenic and portal vein thrombosis occur in approximately 10 per cent of patients, most often in the first few weeks after surgery and presumably due to the size of the splenic vein remnant. However, there is no correlation between splenic or portal vein thrombosis and the platelet count. Surveillance by sonography or computed tomography may useful in identifying this complication with the intent of administering anticoagulants or thrombolytic agents. Although most patients tolerate splenectomy well, the incidence of the transformation of idiopathic myelofibrosis to acute leukaemia is increased postsplenectomy, for unknown reasons.

Finally, as mentioned earlier, both autoimmune phenomena and capillary neovascularization are features of idiopathic myelofibrosis. The use of antiangiogenetic and immunosuppressive agents such as thalidomide are currently undergoing clinical trials. Corticosteroids may also be beneficial if autoimmune phenomena are clinically significant. Finally, tuberculosis was a frequent complication of idiopathic myelofibrosis early in the nineteenth century. Thus, constitutional symptoms in these patients should not be attributed to the myeloproliferative disease without first excluding an infectious process.

Further reading

Barosi G (1999). Myelofibrosis with myeloid metaplasia: diagnostic definition and prognostic classification for clinical studies and treatment guidelines. *Journal of Clinical Oncology* **17,** 2954–70.

Cervantes F, *et al*. (1997). Identification of 'short-lived' and 'long-lived' patients at presentation of idiopathic myelofibrosis. *British Journal of Haematology* **97**, 635–40.

Dupriez B, *et al*. (1996). Prognostic factors in agnogenic myeloid metaplasia: a report on 195 cases with a new scoring system. *Blood* **88**, 1013–18.

Elliott MA, *et al.* (1998). Splenic irradiation for symptomatic splenomegaly associated with myelofibrosis with myeloid metaplasia. *British Journal of Haematology* **103**, 505–11.

Frey BM *et al*. (1998). Adenovector-mediated expression of human thrombopoietin cDNA in immune compromised mice. *Journal of Immunology* **160**, 691–9.

Glew RH, Wolfgang HH, McIntrye PA (1973). Myeloid metaplasia with myelofibrosis. The clinical spectrum of extramedullary hematopoiesis and tumor formation. *Johns Hopkins Medical Journal* **132**, 253–70.

Mesa RA *et al*. (2000). Evaluation and clinical correlations of bone marrow angiogenesis in myelofibrosis and myeloid metaplasia. *Blood* **96**, 3374–80.

Reilly JT (1997). Idiopathic myelofibrosis: pathogenesis, natural history and management. *Blood* **11**, 233–42.

Reilly JT, *et al*. (1997). Cytogenetic abnormalities and their prognostic significance in idiopathic myelofibrosis: a study of 106 case. *British Journal of Haematology* **98**, 96–102.

Sterkers Y, *et al*. (1998). Acute myeloid leukemia and myelodysplastic syndromes following essential thrombocythemia treated with hydroxyurea: high proportion of cases with 17p deletion. *Blood* **91**, 616–22.

Tefferi A *et al*. (2000). Splenectomy in myelofibrosis with myeloid metaplasia: a single-institution experience with 223 patients. *Blood* **95**, 226–33.

Truong LD, Saleem A, Schwartz MR (1984). Acute myelofibrosis. a report of four cases and review of the literature. *Medicine* **63**, 182–7.

Varki A, *et al.* (1974). The syndrome of idiopathic myelofibrosis. A clinicopathologic review with emphasis on the prognostic variables predicting survival. *Medicine* **62**, 353–71.

22.3.10

Thrombocytosis

David M. Gustin and Ronald Hoffman

Thrombocytosis refers to a platelet count elevated above the accepted normal range (more than 500 × 10^9/l). The widespread use of automated cell counters has made the identification of platelet count abnormalities a relatively common event requiring further evaluation. The clinical consequences are usually determined by the cause of the thrombocytosis, ranging from the uneventful recognition of a laboratory abnormality, to medical emergencies such as life threatening thrombosis or haemorrhage.

Pathophysiology and classification

An understanding of the disorders of platelet production requires knowledge of the regulatory events that occur during normal megakaryocytopoiesis. Megakaryocyte development is a complex process in which a wide variety of regulatory signals work in concert to direct a highly specific response to thrombopoietic demand. A large number of cytokines including interleukins (IL-3, IL-6, IL-9 and IL-11), c-kit ligand, granulocyte–macrophage colony stimulating factor (GM-CSF), thrombopoietin (TPO) and, possibly, erythropoietin have been shown to stimulate megakaryocyte development, but TPO and its receptor, c-mpl, are the primary physiological regulators of *in vivo* magakaryocytopoiesis. The liver and the kidney contribute most of the basal, constitutive production of TPO. Levels are regulated by the total mass of platelets and megakaryocytes. TPO binding to its receptor on these cells, and its subsequent degradation, represents its main pathway of clearance. During times of thrombopoietic stress, there is increased TPO production by the spleen and bone marrow. Inappropriately elevated levels of TPO may be observed in primary thrombocythaemia. This is probably not due to excessive production but rather impaired TPO clearance associated with decreased expression of the TPO receptor by megakaryocytes and platelets. Molecular abnormalities in the TPO gene, however, have been recently identified in several families with an autosomal dominant form of hereditary thrombocytosis where serum TPO levels are significantly elevated. This syndrome has been shown to be due to a mutation in a portion of the TPO gene which plays a crucial role in regulating its expression.

Thrombocytosis can occur in response to many underlying clinical conditions (secondary or reactive), or as an expression of a primary abnormality in bone marrow function (primary). A classification of the causes of thrombocytosis is offered in Table 1. Reactive or secondary thrombocytosis account for over 80 per cent of all recognized cases of thrombo-cytosis (platelet count more than 500 ×10^9/l). Short-lived, secondary thrombocytosis may be observed in situations such as trauma, acute bleeding, major surgery, or after strenuous physical exercise. Longer-term thrombocytosis is associated with the presence of chronic disorders such as malignancy, inflammation, chronic infections, and iron deficiency anaemia. The pathophysiology of reactive thrombocytosis is not fully understood but probably involves the increased generation of inflammatory cytokines such as IL-1, IL-6, GM-CSF and, possibly, TPO. Primary thrombocytosis by contrast is associated with a group of bone marrow disorders including chronic myeloid leukaemia, primary thrombocythaemia, polycythaemia vera, myeloid metaplasia with myelofibrosis, and the myelodysplastic syndromes. The level of elevated platelet numbers is not helpful in differentiating a reactive from a primary process.

For the most part, the presence of an underlying cause for reactive thrombocytosis can be identified by clinical criteria. A number of laboratory tests can be useful in distinguishing primary from secondary thrombocytosis. C-reactive protein synthesis in the liver is mediated by IL-6, with C-reactive protein levels being high in those patients with elevated IL-6 levels. Elevated levels of both IL-6 and C-reactive protein are strongly indicative of the elevated platelet count being reactive. The presence of 'endogenous erythroid colonies' (erythroid progenitor cells that proliferate *in vitro* without the addition of erythropoietin) and of megakaryocyte progenitor cells that are hypersensitive to certain stimulatory cytokines *in vitro* may be helpful in the identification of myeloproliferative disorders. Assays using probes for genes located on the X chromosome might help identify clonal haemopoiesis. In a number of female patients with a primary thrombocytosis, this can indicate the malignant origin of these underlying disorders.

The natural history and prognosis of reactive thrombocytosis is defined by its underlying cause. The thrombocytosis per se is probably inconsequential and does not require specific therapy. The thrombocytosis usually resolves after the treatment of the underlying cause. In contrast, the thrombocytosis due to underlying myeloproliferative disorders can cause life-threatening thromboembolic phenomena and bleeding episodes, and frequently requires specific cytoreductive therapy, emphasizing the need for accurate recognition. Primary or essential thrombocythaemia is the most important cause of primary thrombocytosis.

Primary thrombocythaemia

Primary thrombocythaemia, also known as essential thrombocythaemia or essential thrombocytosis, is a chronic myeloproliferative disorder characterized by marked megakaryocytic hyperplasia. The clinical course is punctuated by episodes of thrombosis and/or bleeding. In 1951, Dameshek suggested that primary thrombocythaemia represented a myeloproliferative disorder and hypothesized its clonal nature. The myeloproliferative disorders are currently thought to represent primary stem cell disorders.

Aetiology and pathogenesis

The causative factors of primary thrombocythaemia are poorly understood. Its pathogenesis involves the abnormal proliferation of a blood cell precursor that differentiates mainly towards the megakaryocytic/platelet compartment. Current evidence suggests that hypersensitivity to stimulatory cytokines such as IL-3, IL-6, TPO, and GM-CSF, coupled with relative insensitivity to inhibitory soluble factors such as transforming growth factor-b (TGF-b), might provoke the expansion of the megakaryocytic progenitor pool. Its clonal origin was initially established through biochemical isoenzyme characterization of the blood cells of affected women who were heterozygous for glucose-phosphate dehydrogenase (G-6PD). More recently, analysis of X-linked restriction fragment length polymorphisms in affected women has confirmed a clonal pattern in some cases. Recent evidence, however, suggests hat a significant number of patients have polyclonal myelopoiesis and that these non-clonal cases may have a decreased risk for thrombosis. This information therefore suggests that primary thrombocythaemia, as diagnosed by the currently accepted clinical and laboratory criteria, is a heterogeneous disorder in terms of clonality. The pathogenesis of such sustained non-reactive polyclonal thrombocytosis is as yet unclear.

Epidemiology

The true incidence of primary thrombocythaemia is unknown. Approximately 6000 new cases are identified each year in the United States. There seems to be a slight female predominance and the usual age at onset is between 50 and 60 years. Approximately 20 per cent of all cases occur in individuals younger than 40, but it is very rarely seen during childhood.

Pathobiology

The characteristic clinical features are dominated by the thrombocytosis and abnormalities in platelet function. The association between increased numbers of circulating platelets and ischaemic episodes remains unclear, but duration of thrombocytosis may play a role. Microvascular thrombosis results in a variety of clinical

syndromes associated with digital and cerebrovascular ischaemia. Abnormalities in platelet aggregation occur in 35 per cent to 100 per cent of patients, and prolongation of the bleeding time in 7 per cent to 19 per cent. Despite being common, these abnormalities are poor predictors of bleeding and/or thrombotic risk. This is in contrast to the acquired von Willebrand syndrome and erythromelalgia, clinical entities not infrequently seen in association with primary thrombocythaemia. In acquired von Willebrand syndrome, extreme thrombocytosis (more than 1000×10^9/l) induces the adsorption of larger von Willebrand multimers on to platelet membranes and their subsequent degradation, triggering a haemostatic defect that induces a bleeding diathesis quite similar to that observed in Type II von Willebrand disease. Erythromelalgia refers to redness and burning pain in the extremities which results from platelet-mediated thrombosis of the arterial microvasculature. If left untreated it may progress to frank gangrene. Its exquisite response to cyclo-oxygenase inhibitors such as aspirin and indomethacin suggests that prostaglandin endoperoxides produced by the metabolism of arachidonic acid play a major role in the generation of platelet-associated thrombosis.

Clinical features

As many as two-thirds of patients are asymptomatic when diagnosed. Most symptomatic patients present with either a thrombotic episode or a minor bleeding episode. Bleeding can occur spontaneously but is frequently associated with the recent use of a non-steroidal anti-inflammatory drug. Common sites of haemorrhage include the gastrointestinal and the genito-urinary tracts as well as easy bruisability of the skin. Thrombosis leads to the most common presenting symptoms and can occur in arteries and veins, large or small. Occlusion of the splenic vessels and of the superficial and deep veins of the lower extremities is common. Pulmonary emboli may also occur. An occasional patient presents with thrombosis of the hepatic veins causing the Budd–Chiari syndrome or with occlusion of the renal veins manifesting clinically as nephrotic syndrome.

When the microcirculation is involved, a number of clinical syndromes may occur. Palpable lesions with small areas of gangrene indistinguishable from vasculitic lesions of rheumatoid arthritis or systemic lupus erythematosus may be observed. Erythromelalgia may occur in association with transient ischaemic attacks or acute episodes of cardiac angina. Peripheral pulses are usually preserved; this helps differentiate erythromelalgia from atherosclerotic-related ischaemia. Neurological symptoms are common and include headaches and paresthesias of the extremities. Transient ischaemic attacks may present with symptoms of unsteadiness, dysarthria, dysphoria, motor hemiparesis, scintillating scotomas, amaurosis fugax, vertigo, dizziness, migraine headaches, and seizures. On occasion, transient ischaemic attacks may progress to established infarcts. Myocardial ischaemia with normal angiograms occurs occasionally. Thrombotic non-bacterial endocarditis, usually affecting the mitral or aortic valves, may manifest with findings of distal emboli. Splenic enlargement is often seen. Patients unaware of their diagnosis who have undergone splenectomy as part of the diagnostic work-up for splenomegaly will predictably develop extreme increases in their platelet counts with a consequent increased risk for bleeding and/or thrombosis.

Laboratory diagnosis

Elevated platelet counts, often above 600 to 1000×10^9/l are characteristic. The absolute number of platelets, even if higher than 1 million/µl, is not diagnostic of primary thrombocythaemia. Extreme increments have been observed in reactive thrombocytosis. Marked changes in platelet morphology, which include large and bizarre-looking platelets sometimes forming aggregates, are also characteristic and may be more useful in helping distinguishing primary from reactive thrombocytosis. The bone marrow is hypercellular with megakaryocytic hyperplasia. Clusters of megakaryocytes are often observed. Absent or diminished iron stores are seen frequently. This may be an epiphenomenon of an underlying myeloproliferative disorder or a true expression of iron depletion in patients with chronic bleeding. Reticulin is present in one-quarter of bone marrow specimens but collagen is usually absent. Mild leucocytosis is common.

Platelet function abnormalities are commonly found and include defective platelet aggregation in response to adrenaline, ADP, and collagen. Aggregation in response to arachidonic acid and ristocetin is often normal. An acquired platelet storage pool disease also occurs due to abnormalities in the content and release of a granules associated with a state of increased platelet activation. The bleeding time is occasionally prolonged but does not predict bleeding risk. Cytogenetic evidence for a Philadelphia chromosome and/or the molecular identification of the *brc/abl* fusion gene aids in distinguishing primary thrombocythaemia from chronic myeloid leukaemia. The presence of dyspoietic changes in bone marrow precursor cells and of characteristic chromosomal abnormalities suggests the diagnosis of myelodysplasia. The diagnostic criteria and management of the other myeloproliferative disorders associated with thrombocytosis are outlined in other chapters. Cytogenetic abnormalities occur in approximately 5 per cent of patients with primary thrombocythaemia. The most common chromosomal alterations include 1q–, 20q–, 21q–, and 1q+. Elevated vitamin B-12 levels occurs in 25 per cent of patients.

Diagnostic criteria and differential diagnosis

Diagnostic criteria are listed in Table 2. The exclusion of an identifiable cause of reactive thrombocytosis is a necessary condition. Primary thrombocythaemia is mainly a diagnosis of exclusion. Any condition associated with elevations in circulating platelets is part of the differential diagnosis. Thrombocytosis may be the consequence of primary bone marrow disorders associated with increases in platelet production (non-reactive thrombocytosis), or a secondary response to an underlying disorder (reactive thrombocytosis). Table 1 summarizes the most important. Clearly, secondary causes of thrombocytosis occur more frequently (more than 80 per cent). Infection, hyposplenism, malignancy, trauma, and non-infectious inflammation are the most commonly encountered disorders. Chronic myelogenous leukaemia and primary thrombocythaemia are the most frequent causes of primary thrombocytosis.

Risk assessment

Primary thrombocythaemia is associated with a very low risk of life-threatening complications. Most patients enjoy survival fairly similar to that of their unaffected peers. Exposure of every patient with primary thrombocythaemia to myelosuppresive therapy is unwarranted. A risk-based decision approach to therapy is outlined in Table 3. Advanced age (60 years or older) and previous history of thrombosis clearly define a group at high risk for the development of life-threatening complications. The degree of thrombocytosis and the presence of associated cardiovascular risk factors, particularly smoking and obesity, is also taken into consideration when making treatment decisions.

Treatment

A number of agents can lower the platelet count of patients with primary thrombocythaemia. There is now firm evidence that cytoreduction using hydroxyurea, at least in high-risk patients, results in a significant reduction in the number of thrombotic episodes. However, only a handful of randomized clinical trials have been conducted given the relatively small number of patients diagnosed with this disease each year. Alkylating agents have been extensively used in the past to treat primary thrombocythaemia. Within this group of agents, busulfan has been shown to be quite effective and relatively non-toxic, with predictable cytopenias as its major untoward effect. It is usually prescribed at 4 mg/day until a platelet count of 400 000/µl is reached. Additional 2-week courses are given if and when the platelet count rises over 400 000/µl. Extensive experience has also accumulated with radioactive phosphorus (^{32}P). Its advantages include ease of administration and the relative absence of significant, acute side-effects. It is usually given as a single dose of 2.3 mCi/m^2 that may be repeated in 3 to 6 month. Its effects are seen 4 to 8 weeks after administration. Excessive doses of ^{32}P can be associated with cytopenias. Alkylating agents and ^{32}P have been associated with significant increases in the risk of leukaemic transformation.

The use of hydroxyurea, an antimetabolite that interferes with DNA repair, decreased the number of thrombotic events in a randomized study of high risk patients when given at 15 mg/kg initially with subsequent adjustments based on initial response. In this study, the target was a platelet count of less than 600 000/µl, but it is possible that tighter control (less than 350 000) may be more effective. Onset of action is usually 3 to 5 days. Frequent side-effects include dose-related neutropenia, nausea, stomatitis, hyperpigmentation, rash, nail changes, leg ulcers, and hair loss. Its leukaemogenic potential when given as a single agent is still a subject of major controversy although it is clearly less leukaemogenic than alkylating agents or ^{32}P.

Interferon-a, a biological response modifier, also is useful in treating patients with primary thrombocytosis. Ninety per cent response rates with median times to response of approximately 3 months are seen when 3 to 5 million units are administered subcutaneously 3 to 5 days per week. It is non-mutagenic and does not cross the placenta. Frequent side-effects include flu-like symptoms, fatigue, lethargy, and depression. Its long-term use is associated with mild weight loss, alopecia, autoimmune thyroiditis, and autoimmune haemolytic anaemia. Interferon's extensive toxicity profile and the need for parenteral administration limit its use as initial therapy, particularly in elderly patients.

Anagrelide is at present the treatment of choice for younger patients and acts by selectively inhibiting megakaryocytic maturation. Responses have been documented in over 90 per cent of treated patients with a median time to response of 2.5 to 4 weeks and an onset of action of 6 to 10 days. It is non-mutagenic and its use has not been associated with the development of acute leukaemia. The usual initial dose is 0.5 mg, two to four times per day, which is increased at 0.5-mg increments per week according to response. Excessive dosing predictably causes thrombocytopenia. The current recommendation is not to exceed a total of 10 mg/day or 2 mg

(single) doses. The average daily maintenance dose is 2 mg in divided doses. The most common side-effects include headaches, dizziness, fluid retention, palpitations, nausea, abdominal pain, and diarrhoea. They develop within 2 weeks of use and usually improve within 2 weeks of continued treatment. It may also, on occasion, trigger episodes of tachyarrhythmias and heart failure. For this reason, it should be used carefully in the elderly and avoided in patients with known heart disease.

Recent evidence suggests that the administration of low doses of aspirin (250 mg daily, or less) is safe, and may decrease the recurrence of microcirculatory events (erythromelalgia/ transient ischaemic attacks) and prevent the development of other thrombotic phenomena, especially in combination with myelosuppressive agents. These data are still preliminary and require further study in large, randomized trials. In order to minimize the risk of iatrogenic bleeding, only patients with platelet counts less than 1 000 000/µl and without evidence of an acquired von Willebrand syndrome should be considered for low-dose aspirin administration.

Given the number of available therapeutic options and their different toxicity profiles, the choice of the appropriate cytoreductive drug for a given individual requires the consideration of a number of variables. These include age, childbearing potential, projected life expectancy, co-morbid-ities, and cost of treatment. Furthermore, the overall low risk for the development of life-threatening complications that affects patients with primary thrombocythaemia highlights the need for systematic, risk-based approaches to therapeutic decision making (see Table 3). The optimal therapy for patients with primary thrombocythaemia remains unclear. All patients should stop smoking. Indiscriminant use of high doses of non-steroidal anti-inflammatory agents should be avoided. Their excessive use is clearly associated with bleeding episodes.

Low-risk patients have a risk of thrombosis similar to that of the age and sex-matched population and a very low risk of life-threatening bleeding, supporting close observation without cytoreductive therapy as the most sensible approach. Hydroxyurea is an adequate choice for patients 60 years of age or older who are otherwise in good health. For elderly patients with limited projected survival (less than 10 years) and who have problems with either drug compliance or are too ill to comply with the minimum follow-up requirements during cytoreductive therapy, ^{32}P administration might be appropriate. Anagrelide should be offered to younger patients (less than 60) who are at high risk by virtue of a prior history of thrombosis or to patients at intermediate risk who the physician feels the necessity to treat. This drug should be used with extreme caution in elderly individuals and should be avoided in patients with cardiac disease. a-Interferon may be an acceptable option in the younger population of patients but is usually not used initially given the need for parenteral administration and its prominent side-effect profile. Alkylating agents, ^{32}P, and hydroxyurea are usually avoided in younger patients given their known (alkylators and ^{32}P) and possible (hydroxyurea) leukaemogenic potential. If a young patient, however, is resistant or intolerant to a-interferon and/or anagrelide, and requires treatment, we feel comfortable prescribing hydroxyurea at this point. In patients at intermediate risk based on platelet numbers at or more than 1 000 000/µl and who have the acquired von Willebrand syndrome, platelet reduction therapy is indicated to avoid the high risk of haemorrhage.

Smokers and obese individuals, unless symptomatic, should be managed by risk modification. In patients who suffer from thrombotic episodes, especially episodes involving the microcirculation or large vessels, we usually administer low-dose aspirin (100 mg/day). This dose appears safe and is effective in the treatment of thrombotic events, and is usually given in addition to cytoreductive therapy. In severe, life-threatening episodes, rapid cytoreduction may be achieved by plateletpheresis or by the administration of a single dose of 0.4 mg/kg of nitrogen mustard. In patients who present with a life-threatening, acute bleeding episode, the site of bleeding should be promptly identified and any antiplatelet agent should be stopped. Those suffering from an acquired von Willebrand's syndrome, can be treated with desmopressin (DDAVP) and factor VIII concentrates that contain high concentrations of von Willebrand factor. Cytoreductive therapy with hydroxyurea must be promptly initiated. In bleeding patients who fail to respond to DDAVP and factor VIII administration, the bleeding frequently resolves following platelet transfusions.

The management of patients who are or want to become pregnant requires special consideration. The risk of fetal loss is quite high (approximately 40 per cent). No clinical features other than the previous history of a miscarriage are predictive. Patients at low or intermediate risk should be managed by observation. Specific treatment should be considered during subsequent pregnancies if fetal loss were to occur. Uncontrolled studies have suggested that the careful use of heparin, aspirin, or a-interferon may decrease the chance for miscarriages. These data require further confirmation. Patients at high risk (for maternal thrombosis) are candidates for cytoreduction. Despite the lack of endorsement by the manufacturers of a-interferon, it may be the drug of choice during pregnancy given its lack of mutagenic potential and its inability to cross the placenta. Hydroxyurea, given its mechanism of action, could theoretically cause fetal malformations and anagrelide, due to its small molecular size, probably crosses the placenta and may cause life-threatening thrombocytopenia and haemorrhage in the fetus. Despite these concerns, recent case reports have described first trimester exposures to these two drugs resulting in the delivery of normal newborns. We therefore do not consider unintended exposures to hydroxyurea or anagrelide as absolute indications for the termination of pregnancy.

Prognosis

The probability that a patient with primary thrombocythaemia will survive 10 years is within the range of 64 to 80 per cent and is not substantially different from that of a control age- and sex-matched population. The actual risk for the development of a catastrophic thrombotic or haemorrhagic event in an asymptomatic patient is quite low. The majority of deaths come from thrombotic complications. Transformation to myelofibrosis and/or acute leukaemia has been reported with increasing frequency at a rate of transformation of 3 to 10 per cent. Prior administration of cytotoxic therapy is the strongest predictor of evolution to leukaemia but spontaneous transformations also occur, as in other myeloproliferative disorders. In rare instances, primary thrombocythaemia may also evolve into a clinical picture that resembles one of the other chronic myeloproliferative disorders.

Future directions

Better means to establish thrombotic and/or bleeding risk are required, and will help to improve the individualized risk-based selection of therapy. It is now clear that high-risk patients do benefit from cytoreductive therapy. Randomized comparisons to establish the efficacy of a-interferon and anagrelide in younger individuals at high risk is still needed. Although there is clear evidence that low-dose aspirin is safe and useful for the treatment of microcirculatory events such as erythromelalgia and transient ischaemic attacks, its additional therapeutic contribution when used in combination with anagrelide, hydroxyurea, or a-interferon requires further confirmation. The ideal therapeutic target in terms of the most desirable platelet number also requires better definition. A better understanding of the mechanisms involved in the regulation of platelet production and of the molecular abnormalities specifically associated with primary thrombocythaemia will offer rational targets against which to develop new and more specific therapies.

Further reading

Barbui T *et al*. (1996). Treatment strategies in essential thrombocythemia: a critical appraisal of various experiences in different centers. *Leukemia and Lymphoma* **22** (Suppl.1), 149–60.

Buss DH *et al*. (1994). Occurrence, etiology and clinical significance of extreme thrombocytosis: a study of 280 cases. *American Journal of Medicine* **96**, 247–53.

Cortelazzo S *et al*. (1995). Hydroxyurea for patients with essential thrombocythemia and a high risk of thrombosis. *New England Journal of Medicine* **332**, 1132–9.

Greisshamer M, Heimpel H, Pearson TC (1996). Essential thrombocythemia and pregnancy. *Leukemia and Lymphoma* **22** (Suppl.1), 57–63.

Harrison CN *et al*. (1999). A large proportion of patients with a diagnosis of essential thrombocythemia does not have a clonal disorder and may be at lower risk of thrombotic complications. *Blood* **93**, 417–24.

Hoffman R. Primary thrombocythemia. In: Hoffman R, Benz EJ, Shattil SJ, Furie B, Cohen HJ, Silberstein LE, Mc Glave P, eds. *Hematology: basic principles and practice*, pp. 1188–204. Churchill Livingston, Philadelphia.

Kaushansky K (1995). Thrombopoietin: the primary regulator of platelet production. *Blood* **86**, 419–31.

Kondo T *et al*. (1998). Familial essential thrombocythemia associated with one-base deletion in the 5'-untranslated region of the thrombopoietin gene. *Blood* **92**, 1091–6.

McIntyre CJ *et al*. (1991). Essential thrombocythaemia in young adults. *Mayo Clinic Proceedings* **66**, 149–54.

Murphy S *et al*. (1997). Experience of the Polycythemia Vera Study Group with essential thrombocythemia: a final report on diagnostic criteria, survival and leukemic transition by treatment. *Seminars in Hematology* **34**, 29–39.

Ruggeri M *et al*. (1998). No treatment for low-risk thrombocythaemia: results from a prospective study. *British Journal of Haematology* **103**, 772–7.

Silverman MN and Tefferi A (1999). Treatment of essential thrombocythemia with anagrelide. *Seminars in Hematology* **36** (Suppl. 2), 23–5.

Tefferi A *et al*. (1994). Plasma interleukin-6 and C-reactive protein levels in reactive versus clonal thrombocytosis. *American Journal of Medicine* **97**, 374–8.

Tefferi A *et al*. (1997). New drugs in essential thrombocythemia and polycythemia vera. *Blood Reviews* **11**, 1–7.

Tefferi A, Hoagland HC (1994). Issues in the diagnosis and management of primary thrombocythemia. *Mayo Clinic Proceedings* **69**, 651–5.

Van Genderen PJJ *et al*. (1996). The reduction of large von Willebrand factor multimers in plasma in essential thrombocythemia is related to the platelet count. *British Journal of Haematology* **93**, 962–5.

Van Genderen PJJ *et al*. (1996). Acquired von Willebrand disease in myeloproliferative disorders. *Leukemia and Lymphoma* **22** (Suppl.1), 79–82.

Van Genderen PJJ *et al*. (1997) Prevention and treatment of thrombotic complications in essential thrombocythemia: efficacy and safety of aspirin. *British Journal of Haematology* **97**, 179–84.

22.3.11 Aplastic anaemia and other causes of bone marrow failure

E. C. Gordon-Smith

Introduction

The concept of bone marrow failure as a cause of peripheral blood cytopenias is imprecise but convenient. Broadly, it indicates that the cause of the peripheral blood disturbance lies within the dividing pool of cells in the marrow itself. Fundamental to the classification of disorders within this group is the idea that normal development of cells within the bone marrow and release of cells into the peripheral blood depend upon an interaction between haematopoietic cells and the environment in which they proliferate and differentiate (see Chapter 22.2.1). The pathogenesis of most of these disorders is unknown and their separation depends mainly upon morphological criteria. The classification shown in Table 1 attempts to group the syndromes where there is a failure of circulating cell production according to the presumed cell stage involved and indicates inherited or congenital syndromes which mimic the acquired disorders.

Aplastic anaemias

Aplastic anaemia is defined by peripheral blood pancytopenia associated with a hypocellular marrow in which the normal haematopoietic tissue is replaced to a greater or lesser extent by fat cells. Remaining cells, both in the peripheral blood and bone marrow, appear morphologically normal and there is neither fibrosis nor infiltration by malignant cells in the marrow. Vitamin B_{12} and folate levels are normal and the disorder is not associated with other dietary deficiencies.

Classification

As defined above, aplastic anaemia may occur in a number of ways. There is no universally acceptable classification but a number of more or less well-defined entities may be identified (Table 2). In each of these, the pathogenesis seems to involve damage to the early haemopoietic progenitor cells, either stem cells or early lineage-committed progenitor cells.

Inevitable aplastic anaemia

Myelosuppression occurs following exposure to cytotoxic drugs or irradiation. The severity and duration of aplasia depends upon the nature of the cytotoxic agent and is dose related. Recovery usually occurs 1 to 6 weeks after the cytotoxic agent is discontinued. With very-high-dose radiation and certain cytotoxic agents, which do not depend on cell cycling for their action, stem cell killing may be complete and recovery does not occur.

Acquired aplastic anaemia

Incidence and epidemiology

Aplastic anaemia is a rare disease. In Europe and the United States, the incidence is probably between 2 and 4 per million of the population per year. All age groups may be affected, possibly with peaks between 20 and 30 and again in older patients. There is a slight preponderance of males, possibly reflecting the greater risk of exposure to toxic substances at work amongst men. In the Far East, the incidence of aplastic anaemia is two to three times higher and the male preponderance much more obvious. The risk factors seem to be environmental rather than genetic since people of the same ethnic groups in the West have a lower incidence.

Aetiology

In about two-thirds of cases of aplastic anaemia it is not possible to identify any likely cause. Amongst the rest, drugs, viruses, and environmental toxins may be identified as probable causes. There is no test to pinpoint the cause of the aplastic anaemia. The implication of any particular agent depends upon temporal associations and previous reports. Drugs have been implicated in the aetiology for 50 years or more and the list of drugs is long. In many cases, the association is weak and often confounded by the patient having received other drugs at the same time. The agents most commonly implicated are some antibiotics, of which chloramphenicol is the best known, and drugs used in the treatment of rheumatoid arthritis including non-steroidal anti-inflammatory agents (NSAIDS) and disease modifying agents in rheumatoid disorders (DMARDs). Table 3 includes a list of the more frequently reported drugs.

Viruses are also implicated. Up to 10 per cent of patients in Western and Japanese series of aplastic anaemia, particularly in the younger age group, give a history of jaundice and/or hepatitic symptoms some 6 weeks before the pancytopenia develops. In many instances, disturbances of hepatocellular function consistent with viral hepatitis have been demonstrated. However, a specific hepatitis virus is not usually identifiable. Aplastic anaemia has occurred following liver transplantation for fulminant infectious hepatitis but not other causes of liver failure. Occasional reports of aplastic anaemia following Epstein–Barr or other viruses have been published.

Various domestic and recreational drugs and chemicals have been implicated, though the evidence against any particular agent is not always very convincing. Wood preservatives, pesticides, and various organic solvents have been associated with the disease. Benzene is known to produce proliferative dysplasias, including acute leukaemias, when exposure is high. Its part in causing aplasia is not so certain.

Pathogenesis

The way in which the various agents bring about aplastic anaemia is unknown. The main defect is a failure of the pluripotent haematopoietic stem cells in the bone marrow to proliferate and differentiate into mature blood cells. *In vitro* experiments with long-term bone marrow culture show that aplastic marrow stroma cells are able to support haemopoiesis from normal stem cells but aplastic stem cells continue to grow abnormally on normal stroma. There is indirect evidence that cellular immune reactions play a role in the pathogenesis or at least the perpetuation of stem cell damage. The strongest evidence for this is the response of aplastic anaemia to

immunosuppressive treatment.

Diagnosis and pathology

The diagnosis of aplastic anaemia is made on the basis of the peripheral blood and bone marrow findings and by excluding other causes for pancytopenia. In the peripheral blood there is pancytopenia with no abnormal cells present. The anaemia is usually normocytic at presentation but becomes macrocytic, even strikingly so, in chronic cases. The reticulocyte count is low. Neutrophils are invariably reduced and the count may be very low. Circulating neutrophils may have rather heavy granulation, so-called 'toxic' granulation, and have a high alkaline phosphatase content. The eosinophils, basophils, and monocytes are usually also depleted. The reduction in the lymphocyte count is more variable; in children particularly it may be relatively high so that the total white cell count may be normal. The platelet count is reduced.

Aspiration of bone marrow is usually easy, fragments are obtained which are fatty, and there is a reduction of haematopoietic cells in the trails. The cellularity of the marrow may be judged to some extent from the marrow aspirate, but a so-called 'dry-tap' (no material obtained from an aspirate) or 'blood-tap' (no fragments obtained) does not allow an assessment of bone marrow activity. In aplastic anaemia, there may be a patchy loss of cellularity throughout the marrow so that an aspirate may yield relatively normal-looking marrow. The diagnosis cannot therefore be made on a bone marrow aspirate alone. Assessment of cellularity is made on a trephine biopsy, which shows replacement of the normal cellular marrow by fatty marrow (Fig. 1). The reticulin network of the marrow is reduced commensurate with the reduction in the general overall cellularity. Focal areas of preserved cellularity may be seen in the bone marrow trephine, so called 'hot pockets' (Fig. 1). The morphology of remaining haematopoietic progenitor cells is broadly normal though there may be some changes in the erythroid precursors which constitute mild dyserythropoiesis. Erythro-phagocytosis by macrophages may be prominent, especially early in the disease. Megakaryocytes are often absent but when present have normal maturation and morphology.

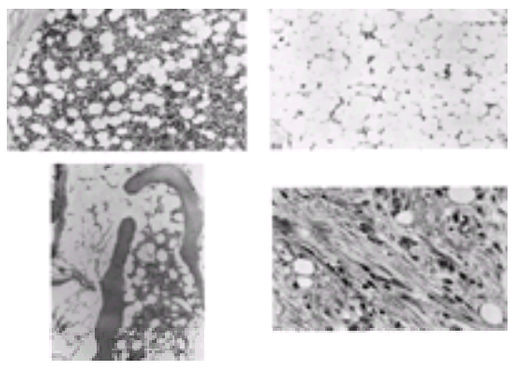

Fig. 1 Trephine biopsies of adult posterior iliac crest: (a) normal marrow; (b) severe aplastic anaemia; (c) cellular focus in severe aplastic anaemia; (d) proliferative dysplasia with fibrosis.

Clinical features

The clinical features of aplastic anaemia arise from the deficiencies of the cellular elements of the blood. There are no specific physical signs. Haemorrhagic manifestations are common at presentation. The development of thrombocytopenia takes place over a matter of weeks or months so that catastrophic haemorrhage as a presenting feature is unusual; minor signs of the bleeding tendency, easy bruising, gum bleeding, or purpuric rash, are more usual. Haemorrhages in the buccal mucosa may occur; retinal haemorrhages may be a portent of serious bleeding. The anaemia also develops slowly and the patient may complain only of mild fatigue or shortness of breath on marked exertion. Infections, particularly of the oropharynx or upper respiratory pathways, may be a presenting feature. Infections anywhere aggravate the effect of thrombocytopenia, particularly in the mouth. If the aplastic anaemia has followed an episode of apparent hepatitis, there may be some residual jaundice with enzyme abnormalities consistent with post-hepatitic cholestasis.

The progression of the disease is variable and depends upon the severity and completeness of the marrow damage. In earlier series of patients with aplastic anaemia where only support in the form of transfusions and available antibiotics was given, about half the patients died within 3 to 6 months as a result of infection or haemorrhage. Patients alive at a year, however, had a better chance of surviving, at least for the next 2 or 3 years. This suggested that there was a group of patients with severe disease with a very poor chance of recovery and another group with a milder disorder. This lead to the establishment of criteria for severe aplastic anaemia (Table 4) which have proven to be useful in stratifying patients when different treatments have been compared. The designations severe aplastic anaemia (neutrophils less than 0.5 but more than 0.2×10^9/l) and very severe aplastic anaemia (neutrophils less than 0.2×10^9/l) are useful in planning therapy.

The natural history of the disease has been so modified by improvement in transfusion support and infection control that it is difficult to decide, in some cases, whether subsequent events are part of the disease or the consequences of treatment. Spontaneous recovery, apparently to complete normality, may occur even after several years of pancytopenia. Other patients may remain stable for many years before haematopoietic activity decreases further. The emergence of abnormal clones, both benign and malignant, transient or progressive, is common. Paroxysmal nocturnal haemoglobinuria (PNH) is the most frequent. PNH arises from a somatic mutation involving a gene on the X chromosome which codes for a protein involved in the assembly of phosphatidyl inositol glycan (PIG), which anchors many proteins to the surface of cell membranes. Two enzymes which inactivate complement complexes, decay accelerating factor (DAF; CD59) and membrane inhibitor of reactive lysis (MIRL; CD55) are absent in PNH red cells, which then become sensitive to lysis by complement. PNH may be recognized by the Ham's test or directly by using fluorescence-activated cell scanning to identify populations which lack the PIG anchored proteins. Myelodysplastic syndromes and acute myeloid leukaemia may also develop following aplastic anaemia and the relationship between these blood diseases is discussed below.

Until platelet transfusions became readily available, the usual cause of death in these patients was haemorrhage. Most patients who fail to respond to treatment now succumb to infection or a mixture of infection and haemorrhage, often after many months of treatment with antibiotics. It is virtually impossible to eradicate infection in the severely neutropenic patient until such time as neutrophil production returns.

Treatment

The treatment of aplastic anaemia has two main components. The first is to protect and support patients from the consequences of pancytopenia. The second is to try to accelerate the recovery of the bone marrow by whatever means without eradicating the chance of spontaneous recovery.

Support and protection

For the aplastic patient this depends upon reducing potential sources of infection to a minimum and replacing deficient cells by transfusion (see Section 7). Infections may arise from the environment or from sources of bacteria and other agents with the patient. As with all immunosuppressed patients, significant and lethal infections may arise from contamination with organisms which are not normally pathogenic. Exogenous infections are more likely in a hospital environment than at home, so any patient with aplastic anaemia admitted to hospital must be nursed in a clean, and preferably sterile, area. Measures to prevent nosocomial infections should be of the highest standards. Virus infections are not in themselves especially likely in the neutropenic host, but if they occur they produce an environment in which secondary bacterial infections may flourish. When the neutropenic patient is also immunosuppressed in other ways, virus infections assume a very important role in causing morbidity. Patients with non-severe aplastic anaemia are not at greatly increased risk of opportunistic infection.

Endogenous infections arise from organisms carried within the patient, particularly the upper respiratory passages and the gastrointestinal tract. Scrupulous attention to oral hygiene minimizes the risk of infection from this source and diminishes gum bleeding. The extent to which potential pathogens should be removed from the gastrointestinal tract is debatable. Mostly these are aerobic organisms which are easily eliminated by antibiotics. Some would argue that removal of anaerobic bacteria may actually be harmful. So-called complete decontamination of the gut is achieved by giving a variety of non-absorbable antibiotics together with antifungal agents such as nystatin or amphotericin. Co-trimoxazole or ciprofloxacin, together with an antifungal agent, may be equally effective in eliminating most aerobic pathogens although it has yet to be demonstrated conclusively that this prevents infections. Recolonization of the bowel by potential pathogens can be avoided by

using freshly cooked, low bacterial food. It should be remembered that patients with aplastic anaemia may require months of protective isolation and therefore measures must be practical as well as effective.

Once an infection is established, it is essential to treat it as soon as possible. Systemic antibiotics, particularly to treat Gram-negative organisms, must be given as soon as fever or signs of infection occur and appropriate samples have been sent to the laboratory. Delay in the severely neutropenic patient may be fatal. Gram-positive infections, mainly with coagulase-negative staphylococci, are now common because of the extensive use of indwelling central venous lines, but rarely lead to rapidly progressive endotoxic shock. Since the most common exogenous infections arise from *Pseudomonas* or *Klebsiella* spp. and the endogenous ones from aerobic organisms of the gastrointestinal tract, the antibiotics used in the first instance must be appropriate to those organisms. Most centres use a combination of aminoglycoside with a second antibiotic likely to have activity against *Pseudomonas* or a third-generation cephalosporin (suitable regimens are described in Section 7). A difficulty in aplastic anaemia is to decide when to discontinue the antibiotics. The patient may become afebrile and apparently well, but when the antibiotics are stopped, infection by the original organism is all too likely to return unless the neutropenia recovers. Granulocyte stimulating cytokines, filgrastim or lenograstim, or granulocyte–macrophage stimulating cytokine, rhGM-CSF, may stimulate the remaining bone marrow sufficiently to raise the neutrophil count enough to eradicate the infection. Granulocyte transfusions do not seem to be helpful (see Section 7).

Transfusion of red cells and platelets is the other main standby in the management of aplasia. Red cell transfusions usually present few problems, but it must be remembered that the platelet count will fall and haemorrhage may occur during such transfusions. Platelets should always be given with red cell transfusion in the severely pancytopenic patient. Repeated platelet transfusion leads to the development of antibodies and resistance to platelet concentrates in about 40 per cent of patients. This complication is reduced by using white-cell-depleted products. The antibodies may be anti-HLA or antiplatelet-specific antigens. Resistance is indicated by an inability to raise the platelet count by platelet transfusion. Conventionally, platelets are only transfused when there is a clinical indication for their use. Indications include the rapid development of purpura, extensive bleeding from the gums and in the buccal mucosa, retinal haemorrhages, and headache. In aplastic anaemia, particularly when the patient is being managed on an outpatient basis, catastrophic and fatal haemorrhage may be the first indication of severe bleeding, particularly so if the patient develops an infection. For this reason centres manage their outpatients with regular platelet transfusions to maintain a count above 15 × 10^9/l.

Further details of the management of patients with marrow failure are given elsewhere.

Specific measures

There are two main approaches to the treatment of aplastic anaemia. Haematopoietic stem cell transplantation is curative but carries a high risk of treatment-related mortality and morbidity and is only available to patients with a suitable HLA-matched allogeneic donor. Immunosuppressive treatment, with antilymphocyte or antithymocyte globulin (ALG, ATG) and/or cyclosporin, is the other treatment option. The choice of treatment for any individual with aplastic anaemia depends on the age of the patient, the severity of the disease, and the availability of a suitable donor.

Haematopoietic stem cell transplantation

Recolonization of the aplastic bone marrow with normal stem cells from a suitable donor has long been considered the most rational treatment for aplastic anaemia. The first successful transplants from HLA-matched siblings for severe aplastic anaemia were carried out in Seattle in 1969 by E. Donall Thomas and colleagues. Subsequent, world-wide experience has shown that such transplants are the most effective treatment for very severe aplastic anaemia and severe aplastic anaemia in patients of suitable age. Patients up to the age of 55 years and, in certain instances, older, with very severe aplastic anaemia should be considered for stem cell transplantation. Children and young adults with severe aplastic anaemia should be offered transplantation as the first choice. The problems of stem cell transplantation for aplastic anaemia are the same as for other conditions, namely graft rejection and graft-versus-host disease. Graft rejection may be increased by sensitization to multiple blood transfusions, so transplants are best carried out early, once the diagnosis has been confirmed, the severity established, and a suitable donor identified. Stem cells for transplantation may be obtained from the bone marrow or from the peripheral blood following mobilization of the stem cells from the marrow by granulocyte colony stimulating factor (G-CSF). Peripheral blood stem cell transplants lead o quicker recovery of peripheral blood counts (at about 14 days compared with 20) but may cause more chronic graft-versus-host disease. On-going trials may establish the superior outcome for one or other source in the future. Conditioning of the patient for sibling transplant is relatively mild in that irradiation is not required. Various regimen have been used, the most wide experience being with intravenous cyclophosphamide 50 mg/kg per day for 4 days before the transplant. Commonly this is combined with ALG given for 4 or 5 days before the cyclophosphomide and continuing up to the transplant. The cell dose given is important. Graft rejection is uncommon after transfusion with greater than 3.0 × 10^6 nucleated cells per kg recipient body weight. Cyclosporin is given for graft-versus-host disease prophylaxis, initially intravenously and subsequently orally, and continued for up to 1 year to prevent late graft rejection. Successful outcome is achieved in about 70 to 80 per cent of cases overall with an incidence of chronic graft-versus-host disease of about 10 to 15 per cent. Children have a success rate of 90 per cent or better. Growth rate, endocrine development, and fertility appear to be normal following this type of transplant and the recovered marrow behaves normally without an increased risk of leukaemia or other clonal disorder.

Problems of bone marrow transplantation are considered further elsewhere.

Immunosuppression

Immunosuppressive treatment for aplastic anaemia was introduced in Europe in 1977, following observations by Georges Mathé in Paris and experimental work by Bruno Speck in Basel. Subsequent controlled trials confirmed that 5 days treatment with ALG was an effective way of achieving remission in all degrees of severity of aplastic anaemia and for all ages. Immunosuppression is the treatment of choice for all patients who are not suitable for transplantation. Recovery following treatment is usually slow with little response before about 3 months and often up to 6 months. Many patients treated in this way still require some transfusion support for this time and may continue with neutropenia and/or thrombocytopenia for many years, though independent of transfusion or hospital care. If the patient fails to respond to the first course of ALG, a second course using an alternative ATG may be given. Some 60 per cent of patients respond to the first course with partial or complete remission and about 40 per cent of non-responders will achieve some improvement with a second. The optimum timing of a second course still has to be determined but most groups wait about 4 to 6 months before a second course. Further courses may be tried in non-responders if they have not been sensitized to the animal protein. There are several preparations of ALG/ATG which are not necessarily bioequivalent so treatment schedules may vary. Reactions during infusion of ALG or ATG are common and may be severe. Serum sickness occurs in some 75 per cent of patients, requiring treatment with corticosteroids. The routine addition of high-dose methylprednisolone (5 mg/kg per day) has no obvious therapeutic advantage and produces a high incidence of avascular necrosis of the hip.

Cyclosporin, 5 mg/kg per day, the dose then adjusted to individual requirement, appears to increase the speed of remission when given after ALG and may also be used alone as an alternative to ALG, though the proportion of responders is less. Recovery, as with ALG, is slow and may be incomplete.

Relapse, or the emergence of PNH or myelodysplastic syndrome clones, leading to a requirement for transfusion, occurs in about 25 per cent of remitting patients over a 10-year period, though some patients respond to further course of immunosuppression. Relapse may follow virus infections, immunizations, or in pregnancy. Some patients seem to be dependent on a continuing dose of cyclosporin post ALG.

Anabolic steroids

Anabolic steroids may be useful in non-severe aplastic anaemia when immunosuppressive therapy fails. The virilizing side-effects make their use unpopular and hepatotoxicity is a problem. A trial of oxymethalone, 2.5 mg/kg per day, or other anabolic steroid in equivalent dose, may be warranted.

Congenital aplastic anaemias

There are a number of inherited disorders which may be associated with bone marrow failure. Table 5 lists the better characterized disorders; familial pancytopenias, which do not fit these diagnoses, sometimes occur.

Fanconi anaemia

The commonest of the inherited disorders which produce aplastic anaemia is that described by Fanconi in 1927. The disorder is inherited as an autosomal recessive and is associated with multiple developmental abnormalities, particularly of the skin and skeleton (Table 6). There is wide genetic and phenotypic heterogeneity.

Cases have been described in all populations.

Genetic basis of Fanconi anaemia

Somatic cell fusion studies have shown that there are at least seven distinct genes identifiable, *FANC(A–G).* Four of these genes have been cloned, *FANCA* (16q24.3), *FANCC* (9q22.3), *FANCF* (11p15), *FANCG* (9p13). For each of these genes, multiple mutations have been described. *FANCG* is identical to a gene, *XRCC9,* which is thought to be involved in cell cycle regulation or postreplication repair but the function of the other genes is unknown and the products of the cloned genes have no homology to each other or to other known proteins. It is presumed that the various gene products are part of a pathway involved in chromosome protection, probably through the formation of a functional complex. The genetic heterogeneity accounts for much of the phenotypic heterogeneity.

Cytogenetic findings

The diagnostic test for Fanconi anaemia is the appearance in metaphase of multiple chromosome breaks in phytohaemagglutinin-stimulated peripheral blood lymphocytes. Breakage rate is increased in baseline cultures but is markedly enhanced when cultures are exposed to clastogens such as diepoxybutane or mitomycin C (Fig. 2). Other inherited disorders which are thought to have underlying chromosome instability and a defect in DNA repair also have an increased tendency to develop acute leukaemia, though not usually with an aplastic phase (see Table 7).

Fig. 2 Chromosomes from a metaphase preparation of peripheral blood lymphocytes from a patient with Fanconi anaemia incubated in the presence of diepoxybutane (DEB) show multiple breaks and rearrangements. ctg, chromatid gap; csg, chromosome gap; ctb, chromatid break; csb, chromosome break. Rearrangements: cte, chromatid exchange; tr, triradial; qr, quadriradial.

Haematological features

Patients with Fanconi anaemia usually have a normal or nearly normal blood count at birth and during infancy. Bone marrow failure develops slowly. The age at which it is manifest clinically depends in part on the underlying genetic cause. In many cases the failure appears between 5 and 10 years, in other families the defect becomes apparent in adolescence whilst in some cases it presents in adult life. A severe form caused by a mutation in *FANCC*, the IVS-4 mutation found in Ashkenazi Jews, has a particularly rapid development of aplasia and a very high transformation to acute leukaemia. In all cases the bone marrow becomes progressively hypocellular, eventually being indistinguishable from acquired aplastic anaemia. Early on, macrophages showing active phagocytosis are prominent, perhaps indicating the removal of cells in apoptosis. Granulopoiesis may be relatively well-preserved. Dyserythropoiesis may be prominent. Evolution to acute leukaemia is common, particularly in some gene types. Patients may present with acute leukaemia, usually myeloid, without a prior period of aplasia. Red cells are macrocytic but there are no specific features in the peripheral blood to suggest the diagnosis.

Clinical features

The features of the full-blown Fanconi anaemia are characteristic (Table 6). There is marked phenotypic variation between patients in different families but considerable similarities within families, which reflects the genetic variation. In some cases, diagnosis may be difficult because of absence of the characteristic skeletal and skin features. Infants are of low birth weight and most remain small-for-age after birth. The skin is often mildly pigmented with areas of deeper pigmentation producing café-au-lait spots, sometimes with areas of depigmentation. Skeletal abnormalities involve particularly the bones of the forearm and thumbs. Abnormalities in the anatomy of the kidneys are also common. Intellectual development is usually normal.

Prognosis and treatment

The outlook in Fanconi anaemia is poor. Untreated, the disease is usually relentless. Despite support with transfusions over many years most patients die of haemorrhage, infection, or of acute leukaemia. Fanconi anaemia should also be suspected in all children and adolescents presenting with acute myeloid leukaemia. Identification of the familial nature of the disease is important for genetic counselling. The median interval between presentation and death is about 2 to 4 years. Patients who survive to adult life have an increased risk of solid tumours of squamous origin, particularly of the tongue, oesophagus, vulva, cervix, and breast. Most adult females are fertile whilst most males are infertile.

Treatment with anabolic steroids may bring about a remission of variable duration. Several years free from transfusion requirements may be obtained, but at the price of virilization and liver toxicity. Hepatocellular carcinoma, often accompanying peliosis hepatis, seems to be particularly common in children treated for years with 17a-alkylated anabolic agents (Fig. 3).

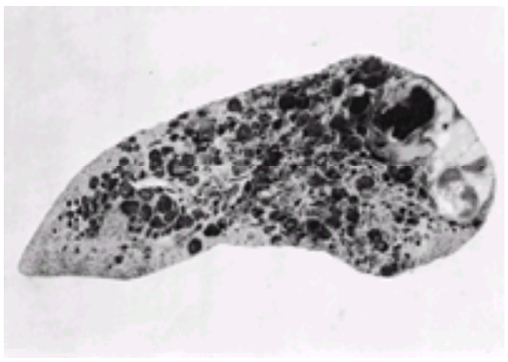

Fig. 3 Hepatocellular carcinoma in the liver of a patient with Fanconi anaemia treated for 4 years with anabolic steroids; the liver also shows multiple venous lakes (peliosis hepatis), another side-effect of anabolic steroids.

Bone marrow transplantation is the only curative form of treatment but carries special risks. The Fanconi anaemia cells are very sensitive to cyclophosphamide and irradiation used to immunosuppress patients prior to transplant and the doses given to these patients have to be greatly reduced. With these modifications, the success of bone marrow transplantation from HLA-matched sibling donors is similar to that for acquired aplastic anaemia. Transplantation from unrelated, HLA-matched donors is less successful but given that no other treatment is effective, should be offered whenever possible.

Dyskeratosis congenita

Dyskeratosis congenita is an inherited disorder involving the mucocutaneous system with the development of bone marrow failure in about 50 per cent of cases. The inheritance is X linked in the majority of cases with the defective gene located at Xq28. The gene mutated in these cases is designated *DKC1*, the protein dyskerin. Dyskerin may take part in the assembly of ribosomes and their export from the nucleus to the cytoplasm. Some families show recessive inheritance with female members also being affected.

Patients have reticular skin pigmentation of the upper body; leukoplakia and nail dystrophy usually appearing in childhood. There is a high incidence of squamous carcinoma of the oropharynx and gastrointestinal tract but not an increased risk of leukaemia. Marrow aplasia develops in the second or third decade. Treatment with anabolic steroids may be temporally effective in some patients as with Fanconi anaemia but most patients become refractory. Stem cell transplantation should be considered as the only possible cure for the haematological problems but results are poor. Late complications post-transplantation include renal failure, pulmonary fibrosis, and diffuse vasculitis.

Aplastic presentation of malignant disease

Aplastic anaemia is one response to bone marrow damage and acute leukaemia is another. At presentation, the distinction between the two may not always be clear, at least on histological and morphological criteria.

Acute lymphoblastic leukaemia (ALL)

ALL may present in a form indistinguishable from aplastic anaemia, usually, but not exclusively, in children. Blasts are not seen in the peripheral blood, and the bone marrow aspirate and trephine are hypocellular without any obvious infiltration by malignant cells. Presentation with severe infection, often of the pharynx, is more common than in acquired aplastic anaemia. The aplasia usually recovers spontaneously, sometimes in response to steroids. Some 6 to 8 weeks later there is the emergence of leukaemic cells in the peripheral blood. Whilst ALL in childhood is the commonest association, aplasia preceding acute myeloid leukaemia in this way has been described, and adults are occasionally affected.

Hypoplastic myelodysplasia

Hypoplastic myelodysplastic syndrome is a disease characterized by a hypocellular marrow in which a small proportion of blasts may be seen, sometimes with occasional blasts in the peripheral blood. The proportion of blasts in the marrow is less than 5 per cent but there may be abnormal aggregations of primitive precursors. Circulating granulocytes, reduced in number, are hypogranular in contrast to the toxic granulation of aplastic anaemia. The condition differs in a number of ways from aplastic anaemia, but the differences may be subtle and there is considerable overlap. The distinction by morphological criteria is subjective. The presence of cytogentically abnormal metaphases has been used to distinguish hypoplastic myelodysplastic syndrome from aplastic anaemia but transient clones may appear in the latter. The condition may remain stable for months or years during which the patient requires transfusions but is otherwise well. The prognosis is in the low-risk group of the international prognostic scoring system. If a suitable bone marrow donor is available, transplantation is indicated. Recent trials have shown that as many as two-thirds of patients may achieve meaningful remissions with ALG, indicating another link with aplastic anaemia.

Proliferative dysplasia with fibrosis

Occasionally, fibrosis of the bone marrow appears without evident underlying cause and in the absence of hepatosplenomegaly or extramedullary haemopoiesis. The condition is characterized by pancytopenia sometimes with the presence of red cell and white cell precursors in the peripheral blood, a leucoerythroblastic picture. Bone marrow aspirate is usually unsuccessful, and a trephine biopsy shows a variable degree of reduction in haemopoietic cells with the marrow replaced by reticulin and fibroblasts. Primitive cells are not seen at this stage of the illness. Some of these patients probably have an unusual form of myelofibrosis, particularly those who present in childhood. Others may have hypoplastic failure in which abnormal megakaryocyte development leads to the fibrosis.

Bone marrow failure affecting single cell lines

There are a number of conditions in which anaemia, neutropenia, or thrombocytopenia develop in isolation as a result of the failure of production by the marrow. The conditions may be inherited or acquired and the main disorders are listed in Table 8. The majority of acquired cytopenias are immune in origin with peripheral destruction so do not represent examples of bone marrow failure.

Amegakaryocytic thrombocytopenia

Thrombocytopenia caused by deficiency of megakaryocytes may be acquired or inherited. Inherited syndromes include amegakaryocytic thrombocytopenia with total absence of radii (TAR syndrome) and other congenital cases with normal skeleton. If children with TAR survive the first year of life, when cerebral haemorrhage is most likely, the platelet count usually increases spontaneously to safe levels and the outlook is good. Acquired amegakaryocytic thrombocytopenia is probably a variant of aplastic anaemia or myelodysplastic syndrome. The condition is rare. About one-third of patients remain thrombocytopenic, another third progress slowly to aplasia, and the remainder develop myelodysplastic syndrome. Immunosuppressive therapy may produce remission and stem cell transplantation may be curative for appropriate patients.

Pure red cell aplasia (PRCA)

PRCA is defined by anaemia with a marked reduction or absence of reticulocytes in which the neutrophil and platelet count are normal. The bone marrow is cellular with normal granulopoiesis and megakaryocytes. There may be a complete absence of red cell precursors or there may be red cell precursors present up to a certain stage of development but not beyond, so-called 'maturation arrest'. Apart from the changes in the red cell series, there are no other abnormalities in the peripheral blood and there is no evidence of peripheral destruction of red cells. The patients are in other respects normal. Both congenital and acquired forms exist.

Congenital pure red cell aplasia—Diamond–Blackfan anaemia

This has also been called rather confusingly 'congenital hypoplastic anaemia' but is better known by its eponym, the Diamond–Blackfan syndrome. In most instances, anaemia is present at birth or is detected shortly afterwards. There is a profound reticulocytopenia often with no reticulocytes present in the peripheral blood. There is macrocytosis and raised HbF. Red cell adenosine deaminase is increased. There is no hepatosplenomegaly. The white count and platelet counts are normal. Skeletal abnormalities may be present, particularly of the head and upper limb. About 50 per cent have no dysmorphic features. There may be disturbances of growth, either inherent in the disease or brought about by anaemia, iron overload, or steroid therapy. There are no abnormalities of the skin or other organs as seen in Fanconi anaemia.

Pathogenesis

There seem to be a number of genetic abnormalities underlying the disorder. About 20 per cent of patients have a family history, usually in earlier generations, suggesting dominant inheritance. The remainder has no such history but adenosine deaminase may be elevated in first degree relatives and sporadic cases may go on to have affected children. One gene in which mutations are associated with Diamond–Blackfan anaemia is located on chromosome 19 and codes for RPS19, a ribosomal protein. The function of the protein in the pathogenesis of Diamond–Blackfan anaemia is unknown. About a quarter of Diamond–Blackfan anaemia families have this defect. At least two other unrelated defects produce the phenotype.

Treatment

Treatment presents many problems. Most of these children, if treated early enough with corticosteroids, will respond. However, if the condition is steroid-dependent, major problems may result from the continued use of corticosteroids in the doses necessary to maintain remission. Patients may become steroid resistant. Some patients fail to respond to corticosteroids, and this seems to be particularly true if the corticosteroids are instituted late in the illness. These patients rely on blood

transfusions for survival. Transfusion will permit normal growth but produces all the problems of iron overload, including delayed or absent puberty. Chelation therapy is required from an early stage. Stem cell transplantation should be considered if there is a matched family donor but the potential donor should be checked for raised adenosine deaminase levels as well as anaemia before being accepted.

During the course of the disease, the spleen may enlarge and transfusion requirements increase. In these patients, splenectomy may reduce transfusion requirements and occasionally is associated with a marked increase in steroid responsiveness or even complete remission. This only seems to apply to those patients whose spleen is enlarged, and relapse may occur after a few years.

Transient erythroblastopenic anaemia in childhood (TEC)

Transient erythroblastopenia of childhood may have a viral aetiology though this is not always clearly demonstrated. The anaemia with reticulocytopenia most often occurs in children from 6 months to 5 years with a peak incidence around 2 years. There is usually a history of preceding viral illness. More than one member of the family may be affected making distinction from Diamond–Blackfan anaemia difficult. The anaemia is normocytic and adenosine deaminase levels are normal. Neutropenia is common. Recovery occurs within a few weeks of diagnosis though the patient may need transfusion in the meantime.

Parvovirus infection

'Aplastic crises' may occur in patients with haemolytic anaemia who develop infection with Parvovirus B 19. The term is confusing because only the red cell series is affected. The virus is tropic for red cell precursors and prevents differentiation. Red cell precursors are large and vacuolated. As antibodies to the virus develop, so the inhibition is removed. The reticulocytopenia lasts for up to 7 days so that the effect is trivial or unnoticed in people with red cell survival of 120 days. In patients with short red cell survival, such as patients with sickle cell disease and other congenital haemolytic anaemias, the effect may be devastating. Anaemia develops rapidly and transfusion may be required urgently. Antibodies to the virus are lacking in the serum initially, followed by an IgM response. The presence of IgG antibodies precludes the diagnosis.

Acquired pure red cell aplasia

This may occur *de novo*, following administration of various drugs, or in association with lymphoma, and about one-third are associated with a thymoma. PRCA may precede, accompany, or follow the development of the thymoma and excision of the tumour has variable effect with no guarantee of recovery of the anaemia. The haematological features of the disorder are similar to that seen in the congenital red cell aplasia, with anaemia and reticulocytopenia associated with absence of red cell precursors or maturation arrest of the red cell series in the bone marrow. There is an unpredictable responsiveness to corticosteroids. Immunosuppression with azathio-prine or cyclophosphamide may be effective. Autoantibodies are thought to play a role in the pathogenesis and occasionally immunoglobulins have been identified which inhibit haem synthesis or prevent the development of red cell colonies *in vitro*. Very rarely antierythropoietin antibodies have been found.

Acquired PRCA may occur in association with common variable hypogammaglobulinaemia and in association with other autoimmune diseases. The presence of such autoimmune phenomena suggests that the patient has a better chance of responding to corticosteroids than in their absence. PRCA may be associated with lymphomas and evidence of an underlying lymphoma may be obtained by finding evidence of immunoglobulin or T-cell receptor gene rearrangement in the marrow even when histological proof is lacking. Occasionally, splenectomy may increase responsiveness to corticosteroids or immunosuppression. An enlarging spleen, which increases transfusion requirements, is an indication for splenectomy. A chronic transfusion regimen with iron chelation may be required for non-responding patients.

Isolated defects in white cell or platelet production

These conditions are described elsewhere and are summarized in Table 7.

Further reading

Barrett J, Sauntharajah Y, Molldrem J (2000). Myelodysplastic syndrome and aplastic anemia: distinct entities or diseases linked by a common pathophysiology? *Seminars in Hematology* **37**, 15–29.

Fanconi G (1967). Familial constitutional panmyelopathy, Fanconi's anaemia (FA). I. Clinical aspects. *Seminars in Hematology* **4**, 233–40.

Gluckman E (1998). Fanconi anaemia. In: Barrett J, Treleaven J, eds. *The clinical practice of stem cell transplantation,* pp. 259–65. Isis Medical Media, Oxford.

Gordon-Smith EC, Issaragrisil S (1992). Epidemiology of aplastic anaemia. *Clinics in Haematology* **5:2**, 475–91.

Marsh J, Gordon-Smith T (1998). Aplastic anaemia. In: Barrett J, Treleaven J, eds. *The clinical practice of stem cell transplantaion,* pp. 238–58. Isis Medical Media, Oxford.

Marsh JC, Gordon-Smith EC (1998). Treatment options in severe aplastic anaemia. *Lancet* **351**, 1830–1.

Schrezenmeier H, Bacigalupo A, eds (2000). *Aplastic anemia. Pathophysiology and treatment.* Cambridge University Press, Cambridge.

Schroeder-Kurth TH, Auerbach AD, Obe G, eds (1989). *Fanconi anemia*. Springer-Verlag, Heidelberg.

Schwartz RS (1994). PIG-A—the target gene in paroxysmal nocturnal hemoglobinuria. *New England Journal of Medicine* **330**, 283–4.

Sieff CA, Nisbet-Brown E, Nathan DG (2000). Review. Congenital bone marrow failure syndromes. *British Journal of Haematology* **111**, 30–42.

Wagner JL, Storb R (1999). Allogeneic transplantation for aplastic anemia. In: Thomas ED, Blume KG, Forman SJ, eds. *Hematopoietic cell transplantation*, 2nd edn, pp. 791–806. Blackwell Science, Oxford.

Wright EG (1999). Inherited and inducible chromosomal instability: a fragile bridge between genome integrity mechanisms and tumourigenesis. Review. *Journal of Pathology* **187**, 19–27.

Young NS, Alter BP (1994). *Aplastic anemia acquired and inherited*. W.B. Saunders, Philadelphia.

22.3.12 Paroxysmal nocturnal haemoglobinuria

Lucio Luzzatto

Definition

Paroxysmal nocturnal haemoglobinuria (**PNH**) is an acquired chronic disorder characterized by persistent intravascular haemolysis, subject to recurrent exacerbations, often associated with pancytopenia, and with a distinct tendency to venous thrombosis. The triad of haemolytic anaemia, pancytopenia, and thrombosis makes PNH a truly unique clinical condition: however, even in the absence of one or more of these manifestations a conclusive diagnosis can be made by appropriate laboratory investigations (see below).

Epidemiology

PNH is encountered in all populations throughout the world, and it can affect people of all socioeconomic groups. The prevalence of PNH is not accurately known: however, it is more rare than the related disorder, acquired aplastic anaemia (**AAA**). A rough estimate of the frequency of PNH is between 1 in 100 000 and 1 in 1 million. It has been suggested that, like AAA, PNH may be somewhat less rare in South East Asia and in the Far East. Most patients present as young adults, but we have seen PNH in a 2-year-old child and in people in their seventies. PNH has never been reported as a congenital disease, and there is no reported evidence of inherited susceptibility. The sex ratio is not far from even.

Clinical features

The patient may seek medical attention because, one morning, she or he has 'passed blood instead of urine'. This distressing or frightening event—the direct evidence of haemoglobinuria—may be regarded as the classical presentation; however, not infrequently the haemoglobinuria may be initially less spectacular, or it is suppressed. Indeed, the patient often presents simply as a problem in the differential diagnosis of anaemia, whether symptomatic or discovered incidentally; this may be associated with jaundice, immediately suggesting it may be a haemolytic anaemia. Sometimes the anaemia is associated from the outset with neutropenia, or thrombocytopenia, or both. Venous thrombosis may be the first clinical manifestation in other patients. Although any vein may be affected, the most common localization is intra-abdominal: indeed, recurrent attacks of severe abdominal pain defying a specific diagnosis, and eventually found to be related to thrombosis, have given to PNH the attribute of being a great impostor. On the other hand, when thrombosis affects the hepatic veins it may produce acute hepatomegaly and ascites—that is to say, a fully fledged Budd–Chiari syndrome.

The natural history of PNH can extend over decades. Without treatment the median survival is estimated to be about 8 to 10 years (see Fig. 1); in the past—but unfortunately even today—the most common causes of death have been thrombosis, or infection associated with severe neutropenia, or haemorrhage associated with severe thrombocytopenia. PNH may evolve into AAA, and PNH may manifest itself in patients who previously had AAA. Rarely (estimated 1–2 per cent of all cases), PNH may terminate in acute myeloid leukaemia. On the other hand, full spontaneous recovery from PNH has been also well documented.

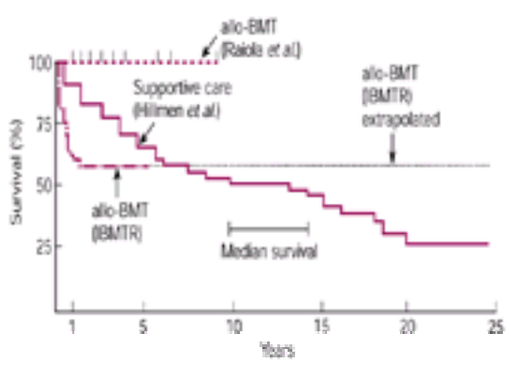

Fig. 1 PNH is a chronic disorder, the time course of which is often measured in decades. From a series of 80 patients who received only minimal supportive treatment we estimate a median survival of about 10 years. Allogeneic BMT has still been associated with significant mortality, and therefore may have reduced the survival of some patients; but recently more encouraging data have been reported on a small series from a single centre.

Laboratory investigations and diagnosis

The most consistent blood finding is anaemia, which may range from mild to moderate to very severe. The anaemia is usually normomacrocytic; if the mean cell volume (**MCV**) is high it is usually largely accounted for by reticulocytosis, which may be quite marked—up to 20 per cent. The anaemia may become microcytic if the patient is allowed to become iron-deficient as a result of chronic urinary blood loss through haemoglobinuria. The red cell morphology is otherwise usually normal. Neutropenia and/or thrombocytopenia may or may not be present from the outset, or may develop subsequently. Unconjugated bilirubin is mildly or moderately elevated; lactate dehydrogenase (**LDH**) is typically markedly elevated; haptoglobin is usually undetectable. All these findings make the diagnosis of haemolytic anaemia compelling. Haemoglobinuria may be overt in a random urine sample: if it is not, it may be helpful to obtain serial urine samples, since haemoglobinuria can vary dramatically from day to day, and even from hour to hour (it is more common, but not always, in the early morning: hence the adjective 'nocturnal'). Obviously, haemoglobinuria must be distinguished from haematuria. Surprisingly, even today a patient may undergo extensive urological investigations before it is realized that the patient has PNH. There may be free haemoglobin in the serum, and sometimes this is so high as to interfere with clinical chemistry. These findings clearly indicate intravascular haemolysis, thus increasing, by an order of magnitude, the likelihood that the haemolytic anaemia is in fact PNH (see Table 1). The bone marrow is usually cellular, with marked to massive erythroid hyperplasia, often with mild to moderate dyserythropoietic features. However, at some stage of the disease the marrow may become hypocellular or even frankly aplastic (see below).

The definitive diagnosis of PNH must be based on the demonstration that a substantial proportion of the patient's red cells have an increased susceptibility to complement, due to the deficiency on their surface of proteins that normally protect the red cells from activated complement. Classically, this is proven by the Ham test (acidified serum test): if appropriately carried out with all the necessary controls this test is still valid. By contrast, the sucrose haemolysis test can give both 'false-negatives' and 'false-positives', and therefore must be regarded as obsolete. Nowadays, the presence of a PNH red blood cell population can be easily demonstrated and quantified by flow cytometry, using anti-CD59 or anti-CD48. This analysis can also be carried out on granulocytes with a higher sensitivity (see Fig. 2).

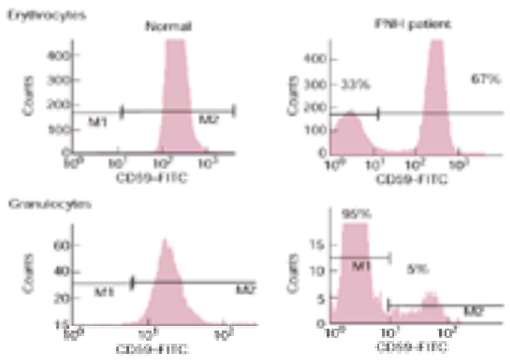

Fig. 2 Flow cytometry analysis of blood cells in a patient with PNH. On the left, red cells and granulocytes from a normal person display a unimodal distribution of surface expression of the GPI-linked protein CD59, which protects red cells against complement-mediated lysis. On the right, a similar analysis reveals, in a patient with PNH, a clearly bimodal distribution: from this analysis the size of the PNH cell population can be quantitated. (Figure by courtesy of Dr David Araten.)

Pathophysiology

Haemolysis

Haemolysis in PNH is due to an intrinsic abnormality of the red cell, which makes it exquisitely sensitive to activated complement, whether it is activated through the alternative pathway or through an antigen–antibody reaction. The former mechanism is probably the reason why there is chronic intravascular haemolysis in PNH. The latter mechanism explains why the haemolysis can be dramatically exacerbated in the course of a viral or bacterial infection. Hypersusceptibility to complement is due to the deficiency of several protective membrane proteins, of which CD59 is the most important, because it hinders the insertion into the membrane of C9 polymers.

The molecular basis for the deficiency of these proteins has been pinpointed not to a defect in any of the respective genes, but rather to the shortage of a unique glycolipid molecule, glycosyl phosphatidyl inositol (**GPI**), which, through a peptide bond, anchors these proteins to the surface membrane of cells. The shortage of GPI is due in turn to a mutation in an X-linked gene, called *PIG-A*, required for an early step in GPI biosynthesis. In virtually each patient the *PIG-A* mutation is different. This is not surprising, since these mutations are not inherited: rather, each one takes place *de novo* in a haemopoietic stem cell (in other words, they are somatic mutations). As a result, the patient's bone marrow is a mosaic of mutant and non-mutant cells, and the peripheral blood always contains both PNH cells and normal (non-PNH) cells (see Fig. 2).

Thrombosis

This is one of the most immediately life-threatening complications of PNH, and yet one of the least understood pathogenetically. It could be due to impaired fibrinolysis, because the urokinase plasminogen activator receptor (**uPAR**) is a GPI-linked protein; alternatively, complement activation could cause hypercoagulability, or hyperactivity of platelets, or both. For instance, it could be speculated that deficiency of CD59 on the PNH platelet could lead to abnormal insertion of the C5b–9 complex in the platelet membrane, as is the case with the red cell.

Bone marrow failure and the relationship between PNH and AAA

PNH has an intimate link with AAA, for several reasons. (1) As stated above, sometimes a patient with PNH becomes 'less haemolytic' and 'more pancytopenic' and ultimately evolves to frank AAA. (2) In terms of pathogenesis, it is believed that AAA is essentially an organ-specific autoimmune disease mediated by 'activated' cytotoxic (CD8+) T lymphocytes, which are able to inhibit haemopoietic stem cells. Recently, skewing of the T-cell repertoire, indicating the presence of abnormally expanded T-cell clones, has also been observed in cases of PNH. (3) Most important, intensive immunosuppressive treatment is the standard of care in those with AAA, and a beneficial response to the same treatment can also be obtained in patients with PNH (see below).

In view of these facts, it seems that an element of bone marrow failure in PNH is the rule rather than the exception: an extreme view is that PNH is a form of AAA, in which bone marrow failure is masked by the enormous expansion of the PNH clone that populates the patient's bone marrow. In other words, it appears that two different mechanisms co-operate in producing PNH (see Fig. 3): autoimmune damage to stem cells, and a somatic mutation in the *PIG-A* gene. This notion is supported by two further lines of evidence. (1) By targeted inactivation of the *pig-A* gene in mouse embryonic stem cells one can produce mice with a PNH cell population. However, this population does not grow further, as it does in patients with PNH. (2) By using refined flow cytometry technology, PNH cells harbouring *PIG-A* mutations can be demonstrated in normal people at a frequency in the order of 10 per million. Both these findings indicate that some other factor is required, in addition to a somatic mutation in the *PIG-A* gene, in order to cause PNH. Most likely, the same cytotoxic damage to stem cells that would otherwise cause AAA spares the PNH stem cells, thus allowing the PNH clone to grow to the size when it gives clinical PNH. The mechanism whereby the PNH cells escape damage is not yet known.

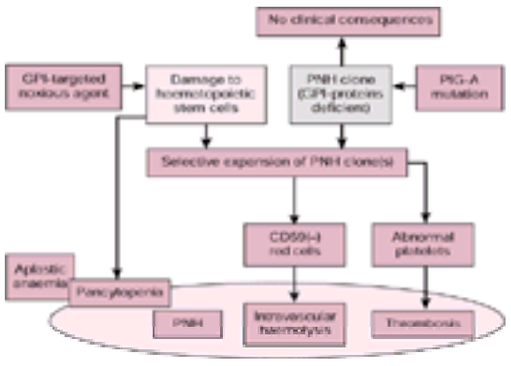

Fig. 3 The role of somatic mutation and bone marrow failure in causing PNH. This cartoon aims to emphasize that two separate factors are required to bring about PNH as a clinical disease. On the one hand, a *PIG-A* mutation on its own will produce a PNH clone, but there will be no basis for it to expand; on the other hand, damage to haemopoietic stem cells (HSC) can cause aplastic anaemia without PNH. When both factors co-operate, and if the damage to HSC is GPI-mediated, then there will be selective expansion of the PNH clone.

Complications

Given the chronic course and the complex nature of PNH, many events can cause concern and sometimes threaten life. The most important complication is certainly thrombosis, which is nearly always venous, and mostly affects the abdominal veins (see Fig. 4). The Budd–Chiari syndrome has already been mentioned: because of its characteristic clinical picture it is usually easy to recognize. However, in PNH it is sometimes associated with portal vein thrombosis, and this may limit the extent of liver enlargement. Thrombosis of the splenic vein should be suspected whenever a patient with PNH has, or develops, splenomegaly. Thrombosis of one of the mesenteric veins is much more difficult to diagnose clinically. Appropriate investigations include Doppler ultrasound, contrast-enhanced computer tomography (**CT**), and magnetic resonance imaging (**MR**I): in our experience, the most sensitive methodology is MR venography. Another life-threatening site of thrombosis is in the cranial veins, particularly the sagittal and transverse sinuses. Assessment of the location and extent of these complications is of great practical importance, because thrombolytic therapy with tissue plasminogen activator (Fig. 4) has been carried out successfully even after 3 weeks from the onset of signs and symptoms.

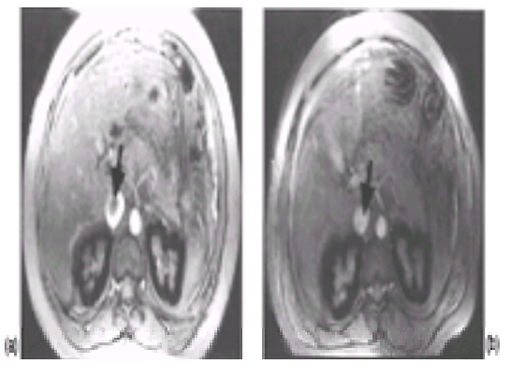

Fig. 4 Abdominal vein thrombosis in PNH can resolve with thrombolytic therapy. (a) shows extensive thrombus in the inferior vena cava in a patient with known PNH who had developed Budd–Chiari syndrome a few days earlier: it is not infrequent in PNH for thrombosis to involve multiple veins in the abdomen all at once. (b) shows a thrombus-free vena cava 2 days after an intravenous infusion of tissue plasminogen activator.(Figure by courtesy of Dr Raymond Thertulien.)

Treatment (see Fig. 5)

In the management of patients with PNH it is important to keep in mind two cardinal points: (1) unlike other acquired haemolytic anaemias, PNH may be lifelong; and (2) in view of the unique pathophysiological features reviewed in the section above, we may have to deal with any or all of three components: haemolysis, thrombosis, bone marrow failure. At the moment, the only form of treatment that can provide a cure for PNH is allogeneic bone marrow transplantation (**BMT**): when an HLA-identical sibling is available, BMT should be offered to any young patient with PNH, especially if there is severe pancytopenia. Results similar to those for AAA can be expected, with long-term disease-free survival ranging from 60 to 100 per cent in the few series that have been published (see Fig. 1: by contrast, the past record of BMT from unrelated donors in PNH is poor). The majority of patients will not have a potential sibling donor, and some of those who do may not wish to undergo BMT. Given the common pathogenesis of PNH and AAA, a logical alternative is immunosuppressive treatment with antilymphocyte globulin (or antithymocyte globulin) and ciclosporin A. Although no formal trial has ever been conducted, this approach has particularly helped to relieve severe thrombocytopenia and/or neutropenia in patients in whom these were the main problem(s): by contrast, there is often little beneficial effect on the haemolysis itself. For all patients, supportive management supervised by somebody who has previous experience of PNH can help the patient to 'live with PNH' for years, sometimes for decades, and sometimes with a good quality of life. The mainstay of support is the transfusion of filtered red cells whenever necessary. Folic acid supplements (at least 3 mg/day) are mandatory; the serum iron concentration should be checked periodically and iron supplements added as indicated. Prednisone (often administered at a dose of 15–30 mg on alternate days) is still quite popular; however, there is no evidence that prednisone decreases the rate of haemolysis, and long-term administration of prednisone, even at a low dosage, is **contraindicated**, in view of the serious potential side-effects. By contrast, a short course of prednisone may sometimes appear helpful during the course of an episode of massive haemoglobinuria associated with intercurrent infection. Any patient who has had a deep vein thrombosis at any one site in the abdomen or in a limb should be given anticoagulant prophylaxis.

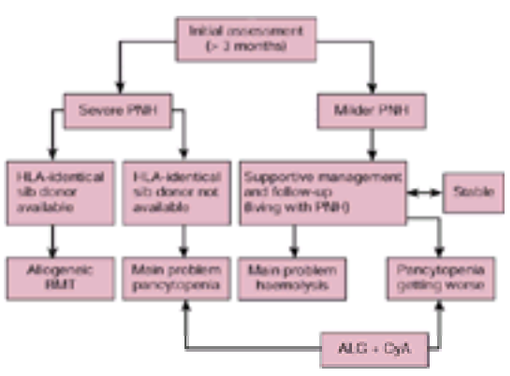

Fig. 5 An algorithm for the management of PNH. This algorithm is based on the consideration that patients with this condition vary considerably (1) in terms of clinical severity, and (2) in terms of the contributions of the PNH clone and of bone marrow failure, respectively, to determining the overall clinical picture. Some patients have been cured by bone marrow transplantation (BMT); other patients who for a long time have been 'living with PNH' have eventually experienced spontaneous recovery (see Hillmen *et al*., 1995).

Further reading

Araten D, *et al.* (1999). Clonal populations of hematopoietic cells with paroxysmal nocturnal hemoglobinuria genotype and phenotype are present in normal individuals. *Proceedings of the National Academy of Sciences USA* **96**, 5209–14.

Dacie JV (1999). *The haemolytic anaemias*, 3rd edn, Vol. 5. Churchill-Livingstone, London.

Hillmen P, *et al.* (1993). Specific defect in *N*-acetylglucosamine incorporation in the biosynthesis of the glycosylphosphatidylinositol anchor in cloned cell lines from patients with paroxysmal nocturnal hemoglobinuria. *Proceedings of the National Academy of Sciences USA* **90**, 5272–6.

Hillmen P, *et al.* (1995). Natural history of paroxysmal nocturnal hemoglobinuria. *New England Journal of Medicine* **333**, 1253–8.

Karadimitris A, *et al*. (2000). Abnormal T-cell repertoire is consistent with immune process underlying the pathogenesis of paroxysmal nocturnal hemoglobinuria. *Blood* **96**, 2613–20.

Luzzatto L (1999). Paroxysmal murine hemoglobinuria (?): a model for human PNH. *Blood* **94**, 2941–4.

Luzzatto L, Bessler M, Rotoli B (1997). Somatic mutations in paroxysmal nocturnal hemoglobinuria: a blessing in disguise? *Cell* **88**, 1–4.

Oni SB, Osunkoya BO, Luzzatto L (1970). Paroxysmal nocturnal hemoglobinuria: evidence for monoclonal origin of abnormal red cells. *Blood* **36**, 145–52.

Raiola AM, *et al*. (2000). Bone marrow transplantation for paroxysmal nocturnal hemoglobinuria. *Haematologica* **85**, 59–62. [See comments]

Rosse W (1995). Paroxysmal nocturnal hemoglobinuria. In: Handin RI LS, Stossel TP, eds. *Blood—principles and practice of hematology*, pp 367–76. Lippincott, Philadelphia.

Rosse WF (1997). Paroxysmal nocturnal hemoglobinuria as a molecular disease. *Medicine (Baltimore)* **76**, 63–93.

Rotoli B, Luzzatto L (1989). Paroxysmal nocturnal hemoglobinuria. *Seminars in Haematology* **26**, 201–7.

Saso R, *et al*. (1999). Bone marrow transplants for paroxysmal nocturnal haemoglobinuria. *British Journal of Haematology* **104**, 392–6.

Takeda J, *et al.* (1993). Deficiency of the GPI anchor caused by a somatic mutation of the PIG-A gene in paroxysmal nocturnal hemoglobinuria. *Cell* **73**, 703–11.

Young NS, Moss J, eds (2000). *Paroxysmal nocturnal hemoglobinuria and the GPI-linked proteins*. Academic Press, New York.

22.4.1 Leucocytes in health and disease

Joseph Sinning and Nancy Berliner

Introduction

Leucocytes perform a critical role in the host defence against pathogens. They mediate inflammation and modulate the immune response. Leucocytes can be divided into granulocytes (neutrophils, eosinophils, and basophils) (Plate 1), monocytes, and lymphocytes. This chapter will focus on the role of granulocytes and monocytes in the normal host response and pathological manifestations of abnormalities of their number and/or function. Lymphocytes are discussed elsewhere.

Neutrophils

Morphology

Under normal conditions neutrophils make up over half of the leucocytes in the peripheral blood. The morphological hallmarks of these cells include heterogeneous granules and a multilobated or segmented nucleus. The two predominant types of granules in the neutrophil's cytoplasm are the azurophilic (or primary) granules and the specific (or secondary) granules. Azurophilic granules arise at the promyelocytic stage of differentiation. They contain myeloperoxidase, proteases, acid hydrolases, and microbicidal proteins. Specific granules and their content proteins are synthesized at the myelocytic stage of differentiation. Their contents include lactoferrin, lysozyme, vitamin B_{12} -binding protein, gelatinase, and neutrophil collagenase. The specific granules are not a uniform population, and vary by their content with the time of their formation. Those formed early in the myelocyte stage contain abundant lactoferrin, while those formed later are enriched for gelatinase, and are often referred to as 'tertiary' granules or gelatinase granules. The specific granule membrane contains the cytochrome b-558 component of the respiratory burst oxidase, as well as chemotactic and opsonic receptors, which are transferred to the plasma membrane upon activation of the neutrophil. Finally, the neutrophil cytoplasm also contains secretory vesicles that are endocytic vesicles containing primarily plasma proteins, and are the most rapidly mobilized fraction of cytoplasmic granules in the neutrophil. The membrane of secretory vesicles is rich in receptors and cytochrome b, and the vesicles contribute these proteins to the plasma membrane upon neutrophil activation.

Common variants of neutrophil morphology include the Pelger–Huet anomaly, hypersegmentation of the nucleus Dohle bodies, and toxic granulations. The Pelger–Huet anomaly is a dominantly inherited defect in nuclear segmentation that results in a dumb-bell- or rod-shaped nucleus. Neutrophils with nuclei similar to this ('pseudo-Pelger–Huet anomaly') may be seen in acquired myelodysplastic syndromes. Hypersegmented nuclei (containing five or more segments) are characteristic of megaloblastic haematopoiesis due to folic acid or vitamin B_{12} deficiency. Dohle bodies are large basophilic inclusions that may be seen in sepsis, pregnancy, and following cytotoxic chemotherapy. Toxic granulations are abnormally staining primary granules that arise when neutrophils are released prematurely from the marrow, as in severe bacterial infections.

Maturation

There are three cellular compartments that contain myeloid cells: the marrow, the intravascular compartment, and the extravascular space. Maturation from the haematopoietic stem cell occurs in the bone marrow and takes from 10 to 14 days. The marrow compartment can be subdivided into the mitotic compartment and the post-mitotic and storage compartment. In the marrow mitotic compartment neutrophils arise through serial division of myeloid precursors. The mitotic compartment contains myeloid cells with the ability to replicate: myeloblasts, promyelocytes, and myelocytes. The marrow post-mitotic and storage compartment contains myeloid elements that have lost the ability to divide, including metamyelocytes, bands, and segmented neutrophils. Neutrophils are released from the storage pool into the intravascular space, where they remain for 4 to 12 h. Within this space approximately half of the neutrophils circulate freely in the peripheral blood while half remain 'marginated' along the vascular endothelium. The marginated and circulating cells are in dynamic equilibrium with one another. Neutrophils then migrate through the vascular endothelium into the extravascular space, where they survive for 1 to 3 days. At any given time approximately 90 per cent of neutrophils are in the marrow compartment and 2 to 3 per cent are in the intravascular space, with the remainder in the extravascular space.

Neutrophilia

Neutrophilia is defined as an elevation of the circulating neutrophil count (greater than 7.5×10^{6}/μl). Although it may reflect a primary haematological process, it usually occurs as a secondary manifestation of an underlying disease process or drug. The causes of an elevated neutrophil count are summarized in Table 1.

Hereditary neutrophilias

Hereditary neutrophilia

This is a dominantly inherited syndrome manifested by leucocytosis, splenomegaly, and widened diploë of the skull. Laboratory evaluation reveals a white blood count of 20 000 to 70 000/μl with a neutrophilic predominance, and an elevated leucocyte alkaline phosphatase. Its clinical course is benign.

Chronic idiopathic neutrophilia

This is a sporadically occurring condition manifest as a white blood count of 11 000 to 40 000/μl with a neutrophilic predominance. Patients are otherwise well and have been followed for up to 20 years without the development of significant pathology.

Leucocyte adhesion deficiency

This is a rare inherited disorder characterized by recurrent life-threatening bacterial and fungal infections, cutaneous abscesses, gingivitis, or periodontal infections. Expression of the CD11b/CD18 integrin is deficient, resulting in the inability of neutrophils to migrate to sites of infection (see below under disorders of neutrophil function for further discussion).

Acquired neutrophilias

Infection

The most common cause of an elevated leucocyte count is infection. Acute infection often causes a modest rise in the white blood count, which may be accompanied

by an increase in circulating immature precursors ('left shift'). This occurs more commonly with bacterial infection but can also occur with viral processes. Along with a left shift, morphological changes in the neutrophil may be seen with bacterial infection, including toxic granulation, Dohle bodies, and cytoplasmic vacuoles. Neutrophilia resolves with treatment or resolution of the infectious process. In chronic inflammation, marrow granulocyte production is stimulated, resulting in moderate neutrophilia, sometimes with monocytosis. Chronic infections such as osteomyelitis, empyema, and tuberculosis can also give rise to a leukaemoid reaction with white blood counts markedly elevated (greater than 50 000/μl), usually associated with a marked left shift.

Drugs

Drugs can cause leucocytosis by several different mechanisms. Steroids increase the release of mature neutrophils from the marrow and should not cause a left shift. b-Agonists acutely raise the neutrophil count by inducing the demargination of neutrophils adherent to the vascular endothelium, and may result in a neutrophil count twice that of baseline. Acute stress also results in demargination of neutrophils, which is probably mediated by adrenergic stimulation. Stresses that can cause this include exercise, surgery, seizure, and myocardial infarction. The cytokines granulocyte colony-stimulating factor (**G-CSF**) and granulocyte–macrophage colony-stimulating factor (**GM-CSF**) stimulate marrow production of neutrophils and can cause dramatic elevations in the white blood count. The majority of white cells formed are neutrophils and a left shift is often seen. The use of these cytokines therefore requires careful monitoring.

Primary haematological conditions

In other situations, neutrophilia may reflect a primary haematological condition. Marrow hyperstimulation in the setting of autoimmune haemolytic anaemia, immune thrombocytopenia, or recovery following chemotherapy or toxic insult to the marrow may result in a reactive leucocytosis. In autoimmune haemolytic anaemia and immune thrombocytopenia, neutrophilia may reflect disease activity, but steroid therapy or splenectomy may contribute. Splenectomy or hyposplenic states (for instance sickle-cell disease) may also result in modest neutrophilia at baseline with more marked neutrophilia at times of stress or infection, reflective of the loss of the spleen as a site of margination and sequestration of leucocytes.

Myeloproliferative disorders

Neutrophilia is a common feature of the myeloproliferative disorders chronic myelogenous leukaemia, polycythaemia vera, and agnogenic myeloid metaplasia, as well as familial myeloproliferative disorders. Elevated eosinophil and basophil counts are also often seen in these disorders. Leucocyte alkaline phosphatase may be low or undetectable in chronic myelogenous leukaemia. The myeloproliferative disorders are discussed in further detail elsewhere.

Non-haematological malignancies

Various non-haematological malignancies including lung and breast tumours may also cause neutrophilia. Tumours may secrete colony-stimulating factors or may cause a leukaemoid reaction. Tumour metastatic to the bone marrow may cause leucoerythroblastic changes, characterized by fragmented erythrocytes, teardrops, and nucleated red cells (myelophthysic changes), as well as leucocytosis with a left shift.

Evaluation of neutrophilia

The evaluation of neutrophilia should take account of the fact that leucocytosis is usually reactive, and that primary haematological aetiologies are relatively rare. The abnormal laboratory value should be verified to rule out laboratory error or a transient unexplained leucocytosis that resolves spontaneously. A careful history and physical examination are essential to evaluate for potential infectious processes, and to obtain a history of medication use. Examination of the bone marrow is usually not necessary for the evaluation of neutrophilia, but examination of a peripheral smear may be very helpful. Evidence of leucoerythroblastic changes warrants examination of the bone marrow to rule out granulomatous disease or tumour infiltration of the marrow. If a bone marrow biopsy is performed, evaluation should include culture of the marrow for fungus or mycobacteria.

Features that raise the question of myeloproliferative disease include concomitant elevation of platelets and haematocrit, basophilia and/or eosinophilia, and splenomegaly. In that setting, evaluation should include stem cell culture of the peripheral blood or bone marrow to assay for cytokine-independent colony growth. Evaluation for myeloproliferative disease is discussed in detail elsewhere.

Neutropenia

Neutropenia is defined as an absolute neutrophil count (**ANC**) of less than 1.5×10^6/μl. In some populations, such as Africans and Yemenite Jews, normal absolute neutrophil counts are lower, with a lower limit of normal of 1.2×10^6/μl. Neutropenia may pose a risk of serious bacterial infection, and this risk is directly related to the degree of neutropenia. In mild neutropenia (ANC 1000 to 1500 × 10^6/μl) the risk of life-threatening infection is not increased, and in moderate neutropenia (ANC 500 to 1000 × 10^6/μl) the risk of severe infection is only mildly elevated. Severe neutropenia (ANC < 500 × 10^6/μl) markedly increases the risk of life-threatening infection. The duration and acuity of neutropenia may also be important, as the acute onset of severe neutropenia is associated with a higher risk of serious infection than is chronic neutropenia of similar severity. Neutropenia in the setting of marrow failure is more threatening than neutropenia with an intact marrow, as the marrow reserve pool may afford protection. Fever of new onset in the setting of severe neutropenia is a medical emergency requiring immediate evaluation and treatment. Common causes of infection in these patients include Gram-negative enteric pathogens such as *Escherichia coli*, *Pseudomonas* spp., and *Klebsiella pneumoniae*, as well as *Staphylococcus aureus.* The causes of neutropenia are summarized in Table 2.

Congenital neutropenia

Congenital agranulocytosis (Kostmann's syndrome)

This is characterized by severe persistent neutropenia, and the early onset of frequent, life-threatening infections. Bone marrow aspirate reveals a maturation arrest at the promyelocyte stage. This syndrome was originally described as an autosomal recessive disorder, but recent evidence suggests that most cases are autosomal dominant or sporadic. These patients respond to G-CSF with increases in their absolute neutrophil count and decreased incidence of infection. Haematopoietic cell transplantation is another viable treatment option.

With the prolongation of life offered by G-CSF therapy, it has become apparent that patients with Kostmann's syndrome have an increased incidence of acute myeloblastic leukaemia (AML) and myelodysplastic syndrome (MDS). These malignancies develop in association with an acquired mutation in the G-CSF receptor. A relationship has been speculated to exist between G-CSF therapy and the development of these mutations in the G-CSF receptor, but this connection remains unproven, as has the pathogenetic role of the mutations in the subsequent development of acute myeloblastic leukaemia (AML) and myelodysplastic syndrome (MDS). Recent studies have established that Kostmann's syndrome is linked to mutations in the gene encoding neutrophil elastase, a neutrophil primary granule protein. How mutations in the elastase gene give rise to agranulocytosis remains to be elucidated.

Cyclic neutropenia (cyclic haematopoiesis)

This is a rare, dominantly inherited, marrow disorder characterized by cyclic fluctuations in neutrophil counts approximately every 21 days and lasting 3 to 7 days. Along with the neutropenia, cyclic drops in the reticulocyte and monocyte counts are also observed. This suggests that the entire pattern of haematopoiesis is cyclic in these patients, although because of the short half-life of the neutrophil, only neutropenia is clinically significant. Episodes of neutropenia may be severe, often with an absolute neutrophil count less than 200 × 10^6/μl, and may be accompanied by fevers, pharyngitis, stomatitis, and other bacterial infections. Cyclic neutropenia has also been linked to mutations in the neutrophil elastase gene, although why some mutations give rise to cyclic haematopoiesis and others to agranulocytosis is still a matter of speculation. Cyclic neutropenia can be treated safely and effectively with G-CSF. Unlike Kostmann' s syndrome, cyclic haematopoiesis is not associated with an increased incidence of AML and MDS.

Acquired neutropenias

Postinfectious neutropenia

This is commonly seen following viral infections. It usually occurs several days after the onset of infection and may last several weeks. Varicella zoster, measles, Epstein–Barr, cytomegalovirus, influenza A and B, and hepatitis A and B are some of the viruses most commonly associated with postinfectious neutropenia. The neutropenia resolves spontaneously. Transient neutropenia may also be seen with parvovirus infection. Neutropenia occurs commonly in patients with HIV. The causes are multifactorial and may be related directly to the viral infection, to opportunistic infections or associated conditions, or to the treatment of the virus or its complications.

Several bacterial infections can cause neutropenia, including rickettsial infections, typhoid fever, brucellosis, and tularaemia. Bacterial sepsis of any cause can result in acute neutropenia. This occurs both as a result of marrow suppression and increased destruction of neutrophils. Acute severe neutropenia in bacterial infections suggest that egress to tissue exceeds the capacity of the marrow reserve pool. The neutropenia may be severe and it portends a poor prognosis. Fungal infections, such as disseminated histoplasmosis, and mycobacterial diseases may also cause neutropenia.

Nutritional deficiencies

Nutritional deficiencies of vitamins B_{12} and folic acid result in megaloblastic haematopoiesis with ineffective myelopoiesis. Deficiency of copper is a rare nutritional cause of neutropenia seen in the setting of severe malnutrition or long-term parenteral alimentation. Mild neutropenia may also be seen with anorexia nervosa.

Drugs and toxins

Numerous drugs and toxins are known to cause neutropenia. Mechanisms of drug-induced neutropenia include: (i) direct marrow suppression, (ii) immune destruction with antibody- or complement-mediated damage of myeloid precursors, and (iii) peripheral destruction of neutrophils. In most cases direct marrow suppression is dose dependent. Common offending drugs that cause dose-dependent neutropenia include cancer chemotherapeutic agents, phenothiazines, anticonvulsants, and ganciclovir. Alcohol can also cause neutropenia by marrow suppression. If a drug is suspected of causing dose-dependent neutropenia, it is best to stop the suspected offending agent when possible. However, if it is not possible to stop the drug and the neutropenia is not severe, the drug may be continued with careful monitoring. Neutropenia is often related to the dose and duration of therapy. In contrast, those drugs that cause immune neutropenia usually cause profound agranulocytosis, resulting from both intramedullary destruction of myeloid precursors and peripheral destruction of mature neutrophils. Such drugs include antithyroid medications, sulphonamides, and semisynthetic penicillins. Examination of the bone marrow shows a maturation arrest of the myeloid lineage, reflecting immune destruction of myeloid precursors. The offending agent must be stopped. Recovery of the neutrophil count can be accelerated by the administration of G-CSF.

Autoimmune neutropenia

This may occur in association with collagen vascular disorders such as systemic lupus erythematosis and rheumatoid arthritis, as well as with immune thrombocytopenia and autoimmune haemolytic anaemia. Destruction may be mediated by IgG or IgM antibodies. The neutropenia may be severe but the degree of neutropenia frequently does not correlate as well with the risk of infection as in other conditions. The marrow typically is hypercellular with a late myeloid maturation arrest. Treatment is indicated in the setting of severe, recurrent infections.

Treatment options include intravenous immunoglobulin, splenectomy, and other therapies directed at the underlying collagen vascular disorder. In Felty's syndrome, neutropenia accompanies rheumatoid arthritis and splenomegaly and neutropenia probably reflects both immune destruction and splenic sequestration. Granulopoiesis is inhibited by either antibodies or T cells. This can lead to severe and recurrent infections. It may be managed with G-CSF. Splenectomy relieves the neutropenia in the majority of cases.

Large granular lymphocytosis

This may cause profound neutropenia accompanied by severe infections. It occurs in an older population, and is frequently seen in association with rheumatological diseases such as rheumatoid arthritis. Because of the association with systemic inflammatory disease, large granular lymphocytosis was originally hypothesized to be a polyclonal abnormal immune response. However, gene rearrangement studies have confirmed that large granular lymphocytosis is frequently a clonal disease representing a form of T-cell lymphoma. There are two distinct subtypes, with cells expressing either an unusual Tg phenotype (CD3+,CD8+, CD56–) or an natural killer phenotype (CD56+). When seen in association with rheumatoid arthritis, the disease may be confused with Felty's syndrome. Neutropenia related to large granular lymphocytosis is associated with a myeloid maturation arrest in the marrow, consistent with immune-mediated neturophil destruction. Surprisingly, however, the neutrophil count will often respond to G-CSF. The course of lymphoma in large granular lymphocytosis varies from indolent to rapidly progressive.

Other causes

Aplastic anaemia reflects a primary failure of haematopoiesis with neutropenia, anaemia, and thrombocytopenia. In the myelodysplastic syndromes and acute leukaemias the marrow does not produce adequate numbers of neutrophils.

Isoimmune neutropenia occurs in 1 in 500 babies born alive. It is caused by placental transfer of maternal IgG directed against fetal neutrophils, and it presents in the first days of life.

Hypersplenism usually causes mild or moderate neutropenia along with anaemia and thrombocytopenia. Normal myeloid maturation is seen in the marrow. The neutropenia is rarely severe.

Evaluation of neutropenia

In contrast to the evaluation of neutrophilia, most patients with confirmed neutropenia require bone marrow examination. A comprehensive history and physical examination may identify the occasional patient with mild neutropenia and no other evidence of disease that may warrant close observation only. However, recurrent infections, including oral and mucosal infections, abnormalities observed in a peripheral blood smear, or severe neutropenia increase the likelihood of significant marrow pathology and marrow aspiration and biopsy is indicated. If neutropenia is accompanied by anaemia or thrombocytopenia, marrow examination is required to rule out aplasia, leukaemia, myelodysplasia, or other primary marrow malignancy. A marrow that shows hyperplastic myeloid precursors and a maturation arrest supports a diagnosis of peripheral neutrophil destruction and/or immune neutropenia, which should lead to a search for an underlying collagen vascular disorder or drug-induced neutropenia.

Management of neutropenia

Fever of new onset in the setting of severe neutropenia (ANC < 500×10^6/µl) is a medical emergency. A careful history and physical examination should be performed in a timely fashion. Because of the lack of neutrophils, sites of infection may be difficult to find as significant inflammation or tissue infiltration by neutrophils may not occur. Blood and bodily fluids should be cultured. Empirical broad-spectrum antibiotics should be initiated without delay. In patients with fever in the setting of neutropenia that is expected to resolve (usually neutropenia induced by chemotherapy or drug reaction), antibiotics should be continued until the neutrophil count recovers to over 500/µl. In patients with chronic neutropenia that is expected to persist indefinitely, antibiotics should be continued for several days past the resolution of fever. If fever persists for more than 1 week despite antibiotic therapy, empirical antifungal therapy should be given. Granulocyte transfusion should be considered in culture-positive Gram-negative sepsis not responsive to antibiotics in the setting of continued neutropenia.

Granulocyte colony-stimulating factor (G-CSF)

G-CSF (Filgrastim) is a haematopoietic growth factor that has effects primarily on the neutrophilic myeloid lineage. G-CSF reduces the time of maturation of committed neutrophil precursors, prolongs the lifespan of mature neutrophils, and primes them for enhanced function of the respiratory burst, phagocytosis, and chemotaxis. Clinically, G-CSF is used in the treatment and prevention of neutropenia. When used in conjunction with myelosuppressive chemotherapy, G-CSF has been shown to reduce the severity of neutropenia, shorten the duration of neutropenia, reduce the risk of developing neutropenic fever, and reduce the length of stay in hospital. G-CSF has also been utilized successfully in the treatment of severe neutropenia secondary to congenital disorders such as cyclic neutropenia and Kostmann's syndrome, and may be useful in the treatment of autoimmune neutropenia as seen in Felty's syndrome and systemic lupus erythematosis. The

neutropenia of marrow failure states, such as the myelodysplastic syndromes, may respond to G-CSF.

Neutropenia secondary to the treatment of HIV infection can also be controlled with G-CSF. The other major use of G-CSF is in the mobilization of haematopoietic progenitor cells from the bone marrow to the peripheral blood. While in the peripheral blood, these cells can be collected by cytopheresis for use in haematopoietic cell transplantation.

Disorders of neutrophil function

Chronic granulomatous disease

Chronic granulomatous disease is a heterogeneous group of rare disorders characterized by defective production of superoxide (O_2^-) by neutrophils, monocytes, and eosinophils. The majority of cases are inherited in an X-linked fashion, but autosomal recessive inheritance also occurs. The genetic lesions causing chronic granulomatous disease have been characterized, and involve mutations in any of four genes encoding the proteins of the respiratory burst oxidase. These include the 91-kDa (X-linked) and 22-kDa (autosomal) components of the membrane cytochrome b-558 complex, and the 47- and 67-kDa soluble components (autosomal) of the oxidase complex. Patients usually present in childhood with severe infections, often with catalase-negative pathogens. The most common infection in patients with chronic granulomatous disease is pneumonia, with *Staphylococcus aureus*, *Burkholderia cepacia*, *Aspergillus* spp., and enteric Gram-negative bacteria often implicated. Other common infections in chronic granulomatous disease include lymphadenitis, cutaneous infections, hepatic abscesses, and osteomyelitis. Aphthous ulceration of the oral mucosa is common, as are chronic mucosal inflammation, perirectal abscesses or fissures, and granulomas of the gastrointestinal and genitourinary tract. The diagnosis of chronic granulomatous disease should be considered in an individual with a history of multiple severe bacterial and fungal infections or a family history of the disorder. The diagnosis is established by confirming abnormal neutrophil oxidative metabolism with tests such as the nitroblue tetrazolium (NBT) slide test or measurements of superoxide or peroxide production. The management of chronic granulomatous disease is based on aggressive prophylaxis and prompt treatment of infection. Prophylactic trimethoprim–sulphamethoxazole or dicloxacillin can significantly decrease the number of bacterial infections in patients with chronic granulomatous disease. Potentially serious infections require the prompt initiation of parenteral antibiotics. Surgical interventions including drainage of abscesses and resection of infected tissue are in important adjunct to antimicrobial chemotherapy. Prophylaxis with recombinant human interferon-g has been shown in a phase III trial to decrease substantially the number of serious infections in patients with chronic granulomatous disease.

Leucocyte adhesion deficiency

Leucocyte adhesion deficiency is an inherited disorder of neutrophil function. Two types of leucocyte adhesion deficiency have been characterized. Type 1 deficiency is a rare autosomal recessive disorder resulting from mutations in CD18, the gene encoding for the b-chain of leucocyte function antigen-1 (LFA-1, CD11a/CD18), Mac-1 (CD 11b/CD18, CR3, the receptor for the opsonin C3Bi), and gp150,95 (CD11c/CD18). Deficient expression of these three integrin complexes on the neutrophil cell surface results in decreased neutrophil adhesion to the endothelium, impaired chemotaxis, and defective C3Bi-mediated pathogen ingestion, degranulation, and respiratory burst activation. Patients with leucocyte adhesion deficiency typically present in early childhood with recurrent pyogenic infections of the skin, respiratory and digestive tracts, and mucosal membranes. A history of delayed umbilical cord separation is also often noted. Common pathogens in patients with type 1 leucocyte adhesion deficiency include *Staphylococcus aureus* and Gram-negative enterics. Foci of infection notably lack neutrophil infiltration. A mild leucocytosis persists due to impaired margination. The diagnosis is confirmed by flow cytometric measurement of neutrophil CD11b/CD18 expression. The treatment of type 1 leucocyte adhesion deficiency includes aggressive use of parenteral antibiotics for pyogenic infections. Prophylactic trimethoprim–sulphamethoxazole may benefit some patients. Patients with a severe phenotype often die in the first 2 years of life, but patients with mild disease may survive to early adulthood. Type 2 leucocyte adhesion deficiency is caused by a deficiency of Sialyl–Lewis X moieties on neutrophil selectins. In addition to neutrophil function abnormalities, this extremely rare syndrome also is characterized by mental retardation, short stature, and the rare Bombay erythrocyte phenotype.

Myeloperoxidase deficiency

Myeloperoxidase deficiency is a relatively common, autosomal recessively inherited, disorder of neutrophil function. Complete deficiency occurs in 1 in 2000 individuals and partial deficiency occurs twice as frequently. Myeloperoxidase catalyses the production of hypochlorous acid, which is an antimicrobial agent. Myeloperoxidase deficiency is often of no clinical consequence because other host defence mechanisms can adequately compensate for the defective myeloperoxidase; however, when myeloperoxidase deficiency coexists with another defect in host defence, such as diabetes mellitus, disseminated candidal or fungal infections may occur. The diagnosis of myeloperoxidase deficiency is made by histochemical staining of neutrophils and monocytes. Therapy consists of aggressive treatment of fungal infections as well as careful control of glucose levels in patients with diabetes. An acquired form of myeloperoxidase deficiency occurs in some myeloid leukaemia.

Chediak–Higashi syndrome

Chediak–Higashi syndrome is a rare disorder of neutrophil function. Neutrophils and monocytes contain giant primary granules and demonstrate impaired degranulation and fusion with phagosomes. Chemotaxis is also defective. Neutropenia results from defective granulopoiesis. Chediak–Higashi syndrome is inherited in an autosomal recessive manner. The gene responsible has been cloned, and is homologous to a murine lysosomal trafficking protein. Chediak–Higashi syndrome manifests in childhood or infancy with infections of the skin, lungs, and mucous membranes. *S. aureus*, Gram-negative enterics, *Candida*, and *Aspergillus* species are responsible for most infections in this syndrome. Non-haematological manifestations of Chediak–Higashi syndrome include partial oculocutaneous albinism, progressive peripheral and cranial neuropathies, and in some cases, mental retardation. The majority of patients will develop an accelerated phase of the syndrome, manifested by lymphohistiocytic proliferation in the liver, spleen, bone marrow, and lymphatics. The diagnosis of Chediak–Higashi syndrome is made by the demonstration of giant peroxidase-containing granules in peripheral blood or bone marrow myeloid cells, outside of the setting of myelogenous leukaemia. Chediak–Higashi syndrome is treated in the early or stable phase with prophylactic antibiotics and aggressive parenteral antibiotics for infections. Ascorbic acid may also be of benefit. The accelerated phase is treated with vinca alkaloids and glucocorticoids, but often responds poorly to these measures. Allogeneic haematopoietic cell transplantation from HLA-compatible donors is the only potentially curative therapy for Chediak–Higashi syndrome.

Specific granule deficiency

An extremely rare disorder, neutrophil specific granule deficiency is characterized by absent or empty neutrophil specific granules. Specific granule deficiency is manifested clinically as recurrent skin and pulmonary infections resulting from the absence of antimicrobial neutrophil granule proteins such as lactoferrin and defensins. An inability to upregulate the expression of integrins stored on the specific granule membrane may also be responsible for the impairment of host defence. The diagnosis of specific granule deficiency is made by microscopic examination of neutrophils. With appropriate antibiotic prophylaxis and aggressive treatment of infections, patients may live to adulthood. A truncation mutation in the transcription factor C/EBPe has recently been demonstrated to be responsible for some, but not all, cases of specific granule deficiency.

Monocytes

Monocytes are large circulating cells with a non-segmented nucleus and cytoplasmic granules. They function as phagocytes both in antimicrobial defence and in clearing cellular debris. Their granules are essentially identical to neutrophil azurophilic granules, and contain acid hydrolases and myeloperoxidase. Monocytes are also capable of producing reactive oxygen and nitrogen compounds with microbicidal activity. Monocytes play a critical role in the immune response as they present antigens in the context of MHC to T cells. They also produce a variety of immunomodulatory cytokines including interleukins 1 and 6, tumour necrosis factor-a, and b-interferon.

Monocytes arise from bone marrow stem cells. They share a common myeloid precursor with granulocytes. The differentiation to the monocyte is modulated by several cytokines, most importantly monocyte colony-stimulating factor and granulocyte–monocyte colony-stimulating factor. The majority of monocytes are marginated to the vascular endothelium. Upon stimulation, they migrate to the tissue where they develop into macrophages. In the tissue they kill bacteria, mycobacteria, fungi, and protozoa. They are especially important in defence against intracellular pathogens. Specialized resident tissue macrophages include the Langerhans cells of the skin, dendritic cells of lymph nodes, Kupffer cells of the liver, and alveolar macrophages.

Monocytosis is defined as a monocyte count of greater than $0.9 \times 10^6/\mu l$. Disorders causing monocytosis are heterogeneous. Recovery of the marrow following chemotherapy or agranulocytosis is heralded by monocytosis prior to the return of neutrophils. Monocytosis is also seen in syndromes such as cyclic neutropenia,

Kostmann's syndrome, and idiopathic neutropenia.

The most common causes of monocytosis include chronic infection, inflammation, or tumour, as well as some primary haematological disorders (Table 3). Chronic infections leading to monocytosis include subacute bacterial endocarditis and mycobacterial diseases. Monocytosis is typically moderate and resolves with treatment of the infection. Autoimmune processes such as systemic lupus erythematosis, rheumatoid arthritis, and vasculitis also cause moderate monocytosis. Monocytosis may arise from primary malignancies of the marrow or in the setting of marrow infiltration with solid tumours (myelophthysis).

Primary marrow disorders causing monocytosis include acute monocytic leukaemia, chronic myelogenous leukaemia and other myeloproliferative disorders, and chronic myelomonocytic leukaemia, which has features of both myelodysplastic and myeloproliferative disorders. Juvenile chronic myelogenous leukaemia is a rare disorder occurring in children less than 4 years of age. Lymphadenopathy and splenomegaly are also prominent features.

Monocytopenia in isolation is uncommon. Monocytopenia is sometimes seen following steroid administration, endotoxaemia, or in marrow failure syndromes such as aplastic anaemia.

Eosinophils

Morphology

Eosinophils have a bilobate nucleus and contain characteristic elliptical granules that stain with eosin. There are three types of eosinophil granules. Primary granules are round in shape. Secondary granules are abundant and contain crystalloid material, and account for the eosinophil's staining properties. The third type of granule is small and contains lysosomal enzymes. Granules contain high concentrations of eosinophil major basic protein, histaminase, eosinophil cationic protein, hydrolases, and peroxidase. Eosinophils are capable of phagocytic function but more commonly release their granule contents to the environment. Eosinophils are also capable of producing reactive oxygen species, and produce prostaglandins, thromboxane A_2, and leukotriene C_4. Eosinophils play a prominent role in defence against helminths and parasites. They arise in the marrow from a common myeloid precursor, and their production is dependent on GM-CSF, IL-3, and IL-5.

Congenital eosinophilia

Job's syndrome is an inherited disorder characterized by recurrent cold abscesses, eczema, and coarse facies. Patients typically have eosinophilia and significantly elevated levels of IgE.

Acquired eosinophilia

Allergic reactions

These are the most common cause of eosinophilia, including allergies to drugs and environmental agents (Table 4). Moderate eosinophilia is common in collagen vascular diseases including rheumatoid arthritis, vasculitis, and eosinophilic fasciitis. Malignancies such as Hodgkin's disease, non-Hodgkin's lymphoma, and various solid tumours may present with eosinophilia. Moderate eosinophilia is commonly seen in chronic myelogenous leukaemia.

Eosinophilic leukaemia

This is an extremely rare disorder. It presents with extreme eosinophilia, and may be difficult to differentiate from the hypereosinophilic syndrome. However, in eosinophilic leukaemia, the eosinophils are clonal and may contain clonal cytogenetic abnormalities. Furthermore, eosinophilic leukaemia is often associated with cytopenias, infections, and an increase in marrow blasts.

Idiopathic hypereosinophilic syndrome

This is characterized by prolonged eosinophilia and end-organ damage secondary to tissue infiltration by eosinophils. Organs commonly involved include the heart, lungs, central nervous syndrome, kidneys, gastrointestinal tract, and skin. Eosinophilia may be severe (> 50–100 × 10^6/μl). Eosinophils deposit toxic proteins in the infiltrated tissues which lead to thrombosis and fibrosis. In the heart the fibrosis results in restrictive cardiomyopathy. The diagnosis of an idiopathic hypereosinophilic syndrome requires eosinophil counts greater than 1.5 × 10^6/μl for 6 months, organ dysfuncton secondary to eosinophilic infiltration, and no other cause to explain the eosinophilia.

Churg–Strauss syndrome

Primary hypereosinophilic syndrome may be difficult to distinguish from Churg–Strauss syndrome, an eosinophilic vasculitis associated with eosinophilia, pulmonary infiltrates, asthma, and neuropathy. The presence of Churg–Strauss syndrome is suggested by the presence of asthma, which is not characteristic of hypereosinophilic syndrome, and the diagnosis can be confirmed by the finding of necrotizing vasculitis of small vessels, often in association with extravascular granulomas.

Eosinophilia–myalgia syndrome

This is a disorder associated with the ingestion of L-tryptophan supplements from a single source. Affected individuals have eosinphilia and severe myositis. Tissues are infiltrated with eosinophils. In many cases the syndrome responds to steroids, but in some cases it is fatal.

Eosinopenia

Eosinopenia is seen in the setting of steroid use, stress, and acute infection. It typically is clinically benign.

Basophils

Basophils are rare circulating cells, accounting for less than 0.1 per cent of white blood cells. They are non-phagocytic granulocytes. Their large heterogeneous granules account for their purple-black staining. Their granules contain histamine, heparin, tryptase, chemotactic factors for neutrophils and eosinophils, leukotrienes, prostaglandins, and platelet-activating factor. They arise in the marrow from the same myeloid precursor as eosinophils. Basophils function in immediate-type hypersensitivity. They are structurally similar to mast cells but the exact relationship between these cell types is not clear. Basophilia (> 0.2 × 10^6/μl) is seen in myeloproliferative disorders such as chronic myelogenous leukaemia and polycythaemia vera, hypersensitivity reactions, and with some viral infections including varicella and influenza. Mast cell leukaemia is a rare disorder with a poor prognosis.

Further reading

Baehner RL (2000). Normal neutrophil structure and function. In: Hoffman R *et al*., eds. *Hematology: basic principles and practice*, pp 667–86. Churchill Livingstone, Philadelphia.

Curnutte IT, Coates TD (2000). Disorders of phagocyte function and number. In: Hoffman R *et al*., eds. *Hematology: basic principles and practice*, pp 720–62. Churchill Livingstone, Philadelphia.

Dale DC *et al*. (2000). Mutations in the gene encoding neutrophil elastase in congenital and cyclic neutropenia. *Blood* **96**, 2317.

Malech HL, Gallin JI (1987). Current concepts: immunology. Neutrophils in human disease. *New England Journal of Medicine* **317**, 687.

Pizzo PA (1993). Drug therapy: management of fever in patients with cancer and treatment-induced neutropenia. *New England Journal of Medicine* **328**, 1323.

Rothberg ME (1998). Mechanisms of disease: eosinophilia. *New England Journal of Medicine* **338**, 1592.

Stock W, Hoffman R (2000). White blood cells 1: non-malignant disorders. *Lancet* **355**, 1351.

Winkelstein JA *et al*. (2000). Chronic granulomatous disease: report on a national registry of 368 patients. *Medicine (Baltimore)* **79**, 155.

22.4.2 Introduction to the lymphoproliferative disorders

Barbara A. Degar and Nancy Berliner

The human immune system has the capacity to identify and respond specifically to invading pathogens. It can also 'remember' the exposure, such that subsequent exposure to the same pathogen results in a more rapid and potent immune response. Lymphocytes play the key role in the adaptive immune response, mediating both specificity and memory.

Lymphocytes

The lymphocytes can be divided into two morphologically indistinguishable types, which play different and complementary roles in the immune system. Both are derived from lymphohaemopoietic stem cells that reside in fetal liver and in adult bone marrow. B cells develop in the marrow (the human equivalent of the avian bursa of Fabricius) and their principal role is to generate immunoglobulin (antibodies). B cells represent about 20 per cent of the lymphocyte population in peripheral blood. T cells mature within the thymus. T cells orchestrate the immune response: they are capable of cell-mediated cytotoxicity, they generate inflammatory cytokines, and they provide help for B-cell function. T cells account for approximately 80 per cent of the lymphocytes in the peripheral circulation. A much smaller population of lymphoid-appearing cells express neither B-cell nor T-cell markers. These null cells, also known as natural killer (**NK**) cells and large granular lymphocytes (**LGLs**), are capable of cell-mediated cytotoxicity, especially against tumour cells and virally infected cells. NK cells are a component of the innate immune response, as they do not demonstrate immunological memory.

Lymph nodes

In their role in infection surveillance, lymphocytes circulate through the body via a network of lymphatic and blood vessels. At strategic locations, lymphoid cells are organized to allow direct interaction among lymphocytes and other specialized cells of the immune system.

These interactions permit the production of specific, functional effector cells. The network includes approximately 500 to 600 discrete lymph nodes, lymphoid populations in the oropharynx (Waldeyer's ring), bronchial tree and gut, as well as in the thymus, the bone marrow, and the spleen.

Within lymph nodes, lymphocytes are arranged in a central medulla surrounded by an outer cortex contained within a connective tissue capsule (Fig. 1) Afferent lymphatics penetrate the cortex and lymphocyte-rich fluid filters toward the medullary sinusoids and the efferent lymphatics at the hilum of the node. The vascular supply to the lymph node includes specialized postcapillary venules that allow the passage of peripheral blood lymphocytes into the node. Lymphocytes are ultimately returned to the bloodstream via the thoracic duct.

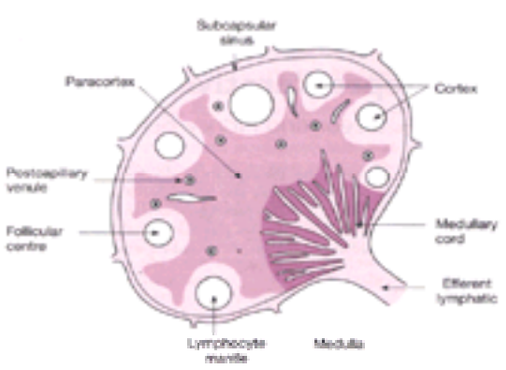

Fig. 1 Functional architecture of a normal lymph node. (Reproduced from Arno J (1980). *Atlas of lymph node pathology*, with permission.)

Roughly spherical follicles are found in the lymph node cortex and predominantly comprise B cells. Primary follicles contain clusters of naïve, unstimulated B cells. Secondary follicles, with pale 'germinal centres' surrounded by a darker 'mantle' zone, represent foci of B cells proliferating and differentiating in the presence of antigen-bearing dendritic cells and activated 'helper' T cells (T_H cells). The interfollicular and paracortical zones of the lymph node are densely populated by T cells. Macrophages, follicular dendritic cells, and interdigitating reticulum cells all process and present antigen to the lymphocytes within the node.

The design of the lymph node facilitates the process whereby the subpopulation of lymphocytes capable of responding to a specific antigen is expanded. Antigens are delivered to the subcapsular sinus of the node via afferent lymphatics, and are taken up by reticulum cells and presented on their surface in the context of the major histocompatibility complex (**MHC**) proteins. Specific T-lymphocyte responses require that peptide antigens, which are derived from 'foreign' proteins, appear on the surface of antigen-presenting cells in close association with a 'self' MHC molecule. B cells, on the other hand, are capable of responding to some antigens in solution. Optimal B-cell responses require the 'help' of T cells both via direct cell–cell contact and in response to cytokines secreted by T cells. Only those T cells and B cells that have been genetically preprogrammed to interact with a specific antigen will proliferate and differentiate in response to it.

Antigen receptors

Both B and T cells express transmembrane proteins on their cell surfaces, these proteins bind antigen and define the antigenic specificity of the cell. In the case of B cells, the immunoglobulin molecule represents the B-cell receptor (Fig. 2) Each immunoglobulin molecule is a bivalent tetramer comprising a pair of heavy chains bound to two light chains (of either kappa or lambda type). Genetic recombination of approximately 400 immunoglobulin gene segments (located on chromosomes 2, 14, and 22), generates about 10^{15} distinct antibody specificities. The expression of recombination activating genes (*RAG1* and *RAG2*) early in B-cell development mediates the random rearrangement of variable (**V**), diversity (**D**), and joining (**J**) gene segments. Terminal deoxyribonucleotidyl transferase (**TdT**) contributes to the diversity of immunoglobulin molecules by inserting additional nucleotides during the splicing of gene segments. This process gives rise to a vast repertoire of antibody molecules, each with a unique antigen-binding cleft. All of the progeny cells of a B cell that has rearranged its immunoglobulin genes have the same antigenic specificity and are referred to as a clone. Most protein antigens are complex and contain many different epitopes (structures capable of binding an antigen receptor). Therefore, most pathogens stimulate many lymphocyte clones to proliferate: that is to say, they result in polyclonal responses. As B-cell clones mature, the isotype of the antibodies they produce 'switches' from IgM/IgD to IgG, IgA, or IgE.

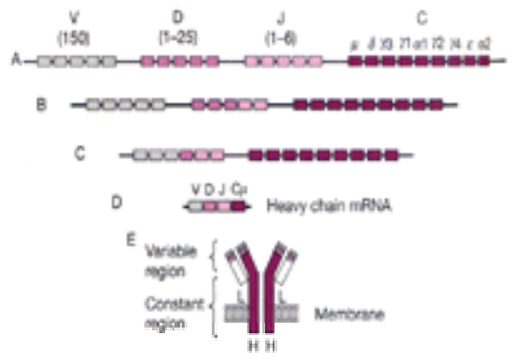

Fig. 2 Immunoglobulin gene rearrangement. The top line (A) represents the germline pattern of the immunoglobulin heavy chain locus found on human chromosome 14. B-cell progenitors express recombination activating genes which mediate the random, sequential rearrangement of gene modules (lines B and C) such that only one of several variable $(V)_1$ diversity (D), and joining (J) segments is expressed by a B-cell clone (line D). As the gene components are spliced, terminal deoxynucleotidyl transferase (TdT) randomly inserts additional nucleotides at splice junctures. Diverse antigenic specificity is thus somatically generated from a relatively small amount of genetic material. The immunoglobulin molecule (line E) is a tetramer of two heavy and two light chains which may be cell-associated (as shown) or secreted. The region of the molecule which interacts specifically with antigen is the variable region. The constant region of the light chain are of either the _ or l types. The constant region of the heavy chain determines the isotype of the antibody (IgM, IgD, IgG, IgA, IgE).

In an analogous fashion, T-cell precursors rearrange the T-cell receptor (**TCR**) genes. The TCR consists of a heterodimer of a and b chains, or g and † chains in a minority of T cells. The a and b genes are encoded on chromosomes 14 and 7, respectively, while the g and † chains are on chromosomes 7 and 14, respectively. T-cell precursors randomly assemble variable, joining, and diversity gene segments to generate a vastly diverse array of antigen-specific T-cell clones. When the T cell encounters antigen to which it can productively bind, the cell undergoes clonal expansion, and generates both activated effector cells and long-lived memory cells.

Lymphocyte ontogeny

As lymphocytes develop and mature from multipotent progenitors to terminally differentiated effector cells, they express a sequential pattern of surface proteins. Some of these cell-surface molecules subserve known, critical functions in the cells that bear them. Others are of less clear biological significance, but are useful markers of cell type and status of differentiation and activation. Malignant lymphomas and lymphoid leukaemias are frequently classified and understood on the basis of their expression of cell-surface markers (Fig. 3) In some cases, the stage of differentiation at which malignant transformation occurred can be inferred from the pattern of the surface antigens expressed by the malignant cells.

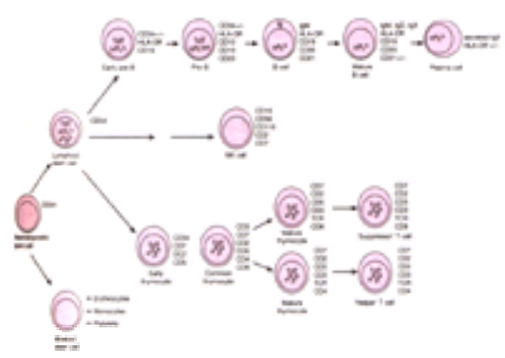

Fig. 3 Simplified depiction of lymphocyte ontogeny. Lymphocytes derive from lymphoid progenitors in the bone marrow, which in turn are derived from multipotent haemopoietic stem cells. B-lymphoid progenitors are recognized by their expression of terminal deoxynucleotidyl transferase (TdT) and the rearrangement of the immunoglobulin heavy chain locus. As B cells mature, the light chain is rearranged and immunoglobulin is expressed first within the cell cytoplasm, then on the cell surface, and is ultimately secreted. T-lymphoid progenitors migrate to the thymus where they express TdT and rearrange the b-subunit followed by the a-subunit of the T-cell receptor (TCR). An overlapping sequence of cell-surface proteins are expressed as the cells differentiate, these have been numerically classified using cluster of differentiation (**CD**) designations. The status of the immunoglobulin and *TCR* genes are represented as follows: H, immunoglobulin heavy chain; L, light chain; a, TCR-a; b, TCR-b; G, germline; R, rearranged.

Lymphocytes develop from haemopoietic stem cells. Although the surface characteristics of these elusive cells are not well understood, it is likely that human stem cells express the cell-surface glycoprotein CD34. The first recognizable sign of commitment to the B-lymphoid lineage is the expression of TdT and the rearrangement of the immunoglobulin heavy chain. As differentiation progresses, B-cell progenitors turn on the expression of class II MHC molecules (HLA-DR) as well as CD19 and then CD10 (the latter is also known as 'common acute lymphoblastic leukaemia antigen', **CALLA**). The immunoglobulin light chain is rearranged and the cells (now termed pre-B cells) express the μ heavy chain within their cytoplasm. As the cells progress to the early B-cell stage, CD34, TdT, and CD10 expression are extinguished, and CD19, CD20, and CD21, as well as IgM are expressed on the cells' surface. Mature B cells express surface IgM and/or IgD, in addition to CD19 and CD20. Plasma cells, the end result of B-cell differentiation, produce cytoplasmic as well as secreted immunoglobulin, but do not express surface immunoglobulin. They lack CD19 and CD20 expression.

Similarly, as T cells mature they progress through an orderly cascade of genetic and cell-surface events. CD34-positive progenitors that are destined for a T-lymphoid fate migrate from the marrow to the thymus and express TdT as well as CD7. Next, the cells express the CD2 molecule, which, among other things, mediates the binding of T cells to sheep erythrocytes. The T-cell receptor genes are then rearranged and subsequently expressed on the surface of the thymocyte in association with the CD3 molecule. Distinct populations of mature thymocytes emerge: those that express CD4 and function as cytokine-secreting 'helper' cells and those that express CD8 and function as cytotoxic 'killer' cells. Rare 'double-positives' (CD4+CD8+) and 'double-negatives' (CD4–CD8–) also exist. The CD4 molecule mediates the binding of T cells to MHC class II molecules, whereas CD8 binds MHC class I proteins.

The third descendant of the lymphoid stem cell, the NK cell, is characterized by its expression of CD7, CD2, CD16, and CD56, in addition to other surface proteins. NK cells are distinguished from T cells by the fact they do not express CD3 (and therefore the T-cell receptor).

Lymphoproliferative disorders

A variety of conditions that span the spectrum of benign, reactive processes to frank malignant transformation result in the expansion of lymphocyte populations. Lymphoproliferative diseases are typically manifested by lymphocytosis and/or lymphadenopathy. Distinguishing these processes clinically and pathologically is not always easy. The lymphoproliferative disorders are a loosely defined group of malignant and non-malignant entities characterized by the autonomous, poorly controlled proliferation of lymphoid cells. Malignant tumours are clonal in nature; they result from the uncontrolled proliferation of a single transformed cell. In contrast, non-malignant lymphoproliferation contains polyclonal lymphocyte populations. Lymphoproliferative disorders may result from chronic antigenic stimulation, certain viral infections, or from an imbalance among interacting lymphocyte populations, as may occur in congenital or acquired immunodeficiency syndromes. In addition, lymphocytes are prone to the acquisition of chromosomal translocations, particularly involving the immunoglobulin and T-cell receptor genes, and such changes may contribute to malignant transformation (see Table 1).

Lymphocytosis

Normal peripheral blood usually contains approximately 1000 to 5000 lymphocytes/μl, accounting for approximately 40 per cent of the circulating leucocytes. Infants and young children typically have higher absolute lymphocyte values. Increased numbers of circulating lymphocytes (lymphocytosis) and/or the appearance of abnormal (or atypical) lymphocytes in the blood are usually caused by either viral infection or lymphoid malignancy. The appearance of the circulating lymphocytes on a peripheral blood smear may provide clues to the pathogenesis of the elevated lymphocyte count. For example, infectious mononucleosis results from primary infection with the Epstein–Barr virus (**EBV**), and gives rise to large numbers of 'atypical' lymphocytes with abundant cytoplasm in the peripheral blood. Chronic lymphocytic leukaemia (**CLL**) leads to an increase in circulating normal-appearing 'mature' lymphocytes. CLL is also frequently associated with the appearance of 'smudge' cells in the peripheral smear, a preparation artefact caused by the destruction of the fragile CLL cells. Follicular lymphoma may be associated with the circulation of characteristic cells with a cleaved nucleus.

Lymphadenopathy

Enlargement of one or more lymph nodes (lymphadenopathy) is an extremely common clinical finding. With the exception of inguinal nodes, normal lymph nodes are non-palpable. Nodes that are palpable and/or exceed approximately 1 × 1 cm on imaging studies are considered pathological. Lymph node enlargement often results from the body's normal and adaptive response to an immunological challenge; however, it may signify a pathological inflammatory or malignant disease. The causes of lymphadenopathy fall into three main categories: infectious, inflammatory (reactive), and neoplastic (Table 1) Younger patients, especially children, are more likely to develop adenopathy as a result of infection, while the likelihood of haematological or metastatic malignancy increases with age.

Approach to the patient with lymphadenopathy

The history of the patient with lymphadenopathy should take into account the age and general health of the patient, the duration of the adenopathy, the coexistence of fever, weight loss, night sweats, pruritis, and cough, and any recent infections, medications, travel, and animal exposures. The physical examination should make note of the location (generalized versus regional), the texture (hard versus rubbery), and the mobility of the nodes (fixed versus mobile), and the presence or absence of signs of inflammation (warmth, tenderness, erythema). The skin and oropharynx should be examined and the size of the liver and spleen should be assessed. Additional screening studies may include a complete blood count, and measurement of the erythrocyte sedimentation rate (**ESR**). Serological studies for certain viral pathogens and for rheumatological diseases can be diagnostically helpful. Radiographs of the chest should be obtained if mediastinal adenopathy is suspected. Ultrasound may demonstrate central suppuration, which is characteristic of acute lymphadenitis. A computed tomography (**CT**) scan is required to diagnose intra-abdominal adenopathy.

Lymph node biopsy

In the absence of an obvious infection or underlying illness associated with lymphadenopathy, or when malignancy is suspected, a lymph node biopsy is recommended. Depending on the clinician's level of concern, a trial of observation with or without empirical antibiotics (usually an antistaphylococcal agent) is sometimes chosen. Empirical treatment with steroids should be avoided because it may undermine the diagnosis and proper therapy of lymphoid malignancy. If there is no resolution within 2 weeks, then a lymph node biopsy should be strongly considered. The largest accessible node is most often selected for biopsy. A fine-needle aspiration of lymph nodes is adequate for diagnosis in a restricted set of clinical circumstances: for example, diagnosis of recurrent disease or metastatic carcinoma or melanoma. Culture of a lymph node aspirate may yield a microbiological diagnosis in infective lymphadenitis. Most pathologists prefer an excisional biopsy, when possible, because nodal architecture is preserved. A portion of the sample should be reserved fresh (that is, not fixed in formalin) for flow cytometry and cytogenetic studies, if indicated.

Histopathology

Histological examination of lymph nodes is the mainstay of diagnostic studies, however non-diagnostic or non-specific inflammatory findings are frequently encountered. Reactive lymph nodes demonstrate characteristic, but by no means specific, histological patterns which involve the three functional domains of the lymph node: the follicles, the paracortex, and the medullary sinuses.

An increase in the size and/or number of lymphoid follicles (which contain proliferating B cells,) is termed 'follicular hyperplasia'. The specific cause is rarely identified. This pattern of lymph node reactivity is characteristic of rheumatological conditions and of HIV infection and Castleman's disease. Castleman's disease is a rare and poorly understood non-neoplastic cause of lymphadenopathy that occurs in localized and multicentric forms. The multicentric form is a systemic illness without defined therapy that is associated with infection with human herpesvirus-8 (HHV-8, also known as Kaposi's sarcoma herpesvirus).

Paracortical expansion accompanies T-cell proliferation and is characteristic of certain viral causes of lymphadenopathy, such as EBV infection. Paracortical expansion with granuloma formation is typical of mycobacterial infections and sarcoidosis. In Kikuchi's disease and Kawasaki's disease (mucocutaneous lymph node syndrome), paracortical necrosis is seen in involved lymph nodes.

Sinus hyperplasia is caused by an increased number of histiocytes in the medullary sinuses. This pattern of lymph node reactivity is seen in the histiocytic syndromes and in the storage diseases. A rare condition known as sinus histiocytosis with massive lymphadenopathy or Rosai–Dorfman disease is characterized by an extreme polyclonal proliferation of macrophages. This entity often involves the cervical lymph nodes, but may occur in virtually any nodal or extranodal site and is usually, but not always, self-limited.

Involvement by a malignant lymphoma leads to effacement of the lymph node structure to a greater or lesser degree. Histology correlates with clinical behaviour and will be described in subsequent sections focused on the classification of lymphoma.

Immunohistochemistry and flow cytometry

Histology alone may be inadequate to distinguish the malignant from the non-malignant lymphoproliferative disorders. Supplemental information from flow cytometry, cytogenetics, and immunoglobulin/TCR gene rearrangement studies demonstrate the clonal nature of malignant disease and provide data with prognostic and therapeutic significance. Immunohistochemistry is used to characterize the pattern of surface marker expression in fixed or frozen tissue samples. Flow cytometry is performed on cells in suspension, such as peripheral blood or bone marrow, or on cell suspensions prepared from a lymph node or other solid tumour. For flow cytometry, solid specimens should not be fixed or frozen but kept refrigerated until processing. Both techniques detect the binding of monoclonal antibodies of known specificity to the clinical sample. Using a panel of antibodies, these studies demonstrate the types of cells present in the sample. Non-haemopoietic metastatic tumours can be identified. The lineage of lymphoid malignancies can be revealed, for example B cell versus T cell versus NK cell. In the case of B-cell lymphoproliferation, the relative expression of kappa and lambda light chains can be measured. As described above, B cells express either the kappa or the lambda light chain, but not both. Predominant expression of either the kappa or lambda light chain by a population of B cells, a phenomenon known as light-chain restriction, suggests a clonal process. Using flow cytometry, lymphoid neoplasms can be placed within the hierarchy of normal lymphocyte ontogeny, and clinical behaviour, such as response to cytotoxic therapy, can often be predicted. These studies may be used to demonstrate the presence of a surface antigen to which monoclonal antibody-based therapy has been developed (for example, CD20 and rituxumab). Sometimes, malignant cells demonstrate lineage infidelity, with expression of a pattern of surface markers that does not correspond to a normal cellular counterpart. This may fortuitously provide an immunophenotypical fingerprint to detect small amounts of disease, early relapse, or minimal residual disease after therapy.

Genetic studies

The high proliferative rate of lymphocytes and the genetic events that occur within them, sets the stage for the development of chromosomal translocations which are aetiologically linked to malignant transformation. Increasingly, haemopoietic cancers are being defined genetically by the presence of specific, non-random chromosomal translocations. The detection and study of these translocations has increased diagnostic precision, has provided insights into the molecular mechanisms of oncogenesis, and has revealed molecular targets for rational therapeutic design. Chromosomal translocations can be demonstrated using classical cytogenetic techniques. When cytogenetics is technically unsuccessful, specific translocations may also be detected using the polymerase chain reaction (**PCR**) and fluorescence *in situ* hybridization (**FISH**). In addition, these methods may be used to identify the presence of specific viral sequences, such as those encoded by EBV and HHV-8. These highly sensitive and specific techniques are increasingly being applied to the detection of minimal residual disease.

As described above, the hallmark of lymphocyte differentiation is the somatic rearrangement of the antigen-receptor genes, immunoglobulin in the case of B cells and

the TCR in the case of T cells. Each lymphocyte clone has a unique arrangement of the components of the antigen-receptor genes, while cells of non-lymphocyte lineage preserve the germline structure of these genes. Lymphoproliferative malignancies are composed of clonal proliferations arising from a single cell with a rearranged antigen-receptor locus. The pattern of gene rearrangement helps to characterize the lineage and stage of differentiation of the tumour. For example, pre-B-cell acute lymphoblastic leukaemia cells usually contain rearranged heavy-chain genes with germline light-chain genes, whereas B-CLL cells usually have a rearrangement of both heavy- and light-chain genes and express surface immunoglobulin. Furthermore, since clonal populations of lym-phocytes all contain the same antigen-receptor rearrangement, these cells possess a 'molecular signature' that is unique to the malignant clone.

Consequently, antigen-receptor rearrangements have become the target of DNA diagnostic techniques for diagnosing and following lymphoproliferative malignancies. Antigen-receptor rearrangements can be detected by several methods including Southern blot and PCR-based techniques. The genetic detection of clonal B-cell populations was first achieved using Southern blotting. For Southern analysis, DNA is digested with restriction endonucleases, blotted on nitrocellulose membranes, and hybridized with probes specific for the immunoglobulin loci. Somatic rearrangement of the immunoglobulin loci, with concomitant excision of the intervening DNA, results in loss of the normal restriction endonuclease sites that lie within the immunoglobulin locus. Clonal populations of B cells then give rise to restriction fragments of altered size when probed with immunoglobulin gene sequences. This technique is still the 'gold standard' for determining B-lymphoid clonality of confusing lymphoproliferations. It has also been used for the detection of minimal residual disease (**MRD**). However, PCR-based techniques have largely supplanted Southern analysis for the detection of MRD. For these studies, PCR is performed using oligonucleotide primers based on conserved sequences within the immunoglobulin heavy-chain locus; approximately 70 to 90 per cent of rearrangements can be detected by this approach. To detect MRD with maximal sensitivity, such rearrangements are then subjected to sequence analysis to determine the antigen-specific sequences unique to the tumour rearrangement. An allele-specific oligonucleotide can then be synthesized and used in a PCR analysis that can detect residual clonal populations representing as few as $1:10^{5}$ cells.

Further reading

Berliner N, Smith B (1991). The pathobiology of lymphoproliferative disease. In: Hoffman R, *et al*., ed. *Hematology basic principles and practice*, pp 897–911. Churchill Livingstone, New York.

Delves P, Roitt I (2000). Advances in immunology 1. *New England Journal of Medicine* **343**, 37–49.

Delves P, Roitt I (2000). Advances in immunology 2. *New England Journal of Medicine* **343**, 108–17.

Foon KA, Todd RF 3rd (1986). Immunologic classification of leukemia and lymphoma. *Blood* **68**, 1–31.

Look A (1997). Oncogenic transcription factors in human acute leukemias. *Science* **278**, 1059–64.

Macintyre EA, Delabesse E (1999). Molecular approaches to the diagnosis and evaluation of lymphoid malignancies. *Seminars in Hematology*, **36**, 373–89.

Sell S (1996). *Immunology, immunopathology, and immunity*. Appleton and Lange, Stamford, CT.

Strauchen J (1998). *Diagnostic histopathology of the lymph node*. Oxford University Press, New York.

Wickremasinghe R, Hoftbrand A (1999). Biochemical and genetic control of apoptosis: relevance to normal hematopoiesis and hematological malignancies. *Blood* **93**, 3587–600.

22.4.3 Lymphoma

James O. Armitage

Introduction

Lymphomas represent malignancies of lymphoid cells and almost always present as solid tumours, ranging from among the least to among the most aggressive of the human malignancies. They have in common a frequent response to available therapies, and a significant subset of patients who develop lymphomas can be cured.

Lymphomas are usually divided into Hodgkin's disease and non-Hodgkin's lymphomas. The non-Hodgkin's lymphomas are much more frequent, with almost 60 000 new cases being diagnosed in the United States annually and approximately 8000 new cases diagnosed each year in the United Kingdom. Non-Hodgkin's lymphomas are increasing in incidence at a higher rate than almost all other malignancies; for instance, in the United States the incidence has increased at a rate of approximately 4 per cent per year since 1950. In contrast, Hodgkin's lymphomas occur approximately 7500 times per year in the United States and 1200 times per year in the United Kingdom. The incidence of Hodgkin's disease appears to be stable.

Presenting manifestation

Patients with lymphoma most commonly present with lymphadenopathy. However, a variety of presentations are possible. These include systemic symptoms such as fevers, night sweats, weight loss, and pruritus, which are believed to be the result of the release of cytokines by normal or malignant cells. Patients can present with symptoms secondary to a mediastinal or retroperitoneal mass such as superior vena cava obstruction, pleural effusion, pericardial tamponade, abdominal or back pain, intestinal obstruction or perforation, gastrointestinal bleeding, or renal failure from urethral obstruction. Central nervous system presentations include primary brain tumours and signs of meningeal involvement and spinal cord compression. Patients might present with cytopenia secondary to either bone marrow involvement or autoimmune destruction of the formed elements of the blood. Symptoms secondary to the overproduction of a monoclonal immunoglobulin or hypogammaglobulinaemia can be seen. In short, the possible presentations of lymphomas are so varied that the diagnosis should be considered in almost all patients and not just restricted to those presenting with lymphadenopathy or splenomegaly.

Establishing a diagnosis

The diagnosis of lymphoma should always be based on evaluation by an expert haematopathologist of, preferably, an adequate lymph node biopsy, or an extranodal tumour mass if lymph nodes are unavailable. Needle aspirates or small biopsies should be avoided as the basis for diagnosing lymphoma whenever possible. As one of the major challenges that pathologists face is the diagnosis of lymphoma, it is important not to handicap the haematopathologist by providing inadequate material. The differential diagnosis that the pathologist considers when diagnosing a lymphoma includes benign proliferations of lymphoid tissue, malignancies of myeloid cells, non-haemopoietic malignancies, viral infections, and unusual disorders such as Castleman's disease and giant lymph node hyperplasia. Having tissue available for immunological studies and/or genetic studies will frequently help to confirm the diagnosis.

Patient evaluation

Once the diagnosis of a type of lymphoma has been established, a series of studies should be carried out to determine the extent of disease. The anatomical spread of disease is usually expressed as an Ann Arbor Stage (Table 1). This staging system was originally developed for Hodgkin's disease and divides patients into those with disease confined to one lymphatic site, multiple lymphatic sites on one side of the diaphragm, lymphatic involvement on both sides of the diaphragm, and those with bone marrow involvement, liver involvement, or other extensive extranodal disease. The Ann Arbor Stage also includes a suffix A or B indicating the absence (A) or presence (B) of unexplained fevers above 38 °C, weight loss of more than 10 per cent of the body weight in the preceding 6 months, or drenching night sweats. Additional factors can also have an impact on a patient's response to therapy and survival. For non-Hodgkin's lymphomas, these factors are incorporated into the International Prognostic Index. In this system, the Ann Arbor Stage represents one factor with an adverse risk associated with stage III or IV. Other adverse risk factors include an elevated serum lactate dehydrogenase (**LDH**) level, age of 60 years or greater, multiple sites of extranodal disease, and a reduced performance status. The International Prognostic Index Score is determined by adding the adverse risk factors.

The laboratory and radiological evaluation of patients with lymphoma typically involves a standardized series of tests, such as a complete blood count, erythrocyte sedimentation rate determination, chemistry studies reflecting major organ function, computed tomography scans of the chest, abdomen, and pelvis, and a bone marrow biopsy (Table 2). In patients with non-Hodgkin's lymphoma, serum LDH, serum b_2-macroglobulin, and serum protein electrophoresis are often useful adjuncts. A gallium scan performed before therapy can identify sites of involvement in the majority of patients with Hodgkin's disease and those with aggressive non-Hodgkin's lymphoma. In addition to potentially altering the diagnostic stage, this technique allows a more accurate restaging of disease at the completion of therapy than can be obtained simply with computed tomography (**CT**) scans. However, CT scans only show anatomical findings, and some patients with lymphomas in the mediastinum and retroperitoneum do not have complete regression of their initial masses because of a sclerotic reaction to the tumour. Moreover, in patients who have actually

achieved a complete remission, the gallium scan (that is to say, a functional as opposed to an anatomical study) will typically have reverted to normal. Other studies can be useful in particular situations. Most patients will be given a chest radiograph. This offers an easy way to follow mediastinal or pulmonary involvement. Magnetic resonance imaging (**MRI**) studies are particularly useful in evaluating suspected bone or central nervous system sites of involvement by a lymphoma. Technetium scans of bone or the liver and spleen are occasionally valuable in detecting occult sites of involvement by a lymphoma. In some patients, abdominal ultrasonography will provide a more economical way to follow intra-abdominal disease.

Bilateral, lower limb lymphangiography was formerly used to evaluate pubic and retroperitoneal node involvement in most patients with lymphoma. However, this study is difficult to perform and is used only occasionally today. Patients with Hodgkin's disease and many patients with non-Hodgkin's lymphoma once routinely underwent staging laparotomy before the initiation of therapy to search for occult intra-abdominal disease. This approach is rarely used today because of the quality of available non-invasive staging procedures and the effectiveness of therapy. The one remaining place for staging laparotomy is for a patient with supradiaphragmatic disease who wants to be treated with limited radiotherapy alone. A laparotomy provides the surest evidence of the absence of intra-abdominal disease in this patient.

The studies necessary to evaluate a new patient with lymphoma and provide prognostic information and a therapy plan are presented in Table 2.

Pathobiology of lymphoma

Introduction

Increased understanding of the biology of the immune system has allowed the various lymphomas to be subclassified, and provided new prognostic information and new potential targets for therapy. Since lymphomas are malignancies of lymphocytes, the surface proteins involved in cell recognition and intercellular signalling can be expected to be important. Although the genetics of lymphomas are complicated, they too are beginning to be unravelled. Information gleaned from all these studies is likely to further change both the classification and therapy of the lymphomas.

Immunology

The recognition of new surface antigens has improved the ability to recognize specific subtypes of lymphoma. For example, discovery of the Ki-1 (CD30) antigen by investigators in Germany provided a marker for the Reed–Sternberg cells in classical Hodgkin's disease. However, it was soon discovered that this antigen was found on the surface of cancers that were previously felt to be undifferentiated carcinomas and malignant histiocytosis. This observation allowed the description of anaplastic large T/null-cell lymphoma as a diagnostic entity and, more importantly, allowed some patients with lymphoma to receive appropriate therapy.

The recognition of specific antigens by standardized antibodies has improved the accuracy of diagnosis. Some of the more commonly recognized antigens are presented in Table 3. A characteristic pattern of occurrence can be a key factor in making an accurate diagnosis. Some types of lymphoma, such as follicular lymphoma and nodular sclerosing Hodgkin's disease, can be diagnosed accurately without immunological studies. Others such as all T-cell lymphomas, diffuse large B-cell lymphoma, and mantle-cell lymphoma can only be accurately diagnosed with immune markers.

Genetics

A theme common to malignant disorders is the abnormal expression of specific genes. The search for these genes was facilitated by the frequent occurrence of chromosomal abnormalities detectable by cytogenetic studies. These abnormalities include chromosomal deletions or deletions of parts of a chromosome, chromosomal duplications, and translocation of genetic material from one chromosome to another. Chromosomal translocations, through studying the sites of chromosome breakage, led to the discovery of a number of genes that appear to be important in lymphomagenesis or in determining the character of a particular lymphoma. The best documented chromosomal translocations associated with lymphomas along with the involved oncogenes are presented in Table 4.

Genetic abnormalities determine the nature of the lymphoma by leading to the overexpression, underexpression, or abnormal expression of specific genes. The genes involved, termed 'oncogenes', are typically those that regulate the cell cycle or differentiation. Since the work of genes is done by the proteins for which they code, the underexpression, overexpression, or abnormal expression of specific proteins is an increasing source for study. In some cases, protein expression might be abnormal despite no obvious translocation. For example, diffuse large B-cell lymphoma displays the t(14;18) in approximately 30 per cent of patients. This translocation involves the *BCL-2* gene on chromosome 18, whose protein product is involved in suppressing apoptosis (that is, the mechanism of cell death usually triggered by chemotherapeutic agents). Tumours can overexpress the BCL-2 protein with or without the t(14;18). Overexpression of BCL-2 protein might be expected to lead to the increased survival of lymphoma cells when they are exposed to therapeutic agents. In patients with diffuse large-cell lymphoma, an increased relapse rate has been associated with overexpression of the BCL-2 protein, rather than with the t(14;18).

Specific chromosomal translocations are highly associated with certain subtypes of lymphoma and thus are useful in diagnosis. These include the t(2;5) and anaplastic large T/null-cell lymphoma, the t(14;18) in follicular lymphoma, the t(8;14), t(2;8), and t(8;22) in Burkitt's lymphoma, and the t(11;14) in mantle-cell lymphoma. Cytogenetic studies in most patients with non-Hodgkin's lymphoma display a large number of chromosomal abnormalities. However, only a few have been shown to be of diagnostic or prognostic significance. No such abnormalities have been consistently identified in patient's with Hodgkin's disease.

Future genetic studies in lymphomas are likely to focus on specific gene expression. The new 'lympho chip' technology will allow the several thousand genes typically expressed in lymphoid cells to be studied simultaneously. Patterns of gene expression may well provide new methods of classifying lymphomas, provide new prognostic information, and direct therapy or provide new targets for therapy.

The general principles of therapy of lymphoma

Those treatments effective in the management of patients with cancer include surgery, radiotherapy, cytotoxic chemotherapy, and a variety of new approaches developed through increasing understanding of the biology of the immune system. The latter include cytokines, antibodies, and attempts to direct an immune reaction against cancer. Because few patients with lymphoma have truly localized disease, surgery has not been a major treatment modality. Radiotherapy, since its utilization in medicine in the first part of the twentieth century, has been a major treatment modality for patients with lymphoma. Radiotherapy is limited in its application by toxicity. Its curative potential depends upon being able to achieve a tumoricidal dose (typically 3000–4000 cGy) without irreversibly injuring normal organs. Thus, the site of involvement by a lymphoma, as well as the number of sites involved, can limit the effectiveness of this treatment, since toxicity increases with the volume of tissue radiated. If a lymphoma is truly localized, radiotherapy is often a curative. Two approaches have been utilized to make radiotherapy a 'systemic' treatment. One involves radiation of the total body. When this is part of a bone marrow transplant regimen, a total dose of 1000–1500 cGy can be administered. More recently, it has been demonstrated that it is possible to give higher doses of radiotherapy to multiple areas by attaching radioactive molecules to antibodies that home to sites of involvement by lymphoma.

Cytotoxic chemotherapeutic agents were first discovered in the 1940s when mechlorethamine (that is, the nitrogen mustard gas used in warfare) and, subsequently, methotrexate were found to cause regressions in immune system malignancies. A wide variety of agents have since been shown to be able to cause regressions in a significant proportion of patients with lymphomas (Table 5). Unfortunately, early studies showed that regressions induced by single agents were almost invariably followed by regrowth of the tumour and eventual death of the patient. In an attempt to circumvent this, combinations of chemotherapeutic agents were first utilized in the 1960s and early 1970s. The drugs were combined by attempting to choose agents with different mechanisms of action and non-overlapping toxicities to allow the administration of doses that were near to the maximum tolerated dose with an individual agent. In both childhood acute leukaemia and Hodgkin's disease this approach was validated by the cure of a significant number of patients. Today, several combination-chemotherapy regimens with acceptable toxicity have been shown to be effective and are widely used worldwide (Table 6). All regimens are not equally good for treating all types of lymphoma.

Increasing knowledge of the immune system has further led to the recognition that a number of biologically active molecules can cause regression of lymphomas and, in some cases, impact on survival. The first such agent to be widely used was interferon-alpha, which has some activity in both non-Hodgkin's lymphoma and Hodgkin's disease. When administered at an adequate dose (at least 36 units per month), it has been shown to prolong survival in patients with poor-prognosis follicular lymphoma who received an anthracycline-containing chemotherapy regimen as their initial treatment. The ability to produce monoclonal antibodies has provided new therapeutic molecules. In B-cell non-Hodgkin's lymphomas, antibodies directed against the CD20 molecule have been incorporated into clinical practice. Rituximab has been shown to be active in a variety of B-cell lymphomas, and the new radiolabelled antibodies such as tositumomab will soon become available. An

antibody directed against CD25 has been introduced into therapy for patients with cutaneous T-cell lymphoma.

Very high doses of cytotoxic chemotherapeutic agents with or without radiotherapy and biologically active molecules have been utilized in the treatment of patients with lymphomas as part of the bone marrow transplantation procedure. This involves the administration of very high doses of antilymphoma therapy in an attempt to overcome presumed treatment resistance. Patients are rescued from the toxicity of treatment by the reinfusion of haemopoietic stem cells. The patient's own haemopoietic stem cells (an autologous transplant) or those from another individual with identical HLA genes (an allogeneic transplant) can be utilized. Cells for this procedure can be obtained from either bone marrow or peripheral blood. Autologous transplantation has been widely used for patients with lymphoma and shown to be able to cure patients with relapsed Hodgkin's disease and aggressive non-Hodgkin's lymphoma. In aggressive non-Hodgkin's lymphomas, a probable increased cure rate has been demonstrated by utilizing adjuvant transplantation following initially effective standard chemotherapy in patients with a poor prognosis. Transplantation is being widely utilized in patients with follicular lymphoma, but the curability of autologous transplantation in this setting remains a point of controversy, and allogeneic transplantation, while apparently curative, has a high mortality rate.

Various new treatments are being studied for patients with lymphoma. These include attempts to stimulate the patient's endogenous immune system to develop antibodies against lymphomas; such 'tumour vaccines' are now in clinical trials. Another approach has been called 'antisense therapy', which involves the use of antisense oligonucleotides aimed at interrupting the transcription and expression of key genes.

A number of factors need to be taken into account when formulating a treatment recommendation for a patient with lymphoma (Table 7). These include the patient's age, general health, extent of disease, likelihood of cure, coexisting illnesses, long-term goals, and concerns about treatment toxicity. This decision should be made in conjunction with the patient, and requires good judgement in addition to technical knowledge. The aggressiveness of the treatment that is finally chosen will often depend upon the physician's interpretation of the chances for cure. It is obvious that more toxicity would be acceptable if the goal was cure rather than palliation. For this reason, patients with definitely curable lymphomas, such as diffuse large B-cell lymphoma and Burkitt's lymphoma, are almost always treated promptly with intensive regimens. In contrast, the best treatment for patients with follicular lymphoma remains a point for intense debate. Since the curability of this disease is in question, many physicians would favour no initial therapy in an asymptomatic patient. However, as discussed below, this is not a simple decision.

For most patients, the goal of therapy is to achieve a complete remission. This implies the disappearance of all symptoms and objective evidence of lymphoma. In practice, a complete remission is documented by repeating all abnormal staging studies after several cycles of therapy or at the completion of the planned therapy. Sometimes, persisting masses visualized on CT scans represent residual fibrosis rather than persisting tumour. In certain patients, a previously abnormal gallium scan reverting to normal would resolve this dilemma, but, on rare occasions, a biopsy might be required. Documentation of complete remission is important. Patients who achieve a complete remission have a chance for cure; those who do not achieve a complete remission with initial therapy will often go directly to second-line treatments.

Patients who fail to be cured with initial therapy, either because they do not achieve an initial remission or they relapse from remission, are candidates for what has been termed 'salvage therapy.' These second-line regimens can regularly cause tumour regression in most patients with lymphoma and can occasionally produce long-term, disease-free survival. However, for most patients, the only curative approach in this setting is bone marrow transplantation. The toxicity of bone marrow transplantation limits its use to patients under certain ages (under 70 years of age for autologous transplantation and under 55 years for allogeneic transplantation), who have a good performance status, without serious compromise of major organ function, and to patients who do not have bulky/chemotherapy-refractory disease.

Hodgkin's disease

In 1832, Thomas Hodgkin of Guy's Hospital, London, reported seven patients who died from a disorder involving lymph node and spleen enlargement. Then, early in the twentieth century, Sternberg and Reed independently described the characteristic giant cells that now bear their name. This made it possible for pathologists to separate the disorder we now know as Hodgkin's disease from other lymphomas.

Incidence and epidemiology

Unlike non-Hodgkin's lymphomas, the incidence of Hodgkin's disease appears to be stable with approximately three new cases per 100 000 per year in western countries. As demonstrated in Fig. 1, this illness displays a peculiar bimodal distribution of occurrence with peaks in young adulthood and older age. This dual peak has led some to propose that Hodgkin's disease actually represents two illnesses, with the earlier peak being related to an infectious aetiology and the latter representing a true malignancy. However, there is little evidence to support this distinction.

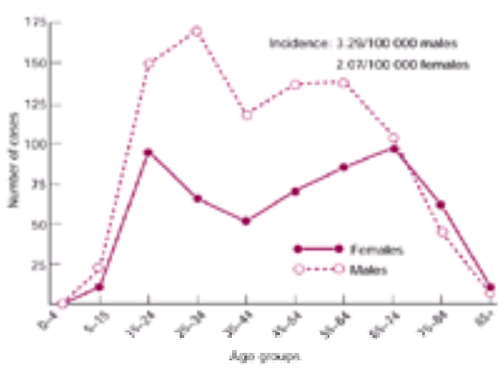

Fig. 1 Age distribution of Hodgkin's disease expressed as new cases registered in England and Wales in 1973.

A strong association has been demonstrated between the occurrence of Hodgkin's disease and infection by the Epstein–Barr virus (**EBV**). Monoclonal or oligoclonal proliferation of EBV-infected cells is found in 20 to 40 per cent of patients with Hodgkin's disease. Patients infected by the human immunodeficiency virus (**HIV**) are at an increased risk for Hodgkin's disease in addition to non-Hodgkin's lymphomas.

The subtypes of Hodgkin's disease vary geographically and by age group. Patients in western countries who develop Hodgkin's disease in young adulthood usually have the nodular sclerosis subtype. Patients from Third-World countries, elderly patients, and those infected with HIV usually have mixed-cellularity or lymphocyte-depletion Hodgkin's disease.

Hodgkin's disease is approximately 100 times more likely in an identical twin of an infected patient. Numerous instances of case-clustering have been described. While these might be taken as evidence of a genetic or infectious aetiology, the cause of Hodgkin's disease remains unknown.

Pathology

The diagnosis of Hodgkin's disease requires the identification of Reed–Sternberg cells in a characteristic cellular background; however, the cell of origin for the Reed–Sternberg cell remains unknown. The subtypes of Hodgkin's disease are presented in Table 8. Nodular sclerosing Hodgkin's disease is characterized by bands of fibrosis that are often visible to the naked eye when a slide is held to the light. Mixed-cellularity Hodgkin's disease typically displays a larger number of Reed–Sternberg cells in a mixed-cellular background including lymphocytes, macrophages, and eosinophils. Lymphocyte-depletion Hodgkin's disease can present with a very large number of Reed–Sternberg cells and atypical mononuclear cells, or a background of diffuse fibrosis with occasional Reed–Sternberg cells. Lymphocyte-predominance Hodgkin's disease is characterized by large numbers of small lymphocytes and histiocytes with occasional Reed–Sternberg cells. The growth pattern can be nodular or diffuse.

It is now clear that patients with nodular lymphocyte-predominance Hodgkin's disease have a different illness that is, in many ways, more like a B-cell non-Hodgkin's lymphoma. The Reed–Sternberg cells in nodular lymphocyte-predominance Hodgkin's disease express the leucocyte common antigen and other B-cell markers including CD20. Typical Reed–Sternberg cells do not express the leucocyte common antigen, but are CD15- and CD30-positive. Staining for CD15 and CD30 can

resolve difficult diagnostic problems in some cases.

Clinical features and evaluation

Patients with classical Hodgkin's disease usually present with palpable non-tender lymphadenopathy. In most patients, lymph nodes are discovered in the cervical, supraclavicular, and axillary regions. More than half the patients have mediastinal lymphadenopathy at diagnosis, and symptoms from a large mediastinal mass are often the initial presentation. Subdiaphragmatic presentation of Hodgkin's disease is unusual and more common in older males. Approximately one-third of patients with classical Hodgkin's disease present with fevers, night sweats, and/or weight loss.

Hodgkin's disease can present as a fever of unknown origin. Presentation as a 'fever of unknown origin' is more common in older patients who have mixed-cellularity or lymphocyte-depletion Hodgkin's disease and who present with disease in abdominal nodes. Fevers associated with Hodgkin's disease occasionally persist for days to weeks, followed by afebrile intervals, and then reoccurrence of the fever. This pattern is known as Pel–Ebstein fever. Unusual presentations of Hodgkin's disease include severe and unexplained pyloritis, paraneoplastic cerebellar degeneration, nephrotic syndrome, immune haemolytic anaemia and thrombocytopenia, hyper-calcaemia, and pain in lymph nodes on alcohol ingestion.

The diagnosis of Hodgkin's disease is based on a review of an adequate biopsy by an expert haematopathologist. Subsequent evaluation should include a careful history and examination, complete blood count, erythrocyte sedimentation rate determination, serum chemistry studies including serum lactate dehydrogenase, chest radiograph, computer demography of the chest, abdomen, and pelvis, and bone marrow biopsy. Gallium scans can be performed to document radioisotope uptake by the tumour, which can then be repeated at the completion of therapy to document remission. Bipedal lymphangiograms can be useful if radiologists expert in carrying out the procedure are available. Staging laparotomies are now rarely indicated.

Nodular lymphocyte-predominance Hodgkin's disease, as noted above, is a different clinical entity from classical Hodgkin's disease. These patients represent less than 5 per cent of all patients found to have Hodgkin's disease. The evaluation of such patients is carried out in a similar way to that for classical Hodgkin's disease. However, nodular lymphocyte-predominance Hodgkin's disease tends to follow a chronic, relapsing course and sometimes transforms to diffuse large B-cell lymphoma.

Prognostic factors

The major factors determining treatment outcome for patients with Hodgkin's disease include the Ann Arbor stage, the presence or absence of systemic symptoms, age, and gender. Patients with asymptomatic, localized disease who are young and female have the best outlook. Histological subtypes do not appear to have major independent prognostic significance. Patients with nodular sclerosing Hodgkin's disease have a better outcome than those with mixed-cellularity or lymphocyte-depleted Hodgkin's disease. However, adverse prognostic factors are more commonly found in patients with the latter histological subtypes. The results of several laboratory studies can predict outcome in patients with Hodgkin's disease. These include anaemia, an elevated erythrocyte sedimentation rate, and a low albumin level. The erythrocyte sedimentation rate is sometimes used to follow the course of patients with Hodgkin's disease since it tends to normalize with successful treatment.

The most important factor in predicting outcome for patients with Hodgkin's disease is their response to therapy. Patients who have a prompt, complete response to chemotherapy and/or radiotherapy have the best outlook and are most likely to be cured. It is important to note that residual masses do not always represent persisting disease. This is particularly true for residual mediastinal and retroperitoneal masses. These sites tend to be associated with a considerable amount of fibrosis that can persist after effective therapy. Normalization of a gallium scan can be used to document remission. Patients who relapse after initial successful treatment for Hodgkin's disease can sometimes be effectively treated with further chemotherapy or radiotherapy. The chances for successful treatment, in part, depend upon the duration of initial remission in addition to other prognostic factors present at relapse. Patients with a longer initial remission are more likely to be successfully retreated.

Primary therapy

Patients with localized Hodgkin's disease (stage I or non-bulky stage II) can be cured with extended-field radiotherapy. Patients with supradiaphragmatic disease are typically treated with a radiotherapy port that is often referred to as a mantle. This involves treatment of the cervical, supraclavicular, axillary, and mediastinal lymph nodes. In the absence of a staging laparotomy before therapy, upper abdominal nodes and spleen are also often treated. A dose of 3500 to 4400 cGY is usually administered in fractions of 175 to 200 cGy daily, 5 days per week for approximately 4 weeks.

Patients with otherwise localized Hodgkin's disease who present with a large mediastinal mass pose special therapeutic problems. A large mediastinal mass is often defined as one whose maximum diameter is greater than one-third of the maximum thoracic diameter. Treatment with radiotherapy alone, or chemotherapy alone, is associated with a high relapse rate. Large mediastinal masses are one definite indication for combined modality therapy.

Patients who present with B-symptoms or stage III or IV disease are best treated initially with a combination-chemotherapy regimen. If complete remission is documented after completing a course of chemotherapy, the majority of patients will be cured. Patients who have large masses often receive adjuvant radiotherapy to the sites of previous bulky disease after completing the chemotherapy regimen. While the original **MOPP** (mechlorethamine, Oncovin (i.e. vincristine), procarbazine, prednisone) regimen was effective in the treatment of Hodgkin's disease, it has now been shown that regimens which include **ABVD** (doxorubicin (Adriamycin), bleomycin, vinblastine, dacarbazine) are associated with a better outcome (Table 6).

Elderly and pregnant patients pose special therapeutic problems. In Hodgkin's disease, the elderly have a much worse prognosis than younger patients: patients over 60 years of age at the time of diagnosis have a survival rate less than half that of younger patients. Elderly patients with localized disease seem to benefit from radiotherapy in a manner comparable to younger patients. However, older patients tolerate aggressive chemotherapy regimens much less well and, even if the drugs can be administered, have a higher relapse rate.

Hodgkin's disease, since it occurs frequently in young adults, is not rarely diagnosed in pregnant women. It is now clear that Hodgkin's disease can be treated with chemotherapy at any point during pregnancy with a chance of a good treatment outcome and a surviving infant. However, the risks are higher in the first trimester. Most physicians would favour delaying therapy past the first trimester, if possible, and would discuss the possibility of a therapeutic abortion with the patient. If the decision is made to treat a pregnant patient with chemotherapy, it must be remembered that the fetus will be myelosuppressed in a manner similar to the mother. This must be taken into account when planning delivery of the baby. Radiotherapy is generally not used in pregnant patients because of its teratogenic potential.

The optimal treatment for lymphocyte-predominance Hodgkin's disease is unclear. Some clinicians favour no initial therapy in asymptomatic patients. However, potentially curative radiotherapy to localized disease seems wise. The clinician must be alert for transformation to diffuse, large B-cell lymphoma.

Treatment of relapse

Approximately 25 to 35 per cent of patients treated with chemotherapy for stage III or IV Hodgkin's disease will suffer relapse after achieving a remission. A small proportion of patients will fail to enter initial complete remission. Patients who fail to enter complete remission or relapse within 1 year of completing therapy have a poor prognosis with further standard chemotherapy. Autologous bone marrow transplantation can be curative in 25 to 50 per cent of such patients, and is the treatment of choice. Patients who have an initial remission of longer than 1 year pose a more complicated therapeutic problem. These patients are likely to achieve a second remission with a standard chemotherapy regimen. However, long-term follow-up has demonstrated that the majority of these remissions are not durable, and many physicians would recommend autologous transplantation to such patients. The occasional patient with a localized relapse after chemotherapy can sometimes be cured with radiotherapy. Patients who relapse after treatment with initial radiotherapy have an excellent result with standard chemotherapy regimens and a high likelihood of cure.

Treatment complications

The treatment of Hodgkin's disease is associated with both short-term and long-term complications. Prominent short-term complications include hair loss, emesis, fatigue, anaemia, and infection due to chemotherapy-induced neutropenia. Hair loss is usually transient. Emesis can be prevented in almost all patients by 5-hydroxytryptamine antagonists such as ondansetron and granisetron. Anaemia and fatigue do not usually limit the administration of therapy. Chemotherapy-induced

neutropenia is a major problem and neutropenic fever needs to be managed aggressively with intravenous antibiotics after cultures are obtained. Even so, treatment for Hodgkin's disease is administered entirely on an outpatient basis.

Delayed toxicity from the treatment of Hodgkin's disease has become a major problem for young patients who are cured of Hodgkin's disease and have been followed for extended periods. In fact, for patients with good-prognosis Hodgkin's disease, long-term complications might lead to a higher mortality rate than the Hodgkin's disease itself.

Most of the serious complications of radiotherapy appear after long follow-up. In the first few months after treatment, some patients will develop an electric shock sensation down the spine and into the legs on flexion of the neck. This represents Lhermitte's syndrome and needs to be recognized so that further evaluation is not carried out. It is usually transient. In some patients, delayed pulmonary fibrosis or cardiac injuries are associated with thoracic radiotherapy. Modern radiotherapy techniques have minimized the risk of these problems, but accelerated coronary artery disease is a significant problem and leads to a number of treatment-related deaths. The major delayed problem with radiotherapy is the development of secondary cancers. This risk begins to appear beyond 10 years post-therapy, and by 20 years after therapy leads to a significant number of deaths. Patients treated with thoracic radiotherapy for Hodgkin's disease should be strongly encouraged not to smoke to reduce the risk of lung cancer. Young women should have screening mammography instituted 5 to 10 years earlier than for women not irradiated, or by 10 years after completing treatment.

Patients who receive radiotherapy to the neck have a high risk of developing subsequent hypothyroidism. Follow-up in such patients should include periodic quantitation of their thyrotropin levels to anticipate this problem. Some patients treated with either radiotherapy or chemotherapy will develop herpes zoster. This diagnosis does not necessarily signify a relapse of Hodgkin's disease.

Long-term problems associated with chemotherapy include treatment-related leukaemia, infertility, and aseptic necrosis of bone. Infertility is most likely in patients who receive alkylating agent-containing regimens. Most young males who received MOPP become infertile. In women, the risks of infertility are age-related. Women over 30 years of age are much more likely to be permanently infertile than those under 30 years. However, in any patient, resumption of fertility is possible and the patient should be aware of this. Infertility is less of a problem in patients who receive the ABVD regimen. Males very anxious to retain fertility can be offered semen storage and women, in extraordinary cases, can be offered egg storage.

Treatment-related leukaemia is most frequent in patients who receive chemotherapy regimens containing alkylating agents and who are treated on more than one occasion. Young patients treated with only one chemotherapy sequence are unlikely to develop leukaemia. The incidence of leukaemia rises dramatically in patients over 40 years of age, and in those who receive alkylating agents on more than one occasion. Leukaemia is unusual in patients treated with ABVD. The combination of chemotherapy and radiotherapy seems to increase the risk of leukaemia. The leukaemias that occur in this setting usually present with myelodysplasia and typically have genetic abnormalities involving chromosomes 5, 7, and 8. Etoposide can lead to the development of acute leukaemia without a preceding myelodysplasia that involves abnormalities on chromosome 11.

Patients who receive corticosteroid treatment as part of a combination therapy are at risk for aseptic necrosis of the femoral heads. Those who develop hip pain on follow-up should be evaluated for this possibility.

Non-Hodgkin's lymphoma

Incidence

In much of the world, it appears that the incidence of non-Hodgkin's lymphoma is increasing, but the incidence still varies widely between countries. The incidence appears to be approximately 2 cases per 100 000 per year in the Orient, 10 per 100 000 per year in the United Kingdom, and more than 15 per 100 000 per year in the United States. In the United States, the disease increased in frequency by approximately 4 per cent per year between 1950 and the mid-1990s. This increased incidence is seen in patients of all ages but more striking in the elderly. However, recent data suggest that the rate of increase may be slowing.

A recent study showed that the specific types of non-Hodgkin's lymphomas vary in occurrence between countries. For example, follicular lymphoma is more common in North America than in Europe or Asia. T-cell lymphomas have been seen more frequently in Asia, and certain types of T/NK-cell lymphomas (**NK**, natural killer) such as angiocentric nasal lymphomas seem restricted to only a few countries in Asia and Latin America. The explanation for this difference in different geographical settings is unclear.

Aetiology

Various aetiological factors, either proven or suggested to be associated with the development of non-Hodgkin's lymphoma, are listed in Table 9. It is now clear that exposure to certain agriculture chemicals does increase the risk of this disease. A variety of immune deficiencies, such as those associated with immunosuppression following organ transplantation and various inherited immune deficiencies, are also associated with an increased risk of developing non-Hodgkin's lymphoma. Patients with rheumatoid arthritis and systemic lupus erythematosus appear to be at increased risk.

A variety of infectious agents have been shown to be associated with the development of non-Hodgkin's lymphoma. Gastric *Helicobacter pylori* infection is associated with the development of gastric **MALT** (mucosa-associated lymphoid tissue) lymphoma. In the case of MALT lymphomas, eradication of the *Helicobacter pylori* infection by antibiotics can lead to regression of the lymphoma in a significant number of patients. **HTLV**-1 (human T-cell lymphoma/leukaemia virus-1) appears to be the cause of a specific type of non-Hodgkin's lymphoma, seen predominantly in southern Japan and the Caribbean, called adult T-cell lymphoma/leukaemia. The Epstein–Barr virus has been associated with Burkitt's lymphoma in Africa, the development of aggressive B-cell lymphomas in immunosuppressed patients, Hodgkin's disease, and certain aggressive T-cell lymphomas. HHV-8 (human herpesvirus-8) has been closely associated with a rare, diffuse, large B-cell lymphoma called effusion lymphoma. HIV (human immunodeficiency virus) infection can lead to the development of aggressive B-cell lymphoma. It is likely that the future will see more associations between lymphomas and specific infectious agents.

REAL/WHO classification

The classification of non-Hodgkin's lymphomas has changed several times during the twentieth century. The first popular classification proposed by Gall and Mallory divided lymphomas into giant follicular lymphoma, reticulum-cell sarcoma, and lymphosarcoma. Both the lack of adequate clinical correlation and clear definitions of the entities led to further proposals. Henry Rappaport recognized the importance of growth pattern in the prognosis of non-Hodgkin's lymphomas, and put forward his system that divided patients into those with nodular (i.e. follicular) or diffuse lymphomas and those with large- or small-cell lymphomas. However, this system was proposed before the recognition that lymphomas were all malignancies of lymphocytes and before the discovery of the existence of subtypes of lymphocytes. The advent of modern immunology led to new classification systems proposed by Carl Lennert and colleagues in Europe and Lukes and Collins in the United States. The Kiel classification proposed by Lennert and colleagues became the most widely used system in Europe. An attempt to unify the classifications of lymphomas led to the development of the Working Formulation. This is a compromise system taking major elements from the Rappaport classification, Kiel classification, and the Lukes/Collins classification. It became widely used in the United States but less so in Europe.

In the 1990s a group of haematopathologists from Europe, North America, and other parts of the world proposed a new system based on not just morphology and immunophenotyping, but taking into account other genetic and biological information that had become available. In the 1990s, a number of 'new' lymphomas were discovered that did not fit into previous classification systems. These included mantle-cell lymphoma, anaplastic large T/null-cell lymphoma, and MALT lymphomas. The Revised European/American Lymphoma classification (**REAL**) classified lymphomas based on clinical pathological syndromes (in other words, real diseases) rather than simply morphology. This system was tested in a large international study and shown to be more accurate than previous systems and to have high clinical relevance. Leaders in the fields of both haematopathology and clinical haematology/oncology agreed on a modified REAL classification to be endorsed by the World Health Organization and published as the WHO classification (Table 10). This, with modifications, is likely to be the major lymphoma classification for at least the next decade. The incidence of major lymphoma subtypes according to the WHO classification are listed in Table 11. Knowledge of 10 to 12 specific subtypes of non-Hodgkin's lymphoma will allow a clinician to care for almost all patients.

International Prognostic Index

Knowledge of the specific subtype of non-Hodgkin's lymphoma is only one of two pieces of information necessary to plan the intelligent management of patients with

these disorders. The other that must be available involves the delineation of the prognostic characteristics of the individual patient. While it is true that follicular lymphoma has a higher median overall survival than diffuse large B-cell lymphoma, individual patients with follicular lymphoma might have a much worse survival because of adverse prognostic characteristics than an individual patient with diffuse large B-cell lymphoma who has good prognostic characteristics. Codification of these prognostic characteristics into a practical clinical tool was accomplished by a large international study that yielded the International Prognostic Index (**IPI**) (Table 12). The IPI is a summation of a number of specific adverse prognostic factors in an individual patient. The important factors include age greater than 60 years, Ann Arbor Stage III/IV, serum lactate dehydrogenase level greater than normal, reduced performance status, and multiple extranodal sites of involvement by lymphoma. The impact of the IPI score on the survival of patients with follicular lymphoma and diffuse large B-cell lymphoma are presented in Fig. 2. As mentioned above, the treatment plan for a patient with lymphoma must always include knowledge of the specific subtype of lymphoma and the patient's prognostic characteristics.

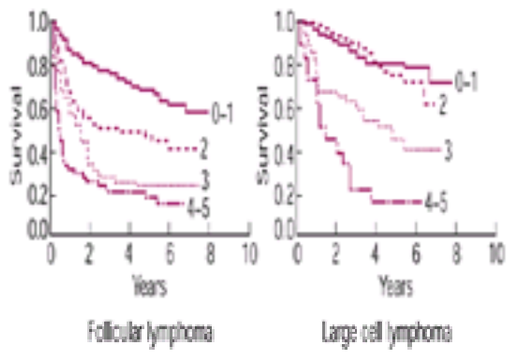

Fig. 2 The survival of 275 patients with follicular lymphoma and 388 patients with diffuse large B-cell lymphoma are shown by International Prognostic Index score.

Lymphoblastic lymphoma of B-cell and T-cell origin

Lymphoblastic lymphoma is a tumour of the precursor cells of T- and B-lymphocytes. It is intimately related to the acute lymphoid leukaemias, with the difference being the method of presentation. Sometimes it is difficult to determine when a patient should be said to have acute lymphoid leukaemia or lymphoblastic lymphoma, since bone marrow involvement is frequent with a lymphomatous presentation and lymphadenopathy and mediastinal mass are common in patients who present with leukaemia.

Most patients with lymphoblastic lymphoma have tumours derived from T-lymphoblasts, but approximately 10 per cent are B-cell in origin. The differential diagnosis of lymphoblastic lymphoma includes a blastic variant of mantle-cell lymphoma, acute myeloid leukaemia, and small round-cell lymphoma in children and young adults.

The median age of patients with lymphoblastic lymphoma is the late twenties and the majority of patients are male. Most patients have widely disseminated disease and an elevated serum LDH level. Approximately 50 per cent of patients will have bone marrow involvement. The IPI predicts outcome in lymphoblastic lymphoma. However, it has been suggested that patients can be divided into two prognostic groups. One group includes patients who have stage IV disease, elevated LDH levels, and bone marrow or central nervous system involvement; adults with these characteristics have a poor outlook. The second group comprises patients without these adverse characteristics, and these patients have a high cure rate.

Most patients with lymphoblastic lymphoma will be treated with a leukaemia-like regimen. This includes an intensive induction therapy along with central nervous system prophylaxis, and an ongoing maintenance or consolidation phase of treatment. The majority of children and young adults can be cured with this treatment approach. Patients who present with adverse risk characteristics or who relapse after initial therapy are candidates for bone marrow transplantation.

Diffuse large B-cell lymphoma

Diffuse large B-cell lymphoma is the most common non-Hodgkin's lymphoma, representing approximately one-third of all patients. It most commonly presents *de novo*, but also can develop after histological transformation of a small-cell lymphoma such as follicular, small lymphocytic, and MALT lymphoma. This tumour can arise in lymph nodes or essentially on any extranodal site including the central nervous system. Rare presentations include pleural effusions from involvement of serosal surfaces (effusion lymphoma) and multiple organ system dysfunction secondary to endothelial involvement (intravascular lymphomatosis).

B-cell lymphomas will display the CD20 antigen. Several cytogenetic abnormalities are frequently associated with diffuse large B-cell lymphoma including t(14;18) t(3;14) and t(8;14). The differential diagnosis includes undifferentiated carcinoma, acute myeloid leukaemia, and Hodgkin's disease. Occasional patients with diffuse large B-cell lymphoma have a large number of infiltrating T cells, and so can be confused with a peripheral T-cell lymphoma.

The clinical characteristics of patients with diffuse large B-cell lymphoma are presented in Box 1. The median age at presentation is 64 years and there is a slight male predominance. Approximately 50 per cent of patients will have stage I or II disease and approximately 50 per cent will have a more widely disseminated lymphoma. Approximately two-thirds of patients will have some sign of extranodal involvement. B-symptoms are seen at presentation in approximately one-third of patients and approximately half of the patients have an elevated LDH. Bone marrow involvement is seen in approximately 15 per cent of patients.

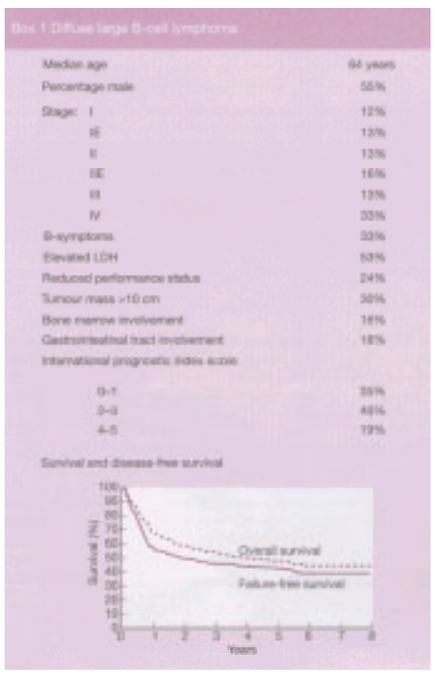

Since the early 1970s it has been known that patients with diffuse large B-cell lymphoma could be cured with combination chemotherapy—even those with disseminated disease. The most popular regimen in use today is **CHOP** (cyclophosphamide, doxorubicin, vincristine (Oncovin), and prednisone), although a number of other regimens including **ACVBP** (doxorubicin, cyclophosphamide, vindesine, bleomycin, and prednisone) are at least as active. A recent randomized trial in older patients found that adding rituximab to CHOP improved both the failure-free and overall survival. When a staging evaluation shows disease confined to one site (that is, stage I) or two nearby sites (minimal stage II) a brief course of chemotherapy followed by radiotherapy gives the highest cure rate. In patients with more disseminated disease, a complete course of one of the combination-chemotherapy regimens mentioned earlier is appropriate. In patients who present with the multiple adverse risk factors listed in the IPI, adjuvant autologous haemopoietic stem-cell transplantation after achieving an initial remission seems to yield a higher cure rate.

Approximately 75 per cent of patients with localized disease can be cured with abbreviated chemotherapy and radiotherapy, and approximately 35 to 40 per cent of patients with more disseminated disease can be cured with combination chemotherapy. Patients who relapse from complete remission can be cured with autologous transplantation. Patients who remain chemotherapy-sensitive after relapse have an approximately 40 per cent cure rate with autologous transplantation, while chemotherapy-resistant patients are cured only approximately 10 per cent of the time.

Follicular lymphoma

The second most common type of non-Hodgkin's lymphoma is follicular lymphoma. The clinical characteristics of patients with this disease are presented in Box 2. The differential diagnosis of follicular lymphoma includes benign follicular hyperplasia and the follicular variant of mantle-cell lymphoma. Patients with follicular lymphoma are subdivided based on the number of large cells in the tumour. In general, a higher proportion of large cells is associated with a higher proliferative rate, more rapid tumour progression, and, perhaps, a better response to anthracycline-containing combination-chemotherapy regimens. The natural history of follicular lymphoma involves a reduction in the degree of follicularity in the tumour over time and an increase in the proportion of large cells. At autopsy, the majority of tumours will be found to have undergone transformation to diffuse large B-cell lymphoma. This is recognized during life in approximately 50 per cent of patients. Histological transformation to diffuse large B-cell lymphoma is associated with a poor prognosis in most patients.

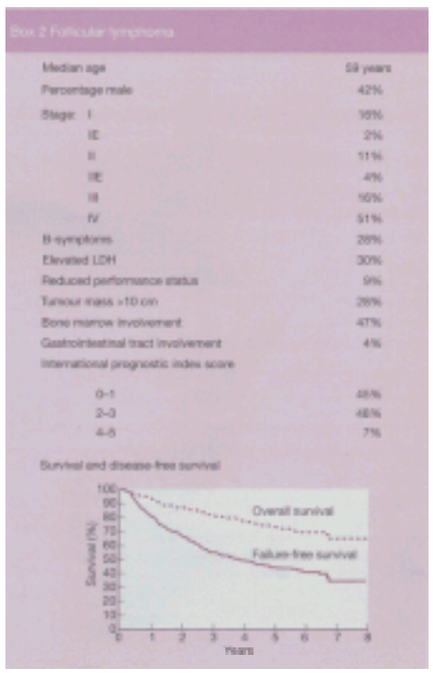

Follicular lymphomas are B-cell lymphomas that display the CD20 antigen. Most patients' tumours will display the translocation t(14;18) and overexpress the *BCL-2* oncogene. Transformation to diffuse large B-cell lymphoma is frequently associated with additional cytogenetic abnormalities and the expression of other oncogenes such as *p53*.

Treatments commonly utilized in the management of patients with follicular lymphoma are presented in Table 13. Asymptomatic patients are often managed with no initial therapy—a strategy that is sometimes called 'watchful waiting'. When followed in this manner, approximately 25 per cent of patients will undergo at least a partial spontaneous regression, although these regressions are not durable. The most popular treatment worldwide for follicular lymphoma is probably single-agent oral chlorambucil. This is particularly true in elderly or infirm patients requiring therapy. Combination chemotherapy with **CVP** (cyclophosphamide, vincristine, and prednisone) or **CHOP** (cyclophosphamide, doxorubicin, vincristine, and prednisone) causes a more rapid response and a higher proportion of complete remissions. Approximately 20 per cent of completely responding patients have remissions that last more than 10 years. Recently, a meta-analysis was conducted of all available trials of interferon incorporated into primary therapy in follicular lymphoma, which demonstrated an improved survival of patients who received adjuvant interferon if they presented with poor risk characteristics and received an anthracycline, such as doxorubicin, as part of their primary therapy. Using **PCR** (polymerase chain reaction) technology, bone marrow specimens from patients with follicular lymphoma frequently test positive for the *BCL-2* gene rearrangement. The goal of some treatment regimens is to eradicate any evidence of *BCL-2* gene rearrangement in the bone marrow, but what the ultimate effect of this will be on survival has not been proven.

Patients with localized follicular lymphoma are often treated with radiotherapy alone. These patients have an excellent outlook with a 10-year survival of 70 to 90 per cent in most series.

Most patients with follicular lymphoma will eventually fail their initial treatment regimen. There is a high response rate to retreatment utilizing single-agent chemotherapeutic agents such as chlorambucil or fludarabine, many combination-chemotherapy regimens, antibodies such as rituximab, and bone marrow transplantation. Both autologous and allogeneic bone marrow transplantation have been demonstrated to produce long-term, disease-free survival in a proportion of patients with follicular lymphoma. Allogeneic transplantation is associated with a higher mortality rate, but there is more convincing evidence that the procedure might be curative.

The median survival in series of patients treated for follicular lymphoma is approximately 10 years, although the median time to relapse is only 3 years. Prolonged survival after relapse is a characteristic of this disease, as are late relapses. This has made study of the treatment of patients with follicular lymphoma difficult and led to controversy of the curability of this illness.

MALT (mucosa-associated lymphoid tissue) lymphoma

This lymphoma, also known as the extranodal marginal-zone B-cell lymphoma of MALT-type, always presents in extranodal sites. A nodal presentation of a similar lymphoma is referred to as nodal marginal-zone lymphoma or monocytoid B-cell lymphoma (see below). Before the recognition of the existence of MALT lymphomas, orbital, pulmonary, and gastric presentations were sometimes referred to by pathologists as pseudolymphoma. The differential diagnosis of MALT lymphoma includes benign lymphocytic infiltration of extranodal organs and other small-cell B-cell lymphomas.

MALT lymphomas are tumours of CD5– and CD23– B cells that express CD20. The commonly seen cytogenetic abnormalities include trisomy 3 and t(11;18). Gastric MALT lymphomas are associated with infection by *Helicobacter pylori*. Thyroid MALT lymphomas are frequently associated with Hashimoto's thyroiditis, and orbital MALT lymphomas are sometimes associated with Sjögren's syndrome. MALT lymphomas can undergo histological transformation to diffuse large B-cell lymphomas. After this transformation, the patient should be treated for diffuse large B-cell lymphoma.

MALT lymphomas have a slight female predominance with a median age at presentation of approximately 60 years. The symptoms of the disorder are those associated with involvement of the extranodal site. The disease is usually localized and the presence of systemic symptoms or elevated LDH is unusual. The characteristics of patients with this lymphoma are presented in Box 3.

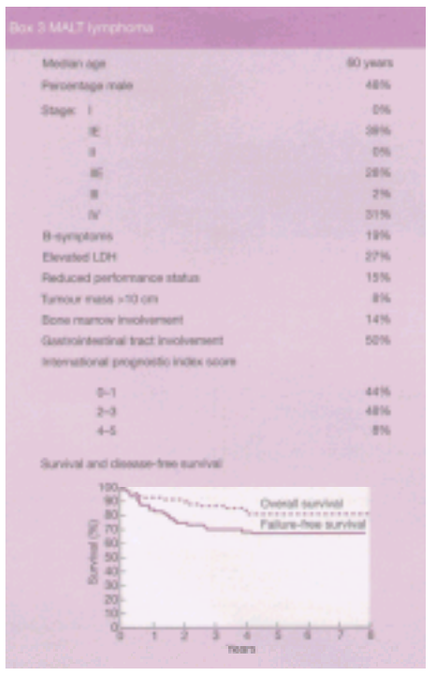

Gastric MALT lymphomas are the first example of a lymphoma that can be treated by eliminating a chronic infection. If the tumour does not transform to a large-cell lymphoma, and has not deeply invaded the stomach, the majority of patients will have their tumour regress with the eradication of *Helicobacter pylori* using antibiotics, proton-pump inhibitors, and bismuth. It appears that in some patients this treatment might be curative. Other local therapies are also effective, and patients with MALT lymphomas can be treated with local radiotherapy or, in some cases, surgery if radiotherapy would be contraindicated. These lymphomas also respond to single-agent chemotherapy or combination chemotherapy. Patients with disseminated MALT lymphoma regularly respond to therapy, but are rarely curable.

The majority of patients with localized MALT lymphoma can be cured, and the 5-year survival in such patients is approximately 90 per cent. However, patients with disseminated disease have a more serious illness and those with a high International Prognostic Index score have a 5-year survival of only 40 per cent.

Small lymphocytic lymphoma/chronic lymphocytic leukaemia

Small lymphocytic lymphoma is the tissue manifestation of chronic lymphocytic leukaemia. Patients who present predominantly with blood and bone marrow involvement will have chronic lymphocytic leukaemia and those who present with lymphadenopathy will have small lymphocytic lymphoma. These are CD5+ B-cell lymphomas. Patients with plasmacytoid differentiation and monoclonal IgM protein in the serum can present the syndrome of Waldenström's macroglobulinaemia. Small lymphocytic lymphoma makes up approximately 7 per cent of non-Hodgkin's lymphomas worldwide, although it is more often seen in western countries.

The differential diagnosis of small lymphocytic lymphoma includes other small B-cell lymphomas. Patients with small lymphocytic lymphoma can undergo histological transformation to diffuse large B-cell lymphoma. This syndrome is seen in approximately 3 per cent of patients and is called Richter's syndrome. It is associated with a poor prognosis.

Chronic lymphocytic leukaemia/small lymphocytic lymphoma is a B-cell neoplasm that typically expresses CD20 and CD23. Approximately 30 per cent of cases have trisomy 12, but there is no specific chromosomal translocation associated with small lymphocytic lymphoma.

The clinical characteristics of patients with small lymphocytic lymphoma are presented in Box 4. Unusual immunological manifestations are sometimes seen, including hypogammaglobulinaemia, autoimmune thrombocytopenia, and autoimmune haemolytic anaemia. When present, these immune abnormalities should be specifically treated, in addition to any treatment given for the lymphoma. Hypogammaglobulinaemia should be treated with intermittent immunoglobulin interfusions and autoimmune cytopenias should be treated with prednisone and/or splenectomy.

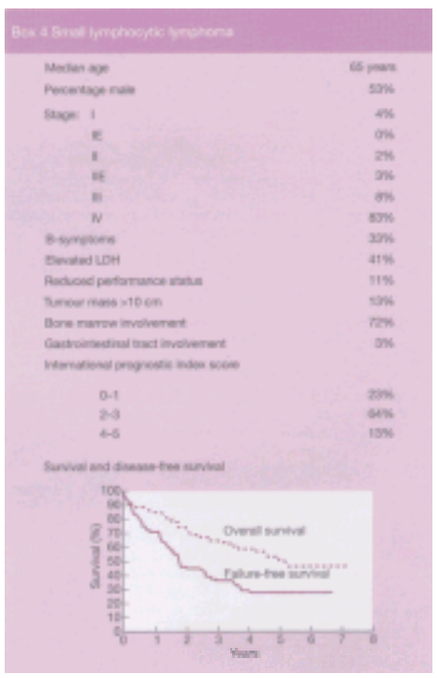

Patients with small lymphocytic lymphoma can be followed with no initial therapy if they are symptomatic, but most patients will require treatment within the first few years. The two most popular treatments for small lymphocytic lymphoma/small lymphocytic leukaemia are single-agent oral chlorambucil and single-agent fludarabine. Fludarabine is associated with a higher complete response rate, but is somewhat more difficult to administer. Neither treatment is curative. Patients frequently respond to further treatment after relapse. Only a small proportion of patients are candidates for bone marrow transplantation. However, occasional patients seem to have a long-term, disease-free survival following allogeneic bone marrow transplantation.

Mantle-cell lymphoma

The clinical characteristics of patients with mantle-cell lymphoma are presented in Box 5. This lymphoma was recognized as a specific entity because of its characteristic cytogenetic abnormality. These tumours regularly express the t(11;14) that involves the *BCL-1* gene on chromosome 11 and the tumour cells overexpress the BCL-1 protein. This can be useful in diagnosis. Before the recognition of mantle-cell lymphoma, patients with this disorder were placed in many other histological categories. In the Kiel classification, mantle-cell lymphoma was usually called centrocytic lymphoma. An expert haematopathologist is important in making the diagnosis, since this lymphoma can be confused with small lymphocytic lymphoma, follicular lymphoma, and lymphoblastic lymphoma.

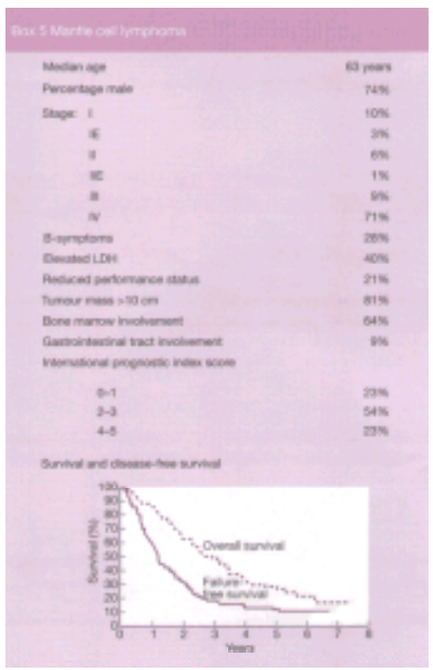

Extranodal sites of involvement by mantle-cell lymphoma are not unusual. Large bowel involvement with mantle-cell lymphoma presents as the syndrome of lymphomatous polyposis. Patients with distal gastrointestinal tract lymphoma often have Waldeyer's ring, and the converse is also true.

Mantle-cell lymphoma responds poorly to available therapies. Combination therapy regimens lead to a complete remission in less than 50 per cent of patients, and most complete responders relapse quickly. The median survival is 3 to 4 years and the 5-year survival for all patients is approximately 25 per cent. Patients who present with a high International Prognostic Index score rarely survive 5 years. Because of the poor outlook, autologous and allogeneic bone marrow transplantation have been increasingly utilized in younger patients.

Less common B-cell lymphomas

Burkitt's lymphoma was originally described by Dennis Burkitt while studying an aggressive lymphoma that occurred in the jaw of children in Central Africa. An association has been demonstrated between infection by the Epstein–Barr virus and this lymphoma. It is much more frequent in children than in adults and in patients infected by HIV. This very rapidly progressive lymphoma is associated with specific chromosomal translocations involving the heavy chain immunoglobulin gene on chromosome 14 or the light chain immunoglobulin genes on chromosomes 2 and 22. In each case, the associated oncogene is the c-*myc* gene on chromosome 8 (namely, t(8;14), t(2;8), and t(8;22)). Burkitt's lymphoma can present as acute leukaemia. This lymphoma can frequently be cured utilizing short courses of very intensive regimens that incorporate high doses of cyclophosphamide.

Nodal marginal-zone lymphoma or monocytoid B-cell lymphoma is immunologically related to MALT lymphoma (see above), but presents in a manner similar to follicular lymphoma. These patients respond to therapy and have an overall survival similar to those with follicular lymphoma. Splenic marginal-zone lymphoma is a rare disorder also known as splenic lymphoma with villous lymphocytes. This rare and indolent lymphoma often responds to splenectomy.

Primary, mediastinal, diffuse large B-cell lymphoma varies from other diffuse large B-cell lymphomas in that it occurs at a younger age and has a striking female predominance. However, the treatment and response to therapy are similar to other diffuse large B-cell lymphomas.

Lymphoplasmacytic lymphoma is a subtype of small lymphocytic lymphoma, which is a tissue manifestation of Waldenström's macroglobulinaemia.

Peripheral T-cell lymphoma

The illnesses classified together as 'peripheral T-cell lymphoma, unspecified type', are a heterogeneous group of non-Hodgkin's lymphomas. The accurate diagnosis of peripheral T-cell lymphoma involves the review of adequate histological material by an expert haematopathologist who has tissue available for immunophenotyping. The diagnosis cannot be made accurately in the absence of immunophenotyping. These tumours are generally CD3- and CD4-positive, although a few will be CD8-positive. Some display CD57 and an NK-cell immunophenotype. Cytogenetic abnormalities are frequent, but there is no consistent genetic abnormality. In occasional cases, demonstrating a T-cell receptor gene rearrangement will help to resolve a difficult diagnostic dilemma and confirm the diagnosis. The differential diagnosis of peripheral T-cell lymphoma includes diffuse large B-cell lymphoma and T-cell hyperplasia as seen in viral infection and drug reactions.

The clinical characteristics of patients with peripheral T-cell lymphoma are presented in Box 6. There are a number of distinctive clinical syndromes grouped together in the category of peripheral T-cell lymphoma. These include the angiocentric nasal NK-cell lymphoma that typically presents with necrotic nasal or facial lesions. Angioimmunoblastic T-cell lymphoma includes most patients who before would have been classed as having angioimmunoblastic lymphadenopathy with dysproteinaemia. These patients present with widespread disease, systemic system disease, skin rash, and polyclonal hypergammaglobulinaemia. Enteropathy-type intestinal T-cell lymphoma is a rare disorder that occurs in patients with gluten-sensitive enteropathy. Patients are frequently wasted and sometimes present with intestinal perforation. Hepatosplenic g,† T-cell lymphoma presents as a systemic illness with sinusoidal infiltration of the liver, spleen, and bone marrow by malignant T cells. Subcutaneous panniculitis-like T-cell lymphoma is a rare disorder that presents with subcutaneous nodules, which are often confused with panniculitis on biopsy. All subtypes of peripheral T-cell lymphoma have a poor prognosis.

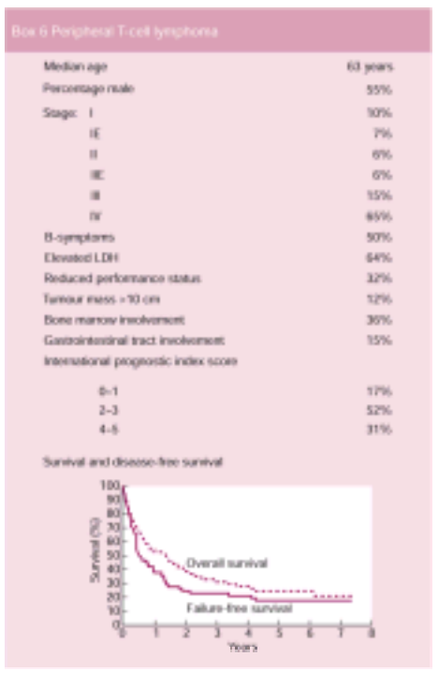

The treatment for patients with peripheral T-cell lymphoma involves the same regimens utilized for diffuse large B-cell lymphoma. Unfortunately, the overall 5-year survival is only 25 per cent and patients with a high International Prognostic Index score have a particularly poor outlook. Bone marrow transplantation can cure some patients who fail primary therapy, and this has been included in the initial treatment of patients with poor prognosis.

Anaplastic large T/null-cell lymphoma

Patients with this lymphoma were previously often diagnosed as having undifferentiated carcinoma, undifferentiated malignant neoplasm, or malignant histiocytosis. Discovery of the Ki-1 antigen (i.e. CD30) led to the recognition that some patients with anaplastic malignancies actually had non-Hodgkin's lymphoma. Subsequent

discovery of the t(2;5) and the resultant overexpression of the ALK protein led to the confirmation of anaplastic large T/null-cell lymphoma as a specific entity. In some patients, the B-cell lymphoma has an anaplastic appearance, but these patients have the same outcome as others with diffuse large B-cell lymphoma.

The clinical characteristics of patients with anaplastic large T/null-cell lymphoma are presented in Box 7. The diagnosis can be made confidently by an expert haematopathologist when facilities for immunophenotyping and staining for the ALK protein are available. The median age of patients with anaplastic T/null-cell lymphoma is approximately 30 years, and 70 per cent of the patients are male. Half the patients have localized (stage I/II) and half have disseminated (stage III/IV) disease. Systemic symptoms are present in 50 per cent of patients and a similar proportion have an elevated LDH level. Occasional patients present with localized disease in the skin and probably have a different and somewhat more indolent disorder.

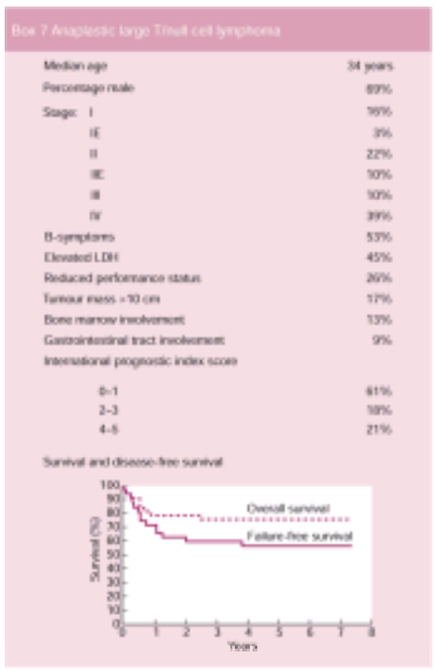

Despite the anaplastic appearance of the lymphoma and frequent poor prognostic characteristics, patients with anaplastic large T/null-cell lymphoma respond well to therapy. The 5-year survival is approximately 75 per cent, and this lymphoma has one of the highest cure rates from combination chemotherapy of any non-Hodgkin's lymphoma. Patients relapsing can respond favourably to bone marrow transplantation.

Mycosis fungoides/Sezary syndrome

Mycosis fungoides or cutaneous T-cell lymphoma is an indolent lymphoma of mature T cells predominantly involving the skin. Patients who present with circulating, atypical (that is, Sezary) cells and erythroderma are said to have Sezary syndrome. The median age is approximately 50 years and the disease is more common in males and Blacks.

Mycosis fungoides often presents with eczematous or dermatitic skin lesions for many years before the diagnosis is firmly established. Frequently, patients will have several biopsies before the diagnosis is confirmed. Lymphoma first manifests itself as superficial lesions in the skin that thicken and eventually ulcerate. In the late stages of the illness, lymphoma can metastasize to lymph nodes and visceral organs.

Treatments utilized for mycosis fungoides include topical corticosteroids, topical nitrogen mustard, phototherapy, **PUVA** (psoralen ultraviolet A-range) therapy, electron-beam radiation, interferon, and systemic cytotoxic therapy. Some patients with localized mycosis fungoides can be cured with radiotherapy. However, the majority of patients will progress. In the end stages of this disease, management is difficult and the ulcerating cutaneous lesions present unpleasant problems for both the patient and the physician. The median survival from diagnosis averages over 10 years.

Adult T-cell lymphoma/leukaemia

The two major manifestations of infection by the human T-cell lymphoma/leukaemia virus-1 (HTLV-1) are tropical spastic paraparesis and adult T-cell lymphoma/leukaemia. Patients can be infected with HTLV-1 through sexual transmission, blood transmission, and transplacentally. The risk of developing lymphoma in a patient infected with HTLV-1 is between 1 and 7 per cent according to various studies. The latency between infection and the development of lymphoma averages approximately 20 years. The diagnosis is established by review of an adequate biopsy by an expert haematopathologist, demonstration of a T-cell immunophenotype, and demonstration of antibodies to HTLV-1. Most patients will have circulating tumour cells with a characteristic pleomorphic histology.

Adult T-cell lymphoma/leukaemia is most frequently seen in the southern islands of Japan and the Caribbean. Most patients seen in Europe and North America will be immigrants from those regions. Blood transfusion provides a possible source for infection, but screening for HTLV-1 has reduced the risk.

The clinical characteristics of patients with adult T-cell lymphoma/leukaemia vary considerably. Some patients present with an indolent disease manifested by lymphadenopathy and skin lesions and survive for extended times without specific therapy. Others present with progressive lymphadenopathy, hepatosplenomegaly, skin infiltration, hypercalcaemia, lytic bone lesions, and elevated LDH levels. Although patients sometimes respond to combination-chemotherapy regimens, complete remissions are unusual and survival is poor.

Lymphoma-like disorders (seeChapter 22.4.2)

Lymphadenopathy caused by infectious mononucleosis, drug reactions to diphenylhydantoin or carbamazepine, autoimmune disorders such as rheumatoid arthritis and lupus erythematosus, and bacterial infections such as cat-scratch disease can all be confused on biopsy with lymphoma. Castleman's disease is a specific condition that can present with localized or disseminated lymphadenopathy and systemic symptoms. The disease appears to be related to an overproduction of interleukin-6. The disseminated form of Castleman's disease is frequently accompanied by anaemia and polyclonal hypergammaglobulinaemia. Patients with localized disease can frequently be treated with local therapy, while systemic disease sometimes responds to systemic glucocorticoids. Sinus histiocytosis with massive lymphadenopathy (Rosai–Dorfman's disease) typically presents with bulky lymphadenopathy in children or young adults. The disease is usually non-progressive and self-limited. Lymphomatoid papulosis is a cutaneous lymphoproliferative disorder that can be confused with T-cell lymphoma in the skin. The cells in lymphomatoid papulosis stain for CD30 and sometimes have a monoclonal T-cell receptor gene rearrangement. The condition is characterized by waxing and waning skin lesions that usually heal leaving small scars. Although these patients have an increased risk of developing lymphoma, aggressive therapy is inappropriate.

Further reading

Armitage JO, Weisenburg ER for the Non-Hodgkin's Lymphoma Classification Project (1998). New approach to classifying non-Hodgkin's lymphomas: clinical features of the major histologic subtypes. *Journal of Clinical Oncology* **16**, 2780–95.

Cheson BD, *et al*. (1999). Report of an. International Workshop to Standardize Response Criteria for non-Hodgkin's lymphomas. *Journal of Clinical Oncology* **17**, 1244–53.

Diehl V, Josting A (2000). Hodgkin's disease. *Cancer Journal* **6**(suppl.2), S150–S158.

Foon KA (2000). Monoclonal antibodies in the treatment of lymphomas for the year 2000. *Principles and Practice of Oncology Updates* **14**, 1.

Godwin JE, Fisher KC (2001). Diffuse large-cell lymphomas; a review of therapy. *Clinical Lymphoma* **2**,155–63.

Jaffe ES *et al*., eds (2001). *World Health Organization classification of tumours, pathology and genetics of tumours of haematopoietic and lymphoid tissues*. IARC Press, Lyon.

Mauch P *et al*., eds (1999). *Hodgkin's disease*. Lippincott Williams & Wilkins, Philadelphia, PA.

Ruzich J, Fisher RJ (2000). MALT lymphoma. *Clinical Oncology Updates* **3**, 1–7.

The International Non-Hodgkin's Lymphoma Prognostic Factors Project (1993). A predictive model for aggressive non-Hodgkin's lymphoma. *New England Journal of Medicine* **329**, 987–94.

Winter JN (1999). High-dose therapy with stem-cell transplantation in the malignant lymphomas. *Oncology*, 1635.

22.4.4 The spleen and its disorders

D. Swirsky

Since Hippocratic times the role of the spleen has been controversial. Galen called it an organ of mystery. Its structure was described during the seventeenth and early eighteenth centuries by Harvey, Glisson, Wharton, Malpighi, and van Leeuwenhoek. In 1777 William Hewson recognized an association with the lymphatic system, and in 1846 Virchow demonstrated that the Malpighian follicles are associated with the formation of white blood cells. In 1885 Ponfick showed that the spleen can remove particles from the blood and might be involved in the destruction of blood cells. Two years later Spencer Wells performed a laparotomy on a 27-year-old woman with a lifelong history of passing dark urine with attacks of jaundice and who had an abdominal tumour thought to be a fibroid. This turned out to be a large spleen and its removal was followed by a complete remission. The retrospective diagnosis of hereditary spherocytosis was made by Lord Dawson of Penn some 40 years later, by which time splenectomy was being performed quite frequently for the treatment of leukaemia, Hodgkin's disease, Banti's haemolytic jaundice, Gaucher's disease, polycythaemia, and thrombocytopenic purpura. The frequent success of the operation led Doan and Dameshek to engage in a lively argument on the mechanisms by which the spleen might destroy blood cells or suppress their formation, a process which Chauffard had earlier called 'hypersplenism'.

There is now a greater understanding of the splenic function in health and of the spleen's involvement in disease. Methods have been developed by which the various functions of the spleen can be defined and measured, sometimes with important clinical application.

Structure of the spleen

At birth the spleen has a mean weight of 11 g. By the age of 1 year the weight is 15 to 25 g; by 5 years it is 40 to 70 g, and by 10 years it is 80 to 100 g. It reaches a maximum weight of 200 to 300 g soon after puberty, and is slightly lighter throughout adult life until the age of about 65 years, when it decreases to 100 to 150 g or less. These figures have been derived from autopsy studies and are probably underestimates. This is mainly due to the splenic red cell pool which will be described later. Ultrasound, computed tomography (**CT**), magnetic resonance imaging (**MRI**) scans and scintigraphic radionuclide scans have shown that, *in vivo*, the normal adult spleen has a length of 8 to 13 cm, a width of 4.5 to 7 cm, a surface area of the order of 45 to 80 cm^2, and a volume less than 275 cm^3. A spleen greater than 14 cm long is usually palpable.

The spleen has a complicated structure (Fig. 1). It consists of a connective tissue framework, vascular channels, lymphatic tissue, lymph drainage channels, and cellular components of the haemopoietic and mononuclear phagocyte systems. There are two main components: (1) the red pulp; and (2) the white pulp. The red pulp consists of sinuses and pulp cords. The sinuses, 20 to 40 µm in diameter, are lined by endothelial macrophages. The white pulp consists of a periarteriolar lymphoid sheath and the adjoining lymphoid follicles (Malpighian bodies), which contain a germinal centre and are structurally similar to lymphoid follicles. From the capsule many lace-like trabeculae extend into the pulp, carrying blood vessels and autonomic nerve fibres. Within the spleen the trabeculae are in direct continuity with a mesh of reticular fibres that supports the pulp vessels and forms the basement membranes of arterial capillaries and the splenic sinuses. Along the reticular fibres lie adventitial reticular cells. These cells have an important role in regulating blood flow through the interendothelial slits of the vascular sinuses.

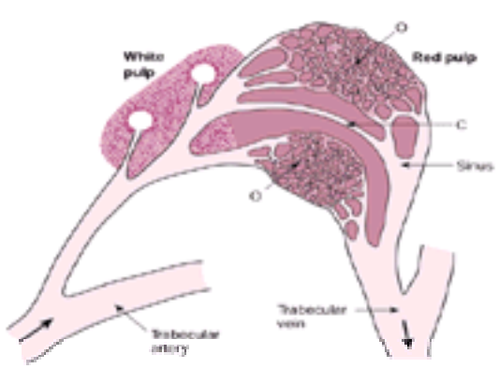

Fig. 1 Diagrammatic illustration of the spleen. The blood passes either directly into a sinus (C, closed system) or first into the cord space in the red pulp (O, open system).

Blood is supplied by the splenic artery and passes through the trabecular arteries, into the central arteries, which are sited in the white pulp. The central arteries run into the central axis of the periarteriolar lymphatic sheaths; they give off many arterioles and capillaries, some of which terminate in the white pulp whilst others go on to the red pulp. There they either connect directly with the sinuses and thence, via the collecting vein, to the trabecular vein (closed system), or they first pass into the cord spaces before joining up with the sinuses (open system).

Thus, as blood flows through the spleen it will come into contact with the reticular fibres, and also with endothelial macrophages, which line the interstices of the reticular mesh.

Blood flow in the spleen

Because the spleen has two vascular systems (open and closed) as described above, there are both rapid- and slow-transit components in the splenic circulation. The rapid transit (closed system) is of the order of 1 to 2 min and the slow transit (open system) about 30 to 60 min. In normal subjects the open system has a minor role and the blood flows through the spleen as rapidly as through organs possessing a conventional vasculature, at a rate of 5 to 10 per cent of the blood volume per

minute, so that each day the circulating blood has repeated passages through the spleen. When the spleen is enlarged, blood flow increases up to 20 per cent or more of the blood volume per minute. At the same time, a proportion of the blood may be pooled in the cord spaces (see below). As blood traverses the spleen, the plasma and leucocytes pass preferentially to the white pulp by a process of plasma skimming, and the plasma rapidly reaches the venous system, whilst blood with a relatively high packed-cell volume remains in the axial stream of the central artery. Some of this blood flows directly through the sinusoids to the venous system, while the remainder passes into the cords of the red pulp. The normally flexible red cells squeeze through the endothelial slits into the sinuses, whilst inflexible cells with fixed membranes or with inclusions remain in the cords where they either become conditioned for later transit or are destroyed.

Functions of the spleen

Haemopoiesis

In the fetus and compared with the liver, the spleen is a minor haemopoietic organ. There is, however, some erythropoiesis and granulopoiesis in the spleen from the 12th week of gestation; this continues until birth, after which there is normally no demonstrable haemopoiesis. However, the potential remains, and under severe haematological stress, in thalassaemia and in chronic haemolytic anaemias for example, extramedullary haemopoiesis may occur together with intense erythroid hyperplasia of the bone marrow. This must be distinguished from myeloid metaplasia occurring in myelofibrosis, chronic myeloid leukaemia, and occasionally other leukaemias and secondary carcinomas. In these conditions, foci of haemopoietic tissue become established in the spleen and elsewhere outside the bone marrow. They represent an abnormal proliferation which is distinct from compensatory haemopoiesis.

Cell sequestration, phagocytosis, and pooling

The spleen has a remarkable ability to 'cleanse' or 'recondition' red cells for recirculation and also to remove from the circulation effete or damaged cells as well as foreign matter. Of particular importance is the trapping of encapsulated bloodborne bacteria. It is important to distinguish between the three mechanisms involved. Sequestration is a temporary (reversible) process whereby cells are held in the spleen before returning to the circulation; phagocytosis represents the irreversible uptake of non-viable cells by macrophages, or the destruction of viable cells that have been damaged; pooling is the presence in the spleen of an increased amount of blood (or some of its component parts). In contrast to sequestration, pooled cells are in continuous exchange with the circulation.

As blood flows through the sinuses and cords, effete and damaged cells, and particulate foreign matter, are promptly phagocytosed by the endothelial macrophages. Intact red cells are held up temporarily, during which time siderotic granules, Howell–Jolly bodies, and Heinz bodies are removed. After the inclusions have been removed the red cells return to the circulation. Sequestration of reticulocytes also occurs, and they are retained in the splenic cords for part of their last 2 or 3 days of maturation while they lose their intracellular inclusions, alter their surface membrane composition, and become smaller. The spleen normally sequesters 30 to 45 per cent of the total circulating platelet content of the blood. This platelet pool is rapidly mobilized under conditions of stress, and normally there is a constant transit between the spleen and vascular pools.

As the blood becomes more viscous in the spleen, red cells are subjected to a further hazard. Because they are packed together in the presence of metabolically active macrophages, they are depleted of glucose and oxygen. This increases their membrane rigidity and reduces their deformability. Cells may become inflexible if: (1) they are metabolically abnormal (as in some congenital haemolytic anaemias) and thus unduly sensitive to the unfavourable environment of the spleen; (2) if they are held up in the spleen for a prolonged period and are thus rendered metabolically abnormal; and (3) if they are already spherical (as in hereditary spherocytosis), fragmented (as in microangiopathic haemolytic anaemia), or misshapen in some other way. This results in their being trapped in the cord spaces where they subsequently undergo phagocytosis.

Immune function

The spleen contains the largest single accumulation of lymphoid tissue in the body; about 25 per cent of the total T-lymphocyte pool and 10 to 15 per cent of the B-lymphocyte pool, with very marked exchange between circulating and splenic lymphocytes. Splenic macrophages are instrumental in antigen presentation to lymphocytes. The spleen is a major, but not unique, site for the conversion of naive circulating B cells into plasma cells which migrate to the bone marrow and into long-lived memory cells.

Micro-organisms or other antigens that find their way to the spleen are taken up and processed by cord macrophages and are presented to immunocompetent cells in the lymphoid tissue. This stimulates antibody production and an increase in size of the lymphoid germinal centres of the spleen. Secondary stimulation with the antigen enhances antibody production, usually IgG.

Antibody-coated red cells lose pieces of their membrane as they come in contact with the Fc receptors on macrophages, and become spherical and less flexible each time they pass through the sinus vasculature, until finally they become too rigid to traverse the endothelial pores and are trapped and destroyed. Red cells sensitized by IgG do not, as a rule, agglutinate in the peripheral blood, but the environment in the spleen promotes local agglutination with consequent sequestration and destruction (autoimmune haemolysis). Antibody-coated neutrophils and platelets are similarly destroyed by splenic macrophages.

Blood pool

The normal red cell content of the spleen is less than 80 ml of red cells, and always less than 5 per cent of the total red cell mass. There is no significant red cell pool in human spleens. However, enlarged spleens are capable of developing remarkably large pools with a relatively slow exchange of red cells with the general circulation. In the myeloproliferative disorders, as much as 40 per cent of the blood volume may be present in the spleen. Increased pools also occur in lymphoproliferative disorders, especially hairy-cell leukaemia and prolymphocytic leukaemia.

In health there is a good correlation between the amount of blood in the spleen and its size. In lymphomas, however, the splenomegaly is greater than can be accounted for by the pool alone; in such cases the increase in spleen size is due primarily to an expansion of the lymphoid components with replacement of splenic sinuses by tumour. In myelofibrosis there is an increase in the reticular element with expansion of the closed system in the red pulp. A similar effect occurs in hairy-cell leukaemia.

Not unexpectedly, the red cell content of the spleen increases with increasing body haematocrit. There is a disproportionately increased pool in primary proliferative polycythaemia compared with secondary polycythaemia, where the pool remains small irrespective of the haematocrit level. Increased pools are also found in patients with hepatic cirrhosis. Here it is the increased portal pressure that leads to an increased splenic blood flow: the splenic arteries are dilated and the splenic pulp becomes expanded with prominent dilated sinuses. Portal hypertension may result from myeloproliferative disorders associated with splenomegaly.

In myeloproliferative disorders and some other conditions an enlarged splenic blood pool may contribute significantly to anaemia. A low venous haematocrit can be present despite a normal red cell mass (pseudoanaemia). Direct measurement of the splenic red cell volume makes it possible to predict the extent to which splenectomy will improve anaemia and reduce transfusion requirements.

There is also a significant reservoir of platelets in the spleen, which is rapidly interchangeable with the circulation. In some cases of thrombocytopenia, destruction occurs mainly in the spleen and it is essential to distinguish this from pooling. As far as granulocytes are concerned, no pool is demonstrable in the normal spleen, but an abnormally large marginal pool has been found in cases of splenomegaly associated with neutropenia.

Plasma volume

Splenomegaly is frequently associated with an increased plasma volume, which may lead to an apparent anaemia (pseudoanaemia or dilutional anaemia), when a reduced venous haematocrit is the result of an expanded plasma volume in the presence of a normal or slightly reduced red cell mass.

Splenomegaly

A palpable spleen is usually enlarged. Occasionally a normal spleen is palpable if it is displaced downwards, by a pleural effusion for example. The spleen has to be

1.5 to 2 times its normal size to be palpable. Ultrasound, CT, and MRI provide reliable methods for measuring the actual spleen size.

Investigation of splenomegaly

The clinical history should include a relevant travel history (for example, to tropical areas) and family history (for example, Gaucher's disease, hereditary spherocytosis). Physical examination should specifically include assessment for hepatomegaly and lymphadenopathy. Laboratory investigations should include a full blood count, liver function tests and hepatitis serology, serum protein electrophoresis, total cholesterol, triglyceride and lipoprotein determinations, as well as immunoglobulin measurements. A bone marrow aspirate and trephine biopsy, and/or lymph node biopsy, with appropriate cytogenetic and immunophenotyping studies, should be done as indicated. These investigations will reveal the diagnosis in most haematological disorders and many chronic infections. CT or ultrasound scanning, with liver biopsy as required, will reveal hepatic or thrombotic causes of portal hypertension. HIV serology should always be done in puzzling cases of splenomegaly; in patients of Ashkenazi Jewish ancestry, Gaucher's disease or Niemann–Pick disease may be reasonably excluded by determining lysosomal hydrolase activity in leucocytes. If all investigations are negative, diagnostic splenectomy may be necessary, and in non-tropical areas the diagnosis will usually be non-Hodgkin's lymphoma or Hodgkin's disease.

Causes of splenomegaly

So many conditions are associated with splenomegaly that it is impossible to give a comprehensive list. It is even more difficult to list the 'common' causes as these depend on geographical pathology. In Western Europe and the United States viral infections and portal hypertension are the most common causes of splenomegaly, and these together with leukaemias, malignant lymphomas, myeloproliferative disorders, haemolytic anaemias, and other infections account for most cases. Isolated splenomegaly is a common manifestation of type I Gaucher's disease. Globally, however, the incidence of these haematological causes of splenomegaly is swamped by the great preponderance of splenic enlargement caused by parasitic infections, particularly malaria, leishmaniasis, and schistosomiasis. Human immunodeficiency virus (**HIV**) infection, particularly in the later stages of the disease, is an increasing cause of mild to moderate splenomegaly. Haemoglobinopathies head the list in some countries. Portal hypertension is an important cause of splenomegaly in most tropical countries but it is especially prevalent in north-eastern India and southern China. The 'tropical splenomegaly syndrome' associated with malaria is seen commonly in New Guinea and Central Africa.

Some of the causes of splenomegaly are listed in Table 1. The conditions which commonly give rise to massive splenomegaly are marked with an asterisk. The spleen sizes indicated are only a rough guide. Most of the conditions listed are described in other chapters.

Hypersplenism

Hypersplenism is a clinical syndrome of varied aetiology. It is characterized by:

1. Splenomegaly, although this may only be moderate.
2. Cytopenias: pancytopenia, single cytopenias, or any combination of anaemia, neutropenia, and thrombocytopenia.
3. A cellular or hypercellular bone marrow, sometimes showing a paucity of mature granulocytes.
4. A premature release of cells into the peripheral blood, resulting in a mild reticulocytosis with nucleated red cells and occasional immature granulocytes.

Other features are:

1. decreased red cell survival;
2. decreased platelet survival;
3. hypervolaemia (that is, increased plasma volume) if splenomegaly is marked.

The haematological features may be obscured or dominated by the primary disease, especially if it involves the marrow. The diagnosis of hypersplenism is ultimately confirmed by the response to treatment of the underlying cause or of splenectomy, although an immediate remission may be followed in the longer term by relapse with a return of cytopenia.

Tropical splenomegaly syndrome 'big spleen disease'

In areas where malaria is endemic, adults may present with moderate to massive splenomegaly, no obvious signs of active malaria, but all the features of hypersplenism including pancytopenia, expanded plasma volume, and haemolysis. The serum IgM level is usually high, and malarial antibody titres are raised. The spleen shows diffuse proliferation of macrophages. The relationship to malaria is evident by the response to long-term antimalarial treatment, which produces a sustained reduction in spleen size and reversal of the cytopenias. It is unclear why this effect is only seen in a proportion of individuals in areas of the world where malaria is endemic.

A similar degree of splenomegaly occurs in schistosomiasis (see Chapter 7.16.1). However, in this condition there is the further complication that the eggs (especially of *Schistosoma mansoni*) have a direct effect on the liver, resulting in hepatic fibrosis and leading to portal hypertension, which may be further exacerbated by splenic vein thrombosis.

Non-tropical idiopathic splenomegaly

Rare patients present with marked splenomegaly and the haematological features of hypersplenism but without exposure to malaria or other parasitic disorders. There may be a positive antiglobulin test and other evidence of autoantibody production. Some of these patients have a malignant lymphoma at the time of presentation, but in others the essential feature is non-neoplastic lymphoid hyperplasia, which probably represents an immunological reaction to as yet unidentified stimuli. The chances of long-term cure after splenectomy appear to be good. However, a lymphoma may appear from months to years after splenectomy. The disorder is diagnosed by the finding of massive splenomegaly in the absence of any other cause and by the non-specific histological appearances in the spleen.

Storage disease

The storage diseases are described in detail in Section 11.7. Some of them, notably Gaucher's disease and Niemann–Pick disease, may be complicated by marked splenomegaly. This may lead to hypersplenism, particularly in Gaucher's disease. The advent of specific enzyme therapy for Gaucher's disease has largely removed the need to consider splenectomy. The clinical picture of Niemann–Pick disease is dominated by hepatosplenomegaly and mental retardation. The disorder presents in infancy, and death often occurs between the second and third years of age, but, as with Gaucher's disease, it may present later in life. Hypersplenism becomes a feature in the older age groups, but anaemia and thrombocytopenia are uncommon in the childhood cases, and, if present, are mild. Several other lipid storage diseases may cause hypersplenism. They include Tangier's disease, in which cholesterol esters fill the histiocytes, and Wolman's disease, which is associated with an accumulation of triglycerides and cholesterol esters. Sea-blue histiocytosis is characterized by splenomegaly, hepatomegaly, thrombocytopenia, and, occasionally, neurological damage. The bone marrow and spleen contain cells that have an accumulation of glycosphingolipids, phospholipids, and mucopolysaccharides.

Rarely, Histiocytosis X (including Hand–Schuller–Christian disease, eosinophilic granuloma, Letterer–Siwe disease, and Langerhans' cell histiocytosis) cause splenomegaly. This is usually moderate, but occasionally it is more marked and may be associated with hypersplenism.

Space-occupying lesions and injury of the spleen

The most common causes of splenic masses are trauma leading to haematoma or rupture, abscesses, tumours, and cysts.

Splenic injury

The spleen is relatively unprotected and easily injured. Spontaneous rupture has been reported in a number of conditions in which the spleen is enlarged: these include typhoid, malaria, Epstein–Barr virus infection, leukaemia, Gaucher's disease, and polycythaemia. This may be restricted to a subcapsular haematoma or there

may be rupture into the peritoneal cavity.

The diagnosis is suggested by the symptoms of shock, left upper quadrant tenderness, guarding, pain referred to the left shoulder, and clinical and laboratory evidence of bleeding. Plain abdominal radiography is not, as a rule, helpful in diagnosis but CT scanning, ultrasound examination, and splenic arteriography are more useful.

Abscess

Although the spleen is frequently enlarged in association with systemic infection, splenic abscesses are rare. They result from direct or haematogenous spread, or when a haematoma becomes infected. Conditions associated with splenic infarction, such as sickle-cell disease, are particularly likely to give rise to splenic abscesses. Almost any organism can be involved.

Tumours

The spleen may be affected by benign tumours such as hamartomas. The very rare littoral-cell angioma is the only tumour confined to the spleen. Metastases in the spleen are uncommon by comparison to other organs, possibly because the spleen, unlike lymph nodes, lacks an afferent lymphatic system. They occur late in the course of carcinoma and are not found in the absence of metastases elsewhere. Metastases in the spleen are most frequently derived from malignant lymphomas, especially Hodgkin's disease. Lung, breast, prostate, colon, and stomach are the primary sites from which carcinoma is most likely to disseminate to the spleen. Melanoma is also a relatively frequent primary source.

Cysts

Splenic cysts are rare. The most frequent cause is *Echinococcus granulosum* (hydatid); other causes include haemangiomas, lymphangiomas, and dermoids. Cysts may also develop in areas of haemorrhage or infarction.

Loss of spleen function and splenic infarction

Splenic hypoplasia or atrophy

Congenital hypoplasia is rare; in some cases it is associated with extensive developmental abnormalities of the heart and gut. Splenic atrophy may occur in a number of conditions—sickle-cell disease, coeliac disease, dermatitis herpetiformis, ulcerative colitis, Crohn's disease, amyloidosis, selective IgA deficiency, and Fanconi's anaemia. There is evidence of reduced splenic reticuloendothelial function in alcoholics. The spleen shrinks in size in old age. Vascular blockade and repeated infarction is the basis for splenic atrophy in sickle-cell disease, and occurs in early childhood. The peripheral blood changes of hyposplenism, when present, are proportional to disease activity in gut diseases. In coeliac disease, withdrawal of gluten from the diet reverses the changes unless splenic atrophy has occurred. The mechanism of the splenic atrophy is unknown.

Splenic hypofunction and atrophy are characterized by changes in the blood film appearances; the main features are the presence of Howell–Jolly bodies and siderotic granules in some of the red cells. This is due to the loss of the spleen's macrophage 'pitting' function. Reduced sequestration (pooling) of red cells also occurs.

Splenic infarction

Splenic infarction occurs quite frequently in patients who have very large spleens from any cause. It is particularly common in association with myelosclerosis and chronic myeloid leukaemia. It also occurs in most patients with sickle-cell anaemia. In this disorder, splenic infarction occurs early in life and repeated episodes result in an autosplenectomy. Occasionally, when there is rapid growth of the spleen in association with an aggressive form of non-Hodgkin's lymphoma there may be multiple infarctions and spontaneous rupture of the spleen. Splenic infarcts are one of the presenting features of chronic myeloid leukaemia.

Splenic infarction causes pain in the left upper quadrant. If the diaphragmatic surface of the spleen is involved, the pain may be referred to the left shoulder tip. The physical signs include tenderness over the spleen, and sometimes a loud splenic rub is heard. Treatment is by rest and analgesia. The occurrence of repeated splenic infarction may be an indication for splenectomy, which may be complicated by adhesions between the spleen and the overlying peritoneum.

Specialized investigation of splenic function

Assessment of splenic function may be helpful, particularly in assessing the likely effect of splenectomy in haematological disorders. In many conditions it is sufficient to assess the spleen size, examine the peripheral blood for evidence of pancytopenia or a reduction in the number of neutrophils and platelets, and to examine the bone marrow to determine whether haemopoiesis is normal. Often this simple approach, combined with a knowledge of the likely effects of splenectomy for a particular haematological disorder, will be all that is necessary to make a decision about whether to proceed to surgery.

Studies with radionuclides provide information about the extent of splenic involvement in a disease process, the role of the spleen in producing anaemia, and the likely benefits of splenectomy. Details of the methods used and analysis of the results obtained in various conditions are to be found in specialized textbooks.

The following list summarizes the various *in vivo* tests useful for investigating splenic function:

- *Delineation of functional splenic tissue.* The spleen can be visualized and its size estimated by scintillation scanning following injection of isotope-labelled, autologous, heat-damaged red cells, which are selectively removed by functional splenic tissue. A gamma camera or rectilinear scanner visualizes splenic tissue (Fig. 2). The technique is most useful for identifying accessory spleens (splenunculi) associated with a postsplenectomy relapse of immune thrombocytopenia. The rate at which heat-damaged red cells are cleared from the circulation provides a rough guide to the competence of splenic function. A slow clearance may identify splenic hypofunction before the blood film shows Howell–Jolly bodies and other morphological changes.
- *Measurement of splenic red cell pool*. Quantitative scanning of the spleen after injection of undamaged, isotope-labelled, autologous red cells allows measurement of the splenic red cell pool. The size of the splenic red cell pool should be taken into account when assessing the significance of anaemia in the presence of splenomegaly. Measuring the pool is particularly useful for distinguishing polycythaemia vera (increased pool) from secondary polycythaemia (normal pool) and assessing the (useless) spleen pool in massive splenomegaly (Fig. 3).
- *Identification of sites of red cell destruction and quantification of splenic red cell destruction*. Surface counting over the spleen, heart, and liver following injection of autologous ^{51}Cr-labelled erythrocytes provides a qualitative indication of splenic red cell destruction in various haemolytic anaemias; quantitative scanning provides a more accurate measurement of the actual proportion of the cells that are destroyed in the spleen and elsewhere. These studies are moderately predictive of the outcome of splenectomy.
- *Identification and quantification of splenic extramedullary erythropoiesis*. Normally, transferrin-bound iron passes to the bone marrow, where the iron is released and enters erythroblasts for incorporation into the haemoglobin of developing erythrocytes. In the normal spleen, iron does not dissociate from transferrin. Hence, the uptake of iron demonstrable by surface counts shortly after administration of radioactive iron (^{59}Fe or ^{52}Fe), indicates that there is erythropoiesis in the spleen. Extramedullary erythropoiesis in the spleen occurs in the majority of patients with myelofibrosis (Fig. 4), and some patients with essential thrombocythaemia, but not in patients with polycythaemia vera. ^{52}Fe studies are useful for detecting early stages of transition from polycythaemia vera to myelofibrosis and for diagnosing the syndrome of transitional myeloproliferative disorder. Extramedullary haemopoiesis can be accurately identified in thalassaemia major or intermedia and sickle-cell disease by positron emission tomography (**PET**) after ^{52}Fe administration, for example paraspinal, mediastinal, or in lymph nodes.
- *Role of the spleen in platelet destruction*. About one-third of an injection of ^{51}Cr-labelled platelets disappears from the circulation during their lifespan, mainly in the spleen pool. Splenomegaly is associated with a marked increase in pooling: by contrast, in asplenia, nearly 100 per cent of the labelled platelets are recovered in the circulating blood. Surface counting and quantitative scanning have been used to assess the role of the spleen in thrombocytopenia, but are less reliable in predicting the response to splenectomy than in autoimmune haemolytic anaemia.

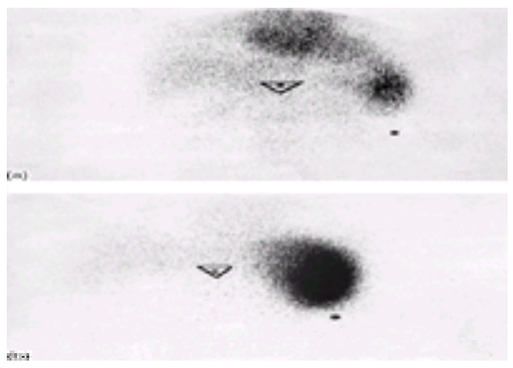

Fig. 2 Images obtained by scintillation camera following administration of (a) ^{99}Tc-labelled red cells and (b) ^{111}In-labelled, heat-damaged red cells.

Fig. 3 Splenic enlargement and increased red cell pool in a patient with myelofibrosis. Demonstrated by scanning after the administration of $^{113}In^{m}$-labelled red cells. The markings indicate the costal margin. The upper pole of the spleen merges with the image produced by labelled blood in the heart.

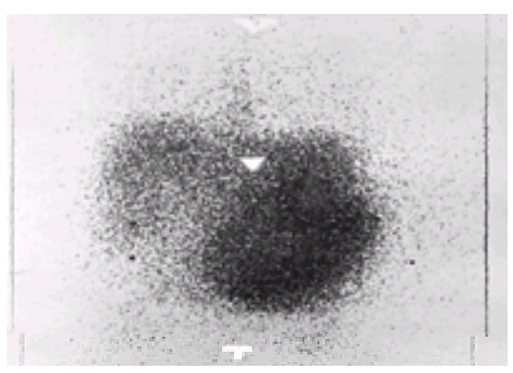

Fig. 4 ^{52}Fe scan in a patient with myelofibrosis showing the extent of splenic erythropoiesis. Vertebral erythropoiesis is markedly reduced.

The combinations of investigations used depends on the particular clinical problem. In many conditions associated with splenomegaly, it is important to distinguish increased macrophage activity causing cell destruction from increased red cell accumulation in a large pool, and to determine to what extent enlargement of the spleen is due to tumour infiltration. In myelofibrosis and hypersplenism, it may be helpful to ascertain the relative importance of the splenic red cell pool, red cell destruction, and extramedullary erythropoiesis, if present.

Indications for splenectomy

The main indications for splenectomy are summarized in (Table 2). Splenectomy should not be undertaken lightly. Where traumatic damage, usually from a blunt injury, has occurred, every effort should be made to preserve the spleen in whole or in part. Ultrasound and CT imaging are vital in assessing the damage as rupture and haematoma can be confused clinically. Surgical techniques for repairing or partially preserving the spleen have improved and should be encouraged. The use of diagnostic laparotomy and splenectomy in Hodgkin's disease has fallen into disuse, partly as a result of improved imaging in the CT and MRI scans, and partly because of the absence of therapeutic advantage, combined with the long-term dangers of splenectomy. In primary haematological disorders, splenectomy is indicated to alleviate complications—repeated infarction, massive red cell pooling, and hypersplenism for example. The decision is usually based on clinical assessment, but radioisotope studies to measure the splenic red cell pool, splenic erythropoiesis, and red cell survival may be helpful in difficult cases. Diagnostic splenectomy may still be required for splenic lymphomas where no other organ is affected. Increasingly, diagnostic splenectomy may be required in HIV-related diseases, particularly for a suspected lymphoma or opportunistic infection. The course of HIV disease is not influenced by splenectomy.

Clinical and haematological effects of splenectomy

Removal of the spleen is associated with certain immediate and delayed clinical complications, and with the presence of permanent changes in the eripheral blood picture.

Clinical complications

Early

In some splenectomies, particularly when the spleen is bound down by adhesions following a previous splenic infarction, there may be difficulty in achieving haemostasis, particularly if the preoperative platelet count is less than 50×10^{9}/l. Platelets should be infused as soon as the splenic pedicle has been ligated. Subphrenic abscess is a significant complication and may occasionally be fatal. Because the platelet count tends to rise immediately after the operation, there is an increased risk of thromboembolic disease in the first 2 or 3 weeks after splenectomy.

Subcutaneous heparin can be used if the preoperative platelet count is 50×10^{9}/l or greater. After wound healing, aspirin 75 mg daily should be given if the platelet count is elevated above 500×10^{9}/l, and continued until this normalizes. There is a small, long-term increase in myocardial infarction in splenectomized patients with persistently raised platelet counts, and in these patients aspirin should be given indefinitely.

Mortality depends on the clinical condition of the patient and the size of the spleen. Laparoscopic techniques for splenectomy are safer and reduce morbidity for small or moderately enlarged spleens. In patients with massive splenomegaly, usually myeloproliferative disorders, the mortality rate is up to 15 per cent in those patients thought fit for surgery.

Long term

All patients, whatever the reason for splenectomy, are at risk of overwhelming postsplenectomy infection (**OPSI**). Classically, OPSI presents with a vague general prodrome, followed by prostration, bacteraemic shock, and frequently disseminated intravascular coagulation. Death may occur within 6 h of the onset. The mortality rate in patients reaching hospital alive is in excess of 30 per cent. By far the most important causative organism is the pneumococcus (*Streptococcus pneumoniae*), but *Haemophilus influenzae*, *Neisseria meningitidis*, *Escherichia coli*, and *Pseudomonas* spp. have all been implicated. In endemic areas, plasmodium and babesia infections are of increased severity in non-immune individuals. Special warnings should be given to splenectomized patients travelling to malarial areas. Viral illnesses may also be of increased severity postsplenectomy.

The risk of OPSI does not decline significantly in the years after splenectomy. Children are at the greatest risk, followed by adults splenectomized for an underlying disorder that itself is immunosuppressive, or who require immunosuppressive treatment. Adults splenectomized for trauma are at least risk, but they still carry a lifelong susceptibility. OPSI has been recorded more than 40 years after splenectomy. The relative risk of severe infection compared with the non-splenectomized population is about 10-fold for traumatic splenectomy and as much as 100-fold for small children and patients with Hodgkin's disease.

Infections indistinguishable from OPSI also occur in non-splenectomized individuals who have hypofunctional spleens. It is well recognized as a cause of death in patients sickle-cell disease, particularly in children. Fatal overwhelming pneumococcal sepsis has been reported in patients with coeliac disease and primary amyloidosis affecting the spleen. Dermatitis herpetiformis and inflammatory bowel disease are also associated with splenic hypofunction. Bone marrow transplant recipients, particularly in the presence of chronic graft-versus-host disease are hyposplenic and have an increased risk of pneumococcal disease. Patients with lymphoproliferative disorders, particularly myeloma, are at increased risk of sepsis with encapsulated bacteria and should be considered for prophylaxis.

Strategies for preventing OPSI

All patients undergoing elective splenectomy should be immunized with polyvalent pneumococcal vaccine ('Pneumovax'), which currently gives variable protection against 23 strains of *S. pneumoniae*. Where possible, it should be given at least 1 month prior to splenectomy to allow IgG antibody production. Antibody responses may be suboptimal in patients with immunosuppressive diseases or in those receiving immunosuppressive treatment, or when it the vaccine is given perioperatively in an emergency. Patients should also be immunized against *H. influenzae* type b. Protection against *N. menigitidis* is of relatively short duration, and vaccination should be reserved for patients travelling to high incidence areas. Pneumovax is not fully protective, and a small proportion of patients fail to make detectable antibodies after vaccination. There are reports of OPSI occurring with strains of *S. pneumoniae* covered by the type of vaccine given, and therefore it should always be combined with life-long prophylactic antibiotics. Revaccination with Pneumovax every 5 to 10 years is recommended. Splenectomized individuals should always carry a card or wear a bracelet stating they have no spleen. At the onset of any febrile illness, particularly upper and lower respiratory infections, penicillin V should be stopped and therapeutic doses of a broad-spectrum antibiotic started. The penicillin V is resumed at the end of the course of antibiotics.

Prophylactic penicillin V, 250 mg twice daily, should be started postoperatively. Erythromycin can be substituted in penicillin-sensitive patients. The lifesaving value of prophylactic penicillin V in children with sickle-cell disease (that is to say, functional asplenia) has been proven beyond doubt, and there are only rare reports of OPSI in splenectomized patients regularly taking penicillin V. Surveys of patients dying of OPSI have identified the failure to follow the above guidelines as the greatest risk factor. While the penicillin does not prevent infection, it prevents the rapid onset of the OPSI syndrome. There is controversy as to the effectiveness of penicillin V in areas where resistant strains of pneumococci are common.

Following spleen rupture, splenic tissue may seed into the peritoneum, giving rise to nodules of recognizable splenic tissue (splenosis). These nodules have been shown to have some phagocytic function. This has led to the deliberate autotransplantation of splenic tissue at the time of splenectomy where partial splenic preservation has not been possible. Although such nodules of splenic tissue can phagocytose damaged red cells and reduce hyposplenic changes on the blood film, their protective capacity from infection is not established. The presence of demonstrable splenosis should not be relied upon to replace vaccination and penicillin prophylaxis.

The safest course is to immunize all patients, counsel them carefully about the dangers of infection, and impress upon them the need for lifelong penicillin prophylaxis. This should be reinforced at outpatient follow-up visits, and every effort should be made to maintain good compliance.

Further reading

Anon (1996). Guidelines for the prevention and treatment of infection in patients with an absent or dysfunctional spleen. *British Medical Journal*, **312,** 430–4.

Berman RS, *et al.* (1999). Laparoscopic splenectomy in patients with hematologic malignancies. *American Journal of Surgery,* **178,** 530–6.

Bowdler AJ (ed.) (1990) *The spleen. Structure, function and clinical significance*. Chapman and Hall Medical, London.

Crane CG (1981). Tropical splenomegaly. Part 2: Oceanian. *Clinics in Haematology,* **10,** 976–82.

Dacie JV, Lewis SM (1995). *Practical haematology,* 8th edn. Churchill Livingstone, Edinburgh.

Fakunle YM (1981). Tropical splenomegaly. Part 1: Tropical Africa. *Clinics in Haematology,* **10,** 963–75.

Frank JM, Palomino NJ (1987). Primary amyloidosis with diffuse splenic infiltration presenting as fulminant pneumococcal sepsis. *American Journal of Clinical Pathology,* **87,** 405–7.

Gaston M, *et al.* (1986). Prophylaxis with oral penicillin in children with sickle cell anemia. *New England Journal of Medicine,* **314,** 1593–9.

Lucas CE (1991). Splenic trauma. Choice of management. *Annals of Surgery*, **213,** 98–112.

O'Donoghue DJ (1986). Fatal pneumococcal septicaemia in coeliac disease. *Postgraduate Medical Journal,* **62,** 229–30.

Oksenhendler E, *et al.* (1993). Splenectomy is safe and effective in human immunodeficiency virus-related immune thrombocytopenia. *Blood,* **82,** 29–32.

Spickett GP, *et al.*(1999). Northern region asplenia register—analysis of first two years. *Journal of Clinical Pathology,* **52,** 424–9.

Tefferi A, *et al.* (2000). Splenectomy in myelofibrosis with myeloid metaplasia: a single-institution experience with 223 patients. *Blood,* **95,** 2226–33.

Traub A, *et al.* (1987). Splenic reticuloendothelial function after splenectomy; spleen repair and spleen autotransplantation. *New England Journal of Medicine,* **317,** 1559–64.

Waghorn DJ, Mayon-White RT (1997). A study of 42 episodes of overwhelming post-splenectomy infection: is current guidance for asplenic individuals being followed? *Journal of Infection,* **35,** 289–94.

22.4.5 Myeloma and paraproteinaemias

Robert A. Kyle

The paraproteinaemias are a group of neoplastic, or potentially neoplastic, diseases associated with the proliferation of a single clone of immunoglobulin-secreting plasma cells. They include: multiple myeloma (**MM**); smouldering multiple myeloma (**SMM**); Waldenström's macroglobulinaemia (**WM**); heavy-chain diseases (**HCD**); solitary plasmacytoma of bone, extramedullary plasmacytoma, plasma-cell leukaemia, osteosclerotic myeloma (**POEMS** syndrome); monoclonal gammopathy of undetermined significance (**MGUS**); and primary systemic amyloidosis (**AL**).

The paraproteinaemias are characterized by the secretion of electrophoretically and immunologically homogeneous (monoclonal) (M) proteins (Table 1). Each M-protein consists of two heavy (H) polypeptide chains of the same class and subclass and two light (L) polypeptide chains of the same type. The heavy polypeptide chains are designated by Greek letters: g in IgG, a in IgA, μ in IgM, † in IgD, and e in IgE. The light-chain types are _ (kappa) and l (lambda).

Recognition of M-proteins

Agarose gel electrophoresis is preferred for the detection of M-proteins. Immunofixation should be used to confirm the presence of an M-protein and distinguish the immunoglobulin class and its light-chain type.

Serum protein electrophoresis should be done when MM, WM, or AL amyloidosis is suspected. A paraprotein is characterized by a narrow peak or spike in the densitometer tracing, or as a dense, discrete band on agarose gel (Fig. 1). In contrast, an excess of polyclonal immunoglobulins (having one or more heavy-chain types and both _ and l light chains) produces a broad-based peak or broad band. It is important to differentiate an M-protein from a polyclonal increase because the former is associated with a malignant process or a potentially neoplastic condition, whereas a polyclonal increase in immunoglobulins is associated with a reactive or inflammatory process. Immunofixation is the preferred technique for identifying an M-protein. Diseases associated with a paraprotein, as found in our practice in 2000, are shown in Fig. 2. Immunofixation of an adequately concentrated 24-hour urine specimen is best for detection of a monoclonal light chain (Bence Jones protein). The presence of a monoclonal light chain in nephrotic urine is strongly suggestive of AL or light-chain deposition disease.

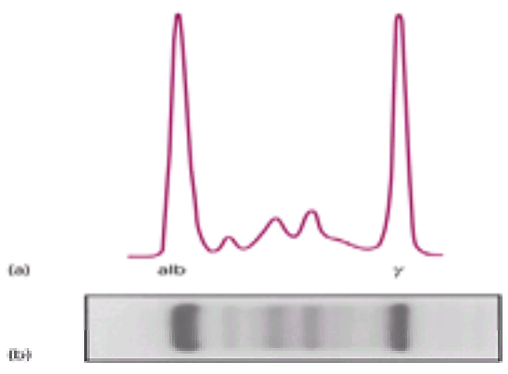

Fig. 1 (a) Monoclonal pattern of serum protein as traced by a densitometer after electrophoresis on agarose gel; tall, narrow-based peak of g mobility. (b) Monoclonal pattern from electrophoresis of serum on agarose gel (anode on left); dense, localized band representing monoclonal protein of g mobility. (From Kyle RA and Katzmann JA (1997). Immunochemical characterization of immunoglobulins. In: Rose NR, *et al*., eds. *Manual of clinical laboratory immunology*, 5th edn, pp 156–76. ASM Press, Washington, DC. By permission of the American Society for Microbiology.)

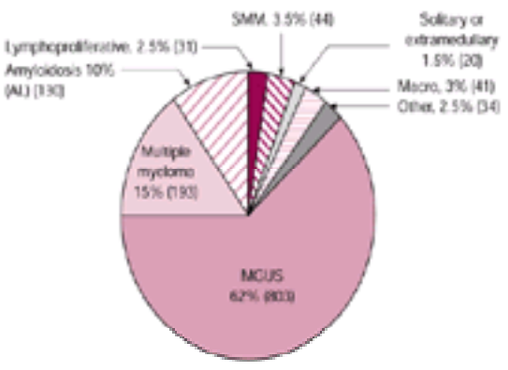

Fig. 2 Types of monoclonal gammopathies in 1296 Mayo Clinic cases in 2000.

Monoclonal gammopathy of undetermined significance (MGUS)

The term 'monoclonal gammopathy of undetermined significance' (MGUS) (benign monoclonal gammopathy) denotes the presence of a paraprotein in persons without evidence of MM, WM, AL, or related disorders. MGUS is characterized by a serum paraprotein concentration of less than 30 g/l; fewer than 5 per cent plasma cells in the bone marrow; no or only small amounts of paraprotein in the urine; absence of lytic bone lesions, anaemia, hypercalcaemia, and renal insufficiency; and, most important, the stability of the paraprotein and the failure of other abnormalities to develop. The prevalence of MGUS is 1 per cent in patients 50 years or older and 3 per cent in those over 70 years of age.

In a series of 241 patients with MGUS followed for 24 to 39 years, 26 per cent developed MM, WM, or AL. The median age at diagnosis was 64 years. Laboratory abnormalities such as anaemia or renal insufficiency were the result of unrelated disorders. The paraprotein concentration ranged from 3 to 30 g/l (median, 17 g/l). The paraproteins consisted of IgG (73 per cent), IgA (11 per cent), IgM (14 per cent), or biclonal (2 per cent). The bone marrow plasma cells ranged from 1 per cent to 10 per cent (median, 3.0 per cent).

After 24 to 39 years of follow-up, the 241 patients were classified into four groups (Table 2). Of these patients, 10 per cent have remained stable and could be classified as having 'benign' monoclonal gammopathy, but they must continue to be observed because serious disease may still develop. More than half of the patients died of unrelated causes without developing MM or a related disorder. In 26 per cent, MM (18 per cent), WM (3 per cent), AL (3 per cent), or related disorders (2 per cent) developed; the actuarial rate was 16 per cent at 10 years, 33 per cent at 20 years, and 40 per cent at 25 years. The interval from the time of recognition of the paraprotein to the diagnosis of serious disease ranged from 2 to 29 years (median, 10 years). In seven patients, MM was diagnosed more than 20 years after detection of the paraprotein.

Differential diagnosis of MGUS from MM and WM

The size of the paraprotein in the serum or urine is of some help. SMM is characterized by the presence of a paraprotein concentration of more than 30 g/l, more than 10 per cent plasma cells in the bone marrow, but no anaemia, renal insufficiency, or skeletal lesions. Affected patients must be recognized because they may remain stable for years and not require therapy. The presence of a urinary paraprotein suggests MM, but small amounts of _ or l paraprotein may persist in the urine and remain stable for years. Large numbers of plasma cells in the bone marrow suggest MM, but some patients may have a plasmacytosis of more than 10 per cent and remain stable. The presence of osteolytic lesions strongly suggests MM, but metastatic carcinoma must be excluded. The plasma-cell labelling index measures the synthesis of DNA, and when increased it is good evidence that the patient has MM. The presence of circulating plasma cells in the peripheral blood usually indicates MM rather than MGUS. No single test will distinguish the patient with MGUS who remains stable from those who develop MM or related disorders. The paraprotein level in the serum and urine should be serially measured, along with periodic re-evaluation of clinical and other features to determine whether MM or a similar disorder is present.

Multiple myeloma (MM)

MM (myelomatosis, Kahler's disease) is characterized by the neoplastic proliferation of a single clone of plasma cells producing a paraprotein. Proliferation of the plasma cells in the bone marrow produces skeletal destruction that leads to bone pain and pathological fractures. The paraprotein can lead to renal failure, recurrent bacterial infections, or hyperviscosity syndrome.

Epidemiology and aetiology

MM accounts for 1 per cent of all malignant diseases and slightly more than 10 per cent of haematological malignancies in the United States. The annual incidence is 4 to 5 per 100 000. The apparent increase during the past few decades is probably related to the increased availability and use of medical facilities, especially in older persons. The incidence in African–Americans is twice that in Caucasians, whereas rates are lower in Asian populations. The median age at diagnosis is about 65 years. Only 18 per cent of patients are younger than 50 years, and 3 per cent are younger than 40 years. The cause of multiple myeloma is unknown, but herbicides, insecticides, and organic solvents may play a role. Human herpesvirus-8 (HHV-8) has been reported in dendritic cells and may play a role in the pathogenesis of MM.

Biological aspects

The plasma cells are phenotypically CIg^{+}, CD38^{+}, PCA-1^{+}, CD56^{+}, with only a minority expressing CD10, CD20, and HLA-DR. Although still unknown, the clonogenic cell in MM appears to arise from the germinal centre, circulates in the peripheral blood, and may home to the bone marrow by means of adhesion molecules. Interleukin-6 (**IL-6**) is a potent plasma-cell growth factor and may be increased in MM, in contrast to MGUS. Overproduction of interleukin-1 (IL-1) and tumour necrosis factor, which have bone-resorbing activity, have been found in MM. Approximately 15 per cent of patients have point mutations of *p53*, a tumour suppressor gene.

Conventional cytogenetic studies reveal an abnormal karyotype in only 40 per cent of patients because of the low proliferative rate of plasma cells. Fluorescence *in situ* hybridization using chromosome-specific probes identifies abnormalities in more than 90 per cent of patients with MM, but no specific pattern has been identified.

Clinical manifestations

Bone pain, frequently in the back or chest, is present at diagnosis in more than two-thirds of patients. Loss of height from multiple vertebral collapses may occur. The most common symptoms are weakness and fatigue, which are often due to anaemia. Fever is rare and, when present, is usually due to an infection. An acute infection, renal failure, hypercalcaemia, or amyloidosis may be the presenting feature. The liver is palpable in about 20 per cent of patients, and the spleen in 5 per cent. Extramedullary plasmacytomas are uncommon and are usually observed late in the course of the disease as large, purplish, subcutaneous masses.

Laboratory findings

If MM is suspected, the laboratory tests listed in Table 3 should be performed. Anaemia is initially present in two-thirds of patients but eventually is found in almost all. The serum protein electrophoretic pattern shows a spike or localized band in 80 per cent of cases, hypogammaglobulinaemia is present in 10 per cent, and no apparent abnormality is found in the remainder. The paraprotein is IgG in about 50 per cent of patients, IgA in 20 per cent, and Bence Jones proteinaemia in almost 20 per cent. IgD occurs in 2 per cent, and biclonal paraproteinaemias are found in 1 per cent, whereas the remainder of the patients have no serum M-protein at diagnosis. Immunofixation of the urine shows a paraprotein in approximately 75 per cent of cases. The _/l ratio is 2:1. A paraprotein is found in the serum or urine at diagnosis in 98 per cent of cases. Hypercalcaemia is initially present in 15 per cent, about one-fifth of whom have a serum creatinine value of 20 mg/l or more.

The bone marrow usually contains more than 10 per cent plasma cells, but involvement may be focal, and repeat bone marrow examination may be necessary for diagnosis. The presence of monoclonal _ or l in the cytoplasm of plasma cells, identified by immunoperoxidase staining, is useful for differentiating monoclonal from reactive plasmacytosis (polyclonal) due to connective tissue disorders, metastatic carcinoma, liver disease, or chronic infections.

Conventional radiographs show abnormalities consisting of lytic lesions, osteoporosis, or fractures in almost 80 per cent of patients at diagnosis. The vertebrae, skull, thoracic cage, pelvis, and humeri and femurs are the most commonly involved sites. Osteoblastic lesions are rare. Technetium-99m bone scanning is inferior to conventional radiography and should not be used. Magnetic resonance imaging reveals abnormalities in 90 per cent of patients with MM. It is particularly helpful in patients who have back pain but no abnormalities on radiography, in whom spinal cord compression must be considered.

Organ involvement

Renal

The serum creatinine value is 20 mg/l or more in 20 per cent of patients at diagnosis. Bence Jones proteinuria is present in 75 per cent. The two major causes of renal failure are myeloma kidney and hypercalcaemia. Myeloma kidney is characterized by the presence of dense, waxy, laminated casts in the distal and collecting tubules. The casts consist mainly of monoclonal light chains. Dilatation and atrophy of the tubules occur, and the entire nephron becomes non-functional. Dehydration contributes to acute renal failure and must be avoided. Hypercalcaemia, present in 15 per cent of patients initially, is a major and treatable cause of renal insufficiency. Amyloidosis occurs in about 10 per cent of patients and may produce nephrotic syndrome and renal insufficiency. Hyperuricaemia, contrast media, antibiotics, and dehydration may contribute to renal failure.

Neurological

Radiculopathy is the most frequent neurological complication and usually involves the thoracic or lumbosacral areas. Compression of the spinal cord from extradural myeloma occurs in 5 per cent of patients. Leptomeningeal involvement is uncommon but is being recognized more frequently.

Other organ systems

The incidence of bacterial infection is increased in MM. Impairment of antibody response, neutropenia, treatment with glucocorticoids, and reduction of normal immunoglobulins increase the likelihood of infection. Coating of platelets by paraprotein may cause bleeding. Occasionally, a tendency to thrombosis is present.

Diagnosis

The diagnosis of MM depends on the presence of an increased number of plasma cells in the bone marrow (usually more than 10 per cent), a paraprotein in the serum (usually more than 30 g/l), Bence Jones proteinuria, and osteolytic lesions. The clinical features of MM must also be present for diagnosis. Metastatic carcinoma, connective tissue disorders, lymphoma, or chronic infections must be considered in the differential diagnosis.

Monoclonal gammopathy of undetermined significance (MGUS), smouldering multiple myeloma (SMM), primary systemic amyloidosis (AL), and metastatic carcinoma are the main conditions considered in the differential diagnosis. In MGUS, the paraprotein value is less than 30 g/l, and the bone marrow contains fewer than 10 per cent plasma cells. There are no osteolytic lesions, anaemia, hypercalcaemia, or renal insufficiency. SMM is characterized by the presence of a paraprotein value of 30 g/l or more and more than 10 per cent plasma cells in the bone marrow but no other findings or symptoms of MM. An increased plasma-cell labelling index strongly suggests that the patient has or soon will have symptomatic MM. However, it must be kept in mind that this value is normal in one-third of patients with symptomatic MM. Monoclonal plasma cells of the same isotype are present in the peripheral blood in 75 per cent of patients with active MM, but patients with MGUS or SMM have few or no circulating plasma cells.

The differentiation of AL and MM is arbitrary because both diseases are plasma-cell proliferative disorders with different manifestations. In AL, the bone marrow plasma-cell content is usually less than 20 per cent, there are no osteolytic lesions, and the amount of Bence Jones proteinuria is modest. Obviously, there is considerable overlap between AL and MM.

Prognostic features

The median duration of survival in MM is approximately 3 years, but there is a great deal of variability from one patient to another. In our experience, the plasma-cell labelling index and the b_2-microglobulin level are the two most powerful prognostic factors. The presence of a low index and a low b_2-microglobulin level is associated with a median survival of almost 6 years when treated with conventional chemotherapy. Cytogenetic abnormalities are an important prognostic factor. The deletion of chromosome 13 and the presence of translocations are predictors of poor outcome. The level of C-reactive protein correlates with the serum IL-6 level and is a useful prognostic factor. Plasmablastic morphology, circulating myeloma cells in the peripheral blood, and increased levels of IL-6 are all associated with more aggressive disease. The Durie–Salmon clinical staging system, in use for almost 25 years, has been superseded by these newer parameters.

Treatment

Although most patients with MM have symptomatic disease at diagnosis and require therapy, some are asymptomatic and should not be treated. All symptoms, physical findings, and laboratory data must be considered in making the decision to begin therapy. An increasing level of the paraprotein in the serum or urine, development of anaemia, hypercalcaemia, or renal insufficiency, and the occurrence of lytic lesions or extramedullary plasmacytomas are all indications for therapy. If there is doubt about beginning treatment, the most reasonable approach is to re-evaluate the patient in 2 months and to delay therapy until progressive disease is evident.

If the patient is younger than 70 years, the physician should discuss the possibility of autologous peripheral blood stem-cell transplantation. Haemopoietic stem cells should be collected before the patient is exposed to alkylating agents.

Chemotherapy is the preferred initial treatment for overt symptomatic MM in persons older than 70 years or in younger patients in whom transplantation is not feasible. Oral administration of melphalan and prednisone produces an objective response in 50 to 60 per cent of patients. A reasonable schedule is melphalan orally in a dosage of 8 to 10 mg/day for 7 days and prednisone 20 mg three times a day orally for the same 7 days. The melphalan should be given when the patient is fasting, because absorption is reduced after food is eaten. Leucocyte and platelet counts must be determined at 3-week intervals after the start of therapy, and the melphalan dosage should be altered until mid-cycle neutropenia or thrombocytopenia occurs. Melphalan and prednisone therapy should be repeated every 6 weeks and the dosage altered depending on the blood counts. Unless the disease progresses rapidly, at least three courses of melphalan and prednisone should be given before therapy is discontinued. An objective response may not be achieved for 6 to 12 months, or even longer in some patients.

Because of the obvious shortcomings of melphalan and prednisone, various combinations of therapeutic agents have been tried. In an overview of individual data in 4930 patients from 20 randomized trials comparing melphalan and prednisone with various combinations of therapeutic agents, the response rates were significantly higher with combination chemotherapy (60 per cent) than with melphalan and prednisone (53 per cent) (p <0.00001). There was no evidence that any subset of patients benefited from receiving combination therapy.

Chemotherapy should be continued until the patient is in a plateau state, or for at least 1 year. A plateau state is defined as stable serum and urine paraprotein levels and no other evidence of progression. Chemotherapy should be discontinued when a plateau state occurs, because continued therapy may lead to the development of a myelodysplastic syndrome or acute leukaemia.

Autologous transplantation

Autologous peripheral stem-cell transplantation has virtually replaced autologous bone marrow transplantation, because engraftment is more rapid and there is less contamination by myeloma cells. Autologous, peripheral stem-cell transplantation is applicable for more than half the patients with MM. The two major shortcomings are that: (1) the myeloma is not eradicated even with large doses of chemotherapy and total body radiation; and (2) autologous, peripheral stem cells are contaminated by myeloma cells and their precursors. Fortunately, the mortality from autologous transplantation is currently 1 per cent if patients are appropriately selected.

Most physicians initially treat the patient with vincristine (Oncovin), doxorubicin (Adriamycin), and dexamethasone (**VAD**) for 3 to 4 months to reduce the number of tumour cells in the bone marrow and peripheral blood. The peripheral stem cells are then collected after treatment of the patient with high-dose cyclophosphamide and granulocyte colony-stimulating factor (**G-CSF**). One can then proceed with the transplant, in which the patient is given high-dose chemotherapy followed by infusion of the peripheral blood stem cells. The other choice is to treat the patient with alkylating agents after stem-cell collection until a plateau state is reached, and then treat with a_2-interferon (**IFN-a**$_2$) or no therapy until early relapse. At that time the patient is given high-dose melphalan or total-body radiation, and the previously collected peripheral blood stem cells are infused. In a French study, 185 patients with primary resistant or relapsed disease were treated with three or four courses of VAD and then randomized to high-dose chemotherapy and autologous stem-cell transplantation or to conventional therapy with high-dose chemotherapy and

autologous transplantation given at relapse (early versus late transplantation); the two groups showed no difference in median overall survival (65 versus 64 months).

The largest single-institution experience with autologous transplantation in myeloma included 496 patients enrolled in a tandem transplant programme. Complete response was obtained in 36 per cent and the transplant-related mortality was 7 per cent. The overall survival from the time of the first transplant was 41 months. This series was heterogeneous and included patients with resistant disease and those with disease sensitive to conventional chemotherapy. In a recent report of 231 patients receiving tandem transplants, the overall median survival was 68 months.

A randomized trial performed by the French Myeloma Group compared high-dose chemotherapy and autologous bone marrow transplantation with conventional chemotherapy in 200 previously untreated patients under 65 years of age. The rates of response (81 per cent versus 57 per cent) and complete responses (20 per cent versus 5 per cent) were superior in the transplant group. The transplant group had a higher rate of 5-year, event-free survival (28 per cent versus 10 per cent) and overall survival (52 per cent versus 12 per cent).

It has been suggested that better results could be obtained with two (tandem) autologous peripheral stem-cell transplants. In a randomized trial of 400 patients from France, there was no difference in event-free or overall survival between double and single autologous stem-cell transplantation at 2 years. Longer follow-up analysis is necessary.

A major hurdle is improvement of the preparative regimen because residual myeloma is the likely source of relapse in most patients. In a comparison (non-randomized) of melphalan (140 mg/m^2) plus total-body radiation or melphalan (200 mg/m^2), no difference was found in remission status, event-free survival, or overall survival. The other major shortcoming of autologous stem-cell transplantation is the presence of myeloma cells and their precursors in the blood. Collection of CD34$^+$ cells produces a lower number of tumour cells, but it remains to be seen whether this results in better responses and survival. Because relapse occurs in nearly all patients, the use of dendritic cells and vaccines after autologous transplantation shows some promise.

Allogeneic bone marrow transplantation

This is advantageous because the graft contains no tumour cells that can lead to relapse. Unfortunately, the mortality rate is approximately 25 per cent within 3 months. Furthermore, more than 90 per cent of patients with MM are ineligible for an allogeneic transplant because of their age, lack of an HLA-matched sibling donor, or inadequate renal, pulmonary, or cardiac function.

The mortality rate for allogeneic transplantation must be reduced before it can assume a major role in the treatment of MM. A preparative regimen using fludarabine and melphalan ('mini-allotransplant') may result in a lower mortality. The use of T-cell-depleted peripheral allogeneic stem cells decreases the incidence of graft-versus-host disease and transplant mortality. The use of donor leucocyte infusions for relapses after allogeneic transplantation produces benefit in about half of patients. However, allogeneic transplantation is currently associated with a high mortality and cannot be recommended as a routine procedure.

Maintenance therapy

It would be desirable to keep the patient in a plateau state indefinitely, but this is not possible. An overview by the Myeloma Trialists' Group revealed relapse-free survival at 5 years in 23 per cent of patients receiving IFN-α_2 and 16 per cent without IFN-α_2 (p <0.001). The overall survival at 5 years was only modestly prolonged. Consequently, IFN-α_2 cannot be strongly recommended for maintenance therapy. Patients should be monitored closely during the plateau state, and the same therapy should be reinstituted if relapse occurs more than 6 months after the plateau state has begun.

Refractory multiple myeloma

The highest response rates have been reported with VAD given by continuous infusion for 4 days and dexamethasone (40 mg daily on days 1–4, 9–12, and 17–20). The response rate is approximately 60 per cent for patients who relapse while not receiving chemotherapy, but only 40 per cent for those who relapse while receiving alkylating agent therapy. Dexamethasone can be used as a single agent in the same dosage and schedule as VAD because the steroids probably account for 80 per cent of the benefit of VAD. Methylprednisolone (2 g three times weekly intravenously for 4 weeks) is helpful for refractory disease with pancytopenia. **VBAP** (vincristine (Oncovin), carmustine [BCNU], and doxorubicin [Adriamycin] on day 1 and prednisone daily for 5 days every 3 to 4 weeks) produces benefit in 30 per cent of patients and is easily administered. If the leucocyte and platelet levels are satisfactory, cyclophosphamide (600 mg/m^2 daily, intravenously, for 4 days) plus prednisone (50 mg twice daily for the same 4-day period) followed by G-CSF has been helpful for patients with refractory, advanced disease. The use of IFN-α_2 as a single agent benefits some patients, but the results have been disappointing.

Thalidomide has shown to be of benefit in approximately 30 per cent of patients with refractory disease. Resistance to chemotherapeutic agents is a major problem.

Supportive care

Skeletal complications

Skeletal involvement often leads to pathological fractures, spinal cord compression, pain, or hypercalcaemia. These complications result from increased osteoclastic bone resorption, which is inhibited by bisphosphonates. In a prospective placebo-controlled study, patients receiving pamidronate had fewer skeletal complications and experienced a reduction in bone pain as well as improved quality of life. Pamidronate in a dosage of 90 mg intravenously every 4 weeks is recommended for patients with MM who have lytic lesions or osteopenia. Its use should be continued indefinitely. Clodronate, an orally administered bisphosphonate, has also been reported to be beneficial.

Patients should be encouraged to be as active as possible, but they must avoid undue trauma. Fixation of fractures or pending fractures with an intramedullary rod and methyl methacrylate has produced good results. Bone pain should be treated with analgesics or narcotics as necessary.

Hypercalcaemia

This is present in 15 per cent of patients at diagnosis and should be suspected in the presence of anorexia, nausea, vomiting, polyuria, polydipsia, increased constipation, weakness, confusion, or stupor. If untreated, renal insufficiency develops. Hydration plus prednisone (25 mg four times a day until the serum calcium level decreases) is effective in most cases. If these measures fail, pamidronate or etidronate is effective.

Renal failure

Two major causes of renal insufficiency are 'myeloma kidney' and hypercalcaemia. Maintenance of a high urine output (3 litres/day) is important for preventing renal failure in patients with Bence Jones proteinuria. Haemodialysis or peritoneal dialysis is necessary in the event of symptomatic azotaemia. Plasmapheresis may be useful in acute renal failure, but patients with severe myeloma cast formation or other irreversible changes are unlikely to benefit. Allopurinol should be administered if hyperuricaemia is present. Patients with acute renal failure should be treated with VAD or dexamethasone to reduce the tumour mass as quickly as possible.

Infection

Appropriate therapy for bacterial infections is essential. Patients should receive pneumococcal and influenza vaccination despite their suboptimal antibody response. Prophylactic daily oral penicillin (500 mg daily indefinitely) often benefits patients with recurrent pneumococcal infections. Because many infections occur in the first 2 months after instituting therapy, trimethoprim–sulfamethoxazole is useful. Intravenously administered gammaglobulin can be used for recurrent infections, but it is very expensive.

Neurological

Spinal cord compression should be suspected in patients with severe back pain who develop weakness or paraesthesias of the lower extremities or bladder or bowel

dysfunction. Magnetic resonance imaging (**MRI**) or CT scans must be done immediately. Radiation therapy and dexamethasone are usually effective, and surgical decompression is rarely necessary.

Hyperviscosity

This is characterized by oral or nasal bleeding, blurred vision, paraesthesias, headache, reduced cerebration, or congestive heart failure. Serum viscosity levels do not correlate well with the symptoms or clinical findings. A decision to perform plasmapheresis depends on the symptoms and changes in the ocular fundus. Plasmapheresis promptly relieves the symptoms and should be done regardless of the viscosity level if the patient has signs or symptoms of hyperviscosity.

Anaemia

Anaemia occurs in almost all patients during the course of MM. Erythropoietin (150 U/kg 3 times weekly or 40 000-U weekly) will increase haemoglobin in 50 to 60 per cent of patients.

Emotional support

All patients with MM need substantial and continuing emotional support. The physician's approach must be positive and emphasize the potential benefits of therapy. It is reassuring for patients to know that some survive for 10 years or more. It is vital that the physician caring for patients with MM has the interest and capacity to deal with an incurable disease over the space of years with assurance, sympathy, and resourcefulness.

Variant forms of multiple myeloma

Smouldering multiple myeloma (SMM)

See above.

Plasma-cell leukaemia

Plasma-cell leukaemia is defined as the presence of more than 20 per cent plasma cells in the peripheral blood and an absolute plasma-cell count of more than 2×10^9/l. It is classified as primary when it presents *de novo* (60 per cent of cases) and as secondary when it is a leukaemia transformation of a previously recognized myeloma (40 per cent). Patients with primary plasma-cell leukaemia are younger and have a higher platelet count, fewer bone lesions, a smaller serum paraprotein, a greater incidence of hepatosplenomegaly and lymphadenopathy, and a longer duration of survival than patients with secondary plasma-cell leukaemia. Cytogenetic abnormalities are more common than in patients with MM. Autologous stem-cell transplantation after high-dose chemotherapy is beneficial for some patients. Those with secondary plasma-cell leukaemia rarely respond to chemotherapy because they already received treatment and are resistant.

Non-secretory myeloma

These patients have no paraprotein in either the serum or the urine and account for only 2 per cent of patients with myeloma at diagnosis. The diagnosis is established by identification of an M-protein in the cytoplasm of the plasma cells by immunoperoxidase or immunofluorescence staining.

Osteosclerotic myeloma (POEMS syndrome)

This is characterized by polyneuropathy (P), organomegaly (O), endocrinopathy (E), M-protein (M), and skin changes (S). The major clinical finding is a chronic inflammatory-demyelinating neuropathy with predominantly motor disability. Sclerotic bone lesions are found in most patients. The cranial nerves are not involved except for the presence of papilloedema. Hepatomegaly occurs in almost half of patients, but splenomegaly and lymphadenopathy occur in a minority. Hyperpigmentation and hypertrichosis are frequent but may be easily overlooked. Gynaecomastia and atrophic testes as well as clubbing of the fingers and toes may be present. Angiomatous lesions of the trunk are often prominent. Pulmonary hypertension has been recognized in several instances. Ascites, pleural effusion, and peripheral oedema may be present. In contrast to MM, the haemoglobin level is usually normal or increased, and thrombocytosis is common. The bone marrow usually contains fewer than 5 per cent plasma cells, and hypercalcaemia and renal insufficiency rarely occur. Most patients have a l light chain, and IgA is the most common heavy-chain type. Castleman's disease may be present. The diagnosis is confirmed by the identification of monoclonal plasma cells obtained from an osteosclerotic lesion. If the skeletal lesions are in a limited area, radiation almost always produces a substantial improvement of the neuropathy. If widespread osteosclerotic lesions exist, chemotherapy or an autologous stem cell transplant should be used for therapy.

Solitary plasmacytoma (solitary myeloma) of bone

The diagnosis depends on histological evidence of a plasma-cell tumour but no evidence of MM. Complete skeletal radiographs, bone marrow aspiration and biopsy, and immunofixation of the serum and urine should reveal no evidence of MM. Occasionally, a small paraprotein may be found in the serum or urine, but it usually disappears after radiation of a solitary lesion. Treatment consists of tumoricidal radiation (40–50 Gy). Overt MM develops in approximately 55 per cent of patients, and new solitary lesions or local recurrence develops in about 10 per cent. MRI scans may be helpful for identifying patients in whom MM will develop in the near future.

Extramedullary plasmacytoma

This is a plasma-cell tumour that arises outside the bone marrow. It is located in the upper respiratory tract in approximately 80 per cent of cases, and the nasal cavity and sinuses, nasopharynx, and larynx are most often involved. The gastrointestinal tract, central nervous system, urinary bladder, thyroid, breast, testes, parotid gland, and lymph nodes have all been reported as the initial site of an extramedullary plasmacytoma. There is a predominance of IgA M-protein in extramedullary plasmacytomas. The diagnosis depends on the finding of a plasma-cell tumour in an extramedullary location and the absence of MM on bone marrow examination, radiography, and appropriate studies of serum and urine. Treatment consists of tumoricidal radiation (40–50 Gy). Regional occurrences develop in approximately 25 per cent of patients, but the development of typical MM is uncommon.

Waldenström's macroglobulinaemia (WM)

This malignant plasma-cell proliferative disorder produces a high concentration of immunoglobulin M (IgM) paraprotein. It bears similarities to MM, lymphoma, and chronic lymphocytic leukaemia. The incidence rate is 0.5/100 000, and in our practice it is one-seventh as common as MM. The median age is approximately 65 years, and 60 per cent of patients are male.

Clinical findings

Weakness and fatigue are the most common features. Chronic nasal bleeding or oozing from the gums is characteristic, but postsurgical or gastrointestinal bleeding may occur. Blurring or loss of vision may be prominent. Dyspnoea and congestive heart failure may develop. Dizziness, headaches, vertigo, nystagmus, ataxia, and diplopia have been seen. Constitutional symptoms including fever, night sweats, and loss of weight may be present. Bone pain is rare. Hepatomegaly occurs in about 25 per cent of patients at diagnosis, and splenomegaly and lymphadenopathy are slightly less common. Retinal vein engorgement and flame-shaped haemorrhages are common and are a better measure of symptomatic hyperviscosity syndrome than is the measurement of serum viscosity.

Pulmonary involvement may be manifested by diffuse pulmonary infiltrates, isolated masses, or pleural effusion. Retroperitoneal and mesenteric lymphadenopathy are common, but they are usually asymptomatic. The most common neurological manifestation is sensorimotor peripheral neuropathy. It is often related to amyloid deposition.

Laboratory findings

Anaemia is found in most patients with symptomatic WM. Spuriously low haemoglobin and haematocrit levels may result from an increased plasma volume due to the

large amount of paraprotein.

Serum protein electrophoresis reveals a tall, narrow spike or dense band usually migrating in the g area. About 75 per cent of the IgM paraproteins are _. The IgM level obtained by nephelometry is often 1000 to 3000 mg/l more than that found with serum protein electrophoresis. A reduction of uninvolved IgG and IgA immunoglobulins is less striking than in MM. About 10 per cent of macroglobulins precipitate in the cold (cryoglobulin). A monoclonal light chain detected by immunofixation is present in the urine in 75 per cent of patients.

Fewer than 5 per cent of patients with WM have lytic bone lesions. The bone marrow aspirate is often hypocellular, but the biopsy specimen is usually hypercellular and extensively infiltrated with lymphoid or plasmacytoid cells.

Diagnosis

The diagnosis of WM depends on the presence of an IgM paraprotein and a lymphocyte–plasma cell infiltration of the bone marrow producing symptoms and physical findings consistent with WM. The differential diagnosis includes MM, MGUS of the IgM type, chronic lymphocytic leukaemia, lymphoma, and undifferentiated lymphoplasma-cell proliferative processes.

Treatment

Patients with WM should not be treated unless they are symptomatic. Symptoms and findings of hyperviscosity are quickly controlled by plasmapheresis with a cell separator. Therapy must be directed against the proliferating lymphocytes and plasma cells because symptoms will recur quickly.

Chlorambucil (Leukeran) is usually given orally in an initial dosage of 6 to 8 mg daily and is reduced when the leucocytes or platelets decrease. It also may be given intermittently at monthly intervals. Patients should be treated until the disease has reached a plateau state, which occurs in about 70 per cent of patients. Patients should be treated for at least 6 months before chlorambucil therapy is abandoned because of a slow response. Cyclophosphamide or combinations of alkylating agents such as the M2 protocol (vincristine, BCNU, melphalan, cyclophosphamide, and prednisone) have also been beneficial. Patients must be followed closely, and chemotherapy of the same type should be reinstituted when the disease relapses. Fludarabine and cladribine (2-chlorodeoxyadenosine) have been reported to produce responses in 80 per cent of patients. However, in a recent report, only one-third of patients with symptomatic WM responded to fludarabine. Rituximab (Rituxan) produces a response in approximately half of patients with refractory disease.

Packed red cells should be transfused into patients with symptomatic anaemia. Erythropoietin may be of help. The median duration of survival for patients with macroglobulinaemia is approximately 5 years.

Heavy-chain diseases

Gamma heavy-chain disease (g-HCD)

The paraprotein consists of a monoclonal g chain with significant amino-acid deletions. The median age of patients is approximately 60 years, but the disease has been recognized in persons under 20 years of age. The initial presentation is often a lymphoma-like illness, but the symptoms and clinical findings are diverse and range from an aggressive lymphoproliferative process to an asymptomatic state. Weakness, fatigue, and fever are the most common presenting symptoms. Hepatosplenomegaly and lymphadenopathy are found in about 60 per cent of patients. Anaemia is present in 80 per cent. The serum protein electrophoretic pattern usually shows a broad-based band more suggestive of a polyclonal than an M-protein. The urinary heavy-chain protein value is usually less than 1 g/24 h. Bence Jones proteinuria is not found. The bone marrow and lymph nodes contain an increased number of plasma cells, lymphocytes, and lymphoplasmacytoid cells.

Only symptomatic patients should be treated. Therapy with cyclophosphamide, vincristine, and prednisone is a reasonable choice. If there is no response, a doxorubicin-containing regimen should be used. The median duration of survival is approximately 1 year.

Alpha heavy-chain disease (a-HCD)

a-HCD is the most common type of heavy-chain disease, with more than 200 reported patients since its recognition. It usually occurs in the second or third decade of life, and about 60 per cent of patients are male. Most patients have been from the Mediterranean region and Middle East. Gastrointestinal tract involvement is most common and is manifested by malabsorption with loss of weight, diarrhoea, and steatorrhoea. It is similar to 'immunoproliferative small intestinal disease' (**IPSID**), but patients with IPSID do not synthesize a heavy chains. The serum protein electrophoretic pattern shows no spike. The diagnosis depends on recognition of a monoclonal a heavy chain in the serum or jejunal fluid. Bence-Jones proteinuria never appears. The bone marrow is not infiltrated with lymphocytes. a-HCD is progressive and fatal without therapy. Surprisingly, antibiotics may produce remission, particularly if given early in the course of the disease. Patients who have advanced disease or who do not respond to antibiotics should be treated with a combination of chemotherapy consisting of cyclophosphamide, doxorubicin (Adriamycin), vincristine, and prednisone.

Mu heavy-chain disease (µ-HCD)

µ-HCD is characterized by the presence of a monoclonal µ-chain fragment in the serum. Most patients have a chronic lymphoproliferative process resembling chronic lymphocytic leukaemia or lymphoma. The serum protein electrophoretic pattern contains a spike or localized band in about 40 per cent of patients. Bence Jones proteinuria, which is usually _, has been recognized in two-thirds of cases. Vacuolization of the plasma cells in the marrow is an important clue for the diagnosis of µ-HCD. There is an increase in lymphocytes, plasma cells, and lymphocytoid cells in the marrow. The course of µ-HCD is variable, with a median survival of approximately 2 years. Treatment with corticosteroids and alkylating agents has produced benefit.

Primary amyloidosis (AL)

Amyloid is a substance consisting of fibrils that appear homogeneous and amorphous under the light microscope and stain pink with haematoxylin–eosin. With polarized light, amyloid stained with Congo red produces an apple-green birefringence. Linear, non-branching, aggregated fibrils 7.5- to 10-nm wide and of indefinite length are seen with electron microscopy. These fibrils consist of various proteins such as monoclonal _ or l light chains in AL, protein A in secondary amyloidosis, transthyretin (prealbumin) in familial or senile systemic amyloidosis, and b $_2$-microglobulin in dialysis-associated amyloidosis. Because paraproteins are associated only with primary amyloidosis, the other types are not discussed here.

Aetiology and epidemiology

The annual incidence of AL is 0.9/100 000. The median age at diagnosis is 64 years, and only 1 per cent of patients are younger than 40 years. The cause of AL is unknown.

Clinical features

Weakness, fatigue, and weight loss are the most common initial symptoms. Light-headedness, syncope, change in the tongue or voice, jaw or hip claudication, paraesthesias, dyspnoea, and oedema are the most frequent symptoms. Macroglossia is present in 10 per cent of patients, and purpura, particularly in the periorbital and facial areas, is found in 15 per cent. The liver is palpable in 25 per cent of patients, but splenomegaly occurs in only 5 per cent. Nephrotic syndrome or renal failure is found in more than 25 per cent of patients at diagnosis (Fig. 3). Congestive heart failure, carpal tunnel syndrome, sensorimotor peripheral neuropathy, and orthostatic hypotension are other important features. The presence of one of these syndromes and a paraprotein in the serum or urine are strong indications of AL, for which appropriate biopsy specimens must be taken for diagnosis.

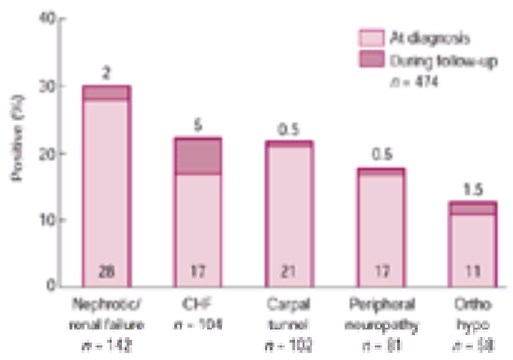

Fig. 3 Frequency of amyloid syndromes at diagnosis of primary systemic amyloidosis. CHF, congestive heart failure; Ortho hypo, orthostatic hypotension. (From Kyle RA and Gertz MA (1995). Primary systemic amyloidosis: clinical and laboratory features in 474 cases. *Seminars in Hematology* **32,** 45–59. By permission of WB Saunders Company.)

Laboratory findings

Anaemia is not a prominent feature of AL and, when present, is usually the result of MM, renal insufficiency, or gastrointestinal bleeding. Thrombocytosis (platelets >500 × 10^9/l) is present in about 10 per cent of cases. Renal insufficiency is present in almost half of patients at diagnosis; 20 per cent have a serum creatinine value of 20 mg/l or more. The serum protein electrophoretic pattern shows a modest localized band or spike in about half of the patients (median, 14 g/l). A paraprotein is found in the serum or urine in 90 per cent of patients, and l light chains are twice as common as _. The bone marrow contains 5 per cent or less plasma cells in almost half of patients. Only one-fifth of patients have more than 20 per cent plasma cells in the bone marrow, but they usually do not have the other features of MM.

An increased serum alkaline phosphatase level is not uncommon. Hyperbilirubinaemia is infrequent, but when present it is an ominous sign. The factor X level is decreased in more than 10 per cent of patients but is rarely the cause of bleeding.

Congestive heart failure is present in about 20 per cent of patients at diagnosis. Electrocardiography frequently reveals low voltage in the limb leads or characteristics consistent with anteroseptal infarction (loss of anterior forces). Arrhythmias, including atrial fibrillation or heart block, are common. Almost two-thirds of patients have an abnormal echocardiogram at diagnosis. Early cardiac involvement is characterized by abnormal relaxation followed by the features of constrictive cardiomyopathy. Amyloid heart disease may closely resemble constrictive pericarditis or hypertrophic obstructive cardiomyopathy. A sensorimotor peripheral neuropathy is present in about 15 per cent of patients at diagnosis. Autonomic dysfunction may be a prominent feature and is often manifested by orthostatic hypotension, diarrhoea, and impotence.

Diagnosis

The diagnosis depends on the demonstration of amyloid deposits. The possibility of AL must be considered in every patient who has a paraprotein in the serum or urine and who has a nephrotic syndrome, congestive heart failure, sensorimotor peripheral neuropathy, carpal tunnel syndrome, giant hepatomegaly, or idiopathic malabsorption syndrome. A paraprotein in the serum or urine or a monoclonal proliferation of plasma cells in the bone marrow occurs in 98 per cent of patients with AL.

The initial diagnostic procedure should be an abdominal fat aspiration, which is positive in about 80 per cent of patients (Fig. 4). A bone marrow aspiration and biopsy should be done to determine the degree of plasmacytosis, and amyloid stains will be positive in more than half of patients. The abdominal fat or bone marrow biopsy is positive in 90 per cent of cases; if negative, a rectal biopsy (including submucosa) or biopsy of a suspected involved organ such as the kidney, liver, heart, or sural nerve is indicated. Specific antisera to _, l, protein A, transthyretin, and b $_2$-microglobulin are useful for identifying the type of systemic amyloidosis. ^{123}I-labelled human serum amyloid-P component is helpful for detecting amyloid deposition.

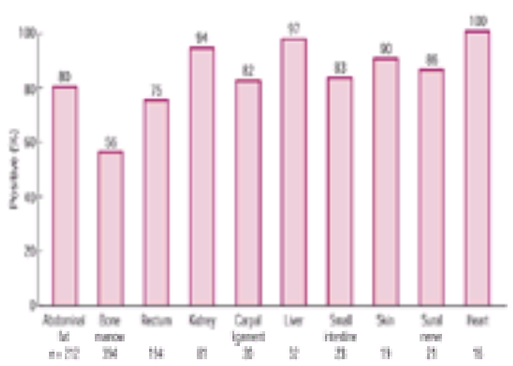

Fig. 4 Diagnosis of amyloidosis on the basis of deposits in tissues. (From Kyle RA and Gertz MA (1995). Primary systemic amyloidosis: clinical and laboratory features in 474 cases. *Seminars in Hematology* **32,** 45–59. By permission of WB Saunders Company.)

Prognosis

The median duration of survival for patients with AL is approximately 13 months. Survival varies greatly depending on the associated syndrome. It is 6 months after the onset of congestive heart failure but more than 2 years in patients presenting with peripheral neuropathy. Almost half of the deaths are due to cardiac involvement.

Treatment

Because amyloid fibrils consist of a monoclonal light chain, treatment with alkylating agents has been a common approach. In a randomized trial, survival of patients receiving the two melphalan–prednisone–containing regimens (17–18 months) was superior to that of patients receiving colchicine (8.5 months). Patients have had substantial clinical improvement after the administration of 4'-iodo-4'-deoxyrubicin (**I-DOX**). This agent appears to bind to the amyloid fibrils and contributes to the resolution of amyloid deposits. Encouraging results have been reported with high-dose intravenous melphalan (100 mg/m ^{2} for 2 days) followed by autologous peripheral blood stem-cell rescue. The impact of this treatment approach needs to be determined because of the short follow-up to date.

This was supported in part by CA 62242 from the National Cancer Institute.

Further reading

Attal M, *et al*., for the Intergroupe Français du Myélome (1996). A prospective, randomized trial of autologous bone marrow transplantation and chemotherapy in multiple myeloma. *New England Journal of Medicine* **335**, 91–7. [A landmark comparison of bone marrow transplantation versus standard chemotherapy.]

Barlogie B, *et al*. (1999). Total therapy with tandem transplants for newly diagnosed multiple myeloma. *Blood* **93,** 55–65.

Berenson JR, *et al*., for the Myeloma Aredia Study Group (1998). Long-term pamidronate treatment of advanced multiple myeloma patients reduces skeletal events. *Journal of Clinical Oncology* **16,**

593–602. [This randomized trial proves the value of pamidronate for multiple myeloma bone disease.]

Bladé J, *et al*. (1998). Renal failure in multiple myeloma: presenting features and predictors of outcome in 94 patients from a single institution. *Archives of Internal Medicine* **158,** 1889–93.

Falk RH, Comenzo RL, Skinner M (1997). The systemic amyloidoses. *New England Journal of Medicine* **337,** 898–909. [A comprehensive review of amyloidosis.]

Fermand JP, *et al*. (1998). High-dose therapy and autologous peripheral blood stem cell transplantation in multiple myeloma: up-front or rescue treatment? Results of a multicenter sequential randomized clinical trial. *Blood* **92,** 3131–6.

Gahrton G, *et al*. (1995). Prognostic factors in allogeneic bone marrow transplantation for multiple myeloma. *Journal of Clinical Oncology* **13,** 1312–22.

Garcia-Sanz R, *et al*. (1999). Primary plasma cell leukemia: clinical, immunophenotypic, DNA ploidy, and cytogenetic characteristics. *Blood* **93,** 1032–7. [A current review of plasma-cell leukaemia.]

Gertz MA, Kyle RA (1995). Hyperviscosity syndrome. *Journal of Intensive Care Medicine* **10,** 128–41. [A comprehensive review of the hyperviscosity syndrome.]

Hallek M, Leif Bergsagel P, Anderson KC (1998). Multiple myeloma: increasing vidence for a multistep transformation process. *Blood* **91,** 3–21. [An excellent review of the biological aspects of multiple myeloma.]

Kyle RA (1975). Multiple myeloma: review of 869 cases. *Mayo Clinic Proceedings* **50,** 29–40. [A good review of the clinical features of a large series of patients with multiple myeloma.]

Kyle RA (1993). 'Benign' monoclonal gammopathy—after 20 to 35 years of follow-up. *Mayo Clinic Proceedings* **68,** 26–36.

Kyle RA (1999). High-dose therapy in multiple myeloma and primary amyloidosis: an overview. *Seminars in Oncology* **26,** 74–83. [A discussion of the advantages and disadvantages of high-dose therapy.]

Kyle RA, Garton JP (1987). The spectrum of IgM monoclonal gammopathy in 430 cases. *Mayo Clinic Proceedings* **62,** 719–31.

Kyle RA, Gertz MA (1995). Primary systemic amyloidosis: clinical and laboratory features in 474 cases. *Seminars in Hematology* **32,** 45–59. [A detailed description of the features of primary amyloidosis.]

Kyle RA, Katzmann JA (1997). Immunochemical characterization of immunoglobulins. In: Rose NR, *et al*., eds. *Manual of clinical laboratory immunology*, 5th edn, pp 156–76. ASM Press, Washington, DC.

Kyle RA, *et al*. (1997). A trial of three regimens for primary amyloidosis: colchicine alone, melphalan and prednisone, and melphalan, prednisone, and colchicine. *New England Journal of Medicine* **336,** 1202–7.

Liebross RH, *et al*. (1998). Solitary bone plasmacytoma: outcome and prognostic factors following radiotherapy. *International Journal of Radiation Oncology, Biology, Physics* **41,** 1063–7.

Myeloma Trialists' Collaborative Group (1998). Combination chemotherapy versus melphalan plus prednisone as treatment for multiple myeloma: an overview of 6,633 patients from 27 randomized trials. *Journal of Clinical Oncology* **16,** 3832–42. [This large meta-analysis fails to show a survival advantage for combination chemotherapy compared with single-agent chemotherapy.]

Susnerwala SS, *et al*. (1997). Extramedullary plasmacytoma of the head and neck region: clinicopathological correlation in 25 cases. *British Journal of Cancer* **75,** 921–7. [A helpful recent review of extramedullary plasmacytoma.]

Waldenström JG (1992). POEMS: a multifactorial syndrome. *Haematologica* **77,** 197–203. [Editorial]

22.4.6

Eosinophilia

Peter F. Weller

Eosinophilia is associated with distinct diseases that include helminth parasitic infections, allergic diseases, and varied diseases of often ill-defined aetiologies. This chapter considers the clinical disorders associated with eosinophilia, with additional information on these diseases available in other chapters and then discusses the idiopathic hypereosinophilic syndrome.

In comparison with other leucocytes, eosinophils are distinguished by their morphologies, constituents, products, and associations with specific diseases. The cytokine interleukin 5 (**IL-5**), specific in promoting the development, differentiation, and release of bone marrow-derived eosinophils, is principally responsible for increases in eosinophilopoiesis. Eosinophils are normally tissue-dwelling cells primarily distributed in those tissues with an epithelial interface with the environment, including the respiratory, gastrointestinal, and lower genitourinary tracts. Eosinophils are distinguished morphologically from neutrophils by their cytoplasmic granules which uniquely contain crystalloid cores visible by electron microscopy. Within these granules are four specific cationic proteins, major basic protein, eosinophil peroxidase, eosinophil cationic protein, and eosinophil-derived neurotoxin. The heavy content of these cationic granule proteins, which bind acidic dyes like eosin, are responsible both for the identifying tinctorial properties of eosinophils and for many of the functional properties of eosinophils. As recently established, eosinophils are sources of over two dozen cytokines; and many, if not all, of these are stored preformed within eosinophil granules. In addition to their content of preformed granule cationic and cytokine proteins, eosinophils also synthesize lipid mediators, including the 5-lipoxygenase pathway-derived eicosanoid, leukotriene C_4. The potential functional roles of eosinophils in parasite–host defence, in the pathogenesis of allergic diseases, and in other immunological responses remain uncertain, due to varied and at times conflicting experimental findings, but are subjects of active ongoing investigations.

Eosinophils normally number less than 450/μl in the blood with a mild diurnal variation, being higher in the morning and falling as endogenous glucocorticosteroid levels rise. Blood eosinophil numbers, however, do not always reflect the extent of eosinophil involvement in affected tissues in various diseases; and at times, as in eosinophilic pneumonias, eosinophils may be recruited into involved tissues without a concomitant increase in enumerable blood eosinophils. Eosinopenia, diminished blood eosinophil levels, occurs with corticosteroid administration and is frequent with active bacterial and viral infections. Thus, even normal blood eosinophil numbers in a febrile patient suggest that an illness is not simply due to a bacterial or viral infection.

Some, but not necessarily all, patients with sustained blood eosinophilia can develop organ damage, especially cardiac, as found in the idiopathic hypereosinophilic syndrome, and patients with sustained eosinophilia should be monitored for evidence of cardiac disease.

Diseases associated with eosinophilia Table 1)

Infectious diseases

Parasitic diseases

Eosinophilia is not elicited by infections with protozoan parasites (with the sole exceptions of the intestinal parasites, *Isospora belli* and *Dientamoeba fragilis*), but rather characteristically by multicellular helminth parasites. Magnitudes of eosinophilia tend to parallel the extent of tissue invasion, especially by helminth larvae. Eosinophilia may be absent in established infections which are well-contained within tissues or are solely intraluminal within the gastrointestinal tract (e.g. *Ascaris*, tapeworms). Even with helminth diseases, superimposed bacterial infections (e.g. in disseminated strongyloidiasis) can suppress expected eosinophilia. In patients with eosinophilia, geographical and dietary histories are pertinent in suggesting potential exposures to helminth parasites. Stool examinations for diagnostic ova and larvae should be obtained. In addition, for several helminth parasites that cause eosinophilia, diagnostic parasite stages are never present in faeces. Hence, negative stool specimens do not necessarily exclude a helminth aetiology for eosinophilia; and examination of appropriate blood or tissue biopsies, as guided by clinical findings and exposure histories, may be needed to diagnose specific tissue- or blood-dwelling infections, including trichinellosis and filarial infections.

Other infectious diseases

The characteristic response in acute bacterial and viral infections is eosinopenia. One fungal disease, coccidioidomycosis, either following primary infection, at times with progressive disseminated disease, or with central nervous system infection (with cerebrospinal fluid eosinophilia), may be associated with eosinophilia.

HIV and retroviral infections

Eosinophilia may be associated with HIV infections: (i) leukopenia may elevate the percentages, but not true numbers, of eosinophils; (ii) adverse reactions to medications may elicit eosinophilia; and (iii) eosinophilia may be due to adrenal insufficiency in patients with AIDS from cytomegalovirus and other infections. In addition, eosinophilia, often modest, is observed in some HIV-infected patients and may accompany eosinophilic folliculitis in HIV infection. Uncommonly, marked hypereosinophilia has developed with HIV infections, including those with a hyperimmunoglobulin E syndrome or exfoliative dermatitis. Eosinophilia frequently develops with HTLV-1 infections.

Allergic and immunological disorders

Common allergic diseases, including allergic rhinitis, asthma, and atopic dermatitis, are accompanied by tissue eosinophil infiltration and usually modest blood eosinophilia. The occurrence of marked blood eosinophilia suggests the presence of other diseases, such as Churg–Strauss vasculitis.

Medication-related eosinophilias

Therapeutic agents, including herbal or 'natural' therapies, can elicit eosinophilia. Eosinophilia may develop without other manifestations of adverse drug reactions, such as rashes or drug fevers. In the absence of organ involvement, blood eosinophilia by itself need not mandate cessation of drug therapy, if such is medically indicated. Drug-induced blood eosinophilia, however, should prompt an evaluation of whether organs, including the lungs, kidneys, and heart, are involved in the eosinophil-associated drug reaction. If organ involvement develops, cessation of drug administration is necessary.

Some cytokines are potential causes of eosinophilia. Granulocyte–macrophage colony-stimulating factor (GM-CSF), but not granulocyte colony-stimulating factor (G-CSF), stimulates eosinophilopoiesis and can cause prominent blood and tissue eosinophilia and, less commonly, eosinophil-associated diseases, including eosinophilic pneumonia and eosinophilic endomyocardial fibrosis. Administration of IL-2 or IL-2-stimulated lymphocytes frequently elicits eosinophilia, most likely due to production of IL-5. Eosinophilic myocarditis and endocardial thrombosis may complicate high-dose IL-2 therapy.

Diverse agents, including many antimicrobial agents and non-steroidal anti-inflammatory agents (**NSAIDs**), may elicit pulmonary eosinophilia. Blood eosinophilia is usually, but not always, present; and if blood eosinophilia is absent, sputum or bronchoalveolar lavage eosinophilia is necessary to help make the diagnosis.

In drug-induced acute interstitial nephritis, eosinophilia is common in the involved kidneys, urine, and at times, the blood. In addition to eosinophilia, fever, rash, and arthralgia support the diagnosis, but these are commonly absent in cases of drug-induced acute interstital nephritis. Agents that elicit acute interstitial nephritis include semisynthetic penicillins, NSAIDs, cimetidine, sulphonamides, ciprofloxacin, and aztreonam. Eosinophiluria is not uniformly present in all with drug-induced interstitial nephritis.

Acute necrotizing eosinophilic myocarditis is a serious but uncommon type of hypersensitivity myocarditis, with reactions to medications, such as ranitidine or penicillin, responsible in some cases. A syndrome of hepatitis with eosinophilia can be a manifestation of drug reactions, including to minocycline, ranitidine, sulpha antibiotics, and trovafloxicin. Other medication-related eosinophilic responses include drug-induced hypersensitivity vasculitis and forms of gastroenterocolitis elicited by medications, including clozapine and NSAIDs. Adverse reactions to contaminated L-tryptophan in 'natural' medications has previously caused a widespread development of the eosinophilia–myalgia syndrome.

Immunological disorders

The hyper-IgE syndrome is characterized by recurrent staphylococcal abscesses of the skin, lungs, and other sites, pruritic dermatitis, hyperimmunoglobulinaemia E, and blood, sputum, and tissue eosinophilia. Eosinophilia is characteristic of Omenn's syndrome, combined immuno-deficiency with hypereosinophilia.

Eosinophil infiltration accompanies rejection of lung, kidney, and liver allografts. Tissue and blood eosinophilia occur early in the rejection process, and eosinophil counts and granule protein levels have correlated with prognosis, severity, and response to rejection therapy.

Myeloproliferative and neoplastic diseases

The hypereosinophilic syndrome is considered below. Eosinophilia may accompany chronic myelogenous leukaemia and the M4Eo subtype of acute myelogenous leukaemia. Blood eosinophils may be elevated in the nodular sclerosing form of Hodgkin's disease. Some patients with carcinomas, especially of mucin-producing epithelial cell origins, have blood eosinophilia. Eosinophilia may accompany angio-immunoblastic lymphadenopathy, mycosis fungoides, Sézary's syndrome, lymphomatoid papulosis, and systemic mastocytosis.

Pulmonary syndromes

Diverse eosinophilic pulmonary syndromes are noted in Table 1.

Skin and subcutaneous diseases

Various cutaneous diseases can be associated with a heightened level of blood eosinophils (Table 1). In episodic angio-oedema with eosinophilia, recurrences are marked by prominent blood eosinophilia, significant angio-oedema, at times with excessive weight gain due to fluid retention, and less frequently by fever.

Gastrointestinal diseases

Eosinophilia is common with eosinophilic gastroenteritis, and tissue eosinophils are found in inflammatory bowel diseases and collagenous colitis.

Rheumatological diseases

The principal eosinophil-related vasculitis is the Churg–Strauss syndrome. Cutaneous necrotizing eosinophilic vasculitis with hypocomplementaemia and eosinophilia, a distinct vasculitis of small dermal vessels which are extensively infiltrated with eosinophils, may occur in patients with connective tissue diseases. Eosinophilia may uncommonly accompany rheumatoid arthritis itself but is more commonly due to adverse reactions to medications or concomitant vasculitis.

Endocrine diseases

Loss of normal adrenoglucocorticosteroid production causes increased blood eosinophilia.

Other disorders

The syndrome of atheromatous cholesterol embolization can be associated with eosinophilia and eosinophiluria. Uncommonly, kindreds with hereditary eosinophilia have been recognized. Irritation of serosal surfaces, as in eosinophilic pleural effusions and peritoneal, and at times blood, eosinophilia developing during chronic peritoneal dialysis, can be associated with eosinophilia.

Idiopathic hypereosinophilic syndrome

Eosinophilia may be associated with a variety of clinical disorders, including common allergic diseases, infections with helminthic parasites, and obvious neoplastic disease. However, in other patients eosinophilia that is pronounced, prolonged, and not associated with identifiable aetiologies has been classified as the idiopathic hypereosinophilic syndrome (HES), in which the heart is the most commonly affected organ due to the development of endomyocardial damage. With time, however, it has become clear that HES includes a clinically heterogeneous and aetiologically diverse group of eosinophilic disorders, with clonal abnormalities in T lymphocyte subsets or eosinophils themselves recognized in some HES patients. Thus, there is a range of syndromes that present with hypereosinophilia and may share some clinical features: as yet these are not fully delineated or differentiated.

Definition

HES is increasingly recognized not to be a single entity but rather a constellation of leucoproliferative disorders, each characterized by sustained overproduction of eosinophils. The three original defining criteria for this syndrome, identified by Chusid and colleagues, need to be expanded to encompass increasing clinical experience with varied eosinophilic syndromes. Contemporary criteria include:

1. Eosinophilia in excess of 1500/μl of blood, persisting for longer than 6 months.
2. Lack of an identifiable parasitic, allergic, or other aetiology for eosinophilia. Thus, varied eosinophil-associated diseases need to be sought and excluded. Amongst the potential parasitic aetiologies of eosinophilia it is especially important to exclude *Strongyloides stercoralis*, which may persist for decades and be difficult to diagnose solely by stool examinations, not only because of its capacity to cause marked eosinophilia mimicking HES, but also because it, unlike other helminthic causes of marked eosinophilia, can develop into a disseminated, often fatal, disease (hyperinfection syndrome) in patients given immunosuppressive corticosteroids.
3. Absence of other idiopathic eosinophilic syndromes clinically distinct from HES. This criterion was not included in the initial analysis of HES, but is needed to help exclude several other defined eosinophilic syndromes whose aetiologies remain unknown. These would include Churg-Strauss vasculitis (see Chapter 17.11.5) as well as syndromes that involve only limited organ systems (e.g., eosinophilic gastroenteritis, eosinophilic pneumonias) and do not have a propensity to cause eosinophil-associated damage to tissues outside their primary target organs. By this criterion, the syndrome of episodic angioedema with eosinophilia would be distinguished from HES.

4. Evidence by symptoms and signs of organ involvement. Not all patients with prolonged eosinophilia develop organ involvement and many have benign courses. These patients are often not reported or subjected to evaluation at referral centres due to the absence of eosinophil-associated disease. Blood eosinophilia *per se* does not warrant therapy in the absence of evidence of concomitant organ involvement.

Aetiology

The current diagnostic criteria for HES encompass a diversity of eosinophilic disorders of varying, and as yet often unknown aetiologies. Clonal abnormalities in the eosinophil lineage have been detected uncommonly: these are based on analyses of X-linked polymorphisms and hence applicable only to the minority of women with HES, but they do indicate that in some patients HES may be a manifestation of chronic eosinophilic leukaemia. Some patients with HES develop blast crises similar to those of chronic myelogenous leukaemia, or they evolve into chronic myelogenous leukaemia-like diseases. Included in the differential diagnosis of neoplastic causes of eosinophilia are chronic myelogenous leukemia with eosinophilia, the M4EO variant of acute myelogenous leukaemia, and acute lymphocytic leukaemia with eosinophilia. Some patients with the typical clinical and haematological features of HES have subsequently developed T cell lymphomas or acute lymphoblastic leukaemia. However, chromosome studies are normal in most patients with HES, with abnormal findings in only 1 of 18 in a British series and 8 of 33 patients at the National Institutes of Health.

Other potential aetiologies for HES might include dysregulated production of eosinophilopoietic cytokines, such as interleukin 5 (IL-5), IL-3, or granulocyte-macrophage colony stimulating factor (GM-CSF). In some patients with HES the disorder has been correlated with either clonal expansions of $CD3^{-}CD4^{-}$ or $CD4^{-}CD3^{+}CD8^{+}$ Th2-like lymphocytes, or other aberrant T cells or NK cells elaborating IL-5. A case with eosinophilia associated with polyclonal expansion of activated $CD3^{+}$ T cells expressing NK cell markers (CD16 and CD56), associated with IL-2 and IL-15 overproduction, has been reported. In other eosinophilic patients, however, it appears that overproduction of IL-5 is not solely responsible for the eosinophilia. In a kindred with autosomal dominant inheritance of eosinophilia, studies mapped the gene close to, but not involving the IL-3, IL-5, and GM-CSF genes, on chromosome 5q31-q33, suggesting an unidentified gene may regulate eosinophil production. However, for most patients with HES the aetiology of the eosinophilia is not currently understood.

Clinical features

HES is more common in men than women (9:1) and tends to occur between the ages of 20 and 50, although cases have developed in children. The initial manifestations may be due to sudden cardiac or neurological complications, but tend to be more insidious and present over months or longer. Eosinophilia may be detected only incidentally. Other frequent presenting symptoms include tiredness, cough, breathlessness, muscle pains, angioedema, rash, sweating, pruritus, or retinal lesions. Patients with HES do not exhibit a propensity to bacterial or other infections. Weight loss and cachexia are not usually seen. Some patients experience alcohol intolerance with abdominal pain, flushing, nausea, or diarrhoea.

Haematological manifestations

The defining haematological abnormality is sustained eosinophilia. Total leucocyte counts are usually less than 25 000/µl, with between 30 and 70 per cent eosinophils, but extremely high leucocyte counts (>90 000/µl) develop in some patients and are associated with a poor prognosis. Eosinophils in the blood may be mature or less commonly can include numbers of eosinophilic myeloid precursors. Eosinophils often exhibit morphological abnormalities including diminished granule numbers, cytoplasmic vacuolization, and nuclear hypersegmentation. At the ultrastructural level there may be loss of granule contents, either of the crystalline core or matrix of specific granules, fewer and smaller specific granules, increased tubulovesicular structures, and increases in cytoplasmic lipid bodies.

Although not often emphasized, many patients with HES will have an absolute neutrophilia along with their eosinophilia, further contributing to elevations in the white blood cell count. Band forms and less mature neutrophilic precursors may be present in the peripheral blood. Leucocyte alkaline phosphatase levels may be abnormally elevated or decreased. Serum vitamin B_{12} and vitamin B_{12} binding proteins may be normal or elevated. Anaemia is present in about 50 per cent of patients.

Bone marrow findings demonstrate increased numbers of eosinophils, often 30 to 60 per cent, with a shift to the left in eosinophil maturation. Increased numbers of myeloblasts are not usually seen. Myelofibrosis is encountered in a minority of patients. Splenomegaly is found in about 40 per cent of cases.

Cardiac manifestations

In HES, the heart is the most commonly affected organ due to the development of endomyocardial damage leading to a restrictive cardiomyopathy. This distinct form of cardiac involvement may also complicate other varied diseases marked by sustained eosinophilia, including Churg-Strauss vasculitis, eosinophilic leukaemia, eosinophilia with carcinomas or lymphomas, eosinophilia from GM-CSF or IL-2 administration or drug-reactions, and eosinophilia from helminthic infections such as trichinosis, visceral larva migrans, and filariasis. However, many patients with eosinophilia do not develop any evidence of endomyocardial damage; hence in addition to increased numbers of eosinophils, the pathogenesis of eosinophil-mediated cardiac damage probably involves some, as yet ill-defined, activating events that promote eosinophil-mediated endomyocardial damage. Patients with sustained eosinophilia should be monitored by echocardiography for evidence of cardiac disease.

Cardiac damage progresses through three stages, the first involving acute necrosis in the early weeks, the second involving the development of endocardial thrombi over many months, the final stage being the fibrotic stage after a couple of years of disease.

The risks of developing cardiac disease in two series of patients with HES were not related to the extent of eosinophilia or duration of disease. Those who developed evident cardiac disease were more likely to be male and to have splenomegaly, thrombocytopenia, elevated levels of vitamin B_{12}, hypogranular or vacuolated eosinophils, and abnormal early myeloid precursors in their blood. HES patients free of cardiac disease tended to be female and have angio-oedema, hypergammaglobulinaemia, and elevated serum levels of IgE.

Neurological manifestations

Neurological complications may be of three types. The first type is due to thromboemboli originating from the left ventricle, which may occur before cardiac disease is demonstrable by echocardiography and can be the presenting manifestation of HES. The second type of neurological disease is primary central nervous system dysfunction, presenting as an encephalopathy including changes in behaviour, confusion, ataxia, and memory loss, and exhibiting upper motor neurone signs with increased muscle tone, deep tendon reflexes, and a positive Babinski. Impaired cognitive abilities may persist for months. The pathological basis for this form of diffuse central nervous system disease remains unknown. Peripheral neuropathies constitute the third type of neurological dysfunction. Symmetric or asymmetric polyneuropathies manifest by sensory deficits, painful paraesthesiae, or mixed sensory and motor deficits are most common, but mononeuritis multiplex occurs with HES, as do radiculopathies and muscle atrophy due to denervation. Biopsies of affected nerves generally show an axonal neuropathy with varying degrees of axonal loss and no evidence of vasculitis or contiguous eosinophil infiltration.

Cutaneous manifestations

The skin is one of the most frequently involved organs, with cutaneous manifestations occurring in more than 50 per cent of patients. The most common skin manifestations are of two types, either angio-oedematous and urticarial lesions, or erythematous, pruritic papules and nodules. Patients who experience angio-oedema and urticaria are likely to have benign courses without cardiac or neurological complications and either do not require systemic therapy or respond to prednisone alone. Some patients with angio-oedema and eosinophilia are now recognized to have a syndrome, episodic angio-oedema and eosinophilia, that is distinct from HES. Particularly incapacitating mucocutaneous manifestations of HES are mucosal ulcers that may occur in the mouth, nose, pharynx, penis, oesophagus, stomach, and anus. Biopsies demonstrate only a non-specific mixed cellular infiltrate without a prominence of eosinophils and no evidence of vasculitis or microthrombi.

Pulmonary manifestations

Pulmonary involvement is reported in about 40 per cent of HES patients, the commonest respiratory symptom being a chronic, persistent, generally non-productive cough. The basis for this may be sequestration of eosinophils in pulmonary tissues, although most symptomatic individuals have clear chest radiographs. As noted by

Spry, asthma is rare in patients with HES.

Pulmonary involvement in HES may be secondary to congestive heart failure, pulmonary emboli originating from right ventricular thrombi, or primary infiltration of the lungs by eosinophils. Infiltrates may be diffuse or focal without a predilection for any region of the lungs, in contrast to the often peripheral infiltrates in chronic eosinophilic pneumonia (see Chapter 17.11.9). Pulmonary fibrosis may develop over time, especially in those with cardiac fibrosis.

Ocular manifestations

Patients with HES can experience visual symptoms, most commonly blurring. Even in those without visual symptoms, fluorescein angiography demonstrates that over 50 per cent of HES patients have choroidal abnormalities, including patchy and delayed filling, and retinal vessel abnormalities.

Rheumatological manifestations

Arthralgias, large joint effusions, cold-induced Raynaud's phenomenon, and digital necrosis of fingers or toes can occur with HES. Although myalgias are frequent, focal myositis or polymyositis occur only uncommonly.

Digestive system involvement

Gastrointestinal tract involvement can accompany HES, and 20 per cent of patients at some time may have diarrhoea. Eosinophilic gastritis, enterocolitis, or colitis may be present. Pancreatitis and sclerosing cholangitis occur rarely. Hepatic involvement with HES includes chronic active hepatitis and the Budd-Chiari syndrome from hepatic vein obstruction.

Diagnosis

There are no specific diagnostic tests for HES. Other diseases associated with eosinophilia need to be excluded. Bone marrow and chromosomal analyses should be performed, and phenotypic studies of blood lymphocyte subsets should be considered. IgE levels should be ascertained. Echocardiography is indicated to search for evidence of endocardial thrombi or fibrosis.

Treatment

For those eosinophilic patients without organ damage, no therapy need be administered. There is no clear threshold value of blood eosinophilia that predicts organ involvement or damage. For those requiring therapy, prednisolone is the initial agent, administered at 60 mg/day in adults. Those more likely to respond to this treatment alone are patients with angio-oedema, urticaria, or elevated serum IgE levels, and also those who experience prolonged eosinopenic responses to single doses of prednisolone. Patients less likely to respond include those with splenomegaly and those with cardiac or neurological dysfunction at the time of presentation. For those not responsive to prednisolone, daily hydroxyurea is one option, but increasing the currently preferred therapy for HES is interferon-a (1–10 million units/day or 3 times a week).

Medical management of cardiac complications, including arrhythmias and congestive heart failure, are important and effective measures in the longer-term management of HES, as is surgical replacement of damaged valves. Although early reports emphasized the mortality due to this disorder, many of the deaths were due to congestive heart failure and complications of endomyocardial damage. If the sequelae of organ damage, especially to the heart, can be managed, the hypereosinophilic syndrome can have a prolonged course in many patients.

Further reading

Bain BJ (2000). Hypereosinophilia. *Current Opinions in Hematology* **7**, 21–5. [A contemporary discussion of eosinophilic syndromes, especially as they relate to leukaemias.]

Chusid MJ, *et al.* (1975). The hypereosinophilic syndrome: analysis of fourteen cases with review of the literature. *Medicine (Baltimore)* **54**, 1–27. [An early analysis of this syndrome.]

Davies J, *et al*. (1983). Cardiovascular features of 11 patients with eosinophilic endomyocardial disease. *Quarterly Journal of Medicine* **52**, 23–39.

Gleich GJ, *et al.* (1984). Episodic angioedema associated with eosinophilia. *New England Journal of Medicine* **310**, 1621–6. [Delineation of a syndrome clinically distinct from HES.]

Guillevin L *et al*. (1999). Churg–Strauss syndrome. Clinical study and long-term follow-up of 96 patients. *Medicine (Baltimore)* **78**, 26–37. [A compromise review of the major eosinophil-associated vasculitis.]

Lim K, Weller PF (1998). Eosinophilia and eosinophil-related disorders. In: Adkinson NF, Jr *et al*., eds. *Allergy: principles and practice*, 5th edn, pp 783–98. Mosby, St. Louis. [A thorough review of the clinical disorders associated with eosinophilia.]

Lin AY, *et al.* (1998). Familial eosinophilia: clinical and laboratory results on a U.S. kindred. *American Journal of Medical Genetics* **76**, 229–37. [An analysis of a kindred with autosomal dominant inheritance of hypereosinophilia.]

Ommen SR, Seward JB, Tajik AJ (2000). Clinical and echocardiographic features of hypereosinophilic syndromes. *American Journal of Cardiology* **86**, 110–3. [A contemporary update on the value of echocardiographic evaluations of potential eosinophil-mediated endomyocardial damage.]

Roufosse F, *et al.* (2000). Clonal Th2 lymphocytes in patients with the idiopathic hypereosinophilic syndrome. *British Journal of Haematology* **109**, 540–8.

Spry CJF, Davies J, Tai PC, Olsen EG, Oakley CM, Goodwin JF (1983). Clinical features of 15 patients with the hypereosinophilic syndrome. *Quarterly Journal of Medicine* **52**, 1–22.

Spry CJF (1988). *Eosinophils. A comprehensive review and guide to the scientific and medical literature*. Oxford Medical Publications, Oxford.

Weller PF, Bubley GJ (1994). The idiopathic hypereosinophilic syndrome. *Blood* **83**, 2759–79. [A thorough review of HES.]

Wilson ME, Weller PF (1999). Approach to the patient with eosinophilia. In: Guerrant RL, Walker DH, Weller PF, eds. *Tropical infectious diseases: principles, pathogens and practice*, pp 1400–19. Churchill Livingstone, New York. [Considerations of the aetiologies of eosinophilia with special emphasis on helminth infections.]

22.4.7

Histiocytoses

D. K. H. Webb

Introduction

The histiocytoses are characterized by the infiltration of affected tissues with cells of monocyte/macrophage lineage. A classification subdividing the disorders into three classes has been proposed (Table 1). However, the boundaries between classes I and II may be blurred, and more than one class of disorder may be present in the same child.

Aetiology and epidemiology

Histiocytes are of bone marrow origin, derived by the migration and differentiation of blood monocytes, although local proliferation in the tissues may occur following contact with antigen, and in disease states. Growth and differentiation are controlled by haemopoietic growth factors produced by bone marrow stromal cells, fibroblasts, macrophages, and lymphocytes.

Normal histiocytes are divided into two subgroups: (1) dendritic cells (Langerhans' cells, dendritic reticulum cells, and interdigitating reticulum cells); and (2) macrophages. Langerhans' cells are normally found in the epidermis, the mucosa of the bronchial tree, in lymph nodes, and in thymus. Characteristic features include the expression of the CD1a surface antigen, and the presence of specific cytoplasmic organelles (Birbeck granules). These arise either by invagination of the surface membrane during endocytosis of antigen, or as secretory organelles derived from the Golgi apparatus. CD1a has considerable homology with HLA class I molecules and may have a role in antigen presentation to T cells. Following stimulation by antigen, Langerhans' cells migrate to lymph nodes, where they present antigen to T cells. Dendritic and interdigitating reticulum cells are localized to lymph nodes, where they present antigen to B and T cells, respectively.

Macrophages occur widely throughout the tissues where they have multiple functions in the immune response, wound healing, bone remodelling, haemopoiesis, haemostasis, the secretion of inflammatory cytokines, phagocytosis of particulate matter/antigens, and the release of proteases, antiproteases, and arachidonate metabolites.

Classification

Class I disorders

Langerhans' cell histiocytosis (LCH)

The term 'Langerhans' cell histiocytosis' has been widely adopted to replace the diagnosis 'histiocytosis X', following the recognition that the presence of Langerhans' cells is characteristic of lesions in the disorder. LCH is rare and can occur at any age, although there is a paucity of epidemiological data regarding the disease in adults. It affects 4 per million children each year, with a peak incidence between 1 and 3 years of age, and is considered to be a reactive disorder resulting from immune activation. Searches for potential triggers have been unsuccessful. There is no evidence for a viral aetiology. High levels of cytokines have been demonstrated both in lesions and in the serum of children with multisystem disease. Two studies of X-linked DNA polymorphisms have demonstrated clonality in lesional Langerhans' cells, but these data do not define LCH as a neoplastic disorder. Clonality has been demonstrated in a variety of non-neoplastic disorders. There is a recognized association between LCH and malignancy. A literature review revealed details of 87 LCH-associated malignancies—39 lymphomas, 22 acute leukaemias, and the remainder solid tumours, including secondary tumours arising within fields of previous irradiation. Amongst 341 children registered in two large international treatment trials, the incidence of secondary malignancy was 1 per cent.

Juvenile xanthogranuloma

Juvenile xanthogranuloma usually presents with single or multiple yellow–red skin lesions in newborns and infants—in one series, the median age at presentation in 36 children was 0.3 years (range from birth to 12 years). Histology shows a cutaneous accumulation of lipid-filled macrophages, Touton giant cells, and fibroblasts. Extracutaneous disease may occur in about 10 per cent of patients involving the CNS, liver, spleen, lung, eye, oropharynx, and muscles.

Class II disorders

Haemophagocytic lymphohistiocytosis (HLH)

This is a rare disorder with typical histology showing tissue infiltration by lymphocytes and macrophages, some manifesting haemophagocytosis (Fig. 1 and Plate 1). It appears likely that HLH is due to the abnormal function of T lymphocytes. Tissue infiltrates and haemophagocytosis is probably due to the dysregulated secretion of cytokines, rather than by a primary disorder of macrophages. The disorder occurs in primary and secondary forms. Primary HLH is an autosomal recessive disorder with an incidence of between 1 and 2 cases per million children each year in the United Kingdom and Sweden. Linkage studies indicate the involvement of at least three genes. In about 20 per cent of cases there is a history of previously affected siblings (familial HLH). Parental consanguinity or onset in early infancy are further supportive features.

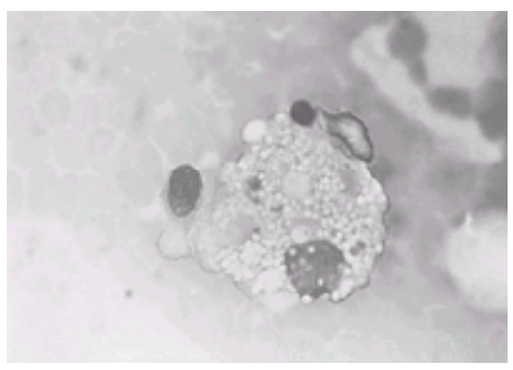

Fig. 1 Macrophage exhibiting haemophagocytosis in the bone marrow of a child with haemophagocytic lymphohistiocytosis. (See also Plate 1.)

Secondary HLH is at least as common as primary disease. Precipitants include viral, bacterial, fungal, or protozoan infections, often in an immunocompromised host. Other precipitants include malignancy, particularly T-cell lymphoproliferative states, autoimmune diseases, and lipid infusions.

Criteria for the diagnosis of HLH are shown in Table 2, although not all features are present in every case. The tissues which are most frequently sampled to substantiate the diagnosis are bone marrow, lymph node, and liver, although fine-needle aspiration of the spleen is reported to have a high diagnostic yield. Diagnostic changes may be difficult to demonstrate, and the bone marrow in particular is hypercellular and reactive in the early stages of the disease. Haemophagocytosis may not be a feature on liver biopsy, but there may be prominent sinusoidal Kupffer cells and lymphoid portal infiltrates similar to those seen in chronic persistent hepatitis.

Sinus histiocytosis with massive lymphadenopathy

This disease was described in 1969 by Rosai and Dorfman as a syndrome of cervical lymphadenopathy with typical histology showing preserved lymph node structure, dilated lymph node sinuses containing mixed inflammatory cells, with vacuolated macrophages manifesting haemophagocytosis and emperipolesis of lymphocytes (that is, a process whereby lymphocytes move straight through the cytoplasm of cells). Fibrosis may be marked. Although most cases are isolated, it has occurred in individuals with malignancy, autoimmune diseases, or other histiocyte disorders, especially LCH.

Class III disorders

Acute myelomonocytic and acute monocytic leukaemia account for 30 per cent of cases of acute myeloid leukaemia. Considerable controversy exists regarding other malignancies of the monocyte/macrophage system, due to difficulties in nosology. Malignant histiocytosis was described as a clinical picture of lymphadenopathy, hepatosplenomegaly, fever, wasting, and pancytopenia, with histology showing tissue infiltration by large cells with copious cytoplasm and irregular nuclei. However, recent studies of cell lineage in such cases indicate that the majority of these tumours are in fact of lymphoid origin, and reclassifiable as anaplastic large-cell (Ki 1-positive) lymphomas. Accordingly, true 'malignant histiocytosis' is an extremely rare entity, requiring careful pathological assessment to substantiate the diagnosis. Both localized and disseminated malignancies of dendritic cells occur, although these are extremely rare.

Clinical features

LCH

Some 60 per cent of children with LCH have single-system disease of skin or bone. The remaining 40 per cent have multisystem disease affecting two or more organ systems. There are inadequate data regarding the patterns of disease in adults, although lung involvement appears common, perhaps due to the association with smoking (see below). However, it is clear that adults may manifest a similar pattern of disease to that seen in childhood.

Skin

The skin rash comprises red or yellow-brown papules to the trunk, erythema in skin folds and behind the ears, and scaling, particularly affecting the scalp. Rarely, young infants manifest a vesicular rash, similar to varicella, which may be present at birth.

Ears

Ear discharge is a classic sign, and may be due either to skin involvement in the external auditory canal, or bone destruction around and in the middle ear, with polyp formation. Such destructive lesions may result in hearing loss, and formal ENT assessment is essential.

Bone

Bone lesions may be occult, but present clinically with pain and soft tissue swelling. They are best seen on plain radiographs as irregular lytic areas sometimes with marked periosteal reaction, most commonly affecting the skull and long bones (Fig. 2). There may be pathological fractures. Involvement of the axial skeleton may result in vertebral collapse and vertebral plana, although spinal cord compression is rare. Orbital disease may cause proptosis, a classical feature, but visual impairment is unusual.

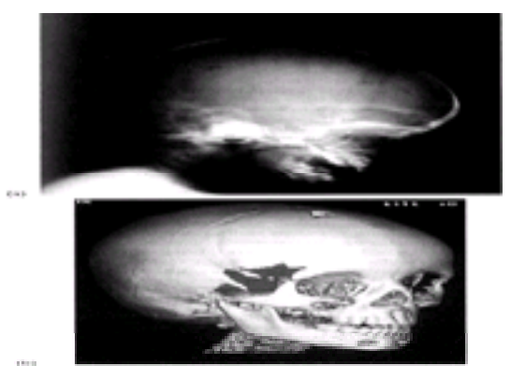

Fig. 2 Plain radiograph and 3D computed tomography scan showing bone lesions of Langerhans' cell histiocytosis in the skull.

Diabetes insipidus

Diabetes insipidus due to involvement of the hypothalamus and pituitary stalk is most common in multisystem disease, reaching 40 per cent in some series. Other risk factors include skull lesions, especially of the orbits or parietal bones. Magnetic resonance imaging (**MRI**) may demonstrate thickening of the pituitary stalk, a suprasellar mass, and/or loss of the posterior pituitary bright signal on *T*2-weighted images due to the absence of vasopressin. Once true diabetes insipidus is established it is irreversible.

Lungs

Lung disease occurs in one-third of children with multisystem disease but is very rare as a single site, it is characterized by cough, chest pain, and tachypnoea, with diffuse micronodular shadowing on chest radiograph. Progression to cyst formation and a honeycomb lung appearance occurs, and hypoxia, pleural effusions, and pneumothorax may occur in advanced disease. Amongst adults, tobacco smoking is a risk factor for pulmonary disease. There may be doubt as to the cause of pulmonary signs, but the diagnosis can be confirmed, and infection excluded, by bronchoalveolar lavage or lung biopsy. As Langerhans' cells are normally present in the bronchial tree, over 5 per cent CD1a-positive cells should be present in lavage fluid to support the diagnosis.

Liver

Hepatomegaly and elevated transaminases may occur without evidence of liver infiltration on biopsy. Jaundice may result from obstruction of the biliary tract by enlarged portal nodes, and therefore is not necessarily diagnostic of liver dysfunction. These provisos emphasize the need for careful assessment of the methods used for clinical findings and of the investigation results in the initial evaluation. Severe liver disease may result in fibrosis, sclerosing cholangitis, and hepatic failure.

Bone marrow and blood

A low haemoglobin due to anaemia of chronic disease is a common finding in active disease. Iron deficiency should be excluded in children with microcytosis and hypochromia. Pancytopenia due to secondary bone marrow infiltration by macrophages and haemophagocytosis may occur. CD1a-positive cells may be demonstrated with monoclonal antibodies by flow cytometry or by the alkaline phosphatase/antialkaline phosphatase technique on slides, but their significance is unclear. Pancytopenia associated with liver disease and splenomegaly carries a poor prognosis.

Gut

Gut involvement with vomiting, diarrhoea, malabsorption, and protein-losing enteropathy occurs in under 5 per cent of children, and requires full investigation, including adequate biopsy. Although infiltrates may be seen in the mucosa and submucosa, biopsy of the muscle wall may be required. Barium studies may reveal alternate dilated and stenotic segments throughout the intestine. Mandibular and maxillary disease may result in floating teeth, and there may be gingivitis and buccal ulceration.

Central nervous system

Disease in the central nervous system, excluding diabetes insipidus, occurs in around 4 per cent of cases, and typically affects the cerebellar and cerebral white matter, with ataxia, dysarthria, nystagmus, and cranial nerve palsies. CNS disease usually develops around 5 years from the original presentation. The mechanism of CNS disease is unknown. Most cases occur in individuals with multisystem disease, but also in the setting of single-system bone disease especially of the skull, and in children with diabetes insipidus. On imaging, several patterns are seen:

- poorly defined changes in the white matter of the cerebellum, cerebrum, and basal ganglia—biopsy in these cases shows perivascular and parenchymal infiltrates of macrophages and lymphocytes associated with oedema and demyelination;
- well-defined lesions in the white and grey matter;
- hypothalamic–pituitary involvement including suprasellar masses;
- extraparenchymal masses, generally not in continuity with skull lesions, which on biopsy comprise xanthomatous histiocytes, lymphocytes and Touton giant cells similar to those found in juvenile xanthogranuloma.

Lymph nodes

Lymphadenopathy occurs in both single- and multisystem disease, and may be gross. Local pressure effects may cause obstruction of the airways, vasculature, or biliary tree, and discharging sinuses may form to the overlying skin. Involvement of the thymus may be detected on chest radiography, and may be present on tissue examination even without enlargement of the organ.

HLH

Clinical manifestations include fever, splenomegaly, hepatomegaly, lymphadenopathy, pancytopenia, abnormal liver function, coagulopathy, and signs and symptoms referable to the central nervous system. Occasionally CNS involvement has been the only evidence of disease at presentation. Initial blood changes may show anaemia or thrombocytopenia, with the development of pancytopenia as the disease progresses. Other features include high fasting triglyceride levels, low fibrinogen, mononuclear pleocytosis and increased protein in the cerebrospinal fluid, high serum ferritin, and reduced or absent natural killer-cell function. Many of these changes result from immune activation with cytokine production. Involvement of the central nervous system varies from asymptomatic cerebrospinal fluid pleocytosis (usually to moderate levels comprising lymphocytes and macrophages) to symptomatic disease with encephalitis, abnormal head movements, fits, cranial nerve palsies, ataxia, regression of developmental milestones, and coma. In children who have died with CNS disease, histology of the brain shows oedema, softening and destruction of tissue, perivascular, parenchymal, and leptomeningeal infiltrates, with necrosis and destruction, especially of white matter.

Sinus histiocytosis with massive lymphadenopathy (SHML)

The lymphadenopathy may be gross. It is often painless, and may wax and wane with time. Involvement of cervical lymph nodes is present in most cases, but other groups are affected, either jointly or alone. Extranodal disease of the head and neck or a variety of other sites is present in almost half of cases, either alone or in association with lymphoid masses. Other features of SHML include: systemic ill health with fever and weight loss; destructive infiltrates in skin, bone, and other extranodal sites; hypergammaglobulinaemia; an elevated erythrocyte sedimentation rate; reactive leucocytosis; and immune dysfunction including autoimmune anaemia and neutropenia.

Management of histiocyte disorders

LCH

Initial work-up requires confirmatory biopsy and an accurate assessment of the extent of disease. Identification of Langerhans' cells within the lesional inflammatory cell infiltrate, with demonstration of either the CD1a surface antigen on immunohistochemistry or the presence of Birbeck granules on electron microscopy, is recommended. Thorough physical examination and investigations are required to determine the extent of disease, and investigations must include a full blood count, liver function tests, serum albumin, a coagulation screen, paired urine and plasma osmolalities, and skeletal survey by plain radiographs. Further investigations should be guided by the need to explain specific symptoms and signs. It is important to carefully assess the function of affected organs, as dysfunction carries prognostic significance (see below).

The majority of cases of LCH eventually resolve spontaneously. No therapy is uniformly effective. Approaches to treatment have varied, with particular controversy regarding the role of chemotherapy for children with multisystem disease. Deaths occur in 10 to 15 per cent of cases, and are largely restricted to children with organ dysfunction. For most cases the primary objectives are control of symptoms, and limitation of long-term disability, which affects 50 per cent of those with extensive disease. For children with skin disease requiring therapy, topical application of corticosteroids may prove beneficial. For those with more severe skin involvement, topical mustine has proved highly effective—in one study rapid improvement within 10 days occurred in each of 16 children, with complete healing in 14. Some children require systemic therapy. Oral corticosteroids result in improvement in over half of cases. Bone lesions may resolve following biopsy, but further local therapies include curettage or injected steroids. Radiotherapy is now uncommon due to concerns over late effects—in particular, secondary malignancies have been described within the radiation field. Systemic therapy is indicated for children with multifocal bone disease (30 per cent of all children with bone disease), or single, symptomatic lesions unsuitable for local therapy. Indomethacin may be effective, but it is unclear whether the drug provides more than symptomatic relief. Further lesions and reactivations of initially responding lesions develop in up to one-third of patients.

Children with multisystem disease may be managed conservatively with systemic therapy reserved for those who have organ dysfunction, pain, systemic upset, or failure to thrive. This issue is controversial. There are claims for better results if initial treatment employs relatively intensive chemotherapy, especially in regard to response rates and late effects. There is no evidence that a more-intensive therapy reduces the 40 per cent mortality rate in children with organ dysfunction. For children who require systemic treatment, the accepted agents are prednisolone, vinblastine, and etoposide in varying combinations. The uncertainty regarding the most effective and least toxic approach to treatment, together with the rarity of LCH, emphasizes the need for international collaborative randomized studies. Hepatic failure due to LCH has been successfully treated by orthoptic liver transplantation—80 per cent of children reported in the literature were alive at a median follow-up of 3 years. For children who fail first-line treatment, further options are limited. There is little evidence that immune modulation with cyclosporin or antilymphocyte globulin, or bone marrow transplantation are effective. Responses have been obtained with cladribine (2-chlorodeoxyadenosine). Further studies are underway.

LCH is associated with a wide range of potential late effects, particularly skeletal deformity and dysfunction, diabetes insipidus, growth hormone deficiency, ataxia,

intellectual impairment, and lung and liver fibrosis. These data stress the need for therapeutic trials to include standardized assessments for late effects which can be compared between treatments.

Juvenile xanthogranuloma

Spontaneous resolution is usual, and no treatment has been proven to be undoubtedly beneficial. Excision may be indicated for lesions resulting in complications.

Class II disorders

HLH

It can be difficult to determine whether a patient has primary or secondary disease, particularly in children with evidence of recent viral infection. Amongst 93 children with HLH reported to the Histiocyte Society registry, there was no difference in outcome between 40 children with and 53 children without evidence of viral infection. These data emphasize the overlap between these disorders, and the need for circumspection in determining the best approach to therapy in each case.

There are two standard approaches to treatment for primary HLH, one using etoposide and corticosteroids, and the other antithymocyte globulin, corticosteroids, and cyclosporin. Around 80 per cent of patients respond, but eventual disease recurrence is usual in primary HLH unless the child receives an allogeneic bone marrow transplant. Inadequate disease control, or reactivation, may also occur in some secondary cases. These children should then be treated in line with strategies for primary disease. Adequate control of CNS involvement is very important, but the role of routine intrathecal methotrexate in the control of CNS disease is controversial. Cranial radiation is no longer recommended. Experience with bone marrow transplantation is greatest using matched sibling donors, with around 80 per cent of children remaining disease-free. Increasing use of alternative donors, including matched and mismatched unrelated donors has demonstrated that similar results may be achieved by this approach. Full engraftment following transplantation is not a prerequisite for cure, as low levels of donor T cells may provide adequate disease control. A particular worry regarding the use of a sibling donor is that in familial disease there is a 25 per cent risk that the donor will also develop HLH. Continued improvement in the results of alternative donor procedures may make these the treatment of choice. It must also be remembered that some children, usually sporadic cases aged over 2 years at diagnosis, remain well following initial therapy. Bone marrow transplantation is not indicated for this group.

Case reports regarding the prognosis for secondary HLH are conflicting. The appropriate approach for these patients is treatment of the associated condition or removal of immunosuppression, with specific HLH therapy for individuals who fail to improve.

Sinus histiocytosis with massive lymphadenopathy

The natural history of the disorder is chronic with spontaneous resolution over months or years in many cases, although approximately 5 per cent of patients have died due to immune-mediated organ dysfunction, amyloidosis, or infection. Few individuals have died directly due to lymphohistiocytic infiltrates. Therapy with steroids, cytotoxic drugs, particularly vincristine and alkylating agents, or radiotherapy has been variably effective, and is unnecessary in most cases.

Class III disorders

The outlook for monocytic variants of AML has improved considerably over recent years, with around a 50 per cent survival at 5 years from diagnosis in children and young adults following standard chemotherapy regimens. The uncertainty over the pathology of reported cases of malignant histiocytosis clouds the issues regarding therapy. It appears appropriate to treat malignancies of macrophages with AML chemotherapy. Due to the rarity of dendritic-cell malignancies, treatment recommendations are anecdotal, but based on excision with or without adjuvant therapy.

Further reading

Egeler RM, D'Angio G (1998). The histiocytoses. *Hematology and Oncology Clinics of North America* **12**, 2. [Editorial]

Favara BE, *et al.* (1997). Contemporary classification of histiocytic disorders. The WHO committee on histiocytic/reticulum cell proliferations. Reclassification working group of the Histiocyte Society. *Medical Pediatric Oncology* **29**, 157–66.

Henter J-I, *et al*. (1991). Incidence and clinical features of familial hemophagocytic lymphohistiocytosis. *Acta Paediatrica Scandinavica* **80**, 428–35.

Henter J-I, Elinder G, Ost A (1991). Diagnostic guidelines for haemophagocytic lymphohistiocytosis. *Seminars in Oncology* **18**, 29–33.

Ladisch S, Gadner H (1994). Treatment of Langerhans' cell histiocytosis—evolution and current approaches. *British Journal of Cancer* **70**(Suppl xxiii), S41–S46.

Schmidt D (1994). Monocyte/macrophage system and malignancies. *Medical and Pediatric Oncology* **23**, 444–51.

22.5.1 Erythropoiesis and the normal red cell

Anna Rita Migliaccio and Thalia Papayannopoulou

Introduction

Mature circulating red cells are specialized cellular elements of the blood responsible for both the delivery of oxygen to and for the removal of carbon dioxide from all tissues of the body. In the adult, their number is constantly maintained by the balance of two ongoing processes: the destruction of old red cells, mainly in the spleen; and the generation of new red cells within the bone marrow by a process referred to as erythropoiesis.

The generation of new red cells, like other cellular elements, is accomplished through a complex interplay between haemopoietic cells, stromal cells, and the extracellular matrix within the bone marrow microenvironment. Unique to the erythropoietic process is its regulation not only by growth factors produced *in situ* in the bone marrow, but also by circulating erythropoietin (**EPO**), a true 'hormone' produced by the kidneys in the adult. Positive or negative alterations of erythropoietin production, whether acquired or congenital, and/or of its signalling pathway, result in quantitative changes in red cell production—that is to say, anaemia or erythrocytosis.

This chapter will summarize key biological features of normal human erythropoiesis, both in the adult and during embryonic/fetal life, and highlight pathogenetic mechanisms that can lead to perturbations of erythropoiesis.

The erythroid compartment

Erythropoiesis is a highly regulated, multistep process through which 10^{11} functional red cells are generated daily from very few haemopoietic stem cells (1:10^{4}–10^{5} marrow cells). Stem cells, after a series of amplification divisions, generate multipotential progenitor cells, then oligo- and finally unilineage erythroid progenitors, which give rise to morphologically recognizable erythroid precursors and mature red cells (Fig. 1).

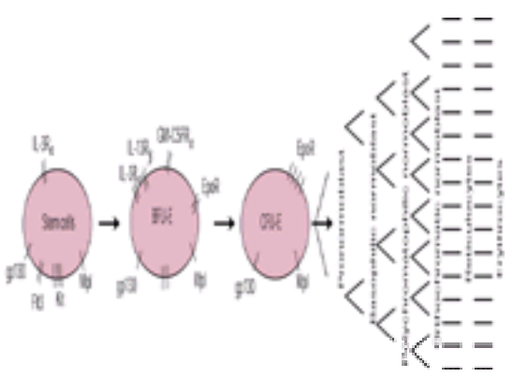

Fig. 1 Adult erythroid progenitor- and precursor-cell compartments. The number of bars on the cell surface is an estimate of the growth factor-receptor concentration/cell responsiveness during erythroid differentiation. A complex of interleukin-6 (**IL-6**) and its soluble receptor (or a fusion molecule called hyper IL-6) may induce growth of burst-forming units-erythroid (**BFU-e**) and erythroid maturation in the absence of exogenously added erythropoietin through activation of an autocrine erythropoietin loop.

Erythroid progenitors

Progenitor cells committed to a specific lineage are defined by their ability to generate colonies of differentiated cells *in vitro*. In these assays, multipotential progenitors generate colonies consisting of cells of several lineages, whereas unilineage cells give rise to colonies of only one lineage. The earliest erythroid progenitor is the burst-forming unit-erythroid (**BFU-e**), which generates colonies containing several thousand erythroid cells. The colony-forming unit-erythroid (**CFU-e**) generates smaller colonies that mature early in culture. Between these two extremes, there exist intermediate classes of BFU-e with less proliferative potential and shorter maturation time in culture. BFU-e tend to have an antigenic profile similar to progenitor cells of other lineages with few exceptions (that is, they are negative for CD45 RA, CD33, and AC133, but express higher levels, or possibly specific isoforms, of the transferrin receptor). CFU-e, in contrast, begin to express more of the markers found specifically in mature erythroid cells (that is, glycophorin A and blood group antigens). The first blood group antigen to be expressed at the BFU-e level is the glycoprotein (**gp**) Kell, followed by the orderly activation of the expression of Rh gp, Landersteiner–Wiener (**LW**) gp, glycophorin A, Band 3, Lutheran gp, and finally, at the erythroblast level, Duffy gp.

Progression of progenitor cells to terminal differentiation is marked not only by the acquisition of phenotypic markers, but also by the acquisition of specific responses to growth factors. The most active of these on early progenitors are stem-cell factor (**SCF**, or Steel factor, or kit ligand), interleukin-3 (**IL-3**), granulocyte/macrophage colony-stimulating factor (**GM-CSF**), and erythropoietin and thrombopoietin (**TPO**) on later progenitors.

The special importance of SCF in erythropoiesis is shown by genetic mutations of SCF (such as *Sl/Sl*^{d}), or of its receptor, kit (such as *W/W*^{v}), that result in mice with anaemia, the severity of which correlates with impairment in kit kinase activity. In humans, heterozygotes with c- *kit* mutations have been reported in individuals with piebaldism, but no homozygotes have been described. Furthermore, mice treated with anti-kit antibodies became anaemic, whereas human kit-antisense containing cultures did not expand the BFU-e compartment.

Erythroid precursors

Erythroid precursors, in contrast to progenitor cells, are morphologically recognizable and include cells at different maturation stages (Fig. 1). The earlier cell is the proerythroblast which, after around 7 days of proliferation and further maturation, gives rise to mature red cells. The maturation process includes haemoglobin synthesis, chromatin condensation (orthochromatic normoblast), and, finally, extrusion of the nucleus and a reduction in cell size (reticulocytes). These morphological changes are paralleled by the accumulation of haemoglobin and by several biochemical changes in the cytoskeleton, which guarantee maximum resistance to stress and flexibility during capillary passage. It has been estimated that about 64 reticulocytes are generated from each pronormoblast, 90 to 95 per cent of which egress

into the bloodstream. Within a day or two in the peripheral circulation, the reticulocytes mature further into red cells.

Ontogeny of erythropoiesis

The ontogenetic development of the erythroid system encompasses a series of well co-ordinated events during embryonic and early fetal life, the timing of which is distinct for each mammal. Erythroid cells are the first differentiated haemopoietic cells to appear during ontogeny, initially in the blood islands of the yolk sac, and later in the fetal liver. The recruitment of new erythropoietic sites during ontogeny (from yolk sac to fetal liver to bone marrow) is accompanied by profound differences in the stem-cell differentiation programme and the phenotypic/functional properties of the erythroid cells being developed in each site. Haemopoietic cytokines, specifically expressed and/or sequestered by each microenvironment, may probably mediate these changes in concert with cell intrinsic transcriptional factors. Furthermore, adhesion receptors on haemopoietic cells and their counter-receptors in the microenvironment ensure patterns of firm adherence, migration, and colonization of the successive haemopoietic sites.

Yolk sac erythropoiesis

In humans, the first erythroid cells (the primitive nucleated erythroblasts) are detected in the yolk sac at 3 to 4 weeks and are the only red cells circulating until week 8 of gestation. They synthesize mainly embryonic haemoglobins, such as Gower I (z_2e_2), Gower II (a_2e_2), and Portland (z_2g_2).

In addition to differentiated primitive erythroid cells, the yolk sac contains progenitor cells of definitive lineage that do not differentiate in the yolk sac but generate BFU-e-like colonies *in vitro*, which are composed of definitive erythroblasts. It is currently undecided whether primitive and definitive progenitor cells derive from the same stem cell.

Gene ablation studies indicate that yolk sac and fetal liver/bone marrow erythropoiesis have clearly distinct growth factor and molecular requirements. Primitive erythropoiesis is not affected by the ablation of several transcription factors (GATA1, AML-1, Rb, Myb, etc.) that affect definitive cells. Also, primitive erythropoiesis *in vivo* is SCF- and erythropoietin-independent, whereas thrombopoietin, but not erythropoietin, is produced *in situ* and may be important for cell survival and partial differentiation. Within the yolk sac, erythroid cells are found in close proximity to endothelial cells and macrophages. Therefore, they are most probably exposed to growth factors produced by these cells, including vascular endothelial growth factor (**VEGF**) and M-CSF. In this regard, yolk sac-derived endothelial cell lines *in vitro* produce leucocyte inhibiting factor (**LIF**), IL-6, flt-3 ligand, SCF, and M-CSF, but not G-CSF, GM-CSF, IL-3, IL-1, erythropoietin, and thrombopoietin. The exchange of trophic signals between endothelial and haemopoietic cells is also mutual at these early stages.

Fetal liver haemopoiesis

The major anatomical site of erythropoiesis in fetal life is the liver, in which newly formed definitive erythroblasts appear at 7 to 8 weeks' gestation within the sinusoidal walls of its parenchyma. Definitive erythroblasts are released into the blood after week 8 and remain the main circulating erythroid cells until birth.

The haemoglobin (Hb) patterns expressed by erythroid cells during ontogeny represent one of the best-studied developmental differences in gene expression. In contrast to primitive erythroblasts, the definitive cells synthesize mainly fetal haemoglobin (Hb F, a_2g_2), together with a small component (10–15 per cent) of adult haemoglobin (Hb A, a_2b_2). The fetal pattern of haemoglobin expression remains stable until 30 weeks' gestation, when a progressive increase in Hb A and a parallel slow decline in Hb F begins. At birth, approximately equal amounts of Hb A and Hb F are synthesized, with the final adult erythroid pattern (adult Hb with <1 per cent fetal Hb) being reached a few months after birth. The fetal to adult Hb switch is strictly related to gestational age and occurs within the same population of cells that undergo an intrinsic modification of its gene expression programme. Its molecular mechanism involves the formation and activation of specific transcriptional complexes within cells at specific stages of ontogeny, but details remain elusive.

Fetal erythroid progenitor cells display a unique phenotypic and functional profile, including higher cell-cycling rates (>30 per cent versus the adult rate of 10 per cent), a lower doubling time (20 h versus the 32 h for the adult cells), and faster kinetics of *in vitro* differentiation (10 days versus 16 days). The average telomeric length is also different (12.8 ± 0.35 kb versus 8.4 ± 0.3 kb). Human fetal BFU-e are exquisitely sensitive to erythropoietin, which is sufficient to sustain their maximal differentiation, whereas SCF complemented by IL-6 alone is sufficient for their maximal *ex vivo* expansion. Whether fetal progenitors do not truly require any other factors, or whether they, unlike adult cells, do not produce autocrine growth inhibitors (transforming growth factor-b (TGF-b)) or are insensitive to them, is unclear.

Unique to the fetal liver stage of haemopoiesis is its almost exclusive erythroid output, whereas its microenvironment is less conducive to myelomonocytic cell differentiation despite the abundant presence of their progenitors.

At the end of fetal development, liver erythropoiesis is suppressed and the organ enters the adult hepatic phase. Oncostatin-M, produced in the liver by haemopoietic CD45+ cells, may facilitate this transition by inhibiting the growth of haemopoietic progenitors while promoting the growth of the hepatocytes expressing its receptor. Increasing glucocorticoid concentrations in fetal liver near term may also contribute to the suppression of liver erythropoiesis.

Bone marrow erythropoiesis

The final site of erythroid cell production maintained throughout life is the bone marrow. Haemopoiesis in the marrow appears between 6.6 and 8.5 weeks after the establishment of a rudimental stroma of cartilagenous and endothelial cells, and is accomplished in four separate phases. Within the bone marrow, granulopoiesis predominates during all the stages of development (fetal and adult).

Erythropoietin and the regulation of erythropoiesis

Erythropoietin was the first haemopoietic growth factor to be identified. Its existence was hypothesized in 1906 by Carnot and Deflandre, but formal proof was obtained by Reissman in 1950 and by Stohlman *et al*. in 1954. Human erythropoietin, a 33- to 38-kDa sialoglycoprotein, was purified to homogeneity in 1977 and its gene, localized on chromosome 7q11, was cloned in 1984. Erythropoietin was also the first growth factor to be used in the treatment of patients because of the clear-cut relationship between its concentration in the blood and the numbers of red cells in the circulation.

Erythropoietin production

Erythropoietin has a short half-life (<5 h; 90 per cent of erythropoietin is rapidly degraded by the liver and 10 per cent secreted in urine), but its blood concentration (0.02 units/ml) is kept constant by continuous production. Erythropoietin levels are exquisitely regulated by changes in O_2 tension, and the kidney is in the ideal anatomical position for sensing these changes (Fig. 2). The kidney is the major producer of erythropoietin, although the liver, which is the main source of erythropoietin in the fetus, retains some capacity of low hypoxia-sensitivity in the adult. Low levels of erythropoietin are also produced in the marrow by the erythroid progenitors themselves.

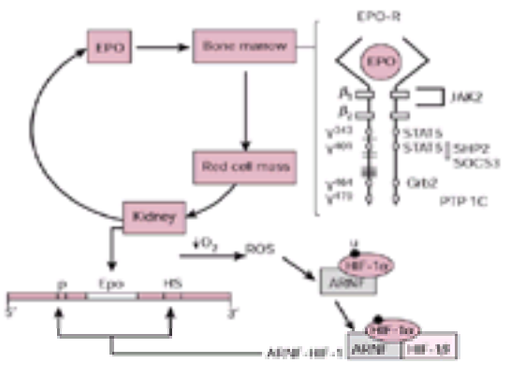

Fig. 2 The regulation of erythroid cell mass. The number of circulating red cells is regulated by the levels of erythropoietin (**EPO**) produced by the kidney under the exquisite control of the hypoxia-sensing machinery. Several pathways are involved in hypoxic recognition and respond by producing reactive O_2 species (**ROS**). The

purpose is to rescue (by ubiquitination) hypoxia-induced factor-1a (**HIF-1a**) from degradation, to facilitate its complex formation with aryl-hydrocarbon nuclear translocator (**ARNF**) and with the ubiquitously expressed b subunit of HIF-1. ARNF–HIF-1 activates the expression of several hypoxia-responsive genes such as *EPO*, glucose transporters, glycolytic enzymes, platelet-derived growth factor (**PDGF**), and vascular endothelial factor (VEFG). (ARNF ablation results in embryonic lethality attributed to the loss of induction of VEGF). ARNF–HIF-1 activates the expression of hypoxia-inducible genes by binding to specific cognate sequences, that, in the case of *EPO*, are present both in its promoter (indicated as p, a typical TATA less promoter) and in its DNA-hypersensitive site (**HS**), which has the function of a hypoxia-inducible enhancer. The enhancer is activated by the binding of a protein complex formed by HIF-1 itself, by the constitutively expressed hepatic nuclear factor-4 (**HNF-4**), and by the general transcriptional activator p300. Of note, the EPO promoter and enhancer also contain binding sites for the steroid hormone receptor, closing the bridge between activation of these receptors and control of the haematocrit level. Another link between a response to stress and EPO expression is provided by p38a, a member of the mitogen-activated protein (**MAP**) family that may be involved in cobalt-induced stabilization of HIF-1a and induction of EPO expression, at least in the fetal liver. EPO induces its effects in the marrow by binding to a specific receptor (**EPO-R**) present on erythroid cells. EPO/EPO-R binding induces tyrosine autophosphorylation and the activation of JAK2 (Janus kinase-2) which is physically associated to its cytoplasmic Box 1 (b1) domain. The activation of the JAK2 catalytic domain is responsible for the phosphorylation of several other proteins, including the EPO-R itself (phosphotyrosine residues involved in docking EPO-R with its transducer proteins are indicated as filled circle, not to scale), STAT5 (signal transducer and activator of transcription-5), and proteins involved in the Ras and PKC (protein kinase C) signal-transduction pathways. On the other hand, at least three other pathways are involved in bringing the receptor complex back to its resting configuration after activation: the phosphatase PTP-1C (protein tyrosine phosphatase-1C) /SHP2 (Src-homology-phosphatase 2), SOCS3 (suppressor of cytokine signalling-3) (overexpression of SOCS3 in transgenic mice results in embryonically lethal anaemia), and the proteasomes, which downregulate the number of activated receptors on the cell surface. PTP-1C becomes physically associated through its SH2 domains with Y479 after EPO stimulation, and SOCS3 becomes associated with the Y401 of EPO-R and with JAK2. The narrow bars on the left side of the EPO-R diagram indicate mutations identified in humans that result in familial erythrocytosis.

As the kidney is the primary site of erythropoietin production, it is not surprising that irreversible damage to this organ results in low or no erythropoietin production and consequently results in anaemia. Anaemia in chronic inflammatory states, such as rheumatoid arthritis, is due instead to the fact that proinflammatory cytokines, such as IL-1 and tumour necrosis factor (**TNF**), inhibit erythropoiesis both directly (by inhibiting the proliferation of the progenitor cells) and indirectly (by inhibiting erythropoietin synthesis by the kidney). At the other extreme, some kidney cancers, by increasing the number of erythropoietin-producing cells, also increase erythropoietin production and then lead to secondary erythrocytosis. Furthermore, disorders that impair O_2 delivery to the tissue are sensed as hypoxia and are also associated with secondary erythrocytosis, such as in chronic lung diseases and congenital heart anomalies. In haemoglobin mutations, because there is inefficient O_2 delivery to the tissues due to altered Hb/O_2 affinity, hypoxia is sensed and causes increased erythropoietin serum levels and secondary erythrocytosis. There are more than 50 different mutations in which changes in either the a- or b-globin gene result in increased Hb/O_2 affinity. Secondary autosomal recessive erythrocytosis (polycythaemias) can also be caused by abnormalities in enzymes such as 2,3-diphosphoglycerate (**2,3DPG**) mutase, involved in the regulation of tissue O_2 delivery (familial 2,3-bisphosphoglycerate deficiency). Secondary asymptomatic polycythaemias are also seen in methaemoglobinaemias. Methaemoglobin is a derivative of haemoglobin in which Fe^{2+}, which binds O_2 reversibly, is replaced by its oxidized form (Fe^{3+}), which binds O_2 irreversibly.

Genetic abnormalities in the O_2-sensing system (summarized briefly in Fig. 2) have not yet been described. Mice lacking hypoxia-inducible factor (HIF)-1a are viable, but are unable to increase their hematocrit levels in response to hypoxia, suggesting that mutations may eventually be found which are not likely to be lethal. In this regard, a congenital polycythaemia common in the Chuvashia region of Russia is characterized both by increased erythropoietin responsiveness of the erythroid cells and by increased erythropoietin concentration in serum. This polycythaemia is not associated with mutations in either *EPO* or *EPO-R* and represents a good candidate for a defective O_2-sensing apparatus such as mutations in the *HIF-1a* gene. Examples of acquired polycythaemias due to defects of the hypoxia-sensing mechanism are renal and neuronal cancers associated with mutations in the von Hippel-Lindau (*VHL*) gene. The product of this gene is a member of the protein complex responsible for the stabilization/degradation of HIF-1a. *VHL* mutations result in overexpression of VEGF and of other O_2-regulated genes.

Erythropoietin signalling pathway

Erythropoietin triggers its biological effect by binding to a specific 64- to 78-kDa (depending on its glycosylation degree) receptor, EPO-R, encoded by a gene on human chromosome 19p. This gene is expressed in erythroid, megakaryocytic, and endothelial cells. Also, marrow cells express high levels of soluble EPO-R species. These may be involved in the fine tuning of erythropoietin concentrations in specific marrow niches, thus allowing erythroid versus myeloid development. In addition to the full-length EPO-R (EPO-RF), immature erythroid cells express a truncated form of the receptor (EPO-RT) that acts as a dominant negative regulator of the EPO-RF mediated signals in mice.

The concentration of EPO-R on the surface of erythroid cells is roughly proportional to their erythropoietin responsiveness *in vitro* (Fig. 1). EPO-R is first detected on BFU-e (around 300 high-affinity erythropoietin binding sites per cell). Its expression increases as the cells progress to CFU-e and proerythroblasts (about 1100 high-affinity sites), but is virtually absent on late normoblasts and reticulocytes. The increase in EPO-R expression observed during erythroid differentiation is also accompanied by a decreased expression of receptors for early-acting growth factors, such as SCF and IL-3 (Fig. 1).

The rapid dimerization induced by erythropoietin binding to its receptor triggers a conformational change that results in autophosphorylation and activation of the kinase catalytic domain of JAK2, a member of the Janus kinase family (Fig. 2). JAK2 phosphorylates and activates several proteins, some of which are responsible for further transmitting the signal to the nucleus, while others, such as the protein tyrosine phosphatases (PTP)-1C (also called HCP or SHP) and -1D (or Syp), dephosphorylate the receptor complex, bringing it back to its resting configuration (Fig. 2). PTP-1C is physically associated with the carboxyterminal domain of EPO-R, a region that is deleted in all of the eight different mutations of *EPO-R* found to be associated with congenital erythrocytosis. 32D cells transfected with one such mutant receptor required erythropoietin to activate the JAK2/STAT5 pathway, but in these cells the activation lasted longer than in cells transfected with the normal receptor. This may explain why congenital polycythaemia/erythrocytosis is characterized by hyper-responsive, erythropoietin-dependent, CFU-e growth. Not all mutations of *EPO-R* cause polycythaemia, however. Two additionally described single-point *EPO-R* mutations are not accompanied by functional consequences or changes in hematocrit.

Signalling through EPO-R triggers several cellular responses, including inhibition of apoptosis, stimulation of cell proliferation, and the induction of expression of erythroid specific genes—that is, the globin genes. The best-studied effect of erythropoietin on erythroid cells is suppression of apoptosis. In fact, mice with targeted deletions of either *Epo*, of its receptor, or of its immediate signal-transduction protein JAK2, die prenatally because of profound anaemia. The livers of these fetuses contain normal numbers of BFU-e and CFU-e, but very few erythroblasts, all of which display signs of apoptosis. These results suggest that these defects impair the final stages of erythroid differentiation by interfering with the erythroblast's ability to survive. Erythropoietin/EPO-R signalling inhibits apoptosis through suppression of caspases (cysteine proteases with aspartate specificity). In erythroid cells, these degrade the erythroid-specific transcription factor GATA1, which is responsible not only for the activation of all the erythroid-specific genes analysed so far, but also of the antiapoptotic protein Bcl-xL. An indirect confirmation of the role of GATA1 in preventing apoptosis of erythroid cells is provided by a recent mutation described in the DNA-binding domain of *GATA1* causing anaemia and thrombocytopenia. As *GATA1* is on the X chromosome, it is possible that other X-linked anaemias or thalassaemia may be due to *GATA1* mutations yet to be identified.

Since survival is a cellular function which is dominant over proliferation, the phenotype of the deletion *Epo*/*Epo-R* mutant mice cannot exclude that erythropoietin may also play a role in the control of proliferation (or differentiation) at later stages of the erythroid differentiation. BFU-e are at least partially dependent on erythropoietin for proliferation and differentiation *in vitro*, but are insensitive to endogenous erythropoietin levels, and neither sustained anaemia nor hypertransfusion alters their frequency and cell-cycle characteristics. In contrast, both the number and cycling characteristics of CFU-e are increased three- to sixfold over the control value. Furthermore, fetal murine CFU-e which constitutively express either Bcl-2 or Bcl-xL survive but do not form colonies in the absence of erythropoietin.

Although the effects of erythropoietin on cell proliferation are not as well characterized as its effects on apoptosis, it can be expected that defects in the proliferation pathway may also result in altered red cell output. In this regard, polycythaemia vera is a human myeloproliferative disorder caused by an acquired presumed mutation at the stem-cell level that results in the formation of CFU-e-derived colonies in the absence of erythropoietin. A familiar autosomal dominant predisposition towards acquiring polycythaemia vera between the ages of 50 and 60 years of age has been described. Since, polycythaemia vera, in contrast to familial polycythaemia, may evolve into leukaemia, it is possible that its defect involves alterations in the proliferative control of erythroid cells.

Red cell homeostasis

Normal red cells have a finite lifespan of 120 ± 20 days. With red cell ageing, metabolic changes decrease their flexibility as they traverse through the microvasculature and promote their lysis or phagocytosis. Thus, the red cell's longevity and ability to carry out its proper function is critically dependent on cell-membrane structure and metabolism. The red cell membrane consists of a lipid bilayer and structural and integral membrane proteins which provide a lattice network under the bilayer and create the red cell cytoskeleton. Inherited defects in protein structure (hereditary spherocytosis, **HS**; hereditary elliptocytosis, **HE**; hereditary pyropoikilocytosis, **HPP**; etc.) lead to haemolytic anaemias. A number of specific receptors and enzymes are also associated with membrane proteins, several of which are important for the maintenance of its structural integrity or for nutrient and ion transport. Enzyme defects in metabolic pathways (pyruvate kinase, **PK**; hexokinase, **G6PD**), or haemoglobin defects (sickle-cell anaemia) can also increase the haemolytic potential of the red cell.

The character of the external red cell surface is defined by its antigenic structure. Over 300 antigens have been identified and many of these contribute to 15 genetic blood group systems. The latter are composed of oligosaccharide prosthetic groups of the integral membrane proteins and complex glycolipids. Nearly all (the Lewis system is an exception) are intrinsic components of the membrane and appear early in the differentiation process. Coating of red cell surface antigens with antibodies in cases of acquired autoimmunity interferes with the membrane functional integrity and allows rapid phagocytosis.

The transport of O_2 by red cells is dependent on their number, their haemoglobin content, and the ability to release O_2 or to increase 2,3DPG, according to tissue needs. However, O_2 transport by red cells is but one of the elements in a multicomponent highly integrated process that is responsible for appropriate O_2 supply to the tissues. A number of other physiological parameters, such as pulmonary function (sufficient O_2 loading in lungs), and haemodynamic factors (cardiac output, blood volume, and viscosity) must be incorporated in an integrated fashion. When anaemia is present, restoration of red cell number is dependent on the degree of erythropoietin stimulation, the ability to increase erythroid proliferation (adequate folic acid, vitamin B_{12} levels, etc.) within the bone marrow, and on the circulating levels and a normal erythroid cell response to iron. Several of these issues are addressed in more details in other chapters.

Finally, several hormones involved in the control of the cellular metabolism, such as corticosteroids, androgens, growth hormone, thyroxine, b-adrenergic agonists, and certain prostaglandins, stimulate erythroid differentiation *in vitro* in synergy with erythropoietin, or can alter hematocrit levels *in vivo*. A direct involvement of glucocorticoids in the control of the red cell mass has recently been provided by the observation that mice lacking the glucocorticoid receptor recover very poorly from haemolytic anaemia caused by phenylhydrazine treatment.

Further reading

Adamson JW (1968). The erythropoietin–hematocrit relationship in normal and polycythemic man: implications of marrow regulation. *Blood* **32**, 597–609.

Arcasoy MO, Harris KW, Forget BG (1999). A human erythropoietin receptor gene mutant causing familial erythrocytosis is associated with deregulation of the rates of Jak2 and Stat5 inactivation. *Experimental Hematology* **27**, 63–74.

Bauer A, *et al.* (1999). The glucocorticoid receptor is required for stress erythropoiesis. *Genes and Development* **13**, 2996.

Broudy VC, *et al.* (1996). Interaction of stem cell factor and its receptor c-kit mediates lodgment and acute expansion of hematopoietic cells in the murine spleen. *Blood* **88**, 75–81.

Carnot P, Deflandre C (1906). Sur l'activitéhématopoietique des serum au cours de la régénération du sang. *Academie des Sciences Medicale* **3**, 384.

Charbord P, *et al.* (1996). Early ontogeny of the human marrow from long bones: an immunohistochemical study of hematopoiesis and its microenvironment. *Blood* **87**, 4109–19.

Dybedal I, Jacobsen SE (1995). Transforming growth factor beta (TGF-beta), a potent inhibitor of erythropoiesis: neutralizing TGF-beta antibodies show erythropoietin as a potent stimulator of murine burst-forming unit erythroid colony formation in the absence of a burst-promoting activity. *Blood* **86**, 949–57.

Ebert BL, Bunn HF (1999). Regulation of the erythropoietin gene. *Blood* **94** 1864–77.

Era T, *et al.* (1997). Thrombopoietin enhances proliferation and differentiation of murine yolk sac erythroid progenitors. *Blood* **89**, 1207–13.

Eschbach J, *et al.* (1987). Correction of the anemia of endstage renal disease with recombinant human erythropoietin. *New England Journal of Medicine* **316**, 73.

Fennie C, *et al.* (1995). CD34+ endothelial cell lines derived from murine yolk sac induce the proliferation and differentiation of yolk sac CD34+ hematopoietic progenitors. *Blood* **86**, 4454–67.

Fleischman RA, Gallardo T, Mi X (1996). Mutations in the ligand-binding domain of the kit receptor: an uncommon site in human piebaldism. *Journal of Investigative Dermatology* **107**, 703–6.

Gallagher PG, Benz EJ, Jr (2001). The erythrocyte membrane and cytoskeleton: structure, function and disorders. In: Stamatoyannopoulos G, *et al.*, eds. *The molecular basis of blood diseases*, 3rd edn, pp. 275–305. WB Saunders, Philadelphia.

Hiyake T, Kung CK-H, Goldwasser E (1977). Purification of human erythropoietin. *Journal of Biological Chemistry* **252**, 5558.

Huehns ER, *et al.* (1964). Human embryonic haemoglobins. *Nature* **201**, 1095.

Ihle JN (2001). Signal transduction in the regulation of hematopoiesis. In: Stamatoyannopoulos G, *et al.*, eds. *The molecular basis of blood diseases*, 3rd edn, pp. 103–125. WB Saunders, Philadelphia.

Iliopoulos O, *et al*. (1996). Negative regulation of hypoxia-inducible genes by the von Hippel-Lindau protein. *Proceedings of the National Academy of Sciences, USA* **93**, 10595.

Jacobs K, *et al*. (1985). Isolation and characterization of genomic and cDNA clones of human erythropoietin. *Nature* **313**, 806.

Kelemen E, Calvo W, Fliedner TM (1979). *Atlas of human hemopoietic development*. Springer, Berlin.

Koury MJ, Bondurant MC (1990). Erythropoietin retards DNA breakdown and prevents programmed death in erythroid progenitor cells. *Science* **248**, 378–81.

Kralovics R, Prchal JT (2000). Congenital and inherited polycythemia. *Current Opinion in Pediatrics* **12**, 29–34.

Lin CS, *et al.* (1996). Differential effects of an erythropoietin receptor gene disruption on primitive and definitive erythropoiesis. *Genes and Development* **10**, 154–64.

Maltepe E, *et al*. (1997). Abnormal angiogenesis and responses to glucose and oxygen deprivation in mice lacking the protein ARNT. *Nature* **386**, 403–7.

Marcus D, ed. (1981). Blood group immunochemistry and genetics. *Seminars in Hematology* **18** 1.

Marine JC, *et al*. (1999). SOCS3 is essential in the regulation of fetal liver erythropoiesis. *Cell* **98**, 617–27.

Migliaccio AR, Papayannopoulou Th (2001). Erythropoiesis. In: Steinberg MH, *et al*., eds. *Disorders of hemoglobin, genetics, pathophysiology, clinical management*. pp. 52–71.

Moore MA, Metcalf D (1970). Ontogeny of the haemopoietic system: yolk sac origin of *in vivo* and *in vitro* colony forming cells in the developing mouse embryo. *British Journal of Haematology* **18**, 279–96.

Moritz KM, Lim GB, Wintour EM (1997). Developmental regulation of erythropoietin and erythropoiesis. *American Journal of Physiology* **273**, 1829–44.

Nakamura Y, *et al*. (1998). Impaired erythropoiesis in transgenic mice overexpressing a truncated erythropoietin receptor. *Experimental Hematology* **26**, 1105–10.

Nichols KE, *et al*. (2000). Familial dyserythropoietic anaemia and thrombocytopenia due to an inherited mutation in GATA1. *Nature Genetics* **24**, 266–70.

Orkin SH, Weiss MJ (1999). Cutting red-cell production. *Nature* **401**, 433–6.

Orkin SH (2001). Transcription factors that regulate lineage decisions. In: Stamatoyannopoulos G, *et al.*, eds. *The molecular basis of blood diseases*, 3rd edn, pp.80–94. WB Saunders, Philadelphia.

Papayannopoulou T, Abkowitz J, D'Andrea AD (2000). Biology of erythropoiesis, erythroid differentiation, and maturation. In: Hoffman R, *et al*., eds. *Hematology. Basic principles and practice*, pp 202–19. Churchill Livingstone, New York.

Ponka P (1997). Tissue-specific regulation of iron metabolism and heme synthesis: distinct control mechanisms in erythroid cells. *Blood* **89**, 1–25.

Raskind WH, *et al*. (2000). Mapping of a syndrome of X-linked thrombocytopenia and thalassemia to band Xp11–12: further evidence of genetic heterogeneity of X-linked thrombocytopenia. *Blood* **95**, 2262.

Reissmann KR (1950). Studies on the mechanism of erythropoietic stimulation in parabiotic rats during hypoxia. *Blood* **5**, 372.

Rico-Vargas SA, *et al*. (1994). c-kit expression by B cell precursors in mouse bone marrow. Stimulation of B cell genesis by *in vivo* treatment with anti-c-kit antibody. *Journal of Immunology* **152**, 2845–52.

Russell ES (1979). Hereditary anemias of the mouse: a review for geneticists. *Advances in Genetics* **20**, 357–459.

Sasaki A, *et al*. (2000). CIS3/SOCS3 suppresses erythropoietin signalling by binding the EPO receptor and JAK2. *Journal of Biological Chemistry* July 5, on-line.

Sato T, *et al*. (2000). Erythroid progenitors differentiate and mature in response to endogenous erythropoietin. *Journal of Clinical Investigation* **106**, 263–270.

Semenza GL (1999). Perspectives on oxygen sensing. *Cell* 98, 281–4.

Shimizu R, Komatsu N, Miura Y (1999). Dominant negative effect of a truncated erythropoietin receptor (EPOR- T) on erythropoietin-induced erythroid differentiation: possible involvement of EPOR-T in ineffective erythropoiesis of myelodysplastic syndrome. *Experimental Hematology* **27**, 229–33.

Southcott MJ, Tanner MJ, Anstee DJ (1999). The expression of human blood group antigens during erythropoiesis in a cell culture system. *Blood* **93**, 4425–35.

Spivak JL (2000). The blood in systemic disorders. *Lancet* **355**, 1707–12.

Spritz RA, Beighton P (1998). Piebaldism with deafness: molecular evidence for an expanded syndrome. *American Journal of Medical Genetics* **75**, 101–3.

Stamatoyannopoulos G, Grosveld F (2001). Hemoglobin switching. In: Stamatoyannopoulos G, *et al.*, eds. *The molecular basis of blood diseases*, 3rd edn, 135–65. WB Saunders, Philadelphia.

Stohlman K Jr, Rath CE, Rose JC (1954). Evidence for a humoral regulation of erythropoiesis: studies on a patient with polycythemia secondary to regional exposure to hypoxia. *Blood* **9**, 721.

Stopka T, *et al*. (1998). Human hematopoietic progenitors express erythropoietin. *Blood* **91**, 3766–72.

Takakura N, *et al*. (2000). A role for hematopoietic stem cells in promoting angiogenesis. *Cell* **102**, 199–209.

Tamura K, *et al*. (2000). Requirement for p38a in erythropoietin expression: a role for stress kinases in erythropoiesis. *Cell* **102**, 221–31.

Vaziri H, *et al*. (1994). Evidence for a mitotic clock in human hematopoietic stem cells: loss of telomeric DNA with age. *Proceedings of the National Academy of Sciences, USA* **91**, 9857–60.

Verdier F, *et al*. (2000). Proteasomes regulate the duration of erythropoietin receptor activation by controlling down-regulation of cell surface receptors. *Journal of Biological Chemistry* **275**, 18375–81.

Verfaillie CM (2000). Anatomy and physiology of hematopoiesis. In: Hoffman R, *et al*., eds. *Hematology. Basic principles and practice*, pp 139–54. Churchill Livingstone, New York.

Winearls C, *et al*. (1986). Effects of human erythropoietin derived from recombinant DNA on the anemia of patients maintained on chronic hemodialysis. *Lancet* **2**, 1175.

Wu H, *et al*. (1995). Generation of committed erythroid BFU-E and CFU-E progenitors does not require erythropoietin or the erythropoietin receptor. *Cell* **83**, 59–67.

Yu AY, *et al*. (1999). Impaired physiological responses to chronic hypoxia in mice partially deficient for hypoxia-inducible factor 1a. *Journal of Clinical Investigation* **103**, 691–6.

Zucali JR, Stevens V, Mirand EA (1975). *In vitro* production of erythropoietin by mouse fetal liver. *Blood* **46**, 85–90.

22.5.2 Anaemia: pathophysiology, classification, and clinical features

D. J. Weatherall

The main function of the red blood cells is oxygen transport. Hence a functional definition of anaemia is 'a state in which the circulating red-cell mass is insufficient to meet the oxygen requirements of the tissues'. However, many compensatory mechanisms can be brought into play to restore the oxygen supply to the vital centres, and therefore in clinical practice this definition is of limited value. For this reason anaemia is usually defined as 'a reduction of the haemoglobin concentration, red-cell count, or packed cell volume (**PCV**) to below normal levels'.

The definition of anaemia

It has been extremely difficult to establish a normal range of haematological values, and hence the definition of anaemia usually involves the adoption of rather arbitrary criteria. For example, the World Health Organization recommends that anaemia should be considered to exist in adults whose haemoglobin levels are lower than 13 g/dl (males) or 12 g/dl (females). Children aged 6 months to 6 years are considered anaemic at haemoglobin levels below 11 g/dl, and those aged 6 to 14 years below 12 g/dl. The disadvantage of such arbitrary criteria for defining anaemia is that there may be many apparently normal individuals whose haemoglobin concentration is below their optimal level. Furthermore, the published 'normal values' for adults (see Chapter 22.1) indicate that there is such a large standard deviation that many adult females must be considered 'normal' even though they have haemoglobin levels below 12 g/dl.

Prevalence of anaemia

Anaemia is a major world health problem and its distribution and prevalence in the developing world are considered in detail in the next chapter.

The prevalence of anaemia has been studied in many populations, but it is difficult to compare data from different sources because of variations in methodology and criteria. Certain patterns emerge, however. An early survey carried out in Great Britain established that haemoglobin levels were low in a significant proportion of the population, particularly susceptible groups being children under the age of 5 years, pregnant women, and those in social classes IV and V. A later random population study in the United Kingdom reported a prevalence of anaemia of 14 per cent for women aged 55 to 64 years and 3 per cent for men aged 35 to 64 years. These and similar studies have shown that anaemia is most common in women between the ages of 15 and 44 years and that it then becomes relatively less frequent, although the prevalence increases again in the 75-and-over age group. Interestingly, it is only in the last group that the prevalence in males and females is almost the same. Where the cause of the anaemia has been analysed in these surveys, the majority of cases have been due to iron deficiency. No doubt these prevalence data vary considerably between the developed countries, but it is clear that nutritional anaemia is relatively common in most populations at certain periods during development and late in life.

Adaptation to anaemia

The function of the red cell is to carry oxygen between the lungs and the tissues. However, tissue oxygenation is the result of a complex series of interactions of different organ systems, of which the red cell is only one (Table 1). Obviously the cardiac output, ventilatory function, and state of the capillaries are of great importance as well. Each of these oxygen supply systems is regulated differently. Ventilation responds to changes in pH, CO_2, and hypoxia. Cardiac output responds to the amount of blood entering the heart, and this is regulated mainly by the effects of tissue metabolism as it modifies the resistance to blood flow in the microvasculature. The erythron itself responds to changes in haemoglobin concentration, arterial oxygen saturation, and to the oxygen affinity of the circulating haemoglobin. Thus a decreased capacity of any of these components may be compensated for by increased activity of the others in an attempt to maintain tissue oxygenation.

Oxygen diffuses across the alveolar membrane and into the blood, which equilibrates with the alveolar gas; the approximate oxygen tension is 100 mmHg, at which the blood is fully saturated with an oxygen content of 20 vol per cent. As blood is pumped through the tissue capillaries oxygen diffuses out. Although the venous oxygen tension varies between organs, the oxygen tension of the pooled venous blood in the pulmonary artery, the 'mixed venous oxygen tension', is remarkably constant at 40 mmHg. At this oxygen tension the oxygen content is 15 vol per cent. Hence, oxygen delivery, as measured by the arteriovenous oxygen difference, is normally 5 vol per cent. By reducing the oxygen-carrying capacity of blood, anaemia tends to reduce the arteriovenous oxygen difference, and this may be compensated for by the following mechanisms: (1) modulation of oxygen affinity; (2) redistribution of flow between different organs; (3) increase in cardiac output; and (4) reduction of mixed venous oxygen tension to increase the arteriovenous oxygen difference.

Intrinsic red-cell adaptation

The consequences of anaemia on the normal oxygen-binding curve of blood are shown in Fig. 1. Anaemia, by lowering the haemoglobin concentration, proportionately reduces the oxygen-carrying capacity of the blood. As a response to this there is an increase in the 2,3-biphosphoglycerate (**2,3-BPG**) concentration in the red cell, shifting the dissociation curve to the right, so significantly enhancing tissue oxygen delivery (Fig. 1).

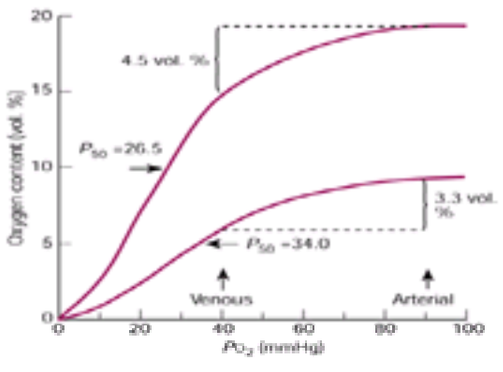

Fig. 1 Enhancement of oxygen loading by decreased red-cell oxygen affinity in a patient with anaemia. An anaemic patient with a 50 per cent reduction in haemoglobin concentration has only a 27 per cent reduction in oxygen unloading. (Based on Klocke RA (1972). *Chest*, **69**, 795.)

With increasing severity of anaemia there is a progressive increase in 2,3-DPG, which may increase oxygen delivery by as much as 40 per cent for the same haemoglobin concentration. It should be noted, however, that a consequence of this adaptation is a lower venous oxygen content and hence a lower reserve of oxygen available for a further increase in oxygen demand, as might occur on exercise for example. Hence the increase in 2,3-BPG in anaemia tends to ameliorate the effects of the diminished oxygen-carrying capacity of the blood, so reducing the adaptation required by other steps involved in tissue oxygen delivery (Fig. 2). 2,3-BPG levels vary in a variety of other clinical conditions, some of which are summarized in Table 2.

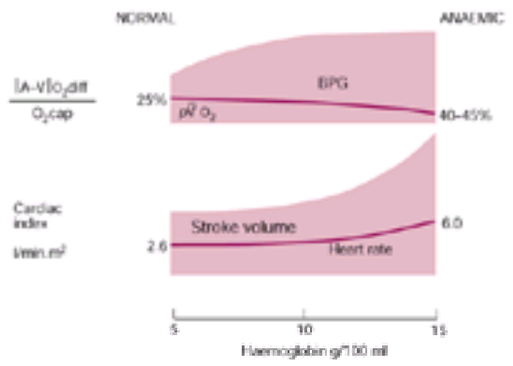

Fig. 2 The changes in factors involved in oxygen delivery with progressive anaemia. As anaemia becomes more severe, cardiac compensation becomes more significant ($P(V)O_2$, mixed venous oxygen tension). (From Bellingham, 1974.)

Local changes in tissue perfusion

The total blood volume does not change greatly in anaemia and therefore increased tissue perfusion has to be achieved by shunting blood from less to more vital organs. There is vasoconstriction of the vessels of the skin and kidney; this mechanism has little effect on renal function. The organs that gain from the redistribution seem to be mainly the myocardium, brain, and muscle.

Cardiovascular changes

It seems likely that mild anaemia is compensated for by shifts in the oxygen dissociation curve. Overall, oxygen consumption is unchanged in anaemia. However, when the haemoglobin level falls below 7 to 8 g/dl, there is an increase in cardiac output, both at rest and after exercise (Fig. 2). The stroke rate increases and a hyperkinetic circulation develops, characterized by tachycardia, arterial and capillary pulsation, a wide pulse pressure, and haemic murmurs. The circulation time is shortened, left ventricular stroke work is increased, and coronary flow increased in proportion to the increased cardiac output. It has been found that there is an acute reversal of the high-output state of chronic anaemia in response to orthostatic stress or pressor amines. This suggests that redistribution of blood volume and vasodilatation with reduced afterload play a dominant role in the hyperkinetic circulatory responses to chronic anaemia. The mechanism of the vasodilatation is not known; it may be a direct result of tissue hypoxia. An additional factor that may be of some importance in increasing cardiac output is the reduction in blood viscosity produced by a relatively low red-cell mass.

While the normal myocardium may tolerate sustained hyperactivity of this type indefinitely, patients with coronary artery disease or those with extreme anaemia may have impaired oxygenation of the myocardium. In such cases, cardiomegaly, pulmonary oedema, ascites, and peripheral oedema may occur, and a state of high-output cardiac failure is established. At this stage the plasma volume is almost always increased.

Pulmonary function

As blood, regardless of its oxygen-carrying capacity, is almost completely oxygenated in the lungs, the oxygen pressure of arterial blood in an anaemic patient should be the same as that in a normal individual, and hence an increase in respiratory rate should not improve the oxygenation of the tissues. Curiously, however, severe anaemia is associated with dyspnoea. Although in some patients this may be related to incipient cardiac failure, in most cases it appears to be an inappropriate response to hypoxia which is centrally mediated.

Clinical manifestations and classification of anaemia

Clinical effects of anaemia

Because anaemia reduces tissue oxygenation it is not surprising that it is associated with widespread organ dysfunction and hence an extremely varied clinical picture. The picture depends, of course, on whether the anaemia is of rapid or more insidious onset.

After acute blood loss the red-cell mass and plasma volume are reduced proportionately and the symptoms are mainly of volume depletion. Depending on the amount of fluid replacement there may be a small fall in the PCV during the first 10 h; volume replacement by the influx of albumin from the extravascular compartment takes between 60 and 90 h. Hence the picture of rapid blood loss is characterized by the typical syndrome of shock, with collapse, dyspnoea, tachycardia, a poor volume pulse, reduced blood pressure, and marked peripheral vasoconstriction.

With anaemia of a more insidious onset, the compensatory mechanisms outlined above have time to come into play. In mild anaemia there may be no symptoms or simply increased fatigue and a slight pallor. As the anaemia becomes more marked the symptoms and signs gradually appear. Pallor is best discerned in the mucous membranes; the nailbeds and palmar creases, although often said to be useful sites for detecting anaemia, are relatively insensitive for this purpose. Cardiorespiratory symptoms and signs include exertional dyspnoea, tachycardia, palpitations, angina or claudication, night cramps, increased arterial pulsation, capillary pulsation, a variety of cardiac bruits, reversible cardiac enlargement, and, if cardiac failure occurs, basal crepitations, peripheral oedema, and ascites. Neuromuscular involvement is reflected by headache, vertigo, light-headedness, faintness, tinnitus, roaring in the ears, cramps, increased cold sensitivity, and haemorrhages in the retina. Acute anaemia may occasionally give rise to papilloedema. Gastrointestinal symptoms include loss of appetite, nausea, constipation, and diarrhoea. Genitourinary involvement causes menstrual irregularities, urinary frequency, and loss of libido. There may be a low-grade fever.

In the elderly, in whom associated degenerative arterial disease is common, anaemia may present with the onset of cardiac failure. Alternatively, previously undiagnosed coronary narrowing may be unmasked by the onset of angina. Other symptoms of arterial degenerative disease may be also exacerbated or unmasked; intermittent claudication and a variety of neurological pictures associated with cerebral arteriosclerosis for example. It is important that anaemia is recognized as a contributing factor to the symptoms of these degenerative diseases as its correction may frequently bring about considerable symptomatic improvement.

Causes and classification of anaemia

A reduction in the red-cell mass can result from either the defective production of red cells or an increased rate of loss of cells, either by premature destruction or bleeding. Decreased production of red cells may result from a reduced rate of proliferation of precursors in the bone marrow or from failure of maturation leading to their intramedullary destruction: that is to say, ineffective erythropoiesis. Based on this approach we can derive a very simple pathophysiological classification of anaemia, as shown in Table 3, in which the causes are divided into failure of red-cell proliferation, defective maturation, haemolysis, and blood loss.

Anaemia due to defective proliferation of red-cell precursors

The major causes of this group of anaemias are an inadequate supply of iron, primary diseases of the bone marrow that involve stem cells or later erythroid

precursors, and a reduction in the amount of erythropoietin reaching the red-cell precursors (Table 4).

Iron deficiency results in defective erythroid proliferation and also in abnormal maturation of the red-cell precursors due to defective haemoglobin synthesis. Red-cell precursors require adequate iron supplies for normal proliferation, and the anaemia of iron deficiency tends to be hypoproliferative as well as dyserythropoietic. Chronic inflammatory disorders and related conditions also interfere with the iron supply to precursors, probably by blocking the release of catabolized red-cell iron from reticuloendothelial cells. The basic defect in iron-deficiency anaemia and that due to inflammation is similar, therefore, in that the supply of iron is inadequate to meet the requirements for erythropoiesis.

Defective proliferation of red-cell precursors can result from any of the causes of bone marrow failure, including infiltration with leukaemic or other neoplastic cells, damage due to ionizing radiation, drugs, or infection, and various intrinsic lesions of the stem cells or red-cell precursors. The intrinsic disorders include the congenital hypoplastic anaemias, involving either all the formed elements or the red-cell precursors alone.

Finally, decreased proliferation of the red-cell precursors may result from erythropoietin deficiency. The most common cause is chronic renal failure. A similar mechanism may be involved in conditions in which the tissue requirement for oxygen is reduced. These include various endocrine disorders such as hypothyroidism and hypopituitarism. It may also explain the mild anaemia associated with haemoglobin variants with decreased oxygen affinity.

As a group, the hypoproliferative anaemias are associated with a low reticulocyte count and defective proliferation of the bone marrow precursors. The red cells are usually normochromic and normocytic, although there may be a mild macrocytosis. If the anaemia is due to iron deficiency, the cells are hypochromic. If granulopoiesis is normal, the defect in red-cell proliferation is reflected by an increase in the myeloid:erythroid (**M/E**) ratio.

Defective red-cell maturation

Defects of red-cell maturation may involve primarily nuclear or cytoplasmic maturation (Table 4). Those involving nuclear maturation include vitamin B_{12} and folic acid deficiency and other causes of megaloblastic anaemia, and some of the primary marrow disorders including erythroleukaemia. The important causes of defective cytoplasmic maturation include the inherited disorders of globin synthesis, the thalassaemia syndromes, and the genetic and acquired defects of iron metabolism that characterize the sideroblastic anaemias. There are other genetic defects of red-cell maturation, the congenital dyserythropoietic anaemias, in which the aetiology is unknown. Furthermore, agents such as drugs, chemicals, and infections may interfere with erythroid maturation.

The main pathological mechanism common to all the anaemias that result from maturation abnormalities is ineffective erythropoiesis. In other words, there is marked erythroid proliferation but many of the precursors are destroyed in the bone marrow before they enter the circulation. Hence, the characteristic finding is marked erythroid hyperplasia with a reduction in the M/E ratio, associated with a low reticulocyte count. Because of the significant intramedullary destruction of precursors there is usually an elevated level of bilirubin and lactate dehydrogenase. Furthermore, there are nearly always morphological abnormalities of the red-cell precursors. The anaemias that are associated with abnormal nuclear maturation, such as those due to vitamin B_{12} and folic acid deficiency, are characterized by megaloblastic erythropoiesis and macrocytic red cells, while those caused by abnormal cytoplasmic maturation are characterized by normoblastic hyperplasia and hypochromic and microcytic red cells. However, even in the last conditions, there is marked anisocytosis and there may be a proportion of macrocytes in the peripheral circulation.

Blood loss

As mentioned earlier, the clinical picture associated with an acute loss of a large volume of blood is that of hypovolaemic shock.

Anaemias due to chronic blood loss may develop very insidiously and cause considerable diagnostic problems. Chronic blood loss from the gastrointestinal tract or uterus of more than 15 to 20 ml per day produces a state of negative iron balance. Assuming that the patient starts with a normal body store of iron, which is usually in the region of 1 g, the bone marrow will be able to maintain a normal haemoglobin level until the iron stores are totally depleted. At this stage there is no demonstrable iron in the bone marrow and the plasma iron level starts to fall but the patient is not anaemic. With a further fall in the plasma iron level, the haemoglobin level starts to fall, although at this stage the erythrocyte morphology may be relatively normal, as are the red-cell indices. It is only when iron-deficiency anaemia is well established that the typical morphological appearances of the red cells develop, and only after extreme periods of iron depletion that the tissue changes of iron deficiency become manifest.

From these considerations it is apparent that there may be prolonged blood loss before a patient presents with the symptoms and signs of anaemia. During the earlier stages the peripheral blood film may not be helpful in diagnosis, even though the serum iron level may be extremely low. Indeed, sometimes a dimorphic blood picture with normochromic and hypochromic cell populations may be seen. With chronic blood loss there is quite often a persistent thrombocytosis, and a hypochromic blood picture with thrombocytosis should always raise the possibility of chronic bleeding. In practice, the most common sites of such bleeding are a hiatus hernia, peptic ulcer, and tumour of the large bowel or the uterus.

Haemolytic anaemia (Table 5)

When the lifespan of red cells is shortened there is a reduction in the circulating red-cell mass, which leads to relative tissue hypoxia. This causes an increased output of erythropoietin with stimulation of the bone marrow and an increased rate of red-cell production. This is reflected by a raised reticulocyte count and a macrocytosis due to the presence of young cells in the peripheral circulation. Because of the increased rate of red-cell destruction, there is an increased production of bilirubin, which leads to mild icterus and the presence of increased amounts of urobilinogen in the urine and stool. Thus the haemolytic anaemias are characterized by a variable degree of anaemia, a reticulocytosis, and hyperbilirubinaemia. Their pathophysiology is considered in detail elsewhere.

Red cells are prematurely destroyed either because of an intrinsic lesion or as a result of the action of an extrinsic agent. The intrinsic abnormalities of the red cells that lead to their premature removal are nearly all genetic defects of either the membrane, haemoglobin, or metabolic pathways. The extrinsic agents that may cause premature destruction of the cells include a variety of antibodies, chemicals, drugs, and toxins, or bacteria and parasites. In addition, red cells may be damaged by direct trauma in the microcirculation or on body surfaces.

Premature destruction of red cells may take place either intravascularly or extravascularly, or, as occurs more commonly, in both sites. The site of destruction depends on the type and degree of damage to the red cell. For example, complement-damaged cells develop large holes in the membrane and are destroyed in the circulation, whereas IgG-coated cells are removed mainly in the reticuloendothelial system.

Clearly, there are numerous causes of premature destruction of red cells. These will be considered in detail later in this section. Usually it is easy to recognize that a particular anaemia has a haemolytic basis, by virtue of the reticulocytosis and macrocytosis associated with erythroid hyperplasia of the bone marrow, hyperbilirubinaemia, and increased urinary urobilinogen. However, it should be remembered that many anaemias associated with the abnormal proliferation or maturation of red cells have a haemolytic component. For example, there may be a slightly shortened red-cell survival in patients with pernicious anaemia or thalassaemia and yet there may be a very poor reticulocyte response. Similarly, there is a haemolytic component in the anaemia due to inflammation or malignancy but again the marrow response is poor. In such cases it may be necessary to measure the lifespan of the red cells directly in order to determine the magnitude of the haemolytic component as compared with defective proliferation or maturation.

General approach to the anaemic patient

Clinical assessment

The clinical assessment of patients with anaemia has two main objectives. First, it is essential to determine the degree of disability caused by the anaemia and hence how quickly treatment must be started. Second, as much information as possible about the likely cause of the anaemia must be obtained from a detailed clinical history and physical examination. There is no place for the 'blind' treatment of anaemia without first establishing the cause.

In assessing the severity of the anaemia and how urgently treatment should be instituted, a detailed history of the patient's exercise tolerance must be obtained. This should include a specific enquiry of symptoms suggestive of cardiac complications including angina, dysrhythmias, positional dyspnoea, cough, or ankle swelling. The clinical examination should include a careful assessment of the degree of pallor, the position of the neck veins, whether there are warm extremities and a bounding

pulse with a large pulse pressure, the presence of ankle or sacral oedema, and whether there are basal crepitations. The finding of profound anaemia with signs of cardiac failure indicates that urgent treatment is required. If the anaemia is associated with marked splenomegaly there will almost certainly be an increased blood volume and, particularly if there are already signs of cardiac failure, the patient may well go into acute left ventricular failure if transfused. Severely ill patients with profound anaemia require immediate treatment in an environment where they can be under constant observation, have regular measurements of their central venous pressure, and where they can be managed by experienced clinical and nursing staff.

An account of history taking and clinical examination in patients with haematological disorders was given earlier in this section (Chapter 22.1). It cannot be emphasized too strongly that in many cases the anaemia is a symptom of a non-haematological disorder. A detailed history and clinical examination will often provide a clue as to the likely cause of the anaemia, and which laboratory investigations are likely to be most productive for confirming the diagnosis.

Haematological investigation

A preliminary blood count and blood film examination should classify anaemia into hypochromic-microcytic, and macrocytic or normochromic, normocytic varieties (Table 6). In middle-aged women with a history of several pregnancies or heavy menstrual loss it is reasonable to assume that a hypochromic anaemia is due to iron deficiency, and to treat them with iron without further investigation. However, hypochromic anaemia in males or young or postmenopausal women always suggests blood loss and should be investigated accordingly. If there is any doubt about a hypochromic anaemia being due to iron deficiency, the serum iron level and total iron-binding capacity should be established. Hypochromic anaemia with a normal serum iron suggests a genetic or acquired defect in haemoglobin synthesis, common causes being thalassaemia and sideroblastic anaemia. The diagnosis of a macrocytic anaemia always requires further investigation and should be followed up with a bone marrow examination. A macrocytosis with a normoblastic bone marrow may result from alcohol abuse, haemolysis, or, occasionally, one of the refractory anaemias with hyperplastic bone marrow (see Chapter 22.5.8). Macrocytic anaemias with megaloblastic bone marrows are usually due to vitamin B_{12} or folate deficiency and should be investigated accordingly. If there is macrocytosis with a reticulocytosis, hyperbilirubinaemia, and a normoblastic marrow, a haemolytic anaemia is likely; an approach to the further investigation of haemolysis is described in Chapter 22.5.9.

The normochromic, normocytic anaemias often cause more diagnostic difficulty. Some help can be gained from a determination of whether the white-cell and platelet counts are normal. If there is associated neutropenia and thrombocytopenia, a primary disease of the bone marrow is likely; hence, bone marrow examination should be made to determine whether there is hypoplasia of the various precursor forms, hypoplastic or aplastic anaemia, or whether the pancytopenia results from infiltration of the bone marrow as occurs in the various forms of leukaemia. If there are nucleated red cells or young white cells on the peripheral film (that is, a leucoerythroblastic picture), a bone marrow examination is essential, as this type of reaction usually indicates infiltration of the bone marrow with abnormal cells, either as part of a primary marrow disease such as leukaemia, or metastatic carcinoma. In the normochromic, normocytic anaemias in which the white-cell count and platelet count are normal, it is also helpful to make a bone marrow analysis. The most common cause is anaemia of chronic disorders, the diagnosis of which is described in detail below. Another particularly common cause is chronic renal failure. After these conditions have been excluded, there remain the chronic anaemias associated with endocrine deficiencies (see Chapter 22.7) or the primary red-cell hypoplasias (Chapter 22.3.11).

The management of anaemia

The management of specific forms of anaemia is described in detail in subsequent chapters. However, a few principles can be outlined here. In general, a cause should always be sought before treatment is instituted. There is no place whatever for treating anaemia 'blind' with multihaematinic preparations. As mentioned above, most cases of iron-deficiency anaemia require further investigation for a source of blood loss. If there is a clear-cut history of poor diet, multiple pregnancies, or obvious uterine bleeding, it is reasonable to start iron therapy and observe the haemoglobin level both during the period of treatment and for some months after iron therapy has been stopped. A rise in the haemoglobin level of approximately 1 g/dl per week indicates a full haematological response. For the megaloblastic anaemias it is quite reasonable to start treatment with vitamin B_{12} and folic acid once a diagnosis has been established and blood samples have been obtained for serum folate and vitamin B_{12} levels. The precise cause of the megaloblastic anaemia can be established at leisure once these samples have been obtained. A brisk reticulocyte response 5 to 7 days after initiating therapy suggests that there will be a full restoration of the haemoglobin level to normal. Failure of response of a hypochromic anaemia to adequate iron therapy should be managed by first finding out whether the iron is being taken by the patient and, if so, by determining the serum iron level. If it is normal, causes of hypochromic anaemia that are not associated with iron deficiency—thalassaemia and sideroblastic anaemia for example—should be sought. Similarly, refractory macrocytic anaemias require detailed analysis of the bone marrow morphology as there may be an underlying preleukaemic state.

Blood transfusion should always be avoided unless the haemoglobin level is dangerously low, in which case it is reasonable to transfuse the patient up to a safe level and then allow the haemoglobin to return to normal following appropriate treatment of the underlying cause. The decision whether to transfuse an anaemic patient depends mainly on the severity of the anaemia and its cause. For example, a young patient with a haemoglobin of 5 g/dl who is shown to have an active duodenal ulcer should probably be transfused because they would be at severe risk from a further brisk bleed from the ulcer. On the other hand, a patient of similar age with a similar haemoglobin level due to chronic nutritional iron deficiency might well be allowed to restore their haemoglobin level by oral iron therapy.

Occasionally, patients present in gross congestive cardiac failure with profound anaemia. This picture is usually seen in elderly patients with long-standing pernicious anaemia or iron deficiency. This type of condition still carries a high mortality and requires urgent treatment. Such profoundly anaemic patients require transfusing up to a safe level, that is a haemoglobin value of 6 to 8 g/dl. This can usually be achieved by the slow transfusion of two or three units of red cells with the intravenous administration of a potent diuretic such as furosemide (frusemide) with each unit; the diuretic should never be mixed directly with the blood. A very careful check on the neck veins and lung bases should be made throughout the period of transfusion. Ideally, a central venous-pressure line should be inserted before the transfusion is started. Occasionally, patients are encountered in such gross heart failure that the administration of packed cells and diuretics worsens the failure. In this situation it is possible to raise the circulating red-cell mass by infusing packed cells or whole blood through one arm while removing an equal volume of blood from the other. By carrying out a two-to-three unit exchange transfusion of this type it may be possible to tide the patient over while treating the heart failure by conventional means.

Further reading

Adamson JW, Finch CA (1975). Haemoglobin function, oxygen affinity and erythropoietin. *Annual Review of Physiology* **37,** 351–69.

Bellingham AJ (1974). The red cell in adaptation to anaemic hypoxia. *Clinics in Haematology* **3,** 577–94.

Bunn HF, Forget BG (1986). *Hemoglobin: molecular, genetic and clinical aspects.* Saunders, Philadelphia.

Hjelm M, Wadman B (1974). Clinical symptoms, haemoglobin concentration and erythrocyte biochemistry. *Clinics in Haematology* **3,** 689–704.

Oski FA (1993). Differential diagnosis of anemia. In: Nathan DG, Oski FA, eds. *Hematology of infancy and childhood*, pp 346–53. Saunders, Philadelphia.

Varat MA, Adolph RJ, Fowler NO (1972). Cardiovascular effects of anemia. *American Heart Journal* **83,** 415–26.

Viteri FE, Torun B (1974). Anaemia and physical work capacity. *Clinics in Haematology* **3,** 609–26.

Weatherall DJ, Bunch C (1985). The blood and blood forming organs. In: Smith LH, Their SO, eds. *Pathophysiology*, 2nd edn, pp 173–320. Saunders, Philadelphia.

Woodson RD (1974). Red cell adaptation in cardiorespiratory disease. *Clinics in Haematology* **3,** 627–48.

22.5.3 Anaemia as a world health problem

D. J. Weatherall

Despite improvements in nutrition and hygiene, which have reduced childhood mortality in many emerging countries, anaemia continues to be a major world health problem. It is not, of course, a disease in its own right but simply a by-product of a wide variety of different disorders, most of which are described in detail elsewhere in this book. However, because of its importance as a source of chronic ill health in many populations, the global aspects of the aetiology and manifestations of anaemia are summarized briefly in this chapter. Readers who wish to learn more of the complex literature on this important topic are referred to the extensive reviews cited at the end of the chapter.

Definition and prevalence

It has been very difficult to produce an adequate definition of anaemia. 'Normal' haematological values vary with age, between sexes, at different altitudes, and, possibly, between races. On the other hand, it is helpful to have a standard set of haemoglobin levels at different ages below which 'anaemia' is defined. The World Health Organization (WHO) have attempted to set out criteria of these kind, summarized in Table 1. Despite their many shortcomings, including methodological vagaries, they at least provide a way of obtaining an approximate comparison of the distribution and frequency of anaemia among the different countries of the world.

The global prevalence of anaemia, based on WHO criteria, was estimated in the 1980s. A review of the epidemiological data available at this time suggested that about 1.3 billion people were affected by anaemia, particularly in the developing countries. Infants, young children, menstruating, and, especially, pregnant women were the most severely affected groups (Table 2). The highest prevalence of anaemia was found in southern Asia and Africa. More recent work suggests that while there has been some improvement, anaemia is still a major public health problem in many developing countries. Though found most frequently in poorer countries, anaemia is still an important problem in richer societies, particularly in infancy, pregnancy, and old age.

The complex and multiple aetiology of anaemia in the Third World

The major causes of anaemia in the developing countries are summarized in Table 3. It is very difficult to determine their relative importance, particularly in tropical countries. Most surveys have focused on one particular mechanism, iron or folate deficiency for example. To obtain a true picture of the cause of anaemia in a particular population it is essential to obtain consecutive data over a long period. For example work in the Gambia has shown that the haemoglobin levels in children vary significantly at different times of the year; anaemia is much more common in the wet season when malaria transmission is at its highest. To complicate matters, this is also the time when diarrhoea and malnutrition are most common. Heavy rains after many dry months have profound effects on the community; sanitation measures are disrupted and food stores are at the lowest level in the annual cycle (Fig. 1).

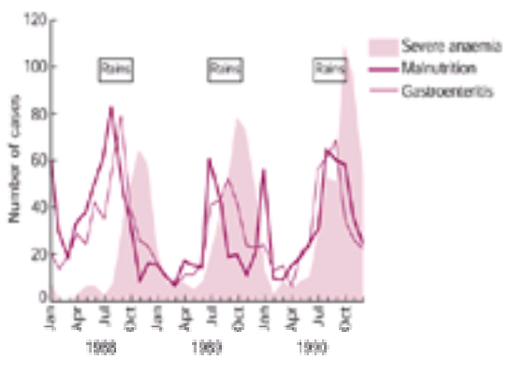

Fig. 1 Admissions to the children's ward in a hospital in The Gambia over dry and rainy seasons. (Data from Brewster DR and Greenwood BM (1993). Season variation of paediatric disease in The Gambia, West Africa. *Annals of Tropical Paediatrics* **13**, 133.)

These observations underline the multifactorial aetiology of anaemia in the developing world. Nonetheless it is clear that iron deficiency, which probably affects at least 20 per cent of the world's population, is the most important factor; the many other diseases that can exacerbate anaemia are often operating in the background of low body-iron stores.

Iron deficiency

It was estimated in the 1980s that some 600 to 700 million individuals suffer from anaemia due to iron deficiency. Surveys using more sensitive indicators of iron status showed that iron depletion is even more prevalent than frank anaemia. The causes of iron deficiency anaemia are extremely complex and vary widely among different populations. The absorption of non-haem iron, except from breast milk, is comparatively restricted, and the content of iron in breast milk is very low. Iron deficiency is particularly common in communities in which food is predominantly of vegetable origin. The three great staples in these populations are rice, wheat, and maize. Sorghum and millet are also important in parts of Africa and Asia. Soy and similar legumes are a major source of protein in many countries. The iron content of these diets is generally low, and, furthermore, absorption is inhibited by fibre, phytates, phosphates, and polyphenols, all of which occur in high levels in vegetarian diets. Populations who have remained as hunter-gatherers, and pastoralists who eat blood and meat, appear to have a lower frequency of iron deficiency anaemia.

Against this background of deficient or borderline dietary iron intake, there are a number of other factors which may exacerbate iron deficiency. Iron requirements are greatly increased during pregnancy because of the expansion of the maternal red cell mass (approximately 500 mg), iron transport to the fetus (approximately 300 mg), and the constitution of the placenta (approximately 25 mg), together with any blood loss at birth. Although there is some compensation by the cessation of iron loss due to menstruation (approximately 200 mg), the total requirements for a single pregnancy are greater than 1000 mg. Iron is also excreted in breast milk and although the concentration is low this loss, particularly with prolonged breast feeding, places a further burden on maternal iron stores.

In many tropical countries, there are important sources of pathological iron loss due to parasitic infection. Hookworm infestation affects millions of people world-wide. These parasites attach themselves to the mucosa of the intestinal tract. With a worm-load of 1000 eggs per gram of faeces, the intestinal blood loss averages about 2.5 ml/day, representing 1 mg of iron. Although some of this is reabsorbed, perhaps up to 40 percent, hookworm infestation is an important source of iron imbalance. Infection with *Schistosoma mansoni* results in intestinal blood loss, while *S. haematobium* results in chronic haematuria. In Kenyan children, for example, mean iron losses in those infected with *S. haematobium* varied from 149 to 652 µg/day, according to the magnitude of the egg counts.

Finally, it should be remembered that chronic ill health due to protein–calorie malnutrition or chronic infection may, by its effect on a patient's appetite, result in further depletion of iron intake.

It must be emphasized that many surveys for assessing body iron stores have used methods which are confounded by associated inflammatory disease or other disorders. These problems are particularly germane to surveys which have been based on serum iron or ferritin levels. More recently, screening methods based on estimation of transferrin receptor levels have been developed but their application to large populations is, as yet, limited.

Folate deficiency

Folate deficiency is thought to be the second commonest cause of nutritional anaemia in the world population. The mechanisms are complex and differ widely between different populations depending in the way in which food is prepared, in particular the temperature at which it is cooked. It is also clear that dietary folate deficiency is not the whole story. Research in Africa suggests that the continuous anorexia which accompanied recurrent infections, such as malaria or tuberculosis, is a major cause of folate deficiency in children. Postinfective malabsorption and the tropical sprue syndrome are also important causes of folate deficiency, particularly in the Indian subcontinent. Folate requirements may be increased in patients with erythroid hyperplasia secondary to chronic haemolytic anaemia, sickle cell anaemia for example, or chronic malarial infection. They also increase markedly during pregnancy. In women with low baseline folate stores, megaloblastic anaemia in pregnancy or the puerperium is particularly common.

Vitamin B_{12} deficiency

Nutritional vitamin B_{12} deficiency is uncommon, although it is observed in true vegans, particularly in the Indian subcontinent. Infants born of mothers with sprue or postinfective malabsorption who are fed on breast or goats milk containing insufficient vitamin B_{12} may develop megaloblastic anaemia with locomotor complications during the early months of life.

Infection

Almost any chronic infection may produce anaemia. Globally, the most important are the parasitic disorders, malaria, visceral leishmaniasis (kala-azar), schistosomiasis, and some forms of trypanosomiasis. The anaemias due to chronic hookworm infestation were considered in Chapter 22.5.2

Malaria is still the most important parasitic illness of humans. Currently it is estimated that it has a global incidence of about 200 million cases per year, with over one million deaths. Its transmission and clinical manifestations are considered in Section 7. Profound anaemia is a major cause of mortality and morbidity during acute attacks of *P. falciparum* malaria in non-immune persons but, from the perspective of health in the developing world, chronic infection with this organism in childhood is an extremely common cause of anaemia. This is most commonly seen in areas of high malarial transmission and is a growing problem in regions of lower transmission because the rise in antimalarial drug resistance prolongs the average duration of infection. The anaemia of chronic malaria has a complex basis involving haemolysis, hypersplenism, and a suboptimal bone marrow response, often set against a background of iron or folate deficiency. In some populations, notably those of Africa, India, and parts of Southeast Asia, chronic malarial infection may be complicated by the hyper-reactive malarial splenomegaly syndrome, in which hypersplenism plays a major role in the generation of chronic anaemia.

The haematological manifestations of the other common parasitic illnesses in the tropics are considered elsewhere.

Malabsorption

A large proportion of people in tropical climates, both indigenous populations and expatriates who have worked in rural areas, have abnormalities of the intestinal mucosa, often associated with impairment of absorption. These structural and functional alterations of the gut have been called 'tropical enteropathies'. It is likely that they result from adaptation to life in the contaminated environment of the tropics, with frequent gastrointestinal infections and differences of diet.

More severe malabsorption syndromes, called sprue and postinfective malabsorption, are associated with chronic diarrhoea, wasting, and a variable degree of anaemia. The pathophysiology and world distribution of these syndromes are considered in Section 14. They are nearly all associated with anaemia which has a complex aetiology including folate deficiency and, in some cases, iron deficiency.

It should also be remembered that in a tropical setting malabsorption can also result from colonization of the small bowel by specific parasites, including *Giardia lamblia*, *Strongyloides stercralais*, *Cryptosporidium*, and others. Abdominal tuberculosis with malabsorption is also common. In Africa, HIV infection is now an important cause of malabsorption.

Inherited anaemias

The inherited haemoglobin disorders are becoming an increasingly common cause of anaemia, particularly in tropical countries. They are described in detail elsewhere.

Because of heterozygote advantage against *P. falciparum* malaria, the important inherited haemoglobin disorders, notably sickle cell anaemia and the thalassaemias, have a high frequency throughout tropical populations of the Old World. Sickle cell anaemia and its variants are particularly common in Africa, some Mediterranean populations, and throughout the Middle East and parts of India. They also occur at a high frequency in the Caribbean and in other regions with large African populations. The thalassaemias occur at a high frequency in parts of Africa, the Mediterranean, the Middle East, the Indian subcontinent, and throughout Southeast Asia. There is now clear evidence that these conditions will produce a major public health problem in these countries in the future. As poorer countries go through the demographic transition, resulting from better hygiene and control of infectious illness, infants with these genetic anaemias are now surviving long enough to present for diagnosis and treatment. Some estimated figures for the annual numbers of new births of babies with sickle cell anaemia or b thalassaemia are shown in Fig. 2.

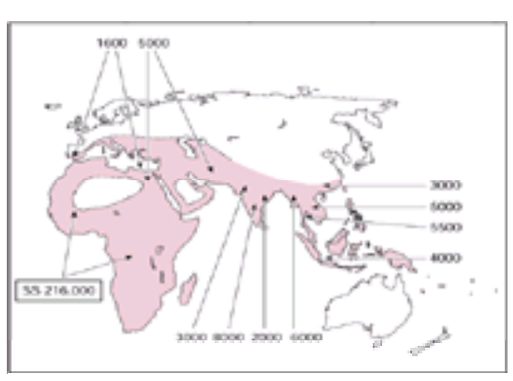

Fig. 2 Estimated annual numbers of births of babies with b thalassaemia and sickle cell anaemia (SS). (Original data in Weatherall and Clegg (2001). *The thalassaemic syndromes*, 4th edn, p.599. Blackwell, Oxford.)

The effect that a high frequency of a disease such as thalassaemia can have on the health economy of an emerging country was shown graphically in the case of Cyprus after is passed through the demographic transition in the 1950s. It was estimated that if every patient with this disease was treated with regular blood transfusion and appropriate medication, within 15 years the management of this one condition would consume up to 40 per cent of the island's health budget. Recent studies in Indonesia indicate that, at a minimum estimate, approximately 1.25 million units of blood will be required each year to treat a proportion of the thalassaemic

population in future years.

In many populations, there are hundreds of thousands of carriers for b thalassaemia or the more common severe forms of a thalassaemia. Although they are asymptomatic they have haemoglobin values which, on average, are 1 to 1.5 g/dl below normal. During pregnancy they retain this difference so that in the midtrimester they have haemoglobin values of approximately 8 g/dl or less. They have increased folate requirements and, in some populations, there appears to be an increased frequency of folate deficiency in pregnancy.

It should be remembered that the inherited anaemias may be exacerbated by other illnesses which are widespread in tropical countries. Folate requirements are increased in all these conditions and secondary folate deficiency is extremely common. They may also be exacerbated by malaria; children may develop malarial infection from infected blood donors. There is also a high frequency of other blood-borne infections, particularly hepatitis C and, in some populations, HIV. Furthermore, there is clear evidence that sickle cell anaemia and thalassaemia can render children more prone to infection. In short, like all forms of anaemia in the tropical world, the inherited disorders of haemoglobin may present with a complex series of complications due to a background of nutritional deficiency and a wide variety of infections.

These complex interactions have a major effect on the prognosis for the important inherited haemoglobin disorders. Early studies in Africa reported a marked paucity of patients with sickle cell anaemia despite a very high carrier frequency, indicating that very few patients with this disorder were surviving beyond early childhood. This may still be the case in parts of rural Africa. On the other hand, in more advanced countries, and with a high quality of medical care, patients with this disease are regularly surviving into adult life; the mean survival time in the United States is now approximately 42 years, with many patients surviving to old age. A similar situation exists for b thalassaemia. In poorer countries, supplies of blood may be limited, there may be difficulties in screening blood for agents such as hepatitis C and HIV, and the prohibitive cost of iron chelating agents means that even children who do receive transfusion die from iron loading before they reach the age of 20 years.

There are other inherited anaemias which are particularly common on tropical countries due to heterozygote advantage against malaria. Glucose-6-phosphate dehydrogenase deficiency is estimated to occur in some 100 million individuals world-wide. Its clinical and haematological manifestations are discussed in Chapter 22.5.12. They include haemolytic reactions to a wide variety of drugs, and, of particular public health significance, to certain foods (favism). There is a form of ovalocytosis which is particularly common in Melanesia which is associated with a mild and well compensated haemolytic anaemia. Recent studies have shown that carriers of Melanesian ovalocytosis are completely protected against cerebral malaria.

Consequences of anaemia

The results of many studies directed at determining the functional consequences of anaemia are still controversial. It is often difficult to distinguish between the effects of anaemia *per se* and the consequences of iron or folate deficiency on other physiological functions. Whatever the mechanism, chronic anaemia is associated with diminished function.

Many studies have suggested that even mild anaemia may reduce near-maximal work capacity. There is some evidence that both in iron-deficient animals and children it reduces mental performance and immune function. There is no doubt that anaemia increases maternal mortality and morbidity. There is a very large literature on the effect of iron deficiency on resistance to infection, as mediated through either immune function or the bacteriostatic and bacteriocidal roles of iron-containing proteins such as transferrin and lactoferrin. The entire complex relationship between iron status and susceptibility of infection requires further work. It is clear that folate deficiency is associated with an increased prevalence of obstetric complications and fetal malformation, although its effect on intellectual and immune function is less clear.

In short, because of the remarkable ability of otherwise healthy individuals to adapt to moderate anaemia it seems likely that many of the associated manifestations which have been observed result from the effects of different deficiency states on other physiological functions rather than the anaemia *per se*. On the other hand, chronic severe anaemia, particularly in childhood, results in a wide variety of complications including failure of growth and development and, possibly, proneness to infection.

Prevention

It is beyond the scope of this brief review to discuss the protean aspects of the prevention of anaemia, particularly in poorer countries. Its high prevalence is a reflection of gross poverty, particularly as manifested by nutritional deficiency, infection, and malabsorption. Its control requires action on many different fronts, including improvements in diet, fortification of commonly eaten foods with iron, the use of modified milk formulae for infants, malaria and hookworm control, iron and folate supplementation in pregnancy, and all round improvements in hygiene. The problem of the population control of sickle cell anaemia and thalassaemia is discussed in Chapter 22.5.7. Good antenatal care helps to prevent anaemia in childhood by reducing prematurity, increasing average birth weight, and improving the nutritional status of the newborn.

Summary

The extremely high prevalence of anaemia in the poorer countries is a reflection of the abject poverty of many of their populations. The anaemia of the developing world, particularly in childhood, seems to reflect a series of vicious circles. Maternal anaemia due to iron or folate deficiency and chronic malaria is associated with the birth of underweight infants who frequently have low iron stores and may also be folate depleted. Anaemia is usually present from about 6 months of age. Such infants are prone to infection, particularly gastrointestinal, and may be further depleted of iron or folate by inappropriately prolonged breast feeding or weaning onto an inadequate diet. They are exposed to hookworm infection as soon as they start to crawl, malaria becomes a major problem after 6 months, and in many populations the increasingly common haemoglobinopathies are a further cause of anaemia after the first few months of life.

Further reading

Beales PF (1997). Anaemia in malaria control: a practical approach. *Annals of Tropical Medicine and Parasitology* **91**, 713–18.

DeMaeyer EM, Adiels-Tegman M (1985). The prevalence of anemia in the world. *World Health Statistics Quarterly* **38**, 302–16.

DeMaeyer EM, Dallman P, Gurney JM, Hallberg L, Sood SK, Srikantia SG (1989). *Preventing and controlling iron deficiency anaemia through primary health care*. World Health Organization, Geneva.

Eskeland B, Hunskaar S (1999). Anaemia and iron deficiency screening in adolescence: a pilot study of iron stores and haemoglobin response to iron treatment in a population of 14–15-year-olds in Norway. *Acta Paediatrica* **88**, 815–21.

Fleming AF (1989). Tropical obstetrics and gynaecology. 1. Anaemia in pregnancy in tropical Africa. *Transaction of the Royal Society of Tropical Medicine and Hygiene* **83**, 441–8.

Flowers CH, Cook JD (1999). Dried plasma spot measurements of ferritin and transferrin receptor for assessing iron status. *Clinical Chemistry* **45**, 1826–32.

Gallacher PG, Ehrenkranz RA (1995). Nutritional anaemias in infancy. *Clinical Perinatology* **22**, 671–92.

Hercberg S, Galan P (1992). Nutritional anaemias. *Clinical Haematology* **5**, 143–68.

Khusun H, Yip R, Schultink W, Dillon DH (1999). World Health Organization hemoglobin cut-off points for the detection of anemia are valid for an Indonesian population. *Journal of Nutrition* **129**, 1669–74.

Morris SS, Ruel MT, Cohen RJ, Dewey KG, de la Briere B, Hassan MN (1999). Precision, accuracy and reliability of hemoglobin assessment with use of capillary blood. *American Journal of Clinical Nutrition* **69**, 1243–8.

Viteri FE, Torun B (1974). Anaemia and physical work capacity. *Clinical Haematology* **3**, 609–26.

Weatherall DJ, Kwiatkowski D (1997). Hematologic manifestations of systemic diseases in children of the developing world. In: Nathan DG, Orkin SH, eds. *Hematology of infancy and childhood*, 5th edn, pp. 1893–914. WB Saunders, Philadelphia.

Wharton BA (1999). Iron deficiency in children: detection and prevention. *British Journal of Haematology* **106**, 270–80.

World Health Organization (1968). *Nutritional anaemias*. Technical Report Series No. 405. WHO, Geneva.

22.5.4 Iron metabolism and its disorders

T. M. Cox

Homeostasis, transport, and storage of iron

Iron is a critically important micronutrient. As a component of metalloenzymes and complexed to form haem, it participates in the transport of oxygen by haemoglobin and myoglobin and in the harnessing of metabolic energy by cytochromes of the electron transport chain.

Iron is abundant in the environment and is the fourth most common element in the earth's crust but iron chemistry poses exceptional problems for living organisms. The metal exists in two readily interconvertible redox states (divalent ferrous and trivalent ferric iron) which are very reactive. In the environment, most iron is oxidized to the trivalent state which, under neutral conditions, is then rapidly hydrolysed to insoluble polyhydroxide complexes that are metabolically inaccessible. High-affinity iron-binding proteins that complex ferric iron have evolved to ensure its transport in the body and delivery to sites of utilization and haem biosynthesis. In bacteria and archaebacteria, a family of compounds termed siderophores have evolved that are secreted into the environment to complex ferric iron competitively for uptake. The availability of iron is limiting for growth for many microbes and for algal growth in the oceans.

In humans, iron deficiency is probably the most frequent organic illness. It affects infants, children, young adults, and the elderly in many populations. Iron deficiency is associated with anaemia and non-haematopoietic disturbances that impair work efficiency and contribute to chronic ill health as well as loss of mucosal integrity; iron-deficiency anaemia is frequently associated with pica, which has important environmental and behavioural associations with hookworm infection. In any event, the prevalence of iron deficiency worldwide provides strong evidence of the critical availability of iron in available nutrients.

Iron may be toxic and excess iron in the tissues is associated with structural injury and functional impairment. The harmful effects of iron are a consequence of its electrochemical properties and potential for the formation of reactive oxygen and nitrogen species which injure cell structures including DNA. There are many causes of excess iron in the tissues but all represent disturbances of iron homeostasis that overwhelm the mechanisms which the body uses to acquire, transport, and store iron safely.

Body iron composition

The total amount of iron in the adult body is between 3 and 4 g—most of which is co-ordinated in protoporphyrin IX as haem (Fig. 1). Haem is found principally as haemoglobin and myoglobin, although appreciable quantities are found in the viscera, especially the liver, kidney, and intestine. Cytochromes of the electron transport chain and of the P-450 system for the metabolism of xenobiotics are abundant in these organs and remarkably selective regions of the brain. In an adult, about 2.5 g of iron is complexed in haemoglobin with an additional 0.5 g as myoglobin in the muscles. In the plasma compartment, very small amounts of iron circulate, bound in the ferric form to the glycoprotein transferrin—this protein is normally only one-third saturated with iron, so that with a mean concentration of 3 g/l for a protein of molecular weight 80 000, it represents less than 2 mg of elemental iron. The normal level of serum ferritin is up to about 250 µg/l, which does not represent an appreciable amount of iron; however, the concentration of ferritin in the serum faithfully reflects the stores of iron in the body. Iron is stored in the mononuclear phagocyte (previously reticuloendothelial) system principally as intracellular ferritin and its proteolytic degradation product, haemosiderin. Body iron stores do not exceed 1.5 g in men and are usually 0.5 g or less in adult women. Non-haem deposits of iron that serve as stores in the iron-rich tissues may be visualized by staining with Perls's reagent (acid potassium ferrocyanide) with which they give a strong Prussian-blue reaction. Faint staining with Perls's reagent may be observed in normal parenchymal liver cells, but in health the principal deposits of storage iron are observed in bone marrow and spleen macrophages as well as in Kupffer cells of the liver.

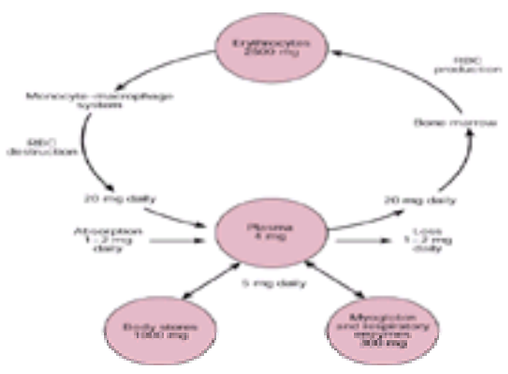

Fig. 1 Daily flux of iron through storage and transport compartments.

Erythropoiesis and iron balance

The mean lifespan of the red cell is 120 days and thus approximately 1 per cent of the steady-state haemoglobin pool is resynthesized each day—this requires *de novo* synthesis of approximately 6 g of haemoglobin into which 20 mg of iron is incorporated. The principal fraction of the iron required for daily haemoglobin production in the basal state is recycled from senescent red cells after their destruction by macrophages; the iron is delivered to the erythron in the plasma by transferrin that binds to cell surface receptors and erythroid precursors (Fig. 1). The transferrin–receptor complex is internalized and, after acidification in endosomes, the iron is released leaving the apotransferrin to be recycled to the surface and re-utilized. Transferrin is the normal mediator of iron delivery and transport in the body.

Under circumstances in which erythropoiesis is stimulated, for example under conditions of reduced oxygen saturation, after bleeding and haemolysis, as well as in dyserythropoietic conditions (including thalassaemia and megaloblastic anaemia), uptake and delivery of iron are greatly increased. Increased delivery of iron occurs in association with an expansion in the number of erythroid precursors that express cell surface transferrin receptors under the influence of the hepatorenal hormone, erythro-poietin.

Iron homeostasis

Iron, an essential nutrient, is fastidiously conserved by the body and only a fraction of that which is utilized in the bone marrow is subject to obligatory daily losses through the exfoliation of epithelia and intercurrent blood loss, such as that incurred in trauma or menstruation. The requirements for iron are met from the diet by the specific absorption of iron in the upper small intestine. The amount of iron available in the diet varies greatly and even under optimal circumstances only a fraction is normally absorbed: in adult men the daily requirement is on average 0.8 mg, whereas in adult women of the reproductive age group, the requirement is usually more than 2 mg daily—the recommended daily allowance in the diet is 10 to 20 mg depending on the bioavailability of food iron components. Inorganic and haem iron complexes are released by digestion; there is a belief that haem iron may be more readily absorbed than inorganic iron in the human intestine and, depending principally on the content in meat, may constitute an important source of iron. Dietary phytates and medication including antacids and tetracyclines, as well as proton pump inhibitors, H_2 antagonists, and prior upper gastrointestinal surgery, all greatly influence the absorption of food iron. The requirement for iron is clearly increased in patients with recurrent bleeding, or in those who are blood donors; iron requirements are also increased during periods of growth in childhood and adolescence. In pregnancy, the daily requirement may be as much as 5 mg and the maternal investment of iron, depending in part on peripartum blood losses, may be as much as 1.5 g—this greatly exceeds the savings due to the cessation of menstruation. Given its iron content, the exsanguination of 1 ml of blood constitutes a loss of approximately 0.5 mg of iron; this relationship facilitates estimates of iron requirements as a result of blood losses, for example those incurred by menorrhagia (more than 80 ml/month) or from other sources.

Iron absorption

In health, iron absorption in the duodenum and upper jejunum is a scrupulously regulated process that matches the acquisition of iron from the diet to body requirements for erythropoiesis and to meet obligatory losses. Under conditions of iron deficiency or on depletion of body iron stores, a greater proportion of bioavailable iron is taken up by the intestine. For reasons that are not fully understood, certain anaemias, particularly those associated with ineffective erythropoiesis and dyserythropoiesis, are also associated with enhanced absorption of iron in the intestine. Where the anaemia is longstanding, for example congenital or acquired sideroblastic anaemia, or in haemoglobinopathies such as b-thalassaemia, intestinal absorption of iron may be such as to cause iron overload leading to tissue injury—secondary haemochromatosis.

The regulation of iron balance by the intestine normally protects the body from iron-rich diets; only under exceptional circumstances, such as the ingestion of alcoholic beverages containing abundant iron as a result of toxic manufacturing processes (for example the kaffir beers that are fermented in iron pots by the South African Bantu), does excess dietary iron lead to iron storage disease. It seems probable that those individuals who develop iron storage disease because of longstanding excessive oral intake do so as a result of the operation of genetic cofactors such as mutant alleles of the adult haemochromatosis gene product, *HFE*, or because of an underlying haematological disorder such as a- or b-thalassaemia trait.

Evaluation of body iron status

Clinical features

The most useful clinical measures of iron status include the detection of pallor and non-erythropoietic manifestations of disease including angular cheilosis, atrophic glossitis, and dystrophy of the nails with longitudinal ridging and koilionychia. Iron deficiency has been associated with behavioural changes in experimental animals. In humans, unusual syndromes of food craving (pica) have been recorded and appear to respond to iron supplementation: this includes craving for soils and the ingestion of silica-rich earths as a cult practice in black populations of the Southern United States—geophagia. Pagophagia (ice-craving) combined with the abnormal taste preferences of pregnancy may account for the bizarre food craving that constitutes part of the folklore of pregnancy. Severe iron deficiency may occasionally be associated with splenomegaly and the signs of underlying disease include peripheral oedema (hypoalbuminaemia associated with massive hookworm infection) and oronasal telangiectasia associated with Osler–Rendu–Weber disease (hereditary haemorrhagic telangiectasia).

Laboratory findings

The most useful measurements apart from those identifying the hypochromic microcytic anaemia associated with abnormal blood cell indices and confirmed by microscopy of the blood film involve surrogate measures of body iron stores. A raised platelet count would suggest a haemorrhagic component associated with iron deficiency anaemia. Iron-deficient erythropoiesis is associated with an elevation of free protoporphyrin in red cell precursors which is reflected in a raised free erythrocyte protoporphyrin in the peripheral blood; this may be easily identified in small samples of blood, taken for example as part of population screening, by the use of portable fluorimeters.

In iron-deficiency anaemia, the absolute concentration of transferrin is raised, with an increase in transferrin iron-binding capacity (**TIBC**). This is reflected by a decrease in serum iron and serum iron transferrin saturation. Such measurements may often serve to discriminate the hypochromic microcytic anaemias of thalassaemia and sideroblastic anaemia from true iron-deficiency anaemia. Such discrimination is necessary before a commitment to iron therapy is ever undertaken.

Measurement of the serum ferritin is often helpful in iron deficiency; serum ferritin concentrations are low, reflecting reduced or absent body iron stores. Neither the serum transferrin saturation nor serum ferritin, however, are absolutely infallible measures of iron deficiency: serum transferrin iron saturation may be artificially elevated with a low transferrin and low serum iron in chronic inflammatory states associated with the anaemias of chronic disorders. Likewise, serum ferritin serves as an acute-phase reactant and may be elevated in malignant disease (especially lymphomas including Hodgkin's disease), or released from the liver in hepatitis and in chronic inflammatory states. Recently there are advocates for the measurement of free circulating transferrin receptors which may be determined by immunoassay. Expression of soluble transferrin receptor protein is enhanced under conditions of iron deficiency and plasma concentrations are elevated in the presence of functionally iron-deficient erythropoiesis; however, greatly increased serum transferrin receptor concentrations are found under conditions of erythroid hyperplasia in the bone marrow and especially when ineffective erythropoiesis occurs (megaloblastic anaemia, haemoglobinopathies, sideroblastic anaemia).

Staining of iron stores in the bone marrow with Perls's reagent is a robust and relatively simple method for resolving difficulties that arise in the investigation of patients with suspected iron-deficiency anaemia. Although an examination of the amount of iron (usually graded semiquantitatively on a scale from 0 to 4, reflecting the strength of Prussian-blue staining) does not provide any information as to the availability of the iron for haemoglobin formation, it does provide useful information as to the appropriateness of iron therapy for hypochromic anaemia. Bone marrow examination, moreover, may be diagnostic in patients suffering from hypochromic anaemias due to primary or sideroblastic change in the marrow, since the characteristic ring sideroblasts with or without other myeloblastic changes will be apparent.

In summary, the laboratory evaluation of patients with suspected iron deficiency should include a full examination of haematological parameters including microscopy of the blood film. Quantification of serum iron, serum transferrin, and transferrin saturation (TIBC) may be valuable in establishing the cause of the hypochromic or microcytic anaemia. Serum ferritin measurements are often confirmatory in the absence of malignant, hepatitic, or other inflammatory diseases—as may fluorimetric red-cell protoporphyrin assays. Determinations of serum transferrin receptor concentration may provide evidence of increased demands for iron by the marrow or indeed expansion of the erythron but this test is not readily available. Microscopic examination of a bone marrow aspirate, including staining with Perls's reagent, may

provide valuable information about iron stores in macrophages and the need for iron supplementation.

Disturbances of iron metabolism

Disorders of iron metabolism are common contributory factors in disease: iron deficiency is rife, particularly amongst the poor and others whose access to meat is limited and those in whom hookworm infestation occurs. At the same time, the prevalence of haemoglobinopathies and other anaemias such as myelodysplasia and sideroblastic syndromes that require transfusion and cause hyperabsorption of iron associated with ineffective erythropoiesis mean that iron storage disease also represents a major world health problem. In addition, as discussed in Chapter 11.7.1, hereditary haemochromatosis occurs at a high gene frequency in certain populations: this includes peoples of North European descent (adult haemochromatosis, due to mutations in *HFE*) and those of sub-Saharan African origin (African iron overload), in whom the nature of the predisposing gene is unknown.

Iron deficiency

About 30 per cent of the world's population, nearly two billion individuals, are anaemic and at least half of this group are believed to have iron-deficiency anaemia. Up to 20 per cent of menstruating females even in rich countries, such as the United States and in Europe, have signs of iron deficiency. In children and young adults, there is a frequency of between 5 and 10 per cent of iron-deficiency anaemia—particularly in deprived socio-economic groups.

Many population studies have in the past been based on erroneous attribution of anaemia solely to iron deficiency: there are many conditions, including the anaemia of chronic disorders and haemoglobinopathies such as b-thalassaemia trait, that lead to hypochromic or microcytic red-cell indices. Population surveys based on the detection of iron-deficient erythropoiesis, especially those using determination of free red-cell zinc protoporphyrin concentrations by fluorimetry, may enhance the detection of true iron-deficiency anaemia; determinations of serum ferritin concentrations also facilitate discrimination between the anaemia of chronic disease and true iron deficiency.

Causative factors

Iron-deficiency anaemia in populations is often attributed solely to an iron-poor diet, but in the absence of significant blood loss or intestinal parasites including hookworm, even the most iron-poor diets rarely cause iron-deficiency anaemia, except in growing children. The amount of iron required to repair obligatory losses is very small so that at least 90 per cent of the iron required for *de novo* haemoglobin formation in erythropoiesis is retrieved from senescent erythrocytes broken down by the mononuclear phagocyte system. Furthermore, once iron deficiency develops, striking adaptive changes occur in the absorptive mechanism for iron in the upper small intestine. In experimental animals with iron deficiency, mucosal expression of the divalent metal transporter 1 (**DMT 1**), on the brush border membrane of the intestinal epithelium, is induced. Iron-deficiency anaemia is also associated with enhanced intestinal expression of mucosal ferrireductase activity.

These changes may not represent the portfolio of adaptive changes that occur. There is evidence that iron deficiency is associated with the recruitment of a greater length of mucosal surface in the upper small intestine for participation in the absorption of luminal iron. Iron deficiency, and the response to the removal of a unit of blood, may increase the overall absorptive efficiency of the intestine for iron up to 10-fold—thus greatly enhancing the bioavailability of dietary iron.

Alcoholic beverages may provide a source of iron and the absorption of haem iron present in red meat, poultry, and fish is usually between 15 and 35 per cent. Between 2 and 20 per cent of non-haem iron present in fruit and vegetable sources is absorbed. Natural enhancers of iron absorption such as ascorbic acid, which maintains ferrous iron in its reduced form in the intestinal lumen, promote direct uptake by DMT 1. Fructose and other organic compounds of low molecular weight also form soluble and reduced complexes with iron released from non-haem sources in food. In the West, normal individuals ingest between about 10 and 15 mg of iron daily. Adult men with normal iron stores absorb approximately 2 per cent of the non-haem iron ingested, whereas men with iron deficiency absorb more than 20 per cent of iron from this source in the diet; the comparable figures for haem iron are 26 and 47 per cent, respectively.

Many compounds present in the diet also inhibit or impede the absorption of iron released by digestion in the lumen. These compounds include tannin, especially present in tea, phytates present in bran and nuts, dietary fibre, and other inhibitory factors such as drugs, including tetracycline and alkalis. Some vegetarians of Asian origin ingest large amounts of phosphate and phytates which inhibit the absorption of iron provided in diets that may contain up to 30 mg of assayable total iron each day. A typical example is spinach which, although rich in iron, leads to the appearance of black stools when consumed in small or moderate amounts; these stools are black because of the passage of iron through the small intestine and its delivery to the colon where it forms insoluble ferrous sulphide complexes through the action of colonic sulphur-reducing bacteria.

Malabsorption of iron

The inability to release and absorb adequate amounts of iron from the diet is an important but unusual cause of iron deficiency. Disease of the stomach, duodenum, and upper jejunum may be responsible for the malabsorption of food iron, which may not be readily detected by studies involving the use of simple radioactive tracer measurements. On the other hand, properly conducted radioactive food labelling studies show that after gastric bypass surgery and after intestinal resection, malabsorption of non-haem and haem food iron sources is the rule. Rarely, iron deficiency may result from inflammatory disease of the upper intestine that causes malabsorption: coeliac disease in infants and adults may be responsible, and the iron deficiency may be combined with deficiency of folic acid. Sometimes large pharmacological doses of iron with or without folic acid may overcome the anaemia caused by coeliac disease but unless a strict gluten-free diet is instituted, the anaemia recurs rapidly after iron therapy is stopped. Although malabsorption of food iron is an important aspect of the iron deficiency associated with coeliac disease, loss of iron exacerbates the effects of malabsorption. In coeliac disease this results from increased exfoliation of the epithelium in association with crypt hyperplasia and bleeding due to ulceration. The abnormal motility and maldigestion associated with upper gastrointestinal surgery compounded by anacidity caused by gastritis or acid-suppressing drugs, also impair the absorption of food iron.

Loss of iron

Women in the reproductive age group lose iron regularly at menstruation. An increased recommended daily allowance for women is higher than in all other groups: the average requirement for healthy menstruating women is approximately 1.4 mg of iron daily to replace losses, compared with normal men who lose about 0.8 to 0.9 mg of iron per day. Pregnancy is often associated with iron deficiency when growth of the fetus is rapid. Twin pregnancies and frequent childbirth, especially in women of low socio-economic groups, are associated with iron-deficiency anaemia. Although anaemia is important, a very large study has been conducted that shows no reliable association between maternal anaemia and the complications of pregnancy, including preterm labour. Pregnancy itself is associated with the development of adaptive responses in the intestine and iron transport proteins that enhance the avidity of the gastrointestinal tract for bioavailable food iron. Clearly socio-economic and sociopolitical considerations are likely to influence the population occurrence of iron deficiency in women of the reproductive age group, particularly since the investment of about 1 to 1.5 g of iron occurs with each pregnancy carried to term. This estimate includes blood loss associated with the birth and the investment of iron placed in human milk, which contains up to 0.5 mg/l of iron bound to the whey protein, lactoferrin.

Other sources of iron loss

Intestinal parasites

Several hundred million people are heavily infested with hookworms. The two common hookworms of humans are *Anclyostoma duodenale* and *Necator americanus*. These helminths attach themselves to the lining of the small intestine by their buccal capsules and cause chronic blood loss by sucking blood from the intestinal villi. Hookworm infestation may be light, so that iron loss is not sufficient to cause iron deficiency. In hookworm disease, involving Old World and New World hookworms, heavy infestation occurs as a result of repeated exposure of the skin to contaminated soil. Mucosal immunity may also be reduced in the susceptible host. Although it is not known exactly what hookworms abstract from human blood, microscopic preparations show red cells expelled from the worm: each *Anclyostoma* induces the loss of up to 300 µl of blood daily, whereas each *Necator* causes the loss of up to 50 µl of blood. Clearly the occurrence of anaemia is dependent on the iron content of the diet, the extent of tissue iron stores, and the duration and intensity of the mucosal helminth infestation itself.

Since up to two-thirds of the haemoglobin iron released by the worms can be reabsorbed in the intestine, significant anaemia requires a very heavy parasite load; none the less, extremely severe anaemia may develop in patients with hookworm disease with all the attendant symptoms of fatigue, dyspnoea, palpitations, and mental changes—including pica. Non-specific abdominal pain may occur and radiographic examination of the intestine or endosccopy may reveal duodenitis with a punctate inflammation associated with partial villus atrophy of the duodenojejunal mucosa. Oedema may result from cardiac failure in severe cases and also in

association with hypoalbuminaemia, since heavy infestation may lead to significant protein-losing enteropathy. Hookworm disease may be associated with other helminth infections such as strongyloidiasis and ascariasis and itself may contribute to poor socio-economic circumstances as a result of incapacity for work due to illness.

Major hookworm parasites are widely distributed in Southern Europe, Africa, the Middle and Far East, and the New World, including the Southern United States. The heaviest infections usually affect rural workers in agricultural communities where repeated exposure occurs in isolated locations and where crops are harvested under conditions of poor sanitation. The iron-deficiency anaemia of hookworm disease may present difficulties for diagnosis when the mucosal inflammation that accompanies heavy infestation is associated with reduction in serum proteins such as albumin and transferrin; this, combined with an acute-phase response, may at first lead to a mistaken diagnosis of the anaemia of chronic disorders.

Other sources of blood loss

The gastrointestinal tract represents an important source of blood loss which should always be considered in patients with iron-deficiency anaemia. Ulcerating lesions of the small and large intestine, including cancers, are frequent causes of iron-deficiency anaemia. However, chronic intermittent bleeding can arise from unusual sources such as Meckel's diverticula, angiodysplastic lesions, hamartomas, and other benign ulcerating tumours such as leiomyomas. Gastric ulcers may be associated with chronic intermittent bleeding, but duodenal ulcers rarely cause chronic gastrointestinal blood loss.

Oesophageal ulceration and inflammatory lesions can cause iron-deficiency anaemia, but precaution is needed in attributing blood loss sufficient to cause iron deficiency to such a source unless other potential sites of bleeding have been excluded. Other unusual sources of gastrointestinal bleeding include multiple telangiectatic lesions of Osler–Rendu–Weber disease (hereditary haemorrhagic telangiectasia)—in which bleeding may occur anywhere from the nasal or oropharynx down to the stomach and upper intestine. The blue bleb naevus syndrome, the Peutz–Jeghers syndrome, and other hereditary gut polyposes are rare causes of chronic gastrointestinal bleeding. Inflammatory disease of the lower small intestine and colon such as Crohn's disease and ulcerative colitis, usually associated with chronic intestinal blood loss, may present with an abdominal history in which iron-deficiency anaemia is prominent. Very occasionally, artefactual iron-deficiency anaemia due to self-bleeding may occur; blood may be removed from any source but bizarre methods may be adopted to conceal it, thus requiring considerable ingenuity, and often detective work, to identify the cause. Because of the striking appearance of expectorated blood, iron-deficiency anaemia associated with frank haemoptosis requires little diagnostic skill, but occasionally recurrent intra-alveolar lung haemorrhage causes unexplained illness and anaemia. Occasionally, iron may be lost in the urine through the kidney in conditions where chronic intravascular haemolysis occurs. Losses may be sufficient to induce iron deficiency in the absence of marked changes in urine colour. Patients with haemolysis due to prosthetic or paraprosthetic cardiac valve malfunction may be revealed by the presence of characteristic red-cell changes; likewise in paroxysmal nocturnal haemoglobinuria, chronic intravascular haemolysis causes chronic urinary iron loss with or without visible haemoglobinuria. In these circumstances, free haemoglobin is released which quickly saturates the capacity of the plasma protein, haemopexin to bind it; free haemoglobin spills into the glomerular filtrate where it is taken up by the proximal tubular cells and degraded. After degradation to haemosiderin, iron is lost in the urine when the iron-loaded epithelial cells are exfoliated.

Clinical and laboratory features of iron deficiency

Symptoms of iron deficiency include fatigue, pallor, palpitations, irritability, and little-recognized mental changes, such as pica. The patient may complain of a sore tongue, deleterious changes in the appearance of hair or hair loss, and angular cheilosis. Examination of the nails may reveal longitudinal ridging and, most often in elderly women with chronic iron deficiency for many years, koilionychia. There may be a complaint of dysphagia associated with the development of an oesophageal web (Patterson–Brown–Kelly or Plummer–Vinson syndrome). This again usually occurs in elderly or middle-aged women with chronic iron deficiency. A small proportion of patients with iron-deficiency anaemia have detectable but modest splenomegaly.

Diagnosis

Blood parameters will reveal microcytic anaemia usually in association with an unequivocal reduction in serum transferrin saturation (below 16 per cent) and a reduced serum ferritin concentration (below 12 µg/l). The absence of these features and of an acute-phase reactive response may suggest dyserythropoietic or sideroblastic anaemia or b-thalassaemia trait. Lead poisoning may be associated with iron deficient indices with or without full-blown sideroblastic changes. A bone marrow aspirate stained with Perls's reagent for iron in marrow macrophages will rapidly confirm reduced or absent stainable iron in the storage compartment and may also be revealing about other aspects of the anaemia, such as the presence of ring sideroblasts, dyserythropoietic features, and/or megaloblastic change.

The presence of immunoreactive serum transferrin receptors may provide additional evidence in favour of iron-deficiency anaemia but because increased serum concentration of these receptors may be observed in several disorders of the bone marrow and the ELISA tests are relatively expensive, the role of this determination in the routine diagnosis of iron deficiency is as yet unestablished. Red-cell zinc protoporphyrin concentrations greater than 35 µg/dl of whole blood are usually observed in patients with iron deficiency; values in excess of 100 µg/dl are generally associated with lead toxicity. Extremely high levels may indicate the presence of erythropoietic protoporphyria or lead poisoning. Modest elevations in erythrocyte protoporphyrin can be observed in patients with haemolytic anaemias, sideroblastic anaemia, and occasionally, the anaemia of chronic disorders.

Investigations and management

The identification of iron-deficiency anaemia should be regarded in a sense as a symptom rather than a diagnosis of a patient's malady: the management of those affected should always include an attempt to determine the cause. Common errors occur when, in elderly patients, iron deficiency is cynically ascribed to the presence of mild oesophagitis or gastritis observed at endoscopy, when the underlying cause is bleeding due to a coincidental but sinister gastrointestinal cancer elsewhere—and for which a diligent search is often required.

A full evaluation of the patient with iron deficiency should include an adequate dietary history including the consumption of drugs, such as aspirin and non-steroidal anti-inflammatory drugs, that may be responsible for gastrointestinal bleeding. An enquiry should be made about additional gastrointestinal symptoms and other signs of blood loss; reasonable attempts should be made to evaluate the extent of menstrual loss, if the bleeding is to be ascribed to menorrhagia in women of the reproductive age group. Attention should be placed on the family and a travel history to exclude causes such as hereditary haemorrhagic telangiectasia or hookworm disease.

Clinical examination should extend from an enquiry about previous gastrointestinal disease or surgery to an examination for visceral enlargement, abdominal lymphadenopathy, splenomegaly, and other features suggestive of intra-abdominal pathology such as portal hypertension and abdominal cancer. Hereditary haemorrhagic telangiectasia may be detected by the presence of the most subtle oronasal lesions.

In patients in which the cause of the iron deficiency is not apparent, further studies may be needed to search for gastrointestinal bleeding, including detection of occult faecal blood on several samples taken consecutively. Endoscopic and radiographic studies of the gastrointestinal tract, and serological studies for the presence of coeliac disease may be required and occasionally there is a need to quantify the amount of blood loss daily in the faeces or during menstrual flow by using radiolabelled chromium red-cell studies. In difficult cases, percutaneous visceral angiography of the coeliac and mesenteric arteries has proved invaluable for detecting sites of active gastrointestinal bleeding that are beyond the reach of conventional endoscopic procedures. In those patients who are actively bleeding, such a procedure can identify local sites of blood loss greater than 0.5 to 1.0 ml/min. Meckel's diverticulum is a potential cause of obscure gastrointestinal bleeding in young adults and children. Some Meckel's diverticula can be diagnosed by scintigraphic studies using technetium-99m labelled pertechnetate which may be concentrated in the ectopic gastric mucosa. Meckel's diverticulum and intestinal strictures, particularly in the ileum, may be occasionally revealed by retrograde colonic contrast radiographic studies. Other diagnostic tests include searching for endomysial antibodies, with confirmatory duodenojejunal biopsy to detect coeliac disease. Examination of the urine and sometimes sputum may be required to detect occult iron loss in exfoliated macrophages or proximal tubular cells, respectively, where intrapulmonary haemorrhage or renal iron loss is suspected.

Sometimes extensive diagnostic procedures fail to identify the cause of iron deficiency when occult gastrointestinal bleeding is responsible. Under these circumstances, it remains appropriate to conduct a diagnostic laparotomy, after consultation with an experienced surgeon, to identify the bleeding lesion. In adults of any age, an appreciable number of obscure gastrointestinal malignancies or treatable benign tumours can be identified by such a procedure which, when combined with angiography with or without enteroscopy, may permit identification of angiodysplastic lesions at remote sites. In younger adults and children, diagnostic laparotomy may be indicated to identify Meckel's diverticula, intestinal stricture, and congenital abnormalities such as duplications that serve as occult sources of

blood loss.

It is not unusual for the patient with recurrent chronic iron-deficiency anaemia to present a challenge for diagnosis. Even the most experienced physician would be well advised to consult widely with colleagues with expertise in radiology, nuclear medicine, and surgery before either prematurely abandoning the search of the causal lesion or requesting an ill-considered laparotomy without a thorough appreciation of the further difficulties it may pose.

Replenishing iron stores is but one aspect of the treatment of iron-deficiency anaemia. Iron should be replaced not only to restore the normal haemoglobin concentration but to replenish body iron stores. It is necessary to replace iron depleted in systemic tissues such as the muscles, where it is an essential component of cytochromes and other enzymes critical for optimal aerobic metabolism. Occasionally a therapeutic trial of oral iron for a defined period may be used to verify the suspected diagnosis of iron-deficiency anaemia. Adequate replacement of iron should be monitored for its effects: a reticulocyte response should be observed in peripheral blood maximally between the 7th and 10th days after initiating treatment and significant increases in blood haemoglobin concentration should be apparent within 2 to 4 weeks. If there is no evidence of continued blood loss, the haemoglobin concentration should come within the normal range within 2 months. Failure to meet these expectations suggests either that the anaemia is not caused by iron deficiency or that there is continued depression of bone marrow function—or that there is bleeding for which further investigation is needed.

Therapeutic preparations of iron

Iron salts should be administered by mouth unless there are overwhelming reasons for using the parenteral route—parenteral preparations of iron are associated with a greatly increased risk of toxicity and hypersensitivity reactions including anaphylaxis. Ferrous salts are better absorbed than ferric salts and show little difference amongst preparations in terms of rate of repair of anaemia at a given dosage of elemental iron.

It is usual to treat iron-deficiency anaemia with at least 100 to 200 mg of elemental iron daily. For full-blown iron-deficiency anaemia, ferrous sulphate is administered three times daily (equivalent to 3 × 65 mg of elemental iron). Some patients are unable to tolerate such a dose of iron because of constipation, diarrhoea, or abdominal pain; the presence of tarry, black stools may interfere with personal hygiene and thus lead to ultimate rejection of iron therapy by the patient. Under these circumstances the dose of iron may be reduced and this, rather than a change of iron salt preparation, usually improves tolerability. The frequency of unwanted effects with ferrous sulphate is the same as that of other iron salts when compared with the amount of elemental iron ingested. Once established, the optimal therapeutic response to oral iron increases the blood haemoglobin concentration by 0.1 to 0.2 g/dl per day. Replenishment of iron has a slow effect on the epithelial changes of iron deficiency and the atrophic glossitis may take several months to improve as iron stores are replenished.

Slow-release oral preparations of iron are available, which the manufacturers often claim release sufficient iron over a 24-h period for optimal haematological responses after once daily dosages. However, these preparations are likely to distribute the iron beyond the upper jejunum and thereby bypass those regions of the intestine in which iron absorption is most avid. Compound preparations of iron including B vitamins and folic acid are available but there is little justification for prescribing these except for prophylactic use in pregnancy (see below). In infants and children, sugar-free preparations of iron complexes are available in the form of polysaccharide iron or iron–sodium EDTA (sodium ironedetate) complexes, which can be used as recommended by the manufacturer. In premature infants, up to 2.5 ml of a syrup containing approximately 5 mg/ml may be used twice daily; up to 5 ml three times daily may be given to children aged 6 to 12 years.

Pregnancy

Prophylactic iron preparations are recommended in pregnant women who have risk factors for iron deficiency such as poor diet, prior menorrhagia, or those in whom gastric surgery has been carried out. Prophylactic iron may also be used in the management of infants of low birth weight including premature babies, twins, and infants delivered by caesarian section. Compound preparations of iron with folic acid may be used for the treatment of iron and folic acid deficiencies in pregnancy. For the prevention of neural tube defects in women planning a pregnancy, the United Kingdom Department of Health advises that a medicinal or food supplement of 400 µg of folic acid daily be taken before conception and during the first 12 weeks of pregnancy. Lone or combined iron compound preparations are not routinely indicated for prophylaxis in patients with chronic haemolysis or in renal dialysis since they may lead, in the circumstances of dyserythropoiesis, to chronic iron overload and secondary haemochromatosis.

Parenteral preparations of therapeutic iron

Given its potential toxicity, the only justification for the use of parenteral iron is in patients who are unable to co-operate with or tolerate oral iron therapy, or those with severe gastrointestinal disease that causes malabsorption or continuing severe blood loss. Provision of iron by the parenteral route does not normally lead to more rapid repair of anaemia than when adequate oral iron preparations are administered. Some patients with renal failure who receive haemodialysis have obligatory blood losses which cannot be treated adequately with oral iron preparations. These patients, and occasional patients receiving peritoneal dialysis, may require intravenous iron regularly. Two parenteral preparations of iron are now available in the United Kingdom: a ferric hydroxide–sucrose complex containing 20 mg/ml of iron (2 per cent) and iron sorbitol citrate that consists of a colloidal stabilized preparation of iron containing 50 mg/ml. Severe sensitivity reactions to these agents may occur and facilities for cardiopulmonary resuscitation should be at hand with the use of iron–sucrose complex. Moreover, administration of these preparations should not be followed by oral iron therapy until at least 5 days after the last injection.

Unwanted and toxic effects of parenteral iron preparations

A history of allergic disorders including asthma, eczema, and prior anaphylaxis are regarded as contraindications to the use of parenteral iron, as is liver disease and concurrent infection. Moreover, these drugs are not recommended for children. Side-effects include nausea, vomiting, taste disturbances, hypotension, parasthesias, abdominal disorders, fever, flushing, anaphylactoid reactions, and the reactivation of inflammatory arthropathies. Injection site reactions, including phlebitis, have been reported. Iron sorbitol is contraindicated in patients with untreated urinary tract infections and early pregnancy as well as liver disease and kidney disease. Parenteral iron should probably be avoided in patients with pre-existing cardiac disease including arrythmias or angina.

Administration

Iron sorbitol is given only by deep intramuscular injection, whereas iron–sucrose complex may be given slowly intravenously or by intravenous infusion. In both instances the total dose is calculated according to body weight and the presumed iron deficit set out in the manufacturer's product literature.

General aspects of iron therapy

Treatment of causes of anaemia, including bleeding, is clearly a critical aspect of the management of iron-deficiency anaemia and its diagnosis. Coeliac disease should be treated with a gluten-free diet; bleeding lesions in the gastrointestinal tract may require definitive surgery directed to their healing. Occasionally, patients with a chronic bleeding disorder for which surgery is not indicated, such as hereditary haemorrhagic telangiectasia, may require long-term iron supplementation at doses less than that required to treat the acute iron-deficiency state. Periodic monitoring is required to ensure that the level of iron replacement is adequate to meet the demands of the bone marrow for *de novo* haem synthesis and that iron overload is not occurring. It should be recognized that relief of iron deficiency will improve many symptoms suffered by a patient even though they may suffer from an incurable underlying disease.

Treatment with iron should be continued until iron stores are replenished: there is no excuse for inadequate therapy—especially in those patients who are likely to suffer recurrent bleeding. Particular attention is needed for iron-deficient patients who have had episodes of acute bleeding treated by blood transfusion and who at the time of therapy are not anaemic. These patients require appropriate iron replacement to replenish iron stores for their long-term restitution of health. Because iron therapy leads to a reduction in the avidity of the transport system of the intestine for iron, it should be continued for several months after the anaemia has been corrected to re-establish appropriate iron stores, ideally as reflected by a serum ferritin determination within the normal range.

Unusual syndromes with iron-deficient erythropoiesis

Congenital deficiency of serum transferrin

There are a few reports of deficiency or virtual absence of serum transferrin in infants with disturbed growth, marked hypochromic anaemia, and disordered iron

metabolism associated with systemic iron storage leading to tissue injury. This disease is extremely rare but holds great fascination for those investigators with an interest in the pathophysiology of iron metabolism. Profound deficiency of serum transferrin disturbs the normal ligand–receptor signalling mechanisms indicated in the overall control of body iron balance and absorption in the intestine. Hypo- or atransferrinaemia in humans appears to be inherited as an autosomal recessive trait; the gene encoding human serum transferrin maps to chromosome 3.

Studies of a naturally occurring mutant mouse, the *hpx* mouse, that also has deficiency of serum transferrin associated with runting and hypochromic anaemia due to iron-deficient erythropoiesis indicate that the disorder responds to infusions of serum transferrin or plasma. These infusions restore normal growth and improve the abnormalities of iron homeostasis; iron-deficient erythropoiesis is also corrected, with resolution of the anaemia. The half-life of transferrin in the plasma is 5 to 10 days and so infusions of plasma or purified preparations enriched with transferrin can be administered at intervals. Since most individuals with transferrin deficiency do express limited amounts of the protein antigen, immune reactions to exogenous human transferrin appear to be either mild or rare. Absolute deficiency of transferrin receptors, for example as occurs in mouse embryos generated as a result of gene disruption technology in embryonic stem cells, is incompatible with normal development beyond the late embryo stage.

Other causes of refractory iron-deficient erythropoiesis

There are sporadic reports of iron deficiency occurring in children and adults for which no cause can be established after intensive investigation. In some instances the expected parameters of iron deficiency associated with iron-deficient erythropoiesis can be demonstrated in individuals who fail to respond to generous oral supplementation with iron salts; administration of parenteral iron, however, leads to an improvement in reticulocytosis with resolution of iron-deficient red-cell indices. Although at the time of writing no molecular lesions have been identified in any of the implicated iron and transport proteins, it is not impossible that disturbed function of DMT 1, ferroportin, haephastin, or as yet uncharacterized moieties involved in the transport of iron across the intestine will be found in these disorders.

Occurrence of iron-deficient erythropoiesis in both females and males that responds only to parenteral iron supplementation is unlikely to be caused by hephastin mutations since this gene maps to the long arm of the X chromosome in humans. It is possible that acquired defects of the intestinal mucosa other than inflammatory disorders may contribute to malabsorption of therapeutic iron. Several young children have been reported with iron-deficiency anaemia refractory to oral therapy but which was corrected by parenteral supplementation. Careful investigation revealed an absorptive defect for iron which was corrected itself by systemic iron supplementation and raises the possibility that severe iron deficiency itself prejudices the ability of the mucosal epithelium in the upper small intestine to carry out its normal absorptive function. However, no further investigations to identify the nature of this acquired metabolic defect have been provided. There is at least one well-documented instance of an acquired defect of iron delivery associated with signs of iron-deficient erythropoiesis caused by loss of human transferrin receptor function. This condition was associated with the development of antinuclear factor and other autoantibodies as part of an autoimmune illness in an adult woman with hypochromic anaemia. Autoantibodies directed against the transferrin receptor were identified in the serum of the patient, but the anaemia with its attendant sideropenia ultimately responded to a combination of steroids and azathioprine therapy; the titre of transferrin receptor autoantibodies of peripheral blood cells diminished. The extent to which this phenomenon occurs generally during the course of autoimmune disorders associated with anaemia is unknown.

Secondary iron storage disease (secondary haemochromatosis)

This is a worldwide problem. It occurs when excess iron is absorbed from the intestine or obtained by the breakdown of transfused red cells in the mononuclear phagocyte system. Each transfused unit of blood contains 200 to 225 mg of iron as haemoglobin. There are instances of iron storage disease occurring in patients who have received oral iron therapy over many years as medicinal tonics or as treatment for refractory anaemia. However, it is unknown if this would occur in the absence of another disorder, such as homozygosity for mutant alleles of the *HFE* gene that predisposes to iron storage disease or underlying bone marrow disease. Conversely, iron excess may develop spontaneously in patients with haemolytic (and especially dyserythropoietic) anaemias alone, although it most commonly results from transfusion with or without underlying bone marrow disease (Table 1).

Each millilitre of human blood contains the equivalent of 0.5 mg of elemental iron complexed with protoporphyrin. Iron present in transfused red cells is eventually retrieved after their breakdown in the macrophage system as a result of the actions of haem oxygenase, which releases bilirubin, carbon monoxide, and one atom of iron per haem molecule; thus each molecule of haemoglobin A yields four iron atoms. Although it is the mononuclear phagocyte system in which significant iron storage is first detected in transfused individuals, continued delivery of iron by its route leads to the excess of iron-loaded ferritin and its breakdown product, haemosiderin, in parenchymal cells throughout the body, with ensuing tissue injury and functional impairment. After the transfusion of 15 to 20 units of blood (representing around 5 g of elemental iron), iron toxicity occurs.

In dyserythropoietic anaemias such as thalassaemia and sideroblastic anaemia, symptoms and signs of iron storage disease may develop early in life and are related to increased dietary iron absorption by the intestine. Although some patients with b-thalassaemia intermedia are treated by occasional transfusion, much of the excess iron stored in the body originates from ingested rather than transfused iron. Iron absorption in healthy adults amounts to 1 to 2 mg/day, but in b-thalassaemia intermedia this may be increased more than fivefold. In regularly transfused patients with b-thalassaemia major, the massive expansion of the erythropoietic marrow may be suppressed to render absorption of iron normal or near normal. However, in patients with thalassaemia who are transfused only intermittently, erythroid hyperplasia persists and excessive absorption of iron from the diet contributes significantly to the iron storage derived from transfused cells; several grams of additional iron may thus be acquired each year.

Patients with hypochromic anaemias due to sideroblastic change in the marrow are particularly at risk because they may be misdiagnosed as suffering from chronic or recurrent iron-deficiency anaemia; they thus receive long-term supplementation with oral iron that serves merely to exacerbate the iron-loading state. It is noteworthy, however, that patients with haemolytic anaemia due to sickle-cell haemoglobin C disease do not commonly develop iron overload as a result of enhanced iron absorption: iron storage disease is thus generally restricted to transfused patients with chronic anaemias. Particular difficulties arise in refractory anaemias in which there is a hyperplastic bone marrow with ineffective erythropoiesis which appears to drive the inappropriate absorption of iron by the intestine.

In the South African Bantu people, the excess iron is ingested in an unusually bioavailable form in beers and other alcoholic drinks prepared by fermentation in iron pots (kaffir beers). Soluble complexes of readily bioavailable iron in these drinks contribute to secondary haemochromatosis, which is common in men in this population and other related sub-Saharan African populations. Although much of the iron is at first detected in the mononuclear phagocyte system (and is seen particularly in Kupffer cells on liver biopsy), associated hypogonadism and vitamin C deficiency later induce scurvy and osteoporosis. Dietary adjustment and iron chelation therapy may relieve the disorder, which is becoming less common after its recognition in the early 1950s. It is of interest that family studies point to a genetic component which predisposes individuals to this secondary iron storage disease within given pedigrees.

The nature of the stimulus leading from the excess iron turnover that accompanies hyperplastic bone marrow to the intestinal disturbance is unknown. The degree of excess iron absorption is however related to the extent of expansion of the red-cell precursor population: blood transfusions, which suppress the marrow, decrease the absorption of food iron. The toxic properties of iron appear to be related to its capacity to participate in free radical-generating reactions that form reactive oxygen and nitrogen intermediates implicated in tissue injury.

Clinical features

The clinical features of secondary iron storage disease in children with chronic anaemias closely resemble hereditary forms of juvenile haemochromatosis (see Chapter 11.7.1). Iron accumulates rapidly in the liver and in the endocrine glands. The several hundred gonadotrophs present within the anterior pituitary gland appear to be particularly susceptible to iron toxicity and hypogonadotrophic hypogonadism results. Iron also accumulates in the b-cells of pancreatic islets, leading to diabetes; in the zona glomerulosa of the adrenal glands, leading to early-onset adrenocortical failure; and in the parathyroid glands, ultimately causing hypoparathyroidism. Secondary iron storage disease also has a predilection for the myocardium. This causes sudden death as a result of tachyarrythmias and injury to cardiac conducting tissue or cardiomyopathy which causes intractable cardiac failure. Secondary iron storage disease in b-thalassaemia and congenital dyserythropoietic anaemias is thus characterized by progressive myocardial disease, endocrine failure, and infantilism.

Untreated iron storage disease is the most common cause of death in these disorders. Similar manifestations of iron toxicity are observed in other patients with secondary iron storage in which the accumulation of iron is less rapid. A picture resembling full-blown adult haemochromatosis ultimately supervenes with complications of diabetes and cirrhosis (sometimes complicated by transfusion-related viral hepatitis and the formation of hepatocellular carcinomas) in the presence of deep skin pigmentation. Secondary iron storage disease represents a significant threat to well-being and prognosis in the chronic anaemias. Once cardiac arrythmias have developed, the outlook is usually bleak and urgent chelation therapy with parenteral desferrioxamine is indicated.

Diagnosis

Secondary iron storage disease should be suspected when the saturation of serum transferrin is greater than 60 per cent. In established secondary iron storage disease, there is a raised non-transferrin iron-binding fraction which may contribute to the tissue injury, since the amount of circulating iron may exceed the binding capacity of circulating transferrin. Under these circumstances, transferrin saturation is usually measured at greater than 90 to 95 per cent and is accompanied by an elevation of serum ferritin which, in the absence of active liver disease, faithfully reflects the extent of iron storage disease and the risks of iron-mediated damage.

Iron chelation therapy should probably be introduced at serum ferritin concentrations greater than 1000 µg/l or if there is biopsy evidence of excess iron storage or a transfusion load of more than 15 units of exogenous red blood cells. Diagnostic evidence of iron storage may be obtained from biopsies of the liver or myocardium; skin biopsy shows excess iron in the sweat gland acini and perifollicular apocrine glands together with increased melanin deposition. In biopsy samples of the liver and heart, histochemical iron storage can be quantified by chemical iron estimations: often the liver iron content exceeds 2 per cent of tissue dry weight (normally less than 0.14 per cent or 7 mg of iron per gram dry weight). Iron concentrations may exceed 5 per cent in affected tissues such as endocrine glands and the pancreas. Liver biopsy may facilitate staging of the disease, particularly in relation to coincidental viral hepatitis where the presence of fibrosis and cirrhosis combined with iron deposits in the parenchymal cell may contribute useful prognostic information. In patients in whom tissue biopsy determinations are not possible, an estimate of body iron overload may be gained by injection of a single dose of 500 mg of desferrioxamine intramuscularly and collection of urine for 24 h in an iron-free plastic container; the daily excretion of more than 2 mg of the coloured ferrioxamine–iron complex indicates iron excess.

Although serum ferritin concentrations generally reflect the amount of iron stored in the tissues, there is a poor correlation between the levels of ferritin in iron-overloaded subjects and clinical outcome. Ferritin concentrations in serum are subject to wide variations; as a result of infection or inflammation (when as an acute reaction it is spuriously elevated), and ferritin concentrations may be reduced when vitamin C is deficient. In contrast, since the liver is the principal site of the iron storage, hepatic iron concentrations provide useful guidance as to prognosis overall, including outcomes from iron-induced cardiac injury, fatal complications of which are usually observed in patients when tissue iron exceeeds 1.5 per cent of dry liver weight. In specialized centres, non-invasive methods have been developed to measure liver iron concentrations, including whole-body magnetic susceptibility techniques but neither this nor sophisticated T_2-weighted magnetic resonance imaging of the heart or liver has been generally accepted in practice. Conventional T_2-weighted imaging may provide a crude assurance that iron storage is either present or under control but is too insensitive to contribute to serial monitoring of secondary iron storage disease—except for the investigation of potential complications such as hepatocellular carcinoma.

Treatment

Patients with homozygous b-thalassaemia and related conditions who are transfusion dependent require adequate blood transfusion to maintain a normal or near normal haemoglobin concentration combined with desferrioxamine as an iron-chelating agent. Long-term studies provide compelling evidence that survival in iron-loaded b-thalassaemic subjects is greatly enhanced by treatment with subcutaneous desferrioxamine, which prevents and reverses the cardiac manifestations of iron storage disease. It must be noted, however, that full compliance with this demanding treatment is required for benefit to accrue, which requires equal commitment from the patient and attending medical and nursing personnel alike. Splenectomy or bone marrow transplantation may be considered in certain cases but is beyond the scope of this article (see Chapter 22.5.7). The overall outcome and prognosis for b-thalassaemia has also been improved by screening donor blood for HIV and hepatitis B and C viruses, as well as other pathogens. These factors are ancillary but may potentiate the development of secondary iron storage disease.

The preferred route for desferrioxamine administration is by slow subcutaneous infusion over 12 to 16 h for up to 7 days per week; this is usually done on an ambulatory basis in adults but nocturnal administration is used particularly in children. Nocturnal administration relies on the use of slow clockwork or battery-operated infusion devices. Although electrical syringe pumps are in common use (such as the Graseby driver device), smaller quieter infusion devices (such as the Cronoject) are now available. Light precharged balloon pumps manufactured by Baxter, though expensive, are also in use. The total daily dose of desferrioxamine is usually set at 20 to 30 mg/kg of body weight with the maximum usually determined by the extent to which near-saturated solutions of the drug can be tolerated by the patient. In patients without cardiac disease it has been shown that the daily oral administration of ascorbic acid at 2 to 3 mg/kg increases the amount of iron that can be chelated by desferrioxamine. Serial determinations of serum ferritin concentrations, combined with regular clinical monitoring and assessment of cardiac, hepatic, and endocrine function assist in the assessment of iron storage disease and the efficacy of iron chelation therapy. Periodic echocardiograms and electrocardiography, with 24-h ECG monitoring, are desirable aspects of management. Urinary excretion of the coloured ferrioxamine complex can be easily measured by light spectroscopy. Desferrioxamine promotes not only urinary excretion of iron but also chelates iron from the body stores, which is excreted into the faeces via the biliary system.

Several studies show that patients with b-thalassaemia maintained on adequate transfusion regimes who are able to tolerate their infusions of subcutaneous desferrioxamine, grow and develop normally and have a better prognosis than those who either default from or do not comply fully with the chelation regimen. When treatment is initiated, careful monitoring is needed using 24-h urine collections for iron measurements to judge the excretion of iron as the dose of desferrioxamine is escalated. Daily doses of desferrioxamine may be increased to about 50 mg/kg of body weight; this usually represents the maximum that can be tolerated. In infants and growing children, unless severe cardiac disease or iron overload is present, the dose should not exceed 35 mg/kg per day over 5 nights each week. Thereafter, most well-transfused patients with b-thalassaemia can be maintained in negative iron balance by the use of not more than 40 mg/kg. For patients who receive blood transfusions, a single intravenous infusion of desferrioxamine given separately from but at the same time as each blood transfusion, at a dose of approximately 150 mg/kg of body weight, also contributes to the control of iron storage disease. In patients who develop endocrine failure, prompt replacement of deficient hormones should be introduced. Sex-steroid hormone replacement may relieve infantilism and improve self-esteem in developmentally arrested adolescents and children.

Desferrioxamine is usually well tolerated and, apart from minor skin reactions, is remarkably non-toxic. These reactions can usually be controlled by lowering the concentration of the drug in the infusion and by alternating sites of infusion; hydrocortisone in doses of up to 10 mg has been reported to reduce severe cutaneous reactions. Very high doses of desferrioxamine, particularly those used for treatment of life-threatening cardiac iron overload and given by intravenous rather than subcutaneous infusion (see below), have been associated with retinal injury and lens opacities as well as hearing loss. Since high-tone hearing loss may occur also, it may be prudent to monitor visual acuity and auditory function at intervals during treatment over the years for which desferrioxamine is required. Minor gastroenterological disturbances, myalgia, and very rarely anaphylaxis may occur; rapid administration of desferrioxamine may be associated with hypotension, especially when given intravenously. Desferrioxamine interacts unfavourably with phenothiazines and coma may result from its use in patients receiving these agents. Some patients receiving desferrioxamine develop infections with micro-organisms such as *Yersinia* and fungi such as *Mucor* that have fastidious requirements for iron. Iron-overloaded patients may also develop other systemic microbial infections and are particularly susceptible to infections with the marine vibrio, *V. vulnificus*. It seems likely that under these circumstances the desferrioxamine may serve, as nature intended, as a source of iron for uptake by microbial siderophore systems.

In patients with acute or subacute cardiac manifestations of iron overload, there are encouraging reports of the effects of high-dose intravenous desferrioxamine: desferrioxamine may reverse cardiac failure and life-threatening tachyarrythmias. An oral iron chelator of a different chemical class from the naturally occurring bacterial agent desferrioxamine has recently been licensed for treatment of iron overload in patients unable to tolerate desferrioxamine or in whom it is contraindicated. This drug, of the hydroxpyridone class, Deferiprone, is used at a dose of 25 to 100 mg/kg of body weight daily in three divided doses; the agent is not recommended for children under the age of 6 years.

Deferiprone appears to induce overall negative body iron balance in patients with severe homozygous b-thalassaemia with attendant reductions in serum ferritin concentrations and clearly represents the first newly licensed oral drug with this important indication. In a proportion of patients, however, negative iron balance does not appear to be maintained and the drug may cause serious toxicity including neutropenia and the occasional incidence of agranulocytosis which appears to be mediated by an immune mechanism. The use of Deferiprone appears to be somewhat controversial following a recent report that its continued administration may be associated with progressive hepatic fibrosis. Conversely, despite the inconvenience of its use, long-term studies of patients receiving desferrioxamine for iron storage disease in homozygous b-thalassaemia show that it is largely safe; moreover desferrioxamine improves cardiac function and life expectancy and arrests hepatic fibrosis in secondary haemochromatosis. Safety information and a side-effect profile on the use of Deferiprone at a daily dose of 75 mg/kg is available.

Other aspects of care

The single most important aspect of care is compliance with iron chelation therapy and monitoring—especially for infants and other young patients with iron-loading anaemias such as thalassaemia. Regular attendance of special clinics is advisable so that wide-ranging professional support from familiar personnel can be given to reinforce medical care delivered with attention to continuity and the nurturing of independence.

Patients with secondary iron overload should be monitored not only for the progression of their iron storage as determined by parameters of iron metabolism but also clinically for the presence of iron-mediated tissue injury. Regular echocardiography, electrocardiography, hormone measurements, and physical examinations are required to search for the presence of endocrine failure, including hypoparathyroidism and adrenocortical failure, both of which may be very difficult to detect. Patients with evidence of hypogonadism should be treated with hormone supplementation to ensure normal sexual characteristics and vigilance should be maintained for the development of diabetes mellitus. Psychological difficulties are prevalent in children and adolescents receiving iron-chelation therapy and transfusion for chronic anaemias and appropriate counselling is often needed over long periods to build up trust with them and their families and to maintain compliance with treatment. Patients with established infantilism and stunted growth frequently develop skeletal disease in addition to that related to their marrow disorder and investigations should be carried out to search for osteopenia and osteoporosis for which additional therapy will be needed. Bone disease and growth arrest may be caused by the overenthusiastic use of desferrioxamine in young infants, and in these patients the daily dose of desferrioxamine should be reduced to below 40 mg/kg, which usually restores growth velocity to normal.

Finally, patients with secondary iron storage disease should be advised to moderate their dietary intake of iron-rich foods such as meat: some investigators advocate the drinking of strong tea at meal times, especially in patients with thalassaemia intermedia. This tannin-rich drink has been shown to decrease bioavailability of dietary iron and should improve overall iron balance in this at-risk group. As far as possible, the blood haemoglobin concentration should be maintained in the normal range to ensure growth and responsiveness to hormone supplements; patients with significant transfusion requirements should be considered for splenectomy when they reach an age of over 5 years. As with patients who are not iron overloaded, splenectomized individuals should be treated appropriately by immunization and antimicrobial prophylactic therapy as far as possible to reduce the risk of intercurrent bacterial infection. This risk is potentiated by systemic iron storage.

Treatment of severe cardiac manifestations of iron storage disease

Continuous intravenous infusions of desferrioxamine not exceeding 50 to 60 mg/kg daily are now recommended for life-threatening heart disease. High-dose intravenous infusions may cause unacceptable toxic injury, especially in the retina and inner ear. Desferrioxamine given continuously through a permanent indwelling portable catheter within the superior vena cava, with careful attention to sepsis, is a satisfactory method for securing reversal of cardiac disease in high-risk patients with serum ferritin concentrations that persist at greater than 2500 µg/l or who have hepatic iron concentrations that exceed 1.5 per cent of dry liver weight. Improved outcomes have been reported with the use of anticoagulation induced by warfarin, and scrupulous attention to cutaneous needle re-siting and skin care to reduce the risk of thrombosis and complicating infections.

Pregnancy

Desferrioxamine therapy is not recommended by the manufacturer during pregnancy but despite this, many successful pregnancies have been reported without fetal injury. The drug should probably be avoided during the middle trimester and should almost certainly be avoided, because of unknown teratogenicity, in early pregnancy or at the time of any planned conception. None the less, it may be reasonable to restart desferrioxamine therapy in the final trimester of pregnancy if the risks to the mother from iron storage disease are high. No information is available on Deferiprone in pregnancy and it should probably not be used until more experience with the drug is forthcoming.

Prognosis and outcome

The principal causes of death in secondary iron storage disease include cardiac failure and arrythmias, endocrine failure and the consequences of diabetes mellitus, infection, and hepatocellular carcinoma. Unless treated, secondary haemochromatosis is a rapidly fatal disease when associated with transfusion therapy and intestinal hyperabsorption of iron in the chronic anaemias. Less than one-third of those unable to comply with iron chelation therapy survive with b-thalassaemia major to the age of 25 years. However, the outcome of iron storage disease in patients with chronic anaemia is now greatly improving, with enhanced life quality and duration. One study has indicated that 95 per cent of patients with b-thalassaemia who administer desferrioxamine subcutaneously more than 250 times each year will survive to 30 years; whereas only 12 per cent of those who do not will survive to this age. In the United Kingdom the overall survival is 50 per cent at 35 years, but at one specialist centre the actuarial survival in more than 100 patients was 80 per cent at 40 years. This again emphasizes the benefits of care administered at a dedicated treatment centre. Several reports also show that the frequency of hypogonadism, diabetes, and growth retardation is significantly reduced by effective iron chelation. Continuous intravenous desferrioxamine can be claimed to reverse life-threatening arrythmias in cardiac iron overload and also improve or reverse left ventricular or biventrical heart failure in a majority of cases. One report describes the actuarial survival of more than 60 per cent at 13 years of patients with life-threatening disease and b-thalassaemia so treated; this outcome appears to be accompanied by improved cardiac tissue iron signals on magnetic resonance imaging.

Further reading

Adamkiewicz TV *et al*. (1998). Infection due to *Yersinia enterocolitica* in a series of patients with b-thalassaemia: incidence and predisposing factors. *Clinical Infectious Disease* **27**, 1362–6.

Andrews NC (1999). Disorders of iron metabolism. *New England Journal of Medicine* **341**, 1986–95.

Bothwell T *et al.* (1989). Nutritional iron requirements and good iron absorption. *Journal of Internal Medicine* **226**, 357–65.

Chen FE *et al*. (2000). Genetic and clinical features of haemoglobin H disease in Chinese patients. *New England Journal of Medicine* **343**, 544–50.

Cohen AR *et al.* (2000). Safety profile of the oral iron chelator Deferiprone: a multi-centre study. *British Journal of Haematology* **108**, 305–12.

Cox TM (1998). Iron salts, iron–dextran complex and iron–sorbitol citrate. In: Dollery CT, ed. *Therapeutic drugs: a clinical pharmacopoeia*, 2nd edn, Vol 2, pp 178–83. Baillière Tindall, Edinburgh.

Dallman PR (1989). Iron deficiency: does it matter? *Journal of Internal Medicine* **226**, 367–72.

De Maeyer EM (1989). *Preventing and controlling iron deficiency anaemia through primary health care*. World Health Organization, Geneva.

De Maeyer EM, Adiels-Tegman M (1985). The prevalence of anaemia in the world. *World Health Statistics Quarterly* **38**, 302–16.

Finch C (1994). Regulations of iron balance in humans. *Blood* **84**, 1697–702.

Gordeuk VR, Boyd D, Brittenham G (1986). Dietary iron overload persists in rural sub-Saharan Africa. *Lancet* **i**, 1310–13.

Kent S (2000). Iron deficiency and anaemia of chronic disease. In: Kiple KF, Ornelas KC, eds. *The Cambridge world history of food*, pp 919–39. Cambridge University Press, Cambridge.

McCance RA, Widdowson EM (1937). Absorption and excretion of iron. *Lancet* **223**, 680–4.

Modell B *et al.* (1982). Survival and desferrioxamine in thalassaemia major. *British Medical Journal* **284**, 1081–4.

Moore DF, Sears DA (1994). Pica, iron deficiency and the medical history. *American Journal of Medicine* **97**, 390–3.

Olivieri NF *et al.* (1994). Survival in medically treated patients with homozygous b-thalassaemia. *New England Journal of Medicine* **331**, 574–8.

Olivieri NF *et al.* (1998). Long-term safety and effectiveness of iron-chelation therapy with Deferiprone for thalassemia major. *New England Journal of Medicine* **339**, 417–23.

Pippard MJ (1989). Desferrioxamine-induced iron excretion in humans. *Baillière's Clinical Haematology* **2**, 323–43.

Pippard MJ, Weatherall DJ (1984). Iron absorption in iron-loading anaemias. *Haematologia* **17**, 407–14.

Pippard MJ, Weatherall DJ (2000). Oral iron chelation therapy for thalassaemia: an uncertain scene. *British Journal of Haematology* **111**, 2–5. [A useful review of iron chelation and a dispassionate evaluation of the emerging role of deferriprone.]

Porter JB (2001). Practical management of iron overload. *British Journal of Haematology* **115**, 239–52. [An excellent contemporary review with abundant practical as well as theoretical and scientific information.]

Roche M, Layrisse M (1966). The nature and cause of hookworm anaemia. *American Journal of Tropical Medicine* **15**, 1029–102.

22.5.5 Normochromic, normocytic anaemia

D. J. Weatherall

A mild normochromic anaemia is one of the commonest findings in every branch of clinical practice. It is important to decide whether the anaemia is of significance and how far it should be investigated.

The first decision to be made is whether the blood findings represent 'anaemia' for the particular patient. The haemoglobin level varies considerably at different ages and there is a wide range of 'normal' values for any particular age. Knowledge of any previous blood count is particularly useful since a haemoglobin value in the lower range of normal may represent anaemia in a patient previously known to have a higher haemoglobin when in good health.

Most of the normochromic, normocytic anaemias are secondary to other diseases; a minority reflect a primary disorder of the blood. The most common causes are summarized in Table 1.

Anaemia of chronic disorders (ACD)

This is the rather unsatisfactory phrase used to cover the most common of the normochromic, normocytic anaemias, namely, those found in association with chronic infection, all forms of inflammatory diseases, and in malignant disease. It is very important for clinicians to be able to identify the main features of this type of anaemia. Although it may be extremely mild and asymptomatic, the presence of this blood picture should always alert the clinician to the possibility of there being a serious underlying disease.

Pathogenesis

The precise mechanism of the anaemia of chronic disorders is still not understood. Several different pathological processes that occur in response to inflammation conspire to cause a defective proliferation of red cell progenitors. In addition, at least in some cases, there may be a mild haemolytic component.

The most constant feature of ACD is a low serum iron level despite adequate iron stores in the reticuloendothelial elements of the bone marrow. This abnormal accumulation of iron in the storage cells, together with a low serum iron level in the blood, suggests that there is a block in the release of iron to the developing red cell precursors. This phenomenon may be observed within 24 h after major surgery, for example. There is also a reduced concentration of transferrin, and turnover studies suggest that this reflects a decreased rate of production.

Several studies have found a mild shortening of the red cell life span in ACD, which appears to be due to an extra corpuscular factor and not to an intrinsic abnormality of the red cells. The red cell survival is not grossly shortened; if marrow function were normal it should be able to compensate for the reduced red cell survival. However, there is a defect in the proliferation of red cell progenitors in ACD. This may reflect inadequate iron delivery or the effect of cytokines produced as a response to infection, or both. There also seems to be a subnormal erythropoietin response for the degree of anaemia, possibly arising from the action of various cytokines.

Recent studies have identified a number of cytokines that inhibit haemopoiesis in bone marrow culture and reduce the output of erythropoietin in hepatoma cell lines. The relevance of these *in vitro* studies to the generation of ACD in patients with such a diversity of associated disorders is uncertain. It is very unlikely that one mechanism will be found to account for such diverse abnormalities. Rather, it appears that ACD is a by-product of the acute phase reaction, probably augmented by a variety of different cytokine responses.

Clinical and laboratory findings

The anaemia of chronic disorders is usually mild. In patients with severe inflammation the haematocrit may fall to levels at which symptoms are experienced. Although the anaemia is usually normocytic and normochromic there may be mild hypochromia with a slight reduction in the MCH and MCV, particularly in children. Occasionally there may be marked microcytosis. Microcytosis should prompt consideration of concomitant iron deficiency, especially in patients who might have gastrointestinal bleeding, for example individuals with inflammatory bowel disease or rheumatoid arthritis on aspirin. The reticulocyte count is in the normal range.

The most important finding is a reduction in the serum iron concentration. Because there is a concurrent reduction in the level of transferrin, the per cent saturation of the iron binding capacity is usually normal or only slightly reduced. This observation clearly distinguishes ACD from true iron deficiency anaemia (Table 2). This distinction can also be confirmed by measuring the serum ferritin level, which is usually in the normal range or slightly elevated in patients with ACD while it is low in those who are iron deficient.

The bone marrow appearance is unremarkable. There may be a slight deficiency of red cell progenitors. Iron staining shows a paucity of iron in the red cell precursors and an accumulation of iron in the storage elements of the marrow. Again, this distinguishes ACD from true iron deficiency in which there is an absence of both sideroblasts and storage iron. The abnormal distribution of iron in an adequately stained sample, together with the low serum iron level, is the true hallmark of ACD and a finding that should always be followed up by a search for an underlying inflammatory or neoplastic condition.

Other forms of normochromic, normocytic anaemia

Other causes of this type of blood picture are summarized in Table 1.

Renal failure

Normochromic, normocytic anaemia is a common presenting feature of renal disease. The features of the anaemia of renal failure are discussed in more detail in the section on blood changes in systemic disease.

Endocrine disease

The hypometabolism observed in hypopituitary and hypothyroid states reduces demand for oxygen in the tissues and, therefore, output of erythropoietin. This is probably the major factor in the development of the mild normochromic, normocytic anaemia which is observed in some patients with these conditions.

Bone marrow failure

Non-specific anaemia is a common feature of bone marrow failure. It may occur in the pure red cell aplasias or as part of aplastic anaemia.

Acute blood loss and early iron deficiency

Blood loss from the gastrointestinal or genitourinary tract may be sufficient to cause anaemia but as long as the iron stores are sufficient to maintain an output of normal red cells the anaemia is normochromic and normocytic. This picture is seen in early cases of bleeding or intermittent bleeding. There is usually a slight increase in the reticulocyte count, reflecting an increased rate of proliferation of red cell progenitors.

Polymyalgia rheumatica and giant cell arteritis

Polymyalgia rheumatica (see Chapter 18.10.4) is nearly always associated with a moderate normochromic, normocytic anaemia together with a marked increase in the erythrocyte sedimentation rate. However, particularly in elderly patients, anaemia may be the presenting feature. The symptoms of polymyalgia or cranial arteritis may be minimal or even absent. This common variant of the polymyalgia syndrome should always be considered in old people with anaemia and a very high sedimentation rate who do not have paraprotein in the blood. The anaemia responds quite dramatically to corticosteroids.

Management

Mild, non-specific anaemias should always be investigated because they may be the first indication of a serious underlying disease. It is important to try to distinguish ACD from iron deficiency or other non-specific normochromic, normocytic anaemias. In ACD, the serum iron level is low and there is a normal saturation of the iron binding capacity. It is worth carrying out a bone marrow examination to study the distribution of iron between the red cell precursors and the storage cells. If the pattern of ACD is observed, it is important to carry out a careful search for chronic inflammation or neoplastic disease. The commonest causes of ACD which give rise to diagnostic problems are low-grade urinary infections, chronic sinus infection, and occult malignancy.

The treatment of anaemias of this type is essentially that of the underlying disease. In the subgroup of elderly patients presenting with this type of anaemia in association with a very high sedimentation rate, in whom underlying blood dyscrasias and paraproteinaemias have been ruled out, it is justifiable to give a therapeutic trial of corticosteroids; and to proceed to further investigations only if there is no immediate and dramatic response characteristic of the polymyalgia syndromes. Early recognition of the true diagnosis may save weeks of fruitless investigation for a non-existent neoplasm.

The major problem for the management of this condition is encountered in those case in which it is impossible to correct the underlying disorder, patients with advanced malignant disease or intractable rheumatoid arthritis, for example. It has been found that the quality of life is undoubtedly improved for many patients of this type if the haemoglobin level is raised. This may be achieved by instituting a regular blood transfusion regimen. As an alternative approach, a number of disorders of this type have been treated with erythropoietin at varying doses. A limited number of trials, some of which were placebo-controlled, have suggested that at least some patients with malignant disease, rheumatoid arthritis, or AIDS experience a useful rise in the haemoglobin level using this approach. In view of the cost of this treatment, further studies of its efficacy are required.

Further reading

Gardner LB, Benz EJ (2000). Anemia of chronic diseases. In: Hoffman R, Benz EJ, Shattil SJ, Furie B, Cohen HJ, Silberstein LE, McGlave P, eds. *Hematology, basic principles and practice*, 3rd edn, pp. 383–8. Churchill Livingstone, New York and London.

22.5.6 Megaloblastic anaemia and miscellaneous deficiency anaemias

A. V. Hoffbrand

Introduction

The megaloblastic anaemias are a group of disorders characterized by a macrocytic anaemia and distinctive morphological abnormalities of the developing haemopoietic cells in the bone marrow. In severe cases, the anaemia may be associated with leucopenia and thrombocytopenia. Megaloblastic anaemia arises because of inhibition of DNA synthesis in the bone marrow, usually due to deficiency of one or other of two water-soluble B vitamins, vitamin B_{12} (B_{12}, cobalamin) or folate. B_{12} deficiency may also cause a severe neuropathy but whether this occurs with folate deficiency is controversial. In a minority of cases, megaloblastic anaemia arises because of a disturbance of DNA synthesis due to a drug or a congenital or acquired biochemical defect that causes a disturbance of B_{12} or folate metabolism or affects DNA synthesis independent of B_{12} or folate. B_{12} and folate are discussed first and the other rare megaloblastic anaemias are mentioned at the end of this chapter.

Biochemical and nutritional aspects of vitamin B_{12} and folate

Vitamin B_{12}

Biochemistry

Four major forms of the vitamin exist in man, all with the same cobalamin nucleus, which consists of a planar corrin ring (hence the term 'corrinoids' for B_{12} compounds) attached at right-angles to a nucleotide portion, 5,6-dimethylbenzimidazole joined to ribose-phosphate (Fig. 1; Table 1). 5'-deoxyadenosyclobalamin (ado-B_{12}) accounts for about 80 per cent of B_{12} inside human and other mammalian cells and is mainly in mitochondria; methyl-cobalamin (methyl-B_{12}) is a minor component in cells but the main form in plasma. Both are extremely light-sensitive and are photolysed to hydroxocobalamin (hydroxo-B_{12}) within 10 s of exposure to daylight; hydroxo-B_{12} is present in small amounts in tissues and plasma and is available commercially for therapeutic use. The fourth form, cyanocobalamin (cyano-B_{12}), is present only in traces in nature, but is stable and is used radioactively labelled with cobalt-57 or cobalt-58 for *in vitro* and *in vivo* studies of B_{12} metabolism. Hydroxo- and cyano-B_{12} are converted, after two reduction steps in cells of the body, to the two biochemically active forms. The fully reduced compounds are termed Cob(I)alamins, and the oxidized compounds Cob(III)alamins. Analogues of B_{12} (pseudo-B_{12}s) exist in nature and have a different sugar (cobamides) or no nucleotide portion (cobinamides), or alterations in the corrin ring. The source and identity of analogues in human serum is unclear. Endogenous production is suggested by their presence in all sera (including fetal serum) and their fall in parallel with physiologically active B_{12} in B_{12} deficiency.

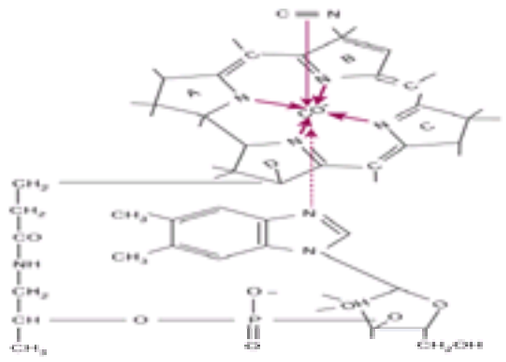

Fig. 1 The structure of cyanocobalamin.

B_{12} is known to be involved in only three reactions in human tissues: as ado-B_{12} in the isomerization of methylmalonyl CoA to succinyl CoA and of a-leucine to b-leucine, and as methyl-B_{12} in the methylation of homocysteine to methionine, a reaction that also requires methyltetrahydrofolate (Fig. 2). In some bacteria, but not

in man, B_{12} has a direct role in DNA synthesis by virtue of its involvement in ribonucleotide reductase.

Fig. 2 Biochemical reactions of vitamin B_{12} in human tissues.

Nutrition

Vitamin B_{12} is synthesized by micro-organisms; animals obtain it by eating parts of other animals or animal produce (milk, cheese, eggs, etc.), or vegetable foods contaminated by bacteria. Clean vegetables, fruit, nuts, and cereals do not contain B_{12}; cooking has little effect on it. A normal mixed diet contains between 5 and 30 µg daily, the amount increasing with the quality. In some species, but not in man, B_{12} is absorbed after synthesis by bacteria in the large intestine. Total B_{12} in man is about 3 to 5 mg, which is mainly stored in the liver (about 0.7–1.1 µg/g). Adult daily losses are related to body stores; to maintain normal body stores, daily requirements are of the order of 2 µg. It takes 3 to 4 years, on average, for deficiency to develop if supplies are totally cut off by malabsorption. There is an enterohepatic circulation for B_{12}, variously estimated at 3 to 9 µg daily, that is intact in vegans, which may partly account for their tendency to maintain low body stores without progressing to severe deficiency. The body is unable to degrade B_{12} and deficiency has not been shown to be due to excess utilization or loss.

Absorption

About 15 per cent of food B_{12} is available for absorption. It is released from protein binding in food by proteolytic enzymes, heat, and acid, and combines one molecule to one molecule with a glycoprotein 'R' B_{12}-binding protein (also called 'haptocorrin') in gastric juice. This protein is related to plasma transcobalamin I. It binds food B_{12} but does not facilitate its absorption. Pancreatic trypsin is needed to degrade this protein and so release B_{12} for attachment to intrinsic factor (IF) and subsequent absorption. The 'R' binder, unlike IF, also binds B_{12} analogues in food. IF is a glycoprotein produced by the parietal, and possibly other, cells of the stomach (Table 2). The IF gene has been localized to chromosome II. Glycosylation of IF is not required for its B_{12} or ileal receptor binding but may play a part in protecting it from digestion by pancreatic proteases in the intestinal lumen. The normal stomach produces a vast excess of IF, measured in units (1 unit binds 1 ng B_{12}). B_{12} in bile is also attached to IF and reabsorbed through the ileum. At neutral pH, in the presence of calcium ions, the B_{12}–IF complex attaches passively to a specific IF receptor, cubulin, on the brush border of the mucosal cells of the terminal ileum. Cubulin is a 460-kDa, 3597 amino acid, peripheral membrane protein present in the epithelium of intestine and kidney. It shows high-affinity, calcium- and cobalamin-dependent binding of IF–cobalamin. The cDNA encodes a precursor protein that undergoes proteolytic processing due to cleavage at a recognition site (Arg ^{7}-Gen^{8}-Lys^{9}-Arg^{10}) for the trans-Golgi proteinase furin. The gene is on the short arm of chromosome 10.

After cubilin-mediated endocytosis, IF undergoes lysosomal degradation. After a delay of 3 to 5 h, B_{12} appears in portal blood, with a peak level 8 h after ingestion, complexed with transcobalamin (TC)II secreted into the circulation from the basolateral side of the intestinal cells. IF itself is digested by the cell and is not absorbed. Ileal absorption of B_{12} is limited, by the number of ileal receptors, to a few micrograms daily and although 80 per cent of a single dose of 1 to 2 µg may be absorbed, the proportion diminishes steeply at higher doses. A small (less than 1 per cent) trace of a large (1 mg or more) dose of B_{12} can be absorbed passively and rapidly through the buccal, gastric, and duodenal mucosae without IF participating.

Transport

Vitamin B_{12} in plasma is 70 to 90 per cent attached to a glycoprotein, TC I, which does not enhance cell uptake of B_{12} (see Table 2). It is one of a group of glycoproteins, the 'R' binders or haptocorrins (see above), that are present in many tissues and fluids (e.g. gastric juice, saliva, tears, milk, and colostrum) and have the same amino acid composition but differ in the composition of the carbohydrate moiety. The haptocorrins may have the role of binding analogues of B_{12} derived from food or intestinal organisms and transporting them to the liver for excretion in the bile. A closely related haptocorrin, TC III, also occurs in human plasma and is probably derived from specific granules of neutrophils. It normally carries only 0 to 10 per cent of plasma B_{12}.

The most important plasma B_{12}-binding protein, TC II, is synthesized in macrophages, liver, the ileum and possibly endothelium. TC II gains B_{12} from the ileum and by release of free B_{12} from the liver and other organs. It is normally almost completely unsaturated; however, it actively enhances uptake of B_{12} by bone marrow, placenta, and other tissues of the body that contain receptors for it. The receptor is a dimer of molecular weight 124 000. TC II–B_{12} is internalized by endocytotosis; B_{12} is split off in lysosomes but TC II is not reutilized (Table 1). TC II accounts for most of cell B_{12} uptake; it has 20 per cent amino acid homology and greater than 50 per cent nucleotide homology with human TC I and with rat IF. The regions of homology common to all three proteins are located in seven domains and it is likely that one or more of these are involved in cobalamin binding. Presumably three-dimensional differences at the ligand-binding site of the proteins exist to explain the different affinities of these proteins (TC I < TC II < IF) for cobalamin analogues. At least five genetic variants of TC II exist, distinguished by their electrophoretic mobility, probably reflecting autosomal polymorphism with many codominant alleles at one locus. Serum TC II is normally higher in women than men and in blacks than whites. The concentration of B_{12} in cerebrospinal fluid is low, with a mean of 10 ng/l in normal subjects. Most of this is attached to TC II. There is virtually no B_{12} in normal urine.

Folate

Biochemistry

This vitamin exists in nature in over 100 forms, all of which are derivatives of folic acid (pteroylglutamic acid), which consists of a pteridine, a *para*-aminobenzoic acid moiety and L-glutamic acid (Fig. 3). Natural folates may differ from folic acid by:

1. being reduced in the pteridine ring to di- or tetrahydo- forms;
2. having a single carbon moiety attached at positions N_5 or N_{10} (e.g. methyl, formyl, etc.); and
3. having a chain of glutamate moieties attached by g-peptide bonds to the L-glutamate moiety.

In human and other mammalian cells, the number of glutamates is mainly four, five, or six. Metabolism of pteroylmonoglutamates to polyglutamates by the enzyme folylpoly-g-glutamate synthetase allows tissues to concentrate folates. In addition, folypolyglutamates are the active coenzyme forms with increased affinity or lowered K_m values for most of the enzymes of one-carbon metabolism. In body fluids (plasma, cerebrospinal fluid, bile, milk, etc.), however, folates are invariably monoglutamate derivatives. In plasma, 5-methyltetrahydrofolate (methyl-THF) predominates.

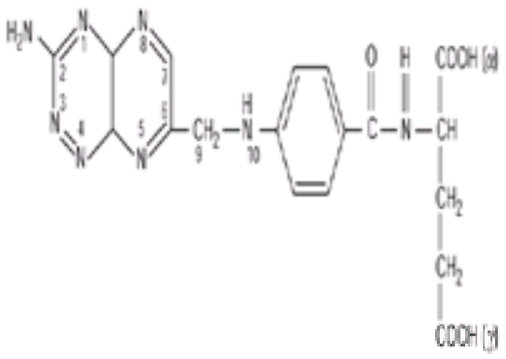

Fig. 3 The structure of pteroylglutamic (folic) acid.

The biochemical reactions of folates are shown in Table 3. In each there is transfer of a single carbon group, methyl ($-CH_3$), formyl (–CHOH), methenyl (ºCH), methylene ($=CH_2$), or formimino (=CHNH), from one compound to another. Three of the reactions are concerned with synthesis of DNA precursors (two in purine and one in pyrimidine synthesis). During thymidylate synthesis, oxidation of folate to the dihydro state occurs; the enzyme dihydrofolate reductase, the major target for the antifolates methotrexate and pyrimethamine, returns folate to the active tetrahydro state (Fig. 4). During its reactions, folate is not completely reutilized, some degradation at the C_9–N_{10} bond occurs to non-folate compounds. Thus, folate ultilization is increased and folate deficiency likely when cell turnover and DNA synthesis are increased.

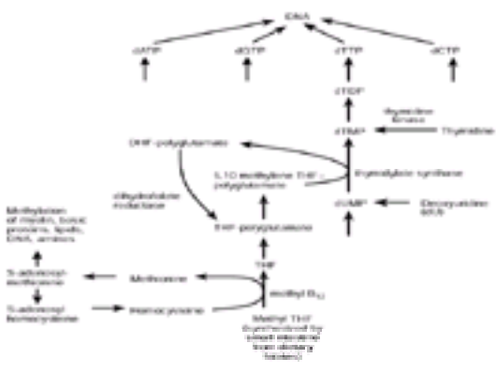

Fig. 4 Suggested mechanisms by which B_{12} deficiency affects folate metabolism and interferes with DNA synthesis. Indirect involvement of B_{12}, as methyl-B_{12}, in DNA synthesis is suggested by the 'methylfolate' trap ('tetrahydrofolate starvation') hypothesis. Methyl-B_{12} is involved in formation of intracellular THF from plasma methyl-THF. THF and/or its formyl derivative, but not methyl-THF, are the 'ground substances' from which all folate coenzymes are made by glutamate addition and single carbon unit transfer (see text). 5,10-Methylene-THF polyglutamate is involved in thymidylate synthesis. D, deoxyribose; A, adenine; G, guanine; T, thymine; C, cytosine; TP, triphosphate; DP, diphosphate; U, uridine; THF = tetrahydrofolate.

Nutrition

Folate occurs in most foods, the highest concentrations (more than 30 µg/100 g wet weight) being found in liver (where, like B_{12}, its is easily destroyed by cooking, particularly if large volumes of water and high temperatures are used); vitamin C protects it from oxidative destruction so reheating of food is particularly likely to reduce the folate content. Recent studies of a Western diet suggest an average normal daily intake of 250 µg, with 50 per cent or more in the polyglutamate form. Body stores are about 10 to 12 mg, with a mean liver concentration of about 7 µg/g. Primitive or rapidly growing tissues have higher folate concentrations than corresponding mature tissues. Normal adult requirements are about 100 µg daily, although estimates as low as 50 µg and as high as 200 µg have been made.

Absorption

Folates are absorbed rapidly, mainly through the duodenum and jejunum. Polyglutamates are deconjugated to the monoglutamate in the intestinal lumen, at the brush border, and possibly in lysosomes of intestinal cells by the enzyme, 'folate conjugase' (g-glutamylcarboxypeptidase, pteroylpolyglutamate hydrolase). They are then completely reduced to the tetrahydro state and methylated at the N_5 position so that methyl-THF enters portal plasma whatever food folate is ingested (Table 1). Folic acid itself, which is not present in food, but is used therapeutically, enters the portal blood largely unchanged at doses of more than 100 to 200 µg, as it is a poor substrate for reduction by dihydrofolate reductase. It does, however, share a specific folate uptake process through the intestine, as in the rare disorder, specific malabsorption of folate, there is failure of absorption of all folates including folic acid. The small intestine has a large capacity to absorb folate; on average 50 per cent of natural folate is absorbed whatever the dose. If excessive amounts are fed, the excess is largely excreted in urine as folates or their breakdown products after cleavage of the C_9–N_{10} bond. There is a substantial enterohepatic circulation for folate, estimated to contain up to 90 µg folate daily. If this is broken, plasma folate falls to about a third within 24 h.

Transport

Folate is transported in plasma, two-thirds unbound and about one-third loosely bound to albumin and possible other proteins. An active transport mechanism exists, however, for getting folates into cells. This is closely linked to the rate of folate polyglutamate synthesis and occurs by a carrier-mediated process, that is saturable, pH and energy dependent, with a greater preference for reduced than oxidized folates. In most cells, the folates then remain until the cells die but the liver can release folate from intact cells. Folate-binding proteins are present in human placenta, brush-border membranes of kidney tubular cells, and other cells. They are linked to membranes by a glycosyl-phosphatidyl anchor and play a part in cell uptake of folates and antifolates. Their level is regulated by extracellular and intracellular folate concentration. Milk protein, however, may enhance intestinal folate uptake. In plasma, the binding protein may take oxidized folates and breakdown products of folates to the liver for excretion or reconversion back to functional folates. Plasma folate is filtered by the glomerulus and mostly reabsorbed unless the renal tubular maximum is exceeded. Normal urine folate is 0 to 13 µg in 24 h. Folate is secreted into cerebrospinal fluid (which has a mean concentration of 24 µg/l) and is present in bile. Human milk has a concentration of 50 µg/l. Prostate-specific membrane antigen is a folate hydrolase carboxypeptidase which can release glutamates in either a or g linkages. The physiological significance of this is unknown.

Biochemical basis of megaloblastic anaemia

All known causes of megaloblastic anaemia, whether drugs, deficiencies, or inborn errors of metabolism, inhibit DNA synthesis by reducing the activity of one of the many enzymes concerned in purine or pyrimidine synthesis or by inhibiting DNA poymerization from its precursors. Folate deficiency, by reducing supply of the coenzyme 5,10-methylene-THF inhibits thymidylate synthesis, a rate-limiting reaction in DNA synthesis. B_{12} does not have a direct role in this or any other reaction in mammalian DNA synthesis. B_{12} deficiency inhibits DNA synthesis indirectly by its effect on folate metabolism.

Clinical, laboratory, and biochemical observations have all shown that B_{12} deficiency disturbs folate metabolism. Patients with severe B_{12} deficiency may show a haemaotological response to folic acid in large doses. Cell folate tends to be low, formiminoglutamic acid (FIGLU) and 5-amino-4-imidazole carboxamide excretion raised, and serine–glycine interconversion reduced in B_{12} deficiency, as in folate deficiency. The deoxyuridine blocking test suggests a defect in thymidylate synthesis in B_{12} deficiency that can be corrected *in vitro* by THF as well as by B_{12}. On the other hand, cell uptake of methyl-THF is reduced and serum folate raised in B_{12} deficiency. The most anaemic patients with B_{12} deficiency, as in folate deficiency, show the lowest levels of serum and red-cell folate, and the greatest disturbance of

folate biochemical reactions.

An explanation for these effects of B_{12} deficiency on folate metabolism is provided by the 'methylfolate trap' or tetrahydrofolate starvation hypothesis. This suggests that in B_{12} deficiency, folate is 'trapped' as methyl-THF, the form circulating in plasma, because of the need for methyl-B_{12} in the conversion of methyl-THF to THF. The 'trap' is supposed to lower the intracellular supply of THF, the most active of the folate compounds from which the other folate coenzymes are made. The natural folate coenzymes are the reduced derivatives of the folate polyglutamates rather than monoglutamates. Methyl-THF cannot act as a substrate for synthesis of these folate polyglutamates in human cells whereas THF (and formyl-THF) can.

The result is that B_{12} deficiency or inactivation puts a block between methyl-THF entering cells from plasma and the formation of intracellular folate polyglutamate coenzymes (Fig. 4). This causes the rise in plasma folate, a low level of intracellular folates, and reduced activity of all reactions requiring folate coenzymes, including those involved in DNA synthesis. The DNA defect has been ascribed to thymidine starvation. Thymidine corrects *in vitro* apoptosis of folate-deficient cells. Misincorporation of uracil, because of the pile up of deoxyuridine monophosphate and hence of deoxyuridine triphosphate, has been proposed to contribute to the cell abnormality.

Clinical features and causes of megaloblastic anaemia

Although pernicious anaemia is only one of the many causes of megaloblastic anaemia (Table 4, Table 5, and Table 6), it is convenient to describe the general clinical features of the anaemia under this heading because it is the most frequent cause of megaloblastic anaemia in Western countries. The laboratory findings and treatment of pernicious anaemia and other megaloblastic anaemias are discussed later.

Acquired pernicious anaemia (addisonian pernicious anaemia, Biermer's anaemia)

Definition

A disease of unknown origin in which there is atrophy of the stomach leading to severely reduced or absent IF secretion with consequent severe malabsorption of B_{12} and B_{12} deficiency.

Aetiology and associated diseases

Although a disease of the stomach, pernicious anaemia is considered with blood diseases because it usually presents with anaemia; it is indeed the most common cause of megaloblastic anaemia in many countries. It is a disease of older persons, less than 10 per cent of patients are under 40 years, with an incidence of 127/100 000 in Caucasians. There is a female:male ratio in most (but not all) series of about 1.6:1. There is a higher incidence (about 44 per cent compared to 40 per cent) of blood group A compared with controls in Britain. No overall association between pernicious anaemia and HLA type has been found, but those with an endocrine disease also have a greater incidence of HLA-B8, -B12, and -BW15. There are regional differences in incidence in the United Kingdom, with over 200 cases per 100 000 in Scotland but less than 60 per 100 000 in south-east England. It occurs in all races including African, Indian, Native American, and Chinese, as well as Caucasians. There is an association with early greying and blue eyes and a higher incidence in close relatives, of either sex, of an affected person. Family history is positive in about 30 per cent of cases. Those with a positive family history exhibit a younger mean age at presentation (55 years) than those without (66 years), but the type of inheritance is not clear.

Carcinoma of the stomach occurs in about 4 per cent of patients with pernicious anaemia, about three times the control rate. Pernicious anaemia may also be associated with other 'autoimmune' diseases, particularly primary myxoedema, thyrotoxicosis, Hashimoto's disease, Addison's disease, and vitiligo. About 55 per cent of patients show thyroid antibodies and 33 per cent with primary myxoedema have parietal-cell antibody. Close relatives also may show these diseases or their associated antibodies. There is probably no significant association with diabetes mellitus. Other evidence for an immune aetiology of the gastritis of pernicious anaemia is the improvement in mucosal appearance and function with corticosteroid therapy, the presence of antibodies in serum and gastric juice directed against parietal cells and IF, and of cell-mediated immunity to IF (see Chapter 5.2). Parietal-cell antibody is present in the serum of 85 to 90 per cent of patients. The autoantigens are the a- and b-subunit of the gastrin proton pump (H^+, K^+ATPase). Two antibodies to IF exist in serum. Type I ('blocking') occurs in about 50 per cent and is directed against the B_{12}-binding site. Type II (to the ileal binding site) occurs in 30 to 35 per cent but only if type I antibody is also present. Antibodies to IF exist in gastric juice and here they may neutralize the action of remaining IF. The incidence of parietal-cell and IF antibodies in serum in pernicious anaemia may be different in different groups of patients, younger patients having a lower incidence of parietal-cell antibody while Blacks and Hispanics may have a higher incidence of IF antibodies.

The antibodies to IF are virtually specific for pernicious anaemia but parietal-cell antibody occurs in many subjects with atrophic gastritis without pernicious anaemia. Cell-mediated immunity to IF can be demonstrated in all cases of pernicious anaemia. Lymphocyte populations in IF antibody-positive patients show a CD4 to CD8 ratio higher than in controls of those negative for IF antibody. Antibody to IF may cross the placenta and cause temporary deficiency of the factor in the newborn infant. An autoantibody to the gastrin receptor may also occur in serum in pernicious anaemia.

The anaemia may also be associated with hypgogammaglobulinaemia or with selective IgA deficiency when it tends to present at an early age, and antibodies to parietal cell and IF may then be absent. The gastric lesion then includes atrophy of the antrum in contrast to classical pernicious anaemia without hypogammaglobulinaemia. Serum gastrin concentrations are normal whereas these are raised above 200 µg/l in 90 per cent of patients with pernicious anaemia and serum pepsinogen (PG) concentrations are below 30 µg/l in 92 per cent of such patients with a low PGI/PGII ratio.

The relationship of pernicious anaemia to simple gastric atrophy, which occurs in about 15 per cent of people between 40 and 60 years and in 20 to 30 per cent of the older population, is not clear. In many cases, simple gastric atrophy does not progress to the anaemia after 10 or more years of follow-up. In a minority, however, deficiency of IF severe enough to cause malabsorption of B_{12} and megaloblastic anaemia, glossitis, or B_{12} neuropathy occurs. In these people, the development of antibody to IF in the gastric juice may be important.

Pathology

There is a gastritis in which all layers of the body and fundus of the stomach are atrophied with loss of normal gastric glands, mucosal architecture, and absence of parietal and chief cells, but mucous cells lining the gastric pits are well preserved. An infiltrate of plasma cells and lymphocytes with an excess of CD8 cells occurs and intestinal metaplasia may be present. The antral mucosa is remarkably well preserved except in hypogammaglobulinaemia, and, like the fundus, shows an increased number of gastrin-secreting cells.

Clinical features

The general features of megaloblastic anaemia are similar, whatever the underlying cause. Particular clinical features may point to the underlying disease, whether pernicious anaemia or some other cause. In pernicious anaemia, the anaemia usually develops gradually, perhaps over several years, and symptoms may not occur until it is severe. The most common complaints are due to the anaemia, while loss of mental and physical drive, numbness, or difficulty in walking suggest neuropathy. Psychiatric disturbances are common and range from mild neurosis to severe organic dementia. They may occur in the absence of anaemia or macrocytosis. Mild jaundice is frequent. Loss of appetite and weight, indigestion, and episodic diarrhoea are frequent. An intercurrent infection may precipitate severe anaemia and thus symptoms. Older patients may present with congestive heart failure. In a few patients, bruising due to thrombocytopenia is marked. On the other hand, many patients are diagnosed because a routine blood test is made.

The typical patient with pernicious anaemia has fair hair (prematurely grey), with blue eyes, and wide cheekbones. Physical signs, if present, are those of anaemia, perhaps with mild jaundice, giving the patient a so-called lemon-yellow tint. A few patients with either B_{12} or folate deficiency develop a widespread brown pigmentation, affecting nail beds and skin creases particularly, but not mucous membranes, which is reversible with the appropriate therapy. The biochemical basis for this is not clear, nor for the depigmentation that also occurs rarely. The tongue may be red, smooth, and shiny, occasionally with ulcers. A mild pyrexia up to 38°C is common in patients with moderate to severe anaemia. The liver may be enlarged while the cardiovascular system shows changes due to anaemia. Patients with pernicious anaemia may also have features of an associated disorder on presentation, most commonly myxoedema. Other thyroid disorders, vitiligo, carcinoma of the

stomach, Addison's disease, and hypoparathyroidism, may precede, occur simultaneously with, or follow the onset of the anaemia.

Vitamin B_{12} neuropathy

B_{12} deficiency may cause a symmetrical neuropathy affecting the lower limbs more than the upper (Section 24), which usually presents with paraesthesiae or with ataxia, particularly in the dark. In some cases, loss of cutaneous sensation, muscle weakness, urinary or faecal incontinence, an optic neuropathy, or psychiatric disturbance dominates. The neuropathy is due to severe B_{12} deficiency judged by serum B_{12} levels or methylmalonic acid excretion, but may occur with mild or no anaemia. It may be due to any cause of severe B_{12} deficiency, most commonly pernicious anaemia. A similar neuropathy has been described in dentists and others repeatedly exposed to nitrous oxide (N_2O) which inactivates methionine synthase. The biochemical explanation for the neuropathy is not clear. A defect in fatty acid metabolism in myelin tissue has been suggested. Studies in N_2O-treated monkeys have also suggested that the neuropathy results from accumulation of *S*-adenosyl homocysteine (caused by the block in conversion of homocysteine to methionine) with inhibition of transmethylation reactions in the brain. Methionine has been shown to prevent the neuropathy caused by N_2O in experimental animals. Defective methylation has yet to be shown in B_{12} neuropathy occurring clinically in man, however, or induced experimentally by dietary deficiency in fruit bats or by N_2O exposure of monkeys or rats. The role of B_{12} and folate in brain metabolism has been extensively reviewed (see Further reading list).

General tissue effects of B_{12} and folate deficiencies

Both deficiencies cause macrocytosis and other abnormal features of proliferating epithelial cells throughout the body (e.g. bronchial, bladder, buccal, and uterine cervix), with glossitis and angular cheilosis, a mild malabsorption syndrome, and reduced regeneration of damaged liver cells. In both sexes, sterility (reversible with B_{12} or folate therapy) may result from effects on the gonads. It is possible that the deficiencies in children affect overall body growth. Nutritional B_{12} deficiency in infants long term causes failure to thrive and poor brain growth with poor intellectual outcome. There does seem to be a real association of maternal folate deficiency with prematurity.

Generalized, reversible melanin pigmentation occurs in a few patients with B_{12} or folate deficiency, the cause of which is uncertain. Defective bactericidal activity of phagocytes due to impaired intracellular killing has been described in B_{12} deficiency but not in folate deficiency. B_{12} deficiency reduces serum levels of the osteoblast-related proteins alkaline phosphatase and osteocalcin but whether clinically important bone disease occurs is unknown.

Neural tube defects (NTD)

Studies by Hibbard and Smithells and coworkers in the 1960s suggested that NTD was associated with reduced maternal folate status and their early studies showed an apparent prevention of NTD by periconceptional vitamin supplementation. It is now established that folic acid supplements at the time of conception and in the early (first weeks) stage of pregnancy reduce the incidence of NTDs (anencephaly, encephalocele and spina bifida) in the first pregnancy and in subsequent pregnancies where such a malformation has occurred previously. The prophylactic daily dose used in a Medical Research Council (MRC) study was 4 mg and this reduced recurrence from 21 to six among 1195 randomized women studied; in a Hungarian first-pregnancy study, 0.8 mg daily preconception reduced the incidence from six to zero among 5000 women.

The explanation for the effect of folic acid on NTD is not certain. Women carrying NTD fetuses have lower serum folate and B_{12} levels and higher serum homocysteine levels than matched controls. There is a linear relationship when plotted on logarithmic scales between prevalence of NTD births and maternal red-cell folate, indicating that an increase in red-cell folate of a given amount is associated with a constant, proportional decrease in the birth prevalence of NTD from any given point on the red-cell folate distribution even within the accepted normal range. Folic acid prevention of NTD despite normal serum and red cell folate levels suggests that folic acid is overcoming a metabolic abnormality in folate metabolism. Only one such defect, a mutated tetrahydofolate reductase, has been identified so far.

Mutated 5,10 methylene tetrahydrofolate reductase (MHTFR)

A common thermolabile variant (677C ® T) (Ala 225 Val) of the enzyme MHTFR is associated with lower serum and red-cell folate levels and with higher plasma homocysteine levels than in controls in the general population. The incidence of the homozygous state in the population is approximately 5 per cent; the incidence in the parents and fetuses with NTD is approximately 13 per cent. The presence of this mutation can therefore account for only a small proportion of NTDs.

Cardiovascular disease

McCully (1969) first implicated homocysteine as a cause of atherosclerosis. This was based on pathological studies of children or young adults with congenital homocystinuria, whether due to a defect of cystathionine synthase, methionine synthase, or MHTFR (Fig. 5). In these children, plasma homocysteine levels are raised to 10 to 100 times normal. It is now apparent that even mild rises in plasma homocysteine are associated with coronary or peripheral arterial disease, with stroke and deep vein thrombosis. Determinants of plasma homocysteine include age, sex, renal function, protein intake, vitamin B_6, folate, and vitamin B_{12} status, the presence of the thermolabile variant MHTFR, smoking, and alcohol consumption, as well as intake of various drugs. Folate deficiency assessed by serum or red cell folate or by dietary folate intake is also associated with coronary vascular disease, myocardial infarct, or peripheral vascular disease. Meta-analysis, however, does not confirm an association of the presence of the thermolabile variant of MHTFR with coronary artery disease.

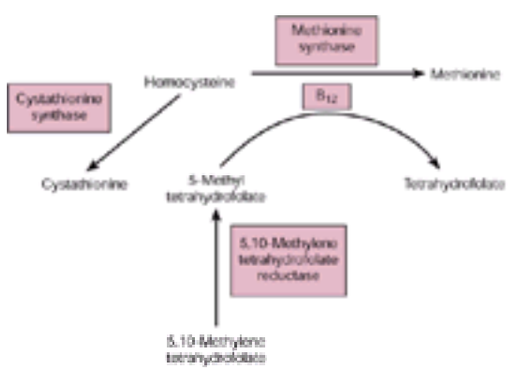

Fig. 5 The role of three enzymes (cysathione synthase, methionine synthase, and MHTFR) and three vitamins (B_{12}, B_6, and folate) in homocysteine metabolism.

It seems likely that homocysteine is a cause of vascular disease rather than an associated abnormality. Folic acid fortification of the diet reduces plasma homocysteine levels. It has been estimated that fortification of the British diet with folic acid will result in 10 000 fewer annual deaths from myocardial infarct. Trials of folic acid therapy aimed at reducing the incidence of myocardial infarct are now in progress. Preliminary results of electrocardiographic studies and of coronary restenosis after coronary angioplasty in patients with coronary arterial disease receiving folic acid are encouraging.

Other causes of vitamin B_{12} deficiency

Juvenile pernicious anaemia

A few cases of pernicious anaemia with gastric atrophy, achlorhydria, and IF antibodies have occurred in children. They may show associated ('autoimmune') conditions, for example myxoedema, hypoparathyroidism, Addison's disease, or chronic mucocutaneous candidiasis.

Congenital deficiency or structural abnormality of intrinsic factor

About 40 cases have been reported of a child being born with absent or non-functioning IF but an otherwise normal stomach on biopsy and normal secretion studies (e.g. of acid). Inheritance is autosomal recessive. In some cases, IF is present in the gastric juice that is susceptible to acid degradation, cannot bind B_{12}, or that binds B_{12} but cannot attach it to ileal receptors. These children tend to present with irritability, vomiting, diarrhoea, and loss of weight, and are found to have megaloblastic anaemia. The usual age of diagnosis is about 2 years, although a few have been diagnosed as early as 4 months and others only in their teens.

Total gastrectomy

All patients who have this operation will develop B_{12} deficiency, which usually presents between 2 and 6 years postoperatively. They should be treated with prophylactic B_{12} injections from the time of the operation.

Partial gastrectomy

Iron deficiency usually accounts for the anaemia that occurs in up to half of subjects after this operation. Subnormal serum B_{12} levels develop in about 18 per cent of patients from about 2 years postoperatively. About 6 per cent develop megaloblastic anaemia due to the deficiency. In most of these patients, malabsorption of B_{12} is due to an abnormal jejunal flora. The exact incidence of B_{12} deficiency depends mainly on the size of the remnant, which tends to be smaller if the operation is subtotal and the peptic ulcer gastric rather than duodenal. Vagotomy and pyloroplasty is not a cause of B_{12} deficiency.

Small-intestinal lesions

Colonization of the upper small intestine with colonic bacteria, if sufficiently heavy as in the stagnant-loop syndrome, leads to malabsorption of B_{12}. The most common causes are listed in Table 4. It appears that the bacteria destroy IF. Infestation with the fish tapeworm (*Diphyllobothrium latum*) has a similar effect but is now almost completely eradicated; infestation is only sufficiently marked in Finland and around the lakes of Russia to frequently cause megaloblastic anaemia.

Human immunodeficiency virus (HIV) infection

Serum B_{12} levels fall progressively in HIV-infected patients and subnormal serum levels occur in 10 to 35 per cent of patients with the acquired immune deficiency syndrome. Increased levels of TC II are usual and malabsorption of B_{12}, normally not corrected by IF, has been found in some of these patients. An abnormal small-intestinal flora is the most likely cause of B_{12} malabsorption. Megaloblastic anaemia due to B_{12} deficiency is, however, rare.

Resection of a metre or more of terminal ileum

This causes severe malabsorption of B_{12}. Other diseases that may affect ileal structure and function include: tropical sprue, in which severe B_{12} deficiency with anaemia or, rarely, neuropathy is a manifestation only in the chronic phase; gluten-induced enteropathy in which megaloblastic anaemia, if it occurs, is always due to folate deficiency (and B_{12} deficiency, if it occurs, is mild); in Crohn's disease malabsorption of B_{12} is frequent but severe B_{12} deficiency is unusual unless an ileal resection, fistula, or stagnant loop occurs.

Selective malabsorption of vitamin B_{12} with proteinuria (Imerslund's disease, Imerslund–Gräsbeck syndrome, recessive megaloblastic anaemia, MGA1)

This congenital disorder with autosomal recessive inheritance is the most common cause of megaloblastic anaemia due to B_{12} deficiency in non-vegan children. The child secretes IF normally but is unable to transport B_{12} across the ileum to portal blood. Most Finnish patients with MGA1 carry the disease specific mutation P1297L (FM1) in cublin. This mutation increases the K_d for IF–B_{12} binding several fold, so that is there is substantial loss of affinity of the FM1 mutant for B_{12}. A second less frequent mutation (FM2) activates a cryptic splice site with insertion of multiple stop codons in the CUB6 domain. The proteinuria, present in over 90 per cent of cases, is benign, non-specific, and persists after B_{12} therapy. The clinical presentation of the disease is identical to that of congenital IF deficiency. The disease may be heterogeneous, encompassing defects of the ileal IF receptor or subsequent steps in transfer of B_{12} to TC II.

Other causes of malabsorption of vitamin B_{12}

A number of other conditions and drugs may cause malabsorption of B_{12} but rarely cause B_{12} deficiency of clinical severity. *p*-Aminosalicylate, colchicine, neomycin, 'slow' potassium tablets, metformin, and phenformin have all been reported to cause malabsorption of B_{12}. In chronic pancreatitis and the Zollinger–Ellison syndrome, there is failure to release B_{12} from R protein due to absence of or inactivation of pancreatic trypsin. Malabsorption of B_{12} also occurs in inherited TC II deficiency.

Malabsorption of B_{12} occurs temporarily after total-body irradiation before bone marrow transplantation. If chronic graft-versus-host disease affecting the gut develops, malabsorption of B_{12} is usual, due to the abnormal gut flora as well as any ileal defect. Irradiation to the ileum during radiotherapy treatment for carcinoma of the cervix has also been reported to cause B_{12} malabsorption.

Dietary vitamin B_{12} deficiency

This occurs most commonly in Hindus who omit all animal produce from their diet. The incidence of overt megaloblastic anaemia is much lower than the incidence of subclinical deficiency assessed by the serum B_{12} assay. These individuals have low B_{12} stores. In India, babies have been born B_{12} deficient with megaloblastic anaemia caused by severe B_{12} deficiency (due to poor diet or sprue) in the mother. Dietary deficiency of B_{12} also occurs in non-Hindu vegans, and rarely in non-vegetarian people living on inadequate diets because of poverty.

Folate deficiency

Clinical features

The main clinical features of megaloblastic anaemia due to folate deficiency are similar to those when the anaemia is due to B_{12} deficiency, except that a severe neuropathy does not occur and the underlying aetiology tends to be different. Folate deficiency may develop rapidly, and although many mildly deficient patients do not progress for months or years, in some patients the deficiency may lead to a severe pancytopenia ('arrest of haemopoiesis') over a short period, particularly if an infection supervenes.

Folate deficiency is less well established as a cause of an organic neuropathy than B_{12} deficiency but in some studies a mild peripheral neuropathy has been found in as many as 20 per cent of patients, with spinal cord dysfunction less frequently. Neurological abnormalities do occur with inborn errors of folate metabolism and may be precipitated by antifolate drugs. The suggestion that folate therapy may precipitate fits in epileptics has not been confirmed by double-blind trials. Methyl-THF donates a methyl group to homocysteine to form methionine, which passes it on to *S*-adenosyl-methionine involved in methylation of biogenic amines (e.g. dopamine), proteins, phospholipids, and neurotransmitters in the brain. This may explain some of the psychiatric disturbances, such as depression, that seem to be equally common in folate as B_{12} deficiency.

Nutritional folate deficiency

Minor degrees of nutritional folate deficiency are frequent in most countries, but severe folate deficiency may account for about 17 per cent of all cases of megaloblastic anaemia in Britain. It occurs mainly in the old and poor and psychiatrically disturbed living alone on an inadequate diet from which liver, fruit, and fresh

vegetables are omitted; in many, barbiturate or alcohol consumption or a physical abnormality—partial gastrectomy, rheumatoid arthritis, or tuberculosis for example—may aggravate the effect of a poor diet. A few cases have developed because a special diet is taken, such as for pheonylkentonuria or for slimming. Scurvy is usually accompanied by severe folate deficiency while goat's milk anaemia is a nutritional folate deficiency due to the low (6 µg/l) folate content of goat's milk. In some countries, nutritional folate deficiency may be the main cause of megaloblastic anaemia, often presenting in pregnancy (e.g. in Burma, Malaysia, Africa, or India). Among Hindus, nutritional B_{12} deficiency is also common, however, and in many countries—Caribbean islands, Sri Lanka, and South-East Asia for example—tropical sprue (see Section 14.9) is an important cause of both deficiencies and is difficult to distinguish from 'pure' nutritional deficiency.

Malabsorption (see Section 14.9)

Gluten-induced enteropathy

Folate deficiency occurs in virtually all untreated patients, the serum folate being subnormal whether or not megaloblastic anaemia is present; red-cell folate is subnormal in 80 per cent or more. Anaemia occurs in about 90 per cent of adult cases, due to folate deficiency alone in 30 to 50 per cent and to mixed iron and folate deficiency in the remainder. Mild B_{12} deficiency may also occur but it is not a cause of anaemia in uncomplicated cases. Spontaneous atrophy of the spleen occurs in most of the patients; in about 10 to 15 per cent of cases, the blood film shows the presence of Howell–Jolly bodies, siderotic granules, and target and crenated cells that do not disappear with either folic acid or a gluten-free diet. Malabsorption of folic acid and of folate polyglutamates has been demonstrated in almost all untreated patients. A gluten-free diet produces a spontaneous rise in serum and red-cell folate and improved folate absorption in those patients who respond.

Malabsorption of folate also occurs in children with gluten-induced enteropathy and virtually all show subnormal serum and red-cell folate levels; anaemia is most often due to combined iron and folate deficiency but 'pure' megaloblastic anaemia also occurs.

Patients with dermatitis herpetiformis almost all show some degree of gluten-induced duodenal and jejunal abnormality; the severity of folate malabsorption and deficiency correlates with the severity of the intestinal lesion.

Tropical sprue (see Section 14.9)

Malabsorption of folate occurs in all severe, untreated patients in the acute phase and megaloblastic anaemia due to folate deficiency may develop within a few months. Not only does the anaemia respond to folate therapy but in many patients all the clinical features, and malabsorption of xylose, fat, B_{12}, and other substances, improve on folate therapy alone. Favourable responses to folic acid are most frequent in the first year of the disease, when about 60 per cent of patients appear to be cured by folic acid alone. Long-standing cases are more likely to be B_{12} deficient and thus to require B_{12} as well as folate and antibiotic therapy.

Congenital specific malabsorption of folate

This is a rare, autosomal recessive abnormality. Affected children all showed features of damage to the central nervous system (mental retardation, fits, athetotic movements) and present with megaloblastic anaemia responding to physiological doses of folic acid given parenterally but not orally. All forms of folate are poorly absorbed. Low levels of folate in cerebrospinal fluid also suggest a defect of folate transport through the choroid plexus.

Other causes

Absorption of folate is impaired by systemic infections. Mild degrees of folate malabsorption have also been reported after jejunal resection or partial gastrectomy, with Crohn's disease, and with lymphoma. In the intestinal stagnant-loop syndrome, folate levels tend to be high due to absorption of bacterially produced folate. Alcohol, anticonvulsants, oral contraceptives, antituberculous drugs, nitrofurantoin, sulphasalazine, bile-salt metabolites, and sodium bromosulphapthalein have been suggested, on variable evidence, to cause malabsorption of folate in some subjects but none is definitely established except sulphasalazine.

Increased folate utilization

A general mechanism of increased folate utilization in conditions of increased cell turnover has emerged. This consists of partial degradation of folate at the C_9–N_{10} band rather than complete recycling of the folate coenzymes required in DNA synthesis.

Pregnancy (see also Section 13)

This, associated with poor nutrition, is probably the most common cause of megaloblastic anaemia world-wide, unless folic acid supplements are taken. The frequency of the anaemia was about 0.5 per cent in most Western cities and up to 50 per cent in some areas of Asia and Africa until the introduction of prophylactic folic acid. The incidence increases with parity, is higher in twin pregnancies, and in some but not all series has been highest at the end of the winter. Folate requirements in a normal pregnancy are thought to be increased to about 300 to 400 µg daily, some 200 to 300 µg above normal. Serum and red-cell folate tend to fall as pregnancy progresses, and to rise spontaneously about 6 weeks after delivery. Lactation may prove an additional cause of folate deficiency, however, which may precipitate megaloblastic anaemia post partum.

The cause of the deficiency in pregnancy is increased degradation of folate due to hydrolysis at the C_9–N_{10} bond. Folate transfer to the fetus may play a minor part. Malabsorption of folate and increased urine folate excretion may be minor factors in some patients; in a few, megaloblastic anaemia of pregnancy is the first sign of adult coeliac disease. The statistical association of iron and folate deficiencies in pregnancy is probably due to a poor quality of the diet in certain women.

Prophylactic folic acid should now be given routinely in pregnancy; 400 µg daily is recommended (see earlier) and intake in women who may become pregnant should be at least this amount daily from food or supplements. Larger doses (4–5 mg daily) should be used if there has been a previous infant with a neural-tube defect. Conventional doses of 5 mg daily are satisfactory generally but have the theoretical drawback of being more likely to mask anaemia in the rare pregnant subject with untreated pernicious anaemia and thus might allow B_{12} neuropathy to develop.

Prematurity

Newborn infants have higher serum and red-cell folate concentrations than adults. These fall to a lowest value at about 6 weeks of age because utilization (and possibly excessive urinary loss) exceed intake. In premature infants, the fall in folate levels after birth is particularly steep and a number of such infants have developed megaloblastic anaemia, particularly if infections, feeding difficulties, or haemolytic disease with exchange transfusion have occurred. Prophylactic folic acid (e.g. 1 mg weekly for the first 3–4 weeks of life) may be given, particularly to those babies weighing less than 1.5 to 1.8 kg at birth.

Malignant diseases

Mild folate deficiency is frequent in patients with cancer (Table 5). In general, the severity correlates with the extent and degree of dissemination of the underlying disease. Patients with megaloblastic anaemia due to folate deficiency are unusual and folic acid might 'feed the tumour'; it should be withheld unless there is a real indication for its use, for example gross megaloblastosis causing severe anaemia, leucopenia, or thrombocytopenia.

Blood disorders

Chronic haemolytic anaemia

Requirements for folate are increased in patients with increased erythropoiesis, particularly when there is ineffective erythropoiesis with a high turnover of primitive cells. Occasional patients, presumably those with a poor folate intake, develop megaloblastic anaemia, particularly in sickle-cell anaemia, thalassaemia major, hereditary spherocytosis, and warm-type autoimmune haemolytic anaemia, Prophylactic folic acid is usually given in these disorders.

Chronic myelofibrosis

Megaloblastic haemopoiesis was reported in as many as one-third of patients in a series in London (England) with this disease but a lower incidence occurred in a large series in the United States. Circulating megaloblasts, increased transfusion requirements, severe thrombocytopenia, or pancytopenia may be the first indication that folate deficiency has developed. Polycythaemia vera is not a cause of folate deficiency.

Sideroblastic anaemia

Folate deficiency, usually mild, may occur in about half of acquired cases. Megaloblastosis, refractory to folate or B_{12}, also occurs in the acquired form as in other myelodysplastic diseases.

Inflammatory diseases

Folate deficiency has been described in patients with tuberculosis, malaria, Crohn's disease, psoriasis, widespread eczema, and rheumatoid arthritis. The degree of deficiency is related to the extent and severity of the underlying disorder. Increased demand for folate probably is a factor but reduced appetite is also important in those who develop megaloblastic anaemia.

Metabolic

Homocystinuria (see Section 11)

Patients with the most common form of this disorder, due to cystathionase deficiency, may show folate deficiency, possibly due to excess conversion of homocysteine to methionine and thus excess utilization of the folate coenzyme concerned.

Excess urinary loss of folate

Urine folate excretion of 100 µg a day or more occurs in some patients with congestive cardiac failure or active liver disease causing necrosis of liver cells. It is presumed that losses are due to release of folate from damaged liver cells. Haemodialysis and peritoneal dialysis remove folate from plasma. Folic acid (e.g. 5 mg weekly) is now usually given prophylactically to patients with renal failure who require long-term dialysis.

Drugs

Dihydrofolate reductase (DHFR) inhibitors

Methotrexate, aminopterin, pyrimethamine, and trimethoprim all inhibit DHFR but have different relative activities against the human, malarial, and bacterial enzymes. Methotrexate is converted to polyglutamate forms, which increases its activity against DHFR and also increases its retention in cells. They cause varying degrees of impairment of folate metabolism in man. Trimethoprim, used as an antibacterial agent, may aggravate pre-existing folate or B_{12} deficiency but does not of itself cause megaloblastic anaemia.

Alcohol

Folate deficiency may occur in spirit-drinking alcoholics. The main factor is poor nutrition and it is likely that alcohol interrupts the enterohepatic circulation for folate. It also has a direct effect on haemopoiesis, causing vacuolation of normoblasts, impaired iron utilization, sideroblastic changes, macrocytosis, megaloblastosis, and thrombocytopenia, even in the absence of folate deficiency. Beer drinkers seem relatively immune to folate deficiency because of the high folate content of beer. The usual macrocytosis in less severe, non-anaemic alcoholics is not related to folate deficiency.

Anticonvulsants, barbiturates

Diphenylhydantoin, primidone, and barbiturate therapy may be associated with some degree of folate deficiency. The more severe deficiency is associated with poor dietary intake of folate and usually prolonged drug therapy at high doses. The mechanism for the deficiency is undetermined. Malabsorption of folate, excess utilization due to induction of folate-requiring enzymes, displacement of folate from its binding protein, or competition for folate-requiring enzymes have all been suggested but not proven.

Other drugs

Nitrofurantoin, triamterene, proguanil, and pentamidine have been suggested to cause folate deficiency. Homofolates and carboxypeptidase G are two folate antagonists that have not been used in man.

Liver disease

Folate deficiency occurs most commonly in alcoholic cirrhosis where alcohol, poor nutrition, poor storage, and excess urine losses may all be important. The deficiency is less frequent in other types of liver disease.

Laboratory investigation of megaloblastic anaemia

This consists of three stages: (i) recognition that megaloblastic anaemia is present; (ii) distinction between B_{12} or folate deficiency (or rarely some other factor) as the cause of the anaemia; (iii) diagnosis of the underlying disease causing the deficiency (Table 7).

Recognition of megaloblastic anaemia

The peripheral blood

There is a raised mean corpuscle volume (MCV) to between 100 fl and 140 fl. Oval macrocytes are seen in the blood film. In mild cases, macrocytosis is present before anaemia has developed. Poikilocytosis and anisocytosis are marked in severe cases. Cabot rings (composed of arginine-rich histone and non-haemoglobin iron) and occasional Howell–Jolly bodies (DNA fragments) may occur due to extramedullary haemopoiesis in the liver and spleen. The MCV may be normal if there is associated iron deficiency, when the blood film appears dimorphic, or if the anaemia (usually due to folate deficiency or antimetabolite drug therapy) develops acutely over the course of a few weeks. The MCV is also normal in some severely anaemic cases involving excess red-cell fragmentation. The reticulocyte count is low for the degree of anaemia, usually of the order of 1 to 3 per cent.

The peripheral blood also shows hypersegmented neutrophils (which have nuclei with more than five lobes) (Plate 1) and the leucocyte count is often moderately reduced in both neutrophils and lymphocytes, although the total leucocyte count rarely falls to less than 1.5×10^{9}/l. The lymphocyte CD4/CD8 ratio is reduced. The platelet count may be moderately reduced but rarely falls below 40×10^{9}/l.

Biochemical changes

These are confined to the anaemic patient and include a slight rise in serum bilirubin (up to 50 µmol/l), mainly unconjugated, a rise in serum lactic dehydrogenase of up to 10 000 IU/l, with less marked rises in serum lysosyme and serum transaminases. The serum iron is also raised and falls within 12 to 24 h of effective treatment;

the serum ferritin is mildly raised and falls over the first few days of therapy. The serum cholesterol is low and alkaline phosphatase mildly reduced. Absence of haptoglobins is usual. In severe cases, free haemoglobin may be present in plasma, Schumm's test for methaemalbumin in serum is positive, and haemosiderin and fibrin degradation products are present in urine. The direct Coombs' test is weakly positive in some patients, due to complement.

Bone marrow

The bone marrow is hypercellular in moderate or severely anaemic cases and expanded along the lengths of the long bones. The myeloid–erythroid ratio is often reduced or reversed. The erythroblasts are larger than normal and show a number of morphological abnormalities; there is asynchronous maturation of nucleus and cytoplasm, nuclear chromatin remaining primitive with an open, lacy, fine granular pattern despite normal maturation and haemoglobinization of the cytoplasm. Fully haemoglobinized cells with incompletely condensed nuclei may be seen. Excessive numbers of dying cells, and nuclear remnants including Howell–Jolly bodies, mitoses, and multinucleate cells may be present. Because of death (by apoptosis) of later cells, there is a disproportionate accumulation of early cells. Giant and abnormally shaped metamyelocytes and megakaryocytes with hypersegmented nuclear lobes are also usually present (Plate 2). Studies with labelled thymidine have shown an increase of cells in G_2 and mitosis, and of cells with intermediate amounts of DNA between 2C and 4C but not synthesizing DNA, and presumably destined to die.

The severity of these changes tend to parallel the degree of anaemia. In milder cases, changes, described as 'intermediate', 'transitional', or 'moderate', are principally in the size and nuclear chromatin pattern of the individual developing erythroid cells, with giant metamyelocytes present; hypercellularity and gross dyserythropoiesis may be absent. In very mild cases, megaloblastic changes are difficult to recognize. In patients with severe anaemia but only mild megaloblastic changes, some additional cause for the anaemia should be sought.

Deoxyuridine suppression test

This is an *in vitro* biochemical test for B_{12} or folate deficiency based on the presence of a block in thymidylate synthesis (see Fig. 2). Deoxyuridine added to normoblastic cells reduces the incorporation of radioactive thymidine into DNA. The deoxyuridine is converted into deoxyuridine monophosphate (dUMP) and hence to mono-, di-, and tri-thymidine phosphates, which inhibit thymidine kinase and so reduce uptake of the labelled thymidine. Uptake of labelled thymidine into DNA is not blocked as much by deoxyuridine in cells from patients with B_{12} or folate deficiency as in normoblastic cells because of the block in conversion of dUMP to dTMP (thymidine monophosphate) in megaloblasts. Correction of the test *in vitro* with B_{12} or methyl-THF can be used to differentiate the two deficiencies as B_{12} will correct in B_{12} deficiency whereas methyl-THF does not; the reverse occurs in folate deficiency. The test is normal in cells from patients with megaloblastosis due to a block in DNA synthesis other than at thymidylate synthetase.

Chromosomes

Changes found in marrow and other proliferating cells include: (a) random chromatin breaks; (b) exaggeration of centromere constriction; and (c) thin, elongated, uncoiled chromosomes.

Ineffective haemopoiesis

The increased cellularity of the marrow with degenerate forms, and the low reticulocyte count, account for the degree of anaemia and suggest that many developing cells are dying in the marrow. This occurs by apoptosis, especially of late erythroblasts. Red-cell survival is moderately shortened, and radio-iron studies show rapid clearance, with increased plasma iron turnover but poor red-cell iron utilization. The raised uncongugated serum bilirubin, lactic dehydrogenase, and lysosyme are all due to ineffective haemopoiesis.

Differential diagnosis

Other causes of macrocytosis include a high reticulocytosis (e.g. haemolytic anaemia or regeneration of blood after haemorrhage), aplastic anaemia, red-cell aplasia, liver disease, alcoholism and myxoedema, the myelodyplastic syndromes, myeloid leukaemias, cytotoxic drug therapy, chronic respiratory failure, myelomatosis, and other causes of a leucoerythroblastic anaemia. Once a bone marrow biopsy has been done, the principal differentiation is from other causes of megaloblastosis, particularly myelodysplasia. Other causes of megaloblastic anaemia not due to B_{12} or folate deficiency are listed in Table 6.

Some patients with rapidly developing megaloblastic anaemia, particularly due to folate deficiency, may develop almost complete aplasia of the red-cell series, and the peripheral blood and bone marrow may resemble that of acute myeloid leukaemia.

Diagnosis of vitamin B_{12} or folate deficiency

The peripheral blood and bone marrow appearances are identical in folate or B_{12} deficiency. Special tests are, therefore, needed to distinguish between the two deficiencies. The deoxyuridine suppression test has been described already (see above), and is used for reliable and rapid diagnosis in some laboratories but it is not widely available.

Vitamin B_{12} deficiency

The assay of the B_{12} content of serum is now usually done by immunoassay. The normal ranges have been reported to be higher with the immunoassays (e.g. 200–1200 ng/l) than the previously used microbiological assays (e.g. 160–900 ng/l). Subnormal levels are found in cases of megaloblastic anaemia due to B_{12} deficiency, being extremely low in B_{12} neuropathy. Subnormal serum B_{12} concentrations in the absence of tissue B_{12} deficiency have been reported in pregnancy, in severe nutritional folate deficiency, in subjects taking large doses of vitamin C, and occasionally in iron deficiency.

Raised serum B_{12} levels, if not due to therapy or a contaminated serum, are most commonly caused by a raised B_{12}-binding capacity due to a rise in TC I as in a leucocytosis due to a myeloproliferative disease—chronic myeloid leukaemia, polycythaemia rubra vera, or in eosinophilic leukaemia for example. Raised levels of 'R' binder also occur in association with some tumours, especially hepatoma and fibrolamellar tumour of the liver. In benign leucocytosis, the rise is mainly of TC III and this is often not accompanied by a high serum B_{12}. Raised levels of TC II occur in conditions where macrophages are stimulated; for example autoimmune diseases such as systemic lupus erythematosus, rheumatoid arthritis, in Gaucher's disease and in some monocytic or monoblastic leukaemias, in histiocytic lymphomas, and inflammatory bowel disease. In active liver diseases, serum B_{12} leaks from the liver with saturation of the serum B_{12} binders.

A third and less widely used test for B_{12} deficiency is the measurement of the serum concentration of methylmalonic acid (MMA) or 24-h urine excretion of MMA. Serum MMA levels and excretion of MMA are raised in B_{12} deficiency but not in folate deficiency but raised levels may occur in renal failure. Rare cases of congenital methylmalonic aciduria have been described, owing to a variety of enzyme defects.

A sensitive method of measuring MMA in serum has been introduced and combined with serum homocysteine assay for the diagnosis of B_{12} or folate deficiency. Savage, Lindenbaum, Stabler, Allen, and others have used the serum MMA concentration to diagnose B_{12} deficiency in the absence of macrocytes or anaemia in patients with neuropathy and serum B_{12} concentration of more than 200 ng/l. Most find that patients with B_{12} neuropathy show haematological changes of B_{12} deficiency and there are not substantial numbers of patients with undiagnosed pernicious anaemia with normal haematological findings and borderline or even normal serum B_{12} levels.

Folate deficiency

Direct tests include the serum and red-cell folate assay. In most laboratories immunoassays are now used. The serum folate is always low in folate deficiency (and is normal or raised in B_{12} deficiency unless folate deficiency is also present). The serum folate does not accurately measure the severity of folate deficiency. Raised levels occur after folate therapy and also in B_{12} deficiency and in the stagnant-loop syndrome. Red-cell folate is a better guide than the serum folate to tissue folate

stores but is also low in a proportion of patients with megaloblastic anaemia solely due to B_{12} deficiency. Serum homocysteine levels are usually raised in folate deficiency, but also in B_{12} deficiency and many other situations.

Diagnosis of the cause of vitamin B_{12} deficiency

Although the clinical and family history and the clinical findings may point to pernicious anaemia or some other cause of B_{12} deficiency, it is important to establish this for certain. A brief dietary history will rapidly establish whether or not the patient is a vegan or takes a very inadequate diet. Radioactive B_{12} absorption tests are valuable to demonstrate malabsorption of B_{12} and to differentiate gastric from small-intestinal lesions as the cause. The patient, after an overnight fast, is fed an oral radioactively labelled dose of cyanocobalamin, usually 1 µg cobalt-57 B_{12}. Absorption can be measured by whole-body counting or by 24-h urinary excretion after a non-radioactive, parenteral flushing dose of 1 mg B_{12} (Schilling test). Hydroxo-cobalamin, instead of cyanocobalamin as originally described, can be used to flush absorbed, labelled B_{12} into urine.

Normal subjects absorb more than 30 per cent of the 1-µg dose. In patients with a gastric cause, malabsorption is corrected when the labelled B_{12} is given with IF, whereas if the lesion is small intestinal, the absorption does not improve with IF. Treatment with broad-spectrum antibiotics may improve the absorption in the stagnant-loop syndrome. In some patients with pernicious anaemia, the absorption with IF only improves substantially after weeks of B_{12} therapy, possibly due to slow recovery of ileal function from the effects of B_{12} deficiency. A combined test 'Dicopac' was available in which B_{12} labelled with cobalt-57 is given simultaneously with cobalt-58 B_{12} attached to IF. This has been withdrawn because human IF cannot be guaranteed to be virus or prion free. A double isotope test has also been developed in which cobalt-58 B_{12} is incorporated *in vitro* in egg yolk; cobalt-58 B_{12} is given in crystalline form. It is aimed to give a more accurate guide to food B_{12} absorption. Some patients with atrophic gastritis, or after partial gastrectomy and low serum B_{12} levels, may show normal absorption of crystalline B_{12} but reduced absorption of food B_{12}. Patients with pernicious anaemia show malabsorption of both forms.

Gastric secretion studies after pentagastrin stimulation in pernicious anaemia reveal achlorhydria (resting pH 7.0 and not falling by more than 1.0 unit on stimulation) and grossly reduced or absent IF in gastric juice.

Endoscopy and gastric biopsy will show features of gastric atrophy and help to exclude gastric carcinoma. Follow-through radiographic examination of the small intestine will help to exclude a small-intestinal lesion, duodenal or jejunal diverticulosis for example.

The serum gastrin level is raised in most patients with gastric atrophy and the serum is tested for antibodies to IF, parietal cells, and thyroid; serum immunoglobulins are measured in view of the association with hypogammaglobulinaemia.

Diagnosis of the cause of folate deficiency

An inadequate diet is usually at least partly implicated, but an exact estimate of dietary intake from the clinical history is impossible because of variation in folate content of foods, losses in cooking, and size of portions. Often it is the general social circumstances that suggest a poor intake. Drug intake, particularly of barbiturates, is important. Many underlying inflammatory or malignant diseases may exaggerate the tendency to folate deficiency in patients with inadequate diets. The main cause of malabsorption of folate is gluten-induced enteropathy; in patients with severe folate deficiency, tests for antiendomysial and antigliadin antibodies and a duodenal biopsy are usually necessary. In certain tropical countries, sprue may cause a generalized malabsorption syndrome in which folate deficiency commonly occurs. Tests of folate absorption have been devised, either by measuring the rise in serum folate after an oral dose of folic acid or of more natural folate derivatives (e.g. folate polyglutamates), or by measuring urinary or faecal excretion of radioactivity after feeding one or other labelled folate compound. None of these tests has achieved routine use.

Treatment of megaloblastic anaemia

Therapy is aimed at correcting the anaemia, completely replenishing the body of whichever vitamin is deficient, treatment of the underlying disorder, and prevention of relapse. In most cases, it is possible to diagnose which deficiency is present before starting therapy.

Vitamin B_{12} deficiency

Hydroxocobalamin 1000 µg intramuscularly given six times at several days' interval over the first few weeks will restore normal B_{12} stores. There is no evidence that patients with B_{12} neuropathy derive greater benefit from more frequent doses, although many physicians use these for 6 months or so.

Response to therapy

The patient feels better within 24 to 48 h, and the mild fever, if not due to infection, falls to normal. A painful tongue and unco-operative, disorientated state may also be improved in 48 h. The reticulocyte count begins to rise on the second day with a peak after 5 to 7 days. The white-cell count becomes normal by the third to seventh day and the platelet count rises and may reach levels of 500 to 1000 × 10^9/l before falling to normal at about 10 to 14 days. The bone marrow reverts to normoblastic by 36 to 48 h, although giant metamyelocytes persist for 10 to 12 days. The serum iron falls within 24 h, usually to subnormal levels, while the serum lactic dehydrogenase falls more slowly during the first 14 days of therapy.

The neuropathy always improves with therapy but residual deficits remain in some patients, particularly those with the longest histories and the most severe manifestations.

Maintenance

Hydroxocobalamin, 1000 µg intramuscularly, is given once every 3 months for life in pernicious anaemia and most other causes of B_{12} deficiency to prevent relapse. The life expectancy in pernicious anaemia once treated, is as good as that in the general population in women, and slightly lower in men, probably due to the increased incidence of carcinoma of the stomach. In a few patients with B_{12} deficiency, the underlying cause can be reversed; for example expulsion of the fish tapeworm, improvement of vegan diet, surgical correction of an intestinal stagnant loop. A few micrograms of B_{12} can be absorbed each day in pernicious anaemia from oral doses of 1000 µg or more by passive diffusion, but this maintenance therapy is usually reserved for those who cannot have injections—for example those with a bleeding disorder, or who refuse them—and for the extremely rare individual who is allergic to all injectable forms of B_{12}. Vegans may be maintained on much smaller oral doses of B_{12} each day, such as 50 µg as a tablet or syrup.

Prophylactic maintenance

B_{12} therapy should be given from the time of operation after total gastrectomy or after ileal resection if a B_{12} absorption test postoperatively reveals malabsorption of the vitamin. Patients with pernicious anaemia tend to develop iron-deficiency anaemia and they may also develop thyroid disorders or carcinoma of the stomach. It is advisable that a regular blood count be made once a year. Routine, regular endoscopy is not warranted but these diseases must be particularly borne in mind if relevant symptoms or signs develop.

Folate deficiency

This is corrected by giving 5 mg folic acid by mouth daily. It is essential to exclude B_{12} deficiency so that precipitation of a neuropathy is avoided. It is usual to continue for at least 4 months until there is a completely new set of red cells, although body stores will theoretically be normal within a few days of therapy. In patients with severe malabsorption of folate, larger oral doses of folic acid (e.g. 5 mg three times daily) may be used but it is not necessary to give parenteral folate except for those unable to swallow tablets. The response to therapy is as described for B_{12}. The decision whether or not to continue folic acid beyond 4 months depends on whether or not the cause can be corrected. In practice, long-term folic acid is usually needed only in patients with severe haemolytic anaemias (e.g. sickle-cell

anaemia and thalassaemia major), myelofibrosis, and in gluten-induced enteropathy when a gluten-free diet is either unsuccessful or not feasible. In patients on a gluten-free diet, assessment of folate status is one simple way of following the improvement in absorption.

Prophylactic folic acid

This should be given to all pregnant women (doses of 300 to 400 µg daily are used, often combined with an iron preparation) and, if the diet is poor, to all women likely to become pregnant. Larger doses are given if there has been a previous neural tube-deficit infant. Folic acid is given to patients undergoing regular haemodialysis or peritoneal dialysis, to premature infants weighing less than 1.5 kg at birth, and to selected patients in intensive care units or receiving parenteral nutrition.

Folate therapy has been shown to improve chromosomal stability in the fragile X syndrome, even though these patients do not have folate deficiency or a demonstrable defect of folate metabolism.

Food fortification

Fortification of cereals and grains with folic acid (140 µg/100 g cereal grain) began in the United States in 1996. Median serum folate in clinical specimens in United States rose from 12.6 to 18.7 µg/l between 1997 and 1998. There was also a reduction of plasma homocysteine in patients with coronary heart disease by consumption of a breakfast cereal fortified with folic acid. These results are consistent with those of the Framingham study which found a rise in serum folate and fall in serum homocysteine in the subjects taking a fortified diet from 1997 to 1998. In one study, addition of 400 µg folic acid daily over a 6-month period resulted in a rise in mean red-cell folate in young adult females from 295 to 571 µg/l, sufficient to reduce the incidence of NTD by 58 per cent. It is also hoped that there will be a significant reduction of deaths from cardiovascular disease by food fortification with folic acid. The theoretical side-effects of fortification are largely in patients with unsuspected B_{12} deficiency who theoretically might present with neuropathy if the extra folate consumed prevents the development of anaemia due to B_{12} deficiency. There is, however, no definite evidence for this at the supplemental doses given in the United States and proposed in Britain. In Britain fortification of flour with folic acid 240 µg per 100 g flour has been recommended but not yet implemented.

Folinic acid (5-formyl-THF)

This reduced folate is used to prevent or treat toxicity due to methotrexate or other dihydrofolate reductase inhibitors.

Severely ill patients

Some patients, usually elderly, are admitted to hospital severely ill with megaloblastic anaemia, perhaps in congestive heart failure or with pneumonia. In this case, it is necessary to commence therapy immediately after obtaining blood for B_{12} and folate assay and aspirating bone marrow, before it is known which deficiency is present. Both vitamins should be given simultaneously in large doses. Heart failure and infection should be treated in conventional fashion but blood transfusion should be avoided, except in cases of extreme anaemia, when 1 to 2 units of packed cells may be given slowly, accompanied by removal of a similar volume of blood from the other arm, and diuretic therapy.

Other therapy

Hypokalaemia may occur during the response to therapy and oral potassium supplements should be given to those with initial heart failure or if severe hypokalaemia is demonstrated, but are not needed routinely. An attack of gout has been reported on the sixth to seventh day of therapy. Most patients develop hyperuricaemia at this stage but the clinical disease probably only occurs in those with a strong gouty tendency. Iron deficiency commonly develops in the first few weeks of therapy and this should be treated initially with oral ferrous sulphate in the usual way.

Megaloblastic anaemia due to inborn errors of folate or vitamin B_{12} metabolism

Folate

A number of babies have been described with congenital deficiency of one or other enzyme concerned in folate metabolism: 5-methyltetrahydro-folate, methylene THF-reductase, FIGLU-transferase, methenyl-THF cyclohydrolase. Some of the babies had multiple congenital defects including the heart and cerebral ventricles and nearly all showed impaired mental development. In the methylfolate transferase deficiency, megaloblastic anaemia was present.

Vitamin B_{12}

Congenital deficiency of TC II was first reported in 1971 in two siblings who developed megaloblastic anaemia requiring therapy with large daily doses of B_{12} at 3 and 5 weeks of age. Similarly affected families have been described in which neuropathy developed in the absence of adequate therapy. A spectrum of loss of TC II occurs and functionally inactive TC II has been detected in some cases, often presenting later in life. The serum B_{12} level is usually normal, B_{12} being bound to TC I. Absorption of B_{12} is impaired. Treatment is with massive doses of B_{12} (e.g. 1000 µg intramuscularly three times each week). In contrast, in subjects with rare, inherited, low levels of TC I, low serum B_{12} levels occur, but haemopoiesis is normal.

Children with one form of congenital methylmalonic aciduria, which responds to B_{12} therapy in large doses, have been shown to have a defect in conversion of hydroxocobalamin to ado-B_{12}. They do not show megaloblastic anaemia. In a few, this defect has been associated with a defect of formation of methyl-B_{12} and with homocystinuria, but some of the children have also surprisingly not shown megaloblastic anaemia. Neurological abnormalities are usual. Homocystinuria and megaloblastic anaemia without methymalonic aciduria have also been reported. In some cases, the defect appears to be in maintaining B_{12} bound to methionine synthase in the reduced state.

Megaloblastic anaemia due to acquired disturbances of folate or vitamin B_{12} metabolism

Folate

Therapy with dihydrofolate reductase inhibitors may cause megaloblastic anaemia. This is usual with methotrexate and less likely with pyrimethamine unless high doses are used or the patient is already folate deficient. Trimethoprim and triamterene are very weak folate antagonists in man, but may precipitate megaloblastic anaemia in patients already B_{12} or folate deficient (see earlier).

Vitamin B_{12}

Nitrous oxide (N_2O)

This anaesthetic gas oxidizes B_{12} from the active fully reduce cob(I)alamin form to the inactive cob(II)alamin and cob(III)alamin forms, inactivating methyl-B_{12} and hence methionine synthase. Megaloblastosis develops within several hours in man and a fault in thymidylate synthase can be demonstrated by the deoxyuridine suppression test in human marrow exposed to N_2O. This recovers over several days when exposure to N_2O is discontinued. After many weeks exposure to N_2O, monkeys develop a neuropathy resembling B_{12} neuropathy in man; peripheral neuropathies have also been described in humans (e.g. dentists and anaesthetists) repeatedly exposed to the gas. When N_2O is used as anaesthetic for patients with low B_{12} stores, megaloblastic anaemia or neuropathy may be precipitated months later, due to failure to replenish B_{12} stores by absorption. Recovery from N_2O exposure needs new cobalamin and also synthesis of new apoenzyme (methionine synthase) because this protein is also damaged by active oxygen derived from the N_2O–cobalamin reaction. Methylmalonic aciduria has not been found in animals or

humans exposed for short periods to N_2O, as methylmalonic CoA mutase does not need reduced B_{12}.

Megaloblastic anaemia not due to folate or vitamin B_{12} deficiency or metabolic defect

Congenital

Orotic aciduria

This is a rare, recessive disorder involving two consecutive enzymes (orotidylic pyrophosphatase and orotidylic decarboxylase) in pyrimidine synthesis and presents with megaloblastic anaemia in the first few months of life. The diagnosis is made if needle-shaped, colourless crystals of orotic acid are found in the urine, daily excretion ranging from 0.5 to 1.5 g. Heterozygotes excrete slightly raised amounts of orotic acid but show no haematological disorder. Treatment with uridine (1–1.5 g daily) leads to a haematological response, restoration of normal haemopoiesis and growth, and reduction in orotic acid excretion.

Lesch–Nyhan syndrome

A few patients with this rare disorder of purine synthesis have shown megaloblastic change but whether this was due to associated folate deficiency or a direct result of reduced purine synthesis is not certain (see Section 11).

Vitamin E deficiency

This has been reported to cause megaloblastosis in a group of children with kwashiorkor. However, many were also folate deficient.

Vitamin C deficiency

Megaloblastic appears to be due to associated folate deficiency.

Thiamine responsive

About 12 cases have been well documented. They have also shown sideroblastic change and a defect in phosphorylation of thiamines has been implicated. Diabetes mellitus and semineural deafness are additional features.

Responding to large doses of vitamin B_{12} and folate

A single patient has been reported who needed both vitamins in large doses but the site of the defect was not elucidated.

Congenital dyserythropoietic anaemia

Some cases of congenital dyserythropoietic anaemia show megaloblastic changes not due to B_{12} or folate deficiency.

Acquired

Megaloblastic changes are often marked in acute myeloid leukaemia/M6 and less commonly in other forms of acute myeloid leukaemia. They also occur in about 50 per cent of patients with primary acquired sideroblastic anaemia and in other myelodysplastic syndromes. The exact site of block in DNA synthesis in these syndromes is unknown.

Drugs that directly inhibit purine or pyrimidine synthesis (e.g. cytosine arabinoside, 5-fluorouracil, hydroxyurea, 6-mercaptopurine, or azathioprine) may cause megaloblastic anaemia. Alcohol has also been found to have a direct effect on the bone marrow, causing megaloblastosis in some cases even in the absence of B_{12} or folate deficiency. On the other hand, drugs that inhibit mitosis (e.g. colchicine or daunorubicin) or alkylate preformed DNA (e.g. cyclophosphamide, chlorambucil, or busulfan) do not cause megaloblastosis.

Other deficiency anaemias

Vitamin C

Anaemia is usual in scurvy but the pathogenesis is complicated. It is likely that vitamin C has a direct effect on erythropoiesis but folate and iron deficiencies, haemorrhage, or haemolysis often complicate the picture.

Biochemical and nutritional aspects

Vitamin C is needed for collagen synthesis by its involvement in the hydroxylation of protein and for maintenance of intercellular substance of skin, cartilage, periosteum, and bone. It may also have a general role in oxidative–reduction systems, for example glutathione, cytochromes, pyridine, and flavin nucleotides. Although vitamin C is also thought to be needed for maintaining body folates in the reduced active state, the exact reactions involved are unclear. Vitamin C has a particular role in iron metabolism, iron excess causing increased utilization of vitamin C and in extreme cases clinical scurvy, whereas iron deficiency is associated with a raised leucocyte ascorbate concentration. Vitamin C is needed for incorporation of iron from transferrin into ferritin and for iron mobilization from ferritin. Vitamin C therapy increases iron excretion in patients receiving subcutaneous desferrioxamine infusions and also, at least in experimental animals, affects iron distribution by increasing parenchymal relative to reticuloendothelial iron. Minimum adult daily requirements for vitamin C are about 10 mg but 30 to 70 mg is recommended; utilization, and therefore requirement, are relatively higher in infants, children, and pregnant and lactating women. Vitamin C may be excreted as such but is also broken down to oxalate.

Vitamin C is present in food as its reduced (ascorbic acid) and oxidized (dehydroascorbic acid) forms, the highest concentrations occurring in greens, fruits, tomatoes, liver, and kidney. Potatoes are not a rich source but provide a substantial proportion of normal dietary intake. Cooking, particularly in alkaline conditions with large volumes of water, destroys the vitamin, which is also lost on storage with exposure to the air. Absorption occurs through the length of the small intestine and deficiency is never solely due to malabsorption.

The anaemia of scurvy is typically normochromic, normocytic with a slightly raised reticulocyte count to 5 to 10 per cent and a normoblastic marrow with erythroid hyperplasia. This suggests a direct role for vitamin C in erythropoiesis but not all patients with clinical scurvy are anaemic. Extravascular haemolysis with mild jaundice and increased urobilinogen excretion occurs in many of the patients. Moreover, in many the anaemia is complicated by folate deficiency (due to inadequate folate intake) with a megaloblastic marrow, or in a few by iron deficiency due to external haemorrhage, reduced diet intake, and possibly reduced iron absorption. In a few patients placed on a low folate diet, response of megaloblastic haemopoiesis to vitamin C alone has been described. In others, response of the megaloblastic anaemia to folic acid alone on a low vitamin C diet has occurred but in most such cases, both vitamin C and folic acid have been found necessary.

Vitamin B_6

This, as its coenzyme form pyridoxal-5-phosphate, is involved in many reactions of the body, especially transaminases and decarboxylases. It is also a cofactor in the important rate-limiting reaction in haem synthesis, †-aminolaevulinic acid (ALA)-synthetase (see Section 11). It occurs in natural tissues in three major forms: pyridoxine, pyridoxamine, and pyridoxal phosphate. Red cells are capable of interconverting them. Anaemia due purely to vitamin B_6 deficiency has been produced in animals. It is hypochromic and microcytic with a raised serum iron and increased iron in erythroblasts, with some partial or complete ring sideroblasts. A similar anaemia has occurred in humans with malabsorption, pregnancy, or haemolysis but has not been fully documented to respond to physiological doses of vitamin B_6

alone. Vitamin B_6-responsive anaemia is, however, well documented among patients with sideroblastic anaemia of all types. Pyridoxine responses occur particularly in the inherited form (when it is assumed that a fault in one or other enzyme of haem synthesis, for example ALA-synthetase, increases the need for pyridoxal phosphate as cofactor) and when sideroblastic anaemia occurs in patients receiving pyridoxine antagonists, such as antituberculous drugs. The value of pyridoxine dietary supplements in lowering serum homocysteine and reducing the incidence of cardiovascular disease has yet to be explored.

Riboflavin

On the basis of studies in experimental animals and humans fed a deficient diet together with a riboflavin antagonist, deficiency of this vitamin is known to cause a normochromic, normocytic anaemia associated with a low reticulocyte count and red-cell aplasia in the marrow, sometimes with vacuolated normoblasts. The exact biochemical basis is undecided. Clinically, a similar anaemia may occur in pure form but is usually associated with the anaemia due to protein deficiency, as in kwashiorkor or marasmus. Other clinical features of riboflavin deficiency—dermatitis, angular cheilosis, and glossitis for example—may be present.

Thiamine

For discussion, see under megaloblastic anaemia not due to folate or B_{12} deficiency or metabolic defect.

Nicotinic acid, pantothenic acid, and niacin

Deficiencies of these vitamins cause anaemia in experimental animals but anaemia purely due to one or other of these deficiencies has not been established to occur in man.

Vitamin E

This vitamin is needed for preventing peroxidation of cell membranes. A haemolytic anaemia responding to vitamin E has been reported in premature infants. Less well documented is a macrocytic anaemia due to vitamin E deficiency in protein-calorie-deficient infants and aggravation of anaemia in patients with thalassaemia major because of vitamin E deficiency.

Protein deficiency (see Section 10)

Anaemia is usual in both 'pure' protein deficiency, kwashiorkor, and in protein-calorie malnutrition (marasmus). It has been reported in many parts of the world where malnutrition, especially in children and pregnant women, is common. The anaemia also occurs in patients with gastrointestinal disease and severe malabsorption. The anaemia is typically normochromic, normocytic, and of the order of 8.0 to 9.0 g/dl. The reticulocyte count is usually reduced and the marrow may show a selective reduction in erythropoiesis. Experimental studies in animals suggest that the anaemia is largely due to reduced serum erythropoietin levels consequent on a lack of stimulus for erythropoietin secretion. Lack of amino acids for synthesis of erythropoietin or globin is not the cause. In many patients, the anaemia is complicated by infection, folate or iron deficiency, and possibly other vitamin deficiencies (e.g. riboflavin, vitamin E) and then it may be more severe and show additional morphological abnormalities in the blood and marrow.

Further reading

General

Chanarin I (1989). *The megaloblastic anaemias*, 3rd edn. Blackwell Science, Oxford. [The major textbook dealing with all aspects.]

Green R (1995). Metabolite assays in cobalamin and folate deficiency. *Clinical Haematology* **8**, 533–66. [Measurements of MMA and homocysteine in plasma are discussed as diagnostic tests for cobalamin and folate deficiencies.]

Green R, Miller JW (1999). Folate deficiency beyond megaloblastic anemia: hyperhomocysteimia and other manifestations of dysfunctional folate status. *Seminars in Hematology* **36**, 477–64. [An excellent review of the non-haematological aspects of folate deficiency.]

Hoffbrand AV, ed. (1976). Megaloblastic anaemia. *Clinical Haematology* **5**, 471–69. [A collection of 12 major reviews dealing with vitamin B_{12} and folate.]

Rosenblatt DS, Hoffbrand AV (1999). Megaloblastic anaemia and disorders of cobalamin and folate metabolism. In: Lilleyman J, Hann I, Blanchette V, eds. *Pediatric hematology*, pp.167–84. Churchill Livingstone, London. [A recent review of inborn errors of B_{12} and folate.]

Savage DG, Lindenbaum J, Stabler SP, Allen RH (1994). Similarities of serum methylmalonic acid and total homocysteine determinations for diagnosing cobalamin and folate deficiencies. *American Journal of Medicine* **96**, 239–46. [Reviews the value of these assays for diagnosing the deficiencies even in patients with normal serum levels of B_{12} and folate.]

Wickramasinghe SN, ed. (1995). Megaloblastic anaemia. Baillieres *Clinical Haematology* **8**, 441–703. A volume containing 12 major articles reviewing different aspects of vitamin B_{12} and folate. [This volume also contains reviews of different aspects of vitamin B_{12} and folate.]

Wickramasinghe SN (1999). The wide spectrum and unresolved issues of megaloblastic anemia. *Seminars in Hematology* **36**, 3–18. [An excellent general update on megaloblastic anaemia.]

Vitamin B_{12}

Carmel R (1995). Malabsorption of food cobalamin. *Clinical Haematology* **8**, 639–56. [This review brings together a large literature.]

Chanarin I, Metz J (1997). Diagnosis of cobalamin deficiency: the old and the new. *British Journal of Haematology* **97**, 695–700. [Discusses whether vitamin assays or measurement of serum methylmalonic acid or homocysteine should be used to diagnose the deficiencies.]

Fish DT, Dawson DW (1983). Comparison of methods used in commercial kits for the assay of serum vitamin B_{12}. *Clinical and Laboratory Haematology* **5**, 272–7. [A useful review of the different methods of assaying serum B_{12}.]

Gleeson PA, Toh BH (1991). Molecular targets in pernicious anaemia. *Immunology Today* **12**, 233–8. [Details the gastric antigens as targets for parietal cell antibody.]

Hewitt JE, Gordon MM, Taggart RT, *et al*. (1991). Human gastric intrinsic factor: characterization of cDNA and genomic changes and localization to human chromosome II. *Genomics* **10**, 432–40. [A major study of genetic aspects of intrinsic factor.]

Kondo H, Kolhouse JF, Allen RH (1980). Presence of cobalamin analogues in animal tissues. *Proceedings of the National Academy of Sciences*, *USA* **77**, 817–21. [Describes the nature and origin of cobalamin analogues.]

Kozyraki R, Kristiansen M, Silahtaroglu A, *et al*. (1998). The human intrinsic factor—vitamin B_{12} receptor, cubilin: molecular characterization and chromosomal mapping of the gene to 10p within the autosomal recessive megaloblastic anemia (MGA1) region. *Blood* **91**, 3593–600. [The identification of cubilin as the IF.B12 receptor.]

Kristiansen M, Aminoff M, Jacobsen C, *et al*. (2000). Cubulin P1297L mutation associated with hereditary megaloblastic anemia 1 causes improved recognition of intrinsic factor—vitamin B12 by cubilin. *Blood* **96**, 405–9. [Identification of the most frequent mutation underlying MGA1.]

Regec A, Quadros EV, Plalica O, *et al*. (1995). The cloning and characterization of the human transcobalamin II gene. *Blood* **85**, 2711–19. [Important report of genetic aspects of TCII.]

Remacha AF, Riera A, Cadafalch J, Grimferrer E (1991). Vitamin B_{12} abnormalities in HIV-infected patients. *European Journal of Haematology* **47**, 60–4. [Describes the incidence of malabsorption of B_{12} and of B_{12} deficiency in HIV-infected patients.]

Rothenberg SP, Quadros EV (1995). Transcobalamin II and the membrane receptor for the transcobalamin II-cobalamin complex. *Clinical Haematology* **8**, 499–514. [A major review of the structure of TCII, the other transcobalamins, and intrinsic factor.]

Savage DG, Lindenbaum J (1995). Neurological complications of acquired cobalamin deficiency: clincal aspects. *Clinical Haematology* **8**, 657–78. [This review deals with the effects of folic acid at different doses on haematological and neurological aspects of cobalamin deficiency.]

Weir DG, Scott JM (1997). Brain function in the elderly: role of vitamin B_{12} and folate. *British Medical Bulletin* **55**, 669–82. [A large review of this important topic.]

Folate

Antony AC (1992). The biological chemistry of folate receptors. *Blood* **79**, 2807–20. [All aspects of folate receptors are discussed.]

Bailey L, ed. (1994). *Folate in health and disease*. Marcel Decker, New York. A collection of articles about all aspects of the vitamin.

Clarke R, Smith AD, Jobst KA, *et al*. (1998). Folate, vitamin B_{12} and serum total homocysteine levels in confirmed Alzheimer's disease. *Archives of Neurology* **55**, 1449–55. [An important study suggesting folate deficiency may predispose to Alzheimer's disease.][

Giovannucci E, Stampfer MJ, Colditz GA, *et al*. (1998). Multivitamin use, folate and colon cancer in women in the nurses' health study. *Annals of Internal Medicine* **129**, 517–24. [A large study implying but not proving that folate deficiency predisposes to colon cancer.]

Heston WD (1997). Characterization and glutamyl carboxypeptidase functions of prostate-specific membrane antigen: a novel folate hydolase. *Urology* **49** (Suppl. 3A), 104–12. [Demonstration that prostate-specific antigen is a folate hydrolase.]

Hoffbrand AV, Weir DG (2001). The history of folic acid. *British Journal of Haematology* **113**, 579–89. [A comprehensive review of all aspects of folate since the original discovery of the vitamin by Lucy Wills.]

Selhub J, Jacques PF, Wilson PWF, *et al*. (1993). Vitamin status and intake as primary determinants of homocysteinaemia in an elderly population. *Journal of the American Medical Association* **270**, 2693–8.[A study showing the importance of folate status in determining plasma homocysteine levels.]

Shane B (1989). Folylpolyglutamate synthesis and role in regulation of one-carbon metabolism. *Vitamins and Hormones* **45**, 263–335. [An important review of folate metabolism with emphasis on folate polyglutamates.]

Neural tube defect

Botto L, Moore CA, Khoury MJ, *et al*. (1999). Neural-tube defects. *New England Journal of Medicine* **341**, 1509–18. [A major review of all aspects of neural tube defects.]

Czeizel AE, Dundas I (1992). Prevention of the first occurrence of neural tube defects by periconceptional vitamin supplementation. *New England Journal of Medicine* **327**, 1832–5. [The first demonstration of prevention of first occurrence of NTD by folic acid.]

Daley LE, Kirke PN, Molloy A, *et al*. (1995). Folate levels and neural tube defects: implications for prevention. *Journal of the American Medical Association* **274**, 1698–702. [Demonstration of close relation between incidence of NTD and red cell folate levels in the Irish population.]

Daley S, Mills JL, Molloy AM, *et al*. (1997). Minimum effective dose of folic acid for food fortification to prevent neural tube defects. *Lancet* **350**, 1666–9. [Study of the effects of different supplemental doses of folic acid daily over a 6-month period on red cell folate levels.]

Hibbard EM, Smithells RS (1965). Folic acid and human embryopathy. *Lancet* **i**, 1254. [The first suggestion that folate deficiency may predispose NTD.]

Lawrence JM, Petitti DB, Watkins M, *et al*. (1999). Trends in serum folate after food fortification. *Lancet* **354**, 915–16. [Analysis of folate levels in the United States population after food fortification with folic acid.]

Molloy A, Sean D, Mills JL, *et al*. (1997). Thermolabile variant of 5,10-methylene tetrahydrofolate reductase associated with low red cell folates: implications for folate intake recommendations. *Lancet* **349**, 1591–3. [Identification of low red cell folate in normal subjects with mutated MHTFR.]

MRC Vitamin Study Group (1991). Prevention of neural tube defects: results of Medical Research Council Vitamin Study. *Lancet* **238**, 131–7. [The first study to establish that folic acid therapy periconception substantially reduces the incidence of recurrence of NTD births.]

Smithells RW, Shephard S, Schorah CJ, *et al*. (1980). Possible prevention of neural-tube defects by periconceptional vitamin supplementation. *Lancet* **i**, 339–40. [The first data to suggest that folic acid supplements may prevent NTD.]

Cardiovascular system

Boushey CJ, Beresford SAA, Omenn GS, *et al*. (1995). A quantitative assessment of plasma homocysteine as a risk factor for vascular disease: probable benefits of increasing folic acid intakes. *Journal of the American Medical Association* **274**, 1049–57. [An important study predicting the benefits on incidence of cardiovascular disease of food fortification with folic acid.]

Brattsom L, Wilcken DE, Ohrvik J, *et al*. (1998). Common methylenetetrahydrofolate reductase gene mutation leads to hyperhomocysteinaemia but not to vascular disease: the result of a meta-analysis. *Circulation* **98**, 2520–6. [Meta-analysis showing mutated MHTFR is not a risk factor for vascular disease.]

Christen WG, Ajoni UA, Glynn RJ, Hennekens CH (2000). Blood levels of homocysteine and increased risks of cardiovascular disease. Causal or casual? *Archives of Internal Medicine* **160**, 422–34. [A discussion of whether or not raised homocysteine levels cause vascular disease.]

Den Heijer M, Koster T, Blom HJ, *et al*. (1996). Hyperhomocysteinemia as a risk factor for deep vein thrombosis. *New England Journal of Medicine* **334**, 759–62. [The only published study showing a raised plasma homocysteine is associated with venous thrombosis.]

Frosst P, Blom HJ, Milos R, *et al*. (1995). A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nature Genetics* **10**, 111–13. [The demonstration of the gene defect underlying thermolabile MHTFR.]

Graham IM, Daly LE, Refsum HNM, *et al*. (1997). Plasma homocysteine as a risk factor for vascular disease: the European Concerted Action Project. *Journal of the American Medical Association* **277**, 1775–81. [A meta-analysis of the association of homocysteine and vascular disease.]

Haynes WG (2000). Homocysteine and artherosclerosis: potential mechanisms and clinical implications. *Proceedings of Royal College of Physicians of Edinburgh* **30**, 114–22. [A useful large review of this topic.]

Homocysteine Lowering Trialists Collaboration (1998). Lowering blood homocysteine with folic acid based supplements: meta-analysis of randomised trials. *British Medical Journal* **316**, 894–8. [An important meta-analysis of trials aimed at lowering plasma homocysteine with folic acid supplements.]

Jacques PF, Selhub J, Borton AG, *et al*. (1999). The effect of folic acid fortification on plasma folate and total homocysteine concentrations. *New England Journal of Medicine* **340**, 1449–53. [The Framlingham study of the effect of adding folic acid to the diet on plasma homocysteine and folate levels.]

McCully KS (1969). Vascular pathology of homocysteinemia. *American Journal of Pathology* **56**, 111–28. [The first suggestion that homocysteine leads to vascular disease.]

Malinow MR, Duall PB, Hess DL, *et al*. (1998). Reduction of plasma homocyst(e)ine levels by breakfast cereal fortified with folic acid in patients with coronary heart disease. *New England Journal of Medicine* **338**, 1009–15. [Study showing that food fortification with folic acid reduces plasma homocysteine levels.]

Morrison HI, Schaubel D, Desmeules M, *et al*. (1996). Serum folate and risk of fatal coronary heart disease. *Journal of American Medical Association* **275**, 1893–6. [A large Canadian study relating folate status to incidence of myocardial infarct.]

Perry DJ (1999). Hyperhomocysteinaemia. *Clinical Haematology* **12**, 451–78. [A major, authorative review of the causes and associations of a raised plasma homocysteine level.]

Rimm EB, Willett WC, Hu FB, *et al*. (1998). Folate and vitamin B_6 from diet and supplements in relation to risk of coronary heart disease among women. *Journal of the American Medical Association* **279**, 359–64. [A retrospective study of the effects of folic acid supplements on the risk of coronary artery disease.]

Robinson K, Arheart K, Refsum H, *et al*. (1998). Low circulating folate and vitamin B_6 concentrations: risk factors for stroke, peripheral vascular disease and coronary artery disease. *Circulation* **97**, 437–43.[A major study relating folate levels to stroke and peripheral vascular disease as well as to coronary artery disease.]

Schnyder MD, Roffi M, Pin R, *et al.* (2001). Decreased rate of coronary restenosis after lowering of plasma homocysteine levels. *New England Journal of Medicine* **345**, 1593–1600. [Demonstrated that a combination of folic acid, vitamin B_{12}, and pyridoxine significantly decrease the rate of restenosis and need for revascularization after coronary angioplasty. There was also a reduction of major cardiac events.]

Verhoef P, Stampfer MJ, Bursing E, *et al*. (1996). Homocysteine metabolism and risk of myocardial infarction: relation with vitamin B_6, B_{12} and folate. *American Journal of Epidemiology* **143**, 845–59. [A major study relating plasma homocysteine to vitamin status and myocardial infarct.]

Vermeulen EGJ, Stenhauwer CDA, Twisk JWR, *et al*. (2000). Effect of homocysteine-lowering treatment with folic acid plus vitamin B_6 on progress of subclinical atherosclerosis: a randomised,

placebo controlled trial. *Lancet* **355**, 517–22. [The first study to show a benefit, albeit on electrocardiographic abnormalities, of dietary supplementation with folic acid and vitamin B_6, in patients with atherosclerosis.]

Wald NJ, Watt HC, Law MR, *et al*. (1998). Homocysteine and ischemic heart disease: results of a prospective study with implications regarding prevention. *Archives of Internal Medicine* **158**, 862–7. [A valuable, prospective study quantifying the increased risk of ischaemic heart disease with incremental rises in plasma homocysteine.]

Welch GN, Loscalzo J (1998). Homocysteine and artherothrombosis. *New England Journal of Medicine* **338**, 1042–50. [A useful review of the association of homocysteine and atherosclerosis.]

Miscellaneous

Adams EB (1970). Anemia associated with protein deficiency. *Seminars in Hematology* **7**, 55–66. An excellent review of the role of protein deficiency in causing anaemia.

Cox EV (1968). The anaemia of scurvy. *Vitamins and Hormones* **26**, 635–52. An excellent review of the role of vitamin C in haemopoiesis.

Rindi G, *et al*. (1994). Further studies of erythrocyte thiamin transport and phosphorylation in seven patients with thiamin-responsive megaloblastic anaemia. *Journal of Inherited Metabolic Diseases* **17**, 667–77. This study shows the mechanism of thiamine responsive anaemia.

22.5.7 Disorders of the synthesis or function of haemoglobin

D. J. Weatherall

Disorders of the synthesis or structure of haemoglobin may be either inherited or acquired. The inherited disorders of haemoglobin are the commonest single gene disorders in the world population. Figures compiled by the World Health Organization suggest that there are hundreds of millions of carriers. Each year 200 000 to 300 000 severely affected homozygotes or compound heterozygotes are born. In many of the developing countries, the very high mortality from infection and malnutrition in the first year of life causes these conditions to be under-appreciated as an important public health problem. However, once economic conditions improve and infant and childhood death rates fall, the genetic disorders of haemoglobin start to place a major burden on the health services. This phenomenon has already been observed in parts of the Mediterranean region and Southeast Asia.

As a result of mass migrations of populations from high incidence areas for the haemoglobin disorders these conditions are being seen with increasing frequency in parts of the world where they have not been recognized previously. Some of them, particularly sickle cell anaemia and the more severe forms of thalassaemia, can produce life-threatening medical emergencies. It is thus important for clinicians to have a working knowledge of their clinical features, management, and prevention.

Haemoglobin disorders have also become of particular interest in recent years because they were the first group of diseases to be analysed by the methods of recombinant DNA technology. More is known about their molecular pathology than any other genetic disorders. Their study has given us a good idea of the repertoire of mutations that underlie inherited diseases in man.

Before describing the haemoglobin disorders it is necessary to discuss briefly the structure, function, and synthesis of haemoglobin and the way that it is genetically determined.

The structure, function, genetic control, and synthesis of haemoglobin

Structure

Human haemoglobin is heterogeneous at all stages of development; different haemoglobins are synthesized in the embryo, fetus, and adult, each adapted to the particular oxygen requirements.

Each human haemoglobin has a tetrameric structure made up of two different pairs of globin chains, each attached to one haem molecule (Fig. 1). Adult and fetal haemoglobins have a chains combined with b chains (Hb A, a_2b_2), † chains (Hb A_2, $a_2†_2$), or g chains (Hb F, a_2g_2). In embryos, a-like chains, called z chains, combine with g chains to produce Hb Portland (z_2g_2), or with e chains to make Hb Gower 1 (z_2e_2), and a and e chains combine to form Hb Gower 2 (a_2e_2). Fetal haemoglobin is itself heterogeneous; there are two kinds of g chains which differ in their amino acid composition at position 136, where they have either glycine (^{G}g) or alanine (^{A}g). The ^{G}g and ^{A}g chains are the products of separate (^{G}g and ^{A}g) loci.

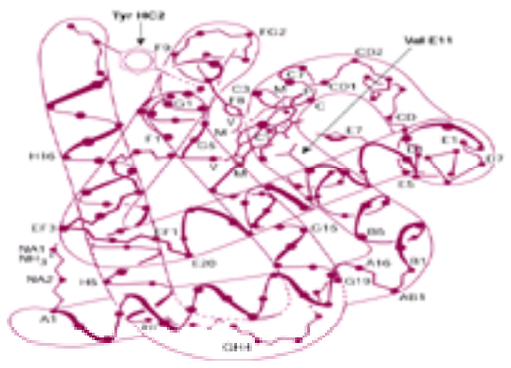

Fig. 1 The a-chain subunit of human haemoglobin showing the position of the haem molecule in a cleft formed by the globin chain. The helical parts of the chain are given letters of the alphabet and each amino acid residue in each helical region has a specific number, for example val E11 is the eleventh amino acid in the E helical region. The non-helical regions of the amino- and carboxyl-terminal ends of the chains are labelled NA and HC respectively. (Reproduced by permission of Dr MF Perutz and the editors of the Cold Spring Harbor Symposia for Quantitative Biology.)

Function

The well-known sigmoid shape of the oxygen dissociation curve, which reflects the allosteric properties of haemoglobin, ensures that oxygen is rapidly taken up at high oxygen tensions in the lungs, and that it is released readily at the lower tensions encountered in the tissues. The shape of the curve is due to co-operativity between the four haem molecules. When one takes on oxygen, the affinity for oxygen of the remaining haems of the tetramer increased dramatically. This is because haemoglobin can exist in two configurations, deoxy (T) and oxy (R) (T and R stand for tight and relaxed states, respectively). The T form has a lower affinity than the R form for ligands such as oxygen. During the sequential addition of oxygen to the four haems, transition from the T to R configuration occurs and the oxygen affinity of the partially liganded molecule increases rapidly.

The position of the oxygen dissociation curve can be modified in many ways. First, oxygen affinity is decreased with increasing CO_2 tensions, the Bohr effect. This facilitates oxygen delivery to the tissues, where a drop in pH due to CO_2 influx lowers oxygen affinity. The opposite effect occurs in the lungs. Oxygen affinity is also modified by the level of 2,3-diphos-phoglycerate (2,3-DPG) in the red cell. Increasing concentrations move the curve to the right, reducing oxygen affinity. Diminishing concentrations have the opposite effect. The 2,3-DPG mechanism plays an important role in response to hypoxia. Increased levels of DPG, with an associated decrease in P_{50} (partial pressure at which haemoglobin is 50 per cent saturated), occur in anaemia, alkalosis, hyperphosphataemia, hypoxic states, and in association with a number of red cell enzyme deficiencies.

Genetic control

The arrangement of the two main families of globin genes is illustrated in Fig. 2. The b-like globin genes form a linked cluster on chromosome 11, that spans about 60 kb (kb = kilobase or 1000 nucleotide bases); they are arranged in the order 5'-e- ^{G}g-^{A}g-yb-†-b-3'. The a-like globin genes form a linked cluster on chromosome 16, in the order 5'-z-yz-ya-a2-a1–3'. The yb, yz, and ya genes are pseudogenes; their sequences resemble the b, z, or a genes but contain mutations which prevent them from functioning as structural genes. They may be 'burnt out' remnants of genes which were functional at an earlier stage of evolution.

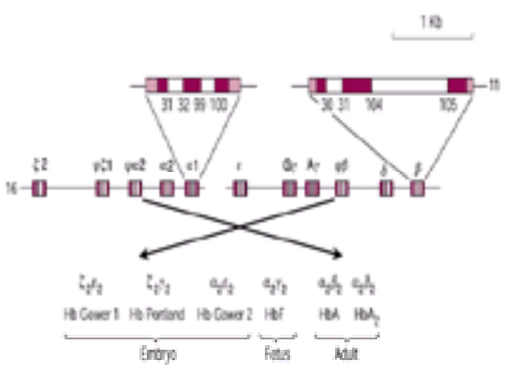

Fig. 2 The genetic control of human haemoglobin. Two of the genes are enlarged to show the introns (unshaded) and exons (dark staining). 1 kb = 1000 nucleotide bases.

Some of the important structural aspects of the globin genes and their flanking sequences are illustrated in Fig. 2 and Fig. 3. Like most mammalian genes, the globin genes are interrupted by one or more non-coding regions called intervening sequences (IVS) or introns. The non-a globin genes contain two introns of 122 to 130 and 850 to 900 base pairs between codons 30 and 31 and 104 and 105, respectively. Similar though smaller introns are found in the a and z globin genes. In the 5' flanking regions of the globin genes there are blocks of nucleotide homology which are found in analagous positions in many species. The first, the ATA box, is about 30 bases upstream (to the left) of the initiation codon. The second, called the CCAAT box, is found about 70 base pairs upstream from the 5' end of the genes. There is a third region of this kind, about 100 base pairs upstream. These regions, called promoters, are involved in the initiation of transcription and hence play an important role in the regulation of the structural genes. As we shall see later, mutations which involve them can reduce the output of the related genes. In the 3' non-coding regions of all the globin genes there is a sequence AATAAA (Fig. 3) which is the signal for polyA addition to RNA transcripts; we shall discuss the significance of this when we consider the disorders of globin chain synthesis.

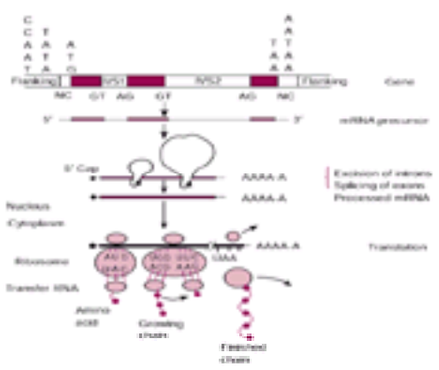

Fig. 3 Globin gene structure, mRNA processing, and globin synthesis. Each of the structures and steps illustrated is described in the text and in Section 5.

The globin gene clusters also contain other types of regulatory elements that interact to promote erythroid-specific gene expression and to co-ordinate changes in globin gene activity during development. They include enhancers, regulatory elements that increase gene expression despite being located at a variable distance from the genes, and master sequences upstream from the clusters which render them transcriptionally active. All these regulatory regions contain sequences to which an array of regulatory molecules called transcription factors are able to bind, some of which are specific for erythropoiesis, while others are ubiquitous in their tissue distribution.

Synthesis

When a globin gene is transcribed a messenger RNA molecule is synthesized from one of its strands by the action of an enzyme called RNA polymerase. The primary transcript of the globin genes is the large messenger RNA precursor molecule which contains both introns and the coding regions or exons. While in the nucleus, this molecule undergoes a number of modifications (Fig. 3). First, the introns are removed and the exons are joined together, a process called splicing. The exon/intron junctions always have the sequence GT at their 5' end, and AG at their 3' end. This appears to be essential for accurate splicing and if there is a mutation in these sites normal splicing cannot occur. The messenger RNAs are chemically modified (capped) at their 5' end, and at their 3' end a string of adenylic acid residues (polyA) is added. The processed messenger RNA now moves into the cytoplasm to act as a template for globin chain production.

Globin mRNA is transported from the nucleus to the cytoplasm where it associates with ribosomes, tRNA, and proteinaceous translation factors. These complexes, called polyribosomes, translate the information encoded in the globin mRNA into the primary amino acid sequence of each globin chain. Individual globin chains combine with haem, which is synthesized through a separate pathway, and with themselves, to form definitive haemoglobin molecules.

Classification of the disorders of haemoglobin

The main groups of disorders of haemoglobin are shown in Table 1. The genetic disorders are divided into those in which there is a reduced rate of production of one or more of the globin chains, the thalassaemias, and those in which a structural change in a globin chain leads to instability or to abnormal oxygen transport. In addition, there is a harmless group of mutations, known collectively as hereditary persistence of fetal haemoglobin, that interfere with the normal switching of fetal to adult haemoglobin production. The acquired disorders of haemoglobin can also be subdivided into those characterized by defective synthesis of the globin chains and those in which the structure of the haem molecules is altered, leading to inefficient oxygen transport.

Like all biological classifications, this way of splitting up the haemoglobin disorders is not entirely satisfactory. For example some structural variants are synthesized in reduced amounts and hence produce the clinical picture of thalassaemia.

The thalassaemias

Historical introduction

The thalassaemias are the commonest of the inherited haematological disorders and, indeed, are the commonest single gene disorders in the world population. The condition was first recognized in 1925 by a Detroit physician called Thomas B. Cooley who described a series of infants who became profoundly anaemic and developed splenomegaly over the first year of life. A milder form was described independently in the same year by an Italian physician, Fernando Rietti. Subsequently, further cases were identified and the disorder was variously called von Jaksch's anaemia, splenic anaemia, erythroblastosis, Mediterranean anaemia, or Cooley's anaemia. In 1936, George Whipple and Lesley Bradford recognized that many of their patients came from the Mediterranean region and hence they invented the word 'thalassaemia' from the Greek qalassa, meaning 'the sea'. Although it was realized later that the disorder occurs throughout the world and is not localized to the Mediterranean region, the name has stuck.

Thalassaemia is extremely heterogeneous. Its clinical picture can result from the interaction of many different genetic defects. This chapter concentrates mainly on the clinical and haematological aspects; readers who wish to learn more about the molecular pathology and population genetics of thalassaemia are referred to several reviews and monographs which are cited at the end of this chapter.

Definition and classification

The thalassaemias are a heterogeneous group of genetic disorders of haemoglobin synthesis, all of which result from a reduced rate of production of one or more of the globin chains of haemoglobin. They are divided into the a, b, †b, or eg†b thalassaemias, according to which globin chain is produced in reduced amounts (Table 2). In some thalassaemias, no globin chain is synthesized at all, they are called a° or b° thalassaemias. In others, the a $^{+}$ or b^{+} thalassaemias, globin chain is produced but at a reduced rate. Thalassaemia occurs in populations in which structural haemoglobin variants are common. It is not at all unusual for an individual to inherit a thalassaemia gene from one parent and a gene for a structural haemoglobin variant from the other. Both a and b thalassaemia occur commonly in some countries and hence individuals may receive genes for both types. These different interactions produce an extremely complex and clinically diverse series of genetic disorders which range in severity from death *in utero* to extremely mild, symptomless hypochromic anaemias.

The thalassaemias are inherited in a simple mendelian fashion. Heterozygotes are usually symptomless, although they can be easily recognized by simple haematological analysis. More severely affected patients are either homozygotes for a or b thalassaemia, compound heterozygotes for different molecular forms of a or b thalassaemia, or compound heterozygotes for thalassaemia and a structural haemoglobin variant. Clinically, the thalassaemias are classified according to their severity in major, intermediate and minor forms. Thalassaemia major is a severe transfusion-dependent disorder. Thalassaemia intermedia is characterized by anaemia and splenomegaly though not of such severity as to require regular transfusion. Thalassaemia minor is the symptomless carrier state. While these descriptive terms do not have a precise genetic meaning, they remain useful in clinical practice.

The b thalassaemias

The b thalassaemias are the most important types of thalassaemia because they are very common and produce severe anaemia in their homozygous and compound heterozygous states (Table 3).

Distribution

The b thalassaemias occur widely in a broad belt ranging from the Mediterranean and parts of north and west Africa through the Middle East and Indian subcontinent to Southeast Asia (Fig. 4). The high incidence zone stretches north through Yugoslavia and Romania and the southern parts of the former USSR and includes the southern regions of the People's Republic of China. The disease is particularly common in Southeast Asia where it occurs in a line starting in southern China and stretching down through Thailand and the Malay peninsula through Indonesia to some of the Pacific islands. In these populations, and in some of the Mediterranean island and mainland countries, gene frequencies for the various forms of b thalassaemia range between two and 20 per cent. It should be remembered that b thalassaemia is not entirely confined to these high incidence regions; it occurs sporadically in every racial group.

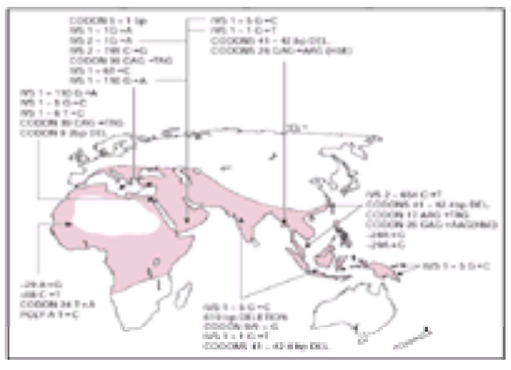

Fig. 4 World map showing the distribution of the different b thalassaemia mutations.

Molecular pathology

The precise molecular lesions responsible for the defective synthesis of the b globin chains have been determined for many patients with b thalassaemia. The disease is extremely heterogeneous. About 200 different mutations can produce the clinical phenotype of b thalassaemia. Some completely inactivate the b globin genes leading to b° thalassaemia; others cause a reduced output from the genes and hence the picture of b $^{+}$ thalassaemia.

The main classes of mutations that cause b thalassaemia are summarized in Fig. 5. With the exception of a deletion of about 600 bases at the 3' end of the b globin gene, which is only found in certain populations of northern India, deletions are an uncommon cause of b thalassaemia. Most of the mutations are single base changes or small deletions and insertions of one or two bases. As shown in Fig. 5, they occur in both introns and exons, and also outside the coding regions.

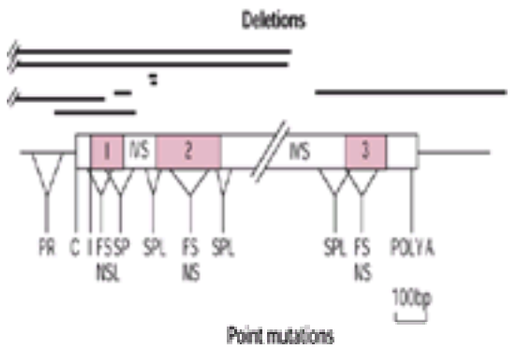

Fig. 5 Some of the mutations that produce b thalassaemia. The b globin gene is divided into three exons (hatched) and two introns (IVS; unshaded). The different deletions are shown at the top of the figure while below the general position of the different point mutations is represented. PR, promoter; C, CAP site; I, initiation site; FS, frameshift; NS, nonsense; SPL, splice-site mutation; polyA, RNA cleavage and polyA addition site.

Many of the exon mutations are nonsense mutations, that is the substitution of a single base in a codon produces a stop codon in the middle of the coding part of the messenger RNA (Fig. 6). Some mutations result in frame shifts; because the information carried by messenger RNA is in the form of a triplet code, the loss of one, two, or four bases throws the reading frame out of phase (Fig. 6). Another important class interfere with splicing. They may alter the invariate GC/AG dinucleotides at the intron/exon junctions, in which case they usually cause b° thalassaemia. Alternatively, they may activate so-called cryptic splice sites, providing an alternate splice site so that both normal and abnormal messenger RNA species are produced (Fig. 7). These lesions cause a b^+ thalassaemia, the severity of which depends on the relative usage of the normal and abnormal splice site and hence the quantity of normal and abnormal b globin messenger RNA that is produced.

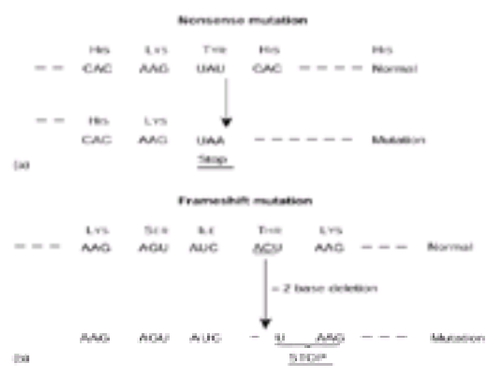

Fig. 6 Point mutations that cause b° thalassaemia: (a) premature stop codon (nonsense mutation); (b) frameshift mutation. See text for further details.

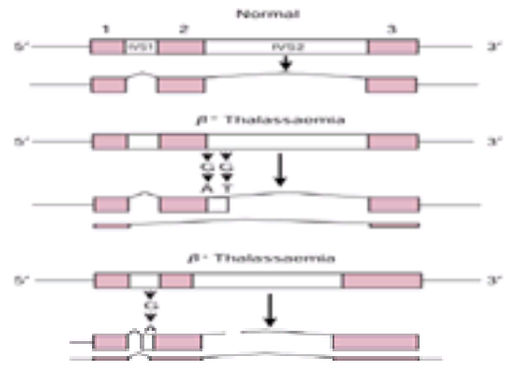

Fig. 7 A representation of the consequences of different splice-site mutations. In b° thalassaemia two different mutations are shown, one that inactivates the normal splice site, and another that produces a new splice site. Two abnormal mRNA molecules are produced. In the b^+ thalassaemia case a new splice site is produced in the first intron. Both normal and abnormal mRNAs are produced, the latter in greater amounts.

Many single base substitutions have also been found in the flanking regions of the b globin genes. They alter either the proximal promoter regions or adjacent sequences, causing down regulation of b globin gene transcription to a varying degree. They are usually associated with milder forms of b^+ thalassaemia.

Because there are so many different b thalassaemia mutations it follows that many patients who are apparently homozygous for b thalassaemia are, in fact, compound heterozygotes for two different molecular lesions.

Pathophysiology

The mutations that cause b thalasaemia result in absent or reduced b chain production. Alpha chain synthesis proceeds at a normal rate and hence there is imbalanced globin chain synthesis (Fig. 8). In the absence of their partner chains the excess a chains are unstable and precipitate in the red cell precursors, forming large intracellular inclusions. These interfere with red cell maturation, and hence there is a variable degree of intramedullary destruction of red cell precursors, that is ineffective erythropoiesis. Those red cells which mature and enter the circulation contain a chain inclusions which interfere with their passage through the microcirculation, particularly in the spleen. These cells are prematurely destroyed. Thus the anaemia of b thalassaemia results from both ineffective erythropoiesis and a shortened red cell survival. The mechanisms of the destruction of red cell precursors and their progeny are extremely complex and are not simply a reflection of mechanical damage to the red cells. Free a chains and their degradation products, particularly haem and iron, cause severe oxidative damage to the red cell membrane proteins. The end result is a dehydrated, rigid erythrocyte with a markedly shortened survival.

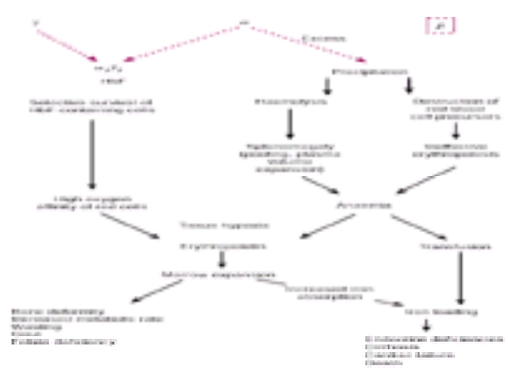

Fig. 8 The pathophysiology of b thalassaemia.

The anaemia acts as a stimulus to increased erythropoietin production, causing massive expansion of the bone marrow which may lead to serious deformities of the skull and long bones. Because the spleen is being constantly bombarded with abnormal red cells, it hypertrophies. The resulting splenomegaly and bone marrow expansion gives rise to an increase in the plasma volume which, together with pooling of the red cells in the enlarged spleen, causes an exacerbation of an already severe degree of anaemia.

As mentioned previously, fetal haemoglobin production largely ceases after birth. However, some adult red cell precursors (F cells) retain the ability to produce a small number of g chains. Because the latter can combine with excess a chains to form haemoglobin F, cells which make relatively more g chains in the bone marrow of b thalassaemics are partly protected against the deleterious effect of a chain precipitation. Red cell precursors which produce haemoglobin F are selected in the marrow and peripheral blood of these patients. Thus, they have relatively large amounts of haemoglobin F in their red cells. Furthermore, because † chain synthesis is unaffected, the disorder is characterized by a relative or absolute increase in haemoglobin A $_2$ ($a_2†_2$) production. These interactions are summarized in Fig. 7.

If the anaemia is corrected with blood transfusion the erythropoietic drive is reduced, growth and development are improved, and bone deformities do not occur. On the other hand, each unit of blood contains 200 mg of iron; with regular transfusion there is steady accumulation of iron in the liver, endocrine glands, and myocardium. Even though well-transfused thalassaemic children grow and develop normally, they die of iron overload unless steps are taken to remove iron.

The severe homozygous or compound heterozygous forms of b thalassaemia

These are the commonest and most important forms of thalassaemia and give rise to a major public health problem in many parts of the world.

Clinical features

Most severe forms of b thalassaemia present within the first year of life, as fetal haemoglobin declines, with failure to thrive, poor feeding, intermittent bouts of fever, or failure to improve after an intercurrent infection. At this stage the affected infant looks pale. In many cases splenomegaly is already present. There are no other specific clinical signs. Diagnosis depends on the haematological changes outlined below. If the infant is established on a regular transfusion regimen at this stage, early development is normal. Further symptoms do not occur until puberty, when the effects of iron loading start to appear. If, on the other hand, the infant is not adequately transfused, the typical clinical picture of homozygous b thalassaemia develops. Thus the clinical manifestations of the severe forms of b thalassaemia have to be described in two contexts, that is the well-transfused child and the child with chronic anaemia throughout early life.

In the well-transfused thalassaemic child, early growth and development is normal. Splenomegaly is minimal. However, there is a gradual accumulation of iron. The effects of tissue siderosis start to appear by the end of the first decade. The normal adolescent growth spurt fails to occur. Hepatic, endocrine, and cardiac complications of iron overloading produce a variety of problems including diabetes, hyopoparathyroidism, adrenal insufficiency, and progressive liver failure. Secondary sexual development is delayed or does not occur at all. Short stature and lack of sexual development may lead to serious psychological problems. By far the commonest cause of death, which usually occurs toward the end of the second or early in the third decade, is progressive cardiac damage. Ultimately these patients die due to either protracted cardiac failure or suddenly due to an acute arrhythmia.

There is now good evidence that children who have been both adequately transfused and chelated may grow and develop normally, pass through a normal puberty, and survive to adult life in excellent condition. However, it is becoming apparent that even children who have been well managed in this way still tend to suffer from complications as they get older, particularly delayed sexual maturation, growth disturbances, and osteoporosis. It seems likely that many of these problems are due to subtle damage to the hypothalamic/pituitary axis with secondary hypogonadism. In addition, some of the growth disturbances may reflect toxicity of the chelating agents used to remove iron (see below).

The clinical picture in children who are inadequately transfused is quite different. Early childhood is interspersed with a series of distressing complications. The overall rates of growth and development are markedly retarded. There is progressive splenomegaly; hypersplenism may cause a worsening of the anaemia, sometimes associated with thrombocytopenia and a bleeding tendency. Because of the bone marrow expansion there may be deformities of the skull with marked bossing and overgrowth of the zygomata giving rise to the classical mongoloid facial appearance of b thalassaemia (Fig. 9(a) and (b)). These findings are reflected by radiological changes which include a lacy, trabecular pattern of the long bones and phalanges and a typical 'hair on end' appearance of the skull (Fig. 10). These bone changes may be associated with recurrent fractures. There is increased susceptibility to infection which may cause a catastrophic drop in the haemoglobin level. Because of the massive marrow expansion, these children are hypermetabolic, run intermittent fevers, lose weight (Fig. 8(b)), have increased requirements for folic acid, and may become acutely folate depleted with worsening of their anaemia. Increased turnover of red cell precursors occasionally gives rise to hyperuricaemia and secondary gout. There is a bleeding tendency which, partly due to thrombocytopenia secondary to hypersplenism, may be exacerbated by liver damage associated with iron loading and extramedullary haemopoiesis. There is also an increased risk of thrombotic complications, possibly reflecting procoagulant properties of the abnormal red cell membranes. The bone deformities of the skull can cause distressing dental complications with poorly formed teeth and malocclusion, and inadequate drainage of the sinuses and middle ear which may lead to chronic sinus infection and deafness. If these unfortunate children survive to puberty, they develop the same complications of iron loading as the well-transfused patients. In this case, some of the iron accumulation results from an increased rate of gastrointestinal absorption as well as that derived from the inadequate transfusion regimen.

Fig. 9 Homozygous b thalassaemia: (a) skull and facial deformity due to bone marrow expansion; (b) gross wasting of the limbs and hepatomegaly in an undertransfused child.

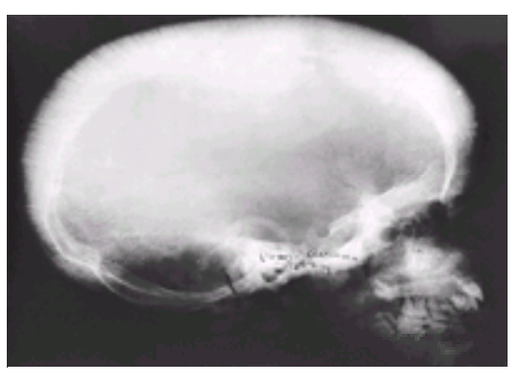

Fig. 10 Radiological changes of the skull in homozygous b thalassaemia.

Haematological changes

There is always a severe anaemia. The haemoglobin values on presentation range from 2 to 8 g/dl. The appearance of the stained peripheral blood film is grossly abnormal (Fig. 11). The red cells show marked hypochromia and variation in shape and size. There are many hypochromic macrocytes and misshapen microcytes, some of which are mere fragments of cells. There is a moderate degree of anisochromia and basophilic stippling. There are always some nucleated red cells in the peripheral blood. After splenectomy, these are found in large numbers. In the postsplenectomy film, many of the nucleated cells and mature erythrocytes show ragged inclusions after incubation of the blood with methyl violet. There is usually a slight elevation in the reticulocyte count. The white cell and platelet counts are normal unless there is hypersplenism in which case they are reduced. The bone marrow shows marked erythroid hyperplasia, with a myeloid/erythroid (M/E) ratio of unity or less. Many of the red cell precursors show ragged inclusions after incubation with methyl violet.

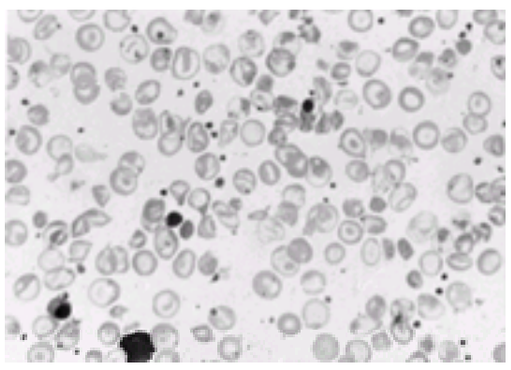

Fig. 11 Peripheral blood film in homozygous b thalassaemia (×630, Leishman stain).

There are biochemical changes of increased haemolysis and progressive iron loading. The bilirubin level is usually elevated and haptoglobins are absent. The ^{51}Cr red cell survival is shortened. The serum iron rises progressively. Most transfusion-dependent children have a totally saturated iron binding capacity. This change is mirrored by a high plasma ferritin level. Liver biopsies show a marked increase in hepatic iron, which may be distributed both in the reticuloendothelial and parenchymal cells (Fig. 12).

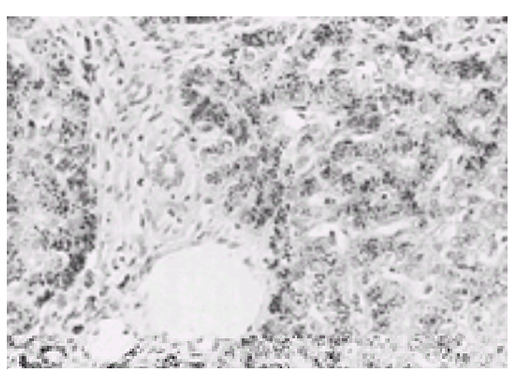

Fig. 12 Histological appearances of the liver in homozygous b thalassaemia showing gross iron deposition (×270, iron stain).

Other biochemical changes

Many thalassaemic children are vitamin E and ascorbate depleted. Folic acid deficiency has already been mentioned. Frank diabetes may develop and endocrine function tests may reveal parathyroid or adrenal insufficiency, or inappropriate response by the pituitary to various release hormones; growth hormone levels are usually normal.

Haemoglobin changes (Table 3)

The haemoglobin F level is always elevated. In b° thalassaemia there is no haemoglobin A and the haemoglobin consists of F and A_2 only. In b^+ thalassaemia the level of haemoglobin F ranges from 30 to 90 per cent of the total haemoglobin. The haemoglobin A_2 level is usually normal and is of no diagnostic value.

Heterozygous b thalassaemia

Carriers for b thalassaemia are usually symptom free except in periods of stress such as pregnancy, when they may become more anaemic than normal women. Splenomegaly is rarely present.

Haematological changes

There is a mild degree of anaemia with haemoglobin values in the 9 to 11 g/dl range. The red cells show hypochromia and microcytosis with characteristically low MCH and MCV values. The reticulocyte count is usually normal. The bone marrow shows moderate erythroid hyperplasia.

Haemoglobin changes

The characteristic finding is an elevated haemoglobin A_2 level in the 4 to 6 per cent range. There is a slight elevation of haemoglobin F in the 1 to 3 per cent range in about 50 per cent of cases. A less common form occurs in which the haemoglobin A_2 is not elevated.

b thalassaemia in association with haemoglobin variants

In many populations where there is a high incidence of both b thalassaemia and various haemoglobin variants it is common for an individual to inherit a b thalassaemia gene from one parent and a gene for a structural haemoglobin variant from the other. Although numerous interactions of this type have been described, in clinical practice only three are of importance, that is sickle cell b thalassaemia, haemoglobin C b thalassaemia, and haemoglobin E b thalassaemia.

Sickle cell b thalassaemia

The clinical manifestations which result from the interaction of the b thalassaemia and sickle cell genes vary considerably from race to race. In African populations, there are mild forms of b^+ thalassaemia which, when they interact with the sickle cell gene, produce a condition characterized by mild anaemia and few sickling crises. This condition is compatible with normal survival and is often ascertained by chance haematological examination. On the other hand, in Mediterranean populations it is quite common for an individual to inherit a b° or severe b $^+$thalassaemia determinant from one parent and a sickle cell gene from the other. These interactions are

often associated with a clinical picture which is indistinguishable from sickle cell anaemia.

The diagnosis of sickle cell thalassaemia rests on the clinical features of a sickling disorder found in association with a peripheral blood picture with typical thalassaemic red cell changes, that is a low MCH and MCV. In the more severe forms of sickle cell b° thalassaemia, there may be an elevated reticulocyte count and sickled red cells are found on the peripheral blood film. The diagnosis can be confirmed by haemoglobin electrophoresis, which in sickle cell b^+ thalassaemia shows haemoglobin S together with 10 to 30 per cent haemoglobin A and an elevated haemoglobin A_2 value. In sickle cell b° thalassaemia, the haemoglobin consists mainly of haemoglobin S with an elevated level of haemoglobins F and A_2. To confirm the diagnosis it is necessary to examine the parents; one should have the sickle cell trait and the other the b thalassaemia trait.

Haemoglobin C thalassaemia

This disorder is restricted to West Africans and some North African and southern Mediterranean populations. It is characterized by a mild haemolytic anaemia associated with splenomegaly. The peripheral blood film shows numerous target cells and thalassaemic red cell changes with a moderately elevated reticulocyte count. Haemoglobin electrophoresis shows a preponderance of haemoglobin C. The diagnosis is confirmed by finding the haemoglobin C trait in one parent and the b thalassaemia trait in the other.

Haemoglobin E b thalassaemia

This is a very common form of thalassaemia in Southeast Asia and throughout the Indian subcontinent. Haemoglobin E is inefficiently synthesized. Thus, when a haemoglobin E gene is inherited together with a b° or severe b^+ thalassaemia determinant, that are the commonest types of b thalassaemia in Southeast Asia, there is a marked deficiency of b chain production. The resulting clinical picture can closely resemble thalassaemia major.

The clinical and haematological changes in haemoglobin E thalassaemia are variable. There is usually a marked degree of anaemia and splenomegaly with typical thalassaemic bone changes (Fig. 13). Although not always transfusion dependent, patients with this disorder usually have low haemoglobin values in the 4 to 9 g/dl range with an average of 5 to 7 g/dl. The blood film shows typical thalassaemic red cell changes and the bone marrow shows marked erythroid hyperplasia with a chain inclusions in many of the red cell precursors. Although very little is known about the natural history of this disorder, it seems likely that in many parts of Southeast Asia and India it causes a very high mortality in the early years of life. Complications include a marked susceptibility to infection, secondary hypersplenism, progressive iron loading, a variety of neurological lesions (due to tumours caused by extramedullary erythropoiesis extending in from the inner tables of the skull or vertebrae), folate deficiency, and recurrent pathological fractures. On the other hand, some patients with haemoglobin E thalassaemia grow and develop normally with few complications and there are many recorded cases of pregnancy in women with this disorder.

Fig. 13 Bossing of the skull in haemoglobin E thalassaemia.

The diagnosis of haemoglobin E thalassaemia is confirmed by finding haemoglobins E and F and little or no haemoglobin A on haemoglobin electrophoresis and by demonstrating the haemoglobin E trait in one parent and the b thalassaemia trait in the other.

Other b thalassaemia variants

It is not uncommon to encounter patients with the clinical and haematological features of heterozygous b thalassaemia who do not have an elevated haemoglobin A_2 level. Many of these individuals are heterozygotes for both b and † thalassaemia. It is important to recognize this interaction because, if it is inherited together with a typical b thalassaemia gene, it can produce a severe transfusion-dependent disorder. Hence this variant is important in antenatal screening programmes. It can only be identified for certain by globin chain synthesis or gene analysis in a specialized laboratory. Families are encountered occasionally in which there is a more severe form of heterozygous b thalassaemia associated with anaemia, jaundice, and splenomegaly. In some of these families it is apparent that the affected individuals are in fact compound heterozygotes for b thalassaemia and the so-called 'silent' b thalassaemia gene, that is a determinant which cannot be identified haematologically in heterozygotes. In other families, a severe form of b thalassaemia behaves as a single gene disorder with full expression in heterozygotes, that is it follows a dominant form of inheritance. In most of these families, the disorder results from the synthesis of a highly unstable b globin chain.

The †b thalassaemias (Table 3)

Molecular genetics and classification

Disorders due to reduced b and † chain synthesis are much less common than those due to defective b chain production. They are remarkably heterogeneous at the molecular level. In some cases they result from deletions of the b and † globin genes, while in others there appears to have been mispaired synapsis and unequal crossing over between the † and b globin gene loci with the production of †b fusion genes. The latter produce †b fusion chains which combine with a chains to form haemoglobin variants called the Lepore haemoglobins (Lepore was the family name of the first patient to be recognized with this disorder). Hence it is usual to classify this group of conditions into the (†b)° thalassaemias and the haemoglobin Lepore or $(†b)^+$ thalassaemias.

Clinical and haematological changes

The (†b)° thalassaemias have been reported in many populations although there are no high frequency areas. In the homozygous state there is a mild degree of anaemia with haemoglobin values of 8 to 10 g/dl. There is often a moderate degree of splenomegaly but these patients are usually symptomless except during periods of stress such as infection or pregnancy. Haemoglobin analysis shows 100 per cent haemoglobin F. Heterozygous carriers have thalassaemic blood pictures, elevated levels of haemoglobin F of 5 to 20 per cent, and normal levels of haemoglobin A_2.The homozygous state for haemoglobin Lepore is characterized by a clinical picture which is usually similar to that of homozygous b thalassaemia although in some cases it may be milder and non-transfusion dependent. The haematological findings are similar to those of b thalassaemia. The haemoglobin consists of F and Lepore only. Heterozygous carriers have thalassaemic blood pictures associated with about 5 to 15 per cent haemoglobin Lepore.

The (eg†b)° thalassaemias

There are several rare forms of thalassaemia which result from long deletions of the b globin gene cluster which, as well as removing or inactivating the b genes, involve the †, g, and embryonic e genes. They also involve the main regulatory sequence upstream of the b globin gene cluster, the locus control region. This means that there is no output of globin chains from this gene cluster at all. Clearly, the homozygous state for these disorders would not be compatible with survival. Heterozygotes often have severe haemolytic disease of the newborn with anaemia and hyperbilirubinaemia. If they survive the neonatal period they grow and develop normally; in adult life they have the haematological picture of heterozygous b thalassaemia with mild anaemia, hypochromic microcytic red cells, and a haemoglobin

pattern consisting of haemoglobin A, no elevation of haemoglobin F, and a normal level of haemoglobin A_2.

Hereditary persistence of fetal haemoglobin

There is a complex family of conditions characterized by persistent fetal haemoglobin synthesis into adult life associated with no major haematological abnormalities. In some cases they result from long deletions of the b globin gene cluster, similar to those which cause †b thalassaemia. Indeed, they form a continuum with this condition; homozygotes have 100 per cent fetal haemoglobin, elevated haemoglobin levels and no clinical findings. Other forms result from point mutations in the promoter regions of the g globin genes. In this case there is increased g chain production together with reduced b chain production on the affected chromosome. Hence, homozygotes have markedly elevated levels of haemoglobin F but also produce some haemoglobin A. Finally, there is a group in which persistent low levels of haemoglobin F, in the 3 to 10 per cent range, are observed. There is increasing evidence that they may result from mutations either within the b globin gene cluster or on other chromosomes.

The only clinical importance of this complex group of conditions is that they may interact with the thalassaemias or structural haemoglobin variants and reduce the severity of different phenotypes by increasing the amount of haemoglobin F that is produced.

The a thalassaemias

Although the a thalassaemias are commoner on a global basis than the b thalassaemias they pose less of a public health problem. This is because the severe, homozygous forms cause death *in utero* or in the neonatal period and the milder forms do not produce major clinical problems.

Distribution

The a thalassaemias occur widely through the Mediterranean region, parts of West Africa, the Middle East, parts of the Indian subcontinent, and throughout Southeast Asia in a line stretching from southern China through Thailand, the Malay peninsula, and Indonesia to the Pacific island populations (Fig. 14). For reasons which will become apparent when we consider the molecular pathology of these disorders, the serious forms of a thalassaemia are restricted to some of the Mediterranean island populations and Southeast Asia.

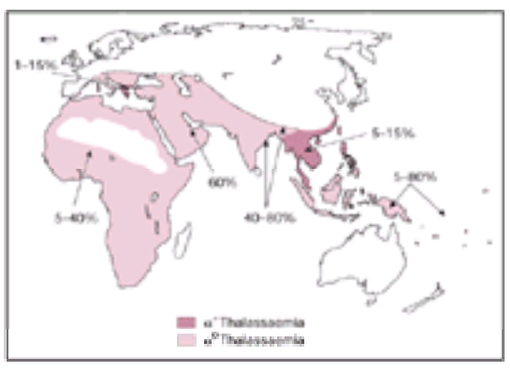

Fig. 14 World map showing the distribution of the a thalassaemias.

Inheritance and molecular pathology

The genetics of a thalassaemia is complicated, and has generated a confusing nomenclature over the years.

Because both haemoglobins A and F have a chains, genetic disorders of a chain synthesis result in defective fetal and adult haemoglobin production. In the fetus, deficiency of a chains leads to the production of excess g chains which form g_4 tetramers, or haemoglobin Bart's (Fig. 15). In adults, a deficiency of a chains leads to an excess of b chains which form b_4 tetramers, or haemoglobin H, the adult counterpart of haemoglobin Bart's. Thus, the presence of haemoglobins Bart's or H in red cells is the hallmark of a thalassaemia. For reasons which are not yet clear, a critical level of globin chain imbalance is required before detectable amounts of haemoglobins Bart's or H appear in the red cells. Unfortunately for clinicians, in persons with mild forms of a thalassaemia this level is not reached; significant amounts of these variants only occur in the red cells of patients who have a severe degree of a chain deficiency. This means that the carrier states for different forms of a thalassaemia are difficult to diagnose.

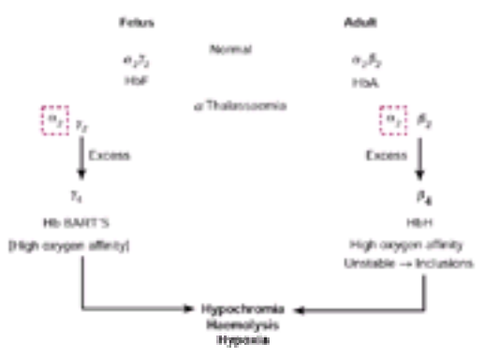

Fig. 15 The pathophysiology of a thalassaemia.

Because normal individuals receive two a globin genes from each of their parents, aa/aa, the genetics of the a thalassaemias is more complicated than that of the b thalassaemia. It is useful to define these conditions in heterozygotes. First, there is a more severe form which is called a° thalassaemia, which results from loss of both of the linked a globin genes, – –/aa. The second type is almost completely silent in carriers; their red cells are normal or are only slightly hypochromic. This condition is due to the deletion, –a/aa, or reduced activity due to a mutation, a ^{T}a/aa, of one of the linked a globin genes. Because there is still some output of a globin from the affected chromosome this is called a^+ thalassaemia. To put it in another way, the terms a° and a^+ thalassaemia describe haplotypes, that is the products of two linked a globin genes on one of a pair of homologous chromosomes 16.

In clinical practice we encounter two symptomatic types of a thalassaemia, the haemoglobin Bart's hydrops syndrome and haemoglobin H disease (Table 4). The former results from the homozygous inheritance of a° thalassaemia. On the other hand, haemoglobin H disease usually results from the coinheritance of both a° and a^+ thalassaemia. We now know that there are many different molecular types of both a° and a^+ thalassaemia. These genetic interactions are summarized in Fig. 16.

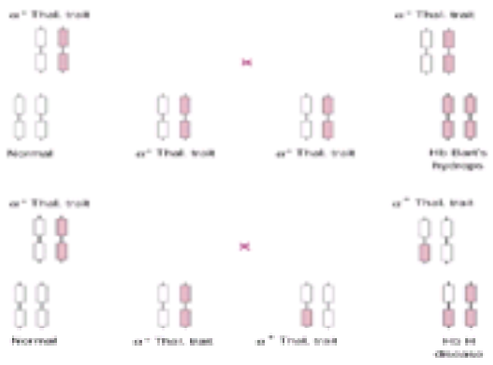

Fig. 16 The genetics of a thalassaemia. The black a genes represent gene deletions or otherwise inactivated genes. The open a genes represent normal genes. a° Thalassaemia and a^+ thalassaemia are defined in the text.

Like the b thalassaemias, the a thalassaemias are extremely heterogeneous at the molecular level. Many different sized deletions can remove either both the a globin genes or the main regulatory regions of the a globin gene cluster and cause a° thalassaemia, but there are only two which are common. One is found in Southeast Asia. The other occurs mainly in Mediterranean populations (Fig. 17). Similarly, there are several different sized deletions that remove a single a globin gene to produce the deletion forms of a^+ thalassaemia; the commonest are those that remove either 3.7 kb or 4.2 kb of the a gene cluster. There are also many different mutations that can produce the non-deletion forms of a^+ thalassaemia. Many of them are similar to those which produce b thalassaemia. A particularly common form of non-deletion a^+ thalassaemia, found in up to 5 per cent or more of some Southeast Asian populations, results from a single base change in the a globin chain termination codon UAA, which changes to CAA. The latter is the code word for the amino acid glutamine. When the ribosomes reach this point, instead of the chain terminating, they read through messenger RNA that is not normally translated until another stop codon is reached. An elongated a chain variant is synthesized, but the messenger RNA is destabilized by read through of sequences which are not normally translated and so the variant is also produced at a reduced rate. It is called haemoglobin Constant Spring after the name of the town in Jamaica in which it was first discovered. Several other chain termination mutations of this type occur, with different base changes in the terminating codon and hence different amino acids at the beginning of the extension of the a globin chain.

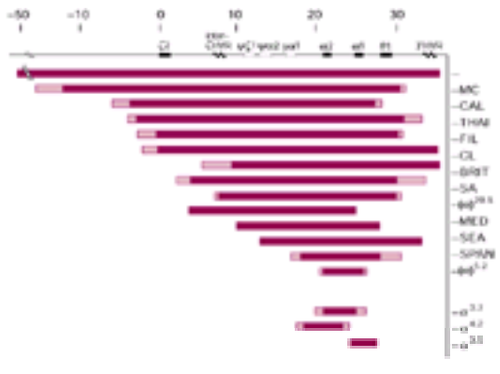

Fig. 17 The different-sized deletions responsible for some forms of a° or a^+ thalassaemia. The a globin gene cluster is shown at the top of the figure. Two highly variable regions (HVR) are shown. The abbreviations on the right-hand side indicate the source of origin of patients with the deletions: MED, Mediterranean; SEA, Southeast Asia. The three smaller deletions at the bottom of the figure show some of the main classes of a^+ thalassaemia. The superscripts 3.7, 4.2, and 3.5 indicate the size of the deletions.

Genotype/phenotype relationships

Molecular studies explain much of the clinical variability of a thalassaemia in different populations. Since the haemoglobin Bart's hydrops syndrome requires the homozygous inheritance of a° thalassaemia (– –/– –), this condition only occurs in populations in which a° thalassaemia is common. It is mainly confined to Southeast Asia and the Mediterranean islands, populations in which the haemoglobin Bart's hydrops syndrome causes a public health problem. Most forms of haemoglobin H disease are due to the inheritance of a° thalassaemia from one parent and a^+ thalassaemia from the other (–a/– – or $–a^T/– –$). Thus, haemoglobin H disease is also restricted mainly to Mediterranean and Oriental populations. On the other hand, a^+ thalassaemia occurs very commonly throughout parts of West Africa, the Indian subcontinent, and the Pacific island populations. a° Thalassaemia does not occur in these regions so that the haemoglobin Bart's hydrops syndrome and haemoglobin H disease are not seen. The homozygous state for a^+ thalassaemia (–a/–a) is characterized by a mild hypochromic anaemia, very similar to the heterozygous state for a° thalassaemia; the results of having only two out of the normal four a genes seem to be the same whether the two genes are missing from the same chromosome or opposite pairs of homologous chromosomes. To complicate matters, sometimes the homozygous state for the non-deletion forms of a^+ thalassaemia, a^Ta/a^Ta, are more severe and cause haemoglobin H disease.

Pathophysiology

The pathophysiology of a thalassaemia is different to that of b thalassaemia. A deficiency of a chains leads to the production of excess g chains or b chains which form haemoglobins Bart's and H respectively (Fig. 15). These more soluble tetramers do not precipitate to any great extent in the bone marrow. Erythropoiesis is thus more effective than in b thalassaemia, that is there is less intramedullary destruction of red cell precursors. However, haemoglobin H is unstable and precipitates in red cells as they age. The large inclusion bodies produced in this way are trapped in the spleen and other parts of the microcirculation leading to a shortened red cell survival. Furthermore, both haemoglobins Bart's and H have a very high oxygen affinity; because they have no a chains there is no haem/haem interaction and their oxygen dissociation curves resemble myoglobin. Thus the pathophysiology of severe forms of a thalassaemia is based on defective haemoglobin production, the synthesis of homotetramers which are physio-logically useless, and a haemolytic component due to their precipitation in older red cells. Furthermore, excess b chains cause a different pattern of damage to red cell membrane proteins than free a chains; the red cells tend to be overhydrated in a thalassaemia.

The haemoglobin Bart's hydrops syndrome

This condition is a common cause of fetal loss throughout Southeast Asia and in Greece and Cyprus. Affected infants produce no a chains and hence can make neither fetal nor adult haemoglobin.

The clinical picture is very characteristic (Fig. 18). Infants are usually stillborn between 28 and 40 weeks. Liveborn infants take a few gasping respirations and then expire within the first hour after birth. They show the typical picture of hydrops fetalis with gross pallor, generalized oedema, and massive hepatosplenomegaly. There is a high frequency of other congenital abnormalities, and a very large, friable placenta, all due to severe intrauterine anaemia. The haemoglobin values are in the 6 to 8 g/dl range and there are gross thalassaemic changes of the peripheral blood film with many nucleated red cells. The haemoglobin consists of approximately 80 per cent haemoglobin Bart's and 20 per cent of the embryonic haemoglobin, Portland (z_2g_2). It is believed that these infants survive to term because they continue to produce embryonic haemoglobin at this level; haemoglobin Bart's is, as mentioned above, useless as an oxygen carrier.

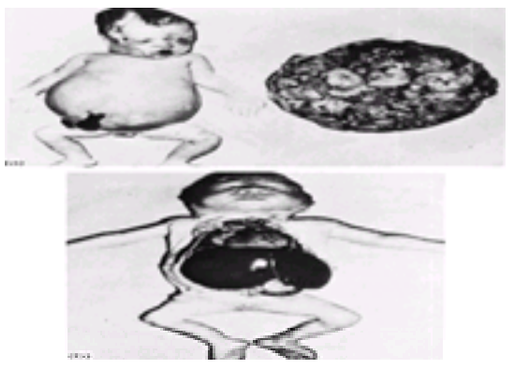

Fig. 18 The haemoglobin Bart's hydrops syndrome: (a) a hydropic infant with massively enlarged placenta; (b) autopsy findings with an enlarged liver. (Reproduced by permission of Professor P. Wasi.)

This syndrome is also characterized by a high incidence of maternal toxaemia of pregnancy and considerable obstetric difficulties due to the presence of the large, friable placenta. Both parents have thalassaemic red cell changes with normal haemoglobin A_2 values, that is the characteristic finding of the heterozygous state for a° thalassaemia.

Haemoglobin H disease

As mentioned earlier, haemoglobin H is a tetramer of normal b chains with the formula b_4. It is produced when there is a marked reduction of a chain synthesis. Haemoglobin H disease usually results from the inheritance of a° thalassaemia from one parent and a^+ from the other. It may also result from the inheritance of a° thalassaemia and haemoglobin Constant Spring or from the homozygous state for a severe, non-deletion form of a thalassaemia. The latter form of inheritance is particularly common in Saudi Arabia.

There is a variable degree of anaemia and splenomegaly but it is most unusual to see severe thalassaemic bone changes or the growth retardation characteristic of homozygous b thalassaemia. Patients usually survive into adult life although the course may be interspersed with severe episodes of haemolysis associated with infection, or worsening of the anaemia due to progressive hypersplenism. Oxidant drugs such as sulphonamides may increase the rate of precipitation of haemoglobin H and therefore exacerbate the anaemia.

Haemoglobin values range from 7 to 10 g/dl. The blood film shows typical thalassaemic changes. There is a moderate reticulocytosis. Incubation of the red cells with brilliant cresyl blue generates numerous inclusion bodies by precipitation of the haemoglobin H under the redox action of the dye. After splenectomy large, preformed inclusions can be demonstrated on incubation of blood with methyl violet. The haemoglobin consists of from 5 to 40 per cent haemoglobin H together with haemoglobin A and a normal or reduced level of haemoglobin A_2.

Usually one parent is heterozygous for a° thalassaemia and the other for a^+ thalassaemia, the deletion or non-deletion varieties. Less commonly, both parents are heterozygous for a non-deletion form of a^+ thalassaemia.

The haematological findings in the a° and a^+ thalassaemia traits are summarized in Table 4. They can only be identified with certainty by analysis of the a globin genes.

a Thalassaemia and mental retardation

There is an increasingly important group of a thalassaemias which are not restricted to individuals from tropical backgrounds. They are observed in all racial groups and have been best characterized in those of north European origin. These conditions are characterized by variable degrees of mental retardation, dysmorphic features, and a thalassaemic blood pictures. They follow a completely different form of inheritance to the commoner genetic forms of a thalassaemia and constitute a increasingly heterogeneous group of disorders. There are two major varieties of this condition. The first is due to lesions that involve the a globin gene cluster on chromosome 16, ATR-16. There is another group that result from mutations on the X chromosome, ATR-X.

The ATR-16 disorders are characterized by a very variable degree of mental retardation and equally variable dysmorphic features. The blood film shows mild a thalassaemic changes and some cells which contain typical haemoglobin H inclusion bodies. It is now clear that they have a heterogeneous molecular pathology. In some cases the condition results from long deletions which remove the end of the short arm of chromosome 16 and extend for one to two megabases. Occasionally they also include the genes that are involved in tuberous sclerosis and adult polycystic disease of the kidney. In other cases, the loss of the end of the short arm of chromosome 16 is the result of an inherited cytogenetic abnormality, including translocations and other rearrangements.

The ATR-X syndrome is characterized by a much more consistent series of dysmorphic features and more severe mental retardation. These infants often suffer from convulsions after birth. They develop typical facial features, genital abnormalities, and a very mild form of haemoglobin H disease. This condition is inherited as a typical sex-linked disorder which affects males. It results from mutations of a gene on the X chromosome called *ATR-X* which is now known to act as a regulator of transcription via an effect on the structure of chromatin. Female carriers may show a very small proportion of red cells containing haemoglobin H bodies. This condition should be thought of in any child with severe mental retardation and dysmorphic features whose blood film shows evidence of a very mild form of a thalassaemia.

Thalassaemia intermedia

Definition and pathogenesis

The term thalassaemia intermedia is used to describe patients with the clinical picture of thalassaemia which, although not transfusion dependent, is associated with a much more severe degree of anaemia than that found in carriers for a or b thalassaemia. Many of the conditions which have been described previously in this section follow this clinical course, for example haemoglobin C or E thalassaemia, the various †b thalassaemias and haemoglobin Lepore disorders, and the wide variety of conditions which can result from the interactions of the different b and †b thalassaemia determinants. However, some children with this condition have parents with typical heterozygous b thalassaemia blood pictures and elevated haemoglobin A_2 levels. These patients appear to be homozygous for b thalassaemia, yet they run a much milder course than is usually the case with this condition. Some of them have inherited an a thalassaemia determinant as well as being homozygous for b thalassaemia. This reduces the overall degree of globin chain imbalance and consequently the severity of the dyserythropoiesis which usually accompanies homozygous b thalassaemia; hence these children run a milder clinical course. In other cases, particularly in African races, relatively mild forms of homozygous b thalassaemia seem to reflect the action of less severe b thalassaemia mutations. Finally, some intermediate forms of b thalassaemia seem to result from the coinheritance of a gene for unusually effective haemoglobin F production.

Clinical and haematological changes

The clinical features of the intermediate forms of thalassaemia are extremely variable. At one end of the spectrum are patients who are virtually symptom free except for moderate anaemia. At the other end there are patients who have haemoglobin values in the 5 to 7 g/dl range and who develop marked splenomegaly, severe skeletal deformities due to expansion of bone marrow, and, as they get older, become heavily iron loaded because of increased intestinal absorption of iron. Recurrent leg ulceration, folate deficiency, symptoms due to extramedullary haemopoietic tumour masses in the chest and skull (Fig. 19), gallstones, and a marked proneness to infection are particularly characteristic of this group of thalassaemias.

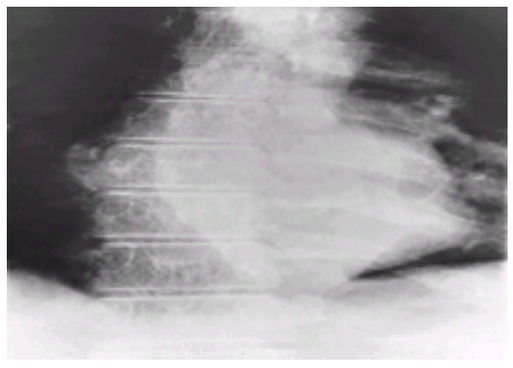

Fig. 19 An extramedullary haemopoietic mass in a patient with b thalassaemia intermedia.

Because of the heterogeneity of these disorders, it is only possible to determine the course that is likely to evolve in any individual patient by following the disorder very carefully from early childhood.

Differential diagnosis of the thalassaemias

There are few conditions which are likely to be confused with the more severe forms of homozygous b thalassaemia or haemoglobin H disease. The racial background of the patient, the presence of anaemia from early life, and the characteristic haematological changes make the diagnosis relatively easy. Once thalassaemia is suspected, the parents and near relatives should be examined for the carrier states for a or b thalassaemia. Both disorders can be distinguished from simple iron deficiency by the finding of a normal serum iron or ferritin level and by the associated changes in the haemoglobin pattern. It should be remembered, however, that in some groups iron deficiency and heterozygous thalassaemia frequently occur together in the same person, particularly during pregnancy. The sideroblastic anaemias can be easily distinguished from thalassaemia by the morphological appearances of the red cells and the presence of ring sideroblasts in the bone marrow. It should be remembered that there are some rare forms of acquired haemoglobin H disease in elderly patients with leukaemia.

The laboratory diagnosis of thalassaemia

The thalassaemias should be suspected when a typical thalassaemic blood picture is found in an individual of an appropriate racial group. The homozygous states for the severe forms of b thalassaemia are easily recognized by the haematological changes associated with very high levels of haemoglobin F; haemoglobin A_2 values vary so much that they are of no diagnostic help. The heterozygous states are recognized by microcytic hypochromic red cells and an elevated level of haemoglobin A_2. The †b thalassaemias are characterized by the finding of 100 per cent haemoglobin F in homozygotes and 5 to 15 per cent haemoglobin F together with a normal level of haemoglobin A_2 in heterozygotes (see Table 3).

When b thalassaemia is diagnosed, a quantitative haemoglobin electrophoresis should be carried out to exclude the presence of an abnormal haemoglobin variant such as haemoglobin E or Lepore.

The haemoglobin Bart's hydrops syndrome is recognized by the finding of a hydropic infant with a severe anaemia, a thalassaemic blood picture, and 80 per cent or more haemoglobin Bart's on haemoglobin electrophoresis. Haemoglobin H disease is identified by the finding of a typical thalassaemic blood picture with an elevated reticulocyte count, generation of multiple inclusion bodies in the red cells after incubation with brilliant cresyl blue, and the finding of variable amounts of haemoglobin H on haemoglobin electrophoresis. There are no really useful, simple diagnostic tests for the different a thalassaemic carrier states although a° thalassaemia heterozygotes usually have typical thalassaemic red cell changes with a normal haemoglobin A_2 value. It is essential for counselling purposes to diagnose the different carrier states for a thalassaemia, blood samples should be referred to a laboratory that can carry out DNA analysis of the globin genes.

Prevention and treatment

Thalassaemia produces a severe public health problem and a serious drain on medical resources in many populations. Since there is no definitive treatment, most countries in which the disease is common are putting a major effort into programmes for its prevention.

Prevention

There are two major approaches to the prevention of the thalassaemias. Since the carrier states for the b thalassaemias can be easily recognized, it is at least theoretically possible to screen populations and provide genetic counselling about the choice of marriage partners. If b thalassaemia heterozygotes marry other carriers, one in four of their children will have the severe, transfusion-dependent homozygous disorder. While large-scale programmes of this type have been set up in Italy, the results are not yet available, and in smaller pilot studies in Greece the outcome has not been encouraging. Until more is known about the usefulness of this form of prospective genetic counselling, most countries are developing screening programmes at antenatal clinics. When heterozygous carrier mothers are found, the husbands are tested and if they are also carriers the couple are offered the possibility of prenatal diagnosis and termination of pregnancies carrying fetuses with severe forms of thalassaemia.

Prenatal diagnosis

Prenatal diagnosis can be offered to couples at risk for having children with severe forms of b thalassaemia. Because of the serious obstetric complications and the trauma of carrying a hydropic fetus to term there is also a good case for prenatal diagnosis for the haemoglobin Bart's hydrops syndrome. Termination of pregnancies at risk for milder forms of thalassaemia is also undertaken, but should only be considered after very careful counselling of the parents. Some children with intermediate forms of thalassaemia are symptom free and develop normally; others have more severe anaemia and bone deformity. There has been some success in determining which particular molecular defects and interactions are associated with these different clinical courses. When in doubt, parents should be referred for expert analysis of their variety of thalassaemia and appropriate counselling.

Prenatal diagnosis of thalassaemia can be carried out in several ways. The diagnosis can be made by globin-chain-synthesis studies of fetal blood samples obtained by fetoscopy at 18 to 20 weeks gestation. The diagnosis can also be made by fetal DNA analysis on amniotic fluid cells obtained by amniocentesis earlier in the second trimester. More recently, it has been possible to carry out prenatal diagnosis of thalassaemia and sickle cell anaemia by direct analysis of fetal DNA obtained by chorion biopsy at about the 12th week of gestation. This approach has largely replaced fetal blood sampling or amniocentesis for the prenatal diagnosis of the thalassaemias. First trimester diagnosis is much more acceptable to many women. This reduces the long period of uncertainty, during which the fetus is growing and the mother and her relatives and friends are coming to accept that she is to have a child, and because late second trimester terminations are often difficult. Prenatal diagnosis of thalassaemia is now carried out in many countries, and in Sardinia, Greece, and Cyprus has significantly reduced the number of new cases of thalassaemia in the community.

Because prenatal diagnosis of thalassaemia is now well established it is very important to discuss the genetic implications of the condition when carriers are detected by chance, regardless of the individuals's racial background. They should also be given a letter explaining, in simple terms, the pattern of inheritance and the dangers for their children. This approach should always be followed, even for sporadic cases in low incidence regions, such as northern Europe. Because of the increasing movements of populations they might still marry another carrier and have severely affected children.

Symptomatic treatment

The symptomatic management of severe b thalassaemia requires regular blood transfusion, the judicious use of splenectomy if hypersplenism develops, and the administration of chelating agents to reduce iron overload. When the diagnosis of severe b thalassaemia is suspected during the first year of life, the infant should be followed for several weeks to make sure that the haemoglobin level is fallen to a level at which regular transfusion will be necessary. It is difficult to be dogmatic about

exactly when transfusions should be started. A severely anaemic infant who is feeding poorly, inactive, or otherwise failing to thrive, will almost certainly need to be transfused. The object is to maintain the pretransfusion haemoglobin level at about 9.5 g/dl. This usually requires transfusion of 10 to 15 mg/kilo red cells every 4 weeks. Washed red cells should be used. Whole blood should be avoided because of the danger of sensitization to serum or white cell components. The rate of transfusion should not exceed 4 to 5 ml/kg per h. In patients who are profoundly anaemic or show evidence of cardiac insufficiency, the rate should be no more than 2 ml/kg per h. It is important to calculate the annual blood consumption by dividing the total volume of blood transfused over 12 months by the patients weight in the middle of the year. If it is higher than 200 ml/kg body weight, splenectomy should be considered. All blood should be screened for hepatitis B and C, and for HIV.

Hypersplenism is becoming much less common if children are maintained on an adequate transfusion regimen. Increasingly blood requirements, or evidence of hypersplenism, pancytopenia for example, should prompt one to consider splenectomy. It should be avoided before the age of 6 years because of the particularly high incidence of infection in asplenic children. Two to three weeks before splenectomy the child should be given: (1) pneumococcal vaccine; (2) *Haemophilus influenzae* type B vaccine; (3) meningococcal A and C vaccine. After the operation the children should be maintained on oral penicillin V, 125 mg twice daily, increasing to 250 mg twice daily for older children. For those who are allergic to penicillin, erythromycin should be given.

The only effective chelating agent for the prevention or treatment of iron overload in thalassaemia is desferrioxamine (Desferal). It is now clear that this drug should not be given too soon because toxic effects are observed at low body iron loads. Ideally, the hepatic iron concentration should be measured at about 1 year after regular transfusion has started. Chelation should be initiated in patients with hepatic iron concentrations of above 7 mg/g liver dry weight. Where this is not possible, the drug should be given when the serum ferritin value has reached or exceeded 1000 µg/l although it is becoming increasingly clear that the serum ferritin level is a very imprecise estimate of body iron load. The initial dose should not exceed 25 to 35 mg/kg body weight per 24 h. Iron excretion is potentiated if children receive 100 mg vitamin C by mouth on the days of the infusion. Ideally, progress should be monitored by regular estimates of the hepatic iron concentration but if this is not possible the serum ferritin level should be maintained below 2000 µg/l. In patients who become iron loaded, it is possible to increase the rate of iron excretion considerably but the daily dose of desferrioxamine should not exceed 15 mg/kg body weight. Patients should be monitored continuously for side effects of desferrioxamine; these include retinal damage, ototoxicity, and interference with growth.

There are no entirely satisfactory alternatives to desferrioxamine as a chelating agent. The most widely studied, the oral chelator deferiprone (L1), does not appear to control iron accumulation in a proportion of patients, and causes neutropenia in about 5 per cent of cases. Its long-term toxicity and true place in the management of thalassaemia remains to be determined.

It is very important to monitor transfusion-dependent patients for hepatitis B and C and HIV infection. The management of these conditions is considered elsewhere in this book. Other complications relating to iron load, including hypoparathyroidism, diabetes, and delayed puberty and hypogonadism, require expert endocrinological assessment with appropriate replacement therapy.

Increasing experience with bone marrow transplantation has suggested that, if done early with adequate HLA matching, the results are extremely good. Patients who have become iron loaded and who have liver damage have a less good prognosis but, as more experience has been gained, there appears to be a place for transplantation in older patients.

The intermediate forms of thalassaemia should be treated by careful observation, folic acid supplementation, and, in the face of a falling haemoglobin and increasing spleen size, the judicious use of splenectomy. It is important to monitor the iron status regularly because some of these patients become iron loaded due to increased intestinal absorption later in life and chelation therapy may be necessary.

Currently, a number of experimental approaches to the treatment of the thalassaemias are being pursued, including the use of intrauterine or later stem cell therapy, the stimulation of fetal haemoglobin production, and, in the longer term, the possibility of somatic gene therapy.

Structural haemoglobin variants

Over 400 structural haemoglobin variants have been described, most of which result from single amino acid substitutions. Many of them are harmless and have been discovered during surveys of the electrophoretic patterns of human haemoglobin. Of course, this approach underestimates the number of variants because it only identifies those in which the amino acid substitution alters the charge of the haemoglobin molecule.

Single amino acid substitutions cause clinical disorders only if they alter the stability or functional properties of the haemoglobin molecule. A classification of these diseases is shown in Table 5. They include the sickling disorders, chronic or drug-induced haemolytic anaemia associated with unstable haemoglobins, and polycythaemia or congenital cyanosis, associated with high and low oxygen affinity haemoglobin variants, respectively. There is a rare group of haemoglobin variants that produce methaemoglobinaemia. We shall consider the different varieties of genetic methaemoglobinaemias at the end of this chapter.

Nomenclature

Originally, the structural haemoglobin variants were named by letters of the alphabet. By the late 1950s there were none left; it was decided to designate new haemoglobin variants by the place of origin of the first patient in whom they were characterized. It is customary to call the heterozygous carrier state the 'trait' and the homozygous condition the 'disease'. For example, haemoglobin S heterozygotes (genotype AS) are said to have the sickle cell trait, while those homozygous for the sickle cell mutation (genotype SS) are said to have sickle cell disease. In practice it is very important to distinguish between the carrier state and the homozygous or compound heterozygous state for a haemoglobin variant; carriers are usually asymptomatic.

The sickling disorders

Sickling disorders (Table 6) consist of the heterozygous state for haemoglobin S, sickle cell trait (AS), the homozygous state or sickle cell disease (SS), and the compound heterozygous state for haemoglobin S together with haemoglobins C, D, E, or other structural variants. Several disorders result from the inheritance of the sickle cell gene together with different forms of thalassaemia.

Pathogenesis

Haemoglobin S differs from haemoglobin A by the substitution of valine for glutamic acid at position 6 in the b chain. Although this has been known for nearly half a century, it is still not absolutely clear how it gives rise to the sickling phenomenon. The latter appears to be due to the unusual solubility characteristics of haemoglobin S which undergoes liquid crystal (tactoid) formation as it becomes deoxygenated. In this state, aggregates of sickled haemoglobin molecules arrange themselves in parallel, rod-like fibres, made up of a complex solid core about 21 nm in diameter, composed of 14 filaments arranged as seven pairs of double filaments. Much is now known about the complex interactions whereby the b6 valine substitution stabilizes the molecular stacks in the deoxy configuration of haemoglobin. There is considerable variation in the extent to which different haemoglobins are able to participate with haemoglobin S in the sickling process. This accounts for some of the clinical variability of the different sickling conditions. For example haemoglobin F is almost completely excluded from the sickling process; increasing concentrations in the red cell reduce the rate of sickling.

The pathophysiology of sickling is an extremely dynamic process. Red cells containing sickle haemoglobin at a high concentration endure a series of cycles of sickling and desickling with progressive membrane damage and loss of plasticity. Finally these dry, rigid cells become irreversibly sickled (Fig. 20). Sickling of this type has two main effects. First, sickled erythrocytes have a shortened survival leading to a chronic haemolytic anaemia. Second, and more importantly, these abnormal red cells tend to adhere to the various receptors on the walls of small blood vessels with the production of aggregates, blockage of the vessels, vascular stasis, and, ultimately, tissue damage.

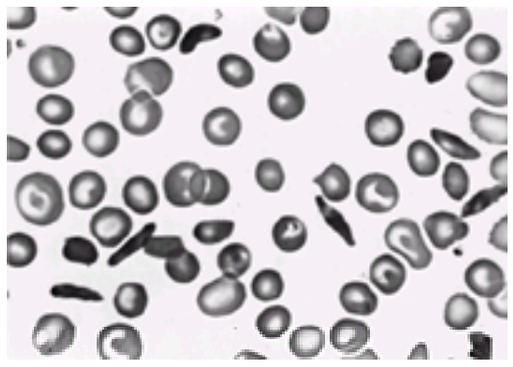

Fig. 20 Irreversibly sickled cells in the peripheral blood (×1000, Leishman stain).

Distribution

The sickling disorders occur very frequently in African populations and, sporadically, throughout the Mediterranean region and the Middle East. There are extensive pockets in India but the disease has not been seen in Southeast Asia. The high frequency of the sickle cell gene occurs because carriers are more resistant than normal individuals to *P. falciparum* malaria.

Clinical features

Except in conditions of extreme hypoxia, such as flying in an unpressurized aircraft the sickle cell trait causes no clinical disability. However, it is possible for individuals to suffer vaso-occlusive episodes if they become oxygen deprived under anaesthesia. Therefore all individuals of the appropriate racial background should have a sickling test (see below) before receiving an anaesthetic. If the test is positive, the anaesthetic should be given with adequate oxygenation and special care should be taken to avoid postoperative dehydration.

Sickle cell anaemia runs an extremely variable clinical course. At one end of the spectrum it is characterized by a crippling haemolytic anaemia interspersed with severe exacerbations, or crises. On the other hand, it may be extremely mild and only found by chance on routine haematological examination. The reason for these remarkable differences in phenotypic expression, which are only partly understood, include the level of haemoglobin F, coinheritance of a thalassaemia, climate, and, probably most important, socioeconomic factors such as availability of early treatment of infection.

Typically, sickle cell anaemia presents in infancy with symptoms related to anaemia or infection. A common presenting symptom is the hand and foot syndrome . It occurs early in infancy and is characterized by a painful dactylitis with swelling of the fingers or feet. Epiphyseal damage during one of these episodes may lead to chronic shortening of a digit. Infants are anaemic from about the third month of life. During early development they often have significant splenomegaly that gradually resolves due to repeated infarction. Indeed, it is most unusual to feel the spleen after the end of the first decade. Typically, the haemoglobin levels are in the 6 to 8 g/dl range with a reticulocyte count of 10 to 20 percent. There is chronic, mild icterus with an elevated bilirubin level. Examination of the peripheral blood film shows anisochromia and poikilocytosis with a variable number of sickled erythrocytes (Fig. 20). As the children grow older the haematological changes of hyposplenism develop with the appearance of pits on the surface of the red cells, Howell–Jolly bodies, and distorted red cells. The white cell and platelet counts are usually normal or slightly elevated.

Growth and development are usually otherwise normal although there may be some skeletal deformities, including frontal bossing of the skull due to expansion of the bone marrow. In some series, children have tended to be short for their age, while postadolescents were usually tall. Inequalities between upper and lower segments, stressed in the early literature, are unusual. The only other physical finding is chronic leg ulceration; this is discussed below.

Complications

The chronic haemolysis of sickle cell disease is interspersed with acute exacerbations of the illness called sickling crises. Furthermore, there are a series of serious and life-threatening, long-term complications which develop in many patients with sickle cell anaemia.

The different forms of sickle cell crises are summarized in Table 7. The commonest is the painful crisis. This is sometimes precipitated by infection, dehydration, or exposure to cold, although quite often no underlying cause can be found. The episode starts with vague pain, often in the back or bones of the limbs. The pain gradually worsens and its bizarre distribution may cause a major diagnostic puzzle. The pain is almost certainly due to blockage of small vessels with sickled erythrocytes; aspiration over areas of bone tenderness has shown infarction of marrow tissue. Occasionally, abdominal pain is the major symptom and this may be associated with distension and rigidity, a picture very similar to an acute abdominal emergency. The diagnostic difficulties in distinguishing between an abdominal crisis and a surgical abdomen are compounded by the fact that the bowel sounds are often diminished during abdominal crises. Two other serious forms of thrombotic crisis are known as the 'chest' and 'brain' syndromes. The 'chest' syndrome, characterized by acute dyspnoea and pleuritic pain together with infiltrates on the chest radiograph, is due to sequestration of sickle cells in the pulmonary circulation. It is sometimes accompanied by a fall in the PCV and platelet count which also may reflect sequestration of sickled cells in the pulmonary vessels. Neurological involvement may present in a variety of ways including fits with or without focal neurological signs. Cerebral infarction is commoner in children, while haemorrhage, due to microaneurysms which develop round infarctions ('moya moya') is commoner in adults.

During painful crises there may be a marked increase in the rate of haemolysis with a fall in the haemoglobin level. Such haemolytic episodes are uncommon. Much more serious are periods of transient bone marrow aplasia called aplastic crises. These seem to result from intercurrent infection, particularly due to parvovirus, and frequently affect more than one sibling in the same family.

Finally, and most serious, are the sequestration crises. Occurring mainly in babies and young children, they are characterized by a rapid enlargement of the spleen or liver, which become engorged with sickled erythrocytes. As the crisis progresses a large proportion of the total red cell mass may be trapped in the spleen or liver. Death may occur due to profound anaemia. These episodes show a tendency to recur in the same individual. Hepatic sequestration, which may occur in adults, is easily overlooked if the liver size is not monitored carefully.

The commonest cause of death in sickle cell anaemia appears to be a sequestration crisis or acute infection, or both. It is not absolutely clear why patients with this disorder are so prone to infection although reduced splenic function may play a role. Abnormalities of the alternate pathway of complement activation have also been described. A variety of organisms are involved, particularly the pneumococcus, and, mainly in tropical countries, typhoid infection of bone infarcts leads to typhoid osteomyelitis. Despite the relative resistance of heterozygotes to *P. falciparum* malaria, deaths due to malaria are extremely common in Africa.

Pregnancy may be uneventful, or associated with an increased incidence of painful crises. There is slightly increased incidence of maternal mortality and a definite increase in the rate of fetal loss.

Chronic complications

The chronic complications of sickle cell anaemia result largely from infarcts following repeated episodes of vascular occlusion. Almost any organ can be involved. Those at particular risk are areas which rely largely on small vessels for their blood supply. The bones are particularly prone to infarction. Aseptic necrosis of the humeral or femoral heads may lead to gross deformity of the shoulder and hip joints (Fig. 21). Bone infarcts may result in chronic sequestra formation which may become secondarily infected with the production of osteomyelitis. Infarction of the bone marrow does not seem to have any long-term sequelae, although occasionally pieces may break off and embolize to the lungs. Chronic leg ulceration is a common problem.

Fig. 21 Aseptic necrosis of the left femoral head in sickle-cell thalassaemia. (Reproduced by courtesy of Dr Graham Serjeant.)

Another organ at particular risk is the kidney. During early childhood renal function may be impaired but this can be corrected by blood transfusion, suggesting that it is due to reversible changes in the renal vasculature. These alterations in renal function are not reversible in later life. Chronic renal failure due to damage of the renal vessels is one of the commonest causes of death in adults with sickle cell anaemia. A typical nephrotic syndrome may develop at some stage during the illness. Pulmonary infarction occurs quite frequently, but repeated episodes leading to severe pulmonary hypertension and right heart failure are unusual, although this complication has been well documented.

There is usually some degree of cardiomegaly. A variety of flow murmurs may be heard but most of these signs seem to be the result of chronic anaemia. Myocardial infarction or fibrosis is not a feature of the disease. Recurrent attacks of painful priapism may lead to permanent deformity of the penis. Occular manifestations are also relatively common in sickle cell anaemia although they tend to be more serious in haemoglobin SC disease; they will be considered with the later disorder in a later section. Finally, there is increased evidence that, unless the neurological crises are treated energetically, permanent brain damage may result.

Course and prognosis

There are still large gaps in our knowledge about the natural history of sickle cell anaemia. Prognosis seems to depend on the racial background of the patient, socioeconomic and ill-defined genetic factors, and, especially the availability of good paediatric care in the early years.

In rural East Africa, the disease still has a high mortality in the first year or two of life. In Jamaica there appears to be a 10 per cent mortality in the early years although survival into adult life and old age is common. This is also the case in some urban parts of Africa and in the United States and Europe. Data from the United States Cooperative Study of Sickle Cell Disease suggest that the median age at death for males is 42 years and for females 48 years. In Saudi Arabia and India, a particularly mild form of the condition occurs. Mortality is extremely low in childhood and a normal survival seems to be common. It is becoming increasingly apparent that the commonest cause of death in the first year or two of life is infection, often associated with splenic sequestration. Later in life infection is still a frequent cause of death, although studies in Jamaica indicate that chronic, progressive renal failure may be responsible for a significant number of deaths. The introduction of prophylactic penicillin has made a major inroad into early deaths from infection (see below).

Laboratory diagnosis

Sickle cell trait causes no haematological changes and is diagnosed by the finding of a positive sickling test together with haemoglobins A and S on electrophoresis (Fig. 22). Sickle cell anaemia is diagnosed by the finding of a variable degree of anaemia, an elevated reticulocyte count, sickled erythrocytes on the peripheral blood film, a positive sickling test, and a haemoglobin electrophoresis pattern characterized by the absence of haemoglobin A and a preponderance of haemoglobin S with a variable amount of haemoglobin F (Fig. 22). The diagnosis is confirmed by finding sickle cell trait in both parents.

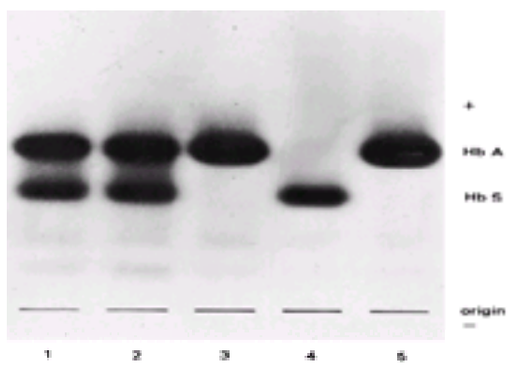

Fig. 22 The haemoglobin pattern in the sickling disorders (starch gel electrophoresis, protein stain, pH 8.5). The following are shown (left to right): (1 and 2) the sickle-cell trait; (3) normal; (4) sickle-cell anaemia; (5) normal.

There is a variety of simple sickling tests available. For ward laboratories the simplest it to take a drop of blood, mix it with two volumes of freshly prepared 2 per cent sodium metabisulphite, place a coverslip over the mixture, seal the edges with vaseline, and examine the slide for sickling after 1 h.

Control and management

There is very little experience of prospective genetic counselling and education of communities as an approach to reducing the number of carriers of sickle cell disease. Although prenatal diagnosis of sickle cell disease can be carried out by DNA analysis following chorion villus sampling, it has not been taken up as extensively as is the case for the thalassaemias. We need to know a great deal more about the factors which modify the clinical prognosis before the place of prenatal diagnosis is clarified.

It is very important that the babies of 'at risk' pregnancies are screened at birth and that the diagnosis is made as early as possible. This is because early deaths due to infection and the frequency of crises may be reduced by the administration of oral penicillin. This should be given to all affected babies at a dosage of 62.5 mg three times a day, up to 1 year of age, 125 mg twice a day from the age of 1 to 3 years, and 250 mg twice a day thereafter. It is also now standard practice for these babies to receive pneumococcal vaccine; in many centres they also receive vaccines against meningococcus and *Haemophilus influenzae*.

Patients with sickle cell anaemia adapt well to their low haemoglobin levels and regular blood transfusion is not required. Particularly in populations in which the diet is low in folate, regular folate supplements should be given. Patients should be given access to a centre that has expertise in the management of this disorder and advised to present at the first sign of a painful crisis. They should also be given a card to carry which states their haemoglobin genotype.

All but the mildest painful crisis should be managed in hospital. Patients should be examined in detail at regular intervals for evidence of underlying infection and given adequate rehydration, oxygen, antibiotics where appropriate, and, in particular, analgesia. The haemoglobin level and reticulocyte count should be estimated at frequent intervals to anticipate an aplastic crisis or pulmonary sequestration episode. While a mild crisis may be managed with first-line analgesics, stronger pain relief is often necessary. There has been concern about the possible dangers of the use of pethedine; it has become fashionable to administer diamorphine by slow, titratrated intravenous infusion. This has to be done under constant surveillance with regular monitoring of respiration and blood gases. It is very unusual for a painful crisis to last more than a few days.

Pulmonary sequestration requires urgent treatment in an intensive care unit. Oxygen should be administered and the blood gases monitored. An exchange transfusion should be initiated unless the haemoglobin level is lower than 4 to 5 g/dl, when the same result can be achieved by rapid transfusion up to 10 to 12 g/dl. Similarly, cerebral complications should be treated by exchange or top-up transfusion. There is evidence that this complication may be prevented by regular Doppler analysis of cerebral blood flow followed, where appropriate, by regular transfusion. Transfusion therapy also prevents recurrence of the cerebral episodes. Hypertransfusion or exchange transfusion should also be used to cover major surgical emergencies or for patients who are having recurrent crises. Occasionally, and most often in young children, the spleen may enlarge to such a degree that secondary hypersplenism develops and splenectomy is required. Splenic sequestration crises require urgent transfusion. Because they tend to recur, they may require splenectomy.

There is no special management required during pregnancy. Occasionally, if the haemoglobin level falls to a value at which symptoms of anaemia occur, or if there are recurrent crises, a regular transfusion regimen should be started to cover pregnancy and delivery.

Ocular manifestations, particularly proliferative retinopathy, require expert ophthalmological treatment. The current place for prophylactic zenon arc or argon laser therapy remains uncertain. Chronic disability due to aseptic necrosis of the femoral head may require hip replacement, although results are often disappointing and this complication requires a great deal more study. Surgical procedures should be undertaken with great caution. It is vital to maintain adequate oxygenation and hydration; limb torniquets should be avoided. Major procedures are best carried out after exchange transfusion. Haeamaturia usually resolves without treatment. Terminal renal failure should be managed as for any other form of renal insufficiency; renal transplantation has been shown to be successful in several studies.

Recurrent priapism is a major problem. Nearly two-thirds of major episodes are preceded by stuttering attacks and therefore it has been suggested that effective therapy at this stage may reduce the risk of sustaining a major attack, with danger of permanent deformity of the penis. Several approaches to management have been suggested although none have been studied in sufficient detail. One approach has been to commence stilboestrol, 5 mg daily, during the stuttering phase. Other forms of treatment at this stage that have been reported to give benefit are the use of opioid analgesics with benzodiazepine or pseudoephedrine hydrochloride. As well as these approaches, the patient should be hydrated, given analgesia, and, possibly, exchange transfusion. Centres with experience of this complication suggest that conservative treatment should be restricted to 24 h at the most. If there is no improvement, surgical correction is recommended, with a cavernosus–spongiosum shunt, a relatively minor procedure that may produce a good cosmetic result.

The management of leg ulcers is unsatisfactory. They may heal with bed rest and debridement but often relapse. Skin grafting does not always give good results and controlled trials have shown that transfusion does not appear to increase the rate of healing.

Experimental forms of treatment have shown some promise. Bone marrow transplantation has been carried out with reasonably good results. However, because of the inherent risks, and the uncertainty of the prognosis, the precise indications are not yet clear. Other studies have been directed at trying to elevate fetal haemoglobin. In a placebo-controlled trial involving adult patients it was found that the administration of hydroxyurea caused a significant reduction in the number of painful crises. This may have resulted from the modest elevation in fetal haemoglobin in response to the drug but other factors such as the reduced white count and increased red cell volume may have played a role. Currently, this drug has been licensed for treatment by the Federal Drug Administration in the United States for adult patients. It has also been used effectively in children but because of its possible leukaemogenic effects its use earlier in life is still restricted to clinical trials.

Other sickling disorders

The other sickling disorders include the interaction of haemoglobin S with haemoglobins C, D, and some of the rarer haemoglobin variants. The interactions with the different forms of b thalassaemia were described above. In many of these conditions, the clinical manifestations are little different from the sickle cell trait, but haemoglobin SC disease and SD disease more closely resemble sickle cell anaemia.

Haemoglobin SC disease

This disease is found in West Africa and less frequently in North Africa. Characterized by a milder anaemia than sickle cell disease, it often goes unrecognized until adult life. It may present with a complication resulting from damage to the microvasculature, probably because of the relatively high haemoglobin level and the combined effects of sickling and red cell rigidity caused by haemoglobin C (see below). Aseptic necrosis of the femoral or humeral heads and unexplained haematuria are the most common complications. Widespread thrombotic episodes, particularly involving the lungs, may occur during intercurrent infection or in pregnancy or the puerperium. Repeated blockage of the retinal vessels may lead to retinitis proliferans, retinal detachment, and permanent blindness.

Haemoglobin SC disease is diagnosed by finding a mild anaemia with splenomegaly and characteristic morphological changes of the red cells, including many target forms, intracellular crystals, and sickle cells. The sickling test is positive and haemoglobin electrophoresis shows haemoglobins S and C in about equal proportions. One parent shows the sickle cell trait and the other the haemoglobin C trait.

Severe thrombotic episodes, particularly in pregnancy, should be treated by exchange transfusion. The role of anticoagulants has never been established. Retinal disease is treated by laser.

Haemolysis due to other common haemoglobin variants

After haemoglobin S the second commonest variant in West Africa is haemoglobin C. Haemoglobin C, because of its relatively low solubility, appears to exist in a precrystalline state in red cells causing their rigidity and premature destruction in the microcirculation. The homozygous state, haemoglobin C disease, is characterized by a mild haemolytic anaemia with splenomegaly, and 100 per cent target cells on the blood film. Haemoglobin analysis shows haemoglobin C with small amounts of haemoglobin F. This is a mild disorder and no specific treatment is required.

The commonest haemoglobin variant throughout Southeast Asia and the Indian subcontinent is haemoglobin E. The homozygous state for this variant, haemoglobin E disease, is characterized by a very mild degree of anaemia with a slight reticulocytosis. The blood film shows mild morphological changes of the red cells which are hypochromic and microcytic, resembling the changes seen in b thalassaemia. No treatment is required.

Haemoglobin variants which migrate in the position of haemoglobin S but which do not sickle have been given the general title of haemoglobin D. There are several different molecular varieties of this variant; the commonest is haemoglobin D Los Angeles. The homozygous state is associated with moderate anaemia, splenomegaly, and a mild degree of haemolysis. The compound heterozygous state with haemoglobin S produces a disorder very similar to sickle cell anaemia. It is diagnosed by finding one parent with the haemoglobin D trait and the other with the sickle cell trait.

The unstable haemoglobin disorders

The unstable haemoglobin disorders are a rare group of inherited haemolytic anaemias which result from structural changes in the haemoglobin molecule that cause intracellular precipitation with the formation of Heinz bodies. Their true incidence is not known. There have been several well-documented families in which patients with one of these haemoglobin variants have had no affected relatives, suggesting that the condition has arisen by a new mutation.

Aetiology and pathogenesis

Most of the unstable haemoglobin variants result from single amino acid substitutions at critical areas of the molecule. For example substitutions in or around the haem pocket can disrupt the normal anatomy and allow in water with subsequent oxidative damage to haem which leads to precipitation of the haemoglobin. Some substitutions, such as those involving proline residues, cause a marked disruption of the secondary structure of a globin chain. A few of these variants result from deletions of either single or several amino acid residues. For example in haemoglobin Gun Hill, five amino acids are missing including the haem binding site. As the unstable haemoglobins precipitate in the red cells or their precursors, they produce intracellular inclusions, or Heinz bodies, which make the cells more rigid causing their premature destruction in the microcirculation (Fig. 23). The degradation products of the precipitated haemoglobin, notably haem and iron, cause oxidative damage to the red cell membrane proteins in much the same way as the excess a and b chains produced in the thalassaemias.

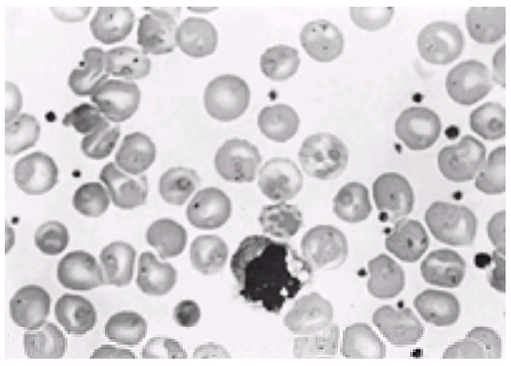

Fig. 23 The peripheral blood film of a patient with an unstable haemoglobin disorder, haemoglobin Hammersmith. This is a postsplenectomy film, which shows small inclusions in many of the red cells (×1000, Leishman stain).

Clinical features

All these conditions are characterized by a haemolytic anaemia of varying severity and splenomegaly. There may be a history of the passage of dark urine, particularly during episodes of infection. Like all chronic haemolytic anaemias, there is an increased incidence of pigment gallstones. The condition may become worse during periods of intercurrent infection. In the more severe forms, such episodes are associated with life-threatening anaemia. Patients with unstable haemoglobins are at particular risk of haemolytic episodes following the administration of oxidant drugs. Apart from intermittent icterus and splenomegaly there are no characteristic physical findings.

Laboratory diagnosis

This condition should be thought of in any familial haemolytic anaemia, particularly if a red cell enzyme deficiency cannot be demonstrated. The peripheral blood film shows the features of haemolysis but the red cell morphology may be relatively normal. Occasionally there is a mild degree of hypochromia and microcytosis. Unless splenectomy has been carried out, Heinz bodies are not seen in the peripheral blood (Fig. 23).

The most characteristic feature of the unstable haemoglobins is their heat instability. If a dilute haemoglobin solution is heated at 50°C for 15 min, most of the unstable haemoglobins precipitate as a dense cloud. A similar phenomenon can be induced by isopropanalol. Some variants can be identified by haemoglobin electrophoresis but others, because they result from the substitution of a neutral amino acid, produce no electrophoretic changes and can only be demonstrated by the heat precipitation test.

Treatment

Because these conditions are so rare there has been very little experience of the effects of splenectomy. From the information that is available, and from the author's personal experience, it appears that if a child has had several life-threatening episodes of anaemia or is running a steady-state haemoglobin level which is impairing development or well-being, splenectomy should be undertaken. It is interesting to note that some of these haemoglobin variants produce a 'right shift' in the oxygen dissociation curve, and a measurement of the P_{50} as part of the presplenectomy assessment may help to decide whether to proceed to surgery; a marked right shift, that is an increased P_{50}, indicates that the anaemia should be more easily tolerated than if the oxygen dissociation curve is moved in the opposite direction with a low P_{50}. An accurate history from the child or its parents is probably more helpful, however.

Haemoglobin variants which cause abnormal oxygen binding

In 1966, an 81-year-old man presented at Johns Hopkins Hospital, Baltimore with mild angina and a haemoglobin value of 19.9 g/dl. No cause could be found for his polycythaemia but it was noted that he had an abnormal haemoglobin. The oxygen dissociation curve of his blood was found to be displaced to the left. This suggested that the abnormal haemoglobin might have a high oxygen affinity and that the patient's increased red cell count might be compensating for a primary defect in oxygen unloading. Further studies showed that this was the case, documenting a new cause for secondary polycythaemia. Since then over 40 haemoglobin variants of this type have been defined, all associated with familial polycythaemia.

Aetiology

The high-oxygen-affinity haemoglobin variants result from single amino acid substitutions at critical parts of the haemoglobin molecule which are involved in the configuration changes that underlie haem/haem interaction and the production of a sigmoid oxygen dissociation curve. Many occur at the junctions between the a and b subunits. Others involve the amino acids which are involved with the binding of 2,3-diphosphoglycerate (2,3-DPG) to haemoglobin. As mentioned earlier, increasing concentrations of 2,3-DPG tend to push the oxygen dissociation curve to the right; fetal haemoglobin has a high oxygen affinity (left-shifted curve) because it cannot interact with 2,3-DPG; mutations of the DPG binding sites have a similar effect.

Pathophysiology

The high-oxygen-affinity variants have a left-shifted oxygen dissociation curve with a reduced P_{50}. Thus the variant haemoglobin holds on to oxygen more avidly than normal haemoglobin. This leads to tissue hypoxia. This in turn causes an increased output of erythropoietin and an elevated red cell mass.

Clinical features

Many patients with high-oxygen-affinity variants are completely healthy and are only found to carry the variant when a routine haematological examination shows an unusually high haemoglobin level or packed cell volume. There have been one or two reports of arterial or venous occlusive disease in these patients. However, this is uncommon. Most patients are asymptomatic. There is no splenomegaly and no other associated haematological findings. Although it might be expected that a high-oxygen-affinity haemoglobin would cause defective oxygenation of the fetus none of the reported families has had a history of frequent stillbirths.

Diagnosis

The condition should be suspected in any patient with a pure red cell polycythaemia associated with a left-shifted oxygen dissociation curve. The diagnosis can be confirmed by haemoglobin analysis.

Treatment

In asymptomatic patients with high-oxygen-affinity haemoglobin variants no treatment is necessary. The difficulty arises if the patient has associated vascular disease with symptoms of coronary or cerebral artery insufficiency. There is insufficient published information to make any dogmatic statements about how this complication should be managed. The author has seen several patients of this type who seem to have responded to venesection; more experience is required before this form of treatment can be recommended however. These patients require a high haemoglobin level for oxygen transport; half their haemoglobin is physiologically useless.

Low-oxygen-affinity variants

At least six haemoglobin variants with reduced oxygen affinity have been reported. The first to be described, haemoglobin Kansas, was found in a mother and son with unexplained cyanosis. The subjects were asymptomatic and had normal haemoglobin levels without any evidence of haemolysis. Like many of the high affinity variants, the amino acid substitution in this variant was at the interface between the a and b globin chains. For reasons which are not clear, some substitutions in this

region give rise to variants with a relatively low oxygen affinity. This condition should be thought of in any patient with an unexplained congenital cyanosis; the differential diagnosis is considered below.

Methaemoglobinaemia, carboxyhaemoglobinaemia, and sulphaemoglobinaemia

Methaemoglobinaemia is a condition characterized by increased quantities of haemoglobin in which the iron of haem is oxidized to the ferric (Fe^{3+}) form. Carboxyhaemoglobinaemia (carbonmonoxyhaemoglobinaemia) results from the binding of carbon monoxide to the haem molecules. Sulphaemoglobinaemia is a rare condition in which there is a mixture of haemoglobin derivatives whose structure is poorly characterized but which can be defined by their specific spectral characteristics.

Pathogenesis

As mentioned earlier, each haemoglobin molecule has four haem moieties. At first sight it is not clear why the oxidation of a proportion of the iron atoms, or the fact that they are liganded to carbon monoxide, should cause such profound changes in oxygen transport. However, oxidation of 30 per cent of the haem molecules has a much more serious effect on tissue oxygenation than a reduction of the haemoglobin level by the same amount. This is because, if a single haem is oxidized, it so alters the conformation of the haemoglobin molecule that the oxygen affinity of the other three haems is increased. Thus methaemoglobin, carboxyhaemoglobin, and cyanmethaemoglobin all have very high oxygen affinities with 'left shifted' oxygen dissociation curves, and hence are associated with impaired unloading of oxygen to the tissues.

Methaemoglobinaemia

Methaemoglobin causes a variable degree of cyanosis. It should be suspected in any patient with significant central cyanosis in whom there is no evidence of cardiorespiratory disease. The degree of cyanosis produced by 5 g/dl of deoxygenated haemoglobin can be produced by 1.5 g/dl methaemoglobin and 0.5 g/dl of sulphaemoglobin. Methaemoglobin concentrations of 10 to 20 per cent are tolerated quite well. It is useless as an oxygen carrier; levels above this are thus often associated with dyspnoea and headache. Much depends on the rapidity at which it is formed; many patients with life-long methaemoglobinaemia are asymptomatic while individuals who have accumulated a similar level of methaemoglobin acutely may be acutely dyspnoeic. For reasons which are not clear, it is unusual for patients with chronic methaemoglobinaemia to have an increased haemoglobin level or red cell count.

Methaemoglobinaemia may arise as a result of a genetic defect in red cell metabolism or haemoglobin structure, or may be acquired following the ingestion of various oxidant drugs and toxic agents.

Genetic methaemoglobinaemia

There are two forms of inherited methaemoglobinaemia. The first results from a deficiency of red cell NADH-diaphorase, the second from a structural alteration in either the a or b globin chains of haemoglobin.

NADH-diaphorase catalyses a step in the major pathway for methaemoglobin reduction. The enzyme reduces cytochrome b_5 using NADH as a hydrogen donor. The reduced cytochrome b_5 reduces, in turn, methaemoglobin to haemoglobin. There are several different molecular forms of NADH-diaphorase deficiency which have been identified by electrophoretic analysis of NADH-diaphorase in the red cells of affected patients. The condition is inherited as an autosomal recessive. Homozygotes have elevated levels of methaemoglobin and are cyanotic from birth. Heterozygotes do not have elevated levels of methaemoglobin but seem to be unusually susceptible to the oxidant action of drugs. For example severe cyanosis has been precipitated by the use of antimalarial drugs.

There are several abnormal haemoglobin variants which are associated with genetic methaemoglobinaemia, all of which are designated haemoglobin M, and further identified by their place of discovery, for example haemoglobin M Boston, M Milwaukee. These variants usually result from amino acid substitutions near the haem pocket. Normally, haem lies between two histidine residues, one called the proximal histidine to which it is attached, and the other called the distal histidine. Oxygen is bound to haem at a site opposite to the distal histidine. If the latter is substituted by tyrosine, as occurs in the a chain variant haemoglobin M Boston and in the b chain variant M Saskatoon, a stable bond is formed between the haem iron and the phenolic ring of the tyrosine. The iron atom is 'fixed' in the Fe^{3+} state. These variants are associated with cyanosis which is present from early life. In the case of the a chain variants it is present from birth, while the b chain haemoglobin variants only produce cyanosis after the first few months of life as adult haemoglobin synthesis becomes established. Unlike NADH diaphorase deficiency, which is inherited as a recessive trait, the haemoglobin Ms have a dominant form of inheritance. Thus it is very simple to make the diagnosis of genetic methaemoglobinaemia and to determine the likely molecular basis by taking a good history; even the affected globin chain can be ascertained!

The diagnosis is confirmed by spectroscopic examination of the blood and by determination of methaemoglobin levels. The precise cause can be established by an assay of NADH-diaphorase or by haemoglobin analysis under appropriate conditions.

Genetic methaemoglobinaemia due to NADH-diaphorase deficiency is readily treated by the administration of ascorbic acid, 300 to 600 mg daily by mouth in divided doses, or by the administration of methylene blue, either intravenously (1 mg/kg body weight) or by mouth 60 mg three to four times daily. On the other hand, the genetic methaemoglobinaemias due to structural haemoglobin variants do not respond to ascorbic acid, methylene blue, or any other treatment. Most affected individuals go through life asymptomatic and require no treatment.

Acquired methaemoglobinaemia

Acquired methaemoglobinaemia usually results from the administration of drugs or exposure to chemicals which cause oxidation of haemoglobin. There are many agents which are capable of exceeding the red cells' ability to reduce methaemoglobin. They include ferricyanide, bivalent copper, chromate, chlorate, quinones, and certain dyes with a high oxidation–reduction potential. Nitrite, often used as a preservative, is one of the most common methaemoglobin-forming agents. Nitrates, after conversion to nitrites in the gut, may cause serious methaemoglobinaemia in infants. Other agents which commonly cause methaemoglobinaemia include phenacatin, primaquine, sulfonamides, and various analine dye derivatives.

If any of the agents listed above is given in low dose over a long period of time it may lead to chronic methaemoglobinaemia with or without a haemolytic anaemia. However, after exposure to a large amount of these agents, and the development of in excess of 50 to 60 per cent methaemoglobin, the symptoms of acute anaemia develop because methaemoglobin lacks the capacity to transport oxygen. Thus the clinical picture may be characterized by vascular collapse, coma, and death.

Methaemoglobinaemia with haemolytic anaemia

The haemolytic action of oxidant drugs is described elsewhere. Chronic methaemoglobinaemia with haemolytic anaemia, characterized by Heinz body formation and fragmented red cells, occurs commonly in patients receiving dapsone, salazopyrine, or phenacatin. This condition is usually innocuous and can be modified by adjusting the dose of the drug.

Occasionally, acute intravascular haemolysis associated with methaemoglobinaemia and intravascular coagulation occurs. It usually follows the ingestion or infusion of a strong oxidizing agent such as chlorate or arsine. There is gross intravascular haemolysis and methaemoglobinaemia together with evidence of disseminated intravascular coagulation. The haemoglobin level may fall very rapidly and may be complicated by renal failure.

Treatment

In cases of chronic acquired methaemoglobinaemia, the drug or chemical agent should be removed where possible. If continued therapy is required, it should be administered at a lower dose.

Acute toxic methaemoglobinaemia presents a serious medical emergency. Methylene blue should be administered in a dose of 1 to 2 mg/kg intravenously over a 5-min period. Repeated doses may be needed. Toxicity is uncommon although doses of over 15 mg/kg may cause haemolysis in young infants. The drug should not be used if the methaemoglobinaemia is due to chlorate poisoning as it may convert the chlorate to hypochlorite which is an even more toxic compound. In cases of

acute methaemoglobinaemia with intravascular haemolysis, haemodialysis with exchange transfusion is the treatment of choice.

Carboxyhaemoglobinaemia

Carbon monoxide (CO) has an affinity for haemoglobin approximately 210 times that of oxygen. Following acute exposure it is so tightly bound that it takes about 4 h for an individual with normal ventilation to expel half of it. At levels of 5 to 10 per cent there may be no symptoms, but above 20 per cent there is usually headache and weakness. Levels of 40 to 60 per cent or more lead to unconsciousness and death.

Carbon monoxide poisoning is discussed in Chapter 8.1 and secondary polycythaemia due to chronic exposure is considered elsewhere in this chapter.

Sulphaemoglobinaemia

This poorly-defined condition derives its name from the fact that it can be produced *in vitro* by the action of hydrogen sulphide on haemoglobin. It has not been reported as a genetic disorder. It is usually associated with the administration of drugs, particularly sulfonamides or phenacetin. It has also been reported in patients with chronic constipation or malabsorption syndromes (enterogenous cyanosis) although its relationship to these disorders is far from clear.

Other acquired abnormalities of the structure or synthesis of haemoglobin

Glycosylated haemoglobin, haemoglobin AIc

Haemoglobin may undergo post-translational modification in patients with diabetes. The abnormal haemoglobin, haemoglobin AIc, is formed by the non-enzymic combination of glucose with the N-terminus of the b chain, forming first a Schiff base which then undergoes a rearrangement to form a stable ketoamine. The level of haemoglobin AIc is raised in diabetics and is related to the blood sugar level over the previous weeks. The value of the estimation of haemoglobin AIc as an index of the control of diabetes is considered elsewhere.

Haemoglobin Pb

Some children with lead poisoning develop a modified haemoglobin which migrates rapidly on alkaline electrophoresis. The precise structural alteration is not known but, if present, this variant is a useful indicator of severe lead poisoning.

Fetal haemoglobin production in adult life

A number of haematological disorders are associated with a reversion to fetal haemoglobin production after the neonatal period. These include juvenile myeloid leukaemia, other forms of leukaemia, and congenital hypoplastic anaemias. Haemoglobin F may also appear transitorily during rapid regeneration of the bone marrow after drug induced hypoplasia, virus infection, or bone marrow transplantation.

Further reading

Ballas SK (1998). Sickle cell disease: clinical management. *Clinical Haematology* **11**, 185–214.

Bunn HF (1997). Pathogenesis and treatment of sickle cell disease. *New England Journal of Medicine* **337**, 762–9.

Cao A, Galanello R, Rosatelli MC (1998). Prenatal diagnosis and screening of the haemoglobinopathies. *Clinical Haematology* **11**, 215–38.

Dover GJ, Platt OS (1998). Sickle cell disease. In: Nathan DG, Orkin SH, eds. *Hematology in infancy and childhood*, pp. 762–801. WB Saunders, Philadelphia.

Forget BG, Higgs DR, Nagel RL, Steinberg MH, eds (2001). *Disorders of hemoglobin*. Cambridge University Press, New York.

Higgs DR (1993). a-thalassaemia. In: Higgs DR, Weatherall DJ, eds. *Baillière's clinical haematology. International practice and research: the haemoglobinopathies*, pp. 117–50. Baillière Tindall, London.

Higgs DR, Sharpe JA, Wood WG (1998). Understanding a globin gene expression: a step towards effective gene therapy. *Seminars in Hematology* **35**, 93–104.

Olivieri N (1998). Thalassaemia: clinical management. *Clinical Haematology* **11**, 147–62.

Weatherall DJ, Clegg JB (2001). *The thalassaemia syndromes*, 4th edn. Blackwell Science, Oxford.

Weatherall DJ, Clegg JB, Higgs DR, Wood WG (2001). The haemoglobinopathies. In: Scriver CR, Beaudet AL, Sly WS, Valle D, eds. *The metabolic basis of inherited disease*, 8th edn, pp. 3417–84. McGraw-Hill, New York.

22.5.8 Anaemias resulting from defective red cell maturation

James S. Wiley

Erythroid cell maturation is specialized towards the co-ordinated synthesis of large amounts of haem and globin necessary to attain the high concentration of haemoglobin found in the mature red cell. Hereditary or acquired defects in the production of either of these cause a maturation block, which leads to ineffective erythropoiesis in which many of the developing nucleated erythroblasts are destroyed in the marrow before they can reach the circulation. Thus in thalassaemia, defective synthesis of either a or b globin leads to unbalanced production of the other chain which precipitates and leads to destruction of the precursor erythroblast. Defective haem synthesis in the sideroblastic anaemias also leads to an anaemia which is characterized by ineffective erythropoiesis (Table 1). Abnormalities of DNA synthesis in the developing erythroid cells, such as produced by vitamin B_{12} or folic acid deficiency, blocks cell division required for erythroid maturation and produces morphological and biochemical evidence of ineffective erythropoiesis.

Ineffective erythropoiesis may be recognized by the characteristic erythroid hyperplasia of the bone marrow with normal or only slight increase in reticulocyte numbers. Some other features of ineffective erythropoiesis may be variably present: a mild increase in bilirubin, decrease in haptoglobin, and increased serum lactic dehydrogenase activity. As a result, iron absorption is increased, serum iron and ferritin become elevated, and, after many years, iron overload develops which is indistinguishable from idiopathic haemochromatosis. However, the degree of iron overload does not depend on either the severity of the anaemia or the presence of the characteristic mutation (Cys282Tyr) of the *HFE* gene associated with genetic haemochromatosis.

The sideroblastic anaemias

Sideroblastic anaemias are a group of hereditary or acquired anaemias of varying severity diagnosed by the finding of ring sideroblasts in the bone marrow aspirate. The peripheral blood film shows hypochromic red cells which are microcytic in the hereditary form (Fig. 1), but are often macrocytic in the acquired forms of the disease. Normochromic and normocytic red cells are also present which gives the film a dimorphic distribution of red cell sizes. The diagnostic procedure is bone marrow aspirate followed by staining of the smear with Prussian Blue iron reagent. Ring sideroblasts are diagnostic (Fig. 2) and are defined as erythroblasts containing iron-positive granules arranged in a perinuclear location around one-third or more of the nucleus. Electron microscopy reveals that the iron-containing granules are mitochondria containing precipitated ferric phosphate and ferric hydroxide. The sideroblastic anaemias have diverse aetiologies (Table 2) but have in common an impaired biosynthesis of haem in the erythroid cells of the marrow. Most sideroblastic anaemias are acquired as a clonal disorder of erythropoiesis, with varying degrees of myelodysplasia. The hereditary forms are uncommon. Most are found in males with an X-linked pattern of inheritance. A number of drugs have been associated with reversible sideroblastic anaemia, chiefly in patients with alcohol abuse (Table 2).

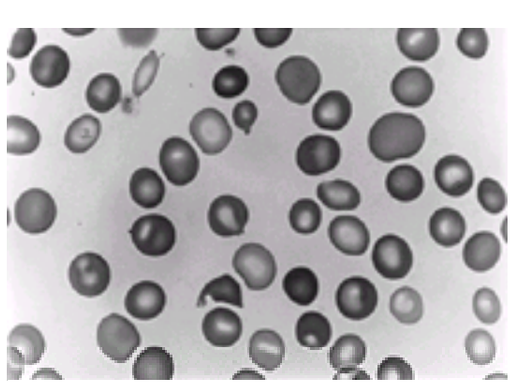

Fig. 1 Peripheral blood smear in hereditary sideroblastic anaemia showing a population of hypochromic and microcytic erythrocytes. (By courtesy of Gillian Rozenberg.)

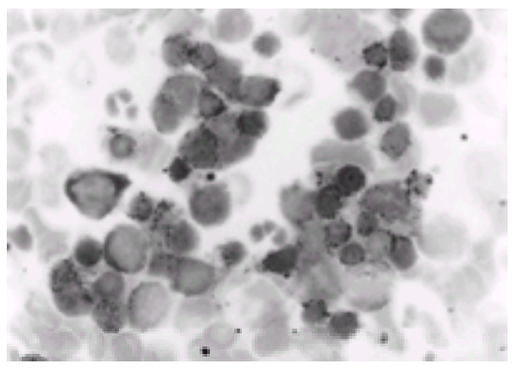

Fig. 2 Bone marrow smear stained with Prussian blue, showing the ring sideroblasts (arrow). (By courtesy of Gillian Rozenberg.)

Hereditary sideroblastic anaemias

Aetiology and pathogenesis

Haem biosynthesis occurs by a cascade of eight enzymes (Fig. 3). In man, mutations affecting the first enzyme of this pathway produce hereditary sideroblastic anaemia. Inborn errors that occur in later enzymes in this pathway result in metabolic disorders known as the porphyrias (Fig. 3). The pathway begins with the condensation of glycine with succinyl CoA to form 5-aminolaevulinic acid (ALA), a step which is under the control of the mitochondrial enzyme ALA synthase. This enzyme requires pyridoxal phosphate as a cofactor. Two isoenzymes of ALA synthase have been identified. One is found in liver and other tissues (ALAS 1). The second is confined to erythroid cells of the bone marrow (ALAS 2). The gene for the erythroid-specific ALAS 2 isoenzyme resides on the X chromosome and is now known to be the site of most mutations giving rise to X-linked hereditary sideroblastic anaemia. Several dozen different mutations have been described in different families. All result from a single amino acid alteration arising from a point mutation in the ALAS 2 coding region of DNA. In nearly half the families with hereditary sideroblastic anaemia, the structure of the ALAS 2 gene is normal, suggesting that other defects may be involved, such as the import of ALAS 2 from the cytosol or its anchoring within the mitochondrial matrix. In most families, males are affected with an X-linked pattern of inheritance consistent with a mutation on the X chromosome (Fig. 4). However, occasionally the disease is transmitted as an autosomal dominant; there are even well-documented families in which only females are affected.

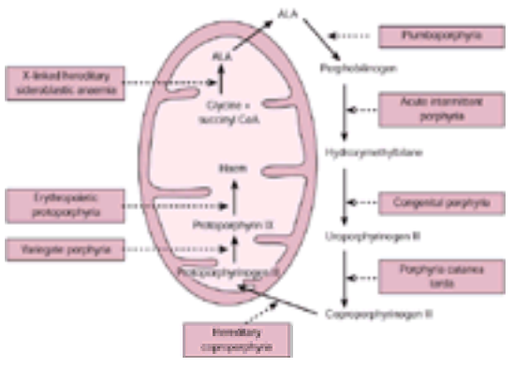

Fig. 3 Pathway of haem biosynthesis in mammalian cells. The first step in the pathway is catalysed by ALAS and occurs within the mitochondrion using pyridoxal 5'-phosphate as a cofactor. ALA then leaves the mitochondrion and is converted by ALA dehydratase to give a monopyrrole, porphobilinogen. Four molecules of this are converted by porphobilinogen deaminase to a linear tetrapyrrole, hydroxymethylbilane. This molecule is then cyclized by uroporphyrinogen III synthase to uroporphyrinogen III, which is then decarboxylated to coproporphyrinogen III. This molecule enters the mitochondrion and is oxidized in succession by coproporphyrinogen III oxidase and protoporphyrinogen III oxidase. The product is protoporphyrin IX, a substrate for ferrochelatase, which catalyses the insertion of Fe^{2+} to form haem. The defective steps associated with specific porphyrias and X-linked hereditary sideroblastic anaemias are shown. (Reproduced from Hoffman R *et al.*, eds. (1999). *Hematology: basic principles and practice*, 3rd edn. WB Saunders Co., Philadelphia, with permission.)

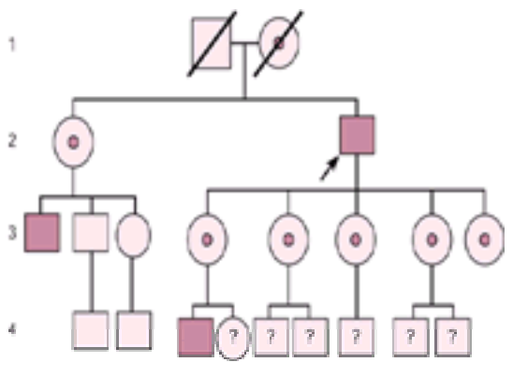

Fig. 4 Pedigree of a family with pyridoxine-responsive sideroblastic anaemia showing X-linked recessive inheritance. n affected; carrier; ? unknown status. Diagonal lines indicate deceased members. The pedigree has been abbreviated to show only the affected branches of the family. The arrow indicates the proband. (Copyright 1994 Massachusetts Medical Society. All rights reserved. Reproduced with permission.)

Clinical and laboratory features

Typically the anaemia presents in infancy or childhood but when the anaemia is mild, the diagnosis may not be made until adult life. Occasionally, such patients may present with features of iron overload such as diabetes or cardiac failure. Others may be found in family surveys, which should be undertaken when this anaemia is diagnosed. Slight enlargement of the liver or spleen may occur. The degree of anaemia is variable, ranging from severe (less than 80 g/l haemoglobin) to mild (more than 100 g/l haemoglobin) but even with mild or no anaemia the mean corpuscular volume (MCV) is below the normal range. Blood film shows a population of cells with hypochromic, microcytic morphology. Female carriers may show the characteristic red cell dimorphism. White cell counts are normal while platelet counts are normal or slightly elevated. Serum iron and ferritin concentrations are invariably increased and transferrin shows an increased percentage saturation with iron. The differential diagnosis includes idiopathic haemochromatosis, since both diseases have evidence of iron overload. Examination of the blood film, the MCV, and the bone marrow should establish the diagnosis.

Treatment and prognosis

A trial of pyridoxine, 100 to 200 mg/day taken orally, is indicated for 3 months in all patients with proven or suspected hereditary sideroblastic anaemia. About 25 per cent of patients experience a full or partial correction. This vitamin should be continued lifelong in responders but at a lower maintenance dosage. Regular transfusions of packed red cells are the mainstay of treatment of severe anaemia. These should be given to relieve symptoms and allow normal childhood development. Splenectomy is contraindicated in this condition. Iron overload progresses rapidly once transfusions begin. Chelation therapy with desferrioxamine should thus be commenced after the first 10 to 20 transfusions. Iron removal may greatly benefit patients with mild or moderate anaemia and evidence of iron overload. Intermittent phlebotomy of 100 to 200 ml blood should be attempted as this is more effective than chelation therapy in removing iron and should be continued if symptoms allow until the serum ferritin becomes normal. Finally, patients should avoid alcohol and ascorbic acid supplements, both of which enhance iron absorption.

Acquired idiopathic sideroblastic anaemia (refractory anaemia with ring sideroblasts)

Acquired idiopathic sideroblastic anaemia is a refractory anaemia with a hypercellular marrow containing ring sideroblasts which may either be idiopathic or develop following chemotherapy or irradiation (Table 2). Since nearly all cases also show evidence of dyserythropoiesis, this anaemia is now classified as one of the myelodysplastic syndromes and termed refractory anaemia with ring sideroblasts by the French–American–British Group. This classification is supported by the demonstration that the defective haematopoiesis is clonal, both in acquired idiopathic sideroblastic anaemia and the other myelodysplastic syndromes. Clonal haematopoiesis has also been shown in acute erythroleukaemia, in which bizarre dysplastic changes and ineffective erythropoiesis are seen in the developing erythroblasts. These comprise a majority (more than 50 per cent) of all nucleated marrow cells. The fact that more than 20 per cent of the myeloid cells are blasts distinguishes acute erythroleukaemia from one of the myelodysplastic syndromes.

Aetiology and pathogenesis

The cause of the defective haem synthesis in acquired sideroblastic anaemia is unclear. Recent reports indicate that levels of ALAS in bone marrow are normal. Indirect evidence points to an acquired defect in the mitochondrial respiratory chain that impairs the reduction of Fe^{3+} to Fe^{2+} since ferrous iron is essential for the terminal ferrochelatase reaction (Fig. 3). Clonal haematopoiesis has been demonstrated in this anaemia by both molecular and karyotypic analysis. Thus a single glucose-6-phosphate dehydrogenase (G6PD) isoenzyme was found in erythrocytes of a woman heterozygous for G6PD who expressed two isoenzymes in her somatic tissues. Clonal chromosome changes are also found in bone marrow cells in many patients with acquired sideroblastic anaemia. Characteristic changes include monosomy 7, trisomy 8, deletions involving chromosomes 5, 7, 11, 13, or 20, and a number of translocations. When sideroblastic anaemia is acquired secondary to chemotherapy or irradiation, chromosomal changes are nearly always found and tend to be multiple. However, they are probably a late event in the course of this anaemia and may be preceded by the expansion of a clone of genetically unstable stem cells.

Clinical and laboratory features

Acquired sideroblastic anaemia typically has an insidious onset. Most patients are middle aged or elderly, but young adults can be affected. Mild splenomegaly may be present. White cell and platelet counts are usually normal; some patients may have thrombocytosis. The bone marrow shows erythroid hyperplasia with varying degrees of dyserythropoiesis, including irregular nuclear contour, nuclear fragmentation (karyorrhexis), bi- or trilobed nuclei, and internuclear bridges. Iron stain of the aspirate shows ring sideroblasts which should total more than 15 per cent of the nucleated erythroid cells to make the diagnosis. Dysplasia of myeloid precursors or megakaryocytes may be present. When associated with leucopenia and/or thrombocytopenia the more descriptive term 'refractory cytopenia' is sometimes used. If the overall blast count exceeds 5 per cent or the peripheral blood monocyte count exceeds 1.0×10^{9}/l, the condition falls within a different category of the myelodysplastic

syndrome. Thus, ring sideroblasts may be seen in other myelodysplastic conditions such as refractory anaemia with excess blasts. Distinguishing acquired idiopathic sideroblastic anaemia from a mild hereditary sideroblastic anaemia presenting in adult life can be difficult. However, careful examination of the marrow for dysplastic changes, the MCV, possible response to pyridoxine, and a family survey all help to distinguish these two entities.

Treatment and prognosis

Transfusions of packed red cells should be given for relief of symptomatic anaemia. A trial of pyridoxine, 100 to 200 mg/day for 3 months, is worthwhile but few patients respond to this vitamin. Acquired idiopathic sideroblastic anaemia and the closely related refractory anaemia have the most favourable outlook among the myelodysplastic syndromes, with a median survival of 42 to 76 months and a 3 to 12 per cent incidence of progression to acute leukaemia. A simple prognostic scoring system has been developed in which two or more of the following place the patient in a poor prognostic category:

1. haemoglobin less than 100 g/l;
2. neutrophils less than 1.5×10^9/l;
3. platelets less than 100×10^9/l; and
4. blasts more than 5 per cent.

Karyotypic analysis of marrow aspirates is valuable since a normal karyotype confers a favourable prognosis. The karyotype is now included in an international prognostic scoring system.

Defective red cell maturation secondary to alcohol and drugs

Alcohol has a direct toxic effect on erythropoiesis, manifested by the macrocytosis which characterizes red cells of subjects chronically ingesting alcohol in excess. Malnourished and anaemic alcoholics may exhibit ring sideroblasts in the bone marrow as well as vacuolation of erythroblasts. These manifestations gradually disappear over 4 to 12 days when alcohol is withdrawn, although the macrocytosis may take several months to normalize. The antibiotic chloramphenicol when given in dosages greater than 2 g/day produces a reversible inhibition of erythropoiesis associated with ring sideroblasts and vacuolation of erythroblasts. This effect, due to inhibition of mitochondrial protein synthesis, is quite separate from the rare idiosyncratic side-effect of aplastic anaemia. Protracted exposure to the antituberculous drug isoniazid has been occasionally associated with development of a sideroblastic anaemia.

Defective red cell maturation secondary to lead, arsenic, or zinc ingestion or copper deficiency

Patients suffering lead poisoning show clinical and laboratory evidence of reduced haem biosynthesis. Basophilic stippling of red cells is prominent. Mild hypochromic, microcytic anaemia may develop. Red cell protoporphyrin, increased due to inhibition of the terminal step in the haem pathway, provides a sensitive measure of lead exposure. The peripheral neuropathy of lead poisoning may be a result of reduced haem biosynthesis, as in the porphyrias. Acute or chronic arsenic ingestion can cause anaemia with marked dyserythropoiesis. Basophilic stippling of red cells is characteristic while neutropenia and thrombocytopenia may be present. Copper deficiency has been described only in malnourished premature infants or in patients receiving long-term, parenteral hyperalimentation. This syndrome consists of anaemia and neutropenia associated with marrow findings of ring sideroblasts and vacuolated erythroid and myeloid precursors. Large quantities of ingested zinc interfere with copper absorption and reproduce the sideroblastic anaemia and neutropenia characteristic of copper deficiency.

Congenital dyserythropoietic anaemias (CDA)

This rare group of inherited refractory anaemias are characterized by gross multinuclearity of erythroid precursors in the marrow, ineffective erythropoiesis, and associated iron overload. Three types have been described based on morphology of the bone marrow and serological features. The most common, Type II, is also known as HEMPAS (hereditary erythroblast multinuclearity with positive acidified serum test) since red cells are lysed by acidified (pH 6.8) serum from about 30 per cent of normal subjects. In CDA Type II, a defect in glycosylation of erythroblast membrane proteins has been identified. Most patients are diagnosed in late childhood or adolescence with mild to moderate anaemia, with intermittent jaundice or in older patients with manifestations of iron overload. Splenomegaly or hepatomegaly may be variably present. CDA carries a good prognosis with few patients requiring transfusions. The degree of iron overload should be monitored and treated when appropriate.

Further reading

Bennett JM *et al*. (1982). Proposals for the classification of the myelodysplastic syndromes. *British Journal of Haematology* **51**, 189–99.

Bottomley SS *et al*. (1999). Sideroblastic anaemia. In: Lee GR *et al*., eds. *Wintrobes clinical haematology*, pp. 832–71. Williams and Wilkins, Baltimore.

Cazzola M *et al*. (1988). Natural history of idiopathic refractory sideroblastic anemia. *Blood* **71**, 305–12.

Cotter PD *et al*. (1995). Late-onset X-linked sideroblastic anemia. Missense mutations in the erythroid †-aminolevulinate synthase (ALAS2) gene in two pyridoxine-responsive patients initially diagnosed with acquired refractory anemia and ringed sideroblasts. *Journal of Clinical Investigation* **96**, 2090–6.

Cox TC *et al*. (1994). X-linked pyridoxine-responsive sideroblastic anemia due to a THR ^{388}- to –SER substitution in erythroid 5-aminolevulinate synthase. *New England Journal of Medicine* **330**, 675–9. Shows a typical response of hereditary sideroblastic anaemia to pyridoxine.

Gattermann N *et al*. (1997). Heteroplasmic point mutations of mitochondrial DNA affecting subunit 1 of cytochrome c oxidase in two patients with acquired idiopathic sideroblastic anemia. *Blood* **90**, 4961–72.

Greenberg P *et al*. (1997). International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood* **89**, 2079–88.

Juneja SK *et al*. (1983). Prevalence and distribution of ringed sideroblasts in primary myelodysplastic syndromes. *Journal Clinical Pathology* **36**, 566–9.

Kibbelaar RE *et al*. (1992). Combined immunophenotyping and DNA *in situ* hybidization to study lineage involvement in patients with myelodysplastic syndromes. *Blood* **79**, 1823–8.

Marks PW, Mitus AJ (1996). Congenital dyserythropoietic anaemias. *American Journal of Hematology* **51**, 55–63. Recent review.

Mufti GJ *et al*. (1985). Myelodysplastic syndromes: a scoring system with prognostic significance. *British Journal of Haematology* **59**, 425–33.

Nusbaum NJ (1991). Concise review: genetic basis for sideroblastic anemia. *American Journal of Hematology* **37**, 41–4.

Raskin WH *et al*. (1984). Evidence for a multistep pathogenesis of a myelodysplastic syndrome. *Blood* **63**, 1318–23.

Roberts PD, Hoffbrand AV, Mollin DL (1966). Iron and folate metabolism in tuberculosis. *British Medical Journal* **2**, 198–202.

Savage D, Lindenbaum J (1986). Anemia in alcoholics. *Medicine* **65**, 322–38.

Weatherall DJ, Abdalla S (1982). The anaemia of *P. falciparum* malaria. *British Medical Bulletin* **38**, 147–51.

Wiley JS, Moore MR (2000). Heme biosynthesis and it disorders: porphyrias and sideroblastic anemias. In: Hoffman R, *et al*., eds. *Hematology: basic principles and practice*, pp. 428–45. Churchill Livingstone, New York,

22.5.9 Haemolytic anaemia - congenital and acquired

Frank J. Strobl and Leslie Silberstein

Introduction

The mechanisms of haemolysis

Following release into the circulation, normal red cells survive for approximately 120 days. As the circulating red-cell mass decreases (anaemia), less oxygen is transported from the lungs to other tissues of the body. In response, the kidneys increase their synthesis and secretion of erythropoietin, which stimulates erythropoiesis, in order to restore normal red-cell mass and oxygen delivery. A deficient red-cell mass results from inadequate production (hypoplasia), loss (haemorrhage), or premature destruction (haemolysis) of the red cells. In cases where red-cell survival is reduced to such an extent that normal bone marrow cannot compensate, a haemolytic anaemia results. The haemolytic anaemias are either genetically determined or acquired. As will be described in this chapter, premature destruction of red cells occurs through two primary mechanisms. First, decreased erythrocyte deformability secondary to membrane defects, metabolic abnormalities, exogenous oxidizing agents, or pathological antibodies provokes red-cell sequestration and extravascular haemolysis in the spleen and other components of the reticuloendothelial system. Second, exposure to pathological antibodies, activated complement, mechanical forces, chemicals, and infectious agents may lead to red-cell membrane damage and intravascular haemolysis.

The consequences of haemolysis

The clinical and laboratory changes associated with haemolysis reflect the physiological mechanisms responsible for restoring red-cell mass and removing free haemoglobin from the plasma. These changes are outlined in Table 1. Within several days of the onset of haemolysis and the development of anaemia, increased erythropoiesis results in erythroid hyperplasia (decreased myeloid/erythroid ratio) in the bone marrow and reticulocytosis (polychromasia and macrocytosis) in the peripheral blood. The peripheral blood film will also often exhibit microspherocytes, fragmented red blood cells, and nucleated red blood cells. If the haemolysis and anaemia begin early in life and persist, extramedullary erythropoiesis can develop in the spleen, liver, and lymph nodes. Chronic anaemia and the resulting marrow hyperplasia can also result in long-bone deformities. Free haemoglobin in the circulation binds to the serum protein haptoglobin. Haptoglobin–haemoglobin complexes are removed from the intravascular space by the reticuloendothelial system. If the rate of haemolysis is greater than the liver's ability to synthesize haptoglobin, serum haptoglobin levels fall. In patients with severe haemolysis, haemoglobinaemia and haemoglobinuria may develop. At low plasma haemoglobin levels, much of the free haemoglobin is reabsorbed in the proximal renal tubules. The renal tubular cells catabolize the haemoglobin converting iron into haemosiderin, which is eventually shed along with renal tubular cells into the urine resulting in haemosiderinuria. Haemosiderinuria is a reliable indicator of chronic intravascular haemolysis. At higher levels, free haemoglobin is found in the urine. Within the reticuloendothelial system, haemoglobin is metabolized and released into the serum as unconjugated bilirubin. The bilirubin is conjugated in the liver, excreted in the gut, converted to faecal urobilinogen, partially reabsorbed, and excreted by the kidneys as urinary urobilinogen. The intracellular enzyme lactate dehydrogenase is released from lysed red cells into the plasma.

Congenital anaemias

Congenital anaemias result from inherited defects in the red-cell membrane, red-cell enzymes, or haemoglobin. The haemoglobinopathies and thalassaemias are discussed in Chapter 22.5.7.

Disorders of the red-cell membrane

Introduction

The strength and flexibility required of the red-cell membrane is provided by the lipid bilayer and a proteinaceous, membrane-bound cytoskeleton. The membrane skeleton forms an underlying lattice which both supports and stabilizes the plasma membrane. Figure 1 provides a schematic model of the erythrocyte membrane and membrane-skeleton. The major, integral membrane proteins are band 3 and the glycophorins. Band 3 functions as a transmembrane channel for the diffusion of anions and glucose. The physiological role of the glycophorins is unknown. Cytoplasmic proteins located adjacent to the plasma membrane include spectrin, actin, band 4.1, ankyrin, and band 4.2. Spectrin, a heterodimeric protein composed of a and b subunits, is the principal component of the membrane-skeleton. The heterodimers of spectrin are bound to each other, head-to-head, to form heterotetramers and larger oligomers. The spectrin tetramers are cross-linked at their ends by actin. This interaction is strengthened by band 4.1. Band 4.1 also binds the cytoplasmic domain of glycophorin to spectrin. Ankyrin binds band 3 to the b chain of spectrin. Band 4.2 probably strengthens this interaction. A deficiency in any of these cytoskeletal proteins would be expected to result in defects in erythrocyte shape and deformability.

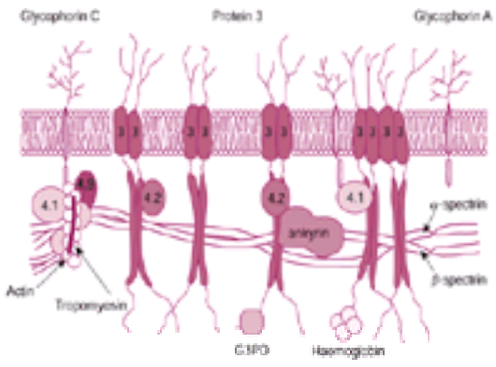

Fig. 1 Schematic illustration of the major components of the red-cell membrane and membrane skeleton. (Reproduced from Lux SE (1989). Hereditary disorders of the red-cell membrane skeleton. *Trends in Genetics* **5**, 222–7, with permission.)

Hereditary spherocytosis

Aetiology

Hereditary spherocytosis is inherited primarily in an autosomal dominant manner. Up to a quarter of cases, however, exhibit a non-dominant or recessive pattern of inheritance. The disorder is characterized by small spherocytic red cells with reduced deformability. The increased rigidity results in entrapment of the spherocytes,

primarily in the microcirculation of the spleen. The aetiology of hereditary spherocytosis appears to be heterogeneous. In most cases, the abnormal deformability is associated with defects in a- and b-spectrin or the proteins that bind spectrin to the plasma membrane: ankyrin, band 3, band 4.1, and band 4.2.

Clinical features

Hereditary spherocytosis occurs in all races but is most common in individuals of northern European descent. The prevalence of hereditary spherocytosis is estimated at 1:5000. Patients usually present in childhood with mild to moderate haemolysis. The main clinical features are anaemia, jaundice, and splenomegaly. The peripheral smear shows reticulocytosis and a variable degree of spherocytosis (Fig. 2). The red cells demonstrate increased osmotic fragility. The persistently elevated serum bilirubin levels often lead to the formation of bilary calculi and recurrent cholecystitis. Infrequent complications include ulcers, dermatitis, extramedullary haematopoietic tumours, cardiomyopathy, mental retardation, renal tubular acidosis, and neurological/muscular abnormalities. Less than 5 per cent of patients have severe disease. Severe haemolytic episodes are associated with infection. Rare aplastic crises are associated with parvovirus infection. Rare megaloblastic crises are associated with folate deficiency, especially during pregnancy.

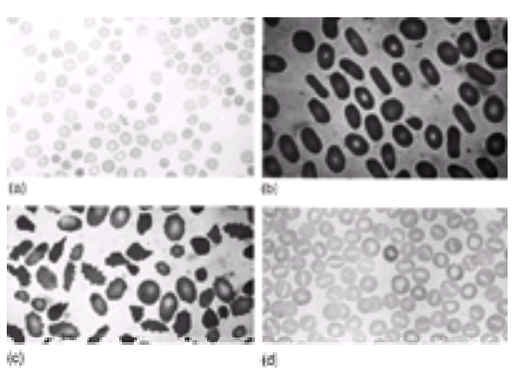

Fig. 2 Altered red-cell morphology: (a) spherocytes; (b) elliptocytes; (c) poikilorytes; (d) stomatocytes.

Treatment

The shortened red-cell survival and resulting anaemia of hereditary spherocytosis can be corrected in almost all cases by splenectomy. In patients with severe hereditary spherocytosis, the haemolysis may only be partially corrected following splenectomy. Except for asymptomatic patients, splenectomy should be performed as early as possible after the age of 3 years. Although erythrocyte survival returns to normal, red-cell morphology and deformability remain abnormal following splenectomy. All patients undergoing splenectomy should receive pneumococcal, meningococcal, and *Haemophilus influenzae* vaccinations several weeks preoperatively. Postsplen-ectomy, antibiotic therapy to protect against pneumococcal sepsis is also recommended. All patients with haemolytic anaemia should take 1 mg of folic acid each day to prevent folate deficiency. Typically, regular red-cell transfusions are only required for patients with severe disease. Phototherapy and/or exchange transfusions can be used to treat hyperbilirubinaemia in the neonatal period.

Hereditary elliptocytosis

Aetiology

Hereditary elliptocytosis is a genetically heterogeneous disorder characterized by elliptical red cells and haemolysis. Autosomal dominant and rare autosomal recessive forms of the disease have been identified. The clinical severity ranges from an asymptomatic condition to a severe haemolytic anaemia (see hereditary pryopoikilocytosis below). In most cases, defects in both a-spectrin and b-spectrin have been implicated. These point mutations or deletions, which occur at the N-terminus of a-spectrin and the C-terminus of b-spectrin, interfere with spectrin self-association. Partial and complete deficiencies of membrane protein 4.1 are also associated, respectively, with mild and severe forms of hereditary elliptocytosis.

Clinical features

Hereditary elliptocytosis is most common in individuals of African and Mediterranean ancestry with a prevalence of approximately 1:2500. Less than 10 per cent of cases exhibit significant haemolysis. Symptomatic individuals with hereditary elliptocytosis exhibit moderate to severe anaemia, splenomegaly, and reticulocytosis. The peripheral blood smear contains elliptocytes, 'pencil cells', and other abnormally shaped red cells (Fig. 2). The osmotic fragility is normal in mild hereditary elliptocytosis but increased in more severe forms of hereditary elliptocytosis with significant poikilocytosis. Rarely, neonates with hereditary elliptocytosis experience severe haemolysis, which gradually improves during the first year of life. This improvement shadows the normal loss of fetal red cells during infancy. Increased concentrations of 2,3 diphosphoglycerate in fetal red cells probably interfere with spectrin-protein 4.1 interactions, thereby further destabilizing the red-cell membrane.

Treatment

Asymptomatic individuals with hereditary elliptocytosis require no treatment. Symptomatic individuals often obtain some degree of benefit from splenectomy. Recommendations for folate administration, presurgical immunizations, and antibiotic prophylaxis noted earlier for hereditary spherocytosis are similar for patients with hereditary elliptocytosis.

Hereditary pyropoikilocytosis

Hereditary pyropoikilocytosis is inherited in an autosomal recessive fashion. Individuals with hereditary pyropoikilocytosis demonstrate severe haemolytic anaemia characterized by marked red-cell fragmentation, microcytosis, poikilocytosis, and spherocytosis (Fig. 2). The osmotic fragility of red cells is increased. These heat-sensitive cells undergo fragmentation and haemolysis when warmed to greater than 41°C. Red cells from normal individuals do not undergo haemolysis or fragmentation until the temperature nears 50°C. Hereditary pyropoikilocytosis results from homozygosity or compound heterozygosity for the a-spectrin defects involved in hereditary elliptocytosis.

Hereditary spherocytic elliptocytosis

This form of hereditary elliptocytosis is characterized by mild to moderate haemolytic anaemia with both elliptocytes and spherocytes. The molecular basis of this disorder has yet to be identified. The osmotic fragility is increased. Splenectomy may be useful in symptomatic individuals.

Hereditary stomatocytosis

Several families have been identified with members exhibiting moderate to severe anaemia and circulating stomatocytes (Fig. 2). This rare, autosomal dominant disorder results from a defect in membrane permeability that allows increased Na^+ and H_2O influx and results in cellular swelling and increased osmotic fragility. An integral membrane protein, stomatin (band 7.2b), has been reported to be decreased or absent in affected individuals. Symptomatic individuals may obtain some benefit from splenectomy. A number of patients with hereditary stomatocytosis, however, have developed hypercoagulability and thrombosis following splenectomy. Stomatocytosis and mild to moderate haemolytic anaemia are also seen in rare individuals who have either absent (Rh_{null}) or markedly reduced Rh antigen expression.

Hereditary xerocytosis

Hereditary xerocytosis is a rare, autosomal dominant disorder characterized by red-cell dehydration and decreased osmotic fragility. The cellular dehydration appears

to be caused by a defect in membrane K+ permeability that leads to intracellular K^+ and H_2O loss. The crenated cells are cleared by the reticuloendothelial system resulting in moderate to severe haemolysis. Splenectomy may provide therapeutic benefit. As with hereditary stomatocytosis, the risk of hypercoagulability and thrombosis are increased following splenectomy in patients with xerocytosis.

Acanthocytosis

Acanthocytes or spur cells are red cells with many thorn-like projections of the membrane surface. Changes in the composition of membrane lipids within the membrane lipid bilayer appear to be responsible for the development of acanthocytosis. Acanthocytosis and haemolytic anaemia are seen in severe liver disease, abetalipoproteinaemia, and the McLoed syndrome. Infrequently, patients with severe liver disease accumulate free cholesterol in the outer leaflet of the red-cell membrane resulting in spur cell shape, trapping of the acanthocytes in the spleen, and rapidly progressive anaemia. Abetalipoproteinaemia is an autosomal recessive disorder characterized by acanthocytosis (>50 per cent on peripheral blood film), fat malabsorption, mild anaemia, retinitis pigmentosa, and progressive ataxia. The McLoed syndrome is characterized by variable acanthocytosis and mild anaemia. The disorder is X-linked and affected individuals appear to lack a membrane precursor of the Kell red-cell antigen that also acts as an integral membrane transporter.

Disorders of red-cell enzymes

Introduction

Erythrocytes circulate throughout the body for approximately 4 months lacking a nucleus, mitochondria, and ribosomes. As a result, red cells cannot synthesize protein nor take advantage of oxidative metabolism. Glucose metabolism is necessary to maintain the integrity of both the erythrocyte membrane and haemoglobin. In red cells, glucose is metabolized to lactate primarily through the anaerobic, Embden–Meyerhof pathway (Fig. 3). Eleven enzymes are required to break glucose down to lactate generating 2 moles of ATP and reducing 2 moles of NAD^+ to NADH. ATP is used primarily by membrane-associated ATPases which pump Na^+ and K^+ against their concentration gradients. NADH prevents the oxidation of iron in haemoglobin. 2,3-diphosphoglycerate (2,3-DPG) which binds to the b-subunits of haemoglobin and facilitates the release of oxygen is also a product of the Embden–Meyerhof pathway. In red cells, the production of 2,3-DPG is exaggerated in order to maintain intracellular concentrations equimolar with the concentration of haemoglobin. The production of 2,3-DPG occurs in a side pathway referred to as the Rapaport–Luebering shunt that branches from the main glycolytic pathway after the formation of 1,3-DPG. The other major red-cell energy pathway is the hexose–monophosphate shunt, which results in the reduction of $NADP^+$ to NADPH (Fig. 4). NADPH maintains adequate levels of reduced glutathione, which protects the red cells against oxidative damage.

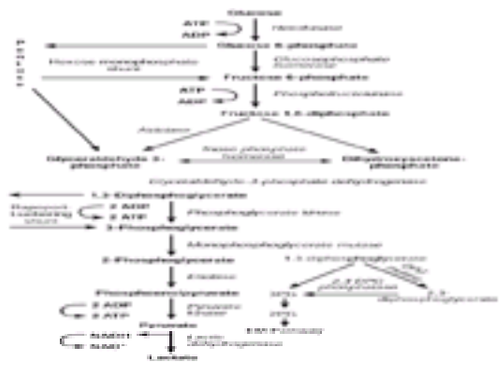

Fig. 3 The relationship between the main red-cell glycolytic pathway (Embden–Meyerhof) and the other metabolic pathways. The insert shows the production of 2,3-DPG in the Rapoport–Luebering shunt.

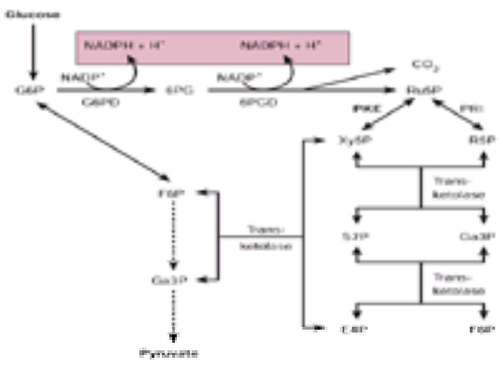

Fig. 4 The hexose monophosphate pathway. Intermediates: G6P, glucose 6-phosphate, F6P, fructose 6-phosphate; Ga3P, glyceraldehyde 3-phosphate; 6PG, 6-phosphogluconate; Ru5P, ribulose 5-phosphate; R5P, ribose 5-phosphate; Xy5P, xylulose 5-phosphate; S7P, sedoheptulose 7-phosphate; E4P, erythrose 4-phosphate. Enzymes: G6PD, glucose 6-phosphate deyhdrogenease; 6PGD 6-phosphogluconate dehydrogenase; PKE, epimerase; PRI, phosphoribose isomerase. Cosubstrates: $NADP^+$ and NADPH + H^+, oxidized and reduced forms of nicotinamide-adenine dinucleotide phosphate.

Many enzyme deficiencies are associated with haemolytic anaemia. Most of these enzyme deficiencies are not limited to the erythrocyte and, thus, are associated with multisystem disease (Table 2). The remaining enzyme deficiencies associated with haemolytic anaemia are clinically specific to the red cell. The majority of these enzyme deficiencies are rare, having been found in only a few families (Table 3). The two most common red-cell enzyme deficiencies—glucose-6-phosphate dehydrogenase deficiency and pyruvate kinase deficiency—are described below.

Glucose-6-phosphate dehydrogenase (G6PD) deficiency

Aetiology

G6PD catalyses the first step in the hexose–monophosphate shunt, which is responsible for reducing $NADP^+$ to NADPH. NADPH along with glutathione reductase maintains adequate supplies of reduced glutathione. Reduced glutathione is used by catalase and glutathione peroxidase to convert hydrogen peroxide to water. Oxygen radicals generated either through normal metabolism or by external oxidizing agents are converted to hydrogen peroxide, a highly oxidative agent. Therefore, through this series of enzyme reactions (Fig. 5) haemoglobin and other red-cell proteins are protected from oxidative damage. On the other hand, red cells deficient in G6PD are extremely sensitive to the oxidative actions of chemicals, drugs, infectious agents, and the bean *Vicia faba* (favism). G6PD deficiency is the most common hereditary enzyme deficiency of man affecting hundreds of millions of people world-wide. The disorder results from the X-linked, recessive inheritance of any one of a number of G6PD variants. The gene encoding G6PD is located on the long arm of the X chromosome (Xq28) and consists of 13 exons and greater than 18 kilobases. The G6PD protein is 514 amino acids long. The wild-type G6PD enzyme has been designated G6PD B. Nearly all G6PD mutations are the result of single point mutations that result in single amino acid substitutions. The rare exception consists of multiple point mutations or larger deletions. The altered protein structure most commonly results in decreased enzyme stability or, less commonly, decreased enzyme function. Males inheriting a mutant G6PD gene are affected. In female carriers, the levels of G6PD vary considerably as a result of random inactivation (lyonization) of the X chromosome. Female heterozygotes have two populations of red cells, those deficient in G6PD and those with normal levels of G6PD.

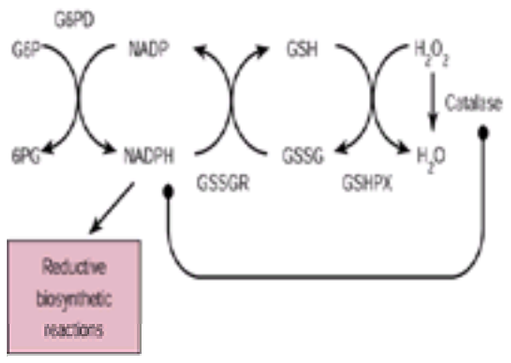

Fig. 5 The role of G6PD in red-cell metabolism: NADPH plays a dual role in (i) regeneration of glutathione (GSH) and (ii) stabilization of catalase (see also Chapter 22.5.12).

Clinical features

Since carriers of G6PD deficiency are more resistant to *Plasmodium falciparum* infection, the disorder has a geographical distribution similar to the haemoglobinopathies. G6PD deficiency is widespread in Africa, the Mediterranean, the Middle East, and Southeast Asia. More than 300 G6PD variants have been identified. Although most are clinically insignificant, a number of G6PD variants are associated with chronic anaemia or acute intermittent haemolysis. The two most common clinical forms of G6PD deficiency are an African variant (G6PD A-) and a family of Mediterranean variants. The African variant is synthesized in normal quantities, but is unstable and its levels decline slowly as the red cells age. Only the most senescent red cells are substantially lacking in enzyme activity. The Mediterranean variant has reduced enzyme activity, which also decreases with age. Newly released reticulocytes have significantly reduced G6PD activity, while mature erythrocytes lack any measurable enzyme activity. Both the African and the Mediterranean type are associated with little to no haematological abnormality under normal circumstances, but severe haemolysis and anaemia occur during periods of oxidant stress. Since only a small fraction of the circulating red cells are G6PD deficient in individuals with the African variant, the disorder is usually self-limited. In contrast, since all of the circulating red cells in the Mediterranean variant are G6PD deficient, these individuals may experience acute, life-threatening haemolysis. The onset of intravascular haemolysis often occurs within 24 h of exposure to the oxidizing agent and is accompanied by malaise, weakness, and abdominal/back pain. Within 2 to 3 days the patient develops anaemia, jaundice, haemoglobinuria, and, on rare occasions, acute renal failure. The peripheral smear often exhibits both anisocytosis and poikilocytosis, including spherocytes and blister cells (Fig. 6). Oxidant damage induces the formation of disulphide bridges within haemoglobin and leads to decreased solubility. Supravital staining with methylviolet demonstrates the presence of Heinz bodies. These intraerythrocytic inclusions consist primarily of denatured and precipitated haemoglobin. Favism is most common in areas where the *Vicia faba* bean grows, such as the Mediterranean and the Middle East. Two fava bean metabolites, divicine and isouramil, are believed to be oxidants. Favism occurs at any age, but is more common in children. Not all G6PD variants are susceptible to favism. Furthermore, every case of fava bean ingestion does not result in haemolysis. Mild haemolysis may also occur during the first weeks of life in G6PD-deficient infants. Occasionally, the jaundice is severe enough to cause kernicterus. Marked haemolysis can also occur in G6PD-deficient individuals during periods of illness, particularly during bacterial and viral infections. Rarely, G6PD deficiency is associated with chronic extravascular haemolysis and anaemia. These individuals retain both their sensitivity to oxidizing agents and, therefore, their risk of acute intravascular haemolysis.

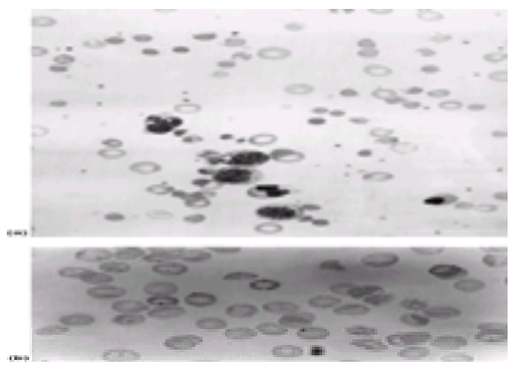

Fig. 6 Blood film in a case of acute haemolytic anaemia in a G6PD-deficient patient (favism). (a) Romanovsky stain, showing marked poikilocytosis, polychromatic macrocytes, bite cells, nucleated red cells, and a shift to the left in the granulocytic series. (b) Supravital stain with methyl violet, showing characteristic Heinz bodies.

Treatment

The diagnosis of G6PD deficiency depends on the demonstration of decreased red-cell G6PD activity. A rapid fluorescent screening assay is available to test at risk or suspected individuals during non-haemolytic periods. False-negative results can occur, especially following the resolution of an acute haemolytic episode when reticulocytes with increased enzyme activity predominate. All positive screening results should be confirmed by a quantitative spectrophotometric test based on the generation of NADPH from NADP. Molecular methods can be used accurately to diagnose female carriers heterozygous for G6PD deficiency. G6PD-deficient individuals should avoid substances known to induce haemolysis (Table 4). In newborns, severe jaundice should be treated with phototherapy and/or exchange transfusion. In all patients, blood transfusion should be considered during severe haemolytic episodes. Splenectomy may provide benefit in rare individuals with chronic haemolysis. Transfusion-dependent individuals should be carefully monitored for haemosiderosis and iron chelation should be initiated early.

Pyruvate kinase deficiency

Aetiology

Pyruvate kinase converts phosphoenol pyruvate to lactate and in the process generates ATP (Fig. 3). Therefore a deficiency in pyruvate kinase impairs the glycolytic pathway resulting in decreased production of ATP. An inadequate energy supply presumably leads to premature destruction of the red cells, particularly in the spleen and liver. There are four pyruvate kinase isoenzymes: M_1, M_2, L, and R. The R isoenzyme is unique to erythrocytes and is a product of the pk gene located on chromosome 1q21. Pyruvate kinase deficiency is inherited in an autosomal recessive fashion. The molecular basis of pyruvate kinase deficiency is quite heterogeneous, including single nucleotide substitutions, deletions, and insertions. Heterozygotes are unaffected, with variable but adequate levels of pyruvate kinase activity in their red cells. Homozygotes have little to no pyruvate kinase activity and exhibit haemolytic anaemia and splenomegaly. Many clinical homozygotes are actually compound heterozygotes with two different genetic lesions.

Clinical features

Pyruvate kinase deficiency occurs most commonly in individuals of northern European descent. The degree of haemolysis varies considerably with individuals and is often exacerbated by physiological stress such as pregnancy or infection. Severe pyruvate kinase deficiency usually presents at birth with haemolysis and jaundice. The haemolysis, anaemia, and jaundice continue throughout life, eventually resulting in splenomegaly, gallstones, and aplastic anaemia. Pyruvate kinase-deficient individuals appear to tolerate their anaemia better than individuals with comparable levels of anaemia due to other aetiologies. It is postulated that since the block in glycolysis occurs after the Rappaport–Leubering shunt, there is increased synthesis of 2,3-DPG in pyruvate kinase-deficient red cells. Increased synthesis of 2,3-DPG encourages the release of oxygen and thus more efficient oxygenation of tissues. The diagnosis of pyruvate kinase-deficiency is best confirmed by the detection of specific mutations at the genomic level.

Treatment

There is no specific therapy for pyruvate kinase deficiency. Periods of active haemolysis are treated with red-cell transfusions. Splenectomy often lessens the degree of haemolysis and anaemia. Splenectomy should be delayed until after the age of 3 years to reduce the risk of pneumococcal and meningococcal infections.

Acquired haemolytic anaemias

Immune haemolytic anaemias

Immune haemolysis may occur when IgG, IgM, or IgA antibodies and/or complement bind to the erythrocyte surface. The red-cell-bound antibodies may induce extravascular haemolysis, intravascular haemolysis, or both. Red cells coated with IgG typically undergo extravascular haemolysis during their transport through the reticuloendothelial system. Interactions between the Fc portion of IgG and surface Fc receptors allow the macrophages to phagocytose the coated erythrocytes. IgM, IgA, and, occasionally, IgG activate and fix complement to the erythrocyte surface. Macrophages also have receptors for the activated complement component C3b and, probably, phagocytose red cells through this pathway. The fixed complement can also induce intravascular haemolysis through activated membrane complex-mediated lysis.

The direct antiglobulin test or direct Coombs' test detects the presence of IgG antibody or complement on the red-cell surface. IgM and IgA antibodies are not directly detectable with standard testing methods. Rather, their presence is indirectly demonstrated by the detection of complement on the erythrocyte. In rare cases, the haemolytic anaemia is due to non-complement-fixing IgM or IgA antibodies. In this situation the direct antiglobulin test will be falsely negative. Eluates can be obtained from the antibody-coated red cells to determine the specificity of the antibody. Alternatively, the antibody may be free in the serum and its specificity determined by the indirect antiglobulin test or indirect Coombs' test. The presence of antibody or complement on the red cell, however, need not reflect ongoing haemolysis. Rather, the diagnosis of haemolytic anaemia rests on clinical findings and other laboratory data, such as red-cell morphology, haemoglobin, bilirubin, haptoglobin, LDH levels, reticulocyte count, and the presence or absence of haemoglobinaemia, haemoglobinuria, or haemosiderinuria. The serological findings provide information as to whether an immune basis exists and what type of immune haemolytic anaemia may be present. Autoantibodies, alloantibodies, and drugs may induce immune haemolytic anaemias.

Autoimmune haemolytic anaemia

Haemolytic antibodies directed against the individual's own red cells may arise as a primary/idiopathic event or may be secondary to lymphoid malignancies, connective tissue disorders, and infection. Autoimmune haemolytic anaemia is best classified according to the temperature at which the antibody optimally binds to the erythrocyte. The four major types of autoimmune haemolytic anaemia are warm autoimmune haemolytic anaemia, cold agglutinin syndrome, paroxysmal cold haemoglobinuria, and mixed-type autoimmune haemolytic anaemia

Warm autoimmune haemolytic anaemia

Aetiology

The offending antibody in warm autoimmune haemolytic anaemia is typically IgG and can be found on the red cell, in the serum, or both. The exact specificity of the antibody is often difficult to determine. With very rare exception, warm-reactive autoantibodies bind to all red cells tested, while others appear to have broad specificity within the Rh system. Occasionally, warm reactive autoantibodies will have relative specificity against an individual antigen such as Rh(D), Rh(C), or Kell.

Clinical features

Warm autoimmune haemolytic anaemia can arise at any age but is more common in older individuals, probably because of its association with lymphoid malignancies. Females are affected slightly more often than males. The direct antiglobulin test is positive for IgG and/or complement. In its mildest form the direct antiglobulin test is positive but red-cell survival is not significantly affected. Symptomatic patients present with anaemia, jaundice, and splenomegaly. Most patients with warm autoimmune haemolytic anaemia have a chronic, stable anaemia (haemoglobin < 8 g/dl). In its severest form, patients present with fulminant intravascular haemolysis, progressive anaemia, congestive heart failure, respiratory distress, and neurological abnormalities. As with other haemolytic anaemias, the peripheral smear often demonstrates anisocytosis and reticulocytosis with spherocytes and macrocytes (Fig. 7). The platelet count is usually normal except in patients with Evan's syndrome where the autoantibody destroys both red cells and platelets.

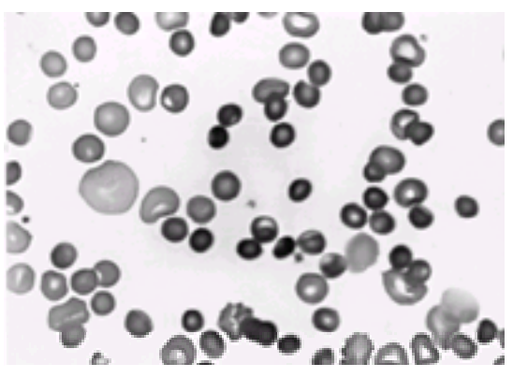

Fig. 7 The peripheral blood changes in autoimmune haemolytic anaemia. There is marked anisocytosis and anisochromia with many macrocytes and microspherocytes. The macrocytes reflect the reticulocytosis (×1000, Leishman stain).

Treatment

Corticosteroids, which presumably block macrophage Fc receptor activity and inhibit antibody production, are the primary therapy for autoimmune haemolytic anaemia. Prednisone at a dose of approximately 1 to 2 mg/kg body weight in divided doses is effective in most patients. Higher doses rarely provide additional benefit, but do increase the number and severity of side-effects. Treatment continues until the haemoglobin levels stabilize. The initial dose of prednisone can then be tapered at a rate of 5 to 10 mg per week. Once a dose of 10 mg/day is reached, the steroid taper should progress more slowly in order to determine the minimum controlling dose. Side-effects may be reduced by using an alternate-day schedule. Splenectomy should be performed only in steroid-refractory patients or patients requiring unacceptably high doses of prednisone to maintain remission. Alternative therapies including azathioprine, cyclophosphamide, intravenous immunoglobulin (IVIG), danazol, and plasma exchange should be reserved for patients unfit for splenectomy or who have failed to respond to steroids and surgery.

Transfusion with ABO and Rh compatible cross-match least incompatible red cells should be performed in patients with symptomatic anaemia. Transfusion should not be withheld because of serological incompatibility. Active serum autoantibodies, however, can mask the presence of clinically significant alloantibodies. Therefore, the most important consideration prior to transfusion is to confirm the presence or absence of alloantibodies in the patient's serum. Various autologous and allogeneic red-cell absorption techniques exist to remove the autoantibody from a sample of the patient's serum and allow identification of any existing alloantibodies. If clinically significant alloantibodies are present, red cells lacking the corresponding antigen(s) should be selected for transfusion. If the autoantibody has a definite specificity, red cells lacking that antigen may be selected. Transfusions in life-threatening situations should not be delayed if the above tests are not readily available or completed.

Cold agglutinin syndrome

Aetiology

Cold agglutinin syndrome accounts for approximately a quarter of all cases of autoimmune haemolytic anaemia. The disorder occurs as an acute or chronic condition. In cold autoimmune disorders the signs and symptoms of disease result from either the agglutination of red cells or from haemolysis. The autoantibodies are typically IgM and are most active at low temperatures. Rare examples of IgG and IgA cold-reactive autoantibodies have been reported. In the lower temperatures of the peripheral circulation, the IgM autoantibodies bind to red cells and activate complement. In warmer areas of the circulation, the IgM dissociates from the erythrocyte leaving activated complement fixed to the red-cell surface. The severity of the disorder depends on both the titre of the antibody and the thermal range at which it is most active. The autoantibody specificity in cold agglutinin syndrome is usually anti-I. Anti –i specificity is associated with infectious mononucleosis.

Clinical features

Acute cold autoimmune haemolytic anaemia is commonly seen in adolescents and young adults following infection with *Mycoplasma pneumoniae* or infectious mononucleosis. Haemolysis occurs approximately 1 to 2 weeks following infection and is most commonly associated with a rise in polyclonal anti-I IgM antibody with *Mycoplasma pneumonia* or polyclonal anti-i IgM antibody with infectious mononucleosis . Chronic cold autoimmune haemolytic anaemia occurs most commonly in the elderly, either idiopathically or associated with lymphoma, chronic lymphocytic leukaemia, or Waldenstrom's macroglobulinaemia. Patients may experience chronic intravascular haemolysis and anaemia that are exacerbated by cold temperature. Patients are often also plagued by episodes of Raynaud's phenomenon. Monoclonal IgM antibodies with kappa light chains and anti-I specificity usually cause the red-cell agglutination and haemolysis in this condition. Examination of the peripheral smear shows red-cell agglutination. The direct antiglobulin test is positive for complement.

Treatment

Treatment involves keeping the patient in a warm environment (>37°C). Steroids and splenectomy are of little benefit. Severe cold autoimmune haemolytic anaemia secondary to a B-cell neoplasm can be treated with chlorambucil, cyclophosphamide, or a-interferon. Blood transfusion should be avoided. In situations of life-threatening anaemia, the blood should be given slowly through a blood warmer. Plasma exchange is often helpful, but its effects should be expected to be temporary. Hypothermia must be avoided during surgery (especially surgical procedures involving extracorporeal circuits) in patients with cold autoimmune haemolytic anaemia.

Paroxysmal cold haemoglobinuria

Aetiology

Paroxysmal cold haemoglobinuria is the rarest form of autoimmune haemolytic anaemia. The disorder is caused by the complement-fixing Donath–Landsteiner IgG antibody. In the cold, this antibody binds to, and irreversibly fixes, complement to the red-cell membrane. Upon return to warmer temperatures, the antibody dissociates from the red cell leaving activated complement to lyse the cell. The Donath–Landsteiner antibody appears to have an anti-P specificity allowing it to bind to practically all red cells.

Clinical features

Patients present with acute intravascular haemolysis, abdominal pain, peripheral cyanosis, Raynaud's phenomenon, haemoglobinaemia, and haemoglobinuria after exposure to cold. In the past, paroxysmal cold haemoglobinuria was commonly associated with congenital syphilis but most cases are now associated with viral infections in children or are idiopathic in adults. During or shortly after a haemolytic episode, the direct antiglobulin test is positive for complement but negative for IgG.

Treatment

No specific therapy for paroxysmal cold haemoglobinuria exists; steroids are not useful. Most postinfectious cases of paroxysmal cold haemoglobinuria are self-limited and require only supportive care. Avoidance of cold ambient temperatures can help prevent recurrent attacks in patients with chronic paroxysmal cold haemoglobinuria. Transfusion is indicated only for severe haemolysis and life-threatening anaemia. Since the Donath–Landsteiner antibody rarely causes agglutination, most random donor blood units will be compatible with patient sera. Transfusions with extremely rare P-antigen-negative blood should be reserved only for those patients who do not respond to random donor blood. The use of a blood warmer should be considered.

Mixed-type autoimmune haemolytic anaemia

Aetiology

Approximately 8 per cent of all autoimmune haemolytic anaemias are of the mixed type. Both IgG and complement are present on the red cells. Both warm-reactive IgG autoantibodies and cold-reactive agglutinating IgM autoantibodies are present in the serum. The warm-reactive IgG autoantibodies are indistinguishable from antibodies encountered in warm autoimmune haemolytic anaemia. The IgM autoantibodies are unlike those in cold-agglutinin syndrome in that they generally have low titres at 4°C and have high thermal amplitudes, reacting at 30°C or above. These IgM autoantibodies usually have no distinguishable specificity, but on occasion have I or i specificities.

Clinical features

Mixed-type autoimmune haemolytic anaemia may be idiopathic or secondary, most commonly associated with systemic lupus erythematosus. The haemolytic anaemia is often severe and chronic with intermittent exacerbations. Exposure to cold does not increase the haemolysis.

Treatment

Steroids, splenectomy, or cytotoxic agents often provide therapeutic benefit in mixed-type autoimmune haemolytic anaemia. If blood transfusions are necessary, selection of blood should adhere to transfusion guidelines outlined earlier for warm autoimmune haemolytic anaemia. Administration of blood through a warmer should be considered.

Drug-induced haemolytic anaemia

Drugs may induce antibodies to bind to the erythrocyte surface resulting in a positive direct antiglobulin test or haemolysis. There are four mechanisms by which drugs can cause a positive direct antiglobulin test: (1) drug hapten, (2) immune complex formation, (3) autoantibody production, and (4) non-specific adsorption. Only the first three mechanisms are associated with haemolysis. Treatment and prevention are as straightforward as drug avoidance.

Drug hapten

Certain drugs bind to the red-cell membrane with a high affinity. Association of the drug with the membrane constituents allows the drug to act as a hapten. The antibodies produced are commonly IgG and are directed predominantly against the drug. Extravascular haemolysis develops gradually, but may be life-threatening if left untreated. After the offending drug is identified and withdrawn, the positive direct antiglobulin test and the haemolysis may persist for several weeks. Serum from these patients will not react with other red cells unless the drug is also present. Penicillin and the cephalosporins are the most notorious examples of this phenomenon. Approximately 3 per cent of patients receiving large doses of penicillin (millions of unit per day) intravenously will develop a positive direct antiglobulin test. Only the rare patient develops haemolytic anaemia.

Immune complex

Other drugs induce the binding of IgM or IgG antibodies that activate complement and cause intravascular haemolysis. The antibodies appear to recognize both the drug and a component of the red-cell membrane. The direct antiglobulin test is often positive for complement but not antibody. Haemoglobinaemia and haemoglobinuria are common. Renal failure occurs in about half of the cases. Once the offending drug is withdrawn, the haemolysis stops. Serum from these patients will lyse normal red cells only in the presence of the drug. Quinine, quinidine, phenacetin, chlorpropamide, and sulfonylureas are examples.

Autoantibodies

Some drugs stimulate the synthesis of red-cell autoantibodies. Patient serum and red-cell eluates react with normal red cells in the absence of the drug. The autoantibodies are indistinguishable from those found in warm autoimmune haemolytic anaemia. The direct antiglobulin test usually becomes positive after 3 to 6 months of drug administration. The haemolysis typically ceases within 2 weeks after the withdrawal of the drug, but the direct antiglobulin test can remain positive for up to 2 years. a-methyldopa, L-dopa, procainamide, mefenamic acid, and sulindac are examples of drugs that can stimulate the production of red cell autoantibodies.

Non-specific protein adsorption

Often a drug-induced positive direct antiglobulin test reflects non-immunological adsorption of protein, including immunoglobulins. First-generation cephalosporins were originally associated with this phenomenon. More recently, other drugs, including suramin, cisplatin, and sulbactam, have also been implicated. This mechanism is not associated with reduced red-cell survival.

Alloimmune haemolytic anaemias

Acute haemolytic transfusion reactions

Aetiology

The most catastrophic cases of alloimmune haemolysis occur following the transfusion of ABO-incompatible red cells. Naturally occurring IgM anti-A and anti-B antibodies bind to the incompatible red cells and activate complement resulting in intravascular haemolysis. Human error leading to the misidentification of patients, their blood samples, or the units of red cells to be transfused is responsible for virtually all cases of ABO incompatibility. Only rarely, do other non-ABO IgG alloantibodies cause acute, severe haemolysis.

Clinical features

Symptoms of an acute haemolytic transfusion reaction may begin after the infusion of as little as 10 ml of incompatible blood. The signs and symptoms include fever, chills, nausea, vomiting, hypotension, respiratory distress, haemoglobinuria, and chest or flank pain. Despite treatment, acute haemolytic transfusions reactions can result in renal failure, disseminated intravascular coagulation, and even death.

Treatment

Once an acute haemolytic transfusion reaction is suspected, the blood transfusion should be stopped immediately. Aggressive treatment of the hypotension with intravenous fluids and pressor agents (that is low-dose dopamine) is crucial. Other critical measures include monitoring the urine output and promoting renal blood flow with diuretics (furosemide or mannitol). Either heparin or the administration of platelets, plasma, and cryoprecipitate can be used to treat organ or life-threatening bleeding secondary to disseminated intravascular coagulation.

Delayed haemolytic transfusion reactions

Aetiology

Approximately 2 to 3 per cent of transfusion recipients become alloimmunized to non-ABO red-cell antigens. Haemolysis is not generally seen during the primary immune response since the transfused red cells often disappear from the circulation before antibody titres reach clinically significant levels. In the absence of further antigenic stimuli antibody titres may diminish to undetectable levels. Subsequent transfusion of red cells possessing the offending antigen, however, will induce an anamnestic response with reappearance of the IgG antibodies within hours to days. Binding of the IgG antibody to the transfused antigen-positive red cells results in a positive direct antiglobulin test and possibly mild to moderate extravascular haemolysis.

Clinical features

Most patients experiencing a delayed haemolytic transfusion reaction present with fever, jaundice, and decreasing haemoglobin levels 1 to 2 weeks following the transfusion of incompatible red cells. Delayed haemolytic transfusion reactions are often discovered during evaluation for fever of unknown origin or when the haemoglobin level fails to increase following transfusion.

Treatment

Treatment is rarely necessary; acute renal failure or disseminated intravascular coagulation are uncommon. If a delayed haemolytic transfusion reaction is suspected, both the patient's serum and an eluate from the circulating red cells should be tested for alloantibodies. If alloantibodies are present, their specificities should be determined. Donor red-cell units lacking the offending antigen should be selected for subsequent transfusions.

Passenger lymphocyte haemolysis

Aetiology

Recipients of a haematopoietic or a solid-organ transplant may experience delayed extravascular haemolysis. In this circumstance, lymphocytes of donor origin produce haemolytic antibodies against ABO or other red-cell antigens possessed by the recipient.

Clinical features

Haemolysis due to passenger lymphocytes is most commonly seen in out-of-group yet ABO-compatible liver and bone marrow transplants (group A or group B recipients of group O tissue) but can also occur in recipients of lung, heart, and kidney transplants. This haemolysis can begin within several days after the transplant and continue for several months.

Treatment

If significant ABO haemolysis occurs, patients should be transfused with group O red cells. If non-ABO haemolysis is present, elution of the patient's red cells may help to identify the antibody specificity and allow transfusion of antigen-negative red cells.

Haemolytic disease of the newborn (HDN)

Haemolytic disease of the newborn occurs when maternal IgG antibodies cross the placenta and bind to fetal red cells resulting in extravascular haemolysis. Usually these antibodies possess specificities within the Rhesus or ABO blood group systems. Occasionally the antibodies are directed against other red-cell antigens such as the Kell, Kidd, and Duffy. In the mildest cases, anaemia develops several weeks after birth and is of little clinical consequence. In more severe cases the neonate develops progressive anaemia and jaundice within the first week of life. If left untreated, bilirubin levels may reach levels associated with kernicterus causing brain

damage and death. In the most severe cases, the fetus develops profound anaemia as early as the fifth month of gestation and may be stillborn or delivered grossly oedematous (hydrops fetalis). An infant with hydrops fetalis also has ascites, hepatosplenomegaly, and erythroblastosis and usually dies shortly after birth.

Rhesus D incompatibility

Haemolytic disease of the newborn is most common and severe in rhesus-D-negative women carrying a rhesus-D-positive fetus. The mother develops anti-D IgG antibodies following exposure to the D antigen during a previous pregnancy, or as a result of the transfusion of D-antigen-positive red cells. One half of all cases of rhesus D alloimmunization are due to transplacental haemorrhage from the fetus at the time of delivery. Spontaneous transplacental haemorrhage can also occur during gestation, particularly during the third trimester. The risk of transplacental haemorrhage increases with ectopic pregnancy, spontaneous or therapeutic abortion, chorionic villus sampling, caesarean section, and trauma. Approximately 8 per cent of untreated D-negative women who deliver a D-positive child will become alloimmunized to the D antigen.

It is essential to identify pregnant women at risk for rhesus D haemolytic disease of the newborn and to prevent sensitization. All pregnant women should have their ABO and rhesus types identified as early as possible. Their serum should be screened for alloantibodies against the D antigen and other red-cell antigens. Pregnant women who are D-antigen-negative and have an initial negative antibody screen should have their serum retested for alloantibodies at 28 weeks gestation. If the initial antibody screen is found positive, antibody titres should be followed at 2 to 4-week intervals to determine whether further sensitization is occurring. The presence of an antibody, however, does not indicate on going haemolysis in all cases.

Naturally occurring IgM antibodies are common during pregnancy but do not cross the placenta. Furthermore, fetal red cells may lack the antigen corresponding to the mother's antibody. Molecular typing of the father's DNA or even fetal DNA is available for several red-cell antigens including D, E/e, C/c, Jk ^{a}/Jk^{b}, and K1/K2. A rising titre of anti-D antibody or other clinically significant red-cell alloantibodies indicates ongoing sensitization and possible haemolytic disease of the newborn.

From 18 weeks of gestation and onward, ultrasonography and fetal blood sampling can be used to assess the severity of haemolysis. After 28 weeks of gestation amniocentesis can be performed. If the fetus is experiencing significant haemolysis and anaemia, clinical intervention must be prompt. Prior to 34 weeks of gestation intrauterine transfusion with blood lacking the offending antigen should be performed. After 36 weeks gestation induced labour should be considered. Upon birth of a 'at risk' fetus a sample of cord blood should undergo a direct antiglobulin test and have measurements of haemoglobin and bilirubin performed. If the direct antiglobulin test on the cord blood sample is positive and the mother's antibody screen remains negative, haemolytic disease secondary to ABO incompatibility or antibodies against low-incidence red-cell antigens should be considered.

Infants with severe anaemia (haemoglobin <12 g/dl) or severe jaundice (bilirubin >14 mg/dl) should undergo exchange transfusion. Phototherapy can also be used to decrease bilirubin levels. A non-sensitized D-antigen-negative mother's blood should also be tested to determine the amount of fetomaternal haemorrhage at delivery. Administration of 300 µg of IgG anti-D (RhIg) within 72 h of delivery will protect 99 per cent of D-antigen-negative mothers from developing anti-D antibodies. Prophylactic administration of RhIg at 28 weeks gestation and following invasive procedures or traumatic events will virtually eliminate the chance of alloimmunization. Patients with large transplacental haemorrhages quantitated by the Kleihauer–Betke acid-elution technique should receive additional RhIg at a dose equivalent to 300 µg for every 15 cc of fetal red blood cells.

ABO incompatibility

Although 15 per cent of pregnancies are ABO incompatible, haemolytic disease of the newborn due to ABO incompatibility is rare. Mild to moderate haemolysis and hyperbilirubinaemia due to ABO incompatibility occurs in about 1.5 per 1000 pregnancies. Group A and group B infants of group O mothers are at greatest risk. Unlike with rhesus D antigen, ABO-haemolytic disease of the newborn occurs during the first pregnancy as often as subsequent pregnancies. Exchange transfusion with group O red cells is rarely required. Hydrops fetalis never occurs.

Non-immune acquired haemolytic anaemias

Red-cell survival may also be reduced by a number of non-inherited, non-immune mechanisms. As red cells circulate they are vulnerable to a variety of insults that may cause structural or metabolic alterations. These changes generally result in reduced red-cell deformability leading ultimately to extravascular haemolysis. These insults include infection, mechanical trauma, and exposure to chemicals, heat, or venom. They often also cause intravascular haemolysis by directly lysing the red-cell membrane. Other less understood causes of acquired non-immune haemolytic anaemias are listed in Table 5.

Infection

Infectious causes of haemolysis are primarily parasites and bacteria. Direct parasitization of red cells by *Plasmodium falciparum*, *Plasmodium vivax*, and *Plasmodium malariae* causes both intravascular haemolysis due to direct membrane destruction and extravascular haemolysis due to membrane alteration and activation of the reticuloendothelial system. Infrequently, *in utero* infection of the fetus with *Toxoplasma gondii* resembles severe haemolytic disease of the newborn. Infants are born hydropic and severely anaemic. Premature delivery and stillbirth are common. *Babesia microti*, endemic in areas of the Northeast and Midwest in North America, is transmitted by ticks and causes severe haemolysis during the erythrocytic phase of its life cycle. Bacterial infections, particularly Gram-negative organisms which produce endotoxin or proteolytic enzymes, may produce mechanical haemolysis by inducing disseminated intravascular haemolysis or red-cell membrane damage via degradation of membrane phospholipids and proteins. *Bartonella bacilliformis* endemic to western South America causes Oroya fever characterized by fever, chills, musculoskeletal pain, and acute intravascular haemolysis.

Chemical

Drugs and chemicals known to cause haemolysis through direct oxidative damage are summarized in Table 6. In most cases the strong oxidant activity of these chemicals overwhelm normally functioning reduction mechanisms responsible for protecting haemoglobin and the red-cell membrane. Variability in the absorption of the chemical or its metabolism determine whether a particular individual will develop chemical-induced haemolytic anaemia. Often it is the chemical's metabolite that is responsible for inducing haemolysis. The red cells of newborns do not have functional reduction mechanisms and thus are more sensitive to oxidant activity.

Mechanical

Mechanical fragmentation of erythrocytes can occur when foreign material is placed within the vasculature, when fibrin strands or platelet thrombi obstruct small blood vessels, or when direct physical forces compress superficial blood vessels.

Foreign material

Mechanical haemolysis occurs most commonly with artificial valvular prostheses, particularly when accompanied by turbulent blood flow. Bacterial endocarditis and associated valvular vegetations can also cause fragmentation of red cells. Haemolysis also occurs in up to 10 per cent of patients with transjugular intrahepatic portosystemic shunts (TIPS). Increased cardiac output as a result of anaemia, exercise, or medications can increase the rate of red-cell fragmentation. The peripheral smear usually demonstrates schistocytes and microspherocytes. Severe haemolysis usually requires surgical repair.

Microangiopathic haemolytic anaemia (MAHA)

MAHA describes a spectrum of disorders characterized by mechanical destruction of red cells resulting from thrombi that occlude the microvasculature. The red cells are probably fragmented during their forced passage through the meshwork of fibrin strands that make up the microthrombi. The degree of anaemia is variable. The peripheral smear reveals findings typical of mechanical haemolysis including schistocytes, microspherocytes, and a reticulocytosis (Fig. 8). The absence of a positive direct antiglobulin test along with significant thrombocytopenia helps to confirm the diagnosis. Two other major forms of MAHA are haemolytic uraemic syndrome and thrombotic thrombocytopenic purpura (TTP).

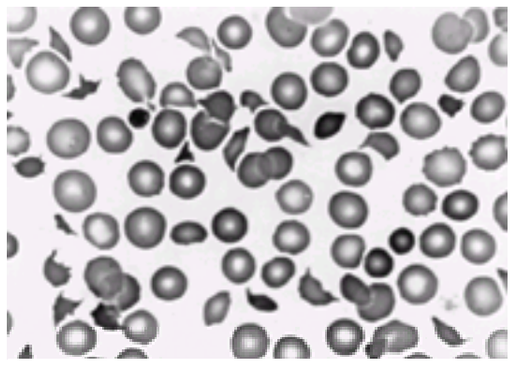

Fig. 8 The peripheral blood changes in microangiopathic haemolytic anaemia. This patient had recurrent thrombocytopenic purpura and the marked fragmentation of the red cells together with microspherocytosis is evident on the blood film (× 1000, Leishman stain).

Haemolytic uraemic syndrome

Haemolytic uraemic syndrome is primarily, but not exclusively, a disease of childhood. The disorder consists of widespread damage to the vascular endothelium and fibrin deposition. These pathological changes are frequently most severe in the renal arterioles and glomerular capillaries. The disorder usually develops following a febrile illness. Numerous reports have documented the development of haemolytic uraemic syndrome following infections with toxin-secreting strains of *Escherichia coli* (strain O157:H7) or shigella. Initial nausea, vomiting, and diarrhoea can develop into severe abdominal pain and bloody diarrhoea. Acutely, the child may develop hypertension, oliguria, purpura, bleeding, and anaemia. If left untreated, convulsions, coma, and death may occur. Mortality rates as high as 10 per cent have been associated with haemolytic uraemic syndrome. The peripheral smear exhibits schistocytosis and thrombocytopenia. Therapy consists mainly of supportive care, transfusion, control of blood pressure, and dialysis.

Thrombotic thrombocytopenic purpura (TTP)

TTP is caused by either a congenital deficiency of, or an acquired inhibitor to, a serum metalloprotease which is responsible for cleaving unusually large multimers of von Willibrand's factor. Left uncleaved, the large von Willibrand's factor multimers induce TTP by causing the agglutination of circulating platelets. Most episodes of TTP occur without an obvious inciting event. However, TTP has been associated with infection, pregnancy, transplantation, AIDS, and drugs such as mitomycin C, ticlopidine, cyclosporine, and tacrolimus (FK506). TTP occurs mainly in adults and more commonly involves the central nervous system. The onset is often sudden with fever, purpura, petechiae, anaemia, thrombocytopenia, and neurological abnormalities. The neurological sequelae include convulsions, coma, paralysis, delirium, and stroke. The peripheral smear demonstrates schistocytes, thrombocytopenia, and a reticulocytosis. During acute episodes front-line therapy includes steroids and daily plasma exchange with fresh frozen plasma or virally-inactivated solvent-detergent plasma (SD plasma). Plasma exchange probably accomplishes one or more of the following:

1. removes the antibody to the protease;
2. removes large multimers of von Willibrand's factor; or
3. replenishes normal protease.

In patients who do not initially respond to plasma exchange with fresh frozen plasma, cryopoor-supernatant is often used as the replacement fluid. Cryopoor-supernatant contains markedly reduced levels of normal von Willibrand's factor which is believed to enhance the formation of microthrombi in some patients. Individuals with drug-induced TTP appear to be less responsive to therapy. Additional therapies in refractory or relapsing patients include vincristine, cytoxan, cyclosporine, and splenectomy. Anecdotal evidence suggests that platelet transfusion can exacerbate the disorder. Therefore, platelet transfusions should be avoided unless absolutely necessary to treat haemorrhage.

March haemoglobinuria

Haemoglobinuria can occur in soldiers or joggers following extended periods of marching or running on a hard surface or in karate or conga drummer enthusiasts following practice. This mechanical haemolysis appears to be the result of red-cell compression in superficial blood vessels during the period of contact between the extremity and the hard surface. The peripheral smear is normal. Treatment is unnecessary as the syndrome is otherwise symptomless and lacks significant clinical sequelae.

Thermal haemolysis

Normal red cells undergo fragmentation and lysis when heated to temperatures of 49°C or higher. The two most common clinical situations associated with heat-induced red-cell lysis are the use of faulty blood warmers during transfusion or patients who have sustained extensive burns.

Venom

Haemolysis has been observed following bee and wasp stings, spider bites, and snake bites. The haemolysis occurs secondary to disseminated intravascular coagulation or as a result of proteolytic enzymes contained within the venom.

Further reading

Agre P *et al*. (1985). Partial deficiency of erythrocyte spectrin in hereditary spherocytosis. *Nature* **314**, 380–3.

Bowman JM (1986). Fetomaternal ABO incompatibility and erythroblastosis fetalis. *Vox Sanguinus* **50**, 104–6.

Brecher ME (1996). Hemolytic transfusion reactions. In: Rossi EC *et al*., eds. *Principles of transfusion medicine*, pp. 747–63. Williams and Wilkins, Baltimore.

Conboy JG *et al*. (1993). An isoform specific mutation in the protein 4.1 gene results in hereditary elliptocytosis and complete deficiency of protein 4.1 in erythrocytes but not in nonerythroid cells. *Journal of Clinical Investigation* **91**, 77–82.

Davidson RJL (1969). March or exertional hemoglobinuria. *Seminars in Hematology* **6**, 150.

Freedman J (1987). The significance of complement on the red cell suface. *Transfusion Medicine Reviews* **1**, 58–70.

Furlan M *et al*. (1998). Von Willebrand factor-cleaving protease in thrombotic thrombocytopenic purpura and the hemolytic-uremic syndrome. *New England Journal of Medicine* **399**, 1578–84.

Garratty G (1987). The significance of IgG on the red cell surface. *Transfusion Medicine Reviews* **1**, 47–57.

Hows J (1986). Donor-derived red blood cell antibodies and immune hemolysis after allogeneic bone marrow transplantation. *Blood* **67**, 177–81.

Judd WJ *et al*. (1990). Prenatal and perinatal immunohematology: recommendations for serologic management of the fetus, newborn infant and obstetric patient. *Transfusion* **30**, 175–83.

Leger RM, Garraty G (1999). Evaluation of methods for detecting alloantibodies underlying warm autoantibodies. *Transfusion* **39**, 11–16.

Liu SC, Palek J, Prchal J (1982). Defective spectrin dimer-dimer association in hereditary elliptocytosis. *Proceedings of the National Academy of Science* **79**, 2072–6.

Marsh GW, Lewis SM (1969). Cardiac hemolytic anemia. *Seminars in Hematology* **6**, 133–45.

Prchal JT, Gregg XT (2000). Red cell enzymopathies. In: Hoffman R *et al*., eds. *Hematology: basic principles and practice*, pp. 561–76. Churchill Livingstone, Philadelphia.

Ramsey G (1991). Red cell antibodies arising from solid organ transplants. *Transfusion* **31**, 76–86.

Savvides P *et al*. (1993). Combined spectrin and ankyrin deficiency is common in autosomal dominant hereditary spherocytosis. *Blood* **82**, 2953–60.

Schrier SL (2000). Extrinsic nonimmune haemolytic anemias. In: Hoffman R *et al*., eds. *Hematology: basic principles and practice*, pp. 630–8. Churchill Livingstone, Philadelphia.

Shepard KV, Bukowski RM (1987). The treatment of thrombotic thrombocytopenic purpura with exchange transfusions, plasma infusions, and plasma exchange. *Seminars in Hematology* **24**, 178–93.

Shulman NR, Reid DM (1993). Mechanisms of drug-induced immunologically mediated cytopenias. *Transfusion Medicine Reviews* **7**, 215–29.

Tsai H-M, Lian EC-Y (1998). Antibodies to von Willebrand factor-cleaving protease in acute thrombotic thrombocytopenic purpura. *New England Journal of Medicine* **399**, 1585–94.

Vengelen-Tyler V *et al*., eds (1999). *Technical manual*, 13th edn. American Association of Blood Banks, Bethesda.

Vulliamy TJ, Beutler E, Luzzatto L (1993). Variants of glucose 6-phosphate dehydrogenase are due to missense mutations spread throughout the coding region of the gene. *Human Mutation* **2**, 159–67.

22.5.10 Disorders of the red cell membrane

Patrick G. Gallagher Sara S. T. O. Saad, and Fernando F. Costa

The red cell membrane

Composition and function

Although the primary structure and a number of the important functions of the red cell membrane have been known for several years, its study continues to yield important insights into our understanding of membrane structure and function. The red cell membrane is composed of three major structural elements: a lipid bilayer primarily composed of phospholipids and cholesterol; integral proteins embedded in the lipid bilayer that span the membrane; and a membrane skeleton on the internal side of the red cell membrane.

The membrane and its skeleton provide the erythrocyte with the ability to undergo significant deformation without fragmentation or loss of integrity during its travel through the microcirculation. The membrane also assembles and organizes the proteins of the lipid bilayer and the membrane skeleton, allowing the red cell to participate in a wide range of functions. These include influencing cellular metabolism by selectively and reversibly binding and inactivating glycolytic enzymes, retaining organic phosphates and other vital compounds, removing metabolic waste, and sequestering the reductants required to prevent corrosion by oxygen. During erythropoiesis, the membrane responds to erythropoietin and imports the iron required for the synthesis of haemoglobin. The lipid bilayer provides an impermeable barrier between the cytoplasm and the external environment and helps maintain a slippery exterior so that erythrocytes do not adhere to endothelial cells or aggregate in the microcirculation. The membrane also participates in erythrocyte biogenesis and ageing. Finally, the membrane participates in the maintenance of pH homeostasis by participating in chloride–bicarbonate exchange.

Interactions of membrane proteins and disorders of red cell shape

Membrane protein–protein and protein–lipid interactions have been classified into two categories, vertical and horizontal interactions (Fig. 1). Vertical interactions stabilize the lipid bilayer membrane while horizontal interactions support the structural integrity of erythrocytes after their exposure to shear stress. The interactions between proteins and lipids of the erythrocyte membrane are more complex than this simplistic model, but it serves as a useful starting point for understanding red cell membrane interactions, particularly in membrane-related disorders. According to this model, hereditary spherocytosis (HS) is a disorder of vertical interactions. Although the primary molecular defects in HS are heterogeneous (see below), one common feature of HS erythrocytes is a weakening of the vertical contacts between the skeleton and the lipid bilayer. As a result, the lipid bilayer membrane is destabilized, leading to release of lipids in the form of skeleton-free lipid vesicles, which in turn results in membrane surface area deficiency and spherocytosis. In this model, hereditary elliptocytosis is a defect of horizontal interactions, primarily those involving spectrin dimer self-association. Defects of horizontal interactions disrupt the membrane skeletal lattice leading to elliptocytic shape in mild cases and skeletal instability and cell fragmentation in severe cases.

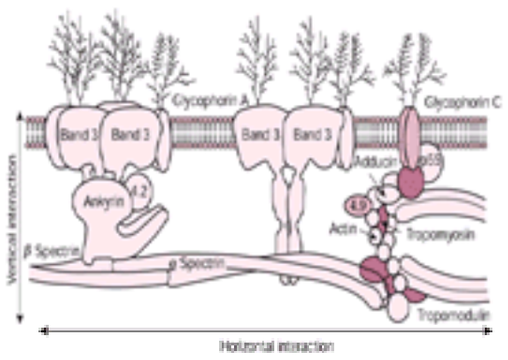

Fig. 1 Schematic diagram of the red cell membrane (not to scale). Membrane–protein and membrane–lipid interactions can be divided into two categories: (1) vertical interactions, which are perpendicular to the plane of the membrane and involve spectrin–ankyrin–band 3 interactions, spectrin–protein 4.1–glycophorin C interactions, and weak interactions between spectrin and the lipid bilayer, and (2) horizontal interactions, which are parallel to the plane of the membrane. (Reprinted from Tse and Lux, 1999, with permission.)

Hereditary spherocytosis

Introduction

Hereditary spherocytosis (HS) refers to a group of inherited disorders characterized by the presence of spherical-shaped erythrocytes on peripheral blood smear. HS occurs in all racial and ethnic groups. It is the most common inherited anaemia in individuals of northern European descent, affecting approximately 1 in 2500 individuals in the United States and England. It is much more common in Caucasians than in individuals of African descent. Clinical, laboratory, biochemical, and genetic heterogeneity characterize the spherocytosis syndromes.

Aetiology and pathogenesis

The primary defect in HS is loss of membrane surface area relative to intracellular volume, accounting for the spheroidal shape and decreased deformability of the red cell. This loss of surface area results from increased membrane fragility due to defects in erythrocyte membrane proteins. Increased fragility leads to membrane

vesiculation and membrane loss. Splenic destruction of these non-deformable erythrocytes is the primary cause of haemolysis experienced by HS patients. Physical entrapment of erythrocytes in the splenic microcirculation and ingestion by phagocytes have been proposed as mechanisms of destruction. Furthermore, the splenic environment is hostile to erythrocytes. Low pH, glucose, and ATP concentrations, and high local concentrations of toxic free radicals produced by adjacent phagocytes, all contribute to membrane damage.

Membrane loss is due to defects in several membrane proteins, including ankyrin, band 3, a-spectrin, b-spectrin, and protein 4.2. Combined spectrin and ankyrin deficiency is the most common defect observed, followed by band 3 deficiency, isolated spectrin deficiency, and protein 4.2 deficiency. The genetic defects underlying HS are heterogeneous. Multiple genetic loci are implicated and various abnormalities, including point mutations, defects in mRNA processing, and gene deletions, have been described. Except for a few rare exceptions, HS mutations are private, that is each individual kindred has a unique mutation.

Clinical features

The clinical manifestations of the spherocytosis syndromes vary widely. The typical picture of HS combines evidence of haemolysis (anaemia, jaundice, reticulocytosis, gallstones, and splenomegaly) with spherocytosis (spherocytes on peripheral blood smear, positive osmotic fragility) and a positive family history (Table 1). Mild, moderate, and severe forms of HS have been defined according to differences in haemoglobin, bilirubin, and reticulocyte counts correlated with the degree of compensation for the haemolysis (Table 2). Initial assessment of a patient with suspected HS should include a family history and questions about history of anaemia, jaundice, gallstones, and splenectomy. Physical examination should seek signs such as scleral icterus, jaundice, and splenomegaly. After diagnosing a patient with HS, family members should be examined for the presence of HS.

HS typically presents in childhood, but may present at any age. In children, anaemia is the most frequent presenting complaint (50 per cent), followed by splenomegaly, jaundice, or a positive family history. Two-thirds to three-quarters of HS patients have incompletely compensated haemolysis and mild to moderate anaemia. The anaemia is often asymptomatic except for fatigue and mild pallor. Jaundice is seen at some time in about half of patients, usually in association with viral infections. When present, it is acholuric, that is there is unconjugated hyperbilirubinaemia without detectable bilirubinuria. Palpable splenomegaly is detectable in most (75–95 per cent) older children and adults. Typically, the spleen is modestly enlarged but it may be massive.

About 20 to 30 per cent of HS patients have 'compensated haemolysis,' that is erythrocyte production and destruction are balanced. Although the erythrocyte life span may only be about 20 to 30 days, these patients adequately compensate for their haemolysis with increased marrow erythropoiesis. They are not anaemic and are usually asymptomatic. Many of these individuals escape detection until adulthood, when they are being evaluated for unrelated disorders or when complications related to anaemia or chronic haemolysis occur. Haemolysis may become severe with illnesses that cause splenomegaly, such as infectious mononucleosis, or may be exacerbated by other factors such as pregnancy. Because of the asymptomatic course of HS in these patients, diagnosis of HS should be considered during evaluation of splenomegaly, gallstones at a young age, or anaemia from viral infection.

Approximately 5 to 10 per cent of HS patients have moderate to severe anaemia. Patients with 'moderately severe' disease typically have a haemoglobin of 6 to 8 g/dl, reticulocytes about 10 per cent, bilirubin 2 to 3 mg/dl, and 40 to 80 per cent of the normal red cell spectrin content. This category includes patients with both dominant and recessive HS and a variety of molecular defects. Patients with 'severe' disease, by definition, have life-threatening anaemia and are transfusion-dependent. They almost always have recessive HS. Most have isolated, severe spectrin deficiency. In addition to the risks of recurrent transfusions, these patients often suffer from haemolytic and aplastic crises and may develop complications of severe uncompensated anaemia including growth retardation, delayed sexual maturation, or aspects of thalassaemic faces.

The parents of patients with recessive HS are clinically asymptomatic and do not have anaemia, splenomegaly, hyperbilirubinaemia, or spherocytosis on peripheral blood smears ('Trait', Table 2). Most have subtle laboratory signs of HS including: slight reticulocytosis and slightly elevated osmotic fragility. The incubated osmotic fragility test is probably the most sensitive measure of this condition, particularly the 100 per cent lysis point (0.43 ± 0.05 g NaCl/dl compared to control 0.23 ± 0.07). It has been estimated that at least 1.4 per cent of the population are silent carriers.

Inheritance

The genes responsible for HS include ankyrin, b spectrin, band 3 protein, a spectrin, and protein 4.2. In approximately two-thirds to three-quarter of HS patients, inheritance is autosomal dominant. In the remaining patients, inheritance is non-dominant due to autosomal recessive inheritance or a *de novo* mutation. Cases with autosomal recessive inheritance are due to defects in either a spectrin or protein 4.2. A surprising number of *de novo* mutations have been reported in the HS genes. A few cases of 'double dominant' HS due to defects in band 3 or spectrin that result in fetal death or severe haemolytic anaemia presenting in the neonatal period have been reported. In general, affected individuals of the same kindred experience similar degrees of haemolysis.

Complications

Gallbladder disease

Chronic haemolysis leads to the formation of bilirubinate gallstones, the most frequently reported complication in HS patients. Although gallstones have been detected in infancy, most occur between 10 and 30 years of age. Management should include interval ultrasonography to detect gallstones, as many patients with cholelithiasis and HS are asymptomatic. Timely diagnosis and treatment will help prevent complications of symptomatic biliary tract disease including biliary obstruction, cholecystitis, and cholangitis.

Haemolytic, aplastic, and megaloblastic crises

Haemolytic crises are usually associated with viral illnesses and typically occur in childhood. They are generally mild and are characterized by jaundice, increased splenomegaly, decreased haematocrit, and reticulocytosis. Intervention is rarely necessary. When severe haemolytic crises occur, there is marked jaundice, anaemia, lethargy, abdominal pain, and tender splenomegaly. Hospitalization and erythrocyte transfusion may be required.

Aplastic crises following virally-induced bone marrow suppression are uncommon, but may result in severe anaemia with serious complications including congestive heart failure or even death. The most common aetiological agent in these cases is parvovirus B19. Parvovirus selectively infects erythropoietic progenitor cells and inhibits their growth. Parvovirus infections are frequently associated with mild neutropenia, thrombocytopenia, or even pancytopenia. During the aplastic phase, the haemoglobin and the production of new red cells fall, the cells that remain age, and microspherocytosis and osmotic fragility increase. Aplastic crises usually last 10 to 14 days (about half the life span of HS red cells), the haemoglobin typically falls to half its usual level before recovery occurs. In patients with severe HS, the anaemia may be profound, requiring hospitalization and transfusion. As the marrow recovers, granulocytes, platelets, and, finally, reticulocytes return to the peripheral blood. Aplastic crisis brings many patients to medical attention, particularly asymptomatic HS patients with normally compensated haemolysis. Because parvovirus may infect several members of a family simultaneously, leading to aplastic crises, there have been reports of 'outbreaks' of HS.

Megaloblastic crisis occurs in HS patients with increased folate demands, for example the pregnant patient, growing children, or patients recovering from an aplastic crisis. With appropriate folate supplementation, this complication is preventable.

Diagnosis

The laboratory findings in HS are heterogeneous. Initial laboratory investigation should include a complete blood count with peripheral smear, reticulocyte count, Coombs' test, and serum bilirubin. When the peripheral smear or family history is suggestive of HS, an incubated osmotic fragility should be obtained. Rarely, additional, specialized testing is required to confirm the diagnosis.

Peripheral blood smear

Erythrocyte morphology is quite variable. Typical HS patients have blood smears with obvious spherocytes lacking central pallor (Fig. 2(a)). Less commonly, patients present with only a few spherocytes on peripheral smear or, at the other end of the spectrum, with numerous small, dense spherocytes and bizarre erythrocyte morphology with anisocytosis and poikilocytosis (Fig. 2(b)). Specific morphological findings have been identified in patients with certain membrane protein defects

such as pincered erythrocytes (band 3) or spherocytic acanthocytes (b spectrin).

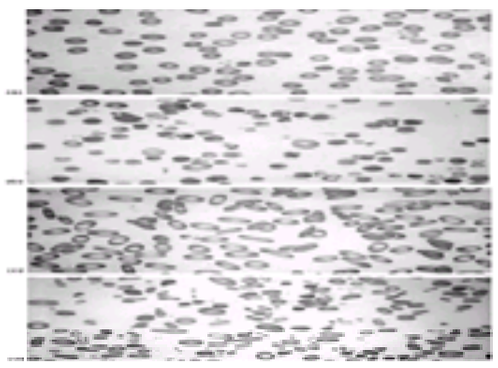

Fig. 2 Peripheral blood smears. (a) Typical hereditary spherocytosis. (b) Severe, recessively-inherited spherocytosis. (c) Hereditary elliptocytosis. (d) Hereditary pyropoikilocytosis.

Erythrocyte indices

Most patients have mild to moderate anaemia. The mean corpuscular haemoglobin concentration is increased (between 35 and 38 per cent) due to relative cellular dehydration in approximately 50 per cent of patients, but all HS patients have some dehydrated cells. The Technicon H1 blood counter and its successors (Technicon, Tarrytown, NY) provide a histogram of mean corpuscular haemoglobin concentration that has been claimed to be accurate enough to identify nearly all HS patients. Finally, the mean corpuscular volume (MCV) is usually normal except in cases of severe HS, when it is slightly decreased.

Osmotic fragility

In the normal erythrocyte, membrane redundancy gives the cell its characteristic discoid shape and provides it with abundant surface area. In spherocytes, there is a decrease in surface area relative to cell volume, resulting in their abnormal shape. This change is reflected in the increased osmotic fragility found in these cells (Fig. 3). Osmotic fragility is tested by adding increasingly hypotonic concentrations of saline to red cells. The normal erythrocyte is able to increase its volume by swelling, but spherocytes, which are already at maximum volume for surface area, burst at higher saline concentrations than normal. Approximately 25 per cent of HS individuals will have a normal osmotic fragility on freshly drawn red cells, with the osmotic fragility curve approximating the number of spherocytes seen on peripheral smear. However, after incubation at 37°C for 24 h, HS red cells lose membrane surface area more readily than normal because their membranes are leaky and unstable. Thus incubation accentuates the defect in HS erythrocytes and brings out the defect in osmotic fragility, making incubated osmotic fragility the standard test for diagnosing HS. When the spleen is present, a subpopulation of very fragile erythrocytes, which have been conditioned by the spleen, form the 'tail' of the osmotic fragility curve; this disappears after splenectomy (Fig. 3). Osmotic fragility testing suffers from poor sensitivity as about 20 per cent of mild cases of HS are missed after incubation. It is unreliable in patients with small numbers of spherocytes, including those who have been recently transfused. It is abnormal in other conditions where spherocytes are present.

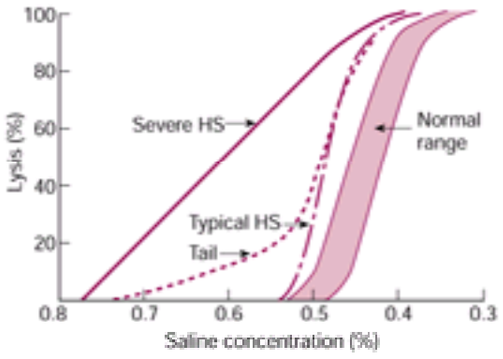

Fig. 3 Osmotic fragility curves in hereditary spherocytosis. The shaded region is the normal range. Results representative of both typical and severe spherocytosis are shown. A tail, representing very fragile erythrocytes that have been conditioned by the spleen, is common in many spherocytosis patients prior to splenectomy.

Additional testing

Other investigations, such as the autohaemolysis test, the hypertonic cryohaemolysis test, and the acidified glycerol test, suffer from lack of specificity and are not widely used. Specialized testing, such as membrane protein quantitation, ektacytometry, and genetic analyses, are available for studying difficult cases or cases where additional information is desired.

Other laboratory manifestations in HS are markers of ongoing haemolysis. Reticulocytosis, increased bilirubin, increased lactate dehydrogenase, increased urinary and faecal urobilinogen, and decreased haptoglobin reflect increased erythrocyte production or destruction.

Differential diagnosis

HS should be able to be distinguished from other haemolytic anaemias by additional diagnostic testing, such as autoimmune haemolytic anaemia via a Coombs' test. Other causes of haemolytic anaemia (Table 3) should be viewed in the appropriate clinical context. Occasional spherocytes are also seen in patients with a large spleen (such as in cirrhosis, myelofibrosis) or in patients with microangiopathic anaemias, but the differentiation of these conditions from HS is not usually difficult.

Treatment

Splenectomy

Splenic sequestration is the primary determinant of erythrocyte survival in HS patients. Thus splenectomy cures or alleviates the anaemia in the overwhelming majority of patients, reducing or eliminating the need for transfusions and decreasing the incidence of cholelithiasis. Postsplenectomy, spherocytosis and altered osmotic fragility persist, erythrocyte lifespan nearly normalizes, and reticulocyte counts fall to normal or near normal levels. Typical postsplenectomy changes, including Howell–Jolly bodies, target cells, and acanthocytes, become evident on peripheral smear. Postsplenectomy, patients with the most severe forms of HS still suffer from shortened erythrocyte survival and haemolysis, but their clinical improvement is striking.

Early complications of splenectomy include local infection, bleeding, and pancreatitis due to injury to the tail of the pancreas incurred during surgery. Overwhelming postsplenectomy infection (OPSI), typically from encapsulated organisms, is an uncommon but significant late complication of splenectomy, especially in the first few years of life. The introduction of pneumococcal vaccines and the promotion of early antibiotic therapy for febrile children who have had a splenectomy have led to decreases in the incidence of OPSI.

Indications for splenectomy

In the past, splenectomy was considered routine in HS patients. However, the risk of OPSI and the recent emergence of penicillin-resistant pneumococci have led to a re-evaluation of the role of splenectomy in the treatment of HS. Considering the risks and benefits, a reasonable approach would be to splenectomize all patients with severe spherocytosis and all patients who suffer from significant signs or symptoms of anaemia including growth failure, skeletal changes, leg ulcers, and extramedullary haematopoietic tumours. Other candidates for splenectomy are older HS patients who suffer vascular compromise of vital organs.

Whether patients with moderate HS and compensated, asymptomatic anaemia should have a splenectomy remains controversial. Patients with mild HS and compensated haemolysis can be followed and referred for splenectomy if clinically indicated. The treatment of patients with mild to moderate HS and gallstones is also debatable, particularly since new treatments for cholelithiasis, including laparoscopic cholecystectomy, endoscopic sphincterotomy, and extracorporal choletripsy, lower the risk of this complication.

When splenectomy is warranted, laparoscopic splenectomy is the method of choice as it results in less postoperative discomfort, shorter hospitalization, and decreased costs. Partial splenectomy via laparotomy has been advocated for infants and young children with significant anaemia associated with HS. The goals of this procedure are to allow for the palliation of haemolysis and anaemia while maintaining some residual splenic immune function. Long-term follow-up data for this procedure are lacking.

Prior to splenectomy, patients should be immunized with vaccines against pneumococcus, *Haemophilus influenzae* type b, and meningococcus, preferably several weeks preoperatively. The use and duration of prophylactic antibiotics postsplenectomy is controversial. Presplenectomy, and in severe cases, postsplenectomy, HS patients should take folic acid to prevent folate deficiency.

Elliptocytosis, pyropoikilocytosis, and related disorders

Introduction

Hereditary elliptocytosis (HE) is characterized by the presence of elliptical or cigar-shaped erythrocytes on peripheral blood smears of affected individuals. The world-wide incidence of HE has been estimated to be 1 in 2000 to 4000 individuals. The true incidence of HE is unknown because most patients are asymptomatic. It is common in individuals of African and Mediterranean ancestry, presumably because elliptocytes confer some resistance to malaria. In parts of Africa, the incidence of HE approaches 1 in 100. HE is typically inherited in an autosomal dominant pattern. Rare cases of *de novo* mutations have been described.

Hereditary pyropoikilocytosis (HPP) is a rare cause of severe haemolytic anaemia with erythrocyte morphology reminiscent of that seen in severe burns. Initial studies of erythrocytes from these patients revealed abnormal thermal sensitivity compared to normal erythrocytes. HPP occurs predominantly in patients of African descent. There is a strong relationship between HPP and HE. Approximately one-third of parents or siblings of patients with HPP have typical HE. Many patients with HPP experience severe haemolysis and anaemia in infancy that gradually improves, evolving toward typical HE later in life.

Aetiology and pathogenesis

The principle defect in HE/HPP erythrocytes is an intrinsic mechanical weakness or fragility of the erythrocyte membrane skeleton due to a defect of horizontal interactions (see above). This is due to defects in the red cell membrane proteins a spectrin, b spectrin, protein 4.1, or glycophorin C. The majority of defects occur in spectrin, the principal structural protein of the membrane skeleton. A variety of mutations in the genes encoding these proteins have been described, with several mutations identified in a number of individuals on the same genetic background, suggesting a 'founder effect' for these mutations.

Clinical features

The clinical presentation of HE is heterogeneous, ranging from asymptomatic carriers to patients with severe, transfusion-dependent anaemia. Most patients with HE are asymptomatic and are typically diagnosed incidentally during testing for unrelated conditions. The erythrocyte life span is normal in most patients. The 10 per cent of patients with decreased red-cell lifespan are the ones who experience haemolysis, anaemia, splenomegaly, and intermittent jaundice. Many of these symptomatic patients have parents with typical HE and thus are homozygotes or compound heterozygotes for defects inherited from each of the parents. Symptomatology may vary between members of the same family, indeed, it may vary in the same individual at different times. To explain these observations, modifier alleles have been hypothesized to influence spectrin expression and clinical severity. One such allele, a ^{LELY} (low expression Lyon), has been identified and characterized.

Diagnosis

The hallmark of HE is the presence of elliptocytes on peripheral blood smear (Fig. 2(c)). These normochromic, normocytic elliptocytes number from a few to 100 per cent. The degree of haemolysis and anaemia do not correlate with the number of elliptocytes present. A few ovalocytes, spherocytes, stomatocytes, and fragmented cells may also be seen. Elliptocytes may be seen in association with several disorders including megaloblastic anaemias, hypochromic microcytic anaemias (iron deficiency anaemia and thalassaemia), myleodysplasic syndromes, and myelofibrosis; however, elliptocytes are generally less than one-third of red cells in these conditions. History and additional laboratory testing usually clarify the diagnosis of these disorders. In addition to the peripheral blood smear findings found in HE, HPP erythrocytes are bizarre-shaped with fragmentation and budding (Fig. 2(d)). Microspherocytosis is common and the MCV is frequently decreased (50–65 mm^3).

The osmotic fragility is abnormal in severe HE and HPP. Other laboratory findings in HE are similar to those found in other haemolytic anaemias and are non-specific markers of increased erythrocyte production and destruction. When indicated, specialized testing, such as membrane protein quantitation, ektacytometry, spectrin analyses, and genetic studies can be performed.

Treatment

Therapy is rarely necessary. In rare cases, occasional red blood cell transfusions may be required. In cases of severe HE and HPP, splenectomy has been palliative. The same indications for splenectomy in HS can be applied to patients with symptomatic HE or HPP. Postsplenectomy, patients with HE or HPP experience increased haemoglobin, decreased haemolysis, and improvement in clinical symptoms.

During acute illnesses, patients should be followed for signs of haematological decompensation. Ultrasonography at regular intervals to detect gallstones should be performed. In patients with significant haemolysis, folate should be administered daily.

South-east Asian ovalocytosis (SAO)

SAO is characterized by the presence of oval erythrocytes with a central longitudinal slit or transverse bar on peripheral blood smears of affected individuals. It is common in parts of the Philippines, Indonesia, Malaysia, and New Guinea and is inherited in an autosomal dominant fashion. Incredibly rigid, SAO erythrocytes are resistant to invasion by malaria parasites. The underlying defect is a mutation in a critical region of band 3. Haematologically, patients with SAO are asymptomatic, with little or no evidence of haemolysis or anaemia. Osmotic fragility is normal. The finding of characteristic ovalocytes in the peripheral blood of an asymptomatic individual from one of the above mentioned ethnic backgrounds is highly suggestive of the diagnosis. Biochemical and DNA diagnostic techniques are available to detect this condition.

Stomatocytosis

The hereditary stomatocytosis syndromes are a heterogeneous group of disorders characterized by mouth-shaped (stomatocytic) erythrocyte morphology on peripheral blood smear (Fig. 4). The clinical severity of stomatocytosis patients is variable; some patients experience haemolysis and anaemia, while others are asymptomatic. The red blood cell membranes of stomatocytosis patients usually exhibit abnormal permeability to the cations sodium and potassium, with consequent modification of intracellular water content, ranging from dehydrated to overhydrated erythrocytes. The underlying defect(s) leading to abnormal cation permeability

and red cell dehydration in these patients is unknown.

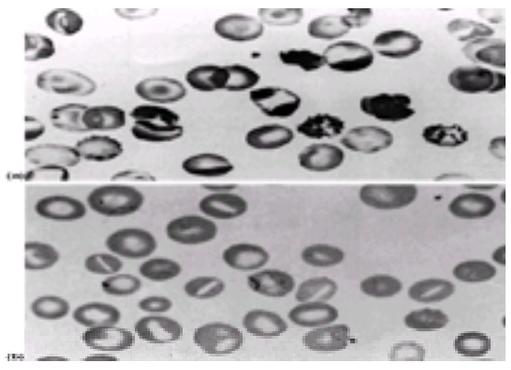

Fig. 4 Peripheral blood smears. (a) Dehydrated stomatocytosis. (b) Overhydrated stomatocytosis. (Reprinted from Lande and Mentzer, 1985, with permission.)

Other conditions

Other conditions associated with hereditary stomatocytosis include the Rh deficiency syndromes and familial deficiency of high-density lipoproteins. Acquired stomatocytosis has been observed in a large number of conditions, particularly hepatobiliary disease and acute alcoholism. Acquired stomatocytosis has also been seen in patients with various malignant neoplasms, cardiovascular disease, and after the administration of vinca alkaloids.

Further reading

Conboy J (1999). The role of alternative pre-mRNA splicing in regulating the structure and function of skeletal protein 4.1. *Proceedings of the Society for Experimental Biology and Medicine* **220**, 73–8.

Delaunay J, Dhermy D (1993). Mutations involving the spectrin heterodimer contact site: clinical expression and alterations in specific function. *Seminars in Hematology* **30**, 21–33.

Delaunay, J, Stewart G, Iolascon A (1999). Hereditary dehydrated and overhydrated stomatocytosis: recent advances. *Current Opinion in Hematology* **6**, 110–4.

Eber SW, Armbrust R, Schroter W (1990). Variable clinical severity of hereditary spherocytosis: relation to erythrocytic spectrin concentration, osmotic fragility, and autohemolysis. *Journal of Pediatrics* **117**, 409–16.

Eber SW, *et al*. (1996). Ankyrin-1 mutations are a major cause of dominant and recessive hereditary spherocytosis. *Nature Genetics* **13**, 214–8.

Gallagher PG, Forget BG, Lux SE (1998). Disorders of the erythrocyte membrane. In: Nathan D, Orkin S, eds. *Hematology of infancy and childhood*, pp. 544–664. WB Saunders, Philadelphia.

Hassoun H, *et al*. (1997). Characterization of the underlying molecular defect in hereditary spherocytosis associated with spectrin deficiency. *Blood* **90**, 398–406.

Jarolim P, *et al*. (1995). Mutations of conserved arginines in the membrane domain of erythroid band 3 lead to a decrease in membrane-associated band 3 and to the phenotype of hereditary spherocytosis. *Blood* **85**, 634–40.

Lande WM, Mentzer WC (1985). Haemolytic anaemia associated with increased cation permeability. *Clinical Haematology* **14**, 89–103.

Lux SE, Palek J (1995). Disorders of the red cell membrane. In: Handin RI, Lux SE, Stossel TP, eds. *Blood: principles and practice of hematology*, pp. 1701–816. JB Lippincott, Philadelphia.

Miraglia del Giudice E, *et al*. (1998). High frequency of de novo mutations in ankyrin gene (ANK1) in children with hereditary spherocytosis. *Journal of Pediatrics* **132**, 117–20.

Morrow JS, *et al*. (1997). Of membrane stability and mosaics: The spectrin cytoskeleton. In: Hoffman J, Jamieson J, eds. *Handbook of physiology*, pp. 485–540. Oxford University Press, London.

Tse WT, Lux SE (1999). Red blood cell membrane disorders. *British Journal of Haematology* **104**, 2–13.

Tse WT, *et al*. (1997). Amino-acid substitution in alpha-spectrin commonly coinherited with nondominant hereditary spherocytosis. *American Journal of Hematology* **54**, 233–41.

Wichterle H, *et al*. (1996). Combination of two mutant alpha spectrin alleles underlies a severe spherocytic hemolytic anemia. *Journal of Clinical Investigation* **98**, 2300–7.

Wilmotte R, *et al*. (1993). Low expression allele alpha LELY of red cell spectrin is associated with mutations in exon 40 (aV/41 polymorphism) and intron 45 and with partial skipping of exon 46. *Journal of Clinical of Investigation* **91**, 2091–6.

Yawata Y (1994). Red cell membrane protein band 4.2: phenotypic, genetic and electron microscopic aspects. *Biochimica et Biophysica Acta* **16**, 131–48.

22.5.11 Erythrocyte enzymopathies

Ernest Beutler

Erythrocytes are living cells that contain a large number of enzymes required to carry out a variety of metabolic processes. Some inherited deficiencies of these enzymes are called red-cell enzymopathies. They may cause haematological disorders, including haemolytic anaemias, polycythaemia, and methaemoglobinaemia. Other deficiencies do not produce haematological disorders, but instead mirror important metabolic disorders such as galactosaemia and are therefore of diagnostic value. Some deficiencies, for example those of lactate dehydrogenase or inosine triphosphatase (ITPase) are, as far as has been determined, 'non-diseases'.

This section deals with those red-cell enzyme defects that cause haemolytic anaemia. Many have been described; most are rare but some are sufficiently common that several hundred cases have been documented. Although the enzymatic bases of these defects are very different, the clinical presentation is similar and relatively nondescript. It is impossible to differentiate the enzymatic defects from one another by clinical or routine laboratory methods.

Red-cell metabolism

The two major pathways of red-cell glucose metabolism are illustrated in (Fig. 1).

Glucose is phosphorylated to glucose-6-phosphate in the hexokinase reaction. It is then either metabolized in the anaerobic Embden–Myerhoff pathway or is oxidized in the glucose-6-phosphate dehydrogenase (G6PD) reaction, entering the hexose monophosphate pathway.

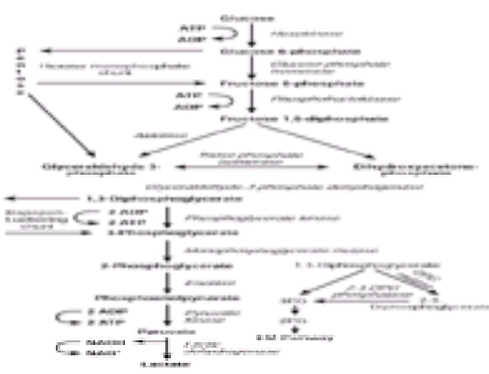

Fig. 1 The relationship between the main red-cell glycolytic pathway (Embden–Meyerhof) and the other metabolic pathways. The insert shows the production of 2,3-DPG in the Rapoport–Luebering shunt.

Anaerobic metabolism of glucose reduces NAD and, by phosphorylating ADP to ATP, provides energy to maintain erythrocyte shape and to transport molecules into and out of erythrocyte. The hexose monophosphate pathway serves to maintain glutathione and protein sulphydryl groups in the reduced state. These pathways are similar in red cells, as in other tissues and in lower organisms. However, the 2,3-diphosphoglycerate (2,3 DPG) shunt is a unique feature of the Embden–Myerhoff pathway in erythrocytes. This 'energy clutch' of erythrocyte metabolism not only allows flexibility in the amount of ATP that is generated in glycolysis, but also provides a source of 2,3-DPG, the key modulator of haemoglobin oxygen affinity. The pathway also reduces NAD to NADH, which serves to reduce methaemoglobin to haemoglobin. The hexose monophosphate pathway, on the other hand, reduces NADP to the NADPH needed to maintain sulphydryl compounds in the reduced state. There are, in addition, many other metabolic functions that the erythrocyte must carry out. Among these are the synthesis of glutathione, the synthesis and degradation of nucleotides and nucleosides, the detoxification of active oxygen radicals, and the transport of small molecules into and out of the cell.

Genetics

Half of the normal activity of red-cell enzymes is generally sufficient for normal function. Thus, haemolytic anaemias due to red-cell enzyme deficiencies occur as autosomal recessive or sex-linked disorders. Only two of the deficiencies, those of G6PD and phosphoglycerate kinase are encoded by genes on the X chromosome; all of the others are autosomal. Extensive mutation analysis at the DNA level has been performed on patients with some of the enzyme defects; notably, deficiencies of G6PD, pyruvate kinase, glucosephosphate isomerase, and triosephosphate isomerase. The vast majority of mutations of the genes encoding these enzymes are missense mutations or nonsense mutations, but a few deletions, insertions, and splicing mutations have been described. So far, with the possible exception of one mutation of triosephosphate isomerase, no regulatory mutations have been documented. DNA-based diagnosis has assumed an increasingly valuable role in the diagnosis of these disorders.

Specific red-cell abnormalities that may cause haemolytic anaemia

Table 1 summarizes a some of the clinical and genetic characteristics of red-cell enzyme deficiencies.

The more common red-cell enzyme abnormalities

G6PD deficiency

This enzymopathy is described in Chapter 22.5.12.

Pyruvate kinase deficiency

Pyruvate kinase deficiency can be considered the clinical prototype of the non-spherocytic haemolytic anaemias caused by red-cell enzymopathies. The severity of the anaemia varies greatly from patient-to-patient. At one extreme, the anaemia can be quite mild; at the other, the patient may be entirely transfusion-dependent. Indeed, the circulating red cells in such patients may have normal pyruvate kinase activity because there are scarcely any of the patient's own cells present; it appears that most of the patient's cells are destroyed before they leave the marrow and it may be only the transfused cells that are sampled and sent to the laboratory

for diagnosis. Pyruvate-kinase-deficient patients have the usual stigmata of haemolytic anaemia; that is, pallor, lack of energy, jaundice, and sometimes gallstones. In those patients who are transfusion-dependent, haemochromatosis occurs with some frequency, probably more so than in patients with many other types of haemolytic anaemia. Patients with pyruvate kinase deficiency usually enjoy a fairly good response to splenectomy. This response is less complete than is observed in hereditary spherocytosis, but may be clinically quite helpful, particularly in reducing the requirement for transfusions.

Pyruvate kinase deficiency is probably the most difficult of all of the red-cell enzymopathies to diagnose, because the enzyme is a complex one with allosteric properties. The residual enzyme activity is not always greatly reduced. Cases have been described in which the residual pyruvate kinase activity is actually higher than is found in normal individuals. In such cases, establishing the diagnosis may depend upon showing that the level of 2,3 DPG or of 3-phosphoglyceric acid in the erythrocytes is greatly elevated, a finding that is characteristic of pyruvate kinase deficiency. It is also useful to measure the thermal stability of the residual enzyme; mutant enzymes are very often unstable on heating. Many different mutations have been documented in patients with pyruvate kinase deficiency. In European populations, the most common of these is a G® A mutation at nucleotide 1529 coding for a Arg®Gly substitution at amino acid 510. This mutation has not been detected among Asians with pyruvate kinase deficiency; among Gypsies the characteristic mutation is a deletion of exon 11.

The less common red-cell enzyme abnormalities

Glucosephosphate isomerase deficiency

Patients with glucosephosphate isomerase deficiency generally have a milder haemolytic disorder than patients with pyruvate kinase deficiency. The response to splenectomy is usually satisfactory. Although milder in general, this enzymopathy seems to be associated with hydrops fetalis more frequently than the other red-cell enzyme defects. Diagnosis is generally straightforward. A fluorescent screening test can be used to detect the deficiency. Several different mutations have been documented. With few exceptions, they are different in each family.

Pyrimidine 5'-nucleotidase deficiency

Basophilic stippling is the hallmark of pyrimidine 5'-nucleotidase deficiency. Interestingly, this enzyme is very sensitive to inhibition by lead. The stippling that is so characteristic of lead poisoning may be the consequences of inhibition of this enzyme. Pyrimidine 5'-nucleotidase is the most age-sensitive all of the red-cell enzymes; this one alone is decreased in activity in aplastic anaemia or other disorders, such as transient erythroblastopenia of childhood, in which the mean red cell age is greatly increased. This can lead to misdiagnosis; while it is not uncommon to encounter enzyme activities of one-half normal in patients with decreased erythropoiesis, these patients do not suffer from clinically significant pyrimidine 5'-nucleotidase deficiency. Accumulation of pyrimidine nucleotides, which can be documented by measuring the ultraviolet absorption spectrum, does not occur in such patients.

Triosephosphate isomerase deficiency

Triosephosphate isomerase deficiency is the most devastating all of the red-cell enzymopathies. With few exceptions, patients with this abnormality die by the time they are 4 years of age. All tissues are affected, and death is usually due to cardiopulmonary complications. It is been suggested, on the basis of enzyme activities and genetic studies, that the heterozygous state for this deficiency is very common among African-Americans. This has not been confirmed. Many different mutations have been detected in patients with triosephosphate isomerase deficiency; one, at genomic nucleotide 1591, accounts for approximately 50 per cent of the patients with this disorder. Polymorphic changes occur in the promoter region of the triosephosphate isomerase gene, but the significance of these mutations is not yet clear. No treatment has been effective.

The rare red-cell enzyme deficiencies

Hexokinase deficiency

Hexokinase deficiency is one of the more difficult red-cell enzymopathies to diagnose, because the activity of this enzyme is much higher in young red cells than in older erythrocytes. As a result, red-cell hexokinase activity is usually increased in patients with haemolytic anaemia of any type. In patients with hexokinase deficiency, this often gives rise to the anomalous finding that the red-cell hexokinase activity in the affected patient is normal, usually higher than that found in the heterozygous parents. The diagnostic hallmark is normal, rather than elevated, hexokinase activity in the face of a high reticulocyte count and high levels of other red-cell enzymes.

Enzymes of glutathione synthesis

Erythrocytes synthesize glutathione from the amino acids glutamate, cysteine, and glycine in two consecutive enzymatic reactions, each of which utilizes ATP. In the first step, catalysed by g-glutamylcysteine synthetase, a peptide bond is formed between the g-carboxyl group of glutamic acid and cysteine. Several patients deficient in this enzyme have been found. In addition to haemolytic anaemia, spinocerebellar degeneration was documented in the initial patient described, but neurological symptoms have not been present in subsequent patients. Defects of the second step of glutathione synthesis, the formation of a peptide link between g-glutamyl-cysteine and glycine, catalysed by the enzyme glutathione synthetase, appear in two clinical forms. In some patients, the deficiency is limited to the erythrocytes. Haemolytic anaemia appears to be the sole clinical manifestation. In other patients, the deficiency is generalized. These patients excrete large amounts of pyroglutamic acid (5-oxyproline); this product of g-glutamylcysteine degradation is overproduced in the absence of the feedback inhibition of g-glutamylcysteine synthetase by glutathione. Patients with the generalized defect have severe neuromuscular manifestations in addition to haemolytic anaemia.

Glutathione reductase deficiency

Only a single family with a severe, hereditary deficiency of glutathione reductase has been described. No haemolysis was present except after the ingestion of fava beans. Low activity of red-cell glutathione reductase, a flavin enzyme, are found when the intake of riboflavine is suboptimal, but this mild or moderate enzyme deficiency has no clinical consequences.

Phosphofructokinase kinase deficiency

Erythrocytes contain two types of genetically distinct phosphofructokinase subunits, L (liver) and M (muscle). Deficiency of the M subunit causes haemolysis, but the haemoglobin level in the blood is often normal or even higher than normal because of the diminished 2,3 DPG levels that are characteristic of this disorder. Muscle enzyme activity is also compromised and a myopathy results. This disorder is sometimes designated Tarui disease or type VII glycogenosis. Deficiency of the L subunit of phosphofructokinase has also been reported, but did not have any clinical consequences.

Aldolase deficiency

A few cases of aldolase deficiency been reported. An association with mental retardation was noted in one case, but it is not clear whether a cause-and-effect relationship exists.

Phosphoglycerate kinase deficiency

Phosphoglycerate kinase shares with G6PD deficiency the distinction of being an X-linked enzymopathy. In addition to haemolytic anaemia, behavioural disturbances have been noted.

Diphosphoglycerate mutase deficiency

The result of diphosphoglycerate mutase deficiency is more frequently erythrocytosis than haemolytic anaemia, because a lack of this enzyme prevents the formation of 2,3 DPG. Consequently, the oxygen affinity of the red cells is increased, stimulating erythropoiesis.

High adenosine deaminase activity

Haemolytic anaemia, inherited as an autosomal dominant disorder, has rarely been found to be associated with greatly elevated red-cell adenosine deaminase levels. The adenosine deaminase that is formed appears to be normal. The abnormality that causes this tissue-specific increase in enzyme activity has not yet been discovered.

Adenylate kinase deficiency

A number of patients with familial haemolytic anaemia have been documented to have markedly decreased levels of red-cell adenylate kinase. However, one very well-studied patient with virtually absent enzyme activity had no clinical disorder. The relationship between this enzyme deficiency and haemolytic anaemia remains unclear.

Specific red-cell abnormalities that do not cause haemolytic anaemia

Severe deficiencies of many red-cell enzymes do not produce haematological abnormality or, indeed, in many cases, no clinical abnormality of all. Included are deficiencies of 6-phosphogluconate dehydrogenase, †-aminolevulinic acid dehydrase, acetylcholinesterase, AMP deaminase, carbonic anhydrase, catalase, galactokinase, galactose-1-phosphate uridyltransferase, glutathione peroxidase, hypoxanthine-guanine phosphoribosyltransferase, ITPase, and phosphoglucomutase. Discussion of these enzyme deficiencies is beyond the scope of this chapter.

Diagnosis

The diagnosis of red-cell enzymopathies has been carried out at four levels: morphological observations, study of autohaemolysis, quantification of red-cell enzyme activity, and DNA analysis.

Morphological observations

The appearance of erythrocytes on a stained blood film may be useful in determining whether haemolytic anaemia is present and in ruling out some causes of haemolysis, such as hereditary spherocytosis, ovalocytosis, or microangiopathic haemolytic anaemia. The presence of prominent red-cell stippling suggests a diagnosis of pyrimidine 5' nucleotidase deficiency.

The autohaemolysis test

The autohaemolysis test is performed by incubating sterile, whole blood with and without glucose for 24 h and observing the degree to which the red cells are lysed. This test outlived its usefulness as a tool for differentiating enzymatic cause of haemolytic anaemia many years ago. Although it is true that the haemolysis of pyruvate-kinase-deficient red cells occurring *in vitro* after incubation for 24 h is not usually corrected by glucose, this is by no means always the case, nor is this pattern specific for pyruvate kinase deficiency.

Qualitative and quantitative estimations of red-cell enzyme activity

The most generally useful means for differentiating red-cell enzyme defects from one another and from defects other than known enzyme deficiencies is to semiquantitate or quantitate the red-cell enzyme activities. Fluorescent screening tests have been developed that allow the non-specialized laboratory to detect decreases in the activity of enzymes such as G6PD, pyruvate kinase, glucosephosphate isomerase, or triosephosphate isomerase with a high degree of reliability. The accumulation of pyrimidine nucleotides can be detected by measuring the ultraviolet spectrum of a perchloric acid extract of red cells. This can be used by non-specialized laboratories to detect this abnormality.

Quantification of red-cell enzyme activities is a more specialized task that can be accomplished by the use of standardized techniques in an experienced laboratory. There are a number of caveats that must be taken into account, both with respect to the performance of red-cell enzyme assays and the interpretation of the results. Leucocyte pyruvate kinase and red-cell pyruvate kinase are encoded by different genes. Moreover, the activity of the white cell enzyme is very high. Thus, contamination of a red-cell suspension with a relatively small number of white cells may obscure the diagnosis of red-cell pyruvate kinase deficiency. The interpretation of the results of red-cell enzyme assays may also be confounded by the fact that the blood of patients with haemolytic anaemia is enriched with reticulocytes and young erythrocytes. Since many of the mutations that cause red-cell enzymopathies result in the production of unstable enzymes, the young erythrocytes that circulate may actually contain normal or near-normal levels of enzyme. It is therefore essential to take into account the age of the circulating cells. It may be helpful to obtain blood samples from parents or children of the patient to determine whether half normal activities can be documented.

Problems in interpretation may also arise when the activity of an enzyme as measured *in vitro* does not accurately reflect its intracellular *in vivo* activity. This comes about because of the necessity of using unphysiologically high substrate concentrations for *in vitro* assays. This difficulty is particularly prone to arise in the case of pyruvate kinase deficiency, because this is a complex allosteric enzyme that not only has binding sites for two substrates, ADP and phosphoenolpyruvate, but also for fructose diphosphate, an allosteric effector.

Finally, there is the confounding effect of red-cell transfusions. It is clearly best to wait until just before a transfusion to draw blood for testing.

DNA-based diagnosis

With the development of PCR-based technologies for the detection of mutations, and for the sequencing of DNA, mutation analysis at the DNA level has played an increasing role in the diagnosis of red-cell enzyme defects. DNA is extracted from peripheral blood leucocytes and the exons of the gene of interest are amplified. Alternatively, RNA may be reverse transcribed and the cDNA amplified. Because the stability of DNA is greater than that of RNA, and samples may need to be transported to distant, specialized laboratories, direct DNA amplification of genomic DNA rather than of cDNA is generally the preferred technology. DNA-based diagnosis is not particularly difficult to perform in laboratories experienced with the techniques involved. It has some advantages over enzyme assay-based diagnosis. First of all, DNA is very stable, even before it is purified. Therefore, shipping of blood is less of a logistical problem. Transfused red cells do not pose a problem in performing DNA-based diagnosis, since transfused leucocytes do not persist in the circulation. Once the mutation has been established, family studies are more readily performed; heterozygote detection using quantitative enzyme levels is often of a dubious reliability. Prenatal diagnosis, too, is more readily accomplished utilizing DNA-based diagnosis.

There are some major disadvantages in DNA-based diagnosis of red-cell enzymopathies. While it is quite straightforward to identify a known mutation in the coding region of one of the red-cell enzymes, doing so by examining the genes that encode all of the enzymes that may be involved as a cause of haemolytic anaemia would be a daunting task. Even if the entire coding region of the enzyme is sequenced, one cannot be certain that the mutation was not be in a promoter, an enhancer, or in a splice site.

A general approach to diagnosis of red-cell enzymopathies

The first step is to make certain that the patient has a haemolytic anaemia. The reticulocyte count should be elevated, unless it has been temporarily suppressed by infection. If the patient's history suggests that the anaemia is chronic in nature, a positive family history can be very helpful. Dominant inheritance suggests that an enzymopathy is not the cause; only the very rare anaemia caused by elevated adenosine deaminase levels falls into this category. Instead, dominant inheritance suggests that the patient either has an unstable haemoglobin or hereditary spherocytosis. Sex-linked inheritance may also appear as though it is dominant, but is excluded if there is father-to-son transmission. G6PD deficiency and phosphoglycerate kinase deficiency are the only red-cell enzymopathies that are sex-linked. Often there is no clear-cut family history. Before trying to establishing whether or not a red-cell enzymopathy is present, hereditary spherocytosis, haemoglobinopathies, and other disorders, such as a paroxysmal nocturnal haemoglobinuria, should be excluded.

Fluorescent screening tests are appropriate starting points for the diagnosis of the red-cell enzymopathies. Screening tests for G6PD deficiency, pyruvate kinase

deficiency, glucosephosphate isomerase deficiency should be carried out. If the patient is a child with neuromuscular disease, a fluorescent test for triosephosphate isomerase deficiency is also indicated. Stippling of the red cells suggests that the patient may have pyrimidine 5' nucleotidase deficiency. In this instance the ultraviolet spectrum of a perchloric acid of extract of the red cells should be examined. A clear-cut positive screening test for one of the red-cell enzymopathies, carried out with appropriate controls, is adequate for diagnosis. Quantitative assays for red-cell enzymes can be performed by specialized laboratories and the may include those enzymes for which no screening tests have been developed.

When a diagnosis has been established, either by performing a screening test or by quantitative assay, it is sometimes useful to identify the mutation at the DNA level. This need not be done in every case, but is particularly useful in the case of young couples who hope to have more children and desire genetic counselling and prenatal diagnosis.

Further reading

Baronciani L, Bianchi P, Zanella A (1998). Hematologically important mutations: Red cell pyruvate kinase (2nd update). *Blood Cells, Molecules, and Diseases* **24**, 271–7. [Compilation of the PK mutations.]

Baronciani L, Zanella A, Bianchi P, *et al*. (1996). Study of the molecular defects in glucose phosphate isomerase-deficient patients affected by chronic hemolytic anemia. *Blood* **88**, 2306–10. [The GPI gene and mutations that affect it.]

Beutler E, Blume KG, Kaplan JC, *et al*. (1977). International committee for standardization in haematology: Recommended methods for red-cell enzyme analysis. *British Journal of Haematology* **35**, 331–40. [Standardized methods for the enzymatic diagnosis of red cell enzymopathies.]

Bianchi M, Magnani M (1995). Hexokinase mutations that produce nonspherocytic hemolytic anemia. *Blood Cells, Molecules, and Diseases* **21**, 2–8. [Description of the first hexokinase mutation at the DNA level.]

Fujii H, Miwa S (1999). Red blood cell enzymes and their clinical application. *Advances in Clinical Chemistry* **33**, 1–54. [A good review.]

Jacobasch G, Rapoport SM (1996). Hemolytic anemias due to erythrocyte enzyme deficiencies. *Molecular Aspects Medicine* **17**, 143–70. [A good review.]

Lenzner C, Nürnberg P, Jacobasch G, *et al*. (1997). Molecular analysis of 29 pyruvate kinase-deficient patients from Central Europe with hereditary hemolytic anemia. *Blood* **89**, 1793–9. [Clinical and molecular analysis of PK mutations.]

Lestas AN, Kay LA, Bellingham AJ (1987). Red cell 3-phosphoglycerate level as a diagnostic aid in pyruvate kinase deficiency. *British Journal of Haematology* **67**, 485–8. [The value of intermediate levels in the diagnosis of pyruvate kinase deficiency.]

Schneider A, Cohen-Solal M (1996). Hematologically important mutations: Triosephosphate isomerase. *Blood Cells, Molecules, and Diseases* **22**, 82–4. [Compilation of the TPI mutations]

Schneider A, Forman L, Westwood B, *et al*. (1998). The relationship of the -5, -8, and -24 mutations in African-Americans to triosephosphate isomerase (TPI) enzyme activity and to TPI deficiency. *Blood* **92**, 2959–62. [The TPI promoter mutations and their effect on enzyme activity.]

Tarui S, Okuno G, Ikura Y, *et al*. (1965). Phosphofructokinase deficiency in skeletal muscle. A new type of glycogenosis. *Biochemical and Biophysical Research Communications* **19**, 517–23. [Original description of phosphofructokinase deficiency as a glycogen storage disease.]

Valentine WN, Paglia DE (1990). Erythroenzymopathies and hemolytic anemia: The many faces of inherited variant enzymes. *Journal of Laboratory and Clinical Medicine* **115**, 12–20. [A good review.]

22.5.12 Glucose-6-phosphate dehydrogenase (G6PD) deficiency

Lucio Luzzatto

Definition

Glucose-6-phosphate dehydrogenase (**G6PD**) is a key enzyme in redox metabolism. G6PD deficiency is an inherited condition in which red cells have a markedly decreased activity of G6PD, which predisposes to haemolytic anaemia.

Epidemiology

G6PD deficiency is distributed worldwide. Areas of high prevalence are found in Africa, Southern Europe, the Middle East, South-East Asia, and Oceania. In the Americas and in parts of Northern Europe, G6PD deficiency is also quite prevalent as a result of migrations that have taken place in relatively recent historical times.

Genetics

The inheritance of G6PD deficiency has long been known to have a mendelian X-linked pattern, and the gene encoding G6PD has been mapped to the telomeric region of the long arm of the X chromosome (band Xq28), physically very close to the genes for haemophilia A, dyskeratosis congenita, and colour blindness. At the genomic level, the G6PD gene consists of 13 exons and spans some 18.5 kb (kilobases). Structural and functional studies have revealed features of a 'housekeeping gene'; this is in accord with the fact that G6PD is found in all cells.

X linkage of the G6PD gene has important implications. First, as males have only one G6PD gene (being hemizygous for this gene), they must be either normal or G6PD deficient. By contrast, females, having two G6PD genes, can be either normal or deficient (homozygous), or intermediate (heterozygous). Moreover, as a result of the phenomenon of X-chromo-some inactivation, heterozygous females are genetic mosaics, and this in turn has clinical implications. Indeed, in most other (autosomal) enzyme deficiencies, heterozygotes are asymptomatic because cells with an enzyme level close to 50 per cent of normal are biochemically normal. But in the case of G6PD, as a result of X inactivation, the abnormal cells of a woman heterozygous for G6PD deficiency are just as deficient as those of a hemizygous deficient man, and therefore just as susceptible to pathology. Thus, although G6PD deficiency is still often referred to as an X-linked recessive trait, this is a mismomer because a recessive trait is, by definition, not expressed in a heterozygote; instead, G6PD deficiency is expressed in heterozygotes both biochemically and clinically—although it is true that heterozygotes are generally less severely affected.

Clinical manifestations

Acute haemolytic anaemia

In view of the large number of people who carry a G6PD deficiency gene, it is fortunate that the vast majority remain clinically asymptomatic throughout their lifetime. However, they are all at risk of developing acute haemolytic anaemia in response to three types of triggers: (i) drugs (Table 1), (ii) infections, and (iii) broad (fava) beans. Typically, a haemolytic attack starts with malaise, sometimes associated with more or less profound weakness, and abdominal or lumbar pain. After an interval of several hours to 2 or 3 days (usually the onset is more abrupt in children) the patient develops jaundice and dark urine, due to haemoglobinuria. In the majority of cases the haemolytic attack, even if severe, is self-limiting and tends to resolve spontaneously. In the absence of additional or pre-existing pathology the bone marrow response is prompt and effective. Depending on the proportion of red cells that have been destroyed (reflected in the severity of the anaemia), the haemoglobin level may be back to normal in 3 to 6 weeks. The most serious threat in adults is the development of acute renal failure (this is exceedingly rare in children). The anaemia is usually normocytic and normochromic, and it varies from moderate to extremely severe (haemoglobin levels of 4 g/dl or less have been recorded); it is due largely to intravascular haemolysis, and hence it is associated with haemoglobinaemia, haemoglobinuria, and low or absent plasma haptoglobin. The blood film shows anisocytosis, polychromasia, and other features associated with acute haemolysis, including spherocytes (Fig. 1); in severe cases the poikilocytosis is very marked, with bizarre forms, numerous red cells that appear to have unevenly distributed haemoglobin ('hemighosts'), and red cells that appear to have had parts of them bitten away ('bite cells' or 'blister cells'). Supravital staining with methyl violet, if done promptly, reveals the presence of 'Heinz bodies', consisting of precipitates of denatured haemoglobin (Fig. 1; apart from the rare cases when they are formed because of a genetic haemoglobin abnormality, Heinz bodies can be regarded as a signature of oxidative damage to red cells). The white blood cell count may be elevated, with predominance of granulocytes. The platelet count may be normal, increased, or moderately decreased. The unconjugated bilirubin is elevated but the 'liver enzymes' are usually normal.

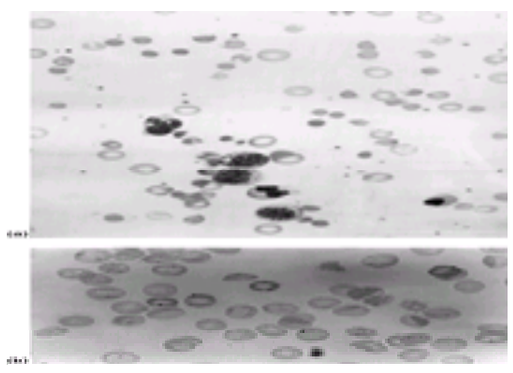

Fig. 1 Blood film in a case of acute haemolytic anaemia in a G6PD-deficient patient (favism). (a) Romanovsky stain, showing marked poikilocytosis, polychromatic macrocytes, bite cells, nucleated red cells, and a shift to the left in the granulocytic series. (b) Supravital stain with methyl violet, showing the characteristic Heinz bodies.

Favism

This is perhaps the most spectacular form of acute haemolytic anaemia associated with G6PD deficiency; it can occur at any age, but more commonly in children. The

child initially becomes very fractious; then he may develop fever, abdominal pain, diarrhoea, and sometimes vomiting; then he may become lethargic. Haemoglobinuria develops within 6 to 24 h from the onset of symptoms. Physical examination reveals pallor, tachycardia, jaundice, and an enlarged spleen; in severe cases there may be evidence of hypovolaemic shock or, more rarely, of high-output heart failure. The cause of favism is the presence in broad beans (or fava beans: *Vicia faba*) of vicine and convicine, two b-glycosides having as aglycones the substituted pyrimidines divicine and isouramil, which produce free radicals in the course of their auto-oxidation. Thus, haemolysis is highly specific for broad beans; other beans are safe. G6PD-deficient subjects (especially when they are adults) do not develop an acute attack of favism every time they eat broad beans; the reasons for this are not yet clear, but important factors are the quantity and quality of broad beans consumed. On the other hands, the widespread notion that favism occurs only with some G6PD-deficient variants and not with others is incorrect. For instance, favism has now been well documented with the 'African' variant A–, and even with G6PD Seattle, a variant associated with milder enzyme deficiency.

Neonatal jaundice

Not every G6PD-deficient baby becomes jaundiced after birth; however, the risk of developing neonatal jaundice is much greater in G6PD-deficient than in G6PD-normal neonates. The extent of the association between G6PD deficiency and neonatal jaundice appears to vary greatly in different populations. The clinical picture of neonatal jaundice related to G6PD deficiency differs from the 'classic' Rhesus-related neonatal jaundice in two main respects: (i) it is very rarely present at birth, and the peak incidence of clinical onset is between day 2 and 3; and (ii) jaundice is more prominent than anaemia, and the anaemia is very rarely severe. The severity of G6PD-related neonatal jaundice varies enormously, from subclinical to overlapping with 'physiological jaundice' to imposing the threat of kernicterus if not treated. The reasons for this are not clear, but prematurity, infection, and environmental factors (for instance, naphthalene–camphor balls used in babies' bedding and clothing) certainly play a part in making neonatal jaundice more severe and more dangerous. From the point of view of public health, it is important to realize that in some parts of the world G6PD deficiency is the commonest cause of severe neonatal jaundice; in addition, if not correctly managed, severe neonatal jaundice can produce permanent neurological damage.

Chronic non-spherocytic haemolytic anaemia

In contrast to the large majority of G6PD-deficient subjects who have minimal and subclinical haemolysis in the steady state, a small minority have chronic anaemia of very variable severity. The patient is always male, and in general he presents because of unexplained jaundice. Frequently the onset is at birth, and a diagnosis is made of neonatal jaundice (Fig. 2), which may be severe enough to require exchange transfusion. Subsequently the anaemia recurs and the jaundice fails to clear completely; or the patients is only reinvestigated much later in life, perhaps because of gallstones in a child or in a young adult. Usually the spleen is moderately enlarged in small children, and subsequently it may increase in size sufficiently to cause mechanical discomfort, or hypersplenism, or both. The severity of anaemia ranges in different patients from borderline to transfusion dependent. The anaemia is usually normochromic but somewhat macrocytic; because a large proportion of reticulocytes (up to 20 per cent or more) will cause an increased mean corpuscle volume and a shifted, wider than normal, size-distribution curve. The red cell morphology is not characteristic, and for this reason it is referred to in the negative as being 'non-spherocytic'. The bone marrow is normoblastic, unless the increased requirement of folic acid associated with the high red cell turnover has caused it to become megaloblastic. There is chronic hyperbilirubinaemia, decreased haptoglobin, and increased lactate dehydrogenase. In this condition, unlike in the acute haemolytic anaemia described above, haemolysis is mainly extravascular. However, the red cells of these patients are naturally also vulnerable to acute oxidative damage, and therefore the same agents that can cause acute haemolytic anaemia in people with the ordinary type of G6PD deficiency will cause severe exacerbations with (sometimes massive) haemoglobinuria in people with the severe form of G6PD deficiency.

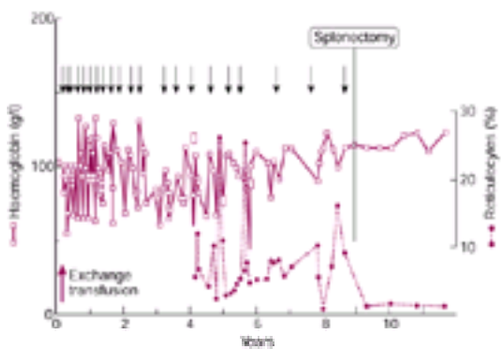

Fig. 2 Clinical course of a patient with chronic non-spherocytic haemolytic anaemia caused by severe G6PD deficiency, illustrating the high transfusion requirement, which was alleviated after splenectomy.

Laboratory diagnosis

Although the clinical picture of favism and of other forms of acute haemolytic anaemia associated with G6PD deficiency is quite characteristic, the final diagnosis must rely on the direct demonstration of decreased activity of this enzyme in red cells. With neonatal jaundice and chronic non-spherocytic haemolytic anaemia the differential diagnosis is much wider, and therefore this test is even more important. The most popular screening tests are the dye decolorization test, the methaemoglobin reduction test, and the fluorescence spot test. Any of these, provided it is properly standardized and subjected to quality control, is perfectly adequate for diagnostic purposes in patients who are in the steady state; but these semiquantitative tests are not adequate for patients in the acute haemolytic or in the posthaematolytic period, or with other complications; nor can they be expected to identify all heterozygotes. Ideally, every patient found to be G6PD deficient by screening should then be retested for confirmation by a quantitative assay. In normal red cells the range of G6PD activity, measured at 30°C, is 7 to 10 iu/g of haemoglobin. In G6PD-deficient males (or homozygous females) the level of G6PD in the steady state is, by definition, less than 50 per cent of normal; but with most variants it is less than 20 per cent and with some it is practically undetectable. In heterozygous females the level is intermediate and extremely variable; therefore, in some cases the diagnosis may be difficult without family studies or DNA analysis. However, for practical purposes it is most unlikely that a woman will have clinical manifestations if her G6PD level is more than 70 per cent of normal.

Biochemistry and pathophysiology

Red cells are very vulnerable to oxidative damage for two reasons. First, oxygen radicals are generated continuously from within the red cells as haemoglobin cycles from its deoxygenated to its oxygenated form. Second, red cells are directly exposed to a variety of exogenous oxidizing agents. Oxygen radicals produced by such compounds are converted by superoxide dismutase to hydrogen peroxide, which is itself highly toxic. G6PD, the first enzyme of the pentose phosphate pathway (Fig. 3), catalyses the conversion of glucose-6-phosphate (G6P) and NADP to 6-phosphogluconolactone and NADPH. The most important product of the G6PD reaction, certainly in red cells, is NADPH because, by producing glutathione via glutathione reductase, it is crucial for the operation of glutathione peroxidase; in addition, it stabilizes catalase: these are the two enzymes able to detoxify hydrogen peroxide (by converting it to water). Normally, G6PD activity in red cells is such that NADPH is maintained at a high level and there is practically no NADP: the NADPH/NADP ratio plays a large part in the intracellular regulation of G6PD activity.

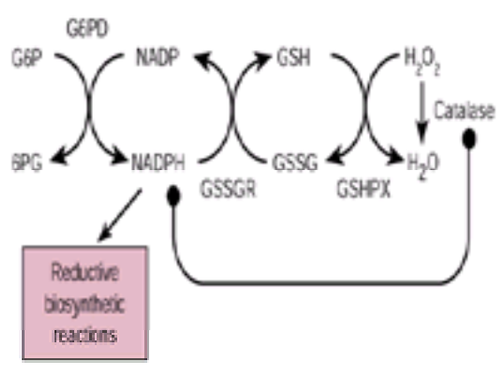

Fig. 3 The role of G6PD in red cell metabolism: NADPH plays a dual role in (i) regeneration of glutathione (GSH) and (ii) stabilization of catalase.

The enzymatically active form of G6PD is either a dimer or a tetramer of a single protein subunit of 514 amino acids with a molecular mass of 59 096 Da. Some regions of the molecule critical for its functions have been identified because they are highly conserved in evolution. The G6P-binding site and the active centre of the enzyme are located near lysine 205. Recently the three-dimensional structure of G6PD has been solved. In the dimer structure the two subunits are symmetrically located across a complex interface of b-sheets. The NADP binding site is near the N-terminus, and bound NADP is important for the stability of G6PD.

Acute haemolytic anaemia associated with G6PD deficiency clearly results from the action of an exogenous factor on intrinsically abnormal red cells. Although the sequence of events ending in haemolysis is not completely understood, we know that oxidative agents cause glutathione depletion in G6PD-deficient red cells. This is followed by oxidation of sulphydryl groups and consequent denaturation of haemoglobin (hence the Heinz bodies) and probably of other proteins, which eventually causes irreversible damage to the membrane of red cells and hence their destruction, partly in the bloodstream and partly through phagocytosis by macrophages. An important feature of haemolysis in G6PD-deficient patients depends on the fact that G6PD decays gradually during red cell ageing (for instance, in normal blood, reticulocytes have about five times more activity than the 10 per cent oldest red cells), and this process is accelerated with many G6PD variants. Thus, a haemolytic attack selectively destroys older red cells because they have a more severe shortage of G6PD. This phenomenon can be so marked with certain G6PD variants that patients in the posthaemolytic state are found to have a significant increase in G6PD activity (hence the risk of misclassification), sufficient to make them relatively resistant to further challenge. By contrast, with some other variants the steady-state level of G6PD is so low that, even in the absence of any oxidant challenge, it becomes limiting for red cell survival: this is the case in patients with chronic non-spherocytic haemolytic anaemia, who may have a red cell lifespan of between 10 and 50 days.

Molecular basis of G6PD deficiency

Since the discovery of G6PD deficiency, one might have expected that some mutations would be located in regulatory regions of the gene, producing a reduction in the amount G6PD produced, without changes in its structure (analogous to thalassaemias); whereas others would be located in the coding region of the gene, thus producing qualitative (or structural) as well as quantitative changes in G6PD (analogous to structural haemoglobinopathies). In fact, whenever G6PD from G6PD-deficient individuals has been subjected to careful biochemical characterization (for example by analysing electrophoretic mobility, substrate affinity constants, or thermostability), qualitative differences have invariably been detected, predicting structural mutations.

By sequencing the G6PD gene from G6PD-deficient subjects it has been verified that all mutations are structural (Fig. 4). The current database of some 130 mutants consists, with few exceptions, of single point mutations in the coding region of the gene, entailing single amino acid replacements in the G6PD protein. The exceptions have been small deletions of one to eight amino acids, and a few instances in which two point mutations rather than one are present (for instance, in G6PD A-, the variant most commonly encountered in Africa). Regulatory mutations have not yet been discovered. Amino acid replacements can cause G6PD deficiency either by affecting its catalytic function or by decreasing the *in vivo* stability of the protein, or by both of these mechanisms. Enzyme instability is the most common mechanism (Table 2). The molecular basis of chronic non-spherocytic haemolytic anaemia associated with G6PD deficiency is highly specific, in the sense that the underlying mutations, while still within the coding region of the G6PD gene, are not the same as those underlying asymptomatic G6PD deficiency. The more severe clinical phenotype can be ascribed in some cases to adverse qualitative changes (for instance, a decreased affinity for the substrate, glucose-6-phosphate); or simply to the fact that the enzyme deficit is more extreme, because of severe instability of the enzyme. For instance, a cluster of mutations map to the dimer interface, and it is clear that they severely compromise the formation of the dimer.

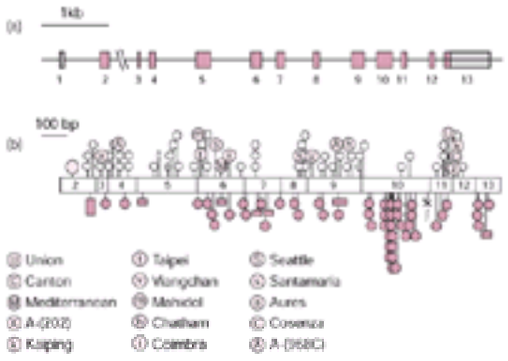

Fig. 4 Hetereogeneity of G6PD deficiency. The 13 exons of the *G6PD* gene are drawn approximately to scale; the introns (not drawn to scale) are shown by thin lines connecting the exons. The location of the mutations for the variants listed in Table 2 are shown; plus that of G6PD Sunderland, as example of an English sporadic variant associated with chronic non-spherocytic haemolytic anaemia and due to a deletion of a triplet of bases, corresponding to codon 35.

Because the G6PD gene is X linked, frequencies of G6PD deficiency in males are identical to gene frequencies, and they are as high as 20 per cent or more in some of the areas just mentioned. The frequency of homozygous females is of course lower, but the frequency of heterozygous females is higher (according to the Hardy–Weinberg rule) than that of G6PD-deficient males. Different G6PD variants underlie G6PD deficiency in different parts of the world: for instance, G6PD Mediterranean on the shores of this sea, in the Middle East, and in India; G6PD A- in Africa and in Southern Europe; G6PD Mahidol in South-East Asia; G6PD Canton in China; and G6PD Union worldwide. It is also important to realize that in some populations several different polymorphic variants coexist. The overall geographical distribution of G6PD deficiency and its heterogeneity, together with findings from clinical field studies and *in vitro* experiments, strongly support the view that this common genetic trait has been selected by *Plasmodium falciparum* malaria, by virtue of the fact that it confers a relative resistance to heterozygotes against this highly lethal infection.

Management

Prevention

The acute haemolytic anaemia of G6PD deficiency is largely preventable by avoiding exposure to triggering factors of previously screened subjects. Of course, the practicability and cost-effectiveness of screening depends on the prevalence of G6PD deficiency in each individual community. Favism is entirely preventable by not eating broad beans. Prevention of drug-induced haemolysis is possible in most cases by choosing alternative drugs. A common practical problem is the need to give primaquine for eradication of malaria due to *Plasmodium vivax* or *P. malariae*; in these cases the administration of a lower dose of the drug for a longer time is the recommended approach: this will still cause haemolysis, but of an acceptably mild degree.

Treatment of acute haemolytic anaemia and favism

A patient with acute haemolytic anaemia may present a diagnostic problem, that once solved, does not require any specific treatment at all; or he or she may be present as a medical emergency requiring immediate action. With severe anaemia, immediate blood transfusion is definitely indicated. If there is acute renal failure, haemodialysis may be necessary. Recovery is the rule.

Management of neonatal jaundice

This does not differ from that of neonatal jaundice due to other causes than G6PD deficiency. In most cases, prompt phototherapy is highly effective and sufficient; but with bilirubin levels above 300 µmo1/1 (or even less in babies who are premature, or who have acidosis or infection), exchange blood transfusion is imperative to

prevent neurological damage.

Management of chronic non-spherocytic haemolytic anaemia

In general terms, this does not differ from that of chronic non-spherocytic haemolytic anaemia due to other causes, for example pyruvate kinase deficiency. If the anaemia is not severe, regular folic acid supplements and regular haematological surveillance will suffice. It will be important to avoid exposure to potentially haemolytic drugs, and blood transfusion may be indicated when exacerbations occur, mostly in conjunction with intercurrent infection. In rare patients the anaemia is so severe that regular blood transfusion is necessary, probably at approximately 2-month intervals, in order to keep the haemoglobin in the 8 to 10 g/dl range. A hypertransfusion regimen aiming to maintain a normal haemoglobin level is not indicated (as there is no ineffective erythropoiesis in the bone marrow). However, in patients requiring regular transfusions, appropriate iron chelation should be instituted by the age of 2 years, and must be continued as long as transfusion treatment is necessary; sometimes the transfusion requirement may decrease after puberty. Although, unlike in hereditary spherocytosis, there is no evidence of selective red cell destruction in the spleen, splenectomy has proved beneficial in severe cases. When a diagnosis of chronic non-spherocytic haemolytic anaemia is made, the family must be given genetic counselling, and an effort should be made to establish whether the mother is a heterozygote; if she is, the chance of recurrence is 1:2 for every subsequent male pregnancy. Prenatal diagnosis can be made by DNA analysis if the mutation is first identified in an affected relative.

Further reading

Beutler E (1978). Glucose-6-phosphate dehydrogenase deficiency. In: Beutler E ed. *Hemolytic anemia in disorders of red cell metabolism*, p 23. Plenum Medical, New York.

Beutler E (1991). Glucose-6-phosphate dehydrogenase deficiency. *New England Journal of Medicine* **324**, 169–74.

Dacie JV (1985). Hereditary enzyme deficiency haemolytic anaemias. Deficiency of glucose-6-phosphate dehydrogenase. In: Dacie JV, ed. *Haemolytic anaemias*, III: *The hereditary haemolytic anaemias*, p 364. Churchill Livingstone, London.

Luzzatto L (1993). Glucose-6-phosphate dehydrogenase deficiency and hemolytic anemia. In: Nathan DG, Oski FA, eds. *Hematology of infancy and childhood*, p 674. Daunders, Philadelphia.

Vulliamy T, Mason P, Luzzatto L (1992). The molecular basis of glucose-6-phosphate dehydrogenase deficiency. *Trends in Genetics* **8**, 138–43.

Vulliamy TJ, Beutler E, Luzzatto L (1993). Variants of glucose-6-phosphate dehydrogenase are due to missense mutations spread throughout the coding region of the gene. *Human Mutation* **2**, 159–67.

22.6.1 The biology of haemostasis and thrombosis

Harold R. Roberts and Gilbert C. White

Introduction

Fluid blood is contained within the vascular tree, but as a result of minor trauma that occurs during the wear and tear of everyday living, leaks occur in the vessel wall that must be sealed by a solid, impermeable fibrin clot in order to prevent significant blood loss. The clot is formed from factors in flowing blood and is located and restricted to the site of the leak without dissemination throughout the vascular tree. This is the process of haemostasis, an exquisitely controlled mechanism that requires components of the vessel wall, blood platelets, and soluble procoagulant and anticoagulant proteins. The haemostatic plug consists of a mass of platelets, red blood cells, and leucocytes enmeshed in interlocking strands of fibrin fibres that plug the leak.

Once formed, the haemostatic plug is gradually replaced by new tissue that results in wound healing. This process requires lysis of the blood clot by the fibrinolytic system and subsequent ingrowth of new cells. Thus, haemostasis is not an isolated phenomenon, but is one component of the defence mechanisms that include inflammation and eventual wound healing.

Thrombosis, as opposed to haemostasis, is a pathological state in which the normal clotting system is disturbed to the extent that a clot is formed that partially or completely obstructs the flow of blood within the blood vessel and sometimes dislodges to become an embolus.

To understand the biology of haemostasis and thrombosis, it is necessary to know the roles of the vessel wall, the platelets, the coagulation and fibrinolytic systems, and their respective inhibitors.

Blood vessel wall

The anatomy of the wall of both an artery and a vein is shown schematically in Fig. 1. All blood vessels are lined by an intima consisting of a monolayer of endothelial cells that rest upon a loose network of tissue called the extracellular matrix. In addition to the intima, larger and intermediate arteries contain two other layers: the media, composed mostly of smooth muscle cells, and the adventitia, consisting largely of connective tissue, nerves, and nutrient vessels. While these three layers also exist in veins, the media and adventitia are much less distinct and are not visible in the smaller arterioles and capillaries.

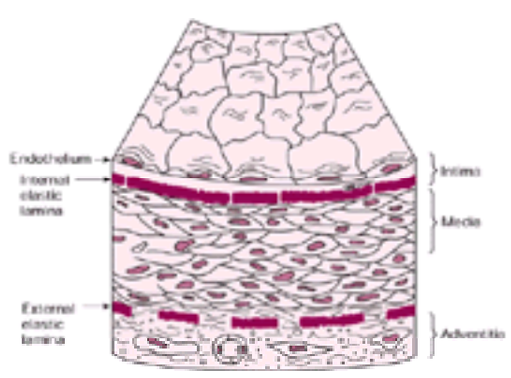

Fig. 1 A schematic diagram of a vessel wall consisting of the intima, the media (smooth muscle cells), and the adventitia. The intima consists of a layer of endothelium that is exposed to the circulating blood. The subendothelial matrix lies below the endothelium and is separated from the media by the internal elastic membrane. See text for detailed description of each layer. (Reprinted by permission, *Robbins pathological basis of disease*, 4th edn, (1989), p.554, WB Saunders Co.)

Endothelial cells

Endothelial cells form the basis of vascular development and are derived from embryonic mesoderm. Embryonic endothelial cells (angioblasts) develop under the influence of growth hormones including basic fibroblast growth factor and vascular endothelial growth factor, both of which interact with receptors on the cell membrane termed receptor tyrosine kinases. These early blood vessels expand into a vascular tree under the influence of two major hormones, angiopoietin 1 and 2, that bind to a family of tyrosine kinase receptors called tie-1 and tie-2 (tyrosine kinase plus Ig and epidermal growth factor-like domains) on endothelial cells. To fully develop into an intact vascular tree, endothelial cells must interact with the extracellular matrix and other cells, a process that requires cell–cell adhesion that is dependent upon cell surface cytoadhesive molecules such as platelet–endothelial cytoadhesive molecule-1, and vascular endothelial cell cadherin. Endothelial cell structure is also dependent upon the integrin family of molecules and interactions with the extracellular matrix.

Endothelial cells are heterogeneous in appearance, function, and genetic regulation. In the brain, endothelial cells form very tight junctions with one another to preserve the blood–brain barrier; in the spleen and liver, the interendothelial gaps are wide, permitting soluble and cellular trafficking between blood and the extravascular space. Not all endothelial cells synthesize the same proteins. Tissue plasminogen activator is synthesized by only about 3 per cent of cells. Von Willebrand factor, often regarded as a specific marker for endothelial cells, is not expressed in all cells. The microenvironment also plays an important role in regulating endothelial cell function. Haemodynamic forces, including hydrostatic pressure, and shear stresses and strains can influence endothelial cell structure and

function. Haemodynamic forces can even regulate endothelial cell gene expression. For example there is a shear stress response element in the gene governing the synthesis of the b chain of the platelet-derived growth factor. Other endothelial cell genes responsive to shear forces include those for: tissue plasminogen activator, intercellular adhesion molecule, and vascular cell adhesion molecule-1.

Endothelial cells contribute to haemostasis by their contributions to vascular tone, procoagulant, anticoagulant, fibrinolytic, and antifibrinolytic activities.

Vascular tone

Vasoregulatory substances produced by endothelial cells are shown in Table 1. The most important vasoregulators are nitric oxide, previously known as endothelial cell-derived relaxation factor and prostacyclin. Nitric oxide and prostacyclin are also important antiplatelet agents. On the other hand, the most important vasoconstrictors are endothelin and angiotensin 2. Endothelin is also a mitogen for smooth muscle cells.

Procoagulant properties

Procoagulant properties of the endothelial cell are depicted in Table 2. von Willebrand factor is synthesized constitutively by endothelial cells and is essential for platelet adhesion to the vessel wall and as a carrier for blood clotting factor VIII. von Willebrand factor is stored in Weibel–Palade bodies, as depicted in Fig. 2. It is released into the circulation in multimers of heterogeneous molecular weight ranging from one to about 20 million. Endothelial cells also secrete very large von Willebrand factor multimers abluminally into the extracellular matrix.

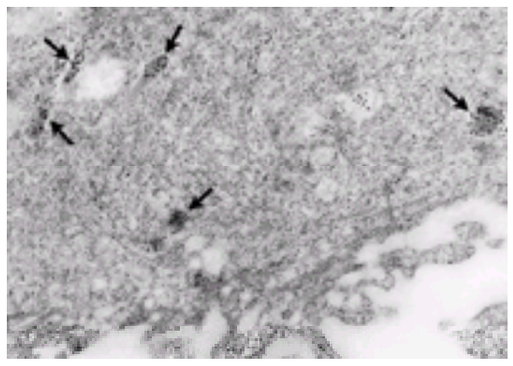

Fig. 2 Electron micrograph of an endothelial cell. Weibel–Palade bodies containing multimers of von Willebrand factor are depicted by the arrows.

Tissue factor acts as a binding protein for factor VII and is essential for the initiation of coagulation. It is not constitutively produced by endothelial cells, but it can be induced by tissue necrosis factor, endotoxin, and other inflammatory substances.

Anticoagulant properties

Anticoagulant properties of the endothelial cells are also shown in Table 2. Prostacyclin not only causes vasodilation, but it is a potent inhibitor of platelet aggregation. Nitric oxide has a similar effect. An important anticoagulant function of endothelial cells is due to the expression of thrombomodulin, a transmembrane-bound protein that acts as a receptor for thrombin. The thrombomodulin/thrombin complex is the physiological activator of protein C, which inactivates clotting factors Va and VIIIa to turn off coagulation. Tumour plasminogen activator (t-PA) is also synthesized by endothelial cells and serves to convert plasminogen to plasmin, the active fibrinolytic enzyme.

Endothelial cells contribute to the control of coagulation by synthesizing tissue factor pathway inhibitor (TFPI), which inhibits the tissue factor-mediated initiation of the clotting reactions. They also secrete glycosaminoglycans, such as heparan sulphate and other proteoglycans that inhibit thrombin. In addition, they express vascular adenosine triphosphate diphosphohydrolase, otherwise known as CD39, on their surface. CD39 acts in concert with 5'ectonucleotidase to convert ATP/ADP to AMP and then to adenosine, which inhibits platelet aggregation.

Receptor function

The receptor function of endothelial cells plays an important role in haemostasis and thrombosis (Table 3). They express a thrombin receptor termed protease-activated receptor 1 (PAR-1). Thrombin cleaves the carboxyterminal end of the receptor, which then binds to the remaining cell-associated protein (a so-called tethered ligand) and triggers intracellular signalling through G proteins, resulting in activation of endothelial cells. The thrombin–thrombomodulin complex not only activates protein C, but also activates a protein known as the thrombin-activatable fibrinolytic inhibitor (TAFI), a procarboxypeptidase that functions to inhibit fibrinolysis. Endothelial cells also express a protein C receptor different from thrombomodulin that acts to modulate the activity of activated protein C. Urokinase plasminogen activator receptors are not found on resting endothelial cells, but are found on those involved in angiogenesis. There are a number of adhesive receptors on the surface of endothelial cells as shown in Table 3. The adhesion of neutrophils is dependent upon the expression of P-selectin. P-selectin is rapidly internalized by the endothelial cell, but this is followed by expression of another cytoadhesive molecule, E-selectin, which is necessary for continued adherence and rolling of neutrophils along the endothelial cell surface. Intercellular adhesion molecule and vascular cell adhesion molecule are receptors for leucocytes and are important for the interaction of leucocytes and the vessel wall.

Extracellular matrix

The extracellular matrix is a complex, heterogeneous structure beneath the endothelium with many interactions related to haemostasis and thrombosis. The matrix consists of a network of collagens, elastins, proteoglycans, and glycoproteins, including fibronectin, vitronectin, laminin, tenascin, thrombospondin, von Willebrand factor, and osteopontin, among others, as shown in Table 4. The matrix proteins promote platelet adhesion, cellular migration, cell proliferation, and endothelial and smooth muscle cell interactions.

Collagens are the most abundant proteins in subendothelial connective tissue. Collagen types I, II, III, IV, V, VI, and VII have been identified in various matrix tissues. The collagens are synthesized by endothelial cells, smooth muscle cells, and by adventitial fibroblasts. The various collagens contribute to the integrity of the vessel wall, but they also play a role in platelet activation and, in some instances, coagulation. For example collagen IV has been shown to be a specific high-affinity binding protein for blood coagulation factor IX, although the function of this complex is not known.

The proteoglycans constitute a heterogeneous group of molecules composed of a core protein attached to a glycosoaminoglycan. These include decorin, biglycan, heparan sulphate, dermatan sulphate, and others. Heparan sulphate, for example, can combine with antithrombin III and inhibit thrombin. The precise role of all of the protoglycans is not known, but some attach to collagen and are necessary for maintaining the structure of the vessel wall.

The matrix also contains elastin, which is secreted by endothelial and smooth muscle cells as tropoelastin that is converted to mature elastin in the matrix where it is assembled into fibres. One function of elastin is simply to maintain the elastic structure of the vessel wall. This substance is found interspersed between smooth muscle cells as well as the matrix. It may also function in cell migration from the vessel wall to the extravascular space.

Fibronectin, vitronectin, and laminins are also components of the extracellular matrix which function in fibrinolysis and platelet adhesion.

Within the extracellular matrix there are a number of matrix metalloproteinases (MMP), which are a group of enzymes useful in matrix degradation and repair. They are secreted as proenzymes and converted to active enzymes that are zinc- or calcium-dependent. They have several functions, as listed in Table 5. Their activities

in matrix degradation and repair are controlled by tissue inhibitors of metalloproteinases.

Smooth muscle cells

The smooth muscle cell layer, found in medium and larger sized vessels and more prominently in arteries, has several functions related to the biology of haemostasis and thrombosis. Smooth muscle cells possess contractile, biosynthetic, and proliferative functions. Contractile properties, governed by such substances as nitric oxide, prostacyclin, and endothelin, play important roles in vasodilation and vasoconstriction, respectively. Smooth muscle cells, like endothelial cells, synthesize growth factors such as vascular endothelial growth factor, insulin-like growth factors, epidermal growth factors, activins, and others that are important in smooth muscle cell generation. Smooth muscle cell proliferation is a hallmark of the atherosclerotic lesion. Biosynthetic products of the smooth muscle cells include various types of collagens, elastin, glycoproteins, and proteoglycans. When exposed to injury, smooth muscle cells can also produce tissue factor.

The adventitia

The adventitia is composed of a loose network of cells consisting of fibroblasts, adipocytes, and mast cells. Collagens I and III, glycoproteins, and elastin are synthesized by fibroblasts. Fibroblasts also contain tissue factor. Adipocytes secrete collagen I and III and synthesize lipids.

Platelets

Platelets are the smallest of the circulating blood cells, about 0.5 µ in diameter. They are derived from the cytoplasm of megakaryocytes and are an essential component of the haemostatic plug. Although they are anucleate and appear to be single cells composed of cytoplasm, a surface canalicular system, and storage granules († or dense granules and a granules); they are, nevertheless, complicated cells with a variety of very important functions essential for normal haemostasis. These can be broadly divided into the following:

1. platelet adhesion, defined as platelets adhering to the damaged area of the vessel wall where subendothelial matrix tissue is exposed;
2. platelet activation, both by agents within the matrix as well as by soluble agonists;
3. platelet aggregation, defined as platelets sticking to one another in an aggregated mass, forming a platelet plug.

The following sections describe each of these broad areas of platelet function in more detail.

Platelet adhesion

The initial platelet response to vascular injury is adhesion to the vessel wall. Resting, non-activated platelets are not attracted to the vessel wall. However, following vascular damage, platelets rolling along the endothelium rapidly adhere to the subendothelial interstitial matrix. A number of matrix proteins, such as von Willebrand factor, fibronectin, fibrinogen, and thrombospondin, are also present in platelet granules and circulate in plasma.

Platelets possess numerous mechanisms for adhering to the subendothelial matrix (Fig. 3). Adhesion is accomplished by a number of protein receptors on the surface of platelets as described below.

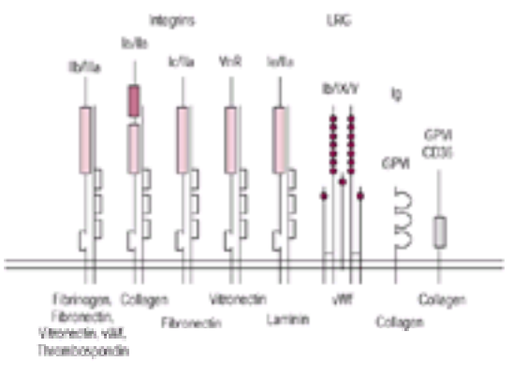

Fig. 3 Receptors mediating the interaction of platelets with subendothelial matrix proteins. Adhesion receptors on platelets include members of the integrin family, leucine-rich glycoproteins (LRG), members of the immunoglobulin (Ig) family, and others. Integrins on the surface of platelets are glycoproteins (GP) IIb/IIIa, which binds multiple ligands; GPIa/IIa, a collagen receptor, GPIc/IIa, which binds fibronectin; VnR, which is a receptor for vitronectin; and GPIc'/IIa, a laminin binding site. Glycoproteins Ib/IX/V are leucine-rich glycoproteins. Glycoprotein VI (GPVI) is a member of the immunoglobulin family and a collagen receptor. Glycoprotein IV (GPIV, CD36) is also a collagen receptor.

Glycoprotein Ib–IX–V (CD42a–d)

The main function of the platelet membrane glycoprotein Ib–IX–V complex is to act as a receptor that mediates von Willebrand-dependent binding of platelets to collagen, resulting in adhesion of platelets to the vessel wall. Glycoproteins Iba, Ibb, IX, and V are members of the leucine-rich glycoprotein family and are characterized by the presence of a common structural motif in the extracellular domain composed of a leucine-rich sequence, PXXLLXXXXXLXXLXLSXNXLXXL *. GPIba contains seven leucine-rich repeats, GPV contains 15 leucine-rich repeats in the extracellular domain, while GPIbb and GPIX each contain a single leucine-rich repeat, all in the extracellular domains. GPIba, GPIbb, GPIX, and GPV are synthesized as separate gene products which coassociate in a ratio of 2:2:2:1 during transit through the endoplasmic reticulum. Coassociation of GPIba, GPIbb, and GPIX, but not GPV, is required for the complex to be expressed on the surface of cells. The role of GPV, a substrate for thrombin, in the function of the complex is uncertain, and mice deficient in GPV bind von Willebrand factor normally and have normal platelet adhesion.

Adhesion to von Willebrand factor-coated surfaces through GPIb–IX–V is increased by shear, which is thought to induce a structural change in the receptor that enhances the interaction with von Willebrand factor. The A1 domain of von Willebrand factor forms the principal site that interacts with GPIb. The site in GPIb that binds von Willebrand factor is less well defined. Binding occurs in the amino-terminal, 45-kDa, tryptic fragment from GPIba. Within this region of GPIba, an anionic site, YDYYPEE^{282}, containing two sulphated tyrosine residues at tyrosines 278 and 279, has been further implicated in von Willebrand factor binding. The A3 domain of von Willebrand factor mediates the interaction with collagens type I and III. A model derived from the crystal structure of the von Willebrand factor A3 domain suggests that the von Willebrand factor–collagen interaction is primarily between negatively-charged residues in the A3 domain and positively-charged residues in collagen. Plasma von Willebrand factor does not interact with unstimulated circulating platelets. For binding to occur, platelets have to be activated and plasma von Willebrand factor must undergo a conformational change. After secretion by endothelial cells, von Willebrand factor binds to underlying connective tissue matrix, providing an active surface for platelet attachment after the vessel wall is damaged.

Glycoprotein IIb–IIIa (aIIb/b3)

Under conditions of low shear, platelets can adhere to matrix-bound von Willebrand factor through a mechanism that involves platelet GPIIb–IIIa. GPIIb and IIIa are members of the integrin superfamily, a conserved family of heterodimeric surface receptors, each composed of a larger two-chain a subunit and a smaller b subunit, bound non-covalently. Integrins were initially identified by an ability to bind adhesive glycoproteins containing a tripeptide sequence, arginine–glycine–aspartic acid (RGD), although subsequent work has identified other ligand sequences recognized by integrins. The interaction of von Willebrand factor with GPIIb–IIIa is mediated by an RGD sequence in the C1 domain of von Willebrand factor. GPIIb–IIIa is also able to bind fibronectin, thrombospondin, and vitronectin and may therefore represent an adhesion receptor with broad specificity.

Glycoprotein Ia–IIa (a2b1)

GPIa–IIa is a receptor for types I and IV collagen and mediates platelet adhesion to the vessel wall independent of von Willebrand factor. The integrin sequences that mediate the interaction with collagen reside in a broad sequence called the I domain in the extracellular portion of the molecule. GPIa–IIa is constitutively active and does not require activation to interact with collagen.

Glycoprotein VI–Fc receptor g-chain complex

GPVI–FcRg is the major platelet receptor mediating collagen-induced activation of platelets. GPVI is a member of the immunoglobulin superfamily and is characterized by immunoglobulin domains, a transmembrane domain, and a short cytoplasmic tail that lacks known signalling components. GPVI is associated on the platelet surface with FcRg, apparently in a 1:1 stoichiometry. The complex binds collagen and mediates collagen-generated signals, presumably through the ITAM (or immunoglobulin receptor tyrosine-based activation motif) of FcRg. Crosslinking of GPVI–FcRg leads to tyrosine phosphorylation of the ITAM sequence by Src kinase. Syk, another tyrosine kinase, binds to the phosphorylated ITAM sequence through Syk sulphydryl domains, initiating a signal that leads to tyrosine phosphorylation of phospholipase Cg2 and the generation of inositol phospholipids.

Glycoprotein IV (CD36)

CD 36 is a highly glycosylated transmembrane protein present on platelets, monocytes, endothelial cells, and nucleated erythrocytes, which binds thrombospondin and collagen. The thrombospondin-binding site has been mapped to a single disulphide loop in the extracellular domain of GPIV, but the collagen-binding site is unknown. Although GPIV is a receptor for collagen *in vitro*, individuals with a deficiency of GPIV have no apparent defect in platelet function.

Other adhesion receptors

Platelets can also adhere to subendothelial matrix through glycoprotein Ic–IIa (VLA-5, a5b1), glycoprotein Ic'–IIa (VLA-6, a6b1), or the vitronectin receptor (avb3, VnR). GPIc–IIa is a constitutively active receptor for fibronectin that does not require cell activation. There are two sequences in fibronectin which interact with GPIc–IIa: an RGD sequence in the tenth type III repeat which interacts primarily with the GPIIa (b1) subunit and a synergy sequence in the adjacent ninth type III repeat which interacts primarily with the GPIc (a5) subunit. GPIc'–IIa is a laminin receptor which is expressed on platelets. Immunoprecipitation studies suggest that GPIc'–IIa may exist on the cell surface in a complex with proteins with four transmembrane domains, so-called TM4 proteins, such as CD9, CD81, and NAG-2. The nature of these interactions is presently unclear. GPIc'–IIa recognizes a sequence in the long arm E8 fragment of laminin obtained after elastin digestion. The binding requires the presence of divalent cations which bind to specific sites on the integrin a subunit. Small numbers of the vitronectin receptor are expressed on platelets.

Current evidence indicates that all of these adhesion mechanisms may be important. The redundancy in adhesion receptors may:

1. provide backup mechanisms to protect against blood loss;
2. generate different signals in response to interaction with different matrix proteins; or
3. represent different systems at work in different parts of the vascular tree.

An example of the latter might be the relative roles of GPIb–IX–V and GPIIb–IIIa in the von Willebrand factor-mediated adhesion of platelets to collagen. Under high shear conditions, as found in capillaries and small arterioles, GPIb–IX–V may be the predominant mechanism mediating platelet adhesion to collagen and von Willebrand factor-dependent adherence whereas, under low shear conditions, like those found in large veins and in arteries, GPIb–IX–V may be less effective and other mechanisms that require a shorter residence time of platelets on the subendothelial matrix, including GPIc–IIa interaction with fibronectin and GPIa–IIa interaction with collagen, may be important. The presence of multiple receptors for collagen on the platelet surface, including GPIb–IX–V, GPIIb–IIIa, GPIa–IIa, GPIV, and GPVI, is interesting and raises the possibility of different collagen responses. Vitronectin also appears to be important for adhesion at high shear, and can bind to both GPIIb–IIIa and specific vitronectin receptors. Recent evidence suggests that platelet adhesion to collagen types I and III in flowing blood is dependent on both von Willebrand factor and fibronectin. Collagen types I, II, and III have been shown to bind von Willebrand factor.

Platelet activation

Following adhesion and in response to soluble agonists such as thrombin, platelets undergo a series of complex biochemical reactions leading to cell activation. As a result, platelets undergo changes in shape, alterations in surface lipid composition leading to the generation of platelet coagulant activity and thrombin generation, and secretion of the contents of intracellular granules leading to the release of ADP. The thrombin generated at the platelet surface and ADP secreted from platelet granules lead to activation of additional platelets. These reactions involve the metabolism of membrane inositol phospholipids, changes in cellular levels of calcium, activation of contractile proteins, stimulation of heterotrimeric and low molecular weight GTP-binding proteins, and tyrosine and serine–threonine phosphorylation of proteins, among other events. These biochemical reactions initiate second messenger signals that drive the functional changes that occur in platelets which transform them from the resting state to an activated one, and which play a crucial role in haemostasis. Some of these signalling pathways are described in the following sections (see Fig. 4).

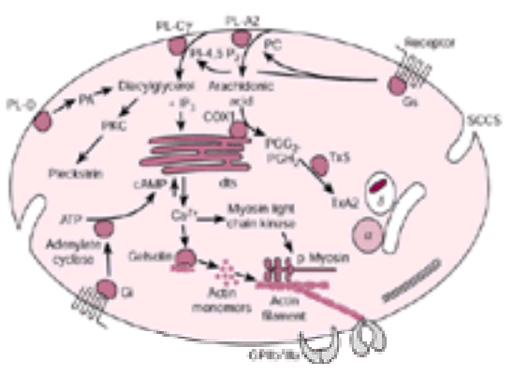

Fig. 4 Signalling pathways involved in platelet activation. Following the interaction of agonist with receptor, there is G protein (Gs)-coupled activation of phospholipid metabolic pathways through phospholipase A_2 (PL-A2), phospholipase Cg (PL-Cg), and phospholipase D (PL-D) leading to generation of thromboxane A_2 (TXA2), inositol trisphosphate (IP_3), diacylglycerol, and phosphatidic acid (PA). Arachidonic acid generated by the action of phospholipase A_2 is converted by cyclo-oxygenase-1 (COX-1) to prostaglandin endoperoxides G_2 (PGG_2) and H_2 (PGH_2) which are, in turn, converted to thromboxane A_2 through the action of thromboxane synthase (TXS). Thromboxane generated through arachidonate metabolism is thought to play a role in secretion, through the fusion of a-granule (a) and dense granule (†) membranes with the membrane of the surface-connected canalicular system (SCCS). Granule contents, including adenosine diphosphate, are emptied into the SCCS and make their way to the outside of the cell. Diacylglyerol stimulates activation of protein kinase C (PKC), resulting in serine–threonine phosphorylation of proteins such as pleckstrin. Inositol trisphosphate stimulates calcium release from storage sites in the dense tubular system (dts). The release of calcium from the dense tubular system is antagonized by cyclic AMP, generated through G protein (Gi)-coupled inhibitory receptor activation of adenylate cyclase. Calcium, released in response to IP_3, activates gelsolin, an actin-capping and -severing protein, which generates actin monomers that then serve as nucleation sites for formation of actin filaments and assembly of the activation-dependent cytoskeleton. Assembly of the cytoskeleton and interaction of the cytoskeleton with surface integrins such as glycoproteins IIb and IIIa (GPIIb/IIIa) may be involved in integrin activation. Calcium also activates myosin light chain kinase which phosphorylates myosin light chain-generating actinomyosin contraction, important for changes in platelet shape and the secretion process.

Phospholipid metabolism

Metabolism of membrane phospholipids is one of the first signalling pathways identified in platelets and remains one of the most important. Platelet stimulation by a

variety of agonists results in activation of membrane-associated phospholipases, including phospholipases C, A2, and D, which cleave fatty acids from the phospholipid. The lipid products generated by these pathways are signalling compounds which are important for changes in cytoplasmic calcium and activation of kinases and phosphatases.

The most intensively studied of these pathways is the metabolism of inositol phospholipids through phospholipase C. Membrane phosphatidylinositol (PI) exists in multiple phosphorylation states: PI, PI-P, PI-P2 which is phosphorylated in the 3,4 or 4,5 positions, and PI-P3 which is phosphorylated in the 3,4,5 positions. Phosphatidylinositol-specific kinases and phosphatases maintain pools of phosphorylated phosphoinositides in a proper concentration range. Platelets contain several isoforms of phospholipase C which are activated by different mechanisms. All cleave phosphatidylinositol 4,5-bisphosphate (PI 4,5-P2) and, later, phosphatidylinositol, as well as phosphatidylinositol 4-phosphate (PI 4-P), to yield diglyceride and inositol trisphosphate (IP3). Phospholipase Ca and Cb are coupled to heterotrimeric G proteins where phospholipase Cg is coupled to growth factor receptors. Inositol trisphosphate (IP_3) generated by phospholipase C cleavage of inositol phospholipids has been implicated in the release of calcium from intracellular storage sites in the platelet-dense tubular system. The other product of phospholipase C cleavage, diacylglycerol, activates protein kinase C, which phosphorylates pleckstrin, a 47 000-dalton protein, and other proteins.

Phospholipase A2 is linked to G-protein coupled receptors and cleaves fatty acids in the *sn*-2 position in membrane phospholipids, primarily phosphatidylcholine. In most individuals in western society, the fatty acid in this position is arachidonic acid. Arachidonic acid, liberated by the action of phospholipase A2, is converted to a variety of possible products by the microsomal enzymes, cyclo-oxygenase and lipoxygenase. Cyclo-oxygenase converts arachidonic acid to prostaglandin endoperoxides, prostaglandins F_2, E_2, and D_2, whose main fate in platelets is rapid conversion to thromboxane A_2 by thromboxane synthase. Thromboxane A_2 is believed to play an important role in the release of intracellular granules by acting as a membrane fusogen, fusing granule membranes with the membrane of the surface connected canalicular system and permitting secretion of the granule contents to the outside of the cell. Thromboxane A_2 is also an exceptionally potent constrictor of vascular smooth muscle and a strong platelet-aggregating agent.

Inhibition of the arachidonate pathway has been a primary target for platelet inhibition. Cyclo-oxygenase is irreversibly inhibited by aspirin, which acetylates serine 340, and reversibly inhibited by non-steroidal anti-inflammatory agents. Inhibition of cyclo-oxygenase inhibits thromboxane formation and results in inhibition of the release of intracellular granules. The mechanism by which aspirin is thought to act as an anti-atherosclerosis agent is by inhibition of the release of platelet-derived growth factor.

Phospholipase D acts primarily on phosphatidylcholine to produce choline and phosphatic acid. Protein kinase C and PI-P2 play an important role in activation of phospholipase D. Phosphatidic acid is an intracellular messenger which is proposed to play a role in platelet activation. In addition, phosphatidic acid can be converted to lysophosphatidic acid through the action of phospholipase A2. Like phosphatidic acid, lysophosphatidic acid is an intracellular messenger which is involved in phospholipase activation, ras signalling, and cytoskeleton reorganization.

Calcium metabolism

In resting platelets, the cytoplasmic concentration of calcium is maintained at a low level by active transport of calcium both outside the cell and into the dense tubular system, a sarcoplasmic reticulum-like fraction in platelets. This transport is accomplished by a plasma membrane sarcoplasmic-endoplasmic-reticulum-like calcium ATPase (SERCA2-b), a dense tubular system SERCA3, a sodium-calcium exchange pump in the plasma membrane, and passive calcium fluxes. During platelet activation, inositol trisphosphate (IP_3), generated by metabolism of membrane inositol phospholipids, induces the rapid release of calcium stored in the dense tubular system. This increase in cytoplasmic calcium is essential for platelet activation, and agents that inhibit increases in cytoplasmic calcium inhibit platelet activation while agents that increase cytoplasmic calcium stimulate platelet activation.

Calcium functions as a major intracellular messenger in platelets, mediating calcium-dependent reactions important in almost all phases of platelet activation. An increase in the concentration of cytoplasmic free calcium activates gelsolin, the calcium-dependent actin capping and severing protein, which plays an important role in reorganization of the cytoskeleton. Calcium also activates the calcium and calmodulin-dependent myosin light chain kinase, leading to phosphorylation of myosin light chains, activation of actin-stimulated myosin ATPase activity, and the development of contractile forces. The contraction generated by actin and myosin mediates changes in platelet shape and is important for events leading to platelet secretion. In the absence of calcium ions, tropomyosin inhibits the interaction of myosin with actin, and this may be an additional regulatory role of calcium in platelets. Calpain, a calcium-dependent thiol protease, hydrolyzes numerous proteins involved in platelet signalling. Activation of calpain is believed to be important both for regulation of cytoskeletal events and integrin-mediated signalling.

Cytoskeletal reorganization

Resting platelets are discoid in shape and feature a cellular cytoskeleton that consists of a network of actin filaments that fill and shape the cytoplasm of the cell and a single microtubule coil at the margin of the disc. Upon activation, platelets undergo remarkable morphological changes (Fig. 5). There is an initial change from the normal discoid shape of the resting platelet to a sphere as calcium levels in the cell increase. Filamentous actin appears in the form of stress fibres, and the cellular content of filamentous actin increases. Membrane ruffles form as well as long cellular projections called pseudopodia. Actin cables are present in these pseudopodia, extending to the end of the projections. Also during activation, microtubules contract and 'squeeze' granules toward the centre of the cell.

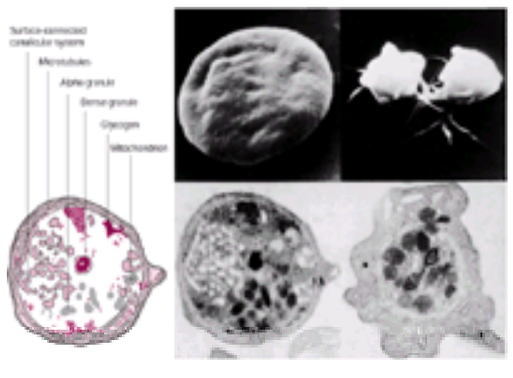

Fig. 5 Platelet morphology. Platelets are small, anucleate cells. In the resting state (a), platelets are ducoid shaped and contain a marginal rim of microtubreles. After activation (b), platelets undergo changes in shape, becoming more rounded and extending cytoplasmic projections, called pseudopods, outward.

The energy for contraction is provided by a magnesium ion-dependent ATPase present in myosin and stimulated by actin. Contraction occurs by actin filaments and myosin rods sliding over one another. Myosin light-chain phosphatase may switch off myosin. Membrane glycoproteins GPIIb–IIIa, GPIb–IX–V, and other membrane proteins are associated with the cytoskeleton and provide direction for the contractile process. This activation-dependent cytoskeleton is more than just a structural scaffold for platelet shape changes. Numerous signalling proteins are incorporated into the cytoskeleton and may function in specialized compartments by virtue of their association with the cytoskeleton.

Platelet coagulant activity (platelet factor 3)

Platelet membranes have an asymmetrical distribution of phospholipids, with almost all of the acidic (negatively charged) phospholipids, such as phosphatidyl serine and phosphatidyl inositol, located in the inner leaflet of the plasma membrane. After platelet activation, the acidic phospholipids are translocated to the outer half of the membrane, while phosphatidylcholine moves to the inner half, in a phenomenon known as a 'flip-flop' reaction. This transbilayer movement of phospholipids in the platelet membrane is not well understood, but evidence for a 'flipase' which enzymatically contributes to it has been presented. There is also a translocase enzyme that works in the opposite way and is capable of restoring the acidic phospholipids to the inner leaflet of the membrane bilayer.

The exposed phosphatidyl serine and other negatively-charged phospholipids account for some of the activity traditionally known as platelet factor 3 by contributing to surface properties for binding of factor X and prothombin activation complexes. This interaction with platelet phospholipids increases the rate of factor X activation and prothrombin activation nearly a thousand fold. In addition to phospholipids on the platelet membrane, there appear to be other specific binding proteins for blood

clotting factors.

Secretion

A primary endpoint of platelet activation is the secretion of platelet granule contents to the outside of the cell. During platelet activation, the granules are 'squeezed' to the centre of the cell where the granules fuse with the surface-connected canalicular system, a series of intracellular canals that are connected to the cell surface. The contents of the granules make their way to the outside of the cell. Secretion requires prostaglandin metabolism and is dependent on contractile events. Products of prostaglandin metabolism, primarily thromboxane A2, may act in the fusion of the granule membrane with that of the surface-connected canalicular system.

Platelets possess two types of storage granules (Table 6), both of which are involved in secretion of active ingredients that modulate platelet function. One type is the dense granule, so called because it is dense by electron microscopy. The other type is the a-granule.

Dense granules contain adenine nucleotides, calcium, and serotonin. Adenine nucleotides are sequestered in the dense granules mainly as adenosine diphosphate (ADP) and adenosine triphosphate (ATP) in a complex with calcium ions and pyrophosphate, and are not interchangeable with the nucleotides involved in general cell metabolism. ADP released from platelet-dense granules activates additional platelets and recruits them to the growing platelet thrombus. Serotonin, a potent modulator of vascular tone and integrity, is also a constituent of dense granules.

a-Granules contain platelet-derived growth factor, b-thromboglobulin, platelet factor 4, fibrinogen, factor V, von Willebrand factor, and thrombospondin. Platelet-derived growth factor is mitogenic for smooth-muscle cells and when released from platelets at a site where the vessel wall is damaged, it stimulates proliferation and migration of smooth-muscle cells in the intima, contributing to the atherosclerotic process. b-Thromboglobulin and platelet factor 4 are basic, lysine-rich proteins which interact with glycosaminoglycans such as heparan sulphate, dermatan sulphate, and chondroitin sulphate, which are components of the endothelial cell surface. Platelet factor 4 has a strong heparin-neutralizing activity and has been implicated in the aetiology of heparin-induced thrombocytopenia. Thrombospondin is a major a-granule glycoprotein, but it is also secreted by fibroblasts, endothelial, and smooth-muscle cells. Thrombospondin is a high-molecular-weight adhesive protein which binds to glycosaminoglycans, fibrinogen, plasminogen, histidine-rich glycoprotein, type V collagen, and calcium ions. It associates with cell surfaces and extracellular matrices and facilitates cell–cell and cell–matrix interactions.

cAMP pathway

The major mechanism for down-regulation of platelet function is the stimulation of adenylate cyclase, which increases cAMP concentrations. Adenylate cyclase is mainly localized in microsomal fractions and is stimulated by adenosine, prostacyclin, and prostaglandin E_1. cAMP inhibits platelet aggregation, platelet secretion, and platelet adhesion to the vessel wall. These effects are probably exerted by inhibiting calcium flux and/or promoting calcium reuptake.

Activation by soluble agonists

In addition to activation through interaction with subendothelial connective tissues, platelets may also be activated by soluble agonists. These include adenosine diphosphate (ADP), epinephrine, and thrombin. In general, this activation occurs through the interaction between soluble agonist and specific receptors on the platelet surface.

Thrombin is one of the most powerful of platelet agonists. Generated during blood coagulation, thrombin activation of platelets occurs through a novel family of receptors called protease-activated receptors (PARs). These are G protein-coupled, seven-membrane-spanning molecules which are activated by proteolysis. Thrombin cleaves the amino terminal exodomain, unmasking a new amino terminal, which functions as a tethered peptide agonist. The tethered peptide binds intramolecularly to the remainder of the receptor to trigger activation. Four members of the PAR family of receptors have been identified. PAR 1 and PAR 4 mediate activation of human platelets by thrombin.

Thrombin interacts with other proteins on the surface of platelets, but the nature of these interactions is uncertain. Glycoprotein V, part of the GPIb–IX–V complex is a substrate for thrombin although the absence of GPV does not appear to inhibit thrombin activation of platelets. GPIb is an equilibrium binding site for thrombin. Patients with a deficiency of GPIb have been reported to have changes in the rate of activation of platelets by thrombin which is overcome at higher concentrations of agonist.

There are at least three receptors for ADP on platelets, all members of the seven membrane-spanning purinergic (P2) receptor family, either P2Y (G-protein-coupled purinergic receptors) or P2X (ligand-gated channel receptors). One receptor, designated P2Y1, is coupled to phospholipase C, probably through Gq. A second receptor, $P2T_{AC}$, is coupled to adenylate cyclase, probably through Gi. The third receptor, P2X1, is coupled to rapid calcium influx and is a member of the intrinsic ion channel family. Full platelet activation by ADP probably involves an interaction of ADP with all three receptors. ADP-induced activation of the GPIIb-IIIa on platelets requires both P2Y1 and $P2T_{AC}$ and concomitant signalling through GTP-binding proteins, Gq and Gi.

Platelet aggregation

Platelet aggregation, the interaction of one platelet with another, is a major function of platelets and is very important in the haemostatic process. The formation of an aggregated platelet-mass at the site of injury provides a physical plug that occludes the defect in the vessel wall and prevents blood loss.

Aggregation is mediated by two glycoproteins on the platelet surface, GPIIb–IIIa, which constitute a receptor for fibrinogen–fibrin. Thus, GPIIb–IIIa on one platelet binds fibrinogen or fibrin which, by virtue of its dimeric structure, interacts with GPIIb–IIIa on another platelet. On resting platelets, GPIIb–IIIa is in an inactive state and is unable to bind fibrinogen. Following platelet activation, GPIIb–IIIa becomes activated through a process that involves calcium, protein kinase C, and heterotrimetric G proteins. Activation of GPIIb–IIIa requires energy and is a multistep process. Fibrinogen binding to GPIIb–IIIa occurs through a carboxy-terminal dodecapeptide sequence, HHLGGAKQAGDV, in the g chain of fibrinogen where the AGDV sequence has been suggested to have structural similarity to the RGD sequence.

Blood coagulation

The blood coagulation system consists of a number of zymogens (proenzymes) that are proteolytically converted to active enzymes in a series of steps involving activators and cofactors. The coagulation reactions are initiated by tissue factor in complex with activated factor VII (VIIa). The tissue factor /VIIa complex then activates both factor IX and factor X, which, in the presence of their respective cofactors (factors VIII and V), lead to the rapid conversion of prothrombin to thrombin. The latter converts fibrinogen into a solid fibrin clot that finally undergoes cross-linking by activated factor XIII to become a stable haemostatic plug. Platelets are essential in several steps of the clotting mechanism and form the surface for activated clotting factors, which lead to the explosive generation of thrombin.

Understanding the modern concept of the clotting reactions requires a detailed knowledge of each factor. Table 7 depicts the clotting factors and their inhibitors, including the vitamin K-dependent clotting proenzymes, the non-vitamin K-dependent zymogens, the cofactors, the inhibitors of the clotting factors, and the structural proteins.

The vitamin K-dependent zymogens

The vitamin K-dependent blood clotting zymogens include: prothrombin, factor VII, factor IX, factor X, and protein C; their characteristics are listed in Table 7 and their schematic structures in Fig. 6. A common feature of all these clotting factors is the presence of gamma carboxyglutamic acid (Gla) domains in the amino terminal region of the molecules. Glutamic acid residues in these proteins undergo carboxylation, a post-translational event that is effected by hepatic carboxylase that requires reduced vitamin K as a cofactor. The vitamin K-dependent factors are highly homologous in terms of amino acid sequence. Factors VII, IX, X, and protein C have a similar domain structure with a Gla domain, two epidermal growth factor-like (EGF) domains, and a catalytic domain (Fig. 6). Prothrombin differs from other vitamin K-dependent factors in that it has two kringle domains (Fig. 6). Both factor X and protein C are secreted as two-chain zymogens while the others are single-chain proteins.

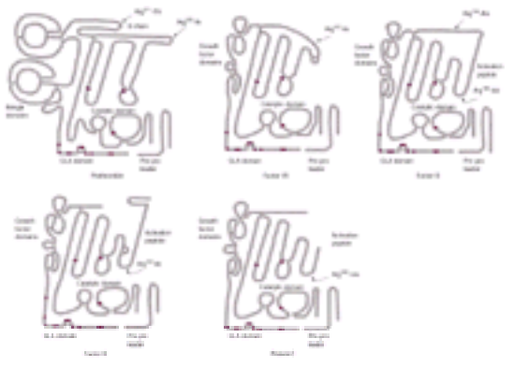

Fig. 6 Schematic diagram of the vitamin K-dependent factors, prothrombin and factors VII, IX, X, and protein C. •, closed circles are gamma carboxyglutamic acid residues; _, represents active site triad of serine, histidine, and aspartic acid; arrows denote cleavage site. (Reprinted by permission of McGraw-Hill Companies from Roberts HR *et al*. (2001). Molecular biology and biochemistry of the coagulation factors. *Williams Hematology*, 6th edn, p.141.)

The Gla domains of these factors are necessary for binding to phospholipid membranes. Calcium ions occupy the Gla domain to result in a conformational change in the protein that favours binding to platelet membrane surfaces. Phosphatidylserine is the major phospholipid in these reactions.

The vitamin K zymogens are all serine proteases with the typical active site: a serine/histidine/aspartic acid triad. Exposure of the active site requires that the zymogen be activated by cleavage of specific arginyl residues. As a result, all the activated vitamin K-dependent zymogens become two-chain enzymes linked by disulphide bonds, as depicted in Fig. 6. Despite the high degree of sequence homology of these proteins, they are highly specific in their interaction with their cofactors and substrates.

Prothrombin

Prothrombin is synthesized in the liver and has a molecular weight of about 72 000 daltons. The molecule has 10 Gla residues that play a role in the binding of the prothrombin to the surface of activated platelets where it is converted to the active enzyme, thrombin, by the so-called 'prothrombinase complex' consisting of factors Xa/Va/Ca^{++} on the platelet surface. Thrombin is a potent enzyme with a molecular weight of about 38 000 that rapidly converts fibrinogen to a fibrin clot. Thrombin also has many other actions including its role as: a potent activator of platelets; an activator of factor V, VIII, and XIII; an activator of protein C in the presence of its cofactor, thrombomodulin; an activator of procarboxypeptidase to form a thrombin-activatable fibrinolytic inhibitor (TAFI); and as a growth factor. The primary inhibitor of thrombin is antithrombin III.

Factor VII

Factor VII is synthesized in the liver and has a molecular weight of about 50 000 daltons. It has a very short half-life of about 3.5 h. The specific receptor for factor VII is tissue factor found on the surface of many cells such as fibroblasts, activated monocytes, and many other cell types. The physiological activator of factor VII is unknown, although it has been suggested that it might be activated by factor Xa. The factor VII/tissue factor complex activates both factors IX and X. The factor VII–tissue factor –Xa complex is inhibited by tissue factor pathway inhibitor (TFPI). Factor VIIa is not appreciably inhibited by antithrombin III except in the presence of heparin.

Factor IX

Factor IX is synthesized by hepatocytes and has a molecular weight of about 57 000 daltons. Its plasma half-life is 18 to 24 h. The molecule has 12 Gla residues. About 40 per cent of the factor IX molecules carry a b-hydroxyaspartic acid at position 64 of the molecule. It is activated by VIIa–tissue factor and by activated factor XI, both of which cleave an arginyl bond at position 145 and 180 of the molecule to release an activation peptide of about 10 000 daltons. Factor IXa in complex with its cofactor, activated factor VIII, cleaves factor X to Xa. Antithrombin III will inhibit factor IXa, but the inhibition is not as rapid as the antithrombin III inhibition of thrombin or factor Xa.

Factor X

Factor X is also synthesized by hepatocytes and has a molecular weight of 59 000. It is secreted as a two-chain molecule linked by disulphide bonds. It has 11 Gla residues. When activated by factor IXa or VIIa–tissue factor, an activation peptide is cleaved from the heavy chain to expose the serine-active site on the heavy chain. Factor Xa, in the presence of its cofactor (factor Va), rapidly converts prothrombin to thrombin. The primary inhibitor of factor Xa is AT III.

Protein C

Protein C, unlike the other vitamin K-dependent zymogens, is not a procoagulant, but, when activated by the thrombin–thrombomodulin complex on the surface of endothelial cells, it becomes an anticoagulant by proteolytically cleaving factors Va and VIIIa, thus inhibiting coagulation. To function in this way as an anticoagulant, activated protein C (APC) requires a cofactor, protein S. Protein C is synthesized in the liver and has a very short half-life of about 6 h. It contains nine Gla residues and has a molecular weight of 59 000 daltons. The primary inhibitor of APC is the protein C inhibitor (PCI), also known as plasminogen activator inhibitor-3 (PAI-3).

The non-vitamin K-dependent zymogens

Factor XI

Factor XI is synthesized in the liver as a dimeric protein composed of identical subunits. It has a molecular weight of 160 000 daltons and a plasma half-life of about 72 h (Table 7). In plasma, factor XI circulates in complex with high molecular weight kininogen (HK), a non-enzymatic cofactor. The physiological activator of factor XI is thought to be thrombin, although *in vitro*, this factor can be activated by factor XIIa. The main function of factor XIa is to boost thrombin generation by activating factor IX on the surface of platelets, over and above the factor IX activated by the VIIa–tissue factor complex. Some patients with factor XI deficiency have no bleeding tendency, and those who do usually exhibit mild bleeding when compared to severely affected haemophilic patients.

Factor XII and prekallikrein (PK)

These factors have been collectively referred to as contact factors since it appears that activation of factor XII is enhanced by contact with a surface. Factor XII and PK are zymogens, which, when activated, expose a serine-active site (Table 7). HK is a non-enzymatic protein cofactor that circulates in complex with factor XI and PK. All of these factors are synthesized in the liver. Unlike the vitamin K-dependent proteins, factors XI, XII, and prekallikrein all possess so-called 'apple domains' that have specific functional characteristics. Deficiencies of factor XII and PK are not associated with bleeding tendencies in patients with complete deficiency of these factors. However, deficiency of each factor is associated with a marked prolongation of the partial thromboplastin time (PTT). In this test and in the presence of glass, ellagic acid, or some inert earth material, factor XII can activate factor XI. These factors may not play a major physiologic role in haemostasis, but there is evidence that they do participate in inflammatory responses that involve blood coagulation, fibrinolysis, and kinin generation.

The cofactors

Some of the cofactors are soluble and exist in circulation, namely protein S, protein Z, factors V and VIII, high molecular weight kiningen (HK), and von Willebrand factor (Table 7). Others are cell-bound, such as tissue factor and thrombomodulin.

Protein S

Protein S is synthesized in the liver and endothelial cells. It circulates in plasma and is also found in platelets. It has a molecular weight of 75 000 daltons and a plasma half-life of about 42 h. It contains 11 Gla residues in the amino terminal region. In structure, protein S differs dramatically from the other vitamin K clotting factors in that the carboxy terminal end is homologous to growth hormone. Protein S acts as a cofactor for activated protein C. Protein S exists in two forms: one form is bound to C4b-binding protein and the other exists as a free form in the circulation. Free protein S is a cofactor for protein C and is in equilibrium with the bound form.

Protein Z

Protein Z is synthesized in the liver and has a molecular weight of 62 000 daltons. There is now convincing evidence that when protein Z is incubated with factor Xa, the activity of the latter is reduced. The inhibition of factor Xa activity is due to the presence of a protease inhibitor that requires protein Z as a cofactor. Whether protein Z has other functions is unknown.

Factor V

Factor V is synthesized in the liver and has a biological half-life reported to be between 12 and 36 h. It is a large glycoprotein with a molecular weight of 330 000 daltons. Factor V is highly homologous to factor VIII. A schematic diagram of the structure is shown in Fig. 7. As can be seen, it is composed of A, B, and C domains. The A domains are homologous to the copper-binding protein, ceruloplasmin, so it is not surprising that this domain of factor V is involved in binding to calcium and copper. The C domains are homologous to fat globule proteins and are involved in the binding of factor V to phospholipid-rich platelet membranes. The A and C domains are homologous to similar domains in factor VIII, but the B domain is completely different from that of factor VIII. For factor V to act as a cofactor for factor Xa, it must be activated by thrombin with cleavage of arginyl bonds at positions 708, 1018, and 1545 as shown in Fig. 7. It is inactivated by activated protein C that cleaves bonds at 306 (slow) and 506 (fast).

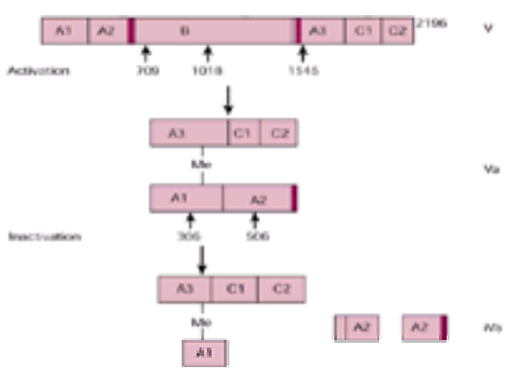

Fig. 7 Schematic diagram of factor V. Factor V is activated by thrombin to factor Va. Factor Va is inactivated by activated protein C (i Va). Activation of factor V by thrombin results in loss of the B chain and formation of a heterodimeric molecule covalently linked by metal ions (Me). Inactivation is by activated protein C that cleaves arginyl bonds at positions 306 and 506. (Reprinted by permission of McGraw-Hill Companies from Roberts HR *et al*. (2001). Molecular biology and biochemistry of the coagulation factors. *Williams Hematology*, 6th edn, p.1419.)

Factor VIII

Like factor V, factor VIII is synthesized in the liver. It is a large glycoprotein with a molecular weight similar to that of factor V. Again, like factor V, factor VIII has A, B, and C domains with the A domain homologous to caruloplasmin and the C domain homologous to fat globule proteins (Fig. 8). The A domain of factor VIII is essential for binding to phospholipid membranes. The B domain of factor VIII is cleaved during activation and has no known function. To act as a cofactor for factor IXa, factor VIII must be activated by thrombin or factor Xa. Unlike activated factor V, activated factor VIII exists as a heterotrimer composed of A1, A2, and C1-C2 domains linked by calcium ions. Factor VIII circulates in a non-covalent complex with von Willebrand factor and has a biologic half-life of 8 to 12 h. In the complete absence of von Willebrand factor, such as occurs with type III von Willebrand disease, the half-life of factor VIII is less than 1 h. When activated factor VIII is released from von Willebrand factor, it binds to the surface of activated platelets where it interacts with factor IXa.

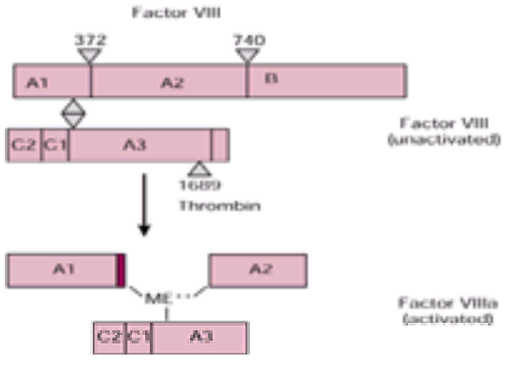

Fig. 8 Schematic representation of factor VIII. Activation by thrombin (or factor Xa) results in a heterotrimer non-covalently linked by metal ions (Me). Like factor Va, factor VIIIa is inactivated by activated protein C. (Reprinted by permission of McGraw-Hill Companies from Roberts HR *et al*. (2001). Molecular biology and biochemistry of the coagulation factors. *Williams Hematology*, 6th edn, p.1420.)

von Willebrand factor

von Willebrand factor is synthesized by endothelial cells and stored in Weibel–Palade bodies. It binds to glycoprotein Ib on platelets and is required for normal platelet adhesion to components of the vessel wall such as collagen. A schematic diagram of von Willebrand factor is shown in Fig. 9. Although synthesized as a prepolypeptide with A, B, C, and D domains, it is secreted into the plasma in multimeric form with molecular weights ranging from 1 million to 15 to 20 million. Higher molecular weight forms of von Willebrand factor are secreted to the abluminal surface of the endothelial cell as one component of the extracellular matrix. The higher molecular weight multimers are very effective in promoting platelet adhesion. von Willebrand factor is also important in platelet aggregation. A major function of von Willebrand factor is to act as a carrier protein for factor VIII. Factor VIII is associated with von Willebrand factor multimers of all sizes.

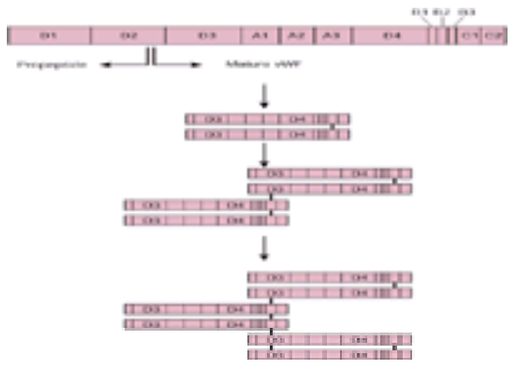

Fig. 9 Schematic diagram of von Willebrand factor. The formation of multimeric forms of von Willebrand factor occurs through links of dimers via D_3 domains. (Reprinted by permission of McGraw-Hill Companies from Roberts HR *et al*. (2001). Molecular biology and biochemistry of the coagulation factors. *Williams Hematology*, 6th edn.)

High molecular weight kininogen (HK)

High molecular weight kininogen circulates in plasma, and part is bound to factor XI and prekallikrein. HK is a cofactor for both of these zymogens. Deficiency of HK is not associated with a bleeding tendency, although the partial thromboplastin times of affected subjects are prolonged.

Tissue factor

Tissue factor, unlike other cofactors, is associated with cell surfaces. It is composed of 263 amino acids and with a 219-amino acid extracellular domain, a 23-amino acid transmembrane domain, and a 21-amino acid intracytoplasmic domain. The characteristics of tissue factor are shown in Table 7. It has a molecular weight of about 46 000 and is constituitively expressed on several extravascular tissues such as fibroblasts and smooth muscle cells. It is not constitutively expressed on cells exposed to the circulating blood, but can be induced in endothelial cells by certain inflammatory cytokines and certain bacterial products such as endotoxin. It can also be induced in blood leucocytes. Tissue factor functions as a receptor for factor VII. When factor VII binds to tissue factor, it is rapidly converted to factor VIIa, although the precise mechanism for its activation is not clear. The VIIa–tissue factor complex is now thought to be the main physiological initiator of blood coagulation by activating both factor IX and factor X, each of which plays a distinct role in subsequent coagulation reactions as described below. On some cells, tissue factor exists in a 'latent' form sometimes referred to as 'encrypted tissue factor' as suggested by the fact that tissue factor antigen levels on cells may be higher than tissue factor functional activity. 'De-encryption' can be accomplished by exposure of cells to agents such as calcium ionophores and various cytokines, but the physiological mechanism by which this process takes place is not known.

Thrombomodulin

Thrombomodulin is a transmembrane protein synthesized by and localized to endothelial cells although it has also been found on mesothelial cells, monocytes, and squamous epithelial cells. It has a molecular weight of about 78 000 daltons. A chondroitin sulphate moiety is attached to thrombomodulin via a serine residue. The major characteristics of thrombomodulin are depicted in Table 7. It serves as a receptor on endothelial cells for thrombin. Thrombin bound to thrombomodulin undergoes a structural transformation such that it no longer activates platelets or clots fibrinogen, but rather activates protein C. The principle function of the thrombomodulin /thrombin complex is to prevent the extension of the haemostatic clot past the site of a break or leak in the vessel wall and as such represents an important control mechanism to restrict the haemostatic plug precisely to the point of injury. Thus, under normal conditions, clot formation does not occur on the endothelial cell surfaces.

Fibrinogen

Fibrinogen is synthesized in the liver and has a molecular weight of 340 000 daltons. It is a dimeric glycoprotein consisting of two sets of identical chains, the a, b, and g chains. The synthesis of each fibrinogen chain is governed by a separate gene as depicted in Table 7. The normal plasma half-life of fibrinogen is about 3 to 5 days. It is also found in the a granules of platelets as a result of endocytosis. Fibrinogen is the soluble plasma precursor of the solid fibrin clot that is so necessary for haemostasis and normal wound healing. The dimeric structure of fibrinogen is composed of two monomers, each containing disulphide-linked a, b, and g chains. A schematic diagram of fibrinogen is shown in Fig. 10. It is a triodular structure with a central E domain that includes the disulphide-linked amino termini of all six polypeptide chains. The E domain is linked to the carboxyterminal domains referred to as the D domains.

Fibrinogen conversion to fibrin is accomplished by thrombin cleavage of two fibrinopeptides (fibrinopeptide A and fibrinopeptide B) from each of the two a and b chains, respectively, leading to the formation of the fibrin monomer. The molecular weight of each fibrinopeptide A and B is about 2500 daltons. The soluble fibrin monomer then undergoes spontaneous polymerization by forming side-to-side and end-to-end anastomoses, resulting in protofibrils that aggregate into a visible fibrin clot composed of thicker, branched fibres. During fibrin clot formation, other proteins are occluded in the clot, including plasminogen, fibronectin, thrombospondin, and von Willebrand factor. The fibrin polymerization is enhanced by calcium ions, but the polymerization process alone does not lead to a stable and impermeable fibrin clot since the fibres are held together weakly by hydrogen bonds and electrostatic forces. A stable fibrin clot requires cross-linking of the a and g chains of fibrin by the action of activated factor XIII.

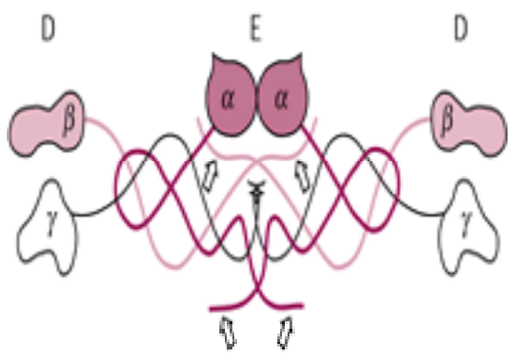

Fig. 10 A diagram of the structure of fibrinogen. The three chains, a, b, g, are shown. The E domain occurs at the amino termini while the D domains are found at the carboxytermini. Arrows represent cleavage sites for fibrinopeptide A from the a chain and fibrinopeptide B from the b chain.

Factor XIII

Factor XIII is a proenzyme which circulates in the plasma as a heterotetramer composed of two A chains and two B chains (Table 7). Factor XIII has a molecular weight of 320 000 daltons and has a half-life of about 10 days. It circulates in plasma in association with fibrinogen. The A chain contains the active site cysteine, while the B chain is enzymatically inactive and serves as a carrier for the A chain. The A chain is found in platelets where it is not associated with the B chain. Upon activation by thrombin, the A and B chains are separated. In addition, thrombin cleaves the A chain so as to expose the active site cysteine. The active component then cross-links the a and g chains of fibrin to form a stable, impermeable fibrin clot resistant to lysis by plasmin.

Inhibitors of the coagulation reactions

Tissue factor pathway inhibitor (TFPI)

TFPI is synthesized by endothelial cells. It has a molecular weight of about 34 000 to 40 000 daltons and serves to inhibit the initiation of coagulation (Table 7). TFPI can inhibit factor Xa in a slow reaction and also inhibits the VIIa–tissue factor–Xa complex. It exists in the circulation in at least three pools. One is bound to plasma lipoproteins; one pool is bound to proteoglycans on the vessel wall; and one exists in platelets. The TFPI bound to proteoglycans can be released by heparin. TFPI is a Kunitz-type inhibitor that is essential for control of coagulation.

Antithrombin III

The characteristics of antithrombin are also depicted in Table 7. Antithrombin belongs to a family of protease inhibitors known as 'serpins' that inhibit many proteases with a serine-active site. It is synthesized in the liver and has a plasma half-life of approximately 65 h. Its major function is to inhibit thrombin and factor Xa, although it will also inhibit the other coagulation serine proteases less well. The inhibitory action of antithrombin III is greatly enhanced by heparin, which accelerates the rate of inhibition of the serine proteases.

Protein Z-dependent protease inhibitor

This inhibitor inhibits factor Xa in the presence of calcium, phospholipids, and protein Z. It has a molecular weight of about 72 000 daltons. It, like antithrombin III, is also a member of the serpin family of serine protease inhibitors.

Other inhibitors of clotting factors

The major inhibitor of factor XIa is thought to be a-1-antitrypsin since it has the highest affinity for the enzyme. However, other inhibitors, namely C-1 esterase inhibitor, will also inhibit factor XIa. The other inhibitors that are of some importance in coagulation are also listed in Table 7.

The coagulation pathways

The coagulation reactions have been viewed as a sequential series of steps in which a enzymatic precursor (zymogen) clotting factor is converted to an active enzyme which, in turn activates another precursor, finally ending in the rapid conversion of prothrombin to thrombin. Early models of the coagulation reactions are shown in Fig. 11. As can be seen, when viewed in this manner, the clotting reactions appear as a waterfall or cascade, hence the terms waterfall or cascade hypotheses. Since tissue factor was extrinsic to the blood stream, the activation of factor X by the VIIa–tissue factor complex was termed the extrinsic system. The intrinsic system consisted entirely of clotting factors within the circulation and, upon conversion of factor IX to IXa by factor XIa, the factor IXa–VIIIa complex could also convert factor X to Xa in the presence of phospholipids. While this concept of coagulation was essentially correct, it did not explain why patients with factor XII deficiency had no bleeding tendency nor why factor XI-deficient patients had only a mild bleeding tendency. It was also pointed out that defects in the intrinsic system could lead to haemorrhage in affected patients even though the extrinsic system was intact and vice versa. The demonstration that the VIIa–tissue factor complex could activate both factor IX and factor X led several groups to conclude that the clotting reactions were, in fact, initiated by VIIa–tissue factor and that the intrinsic and extrinsic systems did not exist *in vivo*. Further work demonstrated that the clotting reactions leading to a haemostatic plug were controlled, in large part, by cell surfaces which modulated the reactions.

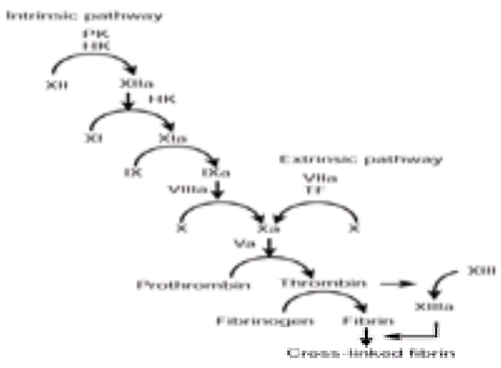

Fig. 11 An earlier model of blood coagulation reactions. The cascade or waterfall hypothesis of coagulation. (Reprinted by permission of McGraw-Hill Companies from Roberts HR *et al*. (2001). Molecular biology and biochemistry of the coagulation factors. *Williams Hematology*, 6th edn.)

The role of the tissue factor cell

When a blood vessel is injured or ruptured, flowing blood is exposed to tissue factor, which is bound through a transmembrane and cytoplasmic tail to cells exposed as the result of injury, for example fibroblasts and other connective tissue cells. Factor VII binds to the tissue factor-bearing cell and is activated. As a result, the VIIa–tissue factor complex on the tissue factor cell activates both factor IX and factor X as shown in Fig. 12(a). The factor Xa and IXa formed in the milieu of the tissue factor cell play very different and distinct roles in subsequent reactions.

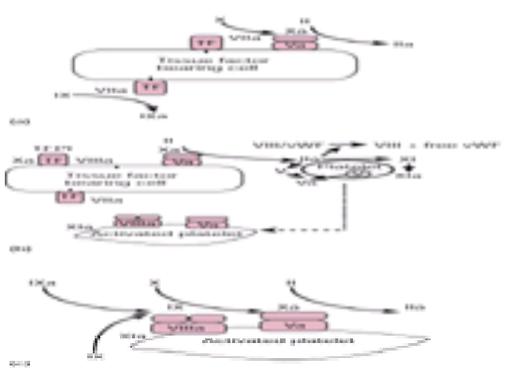

Fig. 12 (a) Tissue factor, a transmembrane protein expressed on tissue factor-bearing cells, acts as a receptor for factor VII, which is rapidly converted to factor VIIa. The tissue factor–VIIa complex then accomplishes two major functions: (1) activation of factor X to Xa and (2) activation of factor IX to IXa. Factor Xa activates factor V on the tissue factor-bearing cell and the resulting Xa–Va complex converts small amounts of prothrombin to thrombin. (b) This small amount of thrombin formed in the vicinity of the tissue factor cell acts as a 'primer' for coagulation by: (1) activating platelets; (2) dissociating factor VIII from von Willebrand factor and activating factor VIII; (3) activating factor V; and (4) activating factor XI. The activated platelets then adhere to the site of vascular injury and bind the cofactors, factors VIIIa and Va. Factor XIa also binds to platelets. The tissue factor–VIIa–Xa complex is then inhibited by TFPI (tissue factor pathway inhibitor). (c) Factor IXa formed by the tissue factor–VIIa complex associates with VIIIa on the platelet surface and recruits additional factor X from plasma to form factor Xa. The factor Xa then associates with its cofactor, factor Va, on the platelet surface to rapidly convert prothrombin to large amounts of thrombin sufficient to clot fibrinogen.

The role of factor Xa on the tissue factor cell

Factor Xa, in concert with its cofactor Va (which is found in the vicinity of tissue factor cells) then converts prothrombin to very small amounts of thrombin as shown in Fig. 12(b). This amount of thrombin, though insufficient to clot fibrinogen, can, however, act as a 'primer' of subsequent coagulation reactions to accomplish the following: activate platelets; activate more factor V; dissociate factor VIII from von Willebrand factor and activate factor VIII; and activate factor XI as shown in Fig. 12(b). Factor Xa alone and in complex with VIIa–tissue factor is then inhibited by TFPI. The activated cofactors resulting from the priming amount of thrombin in the milieu of the tissue factor cell then occupy binding sites on the activated platelet as shown in Fig. 12(b). Thus the main function of factor Xa formed as the result of the VIIa–tissue factor complex is to furnish a priming amount of thrombin sufficient to initiate further subsequent reactions which take place on the activated platelet surface.

The role of factor IXa on the tissue factor cell

Factor IXa formed by the VIIa–tissue factor on the tissue factor-bearing cell diffuses away from the tissue factor cell and occupies a site on the activated platelet adjacent to its cofactor VIIIa (Fig. 12(c)). This factor IXa then plays a primary role in the subsequent burst of thrombin generation on platelet surfaces as noted below.

The role of the activated platelet

The activated platelet mass is the primary site of thrombin generation, which is highly dependent upon the amount of factor IXa formed both by the VIIa–tissue factor cell and factor XIa, which also occupies sites on the platelet. Factor IXa in the presence of its cofactor VIIIa then recruits more factor X from solution and activates it on the activated platelet surface. This factor Xa in the presence of its cofactor Va then converts large amounts of prothrombin to thrombin sufficient to clot fibrinogen. All of these reactions are summarized in Fig. 13. The mass of aggregated platelets upon which these reactions take place is localized to the damaged area of the vessel wall.

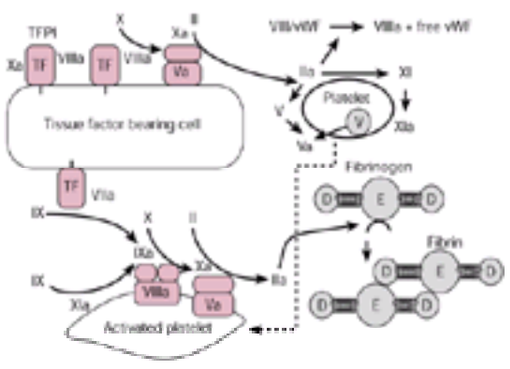

Fig. 13 The clotting reactions summarized. After thrombin formation, fibrinogen is converted to fibrin.

The role of the endothelial cells, vessel wall, and inhibitors

The mass of platelets interspersed with fibrin forms a plug at the site of a leak in the vessel wall where the endothelial cell monolayer is disrupted. The question arises as how the haemostatic plug is confined to the damaged area of the vessel wall. A schematic diagram of these events is shown in Fig. 14. The endothelial cells express thrombomodulin, which traps thrombin to form thrombomodulin–thrombin complex that controls the procoagulant stimulus by activating the protein C system, resulting in inactivation of both factors Va and VIIIa on the endothelial cell surface. In addition, endothelial cells contain glycosoaminoglycans, some of which inhibit thrombin. Antithrombin III also circulates in solution to inhibit any thrombin that escapes from the haemostatic plug. In this way the fibrin clot sealing a leak in a blood vessel wall is confined precisely to that site such that extension of the clot does not occur under normal circumstances.

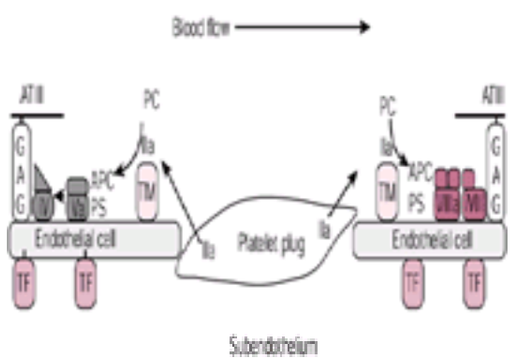

Fig. 14 A diagram of the haemostatic plug and the control mechanisms that restrict this plug to the site of injury and prevent extension of the clot to normal endothelium. TM, thrombomodulin; IIa, thrombin; GAG, glycosoaminoglycans; PC, protein C; APC, activated protein C; PS, protein S; TF, tissue factor; Platelet plug = haemostatic plug. iVa and iVIIIa refer to inactivated Va and VIIIa by APC.

On-going coagulation *in vivo*

It is well known that products of the coagulation reactions are found in the circulation under normal (basal) conditions. Small, but definite levels of fibrinopeptides A and B can be measured. Fragment 1+2 derived from the amino terminal portion of prothrombin after thrombin is formed can also be detected. Activation peptides from several of the coagulation factors as well as complexes of activated factors with their inhibitors can also be found in the circulation. These observations strongly suggest that small leaks in blood vessels that occur during the stress and strain of everyday living are repaired by the on-going formation of haemostatic fibrin clots. This has been termed 'basal' coagulation, a process that allows the blood to remain fluid within the vascular tree and at the same time permitting small, exquisitely controlled and confined fibrin clots to plug small leaks in the vasculature without dissemination. The fibrin plug is then removed by the fibrinolytic system following the formation of new tissue.

The fibrinolytic system

The fibrinolytic system is shown schematically in Fig. 15. The components of the system and their characteristics are depicted in Table 8. The active enzyme in the fibrinolytic system is plasmin, which is derived from its precursor, plasminogen. Plasminogen is activated to plasmin by activators. The physiological activator is single-chain tissue plasminogen activator (t-PA), which cleaves plasminogen into two-chain plasmin. Another activator of plasminogen *in vivo* is single-chain urokinase, but this appears to be more important for degradation of matrix proteins. The physiological inhibitor of plasmin is a $_2$-antiplasmin.

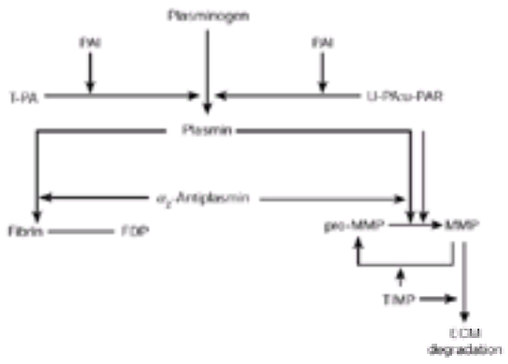

Fig. 15 The fibrinolytic system. Plasminogen is converted to plasmin by activators, including plasminogen activators (t-PA) and urokinase (u-PA). The activators are inhibited mainly by plasminogen activator inhibitor-1 (PAI-1). Plasmin degrades fibrin and activates matrix metalloproteinases (MMP), which degrades extracellular matrix (ECM). Plasmin is inhibited by a_2-antiplasmin. FDP, fibrin degradation product; TIMP, tissue inhibitors of metalloproteinases; U-PAR urokinase protease-activated receptor. (Reprinted by permission of FK Schatter from Collen D. (1999). The plasminogen (fibrinolytic) system. *Thrombosis and Hemostasis* **82**, 261.)

Plasminogen and t-PA associate in the circulation with fibrinogen. Thus when fibrinogen is converted to fibrin, the clot is rich in both of these proteins, which are protected from the inhibitory action of a_2-antiplasmin. Thus, clots can be lysed without interference from inhibitors, yet free plasmin in the circulation will be rapidly inhibited by its inhibitor.

Plasminogen

Plasminogen is synthesized in the liver and has a molecular weight of about 92 000 daltons. It is composed of a single chain and exists in two forms in the circulation: glu-plasminogen and lys-plasminogen. Glu-plasminogen has an amino-terminal glutamic acid and is a larger molecular weight than lys-plasminogen, which is formed in the circulation by plasmin cleavage of an arginyl bond at position 78 of the Glu form, leaving lysine as the amino-terminal residue. Lys-plasminogen rapidly binds to fibrin via lysine binding sites. Thus, lys-plasminogen is in close proximity to fibrin and protected from the action of antiplasmin. When activated, plasminogen is converted to active two-chain plasmin with a serine-active site on the heavy chain that is connected to the light chain by disulphide bonds.

The proteolytic action of plasmin is usually characterized by the proteolysis of fibrinogen and fibrin, but it can also degrade several other proteins including factor VIII, factor V, von Willebrand factor, and others. The cleavage of fibrinogen and fibrin leads to the formation of fibrin(ogen) degradation products. Fibrin(ogen) fragments resulting from plasmin cleavage are shown in Fig. 16. Fragment X is the first and largest fragment of plasmin digestion of fibrinogen. It is still clottable by thrombin, although much slower than native fibrinogen. Fragment X gives rise to fragment Y and D, and fragment Y is further proteolyzed to give rise to a second fragment D plus fragment E. These fragments can be detected in a simple laboratory test using antifibrinogen antibodies coated on latex particles. However, the test is non-specific and does not distinguish between the fibrinogen or fibrin degradation products which are quite similar, since the only difference between fibrin and fibrinogen is the absence of the small fibrinopeptides A and B in fibrin. A better test for detection of fibrin fragments is the so-called D-dimer test, which detects D-dimers resulting from the cross-linking of fibrin by factor XIIIa.

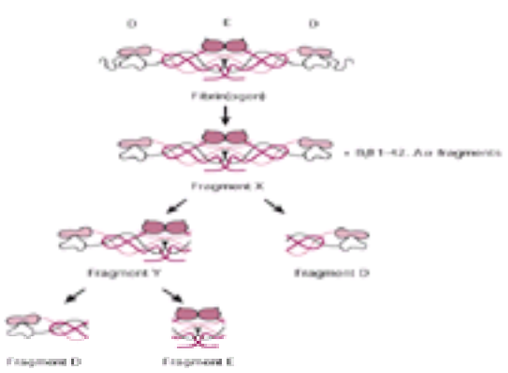

Fig. 16 Plasmin digestion of fibrin(ogen) results in fibrin degradation products: X, Y, D, and E. The final protolytic fragments resulting from plasmin degradation of fibrin are two molecules of fragment D and one of E. (Reprinted by permission from Francis CW, Marder VJ. In: Colman RW *et al*., eds. *Hemostasis and thrombosis: basic principles and clinical practice*, 3rd edn. JP Lippincott and Co.)

Tissue plasminogen activator (t-PA)

Tissue-type plasminogen activator is considered to be the physiological activator of plasminogen. It is synthesized in endothelial cells and has a molecular weight of about 68 000 daltons. It has high affinity for plasminogen. T-PA circulates for the most part in complex with its inhibitor, plasminogen activator inhibitor-1 (PAI-1). The t-PA–PAI-1 complex can be dissociated during the process of coagulation, and free t-PA associates with fibrin, which enhances t-PA activity. Single-chain t-PA has catalytic activity, but when activated to the two-chain form by plasmin, the activity is increased by three-fold.

Urokinase plasminogen activator

Urokinase plasminogen activator exists as a single-chain zymogen and is found in the kidney, the urine, and fibroblast-like cells. It activates plasminogen by proteolysis of an arginyl residue at position 561. Its main function is in wound healing and vasculogenesis, and it is active in proteolysis of the extracellular matrix. Urokinase plasminogen activator associates with the urokinase plasminogen-activator receptor.

Plasminogen activator inhibitor-1 (PAI-1)

PAI-1 is the physiological inhibitor of t-PA. It belongs to the serpin family of inhibitors. It is synthesized in endothelial cells and has a molecular weight of 52 000 daltons. Elevated levels of this inhibitor have been associated with arterial and venous thromboses. PAI-2, found in the placenta, also inhibits t-PA, but not as efficiently as PAI-1. PAI-3 is also known as the protein C inhibitor and inhibits plasminogen activators less efficiently than PAI-1.

a_2-Antiplasmin

a_2-Antiplasmin is the physiological inhibitor of plasmin. It has a molecular weight of about 58 000 daltons and is synthesized in the liver. As an inhibitor, it has three major functions: to inhibit plasminogen binding to fibrin; to inhibit the proteolytic activity of plasmin; and to bind to fibrin in a covalent manner by the action of factor XIIIa. By binding to fibrin, a_2-antiplasmin competitively inhibits the binding of plasminogen to fibrin. However, when plasminogen within the fibrin clot is converted to plasmin, the latter is protected from inhibition by antiplasmin. On the other hand, free plasmin formed in the circulation is rapidly inhibited.

Thrombin-activatable fibrinolytic inhibitor (TAFI)

TAFI is also known as plasma procarboxypeptidase B, and it is activated to carboxypeptidase B by large amounts of thrombin in a reaction dependent upon thrombomodulin. TAFI down-regulates fibrinolysis after clot formation and serves as an important regulatory mechanism for the fibrinolytic system. TAFI acts primarily

by reducing the number of high-affinity plasminogen binding sites on fibrin, the end result of which is decreased fibrinolysis.

The fibrinolytic and coagulation systems are closely interrelated. Under normal conditions, fibrin clot formation is always accompanied by fibrinolysis. The formation of the fibrin clot leads to localization of t-PA and plasminogen leading to activation of the fibrinolytic system. It also appears that activated factor XI and factor XII enhance fibrinolytic activity. The action of the protein C system to decrease thrombin formation down-regulates the thrombin activatable fibrinolytic inhibitor which would favour increased fibrinolysis. Although much is still unknown, it is generally accepted that both the coagulation and fibrinolytic systems are related to the general process of inflammation involving several other host defence mechanisms.

*A – Alanine, C – Cysteine, D – Aspartic acid, E—Glutamic acid, F—Phenylalanine, G – Glycine, H – Histidine, I – Isoleucine, K – Lysine, L – Leucine, M – Methionine, N – Asparagine, P – Proline, Q – Glutamine, R – Arginine, S – Serine, T – Threonine, V – Valine, W – Tryptophan, Y – Tyrosine, X – any Amino acid

Further reading

Andrews RK *et al.* (1999). The glycoprotein Ib-IX-V complex in platelet adhesion and signaling. *Thrombosis and Haemostasis* **82**, 357–64.

Brass LF *et al.* (1997). Signaling through G proteins in platelets: to the integrins and beyond. *Thrombosis and Haemostasis* **78**, 581–9.

Cines DB *et al.* (1998). Endothelial cells in physiology and in the pathophysiology of vascular diseases. *Blood* **91**, 3527–61.

Collen D (1999). The plasminogen (fibrolytic) system. *Thrombosis and Haemostasis* **82**, 259–70.

Coughlin SR (1999). Protease-activated receptors and platelet function. *Thrombosis and Haemostasis* **82**, 353–6.

Fay PJ (1999). Regulation of factor VIIIa in the instrinsic factor Xase. *Thrombosis and Haemostasis* **82**, 193–200.

Fox JEB (1999). On the role of calpain and rho proteins in regulating integrin-induced signaling. *Thrombosis and Haemostasis* **82**, 385–91.

Gimbrone MA (1999). Endothelial dysfunction, hemodynamic forces, and artherosclerosis. *Thrombosis and Haemostasis* **82**, 722–26.

Hartwig JH (1999). The elegant platelet: signals controlling actin assembly. *Thrombosis and Haemostasis* **82**, 392–8.

Kroll MH, Schafer AI (1989). Biochemical mechanisms of platelet activation. *Blood* **74**, 1181–95.

Loftus JC and Liddington RC (1997). New insights into integrin-ligand interaction. *Journal of Clinical Investigation* **99**, S77–81.

Majerus PW (1983). Arachidonate metabolism in vascular disorders. *Journal of Clinical Investigation* **72**, 1521–5.

Mann KG (1999). Biochemistry and physiology of blood coagulation. *Thrombosis and Haemostasis* **82**, 165–74.

Roberts HR *et al.* (1998). Newer concepts of blood coagulation. *Haemophilia* **79**, 306–9.

Roberts HR, Monroe DM, Hoffman M (1999). Molecular biology and biochemistry of the coagulation factors and pathways of hemostasis. In: Beutler E *et al.*, eds. *Williams' hematology*, 6th edn. McGraw-Hill, New York.

Roth GJ (1991). Developing relationships: arterial platelet adhesion, glycoprotein Ib, and leucine rich glycoproteins. *Blood* **77**, 5–19.

Ruggeri ZM (1999). Structure and function of von Willebrand factor. *Thrombosis and Haemostasis* **82**, 576–84.

Schmaier AH, Rojkjaer R, Shariat-Madar Z (1999). Activation of plasma kallikrein/kinin system on cells: a revised hypothesis. *Thrombosis and Haemostasis* **82**, 226–33.

Shattil SJ, Kashiwagi H, Pampori N (1998). Integrin signaling: the platelet paradigm. *Blood* **91**, 2645–57.

Smyth SS, Joneckis C, Parise LV (1994). Regulation of vascular integrins. *Blood* **81**, 2827–43.

Solum NO (1999). Procoagulant expression in platelets and defects leading to clinical disorders. *Arteriosclerosis, Thrombosis and Vascular Biology* **19**, 2841–6.

Ware J (1998). Molecular analyses of the platelet glycoprotein Ib-IX-V receptor. *Thrombosis and Haemostasis* **79**, 466–78.

Ware JA, Heistad DD (1995). Platelet-endothelium interactions. *New England Journal of Medicine* **328**, 628–35.

Watson SP (1999). Collagen receptor signaling in platelets and megakaryocytes. *Thrombosis and Haemostasis* **82**, 365–76.

22.6.2 Evaluation of the patient with a bleeding diathesis

Gilbert C. White, II, Harold R. Roberts, and Victor J. Marder

Introduction

The approach to the bleeding patient should be designed to determine if the bleeding tendency is inherited or acquired, systemic or local, and due to coagulation, platelet, or vessel wall defects (see Box 1). One should use the history, physical examination, and laboratory tests to derive answers to these questions.

Box 1 Questions to answer in the evaluation of a bleeding diathesis

- Is the bleeding acquired or congenital? Patients with congenital bleeding will often give a history of lifelong bleeding. This may be spontaneous or may be in the setting of previous surgical procedures or trauma. With acquired bleeding disorders, the onset of bleeding is usually recent and the patient often remarks that remote surgical procedures and trauma were not associated with bleeding. The family history is also important in determining if the bleeding is acquired or congenital. It is important to determine if unusual bleeding occurs in parents, grandparents, siblings, aunts, uncles, and children. A positive family history for bleeding favours a congenital disorder.
- Is the defect systemic or local? Bleeding from multiple sites suggests a systemic defect, whereas bleeding from a single site may result from local abnormalities.
- Is there a coagulation, platelet, or vessel wall defect? Bleeding due to coagulation defects tends to be in joints, soft tissues, or internal organs. Poor wound healing may suggest a disorder of fibrinogen or factor XIII. Platelet disorders are characterized by the presence of purpura and petechias.

The history

A history of both recent and remote bleeding should be sought. The nature of the bleeding, including the site or sites, severity, duration, association with trauma including surgical or dental procedures, and necessity for treatment of the bleeding, should be determined. Cutaneous, soft tissue, nasal, oropharyngeal, bronchial, genitourinary, gynaecological, joint, and intestinal bleeding should all be included in the history, as well as other more unusual sites. How much blood was lost? Were transfusions required? Was supplemental iron needed? If the bleeding followed surgery, what type of surgery? Excessive bleeding following tonsillectomy is common even in normal individuals and is harder to interpret than excessive bleeding following an appendicectomy. Spontaneous bleeding is more indicative of a bleeding diathesis than bleeding following a surgical or dental procedure, although persistent bleeding following trauma or surgery is often a sign of a coagulation abnormality.

The site of bleeding may provide clues to the cause. Joint bleeding is typical of haemophilia A and B and is much less common in other congenital and acquired bleeding disorders. Umbilical vein bleeding at birth is characteristic of factor XIII and fibrinogen deficiencies. Recurrent intestinal bleeding may be seen in hereditary haemorrhagic telangiectasia (Rendu–Osler–Weber disease) and is also seen in von Willebrand's disease and with certain platelet defects, but intestinal bleeding in general is most commonly due to a structural lesion. Recurrent bleeding from the same site is more consistent with a structural defect, whereas bleeding from multiple sites suggests a systemic haemostatic disorder.

The pattern of bleeding is also important. Delayed bleeding is typically seen with factor XIII deficiency and with a $_2$-antiplasmin deficiency. The initial clot forms normally, but is unstable and breaks down with normal fibrinolysis, causing bleeding after several days. With mild and moderate forms of haemophilia A and B, there is delayed bleeding and wound healing may be impaired, causing wound breakdown 5 to 7 days after surgery or injury.

The whole picture is more informative than a single event. An individual who has a lifelong history of easy bruising, bleeding following extraction of several teeth, recurrent epistaxis, and a long-standing history of menorrhagia is more likely to have a bleeding disorder than an individual that bled excessively following tonsillectomy but has no history of bruising and no bleeding with other dental or surgical procedures.

A family history is very important. It is not enough to ask if individuals in the family bleed. One should ascertain the number of brothers, sisters, aunts, uncles, parents, and grandparents who are known and determine bleeding histories for each person. In the evaluation of a male with possible haemophilia, a single brother who does not have a bleeding history is not very informative; if there are 10 brothers who have no bleeding history, the diagnosis is questionable. It is also important to determine haemostatic stresses in family members. A brother who has undergone an appendicectomy and suffered internal injuries from a motor vehicle accident without bleeding is more informative than a brother who has had no haemostatic stresses. The pattern of inheritance is also important. A sex-linked pattern of inheritance suggests haemophilia A or B. An autosomal pattern suggests other forms of inherited bleeding.

While the bleeding history is important and may provide clues to the nature of the disorder, it can also be misleading. Normal individuals will often report easy bruising that can be indistinguishable from that reported by individuals with congenital disorders. Epistaxis is frequent in children with nasal allergies whether or not there is an underlying bleeding defect and may suggest a congenital defect when there is none. Even bleeding following surgical and dental procedures can be misleading. For example, blood loss during and after tonsillectomy can be especially prominent, even in normal individuals. Blood loss after dental procedures can also be prominent, especially in individuals with poor dental hygiene.

The physical examination

The skin is a window to the blood coagulation system and careful examination of the skin is important in the evaluation of a bleeding diathesis. Ecchymosis, purpura, petechias, telangiectasia, the appearance of scars, and other cutaneous changes may provide an indication of the presence and type of bleeding disorder. The bleeding manifestations of patients with vascular disorders, thrombocytopenia, or functional platelet disorders mainly occur as spontaneous subcutaneous and mucous membrane haemorrhage and petechiae. In contrast, patients with clotting factor defects such as haemophilia develop deep, spreading haematomas, joint bleeding, and retroperitoneal bleeding.

The term 'purpura' is is a general term used to describe cutaneous extravasations of blood (Plate 1). The smallest of these are petechiase, pinpoint extravasations of blood that are round and do not blanch with pressure. They are most commonly found on the lower extremities where hydrostatic pressure is the greatest. Petechias in which there is a characteristic perifollicular distribution are typical of severe scurvy, or vitamin C deficiency. Patients with scurvy may also have excessive bleeding from multiple sites including the gums, alimentary tract, joints, and brain. More extensive cutaneous bleeding is called 'ecchymosis', or bruising, and may be seen with clotting factor abnormalities as well as platelet and vascular abnormalities.

Telangiectases are dilated small blood vessels that may be found in the skin and in the mucous membranes of the nose, lips, mouth, the whole of the gastrointestinal tract, urinary tract, and vagina in various forms of chronic liver disease, hereditary haemorrhagic telangiectasia (Rendu–Osler–Weber disease), and with hyperoestrogen syndromes. Bleeding from telangiectasia may cause recurrent epistaxis or prolonged and progressive gastrointestinal bleeding from multiple sites, which leads to refractory chronic iron-deficiency anaemia. Other cutaneous abnormalities with which bleeding may be associated are cavernous haemangiomas—large thin-walled venous abnormalities. These may precipitate chronic activation of the coagulation mechanism, causing chronic diffuse intravascular coagulation.

Cutaneous changes also characterize connective tissue disorders, including the Ehlers–Danlos syndrome, pseudoxanthoma elasticum, osteogenesis imperfecta, and Marfan's syndrome, that can all present or be associated with defects of primary haemostasis because of abnormal interactions of platelets with vessel wall connective tissue elements. The characteristics of cutanous scars may provide a clue to an underlying bleeding disorder. Congenital dysfibrinogenaemia and factor XIII deficiency may be associated with proliferative scars or keloids. Characteristic 'cigarette paper'-like scars are seen in the Ehlers–Danlos syndrome. Characteristic thickened skin papules and plaques are also seen with pseudoxanthoma elasticum, along with angioid streaks in the ocular retina.

Primary and secondary amyloid can both cause skin purpura, called 'pinch-purpura' because of the characteristic appearance on the cheeks. Amyloid may be found infiltrating the small blood vessels and has also been shown to cause platelet functional abnormalities due to membrane coating by the amyloid fibrils. The lesions often show a propensity for the periorbital tissues and on the skin may be distributed in linear streaks. Splenomegaly may also cause a mild thrombocytopenia. Occasionally there is evidence of factor X deficiency, secondary to binding to the abnormal amyloid fibrils.

Long-term administration of corticosteroids causes atrophy of the collagen fibres that support the blood vessels in the skin. This causes widespread purpura and bruises, usually on the extensor surfaces on the hands, arms, and thighs. The purpura is similar in aetiology to the senile type. A similar distribution may also be seen in Cushing's syndrome.

Certain acute infections may be associated with a purpuric eruption. The bacterial infections include meningococcal septicaemia, streptococcal septicaemia, and diphtheria. In meningococcal infection, the haemorrhage may extend to the adrenal cortex causing the Waterhouse–Friderichsen syndrome associated with acute adrenal insufficiency. If the purpuric lesions are extensive, purpura fulminans may develop with the skin lesions becoming necrotic and the patient entering an acute, shock-like state. An acute diffuse intravascular coagulation may occur, as bacterial products such as endotoxin or immune complexes will directly activate the clotting system. Immune complexes may also coat the platelet membrane causing an immune-mediated platelet destruction and interfering with platelet function. Several acute viral infections, including smallpox, chickenpox, and measles, as well as more recently described haemorrhagic fevers caused by Ebola virus, Rift valley virus, and Lassa virus may also cause similar purpuric lesions. These patients have a grossly prolonged bleeding time, thrombocytopenia, and abnormalities of platelet function.

Various allergic vasculitic purpuras are caused by inflammation and infiltration of the blood vessel wall as an anaphylactic reaction to a variety of agents including chemicals, toxins, infections, and physical stimuli. Henoch–Schönlein purpura is probably the most common and involves skin, joints, alimentary tract, kidneys, heart, and central nervous system. There is often a preceding upper respiratory tract infection caused by a b-haemolytic streptococcus producing a rising antistreptolysin-O titre. Epidemics may occur in young children, with a fever followed by a purpuric rash that is often raised to the touch and classically affects the fronts of the legs, thighs, and buttocks. In addition, the patient may develop acute arthritis, gastrointestinal pain, and nephritis associated with proteinuria. The skin lesions may continue to form over several weeks. The most serious acute complications are central nervous bleeding, acute intussusception, or renal failure. The disease is usually self-limiting but the symptoms and purpura may respond to steroid therapy. Tests of haemostasis including studies of platelet function are usually within normal limits.

Bizarre bleeding and purpuric problems are reportedly associated with several psychological factors, so-called 'psychogenic purpura'. These include self-induced bleeding, hysterical bleeding, religious stigmas, and autoerythrocyte sensitization. Most of these patients have a disturbed or overanxious personality and very often the diagnosis is only suspected after numerous investigations have been made with all the results within the normal range. Autoerythryocyte sensitization is frequently associated with severe pain preceding the onset locally of skin bleeding and bruising. The lesion may be produced by the injection of a weak solution of the patient's own washed red cells or free haemoglobin solution subcutaneously into an area of the body, such as the back, with which the patient cannot directly interfere. To distinguish this condition from self-induced injury, it is important to include a negative saline control injection.

Purpuric lesions frequently occur in normal people, usually women. Single or multiple bruises appear spontaneously, mainly on the arms or legs, which rapidly resolve without any specific treatment. Senile purpura is frequent in older people, usually on areas exposed to mild but recurrent trauma such as the backs of the hands, the forearms, and the face. The purpura is caused by atrophy of the subcutaneous tissue with progressive loss of collagen and elastin fibres in the skin leading to inadequate support of the subcutaneous blood vessels. The lesions retain their dark colour, often for several weeks. There are no abnormalities in the haemostatic screening tests.

Laboratory tests

Laboratory investigations are required to identify the precise nature of an underlying bleeding disorder after a patient with a suspected haemorrhagic state has been clinically evaluated. Laboratory tests can be conveniently divided into screening tests and special tests, with the latter applied to the study of any individual patient according to the nature and clinical circumstances of the bleeding, and following the results of prior screening assays.

Screening tests

The most common screening tests performed in the initial assessment of a bleeding tendency are the activated partial thromboplastin time (**aPTT**), prothrombin time (**PT**), thrombin clotting time (**TCT**), platelet count, and bleeding time. Table 1 and Table 2 show the results of screening tests in inherited and acquired clotting factor abnormalities.

The aPTT measures the intrinsic clotting system and the 'common pathway', the latter being the confluence of intrinsic and extrinsic systems at the point of prothrombin (factor II) conversion to thrombin (IIa). The aPTT is prolonged by deficiencies or abnormalities of high-molecular-weight kininogen, prekallikrein, factors XII, XI, IX, VIII, X, V, and II (prothrombin), fibrinogen, and by inhibitors of blood coagulation, such as the 'lupus inhibitor', heparin, and fibrin/fibrinogen degradation products. The aPTT is sensitive to activities of about 20 per cent or less of the factors listed above.

The prothrombin time measures the extrinsic clotting system. The prothrombin time is normally about 12 ± 2 s. The prothrombin time is prolonged with deficiencies of plasma factors VII, X, V, and II (prothrombin) and fibrinogen, and by inhibitors of these factors, whether iatrogenic (such as with heparin administration—acting through antithrombin III) or pathological (such as with antibodies against factor V). The test is affected by a decrease in factor VII more than by a decrease in prothrombin or fibrinogen in that it is not significantly prolonged with a fibrinogen level of about 100 mg/dl or with a prothrombin concentration above 30 per cent, but is significantly prolonged when factor VII, V, or X is less than 50 per cent of normal.

The thrombin clotting time measures the thrombin-induced conversion of fibrinogen to fibrin. A normal result requires effective release of fibrinopeptides from fibrinogen and unimpeded polymerization of fibrin monomers to form a polymer gel. The TCT is normal even in the presence of severe coagulation abnormalities that involve the coagulation pathway leading to but not including the conversion of fibrinogen to flbrin. Thus, it is normal in haemophilia (factor VIII or IX deficiency) and factor VII deficiency, and even with a decrease in multiple factors associated with vitamin K deficiency. However, the TCT is abnormal in patients with hypofibrinogenaemia or afibrinogenaemia, whether acquired or congenital, or dysfibrinogenaemia, and in the presence of inhibitors such as heparin, myeloma proteins, and fibrin/fibrinogen degradation products, which block either thrombin cleavage of fibrinopeptides or fibrin monomer polymerization.

Mixing studies using either the aPTT or the prothrombin time are used to detect the presence of an inhibitor of coagulation. If normal plasma, containing 1 unit/ml of a given clotting factor, is mixed with an equal volume of deficient plasma containing less than 0.01 unit/ml of that factor, the resulting mix will have 0.5 units/ml of the factor, and the aPTT or prothrombin time of the mixed plasmas will be normal since these tests are not sensitive to clotting factor levels of 40 per cent or above. Conversely, if normal plasma is mixed with plasma containing an inhibitor, the inhibitor will neutralize most of the factor in the normal plasma. As a result, the mix will have less than 0.5 units/ml of the factor, and.the aPTT or prothrombin time of the mixed plasmas will be prolonged. Thus, a prolonged mix is characteristic of an inhibitor, whether in the form of a non-specific inhibitor such as heparin or a lupus inhibitor or a specific inhibitor against a specific coagulation factor.

Although a prolonged mixing study is characteristic of an inhibitor, some inhibitors, especially those against factor VIII and V, display aberrant behaviour in that the aPTT mix may not be immediately prolonged. In such cases, demonstration of the inhibitor may require incubation of the mixed plasmas at 37°C for 1 to 2 h. Thus, inhibitors to factor VIII or V are said to be time- and temperature-dependent. This incubated aPTT forms the basis for the Bethesda assay for quantification of factor VIII inhibitors. Occasionally, the lupus inhibitor may display time and temperature dependence.

The platelet count is a simple first step for evaluating the cellular aspects of haemostasis. Although there is no absolute relationship between the platelet count and the frequency and severity of bleeding, spontaneous bleeding without evidence of trauma is not usually manifest unless the platelet count is less than 10 000/µl, and serious bleeding after trauma or from a local lesion is unusual at counts above 50 000/µl. If platelet-type bleeding occurs when the platelet count is above 50 000/µl, a functional platelet defect of a congenital or acquired nature may be present.

The bleeding time evaluates primary haemostatic competency and therefore reflects both platelet number and function. The bleeding time measures the interval required for haemostasis following a standard superficial incision (1 to 2 mm deep and up to 5 mm long) in the skin of the forearm, while venous pressure is maintained at 40 torr. Prolongation of the bleeding time from a normal between 4 and 7 min usually occurs at a platelet count of 35 000 to 50 000/µl, with progressive prolongation noted with greater decreases in number. At counts below 10 000/µl, the bleeding time is often 15 min or longer: The bleeding time is prolonged in thrombocytopenic states of any aetiology. The bleeding time is also prolonged in various congenital and acquired qualitative platelet disorders, such as Glanzmann's thrombasthenia, Bernard–Soulier syndrome, storage-pool disease, and drug-induced thrombocytopathies (for instance, non-steroidal anti-inflammatory agents, most commonly aspirin). In von Willebrand's disease, the bleeding time is usually prolonged, but to a variable degree, as a result of a decrease or abnormality of plasma von Willebrand factor, which is involved in the binding of platelets to matrix proteins or to other cells. The bleeding time is also prolonged in severe hypofibrinogenaemia, as fibrinogen is the principal ligand for binding platelets to each other.

Although the bleeding time is a useful test of platelet function, it is difficult to standardize. The platelet function analyser (**PFA**)-100 has recently been proposed as a sensitive and reproducible test of platelet function, which may be used as an *in vitro* bleeding time. Citrated whole blood is passed through a standardized hole in a disc made of collagen–epinephrine or collagen–ADP. The time to platelet closure of the hole is measured. The PFA-100 has been reported to be useful in the diagnosis of von Willebrand's disease and qualitative platelet defects.

Some bleeding disorders are not associated with any abnormality of the screening tests. Patients with deficiencies of factor XIII and a $_2$-antiplasmin have normal prothrombin time, aPTT, and TCT values, but specific assays for these respective proteins are available for definitive diagnosis. Since deficient factor levels of 20 to 25 per cent or less may be needed to prolong a screening test, patients with mild deficiencies may also have normal screening tests, and diagnosis is possible only by direct assay of the factor. The screening tests are also normal in disorders of the vessel wall, such as hereditary haemorrhagic telangiectasia.

Specific tests

Depending on the assessment of both the clinical facts and the screening laboratory assays, individual factor assays and other special tests may be chosen to isolate the cause of the haemostatic disorder. Specific assays for all of the clotting and fibrinolytic factors are usually available in standard coagulation laboratories. In general, functional clotting factor assays are based on the one-stage aPTT using deficient plasmas or a two-stage assay using chromogenic substrates. Immunological assays are also used to measure total levels of fibrinogen, von Willebrand factor, and proteins C and S. Interesting differences have been found between the one-stage and chromogenic assays in the measurement of levels of recombinant factor VIII that may be important in the treatment of haemophilia. The fibrinolysis system includes the inactive precursor protein, plasminogen, promoters of fibrinolysis such as tissue plasminogen activator, the serine protease plasmin, and inhibitors of both activator (plasminogen activator inhibitor) and of plasmin (a $_2$-antiplasmin). The total concentration of each protein can be measured immunologically or by functional assay, usually by fibrinolytic or chromogenic assay. Clot lysis times measure the action of plasminogen activators and plasmin in the blood and are usually performed on clots formed from the euglobulin fraction of plasma or whole blood. Abnormally short lysis times (less than 2 h for plasma euglobulin) reflect acute episodes of excessive fibrinolysis that accompany a variety of acquired disorders or that result from the administration of plasminogen activators such as streptokinase, urokinase, or tissue plasminogen activator. The degree of fibrinolysis can be assessed indirectly by the measurement of fibrin degradation products.

Aggregation of platelets is measured photometrically by changes in light transmission of a suspension of platelets in plasma following the addition of selected agonists, usually adenosine diphosphate (ADP), thrombin, adrenaline, collagen, arachidonic acid, and ristocetin. Specific patterns of response to these agonists are observed in patients with storage-pool disease, Glanzmann's thrombasthenia, and Bernard–Soulier syndrome. The evaluation of platelet granule contents and secretion relies on specialized assays for quantifying adenosine nucleotides, serotonin (both dense-body constituents), and trace quantities of platelet factor 4, b-thromboglobulin, or adhesive proteins that have counterparts in plasma such as fibrinogen, fibronectin, thrombospondin, and von Willebrand factor (a-granule components). Tests of a specific platelet enzyme system involved in the arachidonic acid pathway (such as prostaglandin synthesis) and calcium transport are usually not available outside research laboratories.

The assessment of von Willebrand's disease bridges the gap between plasma coagulation factor analysis and platelet function tests, since such patients often present with a long bleeding time and a low plasma factor VIII activity. Most such patients have an abnormal and/or decreased plasma von Willebrand factor. Since the latter serves as a carrier protein for factor VIII, the concentration of factor VIII usually parallels the decrease in plasma von Willebrand factor. The protein concentration and type of von Willebrand factor multimer may be measured immunologically by Laurell rocket electrophoresis and by a combined electrophoresis/radioautograph or immunoblot procedure, respectively. Functional assessment of von Willebrand factor activity is performed by the ristocetin cofactor assay, in which ristocetin is added to a standard source of preserved (formalinized) platelets in the presence of the test plasma, and the degree and rate of platelet aggregation are compared with those of normal plasma. Collagen binding assays are also used to measure the function of von Willebrand factor. Usually a defective von Willebrand factor is associated with a deficiency in 'ristocetin cofactor activity', although some patients have von Willebrand factor with high activity, and rare patients (those with 'pseudo-von Willebrand's disease') have von Willebrand factor-hyperactive platelets. Inhibitors to von Willebrand factor may also account for acquired von Willebrand's disease, and von Willebrand factor may also be depleted by virtue of adsorption on to cells, as in some patients with lymphoid tumours.

Special problems

Sometimes, in patients with central venous catheters or arterial lines that are flushed with small amounts of heparin, blood drawn through the catheter may become contaminated with heparin. It thus becomes necessary to distinguish a systemic coagulopathy from heparin contamination. The first clue that the abnormal clotting screens might be due to heparin contamination is the finding of a prolonged TCT and aPTT with a normal or nearly normal prothrombin time. Mixing studies will be prolonged. The easiest way to demonstrate heparin contamination is to redraw the sample, either through a peripheral vein or through the catheter, discarding the initial 5 to 10 ml. If tests on the repeat sample are normal, it can be assumed that the initial sample was contaminated with heparin. The presence of heparin in the sample can be confirmed by a reptilase time. Reptilase is a snake venom that clots fibrinogen; unlike thrombin, reptilase is not inhibited by heparin. Thus, with the effect of heparin, the TCT is prolonged, but the reptilase time is normal.

The diagnosis of diffuse intravascular coagulation also requires a special set of tests ('DIC screen'). Screening tests in diffuse intravascular coagulation will show prolonged clotting times (prothrombin time, aPTT, TCT), a decreased platelet count, and usually a low fibrinogen. Ancillary tests include fibrin degradation products, fibrinopeptide A, and fibrin D-dimer, all of which will be increased in diffuse intravascular coagulation. Fibrin degradation products are also elevated in many other disorders and are not specific for the diagnosis of diffuse intravascular coagulation. In liver disease, which may be complicated by and confused with diffuse intravascular coagulation, factor VII is decreased while factors V and VIII and fibrinogen are less affected and may even be normal.

Application of clinical observations and laboratory studies to diagnosis: illustrative cases

Case 1

A middle-aged man is admitted for surgery. A careful history reveals that on three previous occasions, dental extraction was associated with rebleeding that required repacking on several occasions. He has had no other operative procedures and no trauma of significance. The family history is negative. Because of the history of mild bleeding, coagulation screens are performed and reveal: aPTT 120 s (control 38 s), prothrombin time 12 s (control 12 s), TCT 12 s (control 12 s), platelet count 200 000/µl, and bleeding time 9 min (normal 4 to 7 min). An incubated aPTT mix reveals patient aPTT 112 s, control 36 s, mix 38 s; specific factor assays reveal: factor XII 105 per cent, factor XI 86 per cent, factor IX 94 per cent, factor VIII 5 per cent, von Willebrand factor antigen 40 per cent, and von Willebrand factor activity

25 per cent. Tests of prekallikrein and high-molecular-weight kininogen are normal.

One of the most common laboratory abnormalities is an isolated increase in aPTT. Initially, the test should be repeated, and it should be determined whether the patient is receiving small doses of heparin (minidose or heparin flushes), which can produce an aPTT prolongation. To induce the marked prolongation of aPTT in this case (120 s), relatively large amounts of heparin would be required, and other screening coagulation tests (prothrombin time and TCT) should also be prolonged. If the prolongation of the aPTT is reproducible, further laboratory studies should focus on components of the intrinsic pathway, including prekallikrein, kininogen, factors XII, XI, IX, and VIII, and inhibitors of these factors. If the patient is asymptomatic, four major possibilities should be considered: the patient may have a 'lupus' inhibitor or be deficient in high-molecular-weight kininogen, prekallikrein, or factor XII. These familial syndromes, especially high-molecular-weight kininogen and factor XII deficiency, can give a remarkably prolonged PTT in patients who have no clinical bleeding. Such an occurrence may unnecessarily deter a surgeon from operating.

If the patient has a bleeding diathesis, one should consider factor VIII, IX, or XI deficiency, or von Willebrand's disease. In this case, an incubated aPTT mix ruled out an inhibitor, and specific factor assays showed diminished levels of factor VIII and von Willebrand factor antigen and activity, data consistent with a diagnosis of von Willebrand's disease. An isolated prolongation of the aPTT is relatively common. In contrast, an abnormal prothrombin time as an isolated finding is unusual and, if the disorder is congenital, signifies factor VII deficiency. In contrast, an abnormal prothrombin time and aPTT is a common combination that may be due to isolated deficiencies of factors II, V, and X or fibrinogen and more complex conditions, such as vitamin K deficiency, warfarin therapy, diffuse intravascular coagulation, therapeutic fibrinolysis, and liver disease. Specific assays for factors II, V, and X and measurement of the thrombin time and fibrinogen degradation products may be used to define these problems. The combination of an abnormal aPTT and a long bleeding time suggests von Willebrand's disease, with the long aPTT caused by a low factor VIII level and the decrease in von Willebrand factor activity accounting for the long bleeding time, while other patients with von Willebrand's disease may have normal factor VIII activity and a normal aPTT but a long bleeding time. Variations in the same patient may even occur, which make this diagnosis more difficult in certain instances. In patients with no abnormalities of the screening tests but clinical evidence of a systemic bleeding state, one must consider factor XIII deficiency, a_2-antiplasmin deficiency, and other mild disorders.

This case raises several other important issues. First, should preoperative coagulation screening tests be performed routinely? The argument has been advanced that if a properly taken history is negative, coagulation screens should not be performed in all preoperative patients. This is a cost–benefit argument that presumes an adequate history is obtained, but too frequently this is not done. Therefore the question of routine screening tests is something that must be worked out individually at each institution. In this hypothetical case, the history was suggestive of a mild bleeding disorder, and this necessitated screening tests. Although all three components of the factor VIII complex were abnormal in this patient, the laboratory expression in von Willebrand's disease is highly variable; had the factor VIII antigen and von Willebrand factor activity been normal, von Willebrand's disease would still have been a possibility. Exclusion of von Willebrand's disease with certainty may require determinations on multiple occasions, as well as analysis of multimeric patterns and testing of family members. As for therapy, newer therapeutic modalities such as desmopressin make it possible to avoid transfusion of plasma products in some patients and thus reduce the risk of transfusion-mediated disorders such as hepatitis and AIDS.

Case 2

A 75-year-old man with severe coronary artery disease is admitted for bypass surgery. He has no history of a bleeding tendency and underwent gallbladder surgery 3 years previously without complication. Preoperative evaluation reveals a normal platelet count and normal coagulation screening tests. Surgery is without complication until the fourth hour, about 30 min after coming off the bypass pump, when increased oozing from the surgical site is noted. The activated whole-blood clotting time, which has been maintained at around 300 s during the course of surgery, is found to be greatly prolonged. Studies indicate a platelet count of 22 000/μl, aPTT more than 150 s, and TCT more than 60 s. A reptilase time is markedly prolonged, indicating that heparin alone does not account for the prolonged clotting tests. The blood smear shows about 20 schistocytes per high-power field. The presumptive diagnosis of diffuse intravascular coagulation is confirmed by factor assay, which revealed fibrinogen 75 mg/dl, factor VIII 15 per cent, and factor V 12 per cent, but factor VII 95 per cent.

The initial step in the evaluation of severe intraoperative or postoperative bleeding is to determine whether a systemic haemostatic disorder is present. The screening tests are very important in this determination because the history is often limited to immediate observations of the site, amount, and time of bleeding. If the screening tests are normal and the patient is bleeding briskly, it is unlikely that the bleeding is due to a haemostatic defect, and attention should then be turned to bleeding due to local causes. On the other hand, if the screening tests are abnormal but the patient appears to have highly localized bleeding, a structural defect may still be the cause, aggravated by the haemostatic defect. If the patient is bleeding from numerous sites, including non-operative sites, it is likely that a haemostatic defect exists.

Cardiac surgery in particular can have several effects on the haemostatic system. Coagulation defects may result from haemodilution, especially in cases in which blood loss is massive and replacement with plasma is inadequate relative to other fluids, from heparin administration during cardiac bypass, from diffuse intravascular coagulation and acute fibrinolysis, and from platelet defects or destruction as a result of interaction with the bypass circuit. In the present case, the screening tests showed marked abnormalities. Although heparin can cause similar abnormalities, the prothrombin time was greatly prolonged, and this occurs only with large heparin doses. The reptilase time, which is not sensitive to heparin, was prolonged. The normal factor VII level provided evidence against haemodilution. Although thrombocytopenia could result from consumption after cardiac bypass, it is more likely to be part of a general consumptive coagulopathy. Treatment for diffuse intravascular coagulation was initiated.

Case 3

A 25-year-old woman is referred for evaluation of lifelong easy bruising. While many of these episodes are associated with mild trauma, bruises develop without known trauma as well. There is a predilection for the extremities, but she also has noticed bruises on the trunk. Menses are heavy, and she has been given supplemental iron. Her only surgical procedure was extractions for orthodontic work at the age of 12. Six teeth were extracted over a 2-month period, and two required repacking because of bleeding. The patient's mother and a sister also bruise easily, although the mother underwent uncomplicated gallbladder surgery at the age of 50. Physical examination reveals several small bruises in various stages of healing on the arms and legs, scattered petechias, and scars on the knees from childhood trauma, which are normal in appearance. Coagulation screens reveal: aPTT 36 s, prothrombin time 12 s, TCT 13 s, platelet count 179 000/μl, and bleeding time 36 min. The blood smear shows adequate numbers of platelets that are normal in size and staining characteristics. Additional studies are obtained as follows: factor VIII 75 per cent, von Willebrand factor antigen 110 per cent with multimeric pattern normal, von Willebrand factor activity 86 per cent, and salicylate level 0. Platelet aggregation studies show a normal response to ristocetin, but adrenaline and low concentrations of ADP produce only a single wave of aggregation, and collagen and arachidonic acid produce shape change but no aggregation response. The patient is referred to a platelet research laboratory. The ADP and ATP content of platelets is normal, electron microscopy reveals normal numbers of dense granules and a-granules, and the production of prostaglandin endoperoxides and thromboxane from radiolabelled arachidonate is greatly diminished.

Easy bruising is a common symptom that is seen in patients with coagulation defects, platelet disorders, and vascular and endocrine disorders, but it is also seen in people with no underlying haemostatic defect. In the present case the initial screening tests revealed a prolonged bleeding time with a normal platelet count, indicating a qualitative platelet defect, von Willebrand's disease, or a vessel wall disorder. Some patients with von Willebrand's disease present with an isolated prolongation of the bleeding time, but this is uncommon. Defects of the vessel wall are rare but possible in patients with a prolonged bleeding time, normal platelet count, persistently normal von Willebrand factor studies (including those of family members), and normal *in vitro* platelet function. The most common cause of a long bleeding time with normal von Willebrand factor evaluation is a platelet disorder.

One can approach the diagnosis of a patient with a normal platelet count and a prolonged bleeding time in two ways. One approach studies platelet aggregation and secretion for a primary platelet abnormality, and the other measures factor VIII activity, von Willebrand factor, and ristocetin cofactor activity to diagnose von Willebrand's disease. A diverse array of normal and abnormal findings in factor VIII activity, von Willebrand factor content and multimer distribution, and ristocetin cofactor activity can exist in patients with von Willebrand's disease, but the key observation is the documentation of a plasma protein abnormality that mitigates against a primary platelet disorder of aggregation and/or release. The finding of a total absence of aggregation response to agonists such as ADP, adrenaline, and collagen is consistent with the diagnosis of Glanzmann's thrombasthenia, which can be confirmed by studies that document a defect or deficiency of the glycoprotein IIb–IIIa complex. An isolated absent response of platelet-rich plasma to ristocetin suggests von Willebrand's disease or the Bernard–Soulier syndrome, while an enhanced response occurs in the type 2B variant of von Willebrand's disease. Simple microscopic examination of the platelets may show the large size that is characteristic of the former. Impairment or lack of a second wave of aggregation and decreased dense-granule secretion (ATP or serotonin) may be due to a deficiency in dense-granule contents that may additionally be accompanied by a-granule disorders or by an abnormality in the mechanism of secretion, such as occurs with defects in thromboxane synthesis. The contents of a-granules and dense granules can be measured directly, as can the pathway of thromboxane synthesis, to identify more precisely the specific nature of the defect.

In this case, platelet aggregation studies suggested a release defect with typical inhibition of collagen and arachidonate responses. Further studies indicated that the granule compartment of the platelets was normal, as evidenced by the normal nucleotide levels and the normal electron-microscopic appearance of the platelets. Studies with labelled membrane lipids indicated that there was a defect in lipid metabolism with failure to generate prostaglandin endoperoxides or thromboxane A_2 in response to appropriate agonists. This suggests a defect in arachidonate metabolism as the cause of the release defect. Since radiolabelled arachidonate was not metabolized normally to prostaglandin endoperoxides, the most likely cause is a cyclo-oxygenase deficiency.

Case 4

A young rubber-company worker presents to the emergency department with a 2-week history of spontaneous bruising, intermittent epistaxis, and gum bleeding after tooth brushing. Four hours before admission he developed haematuria. His only medication is frusemide for hypertension. Coagulation screening tests reveal a prothrombin time of more than 60 s, aPPT more than 90.1 s, and TCT 12.8 s. The platelet count is 302 000/µl and the bleeding time is 3 min. Mixing studies reveal a prothrombin time of more than 60 s, mix 12 s. Liver function tests are normal. Specific factor assays are performed and are normal except for factor II 7 per cent, factor VII less than 1 per cent, factor IX 4 per cent, and factor X 2 per cent. A plasma coumadin level is 0. On further questioning, the patient denies the use of anticoagulants and does not have rat poison at home. Despite administration of vitamin K_1 in a dose of 15 mg subcutaneously and 2 mg intravenously twice daily, the prothrombin time and aPTT remain prolonged. The dose of vitamin K is increased to 5 mg intravenously every 4 h. Further history reveals that a new rodenticide containing a long-lasting coumarin, brodifacoum, has recently been used at the rubber-tyre company. Plasma is obtained that reveals the presence of brodifacoum at a level of 162 µg/ml. Although a supervised search of the patient and his room failed to reveal any suspicious material, his clotting times improved and eventually returned to normal after 25 days of isolation.

This case of surreptitious coumarin abuse illustrates some of the difficulties in establishing the diagnosis. Typically seen in health-care workers, the markedly prolonged prothrombin time and aPTT with a normal TCT are characteristic. The diagnosis is further suggested by the finding of reduced plasma levels of the vitamin K-dependent clotting factors in the absence of liver disease. Confirmation of the diagnosis, although strongly suggested by the previous constellation of findings, may require chemical demonstration of the drug in plasma. It is not unusual in such patients for self-administration of the drug to continue in the hospital. Once the question of surreptitious coumarin abuse is raised, one must be suspicious of the patient and family members, as well as friends and individuals at work. Psychiatric evaluation may be essential and, in cases where it is suspected that someone is trying to poison the patient, police involvement may be advised.

Brodifacoum is a member of a family of coumarin compounds termed 'super-warfarins'. They are used in rodenticides and are so termed because of their potency and very long biological half-life. For example, brodifacoum is about 100 times as potent as warfarin and has a half-life of about 22 days in humans, compared with warfarin's half-life of 1 day. As a result, a single dose of 1 to 2 mg of brodifacoum can produce a prothrombin time prolongation for up to 70 to 80 days.

In this case, plasma levels of brodifacoum were markedly increased, but a warfarin level was undetectable. The chemical structure of the coumarin compounds is such that many assays in commercial use are specific for a single compound and will not detect other coumarins. One must therefore have some clue to the agent in use and request a specific assay.

Case 5

A 60-year-old man with coronary artery disease presents with a 1-week history of easy bruising. He was initially seen 4 months earlier with new onset of atrial fibrillation. Following cardioversion, he was started on quinidine with maintenance of a stable cardiac rhythm. He denied other medications and had no history of viral or other illnesses. Laboratory evaluation reveals normal coagulation screens and a bleeding time of 14 min. The platelet count is 12 000/µl with a haematocrit of 42 and a white count of 6700/µl. A bone marrow examination is performed and is normocellular with adequate numbers of megakaryocytes. A test for antinuclear antibodies is negative. Platelet antibody testing is as follows: direct antiplatelet antibodies, 7450 molecules IgG/platelet (normal 0 to 2500); indirect antiplatelet antibodies, 1130 molecules IgG/platelet (normal 0 to 2500); and indirect antiplatelet antibodies in the presence of quinidine, 7140 molecules IgG/platelet.

Thrombocytopenia is a common cause of bleeding and easy bruising. The first step in the evaluation of a patient with thrombocytopenia is to determine if the defect is a failure of platelet production, increased platelet destruction, or splenetic sequestration. Liver disease and other potential causes of splenic enlargement should be considered and a careful examination should be made for splenomegaly. While a normal haematocrit and white count militate against global bone marrow failure, this condition can be excluded only by a bone marrow aspirate and biopsy. The biopsy in particular may be required to rule out aplastic anaemia and infiltrative processes. Other causes of thrombocytopenia, such as leukaemia, myeloma, and megaloblastic anaemias, can be diagnosed by an aspirate. If the marrow contains a normal number and distribution of cells and a reasonable quantity of morphologically normal megakaryocytes, the cause of the thrombocytopenia is unlikely to be reduced production, and one should look for causes of increased peripheral destruction.

Immune thrombocytopenia is diagnosed by the demonstration of a normal bone marrow, the absence of an enlarged spleen, and the absence of non-immune causes of thrombocytopenia. The presence of elevated levels of antiplatelet antibodies confirms the diagnosis, but elevated platelet antibodies may be seen in other causes of thrombocytopenia, and some cases of immune thrombocytopenia may not have elevated platelet antibodies. Immune thrombocytopenia may be seen in lupus erythematosus, chronic lymphocytic leukaemia, with certain drugs, especially quinidine, and by exclusion in idiopathic thrombocytopenic purpura (ITP).

Quinidine is one of the most common causes of drug-induced ITP. Patients may develop thrombocytopenia months or years after starting the drug, but re-exposure to quinidine may produce explosive thrombocytopenia. Although the diagnosis of quinidine purpura should be considered in any patient with thrombocytopenia and a history of exposure to the drug, the diagnosis can be established by either the demonstration of quinidine-dependent antiplatelet antibodies, as in this case, or by improvement in the platelet count on discontinuation of the drug. Typically, the platelet count improves within 4 to 5 days, but it may take up to 6 weeks.

Case 6

A 38-year-old woman is admitted for gallbladder surgery. Preoperative coagulation screens and platelet count are normal. Surgery is performed without complication. On the seventh postoperative day, she develops pain and swelling in the right leg, and a contrast venogram confirms the diagnosis of deep vein thrombosis. A bolus of 5000 units of heparin is given followed by infusion of 1000 units/h to maintain the aPTT in the range of 45 to 60 s. On the thirteenth postoperative day, 6 days after starting heparin, she experiences the sudden onset of increasing pain in the right arm. The arm is cool and dusky distally and the radial and ulnar pulses are markedly diminished. A platelet count is 40 000/µl. A transoesophageal echocardiogram demonstrates no evidence for a mural thrombus or for akinesis. The patient's plasma, but not normal plasma, induces aggregation of normal platelet-rich plasma in the presence of heparin, but not in the absence of heparin and contains heparin-dependent antibodies to platelet factor 4. Based on these findings, a diagnosis of heparin-induced thrombosis with thrombocytopenia is made and heparin is discontinued. The patient is started on argatroban and on warfarin. A white thrombus is successfully removed from the right brachial artery by Fogarty catheter. The platelet count improves steadily over the next 10 days. Following achievement of a therapeutic prothrombin time on warfarin, the argatroban is discontinued and the patient is discharged on warfarin for 3 months with the admonition that she should never again receive heparin.

Among the various complications of heparin, heparin-induced thrombosis with thrombocytopenia is one of the most catastrophic and dramatic. This uncommon complication of heparin therapy typically occurs 5 to 14 days after starting heparin, occurs more frequently with beef lung heparin than with porcine intestine heparin, and resolves when the heparin is discontinued. The usual presentation is isolated thrombocytopenia, but occasionally (as in this case) there is a complicating (arterial) thrombotic event, either arterial or venous. This thrombosis is unique in that it is typically a white thrombus composed predominantly of platelets and little fibrin.

While thrombosis and thrombocytopenia are dramatic complications of heparin, they are rare. The most common complication of heparin therapy is haemorrhage, which may occur in any tissue. In general, the longer the aPTT, the greater the likelihood of haemorrhage. Heparin may also aggravate haemorrhage from structural lesions, such as colonic carcinoma or gastric ulcer. For this reason, bleeding from the gastrointestinal, respiratory, or urinary tract in patients on heparin should prompt a search for a potential source of the bleeding.

The question of how to maintain adequate anticoagulation in an individual with thrombosis who cannot take heparin is an important one. Alternatives to heparin include warfarin, other antithrombotic agents such as dextran, ancrod, novel antithrombin inhibitors such as hirudin or its synthetic analogues, prepared by recombinant technology or peptide synthesis, low-molecular-weight heparinoids, and the peptide inhibitor of thrombin, argatroban. Argatroban is easiest to use since therapy is monitored using the aPTT, in a manner analogous to heparin. Hirudin and its analogues are effective, but monitoring is with a special assay using the snake venom, ecarin, because of the very high affinity of hirudin for thrombin. Low-molecular-weight heparinoids have a risk of cross-reaction with the

heparin-dependent antibody. The clinical situation largely determines which agent is used. For example, if gastrointestinal bleeding from an ulcer or polyp occurs on preventative doses of heparin (so-called 'mini-dose' heparin), the easiest course may be simply to discontinue the heparin. In the case presented, the presence of a recent deep vein thrombosis and the complicating arterial occlusion are indications to continue antithrombotic treatment.

Further reading

Collen D (1999). The plasminogen (fibrinolytic) system. *Thrombosis and Haemostasis* **82**, 259–70.

Garvey B (1998). Management of chronic autoimmune thrombocytopenic purpura (ITP) in adults. *Transfusion Science* **19**, 269–77.

George JN, *et al*. (1998). Drug-induced thrombocytopenia: a systematic review of published cases. *Annals of Internal Medicine* **129**, 886–90.

Kingston ME, Mackey D (1986). Skin clues in the diagnosis of life threatening infections. *Review of Infectious Diseases* **8**, 1–11.

Kyle RA, Bayrd ED (1975). Amyloidosis: review of 236 cases. *Medicine* **54**, 271–99.

Lak M *et al.* (1999). Bleeding and thrombosis in 55 patients with inherited afibrinogenemia. *British Journal of Haematology* **107**, 204–6.

Levi M *et al.* (1999). Disseminated intravascular coagulation. *Thrombosis and Haemostasis* **82**, 695–705.

Lundblad RL *et al*. (2000). Issues with the assay of factor VIII in plasma and factor VIII concentrates: A brief review and comments. *Thrombosis and Haemostasis* **84**, 942–8.

Mann KG (1999). Biochemistry and physiology of blood coagulation. *Thrombosis and Haemostasis* **82**, 165–74.

McMillan R (2000). The pathogenesis of chronic immune (idiopathic) thrombocytopenic purpura. *Seminars in Hematology* **37** (Suppl. 1), 5–9.

22.6.3 Disorders of platelet number and function

Kathryn E. Webert and John G. Kelton

Introduction

Platelets are the smallest of the circulating blood cells and their numbers in healthy individuals range from 150 × 10^9/litre to 450 × 10^9/litre. Platelets are released from the megakaryocytes in the bone marrow and circulate for 5 to 10 days before being cleared by the cells of the reticuloendothelial system. Disorders of platelet number and function are frequently encountered in medical patients.

Platelets are discoid cells that average 4 μm in diameter. The external membrane is a glycocalyx surface covering a phospholipid bilayer. Penetrating the membrane and traversing the platelet is a tubular system termed the open canalicular system. This system is continuous with the surface membrane and acts as a conduit for the release and uptake of substances. The platelet cytoskeleton is composed of three filamentous systems consisting of microtubules, microfilaments, and intermediate filaments. These tubules maintain the platelet's shape and participate in shape change, a complex process that occurs following platelet activation.

Platelets contain a number of organelles including alpha-granules, dense granules, lysozymes, peroxisomes, and mitochondria. The alpha-granules are the most numerous platelet granules (approximately 50 granules per platelet) and contain proteins synthesized by megakaryocytes, including b-thromboglobulin, platelet factor 4, thrombospondin, and von Willebrand factor. Alpha-granules also contain plasma-derived proteins such as fibrinogen, albumin, immunoglobulin G (**IgG**), and factor V. On the alpha-granule membrane are a variety of proteins including P-selectin and glycoprotein IIb/IIIa. Dense granules are far fewer in number than alpha-granules (four to eight per platelet) and are smaller. Dense granules are important for platelet activation and contain adenosine triphosphate, serotonin, calcium, magnesium, pyrophosphate, and granulophysin. Their membranes also contain a number of platelet proteins including P-selectin, glycoprotein Ib, and glycoprotein IIb/IIIa. Lysosomal granules contain proteolytic enzymes.

Platelet surface structures

Penetrating the platelet membrane are platelet glycoproteins. Most of these glycoproteins can be classified as one of five supergene families: integrins, leucine-rich glycoproteins, immunoglobulin domain molecules, selectins, and quadraspanins. The integrin family is the most common with glycoprotein IIb/IIIa being the most abundant integrin. Glycoprotein IIb/IIIa, also known as $a_{IIb}b_{3,}$ is present in high numbers (40 000 to 50 000 surface copies per platelet) and is the key binding site for platelet aggregation. Glycoprotein Ib/IX complex is the second most abundant platelet glycoprotein with an average of 20 000 surface copies per platelet. Glycoprotein Ib is a binding site for von Willebrand factor. A variety of other platelet glycoproteins are present in lower numbers such as glycoprotein Ia/IIa, the receptor for collagen. Finally, platelets carry 400 to 4000 copies of an IgG crystallizable fragment receptor, which is important in heparin-induced thrombocytopenia.

Thrombopoiesis

Pluripotent stem cells produce precursors of the red and white cells and the platelets. The platelet precursor is the megakaryocyte. Megakaryocytes undergo repeated nuclear replication without cytoplasmic division. This produces very large cells with four to 12 times the nuclear material of other cells of the body. Platelets bud off the cytoplasm of the megakaryocytes and are released into the circulation. The mean platelet volume can be measured on a cell counter and is approximately correlated with the number of nuclei in the megakaryocyte. Thrombocytopenia leads to increased proliferation of megakaryocytes and large platelets.

The primary regulator of megakaryopoiesis and platelet production is thrombopoietin. Thrombopoietin, an erythropoietin-like hormone, is primarily produced in the liver, with secondary sites including the kidney, bone marrow, brain, smooth muscle cells, and testes. The receptor for thrombopoietin, c-Mpl, is present on stem cells, megakaryocytes, and platelets. Binding of thrombopoietin to c-Mpl activates a variety of pathways resulting in the proliferation of megakaryocyte progenitors, an increased rate of megakaryocyte maturation, an increase in megakaryocyte nuclear mass and ploidy, and increased platelet release. The circulating level of thrombopoietin is primarily determined by the platelet mass. Platelets bind the thrombopoietin, internalize it, and degrade it. Consequently, less is available to stimulate platelet production. When the platelet count falls, less thrombopoietin is bound to platelets resulting in increased circulating levels of thrombopoietin and increased platelet production. Platelet production is also regulated by a number of other cytokines including interleukins 6 and 11.

The role of platelets in haemostasis

Platelets play a critical role in haemostasis. When the wall of the blood vessel is damaged, platelets adhere to exposed collagen and other components of the subendothelium. The key receptor is glycoprotein Ib linked to the vessel wall through von Willebrand factor. Other adhesive receptors include glycoprotein Ia/IIa, which binds collagen. The adhesion of the platelets to the vessel wall results in platelet activation. Agonists such as thrombin or adenosine diphosphate are released from their granules. The prostaglandin pathway is also activated during platelet activation; arachidonic acid is released from the platelet membrane where it is converted by a number of enzymes into platelet activating agents including thromboxane A_2. A key, rate-limiting step in this pathway is catalysed by the cyclo-oxygenase enzyme. Aspirin, an antiplatelet agent, irreversibly inactivates this enzyme. Following platelet activation, glycoprotein IIb/IIIa undergoes conformational changes making it able to bind fibrinogen. This process is termed platelet aggregation and results in the formation of the haemostatic plug. Activated platelets also contribute to the clotting cascade by serving as the phospholipid membrane surface needed for many reactions leading to thrombin generation, especially the activation of factor X by a complex of factors IXa and VIIIa and the activation of prothrombin by a complex of factors Xa and Va.

Disorders of platelet number

Thrombocytopenia

Thrombocytopenia is defined as a reduction in the number of circulating platelets to less than the laboratory's normal count (typically < 150 × 10^9/litre). Bleeding is uncommon unless the platelet count falls below 10–20 × 10^9/litre or unless there is an abnormality in platelet function.

Classification of thrombocytopenia

It is convenient to classify disorders of thrombocytopenia into problems of underproduction, increased destruction, and sequestration (Table 1). Since megakaryocytes originate from stem cells it is rare to see a deficit in platelet production without abnormalities also occurring in other cell lines. Although isolated underproduction of platelets can occur, isolated thrombocytopenia usually suggests increased platelet destruction. Platelet sequestration is usually due to splenomegaly and can cause isolated thrombocytopenia, but also causes mild leukopenia or anaemia.

History and physical examination of the thrombocytopenic patient

The physician must investigate the risk of the thrombocytopenia as well as determine the underlying cause. It is important to elicit the duration of the haemostatic impairment to determine if the patient has recently ingested an antiplatelet agent such as aspirin or alcohol, which interferes with platelet function and can trigger bleeding.

The history should be guided by the potential mechanism of thrombocytopenia. For example, if increased destruction is considered, then the patient should be questioned about drugs including prescription drugs, over-the-counter medications, herbal remedies, and illicit drugs. Secondary associations of thrombocytopenia, which include systemic lupus erythematosus, human immunodeficiency virus (**HIV**) infection, and lymphoproliferative disorders (Table 2), will lead to other questions. Finally, one should obtain information about any family members with a history of thrombocytopenia or bleeding disorders.

Physical evaluation focuses on evidence of haemostatic impairment and signs of an underlying cause of the thrombocytopenia. Many patients with thrombocytopenia are asymptomatic. Only at low platelet counts will one see petechiae, which are tiny, red collections of red cells found on dependent parts of the body and sites of trauma. Petechiae are specific for thrombocytopenia. Large bruises or purpura can be observed on the limbs and trunk and have a lower specificity. The risk of bleeding increases progressively from asymptomatic patients, to patients with petechiae and purpura, to patients who have mucous membrane bleeding, which is typically manifest by blood blisters in the mouth. Blood blisters usually occur on the bite margins of the oral mucosa and on the tongue. They indicate that the patient is at significant risk for bleeding and treatment is urgently required. The physical examination should focus on the examination of the joints, lymph nodes, spleen, and liver since abnormalities indicate a secondary cause of the thrombocytopenia.

Laboratory evaluation of the thrombocytopenic patient

One of the most important first steps is to review the peripheral blood film looking for pseudothrombocytopenia. Pseudothrombocytopenia is a laboratory artifact that causes spontaneous platelet agglutination and results in platelet clumps in the peripheral smear. Automated determination of the platelet count will be inaccurate, as the machine will not recognize the larger platelet aggregates as platelets. Pseudothrombocytopenia commonly occurs because of agglutination of the patient's platelets in ethylenediaminetetra-acetic acid. This disorder occurs in 0.1 per cent of blood samples and is caused by a clinically insignificant autoantibody that agglutinates platelets at low calcium concentrations. This can be avoided by using an anticoagulant other than ethylenediaminetetra-acetic acid to collect the blood sample.

The patient's haemoglobin and white blood cell count should be evaluated. Cytopenias involving other cell lines are suggestive of disorders involving the bone marrow such as myeloproliferative or myelodysplastic diseases. The platelet count helps to determine the patient's risk of bleeding. Patients with mild thrombocytopenia (platelet count > 50×10^9/litre) have a low risk of bleeding. Patients with severe thrombocytopenia (platelet count < 20×10^9/litre) have a higher risk of bleeding and can experience spontaneous bleeding. The peripheral smear may lead to the diagnosis of the condition causing the thrombocytopenia. Fragmented red cells or schistocytes may be seen in thrombotic thrombocytopenic purpura, haemolytic uraemic syndrome, disseminated intravascular coagulation, and renal graft rejection. Leukoerythroblastic changes in the peripheral smear, such as teardrop-shaped red blood cells, nucleated red blood cells, and immature white cells suggest infiltration of the bone marrow. The presence of abnormal circulating cells such as lymphoblasts or myeloblasts suggests a malignant process. Typical changes on the peripheral smear such as megaloblastic red blood cells and hypersegmented neutrophils suggest vitamin B_{12} or folate deficiency. The finding of atypical lymphocytes should cause one to consider the diagnosis of a viral infection. Finally, the finding of giant platelets on the peripheral smear suggests the diagnosis of certain congenital thrombocytopenias. Examination of the bone marrow should be considered if the aetiology of the thrombocytopenia is uncertain after the initial evaluation. Additionally, one should perform a bone marrow examination when abnormalities are seen on the peripheral blood smear or when multiple blood cell lineages are affected. The finding of normal or increased numbers of megakaryocytes in the marrow is supportive of a diagnosis of peripheral destruction or sequestration of the platelets. Other laboratory investigations that may be indicated include antinuclear antibody, rheumatoid factor, thyroid stimulating hormone, and testing for HIV infection.

Disorders of increased platelet destruction

Disorders of increased platelet destruction can be subdivided into two major categories: immune and non-immune. Non-immune causes include disseminated intravascular coagulation, and a variety of schistocytic or haemolytic anaemias such as thrombotic thrombocytopenic purpura. For the majority of thrombocytopenic disorders caused by non-immune mechanisms, the underlying cause is apparent and the patient's clinical presentation indicates the correct diagnosis (i.e. fever and clinical septicaemia suggest infectious causes of thrombocytopenia, fragmentation haemolysis suggests thrombotic thrombocytopenic purpura or haemolytic uraemic syndrome).

Immune mediated platelet disorders

Immune mediated disorders can be caused by autoantibodies, for example idiopathic thrombocytopenic purpura; alloantibodies, exemplified by post-transfusion purpura; and immune complexes, as demonstrated in heparin-induced thrombocytopenia. The majority of immune mediated platelet disorders are caused by IgG antibodies that bind to the platelet membrane.

Autoimmune thrombocytopenia

Autoimmune thrombocytopenia is mediated by antibodies that bind to individual platelet glycoproteins, most frequently glycoprotein IIb/IIIa. The autoimmune thrombocytopenia is classified as primary if there are no underlying conditions and secondary if it is associated with a systemic disease.

Primary autoimmune thrombocytopenia (idiopathic thrombocytopenic purpura)

Idiopathic thrombocytopenic purpura (ITP) is one of the most common autoimmune disorders that physicians manage. Idiopathic thrombocytopenic purpura is a disorder of both children and adults. In young children, frequently under the age of 5, the disease presents abruptly with dramatic evidence of haemostatic impairment. At least 80 per cent of children will have a spontaneous remission of their disease. Girls and boys are affected equally. In contrast, 80 per cent of adults who present with idiopathic thrombocytopenic purpura will have chronic disease. The disorder is typically seen in young and middle-aged adult women.

Adults with idiopathic thrombocytopenic purpura can present in one of three ways. Many patients will be asymptomatic and will have thrombocytopenia discovered incidentally. Other patients will give a history of easy bruising that may have occurred for many years and, frequently, worsened with ingestion of a substance which interferes with platelet function, such as aspirin or alcohol. Finally, patients may have an acute onset of petechiae, purpura, and mucous membrane bleeding.

Treatment of idiopathic thrombocytopenic purpura

The most important decision is whether the patient requires any treatment. If the patient has mild or moderate thrombocytopenia (platelet count > 50×10^9/litre) and no history of haemostatic impairment, we would monitor this patient with periodic platelet counts every few weeks. These patients usually maintain a consistent platelet count that tends to drop only if the patient has an immune stimulus such as an infection. The decision is more difficult in patients with more severe thrombocytopenia (platelet count (20–50) $\times 10^9$/litre) and who have modest signs of haemostatic impairment such as occasional bruising. We often do not treat these patients, but would alert the patient that the platelets should be raised before a haemostatic challenge such as a tooth extraction or surgery. Patients with severe thrombocytopenia (platelets < 10×10^9/litre) usually require treatment, especially if they have clinical signs of haemostatic impairment. The first line of treatment is corticosteroids, typically prednisone (1 mg/kg). Corticosteroids are effective in two-thirds of patients, but have predictable side-effects (Cushing's syndrome, hypertension, diabetes mellitus, osteoporosis). Corticosteroids should be given for as short an interval as possible, tapering the dose once the platelet count has reached haemostatically safe levels (> 100×10^9/litre). Patients who have a relapse of their thrombocytopenia may require more definitive treatment such as splenectomy.

Reticuloendothelial blockade through high-dose intravenous immunoglobulins (1 g/kg delivered over 6 h on two consecutive days) or anti-D in a rhesus positive individual (75 µg/kg) will result in a more rapid rise in the platelet count than corticosteroids and are indicated when platelets must be urgently raised. The major disadvantage of these treatments is that they are significantly more expensive than corticosteroids; however, they have fewer side-effects. There is a strong correlation between the response of a patient to high-dose intravenous immunoglobulins and response to a subsequent splenectomy. At least 80 per cent of patients

will respond to reticuloendothelial blockade with the peak platelet count occurring in about a week and lasting for 4 to 8 weeks.

Splenectomy

Splenectomy should be considered for patients who require ongoing medical management. Patients needing splenectomy should be vaccinated 2 weeks prior to the procedure with pneumococcal, meningococcal, and probably *Hemophilus influenzae* vaccines. The platelet count should be raised to safe levels prior to the procedure. Because of its reduced morbidity and significantly shortened hospital stay, laparoscopic splenectomy is the preferred approach. Splenectomy will result in a long-term remission or cure in about two-thirds of patients.

Second-line therapies

As many as one-third of patients will not respond to splenectomy and will require an alternative therapy. Danazol, an attenuated anabolic steroid, will induce a dose-dependent rise in platelet count in some refractory patients. The typical dose ranges from 200 to 1200 mg/day. Unfortunately, it has adverse effects including liver enzyme abnormalities and virilization. Vincristine or vinblastine have been used in refractory patients. However, if a rise in platelet count does occur, it is generally transient. Hence, the drug needs to be given repeatedly, which invariably causes dose-dependent neurotoxicity. Patients with refractory idiopathic thrombocytopenic purpura who require ongoing therapy may need aggressive immunosuppression that includes oral chemotherapy such as cyclophosphamide or azathioprine, intermittent high-dose intravenous immunoglobulins, or intermittent corticosteroids.

Emergency treatment of idiopathic thrombocytopenic purpura

Patients with idiopathic thrombocytopenic purpura who have severe bleeding require aggressive therapy including platelet transfusions, high-dose intravenous immunoglobulins, and high-dose corticosteroids, in addition to standard resuscitation including blood replacement if required.

Idiopathic thrombocytopenic purpura during pregnancy

Idiopathic thrombocytopenic purpura occurs in young women and frequently these young women will become pregnant. The majority of these patients can successfully carry a child without excessive morbidity or mortality. Typically, the platelet count falls across the pregnancy and the mother may require treatment. We use high-dose intravenous immunoglobulins since corticosteroids may be associated with an increased risk of hypertensive disorders in pregnancy. About 10 per cent of the infants born to these mothers will be thrombocytopenic with the platelet nadir occurring several days after delivery. Very severe thrombocytopenia is uncommon (< 1 per cent) and should suggest an alternative diagnosis such as alloimmune neonatal thrombocytopenia. Infant thrombocytopenia cannot be predicted by any maternal factor or serological test with the possible exception of a history of a previously affected infant. We manage these mothers with routine vaginal delivery unless there is an obstetrical indication for caesarean section.

Secondary immune thrombocytopenias

A variety of medical disorders cause secondary immune thrombocytopenia (Table 2). The treatment for secondary immune thrombocytopenia is similar to that of idiopathic thrombocytopenic purpura.

Thrombocytopenia complicating systemic lupus erythematosus

Thrombocytopenia can occur in up to 25 per cent of patients with systemic lupus erythematosus. The thrombocytopenia is usually caused by autoantibodies. Some patients will have concomitant platelet dysfunction characterized by increased bleeding and bruising. The treatment is similar to that for idiopathic thrombocytopenic purpura.

A subset of patients with systemic lupus erythematosus or lupus-like disorders have antibodies which interfere with phospholipid-dependent coagulation reactions, commonly detected by an unexplained prolongation of the patient's partial thromboplastin time. These antibodies are immunoglobulins with specificity for negatively charged phospholipids and are also called lupus anticoagulant antibodies. They tend to be heterogenous in their epitope specificity with most binding protein complexes including b_2-glycoprotein I. Another class of antibodies, the anticardiolipin antibodies, is detected by an enzyme-linked immunosorbent assay using cardiolipin as the antigen. Cardiolipin is the same antigen that is detected in the venereal disease research laboratory (VDRL) test for syphilis, which explains the false positive VDRL test in these patients. The two classes of antibodies are distinct, but have overlapping specificities. Most anticardiolipin antibodies recognize an epitope on b_2-glycoprotein I. The term 'antiphospholipid antibodies' applies to both sets of antibodies.

Antiphospholipid antibodies are associated with venous and arterial thrombosis. The antiphospholipid antibody syndrome includes any combination of arterial and venous thrombosis, recurrent fetal losses and thrombocytopenia. Many of these patients will also have a vascular rash termed livedo reticularis. Patients can have haematological abnormalities including mild thrombocytopenia, platelet dysfunction, autoimmune haemolytic anaemia and leucopenia. As the thrombocytopenia is usually mild, treatment is rarely necessary. Many patients require long-term anticoagulation therapy to prevent recurrent thrombotic events.

Thrombocytopenia secondary to lymphoproliferative disorders

Immune thrombocytopenia commonly complicates chronic lymphocytic leukaemia. This should be differentiated from thrombocytopenia of underproduction, which is seen in the spent stage of chronic lymphocytic leukaemia. Immune thrombocytopenia is often seen in patients with Hodgkin's disease and can predate or postdate the illness and is not a marker of disease activity.

Alloimmune thrombocytopenia

Alloimmune thrombocytopenia is caused by alloantibodies against platelet glycoproteins. There are two typical alloimmune thrombocytopenic disorders, alloimmune neonatal thrombocytopenia and post-transfusional purpura.

Alloimmune neonatal thrombocytopenia

Alloimmune neonatal thrombocytopenia is mediated by alloantibodies in maternal plasma directed against fetal platelet glycoproteins inherited from the father. This disorder causes severe and often life-threatening fetal thrombocytopenia that can occur *in utero*. The most common alloantibody to cause this disorder is targeted against a platelet glycoprotein called PL ^{A1} (HPA-1a) located on platelet glycoprotein IIIa.

Post-transfusion purpura

In cases of post-transfusion purpura the patient, usually a woman, develops severe thrombocytopenia 5 to 12 days after receiving a transfusion of a blood product containing platelets. The thrombocytopenia is often very severe (platelet count < 10×10^{9}/litre). Post-transfusion purpura occurs when a patient produces an alloantibody to a specific platelet antigen that she lacks, usually PL ^{A1}. The syndrome most commonly occurs in multiparous women because previous pregnancies lead to their sensitization. Patients, including men, who have previously been transfused are also at risk.

The diagnosis of post-transfusion purpura is made by the identification of a platelet-specific antibody in a patient with acute onset of thrombocytopenia 5 to 12 days after receiving a transfusion of a blood product. Although post-transfusion purpura is most commonly seen after transfusion of packed red blood cells, all blood products, including plasma, can cause the reaction. Post-transfusion purpura is self-limited with recovery occurring within 1 to 3 weeks. However, because the condition can be lethal, treatment with plasmapheresis or intravenous immunoglobulins should be considered. Platelet transfusions should be avoided except in cases of life-threatening haemorrhage.

Drug-induced thrombocytopenia

Many drugs can cause thrombocytopenia. These medications most commonly implicated include heparin, quinidine, sulfonamides, and gold. However, virtually every medication has been associated with thrombocytopenia.

Patients with drug-induced thrombocytopenia typically have moderate to severe thrombocytopenia. Thrombocytopenia is usually seen 1 to 2 weeks after beginning a medication, but it may occur in patients who have been taking the medication for several years. The platelet destruction is usually IgG-mediated. The thrombocytopenia usually resolves within days of stopping the causative drug. In cases of severe thrombocytopenia, the drug should be discontinued and the patient treated with reticuloendothelial blockade using either intravenous immunoglobulins or intravenous anti-D immune globulin. Treatment with corticosteroids is less effective. In cases of life-threatening haemorrhage, platelet transfusions may be required. Patients should not take the drug causing the thrombocytopenia again as it will cause thrombocytopenia with subsequent exposure.

Heparin-induced thrombocytopenia

Heparin-induced thrombocytopenia develops between 5 and 8 days after the initiation of heparin therapy but if the patient has been exposed to heparin within the last 3 months, it can occur earlier. Patients develop moderate thrombocytopenia (platelet counts (40–80) × 10^{9}/litre). Patients with heparin-induced thrombocytopenia frequently develop thrombotic complications, especially deep venous thrombosis and pulmonary embolism. Other clinical associations include arterial thrombosis, skin lesions, and uncommon thrombotic events such as adrenal gland thrombosis and haemorrhage.

Heparin-induced thrombocytopenia is caused by an IgG antibody, which recognizes a complex of heparin and platelet factor 4. The platelet factor 4/heparin/IgG immune complexes bind to platelet crystallizable fragment receptors causing platelet activation and microparticle formation resulting in activation of coagulation.

The frequency of heparin-induced thrombocytopenia varies among clinical settings. The risk of thrombocytopenia appears to be related to the type, dose, and duration of heparin administration. For example, unfractionated heparin is more immunogenic than low-molecular-weight heparin. Also, different patient populations have different risks of forming heparin-induced thrombocytopenia IgG. For example, the risk of heparin-induced thrombocytopenia IgG is higher in orthopaedic patients than in medical patients.

The diagnosis of heparin-induced thrombocytopenia should be considered in all patients receiving heparin therapy who develop thrombocytopenia or thrombotic complications. Serological tests can be used to confirm the diagnosis of heparin-induced thrombocytopenia. Enzyme assays measure the binding of platelet antibodies to a complex of heparin and platelet factor 4. The gold standard tests are biological assays, such as the serotonin release assay.

Treatment of heparin-induced thrombocytopenia involves discontinuation of heparin. The patient should be treated with an agent that inhibits thrombin generation, such as hirudin or argatroban. Warfarin should not be used to treat acute heparin-induced thrombocytopenia because it can trigger warfarin-induced limb gangrene.

Gold-induced thrombocytopenia

Gold-induced thrombocytopenia occurs in as many as 3 per cent of patients treated. There appears to be a genetic predisposition to the syndrome with HLA DR3 occurring in up to 80 per cent of affected patients. The thrombocytopenia usually occurs within the first several months of therapy and can range from mild to severe. Treatment involves stopping the drug and providing supportive treatment. The thrombocytopenia can persist for many months after the discontinuation of gold. This is probably due to persistence of an autoantibody, but may be due to the prolonged release of gold from tissue stores. Rapid correction of the thrombocytopenia may be achieved with intravenous immunoglobulins; however, a relapse of the thrombocytopenia may occur in 2 to 4 weeks. Patients also respond to corticosteroids. Some patients with persistent thrombocytopenia may respond to splenectomy. There is less experience using gold-chelating agents such as deferoxamine or dimercaprol.

Non-immune platelet disorders

Destructive thrombocytopenia and schistocytic haemolysis

Certain disorders are associated with both thrombocytopenia and schistocytic or fragmentation haemolysis. These disorders include thrombotic thrombocytopenic purpura, haemolytic uraemic syndrome, and disseminated intravascular coagulation.

Thrombotic thrombocytopenic purpura

Thrombotic thrombocytopenic purpura is a syndrome consisting of thrombocytopenia, microangiopathic haemolytic anaemia, renal impairment, fever, and neurological findings secondary to ischaemia. Thrombotic thrombocytopenic purpura is an uncommon disorder, but its recognition is important because it is usually fatal if not treated.

Most patients who develop thrombotic thrombocytopenic purpura are young to middle-aged with slightly more females affected than males. The presentation of illness may be insidious or acute. Typically, the patient has a several day history of generalized malaise, fatigue, or focal ischaemic problems. The focal ischaemic events usually involve the central nervous system and can include sudden weakness, paraesthesiae, and confusion. Approximately 50 per cent of patients will have a neurological event.

Most adult patients with thrombotic thrombocytopenic purpura do not have an associated underlying condition. Nonetheless, the initial evaluation of a patient with thrombotic thrombocytopenic purpura should exclude diseases associated with thrombotic thrombocytopenic purpura (Table 3). Thrombotic thrombocytopenic purpura can develop spontaneously, but is often triggered by an infection, pregnancy, or vaccination.

All patients with thrombotic thrombocytopenic purpura have destructive thrombocytopenia. The thrombocytopenia is the best indicator of disease activity. Additional laboratory investigations demonstrate abnormalities of microangiopathic haemolytic anaemia, such as anaemia, fragmented red blood cells, and increased reticulocyte count. Serum lactate dehydrogenase and bilirubin levels are elevated. Other abnormalities include elevated serum creatinine, proteinuria, and abnormal liver function tests. Investigators have identified the presence of abnormal von Willebrand factor multimers in patients with thrombotic thrombocytopenic purpura.

Thrombotic thrombocytopenic purpura is treated with plasmapheresis. This treatment has reduced the mortality from 80 per cent to 20 per cent. Plasma exchange of at least one to two volumes of plasma should be performed daily. Plasma should be replaced with cryosupernatant plasma or fresh frozen plasma. Cryosupernatant plasma may be more beneficial because it is depleted of von Willebrand factor. Plasmapheresis should be continued until the platelet count and serum lactate dehydrogenase have normalized. This generally occurs after three to ten exchanges. Plasma exchange is better than plasma infusion alone. However, when plasmapheresis is not immediately available, patients should be treated initially with plasma infusion. If the initial response to plasma exchange is poor, other therapies such as glucocorticoids may be added. Additionally, the volume of plasma exchange may be increased. Other treatments, such as antiplatelet agents, are of uncertain benefit. With discontinuation of plasma exchange, exacerbation of disease occurs in about a third of patients. This risk of relapse can be reduced by splenectomy.

Haemolytic uraemic syndrome

Haemolytic uraemic syndrome includes renal failure, microangiopathic haemolytic anaemia, and thrombocytopenia. Different types of haemolytic uraemic syndrome have been identified including classic epidemic, sporadic, hereditary and sporadic in association with non-infectious conditions. Epidemic haemolytic uraemic syndrome is seen primarily in children and occurs after a diarrhoeal illness caused by enterohaemorrhagic or verotoxigenic *Escherichia coli* serotype O157:H7 or *Shigella dysenteriae* serotype I. Haemolytic uraemic syndrome may be also associated with other bacterial, viral, and rickettsial infections. Patients have been reported to develop haemolytic uraemic syndrome after receiving immunizations.

Laboratory investigations demonstrate severe anaemia and thrombocytopenia. Examination of the peripheral smear shows fragmented red blood cells, burr cells, and spherocytes. Haemoglobinaemia and haemoglobinuria may be severe. Serum lactate dehydrogenase levels and other markers of red blood cell destruction are elevated. The serum creatinine is usually increased.

In children, the treatment of haemolytic uraemic syndrome focuses on providing supportive care with careful attention paid to fluid status and electrolyte levels. Plasma exchange should be considered in children with severe haemolytic uraemic syndrome. In adults, treatment of haemolytic uraemic syndrome generally includes plasmapheresis. Other therapies including antiplatelet agents, fibrinolytic therapy, and heparin therapy have not been shown to be beneficial, and are not recommended.

Disseminated intravascular coagulation

Disseminated intravascular coagulation is a disorder in which clotting occurs within the circulation. Disseminated intravascular coagulation is characterized by large amounts of thrombin that overwhelm the physiological inhibitors of coagulation. The thrombin causes platelet aggregation resulting in thrombocytopenia and fibrinogen cleavage into fibrin, which forms the microthrombi. The most common cause of disseminated intravascular coagulation is sepsis, but disseminated intravascular coagulation is associated with a large number of disorders including trauma and obstetric conditions (Table 4). The clinical presentation is variable, but patients with disseminated intravascular coagulation are usually very unwell presenting with fulminant bleeding and organ dysfunction. Some patients have thrombotic events. Occasionally, disseminated intravascular coagulation can be subclinical and detected only with laboratory tests. The diagnosis of disseminated intravascular coagulation is supported by the laboratory finding of thrombocytopenia in association with fragmented red blood cells, decreased fibrinogen level, and elevated fibrinogen and fibrin degradation products such as D-dimers. Coagulation studies often show a prolonged international normalized ratio, partial thromboplastin time, and thrombin time. Disseminated intravascular coagulation is best managed by identifying and treating its cause. If the patient is bleeding, replacement therapy with fresh frozen plasma, cryoprecipitate, and platelets is indicated. Heparin therapy may be of benefit in patients with clinical evidence of ongoing microvascular thrombosis.

Sepsis and infection

Transient thrombocytopenia occurs with systemic infections. Thrombocytopenia occurs in 50 to 75 per cent of patients with bacteraemia or fungal infections. It also occurs in association with viral infections, including HIV. The thrombocytopenia is generally mild to moderate and is not usually associated with symptoms of bleeding. The mechanism leading to the lowered platelet count is multifactorial including activation of platelets by bacterial products or mediators of inflammation; destruction due to immune mechanisms; or destruction due to chemokine-induced macrophage ingestion of platelets. Additionally, severe viral infections may lead to suppression of platelet production. Resolution of the platelet count occurs with eradication of the infection.

Thrombocytopenia associated with HIV is common, occurring in at least 20 per cent of patients with symptomatic disease. Various mechanisms contribute to the thrombocytopenia. Some patients have immune mediated destruction of platelets. Patients also have a defect in platelet production due to direct infection of megakaryocytes and the suppressive effects of medications. The platelet count can improve with antiretroviral therapy. Patients with severe thrombocytopenia should be treated similarly to patients with idiopathic thrombocytopenic purpura including the performance of a splenectomy.

Haemophagocytic syndrome

This rare syndrome is caused by phagocytosis of haematological cells by macrophages. Adult patients can present with an acute illness consisting of fever, weight loss, hepatosplenomegaly, pancytopenia, and increased liver enzymes. Bone marrow aspiration is diagnostic and shows morphological evidence of phagocytosis of platelets, red blood cells, and granulocytes by macrophages. The haemophagocytic syndrome may be associated with infections, particularly with the Epstein–Barr virus, T-cell lymphoma, histiocytosis, or immune disorders such as systemic lupus erythematosus and Still's disease. Treatment is directed at the underlying disorder.

Decreased platelet production

Platelet production is impaired by conditions affecting megakaryocyte progenitor cells, megakaryocytes, or the bone marrow stroma. It is rare to see a deficit in platelet production without abnormalities in the production of other cell lines as well. Decreased platelet production can occur when the bone marrow is aplastic, dysplastic, or infiltrated with other cells. Diagnosis of a defect in platelet production is usually made by evaluation of the bone marrow. Disorders causing decreased platelet production may be classified as congenital or acquired.

Congenital disorders causing decreased platelet production

Thrombocytopenia in infancy is usually due to increased platelet destruction and is only rarely due to decreased production. However, various congenital disorders may result in decreased platelet production. These disorders include congenital amegakaryocytic thrombocytopenia, thrombocytopenia with absent radii syndrome, Wiskott–Aldrich syndrome, May–Hegglin anomaly, Epstein's syndrome, Fechtner's syndrome, and Sebastian platelet syndrome. Bernard–Soulier syndrome is also associated with moderate thrombocytopenia.

Acquired disorders causing decreased platelet production

Toxins

A variety of drugs and toxins may cause bone marrow suppression and subsequent thrombocytopenia. Chemotherapy and irradiation cause direct destruction of megakaryocytes and other cells of the marrow. Other medications causing marrow aplasia are numerous and include chloramphenicol, non-steroidal anti-inflammatory drugs, antiepileptic medications, and gold.

Alcohol

Thrombocytopenia is the most common haematological abnormality associated with alcohol abuse. The thrombocytopenia can be due to hypersplenism (described subsequently) or alcohol suppression of the marrow. Alcohol induced marrow suppression can cause very severe thrombocytopenia requiring treatment by platelet transfusions. Elimination of alcohol intake will result in an increase of the platelet count within days to weeks. Associated haematological abnormalities include megaloblastic anaemia and ringed sideroblasts.

Nutritional deficiencies

Thrombocytopenia may occur with folate or vitamin B_{12} deficiency. The degree of thrombocytopenia is variable and may be severe. Associated haematological abnormalities include megaloblastic anaemia and hypersegmented neutrophils. Replacement of the deficient vitamin will result in recovery of the platelet count. Iron deficiency has also been associated with thrombocytopenia, although more frequently with thrombocytosis. Replacement of iron generally corrects the platelet count.

Infiltration of the bone marrow

The bone marrow may become infiltrated with non-haematopoietic or non-stromal cells. Conditions that may lead to marrow infiltration include metastatic cancer, haematological malignancies (leukaemia, lymphoma, myeloma), myelofibrosis, storage disorders, and granulomatous disorders (sarcoidosis, tuberculosis).

Acquired amegakaryocytic thrombocytopenic purpura

Bone marrow aplasia is characterized by hypocellularity of the marrow. Aplasia involving more than one lineage of haematopoietic cells is called aplastic anaemia. When isolated decreased platelet production occurs, it is called amegakaryocytic thrombocytopenic purpura. This rare condition frequently progresses to aplastic anaemia. Bone marrow examination reveals absent or severely decreased numbers of megakaryocytes. The disorder may be secondary to various aetiologies including drugs, toxins, and infections, but most frequently it is idiopathic. Treatment varies with the suspected aetiology and typically is supportive, but can include intravenous IgG, corticosteroids, and immunosuppressive therapies.

Myelodysplastic syndromes

Myelodysplastic syndrome can present with isolated thrombocytopenia. Examination of the bone marrow usually demonstrates abnormal megakaryocyte morphology and cytogenetic analysis reveals chromosomal abnormalities.

Disorders of platelet distribution and platelet sequestration

Splenomegaly and hypersplenism

Decreased numbers of circulating platelets may be seen in patients with splenomegaly. Normally, one-third of the circulating platelets are pooled in the spleen. With splenomegaly the size of the pool of platelets sequestered in the spleen increases, decreasing the number of circulating platelets. Increased destruction of the platelets may also occur. The thrombocytopenia is usually moderate (platelets > 40 × 10^9/litre). Bone marrow examination reveals normal numbers of megakaryocytes. Other laboratory abnormalities include leucocytosis with a normal differential and mild anaemia. Splenomegaly may be demonstrated by ultrasound or a liver–spleen scan. The diagnosis of hypersplenism can be confirmed by performing an autologous platelet survival test. This test will show a reduced recovery of transfused platelets (usually < 30 per cent) with a normal platelet survival. The thrombocytopenia is rarely severe enough to require treatment; however, splenectomy is curative.

Haemodilutional disorders

A low number of circulating platelets may also be seen in patients who have received large volumes of crystalloid solutions or blood products. This type of thrombocytopenia is commonly seen immediately after surgery and is transient. If treatment is required, the patient should receive platelet transfusions.

Extracorporeal circulation

Patients undergoing cardiopulmonary bypass commonly develop mild thrombocytopenia. The cause of the decreased platelet count is multifactorial; adherence of platelets to synthetic surfaces causes activation and damage to the platelets, haemodilution, and blood loss. The thrombocytopenia is usually mild. Generally, the platelet count recovers within 3 to 4 days to levels greater than the count preoperatively.

Hypothermia

Hypothermia is associated with transient thrombocytopenia. Decreased body temperature results in pooling of platelets in the peripheral circulation. Hypothermia may be seen in cases of environmental exposure, after prolonged surgery, and after transfusions of massive amounts of inadequately warmed blood products.

Thrombocytosis

Thrombocytosis is defined as a platelet count greater than 600 × 10^9/litre. An elevated platelet count may be primary (essential) or secondary to other disorders.

Thrombocythaemia

Primary thrombocytosis also known as thrombocythaemia is a chronic myeloproliferative disorder. Other chronic myeloproliferative disorders such as polycythaemia vera, myeloid metaplasia, and chronic myelogenous leukaemia can also cause an increase in platelet count.

Incidence and epidemiology

The true incidence of thrombocythaemia has been estimated as approximately two patients per 100 000 population per year. The average age at diagnosis is 60 to 80 years with males and females equally affected. Young women in their thirties may present with thrombocythaemia.

Aetiology and pathogenesis

Thrombocythaemia is probably a clonal process originating at the stem cell level leading to sustained proliferation of megakaryocytes with increased umbers of circulating platelets. Thrombopoietin may also play a role in the pathogenesis of the disorder. Studies have shown reduction of c-Mpl protein and messenger RNA expression. This may reflect an intrinsic defect of c-Mpl transcription or decreased receptor expression that results in ineffective clearance of thrombopoietin.

Clinical findings

Two-thirds of patients have symptoms at the time of diagnosis, typically thrombotic or bleeding. Thrombotic events are common, occurring in 20 to 30 per cent of patients. The thrombosis involves the microvasculature and patients present with headache, transient ischaemic attacks or strokes, paraesthesiae of extremities, distal extremity gangrene, and erythromelagia (burning pain and redness of the toes or fingertips). Patients with essential thrombocythaemia have an increased risk of angina pectoris and myocardial infarction. Patients at greatest risk for thrombotic events are older and have a history of thrombosis. Major bleeding complications are rare, but bruising is common.

Laboratory findings

Patients have an unexplained elevation of their platelet count, typically above 800 × 10^9/litre. Examination of the peripheral smear can reveal megathrombocytes and leucocytosis with immature myeloid precursor cells. Mild eosinophilia and basophilia can occur. Bone marrow evaluation shows increased cellularity, marked megakaryocytic hyperplasia, and clustering of megakaryocytes. In addition the megakaryocytes often are morphologically bizarre with nuclear pleomorphism. Bone marrow karyotypes are usually normal. The Polycythaemia Vera Study Group has suggested criteria for the diagnosis of essential thrombocythaemia (Table 5).

Management

Untreated, asymptomatic patients with thrombocythaemia can have a near normal life expectancy. Furthermore, the thrombotic risk in asymptomatic patients younger than 60 years of age with no history of thrombosis is not increased. Therefore, young, asymptomatic patients do not require treatment. Possible indications for treatment to lower platelet count include patients with a history of thrombotic events, patients with cardiovascular risk factors, elderly patients, and patients in whom platelet counts remain very high (> 1000 × 10^9/litre).

Low-dose aspirin can be used to prevent thrombosis and it may relieve symptoms such as headache and erythromelagia. However, aspirin may unmask bleeding tendencies so it should be avoided in patients with a history of bleeding. Hydroxyurea will lower the platelet count and usually reduces thrombohaemorrhagic complications. Adverse affects include myelosuppression and possibly an increased risk of leukaemic transformation. Anagrelide can effectively lower the platelet count, but its efficacy at reducing complications has not been definitively established. Interferon-a may also be used to lower platelet counts. Unfortunately, side-effects including flu-like symptoms, anorexia, and neuropsychiatric symptoms are severe enough to cause discontinuation of therapy in up to 25 per cent of patients.

Prognosis

The life expectancy of patients with thrombocythaemia is near normal. However, patients do have a high rate of morbidity secondary to thrombotic events. Three to four per cent of patients develop leukaemia. This occurs predominantly in patients who have been treated with alkylating agents.

Secondary thrombocytosis

Essential thrombocythaemia must be differentiated from reactive or secondary thrombocytosis. Causes of secondary thrombocytosis include infections, malignancy, chronic inflammatory bowel disease, rheumatoid arthritis, iron deficiency, and hyposplenism. Reactive thrombocytosis is not associated with symptoms related to the

elevated platelet count. Reactive thrombocytosis is not harmful and does not require treatment, although the underlying cause should be determined.

Disorders of platelet function

Congenital disorders of platelet function

Patients with congenital disorders of platelet function often present with a history of easy bruising, epistaxis, menorrhagia, and prolonged bleeding after surgery or dental procedures. Some of these patients may have family members with similar problems. The various platelet abnormalities may be classified functionally into disorders of platelet adhesion, aggregation, secretion, and procoagulant activity.

Disorders of platelet adhesion and aggregation

Platelet function disorders include Bernard–Soulier syndrome which is caused by a deficiency or abnormality of platelet glycoprotein Ib/IX and Glanzmann's thrombasthenia, caused by a deficiency of glycoprotein IIb/IIIa. Both are inherited in an autosomal recessive fashion and are very rare.

Disorders of platelet secretion

Disorders of platelet secretion occur when there are abnormalities of the platelet secretory pathways or if there is a deficiency of platelet granules. Grey platelet syndrome occurs when the alpha granules are decreased or absent. Dense granule deficiency or platelet storage pool deficiency is due to a deficiency of dense granules. In alpha delta storage pool deficiency, both the alpha and dense granules are deficient.

Disorders of platelet procoagulant activity

Platelets play an important role in haemostasis by providing a phospholipid membrane on which various coagulation reactions occur. In disorders such as Scott syndrome, abnormalities of the platelet membrane impair its procoagulant activity.

Treatment

There are no definitive therapies for any of the congenital disorders of platelet function. Administration of 1-desamino-8- D-arginine vasopressin induces the release of von Willebrand factor from endothelial cells and will improve bleeding time and haemostasis. An effect is seen within 1 to 2 h and lasts for up to 12 h. Antifibrinolytic agents, such as aminocaproic acid, may improve haemostasis. Menorrhagia may be controlled by oral contraceptive medications and perhaps by antifibrinolytic medications. In cases of life-threatening bleeding, platelet transfusions may be necessary. However, platelet transfusions can cause immunization against the platelet receptors and should be avoided.

Acquired disorders of platelet function

The most common acquired causes of platelet dysfunction are medications and toxins, systemic disorders, and haematological diseases.

Drugs

There are numerous drugs that have been shown to affect platelet function (Table 6). Aspirin has been demonstrated to cause a significant increase in bleeding. Aspirin acts by irreversibly inhibiting platelet cyclo-oxygenase resulting in decreased formation of thromboxane $A_{2,}$ an agonist for platelet aggregation. Non-steroidal anti-inflammatory agents affect platelet function by inhibiting cyclo-oxygenase. Ticlopidine and clopidrogel inhibit platelet function by inhibiting the action of platelet adenosine diphosphate. Glycoprotein IIb/IIIa inhibitors block platelet aggregation by directly inhibiting the platelet receptor for fibrinogen, glycoprotein IIb/IIIa. b-lactam antibiotics may bind to and modify the platelet membrane resulting in abnormal platelet aggregation with adenosine diphosphate, epinephrine, and collagen. Nitrates inhibit platelet aggregation. Calcium channel blockers and b-blockers affect platelet aggregation by unknown mechanisms. Other drugs that may adversely affect platelet function include antiepileptic medications, tricyclic antidepressants, and phenothiazines.

Chronic renal failure

Patients with chronic renal failure or uraemia have platelet dysfunction including defects in adhesion, aggregation, secretion, and procoagulant activity. The bleeding time may be prolonged. The pathogenesis of the platelet dysfunction is unknown, but is probably secondary to toxins present in the uraemic plasma. Treatment of a bleeding uraemic patient includes prompt dialysis. 1-desamino-8- D-arginine vasopressin will improve haemostasis. Maintenance of a normal haematocrit may decrease the bleeding tendency.

Cardiopulmonary bypass surgery

Excessive bleeding occurs in approximately 5 to 20 per cent of patients undergoing cardiopulmonary bypass surgery. Studies have demonstrated decreased platelet aggregation, altered platelet surface membrane proteins, selective depletion of platelet alpha granules, and evidence of *in vivo* platelet activation. An extrinsic platelet defect may occur resulting from thrombin inhibition by high doses of heparin. The aetiology of these abnormalities could be related to the hypothermia of the procedure and damage to the platelets as they pass through the pump system. The haemostatic abnormalities improve within hours after surgery.

Chronic myeloproliferative disorders and myelodysplastic syndromes

Disorders such as chronic myelogenous leukaemia, essential thrombocythaemia, polycythaemia vera, and myeloid metaplasia may be associated with abnormalities of platelet number and function. Abnormalities of platelet function include impaired aggregation with epinephrine, abnormal arachidonic acid metabolism, and storage pool defects. The bleeding tendency responds to treatment of the underlying disorder and correction of the associated thrombocytosis.

Dysproteinaemias

Patients with a paraproteinaemia, such as multiple myeloma or Waldenström's macroglobulinaemia, can have abnormalities in both platelet number and function. Non-specific binding of the paraproteins to the platelet membrane may interfere with membrane surface receptors. Treatment of the disorder causing the paraproteinaemia will usually correct the bleeding problem. Acutely, measures such as plasma exchange may be necessary.

Further reading

George JN (2000). How I treat patients with thrombotic thrombocytopenic purpura-hemolytic uremic syndrome. *Blood* **96**, 1223–9.

George JN *et al*. (1997). Diagnosis and treatment of idiopathic thrombocytopenic purpura: recommendations of the American Society of Hematology. *Annals of Internal of Medicine* **126**, 319–26.

George JN *et al*. (1998). Drug-induced thrombocytopenia: a systematic review of published case reports. *Annals of Internal Medicine* **129**, 886–90.

Gill KK, Kelton JG (2000). Management of idiopathic thrombocytopenic purpura in pregnancy. *Seminars in Hematology* **37**, 275–89.

Lankford KV, Hillyer CD (2000). Thrombotic thrombocytopenic purpura: new insights in disease pathogenesis and therapy. *Transfusion Medicine Reviews* **14**, 244–57.

McMillan R (1997). Therapy for adults with refractory chronic immune thrombocytopenic purpura. *Annals of Internal Medicine* **126**, 307–14.

Nurden AT (1999). Inherited abnormalities of platelets. *Thrombosis and Haemostasis* **82**, 468–80.

22.6.4 Genetic disorders of coagulation

Eleanor S. Pollak and Katherine A. High

Haemostasis, the physiological process of blood clot formation, involves a co-ordinated interaction between the wall of the blood vessel, platelets, and blood coagulation proteins. The haemostatic mechanism maintains a state of readiness to respond to a multitude of haemostatic stressors to prevent haemorrhage while also preventing inappropriate clot formation. Although acquired diseases of the coagulation system frequently occur with liver disease and other pathological disease states, this chapter will specifically focus on genetic disorders resulting from abnormalities and/or deficiencies of the blood coagulation proteins. More specifically, this chapter will cover haemophilia, von Willebrand disease, and deficiencies/abnormalities of fibrinogen and factors II, V, VII, X, XI, XII, and XIII. The role of an inherited increased risk for excess clotting will also be addressed. These conditions may result from either the loss of function of anticoagulant proteins (antithrombin III, protein C, and protein S) or a gain of function of procoagulant proteins (factor V Leiden and prothrombin 20210 G to A).

The coagulation cascade as a haemostatic mechanism

The human blood coagulation system involves a co-ordinated array of reactions which generate a stable fibrin clot when needed and prevent unnecessary clot formation. The system involves numerous proteins which interact, principally on phospholipid surfaces, to create a meshwork of fibrin fragments entrapping haematopoietic cells (Fig. 1). The majority of coagulation enzymatic complexes involve protease enzymes. Many of these enzymes are serine proteases, and a subset of these have the distinguishing feature that their functional synthesis requires vitamin K to enable post-translational modification of glutamic acid residues in the NH_2 terminal region; this property provides the basis of the therapeutic mechanism by which the drug warfarin prevents proper synthesis of functional factors. The principal enzyme balancing the pro- and anticoagulant forces is prothrombin, thought to be the evolutionary forerunner of the mammalian coagulation proteins. In addition to its procoagulant functions, prothrombin, once activated, provides anticoagulant and cellular mobility functions as well.

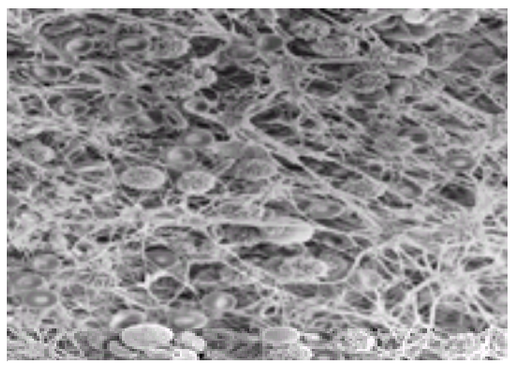

Fig. 1 Scanning electron micrograph of a whole blood clot. There is a meshwork of fibrin fibres emanating from platelet aggregates in which erythrocytes, lymphocytes, and other cells are trapped. (By courtesy of John W. Weisel and Chandrasekaran Nagaswami, Department of Cell and Developmental Biology, University of Pennsylvania School of Medicine, Philadelphia, PA, United States.)

In 1905, Morawitz first described the importance of thrombin, thromboplastin, and calcium in cleaving fibrinogen to create a fibrin clot. In the early 1930s and 1940s laboratory tests were developed that relied on *in vitro* fibrin clot formation to analyse the adequacy of a patient's clotting system. The waterfall cascade of sequential activation steps resulting in a fibrin clot was elegantly described in the early 1960s delineating separate pathways to account for the prothrombin time and the partial thromboplastin time which the earlier laboratory tests measure. However, the set of activation steps is now better described as an interwoven, reinforcing set of reactions (Fig. 2). The unique specificities of the coagulation enzymes summarized in the classical coagulation cascade have been found to be more versatile in activating diverse proteins under varied conditions. However, the separate pathways, now termed the tissue factor (extrinsic) and the intrinsic pathways, help define the steps involved in the principal tests used in clinical medicine for evaluation of haemostatic proteins. For the series of reactions and specific factors involved, the time to clot formation defines the principal parameter used in clinical evaluation of the health of a patient's coagulation system. The assays (the prothrombin time, the activated partial thromboplastin time, and activity levels of specific individual clotting factors) compare the time needed for clot formation in a patient's plasma with that in a control pool of plasma from normal donors.

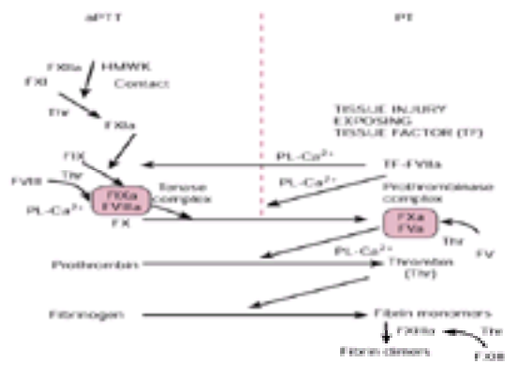

Fig. 2 Schematic representation of the enzymatic reactions involved in blood clot formation. Thr, thrombin; PL-Ca^{2+}, phospholipids/calcium; PT, prothrombin time; aPTT, activated partial thromboplastin time.

Endothelial injury and tissue damage first trigger clot formation. The response of the platelets forms the primary phase of healing by temporarily patching the site of vascular injury. Subsequent to this initial platelet phospholipid patch, a fibrin clot provides a more solid framework for the necessary but slower cellular repair. Secondary haemostasis begins with injury-induced exposure of the integral membrane protein tissue factor to plasma proteins enabling formation of the active enzymatic complex tissue factor–factor VIIa. The generation of tissue factor–factor VIIa then catalyses clotting by activating both factor X to factor Xa and factor IX to factor IXa. This activation primarily involves the cleavage of an arginine–isoleucine bond in a secreted plasma protein zymogen to form a two-chain active protein.

Thus, once tissue injury has signalled the need for fibrin clot formation and tissue factor–factor VIIa has initiated coagulation, the haemostatic process amplifies through the generation of factor IXa, which is ten times more abundant than factor VII and consequently leads precipitously to thrombin generation. Among thrombin's numerous roles is the activation of the essential procoagulant cofactors factors V and VIII. This process then further amplifies clotting by generating more thrombin through the active cofactors Va and VIIIa which then form the tenase (factor IXa/factor VIIIa) and prothrombinase (factor Xa/factor Va) complexes (see Fig. 1). Thrombin also activates the crosslinking enzyme, factor XIII, the fibrinolytic inhibitor, TAFI, and triggers platelet recruitment. Importantly, thrombin generation simultaneously counterbalances its procoagulation activities by inciting lysis of the clot via the release by endothelial cells of tissue plasminogen activator converting plasminogen to plasmin, the enzyme responsible for lysis of fibrin clots. Thrombin also dampens the clotting process by activating protein C that actively breaks down the critical procoagulant cofactors factors Va and VIIIa.

The basis for initiating clot formation in the prothrombin time and activated partial thromboplastin time test is titration of calcium into an anticoagulated plasma specimen along with a source of phospholipid. In addition, in the prothrombin time test, the source of phospholipid is a thromboplastin reagent which provides tissue factor to enable the tissue factor–factorVIIa complex to catalyse clot formation. In the activated partial thromboplastin time test the phospholipid reagent lacks tissue factor and thus prevents formation of the tissue factor–factor VIIa complex. An activator, such as silica particles, also greatly decreases the time required for clot formation through activation of factor XII via the contact activation system.

Deficiencies of specific clotting proteins

Haemophilia

Deficiency of either factor VIII (haemophilia A) or factor IX (haemophilia B), which together make up the factor VIIIa/factor IXa intrinsic tenase enzymatic complex, results in the clinical phenotype commonly known as haemophilia. A sex-linked bleeding diathesis, now thought to be haemophilia, was described in Talmudic writings as a cause of fatal haemorrhage at circumcision. In the modern era, the disease may cause bleeding at circumcision, but haemophilia principally presents with haematoma formation, easy bruising, and bleeding at the site of venepuncture during the toddler period. The disease exists in severe, moderate, and mild forms classified as such on the basis of a clinical laboratory blood coagulation test performed to assess the level of functional coagulant protein (per cent activity of factor VIII or factor IX). The pathological problem in both haemophilia A, factor VIII deficiency, and haemophilia B, factor IX deficiency (also called Christmas disease), is the inability to form a functional tenase complex to activate factor X to factor Xa. Although factor X can still be activated to factor Xa by tissue factor–factor VIIa, the available quantities of factor VII (400 ng/ml) do not allow sufficient activation of factor X to enable clotting to occur in a physiologically timely fashion. Although patients with haemophilia may have some difficulties with immediate haemorrhage subsequent to a cutaneous or superficial injury, they characteristically have joint and deep tissue bleeding problems as discussed below. The severity of disease is very well predicted by an *in vitro* assay for evaluation of the deficient protein level such that patients with severe disease have levels of factor activity of less than 1 per cent, patients with moderate disease have activity levels of 1 to 5 per cent, and patients with mild disease have activity levels of 6 to 30 per cent. Normal factor VIII and IX levels are 50 to 200 per cent and 75 to 125 per cent respectively.

Numerous genetic mutations have been described accounting for the factor deficiencies causing haemophilia. In part because of the considerable difference in size between the factor VIII gene (186 kbp), and the factor IX gene (34 kbp), the ratio of the frequency of factor VIII to factor IX deficiency is between four and five to one (approximately 186/34 kbp). Thus, the frequency of haemophilia A is approximately one in 5000 to 6000 and that of haemophilia B is approximately one-fifth of that. Among affected cases, approximately one in three to one in four patients presents spontaneously without a familial inheritance pattern. One of the only differences between factor VIII and IX deficiencies is the frequency of severe disease, which occurs more commonly in factor VIII deficiency (60 per cent of cases as compared with 45 per cent in haemophilia B). This difference is largely attributed to the frequency of mutation due to a factor VIII gene inversion in intron 22 of the 26 exon long factor VIII gene. At this locus of the factor VIII gene, a region of homology to sequences telomeric to the factor VIII gene, a recombination event results in the inability to synthesize any functional factor VIII, thus leading to severe disease (less than 1 per cent functional protein activity). In both factor VIII and factor IX deficiency, milder disease is commonly due to missense mutations.

The clinical features of haemophilia predominantly include bleeding into joints and soft tissues. The incidence of central nervous system bleeding has dramatically decreased with concentrate therapy. The life expectancy of people with severe haemophilia had increased from 11 years at the beginning of the twentieth century to approximately 60 years in the early 1980s, before the devastating effects of bloodborne viral disease again shortened average life expectancy.

In the untreated patient with severe disease, haemophilic arthropathy and joint deformity are inevitable complications. In decreasing order of involvement, the most commonly affected joints include the knee, elbow, ankle, shoulder, wrist, and hip. Recurrent bleeding episodes create a hypertrophic synovial lining with chronic inflammation; however, the pathophysiology responsible for recurrent joint bleeding remains unknown. Arthropathies commonly necessitate replacement of affected joints for pain control and improvement of mobility. Soft tissue haemorrhages frequently complicate haemophilia; further complications due to these haemorrhages include compartment syndrome, neurological damage, and extensive blood loss from retroperitoneal bleeds. Haematoma formation, a frequent complication of haemophilia, may arise spontaneously or with trauma and require extensive factor replacement and fasciotomy, the necessity for which can be assessed by mean arterial pressure in a compartment. Intracranial haemorrhage, occurring in approximately 5 per cent of patients, warrants immediate evaluation and treatment within the first 6 to 8 h of presentation; however, the majority of children presenting to an emergency room with central nervous system symptoms have not suffered from intracranial haemorrhage.

A pseudotumour, an encapsulated collection of blood most commonly originating in bone or soft tissues, is a very rare but extremely serious consequence of haemophilia occurring in approximately 2 per cent of patients. This complication is difficult to manage but may sometimes be treated with surgery at specialized haemophilia centres.

Frequently, patients with haemophilia have haematuria, the severity of which may range from self-limited episodes to gross haematuria with significant blood loss. Protease inhibitors for HIV therapy may lead to haematuria with flank pain or renal stones. Physicians should be aware of the possibility of nephrotic syndrome in patients placed on immune tolerance regimens.

Dental procedures warrant involvement of a haemophilia specialist. Factor replacement levels of 25 to 100 per cent are suggested depending on the complexity of the dental procedure. Antifibrinolytics such as e-aminocaproic acid or tranexemic acid and fibrin sealants may be a helpful adjuvant to replacement therapy.

Due to the sex-linked inheritance pattern, haemophilia is rarely found in women unless extensive lyonization takes place in the normal gene. Normal vaginal delivery is considered to be relatively safe in the case of a haemophilic infant; however, vacuum extraction and midcavity forceps deliveries and invasive fetal monitoring should be avoided due to the increased risk of formation of subgaleal and cephalic haematomas.

The laboratory diagnosis of haemophilia is based on a modification of the classic activated partial thromboplastin time assay used as a standard test for the haemostatic system. Normally patients are evaluated due to bleeding symptomatology or because of a prolonged activated partial thromboplastin time result. The activated partial thromboplastin time is a very sensitive but poorly specific screening test for haemophilia. All patients, even those with mild disease, will normally have a prolonged activated partial thromboplastin time unless there is a problem with specimen acquisition or the insensitivity of the activated partial thromboplastin time reagent. Once suspected, haemophilia can be evaluated by an inhibitor screen which involves performing a 50:50 mix of patient and normal plasma to evaluate whether the prolongation is due to a deficiency of a clotting protein or alternatively to the presence of an inhibitor. There are many causes of a prolonged activated partial thromboplastin time other than haemophilia (see Table 1). Classically, a phospholipid inhibitory antibody, called a lupus anticoagulant, will cause a prolongation of the activated partial thromboplastin time of the 50:50 mix due to the effect of the phospholipid inhibitory antibody on the normal pooled plasma. A lupus anticoagulant which causes a prolonged activated partial thromboplastin time may also result in a low factor VIII or factor IX level. In such cases, further testing for a lupus anticoagulant is necessary to rule out a low factor VIII due to a lupus anticoagulant as opposed to a deficiency.

Management of haemophilia predominantly involves administering the missing protein (factor VIII or factor IX) to a patient. Factor replacement therapy is most commonly administered in a so-called 'on-demand' regimen, when a patient's symptomatology necessitates treatment. However, prophylaxis is indicated during surgery or at times of expected injury. Prophylactic therapies during early childhood are now recommended when feasible after the first major bleeding episode as a means of preventing arthropathies in patients with severe disease. Prophylactic administration during the first few years of life requires special consideration due to the need for repeated intravenous access generally requiring an indwelling line. These have been associated with high rates of sepsis, particularly in children under the age of 3. Before the development of stringent purification and virucidal procedures, the transmission of viral disease was almost inevitable as each vial of plasma-derived concentrate was pooled from approximately 60 000 to as many as 400 000 donors, although the number has recently been reduced to 15 000. Tragically, the majority of patients with severe disease treated before 1985 developed HIV. Rates of development of hepatitis B and C are also extremely high. Although drastically reduced; the potential for transmission of infectious disease has not been totally eliminated. Many recombinant preparations are prepared with

human serum albumin thus leaving a possible source of transfusion of a bloodborne disease.

Treatment

Acute bleeding episodes

Safe and effective treatment options continue to improve for the management of acute bleeding episodes for patients with haemophilia A and B. Blood products available include fresh frozen plasma which contains both factors VIII and IX, prothrombin complex concentrates containing factors II, VII, IX, and X, activated prothrombin complex concentrates (factors IIa, VIIa, IXa, Xa), monoclonal-antibody purified factor VIII and factor IX, and recombinant factor VIII and factor IX. Recombinant factor VIIa is now approved for use in patients with inhibitors during acute bleeds. Currently trials using gene therapy approaches are under way and may provide a method for continuous prophylaxis against bleeding. Recombinant or highly purified products are the optimal therapy because of the great benefit to risk ratio. Availability, ease of administration, cost, viral safety, and thrombotic risk, particularly in patients undergoing high-dose therapy or procedures with a high risk of thrombotic complications, dictate the choice of product. Cryoprecipitate, made from the precipitate of thawed frozen plasma, contains factor VIII but does not contain factor IX. Cryoprecipitate and fresh frozen plasma should only be used in the haemophilia patient in an emergency setting where concentrates are not available. Inhibitor formation, the development of antibodies to the deficient protein, arises subsequent to transfusion of a blood product or factor replacement and is the major complication of treatment. An inhibitor presents an extremely difficult situation for patient management (see Complications of therapy below).

Several immunoaffinity purified plasma-derived factor VIII and factor IX products are available in the United States and Europe and currently have excellent records of viral safety, efficacy, and lack of thrombogenicity. When concentrate is unavailable, fresh frozen plasma is readily available in most emergency settings. Virucidal methods using solvent detergent treatment may now be applied in production of fresh frozen plasma; furthermore, each unit is from a single screened donor, thus the risk of transfusion-transmitted disease is low.

Recombinant factor VIII and factor IX have been licensed for nearly a decade. These proteins are produced in cultured mammalian cells and purified from conditioned medium. Recombinant factor IX is devoid of human plasma whereas the recombinant factor VIII concentrates utilize human plasma-derived albumin for stabilization. Because *in vivo* coagulant activity of recombinant factor IX is only 80 per cent of *in vitro* estimates used for labelling of product in IU/mg, it is recommended that the calculated factor IX dosage be multiplied by a factor of 1.2 for dose calculation when using recombinant factor IX. A plausible explanation for this discrepancy is a difference in post-translational modifications compared with plasma-derived factor IX.

During severe and critical bleeds it is optimal to achieve 50 to 100 per cent factor activity levels for 7 to 10 days (for example for pharyngeal, retropharyngeal, retroperitoneal, and central nervous system bleeds). More modest levels of 20 to 50 per cent for 2 to 7 days are generally adequate for dental extractions, haematuria, intramuscular or soft tissue bleeds with dissection, or bleeds of the mucous membranes. Levels of 20 to 30 per cent for 1 to 2 days are recommended for uncomplicated haemarthroses, superficial muscle, or soft tissue bleeds. The frequency of dosing is every 12 to 24 h for factor IX concentrates and every 8 to 12 h for factor VIII concentrates. At 24 h for factor IX and 12 h for factor VIII, the calculated amount to infuse would be one-half the initial amount of factor IX, as the half-life of factor IX is approximately 18 to 24 h.

The timing of factor level determination should be 15 to 30 min after the loading dose and immediately prior to subsequent doses for appropriate dose adjustments. When factor concentrates are used for patients with inhibitors, higher doses will most likely be required. Additionally, some authors have also reported good success and reduced cost with constant infusion regimens.

Calculation of the optimal factor concentration for administration

The number of IU (International Units) of factor required is equal to:

$$\text{body weight (kg)} \times \text{desired \% increase in factor VIII or IX level} \times C.$$

C is a constant depending on the product and the source of the product: *C* is equal to 0.5 for administration of plasma-purified and recombinant factor VIII, *C* is equal to 1 for administration of plasma-purified factor IX,* and *C* is equal to 1.2 for administration of recombinant factor IX.†

Surgery in patients with haemophilia

When possible, treatment should be instituted by caregivers aware of major and minor adverse reactions and complications occurring in the haemophiliac population. Care should also be given in association with an experienced reference laboratory able to provide timely evaluation of a patient's response to treatment. Therapeutic factor levels should be obtained prior to surgery. Depending on the type of surgery, the factor level should reach levels of 50 to 100 per cent of normal and should be maintained 2 to 7 days postprocedure. In addition to factor concentrates, fibrin glue has been recommended with circumcision, antifibrinolytics with dental procedures, aprotinin with cardiac procedures, and recombinant factor VIIa and/or apheresis in patients with high-titre inhibitors. The use of aprotinin, however, has been proposed to increase the risk of thrombogenicity. A factor level approaching 100 per cent is recommended for brain or prostate surgery, because of a higher risk of bleeding. In patients with milder haemophilia A administration of the synthetic octapeptide desmopressin (**DDAVP**) may be helpful in increasing factor levels. However, this is not the case for haemophilia B.

Complications of therapy

The main adverse outcomes related to treatment with concentrates include transmission of viruses when using plasma-derived products and development of inhibitory antibodies seen with both recombinant and plasma-derived products. Thrombosis has also been a complication of early complex concentrates used for patients with inhibitors.

The development of purification schemes which inactivate viruses, and the development of recombinant products, has dramatically decreased the incidence of transmission of viral disease. Early preparations of prothrombin complex concentrates presented a significant risk of thrombotic complications, but this risk has now been markedly reduced.

Inhibitor formation—the development of antibodies that inhibit clotting activity—occurs subsequent to transfusion of a blood product or factor replacement. An inhibitor presents a difficult situation for patient management. Therapeutic strategies largely rely on the ability to bypass the factor VIIIa–factor IXa tenase complex. Inhibitor formation almost exclusively arises in severely affected patients and occurs in approximately 7 to 52 per cent of patients with haemophilia A but in only about 1 to 3 per cent of patients with haemophilia B. This difference in inhibitor formation is not completely understood, but possible explanations include the higher incidence of severe disease in haemophilia A, prenatal exposure to maternal factor IX but not factor VIII antigens due to the former's ability to pass the placenta, the structural similarity between factor IX and other vitamin K-dependent proteins, the higher plasma levels of factor IX, and the greater inherent immunogenicity of factor VIII due to its larger size. One very rare but severe complication that may occur with the development of a factor IX inhibitor is the development of a potentially life-threatening anaphylactic complication following first treatment.

Therapeutic strategies for treatment of an inhibitor include acute management of the bleeding episode as well as a longer-term treatment directed toward suppression of antibody production. Quantification of the titre of factor inhibitor involves mixing the patient's plasma with test plasma containing a known amount of factor, normally from a pool of healthy donors. After incubation, factor activity levels present in the patient's incubation mixture are compared with that in a control mixture so that the amount of inhibitory antibody can be calculated. The Bethesda unit (**BU**), the standard unit used to report a titre of factor inhibitor, represents the amount of inhibitor that inactivates 50 per cent of factor activity. Acute management of bleeding in a patient with an inhibitor relies first on quantifying the BU of the inhibitor. With low-titre inhibitors (less than 5 BU), it may be possible to overwhelm the inhibitor with aggressive concentrate therapy. With high-titre inhibitors, it is usually necessary to bypass the inhibitor using either prothrombin complex concentrates, activated prothrombin complex concentrates, porcine factor VIII, or, more recently, recombinant factor VIIa. These bypassing agents, with the exception of porcine factor VIII, largely work by directly activating factor X to factor Xa, thus bypassing the need for the intrinsic tenase complex. With porcine factor VIII, it is wise to first perform testing to ensure that the patient's inhibitor does not crossreact with the porcine factor VIII. Because of the life-threatening bleeding complications in patients with inhibitor, immune tolerance regimens have been designed with the aim of eradicating the inhibitor in the long term. Therapeutic regimens such as the Malmö protocol involve infusion of high-dose factor concentrates along with immunosuppressive agents enabling a tolerance to the deficient factor. This protocol is highly effective in approximately 80 per cent of patients.

Viral diseases

Severely affected patients treated with plasma-derived concentrates before 1985 had an extremely high rate of viral disease. Over the course of a 70-year lifespan, a patient with severe haemophilia may be exposed to donations from 70 million individuals due to the pooling of thousands of donor units for concentrate production. Specific laboratory tests to screen for HIV, hepatitis C, and hepatitis B, in addition to much improved donor screening procedures, have dramatically limited the number of contaminations from individuals carrying viral diseases. Solvent detergent treatment procedures, which inactivate enveloped viruses (HIV and hepatitis viruses B, C, D, and G), and heat treatment procedures used to eliminate non-enveloped viruses (hepatitis A and E viruses and parvovirus B19) have radically decreased the risk of viral infection from plasma-derived products. The viral inactivation procedures in current use include pasteurization, vapour heating, high-dry heating, and nanofiltration. Recently, b-propiolactone ultraviolet inactivation has been discontinued due to ineffective virucidal technique.

HIV

In the late 1970s and early 1980s, HIV, the human immunodeficiency virus, appeared in the blood supply before routine laboratory testing was developed to detect its presence. The leading cause of death in American haemophilia patients in 1982 was haemorrhage; however, contaminated blood products during the period between 1979 to 1983 led to a sharp rise in viral disease shortly thereafter. A large proportion of patients with haemophilia became infected with HIV and have subsequently died from AIDS. Risk factors for infection included the severity of the disease (severely affected patients were much more commonly affected than those with moderate or mild disease), the type of concentrate used (factor VIII versus factor IX concentrate), the viral inactivation procedures used in product preparation, and the geographical location of the patient with regard to percentage of blood products contaminated.

The incidence of HIV infection in American patients who received plasma-derived concentrates between 1979 and 1984 was lower in patients receiving factor IX complex concentrates (55 per cent) than in those receiving factor VIII concentrates (approximately 90 per cent). Despite the devastating consequences of HIV for affected individuals and families, the projected impact on births of patients with haemophilia over the next two centuries is small (1.79 per cent reduction).

Hepatitides

Contaminated plasma-derived products also led to significant morbidity and mortality due to hepatitis viruses. Effective virucidal techniques have greatly reduced the incidence of hepatitic viral disease in this population. In the United States in the late 1980s 87 per cent of the 345 HIV negative, and more than 99 per cent of the HIV-positive patients showed evidence of prior infection with hepatitis B, hepatitis C, or hepatitis D viruses. Infection due to hepatitis A virus has rarely been reported in patients with haemophilia in the United States. Solvent/detergent inactivation of concentrates has been associated with a high prevalence of antibodies to hepatitis A virus.

Hepatitis B was commonly seen in patients with haemophilia until routine screening of liver enzymes and the subsequent availability of hepatitis-specific antibody and antigen tests in the 1980s. Most patients are now vaccinated against hepatitis B so that it is difficult to estimate hepatitis B infection from concentrate administration. The hepatitis delta virus, dependent on coinfection with hepatitis B virus, has also been a significant cause of morbidity in patients with haemophilia; its prevalence is largely attributed to the administration of prothrombin complex concentrates.

Routine testing for hepatitis C, instituted in the early 1990s, has reduced but not eliminated hepatitis C contamination in the donor pool. A variable susceptibility and morbidity is seen in response to hepatitis infection; cirrhosis was estimated at approximately 20 per cent and liver failure at 10 to 20 per cent 20 years after infection. Concurrent infection with HIV can accelerate complications of hepatitis C virus. There is also an increased likelihood of hepatocellular carcinoma with long-term infection with viral hepatitis.

Other infectious agents

The vast majority of patients with haemophilia have antibodies to parvovirus B19. Parvovirus B19 is a small, non-lipid enveloped, highly heat-resistant virus found to contaminate plasma-derived products. Methods that inactivate other non-lipid enveloped viruses in products have not proven to be routinely effective against this virus. Although parvovirus B19 infection is often mild and self-limited, infection with parvovirus B19 has the potential to severely compromise the health of an infected immunodeficient patient.

There is experimental evidence in animal models that cellular blood components, plasma, and plasma components have a potential, though minimal, risk of transmitting the prion disease Creutzfeldt–Jacob disease. To date, no definitive direct infection of a recipient of a blood product or blood product concentrate has been documented, although transmission has been reported from corneal and dura mater grafts and cadaveric pituitary hormones. The American Red Cross currently administers a questionnaire to screen donors for risk of prion disease. There is now concern that transmission of new variant Creutzfeldt–Jacob disease may differ from classical CJD and could potentially be transmitted through plasma-derived concentrates. This new variant has been associated with outbreaks of bovine spongiform encephalopathy, potentially from dietary exposure. Experimental evidence shows that bovine spongiform encephalopathy and new variant Creutzfeldt–Jacob disease are caused by the same infectious agent, and prion-related protein has been found in lymphoid tissue of patients with new variant Creutzfeldt–Jacob disease. In Europe, particular lots of concentrate have been removed from the market due to the development of new variant Creutzfeldt–Jacob disease in product donors.

Treatment of patients infected with hepatitis virus and/or HIV

Vaccination against hepatitis A and B is highly recommended for patients who receive concentrates and lack viral antibodies indicative of past infection. Treatment of hepatitis C with interferon-g is associated with significant improvement in approximately half of patients in many but not all studies. Liver transplantation has been successful in many cases for patients with liver failure who are unresponsive to treatment. The liver transplant fortuitously corrects the deficiency of clotting protein due to synthesis of clotting factors by the orthotopic liver. However, the possibility of reinfection with viral disease is significant and must be included in management decisions.

Drug-related hepatitis in haemophilia patients has been reported subsequent to treatment of HIV, particularly in response to indinavir. Additionally, complications in HIV-positive haemophilia patients taking protease inhibitors include haematuria, intracranial bleeds, and excessive bleeding often requiring hospitalization and administration of higher than expected doses of factor concentrate to correct the bleeding. Protease inhibitor therapy should not be withheld from HIV-positive individuals with haemophilia. A 6-month, prospective study of 20 haemophilia patients receiving protease inhibitors revealed only one unusual bleed which was corrected by factor infusion.

Gene transfer as a method of treating haemophilia

The development of clotting factor concentrates resulted in a dramatic improvement in life expectancy for individuals with haemophilia. None the less this treatment strategy has a number of disadvantages. The protein must be infused intravenously, and has a relatively short half-life in the circulation. This makes chronic prophylaxis difficult, especially in small children where venous access may present a problem. In addition, the product is expensive so that only about one-third to one-half of the world's haemophiliacs (those in the developed world) have access to the product. Although current viral inactivation techniques have largely eliminated the risk of HIV and hepatitis, there are ongoing concerns about the risk of other bloodborne diseases (Creutzfeld–Jacob disease, transfusion-transmitted viruses) that are not easily eradicated using current techniques. These factors have fuelled interest in the development of a gene transfer approach to the treatment of haemophilia. Such an approach, if successful, would result in continuous production of a level of clotting factor adequate to prevent bleeds rather than treating bleeds after they have occurred. The level of clotting factor required for this goal can be predicted based on a generation of experience with clotting factor concentrates. Thus in Swedish prophylaxis studies, it has been shown that maintenance of trough factor levels in the range of 1 to 3 per cent are adequate to prevent all the life-threatening bleeds and most of the joint bleeds in boys with severe haemophilia. The validity of a target of 1 to 3 per cent is further confirmed by the natural history of the disease; individuals with factor levels of less than 1 per cent are severely affected, whereas those with levels of 1 to 5 per cent have a moderately severe phenotype with a considerably lower incidence of spontaneous bleeding episodes.

Successful gene transfer approaches require three elements: a therapeutic transgene, a means of delivering it, i.e. a vector, and an appropriate target cell type in which gene transfer and expression will exert a therapeutic effect. Of the inherited diseases for which gene transfer approaches have been attempted, haemophilia has a number of advantages. First, tissue-specific expression is not required. Although clotting factors are normally synthesized in hepatocytes, biologically active

material can be synthesized in a variety of tissues, including fibroblasts, muscle cells, and endothelial cells. This allows latitude in the choice of target cell. Second, the therapeutic window is wide, since even small increases in circulating levels of factor are likely to result in some improvement in symptoms, and increases to 100 per cent would still leave the patient within normal limits. Excellent small and large animal models of the diseases exist (murine and canine), and determination of therapeutic efficacy is in the case of haemophilia relatively straightforward, since levels of circulating factor correlate well with symptoms of the disease.

A number of different strategies for gene therapy for haemophilia are under active investigation in preclinical studies, and three clinical trials are currently under way, with two more in late planning stages. The plethora of approaches suggests that there will be more than one successful combination of vector and target tissue that is safe and effective for haemophilia. The ongoing trials include an *ex vivo* approach in which fibroblasts from a patient are isolated from a skin biopsy, cells are expanded in culture, transduced with a retroviral vector expressing factor VIII, and then reimplanted onto the patient's omentum. Early data from this trial show safety in the first six patients treated, with evidence of gene transfer and expression in three out of six. In a second trial, a retroviral vector expressing B-domain deleted factor VIII is infused intravenously. This trial has enrolled 13 subjects with severe haemophilia A. There has been no evidence of safety concerns to date. Finally a third trial involves intramuscular injection of an adenoassociated viral vector expressing factor IX. This approach has been demonstrated to be safe and effective in animal models of haemophilia and is currently being tested in patients with severe haemophilia B. Two other planned trials both use liver as the target cell; one uses a gutted adenoviral vector to express factor VIII, and the other uses an adenoassociated viral vector to express factor IX in liver. The trials will determine whether any of these approaches will be safe and effective in patients with haemophilia.

Von Willebrand disease

In 1926 Erik von Willebrand first described what we now know as von Willebrand disease upon finding an autosomally inherited bleeding diathesis in a large kindred on the Aland Islands in the Gulf of Bothnia between Sweden and Finland. Although the bleeding disorder in this family resulted in haemorrhagic death in multiple family members, the bleeding diathesis in patients with von Willebrand disease is usually much milder. Most commonly, patients with von Willebrand disease manifest mucosal platelet-type bleeding tendencies of varying severity. Nose bleeds, menorrhagia, and easy bruising are the most common manifestations.

The pathophysiology of von Willebrand disease involves a functional deficiency of von Willebrand factor, a 270 kDa monomer that forms a large multimeric plasma glycoprotein of several subunits up to 100 subunits. von Willebrand factor, synthesized in the megakaryocyte and endothelial cell and stored in subcellular granules, enables proper two-chain factor VIII formation and serves as a carrier, thus preventing degradation of factor VIII and lengthening the half-life of the labile factor VIII protein to around 8 h. von Willebrand factor secreted by endothelial cells also binds to heparin glycosaminoglycan and to the platelet glycoprotein complex Ib–IX enhancing platelet activation and further platelet recruitment at sites of tissue damage. The interaction between platelets and von Willebrand factor is thought to provide the explanation for the mucosal bleeding phenotype occurring in patients with von Willebrand disease. Patients with von Willebrand disease frequently have reduced levels of factor VIII. However, the remaining factor VIII is normally sufficient to prevent the haemophilia-type symptomatology of arthropathy and deep tissue bleeding.

The gene for von Willebrand factor is located on chromosome 12, is 180 kbp in length, and consists of 52 exons. There are three types of von Willebrand disease: types 1, 2, and 3. Types 1 and 3 are quantitative deficiencies of the von Willebrand factor while type 2 is a qualitative deficiency due to binding defects of the von Willebrand factor. The inheritance of types 1 and 3 are autosomal dominant and autosomal recessive, respectively. However, rare reports of an autosomal dominant inheritance pattern for type 3 have been published. There are four principal subtypes of type 2 classified as follows: 2A, absence of high molecular weight von Willebrand factor species causing decreased platelet-dependent function; 2B, increased affinity of von Willebrand factor for platelet glycoprotein Ib–IX; 2M, platelet functional defect not caused by the absence of high molecular weight multimers; and 2N, a factor VIII binding abnormality.

Laboratory diagnosis of von Willebrand disease involves assaying the plasma for von Willebrand factor. The two principal tests are an antigenic test (von Willebrand factor anitigen) and an activity test (von Willebrand factor RCo) in which formalin-fixed platelet aggregation is induced due to the ristocetin-enhanced von Willebrand factor binding to glycoprotein complex Ib–IX. Comparison of the tests helps identify the enhanced ristocetin-induced aggregation seen in type 2B von Willebrand disease where von Willebrand factor ristocetin cofactor is typically much lower than von Willebrand factor antigen. Other tests performed in the evaluation of von Willebrand disease include the level of factor VIII, which is often decreased, and the activated partial thromboplastin time, which is elevated in approximately half of cases of von Willebrand disease due to the low activity of factor VIII. Non-reducing gel immunoelectrophoresis is employed to assay the distribution of multimeric subunits of von Willebrand factor with a gel containing antibody to von Willebrand factor antigen. This assay is particularly relevant for visualization of the presence of low, intermediate, and high molecular weight von Willebrand factor subunits. The intermediate and high molecular weight species are markedly decreased in subtypes of type 2 disease. A decreased normal pattern is seen in type 1 disease, although the decreased visual intensity may be difficult to quantitate. Type 3 disease shows near absence of all subunit molecular weights. The lower limit of the normal range of von Willebrand factor varies with blood type (A, B, O, AB). Thus, symptomatology must be evaluated based on normal ranges for each blood type. The bleeding time in a patient with von Willebrand disease is most often prolonged; however, the test is no longer routinely necessary because of the non-specific nature of a positive result and the higher specificity of other testing.

The specific treatment for von Willebrand disease varies with a patient's symptomatology, the circumstances of the need for treatment, the subtype of von Willebrand disease, laboratory results indicating the potential success of increased von Willebrand factor with non-protein based treatment, and the clinical experience with a particular patient and his or her biological family members. When possible, treatment based on non-blood products is preferred The mainstay of treatment for mild disease is treatment with the synthetic octapeptide DDAVP, desmopressin. DDAVP causes release of factor VIII and von Willebrand factor from endothelial cells raising the plasma von Willebrand factor by approximately two- to tenfold. Thus treatment with DDAVP relies on a partial quantitative deficiency of von Willebrand factor. Intravenous and nasal preparations are available. The nasal preparation allows a patient to self-administer medication at either regular intervals or on an as-needed basis. The phenomenon of tachyphylaxis, the decreased effectiveness of repeated doses of the compound, does occur, and there is usually little response after three consecutive doses. In the past, DDAVP was considered contraindicated in type 2B von Willebrand disease because of the thrombocytopenia sometimes observed with DDAVP infusion. However, this recommendation is controversial and should be assessed on a case-by-case basis. Patients with type 3 von Willebrand disease may lack sufficient intracellular reserves for effective therapy; thus alternative measures for such patients are usually necessary.

A trial of effectiveness of DDAVP is often indicated, particularly prior to prophylactic surgical use of the compound. The trial is normally performed after subtyping the von Willebrand factor disease to ensure that DDAVP is not contraindicated, as in type 2B. Optimally, the test should not be given within 24 h of the last DDAVP infusion nor at a time of environmental stress in order to minimize problems associated with tachyphylaxis or depletion of intracellular reserves. A therapeutic trial entails measurement of von Willebrand factor antigen before and 1 h after DDAVP infusion of 0.3 µg/kg. The patient should be watched carefully during this period because of possible flushing, mild anaphylactoid reactions, and possible hyponatraemia.

e-aminocaproic acid is frequently administered in the setting of dental surgery to inhibit fibrinolysis. However, care must be taken in administration to patients with a predisposition to thrombosis because of the potential deleterious effects of e-aminocaproic acid in this setting. Other compounds which may be administered include oestrogens in women because of the natural positive regulation of synthesis of von Willebrand factor with oestrogen compounds. This may ameliorate menorrhagia in such patients. Components in cryoprecipitate include factor VIII, fibrinogen, and factor XIII, in addition to von Willebrand factor. Cryoprecipitate had been the mainstay of plasma-based therapy until the recent availability of factor VIII concentrates with preserved von Willebrand factor protein such as Alphanate and Humate P. The use of cryoprecipitate, which does not undergo viral inactivation, has thus fallen out of favour.

Treatment of von Willebrand disease with DDAVP is the method of choice in patients who respond to this therapy. DDAVP for intravenous or subcutaneous use is supplied as either a 4 µg/ml 10 ml vial or a 15 µg/ml 1 or 2 ml vial preparation. The recommended dose is 0.3 µg/kg, mixed in 30 ml normal saline, infused slowly over 30 min or 0.4 µg/kg subcutaneously. This dose may be repeated after 12 to 24 h. A DDAVP nasal spray is available in a metered dose pump which delivers 0.1 ml (150 µg) per actuation. The bottle is at a concentration of 1.5 mg/ml and contains 2.5 ml with a nasal spray pump which can deliver 25 150 µg or 12 300 µg doses. For administration, patients who weigh less than 50 kg should deliver one 150 µg spray in one nostril. For those weighing over 50 kg, one spray should be delivered in each nostril for a total dose of 300 µg. Administration may be repeated after 24 h. Precautions to take with the medication include administration no more than every 24 h or for three consecutive days unless under the supervision of personnel from a haemophilia treatment centre. The medication should not be used in pregnant women or in children under 2 years of age. The medication should be used with caution in the elderly and in individuals with a history of cardiovascular disease.

Factor XI deficiency

Factor XI deficiency is an autosomal recessive bleeding diathesis of variable severity. It was first described in 1953 as a third type of haemophilia and is thus sometimes referred to as haemophilia C or alternatively Rosenthal syndrome. The deficiency predominantly occurs in Eastern European Ashkenazi Jews, accounting for more than 50 per cent the cases. In Ashkenazi Jews the disorder is reported to occur in 5 to 11 per cent of individuals in the heterozygous state and 0.1 to 0.3 per cent in the homozygous state. Genetically, the mutations are grouped into three types: type I, abnormalities in the intron–exon splice boundaries; type II, mutations

that result in a premature stop in translation; and type III, mutations resulting from a missense mutation.

The protein itself is an 80 000 kDa protein that circulates in the plasma as a zymogen in a non-covalent association with high molecular weight kininogen. It contains four apple domains in its protein structure, and although factor XIa is a cleaving protease, its structure differs from the serine protease coagulation proteins. Factor XI is principally activated by factor XIIa in the presence of a negatively charged surface (contact activation). The lack of any bleeding diathesis related to a severe deficiency of factor XII suggests the importance of thrombin as an alternative mechanism of *in vivo* factor XI activation.

The *in vitro* factor XI activity level does not correlate well with clinical phenotype. Family history of the bleeding complications and the specific mutated sites are more predictive. Bleeding manifestations are rare in heterozygotes and occur in approximately 50 per cent of homozygous patients.

Factor XI activity levels are assayed in an activated partial thromboplastin time based test. Bleeding problems include easy bruising, epistaxis, haematuria, postpartum haemorrhage, haematomas, and menorrhagia. Haemophilia symptoms, including haemarthroses and intramuscular bleeding, are rare. Bleeding most frequently occurs after trauma or surgery. Damage to tissues rich in fibrinolytic activity, such as oral mucosa and the prostate, are more commonly associated with bleeding problems.

Therapy for patients with factor XI deficiency is indicated for symptomatic bleeding and prophylactically for surgery in patients with markedly reduced levels (i.e. below 20 per cent), unless there is no personal or family history of any bleeding complication. Fresh frozen plasma should be readily available at surgery for infusion in case of a bleeding emergency. Factor XI has a half-life of 60 to 80 h; 10 ml plasma/kg/day is usually adequate for maintaining haemostasis. Prophylactic therapy for most surgery includes replacement of factor XI with plasma at a loading dose of 15 ml/kg followed by 3 to 6 ml/kg every 24 h. The protective level for surgical prophylaxis is suggested as 45 per cent for major surgery and 30 per cent for minor surgery.

Antifibrinolytic therapy with e-aminocaproic may be a helpful adjunct to plasma therapy; however, antifibrinolytics should be avoided in patients with haematuria or bleeding in the bladder because of possible obstruction by clots.

Deficiencies of proteins in the tissue factor and common pathways

The autosomally inherited deficiencies of factors II, V, VII, and X result in bleeding diatheses of varying severity. Such deficiencies of coagulation factor correlate poorly with tests of *in vitro* factor activity; these are thus quite different disorders from haemophilia, in which *in vitro* assessment predicts the clinical phenotype very well. These factor deficiencies can best be assessed by an initial screen using the prothrombin time as a measurement of the tissue factor pathway. Although the activated partial thromboplastin time may be prolonged with deficiencies of factors II, V, and X, but not VII, the prothrombin time is most often much more sensitive.

Factors II, VII, and X, are structurally homologous containing a signal peptide, a propeptide region necessary for recognition by the post-translationally modifying enzyme g-glutamyl carboxylase, an intermolecular binding region (two epidermal growth factor (EGF) domains in factors VII, IX, and X and two kringle domains in the prothrombin molecule), and a catalytic domain in the carboxy terminal of the molecule.

Deficiency of prothrombin (factor II) results from a lack of prothrombin or a malfunctional prothrombin protein. Deficiencies result in haemorrhagic manifestations. All reported patients with a prothrombin deficiency retain some prothrombin, thus suggesting that complete prothrombin deficiency is incompatible with life. This is consistent with the knockout mouse model which results in embryonic lethality at 9.5 to 11.5 days postcoitum in over 50 per cent of fetuses; however, for some unknown reason, some murine fetuses are able to survive to birth but promptly die within 2 days due to haemorrhage. Patients with heterozygous prothrombin deficiency most commonly are either asymptomatic or have minimal bleeding. Bleeding symptomatology includes easy bruising, soft tissue haemorrhage, excessive postoperative bleeding, epistaxis, and menorrhagia in women. Haemarthroses are uncommon.

Congenital disease is characterized by a lifelong and a family bleeding history. Levels of 20 to 30 per cent prothrombin normally prevent bleeding symptomatology. When necessary, administration of plasma is recommended at doses of 15 to 20 ml/kg followed by 3 ml/kg every 12 to 24 h. Prothrombin complex concentrates can be administered for serious bleeds and as a prophylactic before surgery. Transmission of viral disease and thromboembolic phenomena are risks of the administration of prothrombin complex concentrates.

Factor V deficiency occurs in fewer than one in a million individuals. Approximately 20 per cent of the body's factor V reserve resides in the platelets. Thus, it is not surprising that patients with factor V deficiency tend to have mucosal bleeding manifestations including epistaxis, gastrointestinal bleeds, and menorrhagia in women. Haemarthroses, although a possible complaint, are much less common than in haemophilia. Mild to moderate bleeding may be treated by raising the factor V activity to about 20 per cent of normal with a plasma dose of approximately 15 to 20 ml/kg followed by 3 to 6 ml/kg every 24 h. Because of the large amount of factor V stored in platelet alpha granules, platelet transfusions may be an appropriate therapy. However, patients should be monitored for the possibility of generation of antiplatelet antibodies.

Factor VII deficiency presents as a variable bleeding disorder ranging from mild to severe, with a possibility of fatal intracranial haemorrhage. Patients with homozygous or compound heterozygous mutations manifest symptoms similar to those of a patient with haemophilia. However, unlike the correlation between activity levels and severity of disease in haemophilia, the *in vitro* factor VII activity clotting test provides only a relative indication of possible disease manifestations. Manifestations include haemarthrosis, arthropathies, haematoma formation, and retroperitoneal bleeding. Fatal intracranial haemorrhage is estimated to occur in approximately 16 per cent of patients with severe disease. Levels below 10 per cent activity most often result in bleeding manifestations. Therapy includes replacement of factor VII levels to 10 to 25 per cent for patients undergoing most types of surgery. Therapy includes plasma at 5 to 10 ml/kg for 6 to 12 h for 1 to 2 days for minor episodes. For surgery, the recommended dose is administration of 15 to 20 ml/kg followed by maintenance doses of 3 to 6 ml/kg every 12 h.

Prothrombin complex concentrates may frequently be used to supply the factor VII along with the other vitamin K-dependent proteins. Although thrombogenicity has not been a recent problem, this does remain a potential complication. Recombinant factor VIIa has been used in Europe for several years and was approved for use in the United States in 1999 for treatment of haemophilia with inhibitors. Although the product has not yet been officially approved for use in factor VII-deficient patients in the United States, recombinant factor VIIa is therapeutically effective in this setting at a dose of 25 µg/kg for acute bleeds, a significantly lower dose than that used for treatment of haemophiliac patients with inhibitors. However, the possible development of a factor VII inhibitor must be considered, as this has been reported. The product is administered every 2 h for prophylaxis during surgery for the first 24 h, then reduced to every 3 h 24 to 48 h postoperatively, and then further reduced ccording to patient symptomatology and necessity, depending on the risk of bleeding into the surgical site.

Factor X deficiency may present with symptomatology similar to that of a patient with severe haemophilia. Haemarthroses, soft tissue haemorrhages, retroperitoneal bleed, central nervous system haemorrhages, pseudotumours, and menorrhagia may occur. Therapy with fresh frozen plasma includes a loading dose of 10 to 15 ml/kg followed by approximately 50 per cent of that at 24 h.

Deficiency of the contact activating factors, factor XIII, and fibrinogen

Although the activated partial thromboplastin time is grossly prolonged (often more than150 s) with deficiencies of the contact activating factors—factor XII, high molecular weight kininogen, and prekallikrein—these deficiencies are not associated with bleeding manifestations and will not be covered further here.

Factor XIII deficiency often presents shortly after birth with bleeding of the umbilical cord. Patients with clinical manifestations typically have factor levels of less than 1 per cent. Factor XIII is a transglutaminase that crosslinks fibrin monomers, thus stabilizing a forming fibrin clot. Patients with deficiency of factor XIII therefore have delayed wound healing and often suffer from soft tissue haemorrhages, haemarthroses, haematomas, and excessive bleeding from poorly healed wounds. Up to 25 per cent of individuals deficient in factor XIII may experience intracranial bleeding. For unknown reasons, affected males may have oligospermia and affected women may suffer from repeated spontaneous abortions. Since routine clotting tests are normal in factor XIII deficiency, a physician must specifically request a test for factor XIII deficiency which entails a clot solubility test using 2 per cent chloroacetic acid on a formed clot. Treatment of factor XIII deficiency involves administration of small amounts of factor XIII required to minimize bleeding complications. Prophylaxis includes using 2 to 3 ml/kg of fresh frozen plasma every 4 to 6 weeks or one bag of cryoprecipitate per 10 to 20 kg every 3 to 4 weeks. To prevent spontaneous abortions, products containing factor XIII can be administered every 14 to 21 days.

Afibrinogenaemia may cause dangerous haemorrhagic episodes. However, it is somewhat surprising that the mutation does not lead to embryonic death in light of the fact that the blood is incoagulable *in vitro*. The lack of necessity for fibrinogen during fetal development is supported by the viable fibrinogen knockout mouse model. Prolonged bleeding from the umbilical cord often permits early recognition of an affected child. The leading cause of death in afbrinogaemia is intracranial

haemorrhage. Haemorrhages from mucous membranes occur frequently, and haemarthroses occur in approximately 20 per cent of patients. Pregnancy related problems include first trimester abortion, placental abruption, and postpartum bleeding complications and may be markedly reduced by administration of fibrinogen. However, fibrinogen replacement may cause thromboembolic phenomena. The target fibrinogen level for replacement therapy is approximately 50 to 100 mg/dl. One bag of cryoprecipitate contains approximately 250 mg of fibrinogen; thus dosing of cryoprecipitate usually necessitates 5 to 10 bags per 70 kg person. Therapeutic complications include allergic reactions and the development of antifibrinogen antibodies. Thromboembolic phenomena may occur in conjunction with fibrinolytic inhibitors or oral contraceptives.

Dysfibrinogaemia results from a functional deficiency of fibrinogen associated with a malfunctional molecule, although some degree of antigen remains present. Approximately 55 per cent of patients with dysfibrinogaemia remain asymptomatic, 25 per cent have a bleeding tendency, and 20 per cent may experience thrombotic episodes ranging from mild to fatal events.

Numerous combined deficiencies have been described; the underlying mutation for several of these combined deficiencies has been determined. Combined deficiency of the two structurally similar proteins factor V and factor VIII is an autosomal recessively inherited disorder of variable bleeding severity. The mechanism responsible for the disorder results from a mutation in ERGIC-53, a 53 kDa transmembrane component of the endoplasmic reticulum–Golgi intermediate compartment. Mutations at this site are associated with factor levels of 4 to 30 per cent of normal factor V and factor VIII activity and generally show mucocutaneous and postsurgical bleeding of a severity similar to that seen in individuals with a single protein deficiency at the same level. Other combined deficiencies for which a genetic mechanism has been described include deficiency of factors II, VII, IX, and X caused by a mutation in the g-glutamyl carboxylase gene, required for a critical post-translational modification in vitamin K-dependent factors.

Hypercoagulable disease due to deficiencies of anticoagulant

Pathological diseases resulting from inappropriate clot formation in either the arterial or venous circulation are a major cause of morbidity in the Western world. The genetic contribution to this pathophysiology, particularly to thrombosis in the arterial circulation, is not well understood. Clearly cardiovascular disease represents a complex multifactorial process. The contribution of genetic causes to venous thrombotic disease is better understood; it may be associated with either an isolated deficiency of an anticoagulant protein, a malfunctional procoagulant protein, or a combination of these processes. The functional deficiencies become particularly relevant during times of increased environmental stress such as in the puerperium or in postsurgical, traumatic, or immobilized states. In addition to deficiency states, several common mutations involving a gain of function have also been described which can disrupt the delicate balance of coagulation by shifting the balance toward greater procoagulant function.

Procoagulant and anticoagulant plasma proteins interact with platelets and cellular phospholipids to promote physiological coagulation. Regulation of the formation of thrombin is the key step in the proper balance between pro- and anticoagulant functions. Anticoagulant proteins are particularly important in areas where there may be prolonged exposure of procoagulant factors and platelet phospholipids to the vessel wall, predisposing an individual to thrombotic disease. Deficiencies of anticoagulant proteins thus place a patient at an increased risk for thrombosis in the slowly flowing venous circulation. In the rapidly flowing arterial circulation, laminar flow largely prevents prolonged interaction between platelets and vessel walls.

The principal anticoagulant proteins that keep the procoagulant proteins in check include thrombomodulin, tissue factor pathway inhibitor, antithrombin III, protein C, and protein S. Thrombomodulin, an integral membrane protein expressed by endothelial cells, plays a key role in tempering the action of thrombin. Despite attempts to discover mutations in the thrombomodulin gene, only rare reports have implicated thrombomodulin in the pathophysiology of disease, although some recent studies suggest the existence of polymorphic regulation variants in the promoter region. Recently, a mutation in the small but critical protein known as tissue factor pathway inhibitor, which inhibits procoagulant function by binding to factor Xa either alone or in association with tissue factor–factor VIIa, has been suggested to be associated with a ninefold increased risk of venous thrombosis.

Deficiencies leading to a hypercoagulable state are most frequently caused by deficiencies of antithrombin III, protein C, and protein S. These anticoagulant deficiencies result from either a quantitative deficiency (type I) or a qualitative deficiency (type II). Deficiencies of any of these factors may cause life-threatening deep venous thromboses and pulmonary emboli, or may be asymptomatic. Clinical presentation relates to physical sequelae in the affected organ. In addition to deep venous thromboses and pulmonary emboli, symptomatology may include superficial thrombophlebitis, mesenteric vein thrombosis, and cerebral vein thrombosis.

Antithrombin III deficiency

A deficiency of antithrombin III was the first anticoagulant protein deficiency described which was associated with an increased risk of thrombosis. Antithrombin III is a 60 kDa glycoprotein found at high concentrations in the plasma—150 µg/ml: approximately 15- to 30-fold higher than that of many other pro- and anticoagulant proteins. Antithrombin III primarily inhibits thrombin but also inhibits factors IXa, Xa, XIa, XIIa, kallikrein, and plasmin. The ability to inhibit thrombin requires interaction with heparin, which increases the inhibitory activity several thousandfold. Historically, the risk of thrombosis in individuals deficient in antithrombin III has been thought to be higher than that seen with deficiencies of protein S or protein C, or than that seen with increased functionality of the procoagulant proteins factor V and prothrombin. Clearly the influences of gene–gene and gene–environment interactions contribute to this risk. A normal activity range for most procoagulant/anticoagulant proteins may be as low as 50 per cent. However, the critical requirement for antithrombin III can be surmised from the 80 per cent lower limit of a normal antithrombin III level, significantly higher than that for other coagulation proteins. This makes the diagnosis of antithrombin III deficiency particularly difficult in the post-thrombotic period when patients frequently have lower levels of antithrombin III due either to consumption of antithrombin III during clot formation or to the decreased function seen with heparin administration. Additionally, the presence of homozygous disease of antithrombin III deficiency has only been reported with rare type II deficiencies resulting from impaired heparin binding mutations. No homozygous type I deficiencies have been reported, probably due to their incompatibility with life.

The frequency of antithrombin III deficiency in patients with thrombophilia varies widely between studies. The cause of these widely differing frequencies has recently been carefully addressed by van Boven and colleagues. Their study clearly shows the strong influence of acquired and genetic factors which modulate the baseline risk due to one specific genetic mutation, highlighting the role of additional factors when combined with genetics. In thrombophilic family studies, the risk of thrombosis is 20 times greater than in control populations. The most frequent presentation is deep venous thrombosis with a pulmonary embolism, particularly after an inciting environmental influence such as surgery or immobilization in men or the start of oral contraceptives or pregnancy/postpartum in women. The average age of first onset is 33 years. In patients deficient in antithrombin III without a known acquired risk, the rate of incidence of thrombosis was less than 1 per cent per year.

Therapy for antithrombin III deficiency includes prophylactic treatment with warfarin, low molecular weight heparin, and treatment of an acute event with heparin or another anticoagulant therapy, for example administration of a fibrinolytic agent in the patient presenting early enough during an acute episode. Antithrombin III concentrate may be administered for therapy of deficiency during an acute event or as a prophylactic treatment to prevent further disease.

Deficiencies of proteins C and S

Deficiencies of proteins C and S present with thrombotic manifestations similar to those seen with antithrombin III deficiency. However, in protein C deficiency an additional condition includes warfarin-induced skin necrosis and dangerously life-threatening purpura fulminans in the homozygous or compound heterozygous protein C deficient neonate. A diagnosis of protein C deficiency is found in approximately 33 per cent of individuals with warfarin-induced skin necrosis, a condition which leads to skin necrosis several days after initiation of warfarin therapy. The proposed mechanism for this condition is due to the earlier decrease in protein C compared with decreases in procoagulant proteins following initiation of warfarin therapy (due to the short half-life of protein C, approximately 6 h). It is thus 'normal' clinical practice to begin warfarin only after a patient has first been anticoagulated with heparin or another immediately acting anticoagulant therapy.

Protein C acts in concert with its cofactor protein S to inactivate the active forms of the procoagulant cofactors, factors Va and VIIIa. Protein C is a vitamin K-dependent serine protease structurally similar to factors VII, IX, and X. Protein S is also vitamin K-dependent because of conserved NH2 terminus but lacks enzymatic function because of the existence of a sex-hormone binding globulin domain instead of a catalytic domain at the COOH terminus. Thrombin activates protein C to activated protein C when bound to thrombomodulin, a protein which acts like an endothelial cell receptor for thrombin. Symptomatic manifestations of protein C or protein S deficiencies are similar to that of antithrombin III deficiency. Deep venous thombosis with or without pulmonary embolism occurs in 50 per cent of patients by the age of 30 to 45, depending on the study population. Environmental and gene–gene interactions are particularly important. As with antithrombin III deficiency, superficial thrombophlebitis, cerebral vein thrombosis, and mesenteric vein thrombosis are all possible complications. Postphlebitic syndrome presents as a complication after deep venous thombosis in up to 50 per cent of patients.

Factor V Leiden and the prothromibn 20210 mutation

Since 1994, two additional common mutations have been described leading to an increased risk of thrombosis. These mutations, unlike the anticoagulant protein deficiencies, are due to gain of function mutations causing either an increased resistance to inactivation in factor V (factor V Leiden) or increased levels of a procoagulant protein (prothrombin) which results in higher levels of thrombin formation.

Activated protein C (APC) resistance was first described by Dahlback in a 42-year-old man with a history of recurrent thromboses. Dahlback noted an absence of prolongation of the activated partial thromboplastin time, found after addition of APC, which is normally prolonged due to inactivation of factors Va and VIIIa. Soon thereafter, Poort and colleagues identified a single mutation as the principal cause of APC resistance in the vast majority (over 90 per cent) of patients. The mutation leads to a decreased ability of APC to inactivate the cofactor Va due to an amino acid substitution (arginine for glutamine) at a critical hydrolysis point in the factor Va protein normally enabling inactivation. Other non-factor V Leiden causes of APC resistance include a haplotype in the factor V molecule, the H2 haplotype.

Factor V Leiden leads to thrombotic disease as described for hypercoagulable states due to deficiencies of anticoagulant protein. Because of the extremely high incidence of factor V Leiden in the Caucasian population (approximately 5 per cent), gene–gene interactions play a particularly important role in manifestation of disease. It should be noted that the frequency of factor V Leiden in most non-Caucasian populations is low.

The prothrombin 20210 mutation reported in 1996 results in an increased concentration of prothrombin, also tipping the balance towards excess thrombin formation. The cause of this increase is associated with a guanine to adenine mutation at the last base of the 3' untranslated region in the factor V gene. The mechanism by which this influences prothrombin levels is not understood.

---

*Because factor IX is not confined to the intravascular space, it is customary to double the desired intravascular dosage compared with factor VIII administration to account for this larger volume of distribution in part due to binding of factor IX to endothelial cells.
†When using the only available recombinant protein (BeneFix), it has empirically been shown that it is necessary to multiply this dosage by a factor of 1.2 IU/kg to achieve the projected desired level.

Further reading

General articles about coagulation

Colman RW *et al*. (1994). Overview of hemostasis. In: Colman RW *et al*., eds. *Thrombosis and hemorrhage*, pp 3–18. JB Lippincott, Philadelphia.

Davie EW and Ratnoff OD (1964). Waterfall sequence for intrinsic blood clotting. *Science* **145**, 1310–12.

Furie B, Furie BC (1992). Molecular and cellular biology of blood coagulation. *New England Journal of Medicine* **26**, 800–6.

MacFarlane RG (1964). An enzyme cascade in the blood clotting mechanism and its function as a biochemical amplifier. *Nature* **202**, 498–9.

Roberts HR, Lozier JN (1992). New perspectives on the coagulation cascade. *Hospital Practice* **27**, 97–105, 109–12.

Haemophilia and von Willebrand disease

Djulbegovic B, Goldsmith GH Jr (1995). Guidelines for management of hemophilia A and B. *Blood* **85**, 598–9.

Furie B, Furie BC (1990). Molecular basis of hemophilia. *Seminars in Hematology* **27**, 270–85.

Furie B, Limentani SA, Rosenfield CG (1994). A practical guide to the evaluation and treatment of hemophilia. *Blood* **84**, 3–9.

Ginsburg D, Bowie EJ (1992). Molecular genetics of von Willebrand disease. *Blood* **79**, 2507–19.

Gitscher J *et al*. (1991). Genetic basis of hemophilia A. *Thrombosis and Haemostasis* **66**, 37–9.

Larson PJ, High K (1992). Biology of inherited coagulopathies: factor IX. *Hematology—Oncology Clinics of North America* **6**, 999.

Ljung R *et al*. (1992). Factor VIII and factor IX inhibitors in haemophiliacs. *The Lancet* **339**, 1550.

Management Association of Hemophilia Clinic Directors of Canada (1995). Hemophilia and von Willebrand's disease. *Canadian Medical Association Journal* **153**,147.

Roberts HR (1993). Molecular biology of hemophilia B. *Thrombosis and Haemostasis* **70**, 1.

Sadler JE (1994). A revised classification of von Willebrand disease. For the Subcommittee on von Willebrand Factor of the Scientific and Standardizaion Committee of the International Society of Thrombosis and Haemostasis. *Thrombosis and Haemostasis* **71**, 520–5.

Triemstra M *et al*. (1995). Mortality in patients with hemophilia. Changes in a Dutch population from 1986 to 1992 and 1973 to 1986. *Annals of Internal Medicine* **123**, 823.

Administration of factor concentrates

Aronson DL (1987). Thrombogenicity of factor IX complex: *in vivo* investigation. *Developments in Biological Standardization* **67**, 149.

Blanchette VS *et al*. (1997). Central venous access devices in children with hemophilia: an update. *Blood Coagulation and Fibrinolysis* **8** (suppl. 1), S11.

Ewenstein BM (1997). Nephrotic syndrome as a complication of immune tolerance in hemophilia B. *Blood* **89**, 1115–16.

Federici AB (1998). Optimizing therapy with factor VIII/von Willebrand factor concentrates in von Willebrand disease. *Haemophilia* **4**, 7.

Goudemand JNC, Ounnoughene N, Sultan Y (1998). Clinical management of patients with von Willebrand's disease with a VHP vWF concentrate: the French experience. *Haemophilia* **4**, 48.

Mannucci PM (1993). Clinical evaluation of viral safety of coagulation factor VIII and IX concentrates. *Vox Sanguinis* **64**, 197.

Warrier I *et al*. (1997). Factor IX inhibitors and anaphylaxis in hemophilia B. *Journal of Pediatric Hematology/Oncology* **19**, 23.

White GC 2nd and Nielsen B (1997). Recombinant factor IX. *Thrombosis and Haemostasis* **78**, 261–5.

Infectious diseases associated with haemophilia therapy

Bruce ME *et al*. (1997). Transmissions to mice indicate that 'new variant' CJD is caused by the BSE agent. *Nature* **389**, 498.

Baxter T, Black D, Birks D (1998). New-variant Creutzfeldt–Jakob disease and treatment of haemophilia. *The Lancet* **351**, 600.

Centers for Disease Control and Prevention (1996). Hepatitis A among persons with hemophilia who received clotting factor concentrate–United States, September–December 1995. *Journal of the American Medical Association* **275**, 427.

Craven BM, Stewart GT, Khan M (1997). AIDS: safety, regulation and the law in procedures using blood and blood products. *Medicine, Science and the Law* **37**, 215.

Eyster ME *et al*. (1993). Natural history of hepatitis C virus infection in multitransfused hemophiliacs: effect of coinfection with human immunodeficiency virus. The Multicenter Hemophilia Cohort Study. *Journal of Acquired Immune Deficiency Syndromes* **6**, 602–10.

Kupfer B *et al*. (1995). Beta-propiolactone UV inactivated clotting factor concentrate is the source of HIV-infection of 8 hemophilia B patients: confirmed. *Thrombosis and Haemostasis* **74**, 1386.

Lee CA, Sabin CA (1997). The natural history of chronic hepatitis C in haemophiliacs. *British Journal of Haematology* **96**, 875.

Makris M *et al*. (1996). The natural history of chronic hepatitis C in haemophiliacs. *British Journal of Haematology* **94**, 746.

Gene therapy in haemophilia

Herzog RW *et al*. (1999) Long-term correction of canine hemophilia B by gene transfer of blood coagulation factor IX mediated by adeno-associated viral vector *Nature Medicine* **5**, 56–63.

Herzog RW, High KA (1999). Adeno-associated virus-mediated gene transfer of factor IX for treatment of hemophilia B by gene therapy. *Thrombosis and Haemostasis* **82**, 540–6.

Mannucci PM, Tuddenbam EG (1999). The hemophilias: progress and problems. *Seminars in Hematology* **36** (suppl. 7), 104–17.

Park F, Ohashi K, Kay MA (2000). Therapeutic levels of human factor VIII and IX using HIV-1-based lentiviral vectors in mouse liver. *Blood* **96**, 1173–6.

Thompson AR (2000). Gene therapies for the hemophilias. *Molecular Therapy: the Journal of the American Society of Gene Therapy* **2**, 5–8.

White GC 2nd, Roberts HR (2000). Gene therapy for hemophilia: a step closer to reality. *Molecular Therapy: the Journal of the American Society of Gene Therapy* **1**, 207–8.

Thrombotic disease

Bates SM, Hirsch J (1999). Thrombotic disorders and their treatment: treatment of venous thromboembolism. *Thrombosis and Haemostasis: State of the Art* **82**, 870–931.

Bucciarelli P, Rosendaal FR, Tripodi A (1999). Risk of venous thromboembolism and clinical manifestations in carriers of antithrombin, protein C, protein S deficiency, or activated protein C resistance: a multicenter collaborative family study. *Arteriosclerosis, Thrombosis, and Vascular Biology* **19**, 1026–33.

Koster T, Rosendaal FR, Briet E (1995). Protein C deficiency in a controlled series of unselected outpatients: an infrequent but clear risk factor for venous thrombosis (Leiden thromobophilia study). *Blood* **10**, 2756–61.

Lane DA, Mannucci PM, Bauer KA (1996). Inherited thrombophilia: part 1. *Thrombosis and Haemostasis* **76**, 651–62.

Lane DA, Manucci PM, Bauer KA (1996). Inherited thrombophilia: part 2. *Thrombosis and Haemostasis* **76**, 824–34.

Rivard GE, David M, Farrell C (1995). Treatment of purpura fulminans in meningococcemia with protein C concentrate. *Journal of Pediatrics* **126**, 646–52.

Rohrer MJ, Andrew M, Michelson AD (1998). Hemorrhage, thrombosis, and antithrombotic therapy in children. In: Loscalzo J, Shafer A, eds. *Thrombosis and Hemorrhage*, pp. 1027–63. Williams and Wilkins, Baltimore.

Sanz-Rodriguez C, Gil-Fernandez JJ, Zapater P (1999). Long-term management of homozygous protein C deficiency: replacement therapy with subcutaneous purified protein C concentrate. *Thrombosis and Haemostasis* **81**, 887–90.

van Boven HH *et al*. (1999). Gene–gene and gene–environment interactions determine risk of thrombosis in families with inherited antithrombin deficiency. *Blood* **94**, 2590–4.

22.6.5 Acquired coagulation disorders

T. E. Warkentin

A 'coagulopathy' is a disorder associated with an abnormal coagulation assay result, such as a prolonged prothrombin time (**PT**) (often expressed as the international normalized ratio, or **INR**), activated partial thromboplastin time (**aPTT**), or thrombin clotting time (**TCT**). Coagulopathies can be associated with either bleeding or thrombosis, and have many causes (Table 1). The importance of the clinical context is illustrated by two patient scenarios that have in common a prolonged INR (6.0; usual therapeutic range, 2.0–3.0) during oral anticoagulant therapy: patient A has a life-threatening intracranial haemorrhage complicating warfarin therapy given for a prosthetic heart valve; in contrast, patient B, who was treated for deep-vein thrombosis (**DVT**) complicating heparin-induced thrombocytopenia (**HIT**), has the limb-threatening complication of warfarin-induced venous limb gangrene, caused by microvascular thrombosis.

Table 2 lists common screening tests for coagulopathy. Only a few bleeding disorders give normal results in all these tests (for example, a $_2$-antiplasmin deficiency, factor XIII deficiency, type 2a von Willebrand syndrome associated with aortic stenosis). A drawback of these assays is that only procoagulant *haemo*static pathways are assessed. Thus, deficiency of a natural anticoagulant such as antithrombin or protein C must be determined by specific testing.

Treatment approaches

Blood products are usually indicated for the treatment of patients with coagulopathies who are bleeding or who require a major invasive procedure. *Fresh-frozen plasma* (**FFP**), which contains all the *haemo*static factors at concentrations between 0.7 and 1.0 U/ml, is appropriate for liver disease, *haemo*dilution from massive transfusion, disseminated intravascular coagulation (**DIC**), reversal of coumarin anticoagulation, and replacement of isolated factor deficiency when specific-factor replacement is unavailable. For a 70-kg adult with a 3-litre plasma volume, 1 litre of FFP will increase the coagulation factors by about 0.25 U/ml. In most patients, this should lead to levels greater than the minimum required for adequate *haemo*stasis (>0.30 U/ml for most factors). Repeat FFP transfusion (for example, 500 ml every 6 h) is necessary if the *haemo*stasis defect is ongoing. FFP is being supplanted by cryosupernatant as a replacement fluid for thrombotic thrombocytopenic purpura (**TTP**). Solvent-detergent-treated plasma (**SD**-plasma), in which most blood-borne pathogens are inactivated (but not hepatitis A, parvovirus B19, or the agent that causes Creutzfeldt–Jakob disease, a theoretical bloodborne pathogen), has become available recently, but is limited by its high cost.

Cryoprecipitate contains fibrinogen (0.10–0.25 g per unit), factors VIII and XIII, von Willebrand factor (**vWF**), and fibronectin. Its major indication is the treatment of hypofibrinogenaemia, where it increases fibrinogen levels using just one-quarter of the volume of blood product compared with FFP. Cryoprecipitate is appropriate for patients with significant hypofibrinogenaemia, for example DIC, primary fibrinolysis, congenital hypofibrinogenaemia. For a bleeding patient whose fibrinogen level is about 0.5 g/l, 10 to 20 U of cryoprecipitate would probably increase the fibrinogen to above 1.0 g/l, although a lower than expected increment could occur if the patient had a higher volume of distribution (for example, a cirrhotic patient with ascites).

Specific-factor concentrates are available for use in patients with an isolated deficiency in certain factors, such as VIII or IX. Additionally, some factor IX concentrates, known as prothrombin complex concentrates (**PCC**), contain all four vitamin K-dependent procoagulant factors, and are appropriate for the rapid reversal of severe coagulopathy related to coumarin use. Activated PCC (for example, **FEIBA** (factor VIII inhibitor bypassing activity)) and factor VIIa are other specialized blood products with specific uses, for instance to manage a bleeding patient with an acquired factor VIII inhibitor.

Pharmacological therapies include the antifibrinolytic agents e-amino-caproic acid (**EACA**), tranexamic acid (**TA**), and aprotinin. EACA and TA bind to the lysine-binding sites of plasminogen; paradoxically, although increasing the susceptibility of plasminogen to proteolysis by plasminogen activator, these lysine analogues also prevent plasminogen from binding to fibrin, thus impeding fibrinolysis. Oral dosing for EACA is about 7 g (100 mg/kg) initially, followed by 3.5 g (50 mg/kg) every 4 h; similar doses are used for intravenous administration. For tranexamic acid, 1.0 to 1.5 g is given every 8 h by mouth; the dose is reduced to between 0.5 and 1.0 g every 8 h if given intravenously. Both drugs are available in 500-mg capsules. These drugs are appropriate for the treatment of hyperfibrinolysis, for instance bleeding following thrombolytic therapy or cardiac surgery. These drugs are generally contraindicated in patients with DIC, however, as blocking secondary fibrinolysis could lead to microvascular thrombosis. Aprotinin is discussed below in the section on cardiopulmonary bypass surgery.

Desmopressin, or 1-desamino-8-D-arginine vasopressin (**DDAVP**), a synthetic vasopressin analogue, leads to an increase in factor VIII and vWf levels that peak between 45 and 90 min after intravenous infusion (0.3 µg/kg in 50 ml normal saline over 20–30 min; maximum dose, 20 µg). Although repeat DDAVP can be given at 12- to 24-h intervals, the drug becomes less effective over time (tachyphylaxis) as endothelial stores of vWf are depleted. Blood pressure elevation, free-water retention leading to hyponatraemia, flushing, and angina are occasional side-effects.

Prohaemorrhagic acquired coagulation disorders

Vitamin K deficiency disorders

Vitamin K-dependent coagulation factors

Vitamin K is required for the post-translational modification of six *haemo*static factors, four with procoagulant activity (factors II, VII, IX, and X), and two with anticoagulant activity (protein C, protein S). The physiological relevance of a seventh factor, Factor Z, remains unclear. The enzyme, vitamin K-dependent g-glutamylcarboxylase, adds a carboxyl group to each member of a cluster of glutamyl residues, thereby forming the g-carboxyglutamyl residues crucial for enabling these six haemostatic factors to interact with phospholipid membranes in a calcium-dependent fashion. During this g-carboxylation reaction, the reduced form of vitamin K (vitamin KH_2) is oxidized to vitamin K epoxide; oral anticoagulants inhibit the two enzymes (vitamin K epoxide reductase and vitamin K reductase, respectively) that act in sequence to regenerate the reduced form of vitamin K.

Diet and absorption of vitamin K

Vitamin K_1 (phylloquinone) is exclusively derived from plants; vitamin K_2 (menaquinone) is synthesized by bacteria. Green, leafy vegetables, such as broccoli, lettuce, cabbage, and spinach, are very good dietary sources of vitamin K (100–500 µg/100 mg). Vitamin K is fat-soluble, and absorption occurs primarily in the small bowel. Serum vitamin K levels are only between 150 and 800 pg/ml and, as hepatic storage is limited (*t*½, just a few days), a regular daily intake of about 0.1 to 0.5 µg/kg is

required. Although bacterial synthesis is not a major source of vitamin K to humans, antibiotic treatment nevertheless predisposes to vitamin K deficiency.

Vitamin K deficiency

Malabsorption of fat-soluble vitamins caused by biliary tract disease, or primary bowel disorders such as coeliac or inflammatory bowel disease, can cause vitamin K deficiency. An inadequate diet, particularly when combined with antibiotic therapy, is another cause. Indeed, coagulopathy can arise during a brief period of decreased intake, for example 1-week postoperatively.

A disproportionately prolonged PT/INR in the appropriate clinical setting suggests vitamin K deficiency (Table 3). The diagnosis is usually confirmed by assessing the response to vitamin K administration. Compared with the treatment of a coumarin overdose, small amounts of vitamin K are effective: for example, 1 mg vitamin K_1 given subcutaneously or by slow intravenous infusion (over at least 30 min to minimize risk of an anaphylactoid reaction). For serious bleeding, FFP or especially PCC provides a more rapid correction of the coagulopathy.

Coumarin overanticoagulation

Oral anticoagulants (for example, coumarins such as warfarin and phenprocoumon) are widely used to prevent and treat thrombosis via their vitamin K antagonism. An INR target range between 2.0 and 3.0 is appropriate for most clinical indications, although a higher therapeutic range (INR, 2.5–3.5) is appropriate for patients at very high risk for thrombosis (for instance, mechanical prosthetic heart valves; thrombosis complicating the antiphospholipid antibody syndrome).

Bleeding is the major complication of coumarin, with minor and major bleeding episodes occurring in about 6 to 10 per cent and 1 to 3 per cent of patients/year respectively; the intracranial haemorrhage rate is between 0.25 and 1 per cent/year. Changes in diet or alcohol consumption, poor patient compliance, and the introduction of new drugs (Table 4) can cause bleeding by producing coumarin overanticoagulation. In contrast, recurrent gastrointestinal or urinary tract bleeding at therapeutic levels of anticoagulation often indicates an occult gastrointestinal or renal lesion, respectively.

The treatment of non-therapeutic (elevated) INRs depends on the clinical situation (Table 5). Oral vitamin K_1 use is appropriate in many non-urgent situations as it avoids the risk of anaphylactoid reactions to intravenous use, and has more predictable effects than subcutaneous injection. Much larger and prolonged vitamin K dosing (100–150 mg/day) is required to treat accidental or deliberate overdoses of long-acting, second-generation rodenticides ('superwarfarins'), such as brodifacoum.

Liver disease

Most haemostatic factors are produced exclusively by the liver. Exceptions include factor VIII (hepatic and extrahepatic synthesis), vWf (endothelium, megakaryocytes), and several factors produced by endothelium (for example, plasminogen activator, plasminogen activator inhibitor type I (**PAI-1**)). Table 6 lists the multiple effects on haemostasis caused by liver disease. Often, bleeding is primarily related to anatomical factors, such as oesophageal varices or gastric/duodenal ulcers, though reduced hepatic coagulation-*factor* synthesis can be a contributing factor. Increased susceptibility to DIC via superadded illness (for example, bacterial peritonitis), impaired clearance of activated coagulation factors, and hyperfibrinolysis are other factors.

A prolonged PT/INR is the most frequent laboratory abnormality (Table 3). The fibrinogen level is usually normal or increased; when hypofibrinogenaemia occurs, it generally indicates severe liver disease or hyperfibrinolysis. Fibrin(ogen)-degradation product (**FDP**) and D-dimer levels are often increased; thus, the laboratory picture can resemble that of disseminated intravascular coagulation even in a patient who is otherwise clinically stable.

Management of hepatic coagulopathy should include a trial of vitamin K_1 (for instance, 10 mg subcutaneously for 3 days), although this will not benefit most patients. Fresh-frozen plasma should be given to bleeding patients with a prolonged INR, or who require major invasive procedures. Retrospective studies suggest that minor invasive procedures (for example, paracentesis, pleurocentesis) can usually be performed safely with an INR as high as 1.8. For patients suspected to have significant fibrinolysis, antifibrinolytic therapy can be tried. PCCs should only be used in emergency situations, given their prothrombotic potential in this patient population. Platelet transfusions usually provide minimal increase in the platelet count in patients with platelet sequestration caused by hypersplenism. DDAVP improves haemostasis in patients with prolonged bleeding time secondary to hepatic platelet dysfunction.

Haemodilution and massive transfusion

Coagulopathies occur in most patients who receive crystalloids, colloids, or red cell concentrates (**RCCs**) following trauma, surgery, or fluid resuscitation for other major illnesses. In many patients, no bleeding results despite moderate abnormalities in the INR, aPTT, TCT, and platelet count. The reason is that all the individual coagulation factors remain at haemostatically effective levels, even though the laboratory assays are abnormal when all the factor levels are uniformly reduced.

Massive transfusion is defined as the transfusion of blood products equivalent to the patient's total blood volume within 24 h. RCCs do not provide significant amounts of platelets or coagulation factors. Thus, platelet, FFP, and, sometimes, cryoprecipitate infusions are often needed as well. Although 'formulas' to guide transfusion therapy have been devised, individualized assessment that takes into account clinically evident bleeding, risk factors for haemorrhage, supervening DIC or fibrinolysis, acute liver insult, and laboratory test results, is preferable.

Disseminated intravascular coagulation

Disseminated intravascular coagulation, or DIC, is a group of clinicopathological syndromes characterized by widespread activation of coagulation; there results intravascular generation of thrombin, formation of fibrin, and reactive fibrinolysis. Clinical consequences range from coagulation factor and platelet depletion, resulting in generalized haemorrhage, to widespread microvascular thrombosis, predisposing to multisystem organ dysfunction or limb necrosis. 'Acute' DIC, caused by septicaemia, trauma, and obstetrical complications, is most frequent; 'chronic' DIC, typically caused by malignancy, is often associated with a dramatic hypercoagulable state (Table 7). Although DIC is usually a systemic process, sometimes a localized abnormality (such as a vascular malformation or aortic aneurysm) leads to the regional activation of coagulation and resulting in the depletion of haemostatic factors.

DIC is usually triggered by the extrinsic coagulation pathway: tissue factor and factor VIIa (Fig. 1). The proinflammatory cytokine, interleukin-6 (**IL-6**), is a principal mediator of DIC in septicaemia and other systemic inflammatory responses, and impairs natural anticoagulant and fibrinolytic pathways. For example, a sustained increase in PAI-1 impairs plasmin formation despite intravascular fibrin generation.

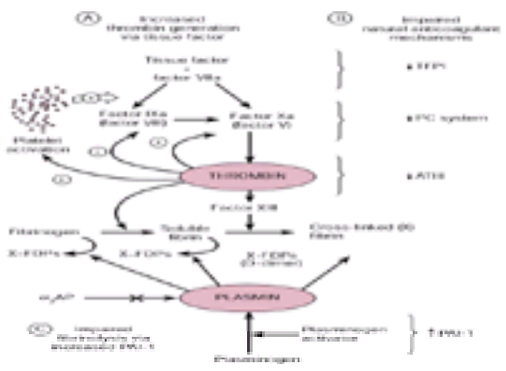

Fig. 1 Pathogenesis of thrombosis in DIC. (A) DIC is usually triggered by tissue factor, which activates coagulation by complexing with factor VIIa, ultimately resulting in the generation of thrombin. (B) Impaired natural anticoagulant mechanisms (e.g. excessive consumption of natural anticoagulants, or cytokine-mediated downregulation of natural anticoagulant pathways) predispose to microvascular thrombosis. (C) Impaired fibrinolysis via increased PAI-1 leads to greater microvascular thrombosis. Sometimes, hyperfibrinolysis is caused by increased plasminogen activator release, or low levels of b $_2$-antiplasmin. Abbreviations: a$_2$AP, a$_2$-antiplasmin; ATIII, antithrombin III; fDPs, fibrinogen degradation products; FDPs, fibrin degradation products; PAI-1, plasminogen activator inhibitor type 1; PC,

protein C; TFPI, tissue-*factor* pathway inhibitor.

Diagnostic and treatment approach to DIC

One or more prolonged clotting times and thrombocytopenia in a patient with one of the disorders listed in Table 7 suggests DIC. However, similar test results are seen in patients following major surgery, emphasizing the need to interpret the laboratory data in the appropriate clinical context. Typically, crosslinked fibrin-degradation products (D-dimers) are significantly increased in DIC. The protamine sulphate 'paracoagulation' test, based on visual detection of gelling of patient plasma after the addition of protamine sulphate, is a rapid test for fibrin monomers; the lower sensitivity of the test is helpful, as a positive result usually indicates clinically significant DIC associated with bleeding or thrombosis. Sometimes, specialized haemostasis assays are useful, for example protein C activity levels in purpura fulminans.

The cornerstone of DIC management is treating its underlying cause and providing supportive measures. For bleeding patients, replacement of depleted haemostatic factors with FFP, cryoprecipitate, and platelet transfusions may be needed. Recently, drotrecogin alfa (recombinant activated protein C) became available in some jurisdictions to treat severe septicaemia. Heparin can benefit patients with large-vessel thrombosis or acral ischaemia. The routine use of vitamin K and folate will avoid coagulation and platelet count disturbances in some patients.

Trauma and shock

Tissue injury due to trauma, burns, or hypoperfusion can cause DIC, especially when organs rich in tissue thromboplastin (for example, the brain) are injured.

Infection

Gram-negative and Gram-positive bacteria can cause DIC, either from procoagulant bacterial components (for instance, endotoxin, *Staphylococcus aureus*-toxin) or via the host response to infection (for example, IL-6). The clinical spectrum ranges from prominent thrombocytopenia with minimal activation of coagulation, to marked coagulation factor and natural anticoagulant depletion. Certain infections, such as meningococcaemia, *Capnocytophaga canimorsus* (from dog bites), sometimes produce severe acquired consumptive protein C deficiency, which leads to widespread ischaemic necrosis of the extremities (purpura fulminans). However, postvaricella purpura fulminans can be caused by acquired antiphospholipid antibodies that interfere with protein S.

Obstetrical complications

Acute DIC can be caused by thromboplastin-like materials released during placental abruption or amniotic fluid embolism. Pre-eclampsia too can be accompanied by DIC, although there can be clinical and laboratory overlap with other life-threatening complications of pregnancy (for example, fatty liver of pregnancy; HELLP syndrome (haemolysis, elevated liver enzymes, low platelets). Bleeding due to hypofibrinogenaemia is often prominent in pregnancy-associated DIC. Chronic DIC can be caused by fetal death.

Acute haemolysis

Haemolysis caused by incompatible blood transfusions, certain infections (for example, *Clostridium perfringens* septicaemia), or microangiopathic disorders such as TTP and HELLP, can sometimes be associated with DIC.

Immunological disorders

Severe allergic reactions (such as anaphylaxis), transplant rejection, glomerulonephritis, and other vasculitic disorders are sometimes associated with DIC.

Vascular anomalies

Giant haemangiomas cause overt DIC in about 25 per cent of affected patients (Kasabach–Merritt syndrome). Although activation of coagulation and fibrinolysis is localized to the vascular anomaly, depletion of haemostatic factors produces a clinical and laboratory profile indistinguishable from DIC. Eradication of haemangioma by radiation, embolization, or surgery is curative. Medical therapies have included heparin, antifibrinolytic drugs (combined with cryoprecipitate to thrombose the vascular tumour), glucocorticoids, and interferon.

DIC also occurs in about 0.5 to 1 per cent of patients with abdominal aortic aneurysms, which usually contain adherent thrombi.

Immunoglobulin-mediated factor deficiency

Coagulation-*factor* inhibitors are usually IgG antibodies that bind to specific coagulation factors, and either neutralize their activity (most coagulation-factor inhibitors) or result in accelerated clearance (for example, antiprothrombin antibodies associated with the antiphospholipid antibody syndrome). Acquired inhibitors against coagulation factors are rare in otherwise normal (non-haemophilic) individuals. Even the most common autoimmune coagulation-factor deficiency (factor VIII) has an estimated incidence of only 1 per 1 000 000 per year.

Acquired factor VIII inhibitor

Acquired factor VIII deficiency should be suspected in a patient with spontaneous bleeding, or bleeding following minor trauma, that occurs in association with a prolonged aPTT and a normal PT/INR (Table 3). Most commonly, muscle or cutaneous haematomas occur, but life-threatening retroperitoneal or intracranial haemorrhages are described; haemarthrosis is uncommon (cf. congenital haemophilia). The disorder occurs most commonly in the elderly (median age, 60 years), affects men and women equally, and is idiopathic in 50 per cent of cases. Other autoimmune disorders (for instance, systemic lupus erythematosus, **SLE**), lymphoid and other malignancies, penicillin treatment, or the postpartum state, have been observed in some patients. About 20 per cent of patients die of bleeding, often from their initial bleeding episode.

A rapid screening test for a coagulation-factor inhibitor is performed by repeating the aPTT after mixing patient plasma 50:50 with normal pooled plasma. An inhibitor is suggested by more than a 4-s prolongation time over the control, although some inhibitors require a 2-h incubation at 37 °C to show inhibition. Confirmation is obtained by a specific-factor assay showing reduced levels of factor VIII; inhibitor quantitation is most often performed by Bethesda assay, in which various dilutions of patient plasma are mixed with normal plasma and incubated for 2 h at 37 °C: a Bethesda unit (**BU**) is defined as the reciprocal of the plasma dilution that yields a 50 per cent reduction in residual factor VIII activity in the test system. Unfortunately, the Bethesda assay tends to underestimate the amount of inhibitor in non-haemophilic patients with acquired factor VIII inhibitors.

Therapy of bleeding depends upon its severity and the amount of inhibitor present, if known. For patients with minor bleeding, and low inhibitor levels (<5 BU), desmopressin (DDAVP) can be tried. Peak factor VIII levels occur between 45 and 90 min post-DDAVP, and repeat levels should be measured to assess efficacy. In other patients with low inhibitor levels but with more severe bleeding, purified human factor VIII concentrates are usually effective. One approach is to give an initial intravenous bolus of 100 U/kg, followed by a continuous infusion of factor VIII at 10 U/kg per h, with factor VIII levels measured again 4 to 6 h later. Careful clinical and laboratory assessment for response is needed, since inhibitor levels may have been underestimated, or higher inhibitor levels stimulated by factor VIII use.

Porcine factor VIII should be considered for bleeding patients in whom the inhibitor titres are not yet known, or to those with high inhibitor levels. This is because crossreactivity of the autoantibodies against porcine VIII is usually substantially less than with human factor VIII. After a loading dose of between 50 and 100 U/kg, repeat doses are given every 8 to 12 h, or further porcine VIII is given by constant intravenous infusion (4 U/kg per h).

Either PCCs or recombinant factor VIIa can be given for patients refractory to human or porcine factor VIII. Activated PCCs (for example, FEIBA or Autoplex) may be

somewhat more effective than non-activated PCCs, but their risk for causing thrombosis is greater. Factor VIIa is preferable for perioperative management, given its low risk for inducing thrombosis. In desperate situations, extracorporeal immunoadsorption using Staphylococcal protein A may be helpful in removing the antibodies.

Spontaneous disappearance of the inhibitor occurs in about 10 to 30 per cent of patients. Nevertheless, the unpredictable clinical course, and the potential for life-threatening bleeding, means that immunosuppressive therapy should be given to most patients, either with high-dose intravenous IgG (1 g/kg for 2 days, or 0.4 g/kg for 5 days) or corticosteroids (for instance, 1 mg/kg per day). For refractory patients, the substitution or addition of cyclophosphamide can be effective. Other options include combination chemotherapy (prednisone, cyclophosphamide, vincristine) or ciclosporin. Even partial remission can help reduce bleeding. Women with postpartum factor VIII inhibitors usually develop remission within 30 months, and usually do not develop recurrent factor VIII inhibitors with later pregnancies. They also may be less likely to respond to corticosteroids or other immunosuppressive therapy.

Other acquired coagulation-factor deficiencies

Hypoprothrombinaemia should be suspected in patients with the antiphospholipid antibody syndrome, particularly if bleeding occurs or the PT/INR is prolonged. Typically, these pathogenic anti-factor II antibodies are non-neutralizing, and therefore mixing patient plasma 50:50 with normal pooled plasma can produce correction of the aPTT, in contrast to other coagulation-factor inhibitors.

Thrombin inhibitors are rare, but may cause severe bleeding. More often, patients have antibodies that react preferentially against bovine thrombin: these are formed following the use of 'fibrin glue', which contains various bovine clotting factors. Patients have prolonged PT/INR, aPTT, and TCT (especially using bovine thrombin). However, it is more likely that any bleeding is the result of clinically significant anti-bovine factor V antibodies.

Factor V inhibitors

Rarely, IgG antibodies against factor V arise spontaneously or following treatment with topical bovine thrombin used at surgery. FFP usually does not provide enough factor V to treat bleeding; however, platelet transfusions are usually effective, as platelet activation causes factor V to be released into haemostatic plugs.

Factor XIII inhibitors

These inhibitors, which sometimes occur in association with isoniazid therapy, cause bleeding via impaired factor XIII-mediated crosslinking of fibrin. Factor XIII should be measured in a patient with unexplained bleeding and normal results of screening coagulation assays.

Factor X inhibitors

Factor X inhibitors are a rare cause of bleeding in patients with prolonged PT/INR and aPTT. The differential diagnosis also includes amyloidosis of the **AL** (amyloid light chain) variety, caused by adsorption of factor X to amyloid fibrils.

Factor IX inhibitors

In non-haemophilic patients, factor IX inhibitors are rare and usually associated with autoimmune disease. Treatment includes PCCs or purified factor IX, and immunosuppression. The differential diagnosis of acquired, isolated, factor IX deficiency includes the nephrotic syndrome (urinary loss of factor IX).

Factor XI inhibitors

These rare inhibitors are most often observed in association with systemic lupus erythematosus, and usually do not cause bleeding or require specific treatment.

Factor VII inhibitors

Factor VII inhibitors are extremely rare, and usually do not cause bleeding or require treatment. The diagnosis is suggested by an isolated prolonged PT/INR in the absence of coumarin or vitamin K deficiency.

Acquired von Willebrand syndrome

Rarely, bleeding is caused by a severe acquired deficiency of vWF, most often in the setting of a monoclonal gammopathy, benign or malignant. Typically, there is disproportional deficiency of the largest vWF multimers due to antibody-mediated clearance (acquired type 2a von Willebrand syndrome (vWD)). Aortic stenosis and obstructive cardiomyopathies are other causes of type 2a von Willebrand syndrome: this probably explains why aortic valve replacement has been reported to cure recurrent gastrointestinal haemorrhage in patients with colonic angiodysplasia.

Heparin and acquired heparin-like anticoagulants

Bleeding is a complication of heparin treatment, particularly when the aPTT is above the therapeutic range. In patients with massive accidental or deliberate heparin overdose, intravenous protamine should be given to treat bleeding complications.

Rarely, patients with spontaneous bleeding and prolonged aPTT and TCT measurements have circulating heparin-like anticoagulants. Usually associated with multiple myeloma and other plasma-cell dyscrasias, the coagulopathy does not necessarily respond even to large-dose protamine infusion, and fatal haemorrhage can ensue. Circulating dermatan sulphate glycosaminoglycan appeared to explain the bleeding in a patient with renal failure.

Coagulopathies secondary to plasma-cell dyscrasias

Multiple myeloma, macroglobulin*aemi*a, and other plasma-cell dyscrasias such as primary amyloidosis can cause various coagulopathies (Table 8). Usually, the TCT is prolonged, most often because of paraprotein-induced interference with fibrin polymerization. A long bleeding time suggests inhibition of platelet function by paraprotein; rarely, acquired von Willebrand syndrome is the cause. Apheresis can improve haemostasis by quickly reducing paraprotein levels, as antineoplastic chemotherapy is initiated.

Hyperfibrinolysis

Activation of fibrinolysis occurs normally when fibrin clots are formed during physiological or pathological haemostasis. However, primary fibrinolysis (Table 3) is sometimes the major cause for bleeding, and requires specific treatment.

Thrombolytic therapy

About 0.5 to 0.7 per cent of patients with myocardial infarction who receive thrombolysis with either streptokinase or tissue plasminogen activator (**t-PA**) develop an intracranial haemorrhage. The thrombolytic agent should be stopped immediately in any such patient, and they should receive cryoprecipitate and an antifibrinolytic drug (for instance, EACA); platelets and FFP can help to increase factor V and VIII levels that may have been reduced by plasmin generated by thrombolysis. It can take between 24 and 36 h for fibrinogen levels to recover after stopping thrombolytic therapy.

Malignancy

Cancer-associated DIC usually causes a hypercoagulable state. However, promyelocytic leuk *aemi*a (**PML**) and prostatic adenocarcinoma are two malignancies commonly associated with prominent hyperfibrinolysis. Laboratory abnormalities include prolonged PT/INR, aPTT, and TCT, and a hypofibrinogen *aemi*a. The use of all-*trans*- retinoic acid (**ATRA**) during induction chemotherapy of PML has reduced the frequency of life-threatening bleeding. Antifibrinolytic therapy can control

bleeding in cancer-associated fibrinolysis, but there is a risk of thrombosis if tissue factor-induced DIC, rather than the release of plasminogen activator by the tumour, is primarily responsible for the coagulopathy.

Cardiopulmonary bypass surgery

Excess bleeding, defined as more than 1 litre per procedure, is a common problem following heart surgery utilizing cardiopulmonary bypass (extracorporeal circulation). About 20 per cent of all red cell concentrates in the United States are given for cardiac surgical bleeding. About 5 per cent of patients require urgent resternotomy for critical rates of blood loss (defined as: >500 ml in the first h; >400 ml/h in the first 2 h; >300 ml/h in the first 3 h; or >1 litre in 4 h). Re-exploration reveals bleeding vessels in two-thirds of patients; the remainder have diffuse oozing.

Thrombocytopenia, transient platelet dysfunction, and hyperfibrinolysis are the major haemostatic defects. Typically, the platelet count falls by between 30 and 60 per cent mainly from haemodilution, although platelet losses from bleeding and within the extracorporeal perfusion device also occur. The thrombocytopenia persists for 3 to 4 days, followed by recovery of the platelet count to values exceeding the preoperative baseline. Marked prolongation of the bleeding time (>30 min) quickly improves to under 15 min shortly after surgery, and to normal several hours later. Some platelet function defects are 'extrinsic' and reversible (for example, hypothermia, heparin), whereas others indicate longer-lasting 'intrinsic' changes (surface glycoprotein deficiency, acquired-granule depletion). Preoperative treatment with aspirin or abciximab also increases bleeding.

The importance of hyperfibrinolysis in postcardiac surgical bleeding is suggested by meta-analysis of studies of high-dose aprotinin, a plasmin inhibitor derived from bovine lung: a two-thirds reduction in blood transfusion, and 50 per cent reduction in resternotomy. Other antifibrinolytic drugs that reduce bleeding include the lysine analogues, tranexamic acid (for example, 10 mg/kg bolus pre-cardiopulmonary bypass (**CPB**); then 1 mg/kg per h, although dosing regimens vary widely) and EACA (total dose, up to 20 g). Although these therapies are usually given before CPB, they may also provide benefit when used postoperatively for bleeding patients.

Management of postcardiac surgical bleeding also includes blood transfusions, especially platelets and fresh-frozen plasma, although their benefit is unproven. Residual heparin, including heparin 'rebound', can respond to additional protamine. Desmopressin probably is ineffective. No universally accepted algorithm for management exists.

Liver disease

Hyperfibrinolysis complicating liver disease is discussed elsewhere.

Venom-induced coagulopathies (snake bites) (see also Chapter 8.2)

Envenomations can harm or kill humans generally through systemic effects, for instance profound hypotension. Sometimes, however, life-threatening coagulopathies result.

Snake bites

In the United States, about 8000 bites from venomous snakes occur each year, resulting in 10 to 20 deaths. This relatively low mortality reflects the less lethal character of New World snakes, as well as the victim's usual close proximity to medical facilities and antivenin therapy. Pit vipers (rattlesnakes, copperheads, cottonmouths, massasaugas) account for 99 per cent of snakebite poisonings in the United States. Worldwide, about 30 000 to 40 000 people die from snakebite, about half in India. Although death usually results from multiple mechanisms (such as circulatory shock, rhabdomyolysis, renal failure, pulmonary failure, neurotoxicity), bleeding is sometimes the major factor.

Venoms contain multiple digestive enzymes with a broad spectrum of activity that can include effects on human haemostasis (Table 9). Within a species, haemostatic effects of envenomation vary with snake age, diet, and other factors. North American rattlesnakes typically cause the 'defibrination syndrome'; despite even profound hypofibrinogen*aemia*, bleeding is uncommon. In contrast, venom from Old World vipers frequently cause generalized activation of the coagulation system (DIC), with a greater chance for bleeding or microvascular thrombosis. Bleeding can also result from platelet inhibitors present within venom; for example, the platelet fibrinogen receptor antagonist, echistatin (*Echis carinatus*), or 'haemorrhagins' such as jararhagin (*Bothrops jararacussu*) that damage endothelium.

Immediate treatment of a snake bite includes efforts to limit the venom spread (immobilizing and placing a constriction band proximal to the bite site). Rapid transport to medical facilities is crucial since antivenin therapy is the mainstay of treatment. Antivenin treatment is indicated for patients with significant pain or swelling, as well as suspected or proven haemostasis abnormalities, as these indicate envenomation rather than a 'dry bite'. Hypersensitivity testing to the antivenin should be performed to rule out pre-existing hypersensitivity to horse serum. The treatment of snake bite is discussed in Chapter 8.2.

Coagulation studies should include: complete blood count (including platelets), PT/INR, aPTT, TCT, fibrinogen, and FDPs. Abnormal results indicate envenomation, and are an indication for antivenin therapy. The bedside assessment of defibrination involves placing a few millilitres of blood in a clean, dry test tube at room temperature for 20 min; incoagulable blood indicates defibrination. Usually, blood products should only be given to patients with bleeding. A small clinical trial found that heparin was ineffective in patients with DIC caused by a Russell's viper bite.

Laboratory and therapeutic uses of snake venoms

Snake-venom fractions are useful for certain laboratory assays. For example, the thrombin-like enzyme, batroxobin (Reptilase®, *Bothrops atrox moojeni*), cleaves fibrinopeptide A from fibrinogen even in the presence of heparin. Thus, a prolonged Reptilase time indicates hypofibrinogen *aemia* even in heparin-containing plasma.

Ecarin activates prothrombin irrespective of its g-carboxylation status; thus, it can be used to detect **PIVKA** (proteins induced by vitamin K antagonists) to document vitamin K deficiency or dysprothrombin*aemia*. An ecarin clotting time (**ECT**) is superior to the aPTT for monitoring therapy with hirudin, particularly the high doses used for heart surgery. Differences in phospholipid dependency of venom prothrombin activators has led to the use of a Textarin®/ecarin ratio to detect lupus anticoagulants; a ratio over 1.3 is a sensitive and relatively specific test for lupus anticoagulants.

Russell's viper venom contains a potent activator of factor X (**RVV-X**); the dilute Russell's viper venom time (**dRVVT**), performed by adding RVV-X and diluted rabbit brain phospholipid to test plasma prior to recalcification, measures the rate of formation and activity of the phospholipid-dependent prothrombinase complex in producing thrombin. The dRVVT is thereby prolonged in the presence of a lupus anticoagulant.

A commercially available protein C activator (Protac®) from *Agkistrodon contortrix contortrix* (the southern copperhead) has greatly simplified assays for protein C activity, as well as in screening for defects in the protein C anticoagulant pathway.

The defibrinogenating snake venom, ancrod (Arvin®, derived from the Malayan pit viper, *Calloselasma [Agkistrodon] rhodostoma*), which proteolyses fibrinopeptide A, has been used for antithrombotic therapy, including the management of heparin-induced thrombocytopenia (**HIT**), acute stroke, thrombotic nephropathy, and priapism. The inability to control thrombin generation is a potential drawback of this therapy. Batroxobin (Defibrase®) is another defibrinogenating venom that has seen limited clinical applications.

Prothrombotic acquired coagulation disorders

Some acquired coagulation disorders are characterized by an increased risk for thrombosis, rather than bleeding. Accordingly, the appropriate treatment usually involves anticoagulant therapy, even if there are abnormal coagulation or platelet count values.

Macrovascular thrombosis

Some acquired coagulation disorders typically cause thrombosis in large veins and arteries, although small-vessel thrombi can also result.

Heparin-induced thrombocytopenia

Heparin-induced thrombocytopenia (**HIT**) is caused by IgG antibodies that recognize multimolecular complexes of platelet factor 4 (**PF4**) and heparin. Thrombosis results from IgG-induced platelet activation (via platelet Fc receptors), resulting in the generation of procoagulant, platelet-derived microparticles, tissue-factor expression by endothelium, and inactivation of heparin by PF4 released from platelets. Increased thrombin–antithrombin complex levels indicate DIC in almost all patients with HIT, although a prolonged INR or low fibrinogen level occur in less than 10 per cent of cases.

Typically, the fall in platelet count begins 5 to 10 days after starting heparin; however, in patients who received heparin within the past 100 days, the platelet count can fall abruptly upon resuming heparin therapy, probably because of residual circulating HIT antibodies. HIT occurs in as many as 5 per cent of certain high-risk populations: for example, postoperative orthopaedic patients receiving unfractionated heparin. HIT is less frequent in patients initially treated with low-molecular-weight heparin (**LMWH**).

Most patients with HIT develop venous or arterial thrombosis (Fig. 2), most commonly a deep-vein thrombosis (**DVT**), pulmonary embolism, major limb artery thrombosis, stroke, or myocardial infarction. Acute or chronic adrenal failure from bilateral adrenal haemorrhagic necrosis has been described. The thrombocytopenia is typically moderate in severity (median platelet count nadir, 60 × 10^9/l), but in only 10 per cent of patients does the platelet count fall to less than 20 × 10^9/l. In at least 10 per cent of patients, the platelet count never drops below 150 × 10^9/l (Fig. 2).

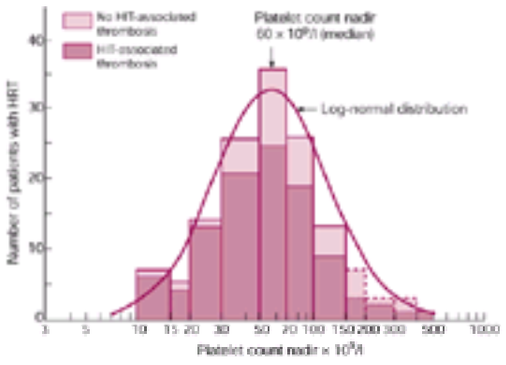

Fig. 2 Thrombosis in relation to the severity of thrombocytopenia in patients with HIT. The magnitude of thrombocytopenia in HIT shows a log-normal distribution. Thrombocytopenia is typically mild-to-moderate (between 20 and 150 × 10^9/l in about 80 per cent of patients; median platelet count nadir, about 60 × 10^9/l). Thrombosis occurs in 50 per cent or more of patients with HIT, irrespective of the platelet count nadir, including patients whose platelet count never falls below 150 × 10^9/l. (Reprinted with permission from Warkentin, 1998.)

Laboratory testing for HIT antibodies includes activation and antigen assays. The former assays detect antibodies via their heparin-dependent, platelet-activating properties. Commercially available antigen assays detect antibodies that bind to surface-immobilized PF4 complexed to heparin or polyvinylsulphonate.

Treatment includes stopping heparin and instituting alternative anticoagulation. Coumarin alone should not be given to patients with acute HIT, particularly to those with associated DVT, as there is a risk for inducing progression to venous limb gangrene. The dramatic natural history of HIT, with a risk for subsequent thrombosis of about 50 per cent even after stopping heparin, means that an alternative anticoagulant, together with DVT surveillance, should be considered for all patients with suspected HIT. Suitable anticoagulants with a rapid-onset of action include danaparoid (a low-molecular-weight heparinoid with predominant anti-factor Xa activity), lepirudin (a recombinant hirudin with potent antithrombin activity derived from leech salivary glands), and argatroban (a synthetic, small-molecule, direct thrombin inhibitor). Among patients with HIT, LMWH treatment has a high risk for clinical crossreactivity, and should be considered a contraindicated treatment for acute HIT. Many patients will benefit from selected adjunctive treatments, for instance surgical thromboembolectomy for acute arterial thrombosis of a limb.

Adenocarcinoma-associated chronic DIC

Metastatic adenocarcinoma sometimes presents as venous or arterial thrombosis accompanied by DIC. The diagnosis is suggested by an unexpected rise in the platelet count during heparin treatment, followed by an abrupt platelet count fall, together with new or progressive thrombosis, when heparin is stopped, despite therapeutic anticoagulation with warfarin. The clinical situation can mimic HIT ('pseudo-HIT'), but HIT antibodies are absent, and the platelet count recovers during resumption of heparin (Fig. 3). Oral anticoagulants are ineffective, and may even cause venous limb gangrene (discussed subsequently). Heparin, especially LMWH, is the preferred treatment. Tissue factor-containing tumour ves-icles, and factor Xa-activating enzymes found in tumour extracts, are two possible explanations for these procoagulant effects of adenocarcinoma.

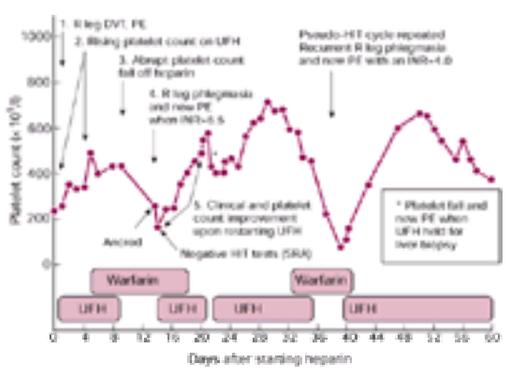

Fig. 3 Pseudo-HIT. Adenocarcinoma with thrombocytopenia and phlegmasia cerulea dolens after stopping unfractionated heparin (UFH). The timing of thrombocytopenia onset suggested HIT, prompting the use of an alternative anticoagulant (ancrod). Heparin was restarted when HIT antibodies were not detected by activation assay (serotonin-release assay, SRA). Subsequently, heparin discontinuation led to the recurrence of thrombocytopenia and warfarin-associated phlegmasia cerulea dolens (repeat of pseudo-HIT cycle). Abbreviations: DVT, deep venous thrombosis; INR, international normalized ratio; PE, pulmonary embolism.

Antiphospholipid antibody syndrome ('lupus anticoagulant')

This clinicopathological syndrome is characterized by large-vessel venous and/or arterial thrombosis, recurrent miscarriages, and thrombocytopenia. An associated 'lupus anticoagulant' (or 'non-specific inhibitor') is a prolonged aPTT that results from the interference by antibodies against phospholipid-dependent coagulation reactions; these antiphospholipid antibodies are usually directed against protein cofactors such as b $_2$-glycoprotein I (**b2GPI**) and prothrombin. Sometimes, a prolonged PT/INR is caused by non-neutralizing antiprothrombin antibodies that cause hypoprothrombinaemia by increased prothrombin clearance.

Despite these laboratory abnormalities, bleeding is unusual, since severe thrombocytopenia or hypoprothrombinaemia is uncommon. More often, antiphospholipid antibodies are associated with intermittent thrombosis; rarely, the abrupt onset of life-threatening multiple vascular occlusions occurs ('catastrophic antiphospholipid antibody syndrome'). The explanation for the paradoxical association with thrombosis remains elusive, but it could be caused by antibody interactions with other protein cofactors described (for example, activated protein C, protein S, thrombomodulin). Many patients have a thrombocytopenia that is typically mild and intermittent. Other less common complications include cardiac valvulitis and microvascular thrombosis, which can manifest as acrocyanosis, digital

ulceration/gangrene, and livedo reticularis.

Antiphospholipid antibodies are detected by enzyme-linked immunosorbent assays (**ELISA**) using purified phospholipids as the target antigen, for example the anticardiolipin antibody assay. Lupus anticoagulant activity is shown by demonstrating inhibition of phospholipid-dependent coagulation assays. Several assays should be performed, as anti-b$_2$GPI antibodies especially interfere with the conversion of prothrombin to thrombin (that is, best detectable by dRVVT), whereas antiprothrombin antibodies interfere most with global coagulation assays (for instance, kaolin clotting time). The coagulation times remain prolonged following mixing with normal plasma; confirmation involves adding excess phospholipid to neutralize the effects of the antiphospholipid antibodies. Not all aPTT reagents are sensitive to antiphospholipid antibodies, and so these phospholipid-dependent coagulation assays should be performed in the appropriate clinical situation, even if the aPTT is normal.

The term 'lupus anticoagulant' refers to the frequent occurrence of these antibodies in patients with systemic lupus erythematosus; nevertheless, most patients with the antiphospholipid antibody syndrome do not have SLE. Some patients have other autoimmune disorders, malignancy, infections, or procainamide treatment, but usually no associated condition is identified (primary antiphospholipid antibody syndrome). Many patients require long-term anticoagulation, although the optimal agents and therapeutic level of anticoagulation remain to be defined. Corticosteroids can benefit patients with bleeding caused by hypoprothrombinaemia.

Microvascular thrombosis

Some disorders of haemostasis are characterized by small-vessel thrombi, affecting either arterioles (for example, TTP) or small venules (for example, coumarin-induced necrosis).

Thrombotic microangiopathy

Thrombotic microangiopathy is a clinicopathological syndrome of microangiopathic haemolysis and thrombocytopenia carrying a risk for arteriolar occlusion by microaggregates of platelets and vWF, particularly affecting the kidneys and central nervous system. Microangiopathic red cell changes are characteristic, for example 'helmet cells' (schistocytes) and small, triangle-shaped, red cell fragments. The prototypic illness is thrombotic thrombocytopenic purpura (**TTP**), which typically affects adults and is idiopathic. However, familial and secondary forms of TTP also exist. The haemolytic-uraemic syndrome is a nephrotropic variant of TTP with a distinct pathogenesis, including its association with verocytotoxin-producing *Escherichia coli* acquired from eating undercooked meat (hamburger disease).

The pathogenesis of TTP involves the formation of platelet–vWF microaggregates in high shear situations (arterioles). Platelet-bound vWF levels are increased during TTP. Patients with familial TTP have ultra-large multimers of vWF during remission; these very large multimers disappear during active disease. Recently, a constitutional deficiency of a vWF-cleaving metalloproteinase has been identified in patients with familial TTP. In patients with non-familial TTP, an IgG autoantibody, which inhibits the vWF-cleaving metalloproteinase, has been identified that disappears in remission.

The mainstays of treatment for acute TTP are corticosteroids and fresh-frozen plasma given by infusion or apheresis. Corticosteroids, often given as prednisone 200 mg/day, may treat the autoimmune component of TTP. Provision of either fresh-frozen plasma, or the cryoprecipitate-depleted fraction of plasma (cryosupernatant), has greatly reduced mortality in TTP, possibly by providing limited disulphide-bond reductase activity that facilitates vWF cleavage. Furthermore, apheresis may help cleave the pathogenic autoantibody and large vWF multimers.

Coumarin-induced skin necrosis

Coumarin-induced skin necrosis (**CISN**) is characterized by necrosis of the skin and underlying subcutaneous tissues that typically begins 3 to 6 days after commencing warfarin or coumarin anticoagulants. CISN results from failure of the protein C natural anticoagulant system to downregulate thrombin generation in the microvasculature. The relatively short half-life of protein C, compared with prothrombin, explains the temporal profile of CISN—that is to say, a transient period of disproportionately reduced protein C activity soon after starting coumarin (Table 10). Furthermore, a relatively high proportion of affected patients have a hereditary abnormality of the protein C anticoagulant pathway, especially protein C deficiency. Other disorders associated with CISN include congenital deficiency in protein S or antithrombin, factor V Leiden, and HIT. The pathology is a predominantly non-inflammatory, small-vessel thrombosis affecting the subcutaneous postcapillary venules and small veins.

CISN characteristically affects central (non-acral) sites with substantial underlying fatty tissues, such as the breast, buttocks, hips, and thighs (Fig. 4). Less common areas include the anterior abdomen, flank, back, penis, legs, arms, and face. About 75 per cent of patients are women; one-third have multiple lesions that can be symmetrical. The earliest features are localized pain, induration, and erythema; over the next few hours, the skin lesions progress to central purplish or black discoloration, with blistering, subsequently demarcating to full-thickness skin necrosis. CISN is rare (1/10 000 patients treated with warfarin).

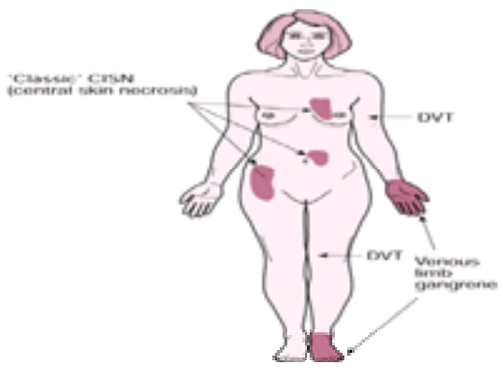

Fig. 4 Coumarin-induced skin necrosis: 'classic' syndrome (usually affecting central tissue sites) and coumarin-induced venous limb gangrene. Typically, an active deep-vein thrombosis (DVT) subtends the distal extremity affected by venous limb gangrene. (Reprinted with permission from Warkentin, 1996.)

Prompt reversal of anticoagulation with vitamin K may prevent incipient CISN if recognized early. However, the diagnosis is usually not made until necrosis is established; at this point, it is unknown whether vitamin K, fresh-frozen plasma, or protein C concentrates alter its natural history. In patients without HIT, warfarin is usually replaced by heparin. Many patients require surgical treatment, such as skin grafting or tissue amputation. Following recovery, it is usually safe to reintroduce warfarin provided certain precautions are taken, for example the gradual initiation of oral anticoagulation.

Coumarin-induced venous limb gangrene

Venous limb gangrene involves the acral (peripheral) regions of the body—most often the toes, feet, and legs, but sometimes also the fingers, hands, and arms—usually in association with DVT. The severity ranges from an initial stage of phlegmasia caerulea dolens ('swollen, blue, painful' limb) to extensive venous limb gangrene requiring limb amputation. Two disorders predispose to coumarin-induced venous limb gangrene: HIT and cancer-associated DIC. Recent data suggest that the supratherapeutic INR (typically, >3.5) that characterizes venous limb gangrene is caused by a severe reduction in factor VII, which parallels a severe reduction in protein C activity that explains the microvascular thrombosis underlying this syndrome. Essentially, coumarin interferes with the protein C anticoagulant pathway, while at the same time it is unable to control the increased thrombin generation characteristic of HIT or cancer-associated DIC.

Purpura fulminans

Purpura fulminans is a rare syndrome of DIC and microvascular thrombosis that results in multicentric ischaemic necrosis of the skin and subcutaneous tissues, predominantly affecting the extremities. The most common cause is overwhelming septicaemia, especially with meningococcus. A severe, acquired reduction in protein C activity complicating DIC is the most likely cause for the microvascular thrombosis, and some experts recommend treatment with protein C concentrates, if available. Autoantibodies against protein S have been implicated in patients with postvaricella purpura fulminans. In other patients with apparent 'idiopathic' purpura

fulminans, autoantibodies that interfere with the protein C anticoagulant system have been described.

P>**Septicaemia and other systemic inflammatory response syndromes**

Multiple organ failure often complicates septicaemia and other systemic inflammatory disease syndromes, including adult respiratory distress syndrome, fat embolism, and acute pancreatitis. Thrombocytopenia and coagulopathy are common, and some patients have DIC that could contribute to organ dysfunction via microvascular thrombosis. However, a prothrombotic basis for organ failure is usually speculative, as microthrombosis is rarely documented pathologically, and non-thrombotic microvascular disturbances that impair tissue oxygen delivery also occur.

Haemostasis in the newborn

Neonatal vitamin K deficiency

Haemorrhagic disease of the newborn caused by vitamin K deficiency was once a relatively common cause of bleeding during the first week of life, particularly in breast-fed infants. Low vitamin K levels in mother's milk, and insufficient colonization of the newborn bowel by vitamin K-produc-ing bacteria, predispose to the inability to meet the infant's vitamin K requirements (1 µg/kg per day). The routine administration of vitamin K, either 1 mg given intramuscularly immediately after birth, or three oral doses of vitamin K, has led to the near-disappearance of this problem. Bleeding within 24 h of birth can occur in certain high-risk settings, for example mothers receiving anticonvulsants or warfarin; in these cases, the mother should receive vitamin K, 10 mg by mouth, each day for 2 weeks prior to delivery. Vitamin K deficiency occurring later in infancy despite appropriate neonatal vitamin K prophylaxis can indicate hepatobiliary or bowel disease.

Neonatal disseminated intravascular coagulation

DIC commonly complicates neonatal infection, asphyxia, respiratory distress syndrome, aspiration of meconium or amniotic fluid, maternal hypertensive syndrome, hypothermia, and brain injury. This condition poses a significant risk of bleeding or thrombosis, as the immature liver has an impaired capacity to synthesize coagulation factors, and the reticuloendothelial system has a limited ability to clear activated coagulation factors. Treatment is aimed at the underlying cause of the DIC, with blood product given for the bleeding.

Neonatal purpura fulminans

Purpura fulminans can begin within hours or days following birth, often first affecting the heels or venepuncture sites. The underlying cause is usually a congenital abnormality affecting the protein C anticoagulant system (homozygous deficiency of protein C or protein S; homozygous factor V Leiden), although infection with group B b-haemolytic streptococcus is described. Fresh-frozen plasma given every few days prevents a recurrence in some patients.

Further reading

Ansell J, *et al.* (2001). Managing oral anticoagulant therapy. *Chest* **119**, 22S–38S. [Discusses the management of non-therapeutic (elevated) INRs in patients receiving oral anticoagulants (Recommendations of the Sixth American College of Chest Physicians Consensus Conference on Antithrombotic Therapy).]

Asherson RA, *et al*. (1998). Catastrophic antiphospholipid syndrome. Clinical and laboratory features of 50 patients. *Medicine* (Baltimore) **77**, 195–207. [Describes clinical presentations of multiorgan failure in patients with antiphospholipid antibodies.]

Bevan DH (1999). Cardiac bypass haemostasis: putting blood through the mill. *British Journal of Haematology* **104**, 208–19. [Excellent review of cardiopulmonary bypass surgery and approaches to bleeding.]

Bossi P, *et al*. (1998). Acquired hemophilia due to factor VIII inhibitors in 34 patients. *American Journal of Medicine* **105**, 400–8. [Summarizes the presentation and clinical course of this disorder.]

Cole MS, Minifee PK, Wolma FJ (1988). Coumarin necrosis—a review of the literature. *Surgery* **103**, 271–7. [Comprehensive review of coumarin-induced skin necrosis.]

Hirsh J, *et al*. (1998). Oral anticoagulants. Mechanism of action, clinical effectiveness, and optimal therapeutic range. *Chest* **114**, 445S–469S. [Lists drugs and foods that interact with warfarin.]

Hutton RA, Warrell DA (1993). Action of snake venom components on the haemostatic system. *Blood Reviews* **7**, 176–89. [Good synthesis of clinical and laboratory aspects of snake envenomation.]

Kitchens CS (1992). Hemostatic aspects of envenomation by North American snakes. *Hematology/Oncology Clinics of North America* **6**, 1189–95. [The focus is on envenomation by North American snakes, resulting in defibrination rather than true DIC syndromes.]

Levi M, Ten Cate H (1999). Disseminated intravascular coagulation. *New England Journal of Medicine* **341**, 586–92. [Recent review.]

Levine JS, Branch DW, Rauch J (2002). The antiphospholipid syndrome. *New England Journal of Medicine* **346**, 752–63. [Recent review.]

Manco-Johnson MJ, *et al*. (1996). Lupus anticoagulant and protein S deficiency in children with postvaricella purpura fulminans or thrombosis. *Journal of Pediatrics* **128**, 319–23. [Provides evidence that purpura fulminans following varicella infection is usually caused by autoantibodies to protein S.]

Marsh NA (1998). Use of snake venom fractions in the coagulation laboratory. *Blood Coagulation and Fibrinolysis* **9**, 395–404. [Discusses many laboratory uses of snake venom fractions.]

Meier J, Stocker K (1991). Effects of snake venoms on hemostasis. *Critical Reviews in Toxicology* **21**, 171–82. [Reviews multiplicity of effects of snake venoms on hemostasis.]

Moake JL, Chow TW (1998). Thrombotic thrombocytopenic purpura: understanding a disease no longer rare. *American Journal of Medical Sciences* **316**, 105–19. [Excellent summary of new concepts of TTP, including role for abnormalities in vWF-cleaving metalloproteinase and the autoimmune pathogenesis of TTP.]

Ortel TL, *et al.* (1994). Topical thrombin and acquired coagulation factor inhibitors: clinical spectrum and laboratory diagnosis. *American Journal of Hematology* **45**, 128–35. [Summarizes acquired coagulation inhibitors that occur following treatment with topical bovine thrombin preparations ('fibrin glue').]

Sane DC, *et al.* (1989). Bleeding during thrombolytic therapy for acute myocardial infarction: mechanisms and management. *Annals of Internal Medicine* **111**, 1010–22. [Describes the management of post-thrombolytic hemorrhage.]

Warkentin TE, *et al*. (1997). The pathogenesis of venous limb gangrene associated with heparin-induced thrombocytopenia. *Annals of Internal Medicine* **127**, 804–12. [Indicates that an oral anticoagulant (warfarin) can cause deep-vein thrombosis to progress to venous limb gangrene in patients with heparin-induced thrombocytopenia.]

Warkentin TE, *et al.* (1995). Heparin-induced thrombocytopenia in patients treated with low-molecular-weight heparin or unfractionated heparin. *New England Journal of Medicine* **332**, 1330–5. [Provides evidence that HIT is a prothrombotic state associated with venous and arterial thrombosis that occurs less frequently with low-molecular-weight heparin.]

Warkentin TE (1996). Heparin-induced thrombocytopenia IgG-mediated platelet activation, platelet microparticle generation, and altered procoagulant/anticoagulant balance in the pathogenesis of thrombosis and venous limb gangrene complicating heparin-induced thrombocytopenia. *Transfusion Medicine Reviews* **10**, 249–58. [Compares and contrasts the pathogenesis and clinical profile of coumarin-induced skin necrosis and coumarin-induced venous limb gangrene.]

Warkentin TE (1998). Clinical presentation of heparin-induced thrombocytopenia. *Seminars in Hematology* **35**(Suppl. 5), 9–16. [Summarizes the clinical features of HIT.]

Warkentin TE (2001). Venous limb gangrene during warfarin treatment of cancer-associated deep venous thrombosis. *Annals of Internal Medicine* **135**, 589–93. [Implicates warfarin in the pathogenesis of venous limb gangrene complicating cancer-associated DIC.]

Warkentin TE (2001). Pseudo-heparin-induced thrombocytopenia. In Warkentin TE, Greinacher A, eds. *Heparin-induced thrombocytopenia*, 2nd edn, pp. 271–89. Marcel Dekker, New York. [Describes disorders that clinically mimic heparin-induced thrombocytopenia.]

Wells PS, *et al*. (1994). Interactions of warfarin with drugs and food. *Annals of Internal Medicine* **121**, 676–83. [Identifies drugs and foods that interact with warfarin.]

22.7 The blood in systemic disease

D. J. Weatherall

There are few diseases that do not produce some alteration in the blood. Here, some of the haematological changes associated with general systemic diseases will be summarized. Many of these topics are discussed elsewhere in this book but they are brought together in order to emphasize how blood changes may give the first indication of the presence of non-haematological disorders. It should be remembered that the haematological consequences of systemic disease vary considerably depending on the age of the patient. Recent reviews which deal specifically with this topic in children and the elderly are cited at the end of this chapter.

Malignant disease

The most common haematological finding in malignant disease (Table 1) is the anaemia of chronic disorders, which was described in Chapter 22.5.3. It may occur together with localized or widespread malignancy and is sometimes associated with an elevated erythrocyte sedimentation rate (ESR). It is found in patients with practically every type of carcinoma or reticulosis, is refractory to haematinics, but may respond to successful removal of a primary tumour.

The anaemia of patients with carcinoma, particularly of the gastrointestinal tract, may be complicated by chronic blood loss and superimposed iron deficiency. Chronic bleeding of this type is often associated with a mild thrombocytosis.

Disseminated malignancy

The most common haematological change with disseminated malignancy is a leucoerythroblastic picture characterized by the presence in the blood of immature myeloid cells together with some nucleated red cells, and, sometimes, a mild reticulocytosis. The red cells often show a moderate degree of anisocytosis and poikilocytosis. This finding is very commonly accompanied by the presence of tumour cells in the bone marrow. Clinically, it can cause confusion with the diagnosis of primary myelosclerosis; but splenomegaly is unusual in patients with disseminated carcinoma.

Occasionally, widespread carcinoma leads to a leukaemoid reaction with white-cell counts in the range seen in chronic myeloid leukaemia. The differentiation between these two conditions was described earlier.

The microangiopathic haemolytic anaemia of disseminated malignancy is most frequently found in association with a mucin-secreting adenocarcinoma, particularly of the stomach, breast, and lung.

Less common forms of anaemia associated with cancer

Autoimmune haemolytic anaemia is sometimes found in patients with an underlying lymphoma. It is much less common in other forms of malignancy except for the association with tumours of the ovary. However, there have been reports of autoimmune haemolysis occurring with a wide variety of tumours, including lung, stomach, breast, kidney, colon, and testis.

Pure red-cell aplasia may occasionally be the presenting feature in a patient with a tumour of the thymus. There have been occasional reports of this type of anaemia occurring in patients with carcinoma of the bronchus or lymphomas.

Finally, it should be remembered that there is an association between pernicious anaemia and carcinoma of the stomach. A patient may present with a megaloblastic anaemia associated with a malignancy of this type. In the early literature on sideroblastic anaemia, an association with carcinoma was well documented. Since acquired sideroblastic anaemia has been classified as part of the myelodysplastic syndrome there seem to have been no further reports of this association and its significance remains uncertain.

Polycythaemia

The relation between secondary polycythaemia and an underlying neoplasm is discussed in Chapter 22.4.14. It has been found in patients with renal tumours, hepatomas, hamartomas of the liver, uterine fibroids, vascular tumours and cystic adenomas of the cerebellum, and carcinoma of the lung.

Changes in the platelets and blood coagulation

An otherwise unexplained thrombocytosis may be the first indication of an underlying malignancy. It is important to remember that this is not always associated with chronic blood loss; bronchial carcinoma may present in this way. Thrombocytopenia may sometimes occur with bone marrow infiltration by tumour cells, but is seen most frequently as a side-effect of chemotherapy. Autoimmune thrombocytopenia has been observed most commonly in association with lymphoid malignancies, but it can also occur in association with tumours of the lung, breast, and testes.

Generalized haemostatic failure associated with disseminated carcinoma is considered in detail elsewhere (Fig. 1 and Fig. 2).

Fig. 1 Disseminated intravascular coagulation in association with carcinoma of the prostate. The patient started to bleed extensively from the iliac-crest marrow biopsy site and from venesection sites. Marrow biopsy showed widespread tumour metastases. (Reproduced from Hardisty RM, Weatherall DJ (ed.) (1982). *Blood and its disorders*, 2nd edn. Blackwell Scientific, Oxford, with permission.)

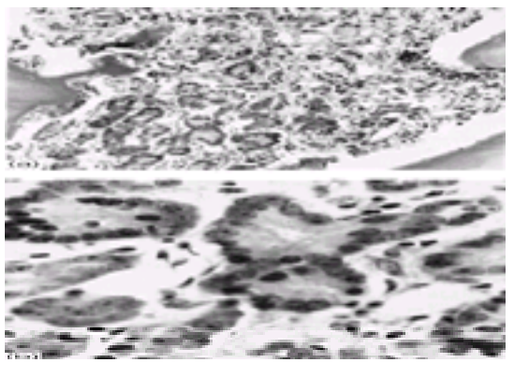

Fig. 2 Section prepared from Gardner-needle biopsies from bone marrow infiltrated with neoplastic cells; the primary tumour was in the prostate. (Reproduced from Hardisty RM, Weatherall DJ (ed.) (1982). *Blood and its disorders*, 2nd edn. Blackwell Scientific, Oxford, with permission.) (a) H and E stain ×230; (b) H and E stain ×920.

Some bleeding disorders associated with cancer seem to be due to selective impairment of coagulation. This may result from pathological inhibitors of different parts of the coagulation system or from isolated factor deficiencies. The mechanism is unknown. If the bleeding disorder is not characterized by consumption of clotting factors or fibrinolysis, a detailed analysis of the activities of the intrinsic and extrinsic pathways must be made in case a correctable lesion is present (Table 2).

Patients with cancer have an increased tendency to thrombosis. Apart from debilitation and periods of prolonged bed rest there is undoubtedly a hypercoaculable state associated with many tumours. This seems to involve a variety of procoagulants including fibrinogen and factors V, VII, VIII, IX, and XI. Low-grade, disseminated, intravascular coagulation can consume anticogulants such as anti-thombin III, protein C, and protein S. Cancer cells can initiate clotting by releasing a tissue factor, a phenomenon which is described in patients with lung, kidney, colon, and breast cancers. The syndrome of non-bacterial thrombotic endocarditis, characterized by cerebral embolic strokes and extensive fibrin/platelet vegetations on the mitral and aortic valves, is most commonly associated with cancers of the lung, prostate, and pancreas.

White-cell abnormalities

In addition to the leukaemoid reaction, there are several white-cell changes that should make the clinician think about an underlying malignancy. For example a persistent monocytosis or eosinophilia may be associated with Hodgkin's disease or with bronchial carcinoma. Persistent lymphopenia may occur in patients with Hodgkin's disease.

Haematological changes due to cancer chemotherapy

Many agents used in cancer chemotherapy depress the bone marrow causing varying periods of neutropenia and thrombocytopenia associated with a variable anaemia. The bone marrow may also show marked myelodysplastic features. Haemolytic reactions have been associated with a number of drugs, including mitomycin C, and bleomycin–cisplatin. In some cases these drugs are associated with a syndrome of microangiopathic haemolytic anaemia resembling the haemolytic uraemic syndrome. Circulating immune complexes have been observed in some cases and there have been reports of response to plasma exchange and immunosuppression. Some chemotherapeutic agents appear to cause a warm antibody type of haemolytic anaemia. In patients who are glucose-6-phosphate dehydrogenase deficient, the administration of doxorubicin can produce a haemolytic reaction.

Haemophagocytic syndrome

This disorder, which is described in a later section in its association with viral illness, has now been reported in patients with cancer, lymphoma, and acute leukaemia. It is characterized by pancytopenia, fever, and splenomegaly; the bone marrow resembles histiocytosis with intense haemophagocytosis by macrophages.

Infection

Most of the important haematological changes in association with infection are considered in Section 7. Just a few points of particular haematological relevance are summarized below.

Acute bacterial infection

Most acute bacterial infections are associated with a neutrophil leucocytosis. This may be so marked, and associated with a 'shift to the left' with production of myelocytes in the blood, that the condition may present a leukaemoid type of reaction. Occasionally, in patients who are severely ill with acute bacterial infection, the neutrophil response seems inadequate. Some may be frankly neutropenic. A number of these individuals will prove to have an underlying haematological disorder or a debilitating condition such as alcoholism, but many who recover from their infection show no such underlying abnormality. A marrow examination usually reveals a paucity of mature granulocytes. This clinical picture is particularly common in newborn infants, especially those born prematurely. Some infections seem to be particularly prone to association with a reduced white-cell count. They include salmonellosis, brucellosis, pertussis, rickettsial infections, disseminated tuberculosis (in some cases), and disseminated histoplasmosis.

Other leucocyte changes are less common in acute infection. Monocytosis has been reported in patients with typhoid fever and sometimes in brucellosis or subacute bacterial endocarditis. In endocarditis a monocytosis may be associated with the presence of undifferentiated reticuloendothelial cells in the blood that show erythrophagocytosis.

Some degree of anaemia is found almost invariably in patients with bacterial infection. It usually presents a picture of the anaemia of chronic disorders. Haemolytic anaemia may occur in severe septicaemias and is usually associated with disseminated intravascular coagulation. Some organisms, *Clostridium welchii* for example, produce an a-toxin that acts as a lecithinase and causes fulminating intravascular haemolysis.

Disseminated intravascular coagulation is a common accompaniment of severe bacterial infection. Many mechanisms have been suggested, including vascular injury with activation of factor XII or the generation of procoagulants from white cells by the action of endotoxin. Thrombocytopenia is also common in patients with septicaemia. Although this may sometimes reflect disseminated intravascular coagulation, the mechanism is probably more complicated. There may be quite dramatic thrombocytopenia without any other evidence of a consumption coagulopathy. Several mechanisms are involved, including suppression of platelet production by the bone marrow, damage to circulating platelets by immune complexes, endothelial damage, and direct interaction of the platelets with bacteria; phagocytosis of bacteria by platelets may provoke the rapid disappearance of platelets from the circulation.

Chronic bacterial infection

Chronic bacterial infection is usually associated with the anaemia of chronic disorders. Some particularly interesting haematological changes are sometimes ascribed to tuberculosis (Table 3). While the most common change is a mild, normochromic, normocytic anaemia with a raised ESR; more spectacular blood changes have been reported, particularly in association with disseminated tuberculosis. These clinical pictures include leukaemoid reactions, pancytopenia, myelofibrosis, and even polycythaemia. The main problem in assessing these associations is whether the reported patients had infections due to atypical mycobacteria superimposed on an underlying blood disease, or whether disseminated tuberculosis can occasionally produce a clinical picture similar to leukaemia or a myeloproliferative disease. In practice, any patient who presents with an atypical myeloproliferative disorder, and who is going downhill for no apparent cause, should be investigated for tuberculosis. Attempts should be made to grow the organism from bone marrow cultures.

Virus infections

Haematological changes occur quite commonly in association with many virus illnesses. Changes associated with specific viral infections such as infectious mononucleosis are considered in Section 7.

Many virus infections are associated with a modest neutropenia and often with a relative or absolute lymphocytosis. Atypical lymphocytes are characteristic of patients with infectious mononucleosis but they may also be found in association with many other virus infections.

Rubella, acquired in childhood or adult life, is often associated with a leucocytosis and an atypical lymphocytosis. A small proportion of patients develop an acute fulminating thrombocytopenic purpura approximately 4 days after the appearance of the rash. It is usually self-limiting but fatalities have been reported. Thrombocytopenia is also common in infants with congenital rubella. This condition is also characterized by a non-immune haemolytic episode shortly after birth. Thrombocytopenia has also been reported in association with measles. In particularly severe forms of rubella and morbilli, severe haemorrhagic states due to disseminated intravascular coagulation have been seen. Similar changes occur occasionally in patients with varicella infections.

Haematological changes very similar to those seen in infectious mononucleosis can occur in patients with cytomegalovirus (CMV) infection. Infants may exhibit hepatosplenomegaly with purpura and anaemia. The anaemia is characterized by a haemolytic picture with many normoblasts in the peripheral blood. This form of anaemia may last for several weeks and may be associated with severe thrombocytopenia. There are well-documented cases of an infectious mononucleosis-like disorder occurring after transfusion with fresh blood or after perfusion for open heart surgery. This self-limiting syndrome usually occurs 1 to 3 months after blood transfusion and resolves within a few weeks. It is characterized by a moderate rise in temperature, with hepatosplenomegaly, lymphadenopathy, and transient maculopapular rashes, and a lymphocytosis indistinguishable from that of infectious mononucleosis.

Haematological problems are common in patients with AIDS. Lymphopenia is particularly common; neutropenia occurs in 0 to 30 per cent of HIV antibody-positive asymptomatic individuals, and in 20 to 65 per cent of patients with AIDS. Thrombocytopenia occurs in 5 to 20 per cent of asymptomatic HIV-1 infected persons and rises to 25 to 50 per cent in patients with AIDS. There is also a strong association of thrombotic thrombocytopenic purpura with HIV infection. Anaemia is also common. Bone marrow examination often reveals dyserythropoiesis with a variable degree of erythrophagocytosis. There have been a number of reports of the presence of lupus anticoagulant in the blood of patients with AIDS. In addition to these haematological complications, there is the added risk of drug-induced marrow hypoplasia, associated particularly with treatment with zidovudine (AZT).

Haematological complications of infectious hepatitis are rare but can be extremely severe. Coombs' positive haemolytic anaemia has been reported. There is also considerable literature on the occurrence of aplastic anaemia. This disorder seems predominantly to affect young males; the onset of aplasia is usually about 9 weeks after the onset of hepatitis. The condition is associated with a mortality in excess of 90 per cent. In those patients who recover, the period to complete haematological normality ranges between 3 and 20 months.

Many viruses are capable of provoking severe bleeding due to intravascular coagulation. Why viruses can fire off the coagulation cascade is far from clear. Activation of factor XII due to vascular injury or damage to platelets with the release of coagulants have been suggested as possible mechanisms.

The human parvovirus has a particular affinity for red-cell progenitors. It probably causes transient red-cell aplasia quite commonly but this only gives rise to a symptomatic anaemia in patients who have a markedly shortened red-cell survival. Thus parvovirus infection appears to be responsible for the aplastic crises in patients with sickle-cell anaemia, pyruvate kinase deficiency, or other congenital haemolytic anaemias. Viruses can cause acute damage to the bone marrow in immunesuppressed patients as part of the virus haemophagocytic syndrome.

The haematological changes associated with the virus haemorrhagic fevers are described in detail in Section 7.

Parasitic disease

The major haematological accompaniments of the parasitic diseases are described in Section 7. Those which produce important haematological changes will be briefly summarized here.

Toxoplasmosis

Congenital toxoplasmosis can produce a condition identical to erythroblastosis fetalis. The clinical picture is of a pale, hydropic infant with a large spleen and liver

associated with severe anaemia, thrombocytopenia, and a leucocytosis, often with a marked eosinophilia. In adult life the acquired forms of toxoplasmosis produce a clinical disorder resembling infectious mononucleosis.

Malaria (Plate 1)

Malarial infection produces a variety of haematological abnormalities. The most severe changes occur during *Plasmodium falciparum* infection. In acute infections in non-immune individuals, there is usually minimal anaemia at the onset of the illness. During the 2 to 3 weeks after treatment there may be a steady decline in haemoglobin level. On the other hand, individuals with chronic malaria, some degree of immunity, and low-level parasitaemias, may be severely anaemic at presentation with an inappropriately low reticulocyte count. The bone marrow is often hyperplastic and shows a marked degree of dyshaemopoiesis (Fig. 3).

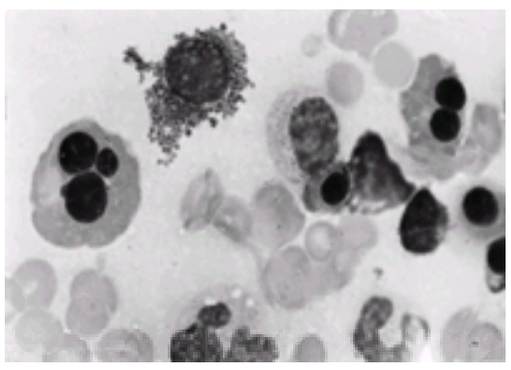

Fig. 3 Bone marrow appearances in *P. falciparum* malaria. There is marked dyserythropoiesis with several multinucleate red-cell precursors (Giemsa stain ×800).

The precise details of the pathophysiology of the anaemia of malaria are still unclear. There is no doubt that in acute attacks there may be massive destruction of parasitized red cells but, curiously, there is strong evidence that the survival of non-parasitized cells is also shortened. Although there is some indirect evidence for an immune basis for red cell destruction, particularly in children, little solid evidence exists for immune destruction in non-immune adults. There is growing evidence that the lack of marrow response may be, at least in part, due to the high levels of tumour necrosis factor (TNF) that are produced during malarial infection. TNF suppresses proliferation of erythroid progenitor cells *in vitro*. Chronic malaria has several features that augment this effect. Large numbers of pigment particles are ingested by the resident macrophages of the spleen and marrow, providing the possibility of a sustained stimulus for TNF production at the site of erythropoiesis.

In some patients with severe *P. falciparum* infections, there may be marked intravascular haemolysis and haemoglobinuria. The mechanism is not certain. Some of these patients may be glucose 6-phosphate dehydrogenase deficient but this is by no means the whole story. It has been suggested that some patients with fulminating malaria have disseminated intravascular coagulation. This is probably uncommon and plays very little part in the pathophysiology of either the anaemia or haemorrhagic phenomena that occur. Thrombocytopenia is extremely common but is only rarely associated with evidence of consumption of blood-clotting factors. In most forms of malarial infection there is a neutropenia. Monocytosis has also been described.

Several interesting haematological manifestations are associated with unusual forms of malaria. In the tropical splenomegaly syndrome, there may be anaemia, thrombocytopenia, and neutropenia, all secondary to hypersplenism. Congenital malaria infection is contracted in intrauterine life from the mother; newborn babies have a febrile illness associated with profound anaemia that appears to result from the combination of haemolysis and bone marrow suppression.

Leishmaniasis

Visceral leishmaniasis, or kala azar, is associated with hepatosplenomegaly, lymphadenopathy, and a pancytopenia, particularly in young children. Early in the course of the disease there is often marked neutropenia. The marrow may be grossly infiltrated with parasitized macrophages. The anaemia is due mainly to a short red-cell survival; there is also an inappropriate marrow response and a variable degree of hypersplenism.

Hookworm

The haematological changes of hookworm infestation are described in Chapter 22.5.3. It is one of the most common causes of iron-deficiency anaemia in the world population. During the systemic phase of the illness, when the larvae invade the lungs, there may be a marked eosinophilia. During this phase the bone marrow shows a remarkable increase in the percentage of eosinophilic myelocytes, which may be out of proportion to the eosinophilia observed in the peripheral blood.

Visceral larva migrans

This condition is characterized by striking haematological changes including anaemia, a marked leucocytosis with eosinophilia, and changes in the titre of anti-A and anti-B blood-group antibodies.

Schistosomiasis

In the chronic phase of *S. mansoni* and *S. japonicum* infections there may be severe portal hypertension, splenomegaly, and the typical picture of hypersplenism.

Other trematode infestations, including clonorchiasis and paragonamiasis, are associated with eosinophilia and anaemia. Antibodies to the P_1 blood-group antigen may be found in grossly elevated titres in the blood of many patients with acute fascioliasis.

Rheumatoid arthritis and related disorders

In patients with rheumatoid arthritis, anaemia is extremely common. It usually follows the general pattern of anaemia of chronic disorders. It is occasionally complicated by genuine iron deficiency, which may result from a variety of causes including poor diet and chronic blood loss due to the effects of treatment, particularly ingestion of salicylates and non-steroidal anti-inflammatory agents or corticosteroids. Furthermore, significant bleeding into actively inflamed joints can occur. It has been estimated that if only two knee joints were affected, the annual blood loss through this mechanism could amount to as much as 2500 ml. It is not certain how much of the iron derived from this blood is available for reutilization for haemoglobin synthesis. The diagnosis of iron deficiency complicating rheumatoid arthritis may not be straightforward; levels of serum iron and iron-binding capacity may be difficult to interpret because of coexisting inflammation. Determination of marrow stores and estimation of serum ferritin may be more helpful. Although the last two are elevated in inflammatory conditions, a low level suggests genuine iron deficiency.

There are no particular changes in the neutrophil response in uncomplicated rheumatoid arthritis; a marked leucocytosis may reflect a response to corticosteroid therapy or a superadded infection such as a septic arthritis. The platelet count is elevated in between 20 and 50 per cent of patients with rheumatoid arthritis. The degree of thrombocytosis parallels the degree of activity of the illness and cannot be accounted for on the grounds of associated intestinal blood loss due to drug therapy.

The haematological changes of Felty's syndrome are summarized in Section 18. There is anaemia, thrombocytopenia, and marked neutropenia. Although many of these changes are features of hypersplenism, recent studies on the neutropenia in this disorder indicate that it has a complex basis. Immune destruction of neutrophils may play a major part.

A variety of haematological changes are due to drug therapy for rheumatoid arthritis and related disorders. Salicylates may produce chronic blood loss, while drugs containing phenacetin produce methaemoglobinaemia and Heinz-body haemolytic anaemia that may sometimes be preceded by a marked eosinophilia.

Phenylbutazone produces pancytopenia, which may be severe and irreversible; this drug has now been discontinued in the United Kingdom. Oxyphenylbutazone and penicillamine may also cause severe marrow depression. The administration of gold occasionally causes marked thrombocytopenia or pancytopenia.

The management of the haematological manifestations of rheumatoid arthritis and Felty's syndrome is unsatisfactory. The anaemia generally reflects the activity of the disease. If there is genuine iron deficiency, iron replacement therapy is indicated. The vexed question of whether intramuscular iron administration has some non-specific effect on the anaemia of rheumatoid arthritis, even in the absence of reduced body iron stores, remains unresolved. Similarly, there is controversy about the best way to manage Felty's syndrome. After splenectomy there is sometimes a dramatic rise in the neutrophil and total leucocyte counts, but this is not always associated with a decreased incidence of infection. Some patients show no change in the white-cell count after surgery. It is difficult to advise about the best approach to the management of this condition; only if there are recurrent, life-threatening infections should splenectomy be done. Patients require extremely careful surveillance after the operation. There may be some place for the use of prophylactic antibiotics in those whose neutrophil counts do not respond.

Recent studies have suggested that the anaemia of rheumatoid arthritis, and related inflammatory states, may respond to erythropoietin given in the higher therapeutic dose range. This treatment is extremely expensive and has only been evaluated in a few clinical trials. Its use should be reserved for those patients who have severe anaemia which is refractory to treatment of the underlying inflammatory disorder by any other means.

Systemic lupus erythematosus and other collagen disorders

It is quite common for systemic lupus erythematosus (SLE) to present with a haematological disorder. This is not the case in the other collagen–vascular disorders.

The most common blood change in SLE is anaemia, which occurs in nearly all patients at some stage of the illness. It is usually a mild anaemia of chronic disorders, which may be complicated by blood loss from analgesics or anti-inflammatory medication, renal impairment, or haemolysis. Acquired autoimmune haemolytic anaemia may be the sole presenting feature in SLE and may antedate the appearance of other typical features by many years. The incidence of this complication varies in reported series but occurs overall in approximately 5 per cent of cases. The Coombs' test is invariably positive with anticomplementary reagents and is positive with anti-IgG during episodes of acute haemolysis. Other forms of anaemia in SLE include those associated with hypersplenism due to splenomegaly, and the occasional occurrence of a hypocellular bone marrow, probably due to involvement of small vessels by the disease process.

The most consistent finding in the white-cell count in SLE is leukopenia, which occurs in up to half the patients at some time during the illness. This is often a combined neutropenia and lymphopenia. Mild eosinophilia occurs occasionally, particularly in association with skin involvement.

A mild thrombocytopenia occurs in 10 to 25 per cent of all cases of SLE. More severe thrombocytopenia, producing a picture almost indistinguishable from idiopathic thrombocytopenic purpura, occurs in a small proportion of patients and may be the sole presenting feature in some. Although early reports indicated that splenectomy might be associated with a flare-up of the systemic symptoms of SLE in patients with thrombocytopenia, this has now been shown to be incorrect.

Lupus anticoagulant

This is an antibody that prolongs phospholipid-dependent coagulation tests *in vitro*. Although it received its name because it was found in patients with SLE, it occurs more frequently in patients without this disease and is associated with thrombosis rather than with bleeding. It is particularly common in patients with lupus-like autoimmune disorders without the associated criteria for the diagnosis of SLE. Originally it was thought to occur in approximately 10 per cent of patients but using more sensitive assays it is now clear that it occurs in about 50 per cent.

Both lupus anticoagulants and associated anticardiolipin antibodies are immunoglobulins which react with phospholipid and other molecules (platelet factor IV). They may be associated with venous thromboembolism, arterial thromboembolism, an increased fate of fetal loss, or thrombocytopenia. They are discussed in detail in Section 18.11.

Other collagen disorders

The haematological changes in the other collagen–vascular diseases are much less impressive. They are all associated with the anaemia of chronic disorders. Polyarteritis nodosa may be characterized by an eosinophilia.

The interesting syndrome of polymyalgia rheumatica and temporal arteritis may present to the haematologist (Section 18.11). Haematological changes are characterized by a severe anaemia of chronic disorders with a marked elevation of the ESR. The leucocyte count is usually normal, although there may occasionally be a mild eosinophilia. There is a marked increase in the a_2- and g-globulins, although this is polyclonal in type. This blood picture can very closely resemble that of multiple myeloma or disseminated malignancy.

Renal disease

Almost all forms of renal disease are associated with haematological changes. However, by far the most important is the severe refractory anaemia that accompanies chronic renal failure.

Anaemia

Anaemia is an important and intractable complication of chronic renal failure. The correlation between the blood urea nitrogen and the haemoglobin level is inconsistent. Although erythropoietin deficiency is an important component, the anaemia has an extremely complex aetiology, which is only partly understood. The red cells of patients with chronic renal disease have a shortened survival, although they survive normally when injected into healthy recipients. Similarly, normal red cells have a shortened survival in uraemic recipients. The nature of the intracorpuscular defect has not been determined. Most red-cell enzymes are present at normal levels and the intracellular level of ATP is elevated. However, changes in membrane function have been demonstrated, in particular decreased activity of the Na^{+}–K^{+} pumps; the toxic substances that cause these changes have not been identified.

There is also impaired red-cell production in the anaemia of chronic renal failure. The fact that the anaemia of chronic renal failure can be corrected by the administration of recombinant erythropoietin suggests that the ineffective production of this hormone due to renal damage is the major aetiological factor in the anaemia of renal failure. However, it has been found that the serum from patients on haemodialysis also inhibits the proliferation of erythroid progenitors. The suppressive activity is found in serum fractions containing material of molecular weights ranging from 47 000 to above 150 000. Interestingly, patients on continuous ambulatory peritoneal dialysis (CAPD) have higher haemoglobin levels than those on haemodialysis. It is possible this reflects the more effective removal of middle molecular-weight molecules of this type by CAPD. Patients on haemodialysis with low haemoglobin concentrations are more likely to have fibrous replacement of their bone marrow. This has been correlated with secondary hyperparathyroidism, suggesting a role for parathyroid hormone in the bone marrow unresponsiveness and fibrosis (see Section 20).

The anaemia of chronic renal failure may be exacerbated by deficiency of iron resulting from blood loss due to excessive blood sampling, incorrect haemodialysis procedures, or bleeding due to defective platelet function (see below). A small proportion of patients with chronic renal failure develop splenomegaly and hypersplenism. Folate deficiency is found occasionally in patients on haemodialysis. There have been a few reports of nephrosis leading to severe urinary loss of transferrin and hence to a low plasma iron-binding capacity. Some patients with renal disease have chronic inflammatory lesions, which may lead to a superadded anaemia of chronic disorders.

The type of renal lesion is also an important factor in determining the severity of anaemia. For example the renal failure of polycystic disease of the kidneys is associated with a relatively higher haemoglobin level than other forms of renal failure. Interestingly, the shrunken kidneys of some patients on long-term dialysis programmes develop cysts and this phenomenon is also associated with a rise in haemoglobin level. It seems likely that both these conditions are associated with a relative increase in the output of erythropoietin.

The anaemia of chronic renal failure is normochromic and normocytic unless there is associated iron deficiency. The red cells show characteristic deformities with multiple tiny spicules and contracted poikilocytes. The capacity of the red cells for oxygen transport does not seem to be impaired. There is often an increased

intracellular concentration of 2,3-diphosphoglycerate (2,3-DPG) in response to anaemia and hyperphosphataemia, and the oxygen affinity of haemoglobin is decreased. This right shift in the oxygen dissociation curve may be augmented by uraemic acidosis. However, part of the advantage of the acidosis is cancelled out by the direct effect of low pH on glycolysis and 2,3-DPG production. Intensive dialysis may cause a reduction in the concentration of intracellular phosphate, which has the effect of increasing the oxygen affinity of haemoglobin. This effect may play a part in the so-called dialysis disequilibrium syndrome.

In patients with chronic renal failure who have associated iron deficiency, the red-cell indices are typical of this condition; the reduced mean corpuscle haemoglobin and volume are corrected by iron therapy.

The bone marrow in chronic renal failure shows normoblastic erythropoiesis but the degree of erythroid hyperplasia is not compatible with the degree of anaemia, indicating suppression of erythropoiesis.

White cells

The total and differential white-cell count is usually normal in patients with chronic renal failure. However, the phagocytic activity of granulocytes may be reduced and complement activation by haemodialysis membranes may cause stasis of white cells in the pulmonary circulation with temporary granulocytopenia. Cell-mediated immunity is also depressed.

Platelets and coagulation

There is a variety of haemostatic defects in different forms of renal disease. Most forms are associated with a bleeding tendency, which is seen in its most florid form in acute renal failure. The main features are purpura, and mucosal and gastrointestinal bleeding associated with abnormal platelet function and a prolonged bleeding time; these changes are reversible by dialysis. Various mechanisms have been proposed. These include a direct action of metabolites on platelet function and a disturbance of prostaglandin balance because of a deficiency of a renal factor that modifies vascular production of prostacyclin and/or platelet endoperoxide and thromboxane synthesis. These changes result in an abnormality of the control of platelet cAMP causing the platelets to become refractory to aggregation agents. Many conditions that lead to renal failure are also associated with thrombocytopenia. For example the circulating immune complexes found in patients with acute glomerulonephritis, polyarteritis nodosa, or lupus nephritis may be responsible for platelet activation and the release of aggregating agents. Thrombocytopenia may also be aggravated by heparin therapy or the use of immunosuppressant drugs in patients who have received kidney grafts. Mild thrombocytopenia is well recognized in patients with functioning renal allografts. This has also been found to be associated with an inability to clear the immune complexes. Graft rejection is associated with enhanced platelet aggregation and thrombocytopenia.

The nephrotic syndrome is characterized by a marked tendency to thrombosis. This also has a complex pathogenesis. Both platelet aggregation and release reactions have been shown to be enhanced in this condition. Protein loss in the urine may also play a part. It has been found that an increased loss of antithrombin III is related to thrombotic episodes. Conversely, coagulation factors IX and XIII are also lost in the urine of patients with a nephrotic syndrome; the deficiency of factor IX may be sufficient to induce bleeding.

The haematological changes associated with the haemolytic uraemic syndrome and thrombotic thrombocytopenic purpura were considered earlier in this section.

Polycythaemia

The polycythaemias associated with renal lesions and following renal transplantation are discussed elsewhere.

Treatment of the haematological complications of renal disease

The management of the anaemia of chronic renal failure, which has been revolutionized by the availability of recombinant erythropoietin, is considered in Section 20. The management of bleeding in patients with acute renal failure is based on correction of uraemia by dialysis and appropriate replacement therapy. Peritoneal dialysis is probably more effective in reversing abnormalities of platelet function, although there is no definite evidence that one form of dialysis is superior to another. If there is severe thrombocytopenia, platelet transfusions should be given.

Gastrointestinal and liver disease

Many of the haematological changes that occur in gastrointestinal and liver disease are described in Section 14. Here we will simply summarize the haematological manifestations of those disorders that present frequently with anaemia or defective haemostasis.

Gastrointestinal blood loss

Blood loss in excess of 20 ml/day will always result in a negative iron balance and ultimately in iron-deficiency anaemia, the time taken depending on the body stores of iron when the bleeding started.

The haematological picture shows the typical changes of iron-deficiency anaemia, with hypochromic, microcytic red-cell morphology. Occasionally, there are some clues that this blood picture is associated with chronic blood loss. Quite frequently there is a mild to moderate thrombocytosis, and if iron is being taken there may be a dimorphic blood picture (Fig. 4), red-cell polychromasia, and a low-grade reticulocytosis. It is always worth examining the peripheral blood film very carefully as it may give some clue as to the site of the blood loss. For example the presence of target cells may indicate liver disease, whereas the presence of distorted cells and Howell–Jolly bodies suggests malabsorption due to adult coeliac disease complicated by hyposplenism.

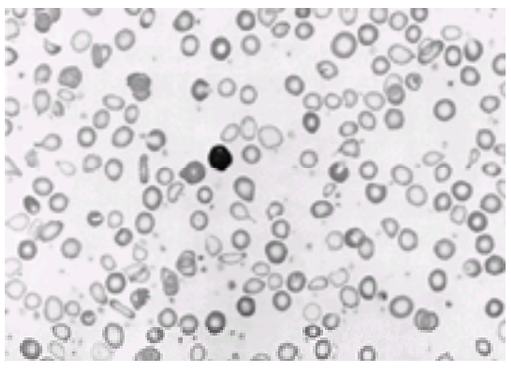

Fig. 4 Peripheral blood picture associated with gastrointestinal bleeding. The red cells show a dimorphic picture with hypochromic and normochromic forms. The platelet count is elevated, a typical finding in bleeding (Giemsa stain × 600).

The diagnosis of the site of acute upper intestinal bleeding is considered in Section 14. The investigation of chronic gastrointestinal blood loss may be difficult. First, it is essential to determine whether iron deficiency anaemia is due to a defective intake or due to excessive loss of iron. If gastrointestinal blood loss is suspected the first step is to confirm that this is occurring, by examination of several stool specimens for occult blood. Currently, the most commonly used method is the Haemoccult card which contains a filter paper impregnated with guaiac, or a similar commercial kit. The peroxidase activity of red blood cells releases a free oxygen radical from hydrogen peroxide (the developer) which then reacts with the guaiac to produce a blue colour. This test is simple, easy to perform following a rectal examination, and is quick. Nevertheless, it is relatively insensitive, since blood loss must exceed 20 ml daily for 80 to 90 per cent of tests to be positive (the normal loss in a healthy individual is 0.5–1.2 ml daily as measured by ^{51}Cr-labelled red cells). False positive results may result from peroxidase or non-specific oxidants in the diet. Thus in many screening programmes, subjects are requested to omit red meat, fresh fruit, cauliflower, swede, turnip, tomatoes, horseradish, and vitamin C supplements from

their diet during the 3 days prior to testing. Non-steroidal anti-inflammatory drugs and aspirin may also give positive results but, since the increased blood loss is in the upper gastrointestinal tract, the haemoglobin is metabolized in the small intestine and is therefore not detected in stools by Haemoccult unless blood loss is considerable. Iron therapy does not affect guaiac-based tests. Many newer tests, which are said to be even more sensitive, are being developed but their role in clinical practice is not yet established.

Once having established that there is gastrointestinal blood loss the next step is to determine the site. This requires a detailed history and clinical examination as outlined in the introduction to this section. The next step is a careful endoscopy, followed by sigmoidoscopy and colonoscopy. If these investigations do not provide a diagnosis, and there is persistent bleeding, the duodenum and small bowel should be studied radiologically. Occasionally it is necessary to resort to coeliac or superior mesenteric angiography, which may be useful for showing duodenal or ileal varices, bleeding from Meckel's diverticulum or non-specific ulcers of the ileum, small bowel tumours, and vascular lesions. However, small lesions can only be visualized if there is active bleeding at the time of the examination, probably at a rate of at least 0.5 ml/min.

Inflammatory diseases of the bowel

A mild anaemia of chronic disorders is a common accompaniment of inflammatory disease of the ileum, caecum, and colon. It is observed frequently in patients with Crohn's disease, ileocaecal tuberculosis, ulcerative colitis, and other forms of proctocolitis. In many of these conditions the anaemia of chronic disorders is complicated by intermittent blood loss or dietetic iron deficiency. In some cases of extensive Crohn's disease there may be an added factor of malabsorption. Anaemia occurs in about one-third of these patients and it may be complicated by reduced vitamin B_{12} or folic acid absorption. In one large survey of patients with Crohn's disease, anaemia was present in 79 per cent of the males and 54 per cent of females. Forty-six out of a total of 63 patients had bone marrow biopsies, and of these 39 per cent were megaloblastic. Of this group, 11 were folate deficient, six vitamin B_{12} deficient, and one had both deficiencies. On the other hand, macrocytic anaemia is unusual in patients with ulcerative colitis and the anaemia is usually hypochromic due to blood loss. Interestingly, there have been occasional reports of autoimmune haemolytic anaemia occurring in association with ulcerative colitis; in several cases the autoantibodies showed rhesus specificity.

The anaemia of intestinal inflammatory disease may be made worse by drugs used in its management. Patients who receive salazopyrine for colitis occasionally develop an acute haemolytic anaemia associated with Heinz-body formation. Bone marrow depression may occur in patients receiving immunosuppressive treatment for colitis or Crohn's disease. Ileocaecal tuberculosis may be associated with any of the bizarre haematological manifestations of tuberculosis described above, and it may be complicated by the side-effects of antituberculous drug therapy.

Whipple's disease may produce a clinical picture and blood changes that can mimic several primary haematological disorders. The typical clinical triad of diarrhoea, arthropathy, and enlarged lymph nodes is usually associated with a mild normochromic, normocytic anaemia, a raised ESR, and a polymorphonuclear leucocytosis. Quite often there is associated lymphopenia or eosinophilia. Some cases present less typically. When the spleen is enlarged the condition may closely mimic a primary reticulosis. Malabsorption of vitamin B_{12} or folic acid may occasionally be encountered in this disorder (see Chapter 14.9.6).

Structural disease of the stomach, and small and large bowel

The structural changes and resulting abnormalities of absorption associated with gastritis are described in detail in Section 14. Similarly, the various anatomical abnormalities of the small-gut and malabsorption syndromes that lead to vitamin B_{12} and folate deficiency are reviewed earlier in this section. The relation between gastric surgery and iron and vitamin B_{12} metabolism is discussed in Section 14.

Most anatomical lesions of the small bowel present to the haematologist as a macrocytic anaemia with a megaloblastic bone marrow due to vitamin B_{12} or folate deficiency or as a refractory iron-deficiency anaemia. Several abnormalities of the small gut are associated with the production of a relatively profuse bacterial flora that utilize vitamin B_{12}. These conditions include surgically produced blind loops, strictures, anastomoses between loops of small bowel, fistulae between various sections of the bowel, diverticula of the small bowel, malfunctioning gastroenterostomies, interference of gut motility in conditions such as scleroderma, Whipple's disease, postvagotomy, and after extensive gut resection, where the disorder may also produce malabsorption. All these conditions are associated with defective vitamin B_{12} absorption, which can be partly corrected by the administration of broad-spectrum antibiotics but not by intrinsic factor.

Megaloblastic anaemia due to intestinal malabsorption is fully reviewed in Section 14.9. It should be remembered, however, that the malabsorption syndromes may present to the haematologist in other ways. For example there is a very high incidence of iron-deficiency anaemia in this group and, particularly in childhood, this is the much the more common form of presentation than a megaloblastic anaemia. The peripheral blood changes of hyposplenism are quite frequently associated with an underlying malabsorption syndrome, which itself may also present with a bleeding disorder due to defective absorption of vitamin K. Patients with malabsorption syndrome frequently have biochemical evidence of vitamin E deficiency; although this may produce a slightly shortened red-cell survival, there is no evidence that vitamin E deficiency alone produces a significant degree of anaemia.

Liver disease

There is usually a moderate degree of anaemia in patients with chronic liver failure (Table 4). The red cells are normochromic or slightly macrocytic with mean corpuscle volume values ranging from 100 to 115 fl. Target cells and a variable degree of polychromasia with a slightly elevated reticulocyte count are often found. The degree of macrocytosis and target-cell formation corresponds reasonably well with the degree of liver failure. The bone marrow tends to be hypercellular with erythroid hyperplasia and macronormoblastic changes.

The actual mechanism of the anaemia of liver failure is uncertain. However, there may be many complicating factors that cause a worsening of the anaemia in this condition. Nutritional folate deficiency is very common in patients with liver disease, particularly the alcoholic form. Secondary iron deficiency is also common and usually results from chronic intestinal blood loss associated with a poor dietetic intake. Interestingly, in patients with severe portal hypertension and cirrhosis, or in those who have undergone portacaval shunt surgery, there may be some increase in iron absorption with marked haemosiderosis of the liver.

A variety of different forms of haemolytic anaemia occur in patients with liver disease. In Zieve's syndrome there is jaundice, hyperlipidaemia, and haemolytic anaemia that follows an excessive alcohol intake. Other forms of haemolytic anaemia may occur. Acute haemolysis has been well documented in patients with viral hepatitis, particularly those who are glucose 6-phosphate dehydrogenase deficient. An acquired haemolytic anaemia with a positive Coombs' test may occur occasionally in patients with chronic active hepatitis. Another form of haemolytic anaemia in liver disease, usually alcoholic cirrhosis, has been observed in which there are marked red-cell abnormalities with burr and spur-shaped forms predominating.

Bleeding and haemostatic failure are extremely common accompaniments of liver failure. They have a complex aetiology including diminished hepatic synthesis of coagulation factors V, VII, IX, X, and XI, prothrombin and fibrinogen. In some forms of liver disease there may be malabsorption of vitamin K with reduction in the K-deficient clotting factors, and reproduction of a dysfunctional form of fibrinogen has been reported in some patients with cirrhosis and hepatocellular carcinoma. In some forms of liver failure there is enhanced fibrinolysis, due to decrease synthesis of a $_2$-plasmin inhibitor. In severe liver failure disseminated intravascular configuration may occur. Thrombocytopenia is extremely common in liver disease, sometimes due to hypersplenism but in other cases its pathogenesis is not clear.

The management of bleeding in liver diseases is considered in Section 14.21.

The haematological effects of alcohol

Because excessive consumption of alcohol is so common, it is important for clinicians to appreciate the remarkably diverse haematological manifestations that it causes.

Anaemia is particularly common in chronic alcoholics. It has an extremely complex aetiology including a deficient diet, chronic blood loss, hepatic dysfunction, and the direct toxic effects of alcohol on the bone marrow.

Macrocytosis is particularly common in chronic alcoholics. An unexplained macrocytic blood picture should always raise the possibility of alcoholism, although its absence does not rule out the diagnosis. It may be associated with normoblastic or megaloblastic erythropoiesis. In moderately severe alcoholics who are maintaining

a reasonable diet, it probably reflects the direct toxic action of alcohol on the bone marrow. The normoblasts may show vacuolation or there may be no specific changes on light microscopy. Megaloblastic anaemia is usually seen in severe alcoholics who are poorly nourished, and is due to folate deficiency. While a folate-poor diet is the major factor, there is some evidence that alcohol plays a more direct part in interfering with folate metabolism by an unknown mechanism. It should be remembered that macrocytosis can also occur in alcoholics during a reticulocytosis in response to bleeding or alcohol withdrawal. It may also reflect coexistent liver disease. The occurrence of sideroblastic anaemia in severe alcoholics was mentioned in an earlier chapter. It is often associated with a macrocytosis or a dimorphic blood picture and occurs in severe alcoholics. The sideroblastic changes revert to normal after stopping alcohol.

Simple iron deficiency is also found commonly in alcoholics and probably reflects both a poor diet and chronic blood loss due to gastritis or bleeding varices. It may be associated with folate deficiency; the blood film is then dimorphic with macrocytes, microcytes, and hypersegmented neutrophils. Alcoholics with chronic pancreatitis may develop iron loading due to increased absorption. These changes, which are specific for alcohol, may be accompanied by any of the haematological manifestations of liver disease.

Alcohol has deleterious effects on the white cells. Severe alcoholics are prone to infection. The neutropenia of alcoholism may reflect both the toxic effect of alcohol on the marrow and folate deficiency. There is also some evidence that alcohol can interfere with neutrophil locomotion and with their ability to ingest foreign material including micro-organisms.

Thrombocytopenia is commonly seen in chronic alcoholics and may occur without accompanying folate deficiency or splenomegaly. Megakaryocytes may be normal or diminished in number. Following withdrawal of alcohol the platelet count usually returns to normal, although it may become markedly elevated for a few days.

Chest disease

(See also carcinoma and tuberculosis, above, and secondary polycythaemia, Chapter 22.3.8).

Pneumonia

Most bacterial pneumonias are associated with a neutrophil leucocytosis. Two relatively common forms of pneumonia are associated with more specific haematological changes. In mycoplasma pneumonia, cold agglutinins can usually be detected in increased amounts towards the end of the first week in up to 80 per cent of cases. The cold antibodies are polyclonal IgM to the red-cell I antigen. Although a positive Coombs' test and an increased reticulocyte count have been described in these cases, serious haemolysis is rare. Occasionally, the condition is complicated by disseminated intravascular coagulation.

There is increasing evidence that in patients with pneumonia caused by *Legionella pneumophila* (Legionnaires' disease) there may be severe thrombocytopenia and, sometimes, lymphopenia. Several cases have been reported to be complicated by disseminated intravascular coagulation.

Pulmonary eosinophilia (see also Section 17.11)

This term refers to a group of disorders that have in common a raised eosinophil count in the peripheral blood in association with pulmonary infiltrates on the chest radiograph. The exact nature of many of the disorders that constitute this syndrome is uncertain. In its simplest form there may be a brief period of respiratory distress in association with eosinophilia. This condition is sometimes called Löffler's syndrome. At the other end of the spectrum there is a severe illness associated with widespread pulmonary infiltrates and eosinophilia, which may culminate with the features of polyarteritis nodosa.

The transient disorder described by Löffler probably represents a heterogeneous group of conditions, which in many cases are associated with parasitic infection. Many parasitic disorders can cause this type of illness, including ascariasis, ankylostomiasis, trichiuriasis, taeniasis, and fascioliasis. A similar condition has been well documented as part of a hypersensitivity reaction to drugs. The most common is *p*-aminosalicylic acid but similar reactions have been observed in patients receiving penicillin, sulphonamides, and nitrofurantoin. A similar clinical picture is associated with the syndrome of allergic alveolitis, including farmer's lung, bird fancier's lung, and a variety of other occupational disorders. Another condition characterized by a marked eosinophilia with pulmonary infiltrates goes under the general term tropical eosinophilia. There is considerable evidence that this disorder is due to occult filarial infection. Pulmonary eosinophilia may also be due to hypersensitivity to fungi, particularly *Aspergillus fumigatus*.

Idiopathic pulmonary haemosiderosis and Goodpasture's syndrome

These disorders occasionally present as a refractory anaemia that has the characteristics of the anaemia of chronic disorders, although it may become markedly hypochromic and microcytic due to chronic blood loss.

Skin diseases

Megaloblastic anaemia and the skin

The whole relation between skin disease and megaloblastic anaemia is extremely complex and much of the work in this field is still controversial. The subject is discussed elsewhere (see Section 23).

A proportion of patients with various dermatoses show evidence of folate depletion, at least biochemically, and in some cases, haematologically. This has been reported in patients with erythroderma, psoriasis, or extensive eczema. There is a well-documented association between malabsorption and dermatitis herpetiformis. Although megaloblastic anaemia is not found frequently in association with disorders of the skin, some patients with these conditions do have mild megaloblastic changes. Earlier reports suggested that a significant proportion of them had abnormalities of small-intestinal function and structure, leading to the descriptive term 'dermatogenic enteropathy'. This concept has been questioned and it is now agreed that a completely flat small-bowel mucosa is rarely seen in these conditions. The relation between dermatitis herpetiformis and malabsorption of the coeliac type seems to be a special case. Several series have shown a high incidence of small-bowel changes of coeliac disease in patients with this condition. Furthermore, there appears to be a high incidence of splenic hypoplasia with typical haematological changes of defective function of the spleen (see Chapter 22.4.4).

Other dermatological disorders

Several dermatological diseases have a major haematological component. Of particular importance are the systemic mast-cell syndromes, Sezary syndrome and cutaneous T-cell lymphomas, hereditary telangiectasia, and some of the inherited disorders of collagen.

Endocrine disease

Pituitary deficiency

A mild, normochromic, normocytic anaemia is very common in patients with anterior pituitary deficiency. The mechanism is not absolutely clear, although the anaemia has many features in common with that of hypothyroidism and is fully responsive to appropriate replacement therapy.

Thyroid disease

Hypothyroidism is associated with a variety of haematological changes. Anaemia is common and may be normocytic, microcytic, or macrocytic.

Severe microcytic anaemia in hypothyroidism is most commonly seen in women who have menorrhagia, which is a frequent complication of this condition. Severe macrocytosis in hypothyroidism usually indicates an associated vitamin B_{12} deficiency; there seems to be a genuine association between pernicious anaemia and myxoedema. It has been suggested that mild macrocytosis may occur in hypothyroidism in the absence of vitamin B_{12} or folate deficiency, although published series of studies have shown a remarkable variability in the incidence of this phenomenon. Some patients with severe hypothyroidism have a small proportion of misshapen red

cells on their peripheral blood films.

The anaemia of uncomplicated myxoedema is normochromic and normocytic. The mechanism is still uncertain. Recent studies have shown that T_3, T_4, and reverse T_3 can all potentiate the effect of erythropoietin on the formation of erythroid colonies *in vitro*. This effect appears to be mediated by receptors with b_2-adrenergic properties. It appears that the thyroid hormones have a direct effect in altering the erythropoietin responsiveness of erythroid progenitors. It has also been suggested that part of the normochromic anaemia of hypothyroidism may be a physiological adaptation to reduced oxygen requirements by the tissues.

Curiously, patients with hyperthyroidism do not have elevated haemoglobin levels. There is some recent evidence that there may be a mild increase in the red-cell mass in hyperthyroidism, but that this is compensated for by an increase in plasma volume. In some patients with severe hyperthyroidism, there is a mild anaemia associated with abnormal iron utilization.

Adrenal disease

A mild, normochromic, normocytic anaemia together with neutropenia, eosinophilia, and lymphocytosis is observed in some patients with Addison's disease. There is a variety of haematological changes following the administration of corticosteroids or endogenous overproduction of these agents. These include granulocytosis, reduced lymphocyte count, involution of lymphatic tissues, and a decrease in the eosinophil and monocyte count.

Parathyroid disease

Primary hyperparathyroidism is occasionally associated with anaemia, which responds to removal of the parathyroid glands. The relation between parathyroid disease and marrow fibrosis is discussed in Section 12.

Diabetes mellitus

The structural changes that occur in the haemoglobin of diabetic patients are discussed in Chapter 12.11.1. There have been recent reports that there may be an increase in the red-cell volume of patients with severe diabetes. The mechanism and significance of this observation remains to be clarified. Severe diabetic acidosis is associated with a marked leucocytosis, even when there is no underlying infection. Hyperosmolarity impairs neutrophil function; reduced neutrophil migration has been observed in patients with diabetic ketoacidosis or poorly controlled hyperglycaemia. Because of the high incidence of atheroma in patients with diabetes, both platelet function and vessel-wall metabolism have been studied in considerable detail in this condition. Synthesis of prostaglandin I_2 in biopsy specimens of forearm veins is reduced. A variety of changes in platelet reactivity and survival have been observed. The relation of these changes to the vascular disease of diabetes requires further clarification.

Neuropsychiatric disease

Anorexia nervosa

About a third of patients with severe anorexia nervosa have a mild, normochromic, normocytic anaemia. In patients who are severely malnourished there may be mild neutropenia. There have been reports of the finding of irregularly shaped red blood cells in this condition. The platelet count is usually normal but there may be mild thrombocytopenia and in one study there was a marked increase in the rate of platelet aggregation.

Trauma

The brain is rich in thromboplastin activity. Acute disseminated intravascular coagulation occurs quite commonly after severe head or brain injury.

Myasthenia gravis

The association between myasthenia gravis and pure red-cell aplasia is described in Section 24.22. An immune neutropenia has also been described as part of the myasthenia–thymoma syndrome.

Lesch–Nyhan syndrome

This X-linked recessive disorder is described in detail in Chapter 11.4. There have been occasional reports of the development of severe megaloblastic anaemia, presumably resulting from defective nucleic acid synthesis; the condition has been reversed by the administration of large doses of adenine.

Abetalipoproteinaemia

This condition is characterized by an ataxic neurological disease, retinitis pigmentosa, fat malabsorption, and the absence of chylomicrons and low-density lipoproteins. It is caused by the failure to synthesize or secrete lipoprotein-containing products of the apolipoprotein B gene. It is characterized by the presence of from 50 to 90 per cent of acanthocytes in the peripheral blood. These are abnormal, spiky red cells, which have a moderately shortened survival. Despite these changes there is only a mild haemolytic anaemia.

Acanthocytosis with neurological disease and normal lipoproteins (amyotrophic chorea-acanthocytosis)

This syndrome is characterized by marked acanthocytosis associated with a progressive neurological disease, beginning in adolescence or adult life, which includes orofacial dyskinesia, lip and tongue biting, choreaiform movements, sensorimotor polyneuropathy, distal muscle wasting, and hypotonia. Because it has been found to follow both dominant and recessive forms of inheritance it is likely that this is a heterogeneous disorder. The cause is unknown.

Cardiac disease

There are several important haematological manifestations of cardiac disease, all of which are dealt with in more detail in Section 15. The severe haemolytic anaemia that occasionally follows the insertion of prosthetic valves, particularly the aorta, is described in Chapter 15.7.

A variety of abnormalities of coagulation are found in patients with cyanotic congenital heart disease. These include thrombocytopenia, low plasma fibrinogen levels, defective clot retraction, a deficiency of factors V and VII, and increased levels of fibrin degradation products. Overall, the severity of these abnormalities correlates with the degree of secondary polycythaemia. The exact mechanism is not known. In addition to the quantitative changes in blood platelets, there may also be qualitative abnormalities of platelet function. These include defects in both aggregation and release. They may be associated with a prolonged bleeding time. Again, the mechanism is not understood.

The striking haematological changes that may accompany bacterial endicarditis were mentioned earlier in this chapter. Dressler's syndrome may be associated with the anaemia of chronic disorders, atypical lymphocytes in the peripheral blood, and, certainly in the earlier descriptions of the disease, an eosinophilia of varying degree. Similar changes have been observed in the postpericardiotomy syndrome.

Further reading

Boxer H, Ellman L, Geller R, Wang C-A (1977). Anemia in primary hyperparathyroidism. *Archives of Internal Medicine* **137,** 588–90.

Castaldi PA (1984). Hemostasis and kidney disease. In: Ratnoff OD, Forbes CD, eds. *Disorders of hemostasis*, pp. 473–84. Grune and Stratton, Orlando, FA.

Colman N, Herbert V (1980). Hematologic complications of alcohol. *Seminars in Hematology* **17**, 164–72.

Dainiak N (2000). Hematologic complications of renal disease. In: Hoffman R *et al.*, eds. *Hematology. Basic principles and practice*, 3rd edn, pp. 2357–73. Churchill Livingstone, New York.

Erslev AJ (1995). Traumatic cardiac hemolytic anemia. In: Beutler E, Lichtman MA, Coller BS, Kipps TJ, eds. *Williams hematology*, 5th edn, pp. 663–5. McGraw-Hill, New York.

Goldsmith GH, Jr (1984). Hemostatic disorders associated with neoplasia. In: Ratnoff OD, Forbes CD, eds. *Disorders of hemostasis*, pp. 351–66. Grune and Stratton, Orlando, FA.

Hamblin TJ, ed. (1987). Haematological problems in the elderly. *Bailliere's Clinical Haematology* **1,** 271–596.

Hardcastle JD, Thomas WM (1989). Screening an asymptomatic population for colorectal cancer. In: Mortensen N, ed. *Ballière's clinical gastroenterology,* Vol. 3, pp. 543–66. Ballière Tindall, London.

Herbert V, ed. (1980). Hematologic complications of anemia in alcoholic patients. *Seminars in Hematology* **17**, vols 1 and 2.

Hoxie JA (2000). Hematologic manifestations of HIV infection. In: Hoffman R *et al.*, eds. *Hematology. Basic principles and practice*, 3rd edn, pp. 2430–57. Churchill Livingstone, New York.

Hughes GRV (1983). The lupus anticoagulant. *British Medical Journal* **287,** 1088–9.

Parker RI, Metcalf DD (2000). Basophils, mast cells and systemic mastocytosis. In: Hoffman R *et al.*, eds. *Hematology. Basic principles and practice*, 3rd edn, pp. 830–46. Churchill Livingstone, New York.

Ratnoff OD (1984). Hemostatic defects in liver and biliary tract diseases. In: Ratnoff OD, Forbes CD, eds. *Disorders of hemostasis*, pp. 451–72. Grune and Stratton, Orlando, FL.

Rosenthal DS (2000). Hematologic manifestations of infectious disease. In: Hoffman R *et al.*, eds. *Hematology. Basic principles and practice*, 3rd edn, pp. 2420–30. Churchill Livingstone, New York.

St John DJB, Young GP, *et al.* (1993). Evaluation of new occult blood tests for detection of colorectal neoplasia. *Gastroenterology* **104,** 1661–8.

Stockman JAI, Ezekowitz RA (1998). Hematologic manifestations of systemic diseases. In: Nathan DG, Oski FA, eds. *Hematology of infancy and childhood*, 5th edn, pp. 1841–91. Saunders, Philadelphia.

Weatherall DJ, Kwiatkowski D (1998). Hematologic manifestations of systemic diseases in children of the developing world. In: Nathan DG, Oski FA, eds. *Hematology of infancy and childhood*, 5th edn, pp. 1893–914. Saunders, Philadelphia.

Weinstein IM, Rosenbloom DE (2000). Hematologic problems in patients with cancer and chronic inflammatory disorders. In: Hoffman R *et al.*, eds. *Hematology. Basic principles and practice*, 3rd edn, pp. 2410–20. Churchill Livingstone, New York.

Zaroulis CG, Kourides JA, Valeri CR (1978). Red cell 2, 3-diphosphoglycerate and oxygen affinity of hemoglobin in patients with thyroid disorders. *Blood* **52,** 181–5.

22.8.1 Blood transfusion

P. L. Perrotta and E. L. Snyder

Blood transfusion is important for the care of patients with severe anaemia, haemorrhage, thrombocytopenia, and coagulation disorders. Advances in the understanding of red cell, platelet, and leucocyte antigen structure, as well as the immune responses to these antigens, have vastly improved transfusion therapy. Routine blood bank procedures, including ABO typing, antibody screening, and compatibility testing, identify most patients at risk for serious immune-mediated red cell transfusion reactions. One of the most important technological improvements in transfusion therapy was the development of sterile, disposable, and flexible plastic containers that allow separation of whole blood into cellular and non-cellular components, including red blood cells, platelets, and plasma. Anticoagulants and additives currently used in blood collection containers allow storage of liquid red cells for up to 42 days. These advances have essentially eliminated the use of whole blood.

Individual components are stored under optimal conditions. Only that portion of blood required by the patient is transfused. Plasma separated from whole blood can be further fractionated into coagulation factor concentrates, albumin, or gamma globulin. Cell separators capable of collecting platelets, plasma, granulocytes, peripheral blood stem cells, and, more recently, red blood cells, are also in widespread use across the United States and Europe. Changes in recruiting and screening blood donors, as well as advances in the testing of donor blood have reduced the risk of viral transmission in Europe and the United States. All units of blood collected in the United States and Britain are tested for hepatitis B, hepatitis C, HIV-1, HIV-2, HTLV-1, and syphilis. Nucleotide testing for HIV and hepatitis C is now performed in most European countries and in the United States. Other risks of transfusion therapy include acute and delayed haemolytic, febrile non-haemolytic, allergic, and septic reactions. Premedication before transfusions, and leucoreduction of blood components have reduced these risks.

Although the hazards of blood transfusion are relatively small, the expected benefit of a transfusion must outweigh any risk to the patient. Therefore, a thorough understanding of the indications and complications of blood transfusion are required to minimize exposure to unnecessary allogeneic blood products and to prevent wastage of limited blood resources.

Blood group systems

ABO system

Over 250 distinct antigens have been identified on the surface of red blood cells. The most clinically important belong to the ABO system. The codominantly expressed A and B genes, located on chromosome 9, code for glycosyl transferases that add either *N*-acetyl-D-galactosamine (A gene) or D-galactose (B gene) to the common precursor H antigen (Table 1). The O gene is structurally similar to the A gene except for a single base deletion that eliminates production of a functional enzyme. The AB antigens are of critical importance because individuals who lack the A and/or B antigens form IgM and IgG antibodies directed against the missing antigen(s). Circulating A and B antibodies can fix complement and cause intravascular haemolysis. Anti-A and Anti-B antibodies are 'naturally occurring', that is, they are formed without prior clinical antigenic stimulation. Presumably, individuals become immunized following exposure to carbohydrate ABO antigenic determinants commonly found in the bacterial environment. Accordingly, group A persons produce anti-B, group B produce anti-A, and group O produce both anti-A and anti-B. Circulating A and B antibodies are of critical importance in blood therapy because they are of high titre, can fix complement to C9, and are responsible for the vast majority of major haemolytic transfusion reactions.

Rh system

The Rh blood group system is composed of at least 44 distinct antigens. The five major antigens in the Rh system (D, C, c, E, and e) are responsible for the vast majority of Rh-related transfusion problems. It is now known that the D polypeptide is encoded at the *RHD* locus, whereas the CcEe polypeptide is coded by alleles at the *RHCE* locus. Based on the D gene frequency in North America and Europe, approximately 15 per cent of individuals will not produce D antigen and are 'Rh negative'. The most common nomenclatures used to classify the Rh antigens include those developed by Weiner and Fisher–Race (Table 2 and Table 3). Each of these systems has its advantages and limitations. Very rare individuals who lack all Rh antigens are termed 'Rh-null'. Rh-null red cells are morphologically abnormal and typically have shortened survival, resulting in a mild haemolytic anaemia. The successful cloning of the RhD gene potentially allows application of molecular techniques to determine fetal RhD status.

The most clinically important Rh antigen is D because it is highly immunogenic—the likelihood of a D-negative person developing anti-D following exposure to as little as 0.1 ml of D-positive red cells is extremely high. Approximately 80 per cent of Rh-negative individuals transfused with a single unit of Rh positive red cells will develop anti-D. Anti-D is responsible for immune reactions including haemolytic disease of the newborn (HDN) and immune-mediated transfusion reactions. Despite widespread use of Rh immune globulin, anti-D remains the most common cause of serious HDN. Rh-negative women most commonly produce anti-D following exposure to D-positive red cells during pregnancy, a miscarriage, or abortion. The anti-D formed is of the IgG class and therefore can cross the placenta where it may cause a potentially fatal intrauterine HDN in an Rh-positive fetus.

Other blood groups

There are a large number of other well-characterized red cell blood group systems. Antibodies directed against some of these antigens may be naturally occurring

and of little clinical significance (e.g. Lewis, Ii). Other antibodies may form following exposure to the corresponding antigen (e.g. Kell, Duffy, Kidd). Some of these antibodies are clinically significant in that they are associated with immune-mediated red cell destruction of transfused cells and HDN (Table 4). In most cases, compatible blood can be found for patients with significant red cell alloantibodies. Based on the high incidence of some red cell antigens on the cells of specific donor populations, some patients may be difficult to transfuse if they have developed multiple antibodies. This is particularly true in patients with sickle cell disease and other red cell disorders who require frequent transfusions. There have been recent advances in the understanding of the molecular genetics of many blood group antigens. These advances will eventually be used to resolve blood group discrepancies, screen for red cells of specific antigen makeup, and to identify fetuses who are at risk for HDN.

The Kell (K) blood group system is clinically important because antibodies to Kell system antigens can cause haemolytic transfusion reactions and HDN. Only the D antigen is more immunogenic than the K antigen, the major antigen of the system. Kell is present in about 10 per cent of white individuals. Therefore, a K-negative blood recipient is unlikely to receive K-positive red cells during transfusion. The Kell blood group is linked to chronic granulomatous disease, a congenital disease resulting in decreased oxidative capacity of neutrophils, which leads to recurrent, severe bacterial infections. The genetic defect seen in chronic granulomatous disease is located on the X chromosome near the Kx Kell locus. The red cells of patients with chronic granulomatous disease are acanthocytic and are prone to mild haemolytic destruction. Systemic abnormalities described in chronic granulomatous disease include cardiomyopathy, areflexia, skeletal myopathies, and muscle wasting.

The Duffy (Fy) blood group is composed of six antigens, of which Fy^a and Fy^b are most important to transfusion practice. Fy^a antibodies have caused severe haemolytic transfusion reactions and severe HDN. Fy^b antibodies have more rarely been associated with these complications, and antibodies against the remaining Duffy antigens are rarely clinically important. Antibodies against Fy^a and Fy^b are reasonably common in diverse populations because the antigen frequency varies dramatically across racial groups. Fy^a and Fy^b antigens are present in 66 per cent and 83 per cent of white blood donors, respectively. These antigens have a much lower incidence in African populations. The Duffy system has an interesting association with malaria. Specifically, $Fy(^{a-b-})$ negative red cells are resistant to *Plasmodium vivax* and *Plasmodium knowlesi* infection. Red cells from most West African blacks are $Fy(^{a-b-})$ and therefore, resistant to these forms of malaria.

The Kidd (Jk) system is composed of three antigens (Jk^a, Jk^b, Jk^3). Jk^a and Jk^b antigens are found on approximately 75 per cent of white donor red cells. Anti-Kidd antibodies have caused severe haemolytic reactions and milder forms of HDN. Kidd antibodies are formed following exposure to Jk antigens during transfusion or pregnancy. They are unusual in that once formed, these antibodies often fall to non-detectable levels, and may not be detected in an already immunized patient. In this situation, transfusion of additional Kidd-positive red cells may cause a rapid immunological response, leading to formation of high-titre anti-Kidd, and subsequent haemolysis.

Detection of blood group antibodies

Antibody screening and antibody identification

Prior to receiving a blood transfusion, patients' red cells are typed for ABO and Rh status using commercially available reagents. During 'front typing', the donor's red cells are reacted with antibodies directed against the A, B, and D antigens. Blood grouping is confirmed during 'back typing' in which donor serum is tested for the presence of anti-A and anti-B antibodies. Following blood grouping, recipient serum or plasma is screened for red cell antibodies. Antibody screening is typically performed by incubating a patient's serum with two to four group O red cells sources that, in sum, contain all the common and clinically significant red cell antigens. If an antibody is present in the serum, it will react with the screening cell and cause red cell agglutination. Naturally occurring ABO antibodies do not interfere with antibody identification because screening cells are type O. Antibody screening is commonly performed at room temperature (immediate phase), after incubating patient serum and test red cells at 37°C, and after incubation with antihuman globulin serum (Coombs phase). Some blood banks screen samples after adding various antigen–antibody enhancing substances including polyethylene glycol, low-ionic strength saline, and albumin. Each has certain advantages and disadvantages in terms of sensitivity and specificity in detecting clinically significant antibodies. Most blood banks still perform 'tube testing' in which red cell agglutinates are identified in standard test tubes. There are a number of newer systems that are being used to detect antigen–antibody reactions. These include gel systems based on the differential mobility of red cell agglutinates through gel columns, and capture systems in which test red cells are immobilized on microtitre plates. Newer automated and semiautomated systems will probably replace tube testing for the majority of ABO grouping, Rh typing, antibody screening, and crossmatching.

If a patient's serum reacts with one or more screening cells, additional tests are performed to identify the antibody(ies). In most cases, the serum is tested against a larger commercial 'panel' of group O red cells of known antigen profile. Based on the reactivity pattern, the antibody can usually be identified. There are a large number of techniques used to identify red cell antibodies. Many are based on using materials that either enhance or suppress the reactivity of a specific antibody. In some cases, a patient's serum may react with all panel cells. These 'panagglutinins' can be caused by: (1) a single antibody directed against a high incidence antigen present on all panel test red cells; (2) multiple antibodies that in total react with all test cells; or (3) an autoantibody. Autoantibodies are often found in autoimmune haemolytic anaemias, in which case the patient's serum will also react with his or her own red cells (see section on autoantibodies, below).

Compatibility testing

Routine compatibility testing is typically performed on red cell units before being transfused to a patient. Specifically, donor red cells are reacted with patient serum and if no reaction is observed, the unit is considered 'compatible'. In emergency situations, there may be insufficient time to perform compatibility testing. Many hospitals will supply group O Rh-negative red cells until a patient sample is obtained and tested. If a patient's ABO Rh status is known with certainty, then type-specific non-crossmatched blood can be provided. In either case, compatibility testing is performed on these transfused units as soon as possible. It is important to realize that supplies of O negative blood are often limited. A 'computer crossmatch' has been instituted at hospitals in North America. Patients with known ABO and Rh typing, and who have a negative antibody screen are provided ABO compatible blood while omitting the crossmatch step described above. Although a true serological crossmatch is not performed, the computer crossmatch is safe in the vast majority of transfusions.

Autoantibodies

Autoantibodies consist of immunoglobulins that react with a wide range of self-antigens including membrane and intracellular components, adsorbed plasma proteins, and nuclear antigens. Patients with warm autoimmune haemolytic anaemia often require transfusion. In this case, the blood bank may have difficulty finding a 'compatible' unit of red cells because the patient's serum not only reacts with his or her own red cells, but also those of all donor red cells. Additional time may be required by the blood bank to exclude the presence of a significant underlying alloantibody that is obscured by the autoantibody. Upwards of 25 per cent of previously transfused autoimmune haemolytic anaemia patients may have an underlying alloantibody. An underlying alloantibody may result in accelerated red cell destruction. Therefore, transfusion therapy must be carefully planned in these patients. Autoimmune antibodies often appear to have specificity for Rh antigens (e.g. anti-e), but the transfusion of antigen negative red cells (e.g. e-negative) is not indicated as *in vivo* red cell survival of antigen negative cells is usually no better than antigen positive cells.

Clinical use of blood components

Red blood cells

Red blood cells account for approximately 75 per cent of the annual cost of transfusion therapy in the United Kingdom. Red blood cells, prepared from whole blood by removing most of the plasma, are indicated for patients with both acute haemorrhage and chronic anaemias (Table 5). Earlier solutions, composed of citrate, dextrose, and phosphate buffers allowed storage of red cells from 21 to 35 days. It was later observed that the addition of adenine to the preservative solution improved cell viability by increasing intracellular ATP levels. The haematocrit of red cell units varies from 70 per cent (citrate, phosphate, dextrose, adenine (CPDA-1)) to 55 to 60 per cent (additive solution, AS). Citrate contained in blood preservatives binds calcium to inhibit clotting and may cause hypocalcaemia and alkalosis in neonates and massively transfused patients. Red blood cells or AS units refrigerated at 1 to 6°C have a shelf-life of 35 (red blood cells) to 42 (AS units) days depending on the ingredients of the preservative. During storage, the following changes are observed in red cell units: (1) a fall in pH; (2) decreases in red cell ATP and 2,3-diphosphoglycerate; (3) increased supernatant potassium; and (4) decreased supernatant glucose. Leucocytes can be removed from the product at the blood centre collection site (prestorage leucodepletion), in the hospital blood bank prior to release, or at the patient's bedside using leucoreduction filters. Red blood cells with uncommon antigen profiles can be frozen within 6 days of collection and stored for up to 10 years. They are frozen with approximately 40 per cent glycerol to avoid cell dehydration and damage during the freezing process. Frozen red cells are no longer used as a leucoreduced product.

The patient's overall clinical status and laboratory parameters should both be considered when deciding to transfuse a patient. A decision should not be based on the

haematocrit alone. Younger patients will usually tolerate a given degree of hypoxaemia and hypotension better than older patients who may have underlying coronary or myocardial disease. Evidence of symptomatic anaemia include excessive fatigue, malaise, headache, tachycardia, hypotension, and end-organ damage. Hypovolaemic shock typically ensues with acute loss (<24 h) of over 30 per cent of total blood volume. Initially, the haematocrit will be falsely elevated in acute haemorrhage, but will then fall with fluid resuscitation. Slowly developing, chronic anaemias are usually better tolerated then rapid onset anaemias due to the ability of the body's fluid compensatory mechanisms. Transfusion is rarely indicated when the haemoglobin (Hb) is greater than 10 g/dl, and is often not considered until the Hb is less than 7 g/dl. A patient's cardiac and pulmonary status must be considered when determining transfusion thresholds. Patients with unstable angina or acute myocardial infarction may require transfusion when the Hb is less than 10 g/dl. In the absence of active red cell destruction, transfusing a single unit will typically increase the Hb by 1 g/dl (haematocrit by 3 per cent).

Platelets

The availability of plastic primary collection bags with attached satellite containers allows the harvesting of platelets as a by-product of red cell separation. In the United States, platelets are prepared by the platelet-rich plasma method, whereas the buffy coat method is used in Europe. Each unit of 'random donor' platelets prepared by differential centrifugation of a single whole blood collection typically contains at least 5.5×10^{10} platelets suspended in 50 ml of plasma. Platelets stored under agitation at 20 to 24°C in plastic containers that allow oxygen diffusion have a shelf-life of 5 days. The risk of bacterial growth and development of platelet function abnormalities (platelet storage defect) preclude storage longer than 5 days. 'Random donor' platelets are usually administered in pools of 4 to 6 units. In the absence of conditions associated with decreased platelet survival, each unit can be expected to raise the recipient's platelet count by 5000 to 10 000/µl. Single donor platelets prepared by apheresis contain more than 3×10^{11} platelets suspended in about 200 ml plasma, equivalent to 6 average random donor platelet units. Platelets are not normally crossmatched with the recipient's serum. ABO type-specific platelets should be provided whenever possible because transfusion of out-of-type platelets may result in a postplatelet increment 10 to 20 per cent less than that expected for ABO type-specific platelets. Rh antigens present on the small number of contaminating red cells found in platelet concentrates are capable of immunizing a Rh negative recipient. If Rh negative platelet concentrates are not available for an Rh negative patient, Rh positive platelets can be transfused followed by administration of Rh immune globulin within 72 h of transfusion.

Platelets are provided to thrombocytopenic patients who are bleeding or to severely thrombocytopenic patients as a prophylactic precautionary measure. Spontaneous bleeding is rare when a patient's platelet count is over 20 000/µl, and studies suggest that patients who receive chemotherapy can tolerate platelet counts as low as 5 to 10 000/µl. Postsurgical patients may require platelet transfusions to control or prevent postoperative bleeding when the platelet count is over 50 000/µl. Overall coagulation status should also be considered because patients with plasma coagulation factor disorders are more likely to bleed at marginal platelet counts. Actively bleeding patients on aspirin, an irreversible inhibitor of platelet function, may require transfusions at higher platelet counts, although transfused platelets will also be affected if the patient remains on aspirin.

Platelet refractoriness is a major problem for patients who are dependent on platelet transfusions. The corrected count increment (CCI) is used to identify patients who are refractory to platelet transfusions through either HLA or platelet (HPA) alloimmunization. The CCI is calculated as follows:

$$\text{CCI} = \frac{\text{Post (/ml)} - \text{Pre (/ml)} \times \text{BSA } (m^2)}{\text{plts} \times 10^{-11}}$$

where Pre = pretransfusion platelet count, Post = post-transfusion platelet count drawn 1 to 4 h after completion of the transfusion, plts = number of platelets transfused (1 U 'random donor' platelets ~ 0.7×10^{11} plt; 1 U single donor platelets ~ 3 to 4×10^{11} plt), and BSA = body surface area in m^2.

Causes of a platelet refractory state include disseminated intravascular coagulation (DIC), sepsis, and circulating immune complexes. After documenting a low CCI (<5000), crossmatch-compatible platelets or HLA-matched single donor platelets should be considered. These products are not readily available in most blood banks. Increasing the dose of standard platelet concentrates can be considered until compatible platelets are identified. Leucocyte reduction filters, as well as UV-B irradiation, decrease the rate of HLA alloimmunization to platelets. Leucocyte reduction filtered blood products should be provided to patients who will require many platelet transfusions.

Granulocytes

There is renewed interest in granulocyte transfusion kindled by improvements in apheresis collection techniques and the use of steroids and/or growth factors to improve granulocyte yields. Granulocytes are primarily transfused to neutropenic oncology patients who develop Gram positive or Gram negative bacterial sepsis unresponsive to antibiotic therapy for a minimum of 24 to 48 h. Granulocytes collected from non-stimulated healthy donors by apheresis contain at least 1×10^{10} neutrophils/unit and can be stored for only 24 h at 20 to 24°C. Higher numbers of granulocytes can be collected when donors are stimulated by steroids and/or growth factors. They contain large numbers of red cells (20–50 ml) and must be crossmatched with the recipient's serum. Granulocytes should be irradiated (2500 cGy) because of the large number of lymphocytes present in the product. They are considered for the above patients provided they also have an absolute neutrophil count less than 500/µl and a reasonable chance of marrow recovery. Because of their short half-life, granulocytes are usually provided daily until the patient can maintain an absolute neutrophil count greater than 500/µl without transfusion or until the infection resolves. Infusion of larger numbers of granulocytes does allow measurable increases in recipient neutrophil counts, but the optimal dose and frequency remain undefined. Febrile reactions to granulocytes are common, the reactions being more severe when amphotericin is infused near the time of granulocyte transfusions. Overall, the additional benefit of granulocyte transfusions for these neutropenic patients as compared to antibiotic treatment alone is unclear. The collection of granulocytes, or any blood component, by apheresis is not an entirely innocuous process. The donor is at risk for uncommon, but potentially serious, adverse reactions including hydroxyethyl starch-related hypertension and anaphylaxis, and citrate-induced hypocalcaemia. Minor, but typically tolerable, side-effects of pretreating granulocyte donors with dexamethasone (insomnia, flushing) and/or G-CSF (bone pain, headaches, insomnia) occur in a substantial number of donors.

Plasma, cryoprecipitate, and plasma derivatives

Plasma therapy began in the late 1940s with the development of fractionation techniques in which large pools of human plasma collected from many donors could be separated into specific plasma proteins. Plasma is separated from whole blood by centrifugation and frozen within 8 h of collection in order to maintain the activity of labile coagulation factors, factors V and VII. Fresh frozen plasma (FFP) contains all coagulation factors, plasma proteins, and complement. FFP should not be transfused for volume expansion because of the risk of transfusion-transmitted disease and the availability of other, safer non-plasma substitutes. The primary indications for FFP transfusion include deficiency of multiple coagulation factors as seen in liver disease and DIC. It is often used to reverse warfarin anticoagulation urgently. One unit of clotting factor activity is defined as the amount of activity in 1 ml of normal plasma. FFP is not particularly effective in replacing individual clotting factors because of the large volumes that would be required to obtain adequate factor levels. The patient's fluid and cardiovascular status may preclude the use of large amounts of plasma.

FFP is no longer the treatment of choice for coagulopathies where virally inactivated or recombinant products exist, such as for deficiencies of factor VIII (haemophilia A) or factor IX (haemophilia B). Fears of transmitting infectious disease with plasma transfusion remain of concern, particularly for pooled products. In addition to donor screening and testing, other strategies to decrease infectious risk include photoinactivation and solvent/detergent treatment technologies. Solvent and detergent treated plasma prepared from 500 l plasma pools (2500 donors) is available in the United States. This treatment effectively removes lipid enveloped viruses including HIV and HCV, but does not eliminate non-enveloped viruses such as parvovirus B-19 and, possibly, other unrecognized non-lipid-enveloped pathogens.

Cryoprecipitate is prepared by thawing FFP between 1 and 6°C. Each 10 to 20 ml unit contains 100 to 350 mg fibrinogen/unit, at least 80 IU/unit factor VIII, and some von Willebrand factor. Use of cryoprecipitate is generally reserved for patients with von Willebrand's disease or those with severe hypofibrinogenaemia (<100 mg/dl). Cryoprecipitate and thrombin are combined to make 'fibrin glue.' This biological sealant works well but exposes the recipient to the risks of transfusion-transmitted disease due to the use of cryoprecipitate.

Albumin is available as a 5 or 25 per cent solution and is used to treat hypovolaemia and hypoalbuminaemia, primarily in surgical settings. Albumin is virally inactivated by heat treatment plus other viral inactivation steps, and is tested for HCV RNA. Properly processed albumin is not considered to transmit viral disease. Readily available non-plasma colloidal solutions have replaced albumin in many situations requiring volume expansion. Intravenous immunoglobulin (IVIg) is used to treat patients with immune thrombocytopenia, Guillain–Barre syndrome, and autoimmune haemolytic anaemias. Prompt and adequate doses of Rho (D) immunoglobulin (RhIG) available in intramuscular and IV preparations are used to prevent alloimmunization in D-negative patients who are exposed to D-positive red cells through transfusion or pregnancy. Rapid advances in molecular techniques led to the cloning and purification of recombinant clotting factors. Recombinant factor

VIII, IX, and factor VIIa are available.

Complications and management of transfusion therapy Table 6)

Acute intravascular haemolytic reactions

Acute intravascular haemolytic transfusion reactions (AIHTR) are one of the most serious transfusion complications. These reactions occur in blood recipients who have developed antibodies directed against antigens present on the transfused red blood cells. ABO incompatibility remains the most common cause of immediate intravascular haemolytic reactions. Donor erythrocytes carrying either A and/or B red cell antigens bind to the recipient's naturally occurring anti-A and/or anti-B antibodies, resulting in complement fixation, formation of the C5b-9 membrane attack complex, and subsequent haemolysis. Biological response modifiers, such as proinflammatory cytokines (IL-1, TNF-a), chemokines (IL-8), and complement fragments (C3a, C5a), also play a role in the pathophysiology of AIHTRs. AIHTRs are typified by the sudden onset of back pain, hypotension, tachycardia, fever, chills, diaphoresis, and dyspnoea. The symptoms usually begin soon after the transfusion is started in immunocompetent recipients. Laboratory studies reveal an increase in unconjugated bilirubin (typically to 2–3 mg/dl) and marked elevation of lactate dehydrogenase. Other evidence of intravascular haemolysis include haemoglobinuria and haemoglobinaemia. The direct antiglobulin test (direct Coombs) becomes reactive due to the coating of donor red cells with the recipient's antibodies.

AIHTRs are usually caused by transfusions of ABO incompatible blood resulting from patient identification or clerical errors, but they can also be caused by incompatibility within other blood group systems (Duffy, Kidd). Proper labelling of clots used by the blood bank for compatibility testing and careful identification of patients are the best ways to prevent potentially fatal reactions. AIHTRs are medical emergencies and treatment consists of immediately stopping the transfusion, close monitoring of vital signs, cardiac and airway support, and maintenance of urine output with saline diuresis with or without a loop diuretic (Table 7). Dialysis must be considered in patients with renal failure.

Delayed extravascular haemolytic reactions

Delayed haemolytic transfusion reactions (DHTRs) generally occur in patients who have a negative antibody screen on pretransfusion testing, but who then experience accelerated destruction of transfused red cells 7 to 14 days post-transfusion. In most cases, red cell destruction is caused by an antibody that is initially of a titre below the limits of detection on routine screening. The antibody then rapidly forms on secondary exposure to the offending antigen. Only rare DHTRs are caused by primary allosensitization in which a patient synthesized a new antibody. The antibodies typically fix complement to C3 and stop, thus, resulting in extravascular as opposed to intravascular haemolysis. Antibodies most commonly implicated in DHTRs include those directed against Rh (E, c), Kell, Duffy, and Kidd blood group antigens. DHTRs can be diagnosed by an unexpected post-transfusional fall in haematocrit, development of unconjugated hyperbilirubinaemia, and appearance of a positive direct antiglobulin test. There is usually a delay of 3 days to 2 weeks between transfusion and onset of extravascular haemolysis. Only rarely do delayed reactions cause intravascular haemolysis with associated haemoglobinaemia and haemoglobinuria.

Febrile non-haemolytic reactions

Febrile non-haemolytic transfusion reactions (FNTRs) to red blood cell and platelet transfusion are very common. They are caused by the development of antibodies in the recipient directed against HLA and/or leucocyte-specific antigens on donor white blood cells and platelets. Reactions between leucoagglutinins present in the transfused product and recipient leucocyte antigens can also occur. Subsequent formation of leucocyte antigen–antibody complexes results in complement binding and release of endogenous pyrogens such as IL-1, IL-6, and TNF-a. Cytokines generated by leucocytes during platelet and red cell storage may also contribute to FNTRs. Symptoms occur during or several hours after the transfusion, and typically include low-grade (> 1°C rise) and high grade fevers, accompanied by shaking chills. Rarely, vomiting, dyspnoea, hypotension, and decreased oxygen saturation may develop. The severity of symptoms is often directly related to the number of leucocytes in the product or the rate or volume of transfusion. Leucoreduction of blood components decrease the frequency of febrile transfusion reactions. Premedication with an antipyretic such as acetaminophen can ameliorate mild febrile transfusion reactions. Antihistamines are not helpful in preventing or treating febrile transfusion reactions. Corticosteroids can also minimize febrile transfusion reactions if they are administered several hours before the transfusion. Intramuscular or subcutaneous meperidine will usually resolve severe rigors in a matter of minutes. If symptoms do not resolve in less than 4 h or are especially severe, other complications such as sepsis due to contaminated blood products or a haemolytic reaction should be considered.

Allergic reactions

Allergic reactions to plasma, platelets, and red blood cells are relatively common. They present as pruritis and/or urticaria in the absence of fever. Allergic reactions are usually IgE mediated and most symptoms are attributed to histamine release. It may be difficult to distinguish allergic and febrile transfusion reactions when urticarial symptoms are accompanied by low-grade fever. Common symptoms and signs include erythema, papular rashes, wheals, and pruritis. Severe anaphylaxis resulting in bronchospasm and hypotension are possible. As in other allergic responses, symptoms are not dose-related and severe manifestations can occur following small exposures. Treatment of mild allergic reactions consists of stopping the transfusion and administering diphenhydramine or other antihistamines. In a mild allergic reaction with only pruritis and hives, it is acceptable to continue transfusing the same unit providing the symptoms promptly resolve and there is no fever or vasomotor instability. If symptoms recur after the transfusion is restarted, a new unit should be obtained. Severe anaphylactic reactions with bronchospasm and cardiovascular collapse are rare and should be treated like any other anaphylactic reaction with steroids, vasopressors, and airway support. Washed red blood cells in which the residual donor plasma has been removed and replaced by saline may benefit patients with repeated or severe allergic reactions. Leucocyte reduction filters are not helpful because they do not remove the implicated soluble mediators.

Septic reactions

Blood products can become contaminated by bacteria if a donor is bacteraemic at the time of collection or if improper arm preparation occurs during venipuncture. Transfusion of blood products contaminated by bacteria is particularly dangerous and can result in profound hypotension and shock. There are no laboratory screening tests commonly used to detect bacterial contamination and contaminated units cannot be easily identified by inspection. The risk of septic transfusion reactions is higher for platelet transfusions than other blood components because platelets are stored at room temperature. Common organisms implicated in septic transfusion reactions include Gram-positive (*Staphylococcus* sp.) and Gram-negative (*Enterobacter*, *Yersinia*, *Pseudomonas* sp.) bacteria. Blood cultures should be obtained from patients who develop high fevers following or during transfusion, especially if they become hypotensive. A Gram stain of the suspected contaminated product may be helpful but is often negative, and the product should be cultured if possible. Other symptoms attributed to preformed endotoxin and cytokines include skin flushing, severe rigors, and rapid-onset cardiovascular collapse. The symptoms may occur during, or minutes to hours after the transfusion is completed. Treatment includes fluids, cardiorespiratory support, and broad-spectrum antibiotics. Febrile transfusion reactions can usually be distinguished from septic transfusion reactions by the former's self-limited nature and lack of profound hypotension.

Transfusion-related acute lung injury

Transfusion-related acute lung injury (TRALI) is a serious complication of blood transfusion therapy that presents as non-cardiogenic pulmonary oedema. It typically occurs within 6 h of transfusion. It is clinically identical to the adult respiratory distress syndrome (ARDS). The most common clinical findings are rapid-onset symptoms including dyspnoea, tachypnoea, cyanosis, fever, and hypotension. Lung auscultation reveals diffuse, crackly, and decreased breath sounds. Invasive cardiac monitoring demonstrates normal cardiac pressures and function with hypoxaemia and decreased pulmonary compliance. Radiographic findings include diffuse, fluffy infiltrates typical of pulmonary oedema. The aetiology is believed to involve immune-mediated reaction of HLA antibodies or other leucoagglutins with white cells resulting in leucocyte activation. Granulocytes are first activated by HLA or other Ag–Ab complexes and then migrate to the lungs. The activated leucocytes bind to the pulmonary capillary bed via integrins and other cell adhesion molecules where they release proteolytic enzymes that destroy tissue, resulting in a capillary leak syndrome and pulmonary oedema. More recently, reactive lipid products released from donor cell membranes have been associated with the development of TRALI. TRALI should be suspected in patients with severe and rapid-onset respiratory distress following transfusion therapy, or, more specifically, pulmonary oedema without hypervolaemia. Definitive diagnosis requires identification of HLA and/or granulocyte antibodies in either the donor's or recipient's serum, as well as the corresponding antigens on the recipient's or donor's leucocytes. This testing is performed in a few specialized laboratories. Approximately 80 to 90 per cent of patients with TRALI will survive with supportive care including aggressive respiratory support, supplemental oxygen, and mechanical ventilation. Based on the presumed pathogenesis of TRALI, leucoreduced blood products could theoretically decrease the incidence of TRALI. Drugs used to treat TRALI have included corticosteroids and diuretics.

Transfusion-associated graft-versus-host disease

Acute graft-versus-host disease (GVHD) is a rare complication of blood transfusion, but is fatal in approximately 90 per cent of patients. TA-GVHD occurs when donor immunocompetent T and NK cells attack immunoincompetent recipient cells because these recipient cells appear foreign due to differences in major or minor histocompatibility antigens. GVHD is commonly seen following allogeneic bone marrow transplant but may also rarely occur in immunodeficient or immunosuppressed patients following blood transfusion. Removal of T cells from a donor graft can prevent acute GVHD in oncology patients, but is associated with increased graft failure and a decrease in a 'graft-versus-leukaemia' effect. The risk of TA-GVHD is related to the number of viable T lymphocytes transfused, the recipient's immune status, and the HLA disparity between donor and host. Therefore, multiply transfused patients who receive cells from donors who share HLA haplotypes with the recipient are at greatest risk. Clinically, TA-GVHD is characterized by the acute onset of rash, abdominal pain, diarrhoea, liver abnormalities (elevated liver enzymes, hyperbilirubinaemia), and bone marrow suppression 2 to 30 days following transfusion. The maculopapular rash seen is similar to that observed in acute GVHD following bone marrow transplant, and biopsy of the skin may help to confirm the diagnosis. Pancytopenia may be severe and is attributed to destruction of recipient marrow stem cells by donor lymphocytes. Immunosuppressive therapy with prednisone and cyclosporine has had little effect in TA-GVHD. Fortunately, TA-GVHD can be prevented by irradiating products prior to transfusion. Specifically, irradiation of cellular blood products with 2500 cGy inactivates donor lymphocytes and is the most effective method for preventing TA-GVHD.

Transfusion-transmitted isease

Despite major improvements in blood safety during the past 20 years, there remains a relatively small risk of transfusion-transmitted disease. The use of volunteer donors and predonation screening questionnaires were the first steps taken to reduce the risk of transfusion-related hepatitis and HIV. These risks continue to drive mandated pretransfusion testing requirements in developed countries. The advent of enzyme immunoassays in the 1970s, and more recent nucleotide testing, have further decreased the risk of transfusion-transmitted disease (Table 8). Transfusion-transmitted disease is a persistent problem in parts of the world that do not have access to screening tests.

Pretransfusion testing typically includes screening for syphilis, hepatitis B (HBsAg, anti-HBc), hepatitis C (anti-HCV), human immunodeficiency virus (anti-HIV-1/2, HIV-1 p24 antigen), human T cell lymphotropic virus (anti-HTLV-1/2), and syphilis. Serum alanine aminotransferase is measured in most European countries as a non-specific surrogate marker of hepatitis. When positive, these tests are typically confirmed by supplemental or confirmatory testing. Current estimates of the risk of transfusion-related HIV range from 1:500 000 to 1:750 000 units transfused. Despite improvements in tests used to detect HIV antibodies in donors, the 'window period' in which HIV could be transmitted by an infected, but HIV seronegative, donor remained in 1996 at about 25 days. The introduction of screening for HIV-1 p24 antigen in 1997 decreased the window period to approximately 15 days.

Genomic testing for hepatitis C virus (HCV) RNA was implemented in the United States and Europe to detect seronegative, yet infectious units. Nucleotide testing (NAT) for hepatitis C and HIV is typically performed on small pools of samples. The importance of hepatitis C transmission in blood therapy has been confirmed in many countries by retrospective review. During these 'look backs', recipients of blood components from donors later found to be positive (since anti-HCV screening was only instituted in 1991) are examined. A large percentage of these recipients, up to 75 per cent, are found to be anti-HCV positive. The majority of those who seroconvert will develop chronic liver disease. NAT testing will decrease the transfusion-related hepatitis C by decreasing the window period, from approximately 60 to 80, to 10 to 20 days. Hepatitis G virus has been transferred by blood transfusion, but its significance is unclear in that transfusion-acquired HGB infection has not been associated with acute or chronic hepatitis.

Several techniques have been developed to inactivate viruses in plasma including solvent and detergent treatment, and photochemical inactivation using psoralens and long wavelength UV light. Methods to inactivate infectious pathogens in cellular blood components, including platelets and red cells, are not currently available but are under development. Due to the low risk of viral infection by transfusion and the fact that most patients who receive plasma also receive cellular blood components, the cost-effectiveness of virally-inactivated plasma is very low. Albumin, immune globulin, factor concentrates, and other plasma derivatives are also virally attenuated, following standard treatment protocols.

Other pathogens, such as CMV and parvovirus B19, are common in the general donor population, and may pose a serious threat in immunocompromised patients. Approximately 40 to 60 per cent of blood donors have been exposed to CMV during their lifetime and subsequently are CMV seropositive. Only about 2 per cent of CMV seropositive donors, however, are actively infected and transfusion of their blood to an immunocompromised recipient could cause potentially serious disease. The actual risk of post-transfusion seroconversion to a CMV negative recipient who receives CMV-untested blood depends on the prevalence of CMV seropositivity in the donor population.

A number of parasitic diseases are known, or are suspected to be transmitted by blood transfusion. These include malaria, Chagas' disease, babesiosis, leishmaniasis, and toxoplasmosis. Transmission of Lyme disease (*Borrelia borgdorferi*) by transfusion has not been documented. The risk of new-variant Creutzfeldt–Jakob disease (nvCJD), first described in the UK in 1996, is unknown. It is unclear whether nvCJD is transmissible by blood transfusion and this form of transmission has not been reported. Fears of transmitting nvCJD, however, have resulted in implementation of a universal white blood cell reduction policy in the United Kingdom. There is a risk of acquiring babesiosis by blood transfusion, and infections with this organism can be dangerous in at risk populations (e.g. splenectomized patients) if untreated.

Use of special blood products

Leucoreduction

Leucocytes contained in blood components can provoke febrile non-haemolytic reactions, induce HLA alloimmunization, and transmit cytomegalovirus (CMV) to at-risk recipients. Leucocytes are most effectively removed from red cell and platelet concentrates by leucocyte reduction filters. Third-generation leucocyte reduction filters remove 3 to 4 $\log_{10}$ of the total intact leucocytes found in red cell and platelet concentrates. American Association of Blood Bank standards require that units labelled leucoreduced in the United States contain less than 5×10^6 white blood cells. Red cells are either leucoreduced shortly after blood collection (prestorage leucodepletion), following refrigerated storage (poststorage leucodepletion), or at the bedside during transfusion. Filters are similarly used to leucoreduce platelet concentrates—apheresis devices have been designed to collect leucoreduced platelets directly (process leucoreduction). Quality control measures must be in place in order to verify adequate leucoreduction of transfused products.

Leucoreduction has been shown to reduce the prevalence and severity of febrile transfusion reactions and to decrease the risk of HLA alloimmunization. Leucoreduced products are less likely to stimulate the HLA alloantibodies implicated in both febrile transfusion reactions and antibody-induced platelet refractoriness. Other generally accepted benefits of white blood cell reduction include reducing platelet refractoriness, and decreasing the risk of transmitting white blood cell-related infectious agents including CMV and HTLV-I/II. Prestorage leucoreduced products are preferable because they are also devoid of cytokines and other biological response modifiers which play a role in transfusion complications. Many of these proteins are not efficiently removed by leucocyte reduction filters. With the dramatic decrease in the risk of viral transmission, investigators are focusing on the immunomodulatory effects of blood transfusion. These effects specifically deal with associations between allogeneic transfusion and bacterial infection, tumour progression, and tumour recurrence. Universal white blood cell reduction of both red blood cells and platelets has been required and/or is being implemented in a number of countries including the United Kingdom, Canada, France, Ireland, Portugal, and the United States.

Irradiation

Blood components are irradiated to prevent potentially lethal transfusion-associated graft-versus-host disease by interfering with the ability of lymphocytes to proliferate. Irradiation of supportive blood components is indicated in bone marrow or peripheral blood stem cell transplant recipeints, patients with congenital immunodeficiency states, neonates, premature infants, and during intrauterine exchange transfusion. Patients with AIDS commonly receive irradiated components, although there is no clear increased risk of transfusion-associated graft-versus-host disease in this population. Standard guidelines recommend irradiating red blood cells, platelets, and granulocytes with a minimum dose of 2500 cGy. Platelets and red cells are not adversely affected by this exposure. It is not necessary to irradiate FFP or cryoprecipitate because they do not contain viable leucocytes. Bone marrow or peripheral blood stem cells must never be irradiated prior to transplant.

Cytomegalovirus-safe

Cytomegalovirus (CMV) infection is a leading cause of morbidity and mortality in marrow and solid organ transplant patients. Most serious CMV infections that develop in these populations are a result of latent reactivation of recipient CMV, but CMV can also be transmitted by blood transfusion. Therefore, blood banks supply products that have a low potential of transmitting CMV. The available products include CMV seronegative units prepared from donors who are CMV antibody negative, and leucodepleted components. The latter refers to blood components leucoreduced in a blood centre or laboratory using cGMP techniques. Depending on the donor population, however, as many as 80 to 90 per cent of blood donors may be CMV seropositive. Thus, the demand for CMV seronegative products may exceed supply. In addition, CMV seronegative products can transmit CMV disease. Studies suggest that CMV seronegative and leucodepleted filtered products are equivalent in preventing CMV transmission. Many transfusion specialists consider cGMP leucodepleted units as CMV 'safe' in that they are unlikely to transmit CMV disease. In addition to CMV seronegative marrow and solid organ transplant recipients, CMV seronegative or safe components are generally indicated for premature infants, during intrauterine transfusions, for patients with congenital immunodeficiencies, CMV seronegative pregnant females, and seronegative patients with HIV. The British Committee for Standards in Haematology has concluded that leucoreduced components are an 'effective alternative' to seronegative products for preventing CMV transmission by transfusion.

Alternatives to blood component therapy

Autologous transfusion

Commonly used forms of autologous transfusion include preoperative blood donation, acute normovolaemic haemodilution, and autologous blood salvage. Many blood centres provide autologous preoperative blood donation services in which a patient's blood is drawn and stored for later use, usually during a surgical procedure. The criteria for autologous donations are less stringent than those for allogeneic donors. Preoperative blood donation can be utilized in elderly patients, although there is a higher risk of anaemia and more serious cardiovascular complications associated with the donation. Although the use of autologous blood decreases the risk of viral infection, the risk of bacterial contamination remains. Acute normovolaemic haemodilution is performed by removing blood from a patient immediately before surgery and replacing the blood volume with crystalloid or colloid solutions to maintain haemodynamic stability. The withdrawn blood is then later reinfused. Autologous blood salvage is performed by collecting and then returning blood lost during or shortly following operative procedures using intraoperative salvage devices. This technique is primarily employed in cardiac and orthopaedic surgery.

Growth factors

Haematopoietic growth factors used in transfusion therapy are designed to limit the exposure of patients to allogeneic blood. The isolation, characterization, and subsequent synthesis of erythropoietin by recombinant technology (rHuEPO) was one of the most important advances in decreasing red cell transfusions. Use of rHuEPO has dramatically reduced the transfusion needs of patients with renal failure and various anaemias. rHuEPO has also been employed to increase the yield of autologous donations and to stimulate erythropoiesis following surgery. Granulocyte colony stimulating factor (G-CSF) has been shown to decrease infection rates in neutropenic patients undergoing chemotherapy, replacing marginally effective granulocyte transfusions. There is rapid growth in the use of other growth factors including FLT-3 ligand, c-MPL ligand (thrombopoietin, TPO) and various combinations of growth factors. These growth factors have been shown to reduce thrombocytopenia following non-myeloablative chemotherapy. Thrombopoietic growth factors also have the potential to stimulate platelet apheresis donors, increase stem cell harvest yields, and to expand progenitor cells *ex vivo*. Development of neutralizing antibodies against endogenous thrombopoietin has plagued clinical testing of thrombopoietic growth factors.

Blood substitutes

Red cell substitutes currently in development include haemoglobin-based oxygen carriers (HBOCs), perfluorocarbon emulsions (PFCs), and liposome-encapsulated haemoglobin. The two major types of blood substitutes, HBOCs and PFCs, are in phase II and III clinical trials. HBOCs are artificially derived products with oxygen carrying properties. They are structurally similar to haemoglobin but do not contain red cell stroma which is toxic and leads to renal damage. Development of HBOCs has been hampered by the relatively short half-life of these oxygen carriers in the circulation. PFCs are synthetic hydrocarbons that have the ability to carry dissolved oxygen. The particles circulate for only a few hours until they are removed by the reticuloendothelial system. Research efforts to modify or remove red blood cell antigens from donor units is proceeding slowly, but a truly universal compatible red cell unit may one day be within reach.

Further reading

BCSH Blood Transfusion Task Force (1996). Guidelines on gamma irradiation of blood components for the prevention of transfusion-associated graft-versus-host disease. *Transfusion Medicine* **6**, 261–71.

Bowden RA, *et al*. (1995). A comparison of filtered leukocyte-reduced and cytomegalovirus (CMV) seronegative blood products for the prevention of transfusion-associated CMV infection after marrow transplant. *Blood* **86**, 3598–603.

Chapman J, *et al*. (1998). Guidelines on the clinical use of leucocyte-depleted blood components. *Transfusion Medicine* **8**, 59.

Cohn E, *et al*. (1950). A system for separation of components of human blood. *Journal of the American Chemical Society* **72**, 465–74.

Contreras M (1998). The appropriate use of platelets: an update from the Edinburgh Consensus Conference. *British Journal of Haematology* **101** (Suppl 1), 10–12.

Corash L (1999). Inactivation of viruses, bacteria, protozoa, and leukocytes in platelet concentrates: current research perspectives. *Transfusion Medicine Review* **13**, 18–30.

Daniels G, *et al*. (1996). Blood group terminology: from the ISBT Working Party. *Vox Sanguinis* **71**, 246.

Dike AE, *et al*. (1998). Hepatitis C in blood transfusion recipients identified at the Oxford Blood Centre in the national HCV look-back programme. *Transfusion Medicine* **8**, 87–95.

Dobroszycki J, *et al*. (1999). A cluster of transfusion-associated babesiosis cases traced to a single asymptomatic donor. *Journal of the American Medical Association* **281**, 927–30.

Dodd RY (1998). Transmission of parasites by blood transfusion. *Vox Sanguinis* **74**, 161–3.

Ereth MH, Oliver WC, Jr, Santrach PJ (1994). Perioperative interventions to decrease transfusion of allogeneic blood products. *Mayo Clinic Procedings* **69**, 575–86.

Friedberg RC, Donnelly SF, Mintz PD (1994). Independent roles for platelet crossmatching and HLA in the selection of platelets for alloimmunized patients. *Transfusion* **34**, 215–20.

Goldberg MA (1995). Erythropoiesis, erythropoietin, and iron metabolism in elective surgery: preoperative strategies for avoiding allogeneic blood exposure. *American Journal of Surgery* **170**, 37S–43S.

Goodnough LT, *et al*. (1999). Transfusion medicine. Second of two parts–blood conservation. *New England Journal of Medicine* **340**, 525–33.

Hasley PB, Lave JR, Kapoor WN (1994). The necessary and the unnecessary transfusion: a critical review of reported appropriateness rates and criteria for red cell transfusions. *Transfusion* **34**, 110–5.

Hebert PC, *et al*. (1999). A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. Transfusion Requirements in Critical Care Investigators, Canadian Critical Care Trials Group. *New England Journal of Medicine* **340**, 409–17.

Heuft HG, *et al*. (1998). Epidemiological and clinical aspects of hepatitis G virus infection in blood donors and immunocompromised recipients of HGV-contaminated blood. *Vox Sanguinis* **74**, 161–7.

Issitt PD, Anstee DJ, eds (1998). *Applied blood group serology*, 4th edn. Montgomery Scientific, Durham.

Jackson MR, *et al*. (1996). Fibrin sealant: current and potential clinical applications. *Blood Coagulation and Fibrinolysis* **7**, 737–46.

Klein HG, Strauss RG, Schiffer CA (1996). Granulocyte transfusion therapy. *Seminars in Hematology* **33**, 359–68.

Krishnan LA, Brecher ME (1995). Transfusion-transmitted bacterial infection. *Hematology and Oncology Clinics of North America* **9**, 167–85.

Kuter DJ (1998). Thrombopoietins and thrombopoiesis: a clinical perspective. *Vox Sanguinis* **74**, 75–85.

Lackritz EM, *et al*. (1995). Estimated risk of transmission of the human immunodeficiency virus by screened blood in the United States. *New England Journal of Medicine* **333**, 1721–5.

Liles WC, *et al*. (1997). A comparative trial of granulocyte-colony-stimulating factor and dexamethasone, separately and in combination, for the mobilization of neutrophils in the peripheral blood of normal volunteers. *Transfusion* **37**, 182–7.

Lo YM, *et al*. (1998). Prenatal diagnosis of fetal RhD status by molecular analysis of maternal plasma. *New England Journal of Medicine* **339**, 1734–8.

Lundberg G (1994). Practice parameter for the use of fresh-frozen plasma, cryoprecipitate, and platelets. Fresh-Frozen Plasma, Cryoprecipitate, and Platelets Administration Practice Guidelines Development Task Force of the College of American Pathologists. *Journal of the American Medical Association* **271**, 777–81.

Marcus DM (1969). The ABO and Lewis blood-group system. Immunochemistry, genetics and relation to human disease. *New England Journal of Medicine* **280**, 994–1006.

Murphy MF (1999). New variant Creutzfeldt-Jakob disease (nvCJD): the risk of transmission by blood transfusion and the potential benefit of leukocyte-reduction of blood components. *Transfusion Medicine Review* **13**, 75–83.

Murphy S, Heaton WA, Rebulla P (1996). Platelet production in the Old World and the New. *Transfusion* **36**, 751–4.

Novotny VM (1999). Prevention and management of platelet transfusion refractoriness. *Vox Sanguinis* **76**, 1–13.

Pehta JC (1996). Clinical studies with solvent detergent-treated products. *Transfusion Medicine Review* **10**, 303–11.

Popovsky MA, Moore SB (1985). Diagnostic and pathogenetic considerations in transfusion-related acute lung injury. *Transfusion* **25**, 573–7.

Przepiorka D, *et al*. (1996). Use of irradiated blood components: practice parameter. *American Journal of Clinical Pathology* **106**, 6–11.

Race R (1944). An 'incomplete' antibody in human serum. *Nature* **153**, 771.

Reid ME, Yazdanbakhsh K (1998). Molecular insights into blood groups and implications for blood transfusion. *Current Opinions in Hematology* **5**, 93–102.

Schreiber GB, *et al*. (1996). The risk of transfusion-transmitted viral infections. The Retrovirus Epidemiology Donor Study. *New England Journal of Medicine* **334**, 1685–90.

Silliman C (1999). Transfusion-related acute lung injury. *Transfusion Medicine Review* **13**, 177–86.

Snyder EL (1995). The role of cytokines and adhesive molecules in febrile non-hemolytic transfusion reactions. *Immunological Investigations* **24**, 333–9.

Turner ML, Ironside JW (1998). New-variant Creutzfeldt-Jakob disease: the risk of transmission by blood transfusion. *Blood Review* **12**, 255–68.

Vengelen-Tyler V, ed (1996). *Technical manual*, 12th edn. American Association of Blood Banks, Bethesda.

Vogelsang GB, Hess AD (1994). Graft-versus-host disease: new directions for a persistent problem. *Blood* **84**, 2061–7.

Wandt H, *et al*. (1998). Safety and cost effectiveness of a 10×10(9)/L trigger for prophylactic platelet transfusions compared with the traditional 20×10(9)/L trigger: a prospective comparative trial in 105 patients with acute myeloid leukemia. *Blood* **91**, 3601–6.

Wiener A (1943). Genetic theory of the Rh blood types. *Procedings of the Society for Experimental Biology Medicine* **54**, 316.

Williamson LM, Warwick RM (1995). Transfusion-associated graft-versus-host disease and its prevention. *Blood Review* **9**, 251–61.

Winslow RM (1999). New transfusion strategies: red cell substitutes. *Annual Review of Medicine* **50**, 337–53.

Yamamoto F, *et al*. (1990). Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229.

22.8.2 Haemopoietic stem cell transplantation

E. C. Gordon-Smith

Introduction

The idea that haemopoietic stem cells from the bone marrow could be transferred from a normal individual to a patient to replace defective bone marrow has a long history. With the exception of rare instances where marrow was obtained from an identical twin, such attempts in humans universally failed until a clear understanding of the immune processes involved in tolerance and rejection became available. Much of the pioneering work in making possible human bone marrow transplantation was carried out by E. Donald Thomas and colleagues in the United States, work for which Thomas received the Nobel Prize jointly in 1990. In the post-Second World War era, experiments on inbred mice showed that lethally irradiated animals could be rescued by transfusion of bone marrow from unirradiated mice and that this protection was the result of engraftment of the normal marrow in the recipient. Successful engraftment depended upon the donor marrow being genetically acceptable by the recipient mouse or the recipient mouse being sufficiently immunosuppressed. Successful engraftment when there was immunological disparity between the donor and recipient was followed after a period of 2 weeks or so by a 'secondary' disease in which the recipient failed to thrive and developed gastrointestinal disorders and skin abnormalities manifest by poor further development and eventual death from infection. This so-called 'runt disease' is the murine equivalent of graft-versus-host disease (**GVHD**) in humans in which immunocompetent cells from the immunologically disparate donor mount an attack against recipient tissues. From these and other experiments in outbred animals it was recognized that transplantation of bone marrow would carry the special risk of GVHD and that histocompatibility would be a critical requirement for successful transplantation.

Further work in animals demonstrated that certain treatments, in particular total body irradiation and cyclophosphamide, were sufficiently immunosuppressive to permit engraftment, and that GVHD could be controlled to some extent, where there was not great disparity between the histocompatibility antigens of donor and recipient, with methotrexate. The elucidation of the major histocompatibility locus on chromosome 6 in humans, with the identification of the histocompatibility antigens at the A, B, or C (class I) and DR (class II) loci of the HLA system, finally allowed the identification of appropriate donors for human transplantation. The paramount importance of histocompatibility in haemopoietic stem cell transplantation has been confirmed subsequently by extensive clinical practice. The first successful transplant from a non-identical, but HLA compatible, sibling was carried out in 1968 for a patient with severe combined immune deficiency where the underlying disease prevented rejection. Successful allogeneic transplantation from sibling donors in patients who required conditioning with total body irradiation and cyclophosphamide to permit engraftment was carried out in 1969 in Seattle by the group led by Thomas. Many thousands of such transplants have been carried out subsequently, though it would be fair to say that the indications for transplantation, particularly in malignant disease, are not always as clear as they might be and the problems of GVHD, graft failure, and infection remain hazards which contribute to transplant-related mortality. On the other hand, better support with blood products and antibiotics, improved tissue typing techniques, and the introduction of less toxic ways of controlling rejection and GVHD, as well as better selection of recipients, have improved outcomes steadily over the last 30 years.

Histocompatibility complex and haemopoietic stem cell transplantation

The organization of the major histocompatibility complex (**MHC**) on chromosome 6, and its importance in transplantation, is described in detail in Chapter 5.7. The closeness of the relevant genes in the complex means that within families there is little crossing-over in germ line cells and inheritance more or less follows the autosomal pattern, so that the chances of a sibling having the same HLA type as a patient is about 1:4. This is genotypic identity, in which many unidentified sequences are identical by descent between siblings. At each HLA locus there are large numbers of possible alleles in humans leading to a potential of many millions of different histocompatibility profiles. However, within populations, certain HLA alleles tend to be associated and segregate together, 'genetic disequilibrium', so that it is theoretically and practically possible to find phenotypically identical pairs within an unrelated population.

The identification of phenotypes was originally based upon serological testing for A, B, and DR antigens. The introduction of molecular techniques for identifying DNA sequences directly has shown that there may be a large number of HLA gene products whose cognate protein molecules are assigned to the same phenotype by serological methods. Some of these differences are moreover of considerable importance in terms of immunological incompatibility. These observations on the MHC within the population have made possible the establishment of large volunteer donor pools of individuals prepared to supply haemopoietic stem cells, but also highlight the difficulties of unrelated transplants from an immunological point of view. Indeed, even where there appears to be close identity in an unrelated pool, there are likely to be many fine genetic variations. Selection of donors by improved typing techniques has reduced the risks associated with unrelated transplants, but selection has also restricted the range of appropriate donors. It has also become clear that there are very wide variations in the linkage disequilibria at MHC loci between different populations of the world so that a donor pool of one ethnic type may have a much reduced chance of providing donors for another.

Where there is a histocompatibility disparity between donor and recipient, haemopoietic stem cell transplants may be possible, but the incidence of complications rises steadily as the degree of disparity increases. It is also apparent that the antigens of the MHC are not the only antigens which are important in determining the presence or absence of GVHD. GVHD is mediated by CD4+ and CD8+ cytotoxic T lymphocytes, but the role of specific HLA antigens and minor antigens in determining the attack, and the part played by recipient antigens in susceptibility to the disease, have not been worked out in detail. As discussed later, the immunological attack on normal tissues which produces GVHD seems to be linked to an ability to attack abnormal tissues, particularly malignant, producing a graft-versus-leukaemia (**GVL**) effect. Much effort has gone into trying to identify the cells which mediate GVL and to see if they can be separated from those that produce GVHD. So far the results are inconclusive. The problems and benefits of immunological disparity obviously only apply in the allogeneic transplantation procedures and are absent when autologous stem cells are used to restore haemopoiesis after intensive chemotherapy.

Haemopoietic stem cells

The idea that there was a cell in the haemopoietic system which was capable of giving rise to all lineages of the haemopoietic system for life through a process of self-renewal, proliferation, and differentiation of progeny became current in the early part of the twentieth century. Experiments by Till and McCulloch in mice demonstrated that there were individual cells which could give rise to colonies of different haemopoietic lineages in the spleen of irradiated and transplanted mice. Subsequently it was shown that the passage of small numbers of early precursor cells could repopulate the haemopoietic system serially in lethally irradiated mice. It seems probable that a single stem cell can repopulate an entire animal in terms of haemopoiesis and the immune system. In animals, stem cells can be identified by immunophenotyping, purifying this population of cells, and showing that they are capable of haemopoietic reconstitution in a series of lethally irradiated animals. Such experiments in humans are impossible, but the best *in vitro* techniques have suggested that the human haemopoietic stem cell is closely related to precursors that carry an antigen designated CD34, lack other haemopoietic markers including CD33, and have no lineage-specific markers. Whether such cells are truly the most primitive cells that are capable of giving rise to both haemopoetic and immunological precursors is not of practical importance since successful haemopoietic reconstitution, both in allogeneic and autologous transplants, is closely related to the number of such cells present in the donation. The CD34+, CD33- cells represent some 1×10^{-3} to 10^{-4} of the cells of normal human haemopoietic marrow.

Sources of haemopoietic stem cells

In the first 20 years or so of haemopoietic stem cell transplantation virtually all donations were collected from the bone marrow. Animal experiments had demonstrated that marrow infused intravenously into a recipient was capable of repopulating the marrow and this method of delivery was practised from the beginning in human transplantation. Within normal marrow, haemopoietic stem cells are located in specific areas, usually close to the bony trabeculas in the haemopoietic spaces. The observation that marrow infused into the circulation could find its way to the marrow cavity indicated that haemopoietic stem cells were capable of trafficking through the circulation and homing to the appropriate part of the marrow microenvironment. It was also recognized that there were small numbers of stem cells in normal circulating blood, and that this number was increased during the marrow recovery following cytotoxic chemotherapy. The discovery of haemopoietic growth factors and their subsequent production by recombinant technology led to their use in clinical practice. Administration of many of these cytokines, particularly granulocyte colony-stimulating factor (**G-CSF**), granulocyte–macrophage colony-stimulating factor (**GM-CSF**), and stem cell factor increases the number of circulating colony-forming cells and CD34+ cells enormously, such that for a period of a few days following treatment there would be more than adequate numbers in the circulation to use as a source of transplant cells. Homing and mobilization of haemopoietic stem cells seems to be a continuous, dynamic process—even under normal conditions.

In the early development of the fetus, haemopoiesis takes place in the liver, and fetal liver cells have been used as a source of haemopoietic stem cells, mainly for the treatment of inherited disorders characterized by severe combined immune deficiency. The logistics of such transplants, which require 11-week-old fetal livers, make this an impractical approach. However, research on embryonic stem cells suggests that there may be other important sources of stem cells, not only for haemopoiesis but for other types of tissue replacement. Of more immediate practical importance was the finding that cord blood contained large numbers of haemopoietic cells with high proliferative potential and characteristics of stem cells. Cord blood has become a third practical source of donor cells. Each of these sources—bone marrow, peripheral blood, and cord blood—have advantages and disadvantages that impinge on clinical management. A critical requirement for successful transplantation is that there should be a sufficient number of stem cells—the ability to expand stem cells *ex vivo* would solve this and other requirements, but so far this has not proved to be practical for clinical use.

Haemopoietic stem cells from bone marrow

Until about 1993 most transplants were conducted using bone marrow stem cells. Much of the data concerning the success and problems of stem cell transplantation are derived from the use of bone marrow and this remains the principal source of stem cells in allogeneic transplants. Bone marrow is harvested with the patient under general anaesthetic by aspiration from the posterior, superior iliac crests, and if necessary the sternum. Experience showed that some 3×10^{8} nucleated cells/kg from the bone marrow were required for successful engraftment and this usually involved collecting 1 to 1.5 litres of bone marrow mixed, of course, with blood. Donors usually have a unit of blood collected before harvesting, which is returned at the end of the procedure to ameliorate the anaemia. The procedure takes 1 to 2 h and the donor usually requires brief admission to hospital to recover. Serious complications are very rare and are those associated with the general anaesthetic or local complications such as osteomyelitis or abscess formation. The advantage of this source of stem cells from the donor's viewpoint is that collection is rapid with a maximum of 48 h involvement. The disadvantage is the need to have an anaesthetic and the pain or discomfort that follows the procedure.

Haemopoietic stem cells from peripheral blood

Haemopoietic stem cells may be mobilized into the peripheral blood following exposure to granulocyte colony-stimulating factor. For allogeneic transplantation, donors receive G-CSF (filgrastim or lenograstim) at a dose of 10 µg/kg subcutaneously daily for 5 days. The peripheral granulocyte count rises to 30×10^{9}/l or higher and CD34+ cells appear in the peripheral blood reaching a maximum 5 to 6 days after the start of treatment. Leucocytes are collected by cytopheresis with the objective of reaching more than 2×10^{6} CD34+ cells/kg body weight of the recipient. Sufficient cells can usually be collected in one procedure. Attempts to increase the circulating stem cell concentration still further using additional cytokines, such as stem cell factor, have not proved to be sufficiently safe for general use. The main disadvantages for donors of this type of stem cell collection is that of bone pain or ache following the injections of G-CSF and the procedure of cytopheresis. The advantage is the avoidance of admission to hospital and an anaesthetic. When autologous collection of stem cells is required, the concentration of CD34+ cells may be increased further by giving cyclophosphamide (or some other chemotherapeutic agents, such as etoposide) before starting the G-CSF. The recovery from the marrow suppression so produced leads to mobilization of stem cells even without G-CSF. This procedure is used mainly for patients with malignant disease for whom the stem cells can be used to rescue them from the effects of further chemotherapy.

The use of peripheral blood for harvesting stem cells for allogeneic transplants provides high numbers of CD34+ cells and more rapid engraftment than that seen with bone marrow-derived stem cells. On the other hand, peripheral blood contains more T cells than bone marrow and, whilst original concerns that acute GVHD would be unacceptably severe unless T cells were removed has not proved to be the case, chronic GVHD does seem to be more prevalent. Nevertheless, the ease of collection and advantages of rapid engraftment have meant that most autologous transplants, and an increasing proportion of allogeneic, are sourced from the peripheral blood.

Haemopoietic cells from umbilical cord blood

Sourcing haemopoietic stem cells from umbilical cord blood has several theoretical and practical advantages.

Umbilical cord blood is widely available with no risk to mother or infant, there is low viral contamination, the immaturity of the immune cells reduces the risk of GVHD, and the cells may readily be stored frozen. Furthermore, a balance of umbilical cord blood stem cells from different ethnic groups to take advantage of genetic disequilibrium can be achieved and specific HLA types can be targeted. A disadvantage is the relatively small numbers of haemopoietic stem cells that are present, so that cells derived from umbilical cord blood are mainly suitable for child recipients rather than adults; a further difficulty is the lack of any back-up source of cells should the transplant fail or relapse occur. There is also the theoretical risk that the umbilical cord blood stem cells carry some latent genetic defect which might appear years after the transplant.

Plasticity of stem cells

It has become apparent that there are present in the bone marrow, and in other tissues, cells which are totipotent in their capacity to develop into differentiated cells depending upon the molecular and cellular microenvironment to which they are exposed. Thus bone marrow-derived cells may differentiate to cardiac muscle cells, nerve cells, striated muscle fibres, and many other tissues, whether they be ectodermal, mesodermal, or endodermal in origin. This potential is also present in embryonic stem cells. The reconstitution of a whole animal from a single somatic nucleus reinserted into an enucleated oocyte (cloning) is the ultimate indication of plasticity. In the future, haemopoietic stem cells may be used to repair neurological or muscle defects and other sources of stem cells used to prepare haemopoietic deficiencies.

Donors for allogeneic stem cell transplantation

Problems of transplant-related morbidity and mortality, graft rejection, GVDH, and infection increase with increasing donor disparity. HLA-matched sibling donors are not only phenotypically matched for the MHC, but have genotypic identity throughout most of the MHC. This does not eliminate transplant-related morbidity and mortality, but reduces the incidence and severity of the problems compared with unrelated volunteer donors matched phenotypically for the MHC. Sibling donors are therefore preferred. Same sex donors are more successful than mismatched, and transplantation from male donors is more successful than female. HLA-matched sibling donors are only available for about 1 in 3 recipients in populations with an average of two or three children per family. To overcome this shortfall, volunteer donor banks have been established, now including some 3 million typed donors worldwide. This pool can provide HLA-suitable matches for about 80 per cent of recipients with the same genetic disequilibrium as the donor pool, though finding the right match may take several weeks. Even with fully matched donors, either sibling or volunteer, extensive immunosuppression of the recipient is required pretransplant to prevent graft rejection and post-transplant to control GVHD. New methods of immunosuppression which allow the stepwise development of donor marrow may produce a greater degree of tolerance and permit successful engraftment of haemopoietic stem cells with some degree of HLA disparity. Volunteer donor stem cells were the source of about a quarter of all allogeneic transplants in 1999 and this proportion is increasing.

Stem cells from umbilical cord blood banks have been used successfully in transplants for genetic abnormalities, particularly Fanconi anaemia, and also for children with malignant disease. However, this source has proved difficult for adults mainly because of the low numbers of stem cells in the cord blood.

Management of recipients for haemopoietic stem cell transplantation

The treatment of recipients pretransplant includes measures to induce immunosuppression and irradication of diseased bone marrow. This was the theory behind the so-called conditioning regimens used during the first 30 years of stem cell transplantation. For haemopoietic stem cell transplantation for malignant disease, most protocols contained cyclophosphamide combined either with total body irradiation (single dose or fractionated) or with busulphan. For non-malignant conditions, particularly acquired aplastic anaemia, cyclophosphamide in higher dosage, either alone or combined with antilymphocyte globulin (ALG), was the major immunosuppressive agent. Some of the more widely used regimens are indicated in Table 1. The incidence and severity of GVHD was reduced by giving methotrexate intermittently post-transplant. The introduction of cyclosporin to reduce graft failure and ameliorate GVHD greatly improved the results of transplantation. Such conditioning regimens, particularly for malignant and genetic disorders, carry considerable delayed as well as acute toxicity, particularly for children. Where radiation is used and to a lesser extent busulphan, infertility is usual, growth is retarded, and other endocrine functions may be impaired. Late onset of solid tumours also occurs. Where transplantation was used for patients who had already received irradiation or chemotherapy to the central nervous system, for example patients with a relapsed acute lymphoblastic leukaemia, intellectual impairment as well as the above problems are common.

Subsequently it has been recognized that much of the success of stem cell transplantation in certain malignant conditions, most notably chronic myeloid leukaemia but also acute myeloid leukaemia, is related to the immunosuppressive attack (GVL), provided by donor lymphocytes. Likewise the repopulation of marrow by donor haemopoietic stem cells does not require the immediate abolition of recipient marrow. Conditioning regimens have been introduced which do not rely on cytotoxic measures to obliterate recipient marrow and immune system, but which have increasing immunosuppressive effects to allow the gradual reintroduction of donor marrow. Such regimens include fludarabine, a highly immunosuppressive drug that is not very cytotoxic, together with antilymphocyte globulin or monoclonal antibodies that have a specific immunosuppressive effect. Depletion of T cells in the donor preparation, with subsequent later add-back of donor lymphocytes, is also employed. Some examples of these so-called non-myeloablative regimens are included in Table 1. Results using this approach have been encouraging, but long-term follow-up will be necessary to confirm these advantages.

Removal of T cells from donor preparations has long been used as a way of preventing GVHD. Unfortunately, survival rates are not generally improved by T-cell depletion. The benefit of reducing GVHD is balanced by an increasing graft failure and, in malignant disease, by an increase in cancer relapse.

Graft-versus-host disease (GVHD)

Acute GVHD may develop at any time within the first 6 weeks post-transplant. The typical features and classification of severity are shown in Table 2. Grades III and IV of GVHD are an important cause of transplant-related morbidity and mortality. The immunosuppressive effect of GVHD may lead to reactivation of latent viruses, particularly cytomegalovirus, as well as death from fungal or bacterial infections. Liver failure, catastrophic diarrhoea, and gastrointestinal haemorrhage are other direct causes of death from GVHD. Chronic GVHD mainly affects the skin. It may follow acute GVHD or arise *de novo* 6 weeks or so post-transplant. The rash may vary from a mild dryness of the skin in localized areas to a major extensive scleroderma-like illness with progressive ulceration and scarring. Extensive and chronic GVHD is associated with a poor outcome. Examples of acute and chronic GVHD are shown in Fig. 1.

Fig. 1 Skin manifestations of acute and chronic graft-versus-host disease. Acute GVHD: (a) Grade I, skin +, showing typical palmer maculopapular rash (recovered); (b) Grade IV, skin 4+, generalized erythroderma with early exfoliation; liver 3+, bilirubin > 250 µmol/l (fatal); (c) Grade III, skin 4+, bullous desquamation (recovered). Chronic GVHD: (a) Scleroderma-like plaques on hands; (b) Sclerotic scarring on back; (c) Severe ulceration and contracting scleroderma-like skin involvement.

Amelioration of GVHD with cyclosporin plus or minus methotrexate has already been discussed and the role of T-cell depletion mentioned. The optimal regime for prevention has yet to be elucidated. Treatment depends on the use of corticosteroids together with specific anti-T-cell monoclonal antibodies. Management of chronic GVHD with thalidomide has also been tried.

Graft versus leukaemia (GVL)

The observation that patients with leukaemia, particularly chronic myeloid leukaemia, who had allogeneic transplants and developed acute and/or chronic GVHD had considerably less relapse, though not better survival, than patients without GVHD led to the idea that there was a specific GVL effect (Fig. 2). This was confirmed when it was found that patients with chronic myeloid leukaemia who relapsed post-transplant could be put back into cytogenetic and molecular remission by giving them donor lymphocytes in increasing dosage. Sometimes this was associated with an increase of GVHD, but by no means in every case. There seems to be a hierarchy of susceptibility to GVL effect: chronic myeloid leukaemia being the most clear-cut, some effect in acute myeloid leukaemia, less in acute lymphoblastic leukaemia, and uncertain in lymphoma and myeloma. It is not yet clear whether the cells responsible for GVL are identical to those which produce GVHD or whether it is a separate population. Donor lymphocyte infusions now form part of the management plan post-transplant both for the management of relapse and for some of the non-myeloablative regimes.

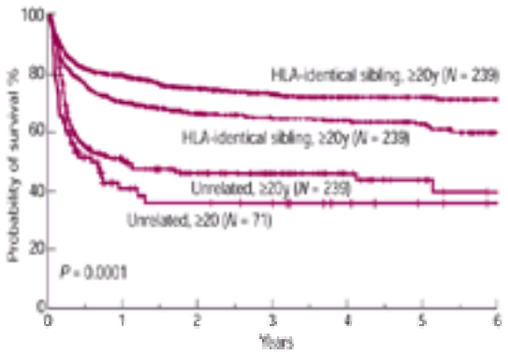

Fig. 2 Probability of survival after allogeneic bone marrow transplantation for severe aplastic anaemia by donor type and recipient age (1991 to 1997). Data from 2064 transplants from the International Bone Marrow Transplant Registry, reproduced with permission.

Indications for haemopoietic stem cell transplantation

The indications for haemopoietic stem cell transplantation fall broadly into two groups. In the first, donor stem cells are used for replacement therapy—a rather crude form of gene therapy for inherited disorders and the re-establishment of marrow function in non-malignant bone marrow failure syndromes. The main indications in

this group are shown in Table 3. In the second group, donor stem cells are used as an adjunct to chemotherapy, both through additional cytotoxicity and biological modification through the GVL effect, in malignant disease. It is in this group that uncertainties remain as to the most appropriate timing as well as effectiveness of allogeneic transplantation. Randomized controlled trials have proved difficult to mount and much of the evidence is placed upon registry data or historical controls. At the same time that the results of haemopoietic stem cell transplantations have improved, the results of chemotherapy have also become better. Nevertheless, particularly in children and younger adults, allogeneic transplantation is widely used with some success, particularly for relapsed conditions. There is a very marked inverse relationship between success of transplantation and age, children having much less transplant-related morbidity and mortality due to reduction in infection and GVHD. Children also tolerate a higher degree of HLA mismatching than adults. The upper age limit for allogeneic transplant has continued to rise as results improve and in some conditions where transplantation is the only hope of cure, for example chronic myeloid leukaemia, patients aged more than 60 years have been successfully transplanted. However, the transplant-related morbidity and mortality at this age is very marked. As would be expected, results of allogeneic transplantation are best in low-risk groups, in first complete remission or with chemosensitive disease, and are worst in relapsed and resistant disease. However, it was in this last group that the potential benefits of allogeneic transplantation were first clearly demonstrated by Thomas and his group in Seattle. In most protocols for the management of leukaemias the inclusion of allogeneic transplantation, where a suitable sibling donor is available, is considered either up-front or as a form of rescue in younger patients. The results of unrelated donor transplants consistently lag behind those of matched sibling donors and whilst HLA antigen-mismatched stem cells are used in desperate situations, success rates decline as transplant-related morbidity and mortality increases.

Indications for autologous transplantation

The use of autologous haemopoietic stem cells for treatment of malignant disease can only be considered a form of rescue from increased chemotherapy since the allogeneic effects which produce GVL do not exist. Where there may be tumour antigens that are amenable to immune suppression, attempts have been made to induce specific immunotoxicity, so far without clear-cut benefit. On the other hand, autologous stem cell rescue does allow greatly increased chemotherapy regimens for lymphoma, myeloma, and a variety of solid tumours with shortening of hospital stay—indeed in some cases treatment can be managed in an outpatient setting—and a prolonged course of therapy with repeated rescue from stored cells. Autologous stem cells will also provide the vehicle for gene therapy once techniques for gene insertion and long-term expression become practical.

Future directions for haemopoietic stem cell transplantation

Transplant-related morbidity and mortality should continue to decline as management of infections, particularly viral and fungal infections, improve. Undoubtedly the plasticity of totipotent stem cells will be explored to treat non-haemological or oncological conditions and both autologous and allogeneic stem cells will be used for specific gene therapy for both acquired and inherited disorders.

Further reading

Laughlin MJ (2001). Mini-Review. Umbilical cord blood for allogeneic transplantation in children and adults. *Bone Marrow Transplantation* **27**, 1–6.

Przepiorka D *et al*. (1995). 1994 consensus conference on acute GVHD grading. *Bone Marrow Transplantation* **15**, 825–8.

Rubinstein P *et al*. (1998). Outcomes among 562 recipients of placental blood transplants from unrelated donors. *New England Journal of Medicine* **339**, 1565–77.

Thomas ED, Blume KG, Forman SJ, eds. (1999). *Haematopoietic cell transplantation*, 2nd edn. Blackwell Scientific Inc., Malden MA.

23.1 Diseases of the skin

T. J. Ryan and R. Sinclair

Introduction

Dermatology is concerned not only with skin diseases but with the Greek ideal of beauty, being confident of 'looking good'. Dermatologists observe the attitudes of parents, schoolteachers, spouses, employers, beauticians, nurses, and others to stigma. Whether it be incipient baldness, the wrinkles of ageing, or tattoos, there are cultural factors to be understood, and a cost of not treating. When is ugliness illness; how much is disfigurement worth in a court of law, and on what does wellbeing depend?

Throughout this section, the impairment, disability, and handicap of skin failure will be emphasized. Dermatology is made difficult by its great variety of physical signs. It is an encyclopaedic subject with more than 3000 named entities. Fortunately, fewer than 10 diseases represent 70 per cent of dermatological practice—acne; bacteriological, viral, and fungal infections; tumours; dermatitis; psoriasis; leg ulcers; and warts.

Good physicians look at the skin while listening to the patient or eliciting physical signs. Recognizing minor details depends not merely on seeing but of knowing their significance. Unfortunately, so much of recognition is the naming of physical signs, and dermatologists have accumulated an enormous amount of jargon. Physicians should know enough to recognize a life-threatening physical sign, such as a melanoma, the malignant pustule of anthrax, or the eroded blisters in the mouth in pemphigus vulgaris. Furthermore, they should recognize signs that are significant indications of systemic disease, such as erythema nodosum, splinter haemorrhages, arsenical keratoses, and the white macules of tuberous sclerosis.

There is no branch of medicine more dependent on observation and less dependent on the laboratory than that of dermatology. However, in few other branches of medicine is there a requirement for the specialist to be so experienced in histopathology. One advantage is that a biopsy can be sent away to experts together with a photograph of the clinical lesion. It is also ideal for telemedicine, easily transmissible to specialist opinion.

In spite of the advances in antimicrobial and corticosteroid therapy, which have completely altered the nature of skin clinics in technically advanced countries, there is no diminution in the number of patients attending for help with skin problems. There is an increase in skin cancer, in the demand for cosmetic treatment, and in the number of environmental agents that damage the skin and cause dermatitis. In developing countries the overwhelming demand is for better management of bacterial and parasitic skin infections, but this is complicated by poverty, malnutrition, poor housing, and water shortage..

The structure of the skin

The skin consists of the epidermis and its supporting dermis lying on a layer of fat (Fig. 1). It is similar to mucosal surfaces where the surface epithelium is separated from its underlying lamina propria by a basement membrane zone which, in turn, is separated from the submucosal fat by the muscularis mucosa.

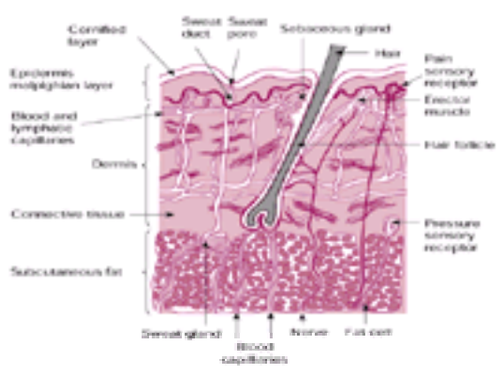

Fig. 1 The structure of human skin.

Unlike in the mucosa or skin of most animals, a subdermal muscle layer only exists in human skin in the areola and scrotum. The epithelium gives rise to all the cutaneous appendages, including: the eccrine sweat glands found over the entire cutaneous surface, but which are more numerous in the palms and soles; the apocrine sweat glands found in the axilla, groin, and beneath the breasts; the hair; and oil-producing sebaceous glands on the upper chest and back.

The epidermis is a stratified squamous epithelium (Fig. 2) comprising a germinative basal layer that is adherent to the basement membrane zone. Through cell division, the basal layer gives rise to successive layers of differentiating cells whose principal function is to synthesize the insoluble protein, keratin. In the process, these cells ultimately die and are shed from the skin surface. Keratins are the intermediate filaments of the epithelial-cell cytoskeleton, which serve as a scaffold in these cells and contribute to cell integrity.

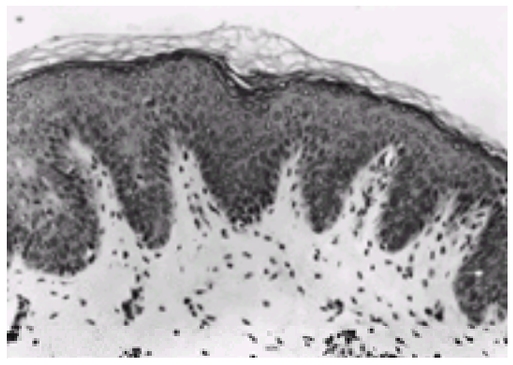

Fig. 2 A section of skin showing the epidermis and upper dermis. Note that the epidermis and the dermal projections interdigitate. The epidermal cells lose their nuclei as they approach the surface.

The epidermis is infiltrated by a number of dendritic cells, including melanocytes, Langerhans, and indeterminate cells. The function of the indeterminate cells is unknown. Skin pigmentation is due to melanin fed into the basal keratinocyte rather than to that stored within the melanocyte. Skin colour is partly due to melanocyte numbers and activity, and partly a reflection of how melanin is stored and processed in the keratinocyte. Melanin is produced from tyrosine and dopamine and acts as a free-radical scavenger.

The rich vasculature has a generous reserve to meet the requirements of wounding and repair, so common at the skin surface. Vasodilatation can increase blood flow by a factor of 200, essential for thermoregulation. Macromolecules and cells leave the dermis through the lymphatic system, which is initiated in an elastic network at the junction of the upper and middle dermis (Fig. 3). The lymphatic system is responsive to hydrostatic forces and to movement of the solid elements of the dermis by massage or compression. The dermis also supports the extremely complex neural network necessary for touch and for sensing danger. All the main constituents of the dermis, collagen, and elastic fibrous proteins embedded in the mucopolysaccharide ground substance, are secreted by fibroblasts.

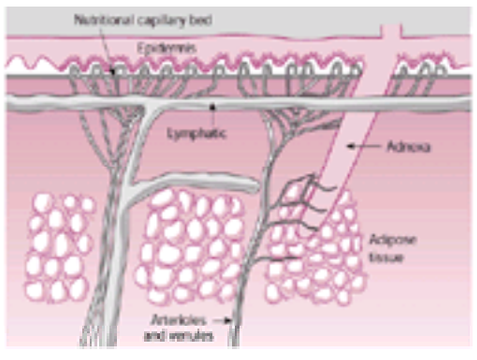

Fig. 3 The lymphatics are the exits from the skin for cells and macromolecules. They control hydrostatic and oncotic pressure and they are one pathway for antigen presentation.

The two functions of basal cells, repair or reduplication and the production of keratin, require that the epidermis should turn over in a controlled way and die in an orderly fashion. Turnover takes about 30 days from the time of reduplication at the basal layer to its loss from the surface. Whereas the lower layers of the epidermis depend on oxygen for mitosis and migration, cells in the upper differentiating layers are anaerobic with no mitochondria. The optimal temperature for epidermal metabolism is probably lower than that for most body cells. Both lipids and carbohydrates provide the energy for epidermal cell metabolism.

Functions of the skin and 'skin failure'

Like the heart, lungs, and liver, the skin can also fail with disastrous consequences. It is the largest organ of the body and, being on the surface, is continuously exposed to injury.

Skin has to be both supple and strong because it is fondled, bent, stretched, trodden upon, and compressed, as well as scratched and prodded. Not only must skin have the capacity to rapidly repair itself to form a physical barrier impervious to excess water loss and to the absorption of damaging environmental agents, it must resist wear and tear. These functions are impaired in people with skin diseases, making them more vulnerable and less able to repair damage, and causing them social embarrass-ment.

Langerhans cells take up antigens and load protein-derived peptides on the major histocompatibility complex (**MHC**), travel to the lymph nodes, and present the antigens to lymphocytes that are preprogrammed to return to the skin. Depletion of these cells, by ultraviolet radiation, prevents contact dermatitis mediated by delayed cellular immunity. This network of cells is the primary immunological defence system of the skin and is termed the 'skin-associated lymphoid tissue' (**SALT**), analogous to the **MALT** (mucosa-associated lymphoid tissue) found in the bowel. The sensation of pain, so finely mediated by the precise innervation of the epidermis, has a similar warning function, helping us to recognize the environment and to itch in the presence of smaller invaders and to follow this with an accurate scratch response. The skin is capable of the presystemic metabolism of drugs and other substances applied topically. It is also capable of forming toxic metabolites. Skin can synthesize vitamin D from calciferol in the presence of sunlight, and contains the enzyme to metabolize it to 1,25-dihydroxycholecalciferol. There is an interaction between the cells in the dermis and those in the overlying epidermis. This interaction is mediated by cytokines, those important in the skin include: interleukins 1, 2, 3, 6, 8, and 10; interferons a, b, and g; and multiple growth factors including epidermal growth factor (**EGF**), fibroblast growth factors (**FGF**) 1 and 2, insulin growth factors (**IGF**) 1 and 2, vascular endothelial growth factors (**VEGF**) 1 to 5, transforming growth factors (**TGF**) a and b, and neurotrophins. Furthermore, a range of peptides, complement factors, eicosanoids, and platelet-activating factors in the epidermis are involved in intracellular communication.

The epidermis contains very high levels of interleukin-1 (**IL-1**), 100 000 times greater than the content of most other tissues. Most IL-1 is produced by keratinocytes and, although this is a continuous process, levels are increased in the presence of ultraviolet light and endotoxins. There is a large intrakeratinocyte preformed pool of IL-1 and a predominantly intracellular inhibitor as a controlling factor by competing for receptors, which are normally scarce in the epidermis but can be induced by ultraviolet rays, trauma, or g-interferon.

The dermis supports the epidermis and its adnexa. Like bone, the skin resists distortion. It is subjected to compression and shearing strains and many mechanical stresses are transduced into biochemical signals. Skin is more supple than bone, and hydrostatic forces or swelling pressure are more finely sensed and distributed.

The dermis, in addition to being a supporting structure, determines many of the characteristics of the epidermis. It is an essential inducer and controller of hair, sweat, and sebum, and provides a selective environment whereby hormones such as oestrogen and testosterone can influence some, but not all epithelial functions; for example, in the pathogenesis of acne hirsutism and androgenic alopecia.

Sexual attraction, being subject to whim and advertising, is an important function on which the fortunes of the cosmetic industry are founded. The social anthropologist has done much to draw attention to what denotes sex appeal; for instance, colouring or decolouring, tattooing, distorting, stretching, and, of course, adorning with jewellery and clothing are all involved. Sex appeal depends on the skin not being too greasy, too matt, or too wrinkled. The White adolescent wants a preparation to reduce a greasy forehead; the Black person wants grease to rid him of any degree of powdery exfoliation. One person must have a beauty spot, and another must not. Some scents attract, while the stink of sweaty feet and rotting shoes repels.

The influence of the psyche

Blushing, cold sweats, and pallor are skin reflections of the mind. Any group of students shown a mite under the microscope will laugh at the sudden awareness of itching it induces in one of their number. The acute inflammatory process mediating a weal or any exudation can be enhanced by anxiety or diminished by relaxation. While a 'neurotic' basis for urticaria, prurigo nodularis, or lichen simplex is no longer overemphasized by terms such as 'angioneurotic oedema' or 'neurodermatitis', modern Western scientific medicine has made such terms unpopular. This is because the influence of the psyche cannot be measured, is mainly subjective, and therefore, by some, is not to be believed. Practitioners of alternative medicine, as well as almost every lay person, recognize a link between anxiety and skin disease.

The principal anxieties resulting from skin disease are the fear of being infectious, unclean, and, ultimately, unwelcome. As with sexually transmitted diseases, a person's upbringing and religious and social mores will often determine their reaction to skin disease.

Few patients will accept that our largest organ can simply wear out or be worn down like the heels of a leather shoe, which after all is only skin. They will, however, believe that their skin disease is due to a malfunction of the liver, an impurity in the blood, to worry, or to a dietary indiscretion. Such beliefs must be countered with tactful explanation.

Occasionally, problems with the psyche will manifest on the skin; for example, Picker's nodules, neurotic excoriations, acne excorieé, trichotillomania, and even onychomania. The relationship of the problem to the psyche is usually acknowledged by the patient; however, patients who refuse to admit to their artefactual dermatitis can be challenging both diagnostically and therapeutically.

A common frustration of patients seeking advice or help is being told their skin problem is trivial. However, when the effect of all chronic or trivial diseases on well being is measured, it is found that the degree of handicap has been belittled. Not all skin disease is psychosomatic.

The handicap of skin disease

Common skin diseases, such as dermatitis and psoriasis, affect the following 'functional specificities' on which personal autonomy depends:

- to move around in and manipulate the environment (Fig. 4 and Fig. 5);

Fig. 4 Psoriasis of the hands interferes with dexterity and makes patients unwelcome in many occupations, such as food handling or public relations.

Fig. 5 A callus or corn is common in ageing skin and pain can make walking very difficult. The patient can be more handicapped than an amputee with a comfortable prosthesis.

- to service oneself;
- to resist normal stresses and traumas;
- to groom oneself;
- to be intimate; and
- to organize oneself emotionally.

Some diseases, for instance leprosy, affect other faculties such as sight. To have personal and economic independence it is necessary to perform effectively in any situation. Skin diseases affecting the hands and feet prevent the patient from getting out and about or from moving around the home (Fig. 4 and Fig. 5). Skin disease, for a variety of reasons, may prevent or threaten the expected care of the home, self, or family, and it often interferes with education and employment.

The threat to life

Absence of skin, as in burns and ulcers, is a common cause of disability and death. Isolation due to rejection by a community is associated not only with poverty, but with infanticide, suicide, and a greater loss of life especially during childbirth. Skin disease does sometimes constitute an emergency and may cause death. Fatal melanoma is not rare, and only human immunodeficiency virus (**HIV**) infection and accidents worldwide are more common causes of death in males aged between 20 and 30 years. There is a 10 per cent incidence of metastasis from squamous epithelioma of the lip, a problem which may increase as actinic damage supersedes pipe-smoking as the major aetiological factor.

Angio-oedema of the upper respiratory tract is the most frightening of dermatological emergencies, accounting for the deaths of most cases of the very rare hereditary angio-oedema due to C1-esterase inhibitor deficiency and other much more common causes of urticaria.

Respiratory obstruction is recorded in other diseases such as epidermolysis bullosa (due to inhalation of 'casts') and Behçet's disease (due to ulceration of the larynx).

Many chronic skin diseases cause death by impairing the skin's ability to protect against adverse climatic conditions, environmental irritants, and infective agents, which all result in fluid loss or increased demands being placed on internal organs such as the heart. Blistering disorders, such as toxic epidermal necrolysis (**TEN**), pemphigus vulgaris, widespread impetigo, or epidermolysis bullosa, are especially threatening.

Erythroderma due to eczema (Fig. 6), psoriasis, or lymphoma commonly results in failure of body temperature control, in a high cardiac output, and, more rarely, in uncontrollable protein-losing enteropathy. Fluid loss and prerenal failure are important and particularly relevant factors in hot countries. In the tropics many die from uncontrolled dermatitis and commonly associated superinfections.

Fig. 6 Some diseases are a threat to life. Exfoliative dermatitis is life-threatening because of fluid loss, heart failure, and loss of temperature control. This patient died following perforation of the small intestine while being treated with steroids.

Restricting employment

Not to be able to resist normal stresses and traumas is a common inconvenience. It accounts for the need for sufferers from atopic eczema, even when in remission, to avoid occupations such as hairdressing, nursing, food handling, and mechanical engineering. Unemployment may be the consequence. Wear and tear of the skin is the most common consequence of work and those who have lowered resistance are unable to work. Some skin diseases present as blisters or as psoriasis in response to even minimal trauma—known as the Koebner phenomenon.

To communicate and to be welcome

The skin is involved in display. Through it we make contact with others. It is observed and touched. If there are defects in a person's skin, observers may not like what they see and will not touch it. Many children with such defects experience insults from other children who refuse to hold hands or play with them. Adults experience more subtle signals, which may prevent a normal sex life and interfere with employment (Fig. 7). Isolation causes premature death.

Fig. 7 Acne vulgaris is a cosmetic disability that makes a teenager feel very self-conscious and unwelcome.

The greatest handicap of all is to be unwelcome. Whether real or merely perceived, it is the commonest social effect of skin disease. The whiteness of the skin of vitiligo, the blood on the sheets, and scale on clothing and furniture left by the person with psoriasis are huge disadvantages. Albinos are outcasts in Africa, while those with severe psoriasis are rejected in the United Kingdom.

Prevalence

An examination of more than 20 000 Americans aged between 1 and 74 years revealed that 60 per cent had significant skin disease (least frequent among children and most common in the old), which often persisted for more than 5 years. In about 10 per cent of cases the condition was a physical handicap: diseases of the hand being the greatest and most frequent handicap. It has been estimated that 6.8 million Americans are handicapped in their social relationships because of a skin condition. Diseases of the skin account for almost half of all reported cases of industrial illness in the United States.

The interview, examination, and investigations

The interview

The following questions form a suitable basis for conducting a dermatological interview:

1. How long have you had it; exactly when did it start; have you had it before?
2. Which part of your skin was first affected; where were you when it started; what were you doing?
3. How did it progress, to what sites, and what was there before?
4. Does it come and go; how long does each individual lesion last?
5. Does it itch; is it painful, tender, numb?
6. Does it develop blisters or clear fluid?
7. Does anything make it better?
8. Does anything make it worse?
9. Have you consulted any one about this? What was their diagnosis?
10. What ointments, creams, lotions, or bath oils have you used? Have you had any medicine or injections?
11. Has anyone else you know got it; does it run in your family; do any other diseases like asthma, eczema, or hay fever run in your family?
12. Have you had any previous illnesses?

The examination

Clinical examination

Undressing, removal of bandages, and, in some countries, even the removal of a hat may be difficult to achieve. In such cases more will be learnt by generally looking at the patient than trying to force compliance. However, the patient must undress when the diagnosis is in doubt.

One should keep looking until something is recognized. Often much of a rash is atypical, but somewhere there should be a classical physical sign. Good lighting is essential; sunlight is best. A magnifying glass is essential for detecting nail-fold telangiectasia, scabies, or crab lice.

Touch assures the patient there is no abhorrence and that contagiousness and uncleanliness are insignificant. Papules are palpable, macules are not. Compression distinguishes purpura from telangiectasia and reveals much about the depth of the lesion and its hardness.

Skin scrapings for fungal mycelia

Skin scrapings are best taken from moist areas since mycelia in dried scales or in the nails may be too desiccated. Scrapings should be placed on a slide and covered with 10 per cent potassium hydroxide, this helps to clear the keratin of extraneous material which obscures the fungus. Gentle heating is helpful, but not essential. In hot climates the rate of evaporation from potassium hydroxide is such that crystals form and it is best to renew the solution regularly.

Finding parasites

A microscope is essential for the diagnosis of mycelia, lice, and other parasites. Vaseline placed over the aperture of a 'boil' raised by bot or tumbu fly larvae may force their emergence since they cannot survive without oxygen. If onchocerciasis is suspected, a new itchy papule can be picked up on the end of a needle, quickly snipped, placed in saline, and examined under the microscope to see whether microfilariae swim out. Scabies mites can be picked out of the end of the burrow on the fronts of the wrists and between the fingers.

Wood's light

Ultraviolet-A rays (UVA, 360 nm) (Wood's lamp) highlight white areas in white skin, as in tuberous sclerosis, and are helpful for identifying *Microsporum audouini* and *M. canis*, which fluoresce green. Erythrasma due to *Corynebacterium minutissimum* fluoresces coral red. Porphyrins in teeth or urine fluoresce pink, and anaerobes such as *Bacteroides melanogenicus* in wounds and ulcers fluoresce red.

Biopsy investigations

The lesion chosen for biopsy should not be modified by excoriation, therapy, or secondary infection. For interpretation, the histopathologist will need some history, such as the site, duration, and appearance of the rash (macular, papular, vasculitic, vesicular), whether it is itchy, and whether the lesion is recent, established, or resolving. A drug history is of particular importance. If possible, a provisional diagnosis and one or two differential diagnoses should be provided. Multiple biopsies from different sites taken from lesions in different stages of development are helpful.

Collection and transport of biopsy samples

Samples for direct immunofluorescence

Direct immunofluorescence is useful for the diagnosis of cutaneous lupus erythematosus (biopsy preferably taken from the centre of the lesion) and vesiculobullous disorders (preferably perilesional skin for suspected bullous pemphigoid, or non-lesional skin for pemphigus or dermatitis herpetiformis).

It is best to provide the biopsy material as a separate specimen rather than dividing it for use in other investigations, this prevents the histological appearance of a crush artefact. The specimen must be received fresh (not in formalin), either in saline-soaked gauze or in an empty container placed inside a larger container containing ice. Ideally, the specimen should be received by the laboratory within 4 h of its removal. A biopsy can remain preserved in saline-soaked gauze for up to 24 h; however, the longer the time, the greater is the risk of a false-negative result. As a last resort, honey is a good preservative.

Samples for microbiology

If infection is suspected, a separate biopsy from the lesion should be submitted fresh in a sterile container. Ideally, it should be placed in sterile, saline-soaked gauze and received by the laboratory within 4 h.

Histological terminology and definitions

The histological report may include the following terms:

- *Hyperkeratosis*: thickening of the horny layer usually resulting from the retention and increased adhesion of epidermal cells.
- *Parakeratosis*: cell nuclei in the horny layer usually resulting from a high rate of cell turnover as in psoriasis.
- *Spongiosis*: separation of cells within the spinous layer by oedema fluid, i.e. the epidermis looks like a sponge—a feature of eczema. Severe spongiosis produces vesicles that may coalesce into blisters.
- *Acantholysis*: loss of cohesion between prickle cells and isolation, and balloon-like appearance of individual epidermal cells, a feature of the blistering disorder pemphigus.
- *Liquefaction*: degeneration and rupture of basal cells—characteristic of lupus erythematosus, lichen planus, and erythema multiforme.
- *Pigmentary incontinence*: the shedding of melanin from the epidermis into the dermis following injury to the basal layer.
- *Elastotic degeneration*: changes in dermal collagen that occur in light-exposed and ageing skin; whorled masses of disorganized elastin-staining fibres replace normal collagen.
- *Fibrinoid degeneration*: deposition of eosinophilic material resembling fibrin.
- *Necrobiosis*: a type of focal necrosis of collagen that leads to the formation of a palisading granuloma, i.e. macrophages lining up like a fence around the necrotic material.
- *Lichenoid*: a heavy infiltrate of white cells hugs the epidermal interface with the dermis and fills the upper dermis.

The basis of rashes

The skin is not homogeneous. It varies in its thickness, rate of epidermal turnover, amount and quality of hair, sebaceous glands, sweat, etc. Some rashes follow the distribution of a particular skin component; for example, of hair follicles in folliculitis (Fig. 8), sweat glands in prickly heat, sebaceous glands in acne vulgaris, dermatomes in herpes zoster, or annular and reticulate patterns as in some rashes determined by the vascular anatomy.

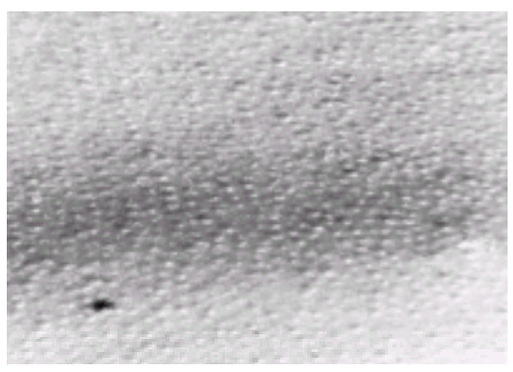

Fig. 8 Perifollicular hyperkeratosis having the distribution of the hair follicle.

Inflammation near the surface of the skin usually damages the epidermis so that vesiculation scaling or erosion become a feature of the response. In contrast, deep dermal or subcutaneous inflammation merely produces 'lumps' known as nodules, and swelling or redness with intact skin markings and no distortion of the epidermis may be the only feature. Rashes may be fundamentally classified into epidermal conditions and dermal conditions: those causing epidermal rashes are included in Table 1, and conditions causing deep dermal rashes are shown in Table 2.

Pityriasis lichenoides, pityriasis rubra pilaris, mycosis fungoides, lichen planus, discoid lupus erythematosus, and Darier's disease are examples of epidermal rashes with specific histology. A biopsy to confirm the diagnosis is recommended if these disorders are suspected. The management of some of these disorders is complex, and a pretreatment biopsy is helpful to document the diagnosis histologically before the morphology of the rash is altered by treatment.

Almost all dermal rashes will require a biopsy to confirm the diagnosis. Occasionally, rashes will be polymorphous with various sites showing epidermal and others dermal change. This is common in epidermal rashes, particularly when they are partly treated. By definition, dermal rashes show no epidermal change whatsoever, even if only 5 per cent of the rash shows epidermal involvement it is still considered an epidermal rash.

The rate of development of the rash is often determined by the type of inflammatory response—oedematous weals or blisters are more acute than white-cell infiltration, purpura, or pustules, and ischaemic necrosis and exfoliation are late responses.

The clinician is a detective and in assessing physical signs must know the sequence of events leading up to what can be seen. The distribution of the rash and its minutest morphology are important. Some classical distributions are shown in Fig. 9 and Fig. 10, while Table 3 illustrates some well-known morphological terms and Fig. 11 and Fig. 12 show some other shapes.

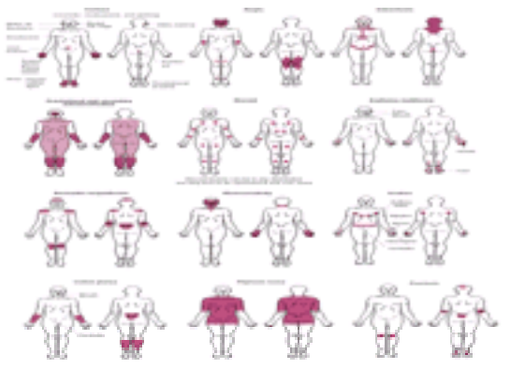

Fig. 9 Distribution of rashes.

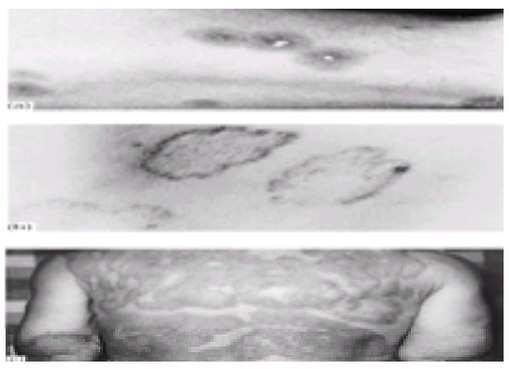

Fig. 10 (a) An example of the 'target' lesion of erythema multiforme. (b) Healing of the centre of the lesion is a feature of many skin diseases, including fungus infections and, in this case, psoriasis. (c) Annular erythema in lupus erythematosus with Ro antibody. This pattern of widespread erythema is also observed in association with underlying malignancy.

Fig. 11 Example of a linear distribution, in this case lichen planus. The distribution does not conform to a dermatome and the exact cause of the linear lesions remains largely unexplained.

Fig. 12 A primary chancre of the lower left eyelid illustrating how skin diseases due to exogenous causes are often asymmetrical.

The management of skin disease requires the elimination of possible agents causing injury and the recognition and treatment of altered host responses. Endogenous

rashes tend to be symmetrical, whereas biting insects and fleas, for example, will produce groups of bites quite indiscriminately. Unlike the rashes of secondary syphilis, the site of the primary chancre is not influenced by host symmetry (Fig. 12). Fungus infections such as cattle ringworm or even *Trichophyton rubrum* in humans are frequently more obvious on one side of the body than another, whereas psoriasis is usually exactly symmetrical.

Injury to the skin from contact dermatitis usually has the distribution of contact; in cases due, for example, to mascara, gloves, or shoes, there will be symmetry (Fig. 13), but casually brushing against a noxious plant will produce bizarre asymmetrical patterns. Scratching spares the centre of the back, and a completely clear area between the shoulder blades when the rest of the body is covered with scratch marks (Fig. 14(a)) suggests that the cause of the rash is the injury done by such scratching. Scabies mites do not seem to like climbing about in hairy areas, so usually spare the head but favour between the fingers, the front of the wrists, or the glans penis.

Fig. 13 Occasionally, symmetry in the distribution of contact may be due to a symmetrical application as in the case of this glove dermatitis.

Fig. 14 (a) The central area of the back is spared from this dermatosis induced by scratching, simple because the patient is unable to reach the site. (b) External irradiation from the sun spares the area beneath the lobes of the ear and under the chin in this case of solar dermatitis. (c) Small islands of unaffected skin scattered throughout a generalized redness and keratoderma are characteristic of pityriasis rubra pilaris.

External irradiation from the sun spares skin beneath the lobes of the ear and under the chin (Fig. 14(b)), whereas an airborne pollen dermatitis will not spare such areas but may have a similar cut-off point below the collar. Small islands of normal skin in a generalized erythroderma are characteristic of pityriasis rubra pilaris (Fig. 14(c)).

Recognizing the signs of exogenous injury make it easier to eliminate the cause. Unfortunately, much skin disease is due to altered host responses, known as 'vulnerability'.

Vulnerability

This is a common characteristic in skin disease, and is seen in dermatitis due to the irritants affecting vulnerable atopic skin. It is seen in the haematogenous localization of immune complexes or other agents at sites altered by previous injury. It is also seen in the Koebner phenomenon, a term used to describe the development of psoriasis, warts, or lichen planus when the skin is injured to a degree that, in most people, would produce no more than a temporary wound but in predisposed individuals results in a recognizable skin disease.

Vulnerability is well worth recognizing because it may be possible to treat the predisposition when it may not be possible to eliminate the trigger. Thus, those whose skin breaks down too easily from unavoidable exposure to solvents may be helped to retain their job by liberally applying emollients.

Recurrent episodes of vasculitis in the legs due to immune complexes may be reduced by more frequent elevation of the legs, the use of supportive bandages, and the avoidance of cold environments. Vulnerability in the legs is due to the chronic stress of blood stasis and venous hypertension, which can be shown to cause inhomogeneity of capillary vessel patterns, adhesiveness of endothelium to leucocytes, and reduced fibrinolysis.

The ecology of the skin, with its integrated, well-balanced interaction between bacteria and surface secretions, also determines the skin's response at its interface with the environment. Erythrasma, pitted keratolysis, pityriasis versicolor, and seborrhoeic eczema are partly constitutional and partly due to exogenous organisms. The seborrhoeic diathesis is poorly understood but such persons seem especially vulnerable to colonization by immunogenic organisms.

Factors determining or modifying skin disease

Changes of skin with age, gender, and race

Newborn–childhood

Birthmarks are usually first noticed in the newborn but some, like cavernous haemangiomas, may not be present on the first day of life. Certain epidermal or so-called 'congenital-type' pigmented naevi do not appear until puberty. In type 1 neurofibromatosis, café-au-lait macules and Lisch nodules may not appear until the child is 5 years of age, axillary freckling is uncommon before the age of 3 years, and neurofibromas tend not to appear until puberty. Only the plexiform neurofibromas present at birth. Some birthmarks have important diagnostic significance indicating serious systemic disease. Examples are the hypopigmented lesions of tuberous sclerosis (see Fig. 65) and the telangiectatic lesion of the Sturge–Weber syndrome.

Fig. 65 Typical oval or 'leaf-like' hypopigmented lesions of tuberous sclerosis. They are present at birth.

Puberty

Secondary sexual characteristics develop at puberty, and at the same time an increase in susceptibility to apocrine diseases, sweating, and blushing is characteristic. Acne vulgaris is mainly a problem for the teenager. Certain diseases such as ichthyosis and eczema tend to improve, while others such as herpes simplex infection and psoriasis are more common. Naevi, particularly pigmented ones, tend to become more prominent.

Pregnancy

See Chapter 13.13.

Old age

Skin diseases in old age are common and reduce the quality of life. Most elderly people have multiple skin problems, including seborrhoeic eczema, intertrigo, and dermatophytosis. Probably the principal characteristic of elderly skin is its inhomogeneity or the increased diversity that develops with age. Some changes are endocrine-related, such as hirsutism and baldness. Others are more specifically age-related, like dryness, decreased sweating, or poor healing of superficial wounds. Dry, scaly, rough skin occurs in about 80 per cent of people over the age of 75, as well as disparities in the size and thickness of the epidermis and in its pigmentation. Seborrhoeic warts are universal and actinic injury, Campbell di Morgan spots, and dilatation and derangement of superficial venules are common. For reasons that are still obscure, some diseases are age-related: for example, pruritus, pemphigoid, and lichen sclerosus et atrophicus.

Degenerative disease and the cumulative exposure to solar radiation explains neoplasia of the skin. Degenerative disease of the vascular system explains venous ulcers and arterial ischaemic diseases. In one study, after controlling for age, sex, and sun exposure, premature wrinkling increased with years of smoking. Heavy smokers were 4.7 times more likely to be wrinkled than non-smokers.

Race

Differences in populations are partly explained by genetic factors, but so much adaptation to the environment occurs that customs and diet may determine some attributes. Although it is frequently reported that certain diseases are absent in tropical climates, this is probably because they have never been looked for or recognized. Erythema is violaceous so that purpura may be difficult to detect in people with dark skins; minor skin problems may not be complained of in the tropics, where many neoplastic and inflammatory diseases are so florid and attendance for advice is so often delayed. A move to a more temperate climate is often associated with urbanization, which can equally influence the skin. The most easily recognized difference between one person and another is their skin colour, and the consequences of sun exposure are much reduced in those with dark skins. Vitiligo is probably more common in the Caucasian races of the Middle East, North Africa, and India. The Japanese rather readily seem to develop a slate-blue or ashy discoloration of the trunk following inflammatory disease. On the other hand, acne vulgaris is uncommon in Japanese people, while both acne and rosacea seem to be uncommon in those with black skins. However, comedone formation due to cosmetics without full-blown acne is common in black skin. Blackness is due to more evenly dispersed, larger and less degradable pigment granules. The stratum corneum of black skin is more compact with a higher lipid content and is subject to less penetration by irritants. Another easily recognized factor is hair size and shape. Facial hirsutism is rare in Japanese women and relative sparseness of hair is a feature of mongoloid races. On the other hand, Mediterranean and some Indian races seem to be particularly hirsute (Fig. 15). The shininess of black skin is partly due to sebum and partly to thermal stress which encourages increased eccrine sweating. Such skins tend to become rather dry in a temperate climate. Scales show up on dark skins. Keloids are a considerable problem for Afro-Caribbeans and can sometimes be massive. Susceptibility to infection depends on immunological factors and on previous exposure. As with malaria or syphilis, some populations seem to acquire a genetic resistance to tuberculosis and leprosy.

Fig. 15 Hair growth on the forehead of an Indian child. This is entirely within normal limits and is of racial origin.

Is it contagious (Table 1, Table 2, Table 3, Table 4, Table 5, Table 6 and Table 7)?

Physicians are often asked whether a skin disease is 'infectious'. The questioner really means, 'Did I catch it?' 'Can I give it to someone else?' 'Is the treatment of choice a simple antiseptic or antibiotic regimen?' The physician may ask, 'Am I missing something which is a danger to other patients in the the ward or to my nursing staff?' Infections are often present in several members of the family at the same time (Fig. 16).

Fig. 16 Infections such as impetigo are highly contagious and tend to be found in more than one member of the family, as in these triplets.

There are many infections, dealt with elsewhere in this textbook, in which a highly virulent organism has broken the defences of a normally resistant host, but there are also organisms that are usually harmless but occasionally, because of immunosuppression or other changes in the host, produce a rash. Pityriasis versicolor, candidiasis, erythrasma, and trichomycosis axillaris, all discussed in Section 7, are examples. More difficult is the relationship with staphylococcal or streptococcal bacteria—although these generally sit in silence on the skin, they are unwelcome in a ward full of more susceptible and vulnerable patients. Psoriasis is not infectious, but the massive exfoliation from such a patient is a great source of cross-infection. Bacterial spread by skin scales is generally considerable and the basis of surgical gowning. Few would feel bound to treat every patient with psoriasis for bacterial infection, but the same degree of infection in atopic eczema is thought to be contributory to the disease, perhaps through a bacterial superantigen effect.

Pathology from skin infection is more common in hot humid climates, and erosions from scratching, prickly heat, and other infections (such as lice or scabies) predispose to boils and other patterns of pyoderma, especially in the groins and axillas.

The primary pathology of infection is often asymmetrical, but an immune response attempting to get rid of it is usually exactly symmetrical and takes 5 to 10 days to develop.

The most difficult diagnostic problem is that of viral disease. The hospital doctor is not well placed to recognize its variety. Rather, it is the general practitioner called to the patient's home who sees virus disease in its early stages or in its transient phase. It is essential to know what rashes are currently endemic.

People with rashes due to infection commonly have an associated fever, lymphadenopathy, coryza, diarrhoea, vomiting, hepatomegaly, or headache. However, the abrupt sterile pustulation of generalized pustular psoriasis (Fig. 17) or the painful deep swelling of delayed-pressure urticaria or vasculitis will also be accompanied by high fever and a neutrophil leucocytosis, but usually there is no lymphadenopathy in these non-infectious processes. Erythema multiforme, Sweet's disease, and toxic epidermal necrolysis similarly show great systemic effects. Since people with widespread skin disease may be unable to control their body temperature, high fever in such persons is not necessarily a sign of infection.

Fig. 17 Pustules are not necessarily due to infection. These pustules are from pustular psoriasis and are sterile.

When the diagnosis is in doubt, good practice is to take adequate swabs and specimens for culture and histological examination, and to treat and touch the patient as his or her comfort requires. Washing the hands suffices to avoid the transmission of scabies, fungus, and most bacterial diseases as well as warts and syphilis. However, practitioners should take the utmost care to avoid inoculating their skin when taking scrapings or biopsies. Patients with much exfoliation should not be nursed on a general ward, but in a single cubicle.

Pustules need not be caused by infection; for instance, the primary lesions are always sterile in psoriasis or an irritant folliculitis from oils. Vesicles need not be due to viruses since they are a feature of papular urticaria, dermatitis hepetiformis, and vasculitis (see Fig. 83). Dark skins exposed to much oil and cosmetics often have a chronic pustular dermatosis of the lower legs that may be sterile.

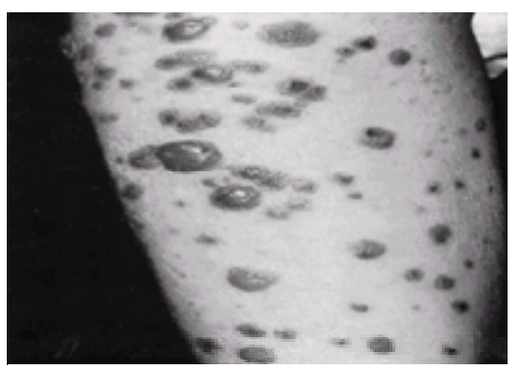

Fig. 83 The presence of blisters in vasculitis is due to the intensity of the oedema in the upper dermis; sometimes it is due to necrosis.

Humidity is a principal cause of profuse skin infection, and treatment by cooling and drying has always been a standard therapy for infected eczema. The fact is that drying is promoted by the use of wet dressings and the consequent evaporation. Wet dressings that are occlusive and changed infrequently encourage infection: ideally, they should be changed every 2 to 4 h. Occlusive surfaces, for instance between the toes, the groins, and the breasts, need to be treated with drying agents (such as those commonly present in deodorants (aluminium chloride)) or with powders. Dry mopping of the ear in otitis externa is similarly helpful.

In some parts of the world, skin clinics are overwhelmed by massive numbers of patients suffering from scabies (Fig. 18), staphylococcal and streptococcal infection, and dermatophytosis. Control is impossible because reinfection is inevitable. Soap and water do much to reduce the incidence of common dermatoses, but water is too valuable to use for washing when there is a drought.

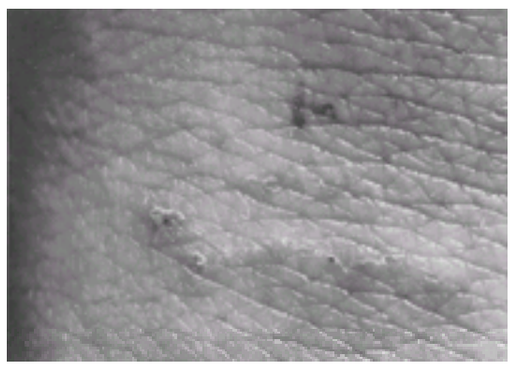

Fig. 18 The diagnostic feature of scabies. The burrow of the mite in the horny layer of the epidermis. The dark spots are haemoglobin in the belly of the mite.

In Mediterranean countries, ringworm of the scalp would be easy to manage (Fig. 19) were it not that the population explosion provides more children for infection than it is possible to treat, and that subclinical infections are difficult to recognize.

Fig. 19 Multiple exudative lesions due to tinea capitis.

Is it hereditary?

See Chapter 23.2 for a complete discussion.

Is it due to malnutrition?

Skin diseases resulting from malnutrition have been termed the 'dermatoses of the poor'. Although they are common in starving communities, they are also seen in those living only on drugs or alcohol, those suffering from malabsorption syndromes, and those debilitated by neoplasia or severe chronic infections. Increasingly, elderly patients suffering from dementia are responsible for more cases in Western urban communities. Poor personal hygiene and lack of, or failure to use, water supplies contribute to some aspects of skin diseases in malnutrition, as well as to the infections of both skin and mouth which often accompany them.

The skin makes up 8 per cent of body weight and uses up about one-eighth of the body's protein; hence it is affected early in malnutrition.

In experimental malnutrition and in studies of people during the Second World War, early signs were dryness of the skin and hyperpigmentation. At birth, malnutrition is seen as loss of vernix and maceration. At all ages the skin is wrinkled and peeling with deficient subcutaneous fat. Older persons proceed to a mild ichthyosis and the associated hyperkeratosis is often a sign of slow turnover. The dry scale is well knit and retains pigment, and histologically may be dense and homogenized. The stratum corneum is unsupple and cracks appear in the horny surface, particularly on the front of the legs (Fig. 20). It is known as eczema craquelée and such eczema that develops is often well marginated, unlike other forms of endogenous eczema.

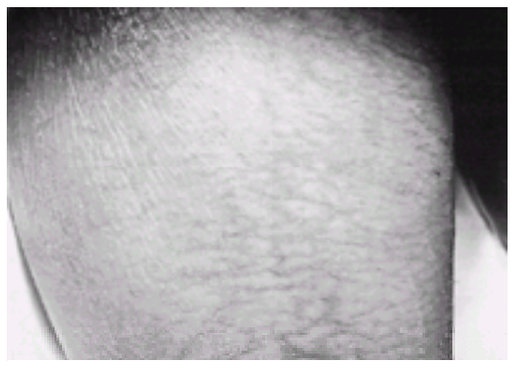

Fig. 20 An early and common sign of malnutrition of the skin, especially in the elderly, is cracking of a well-made stratum corneum giving a pattern of eczema craquelée.

Most malnutrition is a consequence of mixed deficiencies including protein loss. There is weight loss, weakness, and emaciation. Anaemia, oedema, sore tongue, and dry, thin hair are often features.

Vitamin A deficiency should be thought of when there is significant dryness of the eyes and perifollicular hyperkeratosis. It is the commonest preventable cause of blindness.

Vitamin B deficiency causes a dermatitis that has a seborrhoeic distribution, particularly of the nasolabial folds, scrotum, and vulva. The lips are dry, cracked, crusted, or ulcerated; the tongue is sore and smooth.

Nicotinic acid deficiency or pellagra causes the well-known triad of dementia, diarrhoea, and dermatitis. Early signs are prominent sebaceous follicles of the nose. The light-sensitivity dermatosis is also exacerbated by heat, friction, and pressure. The erythema is a characteristic dusky brown and the dermatitis is well marginated. In dark skins the lesions are relatively depigmented but equally well marginated.

Vitamin C deficiency causes perifollicular haemorrhages, painful bruising, or woody oedema of the legs. This means that they look oedematous but are hard to the touch. In a dark skin it may appear that the skin is stretched and shiny. Although coiled hairs are an early sign, they are common in the normal population, especially in the elderly. Swollen and bleeding gums are an important sign but occur only in those with teeth. Vitamin C deficiency should be considered in any non-healing

wound.

Protein deficiency is common in all forms of malnutrition. However, a characteristic disease is recognizable when the protein deficiency is supplemented by carbohydrate and there is no active starvation. In children this is typified by kwashiorkor. Features of protein deficiency include:

1. erythema as in a second-degree burn;
2. dry hyperkeratotic hyperpigmented scales;
3. peeling like enamel paint, cracking like crazy paving;
4. skin signs are maximal over pressure areas; and
5. there is straightening and reddening of the hair.

In some dark skins, raised annular patches of pigmented scales on the trunk are an early sign of malnutrition. This is known as pityriasis rotunda.

Management includes avoiding secondary deficiencies, since the sudden provision of some but not all the necessary foods may precipitate conditions like blindness from vitamin A deficiency. In malnutrition, zinc may be lacking or poorly absorbed, leading to alopecia, diarrhoea, glossitis, and an eroded perioroficial rash called acrodermatitis enteropathica. Some improvement in the rash of kwashiorkor has been described using local zinc ointments, and with the prescription of other trace elements such as selenium.

Is there an association with a systemic disease?

There is a number of associations of skin disease with diseases of the gastrointestinal tract and haematological, cardiovascular, respiratory, renal, and central nervous systems. However, no completely satisfactory system for listing these has been devised. Many of the skin diseases are discussed more fully in other sections.

Gastrointestinal system

Oral mucosa

Leucoplakia may be associated with Darier's disease, pachyonychia congenita, or simple white sponge naevus. It is premalignant in dyskeratosis congenita. In mucocutaneous candidiasis oral candida is associated with nail dystrophy, alopecia areata, and endocrinopathy. Oral hairy leucoplakia is a sign of the acquired immunodeficiency syndrome (**AIDS**). Major aphthous ulcers are a feature of Beçhet's syndrome, while minor aphthae are found in systemic lupus erythematosus, Crohn's disease, as well as iron and folate deficiency states.

Oesophagus

Bullae are common in epidermolysis bullosa, and occasionally the entire epithelial lining of the oesophagus may be coughed up as a cast. Bullae also occur in pemphigus and mucocutaneous pemphigoid. The superficial erosions that follow rupture of the bullae in pemphigus heal without scarring, while scarring and stricture formation complicate healing in mucocutaneous pemphigoid and epidermolysis bullosa. Stiffness and loss of peristalsis frequently occur as an early sign in scleroderma, best demonstrated by a prone barium swallow. Carcinoma of the oesophagus has been associated with plantopalmar hyperkeratosis (tylosis). Webbing of the postcricoid region with anaemia is associated with dyskeratosis congenita—an atrophy of the skin and nails.

Stomach

Pernicious anaemia is an organ-specific autoimmune disease leading to atrophy of parietal cells that clusters with vitiligo and alopecia areata. Carcinoma may present with acanthosis nigricans (Fig. 105) and tripe palms. Gastric polyposis is associated with perioral and finger lentiginoses in the Peutz–Jeghers syndrome, as well as with nail dystrophy and alopecia in the Canada–Cronkite syndrome.

Fig. 105 Acanthosis nigricans: a darkening and thickening of the skin with a tendency to papilloma formation. The angles of the mouth are often involved, as in this patient with carcinoma of the lung.

Gastrointestinal bleeding is a consequence of telangiectasia in hereditary haemorrhagic telangiectasia as well as in acrosclerosis with telangiectasia. It may also occur in disorders of elastic tissues such as Ehlers–Danlos syndrome or pseudoxanthoma elasticum. Henoch–Schönlein purpura usually causes lower gastrointestinal bleeding.

Malignant atrophic papulosis (Degos' disease) is a rare vasculitis of the skin, gastrointestinal tract, and brain. The skin lesion is a porcelain-white punctate scar and the viscera suffer from infarction.

Small bowel

Regional ileitis may present with granulomatous swelling of the buccal mucosa or lips as well as with perianal granulomas and fistulas (Fig. 21). Erythema nodosum, oral aphthous ulcers, and pyoderma gangrenosum are also associated with Crohn's disease, along with any secondary skin changes due to malabsorption.

Fig. 21 Perianal granuloma in Crohn's disease.

Dermatitis herpetiformis is associated with subclinical coeliac disease and may be complicated by small bowel lymphoma. Pigmentation and malnutrition of the skin is particularly recorded in Whipple's disease. Bowel bypass syndrome due to anatomical blind loops may present with widespread pustules and vasculitis ulcers. Metastatic carcinoid syndrome produces characteristic flushing.

Colon

Ulcerative colitis is responsible for many disorders of the skin and mouth, but aphthous ulcers are more common here. Rashes include erythema multiforme, erythema nodosum, and pyoderma gangrenosum. Perianal abscesses and fistulas are also common associations.

Dermatomyositis is most commonly associated with carcinoma of the large bowel.

Pancreas

Paraneoplastic migratory thrombophlebitis (Trousseau's sign) is most likely to be associated with carcinoma of the pancreas. Acute fat necrosis of the trunk or limbs is a consequence of acute pancreatitis. There is an increased electrolyte concentration in the sweat of patients with cystic fibrosis.

Glucagonoma produces the characteristic eruption of necrolytic migratory erythema. The skin lesions are dusky red, annular, and scaly with a vesicopustular element due to epidermal-cell necrosis in the most superficial layers of the epidermis. In addition to the skin changes seen as a complication of diabetes, diabetes mellitis is directly associated with a number of skin disorders. Diabetic dermopathy produces hyperpigmented dull-red papules with superficial scale on the shins of 30 to 60 per cent of patients with diabetes. It heals with atrophic brown scars. Diabetic thick skin is also common, and may manifest as a generalized process in 20 per cent of cases or be localized to the neck and upper back (scleroedema of Bushcke), the fingers (Huntley's papules), or the back of the hands (cheirarthropathy). Acanthosis nigricans and anogenital pruritis are common in obese insulin-resistant diabetic patients, while generalized granuloma annulare, diabetic bullae, lipoatrophy, diabetic yellow skin, and perforating disorders are all uncommon.

Necrobiosis lipoidica diabeticorum (**NLD**) (see Fig. 114) is a characteristic eruption consisting of yellowish to red-brown plaques on the shin, which eventually become atrophic in the centre and may ulcerate. Potent topical steroids or intralesional injection of triamcinolone may be required to control NLD, but care is required as they have a tendency to aggravate the atrophy. NLD occurs in 0.3 per cent of diabetic patients and pre-dates diabetes in 30 per cent.

Fig. 114 Necrobiosis lipoidica usually affects the skin of the shins. The yellowish atrophic plaques are associated with diabetes mellitus.

Although diabetes may be complicated by skin infections such candida or bacterial furuncles, a dermatophytic infection is no more common than in non-diabetic patients. Hyperlipidaemia may be associated with xanthoma, neuropathy with foot deformity and ulceration, and angiopathy with cold, pale hairless legs. NLD, diabetic dermopathy, the erysipelas-like erythema seen on the legs and feet of elderly people with diabetes, and the diabetic rubeosis of the face are all thought to be due to microangiopathy.

Liver

The skin consequences of liver disease include spider naevi, palmar erythema, purpura and bruising, white nails, and clubbing of the fingers. There is loss of hair in the beard, axillas, and pubic region. Gynaecomastia, acne, Dupuytren's contracture, xanthoma, jaundice, pruritus, and pigmentation are other features.

A number of patients presenting with porphyria cutanea tarda or lichen planus will be found to be infected with the hepatitis C virus. Hepatitis B infection is associated with polyarteritis nodosa and the childhood exanthem of pink palpable lesions named the Giannoti Crosti syndrome. A proportion of patients with non-infectious or autoimmune hepatitis will have cutaneous features of lupus or sarcoidosis. Haemochromatosis produces diffuse skin pigmentation and patients may develop hepatocellular carcinoma.

Other systemic manifestations in the skin

These include renal disease in cutaneous vasculitis and lupus erythematosus. Cardiovascular disease and skin disease occur with carcinoid, or secondary to amyloid, scleroderma, as well as subacute bacterial endocarditis, which may produce nodules in the skin—Osler's nodes. Myxoma is recorded with pigment anomalies of the skin in the **LAMB** (lentigenes, atrial myxoma, mucocutaneous myxomas, blue naevi) and **NAME** (naevi, atrial myxoma, myxoid neurofibroma, ephelides) syndromes. Disease occurring in the CNS is observed in Beçhet's syndrome, sarcoid, lupus erythematosus, and with vascular stenosis and livedo reticularis (Sneddon's syndrome). There is also respiratory failure in sarcoid, Churg–Strauss vasculitis, and asthma is common in atopy. Haematological associations include Sweet's acute febrile neutrophilic dermatoses, pyoderma gangrenosum, leukaemia-associated genodermatoses, leukaemia cutis, B-cell lymphoma, and Hodgkin's disease.

Is climate responsible?

Heat, cold, food, and water are all dependent on the climate. The management of skin disease requires washing, soaking, and adequate nutrition as well as control of body temperature. Children and the newborn are particularly susceptible.

Humidity explained why, 70 per cent of lost combat man-days in Vietnam during the rainy season were through skin disease. The distribution of water determines the ecology of many human parasites, such as biting insects that thrive in the rainy season. Wet clothing can cause severe discomfort (particularly inside a boot, around the waist, or between the legs) while marching. Even in the Arctic, occlusive clothing can accumulate much sweat and make walking impossible. 'Immersion foot' and 'paddy foot' can bring a military campaign to an end. In Kuwait, outbreaks of industrial dermatitis were blamed on the absorption of allergens by the skin that become moisturized in certain seasons, while in Scandinavia a low humidity in some factories accounted for drying of the epidermis and consequent irritant dermatitis.

Seasonal variations not only account for increased bacterial injury and epidemics of viral exanthems, but also for eczema; for example, in the atopic patient sensitive to sunlight, or the allergic contact dermatitis due to handling plants seen so often in market gardeners and florists. Sweaty feet in hot weather increase the dermatitis from footwear, and sweat-pore occlusion encourages widespread bacterial infections in extreme heat. The incidence of some disease is influenced by height above sea-level and by the thickness of the atmosphere. People are less likely to be sun-burned at the low level of the Dead Sea where UVA greatly predominates over

UVB, but actinic dermatitis is common in Mexico and in the Andes. Many infections are most exuberant at sea-level. At the slightly higher level of 600 to 1500 m, transmission of leishmaniasis and onchocerciasis by flies is more common. Many of the skin diseases caused by infections with a unique geographical distribution are discussed in Section 7, including pinta, buruli ulcer, or deep mycoses. In this chapter they are only mentioned if they are important in the differential diagnosis of some physical sign, such as depigmentation, wartiness, and blisters.

Both cold weather and low humidity predispose to irritant dermatitis; for instance, the high incidence of dry skin in hospital is explained by the central heating and the very light clothing worn. Pediculosis is encouraged when people huddle together to keep warm.

While much is said about changes in the world's climate, less is said about changes in the skin's microclimate. These are brought about by changes in home heating and bed linen, as well as by clothing, including footwear. Duvets encourage perspiration and can exacerbate nocturnal itch and scratch. Plastic-soled shoes also increase foot temperature and perspiration and lead to tinea, pitted keratolysis, or juvenile plantar dermatosis. The skin, like antique wooden furniture, suffers from contemporary Western overheating and the resultant drying out. Dermatitis is one consequence. The second commonest environmental cause of neonatal mortality is hypo- or hyperthermia.

Cold

Every polar explorer who is inadequately protected will suffer frostbite, snow blindness, and even death. Although the majority of people in more temperate climates do not die of cold, individual susceptibility to its effects varies. A high incidence of skin disease can be attributed to inadequate protection against minor degrees of cold injury. Vasoconstriction and increased blood viscosity mediate internal disease.

It is often noted that residents of the United Kingdom have pink cheeks and blue hands to a degree not seen in, for instance, Australia or the United States. This is because of chronic exposure to cooling. In Canada or Scandinavia where the winters are a danger to the unprotected, there would be no such exposure of the schoolchild or teenager as seen during the winter in the United Kingdom, where 10 per cent of the population are affected by chilblains, acrocyanosis, Raynaud's phenomenon, and the various manifestations of perniosis, an incidence never approached in most other parts of the world.

Chronic cold causes thickening of the subcutaneous and dermal tissues, as in pigs. During the miniskirt era, girls' thighs regularly became fatter in temperate climates. Fat insulates the surface of the skin from the inside, so cooling of the surface is obvious. Chronic cooling causes telangiectasia, which is often perifollicular, and sometimes even angiokeratoma. Pink cheeks are one consequence, but similar changes may be seen over the fat of the calf or upper arm. Cooling causes stasis in the venules so that circulating noxious agents, such as immune complexes and bacteria, usually localize and deposit at such sites.

The anatomy of the skin vasculature is such that cooled skin often shows a pink and blue mottling known as cutis marmorata. If the changes are fixed and do not reverse with warmth, for example in a hot bath, it is then known as livedo reticularis (Fig. 22(a)). This is commonly seen in collagen vascular diseases such as lupus erythematosus. Much disease is localized in the venules of such damaged vasculature. Chilblains or perniosis is essentially a cold-induced ischaemia. Pressure from tight clothing often encourages the damage done by cooling (Fig. 22(b)).

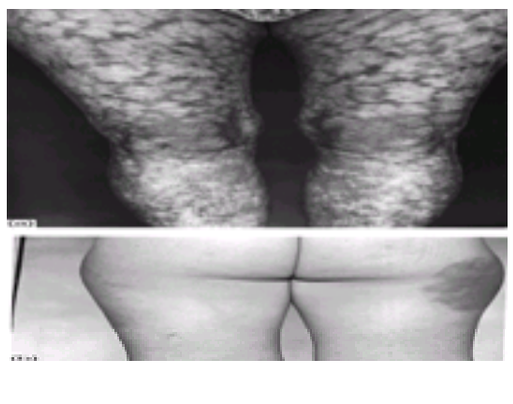

Fig. 22 (a) Chronic vascular disease, especially if inflammatory, summates with the physical effects of cooling to produce livedo reticularis. A non-inflammatory variety associated with cerebrovascular disease is known as Sneddon's syndrome. (b) An equestrian chilblain is due to the combination of the insulating effect of fat and pressure from tight jeans in a young girl riding on a damp and frosty morning.

Ultraviolet radiation and the sun

The sun emits electromagnetic rays comprising a continuous spectrum of short to long waves. Only a narrow range of wavelengths between 400 nm and 770 nm react with photocells in the retina and observed as the various colours of the rainbow. Beyond red (770 nm) is infrared. Heat is due to infrared radiation, which can be felt. Below violet (400 nm) are the ultraviolet (200–400 nm) and X-rays. Most short wavelengths, that can neither be seen nor felt, are filtered out by the Earth's thick atmosphere which includes ozone and water vapour. Therefore, as there is less atmosphere above mountain tops, the danger of radiation exposure is greater. The content of water vapour in the atmosphere varies, which accounts for protection from sunburn in winter, cloudy days, the early morning, or late evening sun. The thick atmosphere of the low-lying Dead Sea in Israel and Jordan is also protective against UVB radiation. Glass filters out wavelengths below 320 nm, so that the closed windows of a car will protect even in a tropical desert unless one is sensitive to the longer wavelengths of ultraviolet radiation. Porphyria is, for example, a disease triggered by UVA radiation and is thus difficult to protect against by shade, cloud, or glass.

Ultraviolet radiation (**UVR**) is arbitrarily divided into UVC (200–280 nm), UVB (280–320), and UVA (320–400 nm). The principal effects of ultraviolet light on the skin are elastic fibre damage and cutaneous ageing, apoptosis, immunosuppression, suntan, sunburn, and carcinogenesis. While each subclass of UVR can produce all of these effects, UVC is relatively more likely to cause cancer, UVB relatively more likely to produce burning, and UVA relatively more likely to produce ageing. UVA is estimated to be 10 times less effective than UVB at producing a suntan and 100 times less effective than UVB at producing non-melanoma skin cancer. This is the basis of so-called 'safe tanning' using UVA light in solariums. However, high-dose UVA does cause non-melanoma skin cancer. In addition, the spectrum responsible for producing melanoma has not been established and may include UVA.

UVR immunosuppression may lead to the reactivation of herpes simplex. It may also suppress tumour surveillance and thereby enhance carcinogenesis. The immunosuppressive, and hence anti-inflammatory, action spectrum has been defined for psoriasis at 312 nm. UVA is only effective for the treatment of psoriasis when the phototoxicity is augmented by a photosensitizer such as psoralen. The action spectrum for the suppression of atopic dermatitis has not been defined, but the skin of patients may improve when exposed to both UVA and UVB. However, a number of related factors influence the response of atopic dermatitis to sunlight and artificial UVR, including ambient humidity (which tends to moisten skin) and heat (which lowers the threshold to itch and scratch).

The diagnosis of UVR damage is determined by recognizing the distribution of the rash as being typical of exposure. Thus the head, nose, and cheeks are principally affected, but there is often sparing below the eyebrows, under a forelock, beneath and behind the ears, and below the chin (see Fig. 14(b)). The sides and back of the neck are picked out, but there is a sharp border to the sun damage where the collar shields the skin from sunlight. Much, of course, depends on the style of clothing as well as on the direction of irradiation. The backs of the hands and dorsum of the feet are often caught by the sun; however, there may be some tolerance of such skin previously exposed and tanned so that skin not so tolerant is clearly more prone to burning. Mediation of sunburn erythema is partly due to the generation of prostaglandins. Plant dermatitis often produces a rash having the distribution of sun exposure. Phytophotodermatitis is a rash in the distribution of actual contact with plant juices on which the sun then acts and produces a burn. The pattern of such casual contact is often streaky and bizarre. Some perfumes containing berloque or musk ambrette are also responsible (see Fig. 53).

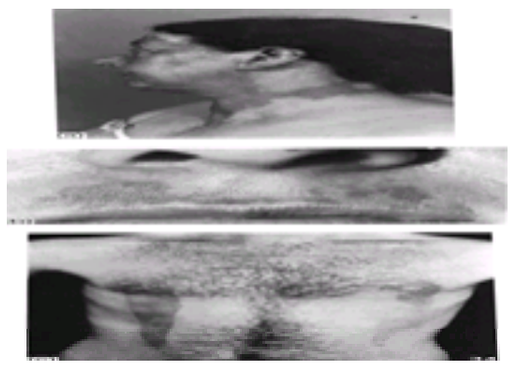

Fig. 53 Pigmentation due to cosmetic agents. Often initially a dermatitis, it is especially induced by exposure to ultraviolet rays. It tends to have the streaky distribution of application. (a) Neck pigmentation from eau de cologne. (b) Lip pigmentation from lanolin. (c) From eau de cologne and sunbathing, the bizarre pattern is characteristic of an exogenous cause.

White skin and the sun

Over the past 50 years, hats, parasols, long skirts, and shawls as well as shady verandas have been replaced by bikinis, solariums, and reckless sun-worshipping. Even redheads and blondes attempt to get a suntan. Only recently through public education has this trend been slowed.

Exposure to sunlight is a major cause of skin ageing and of the epidermal and dermal degenerative diseases that accompany ageing (Fig. 23). In Australia, South Africa, and the south-western United States solar keratosis, basal-cell carcinoma (**BCC**), chronic solar cheilitis, and squamous-cell carcinoma (**SCC**) (Fig. 24) are the commonest reasons for referral to a dermatologist (Table 8). Some two out of every three Australians will develop one or more non-melanoma skin cancers during their lifetime; in Queensland, the lifetime risk of developing a melanoma is 10 per cent. Even children are not completely immune, and persons who burn easily and still persist in exposing themselves regularly to the sun will inevitably suffer gross skin changes, even at an early age. Fortunately, the most common skin cancers have a low potential for metastases.

Fig. 23 Prominent sebaceous glands and comedo formation in solar elastosis.

Fig. 24 Squamous epithelioma of the lower lip as a consequence of sun exposure.

Disorders

Solar keratosis

Solar keratoses are erythematous scaling lesions between 2 and 10 mm in diameter, seen on areas of maximal sun exposure such as the face, dorsum of the hands, forearms, and lower legs. Histologically, these precancerous lesions show dysplasia of the basal keratinocytes. The estimated risk of transformation into either SCC or BCC is very low—1 per cent per annum. Many small lesions resolve spontaneously, particularly with photoprotection. Larger lesions respond to cryosurgery, while patients with multiple lesions may require treatment with topical fluorouracil cream.

Basal-cell carcinoma

Basal-cell carcinoma is a slow-growing invasive neoplasm arising from the basal cells of the epidermis or outer root sheath of hair follicles. Some 50 per cent occur on the head and neck, 30 per cent on the upper trunk, and the remainder elsewhere. BCC can be subdivided into nodular, ulcerated (rodent ulcer), morphoeic, pigmented, and superficial forms. They have a tendency to local invasion, but metastasis rarely, if ever, occurs. Local recurrence is most common with morphoeic BCC. Treatments are influenced both by tumour factors (for example, the size, site, margin definition, and subtype of the BCC) and host factors (for example, general infirmity, coexisting illnesses (such as a bleeding diathesis or susceptibility to bacterial endocarditis), access to local facilities, and patient preference). Options include surgical excision, curettage and electrocautery, radiotherapy, cryosurgery, and injection of interferon-a. Topical immunomodulatory creams, such as imiquimod, are also effective in selected cases.

A number of genetic syndromes of increased susceptibility to BCC have been described. These include the Gorlin syndrome, Bazex syndrome, and Rombo syndrome. Exposure to arsenic also increases BCC susceptibility.

Squamous-cell carcinoma

Squamous-cell carcinomas are faster growing areas of ulceration or tender nodules that occur on areas of maximal sun exposure, such as the dorsum of the hands, balding scalp, face and neck, upper trunk, and lower legs. While UVR is the principal aetiological factor, other predisposing factors include exposure to radiotherapy, chronic leg ulcers, burn scars or sinuses from osteomyelitis, erythema ab igne, and the porokeratosis of Mibelli. Systemic immunosuppression, as used to prevent organ transplant rejection, increases the susceptibility to SCC by 100-fold. In addition, people with xeroderma pigmentosa, dystrophic epidermolysis bullosa,

Rothmund Thomson syndrome, and arsenic exposure have a greater susceptibility to developing SCC. Up to 1 per cent of SCCs will metastasize, most commonly to regional lymph nodes; however, 20 per cent of any metastasis is bloodborne to the liver, lungs, brain, and bone. Survival rates following metastatic SCC are poor, with less than 30 per cent responding to current therapies. High-risk SCCs include rapidly growing tumours, large lesions (>2 cm in size), deeply invading tumours (>4 mm), poorly differentiated tumours or tumours with perineural invasion, recurrent tumours, tumours of the lips or ears or arising within scars, and tumours occurring in immunosuppressed patients.

Following surgery, 75 per cent of local recurrences occur within 2 years, and 95 per cent within 5 years. In addition, patients are at risk of developing a second primary skin cancer. Within 5 years, 12 per cent of patients will have developed a new SCC, 43 per cent a new BCC, and 2 per cent a melanoma.

Melanocytic naevi (moles)

Melanocytic naevi are benign neoplasms of melanocytes. Junctional naevi are localized collections of naeval melanocytes found in the epidermis and superficial dermis, they are flat and pigmented. Compound naevi are localized collections in the epidermis and superficial and deep dermis, these naevi are pigmented and raised. Intradermal naevi are localized collections in the deep dermis with no involvement of the epidermis or junctional dermis, such naevi are flat and flesh-coloured. Most naevi are absent at birth. They tend to first appear and increase in number during childhood, reaching their maximum in early adulthood. Naevi counts have been shown to increase at an earlier age in Caucasians who live closer to the equator, where sun exposure and intensity is greatest.

Large numbers of naevi, both common acquired naevi and atypical naevi, are markers of individuals who are susceptible to developing melanoma. Atypical naevi are large (>6 mm) compound naevi often with a surrounding macular erythematous component, irregular in colour or shape, but nevertheless benign. These lesions may or may not show histological evidence of dysplasia, therefore the old term 'dysplastic naevi', which leads to much confusion, is not recommended. Although the clinical differentiation of atypical naevi from melanoma is difficult, it may be facilitated by regular surveillance and clinical photography. The presence of more than five atypical moles on a person has been shown to be a powerful and independent marker of melanoma susceptibility.

'Congenital naevi' is a term applied both to compound naevi present at birth and to acquired naevi that are clinically and histologically similar to moles present at birth. One child in 100 is born with a congenital pigmented naevus, while 6 to 12 per cent of children and adults have a 'congenital type naevus'. The risk of evolution of a small congenital naevus (<2 cm) into invasive melanoma is unknown, but is thought to be much less than 1 per cent. Prophylactic excision is not universally advocated, but rather considered on a case-by-case basis. Large congenital naevi (20 cm) occur in 1 in 500 000 births and probably carry a 4 to 6 per cent risk of progression to melanoma over a lifetime. The melanoma usually develops after puberty and does not always occur in the naevus itself. Unfortunately, removal of large lesions is rarely easy and may involve numerous surgical procedures.

There may be a natural evolution of acquired naevi from junctional to compound to intradermal naevi. The malignant potential of individual lesions is low, therefore prophylactic excision of acquired melanocytic naevi to prevent transformation into melanoma is not advised. Most melanomas arise *de novo* in the absence of any clinical or histological evidence of a pre-existing naevus. Excision of naevi is advocated where it is not possible to clinically exclude a diagnosis of melanoma. The lesion should be totally excised and submitted for histological assessment. Destructive therapy of melanocytic naevi without histological assessment is a recipe for disaster, for clinical diagnosis is not completely accurate and an opportunity to re-excise an incompletely or inadequately removed tumour may be missed. Patients who subsequently present with metastatic melanoma without an identified primary will rightly or wrongly point the finger at the physician who removed the 'naevus', and claim compensation. Therefore, complete removal and histological examination is advocated even when moles are removed for purely cosmetic reasons.

Melanoma

Melanoma is a malignant neoplasm of melanocytes, which, although initially confined to the epidermis, later invades deeper layers of the skin. Melanoma is one of the most common cancers and cause of cancer-related deaths among adult Caucasians. It is caused by childhood exposure (in particular intermittent exposure) to sunlight resulting in sunburn. The incidence of melanoma increases with age, the lifetime risk for people in the United Kingdom being around 1 to 2 per cent. The greater exposure to sunlight experienced by Australians has increased their lifetime risk to between 3 and 10 per cent, depending on their proximity to the equator. Melanoma is rare during childhood and in Blacks and Asians.

Hutchinson's melanotic freckle occurs on the head and neck of older people who have been heavily exposed to sunlight. These are slow-growing *in situ* melanomas with the potential to progress to invasive melanoma and metastasize. Acral lentiginous melanoma, including subungual melanoma, is equally common in all races and therefore does not appear to be caused by exposure to the sun. A tendency to delayed diagnosis gives this variant a generally poor prognosis.

Superficial spreading and nodular melanomas predominately occur on the trunk in men and on the limbs in women, not necessarily at sites heavily exposed to sunlight. The clinical features of a superficial spreading melanoma include a history of a new mole or a change in the size, shape, or colour of an existing mole. Itching and bleeding are late signs. On examination, melanomas are asymmetrical, irregular in outline and colour, and often stand out as different from other moles on the patient's skin. They are frequently over 6 mm in diameter at the time of diagnosis; however, this feature is now less common than in the past due to improved awareness of the early warning signs by at-risk populations. These signs are illustrated in Plate 1, Plate 2, Plate 3, Plate 4, Plate 5, Plate 6 and Plate 7.

The risk of metastasis, and hence the 5-year survival rate, is related to the depth of invasion of the melanoma measured in millimetres from the granular layer of the epidermis, known as the Breslow thickness. Patients with lesions confined to the epidermis (melanoma *in situ*) have a 5-year survival rate of 100 per cent; those less than 0.76 mm, 98 per cent; between 0.76 and 1.5 mm, 95 per cent; between 1.5 and 3 mm, 80 per cent; and lesions of more than 3 mm in depth, 60 per cent.

At presentation, 10 per cent of cutaneous melanomas will have metastasized. In order to identify the primary lesion it is important to look at the entire cutaneous surface—in the eyes, inside the mouth, and the vulva, etc.—and to examine the skin with a Wood's light (see above), for the primary melanoma may have undergone spontaneous regression after giving rise to metastases and only be identified as a hypopigmented patch under such light.

Metastasis is usually to the regional lymph node basin. If a single lymph node is involved the 5-year survival is 45 per cent, if two lymph nodes are involved the survival rate is 28 per cent, but this rate drops to 9 per cent if more than four lymph nodes are involved. The median survival time following metastasis is between 5 and 16 months. Although systemic metastasis is predominately to the lung, liver, brain, and bone, lesions can arise anywhere including bowel, kidney, and muscle. Localized skin metastasis is also common.

In general, there is a low yield from routine investigations for melanoma; moreover, chest radiographs and CT and/or MRI scans are not routinely ordered in asymptomatic patients. However, the investigation of new symptoms results in a high percentage of positive findings, and this policy is the preferred approach. When metastasis does occur, 50 per cent will be within 1 year, 85 per cent within 2 years, and 95 per cent will be within 5 years. The survival rate of such patients increases with disease-free time following removal of a melanoma; the only caveat is that metastasis following thin lesions, albeit rare, may be delayed.

There is no adjuvant therapy that improves survival, and chemotherapy and radiotherapy are only used for palliation. Surgery, when possible, remains the treatment of choice for metastatic melanoma. Initial optimism regarding the use of adjuvant interferon or elective lymph node dissection has now waned. However, the use of lymphoscintography and sentinel-node biopsy to identify high-risk patients for such therapy now seems helpful in finding lymph node metastases, but its use as a tool to improve long-term survival is still uncertain. The only intervention that improves survival is early diagnosis and complete removal of the primary melanoma. Even re-excision of the scar does not improve survival; but it is still recommended with ever-decreasing margins with the sole intention of reducing the incidence of local recurrence/metastasis. Currently, the margins recommended are 0.5 cm for *in situ* melanoma and 1 cm for invasive melanomas with a Breslow thickness of less than 1.5 mm. A minimum margin of 1 cm and a maximum margin of 2 cm of normal skin is recommended for lesions between 1.5 mm and 4 mm deep. Although re-excision of the scar for lesions greater than 4 mm is unlikely to be worthwhile, a minimum margin of 2 cm is recommended.

Rashes due to sun or artificial light and associated ultraviolet rays

Sunburn

Initially this is an erythema occurring about 6 to 8 h after exposure, which may progress to blistering and later to skin-peeling. However, redness may begin as early as 2 h if the exposure is excessive. Sunburn tends to resolve, often with peeling, after 24 to 72 h, depending on its severity. Topical corticosteroids and oral non-steroidal anti-inflammatory agents may provide partial relief.

Solar urticaria

Here, erythema and wealing occur immediately on exposure to sun, often of sites not habitually exposed to the sun. It is a rare but very disabling condition due to a broad spectrum of wavelengths that rarely responds to sunscreens, although it often does well with plasmapheresis or, paradoxically, phototherapy.

Polymorphic light eruption

Polymorphic light eruption is an altered quality of sunburn. Thus instead of erythema there is an itchy papular or eczematous response about 6 to 8 h after exposure, which may persist for several days (Fig. 25). There are several variants including a lymphoma-like pattern with heavy lymphocytic infiltrates.

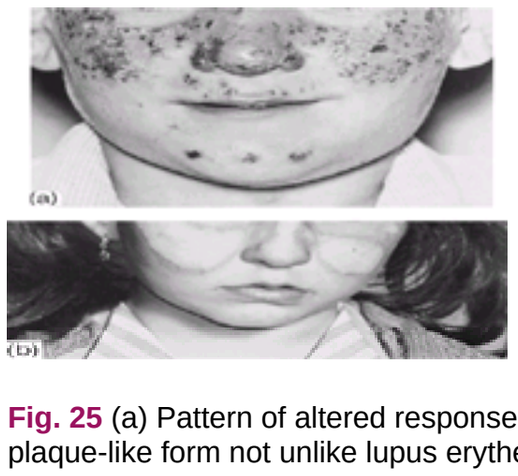

Fig. 25 (a) Pattern of altered response to sunburn wavelengths: an eczematous prurigo with excoriations. (b) Pattern of altered response to sunburn wavelengths: a plaque-like form not unlike lupus erythematosus.

Actinic prurigo

Sun-induced prurigo shares features with atopic eczema, polymorphic light eruption, and persistent light eruption due to photosensitivity agents in the environment. It is uncommon except in genetically susceptible individuals with HLA-A24 and CW4 and the rare subtype DR4 DRB1*0407, which is particularly common among some native Americans. It has a chronic course that is only initially seasonal. Urticarial plaques develop a few hours after ultraviolet exposure and are followed by a persistent eczematous rash, which is not always confined to exposed skin.

Exacerbation or localization of other dermatoses in sun-exposed sites

This is characteristic of pellagra, Hartnup's disease, lupus erythematosus, dermatomyositis, pemphigus erythematosus/foliaceous, Darier's disease, herpes simplex, rosacea, scleroderma, erythema multiforme, actinic lichen planus, and lymphocytoma. It sometimes occurs in psoriasis, seborrhoeic dermatitis, atopic eczema, acne, and bullous pemphigoid.

Ultraviolet rays may diminish antigen surveillance by reducing the population of Langerhans cells.

Porphyrias

See Chapter 11.5 for further discussion.

Drug eruptions and photosensitivity

Various drugs can cause acute eruptions of erythema that swell or blister like severe sunburn, which, since UVA is often responsible, are independent of bright sun. The reactions are often dose-dependent, as with psoralens in PUVA therapy. Ingestion of a spinach (Atriplex) also causes photosensitivity (see Table 9). Some eruptions present only as deep pigmentation.

Xeroderma pigmentosum

See Chapter 25.2 for further discussion.

Persistent light eruption

'Persistent light eruption' is the term given to a sensitivity to light induced by agents previously applied to the skin, often years before. Drugs eliciting light sensitivity are listed in Table 9.

Investigations

Patients should be asked about their family history, their occupation, drug or food ingestion, and exposure to perfumes, as well as how and when any exposure occurred. It is important to know whether glass is protective. Many patients confuse the effects of heat and sunlight, making it important to clarify whether exposure to an open fire, for example, also exacerbates their condition.

Light-testing has become a useful dermatological tool, with a number of centres having access to a monochromator that specifically evaluates each band of light.

Prophylaxis

Health education in schools should emphasize that burning in the sun is not related to heat or wind, rather it is maximal when the sun is directly overhead. Therefore protection is essential between the hours of 10 am and 2 pm, mainly achieved by covering the head and body and keeping in the shade. Sunscreens are effective only if they are properly used; that is to say, they must be applied evenly and thoroughly and well before exposure, and reapplied at appropriate intervals, but they are no substitute for avoidance. Frequent, uninhibited exposure leads to the accumulation of almost inevitable and irreversible injury, which may only become apparent some 20 years later.

Management

The ill-effects of a cool, sunny, and windy noon must be explained, as well as the relative safety of a hot sunny evening. Ensuring protection for those who are sensitive to UVB includes giving advice on the time of day and season likely to be harmful. Most patients can safely take a swim in the early morning or late afternoon. Those with severe sensitivities, such as xeroderma pigmentosum, can be saved from all ill-effects by diligent protection, including clothing. Indeed, clothing and shade are the best protections for children sensitive to the sun, but a wet T-shirt can transmit ultraviolet rays; tightly woven silks, Lycra, and cottons are more effective than loosely woven yarns or wool. Fluorescent lighting emits small amounts of UVR and is safe for all but the most sensitive. Glass windows are protective against UVB and shorter wavelengths, while Perspex and certain plastics also protect against UVA. Natural pigment and thickening of the epidermis accounts for normal tolerance.

Sunscreens are rated, according to UVB protection, by a solar protection factor (SPF). There is an inverse relationship between SPF and sun protection, such that

there is only a minimal difference between SPF 8 and SPF 16 and probably no real difference between SPF 16 and SPF 50. The limiting factor of any sunscreen is correct application and reapplication.

Sun-screening agents include thick reflective pastes or creams such as zinc oxide or titanium dioxide; these filter out both UVA and UVB and also prevent tanning. Sun-screening agents may absorb light, with invisible lotions or creams containing *p*-aminobenzoic acid in 70 per cent alcohol mainly filtering UVB. This allows some tanning due to the effects of UVA.

Is it what I have eaten? Food and drug eruptions

It is generally accepted that much of what we eat is antigenic and that absorption does occur. Usually allergens are complexed with antibodies in the gut wall and tolerance occurs. It is easy to demonstrate that antibodies are made to counteract food allergens and that complexes circulate in the blood as a result of eating. This is not necessarily allergy because resulting inflammation is rare.

Infrequent ingestion is more likely to cause an allergy than tolerance.

Both erythema and acute urticaria can be caused by food, but chronic urticaria is rarely so. In the atopic person, IgE-mediated food allergy is well recognized. Contact urticaria occurs when eggs or milk touch the lips and cause immediate swelling, whereas generalized urticaria and bowel upset result when agents such as fish, nuts, or strawberries are eaten. However, in many patients such eruptions are examples of non-allergic intolerance.

Anaphylactoid reactions are either idiosyncrasies, whereby an individual reacts abnormally to a substance tolerated by most of the population due to some defect in their physiology, or they result from a direct effect or action of a drug on a mast cell, or other cell, often on first exposure to the eliciting substance. Examples of non-allergic responses include C_1-esterase deficiency and angio-oedema, or lactose intolerance and lactase deficiency, causing diarrhoea.

A high iodine level in a diet induces blistering in dermatitis herpetiformis and exacerbates erythema nodosum leprosum. Sources include iodophors used in dairy cleansing, iodine-containing food supplements, and dough improvers in bread, as well as in cough mixtures.

Following the development of toxic erythema and purpura in an extensive epidemic amongst eaters of margarine in The Netherlands and Germany, the possibility that food additives could cause rashes is now well recognized. Urticaria, asthma, and migraine have also been studied in this respect. The total number of food additives exceeds 20 000 and an average person eats about 1.5 kg every year. Salicylates and benzoates as well as many colouring agents present in more than 1000 drugs marketed in the United States partly act through the control of prostaglandin metabolism. Some 10 per cent of people sensitive to aspirin are also sensitive to the colouring agent tartrazine. The mechanism, though suspected to be related to prostaglandin metabolism, has yet to be clarified. It is confusing that some other types of food allergy, perhaps less dependent on IgE and on the release of histamine from mast cells, are prevented by prostaglandin inhibitors: 'If one takes indometacin, one can eat anything'!

Careful studies of food additives have led the authors to the conclusion that food containing additives is very rarely a danger, but food without additives is commonly so. So far as intolerance is concerned, the potentiating effect of psychological tension and unaccustomed or overindulgent eating habits is a probable explanation.

Shellfish and strawberries are well known for not only releasing histamine from mast cells but for causing thrombocytopenic purpura. Usually such agents cause urticaria within hours of ingestion. However, the response is inconsistent, since there may be times or forms of presentation of the same food that avoid this effect. Eggs, nuts, chocolate, fish, shellfish, tomatoes, pork, strawberries, milk, cheese, and yeast are common causes of a sudden transient thrombocytopenia, and sensitivity to food in this way is the basis of the 'thrombo' test. A 20 per cent fall in the platelet count 1 h after ingestion occurs in 70 per cent of persons showing allergy to aspirin, barbiturates, and penicillin. In one series, 203 out of 215 patients with urticaria showed a prolonged bleeding time from the ear lobe 2 h after challenge with a drug, chemical, or food. Bitter lemon or tonic water containing quinine is especially well documented.

In patients with atopic eczema and asthma the problem of food allergy is more complex. These patients seem to be more susceptible to histamine release even from non-allergic sources.

The gut of the newborn with atopy is said to be more immature in its handling of foreign protein; or it may be that the complexing of IgA is less effective so that IgG is brought into play in the immature gut in a manner not usual for the mature immunological system of the adult gastrointestinal system. This is the basis for advocating breast feeding without cows' milk substitution for all babies of atopic parents. Hypoallergic foods are marketed which contain 'predigested' casein, and this is an industry with a strong following. However, minute amounts of antigens from the maternal diet are found in breast milk, so it cannot be assumed that the first experience of a dietary antigen occurs at the time of weaning. Even the fetus is normally supplied by minute samples of the mother's diet. What matters is how much and when.

Another form of food sensitivity is that occurring in nickel-sensitive subjects. Nickel sensitivity is one of the commonest aggravating factors of hand dermatitis, and there seems little doubt that contamination from metal pots and from some green vegetables can contribute to the contact allergic dermatitis in sensitive individuals.

Drug eruptions

The 1975 Boston Collaborative Program of Drug Surveillance observed that adverse reactions accounted for 3 per cent of hospital admissions and 14 per cent of medical resources, and that 30 per cent of hospital patients developed adverse reactions.

It is wise to assume that any drug can cause any rash, but until there are simple, reliable, and specific *in vitro* tests for testing human tissue for hypersensitivity, the diagnosis of drug eruptions will depend entirely on clinical judgement. The physician has to decide whether the rash has some other cause by recognizing certain physical signs, ranging from the mite burrows in scabies to the herald patch of pityriasis rosea. Then if a drug seems a likely cause, there must be an attempt to decide which of the medications currently prescribed, or taken secretly, may be responsible.

Drug rashes are essentially blood borne and therefore often have a symmetrical urticarial, erythematous, or purpuric and ischaemic pattern determined by the vascular anatomy. Less likely is a 'primary epithelial' reaction in the initial stages, and so scaling or even the vesiculation of eczema as a first manifestation of a generalized drug rash would be unusual. The exceptions are well known and include the intraepidermal immunologically induced 'pemphigus' rash of penicillamine, rifampicin, and captopril, especially when the first of these is used to treat rheumatoid arthritis; another is when cell-mediated hypersensitivity to epidermal protein and the drug occurs in a person previously having a contact dermatitis to a local antihistamine or sulphonamide. The psoriasis-like rash of practolol and various other b-blockers as well as the scaly eruption (particularly of the scalp) from methyldopa, are exceptions.

Nevertheless, if a rash looks like eczema, it is probably not caused by a drug. If it is an erythema and urticaria, it may well be. As later stages of the rash are frequently complicated by secondary desquamation and peeling, the diagnosis should be made on the initial manifestation.

Unlikely or likely drugs

It may be helpful to rule out unlikely offenders such as digoxin, paracetamol, steroids, other hormones, and vitamin and electrolyte supplements. In any drug group there are likely and less likely offenders. Thus of the antibiotics, oxytetracycline, nystatin, and erythromycin are not under suspicion, but dichlortetracycline is a common cause of a photosensitivity rash. Moreover, ampicillin is almost invariably responsible for a characteristic bright-pink maculopapular rash in patients with infectious mononucleosis

Sulpha-containing medications such as sulphonamides, thiazide diuretics, oral hypoglycaemic agents, dapsone, and even captopril are capable of causing most rashes. Non-steroidal anti-inflammatory drugs (**NSAIDs**), antibiotics, antiepileptics, gold, penicillamine, allopurinol, halides, and other antihypertensives are all possible candidates worthy of consideration.

Timing

While drug rashes usually relate to newly started medications, this is not always the case. For example, a person may have tolerated multiple courses of a particular antibiotic prior to becoming allergic to it. Maculopapular drug exanthems most commonly begin 7 to 10 days after the drug is commenced, by which time many antibiotic courses will have been completed. The rash often increases in severity with time until the patient stops taking the drug. Once the drug is stopped, the rash may last a further 7 to 14 days and often evolves from a typical maculopapular exanthem to a scaly rash that may resemble the peeling that follows sunburn or even be slightly eczematous. Repeat exposure to that drug, even years later, will usually lead to recurrence of the exanthem within hours to days. Interestingly, some patients do not react on re-exposure, suggesting a toxic rather than an allergic mechanism. This is particularly common in people who react to amoxicillin in the context of infectious mononucleosis.

Other exanthems frequently show a different time course. Anaphylaxis often begins within minutes, while fixed drug eruptions usually begin 30 min to 8 h after rechallenge. Lichenoid drug eruptions often do not begin for many months and sometimes even years. Urticarial reactions are variable: they may start within a few hours with an antibiotic allergy, but may take a week or two to develop following treatment with NSAIDs or angiotensin-converting enzyme inhibitors.

Erythema multiforme, or Stevens–Johnson syndrome, occasionally occurs surprisingly early after a drug's administration, leading to the suspicion that the disease for which the drug was given may have prepared the host in some way.

Many patients do not admit to taking a drug, perhaps because it was never prescribed but borrowed from a family member or neighbour or bought over the counter and therefore considered to be harmless. In general, drug rashes do not persist after withdrawal of the drug—exceptions include pemphigus from penicillamine.

Transient susceptibility

The best example of a susceptibility reaction is urticaria. Many people with chronic urticaria are susceptible to it for a period of many months, during which time the rash may be triggered by prostaglandin synthetase inhibitors, such as aspirin or indometacin. It is possible that certain drug exanthems require both a drug and an infection to provoke the rash, so the underlying disease of the patient should always be taken into account. Patients with infectious mononucleosis or chronic lymphocytic leukaemia are prone to toxic erythema with ampicillin, while people with AIDS are 1000 times more susceptible to developing the Stevens–Johnson syndrome following treatment with sulphonamides. Sometimes immune-complex diseases, such as cutaneous vasculitis from infective organisms, are provoked by interference with immunological mechanisms by certain drugs. The particular set of circumstances, which may not recur, would depend on the formation of antibodies, the nature of the infectious organism, and the taking of the drug at that time. Diseases such as psoriasis or dermatitis herpetiformis may go into spontaneous remission, during which time they are less likely to be provoked by drugs. Dermatitis herpetiformis is provoked by iodine so readily that it should be avoided if possible.

Drug allergy is rarely proven, but overdosage is a frequent well-established fact due to faulty prescribing, attempted suicide, or altered metabolism as in renal or hepatic failure. Interethnic differences in drug metabolism are due to human gene polymorphisms. Drugs may interfere with metabolism, with hormones, they may be deposited, they may react with sunlight, they may modify the ecology of the skin in respect to infective organisms; they may cause reactivity of certain cells such as the mast cell; they may be cytotoxic; and they can act as allergens in the formation of immune complexes, hapten–protein complexes, delayed cellular immunity, and a variety of other mechanisms. Some are not understood, such as the effect of halogens on the formation of granulomas (Fig. 26(a)) and in the causation of an acneiform eruption (Fig. 26(b)). This includes the use of fluoride gel preparations applied to the teeth to prevent dental caries.

Fig. 26 (a) Iodides and bromides are occasionally responsible for a granulomatous eruption with pseudoepitheliomatous hypertrophy. Potassium iodide in a cough mixture was responsible for the eruption in this patient. (b) Prolonged administration of iodides or bromides causes a particularly inflammatory form of acne which, although commonly in the distribution of acne vulgaris, may be more widespread. This eruption was due to an iodide-containing 'tonic' for the blood.

Specific drug eruptions

The most common diagnostic problem is a toxic urticated erythema. It begins like measles without the upper respiratory and conjunctival prodromal signs. It usually develops, over a number of hours, as a bright pink indurated papular eruption (Fig. 27) and, unlike urticaria, persists for days, ultimately involving the epidermis and producing scales (Fig. 28). After the first 2 to 3 days the rash tends to be fixed, with the principal changes due to mild bleeding of the skin with overlying slight peeling. Fever and arthropathy may be associated. Common causes are ampicillin, phenylbutazone, phenothiazine, co-trimoxazole, diazides, and sulphonylureas. Exfoliative dermatitis is the end result of this type of reaction (Fig. 29). Gold, phenylbutazone, indometacin, allopurinol, hydantoins, sulphonylureas, ampicillin and amoxicillin, co-trimoxazole, carbamazepine, phenytoin, cefaclor, gentamicin, and *p*-aminosalicylic acid are causative drugs.

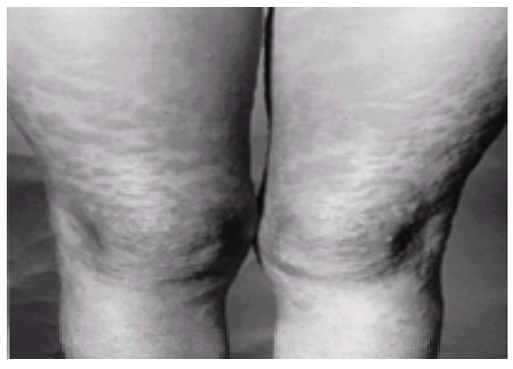

Fig. 27 One of the most common drug eruptions, initially a bright-pink papular eruption, symmetrical, and becoming confluent. This case is due to ampicillin.

Fig. 28 A later stage of acute drug eruption, in this case due to Myocrisin®. The epidermis is reacting to the dermal inflammation by hyperplasia spreading centrifugally to produce an annular scaly lesion, with the scale exfoliating in the centre of the lesion and attached to the spreading margin.

Fig. 29 Severe oedema, crusting, and exfoliation due to dermatitis from arsenicals.

Psoriasis

b-Blockers can cause a psoriaform scaly eruption that may be modified by basal-cell necrosis. The high turnover as in psoriasis, with the slowing down that results from such necrosis, gives rise to a hyperkeratotic scale that is more adherent than in psoriasis and often slightly yellowish. The palms and soles, elbows, and knees are particularly favoured (Fig. 30).

Fig. 30 Hyperkeratosis and slight scaling is a feature of the psoriasiform eruption caused by b-blockers.

Labetalol, propranolol, and oxprenolol cause a partly psoriaform and partly lichenoid rash, which is most marked over bony prominences. There is an itchy hyperkeratosis of the palms and soles. Most b-blockers merely exacerbate ordinary psoriasis. Other drugs capable of this include lithium, antimalarials, non-steroidal anti-inflammatories, potassium iodide, amiodarone, and calcitriol, it also occurs on the withdrawal of oral steroids.

Lupus erythematosus

Lupus erythematosus, like erythemas or necrotizing vasculitis, is most commonly caused by hydralazine, phenytoin, practolol, penicillamine, and isoniazid. Many other drugs have been incriminated. The drug-induced lupus erythematosus is reversed by withdrawal of the drug, but it recurs when it is readministered. The disease is characterized by antinuclear antibody in high titre with normal DNA binding. Inhibition of C4 underlies the immunological disease induced by hydralazine.

Scleroderma

An epidemic originating from denatured rape-seed oil in Spain caused facial oedema, exanthems, and ultimately a scleroderma-like syndrome.

Fixed drug eruption

Although the mechanism is unexplained, the eruption is easy to recognize as it is usually circular and erythematous (Fig. 31), and it frequently blisters. Postinflammatory hyperpigmentation due to pigment incontinence in the dermis is characteristic. It is fixed in site, and whenever the subject takes the causative drug the eruption begins within a few hours in exactly the same site. The tongue and the glans penis are common sites. The affected area can be transplanted without loss of responsiveness in some cases. In pigmented races, very dark pigmentation remains between attacks. In addition to the drugs listed in Table 10, purgatives, blood cleansers and tonics, and many other home remedies containing phenolphthalein, are common causes. Continued ingestion leads to the development of multiple new spots. Skin biopsy is diagnostic in the early stages, and it is safe to rechallenge patients in order to determine the causative agent.

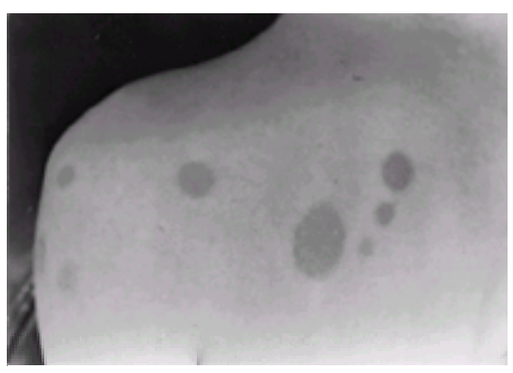

Fig. 31 Fixed drug eruption due to phenolphthalein present in a laxative. Such an eruption characteristically appears within half a day of taking the causative drug and the site affected is the same on every occasion. Violaceous annular lesions are common and may persist for several weeks.

Anticonvulsant hypersensitivity syndrome

This severe reaction is characterized by an extensive morbilliform skin rash that often evolves into an exfoliative erythroderma or toxic epidermal necrolysis (TEN), accompanied by fever, lymphadenopathy, hepatitis, nephritis, and leucocytosis with atypical lymphocytes on the blood film. There is a significant mortality and intensive nursing care is required. Other drugs such s allopurinol can produce a similar syndrome.

Management of drug eruptions

The best way is to stop the use of all drugs likely to cause the eruptions. Readministration of the drug is possible for most drug eruptions other than those that cause anaphylactic shock, but it is usually at a risk of considerable morbidity and therefore should be considered only if essential to the patient. Skin tests are unhelpful, as a risk of dangerous anaphylaxis, false-negatives, and lack of knowledge of the antigen, makes skin testing useless.

Blood tests are of no help in trying to find which drug is causing the problem. Various tests, such as the reaction of basophil cells or the release of lymphokines from lymphocytes, have not proved of routine value. Eosinophils may suggest that an eruption is due to a drug, and, as mentioned above, a fall in the platelet level within 1 h or a prolongation of bleeding time 2 h after injection is helpful for some urticarial rashes.

Dermatitis

Definition

Dermatitis is a non-specific inflammatory response of the skin to a combination of exogenous and endogenous factors. There is no clear distinction between dermatitis and eczema, and the two terms are interchangeable. Endogenous dermatitis comprises discoid, asteatotic, varicose (or stasis or gravitational), vesicular, hand/foot (pompholyx), atopic, and seborrhoeic dermatitis. Exogenous dermatitis includes irritant contact, allergic contact, phototoxic, and photoallergic dermatitis. However, sometimes there is no clear distinction between these two types.

Clinical features

Dermatitis has both dermal and epidermal components. Some signs are confined to the dermis such as swelling, heat, itchiness, tenderness, and redness, but at the same time the epidermis proliferates and therefore thickens and produces scale. The oedema in the dermis extends to the epidermis, swells the cells, and separates them to give the histological appearance of a sponge, known as spongiosis, and frequently this results in vesicles, which distinguish dermatitis from other proliferative states of the epidermis such as psoriasis. Acute weeping exudation occurs when the vesicles burst (Fig. 32). Itching is usually severe in dermatitis.

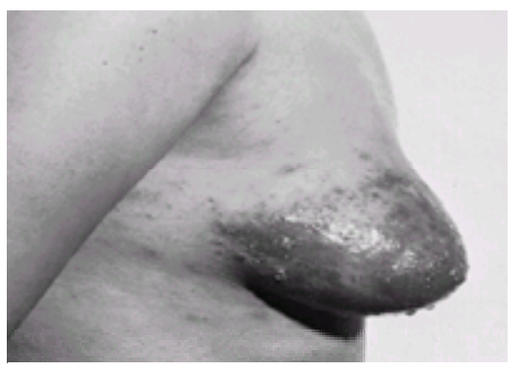

Fig. 32 Acute dermatitis is characterized by an oedematous epidermis in which vesiculation, oozing, and crusting are the principal features. The borders are often ill-defined, while the centre of the lesion is confluent.

The reaction pattern of dermatitis is not homogeneous. It is made up of papular elements of different ages and size, sometimes confluent in the centre (Fig. 33), with widely scattered satellite papules or vesicles. The scales are of varied size and broken by excoriation, exudate, and even pinpoint haemorrhages.

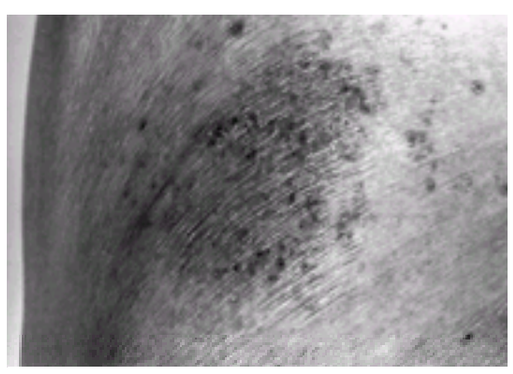

Fig. 33 Dermatitis comprises papules that are confluent in the centre and become vesicular or evidently excoriated. Oedema makes the line markings in the skin more prominent. There are satellite lesions beyond an ill-defined border.

A secondary factor prominent in pigmented skin is a loss of melanin, or at least a failure to retain it, in the acute lesion so that the skin is depigmented. In later or more chronic stages the dermis is darkened by pigment 'incontinence', so that thickened chronic epidermal plaques may contain increased pigment in the underlying dermis. Chronically scratched skin has a brownish violaceous colour due to the combination of pigment, vasodilatation, and epidermal thickening.

For unknown reasons, dermatitis of the foot frequently provokes an autosensitization or 'id' response in the hand. Thus, vesicular eczema of the hands often follows a fungus infection of the feet, varicose eczema of the lower legs often spreads to the forearms and face, and a severe allergic contact dermatitis may generalize to the trunk and limbs.

Contact dermatitis

Primary irritant contact dermatitis

An irritant can be defined as a chemical that in most people is capable of producing cell damage if applied for a sufficient time and in a sufficient concentration. Fibreglass spicules rubbed into the skin are a typical example (Fig. 34). Irritant contact dermatitis is caused by exposure to a single or a few contacts with a highly irritating substance such as a concentrated acid, or by chronic low-grade cumulative exposure to substances that are mildly irritating, such as a very dilute acid. A low-grade, cumulative, irritant contact dermatitis can occur after a few months or even several years, depending on the nature of the irritant and the sensitivity of the skin. This is exemplified by housewives' hand dermatitis, which usually recovers slowly or incompletely because of the inability to fully protect the hands against all irritants (Fig. 35). Many people at home or in industry are in daily contact with various chemicals over long periods. They work in wet or extremely dry conditions with skin cleansers, alkalis, acids, cutting fluids, solvents and oxidants, reducing agents, enzymes, and medicaments. The skin is also worn and irritated by cold and heat, sun, pressure, scratching, or friction of various kinds from tools or clothing. Many variables influence the skin's toughness or vulnerability. It can be immature in the newborn or worn out in the aged. The most important cause of lowered resistance is a constitutional disease, such as the ichthyotic skin of old age (Fig. 36), atopic eczema, or psoriasis.

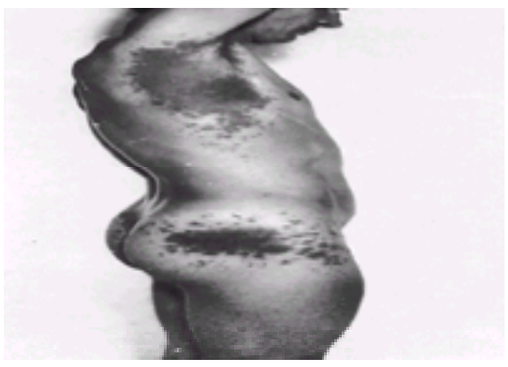

Fig. 34 Primary irritant dermatitis due to small spicules of fibreglass at sites of friction after this patient had insulated a roof.

Fig. 35 Chronic dermatitis causes irregular thickening of an inhomogeneous epidermis. The texture of the stratum corneum varies so that it is firmly attached at some points but exfoliates with small scales at others. Loss of moisture causes decreased suppleness, cracking over joints, and exposure of deeper epidermal cells. This causes irritation of the dermis at the bottom of the deep crevasses.

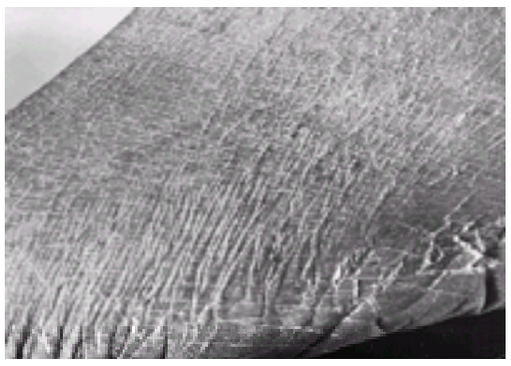

Fig. 36 Chronically thin and slowly turning-over epidermis results in a closely knit stratum corneum which is firmly adherent but cracks excessively. It is characteristic of elderly, malnourished, or ichthyotic skin. Such skin is less resistant to primary irritants.

Contact allergic dermatitis

Allergic contact dermatitis is caused by a type IV delayed hypersensitivity reaction to a chemical in contact with the skin. Initial sensitization can occur 7 to 10 days after the first contact with a potent allergen. However, it is more usually a consequence of many months or years of exposure to small amounts of the allergen. Once sensitized, contact with the allergen can produce dermatitis within 24 to 48 h and all areas of the body are equally susceptible. Sensitivity can vary due to the amount of exposure, the degree of penetration of the skin, and the tolerance of the immune system.

It is believed that certain allergens, such as nickel and chrome, have a greater affinity for the skin than others. This is partly due to the easier recognition and assimilation by the epidermal antigen-presenting cells known as Langerhans cells. The allergen binds to epidermal microsomal protein, or to some cell-surface marker, or to serum proteins that are plentiful in the epidermis. It is a complex of the allergen with such protein that is recognized as foreign. Although the T lymphocyte ultimately recognizes the complex, the macrophage is a necessary intermediary. Suppression of Langerhans cells by ultraviolet rays diminishes cell-mediated immunity. Genetic factors play a part in the recognition process. Once recognized, T-cell proliferation occurs in the paracortical area of the lymph node, and on re-exposure sensitized lymphocytes release lymphokines. The mechanisms of lymphocyte stimulation include some role for suppressor and effector cells. The role of antibodies, some of which are clearly specific for the same antigen, is also unknown. The inflammatory reaction resulting from recognition is variable and dependent on other pharmacological agents, including secretions from the mast cell, and on prostaglandins. Some of the variation in response, such that persons are consequently labelled as more or less allergic, depends on these secondary factors, and these can be modified by various conditional factors, including anxiety and the hormonal status of the monthly menstrual cycle.

Contact dermatitis sensitizers

In the following, we will concentrate on some specific groups of sensitizers and irritants.

Cosmetics

Cosmetics applied to the skin, although more rarely a cause of dermatitis in technically advanced countries where the industry has worked hard to eliminate allergens, are still a source of much disease in developing countries. Perfumes and preparations containing tars, formaldehyde, and Dowicil are increasingly incriminated, and

are as commonly irritant as they are allergic.

Vaseline dermatitis in the Bantu is an example. In technically advanced countries, deodorants are a common cause of dermatitis, while in the hair industry, glyceryl monothioglycollate (acid perms) is the most common allergen. Hair bleaches, such as ammonia persulphate, commonly cause immediate, non-immune wealing. When in doubt, because the constituents are so complex, cosmetics should be tested by direct application to the skin, but this can give rise to false-negative results. Hair dyes are now so common that their relative safety can be expected. However, again in developing countries, the dye paraphenylenediamine may produce an acute dermatitis, often first affecting the eyelids and other aspects of the face before showing much evidence of dermatitis on the scalp.

Clothing and textile dermatitis

On the whole this is rare, but clips containing metal are quite a common cause of dermatitis (Fig. 37); for example, jeans' buttons can cause dermatitis of the skin below the umbilicus. There is also evidence that the rubber in the elastic of many garments is sometimes the cause of dermatitis. Dyes are usually a problem at sites of friction where there is also moisturization by sweat: the majority of which are azodyes or paraphenylenediamine. In the textile industry, chrome and formaldehyde are important agents causing dermatitis.

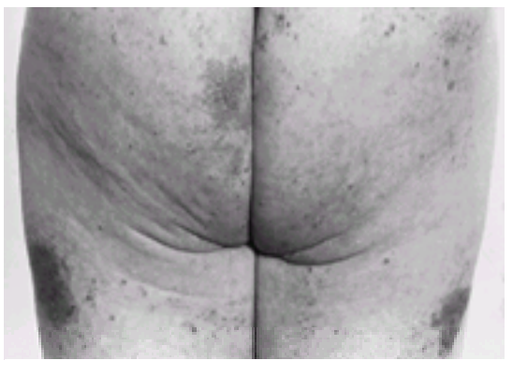

Fig. 37 Contact dermatitis due to garments containing nickel. The diagnosis is made by observing how the distribution of the rash corresponds to the distribution of the contact with the causative agent.

Shoe dermatitis is commonly due to chrome or to rubber additives such as mercaptobenzothiazole or butyl phenol formaldehyde. Adhesives and dyes may also be responsible. This form of dermatitis should be considered in every person with eczema of the feet. It often spares the area between the toes as this is the point where the shoe is not in contact with the skin. Much modern footwear has plasticized toecaps and fails to absorb sweat. Increased sweating encourages a shearing strain on the skin, particularly in the athletic child. Frictional dermatitis of the foot is common in such children and is known as juvenile plantar dermatosis.

Foods

In technically advanced countries, handling animal feeds containing antibiotics gives more trouble than handling food for human consumption. Elsewhere, plants and fruits such as garlic, cinnamon, onions, and lemons and oranges cause much trouble, as do shellfish and various species of fish that are sometimes contaminated by algae. This is an important hazard for fisherman, which, in the United Kingdom, is known as the 'Dogger Bank' itch.

Plastics

An increasingly frequent cause of dermatitis arises from the use of acrylic and epoxy polymers or resins. Acrylics account for dermatitis from adhesive tape, spectacle frames, bonding agents, dentures, hearing aids, bone cement, artificial fingernails, sealants, printing plates, and inks.

Epoxy resins are used as surface coatings for steel pipes and ships, powder paints, electrical insulation adhesives, construction of concrete and steel buildings, and for the surface of roads and bridges. Although they are amongst the most potent sensitizers, they are only active during their initial handling since complete polymerization makes the sensitizing monomer non-available. About 90 per cent of contact dermatitis from epoxy resins is from bisphenol A. Protection in industry depends on common-sense avoidance of handling them and general cleanliness in the workshop, but volatile epoxy resins affecting the face are difficult to avoid.

Rubber

Natural as well as synthetic rubbers require the addition of several agents that are strong sensitizers. They make the rubber more malleable and supple, prevent perishing by oxidization, and some speed up the manufacturing processes.

Accelerators include thiuram, mercaptobenzothiazole, and guanides; antioxidants include the monobenzyl ether of hydroquinone. Most cases of rubber sensitivity are due to clothing such as rubber gloves, or to tyres or rubber linings used in the transport industry. Others include the contraceptive sheath, shoes, fingerstalls, masks (particularly motorbike or scuba-diving masks), elastic bands, bicycle or golf-club handles, and rubber sheets or cushions. Anaphylaxis is not rare from a type I sensitivity to rubber latex surgical gloves.

Colophony

Rosin is made from pine trees and is used worldwide for paper size adhesives, inks, undersea cables, Elastoplast, violin rosin, and cosmetics. Some medicaments like Zam-Buk®, Secaderm® salve, and ilonium also contain colophony, therefore explaining the contact dermatitis arising from the use of these agents. It is responsible for about 3.5 per cent of positive patch tests in the London contact dermatitis clinics.

Plants and wood

Sensitivity to plants and woods accounts for enormous worldwide morbidity and occasional mortality. Some plants release their allergen only when bruised, others when lightly touched, and others by airborne pollen. Some produce a contact non-allergic urticaria (that is, immediate stinging) as with the nettle or cowage (*Mucuna pruritum*); others cause an allergic dermatitis, or even photosensitivity. Many are highly irritant.

In North America the commonest cause is poison ivy, in Europe it is *Primula obconica.* Both produce a severe, streaky blistering eruption from contact allergy mediated by cellular immunity.

Chrysanthemum or ragweed plants, members of the Compositae or daisy family, cause a more diffuse redness and oedema of the face from sesquiterpene lactones. This may look like a photosensitivity and be enhanced by sunlight (see above). Avoidance of the plant may be impossible for someone whose job depends on contact with it, and is a special problem where it is a common environmental weed. In the Poona region of India the weed *Partheneum hysterophorus* can lead to death. Dermatitis resulting from contact with this plant builds up into a severe erythroderma with secondary infection and even pseudolymphoma. In those who are suffering from other diseases it may summate and lead to a very severe illness.

Potentially allergenic plants are most numerous in the cashew family, such as poison ivy, poison oak, poison dogweed, elder or sumac, mango, wax, or lacquer trees, and hence in one form or another they are present worldwide. Attacks can be aborted by washing within an hour of contact. Severe oral dermatitis and acute gastrointestinal systems can be troublesome when sensitivity is due to the mango or cashew nut.

Contamination of other handled agents can also cause outbreaks of dermatitis, as has been recorded with articles of clothing, mail, and even from voodoo dolls.

Wood dermatitis is often due to its resins or its attendant lichens, liverworts, and mosses, and even its insect parasites are occasionally responsible. It is a severe cause of industrial dermatitis in workers in the furniture industry. Furthermore, it can even cause mouth dermatitis in children handling wooden toys, and, in music classrooms, has also been noticed in those playing recorders made of certain woods.

Medicaments

An enormous number of medicaments are now used, often containing unknown constituents. The problem particularly arises where these have been repeatedly applied to the skin over a number of years, and therefore a contact dermatitis is found in patients with leg ulcers, pruritus ani or vulvae, and in those suffering from otitis externa. Local anaesthetics, lanolin and cetylstearyl alcohols, antibiotics and antiseptics, antifungal compounds, and antihistamines are the most significant groups of causative agents. Topical corticosteroids are increasingly recognized as relatively common contact sensitizers (hydrocortisone, hydrocortisone-17-butyrate, budesonide).

An example of the importance of recognizing such sensitivity is illustrated by ethylene diamine; this is a common cause of dermatitis and is present in certain neomycin–nystatin ointment mixtures (for example, Tri-Adcortyl®) and in aminophylline suppositories. It is used as a solvent in many industries and is one cause of coolant-oil dermatitis. This combination of an industrial use and a medicament may mean that a person loses his employment as a result of the previous use of a medicament. Sensitivity to ethylene diamine has serious implications, since it is also sometimes used as a preservative of intravenous aminophylline, and deaths have been recorded.

Metal

Beryllium, used in the manufacture of fluorescent lights, causes skin ulcers, dermatitis, and granulomas. Chrome confers hardness to metals, and dermatitis from it is also common in the leather tanning industry. It is a contaminator of cement. In industrial countries it is one of the most common sensitizers in men; most obtain their sensitivity from cement, but their greatest disability is due to the later inconvenience of being unable to wear leather footwear containing chrome. Ferrous sulphate can be used as an additive in cement to convert hexavalent chromium to the less sensitizing trivalent form.

Cobalt sensitivity is commonly found in association with nickel or chrome sensitivity. Jewellery, and possibly metal prostheses for hip replacements, may be responsible.

Nickel is used in various metal alloys, electroplating, enamels, and glass. It is easily absorbed through the skin and its presence in body-piercing rings or studs and in buttons and clips probably accounts for the high incidence of metal dermatitis. Sensitization is particularly common with jewellery made of nickel-containing alloys; gold of 14 carats and above is considered safer. There is a general worldwide trend towards increased nickel sensitivity, which contributes substantially to hand dermatitis. Simply handling nickel-containing money or pots and pans does not seem to be responsible. However, the abrasive cleaning of such in washing-up water releases nickel, and is a reason for blaming this occupation or for recommending the use of running water.

Employment and contact dermatitis

It will be seen from the above that many people working in industry are liable to contract specific types of 'contact dermatitis'. Some industries are particularly susceptible.

Hairdressers

During their apprenticeship, the hands of hairdressers suffer from the very abnormal wear and tear of frequent shampooing. The skin of atopic subjects almost always break down. Nickel dermatitis is particularly common. When the rash only affects the palmar surfaces, contact dermatitis is more likely than irritant dermatitis. The latter commonly affects the more tender dorsa of the hands and between the fingers.

Bakers

Dough, sugars, and fruit and vegetable peels are irritants often causing considerable skin damage in atopic subjects. Many of the flour additives can cause contact urticaria.

Builders

Cement is highly irritant but skin quickly hardens. Severe alkaline burns of the lower legs from calcium hydroxide in wet cement is now well recognized, especially in the amateur using ready-mixed cement.

Chrome dermatitis may be very similar to constitutional patterns, including seborrhoeic and stasis eczema, and for this reason anyone in the building industry who has any pattern of eczema should be patch-tested.

Agricultural and horticultural workers

Carelessly used fungicides and pesticides are frequent causes of dermatitis. This particularly occurs on isolated farms in developing countries.

Patch-testing

The principle of patch testing is to apply the suspect agent to the patient's skin, but avoiding irritants, and observe its effect on cell-mediated immunity. It involves:

1. applying the agent on a carrier material such as aluminium foil on filter paper, and covering with adhesive tape;
2. using a non-irritant concentration of the agent in white soft paraffin in water or ethyl alcohol; for most chemicals this is 0.1 to 1 per cent (in the case of cosmetics or medicaments the concentration used in the whole product is suitable);
3. applying (1) to the patient's back, which gives a more consistent response than the arms or legs; and
4. removing the covering adhesive tape and filter paper with aluminium foil 2 h before reading at 2 and 4 days.

Most practitioners obtain reagents from Trolab, Karen Trolle-Lassen, Land, Pharm 6B AN, Hansens Alle, 2900 Hellerup, Denmark, and replace them about every 6 months.

False-positives result from sweat gland occlusion, sensitivity to adhesive tape, irritants, and generally increased irritability of the skin, usually due to active eczema but exposure to ultraviolet irradiation can also be causative.

A positive patch test is a papular and a palpable erythema, which may be vesicular (Fig. 38).

Fig. 38 Contact for about 48 h with the allergen to which the patient is sensitive can be used as a test at any site on the skin. This is the basis of the patch-test reaction. In this case a finger dermatitis due to an allergen in cigarette smoke could be proved by applying a smoked filter paper to the patient's back.

Treatment of contact dermatitis

The level of complaint is often lessened by good industrial relations or a happy home. Those who are well satisfied with life may call their problem merely roughness of the skin; those who are unhappy or dissatisfied may well call their problem dermatitis. Especially in those who have atopic eczema or psoriasis, emotional stress is considered to be a factor worth controlling if possible. Such stresses are often no more than the anxieties and irritations of daily living and employment in a complex society.

Elimination of known irritants or allergens must be attempted but, as in the case of poison ivy in the United States or some of the Compositae in Asia, complete avoidance may be impossible. For less severe allergens, such as chrome or nickel, the skin can settle to a tolerable degree merely by removing obvious sources in clothing or jewellery. Dermatologists can encourage cleanliness and ventilation in working environments and the substitution of less allergenic materials in industrial processes. It is not always advisable to make workers change their jobs; this particularly applies to chrome sensitivity in building industry workers, since once they are sensitized, most other jobs are equally difficult. However, most sufferers can manage by taking a little more care at work and with the help of emollients. Anti-inflammatory agents, such as steroid creams, are always of help and can help the affected person stay at work, particularly where exposure is just short-term (for example, during the training of hairdressers or nurses).

Severe chronic allergy can be relieved by immunosuppressive drugs such as azathioprine. Chelating agents, such as Antabuse®, have been used in cases of severe nickel dermatitis.. Nickel-free diets are complicated but much less unpleasant than drug therapy.

The prognosis for contact dermatitis is often good. Thus 30 per cent of nickel dermatitis of the hands is healed in 6 years. Only 25 per cent of apprentice hairdressers with hand dermatitis have to change their job. Only rarely, as with certain plant allergies, is the problem a persistent and intolerable problem affecting many persons in the community.

Contact urticaria—latex gloves

This is an acute swelling developing within a few minutes to half an hour of contact with certain agents. In atopic eczema there is an particular susceptibility to this phenomenon, but it is also well recognized in non-atopic subjects and is particularly common as a result of the application of cosmetics. Many agents commonly applied to the skin will produce irritation in certain sites, such as the eyelids or scrotum, and this is not always an immunological phenomenon. Urticaria from latex occurs within 15 to 30 min of wearing the glove. It is IgE-mediated. The rash may extend beyond the area of exposure. Doctors and dentists are at particular risk, as are patients frequently exposed to latex gloves such as spina bifida patients who self-catheterize. Patients may crossreact to condoms or medical instruments such as intubation tubes, as well as some fruits and vegetables such as bananas and avocados. Severe sensitivity may induce anaphylaxis and such patients should always carry adrenaline (epinephrine). In extreme cases anaphylaxis can occur to airborne contact with latex particles in procedure rooms, and affected doctors may need to seek an alternative vocation. Diagnosis is by radioallergosorbent testing (**RAST**); prick-testing should only be done where resuscitation equipment is available. Prevention by the use of non-powdered, high-quality rubber gloves or vinyl gloves is prudent, especially when the hands are already damaged.

Atopic eczema

This is a constitutional disorder of the skin affecting about 10 per cent of the population. It is one of the most common diseases of childhood and one of the main reasons for loss of work in industry. It accounts for about 50 per cent of cases of hand eczema. Its inheritance is discussed in Chapter 23.2.

Atopic dermatitis is a multifaceted disease, the cause of which is still unknown. Patients with atopic dermatitis frequently have an elevated IgE level. Allergic respiratory disease affects about 50 per cent of eczema sufferers, and 70 per cent of patients are aware of other family members with the disease. Other associations include dry skin, facial pallor, low finger temperature, pronounced vasoconstriction on exposure to cold, white dermographism, and a susceptibility to cutaneous viral and bacterial infections. Ophthalmological manifestations of atopy include infraorbital folds, infraorbital darkening of the skin, conjunctivitis, keratoconus, and cataract formation. Hyperpigmentation of the lateral neck is known as atopic 'dirty' neck. Drug reactions of the anaphylactic type are more common and abdominal symptoms due to food allergy are frequently described. Contact urticaria is common. Alopecia areata may be more severe and less likely to respond to treatment. Around 60 per cent of patients present within the first year of life, and for 90 per cent the disorder starts within the first 5 years of life. In the majority, the eczema gradually improves but the skin remains vulnerable to physical and chemical irritants throughout life.

Data from the International Study of Asthma in Childhood have shown a global prevalence of 2 to 16 per cent of children aged between 6 and 7 years over a 1-year period.

Pathogenesis

Atopic dermatitis is essentially an exaggerated response to environmental irritants and allergens. The basis of this exaggerated response is immunological, with an altered T-cell response and consequent production of proinflammatory cytokines. In atopic patients, CD4-positive T-helper cells, and in particular class 2 helper cells (T_H2), recognize antigens and secrete IL-4, IL-10, and IL-13, which stimulate B-cell growth and IgE and mast-cell production. In addition, T_H2 cells produce IL-5 which leads to eosinophilia. This may provide atopic patients living in tropical areas with relative protection from parasitic infections. T_H2-derived IL-10 inhibits T_H1 cytokine production, while IL-4 promotes T_H2 differentiation. T_H1 cells produce interferon-gamma (IFN-g), IL-2, and IL-3. IFN-g produced by T_H1 cells enhances the development of T_H1 cells and inhibits T_H2 cells. Reduced childhood exposure to infections and intestinal bacteria or parasites, as well as increasing vaccination and immunization, are proposed as causes of T_H1 and T_H2 imbalances favouring atopy.

Epidemiological studies have incriminated a number of allergens that may trigger this immunological response, including pollution, central heating, and house dust mites. Unfortunately, the identification and avoidance of allergens in atopic dermatitis has proven to be more complex than in asthma and hayfever. Skin tests by pricking various antigens into the skin result in weal-and-flare responses that are often multiple and strongly positive. However, there is a poor correlation between the skin-test response and the activity of the eczema, which may even be in remission during, for example, the hayfever season in spite of strong reactivity to skin testing with grasses.

The role of food allergy is difficult to test accurately in such a fluctuating and multifactorial disease. Neither prick tests nor allergen-specific IgE tests can be used to predict those most likely to benefit from dietary elimination. Exclusion diets and food challenges rarely detect meaningful dietary allergens and are therefore reserved for the most severely affected patients. Exclusive breast feeding is of benefit, but is not necessarily due to the avoidance of dietary factors. Breast milk contains much IgA. Complete avoidance of cows' milk during the first 6 months of life in a child born of atopic parents is believed to be beneficial, but nevertheless there are many who fail to benefit.

Humoral immunity

T cells infiltrating the skin in atopic eczema are predominantly IL-4/IL-10/IL-13-producing Thy-2 responses that encourage B cells to produce excessive IgE. IgE reagenic antibody is elevated in over 80 per cent of patients with atopic dermatitis, often to over 2000 units/ml. However, the significance of this finding remains uncertain, since atopic eczema can occur in agammaglobulinaemia and normal levels are found in many actively eczematous patients. Of course, serum levels need not reflect the level of activity of the IgE surrounding the mast cell in the skin itself. Complexes with antigen and the mast cell are often, after all, the basis of the IgE-mediated weal-and-flare response. There is an increased frequency of the presence of specific reaginic IgE antibody also known as the test reagent radioallergoabsorbent test (RAST) to numerous allergens in the sera of atopic people.

Reactivity of the immune system

About 80 per cent of atopic patients demonstrate an excessive reactivity of their immune system, reacting to certain foods and to house dust with immediate itching and swelling of the tissues. The agents to which 90 per cent of atopic patients react differ from those usually encountered in allergic disease—these include a number of animal and vegetable proteins from milk, meat, and corn. Eczema itself is most typically a consequence of delayed or cell-mediated immunity; however, T-cell function seems to be depressed in atopic skin, leading to a greater susceptibility to viral, bacterial, and fungal infections. Herpes simplex, vaccinia, warts, *Staphylococcus aureus*, and *Trichophyton rubrum* infections are most favoured; 90 per cent of atopic subjects carry *Staphylococcus aureus* in their skin, compared to 10 per cent of normal subjects. Fortunately, the atopic subject is not as susceptible to strains responsible for impetigo, toxic epidermal necrolysis, or furunculosis. There is also decreased reactivity to other common allergens such as poison ivy, *Candida* spp., or dinitrochlorbenzene (**DNCB**).

An atopic eczema-like syndrome is also a feature of immune deficiency diseases; for example, the Wiskott–Aldrich syndrome, the hyper-IgE syndrome of Buckley, Jobs, and Jung, and thymic aplasia, as well as the DiGeorge and Nesselof syndromes which also demonstrate high levels of IgE and decreased cell-mediated immunity. It is possible that immaturity of the humoral antibody system results in defective T-lymphocyte regulation, and that IgE production is increased as one consequence.

Characteristics

Atopic patients have an inherently dry and irritable skin. Itch and scratching are responsible for most of the skin changes seen clinically and histologically.

A low itch threshold

A diagnosis should not be made if there is no history of itching. Besides the usual causes of itching, many minor irritants, such as woollen clothing or a change in climate, cause scratching. Scratching causes excoriations and ulceration, as well as thickening of the epidermis and swelling and redness of the underlying dermis. The broken surface is sore and further irritated by soaps, some ointment bases such as sorbic acid, sea-water, or citric fruit juices. It has been found, using intradermal trypsin as a test of itch, that atopic patients have a prolonged itch reaction, although it may be that other patients with other forms of eczema are similarly affected.

Dry and lined skin

In non-excoriated areas the skin is often dry and lined. This is more obvious in hard-water areas in temperate climates. The palms are particularly heavily lined and cause embarrassment even to the fortune teller! About 70 per cent of adult patients have a hand dermatitis that usually spares the palms. In nursing mothers, nipple eczema may be a problem during breast feeding. It seems there is a deficiency in sweating and sebum excretion leading to chapping, wear, and tear, particularly from solvents or water. In industry, people with atopic dermatitis are less tolerant of contact with primary skin irritants and therefore more likely to develop occupational dermatitis. Another feature of atopic dryness is keratosis pilaris, which is a perifollicular hyperkeratosis.

Vasodilatation

The vasculature too readily vasodilates in the popliteal and cubital fossas, thereby heating the skin and hence inappropriately lowering the itch threshold. When scratched, rubbed, or stretched, the skin blanches for a few minutes, beginning 12 to 15 s after injury. This is partly due to upper dermal precapillary shutdown and also to persistent inflammatory oedema. Deeper vessels often dilate so that the skin is warm. This combination of hot but pale skin accounts for the itching as well as the atopic pallor.

Clinical features

Itching is the chief feature and becomes apparent during the first 2 to 6 months of life. The face is usually first affected and scratching begins between the second and third month. Sore lips from licking and chapping as well as conjunctivitis with ectropion are common; 70 per cent of patients have a skin fold or wrinkle just beneath the margin of the lower lid of both eyes (Fig. 39). When the child begins to crawl, exposed surfaces such as the knees and hands become the most involved. The papules are scratched and become exudative and so secondary infection associated with lymphadenopathy is a common finding. Lymphadenopathy can sometimes be so gross as to lead to the suspicion of some dire malignant disease. From 18 months onwards the sites most characteristically involved are the flexures of the elbows, knees, sides of neck, wrists, and ankles (Fig. 40). Local areas of lichenified skin may persist at such sites, and the face, too, may be heavily lichenified. Rubbing the eyes does not fully explain why keratoconus and anterior subcapsular cataracts are featured in severe cases. Seasonal influences on the disease are mainly climatic, due to sunlight and humidity and the use of central heating, but are probably also related to seasonal allergies. Pollen is a feature of spring and early summer, while house dust seems to be a feature of late summer allergies.

Fig. 39 Atopic eczema in an adult, showing the characteristic skin fold just beneath the lower eyelid and the loss of eyebrow hair, as well as thickening of the skin due to rubbing.

Fig. 40 Typically thickened and excoriated skin of the chronic prurigo of atopic eczema.

Prognosis

Most children develop their eczema within the first 6 months of life, but about one-fifth of patients may have a delayed onset, even into adult life. Generally, there is a tendency for gradual improvement, with many children 'growing out' of their atopic dermatitis around the age of 3 to 4 years. Complete clearance without breakdown when in contact with skin irritants is unusual, but most people, in the absence of major irritants, are clear by the time they are teenagers.

Management

In the absence of any treatment that permanently modulates the immune response, the main focus of treatment is to reduce exposure to environmental irritants and to prevent overheating. It is useful for parents to have access both to the doctor and nurse as well as to the literature provided by patient groups. All factors that irritate the skin should be even more avoided by atopic patients, and these include various primary irritants such as soap, wool, and extremes of climate. Moisturization of the skin is good, but evaporation is bad. Wet wrapping is the application of wet dressings over moisturizing creams, which are then covered by dry dressings. Liberal washing with soap substitutes based on emulsifying ointments is mostly helpful, and these are most effective if applied at least four times a day. The common-sense avoidance of jobs involving large amounts of primary irritants should be advised. Attempts should be made to avoid contact with herpes simplex, molluscum contagiosum, and other viruses affecting the skin; however, vaccination poses no particular problems for children with atopic dermatitis and should not be withheld, unless the child is being treated with systemic immunosuppressants. It is still difficult to know how to remedy an immunological defect. While breast feeding is to be recommended for all infants, it may be particularly important for babies with a strong family history of eczema. There is evidence that breast feeding may reduce the incidence of atopic eczema by up to two-thirds, though not all authors agree. It is postulated that a period of transient immune vulnerability exists during early life, during which exposure to food antigens, perhaps by complexing with IgG, IgM, and IgE instead of IgA, results in allergic sensitization and the subsequent development of atopic eczema. The effect of breast feeding may be due to its low antigen load compared to cows' milk. Breast milk is also rich in IgA, which may modify the absorption of food antigens. Breast feeding should continue for an extra 3 to 4 months, since any supplement exposes the immature gut to foreign protein. Especially to be avoided is any supplement given in hospital during the first week of life. Further benefit might be gained by avoiding eggs and cows' milk in the mother's diet, since foreign protein can be transmitted through the mother's milk. When breast feeding is impossible, milk substitutes are second-best since they are expensive and require care to prevent bacterial contamination. Some paediatricians believe cows' milk should be avoided for 1 year and eggs for 18 months.

Some patients appear to benefit from a regime of antigen avoidance, although this is not always the case. Neither RAST tests nor skin-patch tests are reliable in selecting patients who are helped by dietary modification, since even those known to benefit from antigen avoidance may not show specifically raised IgE levels nor positive skin tests.

Elimination diets must be carefully assessed to obtain complete avoidance with, at the same time, adequate nutrition, and since they are not without risk in this respect they should be reserved for the most severely affected children. Some authors recommend the avoidance of eggs, chicken, milk, and artificial colouring agents or preservatives. Goat's milk has lost favour on nutritional grounds. Careful studies of the use of Chinese herbal teas have shown an improvement in generalized dry eczema in children, but users should be aware of the potential hepatotoxicity of these drinks.

Other environmental allergens shown to be important for some children include the house-dust mite. Avoidance includes removing dust-collecting fabrics such as carpets, curtains, bedclothing, and soft-furnishings, and using high-powered vacuum cleaning rather than brushing. A cold and dry environment discourages the mite. Polished floorboards, hydronic rather than ducted heating, blinds rather than curtains, plastic mattress and pillow covers, washable cotton blankets rather than duvets are all appropriate (albeit often expensive) measures in severe cases regardless of whether a house-dust mite allergy is established. Cats and dogs should always be kept off the bed, and avoided all together if contact with them worsens the dermatitis.

Topical therapy

Apart from the liberal use of emollients, steroid creams are effective antipruritic agents. Ointment bases are preferred for dry dermatitis, while creams are used in flexures and for weeping infected or acute dermatitis. Prescribing habits have changed over the past few years. Certainly there was a period when overprescription resulted in systemic side-effects as well as local atrophy of the skin; however, underusage, often encouraged by the dispensing pharmacist, is now far more prevalent. Withholding of steroids deprives the child of the one effective therapy. Short, sharp bursts of effective therapy with strong steroids may be entirely justified, but prolonged daily usage of weaker, partially effective steroids is bound to lead to complications. However, special care is required on the face and flexures. Topical steroids are stored in the skin, and for this reason once-daily application may be sufficient. It seems an inexplicable fact of life that ringing the changes with ointments is of benefit, and a skilled practitioner will always have an alternative preparation on which the worried parents can pin their faith. Secondary infection is so common and bacterial allergy so important that vigorous treatment of infection is justified, and topical antiseptics and systemic antibiotics should be given according to the sensitivities of the bacteria. Erythromycin is particularly valuable; mupirocin, topically, is as effective. The matter of climatic therapy remains unpredictable; undoubtedly a change of climate does effect great improvements in some children, whether it be exposure to sunlight or to the sea or to a mountain top.

Ascomycin macrolactam derivatives formulated as a topical cream inhibit the release of inflammatory cytokines from T cells and show great promise as a topical therapy.

Severe cases of eczema, as with asthma, may have to be controlled by systemic steroids, either in the form of prednisolone or corticosteroid injections. This may be simply to help the patient over an acute period, although a small minority of patients may require long-term therapy for several years. Azathioprine is an effective steroid-sparing agent. Ciclosporin controls severe eczema but provides little long-term benefit. Periodic hospital admission with intensive topical therapy can provide months of respite. Ultraviolet light in air-conditioned cabinets is helpful for many people. An initial worsening often requires cover with prednisolone. Traditional herbal remedies are popular and controlled studies have shown some benefits from Chinese medicinal plants, but it should not be assumed that because they are 'natural' they are safe. Significant hepatic toxicity has been documented and a carcinogenic potential of some of these agents is suspected.

Other patterns of dermatitis

Infected dermatitis

Increasing evidence suggests that bacterial allergy plays a part in the development of an eczematous response in the skin, *Staphylococcus* spp. being particularly implicated. Bacterial allergy may play a part in all types of eczema, but occasionally it is the single cause. This is most frequently seen as a rather well-demarcated patch of eczema with crusting and scaling on an exposed area. There may be small pustules on an advancing edge. It is seen around discharging wounds, around ulcers, and occasionally around a paronychia or in a flexure, subject to sweating and maceration; it is particularly common around the ear or at sites of occlusion such as under a hat-band or between the toes. An underlying pediculosis may be one trigger. Black skin commonly seems to develop a similar condition that principally affects the shins. Management includes the use of local antiseptics and wet soaks, or dyes, such as gentian violet, combined with an appropriate systemic antibiotic.

Herpes simplex infection often presents with a sudden inexplicable flare of atopic dermatitis, often monthly, but this may be difficult to prove as vesicles are rapidly

excoriated. Swabs for immunoperoxidase staining or viral culture may need to be frequently repeated. A trial of prophylactic oral antiviral therapy can be considered in cases with a suggestive history.

Seborrhoeic dermatitis

Adult seborrhoeic dermatitis is different to the infantile disorder bearing the same name. It mainly affects the scalp and face, but can also involve the upper trunk and flexures including the axillas, groins, scrotum, and anus. The aetiology is unknown but the distribution does appear to be in the areas of sebaceous activity. There is a strong association with neuroleptic-induced parkinsonism, idiopathic parkinsonism, spinal injury, as well as with AIDS. *Pityrosporum orbicularis* may be the responsible pathogen. The oval blastosphere is predominant in AIDS, whereas the hyphal form is increased in pityriasis versicolor. The most characteristic lesion is a dull or yellowish-red and greasy plaque with a marginated scale. On the scalp it produces dandruff. On the face it tends to involve the medial cheeks, nose, nasolabial folds, and eyebrows. It is the most common cause of a 'butterfly rash'. Seborrhoeic dematitis affects the axillas and groins with well-defined brownish-red scaly areas deep into the folds, on the front of the chest and in the middle of the back there may be small brown follicular papules covered by greasy scales or multiple discrete patches. Rarely, a widespread eruption resembles pityriasis rosea with oval lesions with peripheral scale. Severe cases of seborrhoeic dermatitis develop marked crusting and scaling, particularly of hair-bearing areas and the genitalia. Otitis externa is one manifestation. The disorder tends to recur and may be chronic.

Management includes an attack on local infection and the removal of crusts with wet soaks. Preparations, such as vioform hydrocortisone, sulphur, and ichthammol in a variety of water-miscible bases, usually in 1 to 2 per cent concentrations, have traditionally been prescribed. Lithium succinate ointment is recently favoured. Imidazoles control pityrosporum overproduction, which is thought to play some part in the diathesis. Antiyeast shampoos with zinc pyrithione or ketaconazole are effective, as are tar shampoos.

Nummular (discoid) eczema

The main feature of this eczema is that it is discoid or composed of rounded lesions scattered, often symmetrically, over the body. They are intensely vesicular and intensely itchy. Undoubtedly endogenous, external influences play little part in their development, although occasionally sensitivity to metals, such as nickel or chrome, may produce a similar picture. Onset is usually in adult life, although Asian children seem prone to this form of dermatitis. Patients are no more likely to be atopic, but dry skin and overheating are important aggravating factors. Secondary infection is common; sometimes nummular eczema is as a reaction pattern to a localized primary irritant such as an insect bite.

Pityriasis alba

This is a pattern of eczema quite common in children, often in those with darker skins, in which a very low-grade dry eczema with shedding of pigment transiently gives rise to a white patch of skin (see Fig. 64). It may be associated with drying out—reduced sebum—around the hair follicles, known as keratosis pilaris.

Fig. 64 Pityriasis alba caused by a mild dry eczema. Slightly scaly areas of depigmentation are a common cause of discoloration.

Itch without rash (pruritus)—mechanisms and causation

'Pruritus' is the term used when itching (the most prominent symptom in skin disease) is the primary complaint, which leads to scratching in the absence of visible evidence of lesions predisposing to itch. Itch is a sensation largely dependent on superficial nerve endings—unmyelinated C fibres—in an intact upper dermis and epidermis. These are very thin fibres, but rich in terminal branching. Thinly myelinated nerves in lateral spinothalamic tracts and secondary neurones to the thalamus relay both pain and itch, and the cerebral cortex can modify these responses. Itch is induced by a number of agents including histamine and histamine releasers such as substance P, opioid peptides, cytokines, bradykinin, bile salts, and proteases, and is potentiated by prostaglandin E. It can be disassociated from pain in hypoalgesia. Central neurological and emotional psychiatric factors control the threshold to itch and pain. Awareness is a complex attribute modifying or intensifying the response to the itch. The itch itself may cause irritability, depression, or invoke the attitude of the masochist who wears a 'hair shirt'.

Itching is usually worse when the skin is heated to normal body temperature and when there is little else to distract the sufferer—a combination common at night. Vasodilatation in the cubital or popliteal fossas partially accounts for the lower itch threshold at such sites in people with atopic dermatitis.

The itching threshold is lowered by isolation, including the common accompaniments of ageing such as blindness, deafness, and loneliness. Endogenous depression is often missed in the elderly and should be treated. Paroxysmal itching may originate in the central nervous system and provoke deep scratching, which is pleasurable but injurious. It is a feature of cocaine addiction.

The itch of different dermatoses evokes different types of scratching. Urticaria is almost never scratched but usually rubbed or pinched, perhaps because the exact site of the itch is difficult to pinpoint. Similarly, excoriations are rare in lichen planus (see 'Localized pruritis', below). Where intense itching is exactly located it is often persistent and deeply excoriated.

A common factor is dryness and desiccation of the stratum corneum, common in the elderly and worse in winter. Sweat retention also causes intense pruritus such as prickly heat. People recently engaged in insulating their roof with fibreglass suffer from pruritus caused by the almost invisible spicules of fibreglass (see above and Fig. 34).

Pruritic conditions

Parasitic causes of itching (scabies) (see Chapter 8.2)

Parasites are an important cause of pruritus, but those experienced at examining the skin will usually observe primary urticarial or papular lesions in amongst the scratches. Onchocerciasis, trichinellosis, and schistosomiasis cause severe pruritus, usually with marked eosinophilia as well as urticaria, prurigo, and depigmentation. In onchocerciasis, loss of elasticity and the development of a leather-like skin hanging in folds is one consequence.

Delusions of parasitosis are common, affected people usually present to pest exterminators and museum entomologists rather than to doctors. This condition is a monosymptomatic delusional disorder, and patients often function well in other aspects of their life. Follie aux deux is common, and the entire family may be drawn into the delusional framework. Referral to a psychiatrist is often resisted. It is best treated with antipsychotic drugs such as pimozide.

Aquagenic pruritus

This occurs after contact with water—fresh, salt, or sweat. In some people it starts from the moment of contact and lasts 15 min. In others, it is less immediate and

longer lasting. Treatment with acetylcholine or histamine antagonists or ultraviolet rays sometimes help. A common similar reaction in the elderly, due to rapid drying out after prolonged hydration, is helped by shortening the period of hydration. It is a premonitory sign of myeloproliferative disease.

Generalized pruritus and systemic disease

Hepatic disease

Obstructive jaundice causes severe pruritus. It is particularly an early feature of biliary cirrhosis, and bile salts rather than bilirubin have been held responsible for the itch. The degree of jaundice need not be great. Bile salts in the skin can achieve a relatively higher concentration than may be indicated by serum levels. A bile-salt concentration of 1 mmol/l causes itching when applied to a blister base; dihydroxy salts, especially chenodeoxycholate, are responsible.

Contemporary studies look to opioid metabolism as a cause of the pruritus, and these early studies on bile salts are only part of the story. Oestrogen-induced pruritus of pregnancy or from the contraceptive pill often shows little or no jaundice in spite of intrahepatic biliary obstruction, severe pruritus, and a much increased alkaline phosphatase level. Chlorpromazine and testosterone can have the same effect.

Blood disease

Iron deficiency has also been blamed for itching even when the patient is not anaemic—though in iron deficiency, which is common, itch is rarely found to be so associated. Some thinning of hair is a frequent complaint. Polycythaemia is frequently associated with itching, particularly after a hot bath. It is believed to be related to blood histamine levels and occasionally to iron deficiency, and hence the reported positive response to iron therapy within 2 to 10 days. Lymphatic leukaemia is another cause of pruritus often long-lasting before it becomes clinically overt.

Carcinoma of the internal organs and lymphoma

Carcinoma of the bronchus in particular may very rarely present with generalized pruritus. Pruritus occurs in 25 per cent of patients with Hodgkin's disease, often burning in quality and associated with icthyosis.

Chronic renal failure

No matter how uraemic is the patient, itching is not a feature of acute renal failure or even of malignant hypertension. Patients with chronic pyelonephritis or chronic glomerulonephritis usually suffer greatly from pruritus, which is not necessarily relieved by haemodialysis. Parathyroidectomy, for reasons that remain obscure, may relieve itching in those in whom removal of the gland is necessitated by secondary hyperparathyroidism. The cause of pruritus in renal failure is unknown, but raised histamine levels, endogenous opioids, and dryness of the skin are factors. Mast cell numbers are also increased in patients with chronic renal failure.

Endocrine disease

Pruritus is sometimes a presenting symptom of diabetes mellitus but generally this is principally localized to the vulva. About 1 in 10 patients with hyperthyroidism complain of itching. Dry skin in hypothyroidism often itches.

Drugs

Morphine, allopurinol, or those causing cholestasis should be enquired about.

Management of pruritus

Overheating should be avoided as should vasodilators such as alcohol and hot drinks. Calamine lotion is used as a cooling agent: all topical therapy is more cooling if kept in a refrigerator. Evaporation is increased by the enhanced surface area provided by the powder. Dryness of the skin should be discouraged by the use of emollients. Oily calamine or 0.5 per cent menthol in aqueous cream are preferred when xerosis coexists with itch. In dry conditions, such as the hospital ward, a moist microenvironment can be enhanced by appropriate clothing. A sensation of cooling can be achieved with 1 per cent menthol or camphor and 1 per cent phenol, both of which have a mild anaesthetic effect. Menthol is also an antihistamine. Nails should be kept short and together with occlusive bandaging may reduce the vicious circle of itch and skin damage. In general, woollen clothing is itchy, cotton clothing is not. Too frequent bathing or showering should be discouraged unless emulsifying ointments are added to the bath as soap substitutes. Bath salts should be avoided. Proprietary bath oils are more cosmetically acceptable but tend to be expensive for regular daily use.

Obviously, any known cause of the pruritis should be treated accordingly. Class I antihistamines, which pass the blood–brain barrier, may principally act through their sedative and anticholinergic effects, and they also reduce awareness. Class II antihistamines, which do not pass the blood–brain barrier, cause no sedation; but as histamine is one of the most prominent mediators of itch, they should always be tried and are often most effective in higher dose than indicated in the National Formulary. The role of increased histamine release and its mediation of itch in senile pruritus justifies the prescription of a class II antihistamine. Chlorpromazine may reduce the reactivity to the itch. Plasma exchange has been used to control sweats and pruritus. The anion-exchange resin cholestyramine, 6 to 8 g daily, or oral activated charcoal helps to relieve the pruritus of liver disease and sometimes its use in patients with chronic renal disease or polycythaemia has been helpful. Suberythema doses of UVB irradiation twice weekly, and even natural sunlight, often ameliorate pruritus, reducing mast cells and some of the cytokine activity inducing inflammation; they have also been used to treat the itching associated with uraemia and with certain acute exanthems such as pityriasis rosea. Use of hydroxyethyl rutosides (Paroven®) and thalidomide have been advocated in the treatment of patients in renal failure. Pentoxifylline (oxpentifylline) is reported to relieve the pain and itch of keloids. Opioid activity can be blocked by opioid antagonists or competitors such as codeine; these have been advocated for the treatment of pruritus of liver disease. The H_2-receptor antagonist cimetidine is sometimes helpful in Hodgkin's disease.

Localized pruritus

Localized intensely itchy areas of skin having no obvious causation are a common problem in the dermatology clinic. The nape of the neck, upper back (Fig. 41), genitalia, lower leg, elbow, and outer thigh are easily accessible sites liable to persistent rubbing and scratching. Such injury to the skin results in thickening, purple-brown violaceous coloration due to dilated vessels and postinflammatory pigmentation. The normal line marks of the skin are exaggerated and excoriations are usually numerous. This is termed 'lichen simplex' or 'neurodermatitis', and the fairly well-defined patches cause paroxysms of itching and emotional upsets with anxiety or irritability, which, in themselves, are also promoting factors. Capsaicin ointment, which reduces substance P in the skin, may relieve localized pruritus after several applications. The burning initially induced by this therapy can be reduced by the prior application of local anaesthetic creams (for example, Emla (lidocaine (lignocaine) and prilocaine cream)).

Fig. 41 Prurigo nodularis is a form of scratched lesion which is very exactly localized. The upper back is a common site for such persistent excoriation.

Nodular prurigo is an unexplained reaction to scratching, evoking severe, very localized pruritus. The nodules are 1 to 2 cm in diameter and scattered over accessible areas. It is sometimes a consequence of a partially resolved, more generalized pruritus arising from atopic eczema or parasitic infestation. Freezing with liquid nitrogen is helpful, but this may lead to depigmentation in pigmented races.

Local steroids are helpful and anything that protects the skin from scratching may eventually allow healing. Occlusive tape or bandaging is occasionally helpful but secondary infection is a problem, especially in hot countries.

Intralesional injection with triamcinolone causes rapid resolution in some cases but this may be only a temporary response, and where the lesions are large or multiple such inoculation is not without the side-effects of steroid therapy. It is always worth admitting such patients to hospital and treating them with traditional dermatological therapies such as tar bandages. A more recent suggestion is that some of these lesions are an immunological response and this has led to therapies as far ranging as azathioprine and thalidomide.

Pruritus ani

Pruritus ani is common in White adult males. It is rare in Blacks except as a manifestation of infestations such as oxyuris, lice, and scabies. Psoriasis, atopic dermatitis, seborrhoeic dermatitis, tinea cruris, candidiasis, and streptococcal infection can all be found in the anal region, albeit rarely, and should be excluded. Extramammary Paget's disease is a rare cause of itch. People with diabetes are prone to pruritis ani for reasons not entirely clear. An important cause is soiling of the perianal skin. Haemorrhoids, fissures, and fistulas contribute to this problem, as can rectal carcinoma. The anal sphincter relaxes in response to anal distension too readily in some sufferers; in others, incomplete bowel evacuation leaves some residual faeces in the folds of the anus. Because soft stools are more likely to cause irritation, fibre intake needs to be carefully balanced to keep the stool firm while at the same time preventing constipation. Pruritis ani is commonly exacerbated by hot, spicy foods: curry and coffee should be limited. Bacterial and fungal contamination is common, and anxiety may lead to excessive hygiene measures that irritate the skin or result in supervening lichen simplex. Both allergic and irritant contact dermatitis may result from the common use of multiple medication.

Often, the perianal skin needs to be cleaned immediately and an hour or two after the bowels have been opened. Weak- or medium-strength local steroids are the mainstay of treatment, and can be mixed with anticandida or antiseptic agents to deal with or prevent secondary contamination and infection.

Pruritus vulvae and vulvodynia

The vulva has a generous innervation and is highly sensitive. Itch or pain or both are common when the skin is irritated or inflamed. It is socially unacceptable to scratch one's vulva in public and these symptoms are often associated with guilt, a sense of being unclean, dyspareunia, and marital disharmony. Furthermore, sexuality is central to a woman's psyche and primary psychological problems or guilt over an extramarital affair may somatize with a vulval or vaginal complaint. When marital problems exist, a vulval problem may be the justification to reject sexual advances, and sympathy from the doctor rather than a cure may occasionally be sought.

The vulva is a favoured site for candidiasis, tinea cruris, atopic dermatitis, psoriasis, seborrhoeic dermatitis, lichen sclerosis, and lichen planus. In addition, physiological vaginal discharge or semen may irritate inflamed skin. This is much more common when the discharge is altered in nature or amount by an infection such as candidiasis, trichomoniasis, or another sexually transmitted disease. Pruritus ani and vulvae are rare in regions where malnutrition is common, except as a manifestation of an orogenital syndrome due to vitamin B deficiency: the pruritus is then associated with dermatitis and angular cheilitis. Threadworm infection is common in children, while urinary tract infections are common in adults and children and are associated with urethral stinging and burning and will also aggravate a pre-existing complaint of vulval itch or pain. Sexual intercourse, in particular unaroused sexual intercourse, produces local mechanical trauma that rarely initiates dermatitis but invariably aggravates it. Anticipation that vaginal intercourse will be painful leads to further complications such as vaginismus that may continue for many years after the precipitating event has resolved.

Sufferers use a large number of agents to relieve their pruritus, some of which cause contact dermatitis. Pruritus vulvae may also be caused by sensitivity to the rubber of condoms or spermicidal jelly, or even to deodorants. Local anaesthetics and local steroids are much used; however, the latter encourage secondary infections, with fungus usually spreading on to the buttocks and down the thighs. Potent fluorinated steroids may lead to a rosacea-like pustular perioral facial dermatitis.

Management includes a thorough examination to exclude the above and to recognize skin diseases such as psoriasis, seborrhoeic dermatitis, lichen sclerosis, atopic dermatitis, and lichen planus. Vaginal swabs should routinely be taken to look for candidiasis and trichomonas infection, and other sexually transmitted diseases when indicated. Skin swabs and scrapings for the detection of secondary fungal, bacterial, and viral (herpes simplex) infection are useful. A full history of topical medicaments and other preparations should be elicited, and patch-testing performed if required. Detailed instructions regarding local hygiene should be given to patients, including advice on the need to avoid excessive ritual washing, particularly with soap. Counselling regarding sexual intercourse may be needed.

Pruritis vulvae usually responds well to potent topical steroids. The risk of atrophy is small if the itch is confined to the vulva. Secondary infection may be prevented by the concomitant use of a topical antiyeast preparation such as nystatin or antiseptic such as Vioform (clioquinol), which tend to be non-irritating. Maintenance treatment with a weak topical steroid may be required.

Recurrent vaginal candidiasis may be recalcitrant to therapy. It relates to oestrogen levels and is not seen before the menarche or after the menopause, except in women receiving hormone replacement therapy. Long-term oral antifungal agents may be required, particularly when topical preparations irritate the skin. Occasionally, menstruation and endogenous oestrogen production may need to be interrupted by depot medroxyprogesterone acetate (Provera).

Atrophic vaginitis should be sought and treated in postmenopausal women presenting with dyspareunia. Primary vulvodynia has been described with vulval vestibulitis, vulval pain syndrome, and pudenal neuralgia. These are complex and controversial entities for which there is no satisfactory treatment. Tricyclic or one of the newer antidepressants in full dosage may help either in altering sensory perceptions or treating an underlying depression. Compliance is often difficult.

Psoriasis

In temperate zones, psoriasis affects between 1.5 and 3 per cent of the Caucasian population. It is less common in sunny climates and among certain ethnic groups, such as Asians, native Americans, and Samoans. Around 30 per cent of patients have a positive family history and there is 75 per cent concordance among monozygotic twins, compared with 20 per cent concordance with dizygotic twins. It is a polygenic trait with two patterns of inheritance. Type 1, with a strong family history and early onset, shows linkage disequilibrium for human leucocyte antigens CW_{6}, B_{13}, and BW_{57}. It affects 30 per cent of patients and is possibly an IL-Ira gene defect. Type 2 occurs as a late-onset disorder and is linked with CW_{2}, B_{27}, and CW_{6}.

Known triggers for expression in genetically susceptible people include trauma, stress, infection, pregnancy, hypocalcaemia and dialysis, HIV infection, alcohol, and certain drugs such as lithium, b-blockers, antimalarials, NSAIDs, and steroid withdrawal. Paradoxically, sunlight may worsen psoriasis in some individuals.

Pathogenesis

The pathogenesis of psoriasis includes a tenfold increase in the speed of epidermal-cell proliferation. Since the cells pass upwards through the epidermis at a faster rate and do not seem to have time to produce a horny layer, they remain nucleated even when exfoliated. Numerous problems beset the measurement of the cell-cycle time in human epidermis. For example, there are the technical difficulties of counting and the exact recognition of different stages of the cell cycle and differentiation. Do all cells in the germinal layer have the potential to divide or is the potential greater in psoriasis? Is the actual cell cycle faster in psoriasis? The answer is probably that it is and that more cells enter the cycle per unit time in psoriasis. Moreover, cell-cycle inhibitory factors may be reduced and stimulatory factors increased. The kinetics of keratinocyte turnover are clearly essential to our understanding of psoriasis. Probably there is no single cause, but neutrophils, which are attracted in large numbers into the epidermis, may play a part. Streptococcal antigens crossreact with skin antigens, thereby stimulating an autoimmune response. The role of the lymphocyte has long received consideration and has been encouraged by the observations of exacerbations of psoriasis in patients with

AIDS, its control by ciclosporin, and the possible immunosuppressive effects of other effective therapies, such as corticosteroids or PUVA. Psoriasis in those with AIDS is most pronounced at intermediate levels of immunodeficiency, and is diminished or lost in terminal profound immunodeficiency states. At the biochemical level, almost every aspect of cell kinetics is a candidate, including the availability of cyclic AMP, increased cyclic GMP, fatty acid deficiency, eicosanoids, phosphorylating mechanisms, polyamines, putrescine, spermidine, and calcium-modulating enzymes such as calmodulin and vitamin D analogues. How cells stick together and how adhesion is modulated during migration and mitosis introduces many other concepts, ranging through the role of interleukins (IL-1, -2, -6, -8), interferons (IFN-g), transforming growth factors-a and -b, to the interaction with proteases, since in psoriasis the proteinase–antiproteinase balance seems to be disturbed. Psoriasis is not merely a disorder of the keratinocyte. Thus arthritis cannot be explained on such a basis. Within the dermis, some hypothesize that the fibroblast, the mast cell, or even the endothelial cell are prime targets for whatever it is that fires the psoriatic process. Recent greater understanding has added neuropeptides to the list of potential triggers, since they induce mast-cell degranulation and fire the interaction between the fibroblast and keratinocyte.

The lesion-free skin in people with psoriasis is not normal. Psoriasis is readily induced and various medications, such as chloroquine, practolol, and lithium, can produce a flare-up. Oddly, glycogen levels, so high in the psoriatic lesions, may be lower than in the normal surrounding skin. The dermis is not normal, and the earliest signs of any abnormality following injury are infiltration by mast cells and macrophages.

The microvasculature in psoriasis is characterized by tortuous and leaky capillaries, generous protein exudation, and poor clearance through immature lymphatics.

Koebner phenomenon

The 'Koebner phenomenon' is a term given to psoriasis developing in traumatized skin; this occurs at some stage in about half of all the patients with psoriasis, and is most common when the psoriasis is active. It is an all-or-nothing phenomenon on the skin surface. After the initial repair stimulus, the epidermis gradually thickens and there is accentuation of the papillary interdigitations and the rete ridges. An early heavy infiltrate by neutrophils forming microabscesses within the epidermis is preceded by increased numbers of mast cells and macrophages in the dermis. High turnover of the epidermal cells results in a less-compact and still partially nucleated scale known as parakeratosis. The reverse Koebner phenomenon, where trauma initiates clearing of psoriasis can also occur, and is probably more common than the true Koebner phenomenon.

Clinical appearance

Psoriasis can affect all age groups, but has a peak age of onset between the ages of 5 and 9 years in girls and 15 and 19 in boys. The commonest lesion is a sharply marginated plaque with silvery scales (Fig. 42). These mask the underlying redness from the tortuous convoluted capillaries that lie close to the surface of the skin. The edges of the lesion are usually the most active and there is commonly clearing in the centre (Fig. 43). Since itching is common, scratching may modify the appearance, increasing shedding of scale and causing bleeding.

Fig. 42 A psoriasis plaque showing the silvery scales, well-defined border, and predilection for the elbow.

Fig. 43 Psoriasis that is less stable than in Fig. 42. The lesions are erupting and more active at the periphery while healing in the centre.

Sites most commonly affected are the elbows, knees, and scalp, areas that normally have a higher rate of epidermal turnover. The face is less often affected. Spontaneous fluctuations are common and remissions occur in about one-third of cases per annum. As there are several well-recognized patterns, it is important to examine the patient thoroughly until a completely recognizable lesion of psoriasis can be detected. Many lesions and some patterns may be quite atypical, especially during the development of psoriasis (Fig. 44) or during its resolution.

Fig. 44 A still more unstable form of psoriasis, tending to be more exudative and exfoliative and not retaining its rapidly produced scale. While tending to be symmetrical, new ill-defined lesions are erupting. There are some linear lesions on the trunk suggestive of the Koebner phenomenon or a reaction to a skin injury, in this case probably from scratching.

Guttate psoriasis

This term is derived from the Latin *gutta*, meaning a drop. The skin looks as though it has been splashed by the psoriasis. It often follows a streptococcal sore throat or vaccination and is especially common in children. The lesions are scattered over the entire body and tend to be no more than a few millimetres in diameter. They may include the face and are often red slightly scaly spots. Guttate psoriatic lesions appear less well defined and less obviously covered by silvery scales than in classic types of psoriasis. In the absence of a family history the prognosis tends to be good.

Nummular discoid

This is probably the commonest form of psoriasis, in which coin-shaped lesions of various sizes (Fig. 45) are scattered over the body in a completely symmetrical distribution. Such lesions are usually well defined and chronic.

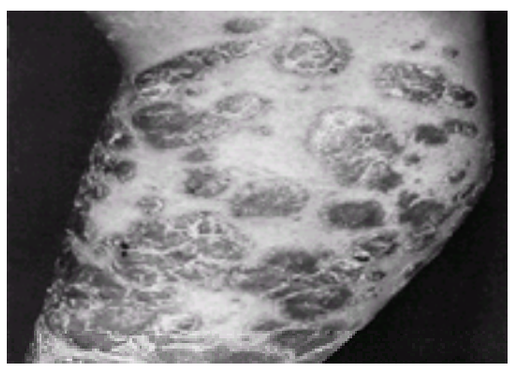

Fig. 45 Discoid lesions still well defined but becoming almost confluent.

Palmar and plantar psoriasis

This may be typical of lesions elsewhere (see Fig. 4), but the psoriasis is often modified due to the nature of the palmar and plantar skin. The scales tend to be more adherent and less silvery and are more likely to develop deep cracks because of the thickness of the epidermis at these sites (Fig. 46). Neutrophils tend to collect into larger abscesses trapped by the thicker surface layers of the stratum corneum, the sterile pustules so formed are often the most obvious feature. Although this pattern may be seen as part of a more generalized disease, in many cases it only affects the hands and feet. There is some evidence that it is a different disease, since the above-mentioned HLA associations are absent and there is no obvious increase in the rate of epidermal turnover. Where the psoriasis is an occasional and acute response to infection, it is known as pustular bacterid.

Fig. 46 Psoriasis of the palms may not have the typical scale. It is sometimes pustular or, as in this case, hyperkeratotic with a tendency to form deep cracks (see Fig. 4).

Psoriasis of the nails

Some 25 per cent of patients with psoriasis have nail disease. Among those with psoriatic arthritis this figure is in excess of 75 per cent. Pin-point pitting is usual but can be seen in other disorders affecting nail growth (see Fig. 66). Onycholysis with brown discoloration of the base of the uplift of the nail, known as brown onychodermal band, is probably even more characteristic. Salmon-pink circular discoloration in the nail-plate, likened to oil drops, are only seen in psoriasis. Sometimes the nail growth is distorted, thickened, and friable and difficult to distinguish from a fungus disorder affecting the nail (see Fig. 69).

Fig. 66 Pits of the nails are due to very localized accelerations in growth such that the nail keratin is less well knit. Such pits are very common in psoriasis.

Fig. 69 Severe growth changes of the nail are often a consequence of psoriasis or eczema of the fingertips. The latter may resolve while nail growth disturbance may

persist for many months.

Flexural psoriasis

When psoriasis affects the groins, natal cleft, or axillas, it is usually less scaly. The bright-red plaques are shiny and liable to cracking and maceration. They may be very well defined.

Erythroderma

This may present as a medical emergency due to fluid loss, septicaemia, or lowering of body temperature. The elderly may develop high-output cardiac failure. Oedema is a consequence of capillary leak, low albumin, and heart failure. When psoriasis affects the entire skin there is generalized redness, the well-defined margins are lost, and the scales are profusely exfoliated. The erythroderma may be indistinguishable from that found in eczema or lymphoma. Bacteraemia commonly ensues when the normal protective function of the skin is lost. The loss of water is difficult to estimate and prerenal failure can develop very rapidly. The vasodilatation and the obstruction to the sweat ducts by the proliferating epidermis results in impaired thermoregulation. Hyperthermia is very common in hot climates, while hypothermia can occur in cold climates. Internal organs such as the gut and liver may be impaired and loss of protein both from the skin and the gut is an important complication.

Generalized pustular psoriasis

In this condition, which is relatively rare, waves of bright erythema develop within a few hours together with a fever, arthropathy, and leucocytosis. Myriads of pustules (see Fig. 17) quickly develop and equally quickly disappear. This disorder may occur in the absence of a previous history of psoriasis and even occasionally as a viral exanthem. However, most commonly it is only a complication of psoriasis that has been treated by systemic or local steroids. It is an acute rebound phenomenon of steroid withdrawal.

Acute generalized exanthematous pustulosis occurs in patients with no previous or family history of psoriasis and has been blamed on mercurial drugs and antibiotics. Another rare cause of pustular psoriasis is hypoparathyroidism. Cutaneous drug eruptions may mimic this condition, and any suspected drug should be stopped. Generalized pustular psoriasis is potentially life threatening and therefore admission to hospital is required. Oral retinoids are helpful, as is methotrexate. Oral steroids are best avoided due to the potential for rebound flare when they are stopped.

Arthropathic psoriasis (see Section 18)

The incidence of polyarthritis in those with psoriasis is about 7 per cent in hospital series; 4 per cent of all patients with inflammatory polyarthritis have psoriasis. Up to one-third of patients with pustular psoriasis develop a polyarthritis. There is a long-standing debate concerning the association of psoriasis with inflammatory polyarthritis; it is still uncertain whether it is a chance association, possibly related to a genetic linkage disequilibrium. Since psoriasis is a common disorder, patients with a positive Rose–Waaler test can have coincidental rheumatoid arthritis.

Management

By far the most disabling aspect of psoriasis is its appearance, and patients' lives can be completely taken over by manoeuvres designed to avoid exposing the affected skin to the public eye. Management includes a sympathetic hearing and, when necessary, admission to an outpatient or inpatient unit where others with psoriasis are being treated.

Psoriasis can usually be controlled with therapy. While the skin can be made to return to normal, the inherited susceptibility is fixed, and so patients remain vulnerable to relapse. Some environmental triggers have been identified and should be avoided. These include infection (streptococcal, HIV), trauma to the skin (Koebner phenomenon), psychological stress, and drugs such as lithium, chloroquine, and b-blockers; however, many triggers are unknown and therefore unavoidable.

It is common to combine treatments, either different topicals together, or topical agents with oral preparations or ultraviolet therapy. The response to treatment is variable, and those that were previously effective may no longer be so, and vice versa.

The aims are to depress epidermal cell turnover, suppress skin inflammation, reverse angiogenesis, and remove hyperkeratosis without irreversibly damaging the skin or other organs. It is important that the treatment chosen is appropriate for the type and site of the psoriasis. In mild cases, emollients (for example, 10 per cent glycerine in sorbolene cream) or keratolytics (for example, salicylic acid 3–10 per cent in aqueous cream) may suffice, but disabling or disfiguring psoriasis may warrant the use of systemic antimetabolite or immunosuppressive drugs.

Local steroids

Topical corticosteroids are anti-inflammatory, immunosuppressive, and antiproliferative. Steroid creams and ointments are often used as the first-line treatment for all types of psoriasis because of their ease of use. The stronger halogenated steroids are the most effective. A response may be seen as early as 1 to 2 weeks. Complete resolution generally takes 4 to 6 weeks and occurs in around two-thirds of those who use it. Relapse on cessation of therapy is common and one-third of patients need to continue once- or twice-weekly application. Care is required on the face and flexures to prevent the formation of stria and telangiectasia. Other side-effects associated with continued use are skin atrophy, gradual extension of the psoriasis, and greater instability of the skin so that psoriasis erupts whenever the therapy is partially or completely withdrawn. Eventual widespread usage and systemic absorption complicates the increasing addiction of the skin for the stronger steroids. Some patients so treated show no remission until all such therapy is withdrawn; although in a few this can be done with no immediate worsening of the psoriasis, in most patients withdrawal leads to a rapid worsening of their skin condition.

Tar

Tar is antiproliferative, antipruritic, and anti-inflammatory. Tar has been known to be effective and safe for more than 50 years. Follow-up of patients treated with tar between 1917 and 1937, using the Danish Cancer Registry of all cell cancers from 1943 to 1990, found no overall risk of skin cancer; however, recent rulings by the European Commission have limited its availability in some countries. The smell, colour, and stain of crude coal tar (**CCT**) make it cosmetically unsatisfactory and it may be irritant on more vulnerable skin such as the face and flexures. Less irritating and more cosmetically acceptable varieties of tar include liquor picis carbonis (**LPC**), ichthammol, and oil of Cade; however, these are comparatively less effective than crude coal tar. CCT (2.5–6 per cent) or LPC (2–10 per cent) may be prescribed in a base of aqueous cream, sorbolene cream, or liquid paraffin. A keratolytic such as salicylic acid can be added to augment its efficacy. Many proprietary formulations are available in most countries.

The preparations are applied once or twice daily; but diluted by 50 per cent if they are irritant. Generally little response is seen for 1 to 2 weeks. Complete resolution takes between 6 and 8 weeks; however, subsequent remissions may be prolonged. Occasional patients are allergic to one or more of the constituents. The general principles of when to use a lotion, ointment, or paste are discussed below. Acute or inflamed psoriasis responds to ichthammol, which has a milder action than coal tar.

Dithranol

Dithranol is antiproliferative and mildly anti-inflammatory. It is more effective than coal tar in the treatment of psoriasis, especially if the plaques are large, well-defined, and few in number, but is more irritant, especially to the eyes and genitalia. However, an acceptable diluted concentration can generally be found, and the concentration can then usually be gradually increased. The dithranol is mixed in zinc oxide and salicylic acid paste (Lassar's paste) in a concentration of 0.1, 0.25, 0.5, 1, or 2 per cent. Weaker preparations are used in the more sensitive occlusive flexures. It is safe in pregnancy.

Non-lesional skin may be protected by Vaseline, and it is usual to apply dithranol paste accurately to the active parts of the skin lesion. Powdering fixes the paste and a gauze or nylon dressing protects the overlying clothing from dithranol's staining property. Patients tolerating this regimen are cleared in about 3.5 weeks on average. Staining is inevitable and short lived and is a sign of effectiveness. Irritation, feeling like a mild burn, is treated by stopping treatment for 1 or 2 days. Various proprietary brands are slightly easier to manage and include Vaseline-based preparations and creams or sticks. The 'minute regimen' is a system using 1 to 5 per cent dithranol in Vaseline and 2 per cent salicylic acid. It is only applied for about 80 min, then removed with an oil. Dithranol is suitable for those whose employment requires their skin to be free of ointments for most of the day.

Calcipotriol

Calcipotriol, a vitamin D analogue, is a safe and cosmetically acceptable, effective topical treatment, which inhibits proliferation and enhances epidermal differentiation. It is applied twice daily at a rate of no more than 100 g/week. However, calcipotriol is irritant for about 10 per cent of users, particularly when used on the face. Its combination with betamethasone valerate in popular.

Tazarotene

Tazarotene is a topical synthetic retinoid that also has some efficacy in the treatment of psoriasis. In general, it works best in combination with other topical agents, but is commonly irritating on the skin.

Phototherapy

Natural sunlight is helpful in about 75 per cent of patients and probably accounts for a decreased incidence of psoriasis in sunny climates. Suberythema doses of UVB are a useful substitute, and its effectiveness can be increased by prior bathing or an application of tar which sensitizes the skin to the UVB.

PUVA therapy, which is a combination of long-wave ultraviolet rays (UVA or black light) and 8-methoxypsoralen tablets (0.5 mg/kg) taken 2 h before exposure, produces effective clearance and a bronze skin in most patients. A 15- to 30-min exposure two or three times per week succeeds in clearing the psoriasis in 6 to 10 weeks. The treatment is stopped on clearance of the psoriasis. Patients who relapse quickly may be considered for maintenance therapy either once a week or once every 2 weeks. Recurrences are no less frequent than with other forms of therapy. Dryness, atrophy, and other expected changes of irradiation are a consequence. The risk of skin cancer is as yet difficult to estimate, but it is not insignificant. In male patients receiving more than 200 treatments, the incidence of squamous-cell carcinoma was 30 times greater than that found in the general population in Sweden. As the risk of melanoma in these patients may also be increased, further long-term studies are required. The risk is especially great in those who have arsenical keratoses or other evidence of a previous intake of arsenic. Concurrent treatment with methotrexate is also considered a risk factor.

Broad-band UVB therapy has been used for much longer and is similarly effective as PUVA. Increased carcinogenicity, although not proven, is suspected. Narrow-band UVB (311 mm) is thought to be less irritant and highly specific for psoriasis. By screening out the shorter wavelengths within the UVB spectrum it is assumed to be safer than broad-band treatment.

Climatotherapy is the combination of natural sunlight (or a filtered form of sunlight in the case of the Dead Sea area) with sea salts and/or other constituents—sulphur, black mud, and bromides—present in different geographical regions. It is effective, but no more so than other regimens that provide topical medicaments and mental relaxation. The Dead Sea may be exceptional in the number of its peculiar features, such as low humidity, increased atmospheric pressure, UVB filtration, and minerals in high concentration.

Systemic therapy

When psoriasis is widespread, severe, or causing disfigurement or disability, systemic therapy is indicated. This may take the form of drugs that have antiproliferative or anti-inflammatory, or immunomodulatory effects. The main agents used are acitretin, methotrexate, and ciclosporin. A host of other agents such as hydroxyurea, sulfasalazine, tioguanine, mycophenolate mofeate, tacrolimus, and azathioprine have some efficacy in psoriasis. Human anti-IL-8 antibody may prove to be a new approach to the management of psoriasis.

Acitretin affects cell proliferation and differentiation mechanisms and is also an anti-inflammatory agent. It may be used in psoriasis as monotherapy in the control of palmoplantar and other hyperkeratotic forms of the disease, as well as pustular erythrodermic and atypical presentations of psoriasis, or in combination with phototherapy to augment its efficacy. The usual dose is 0.5 to 1 mg/kg daily; however, many patients cannot tolerate this dose but lower doses may be similarly effective. The major adverse effect of acitretin is teratogenicity, hence contraception is mandatory for women of childbearing age for the duration of therapy and for at least 2 years after the completion of a course of treatment. This makes acitretin undesirable for use among premenopausal women. Other adverse effects of acitretin include all those of hypervitaminosis A; cheilitis, hair fall, photosensitivity, elevation of liver enzymes, and increase of serum lipids.

Methotrexate slows epidermal-cell proliferation and is an immunomodulator. It is the most commonly prescribed antipsoriasis drug. The usual dose is 0.2 to 0.4 mg/kg orally once weekly. Toxicity is more common in the elderly and in patients with reduced renal clearance, therefore lower doses may be required. Patients should abstain from alcohol while being treated with methotrexate. Adverse effects include nausea, pancytopenia, and elevation of liver enzymes. Nausea may be controlled by the concomitant administration of folic acid. Long-term use of low-dose methotrexate may induce liver fibrosis, which seems to be more common in patients treated for psoriasis with methotrexate than in those being so treated for rheumatoid arthritis. Monitoring of patients includes regular full blood and liver function tests. Liver biopsy may be considered in patients with abnormal liver function tests after a cumulative dose of methotrexate between 2 and 4 g to exclude the liver cirrhosis that occasionally complicates the fibrosis.

Ciclosporin A inhibits selectively activated T-helper cells and reduces the production of cytokines. It is very effective in controlling psoriasis at doses between 2 and 5 mg/kg per day. However, ciclosporin A does not produce long remissions and recurrence follows discontinuation. It is not recommended for continuous use beyond 12 weeks, twice per year, because of long-term adverse effects such as hypertension, deterioration of renal function, hypertrichosis, gingival hyperplasia, and the development of neoplasia (specifically, skin squamous-cell carcinoma and lymphoma). It is reserved for the most difficult cases when other treatments have failed or been deemed unsuitable.

On the basis of preliminary reports, one can predict that monoclonal antibody therapy against adhesion molecules and against interleukin-2 receptors will have increasing advocates.

Pityriasis rosea

Pityriasis rosea is a relatively common, self-limiting, inflammatory skin condition that characteristically affects young adults. The cause is thought to be a virus (possibly human herpesvirus-7). The eruption begins with a herald patch that may be mistaken for ringworm. This is followed about 2 weeks later by the development of multiple, scaly, salmon-coloured macules, each about 1 to 2 cm in size and oval in shape. The macules are confined to the trunk and upper limbs and are arranged along the skin creases to create an appearance reminiscent of a Christmas tree. Itch is variable. The rash disappears spontaneously, usually within 6 to 8 weeks.

The aim of treatment is the palliation of any associated itch while awaiting spontaneous resolution. In the absence of itch, reassurance may suffice. First-line therapy to relieve itch consists of topical antipruritic agents such as calamine lotion, 1 per cent menthol in aqueous cream, or a moderately potent topical corticosteroid cream. UVB phototherapy can be used to relieve itch and may hasten resolution of the rash.

Lichen planus and lichenoid eruptions

Lichen planus is an idiopathic inflammatory condition. It may affect the skin, hair, nails, and oral and genital mucosa. Alopecia occurs if the scalp is involved, and nail damage and destruction are seen when the nail matrix is affected. Occasionally lichen planus can be triggered by drug ingestion or hepatitis C infection.

Lichen planus and lichenoid eruptions are characterized by violaceous papules that are usually flat-topped and shiny and heal leaving pigmentation. Histology shows

damage to the basal layer of the epidermis and an intense infiltration of lymphocytes and a few histiocytes situated immediately below the epidermis (Fig. 47). A T-cell-mediated CD4+ attack on the epidermis may be triggered by viral, drug, or neoplastic processes. Lichen planus thus presents a model for the elimination of damaged and normal keratinocytes. Cytokines, interferon-g, and tumour necrosis factor-b play a critical role. Lymphocyte and keratinocyte molecules subserving adhesion are activated and, in the mouse, lichen planus can be blocked by monoclonal antibodies. The disturbances in epidermal growth that result from this damage range from extreme atrophy with ulceration and almost no epidermal-cell turnover to considerable hypertrophy and hyperkeratinization, giving rise to thick nodules meriting the name 'hypertrophic lichen planus'. Most patients are between the ages of 30 and 60 years and it is extremely rare in children. More-erosive forms are seen in the elderly, and pigmented skin tends to develop more hypertrophic varieties.

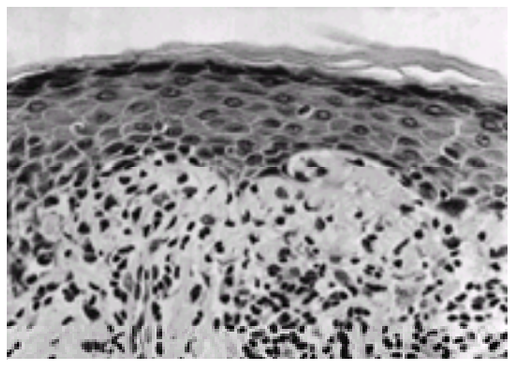

Fig. 47 Lichen planus. There is thickening of the granular layer and necrosis of the basal-cell layer. Pale cells at the lower edge of the epidermis are destroyed epidermal cells. The predominant white cell is a lymphocyte. The infiltrate is often confined to the upper dermis.

HLA-A3 and -A5 occur more often in patients with lichen planus than in controls. The graft-versus-host skin reactions following bone marrow transplantation often present with an identical pattern of pathology to that of lichen planus. There is some clinical pathological overlap with the appearance seen in lupus erythematosus. Although immunofluorescence studies show heavy deposits of fibrin and immunoglobulin, these could be entirely non-specific. There is some evidence of defective carbohydrate metabolism and abnormal glucose tolerance, but the basis of this association still has to be explained. A lichen planus-like drug eruption may also be produced by a variety of medications, particularly antihypertensive and antimalarial agents, gold and organic arsenicals, antituberculous therapy, chlorpromazine, as well as from contact with colour developers.

Clinical features

The classical lesion is a shiny, flat-topped papule (Fig. 48) described as polygonal and violaceous. Small white dots or lines in such papules (termed 'Wickham striae') are due to a mixture of oedema, white-cell infiltrate, and vasculature disturbances. The papules may become confluent and heal in the centre, giving rise to annular (Fig. 49) lesions or plaques with varying degrees of epidermal response. This may result in either atrophic skin or extreme hypertrophy. A lacy-white appearance (Fig. 50) is common in lesions of the mouth or the glans penis. Hair-follicle involvement may give rise to keratosis pilaris and actual destruction of the hair follicle. Thus, lichen planus is one cause of scarring alopecia. Healing of the lesion is often followed by pigmentation due to melanin in the dermis. Warty hyperkeratotic lesions may be very persistent, as may ulceration, particularly of the peripheries or of the oral mucosa. The initial lesions are commonly found on the front of the wrists, in the lumbar region, or around the ankles. The palms and soles may be involved, in which case the appearance may even suggest a vesicular eruption. Involvement of the mucosa and tongue, which occurs in between 30 and 70 per cent of cases, may extend to the genitalia and perianal area, and has even been described in the rectum, stomach, and larynx. Severe itching is common. Ridging of the nails is essentially due to cessation of nail growth, producing longitudinal linear depressions.

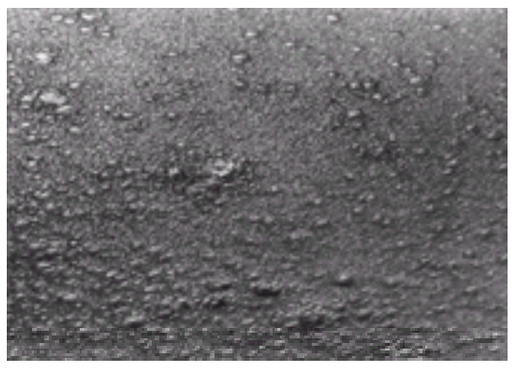

Fig. 48 A black skin affected by the shiny, flat-topped, often polygonal, papules of lichen planus.

Fig. 49 Lichen planus may heal in the centre, leaving atrophy and pigmentation. The edge is slightly raised; which gives rise to the annular form of the condition.

Fig. 50 Lichen planus of the mouth is one cause of mucosal whiteness. Unlike candidiasis the lesions cannot be removed by scraping. They are characteristically

'lacy' in appearance.

Prognosis

The mean age of onset is during the fifth decade. It may be very explosive or insidious in onset. Most cases clear slowly: 66 per cent of cases take up to a year to clear; 85 per cent clear in 18 months. Mucous membrane lesions or extremely hypertrophic lesions on the legs often persist for years, and there is a risk of squamous epitheliomatous changes, particularly in ulcerated mucosal lesions.

Treatment

Treatment is mainly aimed at relieving the itching, and local steroid creams are perhaps the most effective. Occasionally, a course of prednisolone or ACTH is justified for the treatment of a very severe widespread lichen planus. Probably it does not influence the course of the disease but merely its intensity. As with all itching conditions of the skin, cooling evaporating lotions such as calamine lotion may be helpful. Persistent ulcerated lesions can be excised or grafted. Lesions in the mouth can be treated with local steroid creams, those containing Orabase (a carmellose and gelatin paste) being particularly effective. An asthma-type spray may be a more effective way of administering steroids to the oral mucosa, but must be used without inhalation.

Acne vulgaris

Acne is a common, chronic inflammatory disease of pilosebaceous ducts, that affects a high proportion of the population. An onset around puberty is usual and it tends to resolve within 10 years if untreated. Its onset coincides with major changes in the life of an adolescent and may have a negative impact on the psyche of the sufferer and cause profound psychological effects. Physicians should not underestimate the effects acne may have on an individual's body image. Even minor acne may cause depression and negative feelings in particular individuals. It is strongly familial and monozygotic twins show high concordance. It is seen less in some races (Blacks, Japanese, Inuits). The disease may vary at presentation from one to two evanescent lesions occurring at a time, to a severe inflammatory disorder that can result in disfiguring scarring.

Commonly involved areas are those with the highest concentration of sebaceous glands, that is the face, neck, chest, shoulders, and upper back. The primary lesion of acne is the comedo. This non-inflamed lesion may be open (blackhead) or closed (whitehead), and may be accompanied by inflammatory lesions (papules, pustules, and cysts).

Aetiology

The three main processes contributing to acne are: increased sebum secretion; pilosebaceous duct obstruction; and inflammation.

Under the influence of local sex-hormone metabolism, androgens stimulate an increase in the size of the sebaceous glands and, hence, more sebum is produced (see below). These large glands themselves produce more active androgen metabolites through the activity of type 1, 5a-reductase; one effect of these metabolites is to further enlarge the sebaceous glands. Sebum acts in partnership with bacteria to produce keratinization and hence blockage of the pilosebaceous duct and comedo formation.

The principal organism responsible is *Propionibacterium acnes*, which increases in number during flare-ups and is important in the change from non-inflammatory to inflammatory acne. This bacterium produces many inflammatory substances, such as lipases, proteases, hyaluronidases, and chemotactic factors that play a role in producing lesions. Therapy that lowers the *Propionibacterium acnes* count plays a pivotal role in management, but resistance of the bacterium to some antibiotics, especially erythromycin, is an emerging problem in acne therapy.

Acne may also be drug-induced, particularly secondary to anabolic and corticosteroids, iodides, lithium, phenytoin, streptomycin, and isoniazid. Sebum consists of triglycerides, wax esters, squalene, and sterol esters. The fatty acids in sebum are inflammatory and are formed by the lysolytic enzymes of bacteria, even in healthy skin, from unsaturated 14- to 16- or 18-carbon components of the triglycerides. It is possible that acne in people living in the tropics is due to a secondary response in the rate of turnover of the follicular lining, perhaps induced by occlusion under a belt or braces in such hot environments. The acne of Cushing's disease may also be due to an increased rate of such turnover. Chlorinated hydrocarbons also cause acne. Chloracne is an important symptom of poisoning and was present in 168 cases of poisoning in the Seveso industrial accident. The exact way in which the inflammation is produced is uncertain; as the follicle contains fatty acids and bacterial proteases which activate the classical alternative pathway of complement and attract neutrophils, this may be one mechanism.

Sebaceous gland activity is regulated by hormones and, in particular, by androgens from the testes and adrenals, which stimulate, and oestrogens, which seem to suppress activity. In the adult male the glands are normally maximally stimulated, leading to more severe in boys than in girls. The skin itself is a major site for androgenic conversion similar to that observed in the prostate gland and in the male genitalia. Dihydrotestosterone, rather than testosterone, may be the end-organ effector and is formed within the target cells where it stimulates lipogenesis as well as mitosis. Eunuchs do not develop acne. Oestrogens reduce the size of sebaceous glands and sebum production is diminished.

Clinical features

Closed comedones (whiteheads) are the first stage of acne, seen as tiny white nodules below the surface especially when the skin is stretched (see Fig. 7). These may rupture giving rise to irritation of the dermis (that is, inflamed papules) or form the open comedo (blackhead) by pushing open the mouth of the follicle (Fig. 51). Because the black material is melanin, blackheads are blacker in dark skins and white in the albino; melanin is transferred to the keratinocytes before these cells are shed into the sebaceous follicle. Acne cysts (Fig. 52) occur as the result of ballooning of the distended follicle, often leading to destruction of the walls of the cyst and hair and sebaceous apparatus. Adjacent cysts often form fistulas and sinuses, which rupture, displacing epithelium in the dermis, and forming irregular channels or foreign-body reactions.

Fig. 51 greasy skin is a common accompaniment of acne, as is comedo formation, or 'blackheads', as seen on the forehead of this young man. Whether either are wholly responsible for the consequent inflammation and scarring also seen in this photograph is debatable.

Fig. 52 Large cystic lesions are the most disfiguring aspect of acne vulgaris.

Atrophic or hypertrophic scars of all types may be seen; frequently, excoriations, picking, and squeezing contribute to the irregularities of pigment and to the epidermal or dermal thickening. Rarely, young males develop suppurative and highly inflamed lesions in the skin over the chest with pain, fever, and accompanying polyarthralgia, probably mediated by immune mechanisms and the activation of complement.

Acne usually presents before the menarche and without treatment usually resolves within 8 to 10 years. A more persistent form—particularly affecting the chin—may last until middle age, especially in women who have premenstrual exacerbations. Some women with adult acne also have the polycystic ovary syndrome.

Cosmetics

In many parts of the world cosmetics contribute to acne, the lesions of which may be confined to the site of application. Vaseline-type preparations or medicated oily shampoos used by young men and women with long hair, are a well-known cause.

Management

The withdrawal of aggravating factors such as cosmetics and drugs is paramount where they appear to be involved in the aetiology of acne. Most patients with acne, however, have no such triggers. Trauma, such as picking and vigorously squeezing acne lesions, can aggravate the condition; in some cases of acne excoriée the effect of trauma dominates the clinical picture. Large superficial pustules can be evacuated by gentle pressure without deleterious effect, and the removal of loose comedones with a comedo extractor is commonplace in the practice of dermatology.

Local preparations

All local preparations produce some erythema and occasional pustulation before the acne comes under control. Sulphur is a time-honoured agent, producing local irritation and causing peeling. It is helpful for the treatment of pustules, but may not be so good for comedones which precede pustulation, often by several months. Comedones are reduced by topical retinoids, azelaic acid (20 per cent cream), and by 10 per cent salicylic acid in ethanol. Topical retinoids help the follicle-lining cells to slough off without plugging the follicle. They can be applied in cream, lotion, or gel formulations, and are indicated for comedones rather than for pustules or cysts. Side-effects of the topical retinoids include irritation, redness, and peeling.

Long-acting oxidizing antiseptics such as benzoyl peroxide reduce sebum excretion, reduce comedo production, and inhibit *P. acnes in vitro.* They are the topical treatment of choice, are a mild irritant, and may produce mild peeling after several days' application. It is best to start sparingly with 2.5 per cent preparation, and later to increase the amount applied or the concentration to 5 or even 10 per cent.

Topical antibiotics such as clindomycin (1 per cent), erythromycin (2 per cent), and tetracycline can be used to reduce the population of *P. acnes* in the pilosebaceous duct. In view of emerging bacterial resistance and the possibility of transfer of resistant genes between bacteria via plasmids or transposons, prolonged therapy, and in particular prolonged monotherapy, with these agents is not recommended. When used, they should be combined with topical benzoyl peroxide to inhibit bacterial resistance.

Oral therapy

Antibiotics used include tetracycline, minocycline, doxycycline, erythromycin, and trimethoprim (either trimethoprim alone or as co-trim-oxazole). The treatment is prophylactic and the onset of action is delayed for 6 to 8 weeks. Patients can expect a 60 per cent improvement after 3 months and an 80 per cent improvement after 6 months. Relapse is common on cessation of the treatment, however maintenance topical therapy may suffice.

Gastric upset, diarrhoea, and vulvovaginal candidiasis are the most common problems encountered with antibiotics. Photosensitivity to doxycycline and drug-induced hepatitis from minocycline are both uncommon side-effects. Morbilliform drug eruptions due to co-trimoxazole are relatively frequent. Tetracycline may produce discoloration of the teeth in the fetus and in children, hence oral erythromycin is the treatment of choice during pregnancy. Rarely, Gram-negative folliculitis occurs, particularly around the nose or in persistent cysts. This results in sudden worsening with considerable inflammation and may warrant a course of ampicillin.

The oral contraceptive pill is frequently prescribed for women with acne, as the oestrogen component suppresses sebaceous gland activity and decreases the formation of ovarian and adrenal androgens. However, the progestogen component can aggravate acne and therefore contraceptive agents with a low-progestogen dose are preferable. The newer progestogens, such as norgestimate, gestodene, and desogestrel, have less androgenic effects and are therefore the agents sought in a contraceptive pill where an anti-acne effect is desired. Oral contraceptives containing a combination of cyproterone acetate (see below) and an oestrogen (for example, 2 mg cyproterone acetate and 35 µg ethinyl oestradiol), are also an effective treatment of acne in women. While they are more effective than the low-progestogen oral contraceptives, they also have more adverse effects. The onset of action is slow, however they may be suitable for long-term therapy. They are usually more effective when used in combination with other therapies rather than as an isolated treatment. Cyproterone acetate and spironolactone can be used for their antiandrogenic properties in the control of severe acne in women not planning to become pregnant. They are most commonly used when there are associated symptoms of virilization, for example in the polycystic ovarian syndrome.

Where the acne is resistant to treatment or persistently relapses, and in cases of severe cystic acne and acne where scarring is likely, oral isotretinoin (0.5–1 mg/kg) is the treatment of choice. This vitamin-A derivative acts by correcting the keratinization defect in acne, decreasing sebaceous gland activity, and reducing the population of *P. acnes*. It is also anti-inflammatory. A cumulative dose of 120 to 150 mg/kg over a period of 4 to 6 months is usually required and one course is usually sufficient for most patients, although approximately 10 per cent of cases need a further course. Isotretinoin is teratogenic, so strict avoidance of pregnancy is of paramount importance whilst taking the drug and for one full reproductive cycle after discontinuation. Prescription is usually limited to dermatologists. Adverse effects of some degree are universal, including dry skin, eyes, and lips, as well as epistaxis, facial erythema, joint stiffness, myalgia, headaches, in addition to an initial flare of the acne on commencement of the treatment. Prednisolone can be used to ameliorate this initial flare. Other side-effects such as depression and paronychia are rare. The severity of the side-effects are dose-dependent and are generally ameliorated by a reduction in dose. Most adverse effects settle within 2 weeks of discontinuing the drug, and many patients cope better with the annoying adverse effects of cheilitis, xeroderma, and photosensitivity as their course progresses.

Diet

Studies of the effects of starvation in the obese or the malnourished show little evidence of the effect of diet on acne vulgaris, even in pellagra where some plugging of the follicles around the nose is an early sign. Chocolate worsens acne in some individuals, but this has not been shown in trials of larger populations. Nutrition may influence the age of onset of puberty and hence overeating may result in earlier acne.

Acne surgery

Comedo extraction is the expression of a follicle's contents by the application of pressure on the surface, often with a special device called a comedo extractor. The benefits of active attack on the lesions, thereby counteracting stasis and build-up of the contents, has to be weighed against the fact that suppression is always incomplete and a tendency to rupture into the dermis may promote inflammation. Cryotherapy destroys the lining of large cysts. Deep sinuses may require externalization. Solitary inflamed lesions benefit from intralesional inoculation of steroids. Persistent acne cysts sometimes resolve with the injection of small amounts of intralesional triamcinolone.

Pigmentation

The principal pigments in the skin include melanin (black), phaeomelanin (reddish-yellow), haemoglobin (red) and its by-products bilirubin (red) and biliverdin (green), as well as haemosiderin which produce colours of yellow, green, red, and brown. Longer wavelengths such as red penetrate deeper and are absorbed by melanin. Since blue does not penetrate so deeply it is not absorbed and is reflected back, which is why dermal pigment appears blue—hence blue naevus. Although racial causes of pigmentation are common, physical causes are also important since white skin may visibly tan in the sun. Tanning and burning are independent events, with different time courses (albeit overlapping), due to the interaction of ultraviolet radiation with distinct chromophores in the skin. Tanning consists of two phases: (1) immediate pigmentary darkening and (2) delayed pigmentation. Immediate darkening occurs within seconds to minutes of exposure, and is due to oxidation of melanin granules stored within keratinocytes in the skin. Delayed darkening is due to the increased melanin production within melanosomes, increased transfer of melanosomes from melanocytes to keratinocytes, and acanthosis of the epidermis. It begins 8 h after exposure, reaches a maximum at 24 h, and in the absence of re-exposure resolves over days to weeks.

Some pigmented lesions are naevi (Table 11, Table 12, and Table 13, Plate 1, Plate 2, Plate 3, Plate 4, Plate 5, Plate 6 and Plate 7, Fig. 53, Fig. 54, Fig. 55, Fig. 56, Fig. 57 and Fig. 58, and see Fig. 106), others result from pigment 'incontinence' which increases the amount of pigment in the dermal macrophages and is commonly postinflammatory.

Fig. 54 Becker's naevus is due to melanin in the dermis and is often segmental and usually hairy. It may only become overt after childhood.

Fig. 55 Crossbow pattern of chloasma in encephalitis, an association that has been described but may be incidental.

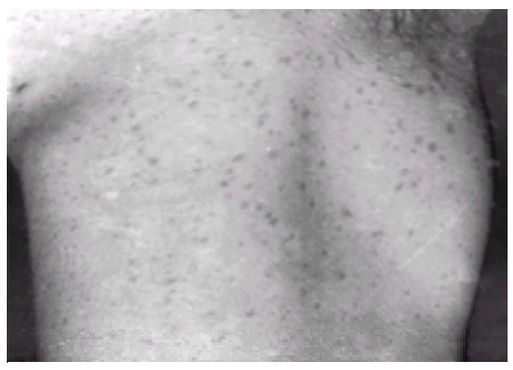

Fig. 56 In neurofibromatosis freckles extend into the axilla. This is a diagnostic feature in incomplete penetrance, important in the genetic counselling of white-skinned, but less reliable in black-skinned, patients.

Fig. 57 Syndrome of progressive darkening of numerous lentigos associated with cardiovascular and neurological abnormalities.

Fig. 58 Pityriasis versicolor due to the organism *Malassezia furfur* causes redness and slight brownish coloration of very pale, white skin, and depigmentation in dark or sallow skin. It favours the upper trunk.

Fig. 106 Erythema gyratum repens in a patient with adenocarcinoma of the colon.

Pigmentation as a feature of systemic illness is most significantly due to endocrine dysfunction affecting the melanocyte-stimulating hormone. In countries where malnutrition and infections are common, protein and vitamin deficiency, as well as cachexia from a variety of causes, account for disturbances in skin colour.

'Tinea' or pityriasis versicolor is due to a superficial fungus known as *Malassezia furfur.* It usually affects the upper trunk and may spread on to the neck or arms. The lesions are slightly scaly, off-white, pink, and brown. 'Pityriasis' is the term for a bran-like powdery scale and 'versicolor' implies the variation in the colouring (Fig. 58).

In leprosy, hypomelanosis is a feature of tuberculoid and borderline tuberculoid types (See Fig. 59 and Chapter 7.11.24). Light touch and later pinprick sensation are impaired. There is often a lack of sweating and there may be loss of hair. An adjacent enlarged peripheral nerve may be palpable, which may be mistaken for an enlarged lymph node.

Fig. 59 Hypopigmentation, especially with a hyperpigmented border, should always be tested for loss of sensation. This lesion is typical of tuberculoid leprosy.

Pigmentation of the buccal mucosa, tongue, or fingernails is significant only in white skin since it is a normal finding in dark races.

Depigmentation

'Leucoderma' is a term used for any whiteness of skin, and ranges from a partial hypopigmentation to the complete depigmentation characteristic of vitiligo. Microbial diseases such as pityriasis versicolor (Fig. 58), leprosy (Chapter 7.11.24) (Fig. 59), and syphilis are important infectious causes of hypopigmentation. Pinta should be suspected in people from central or southern America showing a succession of erythematous hyperpigmented lesions progressing to warty or atrophic plaques of depigmented skin. The late stage resembles vitiligo. Naevus anaemicus is a hypovascularity of the skin observed in white skins (Fig. 60); it is not a disorder of melaninization.

Fig. 60 Not all whiteness is due to loss of pigment. Decreased vasculature has been present since birth in this case of naevus anaemicus.

Vitiligo

This common, autoimmune skin disease results in the destruction of melanocytes and the total depigmentation of affected skin. However, an association with other autoimmune diseases and a family history of such is found in one-third of cases. The cause is unknown, but the melanocyte seems to be damaged by some, as yet,

unidentified antibody or toxin. Although vitiligo is most likely due to an immunological attack on the melanocyte, there are theories based on oxygen free-radical metabolism. In many parts of the world where skins are deeply pigmented, vitiligo is a principal cause of attendances at a dermatology department. Vitiligo affects up to 1 per cent of the United Kingdom population but 8.8 per cent in India. It presents during the first decade of life in 25 per cent of those affected. Except in those people unable to protect themselves from bright sunlight the disability is purely cosmetic, but it causes more concern and social handicap than almost any other common disease.

Occupational vitiligo is a well-recognized effect of exposure to certain chemicals used in the rubber industry—monobenzyl ether of hydroquinone, *p*-tertiary butylcatechol, and *p*-tertiary butylphenol. The hands are usually the first part of the body to be involved along with the genitalia, presumably through contact with the chemical excreted in the urine.

Clinical features

The initial depigmentation often occurs at sites of trauma, particularly the knuckles of the hands, and sometimes forms a white halo around a naevus (Fig. 61 and Fig. 62). The face and neck are usually affected early. In white-skinned people the first complaint often occurs during the summer when the unaffected skin is at its darkest from sun exposure. There is usually marked symmetry; the axillary folds and genitalia are commonly affected; the eye is not involved. Transient hypopigmentation is sometimes observed in evolving lesions; however, depigmentation of the lesion is ultimately total (Fig. 63) and should cause no confusion with the hypopigmentation of diseases such as leprosy. Such areas are never anaesthetic as in leprosy. Pigment may accumulate and be well defined at the borders of the lesion, giving a hyperpigmented edge. Melanocytes of hair follicles are usually unaffected and repigmentation, when it occurs, is often from such sites (Fig. 63(b)).

Fig. 61 Vitiligo is complete depigmentation and not merely hypopigmentation; it often begins at sites of minor trauma such as the knuckles. As with all essentially endogenous disorders it is symmetrical.

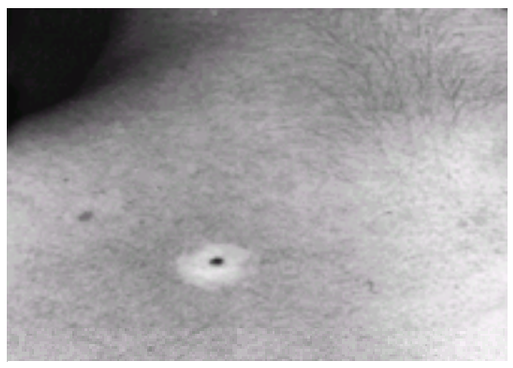

Fig. 62 Vitiligo often begins around a pigmented naevus—a halo naevus.

Fig. 63 (a) The pigment loss in this once dark-skinned woman is almost complete. Satisfactory cosmetic management would be depigmentation of the few residual areas of normal skin. (b) Although repigmentation of the skin in vitiligo is usually from the follicles, it is slow, unpredictable, and incomplete.

The clinician should be aware of associations with diabetes mellitus, pernicious anaemia, Addison's disease, myxoedema, and thyrotoxicosis. Less than one-third of patients show spontaneous repigmentation. In most cases the loss of pigment gradually extends. Depigmentation of the vulva, penis, and neck is sometimes persistent and of a localized variety, but need not necessarily progress to generalized vitiligo. It should be distinguished from the more atrophic lichen sclerosis which may also cause whitening of the skin at those sites.

Management

Patients are usually much distressed by the cosmetic disability. It is helpful to explain that there is a 30 per cent chance of spontaneous cure. Offering advice on camouflage make-up (many dermatology units have advisors specially skilled in this art) gives the patients an opportunity to help themselves, especially for important social occasions. But such camouflage is tedious and difficult to apply effectively for everyday use. There is no special advantage in the purchase of more expensive cosmetics since the basic constituents are cheap. The best effect is achieved from powder and grease mixtures with a powder finish patted gently into the skin after application. Dihydroxyacetone is the basis of many artificial (fake) suntan lotions, but again these are difficult to apply satisfactorily without overpigmenting the adjacent unaffected skin.

Patients should be told to avoid occupations which injure the skin (such as playing with animals that scratch) to prevent vitiligo from the Koebner phenomenon. The cosmetic effect of removing the remaining pigment is sometimes preferred by those whose skin is almost completely depigmented. The formulation used by the Sheffield Royal Infirmary is prepared as follows: hydroquinone 30 g; hydrocortisone BP 6 g; retinoic acid 600 mg; butylated hydroxytoluene 300 mg; and methylated spirit and polyethylene glycol in equal parts to 600 ml. Monobenzyl ether of hydroquinone may also be used to induce permanent hypopigmentation.

Psoralens and sunlight are one of the most ancient remedies used in medicine, UVA (black light) and selective UVB are recently developed extensions of the older remedy. Psoralens, methoxy- or tri-psoralen, are taken by mouth 2 h before exposure to light, or they may be applied topically 30 min before exposure. The simplest

regimen is a combination of meladinin paint and sunlight. It is necessary to test reactivity with short-time exposure and always to expose the skin at the same time of the day to avoid burning the skin by an unexpectedly high intensity of UVA. The chances of remission are in the order of 50 to 60 per cent, but success may take 2 to 3 years to accomplish. As might be expected from an autoimmune disorder, local steroid preparations are sometimes helpful, particularly in early or evolving disease. Although they have been advocated in combination with psoralens, therapeutic triumphs are difficult to assess, and the requirement to use these agents for years rather than days makes side-effects very likely. Local steroids are sometimes used to stabilize a rapidly progressive early stage of the disease. Cosmetically disabling local patches can be treated with light dermabrasion and the application of autologous split-skin grafts from pigmented skin. In Asia many clinics have treated thousands of patients in this way and there are several traditional herbal remedies prescribed for widespread lesions.

Other forms of pigment loss

Albinism

This is a group of at least 15 genetically distinct syndromes, determined not by absence of melanocytes but by their inability to synthesize melanin (see Chapter 23.2). Since melanin is important not only in the skin but also at such sites as the cochlea and retina, and also because the body's capacity to transfer organelles other than melanin is sometimes impaired, there are a number of associated defects affecting vision, hearing, and the delivery of lysosomes. In some societies where inbreeding is usual, albinism is common (San Blas Islands, Tanzania, southern Nigeria). Albinos in many countries are outcast, poor, underfed, and often die from skin cancer.

Albinism can occur in two forms: partial and complete albinism. Partial albinism occurs in piebaldism and albinoidism, a group of conditions where pale skin is seen without associated eye changes. Complete albinism is subdivided into ocular and oculocutaneous albinism. Skin changes are not seen in ocular albinism, but deafness is common in the autosomal dominant and autosomal recessive forms as well as one of the X-linked recessive forms. Oculocutaneous albinism (pale skin together with eye changes) is subdivided into tyrosinase-positive and tyrosinase-negative forms, according to whether incubation of hair bulbs with tyrosine induces pigmentation. In practice, this test is rarely performed. The tyrosinase-negative forms of albinism are all inherited as an autosomal recessive trait, while the tyrosinase-positive forms may be dominant or recessive.

Phenylketonuria

This cause of whiteness should not be forgotten. It results in elevated levels of phenylalanine which compete for tyrosinase.

Halo naevi

These are characterized by a loss of pigment around benign (see Fig. 62), or very rarely, malignant melanocytic naevi. It is a common first sign of vitiligo (see above). The central naevus should be assessed on its merits and need only be removed if there is a progressive enlargement, bleeding, and irregularities in the pigment, shape, or size of the naevus.

Tuberous sclerosis

The oval macules, which look like a thumbprint and are tapered at one end, sometimes known as leaf-like (Fig. 65), are present in about 90 per cent of affected babies. They are easier to see using Wood's light (UVA; see above).

Idiopathic guttate hypomelanosis

This is characterized by small, depigmented, sharply defined, often polygonal macules in light-exposed areas. It is probably caused by sun damage, and is almost universal amongst ageing Whites living in sunny climates.

Postinflammatory depigmentation

This is probably the commonest cause of leucoderma. Pigmented skin retains less pigment when there is accelerated epidermal turnover as in wound repair, eczema, or psoriasis, but the lesions are not as white as in vitiligo. One particular form commonly seen on the face but here illustrated in the cubital fossae is known as 'pityriasis alba' (Fig. 64), which is, in fact, a variant of a dry eczema causing mild hypopigmentation. It may be sharply circumscribed and have a halo of surrounding inflammation. Pityriasis alba is usually slightly scaly and there is follicular prominence due to hyperkeratosis. The cheeks and upper arms are most commonly affected and atopic children are the most frequent sufferers.

Discoid lupus erythematosus (**DLE**) is a common cause of depigmentation in some parts of the world. It is preceded by the itching, deep-violet erythema of light-exposed skin. Hair loss from the scalp is common, usually of the scarring type.

Diseases of nails, hair, and sweat glands

Nails and hair adorn our bodies. In addition, nails aid in picking up small objects. The handicap of disease of hair or nails is greatest in those who are most conscious of their 'looking good'. Beauty is not only in the eyes of the beholder but it is also an image in the mind. Hence, most consultations concerning hair are about too much or too little in comparison to the norm for a particular population. Excessive sweating and body odour is also a cause of great distress.

Nails

Nails grow continuously throughout life. The germinative epithelium is in the nail matrix, which is protected from the environment by a waterproof seal created by the cuticle. The structural integrity of the nail unit requires an intact cuticle and solid adhesion between the nail-plate and the nail-bed. Normal fingernails grow at the rate of approximately 1 cm in 3 months, with toenails taking anything from 9 to 24 months to grow as much. The nails grow more rapidly in psoriasis and more slowly in cold, ischaemia, or severe systemic illness.

There are only a limited number of ways in which injury, infection, inflammation, metabolic disease, and neoplasia may present in a nail. The important physical signs are alterations of the nail-plate, such as thickening, thinning, abnormal curvature, onycholysis (lifting of the nail-plate), pitting (Fig. 66), ridging (Fig. 67), discoloration (Fig. 68), and growth changes (Fig. 69). Other possible presentations include paronychia (or inflammation of the skin of the nail-folds) and destruction of the nail unit (Table 14).

Fig. 67 Longitudinal ridging is usually due to decreased growth, and the nails are often thin and poorly made (idiopathic dystrophy).

Fig. 68 Transverse white nails, in this case idiopathic, may also indicate arsenical poisoning.

Paronychia

Paronychia may be acute or chronic. Acute paronychia is painful and is due to bacterial (usually staphylococcal) or viral (herpes simplex) infection. Chronic paronychia is painless and is a traumatic nail dystrophy. Pushing back the cuticles, or removing the cuticles using keratolytics (as used by manicurists), damages the waterproof seal between the proximal nail-fold and the nail-plate that protects the nail matrix. Once damaged, water and debris can enter the nail matrix and produce inflammation of the undersurface of the proximal nail-fold. Cuticle loss is an essential feature of the diagnosis; if the cuticle is intact, consideration should be given to other causes of swelling of the proximal nail-fold. Habit tic deformity is commonly associated.

Secondary infection with *Candida* spp. is common and for a long time was thought to be important in the initiation of paronychia. Whilst it may aggravate the problem, it does not cause it, and paronychia should be thought of as essentially a problem caused by nail manicure. Chronic paronychia may be aggravated by episodes of acute paronychia caused by secondary infection with staphylococci. The proximal nail-fold becomes painful and pus may be expressed.

Treatment involves addressing any associated infection with either topical antifungal preparations or oral antibiotics and antiviral agents.

Tinea unguium (onychomycosis)

See Chapter 7.12.1 for further discussion.

Nail psoriasis

See above 'Psoriasis'

Lichen planus of the nail

See above 'Lichen planus'.

Hair

The hair cycle

Knowledge of the normal hair cycle is fundamental to understanding hair disorders. Hair growth on the scalp is cyclical, with each follicle producing a number of different hairs during a person's lifetime. The anagen growth phase lasts about 3 to 5 years on the scalp, during this phase hair grows at a rate of approximately 1 cm per month. The duration of anagen varies from person to person and is the prime determinant of how long one's hair will grow if not cut. The length of anagen decreases in androgenetic alopecia on the scalp and increases with hirsutes on the face and body. Lengthening of the eyelashes and eyebrows is seen in AIDS, malnutrition, and in chronic liver disease.

The anagen phase is followed by an involutional stage known as 'catagen', which lasts 2 weeks and leads into a 3-month long dormant phase known as 'telogen'. During telogen the hair remains anchored into the follicle but no longer grows. At the end of telogen the follicle awakens and commences production of the next anagen hair. As the new hair grows it displaces the old telogen hair from the follicle. Thus every 3 to 5 years each of the 100 000 hairs on the scalp is shed and replaced. In animals the growth cycle is synchronized leading to a scheduled moult. In contrast, human hair growth is unsynchronized and so, rather than replacing all scalp hair at the end of the growth cycle, between 50 and 100 hairs are lost every day, most of which go unnoticed.

Examination for other causes of hair loss

The scalp should be examined for evidence of disease such as scaliness, redness, injury, or scarring with its associated loss of follicles (Fig. 70). Severe seborrhoeic eczema, which produces diffuse and excessive dandruff, is often associated with hair thinning. Psoriasis, on the other hand, tends to leave some scalp unaffected and mostly the hair grows well. In very thick plaques, hair may get broken off or its growth is occasionally inhibited. Lichen planus and discoid lupus erythematosus both destroy the hair follicles and, respectively, produce a violaceous or red colour as well as scarring. Tinea capitis (see Fig. 19) is a common cause of hair loss in many parts of the world. The acute, painful, boggy, inflammatory swelling of cattle ringworm, known as kerion, is sometimes mistaken for a bacterial abscess and is inappropriately incised—closer examination would show satellite lesions which are clearly not abscesses. Kerions of the head often heal with scarring and some permanent loss of hair. Equally classic in presentation are groups of children with discoid patches of slightly scaly red areas of broken hairs due to other forms of animal ringworm. Fortunately, most adult scalps are resistant to these fungi, though *Trichophyton tonsurrans* infection is on the increase—as are all types of infections—in HIV-positive patients. Tinea capitis is, however, particularly a problem for children. In many parts of the world *T. violaceum* is responsible in black-skinned children, and *M. canis* in white. White scales and scarring in dark heads is often due to favus (a type of tinea capitis). In Africa and the Middle East, favus is due to *T. schoenleinii.* Infection of the scalp with *Streptococcus* spp. or *Staphylococcus* spp. is common in parts of the world where generalized impetigo is common. The scalp may carry a persistent staphylococcus in persons who scratch or pick at their scalp, sometimes known as 'tycoon scalp'. The rash of secondary syphilis often causes a patchy pattern of hair loss scattered over the scalp like numerous 'glades in a wood'. Loss of eyebrows is a feature of lepromatous leprosy (Fig. 71).

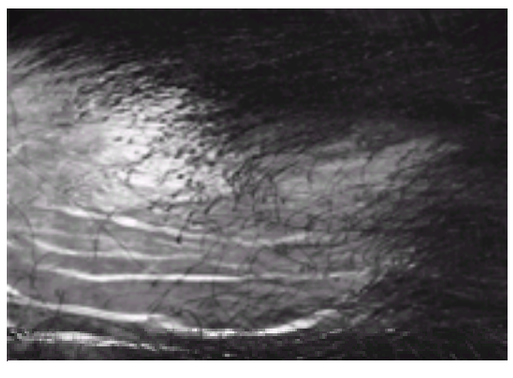

Fig. 70 Scarring is an important prognostic feature because hair loss is irreversible. This pattern of hair loss may be due to a number of chronic inflammatory processes, including lichen planus.

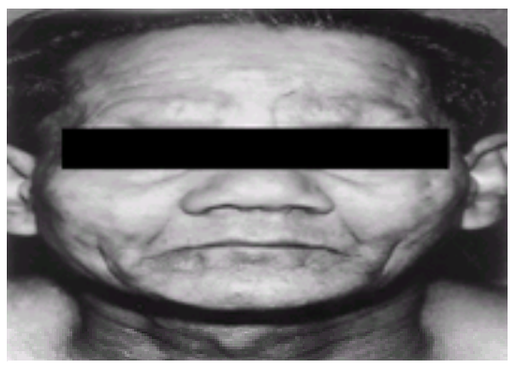

Fig. 71 Loss of eyebrows is a feature of lepromatous leprosy. The skin of the face is thickened but the patient may be unaware of the disease, and for this reason the loss of eyebrows could be the first reason for complaint. Nasal mucosal swelling and erosion is usually obvious.

Hair pulling (trichotillomania) or twisting with subsequent breaking is a common habit in infants and in people with learning difficulties, or occasionally in those who are psychotic (Fig. 72). It is not always consciously done, and if the hair is then eaten there may be little evidence of where the hairs have disappeared to!

Fig. 72 Trichotillomania, or hair pulling, gives rise to hair loss and the hair length varies because it is broken irregularly. It is a common habit of children, but in the adult it is usually an indication of a personality disorder.

Alopecia areata, alopecia totalis, and alopecia universalis

'Alopecia' is a generic term for hair loss; areata is the plural of area. 'Alopecia areata' is a descriptive term for a disorder characterized by one or more discrete circular areas of hair loss (Fig. 73(a) and (b)) which can occur anywhere on the body. Alopecia totalis and universalis are variants of alopecia areata, differentiated only by the extent of the hair loss.

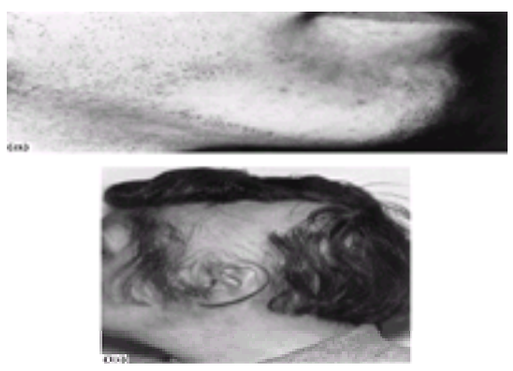

Fig. 73 (a) Alopecia areata is characterized by abortive hair growth and the formation of a short, stubby, 'exclamation mark' hair at the edge of the area of hair loss. (b) Alopecia of the temple or occiput has a poorer prognosis for regrowth.

Telogen effluvium

Diffuse shedding of telogen hairs from the scalp is common. It occurs when anagen hair growth is interrupted, causing follicular hair production to cease and hairs to enter the telogen rest phase of the hair cycle. The telogen phase lasts 3 months before anagen hair growth is resumed. The new anagen hairs push out the old telogen hairs which are shed *en masse*. An acute telogen effluvium almost invariably occurs after pregnancy and may also follow acute illness, major surgical procedure, a crash diet, or on discontinuing or changing the type of oral contraceptive pill. The shedding occurs 3 months after the trigger and may persist for up to 6 months.

Anagen effluvium

Anagen effluvium is the dramatic and rapid hair loss most commonly seen in association with cancer chemotherapy and radiotherapy to the scalp. Other causes are: high-dose colchicine; thallium, mercury, and arsenic poisoning; and cantharadin. The insult to the hair follicle is sufficiently severe to cause an immediate metabolic arrest with complete cessation of hair production. The hair may come out by the root or, if the insult is brief, the hair shaft will narrow, providing a point of weakness that subsequently snaps off. The follicle may remain in anagen, in which case recovery is quick, or move into telogen, in which case regrowth will be delayed by about 3 months.

Androgenetic alopecia

Androgenetic alopecia is a progressive patterned baldness, which is sufficiently common among both men and women to be considered a secondary sexual phenomenon. It is also called common baldness, male-patterned baldness, and female-patterned alopecia. When it occurs prematurely in a man, it can be an unwanted and distressing event and patients may present for treatment. Among women it is usually both unwanted and unexpected at any age and women commonly present for both diagnosis and treatment.

Hirsutes

While unwanted hair occurs in both men and women, men rarely complain of this to doctors. Unwanted hair in a female or hirsutes is difficult to define objectively due

to racial, cultural, and fashion norms. Superfluous hair on a woman in a distribution that mirrors the development of secondary sexual hair in a male is common. Most cases are due to hair-follicle hypersensitivity to normal levels of circulating androgens, while a proportion will be due to elevated levels of circulating androgens. Such patients will usually have other features of systemic androgen excess, such as menstrual irregularity, severe acne, and premature androgenetic alopecia.

Hirsuitism should be distinguished from hypertrichosis (Table 15), which is the widespread overgrowth of non-androgen-dependent hair and that does not respond to antiandrogen therapy. Hypertrichosis may be primary (in which case it is usually apparent prior to puberty) with the hair evenly distributed over the back and limbs. Although prepubertal hypertrichosis is commonly familial, a positive family history is not always found. Secondary causes of hypertrichosis include drugs such as minoxidil, diazoxide, ciclosporin, and phenytoin.

Sweat glands

Apocrine

Apocrine glands occur throughout the skin surface in the embryo, but subsequently disappear to leave just those in the axillas, areolas, and anogenital region in adults. The secretions are formed by the dissolution of apocrine gland cells which are discharged in the hair follicles close to the surface of the skin. They are not active until puberty. Bacterial decomposition accounts for body odour, while the secretions are important in animals for identity and marking out territorial areas. They are also important sexual organs. All such functions are vestigial in man. People sometimes present complaining of body odour. Washing with soap and water is the first phase of management. Deodorants reduce the bacterial flora. Eating garlic and betel nuts should be discouraged since these 'perfumes' are excreted in apocrine sweat. Apocrine sweat is sometimes coloured. If staining is severe and uncontrolled by deodorants, it may be necessary to excise the glands. Retention of apocrine sweat and extreme irritation, known as Fox–Fordyce disease, is similar to prickly heat. Treatment may include the use of topical steroids, destruction with cryotherapy, or excision.

Hydradenitis suppurativa

This is a relatively common and often misdiagnosed condition of the skin characterized by boils, pimples, sinus formation, and comedones (Fig. 74) in the axilla, groin, and submammary areas. It is also known as apocrine acne, as these are the sites of apocrine glands. The lesions often heal with scarring, which may be pitted and cribriform. It may occur alone or in association with severe cystic acne of the trunk, pilonidal sinus, and dissecting cellulitis of the scalp, which are all part of the follicular occlusion tetrad. Dissecting cellulitis of the scalp is a severe folliculitis rather than a true bacterial cellulitis. Hydradenitis suppurativa is very commonly associated with obesity, and may be seen in women along with other features of androgen excess.

Fig. 74 Axilla showing dusky or violaceous erythema overlying cystic 'blind' boils. The follicular prominence and comedone formation is characteristic of early hydradenitis.

Boils should be treated as they develop and prophylaxis given to guard against the development of new lesions. However, there is considerable person-to-person variation in the severity and frequency of the development of the boils, which will determine the need for prophylaxis. Boils may be incised and drained or injected with steroid as they tend to be inflammatory rather than infective, although a short course of oral antibiotics may be indicated based on the results of culture and sensitivity testing. Antibiotic prophylaxis is modelled on the principles of acne therapy. However, because antibiotic resistance does seem to be a problem, they are generally rotated every 4 to 6 months to prevent the emergence of resistance. Those antibiotics shown to be effective include minocycline, co-trimoxazole, metronidazole, erythromycin, tetracycline, cefalexin, and clindamycin. Other therapies that can be used for resistant cases include cyproterone acetate, spironolactone, and isotretinoin. If these treatments are unsuccessful, wide surgical excision or liposuction (which disrupts the glands) may be considered.

Eccrine

Humans possess about 3 to 4 million sweat glands, equivalent in weight to one kidney. The secretory coil produces a plasma-like fluid, which can be secreted at a maximum rate of 2 to 3 l/h. Sodium is reabsorbed in the sweat duct.

While eccrine sweat glands occur in all areas, those in the hands, feet, axillas, and face frequently sweat profusely in the absence of general sweating. Humans rely on evaporation rather than insulation or panting for protection against a hot environment. Generalized sweating occurs when the body temperature rises and is a feature of fever as well as thermoregulation in a warm climate, and when the metabolic rate is increased, as in exercise or thyrotoxicosis.

Eccrine sweat glands are largely innervated by unique postganglionic sympathetic fibres that release acetylcholine at the neuroglandular junction. The control centre is located in the hypothalamus. It is important to consider whether the sweating is appropriate for the degree of stimulus (Table 16).

Emotional- or anxiety-induced sweating is commonly inappropriate for the degree of anxiety. Many teenagers complain of sweaty hands and feet as well as the smell resulting from the bacterial breakdown of skin and clothing. The fear of being unwelcome increases their anxiety and subsequent sweating. It may summate with thermoregulatory sweating and therefore be worse in hot weather or at a dance. Winter clothing is often more troublesome than the loose garments of summer. Sweating of the hands and feet occurs with acrocyanosis and with some forms of keratoderma.

Segmental, unilateral sweating is often due to irritating lesions of the spine and requires a neurological opinion.

Excessive sweating (hyperhidrosis) contributes to tinea pedis and to eczema from footwear.

Treatment of hyperhidrosis

A sympathetic listener is helpful, as is simple advice on hygiene, washing, keeping cool, and appropriate clothing. Basic points of management include the avoidance of obesity, relaxation exercises are helpful if the patient is unduly self-conscious. Cotton-fibre clothing, for example, is more appropriate than non-absorbent fibres. Many shoes, including their linings, are now made from materials that prevent the absorption of sweat and keep the foot and sock wet. Frequent changes of socks prevents bacterial overgrowth, and wearing leather shoes or sandals reduces discomfort. Tranquillizers are sometimes helpful and propantheline, 15 mg every few hours, may be of benefit or can be reserved for a social occasion. The abolition of sweating carries a risk of hyperthermia.

Hypohidrosis

This occurs in the newborn and in premature children during the first month of life, and is also seen in infants as a result of sweat-duct occlusion, especially in the flexures of skin folds on the neck. It may also result from exfoliative dermatitis or erythroderma and is a feature of hypohidrotic ectodermal dysplasia (see Chapter 23.2). Such patients are usually male and are susceptible to heat stroke, and therefore early diagnosis of the affected baby in a hot climate is important. Absence of

sweating causes loss of skin moisturization and impaired grip. Examination of the palmar surface of the fingers with a magnifying lens will show the absence of duct orifices.

Miliaria, prickly heat

Miliaria crystallina is a superficial obstruction of sweat glands, producing clear vesicles. This may occur when more sweat is produced by the glands than the ducts can absorb. Miliaria is commonly seen in patients with high fever. Deeper obstruction gives rise to red, itchy papules known as prickly heat. Around one in three people exposed to hot climates are affected; while prickly heat sometimes begins within a few days of arrival in hot climes, it is commonly a problem 2 to 6 months later. Occlusion of the skin by impermeable clothing aggravates the condition. Bacteria may play some part in its generation and in the complication of staphylococcal abscesses. It is a contributory factor to extreme thermal stress in unacclimatized people. This is also a feature in people working in hot industries around furnaces or in underground mines. Relief from sweating even for a few hours is essential. Loose, non-occlusive clothing and exposure of the skin folds as much as possible is beneficial. Vitamin C (1 g daily) was advocated by dermatologists in the British Army in Malaysia. It is important to realize that severe hypohidrosis of the trunk and limbs may be missed as a cause of asthenia if the face is sweating. The small particles of powder in calamine shake lotions promote cooling by increasing the surface area. A localized loss of sweating ability may be due to tuberculoid leprosy, syringomyelia, and diabetes mellitus.

Skin disorders affecting the genitalia

A diagnosis of disorders affecting the genitalia cannot be made without looking. Natural shyness on the part of the patient or lack of zeal on the part of the doctor are common. Racial and religious grounds for incomplete examination must be overcome by appropriate selection of the examiner and chaperone, as well as an interpreter.

It may be inappropriate to delve into the sexual, gynaecological, or medical history of the patient, and so initial questioning may be limited until an examination has indicated the nature of the disease. A contact dermatitis may require the most detailed and searching questioning.

Many skin conditions of the vulva or penis can be best diagnosed by examining the rest of the skin. For example the knees, scalp, and elbows in psoriasis, the front of the wrists and shins in lichen planus, the mouth in pemphigus, or the neck, breasts, and wrists in lichen sclerosis et atrophicus.

Infections

Infections are commonly transmitted by sexual intercourse and tend to be associated with vaginal or urinary meatal symptoms (see Chapter 21.7).

The pubic region is commonly affected by nits and crab lice, and viral molluscum contagiosum causes smooth, pearly, umbilicated papules (see Fig. 111).

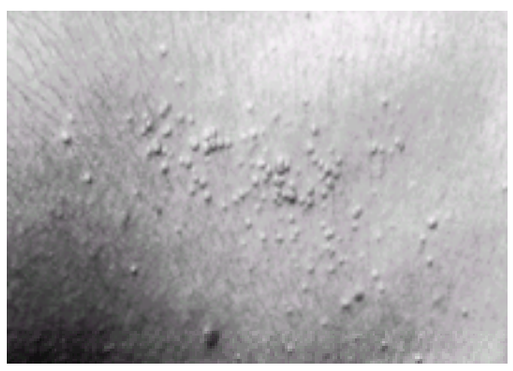

Fig. 111 Molluscum contagiosum: groups of virus-induced papules characterized by a central punctum.

Genital warts

See Chapter 7.10.18 for further discussion.

Candidiasis

See Chapter 7.11.1 and Chapter 21.4 for further discussion.

Herpes simplex

See Chapter 7.10.2 for further discussion.

Adult intertrigo

Obesity and sweating predispose to mixed irritation and infection in the occluded skin under the breasts and in the axillas and groins. The affected area is moist, red, fissured, and malodorous. Attempts at keeping the site dry and free of excessive infection have been improved by preparations such as miconazole and hydrocortisone, and ZeaSORB® powder which acts as a drying agent without too much caking. Washing and gentle drying is the most important therapy. Blind boils and comedones are likely to be due to hidradenitis suppurativa.

Psoriasis in the flexures is usually well defined, bright red, and, unlike at other sites, is non-scaly. It is worth treating initially for 3 days with a strong steroid because this sometimes clears the psoriasis. More often the lesions persist, in which case strong steroids are then harmful since they cause so much atrophy. Hydrocortisone can be used but is only mildly effective. It is important to protect the skin from excessive infection by regular washing.

Dermatitis of the genitalia is commonly due to contact with the agents listed in Table 17, some of which are added to the bath and inadequately mixed with the water. Certain deodorant sprays may cause a considerable immediate contact swelling. Mixed-type infections may contribute to the problem. Persistent pruritus and scratching is a very common disorder, producing thickening of the skin and a range of colours from white-fissured areas to pigmented and violaceous plaques.

In uncircumcised adult males a persistent reddish brown, somewhat fixed balanitis, is heavily infiltrated with plasma cells. This benign condition, known as Zoon's balanitis, is cured by circumcision.

Leucoplakia versus the atrophy of lichen sclerosus et atrophicus

This is often confusing, since it is possible for the skin of the genitalia to be thinned but nevertheless to be covered by a thickened scale. The lesion of lichen sclerosus is well defined (Fig. 75), white, and may have a violaceous border. Small haemorrhagic blisters are common, especially in children, and should not be mistaken for sexual abuse. The perianal areas are always involved, especially in children (Fig. 76). Intractable itching, burning, or soreness of the perineum or genitalia is unfortunately common. In young women it usually slowly improves; in older persons it persists. Lichen sclerosus et atrophicus responds well to high-potency local steroids.

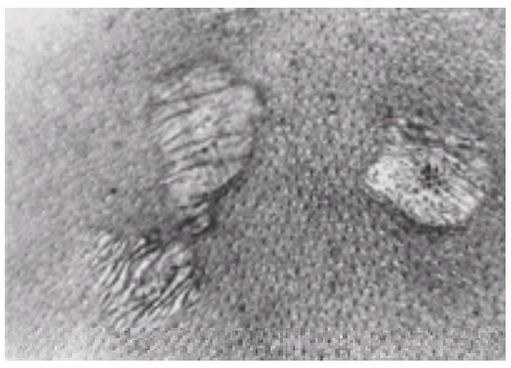

Fig. 75 Lichen sclerosis causes tissue-paper-like crinkling or atrophy of the skin. The border is often violaceous but the centre of the lesion is white.

Fig. 76 Lichen sclerosis of the vulva in a child tends to clear at puberty. In the elderly it is a persistent cause of irritation.

Histologically, lichen sclerosus et atrophicus is characterized by an extremely thin epidermis with a thickened scale and an acellular homogenized upper dermis. By contrast, leucoplakia and lichen simplex are composed of a greatly hypertrophied epidermis and usually thickening of the underlying dermis as well. A biopsy should be performed where there is diagnostic doubt, but leucoplakia and lichen sclerosus may coexist in as many as 24 per cent of patients. Although lichen sclerosus et atrophicus, especially when damaged by scratching, may develop a squamous-cell epithelioma, there is no advantage nor relief of discomfort by prospective vulvectomy.

In the case of leucoplakia, vulvectomy is usually advocated as there is much greater chance of the development of a squamous epithelioma. However, since the skin is predisposed to lichen sclerosus in areas well beyond the genitocrural folds, simple vulvectomy is not a satisfactory treatment for lichen sclerosus itself. Attention to hygiene is important, as is exclusion of mixed infections or contact dermatitis. Rarely, perineal discomfort and eczema may be due to vitamin B_2 deficiency. The main treatment of lichen sclerosus is with strong topical steroids.

Well-defined asymmetrical plaques of red pigmented skin should be biopsied to exclude intraepidermal carcinoma to which this site is predisposed.

The term 'kraurosis vulvae' is now obsolete since it does not differentiate senile atrophy from the now well-recognized lichen sclerosus.

Urticaria

Urticaria is a transient swelling and/or flushing of the skin. The underlying vasodilatation and accumulation of tissue fluid in the dermis is due to a succession of inflammatory mediators acting mainly on the small blood vessels. The time taken to bring their effects under control varies, and thus the inflammatory response varies from the very transient to a more persistent inflammation overlapping with vasculitis.

The knowledge that histamine plays a part in immediate-type (anaphylactic) hypersensitivity has led to the widespread misconception that all urticaria must be allergic. A non-immunological pharmacological explanation is more likely in most cases.

Immunology

Allergens of the type commonly incriminated in sufferers from atopic disease bind to IgE antibodies attached to the surface of mast cells or basophils, these cells then release various mediators including histamine, serotonin (5-hydroxytryptamine), and the slow-reacting substance (leucotriene) of anaphylaxis. Such allergens include egg white, cows' milk, house dust, dandruff, feathers, and tomatoes. It is commonly a contact-type reaction, affecting the lips during eating or some other parts of the skin when in contact with animals or house dust. Transfusion reactions and some drug rashes are due to complement-fixing antibodies attached to blood cells.

The urticaria of serum sickness, penicillin reactions, the acute illness of systemic lupus erythematosus, and many infectious diseases are partly due to immune complexes of immunoglobulins and allergen with subsequent complement activation.

Complement activation

Although complement is activated by immunological reactions, it is also activated enzymatically by proteases (such as plasmin) when there are insufficient natural inhibitors of this mediator in the serum and tissues. Congenital or acquired deficiencies in inhibitor levels account for some forms of angioedema, and for hereditary angioedema in particular. The activation of complement by the alternative pathway may explain the aetiology of some non-familial cases.

Histamine liberators

Some drugs and foods release histamine from mast cells, or at least make such release more likely by inhibiting controlling factors: inhibition of prostaglandin activity may be one such mechanism. Examples of mast-cell stimulators include morphine, codeine, thiamine, polymyxin, and D-tubocurarine. Bee venom, strawberries, and shellfish as well as aspirin, salicylates, benzoates, and tartrazine are enhancers of an urticarial tendency, bringing it to the fore in susceptible subjects as well as occasionally initiating the eruption.

Genetic factors

Familial urticaria is a well-recognized phenomenon. There are reports of hereditary angioedema affecting many members of large families. In this condition the angioedema occurs without associated urticaria. The autosomal dominant pattern of inheritance is mediated through an absence of the C_1-esterase inhibitor. Familial cold urticaria is another autosomal dominant disease, and has been described in several families in the United States, France, and The Netherlands. A low level of chymotrypsin inhibitors was detected in one family. Studies of HLA antigens have been rewarding. BW35 has been associated with acute ordinary urticaria, while HLA-B1*04 (DR4) and its associated allele, DQB1*0302 (DQ8), are raised in patients with chronic idiopathic urticaria compared with a control population.

Careful studies have incriminated candidiasis, though it is not so important a factor in the experience of the majority of practitioners.

Types of urticaria

Variations are observed in the number, size, and depth of weals, as well as in the sensation experienced by the patient. Moreover, the lesions vary in their degree of persistence. Such features allow the classification of different types of urticaria, but there is considerable overlap in their aetiology and pathogenesis.

Only a minority of cases are due to an early, avoidable, and identifiable cause. Although some are due to autoantibodies, non-immunological pharmacological causation is not rare. The majority remain of unknown aetiology.

Contact urticaria

This is a weal-and-flare reaction lasting for 20 to 40 min after the application of certain agents to the skin. Some reactions may be IgE-mediated, such as those produced by animal dander, saliva, or seminal fluid, but most are probably non-immunological, as with the nettle or jellyfish sting, or the solar or aquagenic varieties of urticaria. Often there is a consequent or associated dermatitis, as in atopic eczema. Many of the ointments used to treat dermatitis contain bases such as sorbic acid or polyethylene glycol which cause immediate stinging and a slight swelling.

Cholinergic urticaria

Cholinergic urticaria is characterized by numerous, superficial, small swellings that sting, smart, or itch and are surrounded by a blush lasting for only a few minutes (Fig. 77). The cause is unknown. The commonest pattern is found in adolescents and young adults and, like blushing, is brought on by emotion, exercise, and hot baths.

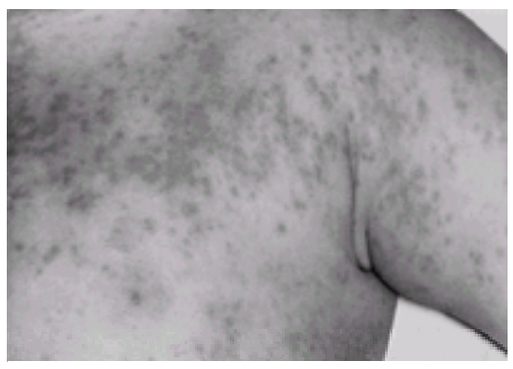

Fig. 77 Cholinergic urticaria is like blushing brought on by emotion, exercise, and heat. It is transient, lasting no more than about 15 min, and may be associated with small, superficial weals with a prominent flush. These tend to sting rather than itch.

Heat urticaria

This is a rare local response to heat in which histamine is released or complement activated.

Angio-oedema

A few, deep, large swellings, which may be tender and often itchy, sometimes preceded by redness, lasting several hours or even days are characteristic. Proteases such as complement, plasmin, and kinins are incriminated.

Ordinary urticaria or hives

This is characterized by numerous weals of all sizes, and varying degrees of pallor or redness, which itch and last for one or more hours, but not usually for more than a day. Successive lesions may account for long illness: chronic urticaria is arbitrarily defined as continuous or recurrent lesions of more than 3 months' duration. Histamine is the principal mediator. Current evidence supports the view that skin blood vessels have both H_1 and H_2 receptors.

Time of onset

Cholinergic urticaria (see above) develops abruptly and instantaneously within minutes of the triggering event. Ordinary urticaria also develops within minutes of the release of the mediator. However, foodstuffs and certain allergens, or drugs such as aspirin, have to be digested and absorbed before the mediators are released. Ordinary urticaria is often difficult to relate to events in the patient's life for this reason. Delayed onset is a well-recognized phenomenon of some of the physical urticarias (see below). Thus delayed dermographism is the development of redness and slight wealing occurring several hours after scratching the skin. Delayed-pressure urticaria is a tender swelling appearing 2 to 12 h after localized pressure injury to the skin. It is possible that the insult localizes noxious agents such as soluble immune complexes, or that mechanisms such as transient ischaemia and the release of proteases bring to light homeostatic defects (for instance, a deficiency of inhibitors of complement or other proteases).

Physical urticaria

Several types of urticarial eruptions are only caused by specific physical insults: for example, from sunlight, cold, heat, pressure, scratch, or stretch.

Solar urticaria is uncommon. A weal develops within 30 s to 3 min of exposure to the sun; however, tolerance may develop in such habitually exposed sites as the hands and feet. The differential diagnosis of porphyria, lupus erythematosus, or photosensitivity following drug ingestion has to be considered, in which case the urticaria is more persistent, and because the longer, more penetrating ultraviolet rays are responsible, it can occur even on a cloudy day or when the skin is protected by glass, clothing, or sunscreens.

Familial cold urticaria is an autosomal dominant disease, usually presenting in infancy, in which the rash develops up to several hours after exposure to, for example, cold winds. Fever and joint pains accompany the rash and there is a leucocytosis. Low levels of a chymotrypsin inhibitor have been demonstrated. It may not be induced by the application of ice to the skin.

Acquired cold urticaria is the most common form of cold urticaria and occurs within a few minutes of plunging into cold water or after applying ice to the skin; this is one cause of sudden death in young people plunging into ice-cold water. Mast-cell degranulatation is a feature.

Papular urticaria

This is the only form of urticaria to have a persistent epidermal component. Most often papular urticaria is caused by insect bites. The epidermis is either damaged directly or by mediators in the upper dermis which evoke an eczematous response, so that oedema of the epidermis and a proliferative repair effect results in a typical itchy and persistent papule. Such lesions are usually excoriated, whereas most urticarias are not deeply scratched but merely rubbed. They often blister (see Fig. 90).

Fig. 90 Typical multiple blisters due to insect bites during the rainy season in India.

Scaling is not a feature of urticaria. While acute dermatitis and some erythema initially appear to be urticarial in nature, the development of scaling immediately excludes such diagnoses.

Distribution of the rash

Cholinergic urticaria favours the head and upper trunk. Angio-oedema most commonly involves mucocutaneous junctions such as the lips, eyes, and penis. The physical urticarias clearly relate to sites of exposure; solar urticaria affects the face and the dorsum of the hands, or, if tolerance has developed, at sites exposed for the first time during the summer. Pressure urticaria favours the soles of the feet when walking or digging, or the backs of the thighs or lumbar region when sitting.

Bizarre patterns

Evolving and resolving urticaria inevitably exhibits a changing morphology. The redness of the vasodilatation merges with the veiling pallor of the oedema. Healing in the centre and peripheral spread often produces bizarre gyrate or circinate and serpiginous patterns, but they are transient and never scaly, unlike similar patterns in the erythemas or epidermal diseases such as psoriasis.

Investigations

History-taking is the most effective investigation of urticaria. It is important to establish whether there is urticaria and angioedema, or angioedema alone. In simple cases, the individual weals last for less than 24 h. Persistence beyond 24 h, and in particular resolution with bruising, is highly suggestive of urticarial vasculitis and justifies a biopsy. The chronicity of the urticaria is also important. Acute urticaria lasts less than 6 weeks and investigations are not required, a drug and dietary history generally suffice. If the lesions have been appearing for more than 6 weeks, further questioning is important, and some investigations may be considered. Trigger factors should be specifically asked for to detect the physical urticarias.

Exercise or a hot bath reproduces the lesions of cholinergic urticaria. Intradermal histamine, 1 µg in 0.1 ml saline, produces a weal that should disappear within 1 h. This disappearance is delayed in pressure urticaria and in immune-complex disease. Localization of a noxious agent results in a persistent lesion. In practice, avoidance of cause can be advised only if this is recognized after taking a history and examining the patient. The two most helpful investigations are a full blood count and the erythrocyte sedimentation rate (ESR). An eosinophilia should alert one to parasites such as microfilaria or trichiniasis, while a raised ESR is due to a systemic illness such as sepsis, malignancy, or 'collagen' disease.

Rubbing or scratching the skin with a fingernail produces a weal and flare in the dermographic subject within 5 min. A 4- to 6-kg weight hung for 10 min over the shoulder on a bandage or belt causing a tender swelling 2 to 8 h later, reveals delayed-pressure urticaria. A biopsy at this stage for immunofluorescence studies may confirm the localization of immune complexes. The white cell count at the time of the biopsy may show neutrophilia, especially if there is accompanying fever. A reduction of serum C_1-esterase inhibitor should be looked for in patients with angio-oedema, especially if initiated by minor surgery, associated with abdominal pains, and if other members of the family affected. Complement levels are not a reliable guide to the participation of proteases and hardly influence the management of urticaria.

Chronic urticaria is a known symptom of filariasis and strongyloidosis, but in ascariasis and enterobiasis it occurs, if anything, more often in controls. Urticaria is such a difficult disease to assess that possible aetiological factors, such as parasitic disease, are worth treating in their own right rather than in the expectation of resolution of the eruption.

About 25 per cent of cases of acute hepatitis B present with urticaria.

Foci of infection as a cause of urticaria are statistically difficult to support, but dental and sinus infections continue to be described as aetiological factors based on impressive case histories.

Bee stings

See Chapter 8.2 for further discussion.

When should urticaria be taken more seriously?

Urticaria is life-threatening when it is part of anaphylaxis, when angio-oedema involves the upper respiratory tract, or when it is part of a systemic immune-complex disease and is associated with more dire pathology such as meningococcal septicaemia or lupus erythematosus. The latter type of urticaria is recognized by its more persistent lesion, lasting at least 1 to 2 days, and often tender and ultimately purpuric. It should be remembered that all acute urticaria may be very widespread and be accompanied by joint pains, stomach aches, and fever. However, if the individual lesion lasts for only a few hours it is less likely to be due to a noxious circulating trigger such as an immune complex or infective organisms.

Management

Removal of the known physical factor and the known trigger is helpful, but no cause is found in the majority of adult cases of chronic urticaria. Insect repellents and topical steroids play some part in the management of papular urticaria. Elimination diets require the exclusion of suspected foods for at least a week, followed by their reintroduction for a week; this cycle should be repeated at least three times to be confident of a real effect, which, however, is convincing in only about 5 per cent of patients with urticaria. Functional autoantibodies against the high-affinity IgE receptor (FceRI), or less commonly IgE, have been demonstrated by autologous-serum skin testing and *in vitro* histamine release from basophils, and Western blotting and an enzyme-linked immunosorbent assay (**ELISA**), in over 30 per cent of patients with chronic 'idiopathic' urticaria. It may respond to immunotherapy. Some European centres rely on studies of the gastrointestinal tract to reveal the presence of *Candida albicans*, *Campylobacter* spp., and other infections as possible causes of urticaria. Food, medicines, and infectious or parasitic diseases are the commonest suspected factors. However, in Europe and the United States physical urticaria accounts for more than half of the patients in some series. Cold, heat, and solar urticaria often respond to the induction of tolerance by subthreshold desensitization. It is useful to try different antihistamines because of much individual variation response and in their side-effects. Antihistamines are often prescribed in too low a dosage to be effective, and patients should be encouraged to rest at home taking a rather higher dosage. The evidence that skin blood vessels possess both H_1- and H_2-receptors has encouraged trials with a combination of their antagonists. At present, the financial cost and large number of pills that have to be taken every day is a disadvantage. Most H_1-antihistamines are cheap and free of serious side-effects. Drowsiness is often troublesome, but the variations in response are considerable. Otherwise, they are effective in the majority of patients, providing they are taken regularly to prevent the urticaria rather than to treat the existing weals. Long-acting antihistamines are worth trying when short-acting ones fail. Although the value of combined H_1- and H_2-blockers (4 mg cyproheptadine and 300 mg cimetidine, four times daily) remains unproven, like all regimens it has resulted in some

individual successes. Hydroxyzine (10–25 mg, three times a day) or cetirizine 10 mg, once or twice a day, is often effective in treating dermographism or cholinergic urticaria. Nifedipine in conjunction with antihistamines has its advocates. Avoidance of known urticaria triggers, such as aspirin, tartrazine, benzoate, and other salicylates, often requires a rather complex diet. Prednisolone should be reserved for a few days' treatment for severe acute urticaria or exacerbation of chronic urticaria.

For impending upper respiratory obstruction, adrenaline (epinephrine) 1/1000, 0.5 ml (adult dose), is first given intramuscularly and hydrocortisone 100 mg intravenously. Such an acute emergency requires maintenance of the airway, if necessary by intubation or tracheotomy, and the administration of oxygen.

Management of autosomal dominant hereditary angio-oedema

This should be suspected if there is a family history of angio-oedema or a few long-lasting swellings precipitated by trauma. Signs include a transient erythema followed by the oedema, and a recurring colicky abdominal pain is often a feature. The diagnosis should be confirmed by looking for low serum levels of functional a-neuroaminoglycoprotein, C_1-esterase inhibitor (normal 18 ± 5 mg/100 ml, although levels vary).

Prophylaxis includes care to protect against trauma, especially in the region of the mouth and neck after dental manoeuvres. Danazol or stanazolol prophylaxis may be appropriate if episodes are frequent. Methyltestosterone, 10 mg as linguets after breakfast and another when there is a suspicion of developing oedema, often aborts attacks. Fresh plasma, containing C_1-esterase, may be given before surgery or at the initiation of an attack. Unlike other forms of chronic urticaria, adrenaline (epinephrine), antihistamines, and corticosteroids are of only little benefit. Intravenous Trasylol is an inhibitor of proteases and is sometimes helpful, as is epsilon-aminocaproic acid (**EACA**) 12 to 18 g daily in divided dosage. The most effective treatment is danazol, which may be supplemented by fresh plasma during an attack or, if available, a C_1-esterase inhibitor concentrate.

Cutaneous vasculitis

The broadest definition of vasculitis is 'the response of small blood vessels to injury'. No other definition encompasses its great variety, which ranges from a transient increase in permeability or wealing, to coagulation and necrosis of the vessel wall. It is caused by agents such as immune complexes, toxins, or by physical stimuli such as cold and heat, as well as impaired perfusion.

The term 'vasculitis' includes many diseases described elsewhere in this book, such as Henoch–Schönlein purpura, polyarteritis nodosa, nodular vasculitis, Wegener's granulomatosis, hypersensitivity angiitis, and allergic granulomatosis. Examples of diseases sometimes included within this term are Behçet's syndrome, pyoderma gangrenosum, purpura fulminans, thromboangiitis obliterans, erythema nodosum, chilblains, atrophie blanche, and livedo reticularis.

Vasculitis overlaps with urticaria and with infarction or gangrene. To use some of the older terminology, some authors have described urticaria as the predominant feature of Henoch–Schönlein purpura in children, and a number of vasculitic syndromes have more recently been described in which urticaria is the only skin manifestation. At the other end of the spectrum, 'necrotizing angiitis' and 'polyarteritis nodosa' are labels often given to infarctive or more destructive patterns of vasculitis.

The physical signs of the skin disease are more or less recognizable as distinct patterns. The names they have been given are of dubious value when it comes to managing the disease. It is possible to explain these patterns and to decide what aspects of the physical signs are the most useful clues to pathogenesis.

Pathology and nomenclature

When a vessel is injured a response ensues which removes or neutralizes the cause of the injury, followed by repair. The response depends on the intensity of the injury, on the efficiency of the inflammation, and on the rate and effectiveness of the repair. Herein lies one of the main reasons why a particular injury does not always produce the same rash. The inflammatory response is a very complex sequence of events (Fig. 78), subject to considerable modification by each individual's particular range of mediators. Important factors explaining variability are local-tissue architecture and previous disease experience, such as scarring or even temporary exhaustion of mediators by prior injury. Formerly, authors used to describe sites of lowered resistance. We now know that such sites comprise areas of non-homogeneous blood supply with hypoxia, leakiness, and blood stasis, as well as exhausted mast cells and endothelial cells undergoing various stages of repair, or upgrading of cell-wall adhesion factors for white cells. Paralysis of the mononuclear phagocytic system is also important.

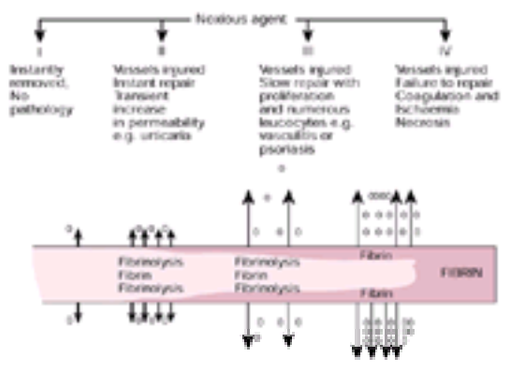

Fig. 78 A spectrum of inflammatory events ranging from a transient wealing, or almost physiological permeability through inflammatory neutrophilic infiltration and oedema, to complete necrosis of the vessel, due to thrombosis, and coagulation or destruction of its wall.

This variation in the inflammatory response can be superficial or deep, and is modified by the distribution of the injury—gravitational, light-exposed, cold-exposed, or at sites of pressure or abrasion; this is one explanation of the physical signs that make up classifiable rashes. A rash with urticarial, purpuric, nodular, pigmented, and necrotic lesions is better not given an eponymous name to any particular combination of physical signs.

Another reason for the later delineation of different syndromes such as Behçet's triad, cutaneous polyarteritis nodosa, or limited Wegener's granulomatosis was the recognition that vasculitis may be confined to either one or more organs. But this, too, is of dubious value: Behçet described mouth, eye, and genital lesions, but the disease is often more widespread; Wegener's granulomatosis affects the respiratory tract but the skin and kidneys are also frequently affected.

One other point of debate is the value of the term 'arteritis'. Smooth muscle in the arterial wall is damaged by ischaemia, which, in most small vessels, is attributable to vasoconstriction, coagulation, and thrombosis, or to obstructed flow due to a more distal vasculitis. Thus 'malignant hypertension', 'coagulation', 'embolism' (Fig. 79), 'thrombotic thrombocytopenia', or 'vasculitis' are often sufficiently appropriate diagnostic terms and are more helpful when considering prognosis, aetiology, or management. The histological diagnosis of arteritis is similarly unsatisfactory. Damaged venules can themselves look like small arteries; even when an artery is clearly involved, it is rarely possible in the same section to see the more distal vasculitis responsible for the obstruction to blood flow and consequent ischaemia.

Fig. 79 The white infarct characteristic of embolization or arterial block. Most cutaneous vasculitis labelled as 'arteritis' is, in fact, venular, but the white infarct is characteristic of arterial occlusion.

Harmful agents responsible for vasculitis

Immune complexes, infective agents, drugs, food additives, and circulating particulate matter all injure blood vessels. It is probable that these are present in small amounts in all of us some of the time, but it is likely that the injuries are often so mild as to be imperceptible and quickly repaired.

Immune complexes

The 'defensive' system of antigen-clearing involves complexing the antigen with antibody and complement to make phagocytosis by macrophages more inevitable. As this is a system that is used to remove damaged tissue, it is often difficult to distinguish whether damage preceded or is a consequence of the immune complex. It is the process of complexing that activates complement not the complex itself. Free and poorly complexed antigen may indeed have a greater potential to cause damage than well-complexed material. Trapping antigen in a tissue and its exposure to immunoglobulin and complement is determined by local events having little to do with what circulates in the bloodstream.

Immune complexes follow the ingestion of food, the presence of a fetus in the pregnant, invasion of the body by parasites such as scabies, or a neoplasm such as breast cancer, and can be demonstrated in everyone following the most mild of virus infections. For this reason, the mere demonstration of immune complexes is not enough to blame them for coincidental vasculitis. Excess antigen is released when infections are overwhelming or when tissues are broken down by immune attack or neoplasia, and sometimes as a consequence of food or drugs.

Immune complexes are mostly harmless but become harmful when they are of certain size and shape or composition. Actual harm is observed only when, and as, they are localized at a site ill-equipped to deal with them and slow to repair the damage done by them.

Often an alteration in host response prevents adequate neutralization of even mildly noxious agents. Recent exposure to similar noxious agents, exhaustion, and insufficient time for recovery of fibrinolytic mechanisms or the phagocytic potential, and the secretions of the mononuclear phagocytic system explain why repeated or continuous exposure to harmful agents precipitates vasculitis. Such an explanation most often explains localized recrudescence in the nose, in Wegener's granulomatosis, or in legs affected by gravitational eczema, ulcers, or atrophie blanche. Factors such as severe infection, foods, drugs, diseases promoting coagulation (for example, hepatitis), and malignancy often alter the inflammatory response and need not act as specific triggers. Immune complexes circulating in a patient known to have disseminated lupus erythematosus may suddenly become damaging when any of these other factors affect the patient. Localization of the defective inflammatory process is often determined by environmental factors: cold exposure, abrasion, or pressure on the skin causing mild stasis and ischaemia are well-known examples met clinically; these conditions are also used experimentally to demonstrate the localization of harmful agents from the bloodstream.

In recent years the discovery of congenital defects in complement and other protease inhibitors has explained why harmful agents are inadequately neutralized in some people. Hereditary angio-oedema is a good example of such a disease. There is a congenital absence of a complement inhibitor, but only in certain circumstances is this important. Trauma or infection sets off the sequence of events, which in this case includes the activation of complement by plasmin, dependent in its turn upon the secretion of plasminogen activator by damaged endothelium. Normally this is balanced by other inhibitors and by absorption into small amounts of fibrinogen and fibrin.

Local deficiencies in the sequence of events triggered by injury depend upon blood flow and diffusion, not only of activators but also of inhibitors. Injury usually releases fibrinolysis activators from endothelium, and heparin, histamine, and hyaluronidase, etc., from adjacent mast cells. These increase permeability and diffusion but prevent coagulation and complement activation. Repeated inoculation of histamine can itself cause vasculitis. It is not always necessary to invoke an immunological mechanism. When the mast cells and endothelium are more or less exhausted, and when activating or inhibiting products are released by adjacent epidermal injury, the proteases, complement, kinins, and materials like fibrin or C-reactive substance occur in sufficient quantities to perpetuate the inflammation and to attract white cells in large numbers.

Diagnosis

Almost essential to the diagnosis is the presence of purpura (Fig. 80), but a tender urticarial lesion that lasts for more than 12 h and leaves a slight bruise on resolution falls within the term 'vasculitis' (Fig. 81). The histology of such a lesion often shows more perivascular neutrophils than the common short-lasting urticarial weal. At the other end of the spectrum is obliterative or sclerosing thromboangiitis in which total occlusion of a small vessel often prevents exudation, and because there is no neutrophilic infiltrate, there may not be such acute destruction of the vessel wall. A similar appearance may be observed in disseminated intravascular coagulation (**DIC**) (Fig. 82) and in platelet embolic diseases.

Fig. 80 Typical purpura of the lower leg of adult Henoch–Schönlein purpura. The lesions are palpable and inflammatory.

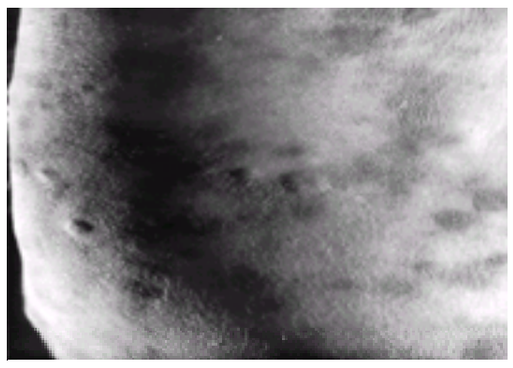

Fig. 81 Typical urticarial initial lesion of vasculitis, often proceeding to purpura and later to necrosis. However, in some types, hypocomplementaemic vasculitis, a persistent urticaria, is the only lesion.

Fig. 82 The blue-to-black discoloration of the extremities in disseminated intravascular coagulation meningococcaemia.

Vasculitis affecting the deep dermis or fat and subcutaneous tissues most commonly produces a nodule, sometimes overlaid with redness or violaceous skin. Blistering and pustulation, when they occur as manifestations of vasculitis, are usually part of a superficial polymorphic eruption in which at least some of the lesions are palpable purpura, thus distinguishing the eruption from other more monomorphic blistering diseases. These physical signs are illustrated in Fig. 83. Very heavy infiltration with eosinophils is a feature of eosinophilic cellulitis, which is sometimes a reaction to arthropod bites.

So far as the diagnosis of purpura is concerned, the traditional classification of thrombocytopenia (Fig. 84) versus non-thrombocytopenia is useful. Vasculitis is included within the latter term, and the purpura is more often palpable.

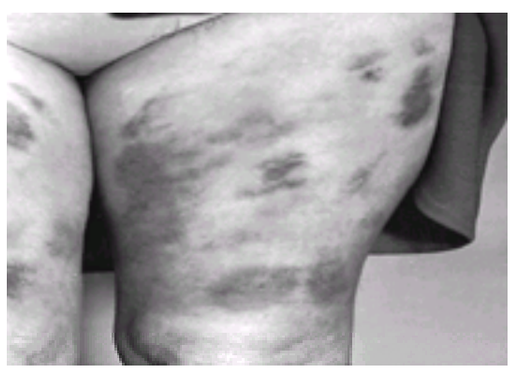

Fig. 84 Bruising or ecchymosis is most commonly due to thrombocytopenic purpura, but it is also a feature of the painful bruising syndrome.

Vasculitis in which immune complexes have played a causative role is more likely to be 'leucocytoclastic'—a term used to describe numerous disrupted neutrophils at the site of the damaged vessel (Fig. 85). It is now well recognized that a mononuclear variety of vasculitis also occurs, sometimes termed 'lymphocytic vasculitis', in which complement activation seems less significant but where upgrading of cell-wall adhesion factors is a response to cytokines and other pharmacological agents, such as histamine. It is a feature of drug eruptions and also of damage to vessels sometimes prior to the deposition of immune complexes, that is within 2 h of a cutaneous capillary fragility test. In fact macrophages rather than lymphocytes could be the more injurious infiltrate, depending on the stage of maturation and their secretion. Much of the pathology of vasculitis is that of ischaemia. Whatever the cause of the vessel damage, there is usually some impairment of blood flow. Hypoxia and infarction are quite capable of causing equally extensive pathology.

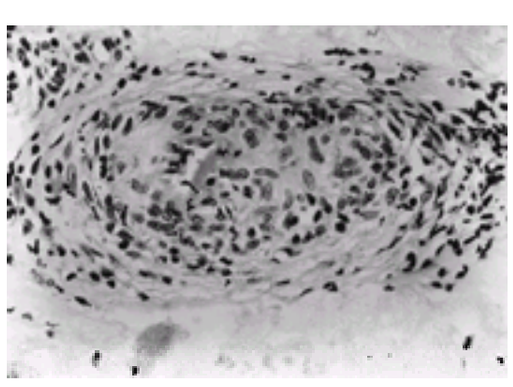

Fig. 85 A damaged vessel surrounded by broken-up neutrophils is typical of the hypocomplementaemic pattern of vasculitis.

The significance of eosinophils in some forms of vasculitis is not known and they are no guide to prognosis or therapy.

Detection of cause

The first investigation required is a full blood count to exclude thrombocytopenic purpura. Efforts should also be made to exclude septic vasculitis due to meningococcaemia and gonococcaemia. Other viral, bacterial, and fungal diseases will also need to be considered in the context of AIDS. Allergic vasculitis has a characteristic histology, showing leucocytoclasis and endothelial damage with fibrin deposition. A skin biopsy should be taken for direct immunofluoresence studies at the same time. Demonstration of IgA deposition (usually IgA1) is required for a diagnosis of Henoch–Schönlein purpura. A serum antineutrophil cytoplasmic antibody

(**ANCA**) test should be ordered. A perinuclear ANCA indicates microscopic polyarteritis and a need to monitor renal function carefully. A diffuse cytoplasmic ANCA is associated with Wegener's granulomatosis.

From the patients' point of view the most worthwhile investigation is that which results in improved management. Certain noxious agents should always be looked for in order to eliminate them therapeutically. Bacteria are the most important of these: the streptococcal sore throat is still the commonest precursor of vasculitis in children, and otitis media, dental caries, cystitis, and sinusitis occasionally play a role; in many countries tuberculosis or leprosy is the commonest cause; bacterial endocarditis and meningococcal septicaemia are often missed. Other treatable infections occasionally causing vasculitis are syphilis, neisseria, rickettsiae, and mycoplasma. Although viruses cannot usually be eliminated, any history of a recent 'flu-like illness or vaccination may be relevant. Hepatitis B virus is a well-recognized cause.

Screening for connective tissue disease with an ANA and rheumatoid factor is useful. Lupus erythematosus and rheumatoid arthritis are common causes of immune-complex vasculitis. But, as mentioned above, it is the exposure of antigen so that its complexing can activate complement that is important, not the complex itself.

Drugs such as penicillin and sulphonamides, allopurinol, amiodarone, streptokinase, thiazides, and warfarin have often been incriminated, but many other drugs appear less likely to be allergens than modifiers of the host response. Similarly, some foods may act as allergens, while others have a more obscure enhancing action. Enquiries should be directed towards headache pills, throat lozenges, purgatives, health foods, any medicine given for a specific illness, and any recent intake of food or drugs to which the patient is known to be sensitive. Questions should also be asked about recent vaccination, radiological investigation with contrast medium, or radiotherapy.

Cold agglutinins, cold fibrinogens, and cryoglobulins should also be looked for. It is now clear that many patients have cold-precipitated immune complexes. Perhaps even more have soluble immune complexes that are localized by blood stasis due to cold, pressure, vasoconstriction, or prior inflammation. The number of patients in whom an antigen has been isolated is small, and in still fewer has it been possible to eliminate the antigen, except in the case of bacteria sensitive to antibiotics. It is not particularly helpful to find a cryoglobulin; it does not alter management since, in any case, every patient should be kept warm. One of the difficulties is that it is often the antibody to the infective agent that in its turn becomes recognized as foreign. Infection may initiate this problem but after elimination of the organism the antibody persists as an autoantigen.

Even after extensive investigations there will remain a proportion of patients in whom everything appears to be normal and no cause for the vasculitis is found. Most often this is seen in women with cold, fat legs, and some degree of venous insufficiency. In these women there is decreased resistance to the deposition of antigen–antibody complexes and decreased clearance. Sometimes even the normal load of complexes generated by a meal or a trivial viral infection overwhelm the cutaneous defences. In such patients rest and compression stockings are particularly valuable.

Factors that modify the inflammatory response

These are very numerous and include any known chronic illness, such as malnutrition, diabetes mellitus, blood disorders, rheumatoid arthritis or other forms of collagen disease, chronic respiratory disease, disorders of the bowels or liver, and hypertension. Malignancy, whether carcinoma or lymphoma, is not an unusual factor and recent surgery, pregnancy, and unusual anxiety are also included in this list.

The mechanisms involved include coagulation and thrombosis, and, since these are treatable, a full blood count should always include a platelet count and other simple relevant tests, especially estimation of prothrombin time, fibrinogen titre, and fibrin degradation products.

Prognosis

The difficulty of naming a constellation of physical signs may force the physician to produce labels traditionally linked to a poor prognosis. One example is the term 'polyarteritis nodosa', another 'Wegener's granulomatosis'.

For all patterns it is useful to use the term 'vasculitis' supplemented by the terms 'limited' or 'local', implying a mild process affecting one locale or organ, and 'complicated' meaning severe and affecting many organs. Such adjectives, by describing the severity of the disease, give a lead to its management.

Management

Avoid all further injury

This allows healing to take place so preventing further damage to already inflamed tissue. Rest is essential for all cases of acute inflammation, but blood stasis should be counteracted by adequate elevation and movement of the limbs. Cold and direct sunlight should also be avoided since both injure the skin and affect blood flow. Women's legs are particularly at risk, depending on the fashion for long or short skirts or trousers.

Scratching, pinching, pressure, and constriction of the skin by ill-fitting clothing or bandages should be discouraged. Patients lying in bed will develop vasculitis on the buttocks, elbows, and over the greater trochanter unless they shift their position every few minutes. Venepuncture sites become inflamed in some forms of vasculitis, particularly in Behçet's disease and pyoderma gangrenosum, and also in severe generalized leucocytoclastic angiitis.

Eliminate circulating noxious agents, especially if they are antigens

Vasculitis following a severe streptococcal sore throat, meningococcal or gonococcal septicaemia, or tuberculosis should be treated with the appropriate antibiotics. Foci of infection, once so popular, are now too rarely thought of; when found they need to be eliminated, sometimes even by surgery. Certain bacterial diseases, such as bacterial endocarditis or leprosy, are not easily eliminated and require prolonged supervision. Although viral diseases are increasingly incriminated, there is no satisfactory way of dealing with them as yet. Immune-complex disease, sometimes as a manifestation of rheumatoid arthritis or lupus erythematosus but more often having no particular association, is now the most often suspected cause of vasculitis. Usually there are no specific measures for dealing with the problem, but immune complexes become less damaging if the factors localizing them are eliminated. Plasmapheresis is practised by a few specialized units. Drugs and food thought to be responsible can be omitted.

Provide specific treatment

Acute short-lasting itchy weals often respond to antihistamines. Acute tender swelling due to progressive tissue oedema may need to be treated with steroids. Painful swollen joints, acute optic neuritis, temporal (giant cell) arteritis, erythema nodosum, tender persistent weals, and tense painful swellings at the edge of pyoderma gangrenosum all usually respond to corticosteroids.

Fulminant vasculitis affecting more than one organ and brought about by a known trigger (allergens, drugs), should be covered by steroids once the cause has been eliminated. Immunosuppressive drugs such as azathioprine, cyclophosphamide, and methotrexate are used as a last resort in persistent chronic vasculitis, but they are of doubtful value except in granulomatous forms affecting the lung where they are the treatment of choice. Necrosis and gangrene are usually due to ischaemia. While inflammation alone may account for this in small vessels supplying superficial lesions, hypertension, coagulation, and thrombosis, as well as the cause of cardiac or peripheral vascular disease in general, sometimes underlie large areas of necrosis. The causes of vasculitis are also, for the most part, the causes of local or disseminated intravascular coagulation. Heparin is probably the drug of choice when fibrinogen, platelets, and prothrombin have been consumed and the levels of fibrin degradation products are raised. It is probably the most useful anticoagulant in malignant disease. Aspirin's anti-inflammatory effect is well known and is particularly effective when platelet aggregation is suspected. Dapsone, enhancers of fibrinolysis, and potassium iodide have had their successes and failures in recurrent nodular forms of vasculitis. Good management includes giving advice on smoking cessation, oral contraception, high blood pressure, and hyperlipidaemia. The prognosis depends on the presence of complications: particularly important are those affecting the eyes, central nervous system, and kidneys. Examination of the eyes for papilloedema and of the urine for red cells, protein, and casts is imperative.

Other vasculitides

Erythema nodosum

It is convenient to include erythema nodosum under the heading 'Vasculitis', though some still prefer to call it panniculitis. There is injury to small blood vessels in the deep dermis and subcutaneous tissue. However, primary injury to the blood vessels from a noxious agent, such as soluble immune complexes, is difficult to prove. Erythema nodosum is characterized by tender red swellings on the front of the shins (Fig. 86) and often also on the thighs and forearms. Bruising is common, but necrosis, scarring, and atrophy of the tissues are not features.

Fig. 86 Tender erythematous swelling on the front of the legs with ill-defined borders is characteristic of erythema nodosum.

Erythema nodosum is a reaction pattern to infection (viral, bacterial, and mycotic) and sometimes to drugs. Neoplasia, pregnancy, and sarcoidosis are other causes. The causes are listed in Table 18. By far the commonest is a streptococcal sore throat. Sarcoidosis and tuberculosis are common causes where the incidence of these diseases is high. Ulcerative colitis and Crohn's disease are common associations seen in teaching hospital practice. Worldwide, erythema nodosum is commonly due to lepromatous leprosy. This is a widespread and often very persistent reaction to local antigen and is not typical of erythema nodosum in general. It may become pustular and necrotic. Erythema nodosum is often preceded by or accompanied by fever, malaise, fatigue, loss of weight, and arthralgia. Although it sometimes resolves in 2 to 3 weeks, the presence of persistent and recurrent forms over several months may suggest an alternative diagnosis. It is important not to label the disease as polyarteritis nodosa or rheumatic fever, for instance, merely because it is persistent and the patient is ill for several months, or the ESR is unusually high. The number, size, and chronicity of the lesions is variable. They can be few and as large as a hand, or multiple and the size of a thumbnail. They can be acute, tender, and last only a few days; or they can be chronic, less tender, and migratory, tending to heal in the centre and spread peripherally as a swollen ring. The more chronic lesions are less red and may be violaceous, or any of the colours of a resolving bruise. The front of the leg is a site of poor lymphatic drainage where foreign protein and bacteria are only slowly removed, especially in the deep dermis and adipose tissue. The underlying tibia splints the overlying tissue so that massage of the lymphatics is reduced. Pretibial cellulitis, pretibial myxoedema, and erythema nodosum have similar localizing factors explaining their pathogenesis.

Investigations

Investigations are required to confirm the diagnosis and extent of disease, to look for a cause, and to ensure the safety of treatment. A skin biopsy will show a lobular panniculitis. Occasionally, granulomas may be seen in patients with sarcoidosis, thus enabling a specific diagnosis to be made. A chase for the source of infection should include a history of possible contacts at home and abroad, human or animal. An anti-DNAase B will demonstrate a recent streptococcal infection. A Mantoux test may have a place, but intradermal testing for sarcoidosis is no longer available. Chest radiography is essential and the most useful for a diagnosis of sarcoid, tuberculosis, or mycoplasma pneumonia. Exclusion of tuberculosis by chest radiography is also useful prior to corticosteroid treatment. A fall in the ESR, which is often initially above 100 mm/h, is a useful guide to complete recovery.

Treatment

This is one of the diseases in which ultimate recovery is to be expected. While for the first 2 to 3 weeks it is possible to keep the patient at rest and to prescribe acetylsalicylic acid, the difficult period is often several weeks after the initial illness when the patient has to be mobilized. Firm support bandages or stockings give some relief for persistent aching and swelling legs. Steroids reduce swelling and fever but do not affect the length of the illness.

Pyoderma gangrenosum

As the name implies this is a necrosis of the tissues, often with a heavy neutrophilic infiltrate; but it is not primarily an infection, rather it is a reaction pattern in which venous and capillary engorgement, haemorrhage, and coagulation feature prominently. The exact pathogenesis is uncertain. In many cases there is an associated depression of the immune system demonstrable by *in vitro* and clinical tests. Failure of macrophages to respond to tissue injury or to clear noxious agents is another feature. Its associations are an important guide to its possible causation, these include ulcerative colitis, Crohn's disease (particularly of the colon), rheumatoid arthritis, seronegative arthritis with paraproteinaemia, Wegener's granulomatosis, and plasma-cell dyscrasias including myeloma. A bullous variety is associated with leukaemia, primary thrombocythaemia, and with myelofibrosis. Nevertheless, up to half of the cases seen in dermatology clinics have no significant association. The clinical features are initially varied, but all ultimately become turgid and ulcerate. These features include a tender red or blue nodule suggestive of erythema nodosum, vesico pustules, or an acneiform folliculitis. The swollen red or blue edge is often acutely tender; blistering may be considerable, especially in the leukaemic variety. The necrosis follows no particular pattern and, like a carbuncle, may have multiple centres. It is usually undermined, and exuberant granulation tissue sprouts from the base of the ulcer. Although the calves, thighs, buttocks, abdomen, and face are favoured, no site is immune.

There is considerable toxicity associated with the acute varieties. Dermatologists see the chronic variety, which is not obviously associated with underlying disease and in which the general health of the patient is not impaired. The ulcerated lesions are not necessarily tender but they are irregular and persistent, often for years. Dermatitis artefacta is often suspected, and the personality of the patient disabled for many months may be consequently affected and encourage the suspicion.

Synergistic gangrene is one cause of very similar acute pathology. Unlike pyoderma gangrenosum, which is often multiple, synergistic gangrene is more clearly associated with a recent wound (for example, an operation on the gastrointestinal tract), and the area of gangrene is acute, solitary, and an extension of the wound. Aerobic and anaerobic culture should be performed in any form of pyoderma gangrenosum for amoebiasis, tuberculosis, buruli ulcer, and deep fungus infections such as nocardiosis or blastomycosis.,

The treatment of choice is oral high-dose corticosteroids. The management of underlying diseases, such as ulcerative colitis or leukaemia, is essential. Any suspicion of an infective causation such as amoebiasis requires the appropriate investigations and treatment. For cases responding poorly to steroids, dapsone 100 mg daily or clofazimine is worth a try. Colchicine, cyclophosphamide, and ciclosporin also have their advocates. Various subacute presentations respond to locally applied steroids by inoculation or under an occlusive dressing.

Behçet's disease

See Chapter 18.10.5 for further discussion.

Vesicoblistering diseases

A vesicle is an elevated circumscribed lesion, usually no larger than 0.5 cm in diameter, filled with serum and sometimes blood and pus. Above this size a vesicle is called a bulla or blister.

Predisposing factors

These include congenital diseases such as epidermolysis bullosa and metabolic disorders such as porphyria.

Causes

Friction and minor knocks can produce blisters in the predisposed person or at sites unaccustomed to wear and tear. The hands and feet are most often affected. Friction is increased by damp, sweating skin.

Ischaemia

Prolonged pressure obliterating the blood supply for more than 2 h causes damage to the smooth muscle of small arterioles and to underlying fat. The epidermis can survive more than 6 h of ischaemia, and much longer periods may be survived in cool skin with a decreased metabolism. Unconscious patients or those with sensory loss, especially from barbiturate poisoning, are particularly vulnerable, but most cases occur from peripheral vascular disease with acute interference of blood supply.

Acute sweat-pore occlusion

This occurs especially with fever or in hot climates. Numerous small transparent vesicles are seen, especially in the flexures or in parts of the body in which the stratum corneum is unduly thick. In the fingers or feet this is called pompholyx.

Burns

Burns can occur from cold (Fig. 87), as in frostbite, and by cryotherapy, heat, or ultraviolet irradiation (photosensitivity from plants, porphyria, or pellagra) (Fig. 88). Dermatitis artefacta is often induced by burning the skin; it is clearly self-induced but usually denied, and is often of a bizarre pattern. Cigarette burns are amongst the commonest induced lesions.

Fig. 87 Urticarial lesions, in this case due to cold; they often blister, especially on the lower legs.

Fig. 88 Blistering on the front of the neck due to an ultraviolet light burn. Self-induced by a home lamp.

Chemicals

These may be toxic, for example mustard gas and cantharidin. Sometimes an allergic dermatitis from contact with chemicals also produces vesicles due to separation of the epidermal cells by inflammatory oedema. Plant dermatitis is amongst the most common of causes; for example, from the primula in Europe and poison ivy in the United States.

Fixed drug eruptions

These can cause erythema and blistering and appear and reappear at the same site whenever the causative drug is ingested; usually itching occurs within 6 h of ingestion.

Infections

(See also under the appropriate chapters in Section 7.)

Viral disorders including herpes simplex (Fig. 89), herpes zoster, chickenpox, and smallpox, or bacterial diseases most commonly cause blisters, particularly *Staphylococcus* and *Streptococcus* spp.

Fig. 89 Blisters on the cheek due to herpes simplex virus infection.

Fungus infections commonly present as blistering on the soles of the feet, and insect bites give rise to papular urticaria that often blisters on the lower legs (Fig. 90). Blisters, pruritus, and fever have been described in ornithologists bitten by ticks carried by marine birds on the Middle East coastline. Arthropods, like the brown recluse spider, give rise to necrotic blisters, while the hairy caterpillar, for example, secretes a toxin in its hairs that can produce blistering. Some infarctions can produce a vasculitis or disseminated intravascular coagulation, which may also present as vesicular or haemorrhagic blisters. Where the Cantharidine beetle is common, avoid injury to it when alighting on the skin and let it fly away.

Specific skin disorders

Erythema multiforme

This, as the name implies, can present with a variety of patterns. The classic pattern affects the hands and feet more than the trunk. Such lesions have an erythematous and coin-shaped presentation which is more intense and blistering in the centre—a target-shaped lesion (see Fig. 10(b)). Several toxic erythematous eruptions overlap with the classic pattern and sometimes the classic distribution and even the target lesions are missing. Involvement of the mucosa is common so that the mouth, eyes, and genitalia may be affected in varying degrees. Where the blistering and mucosal lesions are severe, the disease is termed the 'Stevens–Johnson syndrome' (Fig. 91). This is usually associated with high fever and sometimes also with anterior uveitis, pneumonia, renal failure, polyarthritis, or diarrhoea.

Fig. 91 Stevens–Johnson syndrome, or severe erythema multiforme, resulting in severe erosions of the mouth and conjunctivitis.

Aetiology

The commonest cause is herpes simplex virus infection. Herpes simplex DNA has been demonstrated by polymerase chain reaction (**PCR**) in around 75 per cent of cases. Other infections such as mycoplasma, orf, streptococcus, typhoid, and diphtheria may be incriminated. Drugs also cause this disorder and sulphonamides are amongst the most common. In fact, any infection and any drug can probably give rise to erythema multiforme, usually after a latent period of 1 to 3 weeks. Other causes include neoplasm and its treatment with drugs or radiotherapy, as well as certain other systemic diseases such as AIDS, rheumatoid arthritis, lupus erythematosus, and ulcerative colitis. One of the difficulties is the overlap with the other patterns of toxic erythema and their causation. The erythema of pregnancy may sometimes be called erythema multiforme.

Pathology

There is vacuolar degeneration and apoptosis of the basal cells of the epidermis; vesicles develop between the cells and the underlying basement membrane. Vasodilatation and a lymphocytic infiltrate around the upper dermal vessels are observed.

Treatment

Any known cause should be removed, and systemic steroids prescribed if the patient is very uncomfortable and toxic. Recurrent attacks should also be treated by eliminating the cause if known: for instance, treating the earliest stage of herpes simplex with aciclovir, famciclovir, or valaciclovir and avoiding triggers like bright sunlight. Viral resistance and long-term side-effects of frequent or long-term usage have not, so far, been demonstrated. In the absence of randomized clinical trials, strong debate continues as to whether steroids, plasmapheresis, ciclosporin, and intravenous immunoglobulins are helpful or harmful. Indeed, a combination of high-dose prednisolone with cyclophosphamide given for 3 days each month seems to be effective and almost without side-effects. It is certain that fluid replacement, antimicrobial management, nutritional support, local comfort, and keeping the patient warm are essential life-saving manoeuvres.

Toxic epidermal necrolysis (TEN)

This is a rare variety of erythema with acute epithelial necrosis apoptosis affecting all areas of the skin. This is sometimes called 'scalded skin syndrome' because of its clinical appearance. It is usually acute in onset and may be preceded by various patterns of toxic erythema or blistering. Pressure and shearing stresses on the skin tend to encourage the extension of the blisters. There are two varieties of the disease: the first, originally described by Ritter, is due to a staphylococcus, often phage type 71, and particularly affects children—the blistering and resulting erosions are very superficial and they are due to a split at the level of the stratum granulosum; the second is a drug reaction or a toxic consequence of malignant disease or its therapy. The entire epidermis is necrotic. European physicians have found it of epidemiological value to define TEN as more than 30 per cent of skin detachment and the Stevens–Johnson syndrome as less than 10 per cent, with an overlap syndrome of between 10 and 30 per cent.

Sometimes sulphonamides, barbiturates, phenytoin, pyrazolone derivatives, and phenolphthalein are responsible, while a number of other drugs are also blamed, albeit more rarely.

The use of human intravenous immunoglobulin containing anti-Fas antibodies, and therefore protective against apoptosis, shows promise. Pulse therapy with intravenous gammaglobulin is also under investigation, using 0.4 g/kg for 5 days.

Rarer blistering disorders

These include diseases like pemphigus, pemphigoid, and dermatitis herpetiformis. At one time these were all grouped together, for their pathogenesis has only recently become clearer. The main distinction is in the level of the blister, which determines both clinical and histological features—pemphigus is an intraepidermal blister, whereas the other disorders tend to be subepidermal. Resulting cleavage within the dermis produces dermal inflammation, oedematous papules, infiltration with white cells, as well as bleeding into this blister. The more superficial the blister, the more erosive the appearance: the skin lesions may be red and glistening, whereas deeper dermal blisters tend to be tense and less easily broken. The type and site of immunoglobulin deposition is a further diagnostic feature.

Pemphigus

Pemphigus vulgaris

This is a blistering condition favouring the mucosa as much as the skin. Separation of epidermal cells above the basal layers of the epidermis always occurs in association with an antibody having an affinity with intercellular material in the epidermis. The separated epidermal cell is large, basophilic, and rounded and is termed an 'acantholytic cell'.

Aetiology

The pemphigus antibody will cross the placental barrier and promote neonatal blistering. It is also pathogenic *in vitro.* This antibody reacts with a specific 85-kDa protein, plakoglobin, and a 130-kDa protein in pemphigus vulgaris, but in pemphigus foliaceous (see below) it reacts with a 160-kDa protein, an extracellular epitope of desmoglein. The 130-kDa polypeptide is an epidermal cadherin. The loss of adhesion occurs because of the important adhesive role of these components in the desmosome. (See Chapter 23.2, Fig. 1.)

It is assumed to be an autoimmune disease, possibly associated with HLA-A10 and DR4, and is found more commonly in the Jewish race. Moreover, in Asia, it is one of the commonest causes of admission to a skin hospital. The more superficial variety that affects Brazilians may or may not be a separate, genetically determined reaction pattern. The antibody that binds with complement both *in vivo* and *in vitro* is specific for an, as yet unidentified, intercellular material which activates proteases that lyse intercellular adhesive materials. Several investigators have found that the antibody can frequently cause intraepithelial clefting *in vitro* in human, rabbit, and monkey epithelium. There is an association with thymoma as well as with lymphoma and carcinoma. Not surprisingly, therefore, it occurs with lupus erythematosus and myasthenia gravis.

Penicillamine has been responsible for the development of pemphigus in about 9 per cent of patients treated for rheumatoid arthritis. Captopril and rifampicin as well as meprobamate have also been incriminated.

Clinical features

Erosions of the mucosa of the mouth are the initial problem in more than half the cases. The erosions are often misdiagnosed as mouth ulcers, but close examination reveals a friable mucosa with no well-defined aphthous ulcers. Actual blisters may be missed because they are so quickly eroded. On the skin, the superficial nature of the blisters also determines that the principal lesion is a more painful erosion and the flaccid blisters quickly burst. The base is red and bleeds easily. The epidermis at the edge of the blister is easily dislodged by sliding pressure (Nikolsky sign). There are many reports of clinical and histological overlap with pemphigus foliaceous or pemphigoid. In all such cases pemphigus vulgaris proves to be the final diagnosis.

Treatment

Corticosteroids are lifesaving; without them the disease is one of the most dangerous in dermatology. Very high dosage is required; prednisolone 120 mg daily is a common starting dose and failure to control the eruption within a week merits doubling of even this high dose. As soon as there are no new blisters the steroids are reduced by large increments about every 3 days. Withdrawal is more gradual below 30 mg daily. Most practitioners now add azathioprine, methotrexate, or cyclophosphamide as a steroid-sparing immunosuppressant.

In contrast to the presteroid era, cure now seems possible and many patients are off all treatment after 2 years. However, the side-effects of the therapy are considerable. Death from gastrointestinal haemorrhage is not infrequent. Thromboembolic disease is probably a consequence of the disease as much as the therapy. Steroid-induced osteoporosis with consequent vertebral collapse is a frequent and irreversible side-effect. Bacterial infection of the eroded skin is inevitable and septicaemia is common. The sore mouth and eroded skin require expert nursing—dressings tend to stick to the skin and their removal causes further skin damage. Oral fluids should not be strongly osmotic and soft diets should not include particles that lodge under blister roofs or in crevices.

Pemphigus vegetans

Pemphigus vegetans is a reaction to the erosions in which the repairing epidermis becomes hypertrophic and the dermis is granulomatous; small pustules surround the vegetations. This disease is common in the axillae and groin and the angles of the mouth and nose (Fig. 92). It may be encouraged by steroid dependency.

Fig. 92 Pemphigus vegetans showing the typical granulomatous hypertrophy underlying erosions at the angles of the mouth.

Pemphigus foliaceous

This is a more benign variant of pemphigus in which the blisters are more superficial. The bullae are subcorneal and scaling and crusting may be a principal feature (Fig. 93). The face and upper trunk are most often affected. Localized forms may look more like seborrhoeic warts because of their chronicity and definition. Oral lesions are unusual. Antibodies against intercellular epithelial material are present as in pemphigus vulgaris, but basement membrane antibody and antinuclear antibody are also frequently observed. An association exists with lupus erythematosus and with thymoma and myasthenia gravis.

Fig. 93 Pemphigus foliaceous blisters, so superficial that they merely look like crusting of the erosions. In this case it would have to be distinguished from an intertrigo and secondary monilial infection.

Fogo selvagem

This is a form of pemphigus foliaceous is commonly seen in people working in the rural peanut farms of Brazil. Many members of one family may be affected. Progression to a generalized erythroderma is usual and the mortality is almost 50 per cent, due as much to treatment as from the disease. An immunological reaction to an insect vector has been proposed, based on the study of the black-fly bites and the hypothesis of crossreactivity between the epidermal antigens and the antigen of the fly. Topical steroids may be preferable to high-dose systemic therapy in those who cannot be closely supervised.

Pemphigoid

The bullae in pemphigoid are subepidermal and acantholysis is not a feature. About 80 per cent of patients are over 60 years of age. Pemphigoid is about twice as common as pemphigus. There is a specific antibody (usually IgG) for the basement membrane zone of the epidermis and this is present in about 70 per cent of patients. Complement is bound *in vivo.* The basement membrane remains in the floor of the bullae in most cases. Two large epidermal polypeptides are the major antigenic target of bullous pemphigoid (**BP**) antibodies. The *BP230* gene is localized to the short arm of chromosome 6 and the *BP180* gene to the long arm of chromosome 10. Both protein products of these genes are components of the hemidesmosome. BP180 is a transmembrane glycoprotein with an external terminal ectodomain consisting of collagen triple-helical domains, which bind keratin to the hemidesmosome. (See Chapter 23.2, Fig. 1.)

Clinical features

Initial features of pemphigoid are often non-specific and confusing. It can be eczematous or urticarial. The lesions often begin around a site of damage such as a leg ulcer or burn. After 2 or 3 weeks blisters may erupt abruptly. They favour the flexures and are tense and dome-shaped, often containing blood. Small blisters in the mouth are rare and tend not to erode as in pemphigus. Patients with pemphigoid are distressed by itching, and oedema of the skin may be troublesome, but their general health is usually unaffected.

Treatment

The treatment of choice is prednisolone, 60 to 80 mg daily, until there are no new blisters. Since morbidity in the elderly is great, azathioprine, methotrexate, dapsone, minocycline, or cyclophosphamide may be used to allow a lower maintenance dose of the steroid. Osteoporosis, gastric ulceration, and diabetes mellitus are particularly common complications of steroid therapy. However, complete remission after 1 year is common.

Specific disorders

Cicatricial pemphigoid

Also called 'benign mucosal pemphigoid', the cause of this disorder is unknown, but its immunology includes autoantibodies to an 180-kDa protein. Although mortality is low, cicatricial pemphigoid is responsible for great discomfort. It affects older adults, and the subepidermal bullae favour the mucosa of the mouth, conjunctiva, and the perineal orifices. The base of the lesions are heavily infiltrated with lymphocytes and plasma cells and there is eventual fibrosis. Those adhesions occurring between the bulbar and palpebral conjunctiva result in eventual shrinkage, and entropion is followed by blindness. The skin is less often involved and the sparse lesions often heal by scarring. The scalp is more often affected than other sites.

Treatment

No treatment is very effective, but steroids and azathioprine are usually prescribed.

Dermatitis herpetiformis

This is a vesicobullous disorder associated with the granular deposition of IgA in the dermis and a usually symptomless subtotal villous atrophy of the small intestine. The IgA is believed to be derived from plasma cells in the intestine. As in coeliac disease, HLA-A8/DRW 3 is associated and may be responsible for a defective Fc receptor status. It is probable that gluten hypersensitivity results in circulating immune complexes with an affinity for material in the upper dermis (this is possibly reticulin or transaminases related to gliadin), and that the Fc-receptor dysfunction impairs the removal of the immune material by macrophages. Histology of the skin shows fibrin, neutrophils, and eosinophils in the dermal papillae.

Clinical features

The eruption is characterized by intensely itchy, grouped papular or vesicular lesions that lie on an urticarial or erythematous base. The elbows, knees, sacrum, and shoulders are favoured (see Fig. 9), with the face and scalp more commonly affected than in the case of pemphigus or pemphigoid. The itchy vesicles are quickly excoriated since this relieves the pruritus. The eruption waxes and wanes, sometimes being in remission for many months. However, for most people it remains a lifelong disorder. An increased incidence of lymphoma is well documented.

Treatment

Dapsone (100–200 mg daily) or sulphapyridine (0.5 g, three times daily) are remarkably effective and can be used as a diagnostic test since itchiness is relieved within 48 h. The maintenance dose should be titrated to suit each patient: it may be as low as 50 mg of dapsone per week. Haemolytic anaemia is common on higher dosage and especially when, in some cases, 400 mg of dapsone per day is needed to control the eruption. A gluten-free diet strictly adhered to controls some but not all disease; 70 per cent pf patients can stop taking dapsone after 2 years of such dieting.

Steroid therapy is strangely ineffective and heparin oddly effective. Inorganic arsenicals (Fowler's solution) are effective and were once very popular, and are probably justified in elderly patients much troubled by the disease and unable to tolerate dapsone or sulphapyridine.

Specific disorders

Juvenile bullous pemphigoid

This is a bullous disorder characterized by a predilection for the face and perineum. Linear IgA is deposited on the basement membrane of the epidermis. It is neither associated with enteropathy nor with HLA-A8. The response to dapsone, sulphapyridine, or steroids is unpredictable.

Pemphigoid gestationis

This differs from the common toxic erythema of pregnancy in having large blisters, often periumbilical, beginning as a degeneration of the epidermal cells. Pemphigoid gestationis is associated with HLA-B8/DR3 and an IgG1 autoantibody that avidly binds C3. Thus, it is a blister above the basement membrane, believed to be due to a specific antibody to the basal cell; the antibody is present in umbilical cord blood and binds with a 180-kDa glycoprotein in the basement membrane of the amnion. This disorder occurs during or immediately after pregnancy and usually ceases fairly abruptly within weeks of parturition. It recurs in subsequent pregnancies and as an effect of oral contraceptives.

Other causes of blistering

Lichen planus and lichen sclerosis et atrophicus both rarely blister. Bullous disease and malignancy are a debated association, since so much bullous disease occurs in an age group in which malignancy is common. However, individual case histories of uncontrollable bullous disease with atypical immunofluorescence are

impressive.

Trophoneurotic blisters are another debated association. Unconscious patients seem predisposed to produce these subepidermal blisters even at sites not affected by pressure or shearing forces. This is an important hazard often causing unjustified accusation of mismanagement in the nursing care of such patients.

Diabetes mellitus is a cause of intraepidermal blisters without immunofluorescent material and showing no acantholysis (Fig. 94).

Fig. 94 A haemorrhagic blister on the foot in a patient with diabetes mellitus.

Abnormal vascularity of the skin: angioma and telangiectasia

Patterns of blood vessel development that are inappropriate for the needs of the skin or for thermoregulation include both overgrowth and atrophy. An excess of capillary and venular vessels is a characteristic of wound healing and of many hyperproliferative conditions of the skin (for example, psoriasis). These usually present as redness and individual vessels cannot be seen by the naked eye. Proliferation is still more extreme in strawberry haemangioma, granuloma telangiectaticum, also known as pyogenic granuloma, and in certain malignancies such as Kaposi's sarcoma and angioendothelioma.

On the other hand, telangiectasia is characterized by the dilatation of individual capillaries or venules so that they are visible to the naked eye. There is little evidence that the endothelial cell is at fault and it is more likely that the basic defect is an atrophy or loss of supporting tissue.

Proliferative vasculature is more unstable than that observed in telangiectasia and the natural history is for it to resolve, often completely. Vessels and wounds or angiomas are vulnerable, and the growing phase may be associated with necrosis as a result of thrombosis secondary to a surface injury of the skin. On the other hand, telangiectasia has no tendency to thrombose and the overlying skin rarely ulcerates. The natural history of such dilated vessels is to persist until extreme old age when they may be partially absorbed.

The pulsed dye laser has revolutionized the management of vascular lesions of the skin, but it has also shown the need for a multidisciplinary approach to the diagnosis of cellularity, depth of lesion, haemodynamic changes, and involvement of organs other than the skin. The perfect result is elusive.

Naevi

Strawberry naevi

Although these are almost never present at birth, they may be preceded by a small area of blanching observed at birth. From a few days after birth the lesion consists of rapidly proliferating nests of granulation tissue. After a few weeks the rate of growth becomes less rapid and some vessels become dilated and cavernous. A stable period of no growth often occurs between about 9 months and about 1 year, after which gradual absorption by fibrosis is to be expected. Management consists of reassuring the parents and emphasizing its satisfactory natural resolution (Fig. 95).

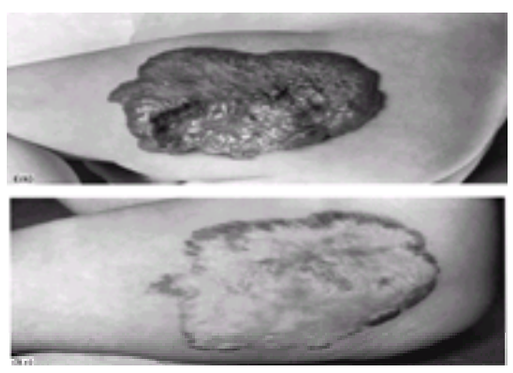

Fig. 95 Strawberry' naevus. (a) A proliferative but benign neoplasia which, after a rapid phase of new growth, stabilizes and eventually regresses. The lesions often ulcerate if traumatized. In this case, ulceration has hastened resolution but there is more residual scarring (b) than usual.

However, there are exceptions to this policy of non-intervention. For example, plastic surgery may be advised where involvement of the eyelid interferes with sight. Some large haemangiomas sequester platelets, thus giving rise to a bleeding tendency (the Kassabach–Merrick syndrome). High-dose steroids (3 mg/kg) are lifesaving. On withdrawal, rebound overgrowth may be observed, justifying a second or third course of treatment. Ulceration of the haemangioma is common, especially in the nappy area and when there is a primary irritant rash. Bleeding is easily controlled by light pressure. The ulceration often accelerates resolution.

Sometimes haemangiomas have a deep element in which arteriovenous shunts are a complication. Interference with underlying structures is not common, but joint involvement warrants surgical advice and management.

The treatment of haemangiomas has included radiotherapy; more recently, systemic steroids, interferon-a, pressure pads, excision, and, currently, embolization. The latter requires angiographic control and siting of sclerosing adhesives at the appropriate site. The delineation of vascular endothelial growth factors and their receptors is opening up new approaches to therapy.

Port wine naevi

This is a pattern of vascular birthmark present at birth and is usually segmental. It is unwise to make a prognosis at birth because pale naevi and segmental patterns of erythema may look similar and often fade. The majority of port wine naevi persist for life. Arteriovenous shunts and gravitational stasis often cause some increase in the vasculature during adult life. A pale plaque of macular telangiectasia in the nape of the neck is present in the majority of normal babies, and persists in more than half of those affected.

Variants of port wine naevi affecting deeper vasculature range from the Klippel–Trenaunay disorder causing enlargement of the limb, to a reticulate and more atrophic pattern sometimes associated with shortening of the limb. Asymmetrical gigantisms and disturbances of pigment are associated with some developmental patterns of

widespread segmental telangiectasia. Telangiectatic vessels on the upper face and nose may be associated with eye or brain defects, presenting as glaucoma or epilepsy.

Telangiectasia

Telangiectases are enduring dilatations of blood vessels. They are usually less than 1 mm in length and may be point-like or punctate, linear, spider, or stellate, forming flat, square, oblong, or oval plaques, or mat-like with an eccentric punctum. They blanch completely when compressed. Telangiectasia is not new-vessel formation—indeed, new vessels in wounds are not unduly dilated.

Telangiectases are probably always secondary to mesenchymal connective tissue dysplasias. However, they can be congenital and naevoid, acquired and genetic (in other words, familial or inherited), as well as being secondary to 'collagen' diseases such as lupus erythematosus, scleroderma, or dermatomyositis, or the result of radiation damage.

All dilatations of small vessels are made worse by blushing, as is seen in rosacea, carcinoid diseases, and oestrogen and related hormonal imbalances (for example, in pregnancy or liver disease). They are also made worse by a loss of supporting tissue, as in steroid atrophy, solar elastosis, ageing (Fig. 96), and Cushing's syndrome.

Fig. 96 Ageing is accentuated in light-exposed skins. One feature of ageing is the poor collagen support of skin vasculature. Telangiectasia is a common consequence.

Telangiectasia is often associated with increased melanin pigmentation and brown spots may be predominant, even in hereditary haemorrhagic telangiectasia; but poikiloderma is a typical example of atrophy, telangiectasia, and pigmentation. Some telangiectasias may be insufficiently dilated to be recognized by the naked eye. However, if they involve most of the vessels in an affected area they may appear as a persistent erythema (for instance, the red cheeks of young children), or as capillary naevi affecting the eyelids, nape of the neck, or forehead, known as salmon patches or stork bites. The erythema may be pale pink or deep purple in colour. The darker the lesion, the more likely the dilated vessels will be inhomogeneous, making some visible to the naked eye. Naevoid lesions usually affect well-defined segments of the skin, though not necessarily dermatomal or unilateral in pattern (Fig. 97). The best known are naevi affecting the trigeminal nerve (Sturge–Weber syndrome) or sometimes an entire limb.

Fig. 97 Segmental telangiectasia of punctate-spider naevoid type.

Diffuse polymorphic patterns that develop in childhood or in young adults favour exposed areas such as the face and forearms, probably because sunlight exaggerates connective tissue dysplasia. However, haemodynamic factors such as gravitational stasis of the venous system also play a part, and the distribution of spider naevi may depend on drainage into the superior vena cava. Gravitational stasis particularly determines the patterns of stellate and arborizing telangiectasia on the legs. While 5 per cent of the population has between two and ten telangiectases on the lips, fingers, palms, and soles, grosser patterns of telangiectasia in disease may involve these sites. Dermatomal or unilateral patterns are rare. A high incidence, up to 40 per cent, of telangiectasia affecting the trunk is described in aluminium workers, apparently associated with the electrolytic processes used in the industry.

Although diffuse and acquired patterns of telangiectasia are commonly familial, sporadic cases account for about 20 per cent, even in the well-known hereditary haemorrhagic telangiectasia. A benign variety of this disease is not associated with severe bleeding and is also probably dominantly inherited. Telangiectasia confined to the lips may also present a dominant pattern of inheritance. However, no large-scale study has been undertaken to rule out polygenic inheritance in any of these disorders. The haemorrhagic diathesis has been recorded as a dominant gene, with many large pedigrees described; but even so, 10 per cent of probands with telangiectasia do not bleed. Severe epistaxis and severe bleeding after tooth extraction or cuts and even heavy menstrual bleeding are characteristic of hereditary haemorrhagic telangiectasia. Minor but frequent nose bleeds are common with even the benign forms. Arteriovenous shunts are described commonly in association with hereditary haemorrhagic telangiectasia. These result in pulsating nodules of the skin, which may have severe consequences in the lung or brain. Arteriovenous shunts are occasionally seen in the non-haemorrhagic telangiectatic forms of the disease.

Histology

The dilated vessels are unremarkable. However, special studies have helped to show that the vessels are venules (that is, alkaline phosphatase-negative) and that they secrete generous amounts of plasminogen activator. The supporting tissue and overlying epidermis is usually atrophic.

Treatment

Although telangiectases are easy to camouflage with 'covermark' types of preparations, advice may need to be given on its application with respect to the use of cream and powder, blends, and matching of skin colour.

Telangiectases when small and localized can be destroyed by cryotherapy, cautery, electrolysis, or laser therapy; the latter can be endoscopic. Since the laser specifically burns haemoglobin, it is therefore more successful in the treatment of the larger blood-containing dilatations. Sclerotherapy is also possible.

Bleeding should first be treated by the simple first-aid measure of elevation and local pressure. Cautery is most effective only on a dry blanched area controlled by

compression. Patients can be taught to inflate a lubricated finger cot tied over the end of a small catheter to immediately control severe epistaxis.

Oestrogen therapy

Oestrogen therapy is sometimes advocated: for instance, ethinyloestradiol 0.25 mg daily, increased to 0.5 mg per day at the end of 4 weeks if epistaxis continues. However, its effectiveness has not been proved in controlled trials.

Percutaneous embolization

This is increasingly used to close unwanted vasculature. However, it requires careful angiographic control and skilled surgeons, and has not overcome the problem of rapid recanalization and opening up of collaterals. Nevertheless, percutaneous embolization is the treatment of choice for severe uncontrolled bleeding from arteriovenous shunts.

Facial erythema (flushing)

In temperate climates the weather-beaten face of the farmer or fisherman is largely a reflection of exposure to cold, as are the rosy cheeks and 'shiny morning face' of the schoolchild. The 'butterfly' area of the cheeks and bridge of the nose pick up hot and cold thermal irradiation. Marked telangiectasia of the cheeks in adults is a common consequence of rosy cheeks in childhood. It may be associated with thickening of the subcutaneous tissues.

Rosacea is usually associated with acneiform pustulation and lymphoedema. For unknown reasons, a keratitis is sometimes associated. Some physicians advocate the elimination of *Helicobacter pylori* with a 1-week course of 20 mg omeprazole twice daily, 400 mg metronidazole twice daily, and 250 mg clarithromycin twice daily. Tetracycline, 250 mg twice daily, or metronidazole gel and lotion are effective therapies for rosacea. A somewhat similar appearance may be seen in sarcoid, especially the lupus pernio variety in which a diffuse granuloma underlies dilated blood vessels filled with slow-flowing blue blood. Mitral or pulmonary stenosis also causes a persistent malar flush. In discoid lupus erythematosus, telangiectasia (Fig. 98(a)) is accompanied by atrophy (Fig. 98(b)), often with well-defined margins and follicular plugging. The borders of rosacea and sarcoid lesions tend to be more diffuse. Asymmetry can be a feature of all disorders.

Fig. 98 In lupus erythematosus—chronic discoid type—(a) telangiectasia and (b) erythema is common. There is follicular plugging and destruction of the skin, resulting in pigment loss and scarring.

Telangiectasia due to other collagen diseases varies from the redness and oedema of the orbit, characteristic of dermatomyositis (Fig. 99), to erythema of the backs of the hands and nail-folds, with persistent erythema such that vessels become increasingly inhomogeneous, and quite large dilated forms may be observed.

Fig. 99 Bright erythema and oedema of the face, especially periorbitally, is characteristic of dermatomyositis.

In scleroderma, especially of the adult acrosclerotic variety, telangiectasia in the form of flat macules, often with square or oblong shapes and fairly well-defined margins, are characteristic. Systemic sclerosis of the face usually causes stiffness and tethering of the skin to deeper structures over the nose and loss of suppleness in the perioral skin. There is, of course, the associated Raynaud's phenomenon and digital ischaemia.

A plethoric complexion is a feature of superior vena cava obstruction, and also of polycythaemia rubra vera and Cushing's syndrome.

Flushing of the face is a common complaint, particularly in the young who may suffer from transient blushing; in older age groups it is characterized by persistent rosacea. Carcinoid should be thought of when there is a prolonged blush associated with a bounding pulse, asthma, abdominal pain, and diarrhoea. Frequent applications of steroids to the face for atopic or seborrhoeic eczema produces gross diffuse telangiectasia; rebound eczematous changes often occur on withdrawal (Fig. 100) when the irritation may be severe. Treatment of this condition is the complete withdrawal of all steroids for 3 to 4 weeks, after which the condition tends to settle.

Fig. 100 Perioral dermatitis is a common consequence of the application of fluorinated steroids to the face.

Neurogenic flushing is accompanied by sweating, but this can be dampened with b-blockers. Local heating of the oropharynx is one physiological cause, for which sipping iced water gives relief. Nocturnal overheating causing restlessness and facial flushing; rubbing is another contributory factor.

Disorders of collagen and elastic tissue

The metabolism and diseases of collagen are described in Section 19. The fundamental defects are in its chemical structure, its crosslinkage between fibres, and its distribution and quantity.

There are at least 18 collagen types, hence there are many genetic diseases.

Signs of collagen defects

These include:

- Diminished skin thickness and increased transparency mean that deeper structures such as veins and nerves are visible and the sclerae are blue.
- Diminished resistance to shear allows the skin to split and tear, sometimes even without surface breaks. Purpura is usually associated and healing results in white stellate scars. Cutaneous striae are another pattern of skin stretching, with separation in this case. Diminished resistance of the skin is a feature of age; osteoporosis and rheumatoid arthritis are other recognized associations. Steroids are responsible for both stellate scars and cutaneous striae, which is the case whether they are endogenously produced, as in Cushing's disease, or prescribed for other diseases. Local application of steroid cream is probably now the commonest cause of these changes.
- Laxity is the failure of the skin to return rapidly to its former state after distortion by stretch. In some way it is caused by degeneration of elastic tissue but changes in water content and cellularity, as well as increased crosslinkage of collagen, play some part even when the total collagen is reduced.

Diseases due to defective collagen

Solar and senile elastosis

This affects white-skinned races, especially those employed in agricultural or marine work. Chronic exposure to ultraviolet radiation causes abnormal collagen having the histological staining characteristics of elastic tissue but not its properties. It is broken and aggregated and contributes to a thickened, yellow and wrinkled skin, especially on exposed areas during old age. The yellow plaques may be sharply marginated on the face. In the neck deep furrows form a rhomboidal network. The sebaceous glands and ducts are poorly supported, dilated, and patulous, forming giant comedones. On the neck the goose pimple, or plucked-bird, appearance is due to the protection provided by hair follicles shading the dermis against ultraviolet rays. Colloid milium is the abnormal production of a scleroprotein by fibroblasts giving rise to yellowish translucent papules or plaques in light-exposed skin. It may begin in childhood.

Striae

These are common but imperfectly understood; stretch is always a factor. The epidermis is thin and elastic fibres are scanty. Striae are seen on the back and thighs of adolescents during growth, especially when there has been a spurt and the child is athletic. It occurs more in girls than in boys. Striae are a feature of pregnancy and especially affect the abdomen and breasts, usually caused by excessive adrenocortical activity. Incomplete fibroblasts inhibition causes atrophy of collagen in response to glucocorticoids. When the collagen is ageing or degenerate, as follows irradiation or in diseases such as cutis laxa or the Ehlers–Danlos syndrome, striae are uncommon and may not appear even in the pregnant or those with Cushing's syndrome. Striae have also been described in those with chronic infections such as tuberculosis. They are only a diagnostic problem when they are newly formed, in which case they may appear to be weal-like and raised. Later they flatten and become bluish-red and still later, white and depressed.

Localized fibrosis, keloids, and hypertrophic scars

The connective tissue response to cutaneous injury exceeds the need for appropriate repair at that site, commonly giving rise, a few weeks later, to hypertrophic scars. If the scar continues to hypertrophy and extends beyond the limits of the injured skin site, especially after a period of 3 months after the injury, it is often then termed a keloid. Such scars tend to be more tender than hypertrophic scars. Keloids tend to be familial and are commoner in Blacks. They are rare in infancy and old age and tend to be less severe after the age of 30. Significant factors are the presence of foreign material in the wound and tension. Preferred areas are the ear lobes, chin, neck, shoulders, upper trunk, and lower legs. Keloids in their early stages may respond to strong local steroids applied locally or intralesionally. Compression therapy is sometimes helpful, as is cryotherapy in the early stages. Re-excision and radiotherapy to the edges of the wounds is now the treatment most preferred.

Pseudoxanthoma elasticum

This is a hereditary disorder of elastic tissue; of which there are four distinct types:

Dominant type I

Here, small, yellowish papules form linear or reticulate plaques, which in older persons are soft, lax, and hang in folds, they are flexually distributed, especially in the groins, axillas, and neck (Fig. 101). There is a severe degeneration of Bruch's membrane giving rise to the slate-grey, poorly defined 'angioid' streaks that form an incomplete ring or radiating lesions around the optic disc of the retina. There is early blindness. Vascular complications include intermittent claudication and coronary artery disease.

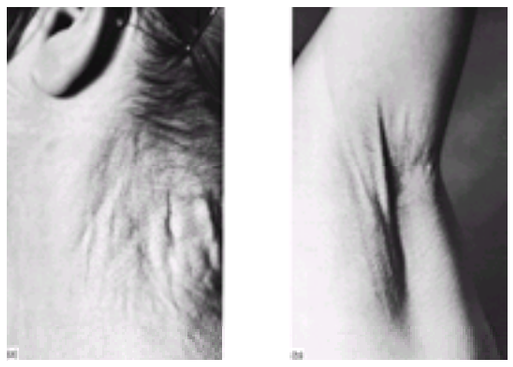

Fig. 101 (a) Yellowish papules and loss of elasticity of the skin of the neck in pseudoxanthoma elasticum. This may be a clue to gastrointestinal bleeding or even blindness. (b) Yellowish papules and loss of elasticity of the skin of the axillas in pseudoxanthoma elasticum. This may be a clue to gastrointestinal bleeding or even blindness.

Dominant type II

The small, yellowish papules are fewer and flatter than in dominant type I disease. There is increased extensibility of the skin. Vascular and retinal changes are mild.

The sclerae are blue and there may be a high arched palate and myopia.

Recessive type I

This resembles dominant type I pseudoxanthoma elasticum, but the vascular and retinal degeneration is mild. Haematemesis is especially common, and women are more often affected than men.

Recessive type II

This is a very rare form, but the skin changes are extensive and generalized. There tends to be no systemic complications. The pathology of pseudoxanthoma elasticum includes a deposition of calcium on the elastic fibres. The mid-dermal elastic tissue is fragmented and swollen.

Perforating elastoma

This is a condition of elastic fibre degeneration in the upper dermis, with a resulting foreign body reaction and extrusion through the overlying epidermis. This reaction gives rise to papules that develop a central plaque of extruded material. There is a tendency for the formation of annular and serpiginous patterns, particularly over the back and neck region. The disorder is associated with mongolism, Marfan syndrome, Ehlers–Danlos syndrome, pseudoxanthoma elasticum, and osteogenesis imperfecta.

Ehlers–Danlos syndrome: cutis hyperelastica

Ehlers–Danlos syndrome is a rare inherited disorder of connective tissue (see Chapter 19.2). The condition is usually recognized when the child begins to walk since there is hyperextensibility of the joints. Trivial cuts form gaping wounds and heal poorly. The skin feels soft and can be stretched, particularly over the knees and elbows. Arterial rupture, aortic dissection, and intestinal perforation have been described in severely affected individuals with deletions in the *COL3A1* gene on chromosome 3 affecting the length of collagen type 3.

Cutis laxa

This is a rare disease in which the skin hangs in loose folds due to the loss of elastic tissue. Severely affected individuals have associated pulmonary emphysema. Both dominant and recessive patterns of inheritance are seen (Fig. 102).

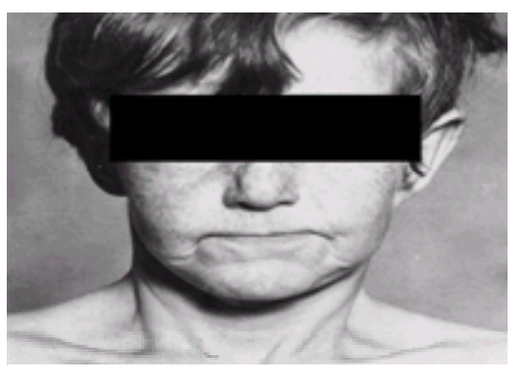

Fig. 102 Drooping of the facial skin of a 9-year-old boy is due to premature loss of elasticity. The diagnosis is cutis laxa.

Atrophy

Atrophy is characterized by thinning, loss of elasticity, loss of hair follicles, and a smooth surface to the skin. When pinched gently the skin produces fine wrinkles and may be compared to tissue paper. The upper dermal atrophy causes poor support to an atrophic vasculature and telangiectasia is often observed. At the same time there tends to be increased pigmentation within the dermis. Atrophy may be a consequence of inflammation following acute bacterial (particularly elastase-producing organisms) infection vasculitis or pancreatitis. It may be widespread, as in the chronic scarring of leprosy or onchocerciasis. Some circumscribed atrophies follow an urticarial vasculitic process, probably caused by an infection that destroys elastic tissue. Perifollicular atrophy or postacne atrophy is similarly due to elastase-producing strains of staphylococci. Syphilis is another cause of destruction of elastic tissue. Non-infectious causes include lupus erythematosus and localized scleroderma with its variants.

Poikiloderma

The combination of pigmentation, telangiectasia, and atrophy is known as poikiloderma (see Fig. 108), causes of which include irradiation, lymphoma, and collagen diseases such as lupus erythematosus and dermatomyositis. A congenital form is associated with light sensitivity, skin cancers, and dwarfism. It may follow lichen planus or stasis eczema. Poikiloderma is common on light-exposed areas of the neck and may be aggravated by cosmetics. It is also described in graft-versus-host diseases.

Fig. 108 Poikiloderma; atrophy, pigmentation, and telangiectasia, commonly preceding the development of lymphoma in the skin. The clinical appearance resembles radiodermatitis.

Morphea

Morphea is a localized form of scleroderma with a good prognosis for complete recovery (Fig. 103). It is not associated with any systemic disease, though subsets in Europe due to *Borrelia burgdorferi* cannot be ruled out. Occasionally a generalized form produces such tightness of the chest wall that breathing may be impaired. The generalized form of morphea also greatly restricts the limbs, and a combination of ischaemia and lymphoedema may result in ulceration of the peripheries.

Fig. 103 (a) Widespread hardness of the skin and brownish or violaceous plaques that are often atrophic are features of morphea—a localized form of scleroderma. (b) Pseudoscleroderma, identical to morphea, is also a consequence of post-thrombophlebitic fibrosis of the lower limbs.

Deep dermal and subcutaneous atrophy

The skin loses its subcutaneous or deep dermal tissue in a number of conditions. Such skin is waxy in colour and may be yellow, pigmented, or bluish with a loss of connective tissue. Deeper vessels may become more obvious, resulting in either telangiectasia or obvious cutaneous atrophy and linear stretch marks that are initially red and which sometimes protrude above the surface of the skin, but later there is always marked atrophy.

The atrophic skin may be tethered to underlying tissue or more obviously scarred. Such skin may feel hard or sclerosed (Table 19).

Other causes of deep dermal atrophy include the injection of insulin—this is commonly seen on the thighs or arms of diabetics. 'Anetoderma' is a term used for very discrete, round idiopathic losses of dermis.

Hemi- or generalized atrophy of a non-inflammatory origin is mainly of unknown aetiology. Partial lipodystrophy is associated with glomerulonephritis, hypocomplementaemia, and protease inhibitors. The Lawrence–Seip syndrome, or total lipoatrophy (with acanthosis nigricans, genital hypertrophy, resistant diabetes, and hepatomegaly), is a condition affecting infants. Atrophie blanche (Fig. 104) is an obliteration of single capillaries in the upper dermis, leading to very localized scarring, the causes of which are listed in Table 20.

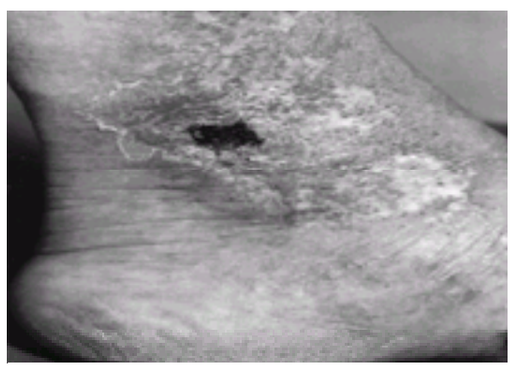

Fig. 104 Atrophie blanche is an obliteration of the capillaries in the upper dermis, causing sclerosis. Residual vessels are elongated and coiled. They are liable to thrombosis and overlying ulceration is a consequence.

Malignant disease

Infiltrations of the skin presenting as papules, *peau d'orange*, nodules, plaques, or ulcerating tumours of the dermis with destruction of overlying epidermis are a common terminal event of malignancy. Such lesions may arise from localized spread, as from a carcinoma of the breast, when they tend to be single or grouped and asymmetrical. More widespread haematogenous spread but with multiple lesions are a still more common terminal event. Certain metastases have diagnostic features, such as the scarring alopecia of breast carcinoma affecting the scalp or the pedunculated tumour of hypernephroma.

Signs of underlying malignancy

The three 'P's of pallor, pigmentation, and pruritus are common terminal events in malignant disease, any of which can also be a presenting sign, albeit rarely in the case of pruritis. Defective immunosurveillance predisposes to infections such as candidiasis or herpes simplex and herpes zoster. Although disseminated intravascular coagulation is a common terminal event of malignancy, it may be a presenting sign of lymphoma, leukaemia, or carcinoma of the pancreas. Rarer diseases associated with malignancy are given in Table 21, and include:

- Acquired ichthyosis, in which the skin becomes progressively drier and more scaly. The surface stratum corneum may crack, giving rise to reactive patterns of eczema craquelée (Fig. 20). Increasing scale eventually overlaps with exfoliative dermatitis but, unlike the exfoliative dermatitis due to drugs or psoriasis, the scale is more adherent, in other words it is less exfoliative. There is usually accompanying atrophy of the skin.
- Dermatomyositis is commonly caused by malignancy in white-skinned adults. In children or in Blacks it is more often a manifestation of autoimmune (collagen) disease. The muscle weakness is proximal. The skin signs include erythema (see Fig. 99), lichenoid, or psoriaform eruptions, and itching or tenderness may be considerable. Periorbital swelling and redness, as well as a streaky erythema on the backs of the fingers and ragged telangiectatic nail-folds, are other features.
- Acanthosis nigricans is the pigmentation and wartiness of the axillas and groins. There is a velvety brown thickness of the skin of the hands and at mucocutaneous junctions such as the lips (Fig. 105).
- Acquired hypertrichosis lanuginosa is a generalized increase in terminal hair. It should be distinguished from hirsutes, which is an increase in hair in sites normally associated with hair growth, such as the chin.
- Acute onset of multiple irritable seborrhoeic warts is known as the sign of Leser–Trélat.
- Superficial thrombophlebitis or migrating thrombophlebitis is particularly associated with carcinoma of the pancreas.
- Bullous pyoderma gangrenosum is a feature of leukaemia and myeloma.
- Bullous disease of erythema multiforme type, or occasionally more suggestive of pemphigoid, is more likely to be associated with malignancy if the oral mucosa is involved or if immunofluorescence studies are negative.
- Erythema gyratum repens (Fig. 106) is one of many patterns of erythema forming repeated concentric rings. The more bizarre and rapidly evolving the process, the more likely it is to be associated with malignancy. This is particularly so when it is generalized, oedematous, or scaling (see also Fig. 10(c)).
- Palmar keratoses are found in association with cancer of the bladder or lung.

Cutaneous lymphoma

Few aspects of dermatology have been more confusing than those concerning lymphoma. In some respects it has been simplified by the recognition and identification of B and T lymphocytes and the realization that terms such as 'reticulosis', 'prereticulosis', and 'reticulum cell' were misapplied. T cells normally traverse through the epidermis and upper dermis, producing a horizontal infiltrate and flat patches. B cells prefer the mid and deep dermis, and produce nodules. Expression on the keratinocyte of a3b1 integrin chains possibly explains epidermotropism. The Hassall's corpuscle in the thymus may have features in common with the cells of the epidermis, which explain helper T-cell, Langerhans-cell, and epidermal interaction. The epidermis also produces cytokines, which may account for lymphocyte

behaviour—attraction, differentiation, and proliferation. Classification and staging procedures are complex, and evolving with improved technology. The simplest classification into B-cell, T-cell, and non-B and non-T-cell lymphoma, with a single subgrouping into small cells indicating low-grade malignancy, and large cells indicating high-grade malignancy, is now largely accepted. However, neither morphology nor a panel of monoclonal antibodies have provided a system of identification by which treatment can be planned with absolute certainty and prognosis reliably determined.

Enzymatic, cytochemical, immunological, monoclonal antibody, gene rearrangement analysis, functional, and ultrastructural methods used in research are not routinely available for aiding clinical diagnosis, but they indicate that conditions in which the epidermis is eczematous, scaly, or crusting are usually infiltrated with T cells, especially of Thy-2 lineage and with reduced IFN-g signalling, as in mycosis fungoides, Sézary syndrome, and pagetoid reticulosis. This may explain the reduced capacity to control infection and secondary malignancies. Mononuclear phagocytes, including the Langerhans cells and eosinophils, often infiltrate the upper dermis. Most of the purple-red tumours that show no involvement of the epidermis and produce sharply demarcated infiltrates in the middle or deep dermis are due to B-cell proliferation. The late tumour stage is reflected in the progression to large, anaplastic, lymphoblastic cells.

The previously labelled reticulosarcomas starting as dome-shaped, deep-red solitary tumours are now thought to be lymphoblastic, more often B cell than T cell. Such less differentiated blast cells also give rise to heavy infiltrates in the whole dermis, and are less inclined to produce the nests of cells within the epidermis that are a feature of mycosis fungoides. There is destruction of blood vessels and fibrous tissue, whereas blood vessels are well preserved in mycosis fungoides and often characterized by prominent epithelioid endothelial cells as in postcapillary venules of lymph nodes. Benign lesions show a well-defined germinal centre.

All types of lymphoma may affect any organ, including lymph nodes and the blood, but mycosis fungoides, Sézary's syndrome, and pagetoid reticulosis favour the skin and often seem confined to it. The leonine facies is a peculiar, diffuse, deep nodular feature more often seen in B-cell lymphoma, as in chronic lymphatic leukaemia.

During the early stage of mycosis fungoides, the behaviour of the T cell cannot be proven to be malignant, and some suggest that it is merely hyper-reactive or overstimulated. The source of this stimulus could even be the skin macrophage known as the Langerhans cell, which increase in number in mycosis fungoides. Exactly when 'overstimulus' becomes 'lymphoma' has been debated, but no conclusions can be reached. Environmental infections such as retroviruses and *B. burgdorferi* as well as genetic susceptibility are possible incriminating factors.

The distinctive cell found in the tissues and blood of patients with the Sézary syndrome and in the epidermis of those with mycosis fungoides is a T cell with an usually, but not invariably, hyperconvoluted cerebri-form nucleus. This cell is also observed in a variety of non-lymphomatous dermatoses and should be equated more with the 'overstimulus' concept rather than with malignancy.

Clinical features of lymphoma

Dermatologists have long grappled with the problem of skin diseases suspected of culminating, often years later, in a malignancy of the lymphoid tissue. These diseases have features of chronic dermatitis and psoriasis (parapsoriasis) because there is a chronic reaction in the dermis and epidermis that is often indistinguishable from other causes of such a reaction. One feature that causes anxiety is a lack of symmetry in an atypical distribution; there is also inhomogeneity within the lesion. Infiltration with white cells suggesting tumour formation is another feature. Yet another, is the combination of atrophy or thinning of the dermis, telangiectasia, and pigmentation known as poikiloderma. Persistent superficial dermatitis, previously known as parapsoriasis in plaque (benign type), consists of flat, symmetrical, slightly scaly, red patches on the trunk or limbs that persist for years. They are round, oval, or finger-like and sometimes yellowish (Fig. 107). This is now thought to be a benign condition.

Fig. 107 Lower abdominal, persistent superficial dermatitis (parapsoriasis). These are fixed and persistent digitate (finger-like) patterns, erythematous, and slightly scaly.

Poikiloderma atrophicans vasculare, previously known as parapsoriasis (large plaque or lichenoides), resembles radiodermatitis in that there is atrophy, telangiectasis, and reticulate pigmentation (Fig. 108). It favours areas not exposed to natural sunlight such as the breasts or buttocks. Poikilodermal lesions may be composed of small papules or large plaques of any shape. Although the expected outcome, often many years later, is the cutaneous T-cell lymphoma known as mycosis fungoides, Hodgkin's disease is also a possibility.

B-cell lymphomas, when present in the skin, form firm pink-red or skin-coloured tumours, often in groups coalescing to produce annular or other patterns (Fig. 109).

Fig. 109 Fleshy tumours grouped and arising in the dermis without epidermal hyperplasia. This is characteristic of B-cell lymphomas.

Lipomelanic reticulosis is a non-specific enlargement of lymph nodes associated with widespread dermatitis or erythroderma.

Although mycosis fungoides is often initially no more than a non-specific dermatitis or, more commonly, poikiloderma atrophicans vasculare, occasionally it is a tumour from the start. The lesions may be symptomless, but severe pruritus is common. Affected areas become more infiltrated, scaly, and reddened (Fig. 110), and often they are annular, serpiginous, or have other bizarre shapes. Erythroderma and widespread ulceration is the final stage of the disease. The diagnostic histological feature is invasion of the epidermis by atypical lymphocytes, often in clusters—Pautrier abscesses—and a heavy pleomorphic infiltration of the upper dermis hugging the epidermis but causing less necrosis of individual epidermal cells than in lichen planus.

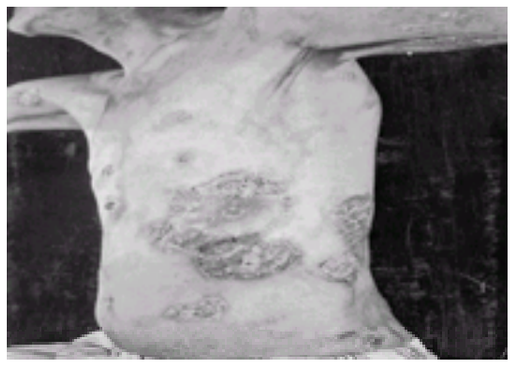

Fig. 110 Marked irregular epidermal reactivity is a characteristic response to T-cell lymphoma of the mycosis fungoides type.

Skin manifestations of Hodgkin's disease include infiltration of the skin with nodules of the disease. Pigmentation and pruritus are common. Prurigo with deep excoriations and secondary infection is one of its most distressing manifestations. Ichthyosiform atrophy as part of the terminal wasting disease is common. The scaling is often as severe as an exfoliative dermatitis, but shedding of the scale is less than that of psoriasis. Hair loss, herpes zoster, and, rarely, erythema nodosum are other complications.

Management of cutaneous lymphoma

The rate of progression is highly variable but usually slow. There is still no clear picture of the natural history of mycosis fungoides. With so many new therapeutic possibilities it should not be forgotten that Samman treated a series of patients conservatively and that only 45 out of 212 patients died of mycosis fungoides. Most of those who died had tumours, skin ulcers, or palpable lymph nodes at the time of presentation, and in the absence of these the prognosis tends to be very good. In patients with benign patterns of the disease it is important not to overtreat. 'Staging' is an attempt to record the extent and progression of the disease in the skin and lymph nodes using multiple biopsies and scanning of internal organs. Some localized forms of lymphoma, even when anaplastic, can be cured by excision. Extensive disease, when confined to the skin, responds well to radiotherapy and topical medication.

Radiotherapy

Small-field orthovoltage radiation has been standard therapy for many years and is very useful for controlling plaques and tumours resistant to other modalities. It is not unusual for patients to require a small dose of radiation to only one area at as little as yearly intervals. Electron-beam therapy is recommended for most patients with extensive infiltrated plaques or tumours. Although a high initial response rate can be expected in the majority of patients, they only remain free of disease for about 3 years.

Topical nitrogen mustard (mechlorethamine, HN_2)

This is a useful treatment for patients who have less-infiltrated skin lesions. Clinical response may be slow and maintenance therapy may be required for at least 2 years. The chief side-effect is allergic contact dermatitis, occurring in about 30 to 60 per cent of patients. Desensitization can be attempted but is difficult to effect. Although there is some debate as to whether an aqueous or ointment-based preparation is best, the latter probably produces fewer hypersensitivity reactions.

PUVA

Several reported series of the good effects of PUVA have resulted in most academic departments using this as a first-line therapy for widespread superficial lesions. However, PUVA penetrance is limited so that deep tumours are unlikely to be cleared. Erythrodermic (Sezary) patients do well with extracorporeal photochemotherapy.

Systemic chemotherapy

On the whole this is reserved for palliation in patients with systemic disease and deep tumours. There is usually some initial response, but clearance for more than 1 year is unusual. Cutaneous lymphoma is susceptible to immunosuppression and this is a rich field of investigation at the present time.

Viral warts

Warts are caused by the papovavirus (see Chapter 7.10.17), which enters the skin through small abrasions, particularly if the skin is moist and warm. Virus is found by electron microscopy in the differentiating cells of the upper epidermis rather than in the proliferating basal-cell layer. The incubation period is probably several months. A number of strains of wart virus give rise to different types of warts—common, plantar, mosaic, plane, and anogenital. Molluscum contagiosum is caused by a pox virus (Fig. 111).

The incidence of warts is increased in immunosuppressed patients, either from drugs or associated with lymphoma. However, cell-mediated immunity is more certainly a factor than humoral immunity. The peak incidence is in children between 12 and 16 years of age, and in recent years there seems to be an increase in infection rate in people living in Europe and the United States compared to Asia, Australia, and Africa.

Trauma may account for the distribution of warts on the hands and feet. Nail-biting in children and shaving in men, as well as ill-fitting shoes in adults, are all relevant factors. Some 20 per cent of warts disappear within 6 months and 65 per cent in 2 years, although plane warts and mosaic warts are slow to clear.

Common warts are firm papules with a rough horny surface. They occur singly or coalesce into large masses. The knuckles and nail-folds are particularly favoured, as are the knees and, more rarely, the penile shaft. They should be differentiated from warty tuberculosis, which is usually a solitary plaque with an erythematous border. Granuloma annulare of the knuckles does not have a horny surface. A persistent wart on the toes or fingers may be a reaction to a subungual exostosis. Squamous epitheliomas or keratoacanthomas are usually solitary and found in an older population.

Plane warts are smooth, flat, or slightly elevated and affect the face or back of hands. They may coalesce or form linear lesions in scratch marks. Lichen planus may be difficult to distinguish from plane warts but is unusual on the face and prefers the flexor surface of the wrists as well as the oral mucosa. The histology of plane warts is unexciting, whereas lichen planus shows destruction of the basal-cell layer of the epidermis and a heavy infiltrate of mononuclear cells.

Filiform and digitate warts are common in the beard area, on the lips, and in the nasal vestibule.

A plantar wart begins as a small 'sago-grain' papule. As it enlarges, paring the surface with a scalpel distinguishes the wart from the surrounding horny ring of normal epidermis and reveals the small capillaries in the tips of the elongated papillae. Most warts occur over pressure points. Clusters of small warts make up a mosaic. A wart which shows numerous thrombosed capillaries and is darker than usual is probably regressing. The fourth interdigital space is a common site for soft corns due to pressure of the little toe on the head of the metatarsal caused by a tightly fitting shoe, these are often seen in ballet dancers. Soft warts or even condylomata lata have been described at such sites.

Treatment

Most human papillomavirus (**HPV**) infections in the vagina and cervix are invisible to the naked eye but can be identified by painting the area with acetic acid (3–5 per cent); however, the demonstration of latent virus is of uncertain biological significance. Although infection with HPV is not sufficient to cause cancer, it may act as a

promoter. There is no successful means of eliminating HPV. Warts should be treated on the basis that they are unaesthetic and uncomfortable.

Spontaneous resolution is to be expected. Overall, 12 weeks is the usual time required to cure warts irrespective of the treatment used, and most standard treatments do no better or worse than this. Podophyllin and formalin or salicylic acid are standard therapies.

Podophyllin, 10 to 20 per cent in liquid paraffin or in tincture benzoic compound, is painted on to anogenital warts and the area then powdered. The podophyllin is irritant and some patients need to wash it off in 2 h, others feel no such discomfort. However, podophyllin should not be used during pregnancy since absorption sufficient to damage the fetus is a possibility. The treatment is repeated at intervals of 1 to 3 weeks.

Formalin (as a 10 per cent solution) can be applied as a soak to multiple warts on the soles of the feet, but dryness and fissuring may be troublesome.

Salicylic acid is the most reliable chemical for treating warts. Paints or plasters containing 20 to 40 per cent salicylic acid are best applied after a 5-min soak with warm soapy water and preferably after excess surface keratin has been removed.

Imiquimod, which induces interferon-a and other cytokines, is now available as a topical agent and is effective against anogenital warts and molluscum contagiosum.

Freezing is with liquid nitrogen, either in a special spray or by application from a cotton-wool bud on the end of an orange stick. The wart should be whitened for at least 20 to 30 s and blistering is a common consequence.

Local anaesthetic injected into the base of a wart to lift it up from the dermis can be followed by curettage. Compression with the thumb on immediate adjacent tissue prevents bleeding while silver nitrate is applied. The rim of horny tissues around the wart side should be cut away using scissors.

Curettage of molluscum contagiosum may be made painless in the majority of children, except at mucocutaneous junctions, by the use of a eutetic mixture of local anaesthetics (**EMLA**) applied under occlusion for 60 min.

Granulomas and other infiltrations of the skin

A granuloma is a compact accumulation of cells, comprising mainly monocytes or their variants, macrophages, epithelioid cells, and giant cells. Often there is subsequent fibrosis. Lymphocytes are more numerous in granulomas due to allergens to which the host is sensitive. Degeneration or the presence of foreign bodies encourage neutrophil and eosinophil participation.

Granulomas are classified as high or low turnover:

- *high turnover:* tissue-destructive, induced by toxic irritants or delayed hypersensitivity, continuous recruitment of macrophages and many mitoses, and frequent epithelioid and giant cells;
- *low turnover:* space-occupying but not destructive, induced by inert (bacterial) and non-degradable irritants, no continued recruitment but long survival of macrophages and few mitoses, and few epithelioid and giant cells.

The clinical features of granulomas are either space-occupying nodules lying in the dermis or, if close to the skin surface, they may be seen as yellow or brownish-red and sometimes translucent areas. The chronic changes in blood supply associated with the lesion cause a bluish colour and sometimes telangiectasia. If they are located in the upper dermis, there may be overlying epithelial hyperplasia or ulceration with extrusion of some of the granulomatous material. On the other hand, thinning of the epidermis may be considerable. In dark skins, pigmentary changes may include hypo- or hyperpigmentation.

A common cause of granulomas is persistent irritation of the skin by external trauma causing ulceration and pseudoepitheliomatous hyperplasia. Examples include granuloma fissuratum of the ear or nose due to ill-fitting spectacles. The ingrowing toenail, the pilonidal sinus, or the presence of extrafollicular but intradermal hair (as is seen in the interdigital clefts of barbers, and cattle or horse dealers) are other examples.

Granuloma gluteal infantum is seen in the nappy area due to incomplete resolution of an irritant rash to which steroid creams have been too extravagantly applied. Numerous agents acting as foreign bodies are causes of chronic granulomas in the skin. They include sea-urchin spines, silicates, cactus allergen, grit, and various chronic infections such as *Candida albicans*, *Trichophyton verrucosum* (Fig. 112(a)), coccidioidomycosis, atypical mycobacteria from fish tanks or swimming pools, tuberculosis (Fig. 112(b)), leprosy (Fig. 112(c)), leishmaniasis (Fig. 112(d)), and halogen granulomas (see Fig. 26).

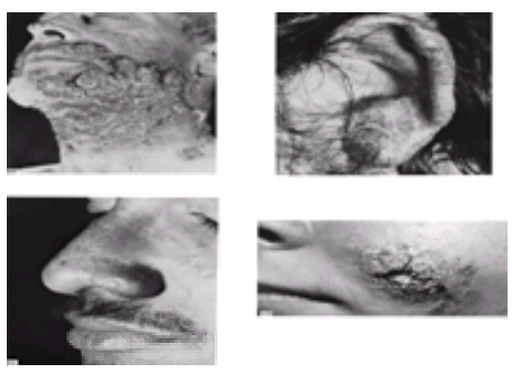

Fig. 112 (a) Chronic granulomas due to cattle ringworm: *Trichophyton verrucosum.* (b) The ear involved by lupus vulgaris (tuberculosis); a brownish-red granuloma inducing irritation in the overlying epidermis. (c) The nose is a common site for the granuloma of lepromatous leprosy. (d) Chronic granuloma due to leishmaniasis following a sandfly bite presenting as an ill-defined cresting and wartiness.

Sarcoidosis

See Chapter 17.11.6 for further discussion.

Urticaria pigmentosa or mastocytosis

Mast cells are normally present in the skin but the numbers vary greatly, with up to 80 per mm^3 found in the upper dermis. In mastocytosis they are greatly increased in number, and may be found as a single isolated mastocytoma, or as numerous nests scattered over the entire body (that is, the classic urticaria pigmentosa) or diffusely throughout the entire skin. Occasionally there is systemic infiltration of all tissues including the liver, spleen, and bone marrow. A very rare leukaemic variety is also recognized.

The mast cell releases histamine, leucotrienes, and heparin, all of which may have systemic effects. However, it is increasingly realized that the local contribution to mastocytosis is through the secretion of proteases from these cells.

In the infant, mastocytosis may present as blisters; more commonly, the lightly pigmented swellings in the skin are noted and observed to swell when scratched or following a hot bath or exercise. Rarely, there is generalized flushing and itching. The condition is most common during the first year of life or at birth, and an onset at this age is a good prognosis for eventual complete resolution by adolescence.

In adults, a late onset is associated with diffuse plaques and telangiectasia.

The systemic variety presents in 10 per cent of adult cases, causing osteoporosis or osteosclerosis. The spleen may be enlarged, and bleeding disorders are the consequence of either thrombocytopenia or from the effects of heparin. Involvement of the gut causes a variety of symptoms including colic and diarrhoea. Right-sided heart failure due to pulmonary hypertension is recorded. Urinalysis for histamine or prostaglandin D_2 may help to confirm the diagnosis when the skin lesions are absent.

Treatment

Treatment is unsatisfactory, but the increasing use of various combinations of H_1- and H_2-antagonists is proving beneficial in some cases. The prognosis for eventual resolution is good in children. A solitary or troublesome single lesion in an adult can be excised. The cosmetic appearance of pigmented lesions is helped by sun exposure or by the use of UVA and psoralens, but the number of mast cells is not reduced. Disodium cromoglycate helps some patients with systemic mastocytosis. The number of mast cells can be suppressed by high-potency steroids applied under occlusive dressings.

Cutaneous manifestations of histiocytosis X

The cutaneous lesions of histiocytosis X are small yellow-brown keratotic scaling papules. These coalesce to form a diffuse seborrhoeic dermatitis that is ulcerative, crusting, and purpuric. Granulomatous eroded plaques that are particularly found in the flexures and in the external auditory meatus cause great discomfort. The hair margins are commonly involved. The common association of diabetes insipidus and hepatosplenomegaly is described elsewhere. The diagnosis is confirmed by demonstrating pale-staining histiocytes devoid of lipid, which contain the Langerhans-cell granules.

Fibrosis, eosinophils, and giant cells are features of a more benign process.

Granuloma annulare and necrobiosis lipoidica

A partial necrosis of collagen- and connective-tissue cells associated with immunoglobulin and complement deposition results in a lymphocytic and histiocytic response known as a 'palisading granuloma'. This is entirely reversible over many months and years in patients with granuloma annulare, but in those with necrobiosis lipoidica it tends to result in fibrosis and scarring. The association with insulin-dependent diabetes mellitus and with AIDS is unpredictable, but is to be expected in more widespread forms, in older age groups with granuloma annulare, and in about 75 per cent of cases of necrobiosis lipoidica.

In children, granuloma annulare commonly appears on the knuckles (Fig. 113), fingers, and dorsum of the feet. Ears and elbows are quite frequently affected. Granuloma annulare may be mistaken for warts but the overlying epidermis, if closely inspected, is rarely papilliferous. The tendency to heal in the centre and spread centrifugally over many weeks gives rise to the annular appearance.

Fig. 113 Papules of granuloma annulare forming a ring around a now healed, but previously affected, area on the knuckle.

Necrobiosis lipoidica is commonly found on both shins (Fig. 114). Widespread forms of granuloma annulare may often be of the giant type, forming large violaceous plaques or rings. No treatment is necessary since eventual resolution of granuloma annulare is expected in 75 per cent of cases within 2 years, but intralesional steroids probably speed resolution, particularly sometimes aborting necrobiosis lipoidica. These disorders may respond to PUVA therapy.

Cutaneous amyloidosis

Systemic amyloidosis is described elsewhere. A waxy appearance of the skin and the ease with which purpura develops within the lesions on slight trauma are suggestive of the diagnosis.

Lichen amyloidosis consists of discrete, firm, hemispheroidal papules. Hyperkeratosis and pigmentation is common, suggesting a waxy infiltrated lichen simplex. The lower legs and outer thighs are involved (Fig. 115(a)). There is no systemic implication.

Fig. 115 (a) Lichen amyloidosis commonly affects the shins or outer aspects of the thighs. The brownish, waxy, lichenified skin is pruritic. (b) Rippled pigmentation of macular amyloid.

Macular amyloid is a common pigmented variant, often in a rippled pattern (Fig. 115(b)), affecting the shoulders and backs of Asian peoples. Local high-potency steroids under occlusive dressings have a temporary good effect.

Crohn's disease

This is a well-recognized cause of chronic granulomatous infiltration occurring perianally (see Fig. 21) or in the buccal mucosa.

Management of skin disease

Recent advances in treatment have been many. The understanding of cell recognition, adhesion, and the role of interleukins, tumour-necrosis factor, and interferons have led to the development of significantly successful antagonists, though they are not without side-effects. The skills of the dermatological surgeon have been refined, and laser therapy has provided a tool that can be both well and badly used. There will be gene guns in the future. The long reign of steroids may be coming to an end as more specific anti-inflammatory agents with less side-effects are trialled. It has to be said, however, that the basis of dermatological therapy is still empirical and randomized controlled trials rare. The principal symptom of itch remains poorly understood, and the 'look good, feel good' factor that contributes to well being is elusive.

Conventional dermatology is not the only resource for patients with skin problems: alternative and complementary medicine, some of it based on long tradition, has become increasingly popular.

General principles

Acute lesions require rest and elevation because this reduces swelling. Dressings should be applied lightly and evenly to the surface, and should support the inflamed area without drag or compression. All agents applied to the skin should be no stronger than necessary. It is very easy for instance, to make a potassium permanganate solution too strong or to use poorly mixed and inhomogeneous medicaments. Preferably, only substances whose components are known and prescribed by experts should be applied to the skin. This is particularly relevant to indigenous medicine. A simple agent used well and which is familiar to the prescriber is always safer than the haphazard prescription of a range of new and poorly tested agents.

Some of the oldest remedies, such as calamine lotion and tar preparations appearing in the *British national formulary* or *The WHO essential drugs list*, do least harm, but patients with chronic skin disease are unlikely to persevere with these unless they are educated and encouraged. This takes time. Remember, patients spend more time listening to other patients in the waiting room than they do to the doctor—and are often in a more receptive frame of mind then. Make sure that the advice given in the waiting room is correct by handing out leaflets and teaching groups of patients, for example with audiovisual aids. Many patients do best with printed sheets of instructions so that they can read at their own pace quietly, and away from the physician. When appropriate, they should be told that their disease is neither contagious nor cancer, nor passed on to their offspring by their genes.

There are a number of supporting organizations of value in the management of skin disease. A high proportion of patients with skin problems are found to be suffering from the stresses of their domestic environment (confused elderly relatives or delinquent teenage children, for example). For these patients the help of a medical social worker is often very successful. The British Red Cross Society in the United Kingdom now provides a 'beauty care' service that has been extended to the provision of camouflage make-up. Patients' associations often provide a great deal of education and support.

In hospital practice, inpatient treatment is only occasionally necessary and can be greatly reduced by adequate provision of a good outpatient service. Some units link this to their inpatient department in order to provide a service outside the unit's normal working hours—it is clearly more convenient for a working person to attend a department that is open in the evening.

Elimination of primary irritants or known allergens, or of infection, should be the aim of treatment. However, it is fruitless to attempt to make the skin and its diseases sterile at all times. Restoration of the skin barrier and the prevention of cracks and heaping up of scales and exudate reduces bacterial penetration. Scratching, rubbing, wrongly applied dressings, and unsuitable local medications are reasons for worsening of the skin condition and for the impairment of natural defences and repair mechanisms.

Chronic skin conditions are more difficult to cure, and a correct diagnosis is even more important. Spontaneous healing may not occur without the cause being eliminated, and therefore the chronic inflammation may be self-perpetuating. A biopsy for histological analysis is often helpful and bacterial or mycological analysis is clearly indicated where chronic infection is a possibility.

The handicap of the chronic lesion includes a feeling of being unwelcome, often leading to the accusation of being 'unclean' and to being 'outcast'. The physician and nurse can sometimes do more to alleviate the skin condition by paying attention to this aspect of the handicap, rather than using more specific measures. Sympathetic questioning along the lines indicated earlier will help to relieve the patient's suspicion that the dermatology team is not interested in the problem. Touching the skin during examination often does more to make the patient feel that the physician cares than any other manoeuvre.

When treating any skin disease, the overall objective is to create the body image that the patients hope for—not that which is perceived as ideal by the attendant—and when this fails, to help them to live with their problem and expect less. Every effort should be made to ensure that children remain willing to go to school, adults stay at work, and the old are kept comfortable. Remedies should be convenient and cosmetically acceptable, and if this is not possible, the social support should be that much more intensive. Attention to diet, camouflage measures, and regular careful attention to the skin can be very irksome for the patient, but can be made less so through cooperation with the doctor, nurse, or patient association, for example the obesity clinic supported by a weightwatchers' group.

It is essential that the skin be protected from further injury. Homely advice such as 'keep warm and out of the sun' and 'rest as much as possible with your feet up' is appropriate for much acute or severe skin disease, but the skin does not normally experience sustained rest. It is best to encourage movement such as intermittent stretching or compression. For a normal and healthy skin, movement promotes blood supply, lymphatic drainage, and a well-balanced structure that is water-containing and to which a well-attached epidermis is attached and supported by appropriately distributed fibres as skeletal support. The best movements are natural and not too vigorous, but they should provide a full range of flexion and extension over the joints.

Some chronic and incurable skin diseases become easier to tolerate when time is given to the education of the patient, thereby increasing patient understanding, satisfaction, and self-help.

The priorities of the patient with skin disease differ from those of the therapist. Thus the patient equates severity with social ostracism or the subjective itch, rather than with the percentages of body involved or systemic complications. Other people who should be drawn into the management of the patient with skin disease include parents, school teachers, employers, hairdressers, and the sporting fraternity—for example, swimming-pool attendants. A school report that, without these discussions, refers to weeping sores and scratching which interfere with the work of other children, should be a thing of the past.

Many chronic skin conditions are either hypertrophic or atrophic. Hypertrophic conditions require suppressive therapy, as described above in the section on psoriasis. Common suppressants are corticosteroids or radiotherapy. However, these are less suitable for treating atrophy. Another adage worth paying attention to is: 'If it is wet, dry it, and if it is dry, wet it!'

The management of skin disease in technically developed countries usually assumes regular follow-up. In developing countries in Africa or India, or with the nomads of the Middle East, there is almost no possibility of follow-up. Treatments with a high risk of side-effects or exacerbations on withdrawal are therefore not ideal. Because malnutrition and infection are so common, dietary supplements and antibiotics are of value in almost any disease in which host or constitutional vulnerability is a factor. Health education is always difficult, but the mother with children is usually the most receptive pupil.

New technologies—lasers and narrow-band UVB

The management of skin diseases is usually low-technology and low-cost. Economic forces have encouraged dermatologists to purchase high-technology equipment, some of which provides an advance in management for those who can afford it: laser therapy has proved excellent for vascular and pigmented lesions, and for the removal of hair. The technology requires a wavelength that is selective for a specific subcellular target, delivered in pulses with sufficient energy to be destructive for that target without at the same time damaging surrounding tissues. As the pulse is so short, there is no time for surrounding tissues to be heated. Many treatments are needed for some of the more troublesome conditions (for example, pigmentation), while some therapies (such as hair removal) are not permanent but partial. The technology has been improved by photodynamic therapy, whereby an agent such as protoporphyrin is added to the subcellular structure, which is then irradiated. Several tumours have been treated in this way. Dermatologists have always used ultraviolet light, either from sunlight or from ultraviolet lamps. The most recent innovation is the use of narrow-band UVB, which is as effective as UVA in the prophylactic management of polymorphic light eruption and is used in a number of other sun-induced dermatoses. While narrow-band ultraviolet light has been found to induce malignancy in animals, the lower dose required in humans compared to other

forms of light seems to induce remissions without significant tumour risk. It has been given to children and patients who are pregnant, but it is probably too early to say that it is entirely safe.

Local topical treatment

The fingertip unit

The 'fingertip unit' is the amount of ointment expressed from a tube with a 5-mm diameter nozzle, applied from the distal crease to the tip of the index finger—1 unit weighs 0.49 g and covers 286 cm^2 in men, and 0.43 g which covers 257 cm^2 in women. One unit will cover an area equivalent to twice the area of the hand, thus four hands is equivalent to two finger units, which is equivalent to about 1 g.

Drugs are dissolved or suspended in bases, which have properties of their own quite independent of the active ingredient. As shown in Table 22, bases were originally either powder, water, or grease. However, modern processes have prepared bases that are essentially much more complex than this, although they still retain the objectives of the primary agents. Powder may repel water or absorb it and allow further evaporation. Modern powders tend not to cake and abrade the skin as much as the original talc or starch. Watery lotions evaporate and cool, as well as wet and dissolve. Various agents, such as alcohol or glycerine, may be added to increase any one of these properties. Creams and emulsions of oil and water (aqueous or milky) or water in oil (butter or oily) are cooling, moisturizing, and emollient. The penetration of active agents through the skin is aided by the aqueous (vanishing) oil and water creams. Ointments based on Vaseline or paraffin are more occlusive and less quickly absorbed. They are better at softening dry surface scales.

There are various other water-soluble preparations, such as macrogels or emulsifying ointment in which a wax or animal fat is mixed with mineral oil. Pastes are powder and oil mixtures, such as talc and Vaseline, which are more occlusive and protective. They are useful for allowing the slow release of agents (such as dithranol) at the skin surface. As the addition of an active ingredient to a base often makes it unstable, various other agents are added as a preservative or to control pH. Further dilution usually makes the preparation still more unstable (that is, shortens its shelf-life). Much of the skill in preparing an ointment, cream, or paste lies in the use of the homogenizer by the pharmacist.

The actions and side-effects of topical steroids are listed in Table 23. Tar and dithranol preparations were discussed above in the section on psoriasis. Tacrolimus is a new topical immunomodulating agent related to ciclosporin. However, unlike ciclosporin, it can be used topically to treat conditions such as pyoderma gangrenosum or the more common psoriasis and atopic eczema. A new oral agent, mycophenolate mofetil—another immunomodulatory agent (1 g orally, twice daily) for psoriasis, pyoderma gangrenosum, bullous pemphigoid, pemphigus vulgaris, and systemic vasculitis—also shows some promise as a topical agent.

Skin cleaning

It is naïve to attempt to sterilize the skin, and the long-term consequences of obsessive washing or the use of local antibiotics are always worse than the original state of the skin. Antibiotic regimens are essential for acute complications such as cellulitis, or for specific infections such as erysipelas. Washing is important for reducing smell and for removing debris, but this is best done by soaking rather than scrubbing. Soaking is, in fact, one of the most effective of skin treatments for oozing exudative conditions. Management of skin conditions requires clean water fit for drinking; this can be obtained by a variety of sterilizing procedures, including boiling, pasteurizing for prolonged periods in the sun, adding antiseptics, charcoal, and other forms of filtering. Soaps irritate because they are alkaline and degreasing agents, and some patients are sensitive to perfumes and other additives in the soaps. Most skin will tolerate some soaping, but the amounts needed to degrease in hard-water areas can cause considerable dryness and cracking of the skin. Soft rainwater or boiled milk should not be despised. Bran is an ancient and harmless water softener: about a pound of bran or oatmeal tied into a muslin bag and soaked in boiling water produces a very thin, starchy emulsion. Because cold causes the stratum corneum to dry, shrink, and crack, cold water is not therapeutic for the skin. Dry skin is best treated at body temperature.

Emulsifying ointment is a useful soap substitute; it can be made into 'cakes of soap' or spooned out of a pot and mixed with hot water to soften it. Liberally applied it is a useful softener of crusts.

Bathing in water, often for prolonged periods several times a day, is an effective remedy for a generalized sore skin. However, the skin tends to dry excessively when all grease is removed and this can enhance pruritus. For this reason, emulsifying ointments and proprietary (oilatum) oils are usually added to the bath.

Although the use of antiseptics in the bath is of dubious value, weak solutions of potassium permanganate 1/8000 to 1/16000 are often used. Patients often use these agents too extravagantly. High concentrations of antiseptics poured into a bath may lie on the bottom of the bath and burn the skin, especially in sensitive areas such as the scrotum.

Bacteria are best dealt with by removing crusts and other debris, soaking does this even without the addition of antiseptics. In intertriginous areas soaking should be followed by drying. Organisms thrive in moist crevasses, for instance under the breasts, in the groins, and between the toes. Non-caking powders are helpful in drying such areas, and many proprietary brands of powders such as ZeaSORB® can be recommended. Like leather, treatment with grease prevents cracking and penetration by undesirable agents.

Softening crusts and exudates

Crusts and exudates (for example, as observed in impetigo, fungus infections, acute dermatitis, and psoriasis) require softening by prolonged contact with a wetting or greasing agent. The problem with wetness is that it quickly evaporates and dries. Wet soaks require absorbent dressings, but modern hospitals often supply only gauze, which is not particularly wettable. Old-fashioned linen or cotton sheets are ideal for wet dressings; usually these are applied in several layers and covered with a light ventilated dressing such as a hand towel or tubular gauze. Occlusive dressings are too heating. Polythene occlusion should not be used. Less rapidly drying agents include the ancient boric and starch poultice—30 g of starch and 4 g of boric acid mixed in 568 ml (1 UK pint) of water, cooled, and smeared on to linen strips to thickly impregnate them. This wet dressing is applied to the crusted area and changed every 4 hours. Vaseline ointments are also suitable for softening scales, as in psoriasis. Where the skin is dry and non-exudative, scaling is best softened with a paste made of 50 per cent Vaseline and 50 per cent talc (talc slowly releases the Vaseline over a number of hours).

Other treatment

Smell

Malodorous necrotic skin is always very difficult to deal with, but the removal of debris and dead tissues is essential. Antibiotics and local antiseptics cut down bacterial degradation, which is a cause of smell. Metronidazole 400 mg three times daily is sometimes used to reduce the smell of tumours as it is effective against various anaerobic bacteria. Charcoal dressings are also helpful. Social intercourse out of doors can be encouraged, while perfumes may be helpful indoors. Washing removes dirt debris and excess bacteria, making the skin less attractive to biting insects.

Diet

Widespread skin disease is a cause of water and protein loss. The tongue and degree of thirst are a good guide to water loss, as is the specific gravity of urine. High-protein diets are necessary, especially when there is great exfoliation.

Retinoids

The greatest advance in dermatological therapeutics of the last decade has been the introduction of systemic retinoids. The main retinoids available are *cis*-retinoic acid and acitretin. These modulate cell differentiation and growth, inhibit polymorphonuclear-cell chemotaxis, inhibit polyamine formation, and inhibit eicosanoid formation. They are effective in the treatment of acne, psoriasis, and genetic dyskeratoses. Their side-effects include teratogenesis, hyperostosis, lipidaemia, hepatitis, and the various minor skin, bowel, and neurological problems previously recorded with vitamin A prescribing. Retinoids are prescribed in combination with other topical therapies or PUVA therapy in an attempt to reduce their dosage and consequent side-effects.

Mood-controlling drugs

Apart from the value of psychotherapeutic drugs to control the secondary emotional reactions to skin disease (many such drugs are used to control skin symptoms), antipruritics are often sedative, but hydroxyzine and alimemazine (trimeprazine) are favoured. Delusions of parasitosis and obsessive concern with pruritus is relieved by pimozide.

The placebo

For many skin conditions, such as alopecia areata, there is no specific treatment, and for some patients the available effective remedies are inappropriate, such as the painful treatments for warts in very young children. Nevertheless, to do nothing at all is to encourage despair. Placebos such as mild lotions for alopecia areata or warts should be harmless, cheap, and given knowingly without self-deception.

Subjective symptoms such as itchiness are intolerable for some patients, but can sometimes be relieved by inert agents such as calcium lactate or vitamin B pills given with assurance and confidence.

Bed rest

It is hard for patients with skin disease to play the sick role, especially when they are otherwise well. When told to rest at home, sitting on a couch and pottering about are often only a partial acceptance of the sick role. Admission to hospital and complete bed rest often switches off skin disease such as atopic eczema within 2 or 3 days; for psoriasis, 2 or 3 weeks is often required.

Irritable and distraught, the patient and his family may need rest from each other as well as the knowledge that the illness is genuine and severe enough to take to bed. The patient's bed should be placed where there is no danger to other patients from cross-infection (exfoliative skin disease is a rich source of staphylococcal infection), which means that neighbouring patients should be selected both for their likely resistance to infection and for their likely good companionship. Modern hospitals include a high proportion of single cubicles for patients with skin disease, together with a day room for social mixing once the skin dressings have been completed.

Intertrigo

The treatment of intertrigo is essentially undertaken to protect the area from chafing and secondary infection, and to encourage dryness. Underlying disorders such as diabetes mellitus or obesity must be managed along traditional lines. Infection from bacteria, *Candida* spp., or fungi requires monitoring by appropriate swabs and scrapings. Bed rest and nudity are helpful. In hot climates a fan encourages evaporation and drying. Skin folds should be kept apart by ventilated loose-weave dressings. Acute eczema requires bland wet lotions, steroid creams, and simple antibacterial agents such as gentian violet or vioform cream. When dry, powdering is to be encouraged. Frequent bathing is always helpful.

Hand dermatitis

To provide healing and to prevent relapse of dermatitis of the hands, patients should use lukewarm water and emulsifying ointments when washing. If possible, running water is better than a prolonged soak in a bowl of detergent soap. Soap should be used sparingly and the hands thoroughly rinsed and dried carefully with a clean towel. As far as possible there should be no direct contact with detergents and other strong cleansing agents, shampoos, polishes, and stain removers. Oranges, lemons, grapefruit juices, and various other irritant vegetables should be avoided. Rings should not be worn during housework or other work even when the dermatitis has healed, because irritants often collect under them. Rings should be frequently cleaned on the inside with a brush and left in ammonia (one tablespoon (20 ml) to 500 ml of water) overnight, then rinsed thoroughly. If gloves are used for washing dishes and clothes, they should be made of plastic and not rubber since the latter often causes dermatitis. They should not be worn for more than 15 to 20 min at a time. If water happens to enter a glove, it must be taken off immediately. The gloves should be turned inside out and rinsed several times a week. Sprinkling with talc before they are used helps to dry them. Cotton gloves can be used under the plastic ones. They should only be worn a few times before they are washed.

Vulnerability

The skin in many diseases is vulnerable. This is manifested as the tendency to produce disease even from minor trauma. It is seen in the primary irritant dermatitis of the atopic eczema sufferer, the Koebner phenomenon of psoriasis, lichen planus, the hyperreactivity of the skin to needle puncture or pressure localization in vasculitis, and in the skin ulceration that results from minor knocks to the legs in gravitational stasis. The skin's vulnerability is severe in epidermolysis bullosa.

In all these diseases advice has to be given about protection of the skin. It can be given in the form of information sheets detailing the care of the hands or legs, or in booklet form for mothers of children with epidermolysis bullosa. The various patient associations produce excellent literature in this respect.

Management of leg ulceration

The cause of ulceration should be identified and, if possible, eliminated. However, the aetiology of an ulcer in an elderly woman living in a city is likely to be different to that of a young man in a rural area in Central Africa. The causes of ulceration are listed in Table 24.

Elevation

The leg being below the level of the heart, there is always a tendency to develop venous hypertension and stasis. Deep vein thrombosis, absence of valves in the deep veins, and shunting of blood from the deep veins to the superficial veins via perforators are a significant cause of congestive changes in the microcirculation supplying the epidermis.

Healing of leg ulceration is always helped by elevation. If there is a major degree of arterial disease, so that there are absent peripheral pulses, then the leg should not be raised more than about 23 cm above the level of the heart. In every other case, emptying of the distended veins and superficial venules is helped by lying the patient in a prone position and elevating the legs to an angle of at least 45°. This is best done by placing an object such as a chair under the mattress (Fig. 116). It is also best to elevate the leg during the day, because the patient cannot sustain such a position when asleep and will curl up into a bundle at the top of the bed.

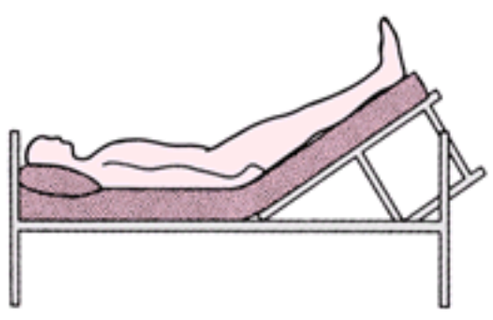

Fig. 116 Elevation of legs above the heart, necessary in the treatment of leg ulcers. A chair can be used to prop up the end of the mattress.

When stiff hips, heart failure, and obesity are factors preventing elevation of the legs, a compromise includes the use of compression bandages and attempts to make the most of any muscle pump in the lower leg. Intermittent positive pressure inflatable bags are commercially available. These are leggings that blow up and squeeze the legs at a pressure and rate that can be regulated.

Elevation is also a requirement for the treatment of lymphoedema, but the protein collecting in the tissues of a swollen leg usually fails to clear satisfactorily via the lymphatics. Movement such as massage, vibration, or ankle exercise is necessary to hasten the passage of protein out of the tissues into the lymphatics. Provided the solid elements of the skin (that is, collagen fibres) are moved, the massage or vibration need not be very sophisticated. In the absence of lymphatics the only other way protein can be removed from the tissues is by proteolysis and macrophages. To control venous hypertension, the pressure at the ankle should be 40 mmHg and there should be a gradient of decreasing pressure as one moves proximally. This gradient is aided by the increasing circumference of the calf. Each layer of bandage doubles the compression.

Arterial disease should always be excluded by examining the arterial system, especially by taking the blood pressure in the legs and arms or by an arteriogram. Stopping smoking and keeping walking is essential therapy.

Wound dressings should encourage moist wound healing, which favours epithelial migration and the development of granulation tissue.

Movement

Inflammation is aimed at removing injurious agents and promoting healing. In acute infection or injury there is often a need for immobilization. However, such immobilization should be localized to the site of injury. Gentle passive movement of the joints and active movement of the main muscles of the leg are encouraged by wriggling the toes or ankles, or by quadriceps exercises.

Unfortunately, many patients with leg ulceration continue to be immobile and to dread any movement that causes pain, long after any need for immobilization to contain the inflammation. Stiffness of the ankle and contractures at the joints are common and delay healing as well as considerably add to crippling.

Deep vein thrombosis is a consequence of immobilization, and this too contributes to morbidity as well as mortality. Thus, in general terms, the maintenance of mobility is essential in the management of leg ulcers.

Exercises in bed can be followed by exercises in the standing position aimed at maintaining an upright posture and ankle mobility, as well as strengthening the muscles of the calves which are so important in pumping blood through and away from the deep veins. For those who are able, walking should be encouraged. Special instruction should be given concerning the harm done by sitting with the legs dependent or crossed, or standing without movement at the ankle. Even a soldier standing to attention needs to maintain venous return by imperceptible, but nevertheless effective, wriggling of the toes. 'March at the sink' is a suitable war cry in the home.

Dressings and bandages

The objective of bandaging is to hold the dressing in position, to protect the leg from further injury, and to provide a sleeve against which movement of the underlying muscles can compress and empty the superficial veins.

The superficial vessels of the skin of the leg are often distorted and congested from chronic inflammation and gravitational stasis. Such vessels are often damaged and cause ischaemic ulceration as a result of external injury, arising from the kinks, wrinkles, and inequalities of an ill-fitting bandage, etc. It should be remembered that a leg that swells can develop severe ischaemia beneath a constricting bandage, even when it was quite loose before the leg became swollen. Large, swollen legs, as in lymphoedema or after deep vein thrombosis, can tolerate unskilled bandaging and tight compressive bandages. By contrast a thin, ischaemic or ageing skin suffers greatly from carelessly applied bandages. The leg is an awkward shape, particularly around the ankle, and many of the twists and the bulk of the bandage tends to be over the bony prominences where the skin is thin and ulcers are common. One system of bandaging is to use two layers—one as a dressing and the other as a cover. The bandage used as a dressing is made from strips of material, no longer than 1.5 times the leg circumference, that are impregnated with a paste. These strips may be applied from a bandage that is cut whenever the direction of the bandaging requires a change. Above the ankle, the bandage is folded at the side of the leg and reversed so as not to completely encircle the leg. Any strips of materials such as calico or linen similarly impregnated will do as well. The overlying covering bandage should be more stretchable than cotton—plasticity rather than elasticity is necessary for a thin leg. Large lymphoedematous legs may be covered by a stronger inelastic material (short stretch). This provides a low resting pressure that increases on desirable activity. Immobile patients may do better with a frequently reapplied, long-stretch elastic bandage, such bandages should be reapplied over orthopaedic padding.

The control of infection—what should be put on the ulcer?

If the cause of ulceration is eliminated, then healing should take place. However, healing depends on healthy epidermis at the edge of the ulcer. Often this is damaged by proteolytic enzymes from slough and infection. Unhappily common is the damage resulting from irritation or sensitivity to medicaments. For this reason, simple bland therapy aimed at reducing debris should be used. Debris will float off if softened. Hard adherent crusts are usually dry. The most effective remedy is wetness: saline is perfectly adequate, provided it is applied very frequently as a wet dressing. Surgical debridement should be performed with a scalpel or scissors if there is any non-viable tissue. This is not difficult as long as it is remembered that dead tissue has no sensation. In other words, trim away anything the patient is unaware of, provided the diseases causing neuropathy (for example, leprosy and diabetes mellitus) have been excluded, in which case only necrotic and non-adherent tissue should be removed. Antiseptics such as eusol, 0.5 per cent acetic acid (a 5-ml teaspoon of vinegar in 500 ml (about a UK pint of water)), or 0.5 per cent silver nitrate in aqueous solution are other wetting agents; however, while helpful for removing slough, they inhibit granulation tissue and are less often recommended.

Many antibiotics are applied to ulcers and they rarely control infection. It is naïve to believe that an ulcer can be made sterile. Antibiotics are commonly and rapidly inhibited by serum and debris under a bandage. It is also common for such agents to do damage to the epidermis. In this respect, it is sometimes forgotten that the health of the surrounding epidermis is more important for healing than the state of the ulcer bed. Tropical phagedenic ulcers often follow trauma, and relative avascularity encourages invasion by fusospirochaetal organisms.

Contact dermatitis

Healing is often delayed and ulcers may be enlarged by damage to the surrounding epidermis. Such dermatitis, so often evident around the ulcer, occurs either because of medicaments, bacterial toxins, or allergy. Table 25 is a list of common causes of contact dermatitis.

Toenails

Poor sight, apathy, stiff hips, and obesity are all reasons why toenails are uncut. It is surprising how the Western world has come to expect 'professionals' to deal with this problem when, in fact, toenail cutting is something any good neighbour can do. Clippers rather than scissors are to be recommended because they cut or fracture hard, thickened nail more effectively. The nails should be softened by soaking them in warm water for 10 min. Only the distal part of the nail protruding beyond the toe needs be cut. Good positioning of the cutter and patient and adequate light are essential. Only very distorted nails or the foot with arteriosclerosis and the consequences of diabetes mellitus need the attention of a chiropodist, where such is available.

Corns

These are due to thickening of the epidermis due to external pressure, or from pressure from underlying bony prominences. It should be possible to avoid external pressure by making adjustments to footwear and skilful padding, so that the weight is taken on less bony areas of the foot. Surgery is sometimes necessary to remove bony prominences. Skin thickening is self-perpetuated, but is greatly helped by careful paring away of excess keratin, avoiding damage to underlying blood capillaries which often project upwards to near the surface.

Carcinoma

Carcinoma develops in the hypopigmented margin of ulcers that have persisted for many years. Such ulcers often invade bone but rarely metastasize. Local excision and grafting is often preferable to amputation.

Surgery

Whereas in technically developed countries, amputation and a good prosthesis may help a disabled person regain their mobility, amputation is objected to by certain races (for example, the Bantu) and religions (such as the Hindu). It inevitably causes them to be rejected so that they are unemployable and outcast. Consideration of amputation must take into account social circumstances and the degree of subsequent aftercare, both available and needed for successful rehabilitation.

Injection of superficial veins with sclerosants (or their removal) is indicated only when the deep veins are patent. If the deep veins are blocked, then the superficial veins are the only venous drainage of the legs and should be preserved. Assessment of the proportion of flow returned through the superficial veins is greatly facilitated by the Doppler flowmeter. Surgical debridement and skin grafting is often a means of quickly healing ulceration but is outside the scope of this textbook.

The decubitus ulcer

The decubitus ulcer is a consequence of tissue distortion, often due to pressure obstructing blood flow. It occurs especially in patients with neurological disease, where painful stimuli from tissue distortion or ischaemia is not recognized. Because the pathogenesis of decubitus ulcers includes impaired blood perfusion, anything that affects the blood supply can contribute to the problem. Thus, in general, old patients who are ill or dying, and especially if they have vascular disease, are most likely to develop sores. Such sores are unusual in those with a purely motor neurological disease and no sensory loss, or in the very old who have a healthy vascular system.

Because it is distortion of the tissues rather than simply pressure that induces sores, shearing forces on the sacral area and heels also need to be taken into account. Such forces are increased by moisture from sweating or incontinence. Distortion of the tissues is enhanced by deformities such as kyphoscoliosis or contractures.

While the basis of management is the relief of tissue distortion by the frequent relaxation of stresses and strains on the tissues, best brought about by movement, attention must also be paid to factors contributing to poor perfusion. These include intercurrent illness causing hypoxia, hypotension, immobility, dehydration, and impairment of consciousness or peripheral sensation. Most acute illnesses requiring admission to hospital provide the necessary criteria for the development of a decubitus ulcer within the first hours or days. The chronic sickness that determines prolonged bed rest at home rarely produces this degree of tissue ischaemia, and thus the hospital nurse gets blamed for what has never occurred at home. The blame is partly misdirected because, for example, an old woman with a fractured hip may develop decubitus ulcers while immobile in the ambulance, on a hard trolley waiting for a bed, or on the operating table. All attendants should be taught the causes of decubitus ulcer.

All ill patients are best nursed in bed. Some of the worst pressure sores can occur while sitting in a collapsed state in a day room or during the postoperative phase of 'mobilization'. The basis of management is regular shifting of position, which has arbitrarily become the practice of turning the patient every 2 hours or intermittent shifts in the surface on which the patient is lying. The combination of heavy and uncooperative patients, together with inadequate staffing levels, especially at night, often result in a failure to prevent bed sores. Good equipment to modify pressure on the mattress is important, and may include a fleece under the patient's heels and buttocks, a bed cradle to take the weight of the bedclothes, a variety of soft surfaces, and a ripple bed to provide alternating pressure, preferably one with large ripples.

In countries where such equipment is unavailable, there is often a large contingent of relatives at the bedside whose attentiveness can be mobilized to assist in frequently turning and massaging the patient.

A long, severe illness is difficult to manage outside an intensive care unit because, as implied by the word 'intensive', there is a requirement for vigilance of all aspects of the patient's physical, conscious, and activity level. It is for this reason that some nursing schools demand that a checklist, known as the Norton pressure sore score, is regularly completed. But even this becomes less accurate each day as a long illness drags on. Once an ulcer has developed it becomes a problem of wound healing. Removal of dead tissue is essential, often by surgical debridement. The development of granulation tissue and re-epithelialization will follow only if the patient's general health improves and their blood supply is adequate. There are so many agents advocated for the healing of wounds that it is important to realize that none are essential and many are harmful to healthy tissue. Those that will remove slough may inhibit living cells. Granulation tissue is the best protection against infection but is discouraged by many of the strongest antiseptics. All wound dressings should aim to reduce debris, to keep the wound moist, and to promote granulation tissue perfused by an adequate supply of normal blood.

Further reading

General

Champion RH, *et al*, eds (1999). *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn. Blackwell Scientific, Oxford.

Cotterill JA, Millard (1999). Psychocutaneous disorders. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 2715–814. Blackwell Scientific, Oxford.

Demis JD (1999). *Clinical dermatology*, 26th revision. Lippincott, Williams and Wilkins, Baltimore.

Fitzpatrick TB, *et al*. (1999). *Dermatology in general medicine*, 6th edn. McGraw-Hill, New York.

Goldsmith LA (1991). *Physiology and biochemistry of the skin*, 2nd edn. Oxford University Press.

Graham-Brown R, Bourke JF (1998). *Mosby's color atlas and text of dermatology*. Mosby London.

McKee PH (1999). *Pathology of the skin with clinical correlations*, 2nd edn. Mosby Wolfe, London.

Oumeish YO (1998). Environmental dermatology. *Clinics in Dermatology* 16, 1–184.

Panconesi E (1985). Stress and skin diseases. Psychosomatic dermatology. *Clinics in Dermatology* **2**, 1–282.

Parrish LC, Millikan LC (1994). Contemporary tropical dermatology. *Dermatology Clinics* **12**, 1–840.

Schaefer H, Redelmaier TE (1996). *Skin barrier, principles of percutaneous absorption*. Karger, Basel.

Weinstock MA (1995). Dermatoepidemiology. *Dermatology Clinics* **13**, 1–716.

The interview, examination, and investigations

Ackerman AB (1978). *Histologic diagnosis of inflammatory skin diseases*, p 863. Lea and Febiger, Philadelphia.

Ackerman AB (1995). *Resolving quandaries in dermatology, pathology and dermatopathology*, pp 1–327. William and Wilkins, Baltimore.

Ashton RE (1995). Teaching non-dermatologists to examine the skin: a review of the literature and some recommendations. *British Journal of Dermatology* **132**, 221–5.

Elder D, Jaworsky C, Johnson B, eds. (1997). *Lever's histopathology of the skin*, 8th edn, pp. 1–1073. Lippincott, Philadelphia.

McKee PH (1999). *Pathology of the skin with clinical correlations*, 2nd edn. Mosby Wolfe.

Weedon D (1992). *The skin: systemic pathology*, 3rd edn, Vol. 9, pp. 1–1095. Churchill Livingstone, New York.

Factors determining or modifying skin disease

Breathnach SM (1999). Drug reactions. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 3349–518. Blackwell Scientific, Oxford.

La Ruche G, Cesarini JP (1992). Histologie et physiologie de la peau noire. *Annales de Dermatologie et de Venerologie* **119**, 567–74.

Parish LC, Millikan LE (1994). *Global dermatology*. Springer-Verlag, New York.

Dermatitis

Burton JL, Holden CA (1999). Eczema, lichenification and prurigo. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 629–80. Blackwell Scientific, Oxford.

Holden CA, Parish WE (1999). Atopic dermatitis. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 681–708. Blackwell Scientific, Oxford.

Rietschel RL, Fowler JF (2001). *Fischer's Contact dermatitis*, 5th edn., pp. 1–862. Lippincott Williams and Wilkins, Philadelphia.

Rycroft RJG (1999). Occupational dermatoses. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 861–82. Blackwell Scientific, Oxford.

Wilkinson JD, Shaw S (1999). Contact dermatitis, allergic. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 733–820. Blackwell Scientific, Oxford.

Williams HC (2000). *Atopic eczema. The epidemiology, causes and prevention of atopic eczema*. Cambridge University Press, Cambridge.

Pruritus

Fleischer AB (2000). *The clinical management of itching.* The Parthenon Publishing Group.

Psoriasis

Camp RDR (1999). Psoriasis. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 1589–650. Blackwell Scientific, Oxford.

Acne

Cunliffe WJ, Simpson NB (1999). Disorders of the sebaceous glands. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 1927–84. Blackwell Scientific, Oxford.

Pigmentation

Behl PN (1994). *Asian clinics in dermatology*, Vol. 1, No. 1. Vitiligo Update. The Skin Institute, Greater Kailash, New Delhi.

Bleehan SS (1999). Disorders of skin colour. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 1753–816. Blackwell Scientific, Oxford.

Njoo MD (2000). Treatment of vitiligo. Thela thesis. University of Amsterdam.

Wasserman HP (1974). *Ethnic pigmentation.* Excerpta Medica, Amsterdam.

Nails

Baran R, Dawber RPR, eds. (1994). *Diseases of the nails and their management*, 2nd edn. Blackwell Scientific, Oxford.

Samman PD (1986). *The nails in disease*, 4th edn. Heinemann, London.

Hair

Rook A, Dawber RPR (1991). *Diseases of hair and scalp*, 2nd edn. Blackwell Scientific, Oxford.

Sinclair R, Banfield C, Dawber R (1999). *Handbook of diseases of the hair and scalp*. Blackwell Scientific, Oxford.

Skin disorders affecting the genitalia

Ridley, M. and Neill, S. (1998) *The vulva*. Blackwell Scientific, Oxford.

Vesicoblistering diseases

Wojnarowska F, Eady RAJ, Burge SM (1999). Bullous eruptions. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 1817–99. Blackwell Science, Oxford.

Angioma and telangiectasia

Colver GB, Ryan TJ (1996). Vascular disorders. In: Harper J, ed. *Inherited skin disorders*, pp 182–200. Butterworth and Heinemann, Oxford.

Mulliken JF, Young AE (1988). *Vascular birthmarks*. WB Saunders, Philadelphia.

Connective tissue

Burton JL, Lovell CR. (1999). Disorders of connective tissue. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 2003–171. Blackwell Science, Oxford.

Lymphoma

MacKie RM (1999). Cutaneous lymphomas and lymphocytoma. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 2373–402. Blackwell Scientific, Oxford.

Worret IF (1993). Skin signs and internal malignancies. *International Journal of Dematology* **32**, 1–5.

Warts

Storling JC, Kurt JB (1999). Human papilloma virus. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 1029–50. Blackwell Scientific, Oxford.

Granulomatous disease

Chu AC (1999). Histocytoses. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 2311–36. Blackwell Scientific, Oxford.

Cunliffe WJ (1999). Necrobiotic disorders. In: Champion RH, *et al.*, eds. *Rook, Wilkinson and Ebling textbook of dermatology*, 6th edn, pp 2297–310. Blackwell Scientific, Oxford.

Ryan TJ (1978). Lymphatics of the skin. In: Jarrett A, ed. *Physiology and pathophysiology of the skin* Vol 5, pp 1755–808. Academic Press, New York.

Management of skin diseases

Arndt KA (1995). *Manual of dermatologic therapeutics*, 5th edition. Little Brown, Boston.

Lebwohl MG, Heymann WR, Berth-Jones J, Coulson I (2000). *Treatment of skin diseases*, pp. 1–693. Mosby, London.

Maddin S, McClean D (1993). Dermatologic therapies. *Dermatology Clinics* **11**, 1–224.

Management of leg ulcers

Westerhof W (1993). *Leg ulcers. Diagnosis and treatment*. Elsevier, Amsterdam.

Decubitus ulcers

Parish LC, Witkowski JA, Crissey JT (1997). *The decubitus ulcer in clinical practice*, pp 1–241. Springer, Berlin.

US Department of Health and Human Services (1994). *Treatment of pressure ulcers*. Clinical practice guideline number 15. AHEPR Publication No 96 –0652.

23.2 Molecular basis of inherited skin diseases

Irene M. Leigh and David P. Kelsell

We would like to thanks Drs Colin Munro and John McGrath for clinical photographs and to Gareth Magee for help with Fig. 1.

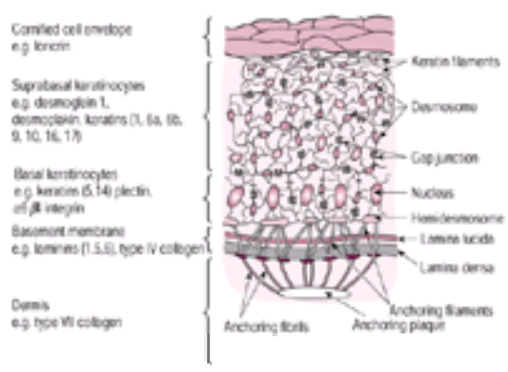

Fig. 1 A schematic representation of the epidermis indicating its organization, important structures, and site of expression of a number of skin-disease-associated proteins.

Introduction

Most patients referred by the general practitioner to the dermatology clinic will be seeking advice and treatment for a few common skin disorders including psoriasis, eczema, and acne. Although the genetic basis of these disorders is complex, putative susceptibility gene variants for some of these diseases have been identified, including the mast cell chymase gene in atopic eczema and the corneodesmin gene in the HLA-linked component of psoriasis vulgaris. In addition, several genes have been identified which predispose to malignancies of the skin including p16 and CDK4 in melanoma and the Patched gene involved in Gorlin's syndrome, in which mutation carriers have an elevated risk of basal cell carcinomas. There are a myriad of rarer epidermal disorders and syndromes for which the genetic basis has been elucidated. Here, we describe the current molecular understanding of these rarer traits which include blistering disease, ichthyosis, palmoplantar keratodermas, and the ectodermal dysplasias.

Structure of the epidermis

To understand the diseases it is first necessary to understand the basic biology of the epidermis and associated basement membrane zone. The basic structure of the epidermis is illustrated in Fig. 1. The epidermis is a stratified squamous epithelium comprised predominantly of keratinocytes. The remaining small percentage of intraepidermal cells are resident melanocytes, Langerhans cells, and migratory leucocytes. The keratinocyte undergoes a process of terminal differentiation which results in a stratum corneum, the critical component for the function of the epidermis as a barrier. It is a highly insoluble, non-viable cell made up of a cornified envelope enclosing keratin macrofibres separated by a highly lipid-rich, intercellular layer. This lamellated lipid is the predominant component of the barrier and is secreted into the extracellular space from membrane-coating granules synthesized in the stratum granulosum. The epidermis is separated from the underlying dermis by a complex basement membrane zone.

The basement membrane zone

When studied ultrastructurally, the basement membrane zone of the epidermis contains four distinct layers:

1. the basal cell membrane of the basal keratinocyte which contains electron-dense adhesion plaques called hemidesmosomes;
2. the electronlucent lamina lucida which is traversed by anchoring fibrils;
3. the electron dense lamina densa;
4. within the sublamina densa region the lamina fibroreticularis contains distinct anchoring structures called anchoring fibrils which insert into the lamina densa and loop around bundles of connective tissue collagens.

The hemidesmosome

Although ultrastructurally this organ appears to resemble desmosome morphology (see below), there are clearly differences in biochemical composition. Two major hemidesmosomal-associated proteins were identified initially by the characterization of autoantibodies arising in bullous pemphigoid, an autoimmune mechanobullous disease of late adult life. These proteins are known as bullous pemphigoid antigen 1 (230 kDa) and bullous pemphigoid antigen 2 (180 kDa). Bullous pemphigoid

antigen 2 has been identified as a unique transmembrane collagen, type XVII collagen, which has an extracellular domain containing the immunodominant epitope of bullous pemphigoid. Bullous pemphigoid antigen 1, like plectin, a further component of hemidesmosomes, is a member of the plakin family of proteins. These proteins have been thought to contribute to plaque structures within hemidesmosome. Other members of the plakin family, including desmoplakin, envoplakin, and periplakin, are found associated with the desmosome. Plectin and bullous pemphigoid antigen appear to interact with keratin intermediate filaments as they course towards the hemidesmosome and bind them into the hemidesmosome structure, acting as a protein clamp. This appears to provide a stable link between the intermediate filament cytoskeleton and the basement membrane zone. Basal keratinocytes also express a number of integrins which are a super family of receptors for extracellular matrix proteins. The major hemidesmosomal integrin is a6b4 integrin, although other areas of the basal cell membrane express a6b1, a5b1, a3b1, and a2b1 integrins.

The lamina lucida appears to contain a complex of laminin molecules, particularly laminins 5 and 6. It is thought that laminin 5 is the major component of the anchoring filament. The lamina densa is constructed of a meshwork of interacting type IV collagen. From this arise the anchoring fibrils of the basement membrane complex which are made of aggregates of antiparallel dimers of type VII collagen.

Keratins

The cytoskeleton of all epithelial cells contains a number of filamentous systems including actin, microfilaments, microtubules, and intermediate filaments. The protein characteristic of intermediate filaments of all epithelial cells is the keratin family of proteins. Keratin polypeptides segregate in two dimensional gel electrophoresis into acidic and basic polypeptides. Nineteen keratins can be identified by this procedure, with each protein being the product of an independent gene. The type II keratin gene family encodes the basic keratin polypeptides and the type I family the acidic keratin polypeptides. Each keratin is expressed in a body-site and cell-type-specific manner, for example K9 is only expressed in the suprabasal layer of the palmoplantar epidermis. The keratin genes are clustered in two chromosomal regions in the human genome: the type I keratins mapping to 17q12–q21 and the type II keratins to 12q11–q13.

A keratin filament comprises both type I and type II keratins. The fundamental building block of a keratin filament is a heterodimer, aligned along the length of its helical backbone, comprising four helical regions separated by non-helical linker regions with a non-helical head and tail domain. These heterodimers aggregate in a complex, antiparallel fashion to form the intact intermediate filament which associates with both hemidesmosomes and desmosomes to provide stability and integrity of the cell. In addition to keratin mutations associated with human disease, *in vitro* and transgenic models of keratin genes harbouring mutations have shown that there are critical regions for filament assembly; these are located particularly at the helix initiation and termination motifs. The function of the head and tail domains is not entirely clear.

Desmosomes

Desmosomal proteins form a complex structure which interfaces between adjacent epithelial cells. The desmosomal plaques of electron dense material run along the cytoplasm parallel to a junctional region in which three ultrastructural bands can be seen. The plaques contain plakoglobin (which is also found in adherens junctions and also thought to be important in cell signalling), desmoplakin, and plakophilin 1. In addition, the desmosomal cores are enriched in calcium binding glycoproteins, called desmogleins and desmocollins. The desmogleins and desmocollins are the adhesive proteins of the desmosome and are similar to the classical cadherins in their general structure, with five extracellular repeats that contain Ca^{2+}-binding sites, a single transmembrane region, and a cytoplasmic domain. To date, six human desmosomal cadherins have been identified and these are clustered in the chromosomal region 18q11–q12. The cytoplasmic domain has binding sites for plakoglobin, plakophilin 1, and desmoplakin, linking them to the intermediate filaments. Desmoglein 1 has been identified as the target antigen for the autoimmune bullous disease pemphigus variatious, and desmoglein 3 for pemphigus vulgaris.

Gap junctions

Gap junctions provide a mechanism of synchronized cellular response to a variety of intercellular signals by regulating the diffusion of small molecules (< 1 kDa), such as metabolites and ions, directly between the cytoplasm of adjacent cells. Connexins are the major proteins of gap junctions and these are encoded by a large gene family. All connexins have four transmembrane domains and two extracellular loops with the amino- and carboxy-terminus located in the cytoplasm. Each connexin assembles into hexameric hemichannels (termed connexons) in the endoplasmic reticulum and these are then transported into the lipid bilayer of the plasma membrane. A connexon then docks with a connexon of an adjacent cell to form a dodecameric, aqueous channel. These intercellular channels cluster in the cell to form the gap junctions. Connexons can form either homotypic or heterotypic channels, with various channel types having distinct molecular permeabilities. The majority of connexins have a wide tissue distribution. Those expressed in the skin include connexin 26, 31, and 43.

Diseases associations with the above structural proteins have provided a molecular classification of disease to complement the classical, morphological description of hereditary blistering diseases and disorders of keratinization, some of which are described below.

Epidermolysis bullosa

The genetic analysis of the heterogeneous group of mechanobullous disorders has given rise to enormous progress in understanding the function of proteins in the basement membrane at the dermo–epidermal junction and the role of keratins in the cytoskeleton (Table 1). The clinical phenotypes of epidermolysis bullosa correspond to different levels of separation within the basement membrane zone or basal keratinocyte, identified via electron microscopic examination. All cases of epidermolysis bullosa are skin disorders characterized by blistering of mucocutaneous sites following minor trauma, which are classified by a combination of laboratory and clinical criteria.

Epidermolysis bullosa simplex

In epidermolysis bullosa simplex, tissue separates within the basal keratinocyte, with or without aggregation of keratin intermediate filaments. This is the most common form of epidermolysis bullosa and is usually autosomal dominantly inherited. Mutations in the genes encoding the basal-cell-specific keratins K5 and K14 have been found to underlie epidermolysis bullosa simplex described, and these probably lead to cytoskeletal weakness and a tendency of cells to rupture on pressure. Three types of epidermolysis bullosa simplex are described below and clinical pictures are shown in Fig. 2 and Plate 1.

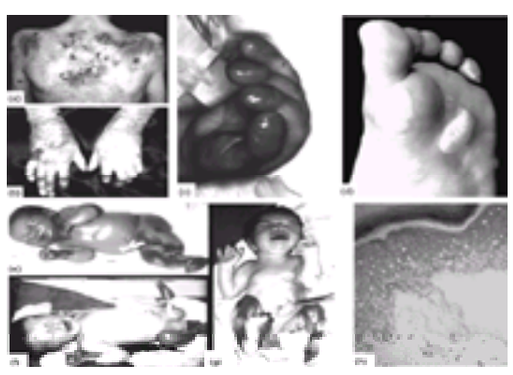

Fig. 2 Clinical photographs of the different forms of epidermolysis bullosa. (a) and (b) a patient with Hallopean–Siemans dystrophic epidermolysis bullosa; (c) the hand of an infant with Herlitz junctional epidermolysis bullosa; (d) blister on the foot of a patient with epidermolysis bullosa simplex?; (e) baby with epidermolysis bullosa simplex Dowling–Meara; (f) baby with Herlitz junctional epidermolysis bullosa; (g) baby with Hallopean–Siemans dystrophic epidermolysis bullosa; (h) intraepidermal blister from a Weber Cockayne epidermolysis bullosa simplex patient. Skin section stained with Richardson's stain. (See also Plate 1.)

Epidermolysis bullosa simplex Weber Cockayne

The soles and palms are mainly affected, rarely in other sites. The blistering occurs from infancy (with walking) and is exacerbated by heat and ameliorated by cold.

The blister heals without scarring.

Epidermolysis bullosa simplex Koebner

The blisters are widespread on scalp, trunk, arms, and legs in addition to the palmoplantar. These cases may represent autosomal recessive inheritance. Nail dystrophy, oral blisters, and dental caries are common.

Epidermolysis bullosa simplex Dowling–Meara (herpetiform epidermolysis bullosa simplex)

The blistering may be very severe and is potentially fatal in infancy. The blisters occur in groups on an erythematous bed, which heals without scarring but hyperpigmentation and milia formation may occur. Patchy keratoderma develops in later life.

Junctional epidermolysis bullosa

In junctional epidermolysis bullosa, the epidermis separates from the dermis through the lucida of the basement membrane zone. Most mutations lie within genes encoding the three subunit polypeptides of laminin 5 (*LAMA3*, *LAMB3*, *LAMC2*). Clinically this disease has been subdivided into two main categories—Herlitz (lethal) and non-Herlitz (non-lethal) forms.

Herlitz junctional epidermolysis bullosa

Blistering and erosions are present at birth and become widespread as the skin is so fragile that it peels away on contact. The resulting lesions are slow healing and tend to persist, becoming infected. The oropharyngeal mucosa is involved, often making feeding difficult. If the infant survives for a few months typical, crusted lesions will be seen on the nose, mouth, and jaw and across the rest of the skin in patches. The teeth have abnormal enamel and are lost easily, as are the nails. Infants usually die from overwhelming infection.

Non-lethal junctional epidermolysis bullosa

The patients show generalized skin fragility and blistering but mucosae are less severely affected. The lesions do heal leaving atrophic scars (generalized atrophic benign epidermolysis bullosa (GABEB)). Poor hair and tooth development occur and nails are dystrophic. Large hyperpigmented patches are seen.

Dystrophic epidermolysis bullosa

In recessive and dominant forms of dystrophic epidermolysis bullosa, separation occurs below the dermoepidermal region at the level of the anchoring fibrils and a large number of mutations have been discovered in the type VII collagen gene (*COL7A1*), which encodes the constituent protein of anchoring fibrils. In dystrophic epidermolysis bullosa, scarring and dystrophy are prominent features in addition to skin fragility and blistering.

Severe generalized recessive dystrophic epidermolysis bullosa

This is the most severe form of dystrophic epidermolysis bullosa and is very disabling in view of the deformities produced by scarring. Blisters are present at birth and recur readily at sites of trauma especially the hands, feet, neck, shoulders, and sacrum. They heal slowly with scarring and milia formation producing a 'mitten-like deformity' and clubbed feet. The severe oral lesions lead to microstoma and inability to protrude the tongue or open the mouth. Poor dentition leads to feeding problems. Scalp blistering and scarring gives permanent hair loss; eye involvement gives corneal erosions and opacities and general physical development is retarded. Oesophageal and perianal strictures lead to difficulty in swallowing and constipation. Although children often survive into adult life, multiple squamous cell carcinomas may develop in the chronically scarred skin and progress rapidly.

Dominant dystrophic epidermolysis bullosa

The skin is less fragile than recessive dystrophic epidermolysis bullosa and blistering much more difficult to provoke and so tends to be localized to bony prominences—knees, elbows, hands, and feet. Localized scarring with milia may replace the nails. Other areas (oral and anal) are much less affected.

Hemidesmosomal epidermolysis bullosa

Rarer forms of epidermolysis bullosa result from inherited defects in three hemidesmosonal components: plectin mutations in epidermolysis bullosa simplex with muscular dystrophy (epidermolysis bullosa simplex muscular dystrophy); type XVII collagen mutations in generalized atrophic benign epidermolysis bullosa; and a6b4 integrin mutations in epidermolysis bullosa with pyloric atresia.

Diagnosis and management of blistering in childhood

Early diagnosis is the key to management and prediction of prognosis. Diagnosis of a baby born with blisters is often difficult on clinical grounds, so diagnosis will rest on electron microscopy of a shave skin biopsy. Immunohistochemistry is likely to be indicative in recessive cases of gene knockout—LH7.2 antibody to type VII collagen and GB3 to laminin 5 being diagnostic reagents. There is no specific treatment for any form of epidermolysis bullosa so the management centres on wound care, avoidance of physical trauma, and general physical and psychological support. Specialist nurses can advise on nursing babies with silk-covered dressing pads and vaseline gauze dressings. Oral hygiene and dental care needs to be life long. High calorie and fibre diet is essential to improve growth. Gastrostomy feeding can also help maintain body weight in children unable to eat. Finger and hand contractures require splinting at night and regular surgical release by an expert surgeon. Now that the genetic basis of epidermolysis bullosa has been identified, prenatal diagnosis by DNA-based techniques and gene therapy by *ex vivo* techniques are being actively explored.

Hailey–Hailey disease and Darier's disease

The genetic basis if these two rare, autosomal dominant, intraepidermal blistering diseases have recently been elucidated and shown to be due to mutations in calcium pumps, *ATP2C1* in Hailey–Hailey disease and *ATP2A2* in Darier's disease.

Ichthyoses

Ichthyoses manifest as dry, rough skin with persistent scaling over most of the body which may resemble fish scales (ichthys, Greek fish). Congenital ichthyosis may be bullous or associated with other abnormalities (ichthyosiform syndromes). Ichthyosis can be acquired in later life, due to drugs such as hypercholesterolemic agents, chronic hepatic disease, lymphoma and other malignancies, thyroid disease, chronic renal hepatic failure, and malabsorption. It is sometimes difficult when a patient's ichthyosis has improved in adult life and then worsened again in late adult life to be absolutely sure whether they have a congenital or acquired ichthyosis. The progressive understanding of the molecular and cellular biology of the ichthyoses will aid in establishing their classification and potential treatment.

Autosomal dominant ichthyosis vulgaris

This commonest ichthyosis is associated with atopic eczema in up to 50 per cent of individuals. The condition improves in teenagers and young adult life and often worsens again with age. The clinical features present with dryness and scaling in the neonatal period but becoming progressively more obvious in childhood. Scaliness is small, flaky or brawny, and is most pronounced on the extensor arms and lower legs. Facial involvement is often minimal, although patients may have dandruff and they have increased markings on the palms and soles. Hyperlinearity of palm creases may be seen. It is usually very well tolerated symptomatically, with the dryness and roughness only being a problem. Treatments have therefore been aimed at removing the keratotic, retained stratum corneum with keratolytic agents such as salicylic acid (1–5 per cent lactic acid) and other hydroxy acids or buffered urea creams.

Histopathology shows hyperkeratosis with a diminished or absent granular cell layer but otherwise very little abnormality at both light and ultrastructural level. This disease is inherited as an autosomal dominant. However, genetic analysis has been hampered by variable penetrance and difficulties in ascertainment. The molecular basis of autosomal dominant ichthyosis vulgaris has centred on studies of filaggrin, the filament aggregating protein expressed in the keratohyalin granules in the stratum granulosum as a precursor protein, profilaggrin. Studies have shown alterations in the expression of profilaggrin although no disease-associated mutations have been identified in the gene encoding this protein. Therefore the genetic defect(s) underlying ichthyosis vulgaris may be in proteins involved in the synthesis or the degradation of profilaggrin rather than in the profilaggrin gene itself. Further studies are awaited.

X-linked recessive ichthyosis

This disorder is much less common than the autosomal dominant form and predominantly affects the male children of female carriers. The scaling is absent usually within the first week of life and progressively increases. The scaling tends to be prominent on the arms, thighs, and lower leg and very large, adherent, brown scaling may involve the flexures and the face. The pathology shows a normal granular cell layer and normal keratohyalin granules ultrastructurally. The molecular basis of this form of ichthyosis was derived from observations of low urinary oestriol secretion in the third trimester of pregnancy and the presence of reduced steroid sulphatase activity. Subsequently, the steroid sulphatase gene was mapped to the X chromosome and disease-associated mutations in this gene have been identified in the vast majority of patients. Steroid sulphatase mutations lead to the abnormal breakdown of cholesterol sulphate in the stratum corneum lipids, resulting in an increase in stratum corneum thickening. A small proportion of patients will have other manifestations of Kallman's syndrome, hypogonatrophic, hypogonadism, and neurological abnormalities. These are due to contiguous gene defects, usually a large deletion on the short arm of the X chromosome encompassing the steroid sulphatase locus.

Bullous ichthyosiform erythroderma or epidermolytic hyperkeratosis

This is a rare, autosomal dominant ichthyosis. At birth, there is a mild erythroderma and, at sites of minor trauma, blisters and peeling may occur. Large areas of denuded skin are often apparent after a difficult birth. In infancy, a yellow-brown hyperkeratosis develops, particularly at sites of joint flexures with cobble-stone keratoses present on the hands, feet, and trunk. Ridged scale may accumulate in skin creases, which are highly susceptible to bacterial and/or fungal infections leading to a pungent body odour. Histologically, there is lysis and clumping of the keratin filaments in the granular layer of the epidermis. Intercellular spaces are often apparent due to the rupture of suprabasal keratinocytes. Immunohistochemical studies revealed the specific aggregation of the suprabasal keratins of the epidermis, Keratin 1 and 10. Subsequently, mutations in either K1 or K10 have been shown to underlie the disease in many bullous ichthyosiform erythroderma patients. A clinical photograph of the feet of a bullous ichthyosiform erythroderma patient is shown in Fig. 3(a), and Plate 2.

Fig. 3 Clinical photographs of: (a) bullous ichthyosiform erythroderma (BIE) and three types of keratoderma: (b) focal palmoplantar keratoderma (PPK) associated with a keratin 16 mutation; (c) striate palmoplantar keratoderma associated with a desmoglein 1 mutation; and (d) constriction around the digit from an individual with Vohwinkel's syndrome associated with a Cx26 mutation. (See also Plate 2.)

Ichthyosis bullosa of Siemans

This is a rarer form of bullous ichthyosis. Neonatal disease is much milder with episodic, superficial blistering occurring throughout childhood, sometimes into adulthood. This blisters occur mainly on the flexures, lower limbs, and abdomen. At these sites, a rippled, grey hyperkeratosis may occur. Plate-like caling and a focal peeling (Mauserung) are usually found. There is an absence of palmoplantar keratoderma and erythroderma. Mutations in another suprabasal keratin, K2e, have been identified as the genetic basis of this condition. This type II keratin is expressed in many of the higher suprabasal keratinocytes.

Netherton's syndrome

Netherton's syndrome is a severe, autosomal recessive disorder, often resulting in infant mortality, which is characterized by ichthyosis with erythroderma and trichorrhexis invaginata (hair shaft abnormalities often termed 'bamboo' hair). Hair anomalies are a characteristic feature seen in patients with Netherton's syndrome. From scanning electron microscopy, plucked hairs from patients often display trichorrhexis invaginata, tortion nodule, pili torti, and trichorrhexis nodosa. Also, light microscopy reveals invaginated hair cuticle into the cortex. Recently, a genome scan of families with Netherton's syndrome mapped the disease to a locus on chromosome 5q32. Mutations in the gene encoding SPINK5, a serine protease inhibitor, have recently been identified.

Sjörgen–Larsson syndrome

Sjörgen–Larsson syndrome is inherited as an autosomal recessive trait and is particularly prevalent in north-western Sweden (1 in 10 000), occurring less frequently elsewhere. The syndrome characteristics include ichthyosis, spastic diplegia, and mild to moderate mental retardation. The skin disease presents as mildly erythrodermic at birth with scaling developing in the first few months. These persist with scaling particularly of the face and limbs. In the flexures, neck, and periumbilical folds, an orange/brown lichenification overlaid with hyperkeratosis is a characteristic of Sjörgen–Larsson syndrome. In early life, neurological defects, including upper motor neurone defects of the limbs, mental disability, and often ocular abnormalities (spots on the retina), are observed. Histologically, the affected skin displays orthohyperkeratosis, acanthosis, and papillomatosis. The genetic defect has been shown to be in the fatty aldehyde dehydrogenase (*FALDH*) gene which affects essential fatty acid metabolism.

Keratodermas

The inherited keratodermas are characterized by thickened skin on the palms and soles. Palmoplantar skin is specialized for high levels of weight bearing and friction, so the stratum corneum is much thicker (hyperkeratotic) than the rest of the epidermis. Keratodermas can be classified clinically according to the pattern of thickening on the palm and sole skin. Three distinct clinical patterns have been observed:

- diffuse—the hyperkeratotic thickening is evenly and symmetrically distributed over the palm and sole; it is usually manifest at birth;
- focal—hyperkeratotic plaques develop particularly at sites of weight bearing and friction; these are usually plaque-like callosites or linear thickening (striate keratoderma);
- punctate—multiple, bead-like keratoses which pepper the palmoplantar skin.

They can have autosomal recessive or dominant inheritance and may occur in syndromes. Keratodermas can be further subgrouped into:

- simple—palmoplantar involvement only;
- complex—associated with lesions of non-volar skin, hair teeth, nails, and sweat glands (including ectodermal dysplasias);

- syndromic—with associated abnormalities of other organs including deafness, cancer, cardiomyopathy, and adrenal insufficiency.

They can also be classified biologically by their recently-discovered genetic defects, for example see Fig. 3 and Table 2 and Table 3. This branch of the genodermatoses is genetically heterogeneous with mutations in genes encoding keratins, desmosomal proteins, connexins, and a protease.

Diffuse palmoplantar keratodermas

Simple: diffuse epidermolytic palmoplantar keratoderma (EPPK)

EPPK is characterized by epidemolytic hyperkeratosis with keratin filament clumping in suprabasal keratins. This autosomal dominant disease presents with symmetrical thickening, giving a cracked, crocodile-skin-like surface due to the underlying epidermolysis, which starts in early infancy. The majority of EPPK pedigrees are linked to the type I keratin cluster on chromosome 17q12–q21 and disease is due to mutations in the palmoplantar-specific keratin, K9, the majority clustering in the helix initiation domain of the protein.

Simple: diffuse non-epidermolytic palmoplantar keratoderma (NEPPK)

NEPPK is also inherited as an autosomal dominant trait and is often difficult to distinguish clinically from EPPK, due to the variability of finding epidermolysis by electronmicroscopy in EPPK. There is a waxy, uniform, yellow thickening over the palms and soles which may spread onto the dorsum of hands and wrists with a sharp cut-off. It is commonly aggravated by secondary fungal infection, which may require intermittent, oral antifungal agents. These often improve the keratoderma. In a number of families the defect has been linked to 12q11–q13. A single family has a mutation in the variable head domain of KRT1. However in the majority of NEPPK families, fine mapping of the 12q11–q13 locus has excluded the type II keratin genes, and a number of candidate genes in the area including a keratinocyte-expressed elastase have been excluded by sequencing.

Complex: erythrokeratoderma variabilis

Erythrokeratoderma variabilis is a rare, autosomal dominantly inherited skin disease characterized by diffuse palmoplantar keratoderma and transient, figurate, red patches at various sites and severity. Germline mutations in connexin 31 (*GJB3*) and connexin 30.3 (GJB4) have been identified in the affected members of some erythrokeratoderma variabilis families.

Focal keratoderma

Most focal palmoplantar keratodermas are characterized by discoid lesions and the majority can be regarded as complex palmoplantar keratodermas as they are often associated with abnormalities of hair, nails, teeth, and glands.

Simple: striate palmoplantar keratoderma

This focal palmoplantar keratoderma is characterized by distinctive linear streaks on palms and soles, over the ventral aspects of fingers and extending onto palms; it is often more extreme on the feet. Variable nail and hair involvement, with fragility or splitting, are seen. Recently, mutations in two desmosomal proteins, desmoglein 1 (18q11–12) and desmoplakin 1(6p21), have been described which result in a hemizygous gene knockout, resulting in haploinsufficiency of the gene product.

Complex: pachyonychia congenita type 1/ focal palmoplantar keratoderma with oral mucosal hyperkeratosis

This clinical overlap syndrome presents in childhood with nail changes (pachyonychia), typically a subungual hyperkeratosis trumpet-shaped nail, especially on the thumb and first finger and toe nails. The sole lesions are painful callosites over weight-bearing areas on the feet, with less prominent callosities on the palms. The mucosa shows milky hyperkeratosis over gingiva. Nutmeg-grater-like follicular keratoses also occur. Variable fragility and blistering can be associated with severe pain on walking. Milder nail involvement shows splinter haemorrhages at the onychocorneal bind. The pathological findings of epidermolytic hyperkeratosis with keratin filament clumping suggested that keratin gene mutations underlie this disorder, confirmed by the identification of mutations in *KRT6a* and *KRT16* in affected individuals from multiple families.

Complex: pachyonychia congenita type 2/ steatocystoma multiplex

The steatocysis palmoplantar keratoderma may be very limited although pachyonychia nail changes present early. Multiple epidermal cysts and steatocysis are seen. Woolly scalp hair and fuzzy eyebrows are seen and natal teeth can occur. The finding of keratin clumps in skin bearing keratin K17, particularly deep outer root sheath, suggested *KRT17* as the candidate gene and autosomal dominant mutations in hot spots have now been described. Mutations in *KRT6b* have also been described. The resulting pathology varies from keratin cysts, vellous hair cysts, to oil filled cysts.

Complex: Papillon–Lefevre syndrome

This focal palmoplantar keratoderma is inherited as an autosomal recessive and is marked by associated severe periodontitis and loss of primary and secondary dentition with opalescent oral mucosa. The inflammatory lesions often result in pocket formation seen pathologically. Linkage to 11q14 was demonstrated in a number of families. Recently, mutations in cathepsin C, a lysosomal protease, have been shown to underlie this disorder. It is postulated that cathepsin C may be important in the processing of key structural proteins, such as keratins, in the epidermis.

Syndromic keratodermas (multiple phenotypic)

Palmoplantar keratoderma and deafness

A number of families with palmoplantar keratoderma and sensorineural deafness have been described, which could have been due to a mutation of a single gene expressed in all affected tissues or cosegregation of two distinct gene mutations. One such disorder is Vohwinkel's syndrome; a mutating palmoplantar keratoderma with constrictions developing around and autoamputating fingers. In a small family with a Vohwinkel's pattern of keratoderma and profound sensorineural deafness, two distinct mutations in the gene encoding the gap junction protein connexin 26 (*GJB2*) were identified. One of the mutations was associated with the profound deafness, the other (*D66H*) segregated with the skin disease. This led to the important discovery that mutations in *GJB2* were causative in both autosomal dominant (DFNA3) and recessive (DFNB1) deafness, accounting for 40 to 60 per cent of hereditary sensorineural deafness world-wide. In other families with Vohwinkel's syndrome, the same palmoplantar keratoderma-associated mutation D66H in *GJB2* has been identified. Mutations in loricrin, a cornified cell envelope component of the stratum corneum, have also been identified in individuals affected with a variant form of Vohwinkel's syndrome, which is associated with ichthyosis and normal hearing. In addition, palmoplantar keratoderma and deafness has also been associated with mitochondrial mutations.

Palmoplantar keratodermas and cancer

Focal NEPPK and oesophageal cancer

Three pedigrees from United Kingdom, United States, and Germany have been studied in which a focal NEPPK with oral hyperkeratosis segregates with a high lifetime risk of squamous cell carcinoma of the oesophagus (40–91 per cent by age 70). In all three families, linkage to DNA markers mapping to 17q24–q25 has localized the disease gene to a region of less than 1 cM. The cornified envelope protein, envoplakin, lies in this region but has been genetically excluded as the candidate gene.

Huriez disease (sclerotylosis)

This is an autosomal dominant disease characterized by palmoplantar keratoderma, nail changes, and scleroatrophy of the distal extremities. Around 15 per cent of individuals develop aggressive squamous cell carcinomas, occurring in their thirties and forties. It is proposed that the scarring resulting from skin fragility predisposes

to the squamous cell carcinomas. Recently, the Huriez disease gene has been mapped to chromosome 4q23.

Punctate palmoplantar keratoderma

A weak association between punctate palmoplantar keratoderma and cancer has been observed in a family with a number of epithelial-derived tumours developing in members under the age of 50. As yet, the punctate palmoplantar keratoderma locus has not been genetically mapped though a number of candidate regions, such as the keratins, have been excluded by linkage.

Other palmoplantar keratoderma syndromes

Diffuse NEPPK, woolly hair, and arrhythmogenic venticular cardiomyopathy (which leads to heart failure and arrhythmias) is due to recessive mutations in plakoglobin. Recessive mutations in another desmosomal protein, desmoplakin, underlie a similar syndrome consisting of a striate PPK, woolly hair, and dilated left ventricular cardiomyopathy. In triple A or Allgroves syndrome, individuals have adrencorticotrophic-hormone-resistant adrenal insufficiency, achalasia, and alacrimia with a PPK due to mutations in aladin, a member of the WD-repeat family of regulatory proteins.

Ectodermal dysplasias

There are a very large number of ectodermal dysplasias which display abnormalities of the skin, hair, teeth, nails, and/or sweating. The clinical lassification is unsatisfactory but may become more transparent when the genetic basis of a significant number of these complexes have been classified. At present, there is limited genetic understanding. Two major subgroups are hidrotic and non-hidrotic ectodermal dysplasia. The concept of dysplasia in these diseases is developmental rather than premalignant.

Hidrotic ectodermal dysplasia (Clouston's syndrome)

Hidrotic ectodermal dysplasia is characterized by nail dystrophy with thick, slow growing, and discoloured, short nails. Diffuse palmoplantar keratoderma is variable but may be severe and spread to knuckles and finger joints. Scalp hair is sparse, fine, pale, and brittle with thin eyebrows and sparse body hair. The disease is inherited as an autosomal dominant. Disease in a large kindred from Canada was mapped to 13q11–12 which harbours a connexin gene cluster. Mutations in the gene encoding connexin 30 underlie this disorder.

Hypohidrotic ectodermal dysplasia

This X-linked, recessively inherited disease is characterized by a loss of sweat glands causing absent or reduced sweating (hypohidrosis) and total or partial loss of teeth. Patients may be very uncomfortable on exertion and are heat intolerant. The teeth are characteristically conical and the mouth dry. In severe forms, the facial appearance is altered with saddle nose, sunken cheeks, and sparse, dry, fine, short hair with absent eyebrows. The disease maps to Xq12–13.1 and is caused by mutations in the ectodysplasin anhidrotic protein. An autosomal recessive form is due to mutations in the ectodysplasin receptor.

Concluding comments

This chapter has focused on the rarer types of genetic skin diseases rather than the more common, genetically complex disorders such as eczema, psoriasis, and acne. This is largely because only a few potential disease-associated genetic variants have been identified with the more common epidermal disorders. In contrast, great advances have been made in understanding the molecular basis of the rarer blistering diseases, ichthyoses, and the keratodermas, with the identification of a number of important proteins involved in epidermal and also non-epidermal biology. In addition, these studies have revealed genetic heterogeneity with mutations in different proteins causing similar clinical manifestations, for example Vohwinkel's syndrome can be due to mutations in either connexin 26 or loricrin. With the imminent completion of the sequencing of the entire human genome, the capability for high throughput genotyping using to high density single nucleotide polymorphism (SNP) maps and new technology development, it is likely that the genetic basis of the more common epidermal disorders will be elucidated in the near future.

Further reading

Review papers

Aumailley M, Rousselle P (1999). Laminins of the dermo-epidermal junction. *Matrix Biology* **18**, 19–28.

Corden LD, McLean WHI (1996). Human keratin diseases: hereditary fragility of specific epithelial tissues. *Experimental Dermatology* **5**, 297–307.

Kelsell DP, Stevens HP (1999). The palmoplantar keratodermas: much more than palms and soles. *Molecular Medicine Today* **5**, 107–13.

Ruhrberg C, Watt FM (1997). The plakin family: versatile organizers of cytoskeletal architecture. *Current Opinions in Genetics and Development* **7**, 392–7.

Uitto J, Pulkkinen L, McLean WH (1997). Epidermolysis bullosa: a spectrum of clinical phenotypes explained by molecular heterogeneity. *Molecular Medicine Today* **3**, 457–65.

White TW, Paul DL (1999). Genetic diseases and gene knockouts reveal diverse connexin functions. *Annual Review Physiology* **61**, 283–310.

Significant research papers

Armstrong D *et al* (1999). Haploinsufficiency of desmoplakin causes a striate subtype of palmoplantar keratoderma. *Human Molecular Genetics* **8**, 143–8.

Chipev CC *et al* (1992). A leucine-proline mutation in the H1 subdomain of keratin 1 causes epidermolytic hyperkeratosis. *Cell* **70**, 821–8.

De Laurenzi V *et al* (1996). Sjogren-Larsson syndrome is caused by mutations in the fatty aldehyde dehydrogenase gene. *Nature Genetics* **12**, 52–7.

Hilal L *et al* (1993). A homozygous insertion-deletion in the type VII collagen gene (COL7A1) in Hallopeau-Siemens dystrophic epidermolysis bullosa. *Nature Genetics* **5**, 287–93.

Hu Z *et al* (2000). Mutations in ATP2C1, encoding a calcium pump, cause hailey-hailey disease. *Nature Genetics* **24**, 61–5.

Kelsell DP *et al* (1997). Connexin 26 mutations in hereditary non-syndromic sensorineural deafness. *Nature* **387**, 80–3.

Kere J *et al* (1996). X-linked anhidrotic (hypohidrotic) ectodermal dysplasia is caused by mutation in a novel transmembrane protein. *Nature Genetics* **13**, 409–16.

Lamartine J, *et al*. (2000). Mutations in GJB6 cause hidrotic ectodermal dysplasia. *Nature Genetics* **26**, 142–4.

Lane EB *et al* (1992). A mutation in the conserved helix termination peptide of keratin 5 in hereditary skin blistering. *Nature* **356**, 244–6.

Maestrini E *et al* (1996). A molecular defect in loricrin, the major component of the cornified cell envelope, underlies Vohwinkel's syndrome. *Nature Genetics* **13**, 70–7.

McKoy G, *et al*. (2000). Identification of a deletion in plakoglobin in arrhythmogenic right ventricular cardiomayopathy with palmoplantar deratoderma and woolly hair (Naxos disease). *Lancet* **355**, 2119–24.

McLean WHI *et al* (1995). Keratin 16 and keratin 17 mutations cause pachyonychia congenita. *Nature Genetics* **9**, 273–8.

Norgett EE, *et al*. (2000). Recessive mutation in desmoplakin disrupts desmoplakin-intermediate filament interactions and causes dilated cardiomyopathy, woolly hair and keratoderma. *Human Molecular Genetics* **9**, 2761–6.

Reis A *et al* (1994). Keratin 9 gene mutations in epidermolytic palmoplantar keratoderma (EPPK). *Nature Genetics* **6**, 174–9.

Richard G *et al* (1998). Mutations in the human connexin gene GJB3 cause erythrokeratodermia variabilis. *Nature Genetics* **20**, 366–9.

Rickman L *et al* (1999). N-terminal deletion in a desmosomal cadherin causes the autosomal dominant skin disease striate palmoplantar keratoderma. *Human Molecular Genetics* **8**, 971–6.

Rothnagel JA *et al* (1994). Mutations in the rod domain of keratin 2e in patients with ichthyosis bullosa of Siemens. *Nature Genetics* **7**, 485–90.

Sakuntabhai A *et al* (1999). Mutations in ATP2A2, encoding a Ca2+ pump, cause Darier disease. *Nature Genetics* **21**, 271–7.

Smith FJD *et al* (1996). Plectin deficiency results in muscular dystrophy with epidermolysis bullosa. *Nature Genetics* **13**, 450–6.

Toomes C *et al* (1999). Loss-of-function mutations in the cathepsin C gene result in periodontal disease and palmoplantar keratosis. *Nature Genetics* **23**, 421–4.

Vidal F *et al* (1995). Integrin beta 4 mutations associated with junctional epidermolysis bullosa with pyloric atresia. *Nature Genetics* **10**, 229–34.

Yen PH *et al* (1987). Cloning and expression of steroid sulfatase cDNA and the frequent occurrence of deletions in STS deficiency: implications for X-Y interchange. *Cell* **49**, 443–54.

24.1 Introduction and approach to the patient with neurological disease

Alastair Compston

Clinical neurology uses intuitive conversation, structured examination, and selective investigation to formulate problems into an anatomical and pathological framework. The competent neurologist instinctively senses relevant components of the history, appreciates the most likely underlying mechanism, reliably elicits the physical signs, knows which investigations are necessary and relevant, and communicates the situation accurately and sensitively. This system has evolved over several centuries during which knowledge has accumulated on structure and function, localization in health and disease, the reliability of physical signs and laboratory investigations, and the nosology of disease.

The neurological history

Although patients usually start with an account of that which troubles them most, the neurologist prefers a history of the components in the order in which they occurred. It may take some time to establish this chronology. The first task is to assess the core symptoms and how they cluster. The neurologist asks enough questions to settle whether, for example, a reported episode of difficulty with speech refers to a disturbance of language (aphasia) or articulation (dysarthria); whether there are motor or sensory deficits in a 'heavy' limb; whether alterations of sensation are positive (tingling and paraesthesiae) or negative (numbness) symptoms; whether a disturbance of bladder function suggests neurological or urological disease; and whether double vision actually refers to diplopia or altered acuity. Some questions reflect the peculiarities of neurological anatomy; it may surprise the patient complaining of impaired vision on the right that the symptom is in fact unaltered by sequential closure of either eye—because it is hemianopic; or that awareness of temperature and the appreciation of pain may be disturbed in the 'good' leg in some forms of spinal cord disease (the Brown–Sequard syndrome).

Once the individual symptoms are accurately defined, they can be grouped and from this follows an interpretation of their anatomical basis suggesting the involvement of one or more sites. Recognizing these patterns is fundamental to interpretation of the neurological history and this synthesis directs attention to specific components of the subsequent examination. It is easy to conclude that the patient with cognitive impairment has disease of the cerebral cortex but a more detailed history will additionally indicate whether this is diffuse or focal and reflects involvement of the dominant or non-dominant hemispheres and the frontal, temporal, or parietal cortices. Inco-ordination of more than one motor skill (eye movement, speech, the limbs, and balance) necessarily indicates involvement of brainstem–cerebellar connections. The process causing a hemianopic field defect lies above and that resulting in lower cranial nerve palsies below the tentorium. The combination of motor and sensory symptoms in the limbs with altered sphincter function indicates spinal cord disease; for the male patient with an unreliable bladder, the significance of linking urgency and frequency to impotence and constipation may seem strange. In turn, the coexistence of diffuse distal symmetric motor and sensory symptoms, shoulder and pelvic girdle weakness, or ocular, bulbar, respiratory, and upper limb weakness steers the thinking towards peripheral nerve, primary muscle, and neuromuscular junction disease respectively.

As a generalization, abrupt events are vascular or electrical in origin; subacute symptoms are demyelinating or inflammatory; and symptoms which develop slowly suggest structural deficits or degeneration. The subsequent course also reveals the underlying process; self-limiting events are often vascular; paroxysmal symptoms tend to be electrical or demyelinating, depending on their duration; and progressive syndromes are compressive or degenerative. The circumstances may be suggestive of a particular pathophysiology: trauma, preceding infection, drug exposure, or pregnancy alert the observer to structural, demyelinating, toxic, and venous thrombotic mechanisms respectively. Dangerous for the beginner but nevertheless important to recognize are the inconsistencies of exaggeration, mismatch between the severity of symptoms and altered function, and anatomical impossibilities which usually feature in non-organic neurological disease. Together, these pattern recognitions are the stuff of neurological diagnosis.

The neurological examination

Examination of the patient with neurological disease needs to be structured and organized without exhausting the patient and examiner through obsessive attention to irrelevant detail. Much can be learned by astute observation without formal assessment. Gross defects of cognition do not need to be confirmed by reciting telephone numbers in reverse or assembling lists of former prime ministers; defects of speech will usually be evident in conversation; many neurological diagnoses are immediately apparent from the patient's gait; movement disorders can be observed whilst taking the history. That said, it is best routinely to adopt a basic core examination and do things in order since the detection of one abnormality will determine the interpretation of another. It takes only a few minutes for the experienced and adequately equipped examiner to confirm that corrected visual acuity is normal in each eye, that there is no gross field defect, and that the optic fundi are normal. Although more detailed assessment will sometimes be necessary, a full range of smooth following (pursuit) eye movements in the horizontal and vertical planes can rapidly be established: this will detect obvious ophthalmoplegia and can be supplemented by cover testing of each eye during fixation on the examiner's nose, and rapid gaze from right to left—very few significant defects of eye movement will escape this rapid screen. Movement of the lower face during forced eye closure, voluntary elevation of the palate, and rapid protrusion or side-to-side movement of the tongue take a few seconds to observe and effectively cover all the lower cranial nerves. It is rarely necessary to test the sense of smell or hearing and a tuning fork is most useful for establishing that deafness is conductive and therefore probably not relevant. Before moving to the limbs, it is worth testing neck flexion in patients where the history suggests muscular or neuromuscular disease.

A sufficient routine examination of the arms would start with posture (outstretched with the eyes open and then closed); a quick look for selective muscle wasting; tone in flexion–extension and supination–pronation at the elbow and wrist respectively; strength in flexion and extension at the elbow and wrist, spreading the fingers and abduction of the thumb; co-ordination during movement between the patient's nose and examiner's finger (or both hands if there is gross inco-ordination so as to avoid accidental ocular injury); and the tendon reflexes. This will take the experienced examiner less than a minute. It may be necessary to establish specific patterns of muscle weakness: global loss affecting the hand in cortical disease; selective involvement of extensor groups in upper motor neurone disease; the patterns of C5–T1 nerve root lesions; diffuse distal weakness of both extremities in peripheral neuropathy; and the subtle distinctions between radial, median, and ulnar neuropathies and C7, C8, and T1 root lesions respectively. Detailed sensory examination of the arms rarely achieves more than can be learned from establishing that crude protective sense (recognition of a sharp pin) or discrimination (position sense and the ability to distinguish two points or perform a simple task such as manipulating a button) are intact.

Although this may involve some rearrangement of clothing, it otherwise takes almost no time to swipe the abdominal reflexes in passing, before examining the legs. Here, the structured motor examination is as for the arms although increased tone is more easily detected by lifting the relaxed leg from the couch at the thigh, and testing internal and external rotation at the hip. Characteristic patterns of weakness are the involvement of flexors at all joints and eversion at the ankle in upper motor neurone lesions; the usual diffuse symmetrical distal involvement in peripheral neuropathy at a time when the hands may be normal; and difficulty in distinguishing injury of the lateral popliteal nerve from an L5/S1 root lesion (in which the ankle jerk is lost) in the context of unilateral foot drop. Proximal weakness is best detected by watching the patient walk, and the calf muscles are normally so strong as to be untestable except with the patient standing. As in the arm, co-ordination can only be assessed once the degree of weakness is established. Tendon reflexes in the legs may be brisk in isolation and often spread, so that in an upper motor neurone lesion when one is tapped several may respond—and in either leg. Even non-neurologists rarely forget to elicit the plantar responses.

Sensory examination of the legs tends to be more reliable for protective than discriminative sensation. In mapping a sensory level it is best to move from the relatively anaesthetic to the normal zone noting the band of hypersensitivity which usually exists at the boundary. It is a matter of fact that many patients confuse the examination by exaggeration or elaboration of physical signs; this most usually affects power and the usual clues are a mismatch between the ability to walk and findings on formal assessment of muscle strength (or vice versa) and simultaneous contraction of agonist and antagonist muscles. Sensory testing is subjective and so necessarily vulnerable to inaccurate reporting, but confirming that a sensory level is present both on the abdomen and back, and on the same side on each with a slightly higher level on the trunk, is a simple manoeuvre which may yield surprising discrepancies in the patient with non-organic deficits.

The overall purpose of the history and examination is to assess where and through what mechanism structure and function have been affected. Detecting these patterns becomes routine for the experienced neurologist but the process represents more than just a ritual of clinical neurology. From anatomical localization follows

a formulation of likely mechanisms and pathological conditions underlying the patient's symptoms and signs.

Investigation of neurological disease

The investigation of patients with neurological disease was revolutionized in the early 1970s with the introduction of computed axial tomography. Before then, only the most primitive structural details of the central nervous system could be detected by demonstrating indirectly the shape and placement of the ventricles and blood vessels, and usually at some discomfort to the patient. Function in the central and peripheral nervous systems was measured using neurophysiological techniques. Disruption of the blood–brain barrier and immunological activity in the central nervous system were assessed through analysis of the cerebrospinal fluid.

Investigation still does not replace clinical assessment but, as the sections which follow make clear, it is now possible to detect structural changes in most parts of the brain and spinal cord at high resolution; to distinguish many pathological appearances at these sites on the basis of differences in the magnetic resonance signals; to map function within regions of interest using changes in blood flow and the use of metabolic substrates; to show variations in efferent and afferent electrical activity in the central and peripheral nervous systems; and to detect an increasing range of soluble mediators of normal and pathological function in the cerebrospinal fluid. Taken together, these laboratory investigations still do no more than supplement clinical assessments and, in one sense, the high expectations of diagnosis make for additional difficulties in interpreting neurological illness when the images are normal compared with the era when authoritative statements from neurologists could never be validated and necessarily went unchallenged.

The value of many routine investigations lies in confirming normality and endorsing abnormalities already strongly suspected on clinical grounds. Given the increasing sensitivity of techniques for brain imaging, altered appearances which are not necessarily of pathological significance and genuine lesions which are not relevant in the particular clinical context need to be interpreted with common sense. Overall, the trend has been for the pendulum to swing from diagnosis without adequate laboratory evidence to diagnosis made in defiance of clinical intuition. Even when an imaging abnormality has been identified, its nature may require clinical discussion in order to resolve the most likely pathological substrate—the distinction between ischaemic and inflammatory tissue often proving difficult and not all neoplastic tissue being easily identified as such.

The management of neurological disease

The first issue that confronts the doctor looking after a person with neurological disease is when to discuss and name the diagnosis. Most wait until there is sufficient clinical or laboratory evidence to rule out misdiagnosis; telling people that they have a condition when they do not is bound to cause distress and has landed some specialists in the law courts. However, overcaution and avoidance of discussion can be equally damaging and there are many more patients who harbour bitterness over delay in learning the true nature of their illness than those who wish they had not been told so soon, or at all. The majority of individuals cope extremely well even with the prospect of conditions which are known to be life threatening or have a poor prognosis for disability. Advice may be needed on alterations in lifestyle resulting from neurological disease—for example driving in epilepsy, and the use of drugs in pregnancy. There is a basic human need to know why a thing has happened and most patients enquire about causation but, naturally, the uppermost question is whether symptoms can be treated or the natural history of disease usefully modified.

The chapters which follow document specific treatments for particular conditions but judgement is often required in deciding whether to deploy these remedies depending on age, significance of the symptoms for the individual, level of disability, security of the diagnosis, adverse effects, and the patient's own views. Drug treatment may be used, on an intermittent or regular basis, to suppress symptoms—for example, intravenous methyl prednisolone to educe inflammation, anticonvulsants to suppress epilepsy, g-aminobutyric acid agonists to deal with spasticity, or anticholinesterases to enhance transmission at the neuromuscular junction. Pharmacological options also exist for interfering with the mechanism of disease, again on an intermittent or routine basis—such as the use of triptans to relieve migraine, or the replacement of dopamine in Parkinson's disease. In other situations, the rationale of treatment is to modify the underlying disease process, for example by suppressing inflammatory processes in acute postinfectious polyneuritis using intravenous gammaglobulin, treating patients with multiple sclerosis using b-interferon, and using immunosuppressants such as methotrexate and cycophosphamide in polymyositis and vasculitis respectively. Many other illustrations could be given confirming that the age-old witticism concerning the therapeutic nihilism of clinical neurology is at best now only of historical interest and was always generally rather ill-informed. Beyond the present pharmacological achievements in drug treatment lie many opportunities for improving handicap and disability through the use of rehabilitation which increasingly assumes centre stage in the management of neurological disease through attention to the person with impairments in a particular social and cultural setting rather than focusing on the pathophysiology of disease in an individual void. For the future, there is the prospect of enhanced regeneration in the context of diseases affecting the central and peripheral nervous systems, restoring structure and function and thereby both limiting and repairing the damage.

24.2 Electrophysiology of the central and peripheral nervous systems

*Christian Krarup**

Introduction

In clinical neurophysiology, the core investigations in electrophysiological studies of the central nervous system (**CNS**) and peripheral nervous system (**PNS**), comprise electroencephalography (**EEG**), evoked potentials, electromyography (**EMG**), and nerve conduction studies. However, since these provide no direct information about pathological changes, it is often necessary to supplement findings by imaging or other laboratory studies, and it is mandatory to view the results in a clinical context. Furthermore, electrophysiological parameters provide information about changes over time obtained from various anatomical regions that may not be accessible to direct pathological examination.

Additional methods (including cardiovascular reflexes in the study of the autonomic nervous system, respiratory movements and oxygen saturation in polysomnographic studies of sleep disturbances, and recording of force in the study of voluntary muscle) are becoming increasingly important in clinical neurophysiology, recognizing that electrophysiological methods must often be supplemented by other investigations.

Electroencephalography (EEG)

At EEG the spontaneous ongoing activity from the cerebral cortex is recorded through electrodes placed over the scalp. In most routine studies, recordings are carried out over 30 to 60 min. In addition, specialized studies may be performed to diagnose patients with particular types of epilepsy, during carotid artery endarterectomy, or brain death. In patients with poorly described epileptic fits, the clinical features may require that both visual information and EEG evidence are obtained simultaneously (video-EEG). This may, in patients who are candidates for surgical treatment of medically intractable epilepsy, be carried out over many days. In some patients, additional information may be obtained with intracranial subdural or intracerebral depth electrodes. During epilepsy surgery, an electroencephalogram is recorded directly from the cortex, so-called electrocorticography.

Indications

The main indications for obtaining an EEG include paroxysmal events, convulsions, disturbed levels of consciousness, and neuroinfections. EEG is not well suited as a screening procedure in patients with suspected focal cerebral lesions, since deep-seated lesions show no abnormalities at EEG if the cortex itself or its afferent projections are unaffected. However, when the clinical picture in patients with focal brain lesions is complicated by periodic changes in consciousness, convulsions, or unexplained changes in focal weakness, EEG is necessary to establish the presence of secondary paroxysmal events. Furthermore, EEG is often indicated in patients with encephalopathy to ascertain whether the clinical features are complicated by additional ictal discharges.

Serial EEGs are often necessary to assess the prognosis in patients with diffuse brain lesions. When abnormalities obtained early during a cerebral disorder (for example, cardiac arrest associated with cerebral ischaemia) are followed by further deterioration of the EEG pattern, a poor prognosis is indicated.

Method

The recording takes place with the patient in a comfortable position in a quiet room. After placing surface or needle electrodes over the scalp according to an international, standardized system (the 10–20 system Fig. 1), the technician ensures that the impedance of the electrodes is less than 5 kW. The patient's age, clinical state, and medication is indicated on the record, in particular whether the level of consciousness is normal. The session includes recordings while the patient is awake, during activation procedures, and, if possible, during drowsiness and sleep. During the recording the technician makes notes about the patient's awareness and state of consciousness, but fully describes any events that occur during the recording.. Activity is evaluated at different electrode montages, including both bipolar and unipolar recordings. Bipolar recordings are of value for localizing abnormalities in focal brain areas, whereas unipolar recordings are necessary for examining more widespread and generalized disturbances.

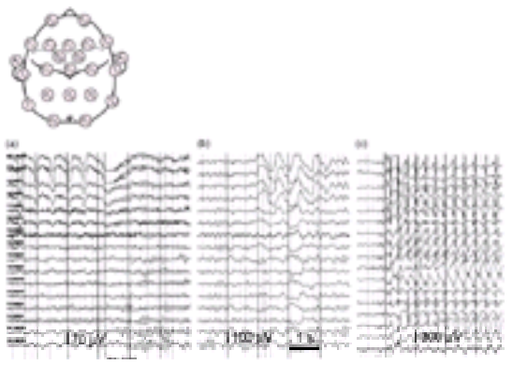

Fig. 1 Examples of EEG curves recorded using a common reference recording montage. The top panel shows the electrode placement using the 10–20 international system (with permission from Niedermeyer and Lopes da Silva). (a) EEG from a 20-year-old normal man (aviation candidate). Eye blinking was carried out to the left of the stippled line. At the stippled line he closed his eyes and the 10-Hz background activity became prominent, mainly over the posterior regions. (b) A 72-year-old female with progressive gait abnormalities, dementia, and urine incontinence. The CT scan showed cerebral atrophy and hydrocephalus. The EEG showed high-amplitude delta waves over the frontal regions. The background activity was slowed to 7 Hz. (c) An 8-year-old boy with absences. The EEG showed generalized 3-Hz spike–wave paroxysms.

During recording the awake patient is asked to relax with closed eyes to assess the background activity. The EEG waveforms are characterized by summated continuous postsynaptic de- and hyperpolarizations of large numbers of cortical cells by input from other brain areas and are, on an empirical basis, described in

terms of their frequencies. The frequency contents of the EEG are classified into activity with frequencies of 8 to 13 Hz (a-activity), 3.5 to 7.5 Hz (q-activity), 3 Hz or less (†-activity), and activity above 13 Hz (b-activity). Interpretation of the EEG should include a description of the background activity, the presence of abnormal wave forms ('transients') during rest and activation procedures, and whether any changes in the background or the occurrence of abnormal waveforms occur diffusely, synchronously, or in a focal pattern (Fig. 1(b), Fig. 1(c), and Fig. 2). Advanced algorithms to localize the distribution of waveforms (brain mapping) are now used for both diagnostic purposes and research on epileptic and non-epileptic syndromes, and they are of particular relevance in the temporal and spatial development of transient abnormalities.

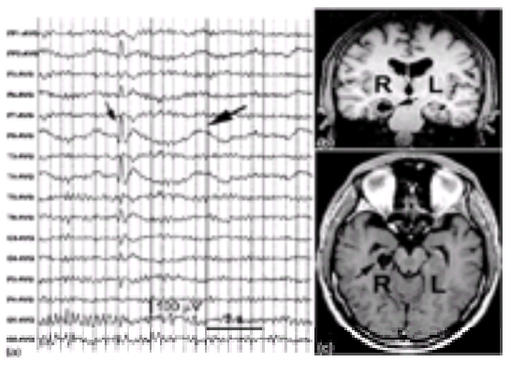

Fig. 2 A 31-year-old man with a history of complex partial seizures. (a) The EEG showed a spike focus over the right pre- and midtemporal regions (small arrow). In addition, there was a slow wave (1–2 Hz) focus over the same regions (large arrow), highly suspicious of a focal brain lesion. An average reference electrode montage was used. (b) Coronal *T*1-weighted MRI after contrast injection. (c) Transverse reconstructed section. Arrows indicate the site of a cystic-ring enhancing lesion in the right hippocampal region. (MRI by courtesy of the Danish Research Centre for Magnetic Resonance, Hvidovre Hospital, Copenhagen University Hospital.)

Activation procedures include hyperventilation: here the patient breathes deeply at a rate of 20/min for 2 to 4 min and changes are followed up to 2 min after hyperventilation. In children and young adults, this may elicit q- and †-activity (activities considered to be abnormal in individuals over 30 years of age), while patients with absences may develop spike-and-(sharp)-wave patterns during hyperventilation (see Fig. 1(c)). The possible epileptogenic effect of photic stimulation is evaluated by stimulating with variable frequencies when the eyes are opened and closed. In susceptible individuals, spike activity limited to the occipital regions is not associated with epilepsy, but spikes or sharp waves in a more widespread distribution are indicators of a lowered epileptogenic threshold.

The normal EEG

The frequency content in the normal subject is highly dependent on age, the level of awareness, and medication. In the normal awake adult with closed eyes, the EEG is dominated by a-activity. The a-activity is most pronounced over posterior parts of the head (Fig. 1(a)), and is subject to modulations by changes in vigilance; for example, it disappears when the subject opens the eyes.

In the newborn, the EEG is characterized by low-frequency activity and variable amplitudes. In premature children, the EEG may be dominated by burst-suppression activity which does not occur in the normal, full-term baby. During maturation, the background frequencies move into the a-range by the early teens. Even in normal young adults, intermittent posterior slowing may be seen over the occipital regions, which becomes enhanced and spreads to other regions during hyperventilation. This slow activity during hyperventilation is augmented by low glucose levels in the blood, so that glucose should be given by mouth to subjects with excessive amounts of slow activity during hyperventilation.

During drowsiness the a-activity is diminished and disappears, first intermittently and subsequently completely, to be replaced by q-activity (stage-1 sleep). During stage-2 sleep (light sleep), sleep spindles (bursts of 12–14 Hz activity), sharp waves over the vertex, and K-complexes (high-amplitude, slow-wave activity) are recorded in addition to q-activity. At deeper levels of sleep (stages 3 and 4), increasing amounts of high-amplitude †-activity are recorded (often designated 'delta sleep'). The EEG may be badly misinterpreted if drowsiness is not recognized during the recording session, therefore the technician must monitor the level of awareness at all times.

The abnormal EEG

The abnormal EEG may be characterized by changes in the background activity, the presence of low-frequency waveforms, epileptiform activity, or by periodic phenomena. The EEG is evaluated for the presence of abnormal frequencies or wave forms, either intermittently or constantly, whether these are localized diffusely or focally, and whether they occur in a synchronous or an asynchronous distribution. Preservation, distortion, or loss of normal background patterns are evaluated.

Abnormal frequencies

Slowing of the normal background activity occurs diffusely in patients with encephalopathy (for example, ischaemic or metabolic brain disease) or degenerative brain disease (for instance, Alzheimer's disease). Focal slowing (see Fig. 2) and attenuation of background activity is highly suggestive of focal brain disease (for example, stroke, tumour, or subdural haematoma).

Generalized, diffuse, and focal abnormalities

Generalized abnormalities occur synchronously throughout the brain, though the amplitudes and wave forms may vary at different recording sites (see Fig. 1(c)). Diffuse abnormalities are also present over large brain areas, but the low-frequency activity or spikes/sharp waves may occur independently. It should be considered if the generalized changes occur in recordings with a single reference electrode, since this may erroneously give rise to the impression of generalization.

These abnormalities are considered to have a central origin if the generalization occurs from the onset, but they may also be due to a focal cortical lesion if generalization occurs as a secondary phenomenon. So-called intermittent rhythmic †-activity (Fig. 1(b)) may occur over widespread areas of the brain due to raised intracranial pressure and have little localizing value. Diffuse low-frequency abnormalities, often associated with triphasic waves, indicate widespread cortical abnormalities in metabolic encephalopathies.

Spikes, sharp waves, and periodic complexes

The central role of EEG in the diagnosis and follow-up of patients with epilepsy justifies the attention paid to the identification and localization of epileptic discharges. The features characteristic of epileptiform events consist of waves of various forms, usually spikes (potential duration, 70 ms or less) or sharp waves (potential duration, 70 to 200 ms) in a rhythmic pattern, that are of high voltage compared to the background activity and reflect hypersynchronization of neuronal discharges (Fig. 1(c)). Spikes may be followed by a negative wave, the so-called 'spike–wave complex'. Since it is unusual for an epileptic seizure to coincide with the EEG, the diagnosis therefore relies on the presence of epileptiform discharges during interictal recordings: the examination at the first EEG may be negative in up to 50 per cent of patients. Repeat studies or prolonged recordings (possibly under video control) are frequently indicated, and proper activation procedures, such as hyperventilation, photic stimulation, and possibly sleep deprivation or the use of sedatives to ensure sleep during the study, may be needed. The diagnostic yield of repeated EEG studies has accordingly been found to show abnormalities in more than 90 per cent of patients with a clinically established diagnosis of epilepsy.

Paroxysmal discharges may be focal or generalized in distribution according to the underlying aetiology. Recently developed focal epileptic symptoms (Fig. 2) may be evidence of a brain tumour and should be thoroughly investigated with appropriate imaging studies. Epileptic activity may develop abruptly in patients with primary generalized seizures (Fig. (c)). It is, however, important to evaluate this development closely to distinguish primary from secondary seizures that develop focally and then spread to adjacent cerebral regions, and possibly with generalization to the whole brain. Detection of focal epileptic activity may require specialized electrode montages. For example, an epileptic focus in the temporal lobe may require recording through electrodes placed over the zygomatic arch or through a needle

sphenoidal electrode placed at the foramen ovale. Such focal epileptic activity may, moreover, not become apparent until the patient becomes drowsy or goes to sleep.

The electrophysiological activity is usually not unambiguous for subgroups of epileptic seizures, though the particular combination of clinical characteristics, the EEG changes during seizures, and the interictal activity may be distinctive for epileptic syndromes, hence the term 'electroclinical diagnosis' has been coined. Generalized 3-s spike–wave complexes are considered pathognomonic for generalized absence seizures (petit mal, Fig. 1(c)), and hypsarrhythmia (high-voltage, irregular slow waves interspersed with spikes) occurs almost exclusively in infantile spasms. Periodic, lateralized epileptiform discharges (**PLED**) give a trace of continuous focal spike activity with a frequency of 0.5 to 3 s, seen in connection with acute severe brain disease. Periodic generalized complexes of sharp waves are characteristically seen in patients with Creutzfeldt–Jakob disease, herpes simplex encephalitis, and subacute sclerosing panencephalitis, and may be present in patients with severe brain anoxia.

Evoked potentials

Evoked potentials are specific CNS potentials obtained by stimulation of particular sensory receptors or fibre tracts and are carried out to examine the integrity of afferent and efferent pathways. The sensory pathways routinely examined include the visual system (visual-evoked potentials, **VEPs**), fibre tracts in the brainstem (brainstem auditory-evoked potentials, **BAEPs**), and somatosensory pathways in the dorsal columns (somatosensory-evoked potentials, **SSEPs**). Motor-evoked potentials (**MEPs**) are elicited by magnetic stimulation of the motor cortex and used to study the corticospinal tracts. Methodological questions and pathophysiological findings are discussed below; however, a detailed description of the methods used are considered beyond the scope of this chapter and the reader is referred to the Further reading list.

Near-field and far-field responses

The responses discussed in this chapter are the modality-specific components of the evoked potentials (**EPs**) that reflect the propagation of action potentials in fibre tracts and cortical areas. The so-called 'event-related' potentials, although time-locked to a stimulus, are not modality specific but reflect the activity in neuronal networks involved in cognitive processing (for example, P300), they will not be described further.

The electrical responses recorded close to the source are the so-called 'near-field potentials', which may arise from axons or be of postsynaptic origin. These include action potentials recorded from peripheral nerves, the spinal cord, or cortical areas. However, the activity recorded from scalp electrodes with a non-cephalic reference also reflect activity in deeply located structures, known as 'far-field potentials'. The origin of a number of these EP components is incompletely known, and the latencies and amplitudes of only some of these are of clinical relevance in routine practice.

Indications

The main purpose of evoked-potential studies is to ascertain the presence of pathological processes localized to myelinated fibre tracts or to the synaptic connections through which messages are relayed. The conduction velocity of the fibres is gauged by the latencies of the responses, and these are particularly susceptible to abnormalities in the myelin sheath. Hence, evoked potentials are of particular use in demyelinating disorders, for example multiple sclerosis. However, conduction abnormalities are not specific for a particular disease, and delayed conduction may be seen in a variety of disorders including hereditary diseases, compression of nervous tissue (such as spondylotic myelopathy), and infectious diseases (such as in patients with the acquired immunodeficiency syndrome, **AIDS**). Thus, the constellation of EP abnormalities, the clinical setting, and other paraclinical or laboratory findings are all important factors in a diagnosis.

The amplitudes of responses are influenced by the number of conducting fibres; in disorders characterized by fibre loss without involvement of the myelin sheath, abnormalities may be confined to a reduction in amplitude. However, because of the amplification that occurs at synaptic relays, the amplitude of the evoked potential is a poor indicator of the degree of fibre loss. Thus, a cortical response may still be recordable at SSEP testing in patients with a severe neuropathy and an absent peripheral nerve response. Finally, conduction may be delayed in disorders characterized by axonal loss, possibly related to delays at synaptic transmission. Thus, in patients with amyotrophic lateral sclerosis (**ALS**), the MEP recording often shows a delayed central conduction time even though the disorder is characterized by fibre loss rather than demyelination.

SSEP and MEP investigations have proved to be valuable intraoperative monitoring tools during surgery on the vertebral column for scoliosis and on the spinal cord for tumours or vascular malformations. Both a reduction in amplitudes and a prolongation of latencies have proved reliable indicators of impending damage to the spinal cord, and hence the need to take measures to avoid permanent damage. Additionally, the use of electromyographic recordings from relevant muscles is helpful during scoliosis operations in warning the surgeon that a root may be in danger of damage from screws or other hardware.

Visual-evoked potentials

Method

Visual-evoked potentials (VEPs) are evoked by either a diffuse stroboscopic flash (flash-VEP) that stimulates the whole retina or by pattern-reversal stimulation (pattern-VEP), where a black-and-white checkerboard reverses position at a frequency of 2/s. The patient is seated at a distance of 1 m and gazes at the centre of the checkerboard projected either on a screen or a TV monitor. The pattern-VEP is sensitive to the co-operation of the patient, whereas the flash-VEP can be used to ascertain whether functional connections exist between the retina and the occipital lobes. The size of each square is either 9 mm or 18 mm, depending on the visual acuity. The pattern-VEP is generated mainly by the central 10° of vision. The responses are recorded over the occipital lobes. At least 100, and preferably 200, sweeps are averaged to yield responses of adequate resolution.

Each eye is stimulated in turn, and in routine studies the whole visual field is stimulated. In some conditions, however, it is more revealing to stimulate part of the visual field: in which case, the half-field is usually stimulated. Half-field stimulation is particularly useful in conditions where lesions are localized in the visual projections behind the chiasm, but its interpretation requires considerable expertise.

Measurements

The pattern-VEP comprises three main components, of which the positive phase at a latency of about 100 m is the most constant (Fig. 3). In clinical practice, the latency of this phase and the amplitude of the response are measured. In flash-VEP, the latencies of the N70, P90, and N120 phases are measured.

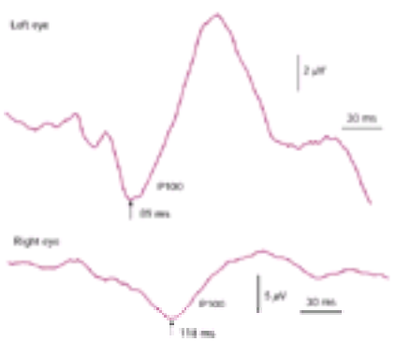

Fig. 3 Pattern-reversal, visual-evoked potentials (checkerboard stimulation) from a 64-year-old man showed a normal pattern-VEP with a P100 latency of 89 ms from the left eye and an abnormal pattern-VEP from the right eye with a latency of 118 ms (27 per cent prolonged, upper normal limit 103 ms). The findings indicated the presence of a right-sided optic nerve lesion.

Clinical correlations

Flash-VEP is reduced in patients with retinal disease. It is a useful test in patients with retinosa pigmentosa, and in those who cannot co-operate, in particular in children.

Monocular, full-field, pattern-VEP with prolonged latency to the P100 component in one eye indicates that the lesion is localized anterior to the optic chiasm (Fig. 3). Although this is most frequently due to demyelination of the optic nerve, it may be due to retinal degeneration, optic nerve compression, or glaucoma. In some patients with mild optic neuritis, the latencies only show an abnormal interocular difference. The interpretation is more uncertain if bilateral prolonged latencies are found, since this may be due to lesions anywhere along the visual pathways. Interocular differences in patients with bilateral retrochiasmal lesions (for example, spinocerebellar syndromes) are within the normal range. Marked latency differences in patients with bilateral abnormalities suggest bilateral optic nerve lesions. Retrochiasmal lesions may be further examined by partial-field (half-field) studies of the individual eye.

Electroretinography

The electroretinogram (**ERG**) is the electrical response evoked in the retina by a flash and is due to depolarization of the interstitial Müller cells and pigmented epithelium. The ERG is recorded by a contact-lens electrode or with an infraorbital surface electrode. The state of light and dark adaptation can be used to separate the function of the rods and cones. The ERG is usually carried out when the VEP pattern is missing and it is uncertain whether this may be due to a retinal problem. In such cases, the ERG may be evoked by a routine flash. Dark-adapted ERG is carried out in the differential diagnosis of retinal degenerations and is a specialist task usually carried out in collaboration with a neuro-ophthalmologist.

Brainstem auditory-evoked potentials (BAEPs)

Method

The auditory brainstem response is elicited by passing short-lasting clicks, at an intensity of 75 to 100 dB, through earphones to each ear separately. The responses of interest include the time-locked far-field responses with latencies of less than 10 ms. The brainstem-derived response consists of several phases (usually numbered as positive waves PI–VI), which indicate conduction along peripheral pathways in the cochlear nerve and different relay stations of the lateral lemniscus pathway within the brainstem.

Measurements

The waves of interest are the positive peaks PI to PVI; PI, PIII, and PV are usually recorded, whereas the remaining waves may be missing even in normal subjects. PI is generated in the cochlear nerve, PII in the cochlear nucleus, PIII in the pons, and PV at the midbrain level. For clinical purposes, the latency of PI is measured to ascertain peripheral conduction and that of PI to PIII, PI to PV, and PIII to PV to ascertain the central conduction time within the brainstem.

Clinical correlations

BAEP recording is helpful in investigating the integrity of the brainstem; it is used to confirm brain-death in some laboratories. Its main usefulness lies in the localization of lesions at the cochlear nerve, at the entry into the brainstem (cochlear nucleus), and at different sites within the brainstem. Central abnormalities are found in 50 per cent of patients with multiple sclerosis.

Somatosensory-evoked potentials (SSEPs)

Method

These responses are evoked by repetitive stimulation (2–5/s) of the median nerves at the wrists and the tibial nerves behind the medial malleolus (Fig. 4), using a stimulus duration of 0.2 ms at a strength just sufficient to elicit a slight motor response. Differentiation between peripheral and central disease is obtained after stimulation of the median nerve by recording peripheral nerve responses through surface or subcutaneous needle electrodes at the supraclavicular fossa (Erb's point, designated the N9 response), from the spinal cord at C6 (designated the N13), and over the contralateral hemisphere (the potential is designated N20). On stimulation of the tibial nerves, recordings are carried out from the peripheral nerve at the popliteal fossa (or at the gluteal fold), at Th12 (designated the N23), and over the brain (onset response and P40 response). Up to 1000 responses are averaged depending on the level of noise and the size of the response. The average is carried out in two bins to ensure reproducibility.

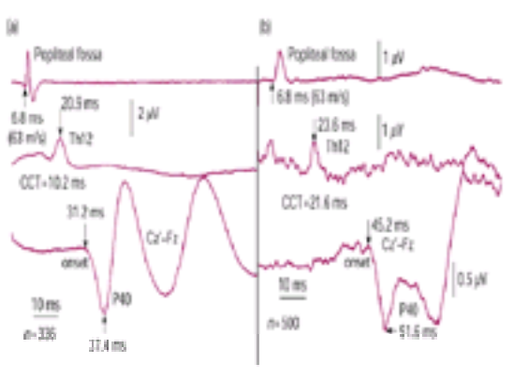

Fig. 4 Somatosensory-evoked potentials from (a) the right leg of a 25-year-old normal woman (a) and (b) the left leg of a 55-year-old man with signs of myelopathy. The tibial nerve was stimulated at the medial malleolus and responses recorded from the peripheral nerve at the popliteal fossa, from the spine (Th12), and the scalp. In both subjects, the peripheral conduction velocities and spinal latencies were normal. The latencies of the cortical responses to both the 'onset' and the P40 were normal in (a) and 31 per cent and 25 per cent prolonged, respectively, in (b). The central conduction time (CCT) was calculated as the difference between the spinal latency and the onset latency. The central conduction time was normal in (a), whereas it was 98 per cent prolonged in (b), consistent with myelopathy.

Measurements

The latencies to the onsets of the peripheral nerve responses are measured to calculate the peripheral conduction velocities (Fig. 4), to the negative peak of the spinal response, and to the first negative response at the cortex after the median nerve and to the onset and the P40 responses after tibial nerve stimulation.

The central conduction time is calculated as the differences in latencies between the spinal responses at C6 and the peak of the N20 response after median nerve stimulation, and between the spinal responses at Th12 and the onset latency (or the P40 latency) after tibial nerve stimulation. The values are compared to height-matched controls.

Clinical correlations

SSEPs from median and tibial nerves are helpful when very proximal nerve disorders or central nervous system disorders are suspected. A prolonged latency of the spinal responses evoked from the tibial nerves indicates the presence of proximal lumbosacral plexus or root lesions, and may be differentiated from a peripheral neuropathy by a normal peripheral nerve response at the popliteal fossa or the gluteal fold. Similar information is obtained regarding the brachial plexus and cervical roots at C6 from median nerve stimulation. SSEP studies may be extended by dermatomal stimulation in the legs and the arms to diagnose monoradicular lesions.

SSEPs are, however, particularly useful for identifying spinal cord disease; the central conduction time obtained separately from the upper and lower limbs may be used to determine the probable localization of myelopathic lesions (Fig. 4(b)). The central conduction time often shows marked prolongation in patients with multiple sclerosis.

Motor-evoked potentials (MEPs)

Motor-evoked potentials are obtained by activating focal motor cortical areas by a short-lasting, strong magnetic pulse of up to 2 tesla, which induces a current within the excitable tissue of the cortex. In some laboratories, electrical stimulation rather than magnetic stimulation is employed. However, electrical stimulation is painful; moreover, the electrical stimulus activates fibres deeper within the cerebrum than does the magnetic stimulus. The two methods therefore yield results that cannot be directly compared. The descending waves from the cortex consist of D-waves from cortical neurones followed by I-waves that arise transynaptically. In addition, stimulation is performed at cervical and lumbar spinal levels. At magnetic stimulation, excitation occurs at the proximal spinal nerves rather than at the spinal cord.

The motor responses are recorded from muscles of the upper and lower limbs (including proximal and distal muscles) using surface electrodes to evaluate the compound muscle action potential (**CMAP**). Facilitation of cortical neurones by slight voluntary contraction is necessary for obtaining 'maximal' motor responses.

Measurements

The amplitudes and latencies of the CMAPs are measured at both cortical and spinal stimulation (Fig. 5). The central conduction time is obtained by calculating the differences of these latencies; however, since excitation at spinal stimulation occurs at the proximal peripheral nerve, the central conduction time includes conduction along the roots. The central conduction time has therefore also been calculated using the F-wave latency to obtain a measure of the peripheral conduction time. In central lesions, the central conduction time is prolonged compared to height- and age-matched controls (Fig. 5(b)). In addition, the CMAP amplitudes of the cortical response may be reduced, indicating central axonal loss, conduction failure, or increased temporal dispersion along corticospinal fibres. The shape of the CMAP recording in patients with multiple sclerosis is often polyphasic, indicating dispersion along demyelinated central pathways.

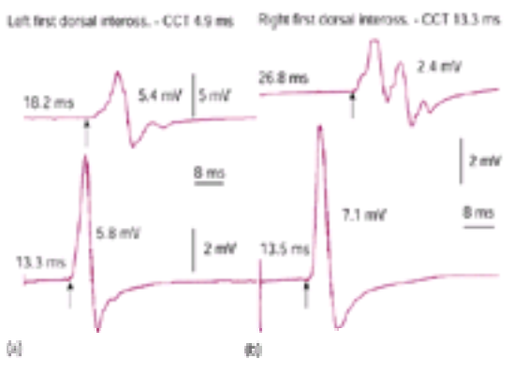

Fig. 5 Motor-evoked potentials obtained by magnetic stimulation from (a) the left and (b) the right first dorsal interosseous muscles in a 60-year-old woman suspected of having multiple sclerosis (the pattern-VEP and the SSEP were also abnormal). Lower traces: compound muscle action potentials (CMAPs) evoked by stimulation of the cervical spine. Upper traces: CMAPs evoked by cortical stimulation. The latencies of the responses (shown above the traces) at spinal stimulation were normal on both sides, whereas cortical latency was normal in (a) and 36 per cent prolonged in (b). The central motor conduction time (CCT, indicated above traces) was calculated as the difference between the cortical and peripheral latencies. The central conduction time in the left arm was normal in (a), whereas it was 86 per cent prolonged in (b), consistent with a central lesion.

Clinical correlations

The central motor conduction time is prolonged in demyelinating disorders and the investigation is of particular value in patients suspected of having multiple sclerosis (Fig. 5(a) and Fig. 5(b)). However, slowing of central conduction is a non-specific abnormality that may also be seen in patients with other causes of CNS motor disorders. In amyotrophic lateral sclerosis for example, the central conduction time is often abnormal in an irregular pattern, though the prolongation is usually only slight. The MEP should therefore be supplemented with other evoked potentials (pattern-VEP and SSEP), MRI, and spinal fluid examinations.

In some patients with peripheral nerve disorders, in particular acute or chronic inflammatory demyelinating neuropathy, the MEP examination may show abnormalities indicating central as well as peripheral nervous system involvement. This may be an erroneous finding due to slowed conduction along spinal roots. Due to the stimulation of peripheral nerves distal to the intervertebral foramen, the conduction along the spinal roots is included in the central conduction time, and slowing at this segment may therefore erroneously be localized to the CNS.

Studies of the peripheral neuromuscular system

Indications

Electromyography (**EMG**) is used to establish whether weakness is due to a primary disease of muscle fibres (myopathy) or to a loss of a-motor fibres (neurogenic disorders). Nerve conduction studies are carried out to ascertain the loss of motor or sensory axons or the disturbed function of myelinated fibres. Both types of studies are usually needed for a differential diagnosis, and since the findings are rarely specific for particular disorders, the interpretation relies on inferences from several criteria of abnormality. The degree to which the findings should be supplemented and confirmed by light or electron microscopy of nerve- or muscle-biopsy specimens and other laboratory studies depends on the clinical setting. EMG and nerve conduction studies should be viewed as an extension of the clinical examination and form part of a neuromuscular consultation. EMG and nerve conduction studies assist in answering specific differential diagnostic questions relating to focal or generalized disorders of the peripheral neuromuscular system. Even though subclinical involvement has important implications in the diagnostic interpretation in several conditions (for example, neurogenic changes in non-weak muscles where amyotrophic lateral sclerosis is a possibility), the use of these studies to 'rule out' neuromuscular involvement in diffuse or focal pain problems should be discouraged.

Disturbances of neuromuscular transmission require specialized studies, including the recording of compound muscle responses evoked by repetitive stimulation of motor nerve fibres. In single-fibre EMG, the action potentials from individual muscle fibres are recorded during voluntary activity or during repetitive stimulation to measure the stability of neuromuscular transmission.

Electromyography

Method

EMG is carried out using needle electrodes. In routine studies, most laboratories either use concentric needle electrodes and a core recording lead with a surface area of 0.07 mm^2 referenced to the cannula, or insulated monopolar needles with a surface recording area of 0.17 mm^2 referenced to a surface electrode. The results obtained with these electrodes differ in regard to the amplitude of the motor-unit potential (**MUP**), whereas its duration only differs slightly. The baseline is somewhat more unstable when recorded with a monopolar electrode than a concentric needle. The signals should be recorded via a high-impedance amplifier with a frequency range of between 2 and 10 kHz.

Measurements

Recordings are carried out at rest, during weak voluntary activity, and during maximal voluntary activity (Table 1).

During rest, both the presence and type of spontaneous activity are characterized. During weak voluntary effort, individual MUPs are recorded without disturbance by other MUPs. The duration and amplitude of the MUP are also measured. The duration of the MUP reflects the activity of muscle fibres of the motor unit at a distance from the recording electrode. In contrast, the very few muscle fibres placed at the tip of the concentric electrode determines the amplitude of the MUP. The mean amplitudes and durations of at least 20 different MUPs recorded from 10 different sites at three different insertions are obtained. The shapes of individual MUPs are evaluated and designated as simple (less than five phases) or polyphasic (five or more phases). The findings are compared to age-matched control values from the investigated muscle, and the percentage deviation is calculated to ascertain whether the findings are normal or consistent with myopathy of chronic partial denervation (Table 1 and Table 2).

During a maximal voluntary contraction, all motor units in the muscle are activated. The MUPs from individual motor units cannot therefore be distinguished, but they form an interference pattern which, according to the degree of overlap, is measured semiquantitatively as a 'full', 'reduced', or 'discrete' recruitment pattern. Although a full recruitment pattern occurs in normal muscle, it requires the patient's full co-operation. It should therefore be noted whether the activity is recorded during maximal or submaximal effort. In addition to the degree of overlap, the envelope amplitude of the main activity is measured to distinguish myopathy and neurogenic involvement.

Clinical correlations

Recordings at rest

The sarcolemma of the denervated muscle fibre undergoes changes, including a gradual spread of acetylcholine receptors, and the resting membrane potential is reduced. Spontaneous discharges of denervated fibres occur in a cyclical pattern, and they appear as fibrillation potentials or positive sharp waves (Fig. 6). Fibrillation potentials have a triphasic shape and a duration of 5 ms at most, and the discharge pattern may be regular or irregular (Fig. 6(a) and Fig. 6(b)). Positive sharp waves are considered as arising from damage to the cell membrane and indicate a propagation block at the needle recording electrode (Fig. 6(b)). No particular pathological significance is assigned to whether the denervation activity consists of fibrillation potentials or positive sharp waves.

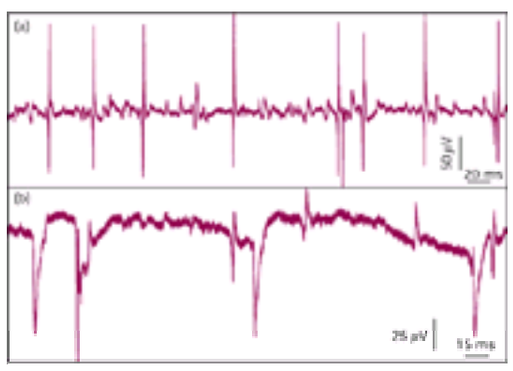

Fig. 6 Fibrillation potentials (a) and (b) and positive sharp waves (b) recorded from a muscle with profuse denervation activity. Fibrillation potentials arise from single muscle fibres and have a triphasic shape, duration of £ 5 ms, and variable amplitudes depending on the distance to the recording electrode. Positive sharp waves are thought to arise from damaged muscle fibres with conduction block at the recording electrode.

Continuous, non-propagated, miniature endplate potentials (**m.e.p.p.**)—and, in addition, irregular, spontaneous, endplate potentials (**e.p.p.**) with negative onset—are recorded from the resting normal muscle within the endplate region. Such single-fibre potentials cannot be distinguished from fibrillation potentials when recorded outside the endplate region. Therefore, spontaneous single fibre activity in normal muscle may be recorded at up to two out of ten recording sites (Table 1).

Denervation activity arises with a delay after the nerve lesion, and the lag is dependent on the length of the distal nerve stump: that is, it occurs within 5 to 10 days at very distal lesions. Similarly, after a nerve root lesion, denervation arises after few days in paraspinal muscles and after 2 to 3 weeks in distal extremity muscles. The presence of denervation activity indicates that the denervation is ongoing, but it may continue for years after occurrence of the lesion if reinnervation does not take place. However, denervation activity also occurs in muscular dystrophy, inflammatory myopathy (polymyositis, dermatomyositis, inclusion body myositis), and some metabolic myopathies (for example, acid maltase deficiency), whereas it is rare or absent in mitochondrial myopathy. In myopathy, denervation is due to segmental muscle-fibre necrosis, thus leaving a segment of the muscle fibre denervated. Denervation activity is a non-specific sign of neuromuscular disease (Table 1).

Whereas denervation activity arises from single muscle fibres, other types of spontaneous activity are due to discharges in groups of muscle fibres, possibly the whole motor unit. These include fasciculations (defined as irregular discharges with short or long intervals of less or more than 3 s (Fig. 7), myotonic discharges (defined as burst of activity with gradually waxing and waning frequencies, often elicited by percussion, needle movement or voluntary activity, and with a decreasing incidence after repeated contractions), complex repetitive discharges (defined as bursts of activity of variable duration with sudden occurrence and drop-out of discharge components that may arise from single fibres), and myokymia (defined as fasciculations, doublets, or triplets occurring with variable and sometimes high frequencies; neuromyotonia belongs in this category of activity).

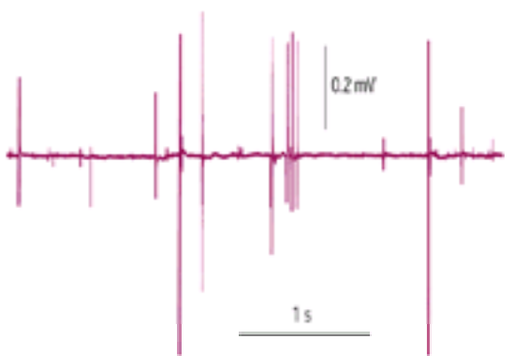

Fig. 7 Fasciculations recorded from the extensor digitorum communis muscle of a patient with multifocal motor neuropathy. The discharges arise from groups of muscle fibres or whole motor units and occur with irregular intervals.

Recording at weak effort

The smallest functional unit in the muscle is the motor unit, which differs quantitatively by several orders of magnitude in different muscles; in lower extremity muscles the motor units have 1000 to 2000 muscle fibres, whereas they have between 5 and 10 in extraocular muscles. The motor units also differ according to the biochemical characteristics of the muscle fibres in fast-contracting motor units (type II fibres) and in slowly contracting motor units (type I fibres). The diagnostic power of EMG mainly depends on the assessment of the structural changes of the motor units as evidenced by evaluation of the MUP. Reliance on the EMG has varied considerably over the years: quantitative measurements of MUPs may be used to gauge the overall size of the motor units and therefore as a diagnostic indicator of whether weakness is due to neurogenic abnormalities or to myopathy; as opposed to qualitative evaluation, which can only give an impression of the MUP changes. The use of quantitative measurements is now more widely used and accepted as the introduction of computerized measurement devices has enabled adequate numbers of MUPs to be sampled. This may increase the use of the MUP to differentiate between neurogenic and myogenic abnormalities.

MUP parameters measured include the duration, amplitude, and shape of the MUPs (Fig. 8). The motor units in myopathic muscle are reduced in size due to the functional loss or degeneration of individual muscle fibres, and this is reflected in reduced durations and amplitudes of the MUPs (Fig. 9 and Fig. 10). In contrast, the motor units in neurogenic lesions are enlarged due to collateral reinnervation of muscle fibres, and the mean duration of the MUPs is prolonged and the amplitude is increased (Fig. 9 and Fig. 10). These changes, which indicate the presence of chronic partial denervation, tend to be more pronounced in very chronic conditions. However, in motor neurone disease, MUPs may be gigantic even in muscles without clinical weakness. In specialized multielectrode studies, the motor-unit territory in myopathy is reduced, whereas the territory in neuropathy and amyotrophic lateral sclerosis is increased.

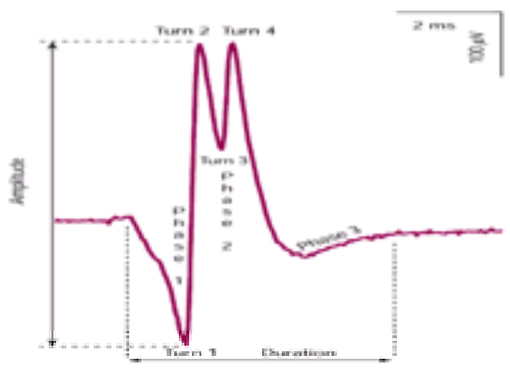

Fig. 8 Motor-unit potential (MUP) to illustrate measurements. The duration is measured from the first deflection from the baseline to the return to baseline. The amplitude (negative sign upwards) is measured peak-to-peak. The MUP has three phases and four turns (potential reversals of >100 µV). This potential is simple in shape (less than five phases). (From Simonetti *et al.*, with permission by Lippincott, Williams & Wilkins.)

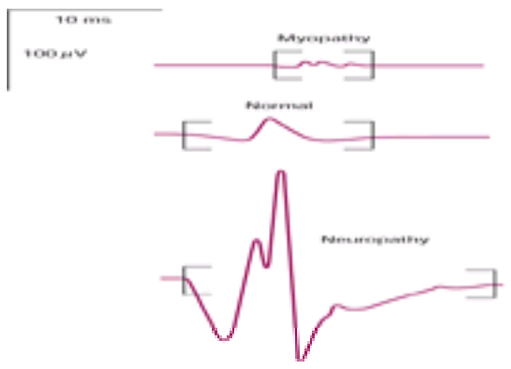

Fig. 9 Examples of MUPs from patient with myopathy (top), normal subject (middle), and neuropathy (bottom).

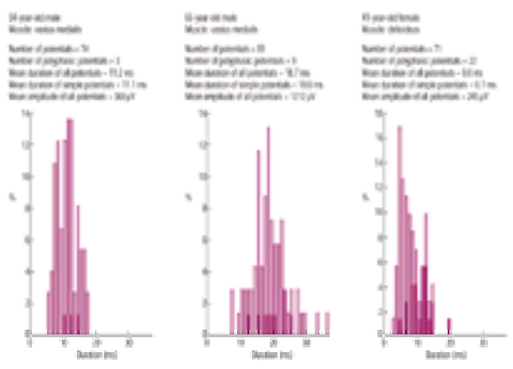

Fig. 10 Quantitative measurements of MUPs from a normal subject (left), a patient with neuropathy (middle), and one with myopathy (right). The total number of MUPs analysed, the number of polyphasic MUPs, the mean duration of all MUPs, the mean duration of simple MUPs, and the mean amplitude of all MUPs are indicated above the histograms. The histograms show the distribution of the durations of simple MUPs (open bars) and of polyphasic MUPs (filled bars). The amplitude and duration were markedly increased (duration, +51 per cent; amplitude, +427 per cent) in the patient with neuropathy. The duration was 29 per cent diminished and the amplitude was normal in the patient with myopathy. The incidence of polyphasic MUPs was slightly (13 per cent and 25 per cent) increased in the patients with neuropathy and myopathy, respectively. (Modified from Simonetti *et al.*, with permission by Lippincott, Williams & Wilkins.)

Whereas the changes in duration and amplitude are specific for either a myopathy or a neurogenic lesion, an increased incidence of polyphasic MUPs occurs both in myopathy and in neurogenic lesions, and is therefore a non-specific sign of neuromuscular involvement (Table 2). It is claimed that long-duration polyphasic MUPs are characteristic of neurogenic lesions, whereas short polyphasic MUPs are seen in myopathy. However, polyphasic MUPs in myopathy have two mechanisms: loss of muscle fibres in the motor unit, which results in short-duration MUPs; and degeneration of muscle fibres followed by regeneration and subsequent reinnervation, resulting in long-duration MUPs. With disease progression, muscle fibre regeneration cannot keep pace with degeneration, and long-duration MUPs become less frequent with advanced disease. The long-duration MUPs in myopathy may obscure the interpretation of the EMG, and it is therefore necessary to calculate the mean duration of simple MUPs (less than five phases) to avoid error.

The EMG examination should include a number of muscles according to the likely clinical diagnosis, since involvement of different muscles may vary in different disorders. In myopathy, proximal muscles in the upper and lower extremities should be examined. Some muscles show clear abnormalities characteristic of the disorder, whereas others show only non-specific changes (for example, fibrillation potentials and increased incidence of polyphasic potentials), and several criteria should therefore be collected. In neurogenic lesions on the other hand, distal muscles are most severely affected and may show abnormalities at an earlier stage than proximal muscles. In this connection it should be considered that some distal muscles may be affected due to focal non-related causes. The extensor digitorum brevis muscle, for example, should be avoided in elderly people due to frequent neurogenic changes caused by compression of the deep peroneal nerve by footwear. In patients suspected of having amyotrophic lateral sclerosis, both clinically weak and non-affected muscles should be studied; as a rule, both show signs of chronic partial denervation often with such pronounced changes that this supports the diagnosis (Table 2). Since amyotrophic lateral sclerosis often has a focal distribution at presentation, it is important to exclude spinal root compression and peripheral nerve lesions as causes of the neurogenic involvement; therefore, it is customary to study several muscles in different extremities to ensure that any changes are widespread.

Maximal voluntary contraction

The number of motor units is normal in myopathy, and therefore the loss of muscle fibres is associated with a full recruitment pattern with reduced amplitude (Table 2). In severely weak myopathic muscle, the loss of muscle fibres may eventually result in a reduced recruitment pattern (see above). Where there is neurogenic involvement, the loss of motor units results in a reduced recruitment pattern, often with increased amplitude due to collateral reinnervation. In advanced denervation, the number of motor units is so depleted that the recruitment pattern becomes discrete (Fig. 11(c)). This is considered a specific sign of neurogenic involvement, whereas a reduced pattern may occur in either myopathy or neurogenic disease. With motor neurone involvement, the reduced or discrete pattern has a markedly

increased amplitude, often of more than 8 mV, which is considered typical of motor neurone disease (Table 2; Fig. 11(c)).

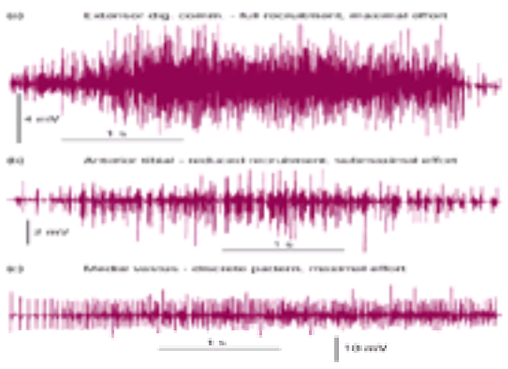

Fig. 11 Electrical activity recorded during voluntary effort. The recruitment pattern in (a) was full and had a normal amplitude of 3.5 to 4 mV, whereas it was reduced in (b) due to incomplete co-operation by the patient (submaximal effort) as shown by the uneven discharges that occurred in bursts. In (c) the recruitment pattern was discrete due to the loss of motor units, and the amplitude was markedly increased to between 8 and 10 mV due to collateral reinnervation (46-year-old man with motor neurone disease).

Specialized recordings

As indicated in Method section, the amplitudes of the MUPs, in particular when recorded with a concentric needle, are markedly variable and dependent on the distance between the recording area and the closest two or three muscle fibres of the motor unit. An increased or decreased amplitude is therefore a relatively insensitive indicator of motor-unit abnormalities. To record more evenly from the whole motor unit and hence obtain a more reliable measure of the MUP amplitude, the macroelectrode (which consists of 15 mm of the non-insulated cannula of the needle electrode to increase the recording surface area), has been introduced. This has been useful in serial studies designed to follow the disintegration of the motor unit in patients with postpolio syndrome.

In contrast to the macroelectrode, the single-fibre electrode has a small recording area of 25 µm, which allows recording of the action potential from individual muscle fibres in the motor unit. The main use of the single-fibre electrode has been in the recognition of disorders of neuromuscular transmission. The timing of the discharges of two (or more) muscle fibres in the motor unit is followed during repetitive activity. Whereas the discharges are quite stable in normal muscle, they become unstable in myasthenia gravis or the Lambert–Eaton syndrome. The larger variance of the discharges is termed 'increased jitter'; in more severely affected neuromuscular transmission disturbances, some of the discharges may fall out altogether, so-called 'blocking'. Increased jitter is also encountered in myopathy. In neurogenic lesions, where the activity of several muscle fibres may be recorded simultaneously, groups of discharges may become unstable, indicating that conduction along immature terminal sprouts may have a diminished safety factor.

Quantitation of the MUP relies on the ability to distinguish the individual MUP from the activity in other motor units. This may introduce a bias regarding the type of motor unit that a diagnosis is based on. Thus small, fatigue-resistant motor units are recruited during weak effort, whereas large motor units are activated at higher levels of activity. Methods have therefore been developed that quantitate the electrical activity during higher levels of activity. These methods have been used to investigate patients with myopathy and neurogenic involvement, and have been found to supplement the findings obtained using quantitative evaluation of MUPs.

Nerve conduction studies

Motor and sensory nerve conduction studies of peripheral nerves are performed by recording the propagated responses evoked by supramaximal electrical nerve stimulation. The responses reflect the summation of action potentials from individual sensory nerve fibres (the compound sensory action potential, **CSAP**) or motor units (the compound motor action potential, **CMAP**), and their amplitudes represent a semiquantitative measure of the number of activated myelinated fibres. The CSAP amplitude is an expression of activity in large fibres with diameters greater than 7 µm, whereas small fibres in the nerve contribute only slightly to the response. The CMAP amplitude is a reflection of the number of a-motor neurones. A reduction of the CSAP or the CMAP amplitudes is an indicator of fibre loss. However, the CMAP amplitude is also influenced by the size of the motor-unit response. During chronic axonal loss, collateral spouting causes an increase of the MUP, which partially or completely may compensate for the fibre loss. Since reinnervation does not have the same effect on sensory nerves, the CMAP is a less sensitive measure for determining the degree of fibre loss in chronic axonal lesions than the amplitude of the CSAP. Conduction studies are supplemented with EMG to ascertain whether motor fibres are affected by the pathological process.

Method

Investigations of motor and sensory fibres are carried out separately.

Motor conduction studies

The nerve is stimulated by an electrical pulse of 0.1- to 0.2-ms duration applied to the nerve at well-defined sites, with the cathode (depolarizing electrode) being placed over the nerve distal to the anode if a longitudinal electrode placement is used. It is essential that the stimulation pulse is supramaximal, defined as being 10 to 20 per cent above the stimulus strength that elicits a maximal CMAP. This stimulation strength requires that the output of the stimulator provides a stimulus at least four times higher than threshold when surface electrodes are used, and ensures that all fibres in the nerve remain activated even though the stimulation electrode may move slightly. The CMAP is recorded from the muscle, preferably using surface electrodes or a subcutaneous needle electrode to ensure that activity from all the muscle fibres in the muscle can be 'seen' by the recording electrode. Usually a belly-tendon montage is used, with the reference electrode being placed over a remote site. Recording of the CMAP with a concentric electrode is usually discouraged since the amplitude is highly dependent on how close the electrode is in relation to the active muscle fibres. However, a concentric needle electrode is useful in very atrophic muscles, as it allows a sharp deflection from the baseline to be measured. The response is amplified at a frequency band between 10 (or 20) Hz and 10 kHz.

Measurements

The parameters include latencies, conduction velocities, amplitudes, areas, durations, and shapes of the CMAPs (Fig. 12). The latency is measured to the first deflection from the baseline, which is negative when the CMAP is recorded from the endplate zone. Motor nerve conduction velocities (**MNCV**) between two sites of stimulation are calculated by dividing the distance between stimulation sites by the difference in latencies (Fig. 12).

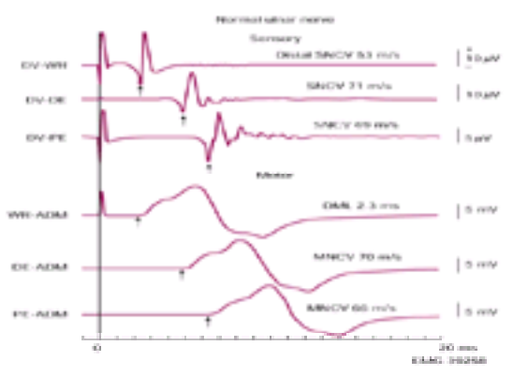

Fig. 12 Motor and sensory conduction studies of a normal ulnar nerve from a patient with diffuse complaints from the arm. Above: compound sensory action potentials, evoked by electrical stimulation at digit V (DV), were recorded via needle electrodes at the wrist (WR), below the elbow (DE), and above the elbow (PE).

The latencies were measured to the first positive peak (arrows) and the sensory nerve conduction velocities (SNCV), indicated above the traces, were calculated as indicated in the Method section. Below: compound muscle action potentials, evoked by stimulation at the wrist (WR), below the elbow (DE), and above the elbow (PE) were recorded via a surface electrode over the abductor digiti minimi muscle (ADM). The latencies were measured to the first deflection from the baseline (arrows). DML (distal motor latency) and MNCV (motor nerve conduction velocities), indicated above the traces were obtained as described in the Method section.

Distance between stimulation sites (mm)Difference between distal and proximal latencies (ms) = MNCV (m/s)

An MNCV is not calculated for the most distal site of stimulation since the latency includes conduction along terminal motor-axon branches and the neuromuscular transmission delay. In this instance, the delay is designated the distal motor latency (**DML**) (Fig. 12).

The CMAP amplitude (in mV) is measured either at the negative phase or peak-to-peak. The negative phase is usually preferred since it is less subject to influence by the positioning of the reference electrode. The area and duration of the negative phase are useful in evaluating temporal dispersion or conduction block.

Sensory conduction studies

In contrast to the CMAP, the compound sensory action potential (in μV) is recorded directly from active nerve fibres, and it is therefore only about 1/500th to 1/1000th of the CMAP amplitude. The response may be recorded through surface electrodes placed on the skin above the nerve or through needle electrodes placed close to the nerve. Surface electrodes are easy to apply but have a lower sensitivity than needle electrodes. In some normal, elderly people a CSAP from the lower limbs may, therefore, not be recorded through surface electrodes that have a resolution of about 1 μV, whereas near-nerve needle electrodes allow the recording of responses with an amplitude as low as 0.1 μV. In addition, the use of needle electrodes allow simultaneous recording from several sites along the nerve, which are usually not possible using surface electrodes (Fig. 12).

Sensory responses may be recorded antidromically (proximal stimulation, distal recording) or orthodromically (distal stimulation, proximal recording). The recording and reference electrodes may be placed longitudinally in relation to the nerve (bipolar recording), or the reference electrode may be placed transversely at a remote site (unipolar recording). The recording arrangements have advantages and disadvantages: the main advantage of bipolar recording being a smaller stimulus artefact; and the main disadvantage that the potential recorded by the reference electrode influences the CSAP shape and amplitude. Due to the low amplitude of the CSAP, electronic averaging is usually required to obtain a clear response that is suitably free of noise.

Measurements

The parameters include latencies, conduction velocities, amplitudes, and shapes of the CSAPs. The latency is measured to the first positive phase of the CSAP. This indicates conduction along the largest fibres in the nerve. Since the conduction path between the sites of stimulation and recording does not include synaptic transmission, a distal sensory nerve conduction velocity (SNCV) may be calculated (Fig. 12):

Distance between stimulation and recording sites (mm)Latency (ms) = Distal SNCV (m/s)

When the orthodromically conducted CSAPs are recorded at several sites along the nerve, the SNCV is calculated using a similar procedure as that used to calculate the MNCV:

Distance between proximal and distal recording sites (mm)Differernce between proximal and distal latencies (ms) = SNCV (m/s)

In some laboratories, the latency of the CSAP is measured to the first negative phase. This is to be discouraged, since this part of the response is a summation of both large and small fibres. A change in summation due to temporal dispersion therefore precludes measurements to the same group of fibres.

The amplitude of the CSAP (in μV) is measured peak-to-peak. The shape is usually bi- or triphasic. In both axonal and demyelinating neuropathies the shape may become dispersed, and when recorded with a needle electrode, the shape may become polyphasic. Due to temporal dispersion, the conduction distance has a marked influence on the shape of the CSAP. At long conduction distances a polyphasic shape may be normal.

Clinical correlations

The motor and sensory nerve conduction velocities (MNCV, SNCV) are measures of conduction of the largest motor and sensory fibres in the nerve. These limitations should be considered when the results of the studies are interpreted as, for example, a small-fibre neuropathy may escape detection. Similarly, if just a single large motor fibre is preserved, the motor conduction velocity may remain normal, indicating that the conduction velocity cannot be used in isolation to establish the presence of a neuropathy.

Focal nerve lesions

The number of patients referred to the clinical neurophysiology laboratory with possible focal nerve lesions due to compression or entrapment at root level, or along the course of the nerve, far outweighs that for other neuromuscular disorders. In entrapment and focal compression neuropathy, the main pathological abnormalities comprise demyelination at the site of the lesion, which is complicated by axonal loss in advanced lesions. Accordingly, the electrophysiological findings display a disproportionate slowing of conduction at the site of the compression, and also, in some cases, a loss of fibres as demonstrated by reduced amplitudes of the CMAP and the CSAP. This is illustrated in Fig. 13 from a patient with ulnar nerve entrapment at the elbow; the MNCV and SNCV across the elbow were markedly reduced compared to the conduction velocities distal to the elbow, consistent with a focal demyelinating lesion. The CSAP amplitudes were, in addition, markedly diminished, thus indicating axonal loss (Fig. 13(b)), and the slight reduction in SNCV distal to the elbow is commensurate with a loss of large fast-conducting fibres. The apparent sparing of the CMAP amplitude is due to collateral reinnervation masking the motor fibre loss revealed by EMG examination (see above). The slight MNCV reduction distal to the elbow is due to the fibre loss (Fig. 13(a)). Similar methods are used to study focal nerve lesions in patients with compression of the peroneal nerve at the fibular head and in patients with Saturday night palsy (compression of the radial nerve in the spiral groove).

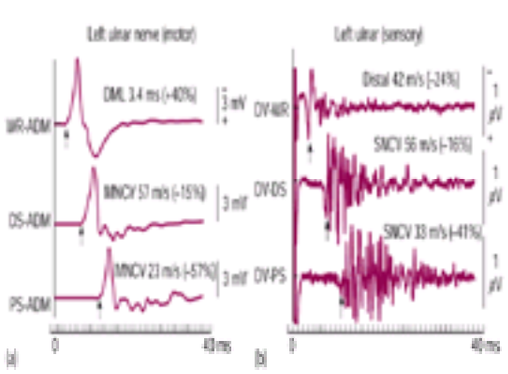

Fig. 13 Conduction studies in a patient with clinical signs of an ulnar nerve lesion. Both (a) motor and (b) sensory conduction showed a marked reduction of conduction velocities across the elbow (57 per cent and 41 per cent reduced between distal to the ulnar sulcus (DS) and proximal to the sulcus (PS), respectively). In addition, there was motor and sensory axonal loss, as indicated by the reduction in amplitudes of the CMAP and the CSAPs. The mild–moderate slowing of conduction distal to the elbow was probably due to a loss of large fibres.

In patients with carpal tunnel syndrome, the distal motor latency of the CMAP to the abductor pollicis brevis muscle, evoked by stimulation at the wrist, is prolonged, indicating a slowing of conduction beneath the flexor retinaculum. However, because it is difficult to stimulate motor fibres distal to the entrapment, differential attenuation of the MNCV is therefore not assessed. To ensure that the median nerve is selectively affected, the latency to the non-affected ulnar nerve should be normal. The SNCV, by contrast, may be tested both distal to and across the carpal tunnel, and is disproportionately reduced along this segment of the nerve compared with that distal to it. Variations on the study paradigm have been devised to increase the sensitivity of the electrophysiological studies.

On the other hand, it is usually not possible to directly study conduction across the compressed nerve segment in patients with very proximal lesions located at spinal roots or the brachial plexus across a cervical rib or band. In these situations, the anatomical distribution of EMG signs of chronic partial denervation is important in the differential diagnosis. For example, EMG findings in patients with apparent involvement of the ulnar nerve due to a C8 lesion or a thoracic outlet syndrome include abnormalities in non-ulnar nerve innervated muscles (for instance, the abductor pollicis brevis and the extensor digitorum communis muscles). This distribution of motor axon loss, and the absence of focal MNCV changes at the elbow (Fig. 14(a)), show that a single nerve lesion is not the cause of the clinical deficit, but it does not distinguish between a radicular and a brachial plexus lesion. By contrast, sensory conduction studies in root lesions usually remain normal, provided that the lesion is located proximal to the dorsal root ganglion, whereas the CSAP from digit V in the thoracic outlet syndrome is diminished due to sensory fibre loss at the level of the brachial plexus (Fig. 14(b)).

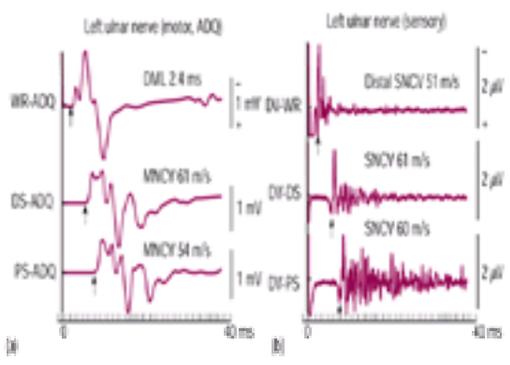

Fig. 14 Conduction studies in a patient with clinical signs of an ulnar nerve lesion that was clinically localized at the brachial plexus or the spinal roots, as indicated by the extent of sensory complaints along the medial forearm and upper arm. Both (a) motor and (b) sensory conduction showed normal conduction velocities distal to and across the elbow. However, the amplitudes of the CMAPs and the CSAPs were markedly decreased, indicating diffuse axonal loss. The loss of sensory fibres is inconsistent with a root lesion and indicates entrapment at the brachial plexus.

Generalized nerve lesions (peripheral neuropathy)

The electrophysiological study in patients with suspected polyneuropathy should document that pathological abnormalities are widely distributed. It is therefore a prerequisite that several motor and sensory nerves in the upper and lower limbs are investigated. However, the study should be individually tailored to delineate the distribution of changes, and the strategy should reflect the symptoms and clinical findings and hence address the question of the differential diagnosis. For example, it may be necessary to investigate certain nerves bilaterally if the symptoms are asymmetrical. This would allow the investigation to show whether the patient may have a mononeuropathy, a multiple mononeuropathy, or a polyneuropathy with asymmetrical features.

Axonal polyneuropathy

The underlying pathology in most patients with peripheral neuropathy is axonal loss in a symmetrical distribution, primarily located at distal nerve segments and more pronounced in the legs than in the arms. The electrophysiological characteristics in these patients are due to a loss of nerve fibres, that is associated with EMG signs of chronic partial denervation and conduction studies that show diminished amplitudes of evoked CMAPs and CSAPs. The remaining fibres in the nerve may conduct normally, and the MNCV and the SNCV in these patients may therefore be normal or show a reduction consistent with a large-fibre loss (Fig. 15). In some patients, the pathological changes may primarily be present at the distal nerve segments, thus studies of these segments will also be required (Fig. 15).

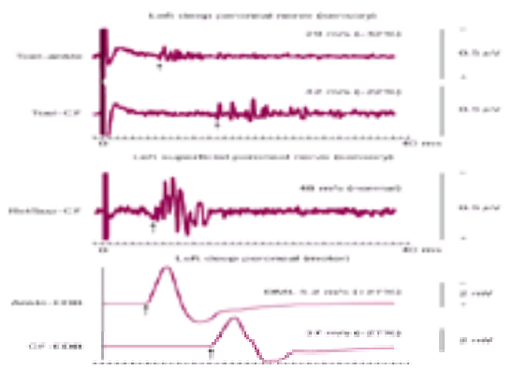

Fig. 15 Motor and sensory conduction studies of the peroneal nerve in a patient with diabetic neuropathy. The findings in this patient were consistent with axonal loss, primarily present in the very distal segment of the nerve. Top panel: orthodromic compound sensory action potentials (CSAPs), evoked at the deep peroneal nerve in the first dorsal interstice (ToeI), were recorded at the ankle (ankle) and the fibular head (CF). The amplitudes were more than 95 per cent diminished. The SNCVs were moderately diminished due to large-fibre loss. Middle panel: CSAP of the superficial peroneal nerve evoked at the superior retinaculum (RetSup) at the ankle and recorded at the CF. The amplitude was slightly diminished, and the SNCV was normal. Lower panel: compound muscle action potentials (CMAPs) of the deep peroneal nerve, evoked at the ankle and the fibular head, were recorded at the extensor digitorum brevis muscle (EDB). The distal motor latency (DML) was prolonged and the MNCV was reduced due to fibre loss.

Only motor or sensory fibres are involved in the rarer types of neuropathy, and demonstration of such a distribution may have important implications in the differential diagnosis. A sensory neuronopathy in a 47-year-old woman with profound sensory ataxia is illustrated in Fig. 16. The motor nerve conduction studies and the EMG were normal, but the sensory conduction studies were profoundly abnormal, and the sural nerve biopsy showed a 95 per cent loss of myelinated fibres (Fig. 16(a) and Fig. 16(b)). The SNCV was at the lower normal range, consistent with the slightly diminished diameter of the largest remaining fibres being around 10 µm (Fig. 16(b) and Fig. 16(c)). These findings were consistent with an autoimmune sensory neuronopathy.

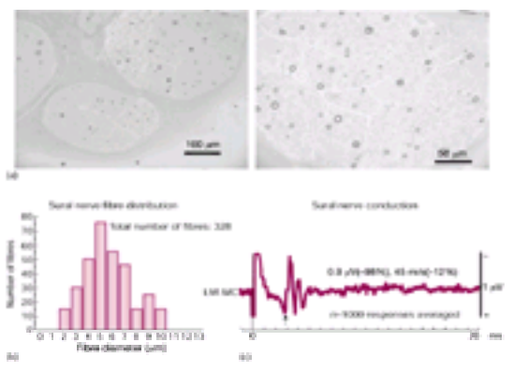

Fig. 16 Morphometric and electrophysiological studies of the sural nerve in a 47-year-old woman with sensory neuronopathy. (a) Transverse semithin sections of the sural nerve showed a generalized severe fibre loss without evidence of degeneration or regeneration. (b) A total of 328 myelinated fibres were counted in the whole nerve. The fibre-diameter distribution showed a loss of both small and large fibres with a maximal diameter of 10 to 11 µm. (c) Conduction studies in the sural nerve showed a pronounced reduction of the CSAP amplitude and a dispersed shape. The SNCV was at the lower normal limit, consistent with the diameters of the largest fibres at morphometry. (Histological studies by courtesy of H. Schmalbruch, University of Copenhagen.)

Demyelinating neuropathy

Primary demyelination usually has a hereditary, inflammatory, or autoimmune basis. Although demyelinating neuropathy is rare compared to axonal neuropathy, it has become increasingly important to be able to diagnose acquired demyelinating neuropathy with a high degree of certainty since acute or chronic inflammatory demyelinating neuropathy may respond to immunomodulatory therapy. The primary electrophysiological sign of demyelination is a markedly reduced conduction velocity that is beyond the diminution caused by large-fibre loss. In hereditary motor and sensory neuropathy, type I and III, the MNCV and the SNCV are markedly diminished to less than 50 per cent of the lower limit of normal throughout the nerves, consistent with primary demyelination. In these conditions the amplitudes of the CMAPs and the CSAPs are, however, also markedly reduced and nerve biopsy confirms a marked loss of myelinated nerve fibres. Pure demyelination without axonal loss does not occur in these hereditary conditions and is extremely rare in acquired demyelinating neuropathy. The distinguishing features in acquired demyelinating neuropathy include widespread demyelination, often in a multifocal pattern, as indicated by focal temporal dispersion or conduction block or both of CMAPs and CSAPs. Criteria have been established to assist in the diagnosis of these demyelinating neuropathies (Table 3).

A weakness or sensory loss does not result solely from a diminished conduction velocity, but also as a consequence of nerve fibre loss or a block of conduction between the CNS and the target muscle. Conduction block is a partial or complete inability of fibres to propagate action potentials along a segment of the nerve, and is demonstrated by recording a larger motor or sensory response more distal to than proximal to the site of the block (Fig. 17(a)). This reduction in amplitude should be greater than that associated with a temporal dispersion of the conducting fibres. Therefore to demonstrate a block of motor fibres, the CMAP should show a reduction of at least 50 per cent (Fig. 17(a)). Conduction block may occur in acquired acute or chronic inflammatory demyelinating neuropathy and in some cases of monoclonal gammopathy. It does not occur in hereditary demyelinating neuropathy and probably not in gammopathy with IgM anti-MAG antibodies. However, demyelination due to compression may also cause a conduction block, and in demyelinating neuropathy a conduction block must therefore be demonstrated outside the usual sites of entrapment or compression. The pathophysiological changes in inflammatory demyelinating diseases usually show a multifocal pattern, with some nerves showing pronounced changes while other nerves or nerve segments have normal conduction.

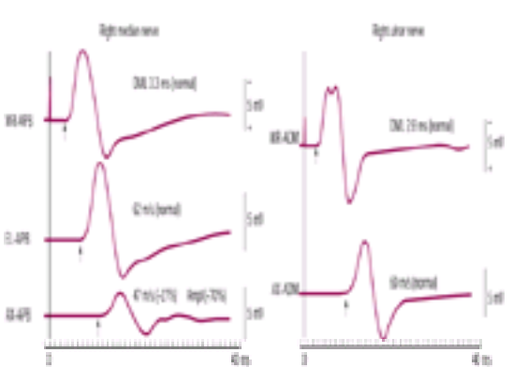

Fig. 17 Motor conduction studies of the median and ulnar nerves of a 29-year-old woman with electrophysiological signs of demyelinating neuropathy and clinical signs of relapsing chronic inflammatory demyelinating neuropathy. Clinical examination showed marked weakness of the thenar and forearm median innervated muscles. The force in ulnar innervated muscles was normal. Left panel: CMAP recorded from the abductor pollicis brevis muscle (APB). Stimulation at the wrist (WR) and elbow (EL) elicited normal CMAPs and the DML and MNCV along the forearm were normal. Stimulation at the axilla (AX) resulted in a markedly reduced CMAP amplitude and a reduced MNCV between the elbow and axilla, consistent with a partial conduction block. Right panel: stimulation of the ulnar nerve at the wrist and axilla evoked CMAPs at the abductor digiti minimi (ADM) of normal amplitudes, DML and MNCV, indicating that motor fibres were not affected.

Apparent conduction block may be found in acute neuropathies due to vasculitis. However, conduction along the nerve segment distal to a focal lesion may continue for several days before Wallerian degeneration takes place, and repeated studies should therefore be carried out to exclude this possibility.

Motor disorders (motor neurone disease and motor neuropathy)

Conduction studies in motor neurone disease are normal at early stages of the disease, but at late stages the CMAP amplitudes are reduced. The distal motor latencies of weak and wasted muscles are often prolonged, and the MNCV slightly to moderately reduced due to a loss of large a-motor axons. In ALS, sensory conduction studies are usually normal, though the CSAP amplitudes may be slightly diminished. Therefore, conduction studies are mainly of use in the diagnosis of ALS in that they are normal in the face of widespread neurogenic changes at EMG. However, patients with X-linked bulbospinal muscular atrophy (Kennedy's syndrome) are characterized by a marked reduction of CSAP amplitudes, while the EMG examination shows abnormalities characteristic of motor neurone disease (fasciculations, widespread denervation, markedly enlarged and prolonged MUPs, and discrete high-amplitude recruitment at maximal effort).

Asymmetrical or focal weakness, atrophy, fasciculations (see Fig. 7), without or with only slight sensory symptoms, due to multifocal motor neuropathy may be mistaken for the early stages of spinal forms of motor neurone disease. The course is, however, prolonged over several years, and there is no involvement of bulbar muscles and no signs of corticospinal involvement. The electrophysiological features in these patients are distinct, with conduction block of motor fibres indicating focal demyelination. Usually there is also EMG evidence of chronic partial denervation, indicating fibre loss. The sensory conduction studies through affected nerve segments are normal. These patients frequently have high titres of anti-GM1 antibodies, and often respond clinically and electrophysiologically to intravenous infusions of immunoglobulin. Patients with chronic inflammatory demyelinating neuropathy may have mainly motor symptoms, with only minimal sensory deficits. Motor conduction studies in chronic inflammatory demyelinating neuropathy show signs of mixed demyelination, including conduction block, and fibre loss; sensory conduction studies are abnormal, with small CSAP amplitudes and often reduced SNCV. This is an important distinction, since patients with chronic inflammatory demyelinating neuropathy may respond to treatment with corticosteroids and other immune-modulating strategies that have no effect on multifocal motor neuropathy.

*I am indebted to Dr H. HØgenhaven MD for comments on the manuscript.

Further reading

Albers JW, Kelly JJ (1989). Acquired inflammatory demyelinating polyneuropathies: clinical and electrodiagnostic features. *Muscle and Nerve* **12**, 435–51.

Binnie CD, *et al*. (1995). EMG, nerve conduction and evoked potentials. In: Osselton JW, ed. *Clinical neurophysiology*, pp. 43–321. Butterworth–Heinemann, Oxford.

Bouche P, *et al*. (1999). Electrophysiological diagnosis of motor neuron disease and pure motor neuropathy. *Journal of Neurology* **246**, 520–5.

Brown WF, Bolton CF, eds. (1993). *Clinical electromyography*, 2nd edn. Butterworth–Heinemann, Boston.

Buchthal F (1957). *An introduction to electromyography*. Scandinavian niversity Books, KØbenhavn, Stockholm.

Buchthal F (1985). Electromyography in the evaluation of muscle disease. Symposium in Electrodiagnosis. *Neurologic Clinics* **3**, 573–98.

Buchthal F, Kamieniecka Z (1982). The diagnostic yield of quantified electromyography and quantified muscle biopsy in neuromuscular disorders. *Muscle and Nerve* **5**, 265–80.

Chiappa KH, ed. (1997). *Evoked potentials in clinical medicine*, 2nd edn. Lippincott–Raven, Philadelphia.

Fuglsang-Frederiksen A (1981). *Electrical activity and force during voluntary contraction of normal and diseased muscle*. Munksgaard.

Fuglsang-Frederiksen A (2000). The utility of interference pattern analysis. *Muscle and Nerve* **23**, 18–36.

Ho TW, *et al*. (1997). Patterns of recovery in the Guillain-Barre syndromes. *Neurology* **48**, 695–700.

Kimura J (1989). *Electrodiagnosis in diseases of nerve and muscle. Principles and practice*, 2nd edn. FA Davis, Philadelphia.

Krarup C (1999). Pitfalls in electrodiagnosis. *Journal of Neurology* **246**, 1115–26.

Mauguière F (1995). Evoked potentials. In: Osselton JW, ed. *Clinical neurophysiology*, pp 325–572. Butterworth–Heinemann, Oxford.

Niedermeyer E, Lopes da Silva F, eds (1993). *Electroencephalography. Basic principles, clinical applications, and related fields*, 3rd edn. Williams & Wilkins, Baltimore.

Nuwer MR (1999). Spinal cord monitoring. *Muscle and Nerve* **22**, 1620–30.

Sandberg A, Hansson B, Stålberg E (1999). Comparison between concentric needle EMG and macro EMG in patients with a history of polio. *Clinical Neurophysiology* **110**, 1900–8.

Simonetti S, Nikolic M, Krarup C (1999). Electrophysiology of the motor unit. In: Younger DS, ed. *Textbook of motor disorders*, pp 45–60. Lippincott Williams & Wilkins, Philadelphia.

Stålberg E, Trontelj JV (1979). *Single fibre electromyography*. Mirvalle Press, Old Woking, Surrey.

24.3 Brain and mind: functional neuroimaging

Richard Frackowiak

Introduction

The sensory input to the brain is generally organized into sets of separate, functionally distinct, maps in the cerebral cortex. That much has been known for many years from neurology and animal studies. Can deductions based on non-human primate experiments be assumed to hold true for the brains of humans? For one thing there are obvious differences in the size of the brain between species, and certain areas such as the frontal lobes seem greatly developed in humans. Certain functions, for example spoken language and silent speech, are apparently unique attributes of the human brain. There may be a species-specific organization of specialized cortical areas to account for such human cognitive attributes.

The aim of functional neuroimaging is to describe the activity of neuronal populations and how they are organized into brain networks and systems. This systems-level approach seeks a biological understanding of how integrated brain functions are embodied in the physical structure of the brain. Important areas of enquiry include an understanding of how sensory inputs map onto the brain, and where subsequent signal processing occurs when complex percepts are experienced. Other examples include how sensory and motor systems interact during sense-guided movement, such as reaching under visual control, or how cognitive functions, such as memory and language, are organized and what is the basis for their modification by emotional state.

Functional neuroimaging methods fall broadly into two classes—those that provide information about synaptic activity and those that provide information about neurochemistry or neurotransmission. The former methods usually depend on measurements of the distribution of local cerebral blood flow (often known as perfusion maps) or, when comparisons between successively recorded distributions in different brain states are made, as activation studies. Local perfusion is a surrogate marker for local synaptic activity because of a tight coupling between it and local cerebral glucose metabolism both at rest and with activation. Such measurements can be accomplished with positron emission tomography and with functional magnetic resonance imaging, although in this method the images depend less clearly on perfusion alone as the signal is dependent on blood oxygen level. Radioactivity is not used in functional magnetic resonance imaging, which can therefore be safely repeated an unlimited number of times in any individual. Both positron emission tomography and functional magnetic resonance imaging localize changes in local brain activity to a millimetre or so. The second class of functional neuroimaging methods relies on mapping the distribution of chemical species of interest with positron emission tomography after injection of appropriately specific radiotracers, or by identification of unique magnetic signatures of compounds of interest with magnetic resonance spectroscopy. It is also possible to map non-invasively the distribution of electrical or magnetic signals coming from the brain. The anatomical precision with which this can be done is markedly worse than with positron emission tomography and functional magnetic resonance imaging, but brain activity can be followed millisecond by millisecond, a temporal resolution that is not possible with the other two methods.

There are clinical uses for functional neuroimaging, such as preoperative assessment of the functional integrity of peritumoural tissue and non-invasive assessment of language and memory dominance in patients undergoing temporal lobe surgery for epilepsy. Positron emission tomography and functional magnetic resonance imaging have had an apparently less spectacular impact on clinical practice than have anatomical imaging methods such as computed tomography and conventional magnetic resonance imaging. This is because functional neuroimaging is used to discover disease mechanisms of general significance, relevant to groups of patients, while computed tomography and magnetic resonance imaging provide information of direct relevance to the diagnosis, prognosis, and management of individual patients. Nevertheless, there are some striking examples of the impact of functional neuroimaging on clinical practice. For example, the definition of the haemodynamic and energetic consequences of preclinical and clinical carotid artery disease and the description of the time course of the natural transition from ischaemia to infarction have modified views about surgical treatment and about pharmaceutical approaches to stroke therapy, and have also affected the design of clinical trial protocols. Functional neuroimaging is able to detect preclinical degenerative brain disease and to distinguish between different clinical disorders. Positron emission tomography has also been invaluable in assessing the therapy of Parkinson's disease by monitoring fetal mesencephalic graft survival and the effects of subthalamic and pallidal lesions and chronic subthalamic nucleus stimulation.

From sensation to cognition

The visual system has been relatively well studied by some of the modern neuroimaging methods described above and will be used to illustrate some general principles. A simple imaging experiment to imagine is measurement of the distribution of brain activity during an eyes-open and an eyes-closed state. A comparison of brain activities recorded in each of the two states identifies areas of the brain in which activity is specifically associated with vision. When early sensory processing is the object of study this comparative approach is relatively free of assumptions. When more complex cognitive functions are studied more sophisticated experiments and analyses must be used.

The visual world is mapped from retina to cortex

The visual world depends on patterns of light hitting the retina. The evoked retinal signals are transmitted to the visual cortex in a point to point manner. Signals coming from adjacent patches of retina and hence adjacent parts of the visual field are mapped onto adjacent patches of cortex, a fact that is also deduced from studies of patients with focal lesions of the visual cortex. This retinotopic organization of primary visual cortex (V1) has been clearly confirmed with functional scanning in normal humans. Activity recorded in the brain with peripheral visual targets can be compared with that recorded with central presentation. Each quadrant of the visual field is located in the opposite cerebral hemisphere and quadrant of the visual cortex. The fovea is represented at the pole of the occipital cortex and peripheral vision activates more anterior parts of the calcarine cortex. One can calculate from such data the magnification factor, i.e. the length of cortex that maps a given 'length' of the visual field, and also the borders between specialized extrastriate visual regions.

The extrastriate occipital cortex receives multiple parallel outputs from area V1. This cortex is functionally heterogeneous, different parts activate in association with different aspects of visual perception (for example form, colour, movement, and face recognition). Area V5 is an extrastriate area in which activity is associated with visually perceived motion. The brains of individuals vary one from another quite markedly, not only in shape but also in the disposition of the gyri and sulci of parts of the cortex. However, accurate alignment of functional and anatomical images is a trivial issue with the use of modern computers. It is possible to show precise relationships between structure and function despite considerable individual variability of the normal anatomy of the occipital cortex. V5 is always found in a circumscribed part of the occipital lobe at the junction with the temporal lobe in the angle formed by two occipital gyri—the inferior occipital and the ascending limb of the inferior temporal. This anatomical site is relatively developed in infants with heavy myelination of the associated white matter fibres. In summary, there is a remarkable correlation between developmental factors, functional specialization, and anatomical location, despite considerable variability in the absolute spatial location of the anatomical structure between different individuals. The heavy myelination that is characteristic of area V5 can be demonstrated by high-resolution anatomical magnetic resonance imaging, as can the stria of Gennari, a unique structural characteristic of primary visual cortex (V1).

Activity in V5 occurs with perceived visual motion, whether that motion is real or illusory. One area of extrastriate cortex (V3) shows equivalent activity when real or illusory objects are perceived. Schizophrenic patients with visual delusions show activation of extrastriate areas that determine the content of their delusions. This evidence suggests that the visual brain is constructing an intrerpretation of the visual scene from the information provided. Sometimes that information is appropriate to reality, at other times it has a configuration that, for unknown reasons, elicits activity in an unexpected specialized visual area. The consequent illusory perception is congruent with that experienced when the area is activated by real stimuli. The Kanizsa triangle is an example of a normal visual illusion of this type for which a partial explanation can be proposed. The perception of an apparent triangle in front of a solid body can be ascribed to the fact that in normal visual life we expect to see objects hidden by others that lie in front of them. The triangle illusion is a reflection of this phenomenon. Similarly, activation of the visual motion area (V5) is seen during the waterfall illusion in which looking at a waterfall and looking away results in a temporary illusion of continuing motion. The stimulus–response characteristic of the visual processing machinery is determined in part by genes and in part by experience. At times the experiential component results in a conflict with reality and

hence an illusion.

Sometimes two perceptual solutions seem equally likely to the brain, as in the case of ambiguous or bistable figures. These figures can be interpreted in one of two ways. The alternatives flip in a deterministic fashion described by a gamma function so that as long as fixation is maintained conscious control over flip frequency is minimal. Studies with event-related functional magnetic resonance imaging, a new technical development that permits examination of changes in the blood oxygen level-dependent signal due to transient neural or perceptual events, has led to investigation of the mechanisms underlying this phenomenon. A perceptual 'flip' results in a transient activation of extrastriate visual areas (amongst other neocortical sites). At the same time there is a transient deactivation of the pulvinar of the thalamus and V1. This result suggests a subcortical–cortical perceptual system in which the thalamus maintains perceptual stability and the cortex effects a reinterpretation of the perceptual content when the stabilizing influence of the thalamus wanes. The importance of the functionally specialized modules to perceptual content can also be demonstrated non-invasively by temporary, localized, functional lesions using transcranial magnetic stimulation. Thus the perception of visual motion can be temporarily disturbed by a shock at an appropriate time after a moving stimulus excites the retina.

Patients with lesions at functionally specialized sites have syndromes of loss of function and provide additional information about the organization of the visual system.. Thus lesions in the lingual gyrus at the site of V4 lead to the syndrome of achromatopsia (cortical colour blindness). Lesions in the V1 cortex provide a substrate for examining the function of retinocortical pathways that access extrastriate cortex without passage through and processing in V1. Up to 10 per cent of retinofugal fibres are of this type. A patient with a V1 lesion has been described who shows residual perception of high contrast, rapidly moving stimuli in his blind hemifield. This is associated with activation of V5, but not V1, which is destroyed. This finding raises issues about just how much cortex is required to produce a conscious experience and whether consciousness is itself organized in a modular fashion. The importance of such 'minor' pathways has not been well understood. In normal subjects, event-related electrical potential mapping in response to a perceptible moving visual stimulus shows a response in area V5 that preceeds the response in V1 by up to 40 ms. As all specialized areas that send projections to another area receive them from it in return, the conditions are established by which preprocessing of a salient visual motion stimulus in V5 might prime V1 to receive the main neural volley from the same stimulus—a potential feedforward regulatory system.

Beyond the extrastriate cortex

The awareness of the position of an object and knowledge of its physical qualities, leading to recognition, are two visual cognitive functions that depend, at least in part, on a recognition of the object's shape, colour, and direction of motion. Imaging studies suggest that the pathways activated in association with these two attributes of objects overlap substantially, but there is also some segregation relevant to each attribute in brain areas in front of the occipital cortex. Activation of posterior parts of the inferior temporal lobes occurs when objects are recognized, for example to be named. Identification of an object's position in space preferentially activates posterior parts of the parietal lobe. A third pathway, in which activity is associated with visually guided reaching for objects, has been demonstrated in the parietal lobes between those areas activated by recognition and those by awareness of position. The recognition of further pathways is to be expected because, in general, integration of visual signals with behaviour occurs at multiple anatomical levels in the human brain. Each specialized brain area that has connections to another specialized area receives signals back. Each area sends and receives signals to and from multiple other dispersed areas and draws on signals from these areas as the behavioural context demands. Yet signal traffic is not chaotic, a remarkable and often ignored result provided by functional neuroimaging.

Synaesthesia are curious experiential phenomena in which signals of one sensory modality elicit experiences in another. The condition is probably developmental and not uncommon early in life. The most frequent manifestation is the visualization of colours whilst reading or speaking. Certain colours are associated with certain stimuli, for example the first letter of a word. The experiences are not unpleasant but bizarre and are therefore frequently unacknowledged or even denied. Such phenomena provide a paradigm for examination of the interactions between sensory modalities. In aural–colour vision synaesthetes, there is abnormal activation of cortical areas beyond V4 and other early extrastriate visual areas. In fact, activity in V4 may be diminished in association with synaesthesia. Another situation in which crossmodality interactions occur is in blind Braille readers. If blindness is congenital, Braille reading is always more proficient than if it is learned after acquired blindness. Touch stimulation in Braille readers elicits visual cortex activation. However, this phenomenon involves V1 only in people with acquired blindness. It is therefore equally plausible that V1 is recruited for touch processing or that in a primed, previously 'seeing', V1 activity may be due to imagery of letters read through the touch sense. Such people can also provide important information about postperceptual processing. For example, sighted people reading visually, blind people reading with Braille, and blind people hearing spoken words all activate a posterior inferior temporal region implicating it in a supramodal function that involves word processing irrespective of input modality.

Conclusion

The brain is organized according to relatively well-ordered principles. Responses are reproducible and common to most humans. The correlation of behaviour, anatomy, and physiology promises much for an understanding of normal brain function and also for understanding better the symptoms of cerebral disease. Functional neuroimaging and advances in computerized analysis of structural images are breakthroughs for the cognitive neurosciences. It is now possible to analyse thoughts, percepts, actions, and emotions at the level of neuronal populations. Brain activity can be followed over time and in different contexts permitting the study of recovery, attention, and pharmacological modulation of brain function. New methods for measuring connection strengths between brain areas and their modification under different conditions add further opportunities for expanding knowledge as to how the human brain works in health and disease.

Further reading

Frackowiak RSJ *et al*., eds (1997). *Human brain function*. Academic Press, San Diego.

Zeki S (1993). *A vision of the brain*. Blackwell, Oxford.

Frackowiak RSJ, Gadian DG, Mazziotta JC (2000). Functional neuroimaging. In: Bradley WG *et al*., eds. *Neurology in clinical practice*, pp 665–75. Butterworth-Heinemann, Boston.

24.4 Investigation of central motor pathways: magnetic brain stimulation

K. R. Mills

The ability to stimulate percutaneously and without pain the central nervous system of awake humans has opened up new areas for neurophysiological investigation both in terms of the early diagnosis of neurological disease and the further understanding of normal and abnormal motor control. Magnetic stimulators are now available that are capable of exciting both upper and lower limb areas of the motor cortex, as well as cranial nerves, motor roots, and deeply sited peripheral nerves.

Magnetic stimulators

The magnetic stimulator is an essentially simple device; a brief pulse of electric current is passed through a coil which then generates an intense magnetic field permeating unattenuated into the surrounding media. Any electrical conductor, such as the brain, in the vicinity of the coil will have currents induced within it; these induced currents are capable of exciting cerebral neurones. Coils are placed on the scalp and may be plane circular, figure of eight, or double cone in geometry, the last being especially effective in exciting leg areas of the motor cortex. Some magnetic stimulators produce a predominantly monophasic field pulse, others produce multiphasic pulses; with the former, the side of the coil next to the scalp determines which hemisphere is predominantly excited, whereas with the latter both hemispheres are about equally excited.

Physiology

If a single anodal shock is applied to the exposed cortex of a monkey and recordings are made from the pyramidal tract, it is seen that, if stimulus intensity is sufficient, an initial wave produced by direct activation of pyramidal tract neurones (the D wave) is followed by a variable number of other waves produced by indirect trans-synaptic activation (I waves) of the same pyramidal neurones. In humans a single weak stimulus to the scalp probably excites pyramidal tract cells trans-synaptically; stronger stimuli may excite the cells directly. The effect of a single stimulus is to cause a high frequency (500 to 1000 Hz) burst of impulses to descend in the fastest fibres of the pyramidal tract; the spinal motoneurones are engaged by these impulses and if their excitability is high enough and there is sufficient temporal and spatial summation, then the motoneurones fire, causing a muscle contraction. There is considerable convergence and divergence of pyramidal tract fibres within motoneurone pools; single spinal motoneurones receive many corticospinal inputs and, conversely, single pyramidal tract fibres branch to supply many spinal motoneurones. Intrinsic hand muscles are the most easily excited from brain stimulation but all voluntary muscles appear to be accessible from cortical stimulation. The amplitude of response of a muscle depends on the intensity of the stimulus, to a lesser extent on coil placement on the scalp, but most potently on the degree of voluntary preactivation of the muscle. Thus the amplitude of response of an intrinsic hand muscle may be 20 to 30 times greater if the subject performs a gentle (5 to 10 per cent maximum) voluntary contraction of the muscle. This facilitation is probably due to both cortical and spinal cord mechanisms, voluntary action increasing the effectiveness of the stimulus at the cortex at the same time as the excitability of spinal motoneurones is increased by other pathways. The latter mechanism predominates in intrinsic hand muscles. Clearly, many factors, including mental set, affect the size of muscle response to the stimulus and it should be emphasized that this phenomenon of response variability contrasts with the identical and reproducible responses obtained from maximal electrical peripheral nerve shocks; central motor conduction studies should not be regarded simply as an extension of nerve conduction measurements.

Single scalp shocks also bring into play inhibitory mechanisms: if a subject maintains a steady voluntary muscle contraction, the initial excitation caused by the stimulus is followed by a silent period. The mechanisms underlying this are still unclear but probably involve inhibition at both cortical and spinal levels.

Safety of magnetic stimulation

A number of studies have looked at the acute effects of magnetic stimuli on animals. It has been shown that magnetic stimuli have little detectable effect on the heart rate, arterial blood pressure, or cerebral blood flow in cats. Magnetic brain stimulation has no acute effects on the human electroencephalogram or on the performance of simple cognitive tests. There have currently been no reports of adverse effects in healthy human subjects, but clearly, workers in the field should remain vigilant, especially for long-term effects.

It has been calculated that the total amount of power dissipated in the brain during magnetic stimulation is 1.8 μJ/cm ^{3}per stimulus and at the maximal rate of stimulation of 0.3 Hz, the average power dissipation is 53 μW, some five orders of magnitude below the basal metabolic rate of the brain.

It was considered prudent for early users of magnetic stimulation to exclude patients who had a history of epilepsy from their studies. Since then, magnetic stimulation has actually been used to attempt to localize epileptic foci in patients with intractable seizures. However, the risks of provoking a fit are considered small since it had been shown in cats that repetitive stimuli direct to the cortex in animals that had lesions induced by penicillin were only effective at rates above 5 Hz. Despite magnetic stimulation devices being used on many thousands of patients, many of whom must have had a predilection for epilepsy, there have been only a few reports of a fit being related to single-pulse brain stimulation.

Measurement of central motor conduction time

The latency of muscle response has a central and peripheral component and a delay due to synaptic transmission in the spinal cord. There is good evidence that, at least with limb muscle, the connection from the pyramidal tract to spinal motoneurone is monosynaptic. The central component of conduction—central motor conduction time (**CMCT**)—can be estimated by subtracting from the cortex to muscle latency an estimate of the peripheral conduction time obtained either from F wave measurement (see Chapter 24.2) or from responses evoked by root stimulation. In healthy subjects, the mean latency (± standard deviation) of responses in intrinsic hand muscle is 19.7 ± 1.2 ms and the CMCT is 6.1 ± 0.9 ms. The amplitude of responses from brain stimulation is usually compared with that obtained from maximal peripheral nerve stimulation; again there is great variability, but in healthy subjects the response from cortical stimuli is usually at least 15 per cent of that from nerve stimulation. Since many factors can influence these values, each laboratory should develop its own normative database.

Motor roots may be excited by both electrical and magnetic stimulators. The former method is preferable since it is not possible to obtain maximal responses in all healthy subjects with magnetic coils, even with optimal coil geometry, coil orientation, and coil position. Both devices activate motor roots at or just outside the intervertebral foramina and so peripheral conduction time estimated by this method omits conduction in the small segment of motor root within the spinal canal and CMCT is slightly overestimated. The method must be used, however, if F waves are unobtainable.

Compound responses from muscle may be recorded with surface electrodes, or single motor unit responses may be recorded with needle electrodes; the former method is used clinically, the latter is useful in research. A number of parameters of the surface-recorded response are useful: the maximum amplitude, the onset latency with the muscle relaxed or contracted, the threshold for evoking a response, and the variability in latency or amplitude in a series of responses.

Prolongation of CMCT has been reported in many conditions and is not specific. Delay can be produced by a variety of pathological processes: demyelination of central fibres can lead to slowing of impulse propagation in the central motor pathway; desynchronization of descending impulses can lead to loss of temporal summation at the motoneurone and delay in its firing; and loss of corticospinal axons can lead to impairment of spatial summation at motoneurones and can again

delay firing.

Multiple sclerosis

In multiple sclerosis, CMCT is prolonged in about 70 per cent of cases when there are clear clinical signs of a pyramidal lesion in the particular limb (Fig. 1). The delay in some cases is very considerable, CMCT may be up to five times longer than in controls. It is likely that, in these cases, demyelination of central fibres is the mechanism leading to delay. In other cases, delay is more modest, only a few milliseconds, and the mechanism is less certain. Abnormal central motor conduction appears to correlate most closely with exaggerated reflexes and spasticity rather than with weakness or cerebellar signs in the limb. Abnormal CMCT from leg areas of motor cortex also correlates with the finding of extensor plantar responses.

Fig. 1 Slowing of central motor conduction in multiple sclerosis. Compound muscle action potentials are recorded with surface electrodes over the left and right abductor digiti minimi muscles. Stimuli are given to the ulnar nerve at the wrist (left), the C7/T1 motor roots (middle), and the motor cortex (right). Onset latencies are shown and the variability of responses from cortical stimulation can be seen. On the left CMCT is 7.4 ms, but on the right is prolonged at 13.9 ms.

Central motor conduction can be abnormal, however, even in the absence of clinical signs. In a large series, it was found that central conduction was abnormal in 20 per cent of cases of multiple sclerosis with no motor signs in the particular limb. The technique can thus be used as a screening test for multiple sclerosis, although it compares unfavourably with visual evoked potentials, which have a higher rate of abnormality in the absence of clinical signs. This may merely reflect the greater accuracy with which the motor system can be examined clinically. Central motor studies may also be helpful in deciding on the importance of equivocal motor signs, such as mild impairment of fine finger movements.

Motor neurone disease

In motor neurone disease, the most common abnormality is a raised threshold for excitation of the motor cortex, although in early cases the threshold may be reduced. In many cases responses cannot be obtained even with the strongest stimuli applied in optimal conditions. CMCT may be prolonged, but usually only modestly, and responses are often reduced in amplitude in comparison with responses evoked by maximal nerve stimulation. The test can be used to confirm an upper motor neurone component to weakness when lower motor neurone signs predominate or for detecting an upper motor neurone lesion in a limb without clinical signs.

Cerebrovascular disease

In stroke, responses in an affected limb may be normal, delayed, or absent, with abnormality grossly paralleling the clinical abnormality. Central motor conduction studies have been used to predict outcome of stroke; if performed within the first 48 h after the ictus, a poor outcome at 6 months is predicted by absent responses and a favourable outcome by normal responses. Whether the prediction is superior to that made purely on clinical grounds is uncertain, but at least the method is quantitative and can be used serially to follow recovery.

Movement disorders

Most studies have shown central motor conduction to be normal in Parkinson's disease, multiple system atrophy, Wilson's disease, Huntington's disease (including at-risk relatives), dystonia, and progressive supranuclear palsy. In some cases of Wilson's disease, central conduction delays have been found. In all these conditions, however, there may be subtle changes in motor cortex excitability detectable as a change in threshold or an abnormal inhibitory response to appropriately timed pairs of cortical stimuli.

Degenerative neurological diseases

A number of rarer degenerative diseases have been investigated with the technique: Friedreich's ataxia often shows delayed and dispersed responses, as does early-onset cerebellar ataxia with retained reflexes, the severity of the abnormalities reflecting disease duration. In late-onset cerebellar degeneration on the other hand, the responses are normal in 62 per cent of cases. In hereditary spastic paraparesis and tropical spastic paraparesis, responses from upper limb muscles are usually normal, whereas those from the lower limbs are delayed or absent. Abnormalities of central motor conduction have also been described in some cases of hereditary motor and sensory neuropathy types I and II, the abnormalities being found especially in those patients with additional upper motor neurone signs. Central motor conduction abnormalities have also been described in a family with hereditary motor and sensory neuropathy with pyramidal signs (HMSN type V).

Spinal cord lesions

Magnetic brain stimulation has been used to assess the completeness of spinal cord injury. A variety of facilitating techniques must be used; the modulation of flexion reflexes by brain stimuli has been shown to be useful in establishing whether injury is complete; in 4 of 26 patients evidence of incomplete lesions was found in patients with clinically complete spinal cord injuries. In compressive myelopathy, by recording from a variety of upper limb muscles, CMCT can be used to localize more accurately the compressed cord segment.

Paediatric applications

The central conduction time in a group of 457 normal subjects between the ages of 32 weeks and 55 years has been determined. It was found that central conduction time decreases rapidly over the first 2 years of life and then remains constant at the adult value. In contrast, peripheral conduction increases in proportion to arm length after the age of 5 years. It is suggested that this constant central delay could be useful during the acquisition of motor skills. Central motor conduction has been studied in a range of neurological diseases in children. For example, in 13 of 20 children with an upper motor neurone syndrome of varied aetiology, the central conduction time was abnormal, but magnetic resonance imaging and/or computed tomography scans showed focal abnormalities in only seven. In 15 children with extrapyramidal syndromes, the central conduction time was normal.

Use of brain stimulation for neurosurgical monitoring

Although somatosensory motoring has been shown to be of use during neurosurgical procedures to alert the surgeon of the possibility of cord damage, the use of motor monitoring is far more relevant since paraplegia is one of the most feared, although rare, outcomes of surgery near the cord. Electrical brain stimulation and recording from the cord by epidural electrodes has been achieved; responses consist of a series of waves analogous to the D and I waves recordable in primates. Magnetic stimulation appears to produce I waves but the responses are very sensitive to the depth of anaesthetic agents, especially nitrous oxide. If the aim of monitoring is merely to stimulate the motor cortex, there seems little to be gained by using magnetic stimuli in favour of the electrical method since the pain of the

procedure is not a factor.

Conclusion

Non-invasive magnetic brain stimulation has shown itself to be a powerful technique in the diagnosis and prognosis of disorders of the central motor system, as well as providing new insights into the normal control of human voluntary movement. It can be used serially to monitor progress or the effects of drugs, can be used safely in neonates and children, and can be used to demonstrate short- or long-term plasticity in the human nervous system after injury.

Further reading

Levy WJ *et al.*, eds (1991). Magnetic motor stimulation: basic principles and clinical experience. *Electroencephalography and Clinical Neurophysiology*, Suppl 43.

Mills KR (1999). *Magnetic stimulation of the human nervous system*. Oxford University Press.

Rothwell JC *et al.* (1991). Stimulation of the human motor cortex through the scalp. *Experimental Physiology* **76**, 159–200.

24.5 Imaging in neurological diseases

Andrew J. Molyneux and Philip Anslow

Introduction

The modern imaging techniques of X-ray computed tomography (**CT**) and magnetic resonance imaging (**MRI**) for the demonstration of structural neurological disease have developed rapidly over the last 20 years. They have done more than any other single development to revolutionize the diagnosis and treatment of structural neurological disease. More recently, a variety of functional imaging techniques have been developed that enable functional and biochemical information to be obtained from the living brain.

Current techniques for neuroimaging

Computed tomography (CT)

X-ray computed tomography was developed by the British scientist and engineer Godfrey Hounsfield during the early 1970s. CT was the first technique to provide non-invasive and cross-sectional images of the brain. It was introduced into clinical usage at the Atkinson Morley Hospital in Wimbledon, London, in 1972 and results were published in 1973. This was the start of a complete revolution in radiological imaging, for which Hounsfield received the Nobel prize for medicine in 1979.

CT rapidly became the mainstay of the diagnosis of structural brain disease until the advent of magnetic resonance imaging (MRI) into widespread clinical use during the late 1980s and early 1990s. However, CT remains an extremely valuable and essential tool, particularly in the acute situation and in countries and regions where the cost of MRI systems is prohibitively expensive.

CT produces a series of cross-sectional images, usually in the axial plane (hence the acronym CAT scan, standing for computed axial tomography). During exposure to an X-ray beam, a detector array spins around the patient and measures the absorption coefficients of tissues within the beam. It is the different coefficients that provide contrast.

Magnetic resonance imaging (MRI)

Magnetic resonance imaging is a fundamentally different method of obtaining images. It relies on a powerful static magnetic field and the various properties of the protons (hydrogen ions) in the different tissues. When a strong magnetic field is subjected to certain radio waves of a specific frequency (radiofrequency) the protons will resonate at an exact frequency that depends on the field strength of the magnet. The radio signal emitted back by the protons when the radiofrequency is switched off can be detected in a receiver coil and a detailed image built up. The resolution of modern MRI scanners is extremely high, as is the sensitivity of the images obtained for the detection of intracranial anatomy and pathology.

Many different radio-pulse sequences are of use in MRI, which are determined by the way the radio signals are timed. They detect different aspects of tissue properties by what is called 'the relaxation times of the protons'; times that will vary according to the proton-containing tissue and the relative mobility of the water molecules. The most commonly used sequences are what are termed '*T*1-weighted and *T*2-weighted sequences'. The appearances of these scans are quite different: for example, cerebrospinal fluid (**CSF**) is white on *T*2 and dark on *T*1 images (Fig. 1 and Fig. 2). Some tissues such as fat and some blood breakdown products (for instance, methaemoglobin) will appear bright on *T*1- and *T*2-weighted images.

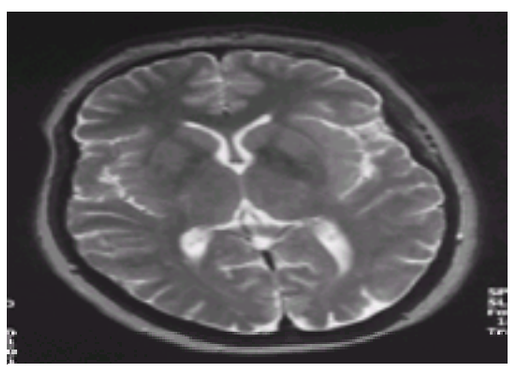

Fig. 1 Normal axial *T*2-weighted image of a brain at the level of the ventricular system. Note that the cerebrospinal fluid is white, the white matter is dark, and the grey matter is lighter than the white matter. This is the most commonly used MRI sequence and it is usually the most sensitive in the detection of pathological processes.

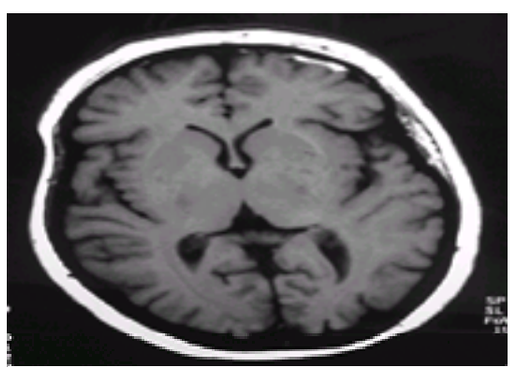

Fig. 2 Axial *T*1-weighted unenhanced MRI image at the similar level of the ventricles as in Fig. 1, showing the CSF dark and the white matter lighter than the grey

matter.

Contrast enhancement in brain imaging

Intravenous contrast in brain imaging is frequently used to determine the vascularity of structures and whether the blood–brain barrier is intact. It will show the extent and patterns of enhancement in tumours, infarcts, and inflammatory lesions. For CT scanning, the same iodinated contrast agents that are used in general vascular imaging are used. In MRI, gadolinium-labelled compounds, which shorten the *T*1 relaxation time, are used. These provide extremely useful information and show the same patterns of enhancement as the iodinated contrast media used for CT scanning, although the sensitivity of MRI contrast agents are significantly greater.

Cerebral angiography

This is used to demonstrate the intra- and extracerebral vessels. The procedure is nearly always performed by transfemoral catheterization of the neck vessels. Before the introduction of CT it was the main means of diagnosing intracranial pathology, particularly of masses and stroke-causing lesions such as vessel occlusions or haemorrhages. The main indication now for angiography is intracranial haemorrhage or suspected extracranial carotid or vertebral stenosis. Non-invasive methods, such as Doppler ultrasound of the neck vessels and magnetic resonance angiographic imaging (which does not require contrast media), has reduced the number of patients requiring invasive intra-arterial angiography, particularly for suspected ischaemic cerebrovascular disease.

Other imaging techniques including functional imaging

Although these are used less frequently, nuclear medicine studies using radioactive labelled isotopes—generally with technetium-99m and **HMPAO** (hexamethylpropylene amine oxide) as the ligand—and a gamma camera can be used to produce perfusion imaging scans of the brain. This technique is known as single-photon computed tomography (**SPECT**). Positron emission tomography (**PET**) scanning requires a cyclotron to produce the very short-lived isotopes of carbon and oxygen, and is primarily used as a research tool for investigation of the functional imaging of the brain.

Functional MRI and spectroscopy

High-field MRI 1.5 tesla or greater, up to 4 tesla, magnetic fields can also be used to provide functional information on brain function and biochemistry (spectroscopy from either hydrogen or phosphorous nuclei). These techniques are not yet in routine clinical use.

Imaging of common neurological diseases

Cerebrovascular disease and stroke

The most frequent neurological presentation is that of acute stroke. Patients presenting with a sudden onset of neurological deficit should be deemed to have suffered a vascular event until proved otherwise. In practice the clinical diagnosis of stroke is very accurate, provided an adequate history is available. The primary role of imaging in patients with acute stroke is to identify whether it is ischaemic or haemorrhagic in origin, or where there is doubt about the underlying pathology based on the clinical history. CT scanning provides a completely reliable way of excluding primary intracerebral haemorrhage as a cause of acute stroke, provided it is performed within about a week of onset. In ischaemic stroke, depending on the timing of the examination relative to the onset of neurological deficit, it will variably detect acute infarction.

Cerebral infarction

Early appearances

Within hours of a stroke, the CT scan may show a vague low attenuation or slight swelling and effacement of the sulci in the area of damage, or it may be normal. Standard MRI imaging may also be normal at this stage. However, certain more recently introduced MRI techniques, known as 'diffusion-weighted imaging', can show abnormalities even within less than an hour of onset that reflect alterations in the state and mobility of the protons in tissue water.

It should be emphasized that a negative CT or MRI scan in a patient with a clinical acute stroke does not mean that the patient has not had a stroke. It just means that imaging has not detected an area of infarction (Fig. 3 and Fig. 4).

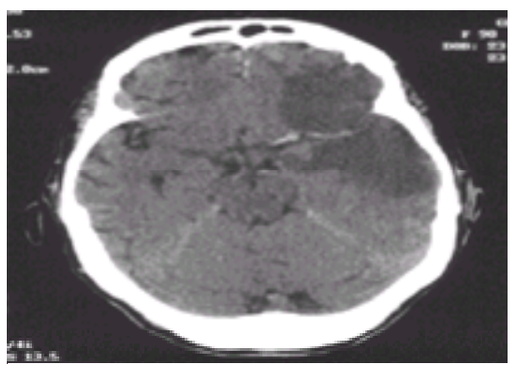

Fig. 3 CT scan showing an acute middle cerebral territory infarction within a few hours of onset of the neurological deficit. (Note the increased density in the left middle cerebral artery representing a thrombus.)

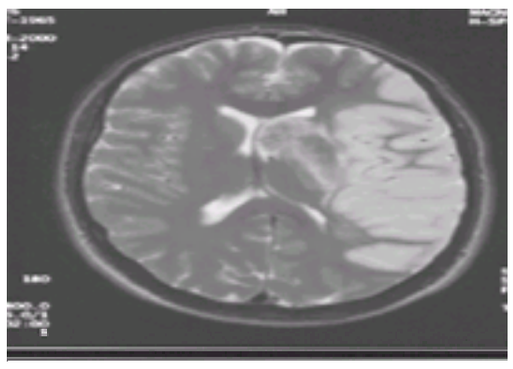

Fig. 4 *T*2-weighted MRI scan of the same patient as in Fig. 3 showing the extensive high *T*2 signal affecting grey and white matter in the left middle cerebral artery territory.

Later imaging findings

There is progressive development of a low-attenuation area on CT scanning, or a high *T*2 signal area on MRI scanning, in the area of damage. There may also be some swelling around the area, representing oedema with effacement of sulci and ventricles. It may be impossible to identify those areas that are truly infarcted and the area, which is ischaemic and may recover, called the 'ischaemic penumbra'. Over a period of a week or more the area of infarction matures with the development of a progressively better defined area of low attenuation, and loss of volume in the damaged area over time.

Intravenous contrast is frequently used when the nature of pathology in patients with cerebral infarction and its differential diagnosis from a mass lesion is in doubt. This will demonstrate increased vascularity of tissues and areas where the normal blood–brain barrier is disrupted; normal cerebral tissue will not enhance and the tight junctions of the normal blood–brain barrier will not allow contrast into the cerebral tissues. In areas of ischaemia or infarction there is diffuse cortical enhancement in ischaemic/infarcted areas. It is possible to confuse these appearances with tumours.

Intracranial haemorrhage

Primary intracranial bleeding (**PICH**) into the brain parenchyma is easily detected on CT scanning. Blood appears as an area of high attenuation on CT imaging (Fig. 5).

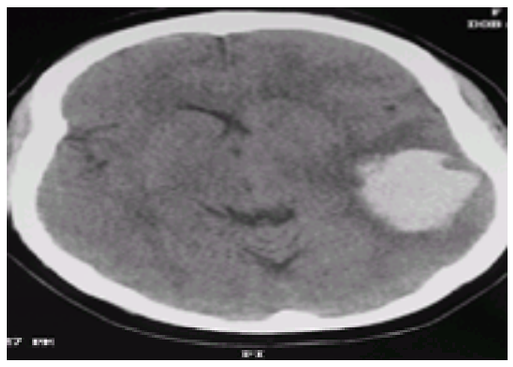

Fig. 5 CT scan of a patient with an acute stroke due to a large intracerebral haemorrhage in the left temporal lobe.

Blood in the subarachnoid space may be visible around the base of the brain as a white layer, in contrast to the normal dark outline of CSF in the basal cisterns. If a scan is carried out within 24 h of a subarachnoid haemorrhage (**SAH**) it will usually be positive for the presence of blood (about 90 per cent of the time); with a large SAH, the scan will remain positive for 3 to 4 days (Fig. 6).

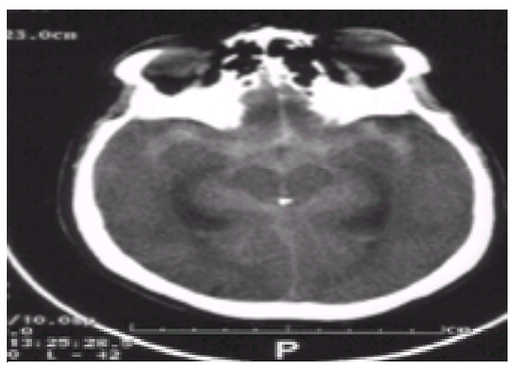

Fig. 6 CT scan of a patient with a subarachnoid haemorrhage. Note that the CSF surrounding the brainstem and in the basal cisterns is high attenuation (bright) compared with normal CSF, which on CT would be dark.

However, the lack of visible blood on CT scanning does not exclude the diagnosis of a subarachnoid haemorrhage. Lumbar puncture is essential if this diagnosis is suspected and if CT is negative or equivocal. The key clinical differential diagnosis is of acute meningitis or a very small subarachnoid haemorrhage undetectable on CT, which may result from rupture of an intracranial aneurysm.

After an intracranial haemorrhage, a decision must be made whether to investigate patients in more detail for the presence of an underlying lesion responsible for the haemorrhage, such as an intracranial aneurysm or vascular malformation. Selection of which patients should undergo cerebral angiography is sometimes difficult. However, all patients under about 50 years of age with primary intracranial bleeding but without a typical hypertensive bleed should probably undergo cerebral angiography to search for an underlying vascular malformation. All patients who survive a primary subarachnoid haemorrhage and who are in good condition should undergo cerebral angiography to detect the presence of an berry aneurysm that may be responsible for the haemorrhage. Recently ruptured aneurysms have a high likelihood of re-bleeding, as much as 30 per cent in the first 4 weeks after the haemorrhage. Without treatment there is a 50 per cent mortality by 6 months. These lesions should be detected and treated if possible, either by surgical clipping or the newer endovascular techniques using detachable platinum coils (Fig. 7 and Fig. 8). A recent large multicentre randomized trial has reported improved 1 year outcomes following coil treatment compared with surgical clipping.

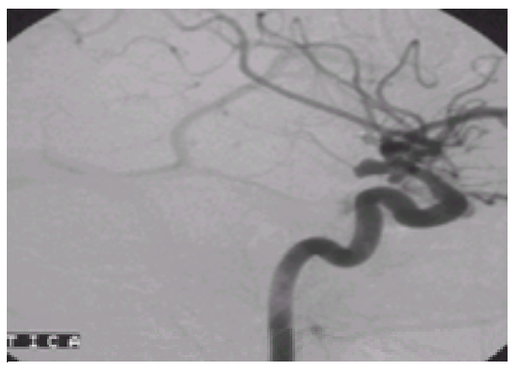

Fig. 7 Digital subtraction cerebral angiogram showing two aneurysms arising from the internal carotid artery that had recently ruptured.

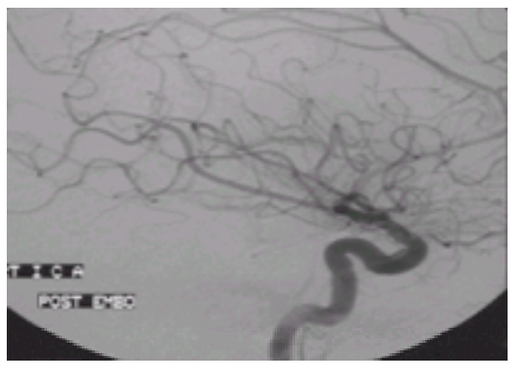

Fig. 8 Digital subtraction cerebral angiogram following placement of detachable platinum coils in the aneurysm to occlude the flow and prevent re-bleeding.

Other vascular diseases

Cerebral venous sinus thrombosis

This uncommon, potentially fatal, condition presents with severe headache, confusion, variable neurological deficits, and sometimes seizures. Since it can be difficult to detect on CT scanning, the diagnosis is best made on MRI scanning, where a lack of flow void is seen on some sequences (*T*2-weighted) and a high signal on *T*1-weighted sequences is seen in affected dural sinuses. Flow-sensitive MRI sequences demonstrate the obstructed sinus well. Anticoagulation treatment is urgently required since progression of the thrombosis can lead to intracerebral haemorrhage and fatal venous infarction (Fig. 9 and Fig. 10).

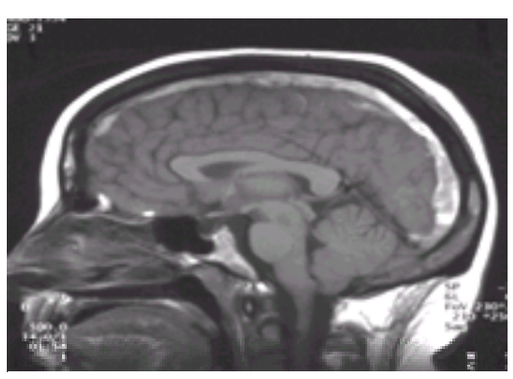

Fig. 9 Sagittal *T*1-weighted MRI scan showing a high signal in the superior sagittal sinus representing an extensive clot throughout the sinus. This 30-year-old woman presented with severe headaches, a depressed level of consciousness, and papilloedema.

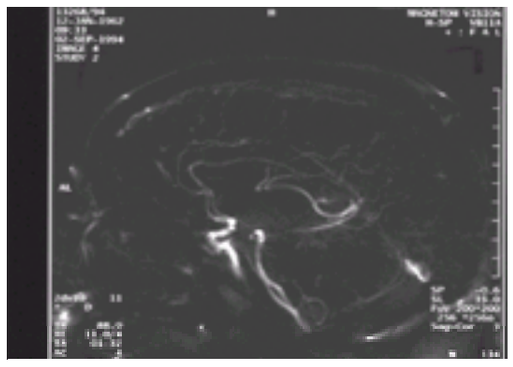

Fig. 10 Sagittal magnetic resonance venogram image showing the lack of flow in the superior sagittal sinus. This sequence is sensitive to flow and is useful in demonstrating whether a venous sinus or artery is blocked, if it is unclear from the normal sequences.

Inflammatory diseases of the nervous system

Multiple sclerosis

The most common neurological disease after stroke is multiple sclerosis. Imaging plays a crucial role in the diagnosis of this condition. However, it is important to understand that MRI alone cannot provide the diagnosis. The investigation must be placed in the context of the clinical presentation and the history and findings on neurological examination.

When patients present with symptoms of spinal cord disease, MRI of the brain and spine is indicated. The whole spinal cord is imaged to ensure that no spinal cord compressive lesion is responsible for the neurological condition. An inflammatory plaque will be seen in the spinal cord, particularly in the cervical cord, in a number of cases (Fig. 11); although failure to identify such a lesion does not mean that one is not present. Current imaging will not detect all spinal or indeed all brain lesions. However, imaging of the brain in patients with a spinal cord lesion is helpful in determining the presence of further lesions in the brain, thereby indicating a multifocal pathology. The difficulty comes in older patients over 45 years of age, where the frequency of incidental lesions in the brain presumed to be due to age-related vascular pathology becomes much more common. The pattern and extent of the white-matter lesions in patients with multiple sclerosis often characteristically affect the corpus callosum and deep white matter. However, it has again to be emphasized that the diagnosis of multiple sclerosis must not be based purely on imaging findings, attacks must be disseminated in time as well as location in the nervous system. An incorrect diagnostic label has profound consequences for the patient (Fig. 12 and Fig. 13).

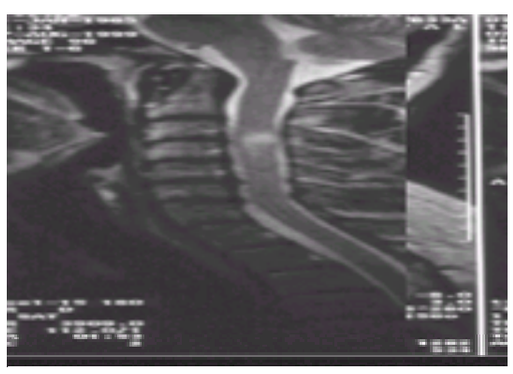

Fig. 11 Sagittal *T*2-weighted image of the cervical cord showing a high-signal lesion lying at the C3 level with some associated swelling of the cord. This is typical for

acute demyelination.

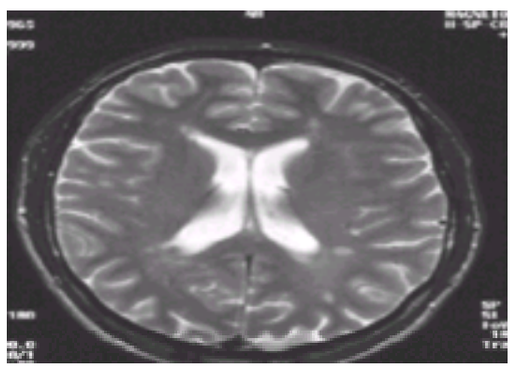

Fig. 12 Axial *T*2-weighted MRI showing high-signal lesions in the white matter around the ventricles and in both hemispheres.

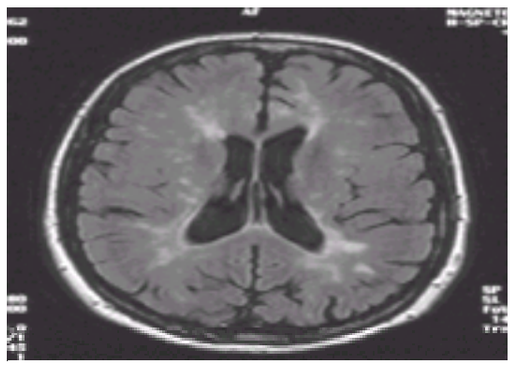

Fig. 13 This is known as a 'Flair sequence'. It detects lesions that give a high *T*2 signal but suppresses the signal from the CSF in the ventricles, so that lesions adjacent to the ventricles are more evident.

Neoplasms

Primary intracranial tumours

Primary intracranial tumours can be divided into those arising outside the brain and those arising in the cerebral substance (intra-axial or extra-axial). The range of pathology of the two locations is fundamentally different, as is often the prognosis. The primary objective of neuroimaging is to establish whether a mass lesion is present and to determine whether the lesion is intra- or extra-axial.

Extrinsic brain tumours

The most common tumours arising from structures outside the brain are meningiomas and acoustic schwannomas arising from the VIIIth nerve (frequently called 'acoustic neuroma'). Both these tumours are usually benign and present with symptoms of local pressure: VIIIth cranial nerve tumours can produce sensorineural deafness and/or sometimes dizziness, while meningiomas over the cerebral convexity may cause seizures.

The imaging characteristics of meningiomas and acoustic neuromas are similar. CT scans usually show a slight hyperdense mass causing local displacement of cerebral tissue. Masses generally enhance uniformly following the administration of intravenous contrast, although they may occasionally contain areas of low attenuation that may represent necrosis or occasionally cyst formation within the tumour (Fig. 14 and Fig. 15).

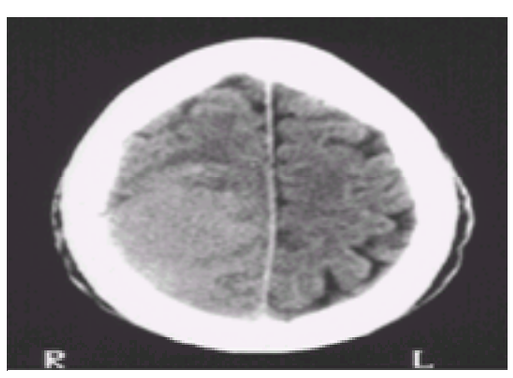

Fig. 14 Typical appearance of a meningioma. CT scan without contrast showing a high-attenuation mass lying over the surface of the brain.

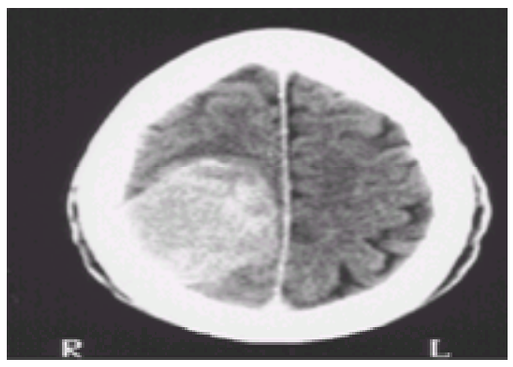

Fig. 15 CT scan after intravenous contrast showing the uniform enhancement of the mass lying over the surface of the brain.

Although MRI shows lesions that give a uniform intermediate signal on *T*1 and *T*2-weighted sequences, gadolinium administration yields uniform similar but intensely enhanced images (Fig. 16). Differentiation between intrinsic and extrinsic lesions is usually easier on MRI than on CT scanning because of the multiplanar capability

of MRI compared with CT. MRI is also superior for surgical planning purposes, to establish the relationship of the tumour to adjacent structures such as the venous sinuses or the skull base.

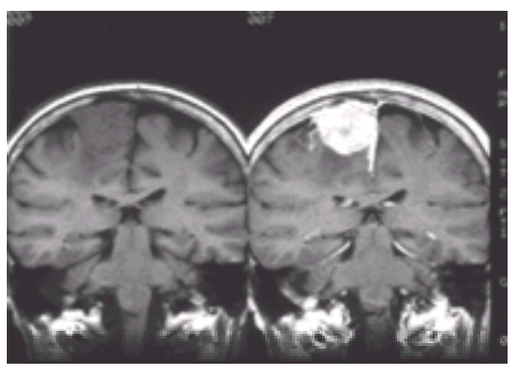

Fig. 16 Coronal *T*1-weighted MRI scan showing the typical appearances of a meningioma before and after intravenous gadolinium.

Screening for acoustic schwannomas

The most frequent presentation of an acoustic schwannoma is sensorineural hearing loss, but only a small percentage of patients with these symptoms will have an acoustic neuroma. Nevertheless, such patients should be screened to exclude the presence of a tumour. This is best achieved by a limited MRI scan, either a high-resolution *T*2-weighted or a 3D sequence. A modern MRI scanner can perform such a scan in less than 15 min.

Intrinsic cerebral tumours

When a mass lesion arises from the brain substance, the range of tissues of origin is completely different. Most tumours arise from glial cells and are therefore classified as gliomas in broad terms. However, the range of biological behaviour of these tumours is very wide. Similarly, the imaging findings can vary greatly, reflecting this biological behaviour. The most common single brain tumour is the malignant glioma or glioblastoma multiforme. These tumours are also referred to as high-grade gliomas.

Glioblastoma multiforme

These tumours show widespread infiltration and a mass effect on CT and MRI, with a high *T*2 signal on MRI and a low attenuation on CT. The distinction between oedema and tumour infiltration may be difficult on imaging grounds.

However, the pattern of glioblastoma multiforme is usually very characteristic on imaging, with extensive enhancement following contrast on both CT and MRI scanning. Irregular margins of low-signal areas due to tumour necrosis are common. It is also common for cysts to form in association with both glioblastoma multiforme and lower grade gliomas (Fig. 17). Figure 18, Figure 19, and Figure 20 show the MRI appearances of a glioblastoma.

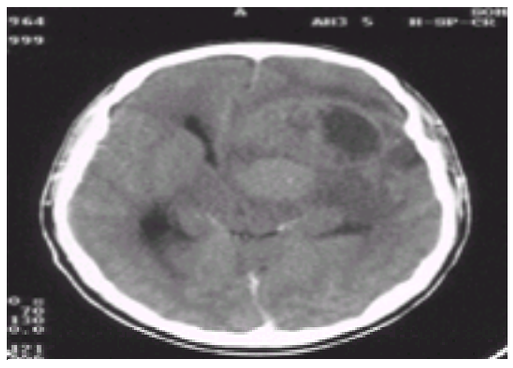

Fig. 17 Contrast-enhanced CT scan showing a large, deeply situated mass in the left hemisphere, with considerable enhancement postcontrast and a cystic or necrotic component with a low-attenuation area. This lesion was confirmed as a glioblastoma multiforme.

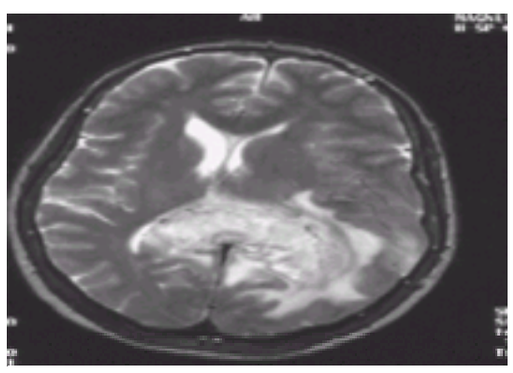

Fig. 18 Axial *T*2-weighted image of a large glioblastoma involving the corpus callosum.

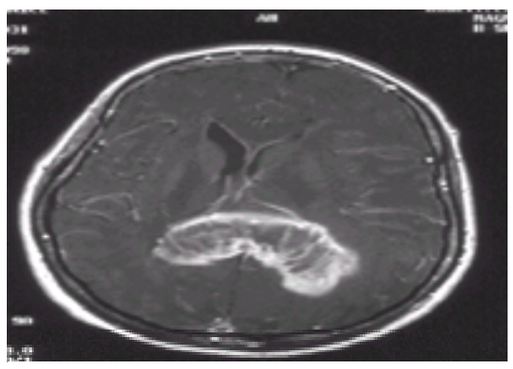

Fig. 19 Contrast-enhanced axial *T*1-weighted MRI image of glioblastoma showing marked irregular contrast enhancement of the margins of the tumour, with lack of central enhancement reflecting the extensive necrosis that is often a feature of these tumours.

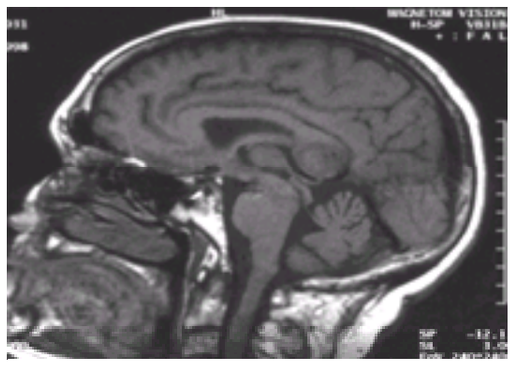

Fig. 20 Sagittal *T*1-weighted image without contrast showing the marked enlargement of the splenium of the corpus callosum depicted in the same patient as Fig. 18 and Fig. 19.

Astrocytoma

These tumours are less aggressive than glioblastoma multiforme and range from benign lesions, which may remain stable for many years, to aggressive malignant lesions. Astrocytomas often form cysts and are seen in children as what are termed 'pilocytic astrocytomas'. These are often benign. The imaging findings show a diffuse mass effect on CT and MRI, with low attenuation on CT and on *T*1-weighted MRI sequences but a high signal on *T*2-weighted MRI. Enhancement is very variable; sometimes there is considerable enhancement in pilocytic astrocytoma in children but often little in the way of enhancement in many low-grade astrocytomas in adults (Fig. 21).

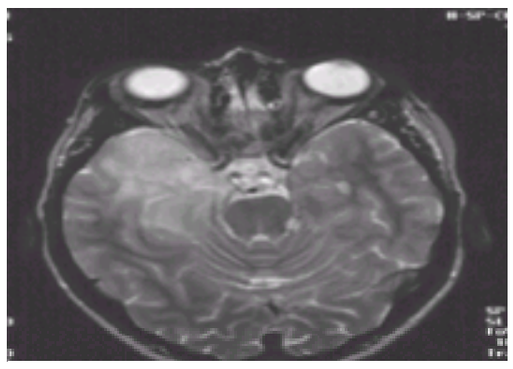

Fig. 21 Axial *T*2-weighted MRI showing a diffuse high-signal lesion in the right temporal lobe with a diffuse mass effect and sulcal effacement. These are typical appearances for a low-grade glioma. Although the differential diagnosis includes herpes encephalitis, the clinical presentation is completely different.

Oligodendroglioma

These tumours are the most benign of the intrinsic cerebral tumours. They often present with seizures rather than neurological deficit. Their radiological hallmark is calcification, best detected on CT scanning. Calcification may be invisible on MRI. The time course of these tumours may be very long, often evolving over 10 to 20 years. Oligodendrogliomas may remain static for long periods (Fig. 22 and Fig. 23).

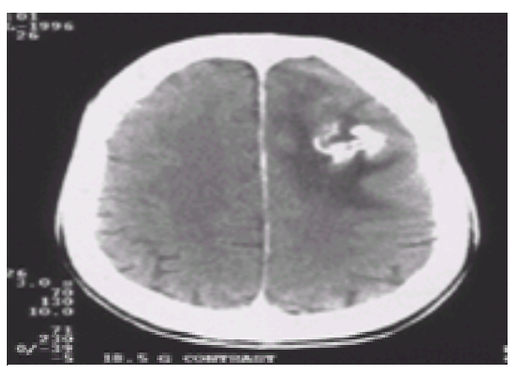

Fig. 22 CT scan of a partially calcified oligodendroglioma in the left frontal lobe.

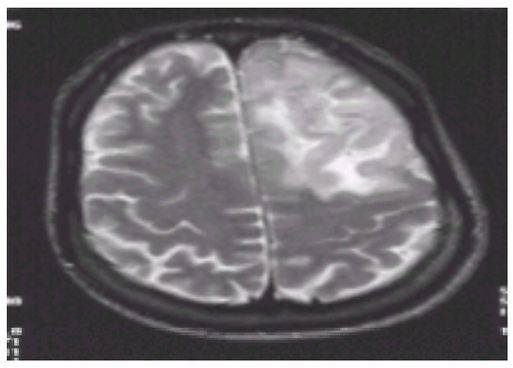

Fig. 23 MRI scan of the same pathology as in Fig. 22 with a diffuse high signal throughout the left frontal lobe.

Posterior fossa tumours

Intrinsic posterior fossa tumours are the most common intracranial tumours in children. The most common lesion is a medulloblastoma (Fig. 24). These usually arise in or near the midline posterior fossa in relation to the cerebellar vermis, they account for about 75 per cent of posterior fossa tumours in children. Other tumours commonly encountered are ependymomas and pilocytic astrocytomas, both of which have a better prognosis than medulloblastomas. Medulloblastomas and ependymomas commonly metastasize down the spinal canal, producing what are known as 'drop metastases' to the lumbar or sacral region.

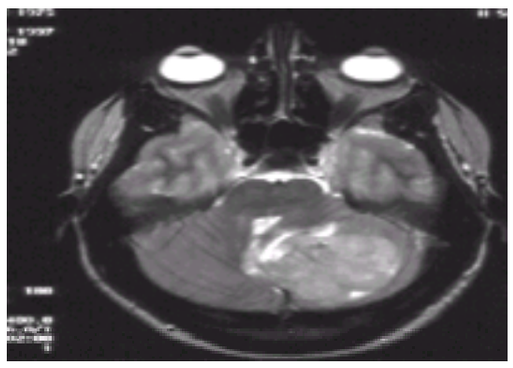

Fig. 24 Large medulloblastoma in the left cerebellar hemisphere demonstrated on *T*2-weighted MRI causing severe compression of the fourth ventricle.

Other intracranial tumours

Colloid cyst

This is a very characteristic benign lesion that arises at the foramen of Munro and causes obstructive hydrocephalus. Colloid cysts are usually readily detectable on CT and MRI and absolutely characteristic in their position, though the attenuation and signal characteristics can vary quite widely (Fig. 25).

Fig. 25 CT scan showing the typical appearance of a colloid cyst arising at the foramen of Munro.

Pituitary region tumours

MRI is the investigation of choice for suspected pituitary lesions. CT is a second best imaging modality.

These tumours, which arise outside the brain itself, are associated with a characteristic range of pathology. The most common lesion is a non-functioning pituitary adenoma, followed by hormonally active tumours, namely Cushing's disease (producing ACTH), prolactinomas, growth-hormone-secreting tumours (acromegaly). All these have similar characteristics, but their size varies widely: ACTH-secreting adenoma are usually very small and may not be detectable even on high-quality MRI imaging. Non-functioning adenomas tend to present late, often with visual loss and/or pituitary failure due to the large size and optic chiasmal compression (Fig. 26 and Fig. 27).

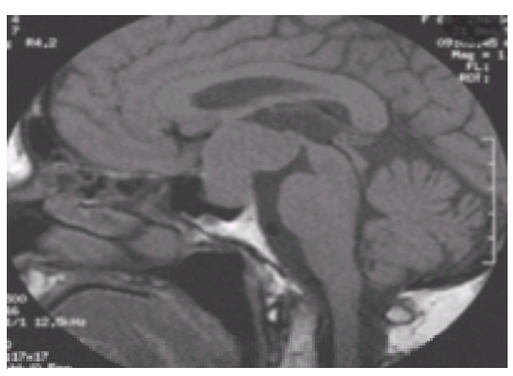

Fig. 26 Sagittal *T*1-weighted MRI showing a large pituitary adenoma extending into the suprasellar cistern.

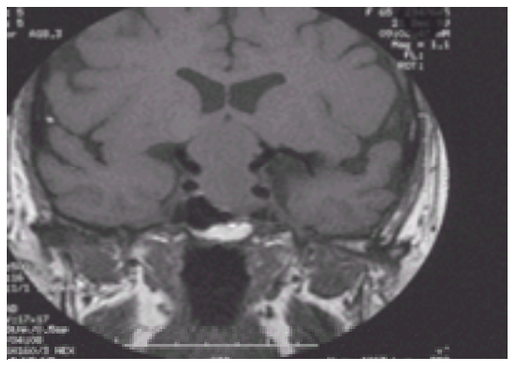

Fig. 27 Coronal *T*1-weighted MRI showing a large pituitary adenoma extending into the suprasellar cistern.

Craniopharyngioma

This specific benign tumour arises in the hypothalamic region usually in young patients, and presents with either visual loss and/or pituitary failure. The characteristic findings on CT are calcification, which may also be seen on plain skull films. There is often a cystic as well as a solid component to the lesion.

Brainstem gliomas

These relatively uncommon tumours occur at a relatively young age. However, because of their location there is no prospect of any surgical approach and if any treatment is appropriate, it is usually radiotherapy. Brainstem gliomas may vary widely in their aggressiveness, from rapidly progressive lesions behaving like glioblastomas to indolent lesions that may remain static for many years (Fig. 28).

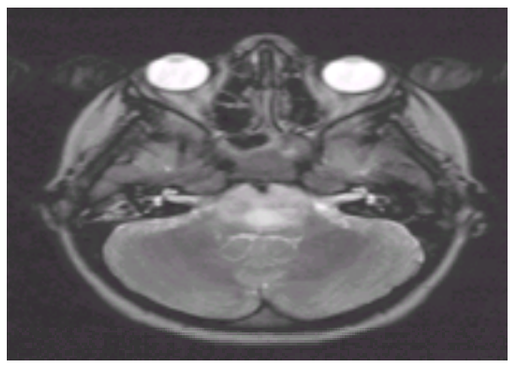

Fig. 28 Brainstem glioma shown on axial *T*2-weighted MRI with a diffuse high signal within the pons causing a local mass effect.

Secondary cerebral tumours

These are amongst the most common intracranial tumours and may be the presenting feature in some patients. Lung, breast, and gastrointestinal tumours as well as melanomas especially metastasize to the brain.

Secondary tumours may be solitary or multiple and are fairly characteristic on the imaging, with intracranial masses surrounded by oedema and frequently with enhancement after intravenous contrast (Fig. 29). The differential diagnosis of multiple ring-enhancing lesions in the brain is between cerebral metastases and abscesses. Sometimes this distinction cannot be made on imaging grounds and other evidence of the underlying cause must be sought; if necessary, biopsy or aspiration may be required.

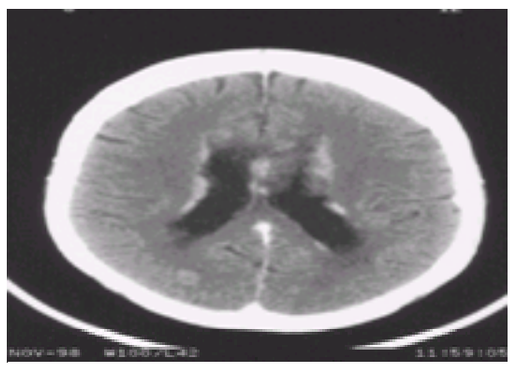

Fig. 29 Enhanced CT of a brain showing extensive secondary deposits around the cerebral ventricles from an oat-cell carcinoma of the bronchus.

Meningeal deposits of primary or secondary tumours are relatively uncommon, but they do occur and may be difficult to detect. MRI with gadolinium enhancement is the most sensitive detection method if the cerebrospinal fluid examination is negative for abnormal cells.

Intracranial infections

Although intracranial infections are less common than tumours, it is vitally important that they be detected as urgent and definitive diagnosis and treatment is essential to their effective management.

Bacterial infections

Pyogenic brain abscesses may be single or multiple. In the early stage they may not be particularly well defined, but by the time they present for scanning they often show a characteristic ring-enhancing mass surrounded by oedema (Fig. 30). If a pyogenic abscess is suspected then burr-hole aspiration is mandatory to establish the diagnosis and to drain the abscess. Abscesses may be seen at various stages of evolution if associated with a septicaemic illness. The source is either due to blood spread, or direct spread from the infection in the paranasal sinuses or the mastoid.

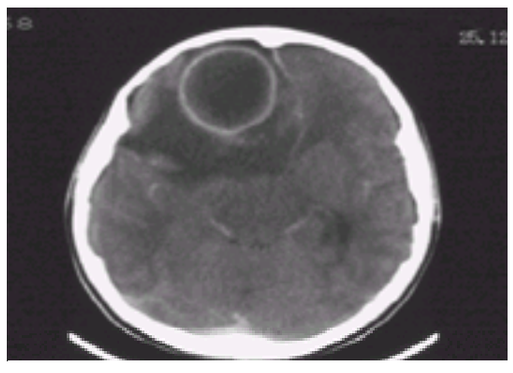

Fig. 30 Enhanced CT scan showing a large ring-enhancing frontal mass due to a pyogenic abscess. The differential diagnosis for these appearances is from a metastasis, or a glioma can occasionally appear similar.

Meningitis is not usually diagnosed by CT or MRI and lumbar puncture remains the method of choice. If imaging is performed, however, it may show a mild communicating hydrocephalus and enhancement of the meninges following intravenous contrast. Note that mild communicating hydrocephalus is not a contraindication to lumbar puncture (see below).

Subdural empyema

This is rare, but important, intracranial infection due to spread from a paranasal sinus infection. Pus accumulates in the subdural space causing a spreading cortical thrombophlebitis. Empyema is usually due to the anaerobic bacterium, *Streptococcus milleri*. Such abscesses are rapidly fatal if they are not treated aggressively with antibiotics and neurosurgical drainage. CT findings are those of a thin subdural collection of fluid, which spreads over the surface, and often alongside the falx. MRI is more sensitive in the detection of the small subdural collections, but it is unnecessary if the diagnosis is clear on CT scans (Fig. 31).

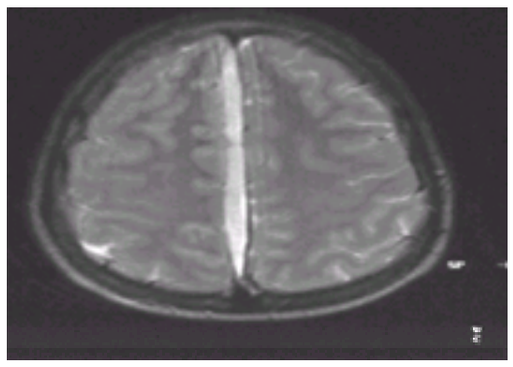

Fig. 31 Axial *T*2-weighted MRI showing a collection of fluid alongside the falx in a patient with a large subdural empyema.

The underlying brain appears swollen and tight with the sulci obliterated, it will show moderate meningeal enhancement following intravenous contrast.

Tuberculosis

This may manifest itself as either abscesses or granulomas in the brain or a basal meningitis. If the meninges are involved there is almost invariably a degree of hydrocephalus.

Viral encephalitis

The most common cerebral viral infection is herpes simplex encephalitis (**HSE**). The imaging findings are often fairly typical, though CT scan changes may be very subtle during the early phase. CT scans show a mild swelling in the temporal region and diffuse low attenuation, as the focus of the pathology in HSE lies in the hippocampal region and medial temporal structures. MRI detects the disease more accurately and earlier and is the investigation of choice. When the disease is advanced or only partially treated it can present with a quite marked mass effect and irregular enhancement simulating an aggressive tumour.

Hydrocephalus

An understanding of hydrocephalus and its two main types is important to knowing whether it is 'safe to carry out a lumbar puncture' in a patient or not.

Obstructive or non-communicating hydrocephalus

This is the term given to enlargement of the ventricles caused by an obstruction, usually a mass lesion in the cerebrospinal fluid pathways within the brain (that is, between where the CSF is produced from the choroid plexus in the lateral ventricles and the outflow from the fourth ventricle). It is usually caused by a tumour pressing on the CSF pathways.

Communicating hydrocephalus

If cerebrospinal fluid escapes from the fourth ventricle but there is disturbance of flow around the basal cisterns over the cortex, or there is a failure of absorption of CSF, this is termed 'communicating hydrocephalus'. Communicating hydrocephalus occurs most commonly with a subarachnoid haemorrhage or meningitis and usually does not require treatment. However, because cerebrospinal fluid escapes from the fourth ventricle and circulates round the spinal CSF spaces, it means that it is safe to perform a lumbar puncture to measure and, if appropriate, lower CSF pressure. Lumbar puncture or drainage may also relieve intracranial pressure after a subarachnoid haemorrhage.

Congenital anomalies and paediatric imaging

Any detailed discussion of this subject is beyond the scope of this chapter, and the reader is directed to a specialist text such as Scott, Atlas.

Where available, MRI is the investigation of choice in infants and children presenting with suspected congenital anomalies of the brain. It provides the most information and avoids exposure of young patients to ionizing radiation. The main drawback in this age group is the need for sedation or anaesthesia. The most common indication for imaging in such patients is developmental delay or seizure disorders. It also plays a vital role in the imaging of a suspected, neonatal, hypoxic ischaemic insult and in elucidating the cause of cerebral palsy. CT is a reasonable alternative, but cannot be relied to detect all relevant pathology, particularly in hypoxic ischaemic injury.

A wide variety of congenital anomalies are possible. These range from minor abnormalities of neuronal migration, or localized areas of dysplastic cortex, to major anomalies of the whole brain and encephaloceles where there is an associated defect of the skull or spine such as a spina bifida. The most frequent is the Chiari malformation of the posterior fossa associated with cerebellar ectopia, the cerebellar tonsils lie below the foramen magnum.

Summary

Modern imaging techniques have revolutionized the diagnosis of neurological disease in the last 20 years. The techniques are likely to become more sophisticated and accurate with further extension into functional imaging techniques, both with magnetic resonance and nuclear medicine single-photon emission tomography (SPECT) and positron emission tomography (PET).

The contribution of these techniques to the efficient and effective diagnosis of intracranial and spinal pathology, together with the ability to effectively exclude structural disease, has had a huge impact on clinical practice. The development of interventional neuroradiological techniques for the treatment of vascular disease in the brain, particularly the endovascular coil treatment for intracranial aneurysms, may soon represent a further revolution in the management of patients with subarachnoid haemorrhages and intracranial aneurysms.

Further reading

Scott, Atlas (1996). *Magnetic resonance imaging of the brain and spine*. Lippincott-Williams and Wilkins, Philadelphia.

24.6.1 Inherited disorders

P. K. Thomas

There are many genetically determined neurological disorders, and other multifactorial disorders in which a genetic component can be detected. Inherited movement disorders, disorders of the peripheral nerves and muscles, and those aminoacidurias that are associated with neurological involvement are considered elsewhere, as is the question of genetic factors in the aetiology of conditions such as developmental abnormalities of the nervous system, epilepsy, migraine, Alzheimer's disease, and multiple sclerosis.

Hereditary ataxias

The classification of the hereditary ataxias remains a matter of controversy. A spinocerebellar degeneration may develop in disorders with a known metabolic basis. This category includes abetalipoproteinaemia (see Section 11), ataxia telangiectasia, and xeroderma pigmentosum (see Section 23). In general, the inherited cerebellar and spinocerebellar degenerations can be divided into examples having an early onset (under the age of 25 years), which are usually of autosomal recessive inheritance and of which Friedreich's ataxia is the commonest example, and the later onset cases of cerebellar degeneration that are most often dominantly inherited.

Early onset hereditary ataxias

Friedreich's ataxia

This disorder is an example of a spinocerebellar degeneration and is dominated by progressive ataxia with an onset in childhood or adolescence. The condition is inherited as an autosomal recessive trait and affects males and females approximately equally. The gene responsible has been localized to chromosome 9q13–q21 and is due to a trinucleotide repeat in a non-coding region of the gene for frataxin, a mitochondrial protein. Degeneration of the larger dorsal root ganglion cells occurs with consequent loss of the larger myelinated fibres in the peripheral nerves and degeneration in the dorsal columns. Degeneration is also evident in Clarke's column, in the spinocerebellar tracts, and in the corticospinal pathways. There is variable loss of Purkinje cells in the cerebellum.

The average age of onset is 11 to 12 years, but cases of later onset may occur. The initial symptom is almost invariably ataxia of gait, although foot or spinal deformity may antedate this. At first it is noted that the child walks awkwardly with a tendency to stumble and fall readily; in cases of early onset, walking may never have been normal. As the disease progresses, the gait slowly becomes more irregular and clumsy. The patient walks on a broad base and tends to lurch from side to side. Involvement of the upper limbs develops later, at first giving rise to clumsiness of fine movements, subsequently for all movements. A coarse intention tremor becomes obvious. The trunk is also affected, leading to oscillation of the body when standing or sitting unsupported. A regular tremor of the head (titubation) occasionally appears. Nystagmus is present in about one-quarter of the cases. Dysarthria of cerebellar type develops and may become severe enough to make speech almost unintelligible.

Initially weakness is not obtrusive, but this develops as the disease advances, beginning in the legs and later involving the upper limbs. It results from degeneration in the corticospinal pathways and tends to vary in severity between cases. The plantar responses become extensor, but tone is not usually increased because of the accompanying disturbance of the afferent fibres from muscle spindles. There may be mild wasting of the anterior tibial and small hand muscles related to loss of anterior horn cells. Bladder and bowel function is usually unaffected.

Loss of the larger dorsal root ganglion cells leads to impairment of the sense of joint position, vibration, and to some extent of touch–pressure sensibility, initially distally in the limbs. The impairment of proprioception superimposes a sensory element on the cerebellar ataxia. The tendon reflexes are depressed or absent.

Apart from occasional nystagmus, the ocular movements are usually intact. The pupils are unaffected. Optic atrophy is present in about one-third of cases and 10 per cent of cases develop sensorineural deafness with particular difficulty in speech discrimination.

Associated skeletal deformities are common, in particular foot deformities (pes cavus and pes equinovarus) and kyphoscoliosis. Contractures of the knees may develop in the later stages. Electrocardiography demonstrates widespread T-wave inversion and ventricular hypertrophy in nearly 70 per cent of patients. Echocardiography may suggest the presence of hypertrophic obstructive cardiomyopathy, but these findings are not specific and the ECG is a more sensitive investigation for the detection of cardiomyopathy. The ECG changes are present early in the disease and tend not to be associated with symptoms. Cardiac failure occurs late and is usually precipitated by supraventricular arrhythmias.

Although progressive dementia is not a feature of the disease, reduced intelligence is present in some cases. There is an increased incidence of diabetes mellitus in Friedreich's ataxia (10 per cent).

The disease is slowly progressive, the average age of death being in the latter part of the fourth decade. The foot and spinal deformities may require orthopaedic

correction. Ultimately patients become bedridden. Death is usually from an intercurrent infection or cardiac failure.

Later onset hereditary ataxias (ADCA)

Autosomal dominant cerebellar ataxia

This comprises the main group of disorders within the adult onset hereditary ataxias. The age of onset usually ranges from the third to the fifth decades. Clinically, these disorders can be divided into three categories. Autosomal dominant cerebellar ataxia (ADCA) type I consists of a multisystem disorder with varying combinations of cerebellar ataxia, dementia, optic atrophy, disturbances of eye movement, pyramidal and extrapyramidal features, and peripheral neuropathy. Autosomal dominant cerebellar ataxia type II consists of these features plus pigmentary macular degeneration. Autosomal dominant cerebellar ataxia type III is a relatively pure cerebellar degeneration. Molecular genetic studies have identified a number of different mutations underlying autosomal dominant cerebellar ataxia, designated spinocerebellar ataxia, at present running from spinocerebellar ataxia 1 to spinocerebellar ataxia 7. Amongst these, spinocerebellar ataxia 1 to 4 and spinocerebellar ataxia 6 correspond to autosomal dominant cerebellar ataxia type I, spinocerebellar ataxia 5 (Lincoln ataxia) to autosomal dominant cerebellar ataxia type III, and spinocerebellar ataxia 7 to autosomal dominant cerebellar ataxia type II. Spinocerebellar ataxia 3 is Machado–Joseph disease. A number of these disorders (spinocerebellar ataxias (SCA) 1, 2, 3, 6, and 7) have been shown to be the result of unstable CAG (cytosine–adenine–guanine) trinucleotide repeats.

The category of Marie's delayed cerebellar atrophy was introduced to describe cases of hereditary ataxia with a later onset than Friedreich's ataxia in which the symptoms develop during the third or fourth decades of life or later. It is clear that Marie collected together a heterogeneous group of disorders, but his description served to emphasize the broad subdivision of the hereditary ataxias into the early and later onset groups. Cerebellar degeneration may be a feature in mitochondrial encephalomyopathies. Many instances of late onset cerebellar ataxia, particularly those without ophthalmoplegia or optic atrophy, are probably non-genetic.

Familial episodic ataxia

This comprises two disorders—episodic ataxia 1 mapping to chromosome 12p13 and episodic ataxia 2 mapping to chromosome 19p13. Both involve transient attacks of ataxia together in some cases with paroxysmal kinesogenic choreoathetosis which lasts for seconds to minutes in episodic ataxia 1 and minutes to hours in episodic ataxia 2. Chronic ataxia may develop in episodic ataxia 2. Episodic ataxia 1 has been shown to be due to mutations in K^+ channel genes. In both forms the attacks respond to acetazolamide.

Hereditary spastic paraplegia

Hereditary spastic paraplegia can be subdivided into a 'pure' form (Strümpell's disease) and others in which a variety of other features coexist, some of which are due to mitochondrial dysfunction. Strümpell's disease is genetically heterogeneous. It may display either autosomal dominant inheritance with mutations on chromosomes 14q and 15q or autosomal recessive inheritance. It may present during childhood or even with delayed motor development in infancy; in other cases the onset does not occur until adult life. It gives rise to difficulty in walking because of weakness and spasticity in the legs. The tendon reflexes are exaggerated and the plantar responses are extensor. Foot deformity may be present in cases of early onset. Some patients show a mild degree of cerebellar ataxia and sensory impairment in the legs of posterior column type. The disease progresses slowly and may later affect the upper limbs. Precipitancy of micturition or urinary incontinence may occur. Pathologically there is degeneration of the corticospinal pathways in the lateral columns of the spinal cord and some fibre loss in the gracile fasciculi. An X-linked form of hereditary spastic paraplegia also exists, related to mutations in the gene for proteolipid protein. It is allelic with Pelizaeus–Merzbacher disease.

Severe spasticity, if present, may be alleviated to some extent by oral baclofen or dantrolene. Continuous intrathecal administration of baclofen by an infusion pump is helpful in selected cases. The precipitancy of micturition may be improved by oxybutynin. Surgical correction of foot deformities is sometimes required.

Genetically distinct disorders in which a spastic paraplegia is associated with other clinical features include the Sjögren–Larsson syndrome, a recessively inherited condition which combines congenital ichthyosis and oligophrenia with a childhood onset spastic paraplegia, and the dominantly inherited disorder in which the paraplegia is associated with distal amyotrophy in the limbs resembling peroneal muscular atrophy. Hereditary spastic paraplegia can represent a mitochondrial disorder affecting the gene for paraplegin.

Disorders of lipid metabolism (see alsoSection 11)

Neurolipidoses

The lipidoses constitute a group of disorders characterized by the intracellular accumulation of a variety of different lipids. Some predominantly involve the nervous system; others primarily affect the reticuloendothelial system, but may also involve nervous tissue. They may be classified in terms of the particular lipid that is stored.

Niemann–Pick disease

This consists of a group of recessively inherited disorders in which there is an accumulation of lipid in 'foam cells' in the reticuloendothelial system. Types A and B, due to mutations on chromosomes p15.1–15.4, are the result of acid sphingomyelinase deficiency resulting in accumulation of sphingomyelin. Types C and D are related to a defect in the intracellular homeostasis of unesterified cholesterol. In type A, progressive mental deterioration and spastic paralysis in association with hepatosplenomegaly appear in the first 6 months of life, leading to death before the age of 3 years. Cherry-red spots are present at the maculae in 50 per cent of cases. Type B does not affect the nervous system. Types C and D resemble type A but the storage material consists of cholesterol and neutral lipids. Sphingomyelinase activity is normal.

Glucosyl ceramide lipidosis (Gaucher's disease)

Three variants exist, all recessively inherited, characterized by hepatosplenomegaly related to the accumulation of glucosyl ceramide in histiocytes as a consequence of a deficiency of the enzyme glucocerebrosidase. The type I adult onset form does not usually affect the nervous system but types II and III, with an infantile and juvenile onset respectively, and a more rapid progression, display widespread cerebral involvement.

Gangliosidoses

These comprise a group of recessively inherited disorders in which there is a combination of progressive dementia, epilepsy, and visual failure. They are related to defective ganglioside degradation. Several GM_1 gangliosidoses exist and are the result of an inherited deficiency of acid b-galactosidase. In the infantile form, which maps to chromosome 3p14.3, there is a generalized storage of GM_1 ganglioside affecting the brain, the viscera, and the skeleton. The onset is at birth or in early infancy, and initially is manifested by a failure to thrive and by hepatosplenomegaly. Later, mental and motor deterioration become evident, and a cherry-red spot may be present at the macula, related to retinal degeneration. Skeletal abnormalities, including abnormal facial features, have led to the condition being referred to as the 'pseudo Hurler syndrome'. Death takes place before the age of 3 years. A juvenile onset variant also exists.

The GM_2 gangliosidoses involve the storage of GM_2 gangliosides, which are largely confined to the nervous system. In the type 1 variety (Tay–Sachs disease), the disorder usually begins within the first 6 months of life. It is encountered most frequently in Ashkenazi Jews. Initially there is retardation of development which is followed by progressive dementia, hypotonic weakness, and blindness. There is a cherry-red spot at the macula. Later, seizures occur and terminally generalized spasticity develops. Death generally takes place in the fourth year of life. The disorder is related to an inherited deficiency of hexosaminidase A and the gene has been localized to chromosome 15q23–q24. Carrier detection is possible by a serum assay, and mass screening programmes have been undertaken in some countries. Antenatal diagnosis by amniocentesis is also possible. There is no specific therapy. The type 2 form (Sandhoff's disease) is similar clinically but involves a combined deficiency of hexosaminidase A and B. A form with juvenile onset also exists, as do phenotypes presenting as spinal muscular atrophy, spinocerebellar degeneration, or as a dystonic syndrome. The gene is localized on chromosome 5q13.

Neuronal ceroid lipofuscinosis

Under this heading are grouped a number of rare disorders in which retinal degeneration, progressive dementia, epilepsy, spasticity, and ataxia occur in various

combinations. The age of onset may be infantile (Santavuori), late infantile (Jansky–Bielschowsky), juvenile (Spielmeyer–Vogt or Batten), or adult (Kufs). There is neuronal storage of lipopigment, but the molecular basis for these disorders has not been established. The infantile, late infantile, and juvenile forms are all of autosomal recessive inheritance. The infantile, late infantile, and juvenile forms map to chromosome 1p, 11p, and 16p respectively.

Leucodystrophies

These disorders are characterized by a diffuse disintegration of white matter in the central nervous system and sometimes also by segmental demyelination in the peripheral nerves. The cell bodies of the neurones are generally spared, although both myelin sheaths and axons show destruction in the white matter lesions.

Metachromatic leucodystrophy (sulphatide lipidosis)

The most common variant is the late infantile type which usually begins in the third year of life with weakness and ataxia in the limbs. Subsequently a progressive dementia supervenes, seizures may occur, and in some instances optic atrophy develops. The tendon reflexes may be depressed or absent in those patients in whom peripheral nerve involvement is prominent. Nerve conduction velocity is reduced. Death sometimes occurs after a course of a few months, but occasionally after as long as 5 or 6 years. Terminally the affected children are demented, with a spastic tetraplegia, and are often blind.

The term metachromatic leucodystrophy is derived from the presence of galactosyl sulphatide in the affected tissues. This stains metachromatically with dyes such as cresyl violet and toluidine blue. It may be demonstrated within cells in fresh specimens of urine and also within Schwann cells and macrophages in biopsies of the peripheral nerves or rectal wall. The disorder, which maps to chromosome 22q13–13qter, is inherited in an autosomal recessive manner, and is due to a deficiency of aryl sulphatase A. This can be demonstrated by assay on leucocytes.

Juvenile and adult onset forms of metachromatic leucodystrophy are also encountered, but are rare. Prenatal diagnosis by amniocentesis and assay of aryl sulphatase activity on cultured amniotic fibroblasts is possible in both the late infantile and juvenile forms. Variants related to multiple sulphatase deficiency and to deficiency of activator protein also occur.

Globoid cell leucodystrophy (Krabbe's disease)

This derives its title from the presence of large multinucleate cells containing galactosylceramide in areas of white matter damage. The condition begins at the age of 3 or 4 months as a failure to thrive. Developmental regression then becomes evident and the tendon reflexes are lost. As the disease advances, generalized hypertonus appears, together with various types of seizure, and optic atrophy. Death often occurs in the first year of life or it may be delayed into the second year. There are also rare late onset cases.

The peripheral nerves are affected, biopsies showing segmental demyelination and inclusions within Schwann cells. Nerve conduction velocity is reduced. This can be helpful diagnostically in suspected cases

The disorder is inherited in an autosomal recessive manner and is due to a deficiency of galactosylceramide b-galactosidase. It maps to chromosome 14q21–q31. This may be demonstrated by assays on leucocytes or serum.

Adrenoleucodystrophy and adrenomyeloneuropathy (see Chapter 24.10)

These are a group of conditions that give rise to widespread demyelination in the brain with an onset during childhood, and cases of adrenoleucodystrophy can be separated by virtue of X-linked inheritance and associated adrenal insufficiency with features resembling those of Addison's disease. The disorder maps to chromosome Xq28. The affected boys exhibit a progressive illness characterized by the development of dementia, cortical blindness, ataxia, and spastic weakness in the limbs. A myeloneuropathy is sometimes the presenting deficit, or other phenotypes. Manifesting female carriers may show a mild spastic paraparesis or adrenal insufficiency.

Other rare demyelinating conditions

Pelizaeus–Merzbacher disease is an X-linked recessive disease that appears in early infancy. It maps to chromosome Xq22. Affected males develop ataxia and spasticity. It is due to the defective production of proteolipid protein in central myelin. Canavan's disease also develops in early infancy, is of autosomal recessive inheritance, and gives rise to progressive mental deterioration and megalencephaly associated with spongy degeneration of the white matter. It is related to a deficiency of aspartoacylase. Affected children show excessive urinary excretion of *N*-acetylaspartic acid. Prenatal diagnosis is possible.

Fabry's disease (a-galactosidase A deficiency) (see also Section 11)

This condition, otherwise known as angiokeratoma corporis diffusum, is an inborn error of glycosphingolipid metabolism. Neutral glycosphingolipids are deposited in various tissues as a consequence of a deficiency of the enzyme a-galactosidase A. The disorder is inherited in an X-linked recessive manner and maps to chromosome Xq21.33–q22. Affected hemizygous males develop a mild peripheral neuropathy which is manifested by the occurrence of severe pains in the extremities, often beginning in childhood. Cerebrovascular lesions also occur, either cerebral infarction or haemorrhage. Non-neurological features include corneal opacification, punctate angiectatic lesions over the lower trunk, buttocks, and upper legs, and cardiac and renal lesions. Heterozygous females may display mild manifestations, most usually corneal opacification.

Hereditary lipoprotein deficiency (see also Section 11)

The occurrence of peripheral neuropathy in hereditary high-density lipoprotein deficiency (Tangier disease) is discussed in Section 11. Hereditary abetalipoproteinaemia (Bassen–Kornzweig disease) is a recessively inherited disorder mapping to chromosome 2p24, in which a spinocerebellar degeneration may develop with features that bear some resemblance to Friedreich's ataxia. Other manifestations of this uncommon disorder include intestinal malabsorption, pigmentary retinopathy, and the presence of acanthocytes in the peripheral blood (see Section 22). In addition to the absence of serum low-density lipoproteins, the serum cholesterol level is substantially reduced. There is evidence that the development of the neurological lesions may be prevented by the administration of vitamin E orally, the absorption of which from the gut is impaired. A spinocerebellar degeneration has also been described in individuals homozygous for hereditary hypobetalipoproteinaemia, a genetically separate condition.

Isolated vitamin E deficiency

A spinocerebellar syndrome resembling Friedreich's ataxia has been described in recent years due to vitamin E deficiency in the absence of generalized fat malabsorption. Titubatory head tremor is a particular feature. The disorder is of autosomal recessive inheritance, maps to chromosome 8q13.1–13.3, and is due to a deficiency of an a-tocopherol carrier protein. Treatment is by vitamin E replacement.

Neurocutaneous syndromes

This category encompasses a number of disorders in which a variety of neurological disturbances are associated with cutaneous abnormalities.

Neurofibromatosis

Two major forms of this disorder exist. Both are of autosomal dominant inheritance. The gene for neurofibromatosis type 1 (von Recklinghausen's disease) is on the proximal long arm of chromosome 17 at q11.2. The gene product neurofibromin is a member of the GTPase activating family of proteins. The disorder has a wide range of manifestations, the most constant of which are focal areas of hyperpigmentation (*café au lait* spots), multiple neurofibromas, and Lisch nodules on the iris. Six or more *café au lait* spots are necessary for them to be considered abnormal. Axillary and inguinal freckling is frequent. The cutaneous fibromas are of varying dimensions and can be extremely numerous. At times, giant plexiform neuromas develop in which there is extensive subcutaneous overgrowth of neurofibromatous tissue. Massive mediastinal, pelvic, or retroabdominal plexiform neurofibromas can occur, as well as cervical paraspinal tumours and astrocytomas of the optic nerve,

cerebellum, or brainstem. Malignant change occurs in a small proportion of peripheral neurofibromas. Mental retardation due to diffuse cortical dysgenesis is encountered in at least 10 per cent of patients. Other manifestations include congenital glaucoma, phaeochromocytoma, spinal deformity, pathological fractures of limb bones with malunion and pseudoarthrosis, and local gigantism of a limb. The neurofibromas are composed of proliferated Schwann cells and fibroblasts in a collagenous matrix through which course nerve fibres in an irregular manner.

Most cases of neurofibromatosis type 1 require no treatment. Neurofibromas causing pressure symptoms may necessitate excision and others may merit removal for cosmetic reasons. Rapid expansion of a tumour, the development of pain, and loss of neural function, will suggest malignant change. This most often occurs during adolescence or in young adults. The development of hypertension will require investigation for phaeochromocytoma, and spinal deformity may need orthopaedic attention.

The gene for neurofibromatosis type 2 (central neurofibromatosis) has been mapped to chromosome 22q12.2. The gene product merlin has homology with proteins at the plasma membrane/cytoskeleton interface. This disorder is characterized in particular by bilateral acoustic neurinomas (Fig. 1) but tumours may occur on other cranial nerves or spinal roots and also paraspinally. Meningiomas and gliomas may develop. A further feature is the occurrence of juvenile posterior subcapsular lenticular opacities.

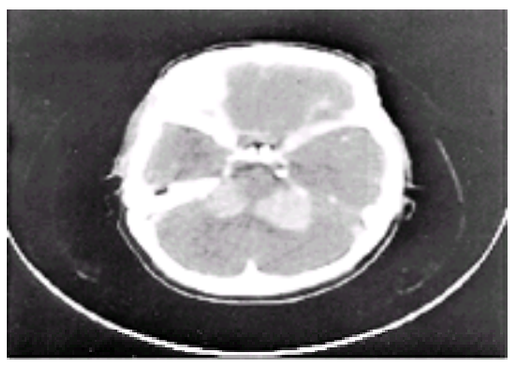

Fig. 1 Computed tomography scan showing bilateral acoustic neurinomas in a patient with neurofibromatosis type 2. She also had an astrocytoma of the thoracic spinal cord.

Segmental neurofibromatosis affecting a restricted area of the body may be due to somatic mutation.

Tuberous sclerosis (Bourneville's disease, epiloia)

The features of this condition are mental retardation, infantile spasms, epilepsy, and the occurrence of retinal hamartomata and characteristic skin lesions. The disorder is dominantly inherited, but may be transmitted by individuals who are asymptomatic and who show only minimal clinical evidence of the disease. Isolated cases are frequent, comprising as many as 80 or 90 per cent of index cases. Many of them probably represent new mutations: others are transmitted by gene carriers with trivial manifestations. Genetic heterogeneity has now been established, with separate loci on chromosomes 9q34 (*TSC1*) and 16p13.3 (*TSC2*). The gene *TSC2*, which has a more severe phenotype, has been identified and its product, called tuberin, has the structure of a GTPase activating protein.

The earliest cutaneous lesions are irregular foliate areas of depigmentation over the trunk. These patches are readily identified when viewed under ultraviolet illumination using a Woods lamp. Facial angiofibromas ('adenoma sebaceum') are a second type of skin lesion which develop over the cheeks in a 'butterfly' distribution and on the forehead (Fig. 2) with multiple small warty elevations. Finally, a 'shagreen patch' may be present over the lower back. This consists of an area of elevated roughened skin with a yellowish tinge which has been likened to shark skin.

Fig. 2 Adenoma sebaceum in a patient with tuberous sclerosis.

The cerebral changes give rise to mental retardation which is evident in early life and which may be static or involve a slowly progressive cognitive decline, often complicated by behavioural disorder. Infantile spasms or epilepsy with recurrent generalized or focal seizures may occur in association with mental retardation or in individuals of normal intelligence. The cerebral lesions, which are demonstrable by computed tomography or magnetic resonance imaging, are typified by nodular or tuberous masses composed of proliferated glial cells and enlarged distorted neurones. They may become calcified. They are found scattered throughout the cerebral cortex and also extend into the ventricles to produce an appearance that was considered to resemble 'candle guttering' when seen in pneumoencephalograms. Gliomas sometimes arise in these lesions.

Retinal tumours, termed phakomas, may be present, and cardiac rhabdomyomas occasionally arise as well as hamartomas of the lungs and kidneys. Polycystic disease of the kidneys may also be associated.

Treatment consists of control of the epilepsy and the management of the mental retardation and behavioural disorder. Many of the more severe cases require institutionalization.

Cerebelloretinal haemangioblastosis (von Hippel–Lindau disease)

This condition comprises the occurrence of vascular tumours in the retina and within the central nervous system, most commonly in the cerebellum and spinal cord. The inheritance is autosomal dominant in pattern. The disorder maps to chromosome 3p25–26.

The retinal lesions consist of angiomatous vascular malformations. The cerebellar lesion is a haemangioblastoma, often cystic, which may slowly expand and present with features of a cerebellar tumour. It may require surgical treatment. Such tumours may be associated with polycythaemia, related to the production of erythropoietin or a similar substance by the tumour. Haemangioblastomas may occur in the spinal cord and rarely in the cerebral hemispheres, as may renal cell tumours and renal and pancreatic cysts. Regular screening for renal tumours in patients and relatives at risk is an important aspect of management.

Ataxia telangiectasia (Louis–Bar syndrome)

The inclusion of this disorder with the neurocutaneous syndromes depends upon the coincidence of a progressive cerebellar degeneration with cutaneous vascular lesions. The inheritance is of autosomal recessive type and the gene is localized on chromosome 11q22.3–q23.1. Ataxia begins in early childhood and choreoathetosis and oculomotor apraxia appear later. Telangiectasia of the conjunctivae is present as a relatively early feature and later becomes evident in the pinnae, over the face, and in the limb flexures. Some patients show an immunoglobulin deficiency and recurrent infections, or the development of malignancies may complicate the clinical picture. Defective DNA repair after irradiation with X-rays is demonstrable in cultured skin fibroblasts. Affected children usually become unable to walk by the age of 12 years and death occurs during the second or sometimes the third decade of life.

Hereditary myoclonic epilepsies (see alsoChapter 24.10)

A number of conditions exist in which generalized epileptic seizures and myoclonus are associated with a progressive degenerative neurological disorder occurring on a genetic basis.

Lafora body disease

This is the most clearly defined form of progressive myoclonic epilepsy. It consists of a combination of major seizures, myoclonus, and progressive dementia with an onset in late childhood or early adolescence, with death usually occurring before adult life is reached. Cerebellar signs may appear later in the illness. The condition is characterized by the presence of intracellular inclusion bodies found most consistently in neurones of the cerebral cortex and in the cerebellar dentate nuclei. They are also detectable in the liver, axillary sweat gland, and in skeletal muscles, all of which are convenient sites for biopsy in order to establish the diagnosis. These Lafora bodies are composed of a polyglucosan. The disorder is caused by an autosomal recessive gene, mapping to 6q23–25. Treatment is directed towards control of the epilepsy.

Progressive myoclonic ataxia (Ramsay Hunt syndrome)

This is a heterogeneous group, the best characterized form being Unverricht–Lundborg disease or 'Baltic myoclonus'. Stimulus-sensitive myoclonus develops from the age of 6 to 15 years, associated with mild mental retardation, followed later by dysarthria and ataxia. The disorder is of autosomal dominant inheritance. The gene maps to chromosome 21q22.3 and encodes for cystatin B, a cysteine protease inhibitor.

Mitochondrial encephalomyopathy

Myoclonus and ataxia, along with other features, may occur in mitochondrial disorders.

Sialidosis

The cherry-red spot–myoclonus syndrome consists of two autosomal recessive disorders (types I and II) with an onset in late childhood, adolescence, or early adult life that combine action myoclonus with mental retardation, ataxia, cherry-red spots at the maculae, and cataracts. Both types I and II are related to a deficiency of sialidase. Type II is associated with dysmorphic features.

The myoclonus in these inherited disorders may respond to combination treatment with clonazepam and piracetam.

Hereditary spinal muscular atrophies

The hereditary spinal muscular atrophies constitute a group of disorders that involve a selective degeneration of anterior horn cells and sometimes also of the motor nuclei of the lower cranial nerves. They can be classified in terms of the pattern of involvement and the age of onset. Only the more common varieties will be described.

Hereditary proximal spinal muscular atrophy

Acute infantile spinal muscular atrophy (Werdnig–Hoffmann disease) almost always begins within the first year of life. It may have a prenatal onset and is one cause of the 'rag doll' child syndrome of hypotonic muscle weakness in infancy. Progressive proximal muscular weakness and wasting, later becoming generalized and involving the bulbar musculature, usually leads to death before the age of 4 years. Cases with prolonged survival also occur (chronic childhood form). In hereditary juvenile proximal spinal muscular atrophy (Kugelberg–Welander disease) the onset is during childhood after the age of 2 years or as late as adolescence. The involvement of the proximal limb and trunk musculature mimics limb girdle muscular dystrophy, but fasciculation may be observed. The course is relatively benign, but progressive disability in adult life occurs in some cases; others remain relatively mildly affected. All three forms are of autosomal recessive inheritance and have been mapped to chromosome 5q12.2–q13.3. They appear to be due to allelic genes.

X-linked bulbospinal neuronopathy

This disorder, otherwise known as Kennedy's syndrome, consists of the development, commonly in the third or fourth decades, of progressive limb weakness and, later, a bulbar palsy. Contraction fasciculation of the facial muscles is usually present. Muscle cramps and upper limb postural tremor are often evident from early adult life. About 50 per cent of cases show gynaecomastia and some have diabetes mellitus. The disorder is due to a trinucleotide repeat expansion within the androgen receptor gene on the proximal long arm of the X chromosome at Xq21.3–q22.

Other miscellaneous disorders

A wide variety of other rare inherited conditions exist, of which the following deserve brief mention.

Hereditary optic neuropathy

Dominantly inherited juvenile optic neuropathy

This disorder gives rise to the insidious bilateral onset of optic atrophy during childhood with either mild or severe loss of vision. A central or centrocaecal scotoma may be detected. Electroretinography and visual evoked potentials demonstrate no loss of retinal receptors and suggest that the lesion affects retinal neuronal elements. There are no associated neurological abnormalities.

Leber's hereditary optic neuropathy

This disorder typically gives rise to acute or subacute bilateral visual loss in males between the ages of 18 and 30 years, although earlier and later ages of onset may be encountered. It may remain monocular for months or years. Initially there is enlargement of the blind spot and later this increases to involve central vision, producing a large centrocaecal scotoma. In the acute phase there is swelling in the nerve fibre layer around the optic disc with tortuous retinal arterioles and peripapillary telangiectasias. Later the disc becomes atrophic. In affected females the age of onset tends to be later and a multiple sclerosis-like syndrome may develop (Harding's syndrome). The disease is only transmitted by females and has recently been shown to be due to mutations of mitochondrial DNA.

Mucopolysaccharidoses (see also Section 11)

The mucopolysaccharidoses constitute a group of disorders related to deficiencies of specific lysosomal enzymes, involving an accumulation in various tissues of acid mucopolysaccharides and gangliosides, and the presence of mucopolysaccharides in the urine. In both the recessively inherited Hurler's syndrome and the X-linked recessive Hunter's syndrome, the skeletal and other manifestations may be accompanied by mental retardation and pigmentary retinal degeneration. Spastic weakness in the limbs may develop in Hurler's syndrome. Mental retardation is also seen in the recessively inherited Sanfilippo's syndrome. Entrapment neuropathies

are a feature in some forms, related to the skeletal changes.

Subacute necrotizing encephalopathy (Leigh's syndrome)

Typically this label is applicable to a fatal encephalopathy that develops during the first 2 years of life with variable combinations of mental retardation, seizures, optic atrophy, cerebellar ataxia, and central respiratory failure associated with lactic acidosis. Pathologically there are necrotic lesions in the brainstem and a prominence of small blood vessels. Haemorrhage does not occur. The distribution of the lesions bears some resemblance to Wernicke's encephalopathy. Later onset cases with similar pathological changes occur. The disorder is genetically heterogeneous **,** but may be related to mutations in mitochondrial DNA leading to a deficiency of cytochrome oxidase. Pyruvate dehydrogenase and pyruvate decarboxylase deficiency may also be responsible.

Progressive neuronal degeneration of childhood with liver disease (Alpers–Huttenlocher syndrome)

This disorder, which is probably of autosomal recessive inheritance, begins at 3 to 15 months of age with developmental delay and failure to thrive. Recurrent vomiting and hypotonia are common, followed by intractable seizures. Death occurs at 10 to 90 months. Pathologically there is extensive neuronal loss and astrocytosis, particularly in the cerebral cortex, and fatty degeneration, cell loss, and fibrosis in the liver. An unidentified biochemical defect is assumed.

Infantile neuroaxonal dystrophy (Seitelberger's disease)

This is a rare, probably recessive condition that develops between the ages of 1 and 3 years and gives rise to progressive motor weakness from both upper and lower motor neurone deficits. It maps to chromosome 22q13–qter. Optic atrophy also occurs and death usually ensues before the end of the first decade. Degenerative changes are present in the brain and the spinal cord, the most striking feature of which is the occurrence of large axonal swellings in the grey matter of the brain and spinal cord.

Menkes' syndrome ('kinky' or 'steely' hair)

Menkes' syndrome is an X-linked recessive disorder in which developmental regression begins within a few months of birth. The gene locus is at Xq12–q13. Muscle hypotonus or hypertonus and seizures appear and are associated with abnormal hair. The scalp hair is sparse, stubbly, and greyish in colour. When examined under magnification, the hairs are seen to be twisted and display partial breaks. The serum copper and caeruloplasmin levels are low and the condition probably results from defective absorption of copper from the gut. Prenatal diagnosis is possible.

Lesch–Nyhan syndrome

This is an X-linked recessive disorder (see also Chapter 24.10 and Section 11) related to the absence of an enzyme of purine metabolism, hypoxanthine–guanine phosphoribosyl transferase. It maps to chromosome Xq25–q27.2. The salient features are overproduction of uric acid and consequent hyperuricaemia which are associated with various behavioural and neurological manifestations, including mental retardation, self-mutilation, choreoathetosis, pyramidal signs, and spasticity in the limbs. The neurological abnormalities develop in childhood and death usually occurs in the second or third decade from renal failure. Allopurinol may reduce some of the non-neurological consequences of the hyperuricaemia, but no treatment influences the neurological abnormalities, the mechanism of which is not understood. Prenatal diagnosis and carrier detection are available.

Cockayne's syndrome

This is of autosomal recessive inheritance and consists of the development in childhood of dwarfism, microcephaly, progeria, mental retardation, cataract, pigmentary retinopathy, and ataxia. There is cutaneous sensitivity to ultraviolet light. The cerebral changes are those of a leucodystrophy. A demyelinating neuropathy coexists.

Further reading

Baraitser M (1997). *The genetics of neurological disorders*, 3rd edn. Oxford University Press, Oxford.

Emery AEH and Rimoin DL (1997). *Principles and practice of medical genetics*, 3rd edn. Churchill Livingstone, Edinburgh.

Harding AE (1984). *The hereditary ataxias and related disorders*. Churchill Livingstone, Edinburgh.

Rosenberg RN *et al.*, eds (1997). *The molecular and genetic basis of neurological disease.* Butterworth-Heinemann, Oxford.

24.6.2 Neurogenetics

Nicholas Wood

Introduction

There has been dramatic improvement in our understanding of inherited neurological disease. It is clearly impossible to be comprehensive here, and a few examples have been chosen to illustrate some of these major developments. The genetics of movement disorders, especially Huntington's disease, are discussed in the context of identified genes causing neurological disease. However, the greatest challenge of the next few years will be to identify the genetic factors and environmental interactions involved in complex disorders such as Parkinson's disease.

Table 1 shows a much wider range of disorders. The rapid progress in this area, however, means that it cannot be comprehensive or up to date and current research publications should be consulted for the latest information.

Huntington's disease

This is one of the most common hereditary movement disorders with a prevalence of approximately 1 in 20 000. Onset is usually between the fourth and sixth decades, but onset in childhood and old age can also be seen. The initial mental and cognitive signs may be very subtle and progress insidiously. Family members often report a change in personality, a coarsening of sensitivities, and the expression of new, often antisocial, behaviours. Usually in parallel, but often not even noticed by the relatives, is the onset of choreic movements. These may start as a slight fidgetiness before semi-purposeful jerky movements develop.

As the disease progresses, dementia becomes more pronounced and the chorea more extreme; this may eventually upset balance and resemble an ataxic disorder. Finally, immobility occurs. The mean time from onset to death is approximately 15 years.

Huntington's disease may present with a parkinsonian syndrome and occasionally epilepsy, especially if the onset is early (under 20 years). Later onset may produce a predominant choreic illness with little cognitive disturbance. Before the identification of the gene it had been noted that age of onset and decreased severity increased in successive generations. This 'anticipation' was initially ascribed to bias but there is now a molecular explanation, which is discussed below.

There is no disease-modifying treatment currently available, but controlled trials are now being undertaken. These are largely based on findings from transgenic animals. Choreic movements may be controlled by sulpiride or tetrabenazine, but often patients prefer chorea to the parkinsonism that can result from this medication. Psychiatric disturbances should be treated as appropriate.

Genetics

The disease is autosomal dominant and therefore offspring are at a 50 per cent risk of inheriting the mutant allele. It is highly penetrant and very seldom are gene carriers ultimately unaffected. The gene was cloned in 1993 and was shown to be due to an expanded CAG repeat in exon 1 of a novel gene, subsequently called Huntingtin. The role of the protein is still unknown but progress has been made in evaluating the role of this abnormal CAG triplet repeat. This is in part because the same mutation is found in a group of neurodegenerative diseases (see Table 2). These diseases not only differ in their clinical features but the genes and proteins have very little in common other than this abnormal repeat. The codon CAG encodes glutamine and the term polyglutamine disorders has been coined. All the CAG repeat disorders so far described are the result of a relatively modest expansion within the coding region of the affected gene. Although the exact number of repeats on both the normal and the abnormal allele varies between the different diseases, the normal range of repeats is in the 20s, whereas for the disease-carrying allele, it tends to be over 40. All of these diseases are predominantly adult-onset neurodegenerative disorders, and most show evidence of anticipation. This is particularly seen with paternal transmission. The exact function of the polyglutamine tract remains unknown, but the repeat length is a major determinant of age of onset and probably also partly determines severity. Recent transgenic animal and cell culture studies strongly support the hypothesis of a gain of function for the allele harbouring the CAG expansion. It is unknown how the neuronal specificity of this disorder is brought about as both wild-type and mutant proteins are widely expressed. Moreover, the common link between an expanded polyglutamine tract and toxicity is unknown. There are at least 14 other proteins that interact with Huntingtin. An insight into the toxic pathway has emerged recently since mice transgenic for exon 1 of the human Huntingtin gene carrying over 100 CAG repeats developed abundant intranuclear inclusions. These pathological changes predated the neurological phenotype. Similar findings are reported in several of the other CAG repeat disorders. Identification of a common downstream pathway which could be manipulated by inhibitory drugs might offer the hope of improving the outcome from these disorders.

Dentatorubropallidoluysian atrophy

This is a dominant disorder reported mainly in Japanese families but is found worldwide. It has a variable phenotype including various combinations of ataxia, dystonia, myoclonus, other types of seizure, dementia, and parkinsonism. It may also closely resemble Huntington's disease. Onset ranges from late childhood to late adult life. The pathological features are incorporated in the name of the disease. This disorder was mapped to chromosome 12p in Japan, and the disease mutation is another expanded CAG repeat. The same mutation has now been described in other populations, but the disease is much less frequent than Huntington's disease outside Japan.

Parkinson's disease

There are several akinetic rigid syndromes with a clearly defined genetic basis (see Table 3). Many are of childhood onset and have additional neurological features. Nevertheless, as the table illustrates, there is a growing list of adult-onset disorders in which parkinsonism may be the presenting feature, in many cases the diagnosis can be confirmed by DNA testing techniques. However, these disorders account for only a small minority of cases of adult-onset parkinsonism, and in the remainder the role of genetic factors is less clear-cut.

Parkinson's disease is a common neurodegenerative disorder among the elderly. In Europe the overall age-standardized prevalence for Parkinson's disease in subjects of 55 years or older is 1.6 per 100, with an increasing frequency up to 4.3 per cent in those aged 85 years and over. Despite intensive efforts, the cause of Parkinson's disease remains largely unknown and treatment is symptomatic with only temporary results. However, there is increasing evidence that genetic factors play an important role in the aetiology of Parkinson's disease. Familial clustering of Parkinson's disease has been reported in many studies. Classic linkage and positional cloning strategies have identified a number of genes and loci responsible for mendelian Parkinson's disease (see Table 3). The a-synuclein gene was the first identified gene in autosomal dominant Parkinson's disease. Since then, a mutation in UCH-L1 and two loci 2p13 and 4p14–16.3 have been implicated in

autosomal dominantly inherited Parkinson's disease.

Juvenile Parkinson's disease

This condition differs from Parkinson's disease not only in the age of onset, but also the classic triad of signs (bradykinesia, rigidity, and tremor) is relatively mild, whereas there is more prominent dystonia, postural instability, and hyperreflexia. Additional features include: (i) early onset, typically before the age of 40; (ii) dystonia at onset; (iii) diurnal fluctuations; (iv) slow disease progression; and (v) early and severe levodopa-induced dyskinesias. A gene for autosomal recessive juvenile Parkinson's disease has recently been linked to chromosome 6q25.2–27 and designated *PARK2*. Pathological examination has shown a massive loss of dopaminergic neurones in the pars compacta of the substantia nigra, in the absence of Lewy bodies, the histopathological hallmark of classic Parkinson's disease. A novel gene designated parkin, in which homozygous exon deletions were detected in four Japanese families with autosomal recessive juvenile Parkinson's disease, has been described. The parkin protein is composed of 465 amino acids, shows moderate homology to ubiquitin at the amino terminus, and contains a RING-finger motif at the carboxy terminus, and it has been shown that it has ubiquitin ligase activity. Subsequently, it has been shown that among the patients with isolated Parkinson's disease, mutations were detected in 77 per cent with onset at 20 years or earlier, but they were much more rare in later onset disease accounting for 3 per cent with onset at over 30 years. Multiple mutations have now been described in this gene and it is numerically the most important locus hitherto described for Parkinson's disease.

Very recently two other autosomal recessive loci have been identified (PARK6 and 7).

Other parkinsonian syndromes

There are a number of syndromes incorporating both parkinsonism and additional features. These are summarized in Table 3. Some of the more recent developments are discussed in more detail here.

Chromosome 17-linked tauopathies

There are a number of diseases consisting of parkinsonism complicated by a variety of features, especially cognitive impairment, see Table 3. Recently the genetic mechanisms underlying chromosome 17-linked parkinsonism dementia have been defined. Mutations of the *tau* exon 10 splice site and coding mutations in *tau* have been described which are predicted to lead to an increase in the transcription of four repeat tau isoforms or a disruption of microtubule binding, respectively. This is consistent with the predominant deposition of four repeat tau isoforms in these types of neurodegeneration.

Other tauopathies

Progressive supranuclear palsy and corticobasal degeneration

Progressive supranuclear palsy is a neurodegenerative condition which affects the brainstem and basal ganglia. It is frequently misdiagnosed, most commonly as Parkinson's disease. Patients present with disturbance of balance, a disorder of vertical gaze, and parkinsonism not resposive to levodopa. They usually develop progressive dysphagia and dysarthria leading to death from the complications of immobility and aspiration. Its prevalence is estimated at 1.4 per 100 000 with a median survival of 9 years, but a pathological study showed that 20 per cent of clinically diagnosed cases of Parkinson's disease had alternative pathological diagnoses and progressive supranuclear palsy accounted for about 6 per cent. Treatment for progressive supranuclear palsy remains largely supportive.

Progressive supranuclear palsy is also a tauopathy characterized by deposition of the four-repeat isoform of tau. Analysis of sporadic cases of progressive supranuclear palsy has demonstrated that one allele (A0) of an intronic polymorphism within the *tau* gene occurs more frequently in patients with progressive supranuclear palsy than controls. It appears that this allele increases the risk of developing progressive supranuclear palsy, but is in itself neither necessary nor sufficient to cause the disease. Interestingly this allele (Ao) and its associate haplotype (H1) is also known to be associated with increased risk of developing corticobasal degeneration, another tauopathy. The link between a genetic predisposition to progressive supranuclear palsy and the differential isoform expression of the *tau* gene may be the key to explaining the pathogenesis of these conditions.

Dystonias

Primary torsion dystonia is a clinically and genetically heterogeneous movement disorder. It is characterized by involuntary muscle spasms causing twisting and repetitive movements and postures, and is distinguished from secondary dystonia by the absence of causative exogenous factors (such as drugs, trauma) or other neurological disorders.

Early-onset dystonia (before 28 years) is the most severe and common form of hereditary primary torsion dystonia. It usually begins in a limb and spreads to other limbs within a few years, usually sparing craniocervical muscles. Most cases of early-onset dystonia are caused by an autosomal dominant gene with reduced penetrance (*DYT1*), on human chromosome 9q34. The *DYT1*-associated phenotype is similar in all ethnic communities, with highest prevalence in the Ashkenazi Jewish population. Recently, a 3-base pair (GAG) deletion in the coding sequence of the *DYT1* gene was found in all affected individuals and obligate gene carriers with 9q34-linked primary torsion dystonia, both in the Jewish and non-Jewish populations. About two-thirds of patients with early-onset dystonia carry the *DYT1* 3-base pair deletion. *De novo* GAG deletions in the *DYT1* gene can also occur rarely.

Dopa-responsive dystonia

In classic cases, the disease manifests in early childhood with walking difficulties due to dystonia of the lower limbs. Some 'parkinsonian' features such as reduced facial expression or slowing of fine finger movements frequently accompany the dystonia. Although rare, patients respond very well to small doses of levodopa without the later motor fluctuations seen in Parkinson's disease, suggesting that dopamine biosynthesis may be disturbed. This has implicated the biochemical pathway producing dopamine. Two different genes have been implicated in this disorder, both of which disrupt dopamine production. Most commonly, heterozygote mutations of the GTP cyclohydrolase I gene are found. This is the rate-limiting enzyme in the synthesis of tetrahydrobiopterin. Tetrahydrobiopterin is an essential cofactor for tyrosine hydroxylase, the rate-limiting enzyme in the synthesis of dopamine. Reduced levels of tetrahydrobiopterin lead to the dopamine-deficit syndrome, dopa-responsive dystonia, because of reduced tyrosine hydroxylase activity. The second, much less common, genetic abnormality is due to recessive mutations in tyrosine hydroxylase itself.

Other genes associated with dystonic syndromes are summarized in Table 1.

Paroxysmal movement disorders

Paroxysmal dyskinesias are rare movement disorders that are currently classified into three groups: paroxysmal non-kinesigenic dyskinesia, paroxysmal kinesigenic dyskinesia, and paroxysmal exercise-induced dyskinesia.

Paroxysmal non-kinesigenic dyskinesia (formerly termed paroxysmal dystonic choreoathetosis) is distinguished by attacks, mainly choreoathetotic, lasting from 5 min to 4 h, which are usually precipitated by alcohol, fatigue, coffee, tea, or excitement, but not by sudden movement. Between attacks, neurological examination is usually normal. This disorder is usually inherited in an autosomal dominant fashion, and sporadic cases are rare. It has been linked to a 4 cM area on chromosome 2q33–35, but the responsible gene has not yet been identified.

In paroxysmal kinesigenic dyskinesia, typically the paroxysms consist of dystonic posturing and choreoathetotic or ballistic movements. All attacks are brief (usually less than 2 min) and are precipitated by sudden movements. Frequency may be as high as 100 attacks each day. Patients with this disorder usually respond to anticonvulsants. Linkage of this disorder to chromosome 16 has recently been described, but no genes have yet been cloned.

In paroxysmal exercise-induced dyskinesia, attacks are mainly dystonic and in most cases predominantly involve the legs. They are precipitated by exercise and are not brought on by sudden movements. Response to medication is generally poor. Recently, a novel autosomal recessive syndrome characterized by rolandic epilepsy, paroxysmal exercise-induced dyskinesia, and writer's cramp has been linked to a region on chromosome 16. There appears to be at least two genes causing

paroxysmal movement disorders on chromosome 16.

These disorders with their overlap with other paroxysmal movement disorders, epilepsy, and some response to anticonvulsant medication implicate an ion channel disorder.

The ataxias

The clinical features of the inherited ataxias are described in Chapter 24.6.1. Inherited ataxic disorders can be divided according to their mode of inheritance. Most autosomal recessive disorders are of early onset (less than 20 years), and autosomal dominant disorders are usually of later onset (over 20 years). X-linked inheritance of ataxia is exceedingly rare.

Autosomal recessive ataxias

Friedreich's ataxia

Friedreich's ataxia is the most common of the autosomal recessive ataxias and accounts for at least 50 per cent of cases of hereditary ataxia in most large series reported from Europe and the United States. The prevalence of the disease in these regions is similar, between 1 and 2 per 100 000.

Cloning of the gene showed that the predominant mutation was a trinucleotide repeat (GAA) in intron 1. Expansion of both alleles was found in over 96 per cent of patients and the remaining alleles carry point mutations. To date no cases have been reported of two point mutations in a singe case, therefore the absence of a least one expansion is very strong evidence against the diagnosis of Friedreich's ataxia. This has permitted the introduction of a specific and sensitive diagnostic test, as it is a relatively simple matter to measure the repeat size. On normal chromosomes the number of GAA repeats varies from 7 to 22 units, whereas on disease chromosomes, the range is anything from around 100 to 2000 repeats. The number of repeats is a determinant of age of onset and therefore to some degree influences the severity.

The exact mechanism of action of this repeat in intron 1 is not known, but it is possible that this huge expansion disrupts normal splicesome binding and therefore exon 1 is not spliced correctly to exon 2. This results in a reduction in mRNA and protein levels accordingly.

Other autosomal recessive ataxias

The other early-onset ataxias listed in Table 4 are rare.

Autosomal dominant cerebellar ataxias

The autosomal dominant cerebellar ataxias are a clinically and genetically complex group of neurodegenerative disorders divided into three types (see Chapter 24.6.1). Type I is characterized by a progressive cerebellar ataxia and is variably associated with other extracerebellar neurological features such as ophthalmoplegia, optic atrophy, peripheral neuropathy, and pyramidal and extrapyramidal signs. The presence and severity of these signs is, in part, dependent on the duration of the disease. Mild or moderate dementia may occur, but it is usually not a prominent early feature. Type II is clinically distinguished from type I by the presence of pigmentary macular dystrophy, whereas type III is a relatively 'pure' cerebellar syndrome and generally starts at a later age. This clinical classification is still useful, despite the tremendous improvements in our understanding of the genetic basis, because it provides a framework which can be used in the clinic and helps direct the genetic evaluation.

Molecular genetics of the autosomal dominant cerebellar ataxias

The classification of the autosomal dominant cerebellar ataxias is potentially confusing. The progress in our understanding of the genes and mutations has led to an additional classification system, but luckily there are many common features between these disorders. The first autosomal dominant cerebellar ataxia to be linked (to chromosome 6p) was labelled spinocerebellar ataxia type 1 (**SCA1**). Thereafter, each new locus was given a subsequent SCA number; SCA2 to chromosome 12q, and so on. At the time of writing there are currently 13 SCA loci. Of these, mutations have been identified definitively for SCA 1, 2, 3, 6, 7, and 12. SCA 12 has only been described in two families to date. The others have all been described in many different populations. They are all caused by an expansion of an exonic CAG repeat. The resultant proteins all possess an expanded polyglutamine tract and there are now at least eight conditions caused by these expansions. This common mutational scheme is discussed above with reference to Huntington's disease.

Autosomal dominant periodic ataxia

Autosomal dominant periodic ataxia is characterized by the childhood or adolescent onset of attacks of ataxia, dysarthria, vertigo, and nystagmus. Not all patients have affected relatives. There are at least two forms of this disorder.

Episodic ataxia type 1

The attacks tend to be relatively brief (minutes and occasionally hours) and clinically and electrophysiologically myokymia may be seen. Mutations in a potassium channel have been found. These patients may benefit from acetazolamide, and phenytoin has also been reported to be useful. Patients tend to be neurologically normal between the attacks.

Episodic ataxia type 2

The attacks tend to be longer, lasting hours or even days. They are usually associated with vertigo and consequent nausea and vomiting. They tend to be more severe in childhood with associated drowsiness, headache, and fever. Although when the disease first begins the patients are well between attacks, an interictal nystagmus can be seen. As the years pass, a slowly progressive ataxia is seen. MRI may reveal cerebellar atrophy. These patients tend to respond better to acetazolamide therapy than patients with episodic ataxia type 1. Mutations in a calcium channel gene on chromosome 19q have been demonstrated. Other mutations in this gene have been described in families with familial hemiplegic migraine and a form of dominant ataxia (SCA6), that is, they are allelic.

Further reading

Bandmann O, Marsden CD, Wood NW (1998). Atypical presentations of DRD mutations. *Advances in Neurology* **78**, 283–90.

Bhatia KP (1999). The paroxysmal dyskinesias. *Journal of Neurology* **246**, 149–55.

Brice A (1998). Unstable mutations and neurodegenerative disorders. *Journal of Neurology* **245**, 505–10.

Conrad C *et al.* (1997). Genetic evidence for the involvement of tau in progressive supranuclear palsy. *Annals of Neurology* **41**, 277–81.

Davies SW *et al.* (1997). Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell* **90**, 537–48.

De Silva R, Khan NL, Wood NW (2000). New developments in the genetics of Parkinson's disease. *Current Opinion in Genetics and Development* **10**, 292–8.

Enevoldson PG, Sanders MD, Harding AE (1994). Autosomal dominant cerebellar ataxia with pigmentary macular dystrophy: a clinical and genetic study of eight families. *Brain* **117**, 445–60.

Harding AE (1984). *The hereditary ataxias and related disorders*. Churchill Livingstone, Edinburgh.

Hutton M *et al*. (1998). Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* **393**, 702–5.

Kitada T *et al.* (1998). Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature* **392**, 605–8.

Ozelius LJ *et al*. (1989). Human gene for torsion dystonia located on chromosome 9q32–34. *Neuron* **2**, 1427–34.

Ozelius LJ *et al*. (1997). The early-onset torsion dystonia gene (DYT1) encodes an ATP-binding protein. *Nature Genetics* **17**, 40–8.

Reddy PH, Williams M, Tagle DA (1999). Recent advances in understanding the pathogenesis of Huntington's disease. *Trends in Neurosciences* **22**, 248–55.

Valente EM *et al.* (1998). The role of DYT 1 in primary torsion dystonia in Europe. *Brain* **121**, 2335–9.

24.7 Lumbar puncture

Robert A. Fishman

Indications

Lumbar puncture should be performed only after clinical evaluation of the patient and consideration of the potential value and hazards of the procedure. The cerebrospinal fluid findings are important in the differential diagnosis of the gamut of central nervous system infections, meningitis, and encephalitis, as well as subarachnoid haemorrhage, confusional states, acute stroke, status epilepticus, meningeal malignancies, demyelinating diseases, and central nervous system vasculitis. Examination of the cerebrospinal fluid is usually necessary in patients with suspected intracranial bleeding to establish the diagnosis, although computed tomography (**CT**), when available, may be more valuable. For example, primary intracerebral haemorrhage or post-traumatic haemorrhage is often readily observed with CT making lumbar puncture an unnecessary hazard. However, in primary subarachnoid haemorrhage lumbar puncture may establish the diagnosis when CT is falsely negative. Lumbar puncture is useful to ascertain that the cerebrospinal fluid is free of blood before anticoagulant therapy for stroke is begun. (However, extensive subarachnoid bleeding is a rare complication of heparin anticoagulation, begun several hours after a traumatic bloody tap. Therefore heparin therapy should not begin for at least an hour after a bloody tap.) Lumbar puncture has limited therapeutic usefulness, for example in intrathecal therapy in meningeal malignancies and fungal meningitis.

Contraindications

Lumbar puncture is contraindicated in the presence of infection in the skin overlying the spine. A serious complication of lumbar puncture is the possibility of aggravating a pre-existing, often unrecognized, brain herniation syndrome (for example uncal, cerebellar, or cingulate herniation) associated with intracranial hypertension. This hazard is the basis for considering papilloedema to be a relative contraindication to lumbar puncture. The availability of CT has simplified the management of patients with papilloedema. If CT reveals no evidence of a mass lesion, then lumbar puncture is usually needed in the presence of papilloedema to establish the diagnosis of pseudotumour cerebri and to exclude meningeal inflammation or malignancy.

Thrombocytopenia and other bleeding diatheses predispose patients to needle induced subarachnoid, subdural, and epidural haemorrhage. Lumbar puncture should be undertaken only if urgently needed when the platelet count is depressed to about 50 000/µl or below. Platelet transfusion just before the puncture is recommended if the count is below 20 000/µl or dropping rapidly. The administration of protamine to patients on heparin and vitamin K or fresh frozen plasma to those receiving warfarin is recommended before lumbar puncture to minimize the hazard of the procedure.

Complications

Complications of lumbar puncture include worsening of brain herniation and spinal cord compression, headache, subarachnoid bleeding, diplopia, backache, and radicular symptoms. Post-lumbar puncture headache is the most common complication, occurring in about 25 per cent of patients and usually lasting 2 to 8 days. It results from low cerebrospinal fluid pressures due to persistent fluid leakage through the dural hole. Characteristically, pain is present in the upright position and is promptly relieved with a supine position. Aching of the neck and low back is common. The headaches are aggravated by cough or strain and may be associated with nausea, vomiting, or tinnitus. They are less likely if a small syletted needle is used and if multiple puncture holes are not made. The management of post-spinal headache depends upon strict bedrest in the horizontal position, adequate hydration, and simple analgesics. If conservative measure fail, the use of a 'blood patch' is indicated. The technique utilizes the epidural injection of autologous blood close to site of the dural puncture to form a thrombotic tamponade which seals the dural hole.

Cerebrospinal fluid

The cerebrospinal fluid pressure should be measured routinely. The pressure level within the right atrium is the reference level with the patient horizontal in the lateral decubitus position. The normal lumbar cerebrospinal fluid pressure ranges between 50 and 200 mmH_2O (and as high as 250 mmH_2O in very obese subjects). With the use of the clinical manometer the arterially derived pulsatile pressures are obscured but respiratory pressure waves reflecting changes in central venous pressures are visible. Low pressures are seen in dehydration, spinal subarachnoid block, following previous lumbar puncture or other cerebrospinal fluid leaks, or may be technical in origin because of faulty needle placement. Increased pressures occur with brain oedema, intracranial mass lesions, infections, acute stroke, cerebral venous occlusions, congestive heart failure, pulmonary insufficiency, and benign intracranial hypertension (pseudotumour cerebri) of diverse aetiology.

Normal cerebrospinal fluid contains no more than five lymphocytes or mononuclear cells per microlitre. A higher white cell count is pathognomonic of disease in the central nervous system or meninges. A stained smear of the sediment is needed for an accurate differential cell count. A variety of centrifugal and sedimentation techniques have been used. A pleocytosis occurs with the gamut of inflammatory disorders. The changes characteristic of the various meningitides are listed in Table 1. The heterogeneous forms of neuro-AIDS also are associated with a wide range of cellular responses. Other disorders associated with a pleocytosis include brain infarction, subarachnoid bleeding, cerebral vasculitis, acute demyelination, and brain tumours. Eosinophilia most often accompanies parasitic infections, for example cysticercosis. Cytological studies for malignant cells are rewarding with some central nervous system neoplasms.

Bloody cerebrospinal fluid due to needle trauma contains increased numbers of white cells contributed by the blood. A useful approximation to a true white cell count can be obtained by the following correction for the presence of the added blood: if the patient has a normal haemogram, subtract from the total white cell count (WBC; per µl) one white cell for each 1000 red blood cells (RBC) present. Thus, if bloody fluid contains 10 000 red cells and 100 white cell/µl, ten white cells would be accounted for by the added blood and the corrected leucocyte count would be 90/µl. If the patient's haemogram reveals significant anaemia or leucocytosis, the following formula may be used to determine more accurately the number of white cells (W) in the spinal fluid before the blood was added:

$$W = \text{blood WBC} \times \text{cerebrospinal fluid RBC/blood RBC} \times 100.$$

The presence of blood in the subarachnoid space produces a secondary inflammatory response which leads to a disproportionate increase in the number of white cells. Following an acute subarachnoid haemorrhage, this elevation in the white cell count is most marked about 48 h after onset, when meningeal signs are most striking.

To correct cerebrospinal fluid protein values for the presence of added blood due to needle trauma, subtract 0.001 g for every 1000 red blood cells. Thus, if the red cell count is 10 000/µl and the total protein is 1.1 g/l the corrected protein level would be about 1 g/l. The corrections are reliable only if the cell count and total protein are both made on the same tube of fluid.

Blood in the cerebrospinal fluid: differential diagnosis and the three-tube test

To differentiate between a traumatic spinal puncture and pre-existing subarachnoid haemorrhage, the fluid should be collected in at least three separate tubes (the

'three-tube test'). In traumatic punctures, the fluid generally clears between the first and the third collections. This is detectable with the naked eye and should be confirmed by cell count. In subarachnoid bleeding, the blood is generally evenly admixed in the three tubes. A sample of the bloody fluid should be centrifuged and the supernatant fluid compared with tap water to exclude the presence of pigment. The supernatant fluid is crystal clear if the red cell count is less than about 100 000 cells/µl. With bloody contamination of greater magnitude, plasma proteins may be sufficient to cause minimal xanthochromia; this requires enough serum to raise the cerebrospinal fluid protein concentration to about 1.5 g/l.

Following subarachnoid haemorrhage, the supernatant fluid usually remains clear for 2 to 4 h and even longer after the onset of subarachnoid bleeding. The clear supernatant fluid may mislead the physician to conclude erroneously that the observed blood is due to needle trauma in patients who have had a lumbar puncture within 4 h of aneurysmal rupture. After an especially traumatic puncture, some blood and xanthochromia may be present for as long as 2 to 5 days following the initial puncture. In pathological states associated with a cerebrospinal fluid protein greater than 1.5 g/l, and in the absence of bleeding, very faint xanthochromia may be detected. When the protein is elevated to much higher levels, as in spinal block, polyneuritis, and meningitis, the xanthochromia may be considerable. A xanthochromic fluid with a normal protein level or a minor elevation to less than 1.5 g/l usually indicates a previous subarachnoid or intracerebral haemorrhage (rarely the xanthochromia is due to severe jaundice, carotenaemia, or rifampin).

Pigments

Two major pigments derived from red cells may be observed in cerebrospinal fluid—oxyhaemoglobin and bilirubin. Methaemoglobin is only seen spectrophotometrically. Oxyhaemoglobin, released with lysis of red cells, may be detected in the supernatant fluid within 2 h of a subarachnoid hemorrhage. It reaches a maximum in about the first 36 h and gradually disappears over the next 7 to 10 days. Bilirubin, is produced *in vivo* by leptomeningeal cells following red cell haemolysis. Bilirubin is first detected about 10 h after the onset of subarachnoid bleeding. It reaches a maximum at 48 h and may persist for 2 to 4 weeks after extensive bleeding. The severity of the meningeal signs associated with subarachnoid bleeding correlates with the inflammatory response, i.e. the leucocytic pleocytosis.

Total protein

The total protein level of cerebrospinal fluid ranges between 1.5 and 5 g/l. While an elevated protein level lacks specificity, it is an index of neurological disease reflecting a pathological increase in the permeability of endothelial cells. Greatly increased protein levels, 5 g/l and above, are seen in meningitis, bloody fluids, or cord tumour with spinal block. Polyneuritis (Guillain Barre syndrome), diabetic radiculoneuropathy, and myxoedema may also increase the level to 1 to 3 g/l. Low protein levels, below 0.15 g/l, occur most often with cerebrospinal fluid leaks due to a previous lumbar puncture or traumatic dural fistula.

Immunoglobulins

Although a vast number of proteins may be measured in cerebrospinal fluid only an increase in immunoglobulins is of diagnostic importance. Such increases are indicative of an inflammatory response in the central nervous system and occur with immunological disorders, and bacterial, viral, spirochaetal, and fungal diseases. Immunoglobulin assays are most useful in the diagnosis of multiple sclerosis, other demyelinating diseases, and central nervous system vasculitis. The cerebrospinal fluid level is corrected for the entry of immunoglobulins from the serum by calculating the IgG index (see Table 1). More than one oligoclonal band in cerebrospinal fluid with gel electrophoresis (and absent in serum) is also abnormal, occurring in 90 per cent of multiple sclerosis cases and variably in the gamut of inflammatory diseases including central nervous system vasculitis.

Glucose

The concentration of glucose in cerebrospinal fluid is dependent upon the concentration in the blood. The normal range of glucose concentration in cerebrospinal fluid is between 2.5 and 4.5 mmol/l in patients with a blood glucose between 4 and 7 mmol/l, i.e. 60 to 80 per cent of the normal blood level. Cerebrospinal fluid glucose values between 2.2 and 2.5 mmol/l are usually abnormal, and values below 2.2 mmol/l invariably so. Hyperglycaemia during the 4 h prior to lumbar puncture results in a parallel increase in cerebrospinal fluid glucose. The latter approaches a maximum and the cerebrospinal fluid/blood ratio may be as low as 0.35 in the presence of a greatly elevated blood glucose level and in the absence of any neurological disease. An increase in cerebrospinal fluid glucose is of no diagnostic significance apart from reflecting hyperglycaemia within the 4 h prior to lumbar puncture. The cerebrospinal fluid glucose level is abnormally low (hypoglycorrhachia) in several diseases of the nervous system apart from hypoglycaemia. It is characteristic of acute purulent meningitis, and is a usual finding in tuberculous and fungal meningitis. It is usually normal in viral meningitis, although reduced in about 25 per cent of mumps cases, and in some cases of herpes simplex and zoster meningoencephalitis. The cerebrospinal fluid glucose is also reduced in other inflammatory meningitides including cysticercosis, amoebic meningitis (*Nagleria*), acute syphilitic meningitis, sarcoidosis, granulomatous arteritis, and other vasculitides. The glucose level is also reduced in the chemical meningitis that follows intrathecal injections, and in subarachnoid haemorrhage, usually 4 to 8 days after the bleed. The major factor responsible for the depressed glucose levels is increased anaerobic glycolysis in adjacent neural tissues and to a lesser degree by polymorphonuclear leucocytes. Thus, the decrease in the cerebrospinal fluid glucose level is accompanied by an inverse increase in the cerebrospinal fluid lactate level.

Microbiological and serological reactions

The use of appropriate stains and cultures is essential in cases of suspected infection. Tests for specific bacterial and fungal antigens (countercurrent immunoelectrophoresis) are useful in establishing a specific aetiology. DNA amplification techniques using the polymerase chain reaction have improved diagnostic sensitivity. Serological tests on cerebrospinal fluid for syphilis include the reagin antibody tests and specific treponemal antibody tests. The former are particularly useful in evaluating cerebrospinal fluid because positive results may occur even in the presence of a negative blood serology. There is no basis for applying the specific treponemal antibody tests to cerebrospinal fluid because these antibodies are derived from the plasma where they are present in greater concentration. The search continues for specific diagnostic markers in cerebrospinal fluid in the heterogeneous degenerative diseases of the central nervous system.

Further reading

Fishman RA (1992). *Cerebrospinal fluid in diseases of the nervous system*,2nd edn. WB Saunders, Philadelphia.

24.8 Disturbances of higher cerebral function

Peter Nestor and John R. Hodges

Introduction

Modern scientific study of higher cerebral function began in the late nineteenth century with the case studies of Broca and Wernicke. Their observations of language disorders associated with damage to the left hemisphere gave rise to the notion that specific mental faculties could be dissociated from each other and localized to specific regions within the cerebral hemisphere. Since that time clinicopathological and, more recently, imaging studies have established associations between specific cognitive disorders and focal brain lesions; these studies also show that some lesions do not give rise to highly specific deficits. The field of neuropsychology has offered complementary insights into this area by providing concepts of how cognitive faculties are organized.

The border between psychiatry and neurology has become less distinct; many patients with brain diseases have psychiatric symptoms, cognitive complaints are prominent in depression and schizophrenia, and a biological basis for many 'functional' psychiatric disorders is now well accepted.

Another critical area has been the study of anatomy: the finding that neocortical histology varies by region led to the development of cytoarchitectonic maps such as that of Brodmann. Brodmann's map has become a shorthand way of discussing regional specialization across the cortex. Meanwhile, anatomical studies of neural tracts have provided insights into how topographically distinct regions may interact.

Handedness and hemispheric dominance

The finding of asymmetric functions in the human cerebral cortex led to the introduction of the term hemispheric dominance. Neuroscientists often refer to cognitive processes being a function of the 'dominant' or 'non-dominant' hemisphere; when such terminology is used, the 'dominant' hemisphere is synonymous with that which underpins language function. In right handers, over 95 per cent have left hemisphere dominance: only rarely does aphasia arise from right hemisphere damage in which circumstance it is referred to as 'crossed aphasia'. In left handers, dominance is more complex and language skills are more often shared between the hemispheres although the left hemisphere is relatively dominant in about 70 per cent of individuals. While the left hemisphere usually specializes in language, the non-dominant hemisphere plays an important role in spatial cognition (with damage to the frontoparietal regions resulting in spatial neglect) and particularly in face processing (with damage to the right occipitotemporal junction producing prosopagnosia).

Primary sensory input and motor output

Motor

The primary motor area lies in the precentral gyrus, immediately rostral to the central sulcus. The body is represented 'somatotopically' along the precentral gyrus; the lower limb at the superomedial, and the face at the inferolateral extremity with the upper limb in between. This is of clinical importance as the vascular supply of the superomedial region is from the anterior cerebral artery whilst the rest of the motor cortex is from the middle cerebral artery. Thus middle cerebral artery territory infarction will affect face and upper limb with relative sparing of the lower limb, and the converse will be the case with anterior cerebral territory occlusions.

Vision

After passing from the retina, via optic nerves and tracts to the lateral geniculate body of the thalamus, visual information passes to the striate cortex of the occipital lobes (primary visual cortex) through the optic radiations (see Chapter 24.11). As images presented to the right visual field are represented on the left retina and conveyed to the left occipital lobe, a lesion of the latter will cause a right homonymous hemianopia (and vice versa for right occipital lesions). Fibres in each optic radiation separate such that input from the superior half of the retina (inferior visual field) runs from lateral geniculate to the striate cortex via parietal white matter whilst that from the inferior retina (superior visual field) loops down into the temporal lobe. Consequently, a lesion of the parietal lobe can cause a contralesional inferior quadrantanopic field defect whilst a temporal lobe lesion can cause a contralesional superior quadrantanopia. Large temporoparietal lesions (for example due to middle cerebral artery occlusion) may also cause homonymous hemianopia which may be distinguished from that due to an occipital lesion by preservation of opticokinetic nystagmus in the latter but not the former.

Bilateral lesions to the primary visual cortex lead to 'cortical blindness' in which vision is lost, but unlike blindness secondary to retinal or optic nerve diseases, pupillary reflexes are preserved. Some cortically blind individuals deny they have any visual disorder at all (namely visual anosagnosia)—a condition known as Anton's syndrome. These cases tend to have more extensive lesions involving both striate and adjacent visual association cortices.

Somatosensory

The primary somatosensory cortex occupies the post-central gyrus of the parietal lobe with a somatotopic representation of the body analogous to that of the primary motor area. Sensory deficits due to lesions of the thalamus, or lower components of the sensory system, cause gross abnormalities in the appreciation of touch, pinprick, temperature, and other sensations, and must be excluded before comment can be made on higher sensory function. Parietal lesions cause specific impairment of 'discriminative' sensation, including joint position sense and two-point discrimination. Parietal drift (the patient is asked, with eyes closed, to maintain the upper limbs outstretched in front of the trunk at 90°) is a sign of impairment of the former ability. It is considered specific for a contralateral parietal lesion when the drift is upward, as a downwards drift may also be a consequence of subtle motor weakness. The normal separation distance at which one can discriminate one point from two varies according to body region: fingertips 3 mm, palm 1 cm, and body surface 4 to 7 cm.

Other signs of parietal sensory impairment are an inability to name numbers traced on the palm of the hand (agraphaesthesia), and an inability to name small objects (such as keys and coins) placed in the patient's hand (astereognosis). Obviously there is potential to confuse true astereognosis with a more general deficit of object naming such as due to loss of semantic knowledge or aphasia (see below). However, ambiguous results on parietal sensory testing can largely be avoided if the examiner adopts a methodical approach of: (i) excluding a lesion below the parietal lobe by establishing that the patient can appreciate, for instance, a pinprick or light touch; and (ii) examining from the suspected normal to abnormal side to exclude a more general impairment of cognitive faculties.

Auditory

Auditory information coming from the cochlear nuclei via the inferior colliculus and the medial geniculate nucleus of the thalamus travels to the primary auditory cortex (Heschl's gyrus) in the posterosuperior temporal lobe. Clinically apparent cortical hearing impairment is uncommon due to the bilateral representation of auditory material from each ear by the cerebral cortex. Bilateral lesions of this area (as a result of strokes, prolonged hypotension, or carbon monoxide poisoning) will cause 'cortical deafness', a rare disorder manifest by inability to understand spoken language or recognize sounds although presence or absence of noise can be determined. Unlike Wernicke's aphasia (see below), subjects can understand written text and their language output is normal.

Cognitive domains

Beyond the primary sensory and motor cortices, the neocortex is made up of unimodal and heteromodal association areas. Unimodal association cortices lie adjacent to their respective primary modality while heteromodal association cortex is found in the prefrontal and temporoparietal regions. Moving from primary through unimodal to heteromodal association cortex, the linkage of topographical region to specific functional attribute becomes progressively less tightly defined. Heteromodal association cortices, as the name implies, receive inputs from multiple unimodal areas but also from non-neocortical areas. Anatomically, as the neocortex approaches the diencephalon, upon which the cerebral hemispheres sit, it transforms into a histologically distinct area: the limbic system. These areas also have critical roles in cognition, particularly in the domains of memory and emotion and have reciprocal projections with heteromodal association cortices.

Other brain regions which have important modulatory roles on cognition include: (i) the basal forebrain nuclei, which contain cholinergic neurones that project extensively to limbic and neocortical regions and are known to be important to the successful encoding of memory; (ii) the basal ganglia, which have reciprocal links to frontal association cortices and have important modulatory roles relating particularly to attention and speed of cognitive processing; and (iii) brainstem reticular formation nuclei which project into the hemispheres via the thalami: the most clearly defined role for these projections being at the level of arousal.

Although the remainder of this section discusses various disorders of higher mental function individually, one should not view these specific deficits as a random and independent collection of phenomena. It cannot be overemphasized that one should always follow a logical sequence in assessing cognitive function so as to avoid false-positive diagnoses due to sequential effects. For example, tests of executive function that utilize analysis of complex verbal material would be beyond the grasp of a patient with Wernicke's aphasia due to the fundamental disorder of language comprehension without needing to implicate frontal lobe damage. Likewise, a patient with an acute delirium may be unable to perform even the most basic memory tasks as a consequence of their attention deficit and therefore ought not to be labelled amnesic. Therefore, regardless of the suspected disorder, one should always bear the following sequence in mind: (i) ensure adequate attention to undergo further testing; (ii) as almost all tests are going to be presented with verbal instruction, assess language comprehension; and (iii) as tests of executive function and praxis often require adequate levels of function in all other cognitive domains, these should be left to last. In summary, always ask 'can this apparent disorder be explained in terms of a more elemental deficit?'

Attention

The ability to attend to a specific sensory stimulus, such as a human voice or passage of text, and to maintain attention is an obligatory first step to any further cognitive processing. Humans are continuously bombarded with sensory stimuli from both within and between individual sensory input modalities; loss of ability to focus and sustain attention (or alternatively, block out irrelevant 'noise') renders the individual incapable of following a specific sensory stimulus (such as a conversation) and at the same time vulnerable to random irrelevant environmental stimuli. Although disorders of the frontal lobes, basal ganglia, and ascending reticular formation are associated with poor attention, it is overly simplistic to consider attention as a localizable brain function. The commonest causes of acute attention failure are diffuse brain insults such as a metabolic encephalopathy or closed head injury; breakdown in attentional processing is the central deficit in delirium or acute confusional states, the main features of which are summarized in Table 1.

Digit span is one of the most simple methods of assessing attention, especially in the backwards condition; normal subjects have a forward span of at least six digits and a reverse span one or two digits less. The digits must be presented as individual items (read the string to be repeated at a rate of one digit per second). A common pitfall is to cluster digits as one does when reciting telephone numbers. This inflates span as each cluster becomes an individual item: compare repeating 6953–8127 with 6 … 9 … 5 … 3 … 8 … 1 … 2 … 7. Ability to persevere at a given task is another way of considering attention; this can be tested by asking the patient to recite the months of the year in reverse order.

Orientation is heavily dependent upon attention and is assessed by questions of time and place. Testing personal orientation adds little, as only profoundly aphasic or hysterical patients are unable to relate their own name. A recent onset of profound disorientation and attention deficit is typical of a delirium. It should be noted that many patients with episodic memory problems (such as early Alzheimer's disease) remain well orientated.

Language and related disorders

Numerous terms are in use to describe aphasic syndromes, although some serve more to confuse than enlighten. The terms 'expressive' and 'receptive' particularly seem to mislead: on the one hand all patients with aphasia have some form of difficulty 'expressing' themselves, and on the other hand 'receptive' aphasia is often, erroneously, taken to mean that patients have difficulty only with incoming language, but can produce their own language perfectly well. Less ambiguous terms for the two principal divisions of aphasia are 'non-fluent' and 'fluent', which correspond in classic aphasia nomenclature to Broca's and Wernicke's aphasia. The classic aphasia syndromes are, however, rarely seen in the acute stages after stroke and do not characterize the language deficits found in the dementias. A better approach is therefore to consider language fluency and paraphasias in spontaneous conversation, comprehension, naming, and repetition.

Examining patients with aphasia

Fluency and paraphasic errors

Speech can be described as fluent if the patient is able to produce some well-formed sentences or phrases even if empty or anomic (such as 'Oh you know, the thing you put the stuff in when you're going somewhere and …'). Non-fluent language, in contrast, is a consequence of breakdown of the language production and syntactic (grammatical) aspects of language and is the hallmark of damage to Broca's area and the insula. Output is laboured or 'telegraphic', with often as few as two or three words per minute; in spite of which they can convey meaning fairly successfully, for example 'I … go … hospital'.

Paraphasic errors are substitutions of a correct word for one related in sound or meaning. The former, known as phonological or phonemic paraphasias, involve the substitution of related sound fragments ('phonemes') such as 'dobble' for 'bottle'. Semantic paraphasic errors involve substitution of words of related meaning; the substituted word is typically a higher frequency example of the same semantic category (such as 'dog' for 'fox') or else of a superordinate category (such as 'animal' for 'fox'). In more extreme circumstances, paraphasic substitutions may not be words at all ('neologisms'); fluent output with virtually continuous neologisms is an utterly incomprehensible state sometimes referred to as jargon aphasia. Patients with lesions to Wernicke's area invariably make a mixture of phonemic and semantic errors. Semantic errors are also very common in Alzheimer's disease and semantic dementia. In Broca's aphasia, phonological errors predominate.

Comprehension

Some degree of impairment of language comprehension can be detected in both fluent and non-fluent aphasia. Patients with fluent aphasia have more overtly impaired comprehension of word meaning (for example ordinary nouns). In mild cases this can be demonstrated with semantically complex language tasks (such as 'Can you point to a source of artificial illumination?') or by defining uncommon words (such as 'What is an aubergine, accordion etc.'). Comprehension of single nouns is preserved in patients with non-fluent aphasia, but comprehension—in addition to production—of complex grammar is impaired. This can be tested with reversible passive sentences (such as 'The lion was eaten by the tiger, who survived?') or by asking the patient to obey syntactically complex commands (such as 'Touch the keys after touching the book').

Anomia

Naming is a complex task that requires the integrity of three basic processes: visual analysis, semantic knowledge (see Memory, below), and word production

(phonology). Virtually all patients with aphasia are anomic when tested using items of low familiarity and late age of acquisition. The type of naming error and the ability to circumvent the deficit varies, however, according to the locus of damage. Patients with visuoperceptive deficits that produce visual errors (a 'head' for a 'mushroom' etc.) have retained tactile naming and can give correct responses when asked to put a name to a description ('What do we call the large grey African animal with a trunk?').

A breakdown in the central semantic process causes impairment in naming from all modalities, while phonological deficits produce phonological errors regardless of the mode of input.

Repetition

Lesions involving any of the peri-Sylvian language structures are almost always associated with impaired repetition, although this may not be apparent unless multisyllabic words ('caterpillar', 'fundamental' etc.) and phrases ('no ifs, ands, or buts') are tested. Certain aphasic syndromes (see below) show either disproportionate impairment or preservation of repetition.

Aphasic syndromes

Broca's aphasia

This classic form of non-fluent aphasia is characterized by grossly distorted speech output with impaired production and comprehension of syntax. Phonological paraphasias are common and there is impaired repetition of phrases. It is associated with lesions to the left ventrolateral frontal lobe (Broca's area); owing to its close proximity to the motor cortex, when focal lesions (such as stroke or tumour) cause a Broca's aphasia it is typically associated with a right hemiparesis. The distortion of language output, often described as speech apraxia, is thought to relate to concurrent damage to structures within the insula, which is almost always affected.

Wernicke's aphasia

In Wernicke's aphasia there is fluent although vacuous output with a mixture of semantic and phonological paraphasic errors and often neologisms. There is also impaired comprehension of word meanings and impaired repetition. In contrast to the fundamental loss of word meaning seen in patients with semantic dementia and destruction of the left inferior temporal lobe after herpes simplex encephalitis, patients with Wernicke's aphasia have breakdown in the mapping between speech and meaning systems. Lesions localize to the posterior portion of the left superior temporal gyrus—known as Wernicke's area. As this area overlies the optic radiation, the commonest neighbourhood sign is a right homonymous hemianopia.

Conduction aphasia

This form of aphasia, as the name implies, is due to a disconnection of the two principle language areas. Comprehension is relatively preserved and output is fluent although phonemic paraphasias occur. The striking abnormality is an impairment of repetition even for single syllable words such that attempts at repeating are laboured and contain phonemic errors. Likewise, naming produces phonemic errors even for high frequency items (such as for 'cup': 'cah … cahb … cub' etc.). Lesions producing conduction aphasia occur in the region of the supramarginal gyrus, and particularly, the underlying arcuate fasciculus, the tract linking the anterior and posterior language areas.

Global aphasia

In this devastating form of aphasia there is derangement of all aspects of language; patients with global aphasia are non-fluent and have impaired word comprehension, repetition, and naming. Language output is restricted to infrequent unintelligible noises or, at best, a single word or clichéd phrase. As the blood supply to both language areas is from the middle cerebral artery, global aphasia is not uncommon secondary to proximal occlusion of this vessel. Consequently these patients are usually also hemiplegic and hemianopic.

Atypical aphasias and the dementias

The term 'transcortical aphasia' is a legacy of an abandoned neural explanation for a distinct category of aphasia. Although the term is meaningless in its originally coined anatomical sense, it is still sometimes used to describe a distinct syndrome in which a patient with aphasia shows preservation of repetition. Patients with so-called transcortical sensory aphasia are fluent and show profound impairment in word comprehension with preserved repetition; this syndrome is also referred to as amnesic aphasia reflecting the loss of word meaning. It is seen in semantic dementia (the temporal lobe variant of frontotemporal dementia or Pick's disease) and advanced Alzheimer's disease. Earlier in Alzheimer's disease impaired naming with intact fluency and word comprehension (sometimes called anomic aphasia) is commonly seen. Impairment in language output with preserved repetition, transcortical motor aphasia, is most often associated with dorsomedial frontal lesions that involve the supplementary motor area.

Dyslexia

Patients with aphasia show dyslexic difficulties in keeping with their type of aphasia, thus those with fluent aphasia will struggle to understand the meaning of words in printed form, while those with non-fluent aphasia have trouble with grammatical aspects or reading (particularly word endings: -ed, -ing etc.). Within acquired dyslexia, however, dissociations have been defined for reading single words, these syndromes are known as deep and surface dyslexia.

Deep dyslexia and surface dyslexia

There may be a dissociation between ability to read orthographically regular (pronounced as they are spelt) words such as mint, flint, and hat, and irregular words such as pint, cellist, and island. Difficulty reading the latter type is known as surface dyslexia and is one of the hallmarks of semantic dementia; for an irregular word such as 'pint' or 'yacht' to be read correctly, the reader must access knowledge of the word meaning as the graphical representation of the word alone (that is, its 'surface' structure) will not lead to correct pronunciation. If the semantic knowledge base (located in the dominant temporal lobe) breaks down, then the word can only be pronounced according to the rules of graphical to phonological translation and thus 'pint' will be pronounced like 'mint' (known as a 'regularization' error); in other words, analogous to how a normal person would pronounce a non-word, such as 'rint'.

A complimentary syndrome is that of deep dyslexia in which patients produce semantic paralexias when reading (reading 'prison' for 'gaol' or 'beer' for 'pint'), are unable to read non-words, and have greater difficulty with abstract than concrete words. This, simplistically, is thought to represent a loss of the grapheme to phoneme route with intact semantic knowledge (that is,. its 'deep' meaning). Deep dyslexia is typically seen in patients with extensive left hemisphere lesions and global aphasia.

Alexia without agraphia

This syndrome represents a classic disconnection syndrome of visual input from language areas due to a lesion in the left occipital lobe and adjacent splenium (disrupting input form the right occipital lobe); as such, although the right occipital cortex is capable of registering text, the information cannot be decoded by the language hemisphere. Patients are not aphasic and can write normally, they cannot read but can say words spelt out loud to them. Visual field testing shows a right homonymous hemianopia. Patients rapidly relearn how to read by identifying individual letters and reconstructing words by a laborious and slow letter-by-letter reading strategy.

Agraphia

Various acquired disorders of writing occur as homologues of other cognitive deficits. For instance, patients with aphasia make writing errors consistent with their aphasic syndrome (for example patients with a Broca's aphasia will make errors in writing syntax), deep and surface dysgraphia give rise to similar errors as deep and surface dyslexia, and ideomotor apraxia (see below) will cause a disorder in motor execution such that writing will be of poor quality.

Visuospatial and perceptual disorders

The regions of the brain concerned with the higher order analysis of visual information can be divided into a dorsal (occipitoparietal) pathway concerned with spatial information and preparation for reaching, and a ventral (occipitotemporal) pathway concerned with identifying visual stimuli. In other words, the dorsal stream is involved in 'where?' and 'how?' and the ventral with 'what?' information for a given visual stimulus. Some of the most striking neuropsychological syndromes are seen following selective damage to one stream.

The dorsal stream and Balint's syndrome

Constructional apraxia, an inability to draw or copy line drawings such as wire cubes and clock faces, is a common finding in parietal pathology, particularly with right-sided lesions. More severe breakdown in spatial cognition causing individuals to misreach for visually guided targets, to trip on steps, or collide with furniture when walking is seen with bilateral parietal diseases (such as watershed infarction, the biparietal variant of Alzheimer's disease, and venous sinus thrombosis) and results clinically in Balint's syndrome; the features of which are simultanagnosia, optic ataxia, and ocular apraxia. Simultanagnosia is the inability to integrate and make sense of an overall visual scene in spite of preservation in the ability to identify individual elements. Such patients are relatively better at identifying small objects; this can also be demonstrated by an inability to read vertically printed words although they can be read when printed normally. Ocular apraxia describes the inability to direct gaze to a novel visual stimulus, while optic ataxia is the inability to reach accurately for a visually guided target.

Spatial neglect

Although considered under the visuospatial heading, spatial neglect is really a cross-modality disorder that typically involves the neglect of all sensory information (visual, tactile, auditory) from the side contralateral to the lesion. Chronic neglect virtually only occurs in the context of right parietal lobe damage. Right hemispatial neglect following an acute left parietal lesion can occur, but is usually less severe and tends to resolve within days. In addition to being a cross-modality disorder it is not correct to define a 'hemispatial field' in purely retinotopic terms. For instance, if a patient who exhibits neglect on visual field testing has the body turned to face the neglected extrapersonal hemispace (with head and eyes fixed in the original position), then the neglected space is reduced.

Visual neglect is best tested by cancellation (crossing off 'A's on a sheet of paper containing randomly arranged letters), drawing (clock, house, flower), or line bisection tasks. Patients with severe visual neglect may even appear to be hemianopic. A milder form of neglect can be elicited by 'sensory extinction' of the neglected side during bilateral sensory (visual and somatosensory at the bedside, although auditory neglect can be demonstrated experimentally) stimulation. Patients often have associated hemiparesis, although as part of their neglect syndrome they may deny this impairment—a phenomenon known as anosagnosia. When presented with the hemiparetic limb they may even deny that it is their own.

The ventral stream

Lesions to the occipitotemporal pathway give rise to difficulty recognizing visual stimuli that is not a consequence of being unable to appreciate where an object is in space as is the case in simultanagnosia. This deficit is known as visual object agnosia and has been divided further into aperceptual and associative varieties. In aperceptual agnosia, basic aspects of vision (acuity, fields, and contrast sensitivity) are intact, but patients cannot identify, or match identical, objects and have grave difficulty copying line drawings, although knowledge of these objects is intact if tested using other inputs such as describing from name. In contrast, associative agnosia describes a state where loss of object knowledge occurs such that although patients can copy line drawings well and match perceptually identical pictures, they cannot match non-perceptually identical images such as different angles of the same face or, for instance, tell that two different types of clock are both clocks. Associative agnosia is a cross-modality disorder such that knowledge of objects is impaired in non-visual modalities—in other words one component of generalized failure of semantic knowledge (see below). Differentiating these agnosias requires the use of test material only found in neuropsychology laboratories. One component of object knowledge is colour, loss of which (achromatopsia) usually accompanies occipitotemporal lesions and is more accessible to bedside evaluation.

A restricted form of impaired object recognition relates to faces. Known as prosopagnosia, the subject can no longer recognize previously familiar faces but can recognize their voices and have access to knowledge from their names. Usually bilateral lesions of the inferior occipitotemporal junction are responsible, although cases with lesions restricted to just the right side have been described.

Memory

Memory is divided by researchers into implicit and explicit subtypes (also known as non-declarative and declarative, respectively). Implicit memory refers to unconscious memory systems such as that responsible for conditioning as well as memory for motor tasks such hitting a golf ball or playing a piece of music 'by heart'. Explicit memory, in contrast, refers to consciously apprehended emory and is further divided into episodic and semantic memory. In clinical terms, when one refers to memory, it is only the explicit type of memory which is considered. When assessing memory complaints it is useful to apply a theoretically motivated approach to analysing symptoms according to the subcomponent of memory involved. In broad terms, memory subtypes can be considered under the following headings.

Working memory

Working memory refers to the amount of information that can be held by the brain 'on-line' (such as reading a phone number then holding it as the object of one's attention until the number can be dialled, or solving mathematical problems in the head); in the absence of rehearsal, when the focus of one's attention has moved to a novel topic for more than a few seconds, such items are lost. Working memory is also referred to as 'short-term' memory by psychologists, although this latter term is often used by patients and their doctors to describe recently acquired episodic memory (see below); it also involves aspects of attention (see above) so, to avoid confusion, the term 'working memory' is preferable. Slips of working memory are often erroneously seen by patients as the harbinger of dementia and thus these individuals are commonly referred to memory clinics: these lapses of attention (such as forgetting why you opened the refrigerator door or went into the study, or immediately forgetting a new telephone number) are common everyday symptoms which are increased with anxiety, depression, and also occur more commonly with advancing age. Complaints of this type are also common after head injury and in basal ganglia disorders.

Semantic memory

Semantic memory refers to the brain's knowledge store of, for example, objects and word meanings; it is also the term applied to knowledge of facts, such as that Paris is the capital of France, canaries are small yellow birds kept as pets, or that Ronald Reagan was a president of the United States. The inferolateral and polar regions of the temporal lobes (left for word and object meanings, right for face knowledge) are particularly critical to supporting semantic knowledge. Loss of memory for words is the usual complaint in patients with a primary disorder of semantic memory such as semantic dementia (also known as progressive fluent aphasia) and after herpes simplex virus encephalitis. However, it is important to distinguish between the occasional word finding lapse, usually for proper nouns, which occurs normally (especially in later life), and the relentlessly progressive loss of vocabulary, which occurs in association with left temporal lobe pathology. Low frequency words are the most vulnerable and patients with semantic dementia often have some insight into this problem in the early stages. For instance, a carpenter may complain that he can no longer remember the names of tools. Alzheimer's sufferers show a similar phenomenon, although it is usually overshadowed by their profound episodic memory deficit.

Breakdown in semantic memory manifests as inability to name objects or drawings with the production of broad superordinate responses (such as 'animal' for 'elephant') and the inability to define the meaning of words. Category fluency (the ability to generate exemplars from a given semantic category such as types of animals or kitchen utensils or birds) is another sensitive measure of semantic memory. Knowledge of famous people can be tested by identifying photographs and names or asking the patient to list prime ministers in chronological order.

Episodic memory

Episodic memory refers to the event-based memories unique to each individual, in other words our recollection of personally experienced episodes (indeed, it is sometimes termed 'autobiographical' memory). Difficulty with the acquisition of new event-based memories (such as inability to recall details of a television programme or conversation with a friend despite good attention at the time) is the hallmark of early Alzheimer's disease and other causes of the amnesic syndrome (Table 2). Lesions that give rise to amnesia involve the limbic system of the brain (especially the hippocampi and their connections; Fig. 1). Although bilateral involvement is usually required to cause a full-blown amnesic syndrome, neuropsychological testing can often reveal a selective deficit in verbal or non-verbal

memory in cases of left- or right-sided damage, respectively. Retrograde memory (established prior to the amnesic insult) is typically better than anterograde (established any time after) in amnesic syndromes and within retrograde memory, very remote memory is classically (although not universally) better preserved than recent memory.

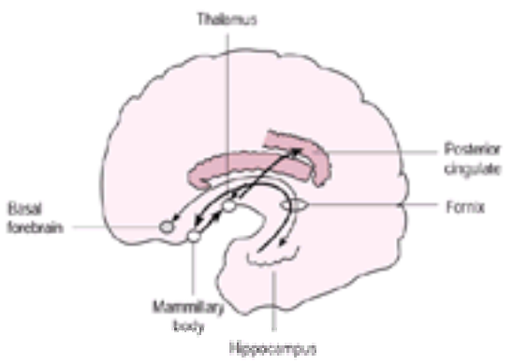

Fig. 1 Principal connections of structures critical to sustaining human memory.

On examination, patients with amnesia have a striking inability to relate anecdotes from their recent life, although in cases of basal forebrain amnesia they may offer confabulations. Amnesia can be assessed in the clinic by asking the patient to learn some information such as a new name and address; patients with amnesic syndromes (including early Alzheimer's disease) typically repeat a name and address perfectly after two to three trials, but show very rapid forgetting and recall little or nothing after a delay of a few minutes of a distracting task.

Amnesia may occur as a temporary state as is seen with transient global amnesia, in which there is a sudden onset of severe amnesia that lasts several hours before resolution; afterwards the patient, characteristically elderly, is left with an islet of amnesia for the hours of the episode. Transient global amnesia typically occurs as a solitary episode; recurrent attacks of self-limiting amnesia occasionally occur as a consequence of epileptic activity, hence the term transient epileptic amnesia.

Apraxia

Apraxia is defined as a loss of ability to carry out skilled motor tasks that cannot be explained in terms of an elementary disorder of motor control (weakness or ataxia), primary sensory disturbance, or a global impairment of cognition. In the early twentieth century Leipmann distinguished three types of apraxia—limb-kinetic, ideomotor, and ideational—and although these terms have suffered from a lack of universally accepted definition they are still widely used today. In an attempt to clear up ambiguity, the terms 'production' and 'conceptual' apraxia are also now used to indicate ideomotor and ideational apraxia according to the definitions below.

Limb-kinetic apraxia refers to the loss of fine motor dexterity that can be seen, for instance, with mild pyramidal lesions (such as after recovery from stroke). In spite of apparently good strength and co-ordination, the subject cannot manage tasks requiring fine motor control such as tying a shoelace or buttoning a shirt. As such, according to the above definition, this is not a 'true' apraxia but rather an artefact of the insensitivity of bedside tests of the motor system: in other words a primary motor deficit is only unmasked by tasks more demanding than routine tests of power and co-ordination.

Ideomotor (production) apraxia refers to the inability to execute the motor programme for a given task (the temporal and spatial organization of movement) in spite of adequate comprehension, as demonstrated, for instance, by the ability to describe the correct execution of the task (such as sharpening a pencil: 'you put the pointed end of the pencil into the hole then turn it') or to identify correctly a task when done by someone else. Patients with ideomotor apraxia also have problems performing meaningless (non-symbolic) gestures.

Ideational (conceptual) apraxia, in contrast, is a loss of knowledge of actions: there is an inability to either perform or recognize a given motor task. There is also an inability to match tools correctly to their actions, thus a subject may select a screwdriver to hammer a nail. Unlike patients with ideomotor apraxia, they do not show disorders of the spatial and temporal aspects of action and thus their tool use, although incorrect, is fluent.

To screen for apraxia, patients should be asked to perform skilled motor tasks to verbal instruction or to imitation including both meaningful and meaningless gestures. If deficits are uncovered, tests such as correctly identifying mimes performed by the examiner and matching tools to functions should be given. Subtle disorders of praxis may be evident only with low frequency tasks (such as using a vegetable peeler or a pencil sharpener as opposed to a knife or a hairbrush). When asked to pantomime an action (such as hair combing or brushing teeth), 'body part as tool' errors are often cited as evidence for apraxia: the patient uses his or her hand as the tool (for example rubbing an extended index finger over the teeth as a toothbrush). It is, however, not uncommon for normal subjects to make these 'body part as tool' errors when asked to perform such tasks, hence it is essential when this type of error is committed to draw it to the subject's attention and reinstruct them accordingly. Normal subjects are able to correct these errors, whilst those with apraxia cannot.

In terms of the neural substrate for production (ideomotor) apraxia, the overwhelming majority of cases follow damage to the left (dominant) hemisphere. More specifically, there is evidence that a motor system incorporating the superior parietal lobule (Brodmann areas 5 and 7) and the premotor area of the left frontal lobe is particularly critical to the temporal and spatial organization of motor programmes. Conceptual apraxia is also indicative of left hemisphere dysfunction in most cases, although whether a more specific site can be identified is contentious. It is also important where a conceptual apraxia is suspected to ensure it is not just one manifestation of a more generalized breakdown of semantic knowledge (see Memory, above)

Buccofacial apraxia represents a specific form of apraxia in which patients are unable to perform tasks such as licking lips or blowing out matches to command. It is particularly associated with non-fluent aphasia, presumably as the motor programming of articulation and non-linguistic buccofacial movements share a common pathway.

Personality and behavioural change

So far, disorders of higher mental function have been considered in quite discrete terms, both in the sense of the cognitive deficit and the cerebral location. Alterations in complex behaviour, personality, and social comportment, however, cannot be so simply defined, but are broadly associated with frontal or anterior temporal lobe pathology. The key to identifying such disorders is the presence of a sustained change from a previous state (thus differing from a lifelong eccentric personality) which cannot be explained by a primary psychiatric diagnosis. The only reliable way to confirm such changes is by taking a separate history from a spouse or other close personal acquaintance with knowledge of the patient's premorbid personality.

Prefrontal syndromes

The prefrontal cortex comprises that part of the frontal lobe rostral to the premotor area; it is classified as heteromodal association cortex and receives extensive inputs from unimodal association areas posterior to the central sulcus. The frontal lobes also have loop projections running to the basal ganglia, then the thalamus, and back to the frontal lobes. Thus lesions along this loop (as seen in conditions such as Huntington's disease or progressive supranuclear palsy) may also share deficits in common with primary frontal lobe disorders. Anatomically, the prefrontal cortex can be divided into dorsolateral, orbital, and medial surfaces; although in many cases damage will not be restricted to just one of these regions, they provide a useful framework for considering prefrontal functions. Broadly, lesions to the dorsolateral surface are responsible for the frontal 'dysexecutive' syndrome, to the orbital surface for the classic frontal behavioural syndrome, and to the medial surface (anterior cingulate) for a profound amotivational state.

The dysexecutive syndrome

The term 'executive' refers to aspects of higher-order brain function, such as problem solving, reasoning, and mental abstraction, which rely upon the dorsolateral prefrontal lobes. It is also associated with impulsivity, susceptibility to distraction, and failure to persevere with the task at hand. Various methods are available to

measure these phenomena although no single test offers foolproof sensitivity in this domain, so one should apply as many as ossible if the index of suspicion is high.

The combination of letter- and category-based verbal fluency provides much useful information. In letter fluency, the patient is asked to generate as many words as they can think of beginning with a given letter in 1 min. They are instructed not to use proper nouns and not to just change the endings to create new exemplars ('go, goes, going' etc.). Neuropsychologists typically use the letters F, A, and S for this test, so it is best to choose another letter if it is likely that patients are also going to have a formal neuropsychological assessment. In category fluency, patients are asked to produce as many exemplars as possible from a given category in 1 min. Normal subjects usually generate 15 or more words on letter fluency and do slightly better on the 'animal' category. Patients with executive deficits secondary to frontal (or the subcortical loop) pathology show an exaggeration of this relationship, doing poorly on category fluency but even worse on letter fluency (patients with semantic impairments related to temporal lobe diseases such as semantic dementia and Alzheimer's disease typically show the reverse pattern of relatively worse performance on category fluency).

The 'go–no go' test offers a way of assessing impulsivity: the patient is asked to tap the desk once if the examiner does so, but if the examiner taps twice, he should not tap at all. Patients with frontal pathology are often unable to stop themselves from tapping in both conditions. Failure to abstract meaning from proverbs ('What does “too many cooks spoil the broth” mean') is a common test but is influenced by background intellectual ability and is culture bound. The so-called 'cognitive estimates' test is also useful ('What is the height of the post-office tower in London?' or 'How fast does a racehorse gallop?'), as are 'differences and similarities' ('What's the difference between a child and a dwarf?' or 'In what way are a sculpture and a piece of music similar?'). Finally, the susceptibility to irrelevant stimuli mean that the tests of attention discussed above may also be impaired.

Orbitofrontal syndrome

The striking changes in behaviour seen in patients with prefrontal lesions relate particularly to orbital (or ventral) surface damage. Although devastating in their effects on social function, such lesions are notoriously difficult to detect using standard psychometric tests. Patients lack empathy and emotional warmth: for instance, if confronted with something as serious as the admission to hospital of their spouse, their primary concern may be that their mealtime routine will be disturbed. They are disinhibited and oblivious to social mores such that they may be overly familiar with strangers, disregard personal space, and make inappropriate (often of a sexual nature) comments or gestures. They often make rash and irresponsible decisions such as spending money above their means. They may develop stereotyped and ritualistic behaviours such as insisting on always taking a particular route when shopping or repetitively closing doors in the home: these behaviours can be so severe as to constitute a secondary obsessive–compulsive disorder syndrome.

A useful clue is often the presence of a change in eating behaviour. Patients may become fixated on one dish; often they develop a preference for sweet foods. A lack of normal satiety means that they may overeat, often with secondary weight gain.

Imitation and utilization behaviour are dramatic phenomena related to orbital frontal lobe damage. The patient with imitation behaviour unconsciously mimics the examiner's posture and mannerisms regardless of how absurd they are: raising an arm in the air, placing a leg on the desk, or sitting on the floor. Utilization behaviour is even more striking: patients will use any object placed in their grasp. The classic example is the patient offered multiple pairs of spectacles who attempts to wear them all, one on top of another.

Amotivational states

Medial frontal lesions are particularly associated with apathy. Patients lack spontaneity, they will not initiate conversation although can reply to specific questions. Likewise, if left to their own devices, they may not spontaneously move, preferring to sit in a chair staring blankly into space. This apathy has also been termed abulia in the past, in its most extreme form where the individual lies motionless with no speech the term akinetic-mutism has also been applied. The catatonic phenomenon of maintaining postures when the limbs are moved by the examiner may also be seen. Patients with depression also show marked apathy, although it is accompanied by both the biological features of depression (anorexia, diurnal variation etc.) and internal symptoms of mood disturbance (pessimism, suicidal thoughts, anhedonia etc.).

Temporal lobe syndromes

In addition to the cognitive deficits that can occur with temporal lobe lesions such as amnesia and loss of semantic knowledge, behavioural distrubances can also occur. The most severe, secondary to bilateral anterior temporal damage (including the amygdala), is the Klüver–Bucy syndrome, which comprises three characteristic features: placidity, even in threatening situations, indiscriminant hypersexuality, and oral exploration of objects. Other behaviours have been described in temporal lobe dysfunction, particularly, although not exclusively, in association with interictal temporal lobe epilepsy. These include preoccupation with religious or philosophical issues and a tendency to excessive writing.

Further reading

Baddeley AD (1999). *Essentials of human memory*. Psychology press, Hove.

Berrios GE, Hodges JR (2000). *Memory disorders in psychiatric practice*. Cambridge University Press, Cambridge.

Cummings JL (1995). Anatomic and behavioral aspects of frontal-subcortical circuits. *Annals of the New York Academy of Sciences* **769**, 1–13.

Driver J, Mattingley JB (1998). Parietal neglect and visual awareness. *Nature Neuroscience* **1**, 17–22.

Garrard P, Perry R, Hodges JR (1997). Disorders of semantic memory. [Editorial.] *Journal of Neurology, Neurosurgery and Psychiatry* **62**, 431–5.

Graham KS, Patterson K, Hodges JR (1999). Episodic memory: new insights from the study of semantic dementia. *Current Opinion in Neurobiology* **9**, 245–50.

Hodges JR (1994). *Cognitive assessment for clinicians*. Oxford University Press, Oxford.

Hodges JR, Spatt J, Patterson K (1999). 'What' and 'how': evidence for the dissociation of object knowledge and mechanical problem-solving skills in the human brain. *Proceedings of the National Academy of Sciences USA* **96**, 9444–8.

McCarthy RA, Warrington EK (1990). *Cognitive neuropsychology: a clinical introduction*. Academic Press, San Diego.

Mesulam MM (1998). From sensation to cognition. *Brain* **121**, 1013–52.

Patterson K, Lambon-Ralph MA (1999). Selective disorders of reading? *Current Opinion in Neurobiology* **9**, 235–9.

Rothi LJG, Heilman KM, eds (1997). *Apraxia: the neuropsychology of action*. Psychology Press, Hove.

Tulving E, Craik FM, eds (2000). *The Oxford handbook of memory*. Oxford University Press, New York.

Ungerleider LG, Haxby JV (1994). 'What' and 'where' in the human brain. *Current Opinion in Neurobiology* **4**, 157–65.

Walsh K, Darby D (1999). *Neuropsychology: a clinical approach*. Churchill Livingstone, Edinburgh.

24.9 Brainstem syndromes

David Bates

Most of the brainstem syndromes involving both the long tracts of the brainstem and the cranial nerve nuclei occur due to vertebrobasilar ischaemia but were originally described in relation to tumours and other non-vascular disorders. Vascular disorders by their very character often have a rostrocaudal and patchy localization rather than the simplified transverse localization that is usually demonstrated in diagrams. It may not always be possible to identify the cause of specific symptoms and signs in an individual patient and therefore the diagnosis of vascular disorders in the brainstem is more profitably identified by knowledge of brainstem anatomy.

The classic presentation of brainstem syndromes, including the long tracts and deficits of cranial nerve nuclei, commonly causes crossed cranial nerve and motor or sensory long tract deficits; the cranial nerve palsy is ipsilateral to the lesion and the long tract signs contralateral. It is important to assess the.extracranial vascular supply to the posterior circulation, especially to listen for bruits over the subclavian vessels and to record the pulse and blood pressure in both upper limbs, remembering that the vertebral arteries arise from the subclavian vessels. Apart from the crossed cranial nerve and long tract deficits, there may be ataxia, vertigo, the presence of an internuclear ophthalmoplegia and unreactive pupils, the symptoms of diplopia and oscillopsia, and the finding of nystagmus or ocular paresis.

The circulation to the brainstem is supplied by the vertebral arteries, which are the main arteries to the medulla, then the basilar artery, which supplies the pons and midbrain. The vertebral arteries are frequently asymmetrical and commonly give rise to the large posterior inferior cerebellar arteries shortly before they join to form the basilar. The vertebral arteries are susceptible to trauma within the cervical spine, but the most common lesions affecting the vertebral arteries are dissection, which is probably under-recognized, or thrombosis.

The basilar artery supplies branches which may be described as paramedian, supplying the area of the pons close to the mid-line, the short circumferential which supply the lateral two-thirds of the pons, the long circumferential which are the superior and anterior inferior cerebellar arteries, and several interpeduncular branches which arise at the bifurcation of the basilar artery and supply the sub-thalamic and high midbrain regions.

The brainstem syndromes

Thalamic syndrome

Sometimes spontaneously but commonly following a recognized hemiplegic and hemianaesthetic stroke, the patient develops altered sensation in a hemisensory distribution together with unpleasant dysaesthetic burning pain (thalamic pain). The pain may be worsened by stimulation and is associated with hemianaesthesia, sometimes proprioceptive loss, and some evidence of hemiparesis. Anatomically the lesion is usually in the ventroposteriolateral nucleus of the thalamus and is commonly caused either by a vascular event or by a tumour. The investigations required are imaging and therapy is with centrally acting analgesic agents.

Tectal deafness

There is a rare syndrome associated with damage at the level of the inferior colliculi, either due to neoplasia or vascular lesions resulting in bilateral deafness and sometimes associated difficulty in co-ordination, weakness, and vertigo. The condition must be differentiated from conductive bilateral hearing loss, cochlear disorders, bilateral eighth nerve lesions, and pure word deafness. The lesion can be identified by brain imaging.

Thalamic stroke syndrome

Lesions affecting the thalamus are commonly vascular and arise from infarction within the distribution of the posterior communicating artery, the basilar and the anterior and posterior choroidal arteries. There is usually hemiparesis with hemianopia, hemianaesthesia, and sometimes hemiataxia. There is often confusion, disorientation, and there may be language disturbance. On occasion there may be vertical-gaze ophthalmoplegia, loss of pupillary reflexes, and an inability to converge the eyes. There may also be memory impairment and on occasions visual perceptual disturbances are recorded.

Midbrain syndromes

Damage to areas of the midbrain is characterized by long tract signs contralateral to the lesion with defects of the third and fourth cranial nerves ipsilaterally. They can been seen with lesions in the brainstem or as the evolution of symptoms of rostrocaudal deterioration associated with supratentorial brain swelling (Fig. 1 and Fig. 2). They are characterized by ipsilateral third and fourth cranial nerve palsies together with contralateral hemiparesis, loss of vibration, propriception, and stereognosis, contralateral loss of pain and temperature, and an ipsilateral Horner's syndrome. Ataxia may occur and there can be eyelid ptosis, diplopia, supranuclear horizontal-gaze paresis, and an internuclear ophthalmoplegia. The association of an ipsilateral oculomotor palsy with a crossed hemiplegia due to a lesion at the base of the midbrain is called a Weber syndrome. Claude syndrome causes an ipsilateral oculomotor palsy with contralateral cerebellar ataxia and tremor and is due to a lesion in the tegmentum of the midbrain involving the red nucleus and the third nerve nucleus. Benedikt's syndrome also involves the tegmentum of the midbrain resulting in an oculomotor palsy with contralateral cerebellar ataxia, tremor, and corticospinal signs and can be regarded as a combination of Claude and Weber's syndrome. Nothnagel syndrome occurs with unilateral or bilateral involvement of the third nerve nucleus together with the superior cerebellar peduncles and causes bilateral ptosis, paralysis of gaze, and cerebellar ataxia.

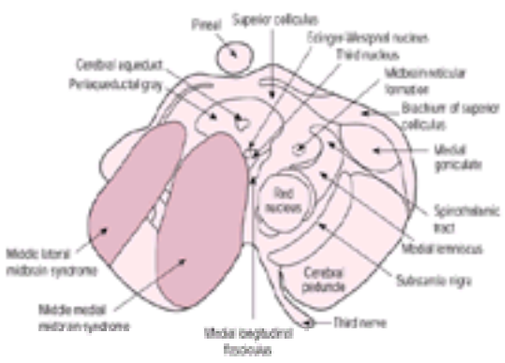

Fig. 1 Midbrain at the superior colliculus level, showing the medial and lateral territories involved with occlusive stroke in this region. (Reprinted with permission from DeArmond SI, Fusco MM, Dewey MM, 1976, *Structure of the human brain*, 2nd edn. Oxford University Press, New York.)

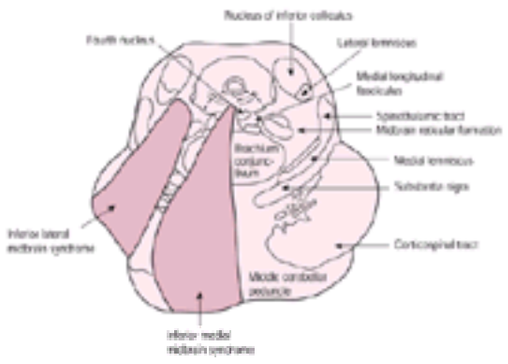

Fig. 2 Midbrain at the inferior colliculus level showing the medial and lateral territories involved with ischaemic stroke syndromes in this area. (Reprinted with permission from DeArmond SI, Fusco MM, Dewey MM, 1976, *Structure of the human brain*, 2nd edn. Oxford University Press, New York.)

Damage in the region of the dorsal midbrain results in the Parinaud syndrome in which there is paralysis of upward gaze due to damage to the supranuclear mechanisms for upward gaze, loss of accommodation, and fixed pupils. Although this may be seen with ischaemic lesions, it is more commonly seen with pineal tumours.

Pontine syndromes

Lesions in the pons and medulla are commonly identified as involving either the medial or the lateral aspect of the brainstem, depending upon whether the paramedian or short circumferential vessels from the basilar have been involved. In the pons the following three levels of damage can be detected and the basal syndrome can occur at any level.

Superior pontine syndrome

The medial superior pontine syndrome results in ipsilateral cerebellar ataxia, internuclear ophthalmoplegia, and palatal and pharyngeal myoclonus with contralateral paralysis of face, arm, and leg and sometimes loss of sensation contralaterally. The lateral superior syndrome causes ataxia of limbs and gait with dizziness, nausea, and vomiting; there is horizontal nystagmus, paresis of conjugate gaze towards the side of the lesion, loss of optokinetic nystagmus, and sometimes skew deviation of the eyes. There may also be an ipsilateral Horner's syndrome and there is contralateral loss of pain and thermal sensation on the face and limbs with impaired touch, vibration, and position sense (Fig. 3).

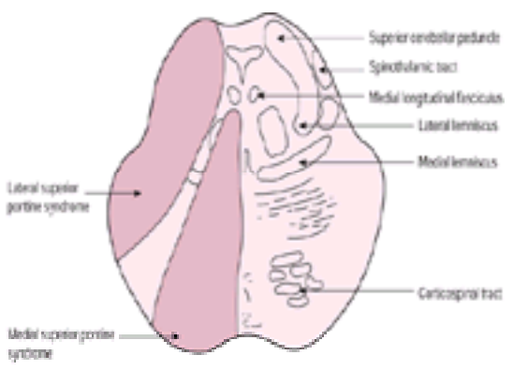

Fig. 3 Superior pontine level, showing the medial and lateral territories involved with occlusive stroke in this region. (Reprinted with permission from Adams RD, Victor M, 1993, *Principles of neurology*, 5th edn. McGraw-Hill, New York.)

The mid-pontine syndrome

The medial, mid-pontine syndrome causes ipsilateral ataxia of the limbs and gait with contralateral paralysis of the face, arm, and leg, deviation of the eyes away from the lesion, and variably impaired sensation contralaterally. The lateral syndrome at this level causes ataxia of the limbs on the side of the lesion together with paralysis of the muscles of mastication and impaired sensation over the face on the same side due to damage to the fifth nerve (Fig. 4 and Fig. 5).

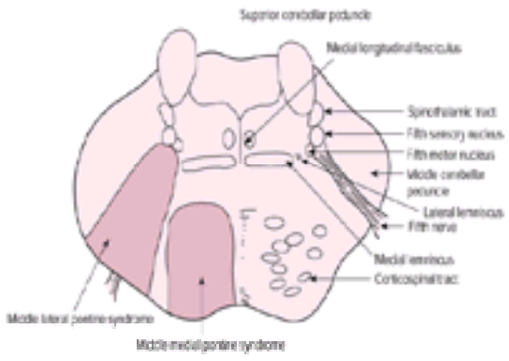

Fig. 4 Mid-pontine level, showing the medial and lateral territories involved with ischaemic stroke syndromes in this locality. (Reprinted with permission from Adams RD, Victor M, 1993, *Principles of neurology*, 5th edn. McGraw-Hill, New York.)

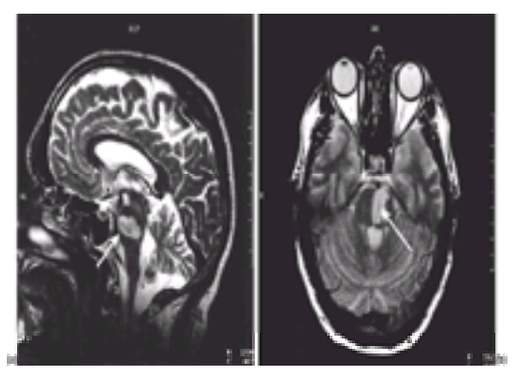

Fig. 5 MRI scan of a mid-pontine infarction.

The inferior pontine syndrome

The medial syndrome causes paralysis of conjugate gaze to the side of the lesion, nystagmus, ataxia of limbs on the same side, and double vision on gaze to that side. Contralaterally there is paralysis of the face, arm, and leg with impaired touch and proprioception over the opposite side of the body. The lateral syndrome involves ipsilateral, horizontal, and vertical nystagmus with vertigo and nausea, ipsilateral facial paralysis, paralysis of conjugate gaze to the side of the lesion, deafness, tinnitus, and ataxia on the side of the lesion with impaired sensation of the face on that side. On the opposite side there is impaired sensation over half of the body (Fig. 6).

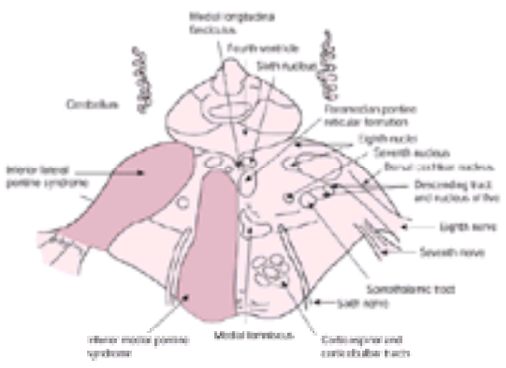

Fig. 6 Inferior pons at the level of the sixth nerve nucleus, showing the medial and lateral territories involved with occlusive stroke in this area. (Reprinted with permission from Adams RD, Victor M, 1993, *Principles of neurology*, 5th edn. McGraw-Hill, New York.)

Basal pontine syndrome (locked in syndrome)

Bilateral lesions of the paramedian vessels from the basilar, commonly seen in patients with hypertension, result in infarction of the basis pontis causing quadriplegia with loss of the ability to speak. The ascending reticular activating system is intact and consciousness is therefore preserved. Vertical eye movements and eye closure are all that are possible and under voluntary control in the 'locked-in syndrome'.

Pseudobulbar palsy

Bilateral lesions of the long descending tracts in the brainstem can result in pseudobulbar palsy, although this condition is more commonly seen with lesions higher in the cerebrum. The symptoms are those of spastic dysarthria, dysphagia, bilateral facial weakness with quadriparesis, and emotional lability.

Medullary syndromes

The medial medullary syndrome may occur with occlusion of the vertebral artery or a branch of the lower basilar artery; it causes paralysis and atrophy of the tongue on the side of the lesion with contralateral paralysis of the arm and leg but sparing the face and impaired tactile proprioceptive sensation over the contralateral half of the body. The lateral medullary syndrome occurs most commonly with dissection or occlusion of the vertebral artery, resulting in ischaemia into the posterior, inferior cerebellar artery; this causes pain, numbness, impaired sensation of the ipsilateral half of the face with ataxia of limbs on that side, the symptoms of vertigo and nausea, double vision, and oscillopsia, and the signs of nystagmus. There is an ipsilateral Horner's syndrome, often dysphagia with paralysis of the vocal cord ipsilaterally, and loss of sensation on the arm, trunk, and leg. There is contralateral impaired pain and thermal sensation over half the body and possibly the face (Fig. 7).

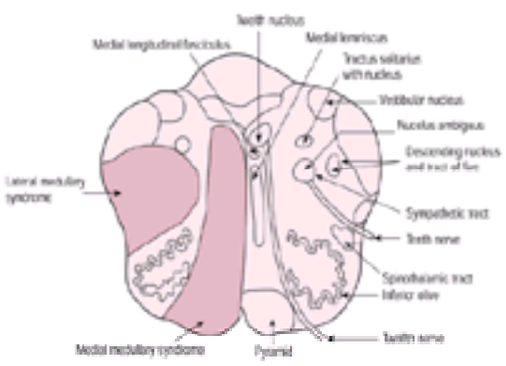

Fig. 7 Cross-section of medulla at the level of the inferior olivary complex, showing the medial and the more common lateral territory involved with ischaemic stroke in this brainstem site. (Reprinted with permission from Adams RD, Victor M, 1993, *Principles of neurology*, 5th edn. McGraw-Hill, New York.)

A syndrome involving ipsilateral seventh and sixth nerve palsies with a contralateral hemiplegia is called the Millard–Gubler syndrome; the involvement of the tenth cranial nerve causing paralysis of the soft palate and vocal cord with contralateral hemianaesthesia is termed the Avellis syndrome and is due to a lesion in the tegmentum of the medulla. The lateral medullary syndrome is eponymously called the Wallenberg syndrome.

Investigations and treatment

The clinical identification of lesions lying within the brainstem by a combination of cranial nerve and long tract signs, though important is only the beginning of diagnosis. Such syndromes commonly occur with ischaemic lesions within the brainstem but can also be seen with neoplasia, demyelination, and infective and hamartomatous lesions. The identification of a brainstem syndrome makes imaging, ideally with MRI rather than CT, obligatory and only then, and possibly following other investigations to identify systemic abnormality or cerebrospinal fluid changes, can appropriate therapy be introduced. Vascular lesions within the brainstem often carry a remarkably good prognosis, but if the syndrome appears to be evolving, the possibility of anticoagulation must be considered. In those lesions in which damage affects the medulla it may be important to protect the airway and avoid aspiration during the early phases of the illness.

Further reading

Adams RD, Victor M (1989). *Principles of neurology*, 4th edn. McGraw-Hill, New York.

Caplan LR (1988). Vertebrobasilar system syndromes. In: Vinken PJ, Bruyn GW, Klawans HL, eds. *Handbook of clinical neurology*. Elsevier, Amsterdam.

24.10 Subcortical structures—the cerebellum, thalamus, and basal ganglia

N. P. Quinn

The cerebellum

Structure and function

The cerebellum occupies the greater part of the posterior fossa, reaching from the tentorium rostrally to the foramen magnum caudally, and lying dorsal to the lower pons and medulla, from which it is separated by the fourth ventricle. Its blood supply is derived from posterior circulation via the superior, anterior inferior, and posterior inferior cerebellar arteries. The cerebellum can be divided into cortex, intrinsic nuclei, and interposed white matter (medullary substance). The cortex comprises three cell layers—from the surface inwards these are the molecular layer (3), the Purkinje cell layer (2), and the granular cell layer (1). The only output cells of the cortex are the Purkinje cells. Inputs to the cerebellar cortex comprise either climbing or mossy fibres. The former synapse directly with Purkinje cells. The latter synapse with granule cells in layer 3, whose axons ascend to layer 1 where they form parallel fibres which synapse with Purkinje cell dendrites ascending from layer 2.

The cerebellum can also be divided into: archicerebellum (flocculonodular lobe), with largely vestibular inputs; palaeocerebellum (anterior lobe), with largely spinal cord inputs; and neocerebellum (posterior, largest lobe), with largely pontine inputs from cerebral cortex. Yet another way of dividing the cerebellum is into functional 'units' as follows: the vermal zone, comprising the fastigial nuclei and the midline unpaired portion of cerebellum, projects to vestibular nuclei and controls mainly axial posture, tone and balance, and locomotion; the paravermal zones, comprising the globose and emboliform nuclei and corresponding cerebellar cortex, project via the contralateral red nucleus and other nuclei of the reticular formation to influence ipsilateral limb tone; and the lateral zones, comprising the dentate nuclei and lateral cerebellar cortex, project via contralateral thalamus on to motor cortex to effect ipsilateral motor co-ordination.

The integrating function of cerebellum is evident from the fact that afferent fibres (Fig. 1) heavily outnumber efferent ones by about 40 to 1. Connections travel in three cerebellar peduncles, the lower two mainly afferent and the upper one efferent: the inferior cerebellar peduncle (restiform body) carries spino-, vestibulo-, and olivocerebellar fibres and input from other medullary ('precerebellar') nuclei; the middle cerebellar peduncle (brachium pontis) carries major afferent fibres from the pontine nuclei responsible for relaying and integrating a large input from all areas of the cerebral cortex; the superior cerebellar peduncle (brachium conjunctivum) contains a few afferent spinocerebellar fibres, but most of its bulk comprises cerebellar efferent fibres which originate in the intrinsic cerebellar nuclei and stream up to the contralateral red nucleus, or through it to the thalamus, to influence heavily thalamocortical, and then corticospinal, input. Other cerebellar efferents go to vestibular nuclei and to nuclei in the pontine and medullary reticular formation. These brainstem efferents provide access to the rubrospinal, vestibulospinal, and reticulospinal tracts.

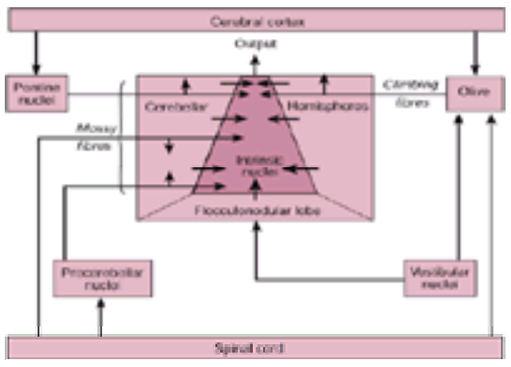

Fig. 1 A simplified diagram showing the principal afferents to cerebellar cortex and to intrinsic cerebellar nuclei. Both mossy (left) and climbing (right) fibre inputs project to both cortex and intrinsic nuclei.

Clinical aspects

There are several theories on the main functions of the cerebellum. Thus it may work as a timing device to control the duration and latency of muscle activities, it may act as a learning device to lay down circuitry for repeating previously performed movements, and it may act as a co-ordinator to harmonize and correctly scale the contribution of several brain areas to an intended movement.

The symptoms and signs resulting from lesions of the cerebellum or its pathways in humans have been based on observations in trauma (particularly gunshot wounds), tumour, stroke, and degenerative and demyelinating diseases.

Midline vermal lesions cause truncal ataxia, often in the absence of limb ataxia. The gait is wide based and particularly precarious on turning or on heel–toe walking. Patients may be unsteady for many reasons, only one of which is ataxia, so the former term is preferable where any doubt exists. Unilateral cerebellar hemispheric lesions cause deviation or falling to the ipsilateral side. Unlike a sensory ataxia, cerebellar ataxia is not made worse by shutting the eyes. Generally, ataxic patients have more problem going down, and those with weakness going up, stairs.

Disease of cerebellar hemispheres or outflow tracts often causes limb ataxia, which is in fact an amalgam of several components. First, there may be dysmetria (misreaching, or past-pointing) evident in the arms on the finger–nose test or in the legs when the heel is first brought to the opposite kneecap. Second, there is the breakdown in force, rate, and rhythm known as dysdiadochokinesia. This can best be sought by asking patients to tap gently, regularly, and rapidly on your hand or a table with their fingers. This breakdown of smooth repetitive movements can even be detected by feel or by sound ('listening to the cerebellum'). The third element is intention tremor. Many individuals claimed to have intention tremor do not actually have it, the principal error being to use this term to describe a tremor that simply appears or worsens terminally. Thus, many postural tremors are positionally dependent, and some are only seen when the hands are in a given posture, particularly either outstretched or held in front of the nose. Other non-cerebellar tremors may be present, or appear only during action (action or kinetic tremor). Only if additional signs of cerebellar dysfunction considered above are also present, is it reasonable to use the term intention tremor. Such tremor should augment throughout a movement from inception to completion. A particular form of tremor may be produced by lesions strategically placed in a small area between cerebellum, midbrain, and subthalamus. This tremor has been variously called Holmes', rubral, midbrain, or peduncular tremor, and combines a tremor at rest with a tremor on posture and an intention tremor on movement. Cerebellar lesions may also cause the 'rebound phenomenon' resulting from impaired damping of limbs when suddenly a load is removed or a displacement applied. Finally, any judgement concerning possible limb ataxia can only be made after taking into account any weakness, sensory loss,

akinesia, or apraxia that may also be present.

Cerebellar dysarthria may often simply manifest as slurred speech, as if intoxicated. However, in addition some patients may have either scanning or explosive speech, due to an inability to modulate its rate, rhythm, and force appropriately. Dysarthria is usually present with lesions of the vermis, whole cerebellum, or its connections, but may be absent if one lateral hemisphere alone is involved.

Eye movements are frequently abnormal in disease of the cerebellum or its connections. The following may be seen: gaze-evoked, rebound, downbeat, or positional nystagmus, dysmetric voluntary saccades and jerky pursuit, square-wave jerks (macrosaccadic oscillations), impaired vestibulo-ocular reflex suppression, and skew deviation. The presence of diplopia usually implies additional pathology outside the cerebellum proper.

The thalamus

Structure and function

The two thalami sit at the head of the brainstem, their medial borders largely separated by the third ventricle, but often partially fused as the massa intermedia. Their blood supply derives from the posterior circulation via the posterior cerebral arteries and perforators from the terminal part of the basilar artery. They constitute the largest nuclear mass in the diencephalon (the others being the hypothalamus and subthalamus), and occupy a strategic position both anatomically and functionally.

The structure of the thalamus, already complex, is further confused by the existence of different nomenclatures (the one used here is that of Walker). Broadly speaking, there are three nuclear groups (anterior, medial, and lateral). The lateral group is divided into the lateral and ventral masses, each of which contain a number of nuclei. The ventral lateral cell mass is the main region where somatosensory afferents terminate.

The thalamus receives inputs from cerebral cortex, sensory tracts, basal ganglia, and cerebellum. Almost all of its output is to the cerebral cortex, either in the form of reciprocal circuits or of more complex loops (see later) from cortex through other subcortical structures to thalamus and back to cortex again, but there is a small output to striatum.

Thalamic afferents

Somatic and visceral afferents

Somatic and visceral afferents from the body pass via the medial lemniscus and spinothalamic tract into the ventral posterolateral nucleus caudalis, where caudal body parts are represented laterally and rostral parts medially. Inputs from the face (via trigeminothalamic tracts) pass even more medially into the ventral posteromedial nucleus. Somatotopic representation is maintained through the connections to the parietal lobe in the form of the sensory homunculus with legs medially, arms high, and face low over the convexity. Taste afferents feed into ventral posteromedial nucleus parvocellularis, and hearing and vision into the medial and lateral geniculate bodies, respectively.

Basal ganglia input

The medial globus pallidus projects to the centromedian nucleus, to ventral anterior nucleus parvocellularis, and to ventral lateral nucleus oralis and medialis, and its homologue the substantia nigra pars reticulata to the mediodorsal nucleus and ventral anterior nucleus magnocellularis.

Cerebellar input

Afferents from intrinsic cerebellar nuclei ascend to the ventral lateral nucleus caudalis, to the ventral posterolateral nucleus oralis, and to the adjacent zone x.

Thalamic efferents

All the thalamic nuclei project to the cerebral cortex with the exception of important outputs from the intralaminar centromedian-parafascicular nuclear complex to striatum.

Clinical aspects

From the above it is clear that, depending on the nuclei involved, thalamic lesions might influence either sensation or motor function or sometimes both. Most commonly an infarct or haemorrhage (10 to 15 per cent of all intracerebral haemorrhages) causes contralateral sensory loss or impairment. A small lacunar infarct in the ventral posterolateral nucleus may give rise to a pure sensory stroke, sometimes sparing the face. A larger lesion may cause the thalamic syndrome of Dejerine and Roussy, in which an initial mild and transient hemiplegia is accompanied by persisting superficial and deep sensory impairment, mild hemiataxia, and astereognosis. These are commonly accompanied by choreoathetoid movements and severe, persistent, paroxysmal, often intolerable pains on the affected side. When mild, the movements may be pseudoathetotic due to deafferentation; when severe, they suggest that the lesion may extend beyond the thalamus to involve basal ganglia connections. Other movement disorders described after thalamic strokes are myoclonus, asterixis, tremor, dystonia, and a delayed-onset syndrome of choreiform and dystonic movements associated with slow rhythmic jerks at 2 to 3 Hz involving the contralateral arm. A significant, persisting hemiplegia implies either a large thalamic lesion also involving internal capsule or the possibility that the stroke is primarily capsular and not thalamic. A particular form of subcortical aphasia has been described in thalamic lesions.

Finally, surgical lesions have been stereotactically placed in the ventral lateral nucleus caudalis (also known as the ventral intermediate nucleus—Vim in Hassler's nomenclature) to relieve tremor in Parkinson's disease and benign essential tremor, and also rigidity (but not akinesia) in the former. Chronic electrical stimulation of the same area is also effective. However, the thalamus as a target for surgery in Parkinson's disease has now largely been superseded by the subthalamic nucleus (see later) because inactivation of the latter target also improves the other features of parkinsonism.

The basal ganglia

Structure and function

There is no uniform agreement on how many of the subcortical nuclei one should include under the terms basal ganglia and extrapyramidal motor system, but all definitions at least include the neostriatum (caudate nucleus and putamen, often together called simply the striatum) and the palaeostriatum (the lateral and medial globus pallidus with the latter's homologue, the substantia nigra pars reticulata). The term corpus striatum refers to neostriatum plus palaeostriatum, and lentiform nucleus to putamen plus globus pallidus. The substantia nigra pars compacta and the subthalamic nucleus and the limbic amygdaloid complex of archistriatum should also be considered part of the basal ganglia. The claustrum, substantia innominata, red nucleus, pedunculopontine nucleus, and even the thalamus are considered in some classifications to be part of the basal ganglia, but will not be dealt with here under that heading.

The putamen lies lateral to the thalamus, separated from it (and from most of the caudate nucleus, except anteriorally) by the internal capsule. The caudate nucleus, whose head lies anterodorsomedial to the putamen, describes most of a circle as it follows, and progressively tapers with, the lateral ventricles through its body posteriorally, its tail swinging forward until its anteriorly pointing tip terminates in the amygdaloid nucleus. The pallidum lies medial to the putamen but still lateral to the internal capsule, and is divided into lateral and medial pallidal segments. The substantia nigra lies in the midbrain, transversely above the cerebral peduncles. Its pars reticulata, the termination of the striatonigral pathway, is homologous with medial globus pallidus, and its pars compacta contains the dopaminergic neurones which form the nigrostriatal pathway. Below the thalamus, medial to the internal capsule and rostral to the midbrain, is the subthalamic nucleus.

Most of the caudate, putamen, and globus pallidus derive their arterial supply from anterior circulation via the lateral lenticulostriate arteries and branches of the anterior choroidal and anterior cerebral arteries. Like the thalamus, the subthalamic region, and also the substantia nigra, are supplied by posterior circulation.

The basal ganglia and their (inter-) connections, rich in neurotransmitters, have been extensively studied. The 'striopallidal complex' receives a wide variety of inputs

from cerebral cortex. Its principal output is to the thalamus, which in turn projects back to the cortex to complete a basal ganglia–thalamocortical circuit. However, it is important to note the existence of additional output to the brainstem in the pallidotegmental tract which terminates in the pedunculopontine nucleus. This structure is believed to play an important role in the control of balance and locomotion, and in the maintenance of rigidity. The caudate and putamen are the afferent, and the globus pallidus and substantia nigra pars reticulata the efferent, parts of the striopallidal complex. There is additional dopaminergic input from substantia nigra pars compacta and the adjacent ventral tegmental area in the midbrain which modulates striatal activity. A highly simplified schema concentrating on the motor circuit is presented in Fig. 2.

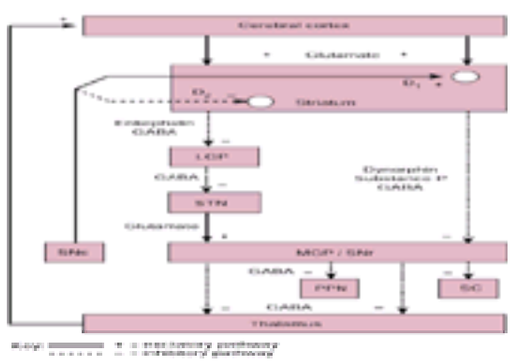

Fig. 2 A simplified schematic diagram showing the principal connections of the basal ganglia. Excitatory synapses are indicated by +, inhibitory ones by –. D [1] signifies dopamine D^1, and D^2 dopamine D^2, receptors. LGP and MGP, lateral and medial globus pallidus; STN, subthalamic nucleus; SNr and SNc, substantia nigra pars reticulata and pars compacta; PPN, pedunculopontine nucleus; SC, superior colliculus.

The massive cortical inputs into the striatum are largely excitatory, using glutamate as a neurotransmitter. Other inputs come from the intralaminar thalamic nuclei, amygdala, and dorsal nucleus of the raphe. In the nigrostriatal pathway from the substantia nigra pars compacta, dopamine preferentially stimulates dopamine D_1 receptors to activate neurones of the direct pathway to the medial globus pallidus which contain dynorphin, substance P, and GABA, and are therefore inhibitory. Dopamine D_2 receptor stimulation preferentially inhibits the first neurones of the indirect pathway to the lateral globus pallidus which contain enkephalin and GABA. These neurones inhibit subthalamic neurones which in turn, using glutamate, excite cells in the medial globus pallidus and substantia nigra pars reticulata. These in their turn use GABA to inhibit thalamic neurones, which finally complete the loop with an excitatory pathway back to the cortex.

This model can be used to predict the functional consequences of over- or underactivity of individual parts, either in human disease states or in experimental animals. In the latter, 2-deoxyglucose autoradiographic studies can be used to confirm such predictions elegantly and validate the model. Thus, the consequences of cell loss in the substantia nigra pars compacta that occur in Parkinson's disease would be as follows: along the direct pathway there is impaired stimulation of the striatal cells that normally inhibit the neurones of the medial globus pallidus/substantia nigra pars reticulata, so the latter are overactive. Along the indirect pathway, there is impaired inhibition (hence overactivity) of the neurones that inhibit the lateral globus pallidus, which is therefore underactive. However, this leads to less inhibition (hence overactivity) of the subthalamic nucleus, which increases excitatory input to the medial globus pallidus/substantia nigra pars reticulata (already overactive via the direct pathway). This overactivity in turn inhibits thalamic, and then cortical, activity. The model would predict that making a lesion in the overactive subthalamic nucleus or medial globus pallidus might relieve parkinsonism, and indeed this has been demonstrated in MPTP-treated primates and patients with Parkinson's disease, respectively.

The model can also be used to understand hyperkinetic movement disorders. Hemiballism (severe unilateral proximal chorea) is classically caused by a destructive lesion in, or close to, the subthalamic nucleus. In this instance an underactive subthalamic nucleus would release the thalamus and hence the cortex from inhibition by the medial globus pallidus/substantia nigra pars reticulata, and again this sequence has been confirmed by 2-deoxyglucose experiments in animals.

All that has been mentioned so far concerns motor function, traditionally equated with the function of the basal ganglia. However, as the complexity and diversity of basal ganglia anatomy, circuitry, and function has become apparent, the concept of multiple basal ganglia–thalamocortical circuits has developed. Although some overlapping of cortical input to the striatum does indeed occur, rather than cortical inputs being simply funnelled through the circuits, thereafter the loops are increasingly non-overlapping, allowing more separate and independent processing throughout the basal ganglia. Five such circuits have been proposed: (i) the motor and (ii) the oculomotor, involved in sensorimotor functions of the body and eyes; (iii) the dorsolateral prefrontal and (iv) orbitofrontal, involved in cognitive aspects of behaviour; and (v) the anterior cingulate circuit, related to limbic functions. A simplified representation of these loops is given in Fig. 3.

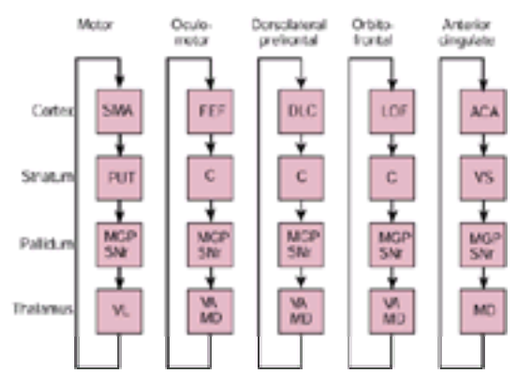

Fig. 3 A highly simplified diagram of proposed basal ganglia–thalamocortical loops (after Alexander *et al*., 1986, with permission). Same abbreviations as for Fig. 2; SMA, supplementary motor area; FEF, frontal eye fields; DLC, dorsolateral prefrontal cortex; LOF, lateral orbitofrontal area; ACA, anterior cingulate area; PUT, putamen; C, caudate; VS, ventral striatum; VL, ventrolateral thalamic nucleus; VA, ventral anterior nucleus; MD, mediodorsal nucleus. Segregated parts of the striatum, pallidum, and thalamus convey largely separate loops.

The last decade has also seen considerable advances in our knowledge of striatal anatomy and neurochemistry. Morphological techniques have long demonstrated a variety of striatal neuronal types. These are either projection neurones (spiny medium type I, the vast majority, and large type II), or intrinsic neurones (aspiny types I to III). However, the striatum as a whole seemed rather amorphous until neurochemical markers gave a new perspective. Thus, the medium type I spiny projection neurones are the major targets of nigral dopaminergic transmission, and make synaptic contact with large aspiny cholinergic interneurones. The former cells are selectively lost in Huntington's disease, resulting in a loss of their inhibitory transmitter GABA, together with colocalized metenkephalin or substance P. In contrast, type I aspiny interneurones containing neuropeptide Y and somatostatin, and type II aspiny interneurones containing acetylcholine, are largely spared in Huntington's disease.

It is now recognized that striatal neurones are also organized into a mosaic pattern comprising patches or striosomes with high levels of μ-opiate receptors and low levels of acetylcholinesterase, suspended in a matrix of cells containing high levels of aceytlcholinesterase, somatostatin, and calbindin. Patch and matrix receive different inputs from the midbrain, thalamus, and cortex. In particular, deeper levels of prefrontal or limbic cortex tend to project to striosomes and more superficial layers of sensorimotor cortex to matrix. Patches mainly project to substantia nigra pars compacta, whereas matrix neurones may take either the direct or indirect route to the medial globus pallidus/substantia nigra pars reticulata.

Clinical aspects

On the basis of the above evidence, we can no longer assume that basal ganglia pathology produces only motor symptoms and signs. Thus, in patients with

Parkinson's disease, Huntington's disease, and progressive supranuclear palsy—all of which principally (but not exclusively) involve basal ganglia— affective disorder, 'subcortical dementia', or 'frontal lobe deficits' may be seen.

Nevertheless, the most striking clinical features of basal ganglia disease remain those in the motor sphere, comprising tremor, rigidity, akinesia, and postural abnormality as evidenced by Parkinson's disease, and hyperkinetic movement disorders such as chorea and dystonia seen, for example, in Huntington's disease and in subjects with Parkinson's disease chronically treated with levodopa preparations.

The classic tremor of Parkinson's disease is slow (4 to 6 Hz), pill-rolling, and disappears and diminishes on movement, to reappear once a new posture has been adopted. In animal studies a nigral lesion seems necessary, but not sufficient, to cause this tremor.

Akinesia is a symptom complex comprising slowness of movement (bradykinesia), poverty of movement, progressive diminution and fatigue of rapid alternating movements, and difficulty in initiating and sequencing movements and in accomplishing simultaneous motor acts. Since changes in neuronal discharge relating to movement seem to occur later in the basal ganglia than in the motor cortex, it has been proposed that the basal ganglia are more concerned with using information from a previous movement to set up the premotor areas to select the correct parameters for running subsequent motor programmes. In Parkinson's disease levodopa strikingly improves akinesia. Lesions or high-frequency deep brain stimulation (which functionally causes inhibition) in the thalamus does not relieve akinesia in humans, but lesions or deep brain stimulation in the internal pallidum or subthalamic nucleus do so.

Rigidity almost always accompanies akinesia. Resistance to passive movement is broadly similar in flexion and extension. It is described as lead-pipe or, if there is superimposed tremor (visible or invisible), as cogwheeling. Abnormalities of the tonic stretch reflex are felt to contribute to its pathophysiology.

As well as akinesia, both tremor and rigidity respond to levodopa. Unlike akinesia, both also respond to thalamotomy, but the lesion needs to be larger for rigidity than for tremor.

Postural instability is another feature of parkinsonism. This is in part due to impairment of anticipatory responses and postural adjustments associated with movement, and problems in controlling body sway. Unlike the above features, this may often be levodopa resistant.

Of the hyperkinetic movement disorders, chorea (in the case of Huntington's disease) and dystonia (when secondary to discernible brain pathology) can be related to basal ganglia disease. Although chorea can be produced by making lesions in the subthalamic nucleus in intact monkeys, or by chronic dopaminergic treatment of primates with a lesioned nigrostriatal tract, it is not usually possible to produce spontaneous chorea in animals solely by making a caudate lesion analogous to that in Huntington's disease. Similarly, although dystonia may be seen in humans after lesions of the putamen (principally), caudate, or thalamus, there is again no good animal model, although it can be induced by dopaminergic drugs in MPTP-treated primates.

Further reading

Albin RL, Young AB, Penney JB (1989). The functional anatomy of basal ganglia disorders. *Trends in Neurosciences* **12**, 366–75. [An excellent synthesis of anatomical and functional aspects of the basal ganglia.]

Alexander GE, DeLong MR, Strick PL (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience* **9**, 357–81.

Carpenter MB (1991). *Core text of neuroanatomy*, 4th edn. Williams & Wilkins, Baltimore. [The best medium-sized all-round textbook of neuroanatomy.]

Crossman AR (1990). A hypothesis on pathophysiological mechanisms that underlie levodopa - or dopamine agonist-induced dyskinesia in Parkinson's disease: Implications for future strategies in treatment. *Movement Disorders* **5**, 100–8.

Gerfen CR (1992). The neostriatal mosaic: multiple levels of compartmental organisation. *Trends in Neurosciences* **15**, 33–9.

Graybiel AM (1989). Dopaminergic and cholinergic systems in the striatum. In: Crossman A, Sambrook MA, eds. *Neural mechanisms in disorders of movement*, pp 3–15. Libbey, London.

Hassler R (1959). Anatomy of the thalamus. In: Schaltenbrand G, Bailey P, eds. *Introduction to stereotaxis with an atlas of the human brain*, pp. 230–90. G. Thieme, Stuttgart.

Krack P *et al*. (2000). Thalamic, pallidal, or subthalamic surgery for Parkinson's disease? *Journal of Neurology* **247** (suppl. 2:II), 122–34.

Lehericy S, *et al*. (2001). Clinical characteristics and topography of lesions in movement disorders due to thalamic lesions. *Neurology* **57**, 1055–66.

Lera G *et al*. (2000). A combined pattern of movement disorders resulting from posterolateral thalamic lesions of a vascular nature: a syndrome with clinico-radiologic correlation. *Movement Disorders* **15**, 120–6.

Limousin P *et al*. (1998). Electrical stimulation of the subthalamic nucleus in advanced Parkinson's disease. *New England Journal of Medicine* **339**, 1105–11.

Marsden CD (1990). Neurophysiology. In: Stern GM, ed. *Parkinson's disease*, pp 57–98. Chapman & Hall, London. [A clear overview of the physiological mechanisms underlying the clinical features of parkinsonism.]

Marsden CD, Obeso JA (1994). The functions of the basal ganglia and the paradox of stereotaxic surgery in Parkinson's disease. *Brain* **117**, 877–97.

Rothwell JC (1994). *Control of human voluntary movement*, 2nd edn. Croom Helm, London. [A short textbook on the physiology of human motor disorders.]

Walker AE (1938). *The primate thalamus*. University of Chicago Press.

24.11 Visual pathways

Christopher Kennard

Introduction

Diagnosis of disturbances of the visual pathways requires both a thorough knowledge of their anatomy and physiology, as well as the ability to carry out a thorough neuro-ophthalmological examination. It is the examination which should enable the character and extent of the visual disturbance to be documented as well as the topographic localization of the lesion, so that the relevant investigative techniques, such as imaging, can be appropriately requested.

Clinical evaluation of visual function

Examination of visual function initially requires an accurate assessment of the visual acuity. Acuity should be tested separately in each eye using the Snellen or some other optotype chart, which contains rows of letters of diminishing size. If an impairment (> 6/6) is noted, the patient should be allowed to wear glasses or alternatively to view the chart through a pinhole which eliminates any significant refractive error or optic media distortion. If the acuity does not improve, it is necessary to try and distinguish media opacities and retinal abnormalities from optic nerve dysfunction using the swinging flashlight test. In a darkened room each eye is alternatively stimulated with a bright light, which is moved rhythmically from one eye to the other. A dilatation of the pupil of the defective eye when the light is swung from the good eye on to it is termed a 'relative afferent pupillary defect', and signifies optic nerve dysfunction. Another good indicator of an optic nerve disturbance is a defect of colour vision, which may be tested using one of several available booklets of colour plates such as the Ishihara pseudo-isochromatic plates.

The photostress test is a useful test to distinguish a maculopathy from optic nerve dysfunction. The retina of the 'normal' eye is bleached by shining a bright light at the pupil for 10 s, and measuring the time for normal acuity to be re-established. The test is repeated in the 'abnormal' eye and if the difference in recovery time between the two eyes is greater than 60 s, the test is considered abnormal, indicating that the impairment is retinal and not due to an optic nerve disturbance.

Careful fundoscopic examination of the eye is essential to identify abnormalities of the optic media, retina, and optic nerve head.

Finally, examination of the visual fields is essential for topographic localization, since due to the invariate ordering of nerve fibres along the visual pathway, lesions at specific sites produce field defects of specific shapes (Fig. 1). Simple confrontation tests provide a qualitative method of investigating the visual fields. The examiner sits opposite the patient, maintaining a constant distance, and each eye is tested separately. With the patient fixating the examiner's nose, he/she is asked to count stationary fingers presented on either side of the vertical meridian in each quadrant in turn. If the patient cannot identify the fingers in a particular area, the fingers are gently wiggled, and the hand moved towards fixation until they are visible to the patient so mapping out the field defect. To examine the central field a red, 5 to 10 mm hatpin is moved away from or towards the central point of fixation. The patient is asked to describe any changes in the perception of colour or brightness, and whether or not the object disappears at any point. Perimetry provides a quantitative technique for measuring the fields, but a full description is beyond the scope of this chapter.

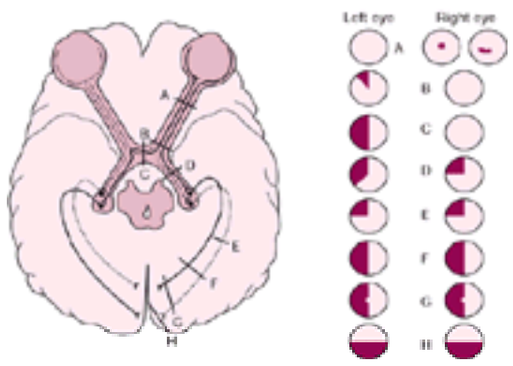

Fig. 1 Patterns of visual field loss due to lesions at different locations along the visual pathway. (a) Optic nerve lesions result in a central scotoma or arcuate defect. (b) Optic nerve lesions just prior to the chiasma produce junctional scotoma due to ipsilateral optic nerve involvement with the inferior contralateral crossing fibres (dashed lines). (c) Chiasma lesions produce bitemporal hemianopia. (d) Optic tract lesions result in incongruous hemianopic defects. (e) and (f) Lesions of the optic radiation result in either homonymous quadrantinopia or hemianopia depending on the extent and location of the lesion (upper quadrant, temporal lobe; lower quadrant, parietal lobe). (g) Lesions of the striate cortex produce a homonymous hemianopia, sometimes with macular sparing, particularly with vascular disturbances. (h) Lesions of the superior or inferior bank of the striate cortex result in inferior or superior altitudinal defects, respectively.

Abnormalities of the optic disc

Optic disc anomalies

Optic nerve hypoplasia

Hypoplasia of the optic nerve may be mild or severe, unilateral or bilateral, and may be associated with normal or impaired visual function. It may occur in isolation, or be associated with central nervous system anomalies, such as the absence of the septum pellucidum in the De Morsier's syndrome (septo-optic dysplasia).

Optic nerve dysplasia

Optic nerve dysplasia presents with a spectrum of abnormalities, including optic nerve colobomas, optic pits, and the morning glory syndrome, all considered to be associated with abnormal closure of the embryonic optic stalk and cup fissure. They are sometimes associated with basal encephaloceles and other forebrain anomalies.

Optic disc colobomas

These are deeply evacuated nerve head anomalies with blood vessels exiting from the margins which are associated with defects in the retinal nerve fibre layer, leading to an appropriate visual field loss.

Optic pits

Optic pits are crater-like depressions in the optic disc with a dark grey hue, usually situated in the temporal disc margin with an accompanying nerve fibre layer defect.

The morning glory syndrome

In this condition, an enlarged dysplastic disc is associated with an elevated centrally retained mass of glial and embryonic glial and vascular material, which radiates outwards in a sunburst pattern.

Tilted discs

An asymmetrically shaped, tilted disc is produced when the optic nerve leaves the globe at an extremely oblique angle. It is often associated with a crescentric zone of exposed sclera along one edge that results in elevation of the superior disc. The disc may appear hypoplastic and patients with this condition often have moderately high myopia and oblique astigmatism.

Optic nerve drusen

Drusen of the optic disc can give rise to an elevation of the optic nerve head. Drusen are intrapapillary, prelaminar refractile concretions that arise from degenerating nerve fibres. Anomalous discs due to drusen are usually smaller than normal, have an absent central optic disc cup, and exhibit an aberrant branching pattern of the central retinal vessels. Initially the drusen are buried with simple elevation of the disc, but become more apparent in later years when they appear to give rise to a typical lumpy disc, with a scalloped margin.

Myelinated nerve fibres

In slightly less than 1 per cent of the population some portions of retinal nerve fibres are myelinated, although normally optic nerve myelination stops at the lamina cribrosa. It appears on fundoscopy as a white area, usually adjacent to the disc, which has a centrifugal feathered edge.

Optic disc swelling

Although optic disc swelling and papilloedema have in the past been used synonymously, it is now usual only to refer to papilloedema as optic disc swelling when it is associated with raised intracranial pressure. Other cases of optic disc swelling are either due to local abnormalities in the optic nerve or orbit, or due to congenital anomalies as described above.

Local causes of optic disc swelling are usually associated with impaired visual acuity and colour vision, central arcuate or altitudinal field defects, and often an afferent pupillary defect. This contrasts with papilloedema when the acuity remains normal, except in the final stages, and is usually bilateral.

Papilloedema

The evolution of the disc changes in papilloedema due to raised intracranial pressure are usually classified into four stages: early, fully developed, chronic, and atrophic.

In early papilloedema there is disc hyperaemia, mild disc swelling with blurring of the fine peripapillary nerve fibre layer striations, dilatation of retinal veins with loss of spontaneous venous pulsations, and occasionally fine splinter haemorrhages at the disc margin.

In fully developed papilloedema, disc elevation is moderate to marked, and there is increased venous distension and tortuosity, an increasing number of peripapillary haemorrhages, cotton wool spots, and dilated capillaries on the disc surface. The retinal blood vessels and disc margin become increasingly indistinct.

In chronic papilloedema, there is resolution of the haemorrhages and exudates leaving a dome-shaped ('champagne cork') disc swelling, which often contains hard exudates. White refractile bodies may appear on the disc surface, known as corpora amylacea. As time goes on there is increasing nerve fibre attrition, leading to progressive visual field loss.

Finally, there is post-papilloedema (consecutive) atrophy, in which the disc acquires a milky opalescence and the retinal vessels are sheathed.

Clinical features

Usually papilloedema is bilateral and there is an absence of visual symptoms. However, unilateral or bilateral transient visual obscurations may occur, which last a few seconds and are often associated with postural changes. Although it has been suggested that such obscurations herald permanent visual loss, there is no evidence to support this view. The longer the papilloedema persists, the more likely there is to be progressive visual field loss, which usually starts as a peripheral field constriction. Occasionally, sudden visual loss occurs in a patient with papilloedema due to ischaemic optic neuropathy.

Pathogenesis

Papilloedema is due to impairment of axonal transport in the retinal nerve fibres, leading to axonal distension which is seen as disc swelling at the level of the prelaminar optic nerve.

Aetiology

There is a vast array of different causes leading to increased intracranial pressure, in particular space-occupying lesions such as tumours (Table 1).

Management

Treatment primarily depends on the underlying cause of the raised intracranial pressure. If due to a mass lesion which cannot be completely removed, or due to a non-surgically remediable cause, then a shunting procedure or medical measures, for example osmotic agents or diuretics such as acetazolamide, may be used. Increasingly, optic nerve sheath fenestration is being used for patients with intractable papilloedema who are developing early visual loss.

Ischaemic optic neuropathy

Ischaemic optic neuropathy is due to infarction of the optic nerve head, and can either be arteritic, as part of giant cell arteritis, or non-arteritic (idiopathic ischaemic neuropathy, anterior ischaemic optic neuropathy), which is the commoner form of the condition.

Non-arteritic ischaemic optic neuropathy

This tends to occur in patients aged between 45 and 80 years, and is characterized by abrupt, painless, and generally non-progressive visual loss, associated with an arcuate or altitudinal visual field loss. In nearly all cases, there is optic disc oedema, often associated with one or more splinter haemorrhages at the disc margin. Although previously considered irreversible, as many as 40 per cent of patients may show some improvement.

There is a 40 per cent chance of involvement of the fellow eye within five years. Optic atrophy rapidly ensues after the ischaemic event. The cause of non-arteritic ischaemic optic neuropathy remains obscure, and there is no treatment of proven benefit. The most important aspect of management is to exclude the possibility of the arteritic form, since in such cases the fellow eye is particularly vulnerable to similar involvement.

Arteritic ischaemic optic neuropathy

The arteritic form of ischaemic optic neuropathy usually occurs in giant cell (cranial, temporal) arteritis, but also rarely occurs in lupus and polyarteritis nodosa. Anyone with non-arteritic ischaemic optic neuropathy over the age of 50 should be suspected of having giant cell arteritis. This often occurs in the context of headache, malaise, weight loss, anorexia, anaemia, proximal muscle ache or stiffness, temporal artery tenderness, jaw claudication, and fever. These symptoms and signs usually precede the visual loss. The disc infarction is similar to that seen in non-arteritic ischaemic optic neuropathy.

A high index of suspicion is required for giant cell arteritis, and if suspected, an urgent erythrocyte sedimentation rate and temporal artery biopsy should be arranged. At the same time as the blood for the erythrocyte sedimentation rate is taken, the patient should be started immediately on systemic steroids (prednisolone at 80 mg daily, plus 200 mg of intravenous hydrocortisone immediately). In most patients the erythrocyte sedimentation rate is markedly elevated, as is the C-reactive protein. Occasionally the erythrocyte sedimentation rate may be normal. A biopsy of the superficial temporal artery should be obtained as soon as possible after the diagnosis has been considered. The biopsy will not be affected by the use of corticosteroids for at least 48 h. A positive temporal artery biopsy confirms the diagnosis of giant cell arteritis, but in 25 per cent of patients skip areas are found in biopsy specimens, and therefore a negative biopsy may sometimes be obtained.

Steroid treatment should not be tapered or withdrawn too early, since a relapse of symptoms is common. The dose of prednisolone can be gradually tapered after 2 to 3 weeks to maintain a normal erythrocyte sedimentation rate and the patient asymptomatic. Treatment should be continued for at least 6 to 12 months.

Optic atrophy

Optic atrophy is the final result of a variety of disturbances to the optic nerve or retina. The disc appears pale, and there is an absence of disc vasculature and retinal nerve fibres (Fig. 1).

Optic atrophy occurs after any disease process that results in death of the retinal ganglion cells with a dying back of their nerve fibres. This can, therefore, be due to diseases that directly involve the ganglion cells themselves or from damage to the axons in the pregeniculate visual pathway, resulting in retrograde atrophy. The development of optic atrophy is usually slow, dependent on its cause. In most instances the optic atrophy is bilateral, the disc appearing chalky-white in colour with clearly defined margins. The differential diagnosis of optic atrophy is considered in Table 2.

Optic neuritis

Optic neuritis is a term used to describe an idiopathic optic neuropathy or one resulting from inflammatory, infectious, or a demyelinating aetiology. In most cases the optic disc is normal on ophthalmoscopy and the term retrobulbar neuritis is used. In those cases in which the optic disc is swollen, the terms papillitis or anterior optic neuritis are used.

Clinical features

It is important to distinguish between those features of typical optic neuritis of idiopathic or demyelinating causation from atypical optic neuritis. In typical optic neuritis there is usually acute unilateral loss of visual acuity and of visual field, which may progress over hours or a few days, reaching its maximal effect within one week. Ninety per cent of patients complain of ocular pain which is noted especially with eye movement, and which may precede the visual impairment by a few days. The visual loss may range from contrast defects with maintained acuity, to no light perception. The patient is usually aged under 40 years, although optic neuritis may occur at any age, and improvement takes place in most patients (90 per cent) to normal or near normal visual acuity over several weeks. There may be persistent subtle residual defects of colour vision, depth perception, and contrast sensitivity, which may continue for several months. Subsequent disc pallor may occur but does not correlate closely with the level of visual recovery. An afferent pupillary defect is present in over 90 per cent of patients with acute optic neuritis. Although optic neuritis is generally associated with a central scotoma, a wide variety of field defects may be found ranging from a central scotoma, to altitudinal and nerve fibre layer defects.

Atypical optic neuritis may involve bilateral simultaneous onset of optic neuritis in an adult patient. There is often lack of pain and there may be other ocular findings suggestive of an inflammatory process, such as an anterior uveitis. Other features include a worsening of visual function beyond 14 days of onset, in a patient outside the 20- to 50- year age span. They may also have evidence of other systemic conditions, particularly inflammatory or infectious diseases (Table 3).

The evaluation of patients with optic neuritis rather depends on whether or not it is a typical or atypical case. Typical optic neuritis probably does not necessitate any additional laboratory investigations, although an abnormal MRI of the brain significantly increases the likelihood of developing multiple sclerosis.

Those patients with atypical optic neuritis should have a chest radiograph, laboratory tests including a blood count, biochemistry, and tests for collagen and vascular disease and syphilis serology. Examination of the cerebrospinal fluid is probably justified in this group of patients.

Management

Although intravenous methylprednisolone leads to a more rapid visual recovery, at the end of 6 months the visual acuity is no better than without the treatment. Therefore, steroid treatment of patients with typical optic neuritis is unnecessary, unless there is severe ocular pain that cannot be managed with analgesics, or if there is already poor vision in the fellow eye due to some other disease process.

Heredofamilial optic neuropathies

The hereditary optic neuropathies can either be those which are autosomal dominant or recessive, or those which are due to point mutations in mitochondrial DNA. The autosomal conditions usually present in childhood with impaired vision and pale optic discs.

Leber's hereditary optic neuropathy

This mitochondrial disorder develops primarily in males (approximately 14 per cent in women) in the second to third decade of life. It is characterized by an abrupt loss of central vision in one eye, usually followed by a loss of vision in the remaining eye which may occur weeks, months, or sometimes years later. Occasionally visual loss may occur simultaneously in the two eyes. There is no associated pain on eye movement in contrast to acute optic neuritis, and the visual loss is usually permanent with optic atrophy and large absolute central scotomas. However, the fundoscopic picture in the acute phase often shows swelling of the papillary nerve fibre layer, circumpapillary telangiectatic microangiopathy, and tortuosity of the retinal vessels.

There is a maternal pattern of inheritance and point mutations in mitochondrial DNA, particularly at the 11778 nucleotide and less frequently at 3460 and 14484, have been identified. The significance of the point mutation at 14484 is that a much higher percentage (37 per cent as opposed to 4 per cent) of patients show some visual recovery when compared with patients who have a defect at 11778. It is, therefore, appropriate to carry out genetic testing in those individuals presenting with atypical optic neuritis of the appropriate sex and age, even if a positive family history is not available. There is no effective treatment for this condition.

Nutritional and toxic optic neuropathies

Bilateral, slowly progressive central visual loss with centro-caecal scotomas, and usually normal or mild temporal atrophic optic discs characterizes optic nerve failure due to either nutritional deficiency or a toxic cause. Once a family history of one of the hereditary familial diseases has been excluded, this condition should be considered, and is usually due to a combination of alcohol abuse, deficiencies within the B vitamin complex, and frequently a high tobacco consumption. With treatment by abstinence of the likely toxic agents and vitamin supplementation, recovery of vision usually occurs, unless the condition is so long standing that optic atrophy has intervened. Recent epidemics of optic neuropathy in Cuba and in West Africa have probably been related to multiple dietary deficiencies.

Toxic optic neuropathy has been associated with ethambutol, chloramphenicol, halogenated hydroxyquinolones, lead, isoniazid, and vincristine.

Tumours of the optic nerve

Optic nerve sheath meningiomas

Although optic nerve sheath meningiomas may arise directly from the optic nerve sheath, usually in the orbital regions of the nerve, they frequently arise from the tuberculum sellae, sphenoid wing, and olfactory groove, leading to secondary invasion or compression of the nerve. Primary optic nerve sheath meningiomas, most frequently found in middle-aged women, are usually unilateral, but if bilateral raise the possibility of central neurofibromatosis (NF-2). Although most patients will have mild (2 to 4 mm) proptosis at the time of their initial consultation, they complain of dimming of vision and decreased colour vision. Visual loss progresses over years with optic disc swelling gradually being supplanted by optic atrophy, with or without the evolution of optociliary venous (retinochoroidal anastamoses) shunt vessels.

The CT picture in patients with these tumours is most often one of diffuse narrow enlargement of the optic nerve, with bulbous swellings of the nerve in the region of the globe and orbital apex. 'Railroad-track' calcification of the optic nerve sheath in the orbit is a characteristic feature. Use of MRI has enabled optic nerve sheath meningiomas to be distinguished from optic nerve gliomas, where the former but not latter shows that the nerve is readily distinguished from the optic nerve sheath.

Management of patients with optic nerve sheath meningiomas is controversial. While there is general agreement that nerve sheath tumours are most aggressive in children and become progressively more indolent with advancing age, there is no consensus as to the best way to treat these lesions. Clinical resection, particularly when there is intracranial spread, is usually incomplete. These patients rarely die from the meningioma and it is probably best to observe. In some instances radiotherapy has shown to result in some visual improvement.

Optic nerve gliomas

Optic nerve gliomas, which may also involve the chiasma, are of two distinct types. By far the commoner is the benign glioma of childhood, and the other the malignant glioblastoma in adults. Approximately a quarter of cases occur in the setting of NF-1.

Benign optic nerve gliomas usually present within the first two decades of life, with a peak incidence from 1 to 6 years of age. The usual presenting manifestations are proptosis and visual loss, which may be so mild as to be undetectable, although a profound reduction in acuity is more common. The fundus may show either papilloedema or optic atrophy.

The clinical course of childhood optic nerve gliomas is highly variable. In some, tumour enlargement proceeds slowly for a time but then reaches a plateau, while in others the enlargement proceeds unabated. Necropsy has suggested that they are in fact hamartomas rather than true neoplasms. Optic nerve gliomas are generally managed conservatively, although some practitioners favour radiation therapy for lesions with chiasmal involvement and surgery for at least those tumours restricted to the orbit.

Optic nerve gliomas of adulthood are malignant gliomas which usually arise in males aged 40 to 60 years. These patients often present with a rapid onset of visual failure, which on some occasions may mimic acute optic neuritis. The tumour rapidly progresses and the patient usually dies within a short period.

Other optic nerve tumours

Metastatic cancer may lead to optic nerve involvement, either as a result of infiltration of the meninges as occurs with cancer of the breast and lung, or by direct tumour infiltration as with lymphoproliferative disorders and certain types of leukaemia and non-Hodgkin's lymphoma. Paraneoplastic optic neuropathy has also been described in patients with small cell carcinoma of the lung.

Disorders of the optic chiasm

Approximately 25 per cent of all brain tumours occur in the chiasmal region and since half of these cases initially present with visual loss, an appreciation of the various field abnormalities is important. Although there are a number of other causes for the chiasmal syndrome, such as trauma and demyelination, these are rare. The neuro-ophthalmological signs of a compressive optic chiasm lesion are primarily a field defect and deterioration of visual acuity, which depend on the relationship of the chiasm to the pituitary. The classic field defect of a chiasmal lesion is a bitemporal hemianopia. This may be complete or incomplete and may or may not be symmetrical. It is unusual to have a bitemporal hemianopia without some reduction in central visual acuity in at least one eye, due to the optic nerve being compromised in addition to the chiasm.

In large series of patients with pituitary tumours the most common field defect is a bitemporal hemianopia (67 per cent); less common are junctional scotoma (29 per cent), homonymous hemianopia (7 per cent), and prechiasmal field loss (2 per cent). Other signs include optic disc pallor, but its absence usually denotes a virtual complete return of visual function with successful decompression.

Other causes of chiasmal compression in addition to pituitary adenomas (50 to 55 per cent) include craniopharyngiomas (20 to 25 per cent), meningiomas (10 per cent), and gliomas (7 per cent).

Optic tract lesions

The optic tract is the first point in the visual pathways where the ipsilateral temporal and contralateral nasal retinal nerve fibres come together, and so the field defect is usually a partial or complete homonymous hemianopia. When partial, there is often gross incongruity between the visual field defects found in each eye, which may also be found with lesions of the lateral geniculate nucleus and more rarely the optic radiations.

The most frequently encountered lesions causing the optic tract syndrome are aneurysms, craniopharyngiomas, and pituitary tumours.

The optic radiations

As the geniculostriate fibres leave the lateral geniculate nucleus, the ventral fibres (subserving the superior visual field) pass anteriorly around the temporal horn of the lateral ventricle to form Meyer's loop. Lesions in this region usually result in a wedge-shaped congruous homonymous field defect, mainly affecting the superior quadrant. The visual acuity and pupillary responses are both normal. Lesions involving the optic radiation are due to vascular occlusion, tumours (intrinsic or metastatic), or abscesses.

Although lesions of the dorsal optic radiation in the parietal lobe may result in a homonymous hemianopia primarily affecting the lower fields, large lesions usually result in a complete homonymous hemianopia with macular splitting. Damage to the parietal or occipitoparietal cortex may result in the phenomenon in the contralateral visual field called unilateral visual inattention or visual extinction. A test object presented in this field is perceived normally, but when an identical object is similarly presented equidistant from the fixation point in the ipsilateral visual field, the stimulus in the field contralateral to the parietal lobe lesion disappears.

Occipital lobe

On reaching the occipital lobe there is a high degree of order in the fibres of the optic radiation and lesions, which are usually due to infarction, trauma, or tumour, produce homonymous congruent field defects. The only features of the field defect which help localize the lesion to the occipital lobe, rather than the anterior optic radiation, are the presence of sparing of the macula or temporal crescent areas in a homonymous hemianopia.

In macula sparing there is preservation of the visual field within a region of 1 to 2° up to 10° around the fixation point in the hemianopic field. In the more usual situation the hemianopic field is split along the vertical meridian through the fixation point (macula splitting).

Altitudinal (dorsal/ventral) field defects involving either the upper or lower occipital poles may occur as a result of trauma or vascular lesions.

Cortical blindness

Cortical blindness usually indicates selective involvement of the occipital visual cortex. The essential features are: (i) complete loss of all visual sensation, (ii) loss of reflex lid closure to threat, (iii) normal pupillary light reactions, and (iv) normal retina and full extraocular eye movements. The commonest aetiology is hypoxia of the striate cortex.

Disorders of higher visual processing

In the extrastriate cortex there is parallel processing of different aspects of visual information before an organized synthesis of the visual scene can be generated. Specific lesions in one or other of these areas might be expected to give rise to an appropriate specific loss of a visual modality such a colour (achromatopsia), movement (akinetopsia), or faces (prosopagnosia).

Acquired disorders of colour vision due to lesions of the central nervous system are of two types. In one type there is an inability to see colours (dyschromatopsia or achromatopsia). These patients have lesions in the region of the lingual and fusiform gyri, which lies in the anterior inferior region of the occipital lobe. They complain that they cannot see colours and that everything looks grey or in varying shades of black and white. They are unable to identify the figures on pseudo-isochromatic test plates although they are able to name the colours of brightly coloured objects correctly. Other functions such as visual acuity, object recognition, and depth perception are all normal, but there is often an associated visual field defect, usually a bilateral superior homonymous quadrantanopia. In the other type of disorder the colour sense is normal but the naming and recognition of colour is impaired. This can occur as part of an aphasia, such as Wernicke's or anomic, in the syndrome of alexia without agraphia, or as one feature of visual agnosia (see below).

Rare cases of patients who exhibit a selective deficit of movement perception (akinetopsia) have been reported. The patients have bilateral lesions involving the lateral occipito-parieto-temporal junction.

Visual agnosia

The term visual agnosia refers to a rare condition in which there is an inability to recognize, name, or demonstrate the use of an object presented visually, in the absence of a language deficit, general intellectual dysfunction, or attentional disturbances. The patient is, however, able to name the object when using other sensory modalities such as touch or sound.

One classification depends on the specific category of visual material that cannot be recognized. A disturbance of recognition of objects (object agnosia), faces (prosopagnosia), or colour (colour agnosia) may occur in isolation or in various combinations. When patients are able to copy and match-to-sample objects that they fail to name or recognize visually, the agnosia is termed associative; but if there is an inability to perform all these tasks, the agnosia is termed apperceptive.

Prosopagnosia is a specific inability to recognize familiar faces despite a normal ability to recognize everyday objects, and is, therefore, different from visual agnosia. Most cases of prosopagnosia are due to infarction, head injury, or hypoxia resulting in bilateral lesions in the ventromedial aspects of the occipitotemporal region.

Visual illusions

Visual illusions occur when the visually perceived target appears altered in size, shape, colour, position in space, and in number of images. The illusory type of defects may occur in the entire field of vision, or may affect only the object or the background. The term 'dysmetropsia' indicates the apparent smallness (micropsia), largeness (macropsia), or irregularity of shape (metamorphopsia) of objects. Dysmetropsia usually occurs as a result of retinal disease due to distortion of the relative distance between rods and cones.

Visual hallucinations

Visual hallucinations occur under many circumstances, such as drug withdrawal, anoxia, migraine, infection, and schizophrenia, in addition to those related to focal neurological disease in the occipital or temporal lobes. Those in the latter category may be unformed, consisting of flashes of light (coloured or white), lines, or simple shapes, or they may be complex highly organized hallucinations of people or objects.

Palinopsia

Palinopsia is a rare disorder in which there is persistence (perseveration) or recurrence of visual images after the exciting stimulus has been removed.

Further reading

Apple DJ, Rabb MF, Walsh PM (1982). Congenital anomalies of the optic disc. *Survey in Ophthalmology* **27**, 3–41.

Beck RW, ONTT Study Group (1992). A randomised, controlled trial of corticosteroids in the treatment of acute optic neuritis. *New England Journal of Medicine* **326**, 581–8.

Boghen DR, Glaser JS (1975). Ischaemic optic neuropathy: the clinical profile and natural history. *Brain* **98**, 689–708.

Chung SM, Selhorst JB (1992). Cancer associated retinopathy. In: Katz B, ed. *Neuro-ophthalmology in systemic disease*. *Ophthalmology Clinics of North America* **5**(3), 587–96.

De Renzi E (1997). Prosopagnosia. In: Finberg TE, Farah MJ, eds. *Behavioural neurology and neuropsychology*, pp 245–55. McGraw-Hill, New York.

Dutton JJ (1992). Optic nerve sheath meningiomas. *Survey in Ophthalmology* **37**, 167–83.

Dutton JJ (1994). Gliomas of the anterior visual pathway. *Survey in Ophthalmology* **38**, 427–52.

Horton JC, Hoyt WF (1991). The representation of the visual field in human striate cortex: a revision of the classic Holme's map. *Archives of Ophthalmology* **109**, 816–24.

Humphreys GW, Riddoch MJ (1993). Object agnosias. In: Kennard C, ed. *Visual perceptual defects*, pp 339–59. Baillière Tindell, London.

Kölmel HW (1993). Visual illusions and hallucinations. In: Kennard C, ed. *Visual perceptual defects*, pp 243–64. Baillière Tindell, London.

Liu GT *et al*. (1994). Visual morbidity in giant cell arteritis: clinical characteristics and prognosis for vision. *Ophthalmology* **101**, 1779–85.

Manford M, Anderman F (1998). Complex visual hallucinations: clinical and neurobiological insights. *Brain* **121**, 1819–40.

McDonald WI, Barnes D (1992). The ocular manifestations of multiple sclerosis. I. Abnormalities of the afferent visual system. *Journal of Neurology, Neurosurgery and Psychiatry* **55**, 747–52.

Neetens A, Smets RM (1989). Papilloedema. *Neuro-Ophthalmology* **9**, 81–101.

Riddoch G (1917). Dissociation in visual perceptions due to occipital injuries, with special reference to appreciation of movement. *Brain* **40**, 15–57.

Riordan-Eva P *et al*. (1995). The clinical features of Leber's hereditary optic neuropathy defined by the presence of a pathogenic mitochondrial DNA mutation. *Brain* **118**, 319–37.

Rosenberg MA, Savino PJ, Glaser JS (1979). A clinical analysis of pseudo-papilloedema: I. population, laterality, acuity, refractive error, ophthalmoscopic characteristics, and coincident disease. *Archives of Ophthalmology* **97**, 65–70.

Sadun AA *et al*. (1994). Epidemic optic neuropathy in Cuba: eye findings. *Archives of Ophthalmology* **112**, 691–9.

Sugishita M *et al*. (1993). The problem of macular sparing after unilateral occipital lesions. *Journal of Neurology* **241**, 1–9.

Thompson HS (1966). Afferent pupillary defects. *American Journal of Ophthalmology* **62**, 860–73.

Zeki S (1993). *A vision of the brain*. Blackwell, London.

24.12.1Eye movements and balance

Thomas Brandt and Michael Strupp

Introduction

The main objective of this chapter is to describe the clinically relevant principles of ocular motor and vestibular disorders. The anatomical and functional overlap of the vestibular and ocular motor systems warrants a joint discussion.

Eye movements

Different types of eye movement can be distinguished, each with particular functions, physiological properties, and specific anatomical substrates. Many abnormalities of eye movement are thus distinctive and often indicate the site and the side of a lesion. This is useful for topographic diagnosis, a method that still frequently proves superior to imaging techniques. It is therefore important that the physician examines in detail the eye movements of a patient suffering from, for example, double vision, oscillopsia, or vertigo, for he can by this means often differentiate between 'peripheral' and 'central' ocular motor or vestibular disorders. In their excellent book *Neurology of eye movements*, Leigh and Zee (1999) correctly state that 'an understanding of the properties of each functional class of eye movements will guide the physical examination; a knowledge of the neural substrate will aid topological diagnosis'.

Normal vision relies on eye movements in two essential ways. On the one hand, eye movements make it possible to shift the gaze and to view objects of interest. On the other, when the head moves during locomotion, the eyes move in a direction opposite to that of the head and compensate for these head movements, thereby preventing involuntary shifts of the visual images projected on the retina. The retinal images are kept steady. Optimal functioning of the eye movements is ensured by the co-operation between the optokinetic reflex and the vestibulo-ocular reflex.

In essence, there are three types of conjugate eye movements: smooth pursuit, saccades, and the vestibulo-ocular reflex. In the following paragraphs we summarize their properties, main function, clinical examination, and pathological findings as well as the typical features of vestibular nystagmus and gaze-evoked nystagmus.

Smooth pursuit keeps the image of a moving object on the fovea. The pursuit system generates smooth tracking eye movements that closely match the pace of a target; its examination is illustrated in Fig. 1. Various anatomical structures (motion-sensitive visual cortex, frontal eye fields, pontine nuclei, cerebellum, vestibular and ocular motor nuclei) are involved in smooth pursuit eye movements. Therefore, topographically impaired smooth pursuit (reduced gain) is an unspecific finding, which may be further influenced by alertness, a variety of drugs, and age. Moreover, vertical smooth pursuit is worse than horizontal, and downward tracking worse than upward. Marked asymmetries of smooth pursuit, however, indicate a central lesion; strongly impaired smooth pursuit is observed in intoxications and degenerative disorders involving the cerebellum or extrapyramidal system.

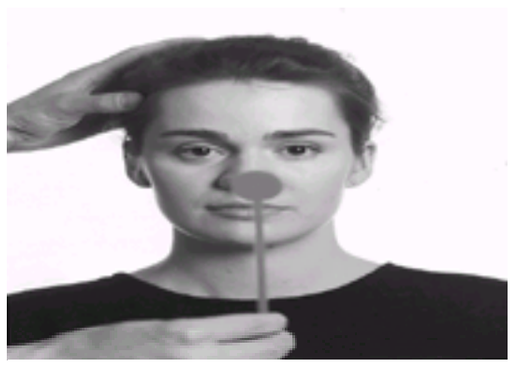

Fig. 1 Clinical examination of smooth pursuit. The patient is asked to track visually an object moving slowly in horizontal and vertical directions (10 to 20°/s) with the head stationary. Look for corrective (catch-up or back-up) saccades; they indicate an inappropriate smooth pursuit gain.

Saccades bring images of objects of interest on the fovea; their clinical examination is illustrated in Fig. 2. Slowing of saccades—often accompanied by hypometric saccades—is also often a side-effect of many types of medications/toxins and is found in neurodegenerative disorders. Slowing of horizontal saccades is observed in brainstem lesions, for example of the ipsilateral paramedian pontine reticular formation; slowing of vertical saccades may be due to a midbrain lesion and is observed in progressive supranuclear palsy. Lesions of the cerebellum or cerebellar pathways may cause hypermetric saccades, followed by corrective saccades that can be easily observed. For example, in Wallenberg's syndrome, a saccadic overshoot toward the side of the lesion is due to an interruption of the inferior cerebellar peduncle; interruption of the superior cerebellar peduncle leads to contralateral hypermetric saccades. In internuclear ophthalmoplegia the adducting saccade is slower than the abducting saccade. Delayed-onset saccades are most often caused by cerebral cortical lesions.

Fig. 2 Clinical examination of saccades. First observe spontaneous saccades to visual or auditory targets. Then ask the patient to glance back and forth between two horizontal and two vertical targets, keeping the head stationary. The velocity, accuracy, conjugacy, and the initiation time of the saccade should be observed. Normal individuals can immediately reach the target with a fast single movement or one small corrective saccade.

The vestibulo-ocular reflex holds images of the seen world steady on the retina during brief head rotations and locomotion. Halmagyi and Curthoys (1988) described an important clinical bedside test of the horizontal vestibulo-ocular reflex, which is illustrated in Fig. 3. This simple test allows the physician to find out whether there is a unilateral or a bilateral peripheral vestibular deficit.

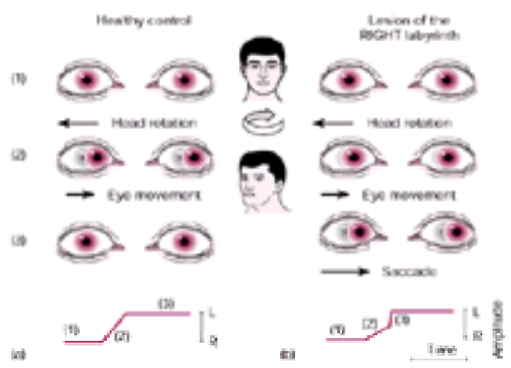

Fig. 3 Clinical bedside testing of the horizontal vestibulo-ocular reflex by the Halmagyi–Curthoys test. Fast 20 to 30° rotations of the head toward the side of the lesion show the dynamic deficit of the horizontal vestibulo-ocular reflex. In contrast to the healthy control (a), the patient is not able to generate a fast contraversive eye movement and has to perform a corrective (catch-up) saccade to fixate the target (b). It is important to instruct the patient to look carefully at the examiner's nose and to apply brief, high acceleration head thrusts to detect a unilateral peripheral vestibular deficit, for example due to vestibular neuritis or acoustic neuroma.

Finally, two clinically important pathological eye movements should be mentioned: gaze-evoked nystagmus and spontaneous nystagmus. As illustrated in Fig. 4 and Fig. 5, gaze-evoked nystagmus should be observed when the patient is fixating with both eyes. It is most often a side-effect of medication/toxins such as anticonvulsants, hypnotics, or alcohol. Horizontal gaze-evoked nystagmus may be due to structural lesions of the brainstem (medial vestibular nucleus and nucleus prepositus hypoglossi, that is, the neural integrator to maintain gaze position after execution of a gaze shift), or the flocculus. Vertical gaze-evoked nystagmus is observed in midbrain lesions involving the interstitial nucleus of Cajal. A dissociated horizontal gaze-evoked nystagmus (greater in the abducting than the adducting eye) and an adduction deficit are the signs of internuclear ophthalmoplegia due to a lesion of the medial longitudinal fasciculus.

Fig. 4 Clinical examination of the eyes in nine different positions to evaluate ocular alignment, fixation deficits, nystagmus, range of movement, and gaze-holding abilities. The examination can be performed using an object (left) or an examination lamp. In primary position look for (a) abnormal eye movements such as nystagmus (for example, peripheral vestibular: horizontal-rotatory, suppressed by fixation; central vestibular: vertical (upbeat, downbeat), horizontal or torsional, poorly suppressed or even increasing with fixation; congenital: usually horizontal, variable in frequency and amplitude, increasing with fixation); square-wave jerks (small saccades of 0.5 to 5° that cause the eyes to move from the primary position, for example in progressive supranuclear palsy or certain cerebellar syndromes); ocular flutter (intermittent bursts of horizontal oscillations); or opsoclonus (combined horizontal, vertical, and rotatory oscillations); the latter two may have different aetiologies, for example encephalitis, tumours, or drugs/toxins and (b) misalignment of the visual axes. Then establish the range of motion with ductions (one eye viewing) and versions (with both eyes viewing) in the eight end-positions; this can indicate, for example, ocular muscle or nerve palsy. Gaze-holding deficits can be evaluated in eccentric gaze position (Fig. 5).).

Fig. 5 Clinical examination of the eye positions/movements using an examination lamp. The advantage of the lamp as opposed to an object is that the reflected light on the eye can be observed and thus ocular misalignments can be easily detected. In addition, the patient can fixate with one or both eyes in the end-positions.

Spontaneous nystagmus indicates a tone imbalance of the vestibulo-ocular reflex which may be central or peripheral; when peripheral—as in vestibular neuritis—it is typically damped by visual fixation. Therefore, spontaneous nystagmus should be examined by Frenzel's glasses (Fig. 6).

Fig. 6 Clinical examination with Frenzel's glasses. The magnifying lenses (+16 dioptres) have light inside to prevent visual fixation, which could suppress spontaneous nystagmus. Frenzel's glasses enable the clinician to observe spontaneous eye movements better. Examination should include spontaneous and gaze-evoked nystagmus, head-shaking nystagmus (instruct the patient to rotate his head about 20 times and observe eye movements following head shaking), positioning and positional nystagmus, as well as hyperventilation-induced nystagmus.

As a general rule, it is often necessary to combine the pathological clinical findings of the different eye movement systems to differentiate between a central and peripheral vestibular disorder and to make an exact topographical diagnosis.

Dizziness and vertigo

Vertigo, dizziness, and disequilibrium are common complaints of patients of all ages, particularly the elderly. As presenting symptoms, they occur in 5 to 10 per cent of all patients seen by general practitioners and 10 to 20 per cent of all patients seen by neurologists and otolaryngologists. The clinical spectrum of vertigo is broad, extending from vestibular rotatory vertigo with nausea and vomiting to presyncope light-headedness, from drug intoxication to hypoglycaemic dizziness, from visual vertigo to phobias and panic attacks, and from motion sickness to height vertigo. Appropriate preventions and treatments differ for the various types of dizziness and vertigo; they include drug therapy, physical therapy, psychotherapy, and surgery.

Vertigo usually implies a mismatch between the vestibular, visual, and somatosensory systems. These three sensory systems subserve both static and dynamic spatial orientation, locomotion, and control of posture by constantly providing reafferent cues. The sensory information is partially redundant in that two or three senses may simultaneously provide similar information about the same action. Thanks to this overlapping of their functional ranges, it is possible for one sense to substitute, at least in part, for deficiencies in the others. When information from two sensory sources conflicts, the intensity of the vertigo is a function of the degree of mismatch; it is increased if information from an intact sensory system is lost, as for example in a patient with pathological vestibular vertigo who closes his eyes. The distressing sensorimotor consequences of the mismatch are frequently based on our earlier experiences with orientation, balance, and locomotion, that is, there is a mismatch between the expected and the actually perceived pattern of multisensory input.

Vertigo may thus be induced by physiological stimulation of the intact sensorimotor systems (height vertigo; motion sickness) or by pathological dysfunction of any of the stabilizing sensory systems, especially the vestibular system. The symptoms of vertigo include sensory qualities identified as arising from vestibular, visual, and somatosensory sources. As distinct from one's perception of self-motion during natural locomotion, the experience of vertigo is linked to impaired perception of a stationary environment; this perception is mediated by central nervous system processes known as 'space constancy mechanisms'. Loss of the external stationary reference system required for orientation and postural regulation contributes to the distressing mixture of self-motion and surround motion.

Physiological and clinical vertigo syndromes are commonly characterized by a combination of phenomena involving perceptual, ocular motor, postural, and autonomic manifestations: vertigo, nystagmus, ataxia, and nausea. These four manifestations correlate with different aspects of vestibular function and emanate from different sites within the central nervous system.

1. The vertigo itself results from a disturbance of cortical spatial orientation.
2. Nystagmus is caused by a direction-specific imbalance in the vestibulo-ocular reflex, which activates brainstem neuronal circuitry.
3. Vestibular ataxia and postural imbalance are caused by inappropriate or abnormal activation of monosynaptic and polysynaptic vestibulospinal pathways.
4. The unpleasant autonomic responses with nausea, vomiting, and anxiety travel along ascending and descending vestibulo-autonomic pathways to activate the medullary vomiting centre.

About 50 per cent of all patients presenting with dizziness, vertigo, or disequilibrium in a neurological dizziness unit will be suffering from one of the five following common syndromes (Table 1):

1. benign paroxysmal positioning vertigo;
2. somatoform vertigo (phobic postural vertigo);
3. basilar migraine (vestibular migraine);
4. Menière's disease; or
5. vestibular neuritis.

A clinician unfamiliar with dizzy patients can most effectively deepen his knowledge by acquainting himself with these five most frequently met and challenging conditions of vertigo. Diagnosis and management of vertigo syndromes always require interdisciplinary thinking, and history taking is still much more important than recordings of eye movements or brain imaging techniques. Although most clinicians welcome the attempts to develop computer interview systems for use with neuro-otological patients and expert systems as diagnostic aids in otoneurology, their application in a clinical setting is still quite limited.

Dizziness is a vexing symptom, difficult to assess because of its purely subjective character and its variety of sensations. The sensation of spinning or rotatory vertigo is much more specific; if it persists, it undoubtedly indicates acute pathology of the labyrinth, the vestibular nerve, or the caudal brainstem, which contains the vestibular nuclei.

History taking allows the early differentiation of vertigo and disequilibrium disorders into seven categories that serve as a practical guide for differential diagnosis:

1. dizziness and light-headedness (such as presyncopal dizziness or drug intoxication);
2. single or recurrent attacks of (rotatory) vertigo (such as Menière's disease, vestibular migraine);
3. sustained (rotatory) vertigo (such as vestibular neuritis, Wallenberg's syndrome);
4. positional/positioning vertigo (such as in benign paroxysmal positioning vertigo, central positional vertigo);
5. oscillopsia (apparent motion of the visual scene, such as in bilateral vestibulopathy, downbeat nystagmus);
6. vertigo associated with auditory dysfunction (such as Menière's disease, Cogan's syndrome); and
7. dizziness or to-and-fro vertigo with postural imbalance (such as phobic postural vertigo, episodic ataxia).

Management of the dizzy patient

The prevailing good prognosis of vertigo should be emphasized because of the following.

1. Many forms of vertigo have a benign cause and are characterized by spontaneous recovery of vestibular function or central compensation of a peripheral vestibular tone imbalance.
2. Most forms of vertigo can be effectively relieved by pharmacological treatment (Table 2), physical therapy (Table 3), surgery (Table 4), or psychotherapy.

There is, however, no common treatment, and vestibular suppressants provide only symptomatic relief of vertigo and nausea. A specific therapeutic approach thus requires recognition of the numerous particular pathomechanisms involved. Such therapy can include causative, symptomatic, or preventive approaches.

The essential characteristics are given for benign paroxysmal positioning vertigo (Table 5; see Fig. 7), Menière's disease (Table 6), vestibular neuritis (Table 7), bilateral vestibular failure (Table 8), and vestibular migraine (Table 9).

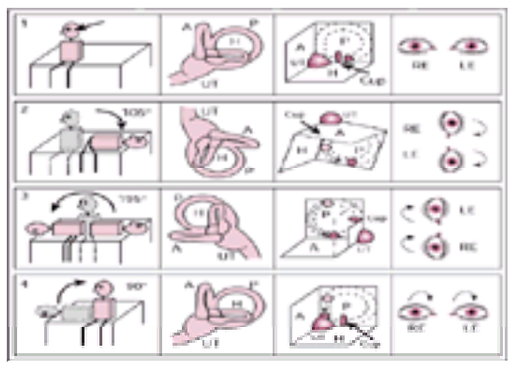

Fig. 7 Schematic drawing of the Semont liberatory manoeuvre in a patient with typical benign paroxysmal positioning vertigo (BPPV) of the left ear. Boxes from left to right: position of body and head, position of labyrinth in space, position and movement of the clot in the posterior canal and resulting cupula deflection, and direction of the rotatory nystagmus. The clot is depicted as an open circle within the canal; a black circle represents the final resting position of the clot. (1) In the sitting position, the head is turned horizontally 45° to the unaffected ear. The clot, which is heavier than endolymph, settles at the base of the left posterior semicircular canal. (2) The patient is tilted approximately 105° toward the left (affected) ear. The change in head position, relative to gravity, causes the clot to gravitate to the lowermost part of the canal and the cupula to deflect downward, inducing BPPV with rotatory nystagmus beating toward the undermost ear. The patient maintains this position for 2 min. (3) The patient is turned approximately 195° with the nose down, causing the clot to move toward the exit of the canal. The endolymphatic flow again deflects the cupula such that the nystagmus beats toward the left ear, now uppermost. The patient remains in this position for 2 min. (4) The patient is slowly moved to the sitting position; this causes the clot to enter the utricular cavity. Abbreviations: A, P, and H: anterior, posterior, and horizontal semicircular canals; Cup, cupula; UT, utricular cavity; RE, right eye; LE, left eye. (From Brandt *et al*. 1994.)

Further reading

Baloh RW, Halmagyi GM (1996). *Disorders of the vestibular system*. Oxford University Press, Oxford.

Brandt T (1999). *Vertigo—its multisensory syndromes*, 2nd edn. Springer, London.

Brandt T, Steddin S, Daroff RB (1994). Therapy for benign paroxysmal positioning vertigo, revisited. *Neurology* **44**, 796–800.

Bronstein A, Brandt Th, Woollacott M (1996). *Clinical disorders of balance, posture and gait*. Arnold, London.

Halmagyi GM, Curthoys IS (1988). A clinical sign of canal paresis. *Archives of Neurology* **45**, 737–9.

Herdman, SJ (2000). *Vestibular rehabilitation*, 2nd edn. F.A. Davies, Philadelphia.

Leigh RJ, Zee DS (1999). *Neurology of eye movements*, 3rd edn. F.A. Davies, Philadelphia.

24.12.2Disorders of hearing

Linda M. Luxon

> Our hearing is a choice and dainty sense, and hard to mend, yet soon it may be marred. Blows, falls and noise...all these...breed tingling in the ears and hurt our hearing.
>
> (Physicians of the Medical School of Salerno)

Hearing loss is the most common sensory impairment and the World Health Organization has estimated that at least 120 million people are affected worldwide. In the United Kingdom, about 20 per cent of the adult population are affected and, of these, three-quarters are over the age of 60 years. Age, gender, occupational group, and occupational noise exposure are factors affecting the prevalence of hearing impairment in adults, while a neonatal intensive-care unit history, a family history of childhood hearing loss, and craniofacial abnormalities explain over half of the population of congenitally hearing-impaired children. The most frequent form of acquired hearing impairment in children is a conductive hearing loss due to chronic secretory otitis media, while meningitis, in particular meningococcal meningitis, is the commonest cause of acquired sensorineural hearing loss. In the developing world, many of the preventable causes of hearing impairment remain common: consanguineous marriages, birth trauma, childhood infections, noise exposure, and the unlicensed sale of ototoxic drugs.

Tinnitus, defined as a noise in the head or ears lasting for more than 5 min, increases with age and affects approximately 20 per cent of the population over the age of 60 years, but only 4 per cent of the population complain of this symptom.

Hearing impairments

Pathophysiology

For clinical purposes the ear is separated into three parts: the external, middle, and internal ear (Fig. 1). The external ear is important in funnelling sound to the tympanic membrane and in the localization of sound. The middle ear ossicles connect the tympanic membrane to the oval window of the cochlea, such that sound waves cause displacement within the fluid-filled compartment of the membranous labyrinth. Within the internal ear, the mechanical activity at the oval window is transduced into neural responses by the hair cells of the Organ of Corti (Fig. 2).

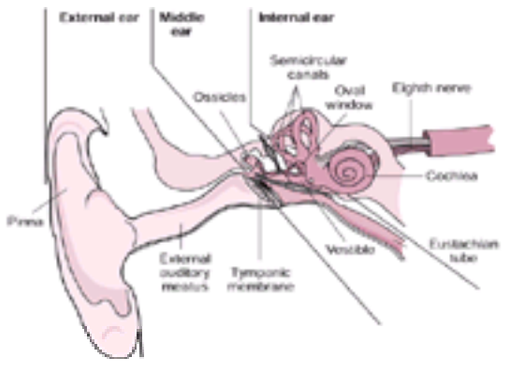

Fig. 1 Diagram to illustrate the anatomy of the peripheral auditory system.

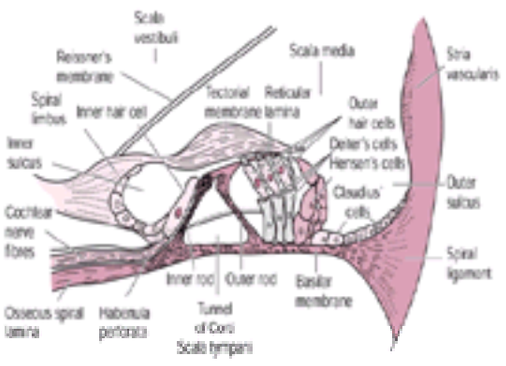

Fig. 2 Diagram of the Organ of Corti.

Disorders of the external and middle ear result in abnormalities of the mechanical transmission of sound from the environment to the internal ear, and give rise to a conductive hearing loss. Common examples include impacted wax, serous otitis media (glue ear), chronic otitis media, and disorders of the ossicular chain, for example otosclerosis, and traumatic discontinuity.

Disorders of the internal ear and the VIIIth cranial nerve characteristically give rise to a sensorineural hearing loss, in which the perception of both bone- and air-conducted sounds is reduced and the appreciation of the intensity of sound and the frequency resolution of complex sounds are impaired. Many conditions may affect the cochlea, ranging from inherited, congenital or iatrogenic non-syndromal or syndromal malformations to ototoxic damage (aminoglycoside, antimalarial, loop diuretics), ischaemia including vertebrobasilar ischaemia, diabetic vasculitis, infections (mumps, rubella, syphilis, cytomegalovirus), autoimmune disorders, degenerative disorders, trauma, and idiopathic conditions such as Menière's disease. Pathology of the VIIIth nerve leading to hearing impairment has been defined in spinocerebellar degenerations, trauma, cerebellopontine angle tumours, bony disorders such as Paget's disease, infective disorders (meningitis), and inflammatory conditions (sarcoidosis). Much doubt has been cast on so-called 'presbyacusis', which may merely reflect an accumulation of toxic/traumatic insults to the ear over many years, and recent advances in molecular biology and genetics have shown the role of genetic mutations/deletions in late-onset/progressive hearing impairments.

Clinical examination

Clinical examination requires examination of the anatomy of the external ear to define visible signs of congenital ear disease (pits, tags, nodules, or malformations)

and evidence of other craniofacial features suggestive of syndromal hearing impairment. In addition, a detailed examination of the tympanic membrane is required to define the presence of pathology within the middle ear. Wax or debris obstructing the external auditory meatus should be removed by or under the supervision of a clinician with experience in this field. Syringing should never be undertaken in the presence of an infection or if it is unknown whether the tympanic membrane may be perforated. Tuning-fork tests remain the most valuable clinical test of auditory function and frequently enable a clinician to distinguish a conductive from a sensorineural hearing loss (Fig. 3). The tests are based on two physiological facts: first, the inner ear is normally more sensitive to sound conducted by air than to that conducted by bone; and second, in the presence of a purely conductive hearing loss the affected ear is subject to less environmental noise, making it more sensitive to bone-conducted sound. A general medical and neurological examination is mandatory to define syndromes and the plethora of general medical conditions associated with hearing impairment.

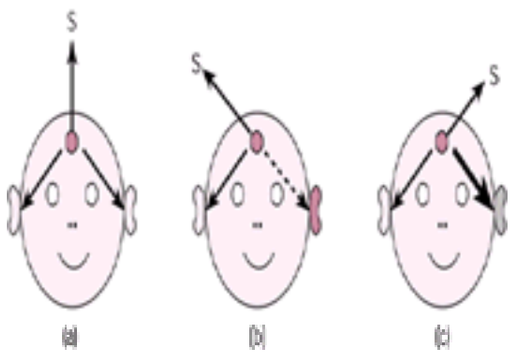

Fig. 3 Diagram to illustrate the Weber tuning fork test in (a) a normal subject, (b) a case of unilateral sensorineural hearing loss, and (c) a unilateral conductive hearing loss, in which the sound is heard more effectively in the affected ear because of the lack of masking by environmental sounds. (s, sound heard; •, tuning fork)

Investigations

A battery of audiological tests is required to:

- quantify audiometric thresholds at each frequency;
- differentiate a conductive from a sensorineural hearing loss;
- differentiate a cochlear from a retrocochlear abnormality;
- identify central auditory dysfunction in the brainstem, midbrain, or auditory cortex; and
- identify a non-organic component.

Each test can be defined as being subjective (dependent upon patient co-operation) or objective (independent of patient co-operation) in terms of providing auditory data. To differentiate a sensorineural hearing loss of cochlear origin from that of an VIIIth nerve dysfunction or neurological disorder, two pathophysiological phenomena are of importance:

1. *Loudness recruitment* is defined as an abnormally rapid increase in loudness, with an increase in intensity of the stimulus, and is characteristic of disorders affecting the hair cells of the Organ of Corti, but is absent in pathology of the VIIIth nerve.
2. *Abnormal auditory adaptation* is a decline in discharge frequency with time, observed following an initial burst of neural activity in response to an adequate continuing stimulus applied to the Organ of Corti. This phenomenon is characteristic of VIIIth nerve and brainstem auditory dysfunction.

Puretone audiometry is the most widely available, subjective, quantitative test of auditory thresholds. Electronically generated pure tones are delivered by earphones and the subject is required to respond to the quietest tone, at given frequencies between 125 and 8000 Hz in each ear. The sound may be delivered by air conduction (**AC**) or, if the tones are delivered by a bone vibrator on the mastoid process, by bone conduction (**BC**). In the latter test condition, because the intra-aural attenuation for a bone-conducted sound is negligible, masking of the ear not under test with narrow-band noise is mandatory. Bone-conduction thresholds significantly better than air-conduction thresholds indicate a conductive hearing loss, whereas similar bone-conduction and air-conduction thresholds are characteristic of sensorineural hearing loss (Fig. 4).

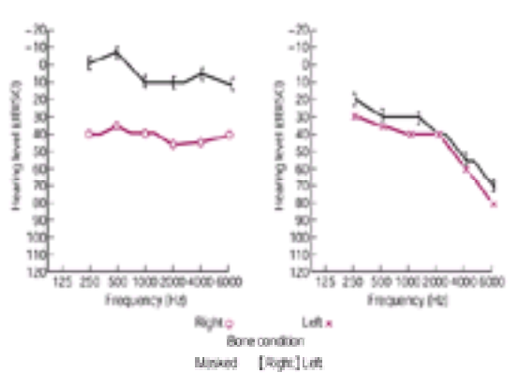

Fig. 4 Pure tone audiograms showing both air- and bone-conduction thresholds and illustrating a right conductive hearing loss (A) and a left sensorineural hearing loss (B).

The stapedius muscle in the middle ear contracts bilaterally in response to loud sound directed into either ear. Using an impedance bridge, the minimum intensity of sound at a given frequency required to produce contraction of the stapedius muscle and thus a movement of the tympanic membrane can be measured (the acoustic reflex threshold). This objective measure enables recruitment and abnormal auditory adaptation to be measured, and allows assessment of middle ear, cochlear, VIIIth nerve, and brainstem auditory function.

Otoacoustic emissions are weak signals that can be recorded in the ear canal and are the result of contractile properties of the outer hair cells of the cochlea. Measurement of otoacoustic emissions thus provides objective information about cochlear function.

Speech audiometry is a subjective test requiring the subject to repeat standard lists of words delivered at varying intensities through headphones. The responses are scored and provide an assessment of auditory discrimination. They are of particular value in assessing the efficacy of hearing-aid provision.

Electrophysiological tests provide the major objective means of assessing auditory function and siting pathology in the auditory system. Electrocochleography enables the measurement of the electrical output of the cochlea and VIIIth cranial nerve in response to an auditory stimulus, while brainstem auditory evoked responses are of particular value in discriminating between cochlear and VIIIth nerve dysfunction. Recordings are obtained by averaging a series of time-locked responses generated by the major processing centres of the auditory system in response to a repetitive sound stimulus (Fig. 5). Analysis of the waveform must be undertaken in conjunction with knowledge of the puretone thresholds if appropriate and valid conclusions about auditory function are to be obtained . Cortical or late-evoked auditory responses are the most effective method of defining auditory threshold at each frequency in an uncooperative patient, and are essential in legal cases in which a non-organic loss should always be excluded.

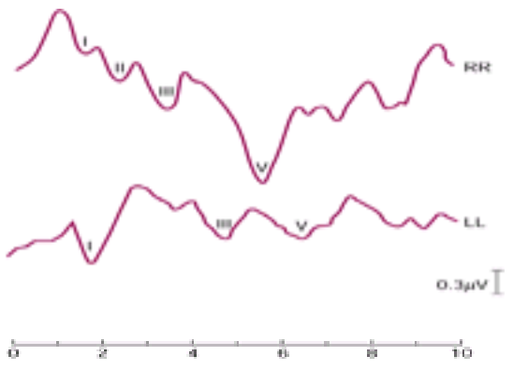

Fig. 5 Illustration of auditory evoked brainstem responses showing normal waves I, II, III, and V from the right ear (RR) and delayed waves III and V from the left ear (LL) in a case of a left acoustic neurinoma.

Management

Appropriate management requires a detailed history and examination to ensure both appropriate management of related general medical conditions and protection from leisure (discotheques) and occupational noise hazards and ototoxic drugs. Auditory rehabilitation is a problem-solving exercise centred on each individual patient, and depends on assessing both the auditory disability of the individual and the relevance of this to other important people in the patient's life. Not only auditory impairment, but also communication skills, including lip-reading ability, the use of visual cues, and the level of speech and language, together with psychological and sociological factors must be considered.

The remedial process may be straightforward in a highly motivated patient in whom there is an uncomplicated hearing loss. However, in the presence of a complicating factor, such as a hearing loss that is difficult to aid or arthritis making hearing aid manipulation difficult, the particular problem must be addressed to ensure optimal use of subsequent hearing-aid provision. In patients who have a negative view of hearing aids, environmental aids and instruction in communication skills prior to the introduction of a hearing aid may facilitate long-term rehabilitation. In general, the provision of a hearing aid is only effective when the patient himself, rather than well-meaning family members, wishes to pursue matters.

Although hearing aids play a pivotal role in audiological rehabilitation, a detailed description of their provision and selection is outside the scope of this short review. For many patients, wearable hearing aids, which bring sound more effectively to the ear, are invaluable, but environmental aids (assisted-listening devices such as amplification systems, alerting warning devices—for example, flashing lights connected to a doorbell or an alarm clock), may be adequate. In addition, sensory substitution systems, for example where visual signals are generated in response to auditory cues such as a telephone or doorbell ringing, or a baby crying, may be helpful to a hearing impaired person.

The general principles of hearing-aid provision include the fitting of a comfortable earmould which provides a secure mounting for the aid and a good acoustic connection between the aid and the ear canal. Hearing-aid selection involves matching the amplification required from the aid at specific frequencies with that required by the user. A particular disability experienced in most hearing-impaired subjects is that of hearing speech in a noisy environment and, although programmable digital processing hearing aids are of some help in this situation, conventional aids provide selective amplification of the frequencies relevant to speech, with minimal amplification at the peak frequency of background noise. Conventional aids may be divided into body-worn and head-worn aids, which can be in spectacles, behind the ear, in the ear, or in the canal in design. The major advantage of body-worn aids is the very high gain and maximum output that can be achieved, whereas the disadvantage is the unsightly nature of the device and the poor microphone placement.

Cochlear implants are electronic devices that convert sound into electrical current for the purpose of directly stimulating residual auditory nerve fibres to produce hearing sensations. The devices are implanted in the cochlea, usually with an electrode array, with an externally worn microphone and processor. Cochlear implants have been used in totally deafened adults and children with good results, and should be considered in all cases of profound acquired hearing loss and in children in whom there is good evidence of auditory nerve preservation in both congenital and acquired hearing impairment.

The value of counselling for the hearing-impaired person by a skilled hearing therapist must be emphasized. Such simple hearing tactics as encouraging the individual to ensure that the light is always on the speaker's face, that he or she places himself so that the better ear is towards the speaker, and sitting close to the sound source thereby minimizing background noise, can greatly improve communication ability. For the profoundly hearing-impaired, psychological problems associated with isolation are significant and it is therefore essential that psychological, medical, and social support are readily available.

Tinnitus

Tinnitus may be defined as a perception of sound, which originates from within the head rather than from within the external world. Rarely, the sound may have an externally detectable component and is then termed 'objective tinnitus' as opposed to the more common 'subjective tinnitus.' The experience of tinnitus is universal, but the complaint of tinnitus is rare.

Many conditions are associated with tinnitus, but it is frequently, although not always, associated with hearing impairment. The proposed pathophysiological mechanisms include:

- decoupling of the stereocilia of the hair cells;
- misinterpretation of auditory neural activity by higher auditory centres;
- self-sustaining oscillation of the basilar membrane;
- spontaneous otoacoustic emissions;
- an abnormality of the spontaneous resting activity of primary auditory nerve fibres, either secondary to the hypo- or hyperexcitability of damaged hair cells or as a direct consequence of the derangement of primary neurones themselves;
- damage to the myelin sheath between auditory nerve fibres allowing ephaptic transmission (cross-talk) between adjacent nerve fibres; and
- derangement of efferent fibres of the vestibulocochlear nerve, producing aberrant auditory behaviour.

A number of studies have demonstrated that tinnitus complaint does not correlate with psychoacoustic features of the tinnitus, but there is a significant correlation between tinnitus complaint and psychological symptoms. Importantly, the onset of tinnitus complaint may be associated with negative life events such as retirement, redundancy, bereavement, and divorce.

The assessment of tinnitus includes a detailed history, clinical examination, and audiometric investigation as outlined for hearing impairment. The commonest causes of objective tinnitus include palatal myoclonus, temporomandibular joint abnormalities, vascular abnormalities such as and arteriovenous fistula, and vascular bruits. Rarely, a patulous auditory tube may give rise to tinnitus in which the patient complains of a blowing sound associated with respiration.

Bilateral subjective tinnitus with evidence of a cochlear hearing loss is associated most commonly with presbyacusis, endolymphatic hydrops, vascular labyrinthine lesions, and noise-induced hearing loss. However, it is also common with head injury, whiplash injury, ototoxicity, barotrauma, surgical intervention, and after such simple clinical practices as syringing. Unilateral subjective tinnitus, with or without an associated sensorineural hearing loss. must be fully investigated to exclude an underlying cerebellopontine angle lesion, in particular an acoustic neurinoma.

Management

The primary management of tinnitus is medical, although surgical intervention is required for the correction of arterial stenoses giving rise to bruits and for glomus jugulari tumours and arteriovenous malformations. Destructive surgery, for example labyrinthectomy or auditory nerve section, has no place in the management of

tinnitus as there is no evidence that destruction of the peripheral cochlear elements brings about improvements in tinnitus complaint.

The medical management of tinnitus can be divided into psychological, pharmacological, and prosthetic intervention.

The psychological aspects of tinnitus management include an explanation of tinnitus, reassurance that the symptom will not progressively deteriorate or indeed remain unchanged, the exclusion of sinister pathology to allay fear, and, if necessary, the appropriate formal psychiatric management of depression/anxiety.

In the presence of a hearing impairment, the provision of hearing aids to 'mask' tinnitus with desirable environmental noise may be of value. In the absence of such a loss, tinnitus maskers and noise generators have been advocated to promote 'adaptation', but it must be emphasized that there is no hard evidence that tinnitus maskers are superior to placebo devices.

Pharmacologically, intravenous lidocaine (lignocaine) has been shown to result in the disappearance or amelioration of tinnitus, but no oral preparation has been found to be equally effective. Psychiatric drugs may be required for psychological management, although no single drug has been shown to be uniformly effective.

Tinnitus retraining therapy is a management strategy based on a neurophysiological model of tinnitus. The retraining is a combination of prosthetic and psychological intervention, which in essence provides a structured framework for the various well-established mechanisms of tinnitus management outlined above.

In conclusion, positive reassurance, appropriate psychiatric management, and prosthetic support remain the mainstays of the medical management of tinnitus.

Further reading

Ludman H, Wright T, eds (1998). *Diseases of the ear*, 6th edn. Arnold, London.

Martini A, Prosser S (2002). Disorders of the inner ear in adults. In: Luxon LM, *et al.*, eds. *A textbook of audiological medicine*. Taylor and Francis, London.

24.13.1 The unconscious patient

David Bates

It is important to distinguish transient unconsciousness occurring with syncope, seizures, cardiac arrhythmias, or metabolic abnormalities from unconsciousness that persists and is coma. Prolonged loss of consciousness is seen commonly in three clinical situations: following head injury, after an overdose of sedating drugs, or in the situation of 'non-traumatic coma' where there are many possible diagnoses but the most common are anoxia, ischaemia, systemic infection, and metabolic derangement. It is important that the physician asked to see a patient in 'non-traumatic coma' should remember that the patient may be harbouring delayed effects of trauma such as subdural haematoma or meningitis arising from a basal skull fracture. The possibility of raised intracranial pressure following a parenchymal haematoma in a patient with hypertension, the decompensation of a cerebral tumour, or the collection of pus means that all causes of loss of consciousness must be considered; in the diagnosis of medical coma it is not easy to exclude the possibility of head injury.

Urgent assessment of the patient in coma is required to identify and, where possible, correct the pathological cause, protect the brain from the development of irreversible damage, and identify those patients in whom the prognosis is hopeless. In this last group the institution and continuation of resuscitative measures is inappropriate and will serve only to prolong the anguish of relatives and carers.

Definition

Normal consciousness

Consciousness is the state of awareness of the self and the environment when provided with adequate stimuli; normal consciousness is exhibited by those patients who are fully responsive to stimuli and show appropriate behaviour and speech. Patients who are asleep can be roused and are then able to perform normally. Normal consciousness depends upon the integration of activity in the ascending reticular activating substance of the brainstem and the neuronal connections between areas of the cerebral cortex. The ascending reticular activating substance determines arousal, which is shown by awakening with eye opening, motor responses, and verbal communication. The content of consciousness, which is the combination of psychological responses to feeling, emotions, and mental activity, is mediated by the cerebral cortex. (Fig. 1)

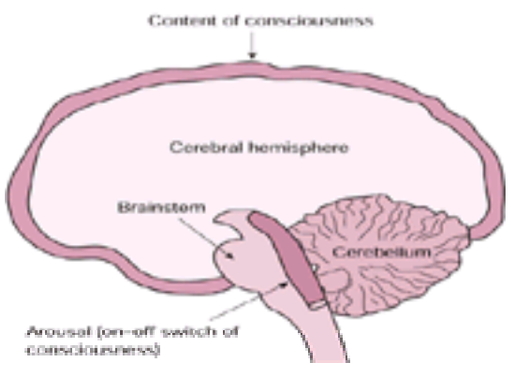

Fig. 1 Normal consciousness.

Coma

Coma is a state of unrousable unconsciousness without any psychologically understandable response to external stimuli or inner need. The patient may appear to be asleep but is incapable of responding normally to external stimuli other than by showing eye opening to pain, flexion or extension of the muscles in the limbs to pain, and occasionally grunting or groaning in response to painful stimuli. It occurs when there is damage to the ascending reticular activating substance or bilateral damage to areas of the cerebral hemispheres, or both (Fig. 2, Fig. 3, Fig. 4, and Fig. 5).

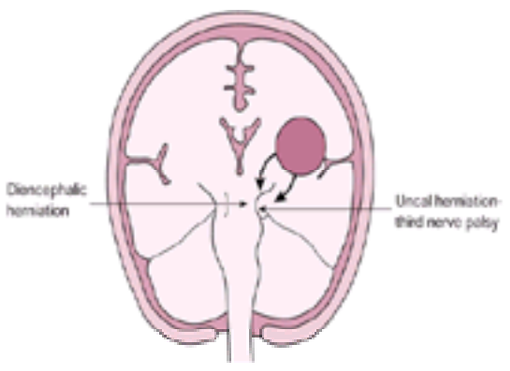

Fig. 2 Supratentorial mass.

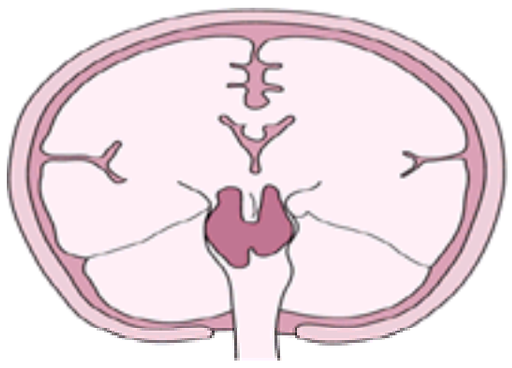

Fig. 3 Brainstem lesion—intrinsic.

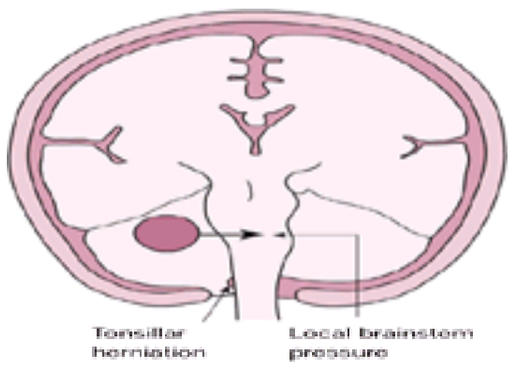

Fig. 4 Brainstem lesion—local pressure.

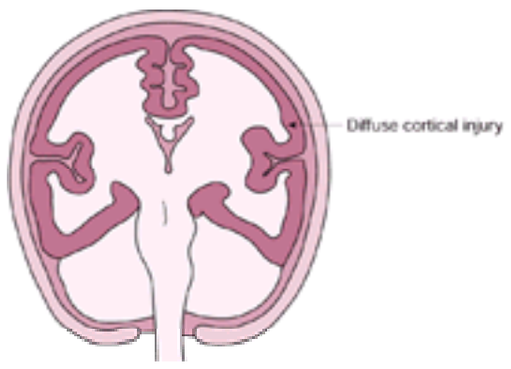

Fig. 5 Bihemispheric damage.

Confusion

Patients are usually disorientated with lowered attention, an inability to express thoughts, drowsiness, and defects in memory. There is a clouding of consciousness characterized by an impaired capacity to think, understand, respond to, and remember stimuli. It is important to differentiate acute confusion from dysphasia, amnesia, acute psychosis, severe depression, or dementia. Confusion is most commonly seen as the result of toxic or metabolic disturbances, particularly in the elderly.

Delirium

There is motor restlessness, hallucination, disorientation, and delusion. The patient is often frightened and irritable and the state can be regarded as a more profound example of confusion; both states should alert the doctor to impending coma. Delerium is most commonly seen in patients with toxic or metabolic disorders but can be mimicked by degenerative brain disease, acute psychosis, and hypomania.

Stupor

The patient appears to be asleep and will show little or no spontaneous activity, respond only to vigorous stimulation, then lapse back into somnolence. It may be difficult to differentiate stupor from catatonic schizophrenia or severe retarded depression, but in stupor due to organic disease the electroencephalogram will always be abnormal.

The vegetative state

The patient breathes spontaneously, has a stable circulation, and shows cycles of eye opening and eye closure which may simulate sleep and waking, but they are unaware of self and environment. It can be seen transiently in the recovery from coma or it may persist to death. This state is usually seen in patients with diffuse bilateral cerebral hemisphere disturbance with an intact brainstem, although it can occur with bilateral damage to the most rostral part of the brainstem. It is most commonly seen following head injury or as the result of hypoxic-ischaemic damage following cardiac arrest. The patient appears to be awake but is unaware, a condition that frequently causes distress to carers and relatives (Fig. 6).

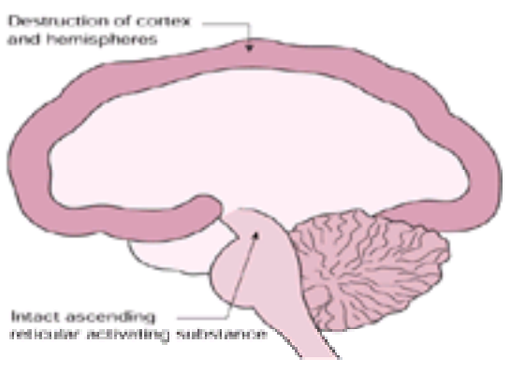

Fig. 6 Vegetative state.

The locked-in syndrome

Damage to the ventral portion of the pons below the level of the third nerve nuclei results in the rare condition of total paralysis of the limbs and lower cranial nerves but intact consciousness (Fig. 7). The patient can open, elevate, and depress the eyes but cannot move the eyes horizontally and there is no voluntary movement or speech. The diagnosis is made when the doctor recognizes that the patient is able to open the eyes voluntarily and allow them to close in response to command and can therefore respond to verbal and sensory stimuli by blinking. The commonest cause is infarction of the ventral pons, usually in a patient with hypertension, although it can also be seen with pontine tumours, multiple sclerosis, in central pontine myelin-olysis following profound hyponatraemia, and after head injury. The prognosis is poor although some patients recover, usually with residual spasticity. An electroencephalogram may help by showing an alert state, reactive to external stimuli and neurophysiology can be used to exclude similar incapacities occurring in myasthenia gravis or the Guillain–Barré syndrome.

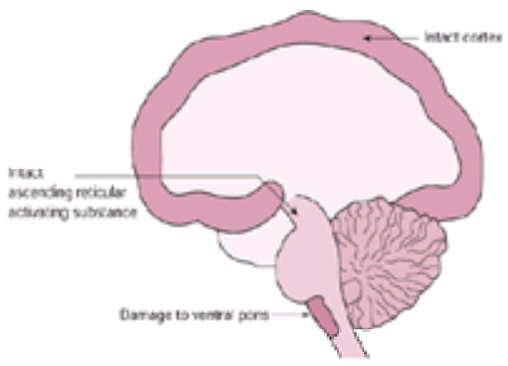

Fig. 7 Locked-in syndrome.

Psychogenic unresponsiveness

The term 'pseudocoma' or psychogenic unresponsiveness is used for patients who appear to be unconscious and in coma but who are not. The simplest way to identify this condition is to undertake oculovestibular testing (see below), which will reveal the presence of nystagmus and indicate that the patient has an intact brainstem and cortex.

The management of patients in coma

History

Once the patient is stable it is important to obtain as much information as possible from those who accompanied the patient to hospital or who observed the onset of coma. The circumstances in which consciousness is lost are of vital importance in helping to identify the diagnosis. Generally, coma is likely to present in one of three ways: as the predictable progression of an underlying illness; as an unpredictable event in a patient with a previously known disease; or as a totally unexpected event. In the first category are patients following focal brainstem infarction who deteriorate or those with known intracranial mass lesions who show similar deterioration. In the second category are patients with recognized cardiac arrhythmia or the known risk factor of sepsis from an intravenous cannula. In the final category it is important to distinguish whether there has been a previous history of seizures, trauma, febrile illness, or focal neurological disturbances. The history of a sudden collapse in the midst of a busy street or office indicates the need for different investigations from those required when the patient has been discovered at home in bed surrounded by empty bottles that previously contained sedative tablets. Where there is uncertainty, a telephone call to relatives and medical attendants may be useful.

Clinical assessment and examination

Estimation of the temperature, pulse, blood pressure, and respiratory rate, and examination of the skin, cardiovascular system, chest, and abdomen may often yield important clues in establishing the cause of a loss of consciousness. Fever, though not diagnostic, will usually indicate the presence of a systemic infection, meningitis, encephalitis, or abscess; seizures increases the likelihood of the latter two diagnoses. Hypothermia is most commonly seen following exposure to low environmental temperatures, intoxication with alcohol or barbiturates, the presence of peripheral circulatory failure, or profound myxoedema. Tachy- or bradyarrhythmias, evidence for valvular heart disease, or peripheral emboli raise the possibility of a cardiogenic cause; bruits over the carotid vessels suggest cerebrovascular disease; and splinter haemorrhages suggest endocarditis or collagen vascular disease. Hypotension raises the possibility of shock, myocardial infarction, or septicaemia and Addison's disease should be considered. Hypertension is less helpful as a clinical sign since it may be seen both as the result of cerebral insult or as an indicator of hypertensive encephalopathy.

The odour of the breath of an unconscious patient may indicate the presence of alcohol, a ketotic fetor raises the possibility of diabetes, and the fetor of hepatic or renal failure provide important clues. Clubbing of the finger nails suggests the possibility of a respiratory or gastrointestinal abnormality, and evidence of tracheal deviation, fluid in the chest, or collapse of the lung suggests the possibility of a respiratory cause. In the abdomen the finding of enlargement of an organ might indicate portal hypertension, polycystic kidneys, and an associated subarachnoid haemorrhage, or abnormality in the blood-forming organs. The general colour of the skin and mucous membranes might reveal anaemia, jaundice, cyanosis, or the pink discoloration of carbon monoxide poisoning. Purpura suggests a bleeding diathesis and bruising around the head indicates the possibility of trauma or a base of skull fracture. A rash may indicate an infective or inflammatory disease and hyperpigmentation raises the possibility of Addison's disease. The evidence of puncture wounds might identify an individual who is diabetic or a recreational drug user.

Neurological examination

This requires observation and an assessment of reflex responses. The position, posture, and spontaneous movements of the patient should be noted; the skull and spine should be examined with testing for neck stiffness and Kernig's sign to identify meningeal irritation. Ophthalmoscopy to identify papilloedema, fundal haemorrhages, emboli, and subhyaloid haemorrhages is important but it must be remembered that the absence of papill-oedema does not necessarily exclude raised intracranial pressure. The ears and fauces should be examined.

Level of consciousness

The level of consciousness must be documented by the initial observer and can then be monitored by medical and nursing staff to determine the progress of the patient and identify the need for further investigation, therapy, and decision. The most useful hierarchical grading scale to assess the level of consciousness is the Glasgow coma scale in which the patient's response to graded stimuli of eye opening, motor response, and verbal response are recorded (Table 1); all four limbs are observed for responses to pain and the best response is recorded, although asymmetry should be noted and may be important in identifying lateralization. The scale measures consciousness and it is possible to score gradations from the fully conscious patient (eye opening—4, motor response—6, verbal response—5) to the totally unresponsive patient (eye opening—1, motor response—1, verbal response—1). If the level of consciousness can be shown to be improving, then urgent decisions may be delayed, but if deterioration is occurring, it is imperative that a decision about management be made.

Brainstem function

The brainstem reflexes identify those lesions which affect the reticular activating substance and determine the viability of the patient. Most of the reflexes involve the eyes and the pattern of respiration, although the latter may be compromised by requirements of ventilation.

Pupillary reactions

Unilateral dilatation of a pupil with lack of a light response suggests an uncal herniation of the temporal lobe over the tentorium entrapping the third nerve, although it may also be seen with a posterior communicating artery aneurysm or other third nerve damage. Midbrain lesions typically cause loss of the light reflex with pupils in the mid-position, lesions in the pons cause small pupils with retained light responses, and fixed dilatation of the pupils suggests significant brainstem damage but must be differentiated from the fixed dilatation caused by atropine-like agents instilled by earlier observers. A Horner's syndrome may be seen with lesions in the hypothalamus or brainstem but can also be seen when diseases affect the wall of the carotid artery. Small pupils that react briskly to light raise the possibility of metabolic causes of coma such as hepatic or renal failure; drug intoxications tend not to affect the pupillary light responses.

Corneal responses

The corneal reflex is usually retained until very deep coma; if absent in a patient who appears to be otherwise in light coma, there is a distinct possibility that the cause may be drug intoxication. The loss of the corneal reflex in the absence of drug overdose is a poor prognostic indicator.

Spontaneous eye movements

Conjugate deviation of the eyes suggests a focal hemispheric or brainstem lesion, depression of the eyes is seen with damage to the midbrain at the level of the tectum, and skew deviation of the eyes suggests a lesion at the pontomedullary junction. Incoordinate eyes suggests damage to the ocular motor or abducent nerve in the brainstem or pathways, but a minor degree of divergence of the eyes is normal in the unconscious patient. Patients in light coma will often have normal roving eye movements, similar to those of sleep, which may be conjugate or dysconjugate. They cannot be mimicked and, when present, exclude the possibility of psychogenic unresponsiveness, when eye movements are likely to be more jerky.

Reflex eye movements are important in assessing brainstem activity. The oculocephalic response obtained by rotating the patient's head from side to side and observing the position of the eyes is likely to show Doll's eye movements when the brainstem is intact but the eyes will remain in the mid-position of the head when the brainstem is depressed. Oculovestibular testing is undertaken by the installation of 50 to 200 ml of ice-cold water into an external auditory meatus. The conscious patient, and those in psychogenic coma, will develop nystagmus with the quick phase away from the side of the stimulation indicating an active pons and intact corticopontine connections. A tonic response with conjugate movement of the eyes towards the stimulated side indicates an intact pons and suggests a supratentorial cause for the coma, whereas a dysconjugate response or no response at all implies a lesion within the brainstem.

Respiration

The techniques of ventilation limit the value of observation of respiration in patients with coma, but if testing is possible before respiration is controlled, then deep breathing suggests acidosis, regular shallow breathing is consistent with drug overdose, long-cycle Cheyne–Stokes respiration suggests damage at the level of the diencephalon, and short-cycle Cheyne–Stokes respiration damage at the level of the medulla. Central neurogenic hyperventilation occurs with lesions in the low midbrain and upper pons, and reflex responses such as yawning, vomiting, and hiccoughing may occur with brainstem disturbances.

Motor function

Motor function is assessed as part of the level of consciousness in the Glasgow coma scale, but lateralizing abnormalities are important and indicate the likelihood of a focal cause, although they may occasionally be seen in the context of hepatic encephalopathy or hypoglycaemia. The presence of generalized or focal seizures implies hemispheric damage and may help in lateralization; multifocal myoclonus suggests a metabolic or anoxic cause with diffuse cortical irritation.

Investigation

After performing the resuscitation, history, examination, and assessment the physician should identify one of the three following states (Table 2).

Coma with focal signs or evidence of head injury

In such patients, whether the focal signs indicate a brainstem or supratentorial problem, a CT scan or MRI should be undertaken. A normal scan may be seen in patients with hypoglycaemia or hepatic coma and the presence of structural pathology will be identified allowing a decision to be made about the indications for surgery or other therapy.

Coma with meningeal irritation but without focal signs

Such patients will most commonly have subarachnoid haemorrhage, acute meningitis, or meningoencephalitis as the cause of their coma. Brain imaging is the ideal investigation to identify the presence of subarachnoid blood and to exclude the possibility of focal collections. Depending upon the results of the scan a lumbar puncture can then be undertaken and may give diagnostic information. If the index of suspicion of meningitis is high then treatment should be commenced and, in the absence of focal signs or pupilloedema, lumbar puncture may precede imaging.

Coma without focal lateralizing neurological signs and without meningismus

Most patients will have suffered diffuse anoxic-ischaemic disease, metabolic derangement, or drug insult. It may be necessary to undertake imaging techniques but the probability of finding a focal abnormality is low and it is more likely that haematological or biochemical tests or a search for toxins in the blood will provide the diagnosis, or help identify an episode of ischaemia or hypoxia in the past. There may occasionally be an indication to undertake a lumbar puncture in such patients to exclude an inflammatory or infective cause. Patients who are in coma as the result of drug overdose will usually be identified from the history and the circumstances of discovery, but the possibility of drug-induced coma should always be considered in patients without focal signs and without meningism. A discrepancy between marked depression of brainstem responses in a patient who appears to be in relatively light coma suggests the diagnosis, the importance of which is that such patients have a good prognosis provided that they are given adequate respiratory and circulatory support during the coma.

Prognosis

The prognosis of individual patients depends upon the aetiology, the depth of the coma, the duration of the coma, and certain clinical signs.

Aetiology

Following head injury, prognosis is dependent upon the presence of intracranial haematoma, the age of the patient, and the severity of systemic injury and its effects. Patients in coma following drug overdose have, in general, a good prognosis provided that they are adequately resuscitated and protected. Patients who are in coma as a result of causes other than head injury or drug overdose for a period of more than 6 h have only a 10 per cent chance of making a good recovery. Those who have suffered subarachnoid haemorrhage or stroke have a less than 5 per cent chance of making such a recovery and those with hypoxic or ischaemic injury, typically following cardiac arrest, about 10 per cent. Those with metabolic or infective causes have almost a 30 per cent change of making a good recovery. A vegetative state is most likely to occur after head injury or hypoxic-ischaemic damage.

Depth of coma

Patients with no response to eye opening, no focal response to pain, and a poor response to pain have a poorer outcome than those who respond with eye opening, grunting, and flexion of the limbs.

Duration of the coma

When patients have been in coma for 6 h about 12 per cent may make a good recovery, those who remain in coma for 24 h have only a 10 per cent chance of recovery, and at the end of a week only 3 per cent of patients can be expected to make a good recovery. In general, patients who remain in coma for more than 7 to 14 days either die or enter a continuing vegetative state.

Clinical signs

Brainstem reflexes are the most important clinical signs in defining prognosis; the absence of corneal or pupillary reflexes or of oculovestibular responses for 24 h, in the absence of sedative drugs, is almost incompatible with recovery to independence whatever the cause of coma. Most brainstem reflexes are useful indicators of a poor prognosis but some, such as the development of nystagmus and oculovestibular testing or vocalization of any recognizable word within 48 h, identify patients with a good chance of recovery.

Continuation of care

The long-term care of patients in coma may be undertaken in an intensive care unit, on a specialist ward, in a rehabilitation unit, or long-stay hospital. It is important that those in whom prognosis is hopeless should not be permanently exposed to the rigors of intensive care medicine but should continue to receive basic care within routine hospital wards or a more long-stay environment. So long as patients are considered to have a potential for recovery they should be looked after in an intensive care unit or in a specialist ward. Their respiration, skin, circulation, and bladder and bowel function need attention, seizures must be controlled, and the level of consciousness should be regularly assessed and monitored. It is important that the mobility of joints and circulation to pressure areas are maintained during the long-term care of the patient and the possibility of aspiration pneumonia, peptic ulceration, and other complications of long-term intensive care need to be avoided. Techniques such as mechanical ventilation and steroid therapy should not be used routinely in the management of comatose patients; they do not improve prognosis and may compromise recovery. Investigations are of little help in identifying long-term prognosis because various types of electroencephalogram pattern have been recorded from patients in prolonged coma and CT scans simply show cortical atrophy with ventricular dilatation. Some somatosensory-evoked responses have been reported to show loss of the cortical component in long-term unconsciousness and positron emission tomography (PET) is reported to show metabolic underactivity, but at present, neither test can provide decisive information as to prognosis.

Further reading

Bates D (1991). Defining prognosis in medical coma. *Journal of Neurology, Neurosurgery and Psychiatry* **54**, 569–71.

Bates D (1993). The management of medical coma. *Journal of Neurology, Neurosurgery and Psychiatry* **56**, 589–98.

Fisher CM (1969). The neurological examination of the comatose patient. *Acta Neurologica Scandinavica* **45**(Suppl 46), 1–56.

Plum F, Posner JB (1980). *The diagnosis of stupor and coma*, 3rd edn. Davis, Philadelphia.

Teasdale G, Jennett WB (1974). Assessment of coma and impaired consciousness: a practical scale. *Lancet* **ii**, 81–4.

24.13.2 Headache

Peter J. Goadsby

Headache is perhaps the commonest of human maladies, and while no less interesting, varied, or biologically based than some of its neurological cousins it has been the subject of less attention than its clinical load demands. If only for cynical reasons of practicality, the general reader will need some basis with which to approach the patient with headache. Diagnosis and management of headache is clinically based, offering the doctor the opportunity to be a physician not a filter for test results, with the chance to treat and improve symptoms. Moreover, there is a sufficient biological basis now for headache to satisfy even the most scientific of inquisitors. Here the principles will be set out: the secrets and enjoyment remain, as with anything truly medical, in the clinic.

General principles

A formal classification system exists for headache, and it might surprise the casual observer that this runs to nearly 100 pages. The International Headache Society system is explicit in the sense that it uses features of the headache to make the diagnosis, summing features to make the diagnosis more certain. In clinical practice a broad categorization that serves well, and is consistent with the International Headache Society system, is the concept that there are primary and secondary forms of headache. Such a system is outlined in Table 1. Primary headaches are those in which headache and its associated features are the disease in themselves, and secondary headaches are those caused exogenously, such as headache associated with fever. Mild secondary headache, such as that seen in association with upper respiratory tract infections, is common but only rarely worrisome. The clinical dilemma remains that while life-threatening headache is relatively uncommon in Western society, it is present and requires suitable vigilance by doctors. Primary headache, in contrast, while not life-threatening is often disabling over time.

Secondary headache

Key clinical features of secondary headache

There are certain issues which are vital to establish in the patient presenting with any form of head pain so that important secondary headaches will not be missed. Perhaps the most crucial is the length of the history. Patients with a short history require prompt attention and may require quick investigation and management, whereas patients with a longer history generally require time and patience rather than speedy consultation. There are some important general features, including associated fever or sudden onset of pain (Table 2), and these demand attention. Unless a benign diagnosis can be positively established, patients with a history of headache of recent onset or neurological signs need referral for specialist neurological assessment. A similar rule can be applied to computed tomography (**CT**) or magnetic resonance imaging (**MRI**). Patients with a history of recurrent headache over a period of a year or more, fulfilling International Headache Society criteria for migraine (Table 3) and with a normal physical examination, have positive brain imaging in only about one in a 1000 images. It should be noted that brain tumour is a relatively rare cause of headache, and rarely a cause of isolated long-term histories of headache. The management of secondary headache is generally self-evident—treatment of the underlying condition such as an infection or mass lesion. An exception is the condition of chronic post-traumatic headache in which pain persists for long periods after head injury. This is an interesting generic problem which may be seen after central nervous system infection, trauma, both blunt and surgical, intracranial bleeds, and other precipitants. While the syndrome is generally self-limiting up to 3 to 5 years after the event, it may require treatment of the headache (see Chronic Daily Headache below).

Primary headache syndromes

The primary headaches are a group of fascinating syndromes in which headache and associated features are seen in the absence of any exogenous cause. The common syndromes (Table 1) are tension-type headache, migraine, and cluster headache and the collection of headaches known as primary chronic daily, or frequent, headache. Some other less well known syndromes will be mentioned because they are easily treated when diagnosed.

Pathophysiology of headache

Understanding of headache has advanced considerably over the last decade. The severe primary headaches—migraine and cluster headache—have been studied extensively. In experimental animals the detailed anatomy of the connections of the pain-producing intracranial extracerebral vessels and the dura mater has built on the classical human observations that it is these structures, and not the brain, that are responsible for generating pain from within the head. The key structures involved in the nociceptive process are:

- the large intracranial vessels and dura mater;
- the peripheral terminals of the trigeminal nerve that innervate these structures;
- the central terminals and second-order neurones of the trigeminal nucleus.

Together these structures are known as the trigeminovascular system. The cranial parasympathetic autonomic innervation provides the basis for symptoms such as lacrimation and nasal stuffiness that are prominent in cluster headache and paroxysmal hemicrania, and which may also be seen in migraine. It is clear from human functional imaging studies that vascular changes in migraine and cluster headache are driven by these neural vasodilator systems so that these headaches should be regarded as neurovascular. The concept of a primary vascular headache should be abandoned since it neither explains the pathogenesis of what are complex central nervous system disorders nor necessarily predicts treatment outcomes.

Migraine is an episodic syndrome of headache with sensory sensitivity, such as to light, sound, and head movement, probably due to malfunction of aminergic brainstem/diencephalic sensory control systems (Fig. 1). The first of the migraine genes has been identified for familial hemiplegic migraine, in which about 50 per cent of families have mutations in the gene for the a_1 subunit of the neuronal P/Q voltage-gated calcium channel. This finding, together with the clinical features of migraine, suggests that it might be part of the spectrum of diseases known as channelopathies—disorders involving malfunction of voltage-gated channels. Functional neuroimaging has suggested that brainstem regions in migraine, and the posterior hypothalamic grey matter, site of the human circadian pacemaker cells of the suprachiasmatic nucleus in cluster headache, are good candidates for specific involvement in primary headache.

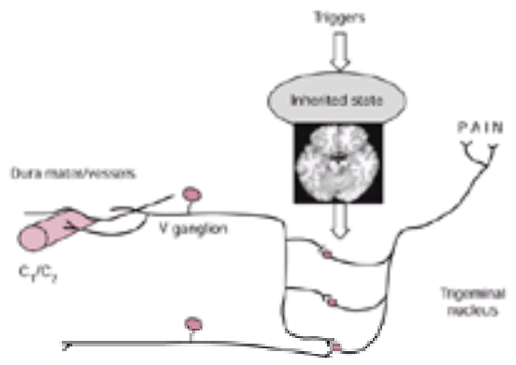

Fig. 1 Illustration of the some elements of migraine biology. Patients inherit a malfunction in brain control systems for pain and other afferent stimuli which can be triggered and are in turn capable of activating the trigeminovascular system as the initiating event in a positive feedback of neurally driven vasodilatation. The trigeminal innervation of pain-producing intracranial structures—dura mater and blood vessels—passes through the trigeminal ganglion (V ganglion) to terminate in the most caudal part of the trigeminal nucleus. Cervical inputs which terminate in the trigeminocervical complex accounts for the non-trigeminal distribution of pain in many patients. Migraine thus has a pain system for its expression and brain centres—modulatory systems—which define the associated symptoms and periodicity of the clinical syndrome. (Brainstem changes after Weiller and colleagues.)

Migraine

Migraine is generally an episodic headache with certain associated features, such as sensitivity to light, sound, or movement, and often with nausea or vomiting accompanying the headache (Table 3). None of the features is obligatory, and indeed given that the migraine aura, visual disturbances with flashing lights or zig-zag lines moving across the visual field or other neurological symptoms, is reported regularly in only about 15 per cent of patients, a high index of suspicion is required to diagnose migraine. A headache diary can often be helpful in making the diagnosis but perhaps more so in measuring the burden of the disease to the individual and then observing the effects of treatment. At a minimum the diary would mark on a calendar each day with headache, the length of the attack, what medication was taken and the doses, and what life events may have been taking place, such as a menstrual period. In differentiating the two main primary headache syndromes seen in clinical practice, migraine at its simplest level is headache with associated features, and tension-type headache is headache that is featureless (see below). Useful rule in practice is that most disabling headache is probably migrainous in biology. As for management, it is preferable to misdiagnose tension-type headache as migraine as opposed to the reverse, since there is so much good that can be done for migraine sufferers and so little for tension-type headache that patients may gain by such a diagnostic bias.

If headache with associated features describes migraine attacks, then 'headachy' describes the migraine sufferer over their lifetime. The migraine sufferer inherits a tendency to have headache that is amplified at various times by their interaction with their environment, the much discussed triggers. The brain of the 'migraineur' seems more sensitive to sensory stimuli and to change; and this tendency is even more notably amplified in females over the course of their menstrual cycle. The migraine sufferer does not habituate to sensory stimuli easily and so is often and adversely stimulated in the world in which they live and work. Migraine sufferers may have headache when they sleep in, when they are tired, when they skip meals, when they are stressed, or when they relax. They are less tolerant to change and part of successful management is to advise them to maintain regularity in their lives in the knowledge of this fluctuating brain sensitivity.

It has been said that migraine can never occur daily, but few biological phenomena respect absolute rules. This author takes the view that there is a very distinct syndrome of Chronic Migraine that is simply the most severe end of a complex phenomenon and often requires referral for specialist advice. Chronic Migraine is part of the group of headaches known as Chronic Daily Headache (see below) whose final nosology will only be settled when there are clearer biologically based development of disease markers. After making a diagnosis the next step in the clinical process is to be sure that the burden of the condition has been understood: how much headache does the patient have and, more important, what can't the patient do; what is their degree of disability? One can ask the patient directly to get a flavour, obtain a diary, or get a quick but accurate estimate using the Migraine Disability Assessment Scale (MIDAS), which is well validated and very easy to use.

Management of migraine

After diagnosis, the management of migraine begins by an explanation of certain things to the patient, notably:

- Migraine is an inherited tendency to headache, and cannot be cured.
- Migraine can be modified and controlled by the adjustment of lifestyle factors and the use of medicines.
- Migraine is not life-threatening nor associated with serious illness, the exception being in females who smoke and receive oestrogenic oral contraceptives, but migraine can make life a misery.
- Migraine management takes time and co-operation when information, such as that from a headache diary, has to be collected or inquiry made concerning the effect of the disease on the patient's life: the disability accrued to the disease.

Non-pharmacological management

Non-pharmacological management of migraine involves helping the patient identify things that aggravate the problem and encouraging them to modify these. Many patients will not find any joy in this approach; they should not be disparaged for this, but for those who do identify such factors, it will be a rewarding strategy. The crucial lifestyle advice is to explain to the patient that migraine is a state of sensitivity of the brain to change. This implies that the migraine sufferer needs to regulate their life with a healthy diet, regular exercise, regular sleep patterns, avoiding excess caffeine and alcohol and, as far as practical, modifying or minimizing changes in stress. A balanced life with fewer extremes will benefit most migraine sufferers. Patients also need to know that the brain sensitivity that is migraine varies, so that triggers will vary in their likelihood of resulting in headache.

Preventative treatments for migraine

The decision to start prophylactic treatment requires consideration from both doctor and patient. The basis for considering preventative treatment from a medical viewpoint is a combination of frequency of acute attacks and the tractability of these attacks. Attacks that are unresponsive to medications for acute management are targets for prevention, while attacks which are simply treated may be less obvious candidates for prevention. The other part of the equation relates to what is happening with time. If a patient diary shows a clear trend to increased frequency it is better to get in early with prevention than wait for the problem to become chronic. A simple rule for frequency might be that for one to two headaches a month there is usually no need to start preventative treatment; for three to four discussion may be needed; and for five or more a month prevention should be considered. Options available for treatment are covered in detail in Table 4 and vary somewhat by country. The problem with preventative treatments is not that there are none, but that they were all developed for other conditions. Often the doses required to reduce headache frequency produce marked and intolerable side-effects. It is not clear how preventatives work, although it seems likely that they modify the brain sensitivity that underlies migraine. Generally each drug should be started at a low dose and gradually increased to a reasonable maximum to determine if there is going to be a useful clinical effect.

Treatments for acute attacks of migraine

Treatments for acute attacks of migraine can be usefully divided into disease non-specific treatments—analgesics and non-steroidal anti-inflammatory drugs—and disease specific treatments—ergot-related compounds and triptans (Table 5). It must be said at the outset that most medications for acute attacks seem to have a propensity to aggravate headache frequency and induce a state of refractory daily or near-daily headache—analgesic-associated Chronic Daily Headache (see below). Codeine-containing compound analgesics are a particularly pernicious problem when available in over-the-counter preparations; the author recommends avoiding their frequent (more than twice a week) use. Many patients who stop taking regular analgesics will have no change to their headache, although some will have a distinct reduction in their headache frequency. Almost all who reduce acute attack medication overuse feel in some way better and will be easier to treat with standard preventatives.

Treatment strategies

Given the array of options for controlling an acute attack of migraine, how does one start? The simplest approach to treatment has been described as 'stepped care'. In this model all patients are treated, assuming no contraindications, with the simplest treatment, such as aspirin 900 mg with an antiemetic. Aspirin is an effective strategy (as proven by double-blind controlled clinical trials) and is best used in its most soluble formulation. The alternative would be a strategy known as 'stratified care', by which the physician determines, or stratifies, treatment at the start based on the likelihood of response to levels of care. An intermediate option may be described as stratified care by attack. The latter is what many headache authorities suggest and what patients often do when they have the options. Patients use simpler options for their less severe attacks relying on more potent options when their attacks or circumstances demand them (Table 6).

Non-specific treatments for acute attacks

Since simple analgesics such as aspirin and paracetamol (acetaminophen) are cheap and can be very effective, they can be employed in many patients. Dosages should be adequate, and the addition of domperidone (10 mg orally) or metoclopramide (10 mg orally) can be very helpful. Non-steroidal anti-inflammatory drugs can very useful when tolerated. Their success is often limited by inappropriate dosing, and adequate doses of naproxen (500 to 1000 mg orally or *per rectum*, with an antiemetic), ibuprofen (400 to 800 mg orally), or tolfenamic acid (200 mg orally) can be extremely effective. Tolfenamic acid has been shown in a double-blind placebo-controlled study to have comparable efficacy to sumatriptan 100 mg, a result that reinforces the general clinical view that non-steroidal anti-inflammatory drugs can be very useful compounds in treating migraine.

Specific treatments for acute attacks

When simple measures fail, or more aggressive treatment is required, the specific treatments are required. While ergotamine remains a useful antimigraine compound, its place as the treatment of choice has slipped in recent years. There are particular situations in which ergotamine is very useful, but its use must be strictly controlled as ergotamine overuse itself produces severe headache and a host of vascular problems. The triptans have revolutionized the life of many patients with migraine and are clearly the most powerful option available to stop a migraine attack. They can be rationally applied by considering their pharmacological, physicochemical, and pharmacokinetic features, as well as the formulations that are available.

Tension-Type Headache

As its name suggests Tension-Type Headache (TTH) is a term that describes the headache form most seeking understanding. One might challenge the reader to define the essence of TTH, which eludes this author, or consider for a moment how hard it is to study something that is commonly considered to be well understood. TTH has two forms, episodic TTH, where attacks occur on less than 15 days a month and chronic TTH where attacks, on average over time, are seen on 15 days or more a month. The latter is part of the broader clinical syndrome of Chronic Daily Headache (see below), but the terms are not equal.

Clinical features

TTH has been defined by the International Headache Society both for its episodic and chronic forms, but by the time this chapter is read that definition will have changed. In the initial classification admixtures of nausea, photophobia, or phonophobia in various limited combinations, without clear biological rationale, were permitted in either the episodic or chronic form of TTH. These are being removed as the classification is being revised. A useful clinical approach is to diagnose TTH when the headache is completely featureless: no nausea, no vomiting, no photophobia, no phonophobia, no osmophobia, no throbbing, and no aggravation with movement. Such an approach neatly divides migraine, which has one or more of these features and is the main differential diagnosis, from TTH. For research I would further divide up the patients with attacks of a TTH phenotype who have migraine at other times, a family history of migraine, migrainous illnesses of childhood, or typical migraine triggers to their attacks, to try and understand what the TTH biology alone imparts to the sufferer.

Pathophysiology

The pathophysiology of TTH is incompletely understood. This results from the fact that the name implies to most that it is a product of nervous tension, for which there is no clear evidence, and the definitions employed have undoubtedly admitted patients with migraine to the studies. It seems likely that TTH will be due to a primary disorder of central nervous system pain modulation, to contrast with migraine which is a much more generalized disturbance of sensory modulation. There are data suggesting a genetic contribution to TTH but one must question these since they applied the current, faulty, diagnostic criteria.

Management

Adopting the clinical approach to TTH outlined above results in diagnosing a headache form that is usually less disabling, and more in the category of irritating. Its episodic form is generally amenable to simple analgesics, paracetamol (acetaminophen), aspirin, or NSAIDs, which can be purchased over the counter. There are clear clinical studies to demonstrate that triptans in TTH alone are not helpful, although germane to the above discussion, triptans are effective in TTH where the patient also has migraine. For chronic TTH amitriptyline is the only treatment with a clear evidence base; the other tricylics, selective serotonin reuptake inhibitors, or the benzodiazepines have not been shown in controlled trials to be effective. Similarly, there is no controlled evidence for the use of EMG biofeedback, relaxation therapy, or acupuncture, and both positive and negative studies using botulinum toxin. At the time of writing botulinum toxin is regarded as experimental. Stress management has been shown to be an effective approach in a controlled trial.

Cluster headache

Cluster headache is a rare form of primary headache with a population frequency of 0.1 per cent. Most standard textbooks cover the topic and the reading list contains specialized books on cluster headache. Cluster headache is part of a spectrum of primary headaches with prominent cranial autonomic activation, lacrimation, conjunctival injection or rhinorrhoea, collectively known as the Trigeminal-Autonomic Cephalgias (TACs). Cluster headache is probably the most painful condition known to humans; of more than 500 patients on our data base we are yet to talk with one who has had a more painful experience, including childbirth, severe burns and multiple limb fracture. A neurologist should manage cluster headache, if possible. Its core feature is periodicity, be it circadian or in terms of active and inactive bouts over weeks and months (Table 7). The typical cluster headache patient is male (male:female ratio 3:1) who has one to two attacks of unilateral pain of relatively short duration (30 to 180 minutes) every day for bouts of 8 to 10 weeks a year. Sufferers are generally perfectly well between times. Patients with cluster headache tend to move about during attacks, pacing, rocking, or even rubbing their head for relief. The pain is usually retro-orbital, boring, and very severe. It is associated with a red or watering eye, the nose running or blocking, and eyelid droop, the cranial autonomic symptoms, on the same side as the pain. Cluster headache is likely to be a disorder involving central pacemaker regions of the posterior hypothalamus, which is likely to share much with the other TACs but may be usually differentiated on clinical grounds from them (Table 8).

Management of cluster headache

Cluster headache is managed using treatments for acute attack and preventative agents. Treatments for acute attacks are usually required by all cluster headache patients at some time, while preventatives can almost be life-saving for those patients with chronic cluster headache and are often needed to shorten the active periods.

Preventative treatments

The options for preventative treatment in cluster headache are a little different depending on whether the patient has the episodic or chronic variety of the condition (Table 9). Most experts would now favour verapamil as the first-line preventative treatment, although for some patients with the episodic variety and short bouts limited courses of oral corticosteroids or methysergide can be very useful.

Verapamil has been suggested as a useful option for the last decade and compares favourably with lithium. What has clearly emerged from clinical practice is the need to use higher doses than had initially been considered and certainly higher than those used in cardiological indications. Although most patients will start on doses as low as 40 mg twice daily, doses of up to 960 mg daily and beyond are now employed. Side-effects, such as constipation and swelling of the legs, can be a

problem, but the issue of cardiovascular safety is more difficult. Verapamil can cause heart block by slowing conduction in the atrioventricular node, as demonstrated by prolongation of the A–H interval. Given that the PR interval on the ECG is made up of atrial conduction, A–H, and His bundle conduction, it may be difficult to monitor subtle early effects as verapamil dose is increased. This question needs study in this group of patients but at present it seems appropriate to do a baseline ECG and then repeat the ECG 10 days after a dose change, usually 80 mg increments, when doses exceed 240 mg daily.

Treatment of acute attacks

Attacks of cluster headache often peak rapidly and thus require a treatment with a quick onset. Many patients with acute cluster headache respond very well to treatment with oxygen inhalation. This should be given as 100 per cent oxygen at 8 to 12 litre/min for 15 to 20 min. It is important to have a high flow and a high oxygen content. Injectable sumatriptan has been a boon for many patients with cluster headache. It is effective, rapid in onset, and with no evidence of tachyphylaxis. Sumatriptan is not effective when given pre-emptively as 100 mg orally three times daily, although the nasal spray has now been shown to be effective in a placebo-controlled study.

Chronic Daily Headache

Daily headache gives the subspecialty a bad name but can be very rewarding when tackled clinically in a methodical manner. Chronic Daily Headache is not one entity but a collection of very different problems requiring different approaches to their management. Certainly not all daily headache is simply Tension-Type Headache (Table 10), and this is the commonest clinical mistake in headache diagnosis confusing the clinical phenotype with the headache biotype . The current definition for 'Daily Headache' requires pain on 15 or more days a month. Both terms are used here because the subject is under intense discussion. Population based estimates of Daily Headache are remarkable, demonstrating that 4.5 to 4.8 per cent of Western populations, notably in Spain and the United States, have daily or near daily headache. Daily Headache may again be primary or secondary, and it seems useful to consider the possibilities in this way when making decisions about clinical management (Table 10). It should be said that population based studies bear out the impression that many patients with refractory daily headache overuse various over-the-counter analgesic preparations.

Chronic daily headache and migraine

While it is widely accepted that chronic variants exist in some of the primary headaches— notably tension-type headache, cluster headache, and paroxysmal hemicrania—chronic migraine is a somewhat controversial entity in some quarters. Most authorities would agree that migraine may sometimes be chronic in terms of frequency but whether this occurs often or not is frequently argued. The issue of whether patients with frequent headache, some of which fulfils standard criteria for migraine and some for tension-type headache, have a single migrainous problem with two phenotypic manifestations is a very vexed one. Given that tension-type headache describes a phenomenon that is indistinct at best it seems unlikely that all such headaches will have a single underlying mechanism.

Considering the population based surveys quoted above, about two-thirds of daily headache patients have chronic tension-type headache and about one-third satisfy the Silberstein–Lipton criteria for 'transformed migraine' (now called Chronic Migraine). The philosophy behind Chronic Migraine is that some patients who inherit a migrainous predisposition end up with Chronic Daily Headache on a migrainous basis. The typical patient will have a dull daily often-featureless pain, punctuated by more severe attacks which would often, in isolation, fulfil standard criteria for migraine. This group is dominant in headache specialty clinics, with about 90 per cent of patients referred to headache clinics having transformed migraine, usually accompanied by overuse of analgesics. It might be that these patients have a more intractable organic problem which explains their over-representation in referral centres. If it is accepted that all other forms of primary headache have chronic counterparts, particularly the typically episodic primary headache, cluster headache, then having frequent migraine is not such a fanciful concept— it can then be called Chronic Migraine, by analogy with the other primary headaches.

Treatment of frequent primary headache with or without migrainous features, whatever view of the biology or nomenclature one takes, requires control of analgesic, ergotamine, or triptan overuse when present and instigation of preventative medicines. It is exceptional for a patient misusing medication to be successfully treated with preventatives unless these other excessive medications are stopped. Preventatives used in episodic migraine are all employed, including tricylics, Valproate (divalproex), Gabapentin, Flunarizine, Methysergide and monoamine oxidase inhibitors. Comorbidity with depression is common in migraine, so that appropriate management of accompanying depression is important. Admission to hospital and treatment with a carefully monitored course of intravenous dihydroergotamine can be a very effective way to break the cycle of persistent headache.

Syndromes responsive to indomethacin

There are number of primary headache syndromes with distinct characteristics that respond to treatment with indomethacin. Many share features with cluster headache, and are collectively known as the trigeminal autonomic cephalgias (TACs) (Table 8), while others do not. They deserve some attention because they can often be very easily treated.

Paroxysmal hemicrania

Sjaastad and colleagues first reported eight cases of a frequent unilateral severe but short-lasting headache without remission, coining the term 'chronic paroxysmal hemicrania'. The mean daily frequency of attacks varied from seven to 22 with the pain persisting from 5 to 45 min on each occasion. The site and associated autonomic phenomena were similar to those of cluster headache, but the attacks of chronic paroxysmal hemicrania were suppressed completely by indomethacin. A subsequent review of 84 cases showed a history of remission in 35 cases whereas 49 were chronic. By analogy with cluster headache the patients with remission have been referred to as having episodic paroxysmal hemicrania, and those without can be labelled with chronic paroxysmal hemcrania. Pareja has recorded attacks which swap sides, just as is known for cluster headache, and attacks with autonomic features without pain. This has been observed in cluster headache after trigeminal nerve section, by this author and others, and is excellent evidence for a disorder which is primarily of the central nervous system.

The essential features of paroxysmal hemicrania are:

- Female preponderance.
- Unilateral, usually frontotemporal, with very severe pain.
- Short-lasting attacks (2 to 45 min).
- Very frequent attacks (usually more than five a day).
- Marked autonomic features ipsilateral to the pain.
- Robust, quick (less than 72 h), and complete response to indomethacin.

Other issues

The treatment of paroxysmal hemicrania is complicated by the gastrointestinal side-effects seen with indomethacin, but thus far there is no convincing replacement. The issue of triptan response is not clearly settled and may be both variable and dependent on the length of the attacks. Injection of the greater occipital nerve is not useful in paroxysmal hemicrania. Piroxicam has been suggested to be helpful, although again it is not as effective as indomethacin. By analogy with cluster headache, verapamil has been used in paroxysmal hemicrania, although the response is not spectacular and higher doses require exploration. Paroxysmal hemicrania can coexist with trigeminal neuralgia—paroxysmal hemicrania-tic syndome—just as in cluster-tic syndrome. Similarly, secondary chronic paroxysmal hemicrania has also been reported with a syndrome like Tolosa–Hunt and in patients with a pituitary microadenoma and a maxillary cyst. If there is any doubt regarding the differential diagnosis between paroxysmal hemicrania and cluster headache then either an oral indomethacin challenge, or a formal placebo-controlled indomethacin test by injection, should be completed.

Hemicrania continua

Sjaastad and Spierings reported two patients, a woman aged 63 and a man of 53, who developed unilateral headache without obvious cause. One of these patients noticed redness, lacrimation, and sensitivity to light in the eye on the affected side. Both patients were relieved completely by indomethacin while other non-steroidal anti-inflammatory drugs were of little or no benefit. Newman and colleagues reviewed the 24 previously reported cases and added 10 of their own, including some with pronounced autonomic features resembling cluster headache. They divided their case histories into remitting and unremitting forms. Of the 34 patients reviewed, 22 were women and 12 men with the age of onset ranging from 11 to 58 years. The symptoms were controlled by indomethacin 75 to 150 mg daily. The essential

features of hemicrania continua are:

- Unilateral pain.
- Pain is moderate and continuous but with fluctuations.
- Complete resolution of pain with indomethacin.
- Exacerbations may be associated with autonomic features.
- Migrainous features, such as nausea, photophobia or phonophobia, are frequently reported.

Apart from overuse of analgesics as a secondary aggravation, and a report in an HIV-infected patient, the status of secondary hemicrania continua is unclear. Injection of the greater occipital nerve is not useful in hemicrania continua.

Time to treatment response

Antonaci and colleagues proposed the 'indotest' by which the intramuscular injection of 50 mg of indomethacin could be used as a diagnostic tool. In hemicrania continua, pain was relieved in 73 ± 66 min and the pain-free period was 13 ± 8 h. The time elapsing between the thrice daily oral administration of 25 to 50 mg indomethacin and relief varied from 30 min to 48 h and thus a response to treatment can be rapidly assessed in the outpatient setting. Acute treatment with sumatriptan has been employed and reported to be of no benefit.

Idiopathic (Primary) Stabbing Headache

P>Short-lived jabs of pain, defined by the International Headache Society as Idiopathic (Primary) Jabbing Headache, are well documented in association with most types of primary headache. The essential clinical features are:

- Pain confined to the head, although rarely is it facial.
- Stabbing pain lasting from one to many seconds and occurring as a single stab or a series of stabs.
- Recurring at irregular intervals (hours to days).
- Cranial autonomic symptoms, such as lacrimation, conjunctival injection and rhinorrhoea, are not reported.

Raskin and Schwartz first described these sharp, jabbing pains about the head resembling a stab from an ice-pick, nail, or needle. They compared the prevalence of such pains in 100 migrainous patients and 100 headache-free controls and only three of the control subjects had experienced ice-pick pains compared with 42 of the migraine patients, of whom 60 per cent had more than one attack per month. The pains affected the temple or orbit more often than the parietal and occipital areas and often occurred before or during migraine headaches. The sites of the ice-pick pains generally coincide with the site of the patient's habitual headache.

Pains in the retroauricular and occipital region are also well described and these respond promptly to indomethacin. Ice-pick pains have been described in conjunction with cluster headaches, and are generally experienced in the same area as the cluster pain. Sjaastad described what he called 'jabs and jolts' lasting less than a minute in patients with chronic paroxysmal hemicrania. These longer attacks are almost certainly part of the spectrum of jabbing headache. It is of interest that jabbing pains generally are not accompanied by cranial autonomic symptoms. The response of idiopathic jabbing headache to indomethacin (25 to 50 mg twice to three times daily) is generally excellent. As a general rule the symptoms wax and wane, and after a period of control on indomethacin it is appropriate to withdraw treatment and observe the outcome.

Benign cough headache

Sharp pain in the head on coughing, sneezing, straining, laughing, or stooping has long been regarded as a symptom of organic intracranial disease, commonly associated with obstruction of the cerebrospinal fluid pathways. The presence of an Arnold–Chiari malformation or any lesion causing obstruction of cerebrospinal fluid pathways or displacing cerebral structures must be excluded before cough headache is assumed to be benign. Cerebral aneurysm, carotid stenosis, and vertebrobasilar disease may also present with cough or exertional headache as the initial symptom. The term 'benign Valsalva's manoeuvre-related headache' covers the headaches provoked by coughing, straining, or stooping but 'cough headache' is more succinct and so widely used it is unlikely to be displaced. The essential clinical features of benign cough headache are:

- Bilateral headache of sudden onset, lasting less than a minute, precipitated by coughing.
- May be prevented by avoiding coughing.
- Diagnosed only after structural lesions, such as a posterior fossa tumour, have been excluded by neuroimaging.

Comparing benign cough with benign exertional headache Pascual and colleagues reported that the average age of their patients with benign cough headache was 43 years more than their patients with exertional headache.

Management

Indomethacin is the medical treatment of choice in cough headache. Raskin has reported that some patients with cough headache are relieved by lumbar puncture which is a simple option when compared with prolonged use of indomethacin. The mechanism of this response remains unclear.

Benign exertional headache

The relationship of this form of headache to cough headache is unclear and certainly much is shared. Indeed the relationship to migraine also requires delineation. Credit must be given to Hippocrates for first recognizing this syndrome when he wrote 'one should be able to recognize those who have headache from gymnastic exercises, or walking, or running, or any other unseasonable labour, or from immoderate venery'.

The clinical features are:

- Pain specifically brought on by physical exercise.
- Bilateral and throbbing in nature at onset and may develop migrainous features in those patients susceptible to migraine.
- Lasts from 5 min to 24 h.
- Prevented by avoiding excessive exertion, particularly in hot weather or at high altitude.

The acute onset of headache with straining and breath holding as in weightlifter's headache may be explained by acute venous distension. The development of headache after sustained exertion, particularly on a hot day, is more difficult to understand. Anginal pain may be referred to the head, probably by central connections of vagal afferents, and may present as exertional headache, so-called cardiac cephalalgia. The link to exercise is the main clinical clue. Phaeochromocytoma may occasionally be responsible for exertional headache. Intracranial lesions or stenosis of the carotid arteries may have to be excluded as discussed for benign cough headache. Headache may be precipitated by any form of exercise and often has the pulsatile quality of migraine.

Management

The most obvious form of treatment is to take exercise gradually and progressively whenever possible. Indomethacin at daily doses varying from 25 to 150 mg is generally very effective in benign exertional headache. Indomethacin 50 mg, ergotamine tartrate 1 to 2 mg orally, ergotamine by inhalation, or methysergide 1 to 2 mg orally given 30 to 45 min before exercise are useful prophylactic measures.

H4>Other interesting primary headaches

Hypnic headache

This syndrome was first described by Raskin in patients aged between 67 and 84 who had headache of a moderately severe nature that typically came on a few hours after going to sleep. These headaches last from 15 to 30 min, are typically generalized, although they may be unilateral, and can be throbbing. Patients may report falling back to sleep only to be awoken by a further attack a few hours later with up to three repetitions of this pattern over the night. In the largest series (Dodick's) of 19 patients, 16 (84 per cent) were female and the mean age at onset was 61 ± 9 years. Headaches were bilateral in two-thirds of cases and unilateral in one-third, and in 80 per cent of cases pain was mild or moderate. Three patients reported similar headaches when falling asleep during the day. None of these patients had photophobia or phonophobia and nausea was unusual.

Management

Patients with this form of headache generally respond to a bedtime dose of lithium carbonate (200 to 600 mg) and in those that do not tolerate this verapamil or methysergide at bedtime may be alternative strategies. Two patients who responded to flunarizine 5 mg at night have now been reported. Dodick and colleagues reported that one to two cups of coffee or caffeine 60 mg orally at bedtime were helpful, and this is well worth trying as the first step in Hypnic Headache. This author has controlled a patient poorly tolerant of lithium by using verapamil at night (160 mg).

Short-lasting unilateral neuralgiform headache attacks with conjunctival injection and tearing (the SUNCT syndrome)

It has been remarked (Lance) that as the duration of the pain in the cluster headache-like syndromes becomes shorter, their names become longer! Sjaastad and colleagues reported three male patients whose brief attacks of pain in and around one eye were associated with sudden conjunctival injection and other autonomic features of cluster headache. Attacks lasted only 15 to 60 s and recurred five to 30 times an hour. Attacks could be precipitated by chewing or eating certain foods, such as citrus fruits, and were not abolished by indomethacin. Most cases have some associated precipitating factors, particularly movements of the neck. Of the patients recognized with this problem males predominate and the paroxysms of pain usually last between 5 and 250 s, although longer duller interictal pains are recognized as well as attacks of up to 2 h in two patients. The conjunctival injection seen with SUNCT is often the most prominent autonomic feature and production of tears may be very obvious. SUNCT is more or less intractable to medical management although there is a modest benefit with carbamazepine that is nothing like its effect in trigeminal neuralgia. Recently, we have found that topiramate and lamotrigine are useful in SUNCT syndrome.

Secondary SUNCT and associations

There have been three reported patients with secondary SUNCT syndromes. The first two patients had homolateral cerebellopontine angle arteriovenous malformations diagnosed on MRI. The third patient had a cavernous haemangioma of the brainstem seen only on MRI. These cases highlight the need for cranial MRI in investigating for secondary SUNCT. Just as there is a reported case of chronic paroxysmal hemicrania associated with trigeminal neuralgia there is a single report of a patient with trigeminal neuralgia who developed a SUNCT syndrome.

Benign sex headache

Sex headache may be precipitated by masturbation or coitus and usually starts as a dull bilateral ache while sexual excitement increases, suddenly becoming intense at orgasm. The term orgasmic cephalalgia is not useful since not all sex headache requires orgasm. Three types of sex headache are discussed: a dull ache in the head and neck that intensifies as sexual excitement increases, a sudden severe ('explosive') headache occurring at orgasm, and a postural headache resembling that due to low cerebrospinal fluid pressure developing after coitus. The latter in the author's clinical experience is simply another form of headache due to low cerebrospinal fluid pressure arising from vigorous sexual activity usually with multiple orgasm and might be usefully considered with the secondary chronic daily headaches (Table 10). The essential clinical features of sex headache are:

- Precipitation by sexual excitement.
- Bilateral at onset.
- Prevented or eased by ceasing sexual activity before orgasm.

Headaches developing at the time of orgasm are not always benign. Subarachnoid haemorrhage was precipitated by sexual intercourse in 4.5 per cent of 66 cases reported by Fisher and 12 per cent of 50 cases studied by Lundberg and Osterman. One young man was reported to have developed a brainstem thrombosis and another a left hemisphere infarction.

Sex headache affects men more often than women and may occur at any time during the years of sexual activity. It may develop on several occasions in succession and then not trouble the patient again, although there is no obvious change in sexual technique. In patients who stop sexual activity when headache is first noticed it may subside within a period of 5 min to 2 h, and it is recognized that more frequent orgasm can aggravate established sex headache. About half of patients with sex headache have a history of exertional headaches but there is no excess of cough headache in patients with sex headache. In about 50 per cent of patients sex headache will settle in 6 months. Migraine is probably more common in patients with sex headache.

Management

Benign sex headaches are usually irregular and infrequent in recurrence, so management can often be limited to reassurance and advice about ceasing sexual activity if a milder, warning headache develops. When the condition recurs regularly or frequently, it can be prevented by the administration of propranolol, but the dosage required varies from 40 to 200 mg daily. An alternative is the calcium channel blocking agent diltiazem 60 mg three times daily. Ergotamine (1 to 2 mg) or indomethacin (25 to 50 mg) taken about 30 to 45 min prior to sexual activity can also be helpful.

Thunderclap headache

Severe headache of sudden onset may occur in the absence of sexual activity and it is appropriate to consider the issue here as it may be the sentinel bleed of an intracranial aneurysm and there are some issues that overlap clinically. Day and Raskin reported a woman with three episodes of very severe headache of sudden onset who was found to have an unruptured aneurysm of the internal carotid artery, with adjacent areas of segmental vasospasm. While headaches of explosive onset may certainly be caused by the ingestion of sympathomimetic drugs or tyramine-containing foods in a patient who is taking monoamine oxidase inhibitors, and can also be a symptom of phaeochromocytoma, the relationship between thunderclap headache and aneurysm in the absence of CT scan or cerebrospinal fluid evidence of subarachnoid haemorrhage is difficult. Wijdicks and colleagues followed up 71 patients whose CT scans and cerebrospinal fluid findings were negative for an average of 3.3 years. Twelve patients had further such headache, and 31 (44 per cent) later had regular episodes of migraine or tension-type headache. Factors identified as precipitating the headache were sexual intercourse in three cases, coughing in four, and exertion in 12, while the remainder had no obvious cause. A history of hypertension was found in 11 and of previous headache in 22. Markus compared the presentation of 37 patients with subarachnoid haemorrhage and 189 with a similar thunderclap headache but normal cerebrospinal fluid examination and could not discern any characteristic which distinguished the two conditions on clinical grounds.

Investigation of any severe headache of sudden onset, be it in the context of sexual excitement or isolated thunderclap headache, should be driven by the clinical context. The first presentation should be vigorously investigated with CT and cerebrospinal fluid examination and where possible MRI angiography. Formal cerebral angiography should be reserved to situations of high clinical suspicion. It is worth noting that of diffuse multifocal reversible spasm may be seen in thunderclap headache without there being an intracranial aneurysm.

Further reading

Antonaci F *et al.* (1998). Chronic paroxysmal hemicrania and hemicrania continua. Parenteral indomethacin: the 'Indotest'. *Headache* **38**, 122–8.

Day JW, Raskin NH (1986). Thunderclap headache: symptom of unruptured cerebral aneurysm. *Lancet* **2**, 1247–8.

Dodick DW, Mosek AC, Campbell JK (1998). The hypnic ('alarm clock') headache syndrome. *Cephalalgia* **18**, 152–6.

Ferrari MD *et al*. (2001). Triptans (serotonin, 5-HT1B/1D agonists) in acute migraine treatment – a meta-analysis of 53 trials. *Lancet* **358**, 1668–75.

Fisher CM (1968). Headache in cerebrovascular disease. In: Vinken PJ, Bruyn GW, eds. *Handbook of clinical neurology*, Vol 5, pp 124–6. Elsevier, Amsterdam.

Goadsby PJ, Olesen J (1998). Diagnosis and management of migraine. *British Medical Journal* **312**, 1279–82.

Goadsby PJ, Silberstein SD (1997). Headache. In: Asbury A, Marsden CD, eds. *Blue books in practical neurology*, Vol 17. Butterworth-Heinemann, New York.

Goadsby PJ (2000). The pharmacology of headache. *Progress in Neurobiology* 62, 509–25.

Goadsby PJ (2002). Chronic tension-type headache. In: Barton S, ed. *Clinical evidence*, Vol. 6. BMJ Publishing Group, London, in press.

Goadsby PJ, Lipton RB, Ferrari MD (2002). Migraine: current understanding and management. *New England Journal of Medicine* **346**, in press.

Griggs RC, Nutt JG (1995). Episodic ataxias as channelopathies. *Annals of Neurology* **37**, 285–7.

Kudrow L (1987). Cluster headache. In: Blau JN, ed. *Headache: clinical, therapeutic, conceptual and research aspects*. Chapman and Hall, London.

Kudrow L, Esperanca P, Vijayan N (1987). Episodic paroxysmal hemicrania? *Cephalalgia* **7**, 197–201.

Lance JW (1976). Headaches related to sexual activity. *Journal of Neurology, Neurosurgery and Psychiatry* **39**, 1226–30.

Lance JW, Hinterberger H (1976). Symptoms of pheochromocytoma, with particular reference to headache, correlated with catecholamine production. *Archives of Neurology* **33**, 281–8.

Lundberg PO, Osterman PO (1974). The benign and malignant forms of orgasmic cephalgia. *Headache* **14**, 164–5.

Markus HS (1991). A prospective follow-up of thunderclap headache mimicking subarachnoid haemorrhage. *Journal of Neurology, Neurosurgery and Psychiatry* **54**, 1117–25.

May A *et al.* (1998). Hypothalamic activation in cluster headache attacks. *The Lancet* **351**, 275–8.

May A *et al.* (1999). Correlation between structural and functional changes in brain in an idiopathic headache syndrome. *Nature Medicine* **5**, 836–8.

Newman LC, Lipton RB, Solomon S (1994). Hemicrania continua: ten new cases and a review of the literature. *Neurology* **44**, 2111–14.

Olesen J, Goadsby PJ (1999). In: Olesen J, ed. *Cluster headache and related conditions*, Vol 9. Oxford University Press, Oxford.

Pareja JA (1995). Chronic paroxysmal hemicrania: dissociation of the pain and autonomic features. *Headache* **35**, 111–13.

Pascual P *et al.* (1996). Cough, exertional, and sexual headache. *Neurology* **46**, 1520–4.

Quality Standards Subcommittee of the American Academy of Neurology (1994). The utility of neuroimaging in the evaluation of headache patients with normal neurologic examinations. *Neurology* **44**, 1353–4.

Raskin NH (1988). The hypnic headache syndrome. *Headache* **28**, 534–6.

Raskin NH (1995). The cough headache syndrome: treatment. *Neurology* **45**, 1784.

Raskin NH, Schwartz RK (1980). Icepick-like pain. *Neurology* **30**, 203–5.

Rassmussen BK (1995). Epidemology of headache. *Cephalalgia* **15**, 45–68.

Silberstein SD, Lipton RB, Sliwinski M (1996). Classification of daily and near-daily headaches: a field study of revised IHS criteria. *Neurology* **47**, 871–5.

Sjaastad O *et al.* (1980). Chronic paroxysmal hemicrania (CPH). The clinical manifestations. a review. *Uppsala Journal of Medical Science* **31**, 27–33.

Sjaastad O *et al.* (1989). Shortlasting unilateral neuralgiform headache attacks with conjunctival injection, tearing, sweating, and rhinorrhea. *Cephalalgia* **9**, 147–56.

Sjaastad O, Spierings EL (1984). Hemicrania continua: another headache absolutely responsive to indmethacin. *Cephalalgia* **4**, 65–70.

Stewart WF *et al.* (1999). Reliability of the migraine disability assessment score in a population-based sample of headache sufferers. *Cephalalgia* **19**, 107–14.

Tfelt-Hansen P *et al.* (2000). Ergotamine in the acute treatment of migraine—a review and European consensus. *Brain* **123**, 9–18.

Tzourio C *et al.* (1995). Case-control study of migraine and risk of ischaemic stroke in young women. *British Medical Journal* **310**, 830–3.

Weiller C *et al.*(1995). Brain stem activation in spontaneous human migraine attacks. *Nature Medicine* **1**, 658–60.

Wijdicks EFM, Kerkhoff H, van Gijn J (1988). Long-term follow up of 71 patients with thunderclap headache mimicking subarachnoid haemorrhage. *Lancet* **2**, 68–70.

24.13.3 Epilepsy in later childhood and adults

G. D. Perkin

Definitions

Using guidelines developed by the International League Against Epilepsy (**ILEA**), epilepsy is defined as recurrent (two or more) epileptic seizures, unprovoked by any immediate identifiable cause. Excluded are febrile seizures and neonatal seizures (the latter defined as those occurring in the first 4 weeks of life). Multiple seizures occurring within a 24-h period are considered to represent a single event. The epileptic seizure itself is defined as the clinical manifestation of an abnormal and excessive discharge of a set of brain neurones. The manifestation is a sudden transient phenomenon that may include alteration of consciousness, motor, sensory, autonomic, or psychic events which are perceived either by the individual or by an observer. Problems arise, when using the term epilepsy, with those individuals who may have had only two or three attacks in a lifetime. To take account of this, the terms active and inactive epilepsy are used, the former referring to patients with at least one seizure in the previous 5 years, the latter to patients who have been seizure free over the same period. The definitions are further qualified, for inactive cases, according to whether the individual is on drug therapy.

The idiopathic epilepsies are defined as those epileptic disorders (partial or generalized) which have characteristic clinical and electroencephalographic features coupled with a genetic predisposition. Cryptogenic epilepsy defines cases of partial or generalized epilepsy in which no aetiological factor has been identified. Symptomatic seizures are those occurring in association with a known risk factor. Epileptic syndromes have been defined by the ILEA on the basis of clinical characteristics, age of onset, and electroencephalographic findings.

Epidemiology

Incidence

Most reported incidence rates lie between 40 and 70 per 100 000. Figures for developing countries usually exceed 100 per 100 000. Age-specific rates show a bimodal distribution, with the highest peak in the first decade, falling thereafter until a second peak in later life.

Prevalence

Prevalence figures are more widely available. For adults, rates lie between 1.5 and 57 per 1000, with an average of 10.3 per 1000. Cumulative incidence (or lifetime prevalence) rates, excluding febrile seizures, are higher, producing a figure between 1.5 and 5 per cent (Fig. 1).

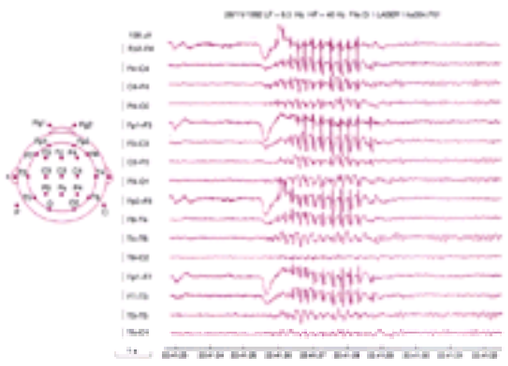

Fig. 1 Electroencephalogram of a typical absence seizure. The first 2.5 s of the record are entirely normal. The event begins with a large downward deflection which records eye closure, immediately followed in all channels by a spike-and-wave discharge at a frequency of 3 cycles/s. The seizure terminates as abruptly as it began. (Record kindly provided by Dr David Fish.)

Sex

Males have slightly higher prevalence rates than females.

Socio-economic status

Higher prevalence rates have been reported in the lower socio-economic groups, both in developed and developing countries.

Pathophysiology

Inherent in any discussion of epilepsy mechanisms is the need to define a homogeneous population. Generalized tonic–clonic seizures, for example, can occur with many different epileptic syndromes. Epileptic seizures are thought to arise at cortical sites. Partial seizures begin focally in the cortex, generalized seizures infer widespread, bilateral cortical involvement from the beginning. An interictal discharge occurs when a group of pyramidal neurones is synchronously activated. During the discharge, the cells develop a large and prolonged depolarization which is terminated by a hyperpolarizing potential. Seizures develop when any of the inhibitory processes suppressing interictal discharges fail.

The underlying mechanisms behind epileptic discharges have been best defined for absence seizures where a thalamocortical circuit is responsible for generating synchronous burst-firing of neurones. The circuit involves neocortical pyramidal neurones, thalamic relay neurones, and neurones of the nucleus reticularis thalami. The last are exclusively g-aminobutyric acid (**GABA**) in type. A voltage-dependent calcium channel (T-channel) appears critical in allowing burst-firing of neurones to appear. Following activation, the T-channels acquire repolarization via $GABA_B$ receptors present on thalamic relay neurones. $GABA_A$ receptors also play an important regulatory role in synchronized thalamocortical burst-firing.

Less information is available on the pathophysiological mechanisms of generalized convulsive seizures. Roles for $GABA_A$ receptors and for altered serotonergic neurotransmission have been suggested.

Classification

The ILEA classification scheme, as revised in 1989, is now widely used for epidemiological, management, and research purposes. The scheme divides seizures into focal, generalized, and unclassifiable forms (Table 1).

Though it is widely used, the classification has disadvantages. The ability to determine whether consciousness is preserved, in order to make the distinction between simple and complex partial seizures, is often limited. Some individuals, though appearing alert, can be shown to have impaired awareness when carefully tested.

An elaboration of the classification consists of a list of epileptic syndromes into which, theoretically, all generalized and partial epileptic seizures can be fitted. The idiopathic generalized seizures are classified according to age of onset and seizure type. The partial seizures, attributed to dysfunction of restricted cortical areas, are predominantly classified according to their clinical features, supplemented by electroencephalographic findings. Much criticism has been made of this syndromic classification. In routine clinical practice many cases (probably the majority) are left in non-specific categories. Moreover, the classification fails to incorporate data derived from CT or MRI.

Partial seizures

Simple partial motor seizures

Any part of the body can be affected by a focal motor seizure, according to the site of origin of the discharge. Sometimes the seizure remains localized to the same area (for example the hand), sometimes it 'marches' along the motor cortex, producing successional jerking of contiguous body parts (jacksonian seizures). During the focal stage, consciousness is preserved. With secondary generalization (that is, diffuse bilateral spread) consciousness is lost. The parts of the body most commonly affected by this type of seizure correlate with their area of representation in the motor cortex. Other focal motor disturbances reflecting epileptic discharges include rotation of head and eyes contralaterally (from the dorsolateral prefrontal cortex), tonic foot movements ipsilaterally (the paracentral lobule), and head turning with arm extension on the same side (supplementary motor cortex). Following such seizures there may be paralysis of the affected part lasting for minutes or hours (Todd's paresis).

Simple partial sensory seizures

Seizures emanating from the sensory cortex produce paraesthesias or numbness. The seizure can march in an analogous fashion to a motor seizure, and similarly, can then become generalized. Where the tongue or face are involved, the symptoms are sometimes felt bilaterally. More complex sensory phenomena may be experienced and with discharges in the second sensory area, the limb sensations can be ipsilateral, contralateral, or bilateral.

Occipital lobe seizures

Visual symptoms predominate, usually as simple rather than complex phenomena. The latter, producing alteration of size, shape, or depth of objects are associated with seizures arising at the occipito-parieto-temporal interface. In addition there may be ocular deviation, jerking, or forced closure of the eyelids. Visual hallucinations may occur.

Frontal lobe seizures

Frontal lobe seizures are commonly nocturnal and frequently associated with turning to a prone position. Vocalization is common and tends to consist of a continuous monotone with moaning or grunting. An aura before the attack is unusual. Other recognized features include pelvic thrusting, rocking of the body, and head movements. Rapid postictal recovery is common.

Simple partial (temporal lobe) seizures

The distinction between simple and complex partial seizures is difficult, based as it is on evidence of altered consciousness with the latter. Olfactory, gustatory, and vertiginous sensations occur. The taste and smell sensations are sometimes pleasurable but often disagreeable. A metallic taste is common. Abdominal sensations also occur, which are typically ill-defined and may ascend to the chest and throat. Psychic symptoms are more often associated with complex partial seizures. There may be intense pleasure or fear ushering in the attack. The patient can experience a sense of loss of personal or environmental reality (depersonalization and derealization, respectively). There may be a sense of intense familiarity (*déjà vu*) or unfamiliarity (*jamais vu*). Epileptic anger is unprovoked and rapidly subsides. Illusions are encountered, in the form of disordered visual perceptions, and visual or auditory hallucinations, sometimes of considerable complexity (Fig. 2).

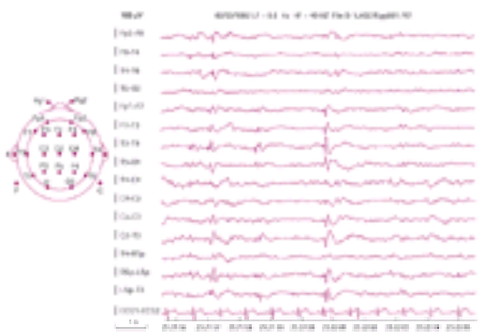

Fig. 2 Interictal spike and slow-wave complex in a patient with complex partial seizures. The discharges are particularly apparent over the left temporal lobe (T3 to T5), but there are some independent discharges over the right temporal lobe (T4 to T6). (Record kindly provided by Dr David Fish.)

Where consciousness is disturbed, various automatic activity or movement may occur of which the patient is unaware (automatisms). These may take the form of eating (chewing or swallowing), speaking, gesture, or more elaborate skilled activities. Some of these automatic movements are also seen with absence seizures. When elaborate, the patient may partly undress, or move about from one room to another. The symptomatology of mesial and lateral temporal lobe discharges has been distinguished, the latter having additional somatosensory, visual, or auditory manifestations to the other features mentioned above.

Other, rarer, focal seizure types are confined to childhood. In benign childhood epilepsy with centrotemporal spikes, consciousness is preserved. The sensory phenomena are usually confined to the mouth where motor activity may also occur. Speech arrest occurs if the dominant hemisphere is affected.

Any of the focal epilepsies can lead to secondary generalization. Consciousness is lost, and a tonic–clonic seizure is the usual outcome. Prolonged focal seizures (epilepsia partialis continua) lead to a repetitive or continuous focal motor activity which may last for weeks or months and is most often the consequence of a focal cortical insult.

Generalized seizures

Tonic–clonic seizures (grand mal epilepsy)

Some patients report a premonition for hours or even days before the attack. The symptoms are usually a vague sense of loss of well being and do not imply a focal origin for the attack. An aura, on the other hand, lasting a few seconds before the onset implies a focal origin for the attack, demanding classification as a focal seizure with secondary generalization. The tonic phase is associated with contraction of axial then limb muscles. If upright, the patient falls heavily. Injury is common. Contraction of the jaw can lead to tongue injury. Forcible contraction of the diaphragm results in a sudden gasp or epileptic cry. Cyanosis results from a loss of respiratory activity. Subsequently clonic movements appear and slowly increase in amplitude. Gradually, periods of relaxation intervene between the clonic contractions until finally all movements cease. The patient is then flaccid. Urinary or faecal incontinence or both may occur at this stage. Subsequently the patient is liable to sleep, often heavily. If the patient wakes, initial confusion and disorientation is the norm. Headache and muscle pains are common. Incomplete forms occur in which the clonic or tonic phase predominates.

In addition to injuries incurred in falling, and those resulting from biting of the cheeks or tongue (typically the lateral margin is affected), the seizures may be of such violence that vertebral compression fractures occur. Sudden death occurring soon after a tonic–clonic seizure is a recognized, though rare, complication. Its incidence lies between 1:500 and 1:1000 deaths per person-year.

Absence seizures (petit mal)

Patients are totally unaware of their absence seizures. Activity suddenly ceases but without loss of posture. Adventitious movements occur, for example slight contractions of the eyes or some lip movement. The head may drop slightly. More typically, the patient simply stares blankly and is unresponsive. Attacks last around 10 to 20 s. In some cases more overt limb movement occurs.

Atypical absences are defined as attacks which begin less abruptly, last longer, and frequently lead to loss of postural tone. They usually coincide with other seizure types. Absence seizures begin in childhood and usually cease in adult life, though some 50 per cent of patients will later develop tonic–clonic seizures.

Myoclonic seizures

Myoclonus consists of brief, shock-like contractions of muscle, occurring either in a generalized or focal distribution. Many forms of myoclonus are non-epileptic. Those associated with epilepsy are accompanied by an ictal electroencephalographic discharge. In primary generalized epileptic myoclonus, the myoclonus is accompanied by diffuse cortical epileptic discharges.

Atonic seizures

Atonic seizures result in sudden loss of muscle tone. If the hypotonus is generalized, falls occur, often with substantial injury. The attacks begin in infancy or childhood. The episodes are brief and recovery rapid unless injury has occurred.

Status epilepticus

Status epilepticus is defined as a single seizure lasting more than 30 min or successional seizures without recovery of consciousness between. The seizures are usually tonic–clonic. Both complex partial seizures and absence seizures can occur in the form of status epilepticus. In such cases, alteration of the conscious level is likely to be the major clinical feature with little motor activity, particularly with the latter.

Epilepsy syndromes

The need to define epileptic syndromes arises from the fact that individual seizure types may be a manifestation of a number of differing conditions, all with individual characteristics and prognosis. The epileptic syndrome is based on a combination of seizure type, presumed localization (according to clinical features and electroencephalographic characteristics) in the case of the partial seizures, and age of onset. In routine, as opposed to heavily specialized, practice only a third of newly diagnosed epilepsy can be fitted into such a classification system.

Causes of epilepsy

In most surveys, only about a quarter to a third of epilepsy cases have been attributable to a specific cause. With modern imaging methods, this proportion is likely to rise significantly.

Genetically determined

In some genetically determined disorders, epilepsy is only one feature of the condition. Many such disorders have features other than epilepsy and typically produce significant neurological disability. Examples include the forms of progressive myoclonic epilepsy associated with Lafora body disease and Unverricht–Lundborg disease. More relevant, in clinical terms, are those genetically determined conditions in which epilepsy is the sole or major manifestation. Among these are some with simple forms of inheritance, and some with complex forms. The epilepsies occurring with such syndromes may be generalized or partial in nature.

Epilepsy syndromes with simple inheritance

Examples in this category include benign familial neonatal convulsions and benign familial infantile convulsions. Linkage to chromosome 20q has been described for the former, coding for potassium channel proteins.

Epilepsy syndromes with complex inheritance

Examples include juvenile myoclonic epilepsy and benign rolandic epilepsy.

Juvenile myoclonic epilepsy

This epilepsy type usually begins between the ages of 12 and 18. Early morning, sudden myoclonic jerks of the arms and shoulders are characteristic. Subsequently, generalized tonic–clonic seizures occur in the majority of patients. Genes for this disorder, inherited in a dominant manner, have been localized to both chromosome 6 and 15.

Benign rolandic epilepsy

This condition presents between the ages of 5 and 10 years. Unilateral motor or sensory seizures occur in sleep. The condition is associated with centrotemporal spikes on electroencephalography. The epileptiform abnormality may be linked to chromosome 15.

Included within the umbrella of the genetically determined epilepsy syndromes (though often classified separately from the epilepsies) are febrile seizures. These occur in between 2 and 5 per cent of children, typically between the ages of 6 months and 3 years. Simple febrile seizures are generalized and last less than 15 min. Complex febrile seizures either have focal features, are longer lasting, or recur within a 24-h period. About two-thirds of children with a febrile seizure do not have a recurrence. Febrile seizures do not have an adverse effect on development but a proportion of children develop epilepsy at a later age. Inheritance of febrile seizures is considered to be as an autosomal dominant trait.

Migration disorders

Alterations of the migration processes that establish the cellular and laminar organization of the neocortex are now considered to underlie a number of cases of epilepsy, particularly those arising in childhood. The migration disorders may be generalized (such as agyria), hemispheric (hemimegalencephaly), or focal (such as cortical dysplasia). Several forms of the migration disorders are genetically determined, and linked to mutations on the X chromosome. Some of these migration disorders are detectable by high-resolution MRI. The epilepsies produced can be either focal or generalized, and typically are difficult to control.

Trauma

Approximately 70 per cent of those individuals who eventually develop post-traumatic epilepsy will have their first seizure within 2 years of the original injury. Risk factors that predict post-traumatic epilepsy include early seizures (those occurring in the first week), a depressed skull fracture, or evidence of intracranial haemorrhage. There is no justification for the use of prophylactic anticonvulsants in the hope of preventing the development of post-traumatic seizures.

Tumour

Though adult-onset epilepsy is often equated with the presence of tumour, the cause in later life of symptomatic epilepsy is more likely to be cerebrovascular or Alzheimer's disease. The likelihood of a tumour producing seizures increases as the tumour is sited more anteriorly in the hemisphere, so that over 50 per cent of patients with frontal lobe tumours have epilepsy. Adult-onset status, in someone without a history of epilepsy, is particularly suggestive of frontal lobe tumour. Epilepsy is more common with slow-growing tumours and may be generalized or focal in nature.

Cerebrovascular disease

The prevalence of epilepsy after stroke has been reported to lie between 6 and 15 per cent, and appears as likely with cerebral infarction as with cerebral haemorrhage.

Infection

In large-scale surveys, infection has been considered the cause of epilepsy in 3 to 5 per cent of cases. Differences in rate between countries are often attributed to the variable prevalence of certain agents, for example cysticercosis. Other tropical infections which have been considered potential contributors to epilepsy prevalence include malaria, schistosomiasis, and trypanosomiasis. Epilepsy is a recognized feature of bacterial, tuberculous, and fungal meningitis, and of viral encephalitis. Epilepsy is often the first symptom of a tuberculoma.

Cerebral degeneration

Epilepsy is more common in patients with Alzheimer's disease or multi-infarct dementia compared with age-matched controls.

Multiple sclerosis

The prevalence of epilepsy in multiple sclerosis is probably of the order of 2 per cent. Both generalized and focal seizures have been attributed to multiple sclerosis. Rarely, status epilepticus and epilepsia partialis continua have been recorded.

Alcohol

Alcohol lowers seizure threshold. Seizures may occur during binge drinking, or during a period of withdrawal after alcohol excess.

Metabolic disorders

Seizures may occur in association with hypocalcaemia, hypercalcaemia, hypomagnesaemia, hypoglycaemia, hyponatraemia, and hypernatraemia. Severe renal and hepatic failure can both precipitate seizures.

Certain drugs are considered to lower the seizure threshold and are relatively contraindicated in patients with epilepsy. The drugs in question include the tricyclic

antidepressants, the phenothiazines, and isoniazid. Rapid withdrawal of barbiturates or benzodiazepines can trigger seizures in those without a history of epilepsy.

Precipitants of epilepsy

Recognized precipitants of epilepsy include inadequate sleep, alcohol abuse, and ingestion of certain drugs. In catamenial epilepsy the attacks are confined to the menstrual period. Seizures confined to sleep are well recognized and indeed sleep electroencephalography recordings are characteristically more likely to register abnormal discharges than recordings made in the alert individual. In reflex epilepsy attacks are virtually inevitably triggered by a particular stimulus. Precipitants include photic stimulation, startle, noise, and movement. Rarer forms of reflex epilepsy have been linked to musical passages, eating, and performance of certain mental tasks.

Differential diagnosis

Syncope

Most individuals who faint experience a characteristic set of symptoms prior to loss of consciousness. These include mental slowing, fading of vision, altered hearing, malaise, and sweating. The process is the result, in varying combination, of bradycardia and profound arterial vasodilatation in skeletal muscle. Unless the individual lies down, loss of consciousness occurs and the patient falls to the ground. Characteristically the fall is gentle, and self-injury relatively uncommon. In falls associated with tonic–clonic or atonic seizures, the fall is precipitate and injury much more likely. Rarely, in complicated faints, there may be brief clonic jerks of the limbs. More commonly, multifocal myoclonus is observed, lasting a few seconds and following the loss of posture. The eyes tend to remain open. Lateral head turns, repetitive movements (such as lip licking), and hallucinations are all recognized features. After the episode there may be brief confusion and feelings of weakness, but these rapidly resolve. If, on the other hand, the upright posture is maintained (typically the individual is a soldier on parade) then stiffness of the limbs or repetitive generalized shaking occurs which is virtually indistinguishable from the movements occurring with epilepsy. Usually, however, a true tonic–clonic sequence does not occur in these circumstances.

Micturition syncope

Micturition syncope occurs predominantly in males, but of any age group. The attacks are almost always nocturnal, typically after an evening of alcohol consumption. Onset is usually during or shortly after micturition. The warning symptoms are often brief. The attacks seldom occur frequently; if they do, then the individual, if male, is advised to micturate in the sitting position.

Cough syncope

Patients with cough syncope effectively perform a Valsalva manoeuvre during a bout of prolonged coughing. Treatment is directed at the underlying chest condition.

Cardiac syncope

A variety of cardiac abnormalities, all having in common the end result of failing output and reduced cerebral perfusion, are associated with syncopal attacks. Mechanisms include complete heart block, paroxysmal ventricular tachycardia or fibrillation, and supraventricular tachycardia or bradyarrhythmia. In addition to disorders of rhythm, abnormalities of ventricular contractility or obstruction of outflow can have a similar outcome, usually when increased output is required during a period of exertion. Rarely, pedunculated masses within the heart, for example an atrial myxoma, cause outflow obstruction when the patient assumes certain postures. Features suggesting that a cardiac lesion may be responsible for a syncopal attack include a history of cardiac disease, palpitations or chest pain in association with the attack, and the finding of cardiac abnormalities on clinical examination.

Separate from these mechanisms are cases of syncope associated with postural hypotension. Autonomic failure resulting in postural hypotension is a feature of: multisystem atrophy; certain neuropathies with autonomic fibre involvement, such as diabetes; and drug therapy, for example with phenothiazines and tricyclic antidepressants. The correct diagnosis is usually readily established from the history.

Carotid sinus syncope

Patients with this condition usually present either with vertigo or with syncopal attacks. The syncopal attacks are sometimes followed by flushing and may be triggered by pressure over the neck, for example during neck rotation. In most patients, the syncope is related to atrioventricular block or asystole. Occasionally, a pure vasodilator reaction occurs, with peripheral pooling of blood.

Transient ischaemic attacks

These attacks should seldom be confused with epilepsy. In some patients with carotid occlusion (or severe stenosis), attacks of limb shaking occur in which involuntary limb movements described as shaking, trembling, or twitching occur, usually for seconds. The movements, which are coarse and irregular, predominate distally. Sometimes the attacks coincide with limb weakness or speech difficulty. The attacks are not influenced by anticonvulsants but can be relieved by endarterectomy where there is an underlying carotid stenosis.

Migraine

Loss of consciousness is a recognized feature of basilar migraine. The condition presents in children or adolescents. The headache is occipital. Visual disturbances are common, along with altered sensations (typically bilateral), ataxia, and dysarthria. Typically the patient, if unconscious, can be roused. Rarely, tonic–clonic seizures are seen with the attacks.

Hyperventilation

Most patients with the hyperventilation syndrome do not develop carpopedal spasm or tetany. Rather, they have a constellation of symptoms which are liable to be confused with other conditions such as epilepsy. Those symptoms include dizziness or vertigo, weakness, paraesthesiase, chest pain, and altered consciousness. Probably some 5 to 15 per cent of patients lose consciousness during hyperventilation, but never with a tonic–clonic progression that would cause real diagnostic difficulty.

Narcolepsy and cataplexy

Narcolepsy is defined as excessive daytime sleepiness, often occurring under unusual circumstances. The onset of sleep is usually preceded by a feeling of tension, tiredness, or a noise in the head. In some patients, onset occurs without warning. At times, patients have periods of semi-automatic behaviour for which they may subsequently be amnesic.

Cataplexy is typically triggered by sudden arousal. Attacks are brief, and may lead to such loss of muscle control that the patient falls. During the attack, the patient is flaccid, the eyes may roll or diverge, and the facial muscles flicker. Despite this, the patient usually remains fully alert.

Drop attacks

Drop attacks are almost confined to women in the last third of life. Typically, while walking, the patient drops to their knees without warning. The patient is aware of the fall, and is usually able to get up quickly, providing there is no injury. The attacks occur in otherwise fit individuals, are not due to vertebrobasilar ischaemia, and eventually remit completely. They are untreatable.

The parasomnias

Parasomnias are largley confined to children. They consist either of abnormal motor activity or excessive autonomic activity. Motor activity includes sleep starts, sleep myoclonus, bruxism, and head banging. Sleep myoclonus produces repetitive leg contraction, typically dorsiflexion of the feet. It increases with age and is usually idiopathic. Head banging, which may coincide with body rocking, is usually only seen in children or infants. The movements, typically occurring in clusters, are often accompanied by various forms of vocalization. In most cases, the child is normal. Sleep terrors usually happen within the first hour or two of sleep, occur in children, and result in a sudden cry followed by anxiety, tachycardia, sweating, and hyperkinesis. The child is not completely aware of the episodes, which sometimes necessitate short-term treatment with benzodiazepines.

Non-epileptic seizures

Non-epileptic seizures sometimes occur in isolation, but sometimes in those with true epilepsy. They account for 20 per cent of the patients referred to specialist epilepsy units, usually with a diagnosis of intractable epilepsy. The vast majority of sufferers are women. They are more likely to have a family history of psychiatric disorders, a past personal history of psychiatric disorder, a history of suicide attempt(s), evidence of sexual maladjustment, and current depressive symptoms. Indeed there is a substantial overlap, in terms of clinical characteristics, between non-epileptic seizures and multiple personality disorder. In addition to the features noted above, up to 90 per cent of patients give a history of sustained trauma, including childhood abuse, which may have been physical or sexual.

Certain features from the history should alert the physician. The attacks usually take place with witnesses present. They develop gradually rather than suddenly and the movements displayed are often unpredictable and bizarre. Attempt to constrain the patient are resisted. Vocalization is common. Incontinence is uncommon and tongue biting particularly so, but self-injury is a recognized feature. Typically the seizures are difficult to control. Serum prolactin levels taken 20 min after the event are normal, in contrast to tonic–clonic seizures where they are commonly, though not inevitably, elevated. Videotelemetry has proved of considerable value in differentiating epileptic from non-epileptic seizures. Management is extremely difficult. Drug withdrawal is resisted by the patient, who often resents suggestions of psychiatric referral and exploration of psychological morbidity.

Investigation

Investigation of a patient with suspected epilepsy (or a single seizure) is performed for three main reasons. The investigation may provide valuable support for the diagnosis, may give an indication as to which part of the brain has initiated the seizure, and finally, imaging may allow a statement as to the underlying structural process, where such exists.

Routine haematological and biochemical tests should be undertaken in all patients with suspected epilepsy although they seldom point to a metabolic disturbance that has not already been recognized.

Electroencephalography

Certain facts about the electroencephalogram must be understood before interpretation is attempted. Epileptiform discharges are encountered in between 0.5 and 4 per cent of individuals who have never had a seizure and who do not do so during a period of follow-up. Furthermore, a routine electroencephalogram in adults with established epilepsy shows epileptiform abnormalities in only some 40 to 50 per cent of cases. With repeat recording, with or without sleep records, the figure rises to 70 or 80 per cent. In other words, some patients with unequivocal epilepsy will have persistently normal or, at least, non-epileptic electroencephalograms. Serial electroencephalographic recording is sometimes helpful in an attempt to define the origin of the seizure and to delineate the seizure type better. If photosensitivity is suspected (10 per cent of individuals with seizures occurring between 1 and 7 years are photosensitive), serial recordings are appropriate, as they are in any individual with atypical status, or in whom cognitive impairment might be due to subclinical epileptic activity. Where surgical intervention is being planned for the epilepsy, routine and sleep recordings are followed by videotelemetry in order to record individual attacks. For some patients, depth electrodes will be needed to establish the seizure source. Magnetoencephalography localizes focal epileptic discharges by measuring the changes in extracranial magnetic fields which these discharges generate. The system costs some 25 times as much as a conventional electroencephalographic system. Depth electrodes are positioned stereotactically at sites determined by clinical and surface electroencephalographic criteria. Depth recordings are more accurate and sensitive in detecting focal discharges than either nasopharyngeal or sphenoidal electrodes.

The electroencephalogram has also been used to attempt prediction of seizure recurrence in individuals after a single seizure of unknown cause. Epileptic discharges, in one series, predicted a seizure recurrence over 2 years of 83 per cent, compared with a 12 per cent rate in individuals with a normal recording. The electroencephalogram has also been used to predict seizure recurrence during or after drug withdrawal in someone whose epilepsy has gone into remission on medication. The predictive value of electroencephalographic abnormalities in such cases has varied widely from series to series.

CT scanning

Neuroimaging is carried out in order to define whether a structural abnormality underlies the patient's epilepsy and, if so, whether some additional treatment, other than anticonvulsants, might be required. CT scanning was originally the most often used imaging process, prior to the more widespread availability of MRI. Some authors advocate MRI in all patients with epilepsy, other than for those epilepsies which are clearly idiopathic (such as absence seizures, juvenile myoclonic epilepsy, and benign rolandic epilepsy). In practice, this is probably unreasonable. For example, a patient with the onset of epilepsy in their 70s or 80s, who has a normal CT (at least, with no evidence of focal pathology) hardly merits MRI if the epilepsy is well controlled.

MRI

MRI is undoubtedly both more sensitive and more specific than CT in detecting small brain lesions and abnormalities of the cerebral cortex thought to be relevant in the genesis of epilepsy (Fig. 3). The most common abnormalities detected are hippocampal sclerosis, malformations of cortical development, vascular malformations, tumours, and acquired cortical damage. MRI is particularly indicated for partial seizures, onset of generalized or unclassified seizures in adult life, patients with fixed focal clinical or neuropsychological deficit, and for those patients with poor seizure control. Quantitative measures of the hippocampi improve the diagnostic sensitivity of MRI for hippocampal sclerosis. MRI is much more sensitive than CT for detecting malformations of cortical development. Magnetic resonance spectroscopy, examining nuclei ^{31}P and ^{1}H, has been used for assessment of patients with complex partial seizures for possible surgery.

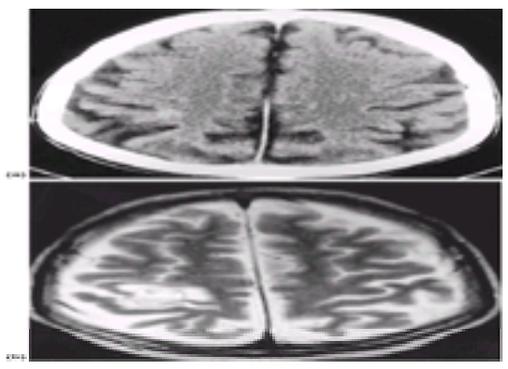

Fig. 3 CT scan (a) and MRI scan (b). The readily visible cavernome on MRI is only just visible on CT.

Single photon emission computed tomography (SPECT)

This technique allows measurement of cerebral blood flow and of specific brain receptors. Both ictal and interictal studies have been performed. Ictal SPECT can achieve a correct localization of over 90 per cent in unilateral temporal lobe epilepsy.

Positron emission tomography (PET)

PET scanning is used to measure cerebral blood flow, regional cerebral glucose metabolism, and the distribution of specific receptors, such as the benzodiazepine–$GABA_A$ receptor complex (Fig. 4 and Plate 1). The spatial resolution achieved is superior to SPECT. Epileptic foci, studied interictally, display reduced blood flow and reduced glucose metabolism. Typically, the abnormalities found are more extensive than the corresponding pathological lesion. Increasingly it appears likely that MRI and functional MRI will largely replace this technique in patient evaluation.

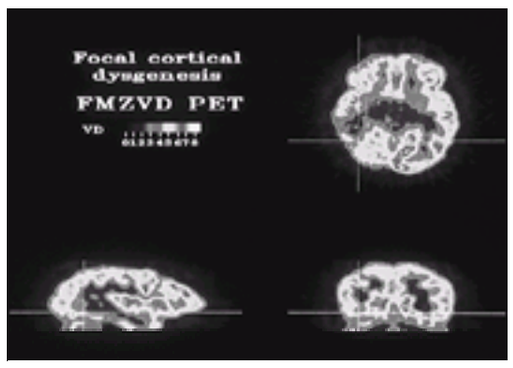

Fig. 4 FMZVD PET scan showing a region of probable cortical dysplasia in the right temporal lobe. The 11C-fluamzenil volume of distribution (FMZVD) is an index of $GABA_A$ receptor density. (See also Plate 1.)

Treatment—drug therapy

Choice of drug therapy

A number of principles can be stated in relation to drug therapy.

Does the patient require anticonvulsants?

The issue of whether isolated seizures should be treated remains unresolved. Seizure recurrence rate after a single seizure reaches 80 per cent in untreated individuals, the vast majority recurring within 2 years of onset. Many patients prefer to defer treatment after a single seizure, a decision substantially influenced by how soon they wish to start driving. For a patient who has very infrequent seizures, say 5 or more years apart, it may seem logical to withhold medication.

Choice of anticonvulsant

An algorithm can provide some guidelines regarding drug treatment (Fig. 5). For generalized seizures (tonic–clonic, absence, or myoclonic) sodium valproate is the drug of choice. Further choices are determined by seizure type. There is no controlled trial data indicating the most appropriate add-on drug or combination of drugs. Myoclonus can be exacerbated by carbamazepine, gabapentin, and lamotrigine and absences by carbamazepine and gabapentin. For partial seizures, with or without generalization, carbamazepine and valproate are probably the drugs of choice, though controlled studies have indicated that phenobarbitone and phenytoin are of comparable value.

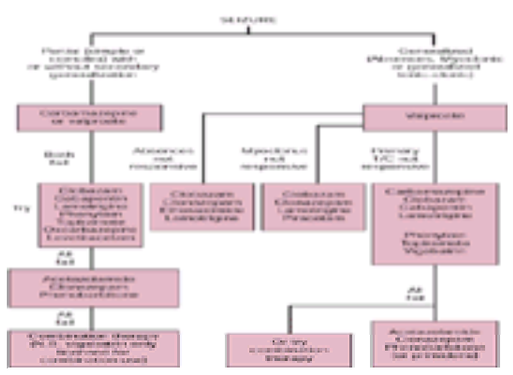

Fig. 5 Choice of anticonvulsant.

In addition, choice of drug will be influenced by the patient's age, sex (regarding the use of oral contraceptives and likelihood of pregnancy), and reliability of adherence to a particular drug regime. The patient should always be started on a single drug.

Dosage

Although standard dose regimes tend to be quoted, many anticonvulsants are sometimes effective in relatively low doses. Accordingly the drug is introduced in low dosage and is then gradually increased according to need and tolerance. Sometimes only dosages that lead to toxic serum levels appear effective. Some patients tolerate such toxic levels without difficulty.

Failure of first drug

When this occurs, a second drug should be gradually introduced without withdrawing the first. If the patient responds, the drug used originally can be slowly withdrawn.

Drug combinations

If drugs given individually have failed then drug combinations should be considered, remembering that they may interact with each other.

Generic prescribing

The bioavailability of the anticonvulsant drugs should be unaffected by whether they are prescribed generically, or as a specific branded product. Patients sometimes disbelieve this assumption and prefer branded products. If they are given generic prescriptions, they should be warned that the appearance of their medication may

change from prescription to prescription.

The problem of non-compliance

Non-compliance is a significant problem with the anticonvulsants and is a potent cause of poor control. A full explanation of each drug's side-effect profile and its potential interactions is essential and appears conducive to improved compliance. Drugs that are given once or twice a day are preferred to ones needing more frequent prescriptions. Slow-release preparations allow drug regimes to be simplified.

Mechanisms of action (Fig. 6)

The prime role of GABA-mediated inhibition in the epileptic process implies that drugs which enhance $GABA_A$-receptor-mediated inhibition will have anticonvulsant activity. The $GABA_A$-receptor complex comprises at least three subunits, a, b, and †, which appear to combine as a five-membered structure forming an anion-permeable channel. Both barbiturates and benzodiazepines act by potentiating $GABA_A$-mediated inhibition. The barbiturates bind to the b-subunit to potentiate action of endogenous agonist GABA and prolong the opening time of the chloride ion channel. Benzodiazepines bind to the a-subunit to potentiate the action of GABA and increase the frequency of opening of the chloride ion channel. GABA is metabolized by GABA transaminase. Vigabatrin irreversibly binds to GABA transaminase to inhibit degradation of GABA and thereby elevate brain GABA levels. GABA-mediated inhibition can also be enhanced by blocking GABA uptake into glia and neurones after its release into the synaptic cleft during synaptic transmission.

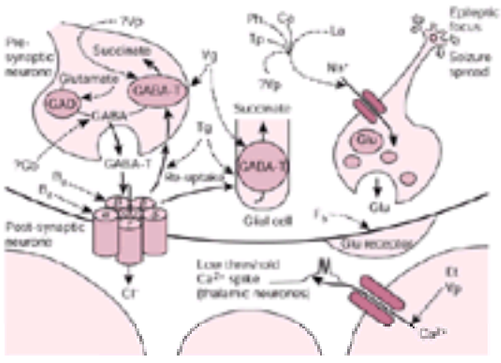

Fig. 6 Mechanism of action of some of the anticonvulsants. Ba, barbiturate; Bz, benzodiazepine; Ca, carbamazepine; Et, ethosuximide; Fb, felbamate; Gb, gabapentin; La, lamotrigine; Ph, phenytoin; Tg, tiagabine; Tp, topiramate; Vp, valproate; Vg, vigabatrin.

Tiagabine blocks uptake of synaptically released GABA into both presynaptic neurones and glial cells, allowing GABA to remain at its site of action for longer periods. Gabapentin acts presynaptically to promote GABA synthesis or release.

The second major neurotransmitter system involved in the genesis of epileptic activity is excitatory utilizing glutamate and, perhaps, aspartate as neurotransmitters. They act on several different receptors including a-amino-3-hydroxy-5-methylisoxazole-proprionic acid (AMPA) and N-methyl- D-aspartate (**NMDA**). The NMDA receptor is activated by glutamate or aspartate in conjunction with glycine. Blockade of the NMDA receptor results in antiepileptic effects.

Voltage-dependent calcium ion currents are thought to be of importance in the genesis of epileptic events. Ethosuximide acts by inhibition of one class of voltage-dependent calcium ion currents (T currents). Valproate may have a similar role.

Regulation of sodium channels also appears of relevance in the modification of the epileptic process. Phenytoin, carbamazepine, and possibly valproate reduce the rate of recovery from inactivation of depolarized voltage-dependent sodium channels, thereby blocking sustained repetitive firing of action potentials in depolarized neurones. Lamotrigine inhibits glutamate and aspartate release, suggesting it may act at voltage-dependent sodium channels to decrease the presynaptic release of glutamate. Lamotrigine may have additional effects on calcium channels. Oxcarbazepine may act by reducing glutamate release via a blocking action on presynaptic calcium channels. Topiramate influences sodium channel activity, suggesting that its anticonvulsant properties are similar to those of phenytoin. Felbamate probably acts primarily through its effects on the NMDA receptor.

Selected drugs

Carbamazepine

Carbamazepine is a first-line drug for both partial seizures and for generalized tonic–clonic seizures. In its standard form, it needs to be given three times per day, but a slow-release preparation allows twice daily prescribing. Dosage ranges from 300 to 1600 mg/day. Sedation is common and the drug should be introduced slowly. A drug rash occurs in perhaps 3 per cent of patients and demands immediate drug withdrawal. Signs of intoxication include drowsiness, blurred vision, and dizziness. Leucopenia occurs and can lead to a frank aplastic anaemia. Hyponatraemia and oedema are recognized features, associated with a mild degree of inappropriate antidiuretic hormone production. The drug influences atrioventricular conduction and should not be given to patients with atrioventricular conduction abnormalities unless they are already paced. The relationship between dosage and plasma concentrations is linear. Carbamazepine is a liver-enzyme inducer and is teratogenic (see below).

Sodium valproate

Sodium valproate is considered, at least by some physicians, to be the drug of choice for all epilepsy types. It is not enzyme-inducing, and therefore does not influence the metabolism of the oral contraceptive. Liver toxicity is a recognized, though rare, hazard. Elevated liver enzyme levels are more common, but usually return to normal without need for drug withdrawal. Thrombocytopenia occurs rarely. Gastrointestinal effects are fairly common. Nausea and weight loss are seen, but appetite stimulation with weight gain is more common. Tremor occurs, as a dose-related effect. Hair loss, of a mild degree, is not uncommon. After a few months, hair regrowth occurs, often more curly than before. Sedation is less troublesome than with other anticonvulsants. Disturbances of menstruation are recognized. It has been suggested that the drug can trigger polycystic ovarian disease. The dose ranges from 600 to 2500 mg/day and it is given twice or three times per day. A slow-release preparation can be given once daily. Plasma levels are not a useful guide to efficacy.

Other drugs

Phenytoin

Experience with phenytoin is vast and despite its side-effect profile and complex pharmacokinetics, large quantities of the drug continue to be prescribed. A 100 mg tablet, in the United Kingdom, costs approximately one-thirtieth of the price of a comparable dose of lamotrigine. The drug is effective in both generalized tonic–clonic seizures and in the partial seizures. It has a long half-life, and can be given once daily, conveniently at bed time. Sedation is common. Toxic effects, generally dose related, include drowsiness, ataxia, confusion, blurred vision, and dizziness. Most patients who are intoxicated with the drug have nystagmus. Permanent cerebellar ataxia and peripheral neuropathy are recorded. Other side-effects or toxic effects include rashes, gum hypertrophy, thickening of the facial features, chorea, and sleep disturbance. The drug is a potent enzyme-inducer and is teratogenic. The relationship between dosage and plasma concentrations is non-linear. Once the dose exceeds 300 mg/day, increments should be pegged to 50 mg or even 25 mg at a time.

Lamotrigine

Lamotrigine is licensed for both generalized and partial seizures. Occasionally it exacerbates myoclonus. Doses seldom exceed 400 mg/day. A drug rash occurs in

about 3 per cent of patients. It interacts with enzyme-inducing anticonvulsants, which lower its plasma level. Valproate enhances lamotrigine levels. The drug can be given once daily. It is said not to be teratogenic.

Phenobarbitone

Phenobarbitone is a very effective anticonvulsant but often badly tolerated. Children may become hyperactive on the drug, adults (and particularly the elderly) heavily sedated. Doses of up to 180 mg/day are used. It has a long half-life and can be given once daily. Rapid withdrawal of phenobarbitone in non-epileptic patients can trigger seizures. Over-rapid withdrawal in someone with epilepsy can trigger status epilepticus. Methyl phenobarbitone is largely converted to phenobarbitone by the liver and phenobarbitone is the main metabolite of primidone, although primidone's other metabolite phenylethylmalonamide probably possesses anticonvulsant activity.

Vigabatrin

Vigabatrin is probably a more potent anticonvulsant than many of the other recently introduced drugs. Increasingly, it has been recognized to cause retinal damage. Up to a third of patients develop concentric constriction of the visual fields, more marked nasally than temporally. The defect is often asymptomatic and probably irreversible. It is now recommended that vigabatrin should only be used as add-on therapy where other combinations have been unsuccessful. Dosage should not exceed 3 g/day. Regular visual field analysis is mandatory.

Gabapentin

Gabapentin is used as add-on therapy for partial seizures with or without secondary generalization. Up to 4.8 g is given in three divided doses. The drug is generally well-tolerated and does not interact with other anticonvulsants. Its anticonvulsant effect appears to be relatively weak.

Ethosuximide

Ethosuximide is seldom used in adults as its role is confined to the treatment of absence seizures. Gastrointestinal disturbances are common along with drowsiness, dizziness, and ataxia. Agranulocytosis or aplastic anaemia have been encountered rarely. The dose range is usually 1 to 1.5 g daily.

Clonazepam

Clonazepam is effective for tonic–clonic seizures but is particularly valuable in the treatment of myoclonic epilepsy. Sedation is a major problem, and the drug must be introduced cautiously. The maximum tolerated dose is about 8 mg/day.

Clobazam

Tolerance to clobazam tends to develop fairly readily. It is sedative. Adult dosage ranges from 30 to 60 mg daily. Used intermittently it can be very effective for the treatment of catamenial epilepsy.

Acetazolamide

Use of this drug is largely confined to childhood epilepsies.

Topiramate

This drug is licensed both for primary generalized tonic–clonic seizures, and as adjunct therapy for partial seizures. It is sedative and must be introduced slowly. The total daily dose (given as a twice daily regime) seldom exceeds 400 mg. Nausea, anorexia, and weight loss are encountered. Behavioural disturbances are reported, including emotional lability, mood change, and aggression. There is an increased incidence of renal stones in those taking the drug.

Tiagabine

Tiagabine is a GABA uptake inhibitor resulting in increased synaptic GABA levels. Initial doses in adults are 4 to 5 mg twice daily. Most studies have used 32 to 56 mg/day, in three divided doses. The drug is licensed as add-on therapy in refractory epilepsy. Side-effects include dizziness, tiredness, tremor, and altered mood.

Oxcarbazepine

This drug is closely related to carbamazepine. It is a less potent hepatic enzyme inducer however. It is licensed as monotherapy or adjunctive therapy, in partial seizures with or without secondary generalization. Its side-effect profile is similar to that of carbamazepine. Patients who are hypersensitive to carbamazepine should not receive oxcarbazepine. The dosage range lies between 600 and 2400 mg daily, in adults.

Levetiracetam

The mode of action of levetiracetam is not understood. It is not metabolized in the liver nor does it inhibit or induce hepatic enzymes. There are no known interactions with the other anticonvulsants. Two-thirds of an oral dose is excreted unchanged in the urine. A quarter is metabolized to an inactive metabolite, also excreted in the urine.

Levetiracetam is licensed as adjunctive therapy in the treatment of partial seizures with or without secondary generalization. The daily dose in adults ranges from 1000 to 3000 mg. The dose needs to be adjusted in the presence of renal impairment. It is not advised for use in pregnancy.

Side-effects include asthenia, somnolence, headache, gastrointestinal disturbances, mood changes, and skin rash.

Other drugs, with very restricted licences, or not yet licensed, include felbamate and zonizimade.

Particular issues

Enzyme-induction

Drugs that induce liver enzymes (phenytoin, phenobarbitone, carbamazepine, topiramate, and possibly lamotrigine) will alter the pharmacokinetics of other agents or drugs which undergo hepatic metabolism. Women on an oral contraceptive pill need to take a preparation containing at least 50 µg of ethinyloestradiol. If breakthrough bleeding still occurs, the dose of oestrogen can be increased to a maximum of 100 µg daily. Alternatively, an injectable long-term contraceptive can be used. The interactions between anticonvulsants are complex, another reason for avoiding drug combinations where possible.

All the enzyme-inducing anticonvulsants have the potential for accelerating vitamin D metabolism. Those individuals at risk for developing vitamin D deficiency (for example due to poor nutrition) are at risk of developing osteomalacia or rickets when taking certain anticonvulsants.

Drug monitoring

Anticonvulsant levels are measured far too frequently. There are specific circumstances where their measurement is of value:

1. to ascertain compliance;

2. to monitor dosage adjustment with phenytoin; and
3. to ascertain the unpredictable effect of combining anticonvulsant preparations.

Phenytoin undergoes saturable hepatic metabolism. Regular monitoring of the serum level is advisable, particularly after dose adjustment. Occasionally, measurement of the levels of carbamazepine, phenobarbitone, and ethosuximide aids management, particularly where epilepsy control has been poor. Carbamazepine epoxide, a metabolite of carbamazepine, can sometimes be the cause of carbamazepine toxicity even when carbamazepine levels are in the therapeutic range. There is no value in the routine monitoring of levels of valproate, vigabatrin, lamotrigine, gabapentin, topiramate, clonazepam, or clobazam.

When measuring levels, the same time after the last dose should be used, wherever possible. Examples of therapeutic serum levels are given in Table 2. The therapeutic ranges of the anticonvulsants should be interpreted with caution. Some patients respond to a drug with subtherapeutic levels. Others need toxic levels to achieve seizure control and can often tolerate such levels without overt difficulty.

Pregnancy

There is an increased risk of congenital malformations in women who have taken anticonvulsants during pregnancy (approximately 4 to 8 per cent overall risk). Most evidence has accumulated for phenytoin, phenobarbitone, valproate, and carbamazepine. There are very few data on the newer anticonvulsants, though lamotrigine is said not to be teratogenic. The critical period for development of the major malformations is from 3 to 8 weeks' gestation.

Phenytoin and phenobarbitone

Both these drugs are associated with cardiovascular malformations (2 per cent risk) and cleft lip/palate syndromes (1.8 per cent risk).

Valproate

Valproate leads to a 2 per cent risk of spina bifida compared with a 0.01 to 0.02 per cent risk for all births. Cardiovascular and urogenital malformations are also recognized to occur.

Carbamazepine

Carbamazepine is associated with spina bifida (1 per cent risk) and hypospadias.

A folic acid supplement of 5 mg daily should be given to women with epilepsy who are taking valproate or carbamazepine and who are contemplating pregnancy. Doses of valproate should be less than 1000 mg/day if possible and slow-release forms of the drug prescribed. For women on other anticonvulsants, a dose of 0.4 mg/day of folic acid suffices.

Seizure frequency increases in pregnancy in about a third of patients with epilepsy. Tonic–clonic seizures are associated with an increased risk of miscarriage. Vitamin K at 20 mg/day should be given in the last month of pregnancy in women on enzyme-inducing drugs to reduce the risk of haemorrhagic disease of the newborn baby.

The epilepsy risk in the offspring of an affected patient is around 2 to 4 per cent but higher where the epilepsy of the parent has a strong genetic basis.

Breast feeding

All the commonly used anticonvulsants are present in low concentrations in breast milk. If the mother is on a barbiturate or a benzodiazepine, significant sedation of the baby is possible. If breast feeding then ceases abruptly, a withdrawal reaction can occur in the infant with tremor and agitation.

Drug withdrawal

Generally medication is continued until at least a 2- to 3-year period free of seizures has been established. Approximately two-thirds of patients remain fit free after drug withdrawal. Factors known to predispose towards recurrence include neurological abnormalities on examination, an underlying structural basis for the epilepsy, the need for multiple drug therapy, and a history of difficulty in establishing initial control. The electroencephalogram is of limited value in predicting outcome although rather better in children than adults. Any drug withdrawal should be gradual, say over 3 to 6 months. Absence seizures usually remit spontaneously in late adolescence, but juvenile myoclonic epilepsy tends to recur after drug withdrawal.

Driving

In the United Kingdom, driving must cease for 1 year after any type of seizure. If a nocturnal pattern of seizures has been established for 3 years, driving can then continue even if nocturnal seizures are still occurring. The Driver and Vehicle Licensing Agency prefers patients not to drive during a period of drug withdrawal, and for 6 months after the withdrawal has been completed. For drivers of heavy goods vehicles a 10-year period of freedom must be established, during which there has been no anticonvulsant use. Furthermore, a continuing liability to epilepsy has to be excluded.

Status epilepticus

Status epilepticus has already been defined. The commonest type is tonic–clonic status. The commonest precipitants are sudden anticonvulsant withdrawal, poor compliance in a patient with known epilepsy, and alcohol abuse. The mortality figures for status epilepticus have varied substantially from series to series. In one recently published, prospective, population-based study, the overall incidence was estimated at 41 to 61 per 100 000 person-years with a mortality of 22 per cent. Incidence rises in the elderly, as does mortality. From other series, overall mortality figures lie between 8 and 37 per cent. At least half the cases occur in the absence of previous epilepsy. Although non-compliance and subtherapeutic drug levels are often quoted as causes of status, several studies have established that the majority of individuals with epilepsy who present in status have therapeutic drug levels at or around the time of presentation. Status in the absence of previous epilepsy is followed by unprovoked seizures in about half the cases.

The diagnosis is by no means straightforward. In one study, half the patients transferred to a specialist centre for management of their status were either in pseudostatus or in drug-induced coma. The diagnosis of pseudostatus should be considered if the attacks are atypical or if the status does not respond to initial therapy.

Analysis of immediate management of patients in status suggests that many are given inadequate loading and maintenance doses of anticonvulsants. The patient should be moved away from possible hazard, such as broken glass, an airway established, and oxygen administered. Lorazepam is probably the drug of choice. It is given in a dose of 0.l mg/kg intravenously at the rate of 2 mg/min. Alternatives included diazepam (Diazemuls) given intravenously in a dose of 10 to 20 mg at a rate of 5 mg/min or clonazepam given in a dose of l mg by slow intravenous injection.

Using the intravenous route, 50 per cent glucose should be administered to a total volume of 50 ml after blood has been taken to establish the glucose concentration. Thiamine in a dose of 250 mg (Pabrinex I/V High Potency) should be given by slow intravenous injection over 10 min if there is suspicion of alcohol withdrawal, but remembering that the infusion can produce an anaphylactic response. In addition to plasma glucose measurement, blood should be taken for urea, electrolytes (including calcium and magnesium), acid-base balance, liver function tests, and full blood count. A serum sample should be stored in case anticonvulsant or alcohol levels are required subsequently. Blood cultures should be performed if the patient is febrile.

If immediate therapy is successful and the patient is receiving phenytoin or valproate, those drugs can be given intravenously before reverting to oral therapy. If the patient is not on anticonvulsants, a phenytoin infusion at 20 mg/kg in 0.9 per cent sodium chloride should be given at a maximum rate of 50 mg/min. An alternative is Fos-Phenytoin, a water-soluble drug, which is metabolized to phenytoin with a half-life of 8 to 15 min. It is given intravenously in the same dose at 150 mg/min in order

to achieve a comparable effect. The drug is more expensive than phenytoin but causes less phlebitis, less hypotension, and is better tolerated.

Midazolam has been developed for intranasal use and may prove of value where immediate intravenous access is difficult, for example in young children.

If phenytoin infusion is unsuccessful, valproate infusions can be used, with 25 mg/kg as a loading dose delivered at 3 to 6 mg/kg/min. If seizures continue phenobarbital can be considered, given at 20 mg/kg intravenously at 50 to 75 mg/min. Intramuscular or rectal paraldehyde is now seldom used, most experts suggesting a move instead to thiopentone, propofol, or midazolam.

Propofol or midazolam are rapidly metabolized and have less hypotensive effects than the barbiturates. The suggested dose of propofol is l to 2 mg/kg followed by a continuous infusion of 2 to 10 mg/kg/h.

For all the therapies used in patients with refractory status, intensive care placement is essential with the patient intubated and haemodynamic monitoring in place.

Sudden death

Patients with epilepsy have an increased risk of death compared with age- and sex-matched controls. Sudden unexpected death in epilepsy predominates in younger age groups and in those with more severe epilepsy. It is likely that most of the deaths are the result of unwitnessed seizures producing either respiratory complications, cardiac arrhythmias, or both.

Surgery

Despite optimal treatment, some 30 per cent of patients with new-onset seizures continue to have attacks. Prerequisite in patient selection for surgery is accurate localization of the epileptic discharge and understanding of circumstances where a resection might prove detrimental in terms of functional deficit.

Assessment for epilepsy surgery demands localization techniques incorporating seizure characteristics, electrophysiological recording, and imaging. Equally important is the recognition by the physician that certain epilepsy syndromes are likely to be resistant to medical therapy and that early rather than delayed referral for surgical opinion is beneficial. Mesial temporal lobe epilepsy, secondary to hippocampal sclerosis, is the commonest cause of medically refractory partial seizures. In most such patients, a unilateral structural abnormality can be confidently established, resection of which leads to a 70 per cent chance of remission. Disabling neurological complications after surgery, such as hemianopia, hemiparesis, or dysphasia, occur in about 2 per cent of patients. Depression and psychosis are recognized complications of temporal lobectomy.

MRI characteristics of mesial temporal sclerosis include atrophy or increased signal on T_2-weighted images. The presence of atrophy is the best predictor for a good surgical outcome. Besides visual inspection, measurement of hippocampal volume and techniques for measuring the T_2 signal change are used to improve sensitivity.

SPECT and PET measure the changes in cerebral blood flow and cerebral glucose metabolism, respectively, which accompany the epileptic process. Both have relatively high sensitivity and moderate specificity for the diagnosis of temporal lobe seizures, but lower sensitivity for epilepsy arising at other sites. Interictal PET and ictal SPECT produce very similar results in predicting outcome after temporal lobectomy.

Proton magnetic resonance spectroscopy can contribute to recognition of the lateralization of the epileptic focus and to the identification of those patients with bilateral changes who are less likely to respond to surgery.

Continuous surface electroencephalographic monitoring is usually undertaken as part of the work-up for patients being considered for surgical intervention. The technique, however, has limitations. It often fails to detect seizure activity arising in areas distant from surface electrodes, such as the orbitofrontal cortex, and may falsely lateralize foci, particularly in the presence of large lesions. For improving electroencephalographic localization, some form of intracranial recording is necessary. Depth electrodes are used to sample deeper structures such as the hippocampus. Electrocorticography is performed at the time of surgery. Subdural electrodes, sometimes with depth electrodes, measure directly from the surface of the exposed brain.

Other less commonly performed surgical procedures include neocortical resections, lesionectomies, hemispherectomies, multilobar resections, and corpus callosotomy. Hemispherectomy is performed when a diffuse epileptogenic region has been localized within one hemisphere, the other hemisphere being normal. Division of the corpus callosum is performed in patients with severe secondary generalized epilepsy who have disabling drop attacks. Cortical dysplasia is increasingly recognized as a cause of intractable epilepsy. MRI criteria have been developed to allow recognition of areas of focal cortical dysplasia and assist in planning the extent of cortical resection.

Vagal nerve stimulation

Vagal nerve stimulation is achieved through the implantation of a small stimulator on the left vagus. The exact mechanism of action remains uncertain. The nucleus of the tractus solitarius, the main terminus for vagal afferents, has projections to the locus caeruleus, raphe nuclei, reticular formation, and other brainstem nuclei. These nuclei have been shown to influence cerebral seizure susceptibility. In patients with chronic partial seizures, there is reduction in the number of seizures, rather than their elimination. The long-term role of this procedure is not yet determined.

Psychiatric aspects of epilepsy

A substantial proportion of patients with poorly controlled epilepsy are likely to have psychiatric symptoms. Those symptoms may partly reflect the underlying structural process in the brain, the effects of repeated seizures, the effects of any social stigma attached to the diagnosis, and as a reaction to the patient's anticonvulsants. Psychiatric symptoms occurring around the time of the seizures tend to be affective or cognitive if before or with the seizure, but psychotic afterwards. Additional psychiatric morbidity is encountered as an interictal phenomenon. It correlates with multiple drug use, the serum concentrations of those drugs, and certain of the anticonvulsants including the newer agents, such as lamotrigine, vigabatrin, and topiramate.

Patients with poorly controlled epilepsy may require referral to a clinical psychologist, partly with a view to helping in the psychological adjustment to the condition, and partly to identify specific areas of cognitive impairment which might require attention.

The role of specialist nurses and the general practitioner

Patients almost inevitably indicate some dissatisfaction with the level of information and support that they receive for their epilepsy. Studies suggest that improvement in these areas can occur using a specialist nurse, working either in general practice or in association with a hospital clinic. Where joint care is to be achieved between general practice and hospital, it is vital that good quality communication and record keeping are achieved. Giving the patient files which document vital information, including their drug regime, is valuable. Patients prefer the continuity of care achievable through seeing the same doctor at each consultation and are more likely to adhere to medical advice under those circumstances.

Prognosis

Prognosis for patients with epilepsy followed in the community is considerably better than for a hospital-based population. Figure 7 records the percentage of patients in remission (defined as being seizure free for 5 years). The top curve indicates the percentage of patients achieving a 5-year period of remission at any time during the 20-year period of follow-up. The middle curve refers to those patients in remission for at least the last 5 years at the time of sampling. The difference between the top and middle curves represents those patients who have relapsed after achieving a 5-year remission. The bottom curve indicates the probability of being in remission whilst not taking anticonvulsants. The curves in Fig. 7 flatten off, indicating that remission becomes less likely the longer the seizures persist. Factors that influence outcome adversely include a combination of complex partial and tonic–clonic seizures, clustering of seizures, abnormal physical signs, and the presence of learning difficulties.

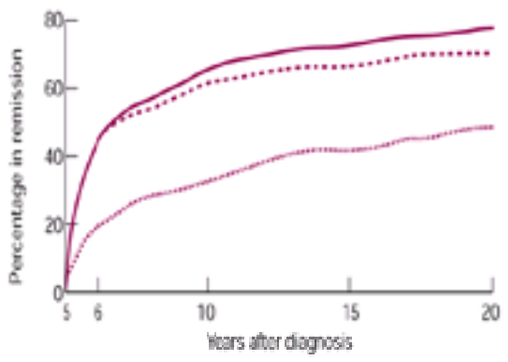

Fig. 7 Probability of seizure recurrence after a first epileptic seizure. (Data from the National General Practice Study of Epilepsy, reproduced by kind permission.)

Overall care

For many patients, shared care between hospital, a specialist nurse, and general practice is ideal. Such an arrangement necessitates a reasonable level of epilepsy experience from the general practitioner, allowing many issues to be resolved without recourse to hospital consultation. The complexities of epilepsy care in terms of new drug developments, issues relating to pregnancy, the question of non-epileptic seizures, and the potential for surgery for many patients with poorly controlled epilepsy makes the case for epilepsy clinics manned by physicians with a particular interest in epilepsy.

Further reading

Arruda F *et al*. (1996). Mesial atrophy and outcome after amygdalohippocampectomy or temporal lobe removal. *Annals of Neurology* **40**, 446–50.

Berg AT, Shinnar S (1994). Relapse following discontinuation of anti-epileptic drugs: a meta-analysis. *Neurology* **44**, 601–8.

Berkovic SF, Scheffer IE (1997). Epilepsies with single gene inheritance. *Brain Development* **19**, 13–18.

Bowman ES (1993). Etiology and clinical course of pseudoseizures. Relationship to trauma, depression and dissociation. *Psychosomatics* **34**, 333–42.

Crawford P *et al*. (1999). Best practice guidelines for the management of women with epilepsy. *Seizure* **8**, 201–17.

Dichter MA (1994). Emerging insights into mechanisms of epilepsy: implications for new antiepileptic drug development. *Epilepsia* **35**(Suppl 4), S51–S57.

Duncan JS (1997). Imaging and epilepsy. *Brain* **120**, 339–77. [An excellent review article, covering all aspects of imaging, including MRI, magnetic resonance spectroscopy, single photon emission computed tomography, and positron emission tomography.]

Goldstein LH (1990). Behavioural and cognitive-behavioural treatment for epilepsy: a progress review. *British Journal of Clinical Psychology* **29**, 257–69.

Handforth A *et al*. (1998). Vagus nerve stimulation therapy for partial-onset seizures. A randomised active-control trial. *Neurology* **51**, 48–55.

Lempert T, Bauer M, Schmidt D (1994). Syncope: a videometric analysis of 56 episodes of transient cerebral hypoxia. *Annals of Neurology* **36**, 233–7.

Manford M *et al*. (1992). The national general practice study of epilepsy applied to epilepsy in a general population. *Archives of Neurology* **49**, 801–8.

Mattson RH *et al*. (1985). Comparison of carbamazepine, phenobarbital, phenytoin, and primidone in partial and secondarily generalized tonic–clonic seizures. *New England Journal of Medicine* **313**, 145–51.

Nashef L, Brown SW, eds (1997). Epilepsy and sudden death. Proceedings of an international workshop. *Epilepsia* **38**(Suppl 11), S1–S76.

Raymond AA *et al*. (1995). Abnormalities of gyration, heterotopias, tuberous sclerosis, focal cortical dysplasia, microdysgenesis, dysembryoplastic neuroepithelial tumour and dysgenesis of the archicortex in epilepsy. Clinical, EEG and neuro-imaging features in 100 adults patients. *Brain* **118**, 629–60.

Ridsdale L *et al*. (1997). The effects of nurse-run clinics for patients with epilepsy in general practice. *British Medical Journal* **314**, 120–2.

Sander JWAS, Shorvon SD (1996). Epidemiology of the epilepsies. *Journal of Neurology, Neurosurgery and Psychiatry* **61**, 433–43.

Saygi S *et al*. (1992). Frontal lobe partial seizures and psychogenic seizures: comparison of clinical and ictal characteristics. *Neurology* **42,** 1274–7.

Shorvon S (1994). *Status epilepticus: its clinical features and treatment in children and adults.* Cambridge University Press.

Sperling MR *et al*. (1996). Temporal lobectomy for refractory epilepsy. *Journal of the American Medical Association* **276**, 470–5.

Van Donselaar CA *et al*. (1992). Value of the electroencephalogram in adult patients with untreated idiopathic first seizures. *Archives of Neurology* **49**, 231–7.

Wallace H *et al*. (1997). *Adults with poorly controlled epilepsy*. Royal College of Physicians of London. [An excellent, short, monograph whose title is deceptive. Many issues are covered, including the role of electroencephalography, neuroimaging, and the issues relating to contraception and pregnancy.]

24.13.4 Narcolepsy

David Parkes

The narcoleptic syndrome

The prevalence of narcolepsy in Western Europe is about four per 10 000. Reported rates worldwide vary 100-fold. This variation may result in part from differences in true case ascertainment and broad definitions of 'narcolepsy'. Narcolepsy is as common in men as women. It usually starts in childhood or adolescence, with a second peak at 30 to 40 years. However, the age at presentation extends from under 1 to over 70 years. Narcolepsy is lifelong and spontaneous recovery does not occur.

Narcolepsy has two key features, cataplexy and daytime sleepiness. Cataplexy is unique to the syndrome but daytime sleepiness has many different causes. The presence of both symptoms is essential for a definite diagnosis. The initial symptom is usually sleepiness, followed by cataplexy within 2 years. Very occasionally this order is reversed and the gap may be prolonged to several decades. Cataplexy is provoked by emotional stimuli such as laughter, startle, excitement, or anger. There is a sudden loss of tone in antigravity muscles with a tendency to fall as well as mouth opening, dysarthria, mutism, and phasic muscle jerking around the mouth. Most attacks are mild and last a few seconds but self-injury can occur in more severe episodes. Several attacks may occur each day. Cataplexy is comparable to the atonia of rapid eye movement sleep but without loss of awareness. As cataplexy is seldom witnessed by the physician, an unequivocal history using clear language is essential to establish its presence.

Sleep, automatic behaviour, and failure of self-monitoring all reduce wakefulness. Sleep pressure increases with monotony and sleep attacks have a 3- to 4-h cyclicity throughout the day. Sleepiness is usually more disabling than cataplexy and causes chronic school and work failure, broken relationships, frustration, embarrassment, poor self-image, and depression. Despite daytime sleepiness, the total sleep period over 24 h is normal as multiple nocturnal arousals shorten night sleep. Insomnia, motor disorders bordering sleep, and vivid dream intrusion at the wake–sleep boundary are common. The old concept of four symptoms—sleepiness, cataplexy, sleep paralysis, and hypnagogic hallucinations—needs revision.

The diagnosis of narcolepsy is totally dependent on a clear sleep–wake history, supported by a sleep–wake diary and a disability rating scale such as the Epworth or Ullanlinna. Study in a sleep laboratory is never a substitute (Table 1).

Aetiology

Narcolepsy has a genetic not a psychological basis. Ninety five per cent of Caucasian subjects with narcolepsy have the HLA D-related (DR) serotype DR2 and oligotype DQ B1*0602. The DR association is slightly different in other ethnic groups. The HLA type is not specific to narcolepsy, being present in 25 to 30 per cent of white subjects, only 1 in 500 of whom have narcolepsy. Despite the HLA association there is no present evidence of an immune defect in narcolepsy, or of definite involvement of HLA systems in normal sleep mechanisms. Attempts to identify additional non-HLA genes in humans, with the possible exception of linkage to g-aminobutyric acid genetic systems on chromosome 4 and the monoamine oxidase X-linked locus, are unconvincing.

An unexpected pathway for sleep has been identified involving hypocretin. The hypocretins are neuropeptides made in the dorsal and lateral hypothalamic areas of the brain, and are implicated in the regulation of feeding and energy (Table 2). A mutation in a hypocretin receptor has been shown in dogs with narcolepsy, but not in DR2 positive humans with the disease. However, in most narcoleptics studied there is a near total absence of hypocretin in the cerebrospinal fluid and brain tissue. Further understanding here, and of the relationship between HLA and hypocretin systems, will be essential in the development of new treatments for narcolepsy.

Familial narcolepsy does occur but accounts for only 5 to 10 per cent of all cases, some of whom show dominant inheritance.

In the few reported studies of monozygotic twins, disease discordance has been found in up to 70 per cent of pairs. Narcolepsy thus appears to be caused by an environmental agent in a genetically susceptible subject. The nature of this external factor is not known.

Pathophysiology of narcolepsy and sleep laboratory investigation

The normal non-rapid eye movement–rapid eye movement sleep cycle of adults is reversed in narcolepsy, where periods of rapid eye movement occur at the onset of sleep. In addition two characteristics of rapid eye movement sleep, atonia and dreaming, are fragmented and intrude into wakefulness, resulting in cataplexy and vivid recall of dreams. Rapid eye movement sleep is generated in ventral pontine areas contiguous with neurones controlling voluntary eye movements. However, no anatomical defect in this area of the brain has been found in narcolepsy, and pursuit and saccadic eye movements are normal.

The mean sleep latency test measures the pressure for sleep using electroencephalographic parameters at 2-h intervals from 10.00 to 16.00 or 18.00 under standard conditions. A mean latency of under 5 min is usually considered abnormal and of over 10 min normal. Results between 5 and 10 min are in a grey area. Results of the mean sleep latency test do not always mirror behavioural tests of alertness. In addition to sleep latency, a mean sleep latency test will show rapid eye movement activity at sleep onset. However, this activity is not diagnostic of narcolepsy. Timing of rapid eye movement sleep varies with recording position, age, previous sleep deprivation, and drug treatment. Overall, polysomnogram findings are less specific and sensitive in the diagnosis of narcolepsy than a definite history of cataplexy. Diagnosis and treatment should therefore depend on the clinical picture rather than on sleep laboratory findings. In cases with an indefinite history, laboratory findings may add confusion rather than clarity and should never be used alone to establish diagnosis. Likewise, HLA tests will not confirm narcolepsy, although the diagnosis is unlikely if the DR2 antigen is absent.

Symptomatic narcolepsy

Narcolepsy has been associated with a wide range of other disorders. These are mostly uncommon and rarely if ever cause problems in differential diagnosis. In many cases there is obvious brainstem pathology and poor resemblance of symptoms with those of true narcolepsy. The occasional association of narcolepsy with multiple sclerosis may result not from a brainstem lesion but from a common genetic predisposition. There is an over-representation of the same HLA D-related antigen in both conditions.

Differential diagnosis

There are a number of narcoleptic syndrome variants with overlapping clinical features (Table 3).

In narcolepsy the commonest mistake is failure to diagnose, rather than incorrect diagnosis. Daytime sleepiness is never normal and is rarely psychological in origin. It is sometimes wrongly attributed to insomnia and most insomniacs have a low, not high, daytime sleep tendency. The symptom is serious and requires investigation rather than a pseudodiagnosis of laziness.

The second most common mistake is to label all forms of sleepiness as due to narcolepsy or sleep apnoea. Real difficulty lies in cases of apparent narcolepsy presenting before cataplexy and where at best the diagnosis can only be possible or probable, not definite narcolepsy. Follow-up is essential here whatever the sleep laboratory findings. Narcolepsy and sleep apnoea sometimes coexist, particularly in overweight males. Cataplexy is sometimes confused with epilepsy or drop attack,

but a careful history will usually separate these.

Hypersomnia without cataplexy or any feature of psychological or physical illness is common. Here exact diagnosis of the cause is often impossible. The idea of idiopathic hypersomnia is based on the concept of abnormal pressure for non-rapid eye movement sleep, with prolonged dream-free deep sleep by night and day, sometimes with a familial or genetic basis. In reality there is little or no distinction in sleep–wake behaviour in different forms of hypersomnia and prolonged follow-up sleep and laboratory studies are needed.

Other causes of sleepiness with medical or psychological illness are unlikely to be confused with narcolepsy. In addition to head injury, hypnotic drug and alcohol abuse, and sleep-related respiratory illness they include:

- depression (insomnia is more common than hypersomnia)
- postviral illness (often Epstein–Barr virus)
- cerebrovascular disease (sometimes with bilateral thalamic infarcts)
- multisystem atrophy
- shift work and circadian delay syndromes
- sleep apnoea treated with continuous positive airway pressure (this rarely completely reverses sleepiness).

Treatment

Treatment of narcolepsy is a problem for both physician and patient. The prescriber may refuse stimulant drugs owing to fear of abuse, or restrict dosage to prevent tolerance. Patients may demand large doses and complete freedom to overdose, not recognizing their own irritability and euphoria.

Most subjects with narcolepsy need a central stimulant drug to improve alertness, and two-thirds need an additional anticataplectic drug to prevent atonia. One drug from each group in Table 4 should be chosen. Dexamphetamine and methylphenidate but not modafinil have a partial anticataplectic as well as an alerting effect. Stimulant drug treatment should be supported by a 15 min nap once or twice a day. Adequate treatment is essential to restore school performance, work, driving ability, and quality of life. This is best achieved with an as-needed, variable dose rather than a fixed dose regime, dependent on factors such as day of the week, activity, and response level. A sleep–wake diary is an important aid to starting and monitoring treatment. Drug response is immediate, while sudden withdrawal may be followed by a severe rebound of sleepiness, cataplexy, or both lasting several days. Stimulant response is the same in hypersomnia as in narcolepsy.

Metabolic tolerance with the need for an increase in dose develops in one-tenth to one-third of subjects. Dose revision, changing to an alternative drug, or a 2-week drug holiday may be necessary. Psychological addiction does not occur and there is no evidence of stimulant abuse in narcoleptics of normal personality. Very occasionally a recreational drug user will feign a history of narcolepsy to obtain stimulants and a urinary drug screen may be appropriate.

Serious dose-related or idiosyncratic side-effects are uncommon. However, sweating and irritability with the stronger stimulants, mild headache with modafinil, and sexual side-effects, increased appetite, and weight gain with clomipramine sometimes limit treatment. Acute amphetamine psychosis is not a problem in narcoleptics and a lifetime of treatment does not cause vascular toxicity or hypertension. A poor drug response should lead to re-evaluation of diagnosis and treatment compliance. Management problems include pregnancy, with the need for the safety of mother and baby to be balanced against potential drug teratogenicity and secretion in breast milk, and cardiovascular disease where low- rather than high-dose treatment is indicated. Conventional treatment is unsatisfactory in about one-fifth of narcoleptics. If disability is severe, a therapeutic trial of morning venlafaxine, 37.5 mg with slow increase to 275 mg per 24 h should be considered but this regime must be carefully monitored.

Further reading

Diagnostic Classification Steering Committee (Thorpy MJ, chairman) (1990). *International classification of sleep disorders: diagnostic and coding manual*. American Sleep Disorders Association, Rochester, MN.

Hublin C *et al*. (1994) The Ullanlinna narcolepsy scale: validation of a measure of symptoms in the narcoleptic syndrome. *Journal of Sleep Research* **3** 52–9.

Johns MW (1991). A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep* **14**, 540–5.

Ling L *et al*. (1999). The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* **98**, 365–75.

Nevsimalova *et al.* (2000). Clinical features of hypocretin (orexin) mutation in human narcolepsy. *Neurology* **54**, A30–A31, Suppl. 3.

Parkes JD *et al*. (1998). The clinical diagnosis of the narcoleptic syndrome. *Journal of Sleep Research* **7**, 41–52.

Peyron *et al.* (2000). A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nature Medicine* **6**, 991–7.

Thannickal *et al.* (2000). Reduced number of hypocretin neurons in human narcolepsy. *Neuron* **27**, 469–74.

24.13.5

Syncope

L. D. Blumhardt

Definition

Syncope (fainting) is a brief loss of consciousness that results from an acute reduction in cerebral blood flow (from the Greek 'synkoptein' to cut or break). It is the most common cause of recurrent episodes of disturbed consciousness.

Pathophysiology

Although there is a seemingly endless list of causes and predisposing factors (Table 1), the sequence of reflex events, once triggered, is relatively constant. Loss of consciousness (probably ultimately due to hindbrain ischaemia) results from a sudden reduction of cerebral perfusion as compensatory mechanisms fail, due either to a reflex reduction of venous return to the heart, or to an inadequate response of the heart when an increased cardiac output is required.

In the common faint (vasovagal syncope), the reduced venous return is caused mainly by a sudden reflex reduction in the resistance of the peripheral blood vessels with pooling, particularly in the lower limbs and abdomen. There is generally some associated, but less important, slowing of heart rate (hence the term vasovagal, originally coined by Sir Thomas Lewis), but in some individuals, an exaggerated reflex bradycardia or even sinus arrest due to vagal hyperactivity may be the dominant factor in reducing cardiac output (cardio-inhibitory syncope). This is a particular feature of loss of consciousness arising from painful pressure on the eyeball (oculocardiac reflex). In cough syncope, the high intrathoracic pressure generated by violent coughing (which may exceed 250 mmHg), triggers a reflex fall in cardiac output. Similar events underlie syncope associated with the Valsalva manouvre and possibly also contribute to the rare defaecation syncope. The proposed mechanisms responsible for carotid sinus syncope include hypersensitivity of the baroreceptors in a diseased carotid sinus and, possibly, contributions from an exaggerated vagal response or oversensitivity of the sinus node. In syncope associated with eating, a vagally induced bradycardia may be provoked by stretching of the oesophagus (swallow syncope) or by very cold stimuli (ice-cream syncope). In other subjects, critical cardiac slowing can be provoked by rectal stretch (for example during proctoscopy or prostatic massage). In micturition syncope, the mechanism is thought to be a combination of high nocturnal vagal tone, postural hypotension, and the sudden loss of the vasopressor effect of a full bladder.

Clinical features

In many subjects syncope is preceded by a characteristic sequence of premonitory symptoms (presyncope). A typical history may include light-headedness, sweating ('clamminess'), sensations of warmth or cold, nausea, and blurring of vision. Subjects may also report tinnitus, receding sounds, 'closing in of peripheral vision', weakness, increased salivation, urgency of micturition, vomiting, diarrhoea, and the need to get fresh air. These symptoms may be proffered in any combination and any, or all, may be missing. It is important to distinguish the usual complaints of dizziness, 'light-headedness', 'muzziness', or giddiness due to hypotension, from vertigo (dizziness with a rotational element) which does not occur in syncope. An eyewitness may report on the sweating, restlessness, excessive yawning, slow sighing respiration, and marked facial pallor. In some individuals the loss of consciousness may be abrupt with little or no warning, whereas in others the symptoms of presyncope may build up slowly. With sufficient warning and insight, some subjects may be able to prevent the syncope by lying down and increasing cardiac return by elevating their legs, but many seem either unaware of this possibility, or are unable to invoke it in time.

If no preventative action is taken, syncope, characterized by a sudden loss of consciousness and collapse with loss of muscular tone, may then follow. The subject is noted to be limp or floppy (no stiffness, rigidity, or tongue biting) and is usually motionless. In some subjects there may be flickering of the eyelids and perhaps an occasional irregular myoclonic twitch or jerk in a limb. Respiration is shallow, the blood pressure low or unrecordable and the pulse thready, rapid, or slow, and often difficult to feel. In uncomplicated syncope, the loss of consciousness usually lasts only seconds and recovery is rapid without confusion. Sweating and profound 'waxy' pallor due to intense vasconstriction of skin vessels may persist well after recovery. If the subject gets up too rapidly, a further episode of syncope may occur.

Most syncope is uncomplicated. Injuries are uncommon, but can occur depending on the circumstances of the faint. If the bladder is full there may be incontinence. If the anoxia is profound, the patient may vomit or be doubly incontinent. If a recumbent posture does not result from the fall, a secondary anoxic seizure may follow (convulsive syncope) in some predisposed individuals. This event is commonly mistaken for epilepsy if the sequence of events is not carefully established from a witness. To complicate matters further, syncope may rarely precipitate a true epileptic seizure.

Main variants of syncope

Vasovagal syncope

The common faint usually occurs for the first time in childhood or adolescence and recurs in well-recognized situations, for example venepuncture, dental procedures, at the sight of blood or injury, sudden emotions, acute pain, postural change, and prolonged standing in stuffy or warm surroundings (school assembly or church). There is almost invariably a postural element—faints occur when standing or sitting and only very rarely when lying (for example in pregnancy). Many secondary factors may increase the risk of faints in the susceptible subject including anaemia, blood loss, convalescence, hypoglycaemia, sleep deprivation, hypotension, cardiac or vascular disease, and drugs. Although there is often a recrudescence of faints later in life, caution should be exercised in diagnosing vasovagal syncope occurring *de novo* in the elderly (check for an earlier history of faints in appropriate circumstances), as the blackouts may be due to cardiac arrhythmias.

Susceptibility to fainting varies widely and some individuals may experience syncope only in association with particular triggers. The diagnosis is usually easy when intense pain such as abdominal colic, glossopharyngeal neuralgia, migraine, or diagnostic manipulation such as venepuncture, oesophagoscopy, or rectal examination provoke attacks. It may be more difficult when syncope complicates a severe vestibular vertigo.

Micturition syncope

The occurrence of loss of consciousness at night during or usually shortly after micturition is highly characteristic of this condition. Contrary to a surprisingly widespread misconception, it is not a complication of prostatism, but occurs almost exclusively in healthy men (some of whom also suffer from vasovagal syncope) with a peak incidence in the third and fourth decades. It does occur in women, but very rarely.

The condition usually responds to advice to mobilize slowly when arising at night and to micturate in the sitting down position.

Cough syncope

This condition is usually associated with chronic obstructive airways disease and smoking. A series of coughs, or sometimes even a single forceful cough, is followed by a collapse with brief loss of consciousness. Afflicted patients often appear unaware of the association between their coughing bouts and the syncope and an account from a witness is required. There may be muscular jerks or twitches during cough syncope and the differentiation from epilepsy is important. Careful clinical assessment is required as cerebellar ectopia with compression of the brainstem and atrioventricular conduction abnormalities may rarely present with similar symptoms. Treatment is usually directed towards the chest condition and education of the patient and relatives into the mechanisms and possible avoidance measures.

Carotid sinus syncope

This is a rare but important cause of syncope as it is potentially treatable. The diagnosis should be considered in elderly patients, usually men with atherosclerotic vascular disease, hypertension, or diabetes when blackouts are associated with a tight collar, head turning, or even the posture or pressure on the neck when shaving. It may also occur with infiltrating cervical tumours or radiation. If suspected, gentle sequential unilateral carotid sinus massage carried out with electrocardiographic control for 6 s per side may result in conduction block or cardiac arrest. Many would accept an asystole of more than 3 s as diagnostic, but the criteria remain controversial. Denervation of the carotid sinus or cardiac pacing is usually effective.

Cardiac syncope

This can be divided into syncope caused by outflow obstruction from the left ventricle and syncope due to disorders of cardiac rhythm. An inability to increase the cardiac output as required, for example during exercise or sexual activity, may cause syncope in hypertrophic cardiomyopathy, aortic stenosis, restrictive pericarditis, left ventricular insufficiency, or atrial myxoma. However, it is more common, even where structural heart disease exists, for loss of consciousness to be caused by an arrhythmia. Syncope may be sudden with no warning (as with with heart block or sinus arrest for example) or there may be dizziness, palpitations, dyspnoea, or chest pain (for example with tachyarrhythmias). The symptoms may closely mimic those of complex partial seizures (temporal lobe epilepsy) which can themselves generate cardiac arrhythmias. After transient cardiac arrest there may be facial flushing as the cardiac output is restored. The clinical presentation of a patient with a slow pulse and syncopal attacks has long been recognized (Stokes–Adams syndrome) and may be due to atrioventricular block, ventricular tachycardia, fibrillation, or standstill. Similar attacks occur with sinus node disease ('sick sinus' or tachycardia–bradycardia syndrome) and the long QT syndrome (see Section 15).

Syncope associated with exercise may also be associated with aortic arch disease, congenital heart disease, and pulmonary hypertension.

Reflex ('vagal') anoxic seizures

This particular form of reflex cardiac arrhythmia can be regarded as one end of the spectrum of fainting disorders. It is important because it is frequently misdiagnosed as epilepsy. In its most common form, a child (the condition may persist into early adult life) has faints that typically are triggered by minor trauma such as a painful knock in the playground. The brief loss of consciousness may be associated with rigidity, pallor, muscular twitching, and sometimes incontinence. There is usually a rapid recovery with little if any confusion. Simultaneous electrocardiographic and electroencephalographic recordings demonstrate that the primary cardiac asystole is followed by secondary anoxic changes (high-amplitude slow waves) on the electroencephalograph with no evidence of epilepsy. The attacks can be provoked in the laboratory by ocular pressure. Treatment if necessary is with long-acting atropine preparations.

Postural hypotension

Convalescent patients may be subject to large reductions of blood pressure on changing their posture. Many drugs, including diuretics, antihypertensives, levodopa, nitrates, major tranquillizers, antidepressants, alcohol, and calcium antagonists may play an important or primary role. Areflexic or 'paralytic' postural hypotension may complicate or be the presenting symptom of diseases affecting the autonomic nervous system, such as diabetes mellitus, idiopathic orthostatic hypotension, extrapyramidal diseases, tabes dorsalis, peripheral neuropathies, and high spinal cord disease. Some otherwise healthy individuals are peculiarly sensitive to mild postural hypotension. A lordotic posture may contribute to syncope in subjects standing to attention on a parade ground.

Diagnostic approach

The essential clues to the diagnosis of syncope are usually found in the history of the immediate circumstances of the collapse and the events leading up to it. A past or family history of similar events, perhaps in more obvious circumstances, will often provide the diagnosis. Factors favouring syncope may be the recent initiation of drugs or a prolonged period of erect posture or postural change. Was there a Valsalva factor present (for example, straining or lifting, or the forceful playing of a wind instrument) or an emotional or painful experience? Are there background factors favouring syncope, such as anaemia, recent convalescence, fatigue, cardiac disease, or blood loss? The history of the attacks themselves needs to be built up in an as much detail as possible including an eyewitness account of events immediately before, during, and after the attack. Was the subject's muscle tone appropriately limp? There should be no tongue biting, rigidity, or rhythmic tonic–clonic movements suggesting seizure activity. If a seizure appears to have occurred in a situation more appropriate to a syncopal episode, a careful history will often establish that the patient's position after the faint did not allow rapid restoration of cerebral blood flow (for example they were propped against a wall or held upright). A secondary anoxic seizure (convulsive syncope) is then the correct diagnosis, rather than epilepsy. Injuries (apart from a bitten tongue) do not particularly favour epilepsy, but confusion is unlikely after syncope, unless there was a complicating head injury or seizure.

Young men who collapse in the bathroom, or on the way back to bed during the night, are at risk of misdiagnosis, particularly if a secondary convulsive syncope has occurred. These circumstances should alert the clinician to the possibility of micturition syncope. The patient with chest disease and blackouts is often peculiarly 'amnesic' for his attacks which he does not associate with his respiratory condition. As for all forms of syncope the eyewitness account is critical to establish the associations, the correct sequence of events, and the diagnosis. The presence of other medical conditions such as chronic airways disease (cough syncope), cardiac disease (arrhythmias), or atherosclerosis (carotid sinus syncope) may provide the main diagnostic clue.

Management

Simple faints require usually require only reassurance, counselling, and education for patients and relatives. They should be instructed in the mechanisms and the avoidance of predisposing situations and trigger factors as well as measures to be taken during presyncopal episodes. Apart from a blood count if anaemia is suspected, or a blood sugar estimation if hypoglycaemia is a possibility, investigations are seldom indicated unless relevant abnormalities are present on history or examination. Investigating all patients with syncope is expensive and produces a low yield. Syncopal episodes occurring *de novo* in adults without obvious triggering factors, or precipitated by exercise, may require electrocardiography, Holter monitoring, treadmill tests, and perhaps sophisticated intracardiac conduction studies. Structural heart disease or failure should be excluded.

Where the diagnosis remains in doubt, or if epilepsy is a possibility, an electroencephalograph and perhaps simultaneous electrocardiographic/electroencephalographic recordings with video monitoring of attacks in a specialist unit may be indicated.

Further reading

deBono DP, Warlow CP, Hyman NM (1982). Cardiac rhythm abnormalities in patients presenting with transient non-focal neurological syndromes. *British Medical Journal* **284**, 1437–9.

Eberhart C, Morgan JW (1960). Micturition syncope. *Journal of the American Medical Association* **174**, 2076–7.

Jaeger FJ, Maloney JD, Fouard-Tarazi (1990). Newer aspects in the diagnosis and management of syncope. In: Rappaport E, ed. *Cardiology update*. Elsevier, New York.

Kapoor WN *et al.* (1982). Syncope of unknown origin. The need for a more effective approach to its diagnostic evaluation. *Journal of the American Medical Association* **247**, 26 787–91.

Kapoor WN, Peterson J, Karpf M (1986). Defecation syncope. *Archives of Internal Medicine* **146**, 2377–9.

Leatham A (1982). Carotid sinus syncope. *British Heart Journal* **47**, 409–10.

Levin B, Posner JB (1982). Swallow syncope. Report of a case and review of the literature. *Neurology* **22**, 1086–93.

Lewis T (1932). Vaso-vagal syncope and the carotid sinus mechanism. *British Medical Journal* **1**, 873–6.

Lipsitz LA (1983). Syncope in the elderly, *Annals of Internal Medicine* **99**, 92–105.

Proudfit WL, Forteza MS (1959). Micturition syncope. *New England Journal of Medicine* **260**, 228–31.

Sharpey-Schafer EP (1956). The mechanism of syncope after coughing. *British Medical Journal* **ii**, 860–3.

Sharpey-Schafer EP (1956). Syncope. *British Medical Journal* **1**, 506–9.

Stephenson JPB (1978). Reflex anoxic seizures ('white breath holding'): non-epileptic vagal attacks. *Archives of Disease in Childhood* **43**, 193–200.

Sugrue DD, Wood DL, McGoon MD (1984). Carotid sinus hypersensitivity and syncope. *Mayo Clinic Proceedings* **59**, 637–40.

24.13.5.1 Head-up tilt-table testing in the diagnosis of vasovagal syncope and related disorders

Steve W. Parry and Rose Anne Kenny

In many patients, the diagnosis of vasovagal syncope or presyncope is evident from their history and physical examination. Where doubt exists, or a definitive diagnosis is needed to clarify atypical presentations or aid decision-making for driving or occupational purposes, head-up tilt-table testing should be considered.

The head-up tilt-table test employs prolonged, controlled orthostatic stress to provoke vasovagal syncope, with characteristic symptoms and concurrent haemodynamic changes confirming the diagnosis. In susceptible patients, head-up tilt causes relative central hypotension through displacement of venous blood to the lower limbs and capacitance vessels, with consequent vigorous left ventricular contraction and inappropriate cardiac mechanoreceptor activation. The resultant afferent neural traffic traverses unmyelinated C-fibres to the nucleus tractus solitarius, which then orchestrates the vagal activation and vasodilatation characteristic of the vasovagal response. In the absence of a 'gold-standard' diagnostic test, estimates of the sensitivity and specificity of head-up tilt-table testing from comparisons with healthy volunteers vary between 32 to 85 per cent and 60 to 90 per cent, respectively. The reproducibility of positive responses in patients with vasovagal syncope is up to 87 per cent in the short term (30 min) and 85 per cent over several days and months. The test's reliability, non-invasive nature, relative cheapness, and safety record make head-up tilt-table testing the current diagnostic test of choice in the investigation of unexplained syncope.

Haemodynamic responses during a positive test may also guide therapeutic strategies, and are divided into vasodepressor (predominant blood pressure fall in the absence of significant bradycardia/asystole), cardioinhibitory (predominant bradycardia/asystole), or mixed (a combination of the two) subtypes. Cardioinhibitory vasovagal syncope may benefit from permanent pacemaker therapy, whereas the predominant vasodepressor subtypes are unlikely to do so. Heart rate responses during tilt may also assist in the choice of medical therapy; for example beta-blockers are useful if pronounced tachycardia antedates symptomatic hypotension/bradycardia.

Head-up tilt-table testing is a relatively benign investigation, with few reported complications worldwide, most of which have anecdotally been related to inappropriate staff training and lack of resuscitation facilities. Relative contraindications include severe coronary or cerebrovascular arterial stenoses, clinically severe left ventricular outflow obstruction, and critical mitral stenosis.

Methodology

Tilt-table testing should be conducted in a quiet, dimly lit environment at a comfortable temperature to minimize confounding autonomic nervous activation. The table should have a foot-plate support, and be capable of smooth, rapid movement between supine and calibrated upright positions between 60° and 80° (lesser or greater angles increase the number of false-negative and false-positive studies, respectively). Subjects should be fasted for no more than 2 h before testing to avoid potential confounding from postprandial hypotension. Intravascular instrumentation is avoided in all but isoproterenol-provoked tilt testing as this markedly lowers the test's specificity. After 10 min rest in the supine position, subjects are strapped securely to the tilt table to prevent injury during syncope and collapse, and then tilted to 70° for 40 min.

A summary of the Newcastle protocols for head-up tilt-table testing is provided in Table 1. Continuous blood pressure monitoring is advised during the test, preferably with non-invasive beat-to-beat digital photoplethysmographic devices such as Finapres (Ohmeda, Wisconsin) or Portapres (TNO-TNM Biomedical, Amsterdam), though sphygmomanometric devices are widely used. Electrocardiography should be undertaken at baseline, continuously during symptoms or haemodynamic changes, and every 5 min otherwise. The test should be supervised by a nurse, physician, or technician trained in the management of the test and its potential complications, with advanced resuscitation equipment and a clinician trained in its use immediately available at all times.

The head-up tilt-table test is deemed diagnostic only if arterial hypotension and/or bradycardia are accompanied by symptoms reproducing the patient's syncopal or presyncopal symptoms. Haemodynamic changes without symptom reproduction should not be construed as vasovagal syncope.

Pharmacological and mechanical provocative agents in head-up tilt-table testing

Where the initial prolonged, passive tilt test is negative, several agents may be used to increase the sensitivity of the test.

Isoproterenol provokes the vasovagal response by simulating the catecholaminergic surge seen prior to syncope in susceptible subjects. Isoproterenol should be used with caution in subjects with known arrhythmias, coronary artery disease, significant aortic stenosis, left ventricular outflow obstruction, and uncontrolled hypertension. Tilt-testing should be terminated if sustained tachycardia (>150 beats/minute), hypertension (>180/100), arrhythmia, or intolerable symptoms supervene.

Nitrates, in particular sublingual glyceryl trinitrate, have more recently been used in this context, with nitrate-induced vasodilatation simulating the venous pooling which provokes spontaneous episodes. Nitrate-provoked tilt testing is as specific and sensitive, and better tolerated than, isoproterenol.

Developments incorporating intravenous adenosine and clomipramine are currently being evaluated, but are not yet in routine use.

Mechanical provocation, using a suction chamber to exert a lower-body negative pressure (again simulating venous pooling), has been successfully used in specialist centres.

Non-vasovagal indications for head-up tilt-table testing

Carotid sinus massage in the head-up tilt position (following an initial supine, bilateral, non-diagnostic massage) may increase the diagnostic rate for carotid sinus hypersensitivity by 30 per cent. Head-up tilt testing is used to diagnose orthostatic hypotension in patients unable to stand unaided for 3 min, while the postural orthostatic tachycardia syndrome may similarly be diagnosed during tilt-testing if the patient's heart rate rises by more than 30 beats/min (or to a maximum of 120 beats/min) during the presence of characteristic symptoms in the absence of hypotension. Tilt-table testing may also be useful in the differential diagnosis of apparently epileptiform events. Prolonged asystole during vasovagal syncope may result in myoclonic jerking, tonic–clonic movements, and (rarely) incontinence (urinary and faecal) which can be mistaken for generalized convulsions. The short duration of the event, rapid recovery, and the absence of postictal confusion and neurological signs should prompt tilt testing as part of the diagnostic work-up. Psychogenic and hyperventilation syncope result in symptom reproduction without haemodynamic changes during tilt, with hypocapnia being demonstrated in the latter. Head-up tilt testing may also be useful in demonstrating neurocardiovascular disorders as attributable causes of unexplained falls in older patients (in whom falls and syncope frequently overlap), with amnesia for loss of consciousness prompting a non-syncopal presentation.

Further reading

Benditt DG, *et al.* (1996). Tilt table testing for assessing syncope. ACC Expert Consensus Document. *Journal of the American College of Cardiology* **28**, 263–75.

Kenny RA, *et al.* (1986). Head-up tilt: a useful test for investigating unexplained syncope. *Lancet* **1**, 1352–4.

Kenny RA, O'Shea D, Parry SW (2000). The Newcastle protocols for head-up tilt table testing in the diagnosis of vasovagal syncope and related disorders. *Heart* 83, 564–9.

24.13.6 Brain death and the vegetative state

B. Jennett

Cardiopulmonary resuscitation and intensive care are now commonplace in the developed world, so that life-threatening brain insults may now be followed by complete recovery. Sometimes, however, these interventions do no more than extend the process of dying for hours or days because the patient is brain dead. In others intervention comes too late after cardiorespiratory arrest to save the cerebral cortex and thalamus which are more vulnerable to hypoxia than is the brainstem. In that case, or when head injury has irretrievably damaged the cortical connections, the patient can survive for a long period in a vegetative state.

Brain death

Much of the controversy about brain death arose because it was not appreciated that death is a process rather than an event, with organs failing in various sequences. Most often the heart stops first, followed within minutes by respiratory arrest due to brainstem hypoxia. After primary respiratory arrest there is rapid hypoxic brain failure but the heart may continue to beat for 15 to 20 min. If the brainstem fails first due to an intracranial catastrophe and a ventilator takes over respiration the heart can beat for days (occasionally weeks)—this is the state of brain death.

The crucial lesion is irreversible loss of brainstem function with subsequent lack of downward drive to maintain respiration and of upward activation of the cerebral cortex by the ascending reticular pathways. When systemic hypoxia has not been the initial insult the cerebral cortex may be structurally intact, and islands of electrical activity may be detected on the electroencephalogram. Early definitions of brain death (for example the Harvard Criteria of 1968) implied that the whole nervous system was dead with a flat electroencephalogram and an absence of all motor activity. But the spinal cord is more resistant to hypoxia even than the brainstem, and is unaffected by intracranial catastrophes—so spinal reflexes often persist after brain death. To resolve this confusion the term brainstem death is now preferred in the United Kingdom. Although in the United States the most commonly used diagnostic criteria are those for brainstem death, most guidelines and statutes still refer to whole brain death. About half the cases of brain death result from head injury, after hours or days of intensive treatment following initial resuscitation. About a third have suffered severe non-traumatic intracranial haemorrhage, and the rest a variety of catastrophic intracranial events, including systemic hypoxia associated with cardiac arrest.

Criteria for diagnosis

Undue emphasis on the final confirmation that no residual brainstem function persists has sometimes distracted attention from the stepwise process of diagnosis, for which these tests are only the last stage. Indeed the most important step is the first one, satisfying the preconditions. These require that the patient be apnoeic and in deep coma due to irreversible structural damage to the brain. This implies that reversible causes of brainstem depression have been adequately excluded. It is usually obvious that structural brain damage has occurred—there has been a recent head injury or a classical history of spontaneous intracranial haemorrhage or of some less acute intracranial condition. Establishing the irreversibility of such brain damage depends on failure to improve after the correction of factors such as systemic hypotension and hypoxia and raised intracranial pressure. Other causes of temporary failure of brainstem function are depressant drugs (including alcohol), neuromuscular relaxant drugs, and physiological factors such as hypothermia and gross metabolic imbalance. The first two of these may complicate cases of structural brain damage but it is only in a minority of cases that serious doubts arise about confusing factors. For example, when a patient is found unconscious and no satisfactory history can be discovered it may be necessary to undertake formal screening for drugs. In all cases the diagnosis of brain death should be delayed until sufficient time has passed for the exclusion of all temporary causes of brainstem depression, usually at least 6 h.

The tests applied to indicate lack of brainstem function are simple to carry out and to interpret. There should be no response of the pupils to light, of the eyelids to corneal touching, of the facial muscles to pain, of the throat muscles to movements of the endoctracheal tube, or of the eyes to syringing each external auditory meatus with ice cold water (the caloric or vestibulo-ocular reflex). Only when there has been a negative response to all of these is the final crucial test applied—to verify that there is still apnoea. There must be no respiratory movement after disconnection from the ventilator for long enough to allow the $P\mathrm{a}\mathrm{CO}_2$ to rise to 6.65 kPa (6.8 kPa in the United States), oxygenation being maintained by delivering 6 litre/min of oxygen down the endotracheal tube. To exclude any possibility of observer error it is required that all these tests are carried out by two doctors and on two occasions.

Action after diagnosis of brain death

There is now wide acceptance of the concept that when the brain is dead the person is dead. In the United Kingdom the legal time of death is when the first tests confirm brainstem death, and not some later time when the heart stops. Some doctors are still reluctant to make this diagnosis explicitly and then to act logically and legally by disconnecting the ventilator. The useless ventilation of a brain dead patient deprives that patient of death with dignity, needlessly prolongs the distress of relatives, wastes resources for intensive care, and is bad for the morale of nursing staff. Moreover the opportunity to offer organs for transplantation is lost because gradual circulatory failure makes such organs useless for donation.

The persistent vegetative state

This term was introduced in 1972 by Jennett and Plum to describe the clinical condition resulting from loss of function in the cerebral cortex with a functioning brainstem. Because of the latter, vegetative patients breathe spontaneously and are not ventilator-dependent; another difference from brain death is that they can survive for many years if adequately fed and nursed. The commonest cause of vegetative survival after acute brain damage is severe head injury, the mechanism being severe diffuse axonal injury severing the subcortical connections over a wide area. Secondary hypoxic brain damage is a contributing factor in some traumatic cases. Most non-traumatic cases result from severe hypoxia/ischaemia of the brain following cardiac arrest, near drowning, or strangulation, whilst a few result from severe hypoglycaemia in diabetics. Other causes are acute intracranial haemorrhage or infection. In adults the vegetative state can evolve gradually during the late stages of chronic dementing conditions, and in children can result from severe congenital malformations of the brain or from progressive metabolic or chromosomal diseases affecting the brain.

At autopsy after acute hypoxic insults there is commonly a widespread loss of cortical neurones. After acute traumatic and non-traumatic damage leading to vegetative survival there is almost always severe bilateral thalamic damage, whilst the cortex may be relatively spared. There is also progressive degeneration over many months of neurones and nerve fibres and their myelin sheaths remote from the site of initial damage, which is reflected during life in progressive enlargement of the ventricles as visualized by computed tomography or magnetic resonance imaging. Findings on the electroencephalogram are variable, but there is often loss of evoked cortical responses to somatic stimuli. Positron emission tomography shows severe depression of glucose metabolism in cortical grey matter, to levels found only in experimental deep barbiturate narcosis.

Diagnosis

In practice the diagnosis depends on characteristic clinical features recorded by skilled observers over a period of time. The patient has long periods of spontaneous eye opening (hence the inappropriateness of calling this condition irreversible or prolonged coma). The eyes may briefly follow a moving object or the head turn reflexly to a sudden noise that may also produce a startle reaction. All four limbs are paralysed and usually spastic, with only reflex posturing and withdrawal from a painful stimulus, and often there is a grasp reflex. The face may grimace and groans may be heard but never words. There is no psychologically meaningful response to external stimuli or any learned behaviour—no evidence of a working mind. There may be emotional behaviours such as smiling, crying, or laughing but these are

not related to appropriate external stimuli. It is concluded that although awake these patients are not aware and do not suffer distress or pain. Misdiagnosis by non-experts is not uncommon, and care is needed to exclude the minimally conscious state in which there are very limited responses indicating some cognitive activity. It must also be ascertained that the patient does not have the locked-in syndrome, due to brainstem damage resulting in full awareness but widespread paralysis, leaving the patient able to communicate only by a yes/no code using the sole remaining motor power, blinking the eyelids or moving the eyes.

Prognosis

Patients in a vegetative state for some time can still make some recovery, and persistent does not mean permanent. Of patients in the vegetative state 1 month after an acute insult, about half of the head injured will regain some consciousness, but only a few of the non-traumatic cases do. Most who recover consciousness remain very severely disabled and dependent, particularly if they have been vegetative for several months. After head injury permanence cannot be declared until 12 months, but after non-traumatic insults after 6 months according to United Kingdom criteria and 3 months in the United States. There is a high mortality in the first year after becoming vegetative but once this period is survived patients can live for many years, if tube feeding and good nursing care is maintained and infective complications actively treated.

Action after permanence declared

There is now a consensus in many countries that survival for years in a permanent vegetative state is of no benefit to the patient, and that it is therefore appropriate to withdraw life-sustaining treatment once permanence is declared. Many courts in the United States and the United Kingdom have agreed that artificial nutrition and hydration is medical treatment that can be withdrawn if judged to be no longer of benefit to the patient. Once this is done a peaceful death occurs in 8 to 12 days, and the cause of death is regarded as the original brain damage. Only in the United Kingdom is it a legal requirement to seek court approval before withdrawing such treatment, although this situation is under review.

Further reading

Brain death

Conference of the Medical Royal Colleges and their Faculties in the United Kingdom (1976). Diagnosis of brain death. *British Medical Journal* **2**, 1187–8.

Conference of the Medical Royal Colleges and their Faculties in the United Kingdom (1979). Diagnosis of death. *British Medical Journal* **1**, 322. [Original descriptions of United Kingdom criteria for brainstem death.]

Health Departments of Great Britain and Northern Ireland (1998). *A code of practice for the diagnosis of brain stem death*. Department of Health, London. [Most recent United Kingdom update.]

Quality Standards Sub-committee of the American Academy of Neurology (1995). Practice parameters for determining brain death in adults. *Neurology* **45**, 1012–14. [Widely accepted United States criteria.]

Youngner SJ, Arnold RM, Shapiro R, eds (1999). *The definition of death*. Johns Hopkins University Press, Baltimore. [Review of controversies, clinical, ethical, legal and social—primarily from an American viewpoint.]

Vegetative state

Adams JH, Graham DI, Jennett B (2000). The neuropathology of the vegetative state after an acute brain insult. *Brain* **123**, 1327–38. [Detailed pathology of 35 traumatic and 14 non-traumatic cases.]

Jennett B (2002). *The vegetative state: medical facts, ethical and legal dilemmas*. Cambridge University Press, Cambridge.[Review of medical facts, ethical issues and details of legal cases in several countries.]

Multi-Society Task Force on PVS (1994). Medical aspects of the persistent vegetative state. *New England Journal of Medicine* **330**, 1499–507, 1572–9. [Review of world literature and prognostic data from an American perspective.]

Quality Standard Sub-Committee of the American Academy of Neurology (1995). Practice parameters: assessment and management of patients in PVS. *Neurology* **45**, 1015–18. [Most recent American criteria.]

Wade DT and Johnston C (1999). The permanent vegetative state: practical guidance on diagnosis and management. *British Medical Journal* **319**, 841–4 (see also Editorial by B Jennett on pp 796–7). [Recent United Kingdom review.]

24.13.7 Stroke: cerebrovascular disease

J. van Gijn

Introduction

Cerebrovascular diseases include many pathological conditions: two main categories being infarction, (through occlusion of major arteries, small arteries, or venous sinuses) and haemorrhage (through rupture of small arteries, arterial aneurysms, or capillaries). Intracerebral haemorrhage was first recorded by the Swiss physician Wepfer, and in more detail by Morgagni in Padua. Non-haemorrhagic stroke, 'serious apoplexy', puzzled the medical community until cerebral softening ('ramollissement') was recognized as a pathological entity in 1820 by Rostan in Paris. Initially it was regarded as an inflammatory condition. The relationship of brain softening with arterial occlusion and atherosclerosis gradually dawned on pathologists in the nineteenth century when it was firmly established by Rokitansky and Virchow. Subarachnoid haemorrhages and their usual source, intracranial aneurysms, were first recognized at the beginning of the nineteenth century; the diagnosis could (sometimes) be made during life from the beginning of the twentieth century. In 1931 the Edinburgh neurosurgeon Norman Dott carried out the first intracranial operation for a ruptured aneurysm, by wrapping it in muscle.

Knowledge of cerebrovascular disease has received great impetus from the advent of computed tomography (**CT**) scanning. Previously, observations depended on postmortem studies and on indirect neuroradiological studies such as angiography and pneumoencephalography. CT scanning has allowed the rapid and reliable distinction between haemorrhagic and ischaemic stroke during life. Subsequently, CT and the newer technique of magnetic resonance imaging (**MRI**) have identified several subtypes of stroke, each requiring specific therapeutic measures. Examples are lacunar infarction, intracranial venous thrombosis, arterial dissection, and complications after aneurysm rupture (rebleeding, ischaemia, or hydrocephalus). The rapid increase in diagnostic accuracy coincided with dissemination of the randomized clinical trial and other methodological innovations in medicine. As a result, stroke research is no longer a backwater of medicine but a bustling area.

The consequences of stroke are often devastating. A sudden loss of a large amount of brain tissue affects much more than specific, localizable functions such as movement, sensation, vision, and language. The greater part of the brain has no defined task, but serves to connect and integrate the separate 'functions'. Mood, initiative, sense of humour, and speed of thought are examples of essential aspects of human life that can be severely affected by stroke, but which are often sadly ignored. There are no 'silent' areas in the brain that can be damaged with impunity.

Epidemiology of stroke

Worldwide, stroke is the third most common cause of death after coronary heart disease and all cancer deaths, and stroke is the most important single cause of disability in the Western world. Although the incidence of stroke is not technically difficult to measure, it does require the expenditure of much time and resources. The few reliable studies, mostly from developed countries, show that age- and sex-standardized annual incidence rates for subjects aged 45 to 84 years are between 300 and 500 per 100 000. The pathological type varies, even between studies with a high rate of CT scanning, but a general estimate is 82 per cent infarction, 15 per cent primary intracerebral haemorrhage, and 3 per cent subarachnoid haemorrhage. The incidence of transient ischaemic attacks (**TIAs**) has been rather consistently estimated at 35 per 100 000 of the entire population, or about 80 per 100 000 of subjects between 45 and 84 years of age.

In terms of an average general practice of 2400 people (1000 patients between 45 and 84 years of age), four patients will have a stroke per annum, whilst one will have a TIA. Intracerebral haemorrhage will occur about twice every 3 years, and subarachnoid haemorrhage once every 8 years.

Arterial occlusive disease

The cerebral circulation and its disorders

Brain tissue is critically dependent upon a constant supply of oxygen and glucose. The cerebral blood flow (800 ml/min) accounts for 15 to 20 per cent of the entire cardiac output, whereas the brain (1350 g) accounts for only 2 per cent of the adult body weight. Neurones in the brain require a constant supply of ATP to maintain concentration gradients of ions across their membranes, necessary for the generation of action potentials. The resting brain consumes energy at the same rate as a 20-Watt light bulb.

Whether occlusion of an artery in the brain or in the neck actually leads to ischaemia depends on collateral pathways. If an end artery is occluded and there is no collateral circulation at all, ischaemic symptoms will occur within seconds. Neurones will start dying within minutes, and within hours the entire supply area of the artery will be irreversibly damaged. In contrast, permanent occlusion of a major artery, such as the internal carotid artery, may be symptomless in the presence of an adequate collateral circulation. Broadly speaking, three levels of collateral circulation can be distinguished (Fig. 1; these can be thought of as three lines of defence):

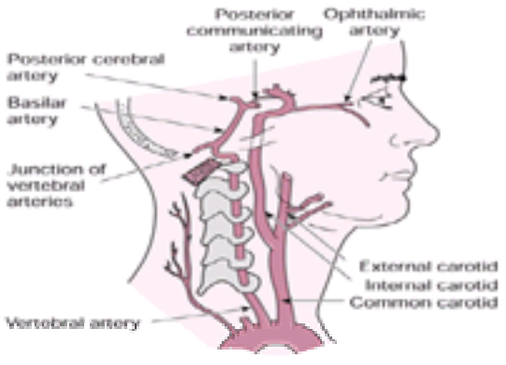

Fig. 1 Arterial supply of the brain. The drawing shows, on the right side, the internal carotid artery, external carotid artery, and vertebral artery. If a main artery is

occluded then collateral flow may occur via the circle of Willis (see also Fig. 2), or through connections between extracranial and intracranial branches.

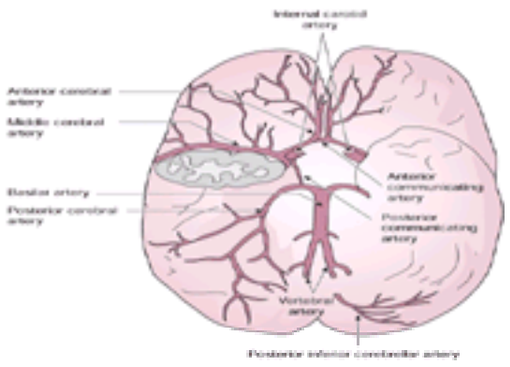

Fig. 2 The arterial circle of Willis, at the base of the brain.

1. *The circle of Willis* (Fig. 2). Even if there is no blood whatsoever flowing to the brain from one or even both internal carotid arteries, collateral flow from the other internal carotid artery or from the basilar artery, via an intact circle of Willis, may ensure an adequate blood supply in the territory of the occluded artery.
2. *Connections between extracranial and intracranial vessels*. If the internal carotid artery is occluded at its origin, collateral channels may develop via the external carotid artery. Branches supplying the outer orbit may connect with branches to the retina, resulting in a reversed flow in the ophthalmic artery. From there, blood reaches the distal part of the internal carotid artery. Similarly, branches of occipital arteries (normally supplying the neck muscles) may fill the basilar artery if this is occluded at its origin.
3. *Leptomeningeal anastomoses*. If, for example, the main stem of the middle cerebral artery is occluded, its terminal branches at the surface of the brain may anastomose with similar branches of the anterior and posterior cerebral arteries; in this way the cerebral cortex in the territory of the occluded artery is spared, partly or wholly, although the deep territory will still be ischaemic.

Atherothrombosis is the major cause of occlusion of the major arteries in the brain or in the neck. However, two important qualifications should be made. First, atherosclerosis of the intracranial arteries is relatively uncommon, at least in White people (other than in Black or Oriental people). This means that brain infarction in the West is usually caused by an embolism, in which a thrombus has been dislodged from an upstream lesion. The source can be the internal carotid artery, the aorta, or the heart. Second, atherosclerosis is not a sufficient cause in itself: not everyone with severe atherosclerotic disease suffers an ischaemic stroke. Other factors are irregularity of the plaque, turbulence of blood, platelet aggregation, and the balance of clotting factors.

Diagnosis of transient ischaemic attacks

Transient ischaemic attacks are important to diagnose because they are potential harbingers of stroke. They precede cerebral infarction probably in only 15 to 20 per cent of cases (it is difficult to be certain of the figure because almost all such information has been retrospectively collected in patients once they had suffered a stroke).

Unfortunately the term 'transient ischaemic attack' is rather imprecise, because it tacitly implies three restrictions. To begin with, it refers only to the brain and not to angina pectoris or intermittent claudication. Also excluded is transient ischaemia of the entire brain, for example in syncope or ventricular fibrillation. It is only ischaemia of a part of the brain that is conventionally covered by the term 'TIA'. Finally, how transient is transient? Traditionally the limit has been set at 24 h. Obviously this threshold has more to do with astronomy than with biology or disease. In fact, most TIAs last minutes, not hours. The longer an attack lasts, the greater the chance that CT scanning afterwards will show a relevant ischaemic lesion. In terms of patient management the essential question is not whether the attack has lasted 3 min, 3 days, or 3 weeks, but what its cause is and how recurrences can be prevented.

What actually happens in the brain during a given period of ischaemia can often only be guessed at. The usual assumption is that an embolus, consisting of platelets or loosely organized thrombus, temporarily blocks an intracerebral vessel and then dissolves into smaller fragments. There is scant evidence for this phenomenon, apart from chance observations during fundoscopy, angiography, or operation. Other explanations, applicable only to a minority of cases, include marginal flow, secondary to severe narrowing or occlusion of arteries.

The diagnosis of a TIA is problematic. That one has to rely on the history alone is a first difficulty (it requires time, skill, and patience), but not a unique one. A greater source of interobserver variation is that the term 'TIA' is an interpretation rather than a description.

Main varieties of transient ischaemic attacks

There are four kinds of symptoms that can safely be regarded as TIAs, given that the onset is sudden (within seconds), that all symptoms appear at the same time, without 'march', and that there is no better explanation.

Transient weakness of one half of the body

Apart from weakness there may also have been numbness. Isolated numbness or pins and needles on one side of the body are a rare manifestation of transient cerebral ischaemia; other causes such as overbreathing are more likely. Weakness and numbness are closely related perceptions, and one should not take these or other expressions ('an arm gone dead') for granted. It is important to make sure the problem had to do with moving the limbs or the face on one side (facial weakness on one side often manifests itself through slurred speech or drooling), and not with what it felt like when those body parts were touched or with spontaneous sensations. It is also critical to verify that the problem occurred in at least two of three body areas, and that, in the elderly, it was not just a leg or arm that had 'gone to sleep' after a mid-day nap. All four limbs may be affected at the same time in TIAs of the brainstem.

Transient loss of the ability to find words or to understand them

The medical term for this type of TIA is 'dysphasia' or 'aphasia'; the problem here is not that patients and relatives may not recognize the episode as representing a problem of language, but describe the attack as 'confusion'. It is helpful to ask specific questions about the patient's ability to put thoughts into words (motor dysphasia), and about having been able to understand what was said (sensory dysphasia). If a patient can write sentences but cannot speak, the cause is almost certainly psychological. A frequent problem is the distinction between dysphasia (disorder of language) and dysarthria (disorder of articulation). To ask whether pronunciation was difficult may not be very helpful. After all, in both cases the patient's thoughts are clear and the difference between finding the right words and forming the right sounds is not great. A useful question is whether the words made sense and whether they were in the right order. Dysphasia relates to the left hemisphere in right-handed people, and in 50 per cent of left-handers.

Transient loss of vision in one eye

The difficulty in this case is to distinguish transient monocular blindness from the loss of vision on one side in both eyes (hemianopia). Either type of attack can be experienced as a problem in one eye. The distinction is not academic, as monocular attacks of blindness should lead to investigation of the internal carotid artery in the neck with a view to angiography and operation, whereas hemianopia mostly (in 80 per cent) reflects a disorder in the posterior circulation, in which case treatment will often be medical. The vital question to ask is whether patients have alternately covered each eye during the attack. A surprisingly large proportion of patients have done so, but they will not always offer this information without prompting. On having covered the 'good eye' in case of hemianopia, the patient should still have been able to see with the 'bad eye', though only the nasal half of the visual field. With a monocular disorder the blindness should have been complete.

Transient loss of vision in one hemifield

Hemianopia reflects dysfunction of the occipital lobe. It is also a common aura in migraine attacks, which may occur without ensuing headache, especially in the elderly. It is therefore important for the physician to enquire about the mode of onset: flashing lights, bright colours, zig-zag lines, and a gradually expanding deficit all argue in favour of a migrainous attack rather than ischaemia in its restricted sense of a stroke warning.

Differential diagnosis of transient ischaemic attacks

Table 1 lists the types of attacks that should not be regarded as TIAs, either because of positive phenomena (such as rhythmic jerking) that are incompatible with the definition of focal ischaemia, or because other causes are much more likely. Especially the tendency to label any episode of 'dizziness' in the elderly as 'vertebrobasilar ischaemia' or, even worse, 'vertebrobasilar insufficiency' should be strongly resisted.

In addition, some specific disorders other than atherosclerosis may cause attacks that are more or less indistinguishable from true TIAs as defined above. They are listed in Table 2. These rare, but important, causes are reason enough for ordering a CT or MRI scan of the brain in patients with cerebral TIAs (not necessarily in those with transient monocular blindness). A chronic subdural haematoma should always be suspected in the elderly, especially if they are taking anticoagulants. Hypoglycaemia should come to mind in diabetic patients. Focal weakness may follow an epileptic seizure (Todd's paralysis) and may be misdiagnosed as TIA if the initial jerking is missed or misinterpreted. Tumours may also cause temporary deficits without focal epilepsy. Transient global amnesia is a disorder of memory probably caused by migrainous vasospasm; although technically ischaemic in nature, it is not associated with an increased risk of stroke or other vascular disease.

Prognostic implications of transient ischaemic attacks

In a population-based study of 184 patients in the United Kingdom after they had one or more TIAs, the actuarial risk of stroke was 12 per cent during the first year after a TIA (13-fold excess risk) and approximately 6 per cent per annum over the subsequent 5 years (sevenfold excess risk). The actuarial risk of death, stroke, or myocardial infarction over the first 5 years after a TIA was between 8 and 9 per cent per annum. Heart disease and stroke each accounted for about one-third of all deaths. Given that most of these patients were treated with aspirin, which reduces the rate of vascular events after cerebral ischaemia by about 15 per cent (see below), the risk of stroke without treatment can be estimated at 14 per cent in the first year and 7 per cent in subsequent years, and the average risk of death, stroke, or myocardial infarction in the first 5 years at 10 per cent per annum. In individual patients, these average risks will be modified by their age, specific risk factors, coexistent disease, and the interval since their first attack.

Investigations in patients with cerebral ischaemia

There is no great difference between searching for the cause of a TIA and searching for the cause of an ischaemic stroke. Very early CT or MRI scanning is mandatory: mainly to exclude intracerebral haemorrhage and the occasional structural lesion mimicking stroke, not so much to demonstrate infarcts. Table 3 lists the major and contributory causes of TIA and ischaemic stroke, with corresponding investigations. In general, first-line investigations are: a full blood count; erythrocyte sedimentation rate (**ESR**); plasma glucose, creatinine, and electrolytes; plasma lipids; treponemal antibodies; urinalysis; ECG; and unenhanced CT scan of the brain.

Diagnosis of cerebral infarction

Distinction from other types of stroke

From a practical point of view, the first step is to distinguish ischaemic stroke from intracerebral haemorrhage. In the past, when a certain distinction could be made only at operation or autopsy, a decreased level of consciousness and headache were considered typical of intracerebral haemorrhage. After CT scanning became available in the 1970s, it was soon clear that smaller haemorrhages were not associated with headache and drowsiness. Given that 3 out of 20 strokes are haemorrhagic, and on the assumption that two-thirds of all haemorrhages lack distinctive clinical features, a diagnosis of cerebral infarction based on clinical features alone is wrong in every tenth case on average. Even complex clinical scoring methods can hardly improve on this error rate.

CT scanning

Acute parenchymal haemorrhage is of a higher density than normal brain tissue on CT scanning (see Fig. 6). The hyperdensity occurs immediately and is caused by the iron molecules in haemoglobin. Signs of infarction are more difficult to detect at an early stage: in the first decade of CT scanning this was not possible until after 3 days, when frank tissue necrosis caused a hypodense lesion on the scan. With improved CT technology, subtle early signs of cerebral infarction have been recognized, at least with a large area of infarction. These features include loss of outline of the insular ribbon and the lentiform nucleus (Fig. 3), loss of normal differentiation between grey and white matter, and effacement of cortical sulci.

Fig. 6 Primary intracerebral haemorrhage in a 52-year-old man. This CT scan shows a hyperdense lesion in the right thalamus; the haemorrhage has ruptured into the ventricular system.

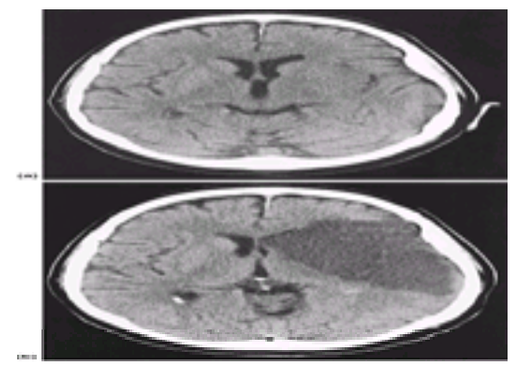

Fig. 3 Acute cerebral infarction in a 78-year-old man. (a) CT scan about 6 h after symptom onset. In the left brain hemisphere (on the reader's right) there are subtle changes in the region of the basal ganglia: other than on the normal side, it is difficult to distinguish the different brain nuclei and their separation by white matter. (b) CT scan 4 days after symptom onset shows marked hypodensity in the entire territory of the left middle cerebral artery.

Within the first few days, the area of infarction changes into a slightly hypodense, ill-defined, and somewhat swollen lesion on CT scanning, to become more clearly demarcated and hypodense towards the end of the first week (Fig. 3). Occasionally there may be massive swelling with brain herniation, or haemorrhagic transformation. During the second week the infarct may again gradually increase in density, because the degradation products of necrotic brain tissue more readily absorb X-rays. In the third and fourth week the infarcted area may even become isodense, thereby almost indistinguishable from normal brain, the so-called 'fogging effect'. Eventually a sharply demarcated, atrophic, hypodense (similar to cerebrospinal fluid) defect remains. It is not always possible to tell with certainty how old an infarct is, nor to distinguish it from the scar of a haemorrhage that occurred weeks or years before. Intravenous injection of X-ray contrast will, in the first week, cause some enhancement of gyri (if the lesion involves the cortex).

The proportion of patients in whom CT scanning shows an appropriate infarct depends not only on the time of scanning and the generation of the scanner, but also on the size of the infarct and on its location. With serial CT scanning, eventually more than 90 per cent of infarcts show up.

Magnetic resonance imaging

MRI is especially useful for demonstrating small infarcts and lesions in the posterior fossa; it is also more sensitive than CT scanning in the early phases of brain ischaemia. Signal changes on *T*2-weighted images occur after 6 to 8 h, and on *T*1-weighted images after 16 h. Infarcts of any size are more often and more quickly visible on fluid-attenuated inversion recovery (**FLAIR**) scans and on diffusion-weighted imaging, but these techniques are not widely available. The distinction from intracerebral haemorrhage is less obvious than on CT, but the paramagnetic effects of deoxyhaemoglobin can be identified after a few hours.

Classification of cerebral infarction

Time was often the guiding principle in the classification of stroke in the era before brain imaging and its positive effects on the accuracy of clinical diagnosis. From the point of view of management and prognosis, however, it is rather irrelevant to distinguish 'progressive stroke' from 'completed stroke', or 'permanent stroke' from 'reversible ischaemic neurological deficit' (**RIND**, a kind of 'extended TIA' with complete recovery within 3 or 6 weeks, depending on local convention). What counts is the eventual severity of the functional deficit and, conversely, what remaining function is at stake.

The anatomical classification distinguishes infarcts according to the territory of major cerebral arteries: in the cerebral hemispheres infarcts can be located in the supply areas of the anterior cerebral artery, middle cerebral artery, posterior cerebral artery, or in the border zones between these three main branches; the cerebellum and brainstem are supplied by the vertebral arteries, basilar artery, and their branches. Problems are that there is little if any relationship with function, that there is no distinction between partial and complete infarcts in a given territory, and that the boundaries between different territories vary substantially between individuals.

Classification according to the cause of ischaemic stroke is of interest for studies aiming to describe or influence the pathophysiological background of strokes. The so-called 'TOAST (Trial of Org 10172 in Acute Stroke Treatment) classification', for example, distinguishes five subtypes of ischaemic stroke:

1. large-artery atherosclerosis;
2. cardioembolism;
3. small-vessel occlusion;
4. stroke with other specific cause; and
5. stroke with undetermined cause.

At present, about 40 per cent of patients would presently end up in the category 'undetermined cause', even in specialized stroke services. Moreover, these distinctions can only be applied after a few days in hospital. Finally, and most important, the system is not suited for assessing the severity of stroke.

Rehabilitation specialists and geriatricians will be more interested in the functional abilities of patients than in the niceties of neurological nosology. They mostly grade patients' disability on a scale for activities of daily life (such as the Barthel scale, which ranks 10 in-house activities in hierarchical order, from bowel continence to taking a bath), or on a scale that includes some elements of social role fulfilment ('handicap'), such as the Rankin scale (Table 4).

A system that strikes a useful compromise between the functional and the anatomical point of view is the classification of the Oxfordshire Community Stroke Project, which distinguishes four categories:

1. total anterior circulation infarcts (**TACI**), with both cortical and subcortical involvement, representing about one-sixth of all ischaemic strokes in the community;
2. partial anterior circulation infarcts (**PACI**), with more restricted and predominantly cortical infarcts (one-third of all infarcts);
3. posterior circulation infarcts (**POCI**), clearly associated with the vertebrobasilar arterial territory (one-quarter); and
4. lacunar anterior circulation infarcts (**LACI**), confined to the territory of the deep perforating arteries (one-quarter).

Although the classes are anatomically defined, they contain important prognostic information: case fatality is highest by far in the TACI group.

Syndromes of cerebral infarction

Occlusion of the internal carotid artery may cause no symptoms at all or infarction in the entire territory of the ipsilateral anterior and middle cerebral artery (and sometimes of the posterior cerebral artery or contralateral anterior cerebral artery as well), depending on the presence of a complete circle of Willis and other collaterals. If arterial dissection is the cause of carotid occlusion, subadventitial bulging of the artery may cause Horner's syndrome and lower cranial nerve palsies, with or without infarction. Occlusion of the anterior, middle, and posterior cerebral arteries may lead to complete or partial infarction in their respective territories, depending on collaterals at the surface of the brain. Obviously, branch occlusions cause smaller infarcts. What follows is a description of syndromes associated with complete infarction in the average territory of the main cerebral arteries.

Infarcts in the area of the anterior cerebral artery

These cause contralateral hemiparesis more marked in the leg than in the arm, with no or only mild sensory deficit. Other frontal lobe features include mutism, incontinence, and apathy or, conversely, disinhibition.

Middle cerebral artery infarcts

If complete, these typically present with contralateral hemiplegia (most marked in the arm), sensory deficit, hemianopia, and cognitive defects such as aphasia (dominant hemisphere) or contralateral neglect (non-dominant hemisphere). Massive infarction of the entire territory of the middle cerebral artery may lead to such a degree of brain swelling that fatal herniation occurs, especially in young patients without cerebral atrophy.

Occlusion of a vertebral artery

Where this involves the origin of the posterior inferior cerebellar artery, such occlusion causes Wallenberg's syndrome, with ipsilateral cerebellar ataxia through infarction of the inferior part of the cerebellum. In addition, it causes a slightly bewildering combination of deficits through infarction of the dorsolateral medulla: decreased skin sensation in the ipsilateral half of the face and the contralateral half of the body, ipsilateral Horner's syndrome, ipsilateral weakness of the soft palate, larynx and pharynx, and rotatory vertigo.

Basilar artery syndrome

The full basilar artery syndrome, with infarction of most of the pons and midbrain, consists of coma, tetraparesis including facial movements, and loss of all eye movements and of pupillary and corneal reflexes. There are two characteristic partial syndromes of the basilar artery. One is the locked-in syndrome (infarction of the base of the pons), with tetraparesis including facial movements and loss of horizontal eye movements. Consciousness is preserved through sparing of the reticular

formation, but patients can communicate only through vertical eye movements; these may not always be correctly interpreted or even noticed. The other is the top-of-the-basilar syndrome, with variable combinations of hemianopia or complete cortical blindness (occipital lobes), amnesia (inferior temporal lobes), as well as vertical gaze palsies, pupillary disturbances, and hallucinations (perforating branches to the midbrain).

Posterior cerebral artery syndrome

This may include hemianopia (occipital lobe), amnesia (lower temporal lobe), and oculomotor disorders or disturbances of language or visuospatial function, through the involvement of perforating branches to the thalamus.

Occlusion of a single perforating artery

Such an occlusion, of one of the many arterioles that originate at right angles from a large parent artery to supply a small area in the deep regions of the brain or brainstem (Fig. 4), may be clinically silent, or cause a so-called 'lacunar syndrome'. A necessary condition for the clinical diagnosis of a lacunar syndrome is the absence of 'cortical' deficits, such as aphasia, neglect, hemianopia, and conjugate deviation of the eyes. The most common and archetypal form is pure motor stroke. In these cases the small, deep infarct strategically involves corticospinal fibres (pyramidal tract) to the motor neurones of the limbs, anywhere in its course. Analogous fibres to the facial nucleus in the pons may be affected as well. The infarct can be located in the corona radiata, adjoining the wall of the body of the lateral ventricle, or slightly more caudally, in the posterior limb of the internal capsule, or, less commonly, in the pons or the medulla. Other 'lacunar syndromes' are sensorimotor stroke (corona radiata or internal capsule), pure sensory stroke (thalamus), and ataxic hemiparesis (usually the base of the pons). Lacunar infarcts in the brainstem may lead to an almost infinite range of syndromes, often with the name of a French nineteenth century neurologist attached to it. Often such syndromes consist of an ipsilateral cranial nerve deficit and a contralateral hemiparesis.

Fig. 4 Small, deep infarct ('lacune') in a 63-year-old woman. CT scanning shows a small area of hypodensity (distinct from sulci) in the left brain hemisphere (on the reader's right), just lateral to the internal capsule.

Treatment of acute cerebral infarction

Several medical interventions aim at dissolving the occluding clot, or at least preventing it from growing: thrombolysis, antiplatelet agents, and anticoagulants. A different strategy, not yet well developed, is to protect ischaemic brain tissue. In addition, some underlying causes of stroke need urgent treatment, such as endocarditis. Before considering these specific measures, it is appropriate to consider the appropriate hospital setting in which stroke patients should be cared for.

Stroke units versus general wards

Specially organized stroke units can be a ward or team that exclusively manages stroke patients (a dedicated stroke unit) or a ward or team that provides a generic disability service (a mixed-assessment or rehabilitation unit). According to a meta-analysis of 20 randomized trials, care in a stroke unit reduces the risk of death or institutionalized care by 13 per cent. The observed benefits are independent of patient age, sex, stroke severity, and types of stroke-unit organization. No single element responsible for the benefits of organized stroke care has so far been identified, and probably there is none. The strength of stroke units lies in the integration of multidisciplinary efforts: stroke physician, nursing staff, physiotherapists, occupational therapists, speech therapists, rehabilitation physicians, and social workers.

Thrombolysis

Restoration of blood flow, to reperfuse the ischaemic brain as soon as possible after the cerebral artery has been occluded, and irrespective of its cause, should theoretically lead to a reduction in the volume of brain damaged by ischaemia and to an improvement in clinical outcome, analogous to myocardial infarction.

The main agents tested so far in the treatment of stroke (17 trials in over 5000 patients) are intravenous recombinant tissue plasminogen activator (**r-tPA**) and intravenous streptokinase, each in about half the patients. Almost all were treated within 6 h of stroke onset. Across all trials there was an excess of symptomatic intracranial haemorrhages (3 per cent in controls versus 10 per cent in treated patients). For every 1000 patients treated this corresponds to 70 extra intracranial haemorrhages, of which 44 are fatal. However, patients who survived the treatment were, on average, less disabled. For the outcome criterion 'death or dependence at the end of follow up' (3 months for most trials) the proportion was 59 per cent in the control group and 55 per cent among treated patients. This corresponds to a net gain of about 40 patients avoiding death or dependency for every 1000 patients treated with thrombolysis within 6 h, despite the excess haemorrhages.

There are still many unanswered questions about the role of thrombolysis in the treatment of ischaemic stroke. The first of these is the time window. The earlier treatment is given the better, which is confirmed by subgroup analysis of patients treated within 3 h, but inevitably this subgroup is biased towards more severe strokes (these get to hospital quickest). Second, is one agent better than another? For r-tPA alone, the balance of risks and benefits seems more favourable than for all agents together: per 1000 patients treated, an excess of 30 fatal intracranial haemorrhages, and a net result of 60 patients avoiding death or dependence, but the difference with streptokinase treatment is not statistically significant. Third, we can roughly identify patients in whom the risk of haemorrhage in the infarcted tissue is great (those with the most severe deficits and those with early signs of extensive infarction), but the potential benefits are also greatest in this group. More controlled studies are needed. Another question is whether the gains in patients with stroke are not offset by unbalancing the 'worried well' who will be also rushed to hospital, once stroke has been recognized as an emergency. There are many contraindications in view of the risk of cerebral haemorrhage, and only a minority of patients admitted with cerebral infarction can be treated with thrombolysis.

Antiplatelet agents

More than 99 per cent of the evidence from randomized trials in this area relates to the use of aspirin. The pooled results of two trials with aspirin (160–300 mg), started within 48 h of the onset of stroke, concluded that 13 fewer patients die or become dependent for every 1000 patients treated. There was no evidence of a net hazard in some 800 patients who had been inadvertently randomized after a haemorrhagic stroke. Only the combination with thrombolytic treatment should be avoided, because there are indications that aspirin enhances the danger of intracerebral haemorrhage.

Anticoagulants

Anticoagulants tested in clinical trials are standard unfractionated heparin, low molecular weight heparins, heparinoids, oral anticoagulants, and thrombin inhibitors. There is no evidence that anticoagulant therapy reduces the odds of being dead or dependent at the end of follow-up.

Neuroprotective agents

There are many steps in the destructive cascade between vessel occlusion and irreversible cell death where pharmacological intervention might be beneficial, at least

theoretically. The pharmaceutical industry has developed several compounds for clinical development and testing. There is no doubt that in animal models many neuroprotective agents, given either before or after the onset of ischaemia, reduce the area of cerebral infarction. So far, none of these agents has been proven to reduce disability in patients, despite dozens of clinical trials. Many other trials are under way, but reduction of disability by neuroprotective drugs is likely to be modest at best.

Surgical decompression of space-occupying infarcts

To prevent brain herniation, some centres in Germany have adopted the procedure of removing large parts of the skull vault in patients with massive supratentorial infarction, but all reports up to now have been uncontrolled. Clearly there is a need for randomized trials, which should take account of the quality of life of both patients and their carers.

With operations for space-occupying infarcts of the cerebellum there is a similar lack of controlled trials, but less controversy. Without operation, swelling of a cerebellar infarct can be fatal, whereas the deficits after surgical evacuation are surprisingly mild. In many patients, however, it is sufficient to relieve obstructive hydrocephalus, by external ventricular drainage.

Secondary prevention of ischaemic stroke

In the management of patients with TIAs or moderately disabling ischaemic strokes, it is often forgotten that the control of primary risk factors is by far the most effective way of diminishing the risk of stroke or other vascular events. First and foremost is blood pressure control, but also important are cessation of smoking, controlling diabetes and hyperlipidaemia, reducing overweight, and daily exercise.

Specific measures to reduce the risk of threatened stroke are discussed below. Drug treatment depends on whether the likely cause is embolism from the heart or arterial disease. With sources in the heart, mostly from atrial fibrillation, anticoagulant therapy (to give an **INR** (international normalized ratio) between 2.5 and 4) is the first choice in the absence of contraindications; no evidence exists for a fixed age limit. In all other patients, aspirin is the mainstay of treatment, but its preventive effect is only modest. Other antiplatelet drugs are only slightly more effective, if at all. Carotid endarterectomy is highly effective in patients with severe, symptomatic carotid stenosis (80–99 per cent reduction of the original lumen diameter), but this is found in less than 10 per cent of all eligible patients with ischaemic events.

Aspirin

The preventive effect of aspirin, in different doses, has been studied in 11 placebo-controlled randomized trials, in over 8000 patients after a TIA or moderately disabling stroke. There is virtually no difference in risk reduction for daily doses between 30 mg and 1300 mg. The overall risk reduction is 13 per cent (95 per cent confidence interval, 6 to 19 per cent). Side-effects, mainly indigestion, nausea, heartburn, or gastrointestinal bleeding are more common as the dose increases, but absolute rates are difficult to compare between studies, owing to differences in criteria.

Thienopyridine derivatives

These antiplatelet agents block the adenosine diphosphate pathway of platelet aggregation. The oldest derivative, ticlopidine, when given in a dose of 250 mg twice daily, is about as effective as aspirin (any dose above 30 mg). The major disadvantages are, however, that it is definitively more toxic (diarrhoea and skin rashes are reported in 15–20 per cent, and neutropenia in about 1 per cent of patients) and that it is much more expensive than aspirin.

Clopidogrel (75 mg daily) was tested in a single trial, which included patients with ischaemic stroke and those who had suffered myocardial infarction or who had peripheral vascular disease. It reduced the risk of the composite outcome event of vascular death, non-fatal stroke, or non-fatal myocardial infarction by 8.7 per cent (95 per cent confidence interval, 0.3 to 16.5). For the stratum of patients with an ischaemic stroke, the advantage was not statistically significant. Again, the cost makes it unattractive as the drug of first choice.

Dipyridamole

The pharmacological actions of this drug on platelets are unclear. Large trials of its efficacy in the secondary prevention of stroke have only tested it together with aspirin. In the analysis of all major vascular events the largest trial showed a benefit for the combination therapy in comparison with aspirin alone, but four earlier, smaller studies did not. The overall analysis shows a marginal difference in favour of the combination therapy, but without further evidence it should not be accepted as standard treatment.

Anticoagulants

So far there is little evidence to support the use of anticoagulants in the secondary prevention of stroke, except in patients with atrial fibrillation. A large trial of patients with cerebral ischaemia who were in sinus rhythm used a target intensity of an INR between 3.0 and 4.5, which is not unusual for preventing arterial thrombosis; patients in the control group were treated with aspirin. The study was prematurely stopped because of a significant excess of bleeding complications, mostly intracerebral. Anticoagulant therapy with an intensity of an INR between 2.0 and 3.0 deserves further study.

Carotid endarterectomy

Although this operation was increasingly performed from the 1960s onwards, it was not until the 1980s that two randomized surgical trials were performed: one in Europe and one in North America. In patients with severe, symptomatic carotid stenosis (80–99 per cent reduction in lumen diameter) the risk of disabling or fatal stroke substantially decreases after surgery. On average, about eight patients need to be operated upon to prevent one ipsilateral ischaemic stroke occurring within 4 years. This basic risk difference varies with age and sex, and it levels off after 3 or more years from randomization (that is, 3½ years after the qualifying event). It should be kept in mind that carotid endarterectomy is indicated in only a minority (less than 10 per cent) of patients with TIAs or moderately disabling ischaemic strokes: the attacks have to be in the carotid territory, the patients should be fit and willing to undergo the operation, and the angiogram should show an accessible stenosis of over 80 per cent at the carotid bifurcation.

Venous occlusive disease

The advent of non-invasive brain imaging methods in the 1970s and 1980s resulted in increased recognition of cerebral venous thrombosis. Before that time, physicians only rarely considered the diagnosis in patients with otherwise unexplained headache, focal deficits, seizures, impaired consciousness, or combinations of these features.

Causal factors

Unlike arterial occlusion, cerebral venous thrombosis is only rarely (some 10 per cent) associated with damage to the vessel wall, by infection, tumour growth, or trauma. Much more frequent causes are inherited disorders of coagulation. The most common form is the factor V Leiden mutation, found in some 20 per cent of patients without other causes. Stagnant flow (completing Virchow's triad of causes of thrombosis), contributes no more than a few per cent. In 20 per cent of patients no causal factors can be identified.

Often there is no single cause but a combination of contributing factors: for example, the postpartum period and protein S deficiency; pregnancy and Behçet's disease; or oral contraceptive drugs and the factor V Leiden mutation. The risk of cerebral venous thrombosis in the postpartum period increases with maternal age and with the performance of a caesarean section.

In neonates, cerebral venous thrombosis is usually associated with acute systemic illness, such as shock or dehydration; in older children, the most frequent underlying conditions are local infection (the leading cause before the antibiotic era), coagulopathy, and, in Mediterranean countries, Behçet's disease.

Diagnosis of cerebral venous thrombosis

The clinical features of cerebral venous thrombosis consist essentially of headache, focal deficits, seizures, and impairment of consciousness, in various combinations and degrees of severity. The symptoms and signs depend on which sinus is affected, and for a large part on whether the thrombotic process is limited to the dural sinus or extends to the cortical veins.

In the case of the superior sagittal sinus, which is affected in 70 to 80 per cent of all cases, cerebral venous thrombosis alone will lead to the syndrome of intracranial hypertension, that is headache and papilloedema. Up to 30 per cent of patients with so-called 'benign intracranial hypertension' may in fact have sinus thrombosis. Papilloedema can cause transient visual obscurations and sometimes irreversible constriction of visual fields, beginning in the inferonasal quadrants. The increased pressure of the cerebrospinal fluid may also give rise to VIth nerve palsies, and sometimes to other cranial nerve deficits. The onset of the headache is usually gradual, but in up to 15 per cent of patients it is sudden and may initially suggest the diagnosis of a ruptured aneurysm.

Involvement of cortical veins causes one or more areas of venous infarction, with or without haemorrhagic transformation. If the affected veins drain into the sagittal sinus the venous infarcts are typically located near the midline in the Rolandic and parieto-occipital regions, often on both sides. In the case of the lateral sinus the venous infarct is usually located in the posterior temporal area.

Clinically, the infarcts manifest themselves through epileptic seizures, or through focal deficits such as hemiparesis or dysphasia. If unilateral weakness develops (with thrombosis originating in the superior sagittal sinus), it tends to predominate in the leg, in keeping with the parasagittal location of most venous infarcts. Obstruction of cortical veins draining into the posterior part of the superior sagittal sinus or into the lateral sinus will commonly lead to hemianopia, dysphasia, or a confusional state. Impairment of consciousness may result from multiple lesions in the cerebral hemispheres, or from transtentorial herniation and compression of the brainstem. Either epilepsy or a focal deficit is a presenting feature in 10 to 15 per cent of patients; in the course of the illness seizures occur in 10 to 60 per cent of reported series, and focal deficits in 30 to 80 per cent.

Involvement of the cortical veins alone, without sinus thrombosis and its associated signs of increased cerebrospinal fluid pressure, is an extremely rare occurrence. Thrombosis of the deep venous system, including the great vein of Galen, may lead to bilateral haemorrhagic infarction of the corpus striatum, thalamus, hypothalamus, the ventral corpus callosum, the medial occipital lobe and the upper part of the cerebellum. In those cases the clinical picture is often dominated by deep coma and disturbance of eye movements and pupillary reflexes.

Investigations

CT

CT scanning will readily show 'venous' infarcts. These do not correspond to a known arterial territory, but often show haemorrhagic transformation (Fig. 5); they are sometimes bilateral, in the parasagittal area, or supra- as well as infratentorial, or they are in the deep regions of the brain. In addition, CT scanning will often provide evidence of the underlying sinus thrombosis: the hyperdense sinus sign or, less reliably, the so-called 'empty delta sign' after injection of intravenous contrast material.

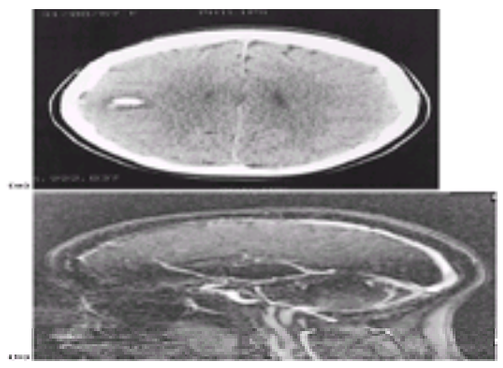

Fig. 5 Cerebral venous thrombosis in a 27-year-old woman. (a) This CT scan shows a small infarct with haemorrhagic transformation in the right brain hemisphere, adjacent to the top of the lateral ventricle. (b) Magnetic resonance imaging, focused on venous structures, shows non-filling of the frontal part (on the reader's left) of the superior sagittal sinus.

MRI

Magnetic resonance imaging has made catheter angiography redundant in the diagnosis of cerebral venous thrombosis. It is not sufficient to rely on non-visualization of a cerebral sinus on MR venography, since this may represent hypoplasia. Demonstration of the thrombus itself is essential, but this greatly depends on the interval from disease onset. Three stages can be distinguished. In the acute stage (days 1 to 5) the thrombus appears strongly hypointense in *T*2-weighted images and isointense in *T*1-weighted images. In the subacute stage (up to day 15) the thrombus signal is strongly hyperintense, initially on *T*1-weighted images and subsequently also on *T*2-weighted images. The third stage begins 3 or 4 weeks after symptom onset: the thrombus signal becomes isointense on *T*1-weighted images but remains hyperintense on *T*2-weighted images, though often inhomogeneously. Recanalization may occur over months in up to one-third of patients, but persistent abnormalities are common and do not signify recurrent thrombosis.

Treatment and prognosis

Anticoagulant treatment is plausible, but the evidence from controlled clinical trials is sparse. In the acute phase, heparin (low molecular weight heparin, either intravenously or subcutaneously) seems preferable to oral anticoagulants, because its intensity can be closely monitored. The totality of the evidence for heparin treatment consists of no more than 80 randomized patients; there is a non-significant trend towards a better outcome in treated patients. At least heparin treatment seems safe, even in patients with haemorrhagic infarcts. Local thrombolysis via endovascular catheters has only been performed in uncontrolled studies.

Death rates in different series range between 5 per cent and 30 per cent, and probably depend more on case mix than on treatment. Residual deficits consist mostly of hemispherical deficits or visual impairment from optic atrophy.

The risk of recurrence has seldom been addressed systematically, but is probably in the order of 10 per cent. It seems wise to advise other means of contraception than 'the pill'. In women with a peripartum episode of cerebral venous thrombosis the available evidence does not warrant the advice to avoid a further pregnancy, although in patients with the factor V Leiden mutation the risk of a recurrent episode is probably higher than average.

Primary intracerebral haemorrhage

Causes of primary intracerebral haemorrhage

In most cases there is no single cause for a primary intracerebral haemorrhage. Even in the classical example of a so-called hypertensive haemorrhage in the region of the basal ganglia, the question is what anatomical or other factors distinguished this patient from others, in whom there were similar degrees and duration of hypertension but no brain haemorrhage. Even a combination of recognized 'causes', such as that of hypertension and anticoagulants, does not invariably lead to intracerebral haemorrhage. In general, therefore, several causal factors combine. These can be broadly distinguished into three categories (Table 5): anatomical factors (lesions or malformations of the brain vasculature), haemodynamic factors (blood pressure), and haemostatic factors (to do with platelet function or with the

coagulation system). Abnormalities of the vascular system account for the vast majority of haemorrhages. The type of underlying abnormality varies with age: below the age of 40 years arteriovenous malformations or cavernomas are the most common single causes, whereas between 40 and 70 years the most frequent sources are ruptured perforating arteries (deep haemorrhages); in the elderly one also finds haemorrhages in the white matter ('lobar' haemorrhages), commonly attributed to amyloid angiopathy.

'Hypertensive' intracerebral haemorrhage

'Hypertensive' intracerebral haemorrhage results from degenerative changes in small perforating vessels, in the deep regions of the brain (basal ganglia and thalamus; Fig. 6), or in the cerebellum or brainstem. Microaneurysms occur on these vessels but are not necessarily the site of rupture. It is probable that rupture of a single small artery leads to a cascade of secondary haemorrhages from adjacent arterioles. This might explain the rapid expansion of intracerebral haematomas seen during a single scanning procedure or on serial scanning. A stable phase is usually reached in a matter of hours.

Deep brain haemorrhages are not always a one-time event. The recurrence rate in the first year is 7 per cent, against 2 per cent per annum over the subsequent 6 years.

Amyloid angiopathy

This condition accounts for about 10 per cent of intracerebral haemorrhages. Its frequency rises with age, but so does that of 'hypertensive' haemorrhage. The underlying abnormality consists of patchy deposits of amyloid in the muscle layer of small- and medium-sized cortical arteries of the occipital, parietal, and frontal lobes. Amyloid can also be found in asymptomatic individuals, the proportion increasing with age. It is not found outside the brain and does not represent generalized amyloidosis. Haemorrhages associated with amyloid angiopathy typically occur at the border of the grey and white matter of the cerebral hemispheres. Recurrent haemorrhage associated with amyloid angiopathy is much more common than with 'hypertensive' small-vessel disease. Autosomal dominant forms of amyloid angiopathy occur in The Netherlands and in Iceland.

Possible manifestations of amyloid angiopathy other than haemorrhage are transient episodes of focal neurological deficits, and also intellectual deterioration, associated with diffuse demyelination of the subcortical white matter (leukoaraiosis).

Cerebral arteriovenous malformations (AVMs)

AVMs are tangles of dilated arteries and veins, without a capillary network between them. On angiography, they are recognizable by large feeding arteries and a rapid shunting of blood to enlarged and tortuous veins via a central nidus of dilated vessels. Haemorrhage is the initial clinical manifestation in 50 to 60 per cent of symptomatic AVMs. Other clinical features include epileptic seizures, headaches, and progressive neurological deficits. Demonstrable AVMs are the most common single cause of intracerebral haemorrhage in patients under 45 years of age (about 30 per cent).

Between 10 and 20 per cent of AVMs are associated with thin-walled saccular aneurysms. These occur on peripheral feeding arteries, not at the classical sites at the circle of Willis, and are likely sources of bleeding. In AVMs in which one or more aneurysms have formed, the annual risk of rebleeding is as high as 7 per cent, against 2 to 3 per cent per annum for other AVMs. If there is no associated aneurysm, the site of rupture is mostly on the venous side of the malformation.

Cavernous angiomas (cavernomas)

Cavernous angiomas consist of sharply demarcated areas with widely dilated and thin-walled vascular channels, without intervening brain tissue. They are often asymptomatic and are encountered in 0.5 per cent of routine postmortems—in the white matter or cortex of a cerebral hemisphere in about one-half of all cases, in the posterior fossa in one-third, and in the basal ganglia or thalamus in one-sixth. If a cavernoma is at all symptomatic, epileptic seizures are at least as common a manifestation as haemorrhage. The annual risk of haemorrhage in patients in whom the lesion presents with seizures or focal deficits is rather low, between 0.25 per cent and 0.6 per cent. After a first rupture, rebleeding is more frequent, around 4.5 per cent per annum. Haemorrhages from a cavernous angioma are rarely fatal.

Familial forms of the disorder occur in several countries around the world, and should be suspected if multiple cavernomas are found.

Diagnosis of primary intracerebral haemorrhage

History

The history sometimes suggests the cause of the haemorrhage. Previous epileptic seizures should raise suspicions about the presence of an arteriovenous malformation, cavernoma, or a tumour. Amyloid angiopathy should come to mind with a history of transient ischaemic attacks, intellectual deterioration, or both. A record of long-standing hypertension indicates small-vessel diseases as the most probable underlying condition in a patient with a haematoma in the basal ganglia or in the posterior fossa; on the other hand, hypertension is so common that it may coexist with other conditions. If the patient is known to have had cancer, haemorrhage into a brain metastasis is a strong possibility. The use of oral anticoagulants is a vital piece of information in patients with intracerebral haemorrhage, because their action should be neutralized as soon as possible. It is equally important to know about the use of recreational drugs, particularly cocaine and amphetamines. Finally, the circumstances preceding an intracerebral haemorrhage may help to identify its cause, such as puerperium (intracranial venous thrombosis, choriocarcinoma), or neck trauma (dissection of the vertebral or carotid artery).

Physical examination

The physical examination will provide rather few clues to the cause of an intracerebral haemorrhage, except for petechiae or bruising, which indicate a generalized haemostatic disorder, signs of malignant disease such as cutaneous melanoma, a collapsed lung or enlargement of the liver or spleen, or telangiectasias in the skin and mucous membranes. Finding a high blood pressure on admission is the rule, but only in about 50 per cent is there is evidence of long-standing hypertension. Retinal haemorrhages indicate intracranial bleeding in general, most often a subarachnoid haemorrhage. Heart murmurs may be coincidental but should at least raise the possibility of infective endocarditis, as should the finding of needle marks in possible drug addicts. The neurological examination will show focal deficits corresponding to the site of the lesion, with or without a decreased level of consciousness.

Investigations

Investigations should start with the usual tests of blood and serum. These will sometimes uncover a cause of intracerebral haemorrhage, such as a low platelet count or massive liver damage. Brain imaging (CT or MRI) is the most important single investigation in patients with suspected intracerebral haematomas. The location of the haematoma may to some extent indicate the underlying cause. Intraventricular extension of the haemorrhage occurs relatively often with deep, 'hypertensive' haemorrhages. The presence of a fluid-blood level within the haematoma strongly suggests an underlying coagulopathy, either iatrogenic or from haematological disease. A grossly irregular margin of a lobar haematoma suggests amyloid angiopathy. Multiple or recurrent haemorrhages in the white matter suggest amyloid angiopathy, at least in the elderly. Intracranial venous thrombosis should be suspected with irregularly shaped haemorrhages in the parasagittal region. Repeat brain CT after injection of contrast may pick up underlying lesions. Sometimes these can only be identified weeks later, when the lesion is no longer obscured by mass effects.

Treatment of primary intracerebral haemorrhage

Factors predicting the prognosis for the survival of patients with primary intracerebral haemorrhage are: level of consciousness (Glasgow Coma Scale); age; volume of haematoma (poor prognosis if supratentorial haematoma >50 ml); and intraventricular extension of haemorrhage (poor prognosis if volume >20 ml). Of course, the possible interventions outlined below apply only to patients who have a chance of survival.

In patients taking oral anticoagulants the first step is the intravenous injection of 10 to 20 mg of vitamin K, at no more than 5 mg/min, followed by infusion of a concentrate of the coagulation factors II, VII, IX, and X, or of fresh-frozen plasma.

Intracranial pressure is often raised. Factors other than the local effects of the haematoma may contribute, such as fever, hypoxia, hypertension, seizures, and elevations of intrathoracic pressure. An unsolved question is the use, in comatose patients, of monitoring and, if judged appropriate, lowering intracranial pressure. There are many believers of this approach but few controlled studies. Insertion of a ventricular catheter may be a definitive measure in patients with cerebellar haemorrhage and no signs of direct compression of the brainstem.

There is insufficient evidence of benefit for the surgical treatment of supratentorial haematoma. Randomized trials have at best been inconclusive, including those employing endoscopic evacuation. In patients with cerebellar haematomas there is no doubt that surgical evacuation can be lifesaving, often with surprisingly few neurological sequelae. Sound indications for evacuation are the combination of a depressed level of consciousness with signs of progressive brainstem compression (unless all brainstem reflexes have been lost for more than a few hours, in which case a fatal outcome is unavoidable), or a haematoma greater than 3 to 4 cm. If the patient has a depressed level of consciousness and hydrocephalus, without signs of brainstem compression and with a haematoma less than 3 cm, ventriculostomy can be carried out as an initial (and sometimes only) procedure.

Subarachnoid haemorrhage

Causes of subarachnoid haemorrhage

Ruptured aneurysms are by far the most common source of non-traumatic subarachnoid haemorrhage (**SAH**), about 85 per cent of cases. Around 10 per cent are non-aneurysmal perimesencephalic haemorrhages, the remaining 5 per cent is made up by rarities (Table 6).

Cerebral aneurysms

Cerebral aneurysms are not congenital, they develop during the course of life. Therefore aneurysmal haemorrhage in a child is extremely rare. The aneurysms are saccular in shape and mostly arise at sites of arterial branching at the base of the brain, at or near the circle of Willis (see Fig. 7). It is largely unknown why some adults develop aneurysms. There are families with two or more affected first-degree relatives, but these account for less than 5 per cent of all SAHs. Many classical risk factors for stroke in general also apply to SAH: smoking, hypertension, heavy drinking, and oral contraceptives. Not all aneurysms rupture. Their prevalence can be estimated from angiographic studies (for other purposes) and autopsy studies at approximately 2 to 3 per cent in middle age, up to 5 per cent at the end of life. On the assumption that this proportion is 1 per cent for a standardized population across all age groups, and given that the incidence of SAH is approximately 6 per 100 000 (of the entire population), the annual risk of rupture of an aneurysm is 0.6 per cent.

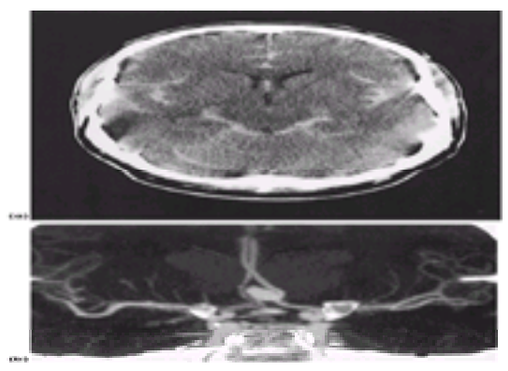

Fig. 7 Aneurysmal subarachnoid haemorrhage in a 31-year-old woman. (a) CT scanning shows evidence of extravasated blood throughout the basal cisterns. (b) CT angiogram, with intravenous contrast, shows an aneurysm at the anterior communicating artery.

Non-aneurysmal perimesencephalic haemorrhage

This is a distinct and benign variety of subarachnoid haemorrhage, in which the distribution of extravasated blood on the brain CT scan is different from that seen with aneurysms, in the cisterns around the midbrain or ventral to the pons. The angiogram is completely normal, and the long-term outcome is invariably excellent. This subtype constitutes 10 per cent of all subarachnoid haemorrhages and two-thirds of subarachnoid haemorrhages with a normal angiogram.

Diagnosis of subarachnoid haemorrhage

History

The key feature in the history is that of a sudden, severe, and unusual headache. However, 50 per cent of patients lose consciousness at the onset, and the headache may emerge only later. The diagnosis is most difficult in patients with headache as the only feature. In general practice, exceptionally sudden forms of common headaches outnumber ruptured aneurysms. The incidence of aneurysmal haemorrhage being about 6 per 100 000 of the population per year, the average general practitioner will, on average, see one such patient every 8 years. There are no single or combined features of the headache that distinguish reliably, and at an early stage, between SAH and innocuous types of sudden headache. The discomfort and cost of referring the majority of patients for only a brief consultation in hospital is a reasonable price to pay for avoiding misdiagnosis of a ruptured aneurysm.

Physical examination

The physical examination is unhelpful in patients with a headache alone, without loss of consciousness or focal deficits. Neck stiffness takes about 6 h to develop, so its absence soon after the onset does not make the diagnosis of SAH more unlikely.

Investigations

CT scanning is the most important investigation. This will show extravasation of blood in the basal cisterns of the brain in at least 95 per cent of patients with a ruptured aneurysm, if the scan is performed within 3 days (Fig. 7)—after this interval the sensitivity of CT scanning quickly decreases. In patients with a negative CT scan but a convincing history, lumbar puncture is indicated. If the cerebrospinal fluid (**CSF**) is blood-stained, it is essential to distinguish SAH reliably from a traumatic tap. For that purpose at least 6 h, and preferably 12 h, should have elapsed from symptom onset. In cases of SAH, sufficient lysis of red cells will have occurred in the meantime for bilirubin and oxyhaemoglobin to have formed. These pigments give the CSF a yellow tinge after centrifugation (xanthochromia); they are invariably detectable until at least 2 weeks later. The 'three tube test' (a decrease in red cells in consecutive tubes in the case of a traumatic puncture) is notoriously unreliable. If the supernatant seems crystal-clear, the specimen should be stored in darkness until the absence of blood pigments is confirmed by spectrophotometry. Cerebral angiography is necessary for demonstrating or excluding an aneurysm as the source of haemorrhage; catheter methods are rapidly being replaced by CT and MR angiography (Fig. 7).

Treatment of aneurysmal subarachnoid haemorrhage

Several complications may occur after a first episode of an aneurysmal SAH, of which rebleeding and cerebral ischaemia are the most dreaded. Despite advances in surgical and medical management, the population-based case fatality rate is still around 50 per cent, with half of the survivors remaining more or less disabled.

As general nursing measures, continuous observation and an intravenous access are essential. A bladder catheter is necessary for monitoring fluid balance. Headache should be relieved in a step-wise approach, with paracetamol and codeine as first steps. Distressing anxiety can be alleviated with short-acting

benzodiazepines. Stools should be kept soft with oral laxatives and also by an adequate intake of fluids.

Prevention of rebleeding is challenging, if only because any effective measure tends to be offset by an increased risk of ischaemia. Moreover, at least 10 per cent of all patients with SAH suffer a further bleed within hours of the initial haemorrhage. Over the next 4 weeks the rate of rebleeding without intervention is at least 30 per cent. The immediate case fatality rate of rebleeding is 50 per cent. Surgical clipping of the aneurysm is the most effective method of treatment, but the earlier the operation is performed, the greater the risk of ischaemic complications. Endovascular treatment ('coiling') is rapidly gaining ground; however, whether the balance between effectiveness and safety is more favourable than with surgery needs to be determined by clinical trials. Antifibrinolytic drugs decrease the rate of rebleeding but do not improve overall outcome.

Delayed cerebral ischaemia occurs in up to 25 per cent of patients with a ruptured aneurysm, mainly between days 5 and 14 after the initial bleed. Understanding its pathogenesis has been impeded by simplistic notions about 'vasospasm' or 'clots around vessels'. Narrowing of the arteries at the base of the brain is a factor, but not a sufficient one. The total amount of subarachnoid blood is a potent risk factor, but only after rupture of an artery, and the distribution of blood in the subarachnoid space does not predict the site of ischaemia. The calcium antagonist nimodipine, in a dose of 60 mg every 4 h by mouth or nasogastric tube, reduces the frequency of cerebral ischaemia and poor outcome by about one-third; its mode of action is incompletely understood. As a rule, hypertension should be left untreated; it is a compensatory reaction to maintain cerebral perfusion. The plasma volume should not be allowed to fall; hyponatraemia is caused by renal sodium depletion, and not, as still often believed, by dilution as a result of inappropriate secretion of antidiuretic hormone. Fluids should therefore be replaced and not restricted. The basic intake should be at least 3 litres per day, with intravenous fluids supplementing oral intake; compensation should be made for fever or a negative fluid balance.

Further reading

Algra A, van Gijn J (1999). Cumulative meta-analysis of aspirin efficacy after cerebral ischaemia of arterial origin. *Journal of Neurology, Neurosurgery and Psychiatry* **66**, 255. [Systematic review of aspirin in the secondary prevention of stroke.]

Antithrombotic Trialists' Collaboration (2002). Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high-risk patients. *British Medical Journal* **324**, 71–86. [Systematic review.]

Bamford J, *et al*. (1991). Classification and natural history of clinically identifiable subtypes of cerebral infarction. *Lancet* **337**, 1521–6. [Proposes a simple classification system for ischaemic stroke that combines anatomical and prognostic information.]

Barnett HJM, *et al*. (1998). Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. *New England Journal of Medicine* **339**, 1415–25. [One of two trials showing that carotid endarterectomy is indicated for patients with a recent, non-disabling, carotid-territory ischaemic event when the symptomatic stenosis is greater than about 80 per cent.]

Bousser M-G, Ross Russell RW (1997). *Cerebral venous thrombosis*. WB Saunders, London. [Comprehensive monograph.]

Brilstra EH, *et al*. (1999). Treatment of intracranial aneurysms by embolization with coils—a systematic review. *Stroke* **30**, 470–6.

De Bruijn SFTM, Stam J, for the Cerebral Venous Sinus Thrombosis Study Group (1999). Randomized, placebo-controlled trial of anticoagulant treatment with low-molecular-weight heparin for cerebral sinus thrombosis. *Stroke* **30**, 484–8. [Controlled clinical trial in 60 patients, showing that low molecular-weight heparin is safe.]

Dennis M, *et al*. (1990). Prognosis of transient ischemic attacks in the Oxfordshire Community Stroke Project. *Stroke* **21**, 848–53. [Provides a population-based, follow-up study about the outcome after transient ischaemic attacks.]

EAFT (European Atrial Fibrillation Trial) Study Group (1993). Secondary prevention in non-rheumatic atrial fibrillation after transient ischaemic attack or minor stroke. *Lancet* **342**, 1255–62. [Proves the effectiveness of oral anticoagulants.]

European Carotid Surgery Trial Collaborative Group (1998). Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European carotid surgery trial (ECST). *Lancet* **351**, 1379–87. [One of two trials, see Barnett *et al*. (1998) for comment.]

Gent M, *et al*. (1996). A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *Lancet* **348**, 1329–39. [Shows a marginal advantage of clopidogrel over aspirin.]

Greenberg SM (1998). Cerebral amyloid angiopathy—prospects for clinical diagnosis and treatment. *Neurology* **51**, 690–4. [Review.]

Gubitz G, Sandercock P, Counsell C (1999). Antiplatelet therapy for acute ischaemic stroke (Cochrane Review). *The Cochrane Library* **Issue 1, 2000**. Update Software, Oxford. [Reviews the evidence from controlled trials (mostly with aspirin).]

Gubitz G, *et al*. (2000). Anticoagulants for acute ischaemic stroke (Cochrane Review). *The Cochrane Library* **Issue 1, 2000**. Update Software, Oxford. [Shows there is no net benefit.]

Hop JW, *et al*. (1997). Case-fatality rates and functional outcome after subarachnoid hemorrhage—a systematic review. *Stroke* **28**, 660–4.

Koudstaal PJ, *et al*. (1992). TIA, RIND, minor stroke: a continuum, or different subgroups? Dutch TIA Study Group. *Journal of Neurology, Neurosurgery and Psychiatry* **55**, 95–7. [Shows how irrelevant it is to strictly distinguish ischaemic episodes of the brain according to their duration.]

Lemesle M, *et al*. (1998). Incidence of transient ischemic attacks in Dijon, France—a 5-year community-based study. *Neuroepidemiology* **17**, 74–9. [A recent study on the incidence of TIAs, summarizing preceding estimates.]

Linn FHH, *et al.* (1996). Incidence of subarachnoid hemorrhage—role of region, year, and rate of computed tomography: a meta-analysis. *Stroke* **27**, 625–9. [Shows the incidence of subarachnoid haemorrhage is lower than was estimated before the introduction of CT scanning.]

Linn FHH, *et al*. (1998). Headache characteristics in subarachnoid haemorrhage and benign thunderclap headache. *Journal of Neurology, Neurosurgery and Psychiatry* **65**, 791–3. [Shows that sudden headaches from a ruptured aneurysm cannot be distinguished from innocuous forms of headache.]

Mathew P, *et al*. (1995). Neurosurgical management of cerebellar haematoma and infarct. *Journal of Neurology, Neurosurgery and Psychiatry* **59**, 287–92. [Narrative review.]

Rinkel GJE, van Gijn J, Wijdicks EFM (1993). Subarachnoid hemorrhage without detectable aneurysm. A review of the causes. *Stroke* **24**, 1403–9. [Lists the causes of subarachnoid haemorrhage other than aneurysms.]

Rinkel GJE, *et al*. (1998). Prevalence and risk of rupture of intracranial aneurysms—a systematic review. *Stroke* **29**, 251–6.

Roos YBWEM, for the STAR Study Group (2000). Antifibrinolytic treatment in subarachnoid hemorrhage—a randomized placebo-controlled trial. *Neurology* **54**, 77–82. [Latest trial of antifibrinolytic drugs after aneurysmal subarachnoid haemorrhage, showing fewer rebleeds but no improvement in outcome.]

Stroke Unit Trialists' Collaboration (1998). Organised inpatient (stroke unit) care for stroke (Cochrane Review). *The Cochrane Library* **Issue 1, 2000**. Update Software, Oxford. [Reviews the evidence from controlled trials.]

Sudlow CLM, Warlow CP (1997). Comparable studies of the incidence of stroke and its pathological types—results from an international collaboration. *Stroke* **28**, 491–9. [Summarizes 11 reliable studies about stroke incidence.]

The Stroke Prevention in Reversible Ischemia Trial (SPIRIT) Study Group (1997). A randomized trial of anticoagulants versus aspirin after cerebral ischemia of presumed arterial origin. *Annals of Neurology* **42**, 857–65. [Shows that high-intensity anticoagulation is not safe for patients with TIA or moderately disabling stroke who are in sinus rhythm.]

Van der Wee N, *et al*. (1995). Detection of subarachnoid haemorrhage on early CT: is lumbar puncture still needed after a negative scan? *Journal of Neurology, Neurosurgery and Psychiatry* **58**, 357–9. [Shows that a few per cent of aneurysmal haemorrhages are missed by CT scanning, even in the first few days after symptom onset.]

van der Zwan A, *et al*. (1992). Variability of the territories of the major cerebral arteries. *Journal of Neurosurgery* **77**, 927–40. [Shows the inter-individual variability of boundaries between the territory of major cerebral arteries.]

Wardlaw JM, del Zoppo G, Yamaguchi T (1999). Thrombolysis for acute ischaemic stroke (Cochrane Review). *The Cochrane Library* **Issue 1, 2000**. Update Software, Oxford. [Reviews the evidence from controlled trials about the balance between risks and benefits.]

Warlow CP, *et al.* (2001). *Stroke—a practical guide to management*, 2nd edn. Blackwell, Oxford. [Comprehensive monograph about cerebrovascular disease, including a chapter about the historical background.]

24.13.8 Alzheimer's disease and other dementias

Clare J. Galton and John R. Hodges

Introduction

The definition of dementia has evolved from one of progressive global intellectual deterioration to a syndrome consisting of progressive impairment in memory and at least one other cognitive deficit (aphasia, apraxia, agnosia, or disturbance in executive function) in the absence of another explanatory central nervous system disorder, depression, or delirium (according to the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn (**DSM-IV**)). Even this recent syndrome concept is becoming inadequate, as researchers and clinicians become more aware of the specific early cognitive profiles associated with different dementia syndromes. For instance, in early Alzheimer's disease there may be isolated memory impairment many years before more widespread deficits develop.

Since dementia is predominantly a disorder of later life, it represents an increasing problem for individuals and society with the projected increase in the elderly population. It is estimated that the 18 million people with dementia worldwide will increase to 34 million by the year 2025. This increase is most marked in the developing countries, where the 11 million people with dementia in the year 2000 will reach 24 million by 2025. In the developed world, the equivalent figures are 7 million in 2000 and 11 million in 2025. In Europe alone, 4 million people will be affected by the year 2004.

Although the incidence of dementia is difficult to establish, community prevalence studies suggest that about 8 per cent of all people over 65 years of age suffer from dementia, this shows a marked increase with advancing age. The prevalence below 65 years is about 1:1000, this rises to 1:50 to the age of 70 and 1:20 from 70 to 80. Over 80 years of age the prevalence is 1:5.

Dementia has numerous causes that can be classified in many ways (Table 1 shows a classification by aetiology). Although a large number of medical and neurological conditions can occasionally cause a dementia syndrome, most of these are rare and have other neurological features that suggest the diagnosis, for example multiple sclerosis, the acquired immunodeficiency syndrome (**AIDS**) dementia complex, and the vasculitides. Routine investigation (see below) focuses on some of these rarer causes because, although rare, they often result in a reversible dementia. An alternative classification, based on the patterns of cognitive impairment, is that of subcortical and cortical dementias as illustrated in Table 2. This classification shows that disease of diverse cerebral structures can result in dementia but that the resultant patterns of cognitive deficits can be very different. Alzheimer's disease is the prototypical cortical dementia. Subcortical dementias are also discussed further below.

The most common causes of dementia before and after the age of 65 years are shown in Fig. 1. The relative frequencies of causes of dementia differ depending on age, but it is notable that Alzheimer's disease is the most common cause in both groups. The genetic forms of Alzheimer's disease and other rarer causes are more common in the younger age group. Before considering the common and treatable causes of dementia, we discuss the main differential diagnoses to be considered as alternatives to dementia.

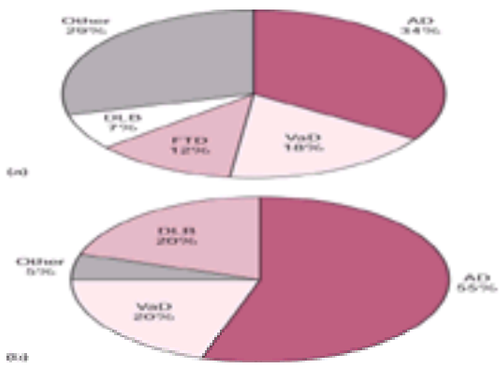

Fig. 1 (a) Relative frequencies of different dementia diagnoses in the under-65-year olds. (b) Relative frequencies of different dementia diagnoses in the over-65-year

olds. AD, Alzheimer's disease; VaD, vascular dementia; FTD, frontotemporal dementia; DLB, dementia with Lewy bodies.

Differential diagnosis

Pseudodementia

This term has been used to describe two disorders namely: depressive pseudodementia and hysterical pseudodementia. Cognitive symptoms are common in depression, particularly in the elderly population. The main complaints are of poor recent memory and concentration with distractibility. There is often a lack of subjective feelings of depression, thereby making the diagnosis difficult. The telltale signs are the so-called biological features of depression, such as sleep disturbance and a loss of appetite and libido. Other common symptoms are low energy and a lack of interest in hobbies and activities. There may be a past personal or familial history of depression. The cognitive picture is of impaired attention and subsequent patchy performance on memory and frontal tasks. There may be some inconsistency in test performance and patients easily give up on a task. Language output may be sparse but paraphrasic errors are not present. Even after detailed testing it may be difficult to distinguish depression from dementia, indeed there may also be some overlap between the syndromes in the elderly. For this reason, ideal practice would be for all newly presenting patients with dementia to undergo psychiatric assessment, and if any doubt remains a therapeutic trial of antidepressants may be warranted.

Hysterical pseudodementia commonly presents with a rapid onset of memory and/or intellectual impairment. There is loss of personal identity and salient personal and life events, which is unlike organic disorders of memory. There may be an obvious precipitant (such as marital problems, financial problems, or trouble with the law) and a past psychiatric history is common. 'Ganser syndrome' is a name for the condition where the patient gives bizarrely wrong answers to questions. For example, when asked 'How many legs does a horse have', they reply three or five. Even with such functional states, the examiner has to be aware of the potential concomitant organic disorder exaggerating the condition, as in other conversion disorders.

Delirium

This clinical syndrome is caused either by a diffuse brain pathology (for example, intracranial infections, head trauma, epilepsy (postictal states and non-convulsive status), raised intracranial pressure, subarachnoid haemorrhage) or is secondary to a large number of systemic illnesses or insults, including infections, metabolic derangements, hypoxia, and drugs.

The clinical features include the acute onset of attentional abnormalities and disturbance of consciousness (from clouding to coma), perceptual distortions, illusions and hallucinations, psychomotor disturbance (hypo- or hyperactivity and rapid shifts between the two), disturbance of the sleep–wake cycle, emotional lability, and marked fluctuations in performance and behaviour. The most consistent abnormality is in attention, with a reduced ability to maintain attention to external stimuli leading to distractibility and difficulty answering questions, and to appropriately shift attention to new stimuli leading to perseverations. The investigation and treatment needs to be focused in each case on the likely precipitants (although in about 5–20 per cent of the elderly no cause is found). Although the course and prognosis depend on the underlying diagnosis, if there is resolution of the precipitant there should be cognitive improvement to the baseline state.

Alzheimer's disease

Definition

Alzheimer's disease (**AD**) is the most common cause of dementia. Of the 5 to 10 per cent of the population aged over 65 years who have some kind of cognitive decline, over 50 per cent of cases will be due to AD and, although accounting for a smaller percentage of presenile cases, AD is still the single largest cause. The initial disease description by Alzheimer in 1907 was of a woman in her fifties with a progressive dementia and behavioural disturbance, who was found to have neurofibrillary tangles and amyloid plaques throughout her cerebral cortex. The term 'Alzheimer's disease' was then applied to similar cases with a presenile dementia, before it was realized that identical pathological changes were seen in the majority of elderly demented patients. Since plaques and tangles are found in a very high proportion of non-demented elderly subjects, debate continues about whether AD represents a continuum or a distinct disease process that increases in frequency with age. With recognition of a number of causative gene mutations (see below) AD is now generally believed to be a multifactorial disease with familial and sporadic forms.

Histological diagnosis remains the 'gold standard', but current research criteria, such as the widely used NINCDS-ADRDA (see Table 3), are accurate in up to 90 per cent of cases. Rather than merely being a diagnosis of exclusion, AD is now recognized as a clinicopathological entity amenable to positive diagnosis. Much recent research has focused on methods of early and accurate diagnosis, which is particularly important in view of the advent of potential disease-modifying treatments.

Epidemiology and risk factors

Age is the most important overall risk factor for AD. A positive family history is also a risk factor, although autosomal dominant presentations account for less than 5 per cent of cases. To date, three major causative gene mutations have been established: mutations in the presenilin genes I and II on chromosome 14 and 1, respectively, and involving the amyloid precursor protein (**APP**) gene on chromosome 21. In these families the onset is invariably at an early age (35–55 years), with remarkable consistency within families and, as with Huntington's disease, penetrance is complete. Dementia is rapidly progressive and seizures and myoclonus are common. Individuals with Down's syndrome (trisomy 21) develop Alzheimer's disease during their third and fourth decades. This is thought to be due to the extra copy of the amyloid precursor gene on chromosome 21.

Apolipoprotein E (**ApoE**) is a risk factor rather than a causative gene for AD in both early- and late-onset cases, which at present is thought to be the single most common genetic determinant of a susceptibility to late-onset Alzheimer's disease. ApoE is a component of several classes of plasma and cerebrospinal fluid lipoproteins. The brain is the most important site of ApoE production outside the liver, and ApoE is thought to be important in lipid homeostasis in the brain. There are three common alleles for the *ApoE* gene: *e2*, *e3*, and *e4*. One or two *e4* alleles confer an increased risk of Alzheimer's disease and lower the age of onset in a 'dose-dependent' fashion.

Many meticulous epidemiological studies have established that women are at an increased risk for AD, even after adjusting for confounding factors such as the increased longevity of women and their over-representation in the elderly population, and the increased vascular disease in men. Possible explanations include hormonal effects and the postmenopausal loss of potential protective effects of oestrogen. Significant head trauma in earlier life is also a risk factor that may summate with ApoE status, and there appears to be a unexplained protective effect of non-steroidal anti-inflammatory drugs.

Pathology

Pathologically, the macroscopic features of Alzheimer's disease are cortical atrophy, particularly involving the medial temporal lobe and parietotemporal association areas with relative sparing the primary sensory motor and visual cortices. The pathological process is thought to start in the entorhinal cortex, hippocampus, and other medial temporal lobe structures before spreading to the temporoparietal neocortex and basal frontal cortex, and then to the other association areas. The histological hallmarks are the senile plaques and neurofibrillary tangles (see Fig. 2 and Fig. 3, respectively). Neither lesion is specific for Alzheimer's disease, as both are found to a lesser extent in the ageing brain; neurofibrillary tangles are also seen in a range of diseases, including progressive supranuclear palsy, encephalitis lethargica, postencephalitic parkinsonism, cerebral trauma, and dementia pugilistica.

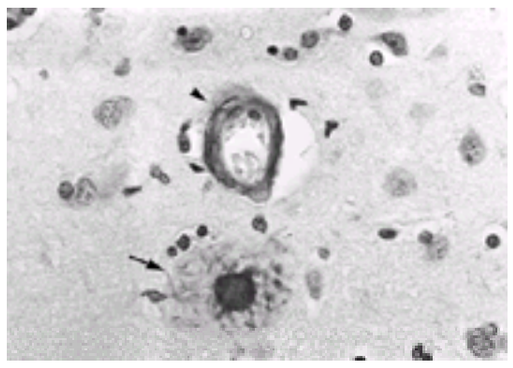

Fig. 2 Amyloid plaque.

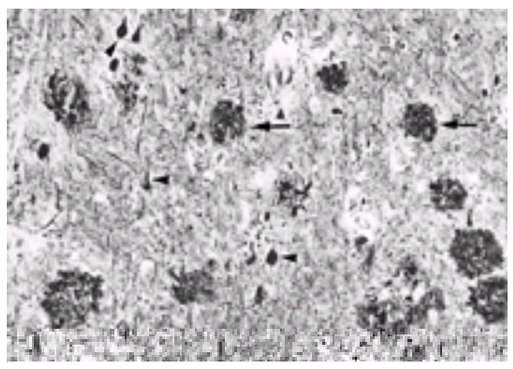

Fig. 3 Neurofibrillary tangle.

Neurofibrillary tangles are formed from bundles of paired helical filaments that replace the normal neuronal cytoskeleton. The central core of the paired helical filaments is the microtubule-associated protein tau. The abnormal phosphorylation of the tau protein causes the microtubular abnormalities and the subsequent collapse of the cytoskeleton. The neurofibrillary tangles are seen as intensely staining intraneuronal inclusions with silver stains or specific anti-tau immunochemistry.

A variety of amyloid plaques are observed in Alzheimer's disease. Diffuse amyloid plaques have a loose accumulation of b/A4 amyloid without surrounding abnormal neurites, and are considered to be precursors to neuritic plaques. The mature neuritic plaque consists of a dense core of b/A4 amyloid surrounded by a halo and ring of abnormal neurites, before this stage the plaque is a loose accumulation of b/A4 amyloid surrounded by abnormal neurites. A hypermature plaque has a dense core of b/A4 amyloid surrounded by reactive astrocytes but without abnormal neurones. The role of microvascular pathology in AD remains controversial. Cerebral congophilic angiopathy can be seen in a high proportion of cases and almost certainly contributes to the hyperintense lesions commonly seen on *T*2-weighted magnetic resonance imaging (**MRI**) scans.

Besides a reduction in synaptic loss from neurones, which may explain some cognitive sequelae of the pathology, there is a major loss of neurotransmitters—especially of acetylcholine. The 'cholinergic hypothesis' of neurotransmitter loss causing attentional and mnemonic dysfunction has been much investigated. There is certainly evidence of severe neuronal loss in the nucleus basalis of Meynert in the basal forebrain, the major site of cholinergic neurones, and the current therapies are armed at improving cognitive function through inhibition of anticholinesterases. There is also disruption to other neurotransmitters including the serotonin system.

Pathophysiology

The increased understanding of the genetics of Alzheimer's disease has led to some advances in theories of the molecular basis of this condition. The b/A4 amyloid is formed from the cleavage of a larger molecule, amyloid precursor protein (APP) of approximately 700 amino acids. Mutations in either the presenilin gene (presenilin I and II) or the amyloid precursor protein gene affect APP and its metabolism, supporting the amyloid cascade hypothesis of Alzheimer's disease pathogenesis (see Fig. 4). APP is cleaved by beta- and gamma-secretases to produce b/A4 amyloid at a length of between 39 and 43 amino acids: the shorter fragments remain in solution, while longer peptides, the result of abnormal cleavage sites, are more prone to aggregate and form insoluble amyloid deposits. Besides mutations in the *APP* gene altering the cleavage site, there is evidence that presenilin genes affect the gamma-secretase-mediated cleavage of the transmembrane section of APP. The amyloid hypothesis suggests that accumulation of beta-amyloid, by overproduction or failure to break down, leads to amyloid deposition, thereby causing amyloid plaques, and leading to neurofibrillary tangles and cell death. There is evidence that insoluble b/A4 amyloid is toxic and that this may disrupt calcium homeostasis and enhance the production of glutamate. This hypothesis is still controversial; it explains rare familial cases and the association with trisomy 21, but the role of tau pathology, the formation of neurofibrillary tangles, and ApoE are not yet fully incorporated into a unifying model.

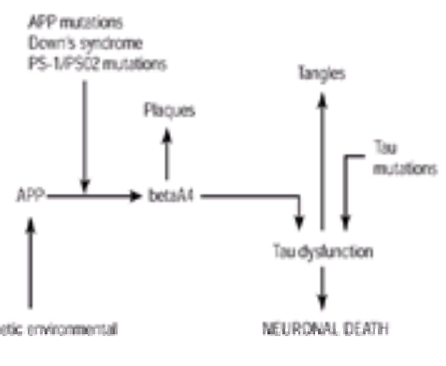

Fig. 4 Proposed pathogenesis of Alzheimer's disease.

Clinical features

The earliest cognitive deficit is impairment of so-called episodic memory (memories for events or episodes, including day-to-day memory and new learning), which is thought to reflect the earliest site of pathology in the medial temporal lobe structures. 'Minimal or mild cognitive impairment' (**MCI**) is a term increasingly used for people who are impaired on episodic memory tasks but who do not otherwise fit the criteria for a diagnosis of dementia. It is becoming clear that many, if not all, such people are in the predementia or early stage of AD, but progression to a full-blown dementia syndrome can take several years. The main clinical features at this stage are severe forgetfulness with often repetitive questioning and impairments in social function or job performance particularly concerning the retention of new information. As the disease progresses to mild Alzheimer's disease, memory function worsens, particularly affecting recall (for example, forgetting recent visits or family events), increasing disability in managing complex day-to-day activities such as finances and shopping, mental inflexibility and poor concentration, which reflects involvement of attentional and executive function. Insight is variably affected, often patients retain a partial awareness into their difficulties but underestimate the extent of the problem. Remote memory is relatively well preserved with a temporally graded pattern (that is, sparing of most distant memories). As the disease continues to progress patients often develop impairments in language, most typically word-finding difficulties, a shrinking vocabulary, and poor understanding of

complex words and concepts. Visuospatial impairments and apraxia, which may develop at this stage, are particularly disabling, causing difficulty in dressing, cooking, and performing other daily activities. In a small subgroup of patients, language or visuospatial difficulties can be the first or most prominent presenting feature. As the cognitive deficits progress there is worsening of language function and semantic memory, and behavioural problems can be prominent.

Neuropsychiatric symptoms are also common in the earliest stages of AD, particularly apathy, anxiety, and mood disturbance. Delusions and hallucinations occur in up to 50 and 30 per cent of patients, respectively, in the later stages. Agitation, restlessness, wandering, and disinhibition also cause considerable carer burden. The final stages of the disease are characterized by reduced speech output (or mutism), ambulatory difficulties, dependence, and incontinence. Seizures and myoclonus are common late features. There is considerable variation in the time to this stage, but the average time from diagnosis to death is around 10 years.

Neurological examination is unremarkable in the early stages, although increased tone (often frontal resistant, or gegenhalten, in type) and mild extrapyramidal features can occur as the disease progresses. Reflex changes such as extensor plantar responses (Babinski reflex) and—in contrast to frontotemporal dementia—pout, snout, and grasp reflexes occur late. In the final stages, there can be greatly increased rigidity and joint contractures.

Investigation

The aims of neuropsychological, imaging, and laboratory investigations in Alzheimer's disease are twofold: first to exclude other potentially reversible causes of dementia, and second to confirm the diagnosis of probable Alzheimer's disease. The extent and nature of investigation obviously needs to be tailored to the individual, but all patients should undergo brain imaging and have a neuropsychological assessment to confirm the diagnosis of dementia. The neuropsychological profile can also be informative in the differential diagnosis of dementia (see Table 2). Particularly characteristic is early impairment in delayed verbal recall of new material, followed by reduced category fluency (in which subjects are asked to generate exemplars from a given category, for example 'animals'), impaired naming of low-frequency words, and difficulty with complex visuospatial tasks such as copying complex figures or block design from the revised Wechsler Adult Intelligence Scale (**WAIS-R**).

The basic laboratory investigations required in all patients, particularly to exclude treatable causes of dementia, and some of the other investigations that may be indicated in certain cases depending on the patient's age, family history, or specific medical history are shown in Table 4. Research into biological markers of AD is yet to yield a consistent biological or surrogate marker. Screening for specific gene mutations in young-onset familial cases is only available in specialist centres.

Magnetic resonance imaging scans of patients with Alzheimer's disease in the earliest stages (including MCI) show evidence of atrophy of the hippocampus and entorhinal cortex (parahippocampal gyrus) reflecting the pathology (Fig. 5). Unfortunately, the variability in size of these structures in normal elderly subjects means that, at present, these imaging abnormalities are not specific enough to be of predictive value. The co-registration of serial MRIs appears capable of detecting abnormal rates of brain atrophy, even before the onset of clear-cut cognitive symptoms in at-risk familial cases, but it remains a research instrument. *T*2-weighted MRI often reveals periventricular high-signal changes even in 'pure' early-onset cases. Single-photon emission computed tomography (**SPECT**) scans similarly demonstrate typical abnormalities in the parietotemporal regions but the specificity is again low in individual cases. More recent technological developments, such as perfusion MRI, magnetic resonance spectroscopy (**MRS**), and photon emission tomography (**PET**) scans may enhance diagnostic accuracy but are expensive and not yet suitable for routine clinical use.

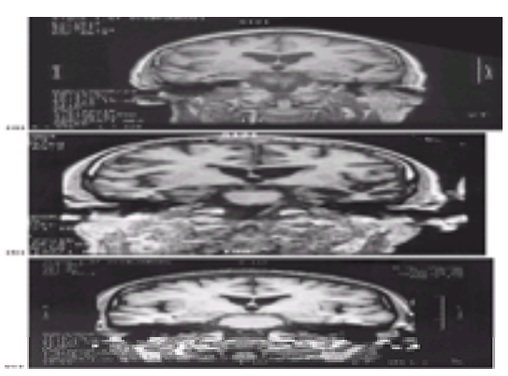

Fig. 5 (a) Coronal *T*1-weighted MRI scan of a patient with early Alzheimer's disease showing bilateral early hippocampal atrophy. (b) Coronal *T*1-weighted MRI image of a patient with FTD showing left temporal atrophy. (c) Coronal *T*1-weighted MRI scan of a normal subject.

Management and prognosis

The management of a patient with dementia involves many sensitive issues. It is crucial to provide medical and psychological support to patients as well as to their families and carers. During the progression of the disease there will be different treatment goals at different stages, ranging from aiding failing memory in the setting of independent living to managing behavioural problems and aggression, and eventually full supportive nursing care. There is great variation in the rate of progression, young-onset cases and those with prominent aphasia appear to deteriorate most rapidly. On average, patients spend several years in the mild or minimal stages (although it can be as long as 5 to 10 years), between 4 and 5 years in the moderate disease stages, and depending on the quality of care in the dependant stages, a year or more requiring full nursing care.

Non-pharmacological treatment

The mainstay of treatment is social support and increasing assistance with day-to-day activities. Issues such as driving and planning for future financial affairs are important and should be discussed early in the course of the disease. Throughout the course of the illness there will be differing requirements for the support services listed below:

- information and education;
- carer support groups;
- community dementia team, including home nursing and personal care;
- community services such as meals-on-wheels, community transport services, home maintenance assistance;
- sitter service;
- day centre;
- respite care; and
- residential/nursing home.

Pharmacological treatment

Pharmacological treatments can be divided into symptom- and disease-orientated approaches. Symptom modification relates to the treatment of depression, agitation, and psychotic phenomena and requires the input from a specialist psychiatrist. The cholinesterase inhibitors donepezil and rivastigmine are the only disease-specific drugs licensed for use in the United Kingdom. The importance of acetylcholine depletion in AD is established; these cholinesterase inhibitors generally achieve modest improvements in cognition in 25 to 50 per cent of the patients studied. However, the disease-modifying effects of these drugs remain controversial. Antioxidants (such as vitamin E), ginkgo biloba, and monoamine oxidase-B (**MAO-B**) inhibitors have shown benefits in some clinical trials, although again their long-term benefit is yet to be established. Pilot trials and experimental studies are being conducted at present in this area. Ideally, the goal is to prevent patients developing further cognitive deficits and to prevent those with MCI from progressing to dementia. The epidemiological findings of protection from cognitive decline in women using hormone-replacement therapy (**HRT**) is of interest in developing preventive strategies, and a trial is in progress to look at the effect of HRT in preventing or delaying the onset of dementia. Further research on the role of the amyloid and tau proteins in the pathogenesis of Alzheimer's disease will be the spur to both curative and preventive treatment for this common dementia.

Frontotemporal dementia

Definition

Frontotemporal dementia (**FTD**) is now preferred to the older term 'Pick's disease', to describe patients with focal frontal and/or temporal focal atrophy, since the underlying pathology of these syndromes can be variable. Arnold Pick (1851–1924) first described patients with both progressive aphasia, associated severe left temporal cortical atrophy at postmortem, and patients with behavioural disturbances associated with frontal lobe atrophy. In 1910, Alzheimer described the histological changes in patients with focal lobar degeneration as distinct from the syndrome that bears his name. Alzheimer described both argyrophilic intracytoplasmic inclusions (Pick bodies) and diffusely staining ballooned neurones (Pick cells). More recently it has become clear that these pathological changes are not the only features that accompany the clinical syndromes of frontal and temporal dementias; some patients have non-specific changes of neuronal loss, spongiosis, and gliosis only, while others have rather different ubiquitin-positive neuronal inclusions. The concept, therefore, has been broadened to accommodate differing underlying neuropathological features.

Epidemiology

FTD is increasingly recognized as a common cause of dementia, particularly in the younger age groups (see Fig. 1(a))—the peak incidence of onset being 45 to 65 years of age. In hospital series, the ratio of FTD to Alzheimer's disease has been found to vary from 1:5 to 1:20, with men and women being equally affected. Many cases are familial with up to 40 per cent having an affected family member.

Pathology and aetiology

The gross pathological appearance of FTD is that of profoundly atrophied frontotemporal regions that may be so severe as to produced the so-called knife-edged gyri and deep widened sulci. The histopathological hallmarks are widespread cortical and subcortical gliosis and loss of large cortical nerve cells. Severe astrocytosis with swollen neurones (Pick cells) and inclusions (Pick bodies), that are both tau- and ubiquitin-positive, are seen in about 20 per cent of cases. Pick bodies are intracytoplasmic argyrophilic neuronal inclusions composed of straight filaments, microtubules, and occasional paired helical filaments. In other patients there may be spongiform degeneration or microvacuolation of the superficial neuropil (cortical layer II) with no inclusions. Frontotemporal dementia can be seen with motor neurone disease, in which the histological changes above are combined with loss of anterior horn cells and motor neurone cells, particularly of the hypoglossal nuclei.

The aetiological basis of this disease is presently unknown. In some familial cases there is a mutation in the microtubule-associated protein tau gene on chromosome 17.

Clinical features

The presentation of frontotemporal dementia mirrors the neuropathological areas of disease; in the early stages frontal and temporal presentations can be distinguished.

Frontal presentations

Patients present with insidiously progressive changes in personality and behaviour that reflect the early locus of pathology in ventromedial frontal lobes. There is often impaired judgement, an indifference to domestic and professional responsibilities, and a lack of initiation and apathy. Social skills deteriorate and there can be socially inappropriate behaviour, fatuousness, jocularity, abnormal sexual behaviour, or theft. Many patients are restless with an obsessive–compulsive behaviour, such as hoarding food. Emotional lability and mood swings are seen, but other psychiatric phenomena such as delusions and hallucinations are rare. Patients become rigid and stereotyped in their daily routines and food choices. A change in food preference towards sweet foods is very characteristic. Of importance is the fact that simple bedside cognitive screening tests such as the Mini-Mental State Examination (**MMSE**) are insensitive at detecting frontal abnormalities. More detailed neuropsychological tests of frontal function (such as the Wisconsin Card Sorting Test or the Stroop Test) usually show abnormalities. Speech output can be reduced with a tendency to echolalia (repeating the examiner's last phrase). Memory is relatively spared in the early stages, although it does deteriorate as the disease advances. Visuospatial function remains remarkably unaffected. Primary motor and sensory functions remain normal. Primitive reflexes such as snout, pout, and grasp develop during the disease process. Muscle fasciculations, or wasting particularly affecting the bulbar musculature, can develop in the FTD subtype associated with motor neurone disease.

Temporal presentations

Temporal lobe degeneration presents with a form of progressive fluent aphasia, also known as semantic dementia, in which there is a profound loss in conceptual knowledge (or semantic memory) causing anomia and impaired comprehension of words, objects, or faces. The patient typically complains of 'loss of memory for words' and has fluent, empty speech with substitutions such as 'thing' 'one of those' etc., but the grammatical aspects are preserved. Naming is impaired with semantically based errors (such as 'animal' or 'horse' for zebra). Patients are unable to understand less frequent words and fail on a range of semantically based tasks such as matching words to pictures and matching pictures according to their meaning. Repetition of words and phrases is normal even though patients are unaware of their meaning. Unlike patients with Alzheimer's disease, day-to-day memory (episodic memory) with good visuospatial skills and non-verbal problem-solving ability is relatively preserved, at least in the early stages.

Another form of progressive focal atrophy, described by Mesulam, produces progressive non-fluent aphasia. Such patients present with a gradual loss of expressive abilities and gross impairments in the phonological (sound-based) and grammatical aspects of language production. This leads to non-fluent, agrammatical, and poorly articulated speech with multiple phonological errors (for example, sitter for sister or fencil for pencil). Repetition of multisyllabic words and phrases is impaired but, in contrast to semantic dementia, word comprehension and object recognition are well preserved.

Diagnosis

The diagnosis of frontotemporal dementia is based on the clinical, neuropsychological, and imaging assessments. The consensus broad clinical criteria are shown in Table 5. The differences between the various syndromes described above is obvious early in the disease, but there is increasing overlap between the temporal and frontal syndromes as the disease progresses. MRI demonstrates a characteristic pattern of frontal and/or temporal lobe atrophy: in contrast to Alzheimer's disease, the changes involve the polar and lateral temporal structures and are asymmetrical, commonly involving the left side to a greater extent (see Fig. 5). The functional imaging (single-photon emission tomography (**SPECT**) or positron emisson tomography (**PET**)) findings mirror the structural imaging results, with reduced frontotemporal perfusion and hypometabolism.

Management and prognosis

There is no curative treatment at present, thus the general management of the dementia sufferer and their family, as discussed above, is of prime importance. The prognosis can be variable with differing rates of progression between individuals. The disease is progressive and the average duration from diagnosis is around 5 to 10 years.

Dementia with Lewy bodies

Definition

Since the discovery in the 1960s that patients with Lewy bodies (ubiquitin-positive inclusions) in the cortex have a distinctive pattern of dementia with features of both Parkinson's and Alzheimer's disease, it has been increasingly recognized as an important cause of dementia. The terminology has been confusing with multiple designations including: Lewy body dementia, dementia of Lewy body type, diffuse Lewy body disease, and cortical Lewy body disease. The consensus clinical criteria for 'dementia with Lewy bodies' (**DLB**), the term now preferred, are shown in Table 6.

Epidemiology

Dementia with Lewy bodies is a common cause of dementia in the elderly population, although the true prevalence remains unclear. As many as 12 to 36 per cent of patients with a clinical diagnosis of Alzheimer's disease reach the pathological criteria for a diagnosis of dementia with Lewy bodies.

Pathology

Pathological criteria require the presence of cortical and subcortical Lewy bodies. Confusingly, there is considerable overlap with the histological features of both Parkinson's and Alzheimer's diseases, although the distribution of pathology is the key to distinguishing these conditions. Lewy bodies are intracytoplasmic eosinophilic neural inclusions formed from altered cytoskeleton components that can be seen on haematoxylin and eosin staining, but are more prominently shown using anti-ubiquitin immunohistochemistry. Cortical Lewy bodies are found in the temporal lobe, insular cortex, and cingulate gyrus, and are always accompanied by typical 'core and halo' Lewy bodies in the substantia nigra (the pathological hallmark of Parkinson's disease). Dystrophic ubiquitin-positive neurites are also seen in the hippocampus, amygdala, nucleus basalis of Meynert, and other brainstem nuclei.

Alzheimer changes—neurofibrillary tangles and amyloid plaques—are seen in up to 50 per cent of cases, raising nosological issues with Alzheimer's disease. The distribution of changes is of importance in distinguishing the conditions: for example, neurofibrillary tangles in DLB commonly spare the hippocampus, which is severely affected in Alzheimer's disease.

The neurotransmitter changes in DLB reflect the areas of pathology, with severe dopamine depletion in the basal ganglia and marked reduction in acetylcholine throughout the cortex.

Clinical features

Patients typically present with a progressive cognitive decline paralleling that seen in those with Alzheimer's disease. There are, however, a number of characteristic and distinguishing features. First, there is a tendency to marked spontaneous fluctuations in cognitive abilities, particularly alertness and attention, producing a delirious state lasting days or even weeks. Second, visual hallucinations, illusions, and fleeting misidentification phenomena occur in 50 to 80 per cent of sufferers even at an early stage and without drug provocation. The hallucinations are commonly well-formed images of people or animals. The marked cholinergic deficit is postulated to be the cause of their tendency to visual hallucinations. Third, is the occurrence of spontaneous parkinsonism, which is usually mild in the early stages. Rigidity, gait disturbance, and bradykinesia are all common, although in contrast to patients with Parkinson's disease the tremor is usually mild and atypical (with postural and action components) and symmetrical. Repeated falls also occur. In the later stages the akinetic rigid syndrome can cause severe disabilities in mobility and swallowing an increase in the number of falls. Fourth, there is often an exquisite sensitivity to neuroleptic medication, producing the malignant neuroleptic syndrome (delirium, hyperpyrexia, muscle rigidity, massive elevation of creatine phosphokinase, and renal failure).

Diagnosis

Neuropsychologically there is a mixture of subcortical and cortical features, with prominent cognitive slowing plus impairment of executive (planning and organizational abilities) and visuoperceptual abilities. Compared with patients with Alzheimer's disease, those with DLB tend to have greater deficits in attention and visuospatial processing. Memory loss may be less prominent than in Alzheimer's disease. There is no diagnostic test for this condition and the diagnosis *in vivo* relies on the clinical features described above and in Table 6. Brain imaging demonstrates similar changes to Alzheimer's disease, although there is a suggestion that medial temporal lobe atrophy is less pronounced.

Management

The symptomatic management of this disorder is complicated by the presence of both hallucinations and an akinetic rigid syndrome. Patients are notoriously sensitive to the side-effects of dopamine-enhancing medications used for the treatment of the akinetic rigid syndrome. However, although dramatic motor improvements are not to be expected, a cautious medication trial is worth attempting. Even though neuroleptic drugs should be avoided whenever possible, neuropsychiatric features, if severe, can be ameliorated with the newer atypical neuroleptics such as clozapine and olanzapine, without exacerbation of the parkinsonism. Thus the main aim is to maintain a balance between the patient being mobile and lucid.

Of considerable interest is the anecdotal improvement in the marked attentional cognitive deficits to treatment with cholinesterase inhibitors such as donepezil and rivastigmine. Although there have been no controlled trial reports as yet, patients with DLB may respond better than those with Alzheimer's disease to this drug therapy.

Vascular dementia

Definition and epidemiology

Vascular dementia can be defined as a dementia resulting from a cerebrovascular disorder. This is obviously a broad categorization and many different aetiologies may be included in this rubric, for example multiple infarcts from cardiac emboli, vasculitides including systemic lupus erythromatosus, primary cerebral amyloid angiopathy, and cerebral autosomal dominant arteriopathy with subcortical infarcts and leucoencephalopathy (CADASIL). The term 'multi-infarct dementia' was introduced in the 1970s to emphasize the contribution of multiple cerebral infarcts to clinical dementia syndromes and to replace the older label of 'atherosclerotic dementia', although it is now apparent that diffuse small-vessel disease contributes significantly in the absence of clinically overt strokes. Traditionally regarded as the second commonest cause of dementia, it is increasingly difficult to estimate the true contribution of vascular disease. Postmortem studies of patients with multi-infarct dementia show that Alzheimer's changes commonly coexist. Conversely, the advent of sensitive instruments for detecting cerebral vascular lesions *in vivo* (magnetic resonance imaging), has revealed that presumed vascular changes are common in patients with the clinical diagnosis of Alzheimer's disease, even in young patients with known gene mutations, and that the presence of vascular lesions may be contributing to the severity of Alzheimer's disease. Finally, it is increasingly apparent that traditional risk factors for vascular dementia—including hypertension, diabetes, hypercholesterolaemia—are also factors which increase the likelihood of developing Alzheimer's disease.

Clinicopathological vascular syndromes

The variety of vascular diseases that affect the brain are legion, and the resultant clinical features and underlying pathology widely different (see Table 1). The most important vascular syndromes will be considered below.

Large infarcts

Recurrent cerebral infarcts involving multiple main arterial territories (for example, posterior or middle cerebral artery territories), resulting from thrombosis or embolism, can cause dementia with a step-wise cognitive decline. There is commonly a history of atherosclerotic risk factors (for example, hypertension, smoking, and hypercholesterolaemia), other evidence of atherosclerotic cardiac or peripheral vascular disease, and neurological signs on examination (for example, spasticity, hyperreflexia, extensor plantor responses, and a pseudobulbar palsy). There are often asymmetries on the neurological examination, and gait apraxia and/or bladder dysfunction can be early features. The cognitive picture is characterized by cortical features and is dependent on the sites of the lesions. There is often severe language impairment, visuospatial disturbance, amnesia, and dyspraxia, related to lesions in the middle and posterior cerebral artery distributions. Specific syndromes can result from discrete lesions: for example, lesions of the left angular gyrus result in a fluent aphasia, agraphia, acalculia, right–left disorientation, and finger agnosia or Gerstmann's syndrome.

Lacunar infarcts

The small multiple lacunar lesions are caused by occlusion in the deep penetrating arterial branches. The underlying pathogenic mechanism is a distinct small-vessel arteriopathy with replacement of the muscle and elastin in the arterial wall by collagen, leading to tortuous vessel and microaneurysm formation as a result of

long-standing hypertension. The basal ganglia, thalamus, and deep white matter are common sites for lesions, due to the nature of the arterial supply. These lacunes may coexist with the larger infarcts (described above) thereby contributing to a mixed picture. However, the typical presentation of the lacunar state is with a more subcortical syndrome causing impaired attention and frontal executive malfunction, forgetfulness, apathy, and emotional lability. Thalamic lacunes can result in a speech disorder and, if bilateral, in amnesia. Examination features are similar to those seen with larger infarcts, with rigidity, gait disturbance, and extrapyramidal and pyramidal signs.

Small-vessel disease (Binswanger's disease)

'Binswanger's disease' (or 'diffuse leucoaryosis') is the term applied to the radiologically defined syndrome of confluent subcortical and corpus callosal demyelination and loss of the cerebral white matter, which again typically complicates severe or accelerated hypertension. The clinical features are similar to those of the lacunar state described above. On CT there is symmetrical diffuse low-density periventricular hypodensity, which can be accompanied by ventricular dilatation. This is visualized with great sensitivity on *T*2-weighted MRI as a diffuse white-matter of high intensity. Pathologically, there is demyelination, axonal loss, and gliosis, thought to be due to diffuse ischaemia in the territory of the long perforating arteries.

Cerebral amyloid angiopathy

Amyloid is deposited in the cerebral vessels both with increasing age and in a proportion of cases with ordinary Alzheimer's disease. However, there is also a rare and sometimes familial form of cerebral amyloidosis that produces recurrent cerebral haemorrhages and an Alzheimer's type dementia. Amyloid deposition in the vessel walls causes structural weakness leading to intracerebral haemorrhages and narrowing of the vessel to produce ischaemia. The haemorrhages tend to be lobar and can be recurrent.

Cerebral autosomal dominant arteriopathy with subcortical infarcts and leucoencephalopathy (CADASIL)

This recently established disorder may be a commoner cause of vascular dementia than previously realized. Patients present in their early twenties with migraine-like headaches and subsequently develop stroke-like episodes, which are sometimes ascribed to migraine or may mimic the attacks of acute demyelination. A subcortical dementia syndrome develops during their fifth and sixth decades. MRI shows multiple subcortical infarcts and diffuse white-matter disease. Other clues to the diagnosis are the absence of risk factors for atherosclerotic disease and the strong family history. Pathologically there is a distinctive non-amyloid, non-atherosclerotic angiopathy of the leptomeningeal and perforating arteries of the brain, with eosinophilic granular substance replacing smooth muscle. The diagnosis can be also confirmed with the finding of the same pathological changes in the cutaneous blood vessels in a skin biopsy. Mutations in the *notch3* gene on chromosome 19 have been reported in patients with CADASIL.

Treatment of vascular dementia

The treatment should be directed to the amelioration of any underlying cause of the vascular disorder, such as reducing cardiac embolism and treating vasculitides and hypertension. The potential for altering the progression of the disease is alluring. Nevertheless, efforts directed at altering atherosclerotic risk factors tend to produce disappointing results. The course of vascular dementia can be as severe as or even more rapid than that of Alzheimer's disease.

Subcortical dementias

Despite shortcomings, the differentiation between cortical and subcortical dementias continues to be useful in clinical practice. This classification highlights the fact that, although disease of diverse cerebral structures can result in dementia, the resultant patterns of cognitive deficits are very different. Alzheimer's disease is the prototypical cortical dementia, vascular syndromes can present with a spectrum of features from cortical to subcortical, as can dementia with Lewy bodies. Purer forms of subcortical dementia result from pathology of the basal ganglia and white matter, the prototypical examples being Huntington's disease and progressive supranuclear palsy (Steele–Richardson–Olszewski syndrome). The typical cognitive pattern is that of attentional and executive dysfunction with marked cognitive slowing (bradyphrenia) causing problems with mentation and information retrieval. Memory is moderately impaired due to reduced attention and poor registration, but is not as severely impaired as in Alzheimer's disease. There is often an associated personality change and mood disturbance with prominent apathy. Spontaneous speech is impoverished and slow.

Huntington's disease

Huntington's disease is an autosomal dominant inherited disorder with an incidence of about 4 per 100 000. The mutation is an expansion of the trinucleotide repeat (CAG) in the IT-15 gene on chromosome 4, which encodes the polyglutamine protein, huntingtin, essential for nervous system development. There is a clear dose-response relationship between the length of the CAG repeat and the age of onset of the disorder. Psychiatric symptoms, such as depression, irritability, and personality changes often precede the motor disorder, which is typically choreiform. The other cognitive changes that develop over the next 10 to 20 years are of a subcortical pattern, with deficits in attention and concentration, executive function, and retrieval from memory.

Progressive supranuclear palsy

Progressive supranuclear palsy (**PSP**) is a rare, but increasingly recognized, disorder with an incidence of 1 to 2 per 100 000. The subcortical dementia is accompanied by an atypical parkinsonian syndrome. The motor deficits are symmetrical in onset, with severe rigidity in the axial muscle groups and bulbar symptoms. A supranuclear gaze palsy invariably develops, but in the early stages the only feature may be slowing of fast downward movement (saccadic slowing). Another early feature is a marked tendency to falls. The pathological features are neurofibrillary tangles, neuropil threads, and neuronal loss and gliosis in the subthalamic nucleus, red nucleus, substantia nigra, and dentate nucleus. The main neurotransmitter deficit is in dopamine. Unlike Parkinson's disease, progressive supranuclear palsy does not respond well to levodopa. The disease progresses rapidly with an average time course of around 5 years.

Parkinson's disease

Subcortical dementia occurs in about one-third to one-half of patients with Parkinson's disease, which develops at a late stage in the motor disorder in contrast to dementia with Lewy bodies.

Corticobasal degeneration

Corticobasal degeneration is a rare cause of a dementia and motor signs. Patients present with an asymmetrical akinetic rigid syndrome together with limb apraxia, and the almost pathognomic feature of alien limb phenomenon in which the hand(s) act as if 'with a will of their own'. Myoclonus and dystonia also occur. Dementia is common in the later stages and there is considerable overlap with frontotemporal dementia. The pathology is focused in the frontal and parietal cortices as well as the substantia nigra, basal ganglia, and thalamus.

Treatable causes of dementia

Normal-pressure hydrocephalus

Normal-pressure hydrocephalus has a classic triad of presenting features: cognitive impairment, gait disturbance, and incontinence. The cognitive features are typically those of a subcortical dementia with frontal features and psychomotor slowing. The gait disorder is a dyspraxia and may show the pathognomonic feature of 'being stuck to the floor', although there is an absence of signs when the patient is examined in the supine position. The condition may be secondary to a prior disturbance of cerebrospinal fluid flow (resulting from, for example, a head injury, meningitis, or a subarachnoid haemorrhage), but often no cause is found in the elderly. Neuroimaging shows ventricular enlargement disproportionate to the degree of cortical atrophy. The presence of periventricular lesions can make the distinction from vascular dementia difficult. The investigation and management of these patients should be undertaken by the neurosurgeons, the definitive treatment being ventricular shunting. If treated early the prognosis is good.

Chronic subdural haematomas

This treatable cause of dementia is caused by head trauma. It is common in individuals at risk of recurrent head injuries, such as the elderly, alcoholics, and people with epilepsy. Risk is also increased by coagulation disorders, either pathological or iatrogenic. The clinical features are of a subacute dementia with symptoms of raised intracranial pressure and fluctuating cognitive performance and focal neurological signs. Diagnosis is confirmed by neuroimaging, the peripheral mass lesions may be of varying signal density on CT, depending on the age of the lesion. If the lesions are isodense with the brain tissue, the diagnosis can be easily overlooked. Treatment is by neurosurgical evacuation, except in clinically insignificant collections. Although the outcome is good, about 10 to 40 per cent of patients have a recurrence that may require a further drainage.

Benign tumours

Subfrontal meningiomas are the classic tumours that present with features of a frontal dementia. The onset is usually insidious with personality changes and other frontal features. Besides the neuropsychological abnormalities there may be anosmia or unilateral visual failure and optic atrophy. Other relatively benign midline tumours occasionally present with hydrocephalus and cognitive impairment secondarily to this (for example, colloid cysts of the third ventricle and non-secretory pituitary tumours).

Metabolic and endocrine disorders

Metabolic derangements can give rise to acute-onset cognitive impairments, but the features are invariably those of a delirium rather than a dementia. Chronic hypocalcaemia and recurrent hypoglycaemia can result in a dementia often accompanied by ataxia and involuntary movements. Endocrine disorders can more frequently present with a dementia syndrome, with or without psychiatric features (for example, hypothyroidism, Addison's disease, and hypopituitism). The prominent complaints common to most disorders are mental slowing, apathy, and poor memory. Cushing's disease can present with psychiatric features, although a dementia syndrome is rarer. Although not strictly an endocrine disorder, Hashimoto's encephalopathy is a recently recognized cause of chronic delirium or dementia, often accompanied by seizures and fluctuating focal neurological signs. The diagnosis is made by finding extremely high levels of anti-thyroid antibodies despite a euthyroid state. Patients respond well to high-dose steroid therapy.

Deficiency states

Vitamin B12 deficiency can cause the classic picture of subacute combined degeneration of the spinal cord and a dementia. The dementia can be variable in severity and it is unusual to present without some features of peripheral neurological disease, at least, diminished vibration sense in the lower limbs and/or sensory ataxia. Reflexes can be increased, decreased, or mixed. Although most patients have a macrocytic anaemia, neurological manifestations can occasionally occur in the absence of haematological features. Severe thiamin (vitamin B1) deficiency results in the Wernicke–Korsakoff syndrome, with delirium, ataxia, and ophthaloplegia. The commonest causes are alcoholism and recurrent prolonged vomiting, such as hyperemesis gravidarum. If not promptly treated a chronic amnesic syndrome can occur.

Infections

Neurosyphilis, once a common cause of dementia, is now rare. The associated neurological features include pupillary abnormalities, optic atrophy, ataxia, and pyramidal signs. The diagnosis is confirmed with serology and examination of cerebrospinal fluid. Treatment with penicillin can result in some improvement. Those at increased risk are people inadequately treated for syphilis and those infected with the human immunodeficiency virus (**HIV**). HIV infection is an increasingly common cause of dementia in some parts of the world. The encephalopathy (AIDS–dementia complex) is characterized by psychomotor slowing, personality change, and other features of a subcortical dementia. Examination of the cerebrospinal fluid can show a pleocytosis and increased protein and oligoclonal bands. White-matter changes are visible on neuroimaging. Cognitive changes in patients with HIV may also be due to opportunistic infections such as cerebral toxoplasmosis and cryptococcal meningitis and progressive multifocal leucoencephalopathy, which all require specific treatment.

Further reading

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders (DSM-IV)*. Washington DC.

Bak TH, Hodges JR (1998). The neuropsychology of progressive supranuclear palsy. *Neurocase* **4**, 89–94.

Berrios GE, Markova IS, Girala N (2000). Functional memory complaints: hypochondria and disorganisation. In: Berrios GE, Hodges JR, eds. *Memory disorders in psychiatric practice*, pp 384–99. Cambridge University Press, Cambridge.

Braak H, Braak E (1991). Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica (Berlin)* **82,** 239–59.

DeLeon M, *et al.* (1997). Frequency of hippocampal formation atrophy in normal aging and Alzheimer's disease. *Neurobiology of Aging* **18,** 1–11.

Galton CJ, Hodges JR (1999). The spectrum of dementia and its treatment. *Journal of the Royal College of Physicians London* **33,** 234–9.

Gauthier S (1999). *Clinical diagnosis and management of Alzheimer's disease*. Martin Dunitz, London.

Goedert M, Spillantini MG, Davies SW (1998). Filamentous nerve cell inclusions in neurodegenerative diseases. *Current Opinion in Neurobiology* **8,** 619–32.

Greene JDW, Hodges JR (2000). The dementias. In: Berrios GE, Hodges JR, eds. *Memory disorders in psychiatric practice*, pp 122–63. Cambridge University Press, Cambridge.

Gregory CA, Hodges JR (1996). Frontotemporal dementia: use of consensus criteria and prevalence of psychiatric features. *Neuropsychiatry, Neuropsychology, and Behavioural Neurology* **9,** 145–53.

Harvey J, *et al.* (1998). Genetic dissection of Alzheimer's disease and related dementias: amyloid and its relationship to tau. *Nature Neuroscience* **1,** 355–8.

Harvey RJ (2001). Epidemiology of pre-senile dementia. In: Hodges JR, ed. *Early onset dementia*, pp. 1–23. Cambridge University Press, Cambridge.

Hodges JR, Patterson K (1995). Is semantic memory consistently impaired early in the course of Alzheimer's disease? Neuroanatomical and diagnostic implications. *Neuropsychologia* **33,** 441–59.

Hodges JR, *et al.* (1992). Semantic dementia: progressive fluent aphasia with temporal lobe atrophy. *Brain* **115,** 1783–806.

Hodges JR, *et al.* (1999). The differentiation of semantic dementia and frontal lobe dementia (temporal and frontal variants of fronto-temporal dementia) from early Alzheimer's disease: a comparative neuropsychological study. *Neuropsychology* **13,** 31–40.

Jellinger K, *et al.* (1990). Clinicopathological analysis of dementia disorders in the elderly. *Journal of Neurological Sciences* **95,** 239–58.

Kalaria RN, Ballard C (1999). Overlap between pathology of Alzheimer's disease and vascular dementia. *Alzheimer's disease and Associated disorders* **13,** S115–S123.

Linn RT, *et al.* (1995). The 'preclinical phase' of probable Alzheimer's Disease. *Archives of Neurology* **52,** 485–90.

McKeith IG, *et al.* (1996). Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB): report of the consortium on DLB International Workshop. *Neurology* **47,** 1113–24.

McKhann G, *et al.* (1984). Clinical diagnosis of Alzheimer's disease: report of the NINDS-ADRDA Work Group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology* **34,** 939–44.

Mesulam MM (1982). Slowly progressive aphasia without generalized dementia. *Annals of Neurology* **24,** 17–22.

Neary D, *et al.* (1998). Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. *Neurology* **51,** 1546–54.

Rahman S, *et al.* (1999). Specific cognitive deficits in early frontal variant frontotemporal dementia. *Brain* **122,** 1469–93.

Reisberg B, *et al.* (1997). Diagnosis of Alzheimer's disease. Report of an International Psychogeriatric Association Special Meeting Work Group Under the Cosponsorship of Alzheimer's Disease International, the European Federation of Neurological Societies, the World Health Organisation, and the World Psychiatric Association. *International Psychogeriatrics* **9,** S11–S38.

Rockwood K, *et al.* (1999). Subtypes of vascular dementia. *Alzheimer's disease and Associated disorders* **13,** S59–S65.

Roman GC, *et al.* (1993). Vascular dementia: diagnostic criteria for research studies. Report of the NINDS-AIREN International Workshop. *Neurology* **43,** 250–60.

Sano M, *et al.* (1997). A controlled trial of selegiline, alpha-tocopherol, or both as treatment for Alzheimer's disease. *New England Journal of Medicine* **336,** 1216–22.

Snowden JS, Neary D, Mann DMA (1996). *Fronto-temporal lobar degeneration: fronto-temporal dementia, progressive aphasia, semantic dementia*. Churchill Livingstone, Hong Kong.

Welsh K, *et al.* (1991). Detection of abnormal memory decline in mild cases of Alzheimer's disease using CERAD neuropsychological measures. *Archives of Neurology* **48,** 278–81.

24.13.9Human prion diseases

R. G. Will

Introduction

Prion diseases, also known as transmissible spongiform encephalopathies, are fatal disorders of the central nervous system affecting both animals and humans (Table 1). The clinical features and patterns of occurrence of these diseases vary, but they are linked by a number of characteristics including experimental and natural transmissibility, shared neuropathological features, prolonged incubation periods measured in years, and the deposition of prion protein, which may be the causal agent, in the brain of the host. Prion diseases have become the subject of intense scientific and public interest because of the likelihood that they are caused by a new disease mechanism and because of the implications for public health following the identification of a new human prion disease, variant Creutzfeldt–Jakob disease, and the accumulating evidence that it is caused by the transmission of the cattle prion disease, bovine spongiform encephalopathy, to humans.

The causal agent

Scrapie was first transmitted experimentally from sheep to sheep in 1936 and to laboratory mice in 1961, but laboratory transmission of human prion diseases was not achieved until 1966 (kuru) and 1968 (Creutzfeldt–Jakob disease). The seminal discovery that apparently neurodegenerative diseases were transmissible stimulated extensive research into the nature of the infectious agent. No bacterium or virus has been isolated in these diseases and there is no immunological response to infection. This is of central importance as there is, as yet, no serological test to identify the presence of infection during the incubation period of any prion disease. The transmissible agent is remarkably resistant to inactivation procedures, including those that disrupt nucleic acids. Prusiner proposed in 1982 that the protein deposited in the central nervous system in these diseases was itself the causal agent. Purified infectious fractions of brain contain prion protein (for proteinacious infectious particle) which is a major, and perhaps the only, component of the infectious agent. This membrane-associated glycoprotein is present in all mammalian species. The normal function of prion protein is unknown. In prion diseases a post-translationally modified form of the protein, partially resistant to protease digestion, is deposited in the brain and is associated with neuronal dysfunction and death.

There is a range of experimental evidence supporting the hypothesis that the disease-associated form of prion protein is the causal agent in prion diseases, most notably a series of elegant studies in transgenic rodents. Cellular expression of prion protein is necessary for the development of the neuropathological changes and the disease. Hereditary forms of human prion disease are associated with, and perhaps caused by, mutations of the prion protein gene. However, the occurrence of multiple strains of the infectious agent and the stability of the transmission characteristics of the bovine spongiform encephalopathy agent in the laboratory following cross-species transmission are not readily explained by the prion theory. The importance of prion protein as a determinant of disease expression has become increasingly clear in human prion disease. The phenotype of different clinicopathological subtypes of Creutzfeldt–Jakob disease is related to the deposition in the brain of different molecular isotypes of prion protein, probably reflecting distinct tertiary protein structures, despite the identical amino acid sequences of the normal and disease-associated forms of prion protein.

In experimental transmission of prion diseases there are a number of key determinants of the efficiency of transmission, as judged by the incubation periods in recipient animals and the proportion of these animals that develop disease. The route of inoculation influences these variables. The intracerebral route is the most efficient. Intravenous, intraperitoneal, and oral routes are decreasingly efficient. The incubation period is inversely related to the infective dose, while the strain of the infectious agent influences both the incubation period and whether recipient animals develop disease. In some transmission studies, for example transmission of bovine spongiform encephalopathy to hamsters or transmission of scrapie to chimpanzees, recipient animals do not develop disease even after intracerebral inoculation of high levels of infectivity. Within-species transmission is more efficient than cross-species transmission and this 'species barrier' to transmission is influenced by characterisitics of both the host and the infective agent. The relative homology of amino acid sequences of prion proteins between species is not the only determinant of the species barrier. The relative efficiency of transmission between species cannot be predicted.

After ingestion, the agent replicates in the lymphoreticular system, including the spleen and lymph nodes, before entering the thoracic spinal cord or brainstem, probably via the autonomic nervous system, and then spreading caudally to the brain. Moderate levels of infectivity plateau in the lymphoreticular system and are not associated with organ dysfunction, while in the spinal cord and brain high levels of infectivity develop, for example 10^{12} infectious units per gram of brain in one model of hamster scrapie, leading to neuronal death and clinical disease. In some experimental and natural prion diseases infectivity in the lymphoreticular system can be detected at about one-third of the total incubation period by inoculation of tissues of the lymphoreticular system, such as spleen, into recipient animals. The implication is that, in the absence of an *in vivo* serological test for the presence of infectivity, animals or humans incubating a prion disease may harbour significant infectivity in some organs or tissues but cannot be identified as being infected. This has important implications for the control and public health implications of prion diseases.

Human prion diseases

Human prion diseases may be classified as sporadic, inherited, or acquired (Table 2).

Sporadic Creutzfeldt–Jakob disease

Sporadic Creutzfeldt–Jakob disease is a rare disease, with an annual incidence of about 1 case per million population. The disease occurs worldwide and the cause is unknown, with no convincing evidence of an environmental source of infection and in particular no proven link with the animal prion diseases. The regional clusters of cases identified in some countries are unusual and may reflect the chance aggregation of a rare phenomenon. Overall the geographical and temporal distribution of cases of sporadic Creutzfeldt–Jakob disease appear to be random and case control studies have demonstrated no consistent risk factors for the development of disease, with no good evidence of an increased risk through occupation, dietary factors, or animal contact. The current favoured hypothesis is that sporadic Creutzfeldt–Jakob disease is caused by a spontaneous mutation of prion protein to the abnormal form, which acts as a template for protein self-replication and eventual disease.

Clinically sporadic Creutzfeldt–Jakob disease presents with a rapidly progressive dementia associated with a range of neurological signs, most commonly myoclonus of the limbs, cerebellar ataxia, and rigidity. Less common features include dysphasia, pyramidal or extrapyramidal signs, primitive reflexes, cortical blindness, and lower motor neurone signs. Despite the predominantly cortical neuropathology epilepsy is rare. The rapidity of the progression of neurological deficits and cognitive decline is distinct from most other causes of dementia and the mean survival is only about 4 months from clinical onset, although in about 10 per cent of cases the illness is more prolonged and a small minority of patients survive for 2 years or more (Fig. 1). Terminally there is often a state of akinetic mutism.

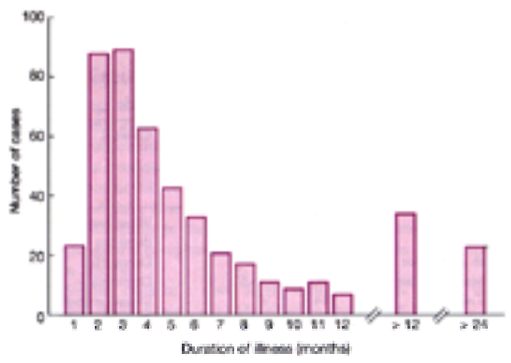

Fig. 1 Sporadic Creutzfeldt–Jakob disease—survival.

Although the clinical presentation in sporadic Creutzfeldt–Jakob disease is relatively stereotyped, a minority of cases present atypically, for example acutely mimicking stroke, with cortical blindness, or with an initially pure cerebellar syndrome.

The neuropathological characteristics of sporadic Creutzfeldt–Jakob disease include spongiform change, neuronal loss, and astrocytosis in the cerebral and cerebellar cortex, in accordance with the neurological signs seen in life (Plate 1). Neuropathological changes are widespread and deposition of prion protein can be detected with immunocytochemical techniques. In about 10 per cent of cases there are cortical deposits of prion protein in the form of amyloid plaques. There is heterogeneity in the distribution and morphology of the neuropathological changes, which correlate in part with the clinical phenotype and with two isotypes of prion protein which can be distinguished on Western blot of brain tissue (Fig. 2).

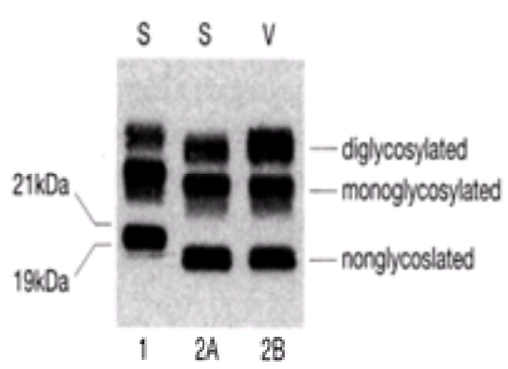

Fig. 2 Western blot of brain tissue showing type 1 and type 2 prion proteins. Lane 1 is sporadic Creutzfeldt–Jakob disease, lane 2A is sporadic Creutzfeldt–Jakob disease and lane 2B is variant Creutzfeldt–Jakob disease. Type 2A and type 2B have the same mobility but are differentiated by the relative proportions of different glycoforms—in variant Creutzfeldt–Jakob disease there is an excess of the diglycosylated form.

Sporadic Creutzfeldt–Jakob disease is mainly a disease of late middle age (Fig. 3) with a mean age at death of 66 years. In most systematic studies males and females are affected with equal frequency.

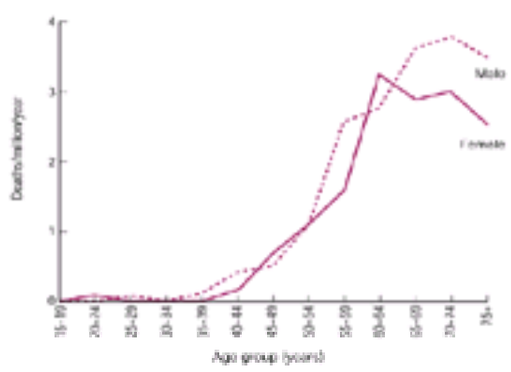

Fig. 3 Age- and sex-specific mortality rates from sporadic Creutzfeldt–Jakob disease in the United Kingdom, 1995 to 2000.

The human prion protein gene is situated on chromosome 20 and contains a polymorphic region at codon 129, which expresses either methionine or valine. Methionine homozygosity (MM) at codon 129 increases susceptibility to sporadic Creutzfeldt–Jakob disease (Table 3). The genotype distribution in sporadic cases is MM 80 per cent, valine homozygous (VV) 15 per cent, and heterozygous (MV) 5 per cent in contrast to the genotype distribution in the normal Caucasian population. There is accumulating evidence that the disease phenotype in sporadic Creutzfeldt–Jakob disease, as well as susceptibility, is influenced by an interplay between the codon 129 genotype and the prion protein isotype. The classical form of sporadic Creutzfeldt–Jakob disease, representing the great majority of cases, is associated with type 1 prion protein and an MM genotype, while alternative combinations of prion protein isotype and codon 129 genotype are often associated with atypical phenotypes.

Hereditary prion diseases

Familial clusters of Creutzfeldt–Jakob disease account for about 10 per cent of all cases and within pedigrees there is a dominant pattern of inheritance. The paradox of a transmissible disease that is also inherited was clarified by the identification of a mutation at codon 102 of the prion protein gene in two families affected by Gerstmann–Straussler–Scheinker syndrome (GSS), a condition known to be a human prion disease on the basis of the neuropathology and laboratory transmissibility. More than 20 prion protein gene mutations, including point and insertional mutations, have now been identified in familial Creutzfeldt–Jakob disease or Gerstmann–Straussler–Scheinker syndrome (Table 4), and all cases of hereditary human prion disease to date have been found to have a mutation of the prion protein gene. Fatal familial insomnia was first identified as a prion disease following the identification of a mutation at codon 178 of the prion protein gene in affected family members and it was only later that transmission in the laboratory confirmed the status of fatal familial insomnia as a prion disease. The current hypothesis is that mutations of the prion protein gene lead to an instabililty in the structure of prion protein and an increased chance of a spontaneous transformation of prion protein to the abnormal self-replicating disease-associated form. With the exception of the prion disease associated with a mutation at codon 200 of the prion protein gene, all hereditary human prion diseases are fully penetrant.

The incidence of Creutzfeldt–Jakob disease in localized areas of Slovakia and in Libyan-born Israelis was discovered many years ago to be 60 to 100 times greater than expected. Possible explanations for these clusters included excessive dietary exposure to sheep scrapie and a high coefficient of inbreeding. Following the identification of the mutations of prion protein in human disease, genetic studies have shown that in both clusters there is a high population frequency of mutations at codon 200 of the prion protein gene and that the excess of cases of Creutzfeldt–Jakob disease is due to an excess of familial cases, with an expected background incidence of sporadic cases.

Overall the age at death in hereditary prion diseases is about 5 to 10 years earlier than in sporadic Creutzfeldt–Jakob disease, but the duration of clinical illness is often more prolonged and the clinical features vary with the underlying mutation. With some mutations, notably the codon 200 mutation, the clinical course is similar to sporadic Creutzfeldt–Jakob disease, but cases of hereditary prion disease may present with ataxia, for example Gerstmann–Straussler–Scheinker syndrome, or with a highly atypical phenotype such as fatal familial insomnia in which the early clinical features include dysautonomia and insomnia. There may be variation in the clinical phenotype both within and between families even if these are associated with the same underlying mutation in the prion protein gene.

Neuropathologically there is great heterogeneity in hereditary prion diseases, and as with the clinical phenotype there is an overall relationship between the neuropathological features and the specific prion protein gene mutation, although there can be great variation within and between pedigrees (Plate 2). The neuropathology can be similar to sporadic Creutzfeldt–Jakob disease but in a significant proportion of hereditary prion diseases there is amyloid plaque formation and in fatal familial insomnia gliosis and neuronal loss may be restricted to the thalamus.

In some forms of hereditary prion disease the codon 129 genotype may influence clinical characteristics, including age at death, and the neuropathology. Variation at this locus has a profound effect on the disease phenotype in association with mutations at codon 178 of the prion protein gene. Cases with a codon 178 mutation and a methionine at codon 129 of the prion protein gene develop fatal familial insomnia, whereas with valine at codon 129 the phenotype is similar to sporadic Creutzfeldt–Jakob disease.

Iatrogenic Creutzfeldt–Jakob disease

Creutzfeldt–Jakob disease has been transmitted accidentally in the course of medical treatment by neurosurgical instruments, corneal grafts, cadaveric dura mater grafts, and human pituitary derived hormones (Table 5). The presumption is that infection from individuals with Creutzfeldt–Jakob disease was transmitted to uninfected individuals via these procedures and there is strong circumstantial evidence that this has occurred. In two of the transmissions by corneal grafts the donors died of sporadic Creutzfeldt–Jakob disease and in the neurosurgical transmissions there was a clear temporal link between surgical procedures on Creutzfeldt–Jakob disease cases and patients operated on using the same instruments who subsequently developed Creutzfeldt–Jakob disease. It is presumed that some human dura mater grafts and human pituitary hormones came from individuals suffering from Creutzfeldt–Jakob disease and there may have been cross-contamination in the production process leading to dissemination of infection. Infection via human pituitary growth hormone has been demonstrated in laboratory transmission studies. All cases of iatrogenic transmission of Creutzfeldt–Jakob disease have involved surgical instruments, grafts, or hormonal products potentially contaminated by central nervous system tissue and, by implication, high levels of infectivity.

There is a distinction between the clinical features in iatrogenic Creutzfeldt–Jakob disease which depends on the route of inoculation. In exposures in or adjacent to the nervous system (neurosurgical instruments, dura mater grafts, and corneal transplants) the majority of cases present with a progressive dementia similar to sporadic Creutzfeldt–Jakob disease. With a peripheral route of exposure to infection (pituitary hormones) there is a progressive cerebellar ataxia and cognitive impairment develops late in the clinical course, if at all.

The incubation period also varies according to the route of exposure to infection. With central exposure the mean incubation period ranges from about 18 months, similar to the incubation periods in primates after experimental intracerebral inoculation, to 6 years with dura mater grafts. With a peripheral route of exposure the mean incubation period is about 12 years, but may extend to over 30 years, which is similar to the extended incubation periods in kuru, a human prion disease also caused by a peripheral route of exposure to infection.

Homozygosity at codon 129 of the prion protein gene, either MM or VV, increases susceptibility to human growth hormone related Creutzfeldt–Jakob disease and heterozygosity may lead to a more prolonged incubation period. In dura mater related Creutzfeldt–Jakob disease 81 per cent of cases have an MM genotype, similar to the proportion of sporadic cases with this genotype, but the codon 129 genotype does not influence the incubation period.

Reducing the risks of iatrogenic transmission of Creutzfeldt–Jakob disease

Measures to reduce the risk of iatrogenic transmission of Creutzfeldt–Jakob disease have been introduced in many countries. There are strict selection criteria for obtaining corneal grafts, recombinant growth hormone replaced human growth hormone in 1985, and human dura mater grafts have not been licensed in the United Kingdom since the early 1990s. There is no evidence that Creutzfeldt–Jakob disease has been transmitted iatrogenically through non-central nervous system tissues such as blood, blood products, or organ transplantation, but continued vigilance is necessary as many of the mechanisms of iatrogenic transmission of Creutzfeldt–Jakob disease were not predicted. The possibility of secondary transmission of variant Creutzfeldt–Jakob disease through blood transfusion, the use of fractionated blood derivatives, organ transplantation, or contaminated surgical instruments is a matter of continuing concern.

Variant Creutzfeldt–Jakob disease

Bovine spongiform encephalopathy was identified in 1986 as a novel prion disease in cattle in the United Kingdom, and is thought to have been caused by feeding cattle material contaminated with sheep scrapie or, perhaps, to have been a previously unrecognized endemic prion disease of cattle. Bovine to bovine recycling of infection through cattle feed amplified the epidemic and there have now been over 180 000 cases of bovine spongiform encephalopathy in the United Kingdom. Small numbers of cases of bovine spongiform encephalopathy have been identified in other countries, mainly in Europe.

In 1996 ten cases of a novel form of human prion disease, variant Creutzfeldt–Jakob disease, were identified in the United Kingdom and a causal link with bovine spongiform encephalopathy was proposed as this was a new disease occurring only in the United Kingdom, the country with the greatest potential human exposure to bovine spongiform encephalopathy. Up to January 2002 there have been 114 cases of variant Creutzfeldt–Jakob disease in the United Kingdom, five in France, and one in the Republic of Ireland. The mean age at death in variant Creutzfeldt–Jakob disease is 29 years (range 15 to 74 years, Fig. 4) contrasting with a mean age at death in sporadic Creutzfeldt–Jakob disease of 66 years. The hypothesis that variant Creutzfeldt–Jakob disease is caused by the bovine spongiform encephalopathy agent has been supported by the consistent disease phenotype, and in particular the neuropathology which is distinct from other human prion diseases, the failure to identify similar cases in the past either in the United Kingdom or elsewhere, and laboratory transmission studies which have shown a remarkable similarity between the transmission characteristics of bovine spongiform encephalopathy and variant Creutzfeldt–Jakob disease in mice.

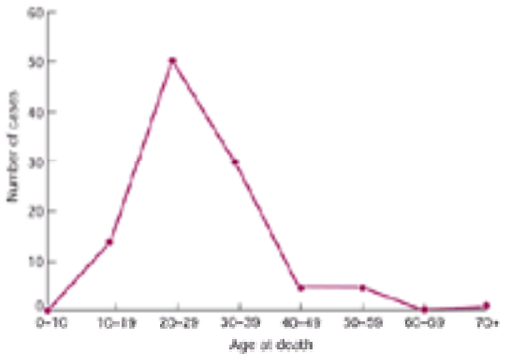

Fig. 4 Age distribution of cases of variant Creutzfeldt–Jakob disease in the United Kingdom (January 2002).

The clinical features of variant Creutzfeldt–Jakob disease are relatively distinct from other forms of human prion disease, including sporadic and iatrogenic Creutzfeldt–Jakob disease. Patients present with psychiatric symptoms, including depression, withdrawal, and anxiety, followed after a period of months by progressive ataxia, dementia, and choreiform or dystonic involuntary movements, which often evolve into myoclonus. The terminal stages are similar to sporadic Creutzfeldt–Jakob disease, but the overall duration of illness, mean 14 months, is significantly more prolonged. The distinctive neuropathological characteristic of variant Creutzfeldt–Jakob disease is the widespread deposition of deposits of prion protein with a halo of spongiform change, so-called florid plaques, throughout the

cerebral and cerebellar cortex, in addition to the spongiform change, neuronal loss, and gliosis seen in other human prion diseases.

Cases of variant Creutzfeldt–Jakob disease have been identified from throughout the United Kingdom and risk factors include residence in the United Kingdom and an MM genotype at codon 129 of the prion protein gene. All the United Kingdom cases and the case diagnosed in the Republic of Ireland had been resident in the United Kingdom during the 1980s to early 1990s when human exposure to bovine spongiform encephalopathy was likely to have been maximal. Three of the French cases had never visited the United Kingdom, implying that exposure to bovine spongiform encephalopathy must have occurred in France from indigenous bovine spongiform encephalopathy or export from the United Kingdom of cattle or food products. A case control study has not yet demonstrated any significant dietary risk factor in variant Creutzfeldt–Jakob disease, although the favoured hypothesis is that transmission of bovine spongiform encephalopathy to humans was through contamination of food, probably with tissues from the central nervous system such as brain or spinal cord which are known to contain high levels of infectivity in cattle affected with bovine spongiform encephalopathy. All tested cases of variant Creutzfeldt–Jakob disease to date have been MM homozygotes at codon 129 of the prion protein gene. This genotype is also present in about 80 per cent of cases of sporadic Creutzfeldt–Jakob disease and may represent a susceptibility factor for the development of variant Creutzfeldt–Jakob disease. Variation at this locus can, however, influence the incubation period and disease phenotype and it is possible that cases of human infection with bovine spongiform encephalopathy may yet be identified in individuals with a VV or MV genetic background.

The possible future number of cases of variant Creutzfeldt–Jakob disease is unknown, but there has been an increase in the annual number of deaths from variant Creutzfeldt–Jakob disease in the United Kingdom (Fig. 5) and statistical analyses, taking into account delays in referral and confirmation, have shown a significant increase in the temporal incidence of cases since 2000. Long-term predictions have estimated a total of 100 to over 136 000 cases of variant Creutzfeldt–Jakob disease in the United Kingdom. This wide range reflects the many uncertainties including the mean incubation period of bovine spongiform encephalopathy in humans, the level of the species barrier between bovines and humans, and the extent of human exposure to the bovine spongiform encephalopathy agent.

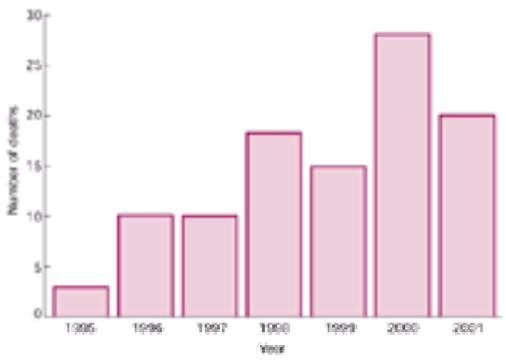

Fig. 5 Variant Creutzfeldt–Jakob disease—number of deaths per annum in the United Kingdom (January 2002).

Kuru

The transmissibility of human prion diseases was first demonstrated in 1966 with the transmission of a spongiform encephalopathy to chimpanzees 18 to 21 months after intracerebral inoculation of a brain extract from a patient who had died of kuru. This seminal experiment followed years of clinical, epidemiological, and anthropological research in the Fore region of Papua New Guinea where kuru was endemic. In the early 1960s kuru caused over half of all deaths in the affected population and there have been more than 3000 deaths from kuru in the at-risk population of 30 000 people.

The epidemiological characteristics of kuru are unusual with familial aggregation of cases and a high incidence of disease in women and children in the early years of the epidemic. Since 1960 there has been a decline in the incidence, particularly in women and children (Fig. 6), and there have been no cases in children born after 1959. After extensive investigation into a possible genetic or toxic origin, anthropological research established that kuru was transmitted in the course of ritual cannibalism. As a mark of respect, relatives consumed affected individuals and virtually all tissues were consumed, including the brain and viscera. Although men took part in these rituals, women and children are thought to have consumed the internal organs such as the brain which contained the highest levels of infectivity. It is also possible that there was transcutaneous transmission through rubbing of tissue on the skin. Detailed investigation of individual cannibalistic events has shown that a number of members of the same family, including those who came from different areas, developed kuru after attending a single cannibalistic rite. Ritual cannibalism ceased by 1960, explaining the subsequent decline in incidence of kuru, but there are still occasional cases with incubation periods exceeding 40 years. It is of interest that at the height of the epidemic many hundreds of women were affected by kuru during pregnancy and breastfed their children, but none of these children later developed kuru.

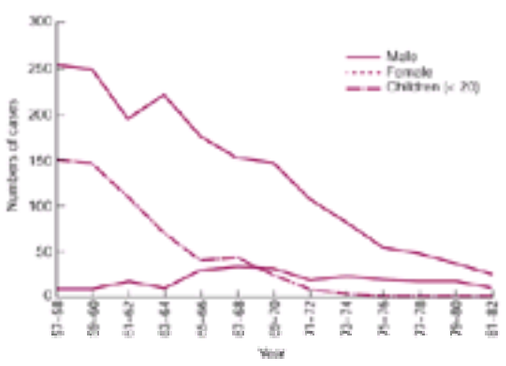

Fig. 6 Kuru—numbers of cases by age, sex, and time.

Clinically kuru presented with a cerebellar syndrome, initially truncal ataxia and titubation, followed by ataxia of gait and dysarthria. A prodromal phase of headache and limb pain was common and hypotonia was a prominent early feature. Involuntary movements such as myoclonus and rigidity of the limbs did not occur, in contrast to other forms of human prion disease. Terminally patients became immobile and communication was often impossible because of severe dysarthria. Dementia did not occur, and even in the terminal akinetic and mute state patients could obey simple commands. In children the clinical features were similar, but in the early stages there were often brainstem signs such as strabismus, nystagmus, and ptosis. The total duration of illness ranged from 12 to 18 months in adults and 3 to 12 months in children.

In kuru, neuropathological changes were most apparent in the cerebellum, consistent with the clinical features. Neuronal loss and intense cerebellar astrocytosis were uniform findings and about three-quarters of cases had amyloid plaque deposition, particularly in the granule cell layer of the cerebellum. The cerebral cortex showed mild spongiform change. The similarity of the neuropathology of kuru to scrapie was commented on by Hadlow in 1959, prompting the transmission studies which later demonstrated that kuru, like scrapie, was experimentally transmissible.

By using stored samples, analysis of the influence of the codon 129 polymorphism of the prion protein gene on susceptibility to kuru has shown that homozygosity, either MM or VV, increases susceptibility and that heterozygotes may have a more prolonged incubation period. The analysis of codon 129 genotype in kuru is complicated by the limited number of tested cases and the possible effect of the high mortality rate on the codon 129 distribution in a closed population.

The diagnosis of human prion diseases

Human prion diseases are rare, but the high public profile of Creutzfeldt–Jakob disease and variant Creutzfeldt–Jakob disease has resulted in an increase in the

number of cases in which the diagnosis of one of these diseases is suspected. Accurate diagnosis of any condition, including patients suffering from a human prion disease, is essential but the exclusion of a diagnosis is also important, particularly for a fatal and untreatable condition. Although symptomatic treatment, for example for involuntary movements, can be helpful in human prion diseases there is currently no available treatment that influences the clinical course nor any treatment to prevent the development of neurological disease after infection. An important objective is to improve diagnostic accuracy in human prion diseases and in particular to allow early diagnosis. In the absence of a test for the presence of the infectious agent, diagnosis depends on the recognition of the clinical characteristics of human prion diseases supported by a range of investigations, some of which have been developed in recent years. Diagnostic criteria for sporadic, iatrogenic, familial, and variant Creutzfeldt–Jakob disease have been formulated and validated (Table 6 and Table 7). In all human prion diseases a definite diagnosis can only be made by the examination of brain tissue, usually at post-mortem.

In the majority of cases of sporadic Creutzfeldt–Jakob disease the diagnosis is made in life because of the multifocal neurological deficits, the development of myoclonus, and in particular the rapidity in the progression of cognitive impairment. The clinical picture is distinct from more common forms of dementia. In forms of sporadic Creutzfeldt–Jakob disease with early focal neurological features such as a cerebellar syndrome the rapid evolution of other neurological deficits and dementia suggests the diagnosis of Creutzfeldt–Jakob disease. Diagnosis can be difficult in cases of sporadic Creutzfeldt–Jakob disease with atypical features such as long duration of illness, and in these cases investigations such as magnetic resonance imaging of the brain can be helpful. There is increasing evidence that cases of sporadic Creutzfeldt–Jakob disease may be atypical if there is an underlying MV or VV codon 129 prion protein genotype.

Hereditary prion diseases are often suspected because of a family history of a similar disorder, but in a significant proportion of cases of Creutzfeldt–Jakob disease associated with a prion protein gene mutation there is a family history of another neurodegenerative disorder or no relevant family history. The gradual clinical progression in many forms of hereditary human prion disease makes accurate diagnosis difficult and the diagnosis may only be recognized in life after prion protein gene analysis. Genetic testing should only be carried out with fully informed consent.

The diagnosis of iatrogenic Creutzfeldt–Jakob disease depends on the identification of a relevant risk factor, for example previous treatment with human growth hormone, and an assessment of the neurological presentation. Most patients with growth hormone related Creutzfeldt–Jakob disease present with a cerebellar syndrome, while after central iatrogenic exposure to infection the clinical picture is usually similar to that of sporadic Creutzfeldt–Jakob disease. The utility of specialist investigation in iatrogenic Creutzfeldt–Jakob disease is uncertain because of the rarity of these forms of Creutzfeldt–Jakob disease and limited information on investigations.

The clinical picture in the later stages of variant Creutzfeldt–Jakob disease is similar to that of sporadic Creutzfeldt–Jakob disease and, although the recognition of the diagnosis in the first cases of this new disease was difficult, the clinical phenotype is now well known and the diagnosis is usually apparent after neurological signs develop, often in young patients in an age group in which dementia is very unusual. Diagnosis in the early stages is, however, difficult as there is a period of many months in which the clinical picture is dominated by psychiatric symptoms, including depression, anxiety, and withdrawal. Clues to the possibility of variant Creutzfeldt–Jakob disease include cognitive impairment, subtle gait ataxia, and persistent painful sensory symptoms in combination with the psychiatric symptoms.

Investigations in human prion disease

Many of the investigations carried out in suspected cases of human prion disease do not show any specific disease related abnormality, but help to exclude other diagnoses, some potentially treatable. The interpretation of the results of investigations depends on the clinical picture because the sensitivity and specificity of surrogate markers for prion disease, such as 14-3-3 cerebrospinal fluid analysis (see below), depend on clearly defining the characteristics of the patients in which the test has been carried out.

Routine haematological and biochemical tests are usually normal. About a third of cases of sporadic or variant Creutzfeldt–Jakob disease may have minor abnormalities in liver function tests.

The electroencephalogram shows periodic triphasic complexes at about 1 per second in 60 to 70 per cent of cases of sporadic Creutzfeldt–Jakob disease (Fig. 7) and in some cases of iatrogenic Creutzfeldt–Jakob disease after central exposure to infection. These electroencephalogram changes are relatively specific, but similar appearances can be seen in hepatic encephalopathy, lithium or metrizamide toxicity, metabolic disturbance, and rarely in other forms of dementia such as Alzheimer's disease.

Fig. 7 Typical electroencephalogram in sporadic Creutzfeldt-Jakob disease.

There is no cerebrospinal fluid pleocytosis in any form of human prion disease, and cerebrospinal fluid protein is elevated in about a third of cases. Elevation of the 14-3-3 cerebrospinal fluid protein, a marker for neuronal damage, has a sensitivity and specificity of about 90 per cent in the diagnosis of sporadic Creutzfeldt–Jakob disease, but is less useful in the diagnosis of variant Creutzfeldt–Jakob disease.

A computed tomography scan of the brain is usually normal, but can show non-specific cerebral atrophy. Magnetic resonance imaging of the brain shows a high signal on T_2-weighted images in the caudate nucleus and putamen in about 70 per cent of cases of sporadic Creutzfeldt–Jakob disease (Fig. 8), but the sensitivity and specificity of these abnormalities has not been formally assessed. In variant Creutzfeldt–Jakob disease about 80 per cent of cases show a high signal on T_2-weighted images (and PD and FLAIR sequences) in the pulvinar region of the posterior thalamus (Fig. 9) and in the appropriate clinical context these abnormalities have a high sensitivity and specificity for the diagnosis of variant Creutzfeldt–Jakob disease. To date all cases of variant Creutzfeldt–Jakob disease classified as 'probable', a diagnosis requiring the abnormalities on magnetic resonance imaging, that have come to post-mortem examination have been confirmed as variant Creutzfeldt–Jakob disease.

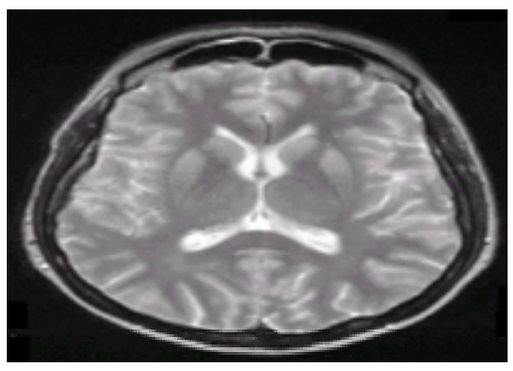

Fig. 8 MRI image of sporadic Creutzfeldt–Jakob disease showing high signal changes symmetrically in the caudate and putamen.

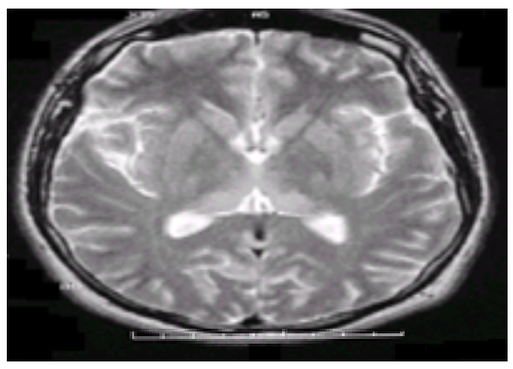

Fig. 9 MRI image of variant Creutzfeldt–Jakob disease showing high signal symmetrically in the posterior thalamus.

Brain biopsy can allow the confirmation of the diagnosis of a human prion disease in life, but this investigation has risks and is mainly carried out when there is a realistic possibility of an alternative diagnosis. Tonsil biopsy in variant Creutzfeldt–Jakob disease can also increase the likelihood of the diagnosis in life, but this procedure is also invasive and, although early diagnosis is important for the relatives of the patient and for clinicians, it does not benefit the patient.

Conclusion

There have been remarkable scientific advances in the understanding of prion diseases and it is hoped that this may lead to the identification of a diagnostic test in life for the presence of infection and to therapies to prevent the development of disease. Human prion diseases have attained a public notoriety disproportionate to the overall burden of disease caused by these rare conditions. However, the transmission of an animal prion disease, bovine spongiform encephalopathy, to humans has been a tragedy and the prolonged incubation periods characteristic of this group of diseases indicate that the eventual consequences of bovine spongiform encephalopathy for public health both in the United Kingdom and in other countries are unpredictable.

Further reading

Brown P *et al.* (2000). Iatrogenic Creutzfeldt–Jakob disease at the millennium. *Neurology* **55**, 1075–81.

Collinge J, Palmer MS (1997). Human prion diseases. In: Collinge J, Palmer MS, eds. *Prion diseases*, pp 18–56. Oxford University Press, New York.

DeArmond SJ, Ironside JW (1999). Neuropathology of prion diseases. In: Prusiner SB, ed. *Prion biology and diseases*, pp 585–652. Cold Spring Harbor Laboratory Press, New York.

Donnelly CA, Ferguson NM (1999). Predictions and scenario analysis for vCJD. In: Donnelly CA, Ferguson NM, eds. *Statistical aspects of BSE and vCJD*, pp 163–94. Chapman and Hall, London.

Gajdusek DC (1990). Subacute spongiform encephalopathies: transmissible cerebral amyloidoses caused by unconventional viruses. In: Fields BN, Knipe DM, eds. *Fields virology*, pp 2289–324. Raven Press, New York.

Gambetti P *et al.* (1999). Inherited prion diseases. In: Prusiner SB, ed. *Prion biology and diseases*, pp 509–83. Cold Spring Harbor Laboratory Press, New York.

Parchi P *et al.* (1999). Classification of sporadic Creutzfeldt–Jakob disease based on molecular and phenotypic analysis of 300 subjects. *Annals of Neurology* **46**, 224–33.

Prusiner SB (1994). Prion diseases of humans and animals. *Journal of the Royal College of Physicians of London* **28**(suppl.), 1–30.

Prusiner SB (1999). Development of the prion concept. In: Prusiner SB, ed. *Prion biology and diseases*, pp 67–112. Cold Spring Harbor Laboratory Press, New York.

Will RG *et al.* (1999). Infectious and sporadic prion diseases. In: Prusiner SB, ed. *Prion biology and diseases*, pp 465–507. Cold Spring Harbor Laboratory Press, New York.

24.13.10 Parkinsonism and other extrapyramidal diseases

Donald B. Calne

The concept of extrapyramidal disease

The concept of extrapyramidal disease arose at the start of the twentieth century. Nineteenth-century anatomists, physiologists, and neurologists established the importance of the pyramidal pathway, but they later recognized that pathology outside this system also led to major disturbances of voluntary movement. It was concluded that alternative pathways within the central nervous system contribute to the control of movement and the term 'extrapyramidal system' was coined. Pathology in this extrapyramidal system was associated with involuntary movements, which became the hallmark of what were called 'extrapyramidal diseases'. Over the course of the twentieth century, however, the notion of a pyramidal pathway was criticized. In consequence, the idea of disease involving an extrapyramidal pathway was also discredited, so the term 'movement disorders' became fashionable, and is now widely used.

In spite of these changes in nomenclature, there is general agreement that certain structures, deep in the brain, play an important part in the planning and execution of voluntary movement. In particular, it has been shown through clinicopathological correlation, that the basal ganglia are linked to motor control. The basal ganglia comprise the caudate nucleus, the putamen, and the globus pallidus. In addition, certain nuclei feeding into this system, such as the substantia nigra, are usually included. Myoclonus and ataxia will not be discussed in this review, because their pathology does not involve the basal ganglia.

Parkinson's disease

Definition

The terms 'paralysis agitans', 'the shaking palsy', 'Parkinson's disease', 'idiopathic parkinsonism', and 'idiopathic Parkinson's disease' have all been used, at various times, to describe the same entity. The definition of this entity is, however, fraught with difficulties. Initially, it was thought that the presence of Lewy bodies in the substantia nigra was the pathological *sine qua non* for the distinction of Parkinson's disease from other disorders. However, it has recently been recognized that Lewy bodies are quite non-specific, occurring in such diverse entities as Hallervorden–Spatz disease, certain forms of dementia, and subacute sclerosing panencephalitis. Furthermore, it has also been recognized that some patients who looked in every way as if they had Parkinson's disease during life, proved to have degeneration of cells in the substantia nigra with tangles rather than Lewy bodies.

There have been different approaches to resolving the dilemma created by this difficulty with definition. Perhaps the most satisfactory solution, for the present, is the use of the term 'Parkinson's syndrome' until we understand the aetiology, or, more likely, the many aetiologies of 'Parkinson's disease'. But the old terminology persists in most of the current literature, so the term 'Parkinson's disease' will be used here for the idiopathic syndrome of parkinsonism.

Aetiology

We already know that several distinct aetiologies are responsible for Parkinson's disease. For example, genetic factors have been identified, including mutations of the a-synuclein gene, the parkin gene, and a gene located on chromosome 2. Genetic causes of parkinsonism tend to produce symptoms at a younger age than sporadic Parkinson's disease. Most cases are sporadic, and are likely to have environmental aetiologies.

We are left with the unproven probability that there exist several different causes for Parkinson's disease, just as there exist several different causes for meningitis, or peripheral neuropathy.

Epidemiology

The prevalence of Parkinson's disease is around 200 to 300 per 100 000. The incidence is about 20 per 100 000 per year. These figures apply to the United Kingdom, Canada, and the United States. Studies on the geographical distribution indicate a low rate in Africa, China, and Japan. Prevalence rises with age, and Parkinson's disease is slightly commoner in males. The prevalence also increases with distance from the equator. A recent Canadian study suggests that Parkinson's disease is commoner in healthcare workers, teachers, and those who share cramped living conditions for sleeping (for example, miners and loggers working in camps)—while those who live more secluded lives have a low prevalence. These finding suggest the possibility that a viral infection might contribute to causation in some cases.

Pathology

The classical hallmark of Parkinson's disease is the relatively selective degeneration of the dopaminergic nigrostriatal pathway. In particular, the ventral tier of the zona compacta of the substantia nigra undergoes degeneration, and there is a reduction of dopamine in the striatum with a characteristic spatial distribution. The loss is predominantly posterior in the putamen, extending, to a lesser extent, forward into the anterior putamen and the caudate. In most cases of Parkinson's disease, Lewy bodies can be found in the substantia nigra.

Clinical features

The principal clinical features of Parkinson's disease are resting tremor, rigidity, and bradykinesia. In addition, postural reflexes are impaired, though this is a rather

non-specific finding. These clinical features can be regarded as the inclusion criteria for a diagnosis. There are also exclusion criteria, for Parkinson's disease should not be considered if pyramidal signs, cerebellar signs, gaze palsies, or autonomic deficits are prominent. As Parkinson's disease progresses, many patients develop dementia.

It is usual to find asymmetry in the clinical presentation of patients with Parkinson's disease, and they almost always respond to dopaminomimetic treatment if there are no dose-limiting side-effects.

Natural history

Parkinson's disease starts several years before patients first notice symptoms. The full extent of this 'preclinical phase' has not been defined. However, we do know that the loss of dopaminergic nigral cells is slow, and that some 50 per cent of the nigrostriatal pathway is lost when symptoms first become apparent. Pathological observations, positron emission tomography, and clinical examination all indicate that the rate of progression of neuronal loss decreases over the course of Parkinson's disease.

The mean duration of Parkinson's disease depends on the time of onset. In general, progression is slower when Parkinson's disease presents in younger patients; it may run a course of over 35 years in these circumstances. In older patients, although the duration is shorter, life expectation is still only slightly reduced. Primarily, the patient is faced with a reduction in the quality, rather than the quantity, of life. Nevertheless, the late stage of the illness can entail cruelly protracted and profound disability.

Differential diagnosis

The most common disorder that can be confused with Parkinson's disease is essential tremor. While the classical description of a resting tremor has been employed to characterize Parkinson's disease, the early stage of the illness is often accompanied by a rather low-amplitude postural tremor, similar to that seen in essential tremor. This phase in the evolution of Parkinson's disease only lasts for a year or two. In contrast, postural tremor is present for many years in essential tremor, and it is only in advanced stages of essential tremor that a resting tremor appears.

Drug-induced parkinsonism is another important consideration in the differential diagnosis of Parkinson's disease. All drugs that block dopamine receptors or deplete dopamine in the brain are capable of producing a parkinsonian syndrome. While the major tranquillizing drugs used to treat psychosis are most frequently responsible, other drugs such as metoclopramide, used to control nausea and modify gastrointestinal motility, are now an important cause of parkinsonism.

The most common serious neurodegenerative disorder to be confused with Parkinson's disease is progressive supranuclear palsy (Steele–Richardson syndrome). It may take a year or two before the importance of gaze palsies, nuchal rigidity, dementia, and impaired balance all lead to the realization that one is not dealing with classical Parkinson's disease.

Multiple system atrophy is also easily mistaken for Parkinson's disease. Again, the clinician may have to wait until the autonomic deficits or cerebellar features of multiple system atrophy become prominent. Drugs used to treat Parkinson's disease tend to induce orthostatic hypotension, compounding the difficulty in differential diagnosis. The task of determining whether orthostatic hypotension is part of the underlying illness or a consequence of treatment often poses a challenge.

Treatment

Medical treatment

There is no convincing evidence that any current treatment slows down the course of Parkinson's disease, so all treatment is directed at reducing symptoms and improving quality of life. The cornerstone of treatment is dopaminomimetic therapy, to overcome the impact of reduced dopamine in the striatum. The standard therapy for many years has been levodopa, combined with a peripheral decarboxylase inhibitor. Evidence is now accumulating that for initial treatment in the younger parkinsonian patient, the new artificial dopamine agonists give the best initial results. A recent 5-year study has shown advantages with ropinirole—because of a substantially reduced prevalence of dyskinesia. Other artificial dopamine agonists seem have similar benefits.

Drugs that inhibit the enzyme catechol-*O*-methyltransferase (COMT) are also useful in the management of Parkinson's disease because of their ability to extend the duration of action of levodopa.

Most patients develop a fluctuating response to levodopa after 3 to 6 years. This may be a deterioration in symptoms at the end of the interval between doses (wearing-off reactions) or unpredictable loss of mobility (on–off reactions). COMT inhibitors smooth out wearing-off fluctuations. The first COMT inhibitor, tolcapone, helped many patients. Unfortunately, it produced very rare fatal hepatic toxicity, so its use is restricted. Another COMT inhibitor, entacapone, has recently been introduced and this drug does not damage the liver.

Ropinirole

Ropinirole, a non-ergot dopamine agonist, is usually started at a dose of 0.25 mg three times a day, after food. The regimen is then increased to a four-times a day schedule, with individual doses gradually built up 8 mg daily. All dopaminomimetics can cause nausea and hypotension. Taking drugs after meals reduces nausea, which can generally be abolished by the addition of 10 mg domperidone, 30 to 60 min before each dose of dopaminomimetic agent. Increasing the intake of salt and fluids usually alleviates hypotension. The response to ropinirole varies from individual to individual, and sometimes the dose has to be increased further, up to 24 mg daily. All dose adjustments should be undertaken slowly. Psychiatric reactions, in particular visual hallucinations, are commoner than with levodopa. The main advantage of ropinirole is that it produces less dyskinesia than levodopa.

A disadvantage of ropinirole is a propensity to cause somnolence during the day, which may exceed the sleepiness associated with levodopa and the ergot derivatives, bromocriptine and pergolide. This problem of daytime sleepiness is shared with the other new dopamine agonist, pramipexole (see below).

Pramipexole

Pramipexole is another non-ergot dopamine agonist. The usual starting dose is 0.375 daily, but this is slowly increased to a range between 1.5 and 6 mg daily in four divided doses. The benefits and side-effects of pramipexole are similar to those of ropinirole, but there has been no controlled comparison of the two drugs. Just as with ropinirole, daytime sleepiness can be a problem. In some countries, physicians have been instructed by government regulatory agencies to warn patients taking pramipexole of the dangers of driving when receiving the drug—several motor accidents have been reported. In rats, pramipexole can increase sleep, and it has been suggested that this effect is produced by excessive stimulation of dopamine D3 receptors.

Bromocriptine

Bromocriptine was the first artificial dopamine agonist to be used for Parkinson's disease. It is an ergot derivative, and, in common with other ergots, can rarely cause pleural or pulmonary fibrosis, pleural effusion, and erythromelalgia. It is useful to obtain a baseline chest radiograph before starting treatment—for comparison if dyspnoea later develops. The usual starting dose of bromocriptine is 0.25 mg once daily, gradually building up to a dose of 20 to 30 mg daily, distributed in four doses, after food, as for all antiparkinson drugs. Occasionally the dose of bromocriptine may need to be increased further, to 40 or 50 mg daily.

Pergolide

Pergolide, another ergot derivative, was the second artificial dopamine agonist to be introduced for the treatment of Parkinson's disease. The usual starting dose is 0.15 mg daily, which may be increased up to 6 mg daily. The benefits and side-effects of pergolide are similar to those of bromocriptine. Pergolide is an agonist at D1 and D2 dopamine receptors, in contrast to bromocriptine, which is an agonist at D2 receptors and an antagonist at D1 receptors. These differences in pharmacological profile do not seem to be reflected by corresponding differences in therapeutic results, so D1 agonism may not be achieved at the dose levels of

pergolide that are employed therapeutically.

Cabergoline

Cabergoline is the newest ergot-derived dopamine agonist. It has the unusual properties of a plasma half-life extending beyond 60 h, so that one dose a day suffices for treatment. The dose range is between 0.25 and 4 mg daily. While it is certainly useful to have a drug with a long duration of action, side-effects will, of course, persist longer than with other dopaminomimetic agents.

Levodopa

Levodopa is prescribed in combination with a decarboxylase inhibitor that does not cross the blood–brain barrier. Sinemet®, a combination of levodopa and carbidopa (co-careldopa), is marketed worldwide. Madopar® (or Prolopa®), a combination of levodopa and benserazide (co-beneldopa), is marketed in many countries. The usual starting regimens for levodopa/carbidopa and levodopa/benserazide is 50 mg of levodopa with 12.5 mg of the decarboxylase inhibitor, administered twice daily, increasing, at 3- to 5-day intervals, to 150 mg of levodopa four times daily. Levodopa, combined with a decarboxylase inhibitor, has the lowest prevalence of early side-effects—but with the artificial dopamine agonists, levodopa can cause nausea and hypotension. Over a longer period, levodopa frequently causes dyskinesia and fluctuations in mobility. While psychiatric side-effects can occur with levodopa, they are less common than with the artificial dopamine agonists.

Preparations of Sinemet and Madopar with prolonged action

Sinemet CR and Madopar HBS have both been marketed as formulations with longer plasma half-lives than standard Sinemet and Madopar. Because these longer acting preparations are often incompletely absorbed, a higher dose has to be given to achieve the same effect as the standard preparations. The increase is about 30 per cent, but there is considerable variation between patients—a usual dose would be 200 mg four times a day The longer acting preparations of Sinemet and Madopar are most useful in patients who have 'wearing-off' reactions.

Entacapone

Entacapone is a new COMT inhibitor given together with levodopa (plus a decarboxylase inhibitor), and the combination extends the plasma half-life of levodopa. Thus patients who have 'wearing-off' reactions get prolongation of the 'on period'. The usual dose is 200 mg with each dose of levodopa. Since entacapone can exacerbate dyskinesia, it is often necessary to reduce the dose of levodopa when introducing entacapone. A reduction of 20 to 30 per cent usually suffices.

Early medical treatment

Most neurologists tend to start elderly patients with Parkinson's disease on a combination of levodopa with a decarboxylase inhibitor, often using the long-acting preparations. This treatment achieves quite a rapid improvement in symptoms with minimal early side-effects. Unfortunately, with prolonged treatment, dyskinesia and fluctuations in mobility may develop and ultimately become significant problems. However, with older and frailer patients, the early period of excellent therapeutic response is given first priority.

With younger patients neurologists tend to place more emphasis on the long-term results. In these circumstances it is usual to start treatment with an artificial dopamine agonist, because of the lower risk for the development of fluctuations in mobility and dyskinesia.

Late medical treatment

After 2 or 3 years, patients who started treatment with an artificial dopamine agonist need the addition of levodopa. For most of the duration of their disease, therefore, parkinsonian patients take a combination of an artificial dopamine agonist and levodopa.

Surgical treatment

Some 30 to 40 years ago, there was widespread interest in achieving suppression of tremor by placing a lesion in the ventrolateral or the ventral intermediate nucleus of the thalamus. With the advent of levodopa this procedure was performed less often, but recently there has been a resurgence of interest in stereotactic surgery. Lesions placed in the globus pallidus consistently alleviate dyskinesia and also help tremor. To a lesser extent, pallidotomy can improve rigidity and bradykinesia.

The trend in surgery is now changing from producing lesions to electrical stimulation. 'Deep brain stimulation' produces good results, with less risk of unwanted consequences. Furthermore, deep brain stimulation of the subthalamic nucleus seems to result in greater benefit than stimulation of the pallidum. At present, the surgical treatment of Parkinson's disease is going through a period of active research extending to the transplantation of cells that produce dopamine or levodopa.

Other parkinsonian syndromes

Progressive supranuclear palsy

The classical diagnostic feature of progressive supranuclear palsy is paresis of conjugate gaze. Initially there is a problem with looking up and down on command, and as the condition advances there is difficulty in following objects up and down. Vertical movement of the eyes can only be achieved by central visual fixation while the neck is flexed and extended. Although the disturbance with gaze is first apparent in the vertical plane, ultimately horizontal gaze also becomes affected. There is over-reaction of the frontalis muscles and extension of the neck to compensate for the weakness of upward conjugate eye movement. Patients with supranuclear palsy usually have an akinetic syndrome involving all limbs, with prominent rigidity of the neck and impairment of righting reflexes. As the condition advances there is intellectual impairment. Progressive supranuclear palsy usually presents in middle or late life, and it advances more rapidly than Parkinson's disease. There is sometimes a family history suggesting autosomal dominant inheritance.

Neuronal loss occurs with neurofibrillary tangles and gliosis. Atrophy of the tectum may be seen on brain imaging. The pathology extends to the substantia nigra, globus pallidus, subthalamus, dentate nucleus, and periaqueductal grey matter.

In the early stages of clinical evolution, there may be some response to dopaminomimetic therapy. Ultimately, however, this is lost. Once it has been appreciated that a patient has an atypical form of parkinsonism with prominent gaze palsies, the differential diagnosis is vascular disease of the brainstem. During life, it is often difficult to distinguish between progressive supranuclear palsy and vascular disease.

Multiple system atrophy

The term 'multiple system atrophy' was coined by Oppenheimer. Included within this entity are three syndromes that usually overlap: (1) strionigral degeneration leading to parkinsonism; (2) autonomic failure; and (3) olivopontocerebellar degeneration.

Multiple system atrophy is a sporadic disorder that generally presents in later life. The pathology may be widespread, involving the basal ganglia, the dorsal nucleus of the vagus, Onuf's nucleus, the cerebellum, and the intermediolateral column of the thoracic spinal cord.

Multiple system atrophy runs a briefer course than Parkinson's disease. While there may be some initial response to dopaminomimetic drugs, these often exacerbate orthostatic hypotension, and hence are seldom useful in advanced illness. It is usually necessary to use agents such as fludrocortisone and midrodrine in an effort to raise the blood pressure, but sometimes the supine blood pressure is so high that treatment becomes extremely difficult.

Stridor, loud snoring, and sleep apnoea are frequently encountered in multiple system atrophy. Sometimes tracheostomy is necessary.

Drug-induced parkinsonism

Drugs that block dopamine receptors, or deplete the storage of dopamine, can produce a syndrome clinically indistinguishable from Parkinson's disease. Almost always, withdrawal of the offending drug leads to a restoration of normal motor function. In patients who are disabled by psychotic symptoms, it may be impossible to stop treatment in order to alleviate the parkinsonian syndrome. In these circumstances, the treatment of choice is clozapine, which has fewer tendencies to produce parkinsonism than any other antipsychotic agent. Clozapine carries a risk of bone marrow depression, so regular blood counts are mandatory when this drug is employed.

Essential tremor

Essential tremor is probably more common than all the other movement disorders put together. Sometimes the adjective 'benign' is attached to the disorder, but this can be quite misleading. Essential tremor may be severe and disabling, though it seldom evolves to produce other neurological deficits—an exception is the quite frequent occurrence of ataxia in long-standing essential tremor among the elderly.

Although essential tremor most often involves the hands, the neck is also vulnerable, so that the head shakes from side to side, or up and down. Sometimes the voice is affected and there may also be a tremor of the chin. Uncommonly, essential tremor can occur in the legs.

The characteristic feature of essential tremor is a postural shake that is sustained during movement. Essential tremor can start at any age, and it may interfere with work, for example in cases where dexterity is important. It is also troublesome when tremor, in front of the public, leads to difficulty in job performance. The impact of essential tremor is amplified through stress, for when a patient particularly wants to control the shaking, circumstantial anxiety leads to exacerbation of the tremor.

Essential tremor can progress to have a serious effect on the quality of life. For example, eating and drinking can become so difficult that patients will no longer eat in the company of others. Essential tremor tends to get worse in the later stages of life, when it can become prominent at rest, mimicking the resting tremor of Parkinson's disease. There may also be some 'cog-wheeling', when testing passive resistance to movement of the wrist. Thus essential tremor can easily be confused with early Parkinson's disease.

In about half the patients with essential tremor it is possible to obtain a family history, indicating autosomal dominant inheritance: genes have been identified on chromosomes 2 and 3. We do not know the site of the pathology of essential tremor, but interest has been focused on the internal olive, the red nucleus, and the dentate nucleus.

The treatment of essential tremor is usually difficult. Some 30 per cent of patients respond to b-blocking drugs, and benzodiazepines such as clonazepam can be helpful. Sometimes primidone is useful. Although essential tremor is characteristically alleviated by alcohol, it is obviously inadvisable for patients to use alcohol on a regular basis to stop the shaking.

When essential tremor is severe, and it is impossible to achieve an adequate response with drugs, improvement can usually be obtained by performing a thalamotomy. Recently, encouraging results have been achieved with deep brain stimulation using an electrode implanted into the ventrolateral nucleus of the thalamus. Deep brain stimulation carries less risk than thalamotomy.

Dystonia

The term 'dystonia' is used to describe a pattern of abnormal movements in which there is sustained contraction of muscles, which may last for several seconds and can induce distorted postures. If a joint can twist, dystonia usually produces torsion—hence the term 'torsion dystonia'. Superimposed on these slow muscle contractions are often quick involuntary movements, which may be regularly repetitive in space and time, or quite irregular.Dystonia is often exacerbated by voluntary movement.

In addition to the term 'dystonia' being applied to these physical signs, it is also used for the 'disease' where no cause can be found. If dystonia is limited to one group of muscles, it is termed 'focal dystonia', and when the entire body is involved it is termed 'generalized dystonia'.

Generalized idiopathic dystonia

This condition usually starts in childhood or early adult life. It is dominantly inherited and the gene responsible, *DYT1*, has been identified on chromosome 9. This form of dystonia is particularly common in Ashkenazim Jews. Dystonia generally progresses over several years and then reaches a plateau. Treatment is difficult; high doses of anticholinergic drugs such as trihexyphenidyl are most effective.

Dopa-responsive dystonia

In any patient with generalized dystonia it is most important to consider the possibility of dopa-responsive dystonia (hereditary progressive dystonia, Segawa's disease), because treatment for this disorder is so effective. Low doses of dopaminomimetic drugs are therapeutic throughout the life of the patient.

The pathogenesis of dopa-responsive dystonia has been worked out with considerable precision. Several mutations—about 60—can be responsible, all on chromosome 14, and all associated with the production of the cofactor for tyrosine hydroxylase, tetrahydrobiopterin. When this cofactor is lacking, tyrosine cannot be converted to dopa, so there is inadequate dopamine available in the striatum. This condition of dopamine depletion leads to dystonia, and in addition these patients usually have some parkinsonian features. Indeed, when dopa-responsive dystonia presents in late life, the parkinsonian features may overshadow the dystonia.

Hemidystonia

Dystonia involving one side of the body is often caused by a structural lesion in the contralateral basal ganglia. Infarcts are most often responsible, but brain imaging should be undertaken to exclude tumours.

Focal dystonia

The commonest form of dystonia is localized to the neck muscles; known as cervical dystonia or spasmodic torticollis. Sometimes there is a genetic basis for this condition, but usually it presents sporadically. Cervical dystonia generally responds to local injections of low concentrations of botulinum toxin. Blepharospasm is another focal dystonia that responds well to botulinum toxin, and dystonia of the vocal cords also responds, in many cases, to small injections of this toxin.

Tardive dyskinesia

Tardive dyskinesia is a syndrome of involuntary movements induced by long-term exposure to dopamine-blocking drugs. Major tranquillizers are usually responsible, but drugs employed to treat nausea, such as metoclopramide, are more recent culprits. While dystonic features usually predominate in tardive dyskinesia, there may also be obvious chorea.

Tardive dyskinesia is a chronic condition, and while it sometimes resolves spontaneously in younger patients, it is generally permanent in the elderly. Paradoxically, there is frequently an exacerbation—or initial appearance—of tardive dyskinesia when the drug responsible is reduced or stopped.

Treatment for tardive dyskinesia is withdrawal of the drug responsible when possible, and the administration of tetrabenazine. Unfortunately, tetrabenazine can cause depression, and is therefore contraindicated when tardive dyskinesia develops in a psychotic patient with depressive features. In this setting, serious psychotic disease with tardive dyskinesia is often managed by increasing the dose of the causal drug.

Chorea

The term 'chorea' derives from the Greek word for dancing. The physical sign of chorea is quick involuntary movement, irregular in space and time, generally most prominent in the extremities. Chorea is closely related to ballism, which is a high-amplitude displacement of the proximal muscles. In patients who develop ballism due to a stroke, the natural history of spontaneous improvement progresses through a phase of chorea, followed by a final resolution of the chorea.

Huntington's chorea

Huntington's chorea is an inexorably progressive disorder featuring a combination of chorea and dementia. The pathology primarily involves the caudate nucleus, but it extends beyond this initial focus.

Huntington's chorea is dominantly inherited; it is caused by a mutation on chromosome 4—the huntingtin gene—characterized by an expansion of CAG repeats.

Huntington's chorea generally starts in mid-adult life, around the age of 35 years. However, it can start in childhood, when it tends to have a more rigid presentation resembling parkinsonism. This is known as the Westphal variant. Sometimes Huntington's chorea can present in late-adult life, over the age of 70.

There is no effective treatment for Huntington's chorea. Tetrabenazine will reduce the involuntary movements, but the progressive dementia is resistant to all therapy.

Other forms of chorea

Neuroacanthocytosis is a rare autosomal recessive disorder associated with hypogammaglobulinaemia. It produces involuntary movements, of which the commonest is chorea—but neuroacanthocytosis can also cause parkinsonism or dystonia.

Sydenham's chorea is associated with rheumatic fever; both are now rare conditions, although they still occasionally occur in children. The chorea runs a benign course, with spontaneous remission usually taking place within 3 to 6 months.

Rarely, chorea can occur during pregnancy, or with exposure to oestrogens. Occasionally other drugs such as phenytoin and digoxin can cause chorea. Severe thyrotoxicosis may produce choreatic movements, as may systemic lupus erythematosus.

Tic

Although tics are usually quick involuntary movements, they may be more prolonged; the term 'dystonic tic' has been coined for the slow variety. Patients feel a compulsion to move, and the characteristic feature of all tics is the patient's ability to suppress them for brief periods of 30 to 60 s.

Tics confined to a single muscle group often persist through life. When tics move around from one part of the body to another, they are termed 'chronic multiple tic', or Gilles de la Tourette disease.

Gilles de la Tourette disease

Gilles de la Tourette disease is dominantly inherited, though the penetrance is variable, so that often a positive family history cannot be obtained. Gilles de la Tourette disease affects boys more often than girls, and it generally starts before the age of 15 years. Tics persist, waxing and waning in severity, over several years, though they often disappear in mid-adult life. In addition to the common movements of the face, shoulders, and hands, there are often truncal movements, and sometimes there are involuntary noises, such as excessive sniffing, throat clearing, or barking. Words may be uttered inappropriately and often these are expletives. Similarly, affected individuals may make obscene gestures.

The pathological basis of tics is not known, but they generally respond to drugs that decrease dopaminergic transmission. In the past, major tranquillizing drugs have been used, but these carry a risk of causing tardive dyskinesia, which, may itself, be expressed as a tic. Tetrabenazine is the most satisfactory treatment for most patients. Unfortunately, sedation can be a troublesome side-effect, particularly for children who are attending school. Depression can also develop and the drug should be withdrawn if this happens. It is therefore best to try to manage patients without using drugs, explaining that educating family, friends, employers, and teachers may help patients to cope. Sometimes, however, Gilles de la Tourette can cause severe tics and noises that are socially disruptive, and in such cases a minimal but therapeutic dose of tetrabenazine must be sought, usually between 12.5 and 100 mg daily.

Gilles de la Tourette disease is often associated with obsessive–compulsive behaviour and attention disorder with hyperactivity; in such cases psychiatric advice is necessary. Clomipramine or fluoxetine may be helpful in these circumstances.

Other conditions

The spectrum of movement disorders is extensive, and space does not permit consideration of all. For a more comprehensive survey, the reader is referred to Jankovic and Tolosa (1998), which provides descriptions of Creutzfeldt–Jakob disease, Hallervorden–Spatz disease, cortical basal ganglionic degeneration, dentatorubropallidoluysian atrophy, olivopontocerebellar atrophy, paroxysmal dystonia, restless leg syndrome, myoclonus, and ataxia. This source also describes the neurology of Wilson's disease, but for a discussion of the metabolic disturbance and treatment of Wilson's disease the reader is referred to Chapter 11.7.2.

Further reading

Calne DB (2000). Parkinson's disease is not one disease. *Parkinsonism and Related Disorders* **7**, 3–7.

Frucht S, *et al*. (1999). Falling asleep at the wheel: motor vehicle mishaps in persons taking pramipexole and ropinirole. *Neurology* **52**, 1908–10.

Gasser T, *et al*. (1998). A susceptibility locus for Parkinson's disease maps to chromosome 2p13. *Nature Genetics* **18**, 262–5.

Higgins JJ, Jankovic J, Patel PI (1998). Evidence that a gene for essential tremor maps to chromosome 2p in four families. *Movement Disorders* **13**, 972–7.

Ichinose H, *et al*. (2001). Dopa-responsive dystonia: from causative gene to molecular mechanism. *Advances in Neurology* **86**, 173–7.

Jankovic J, Tolosa E, eds. (1998). *Parkinson's disease and movement disorders*. Williams and Wilkins, Baltimore.

Johnson RH, *et al*. (1966). Autonomic failure due to intermedio-lateral column degeneration. *Quarterly Journal of Medicine* **35**, 276.

Kitada T, *et al*. (1998). Mutations in the Parkin gene cause autosomal recessive juvenile parkinsonism. *Nature* **392**, 605–8.

Krack P, *et al*. (1998). Treatment of tremor in Parkinson's disease by subthalamic nucleus stimulation. *Movement Disorders* **13**, 907–14.

Kurtzke JF, Goldberg ID (1988). Parkinsonism death rates by race, sex, and geography. *Neurology* **38**, 1558–61.

Largos P, *et al*. (1998). Effects of the D3 preferring dopamine agonist pramipexole on sleep and waking, locomotor activity and striatal dopamine release in rats. *European Neuropsychopharmacology* **8**, 113–20.

Lee CS, *et al*. (1994). Clinical observations on the rate of progression of idiopathic parkinsonism. *Brain* **117**, 501–7.

Marion S (2001). The epidemiology of Parkinson disease. *Advances in Neurology* **87**, 163–73.

Polymeropoulos MH, *et al*. (1997). Mutation in the a-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045–7.

Rajput AH, *et al*. (1989). Parkinsonism and neurofibrillary tangle pathology in pigmented nuclei. *Annals of Neurology* **25**, 602–6.

Rascol O, on behalf of the 056 Study Group (1999). Ropinirole reduces risk of dyskinesia when used in early PD. *Parkinsonism and Related Disorders* **5**, S83. [Abstract]

Rinne UK and the PKDS009 Collaborative Study Group (1999). A five year double blind study with cabergoline versus levodopa in the treatment of early Parkinson's disease. *Parkinsonism and Related Disorders* 5, S84. [Abstract]

Rojo A, *et al*. (1999). Clinical genetics of familial progressive supranuclear palsy. *Brain* **122**(Pt 7), 1233–45.

Samii A, *et al*. (1999). Reassessment of unilateral pallidotomy in Parkinson's disease. A 2 year follow up study. *Brain* **122**, 417–25.

Schulzer M, *et al*. (1994). A mathematical model of pathogenesis in idiopathic parkinsonism. *Brain* **117**, 509–16.

Singer C, Sanchez-Ramos J, Weiner WJ (1994). Gait abnormality in essential tremor. *Movement Disorders* **9**(2), 193–6.

Tanner CM, *et al*. (1999). Parkinson's disease in twins: an etiologic study. *Journal of the American Medical Association* **281**, 376–8.

Tasker RR (1998). Deep brain stimulation is preferable to thalamotomy for tremor suppression. *Surgical Neurology* **49**, 145–53.

Tsui JKC, Calne DB, eds. (1995). *Handbook of dystonia*. Marcel Dekker, New York.

Tsui JKC, *et al*. (1999). Occupational risk factors in Parkinson's disease. *Canadian Journal of Public Health* **90**, 334–5.

24.13.11 Disorders of movement (excluding Parkinson's disease)

*Roger Barker**

Movement disorders typically result from diseases of the basal ganglia and can be classified into one of five main categories: dystonia, chorea, tremor, tics, and myoclonus (see Table 1 for definitions). Each type of abnormal movement may occur in several diseases and many treatments are empirical. However, the study of molecular genetics and the use of functional imaging has revealed subtle neurochemical abnormalities which should facilitate development of more rational therapies.

In this section, attention is drawn to abnormal movements that are a principal manifestation of the disease. Movement disorders have been divided into hyperkinetic and hypokinetic conditions; however, this classification may be misleading because a given disease often evolves with time. It is probably more useful to classify movement disorders by type.

The dystonias

Definition

Dystonias are characterized by prolonged muscle contractions, causing abnormal movements and postures.

Classification

When no symptomatic cause for dystonia can be discovered, the syndrome is described as idiopathic or primary dystonia, and if generalized then the disorder is synonymous with idiopathic torsion dystonia. Secondary dystonia is due to a defined exogenous, structural, or metabolic disorder. 'Dystonia plus' syndrome constitutes dystonia in combination with other abnormalities (for example myoclonic dystonia) and heredodegenerative dystonia occurs when there is an underlying brain degeneration. Dystonia may affect the whole body (generalized dystonia), adjacent parts such as an arm and neck (segmental dystonia), or may be restricted to one part (focal dystonia) as in spasmodic torticollis, dystonic writer's cramp, blepharospasm, oromandibular dystonia, and laryngeal dystonia.

Idiopathic dystonia is frequently inherited (see below), but the focal dystonias usually occur sporadically in middle life. However, focal dystonias may be isolated fragments of the syndrome of idiopathic torsion dystonia.

Aetiology

The many metabolic and other inherited or sporadic diseases that can cause dystonia (Table 2) usually produce other neurological symptoms and signs. A symptomatic cause for dystonia is found in about 50 per cent of children with the condition, but is rare in those with adult onset. In adults, dystonia is most likely to remain confined to its site of origin as a focal dystonia, and the legs are rarely affected. Children often develop symptoms in the legs and frequently develop segmental or generalized dystonia.

The recent identification of mutations in genes responsible for forms of dystonia gives hope for understanding its basis (see Table 2). Abnormalities within the basal ganglia and associated cortical motor areas have been found in some patients with secondary dystonia.

Idiopathic (torsion) dystonia

Symptoms

Idiopathic (torsion) dystonia may present in childhood, when it is frequently inherited as an autosomal dominant trait, or in adult life, when a family history is unusual. In many families with early onset disease, genetic linkage studies have localized the abnormal gene mutation to the *DYT1* locus on chromosome 9q34 which codes for torsin A, a protein of unknown function expressed in the brain (including the substantia nigra). Ashkenazi Jews are particularly prone to this condition. It usually presents in children with dystonic spasms of the legs on walking, or sometimes of the arms, trunk, or neck. The condition is usually progressive when it commences in childhood; the spasms spread to all body parts, leading to severe disability within about 10 years. The intellect is preserved and there are no signs of pyramidal or sensory deficit. Speech is often spared, permitting the pursuit of intellectual employment despite severe physical disability. A spontaneous remission of symptoms occurs in about 10 to 20 per cent of patients, usually within 5 years of onset. There is no way of predicting who will remit or when such a remission will occur. Most remissions are transitory, lasting a matter of weeks or months, but occasionally they may persist.

In adults, the condition usually presents as a focal dystonia (blepharospasm, oromandibular dystonia, spasmodic dysphonia, torticollis, axial dystonia, or dystonic writer's cramp). The legs tend to be spared, and progression is slow, with the dystonia remaining confined to its site of origin. Segmental dystonia develops in some cases.

Treatment

Dystonia is distressing and difficult to treat. Every child and young adult with dystonia should receive a trial of levodopa (for example, Sinemet-Plus up to two tablets

three times a day for 3 months), for they may have the condition of dopa-responsive dystonia-parkinsonism (see below).

The drugs which most patients find helpful, and continue to take to suppress muscle spasm, are benzodiazepines such as diazepam, often in a large dose of 20 to 50 mg daily, and an anticholinergic such as benzhexol, again in large doses (up to as much as 120 mg/day). Fifty per cent of children and 10 per cent of adults will be helped, but adults are more sensitive to anticholinergic side-effects. Phenothiazines and other neuroleptics, such as haloperidol, may also help some patients, as may tetrabenazine, but often at the expense of drug induced parkinsonism. Unfortunately, dystonia is far less responsive to neuroleptics than is chorea. Many other drugs have been tried in dystonia, but none has gained wide acceptance.

The recent interest in neurosurgery for movement disorders has extended to the treatment of dystonia, especially when the disease is advanced and disabling. Originally the thalamus was targeted for surgery but pallidotomy has recently been favoured in the management of patients with generalized dystonia (whereas selective denervation procedures have been used in focal dystonia—see below). In general, patients with generalized torsion dystonia respond erratically to this procedure and at present there is little evidence to support its use.

Dopa-responsive dystonia-parkinsonism (Segawa's syndrome)

This condition is inherited as an autosomal dominant condition with incomplete penetrance and has as its defect mutations in the guanosine triphosphate cyclohydrolase 1 gene. This generates a cofactor for maintaining the normal activity of tyrosine hydroxylase, the rate limiting step in the catecholamine biosynthetic pathway. Homozygous deficiency of this enzyme severely inhibits tyrosine hydroxylase activity and produces mental retardation, seizures, and truncal hypertonia. However, the more common partial deficiency in tyrosine hydroxylase results in dystonia affecting the legs which becomes worse as the day goes on. Rest without sleep does not help, but sleep relieves the dystonia. Many patients also have features of parkinsonism, although focal dystonia may be the presenting feature. The disease can easily be mistaken for cerebral palsy (given its lower limb predominance) or an unexplained 'spastic paraparesis'.

There is a reduction in turnover of dopamine due to the abnormality in tyrosine hydroxylase activity; patients respond well to low doses of levodopa without showing any of the long-term complications encountered in Parkinson's disease.

Spasmodic torticollis

Symptoms

Spasmodic torticollis may be the presenting feature of dystonia in childhood, but isolated spasmodic torticollis usually occurs in the middle aged or elderly. The onset is usually insidious, often with initial pain, and sometimes appears to be precipitated by trauma. The dystonic spasms affect sternomastoid, splenius, and other neck muscles to cause the head to turn to one side (torticollis) (Fig. 1), or occasionally to extend (retrocollis) or to flex (antecollis) the neck. The spasms may be repetitive to cause tremulous torticollis, or sustained to hold the posture. The trunk commonly shows a compensatory lordosis.

Fig. 1 Spasmodic torticollis in a 57-year-old man. The hypertrophy of the sternomastoid muscle is evident.

The condition is usually lifelong, but remissions of a year or more occur in about one-fifth of cases. Most patients are otherwise normal apart from their torticollis, although some may exhibit a postural tremor similar to that of benign essential tremor, and a minority may develop dystonia elsewhere. As with all types of dystonia, the frequency and intensity of the muscle spasms vary considerably, being particularly worse in conditions of mental or emotional stress. A feature characteristic of spasmodic torticollis is the 'geste antagonistique', in which the patient discovers some particular manual act which controls the deviation of the head. A touch of the forefinger to the jaw may suffice, but other more complex and bizarre actions are common.

Treatment

Spasmodic torticollis, like other types of adult onset focal dystonia, does not usually benefit from conventional drug therapy. The best treatment is injection of botulinum toxin A into the most affected muscles. Botulinum toxin prevents the release of acetylcholine and causes functional denervation with localized muscle weakness. Identification of the overactive muscles is a prerequisite to the administration of localized injections of botulinum toxin which, in the case of torticollis, typically involves injections into the sternomastoid and splenius muscles. These injections usually have an effect within a week although the maximum benefit is not apparent until several weeks later. Repeat injections are required approximately every 3 months as relapse, by terminal sprouting, occurs. In about 10 to 20 per cent of patients, antibodies eventually develop to the botulinum toxin A making it less effective with time. In these cases a switch to a botulinum toxin type F or B may be desirable; the long-term efficacy of this manoeuvre is under investigation.

Surgery is sometimes practised and local denervation procedures are still considered in patients with otherwise intractable cervical dystonias.

Dystonic writer's cramp

Symptoms

Inability to write (or to type, play a musical instrument, or wield any manual instrument) has many causes but in most patients no objective neurological deficit is found other than abnormal posturing of the hand and arm on writing. Typically, the pen is gripped with great force and driven into the paper (Fig. 2). However, in some patients the arm adopts a typical dystonic posture and in such cases of dystonic writer's cramp, other manual acts such as wielding a knife or screwdriver are similarly affected. Such dystonic writer's cramp may be the initial symptom of generalized torsion dystonia, but in adults it often remains as an isolated disability. The same considerations apply to other occupational cramps, such as pianist's cramp.

Fig. 2 (a) Dystonic writer's cramp in a 52-year-old man, whose right elbow rises and whose fingers grip the pen so tightly that they slide off. (b) Example of writing and drawing in this patient showing difficulty in executing the task and thus legibility of script and ability to copy simple figures.

Treatment

Writer's cramp, and related conditions, are usually permanent disabilities. Advice to write with the opposite hand allows most to cope with everyday events, but approximately 1 patient in 20 then develops the same problem in the non-dominant hand. Drugs (such as bezhexol and diazepam) are rarely of benefit but botulinum toxin injections into the muscles of the affected forearm may help some patients.

Blepharospasm and oromandibular dystonia (cranial dystonia)

Symptoms

Blepharospasm refers to recurrent spasms of eye closure. The orbicularis oculi forcibly contracts for seconds or minutes, often repetitively and sometimes so frequently as to render the patient functionally blind (Fig. 3). Spasms of eye closure commonly occur while reading or watching television, or in bright light; they often decrease or disappear when the subject is alerted or under scrutiny. Oromandibular dystonia refers to recurrent spasms of muscle contraction affecting the mouth, tongue, jaw, larynx, and pharynx, causing spasms of lip protrusion or retraction, jaw closure or opening (Fig. 4), and difficulty in speech and swallowing. Such patients may lacerate their lips and tongue or even dislocate their jaw, and are usually unable to cope with dentures. Speech may take on a characteristic, forced strained quality, and chewing and swallowing may be impaired.

Fig. 3 Blepharospasm in a 57-year-old woman. Her jaw also is forcibly clamped shut, biting her gums, and some spasm of orbicularis oris is evident, in addition to the obvious spasm of orbicularis oculi.

Fig. 4 Oromandibular dystonia in a 42-year-old woman. The spasm of forced jaw opening with tongue protrusion is evident.

These two conditions are closely related, for the patient with blepharospasm may develop oromandibular dystonia and vice versa. The term Brueghel syndrome is often used when the dominant (or only) feature is a dystonically opened jaw, whilst Meige syndrome has blepharospasm as its central feature. Both conditions may occur in generalized torsion dystonia, or result from drugs; they also appear in isolation in late life without evident cause.

Treatment

Unfortunately, both blepharospasm and oromandibular dystonia are notoriously difficult to control with drugs (for example benzhexol, diazepam, and/or a neuroleptic). Surgery cannot improve oromandibular dystonia but can relieve blepharospasm. The best treatment for blepharospasm is to inject botulinum toxin into the orbicularis oculi which gives relief in about 70 to 80 per cent of cases, thereby restoring normal vision for about 3 months. The injections can be repeated as necessary. Botulinum toxin injections can be used to control some jaw spasms.

Spasmodic dysphonia

Dystonic spasms of the muscles controlling the vocal cords cause spasmodic dysphonia, which impairs speech and singing, and may be severe enough to prevent communication. The most common type involves the adductor muscles, leading to a strangled speech with pitch breaks and stops. Less common is abductor dysphonia which produces a breathy, low-volume voice. The diagnosis can be established by direct non-invasive visualization of the vocal cords during talking. Spasmodic dysphonia may occur in association with cranial or generalized dystonia, or may appear as an isolated focal dystonia in adult life. Speech can be restored by injection of botulinum toxin into the overactive vocal muscles, identified by electromyography.

Paroxysmal dystonia

Focal dystonias often commence with the appearance of a dystonic posture or spasm only on one motor act (action dystonia), but there are rare, usually familial, disorders in which dramatic dystonia occurs intermittently in attacks, the patient being normal in between. These conditions are thought to be caused by mutations in genes encoding ion channels. Several families with paroxysmal dystonic choreoathetosis have now shown linkage to chromosome 2q35–37, where a number of candidate genes have been identified for study.

Chorea

Chorea is seen in many disorders (see Table 3, Fig. 5), but the most common cause other than the treatment of Parkinson's disease is Huntington's disease. There are also several non-inherited conditions in which chorea can occur and in which treatment is beneficial.

Fig. 5 Chorea due to polycythaemia rubra vera in a 57-year-old woman. The characteristic fleeting choreic movements are captured in these three sequential frames.

Huntington's disease

Definition

A dominantly inherited, relentlessly progressive disease, usually of middle life, characterized by chorea and dementia. It was first described in 1872 by George Huntington, a year after he had qualified in medicine, in a handful of families of English descent in a region of Long Island, New Jersey.

Aetiology

The prevalence of the disease is about 1 in 10 000 and it occurs in all ethnic groups worldwide. The condition is inherited as an autosomal dominant trait with full penetrance. The genetic defect is now known, can be tested for, and consists of an abnormal cytosine–adenine–guanine triplet repeat in the gene encoding huntingtin on chromosome 4 and represents one of a number of triplet repeat disorders causing neurological disease. Triplet repeat sequences normally exist in several genes but when an excess number of repeats occurs a disease state is produced. This pathological triplet (or trinucleotide) repeat occurs in the coding region of the huntingtin gene and the consequence of a large unstable DNA sequence is that the triplets can expand during mitosis and meiosis, resulting in longer triplet repeat sequences (dynamic mutation). The most likely time for triplet expansion is during spermatogenesis and subsequent fertilization/embryogenesis; this has two major implications. First, longer repeats tend to occur in the offspring of affected men, and secondly longer repeats tend to occur in the subsequent generations. This results in an earlier onset and more severe form of the disorder in subsequent generations—a phenomenon known as genetic anticipation; i.e. longer repeat sequences are associated with younger onset, more severe forms of the disease.

The abnormal expansion of the cytosine–adenine–guanine repeat in Huntington's disease (more than 36 repeats) causes a new gain of function in the mutant huntingtin. This new protein acquires a unique function that is central to the evolution of the neurodegenerative process. Furthermore this protein is known to interact with a number of other proteins (for example huntingtin associated protein 1), which may be critical for the development of selective pathology. Inclusion bodies resulting from the polymerization of polyglutamine sheets in the mutant huntingtin protein develop in neurones but the exact mechanism by which selective neuronal death at specific sites occurs is unknown.

The chorea of Huntington's disease appears to result from relative overactivity of dopamine mechanisms in the brain, perhaps because the intact dopaminergic nigrostriatal pathway is releasing approximately normal quantities of dopamine on to only a few remaining striatal neurones. Positron emission tomographic studies using ^{18}F-labelled deoxyglucose have revealed a profound reduction of glucose metabolism in the striatum, even in patients without discernible cerebral atrophy.

Pathology

The brain is generally atrophic, with conspicuous damage to the cerebral cortex and corpus striatum, where there is loss of nerve cells and reactive gliosis without inflammatory changes associated with extensive neurotransmitter changes.

Coronal section of the brain characteristically shows dilated lateral ventricles, in which the floor becomes concave rather than convex, due to marked caudate atrophy; commonly, cortical atrophy is also evident. The main histological abnormality is found in the caudate and putamen, where there is extensive loss of small neurones, leading to shrinkage, and a false impression of gliosis, although such changes are found elsewhere. More recently attention has focused on the early stages of the condition, as the identification of the genetic defect involved has allowed for the disease to be diagnosed with certainty. In early cases and in mice transgenic for the mutant human huntingtin protein the earliest histological findings are the neuronal inclusions, which precede the cell loss.

Symptoms

The onset is typically insidious, usually between the ages of 30 and 50 years, and can be with motor, cognitive, and/or psychiatric symptoms and signs. The initial symptoms are frequently those of a change in personality and behaviour, but chorea may be the first sign. At this stage the patient often retains distressing insight, fully aware of what is in store. Serious depression is common and suicide is a risk. Erratic behaviour at work or in society may lead to psychiatric referral, or rarely a frank schizophrenic-like psychosis may develop. As the disease progresses, cognitive deficits become more pronounced and chorea more severe with walking, speech, and the use of the hands are impaired. As the disease progresses, many patients develop increasing rigidity and akinesia, with reduction of the chorea. Finally the patient becomes bedridden with marked weight loss; death occurs on average about 14 years from the onset. Huntington's disease does not always present in this fashion, and a number of variants are recognized including an akinetic-rigid parkinsonian syndrome (the Westphal variant), which is most frequent in children.

Diagnosis

Despite the diverse clinical manifestations of Huntington's disease, with genetic testing the diagnosis is now straightforward. However, in some cases the characteristics of the disease are not obvious and a history is not available, which can mean that the diagnosis is overlooked (see Table 3).

Treatment

There is currently no cure for Huntington's disease. Drugs that modify the dopaminergic input to the striatum can be used (for example tetrabenazine and sulpiride) to treat the chorea but rarely provide a sustained benefit—in fact most patients with Huntington's disease rarely complain of their chorea. Other drugs may be required for psychiatric symptoms including selective serotonin reuptake inhibitors for depression, neuroleptics for psychotic symptoms, and carbamazepine and sodium valproate as mood stabilizers.

A further recent intervention in Huntington's disease has involved the use of neurotrophic factors, such as ciliary neurotrophic factor as well as neuroprotective therapies (for example coenzyme Q); the benefits of these measures are uncertain. Surgical treatment for chorea is poorly documented but there has been increasing interest in the possibility of neural transplantation in Huntington's disease.

Sooner or later chronic hospital care is required for patients with Huntington's disease. Increasing nursing problems may require admission to hospital where dietary advice (including gastrostomy for feeding) and physiotherapy may be of benefit. Particular attention should be directed towards supporting the family, for the nature of the disease poses great ethical and emotional problems.

Genetic testing

Predictive testing programmes are now available for individuals at risk, provided by multidisciplinary teams specializing in this condition and often directed by an

experienced neurogeneticist.

Sydenham's chorea

Definition

Chorea (St Vitus's dance) associated with psychological disturbance due to rheumatic fever (rheumatic chorea) in childhood and adolescence was first described by Thomas Sydenham in 1686.

Aetiology and pathology

Sydenham's chorea was associated with many streptococcal infections, but it is now most frequently associated with acute rheumatic fever, and as a result is rare in most parts of the world. The mechanism is thought to be antibody mediated, against epitopes within structures of the basal ganglia, which would help explain the characteristic radiological lesions seen within the basal ganglia of affected patients. Antineural antibodies have been isolated in cases of Sydenham's chorea and may result from cross-reactivity with elements of group A streptococcal membranes, although the pathogenic nature of these antibodies has not been demonstrated unequivocally.

Symptoms

About three-quarters of cases occur between the ages of 7 and 12 years. The onset is usually gradual, but may be abrupt. The initial symptoms are often psychological, with irritability, agitation, disobedience, and inattentiveness. A frank organic confusional state occurs in about 10 per cent of patients. Generalized chorea then appears and may worsen for a few weeks; speech is impaired in about a third of patients. The chorea is predominantly unilateral in about 20 per cent of patients, and in severe cases is accompanied by flaccidity and subjective weakness (chorea mollis). Although cardiac disease may be found, the child usually has no fever or other manifestations of rheumatic fever.

The chorea and psychological disturbance recover over 1 to 3 months, rarely up to 6 months, but recurrences occur in about a quarter of patients over the next 2 years. About a third of patients will show evidence of rheumatic cardiac involvement at the time of the illness, and about the same proportion later develop chronic rheumatic heart disease. Those who have suffered one or more attacks of Sydenham's chorea are at particular risk of developing chorea in adult life during pregnancy (chorea gravidarum), or when exposed to drugs including oral contraceptives, digoxin, or phenytoin. Although usually self-limiting, more persistent neurological deficits occasionally occur in Sydenham's chorea.

Treatment

Treatment as for rheumatic fever is necessary. The chorea may be controlled with diazepam, haloperidol, or tetrabenazine. A course of penicillin should be given, and prophylactic oral penicillin should be continued until about the age of 20 years to prevent further streptococcal infection.

Hemiballism (hemichorea)

Hemiballism refers to wild flinging or throwing movements of one arm and leg. These movements, like those of chorea, are irregular in timing and force, but involve the large proximal muscles of the shoulder and pelvic girdle. The occasional child or adolescent with Sydenham's chorea may present with hemiballism, but the syndrome usually occurs in elderly hypertensive and/or diabetic patients as a result of a stroke. The vascular lesion usually affects the subthalamic nucleus, although lesions at other anatomical sites may be responsible. It may appear as the hemiplegic weakness improves, when it is often accompanied by thalamic pain, although in other patients the hemiballism appears abruptly without weakness or sensory deficit. The intensity of the movements varies from mild to a severity that causes injury and requires urgent treatment.

Hemiballism due to stroke usually remits over 3 to 6 months. Treatment with a phenothiazine, haloperidol, or tetrabenazine will often control hemiballism until recovery occurs, but interventional neurosurgery is occasionally required and may be beneficial.

Tremor

Three types are generally recognized—static, postural, and action tremors (Table 4). Static tremor occurs when a relaxed limb is fully supported at rest. Postural tremor appears when a part of the body is maintained in a fixed position and may also persist during movement. Kinetic or action tremor occurs specifically during active voluntary movement of a body part. If the amplitude of such an action tremor increases as goal-directed movement approaches its target, this is an intention tremor. Psychogenic tremors are generally rare. They are often of sudden onset with a variable but rarely remitting clinical course and typically affect the trunk or limb with standing and/or using the limb respectively.

Physiological tremor has a frequency in the 7 to 11 Hz band and is typically symptomatic in states of increased sympathetic nervous activity and is increased by stimulation of peripheral b_2-adrenergic receptors in muscle. The fine postural tremors associated with stress and anxiety states along with thyrotoxicosis fall into this category and usually respond to b-adrenergic blocking drugs. Symptomatic postural tremors occur in association with many neurological disorders and can be shown to differ from physiological tremor by frequency analysis.

Benign essential (familial) tremor

Definition

A condition characterized by postural tremor of the arms and head which can present at any age, although usually in early adult life. It is only slowly progressive, generally causes mild disability, and is not associated with dystonia or parkinsonism.

Aetiology

The cause of benign essential tremor is unknown. A positive family history is obtained in over half of such patients with a pattern of inheritance that indicates an autosomal dominant trait.

Pathophysiology

No pathological or biochemical abnormality has been identified in benign essential tremor, but few cases have come to autopsy. Essential tremor is usually of a frequency of 5 to 8 Hz (Fig. 6). Recent functional imaging studies indicate abnormal activation of the cerebellum, red nucleus, and thalamus.

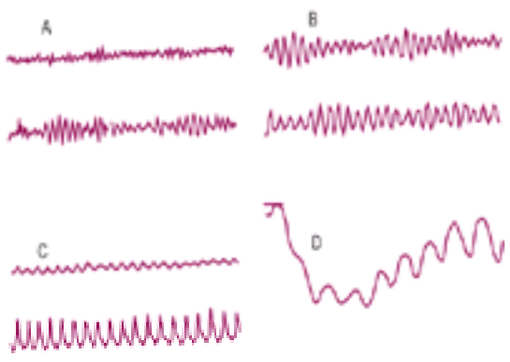

Fig. 6 Recordings of tremor with accelerometers placed on the right arm (above) and the left arm (below) in (a) a patient with enhanced physiological tremor of the outstretched arms (frequency 9 to 10 Hz), (b) a patient with benign essential tremor with the arms outstretched (frequency 7 Hz), (c) a patient with left-sided Parkinson's disease at rest (frequency 5 Hz), and (d) a patient with severe cerebellar disease and marked intention tremor attempting to touch his nose with the right hand (frequency 2 to 3 Hz). All recordings are of 4 s duration. (By courtesy of Dr M. Gresty.)

Symptoms

Tremor is present in one or both hands on maintaining a posture, as when holding a cup or glass. Handwriting becomes untidy and tremulous (see Fig. 7). There is no tremor at rest, but a rhythmic oscillation develops when the patient holds the arms outstretched. On movement, as in finger–nose testing, the tremor continues but does not get strikingly worse, as is the case with cerebellar intention tremor. Tremor of the head (titubation) and jaw is present at about 50 per cent of cases, and tremor of the legs occurs in about a third. Despite the tremor, tests of co-ordination are usually normal, walking is unaffected, and there are no other neurological abnormalities.

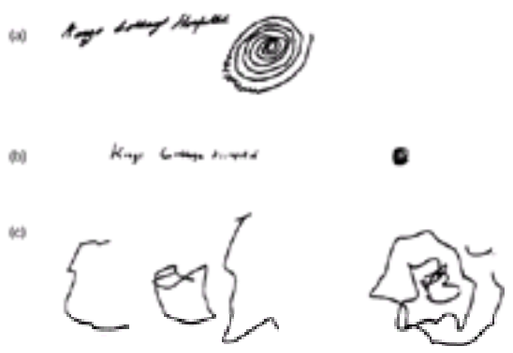

Fig. 7 Samples of handwriting and spiral drawings from (a) a 50-year-old woman with benign essential tremor, (b) a 38-year-old man with Parkinson's disease, and (c) a 20-year-old man with torsion dystonia (who attempts to write COLLEGE).

Two other factors are characteristic of this disorder. First, a family history and, secondly, the observation that small or moderate doses of alcohol may suppress the tremor. The illness is static or only slowly progressive in most patients, causing predominantly a social disability, but individuals dependent upon manual skill may be disabled.

Other variants of the syndrome are occasionally encountered. Thus isolated, inherited head tremor may occur, with either 'yes–yes' or 'no–no' movements, and tremulous 'writer's cramp' (primary writing tremor) is recognized. Tremor of the legs on standing, at around 5 to 8 Hz (orthostatic tremor), may occur as an isolated syndrome or in the context of essential tremor.

Treatment

Although alcohol may suppress the tremor effectively, and can be of value if used wisely, there is a risk of patients becoming alcoholics. Benzodiazepines, such as diazepam, may give some relief at times of stress, but have no major effect on the tremor. Thirty to 40 per cent of patients respond satisfactorily to a b-adrenergic receptor antagonist such as propranolol (up to a dose of 240 mg/day). Primidone, in standard anticonvulsant dosages, also helps some. Stereotaxic thalamotomy or deep brain stimulation of the thalamus may be required in the very small number of patients whose tremor is severe.

Tics

Tics occur in several disorders (see Table 5) and are defined as simple or complex. A simple tic is a sudden, rapid twitch-like movement always of the same nature and at the same site and occurs in about a quarter of all children, disappearing within a year or so (transient tic of childhood). Sometimes these persist into adult life, but are rarely considered as abnormal (chronic simple tic) such that only a minority seek assistance. Tics that are more widespread and severe, and take the form of complicated stereotyped patterns of motor action, are termed complex tics.

Characteristically, complex and simple tics can be suppressed voluntarily although this causes mounting inner tension which can only be relieved by expression of the tics. Complex multiple tics accompanied by vocal utterances, particularly swear words (coprolalia), and compulsive thoughts constitute the Gilles de la Tourette syndrome, which is an organic cerebral disease of unknown aetiology.

Gilles de la Tourette syndrome

Aetiology

George Gilles de la Tourette described the syndrome of chronic multiple tics with vocalizations in 1885, although Itard had described an earlier case. The condition appears to be inherited as an autosomal dominant trait, with variable penetrance. In affected families the genetic abnormality may be expressed as the full syndrome, as chronic multiple tics alone, or as an obsessive–compulsive neurosis (see below). Whilst the cause of this bizarre condition is unknown, the basal ganglia are believed to be involved because abnormalities within the striatum (and paralimbic areas) have been detected using positron emission tomography. Subtle abnormalities within the caudate and pallidum have been reported at post-mortem. Furthermore, drugs acting on the basal ganglia, such as haloperidol, may control the symptoms.

Symptoms

The illness usually begins between the ages of 5 and 15 years, with multiple involuntary repetitive muscular tics which vary in site, number, frequency, and severity with time. These particularly affect the upper part of the body, especially the face, neck, and shoulders, more than the limbs and trunk. Typical initial symptoms are eye blinking, head nodding, sniffing, or stuttering. With time other more complex tics affecting other parts of the body appear. The motor tics are often preceded by a sensory urge (sometimes called a sensory tic), and they can be controlled, albeit temporarily, by an effort of will. Eventually, however, the patient has to 'let the tics go'.

Sooner or later, most patients with multiple tics make involuntary noises, such as grunting, squealing, yelping, sniffing, or barking. Indeed, the coexistence of such noises with multiple tics is an essential feature of Gilles de la Tourette syndrome. In about 60 per cent of cases the noises become transformed into swear words (coprolalia). About a third of patients also exhibit echolalia—an involuntary tendency to repeat words or sentences just spoken to them. A smaller proportion of patients also exhibit copropraxia (involuntary obscene gesturing) and echopraxia (involuntary imitation of the movements of others), as well as palilalia (involuntary repetition of their own words or sounds).

Many patients with Gilles de la Tourette syndrome also exhibit features of an obsessive–compulsive disorder. These psychiatric manifestations of the condition may be even more disabling than the tics. A hyperactive attentional disorder is common in children.

Once established, Gilles de la Tourette syndrome is usually lifelong, although its severity waxes and wanes. A small proportion of cases, probably less than 1 in 20,

experience spontaneous remission of symptoms after adolescence, but in most the multiple tics and the vocalizations persist, although they usually become less prominent in adults. No other neurological abnormality develops; intellect and motor co-ordination are retained.

Diagnosis

Patients with tics and vocal utterances, which are essential components of Gilles de la Tourette syndrome, are often initially considered to have a psychiatric disorder or, if organic disease is considered, chorea or dystonia are diagnosed. Once coprolalia is evident there is no difficulty in establishing the correct diagnosis, although other diseases such as Wilson's disease with onset in childhood require consideration.

Treatment

The multiple tics and the vocalizations of coprolalia cause considerable distress, social isolation, and psychological harm. Neuroleptic drugs such as haloperidol or pimozide may satisfactorily control tics, noises, and coprolalia. The effective dose requires careful and gradual titration as there is risk of side-effects, especially the emergence of a tardive dyskinesia after months or years of therapy (see below). Since treatment must usually be for life, the harm must be carefully balanced against the need for any form of therapy. The obsessive-–compulsive symptoms of the illness may improve with drugs such as clomipramine or fluoxetine. In extreme cases psychosurgery has been undertaken with limbic leucotomies.

Myoclonus

Myoclonus is a feature of many neurological diseases and can be classified according to its aetiology (see Table 6).

Generalized or multifocal myoclonus can occur in four clinical settings:

i. as the solitary feature of the illness (essential myoclonus);
ii. as a dominant feature of a progressive brain disease (progressive myoclonic encephalopathy);
iii. as a residual feature of some transient brain insult (static myoclonic encephalopathy); or
iv. as a feature of obvious epilepsy (myoclonic epilepsy).

Benign essential myoclonus

This condition consists of widespread myoclonus affecting all four limbs, trunk, neck, and face, occurring at about 10 to 50 per minute, enhanced by action and sensory stimuli, often in the context of a positive family history. Onset is usually in childhood or adolescence, but disability is strikingly mild, there is no progression, intellect is normal, fits do not occur, and no other deficit appears. Some patients report that alcohol helps their jerks, and many respond to a b-adrenergic antagonist such as propranolol.

Progressive myoclonic encephalopathies

Most of the diseases causing a progressive myoclonic encephalopathy are described in detail elsewhere, particularly the lysosomal storage disorders and other metabolic disorders as well as the spinocerebellar degenerations. A discussion of other associated conditions lies outside the scope of this chapter.

Static myoclonic encephalopathies : postanoxic action myoclonus (Lance–Adams syndrome)

This is a distinct entity that may appear after a period of cerebral anoxia, typically respiratory arrests in the context of an acute asthmatic attack. After recovery of consciousness, such patients exhibit muscle jerks affecting the face, trunk, and limbs, often provoked by sensory stimuli, and strikingly elicited by willed voluntary action. The condition has been associated with abnormalities of brain 5-hydroxytryptamine, as 5-hydroxytryptophan produces a marked improvement in some patients. However, the side-effects of this therapy, in particular the development of the eosinophilia myalgia syndrome, have meant that other treatments such as clonazepam, piracetam, and sodium valproate are more commonly used

Myoclonic epilepsies

In the myoclonic epilepsies, epileptic seizures are the obvious and dominant feature of the condition. There is some confusion in separating the many conditions that may cause this syndrome, which occurs particularly in children. A convenient, if arbitrary, distinction is based on the age of onset and is discussed in more detail in the section on epilepsy.

Focal myoclonus

There are a number of conditions in which myoclonic muscle jerking may be restricted to one part of the body. Some pathological processes may cause focal myoclonus limited to those segments innervated by the part of brainstem or spinal cord affected (segmental myoclonus). Palatal myoclonus, with rhythmic contractions 60 to 180 per minute is an unusual variant. Sometimes this spreads to the pharynx and larynx and speech is disturbed; the ocular muscles maybe involved. Similar pathologies causing cerebral damage, particularly to the cerebral cortex, may cause rhythmic repetitive focal muscle jerking associated with electrical evidence of epileptic cortical discharge in the electroencephalogram (epilepsia partialis continua).

Spinal myoclonus

Repetitive, often rhythmical, myoclonic jerking restricted to a limb, or even to a few muscles of an arm or leg, may occur with myelitis, spinal cord tumour, or angioma, or after spinal cord trauma. The rhythmic muscle jerking occurs spontaneously, at 20 to 180 per minute, is not affected by peripheral stimuli, often persists in sleep, and is not associated with any change in the electroencephalogram. Anticonvulsants may help, but such segmental myoclonus is often very difficult to control.

Epilepsia partialis continua

Encephalitis, tumour, abscess, infarct, haemorrhage, or trauma to the cerebral cortex may rarely cause repetitive, rhythmic muscle jerking once or twice a second, confined to one collection of muscles, persisting even in sleep for days, weeks, or months. Usually the damage involves not only the cerebral cortex, but also deeper structures including the thalamus. Because of its large cortical representations, the most common site of epilepsia partialis continua is the hand. Typical Jacksonian focal motor fits, and grand mal seizures may also occur in such patients. The surface electroencephalogram usually shows a spike discharge over the opposite motor cortex preceding each jerk by a short interval. Treatment is with anticonvulsants, but may be difficult.

Hemifacial spasm

Hemifacial spasm occurs at a frequency of about 1 in 100 000 people and most commonly affects middle-aged or elderly women, and usually appears without obvious cause. Rarely, it may be symptomatic of demonstrable facial nerve compression. The condition consists of irregular, but repetitive clonic twitching of the muscles of one side of the face. Usually those around the eyes are first involved, producing a feeling identical to the benign myokymia of the lower eyelid which occurs in normal people when fatigued. However, the repetitive twitching spreads slowly to involve the whole face, each spasm closing the eye and drawing up the corner of the mouth. At this stage, a mild facial weakness and contraction becomes evident, but a frank facial palsy never develops. Facial sensation is normal and there are no other physical signs in idiopathic hemifacial spasm. The disorder is so distinctive and unilateral that it is rarely confused with other conditions. True facial myokymia, due to brainstem tumour or demyelination, consists of a continuous rippling contraction of the facial muscles, giving the appearance of a 'bag of worms'.

Treatment with drugs is usually unrewarding. Posterior fossa exploration, with separation of blood vessels from the seventh nerve gives long-lasting relief and failure, when it occurs, is normally evident within the first few months after surgery. However, injection of botulinum toxin into the facial muscles, repeated every 3 to 4 months, is a simpler and effective treatment.

Other movement disorders

Restless leg syndrome (Ekböm's syndrome)

This is a common and poorly understood condition in which patients have a desire to move their extremities often in association with paresthesiae and dysaesthesiae. This is made worse by rest and so is often mainly present at the end of the day and at night and is relieved by activity such as walking around. It is commonly associated with periodic limb movements during sleep. The aetiology of the condition is unknown, but may be due to abnormal cerebellar and thalamic activation. It can be seen with peripheral neuropathies, uraemia, pregnancy, iron deficiency, rheumatoid arthritis, and spinal cord lesions. In some cases there is a family history suggestive of an autosomal dominant inheritance. It responds to a number of drugs including L-dopa, dopamine agonists, baclofen, carbamazepine, clonazepam, clonidine, and opioid drugs.

Stiff man syndrome

This term includes a range of rare conditions which are characterized by muscle rigidity with or without spasms. This includes the stiff man syndrome which is characterized by axial rigidity involving the paraspinal muscles which leads to a hyperlordotic posture of the back and an abnormal gait often described as walking through treacle. These patients often have muscle spasms in response to sensory stimuli with an exaggerated startle response. The patients characteristically have antiglutamic acid decarboxylase antibodies which may account for the high incidence of diabetes mellitus and other autoimmune disorders in this condition. Treatment is usually with baclofen and diazepam, although clonazepam, sodium valproate, and vigabatrin have all been used successfully. Immunosuppressive therapy is often disappointing, although it may benefit some patients.

Hyper-rekplexia

Startle is a sterotypic response that involves a complex series of movements including eye closure, facial grimacing, and a typical body posture. In certain conditions it is exaggerated, and termed hyper-rekplexia. It can be seen in a number of conditions and with a range of central nervous system lesions as well as occurring in isolation. In some cases the condition is inherited in an autosomal dominant fashion and in these cases mutations in the glycine receptor are found. In general the condition responds to clonazepam.

Drug induced movement disorders

The extensive use of antipsychotic neuroleptic drugs has led to much iatrogenic extrapyramidal disease. These drugs, all of which block dopamine receptors in the basal ganglia and elsewhere, are used widely to control acute psychotic behaviour, whatever its cause, and to prevent relapse of schizophrenia. They also are employed as antiemetics, as are other similar drugs such as metoclopramide, and to treat vertigo. The major neurological complications of these drug therapies are summarized in Table 7.

Akathisia refers to an irresistible and unpleasant sensation of motor restlessness, and the inability to sit or stand still, all of which may be mistaken for a recurrence of psychotic behaviour. Akathisia remits if the offending neuroleptic is withdrawn, or if the dose can be reduced sufficiently: It does not usually respond to anticholinergic drugs, but may be helped by a benzodiazepine or propranolol.

Anticholinergic drugs may also be used to treat drug induced parkinsonism if the causative neuroleptic has to be continued for psychiatric reasons, although anticholinergics are not routinely administered to those on neuroleptics.

Acute dystonic reactions commonly consist of oculogyric crises, trismus, neck retraction, or torticollis, and may be mistaken for tetanus or meningitis. Although uncommon, acute dystonic reactions pose a repeated diagnostic problem in casualty departments. Such reactions rapidly disappear after intravenous injection of an anticholinergic drug such as Kemadrine or a benzodiazepine.

Chronic tardive dyskinesias are the most serious of the drug induced movement disorders for they usually persist despite drug withdrawal. About 20 per cent of those receiving chronic neuroleptic therapy will exhibit a tardive dyskinesia. The characteristic syndrome is one of orofacial mouthing, with lip smacking and tongue protrusion (orobuccolingual dyskinesia), accompanied by trunk rocking and distal chorea of the hands and feet. In younger patients the picture may be dominated by axial and cranial dystonia (tardive dystonia), a condition which persists and is largely refractory to treatment although is less likely to occur if the offending drug is discontinued within 5 years of being instituted.

Tardive dyskinesias usually appear after at least 6 months' neuroleptic drug therapy, and their incidence increases with exposure to the drugs and also with the age of the patients, although some of the more recent atypical antipsychotic drugs have less of a propensity to cause this condition. Tardive dyskinesias often get worse in the weeks immediately after stopping the offending drug, or may appear then for the first time. After drug withdrawal, tardive dyskinesias disappear in about 60 per cent or more of patients over the next 3 years, but continue unaltered in the remainder. They are difficult to treat and whilst the offending agent ideally should be stopped this is often not possible. Anticholinergic drugs tend to worsen orobuccolingual dyskinesia, but may relieve tardive dystonia. Baclofen may help some patients.

Other drugs that cause dyskinesias include levodopa in patients with Parkinson's disease. It seems likely that most such drug induced dyskinesias are due to pharmacological effects on dopamine mechanisms in the basal ganglia, resulting in dopaminergic overactivity, in contrast to the akinetic-rigid syndrome produced by dopamine depletion or blockade.

---

*We acknowledge the contribution of the late Professor C. D. Marsden to this chapter in the third edition of the textbook. Much of his text provides a basis for this chapter.

Further reading

General

Marsden CD, Fahn S, eds. (1982). *Movement disorders*, vol. I. Butterworth, London.

Marsden CD, Fahn S, eds. (1987). *Movement disorders*, vol. II. Butterworth, London.

Marsden CD, Fahn S, eds. (1994). *Movement disorders*, vol. III. Butterworth, London.

Watts RL, Koller WC (1997). *Movement disorders. Neurologic principles and practice*. McGraw-Hill, New York.

Dystonia

Marsden CD, Quinn NP (1990). The dystonias. *British Medical Journal* **300**, 139–44.

Berardelli, A *et al.* (1998). The pathophysiology of primary dystonia. *Brain* **121**, 1195–212.

Dauer WT *et al.* (1998). Current concepts on the clinical features, aetiology and management of idiopathic cervical dystonia. *Brain* **121**, 547–60.

Nygaard TG, Wooten GF (1998). Dopa-responsive dystonia. *Neurology 5***0**, 853–5.

Warner TT, Jarman P (1998). The molecular genetics of the dystonia. *Journal of Neurology, Neurosurgery and Psychiatry* **64**, 427–9.

Chorea

Harper PS (1996). *Huntington's disease*, 2nd edn. WB Saunders, Philadelphia.

Nausieda PA *et al.* (1980). Sydenham's chorea: an update. *Neurology* **30**, 331–4.

Reddy PH, Williams M, Tagle DA (1999). Recent advances in understanding the pathogenesis of Huntington's disease. *Trends in Neuroscience* **22**, 248–55.

Vidakovic A, Dragasevic N, Kostic VS (1994). Hemiballism: report of 25 cases. *Journal of Neurology, Neurosurgery and Psychiatry* **57**, 945–9.

Tremor

Bain P (1993). A combined clinical and neurophysiological approach to the study of patients with tremor. *Journal of Neurology, Neurosurgery and Psychiatry* **69**, 839–44.

Elble RJ (1986). Physiological and essential tremor. *Neurology* **36**, 225–31.

Hubble JP, Busenbark KL, Koller WC (1989). Essential tremor. *Clinical Neuropharmacology* **12**, 453–82.

Myoclonus and tics

Brown P *et al.* (1991). The hyperekplexias and their relationship to the normal startle reflex. *Brain* **114**, 1903–28.

Fahn S, Marsden CD, Van Woert M (1986). Myoclonus. *Advances in Neurology* **43**, 1–709.

Lees AJ (1987). *Tics*. Churchill Livingstone, London.

Pranzatelli MR (1994). Serotonin and human myoclonus. Rationale for the use of serotonin receptor agonists and antagonists. *Archives of Neurology* **51**, 605–17.

Robertson MM (1989). The Gilles de la Tourette syndrome: the current status. *British Journal of Psychiatry* **154**, 147–69.

Singer HS, Walkup JT (1991). Tourette syndrome and other tic disorders. Diagnosis, pathophysiology and treatment. *Medicine* **70**, 15–32.

Other movement disorders

Barker RA *et al.* (1998). Review of 23 patients affected by the stiff man syndrome: clinical subdivision into stiff trunk (man) syndrome, stiff limb syndrome and progressive encephalomyelitis with rigidity. *Journal of Neurology, Neurosurgery and Psychiatry* **65**, 633–40.

Gershanik OS (1993). Drug-induced movement disorders. *Current Opinions in Neurology and Neurosurgery* **6**, 369–76.

Rajendra S, Schofield PR (1995). Molecular mechanisms of inherited startle syndromes. *Trends in Neuroscience* **18**, 80–2.

Walters AS, group organizer and correspondent (1995). Toward a better definition of the restless legs syndrome. *Movement Disorders* **10**, 634–42.

24.13.12 Ataxic disorders

Nicholas Wood

Introduction

The term 'ataxia', derived from the Greek, means 'irregularity' or 'disorderliness'. Unsteadiness can result from a number of causes, including poor vision, impaired postural reflexes, or a deficiency of sensory input, that is sensory ataxia. This chapter is devoted to the symptoms, signs, and the pathological and clinical features of the disorders of the cerebellum (and its connections). There are two basic clinical rules which can be applied: (1) lesions of the vermis generally cause ataxia of midline structure (that is, truncal and gait ataxia); (2) output from the cerebellar hemisphere is to the contralateral cerebral hemisphere, which provides output to the contralateral limbs, therefore cerebellar hemisphere lesions are ipsilateral. It should, however, be noted that clinical assessment is complicated by the fact that few patients with ataxia have pure cerebellar disease as there is often additional pathology in the brainstem, spinal cord, or elsewhere.

Symptoms of ataxic disorders

The history is extremely important. A clarification of what patients mean should also be sought. Many refer to 'giddiness' or 'dizziness' when they really mean unsteadiness of gait without associated vertigo or light-headedness. The age and speed of onset and development of other features provides important aetiological clues. Rate of progress and any precipitating or relieving factors should be noted. There has been a tremendous improvement in our understanding of the genetic basis of many ataxia disorders and a detailed family history is paramount.

Disturbances of gait

This is the most frequent presenting feature in ataxic disorders. Patients may report an inability to walk in a straight line and a tendency to bump into things. This may be significantly worse in the dark, thus indicating a sensory ataxia. Sudden changes of direction are particularly difficult and problems turning may be reported. The duration of the gait disturbance should be established, and it is worth asking about early motor milestones and athletic ability at school that may bring out a much longer history than previously appreciated. Collateral history should be sought especially if an insidious onset is suspected, as this may be difficult for a patient to report. A question as to diurnal variation particularly a history of morning unsteadiness that wears off later in the day, often associated with morning headache, suggests raised intracranial pressure even if examination is normal.

Limb incoordination and tremor

Clumsiness of the arms is often noted later in the course of their illness. Generally a tremor that is worse on action is reported and as this worsens patients notice clumsiness carrying objects and deterioration of their handwriting. This is more common in multiple sclerosis than in degenerative disease. Disturbance to the midline structures may result in titubation, and this, in combination with action tremor in the upper limbs and little in the way of gait disturbance, should raise the suspicion of Wilson's disease.

Dysarthria

This may be noted by friends and relatives before the patient. Classically described as having a staccato quality, it is a useful symptom or sign as it points against a purely sensory ataxia.

Visual and ocular motor symptoms

Visual symptoms are relatively rare in pure cerebellar disease and if present is more often associated with brainstem disturbance, especially episodic or persistent diplopia associated with ataxia. Vertical oscillopsia suggests downbeat nystagmus, and a structural foramen magnum lesion should be suspected. Acute or subacute oscillopsia, with chaotic involuntary eye movements observed by relatives, may be mentioned in the history of patients with viral cerebellitis, paraneoplastic cerebellar degeneration, and the dancing-eyes syndrome (opsoclonus). There are some very rare degenerative ataxias with gradual visual loss, due to either optic neuropathy or retinopathy.

Other symptoms

Details of any headache or vomiting should be sought, the presence of which may suggest a posterior fossa mass lesion. An acute history suggests cerebellar haemorrhage. If longer standing then a tumour becomes more likely. It should also be remembered that infections, especially an abscess, can cause similar symptoms. Intermittent symptoms and perhaps associated fever and malaise raise the possibility of posterior fossa cysticercosis, and a detailed travel history over the last 20 years should be sought.

Vertigo is more suggestive of neoplastic, inflammatory, and vascular disease rather than the more slowly progressive degenerative processes.

Direct questioning should cover the urinary system, skeletal deformities, cardiac disease, and assessment of cognitive abilities since many ataxias can be associated

with disease in other systems (see Table 1).

A detailed enquiry of drug ingestion (for both medical and recreational purposes, including alcohol) and occupational exposure is also required.

Signs of cerebellar disease

This section covers the examination in the sequence that it appears to the physician.

Gait and posture

A patient walking into the consulting room may have a broad-based gait with a poor turn, and there is often a lurching quality to the overall sequence. More detailed assessment of mild gait ataxia may be obtained by asking the patient to tandem walk (heel–toe). Asking the patient to stand still may reveal the broad base and unless there is additional proprioceptive loss or vestibular disease, this instability is not aggravated by eye closure.

Speech

It is often stated that cerebellar speech is very distinctive with an explosive quality, so-called scanning dysarthria. Although when this is heard it is characteristic, more often a combination of spastic and cerebellar features can be heard. Additional signs such as a slow moving tongue and brisk jaw jerk support the latter.

Muscle tone

Many textbooks state firmly that cerebellar disease gives rise to hypotonia, and some even include it within the symptoms. Not only do patients never complain of hypotonia but this is rarely detectable clinically in symmetrical slowly progressive or chronic disorders. Pendular knee jerks are also difficult to detect without the eye of faith and many patients with 'cerebellar' ataxic disorders have disease of the spinal cord, peripheral nerves, or both, which complicates the clinical picture.

Limb ataxia

Limb ataxia usually results from a combination of dysmetria and dysdiadochokinesis. Dysmetria refers to errors in the range and force of movement resulting in an erratic, jerky movement which may under- or overshoot the target. Dysdiadochokinesis is demonstrated by asking the patient to tap one hand on the other, alternately pronating and supinating the tapping hand, or rapidly opening and closing the fist. The tapping out of simple rhythms (with the hand or foot) is also useful in assessing both the rhythmicity and force of the tap.

Classically, testing of coordination is undertaken after the motor and sensory tests as the presence of weakness or sensory loss can confuse the picture. There is also a natural asymmetry in cerebellar function, with better performance, particularly for rapid alternating movements, in the dominant limb. About 40 per cent of patients with vermis lesions do not have limb ataxia but have striking gait ataxia.

Tremor

Intention tremor is present if a rhythmical side-to-side oscillation is seen on finger-to-nose testing. A combination of gross intention tremor and a postural component is often called rubral or red nucleus tremor, although peduncular tremor is probably a more accurate label. It is most commonly seen in ultiple sclerosis and occasionally in late-onset degenerative ataxias. A nodding head tremor (titubation) with a frequency of 3 to 4 Hz may be seen with midline cerebellar disease.

Eye movements

Square wave jerks may be seen in the primary position; these are inappropriate saccades that disrupt fixation and are followed by a corrective saccade within 200 ms. Assessment of pursuit may see a jerkiness with saccadic intrusions. Additional isolated or multiple lesions of the third, fourth, or sixth cranial nerves suggests brainstem pathology. Examination of the saccadic system can reveal hypo- or hypermetric saccades. An internuclear ophthalmoplegia may be found whilst examining this system, suggesting a diagnosis of multiple sclerosis, but it can rarely be associated with some degenerative ataxias. The vestibulo-ocular reflex (doll's head manoeuvre) should then be examined to look for any supranuclear component. An inability to suppress this reflex is evidence of pathology involving the vestibulocerebellum.

Acute or subacute presentation of almost any of the above eye movements, especially if associated with alcohol abuse or vomiting, raises the possibility of Wernicke's encephalopathy and requires urgent treatment with thiamin.

Gaze-evoked nystagmus is the most common type of nystagmus associated with cerebellar disease; eccentric gaze cannot be maintained, and the slow phase of the nystagmus is toward the primary position, with rapid corrective movements. It does not have much localizing value. Although downbeat nystagmus should raise the suspicion of a foramen magnum lesion, this is also seen in degenerative cerebellar disease.

Positional nystagmus in a patient with vertigo and unsteadiness should be attributed to benign labyrinthine disease only if it is transient, torsional, and fatiguable; if it does not have these features, a posterior fossa lesion should be suspected.

Other neurological signs and general examination

As the causes of ataxia are numerous, a large variety of other neurological and general physical signs may be found on examination. Table 1 lists the various signs and their possible diagnostic significance.

Disorders of the cerebellum

Numerous pathological processes can affect cerebellar function, some of which such as multiple sclerosis and neoplasia are discussed elsewhere. This section will approach the diseases with approximate reference to the time-course (acute, subacute, chronic) and the nature of the course of the disease.

Developmental disorders

The cerebellum has a long developmental period and is not fully mature until about 18 months of age. It is therefore susceptible to a large number of insults, including intrauterine infections, ischaemic damage, toxins, and genetically determined syndromes (see Table 2). Some of these developmental anomalies, such as dysgenesis or agenesis of the vermis, the cerebellar hemispheres, or parts of the brainstem, give rise to congenital ataxia. These are non-progressive disorders, and in most cases coordination improves with age.

Cerebellar dysfunction in an infant or young child may be overlooked, as it often gives rise to a relatively non-specific abnormal motor development. Later there is nystagmus, obvious incoordination on reaching for objects, and truncal ataxia when first attempting to sit. Associated mental retardation is common but unhelpful diagnostically.

Ataxia of acute or subacute onset

Cerebellar ataxia of extremely acute onset has two main causes: cerebellar haemorrhage (usually associated with headache, vertigo, vomiting, altered consciousness, and neck stiffness); and cerebellar infarction (in which cerebellar signs are usually combined with signs of brainstem ischaemia, and the presentation may mimic that of haemorrhage). Diagnosis should be made as a matter of urgency and imaging is required to clarify these two possibilities.

Subacute, reversible ataxia may occur as a result of viral infection in children between 2 and 10 years of age. There is usually pyrexia, limb and gait ataxia, and

dysarthria developing over hours or days. Although recovery occurs over a period of weeks and is usually complete, it can take up to 6 months.

In older patients the possibility of a postinfectious encephalomyelitis, particularly that related to varicella infection, should be considered. The postinfectious Miller Fisher variant of the Guillain-Barré syndrome may present with a triad that includes subacute ataxia, areflexia, and ophthalmoplegia. Nerve conduction studies and examination of cerebrospinal fluid (**CSF**) may be helpful, but the former are often normal. Other infective agents are shown in Table 3. Viral titres and CSF examination may be helpful, although serological evidence of viral infection may be difficult to establish.

Other causes of subacute ataxia include hydrocephalus, foramen magnum compression, posterior fossa tumour (primary or secondary), abscess, or parasitic infection in any age group. A number of important toxins and drugs also need to be considered, including thallium, lead, barbiturates, phenytoin, piperazine, alcohol, solvents, and antineoplastic drugs.

Vascular disorders of the cerebellum

Cerebrovascular disease is dealt with in detail in Chapter 24.13.7. Transient ischaemic attacks involving the vascular supply to the cerebellum rarely produce ataxia and dysarthria alone, usually there are associated symptoms of brainstem dysfunction. Cerebellar infarction (from embolus or, more commonly, vertebrobasilar occlusive disease) and haemorrhage (usually on a background of hypertension or, less commonly, secondary to a vascular malformation or tumour) are relatively rare. Imaging is often necessary for early diagnosis as the later the diagnosis the worse the prognosis. Both infarction and haemorrhage may be amenable to surgical therapy.

Ataxia with an episodic course

These attacks can be considered bizarre and some patients are misdiagnosed as hysterical. However, a good history can usually distinguish between the main causes (listed in order of approximate frequency): drug ingestion, multiple sclerosis, transient vertebrobasilar ischaemic attacks, foramen magnum compression, intermittent obstruction of the ventricular system due to a colloid cyst or cysticercosis, and dominantly inherited periodic ataxia. Autosomal dominant periodic ataxia is characterized by childhood or adolescent onset of attacks of ataxia, dysarthria, vertigo, and nystagmus; not all patients have affected relatives.

There are at least two forms of this disorder: episodic ataxia-1 and -2. Episodic ataxia-1 (**EA1**) is typified by brief attacks (minutes and occasionally hours) and clinically and electrophysiologically myokymia may be seen. Mutations in a potassium-channel gene (*Kv1.1*) have been found. These patients may benefit from treatment with acetazolamide or phenytoin. Patients tend to be neurologically normal between the attacks. In episodic ataxia-2 (EA2) the attacks tend to be longer lasting, hours or even days, usually associated with vertigo and consequent nausea and vomiting. The attacks tend to be more severe in childhood with associated drowsiness, headache, and fever. Although when the disease first begins the patients are well between attacks, an interictal nystagmus can be seen. A slow deterioration in the ataxia is seen as the disease progresses. MRI may reveal cerebellar atrophy. These patients tend to respond better to acetazolamide therapy than patients with EA1. Point mutations in a calcium-channel gene (*CACNA1A*) have been demonstrated in some families with this disorder.

In children and young adults a metabolic disorder should be suspected, particularly defects of the urea cycle, aminoacidurias, Leigh's syndrome, and mitochondrial encephalomyopathies. Screening investigations include blood ammonia, pyruvate, lactate and amino acids, and urinary amino acids.

Ataxia with a chronic progressive course

Chronic alcohol abuse is the commonest causes of progressive cerebellar degeneration in adults. Thiamin deficiency is probably the main (but not sole) explanation for the chronic progressive cerebellar syndrome found in alcoholics. Patients with this syndrome are almost invariably malnourished. Ataxia may develop during periods of abstinence, and identical cerebellar degeneration has been observed in non-alcoholic patients with severe malnutrition. Cerebellar ataxia is common in the Wernicke–Korsakoff syndrome, and the pathological features of both this syndrome and a cerebellar degeneration frequently coexist. With administration of thiamin some improvement may occur in early cases of alcoholic cerebellar degeneration, but if the patient is already chair-bound the response to treatment is limited.

Other deficiency disorders can give rise to a progressive ataxia. There is a rare syndrome associated with zinc deficiency which responds to oral replacement therapy. Deficiency of vitamin E, either genetic (for example, isolated vitamin E deficiency due to mutations in the a-tocopherol transfer protein, or abetalipoproteinaemia) or acquired, due to malabsorption, may also produce a progressive ataxia. Establishing the diagnosis of vitamin E deficiency is important as treatment with vitamin E may prevent progression of the neurological syndrome and can, in rare circumstances, lead to some improvement.

A number of toxic agents can produce progressive cerebellar dysfunction, including pharmaceutical products, solvents, and heavy metals. The most common cause of a cerebellar syndrome due to drug toxicity seen in neurological practice is that associated with anticonvulsant medication, particularly phenytoin. Transient ataxia, dysarthria, and nystagmus usually develop when serum concentrations of phenytoin, carbamazepine, or barbiturates are above the therapeutic range, and remit when they return to within the therapeutic range. Chronic phenytoin toxicity is reported to cause persistent cerebellar dysfunction, and this is associated pathologically with a loss of Purkinje cells. A persistent cerebellar deficit, with dysarthria and limb and gait ataxia and cerebellar atrophy on CT scan, has also been described as a sequel to the acute encephalopathy of lithium toxicity that is usually precipitated by fever or starvation. Serum lithium levels are not always raised in such cases.

Recreational or accidental exposure to a number of solvents, including carbon tetrachloride and toluene, causes cerebellar ataxia along with other neurological problems, including psychosis, cognitive impairment, and pyramidal signs in the case of toluene. The neurological deficit is potentially reversible but may persist after prolonged exposure in solvent abusers. Exposure to heavy metals including inorganic mercury, lead, and thallium can also produce cerebellar damage.

Structural lesions such as posterior fossa tumours, foramen magnum compression, or hydrocephalus must be excluded by imaging studies. Tumours which may involve the posterior fossa include: astrocytoma, ependymoma, haemangioblastoma, and cranial nerve neuromas.

Paraneoplastic cerebellar degeneration related to carcinomas of the lung or ovary or to the reticuloses usually follows a subacute course, with patients losing the ability to walk within months of onset. A variety of antineuronal antibodies may be found in these patients and help to confirm the diagnosis. Approximately half of patients with paraneoplastic cerebellar degeneration (**PCD**) have demonstrable antibodies directed against neurones in their serum and CSF. The most common antibody seen in PCD is called anti-Yo, and it specifically stains Purkinje cell cytoplasm. If antineuronal antibodies are detected then a search for the underlying malignancy should then be undertaken involving imaging and analysis of tumour markers. Presentation with ataxia precedes diagnosis of the malignancy in 70 per cent of cases and is usually subacute, progressing to severe disability over several months or even weeks, and then arresting. Onset may be acute and is sometimes accompanied by vertigo, mimicking a vascular event. There is severe truncal, gait, and limb ataxia and dysarthria. Opsoclonus may be combined with myoclonus, producing a disorder in adults similar to the dancing eyes syndrome of childhood, and which is sometimes associated with neuroblastoma. There is currently no evidence of a useful response either to immunosuppressant therapy or to plasma exchange. However, there are anecdotal reports of some improvement or stabilization following removal of the primary tumour. The best method of screening for the underlying malignancy is debated, but standard magnetic resonance imaging (**MRI**) may be complemented by whole-body positron emission tomography (**PET**) scanning. Searching for primary tumour markers may also be useful.

Rarely, infectious agents can cause slowly progressive ataxia (see Table 3), these include the chronic panencephalitis of congenital rubella infection in children and, in adults, Creutzfeldt–Jakob disease (**CJD**), the iatrogenic form of which should be particularly considered. A specific enquiry regarding potential risk-factor exposure should be sought, especially growth-hormone replacement. It is now known that the so-called variant form of CJD may also cause ataxia, often in association with psychiatric disturbance. Multiple sclerosis only exceptionally presents as an isolated chronic progressive cerebellar syndrome.

Some conditions that are not generally considered primarily as ataxic disorders may present with clumsiness, tremor, or definite cerebellar signs, particularly in childhood or adolescence. These include Wilson's disease and several inherited neuropathies, such as hereditary motor and sensory neuropathy (**HMSN**; Charcot–Marie–Tooth disease, including the so-called Roussy–Levy syndrome). Although intention and postural tremor are quite frequent in the demyelinating type of HMSN (type I), dysarthria and pyramidal signs do not occur. Other chronic demyelinating neuropathies, such as chronic inflammatory and paraproteinaemic neuropathies and Refsum's disease, may give rise to prominent tremor and ataxia; the same applies to giant axonal neuropathy.

Superficial siderosis is a rare disorder that causes slowly progressive cerebellar ataxia, mainly of gait, and sensorineural deafness, often combined with spasticity, brisk reflexes, and extensor plantar responses. The diagnosis may not be suspected clinically, but the neuroradiological abnormalities are striking, MRI showing a black rim of haemosiderin around the posterior fossa structures and spinal cord, and less often the cerebral hemispheres, on *T*2-weighted images. Superficial siderosis is most commonly secondary to chronic leaking of blood into the subarachnoid space. Treatment relies on identifying the source of bleeding; chelation

therapy does not appear to be effective.

After excluding acquired causes of ataxic disorders, there remains a considerable number of patients with degenerative ataxias, not all of which are overtly genetically determined. The inherited ataxias can largely be classified according to their clinical and genetic features (see below), and in a small proportion of cases a recognizable metabolic defect can be detected. It is important to make as accurate a diagnosis as possible in these disorders for the purposes of prognosis, genetic counselling, and, occasionally, specific therapy.

Progressive metabolic ataxias

Ataxia may be a minor feature of storage and other metabolic neurodegenerative disorders developing in early childhood. Some enzyme deficiencies that usually give rise to diffuse neurodegenerative disorders in which ataxia is a feature, either developing in infancy or early childhood, include the sphingomyelin lipidoses, metachromatic leucodystrophy, galactosylceramide lipidosis (Krabbe's disease), and the hexosaminidase deficiencies. Also included within this group is adrenoleucomyeloneuropathy, a phenotypic variant of adrenoleucodystrophy. This is diagnosed by estimation of very long-chain fatty acids. Although X-linked, approximately 10 per cent of carrier females may manifest neurological abnormalities. The role of diet and dietary supplements (for example, oleic acid and Lorenzo's oil) remains to be established. Ataxia may be prominent in Niemann–Pick disease type C (juvenile dystonic lipidosis), combined with a supranuclear gaze palsy. Sphingomyelinase activity is normal, but foamy storage cells are found in the bone marrow.

Cholestanolosis (also called cerebrotendinous xanthomatosis, **CTX**) is a rare autosomal recessive disorder caused by defective bile salt metabolism, due to a deficiency of mitochondrial sterol 27-hydroxylase. It gives rise to ataxia, dementia, spasticity, peripheral neuropathy, cataract, and tendon xanthomas in the second decade of life. Treatment with chenodeoxycholic acid appears to improve neurological function.

Various phenotypes classifiable as hereditary ataxias have been described in the mitochondrial encephalomyopathies, many of which are associated with a defect of mitochondrial DNA. These include late-onset ataxic disorders associated (for example, the Kearns–Sayre syndrome) with such features as dementia, deafness, and peripheral neuropathy. These features overlap with the syndrome of progressive myoclonic ataxia, which may also be caused by ceroid lipofuscinosis, sialidosis, and Unverricht–Lundborg's disease or so-called Baltic myoclonus. Most of these disorders can now be distinguished with appropriate gene tests or enzyme estimations.

Acquired metabolic and endocrine disorders causing cerebellar dysfunction

These include hepatic encephalopathy, pontine and extrapontine myelinolysis related to hyponatraemia, and hypothyroidism. The last of these is only very rarely a cause of a cerebellar syndrome in both children and adults.

Ataxic disorders associated with defective DNA repair

There are a number of rare conditions associated with a reduced capacity to perform excision repair of DNA damaged by ultraviolet light and some chemical carcinogens. The commonest is ataxia telangiectasia (**AT**). Clinically related conditions include xeroderma pigmentosum and Cockayne's syndrome. Characteristically, motor development is often delayed and ataxia noted at the time of first walking. Growth retardation and delayed sexual development are frequent, and there is mild mental retardation in some cases. A mixed movement disorder may be seen, often with a combination of ataxia, dystonia, and chorea. The cutaneous telangiectasia of AT tend to develop on the conjunctivas between the ages of 3 and 6 years, but occasionally are inconspicuous or absent in adult life. Ataxia telangiectasia is associated with abnormalities of both humoral and cell-mediated immunity. The gene for AT has now been cloned and is called *ATM*.

Degenerative disorders

The degenerative cerebellar and spinocerebellar disorders are a complex group of diseases, most of which are genetically determined. In some there is an underlying metabolic disorder, and it is important to diagnose these as there may be important implications for treatment and genetic counselling. There has been a rapid growth in our knowledge of the genetic basis of many of the spinocerebellar degenerations. The next phase will be to understand how these genes and the abnormal proteins they produce cause cell-specific neuropathology. Inherited ataxic disorders can be divided according to their mode of inheritance (Table 4 and Table 5). Most autosomal recessive disorders are of early onset (before 20 years of age), while autosomal dominant disorders are usually of later onset (over 20 years of age).

Autosomal recessive ataxias

Friedreich's ataxia is the most common of the autosomal recessive ataxias (see Table 4), accounting for at least 50 per cent of cases of hereditary ataxia in most large series reported from Europe and the United States. The prevalence of the disease in these regions is similar, between 1 and 2 per 100 000.

The onset of symptoms, generally with gait ataxia, is usually between the ages of 8 and 15 years. However, an onset between 20 and 30 years of age, but fulfilling all other diagnostic criteria, have been described. In addition to the progressive ataxia, a number of variable features are seen, including dysarthria and pyramidal tract involvement. Initially, this latter feature may be mild, with just extensor plantar responses, but invariably a pyramidal pattern of weakness in the legs is seen after 5 or more years' duration of the disease. Eventually this can lead to paralysis. Distal wasting, particularly in the upper limbs, is seen in about 50 per cent of patients with Friedreich's ataxia. Skeletal abnormalities are also commonly found, including scoliosis (85 per cent) and foot deformities typically pes cavus, in approximately 50 per cent of patients. Additional clinical support for a suspicion of Friedreich's ataxia include optic atrophy, which can be seen in 25 per cent of patients; however, it is rare (<5 per cent) for Friedreich's ataxia to produce major visual impairment. Deafness is found in less than 10 per cent of cases, but rather more have impairment of speech discrimination. Nystagmus is seen in only about 20 per cent, but the extraocular movements are nearly always abnormal, with broken-up pursuit, dysmetric saccades, square-wave jerks, and failure of fixation suppression of the vestibulo-ocular reflex.

Investigation of patients reveals an axonal sensory neuropathy; an abnormal ECG in 65 per cent of patients with widespread T-wave inversion. Diabetes mellitus occurs in 10 per cent of patients with Friedreich's ataxia, and a further 10 to 20 per cent have impaired glucose tolerance.

The gene encoding frataxin (*X25*) was cloned in 1996. The predominant mutation is a trinucleotide repeat (GAA) in intron 1 of this gene. Expansion of both alleles is found in over 96 per cent of patients. The remaining patients have point mutations in the frataxin gene. This was the first autosomal recessive condition found to be due to a dynamic repeat and, as it is a relatively simple matter to measure the repeat size, it has permitted the introduction of a specific and sensitive diagnostic test. On normal chromosomes the number of GAA repeats varies from 7 to 22 units, whereas on disease chromosomes the range is anything from around 100 to 2000 repeats. The length of the repeat is a determinant of the age of onset and therefore to some degree influences the severity, in that early-onset cases tend to progress more rapidly.

There is accumulating evidence that frataxin is mitochondrially located and may be involved in iron transport. Clinically, this fits; a syndrome of ataxia and neuropathy, in association with diabetes, cardiomyopathy, deafness, and optic atrophy, has the hallmarks of a mitochondrial disease.

Other autosomal recessive ataxias are individually rare and are listed in Table 4.

Autosomal dominant ataxias

The autosomal dominant cerebellar ataxias (**ADCAs**) are a clinically and genetically complex group of neurodegenerative disorders (see Table 5). ADCA type I is characterized by a progressive cerebellar ataxia and is variably associated with other extracerebellar neurological features such as ophthalmoplegia, optic atrophy, peripheral neuropathy, and pyramidal and extrapyramidal signs. The presence and severity of these signs is, in part, dependent on the duration of the disease. Although mild or moderate dementia may occur, it is usually not a prominent early feature. ADCA type II is clinically distinguished from the ADCA type I by the presence of pigmentary macular dystrophy, whereas ADCA type III is a relatively 'pure' cerebellar syndrome and generally starts at a later age. This clinical classification is still useful, despite the tremendous improvements in our understanding of the genetic basis (see below), because it provides a framework which can be used in the clinic and helps direct the genetic evaluation.

The genetic loci causing the dominant ataxias are given the acronym *SCA* (spinocerebellar ataxia). At the time of writing 17 *SCA* loci have been identified. Of these, the genes are established for *SCA1*, *-2*, *-3*, *-6*, *-7*, *-10*, *-12*, and *-17*. The last three are all extremely rare. *SCA12*, *-36*, and *-7* are all caused by a similar mutational

mechanism, an expansion of an exonic CAG repeat. The resultant proteins all possess an expanded polyglutamine tract, and at least eight conditions caused by these expansions are now known. Other types of ADCA are exceedingly rare.

Idiopathic degenerative late-onset ataxias

About two-thirds of cases of degenerative ataxia developing over the age of 20 years are singleton cases, and they represent a significant clinical problem; it is difficult even to know how to label them. The literature is confusing, mixing pathological terms such as 'olivopontocerebellar atrophy' (**OPCA**) with clinical terms, I prefer to use the term 'idiopathic late-onset cerebellar ataxia' (**ILOCA**). A proportion of patients in this group, progress to develop the features of multiple system atrophy (**MSA**). These patients may have, or develop, facial impassivity and extrapyramidal rigidity, whilst others present with features of autonomic failure such as postural hypotension, impotence, bladder dysfunction, and a fixed cardiac rate. A cerebellar presentation occurs in about 15 per cent of patients with MSA. The distinction of idiopathic late-onset cerebellar ataxia from MSA may therefore be difficult clinically at presentation.

Most patients with idiopathic late-onset cerebellar ataxia lose the ability to walk independently between 5 and 20 years after onset, and their lifespan is slightly shortened by immobility. Those who go on to develop MSA have a particularly poor prognosis. Investigations, apart from those excluding acquired causes of cerebellar degeneration such as malignancy and hypothyroidism, tend to be unhelpful. Electrophysiological evidence of a sensory peripheral neuropathy is found in about 50 per cent of cases, which can be a useful pointer to the presence of a degenerative multisystem disorder. CT or MRI scan may show cerebellar and brainstem atrophy, or pure cerebellar atrophy. The prognosis is worse in patients with clinical and radiological evidence of brainstem involvement, compared to those with a pure cerebellar syndrome and cerebellar atrophy alone on MRI.

Further reading

Anderson NE, Rosenblum MK, Posner JB (1988). Paraneoplastic cerebellar degeneration: clinical-immunological correlations. *Annals of Neurology* **24**, 559–67.

Bootsma D, *et al.* (2001). Nucleotide excision repair syndromes: xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy. In: Scriver CR, *et al.*, eds. *The metabolic basis of inherited disease*, 8th edn, pp 245–74. McGraw Hill, New York.

Campuzano V, *et al*. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**, 1423–7.

De Michele G, *et al*. (1989). Late onset recessive ataxia with Friedreich's disease phenotype. *Journal of Neurology, Neurosurgery and Psychiatry* **52**, 1398–403.

Enevoldson PG, Sanders MD, Harding AE (1994). Autosomal dominant cerebellar ataxia with pigmentary macular dystrophy: a clinical and genetic study of eight families. *Brain* **17**, 445–60.

Fearnley JM, Stevens JM, Rudge P (1995). Superficial siderosis of the central nervous system. *Brain* **118**, 1051–66.

Hanna MG, Wood NW, Kullmann D (1998). The neurological channelopathies. *Journal of Neurology, Neurosurgery and Psychiatry* **65**, 427–31.

Harding AE (1981). Friedreich's ataxia: a clinical and genetic study of 90 families with an analysis of early diagnostic criteria and intrafamilial clustering of clinical features. *Brain* **104**, 589–620.

Harding AE (1984). *The hereditary ataxias and related disorders*. Churchill Livingstone, Edinburgh.

Harding AE, Diengdoh JV, Lees AJ (1984). Autosomal recessive late-onset multisystem disorder with cerebellar cortical atrophy at autopsy: report of a family. *Journal of Neurology, Neurosurgery and Psychiatry* **47**, 853–6.

Klockgether J, *et al*. (1990). Idiopathic cerebellar ataxia of late onset: natural history and MRI morphology. *Journal of Neurology, Neurosurgery and Psychiatry* **53**, 297–305.

Klockgether T, *et al.* (1998). The natural history of degenerative ataxia: a retrospective study in 466 patients. *Brain* **121**, 589–600.

Marsden CD, *et al.* (1990). Progressive myoclonic ataxia (the Ramsay Hunt syndrome). *Archives of Neurology* **47**, 1121–5.

Muller DP, Lloyd JK, Wolff OH (1983). Vitamin E and neurological function. *Lancet* **i**, 225–8.

Peterson K, *et al*. (1992). Paraneoplastic cerebellar degeneration. I. A clinical analysis of 55 anti-Yo antibody-positive patients. *Neurology* **42**, 1931–7.

Quinn NP, Marsden CD (1993). The motor disorder of multiple system atrophy. *Journal of Neurology, Neurosurgery and Psychiatry* **56**, 1239–42.

Stewart GE, Ironside JW (1998). New variant Creutzfeldt–Jakob disease. *Current Opinion in Neurology* **11**, 259–62.

Woods CG, Taylor AMR (1992). Ataxia telangiectasia in the British Isles: the clinical and laboratory features of 70 affected individuals. *Quarterly Journal of Medicine* **298**, 169–79.

24.13.13 The motor neurone diseases

Michael Donaghy

Introduction

The motor neurone diseases result from selective loss of function of the lower and/or upper motor neurones controlling the voluntary muscles of the limbs or bulbar region. The term 'motor neurone disease' is best used to describe a family of diseases within which there is extensive differential diagnosis; in the past the term has been used synonymously with amyotrophic lateral sclerosis, one of the most serious of these diseases. Precise diagnosis is essential for advising patients about prognosis, for identifying those diseases with genetic implications, and to offer immunosuppressant therapy to patients with some acquired lower motor neurone syndromes.

In practice, differential diagnosis requires clinical and electrophysiological classification as to whether the disease involves the upper or the lower motor neurones, or both. This anatomical differentiation is augmented by the age of onset, the rate of deterioration, and familial occurrence (Table 1). Sensation and cognition are normal on simple clinical assessment in the motor neurone diseases.

The clinical signs of lower motor neurone involvement consist of muscle wasting, fasciculations, and flaccid weakness. The tendon reflexes are often retained until profound denervation or fibrous replacement have affected the muscle. Fasciculations are visible flickerings within the muscle belly which are insufficient to produce movement around the joint; electromyography shows that they correspond to simultaneous discharge of all the muscle fibres within a diseased motor unit. Nerve conduction studies will exclude peripheral neuropathy. Electromyography helps to distinguish denervation from myopathy. Muscle biopsy is often required to exclude myopathy, particularly in syndromes causing slowly progressive proximal limb weakness, and bulbar weakness, such as inclusion body myositis.

Upper motor neurone involvement produces spasticity, clonus, extensor plantar responses, and weakness. Extensor plantar responses or clonus may be obscured by coexisting leg muscle atrophy. The abdominal reflexes are often preserved in motor neurone diseases involving the upper motor neurones. This contrasts with their loss in spinal cord disease due to tumours, compression, or demyelinating disease. Sphincter control and sexual function are usually preserved in motor neurone disease, although trunk and abdominal wall weakness may make excretion slow and awkward.

Motor neurone diseases are incurable for the most part and therefore treatment must aim to overcome, or minimize, the various sources of disability. Malnutrition due to dysphagia can be circumvented by nasogastric tube feeding or percutaneous endoscopic gastrostomy. Various forms of assisted respiration offset respiratory muscle weakness, including continuous positive airways pressure via a facial mask. Limb spasticity can be reduced by baclofen, dantroline, or diazepam. Wheelchairs and arm appliances may overcome inadequate limb function. Electronic communication devices should be supplied to those whose speech is incomprehensible. Amitriptyline may help contain the embarrassing emotional lability of pseudobulbar palsy. Housing and workplace modifications can allow patients to maintain independence despite their disability.

Combined upper and lower motor neurone syndromes

Amyotrophic lateral sclerosis

Amyotrophic lateral sclerosis occurs worldwide, usually with an incidence of 1 to 1.5 per 100 000 population and a prevalence of 4 to 6 per 100 000. It is commoner in men and the incidence increases with advancing age; it is unusual before the fifth decade of life. The cause of the common sporadic form of amyotrophic lateral sclerosis is quite unknown. Its incidence is particularly high in areas of the Western Pacific, particularly in Guam and the Japanese Kii Peninsula where it tends to occur in younger adults and can be associated with dementia or parkinsonism. Autosomal dominant inheritance is evident in approximately 5 per cent of patients with adult onset amyotrophic lateral sclerosis. Roughly 20 per cent of familial amyotrophic lateral sclerosis is associated with the 40 different missense mutations of the Cu/Zn superoxide dismutase (*SOD1*) gene on chromosome 21 which catalyses conversion of toxic superoxide anion radicals to hydrogen peroxide. The disease associated with various *SOD1* mutations shows varying degrees of penetrance and a variable phenotype. It tends to begin earlier in adulthood and may involve minor sensory symptoms. A rare juvenile onset autosomal recessive form of amyotrophic lateral sclerosis with prominent bulbar involvement has been described from North Africa and a rare juvenile onset autosomal dominant form without bulbar involvement occurs in the United States.

Pathology

Lower motor neurones are lost from clinically affected areas of the spinal cord and brainstem. Surviving neurones may show intracytoplasmic inclusions (Bumina bodies) and proximal axonal accumulations of neurofilaments (spheroids). The motor cortex is depleted of Betz cells and the pyramidal tracts degenerate. It is becoming increasingly recognized that other populations of neurones can also degenerate in motor neurone disease even though this is not usually evident clinically. These include peripheral sensory neurones and Clarke's column neurones. Up to 10 per cent of patients may eventually develop a mild dementia, often of frontal lobe type. These recent findings show that amyotrophic lateral sclerosis is either a generalized neurodegenerative disorder, in which the motor neurones take the vast brunt of the disease, or that it sometimes overlaps with other neurodegenerations. However, for the practical purpose of clinical diagnosis early in the disease, amyotrophic lateral sclerosis should be regarded as having purely motor manifestations.

Clinical features

At presentation, patients either have bulbar or spinal symptoms, although both usually become evident as the disease progresses.

The bulbar form causes dysphagia, dysphonia, and inhalation of foodstuffs due to weakness of the tongue, pharynx, and larynx. The tongue is wasted, weak, and fasciculating, palatal movements are reduced, and the ability to cough explosively is lost due to vocal cord paralysis. This bulbar palsy is usually accompanied by, or even preceded by, varying degrees of pseudobulbar involvement. The tongue is spastic and immobile with 'hot potato' speech and difficulty in inhibiting emotional responses such as laughing or crying. Ventilatory respiratory failure may develop due to weakness of the diaphragm and intercostal muscles. Occasionally, amyotrophic lateral sclerosis can present with dyspnoea. Diaphragm weakness can be detected clinically by noting that the upper abdomen is drawn inwards, rather than outwards, during the second half of inspiration. Furthermore, the forced vital capacity is substantially lower when the patient is lying down compared with

standing, because the weight of the liver no longer assists diaphragmatic descent.

The spinal form of amyotrophic lateral sclerosis usually presents with wasting and weakness of one limb, usually as intrinsic hand muscle wasting or foot drop. Occasionally the initial weakness predominantly affects the musculature of the shoulder girdle. Asymptomatic involvement of other limbs is often evident on examination. It is diagnostically important to demonstrate combined upper and lower motor neurone signs in at least two limbs. Wasted fasciculating muscles also exhibiting clonus or hyper-reflexia are a helpful finding. With time the limbs become useless due to progressive denervation. Patients become wheelchair- or bedbound, or unable to use their arms for grooming or feeding. Despite enforced recumbency, decubitus ulcers are relatively unusual because autonomic regulation of skin blood flow and secretion is unaffected. Sphincter control is not affected, although practical difficulties in excretion may result from immobility and because abdominal wall weakness prevents the exertion of intra-abdominal pressure.

Prognosis

Amyotrophic lateral sclerosis progresses relentlessly, both in the severity and the extent of muscular involvement. Death commonly results from ventilatory respiratory failure, from choking, or from inhalational pneumonia; malnutrition often contributes. The median survival from first symptoms in those with bulbar onset is approximately 20 months, with only 5 per cent surviving 5 years. The alternative diagnosis of X-linked bulbospinal neuronopathy should be considered in these long survivors. The median survival for those with spinal onset is approximately 29 months with nearly 15 per cent surviving 5 years. Although a subacute and reversible syndrome resembling spinal amyotrophic lateral sclerosis has been described, this is so extraordinarily rare that it should not influence the physician's prognostications.

Differential diagnosis and investigation

A diagnosis of amyotrophic lateral sclerosis is usually depressingly obvious on simple clinical grounds. Often only electrophysiological investigation is necessary to confirm denervation and to exclude a potentially treatable myopathy or demyelinating neuropathy. Sometimes upper motor neurone involvement is not clinically demonstrable, particularly in patients with absent Babinski responses due to severely denervated toe extensor muscles. Unfortunately measurement of central motor conduction following electromagnetic stimulation of the brain is less reliable for revealing upper motor neurone involvement in such cases than had been hoped. If patients present with the combination of arm denervation and upper motor neurone signs in the legs, the cervical spinal canal should be imaged with magnetic resonance scanning to exclude a compressive lesion, most often spondylitic radiculomyelopathy.

The usual diagnostic problem lies in differentiating amyotrophic lateral sclerosis from other motor neurone diseases. A lack of upper motor neurone involvement should raise the possibility of alternative diagnoses. The post-polio syndrome causes slow deterioration in limb or bulbar function some decades after acute poliomyelitis. X-linked bulbospinal neuronopathy is much more slowly progressive than bulbar amyotrophic lateral sclerosis; grimacing usually evokes characteristic lower facial contractions; gynaecomastia, diabetes mellitus, or abnormal sensory nerve conduction are often evident; and other male family members may be affected. Multifocal motor neuropathy or neuronopathy usually develops insidiously, characteristically produces marked weakness with little wasting, predominantly affects the arms, may be associated with paraproteinaemia or antiganglioside antibodies, and may involve motor nerve slowing or conduction block. Adult onset proximal hereditary motor neuronopathy is very slowly progressive, with early and symmetric involvement of proximal muscles, and rarely involves bulbar muscles.

Giving the diagnosis

Doctors or relatives are sometimes tempted on compassionate grounds not to tell patients about their diagnosis of amyotrophic lateral sclerosis. But when patients eventually detect this conspiracy of secrecy it can lead to serious loss of trust at a time when death looms and trustworthy relationships are of inestimable value. When given the opportunity, patients usually indicate that they wish to know the name of the disease and the likely outcome, and they may even wish a detailed discussion of likely modes of death. Of course questions should be answered honestly, although sometimes it may be preferable to discuss them in stages with the spouse present so as to soften the blow early on in the course of the disease. Once a patient has been told the diagnosis, the doctor must address the particular issues presented by that patient's own brand of amyotrophic lateral sclerosis before they become upset by the summary information which they may glean from lay reference books, journalism, or the Internet.

Treatment

No treatment is known to cure amyotrophic lateral sclerosis. Trials of drug therapy have concentrated upon slowing the downhill progression of disability or improving survival. The antiglutamate agent riluzole, administered orally, has been licensed for treatment of amyotrophic lateral sclerosis. The 100 mg dosage improved the chance of tracheostomy-free survival at 18 months by an extra 35 per cent although there was no significant benefit on muscle function. Criticisms of this study have included the nature of the Cox model statistical adjustment, and it should be noted that more of the placebo group had bulbar features at entry to the study. Riluzole is generally well tolerated by patients; nausea, gastrointestinal upset, and raised transaminase enzyme levels may occur and usually resolve with reduction in dose. Ineffective therapeutic trials have included mixtures of branched chain amino acids, dextromorphan, total lymphoid irradiation, and the free radical scavenger acetylcysteine.

Much can be done to overcome disability and alleviate distress by the care team of speech therapist, physiotherapist, occupational therapist, social worker, and physician. The Motor Neurone Disease Association is often able to provide equipment promptly. Severe dysphagia is most effectively bypassed by percutaneous endoscopic gastrostomy. Preferably the patient or their carer should have good hand function and vision so that they can change nutrient bags at home. If video-swallow shows that cricopharyngeal spasm is responsible for dysphagia, cricopharyngeal myotomy may help. Speech failure can be circumvented by computer-assisted communication devices operated through a practical modality, such as pressure, blowing, head nodding, or blinking depending upon which muscles remain strong.

Decisions regarding the advisability of instituting assisted respiration pose complex practical and ethical dilemmas. Patients with diaphragm weakness and nocturnal dyspnoea may be helped by continuous positive airways pressure delivered by a facial mask. Endotracheal intubation and ventilation are rarely recommendable in a disease causing such ubiquitous irreversible weakness.

Lower motor neurone syndromes

These forms of motor neurone disease generally follow a much more benign course than amyotrophic lateral sclerosis. They include syndromes previously described as spinal muscular atrophy and progressive muscular atrophy. Differential diagnosis within the lower motor neurone syndromes depends principally upon attention to the age of onset, the pattern of the weakness, and a possible family history.

Proximal hereditary motor neuronopathy

Acute infantile form (Werdnig–Hoffmann disease)

This is one of the commonest fatal autosomal recessive disorders of children. The disease frequency of approximately 1 in 25 000 in England results from a gene frequency of 1 in 160. Acute infantile spinal muscular atrophy has been linked to chromosome 5q11.2–13.3. Within this region two candidate genes have been isolated, *SMN* (survival motor neurone) and *NAIP* (neuronal apoptosis inhibitory protein). Mutations in these genes occur in up to 98 per cent (*SMN*) and 20 to 50 per cent (*NAIP*) of patients. Although a valuable aid to diagnosis, and potentially for prenatal diagnosis, it should be noted the *SMN* gene mutations occasionally occur in healthy relatives of an affected proband, and that these same mutations, particularly *SMN*, are also found in milder or later onset forms of spinal muscular atrophy including Kugelberg–Welander disease.

Before the age of 6 months, babies become inactive, weak, hypotonic, feed poorly, and are slow to attain motor milestones. They may be born with limb deformities, and in retrospect, fetal movements have been often absent or sparse. The tongue is weak and may fasciculate. Head control is poor and the infant's areflexic and proximally wasted limbs tend to assume a frog-like position. Respiratory movements are decreased with prominent involvement of intercostal muscles. Half the infants die by 6 months, and almost all have succumbed by 18 months, usually to respiratory complications.

Chronic childhood form (Kugelberg–Welander disease)

This form develops at any time from infancy to the early teens. It is also autosomal recessive, may be genetically heterogeneous, and may commence discordantly within families. It may resemble Werdnig–Hoffmann disease if the onset is early, but follows a comparatively benign course. More than 90 per cent of patients are able to walk or to sit unsupported at some time, although these abilities are often lost eventually. Tongue involvement occurs in only half, and significant dysphagia is unusual. Some patients develop respiratory insufficiency as a result of intercostal muscle involvement. The proximal limb weakness and wasting is only slowly progressive and may stabilize spontaneously. Those with severe early weakness often develop secondary spinal and joint deformities. The prognosis varies, although survival into middle age is usual. It is important, although initially difficult, to differentiate those with infantile onset and no family history from Werdnig–Hoffmann disease.

Adult onset forms

The autosomal recessive adult form starts from 15 to 60 years of age, usually in the fourth decade. Slowly progressive proximal limb weakness ensues, but significant disability for walking does not usually occur until the sixth or seventh decade. Life expectancy is only slightly reduced. Distal muscles can be involved too, the tendon reflexes are usually lost, but bulbar involvement is uncommon. The lack of upper motor neurone signs or of bulbar involvement, and the rather indolent progression, distinguish this from amyotrophic lateral sclerosis.

Autosomal dominant forms are rare, and fall into two groups with onset in childhood and in early middle age respectively. The limb weakness is predominantly proximal. Bulbar involvement does occur, although it is unusual. The childhood form may stabilize at adolescence and some patients retain walking ability into middle or old age. The adult onset form causes more severe disability. The lack of upper motor neurone signs distinguishes these conditions from hereditary amyotrophic lateral sclerosis.

X-linked recessive bulbospinal neuronopathy (Kennedy syndrome)

This disorder occurs only in men, with onset in the third to fifth decades of life. It is due to a mutation causing CAG (cytosine–adenine–guanine) repeat sequences of increased length within the androgen receptor gene. Molecular genetic analysis now forms the basis of a diagnostic test. Weakness usually first affects hand or pelvic girdle muscles and the bulbar symptoms may not be evident until 20 years later, if at all. Fasciculations are usually visible in the limb, tongue, and facial muscles. Characteristically, muscle contractions around the chin are induced by pursing the lips or grimacing. The disorder is only slowly progressive. Most patients survive into their seventh or eight decades except when bulbar involvement is unusually severe. The disorder is often misdiagnosed as amyotrophic lateral sclerosis until the unusually slow deterioration is questioned. Unlike amyotrophic lateral sclerosis, there are no upper motor neurone signs and patients commonly show gynaecomastia, diabetes mellitus, and absent sensory nerve action potentials.

Hexosaminidase deficiency

Autosomal recessive GM2 gangliosidosis presents a variable neurological picture, occasionally as a pure motor neurone syndrome due to lower and, rarely, upper motor neurone involvement. More usually there are also other neurological abnormalities such as cerebellar ataxia or dementia. Hexosaminidase assays should be reserved for those patients with early onset of unusual motor neurone disorders, particularly in Ashkenazi Jews.

Hereditary bulbar palsy of infancy and childhood

The Brown–Violetto–van Laere syndrome presents in the teens with bilateral sensorineural deafness, followed some years later by bulbar, facial, limb, and sometimes respiratory muscle weakness. Fazio–Londe disease is an autosomal recessive bulbar palsy of childhood, without deafness, and respiratory muscle involvement may lead to death within a few years.

Monomelic, focal, or segmental motor neuronopathies

This condition is also known as chronic asymmetric or focal spinal muscular atrophy, or monomelic motor neurone disease. Although most commonly described from Asia, especially Japan, it is seen regularly elsewhere in the world. It usually occurs sporadically and most patients are young adult males. It presents with distal wasting and weakness of one hand or forearm. This progresses steadily for the first 2 years before either stabilizing, or settling to a slow rate of subsequent progression. Initially there may be concern that this is the first presentation of amyotrophic lateral sclerosis, but the expected upper motor neurone and bulbar involvement fail to materialize, and spread to other limbs is unusual. Nerve conduction studies are necessary to exclude focal entrapment neuropathies, or multifocal motor neuropathy with conduction block. Magnetic resonance imaging of the cervical spine will detect syringomyelia or other spinal cord disease.

Post-irradiation lumbosacral radiculopathy

This may follow months or years after inclusion of the lower thoracic and upper lumbar spine in irradiation fields treating testicular tumours or lymphoma. It usually affects both legs, or occasionally one, and later causes mild symptoms affecting the sphincters and sensation. It is painless, and electrophysiology does not reveal the myokymic discharges or abnormal sensory nerve action potentials of irradiation plexopathy. The normal imaging of the lumbosacral plexus and cauda equina, and the absence of pain, exclude tumour recurrence.

Post-polio syndrome

After two or more decades, very slowly progressive weakness may affect muscles previously involved by acute paralytic poliomyelitis. Although this predominantly affects the limbs, approximately half of cases also have mild choking or dysphagia and weakness of the respiratory muscles which may lead to hypercapnic respiratory failure. The sluggish deterioration, lack of upper motor neurone involvement, and previous history serve to distinguish post-polio syndrome from amyotrophic lateral sclerosis. Electromyography reveals the giant motor units typical of extensive reinnervation during recovery from previous acute poliomyelitis. At least equally commonly, late deterioration after polio is due to a secondary degenerative arthritis or fibromyalgia.

Multifocal motor neuropathy and neuronopathy

Patients with these conditions may present at any stage of adult life with multifocal and slowly progressive muscle weakness for as long as 20 years. The clinical picture is immensely variable. Distal limb muscles are mainly involved, often notably asymmetrically. The first symptoms and most severe weakness usually affect the arms. Characteristically, severely weakened muscles show little or no wasting. Reflex loss is generally restricted to affected muscles. The condition is neurophysiologically heterogeneous, ranging from muscle denervation to multifocal conduction block in motor nerves, and occasionally a diffusely demyelinating pure motor peripheral neuropathy. Serum antibodies to GM1 gangliosides are detectable in a third of cases, but are of no proven pathogenetic significance. This antibody assay currently lacks specificity since positives are sometimes found in other neurological diseases. Paraproteinaemia is common, particularly immunoglobulin G. These motor neuropathies usually progress insidiously, sometimes in a stepwise manner, and occasionally spontaneous remissions occur. It is important to detect the subgroup of patients with multifocal motor conduction block, or with diffuse demyelinating neuropathy, because improvement may follow immunosuppressant therapy. Although cyclophosamide is reportedly effective, its potential toxicity should limit its use to those patients with severely disabling and progressive weakness. High-dose intravenous human immunoglobulin therapy can produce dramatic improvement lasting 6 to 8 weeks and repeated administration is the mainstay of treatment in severely symptomatic patients. Unfortunately, steroid therapy does not improve multifocal motor neuropathy, and may precipitate further deterioration.

Upper motor neurone syndromes

The pure upper motor neurone syndromes are the rarest forms of motor neurone disease. They should be considered only after magnetic resonance imaging has excluded structural or demyelinating disease of the spinal cord, foramen magnum, or brain. Spasticity is often severe in the purely upper motor neurone diseases, but unfortunately antispasticity medications are often relatively ineffective. Rarely, similar upper motor neurone syndromes may be seen with syphilis, Lyme disease, and HTLV-I infection.

Primary lateral sclerosis

This rare sporadic form of motor neurone disease has an average age of onset of 50 years, and slow progression thereafter for an average of 15 years. The clinical features are all attributable to symmetric degeneration of the upper motor neurones destined for the spinal cord and the bulbar motor neurones. Spasticity and weakness usually commence insidiously in the legs and ascend ultimately to involve the bulbar muscles. Less commonly patients present with an isolated spastic dysarthria, a symptom of pseudobulbar palsy. Pseudobulbar emotional lability may be distressing for these patients, given their normal cognition, and it often responds well to amytriptiline. Bladder function is generally preserved, at least until the later stages. Electromyography does not reveal the muscle denervation to be expected in predominantly upper motor neurone forms of amyotrophic lateral sclerosis. Magnetic resonance imaging may reveal atrophy of the precentral gyrus motor cortex reflecting loss of the Betz cells from which the pyramidal tract originates. Central motor conduction is notably delayed following electromagnetic stimulation of the motor cortex.

Autosomal dominant 'pure' familial spastic paraplegia

Various forms of slowly progressive symmetric spastic paraparesis may be inherited with linkage to chromosomes 2, 14, or 15, most usually on an autosomal dominant basis with onset in the fourth to sixth decades. The degree of leg spasticity often outweighs the severity of the weakness. Bulbar involvement is very rare, and arm function may be well preserved despite severe leg involvement. The condition is slowly progressive. It may remain asymptomatic in some family members, coming to light only when a familial basis for the disease is sought. Sphincter control is not impaired, but sexual impotence can develop. Sometimes varying combinations of other clinical features have been associated with hereditary spastic paraplegia, particularly with recessively inherited forms: distal amyotrophy, mental retardation, dementia, pigmentary retinopathy, optic atrophy, extrapyramidal features, sensory neuropathy, or ataxia.

Lathyrism

Neurolathyrism is a spastic paraparesis caused by regular consumption of the chickling pea (*Lathyrus sativus*) for some months. It is endemic in parts of India and may be epidemic in times of famine. Patients present either subacutely or chronically with a spastic paraparesis and a characteristic scissoring gait in which the balls of the feet take most of the weight. Once it has developed, neurolathyrism is usually not progressive, but little or no recovery occurs even after chickling pea consumption ceases.

Konzo

Konzo is a form of tropical myelopathy which can occur in epidemics at times of famine in several parts of Africa, including Zaire. It seems to be due to dietary cyanide exposure resulting from insufficient soaking of the cassava roots used to produce flour. There is an abrupt onset of symmetric spastic paraparesis which is non-progressive but permanent. Blood cyanide levels are raised at the onset of disease.

Further reading

Bowen J *et al*. (1997). The post-irradiation lower motor neurone syndrome. Neuronopathy or radiculopathy? *Brain* **119**, 1429–39.

Cochrane G, Donaghy M (1993). Motor neuron disease. In: Greenwood RJ *et al.*, eds. *Neurological rehabilitation*, pp 571–85. Churchill Livingstone, Edinburgh.

Donaghy M *et al.* (1994). Pure motor demyelinating neuropathy: deterioration following steroid therapy and improvement with intravenous immunoglobulin. *Journal of Neurology, Neurosurgery and Psychiatry* **57**, 778–83.

Donaghy M (2001). The motor neuron diseases. In: Donaghy M, ed. *Brain's diseases of the nervous system*, 11th edn, pp. 444–60. Oxford University Press, Oxford.

Gregory RP, Mills KR, Donaghy M (1993). Progressive sensory nerve dysfunction in amyotrophic lateral sclerosis: a prospective clinical and neurophysiological study. *Journal of Neurology* 240, 309–14.

Harding AE (1993). Inherited neuronal atrophy and degeneration predominantly of lower motor neurones. In: Dyck PJ *et al.*, eds. *Peripheral neuropathy*, 3rd edn, ch.55. WB Saunders, Philadelphia.

Howard RS, Wiles CM, Loh L (1989). Respiratory complications and their management in motor neuron disease. *Brain* **112**, 1155–70.

Lacomblez L *et al*. (1996). Dose-ranging study of Riluzole in amyotrophic lateral sclerosis. *The Lancet* **347**, 1425–31.

La Spada AR *et al.* (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–9.

Le Febvre S *et al*. (1998). The role of the SMN gene in proximal spinal muscular atrophy. *Human Molecular Genetics* **7**, 1531–6.

Ludolph AC *et al.* (1987). Studies on the aetiology and pathogenesis of motor neuron diseases. I Lathyrism: clinical findings in established cases. *Brain* **110**, 149–66.

McShane MA *et al*. (1993). Progressive bulbar paralysis of childhood. A reappraisal of Fazio–Londe disease. *Brain* **115**, 1889–900.

Olney RK, Aminoff MJ, So YT (1991). Clinical and electrodiagnostic features of X-linked recessive bulbospinal neuronopathy. *Neurology* **41**, 823–8.

Pestronk A *et al.* (1990). Lower motor neuron syndromes defined by patterns of weakness, nerve conduction abnormalities, and high titres of antiglycolipid antibodies. *Annals of Neurology* **27**, 316–26.

Pringle CE *et al.* (1992). Primary lateral sclerosis. Clinical features, neuropathology and diagnostic criteria. *Brain* **115**, 495–520.

Rabin BA *et al*. (1999). Autosomal dominant juvenile amyotrophic lateral sclerosis. *Brain* **122**, 1539–50.

Riluzole for amyotrophic lateral sclerosis (1997). *Drug and Therapeutics Bulletin* **35**, 11–12.

Rosen DR *et al*. (1993). Mutations in Cu-Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362**, 59–62.

Sonies BC, Dalakas MC (1991). Dysphagia in patients with the post-polio syndrome. *New England Journal of Medicine* **324**, 1162–7.

Tandan R, Bradley WG (1985). Amyotrophic lateral sclerosis: Part 1. Clinical features, pathology, and ethical issues in management. *Annals of Neurology* **18**, 271–80.

Tandan R, Bradley WG (1985). Amyotrophic lateral sclerosis: Part 2. Etiopathogenesis. *Annals of Neurology* **18**, 419–31.

Tylleskär T *et al*. (1992). Cassava cyanogens and Konzo, an upper motor neuron disease found in Africa. *The Lancet* **339**, 208–11.

24.13.14 Diseases of the autonomic nervous system

Christopher J. Mathias

Introduction

The autonomic nervous system has two principal efferent pathways, sympathetic and parasympathetic, that innervate and influence every organ in the body (Fig. 1). Autonomic actions are predominantly involuntary and automatic, as indicated by the term 'autonomic' first proposed by Langley in 1898. The structure of the autonomic system, with numerous synapses centrally and peripherally as well as its multiple neurotransmitters, provides flexible control of organ function locally and in an integrated manner—as in the maintenance of systemic blood pressure and body temperature. Disease of the autonomic nervous system may cause local or systemic effects.

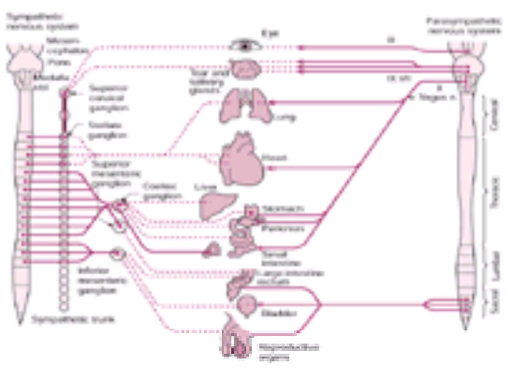

Fig. 1 Sympathetic (thoracolumbar) and parasympathetic (cranial and sacral) pathways that innervate a variety of organs. Janig W (1995). In: Schmidt RF and Thews G, eds. *Physiologie des Menschen*, 25th edn, pp 340–69. Springer Verlag, Heidelberg.

Basic principles

The autonomic nervous system is primarily a visceromotor system, in which each efferent pathway is influenced in a variety of ways. Feedback and central integration is important and virtually every sensory pathway can influence its activity. For example, in spinal cord lesions, activation of visceral, skin, and muscle receptors below the level of the lesion influences autonomic activity and blood pressure through spinal pathways while heart rate responses to classic afferent baroreceptor pathways are retained. Key cerebral autonomic centres are in the hypothalamus, midbrain (Edinger–Westphal nucleus and locus ceruleus), and brainstem (nucleus tractus solitarius and vagal nuclei), and through intracerebral connections. Many other areas affect autonomic activity. Examples are the insular cortex, anterior cingulate gyrus, and amygdala, that are important in processing of emotion and autonomic effects. Parasympathetic efferent pathways are craniosacral and sympathetic efferents are thoracolumbar; each has pre- and postganglionic fibres. The sympathetic ganglia are placed further from target organs than are the parasympathetic ganglia.

Autonomic nerve terminals at target organs vary in complexity; they have the capacity to synthesize neurotransmitters and a host of mechanisms affect uptake and interaction with local or bloodborne chemicals (Fig. 2(a) and Fig. 2(b)). There are differences between organs, especially the gastrointestinal system, in which the enteric nervous system is considered as a third autonomic division. The multiplicity of neural pathways, transmitters, and modulators results in selective control of responses in specific vascular territories and organs, making it a highly complex but precisely regulated and integrated system.

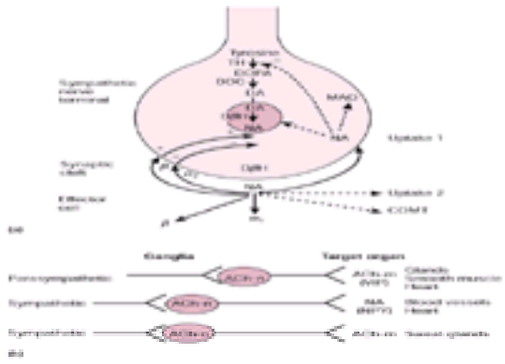

Fig. 2 Schema of some pathways in the formation, release, and metabolism of noradrenaline from sympathetic nerve terminals. Tyrosine is converted into dihydroxyphenylalanine (**DOPA**) by tyrosine hydroxylase (TH). DOPA is converted into dopamine (DA) by dopa decarboxylase (DDC). In the vesicles, dopamine is converted into noradrenaline (NA) by dopamine b-hydroxylase (Dbe). Nerve impulses release both dopamine b-hydroxylase and noradrenaline into the synaptic cleft by exocytosis. Noradrenaline acts predominantly on a_1-adrenoceptors but has actions on b-adrenoceptors on the effector cell of target organs. It also has presynaptic adrenoceptor effects. Those acting on a_2-adrenoceptors inhibit noradrenaline release; those on b-adrenoceptors stimulate noradrenaline release. Noradrenaline may be taken up by a neuronal process (uptake 1) into the cytosol, where it may inhibit further formation of DOPA through the rate-limiting enzyme tyrosine hydroxylase. Noradrenaline may be taken into vesicles or metabolized by monoamine oxidase (MAO) in the mitochondria. Noradrenaline may be taken up by a higher-capacity but lower-affinity extraneuronal process (uptake 2) into peripheral tissues, such as vasular and cardiac muscle and certain glands. Noradrenaline is also metabolized by catechol-*O*-methyl transferase (COMT). Thus, noradrenaline measured in plasma is the overspill not affected by these numerous processes. (From: Mathias CJ (2000). Disorders of the autonomic nervous system. In: Bradley WG *et al*., eds. *Neurology in clinical practice*, 3rd edn, pp 2131–65. Butterworth-Heinemann, Boston.) (b) Outline of the major transmitters at autonomic ganglia and postganglionic sites on target organs supplied by the parasympathetic and sympathetic efferent pathways. The acetylcholine (ACh) receptor at all ganglia is of the nicotinic subtype (ACh-n). Ganglionic blockers such as hexamethonium thus prevent both parasympathetic and sympathetic activation. Atropine, however, acts only on the muscarinic (ACh-m) receptor at postganglionic parasympathetic and sympathetic cholinergic sites. The cotransmitters, along with the primary transmitters, are also indicated—NA, noradrenaline; VIP, vasoactive intestinal peptide; NPY, neuropeptide Y. (From: Mathias CJ (1998). Autonomic disorders. In: Bogousslavsky J, Fisher M, eds. *Textbook of neurology*, pp 519–45. Butterworth-Heinemann, Massachusetts.)

Classification

Diseases of the autonomic nervous system may be primary without a known cause, or secondary with specific abnormalities (dopamine b-hydroxylase deficiency) or strong associations with other diseases (Holmes–Adie syndrome or diabetes mellitus) (Table 1). Drugs are a common cause of autonomic dysfunction (Table 2). Neurally mediated syncope is an intermittent autonomic abnormality. Orthostatic intolerance is listed separately.

Classification may be considered in various ways. Dysfunction may be localized (Table 3) or widespread. Diseases may result from lesions that are central (multiple system atrophy), spinal (spinal cord transection), peripheral (pure autonomic failure), or from a highly specific biochemical deficit (dopamine b-hydroxylase deficiency). Some are age-related, with presentation at birth (Riley–Day syndrome), second decade (vasovagal syncope), or adulthood (familial amyloid polyneuropathy). Autonomic failure commonly causes underactivity, but the reverse, overactivity, causes paroxysmal hypertension during autonomic dysreflexia in high spinal cord injuries. In neurally mediated syncope there is a combination of vagal overactivity and sympathetic withdrawal.

Clinical features

Sympathetic adrenergic failure causes orthostatic hypotension and ejaculatory failure in men, while sympathetic cholinergic failure causes anhidrosis. Parasympathetic failure results in a fixed heart rate, a sluggish urinary bladder and large bowel, and in the male erectile failure. With overactivity there may be hypertension, tachycardia, and hyperhidrosis; while parasympathetic overactivity causes bradycardia. In autonomic disorders there are many clinical manifestations and this may cause diagnostic difficulties, especially when the disorder is generalized.

The presenting complaints often provide clues. Palmar hyperhidrosis or gustatory sweating may indicate a localized disorder, or be a harbinger of widespread autonomic impairment, as the latter may complicate diabetes mellitus. A cardinal feature is orthostatic (postural) hypotension (defined as a decrease in systolic blood pressure of more than 20 mmHg and in diastolic pressure of less than 10 mmHg on standing or head-up tilt, Fig. 3); this impairs perfusion of vital organs, such as the brain. The symptoms vary from fainting (syncope, loss of consciousness) sometimes with ensuing injury, to fatigue and lethargy. Numerous factors in daily life enhance or reduce hypotension (Table 4). Some patients recognize these, with the self-introduction of corrective measures. Large meals, refined carbohydrate, and alcohol, which enhance postprandial hypotension, are avoided. Many sit down, lie flat, or assume curious postures, such as squatting or stooping, that now are recognized to raise blood pressure (Fig. 4). With time, symptoms of orthostatic hypotension wane, for reasons that include improved cerebrovascular autoregulation. In neurally mediated syncope, venepuncture or pain (in vasovagal syncope) or cervical movements and pressure (in carotid sinus hypersensitivity) cause hypotension and bradycardia. A history of impaired sweating and temperature intolerance, urinary disturbances, sexual dysfunction (in men) and gastrointestinal derangement (constipation), especially in combination with orthostatic hypotension, should suggest a generalized autonomic disorder (Table 5).

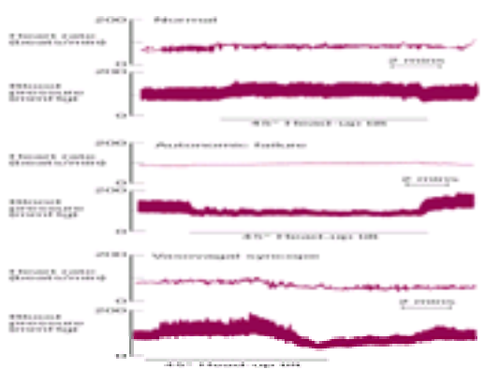

Fig. 3 Blood pressure and heart rate before, during, and after head-up tilt in a normal subject (uppermost panel), a patient with chronic autonomic failure (middle panel), and a patient with vasovagal syncope (lowermost panel). In the normal subject there is no fall in blood pressure during head-up tilt, unlike the patient with autonomic failure in whom blood pressure falls promptly and remains low with a blood pressure overshoot on return to the horizontal. In the patient with autonomic failure there is only a minimal change in heart rate despite the marked blood pressure fall. In the patient with vasovagal syncope there was initially no fall in blood pressure during head-up tilt; in the latter part of tilt, as indicated in the record, blood pressure initially rose and then markedly fell to extremely low levels, necessitating the return of the patient to the horizontal. (From Mathias CJ, Bannister R (1999). Investigation of autonomic disorders. In: Mathias CJ, Bannister R, eds. *Autonomic failure: a textbook of clinical disorders of the autonomic nervous system*, 4th edn, pp 169–95. Oxford University Press, Oxford.)

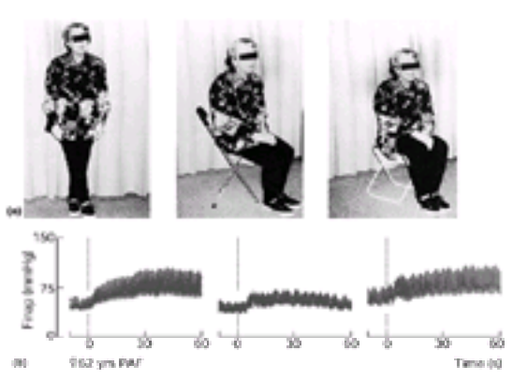

Fig. 4 The effect on finger arterial blood pressure (Finapres) of standing in the crossed-leg position with leg muscle contraction (left), and sitting on a Derby chair (middle), or fishing chair (right) in a patient with orthostatic hypotension. Orthostatic symptoms were present initially when standing and disappeared on crossing legs and sitting on the fishing chair. Sitting on a Derby chair caused the least rise in blood pressure and did not relieve the patient's symptoms completely (From Smith AAJ, Hardjowijono MA, Wieling W (1997). Are portable folding chairs useful to combat orthostatic hypotension? *Annals of Neurology* **42**, 975–8.)

In the Riley–Day syndrome (familial dysautonomia) there is a history of consanguinity, usually in the Ashkenazi Jewish population. A family history often is elicited in vasovagal syncope, and is expected in familial amyloid polyneuropathy. A drug history including exposure to chemicals, toxins, and poisons is important.

A detailed clinical examination is necessary. Pupillary and associated ocular abnormalites occur in Horner's syndrome. To assess orthostatic hypotension, blood pressure should be measured with the patient lying flat, and after standing (or sitting if not possible). A fall in systolic blood pressure of less than 20 mmHg in the prescence of appropriate symptoms does not exclude autonomic failure. Indeed, orthostatic hypotension may be unmasked, or enhanced, by factors such as food ingestion and exercise. Furthermore, in the presence of vascular disease (such as carotid artery stenosis) even a small fall in blood pressure results in cerebral ischaemia. Lack of additional neurological features favour pure autonomic failure (with a good prognosis) while associated parkinsonism or cerebellar dysfunction is suggestive of multiple system atrophy. Several disorders causing a peripheral neuropathy result in autonomic impairment. Basic bedside testing for glycosuria (in diabetes mellitus), or proteinuria (in systemic amyloidosis), provides important information.

Investigation

When an autonomic disorder is suspected, the first step is to determine if autonomic function is normal or abnormal. Autonomic screening tests (Table 6) have their value, but also limitations. The majority are directed towards cardiovascular assessment and to exclude autonomic underactivity. Tests of other systems increasingly are being made available. Normal screening results do not necessarily exclude an autonomic disorder, as on the basis of the history and clinical examination, additional tests such as carotid sinus massage may be needed in patients with syncope. If autonomic tests are abnormal, further evaluation will determine the site and

extent of the autonomic lesion, the functional deficit, and whether it results from a primary or secondary disorder, as an accurate diagnosis is essential for prognosis and for appropriate management. Thus, a 24-h ambulatory blood pressure profile and the effects of stimuli in daily life (such as food and exercise) aid management of orthostatic hypotension; while plasma catecholamine measurements (Fig. 5) and the clonidine growth-hormone stimulation test may separate the different primary autonomic failure syndromes. Investigations may be needed to diagnose underlying diseases, and include neuroimaging studies (MRI or CT scanning), sural nerve biopsy (with specific staining with monoclonal antibodies), and genetic testing. These tests should be combined with non-neurological investigations depending on the suspected diagnosis.

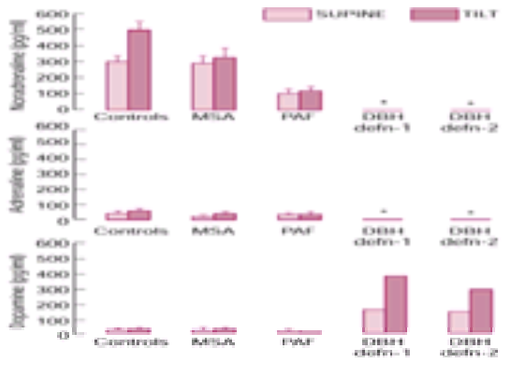

Fig. 5 Plasma noradrenaline, adrenaline, and dopamine concentrations (measured by high-pressure liquid chromatography) in normal subjects (controls), patients with multiple system atrophy (MSA), patients with pure autonomic failure (PAF), and two individual patients with dopamine b-hydroxylase deficiency (DbH defn) while supine and after head-up tilt to 45° for 10 min. The asterisk indicates levels below the detection limits for the assay, which are less than 5 pg/ml for noradrenaline and adrenaline and less than 20 pg/ml for dopamine. Bars indicate ± SEM. (Adapted from Mathias CJ, Bannister R (1999). Investigation of autonomic disorders. In: Mathias CJ, Bannister R, eds. *A textbook of clinical disorders of the autonomic nervous system*, 4th edn, pp. 169–95. Oxford University Press, Oxford.)

Management

This varies depending upon the autonomic disease, the systems affected, the functional autonomic deficit, and whether the disorder is primary or secondary. Treatment should take account of the underlying condition, for example in parkinsonian syndromes, where autonomic features may be worsened by antiparkinsonian therapy. In some diseases simple intervention is effective, such as unblocking a urinary catheter to resolve autonomic dysreflexia in high spinal cord lesions. Complex procedures such as hepatic transplantation are needed to reduce variant transthyretin levels in familial amyloid polyneuropathy. Multidisciplinary expertise may be needed, as in the Riley–Day syndrome and multiple system atrophy, to prevent complications, enhance survival, and improve quality of life. A combined approach is needed to reduce orthostatic hypotension, overcome urinary incontinence, alleviate gastrointestinal disturbances, and treat sexual dysfunction.

The management of orthostatic hypotension is outlined in Table 7 and Table 8; in individual disorders modification is needed.

Individual autonomic disorders

Primary autonomic failure

The onset is usually slow and insidious (chronic autonomic failure) unlike the acute–subacute dysautonomias.

Chronic autonomic failure

The most common is multiple system atrophy where there is additional neurological disease, unlike pure autonomic failure. Patients usually are middle aged at presentation although, with increasing awareness, diagnosis is being made in younger patients.

In pure autonomic failure, diagnosis usually is considered because of orthostatic hypotension. Nocturia (rather than incontinence) is frequent, presumably because fluid shifts from the peripheral to the central compartment elevate blood pressure and improve renal perfusion. Constipation often occurs. In temperate climates, hypohidrosis may not be recognized, unlike tropical areas where heat intolerance and collapse may occur. In the male, impotence is common. The clinical and laboratory findings indicate widespread sympathetic failure, usually with parasympathetic deficits. Physiological and biochemical tests, along with limited neuropathological data, indicate a peripheral autonomic lesion. Management is directed predominantly towards reducing orthostatic hypotension. Although recovery does not occur, the overall prognosis in pure autonomic failure is good.

The most common neurodegenerative disease affecting the autonomic nervous system is multiple system atrophy. It is a non-familial and sporadic disorder with autonomic features and additional neurological (parkinsonian, cerebellar, and pyramidal) features (Table 5) that occur at any stage and in any combination, in an unpredictable manner. Thus, patients initially may consult a range of specialists. It is randomly progressive, which adds to the difficulty of diagnosis. It is synonymous with the previously used term, Shy–Drager syndrome.

In multiple system atrophy the additional neurological features are predominantly parkinsonian; in a smaller number they are cerebellar and as the disease advances there is usually a mixture of features (Fig. 6). The neuropathological findings include striatonigral degeneration in multiple system atrophy (parkinsonian) and olivopontocerebellar degeneration in multiple system atrophy (cerebellar), with both changes often seen in either form. There is cell loss in various brainstem nuclei (that include the vagal nuclei), in the intermediolateral cell mass in the thoracic and lumbar spinal cord, and in Onuf's nucleus in the sacral spinal cord that accounts for the various autonomic and allied abnormalities. The paravertebral ganglia and visceral (enteric) plexuses are spared. A specific feature is the presence of intracytoplasmic argyrophyllic oligodendrocyte inclusion bodies, within the brain and spinal cord. Most patients with multiple system atrophy have parkinsonian features and distinguishing multiple system atrophy from idiopathic Parkinson's disease, especially in the early stages, is difficult. Thus, the true prevalence and incidence of multiple system atrophy is not known. At autopsy up to a quarter of patients previously considered to have Parkinson's disease, have the characteristic neuropathological features of multiple system atrophy.

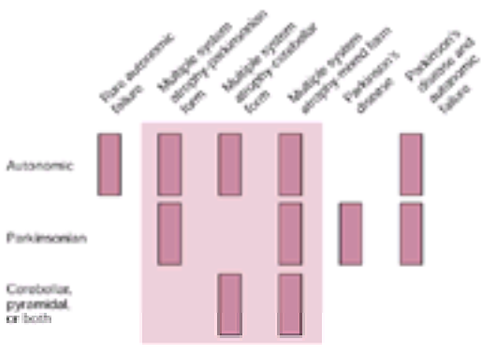

Fig. 6 Schematic representation of the major clinical features of primary chronic autonomic failure syndromes that include the three major neurological forms of multiple system atrophy (adapted from Mathias CJ (1997). Autonomic disorders and their recognition. *New England Journal of Medicine* **310**, 721–4).

In multiple system atrophy (parkinsonian), bradykinesia and rigidity is often bilateral, with minimal or no tremor, unlike Parkinson's disease; however, this may not be a

useful discriminator in an individual. Lack of a motor response to L-dopa is not indicative of multiple system atrophy, as two-thirds respond initially, although refractoriness and side-effects eventually reduce the benefit. The presence of autonomic failure (especially orthostatic hypotension) and unexplained genitourinary symptoms with sphincter disturbance should alert one to the possibility of multiple system atrophy in patients with parkinsonian or cerebellar signs. Oropharyngeal dysphagia and respiratory abnormalities favour multiple system atrophy, although these often occur later. The combination of cardiovascular autonomic failure and an abnormal urethral/anal sphincter electromyogram, with characteristic clinical features are virtually confirmatory of multiple system atrophy. Additional evaluation includes neuroimaging studies using MRI, positron emission tomography, and proton magnetic resonance spectroscopy of the basal ganglia, which are abnormal, at least in established cases. Clonidine growth-hormone testing, based on a $_2$-adrenoceptor stimulation of the hypothalamus with release of human growth-hormone releasing factor, distinguishes central from peripheral autonomic failure and separates Parkinson's disease from multiple system atrophy (Fig. 7); whether this is the case in the early stages of parkinsonism and in patients on dopaminergic agents (that are growth hormone secretagogues), remains to be resolved.

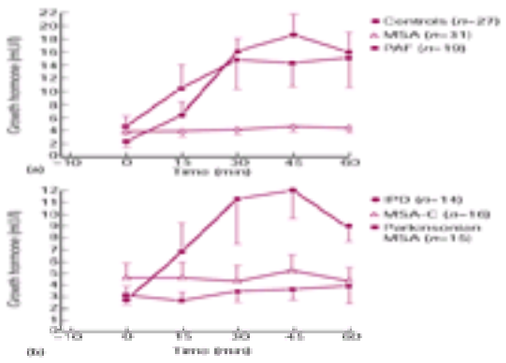

Fig. 7 (a) Serum growth hormone (GH) concentrations before (0) and at 15-min intervals for 60 min after clonidine (2 µg/kg.min) in normal subjects (controls) and in patients with pure autonomic failure (PAF) and multiple system atrophy (MSA). GH concentrations rise in controls and in patients with pure autonomic failure with a peripheral lesion; there is no rise in patients with multiple system atrophy with a central lesion. (From Thomaides T *et al.* (1992). The growth hormone response to clonidine in central and peripheral primary autonomic failure. *Lancet* **340**, 263–6.) (b) Lack of serum GH response to clonidine in multiple system atrophy (the cerebellar form and the parkinsonian form) in contrast to patients with idiopathic Parkinson's disease with no autonomic deficit (IPD), in whom there is a significant rise in GH levels. (From Kimber JR, Watson L, Mathias CJ (1997). Distinction of idiopathic Parkinson's disease from multiple system atrophy by stimulation of growth hormone release with clonidine. *Lancet* **349**, 1877–81.)

The prognosis in multiple system atrophy is poor compared with idiopathic Parkinson's disease and pure autonomic failure. Akinesia and rigidity often worsen, with increasing refractoriness and side-effects (including orthostatic hypotension), to antiparkinsonian therapy. As the disease advances there is often considerable immobility and difficulty in communication. In multiple system atrophy (cerebellar), worsening truncal ataxia causes falls and an inability to stand upright; orthostatic hypotension compounds the disabilities. Incoordination of the upper limbs, speech defects, and nystagmus result in further handicaps.

Respiratory complications include obstructive apnoea (due to laryngeal abductor cord paresis) and central apnoea may necessitate tracheostomy. Oropharyngeal dysphagia enhances the risk of aspiration, especially when vocal cord paresis is present; a percutaneous feeding gastrostomy may be needed. Urinary bladder dysfunction is distressing, and its management, together with that of constipation and, if appropriate, treatment of sexual dysfunction, is important in improving quality of life. There is often a need for specialist therapists, including speech therapists, physiotherapists, dietitians, and occupational therapists. As the neurological decline is inexorable, supportive therapy is crucial in management of multiple system atrophy, and should incorporate the family, carers, and community along with the primary care medical practitioner and therapists.

There is a smaller group of patients with Parkinson's disease, often successfully treated with dopaminergic therapy for many years, who develop severe orthostatic hypotension and other features of autonomic failure. They differ from most patients with Parkinson's disease in whom autonomic dysfunction, if present, is relatively mild and mainly compounded by drug therapy. In Parkinson's disease with autonomic failure, the autonomic lesions appear peripheral (and thus similar to pure autonomic failure)—a conclusion based on low plasma noradrenaline levels and other studies that suggest cardiac sympathetic denervation. Whether in patients with Parkinson's disease complicated by autonomic failure there is a coincidental association of a common condition with an uncommon disease (pure autonomic failure), vulnerability to autonomic degeneration in a subgroup of Parkinson's disease, a link with increasing age, chronic drug therapy, or an inherent metabolic susceptibility—or a combination of these factors—is unknown.

Acute/subacute dysautonomias

These disorders are relatively rare and consist of three main varieties: pure pandysautonomia (with features of both sympathetic and parasympathetic failure); pandysautonomia with additional neurological features usually indicative of a peripheral neuropathy; and pure cholinergic dysautonomia. The prognosis in pandysautonomias is variable, with substantial recovery in some. Recovery in two patients following immunoglobulin therapy favours an immunological basis, and the possibility of a Guillain–Barré syndrome variant. In pure cholinergic dysautonomia, described mainly in children and young adults, there is widespread parasympathetic failure with blurred vision, dry eyes, xerostomia, dysphagia with middle and lower oesophagus involvement, severe constipation, and urinary retention. Clinical findings include dilated pupils, an elevated heart rate, dry and warm skin, a distended abdomen, and a palpable urinary bladder. Anhidrosis may result in hyperthermia. The term 'cholinergic' is used because parasympathetic and also cholinergic sympathetic pathways (to sweat glands) are affected. Sympathetic vasoconstrictor function is preserved and orthostatic hypotension does not occur. Recovery is poor, but the prognosis is good if the condition is detected early. Management includes supportive therapy and adequate fluid and nutrient replacement of losses due to gastrointestinal and sudomotor failure. Barium studies should be avoided because contrast medium accumulates in the atonic colon. The differential diagnosis includes exposure to drugs, poisons, and toxins with anticholinergic effects. Similar autonomic features occur in thorn apple (*Datura stramonium*) seed poisoning; the poisoning is associated with hallucinations, hyperreflexia, and clonic jerking movements and recovery occurs in a few days. Botulism B affects cholinergic but spares motor systems and substantial recovery is expected within 3 months of the exposure.

Secondary disorders

Many disorders are associated with autonomic failure; a few are described.

Riley–Day syndrome (familial dysautonomia)

This is a recessive genetic defect characterized by absent lingual fungiform papillae, lack of corneal reflexes, absence of overflow emotional tears, decreased deep tendon reflexes, and a diminished response to pain and temperature; the disease occurs typically in children of Ashkenazi Jewish extraction. An abnormal intradermal histamine skin test (absent axon flare) and pupillary hypersensitivity to cholinomimetics provide diagnostic confirmation. Prenatal diagnosis is possible with the genetic markers linked to chromosome 9 (q31). Autonomic underactivity and overactivity include lability of blood pressure (hypertension and orthostatic hypotension), intermittent hyperhidrosis, periodic vomiting, dysphagia, constipation, and diarrhoea. The neurological abnormalities include emotional and behavioural disturbances, and sensory deficits that result in injury to skin and joints. Skeletal problems (scoliosis), respiratory (aspiration), and renal failure contribute to a poor prognosis. Anticipation of complications and adequate therapy has extended survival into adulthood.

Amyloid neuropathy

Deposition of amyloid into autonomic nerves can occur in reactive systemic amyloidosis (in chronic inflammatory disorders) or in immunoglobulin light chain (AL) amyloidosis (with lymphomas). In familial amyloid polyneuropathy, sensory, motor, and autonomic abnormalities result from deposition in peripheral nerves of mutated variant transthyretin, produced mainly in the liver. Symptoms of a sensory and motor neuropathy often begin in adulthood in the lower limbs in Portuguese, Japanese, and Swedish forms (familial amyloid polyneuropathy I), and in upper limbs in Indian/Swiss and German/Maryland forms (familial amyloid polyneuropathy II). These and other forms are now classified by the chemical and molecular nature of abnormal fibrillary protein, immunologically related to transthyretin. The most common is based on the first point mutations in the transthyretin gene associated with familial amyloid polyneuropathy—methionine 30 is the Portuguese form. The

cardiovascular system, gut, and gastrointestinal and urinary systems are affected at variable stages, with the disease progressing relentlessly. Autonomic symptoms and signs may be dissociated, leading to underrecognition of the autonomic deficit. Hepatic transplantation reduces variant transthyretin levels and prevents progression of neuropathy. Its ability to reverse neuropathy is unclear, emphasizing the need for intervention before nerve damage occurs.

Dopamine b-hydroxylase deficiency

This rare disorder (with seven patients reported, two of which are siblings) was recognized in the 1980s. Enzymatic deficiency probably occurs at birth but presentation is often in childhood. Orthostatic hypotension has been the clue to recognition. The clinical features indicate sympathetic adrenergic failure, with sparing of sympathetic cholinergic and parasympathetic function; thus sweating is preserved and urinary bladder and bowel function appear normal. In the male, erection is possible but ejaculation difficult to achieve. Basal levels of plasma noradrenaline and adrenaline are undetectable but dopamine is abnormally elevated. Sympathetic nerve terminals, except for the enzymatic and functional defect, are otherwise intact, as demonstrated by electron microscopy, immunohistochemistry, and sympathetic microneurography. Effective treatment is with the prodrug l-dihydroxyphenylserine, that has a structure similar to noradrenaline and is converted by the enzyme dopa-decarboxylase (abundantly present in extraneuronal tissue such as liver and kidneys) to noradrenaline (Fig. 2(a)).

Diabetes mellitus

In patients with long-standing diabetes, especially those on insulin, there is a high incidence of peripheral and autonomic neuropathy. Vagal denervation occurs earlier, impairing heart rate variability. Reduced sympathetic activity, for example in the feet, may increase blood flow substantially at an early stage before detection of neuropathy. Orthostatic hypotension may be enhanced by insulin. There may be sweating abnormalities (gustatory sweating), delayed stomach emptying (gastroparesis diabeticorum), impaired urinary bladder function (diabetic cystopathy), and impotence. Diarrhoea may be extremely distressing.

Spinal cord injuries

Autonomic dysfunction affecting many systems occurs in spinal injuries, depending upon the lesion level and the degree of completeness. Cardiovascular dysfunction may be life threatening, especially in high lesions, in the acute phase in spinal shock, since lack of sympathetic activity with increased vagal tone may cause bradycardia and cardiac arrest (Fig. 8). After a few weeks, spinal shock passes and isolated spinal reflex activity returns; in cervical and high thoracic lesions, abnormal spinal activation results in the syndrome of autonomic dysreflexia. This is induced by cutaneous, skeletal muscle, or visceral stimuli below the level of the lesion. Thus, severe muscle spasms, an anal fissure, or a blocked urethral catheter can result in paroxysmal hypertension (due to increased spinal sympathetic nerve activity, independent of normal cerebral pathways) with associated bradycardia (because of preserved baroreceptor afferents and vagal efferent pathways (Fig. 9). Patients with lesions below T6 are spared. Patients with high lesions also are prone to orthostatic hypotension which compounds difficulties in management, especially shortly after injury.

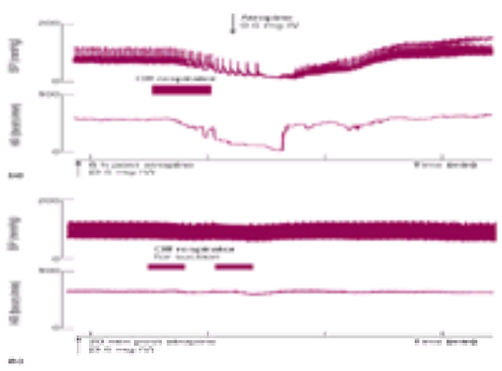

Fig. 8 (a) The effect of disconnecting the respirator (as required for aspirating the airways) on the blood pressure (BP) and heart rate (HR) of a recently injured tetraplegic patient (C4/5 lesion) in spinal shock, 6 h after the last dose of intravenous atropine. Sinus bradycardia and cardiac arrest (also observed on the electrocardiograph) were reversed by reconnection, intravenous atropine, and external cardiac massage. (From Frankel HL, Mathias CJ, Spalding JMK (1975). Mechanisms of reflex cardiac arrest in tetraplegic patients. *Lancet* **ii**, 1183–5.) (b) The effect of tracheal suction 20 min after atropine. Disconnection from the respirator and tracheal suction did not lower either heart rate or blood pressure (From Mathias CJ (1976). Bradycardia and cardiac arrest during tracheal suction—mechanisms in tetraplegic patients. *European Journal of Intensive Care Medicine* **2**, 147–56.)

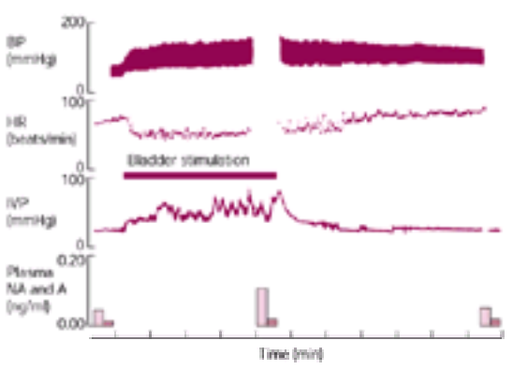

Fig. 9 Blood pressure (BP), heart rate (HR), intravesical pressure (IVP), and plasma noradrenaline (NA) and adrenaline (A) concentrations in a tetraplegic patient before, during, and after bladder stimulation induced by suprapubic percussion of the anterior abdominal wall. The rise in BP is accompanied by a fall in heart rate as a result of increased vagal activity in response to the rise in blood pressure. The level of plasma noradrenaline (open histograms), but not adrenaline (filled histograms), rises suggesting an increase in sympathetic neural activity independently of adrenomedullary activation. (From Mathias CJ, Frankel HL (1986). The neurological and hormonal control of blood vessels and heart in spinal man. *Journal of the Autonomic Nervous System* **Suppl.**, 457–64.)

Drugs

Dysfunction may result from an autonomic neuropathy (as induced by alcohol, vincristine, and perhexiline maleate) or through pharmacological effects. The latter may be expected with the sympatholytic agents, or may be a minor unexpected effect in susceptible individuals. An example of this is the anticholinergic bladder effects of disopyramide, which may cause urinary retention in patients with prostatic hyperplasia. A variety of toxins and poisons, including mushroom toxicity and botulism, as well as nerve gases such as sarin, affect the autonomic nervous system. The first-dose effect of ACE-inhibitors and prazosin may be mediated by the Jarisch–Bezold reflex. Autonomic overactivity occurs during withdrawal of clonidine, alcohol, and opiates.

Neurally mediated syncope

This is an intermittent abnormality with increased cardiac parasympathetic (causing severe bradycardia, cardio-inhibition) and sympathetic withdrawal (causing hypotension vasodepression) that results in fainting. The episodes may be cardio-inhibitory, vasodepressor, or mixed (Fig. 10(a) and Fig. 10(b)). Between episodes, screening autonomic tests usually reveal no abnormalities. In the young, a common cause is vasovagal syncope. This is often familial and more likely in females; it often presents in the early teenage years and is induced by stimuli such as fear, sight of blood, and venepuncture, and at times even discussion of venepuncture. Hypotension is more likely in the upright position and may occur whilst standing still, especially in warm weather when salt and fluid depletion occurs. Testing includes prolonged head-up tilt, or a provocative stimulus such as venepuncture during head-up tilt. A variety of physiological (head-up tilt plus lower body negative pressure) or pharmacological (isoprenaline infusions) stimuli have been used to unmask an episode. Cardiac conduction disorders and other causes of syncope (such as

neurological or metabolic) should be excluded. Treatment includes reducing or preventing exposure to precipitating causes and behavioural psychotherapy in patients with phobias. Added salt, fluid repletion, and exercise are often useful. Drugs such as fludrocortisone, vasopressor agents, and antidepressants such as the serotonin-uptake release inhibitors have been used. The long-term prognosis is favourable.

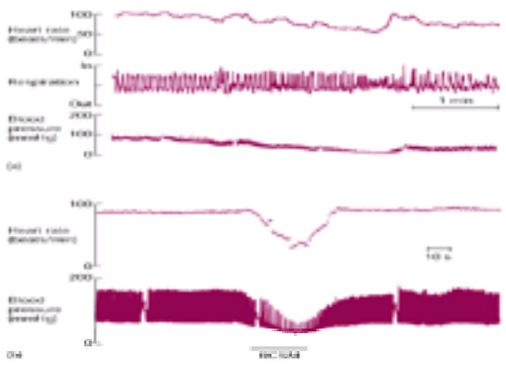

Fig. 10 (a) Blood pressure changes towards the end of a period of head-up tilt in a patient with recurrent episodes of vasovagal syncope. Blood pressure that was previously maintained begins to fall. There is also a reduction in heart rate. Initially there are relatively minor changes in respiratory rate. The patient was about to faint and was replaced to the horizontal (indicated by elevated time signal below) and then to 5° head-down tilt. Blood pressure and heart rate recover but still remain lower than previously. This patient had no other autonomic abnormalities on detailed autonomic testing. Blood pressure was measured non-invasively by the Finapres. (From Mathias CJ, Bannister R (1999). Investigation of autonomic disorders. In: Mathias CJ, Bannister R, eds. *Autonomic failure: a textbook of clinical disorders of the autonomic nervous system*, 4th edn, pp 169–95. Oxford University Press, Oxford.) (b) Heart rate and blood pressure before, during, and after right carotid sinus massage (RCSM) in a patient with syncopal episodes. There is a fall in heart rate and blood pressure during carotid sinus massage, typical of the mixed (cardio-inhibitory and vasodepressor) form of this disorder. The breaks in the record indicate interval calibration by the Finapres machine. (From Mathias CJ (2000). Autonomic dysfunction. In: Grimley-Evans J *et al.*, eds. *Oxford textbook of geriatric medicine*, 2nd edn, pp 833–52. Oxford University Press, Oxford.)

In the elderly, carotid sinus hypersensitivity is increasingly recognized, especially in those with falls of otherwise unknown cause. A classic history of syncope induced by head movements or collar tightening may be provided, although in many the precipitating factors are unclear. Carotid sinus massage should be performed in the laboratory with requisite precautions, ideally using continuous blood pressure and heart rate recordings, with the subject also tilted head-up, because hypotension is more likely to occur when sympathetic activation is needed. Treatment, especially of the cardio-inhibitory forms, includes a cardiac demand pacemaker; vasodepressor forms may require pressor agents. Surgical denervation of the carotid sinus has been used successfully, especially where unilateral hypersensitivity occurs.

A variety of other stimuli, acting through short-lived autonomic mechanisms, also can cause syncope. This may be in conjunction with factors such as heat or drugs that cause vasodilatation or reduce intravascular volume, thus increasing the tendency to hypotension and syncope. Examples include syncope associated with glossopharyngeal neuralgia (caused by swallowing), or induced by micturition, defaecation, coughing, laughing, and playing wind instruments.

Orthostatic intolerance without hypotension

This disorder is increasingly recognized, mainly in women below the age of 50 years. Dizziness on postural change or with modest exertion occurs usually without syncope. The symptoms appear to disrupt their lives, almost disproportionately. There is a substantial rise in heart rate (over 30 beats/min or to 120 beats/min) without orthostatic hypotension, hence the term postural (orthostatic) tachycardia syndrome. Associated features may include those of a partial autonomic neuropathy, chronic fatigue syndrome, mitral valve prolapse, and hyperventilation. It is unclear whether other factors such as vestibular dysfunction contribute. The relationship of this syndrome to previously described psychosomatic disorders such as Soldier's heart (da Costa's syndrome) is not known. Treatment includes salt and fluid repletion, exercise, and b-adrenergic blockers. Many recover within a year.

Further reading

Appenzeller 0, Oribe E (eds) (1997). *The autonomic nervous system*, 5th edn. Elsevier Biomedical, Amsterdam.

Low PA. (ed.) (1997). *Clinical autonomic disorders*, 2nd edn. Little Brown & Company, Boston.

Mathias CJ (2000). Disorders of the autonomic nervous system. In: Bradley WG, Daroff RB, Fenichel GM, Marsden CD, eds. *Neurology in clinical practice*, 3rd edn, pp. 2131–65. Butterworth-Heinemann, Boston.

Mathias CJ, Bannister R (1999). *Autonomic failure: a textbook of clinical disorders of the autonomic nervous system*, 4th edn. Oxford University Press, Oxford.

Mathias CJ, Kimber JR (1999). Postural hypotension —causes, clinical features, investigation and management. *Annual Review of Medicine* **50**, 317–36.

Mathias CJ, Deguchi K, Schatz I (2001). Observations on recurrent syncope and presyncope in 641 patients. *Lancet* **357**, 348–53.

24.13.15 Disorders of cranial nerves

P. K. Thomas

The olfactory nerve

Loss of the sense of smell (anosmia) is most commonly encountered as a sequel to head injury and is probably related to severance of the central processes of the neurones of the olfactory mucosa as they pass through the cribriform plate to the olfactory bulb. It is usually permanent. Distortion of olfaction (parosmia) may occur and may be persistent. The sense of smell is occasionally congenitally absent or may be acutely and permanently lost after a coryzal infection. Bilateral anosmia is frequently accompanied by impairment of taste related to reduced detection of the volatile substances that impart flavours to foods. Unilateral anosmia may occur in olfactory groove meningiomas or other subfrontal tumours. This is usually not detected by the patient.

The central connections of the olfactory pathways are complex and include projections to the temporal lobes, hypothalamus, the septal region, and the amygdaloid nuclei. Olfactory hallucinations are well known to occur as a manifestation of temporal lobe epilepsy. Identification of odours may be impaired after bilateral medial temporal lesions and may be defective in multiple sclerosis, possibly as the result of demyelination in the olfactory tracts. Complaints of hypersensitivity of the sense of smell commonly have a psychoneurotic basis and persistent olfactory hallucinations may be reported by psychotic patients. Persistent parosmia is sometimes produced by lesions of the temporal lobe.

Third, fourth, and sixth cranial nerves

The third, or oculomotor, nerve supplies all the external ocular muscles with the exception of the superior oblique and lateral rectus. It also carries the parasympathetic innervation of the preganglionic pupilloconstrictor fibres of the iris. A complete third nerve lesion produces a dilated and unreactive pupil, complete ptosis, and loss of upward, downward, and medial movement of the eye. The eye becomes deviated downwards and laterally. Diplopia is only experienced when the lid is held up.

The fourth or trochlear nerve supplies the superior oblique muscle. Following a lesion of this nerve, there is extorsion of the eye when the patient looks outwards. When the patient looks downwards and medially, diplopia is experienced. This is particularly disturbing because of its occurrence on looking downwards and produces difficulty in walking and in descending stairs. The patient may compensate for this by tilting the head to the opposite side.

The sixth or abducens nerve supplies the lateral rectus. A lesion of this nerve causes convergent strabismus, inability to abduct the affected eye, and diplopia which is maximal on lateral gaze to the affected side.

The third, fourth, and sixth nerves may be affected singly or in combination, and the paralysis may be complete or partial. In some instances, the lesion is within the brainstem, where it may affect either the nuclei or intramedullary portion of the nerve fibres. In older patients, the commonest causes are vascular disease and neoplasms of the brainstem.

Extramedullary lesions of the third, fourth, and sixth nerves are more frequent and may occur at any point along their course, either intracranially or within the orbit. A third nerve palsy may develop in the region of the tentorial hiatus as a false localizing sign related to displacement of the brainstem produced by supratentorial space-occupying conditions. Unilateral or bilateral sixth nerve palsies may also arise as a consequence of raised intracranial pressure, probably caused by traction, again secondary to brainstem displacement. These nerves can be involved singly or together in conditions such as chronic basal meningitis or carcinomas of the skull base. Gradenigo's syndrome comprises a sixth nerve palsy and pain of trigeminal distribution. It is produced by a lesion at the apex of the petrous temporal bone. As this syndrome was most commonly infective in origin and related to chronic middle ear disease, it is now encountered considerably less frequently.

The third, fourth, and sixth nerves traverse the cavernous sinus, as do the first and second divisions of the trigeminal nerve. In this situation, they are most commonly damaged by an intracavernous aneurysm of the internal carotid artery. The third nerve is affected more often than the fourth and sixth. The consequent internal and external ophthalmoplegia is frequently accompanied by pain, and sometimes sensory loss and paraesthesiae, in the corresponding frontal region related to compression of the first division of the trigeminal nerve, and occasionally in the cheek from damage to the maxillary division. In the superior orbital fissure syndrome, caused for example by a tumour invading the fissure, a total ophthalmoplegia may result, associated with pain and sensory loss in the distribution of the first division of the trigeminal nerve. The eye is often proptosed because of obstruction of the ophthalmic vein. The Tolosa–Hunt syndrome consists of a painful external ophthalmoplegia related to a granulomatous angiitis. Within the orbit, the third, fourth, and sixth nerves may be affected by conditions such as tumours and granulomas. They may be damaged as a result of trauma at any point along their course and may be affected singly or in combination or as part of a cranial neuropathy, of which diabetes, the Miller–Fisher syndrome, Lyme borreliosis, and sarcoidosis are the most important examples. Internal and external ophthalmoplegias are common and this list of causes is by no means exhaustive.

Pupillary abnormalities

Constriction of the pupil (miosis) occurs as a result of paralysis of the sympathetic innervation of the pupillodilator fibres of the iris and may be accompanied by the other features of Horner's syndrome, namely mild ptosis and vasodilatation and anhidrosis of the face on the same side. The ocular manifestations may be encountered alone if the damage is restricted to the intracranial portion of the sympathetic plexus around the carotid artery. Raeder's syndrome consists of these components of Horner's syndrome together with involvement of the first division of the trigeminal nerve. It may be caused by tumours of the skull base. Miosis may also be produced by the local action of cholinergic drugs and by morphine and related compounds.

Pupillary dilatation may be caused by lesions of the third nerve, although it is of interest that the isolated third nerve palsies of presumed vascular origin that may occur in diabetes mellitus, in contradistinction to compressive lesions of the nerve, characteristically spare the pupil. Anticholinergic drugs such as atropine and related substances give rise to pupillary dilatation, as does cocaine.

The Argyll–Robertson pupil is small, fails to react to light, but constricts on ocular convergence, and, if bilateral, the pupils are frequently unequal in size (anisocoria). The pupil may be irregular in outline and it does not dilate fully in response to mydriatics. Argyll–Robertson pupils are almost always related to neurosyphilis but somewhat similar pupils are occasionally encountered in diabetic neuropathy and in some hereditary neuropathies.

The myotonic pupil (Holmes–Adie syndrome) reacts abnormally slowly both to light and on convergence, but particularly so for the response to illumination. A very bright light may be required to demonstrate any pupillary constriction, or if the patient remains in a dark room for some minutes, the pupil slowly dilates. The condition may be unilateral or bilateral and is commoner in women than men. Myotonic pupils may be associated with absence or depression of the tendon reflexes and

occasionally with anhidrosis in the limbs.

Trigeminal nerve

The fifth cranial nerve is predominantly sensory in function, but also innervates the muscles of mastication. It emerges from the pons and runs forwards to the Gasserian ganglion which is situated in Meckel's cave near the apex of the petrous temporal bone. The three sensory divisions of the nerve run anteriorly from the ganglion. The first or frontal division passes through the cavernous sinus and the superior orbital fissure. Its branches supply sensation to the anterior part of the scalp, the forehead, and the eye, including the conjunctiva and cornea. The second or maxillary division leaves the skull through the foramen rotundum, traverses the infraorbital canal, and supplies the cheek. The mandibular division emerges from the skull through the foramen ovale to reach the infratemporal fossa with the motor root with which it unites to form a single trunk. It is distributed to the lower lip, chin, and the lower part of the cheek, and its auriculotemporal branch supplies part of the ear and temporal area. It also supplies the inner aspect of the cheek and the anterior two-thirds of the tongue, and its lingual branch carries taste fibres from the anterior two-thirds of the tongue which leave it in the chorda tympani to join the facial nerve. It is important that the skin over the angle of the jaw is supplied from the second cervical nerve root, and the absence of this 'trigeminal notch' may be useful in distinguishing hysterical or feigned loss of sensation on the face which usually follows the angle of the jaw. The motor root innervates temporalis, masseter, pterygoids, mylohyoid, the anterior belly of the digastric, and also tensor tympani and tensor palati muscles. With unilateral paralysis of the masticatory muscles, the jaw deviates towards the affected side on opening because of the action of the unopposed external pterygoid on the unaffected side.

The trigeminal nerve may be affected by intramedullary lesions, it may be damaged during the intracranial part of its course, or its branches may be compromised extracranially. An acoustic neurinoma may compress the nerve in the posterior fossa or the nucleus of the descending root may be affected by direct compression of the brainstem by this tumour. Loss of corneal sensation is usually the earliest feature. Reference has already been made to involvement of the nerve in association with damage to the sixth nerve at the apex of the petrous temporal bone (Gradenigo's syndrome), as has involvement of the first and second divisions in the cavernous sinus, or the first division in the superior orbital fissure.

Trigeminal neuralgia

Symptoms

This condition is characterized by paroxysms of intense pain strictly confined to the distribution of the trigeminal nerve. In most cases the cause is unknown. It is generally encountered in individuals over the age of 50 years. In younger patients it may be due to multiple sclerosis. Rarely, compression of the nerve, for example by tumours in the cerebellopontine angle, is responsible.

The salient feature of the disorder is pain which is usually unilateral and is felt either within the territory of one division of the nerve only, or may involve two adjacent divisions or affect the whole territory of the nerve. Less commonly it is bilateral.

The pain occurs in brief searing paroxysms, each attack lasting only a matter of seconds. The pain is often described as piercing or knife-like. Its intense quality may cause the patient to screw up their face in agony, hence the use of the term 'tic douloureux' to describe the condition. The paroxysms may be spontaneous or provoked by movements of the face and jaw, by touching the skin, or by draughts of cold air on the face. Eating and speaking may become extremely difficult. 'Trigger spots' on the skin of the face may be present, the touching of which provokes the paroxysms. The attacks may be followed by less severe pain of a dull, boring character and by tenderness of the skin in the affected area. Fortunately the attacks usually cease at night.

The quality of the pain is characteristic, and when trigeminal neuralgia is present, the diagnosis is not usually missed, especially if a paroxysm is witnessed. The usual mistake is to regard as trigeminal neuralgia pain that is due to some other cause, and since there are many conditions that give rise to facial pain, the opportunities for error are numerous. Pain that is of a continuous character is not trigeminal neuralgia and some other cause must be sought. Absence of provocation by eating, talking, or the touching of trigger spots also makes the diagnosis unlikely. Once the diagnosis is accepted, it is essential to exclude compressive lesions affecting the nerve.

In the early stages, remissions lasting for months or years are usual, but in older patients remissions, if they occur, are likely to be brief. In all cases the remissions tend to become shorter as time goes on, and without treatment the condition persists for the rest of the patient's life.

The distribution of the pain is usually in one or two divisions of the nerve. The first division is rarely affected primarily, but pain may spread into it from the second division. If the pain begins in the second division it may, after a time, affect the third, and vice versa.

Treatment

The introduction of carbamazepine revolutionized treatment of this distressing condition. In a high proportion of cases, the paroxysms can be abolished or reduced. A dosage of 200 mg three to five times per day is employed. Ataxia and drowsiness may be troublesome side-effects with higher dosages, and aggravation of ataxia even with modest dosages may impede treatment in cases of multiple sclerosis. Hypersensitivity reactions producing skin rashes or, rarely, bone marrow depression may develop but are, fortunately, uncommon.

If carbamazepine is not successful, or if the patients fail to tolerate it, other drugs such as phenytoin or clonazepam can be tried, but they are rarely effective. In this event, thermocoagulation of the ganglion may have to be considered. This should be undertaken only if the disorder is established so that a prolonged natural remission is unlikely to occur. It should also not be undertaken unless the patient is completely unable to tolerate the disorder, despite analgesics and sedation, and if they are fully aware of the consequences. The persistent analgesia and sometimes dysaesthesiae may subsequently be troublesome, and when the first division is made anaesthetic, damage to the conjunctiva leading to corneal scarring has to be avoided. It may be possible to limit the anaesthesia to the affected area, sparing, for instance, the eye if the first division is not involved by the pain. If thermocoagulation fails, section of the sensory root by a posterior fossa approach employing a microsurgical technique is indicated.

Ophthalmic herpes zoster

In elderly individuals, the fifth nerve is prone to involvement in herpes zoster, the first division being most vulnerable, giving rise to the distressing condition of ophthalmic herpes. The clinical features and treatment of herpes zoster are considered elsewhere (see Section 7). An unfortunate sequel may be visual impairment from residual corneal scarring. Particularly in older subjects, post-herpetic neuralgia may also be a sequel. This gives rise to persistent and unremitting spontaneous pain associated with cutaneous hyperaesthesia in the affected area. Treatment is difficult. Analgesics, sedation, and antidepressive preparations to combat the secondary depression that is frequently present may be of some assistance.

Isolated trigeminal neuropathy

Rarely, a chronic isolated unilateral or bilateral affection of the trigeminal nerve may occur as a manifestation of Sjögren's sicca syndrome, or progressive systemic sclerosis or amyloidosis, although most cases are idiopathic. Extensive nasal scarring and tissue loss may occur secondary to repeated injury from picking and scratching.

Facial nerve

The seventh cranial nerve is largely motor. The nerve traverses the facial canal in the petrous temporal bone in close relationship to the middle ear and emerges at the stylomastoid foramen. Its branches pass forward through the parotid gland to be distributed to the muscles of the face, including the platysma. Within the petrous bone, a branch is given to the stapedius muscle. The chorda tympani, carrying the taste fibres from the anterior two-thirds of the tongue, joins the nerve within the facial canal and a small branch supplies cutaneous sensation to the region of the external auditory meatus. The nerve also carries preganglionic parasympathetic fibres destined for the lachrymal gland.

The distinction between upper and lower motor neurone lesions of the facial muscles is usually easy. In general, with upper motor neurone lesions there is a relative

preservation of power in the upper facial muscles, because these have a bilateral innervation from the cerebral hemispheres. There is no loss of tone with upper motor neurone lesions, so that the sagging of the face that is an unsightly feature of lower motor neurone palsy does not occur.

In common with the trigeminal nerve, the facial nerve may be affected by tumours in the cerebellopontine angle. In the past, it was often involved in middle ear infections. It may be involved in meningeal carcinomatosis, fractures, and tumours of the skull base, in a variety of cranial neuropathies, and cephalic herpes zoster, but the most common lesion by far is Bell's palsy. More peripherally, the nerve may be implicated in tumours of the parotid gland.

Bell's palsy

This term describes a usually unilateral facial paralysis of relatively rapid onset related to a lesion of the nerve within the facial canal. Taste may also be affected. It may develop at any age, most commonly between 20 and 50 years, and affects both sexes equally. Its causation is unknown. In the acute stage, the nerve is swollen and compression within the facial canal may contribute to the damage to the nerve fibres.

The onset is rapid and is frequently heralded or accompanied by aching pain below the ear or in the mastoid region. This clears within a few days and is not present in every case. The paralysis usually reaches its maximum severity after 1 or 2 days but occasionally progresses over the course of several days. Complete paralysis may occur. In the lower face, this may cause a mild dysarthria and some difficulty in eating because of food collecting between the gums and the inner sides of the cheek and the escape of fluid when drinking. The face sags and on smiling is drawn across to the unaffected side. Paralysis of orbicularis oculi renders voluntary eye closure impossible and, particularly in the older subject, ectropion develops. This can result in conjunctival injury from foreign bodies or conjunctivitis. If the paralysis is partial, the lower face is usually affected to a greater extent than the upper.

In the more severe cases, loss of taste over the anterior two-thirds of the tongue is often present, and paralysis of the stapedius muscle may result in a lack of tolerance for high-pitched or loud sounds.

Bell's palsy has to be distinguished from selective lesions of the facial nerve within the brainstem, in which instance taste will not be affected. A facial paralysis superficially resembling Bell's palsy may occur in multiple sclerosis, in which event evidence of more widespread neurological disease may well be detected on examination, or the history may indicate episodes of neurological disturbance in the past. With respect to peripheral lesions, middle ear disease requires exclusion. Facial paralysis related to cephalic herpes zoster is discussed above. A lesion of the facial nerve may represent a mononeuropathy from some generalized disorder of which diabetes, Lyme borreliosis, and sarcoidosis are the most important. Bell's palsy is rarely bilateral and the occurrence of bilateral facial paralysis would raise the possibility of Guillain–Barré syndrome. This may begin with facial weakness, or the weakness may remain restricted to the facial musculature. The occurrence of bilateral facial weakness would also raise the possibility of sarcoidosis.

In approximately 85 per cent of cases of Bell's palsy, the paralysis is the result of a local conduction block within the facial canal without axonal degeneration and this is effectively the situation in all instances of mild weakness. The conduction block is presumably the consequence of segmental demyelination. Providing that such cases do not progress to more severe weakness, all recover fully within a few weeks. In cases where there is total paralysis, a proportion of these will be the result of a conduction block, but in about 15 per cent axonal degeneration will have occurred. Those with a conduction block will again recover satisfactorily within a few weeks. In patients with a degenerative lesion, recovery has to take place by axonal regeneration. Evidence of reinnervation does not appear in under 3 months and the ultimate recovery is often incomplete or may fail to occur altogether. Synkinesis is frequent after reinnervation so that blinking, for example, results in a simultaneous contraction of the angle of the mouth. Aberrant parasympathetic reinnervation may also occur, leading for instance to gustatory lachrymation ('crocodile tears').

Axons remain excitable distal to the lesion for 3 or 4 days after interruption. It is therefore not possible to be certain from electrodiagnostic tests whether axonal degeneration has taken place until after this time. At that stage, electrical stimulation of the facial nerve at the stylomastoid foramen with brief pulses will still elicit a muscle contraction if the paralysis is due to conduction block, whereas none will be obtained if axonal degeneration has taken place.

In the early stages, the main endeavour of treatment should be to prevent either a partial lesion, or complete paralysis related to a conduction block, progressing to a degenerative lesion. There is some evidence that corticosteroids may be advantageous by reducing oedema in the nerve. Thus it is justifiable to treat all cases with corticosteroids if seen within a few days of onset, providing no contraindication to such treatment exists. A course of a week's duration with an initially high dosage is recommended.

Surgical decompression of the nerve has been advised. To be effective, this would have to be performed at the outset, which is not justifiable as 85 per cent of cases will recover satisfactorily without treatment. So far there are no means of predicting which cases will progress to a degenerative lesion. If this were available, decompression could be undertaken selectively in such cases.

It is helpful to perform electrodiagnostic studies at about 1 week after the onset. If this reveals a degenerative lesion, it is then known that recovery will be delayed. A prosthesis attached to the teeth to elevate the angle of the mouth to reduce facial deformity may be helpful. In patients with severe ectropion, a lateral tarsorrhaphy to protect the eye may be required. Electrical stimulation of the paralysed facial muscles has no effect on the ultimate prognosis.

In those cases in which regeneration fails to occur, operation may be desirable for cosmetic reasons to counteract the facial deformity. The angle of the mouth may be elevated by a fascial sling attached to the temporalis fascia, but the result is never highly satisfactory. Restoration of facial tone may be achieved by anastomosis of the hypoglossal nerve to the facial, but at the expense of denervation of the tongue on that side. Any operation should not be contemplated before an adequate length of time has been allowed for regeneration. This should be of the order of 9 months.

Facial paralysis related to 'geniculate' herpes zoster (Ramsay–Hunt syndrome)

Facial paralysis of rapid onset accompanied by severe pain in and around the external auditory meatus and in the throat may accompany 'cephalic zoster'. Vesicles may be detectable in the ear and ulceration in the fauces, or anywhere on the head. Occasionally there is concomitant vertigo, tinnitus, and some deafness with involvement of the eighth nerve ('otic herpes zoster'). Prognosis for recovery of the facial paralysis is stated to be less good than in Bell's palsy.

Hemifacial spasm

This consists of a unilateral disturbance affecting the facial muscles, producing irregular clonic or twitching movements of the facial muscles, usually of insidious onset. It most commonly occurs in middle-aged women. There may be a mild degree of facial weakness, but severe paralysis does not occur. Usually no underlying cause is demonstrable. The condition selectively affects the facial nerve, within the brainstem or in the posterior fossa.

It begins with intermittent twitching of the facial muscles such as around the eye or at the angle of the mouth. These movements gradually become more frequent and extend to involve the rest of the facial muscles, often gradually advancing over the course of some years. If they become severe, the face is contorted by irregular clonic spasms which may keep the eye closed for prolonged periods. The facial distortion is often a considerable embarrassment to the patient, who finds that the spasms tend to be aggravated by emotional stress.

The condition must be distinguished from benign fasciculation of the face, which usually occurs around the eyes, related to fatigue or emotional tension, and from the myokymic twitching that is occasionally encountered as a manifestation of multiple sclerosis. The latter consists of a persisting irregular rippling movement of the facial muscles that usually subsides after a week or two. These conditions can be distinguished by electromyography (see Chapter 24.2.5).

No satisfactory treatment is available. If exaggeration by emotional factors is evident, the administration of diazepam or a similar preparation may produce a marginal improvement. In severe cases, injections of botulinum toxin may be helpful, although these have to be repeated. Neurosurgical intervention to relieve compression of the nerve by aberrant vessels in its intracranial course has been advocated and may be helpful in selected cases.

Glossopharyngeal nerve

The ninth cranial nerve leaves the skull through the jugular foramen, closely related to the tenth nerve. It supplies the stylopharyngeus muscle and the constrictor

muscles of the pharynx. Parasympathetic fibres are supplied to the parotid gland. Sensory fibres are carried from the posterior third of the tongue, the ear, the fauces, and the nasopharynx, and chemoreceptor and baroreceptor afferents from the carotid sinus.

The glossopharyngeal nerve is rarely affected in isolation. Lesions usually occur in conjunction with involvement of the vagus and give rise to some dysphagia, impaired pharyngeal sensation, and loss of taste over the posterior third of the tongue. It may be affected in the jugular foramen syndrome (Vernet's syndrome), along with the tenth and eleventh nerves, of which glomus tumours or metastatic carcinoma are the commonest causes. The nerve may also be involved in diphtheritic neuropathy and in a polyneuritis cranialis.

Glossopharyngeal neuralgia is a rare form of neuralgia within the distribution of the glossopharyngeal nerve. Its features are otherwise strictly comparable with those of trigeminal neuralgia in the quality and severity of the pain, its occurrence in brief paroxysms, its provocation by actions such as speaking or swallowing, and the remissions in its course. As with trigeminal neuralgia, it is most often encountered in elderly subjects, and the pain may initially be confined to individual branches. Thus it may be felt deep in the ear, related to the tympanic branch, or in the throat, related to the pharyngeal branches. It also usually responds to treatment with carbamazepine.

In treatment, carbamazepine may be effective. In instances of severe pain unrelieved by this preparation, surgical treatment, usually avulsion of the nerve, may be required.

Vagus nerve

The tenth cranial nerve is structurally complex. Within the skull it is joined by the cranial division of the eleventh nerve. It leaves the skull through the jugular foramen. Cutaneous sensory fibres are carried from the external ear and visceral afferent fibres are carried from the pharynx, larynx, trachea, oesophagus, and the thoracic and abdominal viscera. Motor fibres are supplied to the striated musculature of the palate and pharynx and through the external and recurrent laryngeal nerves, to the muscles of the larynx. Parasympathetic fibres are provided to innervate the parotid gland (through the glossopharyngeal nerve), the heart, and the abdominal viscera.

The important symptoms of damage to the vagal nerve are those relating to pharyngeal and laryngeal innervation. The cells of origin in the nucleus ambiguus of the medulla may be damaged in the lateral medullary syndrome, in motor neurone disease, and in acute bulbar poliomyelitis, leading to dysphagia and dysphonia. Involvement along with the glossopharyngeal nerve in the jugular foramen syndrome has already been mentioned. The recurrent laryngeal nerve may be damaged during operations on the thyroid gland or by tumours within the neck, or within the thorax, usually by carcinoma of the bronchus. The nerve on the left is vulnerable to damage from aneurysm of the aortic arch. Isolated and unexplained lesions of the recurrent laryngeal nerve are not uncommon.

Nuclear or high vagal lesions, as well as involving the larynx, cause palatal and pharyngeal paralysis. If unilateral, there are no symptoms from palatopharyngeal paralysis. The uvula is pulled up to the opposite side on phonation and pharyngeal sensation is impaired on the affected side. With bilateral paralysis, the palate is paretic leading to nasality of the voice and nasal regurgitation of liquids on attempts at swallowing. Bilateral palatopharyngeal paralysis may be encountered in motor neurone disease, bulbar poliomyelitis, diphtheritic neuropathy, and polyneuritis cranialis.

Unilateral intrinsic laryngeal paralysis from lesions of the recurrent nerve may be asymptomatic or give rise to hoarseness of the voice. If the superior laryngeal nerve is also involved leading to paralysis of the cricothyroid muscle, the affected cord lies in a paramedian or cadaveric position. The effects of bilateral lesions of the recurrent laryngeal nerves depend upon the degree of approximation of the vocal cords. Lesions of insidious onset tend to give rise to dysphonia and also to stridor on exertion. In partial lesions, close approximation of the cords may result from selective paralysis of the abductor muscles, giving rise to limitation of the airway and sometimes necessitating tracheostomy. With bilateral lesions involving both the recurrent and superior laryngeal nerves, both cords are paralysed and in the cadaveric position. Phonation is impossible.

Spinal accessory nerve

The spinal accessory portion of the eleventh cranial nerve arises from the upper cervical cord and the lower medulla. The nerve passes through the foramen magnum and joins the cranial portion of the nerve before emerging from the skull through the jugular foramen. The spinal accessory nerve then separates and supplies the sternomastoid and trapezius muscles, the latter also receiving an innervation from the cervical plexus.

The nerve may be affected by lesions in the region of the jugular foramen, but more commonly it is damaged by injuries to the neck or by operations for the removal of cervical glands, particularly as it crosses the posterior triangle of the neck. Isolated and unexplained lesions of the nerve are occasionally encountered.

Unilateral paralysis of the sternomastoid usually passes unnoticed by the patient. The muscle does not stand out when the head is turned to the opposite side. Paralysis of the trapezius, on the other hand, causes difficulty in lifting the arm above the horizontal, in shrugging the shoulder, and in approximating the scapula to the midline and therefore also in carrying the extended arm backwards. The shoulder droops when the arm is hanging at the side and there is moderate winging of the scapula which is accentuated when the patient attempts to elevate the arm laterally.

The hypoglossal nerve

The twelfth cranial nerve supplies all the muscles of the tongue, both intrinsic and extrinsic. It leaves the skull through the anterior condyloid foramen. A unilateral lesion of the hypoglossal nerve causes weakness and atrophy of the tongue on the affected side. When protruded, the tongue deviates to the affected side. Articulation is unaffected. The nerve may be affected by tumours in the region of the anterior condyloid foramen, or by tumours or penetrating injuries in the neck. If the lesion is the result of a unilateral lower brainstem lesion, it may be combined with a crossed hemiplegia.

Bilateral lesions give rise to generalized atrophy of the tongue. Protrusion becomes impossible and articulation is disturbed. The commonest cause is motor neurone disease (progressive bulbar palsy variant). The wasting of the tongue is usually accompanied by fasciculation.

Further reading

Adour WEK *et al.* (1972). Prednisone treatment for idiopathic facial paralysis (Bell's palsy). *New England Journal of Medicine* **287**, 1268–75.

Asbury AK *et al.* (1970). Oculomotor palsy in diabetic mellitus: a clinicopathological study. *Brain* **93**, 555–66.

Brodal A (1965). *The cranial nerves*, 2nd edn. Blackwell Scientific, Oxford.

Bruyn GW (1983). Glossopharyngeal neuralgia. *Cephalgia* **3**, 143–9.

Cogan DG (1956). *Neurology of the ocular muscles*, 2nd edn. CC Thomas, Springfield, IL.

Cogan DG (1966). *Neurology of the visual system*. CC Thomas, Springfield, IL.

Dyck PJ *et al.*, eds. (1993). *Peripheral neuropathy*, 3rd edn. WB Saunders, Philadelphia.

Esslen E (1977). *The acute facial palsies. Investigations on the localization and pathogenesis of meato-labyrinthine facial palsies*. Springer, New York.

Farrell DA, Medsger A (1982). Trigeminal neuropathy in progressive systemic sclerosis. *American Journal of Medicine* **73**, 57–61.

Katusic S *et al.* (1990). Incidence and clinical features of trigeminal neuralgia. *Annals of Neurology* **27**, 89–95.

Lecky BRF, Hughes RAC, Murray NMF (1987). Trigeminal sensory neuropathy. A study of 22 cases. *Brain* **110**, 1463–86.

Rush JA, Younge BR (1966). Paralysis of cranial nerves III, IV and VI: causes and prognosis of 1000 cases. *Archives of Ophthalmology* **99**, 76–89.

24.13.16 Diseases of the spinal cord

L. D. Blumhardt

Anatomy

The spinal cord extends from its junction with the brainstem (medulla oblongata) at the foramen magnum (opposite the odontoid peg at C1) to the lower border of the first lumbar vertebra. It is enclosed within the arachnoid and dura mater, which extend below the termination of the spinal cord (conus medullaris) into the sacral canal.

The clinically important upper motor neurones run in the corticospinal tract which crosses at the level of the pyramids and then runs caudally in the lateral white matter of the cord. The uncrossed sensory fibres in the posterior columns of the cord convey the sense of joint position and two-point discrimination. Fibres transmitting pain and temperature sensation enter the cord in the posterior spinal roots and ascend three or four segments in the dorsolateral funiculus before decussating through the central grey matter to the contralateral side, where they ascend in the spinothalamic tracts of the anterior and lateral columns. Ascending and descending autonomic fibres involved in sphincter control are in close proximity to the spinothalamic pathways.

The blood supply of the spinal cord is made up of an arterial plexus that is supplied mainly from the anterior spinal artery in the anterior fissure and the two posterior spinal arteries in close proximity to the posterior columns. The anterior plexus is the most extensive and supplies the majority of the cord including the anterior and lateral columns, the corticospinal and spinothalamic tracts, and the anterior grey matter. The posterior system supplies the posterior columns and grey matter. The plexus receives variable supplies from the vertebral arteries, thyrocervical trunk, and multiple spinal medullary arteries arising from the intercostal arteries. The most constant and important contribution to the anterior spinal artery arises from the tenth left intercostal artery ('artery of Adamkiewicz').

Anatomical localization

The sensory and motor pathways of the cord can be damaged by disease anywhere in their long spinal course. Clues to the segmental level of a lesion may be provided by sensory or motor levels, or by involvement of the spinal and superficial cutaneous reflexes. Spastic upper motor neurone weakness in all four limbs arises from involvement of the motor pathways at the cervical level and of the lower limbs at the thoracic level. Loss or reduction of a deep tendon reflex may indicate damage to a particular arc within a segment of the cord, or the appropriate spinal root. By convention, the biceps and supinator jerks are considered C5–C6, the triceps jerk C7, the knee jerk L(3)4, and the ankle jerk S1. Sensory loss, muscle wasting and weakness, or pain within a spinal segment, may also have localizing value. In addition, the loss of superficial reflexes (abdominal D7 to D12, cremasteric L2, plantar S1) has potential localizing value: thus the superficial abdominal reflexes will be abolished by lesions at D6 or above, whereas the upper abdominal reflexes should be retained when the lesion is at D10 or below.

If there is complete transection of the cord, motor power and all sensation is lost below the level of the segment involved. In addition, there will be retention of urine with overflow incontinence. If the condition develops suddenly a state of spinal shock occurs in which there is flaccid paralysis of both skeletal and smooth muscle with loss of sensation and reflex activity below the level of the lesion. Partial lesions result in a variable syndrome of motor and sensory signs. There is often loss of coordination of bladder contraction and sphincter relaxation with consequent urgency and incontinence. If the cord damage is unilateral, the Brown–Séquard syndrome will result in ipsilateral loss of posterior column function combined with contralateral upper motor neurone signs and symptoms and spinothalamic irritation, or impairment. There is a level on the trunk usually several segments below the lesion. Light touch sensation is generally relatively preserved. Superficial abdominal reflexes are generally lost below the level of the lesion. The syndrome is often partial or incomplete.

A spastic paraparesis (upper motor neurone weakness and spasticity in the lower limbs) may be associated with loss of reflexes in the arms, typically asymmetric C5–C6 loss or reduction in cervical spondylotic myelopathy, Damage to the C5–C6 reflex arc may be associated with weakness of shoulder abduction and elbow flexion, reduced biceps and supinator jerks, and enhanced triceps jerks, with spastic weakness below this level. Syringomyelia is usually associated with general upper limb areflexia with a spastic paraparesis and dissociated sensory loss (see below). Lesions at the level of the foramen magnum (meningioma or developmental abnormalities) that involve the cervical enlargement may be confusingly associated with wasting of the small muscles of the hands. Respiratory involvement may complicate high cervical lesions through paralysis of the muscles of the thoracic wall or diaphragm.

Symptoms of spinal cord disease

Lesions of the upper and lower motor neurones cause characteristic patterns of symptoms. Lower motor neurone weakness and wasting often goes unnoticed by patients in the early stages when it will already be obvious to an examiner. By contrast, in the earliest stages of upper motor neurone involvement, symptoms may be present in the absence of signs. Upper motor neurone weakness, which is usually responsible for the first symptoms of a myelopathy, may be described as involving sensations of heaviness, dragging, 'numbness', stiffness, tripping, scuffing, lack of control, clumsiness, or loss of dexterity. Rapid repetitive movements are particularly impaired. Associated spasticity may cause extensor spasms and clonus, which may be described as a vibration of the foot usually noticed when negotiating stairs or kerbs. Lesions of the spinothalamic tracts result in a sensory level on the trunk with loss or reduction of pain and temperature sensation on the opposite side below the lesion (usually two or three segments below, because of the ascending course of the fibres prior to sensory decussation). Partial or irritative spinothalamic lesions cause a variety of unpleasant sensory symptoms, including burning pains, increased sensitivity to touch, feelings of wetness, or the sensation of movements under the skin (formication). Posterior column involvement can lead to tingling paraesthesiae and tight constricting feelings around joints, as well as sensory ataxia. Stretching or movement of the cervical cord on neck flexion may cause shooting paraesthesiae ('electric shocks') down the back into the lower limbs ('Lhermitte's symptom'). This can be associated with any pathology in the cervical canal but is most often seen with multiple sclerosis. Autonomic involvement may cause hesitancy, urgency, and urge incontinence of bladder and bowel.

Acute and subacute myelopathy

Rapidly developing weakness in the lower limbs (over minutes, days, or weeks) requires immediate referral to a neurological unit for diagnosis and management. The urgent problem is to establish the site of the weakness (spinal cord or peripheral nerve) and the cause (intrinsic or extrinsic spinal cord disease or acute polyneuritis). It is essential to exclude a compressive cause requiring emergency decompression. An acute transverse myelopathy, with absent deep tendon reflexes, non-elicitable plantar responses, and paralysed, flaccid limbs (so-called 'spinal shock'), may superficially mimic acute Guillain–Barré syndrome, but retention of urine and a sensory level on the trunk will indicate the correct pathological localization in the spinal cord.

An acute or subacute spinal cord syndrome may be due either to extrinsic cord compression (Table 1) or to intrinsic pathology. If compressive, the prognosis for recovery may depend on how rapidly the spinal cord can be decompressed. There are many intrinsic causes of an acute myelopathy, including multiple sclerosis, acute disseminated encephalomyelitis, viral myelitis, systemic lupus erythematosus, sarcoidosis, Behçet's disease, paraneoplastic syndrome, and spinal cord infarction or haemorrhage. Unless there are other clues from the history, or from general or neurological examination, it is not usually possible to make a reliable distinction between intrinsic and extrinsic causes. It is vitally important to relieve bladder retention and to search for a primary focus of infection or underlying neoplasia that may have led to metastasis. Neuroimaging of the spinal cord (preferably magnetic resonance imaging) must be carried out as an emergency procedure. Lumbar puncture should be avoided until compression has been excluded, as worsening of the myelopathy may be precipitated by reduction of cerebrospinal fluid pressure below an obstruction.

Chronic progressive myelopathy

The first symptoms of a slowly progressing spinal cord lesion, particularly a compressive lesion, are often due to the insidious development of upper motor neurone weakness. There may be a barely noticeable deterioration in ambulation with the onset of subtle difficulties (heavy limbs, dragging foot or leg), at first only apparent when running or during long walks. At first there may be little in the way of objective signs, and symptoms may progress very slowly over months or years A careful history and examination in anyone who complains of walking difficulties is mandatory. Upper motor neurone weakness is frequently missed by incomplete examination as it primarily involves flexor movements in the legs and extensors in the arms. As for acute myelopathy, the cause may be intrinsic or extrinsic, and neuroimaging is indicated without delay. Common compressive causes are cervical spondylotic myelopathy, prolapsed intervertebral dorsal disc, neurofibroma, meningioma, and ependymoma. Multiple sclerosis is the most common intrinsic cause in the United Kingdom, but motor neurone disease, hereditary spastic paraplegia, HTLV-I myelopathy, vitamin B_{12} or folate deficiency, and, rarely, thyroid disease may present in this way and require exclusion.

Cervical spondylotic myelopathy

This is a spinal cord compression syndrome that typically occurs in the sixth decade or later. It usually presents as a chronic progressive spastic paraparesis, but an acute or subacute ascending myelopathy can also occur. Radicular symptoms (pain and sensory impairment or paraesthesiae) may or may not be present, but typically, there is coexisting radiculopathy and the C5–C6 reflexes are found to be depressed or absent and the C7 reflexes brisk ('inverted supinator jerk'). Magnetic resonance imaging of the cervical spine is the investigation of choice. Those affected have constitutionally narrower cervical canals. Coexisting ischaemic demyelination due to compression is often present in the centre of the cord, but a statistical association with multiple sclerosis has also been described.

Spinal epidural abscess

Pain with signs of spinal cord compression may occur acutely with a spinal epidural abscess, and often with little in the way of systemic disturbance. Staphylococcal and streptococcal organisms are frequently responsible and septicaemia may be present. Treatment is with appropriate antibiotics and surgical drainage procedures.

Tuberculosis

Tuberculosis is primarily an infection of the spinal vertebral bodies and intervertebral discs. Compression of the cord often results from vertebral collapse. Infection may spread through the meninges to involve the spinal cord directly, or there may be associated arterial occlusions. Management is by antituberculous drugs and, where appropriate, surgery for decompression or spinal stabilization.

Syringomyelia

A rare condition (prevalence approximately 7 per 100 000) in which an irregular fluid-filled cavity (syrinx) causes a central cord syndrome. The cavity usually begins in the cervical area, initially localized to a segment or two, but may extend the whole length of the cord. Adjacent structures including the decussating fibres destined for the spinothalamic tracts, the anterior horn cells, and the pyramidal tracts may become involved, although the posterior columns are characteristically spared.

Typically in early adult life the patient notices weakness and wasting of an upper limb and/or loss of the sensation of pain. Painless injuries often occur and may lead to dramatic symptoms (one of the author's patients was a butcher who lost the ends of his fingers in a mincer before noticing). There is usually areflexia in the arms, a spastic paraparesis, and a 'dissociated' sensory loss (retention of common touch and proprioception sense with loss of pain and temperature sensation) in the upper limbs with extension on the trunk in the forequarter ('cape') area. Associated features may include Horner's syndrome, upper limb pain, and a mild dorsal scoliosis. The syrinx and the presence of any associated abnormality at the foramen magnum are now easily established by magnetic resonance imaging, sometimes before a typical clinical syndrome develops. Tumours and Tangier's disease can also result in a central cord syndrome. The condition can be static for years, but often progresses irregularly. Rostral extension into the brainstem (syringobulbia) may cause dysarthria, dysphagia, wasting of the tongue, and sensory loss in the face (sparing the central areas such as the mouth and nose until last).

A syrinx is often associated with an Arnold–Chiari malformation, but can be secondary to an intrinsic spinal cord tumour or spinal cord injury. Treatment by surgery is controversial and outcomes are complicated by the variable natural history. If there is an associated cerebellar ectopia and hypodevelopment of the posterior fossa, decompression of the foramen magnum may successfully relieve pain and halt progression. Drainage of the syrinx by various shunt procedures may also be helpful if it is carried out before moderate or severe disability has resulted.

Spinal arachnoiditis

Any inflammatory process can result in a progressive fibrosis of the subarachnoid space. The most common causes include myelographic contrast media, subarachnoid haemorrhage, surgery, trauma, and infection. Complications include cystic compression and ischaemic damage to the spinal cord and/or spinal nerve roots, often with intractable pain and blockage of cerebrospinal fluid pathways. Once initiated, arachnoiditis can be progressive and there is no effective treatment. Attempts at surgical correction may cause further fibrosis and deterioration.

Schistosomiasis (bilharziasis)

A chronic infection with a trematode worm (in South America, Asia, and Africa) occasionally causes spinal cord or root compression or a transverse myelitis due to a granulomatous inflammatory reaction to the parasite's eggs. Granulomas may be intrinsic or extrinsic and there may be associated meningoencephalitis and intracranial granulomas. Treatment is with praziquantel. Other worms including cysticercosis and echinococcus may also (rarely) cause spinal cord compression.

Demyelinating diseases

Involvement of the spinal cord is almost invariable in multiple sclerosis. In the early stages of the disease, acute or subacute episodes of partial myelopathy often occur with partial or complete recovery (relapsing–remitting disease). Lhermitte's symptom is common, but non-specific for demyelination. In the later stages (secondary progressive multiple sclerosis) a chronic progressive myelopathy associated with spasticity, sensory loss, and ataxia is characteristic. Eighty five per cent of patients with primary progressive multiple sclerosis present as an insidious progressive myelopathy, mostly in middle age. Diagnosis is established by clinical history and examination, magnetic resonance imaging of the spinal cord and brain, evoked potentials, examination of the cerebrospinal fluid (intrathecal oligoclonal bands), and blood tests to exclude other causes of myelopathy.

Perivenous inflammation and demyelination also occur in the white matter of the brain and spinal cord following vaccination or an acute childhood viral infection such as measles, mumps, rubella, or chicken pox (acute disseminated encephalomyelitis), but the triggering infection is not always identifiable. The illness is usually monophasic, but recurrent symptoms indicate that the initial bout was the onset of multiple sclerosis.

Acute/subacute transverse myelitis

An acute or subacute spinal cord syndrome may occur as a monophasic, spontaneous illness with no obvious cause, or may follow a viral infection or vaccination. It is characterized by a sensory level on the trunk (usually thoracic) to all modalities with retention of urine and faeces and severe symmetrical paraplegia. The myelogram

may be normal or show a swollen cord. The spinal fluid shows an inflammatory reaction with an excess of white cells and raised protein. High-dose intravenous steroids are usually administered. There may be a full or partial recovery, or permanent paraplegia may result.

Many conditions can give rise to a similar clinical picture, including infections (sarcoidosis, syphilis, Lyme disease, spinal tuberculosis, brucellosis), collagen vascular diseases, viruses (Epstein–Barr, herpes zoster, herpes simplex, rubella), and multiple sclerosis. However, unlike the acute or subacute partial spinal cord syndromes, multiple sclerosis develops in only a small proportion and the cause most often remains obscure, even after many years of follow-up.

Paralytic poliomyelitis

After an incubation period of 1 to 2 weeks, a viraemia is associated with fever, vomiting, and headache. Replication of the virus in the anterior horn cells of the spinal cord and the motor nuclei of the brainstem is usually associated with signs of meningitis and pains in the spine and limbs ('preparalytic phase'). Recovery may then occur, but some patients may go on to develop a highly variable, asymmetric, patchy muscle weakness. Reflex loss, fasciculation, and muscle wasting are early features, but sensation remains normal. A minority of patients develop bulbar complications including respiratory failure. A variable degree of recovery may occur. Management is supportive with control of frequently associated infections. Widespread vaccination has virtually eliminated this condition.

Human immunodeficiency syndrome (AIDS) myelopathy

A subacute to chronic progressive myelopathy characterized by vacuolar changes in myelin sheaths with relative axonal preservation and an emphasis on the dorsolateral thoracic spinal cord is the most common form of myelopathy in AIDS. Spasticity, weakness, loss of the sense of joint position, and sphincter dysfunction develop over weeks or months. Magnetic resonance imaging usually shows the spinal cord to be normal or non-specifically atrophic. There may be coexisting dementia or neuropathy. Treatment is limited to management of symptoms.

Myelitis in AIDS may result from a variety of causes including infection with herpes simplex, cytomegalovirus, varicella zoster, HTLV-I and -II, syphilis, or tuberculosis. Vitamin B_{12} deficiency, lymphoma, and spinal epidural abscess may also cause an acute or subacute myelopathy in this situation.

HTLV-I associated myelopathy ('tropical spastic paraparesis')

Retroviral infection results in a chronic slowly progressive myelopathy characterized by paraesthesiae (often painful) in the lower limbs, sphincter dysfunction, and spastic paraparesis. Some cases may also have cerebellar ataxia, optic neuritis, and signs of a peripheral neuropathy. There is positive HTLV-I serology in the blood and cerebrospinal fluid.

Nutritional deficiencies

Subacute combined degeneration of the spinal cord due to vitamin B_{12} deficiency has become rare with the early diagnosis and treatment of pernicious anaemia. Presenting complaints are persistent paraesthesiae in the feet and later in the hands with unsteadiness of gait. There may be signs of peripheral nerve damage (distal sensory loss and loss of ankle jerks). Unless treatment is initiated early, weakness in the lower limbs becomes more marked with extensor plantar responses and ataxia indicating largely irreversible damage to the pyramidal tracts and posterior columns. There may be additional loss of memory and cognition, reduced vision with central scotomas, and sphincter disturbance. The patients are usually middle aged and there may be no evidence of anaemia or changes in the blood film. The serum level of vitamin B_{12} is almost invariably markedly reduced. Treatment with hydroxycobalamin should be started immediately to prevent irreversible damage. Motor symptoms usually respond better than sensory symptoms, and paraesthesiae may persist indefinitely. Rarely, a similar syndrome may be associated with folate deficiency.

Low levels of vitamin E may cause a spinocerebellar syndrome with pyramidal signs, with or without peripheral nerve involvement and ophthalmoplegia. Patients present with an unsteady gait and weakness with limb ataxia, and loss of reflexes, proprioception, and vibration sense. The cause is often malabsorption due to gastrointestinal disease (see also Section 14).

Vascular disease

Haemorrhage into the parenchyma of the spinal cord is rare and may be spontaneous, or a complication of arteriovenous malformations, trauma, and clotting abnormalities, including anticoagulant therapy . Infarction of the spinal cord may result from aortic disease (atherosclerosis, dissecting aneurysm, surgery), thoracic or cardiac surgery, or cardiac arrest (hypotension). The resulting paraplegia is often ushered in by acute pain at the level of the infarction. As the anterior spinal artery is most often involved, loss of pain and temperature sensation occurs below the level of the infarction with preservation of common sensation and proprioception. Infarction of one-half of the cord results in the Brown–Séquard syndrome. Rarely, the whole cord can be infarcted below the occlusion. Recovery from infarction is variable and unpredictable. Embolism may occur in decompression sickness or bacterial endocarditis and ischaemic damage in the context of autoimmune vasculitic diseases such as systemic lupus erythematosus and Sjögrens disease. Spinal arteries may be secondarily involved during bacterial meningitis or syphilitic leptomeningitis, or by neoplasia.

Drug abuse, particularly of heroin or cocaine, may cause an acute myelopathy. The clinical picture resembles an anterior spinal artery occlusion with paraplegia, sensory level and bladder dysfunction, and sparing of posterior column function. The causes include particulate emboli, vasculitis, and watershed infarction from hypotension.

Decompression sickness arising from bubbles of gas in the bloodstream caused by too rapid decompression following exposure to high atmospheric pressures (usually due to diving) may cause ischaemic damage to the spinal cord. Generally within minutes of a return to normal atmospheric pressure a widely variable syndrome of paraesthesiae, sensory loss, pain, and weakness in the limbs develops, that either recovers on rapid recompression, or progresses to a permanent spastic paraparesis.

Venous thrombophlebitis may cause a patchy, asymmetric, and rapidly ascending subacute necrotic myelitis. This may occur as a result of intra-abdominal or pelvic sepsis, but often no cause can be found. The prognosis is poor.

Spinal arteriovenous malformations are congenital abnormalities of blood vessels supplying the spinal cord and may cause either an acute or chronic progressive spinal cord syndrome. There may be various combinations of spinal cord compression, venous infarction, vascular 'steal' due to shunting, haemorrhage, or progressive fibrosis. Pain may worsen with exercise. Haemorrhage into the cerebrospinal fluid can cause sudden confusion, headaches, and neck stiffness. Cord compression may result from a secondary haematoma. Auscultation over the spine should be performed in all patients with undiagnosed myelopathy, although bruits are only present in a minority. The arteriovenous malformation may be demonstrated by magnetic resonance imaging or myelography and its vascular supply established by spinal angiography, which is not without risk. They may be inoperable, or amenable to embolization or surgical removal in expert hands.

Subacute necrotizing myelitis is a rare condition in which a flaccid paraplegia results from extensive cord involvement below the highest segment. There may be a stepwise development with ascending levels. The lesion can be associated with vascular abnormalities, but no cause can usually be identified, treatment is not effective, and the prognosis is poor.

Spirochaetal infections

Acute vascular cord lesions including an anterior spinal artery occlusion may be the presentation of meningovascular syphilis. Damage to the spinal cord can also arise from chronic inflammatory changes in the cervical meninges (leptomeningitis) involving spinal nerve roots and spinal cord. Symptoms and signs of wasting of the small hand muscles (C8–T1) with and without upper motor neurone signs may be present.

The diagnosis is often suspected when high cell and protein levels are found and confirmed by serology in the blood and cerebrospinal fluid.

The classical chronic progressive neurological disorder tabes dorsalis has become very rare due to improved recognition and early treatment of syphilis. Clinical features may include an ataxic steppage gait, lancinating pains in the legs, skin ulcers, hypotonicity, and painless arthritis. Damage starts in the dorsal root ganglion

cells and particularly involves the central processes extending into the posterior columns. Classic 'lightning' pains or electric shock-like sensations in the lower limbs and eventually loss of deep pain sensation usually occur. Further damage to the spinal roots leads to loss of reflexes and hypotonicity of the limbs. Hypermobility of joints with loss of pain sensation may lead to a progressive painless destruction of knee, ankle, or elbow joints ('Charcot's joints'). Damage to the pyramidal tracts eventually results in extensor plantar responses ('taboparesis'). Additional findings may include analgesic patches on the trunk and inner arms, loss of sphincter control with a dilated atonic bladder, optic atrophy, small light near-dissociated pupils (Argyll–Robertson pupils), and ptosis. Painful autonomic attacks ('crises') affecting the bladder, rectum, stomach, or larynx, may also be a feature.

Toxic damage

X-ray therapy to the thorax or neck, particularly directed at the bronchi or oesophagus, may occasionally result in damage to the spinal cord (radiation myelopathy). The risk is dose related. An acute transient form of myelopathy characterized by paraesthesiae and Lhermitte's symptom comes on soon after exposure and usually remits. The mechanism is probably demyelination of the posterior columns. The delayed myelopathy that comes on 6 months to several years after treatment is due to a vasculopathy with cord necrosis and atrophy leading to a progressive spinothalamic sensory loss and disabling paraplegia. Magnetic resonance imaging excludes other possibilities, including compression from metastases and paraneoplastic subacute necrotizing myelitis.

The neurotoxin b-*N*-oxalylamino-L-alanine, a glutamine receptor agonist, contained in the pulse *Lathyrus sativus* (chickling pea) may cause an acute or chronic paraparesis, often with prominent muscle spasms, which usually occurring in outbreaks during periods of famine (lathyrism). Chronic cyanide poisoning from ingesting the roots of the cassava plant may result in pyramidal signs, usually combined with a painful peripheral neuropathy.

Neuronal degenerations

Hereditary spastic paraplegia

This is a relatively rare condition, in which degeneration of the pyramidal tracts and posterior columns causes a slowly progressive spastic paraparesis of varying severity. It is essential to take a family history and examine relatives carefully, as mild cases may be asymptomatic, causing misdiagnoses and an underestimate of the true prevalence. There are dominant (70 to 80 per cent of cases) and recessive forms. Genetic studies of the autosomal dominant variant have identified loci on at least three chromosomes that account for about half the cases. Type I hereditary spastic paraplegia comes on in childhood or adolescence with clumsiness and poor athletic performance caused by spasticity. Weakness is usually slight or absent. Sensory loss tends to be mild and late and usually involves proprioception. Abdominal reflexes are often retained and significant bladder involvement is relatively uncommon until the later stages. Pes cavus is often present. Type II hereditary spastic paraplegia has similar clinical features, but an onset after the age of 35 years. Recessive forms tend to be more rapidly progressive and are often misdiagnosed. Other causes of spastic paraparesis must be excluded.

Motor neurone disease

The full-blown picture of amyotrophic lateral sclerosis with mixed lower and upper motor neurone features is unmistakeable, but individual cases can sometimes resemble other diseases, particularly cervical spondylotic myelopathy and primary progressive multiple sclerosis. A minority of cases present with a pure upper motor neurone spinal cord syndrome with a spastic paraparesis ('primary lateral sclerosis') and no evidence of lower motor neurone wasting or weakness. The clinical picture is of a slowly progressing spastic paraparesis or quadraparesis. The differential diagnosis is extensive and many of the conditions listed in Table 2 need to be considered and excluded. Investigations may include magnetic resonance imaging, electromyography, muscle biochemistry, evoked potentials, and examination of the cerebrospinal fluid. Prolonged clinical follow-up may be required to clarify the diagnosis.

Friedrich's ataxia

This is a rare, progressive, autosomal recessive condition that begins in childhood or adolescence and leads, usually by the third decade, to limited mobility and death a few years later. Presentation is usually with unsteadiness and clumsiness. Features include dysarthria, inco-ordination of the limbs, gait ataxia, impaired proprioception, and areflexia.

Adrenomyeloleucodystrophy

This is an X-linked recessive disorder in which demyelination in the brain and spinal cord occurs as a result of the accumulation of very long chain fatty acids. Adult males develop a spastic paraparesis with or without mild sphincter involvement and sensory loss. A very slowly progressive mild spastic paraparesis with sphincter and sensory loss may also be found in heterozygote females. Associated adrenal insufficiency may be very mild. The condition may be mistaken for multiple sclerosis, although the magnetic resonance abnormality, when present, usually shows widespread symmetrically distributed lesions often in the posterior white matter more appropriate to a leucodystrophy. A locus has been demonstrated at chromosome Xq28. There are increased levels of very long chain fatty acids in the serum.

Further reading

Fink JK, Heineman-Patterson T (1996). Hereditary spastic paraplegia: advances in genetic research. *Neurology* **46**, 1507.

24.13.17 Spinal cord injury and its management

M. P. Barnes

Introduction

Spinal cord injury is a prime example of the improvement in survival and quality of life that can follow from the application of modern rehabilitation techniques. In the early part of the last century around 9 out of 10 people with a spinal cord injury died within 1 year and only 1 per cent survived in the long term. The situation greatly improved with the advent of spinal cord injury centres, particularly pioneered in the United Kingdom by Sir Ludwig Guttmann at Stoke Mandeville. The co-ordinated, multidisciplinary care provided at these centres significantly reduced mortality and improved quality of life. However, even in the 1960s there was still a 35 per cent mortality associated with tetraplegia, and it has only been in the last decade or so that modern rehabilitation techniques have reduced mortality to less than 5 per cent. Life expectancy has improved such that the major causes of late death in spinal injury are now those experienced by the general population, such as cancer and myocardial infarction. However, there is no room for complacency. Deaths from renal failure or respiratory infection in those with tetraplegia are still too common. Although life expectancy in those with paraplegia is only modestly reduced, people with tetraplegia still have a significantly reduced survival rate. A 20-year-old male would normally be expected to have a further 56 years' life expectancy, but this is reduced to about 45 years in those with paraplegia, and those with tetraplegia only have an expected survival of a further 33 years.

Although this chapter will concentrate on the medical management of spinal cord injury and its complications, it would be incomplete without a mention of the acute and emergency management.

Epidemiology

The annual incidence of spinal cord injury, at least in Western Europe, is around 10 to 15 cases per million of the population per annum. The mean age of injury is about 33 years, although the mode is 19 years. Most injuries occur in males (around 82 per cent). The commonest cause is road traffic accidents (about 40 per cent). Regrettably, spinal cord injury from violence (either self-harm or criminal assault) is increasing, particularly in the United States. In the older age group, falls become a more common cause. Table 1 summarizes the leading causes in the United Kingdom and United States.

The proportion of injuries from road traffic accidents has seen a modest reduction in recent years, probably due to the introduction of seat-belt legislation and improved safety features on cars. Hopefully, the improved safety of vehicles and improved traffic regulation, particularly speed control in urban areas, will further reduce the incidence in coming years. There is some evidence in the last decade that the incidence of spinal injury has plateaued, having been increasing in previous decades. Regrettably, the commonest result of injury is tetraplegia (around 60 per cent), which, compared to paraplegia, is still increasing.

Early acute management

The appropriate management of the individual at the scene of an accident is vital to avoid unnecessary worsening of a spinal cord injury. If the individual is unconscious then it should be assumed there is an injury to the cervical spine until proven otherwise. Until this diagnosis can be ruled out the head and neck should, as far as possible, be held firmly in a neutral position. This is normally achieved at the scene of an accident by immobilization in a semi-rigid collar, but if this is not available alternative improvised methods of stabilizing the head and neck should be initiated. The individual should not be placed in the coma position as this will rotate the cervical spine, but is best placed, if other injuries allow, in a lateral position with the head kept in line with the spine by the underlying arm. If any movement is necessary the person should be 'log rolled' to ensure their spine is kept in a straight and neutral position at all times. Usually, transportation is on a spinal board with a head immobilizer. Speed of evacuation is important, particularly if there are other life-threatening injuries. Preferably, the individual should be transferred to the Regional Spinal Injuries Unit; but obviously the individual may need to be resuscitated, and other life-threatening injuries may need treatment at the nearest casualty department. It is worth recalling that the diagnosis of intra-abdominal injury can be very difficult in people with spinal cord injuries. The initial phase of spinal shock will tend to give rise to paralytic ileus and abdominal distension, which can further confuse the situation if an abdominal injury is suspected.

Obviously, a general and neurological examination is vital—particularly to determine the neurological level of the lesion. Figure 1 illustrates the myotomes, dermatomes, and reflexes as an *aide mémoire*. Table 2 summarizes the likely functional outcome according to lesion level.

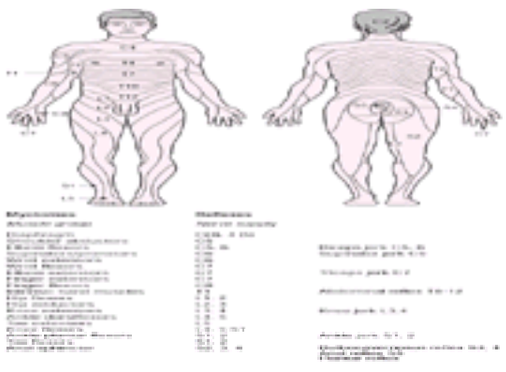

Fig. 1 An *aide mémoire* to examination—summary of the dermatomes, myotomes, and associated reflexes.

Spinal injury, however, can not be determined solely by examination and often there are very few local signs. There may be some bruising, tenderness, or deformity, but equally there may be no clue on examination as to the actual nature and extent of the underlying bone injury. Thus, although radiological investigation is essential, it should preferably only be undertaken in a unit familiar with the management of those with spinal injury. In a radiology department it is still important to remember that spinal movement must be minimal. Usually, radiography will clearly reveal the fracture or dislocation, although bony abnormalities are occasionally minimal or absent. This is particularly true in older people with underlying cervical spondylosis when tetraplegia can result from a hyperextension injury without fracture or dislocation. Radiological examination can also be normal in children when a spinal traction injury can occur without evidence of bony damage.

Initial management of people with injuries to the cervical spine usually consists of skeletal traction applied through skull callipers. Traction will help to stabilize and splint the spine and can also reduce fractures and dislocations. A number of different callipers are available, of which the Gardner–Wells calliper is but one type (Fig. 2).

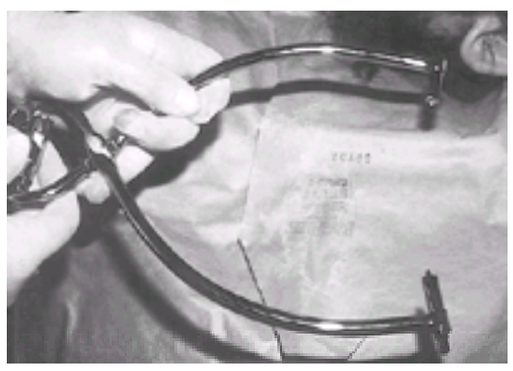

Fig. 2 Skull traction using Gardner–Wells calliper.

The amount of traction applied will vary according to the type and extent of injury, but it will be in the order of 2 kg for upper cervical injuries and somewhat more, around 4 kg, for lower cervical spinal injuries. Sometimes, if the spine is dislocated, reduction is achieved by incrementally increasing the weight of traction every few hours. Once the neck has reduced, a halo brace is a useful alternative to skull traction in many people and will allow early mobilization.

The standard treatment for thoracic and lumbar injuries is simple support of the individual in the correct posture, usually with a pillow under the lumbar spine to maintain the normal lordosis.

Surgical versus conservative treatment

In most cases, skull traction for cervical injuries and conservative postural treatment for thoracolumbar injuries is quite sufficient and operative intervention is unnecessary. It has long been a source of controversy whether operative intervention and fusion aids neurological recovery. Practice varies from country to country and indeed from centre to centre. In a broad-based survey in the United States 60 per cent of people underwent spinal surgery. Most individuals underwent fusion and internal fixation, but increasing numbers also now undergo anterior or posterior decompression of the spinal cord with or without internal fixation and fusion. However, practice in the United States tends to be more oriented towards surgical intervention than in the United Kingdom and other parts of Western Europe. In the United Kingdom, surgical intervention will tend to be reserved for those with unstable displaced fractures, whereas conservative management would be the normal practice for stable and/or undisplaced fractures. However, if the neurological symptoms are deteriorating then many spinal centres would now recommend surgical intervention.

Use of steroids

Another treatment intervention that can be considered in the very early stages after injury is a short course of high-dose methylprednisolone. There is some evidence that such intervention, started within 8 h of injury, improves the neurological outcome. However, this is not totally accepted and such practice not uniform. The results of further trials are awaited.

Management in the spinal cord injury centre

Initial management will consist of resuscitation, treatment of associated injuries, and stabilization of the spine, either conservatively or by surgical intervention. However, the individual should be transferred to a recognized spinal injury centre as soon as possible. There is clear evidence that the outcome is maximized, both physically and psychologically, if individuals are managed in such centres as opposed to a less co-ordinated and less experienced approach in another hospital setting.

Management of the spine

As mentioned above, the injured person will either be managed conservatively or surgically. The advantage of surgery is that the individual can be mobilized more quickly. If a conservative approach is adopted, mobilization and active rehabilitation is obviously difficult in the first few weeks. Cervical spine traction is normally maintained for around 6 weeks and then monitored for signs of bony union and stability. Once the fracture site is stable the individual can be gradually sat up in bed whilst continuing with cervical support. A profiling bed, which enables a more natural seated position to be adopted, is most useful. In the early few weeks a halo brace can be used instead of skull traction. The advantage of this brace (see Fig. 3) is to allow early mobilization. The halo brace is kept on for between 10 and 12 weeks until the site is stable. In those with thoracolumbar injuries, who are usually managed conservatively, the period of bed rest will usually last from 8 to 12 weeks followed by bracing and gradual mobilization, assuming that the fracture site is stable.

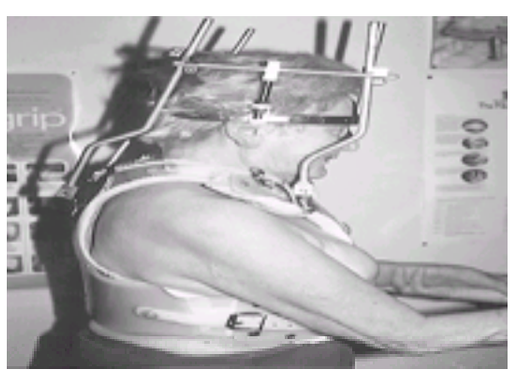

Fig. 3 Illustration of a halo brace.

A number of medical problems can occur over this initial period of immobilization.

Management of medical problems

Respiratory problems

Respiratory insufficiency can occur in people with cervical cord injuries. Intercostal muscles may be paralysed, and in high cervical lesions the diaphragm can also be paralysed. However, even in people with lower lesions, respiratory problems can still occur from associated injuries such as rib or sternal fractures. Since respiratory function can decline several hours or even days after injury, probably due to the development of spinal cord oedema, respiratory function should therefore be monitored carefully and ventilation provided if required. Regular chest physiotherapy is vital at this time. Since pulmonary embolism is also a risk, particularly in those immobilized, anticoagulation is advisable prior to mobilization.

Pressure sores

Regrettably, the development of pressure sores still occurs even with high-quality care. Sores are commonest where there are bony prominences near the skin: such as the ischial tuberosity, greater trochanter, sacrum, heel, and sometimes at the back of the head in those with skull traction. A key to prevention is awareness of the potential problem, with vigilance, regular changes of position in bed, and regular lifting in the wheelchair. A large range of commercial mattresses and wheelchair cushions are available that relieve pressure. Shear forces should be avoided as far as possible when lifting or positioning the patient; and obviously the individual should never be dragged over sheets or from the wheelchair. The skin should be kept clean, with particular care taken to avoid any urine or faecal soiling. If a sore does occur then the area must be kept clean, any dead tissue be removed, and there should be complete relief of pressure from that area until it is fully healed. Occasionally, surgery is indicated for larger or deeply infected sores that otherwise would take too long to heal. Education of the injured person and their family is essential. Despite awareness of the problem around 25 per cent of people still develop a pressure sore during the rehabilitation phase. About 15 per cent of people will develop a pressure sore in the first year following discharge—a figure that increases still further with time, such that by year 10 about 15 per cent of those with incomplete lesions and 28 per cent of those with complete lesions will have developed at least one pressure sore. Septicaemia from pressure sores is still responsible for around 10 per cent of deaths in people with spinal injury.

Bladder problems

In the early part of this century, problems, usually infection, of the urinary system were responsible for at least half the deaths of those with spinal injury. Although very significant progress has been made in the management of bladder and kidney problems, nevertheless urinary tract complications are still the leading residual cause of mortality and morbidity.

During the period of spinal shock the bladder is usually non-contractile, so that catheterization may be appropriate at this time. Once spinal shock begins to wear off the commonest problem is of detrusor hyperreflexia, which usually gives rise to the frequent passage of small quantities of urine associated with urgency. However, other possibilities include detrusor sphincter dyssynergia and detrusor hyporeflexia. The latter will tend to occur when there is damage to the S2, 3, and 4 sacral nerves.

The management of urinary problems usually involves obtaining satisfactory answers to three questions:

- Is there impairment of renal function?
- Is there a failure of bladder emptying?
- Is there detrusor hyperreflexia?

Is there impairment of renal function?

Screening of the upper urinary tract is important both in the short and long term. Intravenous urography should be used in the early months after injury, but long-term follow-up can often be carried out by renal ultrasound scanning or plain abdominal radiography. Late complications are possible, such as renal calculi. Cystometrography is also vital for determining the exact nature of the underlying bladder and sphincter problems.

Is there a failure of bladder emptying?

A residual urine volume greater than 100 ml is generally accepted as the level at which intervention is necessary. Residual urine can predispose to infection and stone formation, and contributes to impairment of renal function, particularly if the failure to empty is associated with a high intravesical pressure and back-pressure up to the kidney. Occasionally, failure of emptying can be managed by artificial stimulation such as suprapubic tapping or perineal stimulation. However, in most cases failure of bladder emptying requires mechanical drainage. The most useful method is intermittent, clean self-catheterization. This is carried out by the disabled person, or sometimes by a carer, four or five times every 24 h such that volumes in the bladder are kept to less than 500 ml. Intermittent self-catheterization has revolutionized the management of bladder problems in those with spinal injury. Anticholinergic drugs, such as propantheline, oxybutynin, or imipramine, may sometimes help to reduce detrusor activity. Condom drainage in the male is helpful in preventing leakage between catheterizations. A silastic indwelling catheter might need to be used if intermittent self-catheterization is impossible. However, suprapubic catheterization is far better in the long term and is associated with less problems. Regrettably, there are many problems of catheterization, including leakage, blockage, stone formation, and infection.

Is there detrusor hyperreflexia?

A small number of people can control minor problems with the detrusor hyperreflexia by rigid bladder drill, emptying the bladder at frequent and regular intervals. However, most people need some form of oral medication and, as above, anticholinergics are the most effective, and oxybutynin the most common—propantheline and imipramine are alternatives. Once again, protection against the embarrassment of leakage is often necessary and is more readily achieved in men with the use of condom drainage. A variety of absorbent pads can be worn by women. Advice from a specially trained nurse continence advisor can be invaluable whatever the nature of the problem.

A whole variety of surgical techniques may be applicable in particular circumstances. An endoscopic distal sphincterotomy can be useful for those with reflex bladder emptying. The technique of bladder augmentation with an ileocystoplasty can also be helpful to allow for sufficient capacity for intermittent, clean self-catheterization. Fortunately, urinary diversion techniques are now needed less frequently. Recent advances include artificial urinary sphincters for treating neuropathic incontinence. Some centres also now employ sacral anterior nerve root stimulators that can be used in some people with suprasacral cord lesions. For instance, the bladder can be emptied by activating a radio-linked implant to stimulate the S2, S3, and S4 anterior nerve roots. Occasionally, a similar implant can also be used to assist in defecation and in obtaining a penile erection.

Incontinence can be a major disability and handicap, and, indeed if not treated properly, the complications can be life-threatening. Long-term follow-up is essential, and proper management can make significant reductions in long-term risks and produce major improvements in the quality of life.

Bowel care

During the initial period of spinal shock the bowel remains flaccid and should not be allowed to overdistend, with the attendant risk of constipation and overflow incontinence. Manual evacuation is usually carried out until bowel activity returns. Eventually, reflex emptying can occur in those with predominant upper motor neurone lesions, but the bowel can remain flaccid in those with lower motor neurone involvement. In the former, bowel evacuation can usually be triggered by glycerin suppository or by anal digital stimulation. In those with flaccid bowel there is a continuing need to evacuate manually or by straining using abdominal muscles. Advice on proper diet is also required, with a good-quality, high-fibre diet being the most helpful.

Autonomic dysreflexia

This is a potentially fatal problem most commonly seen in those with cervical cord injuries above the sympathetic outflow, but it can occur in those with high thoracic lesions above T6. Autonomic dysreflexia is characterized by an exaggerated autonomic response to a stimulus below the level of the lesion. Stimuli can include

distension of the pelvic organs such as the bladder, colon, and rectum: such distension induces sympathetic activity resulting in vasoconstriction and hypertension. Other stimuli include catheterization, urinary infections, sexual intercourse, pressure sores, and even tight clothing; surgical procedures can also induce the reflex. Symptoms include headaches, sweating, vasodilatation, nasal obstruction, paraesthesia, and anxiety. Significant hypertension also occurs. The problem occurs in around 50 to 80 per cent of those at risk, with most cases occurring between 2 and 12 months' postinjury. Other than awareness of the problem and avoidance of the necessary stimuli, attention is directed to reducing blood pressure. Sublingual nifedipine can be used, or intravenous hydralazine in more severe cases. Chlorpromazine, nitroprusside, and diazoxide are also possibilities. Occasionally, the sympathetic reflex activity may have to be blocked by spinal epidural anaesthetic.

Spasticity and contractures

Spasticity occurs in an upper motor lesion with intact spinal reflex arcs below the level of the lesion. It is usually worse in those with incomplete lesions. Spasticity can be functionally useful and the individual can sometimes use flexor or extension spasms as an aid to dressing. However, spasticity usually produces functional problems as well as causing pain. In the long term there is a significant risk of muscle contractures. Initial management focuses on removing any unnecessary exacerbating factors such as pressure sores, tight catheter leg bags, or even urinary infections and constipation. Treatment should always involve an expert neurological physiotherapist who will advise on appropriate positioning and seating. In the early stages, passive stretching of the spastic muscles and regular standing regimes can be helpful; in the longer term, such regimes can often be taken on by the disabled person and their carers. Although antispastic drugs should always be used with care as they induce significant tiredness and weakness, they can provide some useful background antispastic effect. Baclofen, dantrolene, and tizanidine are the commonest prescribed. The latter is a more recently introduced drug, at least in the United Kingdom, and is an effective antispastic agent that appears to produce less weakness than the other available drugs. However, spasticity is often localized and focal treatment is more appropriate. Nerve blocks with phenol and alcohol can be used, but intramuscular botulinum toxin has recently proved very useful in the management of spasticity. The toxin is injected directly into the muscle and blocks the release of acetylcholine from nerve endings, which, over 2 or 3 days, produces a muscle relaxation that lasts for 2 to 3 months. Occasionally, more severe spasticity will need other treatment measures, such as the use of intrathecal baclofen. If contractures have resulted, surgical correction by tenotomy, tendon lengthening, or muscle division is often the only way to get the limb back into a functionally useful position. Aggressive early management of spasticity is important in order to maximize any neurological recovery and prevent unnecessary complications.

Heterotopic ossification

This term is used when bone develops in an abnormal anatomical position in soft tissues. The prevalence in spinal cord injury is reported to vary between 5 and 50 per cent. Heterotopic ossification commonly occurs around the hips and knees, causing a decrease in the range of movement as well as localized swelling and joint effusion. It normally occurs during the first few months after the injury and will only rarely begin later than 1-year postinjury. Unfortunately treatment is difficult. Etidronate disodium (Didronel) is probably the most useful treatment. Surgical intervention can be required in severe cases, but is usually unsatisfactory. Some centres now use prophylactic etidronate disodium for about a year.

Deep venous thrombosis

Deep venous thrombosis still remains a significant complication after spinal injury, with a small risk of death from pulmonary embolism. Heparin is generally used as a prophylactic but some centres now use external pneumatic calf compression.

Pain and dysaesthesia

Peripheral pain is quite common in the early weeks after injury. Although, burning pain can, unfortunately, also continue for some months, it usually responds reasonably well to the use of carbamazepine, tricyclic antidepressants, or gabapentin. Pain from other sources such as osteoarthritis can also occur. It should be remembered that people with spinal cord injury do not always appreciate pain or that it is manifested in different ways, such as autonomic dysreflexia or worsening of spasticity. Treatment modalities—for example, transcutaneous nerve stimulation, acupuncture, and psychological techniques, such as relaxation and hypnotherapy or alleviation of depressive illness—can all help. Spinal cord stimulation is occasionally used, as are surgical techniques, such as dorsal-root, entry-zone radiofrequency coagulation. Other causes of pain such as nerve root compression should also be borne in mind.

Rehabilitation

There is no evidence that rehabilitation can promote natural recovery, but there is ample evidence that a co-ordinated multidisciplinary team can improve functional outcome for the person with a spinal cord injury. The team can ensure that functional abilities are maximized and that physical and psychological complications are kept to a minimum. The co-ordinated team input is vital during the early weeks and months after injury, but it is equally important that the team maintains contact over the period of discharge and indeed into the longer term.

Principles of rehabilitation

There is no room in this chapter to dwell on the basic principles of rehabilitation. However, it is important to state that modern rehabilitation practice is somewhat different from other medical specialties. It is based on the principles of education and is a process in which the disabled person and the family must be involved for it to have any meaning. Rehabilitation should go beyond the narrower confines of physical disease to also deal with the psychological consequences of physical disability and with the social milieu in which the disabled person has to operate.

Rehabilitation is based around the concepts of impairment, disability, and handicap as outlined by the World Health Organization in 1980. 'Impairment' is simply a term that describes a loss or abnormality of psychological, physiological, or anatomical structure or function. Rehabilitation must go beyond impairment and should place such impairment within a functional context—the 'disability'. 'Handicap' describes the social context of disability. Rehabilitation can be defined as an active and dynamic process by which a disabled person is helped to acquire knowledge and skills in order to maximize physical, psychological, and social function.

The basic nature of rehabilitation is to work with the disabled person and their family in partnership. The professional should impart accurate information and advice, give guidance on prognosis and natural history, and help the individual to establish realistic goals in an appropriate social context.

A key to successful rehabilitation is goal-setting. The first goal should be a long-distance strategic aim. In the context of spinal cord injury this could, for example, include enabling the person to return to their previous home fully competent in wheelchair use. The overall strategic goal can also have a number of long-term subgoals in different spheres of life such as employment, home, and leisure. Once the long-term goal has been determined, steps will need to be defined to achieve that goal, which in turn will involve setting a number of short- and medium-term goals. These shorter term aims should be clearly stated. A useful mnemonic is SMART: that is, the goals should be specific, measurable, achievable, relevant, and time-limited. The implication of goal-setting is that the team, and indeed the disabled person, should know when the goals have been achieved. Thus, valid and reliable outcome measures are important tools; but it is neither possible nor desirable to outline such tools. The outcome measures will depend on the goals set. However, it is often useful to employ a general disability measure such as the Functional Independence Measure or, in the short term, the more physically oriented Barthel Score. Some of the standard scales employed in the field of spinal cord injury are frankly of little value in monitoring progress. First to be developed was the Frankel Score. This has now been largely superseded, at least in the United States, by the 1992 revised American Spinal Injury Association Classification (see Table 3). Although this scale is now widely quoted in the spinal cord literature (mainly in terms for helping to determine natural history and prognosis), it is not a tool for monitoring goal attainment.

Rehabilitation team

Medical input is obviously vital to the team, particularly during the early acute stages of management. Spinal cord injury consultants are now trained rehabilitation specialists who do not necessarily have surgical qualifications. However, spinal cord injury centres will always need input from spinal surgeons, as well as assistance from urologists and a variety of other medical consultants. Because of their 24-h daily contact with the injured person, nursing staff on the ward are clearly vital team members, many of whom possess additional spinal cord injury or other specialist qualifications; for example, in giving continence advice or in the management of sexual problems.

The physiotherapist's role during the very early stages of management is to minimize chest complications, particularly in those with high cervical cord lesions. Physiotherapy advice is helpful to ensure the patient is correctly positioned in bed and to prevent the complication of spasticity. Once a patient is beginning to

mobilize the physiotherapist is the key person to advise on the choice of wheelchair and for teaching the individual to become familiar with all aspects of its control. A number of advanced wheelchair skills will eventually be learnt, such as back-wheel balancing, to allow manoeuvrability over rough ground and up kerbs, and sideways jumping for manoeuvrability in a limited space. In people with lower cord lesions the physiotherapists can be involved in providing limited gait training using callipers and crutches. Orthotic devices, such as the reciprocating-gait orthosis (**RGO**) and hip-guidance orthosis (**HGO**), may be considered in some cases. The physiotherapist can also help in the context of handicap by encouraging and assisting with the development of sporting activities. However, this assistance clearly overlaps with the role of the occupational therapist.

The occupational therapist is usually concerned with assisting people to reach their highest level of physical and psychological independence, particularly with regard to personal care and appropriate adaptation of the home, work, and leisure environments. For example, the occupational therapist will be involved in the design of appropriate splinting (for example, writing or typing splints and feeding straps) to assist those with high cord lesions. Many increasingly sophisticated assistive technology devices are now available to enable even those with profound disabilities to remain reasonably independent. For instance, environmental control equipment will enable an individual to operate a door intercom, turn lights on and off, turn the pages of a book, control the television, and use a telephone and computer, etc. Even people with high tetraplegia can control these devices using mouth sticks or breath control. The occupational therapist will often be involved in giving such advice, particularly at the time of discharge back into the home environment.

If necessary, a psychologist may be particularly useful in enabling the person to make an emotional adjustment to their new disability.

The social worker is likely to be involved with the family as a whole, and only a small part of the job is to advise on disability benefits. Most of the social worker's task is to ensure that the disabled person and family integrate and adapt to the new disability as smoothly as possible.

At some point, others, such as vocational advisors, specialist nurses, and dieticians, will all need to be part of the comprehensive spinal injury team.

Long-term issues in spinal cord injury

Discharge home

A particularly difficult time for the injured person is discharge home. Often the person will have spent several weeks or months in a spinal cord centre and returning home can be a traumatic process both for the injured person and their family. Brief, trial home visits will almost certainly have been carried out beforehand. These visits are particularly important for ensuring that the house is appropriately adapted. Obviously in some cases a new house or bungalow will need to be purchased. A number of adaptations regarding access, both internal and external, hoisting gear, adaptations to the toilet, bathroom, and kitchen may all be required before the individual can return home. Environmental control equipment may need to be prescribed and installed. Psychological support is also vital over this period not only for the injured person but for their family. Anxiety and depressive illness are both quite common and will need active intervention. The community services and the primary care team will need to be involved. Planned discharges are vital and should involve a case conference between the hospital and community staff to ensure a smooth hand-over. However, at this time many spinal cord injured people will wish to move away from the more paternalistic hospital care that was important during the first few weeks' postinjury. Most will choose to live as independently as possible, albeit with the help of their family or a personal assistant. Advice on the available financial support is important. If financial compensation from a personal injury claim is ongoing then a solicitor can be helpful at this point to arrange interim payments from the Court towards the costs of home adaptations, transport, and personal care.

Emotional problems

Obviously there are profound changes in one's life following spinal cord injury. The refocusing of life ambitions can be a frustrating, anxious, or depressing time. The attitude of family and friends will have a further bearing over the period of adjustment. Regrettably, clinical depression is common and occurs at some point in at least 50 per cent of individuals. Suicide can also occur. Whilst such problems are not always preventable, anxiety, depression, and adjustment problems can be alleviated by appropriate intervention. Although the role of medication, at least in the short term, can be helpful, probably most assistance can be provided by cognitive therapy or other forms of counselling and psychological support. Contact with others in similar circumstances can often be helpful and may be facilitated through the various peer support groups.

Sexual life

Sexual ability depends on the level and completeness of the spinal lesion. Sexual readjustment is an important part of the rehabilitation process both for the injured person and their partner, regardless of gender. Self-image and self-confidence can be severely affected. Individuals should be counselled about the totality of sexuality as there is a tendency for discussions to focus on penetrative sexual intercourse. In both sexes, absence of genital sensation can be compensated for by the use of other erogenous zones such as the breasts, neck, and mouth. Orgasm is sometimes possible even in those with complete spinal cord lesions. In women, problems can result from the lack of vaginal lubrication. In men, various techniques and devices are available to restore erectile capacity. Most people with complete upper motor neurone lesions will have reflex but not psychogenic erections; however, these are often not always sustained or strong enough for intercourse. Reflex erections are usually impossible in those with parasympathetic lesions. A satisfactory erection can often be achieved either by the use of intracavernosal drugs or mechanical means such as vacuum erection aids and compressive retainer rings. However, the recent introduction of sildenafil (Viagra) is beginning to reduce the need for mechanical or injected assistance.

Fertility

Fertility is not usually reduced in women, although some can go through a time of amenorrhoea. However, fertility is generally reduced in men who have low sperm counts with diminished motility. Sometimes if ejaculation is not possible during intercourse it can be induced by direct stimulation or by electroejaculation. Fertility can also be improved by some of the modern assistive conception techniques such as *in vitro* fertilization and intracytoplasmic sperm injection. Women with spinal cord injury who become pregnant may have some problems in labour, particularly if the lesion is complete above T10. Autonomic dysreflexia is also a risk during labour. However, spinal cord injury is not by itself an indication for caesarean section.

Later medical complications

All the complications listed above in the acute phase can, of course, occur later. This is why it is so important for the multidisciplinary team to keep an overview of the individual in the long term. However, a few other problems are more likely to occur in the long term:

- Pathological fractures—there is a higher risk of osteoporosis in paralysed limbs and thus pathological fractures may occur with minimal trauma. For example, a minor fall from a chair or even a flexor spasm secondary to spasticity can result in a fractured leg. Treatment should usually be conservative.
- Post-traumatic syringomyelia—this occurs in about 4 per cent of people and consists of an ascending myelopathy due to secondary cavitation in the central part of the spinal cord. The problem is commonly delayed several years' postinjury. It presents with pain in the arm with a characteristic disassociated sensory loss: reduced pain and temperature sensation but preservation of proprioception. Motor loss of the lower motor neurone type occurs and occasionally sensory loss can spread up to the face (syringobulbia). Surgical treatment including decompression and drainage of the cavity may be necessary.
- Respiratory management—those with high cervical cord lesions with lost diaphragmatic function obviously require long-term ventilatory support. Modern portable ventilators can be readily mounted on a wheelchair. Speech is entirely possible with an uncuffed tracheostomy tube that allows air to escape to the larynx. In some people it is possible to implant a phrenic nerve stimulator to achieve diaphragmatic ventilation. Regrettably, it is still the case that individuals with long-term, ventilator-dependent requirements have significantly more morbidity and mortality than those with lower lesions.

Leisure pursuits

A wide variety of leisure pursuits are now possible for those with spinal cord injury. Although integration to able-bodied clubs and pursuits is obviously to be encouraged, a reasonable range of sports and other clubs exists for those with spinal injuries. Wheelchair skills can be finely tuned to develop expertise in a variety of sports. Physical access to leisure and social outlets is improving, albeit very slowly. Recent legislation, such as the Disability Discrimination Act in the United Kingdom, should further improve the situation.

Driving

Access to a motor vehicle is vital in modern society. Driving should be entirely possible for people with spinal cord injury, with the probable exception of those with very high cervical cord lesions. Automatic transmission is vital and hand controls are usually essential. Hand controls enable the individual to control the accelerator and brake functions from a lever or other device near the steering wheel. A variety of infrared devices to control secondary functions such as windscreen wipers, lights, and horn are now available. Very light-powered steering makes life easier for those with weak grip. Those with higher cord lesions who can still retain some useful shoulder and upper arm function can still drive a car using a variety of commercial devices attached to the steering wheel. A number of techniques can be taught to stow wheelchairs safely for those with paraplegia, and for those with higher lesions there are a number of mechanical wheelchair stowage devices. It is also quite possible to adapt a suitable vehicle to enable people to drive from their wheelchair. Financial advice is often required, combined with advice on the range and type of possible adaptations. The United Kingdom now has a number of driving assessment centres, often attached to rehabilitation centres.

Employment

Between 25 and 35 per cent of people with spinal cord injuries return to work, either in their original occupation or, after a period of retraining, to a new job. The chances of employment are higher in the younger population and in those who already had a job at the time of injury. There is also a positive correlation with the number of years of education. Fortunately, employment should become more prevalent as the ability to work at home becomes more readily acceptable. The individual should be encouraged to contact disablement employment advisors who can provide both advice and financial help for a return to work. In other cases, careers' advice or retraining to obtain new qualifications may be more appropriate.

Information

The key to independence is access to good-quality information. In most countries there are now voluntary organizations that can provide such information and advice. These organizations can also act as pressure groups, many of which have been instrumental in promoting increased awareness and improved legislation for disabled people. The Internet now provides an excellent source of information and advice, and training in computer literacy should certainly be encouraged by the rehabilitation team.

Conclusions

The management of spinal cord injury can pose a range of challenges to the multidisciplinary rehabilitation team. Although we cannot yet promote natural recovery in spinal cord injury, such interventions may be possible in the future. However, our failure to influence the natural history of a spinal cord injury should certainly not inhibit active and dynamic rehabilitation to enable the individual to resume as normal a life as possible. The application of modern rehabilitation practice, together with greater social awareness and understanding, has led to significant improvements in the overall survival and quality of life of people with spinal cord injuries.

Useful addresses

Spinal Injuries Association, 76 St James Lane, London N10 3DF. Tel: +44 (0) 181 444 2121. This association, for spinal cord injured people and all involved in their care, produces an excellent quarterly newsletter.

Further reading

Archives of Physical Medicine and Rehabilitation (1999). Spinal cord injury—current research outcomes from the model spinal cord injury care systems. *Archives of Physical Medicine and Rehabilitation* **80**, Special issue. [An entire edition of the journal devoted to spinal cord injury and current research outcomes from the model spinal cord injury care systems. A comprehensive review of the whole subject.]

Barbeau H, *et al*. (1999). Walking after spinal cord injury: evaluation, treatment and functional recovery. *Archives of Physical Medicine and Rehabilitation* **80**, 225–35. [A useful review article regarding modern developments to promote ambulation after spinal injury.]

Barnes MP, Ward AB (2000). *Textbook of rehabilitation medicine*. Oxford University Medical Publications, Oxford. [General background text in rehabilitation medicine for the undergraduate and junior postgraduate.]

Berkowitz M, *et al.* (1992). *The economic consequences of traumatic spinal cord injury*. Demos, New York.

Ditunno JF (1999). Predicting recovery after spinal cord injury: a rehabilitation imperative. *Archives of Physical Medicine and Rehabilitation* **80**, 361–3. [An up-to-date source of references regarding natural history.]

Giménez y Ribotta M, Privat A (1998). Biological interventions for spinal cord injury. *Current Opinion in Neurology* **11**, 647–54. [A useful and brief review article covering future possible interventions to promote natural recovery following spinal cord injury.]

Grundy D, Swain A (1996). *ABC of spinal cord injury*, 3rd edn. BMJ Publishing Group, London. [A very useful general text.]

Rehabilitation Institute of Chicago Procedure Manual (1994). *Spinal cord injury—medical management and rehabilitation*. Aspen, Maryland.[A thorough text on medical aspects of spinal cord injury.]

Trieschmann RB (1988). *Spinal cord injuries—psychological, social and vocational rehabilitation*, 2nd edn. Demos, New York. [One of the few volumes to thoroughly discuss the psychological, social, and vocational problems after spinal injury.]

Wade DT (1992). *Measurement in neurological rehabilitation*. Oxford University Press, Oxford. [An invaluable textbook outlining a number of important and useful outcome scales.]

World Health Organization (1980). International classification of impairments, disabilities and handicaps. World Health Organization, Albany, NY.

World Health Organization (1998). *The world health report—1998. Life in the 21st century vision overall*. WHO, Geneva. [A useful reference for a number of world health issues and in particular includes a discussion of the new classification of impairments, activities, and participation.]

24.13.18.1Intracranial tumours

Jeremy Rees

Introduction

Intracranial tumours comprise primary tumours that originate from the brain, cranial nerves, pituitary gland, or meninges and secondary tumours (metastases) that arise from organs outside the nervous system. These tumours present to many different specialists and their management is difficult because of their location and their variable clinical manifestations.

Aetiology

There are no known risk factors apart from prior irradiation to the skull and a few rare neurogenetic syndromes, such as neurofibromatosis (optic nerve glioma, meningioma, vestibular schwannoma) (Fig. 1), von Hippel–Lindau syndrome (haemangioblastoma), and Li–Fraumeni syndrome (glioma).

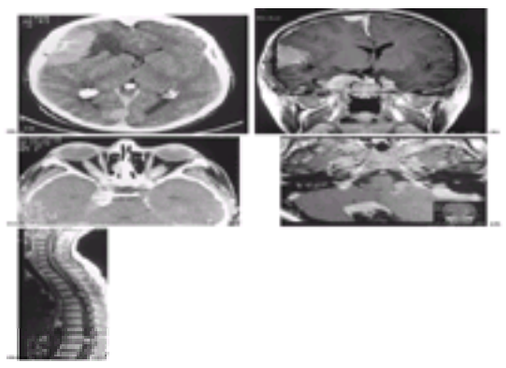

Fig. 1 Contrast-enhanced CT and MR scans of a patient with neurofibromatosis type 2 and multiple intracranial tumours. (a) CT of the brain with gadolinium enhancement showing a large right parietal convexity meningioma surrounded by vasogenic oedema exerting considerable mass effect. There is also a smaller falx meningioma in the right occipital region. (b) Coronal T_1-weighted MRI of the brain with gadolinium enhancement showing multiple meningiomas in the right temporoparietal region, right parafalcine region, and both cavernous sinuses. (c) Contrast-enhanced CT scan of the orbits showing bilateral optic nerve sheath meningiomas with intracranial extension into the right cavernous sinus causing a partial right third nerve and sixth nerve palsies. (d) Axial T_1-weighted MRI of the brain with gadolinium enhancement showing bilateral vestibular nerve schwannomas, and a large fourth ventricle tumour. (e) Sagittal T_1-weighted MRI of the spinal cord with gadolinium enhancement showing three discrete meningiomas encroaching on the spinal column at mid-cervical, mid-thoracic, and upper lumbar levels.

Epidemiology

Intracranial tumours represent the sixth most common neoplasm in adults (approximately 8 per cent of all primary neoplasms) and the second most common neoplasm in children. After stroke, intracranial tumours are the leading cause of death from neurological disease in the United Kingdom.

Based on a recent Scottish study, the crude annual incidence for primary intracranial tumours is 15.3 per 100 000 and for secondary tumours 14.3 per 100 000 population. There is evidence that the incidence is increasing, particularly in elderly patients. Different tumour types present at different ages. Supratentorial gliomas are uncommon below the age of 30 years but become increasingly prevalent thereafter. The most frequent tumours of middle life (third and fourth decades) are astrocytomas, meningiomas, pituitary adenomas, and vestibular schwannomas, while glioblastoma multiforme and metastases are more frequent in the fifth and six decades of life. In contrast, children tend to have infratentorial tumours: 70 per cent of childhood primary intracranial tumours originate below the tentorium cerebelli, whereas in adults the figure is only 25 per cent. There is a strong female preponderance of meningiomas and schwannomas, whereas slightly more men have astrocytomas.

Pathogenesis

Certain genetic lesions are associated with brain tumours. Chromosomal deletions—particularly chromosome 10, which contains multiple tumour suppressor genes—are found in astrocytic tumours, occurring in up to 70 per cent of glioblastomas. Mutations of a tumour suppressor gene *p53*, located on chromosome 17p, have also been reported in approximately 40 per cent of astrocytic tumours. In general the accumulation of predictable genetic alterations is associated with increasing malignant progression.

Clinical features

With increasing sophistication of neuroimaging, tumours are being detected at an earlier stage than before. Patients typically present with one or more of four clinical syndromes:

- progressive neurological deficit
- seizures
- raised intracranial pressure
- altered mental states.

The particular combination of clinical features varies depending on the location, histology, and rate of growth of the tumour. For instance, patients with low-grade gliomas present typically with a seizure disorder that may remain static for many years, while patients with malignant gliomas typically develop a rapidly progressive

neurological deficit and raised intracranial pressure

Progressive neurological deficit

Focal neurological symptoms due to brain tumour are typically subacute and progressive with over 50 per cent of patients having focal signs by the time of diagnosis. Cortical tumours produce contralateral weakness, sensory loss, dysphasia, dyspraxia, and visual field loss depending on their location. Posterior fossa tumours cause ataxia and cranial nerve palsies. Vestibular schwannomas cause progressive unilateral deafness followed by ipsilateral facial sensory loss. Pituitary tumours may cause a bitemporal hemianopia if there is chiasmal compression or endocrine disturbances due to either hypopituitarism or hypersecretion of specific hormones.

Seizure disorder

Brain tumours account for about 5 per cent of epilepsy cases although they are over-represented in cases of intractable epilepsy. Seizures are the presenting symptom in 25 to 30 per cent of patients with gliomas and are present at some stage of the illness in 40 to 60 per cent overall.Approximately half the patients have focal seizures and the other half have secondarily generalized seizures. Low-grade gliomas are associated with seizures in over 90 per cent of cases and these frequently remain the only complaint for many years. Conversely, malignant gliomas have a lower frequency of seizures, presumably because of their more rapid growth and destructive characteristics. In these patients, seizures are associated with a better prognosis. Seizures are also common initial manifestations of meningiomas (40 to 60 per cent) and metastases (15 to 20 per cent). Supratentorial tumours and those superficially located are particularly likely to cause seizures, particularly in the frontal and temporal lobes. A Todd's paresis, which may persist, is an uncommon but characteristic feature of tumour-associated epilepsy. About 10 per cent of patients presenting *de novo* in status epilepticus have an underlying tumour.

Raised intracranial pressure

Intracranial tumours increase intracranial pressure either by a direct mass effect, by provoking cerebral oedema, or by producing obstructive hydrocephalus. The most common symptom of raised intracranial pressure is headache, which is the presenting symptom in 25 to 35 per cent of patients; papilloedema is found in up to 50 per cent of patients with headache due to tumours. The classic picture of headache, vomiting, and visual obscurations (transient fogging of vision usually on rapid changes in posture) due to raised intracranial pressure is well known and easily recognized, but most patients present before this develops. Less than 1 per cent of patients presenting with isolated headache has a brain tumour.

Most brain tumour headaches are intermittent and non-specific and may be indistinguishable from tension headaches. Supratentorial tumours typically produce frontal headaches, while posterior fossa tumours usually result in occipital headache or neck pain. Certain features of a headache are suggestive but not pathognomic of raised intracranial pressure. These include headaches that wake the patient at night or are worse on waking and improve over the course of the day.

Mental state changes

These are an uncommon presentation of brain tumours occurring in about 20 per cent of patients at diagnosis. Personality changes may initially be quite subtle and may show themselves as an inability to cope at work. In these cases it is essential to obtain a collateral history from relatives or colleagues at work.

Pathology

Neuroepithelial tumours (predominantly gliomas) account for approximately 50 to 60 per cent of all primary brain tumours. The other common types are meningiomas (20 per cent), pituitary adenomas (15 per cent), vestibular schwannomas (5 per cent), and primary central nervous system lymphomas (5 per cent). The most common sites of origin of secondary tumours are lung (50 per cent), breast (15 per cent), melanoma (10 per cent), and unknown (15 per cent).

The gliomas are a family of neoplasms that arise from astrocytes, oligodendrocytes, and ependymal cells. Astrocytomas are the most common type of glioma and are infiltrating neoplasms composed of fibrillary astrocytes. Almost all of these tumours have the propensity to undergo anaplastic change to a more malignant lesion. Thus a fibrillary astrocytoma (Fig. 2) progresses to an anaplastic astrocytoma (Fig. 3) and then to the most malignant form, glioblastoma multiforme. This process occurs more often and more rapidly in older patients. The grading systems that have been used have attempted to describe degrees of anaplastic change and thereby correlate the histological appearances with prognosis. The most widely accepted classifications of gliomas are the World Health Organization three-tiered system and the St Anne–Mayo grading system as shown in Table 1. These systems have been retrospectively applied to large series of patients and have been shown to provide reproducible and prognostically useful information. A rare type of astrocytoma is the oligodendroglioma characterized by the presence of uniform round nuclei with small nucleoli. This also has the propensity to undergo anaplastic change but unlike anaplastic astrocytomas, oligodendrogliomas are frequently chemosensitive (see below).

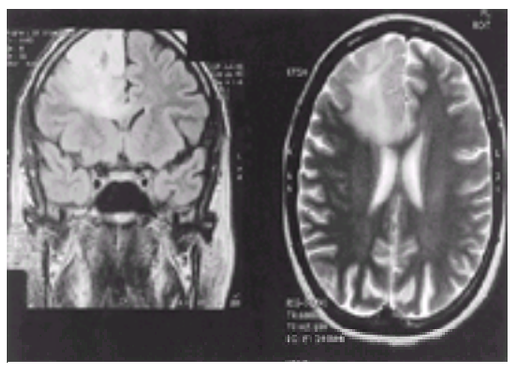

Fig. 2 Coronal and axial T_2-weighted MRI of the brain showing a diffuse lesion in the right frontal lobe which returns high signal. It is seen extending from the cortex into the deep white matter and infiltrating across the corpus callosum. There is mass effect causing compression of the frontal horn of the lateral ventricle. The tumour did not enhance with gadolinium. This patient presented with generalized seizures and has remained well after 3 years of follow-up. Biopsy revealed a fibrillary astrocytoma (WHO Grade II).

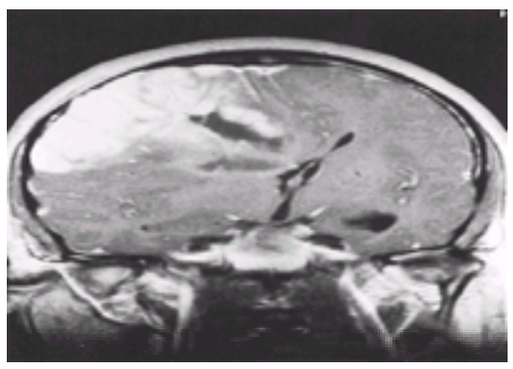

Fig. 3 Coronal T_1-weighted MRI of the brain with gadolinium enhancement showing a large heterogeneous enhancing tumour arising from the right frontal lobe exerting considerable mass effect in a patient presenting with a 2-month history of complex partial seizures, headaches, and papilloedema. The lesion was resected and shown to be an anaplastic astrocytoma.

Diagnosis

The diagnosis of a brain tumour is made by a combination of CT/MR scanning and pathological examination of either a biopsy or resection specimen. Newer techniques include magnetic resonance spectroscopy and metabolic imaging (single photon and positron emission tomography). These may permit a non-invasive method of differentiating between low-grade and high-grade gliomas and between tumour recurrence and radiation damage.

Treatment

The three methods of treatment for brain tumours are surgery, radiotherapy, and chemotherapy. The use of each is dictated by the location of the tumour, the likely histology, and the patient's age and general condition.

Surgery

Recent advances in tumour neurosugery include the use of computerized neuronavigation techniques, improved pre- and intraoperative mapping of eloquent brain areas using functional magnetic resonance imaging (fMRI), and cortical mapping. These have all contributed to improving the morbidity and mortality of neurosurgery but an effect on overall survival has not been demonstrated.

Surgery is indicated as a first-line treatment for meningiomas, non-secreting pituitary adenomas, and vestibular schwannomas. The role of surgery in the management of primary intracranial tumours, particularly gliomas, is more controversial. Some types of glioma, for example pilocytic astrocytomas, can be cured by surgical resection. For most types, however, removal is not curative. While surgery is of undoubted benefit in relieving the symptoms and signs of raised intracranial pressure or an evolving focal deficit, there are no prospective randomized data to support its use for prognostic purposes alone. However, it may be beneficial in a subgroup of patients who are young, fit, and who have a tumour in a non-eloquent region, such as the non-dominant frontal lobe. Overall, about 50 per cent of patients with medically refractory seizures derive considerable seizure reduction from surgery.

There is evidence that a combination of surgery and radiotherapy offers a survival advantage over radiotherapy alone for the treatment of solitary metastases in patients whose systemic cancer is well controlled.

Radiotherapy

Radiotherapy is the only treatment which has been proved to extend survival in patients with primary malignant brain tumours. Radiotherapy provides useful palliation in patients with low-grade gliomas, but there is no evidence to suggest that early radiotherapy prolongs overall survival compared with radiotherapy given at the time of tumour progression. Meningiomas are also partially radioresponsive and should be treated with radiotherapy where there is atypical or malignant histology or where there is recurrent tumour which is not surgically accessible.

Advances in technology have allowed greater accuracy of radiotherapy delivery and, in particular, the use of stereotactic frames which permit the focusing of radiation to a small tumour with minimal dosage to the surrounding normal tissue. This can be done either in a single high dose (stereotactic radiosurgery) or in smaller fractions (stereotactic radiotherapy) and is predominantly indicated for lesions less than 3 cm in diameter which are well circumscribed, extra-axial, and more than 5 mm away from vital structures.

Chemotherapy

There has been increased awareness of the chemosensitivity of certain tumours, particularly anaplastic oligodendrogliomas and primary lymphomas of the nervous system in adults and diencephalic gliomas in children. Approximately two-thirds of anaplastic oligodendrogliomas respond dramatically to a combination of treatment with procarbazine, lomustine, and vincristine, and this is now the first-line treatment for these tumours, particularly in the group who have combined deletions of chromosomes 1p and 19q. The addition of methotrexate-based chemotherapy to cranial irradiation markedly improves disease control and survival of patients with primary lymphomas. Adjuvant nitrosurea chemotherapy is used in patients with malignant gliomas although it offers only a marginal survival advantage. Recently, a new oral alkylating agent, Temozolomide, has been approved for use as a second-line treatment for recurrent malignant glioma. There is no chemotherapy that is effective for the treatment of meningiomas.

H3>Prognosis

Neither earlier diagnosis of tumours nor advances in treatment over the last decade have significantly changed the overall prognosis of primary brain tumours. The median survival of glioblastoma multiforme without treatment is 3 months and with radiotherapy about 1 year. Anaplastic astrocytomas are associated with a median survival of 18 months. Young age and good performance status are the most important prognostic factors.

The outlook for patients with low-grade gliomas is considerably better with a median survival of 5 to 10 years depending on age, performance status, and histology. Oligodendrogliomas are more chemosensitive than astrocytomas and have a more indolent course, so their prognosis is correspondingly better with patients surviving 10 to 15 years after diagnosis.

At least 40 per cent of primary intracranial tumours are extra-axial (not arising from within the brain substance itself) and are thus readily treatable, if not curable. Some tumours such as meningiomas and pituitary adenomas are associated with 10-year survival of over 90 per cent if diagnosed before irreversible neurological damage has occurred.

Further reading

Cairncross JG *et al.* (1998). Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *Journal of the National Cancer Institute* **90**, 1473–9. [First study to show definitive correlation between molecular genetic analysis and chemoresponsiveness of brain tumours.]

Counsell CE, Collie DA, Grant R (1996). Incidence of intracranial tumours in the Lothian region of Scotland, 1989–90. *Journal of Neurology, Neurosurgery and Psychiatry* **61**, 143–50. [Epidemiological study in Scotland showing incidence rates more than twice those previously reported in the United Kingdom.]

Daumas-Duport C *et al.* (1988). Grading of astrocytomas, a simple and reproducible method. *Cancer* **62**, 2152–65. [A 15-year follow-up study of a previously used grading system showing very good correlation between histological critieria and survival.]

DeAngelis LM *et al.* (1992). Combined modality therapy for primary CNS lymphoma. *Journal of Clinical Oncology* **10**, 635–43. [Non-randomized study showing significant improvement in disease-free survival in patients treated with chemotherapy in addition to radiotherapy.]

Forsyth P, Posner JB (1993). Headaches in patients with brain tumours, a study of 111 patients. *Neurology* **43**, 678–83. [Descriptive study of 111 patients with brain tumour headaches showing that the 'classic' early morning brain tumour headache is uncommon.]

Greig NH *et al.* (1990). Increasing annual incidence of primary malignant brain tumours in the elderly. *Journal of the National Cancer Institute* **82**, 1621–4. [Study showing up to a 500 per cent increase in incidence rates of malignant brain tumours in the elderly from the early 1970s to the mid-1980s, which may be despite more extensive uptake of imaging.]

Kleihues P, Burger PC, Scheithauer BW (1993). *Histologic typing of tumours of the central nervous system.* Springer-Verlag, New York. [Definitive pathological typing system for brain tumours.]

Patchell RA *et al.* (1990). A randomised trial of surgery in the treatment of single metastases to the brain. *New England Journal of Medicine* **322**, 494–500. [Randomized trial of surgery and radiotherapy against radiotherapy alone showing increased survival in surgical patients (median 40 compared with 15 weeks).]

Quigley MR, Maron JC (1991). The relationship between survival and extent of the resection in patients with supratentorial malignant gliomas. *Neurosurgery* **29**, 385–9. [Meta-analysis of over 5000 patients with malignant gliomas treated surgically showing little correlation between extent of resection and survival.]

Shaw EG, Scheithauer BW, O'Fallon JR (1997). Supratentorial gliomas, a comparative study by grade and histological type. *Journal of Neuroncology* **31**, 273–8. [Detailed analysis of survival and correlation with histology in over 500 patients with gliomas.]

Walker MD *et al.* (1978). Evaluation of BCNU and/or radiotherapy in the treatment of anaplastic gliomas. A cooperative clinical trial. *Journal of Neurosurgery* **49**, 333–43. [First randomized trial confirming survival benefit of patients with malignant gliomas treated with radiotherapy.]

24.13.18.2Traumatic injuries of the head

Laurence Watkins and David G. T. Thomas

Epidemiology

It is estimated that each year in the United Kingdom approximately 1 million people attend hospital after a head injury. Almost half of these are children under 16 years old. Head injuries cause 9 deaths per 100 000 population per year in the United Kingdom. This represents 1 per cent of all deaths, but 15 to 20 per cent of deaths for those between 5 and 35 years old. Since mainly young people are affected, the prevalence of disability caused is very significant, with an estimated 135 000 people in the United Kingdom dependent on care after brain trauma.

In 1986, the Royal College of Surgeons of England published guidelines on the provision of surgical services for patients with head injuries. More recently, there have been concerns that inappropriate treatment might be leading to unnecessary death and disability. This possibility, together with increasing public expectation, led to a further working party which published updated guidelines in 1999. The availability of CT scanning has also increased, so that now it is considered essential for all hospitals which admit patients with head injuries to have 24-h CT scanning facilities.

Basic concepts

Primary and secondary injury

Primary injury is the damage caused to the brain at the moment of impact. It encompasses diffuse axonal injury and focal contusions. Medicine has little to offer for primary injury; prevention, however, is a major concern for health and safety legislation, town planning, and traffic laws (such as the compulsory wearing of seat belts and crash helmets). The focus of medical intervention is the prevention of secondary damage.

The causes of secondary brain damage can be divided into extracranial (hypoxia and hypotension) and intracranial (haematoma, brain swelling, and infection).

Grading the severity of injury

Only 20 per cent of patients are admitted to hospital and most of these are discharged in less than 48 h. About 1 in 500 of the patients attending hospital will develop intracranial haemorrhage. The doctor's task is to manage patients in such a way that the few with preventable causes of secondary injury are identified and treated effectively.

The British Society of Rehabilitation Medicine defines three broad groups depending on their Glasgow Coma Score (**GCS**) after initial resuscitation:

1. Mild—GCS 13 to 15;
2. Moderate—GCS 9 to 12; and
3. Severe—GCS 3 to 8.

This is a useful categorization for decision-making in head injury management. It should not be confused with other schemes, which are generally retrospective and used for epidemiological and statistical purposes.

The golden hour

Taking into account the practicalities of CT scanning, interhospital transfer, and preparation for theatre, the time available for initial assessment, resuscitation, and treatment of other injuries in the emergency department is less than 1 h. This is sometimes referred to as 'the golden hour' in which rapid action is critical to the patient's outcome.

In a typical series of patients who had surgery for acute subdural haematoma, over 70 per cent had a functional recovery (good recovery or moderate disability) if the delay from injury to operation was less than 2 h. If the delay was between 2 and 4 h, just over 60 per cent made a functional recovery. In contrast, for those whose operation was more than 4 h after the injury, less than 10 per cent made a functional recovery (Fig. 1).

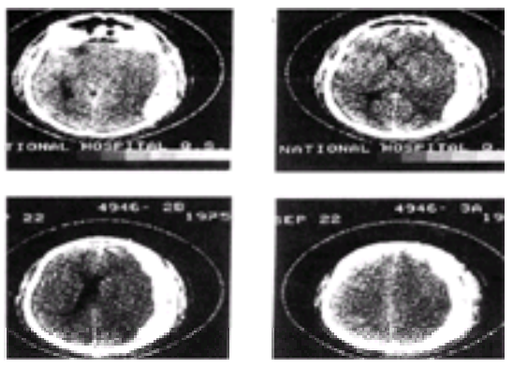

Fig. 1 Typical CT scan appearances of acute subdural haematoma. Fresh haemorrhage appears hyperdense (white). A subdural haemorrhage conforms to the surface of the brain, typically in a thin crescent. There is effacement of the lateral ventricle on the side of haematoma and midline shift away from it. An extradural haematoma, in contrast, usually appears biconvex, with well-defined edges since it is confined between the bone and dura.

Patients who 'talk and die'—the importance of deteriorating conscious level

A classic paper, by Jennett and his team, coined the phrase 'talk and die' to describe patients whose primary injury was mild, but who succumbed to secondary injury: usually an intracranial haematoma. Deterioration in conscious level is an urgent clinical sign that requires immediate action.

The Glasgow Coma Score (Table 1) is now widely used in the United Kingdom and elsewhere, giving objective recording of conscious level, with a high correlation between different observers. Any deterioration is thus more likely to be noticed. When communicating about a patient with head injury, it is good practice to specify observations of each parameter, rather than to use the corresponding numerical scores, which are open to misinterpretation.

Change in conscious level is the most useful clinical sign in head injury assessment. Generally, a patient with primary brain injury shows a gradually improving conscious level. A patient whose conscious level deteriorates is very likely to be suffering secondary brain injury and therefore requires further investigation and treatment. Conscious level must therefore be assessed at the earliest opportunity, and then reassessed at frequent intervals.

Early management of the patient with head injuries

Extracranial injuries

Life-threatening extracranial injuries always take priority over the head injury. However severe the head trauma, the patient needs to be stabilized for safe transfer. Also, hypotension and hypoxia are important causes of secondary brain injury. Time-consuming definitive surgery such as the internal fixation of limb fractures should, however, be postponed if possible.

Airway, breathing, and circulation are the first priorities. Management should follow the general recommendations taught in the Advanced Trauma Life-Support (ATLS) courses. In particular, assessment should include consideration of respiratory problems, shock, and possible internal injuries.

All patients with head injury should be assumed also to have a cervical spine injury until proven otherwise. Cervical immobilization should be established, unless the patient is fully conscious, co-operative, and able to convince the examining doctor that he has no neck pain or tenderness, a full range of cervical movement, and no neurological deficit. There are rare exceptions to this guideline: for example, a patient with a fixed flexion deformity due to ankylosing spondylitis might present with a cervical fracture; in that circumstance placing the neck in a 'neutral' position, in a cervical collar, might actually produce neurological injury.

Initial management of head injuries

After initial assessment, resuscitation, and stabilization of extracranial injuries, the patient is graded for the severity of their head injury. These categories then give a useful broad guide to management.

Severe

If the head injury is severe (GCS 3 to 8) then a member of the team should immediately refer to a neurosurgical unit. If the patient's best motor response is localization or obeying, then they may not necessarily require ventilation, provided that oxygen saturation can be maintained at more than 95 per cent, the PCO_2 at less than 6 kPa, and the PO_2 at more than 12 kPa on 40 per cent inspired oxygen. If the patient's best motor response is flexion or worse, or if any of the above criteria are not met, then the patient should be electively intubated and ventilated prior to transfer. Ventilation should be adjusted to maintain the PCO_2 in the range 4.0 to 4.5 kPa. At this stage, the intracranial pressure is unknown, but should be assumed to be high; therefore a mean arterial pressure of at least 90 mmHg should be maintained.

Whether a CT scan is performed at the referring hospital or on arrival at the neurosurgical unit will depend on the local availability of scanning facilities. This decision is based on whichever pathway is likely to produce the fastest response, given local conditions.

If, after discussion with the neurosurgical unit, a patient is accepted for transfer, they should be accompanied by personnel able to insert and manage an endotracheal tube and ventilation.

Moderate

If the head injury is moderate (GCS 9 to 13), then an urgent CT scan would be advisable. If the CT scan detects an intracranial abnormality, then urgent neurosurgical referral is appropriate and the immediate management will be similar to that for severe head injuries given above. If no abnormalities are detected on CT scan, care should be taken to exclude metabolic and other causes of reduced conscious level (such as hypoglycaemia or drug overdose). If it appears that diffuse brain injury is the only cause of depressed conscious level, then the care of the patient is discussed with the neurosurgical unit. In some cases transfer will be advised, while in others observation under the care of the emergency department will be appropriate. This will depend on local resources and practices. In either situation, if conscious level remains depressed at 48 h, the patient should be transferred to a neurosurgical unit for further assessment.

Mild

Most head injuries are mild (GCS 14 to 15). After initial assessment, the next decision is whether further investigation is required and whether this should be a skull radiograph and/or CT scan.

Patients who have a GCS of 15, have no history of loss of consciousness, and have none of the following criteria for investigation may be considered for discharge according to the local head injury protocol. They must be under the supervision of a responsible adult and written information must be provided concerning symptoms and signs which would warrant seeking further urgent medical advice.

In this context, the criteria for skull radiography include:

- GCS 14
- history of loss of consciousness or amnesia
- scalp swelling
- scalp laceration (to bone or more than 5 cm in length)
- high-energy mechanism of injury
- headache and/or vomiting which is not improving with time
- significant maxillofacial injuries.

In children, additional criteria include:

- fall from a height greater than twice the height of the child
- fall on to hard surface
- tense fontanelle
- any suspicion of non-accidental injury.

If no skull fracture is detected, then the patient should be observed until conscious level is normal and associated symptoms have resolved. If a fracture is seen on skull radiography, then the patient will require CT scanning and admission to hospital.

In addition to the presence of a skull fracture on the plain radiograph, the following are indications for CT scanning:

- focal neurological deficit
- Battle sign (bruising over the mastoid)
- periorbital haematoma ('raccoon eye' bruising)
- subconjunctival haemorrhage with no posterior limit
- blood or cerebrospinal fluid in ears or nostrils
- suspicion of penetrating injury
- seizure
- anticoagulation or known coagulopathy
- difficulty in assessment, whether due to extremes of age (very young or very old) or intoxication.

If the CT scan shows no abnormality, the patient should be admitted for observation for at least one night, and longer if symptoms persist. If the CT scan does show an intracranial abnormality, the care of the patient should be discussed with the neurosurgical unit. In most cases, transfer to the neurosurgical unit will be advised.

Management of intracranial complications

Intracranial haematoma

In almost all cases of intracranial haematoma, urgent evacuation is indicated, bearing in mind that the longer the delay, the greater the risk of death or disability. The above guidelines for observation/skull radiograph/CT scan/transfer to neurosurgical unit are all aimed at the earliest diagnosis of the minority of patients with an intracranial haematoma.

The risk of a traumatic intracranial haematoma depends on conscious level and whether a skull fracture is present (Table 2).

Infection

Meningitis and brain abscess can develop following any head injury in which a communication has been made between the environment and the intracranial contents. The most obvious example is a compound depressed fracture, where comminuted bone fragments have been forced inwards, breaching the dura. With some penetrating injuries (such as a fall on to a sharp object or assault with a pointed weapon) the visible wound may be small and appear insignificant. Since the injury may have been low velocity, the patient may have a deceptively normal conscious level. Such patients should always be referred for neurosurgical assessment.

A closed depressed fracture does not require surgery except for cosmetic reasons if it is on a visible part of the skull.

Cerebrospinal fluid rhinorrhoea or otorrhoea indicates that a skull base fracture has breached the dura. This places the patient at risk of meningitis while the cerebrospinal fluid leak continues. Ninety per cent of such cases close spontaneously within 2 weeks, and usually neurosurgical intervention is not considered until this time has elapsed. An exception is a fracture of the posterior wall of the frontal sinus, visualized on CT scan. Such cases should be discussed with the neurosurgeon or the craniofacial team (if one exists locally) with a view to possible early anterior fossa repair.

The use of antibiotics in cerebrospinal fluid leaks is controversial. In practice, most neurosurgical units still prescribe a penicillin or cephalosporin for 1 week or until the cerebrospinal fluid leak stops (which ever is longer). A working party reviewing the literature, however, concluded that the available evidence does not support the use of prophylactic antibiotics in patients with cerebrospinal fluid fistulas.

Follow-up and late complications of head injury

Cognitive symptoms

After head injury there is a variable period before memory function returns and ongoing memories again begin to be stored. This period is referred to as post-traumatic amnesia and is a useful measure of the severity of brain damage. For example, when questioned after recovery, a patient may not remember the accident but clearly recalls being placed on a stretcher and taken into the ambulance: this would suggest a relatively short post-traumatic amnesia of a few minutes. The post-traumatic amnesia is fixed for a given injury and memories of this period do not later 'recover'.

It is also common for a patient to lose memory of events immediately before the injury. This is known as retrograde amnesia. Unlike post-traumatic amnesia, the period of retrograde amnesia often progressively reduces as the patient recovers.

Incomplete recovery following head injury has behavioural, cognitive, emotional, social, and economic effects. For adults with severe head injuries, 85 per cent remained disabled at 1 year following the accident. In the intermediate group, 63 per cent remained disabled at 1 year. Even those with so-called 'minor' injuries can face considerable problems: at 3-month follow-up 79 per cent still have headaches, 59 per cent have symptomatic memory impairment, and 34 per cent have not returned to work.

The most widely used measure of outcome after head injury is the Glasgow Outcome Scale (Table 3). These are broad categories, which miss the subtleties of impairment in many who have had mild injuries, but its wide adoption and recognition make the Glasgow Outcome Scale invaluable for statistical comparisons.

Even 'mild' injuries, with early brief loss of consciousness and an initial GCS of 14 to 15, can lead to significant symptoms that can interfere with return to previous activities. These 'postconcussional symptoms' include headache, dizziness, poor concentration, memory impairment, and personality change. The patient's relatives often report personality changes, such as 'bad temper' and lack of motivation. Such symptoms usually improve over 6 months, especially if the patient and family are warned to expect such problems and reassured that they are eventually likely to resolve.

Rehabilitation after severe head injury requires multidisciplinary input from rehabilitation neurology, physiotherapy, occupational therapy, speech therapy, and neuropsychology. Other specialists and therapy services are accessed as appropriate for each individual patient. At least as far as the Glasgow Outcome Scale is concerned, 60 per cent of patients reach their final outcome category by 3 months after the injury. Ninety per cent reach their final score by the end of 6 months.

Epilepsy

Epilepsy is more common if there has been an intracranial haematoma, a depressed skull fracture, or post-traumatic amnesia of more than 24 h. A single seizure, within 1 week the injury, is of less significance than repeated seizures or those occurring after the first week. Any patient who has had a seizure, a craniotomy, or depressed skull fracture should be advised not to drive or operate dangerous machinery. They should also contact the Driving and Vehicle Licensing Authority.

Chronic subdural haematoma

The initial injury may have seemed very minor and may have occurred many weeks previously. The most common symptom is headache, progressively worsening and eventually accompanied by vomiting. There may also be a focal deficit, which can vary in severity. Increasing intracranial pressure may lead to cognitive impairment and eventually a depressed level of consciousness.

Whatever the pathophysiology, the treatment of choice is evacuation of the subdural collection and irrigation with isotonic saline at body temperature. This is a relatively small operation, which can even be performed under local anaesthetic, so even advanced age and general frailty do not contraindicate its use.

Hydrocephalus

Hydrocephalus occasionally occurs after head injury, particularly if there has been traumatic subarachnoid or intraventricular haemorrhage. It can be distinguished

from post-traumatic cerebral atrophy by the CT scan appearances: in hydrocephalus, the sulci will be small or effaced relative to the large ventricles and there may be periventricular lucency due to interstitial oedema.

Further reading

American College of Surgeons Committee on Trauma (1997). *Advanced trauma life-support for doctors. Student course manual*, 6th edn. American College of Surgeons, Chicago.

British Society of Rehabilitation Medicine (1998). *Rehabilitation after traumatic brain injury*. British Society of Rehabilitation Medicine, London.

Commission on the Provision of Surgical Services (1986). *Report of the working party on head injuries*. Royal College of Surgeons, London.

Infection in Neurosurgery Working Party of the British Society for Antimicrobial Chemotherapy (1994). Antimicrobial prophylaxis in neurosurgery and after head injury. *Lancet* **344**, 1547–51.

McMillan T, Greenwood R (1991). *Rehabilitation programmes for the brain injured adult: current practice and future options in the UK.* A Discussion Paper for the Department of Health. Department of Health, London.

Mendelow AD, Teasdale GM, Jennett B (1983). Risks of intracranial haematoma in head injured adults. *British Medical Journal* **287**, 1173–6.

Reilly PL *et al.* (1975). Patients with head injury who talk and die. *Lancet* **ii**, 375–7.

Rimel RW *et al.* (1981). Disability caused by minor injury. *Neurosurgery* **9**, 221–8.

Seelig JM *et al*. (1981). Traumatic acute subdural haematoma. Major mortality reduction in comatose patients treated within 4 h. *New England Journal of Medicine* **304**, 1511–18.

Teasdale GM (1995). Head injury. *Journal of Neurology, Neurosurgery and Psychology* **58**, 526–39.

Working Party on the Management of Patients with Head Injuries (1999). Report of the Working Party on the Management of Patients with Head Injuries. Royal College of Surgeons, London.

24.13.19 Benign intracranial hypertension

N. F. Lawton

Synonyms

Pseudotumour cerebri and idiopathic intracranial hypertension are synonymous with benign intracranial hypertension.

Definition

Benign intracranial hypertension is a syndrome of raised intracranial pressure occurring in the absence of an intracranial mass lesion or enlargement of the cerebral ventricles due to hydrocephalus. The synonyms pseudotumour cerebri and idiopathic intracranial hypertension have both been preferred because the outcome is not invariably benign. Although rarely life-threatening, the rise in intracranial pressure may result in permanent visual loss due to optic nerve damage.

Incidence

Benign intracranial hypertension is a rare disease. The incidence is approximately 1 in 100 000 in the general population but rises to 19 in 100 000 in obese women of childbearing age. The disease is certainly more common in females, the preponderance over males ranging from 3:1 to 8:1. Although benign intracranial hypertension may occur in infants and the elderly, it is primarily a disease of young women between the ages of 17 and 44 years. Very rarely, it is familial and may occur in more than one generation.

Clinical features

It is the hallmark of benign intracranial hypertension that presenting symptoms and signs are those of raised intracranial pressure alone. The diagnosis should not be entertained in the presence of neurological features which suggest a focal lesion. Furthermore, there is a remarkable preservation of consciousness and intellectual function rarely encountered in patients with mass lesions or hydrocephalus. A history of epilepsy, either generalized or focal, virtually excludes the diagnosis of benign intracranial hypertension, although seizures may occur in the small group of patients with venous sinus thrombosis (see below). Preservation of cerebral function also distinguishes benign intracranial hypertension from acute viral or bacterial meningoencephalitis. atients with benign intracranial hypertension routinely present to outpatient departments and become a medical emergency when papilloedema is seen.

Headache

This is the most common symptom and is present to some degree in virtually every case. Characteristically the headache is typical of raised intracranial pressure. It is then generalized, throbbing, worse on waking, and aggravated by factors which temporarily increase cerebrospinal fluid pressure such as straining, coughing, or changing posture. Not infrequently, however, headache is mild and non-specific so that the distinction from common tension headache may be difficult. At presentation, headache has usually been present for weeks, although sometimes for months. Although up to 50 per cent of patients complain of nausea, typical early morning projectile vomiting is rare.

Obesity

Among the medical conditions associated with benign intracranial hypertension, obesity is sufficiently common to be a characteristic feature. Up to 90 per cent of patients in reported series are overweight, although a history of rapid weight gain immediately prior to the onset is unusual.

Papilloedema

This is a virtually universal finding and the importance of fundus examination in every patient with headache cannot be overemphasized. Papilloedema is usually moderate and may be unilateral. Occasionally the appearance of the optic discs may be equivocal, and fluorescein angiography is indicated to demonstrate the characteristic leakage of dye in true papilloedema.

The classic symptom of papilloedema, which is not specific to benign intracranial hypertension, is a transient obscuration of vision, often described as a fleeting greyness, a halo, or a more vivid episode of 'Catherine wheels' lasting for a few seconds. Obscurations may be provoked by straining or a change in posture, but may also occur spontaneously. Persistent blurring of vision may also occur and patients may describe scotomas in the field of vision associated with optic nerve damage. Occasionally, sudden and permanent loss of vision results from infarction of the optic nerve.

Visual obscurations, persistent blurring, or scotomas are reported by 30 to 70 per cent of patients. A history of obscurations is often only elicited by direct questioning.

Visual field defects

Visual field analysis is the essential investigation in the examination and follow-up of patients with benign intracranial hypertension. The most common defects are enlargement of the blind spots, generalized constriction of the fields, and scotomas caused by optic nerve damage. There may be a predilection for visual field loss in the inferior nasal quadrants.

Diplopia

About 30 per cent of patients complain of horizontal diplopia due to sixth nerve palsy, which may be bilateral. The cause is a false localizing sign of raised intracranial pressure.

Aetiology

In the majority of patients with benign intracranial hypertension no cause can be identified. Many clinical associations have been reported, but these may have occurred by chance. Preceding minor head injury and intercurrent infections come into this category. Furthermore, the known associations are rare with the exceptions of obesity and the predilection for females. A positive family history, vitamin deficiency, and drugs are each a factor in less than 2 per cent of cases.

Dural sinus thrombosis

Before the advent of antibiotics, benign intracranial hypertension was frequently associated with chronic middle-ear disease complicated by dural sinus thrombosis. The term 'otitic hydrocephalus' was coined to describe this syndrome on the erroneous assumption that ventricular enlargement was present.

Although true 'otitic hydrocephalus' is now rare, the syndrome of benign intracranial hypertension may still occur following venous thrombosis in dural sinuses or in the extracranial jugular system. Sinus thrombosis may complicate pregnancy, the use of oral contraceptives, head injury, venous occlusive disease due to hypercoagulability states, dehydration due to any cause, or mediastinal obstruction. Sinus thrombosis should be suspected clinically when the onset of headache is sudden and accompanied by focal signs or impaired consciousness. Occasionally, the syndrome of benign intracranial hypertension is a late presentation of undiagnosed sinus thrombosis. In the majority of patients, however, venous obstruction is not the predisposing cause and the cerebral venous system is normally patent.

Menstrual disorders

Apart from the complication of venous thrombosis, benign intracranial hypertension is associated with pregnancy *per se*. It is not clear whether an association with menstrual irregularity is more than would occur by chance in obese young women. An association with menarche has been reported.

In spite of the clinical associations which suggest an underlying disorder of female endocrinology, hormonal studies have not shown a consistent abnormality. The pituitary–adrenal axis is intact and occasional abnormal responses may be due to obesity rather than benign intracranial hypertension. Thyroid function and prolactin secretion both appear to be normal. Cerebrospinal fluid vasopressin levels are raised, but this is not specific and may occur in a variety of neurological diseases. Reports of a specific increase in cerebrospinal fluid oestrone, which might link benign intracranial hypertension with obesity because adipocytes are the major source of oestrone, have not been confirmed.

Deficiency states

A rare cause of benign intracranial hypertension in children is hypovitaminosis A due to generalized nutritional deficiency or malabsorption. In such cases the condition responds specifically to vitamin A supplements. Poisoning with vitamin A due to excessive consumption of fish or animal liver may also cause benign intracranial hypertension.

Drug-induced benign intracranial hypertension

Both tetracycline and the retinoids isotretinoin and etretinate, which are vitamin A derivatives, may cause benign intracranial hypertension during long-term treatment for acne. These drugs should not be used in combination. All-*trans*-retinoic acid in the treatment of acute promyelocytic leukaemia is also a cause in this category. Other drugs occasionally responsible for the syndrome include nalidixic acid, nitrofurantoin, and lithium. Corticosteroids may lead to benign intracranial hypertension during their withdrawal after chronic treatment and the syndrome may occur in Addison's disease.

Empty sella

It has been suggested that this association in about 4 per cent of cases is caused by raised intracranial pressure in combination with incompetence of the diaphragma sellae. The theory is supported by the finding of raised pressure at lumbar puncture in some patients with empty sella, suggesting chronic benign intracranial hypertension as the underlying cause. Clinical hypopituitarism does not occur, but occasionally the empty sella may harbour a prolactinoma.

Pathogenesis

The mechanism by which intracranial pressure rises is poorly understood and the contribution of various factors controversial. Since the intracranial contents are housed in a rigid container, an increase in cerebrospinal fluid pressure may result from an increase in blood volume, swelling of the brain parenchyma, or an increase in the cerebrospinal fluid volume due to overproduction or malabsorption. There is little evidence to suggest that increased blood volume or cerebrospinal fluid production are important factors.

Swelling of the brain parenchyma

Direct evidence of a swelling due to cerebral oedema is slight and a single report of oedematous changes in brain biopsies has not been confirmed. However, the tendency for the ventricles to be small may indicate an increase in cerebral volume secondary to leakage from the cerebral vascular bed rather than transudation of cerebrospinal fluid from the ventricular system. Brain imaging in benign intracranial hypertension does not show the periventricular leakage of cerebrospinal fluid that occurs in hydrocephalus. Although there is currently no direct evidence for vasogenic cerebral oedema, this factor cannot be ignored, because it is one mechanism for raised pressure which does not anticipate some degree of hydrocephalus.

Decreased absorption of cerebrospinal fluid

A defect of cerebrospinal fluid absorption is widely regarded as the important factor in the pathogenesis of benign intracranial hypertension. Apart from those cases in which there is dural sinus thrombosis, it is assumed that the defect lies in the arachnoid villi of the superior sagittal sinus where the bulk of cerebrospinal fluid absorption takes place. The delayed clearance of radio-iodinated human serum albumin from the ventricular system after injection into the lumbar subarachnoid space is indirect evidence of reduced cerebrospinal fluid absorption. Simultaneous cannulation of the superior sagittal sinus and the subarachnoid space has shown increased resistance to cerebrospinal fluid absorption in the majority of cases. Manometry has shown consistent hypertension in venous sinuses but it is not clear whether this is primary or secondary to raised intracranial pressure. Finally, vitamin A deficiency in rats and cows produces a rise in intracranial pressure associated with diminished absorption of cerebrospinal fluid and histological changes in the arachnoid villi which are reversible with vitamin A supplements.

The absence of hydrocephalus in benign intracranial hypertension has been cited as an objection to the theory of reduced cerebrospinal fluid absorption. It is probably more significant, however, that sinus thrombosis may cause the syndrome of benign intracranial hypertension by preventing cerebrospinal fluid absorption, and is similarly associated with normal or small ventricles.

Investigations

The diagnosis of benign intracranial hypertension can only be confirmed by measurement of cerebrospinal fluid pressure, but in suspected cases it is essential to exclude a mass lesion or hydrocephalus before proceeding to lumbar puncture.

Radiology

Characteristically, CT scanning shows small and slit-like cerebral ventricles which may increase in volume as intracranial hypertension resolves. A similar appearance is seen on magnetic resonance imaging. Sagittal sinus thrombosis may be visualized on CT scanning as the characteristic 'empty delta' sign due to clot within the sinus. MRI is far superior to CT and provides graphic images of sinus thrombosis. Occasionally MR or CT angiography may be needed to exclude sinus thrombosis or conventional venography if thrombolytic therapy is contemplated.

Cerebrospinal fluid pressure

At lumbar puncture the opening pressure is greater than 200 mm cerebrospinal fluid, but it is important to note that in simple obesity the cerebrospinal fluid pressure may be as high as 250 mm. The diagnostic significance of cerebrospinal fluid pressure must therefore be correlated with the clinical picture. In the few patients whose cerebrospinal fluid pressure is equivocal, continuous monitoring may demonstrate intermittent peaks of raised pressure.

Cerebrospinal fluid analysis

The composition of the cerebrospinal fluid in benign intracranial hypertension is entirely normal, and the presence of white cells or a raised protein concentration cast serious doubt on the diagnosis. An exception to this rule is the rare syndrome resembling benign intracranial hypertension which occurs in association with postinfective polyneuropathy and with spinal tumours. Both conditions may lead to raised intracranial pressure with papilloedema and normal-sized ventricles but a marked rise in cerebrospinal fluid protein. The syndrome may also complicate cryptococcal meningitis and meningoradiculopathy in HIV infection.

Management

Patients given a diagnosis of benign intracranial hypertension are usually bewildered and frightened. It is important to provide a simple explanation of the nature of the condition and the rationale for treatment. Anticoagulation with heparin may be effective in the treatment of acute sinus thrombosis, emphasizing the importance of angiographic diagnosis in this small group of patients. With the further exception of rare cases due to drug treatment, the management of benign intracranial hypertension is aimed at the symptomatic reduction of intracranial pressure to protect vision and relieve headache. The methods available are difficult to evaluate because of the high spontaneous remission rate and the lack of controlled trials. Choice of treatment is further complicated by the absence of reliable risk factors for visual loss. In particular, the height of the cerebrospinal fluid pressure at diagnosis is of no prognostic significance.

In the past, repeated therapeutic lumbar puncture every 2 to 5 days has been shown to reduce cerebrospinal fluid pressure temporarily and may occasionally lead to spontaneous remission. When acute medical treatment is indicated, prednisolone (40 to 60 mg daily) is effective in relieving headache and visual obscuration due to papilloedema. However, steroids are unsatisfactory as long-term treatment because of their complications, especially in obese young females. For this reason diuretics are widely used in patients with mild symptoms. Acetazolamide or a thiazide diuretic may relieve headache, but the efficacy of diuretics in preventing slowly progressive visual loss is unproven.

Because of the limitations of medical treatment, an increasing number of patients are treated surgically, progressive visual field loss and unrelieved headache being the indications for operation. Because of the difficulty of tapping small cerebral ventricles, a lumboperitoneal shunt is usually favoured. Unfortunately, the technical failure rate of lumboperitoneal shunts is high and surgical revision may be required in up to 20 per cent of patients. For this reason the alternative operation, in which the optic nerve sheath is decompressed, has recently been revived, particularly in North America. This procedure produces rapid improvement in papilloedema, occasionally in both eyes after unilateral surgery. It is not clear whether long-term improvement is due to the creation of a cerebrospinal fluid fistula into the orbit or fibrosis of the meninges preventing transmission of the high cerebrospinal fluid pressure to the optic nerve head. However, headache is often unrelieved by this procedure, and lumboperitoneal shunting may not be avoided. There is also a risk of further visual loss postoperatively, which is much less common after a shunt procedure.

Currently it would seem reasonable to begin treatment with diuretics, reserving steroids as a temporary medical treatment in patients with severe symptoms. If surgery becomes necessary, lumboperitoneal shunting seems a logical procedure, with resort to optic nerve sheath decompression in the event of repeated shunt failure. Occasionally, subtemporal decompression may be a last therapeutic resort.

Although the efficacy of weight loss *per se* has not yet been established, a weight-reducing diet is recommended in obese patients. In patients with extreme obesity, gastric bypass surgery has reportedly relieved intracranial hypertension.

Pregnancy

The main threat is to the fetus and the rate of spontaneous abortion is increased. Spontaneous remission of benign intracranial hypertension during pregnancy has been the rule, although the number of reported cases is small. It would seem reasonable to begin treatment with diuretics, although steroids and shunt operations may occasionally be required.

Prognosis

Benign intracranial hypertension is a chronic condition in most patients, but spontaneous relapse and remission of symptoms is common. There is evidence that raised intracranial pressure may be found at follow-up lumbar puncture in patients whose symptoms have been in remission for several years. Mortality from benign intracranial hypertension is nil in most series, although an underlying sagittal sinus thrombosis may lead to a fatal outcome. Permanent visual loss occurs in up to 50 per cent of patients and is a significant disability in 10 per cent. Because the choice of treatment is determined primarily by progression of optic nerve damage, serial visual field analysis is the important yardstick of clinical progression.

Further reading

Ahlskog JE (1982). Pseudotumour cerebri. *Annals of Internal Medicine* **97**, 249–56.

Corbett JJ, Thompson HS (1989). The rational management of idiopathic intracranial hypertension. *Archives of Neurology* **46**, 1049–51.

Janny P *et al*. (1981). Benign intracranial hypertension and disorders of CSF absorption. *Surgical Neurology* **15**, 168–74.

McComb JG (1983). Recent research into the nature of cerebrospinal fluid formation and absorption. *Journal of Neurosurgery* **59**, 369–83.

Rush JA (1980). Pseudotumour cerebri: clinical profile and visual outcome in 63 patients. *Mayo Clinic Proceedings* **55**, 541–6.

Sergott RC, Savino PJ, Bosley TM (1988). Modified optic nerve sheath decompression provides long term visual improvement for pseudotumour cerebri. *Archives of Ophthalmology* **106**, 1384–90.

24.14.1 Bacterial meningitis

*D. A. Warrell, J. J. Farrar, and D. W. M. Crook**

Bacterial meningitis, also known as pyogenic, purulent, or cerebrospinal meningitis, is an inflammation of the leptomeninges with infection of the cerebrospinal fluid (**CSF**) within the subarachnoid space of the brain and spinal cord, and the ventricular system.

Anatomy of the subarachnoid space

In the absence of pathological blockages, bacteria entering the subarachnoid space at any point can spread over the surface of the brain and into the perivascular spaces of Virchow–Robin. After reaching the basal cisterns they can pass through the foramina of Luschka and Magendie into the fourth ventricle, and thence through the cerebral aqueduct of Sylvius to reach the third ventricle, and through the interventricular foramina of Monro to the lateral ventricles. Subdural empyema (pachymeningitis interna) or effusion complicating leptomeningitis can also spread freely because the arachnoid membrane and dura mater are almost entirely separated. By contrast, because of the tight application of the dura mater to the periosteum of the skull, epidural collections of pus are localized. In the spinal column, however, the epidural space is loose and contains fat, permitting the extension of a posterior spinal epidural abscess over several vertebral segments. Within the subarachnoid space and intraventricular system, infection may produce blockages of CSF circulation, especially at the various foramina or in the aqueduct, causing obstructive hydrocephalus or spinal block. If reabsorption of CSF across the subarachnoid granulations is prevented by a subarachnoid haematoma or empyema or thrombosis of the intracranial veins and venous sinuses, communicating hydrocephalus will result. In patients with meningitis, intracranial hypertension may be the result of cerebral oedema, the ventricular dilatation of hydrocephalus, or subdural or epidural collections of pus. Obstructive hydrocephalus and intracranial collections of pus carry a special risk of producing brain herniation after lumbar puncture.

Because they cross the inflamed basal meninges, the cranial nerves and cerebral blood vessels may be damaged. Cranial nerves may be compressed by intracranial hypertension (VI) or brain herniation (III) or suffer ischaemic damage from vasculitis. Cerebral veins and arteries may thrombose.

Acute bacterial meningitis

Classification

Pyogenic bacterial meningitis occurs in a number of clinical situations, each of which is associated with a particular pattern of infecting organisms, clinical presentation, and outcome. Spontaneous meningitis is the most important category and can be divided into neonatal meningitis or meningitis of childhood and adulthood. Post-traumatic meningitis follows neurosurgery or fractures of the skull. Device-associated meningitis complicates the use of CSF shunts and drains. Infection may also be considered as community acquired or nosocomial (hospital acquired) (Table 1).

Aetiological agents and epidemiology

The bacterial species that cause meningitis vary by geographical region and according to the categories mentioned above. Age and local social conditions influence the attack rate and mortality of spontaneous meningitis.

Neonatal meningitis is usually caused by three species: group B streptococci (*Streptococcus agalactiae*); K1 capsulate *Escherichia coli*; and *Listeria monocytogenes*. A wide range of other organisms has been reported to cause the disease. Infection mostly occurs in the postpartum period, but can occur as late as 6 weeks after birth. Prolonged rupture of membranes and low birth weight are important risk factors.

Spontaneous community-acquired meningitis in children (under 14 years of age) is usually caused by *Neisseria meningitidis*, *Streptococcus pneumoniae*, or *Haemophilus influenzae.* However, national implementation of conjugated *H. influenzae* type b (**Hib**) capsular vaccine immunization programmes by many countries during the 1990s has dramatically reduced, or nearly eliminated, Hib meningitis. The introduction of the conjugate vaccine against *Neisseria meningitidis* Group C in 1998 reduced the incidence of meningococcal meningitis in England and Wales. Similarly, pneumococcal conjugate vaccination in the United States is reducing the incidence of pneumococcal meningitis in this age group. The highest attack rate of all three bacterial species is in children under 1 year of age and falls off rapidly with increasing age. The decrease in susceptibility with increasing age results from the acquisition of protective immunity, mainly as a result of nasopharyngeal carriage.

In most countries, more than 50 per cent of cases of spontaneous community-acquired meningitis in adults are caused by *N. meningitidis* and *S. pneumoniae* (Table 1). *Listeria monocytogenes,* aerobic Gram-negative bacilli (such as *Escherichia coli*), *H. influenzae*, and *Staphylococcus aureus* cause most of the remaining cases. The attack rate of endemic *N. meningitidis* meningitis is usually low (1–5 cases/10^5 persons per year), but occasionally the incidence of the infection may increase and even reach epidemic proportions (for example, >300 cases/10^5 persons per year). Crowding is thought to play a role in the epidemics occurring in military recruits, South African miners, and other groups of people crowded together in closed environments. The attack rate of *N. meningitidis* disease may increase secondarily to epidemics of influenza A. However, the precise origin of the major epidemics affecting countries such as Brazil (in the 1970s) and regions such as sub-Saharan Africa remains unexplained. The bacterial capsule plays a role in determining the pattern of invasive disease caused by *N. meningitidis.* Capsulate serogroups A, B, and C occur sporadically and cause outbreaks of invasive disease, while serogroups Y, W, Z, W-135, and 29-E cause only occasional cases.

The attack rate of *S. pneumoniae* meningitis (1–2 cases/10^5 persons per year) is remarkably constant around the world. It increases in patients over 70 years of age. A high proportion of pneumococcal cases exhibit an associated infective focus. Otitis media or sinusitis is found in approximately 30 per cent of cases and pneumonia in up to 25 per cent. Hypogammaglobulinaemia (primary or secondary, for example in nephrotic syndrome and chronic lymphocytic leukaemia), sickle-cell disease,

splenic dysfunction, and previous trauma to the skull (see below) are risk factors for developing pneumococcal meningitis.

Streptococcus suis (Group R haemolytic streptococcus) serotype 2 is an important cause of meningitis (and rarely infective endocarditis, and septicaemia) in the Far East (Hong Kong, Thailand, Vietnam) and in other countries. Infection is related to occupational contact with pigs or pork, but the precise epidemiology remains poorly understood. In Holland the incidence of *S. suis* meningitis among abattoir workers and pig breeders was 3.0/100 000 per year. It is now the commonest cause of adult bacterial meningitis in Hong Kong and Vietnam. Possible routes of entry include skin abrasions, found in 40 per cent of patients, and upper respiratory and gastrointestinal tracts. Splenectomized patients are particularly at risk, as with other capsulated Gram-positive organisms.

Worldwide, *Listeria monocytogenes* accounts for few cases of meningitis, with an attack rate of approximately 0.2 to 0.4 cases/10^5 persons per year or 1 to 5 per cent of the cases of meningitis. Increased attack rates have been associated with contaminated foods such as unpasteurized soft cheeses, pâté, and poorly refrigerated precooked chicken. People at the extremes of age, pregnant women, and those with altered host defence mechanisms from prolonged immunosuppression with corticosteroids or alkylating agents such as azathioprine are at increased risk of listeriosis. *Staphylococcus aureus* causes 1 to 5 per cent of the cases with spontaneous meningitis and usually occurs in association with infective endocarditis. Spontaneous cases of *H. influenzae* meningitis, both capsulate type b and non-capsulate strains, account for up to 5 per cent of adult cases of meningitis. Aerobic Gram-negative (for example, *E. coli*) meningitis occurs especially in aged, debilitated, and diabetic people. The source in these infections is usually thought to be the renal tract.

Post-traumatic meningitis occurs in patients with skull or spinal injuries (for example, skull fractures) or in those who have undergone head and neck or spinal surgery. It usually arises in association with a CSF leak and soon after injury, but may occur many years after the trauma. The risk of developing meningitis is as high as 25 per cent with a clinically apparent CSF leak. The aetiology depends on whether the infection is acquired nosocomially or in the community. The majority of hospital-acquired infections are caused by aerobic Gram-negative bacilli, such as *E. coli, Klebsiella pneumoniae,* other Enterobacteriaceae, *Acinetobacter* spp., and *Pseudomonas* spp. Less commonly, *S. pneumoniae, H. influenzae, S. aureus,* and other normal upper respiratory tract flora cause meningitis in patients in hospital. Post-traumatic meningitis acquired in the community is caused mainly by *S. pneumoniae* (>90 per cent) and *H. influenzae.*

Device-associated meningitis is a well-recognized entity occurring in patients with CSF drains and CSF shunts. Most infections are nosocomial and are caused by coagulase-negative staphylococci (50–60 per cent) and *S. aureus* (15–30 per cent). Aerobic Gram-negative bacilli, *Streptococcus* spp., *Corynebacteria* spp., and *Propriobacterium acnes* are encountered. These infections usually present within a few months of inserting the device. Occasionally, *S. pneumoniae, N. meningitidis,* and *H. influenzae* are responsible.

Recurrent meningitis is an unusual (<10 per cent of meningitis) but well-recognized clinical category. Such cases frequently have either an underlying anatomical defect (for example, CSF leak or spina bifida) or an immunological defect. The immune deficiencies that most often predispose to recurrent meningitis are hypogammaglobulinaemia and complement deficiencies. Consideration should be given to vaccinating such patients against the most common pathogens.

The increasing incidence of human immunodeficiency virus (**HIV**) infection has altered the presentation and pattern of aetiological agents causing meningitis. A large seriesof adult patients withmeningitis who presented totheQueen Elizabeth Central Hospital in Blantyre, Malawi, was reported in 1975. At that time,meningitis comprised 2.5 per cent ofmedical admissions, the most common pathogens being*N. meningitidis* and *S. pneumoniae*. Sincethen, the population ofMalawi has been very severely affected by the **AIDS** (acquired immunodeficiency syndrome) pandemic, and theHIV seroprevalence of antenatalwomen has climbed steadilythrough the 1980s to the present level of more than 30 per cent. Thechanged overall pattern inthis series is probablydue to the influenceof HIV infection; ina survey of 153patients with invasive pneumococcaldisease, HIV seroprevalence was 95 per cent. In South Africa, HIV-infected children have more antibiotic-resistant isolates and a different clinical presentation compared to HIV-uninfected children. In adults, the HIV epidemic was found to be responsible for increasing chronic infections such as tuberculous and cryptococcal meningitides.

Pathogenesis

The acquisition of infection and mode of invasion of the CSF vary with the type of meningitis. However, once infection is established, the inflammatory injury and pathophysiology are remarkably similar in all types of meningitis. Important steps in the pathogenesis of spontaneous meningitis are nasopharyngeal colonization and mucosal adherence and invasion involving receptors, bacterial fimbriae or pili, encapsulation, and other virulence factors that are blocked by secretory IgA and other host defences such as the complement system. Within the subarachnoid space, these defences are inadequate as there is no complement and concentrations of IgG antibodies are low. The cell walls of Gram-positive bacteria and lipopolysaccharides of Gram-negative bacteria cause inflammatory change, thereby increasing vascular permeability and leading to the development of cerebral oedema.

The organisms that cause neonatal meningitis are acquired by the baby from the vagina and perineum during delivery, or from the environment soon after birth. The three main infecting species, *S. agalactiae* (group B streptococci), *E. coli,* and *L. monocytogenes,* invade the host, cause septicaemia, and, as a result, produce meningitis. An unusual feature of *E. coli* and many *S. agalactiae* strains (capsular types K1 and III, respectively) is that their capsules consist of polysialic acid. The association of this unusual type of capsule with two virulent strains suggests a role in the pathogenesis of neonatal meningitis.

Causative organisms of spontaneous meningitis, *S. pneumoniae, H. influenzae,* and *N. meningitidis,* are acquired by person-to-person spread. Replication in the nasopharynx is the essential first step before invasion. Asymptomatic carriage implies a stable and well-adapted relationship between microbe and host. Fortunately, this is the usual outcome. Invasion and infection of the host resulting in disease represents a rare and potentially catastrophic breakdown in the relationship between the bacterium and host. Invasion of the host is particularly likely to occur early after acquisition of the organism, before the host has developed protective immunity. Carriage is sufficient to produce immunity and resistance to disease. The greatest risk of disease, therefore, is in the first few years of life, at a time when the non-immune host first encounters these pathogens. The precise anatomical site of invasion is not known for all three pathogens. Animal studies suggest that the nasopharynx is the probable site of systemic invasion for *H. influenzae* meningitis, associated with escape into the bloodstream of a single organism that multiplies and produces septicaemia. In a proportion of these bacteraemic cases, bacteria then gain access to the CSF. Invasion of the CSF is probably dependent on the concentration of organisms in the blood and on the species causing bacteraemia (for example, bacteria such as enterococci can cause high-intensity bacteraemia under some conditions but the organisms seldom enter the CSF, whereas *N. meningitidis, H. influenzae,* and *S. pneumoniae* frequently invade the CSF). The choroid plexus, a highly vascular tissue, is probably the site of CSF invasion. Once organisms have entered the CSF and multiplied, purulent meningitis is inevitable.

Organisms causing post-traumatic meningitis invade the CSF directly through an anatomical defect. Bacteraemia, which is common, is secondary to the meningitis. Shunt-associated meningitis is caused mainly by organisms that colonize the skin and contaminate the surgical wound and prosthetic material at the time of surgery. The infected shunt becomes coated with a film of adherent bacteria, commonly referred to as a 'biofilm', which is not susceptible to clinically achievable levels of antibiotic. Such infections are usually incurable unless the foreign material is removed.

Once bacteria have gained access to the subarachnoid space they multiply in the CSF relatively uninhibited by host defences. Neutrophils, which rapidly accumulate in the infected CSF, have little inhibitory effect on the growth of the infecting bacteria in experimental animals. Complement is found in such low concentrations in CSF that it is unlikely to have an antibacterial effect. Investigation of the mediators of this inflammatory response is incomplete, but studies of pneumococcal and *H. influenzae* meningitis have identified important features of this inflammatory reaction. The pneumococcal cell wall has been shown to be a potent inducer of inflammation. Since this reaction can be attenuated by cyclo-oxygenase inhibitors, prostaglandins are believed to be important. The role of cytokines in the CSF has not been fully elucidated in this type of meningitis. However, the pneumococcal cell wall is a potent inducer of systemic interleukin-1 (**IL-1**), but not of tumour necrosis factor (**TNF**). The exact role of these mediators in pneumococcal meningitis remains unresolved. The main bacterial component responsible for inducing inflammation in *H. influenzae* type b is lipopolysaccharide (**LPS** or endotoxin), a potent inducer of cerebrospinal fluid TNF and IL-1 which have been shown to mediate the inflammatory response in the subarachnoid space. There is a suggestion of a dose–response effect between lipopolysaccharide and the inflammatory mediators, and so interventions that release lipopolysaccharide may exacerbate the inflammatory reaction. It has also been suggested that certain antibiotics which produce an enhanced lipopolysaccharide release may temporarily exaggerate the inflammatory reaction in a type of Jarisch–Herxheimer reaction.

The inflammatory reaction in meningitis is associated with a number of severe alterations in the normal physiology of the CNS. First, permeability of the blood–brain barrier increases. This is best measured by the increased penetration of the CSF by albumin. Also, antibiotic penetration of the CSF is greatly enhanced. Second, increased intracranial pressure results from cerebral oedema secondary to an accumulation of interstitial fluid, and communicating hydrocephalus is caused by decreased CSF re-absorption and cellular swelling secondary to cell injury. Third, a vasculitis may affect mainly the large vessels traversing the subarachnoid space. This vascular injury may not only disrupt the normal autoregulation of cerebral blood flow, but, in severe cases, the vessel may become obstructed with thrombus, causing a cerebral infarct. The major impact of increased intracranial pressure and vasculitis is decreased cerebral perfusion, causing general hypoxic brain injury.

Pathology

There is diffuse acute inflammation of the pia-arachnoid, with migration of neutrophil leucocytes and exudation of fibrin into the CSF. Pus accumulates over the surface of the brain, especially around its base and the emerging cranial nerves, and around the spinal cord. The meningeal vessels are dilatated and congested and may be surrounded by pus. Pus and fibrin are found in the ventricles and there is ventriculitis, with loss of ependymal lining and subependymal gliosis. Dilatation of the ventricular system may result from obstructive or communicating hydrocephalus. Other abnormalities include subdural effusion or empyema, septic thrombosis of the cerebral venous sinuses, subarachnoid haematomas, compression of intracranial structures as a result of intracranial hypertension, and herniation of the temporal lobes or cerebellum. Gross changes, such as pressure coning, which would provide an obvious cause of death, are rarely found. In some cases death may be attributable to related septicaemia (Fig. 1), although the familiar finding of bilateral adrenal haemorrhage (Waterhouse–Friederichsen syndrome) may well be a terminal phenomenon rather than a cause of fatal adrenal insufficiency as was once imagined. Patients with meningococcal septicaemia may develop acute pulmonary oedema. Myocarditis was a common finding in some series of patients. Histological appearances were of an acute interstitial myocarditis, occasionally with myocardial necrosis and thrombosis of small arterioles. Pericarditis and pericardial effusion were features, particularly of group C meningococcal infections. Myocarditis and Waterhouse–Friederichsen syndrome also occur, less frequently than in meningococcal septicaemia, in septicaemia caused by *H. influenzae,* pneumococcal, streptococcal, and staphylococcal infections.

Fig. 1 Nigerian patient with pneumococcal meningitis and septicaemia who developed 'urea frost' and later died of renal failure. This illustrates the importance of septicaemic complications outside the central nervous system in determining mortality. (Copyright D.A. Warrell.)

Clinical features

Acute bacterial meningitis carries a mortality in untreated patients of between 70 and 100 per cent. Delay in treatment greatly increases the risk of permanent neurological sequelae. The early diagnosis of this condition is, therefore, a formidable challenge to clinical acumen, but early clinical suspicion of meningitis may be impossible in many cases, especially in neonates and small children. When meningitis is secondary to infection elsewhere, such as pneumococcal pneumonia or *H. influenzae* otitis media, the presenting symptoms may be those of the original infection. The incubation period is only a few days. Progression is occasionally so rapid (*N. meningitidis*) that the patient becomes comatose within a few hours after the first symptoms. Early manifestations include non-specific malaise, apprehension, or irritability, followed by fever, usually without rigors, headache, myalgias, and vomiting. Convulsions occur in infants and children and meningitis must always be included in the differential diagnosis of childhood febrile convulsions. Photophobia, drowsiness, or more severe impairment of consciousness usually develop later. Headache quickly becomes more severe and is the dominant symptom. In older children and adults the symptoms most suggestive of meningitis are irritability, severe headache, and vomiting, but in the case of meningococcal infection, diarrhoea is a common non-specific symptom and the vasculitic rash is a crucial sign. An early symptom of meningococcal septicaemia is pain in the calves.

There is rarely any doubt that the child or adult with meningitis is severely ill and distressed. Meningism is best elicited by gentle passive flexion or rotation of the neck with the patient lying supine. If patients can shake their heads vigorously they are unlikely to have meningitis! To elicit Kernig's sign the lower limb is flexed at the hip. The patient with meningism will resist extension of the knee by contracting the hamstring muscles. Brudzinski's neck sign is best elicited while the patient sits up in bed with the legs stretched out. Gentle flexion of the neck will induce a compensatory flexion of the hips, knees, and sometimes the upper limbs. Later, the patient with marked meningism lies in a characteristic position with the neck and back fully extended (Fig. 2) as in tetanic opisthotonos. Local causes of neck stiffness, such as local sepsis (for example, in the nuchal muscles or cervical lymph nodes), cervical spondylitis (particularly common in the elderly), temporomandibular arthritis, dental problems, and pharyngeal lesions, should be considered. Meningism is not uncommon in patients without meningitis who have other febrile conditions such as pyelonephritis. Meningism may be reduced or absent in patients who are immunosuppressed. The optic fundi should be examined as a prelude to lumbar puncture. Papilloedema is suggestive of cerebral oedema or an intracranial space-occupying lesion, such as a cerebral abscess or a subdural or epidural collection of pus. The absence of papilloedema does not, however, exclude cerebral oedema and, if in doubt, and cerebral imaging is available, that investigation must precede a lumbar puncture. Retinal vein pulsation excludes intracranial hypertension. Hypertensive retinopathy will suggest hypertensive encephalopathy, and subhyaloid haemorrhages a subarachnoid haemorrhage. Patients with meningococcal meningitis associated with meningococcal antigenaemia have a petechial rash (Plate 1) which may appear first on the shins or volar surface of the forearms. Petechiae may be visible on the bulbar and tarsal conjunctivas (Plate 2) and palate. An identical rash is occasionally seen in patients with echovirus type 9, leptospirosis, *S. aureus, S. pneumoniae, S. suis* (Plate 3) *H. influenzae, Salmonella typhi,* and other infections, especially in those associated with infective endocarditis. The brownish or reddish geometrical, vasculitic rash of fulminant meningococcaemia is unmistakable (Plate 4 and Plate 5) and, characteristically, the toes and fingers become necrotic (Fig. 3). There is associated profound hypotension, shock with peripheral cyanosis, and spontaneous systemic bleeding. Herpes labialis is commonly seen in all forms of bacterial meningitis, because a fever is the rule and recurrences are provoked by fever (see Fig. 2 and Fig. 4). Physical examination must exclude otitis media, sinusitis, mastoiditis, and nasopharyngeal and other possible sites of sepsis. In patients with recurrent bacterial meningitis, a search should be made for a congenital dermal sinus, which is usually in the midline between the head and coccyx and is often marked by a tuft of long hairs. Watery discharge from the nose or ears should be collected and tested for glucose; the possibility of a basal skull fracture with cerebrospinal fluid leak should be excluded.

Fig. 2 Nigerian girl in coma with severe meningococcal meningoencephalitis. Note head retraction, dysconjugate gaze, and herpes labialis. (Copyright D.A. Warrell.)

Fig. 3 Gangrene of the fingers in a man with fulminant meningococcaemia. Eventually, three of his limbs had to be amputated. (Copyright D.A. Warrell.)

Fig. 4 Nigerian man recovering from meningococcal meningitis. Note right VIth nerve lesion and herpes labialis. (Copyright D.A. Warrell.)

Cranial nerve lesions may become evident as the disease progresses. The commonest are VI (Fig. 4), III, VII, VIII, and II. Patients who become deeply comatose (meningoencephalomyelitis) may lose all signs of meningism and develop focal neurological signs, focal epileptiform convulsions, dysconjugate gaze, upper motor neurone signs, and involuntary movements. Vascular lesions may produce hemiparesis or quadriparesis, speech disorders, and visual field defects. Bilateral sensorineural deafness develops early, 2 to 9 days after the start of symptoms, in the majority of patients with *S. suis* type 2 meningitis. Initially associated with tinnitus and vertigo, this may progress to complete deafness within 24 h. Bacteria probably invade the cochlea via the cochlear aqueduct from the subarachnoid space to produce suppurative labyrinthitis and acute deafness. Associated clinical features of *S. suis* meningitis include third nerve palsy, septic arthritis (Plate 6), and purpuric skin lesions (Plate 3).

Papilloedema, with or without other symptoms and signs of intracranial hypertension (vomiting, postural headache, coma, high blood pressure, bradycardia, etc.) and localizing neurological signs, suggests a subdural effusion or empyema. This is particularly common in children under 2 years old with *H. influenzae* meningitis.

Neonates and infants

Meningitis is particularly difficult to diagnose in this age group. Infants may become irritable or lethargic, stop feeding, and are found to have a bulging fontanelle, separation of the cranial sutures, meningism, and opisthotonos, and they may develop convulsions. These findings are uncommon in neonates, who sometimes present with respiratory distress, diarrhoea, or jaundice.

Post-traumatic meningitis

This is often indistinguishable clinically from spontaneous meningitis. However, in obtunded or unconscious patients who have suffered a recent head injury, few clinical signs may be present. A fever and a deterioration in the level of consciousness or loss of vital functions may be the only signs of meningitis. Finding a CSF leak adds support to the possibility of meningitis in such patients, but this is undetectable in many cases.

Infections of CSF shunts

Patients may present with clinical features typical of spontaneous meningitis, especially if virulent organisms are involved. The more usual presentation is insidious, with features of shunt blockage such as headache, vomiting, fever, and a decreasing level of consciousness. Fever is a helpful sign, but is not a constant feature and may be present in as few as 20 per cent of cases. Shunts can be infected without causing meningitis, in which event the features of the infection will be determined by where the shunt drains. Infection of shunts draining into the venous system produces a disease similar to chronic right-sided infective endocarditis together with glomerulonephritis (shunt nephritis), while infection of shunts draining into the peritoneal cavity produces peritonitis.

Diagnosis

Examination of cerebrospinal fluid (see Chapter 24.7)

Examination of CSF is crucial for the diagnosis of meningitis. The main risk of lumbar puncture, fatal pressure coning, is greatest in patients with space-occupying lesions or post-traumatic cerebral oedema. Fortunately, it is a rare complication in cases of spontaneous meningitis, but caution should be exercised when contemplating spinal puncture. Patients with clinical features suggesting raised intracranial pressure (for example, papilloedema, loss of retinal vein pulsation, focal neurology, bradycardia, and coma) should be examined by CT or MRI of the head, if available, to exclude a space-occupying lesion or severe cerebral oedema. In meningitis, the CSF opening pressure is usually raised (>200 mm of CSF), and occasionally it is markedly raised (>500 mm of CSF), suggesting the potential danger of pressure coning. Other contraindications to lumbar puncture include local skin sepsis at the site of puncture and any clinical suspicion of spinal cord compression.

Frank turbidity of the first drop of CSF emerging from the lumbar puncture needle instantly suggests the diagnosis of bacterial (pyogenic) meningitis. Microscopic examination of CSF for white cells, red cells, and organisms; the measurement of glucose and protein; and culture are important investigations in a case of possible meningitis. A raised CSF white blood cell (**WBC**) count is present in the majority of patients with bacterial meningitis but, rarely, the count may be normal (<6 WBC/µl, all lymphocytes) but the CSF may still appear turbid becuase of the vast numbers of bacteria. A majority of cases (>90 per cent) present with a count of more than 100 WBC/µl. The white cell differential count is helpful. Most cases (>80 per cent) have over 80 per cent of neutrophils. A predominance of lymphocytes is occasionally found and is reported especially in early bacterial meningitis, in association with *L. monocytogenes* infection and in partially treated patients. Red blood cells and xanthochromia are sometimes present. A wide range of non-bacterial infections and non-infectious conditions lead to pleocytosis of the CSF. In this respect, early viral meningitis, parameningeal septic foci, meningeal or cerebral tumours, cerebral infarction, chemical meningitis, aseptic meningitis complicating immunoglobulin replacement, cerebral vasculitis, and demyelination may be indistinguishable from bacterial meningitis.

Detection of bacteria in CSF confirms the diagnosis of bacterial meningitis. Gram-staining of the CSF will reveal organisms in 50 to 80 per cent of cases. The Gram-stain appearance of bacteria may be characteristic of a particular species, but caution must be exercised for up to 10 per cent of smears are misinterpreted. It is prudent, therefore, to administer appropriate empirical therapy initially and to change to specific therapy only when the infecting organism has been isolated and identified. Culture of organisms has a sensitivity of approximately 80 per cent in untreated cases, and is aided by the culture of good volumes of CSF and minimizing the delay between the lumbar puncture and setting up of the culture. Organisms are recovered much less often from partially treated cases. Isolation of an organism is not only helpful in establishing the diagnosis, but allows the identification and susceptibility testing of the aetiological agent. The culture result can also be used to

decide on the need for antibiotic prophylaxis, contact tracing, and other public health control measures.

Measurement of CSF glucose is helpful in making the diagnosis of bacterial meningitis. A glucose concentration below 40 mg per cent is considered low and is found in 50 to 70 per cent of cases. In many cases, glucose may even be undetectable. As the CSF glucose concentration is a function of the serum glucose, a more reliable measure of hypoglycorrhachia (low CSF glucose) is the ratio of serum to CSF glucose concentration. A ratio below 0.31 for glucose concentrations of simultaneously obtained serum and CSF indicates hypoglycorrhachia. A few other conditions may lead to a reduced CSF glucose level. They are: tuberculous, syphilitic, parasitic, fungal, or mumps meningitis; herpes simplex encephalitis; carcinomatous meningitis; meningeal sarcoidosis; post-subarachnoid haemorrhage; severe systemic hypoglycaemia; and rare forms of central nervous system vasculitis.

An elevated CSF protein concentration is a common, but non-specific, finding. Similarly, measurement of CSF lactate is sensitive, but non-specific for meningitis. A range of rapid bacterial antigen tests may be helpful in detecting the presence of bacterial capsular polysaccharide antigens of pneumococci, meningococci, *H. influenzae*, and group B streptococci. These tests may reach a sensitivity and specificity of 90 per cent or greater for detecting specific causes of bacterial meningitis. However, in our experience these tests seldom add to the diagnostic yield of a good Gram stain performed on an adequate volume of CSF. PCR is also used.

The interpretation of the CSF test results depends on the clinical presentation and course of the disease. Acute viral meningitis is the most common differential diagnosis of bacterial meningitis. The CSF in viral meningitis typically contains a preponderance of lymphocytes, less than 1000 WBC/µl, and a normal glucose level. Fortunately, most patients with acute bacterial meningitis will exhibit a combination of clinical and CSF findings sufficiently characteristic to allow a reliable diagnosis. However, in some cases, it may be impossible to distinguish between aseptic meningitis, chronic meningitis, partially treated bacterial meningitis, and early acute bacterial meningitis. In these circumstances it may be necessary to initiate empirical antibiotic treatment. Depending on the clinical course of the patient, it may also be necessary to repeat the spinal tap and monitor the changes in glucose, lactate, and cell count in the CSF. These dynamic changes are particularly helpful now that antimicrobial resistance is an increasing problem and delayed response to treatment might occur requiring an early change of therapy. It is important to appreciate that CSF abnormalities secondary to bacterial meningitis, such as neutrophil pleocytosis, and raised protein levels, may persist for up to a week or longer, although the glucose and lactate levels should show signs of improvement within 48 to 72 h in patients receiving appropriate antibiotics. Since the pleocytosis and hypoglycorrhachia typical of acute bacterial meningitis may persist for a few days after treatment is started, it can be possible to diagnose partially treated meningitis despite negative Gram stain and culture. The difficulty arises when the neutrophil response switches to a lymphocytic one and the glucose starts to normalize after starting treatment for pyogenic meningitis. At this point it can be extremely difficult to distinguish partially treated bacterial meningitis from early tuberculous meningitis (**TBM**)). In these situations the importance of the clinical history (longer than 7 days—TBM), physical signs (lower cranial nerve signs—TBM), peripheral blood white cell count (elevated—pyogenic meningitis), and meticulous Gram and Ziehl–Neelsen stain of the CSF are crucial.

Other tests

Blood cultures should be obtained for all patients with suspected meningitis, as the aetiological agent may be grown. In patients with associated rash, a Gram stain and culture from fluid aspirated from the skin lesions may secure a microbiological diagnosis. A small amount of sterile saline should be injected under the lesion and immediately aspirated for staining and culturing. Radiological imaging of the central nervous system may be helpful. CT scanning can indicate whether or not a shunt is obstructed. Skull fractures or parameningeal septic foci (for example, sinusitis, spinal epidural abscess, or brain abscess) may also be detected by scanning.

Differential diagnosis

Meningeal irritation is seen in many acute febrile conditions, especially in children. Local infections of the nasopharynx, cervical lymph nodes, muscles, and spine may produce convincing neck stiffness. Tetanus may be easily confused with meningitis if the persisting rigidity and recurrent spasms go unnoticed. In all these conditions the CSF will be normal. Subarachnoid haemorrhage can present with sudden headache, neck stiffness, and deteriorating consciousness, and a less dramatic progression of symptoms is seen in patients with some intracranial tumours. Tuberculous and cryptococcal and other fungal meningitides usually develop more slowly than acute bacterial meningitis. They may be distinguished by examining CSF. Cryptococci and free-living amoebae may be mistaken for lymphocytes in the CSF unless an India-ink preparation is examined to reveal the capsule of cryptococcus and the characteristic movements of amoebae. Aseptic meningitis comprises a large number of conditions, many of them caused by viruses, in which there are clinical signs of meningism and the CSF is found to be abnormal. This group includes partially treated bacterial meningitis and the chemical meningitides, resulting from the introduction of irritants into the subarachnoid space (contrast media, antimicrobial agents, and contaminants of lumbar puncture and spinal anaesthesia). The CSF glucose concentration may be very low. Discharge of a tuberculoma may produce a sterile tuberculin reaction, and the discharge of the contents of a craniopharyngioma or epidermoid cyst into the CSF can also cause chemical meningitis. Lead encephalopathy may present with meningism, lymphocyte pleocytosis, and an increase in CSF protein.

Recurrent purulent meningitis

This usually suggests a congenital or traumatic defect providing access to the subarachnoid space, such as congenital occult spina bifida or fracture of the base of the skull. A CSF leak may be apparent in about 50 per cent of the cases with post-traumatic recurrent meningitis. The head trauma may have occurred many years earlier and a connection with the subarachnoid space may be clinically inapparent. *S. pneumoniae* and *H. influenzae* are the predominant aetiological agents for community-acquired cases. Gram-negative aerobic bacilli or *S. aureus* are the main causes in nosocomial cases.

Rarely, recurrent meningitis may arise from episodes of recurrent sepsis of a parameningeal focus (for example, sinusitis or mastoiditis) or from a complement deficiency. Deficiency in a number of the components of the complement pathway has been detected in patients with recurrent meningitis. *N. meningitidis* meningitis caused consecutively by different serogroups is the usual presentation in these cases.

Mollaret's meningitis (benign recurrent aseptic meningitis or benign recurrent lymphocytic meningitis) is mentioned here because it is an important differential diagnosis of recurrent bacterial meningitis. It is a sporadic condition presenting between the ages of 5 and 60 years. The symptoms are typical of acute meningitis—malaise, fever, vomiting, neck stiffness, convulsions, and coma. There is complete spontaneous recovery, usually within a few days, and symptom-free intervals lasting from a few days to years. About half the patients develop other neurological disturbances including hallucinations, diplopia, cranial nerve lesions, and signs of an upper motor neurone lesion. Pleocytosis is usually less than 3000/µl, with a predominance of lymphocytes, monocytes, and large endothelial ('Mollaret's') cells, but occasionally neutrophils are in the majority. The CSF protein level is mildly increased, with increased gammaglobulin. CSF glucose concentration may be decreased. Recently, evidence of herpes simplex type 1 and 2 has been obtained, using polymerase chain reaction (**PCR**) technology. Other causes of recurrent meningitis include Behçet's syndrome, Vogt–Koyanagi–Harada syndrome, sarcoidosis and systemic lupus erythematosus, and undiagnosed viral meningitis (for example, that due to encephalomyocarditis virus).

Management

Bacterial meningitis progresses rapidly and has a high mortality. Antimicrobial treatment must therefore be started as soon as possible after the diagnosis is suspected clinically. This is vitally important in meningococcal meningitis/septicaemia. If this condition is suspected by the practitioner (and in Britain the first doctor the patient meets is usually the general practitioner), the patient should **without delay** be given benzylpenicillin (intravenous or intramuscular (**IV**/**IM**)), cefotaxime (IV/IM), or ceftriaxone (IV). Although it is desirable that antibiotics are given following a blood culture their administration must not be delayed while waiting for the culture to be taken (Table 2). These drugs should be carried by general practitioners in their emergency bags. Antigen may still be detectable in the CSF later in hospital. Patients with no papilloedema or lateralizing neurological signs to suggest a space-occupying lesion, and with no other contraindications (see above), should undergo an immediate lumbar puncture. Again antimicrobial treatment should be started as soon as bacterial meningitis is suspected clinically and, if necessary, before the lumbar puncture is performed.

Antimicrobial treatment

Successful antimicrobial treatment of meningitis (Table 3 and Table 4) depends on the agents crossing the blood–brain barrier and achieving a concentration of more than tenfold the minimum bactericidal concentration in the CSF (a level which predictably sterilizes the subarachnoid space). Before the aetiological agent has been isolated, empirical treatment that will be effective against the likely bacterial causes should be started immediately the diagnosis is made. Once the pathogen has been isolated, specific treatment based on the susceptibility of the isolate can be substituted for the empirical regimen.

Empirical regimens (Table 3) depend on the clinical circumstances of the case and the local pattern of aetiological agents and their antibiotic susceptibility patterns. Neonatal meningitis is largely caused by group B streptococci, *E. coli*, and *L. monocytogenes*. Initial treatment, therefore, should consist of an aminoglycoside and

penicillin or ampicillin; alternatively, a third-generation cephalosporin, preferably cefotaxime or ceftriaxone, **and** penicillin or ampicillin (to cover *L. monocytogenes*) should be used.

In the community, children are at risk of meningitis caused by *N. meningitidis, S. pneumoniae*, and, rarely in Hib-vaccinated children, *H. influenzae.* Antimicrobial resistance has emerged among the three major bacterial pathogens causing meningitis. Recently, chloramphenicol resistance in the meningococcus has been described, and although intermediate penicillin resistance is common in some countries, the clinical importance of penicillin resistance in the meningococcus has yet to be established. b-Lactamase-producing *H. influenzae* are relatively common, and chloramphenicol resistance is emerging. Third-generation cephalosporins are required to treat meningitis caused by these resistant strains. Pneumococci resistant to penicillin and to chloramphenicol are widespread, and resistance to third-generation cephalosporins is found in many parts of the world. Correct management of these strains includes the addition of vancomycin or rifampicin to therapy with third-generation cephalosporins. It is crucial to have up-to-date information on the resistance patterns of these common pathogens within communities to guide antibiotic prescribing. Spontaneous meningitis in adults is usually caused by *S. pneumoniae* or *N. meningitidis*; however, in older patients (>50 years of age) and chronically immunosuppressed patients, there is an increased risk of *L. monocytogenes* and infection caused by Enterobacteriaceae (for example, *E. coli*). If infection with *L. monocytogenes* is possible, penicillin or ampicillin should be used. In all age groups, patients presenting with features of meningococcaemia should receive parenteral penicillin (for example, 2.4 g or 4 million units) immediately. *S. suis* remains sensitive to the b-lactams and should be treated with penicillin, cefotaxime, or ceftriaxone.

Community-acquired, post-traumatic meningitis is caused mainly by *S. pneumoniae, H. influenzae,* and a wide range of other bacterial species. Cefotaxime (2 g intravenously, every 6 h), ceftriaxone (2 g intravenously, every 12 h), or chloramphenicol (1 g, every 6 h) plus ampicillin (2–3 g intravenously, every 6 h) should be given. Nosocomial post-traumatic meningitis is mainly caused by multiresistant hospital-acquired organisms such as *K pneumoniae, E. coli, Pseudomonas aeruginosa,* and *S. aureus.* Depending on the pattern of susceptibility in a given hospital unit, ceftazidime (2 g intravenously, every 8 h), cefotaxime, ceftriaxone, or meropenem should be chosen. If *P. aeruginosa* infection seems likely, ceftazidime or meropenem is the preferred antibiotic.

Device- and shunt-associated meningitis is caused by a wide range of organisms, including methicillin-resistant staphylococci (mostly coagulase-negative staphylococci), multiresistant aerobic bacilli, and *Candida* sp. Cases with shunts and an insidious onset are probably caused by organisms of low pathogenicity, and empirical therapy is a less urgent requirement. For postoperative meningitis the first-line empirical therapy should be cefotaxime (3 g intravenously, every 8 h), or ceftriaxone (2 g intravenously, every 12 h), or meropenem (2 g intravenously, every 8 h). If the patient has received broad-spectrum antibiotics recently or if *P. aeruginosa* is suspected, ceftazidime (2 g intravenously every 8 h) or meropenem should be given. Meropenem should be used if an extended-spectrum, b-lactamase organism is suspected, and flucloxacillin or vancomycin if *S. aureus* is likely. The infected shunt or drain will almost certainly have to be removed urgently.

Once the aetiological agent has been isolated and its susceptibilities determined, the empirical treatment should be changed, if necessary, to an agent or agents specific for the isolate (Table 4). The optimal duration of treatment has not been determined by rigorous scientific investigation; however, treatment regimens that are probably substantially in excess of the minimum necessary to achieve cure have been arrived at through wide clinical experience.

Treatment of brain abscess (see Chapter 24.14.3)

Brain abscesses may arise as a result of direct spread from contiguous anatomical structures, following injury or local infection or metastasis from a distant source. Whenever an abscess is suspected a detailed search for the source of the infection is important. The middle ear and mastoid cavity and frontal, paranasal, ethmoidal, and sphenoidal sinuses are common sites for the primary focus. Metastatic abscesses can spread from foci in the heart (endocarditis), the lungs (bronchiectasis), dental abscesses, the pelvis, and gastrointestinal tract. Once diagnosed, surgical drainage remains the treatment of choice for almost all abscesses. Empirical therapy can be guided by consideration of the likely primary focus of the infection, but it should include penicillin or a third-generation cephalosporin (cefotaxime of ceftriaxone), and metronidazole. For abscesses complicating trauma, flucloxacillin or vancomycin should be added. It is unclear if there are any benefits to be gained by instilling antibiotics into the abscess cavity during drainage. Treatment should continue for a minimum of 6 weeks.

General management and treatment of complications

General treatment

In the conscious patient, the severe headache may need treatment with opiates and the associated pyrexia may require treatment with antipyretics or cooling with tepid sponging or fanning. Many patients are unconscious and should be managed accordingly. Their airway should be maintained and they may need intubation to protect the airway and maintain ventilation. A urethral catheter should be inserted.

The associated septicaemia in patients with meningitis may result in septicaemic shock. These patients require careful monitoring and treatment of shock. A combination of fluid administration to expand the intravascular volume, pressors, and inotropic support is needed. Multiple end-organ failure may develop, requiring intensive medical support, including ventilation and renal dialysis.

Treatment of complications

Raised intracranial pressure is a serious complication of meningitis. Clinically, altered consciousness, poorly reactive unequal or dilated pupils, cranial nerve palsies, bradycardia, and hypertension suggest its onset. Various measures can be used to lower a raised intracranial pressure, including elevation of the head of the bed to 30°, administration of mannitol, and endotracheal intubation and mechanical hyperventilation. These measures may be used to maintain an adequate perfusion pressure monitored by continuous intra-arterial and intracranial pressure monitoring. The effect of these measures on the outcome of patients with raised intracranial pressure have not been evaluated systematically. The role of corticosteroids in reducing intracranial pressure remains unresolved and no consistent approach to their use is accepted.

The use of corticosteroids aimed at reducing the sequelae of meningitis has received support from a number of studies. The most pronounced effect has been the reduction of deafness in children infected with *H. influenzae* and treated with cefuroxime (a less effective cephalosporin for meningitis than cefotaxime or ceftriaxone). In three recent studies of bacterial meningitis (mainly *H. influenzae*) in children treated with ceftriaxone, dexamethasone (0.4 mg/kg intravenously, every 12 h for 2 days starting 10 min before the first dose of antibiotic) reduced the incidence of neurological and audiological sequelae. Studies in Cairo suggested that dexamethasone significantly reduced the mortality and incidence of permanent sequelae in adult patients with pneumococcal meningitis but in Malawi, this drug did not prove to be an effective ancillary treatment in children with acute bacterial meningitis.

Seizures occur in as many as 40 per cent of patients with meningitis and subclinical fitting should be considered in patients with persisting coma. A prophylactic anticonvulsant might be justified in pneumococcal meningitis. Rapid control of seizures can be achieved by administering intravenous diazepam or lorazepam, but respiration may be depressed. Phenytoin should be used for the longer term control of seizure activity. Severe cases that fail treatment with these antiepileptic agents should undergo endotracheal intubation and receive high-dose phenobarbital.

Many patients with meningitis develop hyponatraemia as a result of the syndrome of inappropriate antidiuretic hormone secretion (**SIADH**). Such cases may require fluid restriction. Complications such as cerebral venous sinus thrombosis and cavernous sinus thrombosis may occur, but may be difficult to detect. Anticoagulants or fibrinolytics may be used in such patients with the hope of improving the outcome (Chapter 24.13.7).

Prognosis and sequelae

In Europe and North America the overall mortality of patients with meningitis caused by *N. meningitidis* is about 7 to 14 per cent; for *H. influenzae,* 3 to 10 per cent; *Strep. pneumoniae,* 15 to 60 per cent; and for group B streptococci and *L. monocytogenes* meningitis, above 20 per cent. The mortality is much higher in the very young and old, and in patients with debilitating illnesses. A study in Zaria, Nigeria, demonstrated that the mortality of pneumococcal meningitis was 32 per cent in patients who were fully conscious on admission, 40 per cent in those who were confused, 54 per cent in semiconscious patients, and 94 per cent in those who were comatose. In Vietnam, in a prospective study of 250 cases of adult bacterial meningitis, the overall mortality rate was 13 per cent.

Permanent neurological sequelae include mental retardation, deafness and other cranial nerve deficits, and hydrocephalus. The reported incidence of sensorineural deafness after meningitis ranges from 5 to 40 per cent. A large proportion of patients recover within a few months. *N. meningitidis* and *H. influenzae* are the main causes of this complication. Permanent deafness occurs in more than 50 per cent of patients with *S. suis* meningitis. It may be bilateral, complete, and associated with

vestibular involvement. *H. influenzae* used to be the major cause of acquired mental retardation in the United States. This complication was found in 30 to 50 per cent of children who had suffered from *H. influenzae* meningitis.

Prevention

Vaccination

Vaccines to the three major pathogens causing community-acquired meningitis are available. *H. influenzae* type b capsular conjugate vaccines are in wide use and are dramatically reducing the incidence of *H. influenzae* type b meningitis. In most countries where the *H. influenzae* type b meningitis vaccine has been introduced, Hib invasive disease has been essentially eliminated. A 7-valent pneumococcal conjugate vaccine is now widely available in the developed world, the efficacy of which is supported by a Californian Kaiser Permenete study in children. Studies in adults are being carried out in the developing world. A conjugate, type C meningococcal vaccine is available and has been used in the United Kingdom immunization schedule, producing a dramatic decline in group C meningococcal invasive disease including meningitis.

The purified capsular polysaccharide vaccines directed at *N. meningitidis* (serogroups A, C, Y, and W135) and *Strep. pneumoniae* (23-valent) are less effective than conjugate vaccines, especially in children under 2 years of age. No effective vaccine exists for *N. meningitidis* serogroup B. Vaccination of groups of adults at high risk, such as close contacts, military recruits, and migrant miners in Africa, has been highly effective in preventing meningococcal meningitis caused by serogroups A and C. The trials designed to assess the efficacy of population-based vaccination against *N. meningitidis* (serogroups A, C, Y, and W135) and *Strep. pneumoniae* are awaited with interest. The annual Muslim pilgrimage to Mecca (The Haj) is frequently associated with meningitis in returning travellers. It is advisable for people travelling to Mecca to be vaccinated with the polyvalent vaccine against *N. meningitidis* groups A, C, W135, and Y.

Since 1905, major epidemics of meningococcal meningitis have occurred in sub-Saharan Africa every few years, culminating in a massive epidemic in 1996 in which nearly 200 000 cases were reported. For epidemic meningitis control in sub-Saharan Africa, the World Health Organization recommends a strategy of emergency vaccination with meningococcal A+C polysaccharide vaccine when epidemic thresholds are exceeded. A recently derived model of how to respond to such epidemics has concluded that, given the relatively poor routine-vaccination coverage in this region, current strategies of vaccination campaigns that achieve higher coverage would generally be more effective and less costly than model routine-scheduled programmes, assuming that campaigns can be rapidly implemented. Until a better vaccine is available, countries in this region should aim to speed up the response times to outbreaks, perhaps through improved surveillance, and to bolster existing vaccination infrastructures.

Vaccination should be considered in all patients with recurrent meningitis, traumatic head injury, and in splenectomized patients.

Chemoprophylaxis

The attack rate of meningitis is higher in the immediate contacts of an index case of meningococcal (up to 1000-fold) or *H. influenzae* type b meningitis (500-fold only in children under 4 years of age) than in the population at large. The administration of rifampicin or ciprofloxacin eliminates the carrier state and is assumed to eliminate the risk of secondary cases of meningitis. A major preventive effect of sulfadiazine chemoprophylaxis has been shown in large outbreaks of meningococcal meningitis among military recruits; however, meningococci are now usually resistant to sulphonamides. Rifampicin and ciprofloxacin are assumed to produce a similar effect. Close adult contacts of meningococcal disease are given either rifampicin (300 mg, every 12 h for 2 days) or a single oral dose of ciprofloxacin (750 mg) (ciprofloxacin is still to be avoided in children, but is now accepted for prophylaxis). Contacts of *H. influenzae* type b disease (including adults) are given rifampicin for 4 days, and Hib vaccine is given to unvaccinated children under 4 years of age. However, in those countries where Hib vaccination has been implemented, the need for chemoprophylaxis is under review. Doctors, nurses, and other healthcare workers need not be given chemoprophylaxis unless they have given mouth-to-mouth resuscitation.

Although antibiotics are administered prophylactically to many cases of skull fracture with CSF leak, there is no evidence of benefit. Surgical closure of the leak is the only effective means of preventing meningitis in such cases. However, since many acute leaks heal spontaneously, this is necessary only in cases with large defects or in those with recurrent meningitis.

The prevention of device-associated meningitis relies on rigorous infection control. Ventricular and lumbar drains should be removed as soon as possible. Shunt insertion should be performed while adhering to a strict aseptic technique, and surgical antibiotic prophylaxis may also help to reduce shunt infection.

Tuberculous meningitis

Epidemiology

Tuberculous meningitis (**TBM**) has been a major problem and cause of death in developing countries. Human *Mycobacterium tuberculosis* is responsible for most cases. In Western countries, its incidence has fallen in parallel with tuberculosis as a whole. For example, in the late 1940s there were 2000 cases/year in England and Wales, accounting for 10 to 20 per cent of cases of bacterial meningitis; but by the early 1970s this had fallen to less than 4 per cent.

Most cases of TBM are in young children, but primary infection can be acquired at a later age and, in recent years, a larger proportion of patients with this condition have been adults. The disease is uncommon, but severe, in pregnant women.

There has been an increase in the incidence of tuberculosis in many parts of the world related to the epidemic of HIV infection. HIV-infected patients with tuberculosis are at increased risk of meningeal involvement. In some areas of endemic tuberculosis, TBM is an important complication in patients with HIV infection.

Pathogenesis

Small caseous microtubercles develop in the brain, meninges, or, less commonly, in the bones of the skull and vertebrae close to the meninges. This infection seeds through the bloodstream from the primary lesion or a site of chronic infection. Many patients develop miliary tuberculosis at the stage of haematogenous spread. Infection of the meninges results from rupture of a microtubercle with discharge of tuberculoprotein and mycobacteria into the subarachnoid space. This event will be marked by an episode of fever and meningeal irritation caused by the intrathecal tuberculin reaction. Subacute inflammation, especially of the basal meninges, then develops, producing cranial nerve lesions, cerebral arteritis causing infarction, impairment of CSF absorption, or obstruction to the CSF circulation causing hydrocephalus, and, in the spinal cord, spinal arachnoiditis, producing multiple radiculopathy or myelopathy.

Pathology

Meningeal miliary tubercles may be found on the brain surfaces of most victims of miliary tuberculosis. They are usually few in number and occur in the region of the sylvian fissure, while there may be larger caseous plaques deeper in the sulci. The brains of patients dying of TBM are usually oedematous. A mass of thick, greyish exudate encases the base of the brain, filling the basal cisterns. Within the ventricular system there is ependymitis with a similar exudate choking the choroid plexus. The exudate consists of lymphocytes, plasma cells, giant cells, and foci of caseation, but mycobacteria are usually scanty. At the base of the brain the cranial nerves and the internal carotid artery and its branches are trapped and damaged by the exudate. Arteries are obliterated by an endarteritis, with resulting ischaemia and infarction of superficial areas of the brain, internal capsule, basal ganglia, and brainstem. There is congestion and phlebitis of the meningeal veins. Both the inflammatory exudate and the arteritis are probably a delayed hypersensitivity reaction to the tuberculoprotein. Some degree of hydrocephalus develops in most cases. Usually the hydrocephalus is communicating in type and is caused by obliteration of the basal cisterns. Less commonly, blockage of the foramina of Luschka and Magendie in the fourth ventricle, or the aqueduct, causes obstructive hydrocephalus. The exudate may extend into the spinal cord, enveloping the nerve roots and spinal cord and producing spinal arachnoiditis and spinal block. Tuberculomas, single or multiple and varying in size, may be found in virtually any part of the brain (Plate 7).

Clinical features

Symptoms of meningitis are preceded by 2 to 8 weeks of non-specific prodromal symptoms which are unlikely to raise the suspicion of TBM. This phase of vague

malaise, irritability, insomnia, lethargy, anorexia, headache, abdominal pain, vomiting, and behavioural changes may develop after a head injury, surgical operation, or common childhood infection such as measles, influenza, or otitis media, suggesting that meningeal infection may be precipitated by these conditions. By the time patients have developed obvious symptoms and signs of meningitis, the disease is well advanced. They usually have low-grade fever, but severe pyrexia can occur. Half of the patients will have symptoms and signs of tuberculosis in the lungs or elsewhere. Rarely, TBM presents dramatically as an acute encephalopathy with severe headache, neck stiffness, vomiting, and seizures, mimicking acute bacterial meningitis, viral encephalitis, and subarachnoid haemorrhage.

During the second stage there is evidence of meningeal irritation (headache, vomiting, neck stiffness), cranial nerve damage, evolving hydrocephalus, and cerebral endarteritis. Infants are irritable with opisthotonos and a tense fontanelle. Cranial nerve lesions, seen in 25 per cent of cases, involve one or more of the following: II, III, IV, VI, VII, and VIII. The pupils may be dilated, unequal, and unresponsive, and many patients have a VIth nerve palsy (Fig. 5). Fundal examination reveals papilloedema in 40 per cent of cases and sometimes evidence of optic atrophy. Choroid tubercles are occasionally seen (Plate 8). Bilateral visual failure with optic atrophy is a feature of arachnoiditis in the cistern of the optic chiasm. Raised intracranial pressure, a common and serious complication of TBM, results from obstruction of the CSF circulation in the basal cisterns by the exudates or, less commonly, obstruction of the outlets of the fourth ventricle. Increasing headache, vomiting, and impairment of conscious level are mainly the result of intracranial hypertension rather than the meningeal irritation. Hydrocephalus, which often accompanies the increased intracranial pressure, is common and should be suspected in all patients with TBM. They often have severe headache, ocular palsy, pyramidal signs in the lower limbs, and incontinence of urine. If left untreated, the patients become stuporose or comatose and develop signs of brainstem damage, such as decerebrate rigidity, irregular breathing, and impairment of brainstem reflexes. About 20 per cent of the patients develop focal neurological signs such as hemiparesis, hemianaesthesia, aphasia, and hemianopia. These are the results of cerebral infarct, caused by the arteritis. Convulsive disorders are common in children but rare in adults. About 10 per cent of the patients develop symptoms and signs of spinal arachnoiditis (Fig. 6), which vary from radicular pain, radicular weakness of the lower limbs, and urinary retention, to paraplegia with sensory loss on the trunk. SIADH is common in patients with severe TBM. It produces impairment of consciousness, but decerebrate rigidity or other signs of brainstem damage are not found. Other uncommon neurological abnormalities include internuclear ophthalmoplegia, hemichorea, and hypothalamic disorders leading to loss of control of blood pressure and body temperature and diabetes insipidus.

Fig. 5 Bilateral VIth nerve palsies in a Thai girl with TBM. The cervical node biopsy was positive for acid-fast bacilli. (Copyright the late Prida Phuapradit.)

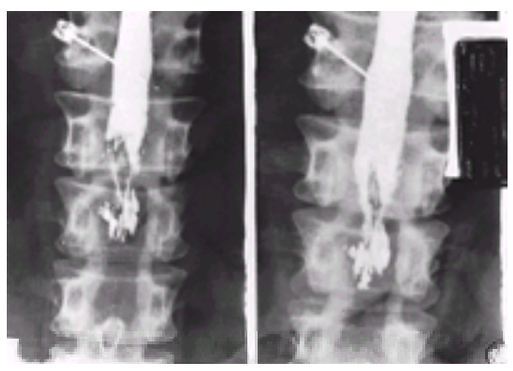

Fig. 6 Spinal arachnoiditis revealed by myelography in a Thai patient with TBM. (Copyright the late Prida Phuapradit.)

HIV-infected patients with tuberculosis are at a higher risk of developing the meningitis than non-HIV-infected patients. In HIV-infected patients, peripheral, intrathoracic, and intra-abdominal lymphadenopathy is more common, but the clinical manifestations of TBM do not seem to be modified by the HIV infection.

Diagnosis

The blood count is usually normal, but marked leucocytosis of more than 20 000/µl ('leukaemoid picture') may occasionally be found in TBM and miliary tuberculosis. Examination of CSF is crucial. Lumbar puncture does not seem to pose the same dangers as in acute bacterial meningitis and, apparently, lumbar puncture has been repeated safely as a treatment of raised intracranial pressure in patients with TBM even when there is papilloedema. The CSF opening pressure is raised in the majority of patients, but it may be low or fall in those developing block from spinal arachnoiditis. The CSF is clear or slightly turbid, and may form a spider's web clot on standing. The mechanism of the cobweb formation is unclear, but it does not appear to be caused by the high protein concentration *per se*. In patients with spinal block the fluid may be xanthochromic with a very high protein concentration and may quickly form a jelly (Froin's syndrome). Total cell counts range between 10 and 1000/µl. Both lymphocytes and neutrophils are present and the latter can be as high as 70 per cent of the total. In 90 per cent of the patients the white cell count is less than 500/µl. Rarely, the cell count exceeds 1000/µl. The CSF glucose concentration of the initial CSF sample is low in about 90 per cent of the patients. Although non-specific, this finding is of great practical use, because it is a simple test that differentiates TBM presumptively from most cases of viral meningitis. Indeed, TBM must be strongly suspected in any patient who presents with lymphocytic meningitis and a low CSF glucose concentration. The CSF protein concentration is usually raised, and ranges from normal to 5 g/l. Levels of more than 5 g/l suggests a spinal block. Success in detecting tubercle bacilli in CSF by Ziehl–Neelson staining can, in some hands, be increased from the usual average of 10 to 20 per cent by centrifuging a large volume of CSF (10–20 ml) and carefully examining the sediment under a microscope. Repeated lumbar punctures increase the chance of finding tubercle bacilli and of observing the characteristic changes in CSF composition. Marked neutrophil pleocytosis of the CSF may be seen transiently in association with abrupt deterioration of the headache, increased neck stiffness, and fever, suggesting bacterial meningitis. The phenomenon is self-limited, and is thought to be the result of a rupture of a microtubercle into the subarachnoid space. Culture of mycobacteria is successful in 40 to 60 per cent of cases. Specimens other than CSF (for example, sputum, gastric washings in children, urine, etc.) should also be cultured. Since it can take up to 2 months or longer for the mycobacteria to grow in the standard cultures, a number of methods for the rapid diagnosis of TBM have therefore been developed. The bromide partition test is not of practical use, because it lacks specificity and is abnormal in other types of meningitis and even in cerebral malaria. It is no more useful than the CSF glucose concentration. Tests for detecting membrane antigen of tubercle bacilli in CSF by enzyme immunoassay and latex-particle agglutination are disappointing because they are not reproducible and do not have sufficient sensitivity and specificity to be of clinical use. Detection of tuberculostearic acid (a lipid component of the mycobacterial cell wall) in the CSF by gas chromatography and mass spectroscopy is highly specific and sensitive, but the apparatus is very expensive and technically too complicated for clinical use. Detection of the genome of *M. tuberculosis* in the CSF by PCR can be specific and sensitive in the rapid diagnosis of TBM. Recent studies in centres of excellence suggest a sensitivity of 75 per cent and a specificity of 94 per cent; however, further work on larger sample sets are still required to establish the utility, reproducibility, and cost-effectiveness of this approach. Cell-mediated immunity should be assessed. The tuberculin test is positive in 50 to 95 per cent of patients with TBM. Reactivity is suppressed in debilitated or immunosuppressed patients. Serological tests for HIV infection and a CD4 cell count should, if possible, be performed in every case. Immunological assays based on the production of g-interferon following stimulation with *M. tuberculosis*-specific antigens (ESAT6 and CFP10) (ELI-SPOT TEST) have shown promise as a diagnostic tool among contacts of patients with TB. Their potential role in TB meningitis has not yet been established. A search for evidence of tuberculosis elsewhere in the body is useful. Chest radiographs are normal in about half of the patients, and miliary mottling is seen in only a minority.

The advent of computed tomography (**CT**) scanning and magnetic resonance imaging (**MRI**) has provided insight into disease progression, and gives prognostic and

diagnostic information. Both CT and MRI of the brain will reveal hydrocephalus, basilar meningeal thickening, infarcts, oedema, and tuberculomas (Fig. 7 and Fig. 8).

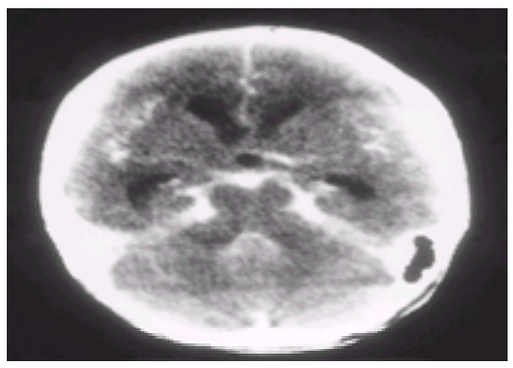

Fig. 7 CT scan with enhancement showing thick basal exudates and hydrocephalus. (Copyright the late Prida Phuapradit.)

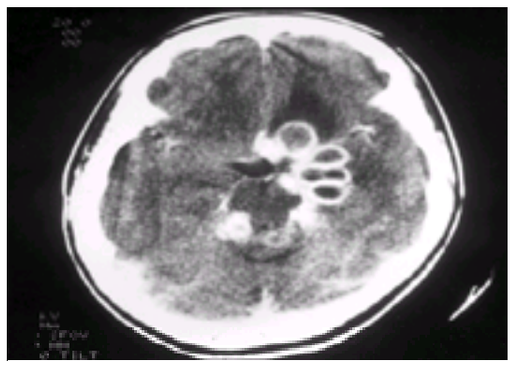

Fig. 8 CT scan with enhancement of a patient with TBM showing multiple tuberculomas developing near the basal cisterns (Copyright the late Prida Phuapradit.)

In a CT study of 60 cases of TBM in adults and children, only three had normal brain scans. Hydrocephalus was reported in 87 per cent of children and 12 per cent of adults. The incidence of hydrocephalus is greater in the young, and increases with duration of the illness. In children, hydrocephalus is invariably present after 6 weeks of illness. Infarcts are seen on CT scanning in 28 per cent, with 83 per cent occurring in the middle cerebral artery territory. The basal ganglia are the most commonly affected region. Poor prognosis has been associated with enhancing basal exudates and periventricular lucency.

MRI has increased sensitivity in detecting the distribution of meningeal inflammatory exudate. Gadolinium-enhanced, *T*1-weighted images highlight the exudate, and reveal parenchymal infarcts as hyperintense areas. MRI may provide more diagnostic information than CT when assessing space-occupying lesions. Cerebral miliary TB, with multiple small intraparenchymal granulomas, produces moderate perilesional oedema and contrast enhancement. Larger tuberculomas are initially non-enhancing, but later demonstrate marked enhancement.

CT and MRI are sensitive for the changes of TBM, particularly hydrocephalus and basal meningeal exudates, but they lack specificity. The radiological differential diagnosis includes cryptococcal meningitis, cyto-megalovirus encephalitis, sarcoidosis, meningeal metastases, and lymphoma. The major role of neuroradiology has been in the management and, in particular, in the diagnosis and follow-up of those complications requiring neurosurgery.

Differential diagnosis

Although a few patients with TBM present acutely, the majority show a subacute or chronic progression. Clinically, the differential diagnosis should include cryptococcal meningitis and various subacute or chronic meningitides, including partially treated bacterial meningitis, parameningeal infections, neoplastic and granulomatous infiltrations of the meninges (for example, carcinomas, leukaemias, lymphomas, sarcoidosis), and cerebral tumours. Fungal meningitides (with *Cryptococcus* spp., Coccidioides, *Histoplasma* spp., *Candida* spp.) may present like TBM. Other conditions that have caused confusion in particular clinical or geographical settings are meningovascular syphilis, toxoplasma meningitis in immunosuppressed patients (notably in those with AIDS), cysticercosis, amoebic meningoencephalitis, African trypanosomiasis, and schistosomal myelopathy. In most of these cases, CSF examination, including cytology, an India-ink preparation, immunodiagnostic tests (for example, cryptococcal latex agglutination), serological tests, and microbial cultures will allow differentiation.

Treatment

The untreated mortality of TBM is close to 100 per cent. Full treatment must be started when the diagnosis is suspected on clinical grounds, immediately after adequate samples have been taken for microscopy, culture, and immunodiagnosis. The choice of antituberculosis drugs has been based mainly on their pharmacokinetic and antibacterial properties. Isoniazid, pyrazinamide, ethionamide, and cycloserine freely distribute into the CSF. Penetration is limited, but still adequate during the first few months of the meningitis, in the case of rifampicin, ethambutol, and streptomycin. *p*-Aminosalicylic acid should not be used because it does not enter the CSF. At least two drugs to which the organism is sensitive should be used. During the first 2 months, however, intensive chemotherapy with three or four antituberculosis drugs should be given: the impairment of the blood–brain barrier during the active stage of meningitis allows most antituberculosis drugs to enter the CSF in amounts sufficient to kill the organism, and offers a good chance of eliminating mycobacteria from the CSF after a short period of treatment. The combination of isoniazid and rifampicin for 12 months, with pyrazinamide and streptomycin during the first 2 months, is an effective regimen that has been widely used. In adults, daily single doses of 300 mg of isoniazid, 600 mg of rifampicin, and 1500 mg of pyrazinamide provide adequate levels in the sera and CSF of patients with active TBM. Higher doses of these drugs are unnecessary and may result in a higher incidence of hepatotoxicity. The incidence of adverse reactions of antituberculosis drugs in TBM is acceptable and is similar to that encountered in the treatment of pulmonary tuberculosis.

In countries where rifampicin cannot be afforded, triple therapy with isoniazid, pyrazinamide, and streptomycin should be tried. TBM in HIV-infected patients should be similarly treated as the meningitis in those not infected. Responses to treatment and the outcome of TBM are similar in the two groups. Drug-resistant isolates of *M. tuberculosis,* which have been found increasingly, poses problems in the treatment now and in the future. Ethionamide, kanamycin or amikacin, and cycloserine should be considered in this situation. Development of new effective antituberculosis drugs is urgently needed.

Response to antituberculosis chemotherapy is slow, particularly in patients who are not given corticosteroids. There may be an increase in temperature and CSF protein concentration as well as a transient neutrophil pleocytosis during the first 2 months after starting optimal chemotherapy. However, some signs of clinical improvement are usually seen within the first few weeks. Early clinical evidence of response is an improvement of the headache, sense of general well being, and a decrease in the elevated intracranial pressure. A rapid return to normal in CSF composition within a few days virtually excludes the diagnosis of TBM, in which case antituberculosis treatment should be stopped. Usually it would take at least a few weeks to a few months for the cells, CSF glucose, and protein levels to return to normal. However, the high protein concentration persists in some patients.

'Trial' of chemotherapy is justified when there is clinical suspicion of TBM, particularly when diagnostic facilities are limited. Treatment should be continued for 12 months unless there is rapid improvement in the patient's condition and CSF composition, suggesting another cause for aseptic meningitis. In some severely ill patients who present acutely with features of acute bacterial meningitis (for example, neutrophil pleocytosis) but in whom initial laboratory results are unhelpful, it may be necessary to initiate 'blind' treatment for acute bacterial and TBM simultaneously. In these, fortunately rare cases, isoniazid, rifampicin, and streptomycin or ethambutol, together with penicillin or a third-generation cephalosporin, can be given.

There is conflicting evidence regarding the length of treatment. The current United Kingdom guidelines recommend 12 months in uncomplicated cases of TBM (including cerebral tuberculoma without meningitis), extending to 18 months should pyrazinamide be omitted. No guidelines exist for the components and duration of treatment in the case of multidrug-resistant TBM. Treatment for 12 months is probably a conservative estimate of the time required for bacterial cure. Different regimes, incomparable patient groups, and the variable use of adjuvant steroid therapy, makes meta-analysis from the trials impossible. Some suggest that TBM should be treated for a minimum of 2 years. Evidence from 781 cases of TBM treated for 2 years revealed that 35 patients had a recrudescence; however, nearly all patients with relapse had received less than 6 months' therapy, indicating that therapy should continue in excess of this period. In South Africa, 95 children were treated for 6 months with a combination of isoniazid 20 mg/kg, rifampicin 20 mg/kg, pyrazinamide 40 mg/kg, and ethionamide 20 mg/kg. Some 96 per cent of these cases presented in either stage II or III TBM. The doses of both isoniazid and rifampicin used were considerably higher than those recommended in the United Kingdom, but no serious adverse reactions were reported. The study provides good evidence for the adequacy of short-course intensive chemotherapy, but the lack of a control group does not allow conclusions to be drawn regarding the optimal dosages. Studies using 9 months' chemotherapy (2 months' isoniazid, rifampicin, pyrazinamide, streptomycin, followed by 7 months' rifampicin and isoniazid) at lower doses produced similar outcomes.

Treatment of complications

The complications of TBM are common, some of which are often serious enough to cause severe morbidity and death in spite of active treatment with antituberculosis drugs. Increased intracranial pressure, found in about 90 per cent of the patients and often associated with communicating hydrocephalus, is usually caused by the basal arachnoiditis. Less commonly, it is caused by diffuse cerebral oedema which compresses the small lateral ventricles. In these patients, conservative treatment by repeated lumbar punctures in combination with acetazolamide, with or without furosemide (frusemide), and corticosteroids should be tried. Within the first 4 to 6 weeks of treatment, the CSF pressures in the majority of the patients return to normal and the transependymal oedema (periventricular lucency on the CT scan) disappears. However, the sizes of the ventricles usually remain unchanged during the first 4 to 6 weeks of treatment. This is a state of arrested hydrocephalus. Usually, it would take up to 6 months for the ventricular size to return to normal. In a few centres, intrathecal hyaluronidase, which might resolve the basal exudates, has been shown to be beneficial in the treatment of communicating hydrocephalus and spinal arachnoiditis. If these conservative measures fail, surgical treatment will be needed for the hydrocephalus. Patients with communicating hydrocephalus who are very ill with impairment of consciousness should receive early temporary external ventricular shunting. Shunt surgery, as a first-line treatment, should be reserved for patients with non-communicating hydrocephalus.

Corticosteroids might be expected to reduce the inflammatory reaction and the fibrotic organization of exudate in the brain and spinal cord. A recent Cochrane review analysed six trials comprising a total of 595 patients. Steroids were associated with fewer deaths (relative risk (**RR**) 0.79; 95 per cent confidence interval (**CI**) 0.65 to 0.97) and a reduced incidence of death and severe residual disability (RR, 0.58; 95 per cent CI, 0.38 to 0.88). Subgroup analysis suggested an effect on mortality in children (RR, 0.77; 95 per cent CI, 0.62 to 0.96), but the results in a smaller number of adults were inconclusive (RR, 0.96; 95 per cent CI, 0.50 to 1.84). There is little evidence that the severity of disease influences the effects of steroids on mortality. The conclusion of the Cochrane reviewers was that adjunctive steroids might be of benefit in patients with TBM. However, existing studies were small, and publication bias may account for the positive results. The data are stronger for the use of steroids in children than in adults. No data are available on the use of steroids in HIV-positive patients. The usual regimen is intramuscular dexamethasone (16 mg/day in adults and 0.5 mg/kg per day in children) in divided doses, or oral prednisolone 60 mg/day for adults, 2 mg/kg per day for children, given in a tapering course over 3 to 6 weeks. There is no evidence that corticosteroids interfere with the penetration of antituberculosis drugs into the CSF. Intrathecal injection of corticosteroids is unnecessary.

Fluid, electrolyte, and acid–base disturbances are common, the result of vomiting, inadequate fluid intake, and SIADH. Progressive loss of vision caused by fibrosing exudate around the optic chiasma may respond to surgical decompression. Cerebral tuberculomas may occasionally develop during the course of the treatment of TBM. Characteristically, the lesions are thick-walled nodules of small or moderate size in clusters, and develop in the surface of the brain near the basal cisterns, such as the interpeduncular cistern and cistern of the optic chiasm (Fig. 8). They should be treated conservatively and the response to antituberculosis treatment assessed by CT scan. Biopsy or surgical intervention is unnecessary and may be harmful. Tuberculomas usually respond very slowly to antituberculosis drugs and it usually takes at least 24 months or longer for the lesions to disappear. Nursing care is very important during the acute illness, when there are the usual problems presented by unconscious patients, and during the prolonged phase of convalescence and rehabilitation. Anticonvulsants are often needed, especially in children.

Prognosis and sequelae

In Western industrialized countries, the mortality from TBM is still high, at about 15 to 30 per cent, and in developing countries it remains between 30 and 50 per cent. The prognosis is worst and the risk of sequelae highest in those admitted in coma with signs of brainstem damage, in the very young and very old, pregnant women, and those with malnutrition or other diseases. The outcome of TBM in HIV-infected patients is similar to that in patients without HIV infection. There are permanent sequelae in 10 to 30 per cent of survivors. Intellectual impairment is especially common in infants and young children. As many as 60 per cent of patients who have seizures during the illness will suffer recurrences. Up to 25 per cent of survivors will have cranial nerve deficits, including blindness, deafness, and squints. Some 10 to 25 per cent of survivors have some residual weakness after hemiparesis or paraparesis. About 10 per cent of patients develop CSF spinal block at some stage of the illness, but this will recover completely in at least half of them. Occasionally neurological deficit may progress or appear months after the illness as the subarachnoid exudate becomes fibrotic and calcified.

Prevention (see also Chapter 7.11.22)

BCG vaccination at birth reduces the risk of infection by at least 80 per cent, but this seems to vary in different countries. It is recommended for all infants born into communities where tuberculosis is prevalent, including Asians living in Britain and expatriates living in tropical countries. To prevent the development of TBM in household contacts of newly diagnosed cases of pulmonary tuberculosis, prophylaxis with isoniazid 10 mg/kg daily for 6 to 12 months is recommended for all Mantoux-positive children under the age of 5 years.

---

*Contains material contributed to OTM3 by the late Prida Phuapradit.

Further reading

Acute bacterial meningitis

Anonymous (2000). The management of neurosurgical patients with postoperative bacterial or aseptic meningitis or external ventricular drain-associated ventriculitis. *British Journal of Neurosurgery* **14**, 7–12.

Bisno AL (1994). Infections associated with indwelling medical devices. In: Bisno AL, Waldvogel FA, eds. *Infections associated with indwelling medical devices*, 2nd edn, pp 93–109. American Society Medical Press, Washington DC.

Christie AB (1980). *Infectious diseases: epidemiology and clinical practice*, 3rd edn, pp 605–46. Churchill Livingstone, Edinburgh.

Durand ML, *et al.* (1993). Acute bacterial meningitis in adults. *New England Journal of Medicine* **328**, 21–8.

Girgis NI, *et al*. (1990). Dexamethasone for the treatment of children and adults with bacterial meningitis. *Reviews of the Infectious Diseases* **12**, 963–4.

Gordon SB, *et al.* (2000). Bacterial meningitis in Malawian adults: pneumococcal disease is common, severe and seasonal. *Clinical Infectious Diseases* **31**, 53–7.

Gray LD, Fedorko DP (1992). Laboratory diagnosis of bacterial meningitis. *Clinical Microbiological Reviews* **5**, 130–45.

Infection in Neurosurgery Working Party of the British Society for Antimicrobial Chemotherapy. (2000). The rationale use of antibiotics in the treatment of brain abscess. *British Journal of Neurosurgery* **14**, 525–30.

Kay R, Cheng AF, Tse CY (1995). *Streptococcus suis* infection in Hong Kong. *Quarterly Journal of Medicine* **88**, 39–47.

Molyneux EM, *et al.* (2002). Dexamethasone treatment in childhood bacterial meningitis in Malawi: a randomised controlled trial. *Lancet* **360**, 211–18.

Rathore MH (1991). Do prophylactic antibiotics prevent meningitis after basilar skull fracture? *Pediatric Infectious Disease Journal* **10**, 87–8.

Schaad UB, *et al.* (1993). Dexamethasone therapy for bacterial meningitis in children. *Lancet* **342**, 457–61.

Scheld WM, Whitley RJ, Durack DT (1997). *Infections of the central nervous system*, 2nd edn. Lippincott-Raven, Philadelphia.

Tedder DG, *et al.* (1994). Herpes simplex virus infection as a cause of benign recurrent lymphocytic meningitis. *Annals of Internal Medicine* **121**, 334–8.

Tugwell P, Greenwood BM, Warrell DA (1976). Pneumococcal meningitis: a clinical and laboratory study. *Quarterly Journal of Medicine* **45**, 583–601.

Tuberculous meningitis

Alarcon F, *et al.* (1990). Tuberculous meningitis: short course chemotherapy. *Archives of Neurology* **47**, 1313–17.

Berenguer J, *et al.* (1992). Tuberculous meningitis in patients infected with the human immunodeficiency virus. *New England Journal of Medicine* **327**, 668–72.

Donald PR, *et al.* (1988). Intensive short course chemotherapy in the management of tuberculous meningitis. *International Journal of Tuberculosis and Lung Disease* **2**, 704–11.

Girgis NI, *et al*. (1991). Dexamethasone adjunctive treatment for tuberculous meningitis. *Pediatric Infectious Disease Journal* **10**, 179–83.

Goel A, Pandya S, Satoskar A (1990). Whither short-course chemotherapy for tuberculous meningitis. *Neurosurgery* **27**, 418–21.

Joint Tuberculosis Committee of the British Thoracic Society (1998). Chemotherapy and management of tuberculosis in the United Kingdom: recommendations 1998. *Thorax* **53**, 536–48.

Kaojarern S, *et al.* (1991). Effect of steroids on cerebrospinal fluid penetration of antituberculosis drugs in tuberculous meningitis. *Clinical Pharmacology and Therapeutics* **49**, 6–12.

Parsons M (1988). *Tuberculous meningitis. A handbook for clinicians*, 2nd edn. Oxford University Press, Oxford.

Phuapradit P, Vejjavjiva A (1987). Treatment of tuberculous meningitis: role of short-course chemotherapy. *Quarterly Journal of Medicine* **62**, 249–58.

Prasad K, Volmink J, Menon GR (2000) Steroids for treating tuberculous meningitis (Cochrane review). *Cochrane Database Systematic Reviews* (3), CD002244.

Schoeman JF, *et al*. (1985). Intracranial pressure monitoring in tuberculous meningitis: clinical and computerized tomographic correlation. *Developmental Medicine and Child Neurology* **27**, 644–54.

Schoeman JF, *et al.* (1991). Tuberculous hydrocephalus: comparison of different treatments with regard to intracranial pressure, ventricular size and clinical outcome. *Developmental Medicine and Child Neurology* **33**, 396–405.

Shankar P, *et al.* (1991). Rapid diagnosis of tuberculous meningitis by polymerase chain reaction. *Lancet* **337**, 5–7.

Tartaglione T, *et al.* (1998). Diagnostic imaging of neurotuberculosis. *Rays* **23**, 164–80.

Thwaites G, *et al.* (2000) Tuberculous meningitis. *Journal of Neurology, Neurosurgery, Psychiatry* **68**, 289–99.

Visudhiphan P, Chiemchanya S (1989). Tuberculous meningitis in children: treatment with isoniazid and rifampicin for twelve months. *Journal of Pediatrics* **114**, 875–9.

24.14.2 Viral infections of the central nervous system

*D. A. Warrell and J. J. Farrar**

Viruses invade and damage the central nervous system in two ways: directly, by infecting the leptomeninges, brain, and spinal cord; and, indirectly, by inducing an immunological reaction resulting in para- and postinfectious diseases. In both cases, the terms 'meningitis', 'encephalitis', and 'myelitis' are used alone or in combination. Meningitis implies inflammation of the meninges without alteration of consciousness, convulsions, or the production of focal neurological abnormalities; in encephalitis there is impairment of cerebral function, usually with an altered state of consciousness and often with convulsions and focal neurological signs; while myelitis indicates involvement of the spinal cord. Retroviral and prion diseases of the central nervous system are dealt with elsewhere (see Chapter 7.10.21, Chapter 7.10.22, Chapter 7.10.23 and Chapter 24.13.9).

Virology

There is considerable geographical and seasonal variation in the kinds of viruses causing meningitis, myelitis, and encephalitis. However, compared with bacterial infections of the central nervous system, there is less variation with age and immunocompetence.

Enteroviruses are responsible for 80 to 90 per cent of diagnosed cases of viral meningitis. Almost all the serotypes have been implicated in sporadic cases, and outbreaks have been associated with coxsackieviruses A7 and 9, all the coxsackie B types, and many of the echoviruses, especially 4, 6, 9, 11, 14, 16, and 30. Recently, echovirus 13, a rare type, has caused cases in the United States, Australia, and Europe and there has been an increase in echovirus 30 cases. Mumps is responsible for about 10 to 20 per cent of cases of viral meningitis. Other less common causes include herpes zoster, herpes simplex virus (predominantly type 2, HSV-2), measles, adenoviruses, Epstein–Barr virus and, in the United States, togaviruses, such as St Louis, eastern and western equine encephalitis, and West Nile and bunyaviruses, such as California (La Crosse) encephalitis viruses.

Poliovirus has long been considered the major cause of viral 'paralytic' myelitis throughout the world, but has now been virtually eliminated from the Americas. A confusingly similar syndrome of acute flaccid paralysis caused by Japanese encephalitis (**JE**) has now been reported from Vietnam. Coxsackie A7 (AB IV) has caused occasional small outbreaks, and other coxsackie A and B viruses, echoviruses, enterovirus 70, and flaviruses (tick-borne encephalitis) have all been implicated as causes of flaccid paralysis. Herpes zoster, paralytic rabies, Epstein–Barr, and *Herpes simiae* B viruses can cause myelitis or ascending paralysis, and HSV-2 can cause lumbosacral myeloradiculitis.

Viruses causing encephalitis vary from country to country. JE virus is the most widespread human togavirus infection in the world and is the major cause of encephalitis throughout Asia. There are at least 50 000 cases of JE with 15 000 deaths annually. The virus is transmitted by Culex mosquitoes and is endemic across much of southern Asia and the Indian subcontinent. It is spreading through the Pacific rim to New Guinea, Torres Strait Islands, and Cape York Peninsula, in northern Australia. Dengue viruses have been implicated as a cause of encephalitis in South-East Asia.

In 1999 an outbreak of an encephalitic illness among pig farm and abattoir workers was reported from Singapore and Malaysia. There were 258 cases of encephalitis, with a case fatality rate of almost 40 per cent. The causative agent was a new paramyxovirus, Nipah virus, closely related to the Hendra and Manangle viruses described in Australia. Nipah virus encephalitis is a zoonosis infecting pigs and flying foxes (*Pteropus* spp.). Almost all patients infected in this outbreak had direct contact with pigs. Hendra virus has caused a few cases of equine and human encephalitis (Chapter 7.10.6.1).

In North America, herpes simplex virus is the most common cause of sporadic fatal viral encephalitis, followed by the California encephalitis group, St Louis encephalitis virus, herpes zoster, enteroviruses, mumps, measles, and, most recently, the West Nile virus. In the United States, herpes simplex encephalitis has an estimated incidence of 2.3 per million population each year; HSV-1 accounts for 95 per cent of cases; HSV-2 causes encephalitis mainly in neonates and those who are immunosuppressed, such as transplant patients and those with human immunodeficiency virus (**HIV**) infection. In 1999 there was an outbreak of West Nile infection in the eastern United States with a cluster of cases of encephalitis in New York and, since then, 16 human deaths. West Nile virus is a mosquito-borne flavivirus closely related JE. It has been known to cause encephalitis in Africa, the Middle East, and southern and eastern Europe, but this was the first appearance of this virus in the New World. In endemic areas, infection with West Nile virus is usually asymptomatic or associated with a mild 'flu-like' illness. Only occasionally does it cause encephalitis, with a case fatality rate for patients admitted to hospital in New York of 12 per cent. The virus has now become established in migrant bird populations along the Atlantic coast of the United States and has killed hundreds of thousands of crows.

In the United Kingdom, mumps is the most frequently diagnosed viral encephalitis, followed by echoviruses, coxsackieviruses, measles, herpes simplex virus, herpes zoster virus, Epstein–Barr virus, and adenoviruses (especially adenovirus 7). Louping ill is the only indigenous arthropod (tick)-borne encephalitis in Britain. In Central and Eastern Europe and Scandinavia, tick-borne encephalitis virus and Russian spring–summer encephalitis viruses are endemic. Usutu, a flavivirus, has been isolated in birds in Austria. In many developing countries rabies is an important cause of viral encephalitis. Other regional causes are Rift Valley fever virus in Africa and the Middle East, arenaviruses (Junin, Guanarito, Sabiá, Lassa, and Machupo) in Latin America and Africa, Marburg and Ebola viruses in Africa, Colorado tick fever virus in North America, and Murray Valley encephalitis virus in Australia.

Postinfectious encephalomyelitis most commonly follows measles, vaccinia, varicella, rubella, mumps, and influenza. Guillain-Barré syndrome, a sensorimotor polyneuropathy (see Chapter 24.19), has been associated with infections by Epstein–Barr virus, cytomegalovirus, coxsackie B, and herpes zoster virus. Nervous-tissue vaccine against rabies may give rise to postvaccinal encephalitis (see below), while vaccination against influenza, rabies, hepatitis B, measles, and

poliomyelitis has been complicated by Guillain-Barré syndrome.

Immunodeficient patients are particularly vulnerable to some viral infections. Those with depressed cell-mediated immunity (Hodgkin's disease) may develop herpes zoster encephalitis, and cytomegalovirus may cause a subacute encephalitis in patients with acquired immunodeficiency syndrome (**AIDS**). In children or adults with hypogammaglobulinaemia, enteroviruses, including live-attenuated polio vaccine, may produce a progressive and fatal meningoencephalitis. Progressive multifocal leucoencephalopathy, a chronic and fatal papovavirus infection in patients with impaired cell-mediated immunity, is described below. HIV infection of the brain and meninges may be responsible for acute meningoencephalitis at the time of seroconversion and for subacute chronic encephalopathies and dementia in patients with AIDS (see Chapter 7.10.23 and Chapter 24.14.4).

Epidemiology

Many viral infections of the central nervous system occur in seasonal peaks or as epidemics, while others, such as herpes simplex encephalitis, are sporadic. Epidemics of Japanese encephalitis occur in the summer or rainy season in northern India, Nepal, northern Thailand, Korea, Taiwan, and China. However, in southern Vietnam, Indonesia, Malaysia, southern India, and the Philippines the disease can occur all the year round, although the peak is at the start of the rainy season. This variation in the incidence of disease is an important consideration when recommending vaccination. In endemic areas it is mostly a disease of children, but as the disease spreads to new regions, or non-immune travellers visit endemic regions, non-immune adults are also affected. The major vector is *Culex tritaeniorhynchus* mosquitoes which have been infected by first feeding on the bird (cattle egrets, herons) or mammal reservoir species. Indigenous children and non-immune (immigrant) adults are most susceptible. Tick-borne encephalitides occur in spring and early summer when the ticks are most active. Mumps encephalitis is commonest in the late winter or early spring, while enterovirus infections occur most often in the summer and early autumn. Rodent-related encephalitides, such as the arenaviruses, are most common when the rodent population is at its peak, either in the fields (Machupo and Junin viruses) or in the home (lymphocytic choriomeningitis virus). Zoonotic viral infections, such as Rift Valley Fever, survive periods of cold weather, during which the invertebrate–vertebrate cycle is suspended by 'overwintering' in their arthropod vectors (for example, in the bottom of dried-up ponds) or hibernating vertebrate reservoirs.

Invasion of the central nervous system seems to be a rare event in most viral infections. In the case of some togavirus infections, such as JE, there may be only one case of encephalitis for every 500 to 1000 asymptomatic infections. Eastern equine encephalitis virus produces a much higher proportion of encephalitic cases than other togaviruses.

Infections by many neurotropic viruses are most frequent and severe in children and the elderly. Herpes simplex encephalitis affects all age groups but shows peaks of incidence in those aged between 5 and 30 years and over 50 years. When HSV-2 invades the central nervous system it is likely to cause a benign lymphocytic meningitis in adults, but in neonates it usually produces a severe encephalitis. Among mosquito-borne epidemic encephalitides, California encephalitis and JE are most common in children, St Louis and West Nile encephalitis in the elderly, while eastern and western equine encephalitis affect both the very young and the elderly. Postinfectious encephalitis is most frequent in children, for it complicates the common childhood exanthematous viral infections. It is the most common demyelinating disease in the world.

Pathogenesis

Most viral infections reach the central nervous system from the primary site of infection and multiplication via the bloodstream, but the rabies virus enters peripheral nerves through acetylcholine and other receptors and travels to the CNS in axoplasm, employing the microtubular dynein motor system. Viruses inoculated through the skin include those transmitted by arthropods, rabies virus, herpes simplex virus, *Herpes simiae* B, and lymphocytic choriomeningitis virus. Arthropod-borne viruses are presumed to replicate in local lymph nodes, vascular endothelium, and circulating fixed macrophages, in order to sustain viraemia. Rabies virus may multiply locally in the cytoplasm of muscle cells before entering peripheral nerves. Viruses that enter through the respiratory tract (for example, measles, mumps, varicella) or gut (enteroviruses) multiply in local lymphoid tissue before entering the bloodstream. Viraemia is a feature of most viral infections, yet invasion of the central nervous system is rare in most cases. The explanation for this is not known, but the CNS contains a number of intrinsic physical barriers to infectious agents such as viruses. These include the blood–brain barrier with its 'tight junctions', virus-resistant cells, and the absence of lymphatic drainage. Non-specific mechanisms at or near the site of virus entry, such as gastric acidity and cilia in the respiratory tract, also play a protective role. In the case of rabies, herpes simplex, and herpes zoster viruses, the virus enters the central nervous system through the peripheral nerves. Although the subarachnoid space surrounding the olfactory nerves projects through the cribriform plate and is directly beneath the nasal mucosa, this route of infection seems to be extremely rare in humans and has been proven only in a few cases of inhaled rabies virus infection and herpes simplex encephalitis. Viruses have been inoculated directly into the central nervous system by infected corneal transplant grafts (rabies) and prions through infected brain-surface electrodes (Creutzfeldt–Jakob disease). Herpes simplex encephalitis may complicate primary herpes simplex virus infection in children and young adults, but in most cases of herpes simplex encephalitis the cause is thought to be reactivation of latent virus (HSV-1) in the trigeminal nerve, autonomic nerve roots, or brain.

Some viruses, such as the enteroviruses and mumps, usually infect the meninges rather than the parenchyma of the central nervous system, whereas others, such as the togaviruses, usually cause encephalitis. Different neural cells are selectively vulnerable to different neurotropic viruses. Examples are the predilection of polioviruses for motor neurones of the anterior horns of the spinal cord, and of rabies for neurones of the limbic system and cerebellar Purkinje cells. The pathological effects of viral infections on the central nervous system include:

1. the destruction and phagocytosis of neurones (neuronophagia) as a result of either viral invasion *per se* or immune lysis;
2. demyelination;
3. inflammatory oedema with the compressive effects of raised intracranial pressure; and, in some cases,
4. vascular lesions.

In rabies, a universally fatal encephalitis, neuronolysis is relatively mild. However, rabies virus may interfere with neurotransmission at central and peripheral synapses. It also produces severe systemic effects, following its centrifugal spread (for example, myocarditis and cardiac arrhythmias) or its focal effects on vasomotor and respiratory centres in the brainstem and in the temporal (lobes and amygdala (cf. Klüver–Bucy syndrome) (Chapter 7.10.9).

Postinfectious encephalitis and the Guillain-Barré syndrome are thought to result from sensitization to central and peripheral myelin, respectively. The animal model for the former is experimental allergic encephalomyelitis, which can be produced in a variety of animals following immunization with myelin basic protein. A similar animal model for Guillain-Barré syndrome is known as experimental allergic neuritis. It is uncertain how the preceding viral infection induces this autoimmune response. In the case of postvaccinal encephalomyelitis resulting from old-fashioned nervous tissue antirabies vaccines containing homogenized animal brain, the mechanism is still not clear. The anti-myelin basic protein is not always present and is probably not the direct cause of demyelination.

The host's immune responses to viruses play a crucial role in combating infection. They may be directed against either the virus particle or the virus-infected cell, and may be humoral- or cell-mediated. An important local immune response at infected surfaces is provided by IgA antibody, which is present in secretions in the gut, saliva, and respiratory tract. This is important, for example, in the early stages of poliovirus infection where the antibody neutralizes the virus by combining with viral surface proteins. The systemic viral infection may also be limited by means of circulating IgG and IgM antibodies, which can neutralize the virus in a variety of different ways. Immune responses may also occur locally within the CNS, where local synthesis of immunoglobulins in response to virus infection, sometimes in an oligoclonal pattern, may be evident. Such antibody elevations may be of considerable diagnostic value (see below). Under certain conditions immune responses to viruses may themselves set in train immunopathological processes leading to disease. This may occur in a number of different ways, such as through the deposition in blood vessels of immune complexes formed between an antiviral antibody and viral antigen. In other cases, such as lymphocytic choriomeningitis virus infection, the induction of virus-specific cytotoxic T lymphocytes is itself responsible for the production of encephalitis.

Pathology

Meningitis

The basal leptomeninges, ependyma, and choroid plexus are infiltrated with mononuclear cells but the parenchyma is normal. In mumps meningitis there may be exfoliation of ependymal cells.

Poliomyelitis

Virus is distributed widely throughout the brain and spinal cord, possibly even in non-paralytic cases, but usually the only cells to suffer chromatolysis and phagocytosis are motor neurones in the anterior horns of the spinal cord, medulla, and grey matter of the precentral gyrus.

Encephalitis

Most viral encephalitides are characterized by lymphocytic infiltration of the meninges and perivascular cuffing (in Virchow–Robin spaces) in the cortex and underlying white matter, by lymphocytes, plasma cells, histiocytes, and some neutrophils, and proliferation of microglia with the formation of glial nodules. Neuronolysis and demyelination are variable in their degree and location. Infected neurones may show characteristic inclusion bodies in their nuclei (measles, herpes simplex virus, and adenoviruses) or cytoplasm (Negri bodies in rabies). Microhaemorrhages and foci of necrosis may be found.

Herpes simplex encephalitis

Characteristic features of this condition are gross cerebral oedema and severe haemorrhagic and necrotizing encephalitis, which is often asymmetrically localized to the inferior and medial parts of the temporal lobe, the insula, and the orbital part of the frontal lobe. Histological sections show eosinophilic Cowdry type A intranuclear inclusions with margination of chromatin in neurones, oligodendrocytes, and astrocytes, inflammatory and haemorrhagic perivascular reactions, but no demyelination. Cowdry type A inclusions are also found in herpes zoster virus and cytomegalovirus encephalitides. The unique cerebral localization of herpes simplex encephalitis has not been satisfactorily explained, but is probably the result of viral spread along specific neural pathways rather than a differential susceptibility of particular cell populations. A popular idea is that herpes simplex virus spreads along olfactory pathways to the base of the brain and temporal lobes, but it is also possible that virus may spread from the trigeminal ganglia through sensory fibres innervating the dura near these regions. This latter mechanism is consistent with the discovery of latent HSV-1 in the trigeminal, superior cervical, and vagal ganglia in a high proportion of normal individuals, irrespective of whether they have a history of mucocutaneous herpes infections ('cold sores'). Latent HSV-1 might be reactivated by a variety of stimuli, such as sunlight, fever, trauma, and stress; however, the actual mechanisms underlying its latency and reactivation in the nervous system are not yet fully understood. If herpes simplex encephalitis is caused by the reactivation of latent virus, its rarity, despite ubiquitous asymptomatic infection in humans, is hard to explain.

Japanese encephalitis

Microscopical appearances are typical of other viral encephalitides: there is oedema, congestion, and focal haemorrhages of the brain and meninges, and perivascular cuffing, neuronophagia, and glial nodules of the brain parenchyma. Neuronolysis and neuronophagia are unusually widespread in the thalamus, basal ganglia, brainstem, cerebellum (where there is marked destruction of Purkinje cells), and the spinal cord. Viral antigen is localized to neurones, especially in the brainstem and thalamus.

Nipah virus encephalitis

Pathological studies on the brains of fatal cases demonstrated that the endothelium of small blood vessels in the central nervous system was particularly susceptible to infection. This led to disseminated endothelial damage and syncytium formation, vasculitis, thrombosis, ischaemia, and microinfarction. There was also evidence of neuronal infection by the virus that may have contributed to neurological dysfunction.

West Nile virus encephalitis

Pathological studies from the outbreak of this encephalitis in New York showed varying degrees of neuronal necrosis in the grey matter, with infiltrates of microglia and polymorphonuclear leucocytes, perivascular cuffing, neuronal degeneration, and neuronophagia. Viral antigens were demonstrated in neurones and in areas of necrosis. No antigen was detected in other major organs, including lung, liver, spleen, and kidney. The major pathological lesions were seen in the brainstem and spinal cord.

Enterovirus 71

Pathological studies of patients who died of enterovirus 71 infection showed severe perivascular cuffing, parenchymal inflammation, and neuronophagia in the spinal cord, brainstem, and diencephalon, and in focal areas in the cerebellum and cerebrum. Although no viral inclusions were detected, immunohistochemistry showed viral antigen in the neuronal cytoplasm. Inflammation was often more extensive than neuronal infection, suggesting that other indirect factors may be involved in tissue damage in addition to the effects of direct viral invasion.

Postinfectious encephalomyelitis

This is a perivenous microglial encephalitis with demyelination. Fibrinoid necrosis of arterioles is an associated lesion in a more severe form designated 'acute haemorrhagic leucoencephalitis'.

Clinical features

Meningitis

A prodromal influenza-like illness, followed by a brief remission of symptoms, is typical of lymphocytic choriomeningitis viral infection and some outbreaks of enteroviral meningitis (for example, echovirus 9), but in most cases of viral meningitis, symptoms start suddenly. As with bacterial meningitis, there is fever, headache, a stiff neck, and vomiting, especially in children. Compared with bacterial meningitis, headache is less severe and tends to be frontal or retrobulbar (eye movements may be painful) and neck stiffness is less marked. Nausea, anorexia, abdominal pain, myalgias, and sore throat are particularly common in enteroviral meningitis. Myalgia is particularly severe with coxsackie B infections. As in acute bacterial meningitis, infants usually present with vague irritability and a tense fontanelle, and young children with fever and irritability or lethargy. Conjunctival injection, pharyngitis, and cervical lymphadenopathy may be found. Macular or petechial exanthems or enanthems are seen with coxsackie A and B and echovirus infections (especially echovirus 9). Vesicles on the hands, feet, and mouth have been reported with coxsackie A16 and enterovirus 71 infections. By definition, the level of consciousness is normal in simple meningitis. Neurological features include vertigo, nystagmus, cerebellar ataxia, facial spasms, and involuntary movements.

The specific cause of viral meningitis may be suggested by characteristic signs outside the nervous system, such as genital or rectal vesicles in the sexually active age group (HSV-2), herpes zoster skin lesions, swelling in the parotid region (mumps, and occasionally coxsackie, lymphocytic choriomeningitis, and Epstein–Barr viruses), orchitis (mumps and lymphocytic choriomeningitis virus), and arthritis (lymphocytic choriomeningitis virus). However, potentially helpful features, such as gastrointestinal symptoms associated with enteroviral infections and parotitis associated with mumps, may be completely absent in patients with meningitis.

Paralytic poliomyelitis

The infection (see Chapter 7.10.7) is acquired by droplet spread from the respiratory tract or by the faecal–oral route. The 'minor illness', coinciding with viraemia, is a non-specific episode of influenza-like symptoms—fever, headache, sore throat, malaise, and mild gastrointestinal symptoms—which resolves in a few days. Most of those infected have no further symptoms but, in a minority, the 'major illness' follows, sometimes after a few days' remission of symptoms. The features are those of viral meningitis: muscle pain, spasms, and sensory disturbances may precede or accompany the development of lower motor neurone (flaccid) paralysis. Any combination of motor unit deficits may be seen (Fig. 1). It is most unusual for paralysis to extend after the first 3 days or after the temperature has fallen (Fig. 2). Respiratory and bulbar paralysis is life-threatening. Encephalitis is rare. The commonest causes of death are aspiration and airway obstruction, resulting from bulbar paralysis, and paralysis of respiratory muscles. Disturbances of respiratory and cardiac rhythm, thought to be the result of damage to medullary vasomotor and respiratory centres, are extremely uncommon. Other complications include impaired control of body temperature and blood pressure, gastrointestinal haemorrhage, aspiration pneumonia, and paralysis of the bladder and bowel.

Fig. 1 Paralytic poliomyelitis in a 3-year-old Thai child. Note systemic illness and paralysis of right arm. (Copyright D.A. Warrell.)

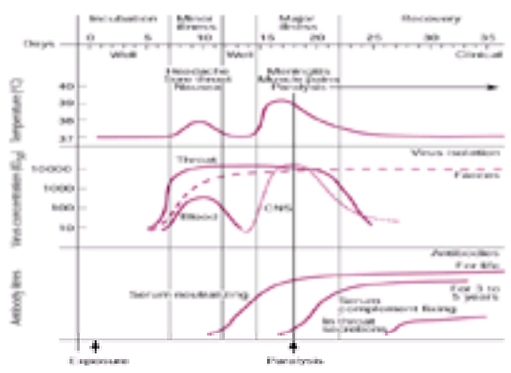

Fig. 2 The course of a paralytic poliomyelitis infection. (Reproduced from Christie (1981). *Medicine International*, **1**, 139. Adapted from Bodian, D. (1957). Mechanisms of infection with polio viruses. In *Cellular biology, nucleic acids, and viruses*. New York Academy of Sciences, by permission.)

Encephalitis

Most patients with viral encephalitis present with the symptoms of meningitis (fever, headache, neck stiffness, vomiting) followed by altered consciousness, convulsions, and sometimes focal neurological signs, signs of raised intracranial pressure, or psychiatric symptoms.

Herpes simplex encephalitis

This relatively common sporadic encephalitis may occur in any age group. In neonates, it is caused by HSV-2.

As well as the usual clinical features of a severe viral encephalitis, patients with herpes simplex encephalitis have symptoms related to the focal nature of the encephalitis (frontal and temporal cortex and limbic system). These include behavioural abnormalities, olfactory and gustatory hallucinations, anosmia, amnesia, expressive aphasia, and temporal lobe seizures. Herpetic skin or mucosal lesions are rarely found, except in the case of acute genital HSV-2 infection, or proctitis, and a past history of 'cold sores' does not affect the chances of the infection being due to herpes simplex virus. Effects of cerebral oedema are unusually severe. Patients usually lapse into coma towards the end of the first week and most deaths occur within the first 2 weeks. This condition is also discussed in Chapter 7.10.2.

Japanese encephalitis

After an incubation period of 7 to 14 days, patients develop non-specific prodromal symptoms (fever, headache, malaise, and nausea) lasting 2 to 3 days. Neurological symptoms begin suddenly, with increasing headache, deteriorating level of consciousness, and generalized convulsions, which may result in status epilepticus. There is meningism and mask-like facies, with upper motor neurone signs or a myelitic pattern, cranial nerve lesions (for example, lower motor neurone VII), flaccid paralysis, ataxia, involuntary movements (Fig. 3), and, in severe cases, prolonged coma, hemi- or quadriparesis, decerebrate rigidity, and respiratory failure. Fever persists for 6 to 7 days and, in survivors, neurological symptoms may persist for several weeks. Parkinsonian and extrapyramidal features occur frequently and choreoathetoid movement disorders or severe dystonias can last for many months. The case fatality rate is 30 per cent in those admitted to hospital. Most deaths occur in the first 7 to 10 days from respiratory failure, aspiration pneumonias, intracranial hypertension, and uncontrolled seizures. Up to 50 per cent of survivors suffer from intellectual impairment, psychiatric problems, persistent epilepsy, or a vegetative state with spastic quadriparesis and evidence of basal ganglia involvement, such as dystonia of the limbs and trunk, rigidity, and tremor (Fig. 3).

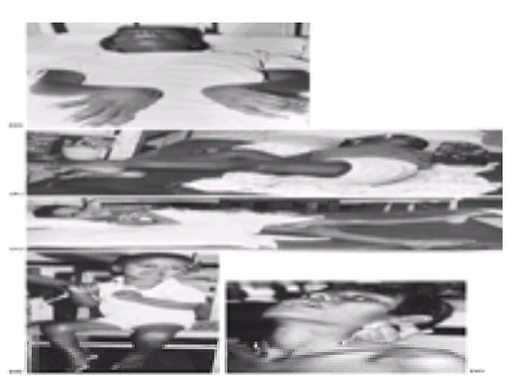

Fig. 3 Japanese encephalitis in Anuradhapura, Sri Lanka. (a) Comatose female patient showing symmetrical chorioathetotic movements of the upper limbs. (b) Comatose child showing dystonic movements of the upper and lower limbs. (c) Convalescent child, conscious but with residual dystonia of all four limbs. (d) Convalescent child with floppy head and involuntary movements of all four limbs. (e) Convalescent boy with residual weakness of the neck flexors. (All figures by courtesy of Dr D.T.D.J. Abeysekera.)

Nipah virus encephalitis

The main clinical features of Nipah virus encephalitis are fever, headache, dizziness, reduced consciousness, and prominent brainstem dysfunction. Distinctive signs included myoclonus, areflexia, hypotonia, hypertension, and tachycardia, suggesting extensive brainstem and spinal cord involvement. MRI imaging during the acute illness shows widespread focal lesions in subcortical and deep white matter and, to a lesser extent, in grey matter on T_2-weighted sequences (Fig. 4).

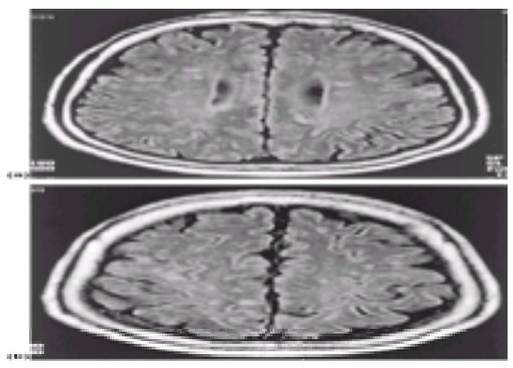

Fig. 4 (a) and (b) Acute Nipah virus encephalitis. MR brain images in a 57-year-old pig farmer showing multiple focal lesions in the grey–white matter junction. These are areas of infarction secondary to vasculitis. (By courtesy of Drs B.J. Abdullah and Sazilah Sarji, Kuala Lumpur.)

West Nile virus encephalitis

The most common clinical features are encephalitis, meningitis, fever, weakness, and headache following an incubation period of 3 to 15 days. Infection usually results in an acute febrile episode with no CNS involvement. In unusual cases or, as in the United States, when the virus is introduced into a naïve population, the incidence of encephalitis rises particularly in the elderly. An erythematous rash of the neck, trunk, and limbs is present in 20 per cent of cases. Patients over 50 years of age were most at risk of developing encephalitis during the New York outbreak, but all age groups are affected in endemic areas. Muscle weakness, areflexia, and diffuse flaccid paralysis in association with an axonal polyneuropathy was also reported. MRI imaging of the brain demonstrated enhancement of the meninges and periventricular areas. There is no specific treatment.

Enterovirus 71

As the goal of poliomyelitis eradication appears more achievable, another enterovirus is emerging as a significant cause of acute neurological disease in Asia. Enterovirus 71 (**EV71**) was first recognized in 1969 and is responsible for a variety of clinical manifestations, including: hand, foot, and mouth disease; aseptic meningitis; meningoencephalitis; and acute flaccid paralysis. In an outbreak of hand, foot, and mouth disease in Malaysia; a number of young children developed fatal encephalomyelitis, dying within a few hours of presentation with cardiovascular instability and severe pulmonary oedema. Postmortem examination in four cases revealed major involvement of the brainstem and spinal cord, with EV71 being isolated from brain tissue in all cases; there was no apparent cardiac pathology and virus was not isolated from the myocardium. Molecular characterization of these four viruses and others isolated concurrently suggest that at least two potentially virulent EV71 strains were circulating during the outbreak. An adenovirus was also thought to have complicated the infection in 60 per cent of the children dying with a similar clinical picture. It is possible that co-infection with the two viruses may have resulted in severe disease.

Postinfectious encephalomyelitis

Sudden convulsions, coma, fever, or pareses appear 10 to 14 days after the start of vaccination (vaccinia or nervous tissue rabies vaccine) or after infection with measles, varicella, rubella, mumps, or influenza. In the case of measles, varicella, and rubella, encephalitic symptoms develop 2 to 12 days after the rash has appeared, and in mumps before or after parotid swelling. Involuntary movements, cranial nerve lesions (VII, III), pupillary abnormalities, nystagmus, ataxia, and upper motor neurone signs are common

Diagnosis

Clinical and epidemiological details

The time of year, known current epidemics, the patient's age, occupation, animal contacts, and countries or states visited recently may help to narrow down the possibilities. A specific diagnosis may be suggested by distinctive clinical features of the encephalitis itself (for example, hydrophobia in rabies, temporal lobe features in herpes simplex encephalitis) or of the associated infection (for example, mumps parotitis; measles rash; skin and mucosal lesions of herpes viruses; and gastrointestinal symptoms associated with enteroviral infections).

Laboratory investigations

These should aim to demonstrate a specific viral agent (particularly important for the potentially treatable herpesvirus infections) or exclude potentially treatable non-viral causes of meningitis or encephalomyelitis (Table 1). The most important investigation is examination of the cerebrospinal fluid (**CSF**). Contraindications to lumbar puncture are the same as for acute bacterial meningitis (Chapter 24.7 and Chapter 24.14.1). If there are lateralizing neurological signs or evidence of raised intracranial pressure, a computed tomography (**CT**) or magnetic resonance imaging (**MRI**) scan should be performed to exclude an intracranial mass lesion before contemplating a lumbar puncture. CSF pressure is especially increased in herpes simplex encephalitis, where there is intense cerebral oedema. Pleocytosis ranges from tens to thousands of cells/µl. Lymphocytes and other mononuclear cells predominate, except in the early stages of some infections (for example, enteroviruses, herpes simplex encephalitis). The CSF contains erythrocytes or is xanthochromic in haemorrhagic encephalitides such as herpes simplex encephalitis and acute necrotic leucoencephalitis. Protein concentration is usually increased in the range of 50 to 150 mg/dl with an increasing proportion of IgG as the disease progresses. Leakage of serum IgG into the CSF and intrathecal IgG synthesis, indicated by a monoclonal band, are responsible. CSF glucose concentration is usually normal or increased towards the level in a blood sample taken simultaneously, but low levels are occasionally reported, especially in mumps and lymphocytic choriomeningitis virus infections. Measurement of lactate, C-reactive protein, lactic dehydrogenase, creatine kinase (CK-BB), muramidase, and various cytokines in the CSF have not proved helpful in distinguishing viral from other infections. CSF examination may be misleading if it is normal: as it is at the first examination in 10 to 15 per cent of patients with herpes simplex encephalitis; if there is a predominantly neutrophil pleocytosis; or if the glucose concentration is low. Myelin basic protein may be found in the CSF of patients with postinfectious encephalomyelitides.

Virology

Full laboratory resources allow a specific virus to be implicated in 70 to 75 per cent of cases of lymphocytic meningitis and in 30 to 40 per cent of patients with meningoencephalitis (Table 2). At appropriate stages of the illness, a rapid diagnosis by direct immunofluorescence may be made of herpes simplex virus (skin and brain), herpes zoster virus (skin lesion scrapings), rabies (skin sections and brain), measles (nasopharyngeal aspirate), and some non-viral causes such as Rocky Mountain spotted fever (skin). Electron microscopy of skin lesions will identify a herpesvirus. Some viruses can be isolated from the CSF (for example, mumps, enteroviruses, lymphocytic choriomeningitis virus, Central European encephalitides, Louping ill, and HIV). Virus cultured from a distant site may help with the diagnosis (for example, polio and other enteroviruses from stool, or arthropod-borne viruses from blood culture), but they may not be related to the neurological symptoms (for example, cytomegalovirus from the pharynx or urine, herpes simplex virus from skin or mucosa or adenovirus seen in stool by electron microscopy). Specific viral IgM can be detected in serum for mumps, Epstein–Barr virus, cytomegalovirus, or measles, or, using a µ-capture technique, in the CSF for JE virus. This method is being used increasingly to detect IgM to other viruses. The viraemia associated with JE is very brief and isolation from CSF difficult. Virus can occasionally be isolated from postmortem material. A viral diagnosis is often delayed until a rising convalescent antibody titre is found by an appropriate technique. This is usually the case for mumps, coxsackieviruses, and most arthropod-borne viruses.

An important diagnostic advance has been the introduction of polymerase chain reaction (**PCR**) technology for the routine diagnosis of a viral infection of the CNS. PCR greatly amplifies the amount of viral nucleic acid in the test sample, enabling the identification of herpes simplex virus in the CSF of suspected cases of herpes simplex encephalitis within a short time of the onset of symptoms. PCR is now the investigation of choice for the rapid diagnosis of HSV encephalitis, having a sensitivity of 95 per cent and a specificity of 100 per cent. The application of microchip and real-time PCR technology may further aid the rapid diagnosis of encephalitis. It is hoped that molecular techniques may aid the early diagnosis of a greater variety of CNS viral infections in the future (Table 2).

Brain biopsy

For the rapid diagnosis of viral encephalitides such as progressive multifocal leucoencephalopathy there is still no substitute for brain biopsy, but few would regard this inherently risky procedure as being justified. Electroencephalography, CT or MRI scans, angiography, or technetium scans can help to direct the surgeon towards the affected area of brain.

Imaging of the brain and spinal cord

CT and MRI scans of the brain and spinal cord can be extremely useful for the diagnosis of the site, nature, and extent of mass lesions and associated oedema, sub- and epidural empyemas, meningitis, cerebritis, and ventriculitis, the presence of intracranial hypertension, hydrocephalus, cerebral and brainstem herniation, demyelination, and other anatomical abnormalities (see Chapter 24.5).

CT scans are superior for bony details and calcifications, are quicker to perform, and are less dependent on the patient being able to lie motionless for prolonged periods. The resolution of MRI is greater for parenchymal lesions, but MRI cannot be performed in patients with pacemakers and may be dangerous in those with metal clips in cerebral blood vessels. Some viral encephalitides do have characteristic lesions on MRI. Some 94 per cent of patients with HSV have high-signal T_2-hyperintense lesions in the medial and inferior temporal regions, and JE is associated with characteristic lesions in the basal ganglia. More discrete high-signal intensity 2- to 7-mm lesions, particularly in the subcortical and deep white matter of the cerebral hemispheres, have been associated with the recently described Nipah virus infection (Fig. 4). However, these classical descriptions often overlap and the general features of oedema, infarction, and high signal on the T_2-weighted images are commonly seen in a variety of viral infections of the CNS.

Differential diagnosis

Viral infections of the CNS must be distinguished from the many other conditions that produce similar clinical features and CSF abnormalities (Table 1). The differential diagnosis of viral meningitis includes the other causes of aseptic meningitis, such as partially treated bacterial meningitis, tuberculous meningitis, spirochaetal infections (leptospirosis, borreliosis, Lyme disease, and syphilis), fungal, amoebic, neoplastic, granulomatous, and idiopathic meningitides. Viral myelitides must be distinguished from other causes of transverse myelitis and the Brown–Séquard syndrome. These include spinal compression by tumours, abscesses, helminths or their ova, or vertebral disease.

The differential diagnosis of paralytic poliomyelitis includes: postinfectious and other immunopathic polyneuroradiculopathies, such as Guillain-Barré syndrome and Landry's ascending paralysis; metabolic neuropathies such as acute porphyria; paralytic rabies; neoplastic polyradiculopathies; and rarities, such as tick paralysis and *Herpes simiae* B virus infection. The lack of objective sensory loss in poliomyelitis usually distinguishes it from these other entities.

The differential diagnosis of viral encephalitis includes other infective encephalopathies: bacterial, fungal, protozoal, and parasitic; intracranial abscesses and neoplasms; toxic and metabolic encephalopathies; and heat stroke. The diagnosis of 'viral encephalitis' should not be made too hastily, as it may condemn the patient with concealed cerebral malaria or some other curable encephalopathy to delayed treatment or even death.

Treatment

Antiviral chemotherapy

Aciclovir and, to a lesser extent, vidarabine (cytosine arabinoside) have proved effective in treating herpes simplex encephalitis. This subject is also discussed in Chapter 7.10.2. The nucleoside analogue acycloguanosine (that is, aciclovir) is only taken up by cells infected by herpes simplex virus and is therefore non-toxic to normal, uninfected cells. In view of this remarkable lack of serious toxicity, treatment can be started as soon as herpes simplex encephalitis is suspected clinically. Although there is still some controversy in the United States regarding the role of brain biopsy in herpes simplex encephalitis, virtually all clinicians in the United Kingdom now start therapy with aciclovir immediately on suspicion of encephalitis without attempting to confirm the diagnosis by brain biopsy. Aciclovir has also been used to treat herpes zoster virus encephalitis, but there is no convincing evidence for its efficacy in cytomegalovirus infections of the CNS. The rare, but very dangerous, encephalomyelitis caused by *Herpes simiae* B virus should be treated with aciclovir. Ribavirin is effective against some RNA viruses, such as those causing Lassa fever, haemorrhagic fever with renal syndrome, and possibly Argentine haemorrhagic fever, Rift Valley fever, and Congo Crimean haemorrhagic fever.

Interferons have been used by intravenous, intrathecal, or intraventricular routes in the treatment of rabies, herpes zoster virus, and other herpesvirus encephalitides, but have not proved effective. Although initial pilot studies of the use of interferon in JE were encouraging, the efficacy of this expensive treatment is awaited from randomized controlled trials.

Hyperimmune plasma given within 8 days of the start of symptoms has reduced the mortality of Argentine haemorrhagic fever (Junin virus) from between 20 and 30 to 1 and 3 per cent. Hyperimmune human globulin has also proved effective in the treatment of Congo Crimean haemorrhagic fever.

Supportive treatment

Corticosteroids have been used in the treatment of most of the viral encephalomyelitides, both in an attempt to combat cerebral oedema (especially in herpes simplex encephalitis) and for their other anti-inflammatory effects. Convincing evidence of benefit from controlled trials is lacking, but the immunosuppressive effects of corticosteroids have not led to obvious clinical deterioration, except perhaps in some cases of diffuse myelitis. Corticosteroids or ACTH have also been used for postinfectious and postvaccinal encephalomyelitides, but the evidence for their efficacy is not convincing and, since they may exacerbate latent rabies in experimental animals, should be used only in life-threatening cases of rabies postvaccinal encephalomyelitis. Severe intracranial hypertension should be treated with intravenous mannitol or mechanical hyperventilation. Nursing and general care are the same as for acute bacterial meningitis (Chapter 24.14.1) and tuberculous meningitis. Seizures must be controlled with phenytoin or phenobarbital, fever lowered by cooling, respiratory failure treated by mechanical ventilation, and attention given to fluid, electrolyte, and acid–base balance. Hyponatraemia is attributable to inappropriate secretion of antidiuretic hormone in some cases.

Paralytic poliomyelitis

Most authorities recommend rest and even mild sedation during the preparalytic stage of the 'major illness', because of the suspicion that exercise increases paralysis. The severe muscle pains and spasms reported in patients in some parts of the world are treated with mild analgesics, such as salicylate, and with hot-water bottles. During the phase of developing paralysis, patients must be observed closely and, if possible, assessed objectively for the development of life-threatening bulbar and respiratory paralysis. Those with weakness of swallowing should be nursed on their sides to prevent aspiration. The need for a cuffed tracheostomy tube may be avoided by careful positioning, frequent observations, and suction. Indications for mechanical ventilation are a progressive decline in ventilatory capacity to less than 30 to 50 per cent of normal, hypoxaemia, or gross disturbances of respiratory rhythm (Cheyne–Stokes respiration, long apnoeic intervals, etc.) suggesting damage to the respiratory centres. Respiratory weakness without bulbar paralysis may be treated in a tank respirator or rocking bed, which do not require tracheostomy. However, patients with severe or rapidly progressing respiratory paralysis need urgent tracheostomy and intermittent positive-pressure ventilation. Overventilation must be avoided. Assisted ventilation may be required for long periods. In the Copenhagen epidemic of the 1940s, this was achieved by manual ventilation. Attempts should be made to wean patients off the ventilator as soon as their condition becomes stable. Severe fluctuations in body temperature and blood pressure, reminiscent of those in severe tetanus and rabies, may require intensive care. The paralysed patient may have to lie in bed for many months and will develop complications from this prolonged immobilization. These include bed sores, osteomalacia, hypercalciuria leading to renal calculi, recurrent urinary tract infections resulting from chronic urethral catheterization, respiratory infections, and contractures of muscles and tendons leading to severe musculoskeletal deformities that will require orthopaedic correction. Some of these can be prevented by passive movement of the joints and splinting. Physiotherapy and psychological support are needed during the prolonged phase of rehabilitation.

Prognosis and sequelae

Viral meningitis has an excellent prognosis, but some patients with HSV-2 infection have recurrent attacks with spinal cord or nerve root involvement. Case fatality rates of some viral encephalomyelitides are as follows: rabies, 100 per cent; herpes simplex encephalitis (untreated), 40 to more than 75 per cent (highest in neonates

and those over 30 years old); eastern equine encephalitis, 50 per cent; Japanese encephalitis, 10 to 40 per cent; measles, 10 to 20 per cent; varicella, 10 to 30 per cent; western equine encephalitis, 8 per cent; St Louis encephalitis, 3 per cent; California encephalitis, Venezuelan encephalitis, and mumps, less than 1 per cent. The mortality of paralytic poliomyelitis increases from 5 per cent in young children to more than 20 per cent in adults. Postinfectious and postvaccinal encephalomyelitides carry case fatalities of 15 to 40 per cent.

Neurological sequelae are found in 5 to 75 per cent of survivors of Japanese encephalitis and herpes simplex encephalitis, and are especially common in infants. They include mental retardation, loss of memory, speech abnormalities (including subtle expressive aphasias), hemiparesis, ataxia, dystonic brainstem and cranial nerve lesions, recurrent convulsions, and various behavioural and personality disturbances. Sequelae are common with postinfectious encephalomyelitis. An unusual sequel to paralytic poliomyelitis developing after an interval of many years is a condition characterized by progressive muscle weakness and wasting, attributable to depletion of anterior horn cells, which has some similarities to motor neurone disease.

Prevention

Prophylactic vaccination against poliomyelitis and measles has virtually eradicated encephalitides caused by these viruses in many communities. Postexposure rabies vaccination has also proved effective in preventing rabies encephalitis, and tissue-culture rabies vaccines are used increasingly for pre-exposure prophylaxis. A formalin-inactivated, adult mouse-brain vaccine is manufactured in Osaka for JE. It is effective and carries a very low risk of objective neurological complications (one in a million courses). An alternative live-attenuated vaccine has been developed in China, and has been shown to be both safe and effective in over 100 million Chinese children. Promising future vaccine candidates are currently being evaluated in non-human primate models, including a chimeric live-attenuated JEV/yellow fever virus combination and two poxvirus-vectored recombinant JE vaccines. Travellers to endemic regions should be vaccinated.

Since the outbreak of West Nile infection in the United States, several vaccine candidates have already been identified and immune protection against infection demonstrated in several animal models: human trials have been planned. There have been no reports of such success against Nipah virus. Vaccines for use in humans have been prepared against a number of other arthropod-borne viruses (for example, European tick-borne encephalitis).

Hyperimmune immunoglobulin has been used for prophylaxis (and in some cases attempted treatment) of measles, herpes zoster virus, HSV-2, vaccinia, rabies, and some other infections in high-risk groups. Immunocompromised patients, such as those with leukaemia, who are household contacts of a case of herpes zoster virus infection, should be given prophylactic hyperimmune globulin and, if they develop skin lesions, they should be treated with aciclovir to prevent the development of severe disease.

Interferons have been used with some success to prevent herpesvirus infections, for example cytomegalovirus in high-risk groups such as renal transplant recipients. However, the evidence does not yet justify their recommendation.

Caesarean section before rupture of the membranes in a full-term pregnant woman with genital herpes may prevent HSV-2 encephalitis in the neonate. If the herpetic lesions are discovered during or after vaginal delivery, topical aciclovir should be applied to the eyes of the neonate, as they are the most likely portal of entry.

Arthropod-borne viral encephalitides can be prevented by avoiding or controlling the arthropod vectors (for example, by the use of mosquito nets, insect repellents, insecticides, etc.), by attempting to control the numbers of wild vertebrate reservoir species, or by immunizing domestic animals, such as horses (eastern and western equine encephalitides) and pigs (JE). To control rabies, the principal wild mammalian vectors can be immunized (for example, wild foxes, racoons, and black-backed jackals have been immunized by distributing oral vaccine in bait). Domestic dogs and cats should be vaccinated. To prevent the viral encephalitides transmissible from laboratory animals (for example, lymphocytic choriomeningitis from mice and rats, *Herpes simiae* B from monkeys) their screening, quarantine, handling, and housing should be strictly controlled.

Reye's syndrome

Reye's syndrome is an acute encephalopathy affecting children between the ages of 2 and 16 years. It is rapidly fatal in 10 to 40 per cent of cases. The defining characteristics are sudden impairment of consciousness, increase in serum aminotransferase concentrations (or, if a biopsy is done, a fatty liver), and the exclusion of other diseases. Symptoms develop a few days after varicella or an upper respiratory tract or gastrointestinal illness. Clusters of cases (median age 11 years) have been associated with influenza B epidemics, while sporadic cases (median age 6 years) have followed varicella, coxsackie, dengue, and other viral infections. Studies in the United States have demonstrated an association between Reye's syndrome and the use of salicylates, but not of paracetamol, during the preceding viral illness. This has led the United Kingdom Committee on Safety of Medicines to recommend that aspirin should not be given to children under 12 years of age, unless specifically indicated for childhood rheumatic conditions. Aflatoxin has been implicated in Thailand. In the United States, the annual incidence of Reye's syndrome in those under 18 years old is 0.42 per 100 000 urban dwellers and 1.8 per 100 000 rural and suburban dwellers.

The child is nauseated and retches or vomits for 1 or 2 days before becoming confused or comatose and requiring admission to hospital. Most are afebrile and have hepatosplenomegaly but no jaundice at presentation. Fever develops later. The CSF is usually normal or contains a few mononuclear cells. Irritability, extreme agitation, aggression, and delirium are succeeded by coma and death in 2 to 3 days. Decorticate and decerebrate posturing and convulsions may be partly attributable to hypoglycaemia, which occurs in the majority of cases. There is rapid neurological deterioration with loss of pupillary and oculovestibular reflexes, evidence of increased intracranial pressure, deepening coma, and death. Neurological sequelae are common in survivors. Blood ammonia is increased above the normal limit of 48 µg/dl in almost all cases. The characteristic histological abnormality is fatty droplets in the liver cells. Mitochondrial abnormalities, but no inflammatory changes, have also been seen in neurones and hepatocytes.

The differential diagnosis includes acute hepatic encephalopathy, especially associated with poisoning, infective encephalopathies such as cerebral malaria (usually distinguishable by a positive blood smear) or bacterial, viral, and fungal meningoencephalitides (distinguished by characteristic CSF abnormalities).

There is no specific treatment, but mortality can be reduced by treating hypoglycaemia, cerebral oedema, respiratory failure, fluid and electrolyte disturbances, and other complications. These measures are also considered in Chapter 24.14.1.

Other viral infections or disorders in which viruses play a role in the pathogenesis of neurological disease

Subacute sclerosing panencephalitis

This disorder (see also Chapter 7.10.6) is a form of subacute encephalitis affecting children and young adults due to persistent infection with the measles virus. The cumbersome title, usually abbreviated to **SSPE**, is derived from the conditions formerly known as subacute sclerosing leucoencephalitis and inclusion-body encephalitis, now known to be the same disease.

Aetiology

An infective cause was long suspected and there is now conclusive evidence to incriminate the measles virus. Measles virus antibody titres are extremely high in the blood and CSF, measles antigen has been demonstrated in the brain, and the virus has sometimes been isolated, but only with difficulty. Most affected children have had measles at an unusually early age and there is a mean interval of some 6 years between infection and the onset of encephalitis. The disease can occur in children vaccinated with live measles virus, but the risk is much lower than that following the natural disease.

The measles virus in subacute sclerosing panencephalitis appears to be incomplete, as the matrix (**M**) protein required to attach the nucleocapsid to the cytoplasmic membrane prior to budding is deficient or absent. It is not known whether the absence of M protein from the brain is the result of an abnormality of the virus or of the host, and, if the latter, whether inborn or acquired. Current thought is that during the long symptom-free interval between infection and appearance of disease, viral material accumulates, eventually leading to cell damage. The paradox of high antimeasles antibodies, except against M protein, and persistent virus has not been explained. The comparatively early age of clinical measles in affected children, often below the age of 2 years, suggests that the immature immune system permits entry and persistence of the virus in the brain.

Pathology

As its name implies, both grey and white matter show the changes of encephalitis, with perivascular cuffing and more diffuse cellular infiltration, neuronal loss and myelin destruction, with variable glial scarring or sclerosis. Acidophilic nuclear inclusion bodies are never profuse and may not be detected. No visceral lesions are found.

Clinical features

In the great majority, the onset is in the first two decades, but young adults may also be affected. The disease is twice as common in boys as in girls. Incidence has fallen sharply in countries where measles vaccination is at a high level; the annual incidence in England and Wales has fallen from 20 to around 5. Subacute sclerosing panencephalitis remains relatively common in parts of eastern Europe, Egypt, and the Lebanon. No convincing predisposing factors have been identified and, in particular, immunosuppressed children are not at special risk but they may occasionally develop acute measles inclusion-body encephalitis.

The speed of onset is extremely variable, but there is usually a prolonged period of altered behaviour, mild intellectual deterioration, and loss of energy and interest, often misinterpreted as sloth or neurosis. After some weeks or months increasing clumsiness or the appearance of focal neurological symptoms draws attention to the organic nature of the disease. Periodic involuntary movements then appear, the commonest form being myoclonus, consisting of a stereotyped jerk or lapse of posture involving the limbs, often asymmetrically, occurring every 3 to 6 s. The myoclonus may result in sudden falls, which are occasionally the presenting symptom. Visual signs may be prominent, with papilloedema, retinitis, optic atrophy, or cortical blindness. Choroidoretinal scarring is present in 30 per cent of cases. In other cases the onset is relatively abrupt with no recognizable prodromal stage. There is no fever or other evidence of systemic infection.

Further progression is marked by intellectual deterioration, rigidity and spasticity, and increasing helplessness. Some 40 per cent of patients die within a year, but a similar proportion survive for more than 2 years. A period of apparent arrest is common and in some patients, particularly at the upper end of the age range, substantial remission and prolonged survival occur. Even in such cases there may be radiological evidence of continued cerebral damage and it is probable that the disease is always eventually fatal.

Investigation

There is no significant pleocytosis in the cerebrospinal fluid and total protein is not increased, but there is evidence of intrathecal synthesis of immunoglobulin and oligoclonal bands of IgG. Although the measles antibody titres in blood and CSF are usually raised to high levels, occasionally they overlap control values. In established disease, the electroencephalogram (**EEG**) shows highly characteristic periodic discharges, synchronous with the myoclonus, but persisting in the absence of the movements. The CT scan shows low-density, white-matter lesions and cerebral atrophy.

Treatment

There is no effective treatment for subacute sclerosing panencephalitis. Inosiplex, 100 mg/kg daily by mouth in divided doses, possibly prolongs survival, particularly in older patients with disease of slow onset, but adequately controlled trials are naturally difficult to mount. Interferon given by intraventricular catheter has been reported to induce partial remission.

Progressive multifocal leucoencephalopathy (see also Chapter 24.14.4)

This disease is caused by opportunistic infection by papovaviruses, most commonly JC virus and the simian virus SV40. A high proportion of normal adults have antibodies to the former and the agent appears to be ubiquitous. The reservoir of SV40 is in monkeys and the agent was apparently transmitted in early types of poliomyelitis vaccine, without evident ill-effects. These viruses are potentially oncogenic, but are non-pathogenic for humans unless the immune system has been compromised.

Progressive multifocal leucoencephalopathy thus occurs in patients already affected by such conditions as lympho- or myeloproliferative diseases, sarcoidosis, and other chronic granulomatous diseases, or, more recently, AIDS, and also in those therapeutically immunosuppressed. Most patients are over 50 years old but, with the spread of AIDS, younger people are being affected, with a male preponderance, and the disease is no longer rare.

Pathology

The virus particularly invades the nuclei of the oligodendroglia and, as a result, there is demyelination of the white matter of the cerebral hemisphere, spreading from numerous foci. The cerebellum and brainstem are less often involved and the spinal cord is spared. Abnormal giant forms of oligodendrocytes with eosinophilic inclusions are seen microscopically, and arrays of intranuclear virus particles can often be identified by electron microscopy. JC virus antigen can be identified by immunofluorescence or immunohistochemistry. DNA probing has revealed unintegrated virus in oligodendrocytes, astrocytes, endothelial cells, and in extraneural organs such as kidney, liver, lung, spleen, and lymph nodes.

Clinical features

The onset is usually with progressive signs of a focal lesion of one cerebral hemisphere; limb weakness, aphasia, or visual field defect such as homonymous hemianopia. More widespread signs gradually develop, leading to personality changes, intellectual deterioration, dysarthria or fluent aphasia, and bilateral weakness. Fits are rare. There is no systemic evidence of infection. Spontaneous temporary arrest or partial remission are common but eventual progression causes death in 6 to 12 months, although much more chronic cases are on record, with survival, exceptionally, to 5 years.

Investigation

The CSF is normal apart from occasionally a mild elevation of protein and slight pleocytosis, and is not under increased pressure. The EEG shows a bilateral excess of slow activity. The CT scan may at first show little abnormality, but eventually large, non-enhancing, low-density lesions appear in the cerebral white matter. MRI is more sensitive. Serum antibodies are of no diagnostic help but the response in the CSF has not been fully evaluated. The diagnosis can be confirmed only by cerebral biopsy, but it is essential that white matter is included in the specimen. This may be important to distinguish lymphoma and, rarely, herpes simplex encephalitis involving white matter.

Treatment

No treatment is of proven value, but cytosine arabinoside has sometimes appeared to induce partial remission.

Progressive rubella panencephalitis

This extremely rare disorder (see also Chapter 7.10.12) may follow congenital rubella or rubella in early childhood. It evolves insidiously some 10 years after the original illness and is characterized by progressive mental retardation with behaviour changes, fits, ataxia, spasticity, optic atrophy, and macular degeneration. Pathological changes are those of encephalitis with perivascular infiltration. The CSF may show a slight rise in white cell and protein content, elevation of gammaglobulin and of antirubella antibodies to an extent greater than the rise in the serum level, suggesting local production of antibody within the CNS. The EEG may show changes similar to those seen in subacute sclerosing panencephalitis due to measles virus. The mechanism responsible for the appearance of this disorder is unknown and there is no effective treatment.

Vogt–Koyanagi–Harada syndrome

The cause of this rare syndrome is thought to be an inflammatory autoimmune reaction to an unidentified viral infection. The disorder affects tissues having a common embryological origin, the uvea and leptomeninges and the melanoblasts, ocular pigments and auditory labyrinth pigments originating from the neural crest. The

dermatological features consist of patchy whitening of eyelashes, eyebrows, and scalp hair, alopecia, and vitiligo. Neurological manifestations include meningoencephalitis, raised intracranial pressure, neurosensory deafness, tinnitus, nystagmus, ataxia, ocular palsies, and focal cerebral deficits. Ocular features are those of uveitis with pain and photophobia, more generalized inflammation of the eye, retinopathy, and impaired visual acuity. The condition tends to be self-limiting but may result in serious permanent ocular and neurological deficits. Steroids and immunosuppressive drugs have been used and are said to arrest the progression of at least some features of the disorder.

Viral causes of psychiatric illness

Mental changes are common in patients with encephalitis. Influenza, infectious mononucleosis, and infectious hepatitis are sometimes followed by psychiatric sequelae, in particular a depressive reaction. Psychosis following encephalitis lethargica has been reported on occasions.

Other possible virus infections in which the nervous system is involved

Acute disseminated encephalomyelitis is considered in Chapter 24.16. Reye's syndrome is discussed above and Behçet's syndrome in Chapter 18.10.5. Mollaret's meningitis is discussed in Chapter 24.14.1.

*Contains some material contributed by PGE Kennedy to OTM3.

Further reading

Boos J, Esiri MM (1986). *Viral encephalitis: pathology, diagnosis and management*. Blackwell Scientific, Oxford.

Cardosa MJ, *et al.* (1999). Isolation of subgenus B adenovirus during a fatal outbreak of enterovirus 71-associated hand, foot, and mouth disease in Sibu, Sarawak. *Lancet* **354**, 987–91.

Christie AB (1980). *Infectious diseases: epidemiology and clinical practice*, 3rd edn. Churchill Livingstone, Edinburgh.

Goh KJ, *et al.* (2000). Clinical features of Nipah virus encephalitis among pig farmers in Malaysia. *New England Journal of Medicine* **342**, 1229–35.

Griffiths JF (1985). SSPE and lymphocytes. *New England Journal of Medicine* **313**, 952–3.

Jackson AC, Johnson RT (1989). Aseptic meningitis and acute viral encephalitis. In: Vinken PJ, *et al.*, eds. *Handbook of clinical neurology*, Vol. 12, ch. 56, *Viral diseases*, pp. 125–48. Elsevier, Amsterdam.

Johnson RT, *et al.* (1985). Japanese encephalitis: immunocytochemical studies of viral antigen and inflammatory cells in fatal cases. *Annals of Neurology* **18**, 567–73.

Krupp LB, *et al.* (1985). Progressively multifocal leukoencephalopathy: clinical and radiological features. *Annals of Neurology* **17**, 344–9.

Nash D, *et al*. (2001). The outbreak of West Nile virus infection in the New York City area in 1999. *New England Journal of Medicine* **344**, 1807–14.

Pattison EM (1965). Uveomeningoencephalitic syndrome (Vogt–Koyanagi–Harada). *Archives of Neurology* **12**, 197–205.

Price RW, Plum F (1978). Poliomyelitis. In: Vinken PJ, *et al.*, eds. *Handbook of clinical neurology*, Vol. 34, *Infections of the nervous system*, pp 93–132. North Holland, Amsterdam.

Scheld WM, Whitley RJ, Durack DT, eds. (1997). *Infections of the central nervous system*, 2nd edn. Lippincott-Raven, New York.

Solomon T, *et al.* (1988). Poliomyelitis-like illness due to Japanese encephalitis virus. *Lancet* **351**, 1094–8.

Solomon T, *et al.* (2000). Neurological manifestations of dengue infection. *Lancet* **355**, 1053–9.

Townsend JJ, *et al.* (1975). Progressive rubella panencephalitis—late onset after congenital rubella. *New England Journal of Medicine* **292**, 990–3.

24.14.3 Intracranial abscess

P. J. Teddy

Intracranial abscesses may occur within the extradural or subdural space, or may be intracerebral. Occasionally, abscesses exist in more than one tissue plane. Intracerebral and subdural abscesses may rupture into the subarachnoid space and be accompanied by meningitis; intracerebral pus may rupture into the ventricular system and produce ventriculitis.

Aetiology

Extradural abscesses are usually related to focal osteomyelitis of the skull, mastoiditis and nasal sinusitis, penetrating injuries of the skull, and are a rare complication of craniotomy.

Subdural empyema is related most commonly to infection of the paranasal sinuses and middle ear. Other causes include septicaemia related to cyanotic congenital heart disease, lung abscess, trauma, and intracranial surgery.

The most common intracranial abscess is found within the intracerebral compartment, with about 60 per cent related to middle-ear infection and 20 per cent to frontal sinusitis. Other established causes are septicaemia related to congenital heart disease with a right-to-left shunt, lung abscess, bronchiectasis, penetrating injuries of the head, and bacteraemia following tooth extraction. In about 10 per cent of cases no primary source of infection can be identified. Owing to their strong connection with sinus and middle-ear disease, most intracerebral abscesses are found within the frontal or temporal lobes, or within the cerebellum. Infection disseminated through the bloodstream from more distant sites may result in multiple abscesses in any part of the brain.

Microbiology

The most common organisms associated with subdural empyema are aerobic, anaerobic, and micro-aerophilic streptococci, *Staphylococcus aureus*, and *Bacteroides* spp.

Cerebral abscesses associated with otitis media, mastoiditis, and nasal sinusitis usually show a mixed growth of anaerobes and aerobic organisms including anaerobic and micro-aerophilic streptococci and *Bacteroides*. *Streptococcus viridans and Staph. aureus* are frequently seen. *Listeria* spp. tend to produce areas of focal cerebritis rather than true abscess.

Pathology

Infection within an accessory air sinus or the petrous bone may cause an area of localized osteitis just above the dura, which can then spread intracranially. Initially it may be entirely confined to the extradural space, but will eventually penetrate the dura and spread subdurally or, if the adjacent arachnoid is stuck to the inflamed patch of dura, then it will spread into the subarachnoid space to give meningitis. If the subarachnoid space has been obliterated, it may penetrate the brain to produce initially a focal cerebritis. Usually after about 10 days the area of cerebritis becomes enclosed within an area of gliotic brain, and after about 3 weeks a firm capsule forms around the pus. Large intracerebral abscesses may rupture into the ventricular system, producing a ventriculitis.

Cerebral abscesses are usually surrounded by areas of oedematous brain, which may exert a considerable mass effect.

Clinical features

These will depend upon the site, size, and number of lesions, and the involvement of neighbouring structures such as the cerebral ventricles and the venous sinuses. The signs are therefore legion, but the diagnosis should be considered in any case where there is an obvious primary source of infection associated with evidence of raised intracranial pressure, focal neurological signs, epileptic seizures or meningeal irritation, or any combination of these.

Extradural abscess may be difficult to detect clinically, but is sometimes manifest by severe, unremitting, localized headache in association with sinusitis or mastoiditis. Patients with subdural empyema frequently appear toxic, with a swinging pyrexia, severe headache, a depressed level of consciousness, contralateral hemiparesis, papilloedema, meningeal irritation, and seizures. There is usually an accompanying frontal sinusitis with tenderness of the forehead and redness and swelling of the eyelids, or mastoiditis or scalp infection.

Diagnosis

If a brain abscess is suspected, predisposing sources of infection, including possible distant sites, should be carefully sought, as intracranial abscesses derived by haematogenous spread are often more fulminating in their course than those associated with local cranial disease. If CT is available, scans of the skull base, including views of the mastoids and other skull sinuses, should be performed. Otherwise, skull radiography with sinus views is necessary. Chest radiographs should be obtained.

The investigations of choice for all forms of suspected intracranial abscess are either CT scanning, with and without contrast, or MRI. CT will normally demonstrate both extradural and subdural empyema, may demonstrate diffuse cerebritis in early cases, and will normally show intracerebral abscesses as ring-enhancing lesions with low-attenuation centres (see Fig. 1). Nevertheless, there are pitfalls, particularly in the early stages both of subdural empyema and of cerebral abscess. Subdural empyema may initially be fairly thinly spread over the cerebral cortex, producing relatively little midline shift, and may be virtually isodense with brain on CT. Under such circumstances, contrast-enhanced MRI (particularly with coronal views) is of great value.

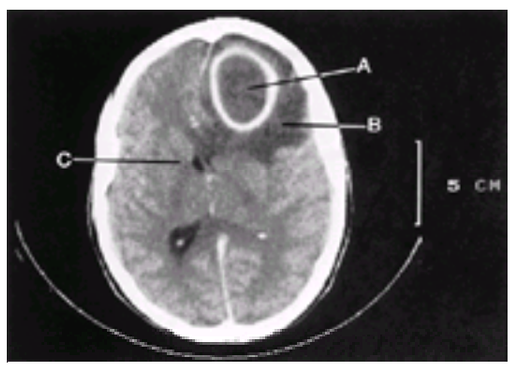

Fig. 1 Contrast CT scan showing large right frontal cerebral abscess (A) with surrounding oedema (B) and ventricular compression (C).

The principal differential diagnoses in an intracranial abscess are meningitis, subdural haematoma, and intracranial tumour. It is not always possible to differentiate between intracerebral abscess and tumour on CT scan, particularly when there is an appearance of ring enhancement, and it is largely for this reason that the biopsy of suspected cerebral tumour is advocated in nearly all such cases. MRI, however, tends to show a low-signal capsule on T_2-weighted images and may be helpful in making this differentiation.

One obvious concern is to differentiate between bacterial meningitis and intracerebral abscess. Both may present with pyrexia, neck stiffness, and with some focal signs, but if there is any evidence whatsoever of raised intracranial pressure, or any other supportive evidence of cerebral abscess, a lumbar puncture should be strictly avoided until a neurosurgical opinion has been sought. Lumbar puncture in the presence of cerebral abscess can lead to tonsillar or tentorial herniation, and in any event, the cerebrospinal fluid can be entirely normal.

Management

Except in a few cases of multiple or inaccessible abscess, and the occasional patient whose general medical condition is such that surgery is precluded, treatment of the intracranial infection requires evacuation of pus and high-dose intravenous antibiotic therapy.

The single, main factor in securing a good outcome is early diagnosis. Early management includes taking specimens for blood culture and culture of any extracranial infective lesion, setting up an intravenous infusion, administration of anticonvulsant agents, and, in cases of grossly depressed level of consciousness and massive cerebral oedema seen on CT scan, giving intravenous dexamethasone.

Pus from the suspected primary site of infection should be collected immediately and both aerobic and anaerobic cultures obtained. The intracranial pus must be similarly cultured. Antimicrobial treatment, using massive intravenous doses, should be commenced immediately without waiting for the culture report, and subsequently changed in the light of the sensitivity findings. The antimicrobial regimen should include penicillin (4 mega units 4-hourly), metronidazole, ampicillin, and either gentamicin or chloramphenicol depending on the likely source of infection and the infective agent. Intravenous antimicrobials should be continued for at least 1 week before reverting to oral medication.

Most supratentorial abscesses can be sterilized by aspiration through a burr hole, and the direct instillation of antibiotics is sometimes employed. Aspiration must usually be repeated several times, but in about 30 per cent of cases a single aspiration will suffice. Once the abscess is sterile, the capsule will shrink and finally form an irregular gliotic scar within the brain. Shrinkage of the abscess must be checked by serial CT scan. Subdural empyema should be evacuated through a craniotomy rather than burr holes, as very often the pus can spread widely, and particularly alongside the falx cerebri. Extradural empyema is evacuated through a burr hole, or through a craniotomy for larger collections.

Cerebellar abscess, when diagnosed early, may be aspirated through a burr hole, but immediate total excision is often recommended because the small volume of the posterior cranial fossa leaves little latitude in terms of tonsillar herniation and death.

Prognosis

The mortality is around 10 per cent, but the main problems remain those of late diagnosis and resistant bacteria. Even with an otherwise good outcome, epilepsy may continue in about 30 per cent of cases, particularly in patients with temporal-lobe abscess and subdural empyema.

Further reading

Lorber B (1997). Listeriosis. *Clinical Infectious Diseases* **24**,1–9.

Mathisen GE, Johnson JP (1997). Brain abscess. *Clinical Infectious Disease*, **25**, 763–79.

Report of the Quality Standards Subcommittee of the American Academy of Neurology.(1998). Evaluation and management of intracranial mass lesions in AIDS. *Neurology*, **50**, 21–6.

24.14.4 Neurosyphilis and neuroAIDS

Hadi Manji

Neurosyphilis

Introduction

Syphilis remains a public health problem in certain areas of the United States, Eastern Europe, and in the developing world. The incidence of primary and secondary syphilis in the United Kingdom (excluding Scotland) increased from 109 cases in 1995 to 259 cases in 2000 – an increase of 138 per cent. Since syphilis, like other ulcerating genital infections such as herpes and chancroid, is an independent risk factor for the acquisition and transmission of infection with the human immunodeficiency virus (**HIV**), the disease has once again come under scrutiny. In addition, there are recent anecdotal reports of *Treponema pallidum* being more neurovirulent and with a greater risk of treatment failure in those dually infected with HIV.

Invasion of the central nervous system occurs early in the course of syphilis infection. *T. pallidum* has been isolated from the cerebrospinal fluid of up to 40 per cent of neurologically asymptomatic patients with untreated primary and secondary syphilis. Despite this, cohort studies of untreated patients suggest that symptomatic late syphilis (neurosyphilis, cardiovascular syphilis, and gummas) occurs in 15 to 40 per cent of such individuals; the Oslo study documented an incidence of clinical neurosyphilis in 9.4 per cent. Thus, it would seem as if, at least in the immunocompetent patient, *T. pallidum* has a low virulence for the central nervous system.

Clinical features (see Table 1)

Acquired syphilis is divided into an early, potentially infectious stage (primary, secondary, and early latent where less than 2 years have lapsed since infection) and a late, non-infectious stage (late latent where more than 2 years have lapsed, gummatous, cardiovascular, and neurosyphilitis). Although there is a rough time course to the development of the various neurological syndromes, there is considerable overlap; these syndromes are, in reality, part of a spectrum of disease.

Neurosyphilis may include meningitis (acute and chronic), a myeloradiculopathy due to a pachymeningitis and granulomatous lesions (gummas) that present as space-occupying lesions within the brain, the spinal cord, or the epidural space causing compression. Meningovascular syphilis involves the small- and medium-sized arteries, typically causing an endarteritis (Heubner's endarteritis obliterans) resulting in infarction. The so-called late manifestations of neurosyphilis result from a low-grade meningoencephalitis. In patients with general paralysis (also called general paralysis of the insane or dementia paralytica) the focus is on the frontotemporal cortex. Therefore, during the early stages, vague symptoms may include personality and mood changes, with impaired faculties of concentration and attention being the presenting features; memory difficulties develop later.

In tabes dorsalis (taboparesis), which may coexist with general paralysis, the clinical presentation results from involvement of the dorsal roots and ganglia as well as the posterior columns within the spinal cord, with the resultant emphasis on a sensory ataxia. Diabetes may produce a similar clinical picture with a neuropathy and pupillary abnormalities (diabetic pseudotabes).

The optic nerve may be involved with or without other evidence of neurosyphilis, but must always be treated as if it were part of a systemic infection. A uveitis, chorioretinitis, optic neuritis, papillitis, and optic atrophy have all been reported at different stages of the disease. Extraocular presentations include nerve palsies involving the eye muscles and a superior orbital fissure syndrome. Although the Argyll Robertson pupil may occur in any form of the disease, it is generally encountered in tabes dorsalis. The pupils are small and irregular, being unreactive to light, but constrict normally to accommodation and convergence. Unilateral involvement is rare. The light/near dissociation is the result of gliosis in the periaqueductal grey midbrain tegmentum, which may also account for the bilateral ptosis seen in some individuals.

Diagnosis

Neurosyphilis has a myriad of neurological manifestations and therefore the diagnosis enters the differential of most neurological conditions (Table 1). Treatment in the early stages of the disease (that is, of the meningitic and meningovascular syndromes) may well result in recovery, whereas the late forms—with general paralysis and tabes dorsalis—may only respond partially, if at all. These common neurological presentations include stroke, especially in younger patients, chorioretinitis, optic neuropathy of unknown cause, and single or multiple cranial neuropathies, particularly those involving the VIIIth nerve with vertigo and sensorineural deafness. Syphilis serology should be routinely performed in patients with dementia and psychiatric illnesses.

The serum reaginic tests, Venereal Diseases Research Laboratory test (**VDRL**) and rapid plasma reagin test (**RPR**), are usually positive in secondary syphilis when the first neurological complications may be encountered. However, a false-negative result may occur due to the prozone phenomenon if undiluted serum is used. This occurs in 1 to 2 per cent of cases of secondary syphilis and is due to blockage of agglutination caused by the saturation of antigenic sites by excess antibody. The specific serological tests (*Treponema pallidum* haemagglutination test (**TPHA**), *T. pallidum* particle agglutination test (**TPPA**), fluorescent treponemal antibody absorption test (**FTA-abs**), and the treponemal enzyme immunoassay (**EIA**)) are invariably positive.

In late syphilis (meningovascular syphilis, gummatous, general paralysis, and tabes dorsalis), the serum VDRL/RPR tests are negative in 30 per cent of untreated cases. All the specific tests have a sensitivity approaching 100 per cent, so that a negative treponemal antigen test has an extremely high predictive value for excluding neurosyphilis.

It is recommended that all patients with positive syphilis serology who have ocular and or neurological symptoms and signs should undergo cerebrospinal fluid (**CSF**) examination, as should patients with latent infection of unknown duration. In order for these tests to be correctly interpreted it is important that the CSF is not significantly (macroscopically) contaminated with blood.

In patients with neurosyphilis there is usually a lymphocytic pleocytosis (>5 cells/µl), with an elevated protein (>0.4 g/l). In the late stages, particularly in tabes, the CSF may be quiescent. A reactive CSF-VDRL establishes the diagnosis of active neurosyphilis, but a non-reactive test does not exclude it. The sensitivity of the CSF–VDRL is 50 per cent, with a specificity of 100 per cent. A non-reactive CSF-FTA-abs or TPHA excludes the diagnosis. However, a reactive CSF-FTA-abs or TPHA does not establish the diagnosis because the presence of treponemal antibodies in the CSF could result from the passive transfer from the blood, or may result from a previous episode of treated syphilis. The sensitivity for the CSF-FTA-abs is 100 per cent, with a specificity of 30 per cent.

The role of the polymerase chain reaction (**PCR**) in the diagnosis of neurosyphilis is unclear at present, for technique cannot discriminate between viable and non-viable organisms. *T. pallidum* DNA has been detected in CSF up to 3 years after intravenous treatment with penicillin.

Treatment

In patients with symptomatic neurosyphilis or ocular disease, the World Health Organization/United Nations Programme on HIV/AIDS (**WHO/UNAIDS**) as well as the Centers for Disease Control (**CDC**) recommend treatment with penicillin G (2–4 mU intravenously every 4 h for 14 days). In the United Kingdom the preference is for procaine penicillin (1.8–2.4 million IU intramuscularly once daily, plus probenecid 500 mg by mouth four times daily, for 17–21 days). The alternative is intravenous benzyl penicillin (3–4 million units intravenously every 4 h for 17–21 days). In patients with a history of penicillin allergy one option is to perform skin testing to confirm the allergy and to then consider desensitization. The other is to treat with doxycycline 200 mg by mouth four times daily for 28 days.

Following treatment of neurosyphilis, a repeat lumbar puncture should be performed at 6-month intervals until the cell count is normal. This should be decreased by 6 months and be entirely normal by 2 years. The CSF-VDRL may take years to become non-reactive.

Syphilis in the era of HIV

Since the onset of the AIDS epidemic there have been numerous reports of an accelerated course of syphilis and of treatment failures in patients who are dually infected. Compared to non-immunosuppressed individuals there certainly does seem to be a higher than expected rate of cases of syphilitic meningitis and meningovascular syphilis. To date, however, there are no denominator data. Since cell-mediated immunity, which is necessary to eradicate *T. pallidum*, may be impaired in HIV infection this seems plausible.

As a result of altered B-cell function there has been concern regarding serological tests. However, these are usually positive or may show a delayed response in the occasional case.

There is still debate as to whether or not patients with HIV and early syphilis who are neurologically asymptomatic should have a CSF examination. Any CSF cytochemical abnormalities could be either be due to HIV or syphilis. In view of the treatment failures reported with benzathine penicillin some authorities suggest that all HIV patients with early syphilis should be treated with neurosyphilis treatment regimens.

NeuroAIDS (or neurological complications of HIV infection)

Introduction

Soon after the onset of the AIDS epidemic in 1981, it became clear that the nervous system was frequently involved. However, opportunistic infections such as toxoplasmosis and cryptococcal meningitis as well as neoplasms (such as primary central nervous system lymphoma (**PCNSL**)) accounted for only 30 per cent of the neurological problems encountered. It also became evident that in the later stages of the AIDS illness, patients developed neurological complications due to the human immunodeficiency virus itself. This included a progressive decline in cognitive function in association with motor abnormalities—the HIV–dementia complex.

Neurological disorders are the AIDS-defining illness in up to 20 per cent of cases. Over the course of the illness the prevalence of neurological complications increases up to 70 per cent. These include other opportunistic infections and tumours, as well as the HIV-related problems of dementia, vacuolar myelopathy, and distal sensory peripheral neuropathy.

At postmortem more than 90 per cent of the brains from patients dying of AIDS show evidence of HIV encephalitis and of one of the opportunistic infections such as cytomegalovirus (**CMV**) and tumours.

During the last 5 years, with the introduction of the highly active antiretroviral therapies (**HAART**), there has been a dramatic decline in the incidence of neurological opportunistic infections as well as HIV related disorders such as HIV dementia. However, these are expensive drugs and are out of reach of the majority of HIV-infected individuals worldwide.

Clinical approach

All areas of the neuraxis are vulnerable in individuals infected with HIV. Not infrequently do differing pathological processes occur simultaneously in various parts of the nervous system. Thus, Occam's Razor—the principle of diagnostic parsimony, often used in medicine—does not always apply. Another aspect is the possibility of simultaneous infection with more than one organism: for example, meningitis due to *Mycobacterium tuberculosis* and *Cryptococcus neoformans*. Mass lesions in the brain, with some not responding to antitoxoplasma therapy, could be due to lymphoma or another infective cause such as a tuberculoma.

The nervous system is involved early in the course of infection, as evidenced by neurological seroconversion illnesses such as an aseptic meningitis, encephalitis, and the Guillain–Barré syndrome. Furthermore, during the asymptomatic phase of the illness (that is, when patients are well) the cerebrospinal fluid shows abnormalities in up to 60 per cent of cases. This may be a lymphocytic pleocytosis of up to 50 cells/mm ^{3}, an elevated protein, or the presence of oligoclonal bands. The CSF glucose level is usually normal. Therefore, these cytobiochemical markers are unhelpful in making a diagnosis of a meningitic or an encephalitic illness. Reliance is therefore placed on specific markers such as the cryptococcal antigen or antibody tests like the CSF-VDRL or TPHA.

As a result of the impaired immune response, a rise in antibody titres to specific infections may not occur, especially during the later stages of HIV infection. Furthermore, the typical clinical picture—the presentation of which, at least in some infections like meningitis, are due to a brisk inflammatory response such as fever—may not occur. In cryptococcal meningitis, only one-third of patients exhibit the classical signs of meningism: namely, neck stiffness, photophobia, and a positive Kernig's sign.

The specific type of opportunistic complications encountered is dependent on a number of factors, including the degree of immunosuppression. During the early stages when subjects are relatively immunocompetent, with CD4 counts above 500/µl, autoimmune disorders such as demyelinating neuropathies may occur. Between CD4 counts of 200 and 500/µl multidermatomal herpes zoster infections may present. Once the level declines below 200/µl, patients are vulnerable to all the major opportunistic infections and the complications due to HIV itself. Symptomatic infection with cytomegalovirus tends to occur at very low levels below 50/µl.

In different parts of the world, some infections may be more prevalent than others. The incidence of toxoplasmosis in France is significantly higher than in the United Kingdom because of a higher background seroprevalence, due to differing dietary habits and the greater prevalence of raw meat consumption.

Opportunistic infections

Toxoplasmosis

Toxoplasma gondii, whose definitive host includes members of the cat family with humans the intermediate hosts, is an obligate intracellular protozoan. Human infection occurs through the ingestion of tissue cysts in undercooked meat. Variations in dietary habits therefore explains the differing seroprevalence rates worldwide—90 per cent in French adults compared to 50 per cent of residents in the United Kingdom. Symptomatic toxoplasmosis is usually due to a reactivation of latent infection in individuals with HIV. The risk of an HIV-infected patient who is seropositive for IgG *T. gondii* antibody developing toxoplasmosis is around 25 per cent.

Toxoplasmosis is the most common cause of mass lesions in the brains of patients with HIV infection. The clinical presentation is variable, but headache, confusion, seizures, and focal neurological deficits such as hemiplegia, dysphasia, and visual field defects are the most common. Other presentations described include: a variety of movement disorders (choreoathetosis, dystonia, and hemiparkinsonism); psychiatric illness such as depression; brainstem syndromes; and a rapidly progressive diffuse encephalitis. Rarely, the spinal cord may be involved with a myelitis or a cauda equina syndrome.

A definitive diagnosis of toxoplasma encephalitis can only be made by brain biopsy. With increasing experience and pragmatism, it is now standard practice to treat any HIV-infected individual who has a low CD4 count and multiple lesions on imaging with antitoxoplasma therapy (Fig. 1). A response, clinically and radiologically, confirms the diagnosis. Although a negative blood toxoplasma serology result makes the diagnosis less likely, it may occur in up to 17 per cent of cases. This loss of

seropositivity may be the result of impaired antibody synthesis with increasing immunosuppression. It is useful therefore to document an individual's toxoplasma serostatus on first diagnosis of HIV-positivity. For similar reasons, the expected rise in IgM and IgG levels does not occur. A single lesion on magnetic resonance imaging (**MRI**) is most likely to be due to lymphoma. A single lesion on computed tomography (**CT**) scanning should, if possible, be followed by MRI, which is a more sensitive method of detecting lesions particularly in the posterior fossa (Fig. 2).

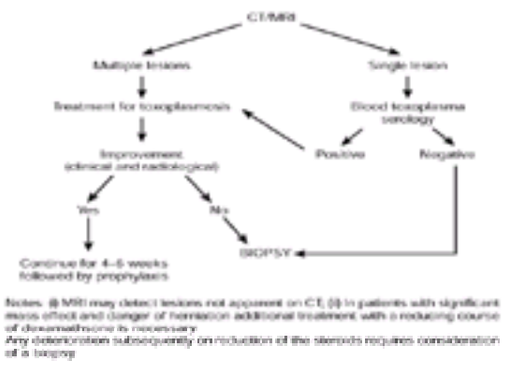

Fig. 1 Management of mass lesions in AIDS

Fig. 2 Cranial CT—multiple lesions with mass effect and cerebral oedema due to toxoplasmosis.

The main differential diagnosis is that of primary CNS lymphoma, which presents at similar CD4 counts and with a similar presentation both clinically and on imaging studies (Table 2).

A response is seen in 90 per cent of patients by the second week of treatment (Table 3). It is necessary to reimage 2 weeks after treatment even if there is clinical evidence of improvement, since it is not uncommon for some lesions to improve but others due to, for example *Mycobacterium tuberculosis*, to enlarge which then makes it necessary to consider a biopsy. The radiological improvement generally lags behind the clinical improvement.

Patients infected with HIV who are seropositive for IgG against *T. gondii* should be offered primary prophylaxis with 980 mg of co-trimoxazole (trimethoprim and sulfamethoxazole) when their CD4 count falls below 100/mm^3. This will confer cross-protection against *Pneumocystis carinii* pneumonia.

Cryptococcus neoformans

This encapsulated yeast is a ubiquitous organism in the environment acquired by humans through inhalation. Although disseminated infection can involve the skin, bones, lungs, eyes, and prostate, symptomatic infection with *C. neoformans* most often presents as a meningitis.

Cryptococcal infection is the most common infectious cause of meningitis in patients with AIDS (Table 4). The presentation may be acute, but it is usually subacute with symptoms of malaise, headache, fever, and vomiting. The classical signs of meningism—neck stiffness, photophobia, and Kernig's sign—are present in only one-third of patients. Other, less common symptoms include altered mental status, seizures, and focal neurological signs. The latter are due to parenchymal cryptococcal abscesses.

Brain imaging is usually normal, although the basal meningitis may result in hydrocephalus or sometimes, particularly on MRI, small abscesses—cryptococcomas—may be visualized.

Cerebrospinal fluid examination is essential for the diagnosis, with culture of the fungus being the 'gold standard'. The cytochemical markers in the CSF may be normal. India-ink staining of the CSF will reveal the fungal hyphae in 70 to 80 per cent of cases and cryptococcal antigen is detected in over 90 per cent. Cryptococcal antigen is also detected in the blood in over 90 per cent of patients, and should be measured in conjunction with the CSF level since in occasionally reported cases of fulminant cryptococcal meningitis the CSF may be negative and the blood positive. The blood antigen measurement may be used as a screening test in patients presenting with symptoms of early infection such as headache. However, it should be appreciated that a negative result does not completely exclude the diagnosis of cryptococcal meningitis.

Treatment with amphotericin B remains the drug of choice for the treatment of severe cases of cryptococcal meningitis. The mortality rate still remains around 10 per cent. Features that have been identified with a poor outcome include a relapse infection, abnormal mental status, CSF cryptoccal antigen titre over 1:1024, CSF white cell count <20 cells/mm^3, positive India-ink staining, hyponatraemia, and positive culture from an extrameningeal site. A CSF opening pressure of greater than 250 mmH_2O is also a marker of poor prognosis. In milder cases, where none of these features are present, oral fluconazole may be used. Although combination with 5-flucytosine has been shown to improve outcome in non-AIDS patients, this has not been confirmed in patients with AIDS. However, the combination should be considered in fulminant cases.

A specific complication that requires close monitoring is the development of raised intracranial pressure due to obstruction of the arachnoid villi and cerebral oedema. This should be managed with repeated lumbar puncture or, if necessary, by the insertion of a lumbar or ventricular drain.

Maintenance therapy is essential, with relapse rates approaching 100 per cent if secondary prophylaxis with oral fluconazole is not adhered to. The serum cryptococcal antigen titre is not useful in predicting relapse.

JC (Jamestown canyon) virus

Progressive multifocal leucoencephalopathy (**PML**) is caused by the reactivation of latent JC virus, which is acquired by the majority of the population during childhood as a banal upper respiratory infection. Prior to the AIDS epidemic, PML was a rare condition encountered in patients immunosuppressed as a result of haematological malignancies, drugs used in the treatment of post-transplant patients, autoimmune disorders such as systemic lupus erythematosus (**SLE**), and granulomatous disorders such as sarcoidosis. Nowadays, underlying HIV infection accounts for 85 per cent of cases.

Prior to the introduction of HAART, the incidence of PML was 4 per cent. The clinical presentation is subacute, with progressive focal neurological deficits such as a hemiparesis, visual field defects, and a cerebellar syndrome. The disorder is not restricted to the white matter since patients may also develop dysphasia and

seizures. Occasional patients may present with a progressive dementia with focal neurological signs.

MRI characteristically shows multiple areas of high signal on T_1-weighted images and a low signal on T_2-weighted ones (Fig. 3). There is little or no enhancement, with no mass effect or oedema around the lesions. Blood serological testing is unhelpful since 80 per cent of the general population is seropositive. Recently it has become possible to confirm the diagnosis of PML by isolating JC-viral DNA in cerebrospinal fluid by PCR techniques. This has a sensitivity of 75 per cent with a specificity of 95 per cent. In PCR-negative cases it may be necessary to either repeat the CSF examination or to perform a brain biopsy. The typical histological features show areas of focal demyelination, bizarre enlarged astrocytes, and abnormal oligodendrocytes with inclusion bodies that stain for JC viral antigens.

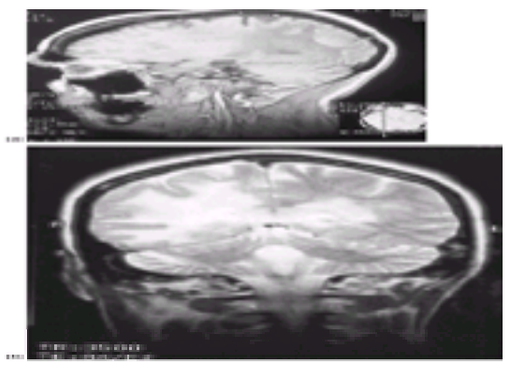

Fig. 3 (a) T_2 weighted and (b) T_1 weighted MRI in a patient with PML.

There is, to date, no specific treatment. Cytosine arabinoside, both intravenous and intrathecal, has been shown to be ineffective. Trials are underway to examine the efficacy of cidofovir, an anti-CMV drug, and interferon-a. However, improving immune function with HAART has been shown to improve survival times from a median survival of 10 weeks to 40 weeks.

Cytomegalovirus (CMV)

The neurological complications from this herpesvirus results from reactivation in severely immunocompromised patients. Almost all patients infected with HIV are seropositive for cytomegalovirus. Postmortem studies of the brains of patients who died from AIDS show evidence of CMV in 25 per cent of cases. However, clinical CMV disease, apart from CMV retinitis, is rare.

CMV retinitis is the most common manifestation of CMV disease and can affect up to 20 per cent of patients with AIDS. The slowly progressive necrotizing retinitis results in characteristic white irregular lesions with central necrosis and haemorrhages—the cheese and tomato ketchup appearance. Retinal detachment may occur in patients with extensive retinal involvement. The retinitis presents with symptoms of reduced visual acuity, floaters, and loss of peripheral vision. Since the condition may be asymptomatic in the early stages, regular ophthalmological screening is recommended for high-risk patients with CD4 counts below 50 cells/μl.

A necrotizing ventriculoencephalitis has been described, usually in patients with evidence of CMV disease elsewhere (Table 5). The onset is subacute over a period of days or weeks with confusion, seizures, and brainstem signs such as internuclear ophthalmoplegia, ataxia, and cranial nerve palsies. Imaging studies typically show periventricular enhancement.

CMV polyradiculopathy presents over a period of days with back pain, leg weakness, sensory impairment, and sphincter disturbance. The differential diagnosis includes syphilitic polyradiculopathy and infiltration with metastatic lymphoma. The cerebrospinal fluid reveals a polymorphonuclear leucocytosis which is unusual for a viral infection. Early recognition and treatment is necessary to stabilize and, in some cases, improve the neurological impairment.

Drugs licensed for the treatment of CMV disease include ganciclovir, cidofovir, and foscarnet. Oral ganciclovir is prescribed for secondary prophylaxis.

Opportunistic tumours

Primary CNS lymphoma (**PCNSL**) is the second most common cause of mass lesions after toxoplasmosis in adults, and the most common in paediatric patients with AIDS. Histologically, this a high-grade, non-Hodgkin's, B-cell lymphoma. Recent evidence suggests that the Epstein–Barr virus is causally linked to PCNSL, with the identification of the viral DNA incorporated into that of the neoplastic cells.

The common presenting symptoms are those of headache with focal neurological deficits, altered level of consciousness, and seizures.

Brain imaging reveals enhancing mass lesions with surrounding oedema and mass effect. These are similar to those found in toxoplasmosis. PCNSL is more likely to present as a single mass lesion than toxoplasmosis and is also more likely to invade the ventricular walls. Recent studies using thallium-201 single-photon emission computed tomography (**SPECT**) suggest that it may be possible to differentiate between an abscess and a tumour, with the former having little uptake compared with the high uptake of the mitotically active lymphoma.

There is no effective treatment for PCNSL. Whole brain radiotherapy provides, at best, only a modest benefit, with most patients succumbing within 2 months.

HIV-associated neurological disorders

HIV–dementia complex

Before the introduction of HAART (and in areas of the world where they are still unavailable) approximately 15 to 20 per cent of individuals infected with HIV developed a variably progressive dementia with associated motor deficits. In children, a similar HIV-1 associated progressive encephalopathy occurs more frequently than with opportunistic infections. This usually occurs within the context of severe immunosuppression in those with a CD4 count of less than 200/mm ^{3}. In around 3 per cent of cases, HIV-dementia is the AIDS-defining illness. Large cohort studies, using clinical, MRI, and neuropsychological methods, have largely discounted the early reports of evidence of cognitive changes in asymptomatic HIV-positive patients.

The clinical presentation in the early stages is with vague symptoms of apathy, mood changes, and difficulty with memory and concentration. These are features of a subcortical dementia with no features of cortical involvement such as language, visuospatial or calculation difficulties. This picture may be mimicked by depression, metabolic encephalopathy, and drugs, both therapeutic and recreational. At this stage, there may be few physical signs apart from brisk reflexes, impaired fine finger movements, and unsteady gait.

Later, the memory impairments are obvious, as is the psychomotor retardation—which may progress to frank mutism and a global dementia. Some patients develop seizures. The motor signs due to the associated vacuolar myelopathy with a spastic paraparesis and sphincter disturbances are also present in a significant number of patients (Table 6). In addition, some patients will have the HIV-related distal sensory peripheral neuropathy. Thus, this group will have absent ankle jerks and extensor plantar responses.

The diagnosis of the HIV–dementia complex is made by clinical assessment—there are usually no focal signs and the tempo of the disorder is an insidious one. Investigations are performed to exclude other infection or neoplastic pathologies, and therefore necessitate imaging, preferably with MRI, and a CSF examination. MRI may show evidence of cerebral atrophy with compensatory ventricular dilatation, a diffuse white-matter high signal on T_2-weighted images with no enhancement. A CSF examination may be non-specifically abnormal with a pleocytosis, elevated protein level, and oligoclonal bands. It important to exclude cryptococcal and tuberculous meningitis as well as neurosyphilis. The HIV RNA-viral load in cerebrospinal fluid correlates with the severity of clinical dementia, but there is too much

overlap between non-demented and demented subjects for the measurement to be of use as a diagnostic aid. There is no correlation between the plasma HIV RNA-viral load and dementia. Electroencephalography (**EEG**) may be normal in the early stages, with non-specific diffuse slowing being shown later.

The pathology of the HIV–dementia complex is a spectrum ranging from diffuse myelin pallor, microglial nodules—which are non-specific and may be found in CMV encephalitis—to multinucleated giant cells that are indicative of productive brain infection and cortical neuronal loss. There is no clear correlation between the clinical and pathological findings.

The mechanisms of disease in the HIV–dementia complex are still unclear. It is, however, evident that HIV predominantly infects the microglial and astrocytic cells rather than neurones or oligodendrocytes. One hypothesis for the entry of the virus into the CNS is the 'Trojan horse' theory, with invasion occurring by infected peripheral blood monocytes penetrating a blood–brain barrier that has been disrupted by damage to the capillary endothelial cells. Neuronal damage is subsequently caused by virotoxins (for example, Gp120) and cytokines (for example, tumour necrosis factor-a) released from activated macrophages.

After the introduction of zidovudine in 1987, there was a dramatic reduction in the incidence of HIV-associated dementia. One clinical study looking specifically at the effect of zidovudine on cognitive function confirmed its beneficial effect, albeit at dosages much higher than those currently used. With the introduction of the newer antiretroviral drugs—the majority of which have poor penetration into the CSF and presumably the brain—there is concern that, despite the reduction of plasma HIV viral loads, the CNS may develop into a safe sanctuary for the virus from which reinfection could occur. However, recently published studies do suggest that these newer therapies improve cognitive function and it seems prudent, until further data become available, to use drugs that best penetrate the CSF to treat HIV dementia complex.

Since macrophage activation resulting in the release of neurotoxic factors also has a role in the pathophysiological mechanism, trials are underway to assess the therapeutic benefits of drugs such as the **PAF** (platelet-activating factor) antagonist, lexipafant, and the **NMDA** (*N*-methyl-D-aspartate) antagonist, memantine.

HIV-associated neuropathy

The most common neurological complication encountered in patients infected with HIV is distal sensory polyneuropathy (**DSPN**), which may occur in 30 per cent of those with AIDS (Table 7). It is a significant cause of morbidity.

Typically, patients complain of numbness of the soles of the feet together with shooting pains and parasthesias developing over a period of months. There is little or no weakness. The hands are infrequently involved. The ankle jerks are depressed or absent. Sensory testing reveals impaired pain and temperature perception as well as vibration.

Further investigations are usually unnecessary in a patient with a CD4 count below 200 and showing the typical clinical picture, but it is always worth checking the blood sugar, vitamin B_{12} level, and syphilis serology. It is important to enquire about alcohol intake and the possibility of an excess intake of vitamin B_{6}.

Neurophysiological and pathological studies suggest this to be a length-dependent axonal neuropathy. Productive HIV infection has not been found in pathological specimens and the underlying mechanisms, like those for HIV dementia, are linked to macrophage activation products.

Since antiretroviral therapy has no benefit, treatment is symptomatic with the use of tricyclic antidepressants and anticonvulsant drugs such as gabapentin.

The nucleoside analogues didanosine (**ddI**), zalcitabine (**ddC**), and stavudine (**d4T**) cause a dose-dependent sensory neuropathy that may be indistinguishable from distal sensory polyneuropathy. Clues to this drug-induced neuropathy include the shorter history of weeks rather than months, and the improvement on stopping the offending drug. However, there may be a continued worsening of symptoms for a period of 4 to 8 weeks after stopping—the phenomenon of 'coasting'. The underlying mechanism appears to be the impairment of mitochondrial protein synthesis.

Further reading

Brew B (2001). *HIV neurology*. Oxford University Press, Oxford.

Clinical Effectiveness Group (1999). National guideline for the management of early syphilis. *Sexually Transmitted Infections* **75**(Suppl 1), S29–S33.

Clinical Effectiveness Group (1999). National guideline for the management of late syphilis. *Sexually Transmitted Infections* **75**(Suppl 1), S34–S37.

Harrison MJ, McArthur JC (1995). *AIDS and neurology*. (*Clinical Neurology and Neurosurgery Monographs*). Churchill Livingstone, Edinburgh.

Seminars in Neurology (1999). **19,**Thieme Medical Publishers. [Whole volume devoted to HIV neurology.]

Swartz MN, Healy BP, Musher DM (1999). Late syphilis. In: Holmes KK, *et al*., eds. *Sexually transmitted diseases*, pp 487–509. McGraw-Hill, New York.

24.15 Metabolic disorders and the nervous system

*Neil Scolding and C. D. Marsden**

In general, the term metabolic diseases of the nervous system covers the neurological consequences of systemic disorders of metabolism. This alone is an enormous field, ranging from common disorders, such as diabetes, chronic and acute alcohol poisoning, and renal disease, to less common but no less important disorders such as pontine myelinolysis and critical illness polyneuropathy.

Metabolic complications of major organ disease

Cardiovascular disease/anoxia

Cerebral anoxia may be due to insufficient cerebral blood flow, reduced oxygen availability, reduced oxygen carriage by the blood, metabolic interference with the utilization of oxygen, or combinations of these events. Thus, acute cardiovascular insufficiency as a consequence of cardiac arrest is a relatively common cause of severe global cerebral anoxia; others as diverse as suffocation, anaesthetic catastrophes, drowning, or acute carbon monoxide poisoning can produce similar results.

A brief period of global ischaemic anoxia causes syncope. If the episode is prolonged, myoclonic jerks or tonic–clonic seizures may occur. Still more protracted insults may precipitate a period of confusion and residual amnesia.

Persisting acute severe anoxia rapidly leads to loss of consciousness, generalized fits, dilated pupils, and bilateral extensor plantar responses. Periods of anoxia up to perhaps 5 min may cause transient coma with recovery of consciousness. A delayed postanoxic encephalopathy, characterized pathologically by demyelination in the hemispheres and in the basal ganglia, may follow within 1 to 2 weeks, often commencing with increasing irritability, apathy, and confusion. Frank dementia may emerge, or an amnesic syndrome in less severe cases, and there may also be pseudobulbar palsy and other pyramidal signs, gait ataxia, and incontinence, and/or an akinetic–rigid syndrome with or without dystonia. Some patients may be severely disabled by action myoclonus—dramatic muscle jerking on attempted movement. The residual deficits following prolonged cerebral anoxia with survival may be permanent; in other patients they gradually recover, often to a very considerable degree, although over months or years. In yet other patients, the condition may progress over a matter of some weeks or months.

If oxygen deprivation lasts longer than a few minutes, permanent or prolonged but reversible brain damage occurs. Irreversible coma is accompanied by flaccidity and loss of all reflex function except heart beat and tendon jerks. The pupils remain fixed and dilated and the electroencephalogram is flat on repeated examination. (Drugs and hypothermia may cause a flat electroencephalogram, but recovery is possible.) Such patients may be said to have suffered irreversible brain death if all signs of brainstem function are absent on repeated examination over 12 to 24 h. Other, less severely affected patients show partial recovery of brainstem reflex function, such as pupillary responses, reflex eye movements, and muscle tone, and may breathe spontaneously. However, no sign of consciousness or intelligent response to the external world occurs, and they may remain in such a 'persistent vegetative state' for months or years.

Subacute or gradual anoxia may occur in severe anaemia, heart failure, pulmonary disease, or exposure to high altitude ('mountain sickness'). It produces inattentiveness, fatigue, headache, and intellectual deterioration, followed by memory difficulties and ataxia.

Cerebral anoxia is also the main cause of neurological symptoms in a number of other systemic conditions. Disseminated intravascular coagulation (see Section 22), resulting from platelet aggregation and fibrin formation, can be produced by a number of illnesses, including sepsis and malignancy. Patients complain of headache and difficulty in concentration, vertigo, blurred vision, and speech difficulties. Such confusion and disorientation may progress to stupor and coma with focal or generalized signs of brain disturbance. Spontaneous bleeding is common, in the form of petechiae in the skin or optic fundus, purpura, and even intracranial haemorrhage. Cerebral malaria (see Chapter 7.13.2) should always be borne in mind as a cause of unexplained coma in patients recently returning from an infective area. Most patients describe chills and fever for a few days prior to the onset of lethargy, stupor, and finally coma. The diagnosis is established by finding the parasite in fixed smears of the blood.

Fat embolism follows severe trauma, particularly to the limbs, but may also be a complication of burns and other severe system disturbance. Multiple pulmonary microemboli of fat may lead to progressive hypoxia and respiratory failure. Multiple cerebral microemboli produce confusion, lethargy, stupor, and finally coma. Symptoms often begin hours to days after the original injury, and are accompanied by fever and hyperventilation. A characteristic petechial rash usually develops over the upper half of the body on the second to third day after injury. There may also be fundal haemorrhages. The respiratory features range from the appearance of linear streaks radiating from the hilar region or patchy opacities on the chest to the fully developed adult distress syndrome (see Section 16). Clotting abnormalities range from mild thrombocytopenia to acute disseminated intravascular coagulation (see Section 22). Management consists of correcting hypoxia, in severe cases with positive end-expiratory pressure (PEEP) ventilation, and correction of the coagulation disorder (see Section 16). Cardiac surgery, at least in the earlier days of bypass pump oxygenation, produced frequent transient neurological damage in many patients. Improvements in technique, such as the introduction of filters in blood perfusion lines to remove debris, have greatly reduced neurological complications of the procedure. However, some patients still emerge from the anaesthetic with signs of diffuse or focal brain damage. If they survive the acute episode, the prognosis usually is good.

Hepatic failure

Patients with liver disease of whatever cause (see Section 14) may develop acute hepatic coma or a more chronic form of hepatic encephalopathy with behavioural disturbance and other neurological symptoms.

Acute hepatic coma

This occurs with massive liver necrosis due to severe hepatitis or poisons such as paracetamol. In other patients, who may have relatively well-preserved liver function but extensive portosystemic shunts, coma may be precipitated by a sudden intake of nitrogenous substances as occurs with gastrointestinal bleeding, infections, or high-protein diets. Personality and cognitive changes proceed if unchecked to confusion, apathy, and lack of concentration, or occasionally excitement requiring sedation, and are rapidly followed by stupor and coma in a matter of a few hours or days. Characteristic findings are asterixis ('liver flap'), in which the outstretched hands show postural lapses or negative myoclonus, and hepatic fetor. Chorea and pyramidal signs may appear as the patient lapses into coma. Decerebrate posturing is common at this stage, and focal deficits such as hemiplegia may occur. Nystagmus, conjugate deviation of the eyes, skew deviation, and even disconjugate eye movements may be evident, but reflex eye movements and pupillary responses are preserved, until the patient becomes totally unresponsive and dies. Paroxysmal and later persistent high-voltage triphasic slow waves are present in the electroencephalogram until death is imminent.

Many metabolic abnormalities may contribute to the cause of hepatic coma, including hyperammonaemia (more than 145 µmol/l or 200 mg/dl), which results from the products of intestinal digestion bypassing the urea-synthesizing mechanisms of the liver. However, hypoglycaemia and hyperventilation producing a respiratory alkalosis are also nearly always present. Altered amino acids and neurotransmitters (especially a-aminobutyric acid), formation of toxic amines such as octopamine, and short-chain fatty acids have also been incriminated. Intravascular coagulation occurs, as do other coagulation defects, leading to secondary vascular damage to the brain; cerebral oedema can raise intracranial pressure seriously or even fatally.

Hepatic coma carries a high mortality, but if the patient can be kept alive, liver regeneration and recovery may occur. Treatment includes correcting where possible the precipitant, sterilizing the bowel, correction of metabolic and bleeding abnormalities, the administration of lactulose, and haemoperfusion or other techniques to remove toxins (see Chapter 14.21.3). The benzodiazepine antagonist flumazenil may have a useful role. Intracranial pressure monitoring is fraught with hazards, not least the coagulopathy, but mannitol (though not dexamethasone) is of proven benefit in lowering intracranial pressure in this context.

Chronic hepatic encephalopathy

This refers to the development of changes in intellect, cognitive function, and consciousness, often accompanied by other neurological signs (such as tremor or chorea, an akinetic–rigid syndrome, ataxia, or even spastic paraparesis) occurring in those with chronic liver failure, and particularly in those with extensive portosystemic anastomoses. (For Wilson's disease see Chapter 11.7.2.) The exact nature of the substances responsible for chronic hepatic encephalopathy has not been established. Characteristically, the disorder fluctuates, with episodes of marked confusion, excitement, or frank hepatic coma. In addition, intellectual changes, parkinsonism, ataxia, or spasticity may gradually progress. Treatment consists of a low-protein diet and antibiotics to sterilize the gut, and the administration of lactulose.

Respiratory disease

Hyperventilation causes hyocarbia and alkalosis, resulting in parasthesias, especially perioral, light-headedness and unsteadiness, visual disturbances, and occasionally carpopedal spasm; syncope may follow.

More seriously, chronic respiratory failure causes what is essentially a low-grade chronic hypoxia and hypercarbic encephalopathy, with the defining features of confusion and headache accompanied by a myoclonic or asterictic tremor and papilloedema. Mechanical devices for delivering domiciliary oxygen have transformed the management of this disorder, and the quality of life of its sufferers.

Obstructive sleep apnoea is characterized by conspicuous snoring and an often obese habitus. Early morning headache and inattentiveness or irritability with excessive daytime sleepiness should suggest this disorder.

Critical illness polyneuropathy

This disorder develops subacutely but often asymptomatically in (often anaesthetized) patients on intensive care units receiving intensive support for multiorgan failure and/or sepsis, only revealing itself as they otherwise improve. It is axonal in nature, but of still unknown aetiology. The prognosis is variable; it may slowly improve in patients whose underlying disease allows sufficient time for recovery.

Renal failure

Renal failure (see Section 20) is associated with a variety of neurological complications. Uraemic encephalopathy was common before the use of dialysis. Patients become progressively drowsy, stuporose, and finally lapse into coma. Hyperventilation, multifocal myoclonus, tremor, asterixis, tetany, and generalized fits are common. Eye movements and pupillary reactions are not affected. Uraemia, metabolic acidosis, hyperkalaemia, disorders of calcium, sodium, and water balance, and hypertensive encephalopathy all contribute to the clinical picture. Dialysis rapidly reverses the metabolic abnormalities of uraemia, but the encephalopathy may take days to clear. Other complications of chronic renal failure include myopathy due to chronic hypocalcaemia, and a symmetrical sensorimotor polyneuropathy, often subacutely progressive and disabling. It may be resistant to dialysis, but renal transplantation has been associated with a slow and sustained improvement.

Iatrogenic disease in renal failure

Some patients develop the dialysis disequilibrium syndrome during correction of their uraemic abnormalities. Rapid correction of the metabolic changes, possibly through osmotic shifts, leads to the emergence of asterixis, myoclonus, delirium, generalized convulsions, stupor, and even coma. Raised intracranial pressure with papilloedema may occur. Chronic dialysis—perhaps 3 to 7 years—may precipitate dialysis dementia if dialysate with a high aluminium content has been used. Such patients begin to develop speech hesitancy and arrest, then intellectual and cognitive abnormalities, convulsions, myoclonus, and sometimes focal neurological abnormalities. Death follows with a year.

Wernicke's encephalopathy (see below) can occur, due to chronic dialysis without thiamine supplements.

Patients with renal disease are particularly prone to develop toxic complications of drugs normally excreted in the urine—peripheral neuropathy due to nitrofurantoin, labyrinthine damage due to streptomycin, or optic atrophy due to ethambutol.

Metabolic disorders due to endocrine disease (seeSection 12)

Adrenal disease

Phaeochromocytoma

Phaeochromocytoma causes paroxysms of anxiety, tremor, headache, and palpitations, together with the consequences of malignant hypertension. Fits may occur. The associations with von Hippel–Lindau disease, multiple endocrine neoplasia syndromes, ataxia telangiectasia, and Sturge–Weber syndrome should not be overlooked.

Cushing's syndrome

Endogenous Cushing's syndrome in two-thirds of cases is due to a pituitary ACTH-secreting adenoma—conventionally termed Cushing's disease. Ectopic

ACTH-secreting malignant neoplasms and ACTH-independent adrenal tumours represent the other principal causes of endogenous disease; iatrogenic hyperadrenalism produces similar neurological symptoms. The systemic features are described in Chapter 12.7.1. Neurological complications include: (i) proximal myopathy, which can be severe and painful; (ii) psychiatric disorders, ranging from mild mood disturbance through moderate depression (common) to severe psychosis; (iii) a benign intracranial hypertension-like picture; and (iv) direct consequences of a pituitary tumour, particularly optic chiasmal compression.

Adrenal insufficiency

Hypoadrenalism due to primary adrenal failure (Addison's disease) or ACTH deficiency (from pituitary disease or chronic steroid treatment) causes weakness, lassitude, nausea and diarrhoea, and stupor or coma may be precipitated by surgical procedures or other acute illness. Hypotension (especially postural), hyponatraemia, hyperkalaemia, and often hypoglycaemia (see Chapter 12.11.1) occur: each may be symptomatic—indeed, attacks of hyperkalaemic periodic paralysis may occur. Amnesic deficits, depression, and impaired concentration progressing to confusion are relatively common. Addisonian crises may be accompanied by generalized convulsions, which are attributed to hyponatraemia and water intoxication. Benign intracranial hypertension with papilloedema and a proximal myopathy may also occur.

X-linked adrenoleukodystrophy is discussed in Section 12.7.

Thyroid disease

Thyroid disease carries one set of neurological complications directly related to abnormal thyroxine levels; another sharing the same autoimmune origin (and eponyms)—Hashimoto's encephalopathy and Grave's ophthalmopathy. Here only the former will be considered.

Thyrotoxicosis

The features of hyperthyroidism include anxiety, tremor, tachycardia, and insomnia. Chorea or mania may occur. A severe proximal myopathy is not uncommon, and rarely myasthenia gravis is seen. Thyroxine-responsive hypokalaemic periodic paralysis is well reported.

Myxoedema

Hypothyroidism may present with lethargy, even progressing to a toxic confusional state or a subacute hypothermic, hypotensive coma. The latter (which may be provoked by infection, trauma, exposure to cold, or sedation), together with the occasionally seen psychosis or dementing illness ('myxoedema madness') responds to (judicious) thyroxine hormone replacement. Ataxia occurs in 5 to 10 per cent of patients with hypothyroidism, and improves with thyroxine replacement.

Hypothyroid myopathy is characterized by proximal weakness with stiffness, aching, and cramps, and pseudomyotonic delayed muscle relaxation evident on tapping tendons or muscle bellies (with percussion-induced muscle ridging). Muscle hypertrophy (Hoffmann's syndrome) is rare. The carpal tunnel syndrome may occur due to deposits of myxoedematous tissue around the median nerve of the wrist, and rarely this may cause a diffuse peripheral neuropathy.

Diabetes mellitus

Diabetes mellitus (Chapter 12.11.1) causes a wide variety of neurological disturbances. Centrally, stupor or coma may be produced by hyperosmolality, ketoacidosis, lactic acidosis, spontaneous (prediabetic) or iatrogenic hypoglycaemia, uraemia, or hypertensive encephalopathy. Transient ischaemic attacks and stroke due to cerebral arteriosclerosis and hypertension are common in patients with diabetes.

Peripherally, nerve damage may occur in patients with established diabetes, or may be the presenting feature of the illness; it is described in more detail in Chapter 24.19. The following syndromes are recognized.

1. Single painful nerve lesions (mononeuritis) such as isolated ocular nerve palsies, Bell's palsy, a lateral popliteal nerve palsy, or an intercostal neuropathy are common and may result from haemorrhage or infarction of the nerve.
2. Carpal tunnel syndrome, an ulnar nerve palsy, or other compression neuropathies may result from the undue susceptibility of peripheral nerves in diabetes to pressure.
3. Mononeuritis multiplex may occur, with a microvascular basis.
4. Diabetic amyotrophy refers to a proximal motor neuropathy causing the subacute weakness and wasting, often with pain, affecting quadriceps muscles, usually asymmetrically. It is probably due to ischaemia or haemorrhage in the femoral nerve or lumbosacral plexus.
5. A distal symmetrical peripheral neuropathy in diabetes may take the form of a mild asymptomatic sensory neuropathy with loss of vibration sense in the feet and absent ankle jerks. Less commonly, there is severe and progressive sensorimotor neuropathy affecting the legs before the arms.
6. Autonomic neuropathy is common, producing impotence, diarrhoea, loss of sweating, and abnormal pupils. The last may be irregular, and unreactive to light, mimicking Argyll Robertson pupils. Autonomic neuropathy causes orthostatic hypotension, syncope, and sometimes abrupt cardiac arrest in patients with diabetes.

It should be recalled that diabetes may occur as a feature of a number of genetically determined neurological diseases, including Friedreich's ataxia, X-linked spinomuscular atrophy, mitochondrial cytopathies, myotonic dystrophy, and the Wolfram syndrome; it is also associated with the stiff man syndrome.

Hypoglycaemia

Hypoglycaemic coma can be difficult to diagnose and dangerous. In any case of coma, stupor, or confusion of unknown cause, and often in newly presenting status epilepticus, blood should be drawn for glucose analysis and insulin levels, and then 25 g of glucose (with thiamine) should be administered intravenously. Such an injection can do no harm and may save life.

The commonest cause of hypoglycaemia is insulin overdose, or excessive hypoglycaemic drug intake. Hyperinsulinism due to an adenoma of the islets of Langerhans in the pancreas is uncommon, as is hypoglycaemia due to prediabetes or a retroperitoneal sarcoma. Hypoglycaemia may also occur in alcoholism and liver disease, after gastric surgery, and in a variety of rare metabolic conditions.

Hypoglycaemia presents in four ways: (i) as an organic toxic confusional state, sleepy confusion, bizarre behaviour, or mania; (ii) as unexplained coma with brainstem dysfunction, including decerebrate spasms and neurogenic hyperventilation, but with preserved oculocephalic reflexes and pupillary responses; (iii) as a stroke-like illness with focal deficit; and (iv) as epilepsy. Hyperinsulinism, very rarely, also causes predominantly motor peripheral neuropathy.

Hypoglycaemia is established by measurement of the blood glucose concentration, and by clinical response to intravenous glucose replacement. Hyperinsulinism is difficult to diagnose on occasion, but can be established by satisfying the criteria for Whipple's triad, namely symptoms of hypoglycaemia, associated with a low blood sugar and a disproportionately high serum insulin, and clinical response to glucose replacement. A 72-h fast, measuring morning blood sugar and insulin levels, will detect nearly all pancreatic islet cell adenomas (see Chapter 12.10).

Metabolic disorders due to ionic or acid–base abnormalities

Hyponatraemia or 'water intoxication'

Sodium is the most abundant serum cation, so that hyponatraemia is almost always the cause of hypo-osmolality. Serum osmolality is approximately equal to double the serum sodium concentration plus 10, provided glucose and urea levels are normal. Normal serum osmolality is 290 ± 5 mosmol/kg; serum osmolality below about 260 or above about 330 mosmol/kg is likely to produce cerebral changes. Hyponatraemia means that body water is increased relative to solute, resulting in water excess in the brain. Rapid changes in serum sodium osmolality produce much greater neurological effects than does slowly developing chronic hyponatraemia. Hyponatraemia occurs in renal disease, as a result of excessive intravenous water infusions, due to excessive diarrhoea, vomiting, or sweating, or may result from the inappropriate secretion of antidiuretic hormone that occurs in bronchial carcinoma, focal hypothalamic damage due to neoplasm or infection, or diffuse acute brain

disease resulting from head injury, meningitis, or encephalitis, or subarachnoid haemorrhage. (It is, however, noteworthy that in the latter acute situations, salt-wasting may also cause hyponatraemia, in which circumstances fluid restriction exacerbates the problem: hypovolaemia distinguishes salt-wasting from the eu- or hypervolaemia of vasopressin excess.) Patients with hyponatraemia become confused and restless, and develop asterixis, multifocal myoclonus, generalized convulsions, stupor, and coma. Symptoms may appear when the plasma sodium drops below about 120 mmol/l, and fits and coma usually are associated with plasma sodium values below 110 mmol/l. A few patients with chronic hyponatraemia may develop the syndrome of central pontine myelinolysis (see below). Treatment is by water restriction; infusions of hypertonic saline are not advised.

Central pontine myelinolysis

This is a rare disease, often associated with hyponatraemia, and in particular with rapid attempts to correct serum sodium by parenteral hypertonic fluids: elevation by no more than 0.55 mmol/l per hour is allegedly safe. It is also seen in alcoholics, in severe liver and renal disease, and other metabolic disturbances. The disease is characterized by a rapidly progressive flaccid or spastic quadriplegia, with involvement of bulbar muscles producing dysarthria and dysphagia. Consciousness and eye movement may remain intact. At worst the patient may be unable to speak or swallow, or to move any muscle except those of the eyes. Death is common but remarkable recovery may occur.

Hypernatraemia

The common cause of hyperosmolality is diabetes, producing severe hyperglycaemia. Hyperosmolality due to hypernatraemia is rare, except in those who dehydrate in hot climates. Chronic uncompensated water loss in untreated diabetes insipidus may result in mild hypernatraemia, but such patients only develop severe hypernatraemia if they fail to drink. Patients with simple diabetes insipidus usually maintain thirst, but if intercurrent illness leads to excessive water loss and restricted water intake, they may become dehydrated, drowsy, stuporose, and unconscious. Simple diabetes insipidus may be due to pituitary surgery, trauma, or pituitary tumours. If pathology extends into the hypothalamic region, not only may secretion of vasopressin be deficient, but thirst regulation may also be abolished. Hypothalamic damage causing severe hypernatraemia may occur in large pituitary tumours, craniopharyngiomas, hypothalamic tumours, sarcoidosis, or Hand–Schüller–Christian disease. Loss of thirst in such patients often precipitates hypernatraemic coma with serum sodium rising above 160 to 170 mmol/l. Hypernatraemia may also occur as a result of severe water depletion, particularly in children with intense diarrhoea.

Hypercalcaemia (see Chapter 12.6)

A high serum calcium concentration may be due to primary hyperparathyroidism, immobilization, sarcoidosis, vitamin D intoxication, or multiple bony metastases. Symptoms include anorexia, nausea, vomiting, intense thirst, polyuria, and polydipsia. Muscle weakness, lassitude, and a mild encephalopathy are common. The latter may produce delusions and changes in mood so that many such patients are initially treated for a psychiatric condition. A toxic confusional state with lethargy and stupor, sometimes with generalized or focal seizures and papilloedema, also may occur. A more severe syndrome with pyramidal signs, ataxia, and an internuclear ophthalmoplegia is also described.

Hypocalcaemia (see Chapter 12.6)

Reduced serum calcium concentration may be caused by parathyroid or thyroid surgery, chronic renal failure, or chronic anticonvulsant drug treatment. It also occurs in primary idiopathic hypoparathyroidism (when the serum parathormone level is low), and in pseudohypoparathyroidism (in which the serum parathormone level is normal or high, and there is no response to parathyroid hormone; skeletal deformities and dysmorphism also are present). Pseudopseudohypoparathyroidism is a syndrome with similar skeletal and dysmorphic abnormalities but normal serum calcium and parathormone levels. Hypocalcaemia causes neuromuscular irritability, tetany with a positive Chvostek's sign, and a mild encephalopathy. Severe degrees of hypocalcaemia produce generalized convulsions, psychotic behavioural disturbances, stupor, and coma. Raised intracranial pressure with papilloedema may occur in hypoparathyroidism. Hypocalcaemia is commonly misdiagnosed as mental retardation, dementia, or epilepsy. Skin changes and cataracts are characteristic. Calcification in basal ganglia on skull radiograph or CT scan may be evident. Rarely, basal ganglia calcification may be associated with extrapyramidal disorders.

Magnesium

Renal disease may impair the ability to excrete magnesium, which is cardiotoxic. Hypomagnesaemia, due to inadequate intake or excessive renal or gastrointestinal loss, causes secondary hypocalcaemia; the former rarely occurs without the latter, and the neurological complications often attributed to low magnesium are precisely those of hypocalcaemia. Hypermagnesaemia may cause an encephalopathy with decreased or absent tendon jerks; the latter may progress to a flaccid paralysis.

Potassium

Hypokalaemia, not uncommonly caused by diuretics, is associated with myalgia and a proximal myopathy; if severe, rhabdomyolysis can occur. Hyperkalaemia can precipitate an areflexic flaccid paralysis, which may be fully reversible with correction of the serum potassium. The periodic paralyses are discussed in Chapter 24.22.5.

Acid–base disturbances

Systemic acidosis and alkalosis (see Section 11) occur in many diseases causing metabolic coma, but of the four disorders of acid–base balance (respiratory or metabolic acidosis or alkalosis), only respiratory acidosis acts as a direct cause of stupor and coma. Hypoxia associated with respiratory acidosis may be important in producing neurological abnormalities. Metabolic acidosis, by itself, usually only causes delirium or, at most, drowsiness. The reason why even severe disorders of systemic acid–base balance usually do not interfere with the function of the brain is that it possesses powerful mechanisms for protecting its own acid–base balance, including respiratory compensation, changes in cerebral blood flow, and cellular buffering in nervous tissue. Coma in metabolic acidosis due to diabetic ketosis or hyperosmolality, lactic acidosis, uraemia, alcohol poisoning, or intake of ethylene glycol, methyl alcohol, or paraldehyde is usually due to associated metabolic abnormalities or direct effects of other toxins in these conditions. Severe respiratory acidosis produces a reduction in alertness parallel to the degree of acidosis. Respiratory alkalosis, although constricting cerebral arterioles and decreasing cerebral blood flow, rarely interferes with cerebral function. A patient in coma with respiratory alkalosis due to hyperventilation has some other condition such as sepsis, hepatic disease, pulmonary infarction, or salicylate overdose. Even severe metabolic alkalosis only produces a confusional state rather than stupor or coma.

Alcohol and the nervous system

Alcohol damages the nervous system in many ways. Some are the result of acute or chronic poisoning, while others are a consequence of associated vitamin deficiency. This section will mainly address the neurological consequences of chronic, excessive alcohol intake, not the acute transient effects of alcohol.

Delirium tremens

This develops several days after ethanol abstinence in chronic abusers. Usually rapid in onset, there is an agitated confused state, with signs of sympathetic overactivity. Circulatory collapse may contribute to the 5 to 10 per cent mortality. It is generally distinguished from the less severe alcohol withdrawal syndrome, characterized by broadly similar symptoms that occur sooner—within hours of withdrawal—and are usually self-limiting. Ethanol withdrawal seizures represent a not uncommon cause of late-onset fits. Benzodiazepines have transformed the management of delirium tremens and the alcohol withdrawal syndrome, and significantly reduced its mortality.

The Wernicke–Korsakoff syndrome

Aetiology

Inadequate intake of thiamine, of whatever cause, may lead to foci of marked hyperaemia with multiple small haemorrhages affecting particularly the upper brainstem, hypothalamus, and thalamus adjacent to the third ventricle, and the mamillary bodies. Histologically there is a proliferation of dilated capillaries with perivascular

haemorrhage in these areas. There may be associated alcohol-induced damage to the cerebral cortex, cerebellum, and peripheral nerves.

Such pathology can be produced in animals by a diet deficient in thiamine. Thiamine deficiency can be demonstrated in patients with the Wernicke–Korsakoff syndrome, and administration of thiamine can reverse many of the symptoms and signs of this syndrome. (Wernicke's and Korsakoff's syndromes probably represent the acute and chronic consequences of the same pathological process.) Thiamine and its pyrophosphate are cofactors to at least four enzymes—pyruvate decarboxylase, a-ketoglutarate dehydrogenase, the branched-chain ketoacid decarboxylase system, and transketolase. Thiamine deficiency results in reduced conversion of pyruvate to acetyl coenzyme A, causing elevated plasma and tissue pyruvate levels, with decreased flux through the Krebs cycle, reducing ATP production, and impairing energy supply. In addition, there is a shortage of one-carbon groups for biosynthetic pathways.

Alcoholism with an inadequate diet is the most frequent cause of the Wernicke–Korsakoff syndrome today. Malnutrition in prisoners of war, or at times of famine, may also be responsible. Chronic vomiting, for example during pregnancy or due to gastrointestinal disease, systemic malignancy, prolonged intravenous feeding, and anorexia nervosa are rarer causes.

Clinical features

The onset may be insidious or subacute with increasing lethargy and inattentiveness, which develops into a typical confusional state with disorientation in time and place, loss of memory, and altered consciousness. Ophthalmoplegia develops with diplopia. The most common eye signs are nystagmus on lateral or vertical gaze, sixth-nerve palsies, or defects of conjugate gaze. Retinal haemorrhages may occur. Most patients who are alcoholics will also have signs of a peripheral neuropathy, and many exhibit ataxia—the third classically described feature. Hypothermia may appear. Wernicke's encephalopathy is a medical emergency: untreated, the patient lapses into stupor and then coma, and dies—the mortality untreated is 20 per cent.

In less acute cases, or in those recovering from the acute confusional phase, the characteristic features of the Korsakoff psychosis or amnesic syndrome will appear. The patient has an often very severe gross defect of memory for recent events, such that new information cannot be retained for more than a matter of minutes or hours. The patient is disorientated in time and place, but alert. Despite the severe defect of recent memory, he or she can recall events in the remote past. Gaps in memory are filled by giving imaginary and often graphic accounts of events (confabulation).

Diagnosis

Diagnosis is essentially clinical. The cerebrospinal fluid is usually normal, although the protein may be slightly raised. Brain scanning can be normal, although patients who are alcoholic (who fall often) may have subdural haematomas. Demonstration of reduced red cell transketolase activity, or of raised plasma pyruvate levels, is often invalidated by intake of food as soon as the patient comes under supervision.

Treatment

The Wernicke–Korsakoff syndrome must be considered in any individual with unexplained confusion, stupor, or coma, particularly in the presence of eye signs, a peripheral neuropathy, or a history of alcoholism or excessive vomiting. Thiamine should be given to all such patients. High-potency vitamin injections should be given daily until oral vitamin B complex preparations can be taken.

A particular problem arises commonly in the emergency department when patients exhibiting stupor or coma of unknown cause are admitted. All such patients should be given high-dose thiamine and glucose parenterally—if glucose is given without thiamine to a patient with Wernicke–Korsakoff syndrome, rapid deterioration and death can follow. Those who are thiamine deficient cannot handle the glucose load.

Effective treatment will restore consciousness and reverse eye signs, the latter usually within hours, but unfortunately the Korsakoff amnesia syndrome frequently does not resolve. The earlier the treatment, the better the chances of recovery, so suspicion or possibility of this diagnosis represents a medical emergency.

Alcoholic peripheral neuropathy (see also Chapter 24.17)

Aetiology

Although thiamine deficiency has long been held responsible for the peripheral neuropathy associated with chronic alcoholism, a direct toxic effect has more recently been proposed. Pathologically the picture of peripheral nerve damage is very similar to that seen in beriberi. There is predominantly axonal neuropathy of the 'dying back' type, affecting the somatic and sometimes the autonomic nerves.

Clinical features

Alcoholic peripheral neuropathy predominantly involves sensory nerves, producing distal parasthesias in the feet, followed by the hands, and characteristic pain. The last may be intense and agonizing. Squeezing the calves or scratching the soles of the feet may cause severe discomfort. At a later stage, weakness and wasting of the distal limb muscles follows. Tendon reflexes are lost. Evidence of autonomic neuropathy may be seen in abnormal pupillary reactions and tachycardia, although postural hypotension is rare and the sphincters are usually spared.

Treatment

Alcohol must be proscribed and high-potency vitamin B given parenterally for some 10 days and then orally. The prognosis depends on how early treatment is initiated. Symptoms may take weeks to subside and in more severe cases recovery may take many months or may be incomplete.

Alcoholic cerebellar degeneration

Some patients who are alcoholic may develop a relatively pure syndrome of midline cerebellar ataxia, with a progressive unsteadiness of gait and of leg movements, and little or no involvement of the arms. Speech is not affected, and nystagmus is not present. Many such patients also have evidence of alcoholic peripheral neuropathy. Pathologically there is degeneration of the cerebellar cortex, particularly of the Purkinje cells, and also of the olivary nuclei. Changes in the cerebellum characteristically affect the anterior and superior parts of the vermis and hemispheres. This complication of alcoholism does not seem to be due to thiamine deficiency. However, withdrawal of alcohol and vitamin replacement can lead to recovery.

Alcoholic dementia (see Section 26)

In the past, there has been much debate over whether alcoholism produces dementia. However, it is now clear that a large proportion of those who habitually take excessive alcohol develop cognitive deficits. These can vary from mild changes to severe diffuse global dementia, and are associated with atrophy of the cerebral cortex and enlargement of the cerebral ventricles.

The dementia has the usual features of personality change, loss of memory, impairment of intellect, and emotional instability. Patients who are alcoholic commonly fail at work or in personal relationships. The gradual drift into destitution is well described in the literature and all too familiar on the streets. Head injuries in alcoholic bouts and epilepsy may occur and contribute to the overall final picture. The fully developed case of the 'down and out' is an antisocial demented individual, with dysarthric speech, tremor, an ataxic gait, and a peripheral neuropathy, who still forlornly or aggressively clutches the bottle and a bag of residual belongings.

The dementia of alcoholism is not directly related to thiamine deficiency alone. Treatment by withdrawal of alcohol, if possible, and vitamin replacement can lead to improvement. Indeed, some degree of reversal of evidence of cerebral atrophy on CT brain scan can be seen after 'drying out'. However, the prognosis is generally poor, not least because of the difficulties of persuading those addicted to alcoholic to stop drinking.

Marchiafava–Bignami disease

This rare disease was first described in Italian drinkers of crude red wine, but occurs in other patients who are alcoholic. It presents as a subacute dementing illness, which progresses rapidly to fits, spasticity or rigidity, and paralysis, culminating in coma and death within a few months. Pathologically there is widespread demyelination and axonal damage in the corpus callosum and the central white matter of the cerebral hemispheres, as well as in the optic chiasma and middle cerebellar peduncles. Abstinence stabilizes but rarely reverses the syndrome.

Alcoholic myopathy

Acute alcohol poisoning can produce a dramatic toxic myopathy. There is severe pain, muscle tenderness, oedema, and weakness, which may be associated with myoglobinuria, renal damage, and hyperkalaemia. Arrhythmias may occur. The syndrome is reversible if the necessary intensive support is available. A subacute painless myopathy resolving after withdrawal of alcohol has also been described. Chronic alcoholism is associated commonly with a painless myopathy, occasionally with coexistent cardiomyopathy; again abstinence can cure the disorder.

Tobacco–alcohol amblyopia

Another uncommon complication of alcohol occurs in combination with strong tobacco. The patient develops sudden or subacute bilateral visual failure, associated with bilateral centrocaecal scotomas. The condition has been attributed to cyanide in tobacco causing a disorder of vitamin B_{12} metabolism. Visual failure and optic atrophy may occur in patients with pernicious anaemia, particularly those who smoke. A related condition is tropical amblyopia, occurring in Africa. This has been related to excessive consumption of cassava root containing cyanide. Treatment of these conditions is with hydroxycobalamine injections.

Superficial siderosis of the central nervous system

Superficial siderosis is an unusual disorder of the nervous system that has been recognized only recently. The four principal clinical manifestations are: progressive ataxia; cranial polyneuropathy, particularly sensorineural deafness; myelopathy causing a spastic tetraparesis; and progressive dementia. Headaches occasionally feature. Cerebrospinal fluid examination reveals xanthochromia which, importantly, persists with repeated examination. MRI is diagnostic, with low signal intensity on T_2-weighted images apparent at the surface of the cerebellum, cranial nerves, brainstem, spinal cord, and more deeply on the borders of the dentate and basal ganglia. Iron deposition can be shown by high-strength MRI, corresponding to pathological descriptions of siderotic deposits in the meninges. Repeated subarachnoid haemorrhage, cerebral tumours, or past surgery are recognized causes of this syndrome, but often none is historically evident; repeated subclinical haemorrhage is postulated but not proven. No treatments are of proven benefit.

Porphyria

Porphyrias affect predominantly the liver (acute intermittent porphyria (AIP), variegate porphyria, hereditary coproporphyria, and porphyria cutanea tarda) or the blood (for example erythropoietic protoporphyria). AIP spares the skin. Certain drugs (sulphonamides, barbiturates, oral contraceptive pill), starvation, alcohol, and other insults can precipitate 'crises' in AIP, hereditary coproporphyria, and variegate porphyria characterized by: (i) a predominantly motor, often proximal areflexic peripheral neuropathy, occasionally mimicking the Guillain–Barré syndrome; (ii) abdominal pain and cardiovascular instability caused by autonomic involvement; and (iii) confusion or psychosis. Fits may occur and acute attacks are commonly accompanied by progressive severe hyponatraemia due to inappropriate secretion of vasopressin. Increased urine and faecal porphyrins lead to the diagnosis. Acute attacks are treated largely symptomatically with benzodiazepines or opiates and major tranquillizers, and cardiorespiratory support. Carbohydrate loading may be beneficial.

---

*It is with regret that we must report the death of Professor C.D. Marsden since the publication of the third edition of this textbook. Much of his text for that edition has been retained here.

Further reading

Metabolic complications of major organ disease

Bolton CF, Young GB (1990). *Neurological complications of renal diseases.* Butterworth Heinemann, Stoneham, Massachusetts.

Jones EA, Weissenborn K (1997). Neurology and the liver. *Journal of Neurology, Neursurgery, and Psychiatry* **63**, 279–303.

Metabolic disorders due to endocrine disease

Shaw P (1998). Neurological complications of thyroid disease. In: Goetz CG, Aminoff MJ, eds. *Handbook of clinical neurology* **26** (70) *Systemic diseases Part II*, pp. 81–110. Elsevier Science BV, Amsterdam.

Watkins PJ, Thomas PK (1997). Diabetes mellitus and the nervous system. *Journal of Neurology, Neursurgery, and Psychiatry* **65**, 620–32.

Metabolic disorders due to ionic or acid–base abnormalities

Abrams GM, Jay C (1998). Neurological complications of mineral metabolism and parathyroid disease. In: Goetz CG, Aminoff MJ, eds. *Handbook of clinical neurology* **26** (70) *Systemic diseases Part II*, pp 111–129.

Gocht A, Colmant HJ (1997). Central pontine and extrapontine myelinolysis: a report of 58 cases. *Clinical Neuropathology* **6**, 262–70.

Alcohol and the nervous system

Harper C (1983). The incidence of Wernicke's encephalopathy in Australia—a neuropathological study of 131 cases. *Journal of Neurology, Neurosurgery, and Psychiatry* **46**, 593–8.

Harper C, Giles M, Finlay-Jones R (1986). Clinical signs in the Wernicke–Korsakoff complex: a retrospective analysis of 131 cases diagnosed at necropsy. *Journal of Neurology, Neurosurgery, and Psychiatry* **49**, 341–5.

Miles MF, Diamond I (1998). Neurological complications of alcoholism and alcohol abuse. In: Goetz CG, Aminoff MJ, eds. *Handbook of clinical neurology* **26** (70) *Systemic diseases Part II*, pp 339–57.

Miscellaneous metabolic and deficiency disorders of the nervous system

Bruyn RPM, Bruyn GW (1998). Superfical siderosis of the central nervous system. In: Goetz CG, Aminoff MJ, eds. *Handbook of clinical neurology* **26** (70) *Systemic diseases Part II*, pp 65–80.

Young GB (1995). Neurologic complications of systemic critical illness. *Neurologic Clinics* **13**, 645–58.

24.16 Demyelinating disorders of the central nervous system

Alastair Compston

Clinicians suspect demyelination when episodes reflecting damage to white matter tracts within the central nervous system occur in young adults. The diagnosis of multiple sclerosis becomes probable when these symptoms and signs recur, affecting different parts of the brain and spinal cord. Demyelination also underlies many postinfectious neurological conditions affecting the central nervous system.

Neurobiology of demyelination

Glial development and myelination

Glial progenitors migrate from germinal zones around the lateral ventricles, the fourth ventricle, and in the ventral spinal cord, and differentiate either into astrocytes or oligodendrocytes. Oligodendrocyte precursors can be recovered from the adult nervous system. These behave as stem cells, dividing asymmetrically (at least *in vitro*) to produce one daughter precursor cell and one oligodendrocyte—providing a potential pool of new oligodendrocytes.

Growth, differentiation, and survival of glial progenitors and their progeny are orchestrated by growth factors. These are produced by neurones, astrocytes, and microglia. Those factors involved in rodent development (where most is currently known: glial-derived nerve growth factor, fibroblast growth factor 2, platelet-derived growth factor, insulin-like growth factors 1 and 2, nerve growth factor, neurotrophin 3, ciliary neurotrophic factor, retinoic acid, glial growth factor, interleukin 6, and leukaemia inhibitory factor) are not yet shown to be relevant in development of the human central nervous system.

Myelination occurs when the membraneous processes of mature oligodendrocytes contact and ensheathe axons and compact to form the myelin lamellae needed for saltatory axonal conduction. The number of surviving oligodendrocytes is matched to local axon density. Compact myelin consists of a condensed membrane, mainly composed of lipid (cholesterol, phospholipid, and galactolipid) with some protein, wrapped spirally many times around axons to form a segmented sheath. The glycoproteins are galactocerebroside, myelin-associated glycoprotein, and myelin oligodendrocyte glycoprotein (**MOG**). The two major proteins are proteolipid protein (**PLP**) and myelin basic protein (**MBP**). A further structural component is the myelin-specific enzyme 2',3'-cyclic nucleotide 3'-phosphohydrolase (CNP-ase). It is periodically interrupted along the course of the axon at the (unmyelinated) nodes of Ranvier, where electrical resistance is low due to the high concentration of sodium channels, and depolarization thereby facilitated. In myelinated axons, the action potential generates electrical currents which preferentially trigger depolarization at the next node of Ranvier. This saltatory conduction is considerably more rapid than continuous propagation of the nerve impulse.

Pathophysiology of demyelination

Myelin injury blocks saltatory conduction through myelinated pathways in the central nervous system. Although function may be preserved by redundancy in individual systems or tracts, strategically placed pathways lose their safety factor for conduction resulting in neurological symptoms and signs. Most clinical manifestations of demyelination merely reflect abnormalities to be expected from any process that disrupts physiological performance at that site. However, saltatory conduction may be compromised by partial demyelination in ways which account for specific features of multiple sclerosis and related disorders.

Partially demyelinated axons cannot transmit fast trains of impulse. This may explain symptoms that reflect physiological fatigue. Depolarization may traverse the lesion but at reduced velocity. This accounts for the characteristic delay in arrival of potentials evoked by sensory stimuli and recorded over appropriate cortical receptor zones. Partially demyelinated axons may discharge spontaneously. This explains distortions of sensation reported by a high proportion of patients. Increased mechanical sensitivity manifests as movement-induced symptoms including flashes of light on eye movement, and the electric sensation that spreads down the spine, limbs, or anterior chest wall after neck flexion—Lhermitte's symptom and sign. Increased temperature sensitivity, with a reduction in the safety factor for conduction in partially demyelinated axons, explains the temporary increase in severity of pre-existing symptoms experienced by many patients after exercise or immersion in hot water. Cold may improve performance—some patients adopting complicated water-cooled systems and others reporting that, for example, vision improves after eating ice cream. Ephaptic transmission occurs between neighbouring and partially demyelinated axons giving rise to paroxysmal symptoms of demyelination usually manifesting as trigeminal neuralgia, ataxia, and dysarthria, or tonic brainstem seizures. These are often triggered by touch or movement.

There are several mechanisms of symptom recovery early in the course of multiple sclerosis. These include the resolution of conduction block in nerve fibres which were never demyelinated, re-establishment of conduction in persistently demyelinated axons, functional reorganization of surviving pathways, and remyelination. Onset and recovery of conduction block and clinical impairments match the phase of acute inflammation. Transient symptoms depend on reversible conduction block caused by direct action of cytokines and inflammatory mediators (especially nitric oxide) on normal or hypomyelinated axons (Fig. 1). Function may be restored after demyelination by rearrangement of sodium channels providing a variety of alternative patterns of ordered or partially disordered conduction. There is probably also a contribution from the remyelination seen in acute lesions. Experimentally, remyelinated axons restore conduction of the nerve impulse and motor function.

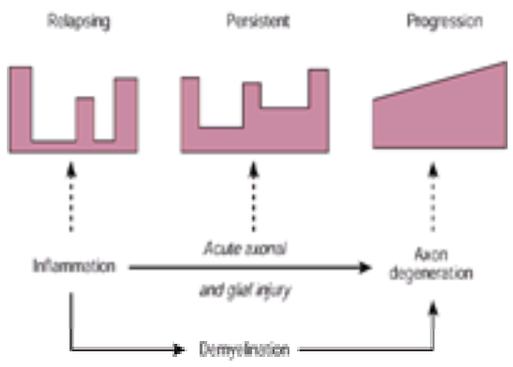

Fig. 1 Inflammatory mediators cause reversible symptoms in multiple sclerosis. Inflammation drives demyelination, which explains persistent symptoms. Progression depends mainly on axon degeneration, which causes acute axonal inury and continues throughout the course, but is conditioned by the amount of early inflammation.

Inflammation and the brain

Demyelinating disease includes conditions in which myelin fails to develop (leucodystrophies) or is lost through inflammatory (multiple sclerosis and related conditions) and metabolic (central pontine myelinolysis) mechanisms. Outstandingly the most common group is inflammatory demyelination. In all but the most severe forms, perivascular inflammation evolves through stages of acute axonal injury, demyelination, oligodendrocyte depletion, remyelination, astrocytosis, and axon chronic degeneration. The order and relationship of these separate components is still debated. The resulting plaques are widely distributed but concentrated around venous networks, the ventricles, and in the corpus callosum, the optic nerve, brainstem, and cervical cord.

Conditions required for inflammatory cell penetration of cerebral vessels include slowing of the circulation and the establishment of electrostatic force which allows adhesion between circulating and lining cells. Endothelial cells extend microvillar processes in response to injury and these entangle inflammatory cells as they pass along the vessel wall. Infiltrating lymphocytes which are not activated against brain antigen either return to the circulation or (in common with immune cells that have outlived their purpose elsewhere) die by apoptosis. Activated T cells that encounter antigen persist within the nervous system.

Outward migration of cells from the inflammatory nidus is promoted by local production of chemokines interacting with specific receptors on migrating cells, and by metalloproteases which degrade tissue barriers. Proinflammatory cytokines (especially interferon-g) amplify the immune response. Microglia are activated, leading to the release of yet more T-cell derived interferon-g, and the recruitment of additional naive microglia. Contact is established between activated microglia and the oligodendrocyte–myelin unit if the latter is opsonized with ligands for (Fc and complement) receptors activated on the surface of microglia. Demyelinated axons are coated with anti-MOG antibody in the lesions of acute multiple sclerosis. This may be a key antigen in attracting degradation of the oligodendrocyte–myelin unit by activated microglia. Adherent activated microglia deliver their lethal signal to the target oligodendrocyte using tumour necrosis factor-a bound to the cell surface. Together, these inflammatory processes lead to disruption of the myelin membrane with increased spacing, vesicular disruption, splitting, vacuolation, and fragmentation of the lamellae.

Much emphasis has been placed on the role of axon degeneration as a pathological feature. In hyperacute multiple sclerosis, large confluent zones of demyelination are associated with extensive axonal loss and surrounding oedema but little inflammation. In many other lesions, immunohistochemical staining for the amyloid precursor protein shows that axonal injury is initiated as part of the acute demyelinating episode. However, it is not certain when axons actually die. Acute damage to axons with transection appears early and the circumstantial evidence suggests vulnerability of recently demyelinated axons to the inflammatory environment of acute lesions; but there is also a chronic attrition which may be degenerative and secondary to loss of trophic support normally provided by myelin.

Acute lesions sometimes show an increase in the number of oligodendrocytes indicating microglia-associated loss of healthy oligodendrocytes and recruitment of new progenitors which then undergo differentiation. Axons are remyelinated in acute shadow plaques. Remyelination is associated with inflammation and reactive astrocytes which deliver cytokines and growth factors. The morphological criteria for remyelination are inappropriately thin myelin lamellae for the corresponding axon, with a short internode and myelin embedded in a satellite cell. Experimentally, remyelination can restore structure, conduction of the nerve impulse, and function, but—in a clinical context—new myelin may not survive repeated injury. The source of remyelinating cells is presumed to be the oligodendrocyte progenitor which is found in the lesions of multiple sclerosis.

Isolated demyelinating syndromes

The clinical expression of demyelination may be focal and monophasic even when imaging shows multiple lesions. The distinction between multiple sclerosis and isolated demyelinating disorders can therefore reliably only be made when more than one episode has occurred, affecting two or more sites, and not merely on the basis of anatomical dissemination of lesions.

Acute disseminated encephalomyelitis

Typically, acute disseminated encephalomyelitis develops within days or a few weeks after an infectious illness. It is usually but not invariably a disease of children. Formerly, acute disseminated encephalomyelitis affected 1 in 1000 children with exanthematous illnesses, the risk being slightly lower following pertussis and scarlet fever than measles and rubella, but these childhood illnesses, and hence their complications, are now less prevalent. A greater variety of causative organisms has been implicated in adult-onset acute disseminated encephalomyelitis, but in both groups a presumptive diagnosis often has to be made in the absence of an identifiable preceding infection.

The disorder is usually diffuse, and with a cerebral flavour, but the clinical manifestations may be restricted to the brainstem, optic nerves, or spinal cord. About 50 per cent of cases occurring after varicella infection present with a pure cerebellar syndrome. Headache, drowsiness, meningeal irritation, signs of systemic infection, focal or generalized fits, and combinations of lesions indicating damage to the cerebrum, optic nerves, brainstem, or spinal cord evolve over the course of a few days. The cerebrospinal fluid contains a mixture of polymorphonuclear cells and lymphocytes with raised protein and slight reduction in glucose; oligoclonal bands may be present. Whilst there is an appreciable mortality, the majority of patients survive, sometimes with persistent neurological deficits. Magnetic resonance imaging shows changes similar to those occurring in multiple sclerosis but the lesions are more extensive and symmetrical; they persist long after recovery of the clinical illness.

The hyperacute form of acute disseminated encephalomyelitis (Hurst's disease) starts with headache and progresses over hours to disorientation, confusion, drowsiness, and coma; events move quickly and the illness often proves fatal before the diagnosis has been established. The combination of pyrexia and a marked cerebrospinal fluid pleocytosis with a predominantly neutrophil response mimics pyogenic infection of the central nervous system, but the course is not influenced by antimicrobial treatment. Occasionally, the clinical and pathological features of acute haemorrhagic leucoencephalitis are focal and suggest a rapidly expanding tumour or herpes simplex encephalitis.

The outcome in acute disseminated encephalomyelitis is probably influenced by early use of high-dose intravenous steroids, but anecdotally, there may be a more favourable response to intravenous immunoglobulin. A proportion of patients recovering from the initial attack subsequently relapse. In some, although the illness remains monophasic, separate sites are involved sequentially over several weeks but the disorder does not recur. In others, the illness is subsequently shown to be the encephalopathic presentation of multiple sclerosis, which then follows the typical relapsing–remitting course. The nosological status of multiphasic disseminated encephalomyelitis—based on a history of episodes and atypical imaging appearances for multiple sclerosis—has not gained general acceptance.

Postvaccinial encephalomyelitis has become a rare disorder and the definitive series were collected several decades ago when vaccination against smallpox was necessary. The illness develops within 2 to 3 weeks of vaccination with a skin rash and systemic symptoms, followed by cerebral or myelitic signs which usually recover spontaneously, in due course.

Optic neuritis

Optic neuritis presents with pain on eye movement, followed by blurred vision which evolves over hours or days, sometimes to complete blindness; patients may be aware of selective loss of colour vision and flashes of light (phosphenes) on eye movement. The pain disappears within a few days; vision improves in 90 per cent of patients over months, but defects of colour perception frequently persist. Optic neuritis may present with progressive visual failure in one or both eyes, but in these situations care must be taken to exclude compression of the anterior visual pathway. Transient visual loss, mimicking optic neuritis, also occurs in ischaemic optic neuropathy, sarcoidosis, or Eales' disease and a family history should be taken since the presentation of visual failure in Leber's hereditary optic neuropathy is similar to bilateral sequential optic neuritis in men. The lesion responsible for optic neuritis can be imaged *in vivo*; inflammation within the intracanalicular portion of the nerve and long lesions are associated with delayed or incomplete recovery of vision. Correlations between imaging, symptoms, and neurophysiological changes indicate that the visual deficits in optic neuritis arise at the time of altered blood–brain barrier permeability. They are associated with conduction block and precede demyelination or axonal degeneration.

The frequency with which the optic nerve is involved leads to anxiety in the informed patient that an episode of optic neuritis is likely to be the first manifestation of multiple sclerosis. The risk is highest in the first 5 years, but the proportion of cases having recurrent demyelination continues to rise thereafter and life-table analysis suggests that up to 80 per cent eventually convert. In children, optic neuritis is commonly bilateral and recurrent demyelination affecting other parts of the nervous system rarely occurs. Bilateral simultaneous optic neuritis in adults, although less common than in children, also carries a low risk of multiple sclerosis. Recurrent optic neuritis is associated with an increased risk and this is marginally higher in females than males. MRI abnormalities in the periventricular white matter are found in more than 60 per cent of patients with optic neuritis, and the risk of developing multiple sclerosis is substantially increased for those having two or more such lesions at presentation; conversely, the absence of cerebral lesions is a good prognostic sign. The presence of oligoclonal bands on cerebrospinal fluid electrophoresis during the acute phase is also a significant risk factor.

Transverse myelitis

The spinal cord is vulnerable to postinfectious inflammatory damage, but as with acute disseminated encephalomyelitis in adults, the precipitating cause is often not identified. Transverse myelitis presents with pain at the site of the lesion, followed by weakness in the legs, sensory symptoms, and sphincter involvement. The weakness increases and the clinical picture is that of spinal shock—features rarely seen in acute cord lesions due to multiple sclerosis. Sphincter control is lost, but unlike patients with multiple sclerosis, there is usually difficulty in emptying rather than filling the bladder. The need to exclude a structural abnormality in patients with transverse myelitis means that many patients undergo radiological investigation which may demonstrate cord swelling. The spinal fluid shows an increased mononuclear cell count, numerically intermediate between the marked pleocytosis of acute necrotizing myelitis and the marginal abnormalities seen in multiple sclerosis; total protein is raised and oligoclonal bands may be present on electrophoresis, but the glucose is usually normal. Transverse myelitis is more common in adults than children; there is a high frequency of persistent disability, but a much lower conversion to multiple sclerosis than following optic neuritis.

Acute necrotizing myelitis causes rapidly progressive flaccid areflexic paraplegia with anaesthesia and loss of sphincter control. The intensity of inflammation results in severe pain with meningism, pyrexia, and systemic symptoms. The condition mimics cord compression; the cerebrospinal fluid changes resemble pyogenic or tuberculous infection of the central nervous system. For these reasons, treatment with high-dose intravenous steroids, which may usefully influence mortality and limit long-term disability, is often withheld. Acute necrotizing myelitis has been described in association with herpes virus infection, and as a complication of acute lymphocytic leukaemias, lymphoma, carcinoma, and acquired immune deficiency syndrome.

Devic's disease (neuromyelitis optica)

Devic's disease is characterized by massive confluent demyelination in the anterior visual pathway together with equally severe spinal cord damage, occurring simultaneously or sequentially and in either order, the episodes usually separated by weeks or months. Cellular reaction in the cerebrospinal fluid more usually involves polymorphonuclear cells than lymphocytes, but often lacks oligoclonal bands. Cerebral white matter abnormalities tend to be frontotemporal and the spinal lesion is long, extending over several segments, in contrast to the several short lesions which characterize spinal MRI in multiple sclerosis.

The distinction from multiple sclerosis is partly confused by definitions. Demyelinating disease often follows the Devic pattern in Japanese and African patients, where multiple sclerosis is otherwise rare. Cases are described complicating pulmonary tuberculosis, especially in African patients. European patients with bilateral simultaneous optic neuritis and transverse myelitis often show manifestations of widespread demyelination and multiple events occur in due course. There is a better prognosis for recovery when the optic nerves and spinal cord are affected in rapid sequence, but the outcome is generally poor with an appreciable mortality, especially in relapsing patients with more than two episodes.

Isolated brainstem syndromes

The clinical symptoms and signs of isolated brainstem syndromes typically consist of disequilibrium, disturbed eye movements, facial numbness, and dysarthria, but there may be severe headache which rightly leads to early investigation in order to exclude a structural lesion. The majority of patients progress to clinically definite multiple sclerosis; as with other isolated demyelinating syndromes, abnormal MRI outside the affected site at presentation is a poor prognostic sign for clinical conversion.

Multiple sclerosis

Aetiology

The aetiology of multiple sclerosis involves an interplay between genes and the environment. It is a disease of northern European people and occurs less frequently in other racial groups. The familial recurrence rate is approximately 15 per cent. Meta-analysis amongst relatives of probands from three population-based series shows that the age-adjusted risk is highest for siblings (3 per cent), then parents and children (2 per cent), with lower rates in second- and third-degree relatives. Recurrence in monozygotic twins is around 35 per cent. Conversely, the frequency of multiple sclerosis in adoptees is similar to the population risk for Europeans. The age-adjusted risk for half-siblings is intermediate between social and biological relatives. Recurrence is higher in the children of conjugal pairs with multiple sclerosis (age-adjusted 20 per cent) than the offspring of single affecteds (2 per cent) (Fig. 2).

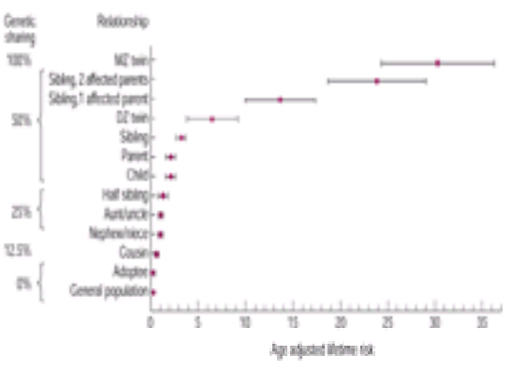

Fig. 2 Lifetime risk for multiple sclerosis amongst European people and in biological and social relatives of affected individuals. The increased risk with relatedness implicates genetic factors whereas the incomplete concordance in identical twins reflects the contribution made by environmental conditions.

Population studies demonstrate an association between the class II MHC alleles (DR15 and DQ6) and their corresponding genotypes. A specifically different association in seen in Mediterranean populations (DR4). Extensive searches, using association and linkage studies over many years, have yielded very few additional candidates for susceptibility. To date, eight genome screens have failed to identify a major susceptibility locus using identity by descent analysis in sibling pairs and enough markers to provide around a 10 centiMorgan map. The possible reasons are that no such gene exists, it has been missed, or heterogeneity has obscured the picture. Whole genome association screens dependent on linkage disequilibirum are in progress. Genetic analyses assume that multiple sclerosis is one disease, but

this may not be true. Mutations of mitochondrial DNA are responsible for a multiple sclerosis-like illness characterized by disproportionate involvement of the anterior visual pathway, although mitochondrial genes do not contribute generally to susceptibility in multiple sclerosis. Conditioning the United Kingdom genome screens for DR15 shows clustering of regions of interest within subsets of families. A major part of future studies in the genetics of multiple sclerosis will be to resolve the question of disease heterogeneity.

The distribution of multiple sclerosis cannot be explained only on the basis of population genetics. In white South African people and in Australia, prevalence rates are half those documented for many parts of northern Europe. There is a gradient in frequency, both in Australia and in New Zealand, which does not follow genetic clines. The risk is higher for English-speaking white people migrating into South Africa as adults than in childhood. Multiple sclerosis occurs at a low frequency in the Caribbean poulation, but the risk increases substantially in their first-generation descendants raised in the United Kingdom. Over and above the effect of racial predisposition, migration influences distribution of the disease. Surveys of multiple sclerosis have prompted speculation on the occurrence of post-Second World War epidemics in Iceland, the Orkney and Shetland Islands, and the Faroes, but others prefer the interpretation that these merely reflect improved case recognition.

The risk of developing multiple sclerosis is increased for individuals exposed to measles, mumps, rubella, and Epstein–Barr virus infection relatively late in childhood or adolescence. These studies suggest that an age-linked period of susceptibility to viral exposure exists in those who are constitutionally at risk of developing the disease. Attempts to implicate specific environmental agents are frustrating. Putative candidates of current interest are human herpes virus 6 and *Chlamydia pneumoniae*.

Clinical symptomatology

Special senses

Visual involvement is almost invariable and most commonly affects the optic nerve (see above). The post-chiasmal visual pathway is occasionally involved resulting in hemianopic field defects. Deafness occurs in multiple sclerosis, sometimes at presentation. Feelings of unsteadiness are common. Acute brainstem demyelination causes severe positional vertigo, vomiting, ataxia, and headache. Taste may be subjectively abnormal but ageusia is rarely described. Anosmia is reported in a high proportion of asymptomatic patients examined with more than usual thoroughness.

Motor symptoms and signs

Impaired mobility affects the majority of patients with multiple sclerosis usually as a result of spinal disease. Movements are slow, weakness differentially affecting extensors in the arms and flexors in the legs, and there are the expected signs of upper motor neurone lesions. Spasticity may be more problematic than weakness and all aspects of immobility are frequently complicated by fatigue. Cerebellar involvement causes incoordination of speech, bulbar control, eye movements, the individual limbs, or balance, usually in combination with corticospinal damage. Damage to the superior cerebellar peduncle or red nucleus produces a disabling proximal wild flinging tremor, and many other movement disorders have been described. Lower motor neurone signs occur when there is extensive demyelination adjacent to the dorsal root entry zone.

Sensory symptoms and signs

Altered sensation occurs at some stage in nearly every patient with multiple sclerosis. Damage to the posterior columns in the cervical cord produces tight, burning, twisting, tearing, or pulling sensations, which are usually unpleasant. Associated loss of proprioception severely compromises function. Spinothalamic tract involvement leads to loss of thermal and pain sensation. Non-specific tingling without accompanying signs is often described and the commonest physical sign found in the absence of symptoms is impaired vibration sense in the legs.

Demyelination of the dorsal or lumbar segments of the spinal cord produces paraesthesias and numbness in the legs, ascending to the trunk, and sometimes associated with sacral sparing, although a characteristic sensory syndrome seen in patients with multiple sclerosis is numbness of the perineum and genitalia with disturbed sphincter function.

Autonomic involvement

Autonomic symptoms occur in most patients with multiple sclerosis. Bladder symptoms are most common in women, whereas impotence occurs frequently in males. Loss of inhibition of reflex bladder emptying, normally mediated by cholinergic neurones that contract the detrusor and relax the internal sphincter, results in urgency and frequency with incontinence when combined with immobility. With conus lesions, the problem is impaired bladder emptying. Failure to fill and empty may coexist, resulting in detrusor contractions against a closed sphincter.

Impaired control of the rectal sphincter is much less of a problem than failure of emptying. Some impotent males with multiple sclerosis retain reflex erections, in which case psychogenic factors are often invoked; others have erectile failure due to spinal cord disease. Mechanical difficulties, spasticity, altered sensation, skin excoriation, and in-dwelling catheters affect sexual performance in both sexes. Other autonomic features in multiple sclerosis include: loss of thermoregulation leading to inappropriate sweating, fever, and hypothermia; Horner's syndrome; abnormalities of cardiac rhythm and vascular responses with acute pulmonary oedema; weight loss; and inappropriate secretion of vasopressin.

Eye movements

Abnormalities of eye movement are routine in multiple sclerosis. They are often asymptomatic but may manifest as double vision and oscillopsia. The commonest sign is first-degree symmetrical horizontal jerking nystagmus. Weakness of the lateral rectus is more common than isolated third and fourth nerve palsy. Internuclear ophthalmoplegia is often bilateral and may coexist with gaze paresis to produce the 'one and one half' syndrome.

Vertical up-beating nystagmus is always associated with bilateral internuclear ophthalmoplegia. Down-beating nystagmus has other important causes which can be confused with multiple sclerosis. Ocular flutter consists of horizontal saccadic oscillations without an intersaccadic interval. Opsoclonus, in which the movements occur in all directions, is equally disabling. Ocular bobbing describes an initial rapid downward eye movement followed by slow return to the neutral position and denotes cerebellar involvement. Abrupt displacement from the primary position during central fixation (square wave jerks) occurs with severe cerebellar deficits.

Other brainstem manifestations

Facial weakness, indistinguishable from Bell's palsy, occurs in patients with multiple sclerosis, alone or in association with other signs of brainstem disease including hemifacial spasm and diffuse rippling of muscle fibres (myokymia). Exceptionally, there may be unilateral involvement of the hypoglossal and recurrent laryngeal nerves. Extensive brainstem demyelination may produce disturbances of consciousness and respiratory failure distinct from the narcolepsy syndrome which is seen more frequently in patients with multiple sclerosis than expected by chance—an observation of immunogenetic interest in view of their shared HLA DR2 association. Occasional manifestations include the locked-in state, persistent hiccup, and the lateral medullary syndrome.

Paroxysmal symptoms are invariably brief but repetitive and last a few months before remitting. Symptomatic trigeminal neuralgia may begin in the first division or bilaterally, at a younger age than the idiopathic condition, and with associated signs of trigeminal involvement including motor weakness and sensory loss. It is usually associated with demyelinating lesions of the dorsal root entry zone, but may coexist with compression of the fifth cranial nerve by ectatic vessels. Other than trigeminal neuralgia, isolated involvement of the fifth nerve is rare. Paroxysmal dysarthria and ataxia with a clumsy arm, complex disturbances of sensation, and painful tetanic posturing of the limbs lasting 1 or 2 min are often triggered by movement and preceded by positive sensory symptoms on the side opposite to the muscular spasm. These are easily recognized and treated. Bursts of pain and paraesthesias, sensory distortion, itching, cough and hiccup, painful extensor spasm, akinesia, kinesogenic choreoathetosis, and complex gaze palsies—any of which may respond to anticonvulsants, especially carbamazepine—also appear to be paroxysmal manifestations of multiple sclerosis.

Cognitive and affective symptoms

Defects of visual and auditory attention occur in multiple sclerosis, sometimes at an early stage, and these are also detectable in patients with isolated demyelinating

lesions. An overall impairment in intelligence quotient relates more to duration of disease, and onset of the progressive phase, affecting memory rather than language skills. Specific cognitive deficits due to hypothalamic involvement including the Korsakoff state and the syndrome of bulimia, lack of social restraint, mental interia, and mutism are sometimes seen. Psychotic behaviour is rare, but depression occurs more frequently than in patients with comparable neurological disability; hypomania is occasionally seen but should not be confused with pathological laughter and crying, arising from loss of central inhibition of facial and bulbar reflexes in association with extensive brainstem disease.

Rare manifestations of multiple sclerosis

The list of rare clinical manifestations (some already described) includes massive cerebral lesions, aphasia, headache, fever, movement disorders, epilepsy, hypothalamic and pituitary symptoms, respiratory failure, and peripheral neuropathy. Narcolepsy, Sjögren's syndrome, ankylosing spondylitis, type I neurofibromatosis, and autoimmune thyroid disease have periodically been associated with multiple sclerosis.

Childhood multiple sclerosis

In retrospect, symptoms attributable to recurrent demyelination often affect individuals with multiple sclerosis as teenagers, but onset in the first decade also occurs; 2 per cent of patients with multiple sclerosis present before the age of 10, and 5 per cent before 16 years. Children with multiple sclerosis are usually girls. The individual episodes are often severe but the long-term prognosis surprisingly good. Fever and meningism, impaired conscious level due to cerebral oedema with swollen optic discs, and seizures are regular features and the distinction from acute disseminated encephalomyelitis can often only be made by the later occurrence of remission and relapse.

Clinical course and prognosis

The majority (80 per cent) of patients present with relapsing–remitting disease. Typically, the illness passes through the three phases of relapse with full recovery, relapse with persistent deficits, and secondary progression. One patient may spend several years or even a few decades in each, whereas another moves rapidly to a condition of fixed progressive disability. About 25 per cent of patients have multiple sclerosis in a form which is not disabling. In 5 per cent, relapses occur frequently and do not recover, leading rapidly to disability and early death from respiratory failure when the medulla is affected and from massive cerebral or spinal demyelination. Up to 15 per cent become severely disabled within a short time.

Episodes occur at random frequency but initially average about 1.5 per year and decrease steadily thereafter. Recovery from each attack is invariably slower than onset and may be incomplete. Self-evidently, secondary progressive multiple sclerosis tends to affect whichever system has previously been involved. Progression may follow directly upon a severe relapse and be interrupted by further episodes. In 20 per cent, multiple sclerosis is progressive from onset. The spinal cord bears the brunt of progressive multiple sclerosis, but optic nerve, cerebral, and brainstem disease may also advance slowly. Primary progressive spinal disease is the usual mode of presentation when multiple sclerosis develops beyond the fifth decade. Life expectancy is at least 25 years, and a high proportion of patients die from unrelated causes.

The prognosis is relatively good when sensory or visual symptoms dominate the illness and there is complete recovery from individual episodes; this pattern in most common in young females. Conversely, motor involvement, especially when co-ordination or balance are disturbed, has a less good prognosis. The outlook is also poor in later-onset patients and these are often males. Frequent, prolonged relapses with incomplete recovery and a short interval between the initial episode and first relapse carry a worse prognosis, but the main determinant of disability is onset of the progressive phase.

Prospective studies show that 9 per cent of upper respiratory (adenovirus) and gastrointestinal infections occurring in patients with multiple sclerosis are followed by relapse and 27 per cent of new episodes are related to infection. The emerging evidence suggests that disease activity is not increased by vaccination. Relapse rate is affected by pregnancy. There is a reduction in the prepregnancy relapse rate for each trimester with approximately a threefold higher risk in the puerperium, and the attacks may be more severe. The clinical course is uninfluenced by breast feeding or epidural anaesthesia. There is no evidence that trauma ever triggers the first or recurrent clinical manifestations of multiple sclerosis in someone who has the underlying disease process, or alters the course in individuals who have already experienced symptoms. A study of 170 patients studied prospectively for 8 years showed that (with the possible exception of electric shock) all forms of trauma are negatively correlated both with clinical exacerbations and disease progression.

Laboratory investigations

Investigations are used for four purposes in patients with multiple sclerosis: to demonstrate the anatomical dissemination of lesions; to provide evidence for intrathecal inflammation; to demonstrate that conduction is altered in a form consistent with demyelination; and to exclude conditions that mimic demyelinating disease. That said, the diagnosis can often reliably be made using clinical criteria and without laboratory support.

Electrophysiology

Demyelination can be detected in clinically unaffected pathways using visual, auditory, somatosensory, central motor, and event-related potentials; their latencies are characteristically delayed whereas, except in acute lesions, the amplitude is unaffected. Evoked potentials add little in situations where the pathway under investigation is clinically affected. Since they provide qualitatively different information, evoked potentials remain useful as an adjunct to diagnosis despite the advent of imaging techniques.

Magnetic resonance imaging

Low-density lesions, corresponding to areas of demyelination, may be seen using contrast-enhanced computed tomography and these occasionally have the appearances of cerebral tumour or abscess, but this technique is insensitive compared with MRI. More than 95 per cent of patients with clinically definite multiple sclerosis have periventricular lesions and more than 90 per cent also show discrete white matter abnormalities. Focal demyelination can be imaged in the optic nerve, brainstem, and spinal cord.

Variations in the imaging protocol are beginning to distinguish separate components of the underlying pathological process. Imaging can distinguish inflammation (gadolinium–DTPA enhancement of T_1-weighted lesions, indicating that the lesion is of recent origin), demyelination (magnetization transfer ratio), astrocytosis (T_2-weighted lesions, the signal arising from increased water content), and axonal damage (reduction in diffusion tensor imaging anisotropy and *N*-acetyl-aspartate spectra with chemical shift imaging, or the presence of focal atrophy and T_1-weighted black holes). The evolving lesion starts with increased blood–brain barrier permeability which lasts for up to 4 weeks and precedes the onset of T_2-weighted magnetic resonance changes. These lesions may disappear but reactivation is sometimes seen, the cycles lasting about 8 weeks. The periventricular lesions, which best characterize multiple sclerosis, correlate with areas of persistent demyelination and astrocytosis. A mixture of new, evolving, and recovering lesions may be seen in an individual patient at any one time. Magnetic resonance lesions occur about 15 times more frequently than new clinical events. Eventually, there is a reduction in the frequency of new lesions as patients switch from the relapsing to progressive phases of the disease and evidence for atrophy is then more apparent. The number or volume of lesions correlates poorly—if at all—with disease severity or course, but there is less cerebral involvement in patients who present with primary progressive disease compared with those having similar disability from secondary progression. The imaging abnormalities of multiple sclerosis are not specific and similar changes occur with inflammatory or vascular lesions and with advancing age (Fig. 3).

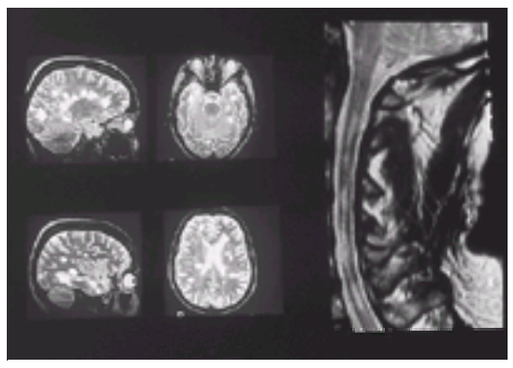

Fig. 3 T_2-weighted MRI abnormalities diffusely affecting the cerebrum and spinal cord in multiple sclerosis.

Imaging is not always necessary for diagnostic purposes in patients with a history of relapsing disease affecting multiple sites within the central nervous system. The major practical use is in the investigation of individuals with isolated demyelinating lesions, recurrent episodes at a single site, or progressive disease affecting the spinal cord. In all these situations the first requirement is to exclude a structural lesion, especially since these can present with relapsing symptoms. Imaging any region of the nervous system, clinically affected in isolation, will reliably exclude a structural lesion that might mimic multiple sclerosis and may show changes consistent with focal demyelination, but will not distinguish the syndromes of isolated demyelination from multiple sclerosis. Once the clinically affected part has proved negative for a structural lesion, the diagnosis of multiple sclerosis also requires the demonstration of anatomically separate lesions, ideally with enhancement to identify recent lesions, or the accumulation of new imaging abnormalities over time even if these are not expressed clinically.

Cerebrospinal fluid

Cerebrospinal fluid analysis provides information which is complementary to imaging abnormalities and specifically useful in elderly patients suspected of having multiple sclerosis. The cell count rarely exceeds 50 lymphocytes/ml, even during periods of clinical activity, and is normal in more than 50 per cent of patients. There is a rise in total protein with a specific increase in the immunoglobulin concentration and the presence of oligoclonal bands on protein electrophoresis in more than 90 per cent of cases, after correction for leakage of serum proteins through the blood–brain barrier, providing evidence for synthesis of immunoglobulin within the central nervous system. As with the imaging abnormalities, these are sensitive but not specific. Although some antibodies are directed against components of the oligodendrocyte or its myelin membranes, and others recognize extrinsic antigens including viruses, collectively these specificities only account for a minority of the bands. In each clinical situation a number of additional investigations will be required to exclude conditions which mimic multiple sclerosis.

Differential diagnosis

The commonest error in clinical practice is to make the diagnosis of multiple sclerosis in patients with progressive spinal disease in whom a structural lesion has not been adequately excluded. Rarely, spinal tumours present with intermittent symptoms creating difficulties for the unwary; it is just not safe to assume the diagnosis of multiple sclerosis in patients with symptoms and signs restricted to a single site whatever the clinical course. Lesions at the foramen magnum are particularly well placed to cause confusion through appearing to produce evidence for independent spinal and brainstem lesions. Errors also arise with progressive and relapsing manifestations of brainstem or spinal arteriovenous malformations.

Care must be taken in the diagnosis of multiple sclerosis when several members are affected within one family. Hereditary spastic paraplegia mimics familial multiple sclerosis and this should also be considered in isolated cases of progressive spastic paraplegia when pyramidal manifestations occur in isolation and with disproportionate spasticity. Other familial disorders confused with multiple sclerosis include the hereditary ataxias, adult-onset leucodystrophies, and vasculopathies (CADASIL). Pedigrees with affected males and maternal inheritance may be examples of X-linked adrenoleucodystrophy and individuals with the phenotype of multiple sclerosis occur in families with the clinical and genetic features of Leber's hereditary optic atrophy (Harding's disease).

Clinical, immunological, and imaging abnormalities indistinguishable from multiple sclerosis occur with granulomatous and vasculitic diseases of the brain, especially the cerebral variant of systemic lupus erythematosus which often occurs in the absence of systemic manifestations or informative serology. Sarcoidosis may present with clinical involvement of the central nervous system, typical magnetic resonance and cerebrospinal fluid abnormalities, and without pulmonary or cutaneous manifestations; uveitis also occurs in multiple sclerosis and so is not necessarily a useful discriminator. Orogenital ulceration in a patient with the clinical manifestations of multiple sclerosis suggests the diagnosis of Behçet's disease.

Alternative diagnoses need to be considered when multiple sclerosis is diagnosed in African or Asian people in whom progressive spinal disease, sometimes with visual involvement, is more probably due to HTLV1-associated tropical spastic paraplegia or neuromyelitis optica (see above). Infections of the nervous system can mimic the isolated demyelinating syndromes and multiple sclerosis. These include tuberculous and other chronic meningitides, and the neurological manifestations of acquired immunodeficiency syndrome and Lyme disease; borreliosis can also cause a chronic or relapsing disorder of the central nervous system, but this is usually preceded by the characteristic painful polyradiculitis and facial palsy that epitomizes Lyme disease. Similarities between multiple sclerosis and neurosyphilis should not be forgotten in the context of opportunistic infection complicating HIV infection. The age distribution and clinical manifestations usually make it easy to distinguish subacute combined degeneration of the spinal cord from multiple sclerosis, but focal spinal lesions, accompanied by Lhermitte's sign, occur in vitamin B_{12} deficiency.

The high public profile which multiple sclerosis currently enjoys leads many individuals with vague sensory symptoms or dizziness to consider the diagnosis for themselves. Many are easily reassured when these transient symptoms are unaccompanied by physical signs; these understandable anxieties differ from the syndromes fabricated by individuals seeking the dignity of a neurological diagnosis in the setting of psychiatric disease, which are much more difficult to manage.

Treatment of demyelinating disease

Management of the acute episode

Corticosteroids are effective in abbreviating acute episodes in multiple sclerosis. There is no difference in early or eventual response to treatment using high-dose oral or intravenous regimens. Most neurologists prefer the intravenous route despite the practical advantage of oral therapy. There may be a role for intravenous immunoglobulin in patients with severe acute deficits which do not respond to corticosteroids, although this strategy has not been assessed in clinical trials. Plasma exchange given up to 1 month after onset in the context of failed response to intravenous corticosteroids may reduce persistent deficits, although this does not prevent subsequent disease activity.

The treatment of symptoms

Several manifestations of multiple sclerosis can be improved symptomatically. Urgency or frequency of micturition respond to drugs with anticholinergic activity (oxybutinin or propantheline). A simple means for intermittently reducing urine volume, and hence the desire to micturate, is to use intranasal desmopressin spray. When detrusor and sphincter function become uncoupled, causing impaired bladder emptying with failure to fill, the preferred treatment is self-intermittent catheterization, which is easily adopted by motivated patients with adequate vision and arm function and ensures complete bladder emptying often with unimagined advantages to social activities and sleep. Other options include subtrigonal injections to reduce bladder sensation, a suprapubic catheter with closure of the lower urinary tract, urinary diversion through an ileal conduit, insertion of an artificial mechanical sphincter, or electrical stimulation of the spinal nerve roots in an attempt to synchronize sphincter contraction and relaxation. These may be preferable to an indwelling urethral catheter or, worse still, constant dribbling incontinence, which usually leads to skin excoriation.

Constipation in multiple sclerosis is managed by dietary alteration and the use of bulk laxatives, avoiding agents that act directly on the bowel wall. Loperamide may be useful where the predominant complaint is rectal urge incontinence. Psychological factors contribute to impotence in males with multiple sclerosis, but in most cases the complaint is a direct consequence of spinal demyelination. Trends in management have shifted from the use of semirigid prostheses and vacuum pump-induced tumescence, and self-administered cavernous injection of papaverine or prostaglandin E_1 applied through the urethra, to oral treatment with sildenafil

(Viagra)—a phosphodiesterase inhibitor which acts by increasing local production of nitric oxide in response to sexual stimulation.

The mainstay of pharmacological treatment for tremor is b-blockers; alternatives include anticonvulsants, isoniazid, ondansetron, and hyoscine. Physical restraint is rarely successful. Stereotactic procedures involving stimulation of the ventrolateral nucleus produce results comparable to destructive procedures, but the dividend is small. Unsteadiness arising from altered vestibular input may improve with the use of a vestibular sedative.

Fatigue may improve with amantadine or modetanil. Use of the aminopyridines in this and other contexts is limited by adverse effects including the risk of convulsions, although they can improve vision and muscle strength. Baclofen is still the most widely used effective antispastic agent. Benzodiazepines also reduce spasticity by increasing presynaptic spinal inhibition. Dantrolene sodium acts by uncoupling excitation–contraction mechanisms in individual muscle fibres. It is claimed that Tizanidine reduces spasticity without increasing weakness. Patients report that spasticity and pain improve with the use of cannabis and this is now formally being evaluated. Intrathecal baclofen carries the potential advantage of selectively reducing muscle tone in affected muscles whilst leaving others intact. It is mainly appropriate for patients with advanced disease and does not seem to have any additional adverse effects compared with systemic administration. Another approach is to use local injection of botulinum toxin. There may be a role for surgical interruption of the reflex pathways or tenotomy and peripheral nerve block with phenol or alcohol.

The paroxysmal manifestations of multiple sclerosis usually stop abruptly with the use of carbamazepine; this and other anticonvulsants, especially gabapentin, may also relieve trigeminal neuralgia or the more refractory forms of pain arising from spinal demyelination. Nerve block and chemical or surgical destruction of nerve fibres are sometimes an acceptable method for reducing pain in multiple sclerosis. All these sensations are coped with less well in the context of impaired mood and can respond usefully to antidepressants.

For those who develop significant disabilities and impairments, comprehensive care includes access to physical and occupational therapists, social workers, and other health-care staff with expertise in the management of chronic neurological illness. Complications are best prevented by awareness and anticipation since they usually develop quickly yet take months to resolve. Minimizing handicap by attention to social, vocational, marital, sexual, and psychological aspects of the illness remains more important to most patients than drug treatment. In situations where the natural history has already led to loss of mobility, the early use of mechanical aids and home adaptations should be encouraged despite the associated stigma.

Immunological treatment in multiple sclerosis

The use of non-specific agents in multiple sclerosis proved the concept that immunosuppression is a valid approach to treatment even if the magnitude of the effect often failed to establish a role for any one drug. The modern era began with meta-analysis of trials evaluating azathioprine showing a reduction in relapse rate but with a more modest effect on disability. The conclusion that the effect was of doubtful value to the individual patient and potentially posed serious long-term risks, meant that azathioprine was never routinely used to treat patients with multiple sclerosis. More recently introduced drugs appear to offer better adverse effects profiles but only achieve comparable efficacy. Some have already disappeared because of a poor showing in initial or confirmatory phase III trials, and due to unexpected toxicity. Casualties include linomide, cladribine, methotrexate, anti-CD4 monoclonal antibody, anti-TNFa antibody, oral myelin, altered myelin basic protein peptide, T cell vaccination, sulfasalazine, and intravenous imunoglobulin. For example, despite the apparent therapeutic rationale, a fusion protein linked to the soluble TNFa receptor produces a dose-dependent increased relapse rate, reduced time to relapse, and longer and more severe episodes—results which presumably have their explanation in the complex interacting networks of pro- and anti-inflammatory cytokines. Only a minority of patients receiving one or more well tolerated courses of a chimeric anti-CD4 antibody, sufficient partially to suppress the CD4 count, demonstrated clinical and imaging effects. Thus, the available anti-CD4 monoclonals used in isolation appear not to have a therapeutic future in multiple sclerosis. The two trials of altered myelin basic protein peptide ligand therapy in multiple sclerosis, designed to tolerize against auto-aggressive myelin basic protein-reactive T cells, either promoted relapses of multiple sclerosis or caused intolerable allergic adverse effects.

Several drugs, however, are now licensed in Europe and the United States for use in defined groups of patients with multiple sclerosis. Mitoxantrone achieves a higher conversion to disease inactivity (clinical and enhanced magnetic resonance imaging) in patients with active disease receiving monthly injections of methyl prednisolone—at least in the short term—but treatment is limited by the cumulative potential for cardiotoxicity. Glatiramer acetate (Copaxone) has been used in multiple sclerosis on the basis that disease activity can be suppressed by mimicking the antigenic challenge initiating brain inflammation. There is a reduction in relapse rate, more relapse free patients, and a delay in time to relapse but a less clear-cut effect on disability. Copaxone is associated with a change in cytokine production from a pro-inflammatory Th-1 to an anti-inflammatory Th-2/3 profile.

The therapeutic rationale for use of the b-interferons rests on the argument that IFN-b may limit inflammation by inhibiting antigen presentation, promoting a Th-2 immune phenotype, restricting migration of cells across the blood-brain barrier, reducing parenchymal cytokine production, and enhancing growth factor protection of axons; but it seems intrinsically unlikely that so many desirable properties, many characterized *in vitro* using experimental systems, are all relevant *in vivo*. BetaferonTM (IFN-b1b) is given by alternate day subcutaneous injection (8 million international units [miu]); AvonexTM (IFN-b1a) by weekly intramuscular injection (30 µg); and RebifTM (IFN-b1a) by alternate day subcutaneous injection (12 or 30 miu). IFN-b1a and IFN-b1b both reduce relapse rate by about one third and significantly reduce the accumulation of lesion load on magnetic resonance imaging in relapsing-remitting multiple sclerosis. The main adverse effects of IFN-b1b and IFN-b1a are local injection site reactions, flu-like symptoms and hyperthermia; contraindications include the use of IFN-b in pregnancy, and in patients with epilepsy or depression. Complications are not necessarily immediate and may occur 2 to 3 years after starting treatment. Between 15 and 40 per cent of patients receiving IFN-b develop neutralizing activity. Antibodies to IFN-b1a and IFN-b1b are immunologically and biologically cross-reactive. Subsequent attention has turned to whether there are dose response effects, a role for patients already in the progressive phase, and clinically useful delays in conversion to multiple sclerosis when IFN-b is given after a first episode of demyelination.

Comparisons of the licensed regimens do not show obvious dose effects or superiorities but some would rank high-dose Rebif above Betaseron and then Avonex for efficacy; whereas Avonex may have the edge over the other products for convenience and adverse effects. Although IFN-b1b was shown to delay progression in patients with secondary progressive multiple sclerosis, it seems likely that this mainly reflects suppression of superimposed new inflammatory lesions rather than an effect on other components of the pathogenesis contributing to secondary progression – and trials of the other products have shown no benefit; at best, IFN-b is only indicated in patients with secondary progressive multiple sclerosis also having frequent and clinically significant relapses. As predicted from the initial demonstrations of reduced relapse rate, IFN-b increases the time to a second and hence defining episode for multiple sclerosis in patients with isolated episodes of demyelination but this does not mean that these drugs have been shown to prevent multiple sclerosis.

Much effort has gone into the design of humanized monoclonal antibody and small molecule treatments for multiple sclerosis which either remove lymphocytes from the systemic circulation or prevent their migration into the central nervous system. Pulsed anti-a4 integrin antibody treatment in relapsing-remitting or progressive disease reduces active magnetic resonance lesions in the short term and is being studied in phase III trials. The humanized anti-CDw52 (CAMPATH-1H) antibody was originally shown to suppress radiological markers of cerebral inflammation by more than 90 per cent after a 1-week course and for at least 18 months during which relapses also stopped. However, 50 per cent of patients became progressively disabled from deficits acquired prior to treatment; these showed brain atrophy with evidence for axon degeneration on magnetic resonance spectroscopy. Clinical progression and atrophy correlated with the amount of brain inflammation in the pretreatment phase but occur in the absence of ongoing disease activity. The reduction in brain inflammation was associated with alteration in the immune response from a Th-1 pattern of cytokine release but, unexpectedly, this exposed autoimmune thyroid disease in one-third of patients. This drug is in phase II trials.

These observations support the hypothesis that inflammation is necessary for new lesion formation and conditions axon degeneration. The implication is that immunological therapies will best prevent progression of disability if given early in the course and before the cascade of events leading to axon degeneration is irretrievably established. This may explain the present limitations of immunotherapy in patients with secondary progressive multiple sclerosis but raises the dilemma of exposing individuals who may never develop disabilities from multiple sclerosis to the unpredictable hazards of prolonged immunosuppression.

Central pontine myelinolysis

Central pontine myelinolysis is associated with metabolic disturbances induced by alcohol with and without Wernicke's encephalopathy, non-alcoholic cirrhosis, Wilson's disease, following hepatic transplantation, as a complication of uraemia and haemodialysis, after prolonged vomiting, and in the context of diuretic therapy. In each of these situations, affected individuals have usually been hyponatraemic before the onset of neurological symptoms. Central pontine myelinolysis seems to result from overzealous correction of a low (and occasionally also a high) serum sodium. Demyelination correlates both with the degree of hyponatraemia and rate at which this is corrected; starting levels of less than 110 mmol/l or rates of correction of more than 2 mmol/l per hour substantially increase the risk of central pontine

myelinolysis. Rapid changes in sodium are better tolerated in acute than chronic hyponatraemia.

The illness affects central pontine pathways and spreads centrifugally. The fully evolved clinical picture is of flaccid paralysis with facial and bulbar weakness, disordered eye movements, loss of balance, and altered consciousness. Features of hyponatraemia, such as epilepsy, are not usually present since pontine demyelination follows correction of the serum sodium. The recent literature emphasizes the extrapontine manifestations including movement disorders and other features of extrapyramidal disease. The clinical features are distinctive and present no diagnostic difficulties unless the reduction in serum sodium has been overlooked; the acute changes of central pontine myelinolysis can be imaged and abnormalities persist after clinical recovery. Prognosis depends on the underlying metabolic disorder. With stabilization of the serum sodium and management of bulbar failure, neurological recovery is often complete and the condition does not recur spontaneously.

Childhood and adult-onset leucodystrophies

The leucodystrophies are characterized by non-inflammatory demyelination. They include a heterogeneous group of conditions. Increasingly, these are being shown to result from mutations affecting genes which determine the synthesis, maintenance, and structure of myelin. Although rare even in paediatric practice, these need to be considered in young adults with atypical syndromes combining physical and intellectual deficits, sometimes with peripheral nerve involvement, in whom imaging shows confluent lesions confined to white matter.

Diffuse sclerosis (Schilder's disease)

The term diffuse cerebral sclerosis was originally used to identify a heterogeneous group of diseases affecting cerebral white matter. Of the diseases previously classified under this heading, familial sudanophilic diffuse sclerosis, Pelizaeus–Merzbacher disease, Krabbe's diffuse sclerosis (globoid cell leucodystrophy), Canavan's diffuse sclerosis (spongy degeneration of the white matter), Alexander's disease, and metachromatic leucodystrophy are dysmyelinating leucodystrophies. Conversely, Binswager's subcortical encephalopathy is now considered a consequence of diffuse cerebral arteriosclerosis—although some cases may have been examples of CADASIL; and Balo's concentric sclerosis is now considered within the spectrum of multiple sclerosis. Many male patients previously classified as having diffuse sclerosis were probably suffering from adrenoleucodystrophy. Some of the relapsing disorders were probably Leigh's disease associated with mutations of mitochondrial DNA. But even after separating these newly recognized conditions, the nosological status of diffuse sclerosis remains uncertain and some consider that, between them, acute childhood multiple sclerosis and adrenoleucodystrophy account for all the cases.

Krabbe's disease

Globoid cell leucodystrophy usually presents as an early infantile disorder. Late-onset globoid cell leucodystrophy is uncommon—almost all patients becoming symptomatic before the age of 5 years and so almost never leading to confusion with childhood multiple sclerosis. The clinical picture is dominated by behavioural changes with startle, progressive intellectual and motor deterioration, epilepsy, visual failure, and peripheral neuropathy leading to severe disabilities; pyrexia and other autonomic features usher in the onset of a vegetative state. Visual evoked potentials are delayed and the spinal fluid has a raised protein level but does not contain oligoclonal bands. MRI shows periventricular lesions subsequently extending into extensive white matter changes. The deficiency of a-galactocerebrosidase, best demonstrated in peripheral blood leucocytes or skin fibroblasts, leads to the accumulation of galactocerebroside, the neurotoxic molecule psychosine, and the myelin-laden macrophages or globoid cells.

Adrenoleucodystrophy

An important group of disorders is characterized by deposition of saturated fatty acids in the brain and other lipid-containing tissues as a result of defective very-long-chain fatty acyl-CoA synthetase activity in peroxisomes. The molecular defect may result from failure of the adrenoleucodystrophy gene product to anchor very-long-chain fatty acids into the peroxisomal membrane or translocate these into peroxisomes.

Four related syndromes share this biochemical abnormality: childhood adrenoleucodystrophy and adult-onset adrenomyeloneuropathy are X linked; neonatal adrenoleucodystrophy and Zellweger's syndrome are autosomal recessive disorders.

X-linked childhood adrenoleucodystrophy presents with behavioural disturbance, dementia, and epilepsy followed by involvement of special senses and motor systems. Although a significant proportion of children later develop adrenal insufficiency, Addison's disease may precede the neurological manifestations by several years. Treatment has been proposed with a dietary supplement containing a 4:1 mixture of glyceryl trioleate and trieructate, popularly known as Lorenzo's oil. This lowers the plasma levels of very-long-chain fatty acids, but does not appear to influence the phenotype in individuals with established neurological disease, although there may be a prophylactic role. Bone marrow transplantation is successful in early symptomatic cases and, in view of the inflammatory reaction, trials of immunosuppression are in progress.

Adrenomyeloneuropathy presents in adult men with spastic paraparesis and sensory loss in the legs; attention is drawn to an unusual cause for this otherwise common neurological problem by the associated peripheral neuropathy, but the diagnosis is frequently overlooked if adrenal insufficiency is not obvious at presentation. Identification of the peroxisomal defect in easily sampled body tissues has led to the description of cases with obscure clinical manifestations; these include focal cerebral lesions, Kluver–Bucy syndrome, dementia, and spinocerebellar degeneration. Mild spastic paraparesis with sphincter involvement and peripheral neuropathy may occur in obligate heterozygote female carriers with elevated very-long-chain fatty acids. The gene has been mapped to Xq28, close to that for glucose-6-phosphate dehydrogenase deficiency and colour blindness.

Autosomal recessive adrenoleucodystrophy presents in infancy with seizures, hypotonia, retardation, retinal degeneration, and hepatic involvement; females are more commonly affected than males. Although the clinical manifestations and mode of inheritance are similar in neonatal adrenoleucodystrophy and Zellweger's syndrome, these are thought to be separate disorders.

The sensitivity and specificity of routine assays for very-long-chain fatty acids show that the level of hexasanoic acid and its ratios to tetrasanoic and docosanoic acids are fully discriminating in homozygote males, irrespective of the clinical phenotype, from the day of birth if dietary supplements have not been given, providing an opportunity for mass screening; there is a false-negative rate of 15 per cent for heterozygotes.

Metachromatic leucodystrophy

The separation of metachromatic leucodystrophy from the heterogeneous group of diffuse sclerosis occurred when metachromatic material was first detected in urinary deposits. It subsequently became clear that the diagnosis can be confirmed by demonstrating increased urinary sulphatide excretion with a deficiency of arylsulphatase A in urine, peripheral blood leucocytes, and skin fibroblasts, or showing metachromatic material in peripheral nerve biopsies having segmental demyelination and remyelination. There is diffuse white matter involvement due to non-inflammatory demyelination with loss of oligodendrocytes, axon preservation, and reactive astrocytes which, together with macrophages, contain the metachromatic material, especially in the most extensively demyelinated areas.

The clinical phenotype varies with the amount of surviving arylsulphatase A depending on heterozygosity of the mutant allele; pseudodeficiency refers to those individuals with low levels of arylsulphatase A that are sufficiently high not to display a clinical phenotype. Some affected individuals have a genetic defect of the arylsulphatase A activator and this is associated with a more complex pattern of sphingomyelin storage, biochemically and in terms of the tissue distribution.

The most common form of metachromatic leucodystrophy develops in late infancy with delayed walking due to the neuropathy, which may be painful. There are also features of brainstem involvement and the emergence of diffuse upper motor neurone signs with reduced intellectual development, optic atrophy, and death within about 5 years from presentation. In later-onset childhood cases, after several years normal development, there are behavioural changes with poor school performance, anticipating cerebellar and upper motor neurone disability which then follows much the same course as in younger patients, although with less evidence for neuropathy. The early adult form of metachromatic leucodystrophy is rare, or perhaps seldom diagnosed, and tends to present with intellectual or emotional abnormalities. Onset with dementia and behavioural disorders is usual with ataxia, paralysis, and optic atrophy only developing at late stages; the presentation is occasionally with paraparesis or cerebellar ataxia and the condition can then more easily be mistaken for multiple sclerosis. Clinical evidence for peripheral neuropathy may be revealed by slowed nerve conduction. Treatments have included dietary manipulation with reduced vitamin A and sulphur-containing substances, and bone marrow transplantation, but the successes are limited.

Multiple sulphatase deficiency combines the features of metachromatic leucodystrophy with mucopolysaccharidosis. It also has neonatal, early childhood, and juvenile forms. The pattern of combined motor and mental regression or lack of development reflecting widespread dysmyelination with peripheral neuropathy is associated with dysmorphic features and organomegaly. The more severe phenotype also reflects extensive neuronal loss due to the combination of stored sulphatide, sulphated steroids, and mucopolysaccharides. The enzyme defects are complex involving many sulphatases including arylsulphatase A.

Pelizaeus–Merzbacher disease

The three phenotypes of X-linked Pelizaeus–Merzbacher disease usually present in childhood. The clinical features which may distinguish the otherwise ubiquitous motor and developmental delay with epilepsy are abnormal eye movements, dystonia and choreoathetosis, and laryngeal paralysis. Affected individuals often stabilize with severe disabilities and live into early adult life. Some cases do not manifest until early adult life, but here the blur with specifically different disorders becomes more apparent. MRI either fails to show myelin or depicts myelin which is immature with an atrophic brain.

The molecular defect is a mutation of the gene for proteolipid protein (encoded on X-q21.2). Proteolipid protein is normally involved in stabilizing the lamellar structure of central myelin. Over 30 mutations have been described resulting in expression of truncated forms of proteolipid protein sufficient to cause extensive oligodendrocyte loss and failure of myelination. The pedigree is not always X linked in the early-onset connatal form and, in these situations, a genetic defect other than proteolipid protein mutation is presumably involved.

Adult-onset dominant leucodystrophies

Forms of dominantly inherited leucodystrophy also occur exclusively in adults and may closely resemble chronic progressive multiple sclerosis. MRI shows diffuse, non-discrete, white matter disease and there are no oligoclonal bands in the spinal fluid. It remains uncertain whether all the adult-onset dominant leucodystrophies are one and the same disorder, and many are difficult to distinguish from the heterogeneous group of hereditary spastic paraplegias. The various phenotypes are gradually being classified as their biochemical and genetic defects are characterized. A family with spastic paraparesis, ataxia, and mild dementia presenting in adulthood, but with onset in childhood, has been described; diffuse white matter abnormalities were present on cerebral magnetic resonance, whereas pathognomic features of the other leucodystrophies were absent. The most recent addition to this group involves two siblings with behavioural abnormalities progressing to dementia with extensive white matter abnormalities on MRI in whom brain biopsy showed glycolipid inclusions in macrophages unlike any other lysosomal storage disease.

Further reading

Barnes D *et al*. (1997). Randomised trial of oral and intravenous methylprednisolone in acute relapses of multiple sclerosis. *Lancet* **349**, 902–6. [No difference between these two regimens for management of acute relapse.]

Bauer HJ, Hanefeld FA (1993). *Multiple sclerosis: its impact from childhood to old age*. Saunders, London. [The definitive monograph on childhood multiple sclerosis.]

Brex PA *et al* (2002). A longitudinal study of abnormalities on MRI and disability from multiple sclerosis. *New England Journal of Medicine* **346**, 158–64. [Long term follow-up of patients with isoloated demyelinating lesions.]

Coles AJ *et al*. (1999). Monoclonal antibody treatment exposes three mechanisms underlying the clinical course in multiple sclerosis. *Annals of Neurology* **46**, 296–304. [Clinical evidence for the complex pathogenesis of multiple sclerosis.]

Comi G *et al*. (2001). Effect of early interferon treatment on conversion to definite multiple sclerosis: a randomised study. *Lancet* **357**, 1576–82. [Delay to second episode in patients with isolated demyelination treated with b-interferon.]

Compston DAS *et al*. (1998). *McAlpine's multiple sclerosis*. WB Saunders, London. [The most recent monograph on multiple sclerosis.]

Compston DAS, Coles AJ (2002). Multiple sclerosis (seminar) *Lancet* **359**, 1221–31. [A comprehensive review of the pathogenesis and treatment of multiple sclerosis.]

Confavreux C *et al*. (1998). Rate of pregnancy-related relapse in multiple sclerosis. *New England Journal of Medicine* **339**, 285–91. [A prospective study of disease activity in pregnancy.]

Confavreux C *et al* (2001). Vaccinations and the risk of relapse in multiple sclerosis. Vaccines in Multiple Sclerosis Study Group. *New England Journal of Medicine* **344**, 319–26. [Evidence that vaccinations do not increase activity in multiple sclerosis.]

Ebers GC *et al*. (2000). The natural history of multiple sclerosis: a geographically based study. 8: familial multiple sclerosis. *Brain* **123**, 641–9. [The clinical features and natural history of multiple sclerosis in a population-based cohort described (to date) in a series of eight apers.]

Edan G *et al*. (1997). Therapeutic effect of mitoxantrone combined with methylprednisolone in multiple sclerosis: a randomised multi-center study of active disease using MRI and clinical criteria. *Journal of Neurology, Neurosurgery, and Psychiatry* **62**, 112–18. [Trial defining the role of mitoxantrone in multiple sclerosis.]

European Study Group on Interferon b-1b in Secondary Progressive MS (1998). Placebo-controlled multicentre randomised trial of interferon b-1b in treatment of secondary progressive multiple sclerosis. *Lancet* **352**, 1491–7. [Suggestive evidence for the role of interferon-b in progressive multiple sclerosis.]

Genain CP *et al*. (1999). Identification of autoantibodies associated with myelin damage in multiple sclerosis. *Nature Medicine* **5**, 170–5. [The putative autoantigen in multiple sclerosis.]

Hohlfeld R (1997). Biotechnical agents for the immunotherapy of multiple sclerosis: principles, problems and perspectives (review). *Brain* **120**, 865–916. [The present and future basis for treatment in multiple sclerosis.]

Jacobs LD *et al*. (1996). Intramuscular interferon b-1a for disease progression in relapsing multiple sclerosis. *Annals of Neurology* **39**, 285–94. [The pivotal trial of interferon-b1a in multiple sclerosis.]

Jacobs LD *et al*. (2000). Intramuscular interferon b-1a therapy initiated during a first demyelinating event in multiple sclerosis. *New England Journal of Medicine* **343**, 898–904. [Delay to second episode in patients with isolated demyelination treated with b-interferon.]

Jeffery ND, Blakemore WF (1997). Locomotor deficits induced by experimental spinal cord demyelination are abolished by spontaneous remyelination. *Brain* **120**, 27–37. [Experimental evidence that remyelination restores function.]

Johnson K *et al*. (1998). Extended use of glatiramer acetate (Copaxone) is well tolerated and maintains its clinical effect on multiple sclerosis relapse rate and degree of disability. *Neurology* **50**, 701–8. [Evidence that copolymer-1 has a clinical effect in multiple sclerosis.]

Luchinetti C *et al*. (1999). A quantitative analysis of oligodendrocytes multiple sclerosis lesions: a study of 117 cases. *Brain* **122**, 2279–95. [New ideas on the cellular pathology of multiple sclerosis.]

McDonald WI *et al* (2001). Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Annals of Neurology* **50**, 121–7. [Revised diagnostic criteria for multiple sclerosis.]

Miller DH *et al*. (1999). Effect of interferon-b1b on magnetic resonance imaging outcomes in secondary progressive multiple sclerosis: results of a European multicenter, randomised, double-blind placebo-controlled trial. *Annals of Neurology* **46**, 850–9. [Interferon-b1b may only suppress residual inflammation in secondary progressive multiple sclerosis.]

Miller HG, Stanton JB, Gibbons JL (1956). Parainfectious encephalomyelitis and related syndromes. *Quarterly Journal of Medicine* **25**, 427–505. [The classic account of acute disseminated encephalomyelitis.]

Moser HW (1997). Adrenoleukodystrophy: phenotype, genetics, pathogenesis and therapy. *Brain* **120**, 1485–508. [Review of adrenoleucodystrophy: the Gordon Holmes lecture.]

Paty DW, Li DKB, The IFNb Multiple Sclerosis Study Group (1993). Interferon b-1b is effective in relapsing–remitting multiple sclerosis. MRI results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology* **43**, 662–7. [The pivotal study of interferon-b1b in multiple sclerosis.]

PRISMS Study Group (1998). Randomised double-blind placebo-controlled study of interferon b-1a in relapsing/remitting multiple sclerosis. *Lancet* **352**, 1498–504. [The second pivotal study of interferon-b1a in multiple sclerosis.]

PRISMS-4 (2001). Long-term efficacy of interferon-b-1a in relapsing MS. *Neurology* **56**, 1628–36. [Late follow-up results for patients in the phase II trial of Rebif.]

Secondary Progressive Efficacy Clinical Trial of Recombinant Interferon-b-1a in MS (SPECTRIMS) Study Group (2001). Randomized controlled trial of interferon-b-1a in secondary progressive MS:

MRI results. *Neurology* **56**, 1505–13. [No evidence for efficacy of b-interferon in secondary progressive multiple sclerosis.]

Sibley WA, Bamford CR, Clark K (1985). Clinical viral infections and multiple sclerosis. *Lancet* **i**, 1313–15. [Prospective study of infections and disease activity in multiple sclerosis.]

Sibley WA *et al*. (1991). A prospective study of physical trauma and multiple sclerosis. *Journal of Neurology, Neurosurgery, and Psychiatry* **54**, 584–9. [Prospective study of trauma and disease activity in multiple sclerosis.]

The IFNb Multiple Sclerosis Study Group, the University of British Columbia MS/MRI Analysis Group (1995). Interferon b-1b in the treatment of multiple sclerosis: final outcome of the randomised controlled trial. *Neurology* **45**, 1277–85. [Final result on the pivotal study of interferon-b1b in multiple sclerosis.]

The Lenercept Multiple Sclerosis Study Group, the University of British Columbia MS/MRI Analysis Group (1999). TNF neutralisation in MS. Results of a randomised, placebo-controlled multicenter study. *Neurology* **53**, 457–65. [Suppressing tumour necrosis factor-a makes multiple sclerosis worse: surprising result.]

Trapp BD *et al*. (1998). Axonal transection in the lesions of multiple sclerosis. *New England Journal of Medicine* **338**, 278–85. [Rediscovery of the axonopathy in multiple sclerosis.]

Turbridy N *et al*. (1999). The effect of anti-a4 integrin antibody on brain lesion activity in MS. *Neurology* **53**, 466–72. [Minimal effect of anti-adhesion monoclonal antibody in active multiple sclerosis.]

van Oosten BW *et al*. (1997). Treatment of multiple sclerosis with the monoclonal anti-CD4 antibody cM-T412; results of a randomised, double-blind, placebo-controlled, MR monitored phase II trial. *Neurology* **49**, 351–7. [Trivial effect of anti-CD4 monoclonal antibody in multiple sclerosis.]

Wingerchuk DM *et al*. (1999). The clinical course of neuromyelitis optica (Devic's syndrome). *Neurology* **53**, 1107–14. [Definitive recent series of Devic's disease.]

Youl BD *et al*. (1991). The pathophysiology of acute optic neuritis: an association of gadolinium leakage with clinical and electrophysiological deficits. *Brain* **114**, 2437–50. [Classic study of the pathophysiology of inflammation in human demyelinating disease.]

24.17 Disorders of the neuromuscular junction

David Hilton-Jones and Jackie Palace

Introduction

Two fundamentally different pathological processes are associated with disease at the neuromuscular junction. First, acquired disorders in which autoantibodies are directed against nerve or muscle ion channels. Second, and much rarer, inherited conditions in which the defect may be pre- or postsynaptic. These acquired and inherited conditions share some symptomatology. The most important are the autoimmune diseases: myasthenia gravis, the Lambert–Eaton myasthenic syndrome, and acquired neuromyotonia—disorders for which therapy is available.

The pharmacological and neurophysiological complexities of the neuromuscular junction can be simplified to a level that permits ready understanding of the pathogenesis and treatment of these various conditions and will reduce the frequency of misdiagnosis and/or mismanagement.

Neuromuscular transmission

Anatomically there are three main components to the neuromuscular junction (Fig. 1). The presynaptic component is the motor nerve terminal, which contains packages (quanta) of acetylcholine, each of which contains several thousand molecules of acetylcholine. This is separated from the postsynaptic acetylcholine receptors, which sit atop the terminal expansions of the junctional folds of the muscle fibre membrane, by the synaptic space. The nerve fibre membrane contains voltage-gated sodium, potassium, and calcium channels. Voltage-gated sodium channels are also present postsynaptically, at the base of the clefts of the junctional folds.

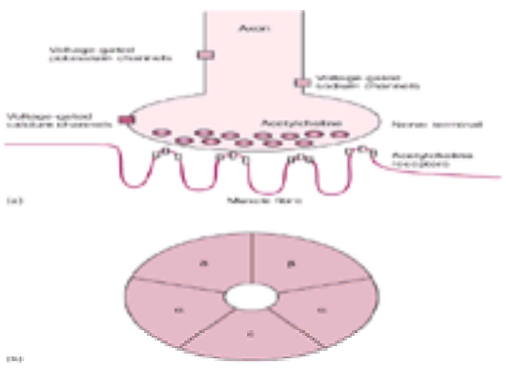

Fig. 1 (a) VGNC, voltage-gated sodium channel; VGCC, voltage-gated calcium channel; VGKC, voltage-gated potassium channel; ACh, acetylcholine. (b) Cartoon of the organization of the subunits of the nicotinic acetylcholine receptor (see text).

The nicotinic acetylcholine receptor is a pentameric structure composed of four different subunits—a, b, g, and † in fetal muscle, and a, b, e, and † in adult muscle. It is configured to produce a central ion channel. Structurally and functionally there are similarities to voltage-gated ion channels, but the acetylcholine receptor is a ligand-gated channel, the ligand being acetylcholine.

Depolarization of the motor nerve terminal is dependent upon voltage-gated sodium channels. Repolarization is the result of inactivation of these sodium channels and opening of voltage-gated potassium channels. During depolarization, voltage-gated calcium channels open—the influx of calcium ions into the nerve terminal triggers release (by exocytosis) of quanta of acetylcholine.

The acetylcholine binds to the a-subunits of the acetylcholine receptors. This alters the conformation of the channel allowing cations (mainly sodium) to enter the muscle fibre. This influx generates the end-plate potential, which in turn activates voltage-gated sodium channels. These trigger the action potential which is propagated through the muscle fibre and initiates contraction. Spontaneous release of individual quanta of acetylcholine, as opposed to mass release triggered by a nerve action potential, gives rise to miniature end-plate potentials, which can be recorded by microelectrode. These are of insufficient amplitude to trigger an action potential in the muscle fibre membrane.

The action of acetylcholine on acetylcholine receptors is terminated by the hydrolysis of acetylcholine by the enzyme acetylcholinesterase, which is anchored to the basal lamina by a collagen-like molecule, ColQ.

The acquired myasthenic disorders are associated with antibodies directed against one of the ion channels (Table 1). That there are three autoimmune disorders known to affect such a small region may be explained by the fact that the neuromuscular junction, unlike the peripheral nerves, is not contained within the blood–nerve barrier, which stops just short of the nerve terminal, and is thus potentially exposed to immune-mediated attack. The inherited disorders may affect presynaptic processes (acetylcholine resynthesis, packaging, or release), acetylcholinesterase binding, or postsynaptic function (point mutations in acetylcholine

receptor subunits). Pathogenetic mechanisms are considered in more detail when discussing individual disorders.

Myasthenia gravis

This is by far the most common of the conditions to be discussed in this section and responds favourably to treatment. In general, over 90 per cent of patients can be returned to normal function, although in most this represents a pharmacological remission and the patient remains dependent on treatment.

Epidemiology

All ethnic groups are affected. The annual incidence is about 4 per million population, and the prevalence about 8 per 100 000. All ages may be affected. Overall, women are more frequently affected, in a ratio of about 3:2. The female bias is even more marked in younger-onset cases, whereas over the age of 50 years male cases predominate. A rather different pattern is seen in people of Asian origin; prepubertal onset is very common, the disease is often purely ocular, and there is a strong association with HLA DRw9.

Pathogenesis

The fundamental disorder in myasthenia gravis is loss of functional acetylcholine receptors consequent upon binding of anti-acetylcholine receptor (**anti-AChR**) antibodies. IgG class antibodies can be detected, by the standard assay used for diagnostic purposes, in 85 per cent of patients with generalized myasthenia and about one-half of those with purely ocular myasthenia (so-called seropositive cases). Antibodies mostly bind to the main immunogenic region of the a-subunits of the acetylcholine receptor. Patients who do not have antibodies detected by this assay are called seronegative. However, there is overwhelming evidence even in these patients that their disease is immune mediated, possibly by IgG antibodies that bind to other regions within the end-plate area or to other IgM class antibodies; their clinical characteristics are similar to seropositive patients, they respond to plasma exchange and immunosuppressant therapy, their plasma can induce neuromuscular transmission dysfunction when injected into animals, and the infants of such mothers may be born with neonatal myasthenia (see below) indicating transplacental transfer of a humoral component.

Loss of functional acetylcholine receptors by antibody binding is due to complement-mediated lysis, acceleration of internalization and degradation, and blocking of acetylcholine binding. Morphological consequences include widening of the synaptic cleft and a marked reduction of the postsynaptic folds of the muscle fibre membrane.

Although the efferent limb of the immune response, described above, has been reasonably well characterized, numerous questions remain to be answered about the afferent limb. Susceptibility to myasthenia gravis is associated with particular immune response genes, with correlation to different haplotypes relating to the age of onset of the disease. These observations are not of immediate relevance to routine clinical practice. In contrast, knowledge about involvement of the thymus is relevant to classification and management.

In early-onset, seropositive patients there is hyperplasia of the thymic medulla, with germinal centres surrounded by a T-cell zone. The acetylcholine receptor is expressed on thymic myoid cells and there is enrichment of acetylcholine receptor-specific T cells. In addition, the thymus has a key role in the process of inducing immune tolerance, by removal of self-antigen T-cell clones. On the basis of these observations, and the beneficial response to thymectomy, there seems little doubt that the thymus is involved in the pathogenesis of myasthenia gravis, but exactly how has yet to be elucidated. Identification of the mechanism may well be important in developing immune-specific treatment.

In late-onset cases and seronegative patients, the thymus is typically normal or atrophic, although some pathological changes have been noted.

Thymomas occur in about 10 per cent of cases. They are locally invasive (notably affecting the pleura and pericardium) and may seed within the pleural cavity. These patients are almost invariably seropositive. Surgical excision is required because of local tumour invasiveness, but in contrast to those patients with thymic hyperplasia, this does not usually ameliorate the myasthenic symptoms.

Based on the presence or absence of antibodies detected by the routine clinical assay and the state of the thymus gland, four main subgroups of patients can be identified (Table 2). Penicillamine-induced myasthenia, which generally recovers following drug withdrawal, is clinically similar to the idiopathic disease and the patients are seropositive.

Clinical features

Myasthenia gravis causes skeletal muscle weakness but the most characteristic feature is fatiguability. The term fatigue causes some confusion because it may have different meanings to a clinician, physiologist, and lay person. Thus, the fatigue of chronic fatigue syndrome is quite different from that of myasthenia. Simply, fatigue in myasthenia gravis manifests itself by increasing, demonstrable weakness precipitated by repeated or sustained muscular activity. Symptoms fluctuate from day to day and week to week, which may in part explain the common delay in diagnosis and suspicions as to its genuineness. Other factors which can exacerbate the weakness include heat, emotional factors, menstruation, intercurrent infections, and drugs that interfere with neuromuscular transmission (aminoglycoside antibiotics, quinine, quinidine, b-blockers, procainamide, and neuromuscular-blocking drugs related to anaesthesia).

In over one-half of patients the presenting symptoms relate to extraocular muscle weakness (diplopia and ptosis); these muscles will be involved in over 90 per cent of patients at some stage during the disease. The next most frequent presentation is with limb-girdle weakness. Typically, as the disease worsens, the weakness spreads from the extraocular muscles to the lower facial and bulbar muscles (causing dysarthria and dysphagia), to the neck, and then to the limbs. However, there are many variations on this theme. A relatively common presentation in older patients, typically men, is with selective weakness of neck extension—as they walk their head drops forwards and they arrive in the clinic holding up their head with a hand under the chin. Relatively selective weakness of finger extension and abduction is common.

On examination, weakness may or may not be evident—fatigue can be demonstrated in limb muscles, but is often most striking around the eyes and with respect to bulbar muscles. Fatiguable ptosis is a striking sign (Fig. 2). As the patients give their history, the fatigue of bulbar muscles may be revealed by increasing dysarthria. A potentially misleading sign is 'pseudo-internuclear ophthalmoplegia', which may be bilateral—failure of adduction due to weakness of the medial recti. Eye movements may show striking fatigue. With increasing severity, the weakness at rest may be so marked that it is difficult to demonstrate fatigue. Respiratory muscle weakness may be out of proportion to limb weakness—it is best assessed by measuring the vital capacity (not peak flow), and the effects of it by monitoring oxygen saturation. Muscle wasting is only seen in undertreated patients with long-established disease. The tendon reflexes are normal, and indeed often rather brisk. There are no abnormal sensory signs.

Fig. 2 Fatiguable ptosis in myasthenia gravis.

There is an increased incidence of other autoimmune diseases, particularly thyroid disease (about 3 per cent of patients) and less frequently rheumatoid arthritis, systemic lupus erythematosus, polymyositis, Lambert–Eaton myasthenic syndrome, and acquired neuromyotonia.

Natural course

This is very variable. In some patients the disorder remains confined to the extraocular muscles (ocular myasthenia gravis). If that is the case for more than 2 years, and particularly if the patient is seronegative, the development of generalized disease is unlikely. Older studies, before the introduction of immunosuppressive therapies, suggest that the disease reaches maximum severity within 7 years. In one study, the interval between onset and the first episode of maximal weakness ('myasthenic crisis') was less than 36 months in over 80 per cent of patients. Permanent, spontaneous remission occurs, but is rare—in the order of 1 per cent per annum. On the other hand, particularly early in the course of the disease, there may be protracted periods of spontaneous remission, sometimes lasting several years.

Diagnosis

This is based on the clinical picture, supported by appropriate laboratory results. For practical purposes, the presence of anti-AChR antibodies is confirmatory and no further diagnostic investigations are required. In seronegative patients, electromyography and the intravenous edrophonium (Tensilon®) test are helpful. Although the edrophonium test has a long pedigree and sound pharmacological basis (it is a short-acting cholinesterase inhibitor), there are concerns about its use, particularly by the inexperienced. The patient is given 600 µg of atropine intravenously—this blocks the potentially unpleasant muscarinic effects of the edrophonium and also acts as a single-blind placebo for the patient. The test dose of 2 mg of edrophonium follows, which in some patients is sufficient to give a diagnostic response. If not, a further 8 mg of edrophonium is given. There must be an easily assessable measure of improvement—most commonly degree of ptosis. The test is therefore likely to be of most use in patients with purely ocular symptoms and signs. Rarely, cardiorespiratory collapse may occur. False-negative and false-positive results are not uncommon.

The conventional electromyographic measure for diagnosing myasthenia gravis is the demonstration of a decremental response of the compound muscle action potential in response to repetitive nerve stimulation at 3 Hz. More sensitive, but not specific and only available in specialist centres, is the presence of increased jitter and blocking, as assessed by single-fibre electromyography.

The presence of a thymoma is best assessed by computed tomography or magnetic resonance imaging of the thorax.

Differential diagnosis

There are few difficulties in the presence of extraocular muscle involvement and readily demonstrable fatigue, although there can be confusion with the Lambert–Eaton myasthenic syndrome and the congenital myasthenic syndromes. Diagnostic difficulties can occur when, as occasionally happens, eye signs and fatigue are absent. Amyotrophic lateral sclerosis with little wasting may be suspected. Conversely, in long-established myasthenia, muscle wasting may be misleading. More difficult is seronegative, purely ocular myasthenia—the most important differential diagnosis is mitochondrial cytopathy, in which increased jitter may also occur. Other diagnoses to consider include oculopharyngeal muscular dystrophy and thyroid ophthalmopathy.

Botulism, caused by food poisoning, an infected wound, or clostridial overgrowth in the gastrointestinal tract in infants, may need to be considered. Features of autonomic malfunction are usually present.

Treatment

As noted, thymomas require excision, but this in itself will not improve the myasthenia. Management of patients with thymic tumours follows the same guidelines given below.

Anticholinesterase drugs, by reducing acetylcholine breakdown, give symptomatic improvement in most patients, and may be sufficient in those with very mild disease. Pyridostigmine is the drug of choice, given orally four or five times daily, starting at 30 mg per dose and increasing if required to 60 mg. Abdominal cramping is a common side-effect, relating to muscarinic overstimulation, and responds to propantheline, ideally taken 30 min before each dose of pyridostigmine. If an adequate response is not obtained at this dose, then further increases should not be made and other forms of therapy should be considered. The management of ocular myasthenia differs somewhat from the generalized form of the disease, the latter also depending upon age of onset and antibody status.

Ocular myasthenia

If anticholinesterase drugs have given an inadequate response, alternate-day prednisolone therapy should be introduced. A suitable starting daily dose is 5 mg, increasing by 5 mg every fourth dose (or weekly) until an adequate response has been obtained (often, for an adult, a dose of around 30 mg) or a maximum acceptable dose (around 0.75 mg/kg body weight) has been reached. Once remission has been achieved, the pyridostigmine can be withdrawn, and then the prednisolone reduced slowly— initially by 5 mg per month, but when down to 20 mg by as little as 1 mg, each month). Azathioprine may be added if there is an inadequate response or the minimal effective dose of prednisolone is deemed to be unacceptably high. Ocular muscle surgery can be beneficial if there is a poor or incomplete response to treatment and if the defect appears to be fixed.

Early-onset, seropositive myasthenia

Many, but not all, of these patients benefit from thymectomy. Up to one-third enter remission, and a further one-half improve. These benefits are occasionally rapid, but more typically develop over the following 1 to 2 years, possibly longer. The conventional approach is through a sternal split. There is concern that less invasive surgical procedures risk leaving behind thymic remnants which will negate the benefits of the operation. Thymectomy should be performed in centres experienced in such surgery and with the support of appropriately trained anaesthetists and neurologists.

For those patients who do not respond adequately to anticholinesterase drugs and thymectomy, immunosuppression with prednisolone and azathioprine is indicated. A controlled trial has shown the benefits of the addition of azathioprine (2.5 mg/kg body weight per day)—the starting dose is 25 or 50 mg daily, increased by 25 or 50 mg daily, each week (or more rapidly as an in-patient) until the target dose is reached. During introduction, weekly tests of full blood count and liver function are required. When established, testing can be reduced gradually to 3-monthly. Introduction of prednisolone may exacerbate myasthenic weakness and should generally be done in hospital. The starting dose is 10 mg on alternate days, increasing by 10 mg per dose until the patient reaches the target dose of 1 to 1.5 mg/kg body weight per dose. When remission has been achieved the dose is slowly reduced, as for ocular myasthenia, until the minimal effective dose has been established.

For those who do not respond to, or are intolerant of, prednisolone and/or azathioprine, other immunosuppressant drugs such as cyclosporin, methotrexate, or cyclophosphamide may be used.

Late-onset and seronegative myasthenia

Although not formally assessed, it appears that these patients do not benefit significantly from thymectomy. Most respond to the immunosuppressant regime described above.

Myasthenic crisis

Intubation and assisted ventilation may be required. Plasma exchange and intravenous immunoglobulin may both lead to a rapid improvement (within 1 to 2 weeks) in strength, but the beneficial effects start to wear off within about 8 weeks. However, this gives useful time in which to establish an immunosuppressant regime, as discussed above.

Plasma exchange and intravenous immunoglobulin are also useful in preparing myasthenic patients for thymectomy and may reduce the likelihood of deterioration

consequent upon the introduction of prednisolone.

Osteoporosis is an important concern in patients receiving long-term, high-dose prednisolone. A bone density determination should be carried out before starting such therapy, and repeated periodically, as appropriate. Local guidelines should be followed—these will advise on general physical measures, assess dietary calcium intake, and indicate the need to introduce calcium/vitamin D or a bisphosphonate, and the place of hormone replacement therapy for postmenopausal women.

Prognosis

The outlook for most patients with myasthenia is good, with over 90 per cent achieving near-normal functional recovery. Death is most likely to occur during a myasthenic crisis early in the course of the disease. The response to thymectomy has been noted. Unwanted effects relating to the immunosuppressant drugs may have an important influence on the outcome.

Myasthenia in pregnancy

Pregnancy has no significant long-term effect on myasthenia, but relapse may be more common in the puerperium. Some 10 per cent of infants born to myasthenic mothers have transient neonatal myasthenia due to transplacental passage of maternal anti-AChR antibodies. Symptoms include feeding and respiratory difficulties, generalized weakness, and less commonly, ptosis. They resolve within a few weeks.

Immunosuppressive treatment should be maintained during pregnancy to ensure good control of the mother's myasthenia and to reduce the likelihood of neonatal weakness.

Much more rarely, the infant is born with arthrogryposis multiplex congenita, secondary to profound intrauterine weakness and lack of movement. This relates to maternal antibodies which target the fetal form of the acetylcholine receptor (see above— Neuromuscular transmission) and in some cases the mother herself has been asymptomatic.

Future research

This may provide a better understanding of the immune processes involved, and thus lead to the development of selective treatments that avoid generalized immune suppression or other unwanted effects of the currently available drugs.

Lambert–Eaton myasthenic syndrome (LEMS)

This is a presynaptic disorder, characterized by limb-girdle weakness and symptoms of autonomic dysfunction, which is often associated with small-cell lung cancer. Delayed diagnosis is common. Symptomatic and immunosuppressant therapies are available.

Epidemiology

Some 60 per cent of patients have cancer-associated LEMS, caused by small-cell lung cancer and the peak presentation is in the fourth to sixth decades. The other 40 per cent have non-cancer-associated LEMS and may present from childhood onwards. It is estimated that 3 per cent of patients with small-cell lung cancer develop LEMS, but that the diagnosis is frequently not made. The weakness is often attributed to non-specific cachectic effects and the disorder is not either suspected or investigated. LEMS may predate the appearance of the cancer by as much as 5 years.

Pathogenesis

Both forms are associated with IgG class antibodies, which reduce the number of functional presynaptic P/Q-type voltage-gated calcium channels by cross-linking adjacent channels. This causes reduced calcium influx and therefore reduced quantal release of acetylcholine. As in myasthenia, patients with LEMS have an increased incidence of other forms of autoimmune disease, including a rare association with acquired myasthenia gravis.

Small-cell lung cancers express voltage-gated calcium channels and it is proposed that the tumour triggers an antibody response to those channels, the antibodies then cross-react with the calcium channels at the neuromuscular junction, causing LEMS.

Clinical features

Most patients present with an abnormality of gait and complain that their legs feel heavy or weak. Symptomatic upper limb weakness tends to present later. Autonomic dysfunction is common, but infrequently volunteered, and includes dryness of the mouth and constipation. In males, impotence may predate limb weakness. Compared with myasthenia gravis, ocular symptoms are rarely severe or particularly troublesome, and bulbar weakness is rare.

On examination, mild ptosis and diplopia may be evident. The abnormality of gait is often more striking than demonstrable weakness when testing on the examination couch. Partly this is because of the phenomenon of postexertional potentiation. Physiologically, with sustained effort there is mobilization of nerve calcium stores and consequently increased quantal release of acetylcholine. Clinically, this augmentation is apparent in two ways. First, strength increases after a few seconds of maximal effort. Second, the tendon reflexes, which are reduced or absent, increase or appear following 10 to 15 s of maximal contraction of the relevant muscle. Sensory testing is normal.

Diagnosis

Single-fibre electromyography, as in myasthenia gravis, shows increased jitter and blocking, and repetitive nerve stimulation studies show decrement at certain frequencies. However, the characteristic neurophysiological finding, which reflects the clinical observations made above, is of a small-amplitude compound muscle action potential which shows potentiation, sometimes enormous, 15 s after voluntary maximal contraction. Diagnosis is confirmed by demonstrating the presence of antivoltage-gated calcium channel antibodies, which are detectable in 95 per cent of cases.

Treatment

Pyridostigmine may offer some symptomatic benefit, but 3,4-diaminopyridine is more effective and the drug of choice (but is only available from specialist centres). 3,4-Diaminopyridine blocks the voltage-gated potassium channels (see Fig. 1), thereby prolonging the duration of the nerve action potential and allows a greater influx of calcium. The maximum dose of 3,4-diaminopyridine is 100 mg daily.

When an associated cancer is unlikely (young patients, non-smokers, more than 5 years since onset and no cancer apparent), treatment with alternate-day prednisolone (up to 1.5 mg/kg body weight per dose) and azathioprine (2.5 mg/kg body weight per day), as in myasthenia gravis, can be highly effective.

In a smoker in whom a cancer is not identified at presentation, it is prudent to repeat chest imaging (CT or MRI) yearly for 5 years.

In cancer-associated LEMS, removal of the tumour often leads to symptomatic improvement. Although there is some reluctance to use immunosuppression in patients with known cancer, alternate-day prednisolone may be used if there has been an inadequate symptomatic response to 3,4-diaminopyridine.

Plasma exchange and intravenous immunoglobulin both give short-term benefit and can be used in cancer-associated and non-cancer-associated LEMS.

Prognosis

In cancer-associated LEMS the prognosis is largely determined by the tumour. In non-cancer-associated LEMS many patients can be rendered symptom free, but

some prove very resistant to treatment.

Congenital myasthenic syndromes

This is a rare group of conditions with an overall prevalence in the order of 1 in 200 000 population. They are genetically determined (usually autosomal recessive—so a history of consanguinity is common), non-autoimmune disorders. Major clinical features include onset in infancy, fatiguable weakness, a decremental response to repetitive nerve stimulation, and absence of anti-AChR antibodies. A significant exception to this generalization is the classic slow-channel syndrome, which may present in infancy or adult life, and is inherited as an autosomal dominant trait. The syndromes may be classified on the basis of the site of the defect of neuromuscular transmission, but this is not always certain. A revised classification is likely to evolve as the molecular basis of each is identified. Diagnosis depends upon electrophysiological tests, morphological studies of the end-plate region in muscle biopsy specimens, and increasingly on identification of the specific genetic defect.

Presynaptic disorders

These are the least well characterized of the myasthenic disorders. They include disorders of acetylcholine resynthesis or packaging (previously known as familial infantile myasthenia, now called congenital myasthenic syndrome with episodic apnoea), and a recently described condition with paucity of synaptic vesicles with reduced quantal release. Symptoms respond to anticholinesterase drugs. The episodic apnoea syndrome has recently been shown to be caused by choline acetyltranasferase mutations.

H4>End-plate acetylcholinesterase deficiency

Fatiguable weakness is usually evident from birth. A single nerve stimulus may give rise to a repetitive compound muscle action potential response. The molecular basis is a mutation within the ColQ polypeptide gene. ColQ anchors acetylcholinesterase to the basal lamina. In the absence of the enzyme the acetylcholine receptors have a prolonged exposure to acetylcholine. Anticholinesterase drugs, not surprisingly, are ineffective. No specific treatments are available.

Postsynaptic disorders

These disorders are associated with mutations in the genes that encode the acetylcholine receptor subunits. They may affect the number of receptors or the kinetic properties of the central ion channel.

The most common in the United Kingdom is acetylcholine receptor deficiency, which is most frequently caused by mutations in the e-subunit gene. Presentation is at birth or within the first few years of life. There is generalized weakness, delayed motor milestones, feeding difficulties, and extraocular muscle involvement. There is a good response to anticholinesterase drugs and 3,4-diaminopyridine.

The low-affinity, fast-channel syndrome is phenotypically similar to acetylcholine receptor deficiency and may be associated with a-, †-, or e-subunit mutations. The mechanism is altered kinetics of the receptor ion channel.

The slow-channel syndrome is also a kinetic disorder, associated with mutations in different subunits and in different domains within those subunits. It is an autosomal dominant disorder with variable penetrance that may not become symptomatic until adult life or may remain subclinical. It tends to be progressive and characteristically produces weakness of periscapular muscles and of finger extensors. As in end-plate acetylcholinesterase deficiency, electromyography may show a repetitive response to a single nerve stimulus. Anticholinesterase drugs are unhelpful, but quinidine may be beneficial.

Neuromyotonia

This term describes a condition in which peripheral nerve overactivity leads to spontaneous muscle activity. It is thus quite different from classic myotonia, which relates to an abnormality of muscle fibre membrane activity. Neuromyotonia may be seen in association with a variety of inherited disorders (notably neuropathies and spinal muscular atrophy), but the commonest form is acquired. The acquired form may be idiopathic, but recognized associations include tumour (thymoma—sometimes also in association with myasthenia gravis; bronchial carcinoma) and acquired demyelinating polyneuropathies. Most acquired cases are autoimmune in origin and relate to the presence of antibodies directed against voltage-gated potassium channels in the peripheral nerve (Fig. 1), for which an assay is now available. As noted above, activation of these channels is an important factor in nerve repolarization—the symptoms of neuromyotonia can be understood in terms of prolonged depolarization and excessive release of acetylcholine.

The main clinical features are muscle stiffness, cramps, and twitching (myokymia), which may be localized or generalized. Voluntary muscle contraction may precipitate or exacerbate the abnormal activity. The myokymia persists during sleep and general anaesthesia. Additional symptoms include peripheral paraesthesias and excess sweating, and rarely, mood change, disturbed sleep, and hallucinations.

Apart from the muscle twitching (which may be confused with the fasciculation of denervation), physical examination may be normal. Mild weakness may be evident, proximally or distally. In long-standing cases, muscle hypertrophy (simply a form of work hypertrophy) may be present. Tendon reflexes may be reduced.

Electromyography shows highly characteristic, and diagnostic, doublet, triplet, or multiplet motor unit discharges, or periods of continuous motor unit discharge, with a high (up to 300 Hz) intraburst frequency. Fibrillation and fasciculation potentials may also be seen. Further confirmation of the diagnosis comes from antivoltage-gated potassium channel antibody assay, which is positive in about 50 per cent of cases using the currently available assay. Chest imaging should be considered to exclude thymoma and bronchial carcinoma.

Most patients gain symptomatic relief from carbamazepine, phenytoin, or lamotrigine. If the benefit is insufficient, immunosuppression with prednisolone and azathioprine is often helpful. Intractable cases may respond to plasma exchange and intravenous immunoglobulin.

Further reading

The neuromuscular junction and neuromuscular transmission

Vincent A (2001). The neuromuscular junction and neuromuscular transmission. In: Karpati G, Hilton-Jones D, Griggs R, eds. *Disorders of voluntary muscle*. Cambridge University Press, Cambridge.

Myasthenia gravis

Aarli JA (1999). Late-onset myasthenia gravis: a changing scene. *Archives of Neurology* **56**, 25–7.

Gajdos P *et al.* and the Myasthenia Gravis Clinical Study Group (1997). Clinical trial of plasma exchange and high-dose intravenous immunoglobulin in myasthenia gravis. *Annals of Neurology* **41**, 789–96.

Newsom-Davis J, Besson D (In press). Myasthenia gravis and myasthenic syndromes: autoimmune and genetic disorders. In: Karpati G, Hilton-Jones D, Griggs R, eds. *Disorders of voluntary muscle*. Cambridge University Press.

Lambert–Eaton syndrome

Elrington GM *et al.* (1991). Neurological paraneoplastic syndromes in patients with small cell lung cancer: a prospective survey of 150 patients. *Journal of Neurology, Neurosurgery and Psychiatry* **54**, 764–67.

Maddison P *et al.* (1999). Favourable prognosis in Lambert–Eaton myasthenic syndrome and small-cell lung carcinoma. *Lancet* **353**, 117–18.

Motomura M *et al.* (1995). An improved diagnostic assay for Lambert–Eaton myasthenic syndrome. *Journal of Neurology, Neurosurgery and Psychiatry* **58**, 85–7.

Motomura M *et al.* (1997). Incidence of serum anti-P/Q-type and anti-N-type calcium channel autoantibodies in the Lambert–Eaton myasthenic syndrome. *Journal of the Neurological Sciences* **47**,

35–42.

Congenital myasthenic syndromes

Beeson D *et al*. (1993). Primary structure of the human muscle acetylcholine receptor: cDNA cloning of the gamma and epsilon subunits. *European Journal of Biochemistry* **215**, 229–38.

Beeson D, Palace J, Vincent A (1997).Congenital myasthenic syndromes. *Current Opinion in Neurology* **10**, 402–7.

Engel AG, Ohno K, Sine S (1999). Congenital myasthenic syndromes. *Archives of Neurology* **56**, 163–7.

Neuromyotonia

Hart IK *et al.* (1997). Autoantibodies detected to expressed K^+ channels are implicated in neuromyotonia. *Annals of Neurology* **41**, 238–46.

Newsom-Davis J (1997). Autoimmune neuromyotonia (Isaacs' syndrome): an antibody-mediated potassium channelopathy. *Annals of the New York Academy of Sciences* **835**, 111–19.

24.18 Paraneoplastic syndromes

Jerome B. Posner

Introduction

The term paraneoplastic syndrome refers to disorders of an organ or tissue caused by cancer but occurring at a site distant from the tumour or its metastases. Organs or tissues commonly involved include the skin (for example paraneoplastic pemphigus), the endocrine system (for example paraneoplastic hypercalcaemia or paraneoplastic Cushing's syndrome), and the central and peripheral nervous system. The nervous system can be damaged directly or the damage may be indirect, such as when paraneoplastic hypercalcaemia or paraneoplastic Cushing's syndrome causes secondary nervous system dysfunction. Table 1 lists paraneoplastic syndromes involving the central and peripheral nervous system. Only the direct paraneoplastic syndromes are discussed here. More extensive reviews of paraneoplastic syndromes and other effects of systemic cancer on the nervous system can be found in the further reading section.

Incidence

Neurological examination of patients with cancer often reveals mild abnormalities such as proximal leg weakness or diminished Achilles reflexes. Most of these mild, usually subclinical, abnormalities result from metabolic or nutritional disturbances associated with advanced cancer. These disorders are not usually classified as paraneoplastic syndromes. In fact, most paraneoplastic syndromes occur in patients not known to have cancer at the time the neurological symptoms develop. Furthermore, most paraneoplastic syndromes are significantly disabling. If one limits oneself to 'true paraneoplastic syndromes' as indicated in Table 1, the disorders are rare. In one series of almost 1500 patients, only 3 had paraneoplastic cerebellar degeneration, and none had subacute sensory neuronopathy. With the exception of the Lambert–Eaton myasthenic syndrome (**LEMS**), a disorder that affects about 3 per cent of patients with small-cell lung cancer, paraneoplastic syndromes probably affect fewer than 1 per cent of patients with cancer.

Despite their low incidence, paraneoplastic syndromes are important for several reasons.

1. They usually develop before the cancer has been identified and so their presence may lead to the detection of small and potentially curable cancers.
2. Certain paraneoplastic syndromes characterized by specific autoantibodies suggest a specific cancer site.
3. The paraneoplastic syndrome is often more disabling than the cancer and may, in some instances, be the cause of death.
4. A paraneoplastic syndrome or an antibody associated with a paraneoplastic syndrome (see below) may predict an indolent course for the cancer.
5. The presence of specific antibodies in the serum of patients with a paraneoplastic syndrome identifies the neurological disorder as paraneoplastic and, as indicated above, sometimes strongly suggests the location of the underlying tumour.
6. Paraneoplastic antibodies identify proteins normally restricted to neurones that are of importance in the development and maintenance of neurones.

Pathogenesis

Current evidence suggests that paraneoplastic syndromes affecting the nervous system result from an autoimmune reaction to the tumour: protein antigens that are normally restricted to the nervous system are expressed in some cancers. The immune system recognizes the antigen in the cancer as foreign and some patients mount an immune response. The immune response has the dual effect of retarding the growth of the tumour but damaging those portions of the nervous system that express the antigen. A few of these so-called paraneoplastic antigens are normally expressed not only in the nervous system but also in the testis which, like the nervous system, is an immunologically privileged site. The best evidence for an immune-mediated mechanism comes from studies of the LEMS. Voltage-gated calcium channel proteins are found in all small-cell lung cancers. Some patients develop antibodies against these proteins. The antibodies react with voltage-gated calcium channels found in the presynaptic neuromuscular junction to prevent calcium from entering the junction, which in turn prevents the release of acetylcholine. The decreased release of acetylcholine causes the weakness that characterizes the syndrome. If one removes the antibody from the patient by plasma exchange, the patient's symptoms improve. If one injects experimental animals with the antibody from the patient, the syndrome is reproduced.

Evidence of immune-mediated mechanisms in most other paraneoplastic syndromes is less firm. However, several lines of evidence suggest that immune-mediated mechanisms are also pathogenetic in syndromes involving the central nervous system.

1. In some but not all patients with a paraneoplastic syndrome, the serum and cerebrospinal fluid contain high titres of autoantibodies that react with both tumour and those portions of the nervous system damaged by the paraneoplastic syndrome.
2. The titre of the antibody relative to total IgG is higher in cerebrospinal fluid than in serum.
3. In some patients, T-cell infiltrates of limited T-cell receptor families are found in the tumour and the nervous system.
4. One report describes intracellular IgG in the brain of patients who die of a paraneoplastic syndrome.

Several different autoantibodies have been found in the serum of patients with different paraneoplastic syndromes. The antibodies often identify patients with a specific clinical syndrome and with a specific underlying tumour (Table 2). For example, the anti-Yo antibody, which reacts with Purkinje cells of the cerebellum and some ovarian and breast cancers, is almost always associated with the syndrome of paraneoplastic cerebellar degeneration and with breast, ovarian, or other gynaecological cancers. In a few patients, the antibody and the paraneoplastic syndrome have been associated with other cancers, including cancers in men. Conversely, most patients with paraneoplastic cerebellar degeneration associated with non-gynaecological cancers either demonstrate other antibodies that react with Purkinje cells or do not have antibodies identifiable by current techniques. For example, some patients with paraneoplastic cerebellar degeneration express the anti-Hu antibody, an antibody usually associated with small-cell lung cancer and paraneoplastic encephalomyelitis or sensory neuronopathy.

Some autoantibodies are associated with specific tumours but widely varying paraneoplastic syndromes. For example, the anti-Hu antibody is almost always associated with small-cell lung cancer (occasionally neuroblastoma, prostate cancer, or mesenchymal chondrosarcoma), but may be associated with several different clinical syndromes usually encompassed by the term encephalomyelitis. The clinical abnormalities include limbic encephalitis (see below), paraneoplastic cerebellar degeneration, brainstem encephalitis, sensory neuronopathy, and autonomic failure. Some or all of these clinical abnormalities may be found in the same patient.

In some patients with paraneoplastic syndromes no antibodies are found. A good example is opsoclonus/myoclonus associated with neuroblastoma in children. Most observers believe that this paraneoplastic disorder is immune-mediated, articularly because it responds to ACTH. The failure to find an antibody does not mean that one is not present, only that current techniques cannot identify it. Another good example is LEMS. Standard histochemical and immunoblotting techniques cannot identify an antibody; only by the special techniques of electron microscopic histochemistry or immune precipitation can the P/Q type calcium channel antibodies be identified. These findings give some hope that by using better techniques, antibodies may be found in some currently antibody-negative paraneoplastic syndromes.

Diagnosis

Certain clinical clues suggest to the physician that a patient with a neurological disorder may be suffering from a paraneoplastic syndrome. These are summarized in

the following paragraphs.

Most paraneoplastic syndromes evolve subacutely. The disorder usually becomes apparent to the patient within a few days and often within several weeks has reached its peak. A few patients develop symptoms overnight. Occasional patients have a more protracted course, slowly evolving over several months. However, disorders that evolve slowly over many months to years are unlikely to be paraneoplastic.

Most paraneoplastic syndromes stabilize after weeks to months. As indicated above, symptoms usually progress rapidly but the progression usually ceases within several months.

The neurological disorders are usually severe. Most patients have substantial disability by the time they first consult a physician. Mild or waxing and waning neurological disorders are usually not paraneoplastic. For example, the patient with paraneoplastic cerebellar degeneration is usually unable to walk, unable to write, and sometimes because of the oscillopsia associated with nystagmus, unable to read or watch television. Many patients cannot sit unsupported.

The neurological findings are often characteristic. Those disorders are listed in Table 1. A subacutely developing pancerebellar disorder, the rapid development of opsoclonus (see below), or the development of LEMS strongly suggests cancer as the underlying cause. However, none of the syndromes, even the most characteristic, is invariably associated with cancer. Thus, only about two-thirds of patients with LEMS have cancer and only about 15 per cent of patients with myasthenia gravis have a tumour (almost always thymoma). Probably about half of patients with subacute cerebellar degeneration have cancer.

Paraneoplastic syndromes involving the central nervous system are often accompanied by cerebrospinal fluid pleocytosis, elevated protein, increased IgG, and oligoclonal bands. The spinal fluid findings are more likely to be abnormal early in the course of disease and revert to more normal values later on. In particular, the pleocytosis, rarely more than 30 to 40 cells, may be gone within a few weeks of the onset of disease. The immunoglobulin abnormalities usually persist.

Paraneoplastic syndromes often affect one particular portion of the nervous system with additional subtle or minor findings suggesting dysfunction in other areas. For example, paraneoplastic cerebellar degeneration selectively affects Purkinje cells of the cerebellum, causing pancerebellar neurological symptoms. Some patients will also be found to be mildly demented and demonstrate extensor plantar reflexes or sensory changes. These widespread changes have led to the term encephalomyelitis when the central nervous system is involved and neuromyopathy when the peripheral nerves and muscles are involved.

The physician encounters the patient with a paraneoplastic syndrome in one of two settings. In the first, the patient is known to have or have had cancer and then develops a neurological disorder. The cancer may be under active treatment, may have recurred after a remission, or may be assumed to have been cured in the remote past. Because paraneoplastic syndromes are the least common of the neurological complications of systemic cancer (Table 3), the physician must rule out all of the other causes before considering the disorder to be paraneoplastic. Unless a paraneoplastic antibody is found in the serum, the diagnosis is one of exclusion.

In the second setting, the patient is not known to have cancer. The physician must consider if the clinical findings fit a paraneoplastic syndrome (Table 1) and order the appropriate antibody studies (Table 2). Although the presence of a paraneoplastic antibody usually establishes the diagnosis, its absence does not. If an antibody is present and the search for an underlying cancer is negative, the physician is obligated to follow the patient carefully, searching periodically for a cancer.

Treatment

Some paraneoplastic syndromes, such as LEMS, respond to immunosuppression or to treatment of the underlying cancer (Table 4). Some syndromes, such as opsoclonus/myoclonus, may remit spontaneously, but for most paraneoplastic syndromes, treatment is unrewarding and most patients remain with severe neurological disability even if the cancer is cured. Most treatment has involved immunosuppression, particularly for those syndromes associated with autoantibodies. It is possible that the rapid onset of the syndromes does not allow sufficient time for accurate early diagnosis and for treatment to begin before irreversible neural damage has occurred. With earlier diagnosis, therapy may be more successful.

Specific syndromes

Brain and cranial nerves (Table 5)

Paraneoplastic cerebellar degeneration

The disorder may complicate any malignant tumour but is most common with lung cancer (especially small-cell), gynaecological neoplasms, and Hodgkin's disease. Males and females are both affected and the age incidence reflects the age distribution of the cancer. Neurological manifestations precede detection of the associated tumour in over one-half of patients, rarely by up to 4 years. Alternatively, paraneoplastic cerebellar degeneration may develop after diagnosis of the neoplasm. In some instances, the tumour is not found until autopsy. Typically, the disorder begins as gait ataxia that over a few days to months progresses to severe truncal and appendicular ataxia with dysarthria and often nystagmus. The nystagmus is frequently downbeating. Vertigo with or without nausea and vomiting is common and many patients complain of diplopia. The cerebellar signs are bilateral but may be asymmetrical. A more rapid onset within a few hours or days or a slower progression sometimes occurs. The cerebellar deficit usually stabilizes but, by then, the patient is often incapacitated. Spontaneous improvement sometimes occurs, particularly when associated with Hodgkin's disease.

The cerebrospinal fluid may be normal, but early in the illness usually shows a mild pleocytosis. Oligoclonal bands may be present. Cytological examination of the cerebrospinal fluid and contrast-enhanced MRI of the neuraxis rule out leptomeningeal metastases. MR scans typically are normal early, but later show signs of progressive cerebellar atrophy with prominent cerebellar folia and a dilated fourth ventricle.

The pathological hallmark of paraneoplastic cerebellar degeneration is loss of Purkinje cells, affecting all parts of the cerebellum. Less striking changes in the cerebellar cortex may include thinning of the molecular layer with microglial proliferation and astrocytic gliosis, proliferation of Bergmann astrocytes, and slight thinning of the granular layer with decreased numbers of granule cells.

When typical, the clinical picture of paraneoplastic cerebellar degeneration is almost pathognomonic. When atypical, the disorder must be distinguished from a cerebellar tumour (primary or metastatic) and from leptomeningeal metastases (by MRI and cerebrospinal fluid examination, respectively), from late-onset, non-paraneoplastic cerebellar degenerative disorders, cerebellar haemorrhage and infarction, abscess, prion diseases, cerebellar ataxia related to 5-fluorouracil or high-dose cytarabine, and metabolic disorders, especially alcoholic cerebellar degeneration. In alcoholic cerebellar degeneration, the ataxia predominantly involves the lower extremities; dysarthria and nystagmus are unusual.

There have been occasional reports of a partial or near-complete remission of paraneoplastic cerebellar degeneration following treatment of the primary tumour. The presence of ocular flutter or opsoclonus indicates a better chance of improvement, but this may be a different paraneoplastic syndrome (see below). Only rarely is immunosuppression beneficial and most patients do not respond. It is possible that if begun early in the illness, before Purkinje cells are irreversibly damaged, plasmapheresis and immunosuppressive drugs might have a beneficial effect. Symptomatic improvement in the ataxia occurs in a few patients using clonazepam in doses varying from 0.5 to 1.5 mg daily. Buspirone may also give modest relief.

Opsoclonus/myoclonus

Opsoclonus, a disorder of eye movements consisting of almost continuous arrhythmic, multidirectional, involuntary, high-amplitude conjugate saccades that are often accompanied by synchronous blinking of the lids, is a paraneoplastic syndrome complicating neuroblastoma in children and a variety of tumours in adults. Opsoclonus may be an isolated neurological sign, but is often accompanied by myoclonus of the trunk, limbs, head, diaphragm, larynx, pharynx, and palate, and ataxia. When opsoclonus is a paraneoplastic syndrome of adults it may be accompanied by paraneoplastic cerebellar degeneration. Opsoclonus/myoclonus is also associated with viral infections, postinfectious encephalitis, trauma, intracranial tumours, hydrocephalus, thalamic haemorrhage, and toxic encephalopathies from thallium or lithium, amitriptyline overdose, and diabetic hyperosmolar coma. Opsoclonus occurs in about 2 per cent of children with neuroblastomas. Neurological symptoms precede identification of the neuroblastoma at least 50 per cent of the time and the tumour often is not obvious on examination; thus, recognition of the neurological syndrome is an important clue to the presence of a neuroblastoma. When a neuroblastoma is associated with opsoclonus/myoclonus, there is a higher than expected incidence of intrathoracic tumours and of tumours with a benign histology. The prognosis of the neuroblastoma is better if opsoclonus/myoclonus is associated than when there

is no neurological complication, an observation not explained by earlier diagnosis when neurological symptoms are present. The neurological disorder responds to ACTH and to intravenous immunoglobulin but not to prednisone. However, most patients suffer residual neurological damage, usually cognitive.

Opsoclonus/myoclonus is less common in adults. Nevertheless, about 20 per cent of adult patients reported with opsoclonus/myoclonus have an underlying cancer. The neurological symptoms usually precede diagnosis of the tumour and commonly progress over several weeks, although more rapid or slower progression may be observed. Opsoclonus often is associated with truncal ataxia, dysarthria, myoclonus, vertigo, and encephalopathy. The cerebrospinal fluid may show a mild pleocytosis and a mildly elevated protein. The MRI is usually normal, but brainstem abnormalities have been reported.

Neuropathological findings have been variable. In some patients there are no identifiable abnormalities. In others, the changes resembled those of paraneoplastic cerebellar degeneration with a loss of Purkinje cells, inflammatory infiltrates in the brainstem, Bergmann gliosis, and loss of cells from the granular layer of the cerebellum.

The prognosis for recovery or partial remission of the neurological disorder is better for paraneoplastic opsoclonus/myoclonus than it is for paraneoplastic cerebellar degeneration. Improvement may follow treatment of the underlying tumour, and spontaneous partial remissions occur. Remissions have been reported to follow treatment of the tumour or immunosuppressive treatment including immunoadsorptive therapy using a protein A column.

Limbic encephalitis

Limbic encephalitis may occur as an isolated finding or as a more extensive encephalomyelitis. The neurological symptoms often precede diagnosis of the tumour by up to 2 years; sometimes the cancer is not detected until autopsy. Symptoms usually progress over several weeks but the course may be more insidious. Anxiety and depression are common early symptoms, but the most striking feature is a severe impairment of recent memory. Other manifestations include agitation, confusion, hallucinations, partial or generalized seizures, and hypersomnia. Progressive dementia usually occurs, but occasionally there may be a spontaneous remission. The cerebrospinal fluid commonly shows a pleocytosis and an elevated protein concentration. MR scans are usually normal but medial temporal abnormalities have been reported.

Inflammatory pathological changes affect the grey matter of the hippocampus, cingulate gyrus, pyriform cortex, orbital surfaces of the frontal lobes, insula, and the amygdaloid nuclei.

No treatment has proved uniformly beneficial although spontaneous remissions have been reported and some patients have improved after treatment of the underlying tumour.

Brainstem encephalitis

Paraneoplastic brainstem encephalitis is often associated with clinical and pathological evidence of encephalomyelitis elsewhere within the central and peripheral nervous systems, but may occur as the dominant or an isolated clinical finding. It is commonly associated with small-cell lung cancer, but an identical clinicopathological syndrome may be seen in the absence of a malignancy.

The clinical features vary according to the brainstem structures involved in the pathological process. Common manifestations include vertigo, ataxia, nystagmus, vomiting, bulbar palsy, oculomotor disorders, and corticospinal tract dysfunction. Less common clinical features include deafness, myoclonus of the branchial musculature, hypoventilation, and movement disorders including chorea or Parkinson's syndrome.

Neurological symptoms may develop before or after discovery of the malignancy. The pathological changes are identical to those observed in other forms of paraneoplastic encephalomyelitis.

Visual loss

Paraneoplastic syndromes can affect retinal photoreceptors, either rods or cones or both. They can cause a retinal vasculitis or optic neuropathy. Paraneoplastic retinal degeneration, also called cancer-associated retinopathy, usually occurs in association with small-cell cancer of the lung, melanoma, and gynaecological tumours. Typically, the visual symptoms include episodic visual obscurations, night blindness, light-induced glare, photosensitivity, and impaired colour vision. Visual symptoms usually precede the diagnosis of cancer. The symptoms progress to painless visual loss. They may begin unilaterally but usually become bilateral. Visual testing demonstrates peripheral and ring scotomas and loss of acuity. Funduscopic examination may reveal arteriolar narrowing and abnormal mottling of the retinal pigment epithelium. The electroretinogram is abnormal. Cerebrospinal fluid is typically normal, although elevated immunoglobulin levels have been reported. Inflammatory cells are sometimes seen in the vitreous by slit-lamp examination.

Pathologically, a loss of photoreceptors and ganglion cells with inflammatory infiltrates and macrophages is usually noted. The other parts of the optic pathway are preserved, although a loss of myelin and lymphocytic infiltration of the optic nerve may occur.

Treatment of cancer-associated retinopathy is usually unsuccessful although a recent report describes improvement in some patients with the use of intravenous immunoglobulin.

Spinal cord and dorsal root ganglia (Table 6)

Necrotizing myelopathy

This is an extremely rare remote effect of cancer. The initial symptoms of muscle weakness and sensory loss in the arms and legs may be asymmetrical, but eventually signs become bilateral and symmetrical. Back or radicular pain may precede other neurological signs. Cerebrospinal fluid abnormalities may include an elevated level of protein and a mild pleocytosis. Swelling of the spinal cord may be apparent on MRI. Typically, the neurological deficit progresses rapidly over days or a few weeks, ultimately leading to respiratory failure and death. There is no effective treatment.

Pathologically, there is widespread necrosis of the spinal cord, often most marked in the thoracic segments. The necrosis involves all components of the spinal cord with white matter usually more affected than grey matter.

Motor neurone disease (amyotrophic lateral sclerosis)

Paraneoplastic syndromes include: (i) amyotrophic lateral sclerosis with both upper and lower motor neurone dysfunction; (ii) progressive muscular atrophy, a pure lower motor neurone syndrome that is sometimes reversible and also associated with lymphoproliferative disorders; and (iii) primary lateral sclerosis, a pure upper motor neurone syndrome associated with solid tumours as well as lymphoproliferative disorders. The clinical and pathological characteristics differ little from non-paraneoplastic motor neurone disease save for the fact that the paraneoplastic disorders are often more rapid in onset and evolution, sometimes reverse spontaneously, and, at autopsy, may be more inflammatory than non-paraneoplastic disorders.

Myelitis

Paraneoplastic myelitis is usually a part of the encephalomyelitis syndrome with inflammatory lesions elsewhere in the brain and dorsal root ganglia as well as the spinal cord. The clinical picture is characterized by patchy wasting and weakness of muscles, sometimes combined with fasciculations. The upper extremities are often more severely affected than the legs, reflecting predominant involvement of the cervical spinal cord. There may be striking weakness of neck and intercostal muscles, resulting in respiratory failure. Sensory symptoms may be present. Autonomic dysfunction results from involvement of autonomic neurones.

Sensory neuronopathy

In contrast to the common axonal or demyelinating sensory neuropathies, paraneoplastic sensory neuronopathy where the dorsal root ganglion is the site of pathology

is a rare syndrome. At least two-thirds of the patients have small-cell lung cancer. Symptoms typically begin before the cancer is identified, with dysaesthetic pain and numbness in the distal extremities or occasionally in the arm(s), face, or trunk. The symptoms may be asymmetrical at onset but progress over days to several weeks to involve the limbs, trunk, and sometimes the face, causing a severe sensory ataxia. All sensory modalities are affected, distinguishing this disorder from cisplatin neuropathy, in which pin and temperature sensation are spared. Deep tendon reflexes are lost but motor function is preserved. The cerebrospinal fluid is typically inflammatory.

Early pathological changes are limited mostly to the dorsal root ganglia, in which both a loss of neurones and the presence of lymphocytic inflammatory infiltrates are noted. About 50 per cent of patients with paraneoplastic sensory neuronopathy have pathological changes that may be clinically inapparent in other regions of the nervous system.

In most patients, treating the underlying tumour or removal of the autoantibody by plasmapheresis or immunosuppressive therapy does not alter the course of the neurological disease, although there are isolated reports of responses to immunotherapy. Occasional patients have a mild and indolent neuropathy.

Peripheral nerves (Table 7)

Sensory and sensorimotor neuropathy

Peripheral neuropathies, particularly mild distal sensorimotor neuropathies, are quite common in patients with cancer. In one study of lung cancer the incidence was 16 per cent. The incidence is even higher if one defines the disorder by electrical evidence in clinically asymptomatic patients. However, many patients may have suffered from the metabolic or nutritional ravages of late cancer and would not be considered by the definitions here to have true paraneoplastic syndromes.

Some patients not known to have cancer, and who are not evidently systemically ill, present to the neurologist with a peripheral neuropathy which may be quite severe and disabling. It is estimated that in those patients whose initial evaluations do not reveal an obvious cause (such as vitamin deficiency, amyloidosis, diabetes), about 10 per cent will eventually prove to have cancer as the underlying reason for the peripheral neuropathy. Therefore, one should seriously consider a cancer diagnosis in such patients. Paraneoplastic peripheral neuropathy may take several clinical and pathological forms. The most common is the distal, symmetrical, subacutely developing, sensory neuropathy which may be either axonal or demyelinating. A relatively pure sensory neuropathy, a mononeuritis multiplex due to microvasculitis, and acute polyradiculopathy, a focal neuropathy such as brachial neuritis, or an autonomic neuropathy may also be paraneoplastic. Most of these neuropathies are not associated with autoantibodies and the diagnosis is often one of exclusion.

Neuromuscular junction and muscle (Table 8)

Paraneoplastic disorders of the neuromuscular junction include the Lambert–Eaton myasthenic syndrome and myasthenia gravis. These disorders have a common pathogenetic mechanism—they are caused by antibodies against ion channels and, whether paraneoplastic or not, they respond to immunological treatment. Another ion channel disorder included in this section is neuromyotonia, which is not confined to the neuromuscular junction. Finally, because of its similarity to neuromyotonia, the stiff person syndrome is also included in this section.

Lambert–Eaton myasthenic syndrome (LEMS)

LEMS results from a reduced release of acetylcholine at presynaptic nerve terminals. The same P/Q-type voltage-gated calcium channels are found in small-cell lung cancers. Interestingly, the richest source of P/Q-type voltage-gated calcium channels is the cerebellum, perhaps explaining the occasional relationship of paraneoplastic cerebellar degeneration and LEMS.

LEMS can be treated either by immune suppression or by treatment of the underlying cancer when present. Patients with small-cell lung cancer associated with LEMS have a better prognosis than patients with small-cell lung cancer who do not develop a paraneoplastic disorder.

Myasthenia gravis

Myasthenia gravis occurs in 30 per cent of patients with thymomas, and approximately 15 per cent of patients with myasthenia gravis are found to have a thymoma.

Polymyositis and dermatomyositis

Only a minority of patients suffering from these disorders have an underlying malignancy as the cause; elderly patients are more likely to have an underlying malignancy. Dermatomyositis with typical cutaneous changes is more likely than polymyositis to be paraneoplastic. Females and males are affected in approximately equal numbers. Symptoms of the muscle weakness generally precede identification of the cancer. The tumour may be at any site, but breast, lung, ovarian, and gastric malignancies are the most common. Hodgkin's disease and prostate and colon cancer are also reported offenders.

Corticosteroids, cyclosporine, and other immunosuppressants have been used successfully. Other reports suggest that high-dose intravenous immunoglobulin is useful in patients unresponsive to other forms of immunosuppression.

Neuromyotonia and stiff person syndrome

Muscle cramps are a common complication of cancer, sometimes related to electrolyte imbalance or induced by chemotherapy. A much rarer but clinically significant paraneoplastic disorder is acquired neuromyotonia. The disorder is characterized by muscle stiffness, cramps, and obviously rippling and twitching muscles, sometimes leading to sustained abnormal postures. Relaxation after voluntary contraction is delayed. Symptoms persist during sleep but are abolished by curare. Sudden prolonged bursts of high-frequency, involuntary, repetitive muscle action potentials are seen on electromyography.

The muscle spasms and rigidity are sometimes precipitated by activity, forcing patients to become sedentary. The disorder arises from peripheral nerves and is sometimes a part of the encephalomyelitis syndrome. The disorder is usually non-paraneoplastic, but may be associated with cancer including thymomas and small-cell lung cancer. Antibodies against voltage-gated potassium channels are often positive. Plasma exchange improves the patient's condition. Some patients respond to anticonvulsants. Injection of IgG from affected patients into experimental animals can reproduce the syndrome.

The stiff person syndrome may superficially resemble neuromyotonia, but has a central origin. The disorder is clinically characterized by stiffness and rigidity with episodic spasms of axial muscles. A variant of the syndrome affects the limbs. Painful reflex spasms can occur in response to tactile stimuli or startle. Muscle action potentials are normal on electromyography but the activity is continuous and excessive and increased by voluntary activity. The disorder is not usually associated with cancer, but in some patients the underlying syndrome is paraneoplastic.

Further reading

Dalmau JO, Gultekin HS, Posner JB (1999). Paraneoplastic neurologic syndromes: Pathogenesis and physiopathology. *Brain Pathology* **9**, 275–84. [Part of a comprehensive symposium of paraneoplastic syndromes in that issue of *Brain Pathology*.]

Darnell RB (1996). Onconeural antigens and the paraneoplastic neurologic disorders: At the intersection of cancer, immunity, and the brain. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 4529–36. [An excellent summary of the biology of paraneoplastic syndromes.]

Das A, Hochberg FH, McNelis S (1999). A review of the therapy of paraneoplastic neurologic syndromes. *Journal of Neuro-oncology* **41**, 181–94. [A review of current treatments.]

Griswold W, Drlicek M (1999). Paraneoplastic neuropathy. *Current Opinion in Neurology* **12**, 617–25. [A summary of peripheral neuropathy associated with cancer.]

Posner JB (1995). *Neurologic Complications of Cancer*. FA Davis, Philadelphia. [A comprehensive review of metastatic and non-metastatic neurological complications of cancer.]

24.19 Diseases of the peripheral nerves

P. K. Thomas

Pathophysiological considerations

The peripheral nerves consist of bundles (fascicles) of unmyelinated and myelinated axons that have their cell bodies in the anterior horns of the spinal cord, dorsal root ganglia, or autonomic ganglia. The fascicles are surrounded by a lamellated cellular sheath, the perineurium, which provides a diffusion barrier that separates the intrafascicular or endoneurial compartment from the extracellular tissues. Peripheral nerve trunks usually consist of several fascicles bound together by the mainly collagenous epineurial connective tissue. The nutrient vessels connect with a longitudinal anastomotic network of arterioles and venules in the epineurium. This in turn communicates through perforating vessels with a longitudinal intrafascicular capillary anastomotic network. This anastomotic system is extremely efficient: experimentally it is very difficult to produce ischaemia of nerve trunks by ligation of nutrient vessels. The occurrence of an ischaemic neuropathy, therefore, implies widespread vascular insufficiency. A blood–nerve barrier, comparable to the blood–brain barrier, exists in peripheral nerves (except in the sensory and autonomic ganglia). This, in conjunction with the diffusion barrier provided by the perineurium, probably regulates the composition of the endoneurial connective tissue fluid and thus the ionic environment of the nerve fibres.

All nerve fibres, whether myelinated or unmyelinated, are closely related to satellite cells, the cells of Schwann. There is evidence that they may provide metabolic support for the axons, which often extend for very considerable distances from their perikarya. In myelinated fibres the myelin segments are derived by the spiralling of Schwann cell surface membrane around the axons. The axon is exposed at the nodes of Ranvier, which represent the gaps between adjacent myelin segments. Conduction in unmyelinated axons takes place by the spread of a continuous wave of depolarization, the action potential, that migrates along the axolemma. In myelinated fibres, because of the high electrical resistance of the lipid in the myelin lamellae, the generation of the action potential is restricted to the region of the nodes of Ranvier. Conduction is therefore saltatory, jumping from one node to the next by local currents that traverse the axon and the extracellular tissue fluid. By this means, conduction velocity is increased from about 1 m/s in unmyelinated axons to 60 to 70 m/s in the largest myelinated fibres in human nerves.

The majority of the synthetic mechanisms in neurones are sited in the cell bodies. Synthesized materials are then transported down the axons to the termination of the fibres by an active transport system. This involves a fast system with a rate of about 400 mm/day, and a slow system, in which the structural proteins travel at 1 to 2 mm/day. The system is bidirectional: apart from the two anterograde fluxes, there is a retrograde system transporting materials, including neurotrophic factors, back from the periphery to the cell body. The retrograde system may be involved in the regulation of protein synthesis in the cell body and probably carries the signal for chromatolysis which ensues on transection.

Disorders of peripheral nerve function can be categorized in terms of the site of the primary disturbance. Conditions that lead to the death of the neurone as a whole, with the loss of the cell body and the axon, are categorized as neuronopathies. Conditions that have a selective effect on axons are termed axonopathies. A selective effect on axonal conduction is seen in poisoning with tetrodotoxin, which blocks the sodium channels at the nodes. Axonopathies may be focal or generalized. Focal axonopathies occur as a result of insults such as trauma or ischaemia. Axonal interruption leads to wallerian-type degeneration below the site of the injury. Recovery has to take place by axonal regeneration which is a slow process: the rate of axonal regeneration is about 1 to 2 mm/day.

Generalized axonopathies often lead to a selective degeneration of the distal portion of the fibres which then extends proximally. The axons are said to 'die back' towards the cell bodies. This pattern is seen in many toxic neuropathies and neuropathies due to nutritional deficiency. It has been suggested that in these conditions the axonal breakdown may result either from interference with enzymes involved in glycolysis which provide the metabolic energy for axonal transport mechanisms, or from cofactor deficiency or inactivation. As the enzymes are synthesized in the cell bodies and then transported down the axons, the further the distance from the cell body the greater will be the likelihood of the occurrence of metabolic insufficiency. This probably accounts for the distal distribution of many such neuropathies, as longer axons will be more vulnerable. Recovery again has to take place by axonal regeneration. In many distal axonopathies that involve the peripheral nervous system, not only does the degeneration affect the distal parts of the motor and sensory axons in the periphery, but the terminal parts of the centrally directed axons derived from the dorsal root ganglion cells also degenerate. Thus degeneration may be found in the rostral portions of the posterior columns. This process has been referred to as central–peripheral distal axonopathy. Neuropathy from iminodipropionitrile blocks the slow axonal transport system and leads to large swellings in the proximal parts of the axons that contain aggregations of neurofilaments (proximal axonopathy).

Other neuropathies primarily affect the myelin, either directly, or through an interference with Schwann cell function. The consequence is a selective demyelination with relative preservation of axonal integrity. This may be restricted to the region of the nodes of Ranvier (paranodal demyelination) or involve whole internodal segments (segmental demyelination) with consequent conduction block. The selective myelin damage may occur, for example, as the result of a cell-mediated attack on myelin by sensitized mononuclear cells, which is the likely explanation of the Guillain–Barré syndrome. Another instance is in diphtheritic neuropathy where the demyelination is considered to be secondary to an interference with Schwann cell protein metabolism. Local compression by a tourniquet also gives rise to selective damage to myelin through mechanical effects, although more severe pressure also causes axonal interruption. In diffuse demyelinating neuropathies, the distribution of the clinical effects, as for distal axonopathies, is often maximal peripherally. Presumably, this is a statistical effect: the longer the nerve fibre, the more likely it is to develop a region of demyelinating conduction block.

Recovery after paranodal or segmental demyelination occurs by remyelination. Initially, the newly formed myelin is thin, which results in an abnormally slow conduction velocity. Such reductions in conduction velocity may be focal, for example in relation to localized myelin damage in entrapment neuropathies, or widespread as in the Guillain–Barré syndrome or the inherited demyelinating neuropathies. In the latter, motor nerve conduction velocity is sometimes reduced to 10 m/s or less.

Finally, in other neuropathies the nerve fibres may be secondarily damaged by processes that primarily affect the connective tissues of nerve or the vasa nervorum. Usually a combination of demyelination and axonal loss occurs.

Clinical categories of neuropathy

Mononeuropathy, multifocal neuropathy, and polyneuropathy

Peripheral neuropathies may be divided into two broad categories depending upon the distribution of the involvement. The first category comprises lesions of isolated peripheral nerves or nerve roots termed mononeuropathy or multiple isolated lesions termed multifocal neuropathy (multiple mononeuropathy or 'mononeuritis multiplex'). The lesions in a widespread multifocal neuropathy may summate to produce a symmetrical disturbance, but the history or a careful examination may indicate the involvement of individual nerves. Isolated or multiple isolated peripheral nerve lesions arise from conditions that produce localized damage, such as mechanical injury, nerve entrapment, thermal, electrical, or radiation injury, vascular causes, granulomatous, neoplastic, or other infiltrations, and nerve tumours.

Secondly, there may be a diffuse and bilaterally symmetrical disturbance of function which can be designated polyneuropathy. When such a process affects the spinal roots, or affects the roots and the peripheral nerve trunks, the terms polyradiculopathy and polyradiculoneuropathy are sometimes employed. In general terms, polyneuropathies result from causes that act diffusely on the peripheral nervous system, such as metabolic disturbances, toxic agents, deficiency states, and certain instances of immune reaction. Isolated nerve lesions may sometimes be superimposed upon a symmetrical polyneuropathy, as a consequence, for example, of pressure lesions in a patient confined to bed. In certain polyneuropathies, there is an abnormal susceptibility to pressure lesions.

Symptomatology

Weakness or paralysis may be due either to conduction block in the motor nerve fibres or to axonal degeneration. Conduction block is related to demyelination with preservation of axonal continuity (neurapraxia). Recovery may occur by remyelination and may be rapid and complete. This can be the situation in localized nerve lesions, for example 'Saturday night' paralysis of the radial nerve, or in more widespread polyneuropathies, such as in acute inflammatory demyelinating polyneuropathy (Guillain–Barré syndrome). If axonal interruption takes place, axonal degeneration occurs below the site of interruption. The muscle weakness is accompanied by denervation atrophy and electromyographic signs of denervation. In a reversible process, recovery has to take place by axonal regeneration which is often slow and incomplete. An important recovery mechanism in conditions in which muscles become partially denervated is reinnervation of denervated muscle fibres by collateral sprouting from the remaining intact axons.

In generalized symmetrical polyneuropathies, the muscle weakness and wasting are commonly peripheral in distribution with an onset in the lower limbs. This results in bilateral footdroop and a 'steppage' gait in which an affected individual lifts his feet to an abnormal extent to avoid catching his toes against the ground. Involvement of the upper limbs begins with weakness and wasting of the small hand muscles and usually weakness of the finger and wrist extensors before the forearm flexor muscles are affected. At times, a symmetrical involvement of the proximal musculature in the limbs occurs in polyneuropathies, for example in the Guillain–Barré syndrome or porphyric neuropathy. Fasciculation due to spontaneous contraction of isolated motor units is most often a feature of anterior horn cell disease but may be encountered in peripheral neuropathies, as may muscle cramps. Postural tremor, mainly affecting the upper limbs and resembling essential tremor, may be seen in patients with chronic demyelinating polyneuropathies with slow conduction velocity. This 'neuropathic tremor' is most often encountered in type I hereditary motor and sensory neuropathy, chronic inflammatory demyelinating polyneuropathy, and IgM paraproteinaemic neuropathy. A rare manifestation of peripheral neuropathy is the occurrence of continuous repetitive discharges in motor nerve fibres leading to generalized muscular rigidity or 'neuromyotonia' (Isaacs' syndrome, continuous motor unit activity syndrome).

Loss of the tendon reflexes is a frequent accompaniment of a peripheral neuropathy, and usually first affects the ankle jerks. In assessing the clinical findings, it is important to remember that the ankle jerks may be lost in later life, probably as a result of senile changes in the peripheral nerves.

Sensory symptoms and sensory loss in symmetrical polyneuropathies are usually distal in distribution, giving rise to the 'glove and stocking' pattern of involvement. Only rarely is a proximal pattern encountered. The sensory loss may affect all modalities or be restricted to certain forms of sensation. If the loss is restricted, two broad patterns are discernible. In the first, the impairment predominantly affects the sense of joint position and vibration and touch–pressure sensibility, corresponding to a predominant loss of function in the larger myelinated nerve fibres. Sensory ataxia is the salient manifestation in 'large fibre' sensory neuropathies. Loss of postural sensibility may lead to sensory ataxia in the limbs and to 'pseudoathetosis', that is, involuntary movements, most often of the fingers and hands, that occur, for example, when a patient holds their arms outstretched with their eyes closed. In the second pattern of selective sensory loss, pain and temperature sensibility are predominantly affected, often associated with loss of autonomic function, corresponding to a predominant loss of small myelinated and unmyelinated axons. 'Trophic changes' and pain may complicate such small fibre neuropathies. The most important factor in their genesis is the loss of the protective effect of pain sensation with the consequent development of persistent ulceration or more extensive tissue loss, most commonly in the feet, and neuropathic joint degeneration.

Paraesthesiae are a frequent feature in peripheral neuropathy. These are usually of a tingling nature ('pins and needles'), but, especially in 'small fibre' neuropathies, may involve thermal sensations, most often with a burning quality. The paraesthesiae may be aggravated by touching or stroking the skin. Stimuli that are normally not painful may acquire an unpleasant quality (allodynia) and painful stimuli may give rise to an excessive or hyperpathic response, in which the stimulus, for example a pinprick, is abnormally intense. With repeated stimulation at the same site, the pain that is felt may spread widely and reach an intolerable intensity. An unusual symptom encountered most often in uraemic neuropathy is that of 'restless legs' (Ekbom's syndrome). Affected individuals experience sensations in the feet and legs that they find difficult to describe but which are temporarily relieved by movement of the feet and legs. 'Ekbom's syndrome' may also occur in the absence of any detectable disease process.

Spontaneous pains of an aching or lancinating character may complicate a number of generalized polyneuropathies. Severe paroxysms of lancinating pain occur in trigeminal neuralgia, but here the responsible lesion may well lie within the central nervous system. Causalgia constitutes a particularly troublesome painful syndrome, most often following gunshot wounds injuring the median nerve, the lower trunk of the brachial plexus, or the tibial nerve. It is a severe persistent pain, often with a burning quality that is characteristically aggravated by emotional factors. Sympathectomy relieves a high proportion of such cases.

Disturbances of autonomic function are occasionally the salient abnormality in a peripheral neuropathy, as in the Riley–Day syndrome, or they may accompany other manifestations, and can be observed both with localized peripheral nerve lesions and in generalized neuropathies.

Diagnosis and investigation

The history and physical examination frequently indicate that the disturbance has affected the peripheral nerves. If confirmation is required, this may usually be obtained by nerve conduction studies. Conduction may be examined in motor and sensory nerve fibres, and can give evidence of both localized and generalized neuropathies. Severely reduced conduction velocity may occur as a result of segmental demyelination or because of conduction in regenerating axons of small calibre after axonal degeneration.

Examination of the cerebrospinal fluid is not commonly of value in the diagnosis of peripheral neuropathies, although the substantially elevated protein content that often occurs in the Guillain–Barré syndrome may be helpful, as may inflammatory changes in some neuropathies.

Nerve biopsy is rarely required in establishing the existence of a peripheral neuropathy, but may be of diagnostic value in establishing the cause of the neuropathy, particularly in conditions that affect the vasa nervorum or neural connective tissues, in some 'storage' disorders and in inflammatory demyelinating neuropathies.

Individual nerves

Phrenic nerve (C2–C4)

This nerve innervates the diaphragm. When the diaphragm is totally paralysed, the normal protrusion of the upper abdomen during inspiration is lost, or is replaced by retraction (paradoxical movement). Radiographically, paralysis may be detected by unilateral or bilateral elevation of the diaphragm in a chest radiograph and its failure to descend on inspiration. The phrenic nerve may be involved in its course through the neck or thorax by wounds or tumours such as bronchial carcinoma, and it is sometimes affected in idiopathic brachial plexus neuropathy (neuralgic amyotrophy).

Nerve to serratus anterior (C5–C7)

The serratus anterior acts as a fixator of the scapula, holding the scapula against the chest wall when forward pressure is exerted by the arm. It is involved in forward movement of the shoulder as in a rapier thrust and in elevation of the arm, when it rotates the scapula. When serratus anterior is paralysed in isolation, the position of the scapula is normal at rest but if the extended arm is pushed forwards against resistance, 'winging' of the scapula becomes evident. The vertebral border, particularly in its lower portion, stands away from the chest wall. The nerve to serratus anterior may be involved in penetrating wounds, but usually in association with damage to the brachial plexus. It may be injured by forcible depression of the shoulder. Serratus anterior weakness is a common component of idiopathic brachial plexus neuropathy (neuralgic amyotrophy) and it is not infrequently encountered as an isolated and unexplained lesion.

Brachial plexus

The brachial plexus may be affected by penetrating wounds of the neck, in fractures and dislocations of the shoulder and clavicle, as a result of traction on the arm, by pressure from an aneurysm or a cervical rib, and by neoplastic involvement.

Traction lesions

Traction on the arm may result in damage to the plexus itself or may lead to avulsion of the spinal roots from the cord. If the roots are avulsed, sensory nerve action potentials will be preserved if recorded from affected fingers despite total anaesthesia, and the histamine flare response will be preserved in anaesthetized skin. This follows from the fact that the nerve fibres are interrupted proximal to the dorsal root ganglia and therefore the peripheral sensory axons do not degenerate.

In severe traction lesions, commonly encountered in current medical practice as a result of motorcycle or aircraft accidents, the whole of the plexus may be damaged. With forcible downward displacement of the shoulder, as when someone is thrown forwards and the shoulder strikes against an obstacle, only the upper part of the plexus, involving the contribution from the fifth and sixth cervical nerve roots, may be damaged. This may also be encountered as a birth injury from traction on the head, or on the trunk in a breech presentation (Erb's palsy), and rarely in anaesthetized patients during operation or in individuals carrying heavy rucksacks. Selective injury to the lower part of the plexus involving the contributions from the eighth cervical and first thoracic nerve roots occurs as a result of traction with the arm extended, as when an individual falls from a height and tries to save himself by hanging on to a ledge. It may also occur as a birth injury following traction with the arm extended (Klumpke's paralysis), but is less common than upper plexus damage.

Selective damage to the upper portion of the plexus (C5 and C6 roots or upper trunk) results in paralysis of deltoid, biceps, brachialis, brachioradialis, and sometimes supraspinatus, infraspinatus, and subscapularis. If the roots are avulsed from the cord, the rhomboids, serratus anterior, levator scapulae, and the scalene muscles will be affected. The arm hangs at the side, internally rotated at the shoulder, with the elbow extended and the forearm pronated in the 'waiter's tip' position. Abduction at the shoulder and flexion at the elbow are not possible. The biceps and brachioradialis jerks are lost. Sensory loss affects the lateral aspect of the shoulder and upper arm and the radial border of the forearm. Selective paralysis of the lower brachial plexus (C8, T1) results in paralysis of all the intrinsic hand muscles and a consequent claw-hand deformity, weakness of the medial finger and wrist flexors, and sensory loss along the medial border of the forearm and hand and over the medial two fingers. Cervical sympathetic paralysis, giving rise to Horner's syndrome, is frequently associated.

When the spinal roots are avulsed from the cord, regeneration is impossible and intractable spontaneous pain may be a highly troublesome sequel. Where the injury is distal to the dorsal root ganglia, lesions of the upper portion of the brachial plexus recover more satisfactorily than lower plexus lesions. The value of surgical repair is still a controversial issue. In the Erb's form of birth injury, weakness of abduction at the shoulder and flexion at the elbow often persist, although there may be little residual sensory loss. Full recovery takes place in about a third of the cases. It is less likely to occur with lower plexus injuries or if the whole plexus is involved. Early recognition and the application of measures to reduce the risk of joint contractures are important. Surgical treatment is of limited value.

Thoracic outlet syndromes

The contribution of the eighth cervical and first thoracic roots to the brachial plexus may be damaged by angulation over an abnormal rib or, more usually, a fibrous band arising from the seventh cervical vertebra and attached to the first rib. Although local structures such as the tendon of scalenus anterior may be involved in the production of symptoms, the isolation of a separate 'scalenus anterior syndrome' or of 'costoclavicular compression' is not justified. The subclavian artery may be affected by cervical ribs giving rise to aneurysmal dilatation and vascular symptoms such as Raynaud's syndrome and embolic phenomena, but the simultaneous occurrence of both neural and vascular phenomena is rare.

Damage to the lower part of the brachial plexus leads to weakness and wasting of the small hand muscles, and of the medial forearm wrist and finger flexors. Occasionally, there is selective wasting of the thenar pad in the hand, mimicking to some extent the appearances of the carpal tunnel syndrome. Numbness, pain, and paraesthesiae occur along the inner border of the forearm and hand, extending into the medial two fingers. The pain tends to be provoked by carrying heavy articles with the hand on the affected side. Horner's syndrome may be a feature. Nerve conduction studies are helpful when there are difficulties in distinguishing a cervical rib syndrome from a lesion of the ulnar or median nerves on clinical grounds. Surgical removal of the rib or fibrous band often leads to abolition of the pain and paraesthesiae, but recovery of power in the small hand muscles is usually disappointing.

Neoplastic involvement

Tumours may arise locally in the brachial plexus, such as neurofibromata in von Recklinghausen's disease (type I neurofibromatosis) or a solitary neurinoma, or the plexus may be invaded by tumours arising in other structures. In the latter case the commonest situation is involvement of the lower part of the plexus by an apical carcinoma of the lung (Pancoast's syndrome), which gives rise to wasting and weakness of the small hand muscles and of the medial forearm wrist and finger flexors, pain and sensory loss affecting the medial border of the forearm and hand, and cervical sympathetic paralysis. Other tumours that may invade the brachial plexus include carcinoma of the breast and malignant lymphomas affecting the lymph glands in the root of the neck.

Neuralgic amyotrophy

This condition was not clearly differentiated from the other painful paralytic disorders of the shoulder and upper arm, such as root compression from disc prolapse, until the Second World War. It has been described in a variety of terms, including 'idiopathic brachial plexus neuropathy' and 'paralytic brachial neuritis'. It may follow immunizing procedures, in particular the administration of antitetanus serum or operations, or occur without recognizable antecedent event. It can occur on a genetic basis as an autosomal dominant disorder, hereditary neuralgic amyotrophy (HNA), with variable penetrance. This has been mapped to chromosome 17p.

The disorder develops acutely with intense pain in the shoulder region which may take some weeks before it subsides completely although generally it ceases after a few days. Paralysis of the muscles of the shoulder girdle becomes evident within a day or two of the onset of the pain, sometimes also of the arms or of the diaphragm. It may be unilateral or bilateral and may be associated with sensory loss. More distal upper limb muscles may at times be affected, as may the phrenic nerve, and, occasionally, the recurrent laryngeal nerve. The cerebrospinal fluid is consistently normal. The affected muscles show electromyographic evidence of denervation. Recovery is usually ultimately satisfactory. Not all cases recover fully and recurrences may occur. A comparable disorder can affect the lumbosacral plexus (idiopathic lumbosacral plexus neuritis).

The pattern of muscle involvement and sensory disturbances suggests that the neuralgic myotrophy affects the brachial plexus in a patchy manner. An immune reaction is assumed but not established. The condition takes the same course whether or not it follows an immunizing procedure. Corticosteroids do not influence

either the initial pain or the ultimate outcome.

Post-irradiation brachial plexus neuropathy

Brachial plexus damage may occur as a sequel to radiotherapy for breast carcinoma or tumours in the neck. The onset of symptoms is usually several years after treatment, but may be within months. It can be difficult to distinguish from tumour recurrence but is less likely to be painful. Magnetic resonance imaging may be helpful in diagnosis.

Radial nerve (C5–C8)

The long course of the radial nerve and its position in relation to the humerus make this nerve unusually susceptible to external compression. It is a continuation of the posterior cord of the brachial plexus. In the upper arm it supplies triceps and anconeus and the skin on the back of the arm just above the elbow through the posterior cutaneous nerve of the arm. The lateral aspect of the lower part of the upper arm is supplied by the lower lateral brachial cutaneous branch and the dorsal aspect of the forearm by the posterior cutaneous nerve of the forearm. Muscular branches of the radial nerve innervate brachioradialis and extensor carpi radialis longus and brevis. The superficial branch of the nerve is its continuation. It descends along the radial border of the forearm and supplies the skin over the dorsum of the hand and the thumb, index, and middle fingers. The deep branch posterior interosseus nerve winds around the lateral aspect of the radius, passes through supinator, which it supplies, and innervates extensor digitorum, extensor digiti minimi, extensor carpi ulnaris, and often extensor carpi radialis brevis, abductor pollicis longus, extensor pollicis longus and brevis, and extensor indicis.

The nerve may be injured in wounds of the axilla so that the paralysis includes triceps, resulting in loss of extension at the elbow. The most frequent type of injury is compression of the nerve in the middle third of the arm against the humerus. This is encountered as 'Saturday night paralysis' in which an individual falls asleep when intoxicated with their upper arm over the arm of a chair. Triceps is spared, but brachioradialis, supinator, and all the forearm extensor muscles are paralysed. Sensory impairment is limited to the dorsum of the hand. Commonly the lesion consists of a localized conduction block (neurapraxia) so that muscle wasting does not occur and a muscle response can be obtained on electrical stimulation of the nerve below the level of the lesion. Recovery is complete within a matter of weeks. A cock-up wrist splint may be helpful while recovery is awaited. At times, there is some associated axonal degeneration so that electromyographic evidence of denervation is detectable and full recovery is correspondingly delayed.

Many muscles not supplied by the radial nerve work at a disadvantage when the wrist and finger extensors are paralysed. These defects must not be mistaken for signs of injury to other nerves. Owing to the flexed position of the wrist, gripping is impaired, but if the power of the wrist and finger flexors is tested with the wrist extended, it can be shown to be normal. The action of the interossei in abducting and adducting the fingers is also feeble when the wrist is flexed, but full power is demonstrable if these muscles are tested with the hand resting flat on a table.

A lesion of the posterior interosseus nerve gives rise to weakness confined to abduction and extension of the thumb and extension of the index finger. Supinator is spared, together with brachioradialis and the radial wrist extensors, and there is no sensory loss. The nerve may be compressed under the arcade of Frohse or during its transit through supinator.

Axillary nerve (C5, C6)

This is a branch of the posterior cord of the brachial plexus. It supplies deltoid and teres minor and the skin over deltoid through the upper lateral brachial cutaneous nerve. It may be damaged in injuries to the shoulder and the chief symptom is an almost complete inability to raise the arm at the shoulder. In the past, it was sometimes injured by pressure from a crutch ('crutch palsy').

Musculocutaneous nerve (C5, C6)

This nerve is rarely damaged alone, but may be involved in injuries to the brachial plexus. It supplies coracobrachialis, biceps, and brachialis and the skin over the lateral aspect of the forearm through the lateral cutaneous nerve of the forearm. Flexion at the elbow is still possible by brachioradialis, but is weak, and sensation may be impaired along the radial border of the forearm.

Median nerve (C6–C8, T_1)

The median nerve arises from the medial and lateral cords of the brachial plexus and descends with the brachial artery through the upper arm, entering the forearm deep to the bicipital aponeurosis. It has no muscular branches above the elbow. It supplies all the muscles in the anterior aspect of the forearm except flexor carpi ulnaris and the medial half of flexor digitorum profundus. The main trunk of the nerve supplies pronator teres, flexor carpi radialis, palmaris longus, and flexor digitorum superficialis. Through the anterior interosseus branch, it also supplies the lateral aspect of flexor digitorum profundus, flexor pollicis longus, and pronator quadratus. The main trunk passes deep to the flexor retinaculum of the wrist and its recurrent muscular branch supplies abductor pollicis brevis, opponens pollicis, and contributes to the innervation of flexor pollicis brevis. It also supplies the lateral two lumbrical muscles and the skin of the lateral aspect of the palm and the lateral three and a half digits over their palmar aspects and terminal parts of their dorsal aspects.

Lesions in the forearm

The median nerve may be injured in the region of the elbow or compressed at the level of the pronator teres muscle. Entrapment neuropathies in the upper forearm, however, are uncommon. Occasionally the anterior interosseus branch is involved in isolation.

Complete lesions of the median nerve at the elbow give rise to paralysis of pronator teres, the radial flexor of the wrist, the long finger flexors except the ulnar half of the deep flexor, most of the muscles of the thenar eminence, and the two radial lumbricals. In brief, there is an inability to flex the index finger and the distal phalanx of the thumb, flexion of the middle finger is weak, and opposition of the thumb is defective. The appearance of the hand has been described as simian; it shows ulnar deviation, the index and middle fingers are more extended than normal, and the thumb lies in the same plane as the fingers.

In more detail, pronation is incomplete and defective. The patient attempts to overcome this by rotating the whole limb at the shoulder. Paralysis of the wrist flexors is evident when attempts are made to flex against resistance. The tendon of flexor carpi ulnaris stands out alone and the hand goes into ulnar deviation. Flexion of the fingers is good in the ulnar two fingers, although weaker than normal. The index finger cannot be flexed, and the middle finger only incompletely. Flexion at the metacarpophalangeal joints is possible in all fingers, including the index, and flexion at these joints with extension at the interphalangeal joints is accomplished by the interossei and lumbricals. If the proximal phalanx of the thumb is immobilized, it will be found that flexion of the terminal phalanx is abolished because of paralysis of flexor pollicis longus. Paralysis of the thenar muscles gives rise to defective abduction and opposition of the thumb. By means of the adductor, the thumb can be drawn into the palm, but as the radial fingers cannot be flexed or the thumb opposed, it is impossible to place the tip of the thumb on the fingers.

Sensory loss is evident over the lateral three and a half digits and the lateral aspect of the palm, although individual variations occur. There is almost complete anaesthesia over the two terminal phalanges of the index and middle fingers. This degree of sensory loss, combined with the motor deficit, renders the thumb and index fingers almost useless and makes paralysis of the median the most serious single nerve lesion in the upper limb.

Vasomotor and trophic changes often ensue. The skin in the distribution of the median nerve tends to become reddened, dry, and atrophic. The pulp of the affected fingers becomes atrophic and ulceration occasionally develops in the tip of the index finger. The nails may become white and atrophic.

After a total transection of the nerve in the region of the elbow, even with a satisfactory surgical repair, recovery is slow and rarely complete, particularly with respect to the innervation of the hand.

With partial lesions of the median nerve in the arm or forearm, causalgia may be a troublesome consequence. This most often follows gunshot wounds. The pain develops at any time from a few hours to 45 days after the injury. The pain is severe and unremitting, frequently of a burning or smarting quality. Upon this may be superimposed severe paroxysms of pain provoked by touching or jarring the limb or by emotional factors. Vasomotor and sudomotor changes may be associated. The skin usually becomes dry and scaly, but excessive sweating may be a feature. The patient adopts a protective attitude towards the limb, so that fixation of the joints of

the fingers and wrist may develop, together with atrophic changes in the skin and subcutaneous tissue. About 80 per cent of cases of true causalgia are relieved by sympathectomy. Untreated, the pain gradually subsides over months or years.

Lesions at the wrist

The superficial situation of the median nerve at the wrist renders it liable to injury in lacerations sustained by falling against a window with the hand outstretched or in suicidal attempts. It may also be damaged as an occupational hazard by individuals who exert repeated pressure on the butt of the hand.

Much the most common lesion at this site is the carpal tunnel syndrome, in which the median nerve is compressed as it passes deep to the flexor retinaculum. The usual presentation is with acroparaesthesiae. These consist of numbness, tingling, and burning sensations felt in the hand and fingers, the pain sometimes radiating up the forearm as far as the elbow or even as high as the shoulder or root of the neck. The paraesthesiae are sometimes restricted to the radial fingers, but may affect all the digits as some fibres from the median nerve are distributed to the fifth finger through a communication with the ulnar nerve in the palm. The attacks of pain and paraesthesiae are most common at night and often wake the patient from sleep. They are then relieved by shaking the hand. The hand tends to feel numb and useless on waking in the morning but recovers after it has been used for some minutes. The symptoms may recur during the day following use, or at times if the patient sits with the hands immobile. Such symptoms of acroparaesthesiae may persist for many years without the appearance of symptoms of median nerve damage. In other patients, weakness of the thenar muscles develops, particularly of abduction of the thumb, and is associated with atrophy of the lateral aspect of the thenar eminence (Fig. 1). Sensory loss may appear over the tips of the median innervated fingers. Occasionally patients present with symptoms of median nerve deficit in the hand without attacks of acroparaesthesiae having occurred, or motor and sensory signs may be discovered incidentally in the absence of symptoms, particularly in older individuals.

Fig. 1 Thenar wasting in a patient with the carpal tunnel syndrome.

The symptoms are usually characteristic. If confirmation is required in atypical cases, this can generally be obtained by nerve conduction studies. In patients who are experiencing frequent attacks of acroparaesthesiae, the symptoms may be reproduced by inflating a sphygmomanometer cuff around the arm above arterial pressure for 2 min. At times percussion over the carpal tunnel may elicit a Tinel's sign, or symptoms may be provoked by hyperextension of the wrist or sustained flexion (Phalen's sign).

The majority of cases occur in middle-aged and often obese housewives. In younger women it is commonly associated with excessive use of the hands, and it may develop in males after unaccustomed use of the hands, such as in house painting or fly fishing. In these instances, tenosynovitis of the flexor tendons is responsible. It may also be caused by tuberculous tenosynovitis at the wrist or involvement of the wrist joint in rheumatoid arthritis. It may develop as a consequence of osteoarthritis of the carpus, perhaps related to an old fracture. Other predisposing causes are pregnancy, myxoedema, acromegaly, and infiltration of the flexor retinaculum in primary and hereditary amyloidosis.

In cases in which muscle weakness and wasting, or sensory loss, are present when the patient is first seen, treatment should be decompression of the nerve by section of the flexor retinaculum. In patients with acroparaesthesiae alone and in which the cause is probably tenosynovitis at the wrist, reduction in the amount of activity engaged in with the hands may be sufficient to allow the symptoms to subside. Injection of the carpal tunnel with a long-acting corticosteroid preparation may give temporary relief. Splinting of the wrist to reduce movement during the day may also be useful. If the symptoms persist despite conservative measures, decompression is then advisable.

The majority of patients with acroparaesthesiae are relieved by decompression. In patients with sensory impairment and cutaneous hyperaesthesia, such symptoms may persist for prolonged periods despite decompression. If denervation of the thenar muscles has been present for a long time, full recovery may not occur.

Ulnar nerve (C7, C8, T1)

This nerve arises from the medial cord of the plexus, usually with a contribution from the lateral cord. It descends in the medial side of the upper arm, passes around the elbow in the ulnar groove, and enters the forearm under an aponeurotic band between the humeral and ulnar heads of flexor carpi ulnaris. It then runs superficial to flexor digitorum profundus to the wrist and enters the hand between the pisiform bone and the hook of the hamate, superficial to the flexor retinaculum. After penetrating the hypothenar muscles, its deep branch crosses the palm and ends in flexor pollicis brevis.

In the upper arm, branches arise that supply flexor carpi ulnaris and the medial part of flexor digitorum profundus. In the forearm, the dorsal branch arises that winds around the ulna and supplies the skin over the dorsal aspect of the hand and the medial one and a half fingers. In the hand, a superficial branch supplies palmaris brevis and the skin over the medial aspect of the palm and the medial one and half fingers. The deep branch, after supplying the hypothenar muscles, innervates the interossei, the third and fourth lumbricals, adductor pollicis, and part of flexor pollicis brevis.

Lesions at the elbow

Total paralysis from lesions at this level, including the branches to flexor carpi ulnaris and flexor digitorum profundus, gives rise to wasting along the medial side of the forearm flexor mass. There is weakness of flexion of the fourth and fifth fingers. If the proximal portions of these fingers are held immobilized, flexion of the terminal phalanges is not possible. When the hand is flexed to the ulnar side against resistance, the tendon of flexor carpi ulnaris is not palpable. Paralysis of the hypothenar muscles abolishes abduction of the fifth finger. Paralysis of the interossei and the medial two lumbricals gives rise to the 'claw hand' deformity (Fig. 2). The action of these muscles is to flex the fingers at the metacarpophalangeal joints with the fingers extended at the interphalangeal joints. In the claw hand, the posture of the fingers is opposite to this, namely, extension of the metacarpophalangeal joints with flexion at the interphalangeal joints. Although all the interossei are paralysed, the defect is seen mainly in the ulnar fingers since the radial lumbricals supplied by the median nerve are still active. The long extensors of the fingers, being unopposed, overextend the proximal joints, and the flexor digitorum superficialis flexes the proximal interphalangeal joints.

Fig. 2 'Claw hand' deformity in a patient with an ulnar nerve lesion.

In the hand, there is wasting of the hypothenar muscles, of the interossei, and of the medial part of the thenar eminence. Movements of abduction and adduction of the fingers are weak, as is adduction to the extended thumb against the palm. Sensory loss affects the dorsal and palmar aspects of the medial side of the hand and the medial one and a half fingers.

The ulnar nerve may be damaged by dislocations or fracture dislocations at the elbow and is sometimes compressed in individuals who habitually lean on their elbows. Entrapment may occur in the cubital tunnel as the nerve underlies the aponeurotic band between the two heads of the flexor carpi ulnaris. This is most likely to occur in those performing heavy manual work or if there is an excessive carrying angle at the elbow, as may occur following a previous malunited supracondylar fracture of the humerus ('tardy ulnar palsy'). The medial wall of the cubital tunnel is formed by the elbow joint; osteoarthritis of the elbow can lead to osteophytic encroachment on the tunnel and compression of the ulnar nerve. In the cubital tunnel syndrome, the ulnar nerve is often palpably enlarged in the ulnar groove and for a short distance proximally. Ulnar nerve lesions are not infrequent in leprosy. Here the enlargement of the nerve tends to be maximal at a little distance above the elbow.

When it is suspected that the nerve has been subjected to repeated compression at the elbow, surgical transposition to the front of the medial epicondyle should be considered. If the nerve is compressed in the cubital tunnel, decompression by slitting the aponeurosis may suffice.

Lesions at the wrist or in the hand

Damage to the nerve at the wrist will spare the dorsal branch, so that cutaneous sensation over the dorsum of the hand and fingers is spared. A lesion just proximal to the wrist will give rise to sensory impairment on the palmar aspect of the hand and fingers alone, and weakness of all the ulnar-innervated intrinsic hand muscles. A slightly more distal lesion spares the superficial branch of the nerve and therefore produces no sensory deficit. Finally, damage to the deep palmar branch spares the hypothenar muscles, but causes weakness of the other ulnar-innervated small hand muscles. Lesions at the wrist or in the hand are usually the result of compression by ganglia or by repeated occupational trauma. Damage to the deep palmar branch, for example, may be caused by firm pressure in the palm from a screwdriver or drill. If occupational pressure is the cause, recovery follows cessation of the precipitating cause. Should improvement fail to occur after an appropriate interval, surgical exploration to establish whether a ganglion is present is merited.

It is not always easy on clinical grounds to decide whether the lesion is at the elbow or the wrist. Compression of the nerve in the cubital tunnel, for example, may spare the branches to the flexor carpi ulnaris and flexor digitorum profundus. In these circumstances, nerve conduction studies may be helpful, as they may in distinguishing between lesions of the ulnar nerve and damage to the eighth cervical and first thoracic spinal roots.

Lumbosacral plexus

Lesions of the lumbosacral plexus are not common. The plexus may be involved in pelvic malignancy, such as from carcinoma of the uterine cervix, bladder, prostate, or rectum, or be the site of a local neural tumour. It may be compressed by a haematoma in patients receiving anticoagulant therapy or suffering from haemophilia, or be involved in fractures of the pelvis. The lumbosacral cord may be compressed against the rim of the pelvis by the fetal head during parturition, with consequent weakness of the anterior tibial and peroneal muscles, and sensory impairment in the distribution of the fourth and fifth lumbar dermatomes. The superior gluteal nerve may also be affected. Recovery is initially good but may not be complete. The plexus may be affected in diabetic amyotrophy. Rare instances of idiopathic lumbosacral plexus neuropathy are encountered, comparable to the corresponding disorder that affects the brachial plexus.

Femoral nerve (L2–L4)

This nerve arises from the lumbar plexus, crosses the iliac fossa between the psoas and iliacus muscles, and enters the thigh deep to the middle of the inguinal ligament. In the iliac fossa it supplies the iliacus, and in the thigh, pectineus, sartorius, and quadriceps femoris, and anterior cutaneous branches to the front of the thigh. The continuation of the femoral nerve is the saphenous which supplies the skin over the medial aspect of the lower leg as far as the medial malleolus.

Damage to the femoral nerve causes weakness of knee extension, wasting of quadriceps, loss of the knee jerk, and sensory impairment over the front of the thigh and in the distribution of the saphenous nerve. With a proximal lesion, there may also be weakness of hip flexion from paralysis of iliacus.

The femoral nerve may be injured in fractures of the pelvis or femur, in dislocations of the hip, and at times during operations on the hip. It may be involved by psoas abscesses, tumours, or implicated in wounds of the thigh. It is commonly involved in large psoas muscle haematomas in haemophiliacs (see Section 22) and in diabetic amyotrophy. Owing to the rapid dispersion of the branches in the thigh, partial lesions are common from wounds at this site. The nerve to quadriceps is most often injured. The resulting paralysis causes considerable difficulty in walking as the knee cannot be locked in extension and gives way, especially when descending stairs. The saphenous nerve is sometimes damaged in operations for the treatment of varicose veins.

Obturator nerve (L2–L4)

The nerve emerges from the lateral border of psoas, crosses the lateral wall of the pelvis, and enters the thigh through the obturator foramen where it supplies gracilis, adductor longus and brevis, adductor magnus, obturator externus, and sometimes also pectineus, and the skin over the lower medial aspect of the thigh.

Damage to the obturator nerve results in weakness of adduction and internal rotation at the hip, pain in the groin, and sensory impairment on the medial part of the thigh. The nerve may be involved in neoplastic infiltration in the pelvis and can be damaged by the fetal head or by forceps during parturition.

Lateral cutaneous nerve of the thigh (L2, L3)

This nerve arises from the lumbar plexus, passes obliquely across iliacus and enters the thigh under the lateral part of the inguinal ligament. It supplies the skin over the anterolateral aspect of the thigh.

Meralgia paraesthetica is an entrapment neuropathy resulting from compression of this nerve as it passes under the inguinal ligament. It is more common in men, who are often obese, and may be unilateral or bilateral. The symptoms consist of numbness in the territory of the nerve combined with tingling or burning paraesthesiae provoked by prolonged standing, or following excessive walking. Weight loss may be helpful, and in many instances the condition subsides spontaneously. Decompression of the nerve is rarely necessary.

Sciatic nerve (L4, L5, S1–S3)

The sciatic nerve enters the thigh through the sciatic notch. It is composed of the tibial and peroneal divisions which are usually bound together within a common sheath, the tibial division lying medially. It descends through the posterior aspect of the thigh, initially deep to gluteus maximus, and supplies semitendinosus, semimembranosus, and the long head of biceps through its peroneal division. It separates into the tibial and common peroneal nerves in the lower thigh, which supply all the muscles below the knee, and both nerves contribute to the formation of the sural nerve.

Total interruption of the sciatic nerve gives rise to foot drop. Walking is possible, but the patient cannot stand on the toes or the heel of the affected foot and the ankle is unstable. All movement below the knee is paralysed. If the injury is in the upper thigh, flexion of the knee is also weak. The skin is completely anaesthetized over the entire foot except for the medial border which is supplied by the saphenous nerve. Pressure sores may develop. The anaesthesia extends upwards on the posterolateral aspect of the calf in its lower two-thirds. The sense of joint position is abolished in the foot and toes. Beyond this area of complete anaesthesia, there is a wide zone in which sensibility may be diminished. Sweating is absent on the sole and dorsum of the foot, but is preserved on the medial side. The ankle jerk is lost

but the knee jerk is retained.

The sciatic nerve may be involved in pelvic tumours and can be injured by fractures of the pelvis or femur or during hip replacement operations. After the radial and ulnar, it is implicated in gunshot wounds more frequently than any other nerve. Partial injury of the tibial division may be followed by causalgia. Incomplete lesions of the nerve may be caused by pressure of the nerve against the hard edge of a chair in individuals who fall asleep while intoxicated. Similar lesions may occur in diabetic subjects, in whom the peripheral nerves are more susceptible to pressure neuropathy.

The syndrome of root pain and sciatica is considered in Chapter 24.3.11.

Tibial nerve (L4, L5, S1–S3)

After separating from the peroneal division of the sciatic nerve in the lower thigh, this nerve passes through the popliteal fossa and enters the calf deep to gastrocnemius through the fibrous arch of soleus. It descends through the calf to the medial side of the ankle, passes beneath the flexor retinaculum, and divides into the medial and lateral plantar nerves. It supplies popliteus, all the muscles of the calf, and, through the plantar nerves, the small muscles of the sole of the foot and sensation to the sole.

When the nerve is interrupted, the patient is unable to plantarflex or invert the foot, to flex the toes, or to stand on the ball of the foot. Paralysis of the interossei leads to a claw-like deformity of the toes. Sensation is lost over the sole. Causalgia may arise after partial lesions. Injury to the distal portion of the nerve by a penetrating injury or deep wound of the calf gives rise to paralysis of the intrinsic muscles of the foot but spares the muscles acting at the ankle. Sensation is lost on the sole of the foot and this may be accompanied by pain. If the injury is distal to the origin of the branches to flexor hallucis longus and flexor digitorum longus, the lesion may escape detection since paralysis of the small foot muscles and sensory loss over the sole may be overlooked.

The tibial nerve is occasionally compressed under the flexor retinaculum (tarsal tunnel syndrome), usually precipitated by osteoarthritis or post-traumatic deformities at the ankle or by tenosynovitis. Burning pain and tingling paraesthesiae occur in the sole, usually following prolonged standing or walking. The condition is generally unilateral. Careful examination may demonstrate wasting of the intrinsic muscles in the medial aspect of the foot, and sensory impairment over the sole. Nerve conduction studies may be helpful diagnostically. Treatment is by surgical section of the flexor retinaculum.

Painful neuromas sometimes develop on the digital branches of the plantar nerves. These give rise to the syndrome of Morton's metatarsalgia in which pain occurs in the anterior part of the foot on standing. A localized area of tenderness is detectable on palpation. The condition is relieved by excision of the neuroma.

Common peroneal nerve (L4, L5, S1, S2)

After separating from the tibial division of the sciatic nerve in the lower part of the thigh, this nerve descends through the popliteal fossa, winds around the neck of the fibula, and divides into its superficial and deep branches. The superficial peroneal nerve passes down in front of the fibula, supplies peroneus longus and brevis, and emerges in the lower leg, supplying the skin on the lateral aspect of the lower leg. It crosses the extensor retinaculum and supplies the skin on the dorsum of the foot and the second to the fifth toes. The deep peroneal branch continues to wind around the fibula, pierces the anterior intermuscular septum, and descends on the anterior interosseous membrane. It innervates tibialis anterior, extensor digitorum longus, extensor hallucis longus, and peroneus tertius. It passes deep to the extensor retinaculum after which it supplies the extensor digitorum brevis and the skin of the adjacent sides of the first and second toes.

Damage to the common peroneal nerve is more frequent than injury to its two branches because of its vulnerable superficial position at the neck of the fibula. It gives rise to foot drop with paralysis of dorsiflexion and eversion at the ankle and of toe extension. Cutaneous sensation is impaired over the lateral aspect of the lower leg and ankle and on the dorsum of the foot.

The common peroneal nerve may be compressed at the neck of the fibula by habitual sitting with the legs crossed, prolonged squatting, pressure during sleep or while anaesthetized, and various other events. It can be damaged by traction caused by fractures of the tibia and fibula and is sometimes damaged by ischaemia in the anterior tibial compartment syndrome. Paralysis caused by external pressure frequently gives rise to a local conduction block (neurapraxia) with satisfactory recovery within a few weeks. If electromyography indicates that nerve degeneration has taken place a foot drop support should be provided while axonal regeneration is awaited.

Sural nerve (L5, S1–S2)

This arises from the sciatic nerve and descends to the back of the calf, winds around to the lateral side of the ankle, and reaches the lateral border of the foot. It supplies the skin in this distribution. Sensory impairment occasionally results from pressure on the nerve as it lies in a superficial situation in the back of the calf.

Generalized neuropathies

Neuropathies related to metabolic and endocrine disorders

Diabetes mellitus

A significant degree of peripheral neuropathy develops in about 15 per cent of patients with diabetes, although a substantially greater number either have minor symptoms without signs, or evidence of a subclinical neuropathy either on clinical examination or on the basis of abnormalities of nerve conduction. In general, the neuropathies that appear can be divided into symmetrical sensory and autonomic polyneuropathies on the one hand, and isolated peripheral nerve lesions or multifocal neuropathies on the other. Mixed syndromes are common.

The commonest form is a symmetrical sensory polyneuropathy, giving rise to numbness and tingling paraesthesiae in the toes and feet and less often in the fingers. Aching or lancinating pains in the feet and legs, particularly at night, may be a troublesome feature. Examination reveals loss of vibration sense in the feet, depression of the ankle jerks, and distal cutaneous sensory impairment. Neuropathic plantar ulcers and occasionally Charcot joints are an important complication. Loss of pain sense results in perforating ulcers on the feet and neuropathic joint degeneration, particularly in the toes and in the tarsal joints; impaired postural sense may give rise to an ataxic gait. An acute painful diabetic neuropathy also occurs that predominantly affects the lower limbs. The onset is often associated with poor diabetic control and precipitate weight loss ('diabetic neuropathic cachexia').

Autonomic neuropathy frequently accompanies the sensory neuropathy and may be the salient manifestation. It rarely occurs in isolation. Pupillary disturbances usually take the form of a reduced response to light. Gustatory facial sweating provoked by the smell and taste of food can be troublesome. Anhidrosis may occur distally in the limbs; if it is extensive and also affects the trunk, heat intolerance may result. Symptoms referable to the alimentary tract include dysphagia from oesophageal involvement, episodes of vomiting related to gastric atony, and episodic nocturnal diarrhoea, often alternating with periods of constipation. Those related to the genitourinary system include impotence, retrograde ejaculation, and bladder atony with difficulty in voiding and urinary retention with overflow. Vascular denervation sometimes results in orthostatic hypotension, and cardiac denervation may be demonstrable by an elevated resting heart rate and the absence of beat-to-beat variation with respiration. The risk of diabetic polyneuropathy is reduced by strict glycaemic control.

Isolated nerve lesions tend to occur more commonly in elderly diabetic subjects. At times they develop insidiously, at others they have an abrupt onset with pain. Of the cranial nerves, the nerves to the external ocular muscles, particularly the third and sixth, and also the facial nerve, are the most often affected. In contradistinction to the effects of compression of the third nerve by a carotid aneurysm, the pupillary innervation is often spared. In the limbs, the lesions tend to occur at the common sites of compression or entrapment. It seems likely that the nerves of diabetics exhibit an excessive vulnerability to damage from pressure.

Diabetic amyotrophy, or proximal diabetic neuropathy, represents a particular example of a multifocal neuropathy that develops usually in elderly obese diabetics. It consists of an asymmetric proximal motor syndrome that affects the anterior thigh muscles and hip flexors, and sometimes also the anterolateral muscles of the lower leg. Less commonly it is symmetric. Its onset may be acute or insidious and is often accompanied by pain, particularly at night. There is generally little or no associated sensory loss. The knee jerks are usually depressed or absent. Inflammatory lesions including vasculitis have recently been demonstrated in peripheral

nerves in proximal diabetic neuropathy, leading to trials of immunomodulatory therapy.

The causation of diabetic neuropathy is uncertain. It tends to occur more often in poorly controlled diabetics, but the correlation is not close. It may occur for the first time on initiation of treatment with insulin, or be the presenting symptom in maturity onset diabetes. There is evidence to suggest that diabetic microangiopathy is important in the genesis of isolated nerve lesions. Metabolic factors are probably more important in the origin of the symmetric polyneuropathies, but their nature is uncertain. An increased concentration of sorbitol in nerves secondary to hyperglycaemia may be involved in causing nerve fibre dysfunction.

Focal peripheral nerve lesions and diabetic amyotrophy, if of acute onset, often recover adequately, as does acute painful diabetic neuropathy when satisfactory glycaemic control is achieved. Symmetric sensory and autonomic neuropathy, once established, recovers less satisfactorily, even with good diabetic control. Correcting the hyperglycaemia by continuous subcutaneous insulin infusion or pancreatic transplantation will stabilize the neuropathy. Trials of aldose reductase inhibitors to reduce sorbitol accumulation have, so far, not given clear evidence of improvement in neuropathy.

Care of the feet is vitally important in diabetic sensory neuropathy, to prevent the development of chronic ulceration. Pain may sometimes be helped by carbamazepine, tricyclic antidepressants, phenothiazines, or mexiletine. Hypotension can be improved by raising the head of the bed at night or by support bandages to the legs; more severe cases may require treatment with fludrocortisone. Gastroparesis may respond to metoclopramide, domperidone, or erythromycin; persistent vomiting may necessitate a Roux-en-Y gastroenterostomy. Diabetic diarrhoea can be helped by low-dosage tetracycline or diphenoxylate, loperamide, or codeine phosphate. Diabetic cystopathy can be managed in the earlier stages by regular voiding and cholinergic treatment with bethanechol. Urinary tract infections should be treated promptly. Bladder neck resection can be useful in carefully selected cases. Penile papaverine injections can be employed for erectile impotence, and sildenafil (Viagra) may be helpful in early cases. Silicone implants should be avoided because of the risk of infection.

Amyloidosis

The various forms of amyloid disease are described in Section 11.12. The peripheral nerves may be involved in primary amyloidosis due to a benign plasma cell dyscrasia and in amyloidosis related to myeloma (light chain amyloidosis). There are also several dominantly inherited forms of amyloid neuropathy, the most important of which are due to mutations in the gene for transthyretin (*TTR*), including the Portuguese type (see later). Isolated lesions may occur from the infiltration of amyloid into nerves or from compression of the median nerve in the carpal tunnel because of deposits in the flexor retinaculum. More strikingly, a generalized neuropathy may develop. It begins with selective loss of pain and temperature sensation in the feet and later in the hands. Motor involvement, loss of tendon reflexes, and impairment of other sensory modalities occur later. Autonomic involvement is an early feature, causing impotence, orthostatic hypotension, bladder atony, and disturbances of alimentary function. Amyloid deposits are present in the peripheral nerve trunks, which may be enlarged, and in the dorsal root and sympathetic ganglia.

No treatment influences the progress of the neuropathy apart from liver transplantation in neuropathy due to *TTR* mutations (see Section 11). The use of stem cell transplantation is being explored in amyloidosis related to malignant plasma cell dyscrasias. The spontaneous pains are sometimes improved by carbamazepine or tricyclic antidepressant drugs. Care must be taken to prevent damage to the anaesthetic feet, lower legs, and hands. Autonomic symptoms may require treatment as described for diabetic neuropathy.

Carpal tunnel syndrome is frequent in patients on long-term haemodialysis related to deposition of amyloid in the flexor retinaculum derived from retained b_2-microglobulin.

Uraemia

Uraemic neuropathy did not become a clinical problem until the advent of treatment of endstage renal failure by haemodialysis. It occurs in patients with severe chronic renal failure. It was most often seen in patients under treatment with periodic haemodialysis but is now much less frequently a problem. The symptoms are usually predominantly sensory, with numbness and tingling paraesthesiae in the feet. 'Restless legs' (Ekbom's syndrome) are often a conspicuous feature (see Section 24.22). A distal motor neuropathy may be associated and occasional cases are purely motor. The condition is improved by increased haemodialysis and more effectively by renal transplantation. A retained metabolite is assumed to be the cause, but this has not so far been identified.

Myxoedema

Compression of the median nerve in the carpal tunnel in myxoedema has already been discussed. Rarely a generalized mixed motor and sensory neuropathy develops. This improves on treatment of the hypothyroidism.

The slow contraction and relaxation observed in the tendon reflexes is not due to a disturbance of peripheral nerve function, but to an alteration in the contractile mechanism of the muscle fibres.

Acromegaly (see Section 12)

The occurrence of the carpal tunnel syndrome in acromegaly has also been mentioned. A rare manifestation of this condition is a sensorimotor polyneuropathy in which the peripheral nerves are thickened because of an overgrowth of the neural connective tissues. A similar neuropathy is occasionally observed in pituitary gigantism.

Critical illness polyneuropathy

A generalized polyneuropathy involving widespread axonal degeneration may be encountered in patients in intensive care units with sepsis and multiple organ failure. The neuropathy is discovered when attempts are made to wean them from the ventilator. The precise cause of this condition which has been termed critical illness polyneuropathy is unknown. Slow recovery occurs.

Other metabolic disorders

It has been claimed that a generalized peripheral neuropathy may be caused either by acute or chronic hepatic failure, but this is probably uncommon. A mild painful sensory neuropathy is occasionally encountered in primary biliary cirrhosis, sometimes related to xanthomatous deposits in the cutaneous nerve trunks. A motor neuropathy is a rare sequel to severe recurrent hypoglycaemia.

Toxic neuropathies

Industrial, environmental, and pharmaceutical substances

Acrylamide

This substance is widely employed industrially. The monomer is neurotoxic and causes peripheral neuropathy with mixed motor and sensory features. Ataxia is prominent and is possibly the result of concomitant cerebellar damage. Distal axonal degeneration occurs and slow recovery takes place on cessation of exposure.

Arsenic

Arsenical poisoning is occasionally seen as a result of accidental or homicidal ingestion of insecticides containing arsenic, or from indigenous medicines in India. Gastrointestinal symptoms develop after acute ingestion, followed by a mixed sensory and motor neuropathy after 1 to 3 weeks. Desquamation of the skin of the feet and hands takes place after about 6 weeks and white lines (Mees' lines) appear in the nails. With ingestion of smaller quantities on a chronic basis, gastrointestinal symptoms are less obtrusive and a slowly progressive neuropathy makes its appearance. The skin may become generally pigmented or show focal 'raindrop' pigmentation, and hyperkeratosis of the palms of the hands and soles of the feet may appear.

Slow recovery in the neuropathy occurs with removal from exposure. Chelating agents are of value in treating the non-neurological complications, but it is uncertain whether they are effective for the neuropathy.

Lead

Lead neuropathy is now a rare occurrence in Britain, although it was encountered as a consequence of the contamination of drinking water by lead pipes in old buildings. Subclinical neuropathy may be detectable in lead workers. It remains a hazard in certain parts of the world from the use of lead glazes in pottery. Lead neuropathy is predominantly motor, typically giving rise to wrist and foot drop. The 'lead colic' that may occur is probably a manifestation of autonomic involvement. Other features of lead poisoning that may be associated include a sideroblastic anaemia and a 'lead line' on carious teeth. The neuropathy improves on cessation of lead intake; the utility of treatment with BAL (dimercaprol), edetate, or penicillamine is uncertain.

Mercury

Exposure to inorganic mercury salts and to organic mercurial compounds may lead to neurological damage, as in 'Minamata disease' which was related to consumption of fish contaminated by organic mercury. Dementia, cortical blindness, and ataxia occur, together with sensory changes in the limbs attributed to a sensory neuropathy, although how far these have a peripheral origin is uncertain. Historically, a peripheral neuropathy was an important component of 'pink disease' which was caused by the administration of mercury-containing purgatives.

Thallium

This is present in certain pesticides and rodent poisons and was formerly used as a depilatory agent. Accidental or homicidal poisoning is occasionally encountered. Abdominal pain and diarrhoea are followed after 2 or 5 days by the development of a mixed motor and sensory neuropathy which is often painful. Evidence of central nervous system damage may be present with behaviour disorder, optic neuropathy, and choreiform movements. Alopecia develops later, after about 2 or 3 weeks, and renal damage may be produced. Diethyldithiocarbamate, which binds thallium, has been employed in treatment.

Triorthocresyl phosphate

This substance is used industrially as a high-temperature lubricant. Outbreaks of a sensorimotor neuropathy, often accompanied by evidence of damage to the central nervous system, occur periodically, usually as a consequence of the contamination of cooking ingredients or utensils. The original description was in relation to illegal liquor distillation (ginger jake paralysis) in the United States during the prohibition era. In more recent years, a large outbreak occurred in Morocco from the use of contaminated cooking oil. Recovery is slow and often incomplete.

Other industrial substances

Carbon disulphide, used in the manufacture of rayon, occasionally gives rise to a mild sensory neuropathy. Neuropathy may occur as a result of industrial exposure to the organic solvents *n*-hexane and methyl-*n*-butyl ketone. The former is also encountered as a consequence of solvent abuse; *n*-hexane, which has an intoxicant action, has been used as a solvent in certain glues. Other industrial agents causing neuropathy are ethylene oxide and methyl bromide. Trichlorethylene (or an impurity) has caused trigeminal neuropathy.

Iatrogenic

Cisplatin

This platinum derivative (*cis*-diaminedichloroplatinum) is used in the treatment of malignancy, including carcinoma of the ovary. A predominantly sensory neuropathy that recovers poorly may develop after the administration of several courses. Ototoxicity is more frequent, causing high-tone deafness and tinnitus.

Isoniazid

A mixed motor and sensory neuropathy may be produced by isoniazid and is more likely to occur in individuals who acetylate the drug slowly. The neuropathy is related to an interference with pyridoxine metabolism. Axonal degeneration occurs in the peripheral nerves. The neuropathy recovers slowly on cessation of administration of the drug and may be prevented by giving pyridoxine, which does not interfere with the antituberculous action of the isoniazid.

Nitrofurantoin

Excessively high blood levels of this preparation, as may occur in patients with reduced renal function, can cause a mixed motor and sensory neuropathy.

Vincristine

A neuropathy will occur in all subjects if sufficient amounts of this cytotoxic agent are administered. Mild sensory symptoms and the loss of tendon reflexes may have to be accepted if a satisfactory therapeutic effect of the drug is to be achieved. If the neuropathy advances, bilateral weakness of the extensors of the wrist and fingers develops, followed by more widespread weakness. The neuropathy improves satisfactorily if the drug is withdrawn or if the dosage is reduced.

Other substances

Less important drugs that may give rise to neuropathy are adriamycin, amiodarone, dapsone, disulfiram, gold, metronidazole, misonidazole, nitrous oxide (with a myelopathy), suramin, and zimeldine. A mild sensory neuropathy may develop after prolonged administration of phenytoin, and neuropathy was one of the complications produced by thalidomide. Pyridoxine, if taken in large doses, as 'megavitamin therapy', causes a sensory neuropathy.

Deficiency neuropathies

Beri beri neuropathy (see also Section 10)

This disorder is predominantly encountered in populations subsisting on diets composed largely of polished rice, but a similar neuropathy may be observed in other malnourished communities. Thiamine deficiency is probably involved, but a deficiency of other vitamins of the B group may also be implicated. A distal motor and sensory neuropathy develops which is frequently accompanied by spontaneous aching pain in the extremities, cutaneous hyperaesthesia, and tenderness of the soles of the feet and calves. Involvement of the recurrent laryngeal nerves may lead to hoarseness of the voice. The neuropathy may be associated with a cardiomyopathy ('wet beri beri'). Thiamine deficiency is established by the finding of reduced activity of erythrocyte transketolase. This enzyme requires thiamine as a cofactor.

Distal axonal degeneration occurs in the peripheral nerves and slow recovery ensues with vitamin replacement.

Strachan's syndrome

Strachan's syndrome, originally described in Jamaica but also observed in other parts of the world under conditions of nutritional deprivation, is characterized by the combination of a painful sensory neuropathy with amblyopia and at times deafness, in association with an orogenital dermatitis. It is assumed to be due to B vitamin deficiency, but the precise deficit has not been identified. It improves with B vitamin supplementation.

Alcoholic neuropathy

This always occurs on a background of nutritional deficiency. The dietary intake of the alcoholic is high in carbohydrates and low in vitamins. Moreover, alcoholics are known to have a reduced capacity to absorb thiamine. A direct toxic effect of alcohol on peripheral nerves may also be involved. The clinical features of alcoholic neuropathy are similar to those of beri beri. Other deficiency states may coexist, such as the Wernicke–Korsakoff syndrome. Improvement may take place with vitamin replacement and reduced alcohol intake, but it is beset with the usual difficulties met in treating alcoholic patients.

Pyridoxine deficiency

Attention has already been drawn to the fact that isoniazid neuropathy is related to an interference with pyridoxine metabolism. Pyridoxine deficiency may contribute to the neuropathy that occurs in nutritional deficiency states, and possibly accounts for the mild neuropathy of pellagra.

Pantothenic acid deficiency

Experimental deficiency of pantothenic acid in human volunteers is known to give rise to a sensory neuropathy, and the administration of pantothenic acid has been reported to alleviate the 'burning feet' syndrome which sometimes develops in deficiency states.

Vitamin B_{12} deficiency

Vitamin B_{12} deficiency, from whatever cause, may be responsible for the development of a distal sensory neuropathy, with 'glove and stocking' sensory loss and paraesthesiae, and areflexia, either in isolation or in association with a myelopathy or other central nervous system manifestations. Haematological changes are not always present. The peripheral neuropathy improves more satisfactorily with treatment than the central disturbances. This condition is considered in detail in Chapter 24.3.9.

A peripheral neuropathy is one component of Nigerian ataxic neuropathy, in which the other features are posterior column degeneration, sensorineural deafness, and optic atrophy. It has been suggested that an interference with vitamin B_{12} metabolism by cyanide derived from cassava in the diet, combined with nutritional deficiency, may be responsible.

Chronic severe vitamin E deficiency has recently been established as a cause for peripheral neuropathy in combination with a spinocerebellar degeneration. This may occur in abetalipoproteinaemia, and isolated vitamin E deficiency, both of autosomal recessive inheritance, and in congenital biliary atresia, cystic fibrosis, and occasional adults with chronic intestinal malabsorption.

Inflammatory and post-infective neuropathies

Leprous neuropathy

Peripheral nerve involvement in leprosy is considered in Chapter 7.11.24.

Guillain–Barré syndrome (acute idiopathic inflammatory polyneuropathy)

Guillain–Barré syndrome is characterized by a polyneuropathy that develops over the course of a few days up to maximum of 4 weeks. Cases that progress for up to 8 weeks (subacute Guillain–Barré syndrome) are probably distinct. An identifiable infection may precede the onset of the neuropathy by 1 to 3 weeks. This is commonly an upper respiratory tract infection or an infection with an enterovirus, Epstein–Barr virus, or mycoplasma. More recently *Campylobacter jejuni* has been recognized as an important cause, as has human immunodeficiency virus (**HIV**) infection. Other cases may follow surgical operations. In approximately 40 per cent of cases no antecedent event is identifiable.

The neuropathy may be ushered in by severe lumbar or interscapular pain. Motor involvement usually predominates over sensory loss and may be of a proximal, distal, or generalized distribution, and in severe cases affects the respiratory musculature. Distal paraesthesiae in the limbs are common and, if sensory loss occurs, it tends to affect tactile, vibratory, and postural sensibility. The cranial nerves may be affected, in particular the facial nerves, but bulbar involvement also occurs sometimes. A complete 'locked-in' state may develop. Autonomic disturbances may be associated, including bladder atony, ileus, hypertension (possibly the result of denervation of the carotid sinus), and orthostatic hypotension. Associated central nervous system involvement is occasionally encountered, particularly after infectious mononucleosis, and such cases are sometimes excluded from the Guillain–Barré syndrome as such. Papilloedema sometimes develops, possibly related to impaired resorption of cerebrospinal fluid because of the elevated protein content. Further variants are a combination of an external ophthalmoplegia, ataxia, and tendon areflexia (Miller Fisher syndrome), as possibly are instances of acute sensory neuropathy or pandysautonomia.

Nerve conduction studies reveal evidence of demyelination in most cases, but at times the findings indicate an axonopathy ('axonal' Guillain–Barré syndrome), as is seen in the acute motor axonal neuropathy or motor and sensory axonal neuropathy that occurs as an annual epidemic in children in northern China. Cerebrospinal fluid protein is usually raised, often to a substantial degree, but it may be normal, particularly in the early stages. The cell content is usually normal, but there may be a mild lymphocytic pleocytosis; this is more likely to occur in cases related to HIV infection or infectious mononucleosis. The Miller Fisher syndrome is frequently associated with circulating anti-GQ1b antibodies. Histologically, the abnormalities are maximal in the spinal roots but also occur diffusely throughout the peripheral nerves. In the demyelinating form, focal perivascular accumulations of inflammatory cells are associated with segmental demyelination of the nerve fibres and relative preservation of axonal continuity. Recovery occurs by remyelination. The disease probably represents a cell-mediated hypersensitivity reaction in which myelin is stripped off the axons by mononuclear cells. Whether antibody-mediated demyelination is also involved is not yet established. Severe axonal loss may be a 'bystander effect' or represent direct axonal damage in cases of axonal Guillain–Barré syndrome.

Most cases of Guillain–Barré syndrome recover satisfactorily within weeks or months. Severely affected patients, particularly those that require assisted respiration and in whom extensive axonal degeneration occurs, recover slowly and often show residual muscle weakness. Occasional patients have recurrences, which are sometimes multiple.

Although widely employed in the past, controlled trials of treatment with corticosteroids have shown no beneficial effects. Plasma exchange and high-dose intravenous human immunoglobulin have both been shown to improve the rate of recovery if given before the nadir of the disease. Because of significant morbidity, particularly with plasma exchange, and cost, these forms of treatment are best reserved for more severe cases. Severely affected patients may require extensive support in an intensive care unit because of respiratory failure and autonomic dysfunction.

Chronic inflammatory demyelinating polyneuropathy

Instances of peripheral neuropathy occur that resemble Guillain–Barré syndrome in that the neurological involvement is predominantly motor and the cerebrospinal fluid protein level is elevated, but which pursue either a chronic relapsing or chronic progressive course. They are also associated with widespread demyelination in the spinal roots and peripheral nerves and with inflammatory infiltrates. Nerve conduction velocity is usually markedly reduced and conduction block may be evident. Cases with a purely sensory ataxic neuropathy also occur, as do others with localized involvement, most often of the brachial plexus. Both the generalized and focal cases may respond to treatment with corticosteroids, plasma exchange, or high-dose intravenous human immunoglobulin. Cytotoxic drugs may be required in refractory cases. The response is less satisfactory in the chronic progressive cases.

Patients have recently been identified with a chronic multifocal motor neuropathy with persistent conduction block associated with GM1 ganglioside antibodies. They probably represent a variant of chronic inflammatory demyelinating polyneuropathy. They may respond to immunosuppressive therapy or plasma exchange.

Lyme borreliosis

Lyme borreliosis is a multisystem disease caused by a tick-borne spirochaete *Borrelia burgdorferi* (see Chapter 7.11.30). The peripheral nervous system is frequently affected both during the phase of early disseminated infection and during the late stage. Cranial neuropathies or an acute or subacute radiculoneuritis characterize involvement in the early stages, and a mild, predominantly distal, neuropathy characterizes the late stage. Nerve biopsies show perivasculitis and nerve fibre

degeneration, but spirochaetes are not identifiable. Laboratory diagnosis is based on the detection of specific antibodies to *B. burgdorferi* but seronegative cases occur, as may false positive reactions. Treatment, which is with doxycycline and amoxicillin, may therefore have to be given on clinical suspicion of the disease.

Human immunodeficiency virus infection (see Chapter 24.14.4)

A variety of neuropathies may be related to HIV-1 infection, particular types tending to occur in different phases of the disease. Characteristically, Guillain–Barré syndrome or chronic inflammatory demyelinating polyneuropathy occur at the time of seroconversion, when the patient is otherwise well, and a multifocal vasculitic neuropathy occurs in the early symptomatic stage. A distal, often predominantly sensory and painful neuropathy occurs mainly in the later AIDS phase, and an aggressive lumbosacral polyradiculoneuropathy from cytomegalovirus infection is encountered in advanced cases. Neuropathy may also occur in patients with human T-cell leukaemia virus (HTLV-I) infection. The treatment of HIV infection is discussed in Chapter 7.10.21.

Sarcoid neuropathy

Sarcoidosis (see Section 17) may give rise to a multifocal neuropathy with a particular tendency to involve the facial nerves, or to a generalized neuropathy. The neuropathy may be restricted to the cranial nerves (polyneuritis cranialis). Evidence of involvement of other systems is not always present and sarcoid tissue may or may not be detectable on biopsy.

Diphtheritic neuropathy

The neuropathy of diphtheria (Chapter 7.11.1) is caused by the exotoxin which produces segmental demyelination by interfering with Schwann cell function, probably by affecting protein synthesis. The nerves are not invaded by the bacteria.

Palatal weakness tends to develop after 2 to 3 weeks following pharyngeal diphtheria, and local muscle paralysis after a similar interval following cutaneous diphtheria. Paralysis of accommodation and sometimes of the external ocular muscles appears after an interval of 4 to 5 weeks. A generalized predominantly motor neuropathy of distal distribution may develop after 5 to 7 weeks. In severe cases the respiratory muscles are affected, but if death occurs it is usually as a result of an associated myocarditis.

Neuropathy in autoimmune connective tissue disorders

Peripheral nerve involvement may be encountered in a wide range of the 'collagen-vascular' disorders. Polyarteritis nodosa characteristically gives rise to a multifocal neuropathy, often with considerable pain. Wegener's granulomatosis may similarly be associated with a florid neuropathy and, in both instances, the peripheral nerve damage is related to necrotizing angiitis of the vasa nervorum. Such changes may also occur in rheumatoid arthritis in association with a florid multifocal neuropathy; at other times, a less aggressive neuropathy is observed, either in the form of a distal sensory neuropathy or one restricted to the digital nerves. Entrapment neuropathies also occur in rheumatoid arthritis, for example median nerve compression in the carpal tunnel, related to inflammatory changes in articular synovial tissues or tendon sheaths.

An ataxic sensory neuropathy related to a sensory ganglionitis can complicate the Sjögren sicca syndrome. A clinical constellation that combines a distal ensory neuropathy with a trigeminal sensory neuropathy, and myotonic pupils with the sicca syndrome is particularly characteristic. A multifocal neuropathy can also be seen in patients with Sjögren's syndrome.

The neuropathy of polyarteritis nodosa or rheumatoid arthritis may respond to corticosteroids or cyclophosphamide. The neuropathy of Sjögren's sicca syndrome is largely refractory to treatment.

Neoplastic and paraneoplastic neuropathy

Peripheral neuropathy may develop as a non-metastatic complication of carcinoma, most often bronchial or gastric, or lymphoreticular proliferative disorders. The precise mechanism of production of the neuropathy is uncertain. The neuropathy may antedate the discovery of the carcinoma by as much as 2 or 3 years. In relation to carcinoma of the bronchus, the neuropathy may be purely sensory, either subacute or chronic, often with troublesome distal dysaesthesiae, or mixed sensory and motor. The sensory neuropathy is associated with circulating antineuronal Purkinje cell and anti-Hu antibodies. Guillain–Barré syndrome may be encountered in Hodgkin's disease and in chronic lymphocytic leukaemia, and a subacute, mainly motor neuropathy in relation to lymphoma. Non-metastatic carcinomatous neuropathies may regress following removal of the underlying tumour, or may remain unaffected.

Direct invasion of cranial nerves or spinal roots may occur in cases of malignant infiltration of the meninges and of the cervical and lumbosacral plexuses from local malignancies. Infiltration of peripheral nerve trunks is seen most commonly from malignant lymphomas.

Paraproteinaemic neuropathy

A sensory or sensorimotor polyneuropathy related to benign monoclonal paraproteins (monoclonal gammopathies of undetermined significance) has emerged n recent years as an important cause of late onset neuropathy. The neuropathy is usually demyelinating, and in some with features similar or identical to chronic inflammatory demyelinating polyneuropathy. A postural upper limb tremor is often a prominent feature. The paraprotein is most commonly immunoglobulin M, less frequently immunoglobulins G or A. The immunoglobulin M paraproteins can be demonstrated on surviving myelin sheaths in nerve biopsies, where they are probably acting as demyelinating antibodies. Neuropathies associated with immunoglobulin G or A paraproteins may respond to corticosteroids, intravenous immunoglobulin, plasma exchange or immunosuppressive drugs; the response in immunoglobulin M paraproteinaemic neuropathy is disappointing. Although a distal sensorimotor and often painful axonal neuropathy may be associated with myeloma, a demyelinating neuropathy accompanied by a dermatoendocrine syndrome may be encountered, referred to as the Crow–Fukase or POEMS syndrome (**P**olyneuropathy, **O**rganomegaly, o**E**dema, **M** protein, **S**kin changes). A mixed sensorimotor neuropathy occurs which may be associated with papilloedema. The skin changes consist of excessive pigmentation and hypertrichosis. Peripheral oedema develops. Partial syndromes may occur in which all features of the syndrome are not present. The disorder is most often related to osteosclerotic myeloma.

Multifocal, predominantly lower limb, neuropathy may be caused by single or mixed cryoglobulins in myeloma, lymphoma, systemic lupus erythematosus, rheumatoid arthritis, or Waldenström's macroglobulinaemia. They may be the result of vasculitis produced by the deposition of immune complexes in the walls of the vasa nervorum, or to intravascular precipitation of cryoglobulin. These neuropathies occasionally respond to treatment either with immunosuppressive or cytotoxic drugs, or to repeated plasma exchange.

Genetic neuropathies

Porphyria (see also Section 11)

A predominantly motor neuropathy may complicate acute attacks in the autosomal dominant disorders of acute intermittent and variegate porphyria and hereditary coproporphyria, and in the recessively inherited †-aminolaevulinic acid dehydratase deficiency. It tends to affect the proximal muscles to a greater extent. There may be associated sensory loss which, although sometimes distal in distribution, can affect the trunk and the proximal portions of the limbs. The tendon reflexes are lost, with occasional paradoxical sparing of the ankle jerks. Accompanying autonomic features include abdominal pain and vomiting, tachycardia and hypertension; mental confusion, psychotic behaviour, and epilepsy.

The explanation of the neurological damage has not been established. Axonal degeneration occurs in the peripheral nerves so that recovery is slow and often incomplete.

Attacks may be provoked by a variety of drugs, including barbiturates, sulphonamides, and the contraceptive pill, and by alcohol, probably by enzyme induction in the liver (see Section 14).

Treatment with oral or intravenous glucose, or by infusions of laevulose or haematin, has been shown to reduce the urinary excretion of porphyrin precursors, but a

beneficial effect on the neurological disturbances has not been established.

Familial amyloid polyneuropathy

A number of inherited amyloid neuropathies have been recognized, the commonest being those related to point mutations in the gene for transthyretin, formerly known as prealbumin, which is on chromosome 18. The commonest is the Portuguese type where there is a substitution of valine for methionine in the transthyretin molecule. The neuropathy begins with the involvement of small nerve fibres, leading to a distal loss of pain and temperature sensation and autonomic failure. Spontaneous pain is often a feature and a mutilating acropathy frequently develops. The onset is commonly in the fourth or fifth decades and the disorder is slowly progressive, leading to death within about 10 years. Transthyretin is produced mainly in the liver and liver transplantation may halt the progression of the disease. In other types of hereditary amyloid neuropathy with differing clinical features, the amyloid is derived from a variant apolipoprotein A1 (Iowa form) or plasma gelsolin (Finnish form).

Hereditary motor and sensory neuropathy types I and II and X-linked (Charcot–Marie–Tooth disease, peroneal muscular atrophy); hereditary neuropathy with liability to pressure palsies

Hereditary motor and sensory neuropathies type I and II (or Charcot–Marie–Tooth 1, Charcot–Marie–Tooth 2) usually present during childhood or adolescence with difficulty in walking or because of foot deformity. The deformity is most commonly pes cavus associated with clawing of the toes and sometimes with an equinovarus position of the foot. Muscle weakness tends to affect the lower leg muscles and may give rise to bilateral foot drop with a 'steppage' gait. The muscle wasting is often restricted to below the knees, producing a 'stork leg' appearance (Fig. 3). Weakness and wasting of the small hand muscles may appear later. The tendon reflexes become depressed or lost, and there is a variable degree of distal sensory loss. This is the Charcot–Marie–Tooth phenotype. Progress of the disease is slow and cases with little disability or which are asymptomatic are common.

Fig. 3 Patient with type I hereditary motor and sensory neuropathy (Charcot–Marie–Tooth disease) showing symmetrical distal lower limb muscle wasting.

In the commoner type I families, there is a diffuse demyelinating neuropathy. The onset is most frequently in the first decade. Foot deformity and scoliosis occur more often than in the type II disease. Sensory loss and ataxia tend to be greater and generalized tendon areflexia is usual. Weakness in the hands appears earlier. The peripheral nerves may be thickened. Cases with ataxia and upper limb tremor are sometimes referred to as the Roussy–Lévy syndrome. The onset in the type II form, which is an axonal neuropathy, is most often in the second decade but it may be delayed until middle or even late adult life. Inheritance in both types I and II hereditary motor and sensory neuropathy is usually autosomal dominant. The disorder in type I hereditary motor and sensory neuropathy is most often caused by a segmental duplication on chromosome 17p11.2 (hereditary motor and sensory neuropathy HMSN Ia) which includes the gene for peripheral myelin protein 22 (*PMP22*). Other cases are related to mutations in the gene for myelin protein zero (hereditary motor and sensory neuropathy Ib). X-linked hereditary motor and sensory neuropathy is due to mutations in the gene for connexin 32. The clinical features resemble hereditary motor and sensory neuropathy I but female carriers are asymptomatic or only mildly affected. Several separate loci have so far been identified for hereditary motor and sensory neuropathy II but the gene products are not known.

Nerve conduction velocity is severely reduced in type I cases, moderately reduced in the X-linked form, and either normal or only slightly reduced in type II.

Affected individuals may be helped by the use of orthotic appliances and sometimes by surgical correction of foot deformity or tendon transfer.

Hereditary neuropathy with liability to pressure palsies is an autosomal dominant disorder in which affected individuals develop recurrent focal peripheral nerve or brachial plexus lesions produced by compression or stretch injury. It has been shown usually to be due to a segmental deletion on chromosome 17p11.2, i.e. it is the reciprocal of hereditary motor and sensory neuropathy Ia. Nerve fibres show focal regions of myelin thickening (tomacula).

Hereditary motor and sensory neuropathy type III (Dejerine–Sottas disease and congenital hypomyelination)

The Dejerine–Sottas phenotype consists of a severe slowly progressive mixed motor and sensory polyneuropathy with an onset in childhood. There is hypomyelination and extensive demyelination in the peripheral nerves, and there may be accompanying hypertrophic changes (concentric Schwann cell proliferation). Striking enlargement of the peripheral nerve trunks may be evident. These cases are most often due to de novo *PMP22* or P zero (*P0*) mutations. Some cases result from mutations in the early growth response gene 2 (*EGR2*). A severe congenital hypomyelination neuropathy can also result from *P0* or *EGR2* mutations.

Refsum's disease

This is a rare disorder inherited as an autosomal recessive trait that gives rise to a mixed motor and sensory polyneuropathy accompanied by a variety of other clinical features, including ataxia, anosmia, pigmentary retinal degeneration, pupillary abnormalities, deafness, cardiomyopathy, and ichthyosis. The presentation is usually during adolescence or early adult life and the course may be steadily progressive or relapsing. The peripheral nerves become thickened and display hypertrophic changes. Nerve conduction velocity is usually severely reduced.

The disorder is due to an inability to metabolize phytanic acid, a long-chain fatty acid, which accumulates in the blood and tissues. Phytanic acid is largely of dietary origin, and clinical improvement may be achieved with diets low in phytanic acid. Plasma exchange is effective for acute episodes of deterioration.

Hereditary sensory and autonomic neuropathies

Predominantly sensory neuropathies may occur with either an autosomal dominant or recessive inheritance. The symptoms in the latter instance are usually present from birth; in the former they generally develop during the second or third decades. In both, the sensory loss often leads to a mutilating acropathy, with neuropathic joint degeneration and chronic cutaneous ulceration, particularly of the feet (Fig. 4). Autonomic features are dominant in the recessive disorder of familial dysautonomia (Riley–Day syndrome). A further rare recessive neuropathy combines congenital insensitivity to pain and anhidrosis. Most cases of 'congenital insensitivity to pain' are probably examples of small-fibre neuropathies.

Fig. 4 Chronic foot ulceration and deformity in a case of hereditary sensory neuropathy.

Familial dysautonomia

Otherwise known as the Riley–Day syndrome, this recessively inherited disorder is encountered most often in Jewish populations. There is an aplasia of peripheral autonomic neurones that leads to a variety of symptoms, including absence of tears, unexplained pyrexia, cutaneous blotching, and episodic sweating attacks. These symptoms are present at birth and are accompanied by congenital insensitivity to pain related to an associated sensory neuropathy. In early infancy there is usually difficulty in feeding because of poor sucking, and repeated episodes of aspiration pneumonia. Later, stunted growth and often kyphoscoliosis become evident. The disorder has been mapped to chromosome 9.

Other hereditary neuropathies (see also Section 11)

Peripheral nerve involvement occurs in metachromatic and globoid cell leucodystrophy, adrenomyeloneuropathy, Fabry's disease, hereditary high-density lipoprotein deficiency (Tangier disease), hereditary abetalipoproteinaemia, and cholestanolosis. Giant axonal neuropathy is a rare autosomal recessive disorder with an onset in childhood, characterized by segmental axonal enlargements containing accumulations of neurofilaments. Affected children usually have abnormally curly hair and may have enlarged tangerine-coloured tonsils.

Cryptogenic neuropathy

Despite extensive investigation, the cause of a substantial number of peripheral neuropathies remains unknown. This applies in particular to examples of chronic progressive axonopathies, some of which may be instances of late onset type II hereditary motor and sensory neuropathy. A careful family history in such cases may reveal evidence of undetected neuropathy in relatives. Prolonged follow-up in other cases may disclose underlying malignancy.

Further reading

Asbury AK, Thomas PK (1995). *Peripheral nerve disorders. A practical approach*, 2nd edn. Butterworths, London.

Birch R, Bonney C, Wynn Parry CB (1998). *Surgical disorders of the peripheral nerves.* Churchill Livingstone, Edinburgh.

Dawson DM, Hallett M, Millender LH (1990). *Entrapment neuropathies*, 2nd edn. Little, Brown, Boston.

Dyck PJ, Thomas PK (1999). *Diabetic neuropathy*, 2nd edn. WB Saunders, Philadelphia.

Dyck PJ *et al.* (1993). *Peripheral neuropathy*, 3rd edn. WB Saunders, Philadelphia.

Harding AE (1995). From the syndrome of Charcot, Marie and Tooth to disorders of peripheral myelin proteins. *Brain* **118**, 809–18.

Hughes RAC (1990). *Guillain–Barré syndrome*. Springer, London.

Kimura J (1980). *Electrodiagnosis of diseases of nerve and muscle. Principles and practice*, 2nd edn. FA Davis, Philadelphia.

Stewart JD (2000). *Focal peripheral neuropathies*, 3rd edn. Lippincott, Williams and Wilkins, Philadelphia.

24.20 Neurological complications of systemic autoimmune and inflammatory diseases

Neil Scolding

Introduction

The range and breadth of diseases of the nervous system caused by immunological, infective, or inflammatory disturbances is very large. It includes 'primary' or idiopathic neuroimmune disorders, which may affect any part of the neuraxis (for example multiple sclerosis and Guillain–Barré syndrome) and which are very familiar to neurologists. However, 'secondary' disorders, where the neurological disturbance reflects involvement of the nervous system in a systemic inflammatory disease, are often no less common than idiopathic immune disorders, but most neurologists are rather less familiar and possibly less comfortable with them.

Systemic lupus erythematosus

Systemic lupus erythematosus, like many autoimmune diseases, occurs more in women than men—perhaps 20 times more commonly. Black people are more commonly affected than white. The neurologist should not (but usually does) omit direct enquiry and focused systemic examination to exclude fever and general malaise, skin changes—classically, the malar butterfly rash and/or photosensitivity—and large and small joint arthritis. Glomerulonephritis, pleurisy and pneumonitis, pericarditis and (so-called) Libmann–Sachs endocarditis, and haematological disorders—anaemia, thrombocytopenia, leucocytopenia, and the generation of circulating anticoagulants—also occur. Other laboratory abnormalities include the presence of a variety of autoantibodies, including antinuclear antibodies and anti-native DNA antibodies. The diagnosis—particularly for research and therapeutic trial purposes—is now commonly based on the widely accepted revised diagnostic criteria suggested by the American College of Rheumatology. The presence of any four (or more) of the listed features, 'serially or simultaneously, during any interval of observation' are sufficient for the diagnosis, with an estimated specificity and sensitivity of 96 per cent.

Neurological complications

Neurological involvement in systemic lupus erythematosus is seen in perhaps 50 per cent of cases; neurological presentation, in perhaps 3 per cent of cases. Central nervous system disease is much more frequent than neuromuscular involvement, and is a poor prognostic sign, reducing the overall survival figures, and representing the third commonest cause of death (after renal involvement and iatrogenic causes).

An enormous variety of central nervous system disease complications can occur, reflecting two broad pathogenetic mechanisms—thromboembolic (triggered either by changes in endothelial surfaces or by coagulation disturbances, including lupus anticoagulant activity) and more direct autoimmune events affecting the target tissue—neurones or glia—in which soluble and cellular mediators are implicated.

Headache (including that associated with dural sinus thrombosis), acute or subacute encephalopathy, fits, myelitis, strokes and movement disorders (especially chorea), ataxia and brainstem abnormalities, and cranial and peripheral neuropathies are all seen in the context of systemic lupus erythematosus. Psychiatric and cognitive disturbances have also long been associated with lupus.

Stroke, the lupus anticoagulant, and the primary phospholipid syndrome

The thrombotic tendency in patients with systemic lupus erythematosus and lupus anticoagulant manifests itself principally in the form of stroke and recurrent spontaneous abortion. Intra-abdominal and deep venous thrombosis, and peripheral arterial thrombosis are also seen. Thrombocytopenia is a key additional feature. Importantly, Hughes also showed that a similar clinical picture was associated with the presence of anticardiolipin antibodies (ACA) and/or lupus anticoagulant in patients without serological or clinical evidence of lupus, and introduced the term 'antiphospholipid syndrome'.

ACAs represent an independent risk factor for stroke. Central nervous system thrombosis in patients with primary or secondary antiphospholipid syndrome takes the form of completed arterial stroke, repeated transient ischaemic attacks, multi-infarct dementia, and cerebral venous sinus thrombosis. Vascular visual problems, including amaurosis fugax and ischaemic retinopathy, also occur. Chorea too is associated with antiphospholipid antibodies; the putative link with migraine may be factitious.

A severe acute ischaemic encephalopathy is also described, with confusion, obtundation, and a hyperreflexic quadriparesis (usually asymmetrical), with or without systemic disturbances (dermatological and renal). Cerebrospinal fluid examination may show only a raised protein; a fatal outcome is common. The disorder may represent a focal variant of the recently described 'catastrophic antiphospholipid syndrome', in which there is severe multi-organ failure and a mortality of the order of 60 per cent.

There are both clinical and pathological similarities between microangiopathic complications of lupus and the syndrome of thrombotic thrombocytopenic purpura. In this latter uncommon disorder, multi-organ involvement is also seen, with hepatic and renal disease, and fever, together with thrombocytopenia and an associated purpuric rash and other haemorrhagic complications. Neurologically, an encephalopathy occurs, often with fits, with or without focal deficits. Pathologically, there are widespread microangiopathic changes in the brain and systemically. Plasma exchange is commonly recommended.

Diagnosis of central nervous system lupus

Cerebrospinal fluid examination may reveal a raised protein level and a neutrophil or lymphocyte pleocytosis. It is clearly vital in such cases to exclude infectious complications of immune suppressants or steroids, now a major cause of death in patients with lupus. Serological tests are discussed elsewhere. MRI changes are common, though neither invariable nor specific. Cerebrospinal fluid oligoclonal band analysis is positive in up to 50 per cent of patients with central nervous system lupus and, interestingly, these changes can resolve with successful immunotherapy. A skin biopsy can be extremely helpful in suspected lupus (see Chapter 18.10.2).

The management of neuropsychiatric lupus

Symptomatic therapies are important in patients with encephalopathies, epilepsy, and/or psychiatric ailments. Disease-modifying therapeutic efforts fall into two categories depending on the presumed underlying mechanisms: (i) stroke prevention in cerebral ischaemia, particularly that associated with ACA, probably best achieved with moderate- to high-dose warfarin, and (ii) immunotherapy of 'other' central nervous system complications. Here, intravenous methyl prednisolone followed by oral steroids is the mainstay of treatment. Cyclophosphamide may be given for severe or steroid-resistant disease, with azathioprine to maintain remission and spare steroids. Plasmapheresis synchronized with cyclophosphamide, and intravenous immunoglobulin, may prove useful.

Rheumatoid arthritis

An inflammatory peripheral neuropathy occurs in approximately 30 per cent of seropositive rheumatoid cases. A relatively benign mononeuritis is typical, but a more severe and aggressive axonal polyneuropathy or mononeuritis multiplex may be seen when rheumatoid arthritis is accompanied by a vasculitis. More common than either are entrapment neuropathies of conventional distribution, precipitated by synovial swelling. Pannus formation and cervical spine subluxation with resulting cord compression represent the commonest cause of central nervous system involvement. More rarely, rheumatoid vasculitis, or deposition of rheumatoid nodules, may involve the central nervous system; the former warrants treatment with cyclophosphamide and steroids.

Sjögren's syndrome

Sjögren's syndrome characteristically comprises a triad of: (i) keratoconjunctivitis sicca, and (ii) xerostomia, occurring in approximately 50 per cent of cases (iii) in the context of another connective tissue, usually rheumatoid arthritis. Speckled antinuclear antibodies of the anti-Ro (SS-A) or anti-La (SS-B) type are present in up to 75 to 80 per cent of patients. Conventionally, the principal neurological manifestations have been held to be peripheral, with descriptions of both a mainly sensory neuropathy and of myositis. Trigeminal sensory neuropathy is also classically described.

More recently, attention has been drawn to various central nervous system complications of the disorder, with seizures, focal stroke-like or brainstem neurological deficits, and encephalopathy with or without an aseptic meningitis, often with raised cerebrospinal fluid pressure, protein, and white cell count, together with oligoclonal immunoglobulin bands. Psychiatric abnormalities may occur; spinal cord involvement may take the form of an acute transverse myelitis, a chronic myelopathy, or intraspinal haemorrhage. Occasionally, the features resemble those of multiple sclerosis (optic neuropathy is particularly associated) although most such patients have additional features of peripheral neuropathy or myositis.

Steroids may be insufficient for the treatment of patients with central nervous system complications of Sjögren's syndrome; more powerful immunosuppressants are probably more useful, although, as is so often the case, their value is yet to be proved objectively.

Systemic sclerosis

Systemic sclerosis results from the excessive deposition of collagen in the skin and other affected tissues. The cutaneous manifestation, scleroderma, may exist in isolation, but in multisystem disease, it is accompanied by Raynaud's phenomenon, calcinosis and atrophy of subcutaneous tissues, telangiectasia, and oesophageal strictures. Neurological complications are not common. Peripheral nervous system disease predominates, particularly painful trigeminal neuropathy; myopathy with an elevated creatine phosphokinase also occurs. A myelopathy may be associated. No treatment is of proven benefit.

Mixed connective tissue disease

In this disorder, features of scleroderma, polymyositis, and systemic lupus erythematosus coincide, and high levels of antibodies directed against extractable nuclear antigens—ribonucleoproteins—are found. Rheumatoid factor is also often present. In common with both systemic sclerosis and Sjögren's syndrome, trigeminal neuralgia and/or sensory neuropathy are described.

Seronegative arthritides

Ankylosing spondylitis

Neurological disease in the setting of ankylosing spondylitis usually reflects advanced bony disease; a cauda equina syndrome is well reported, unexplained, and difficult to treat.

Reiter's disease

The clinical triad of seronegative arthropathy, non-specific urethritis, and conjunctivitis, usually following venereal or dysenteric infection, constitutes Reiter's syndrome. As many as 25 per cent of patients are reported to have neurological features. Peripherally, radiculitis and polyneuritis occur; central nervous system disorders include aseptic meningoencephalitis, seizures, and psychiatric disturbances, particularly paranoid psychosis. Cranial neuropathies, pyramidal signs, and myelopathy are also reported. A recent report suggests that cyclosporin may be of value in severe Reiter's disease.

Psoriasis

Included as the third seronegative arthropathy, the neurology of psoriasis is not extensive. Cord compression from cervical psoriatic spondylosis is described, but reports of a complicating polyneuritis have not been substantiated.

Vasculitis

The vasculitides are a heterogeneous group of disorders which share certain pathological features, particularly intramural inflammation and necrotic changes within the walls of blood vessels. Their classification is complex, with subdivisions into: (i) idiopathic vasculitic disorders, for example giant cell arteritis and Wegener's granulomatosis; (ii) vasculitis secondary to collagen diseases, malignancy, viral infection, and so on; and (iii) vasculitis according to pathological features, largely vessel size (see Section 18.10). Nervous system involvement can occur in any of the systemic vasculitides. Additionally, isolated vasculitis of the central or peripheral nervous system is recognized, where little or no inflammation is apparent outside the nervous system—primary central (or peripheral) nervous system angiitis. In both primary and secondary vasculitis of the nervous system, neurological features arise from inflammation and necrosis of the vasculature, principally through infarction.

The clinical features of vasculitis of the nervous system

The picture of peripheral nerve vasculitis is relatively straightforward: a mixed sensory and motor neuropathy, usually rapidly progressive, and often painful. About 50

per cent of patients present with mononeuritis multiplex, the remainder with a more diffuse asymmetrical polyneuropathy or a distal symmetric neuropathy.

Central nervous system disease is infinitely more varied; focal or multifocal infarction, or diffuse ischaemia, affecting any part of the brain, explaining the protean manifestations, wide variation in disease activity, course, and severity, and the absence of a pathognomic or even a typical clinical picture. Thus, in primary and secondary intracranial vasculitis, the following are seen: headache, focal and generalized seizures, stroke-like episodes causing hemispheric or brainstem deficits, acute and subacute encephalopathies, progressive cognitive changes, behavioural disturbances, chorea, myoclonus and other movement disorders, and optic and other cranial neuropathies. The course is commonly acute or subacute, but monophasic, chronic progressive, and spontaneously relapsing–remitting presentations all occur. Despite this range, three broad clinical categories of presentation may be delineated: (i) phenotypically resembling atypical multiple sclerosis ('MS-plus'), with a relapsing–remitting course, and features such as optic neuropathy and brainstem episodes accompanied by other features less common in multiple sclerosis—seizures, severe and persisting headaches, encephalopathic episodes, or hemispheric stroke-like episodes; (ii) acute or subacute encephalopathy, with headache with an acute confusional state, progressing to drowsiness and coma; and (iii) intracranial mass lesion—with headache, drowsiness, focal signs, and (often) raised intracranial pressure. This grouping carries neither pathological nor therapeutic implications, but may help improve recognition of the condition. Systemic features, such as fever and night sweats, livedo reticulares, or oligoarthropathy, may be present (although often only revealed on direct enquiry) even in so-called isolated central nervous system vasculitis.

Diagnosis and management

The diagnosis of cerebral vasculitis involves the exclusion of alternative possibilities (Table 1), the confirmation of intracranial vasculitis, and pursuit of the causes of vasculitis.

Confirming cerebral vasculitis

No single simple investigation is universally useful in confirming cerebral vasculitis. Serological markers, including antineutrophil cytoplasmic antibodies (**ANCA**), are important. Spinal fluid examination is, like the erythrocyte sedimentation test, often abnormal, but lacks specificity, with changes in cell count and/or protein in 65 to 80 per cent of cases; oligoclonal immunoglobulin bands may be present. Magnetic resonance imaging may disclose ischaemic areas, periventricular white matter lesions, haemorrhagic lesions, and parenchymal or meningeal enhancing areas, but lacks both specificity and sensitivity. Contrast angiography may show segmental (often multifocal) narrowing and areas of localized dilatation or beading, often with areas of occlusion, rarely also with aneurysms. Again, these changes are not specific, and angiography carries a false-negative rate of perhaps 50 per cent, and a risk of 10 per cent for transient neurological deficit, and 1 per cent for permanent deficit. Nuclear imaging of labelled leucocytes and examination of the ocular vasculature may be useful.

Histopathological confirmation, taking a biopsy of an abnormal area of brain where possible, or 'blind' biopsy, incorporating meninges and non-dominant temporal white and grey matter, is important. Biopsy may reveal an underlying process not otherwise suspected with profound therapeutic implications, such as infective or neoplastic (principally lymphomatous) vasculopathies, but is not a trivial procedure, carrying a risk of serious morbidity estimated at 0.5 to 2 per cent—although immune-suppressant treatment may have a higher morbidity than biopsy, emphasizing the rationale behind this procedure.

Once a vasculitic process has been confirmed, the specific defining characteristics of the primary and secondary vasculitides must be painstakingly sought.

Neurological vasculitis complicating systemic vasculitides

Wegener's granulomatosis predominantly affects the upper and lower respiratory tracts—the nose (often with destructive cartilaginous change causing saddle nose deformity), sinuses, larynx, trachea, and lungs. Ocular involvement may occur; renal disease is usual. cANCA is positive, with proteinase-3 specificity, and the biopsy is characteristic, with granulomatous vasculitis. Microscopic polyangiitis is a multisystem small-vessel vasculitis which can involve almost any organ, or may rarely be confined to a single organ. Renal involvement is almost invariable. The diagnosis usually rests upon a combination of renal biopsy and ANCA serology (commonly pANCA). Classic polyarteritis nodosa is now recognized as an unusual disorder which may have some overlap and coexist with microscopic polyangiitis, but often occurs alone. Medium-sized vessels are affected in polyarteritis nodosa, and the kidneys are again commonly involved; renal angiography may reveal microaneurysms. pANCA testing is also often positive in Churg–Strauss syndrome, a multisystem disease characterized pathologically by a granulomatous necrotizing vasculitis, and clinically by prominent asthma with an eosinophilia. Small-vessel vasculitis commonly affects postcapillary venules. The skin is most commonly involved, usually with purpura or urticaria; the common presence of an allergic precipitant has led historically to the term hypersensitivity vasculitis often being used synonymously in this context; cutaneous leucocytoclastic vasculitis is the currently preferred epithet.

In all these disorders, peripheral nervous system involvement, with mononeuritis multiplex, is considerably more common than central nervous system disease, ranging from up to 70 per cent in patients with classic polyarteritis nodosa and microscopic polyangiitis, to around 30 per cent in patients with Wegener's disease. Central nervous system disease can, however, also occur. Direct effects of the granulomatous process—either by contiguous invasive spread or from remote metastatic granulomas—represent a mode of neurological involvement unique to Wegener's disease.

Neurological vasculitis complicating non-vasculitic systemic disorders

Although the clinical picture of cerebral vasculitis may closely be mimicked by systemic lupus erythematosus, a non-inflammatory vasculopathy is far more commonly responsible, but rare instances of vasculitis are described. In contrast, seropositive rheumatoid disease is a well-recognized precipitant of vasculitic mononeuritis multiplex and of central nervous system vasculitis. There are rare reports of central nervous system vasculitis in the context of systemic sclerosis, Sjögren's syndrome, and mixed connective tissue disease. The clinical features of cryoglobulinaemia represent the combined consequences of hyperviscosity and of immune complex deposition-triggered vasculitis, particularly in mixed cryoglobulinaemia, when associated with hepatitis C infection. Skin disease, with purpura progressing to necrotic ulceration, and renal and joint involvement are common. However, the diagnosis will only be made if blood is collected into a plain tube, immediately placed in water in a vacuum flask at 37°C, taken to the laboratory, and tested immediately. Peripheral neuropathy occurs in a quarter of patients with essential cryoglobulinaemia; central nervous system involvement is rare. Peripheral nerve disease, and/or histologically and angiographically evident vasculitis of the central nervous system, usually in the context of granulomatous meningitis, may occur in sarcoidosis.

Drug-induced vasculitis

The issue of vasculitis and drugs is complex. The most compelling evidence of a direct association relates to amphetamines, with clinical and histological evidence of multisystem necrotizing vasculitis. The majority of strokes occurring with cocaine abuse are associated with arterial spasm, platelet aggregation, severe abrupt hypertension, or migrainous phenomena, not vasculitis, although histologically proven cerebral vasculitis does occur.

Infections

At least three mechanisms may underlie microbe-related vascular damage—direct invasion, immune complex formation and deposition, and (in part related to the second) secondary cryoglobulinaemia. Although the association of hepatitis C infection with cryoglobulinaemia and small-vessel vasculitis has been stressed above, other infections, including hepatitis B, Epstein–Barr virus, cytomegalovirus, Lyme disease, syphilis, malaria, and coccidiomycosis have also been linked to mixed cryoglobulinaemia.

Primary invasion of the vascular wall by the infectious agent is, however, the commonest precipitant of infection-associated vasculitis. *Histoplasma*, *Coccidioide*s, and *Aspergillus* spp. are among the fungal causes of this picture, usually confined to immune-suppressed patients—although this includes diabetes mellitus. In HIV infection, cytomegalovirus and *Toxoplasma* may precipitate vasculitis, and syphilitic cerebral vasculitis has re-emerged in the context of HIV. More general bacterial causes of meningeal or cerebral infection—mycobacteria, pneumococci, and *Haemophilus influenzae*—may also trigger intracranial vasculitis.

Herpes zoster can precipitate cerebral vasculitis in approximately 0.5 per cent of cases, usually causing a monophasic illness, with hemiparesis contralateral to the eye disease. However, more generalized necrotizing and granulomatous vasculitis can also occur.

Malignancy, lymphomatoid granulomatosis, and malignant angioendothelioma

Leucocytoclastic vasculitis (often dermatological) may occur in association with a variety of cancers as a paraneoplastic phenomenon. Central nervous system

disease in the context of Hodgkin's disease with a pathological picture indistinguishable from conventional isolated central nervous system angiitis is reported. Lymphomatoid granulomatosis is a lymphomatous disorder centred on the vascular wall, with destructive change and secondary inflammatory infiltration lending the appearance of true vasculitis; the infiltrating neoplastic cell is of T-lymphocyte derivation. Cutaneous and pulmonary involvement are common, with nodular cavitating lung infiltrates, and neurological manifestations occur in 25 to 30 per cent of cases; they are the presenting feature in approximately 20 per cent. Neoplastic or malignant angioendotheliosis is also a rare, nosologically separate disorder, wherein the neoplastic process is intravascular (within the lumen) and the lymphomatous cells are B-cell derived and characteristically do not invade the vascular wall. The neurological features of each disorder are similar, largely representing those of cerebral vasculitic disease; in malignant angioendotheliomatosis, lung involvement is not the rule; characteristic skin manifestations occur.

Treatment of cerebral vasculitis

Prospective controlled randomized trials remain conspicuous by their absence, but retrospective analyses support the use of cyclophosphamide with steroids in vasculitis. In proven cerebral vasculitis—as in lupus—a 3- to 4-month induction regime might comprise high-dose intravenous then oral steroids, with oral or pulsed intravenous cyclophosphamide; this is followed by a maintenance regime of alternate-day steroids with azathioprine. In resistant disease, methotrexate (10 to 25 mg once weekly; again, with steroids) or intravenous immunoglobulin may be useful.

Two eponymous primary disorders may involve the central nervous system. Cogan's syndrome is an unusual disorder, mostly affecting young adults, characterized by recurrent episodes of interstitial keratitis and/or scleritis with vestibuloauditory symptoms, which may be complicated by central or peripheral nervous system or systemic vasculitis. In Eale's disease, an isolated retinal vasculitis occurs, causing visual loss; again, neurological complications are well described.

Giant cell arteritis

Giant cell arteritis, the most common large-vessel vasculitis, rarely affects individuals under 55 years of age. It affects women twice as commonly as men, with an overall prevalence of 100/10 000. Generally it presents with uni- or bilateral scalp pain, often severe, with exquisite tenderness. Additional symptoms include jaw claudication and polymyalgia rheumatica, with stiffness and aching of the shoulder girdle, worse in the mornings, and occasionally general malaise. The affected temporal artery (-ies) may be thickened and cord-like, often non-pulsatile, and tender. A raised erythrocyte sedimentation rate, often accompanied by a normochromic normocytic anaemia, must be followed by temporal artery biopsy—a specimen of several centimetres in length is recommended to help avoid false-negative results, which may occur because of the focal or multifocal nature of the disorder.

Histopathological examination of the vessel reveals changes of vasculitis, with an inflammatory infiltrate comprising mononuclear and giant cells; the latter phagocytose the elastic laminae, causing characteristic fragmentation. Immunoglobulin and complement deposits are apparent in lesions, but activated T cells predominate in the inflammatory infiltrate, suggesting cell-mediated immune damage. Vasculitic changes may still be apparent in biopsies taken 14 days or more after the commencement of steroids.

Neurological complications

Blindness occurs in approximately one-sixth of treated patients with temporal arteritis, as a consequence of anterior ischaemic optic neuropathy following vasculitic involvement of the posterior ciliary arteries and/or the ophthalmic artery, from which they are derived. A typical picture comprises (locally) painless loss of acuity, commonly severe, often with an altitudinal field defect. The fundal appearances may be normal, although swelling (usually mild) may be seen. Intracranial involvement is much less common; vertebral artery involvement is typical.

Treatment

Oral steroids should be used immediately there is serious suspicion of the disease, and in high doses (60 to 80 mg a day) in view of the risk of permanent blindness. The dose is generally reduced slowly (5 mg decrements weekly) after 4 to 7 days to a maintenance dose of perhaps 10 mg daily; thereafter, some would suggest continuing for 12 to 24 months before closely monitored phased withdrawal. Such a duration of steroid therapy, particularly in this elderly population, should direct attention to the treatable or preventable long-term consequences of corticosteroids, particularly osteoporosis, diabetes, cataract, and peptic ulceration.

Behçet's disease

Behçet's disease is a chronic relapsing multisystem inflammatory disorder whose clinical manifestations vary. The classic triad of recurrent uveitis with oral and genital aphthous ulceration remains clinically useful, although formal diagnostic criteria have now been proposed and generally adopted. Recurrent oral ulceration (at least three times in one 12-month period) is an absolute criterion; any two of (i) recurrent genital ulceration, (ii) uveitis (anterior or posterior) or retinal vasculitis, (iii) skin lesions, including erythema nodosum, or acneform nodules, pseudofolliculitis, or papulopustular lesions, or (iv) a positive pathergy test (read at 24 to 48 h) are also required to confirm the diagnosis.

Approximately one-third of patients develop neurological involvement, although this includes the very common occurrence of benign headache. Cerebral venous sinus thrombosis is one of the more specific serious complications; others include sterile meningoencephalitis, encephalopathy, brainstem syndromes, cranial neuropathies, and cortical sensory and motor deficits. Psychiatric and progressive cognitive manifestations are reported. Investigation may reveal an active cerebrospinal fluid, and oligoclonal IgA and IgM bands—but apparently not IgG—may be present. Evoked potentials may be diagnostically useful. MRI abnormalities are non-specific.

Treatment of Behçet's disease

Recent retrospective studies indicate an improved survival in patients with Behçet's disease of the central nervous system treated with steroids and immunosuppressants. The place of thalidomide in steroid-unresponsive Behçet's disease is currently under review; chlorambucil is often advocated. Monitoring treatment is difficult—neither the erythrocyte sedimentation rate nor C-reactive protein levels are useful; MRI might have such a role.

Sarcoidosis

Sarcoidosis is a multisystem granulomatous disease of unknown aetiology commonly affecting the lungs and, in approximately 5 per cent of patients, the nervous system. Optic and other cranial neuropathies (especially involving the facial nerve), often due to meningeal infiltration, and brainstem and spinal cord disease are the commoner manifestations. Cognitive and neuropsychiatric abnormalities are reported. Peripheral nerve and muscle involvement is also well described.

The diagnosis can be difficult. Serum and cerebrospinal fluid ACE levels may be elevated; the cerebrospinal fluid may reveal more general abnormalities of protein or cell count and oligoclonal bands may be present. Whole-body gallium scanning remains a useful indicator of systemic disease. Cranial MRI may show multiple white matter lesions or meningeal enhancement. The diagnosis is confirmed where possible by biopsy, either of cerebral or meningeal tissue, or of lung or conjunctiva where appropriate.

The mainstay of medical treatment in neurosarcoidosis is corticosteroids, although response rates as low as 29 per cent have been reported. Methotrexate, azathioprine, hydroxychloroquine, and cyclophosphamide have been used in steroid-resistant cases.

Organ-specific autoimmune disease

Ulcerative colitis and Crohn's disease

While differences in the frequency and type of, for example, dermatological or articular complications may occur between ulcerative colitis and Crohn's disease, the neurological complications, seen in around 5 per cent of patients, are similar. Three types of central nervous system disease have been associated: cerebrovascular accidents, mostly precipitated by the hypercoagulable state, and including venous or arterial thromboembolism, cerebral sinus venous thrombosis, and (more rarely and less explicably) vasculitis; epileptic seizures, focal and generalized, and not always in connection with dehydration or sepsis; and, in some reports, a slowly progressive myelopathy.

Peripheral neuropathy is seen in 0.5 to 1.0 per cent of cases and an acute Guillain–Barré syndrome is the commonest phenotype. Lastly, myopathy, sometimes of metabolic origin but mostly inflammatory, is also reported.

Whipple's disease

Whipple's disease is an uncommon multisystem disorder characterized by arthropathy, respiratory symptoms, anaemia, fever, erythema nodosum, and severe wasting in addition to steatorrhoea and abdominal distension, caused by *Tropheryma whippelii*. Approximately 10 per cent of patients have neurological involvement; 5 per cent present in this way. A wide variety of features is seen (Table 2).

Diagnosis and management

Up to 20 per cent of cases of cerebral Whipple's disease occur in the absence of gastrointestinal or indeed other systemic symptoms. CT and MR scanning may be normal, although the latter can also reveal non-specific abnormalities—multiple high-signal intensity areas on T_2-weighted images, or more striking enhanced mass lesions warranting biopsy. Similarly, the cerebrospinal fluid may be normal, or show an elevated protein and/or raised cell count; widely varying ratios of monocytes and polymorphonucleocytes are reported. One-third of cerebrospinal fluid samples may reveal pathognomic periodic acid–Schiff-stained bacilli; repeat spinal fluid examination increases this yield. Approximately 30 per cent of cases have a non-informative small bowel biopsy, although electron microscopy increases the sensitivity. Lymph node biopsy can also be useful. Polymerase chain reaction analysis of blood, lymph node, spinal fluid, small bowel tissue, or brain is increasingly used.

Whipple's disease usually responds to tetracyclines, penicillin, or more commonly nowadays, co-trimoxazole. Prompt treatment is vital in patients with neurological disease, which may (if untreated) run a profoundly aggressive and, not unusually, rapidly fatal course. Successful reversal of neurological deficits, including cognitive impairment, may follow antibiotic treatment.

Coeliac disease

Coeliac disease (non-tropical sprue) is an immunologically mediated disorder resulting from intolerance to dietary gluten; it causes weight loss with steatorrhoea and/or diarrhoea, and malabsorption. In common with other enteropathies, neurological sequelae of a predictable nature may complicate coeliac disease as a direct consequence of malabsorption. Central nervous system complications apparently unrelated to deficiency states may also occur in perhaps 10 per cent of patients. Rarely, vasculitis is responsible, but the cause of the most commonly described and distinctive central nervous system association, a progressive cerebellar or spinocerebellar degeneration, with eye movement disorders, myoclonus, and occasionally epilepsy, remains unresolved.

Major psychiatric complications and dementia are well described as a significant cause of morbidity, and have been studied in detail.

Thyroid disease

Hyperthyroidism and myxoedema both carry neurological complications generally considered direct consequences of abnormal thyroxine levels—anxiety, tremor, and occasionally chorea in thyrotoxicosis; and lethargy, myopathy, and dementia in hypothyroidism (see Chapter 12.4). By contrast, Grave's ophthalmoplegia and Hashimoto's encephalopathy are both thought to be immunologically driven.

In dysthyroid eye disease, the orbit and extraocular muscles are oedematous and infiltrated with inflammatory cells and glycosaminoglycans, resulting in proptosis and a restrictive ophthalmopathy. Up-gaze limitation is the commonest presenting sign. Vision is occasionally threatened by a complicating infiltrative or compressive optic neuropathy. Circulating thyroid-stimulating hormone (TSH) receptor-stimulating antibodies cross-reactive with orbital fibroblasts are found. Steroid treatment and radiotherapy are equally effective.

Hashimoto's encephalopathy exhibits a female:male ratio of up to 9:1. Most cases are clinically and biochemically euthyroid at presentation, and two modes of presentation occur. The relapsing–remitting variety causes stroke-like episodes, with or without mild cognitive impairment, focal or generalized seizures, and episodes of encephalopathy. The second group present with a more diffuse progressive disease, with dementia, psychotic features, seizures, and occasionally myoclonus, tremor, and/or ataxia; focal neurological deficits are uncommon.

Imaging by CT or even MR is often normal, as is angiography, though isotope brain scanning may show patchy uptake. Spinal fluid examination may reveal a raised protein level but typically a normal cell count. Very high titres of antithyroid antibodies are found, usually antimicrosomal. Most patients respond well to steroid treatment; some have received further immunosuppressive therapy, such as cyclophosphamide or azathioprine.

Stiff man syndrome

Stiffness on examination of muscles occurs in an enormous number of disorders (Table 3). In the great majority, a clinical diagnosis can be made from the nature of the stiffness, the context, and the associated neurological and general clinical findings. Muscle stiffness as a primary disorder is much less common. It occurs in three major conditions – tetanus and neuromyotonia and the stiff man syndrome. Neuromyotonia has a peripheral origin in the distal axon and neuromuscular junction, while the stiff man syndrome is a CNS disease.

Stiff man (or person) syndrome is an uncommon disorder, relatively recently recognized and generally now agreed to be of autoimmune origin. It appears to be a disorder unique among CNS diseases – a primary, non-malignant immune-mediated process caused by antibodies directed against a specific subpopulation of (spinal) neurones. It may be associated with diabetes mellitus, or with systemic autoimmune diseases, particularly lupus, and rarely is seen as an apparent paraneoplastic disorder, but in the majority of cases it occurs in isolation.

Clinical features

It presents with adult onset slowness, aching discomfort and stiffness of muscles, mainly but not exclusively, axial, and with painful muscle cramps, progressing slowly over months and years. Spasms, often noise-, startle-, or action-induced, may be very severe – tendon and muscle rupture may occur. Walking may become clumsy and unsteady. There is no disturbance of sphincter function. Examination reveals normal power and tendon reflexes, downgoing plantar responses, and no abnormalities either of sensation or (barring spasms) coordination. However, axial and abdominal wall rigidity is apparent, and there may be proximal limb muscle stiffness, agonists and antagonists acting simultaneously. An hysterical origin for the symptoms is often wrongly assumed. Asymmetrical contraction of the paraspinal muscles causes a characteristic lordotic and often scoliotic posture.

Investigations

Brain and spinal cord imaging is normal. The spinal fluid is usually normal, but for the common finding of oligoclonal immunoglobulin bands. Electrophysiological muscle examination reveals continuous muscle activity despite invitation to relax, with normal motor unit morphology. ('The patient was unable to relax during the examination' should raise suspicion.) Importantly, voluntary contraction of antagonists fails to inhibit the activity in the muscle under examination. Abnormal activity – and likewise spasms – does not persist during sleep; its central origin is confirmed by its disappearance following pharmacological peripheral nerve block or spinal or general anaesthesia, in contrast to the abnormal activity demonstrable in neuromyotonic syndromes.

Pathogenesis and detection of specific antibodies

The syndrome is thought to result from an imbalance between excitatory (catecholaminergic) and descending inhibitory (g-amino butyric acid or GABA-ergic) influences on spinal motor neurones. Antibodies directed against glutamic acid decarboxylase (GAD), the enzyme responsible for producing GABA from glutamic acid, which therefore react with GABA-ergic neurones (and with pancreatic islet b-cells) are present in 60 per cent of patients. A clonal B cell response against GAD is apparent within the CSF, partly accounting for the oligoclonal immunoglobulin bands.

In patients with cancer and stiff man syndrome, antineuronal antibodies of a different specificity to a synaptic vesicle-associated protein amphiphysin, may be found. Additionally, in the more serious stiff or progressive encephalomyelitis with rigidity (**PEWR**), with stiffness accompanied by cranial neuropathies, myoclonus, ataxia, diminished tendon jerks, and (especially) extensor plantar responses, MRI brain stem and spinal cord changes occur, and the CSF shows a pleomorphic leucocytosis. The course is substantially more aggressive, with death in 3 to 10 years. GAD antibodies may be found, but in PEWR seen in the context of cancer, antibodies against amphiphysin can also be present.

Treatment

Benzodiazepines (particularly), tizanidine, and also baclofen and occasionally sodium valproate are used therapeutically. More experimental treatments have included intrathecal baclofen and paraspinal botulinum toxin. There is now Class 1b evidence for the value of intravenous immunoglobulin.

Further reading

Adams M *et al.* (1987). Whipple's disease confined to the central nervous system. *Annals of Neurology* **21**, 104–8.

Adelman DC, Saltiel E, Klinenberg JR (1986). The neuropsychiatric manifestations of systemic lupus erythematosus: an overview. *Seminars in Arthritis and Rheumatism* **15**, 185–99.

Akman-Demir G, Serdaroglu P, Tasci B (1999). Clinical patterns of neurological involvement in Behçet's disease: evaluation of 200 patients. The Neuro-Behçet Study Group. *Brain* **122**, 71–82.

Alexander EL (1986). Central nervous system (CNS) manifestations of primary Sjögren's syndrome: an overview. *Scandinavian Journal of Rheumatology Supplement* **61**, 161–5.

Andonopoulos AP *et al.* (1990). The spectrum of neurological involvement in Sjögren's syndrome. *British Journal of Rheumatology* **29**, 21–3.

Averbuch Heller L, Steiner I, Abramsky O (1992). Neurologic manifestations of progressive systemic sclerosis. *Archives of Neurology* **49**, 1292–5.

Bathon JM, Moreland LW, DiBartolomeo AG (1989). Inflammatory central nervous system involvement in rheumatoid arthritis. *Seminars in Arthritis and Rheumatism* **18**, 258–66.

Bennett RM, Bong DM, Spargo BH (1978). Neuropsychiatric problems in mixed connective tissue disease. *American Journal of Medicine* **65**, 955–62.

Brain L, Jellinek EH, Ball K (1966). Hashimoto's disease and encephalopathy. *Lancet* **ii**, 512–14.

Caselli RJ, Hunder GG (1994). Neurologic complications of giant cell (temporal) arteritis. *Seminars in Neurology* **14**, 349–53.

Cerinic MM *et al.* (1996). The nervous system in systemic sclerosis (scleroderma). Clinical features and pathogenetic mechanisms. *Rheumatic Diseases Clinics of North America* **22**, 879–92.

Cooke WT, Smith WT (1966). Neurological disorders associated with adult coeliac disease. *Brain* **89**, 683–722.

Dalakas MC, Li M, Fujii M, Jacobowitz DM (2001). Stiff person syndrome: quantification, specificity, and intrathecal synthesis of GAD65 antibodies. *Neurology* **57**, 780–4.

Dalakas MC, *et al.* (2001).High-dose intravenous immune globulin for stiff-person syndrome. *New England Journal of Medicine* **345**, 1870–6.

Dresner SC, Kennerdell JS (1985). Dysthyroid orbitopathy. *Neurology* **35**, 1628–34.

Dropcho EJ (1996). Antiamphiphysin antibodies with small-cell lung carcinoma and paraneoplastic encephalomyelitis. *Annals of Neurology* **39,** 659–67.

Dyck PJ *et al*. (1987). Nonsystemic vasculitic neuropathy. *Brain* **110**, 843–54.

Eldor A (1998). Thrombotic thrombocytopenic purpura: diagnosis, pathogenesis and modern therapy. *Baillière's Clinical Haematology* **11**, 475–95.

Ellis SG, Verity MA (1979). Central nervous system involvement in systemic lupus erythematosus: a review of neuropathologic findings in 57 cases, 1955–1977. *Seminars in Arthritis and Rheumatism* **8**, 212–21.

Elsehety A, Bertorini TE (1997). Neurologic and neuropsychiatric complications of Crohn's disease. *Southern Medical Journal* **90**, 606–10.

Giang DW (1994). Central nervous system vasculitis secondary to infections, toxins, and neoplasms. *Seminars in Neurology* **14**, 313–19.

Good AE (1974). Reiter's disease: a review with special attention to cardiovascular and neurologic sequelae. *Seminars in Arthritis and Rheumatism* **3**, 253–86.

Hankey G (1991). Isolated angiitis/angiopathy of the CNS. Prospective diagnostic and therapeutic experience. *Cerebrovascular Disease* **1**, 2–15.

Jain R *et al.* (1994). Systemic lupus erythematosus complicated by thrombotic microangiopathy. *Seminars in Arthritis and Rheumatism* **24**, 173–82.

Johnson RT, Richardson EP (1968). The neurological manifestations of systemic lupus erythematosus. *Medicine (Baltimore)* **47**, 337–69.

Leonard TJ *et al.* (1984). Graves' disease presenting with bilateral acute painful proptosis, ptosis, ophthalmoplegia, and visual loss. *Lancet* **2**, 431–3.

Levine SR, Brey RL (1996). Neurological aspects of antiphospholipid antibody syndrome. *Lupus* **5**, 347–53.

Louis ED *et al.* (1996). Diagnostic guidelines in central nervous system Whipple's disease. *Annals of Neurology* **40**, 561–8.

Matthews WB (1968). The neurological complications of ankylosing spondylitis. *Journal of the Neurological Sciences* **6**, 561–73.

Moersch FP, Woltman HW (1956). Progressive fluctuating muscular rigidity and spasm (“stiff man syndrome”): report of a case and observations in 13 other cases. *Mayo Clinic Proceedings* **31,** 421–7.

Moore PM (1994). Vasculitis of the central nervous system. *Seminars in Neurology* **14**, 307–12.

Moore PM, Lisak RP (1995). Systemic lupus erythematosus: immunopathogenesis of neurologic dysfunction. *Springer Seminars in Immunopathology* **17**, 43–60.

Neuwelt CM *et al.* (1995). Role of intravenous cyclophosphamide in the treatment of severe neuropsychiatric systemic lupus erythematosus. *American Journal of Medicine* **98**, 32–41.

Nishino H *et al.* (1993). Neurological involvement in Wegener's granulomatosis: an analysis of 324 consecutive patients at the Mayo Clinic. *Annals of Neurology* **33**, 4–9.

Oksanen V (1986). Neurosarcoidosis: clinical presentations and course in 50 patients. *Acta Neurologica Scandinavica* **73**, 283–90.

Pincelli C *et al.* (1994). Psoriasis and the nervous system. *Acta Dermato-venereologica Supplementum (Stockholm)* **186**, 60–1.

Puechal X *et al.* (1995). Peripheral neuropathy with necrotizing vasculitis in rheumatoid arthritis. A clinicopathologic and prognostic study of 32 patients. *Arthritis and Rheumatism* **38**, 1618–29.

Scolding NJ (1999). Cerebral vasculitis. In: Scolding NJ, ed. *Immunological and inflammatory diseases of the central nervous system*, pp 210–58. Butterworth-Heinemann, Oxford.

Scolding NJ (1999). Neurological complications of rheumatological and connective tissue disorders. In: Scolding NJ, ed. *Immunological and inflammatory diseases of the central nervous system*, pp 147–80. Butterworth-Heinemann, Oxford.

Scolding NJ (1999). Organ-specific autoimmune and inflammatory disease and the CNS. In: Scolding NJ, ed. *Immunological and inflammatory diseases of the central nervous system*, pp 181–92. Butterworth-Heinemann, Oxford.

Scolding NJ, ed (1999). *Immunological and inflammatory diseases of the central nervous system*, pp. 138–46. Butterworth–Heinemann, Oxford.

Scolding NJ *et al.* (1997). The syndrome of cerebral vasculitis: recognition, diagnosis and management. *Quarterly Journal of Medicine* **90**, 61–73.

Solimena M, DeCamilli P (1991). Autoimmunity to glutamate decarboxylase (GAD) in stiff man syndrome and insulin-dependent diabetes. *Trends in Neurosciences* **14**, 452–57.

Stern BJ *et al.* (1985). Sarcoidosis and its neurological manifestations. *Archives of Neurology* **42**, 909–17.

Zajicek JP (1999). Sarcoidosis and the nervous system. In: Scolding NJ, ed. *Immunological and inflammatory diseases of the central nervous system*, pp 193–209. Butterworth-Heinemann, Oxford.

24.21 Developmental abnormalities of the central nervous system

*C. M. Verity, H. Firth, and C. ffrench-Constan**

Normal development of the human central nervous system (CNS)

The human CNS, like that of all vertebrates, develops from a two-dimensional sheet of cells into a complex three-dimensional structure. Within the CNS individual neurones establish precise connections over long distances. The resulting neural networks control and support behaviours ranging from simple reflex activities to the most complex functions of the brain. Given this complexity, it is not surprising that a range of abnormalities results from failures in distinct stages of development. This chapter describes normal development of the human CNS and uses it as a framework to discuss disorders of each phase as well as those that result from multiple disturbances of these complex processes. We have included only structural abnormalities of the CNS that are present at birth and have not reviewed the numerous metabolic and degenerative disorders that can affect the infant brain.

Induction

This is a process by which one group of cells acts upon another group so that the second group differentiates or in some way alters its behaviour. Following the development of the three cell layers of the early embryo (ectoderm, mesoderm, and endoderm), signals from the underlying mesoderm (the 'inducer') instruct a region of the ectoderm (the 'induced tissue') to adopt a neural fate.

Neural tube formation

The neural ectoderm folds to form a tube which separates from the ectoderm and runs most of the length of the embryo. Closure starts at a level corresponding to the future hindbrain/spinal cord junction and then proceeds both towards the head (rostrally) and the tail (caudally). The most caudal part of the neural tube is formed by the thickening of the neural plate and the subsequent formation of a cavity (rather than a tube as is seen rostrally).

Regionalization and specification

Once the neural tube has developed, specification of different regions and individual cells within these regions occurs. This patterning occurs in both the rostrocaudal and dorsoventral axis. The three basic regions of the CNS (forebrain, midbrain, and hindbrain) develop at the rostral end of the tube, with the spinal cord more caudally. Within the developing cord the specification of the different populations of neural precursors (neural crest, sensory neurones, interneurones, glial cells, and motor neurones) is observed in progressively more ventral locations. This process reflects the interaction between genes whose expression defines individual territories or cell types, and diffusible signalling molecules secreted by adjacent areas of the embryo.

Some of the genes and signalling molecules involved in these processes were originally identified by genetic studies in the fruit fly *Drosophila melanogaster*, and appear to be very highly conserved throughout evolution. Of particular importance are a family of genes, called homeotic genes, most of which contain a conserved sequence called a homeobox encoding a protein motif called a homeodomain. This motif binds DNA sequences involved in the regulation of expression of other genes important in development, and so can regulate cell differentiation. For example, homeotic genes have been shown in *Drosophila* to define the identity of individual segments of the fly's body, so that homeotic mutations alter segment identity. The same basic mechanism patterns the vertebrate CNS. In addition, two of the key extracellular signalling molecules, encoded by the genes *wingless* and *sonic hedgehog* (*shh*), were first identified in *Drosophila*.

An example of how these genes and signalling molecules interact is provided by the process of ventral induction. This is illustrated by the specification of motor neurones and glial cells in the spinal cord. Initially the cells of the tube have no axis. The notocord, a transient structure ventral to the neural tube, then produces *shh* protein. This induces the ventralization of the adjacent tube and the expression of two homeobox-containing genes (*PAX6* and *NKX2.2*) in this region. These genes in turn allow motor neurones to develop in the ventral tube, but this requires further exposure to *shh*. Because it is produced in ventral sites, the *shh* molecule is present in a gradient from ventral to dorsal, and this provides further patterning information. Motor neurone specification occurs at a higher concentration of *shh* than that required for the specification of the glial precursor cells. As a consequence, motor neurones develop correctly in a location ventral to the glial cells.

Proliferation and migration

Following the establishment of the basic plan of the CNS, the most dorsal cells of the tube (the neural crest) migrate away to form much of the peripheral nervous system. At the same time, cell proliferation and migration within the tube leads to thickening of the wall and the movement of the many different cell types to their correct locations. The developing forebrain cortex provides a good example. An area called the germinal matrix adjacent to the lumen of the neural tube (which will become the ventricular system) contains populations of neural stem cells that divide to generate committed precursor cells of the different cell lineages of the CNS—neurones and the two glial cell types, oligodendrocytes and astrocytes. These precursor cells then migrate to their final locations. Many of the neurones migrate along specialized cells called radial glial cells, whose processes span the entire thickness of the developing cortex, before leaving the glial fibre to join a specific cortical layer. The time at which the precursor cells are generated defines which type of neurone they will become, and the cortex is built in an 'inside-out' pattern with each neuronal type then migrating beyond its predecessors before leaving the radial fibre. In this way, the six layers of the cortex each containing different neuronal populations are established. Abnormalities of migration in mice and humans have been linked to specific genetic mutations. For example mice with abnormalities in an extracellular matrix molecule called reelin or in its receptors show aberrant patterns of cortical lamination and cerebellar development.

Connection and selection

Once each cell type is specified and in an appropriate location, axon outgrowth and the formation of synapses occurs. These connections are made both locally and over considerable distances, as for example in the contralateral connections made by axons running in the corpus callosum (a fibre tract crossing from one side of the brain to the other). The mechanisms by which the vast number of connections are made are complex and incompletely understood. There are both attractive and repulsive cues that guide growth cones within the CNS. Examples are the netrins, that are attractive to most outgrowing axons and are produced by the ventral floor

plate of the spinal cord before axons grow into it, and the semaphorins that can repel growth cones. In addition there is evidence that recognition molecules such as cadherins and ephrins can regulate cell–cell adhesion and intracellular signalling and allow axons to select precise targets once they have arrived in the appropriate location. In many cases such signalling and recognition molecules exist as large families of related molecules, a diversity that may underlie the development of such complex connections. Finally, cells that fail to establish the correct connections undergo programmed cell death (apoptosis) as a result of a failure to obtain factors produced by the target cells that are required for survival. In this way, errors in pathfinding or cell production can be corrected at a later stage. At the same time, neurones establish close interactions with glial cells that either form the myelin sheaths essential for rapid impulse conduction (oligodendrocytes) or regulate the extracellular environment of the neurone and so play a homeostatic role essential for correct function (astrocytes).

Although developmental abnormalities can occur at any stage of gestation, some developmental events such as neural tube formation and specification occur in the first month after conception. This is important from a clinical standpoint, as developmental abnormalities resulting from environmental factors can occur before pregnancy is confirmed. Dietary folic acid supplements can reduce the risk of neural tube defects but need to be taken from the very earliest stages of pregnancy and should be started before conception.

Structural abnormalities resulting from disturbances of normal CNS development

Disorders of neural tube formation

Introduction

The neural tube usually fuses completely between 18 and 26 days after ovulation. Failure of closure leads to malformations that include anencephaly, encephalocele, spina bifida, and spina bifida occulta. These neural tube defects are aetiologically related and if one member of a family is affected there is an increased risk in the relatives for all types of neural tube defect. They are malformations of the neuroectoderm which are associated to a variable extent with abnormalities of the surrounding mesodermal structures. The term dysraphism is used when there is continuity between the posterior neuroectoderm and cutaneous ectoderm.

Epidemiology

The prevalence of neural tube defects varies according to geography and race, although they are among the most common congenital abnormalities in most countries. High rates (more than 8 per 1000 births) have been reported in Northern Ireland, Egypt, India, and China. There are worldwide reports of decreasing prevalence rates. In England there was a peak in 1954 to 1955 followed by a substantial decline which started in the early 1970s. The epidemiological evidence suggests that this decrease was not entirely due to prenatal screening, and it also preceded the widespread use of periconceptual vitamin supplemention, so some of the decrease remains unexplained. By 1994 the prevalence of neural tube defects in England and Wales was just under 0.8 per 1000 total births. In the United Kingdom anencephaly and spina bifida are of approximately equal prevalence and together make up 95 per cent of all neural tube defects.

Aetiology

Genetic factors

Most neural tube defects are a complex interaction between several genes and environmental factors. Major genes have been identified in the mouse that may mutate and cause neural tube defects, but their relevance to human defects is still not clear. More is being discovered about the identity of modifier genes. For instance mutations in the methylene tetrahydrofolate reductase gene are associated with elevated blood homocysteine levels in pregnant women and and increased risk of neural tube defects. However, the relative risks are low (about twofold) and mutation analysis is not used in routine clinical practice. Neural tube defects occur in many different syndromes and many chromosome disorders, but if a neural tube defect is the only anomaly, karyotyping is not indicated.

Environmental factors

Periconceptual multiple vitamin supplements containing folic acid have been shown to reduce the incidence of neural tube defects. In England it is currently recommended that, to prevent a first occurrence of neural tube defect, women who are planning pregnancy should take 400 µg of folic acid daily before conception and during the first 12 weeks of pregnancy. To prevent recurrence of neural tube defect the dose should be 4 to 5 mg per day.

Some drugs taken during pregnancy may increase the risk of neural tube defects in the fetus. These include sodium valproate and folic acid antagonists such as trimethoprim, triamterene, carbamazepine, phenytoin, phenobarbitone, and primidone.

Prenatal diagnosis

a-Fetoprotein levels in maternal serum

The fetal liver is the main source of a-fetoprotein (AFP), which leaks through open neural tube defects into the amniotic fluid and then into the maternal blood. This abnormal increase in maternal serum a-fetoprotein is best detected at 16 to 18 weeks of pregnancy. Maternal serum screening does not detect closed defects (those covered by skin) and is less sensitive in women taking the antiepileptic drug sodium valproate.

Ultrasonography

This is recommended for all at-risk women—those with positive serum a-fetoprotein screening, those who have had one or more affected child, and those taking drugs associated with neural tube defects in the fetus. Anencephaly can be detected by ultrasound from the 12th week of gestation and spina bifida from 16 to 20 weeks (Fig. 1(a) and Fig. 1(b)). However, even the best ultrasonographers may occasionally miss spina bifida, particularly in the L5–S2 region.

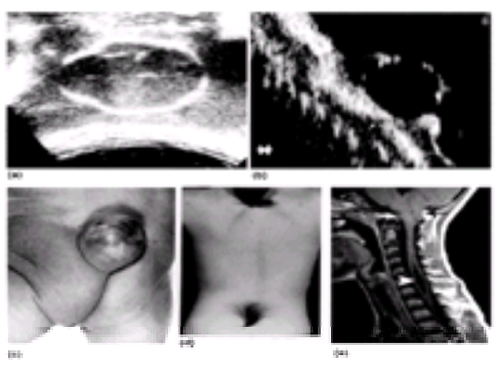

Fig. 1 (a) Prenatal ultrasound of a child with a neural tube defect, showing the 'lemon sign' resulting from the change in shape of the back of the skull (on the left-hand side in the image) which is associated with the Chiari II malformation described in the text. (b) Prenatal ultrasound of a child with a neural tube defect, showing a cystic lumbar meningomyelocele in the caudal neural tube. (c) Lumbar meningomyelocele. Photograph of a newborn infant. (d) Chiari I malformation and syringomyelia in an asymptomatic girl aged 11 years. Photograph of tuft of hair seen over the lumbar region at birth. The associated CNS malformations are shown in scan (e). (e) Chiari I malformation and syringomyelia. T_1-weighted sagittal MRI shows there is herniation of the cerebellar tonsils through the foramen magnum (arrow) and a syrinx of the lower cervical spinal cord (C5–C7) (arrow head). The associated tuft of lumbar hair is shown in photograph (d).

Amniocentesis

This has been largely superseded by detailed ultrasound imaging. When adequate ultrasound images cannot be obtained, amniocentesis with measurement of of a-fetoprotein and assay of neuronal acetylcholinesterase provides an alternative method of prenatal diagnosis.

Cranial dysraphism

Anencephaly

This is a lethal defect that results from failure of fusion of the rostral folds of the neural tube between days 18 to 25 of embryonic development. The cranial vault is absent and an angiomatous membranous mass lies on the floor of the cranium. The eyes are protruberant because of shallow orbits and there is variable involvement of the spinal cord. Before the advent of prenatal diagnosis by ultrasound most anencephalic babies were born in the last 3 months of gestation; now an increasing number of such pregnancies are terminated. In liveborn anencephalic babies the initial neurological examination may be surprisingly normal if brainstem structures are reasonably intact and seizures may be seen despite the absence of cerebral hemispheres. However, the infants usually die in hours or days.

Cephaloceles

A cephalocele is a herniation of cranial contents through a skull defect. There are several subtypes: a cranial meningocele contains only meninges, an encephalocele contains brain tissue, and a ventriculocele contains part of the ventricle within the herniated portion of brain. Cephaloceles are less common than anencephaly or spina bifida, occurring in 1 to 3 per 10 000 live births. They are associated with other brain abnormalities such as agenesis of the corpus callosum or abnormal gyration. They may be part of a recognized syndrome, so it is important to look for abnormalities in other parts of the body. Sometimes neurosurgery is indicated.

Posterior cephaloceles

This is the commonest group in Western countries and the majority are occipital encephaloceles. They are of variable size and may be above or below the tentorium. The latter are often associated with severe cerebellar defects, such as the Chiari III malformation (see under posterior fossa abnormalities). The prognosis depends on the size of the encephalocele, the site, and the associated abnormalities, such as hydrocephalus. Visual impairment may result from involvement of the occipital lobes. There may be motor and intellectual impairment and seizures.

Anterior cephaloceles

These are more common in some parts of Asia. Frontoethmoidal cephaloceles may protrude into the nose, the ethmoid, or the orbit. They often include olfactory tissue and frontal lobe tissue and can present with nasal blockage or cerebrospinal fluid leakage. Sphenoidal cephaloceles can cause pharyngeal obstruction and recurrent meningitis and may be associated with abnormalities in the secretion of somatotrophin, gonadatrophin, or antidiuretic hormone.

Spinal dysraphism

Spina bifida

This can be divided into spina bifida occulta, which consists of failure of closure of the vertebral arches without an external lesion, and spina bifida cystica in which there is a cystic lesion on the back. The lesion may be either a meningocele without neural tissue or a myelomeningocele in which the spinal cord is a component of the cyst wall.

The term rachischisis is used for the most severe defect, which is a widely patent dorsal opening of the spine, often associated with anencephaly.

Myelomeningocele

The spinal defect

This is the abnormality found in 80 to 90 per cent of children with spina bifida cystica. It is lumbosacral in about 80 per cent of cases and consists of a sac covered with a thin membrane which may leak cerebrospinal fluid (Fig. 1(c)). Neurological abnormalities depend on the level of the lesion, which is best judged clinically by determining the upper limit of sensory loss. There is usually a mixture of upper and lower motor neurone signs depending on the level. Whatever the level of the lesion, there is disturbance of bladder and bowel sphincters and also bladder detrusor dysfunction. The sensory level correlates with the severity of abnormalities in the urinary tract and is also related to long-term disability. Higher lesions of the cord are associated with bladder outlet obstruction, dilatation of the upper urinary tract, and chronic pyelonephritis.

Hydrocephalus

This complicates most cases of lumbosacral meningomyelocele. Ultrasound studies show hydrocephalus in about 90 per cent of cases at birth, even though the head circumference is normal. Usually it is associated with the Chiari II malformation (see below), although it may be due to aqueduct stenosis or have no clear structural cause. If there is evidence of progressive ventricular dilatation (often detected with ultrasound) or signs of increasing intracranial pressure, insertion of a ventriculoperitoneal shunt is usually necessary.

Chiari II malformation

This is present in about 70 per cent of cases of meningomyelocele. It is the most common of the four types of Arnold–Chiari malformation. It consists of downward protrusion of the medulla below the foramen magnum to overlap the spinal cord. The medulla is kinked and the cerebellar vermis indented by the posterior lip of the foramen magnum. The fourth ventricle is elongated and the midbrain distorted, which can cause early or late problems. These include palsies resulting from involvement of the lower cranial nerves and central apnoea (which may be misdiagnosed as epilepsy in older children).

Almost all cases of Chiari II malformation are associated with meningomyelocele. In contrast the other types of Arnold–Chiari malformations (I, III, and IV) are not associated with spina bifida and are dealt with in the section on posterior fossa abnormalities.

Management

Treatment of infants with meningomyeloceles became possible with the development of ventriculoatrial and ventriculoperitoneal shunts. In the early 1960s it was argued that closure of the defect within 24 h of birth reduced mortality and morbidity by avoiding infection and reducing trauma to the exposed neural tissue. The early active management of all cases was questioned by Lorber who proposed that surgery should be selective. He reported four adverse criteria that he thought were contraindications to treatment: a high level of paraplegia, clinically evident hydrocephalus at birth, lumbar kyphosis, and the presence of other major malformations. However, using these criteria the outcome was uncertain; many infants survived even though they did not have closure of the defect within 24 h, and some children with a supposedly good prognosis were left with major disabilities after surgery. Selective surgical management is therefore not universally practised and this remains a controversial area.

Now the emphasis has moved towards prevention. It is recommended that women planning to conceive supplement their diet with folic acid, which reduces the risk of neural tube defects. Screening of maternal serum for a-fetoprotein is possible and prenatal diagnosis by ultrasound and amniocentesis is available. This is discussed above in the section on neural tube defects.

Meningocele

Here there is protrusion of the meninges outside the spinal canal: the sac does not contain any neural tissue. Meningoceles account for about 5 per cent of cases of spina bifida cystica. There is no associated hydrocephalus and the neurological examination is usually normal. They must be distinguished from meningomyeloceles

because the prognosis is so different.

Occult spinal dysraphism

The term spina bifida occulta is often applied to a defect of the posterior arch of one or more lumbar or sacral vertebrae (usually L5 and S1). It is found incidentally by radiography in 25 per cent of children admitted to hospital and may be regarded as a normal variant. However, it must not be assumed that spina bifida occulta is always benign. If examination of the skin over the spine reveals a naevus, hairy patch (Fig. 1(d)), dimple, sinus, or subcutaneous mass, further evaluation is necessary. Even if there are no associated abnormalities of sphincter or limb control on MRI of the spinal cord is indicated. A spinal cord malformation may cause an asymmetrical lower motor neurone weakness with wasting, deformity, and diminished reflexes in the lower limb. Alternatively there may be a progressive gait disturbance with spasticity. Either presentation may be associated with disturbed bladder control. Several different abnormalities may be found.

Dorsal dermal sinuses may connect the skin surface to the dura or to an intradural dermoid cyst. They are most commonly found in the occipital and lumbosacral regions. An open sinus tract can cause recurrent meningitis so ideally they should be explored and excised before infections occur. Lipomyelomeningoceles present as a bulge in the lumbosacral region, usually lateral to the midline. They consist of a lipoma or lipofibroma attached to a low lying abnormal spinal cord and are often associated with a meningocele. Diastematomyelia is the presence of a sagittal cleft which divides the spinal cord into two halves, each surrounded by its own pia mater. A bony or cartilaginous spur may transfix the cord, fixing it in a low position as the child grows. The cleft is usually in the low thoracic or lumbar region, but cervical clefts have been reported. In 75 per cent of cases of diastematomyelia an overlying midline skin abnormality is present and plain radiographs show abnormalities in most cases—these include abnormal segmentation of vertebrae, spina bifida, and scoliosis.

Treatment of occult spinal dysraphism

If an abnormality involving the cord or nerve roots is found, there is often a good case for neurosurgical intervention. The aim is to free the spinal cord from its abnormal attachments to allow for growth and prevent further damage. Early intervention may prevent worsening motor deficits and urological complications, but the indications for intervention are controversial.

Other developmental abnormalities of the spinal cord

Syringomyelia

This is a tubular cavitation of the spinal cord as opposed to hydromyelia which is dilatation of the central canal of the cord—a distinction that may be difficult to make clinically. Syringomyelia is often associated with the Chiari I malformation and hydrocephalus (Fig. 1(e)). It tends to be in the cervical region but may involve the whole cord. It rarely becomes symptomatic in children. Treatment is controversial. Shunting of the abnormal cavity is sometimes performed and posterior fossa exploration may be undertaken if there is a Chiari I malformation.

Sacral agenesis

This is strongly associated with maternal diabetes mellitus. Absence of the acrum and coccyx is usually associated with abnormalities of the lumbosacral cord. There may be arthrogryposis at birth (defined as a fixed deformity of one or more joints). A flaccid neurogenic bladder causes incontinence and there are sensory and motor deficits in the legs.

Disorders of regionalization

Failure of normal development of the most anterior portion of the neural tube (the mediobasal prosencephalon) and associated structures due to disturbances in the process of ventral induction described above may result in various abnormalities of the brain and face. The most severe CNS abnormality is holoprosencephaly in which there is failure of the prosencephalon to separate into two cerebral hemispheres. The mildest is olfactory aplasia without other cerebral malformations. The severity of the associated facial abnormalities tends to parallel those in the brain. In the most severe facial abnormality there is anophthalmia and absence of the nose. However there may be just mild hypotelorism (closely set eyes), a single central incisor tooth, or the face may be normal.

Holoprosencephaly (prosencephaly)

This occurs with a frequency of approximately 1 in 14 000 births. There is failure of formation of the two cerebral hemispheres resulting in abnormalities of varying severity. The causes may be chromosomal, multifactorial, or monogenic and there may be associated facial and midline malformations.

Aetiology

This is time specific and stimulus non-specific. There is a very short vulnerable period, because ventral induction probably occurs prior to 23 days, just before the elaboration of the optic vesicles (24 days), which explains its relative rarity (6 per 10 000 live births in one study). Environmental factors may be important: it is at least 20 times more common in the infants of mothers with diabetes than in the general population.

Genetic factors are also important. At least 12 genetic loci and several holoprosencephaly (*HPE*) genes have been identified in humans. Some of these genes can be linked to signalling pathways described above in the section on CNS development. One (*HPE3* on chromosome 7q36) is the *sonic hedgehog* gene encoding a secreted protein required for ventral induction throughout the neuraxis. Mutations in PATCHED-1, the receptor for *sonic hedgehog*, have been found in individuals with holoprosencephaly. In addition *ZIC2*, present on chromosome 13q32 and encoding a human homologue of the *drosophila odd-paired* gene is implicated in *wingless* signalling. Two other genes, *SIX3* and *TGIF*, both encode homeodomain proteins of undefined function. A number of chromosomal abnormalities have also been recognized in association with holoprosencephaly. These include trisomy and other abnormalities of chromosome 13, partial deletion of the short arm of chromosome 18, ring chromosome 18, and partial trisomy of chromosome 7.

Clinical features

In alobar holoprosencephaly the completely undivided forebrain is in the shape of a horseshoe surrounding a single cavity. The thalami are fused but the brainstem and cerebellum are well developed. The associated facial abnormalities are severe—there may be anophthalmia or cyclopia in which there is a single orbit. In holoprosencephaly with median cleft lip there is marked hypotelorism. In semilobar holoprosencephaly the brain is divided into two hemispheres posteriorly but anteriorly the two hemispheres are fused (Fig. 2). In lobar holoprosencephaly there is almost complete separation of the hemispheres and the face may be normal. The head is usually microcephalic unless there is associated hydrocephalus.

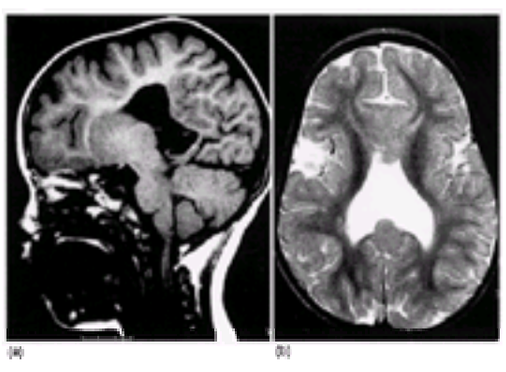

Fig. 2 (a) Semilobar holoprosencephaly in a girl aged 2 years imaged with T_1-weighted sagittal MRI. This midline view shows absence of the corpus callosum and fusion of the frontal lobes. (b) Semilobar holoprosencephaly in the same patient using T_2-weighted axial MRI. There is fusion of the frontal lobes of both cerebral

hemispheres and a common central ventricle.

The most severely affected infants die in the neonatal period. Less severely affected patients may live for months or years. The survivors often develop infantile spasms or other seizures. Some patients with significant structural abnormalities may survive to adulthood but usually there are severe learning difficulties. Associated anomalies include congenital heart disease, scalp defects, and polydactly. Holoprosencephaly may be part of syndromes such as the Meckel–Gruber syndrome and the Aicardi syndrome.

Genetic counselling

Families have been described in which there is dominant inheritance of minor features and one or more children with holoprosencephaly. It is therefore important to look for minor signs in both parents of an affected child. These include orbital hypotelorism, median cleft lip, flat nose with or without a single nostril, anosmia, and a single central incisor.

Prenatal diagnosis can be made by ultrasound from the 16th week of pregnancy and holoprosencephaly accounts for a proportion of the cases of hydrocephalus diagnosed antenatally. Orbital hypotelorism is a reliable diagnostic feature for antenatal diagnosis by ultrasound.

Disorders of cortical development

Modern brain imaging, in particular MRI, has resulted in the identification of many previously unrecognized developmental abnormalities of the cerebral cortex. Following the classification used by Aicardi this section is divided into disorders of proliferation, migration, and cortical organization.

Disorders of proliferation and differentiation

Microcephaly

This is an abnormally small head, which is disproportionately small in relation to the rest of the body. A child is microcephalic when the head circumference is below the normal range (less than the 0.4th centile), defined by head growth charts appropriate for sex and race. Some children with microcephaly are neurologically normal.

Genetic causes of microcephaly

There is a genetically determined type of microcephaly in which the inheritance is usually autosomal recessive with at least three loci currently identified, but may be autosomal dominant or X linked. Characteristically there is marked microcephaly but the neurological problems are relatively mild. They consist of fine motor incoordination and hyperkinetic behaviour with moderate learning difficulties and seizures. However, it is more usual to find that patients with microcephaly have more significant abnormalities of the nervous system such as pyramidal tract signs and profound learning difficulties. Microcephaly is a feature of more than 450 syndromes listed in the Oxford Dysmorphology Database, so it is a challenge to differentiate genetic from other causes.

Non-genetic causes of microcephaly

Ionizing radiation in the first two trimesters of the pregnancy, intrauterine infections, drugs and other chemicals, circulatory disturbance, and perinatal hypoxic–ischaemic events can all cause microcephaly. Poor dietary control in mothers with phenylketonuria is also an important cause of microcephaly, as the fetal brain is very sensitive to the toxic effects of phenylalanine. Sometimes serial head circumference measurements are necessary to make the diagnosis. When there is a significant perinatal insult to the brain the head circumference may be normal at birth with subsequent failure of growth in the first few months of life. However, the first head circumference measurement at birth may be misleading because of skull moulding during delivery.

It is important to perform a skull radiograph to look for evidence of early closure of all the cranial sutures (total craniosynostosis). In some types of genetic microcephaly the head size falls off as late as 32 to 34 weeks of gestation or even after birth. Prenatal diagnosis by ultrasound may be difficult.

Megalencephaly

The term macrocephaly is used when the head circumference is above the normal range for the age, sex, and race of the child. This may result from abnormalities outside the brain parenchyme such as hydrocephalus, arachnoid cysts, congenital abnormalities of the cerebral veins, fluid collections over the surface of the brain, or abnormalities of the skull. Cranial imaging is necessary to make the diagnosis. This discussion deals only with megalencephaly, which is increased size of the brain itself.

Many normal individuals have large heads. When assessing a child with a large head it is important to exclude a developmental or neurological abnormality. A large head may be part of a specific disorder, for example one of the neurocutaneous syndromes, or an overgrowth disorder such as Sotos syndrome. Large heads can run in normal families ('familial megalencephaly') and it is important to check the head circumference of the parents. However, there are kindreds in which some of the family members with large heads have learning problems.

Sometimes megalencephaly is associated with significant learning difficulties, autism, neurological abnormalities, and seizures, and this combination of features can have a genetic basis. The brains may have bulky gyri and usually all parts of the cerebrum are diffusely enlarged, with normal-sized or mildly enlarged ventricles. Occasionally particular parts of the brain such as the cerebellum are disproportionately large. No consistent microscopical alterations are reported in the cortex, but minor anomalies such as small heterotopias may be found. The pathogenesis of this type of megalencephaly is not clear: it may be caused by overproduction of CNS cells, possibly combined with failure of apoptosis.

Hemimegalencephaly or unilateral megalencephaly may involve all parts of the brain on the same side or there may be enlargement of one hemisphere only. In the affected hemisphere there are broadened and coarse gyri and sometimes areas of polymicrogyria. The microscopic abnormalities vary. There may be disorganized masses of grey matter without laminar organization or nodular heterotopias of grey matter: giant neurones may be found.

The neurological problems associated with hemimegalencephaly can be severe. Sometimes there are intractable seizures, which may start in infancy, associated with marked developmental delay and sometimes with hemiparesis. There may be overgrowth of one side of the face or of one side of the whole body. Some infants may present with prenatal or perinatal onset of seizures and die early. Others may develop seizures later which may be refractory to drug treatment, and they may be candidates for hemispherectomy.

Disorders of migration

Migration defects occur when neurones of the subependymal matrix zone lining the ventricular cavity (the ventricular zone) fail to reach their intended destination in the cerebral cortex. This results in either major or minor disturbances of development which may be focal or diffuse. If neurones fail to leave the ventricular zone entirely, periventricular heterotopias result. If neurones leave the ventricular zone but then fail to complete their migration in the cortex, this causes lissencephaly. If, however, only a subpopulation of neurones are affected and the others complete their migration normally, then this results in nodular or band heterotopias. Migration disorders are found as part of recognized syndromes and there are also acquired types resulting from intrauterine infections, circulatory disturbances, and toxins (alcohol or phenytoin for example).

Agyria–pachygyria (lissencephaly)

This is a group of disorders of varying severity. There may be complete absence of gyri, in which case the terms agyria or lissencephaly (greek: 'smooth brain') are

used. Pachygyria describes a reduced number of broadened and flat gyri with less folding of the cortex than normal. There may be varying degrees of agyria–pachygyria in the same brain.

Type I lissencephaly

Here the brain is small with only the primary and sometimes a few secondary gyri. The cortex is thick with the white matter forming a thin rim along the ventricles (Fig. 3(a)). In the brainstem the olivary nuclei are ectopic and the pyramids are hypoplastic or absent. In the cerebellum the dentate nuclei are abnormally convoluted. There may be associated agenesis of the corpus callosum. Infants with type I lissencephaly may be divided into the majority who have no dysmorphic features ('isolated lissencephaly sequence') and those with the dysmorphic features of the Miller–Dieker syndrome.

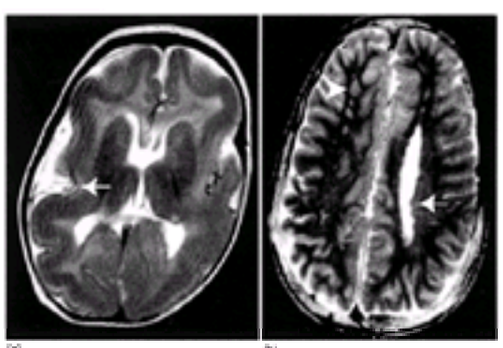

Fig. 3 (a) Lissencephaly type I in a girl aged 5 months who is visually and socially unresponsive, displays poor feeding, and increased tone. The T_2-weighted axial MRI shows a smooth cerebral cortex with absence of normal gyri and sulci. On the right a vertically orientated shallow sylvian fissure is seen (arrow). (b) Nodular heterotopias in a boy aged 13 years. The T_2-weighted axial image shows the nodular heterotopias are subependymal (arrow) and subcortical (arrow head).

The Miller–Dieker syndrome

Characteristically in this syndrome there is postnatal growth deficiency and occasional microcephaly. The dysmorphic features include a tall narrow forehead, a depressed nasal bridge, anteverted nares, midfacial hypoplasia, a prominent upper lip with a thin vermilion border, retrognathism, and hypervascularization of the retina. Hypotonia and seizures are associated. MRI and CT scans show lissencephaly together with a midline focus of calcification in the callosal remnant in about 40 per cent of cases (this is rarely seen in the isolated lissencephaly sequence). About 50 to 70 per cent of cases have a deletion of 17p13.3 by light microscopy and almost all the remainder have a submicroscopic deletion demonstrable by fluorescent *in situ* hybridization of chromosomes using gene probes. This region includes the *LIS1* gene, mutations in which are associated with isolated lissencephaly and which is also required for correct neuronal migration in mice. It has been suggested, therefore, that deletions of the *LIS1* gene cause the lissencephaly seen in the Miller–Dieker syndrome and that the facial dysmorphism is caused by loss of adjacent genes.

Isolated lissencephaly sequence

This is a heterogeneous group. Many have a deletion of, or mutations within, the *LIS1* gene. Mutations in a second gene, *doublecortin* (*DCX*) have also been shown to cause lissencephaly. *DCX* is on the X chromosome, explaining why inheritance of isolated lissencephaly sequence in some cases is X linked. While affected males in these families show the full isolated lissencephaly sequence phenotype, carrier females can show band heterotopia in which a subset of neurones fail to complete migration and form bilateral symmetrical ribbons of grey matter in the centrum semiovale. This is thought to reflect X inactivation of the normal *DCX* gene in these neurones, while those that inactivate the mutation-containing X chromosome migrate normally. In addition to *LIS1* and *DCX*, mutations in the human gene encoding *reelin* (RELN) have recently been described in lissencephaly associated with cerebellar hypoplasia. As described above, *reelin* gene mutations are found in some naturally-occurring mouse mutants with abnormal neuronal migration, and so this provides a direct link between the animal and human studies. Possible non-genetic causes include intrauterine cytomegalovirus infection and early placental insufficiency. Clinically there is severe mental retardation and diplegia together with partial seizures and infantile spasms.

Diagnosis of type I lissencephaly

The diagnosis of type I lissencephaly is made by CT or MRI, which show a thick cortical plate with no or few sulci separated from the white matter by an undulating border. The differential diagnosis includes peroxisomal disorders such as Zellweger syndrome, but these have their own specific features. Some cases are due to cytomegalovirus infection and there may be associated periventricular calcification. Prenatal diagnosis is not possible by ultrasound before 24 weeks as tertiary sulci do not appear before then.

Genetic counselling

In the Miller–Dieker syndrome, if a deletion is found in the *LIS1* gene and the parental chromosomes are normal, the recurrence risk is less than 1 per cent. In the isolated lissencephaly sequence, if chromosome and DNA studies are negative and the other differential diagnoses are excluded, the empirical recurrence risk is 5 to 7 per cent.

Type II lissencephaly or Walker–Warburg syndrome

This is also called 'cobblestone lissencephaly' and is a completely different malformation from type I lissencephaly. The smooth cortex has a granular surface and is covered with meninges that are thickened due to mesenchymal proliferation. The cerebellum is small with an absent vermis and the pyramidal tracts are usually absent. Hydrocephalus is present in 75 per cent of cases. Microscopically there is complete disorganization of the cortex which consists of neurones separated by bundles of gliomesenchymal tissue continuous with the meninges. More deeply there is a thin layer of white matter lying above islands of heterotopic grey matter. These abnormalities probably result from overmigration of neuroglial precursors through a disrupted pial–glial limiting membrane.

Clinical features

The clinical features include both nervous system and muscle abnormalities. The infants are very abnormal at birth. The eyes show retinal dysplasia, microphthalmia, and anomalies of the anterior segment. There may be hydrocephalus or sometimes microcephaly. Usually there is an elevated creatine kinase and necrosis of fibres in all muscles, similar to that seen in severe muscular dystrophy.

Diagnosis

The diagnosis of type II lissencephaly is confirmed by MRI or CT scan, which show that the cortex is thinner than in type I lissencephaly and there are characteristic trabeculae penetrating the cortex from the white matter with the appearance of a double cortical layer due to the subcortical heterotopic islands. There is agenesis of the cerebellar vermis and often a posterior encephalocele or large posterior fontanelle.

Differential diagnosis

There are a group of conditions involving muscle, eye, and brain that all follow recessive inheritance. It seems likely that the cerebro–ocular–muscle syndrome (COMS) is identical to the Walker–Warburg syndrome for which the gene locus is not yet identified. Fukuyama-type muscular dystrophy is associated with severe mental retardation. It is relatively common in Japan and is due to mutations in the Fukutin gene at 9q31–33. Muscle–eye–brain disease is found in Finland. The eye

and brain abnormalities are less severe and the gene locus is on 1p32–34.

Non-lissencephalic cortical dysgenesis

Polymicrogyria (microgyria) is the most important type of abnormality in this section, although the aetiology remains poorly understood. Sometimes the surface of the cortex is relatively smooth resembling pachygyria because the small gyri pile upon each other to form a thickened cortex. The histology of polymicrogyria varies. In unlayered microgyria there is a single cell layer between the white matter and the molecular layer. In classic four-layered microgyria the cortex consists of a molecular layer, an upper dense layer, a layer of low cellular density containing myelinated fibres, and a deep cellular layer. It is suggested that the developmental disturbance occurs near the fifth month of pregnancy. Case reports of polymicrogyria in the infant brain after maternal trauma or asphyxiation during the pregnancy suggest that the abnormality may sometimes be due to failure of cerebral perfusion with resulting hypoxia.

The clinical manifestations of polymicrogyria depend on the location and extent of the abnormalities. Small patches may be found incidentally in the absence of symptoms, but there may be involvement of the whole cortex, or areas of polymicrogyria may border porencephalic cysts in patients with neurological disabilities. There is a bilateral perisylvian syndrome (or anterior operculum syndrome) in which bilateral opercular abnormalities are seen on MRI, some of which have the appearance of polymicrogyria. These patients have a pseudobulbar palsy with dysarthria, and loss of voluntary control of the face and tongue leading to drooling and difficulty feeding. Familial occurrence has been reported on several occasions.

Heterotopias

Periventricular heterotopias are abnormal collections of neurones in the subependymal region. They may be part of a complex malformation syndrome such as the Aicardi syndrome. They may be isolated and clinically silent or associated with seizures. A gene responsible for periventricular heterotopia, the *filamin 1* (FLN1) gene, has been identified on the X chromosome. Filamin protein reorganizes the cytoskeleton, consistent with a role in cell migration. Families with periventricular heterotopia have been described in which females are affected while affected males appear to die before or soon after birth. It is likely that the heterotopias present in affected females result from X inactivation of the normal *FLN1* gene in those cells, while those cells inactivating the abnormal *FLN1* gene migrate normally. Males have only one X chromosome and so all cells will fail to migrate—a lethal phenotype in these families. However, in other families some surviving males with periventricular heterotopias have now been shown to have *FLN1* mutations.

Subcortical heterotopias can be divided into two groups. Nodular heterotopias of grey matter are found in association with other migration disorders and may be the cause of partial seizures (Fig. 3(b)). Subcortical laminar heterotopias are also known as band heterotopias or 'double cortex' and may be inherited as an X-linked trait, as discussed above. Patients with subcortical heterotopias often have seizures, which may be focal or generalized, and they may also have intellectual problems although some do develop normally.

Disorders of cortical organization

There is increasing interest in developmental abnormalities within the cerebral cortex that are relatively subtle compared with those described above, as they may represent important causes of epilepsy and developmental delay. It is likely that this group of disorders will become increasingly well recognized as imaging and other investigative techniques improve.

Cortical microdysgenesis

In 1971 Taylor and colleagues described localized cortical abnormalities in the brains of 10 patients with intractible epilepsy. The lesions could only be seen microscopically and consisted of an excess of large abnormal cells scattered throughout the cortical layers. The abnormal neurones were restricted to sharply delineated areas of cortex and also formed foci in the depths of sulci.

Experiments have been performed in rats using microelectrodes to freeze small areas of cortex, leading to microgyri at these sites. They show that focal cortical abnormalities are associated not only with alterations in the membrane properties and synaptic connections of the neurones directly involved, but also with much more widespread abnormalities of neuronal circuits. Thus there may be more global effects on cerebral function than expected from the size of the lesion.

Microscopic abnormalities of cortical arrangement have been described in the brains of patients with epilepsy or learning difficulties. They include persistence of the subpial layer, aggregates of large neurones in the plexiform zone, a fragmented appearance of the superficial neuronal layers, excess ectopic cells in the cortex, and excess numbers of cells in the molecular layer. Such abnormalities can be found in normal individuals. These abnormalities may cause cortical excitability in generalized epilepsy and localized temporal lobe cortical dysgenesis might lead to dyslexia. Cortical dysgenesis has been reported in autism, schizophrenia, and fetal alcohol syndrome. The extent to which these findings explain abnormal brain function is an area of active research.

Focal cortical dysplasia

It is now known that cortical dysplasias are an important cause of early-onset seizures that may be focal or generalized (Fig. 4(a), Fig. 4(b) and Fig. 4(c)). Patients with refractory epilepsy should therefore have the best possible imaging, even if they have generalized seizures. Resection of cortical dysplasias may improve seizure control, but this is a challenging field because the preoperative assessment must take into account the widespread anatomical and functional abnormalities associated with cortical dysplasias.

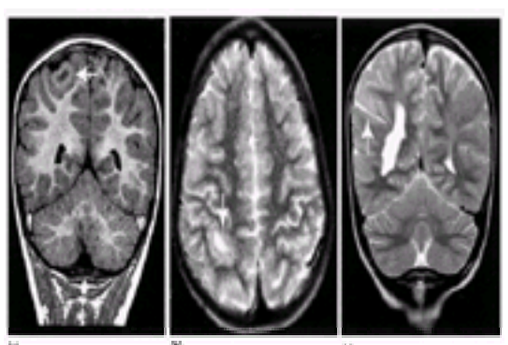

Fig. 4 (a) Cortical dysplasia in a boy aged 4 years. Focal seizures started at 1 year of age and consisted of a giggle, flexion of the left arm, and a vacant stare. T_1-weighted coronal MRI shows cortical dysplasia in the right parietal region (arrow). (b) The same patient. On T_2-weighted axial MRI, the cortical dysplasia is marked with an arrow. (c) Cortical dysplasia in a boy aged 3 years. Seizures commenced at 9 months of age and consisted of daytime absences and nocturnal generalized tonic–clonic seizures. A T_2-weighted coronal MRI shows an abnormal fissure in the cortex on the right (arrow). The right lateral ventricle is abnormal in size and shape.

Combined and overlapping cerebral malformations

There is overlap between the different malformations even though they are often described as distinct entities. This is not surprising—the teratogenic periods are so closely spaced that overlaps are likely if there is an environmental cause. Also, in genetically determined syndromes, more than one developmental process may be affected giving predictable combinations of cerebral malformations.

Agenesis of the corpus callosum

The true prevalence of this abnormality is not accurately known because it can be present without any symptoms. Estimated prevalence has varied from 0.05 to 70 per 10 000 in the general population, increasing to 230 per 10 000 in children with developmental disabilities.

Embryology

The corpus callosum forms within the commissural plate, a thickening of the lamina terminalis which is the frontal boundary of the neural tube. At 11 to 12 weeks the first fibres cross the midline to form the corpus callosum, which displaces the fornix and extends back in the occipital direction to assume the adult form by 18 to 20 weeks. There are two types of 'true' callosal agenesis. These are: (i) defects in which axons are unable to cross the midline and become large aberrant longitudinal fibre bundles, called Probst bundles, along the medial walls of the cerebral hemispheres; and (ii) defects in which the commissural axons or their parent cell bodies fail to form in the cerebral cortex. The former is probably the most common type, the latter is seen in the Walker–Warburg syndrome and other types of lissencephaly. There are also two secondary types: (i) absence associated with major malformations of the embryonic forebrain, such as holoprosencephaly; and (ii) degeneration or atrophy, as is seen in some syndromes in which the corpus callosum is thinned but not shortened.

Pathology

Agenesis of the corpus callosum may be complete or partial. Either the anterior or the posterior part may be missing. When there is complete absence there is no cingulate gyrus. There is associated enlargement of the occipital horns of the lateral ventricles, known as colpocephaly. Other associated abnormalities include cysts dorsal to the third ventricle, heterotopias, gyral abnormalities, cephaloceles, lipomas of the corpus callosum, eye abnormalities, nd hydrocephalus.

Aetiology

Isolated callosal agenesis may be inherited as an autosomal recessive, autosomal dominant, or X-linked recessive trait, but none of these loci have been mapped and non-syndromic genetic transmission is rare. It has been associated with several chromosomal rearrangements. These include trisomy 18, trisomy 13, and many deletions and duplications. Also it has been reported in more than 20 autosomal and many X-linked malformation syndromes. Callosal agenesis is part of the fetal alcohol syndrome and is seen in metabolic disorders including glutaric aciduria type 2, peroxisomal disorders, pyruvate dehydrogenase deficiency, and non-ketotic hyperglycinaemia.

Clinical findings

Non-syndromic forms are the most common. When agenesis of the corpus callosum is the only lesion there may be no symptoms, although tests of perception and language may demonstrate disturbances of integration of hemispheric function. However, even if there is no clearly defined syndrome, some patients have mental retardation, seizures, or cerebral palsy. Sometimes there is macrocephaly, which may be due to cysts lying dorsal to the third ventricle or to hydrocephalus.

Diagnosis

This depends on brain imaging (Fig. 5(a), Fig. 5(b) and Fig. 5(c)). The abnormalities that can be found are widely spaced parallel lateral ventricles, colpocephaly (enlarged posterior horns of the lateral ventricles), upward displacement of the third ventricle, absent callosal tissue, or midline dorsal cyst. Prenatal ultrasound allows diagnosis from the 20th week of gestation. After birth, MRI is best because it gives sagittal views of the corpus callosum. The scan should be carefully reviewed for other midline anomalies (such as agenesis of the septum pellucidum) or generalized defects (such as lissencephaly).

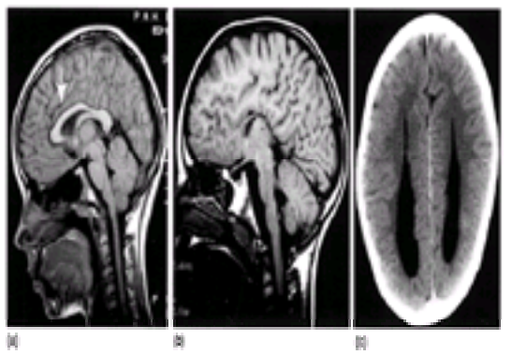

Fig. 5 (a) Normal brain in a girl aged 2 years. A T_1-weighted sagittal MRI shows normal corpus callosum and cingulate gyrus (arrow). (b) Agenesis of the corpus callosum in a girl aged 6 years who has microcephaly and moderate learning difficulties. A T_1-weighted sagittal MRI shows absence of the corpus callosum and of the cingulate gyrus, which normally runs parallel to the corpus callosum. (c) Agenesis of the corpus callosum in the same girl as (b). Axial CT shows typical appearance of parallel lateral cerebral ventricles, with divergence of the anterior horns of the ventricles and colpocephaly (dilated posterior part of the lateral ventricles).

The eyes should be examined to look for optic nerve hypoplasia (as seen in septo-optic dysplasia) or choroidal lacunae (as seen in the Aicardi syndrome). In neonates with seizures or other significant neurological problems the cerebrospinal fluid should be taken to measure glycine (raised in non-ketotic hyperglycinaemia) and lactate (raised in mitochondrial encephalomyelopathy). A karyotype should be performed and urine sent to measure amino and organic acids. If there is evidence of septo-optic dysplasia (see below), pituitary function should be checked.

Genetic counselling

This depends on the specific diagnosis. When callosal agenesis is discovered on antenatal scan the prognosis is difficult to assess because the isolated lesion can be associated with normal development. A decision to terminate the pregnancy may depend on the demonstration of associated abnormalities.

Porencephaly

The term porencephaly is often used indiscriminately for all large cavities in the brains of infants. Friede recommends using the term only for circumscribed hemispheric necrosis that occurs *in utero* before the adult features of the hemisphere are fully developed (Fig. 6(b)). The developmental origin of such lesions is shown by their smooth walls and from disturbances in the development of the adjoining cortex. These disturbances may take the form of polymicrogyria or of local distortion of the gyral pattern. In contrast, areas of damage resulting from insults in the terminal phase of the pregnancy or in postnatal life have irregular shaggy walls and do not alter the gyral environment except by atrophy or scarring.

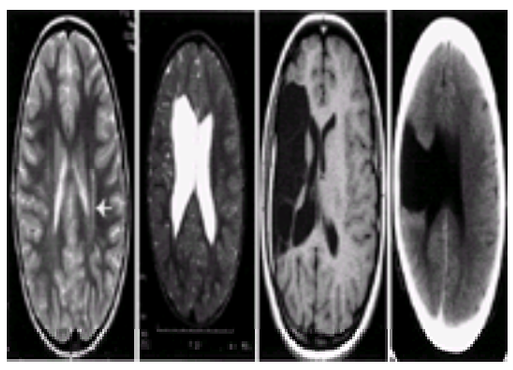

Fig. 6 (a) Cerebral palsy: spastic diplegia. Probable periventricular leucomalacia in a girl aged 6 years. There was threatened premature labour at 29 and 32 weeks, but she was born at term with no perinatal problems. She walked late with a diplegic gait. T_2-weighted axial MRI shows abnormal signal change lateral to the body of the left lateral ventricle and posterolateral to the posterior horn of the left lateral ventricle (arrow). This a characteristic distribution of periventricular leucomalacia, but such scan appearances should be interpreted with caution because there are other causes of white matter abnormalities in children (e.g. the leucodystrophies). (b) Cerebral palsy: left hemiplegia. Porencephalic cyst in a boy aged 18 months. He was delivered by forceps at 38 weeks with no resuscitation, but nasogastric feeding for several days after birth. At 10 months of age he was not moving the left arm normally. T_2-weighted axial MRI shows there is dilatation of the anterior horn of the right lateral ventricle with loss of overlying cerebral cortex and a small periventricular cyst adjacent to the anterior horn of the right lateral ventricle. These abnormalities may result from periventricular leucomalacia. Such loss of tissue due to *in utero* damage of the developing brain is called a porencephalic cyst. (c) Cerebral palsy: left hemiplegia. Tissue loss in middle cerebral artery territory in a young woman aged 17 years. There were no perinatal problems; reduced movement of left arm from 6 months of age; nocturnal generalized tonic–clonic seizures from 4 years of age; normal intelligence; abnormal posture of left hand; and shortening of the left leg. T_1-weighted axial MRI shows there is a loculated cystic lesion in the distribution of the supply of the middle cerebral artery. Also *ex vacuo* enlargement of the right lateral ventricle and small ipsilateral left hemicranium. (d) Open-lipped schizencephaly in a 49-year-old female. An axial CT scan shows there is a wide cleft joining the right lateral ventricle to the subarachnoid space.

Schizencephaly

This term is used to describe clefts which traverse the full thickness of the hemisphere, connecting the ventricle to the subarachnoid space. They are described as type I or 'fused-lip' when the walls of the cleft are opposed, and type II or 'open-lip' when cerebrospinal fluid separates the walls (Fig. 6(d)). Some authors think that the clefts are usually the result of destruction of brain tissue and the term porencephaly should be used for them all. However, there is now evidence that some of them are genetic—familial and sporadic cases have been recognized in association with mutations in the homeobox gene *EMX2*. This is one of the vertebrate homeobox genes expressed in the extended regions of the developing rostral brain of mouse embryos that are thought to play a role in patterning the forebrain. The clefts are frequently bilateral and symmetrical, the most severe form being large bilateral defects. Even when unilateral, they are often combined with cortical dysplasia of the opposite hemisphere.

Clinical features

These are variable, depending on the site and size of the lesion. Epilepsy is common and sometimes the only problem is isolated partial seizures.

There may be hemiplegia, quadriplegia, and learning difficulties of variable degree. If there is bilateral involvement of both opercular regions, there may be facial apraxia and speech difficulties. The diagnosis is best made by MRI.

Hydranencephaly

In this condition the cerebral hemispheres are mostly replaced by fluid-filled sacs. The defect typically corresponds to the territory of the anterior and middle cerebral arteries, although the major cranial arteries do not usually show evidence of obstruction. Preservation of the temporal lobes and the tentorial parts of the occipital lobes is common. The extent of preservation of the basal ganglia varies. The cause of hydranencephaly is not clear in many cases. It has been described after intoxication of pregnant women with gas at about the 25th week of gestation. It can result from intrauterine infections and has been described in association with a proliferative vasculopathy.

Affected infants may be born after a normal pregnancy and be surprisingly normal on neurological examination for the first few weeks of life. Gradually the infants become hypertonic and irritable. They may develop infantile spasms, which is surprising because of the almost complete lack of cerebral hemispheres. The head may enlarge because of associated hydrocephalus. The diagnosis can be made by transillumination of the skull, which lights up like a lantern in a darkened room. Similar appearances can be caused by hydrocephalus with a very thin cortical mantle, so MRI is indicated to confirm the diagnosis. Infants with hydranencephaly often die in a few months, but they may survive for several years and may need a ventriculoperitoneal cerebrospinal fluid shunt if there is progressive hydrocephalus.

Septo-optic dysplasia

This is the association of optic nerve hypoplasia with absence of the septum pellucidum. Disturbances of the hypothalamopituitary axis may occur. The most severely affected patients are blind and have severe learning difficulties. The optic discs have a characteristic double contour: the true disc at the centre is small and there is a peripheral ring about the size of a normal optic nerve head. It important to search for evidence of endocrine disturbances when these abnormal discs are identified—deficiences of growth hormone, corticotrophin, luteinizing hormone, and follicle-stimulating hormone have been described, together with hypoglycaemia and diabetes insipidus. Most cases are sporadic, but a homozygous mutation in the homeobox gene *HESX1* has been identified in familial septo-optic dysplasia. Also some sporadic cases of the more common mild forms of pituitary hypoplasia are associated with heterozygous mutations of the *HESX1* gene.

Malformations of posterior fossa structures

These malformations are now identified more accurately using prenatal ultrasound and MRI, which is superior to CT scanning for showing posterior fossa structures.

Cerebellar and brainstem development

Embrology

At the end of the fourth week of gestation the neural tube divides into the three primary brain vesicles—the prosencephalon, the mesencephalon, and the rhombencephalon. The last further subdivides into the metencephalon and the myelencephalon. The cerebellar hemispheres (neocerebellum) are derived primarily from the metencephalon, while the vermis (palaeocerebellum) is derived from the mesencephalon.

Experiments in animals demonstrate that a critical area for vertebrate cerebellar development occurs at the junction of the mesencephalon and the metencephalon in the region of the isthmus. An early abnormality in the isthmus that affects the mesencephalic contribution to midline cerebellar structures could contribute to the pathogenesis of syndromes that involve agenesis or hypoplasia of the cerebellar vermis, such as Joubert syndrome.

Aplasia and hypoplasia of the cerebellum

This is a heterogeneous group of conditions that affect cerebellar development in various ways—total cerebellar aplasia is exceptional and unilateral hypoplasia occurs. Neocerebellar aplasia (Fig. 7(a)) is characterized by a small vermis and extreme smallness or absence of the cerebellar hemispheres except for persistent flocculi. There may be associated anomalies in the brainstem such as dysplasia of the inferior olivary nucleus and other brainstem nuclei.

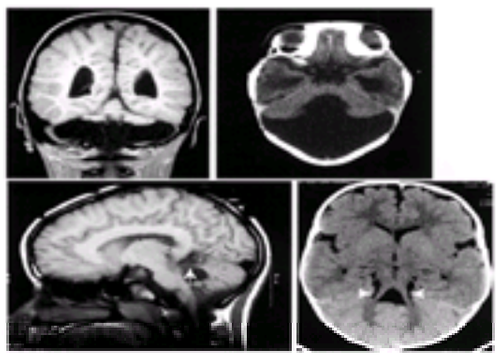

Fig. 7 (a) Cerebellar hypoplasia in a boy aged 5 years, who was born preterm at 26 weeks of gestation, with no neurological problems apart from absence seizures of unknown cause. T_1-weighted coronal MRI shows almost complete absence of the cerebellar hemispheres and hypoplasia of the cerebellar vermis. (b) Dandy–Walker malformation in a 1-year-old girl. Axial CT scan shows absence of the roof of the fourth ventricle. A large cyst is continuous with the fourth ventricle and fills the posterior fossa. (c) Joubert syndrome in a girl aged 10 years. T_1-weighted sagittal MRI shows the superior cerebellar peduncles run horizontally (arrow) and the cerebellar vermis is absent. (d) Joubert syndrome in a girl aged 9 months who is hypotonic and visually unresponsive, with 'wandering' nystagmus. Axial CT shows the superior cerebellar peduncles (arrows) run horizontally and stand out because of the absence of the vermis ('molar tooth sign'). The prominent fourth ventricle has a typical shape (sometimes looking like a 'bat wing').

Some cases are associated with genetic syndromes, many of which are poorly defined. Recent attention has been drawn to a group of disorders classified under the broad headings of pontocerebellar hypoplasia or olivopontocerebellar atrophy. Pontocerebellar hypoplasia is found in carbohydrate-deficient glycoprotein syndrome type 1 and cerebromuscular dystrophies (Walker-Warburg syndrome, Fukuyama syndrome, and muscle-eye-brain diseases—see above). There are at least two types of autosomal recessive pontocerebellar hypoplasias (type I and type II). MRI demonstrates cerebellar hypoplasia together with a hypoplastic ventral pons. It has been speculated that some of these cases are due to mutations of *engrailed* genes which are essential for the development of the segmental precursors of the pons and cerebellum in the mouse brain.

The Chiari malformations

There are four types. The Chiari II malformation is usually associated with a meningomyelocele and is dealt with in the section on neural tube defects. In the Chiari I malformation there is downward displacement of the lower cerebellum, including the tonsils. It rarely cause symptoms in childhood but may be associated with hydrocephalus and syringomyelia. The Chiari III malformation consists of downward displacement of the cerebellum into a posterior encephalocele and the Chiari IV malformation is a form of cerebellar hypoplasia.

Abnormalities of the vermis

Dandy–Walker malformation and Dandy–Walker variant

Two main complexes have been described. The Dandy–Walker malformation (Fig. 7(b)) consists of the following triad: (i) complete or partial agenesis of the vermis, (ii) cystic dilatation of the fourth ventricle, and (iii) enlarged posterior fossa with upward displacement of lateral sinuses, tentorium, and torcula. A Dandy–Walker variant has been recognized more recently and consists of variable dysplasia of the vermis without enlargement of the posterior fossa. There is an association between Dandy–Walker malformations and chromosomal abnormalities including trisomies 13 and 18. Dandy–Walker malformation is listed as a feature of 80 syndromes in the London Dysmorphology Database.

Prenatal ultrasound studies show that the majority of fetuses with both complexes have other anomalies. The commonest CNS anomaly was ventriculomegaly and the commonest non-CNS anomalies were structural heart defects. Other associated CNS abnormalities include holoprosencephaly, agenesis of the corpus callosum, occipital encephaloceles, and abnormal migration of the inferior olive.

Clinical features

The outcome of fetuses with both Dandy–Walker malformation and Dandy–Walker variant ranges from severe mental and physical handicap to normal development. In general the outcome is worst if there are associated abnormalities and best for isolated Dandy–Walker variant. Intrauterine deaths occur and some infants do not survive the neonatal period, the poor early outcome being related to the extra-CNS abnormalities. However, the abnormality is often recognized only when the infant is investigated for the signs of hydrocephalus, which may not become apparent until late in the first year of life. Sometimes there is a bulging occiput which alerts suspicion. Another presentation may be later in life with learning difficulties. Cerebellar signs tend not to be prominent, but cranial nerve palsies, nystagmus, and truncal ataxia have been described.

Diagnosis

The cerebellar vermis starts to form in the ninth week of gestation, beginning superiorly so that fusion of the two cerebellar hemispheres is completed when the inferior part of the vermis is formed at 15 weeks. However, in a small proportion of infants this happens later, so that conclusive prenatal ultrasound diagnosis cannot be made until 18 weeks. When the diagnosis is made there should be an exhaustive screen for associated structural and chromosomal anomalies because these determine the prognosis. Radiological diagnosis is relatively straightforward for the Dandy–Walker malformation. Dandy–Walker variant may be difficult to distinguish from a prominent cisterna magna or a retrocerebellar arachnoid cyst. Sagittal MRI can then help to differentiate whether the vermis is partly absent or alternatively is fully present but displaced.

Genetic counselling

If a chromosomal anomaly is identified or a syndrome diagnosis is made, appropriate counselling is given. If there are no associated abnormalities and the chromosomes are normal, the recurrence risk is 1 to 5 per cent.

Joubert syndrome

This rare autosomal recessive disorder is characterized by brainstem and cerebellar malformations. The disease is genetically heterogeneous, and one locus has been mapped to chromosome 9q.

Neuropathology

The neuropathological features include absence or hypoplasia of the posteroinferior part of the cerebellar vermis. In some cases enlargement of the fourth ventricle and the cisterna magna has been reported. Microscopically, heterotopias have been seen in the cerebellar hemispheres with fragmentation of the dentate nuclei. Brainstem abnormalities include absence of the pyramidal decussation, abnormal inferior olivary nuclei, and subtle dysplasias in the nuclei of the solitary and descending trigeminal tracts and of the dorsal columns.

Clinical features

The common abnormalities are marked hypotonia (particularly in the neonatal period and infancy), poor balance (walking occurs in 50 per cent of cases and is late—at approximately 4 years), and variable cognitive problems (some affected children are unable to talk but others develop language, read, and write). Typically CT or MRI shows the 'molar tooth' sign in the axial plane, which consists of: (i) a deeper than normal posterior interpeduncular fossa, (ii) prominent or thickened

superior cerebellar peduncles, and (iii) vermian hypoplasia or dysplasia. MRI in the coronal and axial plane shows clefting of the vermis; in the sagittal plane it shows an abnormally shaped and rostrally placed fourth ventricle (Fig. 7(c) and Fig. 7(d).

The associated abnormalities are dysmorphic facial features (high rounded eyebrows, broad nasal bridge, epicanthus, anteverted nostrils, triangular-shaped open mouth with irregular tongue protrusion, low-set coarse ears), episodic hyperpnoea and/or apnoea in up to 75 per cent of patients (most marked in the neonatal period), eye abnormalities (retinal dysplasia, colobomata, nystagmus, strabismus, ptosis, and oculomotor apraxia), and microcystic renal disease.

Neurological syndromes resulting from disturbances of brain development

The cerebral palsies

Introduction

The cerebral palsies are defined as a heterogeneous collection of non-progressive disorders of movement and posture due to defects or lesions of the immature brain.

It can be seen that this is a broad definition, but it does need to be used carefully. It is not satisfactory to label a child with neurological problems as suffering from 'cerebral palsy' and go no further. It is important to determine the type and distribution of the abnormality of motor control. In addition to motor deficits, patients with cerebral palsy may suffer with other neurological problems, such as learning difficulties, epilepsy, and hearing or visual loss. It is important to evaluate these, although they are not present in all cases.

Although the underlying causes of the cerebral palsy syndromes are by definition not progressive, the symptoms and signs of cerebral palsy do change with age. For instance some children who are destined to have major problems with spasticity are initially very hypotonic. In some cases it can be difficult to be sure whether or not a child with suspected cerebral palsy has a progressive underlying disorder. It may be necessary to allow the passage of time and children may be 3 or 4 years old before the diagnosis of cerebral palsy can be made with confidence.

Classification

Patients may be classified according to the type of motor abnormality as follows: spastic, dyskinetic (dystonic or athetoid), ataxic, or hypotonic. The clinical picture is rarely clear-cut and individuals may exhibit complex mixtures of motor disability.

Patients are subclassified according to the distribution of motor abnormality—in diplegia the legs are involved more than the arms, in quadriplegia all four limbs are involved, and in hemiplegia just one side of the body is involved.

In practice there is considerable interobserver disagreement when classifying patients with cerebral palsy. However, it is important to use a classification system for research and management because different types of cerebral palsy tend to have distinct causes, different associated deficits, and different prognoses.

Epidemiology

There is a relative shortage of data from developing countries, so the statistics quoted here are for developed countries. Although the risk of cerebral palsy is higher for preterm infants, most children with cerebral palsy are born at term. Overall cerebral palsy rates are therefore determined mainly by the numbers of term infants born with cerebral palsy.

Overall cerebral palsy rates since the mid-1950s have remained remarkably constant at about 2 to 2.5 per 1000 live births, although there have been some fluctuations with time. In 1970 the rate fell to 1.5 in Sweden, Western Australia, and Mersey (United Kingdom), rising again in the 1980s.

Cerebral palsy rates stratified by birth weight show marked changes with time. Most population-based registers have shown increases in rates in infants of very low birth weight (less than 1500 g) since the 1970s. For instance, in Mersey in the early 1970s the rate in infants of very low birth weight fluctuated around 10 per 1000 live births. In the late 1970s the rate increased sharply to about 50 per 1000 live births, presumably because more children of very low birth weight were surviving with neurological deficits. This increase was seen for all cerebral palsy types. However, to put the increasing cerebral palsy rates in survivors of very low birth weight in perspective, during this time an increasing proportion of patients were also surviving unimpaired. Another important point is that very low birth weight survivors with cerebral palsy are only a small proportion of the total number of children with cerebral palsy and therefore have very little effect on overall cerebral palsy rates.

Relative rates of the different types of cerebral palsy vary according to the series. In the Western Australian Cerebral Palsy Register 1980–1992 cohort the proportions of cases of congenital (as opposed to postneonatal) cerebral palsy were as follows: spastic hemiplegia 33.6 per cent, spastic diplegia 29.7 per cent, spastic quadriplegia 18.1 per cent, ataxic 7.6 per cent, dyskinetic 10.0 per cent, and hypotonic 0.9 per cent.

Aetiology

Although it has been thought that cerebral palsy results primarily from 'birth asphyxia', recent studies suggest that abnormal events around the time of birth play only a limited role. Genetic causes are clearly important as there can be a significant recurrence risk to future children, particularly in populations where consanguineous marriage is relatively common. Families have been reported in which spastic diplegia and quadriplegia (often with associated mental retardation) appear to be inherited in autosomal recessive, autosomal dominant, or X-linked recessive patterns. It is said that the highest risk of recurrence is in the category of children with ataxic cerebral palsy: both autosomal dominant and autosomal recessive inheritance have been reported. However, there are many conditions which cause ataxia in children. It is therefore important to search for an underlying cause before giving genetic advice, rather than to 'lump' this group and give an overall recurrence risk.

Other possible causes before conception or in early pregnancy

Maternal iodine deficiency in early pregnancy may cause endemic cretinism (which causes spastic diplegia and deafness): this is the most important cause of cerebral palsy worldwide. Abnormal thyroid function in pregnancy may play a role in developed countries. Exposure to toxins during pregnancy may cause cerebral palsy—recognized examples are methylmercury, alcohol, and carbon monoxide poisoning. Viral infections which are vertically transmitted before the third trimester often result in cerebral malformations—those best known are toxoplasmosis, rubella, and cytomegalovirus. Finally, there are some fetal malformation syndromes that include brain abnormalities and cause cerebral palsy.

The role of very preterm birth

The rate of cerebral palsy among neonatal survivors born before 33 weeks is up to 30 times higher than among those born at term. This may be because of increased survival of very preterm infants whose brains are already damaged or because preterm infants are vulnerable to cerebral damage after birth. It seems likely that there is a combination of these mechanisms. Cerebral ultrasound scans performed in newborn babies have shown that the strongest predictor of cerebral palsy in these infants is periventricular leucomalacia. This term is used for abnormal echolucency, often associated with cystic change, which is found particularly in the white matter dorsolateral to the lateral ventricles (Fig. 6(a)). It is difficult to time the onset of the pathological processes that lead to the appearance of these lesions. At present there is evidence that many different factors result in preterm birth and it is not known how they contribute to cerebral palsy.

The role of intrauterine growth restriction

Babies born small for their gestational age are at increased risk of cerebral palsy and the risk increases with the degree of birth weight deficit. The underlying mechanism is not clear and the majority of small-for-dates infants do not have cerebral palsy.

The role of multiple births

The prevalence of cerebral palsy is much higher in twins than in singletons, particularly in those who survive after the other twin has died *in utero* and in monochorionic twins. The risk rises with the number of fetuses carried. Causes include low birth weight, congenital anomalies, cord entanglement, and abnormal vascular connections.

The role of birth asphyxia

In the 1970s it was expected that more intensive monitoring of the fetus during labour, coupled with earlier obstetric intervention, would improve neonatal outcome. The major effects of electronic monitoring of the heart rate during labour have been an increase in caesarean section rates and a reduced rate of neonatal seizures, however it has had no impact on the rates of cerebral palsy. The proportion of cerebral palsy cases associated with intrapartum events has been estimated by several epidemiological studies to be only about 10 per cent. However, there may be some cases caused by intrapartum events that are preventable.

In 1999 a consensus statement for the International Cerebral Palsy Task Force outlined a template for defining a causal relationship between acute intrapartum events and cerebral palsy. The statement emphasized the difficulty of retrospectively identifying the antenatal causes of cerebral palsy in the individual case and the non-specific nature of the clinical signs that lead to the suspicion of fetal hypoxia in labour. It proposed that the terms 'fetal distress' and 'birth asphyxia' should be replaced by the term 'non-reassuring fetal status' and suggested eight criteria for defining an acute intrapartum hypoxic event. The hope is that more general use of these criteria will reduce the number of cases of cerebral palsy that are wrongly attributed to an acute event during labour.

Cerebral palsy acquired after the neonatal period

Cerebral palsy registers in Sweden, Mersey, and Western Australia report rates varying from about 1 to 6 per 10 000 live births. Causes include CNS infections, accidental and non-accidental head injuries, cerebrovascular accidents, and hypoxia (suffocation, near drowning).

Brain imaging in children with cerebral palsy

The pathological processes that cause cerebral palsy have been investigated using ultrasonography, MRI, and CT scanning. One study found that about a quarter of children with a hemiplegia had normal CT scans, whilst an MRI study of a heterogeneous group of children with cerebral palsy found abnormalities in 93 per cent of the patients.

The studies have in many cases been performed years after birth on heterogeneous groups of children with cerebral palsy attending specialized clinics. The commonest lesion in preterm infants is periventricular leucomalacia, which is necrosis of periventricular white matter in the watershed regions dorsal and lateral to the lateral ventricle (Fig. 6(a)). This is said to be characteristic of damage in the early third trimester. In term infants there are a number of different findings, which are said to occur only in infants born at or near term. These are infarcts in the arterial border zones in the parasagittal regions leading to cortical and subcortical injury, bilateral lesions of the basal ganglia and the thalamus, areas of subcortical leucomalacia, and multicystic leucomalacia (replacement of the brain tissue by fluid-filled cysts). Periventricular leucomalacia and localized gyral abnormalities are also seen in infants born at term.

Children with hemiplegias are sometimes found to have periventricular leucomalacia, porencephalic cysts (Fig. 6(b)), or cortical/subcortical lesions in the middle cerebral artery territory distribution (Fig. 6(c)). The lesions tend to be unilateral, but bilateral lesions are seen. Rarely they may have schizencephaly (Fig. 6(d)), focal pachygyria, or focal heterotopia.

Although brain scans in children with cerebral palsy are performed a long time after the insult, the nature of the lesions allows some assessment of the timing of cerebral damage. Like the epidemiological studies the brain imaging studies suggest that perinatal brain injury occurs in a relatively small proportion of cases.

Life expectancy

Different studies have followed patients with cerebral palsy for 10 years or more and have yielded similar findings. The survival rates have been about 90 per cent when all the types of cerebral palsy are considered together. The prognosis is best for those with hemiplegia and worst for those with quadriplegia. A population-based study in Canada found that 30-year survival rates were: hemiplegia 96 per cent, diplegia 95 per cent, and quadriplegia 83 per cent. The factors associated with reduced survival rates are severe mental retardation, lack of basic functional skills (mobility, independent feeding), and epilepsy.

Hydrocephalus

This results from expansion of the ventricles secondary to a block in the normal flow pathway of cerebrospinal fluid. Cerebrospinal fluid is produced by the choroid plexus in the lateral ventricles, from where it flows through the foramen of Munro into the third ventricle and then the fourth ventricle via the aqueduct of Sylvius. It leaves the ventricular system via small openings in the roof of the fourth ventricle, the foramina of Magendie and Luschka. From here the fluid flows in the subarachnoid space before being reabsorbed into the blood supply via arachnoid villae.

Two major forms of hydrocephalus are recognized. In communicating hydrocephalus the ventricular pathways are clear and a failure of reabsorption (following, for example, bleeding into the subarachnoid space) results in increased cerebrospinal fluid volume. In obstructive or non-communicating hydrocephalus the blockage occurs at one of the ventricular levels, with expansion of the ventricular system above the block (Fig. 8). The major clinical sign that results is increasing head circumference following the ventricular enlargement, and this allows the distinction from cases in which increased ventricular size reflects cerebral atrophy. Mental retardation can result from both the damage associated with ventricular expansion and other abnormalities associated with the underlying cause of the problem.

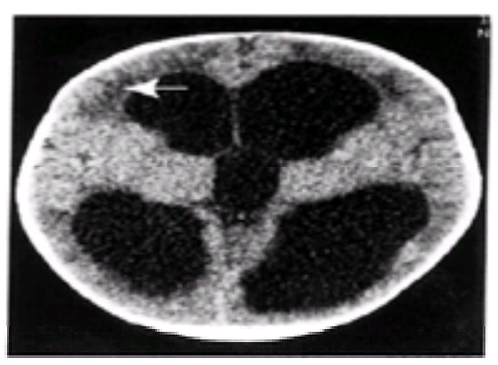

Fig. 8 Aqueduct stenosis in a boy aged 1 month with a bulging anterior fontanelle and increasing head circumference. Axial CT shows a gross dilatation of the third and lateral ventricles (the fourth ventricle is not shown, but was normal in size). Note the periventricular low density due to transependymal exudation of cerebrospinal fluid under pressure (arrow).

Sometimes hydrocephalus is genetically determined; stenosis of the aqueduct between the third and fourth ventricle can result from mutations in the cell adhesion molecule L1-CAM. Hydrocephalus then occurs in association with hypoplasia of the corpus callosum, mental retardation, spastic paraplegia, and adducted thumbs. This X-linked syndrome has been given the extremely unfortunate acronym CRASH syndrome and mutations in *L1-CAM* are found in as many as 75 per cent of cases with a family history and 15 per cent of apparently isolated cases. The developmental abnormalities of the cerebellum in both the Dandy–Walker syndrome and the Arnold–Chiari malformation (see above) may also be associated with obstructive hydrocephalus. While treatment via a ventriculoperitoneal shunt can relieve the obstruction, the other abnormalities associated with these developmental problems remain.

Effects of alcohol on the developing nervous system

Worldwide, alcohol is one of the commonest causes of learning difficulty and neurobehavioural disturbance in young children. The incidence of fetal alcohol sydrome depends on geographical location. An international survey in 1997 found that in the United States the incidence per 1000 live births in Seattle was 2.8 and in Cleveland was 4.6. The combined rate of fetal alcohol sydrome and alcohol-related neurodevelopmental disorder in Seattle was estimated at nearly 1 per cent of all live births. The reduced brain mass and neurobehavioural disturbances associated with fetal alcohol syndrome may be related to the recent observation in rats that ethanol can trigger widespread apoptotic neurodegeneration.

Regular and binge drinking can both cause fetal alcohol syndrome and alcohol-related neurodevelopmental disorder. Unlike many other teratogens, alcohol has harmful effects throughout pregnancy. There is a significant risk of fetal alcohol syndrome associated with high-dose exposure (estimated blood alcohol concentrations of 150 mg per decilitre or more, at least weekly for several weeks in the first trimester).

In addition to microcephaly, structural anomalies of the brain such as partial or complete agenesis of the corpus callosum or cerebellar hypoplasia may occur. Children with fetal alcohol syndrome may have impaired fine motor skills, sensorineural deafness, poor hand–eye co-ordination and a poor tandem gait. A complex pattern of behavioural and cognitive abnormalities is observed following exposure to teratogenic levels of alcohol in pregnancy. These include learning difficulties, poor impulse control, problems with social perception, deficits in higher level receptive and expressive language, poor capacity for abstraction, and difficulties with memory, attention, and judgement.

Congenital infections

Cytomegalovirus, herpes simplex, parvovirus, rubella, syphilis, toxoplasmosis, and varicella are all recognized as teratogens. Primary infection rather than reinfection of the mother during pregnancy is more likely to result in congenital infection. The risk of congenital infection and the outcome of such infection is crucially dependent on the stage of pregnancy. This brief account focuses on the effects of congenital infection on the developing nervous system.

Congenital infection should be considered in the differential diagnosis of microcephaly. Intracranial calcification identified on a cranial ultrasound or CT scan during the investigation of developmental delay or seizures should arouse suspicion of congenital infection, especially cytomegalovirus or toxoplasmosis (calcification is not picked up well by MRI). Detailed ophthalmological assessment may reveal clues such as chorioretinitis or cataract that may help in the retrospective diagnosis of congenital infection. Chorioretinitis (pigmentary retinopathy) is characteristic of intrauterine infection by cytomegalovirus or toxoplasmosis. Audiometry is important since sensorineural deafness is a common sequel to congenital infection with cytomegalovirus, rubella, and toxoplasmosis.

The risk of maternal–fetal transmission with primary cytomegalovirus infection is as high as 40 per cent, however fewer than 10 per cent of infants with intrauterine infection are symptomatic at birth. Of those who are symptomatic as neonates, approximately 90 per cent have some of the characteristic features including microcephaly, periventricular calcification, chorioretinitis, optic atrophy, and sensorineural deafness. Of the 90 per cent of infants who are asymptomatic at birth, approximately 15 per cent have sequelae including sensorineural deafness and/or developmental delay.

Intrauterine infection with herpes simplex virus is rare. Congenitally affected infants may have microcephaly, chorioretinits, and microphthalmos. Neonatal infection acquired at the time of delivery, which may occur with primary infection of the mother with herpes simplex virus in the third rimester, is a commoner cause of neurodisability than congenital infection. Neonatal infection may cause meningitis and encephalitis with resulting neurological damage. The risks of perinatally acquired infection may be reduced by appropriate obstetric intervention (such as delivery by caesarian section for women with active genital lesions resulting from herpes simplex virus) and by treatment of affected neonates with acyclovir.

The classic triad of defects associated with congenital rubella syndrome is sensorineural deafness, congenital heart disease, and eye abnormalities (retinopathy, cataracts, microphthalmos, and congenital glaucoma). Microcephaly and developmental delay may also occur. The spectrum of defects in an individual child is determined by the stage of pregnancy at which intrauterine infection occurs. The risk of congenital infection is more than 90 per cent below 10 weeks and falls to zero beyond 18 weeks.

The risk of intrauterine infection with toxoplasmosis increases with the stage of pregnancy at which the mother acquires her primary infection; however the sequelae of intrauterine infection diminish with advancing gestation. Congenital toxoplasmosis syndrome includes hydrocephalus, intracranial calcification, microcephaly, seizures, and developmental delay. There may also be sensorineural deafness and chorioretinitis with visual impairment.

Congenital varicella syndrome follows primary maternal varicella occurring at 1 to 20 weeks of gestation, but the risk of sequelae is small at around 2 per cent. Cataracts and chorioretinitis may occur together with hypoplasia of the optic disc. Microcephaly and porencephaly have been described.

Clinical approach to diagnosis and genetic counselling

Assessing the nervous system in children

History

General

The importance of the history cannot be overemphasized. Children may give a history themselves, but usually the parents or carers are an essential source of information and this may be amplified by teachers, therapists, and other health professionals.

Past history

The pregnancy details are important. Significant events in the first trimester may be a threatened miscarriage, hyperemesis, or a viral infection—the mother may have been taking medication. Later there may have been unsatisfactory fetal growth (perhaps poor head growth assessed by ultrasound) or reduced fetal movements. The perinatal history is relevant, including weeks of gestation at the time of delivery, details of labour and delivery, birth weight, and head circumference. In the neonatal period the infant may have required treatment for early hypoglycaemia, seizures, or breathing or feeding difficulties. A developmental history is essential—particular areas of concern in infants are lack of social response, absence of a social smile, poor fixing and following of the eyes, and lack of symmetrical organized limb movements. Later a characteristic pattern of delayed development may emerge—for instance global delay is found in the most severe brain abnormalities or there may be mainly motor delay in the milder forms of cerebral palsy.

Family and social history

Information should be obtained about first-degree and more distant relatives. Considerable effort may be needed to obtain relevant facts—some families conceal or do not know about relatives with severe disability, perhaps because they are in instituitions. It is important to know about consanguinity, also about epilepsy, motor disorders, and severe or mild learning disabilities. Social factors are important in determining the environment in which the child grows up and they also determine the quality of care available for a child with significant disability.

Examination

Observation of spontaneous activity is essential. The form of this depends on the degree of disability, but if the child is able to play this should be encouraged. It may be helpful to use toys, bricks, beads for threading, paper, and crayons. The quality and symmetry of spontaneous movements should be noted and also any abnormal movements. It is best to assess muscle power by watching the child run, jump, and climb stairs. Fine motor function can be assessed whilst the child is drawing or threading beads.

Developmental assessment may be formally undertaken in infants using one of the standardized schedules, such as the Bailey Scales of Infant Development or the

Denver Developmental Screening Test. Later the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) and the Wechsler Intelligence Scale for Children (Revised) (WISC-R) may be used. The latter assessments are usually undertaken by a clinical or educational psychologist.

The conventional examination of the nervous system may be difficult in infants or young children. The examiner may need to adapt the order of events or even come back later—a useful assessment cannot be made if an infant is deeply asleep or upset and crying. Examination of the cranial nerves should be made as much like play as possible by using a toy to observe eye movements and by encouraging the child to smile, whistle, close the jaw tight, stick out the tongue, and so on.

Dysmorphic features are particularly relevant in the context of a suspected abnormality of the nervous system. There may be birth marks (capillary haemangiomata of the face in Sturge–Weber syndrome or midline skin abnormalities such as hairy patches or dimples over the spine in cord abnormalities). Other important skin abnormalities may not be very obvious at birth but become so in infancy or early childhood. Examples are the pale ash-leaf patches, shagreen patches, and angiofibromas of the face ('adenoma sebaceum') that are found in tuberous sclerosis or the café-au-lait patches found in neurofibromatosis type I.

A full eye examination is essential. In babies it may be necessary to dilate the eyes and come back to perform fundoscopy whilst the child is feeding (and therefore quiet!). Indirect ophthalmoscopy by an experienced ophthalmologist is probably best for older children. There may be hypo- or hypertelorism which may be associated with midline defects of the brain (for example hypotelorism in holoprosencephaly and hypertelorism in agenesis of the corpus callosum), so the interpupillary distance and the distance between the inner canthi should be measured and checked on standard charts. There may be abnormalities of the iris (such as colobomata in trisomy 13, the CHARGE association, and other syndromes that involve the nervous system; Lisch nodules in neurofibromatosis type I; or Kayser–Fleischer rings in Wilson's disease). Pale hypoplastic optic nerve heads are seen in septo-optic dysplasia and other congenital and acquired conditions. Significant retinal abnormalities include the chorioretinitis seen in congenital toxoplasmosis or cytomegalovirus infections and the retinal 'lacunae' seen in Aicardi syndrome.

Growth should be assessed by measuring weight, length (height), and head circumference and plotting them on standard charts. In particular the head circumference should be related to the age of the child and to the other measurements (see sections on microcephaly and megalencephaly above). Changes with time may be significant—for instance, after a severe perinatal insult, the head circumference may initially be in the normal range and then fall progressively further below the expected centile line in the first few months of life, which can be important in dating the insult to the brain. Babies may be upset by having their heads measured, so this is best left to the end of the examination.

Investigations

The cornerstone of investigations in children or adults with suspected disorders of CNS development is MRI to investigate brain structure. It is important to discuss the investigation with a neuroradiologist as special imaging sequences not normally performed may be required to visualize relevant abnormalities, for example subependymal nodules in tuberose sclerosis. Infants and young children may require sedation or anaesthesia for the procedure. CT scanning does not provide the resolution of CNS structure obtained with MRI, but may be valuable if intracerebral calcification is suspected (as in tuberous sclerosis or cytomegalovirus infection).

Further investigations will depend on the specific diagnosis in question. Metabolic disorders can cause structural abnormalities in the developing CNS, and routine investigations that may be appropriate include plasma and urine amino acids, together with urine organic acids. In addition, further specific investigations may be indicated, for example in suspected Zellweger's syndrome which is associated with pachygyria and which is caused by abnormalities of very long chain fatty acid metabolism. Mutation analysis of specific genes may confirm a clinical diagnosis. Fluorescent *in situ* hybridization studies of chromosome regions using labelled probes that will bind (hybridize) to specific gene sequences may detect microdeletion syndromes such as Miller–Dieker by revealing an absence of fluorescent labelling on one of the pair of chromosomes.

These investigations may then allow diagnosis and accurate assessment of risks for other family members following extended family testing. Molecular genetic techniques are improving very rapidly and new tests will become available. It is therefore valuable to take blood in order to extract and store DNA or establish a lymphoblastoid cell line if no precise diagnosis can be reached, especially if life expectancy is short. Immediately after death it may be appropriate to obtain a muscle or liver biopsy to help establish a diagnosis. Also skin may be obtained to establish a fibroblast culture. Later other tissues can be frozen if a full postmortem examination is performed. The ability to perform new tests many years after the death of the index case may be extremely valuable to other family members concerned about risks to their own offspring.

Risk assessment, prenatal diagnosis, and genetic counselling

When families request genetic advice regarding a developmental disorder of the nervous system they usually have four questions in mind: what is it?, why did it happen?, will it happen again?, and what can be done to reduce the chance of it happening again, or to detect it if it does? If it is possible to make a specific diagnosis, these questions can often be answered with some accuracy.

Risk assessment

An important component of risk assessment is the construction of a three-generation family tree, with detailed enquiry and if necessary examination and investigation of close relatives for subtle expression of a disorder, or evidence of carrier status. Sometimes the diagnosis will immediately identify the recurrence risk. For example Zellweger syndrome always follows an autosomal recessive pattern of inheritance (hence there is a 1 in 4 risk of recurrence in future pregnancies). For many of the developmental anomalies discussed in this chapter, the causes are heterogeneous with a variety of different mechanisms resulting in similar clinical endpoints.

The assessment of holoprosencephaly provides an example of the steps involved in risk assessment. It may occur in an individual with a chromosomal anomaly such as trisomy 13, or a variety of subtle chromosome deletions such as del (18p), del (7q). It may also follow an autosomal dominant pattern of inheritance with incomplete penetrance and variable expression (mutations in the *sonic hedgehog* gene in some families), or be a feature of a recognizable syndrome such as Smith–Lemli–Opitz, which follows an autosomal recessive pattern of inheritance. Assessing the risk for a particular family depends upon careful integration of the clinical picture in the affected individual with information from the family tree and the results of investigations.

If it is not possible to identify the precise aetiology, and common causes have been excluded as far as possible by appropriate investigation, it is usually possible to offer an empirical recurrence risk after examination of the parents. For example, in a family where the child has holoprosencephaly with no additional features suggestive of a syndromic cause and who has normal chromosomes, the next step is a careful examination of both parents to look for the subtle features of autosomal dominant holoprosencephaly (such as single central incisor, hypotelorism). They should also both have a cranial MRI. If these assessments are normal, then the empirical recurrence risk is 5 to 6 per cent.

Prenatal diagnosis

Prenatal diagnosis and termination of affected pregnancies is only one of a range of reproductive options open to parents at increased risk of having children with neurodevelopmental abnormalities, but for many couples it is the option of choice. Other options include embarking on a further pregnancy and accepting the risk of recurrence, or electing against any further pregnancies and perhaps considering adoption. For the majority of developmental disorders of the nervous system, preimplantation genetic diagnosis is not yet feasible. For a condition following mendelian inheritance the option of donor gametes could be discussed. For a condition with a strong environmental component it is imperative that measures are taken to minimize the risk of exposure in a future pregnancy. For neural tube defects, periconceptual supplementation with high-dose folate has been shown to reduce the risk of recurrence in future pregnancies (see above).

When a specific diagnosis has been made and a chromosomal anomaly, genetic mutation, or biochemical defect has been identified, it is usually possible to offer prenatal diagnosis by chorionic villus sampling at 11 weeks of gestation in a future pregnancy. If this is not the case, detailed ultrasound scanning may be helpful in some instances; for example, neural tube defects where anencephaly can be clearly visualized by 13 to 14 weeks of gestation, and spina bifida by 18 to 20 weeks. For other conditions such as isolated lissencephaly, no features are likely to be visible on an ultrasound scan before 24 weeks of gestation, and for isolated microcephaly often not until 32 to 34 weeks of gestation or later. The limitations of detailed ultrasound scanning in these circumstances will need to be discussed frankly with the parents.

Genetic counselling

Providing accurate genetic advice about developmental anomalies of the nervous system is a challenging task. Referral for specialist advice is strongly

recommended.

---

*We are very grateful to Dr Nagui Antoun (Addenbrooke's Hospital, Cambridge), Dr Fred Pickworth (Norfolk and Norwich Hospital), and Mr Paul Chamberlain (John Radcliffe Hospital, Oxford) for the images shown in this chapter and for advice on their interpretation.

Further reading

Aicardi J (1998). *Diseases of the nervous system in childhood*, 2nd edn. Mac Keith Press, London.

Baraitser M. (1997). *The genetics of neurological disorders*, 3rd edn. *Oxford Monographs on Medical Genetics 34*. Oxford University Press, Oxford.

Bock G, Marsh J, eds (1994). *Neural tube defects. Ciba Foundation Symposium 181*. John Wiley, Chichester.

Faerber EN, ed. (1995). *CNS magnetic resonance imaging in infants and children. Clinics in Developmental Medicine No. 134*. Mac Keith Press, London.

Friede RL (1989). *Developmental neuropathology*, 2nd (revised and expanded) edn. Springer-Verlag, Berlin.

Gleeson JG, Walsh CA (2000). Neuronal migration disorders: from genetic diseases to developmental mechanisms. *Trends in Neuroscience* **23**, 352–9.

Govaert P, de Vries LS (1997). *An atlas of neonatal brain sonography. Clinics in Developmental Medicine No. 141–2.* Mac Keith Press, London.

Maclennan A, for the International Cerebral Palsy Task Force (1999). A template for defining a causal relationship between acute intrapartum events and cerebral palsy: international consensus statement. *British Medical Journal* **319**, 1054–9.

Miller G, Clark GD, eds (1998). *The cerebral palsies. Causes, consequences, and management*. Butterworth-Heinemann, Boston.

Milunsky A, ed. (1998). *Genetic disorders and the fetus. Diagnosis, prevention and treatment*, 4th edn. Johns Hopkins University Press, Baltimore.

Norman MG *et al.* eds (1995). *Congenital abnormalities of the brain. Pathologic, embryologic, clinical, radiologic and genetic aspects*. Oxford University Press, New York.

Pless IB, ed (1994). *The epidemiology of childhood disorders*. Oxford University Press, New York.

Stanley F, Blair E, Alberman E. (2000). *Cerebral palsies: epidemiology and causal pathways. Clinics in Developmental Medicine No.151*. Mac Keith Press, London.

Swaiman KF, Ashwal S, eds (1999). *Paediatric neurology. Principles and practice*, 3rd edn. Mosby, St Louis.

Wallis D, Muenke M (2000). Mutations in holoprosencephaly. *Human Mutation* **16**, 99–108.

24.22.1 Introduction: structure and function

M. Hanna

Basic anatomy of skeletal muscle

We possess more than 150 voluntary (skeletal) muscles most of which are attached to the skeleton at both ends through tendons. Complex voluntary movements of the body are achieved by integrated activity of different skeletal muscle groups. To the naked eye a transverse section of any skeletal muscle reveals small units known as muscle fascicles. Each skeletal muscle fascicle is composed of many basic structural units known as muscle fibres (Fig. 1). Muscle fibres are cylindrical structures that may be several centimetres long and 50 to 100 µm in diameter. A muscle fibre is a highly specialized cell. Like any other cell it has a membrane (the sarcolemma), it contains cytoplasm (the sarcoplasm), and it has an endoplasmic reticulum (the sarcoplasmic reticulum) as well as other subcellular organelles such as mitochondria. However, unlike cells from many other tissues, muscle cells are multinucleate. Typically the nuclei are positioned at the edges of the muscle fibre. The sarcolemma of muscle fibres possesses specialized regions known as motor endplates. These endplate regions are the points at which the axon innervating a muscle fibre forms synapses. Release of acetylcholine from the presynaptic region transmits the axonal action potential to the muscle fibre membrane by binding to postsynaptic acetylcholine receptors located in the sarcolemma at the endplate. The sarcolemma is differentially permeable to ions. This allows different concentrations of ions to be maintained inside and outside the membrane and this is critical in maintaining the resting membrane potential. A chain of important structural proteins maintain the integrity of the sarcolemma by linking intracellular muscle fibre cytoskeletal proteins to the extracellular matrix. These structural proteins include dystrophin (located in a subsarcolemmal distribution), the dystrophin-associated glycoprotein complex (a trans-sarcolemmal protein complex), and laminin (located extracellularly). These important proteins may be dysfunctional in certain forms of genetic muscle diseases (see Chapter 24.22.2).

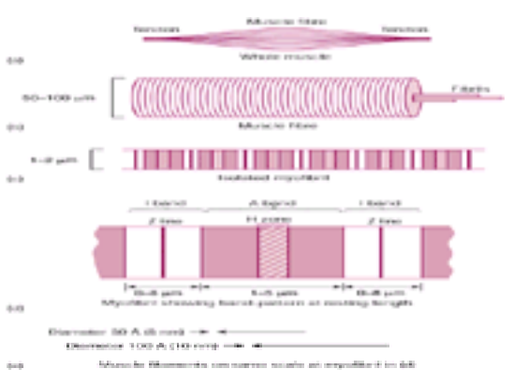

Fig. 1 The dimensions and arrangement of the contractile components in a muscle. The whole muscle (a) is made up of fibres (b) which contain cross-striated myofibrils (c, d). These are constructed of two types of protein filaments (e), put together as shown in Fig. 2. (Reproduced from Huxley and Hanson, 1960, with permission.)

After staining, or if suitably illuminated, muscle fibres are seen to have regular cross-striations that extend right across the inside of the fibre, dividing it up into sarcomeres (Fig. 1). The parts of the cross-striations are identified by letters. The light I band is divided by the dark Z line and the dark A band has the lighter H zone in its centre. The region between two adjacent Z lines is called a sarcomere. The cross-striations are due to the presence of the principal contracile filamentous proteins, actin and myosin, in the sacroplasm. These filamentous proteins are arranged in rod-like structures known as myofibrils. A single myofibril contains many protein filaments. In life, myofibrils are transparent on routine light microscopy, but if viewed with a polarizing microscope, a typical pattern of cross-striations can be seen within individual myofibrils. The correct understanding of the basic microscopic anatomy of this pattern of cross-striations was critical to the discovery of the sliding filament theory of skeletal muscle contraction.

The sliding filament theory of skeletal muscle contraction

The protein filaments contained within myofibrils are of two types; the thin filaments are composed of actin, tropomyosin, and troponin and the thick filaments are composed of myosin (Fig. 2). The thick filaments are approximately twice the diameter of the thin filaments.

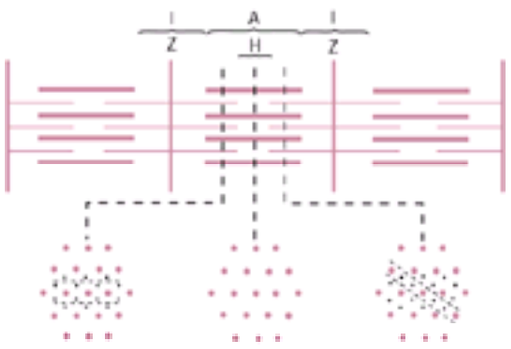

Fig. 2 Diagram illustrating the arrangement of the different kinds of protein filament (thick filaments: myosin; thin filaments: actin) in a myofibril. At the top are three sarcomeres drawn as they would appear in longitudinal section. Below are transverse sections through the H zone and other parts of the A band where the thick and thin filaments interdigitate. The plane of section determines whether, in electron micrographs, there seem to be one or two thin (actin) filaments between two thick (myosin) ones. (Reproduced from Huxley and Hanson, 1960, with permission.)

The thick filaments are lined up to form the A bands, whereas the array of thin filaments forms the less dense I bands. The lighter H bands in the centre of the A bands are the regions where, when the muscle is relaxed, the thin filaments do not overlap the thick filaments. The Z lines transect the myofibrils and connect to the thin filaments. If a transverse section through the A band is examined under the electron microscope, each thick filament is found to be surrounded by six thin filaments in a regular hexagonal array (Fig. 2). The myosin molecules have large globular heads at their C-terminal portions (Fig. 3). The heads contain an actin-binding site that hydrolyses ATP. During muscle contraction, cross-linkages occur between the heads of the myosin molecules and the actin molecules (Fig. 3).

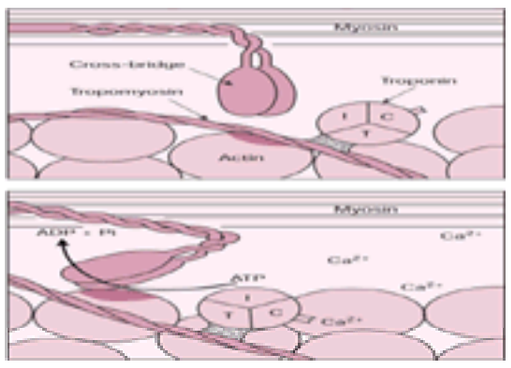

Fig. 3 Initiation of muscle contraction by Ca^{2+} ions. The cross-bridges (heads of myosin molecules) attach to binding sites on actin (striped areas) and swivel when tropomyosin is displaced laterally by binding of Ca^{2+} ions to troponin C. (Modified from Katz AM, 1975, Congestive heart failure. *New England Journal of Medicine* **293**, 1184.)

The thin filaments are composed of two chains of actin that form a long double helix. Tropomyosin molecules are long filaments located in the groove between the two chains of actin. Troponin molecules are small globular units located at intervals along the tropomyosin molecules. Troponin has three components: troponin T, responsible for binding to tropomyosin; troponin I, which inhibits the interaction of actin and myosin; and troponin C, which contains the binding sites for the Ca^{2+} ions that initiate contraction (Fig. 3).

The process by which shortening of the contractile elements of muscle is brought about is sliding of the thin filaments over the thick filaments. The width of the A band is constant, whereas the Z lines move closer together when the muscle contracts and further apart when it is stretched. The sliding during muscle contraction is produced by breaking and reforming of the cross-linkages between actin and myosin. The immediate source of energy for contraction is hydrolysis of ATP localized to the myosin head.

Neural activation of muscle fibres—the motor unit

The motor unit is the final common pathway for all voluntary muscle activity. The motor unit is composed of an anterior horn cell (located within the spinal cord), its peripheral axon, the axon terminal branches, the associated neuromuscular junctions, and the muscle fibres innervated. The muscle fibres of a single motor unit are spatially dispersed throughout a muscle and only a few fibres innervated by the same anterior horn cell are contiguous. The number of motor units varies greatly between muscles, from approximately 1000 in leg muscles to 100 in intrinsic hand muscles. The number of muscle fibres per motor unit also varies greatly. Motor units also differ in physiological and biochemical characteristics. Two main types of motor units are recognized, each composed of a single muscle fibre type. Type 1 muscle fibres contain many mitochondria and are slightly smaller than type 2 muscle fibres as they contain myofibrils, which are more slender. Type 1 fibres contain a high concentration of oxidative enzymes and more fat. Type 2 fibres are larger, contain fewer mitochondria, but have a higher concentration of glycogen and enzymes involved in anaerobic metabolism such as myophosphorylase. All skeletal muscles contain a mixture of both fibre types, typically in a chequerboard pattern when stained appropriately (with the myofibrillar ATPase reaction) and visualized under light microscopy (Fig. 4). Type 1 fibres are also known as slow fibres since they contract and relax slowly and are abundant in muscles concerned mainly with maintaining posture. In contrast, type 2 fibres contract and relax quickly and are also known as twitch fibres. Type 2 fibres can be further subdivided into type 2a and 2b based on their intensity of staining with myofibrillar ATPase reaction at different pHs (Table 1).

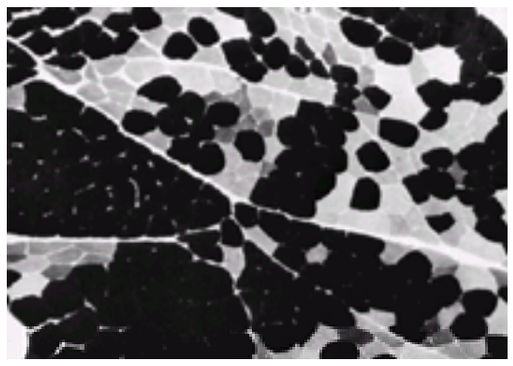

Fig. 4 A transverse section of human skeletal muscle obtained by biopsy from a patient with spinal muscular atrophy stained for the myofibrillar ATPase reaction after preincubation at pH 4.6. There is extensive evidence of fibre type grouping, particularly of the type 1 fibres, resulting from reinnervation. Magnification ×150. (Kindly supplied by Dr Margaret Johnson.)

Normally muscle fibres do not contract in isolation, rather the muscle fibres which comprise the motor unit contract together in response to depolarization of an anterior horn cell. Such depolarization is transmitted along the axon until it invades the nerve terminal. This results in opening of voltage-gated calcium channels located in the presynaptic membrane. Calcium enters the nerve terminal down an electrochemical gradient. The resulting increase in presynaptic calcium concentration promotes fusion of acetylcholine-containing vesicles normally present in the nerve terminal with the presynaptic membrane. Quanta of acetylcholine are released into the synaptic cleft and diffuse to the postsynaptic membrane to bind to and activate acetylcholine receptors. Acetycholine binding causes opening of its receptor channel allowing cations to enter the muscle fibre in the endplate region. This cation flux depolarizes the postsynaptic membrane resulting in a mini-endplate potential. The summation of endplate potentials results in the excitation of the postsynaptic membrane, which is then conducted along the muscle fibre membrane. The excitation is transmitted into the muscle fibre by invaginations of the sarcolemma known as the T-tubule system. Activation of calcium channels in the T-tubule system membrane results in opening of calcium channels in the sarcoplasmic reticulum. Calcium is then released into the muscle fibre cytoplasm, initiating muscle contraction.

Energy production in skeletal muscle

Resting skeletal muscle requires remarkably little energy. However, the requirement for energy production may increase dramatically in response to exercise, since energy is required for muscle contraction. Adenosine triphosphate (ATP) is the main source of energy in muscle. ATP is required for shortening of the contractile filaments and also for the active reuptake of calcium into the sarcoplasmic reticulum after each muscle contraction. Maintenance of electrochemical gradients across the sarcolemma also requires ATP. Resynthesis of ATP from ADP is essential for normal muscle function. The two main energy-producing pathways in muscle are glycolysis in the sarcoplasm and oxidative phosphorylation in the mitochondria. Resynthesis of ATP from ADP is also aided by phosphocreatine and the creatine kinase reaction. Creatine kinase catalyses the transfer of high-energy phosphate from phosphocreatine to ADP in circumstances in which ATP demand may outstrip ATP production, for example at the very beginning of exercise before oxidative phosphorylation or glycolysis is activated. Glycolysis is the main pathway of ATP synthesis in anaerobic conditions and results in the generation of lactate. Oxidative phosphorylation is the major ATP-generating pathway in aerobic conditions. The main fuel sources in skeletal muscle are glucose, glycogen, and fatty acids. In anaerobic conditions, glycogen is the main energy source. In aerobic exercise, glycogen and glucose are utilized initially, but after approximately 30 min, fatty acids are the main energy source. In resting aerobic muscle, fatty acids provide the principal source of fuel. Several muscle diseases are recognized in which energy metabolism is impaired and are known as the metabolic myopathies.

Diseases of human skeletal muscle—overview

Human muscle diseases may be conveniently divided into those which are genetically determined and those which are acquired (Table 2).

The clinical history in muscle diseases

Although a muscle biopsy is usually needed to determine the exact type of muscle disease, the clinical history and examination are usually sufficient to determine whether a muscle disease is present or absent. Since many muscle diseases are genetically determined, it is particularly important to consider the family history. A careful drug history is also essential.

Although many diseases may affect skeletal muscle (Table 2), there are three main symptoms with which patients may present: muscular pain, muscular weakness, and fatiguability. A further important but less common symptom is darkening of the urine (pigmenturia) due to release of myoglobin from damaged muscle. This occurs particularly in the metabolic myopathies. Unless pigmenturia has been dramatic, patients may not volunteer this symptom. Muscle pain is a common symptom, but in only about one-third of patients presenting with this symptom will an underlying muscle disease be identified. In those without a definable muscle disease, many are considered to have a psychogenic cause for their muscle pain, although some may have as yet undefined disorders of muscle metabolism. Sometimes it can be difficult for the patient and the physician to distinguish between pain originating in muscle and pain originating in joints or bones. Certain rheumatological diseases may result in joint pain as well as muscle pain. For example, systemic lupus erythematosus may cause arthritis and polymyositis. Muscle pains may take the form of cramps, which are involuntary contractions of muscle groups. Simple muscle cramps are not uncommon in the elderly and frequently occur at night. There is usually no underlying muscle disease but drugs such as diuretics (which induce hypokalaemia) may be implicated. In younger patients, muscle cramps may be the presenting feature of a metabolic muscle disease such as McArdle's disease. Muscle pain brought on by exertion is a particular feature of the metabolic muscle diseases. Muscle contractures may also be a source of muscle pain in patients with metabolic myopathies. Patients experience a pain similar to a cramp, but unlike a cramp, electromyography reveals that a contracture is electrically silent.

Muscle weakness is a common feature of muscle diseases and the distribution of weakness in most is in the proximal limb muscles. Patients may complain of difficulty performing tasks which involve lifting their arms up to or above their head, such as brushing hair. Proximal lower limb muscle weakness causes difficulties getting out of low chairs and in climbing stairs. Muscle diseases often affect the limb musculature symmetrically—although there are important exceptions to this. For example, one of the common autosomal dominant muscular dystrophies, fascioscapulohumeral muscular dystrophy, often affects the limb muscles in an asymmetrical fashion. Some muscle diseases may affect the facial musculature as well as that of the limb. Symptoms may include difficulty in whistling, in closing the eyes, or in articulating. Respiratory muscle disease may cause breathlessness. It is important to determine the natural history of muscle weakness. In most genetically determined muscle diseases, weakness progresses slowly over years; occasionally the patient may experience attacks of weakness separated by periods when they seem to have normal strength, as in the periodic paralyses. The muscle weakness in the inflammatory muscle diseases usually develops more rapidly.

Fatiguability is defined as an increase in weakness with exercise. Patients may describe that they can start a particular physical activity but the longer they continue the weaker they become. They may also complain that they become weaker as the day goes on. Myaesthenia gravis, a disorder of neuromuscular transmission, is the principal cause of fatiguability. In patients with myaesthenia gravis, fatiguability can usually be demonstrated at the bedside. Patients with metabolic muscle diseases may also experience fatiguability.

The physical examination in muscle disease

The examination may be broadly divided into two aspects. First, an examination is made to establish whether there are any clues to the cause of the muscle disease. In this context, the general physical examination is very important. Particular attention is paid to eliciting signs which might indicate an underlying endocrine or rheumatological disorder. For example, signs of hyper-/hypothyroidism, Cushing's syndrome, or of rheumatological disorders such as systemic lupus erythematosus. Inspection of the skin may reveal the appearances of dermatomyositis. The second part of the examination involves examining the muscular system to determine the extent and severity of the condition; this may in addition give further clues to the aetiology. The muscles are inspected for any atrophy or hypertrophy (as occurs in some muscular dystrophies) or for any spontaneous activity of the muscle fibres (such as fasciculation, which might indicate an anterior horn cell disorder). The muscles should be palpated for any tenderness or swelling, which may occur in inflammatory muscle diseases. Myotonia is a delayed relaxation of muscle after contraction. This may be observed by asking the patient to clench their fist and then to open it rapidly. A patient with myotonia is unable to open the clenched fist rapidly due to an inability to relax the contracted muscles quickly. Myotonia may also be evident on percussion of muscle. The examination of muscle power is carried out systematically starting with the cranial musculature before proceeding to the arms and legs. The degree of weakness is assessed with reference to the Medical Research Council grading scale (0 to 5). The distribution of weakness is also noted, since different muscle diseases have characteristic patterns of weakness. Bedside assessment of respiratory muscles including the diaphragm is also important, although detailed assessment of these muscles requires formal spirometry. Finally, the tendon reflexes are elicited. These are generally preserved in acquired muscle diseases, except when there is advanced weakness; however, they may be lost relatively early in the course of dystrophies.

Investigating the patient with muscle disease

Investigations are generally only instituted when the history and examination have provided clear evidence that the patient has symptoms and/or signs of muscle disease. The investigations are aimed primarily at determining the exact type of muscle disease as it is essential to establish whether the patient has a treatable muscle disease, such as an inflammatory myopathy. Many investigations of increasing complexity and invasiveness are available.

Simple blood tests allow an assessment of the endocrine and nutritional status of the patient (such as thyroid function, the consumption of excess alcohol, or the presence of vitamin D deficiency). Measurement of the blood creatine kinase is important as this can be an indicator of the degree of muscle fibre damage or necrosis. The creatine kinase is generally elevated in the inflammatory muscle diseases and in many of the muscular dystrophies.

Increasingly, DNA-based testing is available from simple blood samples. This can be particularly helpful and in some situations may obviate the need for further more-invasive tests, such as a muscle biopsy. For example, if analysis of the dystrophin gene on the X chromosome identifies a pathogenic mutation known to associate with Duchenne muscular dystrophy, the diagnosis is confirmed. It is likely that there will be greater availability of DNA-based tests for genetic muscle disease in the future and this will become an increasingly important aid to diagnosis.

The diagnosis of metabolic muscle diseases may be achieved by specific dynamic tests. For example, McArdle's disease can be diagnosed using the ischaemic lactate test, and mitochondrial disease may be suspected on the basis of subanaerobic exercise tests (both these tests are described in the relevant section).

Detailed nerve conduction studies and electromyography are useful in determining whether a patient has a neuropathy, a defect in neuromuscular junction transmission, or a myopathy. Electromyography is useful in characterizing any spontaneous activity of muscle, such as fasciculations or myotonia. Although electromyography is generally useful in confirming the presence of a myopathy, it is less useful in determining the cause.

Muscle biopsy allows a detailed analysis of the internal architecture of muscle and is an extremely valuable and safe investigation in carefully selected patients. Using a range of histochemical stains, histochemical enzyme reactions, and immunological techniques on frozen muscle biopsy sections, much information of diagnostic use can be obtained. Different muscle diseases often reveal characteristic patterns of abnormalities, which are usually identified by light microscopic techniques. Using basic histochemical stains the features of different muscular dystrophies are generally similar; the most common features being marked variations in fibre diameter, internal nuclei, fibre splitting, fibre necrosis and regeneration, and increase in connective tissue. However, a more precise diagnosis of the type of muscular dystrophy can now be obtained by immunostaining techniques. Antibodies which are raised against specific membrane proteins allow quantitative analysis. For example, staining using antibodies directed against dystophin reveals no or very little dystrophin in cases of Duchenne muscular dystrophy (Fig. 5). Prominent inflammatory infiltrates are typically seen in muscle sections from patients with inflammatory myopathies. Figure 5, Figure 6, and Figure 7 show the muscle biopsy features of some of the metabolic myopathies. In some cases the changes seen on the biopsy clearly indicate a myopathic process, but it is not possible to be more specific in the absence of typical immunological, inflammatory, or metabolic changes.

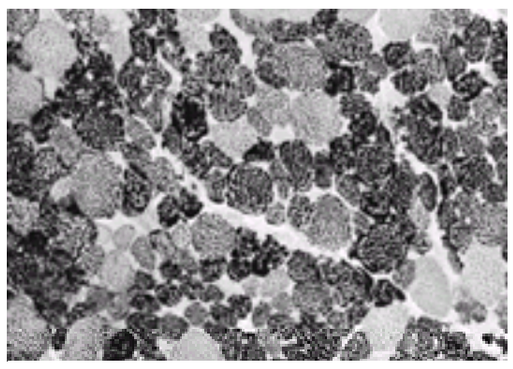

Fig. 5 A transverse section of human skeletal muscle obtained from a patient with carnitine deficiency and stained with Sudan black B. The massive accumulation of neutral fat, especially with the type 1 fibres, is evident. Magnification ×196.

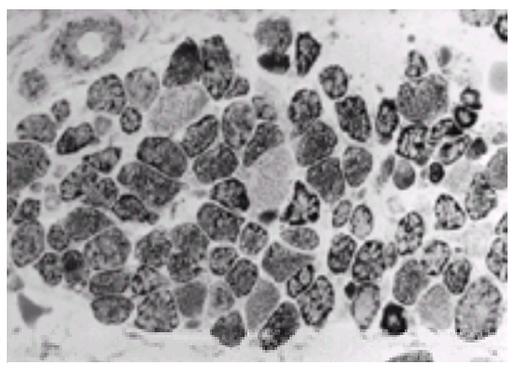

Fig. 6 A transverse section of skeletal muscle obtained from a patient with mitrochondrial myopathy, stained for the MADH-TR reaction. The type 1 fibres are darkly stained and show the typical reticulated appearance of so-called 'ragged-red fibres' with massive mitochondria, particularly in many fibres just deep to the sarcolemma. Magnification ×384. (Kindly supplied by Dr Margaret Johnson.)

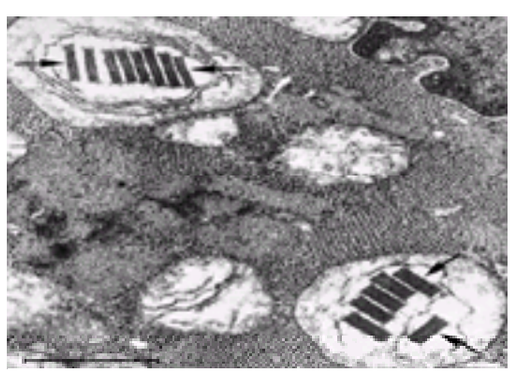

Fig. 7 A transverse section of a biopsy specimen obtained from one quadriceps muscle in a patient with mitrochondrial myopathy showing arrays of paracrystalline inclusions in the damaged mitochondria. Bar = 1 µm. (Kindly supplied by Dr Michael Cullen.)

Further reading

Huxley HE, Hanson J (1960). In: Bourne GH, ed. *The structure and function of muscle*, Vol.1. Academic Press, New York.

Walton JN, Mastaglia FL (1980). The molecular basis of muscle disease. In: Thompson RHS, Davison AN, eds. *The molecular basis of neuropathology*. Edward Arnold, London.

24.22.2 Muscular dystrophy

K. Bushby

Introduction

Muscular dystrophy is not a single disease. Many different types of muscular dystrophy can be recognized: all are primary, genetically determined disorders of muscle and all cause muscle weakness, which is usually progressive. The various types of muscular dystrophy share several characteristic findings on muscle biopsy, most notably a variation of fibre size, evidence of muscle fibre necrosis, and usually replacement of muscle tissue by fat and fibrous tissue. These pathological findings are often but not always accompanied by elevation of the serum creatine kinase. While the key clinical sign in muscular dystrophy is muscle weakness, the distribution of that weakness and the association with other features such as wasting, hypertrophy, and joint contractures are the defining features which are most helpful in making a clinical diagnosis, together with age at presentation and rate of progression. Unusual manifestations of muscular dystrophy are muscle pain, rhabdomyolysis, and myoglobinuria. Compli-cations may include cardiac and respiratory failure or anaesthetic problems. These complications may be specific to particular types of muscular dystrophy. Taken in conjunction with the clinical findings in any patient, precise diagnostic tests (either through DNA analysis or protein analysis of a muscle biopsy sample) are available for a growing number of these disorders, as knowledge of the underlying mechanism of disease for each of these entities has increased. Confirmation of the type of muscular dystrophy in any individual patient is critical to the provision of appropriate management, prognostic advice, and genetic counselling. No form of muscular dystrophy is currently curable, although various experimental therapeutic procedures are under investigation.

Classification of the muscular dystrophies

Various classifications of the muscular dystrophies have been proposed, reflecting historical advances in the understanding of this group of diseases (Box 1). The current basis for classification combines an appreciation of the clinical features with the ability to determine the molecular basis for the disease. Therefore the eponymous names (for example Duchenne muscular dystrophy) still in common usage reflect the detailed clinical descriptions provided by early clinicians: other disease names reflect the recognized pattern of muscle involvement in a particular condition (for example facioscapulohumeral muscular dystrophy). Disease designations based on the genetic or protein defect in a particular disorder (for example dystrophinopathy) are becoming more widely used, reflecting the fact that some disorders previously believed to be clinically distinct actually represent different manifestations of lesions at the same locus. Genetic analysis has also revealed an unsuspected level of heterogeneity with different genetic causes for disorders which show superficial clinical similarities. This can be seen most strikingly within the 'limb girdle' group of muscular dystrophies.

Box 1 Basis of the classification of muscular dystrophies

- Clinical description.
- Genetics (autosomal dominant/recessive/X-linked).
- Underlying gene/protein defect.
- Localization or function of the protein involved.

The pathophysiology of the muscular dystrophies

Biochemical and physiological experiments failed to shed any light on the mechanisms by which muscular dystrophy could arise, and it has only been since the cloning of the *dystrophin* gene in 1987 that progress has been made. It is now quite clear that proteins involved in several different functions within the muscle cell can, when altered or absent, cause muscle damage and account for the pathological and clinical features of a muscular dystrophy. Some of these proteins are components of the membrane of the muscle fibre which may have a structural or signalling role, others are components of the nuclear envelope or are muscle-specific enzymes (see Fig. 1).

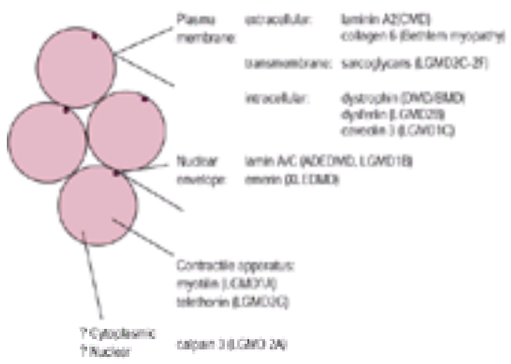

Fig. 1 Schematic diagram to show localization within the muscle fibre (where known) of the proteins known to be involved in producing a dystrophic phenotype. In particular, groups of proteins localizing to the plasma membrane or nuclear envelope have been implicated in a number of different types of muscular dystrophy.

General points on the diagnosis of muscular dystrophy

Box 2 summarizes some of the major considerations in arriving at the correct diagnosis of a muscular dystrophy. History taking at the time of presentation (Box 3) may be particularly informative. The clinical history may be pathognomonic. Detailed diagnostic information is given in the following text relating to specific diseases. The main tools for specific diagnosis in muscular dystrophy are the use of antibodies for the immunolabelling of muscle biopsy sections and/or the application of specific DNA-based genetic analysis.

Box 2 Diagnosis of muscular dystrophy

- History (especially motor milestones, age at onset, physical prowess as a child).
- Age of patient (congenital/childhood/teenage/adult presentation).
- Pattern of muscle involvement on examination (predominantly proximal/distal, upper limb/lower limb, symmetrical/asymmetrical).
- Pattern of associated features on examination (contractures, muscle wasting, hypertrophy).
- Level of serum creatine kinase in active disease (massive elevation in dystrophinopathy, sarcoglycanopathy, dysferlinopathy, calpainopathy, congenital muscular dystrophy (some); moderate elevation in facioscapulohumeral muscular dystrophy, Emery Dreifuss muscular dystrophy, congenital muscular dystrophy (some); normal to mild elevation in autosomal dominant limb girdle muscular dystrophy, facioscapulohumeral muscular dystrophy (some)).
- Electromyography (to exclude neurogenic causes of weakness, especially if serum creatine kinase is not markedly elevated).
- Muscle imaging (ultrasound scans can confirm muscle involvement, but to confirm pattern of muscle involvement need MRI/CT).
- Muscle biopsy, histology, and storage of frozen biopsy material for further analysis.
- Specialized analysis of muscle biopsy (immunocytochemistry, immunoblotting, electron microscopy).
- DNA analysis.

Box 3 History taking in muscle disease

- Question in detail about early motor development.
- Eliciting what actually were the first symptoms experienced by a patient may be difficult but is important in highlighting the initial pattern of muscle involvement—lower limb versus upper limb/proximal versus distal musculature.
- Asking 'when were you at your fastest' may be informative in determining age of peak motor performance.
- Ask about performance at school sports.
- Particularly useful indicators in that respect are the ability to climb ropes (upper girdle weakness), muscle pain on running, a tendency to spend all the time in goal at football(!).
- Do not assume that difficulty climbing stairs always indicates proximal muscle weakness—it may reflect an inability to push up on the toes.
- Ask specifically about the ability to stand on tiptoe/stand on heels. The need to wear heels on shoes at all times may indicate Achilles tendon contractures.
- Patients who had early Achilles tendon contractures may have had them operated on before being referred for diagnosis. Ask about this.

General points on the management of the muscular dystrophies

Despite the fact that no cures for muscular dystrophy are established, there are many issues for management which may be important or specific to the various types. However, there is as yet little systematic or comprehensive clinical research into management and randomized trials of management regimes are few and far between. It is nonetheless appropriate that where possible, patients with a diagnosis or a suspected diagnosis of muscular dystrophy should be referred to a specialist clinic. The multidisciplinary approach of these clinics ensures that patients have access to the full range of diagnostic facilities, are able to obtain specialized physiotherapy advice, and can obtain accurate genetic counselling where this is required. Access to patient support organizations and their staff is also of paramount importance. The diagnosis of any kind of muscular dystrophy, in that it inevitably implies a progressive and incurable disease, possibly with implications for children or other relatives, is a considerable burden and one which needs to be recognized and supported.

The congenital muscular dystrophies

The congenital muscular dystrophies (**CMD**) are defined by their very early childhood onset. They comprise a number of disorders with different molecular pathological bases for the diseases.

Presentation

1. Neonatal presentation:
   - hypotonia, which may be prenatal
   - feeding problems (usually mild)
   - joint contractures, especially knees, hips, and ankles (see Fig. 2).

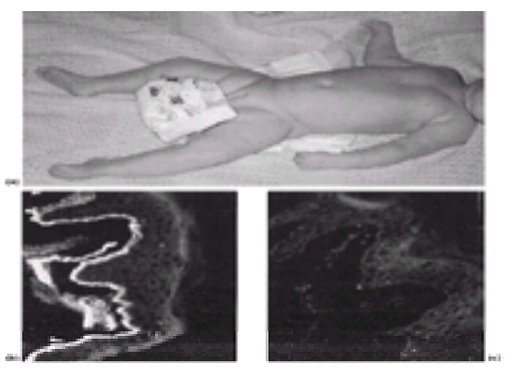

Fig. 2 (a) Typical clinical picture of a baby presenting with merosin negative muscular dystrophy. Note the hypotonic posture, and mild contractures of the hips, knees, and ankles. (b), (c) Immunofluorescence picture of skin biopsy labelled with an antibody to laminin A2 (merosin) showing normal and absent

labelling patterns. This investigation can be carried out on a variety of tissues including skin, muscle, or placenta.

2. Early childhood presentation:
   - delayed motor milestones
   - failure to thrive
   - repeated respiratory infections.
3. Later childhood presentation (rare):
   - mainly proximal muscle symptoms
   - history of delayed motor milestones
   - rigid spine, contractures of ankles, hips, and knees.

Differential diagnosis

In the neonatal and early childhood presentation the main clinical diagnostic confusion (after excluding central causes of hypotonia) – may be with spinal muscular atrophy (check *SMN*/*NAIP* genes), congenital myotonic dystrophy (facial weakness is usually more pronounced and diagnosis can be excluded on genetic testing), and congenital myopathy (may be distinguished on muscle biopsy). In all of these conditions, serum creatine kinase is either normal or much lower than seen in many congenital muscular dystrophies.

With later childhood presentation the differential diagnosis is as above, plus Duchenne muscular dystrophy (though calf hypertrophy is usually more pronounced and serum creatine kinase is typically higher—biopsy will exclude the diagnosis) or childhood presentation of a limb girdle type of muscular dystrophy.

Classification

There are several recognized forms of congenital muscular dystrophy, and as there is considerable heterogeneity in the group which remains, many more different entities are ultimately likely to be distinguished at the genetic level. The first level of subdivision is on the basis of whether clinically there is a 'pure' muscle phenotype, or whether there is also prominent involvement of the central nervous system and eyes. In the types of congenital muscular dystrophy with a 'pure' muscle phenotype clinically, the next major subdivision is based on the presence or absence of a muscle protein merosin or laminin A2. Examination of collagen VI labelling in the muscle fibre may also be informative (see Table 1).

Establishing the diagnosis

Serum creatine kinase may in some kinds of congenital muscular dystrophy be normal, but is typically elevated at least twofold, and up to 20-fold or more in the laminin A2 deficient group. Muscle biopsy shows dystrophic changes and examination of LAMA2 or collagen VI in muscle or skin is used to distingush cases with normal and abnormal or absent protein. Secondary changes in laminin A2 probably reflect the involvement of a number of different primary mutations. Mutations in the gene *FKRP* are responsible for some of these cases. Magnetic resonance imaging of the brain is a useful adjunct to diagnosis as it will confirm the presence of white matter changes always present after 6 months of age in primary LAMA2 deficiency, and the characteristic brain malformations in the other types of congenital muscular dystrophy (see Table 1).

Prognosis and management

The muscle weakness in congenital muscular dystrophy may be relatively static, but the complications of that weakness can be severe, and vary according to the precise diagnosis. The degree of muscle weakness is quite variable. In primary laminin A2 deficient congenital muscular dystrophy, the severity of the disease correlates roughly with the abundance of laminin A2 in the muscle, with children completely lacking laminin A2 rarely achieving independent ambulation. Others may learn to walk independently but this is usually much later than usual, and these children may later lose this ability. Joint contractures and scoliosis are major complications of the disease and cause much additional disability, requiring careful management by physiotherapy, standing regimes, splinting, bracing, and surgery where appropriate. Feeding problems may be intractable and lead to chronic malnutrition unless treated by nasogastric or gastrostomy feeding. Malnutrition may contribute to susceptibility to chest infections, which is also heightened by weakness of the respiratory muscles. These children are at risk of respiratory failure and their follow-up should include monitoring for this complication which can be effectively managed by the provision of non-invasive home nocturnal ventilation. Cardiac failure is reported in some children, mainly in the group lacking laminin A2.

The overall prognosis depends on the type of congenital muscular dystrophy. Children with the most severe forms are at risk of dying in early childhood. If they survive this period, with appropriate management of feeding problems, respiratory, and cardiac complications, survival into adult life is the norm. In the 'pure' forms of congenital muscular dystrophy the intellect is normal, and these children should be encouraged to pursue the best possible education, with appropriate support for their physical difficulties. Fukuyama congenital muscular dystrophy (FCMD), muscle–eye–brain disease (MEBD), and Walker–Warburg syndrome (WWS) may be dominated by intellectual and visual handicap. General management issues remain the same; however, on the whole all of these groups carry a much poorer prognosis (see Table 1).

Genetic counselling

All these disorders are autosomal recessive in inheritance. As the molecular basis for these disorders becomes better established, specific prenatal and carrier testing will become more widely referable in this group of conditions.

Dystrophin deficiency

This group, including two of the most common forms of muscular dystrophy, Duchenne and Becker muscular dystrophy, involve the same gene and protein. These are X-linked diseases, caused by mutations, most of which are deletions, in the *dystrophin* gene.

Presentation

Duchenne muscular dystrophy

- All patients are symptomatic within the first 3 years of life though the mean age at diagnosis is 4 years 10 months.
- Motor milestones are often delayed (half of cases are not walking by 18 months).
- Speech is also frequently delayed.
- Patient is unable to run—there is a pronounced waddling gait on attempting to rush.
- Patient is unable to jump with both feet together or to hop—there is no spring in the step.
- 'Climbs up legs' on rising from the floor—Gower's manoeuvre.
- Rarely presents with anaesthetic complications.

Becker muscular dystrophy

- The mean age at onset of Becker muscular dystrophy is 11 years, though the range of age at presentation is extremely wide and the diagnosis may be made at any age.
- A proportion will have had delayed motor milestones (this may correlate as much with reduction in IQ as with major motor problems at that age).
- Many describe being unable to keep up with peers at school.
- Difficulty with high steps, climbing hills.
- Muscle pains after exercise are a common complaint especially in teenagers (rarely myoglobinuria).

Manifesting carriers of Duchenne muscular dystrophy/Becker muscular dystrophy

A highly variable group, who may occasionally be as severely affected as those with Duchenne muscular dystrophy or as mildly or more mildly than those with Becker muscular dystrophy.

Dystrophin-associated cardiomyopathy

Symptoms and signs of hypertrophy progressing to dilated cardiomyopathy in the absence of major muscle symptoms. Some patients have elevated serum creatine kinase.

Establishing the diagnosis

The clinical presentation of Duchenne muscular dystrophy is very characteristic. Hypertrophy of the calf muscles is almost universal (see Fig. 3), sometimes accompanied by muscle hypertrophy elsewhere, most frequently involving deltoid, parts of the quadriceps, the tongue, and masseters. Wasting of the pectoral and scapular muscles leads to hypotonia around the shoulders detected as the child 'slipping through the hands' on being lifted. In the lower limbs, quadriceps power is weaker than that of the hamstrings. Formal examination of a small child may be difficult, and the main clinical tool is observation of walking, attempting to run, jump, and climb stairs, and to rise from the floor. It is imperative to give the child space to attempt to run, as this will bring out the lack of spring in the step and the lack of fluidity of the attempted running.

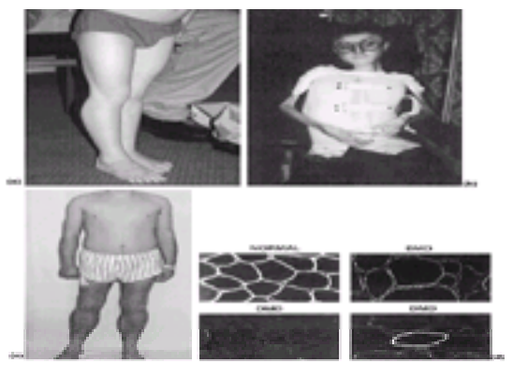

Fig. 3 (a) Child with Duchenne muscular dystrophy at presentation, showing the marked calf and quadriceps hypertrophy and tendency to rise onto the toes. (b) Teenage boy in the later stages of the disease, showing the complications of marked immobility, scoliosis, and muscle wasting. This young man has now been maintained on home nocturnal ventilation successfully for more than 7 years. (c) Clinical pattern at presentation in a young man with Becker muscular dystrophy. Note hypertrophic muscles in calves and quadriceps and mild wasting around the shoulder girdle. (d) Immunocytochemical analysis of dystrophin in normal muscle, Becker muscular dystrophy muscle, and Duchenne muscular dystrophy muscle. In normal muscle, dystrophin labels evenly around the periphery of the muscle fibres. This labelling is typically patchy and reduced in Becker muscular dystrophy, and is either completely or nearly completely absent in Duchenne muscular dystrophy.

Becker muscular dystrophy has been described as a 'slow motion version of Duchenne muscular dystrophy' in that the pattern of muscle involvement in these two allelic disorders is essentially identical (see Fig. 3), but progresses at a much slower rate in Becker muscular dystrophy. Patients with Becker muscular dystrophy may be quite strong on formal muscle examination, but tend to show subtle signs of proximal muscle weakness on climbing stairs or running. They frequently have hypertrophy involving the same muscle groups as seen in Duchenne muscular dystrophy. Some patients have pes cavus.

Serum creatine kinase is always massively elevated, even to more than 200 × normal, but levels of serum creatine kinase do not distinguish the severity of the disease. Muscle biopsy and electromyography are non-specifically dystrophic. Molecular confirmation of the diagnosis is essential to assist in defining prognosis and to provide appropriate genetic counselling. Genetic analysis readily confirms the diagnosis in the 60 to 80 per cent of patients in whom a deletion of the *dystrophin* gene is present: in all patients the diagnosis can be established by the finding of absent or reduced dystrophin in the muscle biopsy (see Fig. 3). This analysis also allows the distinction of dystrophinopathy from the rarer (in most populations) limb girdle types of muscular dystrophy.

Prognosis

Within the 'dystrophinopathy' group the prognosis is highly variable. By definition, those patients with Duchenne muscular dystrophy lose the ability to walk by the age of 12. The development of scoliosis, respiratory failure, and cardiomyopathy (see Box 4) during the teenage years can all be managed so that survival into or beyond the late 20s is becoming more common. Patients with Becker muscular dystrophy are ambulant beyond 16 years of age, and may remain able to walk independently into their fifth decade or longer. These patients are susceptible to cardiac failure at any age from the teens onward and should be monitored for this complication on a regular basis (see Box 4). Respiratory failure is a late complication in Becker muscular dystrophy and correlates with very late stage disease. Lifespan in Becker muscular dystrophy may be normal, or reduced in more severe disease. An 'intermediate' group is also recognized who lose ambulation between 12 and 16: their overall prognosis is also intermediate between Duchenne muscular dystrophy and Becker muscular dystrophy. Around 8 per cent of carriers of Duchenne muscular dystrophy or Becker muscular dystrophy may develop some signs of the disease: rarely this is in a full blown form comparable to the disease in the boys. In practise, there is a continuum of severity with the highest incidence in the Duchenne muscular dystrophy group (birth incidence 1 in 3500 male live births). As lifespan is so much longer in the Becker muscular dystrophy group, however, the prevalence of the two conditions is roughly similar (about 24 per million population in northeast England).

Box 4 Practice point: cardiac involvement in dystrophinopathy

- All patients with dystrophinopathy are at risk of developing cardiomyopathy which progresses with age. It is frequently asymptomatic, and needs to be sought through full cardiac assessment including echocardiography, as treatment with antifailure medication may improve function and prognosis.
- Cardiac transplantation has been used successfully in patients with Becker muscular dystrophy and manifesting carriers of dystrophinopathy.
- Cardiac compromise is the major determinant of operative risk in boys with Duchenne muscular dystrophy, and all should have a full cardiac assessment in advance of any surgery at any age.

Over the whole group, there is a correlation between dystrophin abundance (as measured in a muscle biopsy sample) and severity: children with completely absent dystrophin tend to be confined to a wheelchair slightly earlier than children whose biopsies contain low levels of dystrophin. Patients with Becker muscular dystrophy have much higher levels of dystrophin (Fig. 3). These dystrophin levels also correlate in most cases with the type of mutation found in the *dystrophin* gene—in Duchenne muscular dystrophy most deletions are out of frame, not supporting the production of dystrophin, while Becker muscular dystrophy patients typically have in-frame deletions, allowing the production of a reduced amount of dystrophin of a slightly smaller size.

While these correlations are useful in a general sense, they are not absolutely predictive of outcome in an individual case, and must always be taken in the context of the clinical features of the patient. They can be useful though in giving the best possible guide to prognosis, especially in those patients with Becker muscular dystrophy who present early or who are identified by neonatal screening or the incidental finding of a high serum creatine kinase level.

Management

Duchenne muscular dystrophy—the early stages

Proper management of a child with Duchenne muscular dystrophy starts with awareness of the possibility of the diagnosis in any boy who is not walking by the age of 18 months or whose mobility is poor compared with his peers. The current mean age at diagnosis of nearly 5 years reflects a typical but unacceptable delay of at least 2 years since the onset of disease is noticed by the parents. The principal impetus to early diagnosis at present is the ability to offer parents the option of prenatal diagnosis in subsequent pregnancies. When specific treatments become available, there will also be a need to implement such treatments before the disease is too advanced.

Once the diagnosis has been considered, measurement of the serum creatine kinase will confirm the suspicion and ideally a referral into a specialist unit should be made at this stage. The specialist unit should have rapid-track access to DNA diagnostic and muscle biopsy facilities to confirm the diagnosis as quickly as possible. Duchenne muscular dystrophy is a devastating diagnosis, and should be given to the family following guidelines for the best practice for disclosure of bad news—the parents should be seen together wherever possible in complete privacy, they should have time to sit and ask questions, and have access to experienced staff for support and further information. Access to support groups and the relevant national charity is also appropriate. Supporting information should also be passed immediately to the general practitioner, health visitor, and school who may never have looked after a child with this type of condition before.

Since Duchenne muscular dystrophy is an X-linked condition, early access to genetic counselling is also vital shortly after diagnosis (see Box 5).

Box 5 Genetic counselling in dystrophinopathy is an essential part of the management of any family where a diagnosis of dystrophinopathy has been made because the potential implications go far beyond the index case

- These are X-linked diseases.
- The new mutation rate in the dystrophin gene is high.
- Most cases of Duchenne muscular dystrophy are born now in families with no prior history of the disease.
- None the less, even in these families, other female relatives (through the maternal line) are at risk of being carriers.
- The essential piece of information is the delineation of the dystrophin mutation in the affected child (easy to find in the 60 per cent in whom the mutation is a deletion, harder and much more specialized if it is not).
- In the presence of a known mutation, female relatives can be offered testing directly to see if they are carriers or not.
- They may choose to have prenatal diagnosis on the basis of that testing.
- Even if mothers of boys with Duchenne muscular dystrophy can be shown not to be somatic carriers of the mutation in their son, they still may have a proportion of egg cells containing the mutation (a situation known as 'germline mosaicism'). They therefore remain at a 10 to 20 per cent risk of having another affected child in a future pregnancy.
- Boys with Duchenne muscular dystrophy do not often have children, but men with Becker muscular dystrophy often do (overall fitness reduced to around 2/3). All of their daughters are obligate carriers of Becker muscular dystrophy, but none of their sons are at risk.

In the early stages of the disease it is advisable for the child to be introduced to a community physiotherapist for advice on stretching, which at this stage can usually concentrate on the ankles and hips, with the emphasis on the maintenance of symmetry. Boys frequently develop a toe-walking gait which is partially compensatory for their proximal muscle weakness—walking splints or ankle–foot orthoses are therefore not appropriate at this stage and any early Achilles tendon contractures are better managed through passive stretching and night-time below-knee splints. At the point at which walking becomes impossible independently, the child can often be rehabilitated in long leg callipers or knee–ankle–foot orthoses. Lengthening of the Achilles tendon is often necessary to allow the child to do well with knee–ankle–foot orthoses. The length of time children walk in knee–ankle–foot orthoses varies from child to child—residual muscle strength is probably the best predictor of how long a child will use them for mobility, but motivation on the part of the child's family and school is also a key factor.

Despite a consensus that use of corticosteroids does prolong ambulation for up to a couple of years, various questions about side-effects and long-term complications remain and thus corticosteroids are not universally used. Where steroids are considered, intermittent treatment probably offers the best balance between efficacy and side-effects.

Duchenne muscular dystrophy—after mobility is lost

Inevitably there comes a point in Duchenne muscular dystrophy where even the ability to stand supported in knee–ankle–foot orthoses is lost and the child is confined permanently to a wheelchair. The provision of an electric chair with indoor and outdoor access is critical to the best possible maintenance of independence. Seating in the chair should also be carefully addressed to promote an upright and symmetrical posture. Scoliosis is an almost universal complication of Duchenne muscular dystrophy and close liaison with the orthopaedic department is necessary to co-ordinate management which in the long term is likely to include spinal surgery. Physiotherapy priorities shift towards postural support, the prevention and containment of contractures, and respiratory maintenance. Measurements of forced vital capacity carried out regularly provide an indication of the trend of respiratory function—forced vital capacity usually plateaus soon after confinement to a wheelchair and thereafter falls. The timing of surgery for scoliosis therefore needs to take this variable into account, though cardiomyopathy is probably an even greater risk factor in the timing of surgery (see Box 4).

As forced vital capacity falls further, boys are at serious risk of chest infections and ultimately nocturnal respiratory failure. Symptoms of this respiratory failure may be extremely insidious and totally missed unless explicitly sought (see Box 6). Routine overnight pulse oximetry (which can readily be carried out at home provided the equipment is available) can show a trend of deteriorating overnight oxygenation and, together with the monitoring of symptoms, highlight the time at which elective nocturnal respiratory support, ideally initially at least through non-invasive means, should be provided. Such respiratory support abolishes symptoms, reduces the tendency to chest infections, and undoubtedly improves lifespan.

Box 6 Respiratory failure in neuromuscular disease is a complication which needs to be specifically sought

- It may be the result of intercostal muscle or diaphragmatic weakness or a combination of the two. The presence of a scoliosis or other spinal deformity may be an additional factor.
- Nocturnal problems tend to dominate.
- Frank symptoms of morning CO_2 retention may be seen (poor colour, morning sickness, headaches, confusion) but these are late symptoms and the problem should be detected by investigation or careful history taking before this stage.
- Increasing frequency of chest infections may indicate incipient respiratory failure.
- Subtle signs include loss of appetite and weight loss, loss of energy and enthusiasm.
- Poor sleep, increasing wakefulness at night, inability to lie flat may also be seen together with a tendency to fall asleep during the day.
- Difficulties swallowing and difficulty completing sentences may also be seen.
- In many muscle diseases, the main risk of respiratory failure is when the patient is no longer able to walk independently and weakness is pronounced (for example Duchenne muscular dystrophy, Becker muscular dystrophy, congenital muscular dystrophy, facioscapulohumeral muscular dystrophy, limb girdle muscular dystrophy, etc.).
- In other muscle diseases, respiratory failure may be an earlier feature and present while the patient is still ambulant (for example multicore and other congenital myopathies, some forms of congenital muscular dystrophy).

In the late stages of Duchenne muscular dystrophy, nutrition may be of concern. Loss of weight occurs in most boys as the disease progresses, and issues of diet and the possibility of supplemental nutrition need to be addressed.

The actual cause and timing of death in Duchenne muscular dystrophy is hard to predict. Some patients will die of a particularly severe chest infection. In others cardiomyopathy may be difficult to control or a cardiac arrhythmia may arise. Early onset of cardiomyopathy is a poor prognostic sign. Talking about death to these patients and their parents, helping them to prepare and also addressing their fears and uncertainties is another important but easily neglected aspect of management.

Education

On average, children with Duchenne muscular dystrophy have an IQ around one standard deviation below the normal mean; often a striking verbal-performance deficit is also observed. Schooling should offer the best possible environment for learning, including full attention to information technology equipment, while

supporting the very real physical needs of the child. Families and areas vary as to whether this will be best provided through mainstream or special schooling. With a good education and medical support, boys with Duchenne muscular dystrophy and the appropriate intellectual potential can go on to higher education, and where possible, should be encouraged to do so.

Becker muscular dystrophy

Management issues in Becker muscular dystrophy tend to cover the same broad areas as Duchenne muscular dystrophy, but with the deterioration in muscle function over a much longer timescale. Certain complications, such as scoliosis, are very unusual. Other complications, such as cramping muscle pains after exercise, which can be a particular problem in the teenage years, are more common. Despite the fact that Becker muscular dystrophy is much milder than Duchenne muscular dystrophy it can represent a considerable and insurmountable disability for the person who has it, and problems with adjustment, poor self-esteem, and poor body image are all fairly common in this group. No hard data exist to define completely any intellectual problems in Becker muscular dystrophy, but on average it is likely that this group has a general reduction in IQ, though probably not to the extent seen in Duchenne muscular dystrophy. Cardiac complications may occur at any age in Becker muscular dystrophy (see Box 4): respiratory complications tend to be a feature of the late stages of the disease.

Facioscapulohumeral muscular dystrophy

Facioscapulohumeral muscular dystrophy is an example of a muscular dystrophy named for the most characteristic pattern of muscle involvement observed (that of involvement of the facial, scapular, and humeral muscles predominantly). However, other muscle groups usually become involved with time and may even be involved at onset.

Presentation

- Age at presentation is variable. Most affected individuals manifest some symptoms by their teens or twenties. Occasionally symptoms may be very minor, even late in adult life.
- Symptoms may, unusually for a muscular dystrophy, be very markedly asymmetrical.
- Early symptoms typically include facial weakness (inability to bury eyelashes or puff cheeks; this often goes unnoticed), shoulder girdle weakness manifesting as problems in reaching high shelves, changing lightbulbs, or climbing ropes, and foot drop.

An infantile form of facioscapulohumeral muscular dystrophy is recognized with early childhood onset, extremely marked facial weakness and progressive weakness of both the shoulder and pelvic girdle musculature. Lumbar lordosis may be profound. Hearing loss and retinal telangiectasia may be seen in any patient with facioscapulohumeral muscular dystrophy but are particularly associated with this most severe form of the disease.

Differential diagnosis

The clinical pattern of facioscapulohumeral muscular dystrophy can be very distinctive, and the asymmetry of muscle involvement is a major clue. However, facial weakness may be very variable, and if it is absent or subtle, confusion can arise with forms of limb girdle muscular dystrophy.

Diagnostic investigations

Serum creatine kinase may be normal or mildly elevated. Muscle biopsy and electromyography (EMG) provide supportive evidence for a muscular dystrophy; some inflammatory features are sometimes also seen in the biopsy. Most cases (if not all) of facioscapulohumeral muscular dystrophy are linked to chromosome 4q35. Although the nature of the gene responsible for facioscapulohumeral muscular dystrophy is not yet known, a DNA-based test is available which can confirm the diagnosis in 95 per cent of cases. This test involves the demonstration of a DNA deletion which is consistently associated with the disease. It is likely that this deletion alters the expression of an unknown gene close to the telomere of chromosome 4q (position effect variegation).

Prognosis and management

Infantile facioscapulohumeral muscular dystrophy is a progressive disease which leads to early confinement to a wheelchair and the development of such complications as scoliosis and respiratory failure. This condition is most frequently seen as a result of a new dominant mutation in cases with no family history, and these children often have particularly large DNA deletions on chromosome 4. The development of a lumbar lordosis, seen also in later onset facioscapulohumeral muscular dystrophy, together with secondary hip flexion contractures, can be very disabling. Bracing may be partially successful at controlling the lordosis, but at the expense of some loss of mobility.

More typically, facioscapulohumeral muscular dystrophy is a slowly progressive disease. As the disease progresses it can involve the proximal as well as the distal lower limb muscles. Around 20 per cent of patients with facioscapulohumeral muscular dystrophy will become unable to walk independently, most over the age of 40. Involvement of the proximal lower limbs before the age of 20 years is a poor prognostic sign, indicating an increased likelihood of needing to use a wheelchair. Some patients describe progression as being stepwise in nature, with periods of faster deterioration alternating with phases of plateauing of their symptoms. Footdrop is a common complaint, which can be helped by the provision of daytime ankle–foot orthoses. A significant proportion of patients with facioscapulohumeral muscular dystrophy complain of painful muscles, for which no cause can be found, and for which pain relief may be difficult. Some patients find swimming or a small dose of antidepressants useful for this symptom. More severely affected patients with facioscapulohumeral muscular dystrophy may develop respiratory failure or swallowing problems and these complications should be sought. Cardiomyopathy is rarely reported.

Genetic counselling

Facioscapulohumeral muscular dystrophy is an autosomal dominant disease and as such an affected person has a 50 per cent chance of transmission to his or her offspring, regardless of sex. Use of the new DNA diagnostic techniques have shown that up to 30 per cent of cases of facioscapulohumeral muscular dystrophy may represent new dominant mutations. Germline mosaicism is also common. Genetic analysis has also shown a higher proportion of asymptomatic gene carriers than expected, with females overrepresented in this group. The availability of a relatively straightforward genetic test in this disorder has opened up the possibility of presymptomatic and prenatal testing, which were previously impossible. However, despite an overall correlation between the size of the deletion found and the severity of the symptoms, the DNA test is not useful in predicting the severity of the disease—people in individual families with apparently the same sized deletion may have a very variable experience of the disease (see Fig. 4).

Fig. 4 Mother and daughter with facioscapulohumeral muscular dystrophy. The mother is extremely mildly affected and has minimal symptoms. By contrast the daughter was affected from early childhood and has been wheelchair dependent outside from her early teens. Note the daughter's expressionless face and her posture—she is leaning forwards due to a combination of her marked lumbar lordosis, a major feature of the condition, and hip flexion contractures.

Emery Dreifuss muscular dystrophy

Emery Dreifuss muscular dystrophy has a highly characteristic phenotype. X-linked recessive, autosomal dominant, and autosomal recessive forms are recognized, and the genes involved in these conditions both encode proteins which are components of the nuclear envelope (see Fig. 5). The gene involved in X-linked Emery Dreifuss muscular dystrophy is *emerin*, and the gene involved in autosomal dominant and autosomal recessive Emery Dreifuss muscular dystrophy is *lamin A/C*.

Fig. 5 Muscular dystrophy phenotypes characterized by prominent contractures. (a) This patient has autosomal dominant Emery Dreifuss muscular dystrophy, with a proven mutation in his *lamin A/C* gene. The elbow and Achilles tendon contractures seen here, combined with his markedly rigid spine, are very similar to the pattern of contractures and weakness seen in the X-linked form of the disease. (b) Bethlem myopathy in a woman with marked contractures of the elbows, ankles, and spine. In addition she has finger flexion contractures, demonstrated here by attempting to straighten the fingers with the wrist extended.

Presentation

- Patients may present at any age, most typically in the early teens, though symptoms may be present much earlier than that.
- Contractures of the ankles and elbows and rigidity of the spine often predate any clear weakness.
- Consequently, these patients have frequently had Achilles tendon release before the diagnosis is suspected.
- Weakness and wasting are typically humeroperoneal in distribution.
- A key part of these conditions, which may rarely be present at presentation, is cardiac involvement, most typically arrhythmias (see below). An alternative phenotype (limb girdle muscular dystrophy 1B, see Box 7 and Fig. 5) exists in combination with mutations in the same gene as autosomal dominant Emery Dreifuss muscular dystrophy (*lamin A/C*). These patients may present with purely cardiac disease, or with a more proximal 'limb girdle muscular dystrophy' presentation, without prominent contractures. *Lamin A/C* mutations are also described in patients with partial lipodystrophy without prominent muscle weakness.

Box 7 Autosomal dominant Emery Dreifuss muscular dystrophy and limb girdle muscular dystrophy 1B—the 'laminopathies'

- These disorders are caused by mutations in the lamin A/C gene. Lamin A/C, like emerin, is a component of the nuclear envelope.
  - The phenotype is variable, depending on the presence or not of contractures as a major component of the phenotype.
  - Where contractures are present, these typically involve the elbows, Achilles tendons, and spine. In these patients, there is often a humeroperoneal pattern of muscle weakness as in X-linked Emery Dreifuss muscular dystrophy.
  - Where contractures are less of a feature, patients typically present with proximal muscle weakness.
  - In both groups, cardiac involvement is the most important complication. Arrythmias may lead to sudden death and should be sought and treated appropriately.
  - A phenotype with exclusively cardiac involvement has also been described.
  - New mutations and germline mosaicism is common in this group.

Confirming the diagnosis

Serum creatine kinase is typically mildly elevated in Emery Dreifuss muscular dystrophy. Muscle biopsy shows non-specific histological features: in X-linked Emery Dreifuss muscular dystrophy, emerin is absent in muscle and skin. Detection of mutation in the *emerin* gene is necessary to offer genetic counselling to female relatives at risk of being carriers.

The involvement of *lamin A/C* (the gene responsible for autosomal dominant Emery Dreifuss muscular dystrophy) cannot be determined by antibody analysis in muscle, but requires the demonstration of a *lamin A/C* mutation. A secondary deficiency of laminin b-1 may be seen in muscle from some patients with autosomal dominant Emery Dreifuss muscular dystrophy, but is not specific to this condition. Many *lamin A/C* mutations arise *de novo*, and germline mosaicism is common.

Differential diagnosis

Other muscular dystrophies may present with contractures as an important component (Fig. 5). Some forms of congenital muscular dystrophy may be associated with contractures and a rigid spine. Bethlem myopathy may present congenitally (often with torticollis) or in early childhood: here finger flexion contractures, elicited especially on wrist extension, are more prominent and cardiac involvement is not associated. Bethlem myopathy is itself genetically heterogeneous, involving mutations in any of the genes for collagen 6A1, 6A2, and 6A3. Like autosomal dominant Emery Dreifuss muscular dystrophy, some cases of Bethlem myopathy show a secondary reduction in laminin b-1 staining in muscle biopsy: the significance of this is uncertain.

In some cases, calpainopathy (limb girdle muscular dystrophy 2A) may be associated with contractures of the ankles, elbows, fingers, and paraspinal muscles. However, the associated weakness here is predominantly proximal and of a characteristic distribution (see below). These patients have typically a higher creatine kinase, absent calpain 3 on biopsy and *CAPN3* mutations.

Prognosis and management

The prognosis in Emery Dreifuss muscular dystrophy relates almost directly to the ability to manage the life-threatening arrhythmias to which every patient with either the X-linked or dominant form is susceptible. Severe arrhythmias are inevitable by the third decade. All patients with this diagnosis should therefore be under regular cardiological review, and cardiac pacing once a rhythm disturbance is detected may be life-saving. However, in autosomal dominant Emery Dreifuss muscular dystrophy the risk of cardiomyopathy remains and may be less amenable to routine treatment.

Management of the contractures in Emery Dreifuss muscular dystrophy is the other main issue, and will involve close liaison with a physiotherapist. Operative treatment of contractures, especially at the Achilles tendons, is commonly performed, but while such surgery does work in the short term, contractures often recur. With increasing age, however, contractures often stabilize. Muscle weakness may worsen but progression is usually very slow.

The limb girdle muscular dystrophies

The broad definition of limb girdle muscular dystrophy comes from the classification of Walton and Nattrass in 1954 when the term was suggested to describe those patients with weakness of the proximal musculature who did not fulfil the criteria for the X-linked muscular dystrophies or for facioscapuluhumeral muscular dystrophy. The term has always encompassed a heterogeneous group of disorders: now that many of them can be distinguished at the gene or protein level it is no longer sufficient to use it without qualification as to the specific type of disease (see Table 2 and Table 3). The type of limb girdle muscular dystrophy may be suggested by the precise pattern of muscle involvement, with confirmation from a combination of genetic and protein analysis. The ability to provide a precise diagnosis in limb

girdle muscular dystrophy has greatly improved the prognostic and genetic information which can be given to these patients.

The approach to diagnosis in limb girdle muscular dystrophy

Could it be dominant disease?

Autosomal dominant limb girdle muscular dystrophy represents only around 10 per cent of the total limb girdle muscular dystrophy population, and limb girdle muscular dystrophies 1A, 1C, 1D, and 1E have been very rarely reported (see Table 2). In families with a dominant history the most likely diagnoses are facioscapulohumeral muscular dystrophy (exclude facioscapulohumeral muscular dystrophy on DNA analysis especially if there is any suspicion of facial weakness), limb girdle muscular dystrophy 1B (allelic with autosomal dominant Emery Dreifuss muscular dystrophy—see Box 7), and Bethlem myopathy. New mutations are common, however, so that if the clinical features are suggestive of one of these disorders, then the diagnosis should be pursued even in the absence of a family history. Features which should raise the suspicion of dominant disease are less marked elevation of creatine kinase (typically normal to fives times normal in dominant disease and much higher than this in active recessive disease) or the presence of early and prominent contractures.

What is the age and nature of the presentation?

Variability in the age of presentation and of rate of progression is usual in the various autosomal recessive types of limb girdle muscular dystrophy. However, some broad conclusions can be helpful (Fig. 6). Childhood presentation is most common in sarcoglycanopathy, which may superficially resemble dystrophinopathy, with frequent calf (and other muscle) hypertrophy. Adult onset cases are less frequent and are essentially 'Becker like' in presentation. However, whatever the age at presentation, in sarcoglycanopathy, quadriceps is almost always stronger than the hamstrings. This is the reverse of the pattern seen in dystrophin deficiency. Calpainopathy may present with early childhood symptoms, especially contractures of the Achilles tendons, but onset is most commonly between 8 and 15 years of age. Dysferlinopathy typically presents in the late teens or early twenties, and early features may include proximal weakness or distal involvement (usually manifesting as difficulty standing on tiptoe).

Fig. 6 Typical clinical pictures of patients with different types of autosomal recessive limb girdle muscular dystrophy. (a) Calpainopathy or limb girdle muscular dystrophy 2A; note the predominantly atrophic pattern of muscle involvement and Achilles tendon contractures. The stance is often wide-based due to the imbalance of the hip abductors and adductors. (b) Dysferlin deficiency or limb girdle muscular dystrophy 2B. Note the wasting of the posterior calf muscles and flat-footed stance. (c) Child with g sarcoglycanopathy or limb girdle muscular dystrophy 2C. Note the lordotic posture and scapular winging, both of which may be more marked at presentation in sarcoglycanopathy than in dystrophin deficiency. (d) Adult with g sarcoglycanopathy, to illustrate the variability in severity of sarcoglycan deficiencies and the muscular hypertrophy which may be as marked or more marked than in dystrophin deficiency.

Which investigations should be performed?

Serum creatine kinase is greatly elevated in all forms of autosomal recessive limb girdle muscular dystrophy, but may be only marginally elevated or within the normal range in autosomal dominant limb girdle muscular dystrophy. EMG confirms a primary myopathic process and standard analysis of the muscle biopsy shows dystrophic changes (which especially in dysferlinopathy can be accompanied by evidence of inflammation). Specialized investigations are always necessary to attempt to determine the type of limb girdle muscular dystrophy, and require a muscle biopsy as the starting point for immunologically based diagnosis (see Fig. 7), taken in conjunction with the clinical features.

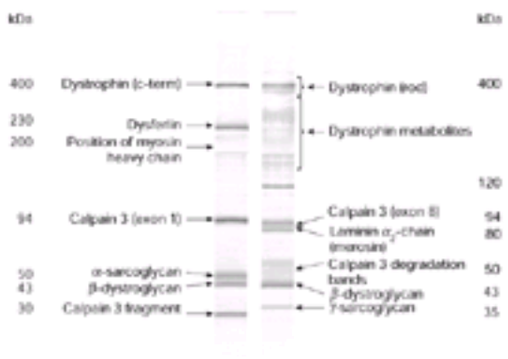

Fig. 7 Multiplex Western blotting as an approach to diagnosis in limb girdle muscular dystrophy. Two strips of a Western blot of control human skeletal muscle protein extracts immunostained with a mixture of antibodies to the proteins indicated. Absence or reduced intensity of a particular species compared with the other proteins labelled in the same lane can indicate which gene and protein is implicated in that patient's disease. (By courtesy of Dr L. V. B. Anderson, University of Newcastle upon Tyne.)

Scheme for specialized investigations

Do the clinical features or family history suggest a specific disorder (see Box 8, Box 9, and Box 10)? If so look for that first.

Box 8 Clinical features of sarcoglycanopathies

- Most frequently present in childhood, but may present at any age. Intrafamilial variability is common.
- These conditions are most closely related clinically to dystrophinopathy which will be the major differential diagnosis and have a similar spectrum of severity.
- Typically motor milestones are less delayed than in dystrophinopathy.
- Muscle hypertrophy is common.
- Intelligence is not affected.
- Cardiomyopathy is an important complication, though not universal, and should be sought through careful surveillance.
- Respiratory failure is an important late complication.
- Scoliosis is seen in the most severely affected individuals.
- Prognosis overall is typically better than dystrophinopathy presenting at a similar age.

Box 9 Clinical features of calpainopathy

- This is the most common form of limb girdle muscular dystrophy in most populations.
- May present at any age but typically 8 to 15 years.
- Highly selective muscle involvement: posterior thigh weakness, and wasting; scapular winging common at onset.
- Muscle hypertrophy rare—tends to be predominantly atrophic pattern.
- Preservation of hip abductor muscles even at late stages contributes to characteristic wide-based stance.
- Most have Achilles tendon contractures: a subgroup presents with much more prominent contractures in an Emery Dreifuss muscular dystrophy-like pattern.
- Progression is variable but never as fast as Duchenne muscular dystrophy.
- Cardiac involvement is not common but respiratory impairment may be seen in late stages.
- Prognosis in all but the most severe and early onset cases is good.

Box 10 Dysferlinopathy: clinical features

- Presentation most commonly in late teens or early twenties.
- Patients often report good muscle prowess before onset of disease. Serum creatine kinase may not be massively elevated in presymptomatic cases.
- Occasional patients present with unilateral calf swelling which may be tender and lead to the clinical diagnosis of myositis.
- Primary muscle involvement is always in the lower limbs, with absence of upper girdle involvement at onset.
- Lower limb involvement may be of proximal muscles or distal muscles. The distal muscles involved first are typically posterior (leading to difficulty standing on tiptoe as an early feature) but may be anterior.
- Progression is typically slow and life expectancy is not reduced. This is the usually mildest type of limg girdle muscular dystrophy.
- Cardiomyopathy is not reported and respiratory involvement is usually mild and at a very late stage only.
- This phenotype is genetically heterogeneous, with another locus for Miyoshi myopathy on chromosome 10.
- The main differential diagnosis, especially in patients presenting with distal weakness, may be an alternative form of distal myopathy. Typically here the creatine kinase is not so high.

The sarcoglycanopathies

Dystrophin staining may be mildly abnormal in these patients reflecting the close and interdependent relationship between the proteins of the dystrophin associated complex; but the predominant abnormality on immunolabelling or immunoblotting will be the absence or reduction of one or more of the sarcoglycans. The pattern of reduction of these proteins may give a clue as to the primary gene involvement. In g sarcoglycanopathy, there may be preservation of some or all of the other sarcoglycans, with specific loss of g sarcoglycan. a Sarcoglycanopathy may similarly present with a selective reduction, or with more widespread loss or reduction in labelling for all members of the complex. In † and b sarcoglycanopathy there is typically loss of the whole sarcoglycan complex. Determining the primary protein involved is important to direct genetic analysis. Detection of the mutation is necessary to offer prenatal diagnosis and specific genetic counselling, but as there are few recurrent mutations in the *sarcoglycan* genes the search for mutations can be very time-consuming: hence the need to obtain the best possible information from immunolabelling before embarking on detection of mutations.

Calpainopathy

Here the sarcoglycans are normal, as is dystrophin. Currently available antibodies to calpain 3 do not work on tissue sections but need to be used on immunoblotting. Detection of reduced or absent calpain on immunoblotting (Fig. 7) indicates the need to search for *calpain 3* mutations, which are highly variable, generally non-recurrent, and which may involve any part of the large (24 exons) gene. However, a secondary reduction in calpain 3 may also be seen in some cases of dysferlin deficiency. Hence the need to use all antibodies in combination to ensure that the primary problem can be identified correctly and the appropriate gene searched for mutations.

Dysferlinopathy

Here all other proteins with the possible exception of calpain 3 are within the normal range and deficiency of dysferlin can be demonstrated on tissue sections or immunoblotting. Decreased or absent dysferlin in muscle is an indication to proceed to mutation detection. The *dysferlin* gene is very large (55 exons), and as with the other forms of limb girdle muscular dystrophy, mutations are highly variable.

Laminopathy

Some (but by no means all) patients with *lamin A/C* mutations have a secondary reduction of the muscle protein laminin b-1 on immunolabelling of tissue sections. The confirmation of the diagnosis is by demonstration of a *lamin A/C* mutation.

Other forms of limb girdle muscular dystrophy

Limb girdle muscular dystrophy 2G and 2H appear relatively restricted in their geographical distribution. Limb girdle muscular dystrophy 2I is seen worldwide and is due to mutations in *FKRP*. This type of locus can be associated with cardiomyopathy.

Management

Once the diagnosis is secure, management should include monitoring and treatment for the specific complications of the various subtypes. If a clear diagnosis is not possible (for example where appropriate samples are not available or where the diagnosis cannot be reached even after exhaustive investigation) the management should as a minimum include physiotherapy and regular cardiac and respiratory surveillance.

Oculopharyngeal muscular dystrophy

Oculopharyngeal muscular dystrophy is unusual in that it has an exceptionally late presentation.

Presentation

- Presentation is typically in the sixth decade.
- It commonly presents with ptosis, dysphagia to solids, and dysphonia which may be as severe as in myotonic dystrophy.
- Other features include ophthalmoparesis, facial weakness, and proximal muscle weakness.

Diagnosis

The muscle biopsy in oculopharyngeal muscular dystrophy typically shows the presence of rimmed vacuoles and intranuclear inclusions. DNA analysis confirms the presence of an expanded guanine–cytosine–guanine repeat in the *poly(A) binding protein 2* gene.

Prognosis and management

Ptosis can be managed surgically, but frequently recurs. Dysphagia may respond, at least partially, to surgical intervention with myotomy of the cricopharyngeal muscle and other annular muscle fibres. Potentially life-threatening complications may include aspiration pneumonia and regurgitation. Progression of the limb muscle

weakness is highly variable.

Genetic counselling

Oculopharyngeal muscular dystrophy is an autosomal dominant disorder. Genetic analysis offers the potential for presymptomatic testing if this is specifically sought.

Prospects for specific treatment in muscular dystrophy

Drug treatments have a limited place in the treatment of muscular dystrophy at present. Despite many small-scale studies reporting the beneficial use of steroids in prolonging ambulation in Duchenne muscular dystrophy and possibly the sarcoglycanopathies as well, these treatments are not universally applied due to reservations about side-effects and the true cost–benefit equation for this intervention. Steroids do not appear to be beneficial in facioscapulohumeral muscular dystrophy: b agonists are currently under trial in this condition after some fairly positive findings in a pilot study.

Treatments to modify the underlying disease are not yet available for clinical application. Gene transfer experiments in animal models have proved the general feasibility of this approach to these genetic diseases, at least on a small scale. Modification of mutations, either by drugs or other means, is an area of research, as is the concept of upregulating the production of ancillary proteins.

Supportive treatment for patients and their families remains the mainstay of treatment at present, and this is likely to be the case for the current generation of patients at least. This treatment is ideally provided through a specialized multidisciplinary team, bringing together with the 'myologist' the skills of medical and associated colleagues from physiotherapy, occupational therapy, genetics, cardiology, respiratory medicine, and orthopaedics.

Further reading

Bushby K, Anderson LVB, eds (2000). *Molecular methods in medicine: the muscular dystrophies*. Humana Press, New York. [A practical guide to DNA and protein based diagnosis in many types of muscular dystrophy.]

Emery AEH, ed (1998). *Neuromuscular disorders: clinical and molecular genetics*. Wiley, Chichester. [A 'state of the art' review of the level of knowledge about a variety of neuromuscular disorders in 1998.]

Emery AEH, ed. (2001). *The muscular dystrophies*. Oxford University Press. [Authoritative reviews of the different types of muscular dystrophy.]

Karpati G, Hilton-Jones D, Griggs RC, eds (2001). *Disorders of voluntary muscle*, 7th edn. Cambridge University Press, Cambridge, UK. [The book contains sections on basic physiology and anatomy of muscle as well as approaches to clinical diagnosis and management.]

With the rate of change in the last few years in the information we have available about genetically determined diseases, the most up to date reviews of the subject may be found on the Internet rather than in traditional textbooks. Online Mendelian Inheritance in Man (http://www3.ncbi.nlm.nih.gov/) provides a good starting point for up to date information on a range of subjects. The Leiden muscular dystrophy database (http://www.dmd.nl/) offers specific information on the muscular dystrophies.

24.22.3 Myotonia

David Hilton-Jones

Introduction

The term myotonia, and related terms such as paramyotonia and neuromyotonia, cause much confusion. Various diseases accompanied by myotonia have different molecular origins and many associated symptoms and signs. Myotonia can be considered as a symptom, as a physical sign, or as a neurophysiological phenomenon, but understanding is perhaps best served by discussing these in reverse order.

The basic neurophysiological finding is of repetitive muscle-fibre action potentials following a stimulus, which may be voluntary contraction or muscle percussion. The repetitive electrical activity causes muscle contraction, and thus myotonia is characterized by delayed muscle-fibre relaxation following such a stimulus. Electromyography demonstrates the repetitive firing. Characteristically, the discharge gradually declines in amplitude and frequency, producing the so-called 'dive-bomber' sound in the monitoring loudspeaker.

As a physical sign, myotonia is demonstrated either as delayed muscle relaxation following voluntary contraction (for example grip myotonia, Fig. 1), or as persistent muscle dimpling following percussion (percussion myotonia, Fig. 2).

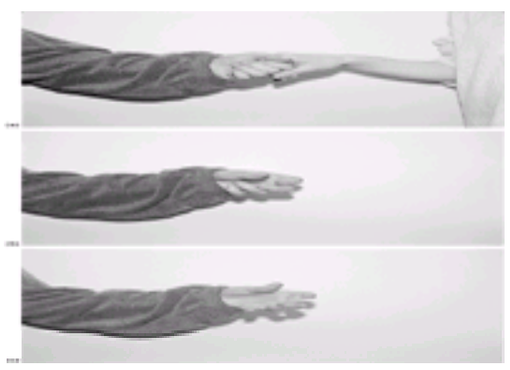

Fig. 1 Grip myotonia. The patient was asked to grip the examiner's fingers tightly for 3 s, and then to release the grip as rapidly as possible. The following two photographs were taken at 3-s intervals.

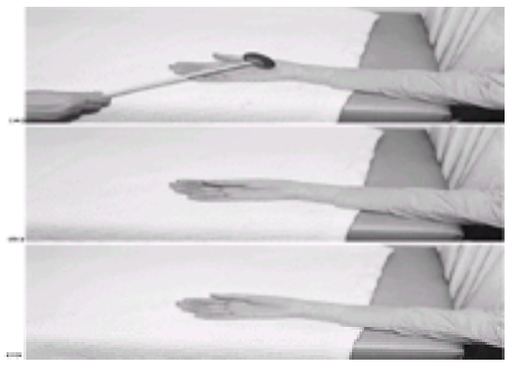

Fig. 2 Percussion myotonia. Following a sharp tap, the thenar eminence muscles contract and then relax slowly (subsequent photographs taken at 3-s intervals).

As a symptom, complaints relating to myotonia differ between patients with myotonic dystrophy, which is by far the most common cause of myotonia, and those with myotonia congenita. In myotonic dystrophy, even when grip myotonia is readily evident on examination, the patient may offer no symptoms. They are more likely to complain of hand weakness than of myotonia. When the myotonia is symptomatic, the patient complains of difficulty releasing objects after a tight grip. This is sometimes striking. One patient first noted grip myotonia in early adult life, when he was appointed as a teacher at a school—as his future headmaster shook his hand to congratulate him, he was embarrassingly unable to release his grip. In myotonic dystrophy, bulbar symptoms relating to myotonia are quite common—patients complain of their tongue or jaw 'locking' when speaking or swallowing and tongue myotonia on percussion may be demonstrated.

By contrast, in myotonia congenita weakness is absent and the myotonia, which is generalized, is problematic, particularly in the lower limbs. Patients complain of stiffness which is most evident on trying to initiate movement after rest. Thus, the patient who has been sitting in the waiting room rises and walks with profound leg stiffness, somewhat reminiscent of spasticity, into the consulting room. A classic presentation is the soldier on the parade ground—after a prolonged period 'standing to attention', the order to march results in his falling due to leg muscle stiffness. One such patient also demonstrated marked grip myotonia—on an unfortunate occasion he alighted from a bus but, unable to release his grip from the handrail before the bus departed, was dragged along the road.

In most disorders, myotonia lessens with repeated activity of the muscle. Thus, the sign becomes less striking with repeated percussion of the thenar eminence or attempts to demonstrate grip myotonia. As a symptom, for example, the leg stiffness in myotonia congenita lessens as the patient continues to walk. In paramyotonia the reverse is seen, with myotonia increasing with activity—so-called paradoxical myotonia. Some, but by no means all, patients complain that their myotonia is worse in the cold. This is again a particular characteristic of paramyotonia.

Classification of myotonic disorders

As with many other inherited neuromuscular disorders, nomenclature and classification are currently in a state of flux as molecular mechanisms are being unravelled. For clinical purposes a useful distinction is between those multisystem disorders in which weakness is a significant feature, and which are therefore referred to as dystrophies, and the non-dystrophic myotonias (Table 1).

The nomenclature for the gene loci of myotonic dystrophies has been revised recently. Classic myotonic dystrophy was previously called dystrophia myotonica, which gave rise to the abbreviation DM. It shows no genetic heterogeneity, all cases being associated with a trinucleotide repeat expansion in the 3' untranslated region of a

novel protein kinase gene on chromosome 19q. This locus is referred to as DM1. Many cases of proximal myotonic myopathy, a recently described disorder, are linked to chromosome 3q, and that locus is designated DM2. However, some cases of proximal myotonic myopathy do not link to DM2. As yet, other loci have not been identified, but as they are they will be designated DM3, DM4, and so on.

The most common non-dystrophic myotonias are the autosomal dominant and recessive forms of myotonia congenita, both of which are caused by mutations of the skeletal muscle chloride channel (*CLC-1*) gene. Different mutations of the skeletal muscle sodium channel gene (*SCN4A*) give rise to hyperkalaemic periodic paralysis and related disorders, including paramyotonia congenita. These chloride and sodium channelopathies, together with the calcium channel disorders causing hypokalaemic periodic paralysis, are discussed further in Chapter 24.22.5.

Schwartz–Jampel syndrome is a very rare recessive disorder of infantile onset, characterized by skeletal abnormalities, abnormal facial appearance, and abnormal muscle electrical activity. Electromyography may show typical myotonia as well as periods of continuous electrical activity which are probably neural in origin. It is genetically heterogeneous. Some cases are linked to chromosome 1p. One patient with a clinical diagnosis of Schwartz–Jampel syndrome was subsequently found to have a sodium channel mutation.

Myotonic dystrophy

Myotonic dystrophy is the most frequent cause of myotonia and indeed is also the most prevalent muscular dystrophy in adults. It is a multisystem disorder that has very important (but sometimes rather neglected) manifestations other than skeletal muscle dysfunction, involving cardiac conduction tissues, smooth muscle, eyes, and the central nervous system. Clinical severity ranges from death *in utero* to a condition so mild that it may be asymptomatic and without abnormal physical signs in old age. The molecular basis is now known (an expansion of an unstable trinucleotide repeat in a gene coding for a novel protein kinase), but it is not yet clear how this leads to the various manifestations of the disease. Myotonic dystrophy provides a dramatic example of the phenomenon of 'anticipation', by which succeeding generations may be much more severely affected than their predecessors, and this correlates with the size of the genetic expansion.

Epidemiology

The disease is seen worldwide, with a particularly high frequency in French Canadians in Quebec (originating from a single immigrant couple). Incidence and prevalence figures are unreliable, and probably mostly underestimates, because of the difficulty in ascertaining asymptomatic individuals. A generally accepted prevalence value is 5/100 000 population.

Pathogenesis

The molecular basis is the expansion of a trinucleotide (cytosine–thymine–guanine, **CTG**) repeat sequence in the 3' untranslated region of the myotonic dystrophy protein kinase (*DMPK*) gene on chromosome 19q. The function of this novel putative serine–threonine protein kinase is unknown. In the normal population the size of the repeat is in the range $CTG_{5–37}$, with a trimodal distribution of 5, 11 to 17, and 19 to 37 repeats. Expansions in the range $CTG_{37–49}$ are believed to represent premutations. Individuals with myotonic dystrophy have repeats in the range $CTG_{50–5000}$ and, as noted below, there is a correlation between the size of the repeat and clinical severity, and an inverse correlation between repeat size and age of onset.

A fundamental concept is that the expanded gene is unstable. It is mitotically unstable, and so the size of the gene increases with age. There is somatic mosaicism, so that the expansion is not the same size in different tissues. Diagnostic studies are based on measurement of the expansion size in blood lymphocyte DNA.

More important is intergenerational CTG-repeat instability, which explains why the disease tends to increase in severity in subsequent generations. The gender of the parent of origin is important. In most transmissions the allele size increases. However, there appears to be a threshold limit for sperm, and males never transmit the very large expansions associated with congenital myotonic dystrophy (see below), which only occurs when the mother is the gene carrier. There is some evidence of meiotic drive, which leads to preferred transmission of the abnormal expanded allele.

It remains uncertain as to how the basic molecular abnormality causes the various manifestations of the disease. One theory is that the expansion affects the expression of other genes in the DM locus apart from *DMPK*, such as the DM locus-associated homeodomain protein (*DMAHP*) gene. Another theory is that nuclear accumulation of RNA transcripts from the expanded *DMPK* gene affects processing of other mRNAs, thus having a knock-on effect on the translation of many proteins.

Clinical features

From the above it is apparent that there is a continuous distribution of expanded allele size, and that there is a relationship between allele size and disease severity, and between allele size and age of onset. While accepting that some patients will fall between these categories, for practical clinical purposes myotonic dystrophy can be considered to give rise to three main patterns of disease:

- congenital
- classic or early adult onset
- late onset, asymptomatic, or oligosymptomatic.

Because it is the best known, and illustrates the multifarious manifestations of myotonic dystrophy, the classic form will be discussed first.

Classic form

Onset is in adolescence or early adult life. The principal manifestations are summarized in Table 2. A number of rarer or clinically less important associations are also recognized. These include reduced fertility, testicular atrophy, insulin resistance (but rarely overt diabetes), retinopathy, eye movement disorder, peripheral neuropathy, disturbed tests of endocrine function, hypotension, pilomatrixomas, and reduced levels of immunoglobulins and complement.

Skeletal- and smooth muscle

The features of myotonia have already been discussed. The distribution of muscle weakness is highly characteristic. Wasting and weakness of the facial muscles, combined with premature male-pattern balding, gives rise to the typical facial appearance of the condition (Fig. 3). The temporalis muscle is atrophic, giving a sunken appearance over the temples. There is ptosis. Eye closure is weak and in severe cases the sclera may remain visible. The jaw tends to hang down. Neck flexion is weak and in some, but not all patients, there is evident atrophy of the sternomastoid muscles. In the limbs, and in marked contrast to most other myopathic disorders, the weakness is predominantly distal. In the upper limbs there is weakness and wasting of the small hand muscles and of the long wrist and finger flexor and extensor muscles in the forearm. There is often profound weakness of grip and the patient complains of difficulty with tasks such as wringing-out a cloth and removing the lid from a bottle. A simple hand-held dynamometer reveals the extent of the weakness—whereas a normal female would easily exceed 35 kg, patients of either sex may manage only 1 or 2 kg. In the lower limbs there is weakness of ankle dorsiflexion, presenting as tripping easily and foot drop. As the disease advances, weakness becomes evident more proximally, but the marked distal predilection remains throughout.

Fig. 3 Adult-onset myotonic dystrophy. Typical facial features (see text).

Bulbar muscle weakness presents with dysarthria and dysphagia. Smooth muscle involvement contributes towards the dysphagia. Symptoms akin to those of irritable bowel syndrome are frequent. Constipation is common, pseudo-obstruction rare. There may be evidence of incoordinate uterine contraction in labour.

Ocular

Cataracts develop at an early age. The initial manifestation is multicoloured opacities in the subcapsular regions, readily seen on slit-lamp examination. Identification of cataracts used to be important in screening asymptomatic family members for the disease, but that has now been replaced by DNA testing. In practice, the cataracts are managed as any other cataracts, being operated on when vision is significantly impaired. Early-onset cataracts should always raise the suspicion of myotonic dystrophy.

Central nervous system

Central nervous system disease is expressed in two main ways. As a group, patients with myotonic dystrophy have a lower IQ than average, but many mildly affected patients have intelligence within the normal range. They are often perceived as apathetic or lacking self-motivation. There is some neuropsychological evidence of specific defects of frontal lobe functioning. The second principal feature is excessive daytime sleepiness which affects over three-quarters of patients, some profoundly. This appears to be a central phenomenon and is only rarely attributable to obstructive sleep apnoea/nocturnal sleep disturbance.

Cardiovascular

Cardiovascular dysfunction is arguably the most important extramuscular manifestation of myotonic dystrophy and is probably responsible for most of the not infrequently reported cases of sudden unexpected death. The most commonly recognized pattern is of progressive conduction disturbance. Thus, in very early cases the ECG is normal. Subsequently, the P–R interval gradually lengthens until first-degree block is present. Later features include bundle branch and complete heart block. Tachyarhythmias also occur, most frequently atrial flutter or fibrillation. Symptoms include palpitation, dizzy spells, and fainting. Prolonged ECG monitoring and sometimes intracardiac electrophysiological studies are indicated if such symptoms are reported, or the standard ECG shows significant change. All patients should have an ECG annually, and be advised to report any cardiac symptoms immediately. Rhythm disturbances precipitated by anaesthesia or surgery are common, as are respiratory problems. For these reasons, patients should carry a medical alert bracelet/medallion and, for elective admissions for surgery, be reminded to inform the anaesthetist of their diagnosis. The latter is particularly important for asymptomatic individuals diagnosed on the basis of DNA studies following family screening, as they may not consider themselves to be at risk—they are. Although there is some correlation between cardiac involvement and overall severity of the myotonic dystrophy, it is not absolute and individuals with minimal muscle involvement may have significant ECG changes.

Heart muscle disease, as opposed to disordered cardiac conducting tissues, is not clinically significant and routine echocardiography is not required.

Respiratory

Recurrent chest infections are common and relate to respiratory muscle weakness and the tendency to aspirate. In advanced disease, death is often secondary to pneumonia. Respiratory insufficiency may become apparent following anaesthesia, with difficulty in weaning from the ventilator. Chronic hypoventilation and hypercapnia may cause excessive daytime sleepiness, but in practice is much less common than the presumed central mechanism already mentioned. However, it must be considered and excluded (for example by overnight oximetry and blood gas measurements) if felt to be a possibility. Particular warning features would include a history of disturbed night-time sleep, snoring, waking with headaches, and the development of secondary polycythaemia.

Congenital myotonic dystrophy

By simple definition, this form of myotonic dystrophy is evident at birth, but in reality the spectrum of early-onset myotonic dystrophy is much wider. The exclusive (with only very rare exceptions) maternal transmission of congenital myotonic dystrophy has already been discussed. Many fetuses carrying large expansions are aborted spontaneously in early pregnancy and there is a high rate of fetal wastage. Because of the unstable nature of the CTG-repeat and the associated phenomenon of anticipation, it is not uncommon for the mother to be unaware of her own diagnosis at the time of birth. In that situation, the diagnosis in the infant is not always immediately apparent, as there are no entirely specific clinical features. In other infants the disorder is not evident at birth and thus is not strictly congenital, but symptoms become apparent within the first few years of life.

There is often a history of polyhydramnios and poor fetal movement in the pregnancy. The child is born hypotonic ('floppy') and talipes is present in about one-half. Respiratory and feeding difficulties may necessitate assisted ventilation or an oxygen tent, and feeding by nasogastric tube. Some die in the neonatal period from respiratory complications, but somewhat surprisingly, there are few further deaths in the survivors until the late teens and early adult life. There is generalized weakness, including the face—the jaw hangs open and the mouth has a characteristic tented or carp-like (as in fish) appearance. Myotonia is not evident clinically and even electromyographically may not appear for several years.

In those who survive, hypotonia resolves and motor function improves over the following few years, but during adolescence the features of the classic adult form of the disease appear (Fig. 4).

Fig. 4 Myotonic dystrophy. The affected mother's two children have the congenital form of the disease.

Mental retardation is invariable and may be severe. Most require special-needs schooling. Bowel involvement is common with faecal soiling and irregular bowel habit. Curiously, cataracts are relatively uncommon.

The overall prognosis is poor. Some 25 per cent die in the first 18 months of life, most in the neonatal period. One-half survive into the mid-30s, death most commonly resulting from respiratory involvement, but with a proportion of sudden deaths almost certainly due to cardiac conduction defects. Few achieve an independent adult life.

Late-onset form

This form is associated with a small CTG-repeat expansion. It is typically asymptomatic or oligosymptomatic and is diagnosed during family studies or by an alert ophthalmologist when the patient presents with cataracts. Skeletal muscle disease may be absent, or confined to mild myotonia and weakness confined to the hands. Balding may be a feature. It is not uncommon to see the parents of a patient with the classic adult form of the disease and not be able to identify the transmitting parent on clinical examination.

Importantly, even patients with such minimal symptoms may occasionally develop significant cardiac conduction problems and they should have annual electrocardiograms.

Management

The essential management issues in myotonic dystrophy are:

- genetic counselling
- annual ECG
- anaesthetic risks
- physical therapies
- cataract surgery.

A particular concern relates to the genetic phenomenon of anticipation and the potential for an asymptomatic mother, ignorant of the diagnosis, to give birth to a congenitally affected child. When the diagnosis of myotonic dystrophy is established in a family member it is imperative that at-risk relatives are offered screening. Prenatal diagnosis, by chorionic villous sampling, can then be offered.

Annual ECG should be performed in all patients. They and their medical attendants must be aware of the cardiorespiratory complications associated with anaesthesia. They should be encouraged to wear an appropriate medical alert bracelet or medallion. A few patients require nocturnal positive-pressure ventilation by face mask, but most excessive daytime sleepiness is not related to respiratory insufficiency. Recurrent chest infections are common. Annual influenza vaccination should be advised. Pneumococcal vaccination is also given but is of uncertain value.

Physiotherapy and occupational and speech therapy all have a role, as does the use of orthotic devices (for example for foot drop). Bowel problems in the congenital form require specific advice and counselling.

Cataract surgery is required when vision is significantly impaired.

Proximal myotonic myopathy

In the last few years a number of families have been described with an autosomal dominant multisystem disorder similar to myotonic dystrophy, but who do not have the chromosome 19 CTG expansion. Many, but not all, show linkage to chromosome 3q. In some countries (such as Germany) these conditions seem to be almost as common as myotonic dystrophy, but in others (such as the United Kingdom) they appear to be rare. Some issues concerning classification were discussed earlier.

Various names and abbreviations have been given to these conditions, including proximal myotonic myopathy, proximal myotonic dystrophy and DM2.

Despite the superficial similarities to myotonic dystrophy, there are also differences. Onset is usually in adult life. Myotonia tends to be more symptomatic than in myotonic dystrophy and patients complain of stiffness and muscle pain. Unlike myotonic dystrophy, the weakness tends to be more marked proximally, and may show significant fluctuation from day to day. Cataracts may be indistinguishable from those seen in myotonic dystrophy. Cardiac conduction problems appear to be less common. Male hypogonadism and deafness occur. There is considerable clinical variability even between families linked to the same chromosome 3q locus, indicating allelic heterogeneity.

Further reading

Myotonic dystrophy

Brook JD *et al*. (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat in the 3' end of a transcript encoding a protein kinase family member. *Cell* **68**, 799–808.

Harper P (2001). *Myotonic dystrophy*, 3rd edn. WB Saunders, London.

Koch MC *et al.* (1991). Genetic risks for children of women with myotonic dystrophy. *American Journal of Human Genetics* **48**, 1084–91.

Lazarus A *et al*. (1999). Relationships among electrophysiological findings and cardiac status, heart function, and extent of DNA mutation in myotonic dystrophy. *Circulation* **99**, 1041–6.

Reardon W *et al.* (1993). The natural history of congenital myotonic dystrophy: mortality and long term clinical aspects. *Archives of Disease in Childhood* **68**, 177–81.

Proximal myotonic myopathy

Moxley RT (1996). Proximal myotonic myopathy: mini-review of a recently delineated clinical disorder. *Neuromuscular Disorders* **6**, 87–93.

Ranum LPW *et al.* (1998). Genetic mapping of a second myotonic dystrophy locus. *Nature Genetics* **19**, 196–8.

Udd B *et al*. (1997). Proximal myotonic dystrophy—a family with autosomal dominant muscular dystrophy, cataracts, hearing loss and hypogonadism: heterogeneity of proximal myotonic syndromes? *Neuromuscular Disorders* **7**, 217–28.

Wieser T *et al.* (2000). A family with PROMM not linked to the recently mapped PROMM locus DM2. *Neuromuscular Disorders* **10**, 141–3.

24.22.4 Metabolic and endocrine disorders

David Hilton-Jones and Richard Edwards

Introduction

This section deals with disorders of voluntary muscle that arise either as the result of a disturbance of muscle metabolism, or disordered ion flux. In many cases precise mechanisms have yet to be defined.

The term 'metabolic myopathy' is applied to those disorders in which there is a primary defect, usually an enzyme deficiency, in the biochemical pathways associated with energy generation (ATP synthesis). This group includes the mitochondrial disorders, which are some of the most common causes of primary metabolic myopathy seen in clinical practice.

Endocrine myopathies and nutritional and toxic myopathies, including those that are drug-induced, can be considered as secondary (acquired) metabolic myopathies.

Defects in genes coding for subunits of the skeletal-muscle sodium and calcium channels underlie primary hyperkalaemic and hypokalaemic periodic paralysis, respectively. Both autosomal dominant and autosomal recessive myotonia congenita are caused by mutation in the skeletal-muscle chloride channel. Mutations affecting two skeletal-muscle calcium channels, the dihydropyridine (**DHP**) and ryanodine (**RYR1**) receptors, are associated with malignant hyperthermia (**MH**). The congenital myopathy central core disease is allelic to MH and is associated with *RYR1* mutations.

The cardinal symptoms of myopathy are weakness, fatigue, and/or pain; altered excitability may also occur. It is important that the physician appreciates several points:

- There are non-specific effects, such as loss of muscle, that may be far more important as a cause of weakness than the energetic consequences of the biochemical defect. Visual inspection and circumference measurements tend to underestimate the extent of wasting, which may be better documented by quantitative scanning methods (magnetic resonance imaging (**MRI**) or computed tomography (**CT**)).
- Not all the biochemical abnormalities cause symptoms. Clinical expression of the underlying defect depends on the habitual demands on the muscle for movement and weight-lifting.
- A patient with a metabolic myopathy may have common, non-myopathic, musculoskeletal complaints which have no relation to the inherited or acquired defect.
- Muscle symptoms may have no physiological connection with the underlying defect and may be consequences of somatization or other psychological process.
- The practical assessment of metabolic myopathy should include consideration of the WHO ICIDH-2 (2000) classification of the functioning and disability criteria of impairment, activities, and participation (revised from the International Classification of Impairments, Disabilities, and Handicaps (**ICIDH**) of WHO 1980). In this generic consideration, the relationship between antigravity muscle strength and the body weight to be carried is crucial: performance may be improved as much or more by weight reduction as by therapeutic attempts to reverse the myopathy, providing that calorie restriction does not aggravate the metabolic defect—for example, in the case of carnitine palmitoyl transferase deficiency, where carbohydrate starvation may exacerbate the energy supply problem of the underlying enzyme defect.
- An objective assessment of a response to treatment requires the measurement of individual muscle strength and/or timing of the performance of tasks relevant to the patient's symptoms and the everyday life demands placed on the diseased muscles.
- Metabolic myopathies are unusual or rare conditions that are very variable in presentation. They are not easy to discuss in the light of current, evidence-based healthcare philosophies, which are largely based on the results of randomized controlled trials (**RCTs**) of therapeutic interventions. The treatments of the metabolic myopathies tend to fall under the general rubric of 'orphan drugs' and 'orphan diseases', since, as with other rare diseases, a commercial return on the investment in research and development to deliver effective treatments is unlikely. Furthermore, in view of their rarity, there is little or no chance of formal treatment evaluation by RCTs. These conditions are therefore still to be evaluated by thoughtful clinical research employing the most relevant modern biochemical and physiological approaches.
- The patient with a metabolic myopathy is a person, and therefore far more important and complex to understand and help than the underlying metabolic diagnosis, difficult though that may be. It is essential to the humane and effective management of such a patient to see the individual as coping in a personal and social sense despite the metabolic impairment. The aim is to determine what is likely to best improve the patient's overall quality of life. Here, as with other disabilities, the constructive analysis and recommendations of the World Health Organization are useful as a basis for working with the patient to determine an individual management plan. (Table 1).

Primary metabolic myopathies

The principal energy currency of living cells is adenosine triphosphate (**ATP**). Whereas in most organs the rate of ATP utilization is fairly constant, in voluntary muscle the change from rest to strenuous activity may increase the demand on ATP generation several thousand-fold. If that demand is not met, contractile failure (that is to say, fatigue or weakness) will develop and may be accompanied by the destruction of muscle fibres. In many of the primary metabolic myopathies it is often assumed that exercise-induced symptoms relate to a failure of ATP generation, and although this is probably not always correct, it is a useful generalization. Whilst exercise-induced symptoms are often a striking feature of this type of metabolic myopathy, they are not always present. Some patients develop a chronic progressive myopathy.

The main fuels providing energy for ATP generation in skeletal muscle are glycogen, fatty acids, and glucose (Fig. 1). Their relative contributions depend upon the state of nutrition and, more importantly, the level and duration of exercise. A gross oversimplification of these pathways aids understanding of the clinical features of the different forms of metabolic myopathy.

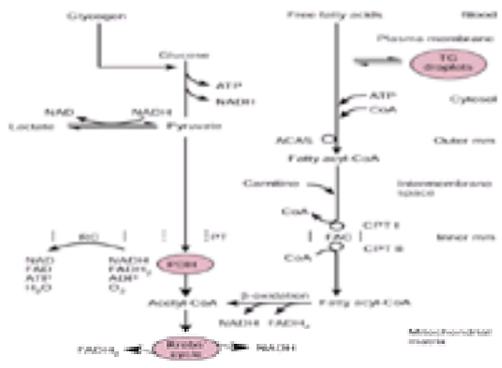

Fig. 1 Major pathways associated with energy production in skeletal muscle. ACAS, acyl-CoA synthetase; ADP, adenosine diphosphate; ATP, adenosine triphosphate; CoA, coenzyme A; CPT, carnitine palmityl transferase; FAC, fatty acyl-carnitine; FAD, flavin adenine dinucleotide; $FADH_2$, reduced FAD; mm, mitochondrial membrane; NAD, nicotinamide adenine dinucleotide; NADH, reduced NAD; PDH, pyruvate dehydrogenase complex; PT, pyruvate translocase; RC, respiratory chain; TG, triglyceride.

At rest, the main fuel source is circulating free fatty acids, with a lesser contribution from circulating glucose. Small amounts of ATP may be generated directly from glycolysis, but the production of the energy-rich electron carriers (reduced nicotinamide adenine dinucleotide (NADH) and reduced flavin adenine dinucleotide ($FADH_2$)) from fatty acid b-oxidation, and the Krebs' cycle are more important. Transfer of electrons to molecular oxygen through the electron transport chain of the mitochondria releases energy for the generation of ATP (oxidative phosphorylation)

The increased demand on ATP generation during early strenuous exercise cannot be met by oxidative pathways. The resting blood flow provides an inadequate delivery of oxygen and substrate, and compression of blood vessels by the contracting muscle exacerbates the problem. ATP is therefore generated by the breakdown of muscle-fibre stores of glycogen (anaerobic glycolysis). The relative lack of oxygen leads to increasing levels of NADH and pyruvate. NADH accumulation would inhibit glycolysis, and thus ATP generation, and is avoided by the reduction of pyruvate to lactate, explaining the lactic acidosis seen in disorders of oxidative metabolism.

Adaptive processes occur as exercise continues; muscle blood flow increases, the respiratory rate rises, and free fatty acids are mobilized from adipose stores. Glycogen stores in muscle become depleted and circulating free fatty acids become the main energy source, with a very small contribution from circulating glucose.

Certain deductions can be made from the above that are largely borne out in clinical practice. Disorders of glycogen and glucose metabolism are typically asymptomatic at rest, but produce symptoms early in exercise when anaerobic glycolysis is important for energy supply. If low levels of exercise can be sustained, symptoms can improve as fatty acid oxidation increases ('second wind' phenomenon in McArdle's disease). Disorders of fatty acid metabolism, insufficient to cause symptoms at rest, are likely to be exposed by sustained exercise and fasting. The central role of oxidative phosphorylation explains why disorders of the respiratory chain may be symptomatic at rest. The clinical presentation will also depend upon whether the enzyme defect is restricted to skeletal muscle or is more generalized, thereby causing dysfunction of other tissues and organs. Systemic features may dominate in disorders of b-oxidation and in mitochondrial disorders but are absent in McArdle's disease because the defective enzyme is muscle-specific.

Disorders of glycogen and glucose metabolism (see also Section 11.3)

Several of the glycogenoses show significant skeletal-muscle involvement. The major pathways of metabolism, and the enzymes associated with these disorders, are shown in Fig. 2. They are autosomal recessive disorders, except for the X-linked recessive, phosphoglycerate kinase deficiency. In most of these disorders the serum creatine kinase is elevated at rest, and massively so after exercise-induced muscle damage.

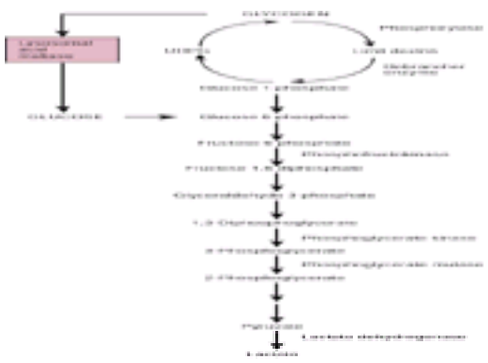

Fig. 2 Pathways of glycogenolysis and glycolysis. Enzymes known to be associated with particular clinical syndromes are shown.

Acid maltase deficiency (type II glycogenosis)

Acid maltase is a lysosomal enzyme not directly involved in energetic pathways, and exercise-induced symptoms are absent. In the infantile form (Pompe's disease) there is widespread organomegaly as well as skeletal-muscle involvement, and death occurs by the age of 2 years due to cardiac or respiratory failure. The adult form is of considerable importance and has probably been underdiagnosed. The most obvious feature is a slowly progressive, painless, proximal myopathy. Diaphragmatic involvement is an important characteristic, and many of these patients first present with respiratory failure. Nocturnal assisted ventilation alleviates sleep-disordered breathing and may prolong survival for many years. Muscle biopsy showing glycogen-containing vacuoles is usually suggestive; but the definitive diagnosis is established by enzyme assay in muscle, fibroblasts, or leucocytes, or by demonstrating glycogen granules by Periodic acid–Schiff (**PAS**) staining in lymphocytes on a peripheral blood film.

There is an intermediate juvenile form, with limited survival.

Myophosphorylase deficiency (type V glycogenosis—McArdle's disease)

The onset of symptoms is usually during childhood, and the cardinal features are pain, weakness, and stiffness of muscles early in exercise, relieved by rest. Strenuous exercise, such as helping to push a car or lift heavy furniture, may induce painful, electrically silent, muscle contractures. Muscle fibre breakdown is reflected in myalgia and myoglobinuria (dark red/black urine), which, if severe, may cause renal failure. Muscle breakdown is accompanied by a large release of creatine kinase (**CK**) into the blood, and a failure to see such a rise in serum CK levels should cast doubt on a diagnosis of myoglobinuria. Conversely, if renal failure is present then no myoglobinuria may be seen and the only evidence of rhabdomyolysis is the raised CK level. However, symptoms may ease ('second wind' phenomenon) if low levels of activity are maintained, as circulating free fatty acids and glucose become available as alternative fuels.

Progressive proximal weakness frequently develops and is sometimes the mode of presentation in late-onset cases.

Failure of lactate generation (accompanied by increased blood ammonia and hypoxanthine concentrations) during ischaemic forearm exercise is consistent with the diagnosis. However, this is not specific since it also occurs in other glycogenolysis disorders, and may be seen to some extent in acquired conditions such as alcoholic myopathy or hypothyroidism. Also, the test may give a misleading ('false-negative') result if the myophosphorylase deficiency is only partial. The definitive

diagnosis is established by histochemical demonstration of the absence of phosphorylase staining (or by enzyme assay) on muscle biopsy, or by genetic studies of the coding and expression of muscle phosphorylase.

Debrancher enzyme deficiency (type III glycogenosis—Cori–Forbes disease)

In infancy and childhood the main features of this disorder are hepatomegaly, hypoglycaemia, and failure to thrive. During adolescence muscle symptoms become more prominent. A small group of patients first present during adult life with muscle symptoms, but may give a history of a protuberant abdomen in childhood. Both exercise intolerance and a slowly progressive proximal myopathy are present.

It has recently been recognized that most patients develop a potentially fatal cardiomyopathy.

The ischaemic forearm test shows impaired lactate generation, muscle biopsy shows glycogen accumulation, and the administration of glucagon fails to produce a hyperglycaemic response. Enzyme assay can be performed on muscle, liver, erythrocytes, and leucocytes.

Phosphofructokinase deficiency (type VII glycogenosis—Tarui's disease)

The clinical picture is very similar to that of myophosphorylase deficiency, but a phosphofructokinase (**PFK**) deficiency in erythrocytes leads to the additional features of haemolytic anaemia and gout. Unlike patients with myophosphorylase deficiency, ingested glucose does not improve exercise tolerance in those with PFK deficiency because of the position of PFK in the sequence of enzymes in the glycolytic pathway (Fig. 2). Diagnosis is established by enzyme assay in muscle.

Defects of distal glycolysis

Deficiencies of phosphoglycerate kinase, phosphoglycerate mutase, and lactate dehydrogenase have been described recently. All three are associated with exercise intolerance and myoglobinuria. It is possible that other defects of glycolysis, causing similar symptoms, remain to be discovered.

Treatment

There is, as yet, no specific treatment for any of the disorders described above. Attempts at dietary manipulation have generally proved unsuccessful. Patients must be aware of the risk to renal function from myoglobinuria, and try to avoid intense exercise. There is evidence, in patients with muscle pain due to McArdle's disease and other metabolic myopathies, that maintaining a reasonable level of aerobic fitness is beneficial, by sustaining sufficient activity of muscle mitochondria to provide energy from oxidative phosphorylation to adapt for the deficiencies in energy availability from glycogenolysis.

Disorders of lipid metabolism

Unlike glycolysis, lipid metabolism is entirely dependent upon oxidative processes. Moreover, there is a close relationship between the disorders described below and defects of the mitochondrial respiratory chain; for example, lipid accumulation in muscle is a common histological feature in respiratory chain disorders.

Free fatty acids, mainly from the blood but also from triglyceride droplets stored within muscle fibres, are a major fuel at rest and during sustained exercise (see Fig. 1). They are converted to fatty acyl-coenzyme A (**CoA**) at the outer mitochondrial membrane which, within the mitochondrial matrix, can undergo b-oxidation. A transport system involving carnitine and the enzyme system carnitine palmityl transferase (**CPT**) is required to enable fatty acyl-CoA to cross the inner mitochondrial membrane. Defects involving carnitine, CPT, and b-oxidation are recognized.

Carnitine deficiency

Muscle carnitine deficiency, causing fluctuating weakness and lipid storage (intracellularly in myocytes), and systemic carnitine deficiency causing weakness and recurrent, often fatal, Reye-like episodes, have been described. It has become apparent that in most cases of so-called systemic carnitine deficiency, and perhaps even in some purely myopathic cases, the carnitine deficiency is secondary to another metabolic disorder: most commonly defects of b-oxidation or of the mitochondrial respiratory chain.

Defects of b-oxidation

Many enzyme deficiencies have been described, but clinical features are limited. They may present during the neonatal period with hypotonia, hypoglycaemia, cardiomyopathy, failure to thrive, and early death. Such defects may be a cause of some cases of sudden infant death syndrome. Later-onset cases develop Reye-like crises, muscle weakness, and cardiomyopathy. Secondary carnitine deficiency is common. A high carbohydrate and low fat diet may help.

Carnitine palmitoyl transferase deficiency

This rare disorder shows a male predominance. Symptoms are precipitated by sustained exercise (for example, a route march) or prolonged fasting, and consist of muscle pain followed by myoglobinuria, which may cause renal failure. The diagnosis may be suggested by showing impaired ketone body production during fasting, but is proven by enzyme assay in skeletal muscle. A high carbohydrate, low fat diet may reduce the number of attacks.

Myoadenylate deaminase deficiency

Deficiency of myoadenylate deaminase has been suggested as a cause of exercise-induced myalgia, weakness, and cramps but its exact status remains controversial. It is almost certainly an unusual cause of significant myalgia. It has been described as an incidental finding in muscle needle biopsies taken from normal volunteers to study muscle chemistry in sports science research. The enzyme catalyses the reaction adenosine monophosphate (**AMP**) ® inosine monophosphate (**IMP**) + ammonia (NH3). Theoretically, this reaction may aid ATP production by removing AMP and increasing flux through the adenylate kinase reaction 2ADP® ATP + AMP. The diagnosis is established from the absence of a rise in the plasma ammonia level during forearm exercise testing and from the histochemical demonstration of absent enzyme activity.

Endocrine myopathies

Although weakness is a common symptom in many endocrine disorders, the mechanisms are generally poorly understood. However, the myopathy responds to treatment of the underlying hormonal disorder, and extensive investigation of the myopathic component is rarely required. The commonest pattern is limb-girdle weakness.

Thyroid disorders (see also Chapter 12.4)

Thyrotoxicosis

Typically, weakness develops shortly after the onset of other thyrotoxic symptoms, and 80 per cent of patients have demonstrable weakness at presentation. The shoulder-girdle muscles tend to be involved before the pelvic musculature. Muscle atrophy is usually slight. Asymmetrical and distal weakness, myalgia, cramps, and fasciculations are rare findings.

The serum creatine kinase level is usually normal, but electromyography shows features consistent with muscle disease. The myopathy responds to treatment of the thyrotoxicosis.

Thyrotoxic periodic paralysis

Most cases have been reported in individuals from the Orient, with a strong male predominance. Clinical features closely mimic those of familial hypokalaemic periodic paralysis. The weakness is disproportionate to any muscle wasting. The onset of paralytic attacks usually follows the development of hyperthyroid symptoms but the attacks cease when the patient is rendered euthyroid.

Thyroid ophthalmopathy (Graves' ophthalmoplegia)

The classic features of this condition include eyelid lag, retraction, and swelling, as well as progressive swelling of the extraocular muscles and orbital soft tissues leading to proptosis and diplopia and, in severe cases, corneal ulceration, papilloedema, and optic atrophy. An extremely important but often missed variant is the patient who presents with minimal diplopia only.

In mild cases, MRI or CT imaging is useful for detecting extraocular muscle swelling. Simple tests of thyroid function may be normal. Estimation of antithyroglobulin and antimicrosomal antibodies, and the performance of a thyrotrophin-releasing hormone (**TRH**) stimulation test may be required. Thyroid-stimulating immunoglobulins are present in most patients.

If thyrotoxic, the patient should be rendered euthyroid. Lid retraction may respond to topical 10 per cent guanethidine. Persisting major eye problems may require high-dose prednisolone, plasma exchange, or orbital decompression. Tarsorrhaphy protects the cornea.

Thyroid disease and myasthenia

Patients with myasthenia gravis have an increased incidence of thyroid disease, including hyperthyroidism, hypothyroidism, Hashimoto's thyroiditis, and increased antibodies to thyroglobulin or microsomal fractions. Thyroid disease may pre-date or follow the onset of myasthenia and must be considered as a cause of deterioration in an otherwise stable patient with myasthenia. Some 5 per cent of patients with myasthenia will develop thyroid disease, but only about 0.1 per cent of thyrotoxic patients develop myasthenia.

Hypothyroidism

Although hypothyroid myopathy may be asymptomatic, mild weakness is probably present in most patients. Muscle biopsy characteristically shows evidence of type II (fast twitch, glycolytic, high intrinsic force) muscle fibre atrophy with type I fibre dominance. Even in the absence of weakness the serum creatine kinase level is often markedly raised. Slow relaxation of the tendon jerks may be present in isolation. Muscle pain and cramps are common. In children, the combination of hypothyroidism, weakness, and muscle hypertrophy is referred to as the Kocher–Debré–Semelaigne syndrome. In adults, Hoffman's syndrome describes the combination of hypothyroidism, weakness, muscle hypertrophy, cramps, and myoedema (the formation of a localized ridge of muscle following direct percussion). They probably represent variants of the same disorder.

All hypothyroid myopathic symptoms respond to thyroxine replacement.

Pituitary–adrenal axis disorders

Clinically, the most important of these is iatrogenic steroid myopathy, discussed below under 'Glucocorticoid excess'.

Acromegaly

Proximal weakness, pelvic more than shoulder girdle, is present in about one-half of patients. Common complaints include tiredness, weakness, and myalgia; muscle wasting is slight. Serum creatine kinase levels are normal or slightly raised. Normalizing growth-hormone levels improves the myopathy, but recovery may be incomplete.

Hypopituitarism

Growth-hormone deficiency in childhood impairs muscle and skeletal development proportionately; weakness is not usually a feature. In adults, panhypopituitarism causes generalized weakness and fatigue, which usually responds to thyroxine and cortisone-replacement therapy. Replacement of growth hormone in growth hormone-deficient adults has been associated with varying degrees of improvement in the strength of wasted muscles.

Glucocorticoid excess

ACTH excess, from a functioning pituitary adenoma or from ectopic production, is usually associated with high glucocorticoid levels, producing pituitary or ectopic Cushing's syndrome. Weakness is common and thought to relate to glucocorticoid excess. Weakness may occur in Nelson's syndrome, in which there is a high level of ACTH but no glucocorticoid excess.

The myopathy associated with Cushing's syndrome is probably related to glucocorticoid excess, and the clinical features are essentially the same as those of iatrogenic steroid myopathy. The 9a-fluorinated steroids, including dexamethasone, triamcinolone, and betamethasone, appear to have the greatest myopathic potential. Topical steroids can cause myopathy.

The most common picture is of a slowly progressive limb-girdle wasting and weakness, pelvic more than shoulder girdle, often accompanied by myalgia. The drug-induced form may have a more acute onset. Myopathy without other features of glucocorticoid excess is unusual. The serum creatine kinase level is usually normal and muscle biopsy shows non-specific type II fibre atrophy.

Steroid withdrawal is followed by recovery over several months. If steroid therapy for the primary disorder has to be continued then a non-fluorinated compound such as prednisolone should be used, preferably on an alternate-day basis. Successful treatment of Cushing's syndrome leads to recovery.

Conn's syndrome

Weakness is present in about 75 per cent of patients and is due to the associated hypokalaemia. Secondary hypokalaemic periodic paralysis may occur.

Addison's disease

Weakness, fatigue, and myalgia occur in up to one-half of patients. Rare myopathic presentations include progressive flexion contractures and secondary hyperkalaemic periodic paralysis.

The serum creatine kinase level is normal or slightly increased. Glucocorticoid replacement therapy is curative.

Disorders of calcium, vitamin D, and parathyroid hormone metabolism (see also Chapter 12.6)

There are complex interactions between vitamin D metabolism, calcium and phosphate homeostasis, and parathyroid hormone activity. Myopathy occurs in several clinical situations, but the precise pathophysiological mechanisms are unclear.

Osteomalacia

Weakness is the presenting symptom in one-third of patients, affecting predominantly the pelvic girdle musculature. Bone pain is prominent. The serum creatine

kinase level is usually normal. Muscle biopsy may show type II fibre atrophy, sometimes severe.

The pain responds fairly rapidly to vitamin D treatment, but the weakness recovers more slowly and may be incomplete.

Primary hyperparathyroidism

Myalgia, stiffness, and complaints of fatigue are common, but overt weakness is rare. Symptoms resolve when the underlying parathyroid adenoma is removed and serum calcium levels fall.

Renal osteodystrophy

Endstage renal failure is frequently accompanied by a predominantly pelvic girdle myopathy, sometimes with buttock and thigh pain. Symptoms respond to dialysis, transplantation, or vitamin D treatment.

Dialysis osteodystrophy

Some patients undergoing dialysis develop a severe myopathy with bone pain, fractures, and vitamin D resistance. It probably relates to aluminium toxicity. Fatigue and muscle weakness are common. Objective muscle testing is needed to distinguish true changes in muscle function from the non-specific causes of fatigue and ill health seen in patients on dialysis.

Ischaemic myopathy

Rarely, a painful ischaemic myopathy with arterial narrowing due to calcium deposition complicates renal failure. Skin ulceration and bowel infarction may also occur.

Nutritional and toxic myopathies

Whilst malnutrition causes muscle wasting, specific myopathic effects of nutritional deficiencies are uncommon, a notable exception being vitamin D deficiency, discussed above. Myopathies due to ingested toxins are relatively more common than the inherited metabolic myopathies and include those due to alcohol, and therapeutic-drug excess or idiosyncrasy.

Alcoholic myopathies

Chronic alcoholics may develop subacute or slowly progressive, proximal muscle weakness with mild-to-moderate wasting, and with muscle-biopsy evidence of type II fibre atrophy, mainly affecting the lower limbs. Occasionally the wasting is more generalized, since alcoholism may be associated with neurogenic muscle atrophy secondary to concomitant thiamin deficiency and more generalized malnutrition. It is thus still debated whether the so-called chronic alcoholic myopathy is purely myopathic, neuropathic, or both, and whether the cause is a direct toxic effect of alcohol or a secondary phenomenon, perhaps relating to malnutrition. Abstinence may lead to some degree of recovery.

Much more dramatic is acute alcoholic myopathy ('alcoholic rhabdomyolysis'), which usually occurs during or shortly following a binge. There may be widespread cramps, pain, and weakness. However, the most striking feature is the development of extremely painful muscle swelling, which may be localized or generalized. Myoglobinuria presents a threat to renal function, and hyperkalaemia may be present in severe cases. The serum creatine kinase is elevated and muscle biopsy shows acute necrosis. Recovery, which may be incomplete, occurs over several weeks.

Vitamin E deficiency

Vitamin E deficiency probably causes a myopathy, but interpretation is confused by the presence of additional neurological problems including neuropathy and ataxia.

Drug-induced myopathies

Drug-induced neuromuscular disorders are common, under-recognized and under-reported. Numerous drugs have been implicated, several mechanisms are responsible (Table 2), and some drugs can affect both muscle and peripheral nerves (for example, vincristine, D-penicillamine, and perhexiline).

Skeletal-muscle channelopathies

The term 'channelopathy' has come into common usage since the last edition of this textbook. Ion channels may be ligand-gated or voltage-gated. In the field of muscle diseases, the most important ligand-gated channel is the skeletal-muscle nicotinic acetylcholine receptor, at the neuromuscular junction. Antibody-mediated destruction underlies acquired myasthenia gravis, whereas inherited mutations of genes coding for the subunits of the receptor are the basis of several forms of congenital myasthenic syndrome. Acquired neuromyotonia and Lambert–Eaton myasthenic syndrome are caused by antibody-mediated damage to the voltage-gated potassium and calcium channels, respectively, of the terminal axon, and are discussed, together with myasthenia gravis and the myasthenic syndromes, in Chapter 24.17.

The following section is concerned with inherited disorders of skeletal-muscle voltage-gated sodium, calcium, and chloride channels. In passing, it should be noted that channelopathies are not confined to muscle, and note was made above of two neuronal channelopathies. Other disorders caused by an inherited channel defect include certain forms of epilepsy (nocturnal frontal lobe epilepsy, benign neonatal convulsions), episodic ataxia, hemiplegic migraine, deafness, night-blindness, cardiac long-QT syndromes, and nephrolithiasis.

Periodic paralyses

Marked hypokalaemia and hyperkalaemia from whatever cause may produce weakness or paralysis (secondary periodic paralysis). The primary periodic paralyses are familial, dominantly inherited disorders characterized by recurrent attacks of paralysis. These have previously been subdivided into hyperkalaemic, hypokalaemic, and normokalaemic forms on the basis of changes in the serum potassium level during attacks. Recent evidence has shown that the primary abnormality in the hyperkalaemic and normokalaemic forms is a mutation affecting the adult skeletal-muscle sodium channel, whereas the hypokalaemic form is caused by a mutation affecting the skeletal-muscle calcium channel.

Hypokalaemic periodic paralysis

Attacks usually start during the second decade of life and then vary in frequency from daily to years between episodes. Weakness may be present on waking or develop during the day, typically in response to a heavy carbohydrate meal or during rest after strenuous exercise. The weakness involves the legs more than the arms, proximal muscles more than distal, and it may be asymmetrical. Bulbar and respiratory muscle weakness is rare. Attacks last from hours to several days. The tendon reflexes may be depressed or lost during an attack. Permanent and progressive proximal weakness often develops by middle age. The serum potassium level typically falls during an attack, but not necessarily outside the normal range.

The disorder is caused by a mutation in the *CACNA1S* gene (on chromosome 1) encoding the dihydropyridine receptor (**DHPR**) component of the skeletal-muscle calcium channel. The DHPR is located within the transverse tubular system, and acts as a voltage-sensor for the ryanodine receptor (RYR1) component of the calcium channel, which is located in the sarcoplasmic reticulum and is responsible for triggering calcium release and thus muscle contraction. Different mutations in the same gene, and mutations in the *RYR1* gene, are associated with malignant hyperthermia (see below).

Acetazolamide is the treatment of choice to prevent attacks. Acute attacks respond to oral potassium, given as an unsweetened aqueous solution.

Apparently-identical attacks may occur in association with thyrotoxicosis and resolve when the patient is rendered euthyroid.

Hyperkalaemic periodic paralysis

Attacks tend to start at an earlier age than in the hypokalaemic form, and do not last as long. Precipitants include cold, fasting, rest after exercise, pregnancy, alcohol intake, and potassium loading. Readily utilized carbohydrate sources, such as a sweet drink, may abort an attack. A progressive proximal myopathy may also develop. Myotonia is present in some patients (see below). The serum potassium level may rise during an attack, but the change is often slight.

The underlying abnormality is a mutation within the *SCNA4* gene (on chromosome 17) encoding the a-subunit of the skeletal-muscle sodium channel.

Mild attacks respond to carbohydrate ingestion. Kaliuretic diuretics usually prevent attacks.

Paramyotonia congenita

Paramyotonia congenita describes a dominantly inherited condition characterized by cold-induced weakness and muscle stiffness (paramyotonia), and which is sometimes accompanied by periodic paralysis. The relationship between this disorder and primary hyperkalaemic periodic paralysis had been much debated, but recent evidence has shown that hyperkalaemic periodic paralysis, hyperkalaemic periodic paralysis with myotonia, paramyotonia congenita, and paramyotonia congenita with periodic paralysis are allelic disorders involving the *SCNA4* gene (on chromosome 17) encoding the a-subunit of the skeletal-muscle sodium channel.

Myotonia congenita

Autosomal dominant (Thomsen's disease) and recessive (Becker-type) forms of this condition are recognized, with the recessive type being much commoner. Onset tends to be earlier in the dominant form but both usually become apparent in childhood. There is muscle stiffness, worse after rest and exacerbated by cold; minimal or no weakness, readily demonstrable percussion myotonia; and muscle hypertrophy, which tends to be more marked in the recessive form.

Both the recessive and dominant forms are caused by mutations in the *CLCN1* gene (on chromosome 7) encoding the skeletal-muscle chloride channel.

Malignant hyperthermia (MH)

The main features of this autosomal dominant disorder are a rapidly rising body temperature and generalized muscular rigidity during anaesthesia. Additional features include skin mottling, cyanosis, tachypnoea, tachycardia, cardiac dysrhythmias, and autonomic instability. Attacks in susceptible individuals may be triggered by suxamethonium and anaesthetic agents (halothane, cyclopropane, enflurane, ketamine). A similar, but probably different, disorder may be associated with heavy exercise in very hot conditions (for example, recruits undergoing route marches on mountains during a hot summer).

Attacks are life-threatening. Treatment consists of withdrawing the offending agent and providing general supportive measures and intravenous dantrolene (2 mg/kg body weight).

Disturbed calcium homeostasis underlies the attacks, with excessive calcium-ion influx into the sarcoplasmic reticulum. The disorder is genetically heterogeneous. In many families the underlying abnormality affects the skeletal-muscle calcium channel with either a mutation in the ryanodine receptor (*RYR1*) gene (on chromosome 19), or in the *CACNA1S* gene (on chromosome 1). *RYR1* mutations may also cause central core disease (**CCD**)—CCD and MH are allelic disorders and may occur together in the same individual or independently. Other *CACNA1S* gene mutations cause hypokalaemic periodic paralysis.

Screening for MH susceptibility involves muscle biopsy and *in vitro* testing for a reduced contractile threshold to halothane and caffeine. False-positive results are common but false-negative results appear to be very rare. It is hoped that specific molecular biological tests will become available. A significant practical problem is the management of family members who fear they may be at risk. As with those patients who have suffered hyperpyrexia under anaesthesia (even in whom repeated exposure has not led to a consistent re-occurrence), it is advisable for those individuals of proven or suspected risk to wear at all times some form of bracelet or locket giving details of the risk, in case they are casualties in an emergency such as a road accident.

Myoglobinuria

This important symptom and sign must be differentiated from haematuria and haemoglobinuria. Red cells are visible on microscopy in the former but not in the latter. In all three conditions, the haemoperoxidase stick test is positive.

Myoglobin is a protein that acts as an oxygen store within skeletal-muscle fibres. Myoglobinuria causes a dark-brown/red discoloration of the urine, the main concern being that the protein can cause renal tubular necrosis and thus renal failure. Numerous disorders are known to be associated with myoglobinuria (Table 3). In the metabolic disorders, the presumed mechanism is failure of substrate utilization or supply when energy demands increase during exercise or starvation. In other disorders, there is disruption of the plasma membrane. Apparently idiopathic cases are probably due to an unidentified metabolic defect or infection.

Rhabdomyolysis is considered further in Section 28.

Further reading

General

Brooke MH (1986). *A clinician's view of neuromuscular diseases*, 2nd edn. Williams and Wilkins, Baltimore.

Engel AG, Franzini-Armstrong C, eds (1994). *Myology*, 2nd edn. McGraw-Hill, New York.

Karpati G, Hilton-Jones D, Griggs R, eds (2001). *Disorders of voluntary muscle*, 7th edn. Cambridge University Press.

Lane RJM, ed. (1996). *Handbook of muscle disease*. Marcel Dekker, New York.

Scheinberg IH, Walshe JM, eds (1986). *Orphan diseases and orphan drugs.* Fulbright Papers 3. Manchester University Press.

WHO (1980). *International classification of impairments, disabilities, and handicaps*. World Health Organization, Geneva.

WHO (2000). *International classification of functioning and disability ICIDH-2*. http://www3.who.int/icf/icftemplate

Metabolic myopathies

Angelini C (1990). Defects of fatty-acid oxidation in muscle. *Ballière's Clinical Endocrinology and Metabolism* **4**, 561–82.

Barsy T, Hers H-G (1990). Normal metabolism and disorders of carbohydrate metabolism. *Ballière's Clinical Endocrinology and Metabolism* **4**, 499–522.

Bartram C, *et al.* (1994). McArdle's disease: a rare frameshift mutation in exon 1 of the muscle glycogen phosphorylase gene. *Biochimica et Biophysica Acta* **1226**, 341–3.

Hilton-Jones D, *et al.*, eds (1995). *Metabolic myopathies*. WB Saunders, London.

Layzer RB (1990). Muscle metabolism during fatigue and work. *Ballière's Clinical Endocrinology and Metabolism* **4**, 441–59.

Wagenmakers AJM, Coakley JH, Edwards RHT (1988). The metabolic consequences of reduced habitual activities in patients with muscle pain and disease. *Ergonomics* **31**, 1519–27.

Endocrine myopathies

Fells P (1991). Thyroid-associated eye disease: clinical management. *Lancet* **338**, 29–32.

Ruff RL, Weissmann J (1988). Endocrine myopathies. *Neurologic Clinics* **6**, 575–92.

Weetman AP (1991). Thyroid-associated eye disease: pathophysiology. *Lancet* **338**, 25–8.

Nutritional and toxic myopathies

Argov Z, Mastaglia FL (1994). Drug-induced neuromuscular disorders in man. In: Walton JN, Karpati G, Hilton-Jones D, eds. *Disorders of voluntary muscle*, 6th edn, pp 989–1029. Churchill Livingstone, Edinburgh.

Channelopathies

Ebers GC, *et al.* (1991). Paramyotonia congenita and hyperkalemic periodic paralysis are linked to the adult muscle sodium channel gene. *Annals of Neurology* **30,** 810–16.

Fontaine B, *et al.* (1990). Hyperkalemic periodic paralysis and the adult muscle sodium channel a-subunit gene. *Science* **250,** 1000–2.

Fontaine B, *et al.* (1991). Different gene loci for hyperkalemic and hypokalemic periodic paralysis. *Neuromuscular Disorders* **1,** 235–8.

Greenberg DA (1997). Calcium channels in neurological disease. *Annals of Neurology* **42**, 275–82.

Gronert GA (1980). Malignant hyperthermia. *Anesthesiology* **53,** 395–423.

Koch MC, *et al.* (1992). The skeletal muscle chloride channel in dominant and recessive human myotonia. *Science* **257,** 797–800.

Lehmann-Horn F, Jurkat-Rott K (1999). Voltage-gated ion channels and hereditary disease. *Physiological Reviews* **79**, 1317–72.

MacLennan DH, *et al.* (1990). Ryanodine receptor gene is a candidate for predisposition to malignant hyperthermia. *Nature* **343,** 559–61.

Myoglobinuria

Penn, AS (1986). Myoglobinuria. In: Engel AG, Banker BQ, eds. *Myology*, pp 1785–805. McGraw-Hill, New York.

24.22.5 Mitochondrial encephalomyopathies

P. F. Chinnery and D. M. Turnbull

Introduction

Mitochondria are ubiquitous intracellular organelles that are involved in many different metabolic pathways. Disorders of intermediary metabolism (such as fatty acid b-oxidation or tricarboxylic acid cycle defects) involve mitochondrial enzymes, but the term 'mitochondrial encephalomyopathy' usually means a disease which is due to an abnormality of the final common pathway of energy metabolism—the mitochondrial respiratory chain. The respiratory chain is essential for aerobic metabolism, and respiratory chain defects characteristically affect tissues and organs that are heavily dependent upon oxidative metabolism (such as the central nervous system, the eye, skeletal muscle, myocardium, and endocrine organs).

Recent studies have demonstrated the central role of the mitochondrion in the pathophysiology of well-established diseases such as Friedreich's ataxia and Wilson's disease, but these are not primarily disorders of the mitochondrial respiratory chain and are not considered here.

Biochemistry and genetics of the respiratory chain

The intermediary metabolism of carbohydrates, amino acids, and fatty acids generates the reduced cofactors NADH, NADPH, and $FADH_2$. These cofactors transfer electrons to the mitochondrial respiratory chain. As the electrons are passed through complexes I to IV of the respiratory chain along the inner mitochondrial membrane, protons are pumped out of the mitochondrial matrix into the intermembrane space. This creates an electrochemical gradient that is harnessed by complex V (ATP synthase) to generate ATP from ADP. Each respiratory chain complex contains many polypeptide subunits, some of which are coded by genes within the nucleus and some of which are encoded by the mitochondrial genome. Although all of the polypeptides encoded in mitochondrial DNA (**mtDNA**) have been known for over a decade, many nuclear genes involved in mitochondrial biogenesis have yet to be characterized.

The mitochondrial genome encodes seven complex I subunits (NADH-ubiquinone oxidoreductase), one of the complex III subunits (ubiquinol-cytochrome *c* oxidoreductase), three of the complex IV (cytochrome *c* oxidase, or **COX**) subunits, and the ATPase 6 and ATPase 8 subunits of complex V. Interspaced between the protein-encoding genes are two ribosomal RNA genes (12S and 16S rRNA), and 22 transfer RNA genes that provide the necessary RNA components for the mitochondrial translation machinery. The remaining polypeptides, including all of the complex II subunits, are synthesized from nuclear gene transcripts within the cytosol. These are subsequently imported into the mitochondria through the inner and outer membrane translocation complexes. There are many additional proteins that are essential for the normal assembly and function of the mitochondrial respiratory chain. As a result, mitochondrial respiratory chain disorders can be due to mutations affecting both nuclear and mitochondrial genes.

The classification and investigation of mitochondrial respiratory chain disorders has been revolutionized by the recent advances in our understanding of the underlying genetic defects. By the year 2000, over 70 different point mutations and over 100 different deletions of mtDNA had been associated with a wide variety of different diseases. One factor that contributes to the frequency of mitochondrial mutations is the absence in mitochondrial DNA polymerase of the 'proof-reading' property of nuclear DNA polymerase where exonuclease activity greatly enhances the fidelity of replication. Very recently a number of nuclear genetic defects have been identified in some patients with respiratory chain defects (Table 1).

Basic mitochondrial genetics

There are two main differences between nuclear DNA and mtDNA that are important for the expression and transmission of mitochondrial genetic disease, as follows.

Heteroplasmy and the threshold effect

Each mammalian cell contains over 1000 copies of the small (16.5 kb) mitochondrial genome; there are on average 5 to 15 mtDNA molecules in each organelle. Individuals with mtDNA disease often harbour a mixture of mutated and wild-type (normal) mtDNA—a situation known as heteroplasmy. Single cells only express a respiratory chain defect when the proportion of mutated mtDNA exceeds a critical threshold. Different organs, and even adjacent cells within the same organ, may contain different amounts of mutated mtDNA. This variability, coupled with tissue-specific differences in the threshold and the varied dependence of different organs on oxidative metabolism, explains in part why certain tissues are preferentially affected in patients with mtDNA disease. In general, postmitotic (non-dividing) tissues such as neurones, skeletal and cardiac muscle, and endocrine organs harbour much higher levels of mutated mtDNA and are often clinically involved. In contrast, rapidly dividing tissues such as the bone marrow are only rarely clinically affected (one example is Pearson's syndrome—see below).

Maternal inheritance and the transmission of heteroplasmy

After fertilization of the oocyte, sperm mtDNA is actively degraded. As a consequence, mtDNA is transmitted exclusively down the maternal line. This means that affected males with mtDNA disease cannot transmit the genetic defect. Deleted molecules are rarely, if ever, transmitted from clinically affected females to their offspring. By contrast, a female harbouring a heteroplasmic mtDNA point mutation, or mtDNA duplications, may transmit a variable amount of mutated mtDNA to her children. Early during development of the female germ line, the number of mtDNA molecules within each oocyte is reduced before being subsequently amplified to reach a final number of around 100 000 in each mature oocyte. This restriction and amplification (also called the mitochondrial 'genetic bottleneck') contributes to the variability between individual oocytes, and the different levels of mutant mtDNA seen in the offspring of a single female.

Clinical presentation of respiratory chain disorders

Mitochondrial encephalomyopathies are highly variable both clinically and at the genetic level. The same clinical syndrome can be caused by different genetic defects (which may be within nuclear or mitochondrial genes), but the same genetic defect may present in a variety of different ways. In general, adults who present with mitochondrial disease are often found to have a defect of mtDNA. Children often present with different clinical features and are more likely to have a nuclear genetic defect. It is often possible to identify well-defined clinical syndromes, but many patients present with a collection of clinical features that are highly suggestive of respiratory chain disease but do not fit into a discrete clinical category.

Defined clinical syndromes (Table 2)

Large-scale deletions can cause chronic progressive external ophthalmoplegia and bilateral ptosis. Some of these patients have minimal disability and may have limited skeletal muscle involvement. In contrast, similar deletions may also cause chronic progressive external ophthalmoplegia with bilateral sensorineural deafness, cerebellar ataxia, pigmentary retinopathy, diabetes mellitus, and cardiac conduction defects leading to complete heart block. When this begins in teenage years and is associated with a raised cerebrospinal fluid protein, it is called the Kearns–Sayre syndrome, which is a progressive neurological disorder associated with severe disability. Hypopara-thyroidism and hypothyroidism are well-recognized features of Kearns–Sayre syndrome. The vast majority of cases of chronic progressive external ophthalmoplegia and Kearns–Sayre syndrome are sporadic. These two syndromes are the extremes of a spectrum of disease and many individuals lie somewhere between the pure extraocular muscle and severe central neurological phenotypes.

Pearson's syndrome of exocrine pancreatic failure, sideroblastic anaemia, and marrow panhypoplasia is usually due to a mtDNA deletion. Pearson's syndrome usually presents in infancy and a number of individuals who have survived into later childhood subsequently developed the Kearns–Sayre phenotype.

Pathogenic point mutations of mtDNA are more common than rearrangements. This is partly because mtDNA deletions cause sporadic disease, whereas many mtDNA point mutations are transmitted down the maternal line. The A3243G mutation in the leucine (UUR) tRNA gene was first described in a patient with mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes (MELAS). Different families harbouring the same genetic defect may have different phenotypes. For example, some families harbouring A3243G have predominantly diabetes and deafness, some families have chronic progressive external ophthalmoplegia, and some present with a cardiomyopathy. It is currently not known why this is the case but it is likely that additional nuclear genetic factors play an important role in modifying the expression of the primary mtDNA defect. This single mutation is important since it has been estimated that between 0.5 and 1.5 per cent of cases of diabetes mellitus in the general population are associated with the A3243G mutation.

Patients may present with myoclonic epilepsy, ataxia, optic atrophy, and have ragged-red fibres in skeletal muscle (MERRF) and this may also be due to a point mutation of mtDNA (for example A8344G).

mtDNA mutations are the major cause of visual loss in young adult males. About half of all males who harbour one of three point mutations of mtDNA (G11778A, T14484C, G3460A) develop bilateral sequential visual loss in the second or third decade—a disorder known as Leber's hereditary optic neuropathy (LHON). The majority of individuals with these mutations are homoplasmic—harbouring only mutated mtDNA. It is not clear why the disease only affects approximately half of the males and 10 per cent of females who inherit the primary mtDNA defect. Environmental factors, such as alcohol and tobacco, partly explain the variable penetrance of this disorder; however, additional, as yet unknown, nuclear genetic factors may also be important.

Leigh syndrome (subacute necrotizing encephalomyopathy) is a relapsing encephalopathy with prominent cerebellar and brainstem signs that usually presents in childhood and is associated with characteristic neuroimaging abnormalities involving the basal ganglia. Leigh syndrome can be due to an X-linked pyruvate dehydrogenase deficiency or a defect of the mitochondrial respiratory chain. Complex I deficiency or COX deficiency are common findings in Leigh syndrome. In these patients it may be possible to identify recessive mutations in nuclear complex I genes, or genes involved in the assembly of the respiratory chain complexes (for example *SURF 1*). Point mutations at position 8993 in the ATPase 6 gene of mtDNA may cause neurogenic weakness with ataxia and retinitis pigmentosa. These particular mutations are also associated with some forms of childhood Leigh syndrome.

COX deficiency may also present in childhood with an infantile myopathy and a severe lactic acidosis, which may also be associated with a cardiomyopathy and the Toni–Fanconi–Debre syndrome. Despite maximal supportive intervention, this is usually a fatal disorder and a severe depletion of mtDNA occurs in a proportion of these cases. It is important to recognize that isolated myopathy and lactic acidosis may be self-limiting, often with a significant improvement by 1 year of age and complete resolution by the age of 3 years.

Non-specific clinical presentations

Although the foregoing diseases and numerous other syndromes may strongly suggest a mitochondrial aetiology (Fig. 1 and Table 2), many patients do not present with a characteristic phenotype. Children may present in the neonatal period with a metabolic encephalopathy and systemic lactic acidosis, often associated with hepatic and cardiac failure. This may be associated with a depletion in the total amount of mtDNA within affected tissues. Although this syndrome may be fatal, in some it is a self-limiting disorder. Childhood presentations may be even less specific, with neonatal hypotonia, feeding and respiratory difficulties, and failure to thrive. A respiratory chain defect should be considered in any patient who has a disease with multiple organ involvement, particularly if there are central neurological features (such as seizures and dementia), a myopathy, cardiomyopathy, and endocrine abnormalities such as diabetes mellitus (Fig. 1). Bilateral sensorineural deafness and ocular features (retinopathy, optic atrophy, ptosis, and ophthalmoparesis) are common. Renal tubular defects, gastrointestinal hypomotility, cervical lipomatosis, and psychiatric features are also well described in patients with respiratory chain disease.

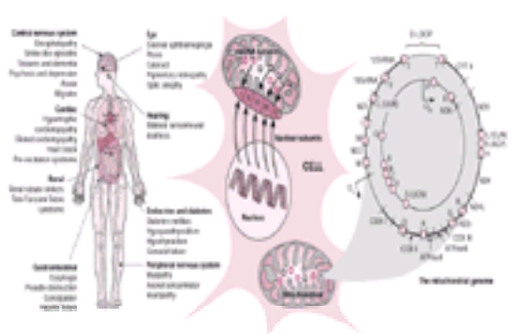

Fig. 1 The clinical features and biochemical and molecular genetic basis of mitochondrial encephalomyopathies.

Investigation of respiratory chain disease

The investigation of patients with a suspected mitochondrial encephalomyopathy involves the careful assimilation of clinical and laboratory data. In a significant proportion of cases (such as Leber's hereditary optic neuropathy), it is possible to identify a specific clinical syndrome with a clear maternal family history. Under these circumstances it is appropriate to carry out a molecular genetic test on a blood sample. In many situations, particularly in sporadic cases, this is not appropriate because the clinical features overlap with those of many other disorders. Even if the patient has a mitochondrial disorder, numerous different genetic defects may be responsible, some of which will not be detectable by analysis of blood samples.

Investigations fall into two main groups: clinical investigations used to characterize the pattern and nature of the different organs involved, and specific investigations to identify the biochemical or genetic abnormality.

General clinical investigations

It is essential to search for the more common features of respiratory chain disease. This includes cardiac assessment (ECG and echocardiography) and endocrine assessment (oral glucose tolerance test, thyroid function tests, alkaline phosphatase, fasting calcium, and parathyroid hormone levels). The organic and amino acids in urine may be abnormal even in the absence of overt tubular disease. Measuring blood and cerebrospinal fluid lactate levels is more helpful in the investigation of children than adults. These measurements must be interpreted with caution because there are many causes of blood and cerebrospinal fluid lactic acidosis, including fever, sepsis, dehydration, seizures, and stroke. The cerebrospinal fluid protein may be elevated. The serum creatine kinase level may be raised but is often normal. Neurophysiological studies may identify a myopathy or neuropathy. Electroencephalography may reveal diffuse slow-wave activity consistent with a subacute encephalopathy, or evidence of seizure activity. Cerebral imaging may be abnormal, showing lesions of the basal ganglia, high signal in the white matter on MRI, or generalized cerebral atrophy.

Specific investigations

A skeletal muscle biopsy is invaluable in the investigation of respiratory chain disease. Histochemical and biochemical investigations, in conjunction with the clinical assessment, often indicate where the underlying genetic abnormality must lie.

Histochemistry and biochemistry

Histochemical analysis may reveal subsarcolemmal accumulation of mitochondria (so-called 'ragged-red' fibres), or COX deficiency. A mosaic of COX-positive and COX-negative muscle fibres suggests an underlying mtDNA defect. Patients who have COX deficiency due to a nuclear genetic defect usually have a global deficiency of COX affecting all muscle fibres. Electron microscopy may identify paracrystalline inclusions in the intermembrane space, but these are non-specific and may be seen in other non-mitochondrial disorders. Respiratory chain complex assays can be carried out on various tissues. Skeletal muscle is preferable, but cultured fibroblasts are useful in the investigation of childhood mitochondrial disease. Measurement of the individual respiratory chain complexes determines whether an individual has multiple complex defects that would suggest an underlying mtDNA defect, involving either a tRNA gene or a large deletion. Isolated complex defects may be due to mutations in either mitochondrial or nuclear genes.

Molecular genetic investigations

Under certain circumstances, the clinical and biochemical features may point towards a specific genetic defect, and it may be possible to detect this abnormality in a blood sample. Children presenting with Leigh syndrome and who have an isolated deficiency of one of the respiratory chain subunits may have a point mutation within the nuclear-encoded respiratory chain subunit or assembly genes. These have been identified by direct sequencing of the appropriate exons.

For some mtDNA defects (particularly mtDNA deletions) the abnormality is not detectable in a DNA sample extracted from blood, and the analysis of DNA extracted from muscle is essential to establish the diagnosis. The first stage is to look for mtDNA rearrangements or mtDNA depletion by Southern blot analysis and long-range polymerase chain reaction (**PCR**). This is followed by PCR or restriction fragment length polymorphism analysis for common point mutations. Many patients with mitochondrial disease have a previously unrecognized mtDNA defect and it is necessary to sequence directly the mitochondrial genome. Interpretation of the sequence data can be extremely difficult. mtDNA is highly polymorphic and any two normal individuals may differ by up to 60 base pairs. In the strictest sense, a mutation can only be considered to be pathogenic if it has arisen independently several times in the population, it is not seen in controls, and it is associated with a potential disease mechanism. These stringent criteria depend upon a good knowledge of polymorphic sites in the background population. If a novel base change is heteroplasmic, this suggests that it is of relatively recent onset. Family, tissue segregation, and single cell studies may show that higher levels of the mutation are associated with mitochondrial dysfunction and disease, which strongly suggests that the mutation is causing the disease.

Management

There is currently no definitive treatment for patients with mitochondrial disease. Management is aimed at minimizing disability, preventing complications, and genetic counselling.

Supportive care and surveillance

Many patients with mitochondrial disorders require follow-up over many decades. An integrated approach is essential involving the primary physician, other specialist physicians (ophthalmology, diabetes, and cardiology), specialist nurses, physiotherapists, and speech therapists. Vigilant clinical monitoring over many years can prevent the development of complications, such as those secondary to cardiac and endocrine involvement. Specific procedures may be indicated at various stages of disease. These include cardiac pacing, ptosis correction, cataract surgery, and percutaneous gastrostomy.

Genetic counselling

The detailed investigation of patients with respiratory chain disease usually leads to a specific molecular genetic diagnosis, particularly in adults. This has profound implications on the counselling given to patients and their families. Most children with respiratory chain disease are compound heterozygotes with recessive nuclear gene mutations. If it is possible to identify the causative mutations in both the offspring and parents, then this will allow confident genetic counselling for the whole family. If, as in many cases, it is not possible to identify the underlying gene defect, or the genetic defect in the affected child cannot be traced back to the parents, then counselling is less straightforward.

If a causative mtDNA defect is identified, then the implications for counselling are distinctly different. Males cannot transmit pathogenic mtDNA defects. Patients who carry mtDNA deletions rarely have a family history suggestive of mtDNA disease, and there is no significant risk that they will transmit the mtDNA defect to any offspring. There are a few rare exceptions to this rule where the propensity to develop mtDNA deletions is transmitted as an autosomal dominant or autosomal recessive trait. By contrast, women harbouring pathogenic mtDNA point mutations may transmit the genetic defect to their offspring. The mitochondrial genetic 'bottleneck' leads to a variation in the proportion of mutated mtDNA that is transmitted to any offspring (see above). It is therefore possible for a female to have mildly affected as well as severely affected children. The risk of having affected offspring varies from mutation to mutation, and although there does appear to be a relationship between the level of mutated mtDNA in the mother and the risk of affected offspring, there are insufficient data from prospective studies to allow accurate risk prediction.

Prognosis

In general the prognosis depends upon the extent of central neurological involvement. Patients with Leber's hereditary optic neuropathy rarely have significant central neurological features and have a normal lifespan. The prospect for visual recovery varies. After the initial nadir, individuals harbouring the G11778A mutation are the least likely to regain functional vision, whilst those harbouring the T14484C mutation are the most likely to regain their sight.

Children presenting with an encephalopathy have a poor prognosis. Although residual neurological deficits are common after repeated childhood encephalopathic episodes, the disease may enter a more stable 'chronic' phase during teenage years and adulthood. A similar course may be seen in adults presenting with a relapsing encephalopathy. In contrast, a large proportion of adults with mtDNA defects and chronic progressive external ophthalmoplegia have very mild disease that may remain limited to the extraocular muscles for many decades. For certain mutations, there also appears to be a relationship between the proportion of mutated mtDNA in skeletal muscle and the severity of the disease. Although the proportion of mutated mtDNA in muscle may give some guide to prognosis, there is insufficient information available to allow accurate prognostic counselling based upon these determinations. A significant proportion of patients have distinct phenotypes associated with unique genetic defects and the prognosis must be guarded in these families.

Pharmacological treatments and novel approaches under development

Standard doses of vitamin C and K, thiamine, riboflavin, and ubiquinone (coenzyme Q10) may be of some benefit. These treatments have no significant side-effects and are relatively cheap, but their efficacy is largely based upon anecdotal reports. Novel treatments are, however, under development. Dichloracetate can be used to reduce lactic acid levels but may cause an irreversible toxic neuropathy. The efficacy of dichloracetate is currently being assessed in clinical trials. Exercise is important for patients with mtDNA disease, and isometric muscle contraction may lead to an improvement in muscle strength. Drug-induced muscle necrosis followed by proliferation of myoblasts may also be important for the treatment of mitochondrial myopathy and ptosis, but this approach is only at the experimental stage. Finally, several centres are investigating methods for correcting the underlying mtDNA defect by gene therapy.

Further reading

Anderson S *et al*. (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–65.

Andrews RM *et al*. (1999). Reanalysis and revision of the Cambridge Reference Sequence. *Nature Genetics* **23**, 147. [Benchmark reference sequences for normal human mtDNA.]

Chinnery PF *et al*. (1998). MELAS and MERRF: the relationship between maternal mutation load and the frequency of clinically affected offspring. *Brain* **121**, 1889–94. [First paper to show a

relationship between maternal mutation load and the outcome of pregnancy.]

Chinnery PF *et al*. (1999). Clinical mitochondrial genetics. *Journal of Medical Genetics* **36**, 425–36. [A detailed description of the clinical aspects of mitochondrial disease.]

Dahl H-HM (1998). Getting to the nucleus of mitochondrial disorders: identification of respiratory chain-enzyme genes causing Leigh syndrome. *American Journal of Human Genetics* **63**, 1594–7. [A review of the nuclear genes causing Leigh syndrome.]

DiMauro S, Schon EA (1998). Nuclear power and mitochondrial disease. *Nature Genetics* **19**, 214–5. [An excellent introduction to nuclear genes and diseases involving mitochondria, including Wilson's disease, Friedreich's ataxia, and hereditary spastic paraparesis.]

Harding AE *et al*. (1995). Pedigree analysis in Leber hereditary optic neuropathy families with a pathogenic mtDNA mutation. *American Journal of Human Genetics* **57**, 77–86. [Important paper summarizing the risks of blindness for the most common mutations causing Leber hereditary optic neuropathy.]

Howell N *et al*. (1998). Mitochondrial DNA mutations that cause optic atrophy: how do we know? *American Journal of Human Genetics* **62**, 196–202. [A succinct discussion of the molecular genetics and disease mechanisms in Leber hereditary optic neuropathy.]

Jackson MJ *et al*. (1995). Presentation and clinical investigation of mitochondrial respiratory chain disease. *Brain* **118**, 339–57. [Clinical features and investigation of a large series of adults with mitochondrial disease.]

Larsson N-G, Clayton DA (1995). Molecular genetic aspects of human mitochondrial disorders. *Annual Review of Genetics* **29**, 151–78. [A review of basic mitochondrial genetics.]

Lightowlers RN *et al*. (1997). Mammalian mitochondrial genetics: heredity, heteroplasmy and disease. *Trends in Genetics* **13**, 450–5. [A discussion of the basic principles of mitochondrial genetics.]

Poulton J, Macaulay V, Marchington DR (1998). Mitochondrial genetics '98: Is the bottleneck cracked? *American Journal of Human Genetics* **62**, 752–7. [A contemporary review of the approaches to the inheritance of heteroplasmic mtDNA defects.]

Smeitink J, van den Heuvel L (1999). Human miotochondrial complex I in health and disease. *American Journal of Human Genetics* **64**, 1505–10. [Comprehensive review of nuclear complex I genes and human disease.]

Taylor RW *et al*. (1997). Treatment of mitochondrial disease. *Journal of Bioenergetics and Biomembranes* **29**, 195–205. [A discussion of current therapy and novel treatment approaches under development.]

Wallace DC (1999). Mitochondrial diseases in mouse and man. *Science* **283**, 1482–8. [A review of recent scientific developments and mouse models for mitochondrial disease.]

24.22.6 Tropical pyomyositis (tropical myositis)

D. A. Warrell

Definition

The term 'tropical pyomyositis' should be restricted to primary muscle abscesses arising within skeletal muscles. This condition must be distinguished from abscesses extending into muscle either from subcutaneous sites following infection through the skin, or from osteomyelitis or suppuration originating in tissues other than muscle.

Geographical occurrence

Tropical myositis has been reported from most parts of tropical Africa, Malaysia, Thailand, India, Indonesia, Oceania, Central and South America, and the Caribbean. It is common in many tropical countries, accounting for 4 per cent of admissions to a hospital in Uganda and for 2.2 per cent of all surgical admissions to a hospital in eastern Ecuador. In temperate climates, pyomyositis was extremely rare, but is becoming more common in patients immunosuppressed due to the human immunodeficiency virus (**HIV**), lymphomas, chemotherapy of malignant diseases, asplenia, Felty's syndrome, and other conditions.

Aetiology

Staphylococcus aureus is the organism most commonly cultured from the abscesses. *Streptococcus pyogenes* (usually group A) is responsible for a few cases, but tropical pyomyositis must be distinguished from streptococcal necrotizing myositis (also known as peracute streptococcal pyomyositis or spontaneous streptococcal gangrenous myositis) which is more fulminant and diffuse and has a very high mortality. Other isolates have included *S. pneumoniae*, *Haemophilus influenzae*, *Escherichia coli*, *Pseudomonas* species, and anaerobes. In Thailand, most cases of pyomyositis are caused by *Burkholderia pseudomallei*. The strikingly different incidence of pyomyositis in tropical and temperate countries has not been explained. In Africa and South America, the condition appears to be relatively more common in indigenous peoples. A history of preceding trauma to the affected muscle is obtained from more than 20 per cent of patients in most series. It has been suggested that, by analogy with osteomyelitis, a muscle haematoma provides a nidus for blood-borne infection. A number of predisposing causes has been suggested: preceding viral infection (for example, an arbovirus), general debilitation, and nematode infections—particularly toxocariasis, *Lagochilascaris minor*, and filariasis. None has been supported by convincing evidence, but sickle-cell disease may be a genuine predisposing cause in a minority of cases. Most of the abscesses associated with helminth infections should not be termed 'pyomyositis' as they are inter- rather than intramuscular. For example, *Dracunculus medinensis* can give rise to deep intermuscular abscesses secondarily infected with *Staphylococcus aureus*.

Pathology

The abscesses may be large, are usually loculated, and are situated within skeletal muscles beneath the deep fascia. Histologically, there is focal muscle necrosis with an infiltration of mononuclear cells and inflammatory oedema.

Clinical features

Tropical pyomyositis can occur at any age but its highest incidence is in the second decade. It is commoner in males. The earliest symptom is pain and tenderness of the affected muscle. Any of the skeletal muscles may be involved, but those of the trunk and lower limbs are the most commonly affected. Usually there is a single localized abscess, but multiple abscesses in distantly separated muscles can occur. At an early stage, an ill-defined tender and thickened area may be palpable in the muscle. Later, a localized, very tender, and hot swelling is palpable. There may be redness and oedema of the overlying skin, but the skin is not primarily involved. The swelling is usually non-fluctuant and there is no local lymphadenopathy. Symptoms and signs usually develop over a few days. Peripheral leucocytosis is not invariable. Eosinophilia is frequently described but is usually common in the populations most affected by tropical pyomyositis. In spite of considerable muscle destruction at the site of the abscess, serum concentrations of muscle enzymes may not be elevated, but in some cases there is myoglobinaemia, myoglobinuria, and acute renal failure. Complications are uncommon, but consist of spread of infection from the affected muscle to other structures such as joints resulting in septic arthritis, to the pleural cavity resulting in empyema, or by haematogenous spread to the heart valves. Mortality in inpatients is said to be less than 1.5 per cent.

Diagnosis

The differential diagnosis is from pus tracking from abscesses in other organs and tissues, muscle haematomas, torn muscles, certain highly vascular or necrotic tumours of connective tissue or muscle (such as rhabdomyosarcoma), and the inflammatory and allergic swellings resulting from the migration of helminths such as *Loa loa* and *Gnathostoma*, *Paragonimus*, and sparganum spp. *Staphylococcus aureus* is usually cultured from the pus, but blood cultures are positive in less than 5 per cent of cases. Ultrasound, computed tomography (**CT**), and especially magnetic resonance imaging (**MRI**) scans are useful for localizing abscesses and guiding needles for diagnostic and therapeutic aspiration.

Treatment

Full surgical exploration, debridement, and drainage are essential. Because the abscesses are usually loculated, needle aspiration is inadequate. Parenteral treatment with a b-lactamase-resistant penicillin (flucloxacillin) should be started immediately, but if group A *Streptococcus* is cultured, benzyl penicillin or clindamycin are the drugs of choice.

Further reading

Chiedozi LC (1979). Pyomyositis: review of 205 cases in 112 patients. *American Journal of Surgery* **137**, 255–9.

Gibson RK, Rosenthal SJ, Lukert BP (1984). Pyomyositis: increasing recognition in temperate climates. *American Journal of Medicine* **77**, 768–72.

Hossain A, *et al.* (2000). Nontropical pyomyositis: analysis of eight patients in an urban center. *American Surgeon* **66**, 1064–6.

Levin MJ, Gardner P, Waldvogel FA (1971). 'Tropical' pyomyositis. An unusual infection due to *Staphylococcus aureus*. *New England Journal of Medicine* **284**, 196–8.

Marcus RT, Foster WD (1968). Observations on the clinical features, aetiology and geographical distribution of pyomyositis in East Africa. *East African Medical Journal* **45**, 167–76.

Norrgren H, *et al.* (1997). Increased prevalence of HIV-2 infection in hospitalized patients with severe bacterial diseases in Guinea-Bissau. *Scandinavian Journal of Infectious Diseases* **29**, 453–9.

Smith PG, *et al*. (1978). The epidemiology of tropical myositis in the Mengo districts of Uganda. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **72**, 46–53.

Soler R, *et al.* (2000). Magnetic resonance imaging of pyomyositis in 43 cases. *European Journal of Radiology* **35**, 59–64.

Vassilopoulos D, *et al.* (1997). Musculoskeletal infections in patients with human immuno-deficiency virus infection. *Medicine (Baltimore)* **76**, 284–94.

25 The eye in general medicine

Peggy Frith

The significance of disorders of the eye in general medicine

Because the eye may be involved in so many diseases, it is essential that clinicians are familiar with ocular manifestations, learn how to examine the eye, and in particular, are proficient with the ophthalmoscope. Ocular findings may point to the diagnosis of a particular systemic disorder and in some cases an eye complication may need urgent and specific treatment.

Red eye

The pattern of redness suggests a possible diagnosis and other features help to confirm this, as shown in Table 1. The slit lamp shows specific features found in some types of conjunctivitis, the staining pattern of corneal lesions, cells diagnostic of uveitis within the eye chambers, and raised eye pressure of glaucoma. Iritis is described with ankylosing spondylitis, and scleritis with rheumatoid arthritis.

Dry eye

Lack of tears may have a systemic cause, particularly if there is also dryness of the mouth—sicca syndrome. Sicca with an identifiable systemic association is known as Sjögren's syndrome (as with rheumatoid arthritis, systemic sclerosis, mixed connective tissue disease), graft-versus-host disease, or sarcoidosis. The eyes feel gritty and are red or sticky (see rheumatoid arthritis below). Artificial tear drops can help, but severely dry eyes are a miserable problem which can be very difficult to manage.

Loss of vision

A clear history of the visual loss is important. Visual acuity should be measured in each eye, using glasses or a pinhole to correct for any error of focus. Major impairment of vision in one eye gives an asymmetrical pupil response to a bright light. There is a limited number of important causes within the retina or optic nerve (see Table 2).

The Ophthalmoscope

The optic nerve head is usually visible through an undilated pupil, but it is impossible to assess the retina reliably without using a mydriatic. Short-acting drops, such as tropicamide 1 per cent, work within 15 min and last about 2 h, with no risk of causing acute glaucoma. The pupil should not be dilated if the patient has a suspected subarachnoid haemorrhage, coma, or recent head injury. The central fovea is seen if the patient looks directly into the light, and in patients with diabetes the area temporal to the fovea should also be examined. A clouded view suggests opacity in the lens or vitreous, best seen by adjusting focus on to the pupil margin to give a red reflection against which opacities stand out as black shadows, especially if the pupil is dilated. The peripheral retina, for example in sickle-cell retinopathy, is best seen with the indirect ophthalmoscope.

Superficial flame-shaped haemorrhages, though not unique to hypertension, demand measurement of blood pressure. Deeper dot and blot haemorrhages temporal to the macula suggest diabetes, vein occlusion if unilateral or localized, or a haematological disorder if bilateral. Subhyaloid haemorrhage, confined in front of the retina and behind the vitreous, forms a dense focus which may sediment into a characteristic flat-topped shape, typical of bleeding from new vessels—as in diabetes or after retinal vascular occlusion—or secondary to a bleeding diathesis or trauma, including non-accidental injury. Most haemorrhages are asymptomatic and will resolve, but vision falls if the fovea is involved or if blood leaks into the vitreous itself, causing floaters.

Shiny hard exudates consist of protein and lipid; if in circles (circinate), focal vascular leakage may be associated with diabetes, whereas a star around the fovea forms with resolving retinal oedema, as in treated hypertension or papilloedema. Commonly confused with exudates are retinal drusen, which are more uniform and discrete, usually scattered around the retina or congregated close to the fovea. Cotton wool spots are fluffy pale patches indicating swollen nerve fibre axons at sites of microvascular closure. They are always significant (see Table 3). Retinal infiltrates look like cotton wool spots but consist of cells spilled into the vitreous. These are visible with the slit lamp, as in active toxoplasmosis, cytomegalovirus retinitis, sarcoidosis, Behçet's syndrome, or ocular lymphoma. Discrete punched-out scars suggest healed foci of inflammation of the retina and underlying choroid, of which the most common cause is toxoplasmosis.

Fluorescein angiography

This can define the type and severity of retinal vascular disorders. The dye, injected intravenously, demonstrates patterns of perfusion both in the retina and underlying choroidal circulation, outlines abnormalities of the vessels such as microaneurysms, and identifies sites of leakage indicating damage to retinal vessels or the formation of new vessels. Angiography is valuable diagnostically and indicates where laser treatment is needed. Anaphylactic reactions and even fatalities have been reported, so patients must be carefully selected.

Visual fields

The visual fields should be examined in patients with visual loss. Even large defects may go unnoticed. A unilateral central or altitudinal (top or bottom of field) defect suggests an anterior lesion, in retina or optic nerve; a bitemporal defect implicates the optic chiasm; and a homonymous defect the visual path posterior to the chiasm. With bilateral occipital infarction, visual loss may be difficult to define and pupil reactions are normal. A CT scan may be advisable. If there is unaccountably poor vision in one eye, a defect of focus or an amblyopic (lazy) eye resulting from a squint or refractive error in childhood may be responsible: the first should improve with a pinhole device but the second will not.

Eye changes in diabetes (Plate 1, Plate 2, Plate 3, Plate 4 and Plate 5)

Retinopathy, the most common serious eye complication in diabetes, is the principal reason for blind registration of younger adults from industrial countries. Annual retinal screening is essential, as early treatment can prevent blindness and patients with sight-threatening changes are often asymptomatic.

Older patients especially may have cataract, glaucoma, retinal vein occlusion, and occasionally ischaemic optic neuropathy. Diabetic eye disease is discussed fully in Chapter 12.11.1.

The eye in hypertension

In the hypertensive patient retinal changes will help determine if treatment is necessary, if it is adequate, or if it is needed urgently. Description of individual features is preferable to grading. Unless the pupil is dilated, it is easy to miss or to underestimate retinal changes. A bright halogen bulb and a green or 'red-free' filter helps accentuate vessels and haemorrhages. Haemorrhages or cotton wool spots, indicating acute changes, are most likely to be seen temporal and nasal to the optic nerve head, around the major vessels. Blurring of the margins of the optic nerve head, indicating disc oedema, must be excluded.

Long-standing hypertension and ageing produce similar changes. Arterioles are narrowed, irregular, or tortuous and the wall may be thickened, showing an increase in reflected light described as copper or silver wiring. There may be nipping at the arteriovenous crossings so that the underlying vein appears to be constricted. Long-standing changes often persist with treatment of hypertension but may be reversed in younger patients.

Classification of hypertensive retinopathy

Perhaps the best known grading of hypertensive retinopathy is the Keith–Wagner classification:

- Grade 1, mild narrowing or sclerosis of retinal arteries;
- Grade 2, moderate to marked narrowing or sclerosis with light reflex and arteriovenous crossing changes;
- Grade 3, in addition, haemorrhages or cotton wool spots; and
- Grade 4, in addition, swelling of the optic nerve head (papilloedema).

Grades 3 and 4 indicate severe, accelerated, or 'malignant' retinopathy, but disc swelling is no longer regarded as a reliable feature in assessing urgency for treatment—the prognosis is similar for grades 3 and 4.

Blood pressure high enough to damage the renal and cerebral circulation is best recognized by inspecting retinal vessels, even in the absence of visual symptoms. Diastolic pressure likely to be associated with severe retinal changes is usually 110 mmHg or higher at some stage. Proteinuria is almost invariable.

Acute, severe retinal changes indicate either leakage or closure of smaller vessels. Flame haemorrhages are seen particularly around the vessel trunks above and below the macula, temporal to the optic disc (Plate 6). These indicate leakage of blood from fine superficial capillary branches supplying the nerve fibre layer. Bleeding deeper in the retina forms blot-like haemorrhages which, if widespread or in a wedge shape from an arteriovenous crossing, indicate occlusion of the central or a branch retinal vein. Haemorrhages resolve with treatment of hypertension. Only foveal haemorrhage causes visual impairment.

Cotton wool spots indicate closure of capillaries supplying the nerve fibres. Microinfarcts cause stasis of axoplasmic flow and intra-axonal contents accumulate, distending and opacifying the fibres and producing the pale fluffy appearance. Spots gradually resolve with treatment. In the absence of hypertension, an

inflammatory vasculitis such as systemic lupus erythematosus should be suspected.

Hypertensive damage causes disruption of endothelial tight junctions in retinal vessels so that fluid, protein, and lipid leak into the extracellular spaces within the retina. These are removed by macrophages and processed into shiny hard exudates which may persist for many months. Hard exudates imply leakage for more than a matter of days. They are common in resolving hypertensive retinopathy, forming a characteristic star around the fovea. In hypertension, exudate rarely forms in the ring-shaped circinate pattern typical of diabetes.

Papilloedema implies hypertensive damage to the disc capillaries or cerebral oedema with raised intracranial pressure, attributable to hypertension. Sudden reduction of blood pressure may cause acute, sometimes irreversible, loss of vision also with a risk of stroke.

Histopathology

In the early phases of severe retinopathy there is disruption of endothelial cells or tight junctions followed by vessel wall damage leading to occlusion, sometimes with fibrinoid necrosis or frank thrombosis.

Non-retinal eye changes in hypertension

Occasional patients with severe hypertension, especially those with eclampsia or renal failure, may suffer pronounced visual loss secondary to occlusive changes in the vessels supplying the optic nerve head or in the choroid underlying the retina itself. The tissues become swollen and pale and the retina may even become detached by fluid. Rarely, patients with secondary hypertension may have eye manifestations of genetic disorders such as neurofibromatosis type 1, von Hippel-Lindau, or Sipple's syndrome (see below).

Ocular vascular occlusion

Retinal vascular occlusion is a common cause of blindness, especially in elderly patients. This can be a valuable warning of vascular disease elsewhere, particularly affecting the cerebral circulation.

Retinal artery occlusion

Occlusions of central retinal arteries or their branches, are almost always embolic, arising in the carotid system in the neck, or intracranially, or in the heart. Usually, they start abruptly with permanent loss of function once the retina has infarcted. Central artery occlusion results in profound loss of vision. Branch occlusion causes a visual field defect that often has a horizontal (altitudinal) edge which the patient may be able to define. Recovery of vision is unlikely but prognosis for the other eye is good, especially if an underlying cause can be corrected (Table 4).

Amaurosis fugax

Transient retinal ischaemia causes brief episodes of blindness limited to one eye (amaurosis fugax), usually lasting for a few seconds, some for up to a few minutes, rarely longer. Loss of vision may be total or partial, affecting the upper or lower half of the field like a blind moving up or down. Recovery is usually complete. Most attacks are painless and associated symptoms rare. There may be a history of cerebral transient ischaemic attacks on other occasions. Emboli have been seen passing through the retinal circulation during an attack, moving from central to branch arterioles where they may disperse or become permanently trapped.

Clinical findings

Initially, the infarcted retina swells and becomes opaque leaving a 'cherry' red spot at the fovea where the intact underlying choroidal circulation shows through. The territory of an occluded branch artery may become whitened for several days or a few weeks (Plate 7), then subside to leave thinned retina and narrowed, often sheathed, vessels. The optic nerve head may atrophy over ensuing months as the nerve fibres die. If much of the retina is infarcted, there is a defect in the afferent pupil response which persists when other signs have subsided. Emboli may be visible at any stage, in the central or branch vessels, often at a bifurcation. Most emboli are small, glistening white or yellow pieces of cholesterol from atheromatous plaque. Larger, round, solid white emboli, which usually lodge proximally within or near the disc in the larger vessels, may have come from a calcified heart valve, whereas fibrin and platelet emboli from thrombosed plaque or cardiac thrombus may look dark or grey.

Associated findings include an ipsilateral carotid bruit, heart murmur, or dysrhythmia (particularly atrial fibrillation), absent pulses or bruits at other sites, and hypertension.

Investigation and management

No treatment is worthwhile acutely, apart from firm ocular massage which might dislodge an unstable central embolus. Fluorescein angiography is necessary only if the clinical picture is not typical. Risk factors are assessed by measuring blood pressure, full blood count, blood sugar, lipids, and renal function. Even in the absence of a bruit, carotid Doppler ultrasonography is useful for detecting atheromatous plaque at the bifurcation, with a view to carotid surgery.

Management involves reducing risk factors such as smoking, hypertension, obesity, or other abnormalities. Patients unsuitable for surgery are given long-term aspirin.

Retinal vein occlusion

Retinal venous occlusion usually occurs *in situ*. Risk factors include age, hypertension, diabetes, haematological disorders, and glaucoma (Table 5). Symptoms develop less abruptly than with arterial occlusion. Commonly the patient wakes with blurred vision. With a central vein occlusion, haemorrhages are scattered throughout the fundus, the characteristic 'bloodstorm' pattern (Plate 8), often with cotton wool spots. Less complete block causes sparse scattered haemorrhages, but foveal oedema may impair vision. Branch vein occlusion, usually at an arteriovenous crossing, causes a wedge-shaped sector of haemorrhage in the area of drainage, its apex towards the optic disc. Vision may improve, depending on the state of the fovea, and the outlook for the opposite eye is good if risk factors are minimized. If the retina is ischaemic, the risk of retinal new vessel formation may be prevented by laser treatment. Acute signs of occlusion may persist for many weeks or months. Curly collateral vessels may develop at the disc or peripheral retina.

Investigation and management

Blood pressure, blood sugar, full blood count, and erythrocyte sedimentation rate should be measured. Ocular pressure is checked as glaucoma is a treatable risk factor. Plasma protein electrophoresis, viscosity, and blood coagulation (including antiphospholipid antibodies or lupus anticoagulant) are tested in younger patients, particularly if the changes are recurrent, bilateral, or associated with thrombosis elsewhere. Other possibilities include sarcoidosis or Behçet's syndrome, particularly if the patient describes floaters or the slit lamp shows inflammatory cells within the eye. Risk factors such as blood pressure, smoking, and obesity must be addressed. The benefit of long-term aspirin in patients with venous occlusion is unproven.

Chronic ocular ischaemia

Eye ischaemia is associated with arterial disease anywhere from the aortic arch to the ophthalmic artery. The eye is often painful and red with impaired vision. On slit lamp examination, intraocular pressure is low and there are dilated vessels on the iris with protein flare in the anterior chamber. Cataract may obscure dilated and tortuous retinal vessels, often with scattered haemorrhages. New vessels may form. In younger patients this syndrome may suggest congenital or acquired proximal arterial occlusion, particularly Takayasu's arteritis. In older patients, surgical relief of stenosis may save and even improve vision.

Occlusion of vessels supplying the optic nerve

Giant cell arteritis is associated with occlusion of ciliary (rather than the central retinal branches of the ophthalmic artery) causing acute ischaemia of the optic nerve

(see below). Some patients have non-inflammatory occlusion from atheroma. Rarely, acute optic nerve ischaemia, sometimes bilateral, is associated with catastrophic postpartum or gastrointestinal haemorrhage; prognosis for recovery of vision is poor.

The eye in systemic inflammatory diseases

Sarcoidosis

External eye

Asymptomatic sarcoid granulomas may occur in the eyelids or conjunctiva. They are solid, raised, of variable size, often clustered, and characteristically yellowish in colour. Biopsies may be made of them at slit lamp examination. 'Blind' biopsy of normal-looking conjunctiva is not fruitful. Dry eye is common causing grittiness, reduced Schirmer's test, fluorescein staining of the cornea, perhaps with enlarged lacrimal glands.

Uveitis

Iritis (anterior uveitis) is common. It is usually bilateral and recurrent; sometimes severe and damaging. Acutely, the eye is red, painful, and photophobic. Large, greasy precipitates, said to resemble mutton fat, are seen on the internal surface of the cornea by slit lamp. Granulomas may be visible; Busacca's (in the iris), Koeppe's (at the pupil margin). Repeated attacks of iritis cause cataract, glaucoma, or particularly if there is also hypercalcaemia, calcified corneal band keratopathy.

Posterior uveitis causes cells to appear in the vitreous, noticed by the patient as floaters. These may obscure the view of the fundus and may aggregate into characteristic strands or 'snowballs'. Granulomas in retina or choroid are visible as pale foci behind a haze of cells. Vision may be further impaired by fluid leaking from inflamed retinal vessels and collecting around the fovea. Inflamed retinal branch veins look dilated, irregular, later sheathed, and may be surrounded by inflammatory cells; fluorescein angiography shows a segmental pattern of leakage. Occlusive retinal phlebitis of branch veins, causing focal retinal haemorrhages, strongly suggests sarcoidosis. Differential diagnoses are Behçet's syndrome or 'idiopathic' retinal vasculitis without systemic features.

Sarcoid can also cause an optic neuropathy. Sarcoid granulomas behind the eye may cause exophthalmos or cranial nerve palsy.

Corticosteroids are given topically for anterior and systemically for posterior lesions.

Behçet's syndrome (see also Chapter 18.10.5)

Uveitis (iritis and retinal vasculitis) is a defining feature of Behçet's syndrome. Recurrent attacks progressively damage the eye with risk of blindness. In Japan and Turkey, Behçet's syndrome is the commonest cause of uveitis. The incidence in northern Japan is about 1 per 10 000 population, of whom roughly 75 per cent have ocular inflammation at some stage. Untreated, blindness results in 50 per cent of eyes within 5 years of the first ocular attack. Males with the HLA B5 haplotype, particularly the BW51 subtype, are at highest risk of eye disease.

Ocular inflammation is usually bilateral, sometimes with a gap of many years between involvement of each eye. Iritis is typically acute with pain, redness, and photophobia. Hypopyon is characteristic but not unique to Behçet's syndrome (Plate 9). Attacks may settle spontaneously, but the eye may be damaged so short intensive courses of corticosteroid drops and mydriatics are recommended.

Retinal vasculitis particularly involves capillaries and branch veins. Inflammatory cells spill into the vitreous giving rise to floaters. Inflamed vessels leak, causing foveal oedema and an increase in visual impairment with distortion of central vision and risk of permanent foveal damage. Occlusion of inflamed branch veins causes haemorrhages and cotton wool spots (Plate 10). Vision is permanently affected if occlusion involves the fovea. Fluorescein angiography indicates severity of leakage and closure. With recurrent attacks, the retina gradually dies, vessels become sheathed, and the optic nerve head atrophies. Neovascularization may cause vitreous haemorrhage. Acute retinal infiltration with polymorphonuclear leucocytes causes white fluffy patches which indicate active inflammation, strongly suggestive of Behçet's syndrome.

Management is difficult. No regime of immunosuppression tolerable in the long term will prevent all inflammation. Systemic corticosteroids limit acute damage. Longer-term agents such as azathiaprine, clorambucil, colchicine, or cyclosporin A seem to be helpful but their impact on blindness is uncertain.

In a patient dying of cerebral involvement 10 years after the onset of treated eye disease there were collections of T_4 lymphocytes within and around walls of retinal vessels. Many cells were positive for interleukin 2, and HLA DR-positive cells were found in eye tissue, despite heavy immunosuppression.

Giant cell arteritis

Ischaemic, irretrievable visual loss is a feared complication. On systemic corticosteroid treatment the risk of visual loss and blindness falls, provided initial doses are adequate. Patients presenting with visual loss are at high risk of further loss in that eye or of rapid involvement of the second eye; they should be started on high doses until the symptoms, erythrocyte sedimentation rate, and C-reactive protein are controlled. This usually takes days rather than weeks. In patients presenting with bilateral involvement, intravenous methylprednisolone is justified. Temporal artery biopsy is val-uable and helps to confirm the diagnosis and the need for continued treatment in patients with visual loss.

Vision may be lost overnight or during the daytime. Patients may have experienced episodes of transient visual loss in the preceding weeks or days. Initially loss may be partial involving either the top or bottom half of the visual field, but often becomes total. The optic nerve head is characteristically pale and swollen (Plate 11). The afferent pupil response is usually decreased compared with the normal eye. Occlusion of the central retinal artery, producing a pale retina and cherry-red foveal spot, is uncommon.

Takayasu's arteritis

This is an inflammatory disorder involving large arteries which can cause an aortic arch syndrome with raised erythrocyte sedimentation rate and C-reactive protein, typically in younger patients. There may be amaurosis fugax or retinal vascular changes of chronic ocular ischaemia, sometimes with anastomoses at the optic nerve head, or scleritis or iritis.

Wegener's granulomatosis

Episcleritis is very common in the active stages and many patients notice that their eyes become red when the disease flares. Occasionally there is a more severe painful scleritis involving the cornea with the risk of corneal thinning and even perforation. Acutely the eye is red and the slit lamp may reveal infiltrates of inflammatory cells in the peripheral cornea. Scleritis and sight-threatening sclerokeratitis respond poorly to topical treatment and require systemic immunosuppression. A pulsed intravenous regime may be needed for initial control.

In patients with 'limited' Wegener's granulomatosis of the upper airway, the orbit may be involved, usually secondarily to disease in the adjacent sinuses but sometimes in isolation. Retro-orbital granuloma produces proptosis, usually painful, and may involve cranial nerves including the optic nerve, with an acute threat to vision (Plate 12). Some patients have a positive antineutrophil cytoplasmic antibody (ANCA) test, although the titres may be low. Many respond to immunosuppression, but high doses may be required for local control.

Polyarteritis nodosa

Inflammation of the eye coat, similar to Wegener's granulomatosis, may produce episcleritis, scleritis, or keratitis. Complications may be severe, requiring systemic immunosuppression. The retinal vessels may be involved, with or without hypertensive changes, and branch arteriolar closure is characteristic. Uveitis is not a

feature.

Relapsing polychondritis

This systemic inflammatory disorder involving cartilage is associated with inflammation of the eye coat, similar to rheumatoid arthritis. Half the patients will have eye features at some stage. Episcleritis and scleritis are most common, sometimes with severe corneal features similar to Wegener's granulomatosis and polyarteritis nodosa. Uveitis, retinitis, Sjögren's syndrome, and ischaemic optic neuropathy also occur.

Systemic lupus erythematosus

Episcleritis may occur with exacerbations in systemic lupus activity; the more serious scleritis is uncommon. Retinal vascular occlusions—particularly venous but sometimes branch arterial—may be linked with the antiphospholipid syndrome. Retinopathy with cotton wool spots is associated with active vasculitis, anaemia, and perhaps, moderate hypertension. Ischaemic optic neuropathy with acute irretrievable loss of vision is unusual. Some patients, particularly those with mixed connective tissue disease, have Sjögren's syndrome. Uveitis is not a feature of systemic lupus. Cutaneous lupus may involve the eyelids with oedema and the lid margins may develop scarring inflammatory plaques.

Dermatomyositis

Purple coloration of the eyelids and oedema of the lids and conjunctiva are typical findings. Less common is retinal ischaemia with microinfarcts.

Kawasaki's disease

Bilateral conjunctivitis without discharge is a cardinal feature of Kawasaki's syndrome, characteristically found in young children. Other features are fever, rash, lymphadenopathy, and changes in the other mucosa and nails. Cells may be found in the anterior chamber (mild iritis) and cornea (keratitis) using the slit lamp. The eyes do not need specific treatment.

Multiple sclerosis

Uveitis can occur in multiple sclerosis. Low-grade subtle changes, such as sheathing of the peripheral retinal veins with inflammatory cells within the vitreous, are common in patients with optic neuritis. Fluorescein angiography reveals inflammation and leakage of the retinal veins even though the retina itself does not usually contain myelin. Association of ocular and neurological features also occurs in sarcoidosis and Behçet's syndrome, but in these conditions eye inflammation is usually more pronounced.

Vogt–Koyanagi–Harada syndrome

This curious and uncommon clinical syndrome comprises deafness or meningoencephalitis with cerebrospinal fluid lymphocytosis in the acute stages and bilateral pan uveitis, almost exclusively in patients of Asian origin. There are HLA associations. Inflammatory cells collect within the retinal pigment epithelium and may cause fluid detachment of the retina or deeper layers, associated with decreased vision. There is a response to systemic corticosteroids, with relapses if the dose is reduced. In the chronic phase of the disease there is depigmentation of skin, hair, or eyelashes (poliosis).

Cogan's syndrome

In young adults, especially males, eye inflammation may be associated with deafness or vestibular dysfunction and proximal aortitis or inflammation of medium-sized arteries. Keratitis with patchy cell infiltration in the corneal stroma may lead to corneal vascularization, as in syphilitic keratitis. Some patients have anterior uveitis, scleritis, inflammation of the eye coat, or retinitis. The disorder responds to systemic corticosteroid, which may prevent total deafness if given early enough.

Inflammatory bowel disease

About 10 per cent of patients with Crohn's disease have episcleritis, scleritis, or iritis. Corneal or retinal inflammation is uncommon. Episodes of episcleritis may be associated with exacerbation of the bowel disorder. Eye problems are less common in ulcerative colitis.

Pancreatitis

Ischaemic retinopathy and acute visual loss may occur. There is retinal oedema with cotton wool spots. Fluorescein angiography shows closure of branch retinal arterioles and capillaries, with patches of retinal non-perfusion.

Whipple's disease

This rare disorder is suggested by the association of a malabsorbing enteropathy with arthritis, ocular inflammation, or particular neurological features. There are retinal haemorrhages, diffuse retinal and choroidal vasculitis with cells in the vitreous, or keratitis. Central nervous system features include cranial nerve palsies, papilloedema, and brainstem involvement. The diagnosis is confirmed by small bowel biopsy (see Chapter 14.9.6).

Rheumatoid arthritis

The commonest problem is keratoconjunctivitis sicca, apparently due to autoimmune damage to lacrimal tissue with lymphocytic infiltration and destructive fibrosis. The eyes are uncomfortable, gritty, and often sticky due to low-grade lid infection and poor flushing of the eye surface. Signs include reduced Schirmer's test and staining of the conjunctiva with fluorescein where epithelial cells are shed, particularly of the surface exposed between the eyelids. Symptoms may respond to topical tear substitutes containing methylcellulose: most common is hypromellose. In some patients filaments of adherent mucus may disperse with topical acetyl cysteine treatment (Ilube). Severe dry eye is best managed by a specialist who will watch for complications, particularly corneal ulceration. Other systemic conditions associated with Sjögren's syndrome include systemic sclerosis, mixed connective tissue disease, lupus erythematosus, and sarcoidosis.

Episcleritis is common and may indicate an exacerbation of systemic activity. It rarely needs treatment, but may respond to oral non-steroidal anti-inflammatory agents.

Scleritis, usually found in patients with active vasculitis, is more serious. The eye is usually painful, red, and boggy. The inflammatory process may spread from the posterior eye into the internal eye or to the orbit. Any patient with rheumatism and a painful eye should be referred for specialist assessment, even if the eye is white. Scleritis is an ischaemic vasculitic process involving the vessels which supply the sclera. It responds best to systemic immunosuppression with corticosteroids, sometimes with a cytotoxic agent. Pulsed intravenous treatment may be needed to control the acute attack. Untreated scleritis may cause scleral thinning (Plate 13), corneal ulceration, and perforation of the eyeball. Patients are rarely suitable for corneal grafting.

Ankylosing spondylitis

Ankylosing spondylitis is the most common association with iritis in young patients, particularly men, who should be asked about pain and stiffness of the spine or sacroiliac joints. Radiographs of lumbar spine or sacroiliac joints, or HLA B27 haplotype may be positive. One-third of patients with ankylosing spondylitis will develop eye features at some stage.

The eye is painful, aching, photophobic, and red. The slit lamp shows the cells diagnostic of iritis floating in the anterior chamber and sedimented on the back surface of the cornea as keratic precipitates (Plate 14). Posterior synechias may form, often with the iris constricted: a mydriatic may break these adhesions. This treatment should be continued until inflammation has settled so that the pupil remains large and the iris mobile. Inflammatory cells usually clear quite rapidly with 1 to 2 weeks of topical corticosteroids. Patients with ankylosing spondylitis should be warned that recurrent iritis should be treated early and effectively; there is a 50 per cent chance

of recurrence.

Reiter's syndrome

Arthritis, urethritis, cervicitis, or colitis together suggests Reiter's syndrome, especially in HLA B27-positive patients. A self-limiting sterile conjunctivitis is common in the early stages, causing a red sticky eye. Later, iritis may be the dominant recurrent feature. Features and management are similar to ankylosing spondylitis. Other differential diagnoses of arthritis with iritis include sarcoidosis, Behçet's syndrome, psoriasis, and gonorrhoea, with intestinal involvement, inflammatory bowel disease, or Whipple's disease. Sometimes, posterior uveitis, scleritis, or keratitis develop.

Juvenile chronic arthritis

Children most at risk have chronic, seronegative, pauciarticular disease, perhaps involving only one digit or an ankle, especially younger girls positive for antinuclear antibody. The picture is usually one of a low-grade recurrent iritis over several months or years. There may be no symptoms or redness of the eye in the early stages; slit lamp examination is essential. Cells appear in the anterior chamber when inflammation is active. Untreated inflammation can damage the cornea, lens, and aqueous drainage causing band keratopathy, cataract, glaucoma, and risk of blindness. Topical treatment may prevent secondary problems, but some patients will lose useful vision in both eyes. Other causes of iritis in children include sarcoidosis, ankylosing spondylitis, toxocariasis, leukaemia, and retinoblastoma.

Ocular features of blood disorders

Retinal changes are common in haematological disorders even if vision is normal. Bilateral changes result from anaemia, hyperviscosity, and haemostatic abnormalities. The signs are easily missed unless the pupils are dilated.

Hypoxia or hyperviscosity cause retinal vein enlargement, scattered retinal haemorrhages, and cotton wool spots: 'slow flow' or 'stasis' retinopathy. If blood pressure and blood sugar are normal, bilateral retinal haemorrhages suggest a blood disorder. Full blood count, erythrocyte sedimentation rate, and plasma protein electrophoresis should be checked. Roth spots, haemorrhages with a white centre, occur in leukaemia and hyperviscosity. Blood may leak in front of the retina to form a dense, rounded, often boat-shaped, subhyaloid blotch; leakage into the vitreous will cause floaters or clouding of vision.

Leukaemias

Although retinopathy is common in acute and chronic leukaemias, visual symptoms are unusual. The retinal haemorrhages are non-specific, though Roth spots represent focal collections of white cells (Plate 15). If the white cell count is very high, frank infiltration of the retina or optic nerve head causes pale fluffy areas. Leukaemic cells may spill into the vitreous. Chronic leukaemias may cause a slow-flow picture from chronic retinal hypoxia. Retinopathy may improve with chemotherapy. In acute lymphoblastic leukaemia, collections of cells may form a mass retro-orbitally (causing proptosis), or on the iris masquerading as iritis. Ocular infiltrations may respond to radiotherapy. Associated infections (such as orbital mucormycosis), chemotherapy, or radiotherapy may affect the eye. Bone marrow transplantation is associated with cataract and dry eye whilst graft-versus-host disease may cause conjunctival scarring and severe dry eye.

Lymphomas

Lymphoma may occur around or inside one or both eyes, in isolation, in disseminated disease, or in relapse; usually non-Hodgkin, low-grade, B-cell lymphomas. Externally, they form firm swellings in the eyelid or conjunctiva, resembling smoked salmon. In the orbit lymphomas may cause neuro-ophthalmic signs. T-cell tumours may infiltrate the internal eye, particularly iris or choroid. The rare ocular reticulum cell sarcoma ('histiocytic' lymphoma), can masquerade as uveitis, but with a pale mass in the choroid or retina visible through a cloudy vitreous. The monoclonal cells may spread from or to the brain, often the frontal or temporal lobes, so repeated cranial scanning is necessary. Immunocytochemistry of cells obtained by vitreous biopsy is diagnostic. Ocular lymphomas usually respond to local radiotherapy.

Bleeding tendencies

Pronounced or repeated subconjunctival haemorrhage, hyphaema, or vitreous haemorrhage suggests a bleeding diathesis. Bleeding may be spontaneous or follow minor trauma or eye surgery. In haemophilia, bleeding around or inside the orbit may compress the optic or other cranial nerves.

Clotting tendencies

Thrombophilias, including factor V Leiden, protein S, protein C, or antithrombin III deficiencies, are associated with retinal vascular occlusion, especially in the veins. Closure of choroidal vessels affects vision if fluid exudes to detach the retina; this pattern suggests thrombotic thrombocytopenic purpura or disseminated intravascular coagulation. Retinal venous or arterial occlusions occur in systemic lupus erythematosus with lupus anticoagulant or antiphospholipid antibodies. The optic nerve head may be involved, causing amaurosis fugax or ischaemic optic neuropathy. A full clotting screen is indicated in patients with retinal vein thrombosis if another site is involved, episodes are multiple, or there is a family history of juvenile thrombosis at any site.

Sickling disorders

Minor eye features are common in the sickling haemoglobinopathies and may assist diagnosis. Major eye features occur in less than half the patients. They are more common with haemoglobin SC and sickle-cell thalassaemia (**SThal)** than in haemoglobin S homozygotes (**SS**). Unilateral blindness is uncommon, even in SC patients; bilateral blindness is rare. As early treatment improves prognosis, screening the retina of high-risk patients is important. The risk of acute painful or chronic painless glaucoma is increased. Orbital infarction or pneumococcal ophthalmitis are rare. Patients may suffer a stroke affecting the visual field.

Conjunctival signs are more marked in haemoglobin SS than SC. Small conjunctival vessels develop linear, saccular, or comma-shaped dilatations, more prominent in children and after topical phenylephrine drops.

Bleeding into the retina causes a round 'salmon patch'. After resolution over several weeks, haemosiderin is left as iridescent spots in the superficial retina. Deeper haemorrhages damage the underlying retinal pigment layer leaving a permanent black 'sunburst' scar. Bleeds are usually asymptomatic and do not threaten vision.

Sickling in terminal branches of retinal arterioles produces signs in about half the patients. Their prevalence is related to age as most new vessels form between the ages of 10 and 25 years, rarely after 40. In SC patients blindness usually occurs between the ages of 20 and 30 years. Severe retinal ischaemia with new vessel formation is twice as common in haemoglobin SC and SThal as in SS. The risk of sight-threatening vitreous haemorrhage is related to the number and size of new vessels.

After their pupils are dilated, patients must be screened using the indirect ophthalmoscope or an accessory lens at the slit lamp. The earliest sign of peripheral ischaemia is closure of arterioles in the superior temporal sector (stage I) with a paler background and narrow white vessels. If these areas extend and become confluent, anastomotic loops form (stage II), then tufts of new vessels (stage III). As the tufts grow forwards into the vitreous they often look like coral 'seafans'. Fluorescein angiography reveals profusely leaking new vessels whose size can be assessed before or after treatment. Many new vessel tufts will autoinfarct from sickling in the feeder arteriole; they will not then bleed but others may form, so the patient must still be observed.

There are no symptoms until vitreous haemorrhage occurs (stage IV). Small haemorrhages produce a sudden shower of many small floaters, like 'midges'. Large haemorrhages cause sudden marked cloudiness with reduced red reflex. If the retina is distorted or detached (stage V) there may be flashes of light and a visual field defect. Some patients lose central vision, gradually with macular ischaemia or suddenly with foveal haemorrhage or central retinal artery occlusion. Retinal vein occlusion is not associated with sickling.

Screening and treatment

Annual retinal screening is recommended for patients aged 20 to 30 years, especially for SC and SThal diseases. New vessels should be reassessed every few months. If they do not autoinfarct and their size increases or vitreous haemorrhage occurs, laser treatment should cause regression and reduce the risk of early

blindness.

Infectious diseases and the eye

Organisms on the surface of the eye can be identified from swabs. Those inside the eye are identified from their pattern of involvement; it is rarely necessary to aspirate material from inside the eye. Treatment for superficial eye infections is by topical antimicrobial drops or ointments, whereas internal infections demand systemic therapy; the choice is partly dictated by penetration into the eye cavities. Rarely, drugs are injected directly into the vitreous to supplement systemic treatment in achieving high intraocular levels.

Bacterial infections

The commonest bacterial eye infection is conjunctivitis caused by *Staphylococcus aureus*, *Haemophilus* spp., or the pneumococcus. Topical chloramphenicol is effective as drops (hourly for the first 24 h then three times daily for several days) with ointment at night. Cellulitis is usually caused by the same organisms; *Haemophilus* is common in children; systemic amoxycillin is the regimen of choice. Retro-orbital spread is an ophthalmic emergency requiring admission to hospital for investigation. If the patient is systemically ill or has orbital signs (proptosis, double vision, or loss of vision), intravenous treatment is warranted.

Metastatic endophthalmitis results when bloodborne bacteria seed to the internal eye (Plate 16). The commonest sources are meningeal, urinary, and endocardial; the most likely organisms are staphylococci, *Neisseria* spp., or streptococci. *Bacillus cereus* and fungi (see below) may complicate intravenous drug abuse and unusual opportunistic organisms must also be considered in immunosuppressed patients. There is visual impairment with pain. Cells and debris within the eye chambers blur the ophthalmoscopic view, though a pale chorioretinal focus of infection may be visible. Blood, urine, and cerebrospinal fluid cultures are necessary. Tapping of the internal eye for vitreous microscopy and culture is justified in some cases. In infective endocarditis, retinal haemorrhages and microembolic infarcts are common, classically in the form of Roth spots.

Tuberculosis can cause indolent granulomatous uveitis, either of the iris or choroid. The eye is frequently involved in leprosy with a risk of blindness; specific iritis and cataracts occur in the lepromatous form. Corneal scarring complicates facial palsy and/or reduced corneal sensation.

In syphilis, uveitis or neuroretinitis occur in the secondary stage and optic neuropathy or Argyll Robertson pupils in the tertiary stage. Congenital syphilis is associated with interstitial keratitis and a salt-and-pepper retinopathy. *Leptospira icterohaemorrhagiae* commonly causes an early conjunctivitis with subconjuctival haemorrhages and a late uveitis. Late Lyme disease may cause ocular inflammation.

The gonococcus causes a marked purulent conjunctivitis in the newborn baby or in those sexually exposed. Tularaemia is associated with a severe granulomatous conjunctivitis with local lymph node enlargement. Botulism causes paralysis of the ocular muscles, sometimes with autonomic signs, with diphtheria as a differential diagnosis. Brucellosis can cause optic neuritis or uveitis—consider this especially in slaughterhouse or farm workers. Actinomycetes can infect the tear canaliculi. Rarely, *Nocardia* spp. can infect the internal eye with a focal chorioretinitis.

Chlamydial eye infection (see also Chapter 7.11.40)

Trachoma is the most common cause of chronic conjunctivitis and worldwide a preventable cause of blindness. At least 600 million people are infected, and about 6 million blinded. *Chlamydia trachomatis* serotypes A to C cause a chronic conjunctivitis with follicles which look like pale grains of rice in the conjunctiva. Scarring of the lids associated with inturning of eyelashes may accelerate corneal scarring. The diagnosis is confirmed by seeing inclusions in conjunctival scrapes or by culturing the organism from swabs. World Health Organization recommendations for control are topical tetracycline ointment twice daily for 7 days six times a year, or six doses of oral doxycycline at 5 mg/kg given monthly.

The genital serotypes of chlamydia cause acute conjunctivitis. The eye is red and sticky, and lymphoid follicles are found in the conjunctiva lining the eyelids. Infection is persistent, responding only partially to topical chloramphenicol. Systemic tetracycline (or erythromycin) with topical tetracycline is effective.

Viral infections and the eye

Adenovirus conjunctivitis is the most common viral infection. It is usually caused by highly contagious, potentially epidemic types 3, 4, 7, 8, or 19. The eye is acutely red and uncomfortable with scanty discharge. Lymphoid follicles may be visible in the conjunctiva lining the eyelids and the preauricular node may be enlarged. Symptoms may continue for some weeks, but recovery is usually uneventful and treatment rarely necessary. The most important differential diagnosis is chlamydial infection, in which eye discharge is more profuse.

In systemically ill patients with fever, rash, and red eyes, measles, meningococcal, or disseminated gonococcal infection may be implicated and in some parts of the world, relapsing fevers (borreliosis) or rickettsoses. Conjunctivitis with marked local lymphadenopathy (oculoglandular syndrome) may be attributable to adenovirus, chlamydia, mumps, or other rarer causes.

Primary herpes simplex can cause conjunctivitis. Secondary herpes infection is associated with recurrent attacks of dendritic corneal ulceration which can result in corneal scarring and poor vision, especially if treated with topical corticosteroid. Herpes zoster can cause corneal ulceration, iritis, glaucoma, and delayed cranial nerve palsies including optic neuropathy. Patients with ophthalmic shingles should be referred for slit lamp examination if there is red eye or visual impairment. All herpes viruses can cause retinal infection, particularly in the immunosuppressed patient; simplex and zoster can cause potentially blinding necrotizing retinitis, which progresses rapidly and may respond poorly to systemic antiviral therapy.

Cytomegalovirus causes progressive retinitis with characteristic haemorrhages and patchy retinal necrosis (see below). In transplant recipients, cytomegalovirus infection may respond to reduction of immunosuppression.

Measles may cause a scarring corneal inflammation, an important cause of blindness in undernourished children. Inflammation of the internal eye is less common. Neuro-ophthalmic associations occur in subacute sclerosing panencephalitis. Congenital rubella and varicella are associated with cataract and retinopathy in infancy. Iritis is characteristic of mumps. Molluscum contagiosum, cowpox, and orf can cause lid infection.

Fungal infections

Indolent fungal keratitis is associated with contact lens wear, diabetes, exposure to inoculation in the garden or field, and intravenous opiate abuse. Metastatic endophthalmitis is usually caused by *Candida albicans* in association with immunosuppression, irradiation, intravenous drug use/abuse, and poorly controlled diabetes. Small, dense, white 'snowballs' are seen in the vitreous with white foci of infection visible in the choroid and overlying retina (Plate 17). Retinal haemorrhage is uncommon. Diagnosis can be confirmed by vitreous biopsy which provides the opportunity to inject antifungals into the eye. The differential diagnosis of a white focus with hazy vitreous full of cells includes purulent endophthalmitis, toxoplasmosis, intraocular lymphoma, sarcoid, or tuberculosis.

Peri- and retro-orbital infection is characteristic of invasive mucormycosis in the same groups at risk of candida, especially debilitated patients receiving treatment for haemtological malignancies and in severe diabetic ketoacidosis. The infection spreads rapidly, often involving the vascular supply and producing tissue necrosis, particularly blackening of the hard palate. Medical treatment is combined with surgical debridement.

In endemic areas such as the Mississippi basin, *Histoplasma capsulatum* produces a multifocal scarring chorioretinitis described as 'histo spots'.

Rickettsial infection

There are petechial haemorrhages of the bulbar conjunctiva with marked redness. Retinal haemorrhages also occur.

Protozoal infections

Toxoplasma gondii causes congenital infection of the retina and underlying choroid, if the mother acquired a primary infection in pregnancy. This is especially common in France and Brazil. The primary scarring focus in the eye may involve the macula or optic nerve head resulting in congenitally poor vision. More commonly, an asymptomatic scar reactivates later in life, releasing cells into the vitreous (Plate 18). An inactive scar may be visible in the other eye. Patients presents with visual blurring, often describing 'floaters'. Presumptive diagnosis is based on clinical findings and positive blood serology. Acute attacks are best treated promptly by an ophthalmologist with several weeks of combined systemic clindamycin or co-trimoxazole and corticosteroid. Some infants have associated cerebral toxoplasmosis.

Retinal haemorrhages are commonly found in patients with cerebral malaria (see Chapter 7.13.2). Retinal toxicity has not been reported with standard use of antimalarials, but has been reported with chloroquine abuse.

Ulcers and nodules of cutaneous leishmaniasis may be seen on the eyelids in endemic areas. Retinal haemorrhages are common in kala-azar, particularly when there is associated anaemia.

Keratosis may occur in African trypanosomiasis (Chapter 7.13.10). In Latin America, oedema of the eyelids, lacrimal gland, and local lymph nodes (Romaña's sign) develops in the weeks following a periocular bite by a reduviid ('kissing') bug transmitting *Trypanosoma cruzi*, the causatial agent of Chagas disease (Chapter 7.13.11).

Acanthamoeba can cause an indolent and potentially blinding keratitis in contact lens wearers or after corneal abrasion.

Pneumocystis choroiditis is discussed in Chapter 7.10.21 and Chapter 7.12.6.

Helminth infections

Onchocerciasis ('river blindness') (see Chapter 7.14.1) is a common cause of blindness, particularly in Africa. Microfilarias lodge particularly in the choroid causing insidiously progressive destructive and scarring chorioretinitis. Lymphatic filariasis rarely affects the eye.

Nematode worms of *Toxocara canis* (see Chapter 7.14.7) form a visible mass beneath the retina with uveitis and sometimes whitening of the pupil. The differential diagnosis is a tumour such as retinoblastoma.

Some adult worms invade the eye surface. *Loa loa* (see Chapter 7.14.1) may be felt by the patient and be visible to an observer beneath the bulbar conjunctiva; it may be removed surgically. In Japan, the fly-transmitted 'oriental eye worm' *Thelazia* occurs in the conjunctival sac. In South-East Asia, *Gnathostoma* infects the eyelids or internal eye where the larvas may be visible with the slit lamp.

When pork is eaten, trichinosis (*Trichinella spiralis* infection) affects extraocular muscles, causing pain, periorbital oedema, proptosis, and defective eye movements. There may be internal eye involvement.

Sparganosis (see Chapter 7.15.4) can cause conjunctivitis, swelling, itching, proptosis, and blindness.

The larval form of cysticercosis may be visible inside the eye in either chamber, looking like a motile pearl or toxocara-like mass. Posterior uveitis and retinitis may occur. Orbital involvement is rare. Orbital cysts, perhaps calcified, may occur in patients with hydatid disease.

In schistosomatasis an urticarial conjunctivitis is associated with egg deposition, but the interior of the eye is rarely involved.

Myiasis can involve the eye and orbit (ophthalmomyiasis externa) (Chapter 7.17).

Human immunodeficiency virus infection and AIDS (see also Chapter 7.10.21)

Eye signs are common, particularly in the later stages of AIDS. The retina and optic nerve head are most commonly involved. In patients with coexisting central nervous system infection, eye features may prove an important clue. Definitive diagnosis in life is possible only by retinal biopsy, which is rarely if ever justified. The cellular response within the eye is much scantier than usual. Opportunistic pathogens tend to be facultative intracellular parasites. Cytomegalovirus retinitis is the most common problem in patients with AIDS in the United Kingdom and United States, but is rarely seen in patients with haemophilia or in Africa where non-specific retinopathy and herpes zoster ophthalmicus are more common.

Retinal microvascular disease is common. Cotton wool spots (Plate 19) are commonly seen in relatively early HIV infection with haemorrhages, Roth spots, and microaneurysms. There may be closure and sheathing of the peripheral retinal venous branches and occasional microvascular closure around the fovea producing visual loss; this retinal pattern is common in patients in Africa, particularly children.

Cytomegalovirus retinitis

This was the most common ocular complication in sexually acquired HIV infection in the United Kingdom and United States, affecting about one in three of such patients in the later stages of AIDS when the CD4 T-cell count fell below 100. The incidence has fallen and the prognosis improved strikingly since the introduction of highly active antiretroviral therapy (**HAART**) (see Chapter 7.10.21).

The typical appearance is of patchy areas of pale crumbled-looking retina with associated scattered haemorrhages (sometimes said to look like pizza), most commonly around branch vessels (Plate 20). These are areas of cytomegalovirus replication with oedematous and necrotic retina and some cells, mostly neutrophils and macrophages, although the cell response within the eye cavity is usually scanty. The patches spread contiguously from their borders and, untreated, may enlarge over the course of several weeks. Retinal death results eventually in blindness. Before HAART, the average life expectancy of patients with AIDS in conjunction with cytomegalovirus retinitis was about 9 months.

Differential diagnoses of these appearances include branch retinal vein occlusion, toxoplasmosis, early acute retinal necrosis, candida, or cryptococcus. Definitive diagnosis is difficult. Culture and serology discourage a diagnosis only if they are repeatedly negative.

Adequate doses of virustatic (cidofovir, ganciclovir, or foscarnet) damp down the infection. The lesions become atrophic with resolution of the pale and haemorrhagic features and arrest of spread. Ganciclovir treatment is initiated with 2 weeks of twice daily intravenous doses of 5 mg/kg. The dose is reduced if there is renal impairment. The daily maintenance dose is 5 mg/kg intravenously or 3 g orally. The most important complication is bone marrow suppression. To avoid toxicity, ganciclovir may be given by direct injection into the eye (vitreous), but this is rarely justified as the treatment must be repeated perhaps weekly and the infection is usually bilateral and elsewhere in the body. If the CD4 cell count remains low, breakthrough of retinitis is common and is treated by repeating the induction course of ganciclovir, or by switching to foscarnet or cidofovir.

Spread of retinitis will occur if treatment is interrupted and will often smoulder on during the treatment course; breakthroughs after several weeks are treated by increasing the dose or changing to another agent. Ophthalmic supervision is important as it is more accurate to assess the lesions by indirect ophthalmoscopy, preferably with serial retinal photographs to document progression at the edges of lesions.

In patients with cytomegalovirus retinits, the low CD4 cell count must be improved with HAART and once the count is securely above 100/µl, maintenance therapy for cytomegalovirus may usually be suspended.

Other infections

Especially in non-industrial countries, HIV infection is a common cause of herpes zoster ophthalmicus. Retinal infection with herpes zoster or simplex may also cause rapidly spreading retinal death, as acute retinal necrosis (**ARN**) with pale oedematous areas lacking the crumbled texture and haemorrhages characteristic of

cytomegalovirus retinitis. Often there is involvement of the optic nerve with optic neuritis and sometimes there may be encephalitis. Vision can be lost bilaterally, within days if untreated. Intravenous acyclovir may halt spread within the retina.

Syphilis in HIV infection may cause iritis or optic neuritis, often bilateral. There may be retinitis with a vitreous cell reaction, abnormal cerebrospinal fluid, or other signs of central nervous system involvement. Non-specific treponemal serology may be negative, so specific tests must be done. Eye complications are treated with benzathine penicillin, using a regime suitable for central nervous system infection.

Toxoplasmal choroidoretinitis in HIV-infected patients may show a fluffy focal retinal lesion with cells in both chambers. These signs may explain accompanying optic neuritis or encephalitis. Treatment is with clindamycin, without corticosteroids.

Pneumocystis pneumoniae may occasionally cause a multifocal choroidoretinitis with multiple, pale, rounded patches visible beneath the retina, and cryptococcus an acute optic neuropathy associated with meningitis.

Disorders of the thyroid and parathyroid

Eye signs result either from imbalance of thyroid hormones or from an immunological disorder of both the thyroid (Graves' disease) and retro-orbital tissues; the most common cause of proptosis/exophthalmos, referred to as 'Graves' ophthalmopathy' or 'ophthalmic Graves' disease'.

Cosmetic problems and eye discomfort are common, but a threat to vision may be an acute emergency. Evolution of eye signs is often independent of current thyroid status; the patient may be euthyroid, hyperthyroid, or hypothyroid, and correction of hormone imbalance may not affect eye features. Commonly, orbital disease appears in patients who have become hypothyroid after treatment for hyperthyroidism.

Orbital disorders result from infiltration of orbital tissues by T cells, stimulated by autoantibodies which cross-react with adipocytes. Initially, fibroblasts are stimulated to produce mucinous material and oedema within muscle or fat; later this leads to fibrosis and atrophy.

Werner's classification is as follows.

- Class 0, signs and symptoms both absent;
- Class 1, signs without symptoms;
- Class 2, both symptoms and signs of soft tissue involvement;
- Class 3, proptosis indicating orbital involvement;
- Class 4, eye muscle involvement with double vision;
- Class 5, secondary corneal involvement; and
- Class 6, optic nerve involvement with loss of vision.

The eyelids may be swollen in both hyper- and hypothyroidism, especially on waking. In hyperthyroidism, upper lid retraction reveals white sclera above the upper cornea and there is lid lag. Raised orbital pressure causes congestion and redness of the eye, particularly over the visible tendon insertions of the lateral rectus muscles and conjunctival swelling (chemosis) (Plate 21). Diplopia is common; it is usually vertical, worse on waking and looking upwards.

Proptosis causes white sclera to appear, often asymmetrically, below the lower corneal margin. This can be measured from the bony rim of the orbit using an exophthalmometer, which gives a useful impression of progression. Thyroid function and autoantibody tests may be normal in patients with typical eye disease. Scans, especially coronal MRI views, can show enlargement of ocular muscles. Severe protrusion with upper lid retraction exposes the cornea to damage; corneal abrasion, ulceration, and perforation can develop rapidly and so patients with a protruding eye, impaired blinking, pain, or fluorescein staining of the cornea are an ophthalmic emergency.

Orbital pressure may be highest in patients without much proptosis as protrusion has a decompressing effect. Optic nerve compression causes visual blurring, perhaps with a central scotoma, loss of colour definition, or relative afferent pupillary defect. The optic nerve head may be swollen. Scanning shows enlarged extraocular muscles. Urgent management is necessary.

Cosmetic orbital surgery is rarely justified, but upper lid surgery is sometimes worthwhile. Discomfort is difficult to treat; simple artificial tears may be tried. Immunosuppression seems justified for active inflammation. Initially, diplopia is best managed with a plastic Fresnel prism stuck on to a spectacle lens. Stable diplopia may need a permanent spectacle prism, surgery, or botulinum toxin injection. Corneal exposure demands lateral tarsorrhaphy or temporarily taping the lids or single lid suture under local anaesthesia.

Optic nerve compression needs urgent orbital decompression; medically, using high-dose systemic corticosteroid; surgically, by removing bone from the orbital walls, or by orbital radiotherapy in severe cases.

Another cause of a congested protuberant eye is an orbital mass, usually unilateral. Few conditions mimic thyroid eye disease in having bilateral if asymmetrical signs. Upper lid signs are particularly helpful in suggesting this diagnosis. Orbital pseudotumour or myositis is characteristically more painful and a caroticocavernous arteriovenous fistula may cause a frontotemporal bruit. The conditions are differentiated neuroradiologically.

Parathyroid disorders

Hyperparathyroidism producing hypercalcaemia may cause calcium deposition (band keratopathy), a lacy opacity spreading horizontally from the margins inwards. The eyes may be red and feel gritty.

In hypoparathyroidism and pseudohypoparathyroidism, hypocalcaemia causes lens opacities. Small white or coloured crystals beneath the lens capsule may not impair vision. Papilloedema from intracranial hypertension is rare and reversible by correcting hypocalcaemia.

Multiple endocrine neoplasia syndrome

In type IIb of this rare autosomal dominant condition, prominent corneal nerves are easily detected by slit lamp. Conjunctival neuromas or thickened eyelids may occur.

The eye in diagnosis of inherited conditions Table 6)

Marfan's syndrome (see Chapter 19.1)

Most patients have reduced vision, commonly due to myopia and astigmatism which may be inferred from their spectacle lenses. Slit lamp reveals lens dislocation upwards (Plate 22). The iris trembles with eye movement (iridodonesis), because it is poorly supported by an abnormally mobile lens. The dislocated lenses are best retained. Careful correction of focus can improve vision dramatically in early childhood, preventing permanent amblyopia. Differential diagnoses of dislocated lenses are isolated ectopia lentis (without other marfanoid features) and homocystinuria (with marfanoid habitus).

Neurofibromatosis type 1 (see Chapter 24.6.1)

Most patients have raised, yellowish/brown, multiple, Lisch nodules of the iris visible by slit lamp which must be distinguished from common, simple, flat iris freckles. Corneal, retinal, or orbital neurofibromas/schwannomas may be found. There is an increased incidence of glaucoma. Screening and management of intracranial tumours associated with neurofibromatosis type 1, including optic nerve or chiasmal gliomas, is controversial. Neurofibromatosis type 2 is associated with posterior

subcapsular cataracts.

von Hippel–Lindau disease

Retinal angiomas may be the presenting and sole features. They are usually bilateral and multiple in the mid-peripheral retina, so indirect ophthalmoscopy is advised. They start as a very small lesion no bigger than a microaneurysm which enlarge, later developing dilated, tortuous feeder and draining retinal vessels. Early peripheral angiomas may be destroyed by laser, preferably before they bleed or cause retinal detachment. Juxtapapillary angiomas at the optic nerve head may affect vision early in their development and are difficult to treat (Plate 23). Prolonged eye follow-up is needed.

The eye in diagnosis of inherited premalignant conditions

Gardner's syndrome is deep retinal pigmentation with polyposis coli; the dark retinal patches, multiple and usually bilateral, are best seen by indirect ophthalmoscopy. Absence of the iris (aniridia) is associated with renal Wilm's tumour, especially if there is a chromosome 11p deletion. Retinoblastoma, when associated with deletion of chromosome 13q, is inheritable as a dominant trait with high penetrance, together with other tumours such as osteosarcomas. In multiple endocrine neoplasia syndrome type IIB, associated especially with thyroid malignancy, there are enlarged corneal nerves and other ocular abnormalities (see above). von Hippel–Lindau disease is also associated with malignant renal tumours.

Ocular drug toxicity

Few drugs require ophthalmic screening but visual loss can result from poisoning.

Antimalarials

The risks of chloroquine are discussed in Chapter 7.13.2. Patients may develop a dose-dependent retinal toxicity of the 'bull's eye' type with permanent reduction in central vision (Plate 24).

Hydroxychloroquine (Plaquenil) carries a lower risk of retinal damage. Monitoring is recommended only in patients taking more than the standard dose of up to 400 mg daily.

Ethambutol (see Chapter 7.11.22)

Toxicity, rare at doses of 15 mg/kg or less, is usually related to total dose, and is therefore least likely in the first 6 months of treatment. However, idiosyncratic toxicity in the first few weeks has been reported. Patients with symptoms of optic neuropathy should stop taking the drug immediately and should be warned to report any change in visual clarity or colour vision. Visual acuity, colour vision (100 Hue test), and optic disc appearance are monitored 3-monthly during treatment. Previous optic nerve damage or renal failure increase the risk. The early changes usually recover if the drug is stopped.

Corticosteroids

Lens opacities, typically posterior subcapsular, producing light scatter and glare, are visible by slit lamp or the ophthalmoscope set to catch the red reflex in focus, particularly if the pupil is dilated. Surgery may be necessary.

Eye signs in poisonings (see Chapter 8.2)

Quinine poisoning can blind by damaging the retina and optic nerve head. Ethyl alcohol (antifreeze) damages the optic nerve. Cyanide in raw cassava can blind.

Organophosphate pesticides and other anticholinesterases constrict the pupil by parasympathetic stimulation. Vision is not affected. Opiates also constrict the pupil. Atropine-like compounds, including nightshade berries, dilate the pupils ('gardeners's mydriasis') (see Chapter 8.3).

The external eye is vulnerable to ammonia, alkalis (including lime, cement, and plaster), acids, and some riot control agents. Primary treatment is prolonged irrigation for up to 20 min, using tap water (or even milk if necessary), followed by specialist referral.

Blindness worldwide

The World Health Organization defines blindness as binocular vision of Snellen 6/60 or less. More than 45 million people are estimated to be blind worldwide; perhaps two-thirds of these have visual impairment that prevents self-sufficiency.

Trauma

Physical or chemical eye injuries are a common cause of visual loss where prevention and treatment are poor. Emergency eye surgery and antibiotics can save vision.

Cataract

There is no known method of preventing cataract, which remains the most common cause of blindness in populations with limited surgical services. As the lens ages, its protein structure changes and the lens opacifies. Malnutrition, dehydration, diabetes, and perhaps sunlight accelerate lens ageing. In some populations, primitive surgical 'couching' or dislodging of the lens within the eye makes matters worse. Implanting artificial lenses is impracticable in many countries and, even if surgery is successful, correction by spectacles needed afterwards is often unsatisfactory. In poorer countries attention is focused on organizing the training and deployment of mobile surgical teams to carry out as many effective operations as possible.

Glaucoma

Nerve fibres within the rim of the optic nerve head are damaged by relatively high intraocular pressure and impaired blood supply. The central cup of the nerve head enlarges and visual field is irretrievably lost long before central acuity is affected, so the early stages are usually asymptomatic and painless. Even in wealthy populations, screening for early glaucoma is difficult and some patients progress to blindness despite all efforts. Those most at risk have a first-degree relative with glaucoma.

Age-related macular degeneration

The central retina around the fovea has an extraordinarily high metabolic turnover. In some patients, the efficiency of recycling metabolic products fails, abnormal material is deposited in the retina, and tissue integrity breaks down. Drusen may form around the fovea. They do not impair vision themselves but may herald formation of aberrant vessels which grow into the fovea from the choroid beneath, and may leak and may leak or bleed to form a scar with an irregular 'disciform' shape. This permanently damages the fovea. The patient loses detailed central vision, although peripheral vision allows them to navigate independently. Only a few patients at an early stage are ever likely to benefit from laser coagulation.

Diabetic retinopathy

This remains an important cause of blindness in younger patients. Major problems are organization of effective screening programmes and deployment of laser treatment. Over 50 per cent of blindness is preventable if laser treatment is given early enough.

Trachoma

Blindness occurs in populations who have poor eye hygiene and repeated fly-borne infection. Vaccination is not feasible. The only effective means of control is intermittent topical or systemic tetracycline treatment. Surgery for established scarring is less effective on a mass scale.

Vitamin A deficiency (xerophthalmia)

This most commonly affects young children and may destroy the whole eye. Lack of vitamin A causes conjunctival and corneal dryness and keratinization (xerosis) with Bitot's spots. The cornea softens (keratomalacia), ulcerates, and may perforate and become infected, resulting in endophthalmitis. Vitamin A is found in many green leafy vegetables and palm oil. Measles keratitis increases the risk of corneal scarring in malnourished children.

Onchocerciasis (river blindness)

This is estimated to cause blindness in 1 million people. Recently the disease has been largely controlled, thanks to vector control and the widespread use of ivermectin.

Further reading

Easty DL, Sparrow JM, eds (1999). *Oxford textbook of ophthalmology*. Oxford University Press. [For reference purposes.]

Fraunfelder FT, ed. (2000). *Drug-induced ocular side effects and drug interactions.* Lea & Febiger. [A compendium, regularly updated.]

Frith PA, ed. (2001). *The eye in clincal practice*, 2nd edn. Blackwell Scientific Publications. [A basic practical exposition for non-specialist clinicians.]

Nussenblatt RB, Palestine AG (1995). *Uveitis, fundamentals and clinical practice*, 2nd edn. Year Book Medical Publishers Inc. [For clinical aspects and conundrums.]

Taylor D, ed. (1997). *Paediatric ophthalmology*, 2nd edn. Blackwell Scientific Publications. [For comprehensive clinical coverage.]

26.1 General introduction

Michael Sharpe

Further reading

Modern psychiatric medicine represents a substantial body of knowledge and skills, much of which is of potential value to the physician. It is therefore unfortunate that psychiatry and medicine have become so divorced from one another. Based on the questionable intellectual foundations of mind–body dualism, this separation has shaped research, planning, and services. In recent years, the split seems to have widened as medicine has focused increasingly on the basic biology of disease and psychiatric services have tended to make psychoses such as schizophrenia their central concern. Despite this, there is much evidence to show that physicians are faced every day with diagnostic and management problems for which the larger body of psychiatric knowledge and skills are relevant. Indeed, it could be argued that much of the criticism of clinical medicine in recent years reflects a lack of attention to the non-biological aspects of patient's illnesses. Examples include failures to establish therapeutic relationships with patients, to properly manage distress, and to effectively understand and manage complaints that are medically unexplained.

Psychiatric knowledge of relevance to medicine includes a practically useful, if imperfect, classification system, with an increasingly sophisticated psychological and neurobiological underpinning, and a range of pharmacological and psychological treatments, each with a substantial evidence base. Skills of relevance to the physician include the ability to assess the patient's mental as well as physical state, the detection of symptoms of depressive and anxiety disorders, and the eliciting of the patient's own understanding of their illness. These factors can be of substantial importance in the patient's management. Attitudes are also important, it being no secret that some physicians are dismissive toward patients who are perceived as 'psychiatric'. The acceptance of patients' concerns, even if apparently illogical, the tolerance of difficult behaviours, and a degree of reflectiveness in moderating one's own response; all are valuable in all medical settings.

Although physicians do a great deal of 'psychiatry' themselves, specialist help is not infrequently required, but psychiatric and psychological services can be problematic for the hospital physician to access because of the separation of psychiatric and medical services. In recent years, however, there has been a slow but steady growth of general hospital-based psychiatry and psychology services dedicated to the needs of medical patients, often termed 'liaison psychiatry' or 'health psychology', respectively. Such services offer improved integration of medical and psychiatric practice.

The sections that follow hopefully provide a practical and accessible summary of those aspects of assessment and management conventionally deemed 'psychiatric', but which are in fact central to the practice of medicine: They include:

- guidance on taking a psychiatric history from a medical patient in a way that is manageable within a medical setting;
- information about relevant psychiatric diagnoses, including organic mental disorder, depression and anxiety, reactions to stress, somatoform disorders and eating disorders, as well as basic coverage of the less commonly encountered but important psychiatric diagnosis of bipolar disorder, schizophrenia, and obsessive–compulsive disorder;
- practical advice on the management of depression and anxiety when it coexists with disease, somatic symptoms that are unexplained by disease, deliberate self-harm, and on how to cope with acute behavioural emergencies;
- a substantial section on the highly prevalent and clinically significant problem of alcohol and substance misuse.

Some might regard this section as an 'add on' that is of questionable relevance to the practising physician. Rather, it is better considered as a more detailed look at some aspects of medicine that are of relevance to the management of many of the medical conditions described in this book. It therefore forms part of an integrated approach to illness in which biological, psychological, and social strands of patient assessment and management run in parallel.

For the interested reader who wishes to find out more about psychiatry in general or liaison psychiatry in particular standard texts are listed below.

Further reading

Gelder M *et al*. (1996). *Oxford textbook of psychiatry*, 3rd edn. Oxford University Press, Oxford.

Royal Colleges of Physicians and Royal College of Psychiatrists (1995). *The psychological care of medical patients; recognition of need and service provision*. Royal College of Physicians, London.

Rundell JR, Wise MG, (1996). *Textbook of consultation–liaison psychiatry*. American Psychiatric Press, Washington DC.

Sharpe M (1998). Psychiatry in relation to other areas of medicine. In: Johnstone JC, Freeman CPL, Zealley AK, eds. *Companion to psychiatric studies*, 6th edn, pp 785–806. Churchill Livingstone, Edinburgh.

26.2 Taking a psychiatric history from a medical patient

Eleanor Feldman

> Listen to the patient, he is telling you what is wrong
>
> William Osler

This chapter covers issues that physicians and surgeons need to know about concerning psychiatric history-taking in general hospital patients. It would not be appropriate for a non-psychiatrist to attempt to take a full psychiatric history, involving as it does at least one hour's discussion covering relationships in the family of origin and a detailed biography to establish premorbid personality and aetiological factors, plus further discussion with at least one other informant. However, all patients should be screened for the most common problems: cognitive dysfunction, mood disorder, anxiety states, and alcohol and substance misuse. The assessment of cognitive dysfunction is predominantly a matter of mental state examination rather than taking a history and is covered in Chapter 26.4 and Chapter 30.2. Substance misuse is covered in some detail in Section 26.7. Chapter 26.5.2 covers how to assess a patient following attempted suicide, and the diagnostic features of patients with eating disorders are discussed in Chapter 26.5.5.

It is not necessary to screen routinely for psychotic symptoms as functional psychosis rarely presents for the first time in general hospital cases. If hallucinations and delusions do emerge during an inpatient's stay, then the most likely cause is an acute organic brain syndrome, and careful testing of orientation in time and observation for fluctuations in conscious level will usually confirm delirium. If in doubt, psychiatric advice should be sought.

Screening questions in routine assessment

Depression and anxiety

Significant proportions of general hospital patients will have diagnosable mental health problems. (Table 1 and Table 2) Frequently these will impinge on the physical health and well being of the patient. Stress affects the immune response; depression and anxiety are often comorbid with physical illness, either preceding it, or arising largely as a result of it; depressed and anxious people frequently have increased worry about physical health and experience minor physical symptoms as severe and intolerable. Antidepressants may help.

Screening for depression and anxiety need not take much time in itself (Box 1). However, if the patient gives positive answers it is helpful for these to be explored in more depth when time allows and in a private interview room. Patients will be aware when you are under pressure or in a rush, and this will inhibit them from telling you important things and you from wanting to hear about them. Therefore, indicating that you think something is important and will come back to them when you can set aside more time is very helpful and reassuring to the patient.

Questions about mood disorder are best construed as part of an enquiry into general health and the 'person as a whole', and most patients are pleased to discover that their physician takes an interest in their general well being. Starting with a non-directive enquiry about sleep **before** coming into hospital (most patients have sleep disturbance **in** hospital) is a natural way to link physical and emotional health: difficulty sleeping is a common denominator in stress, anxiety, and depression and a description of a disturbed sleep pattern may also assist you in distinguishing endogenous depression characterized by early morning wakening and diurnal mood variation. The reasons for any sleep disturbance, whatever they may be, are important in general health and will often reveal what troubles and worries a patient. Again ask this non-directively; do not be tempted to offer the patient a multiple choice of explanations. By open questioning, you will guide the patient into revealing what difficulties they are facing and most will find it a relief to tell you. At this stage, it is best to listen empathically and let the patient tell you their story. It may seem to you that you are doing nothing, but the patient gains great relief by being heard and understood, and you are gaining their trust and eliciting valuable information Box 1.

Box 1 Screening and probing questions for mood and anxiety disorder

- **Screening for current problems**
- How have you been sleeping (before you came into hospital)?
- **Probing: sleep, worry, and mood**
- *If not sleeping well, ask about the pattern of, and perceived reason for, sleep disturbance:*
- Is it difficulty getting off to sleep?—If Yes: how long before you fall asleep?
- Are you woken intermittently? Why? (may be due to physical symptoms)—Do you get back to sleep easily?
- Do you wake early and find you can't get back to sleep?—If Yes: how early?
- *If they have good nights and bad nights:*
- What proportion are good or bad; is it 50:50 or better or worse than that?
- *If sleeping badly:*
- Why do you think you are sleeping badly? (This is a natural opportunity for patients to reveal what is worrying them)
- Are you kept awake with worries going round and round your mind?
- At this point you will find out what is bothering the patient—after they have confided this in you, it is then empathic to ask:
- How have you been feeling in your spirits?
- If a patient reveals problems, it is appropriate now to ask about their previous mental health.
- **Screening for past mental health problems:-**
- Have you ever had trouble with your nerves?
- Ever seen a doctor about your nerves?
- Ever taken tablets for your nerves or to help you sleep?

The rare patient who objects to a sensitive enquiry intended to be helpful will have a reason to be defensive, and a history from the family practitioner or another informant will usually explain all. If you find the patient defensive and there always seem to be reasons why there is no-one else you can talk to, so that you cannot even confirm the patient's identity, then this is characteristic of those with factitious disorder using a false identity.

It is also advisable to screen for a past history of mental disorder, either when screening for the patient's past medical history, or if you have already elicited current mental health problems. Use the vernacular language of the culture of that patient; for example, in my own culture it would be: 'Have you ever had any troubles with your nerves or had to see a doctor about your nerves?'. Positive answers should then be explored non-directively by saying something like: 'Tell me more about that'. Hence, there need be only a few screening questions to lead effectively into a discussion of most mood disorders, worries, and stress. As important as the questions themselves will be, the way in which they are asked, the time available for discussion, and the physician's own willingness to listen and take note, are equally—if not more—important.

The importance of the alcohol use history

Insomnia and mood and anxiety disorders may be associated with heavy alcohol use. This alone may depress mood and give rise to early morning wakening, appetite changes, and weight loss. For this reason, a diagnosis of depressive illness cannot be made until the patient has been through alcohol withdrawal and been dry for a few weeks. If the alcohol problem is missed, incorrect advice and treatment will be given.

Patients are often defensive and evasive when they have an alcohol problem and persistent probing is needed to elicit the precise amount. Questions need to be asked as a routine and in a non-judgemental friendly way (Box 2). It is important to know the average intake in terms of units, rather than a vague qualitative response such as 'social drinking' or 'moderate drinking'. The CAGE questions may also be asked as a routine in anyone who drinks excessively (also in Box 2).

Box 2 Screening for alcohol problems

- **Routine questions**
- On an average week, how much alcohol do you drink?
- What do you like to have (spirits, wine, beer, etc.)?
- How many measures/glasses/pints?
- How often do you drink that much?
- Did you used to drink much more than that?
- If so, when and how much?
- **CAGE questions**:
- Do you feel you should **C**ut down on your drinking?
- Does anyone **A**nnoy you or get on your nerves by telling you to cut down your drinking?
- Do you feel bad or **G**uilty about your drinking?
- Do you have a drink first thing in the morning to steady your nerves or get rid of a hangover (**E**ye-opener)?
- *A positive answer to one or more questions is indicative of problems*

Recognizing and dealing with somatized anxiety and depression

Stress, anxiety, and depression can also be the main reasons for physical symptoms (Table 3). When patients experience the unpleasant physical symptoms that arise from their bodily reactions to emotional states, it is natural for them to complain of these to their medical attendants rather than present their primary complaint as an emotional disorder, which they will often regard as secondary to their physical symptoms, or make no connection at all. This is the common phenomenon known as somatization.

Misinterpretation by patients and their doctors of the symptoms of chronic tension, as well as sympathetic hyperarousal and hyperventilation in panic disorder, may lead to inappropriate extensive searches for organic abnormalities and provide no relief from suffering for the patient. Indeed, the patient's suffering increases as they are left with continuing uncertainty as to the cause of very distressing symptoms, and they become sensitive to the increasing scepticism and exasperation displayed by their doctors. In a study at a cardiac clinic in London, 50 to 60 per cent of patients had normal cardiac function on investigation, and many of these were experiencing the palpitations, dyspnoea, and chest discomfort of anxiety; 21 per cent showed evidence of hyperventilation. The common physical effects of hyperarousal and hyperventilation affect most organ systems and parts of the body from head (ache) to toe (tingling) (see Table 3), and so these patients find their way into every specialist clinic in the hospital.

If you have patients with such unexplained physical symptoms, then the most acceptable way of exploring the possibility of these syndromes is to enquire systematically about all the common symptoms listed in Table 3. Patients often have a few more symptoms that are not on the list and are usually not bothered by all. Once organic causes have been excluded, the diagnosis of panic disorder can be made on the history of physical symptoms alone, and an explanation can be given to the patient in terms of the physiological effects of adrenaline plus overbreathing. It is particularly important that the physician makes it clear that the symptoms are genuine, and that the tests only show that nothing is wrong because we do not use tests for the transient physiological changes accounting for the symptoms.

Treatments for panic disorder include low doses of antidepressants and cognitive–behavioural therapy, and give very good results. The majority of patients will accept psychiatric referral if it is made clear: that you **do** believe their symptoms are real (which they are); you are **not** saying that they are mad, or making it up, or it is all in their minds; and that there **is** a treatment that is effective and not addictive or harmful. It helps if you know the psychiatrist and know what the patient is likely to experience. Left untreated these patients may become severely disabled and continue to undergo expensive and unnecessary medical investigation.

Recognizing depression in someone with good reasons to be unhappy

To be ill and in discomfort is generally to be unhappy, and the more serious the illness, the greater the pain, the greater the loss and the tragedy, so much more the misery. Worry about health and fears for the future are to be expected. How can a doctor tell when an ill person's unhappiness amounts to a depressive illness requiring specific intervention, and when is it an appropriate adjustment reaction to grievous circumstance? Some of the cardinal diagnostic features of depressive illness much emphasized in general psychiatric practice are of little use in these circumstances, so that in patients with physical reasons for poor appetite and weight loss, and sleep disturbed by pain or other physical symptoms, we must look for other indicators of mood disorder. Endicott has thus suggested modifications to the diagnostic criteria for depression (see Box 3).

Box 3 Endicott's criteria for depression in the medically ill

- Presence of five out of these nine symptoms for at least 2 weeks:
- *Fearful or depressed appearance*
- Social withdrawal or decreased talkativeness
- Psychomotor retardation or agitation
- Depressed mood, subjective or observed
- Marked diminished interest or pleasure in most of the activities, most of the day
- *Brooding self-pity or pessimism*
- Feelings of worthlessness or excessive or inappropriate guilt
- Recurrent thoughts of death or suicide
- *Mood is non-reactive to environmental events*
- Symptoms in italics replace DSM-III-R* symptoms as follows:
- weight change
- sleep disturbance
- fatigue or energy loss
- diminished ability to think, concentrate; indecisiveness
- **Diagnostic and statistical manual of mental disorder, third edition, revised* (1987).

The emphasis is placed on the predominantly negative thinking style that is pervasive. A depressed person cannot be cheered up when nice things happen, will not be interested in things he/she used to enjoy, will be excessively pessimistic, and have an exaggerated sense of guilt and worthlessness.

Final comment

It is important to screen for mental health problems: they are common in general medical patients and failure to recognize and deal with them will often interfere with the management of the physical health of the patient. Moreover, depression itself can kill, and if by screening you identify depression, you should ask about hopelessness and suicidal ideation. How to do this sensitively is covered in Chapter 26.5.2.

Further reading

American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorder, third edition, revised*. APA, Washington, DC.

Bass C, *et al*. (1988). Panic anxiety and hyperventilation in patients with chest pain, *Quarterly Journal of Medicine* **69**, 949–59.

Endicott J (1994). Measurement of depression in out-patients with cancer. *Cancer* **53**, 2243–8.

Feldman E, *et al*. (1987). Psychiatric disorder in medical in-patients. *Quarterly Journal of Medicine* **63**, 405–12.

Mayfield DG, Johnstone RGM (1980). Screening techniques and prevalence estimation in alcoholism. In: Fann WE, *et al*., eds. *Phenomenology and treatment of alcoholism*, pp.33–44. Spectrum, New York.

Van Hemert AM, *et al*. (1993). Psychiatric disorders in relation to medical illness among patients of a general medical outpatient clinic. *Psychological Medicine* **23**, 167–73.

26.3 Neuropsychiatric disorders

Laurence John Reed, Tom Stevens, and Michael D. Kopelman

Introduction

'Neuropsychiatry' is concerned with disorders of affect, cognition, and behaviour that arise from overt disorder in cerebral function, or from indirect effects of extracerebral disease. The term has largely replaced the earlier expression 'organic psychiatry', which originated in the classification of mental disorders as either 'organic' or 'functional' on the basis of the presence or absence of pathological changes in the brain. The latter distinction has become increasingly ambiguous as a result of the development of new methods for detecting abnormal brain pathology and pathophysiology in so-called 'functional' disorders such as depression and schizophrenia. Indeed, the most recent version of the *Diagnostic and statistical manual of mental disorders, fourth edition* (DSM-IV, American Psychiatric Association) states, 'the term organic mental disorder is no longer used in DSM-IV because it incorrectly implies that 'non-organic' mental disorders do not have a biological basis'. Nevertheless, a creative conflict has always pertained between neurological and psychological theories of behaviour and, in the absence of satisfactory alternatives, these terms have retained a place in clinical practice. In part, this serves to demarcate the uneasy and shifting boundary between disorders predominantly diagnosed and managed by physicians and psychiatrists, respectively.

This chapter is divided into two parts. First, we provide a consideration of practical issues related to the assessment, investigation, and management of patients manifesting cognitive and behavioural change. Second, we discuss specific cerebral and extracerebral disorders that commonly involve or are accompanied by cognitive or behavioural change. This bipartite organization is intended to help in the identification of possible diagnoses causing particular behavioural features, and also to alert clinicians to the likely neuropsychiatric sequelae of specific medical disorders.

Assessment, investigation, and management of patients with cognitive and behavioural change

The assessment and classification of mental and behavioural disorders is a frequent source of misunderstanding and confusion for clinicians, the process being undermined by the absence of robust clinical and laboratory markers for these conditions. Moreover, the clinical terminology used to describe certain symptoms and signs (such as 'confusion') is often unsatisfactory and unreliable. Although the major systems of classification are broadly similar, they continue to use different terminology: for instance, the World Health Organization's *International classification of diseases, tenth edition* (ICD-10) retains the term 'organic disorders', whereas DSM-IV uses the broad grouping 'delirium, dementia, amnesic and other cognitive disorders'. In addition, some of the operational diagnoses may have little validity in assisting the clinician to determine the appropriate investigation and treatment, for example the ICD-10 'unspecified organic personality and behavioural disorders due to brain disease, damage and dysfunction'. Nevertheless, there is consensus on the essential clinical features of these disorders, and in this section we shall describe their assessment on the basis of a number of core features underpinning the differential diagnosis.

Differential diagnoses

Acute versus chronic disorder

The differentiation of acute and chronic cognitive disorder essentially determines the boundary between delirium and dementia. This distinction should be apparent from the history and mode of presentation, although difficulties may arise where a clear history is lacking due to disturbed communication or the absence of an adequate informant. However, they can usually be distinguished on the basis of the characteristic clinical features described below. Essentially, a conspicuous impairment of attention is typical of an acute disorder, together with a fluctuating course and prominent perceptual disturbance. However, the 'acute on chronic' disorder, where there is a delirium superimposed on a chronic cognitive disorder, should not be overlooked.

Cognitive versus psychiatric disorder

This distinction between a cognitive and a psychiatric disorder is not always easy. It is important to recognize that an apparent cognitive abnormality may be seen in psychiatric disorders such as schizophrenia and depression. In depression, impairment of memory and concentration together with somatic complaints may lead to a misleading impression of dementia, so-called 'pseudodementia' or 'reversible dementia'. Likewise, the distinction between acute psychotic disorders and delirium can be difficult where both conditions show behavioural disturbance and disturbed communication. The risks associated with the wrongful categorization of delirium as psychosis are high: delirium is a medical emergency with high morbidity and mortality, and it is potentially reversible. Likewise, the attribution of a psychiatric disorder as delirium or dementia bears costs in terms of performing unnecessary investigations and pursuing the wrong therapy in an inappropriate setting, thereby compounding any illness behaviour.

Specific versus generalized cognitive impairment

If cognitive impairment is identified, it needs to be determined whether this is generalized to many cognitive functions or affects a specific function such as memory, planning, perception, language, or attention. Identification of a specific impairment, such as the amnesic syndrome, offers important clues as to the aetiology and management and is more likely to result from a focal brain lesion (as opposed to an extracerebral disorder).

Reversible versus irreversible

The range of causes of any cognitive impairment needs to be fully assessed. In particular, it is essential that those conditions that can be reversed or arrested should be specifically considered, for example human immunodeficiency virus (HIV) infection and cerebral neoplasms. It is equally important that any treatable psychiatric disorder is identified.

Acute cognitive and behavioural change

Assessment

A wide range of disorders may cause acute emotional and behavioural disturbance (Table 1). One of the most problematic aspects of assessment is the distinction between an acute psychotic episode and delirium. The clinical features of delirium (also known as 'acute organic brain syndrome' or 'acute confusional state') and of the 'functional' psychoses typical of schizophrenia or affective disorder share a number of characteristics. First, both involve a pervasive disruption of thought, cognition, communication, and behaviour in the patient, hence presenting particular difficulties in assessment. Second, both conditions may involve abnormalities of

perception in the form of hallucinations or illusions; abnormalities of belief, in the form of delusions or overvalued ideas; psychomotor abnormalities, including hypo- or hyperactivity; disturbance of the sleep–wake cycle; and emotional disturbance encompassing the range from depression to irritability and euphoria. These similarities cause practical difficulties in diagnostic differentiation, and they also hint that an absolute distinction between 'functional' psychosis and 'medical' delirium is probably untenable.

Delirium

A thorough history of the antecedents and onset of any behavioural and mental disturbance, as well as details of any past medical or psychiatric contact, will yield important clues as to the likelihood of an organic aetiology to behavioural change. The unco-operative or mute patient presents a particular challenge as important historical details may not be forthcoming, such as head injury, substance misuse, foreign travel, diabetes, or other medical disorders. Furthermore, accurately eliciting a mental and cognitive state is problematic, and it is in this group that a history from an informant, ward staff, or relatives is especially important. The diagnosis of delirium should be suspected where the history of behavioural disturbance is of recent onset, fluctuating, and there is evidence of deterioration at night. Difficulty in communicating with a patient is frequently the first indication of an underlying delirium.

The elderly and general hospital inpatients are particularly vulnerable to delirium. Any change of environment such as a recent admission to residential care or pre-existing cognitive impairment will heighten this vulnerability. Visual and hearing impairments are also more frequently observed in this group. Amongst inpatients the problem is compounded by inadequate information, impersonal environments, and confusing exposure to a myriad of different professionals. It is common for the diagnosis to be missed where there is no overt agitation or antisocial behaviour.

Behavioural changes seen with delirium include irritability, repetitive purposeless movements, and disorganization or difficulty performing routine tasks such as undressing. It is important to recognize that patients may be both overactive and noisy or inactive and slow.

The predominant clinical feature of delirium has been described as 'clouding of consciousness' or 'clouding of the sensorium'. These terms lack consensus definitions or clarity in practice, but have traditionally been used to describe a combination of orientation, attention, and memory deficits. Consequently, deficits of attention are stressed in diagnostic criteria, which in the delirious patient may range from distractibility and inability to follow complicated conversations, through an almost complete inability to register information or to concentrate, progressing in the extreme case to diminished consciousness and coma. Furthermore, such attentional difficulties tend to have a sudden onset and to fluctuate over time. Thinking tends to be muddled and speech may show considerable perseveration. The illusions and hallucinations associated with delirium tend to include a strong visual component, although auditory hallucinations and misperceptions are common. Delusions are usually simple, persecutory in nature, fluctuating, and transient.

It should be noted that if delirium and cognitive impairment are simply assessed by orientation in time, place, and person, then 'mild' or 'early' delirium may be missed, and it is therefore important to use additional tests of concentration and memory. All patients should be screened with a small battery of bedside cognitive tests that include specific tests of concentration such as serial subtractions and an assessment of memory for recent events and new information. The Mini-Mental State Examination (**MMSE**), supplemented with a few additional memory tests, is often used.

Psychiatric disorder

The characteristic clinical features of psychiatric disorders are covered in Chapter 26.5.6. Here we will discuss the features of acute behavioural disturbance that are suggestive of a psychotic illness or other psychiatric disorder. This issue is especially important in the emergency medical setting where such patients may be perceived as 'time-wasting' and 'not medical', and their medical needs may be crucially neglected.

A past history of psychiatric contact or treatment should be sought in all those with behavioural disturbance, as this is an indicator of putative psychiatric causation. In those with an underlying psychiatric disorder there is usually a background of insidious behavioural disturbance or personality change, and this will often become apparent from any informant. Delusions in psychotic disorders tend to be complex, bizarre, and consistently held, but this may not be so in early cases. Visual hallucinations are rare in psychosis. Marked attentional and memory deficits are not typical of psychosis, although more subtle attentional problems and a range of other cognitive deficits may be present. Distractibility as a consequence of internal experiences may give the impression of confusion and attentional impairment, although careful cognitive assessment will usually indicate preserved function.

Delirium in those with psychiatric disorder

The diagnosis of delirium is particularly difficult in those with a history of severe psychiatric disorder and/or learning disability. Difficulty in communicating with and examining such patients, who may have baseline cognitive impairment, means that delirium is particularly likely to be overlooked. Patients with severe mental illness will often attend for emergency consultations where the initial impression is of deterioration in their mental state, often coupled with a recent history of failing to comply with prescribed treatment or a disengagement from services provided. It should always be remembered that there is a high rate of undiagnosed physical illnesses in this population, and their risk of delirium is also raised because of serious side-effects from psychotropic medication, including neuroleptic malignant syndrome and lithium toxicity that can result in a deteriorating mental state. In addition, other aspects of these patients' behaviour place them at risk of physical illness, such as coexisting substance and alcohol dependency.

Investigation

It is necessary to exclude the wide range of medical conditions (Table 2) that may lead to delirium. A history of alcohol and/or illicit substance misuse may offer important indicators of aetiology. Although not always easy, a thorough physical examination with particular attention to neurological examination is essential in the assessment of all patients with acute disturbance. In addition, a routine screen—including blood count, electrolytes, liver and thyroid function, and C-reactive protein (**CRP**)/erythrocyte sedimentation rate (**ESR**)—is required, as this might indicate delirium where the diagnosis is in doubt. Infection is implicated in around one-third of hospital inpatients who are delirious, and a mid-stream urine sample (**MSU**) and chest radiograph are usually warranted in addition to routine blood testing in these patients. Relevant history and findings on physical examination usually guide more specific investigation. Encephalitis and intracerebral haemorrhage sometimes present with acute disturbance and cognitive impairment with no additional abnormalities in the history and clinical examination. An urgent computed tomography (**CT**) head scan or magnetic resonance imaging (**MRI**) is indicated where the immediate cause of acute cognitive impairment is not apparent or there are focal neurological signs. Appropriate tests for infectious diseases such as malaria, trypanosomiasis, typhoid fever, and typhus will also need to be considered when there is a history of foreign travel. An electroencephalogram with evidence of progressive cortical slowing may suggest a delirium and the need for a more extensive investigation where the diagnosis is in doubt.

Management

The management of delirium essentially consists of treating the underlying cause. Containment of any behavioural disturbance should involve general measures in the first instance, rather than psychotropic drug treatment, although sedation is necessary in some cases. Careful and repeated explanation of the diagnosis, investigations, and treatment to the patient and relatives is important. The patient should be nursed in a bright, simple room with minimal changes in staff and good lighting at night to reduce perceptual disturbance. Drugs, especially psychoactive and anticholinergic agents that may exacerbate confusion, should be reduced to a minimum. Where sedation is required then a regular oral antipsychotic such as haloperidol or chlorpromazine can be administered, although the clinician should be alert to the powerful antidopaminergic side-effects of these drugs. (See Chapter 26.4 for further discussion of these issues.)

Behavioural disturbance in patients with alcohol and substance misuse

This group of patients often present considerable demands on clinicians due to the wide range of associated physical morbidity that follows from substance and alcohol misuse. Approximately one-quarter of all male medical admissions have been found to have a current or previous alcohol problem. Such individuals commonly attend accident and emergency departments in a state of withdrawal or intoxication that engender negative attitudes from clinical staff. Often there is an expectation that the behavioural disturbance is due to intoxication or a withdrawal syndrome, without adequate assessment of any other physical pathology. Alternatively, such patients may attempt to minimize their alcohol and drug history so that the contribution of these to their complaints may not be immediately apparent.

Patients with a history of excessive alcohol consumption are vulnerable to a large number of complications that may precipitate a delirium (see Table 3) and care is needed to assess all of these possibilities. The onset of hallucinations may be mistakenly labelled as a consequence of delirium tremens without consideration of

other 'organic' or 'functional' disorders.

Delirium tremens carries a mortality risk of approximately 5 per cent and, furthermore, there is the danger that a withdrawal or intoxication syndrome may mask the emergence of other complications of alcohol and substance misuse. A history of recent blackouts or seizures should alert the physician to the possibility of hypoglycaemia or epilepsy. A careful assessment of the mental state is needed to differentiate 'functional' disorders, such as alcoholic hallucinosis, from schizophrenia as treatment of a mental disorder may be overlooked. Physical examination should include a careful assessment for signs of cirrhosis or acute hepatic encephalopathy. Investigation is essential to distinguish underlying conditions such as hepatic encephalopathy and hypoglycaemia from delirium tremens and should include full blood count, liver function tests, electrolyte and g-glutamyl transferase (**GGT**) measurements, glucose estimation, chest radiographs, and a CT or MRI scan of the brain where there is a suspicion of head trauma contributing to the disturbance. In particular, Wernicke's encephalopathy should be considered since it may be seen in up to 3 per cent of all admissions for alcohol complications. (See Chapter 26.7.1, Chapter 26.7.2, and Chapter 26.7.3 for further discussion of these issues.)

Chronic and subacute cognitive and behavioural disturbance

Assessment

In the assessment of patients with a more insidious onset of cognitive or psychiatric disturbance there can again be uncertainty as to the relative aetiological roles of organic or psychiatric factors. This may lead to unnecessary investigations at both considerable expense and discomfort to the patient, with attention diverted from appropriate management. Moreover, failure to consider a treatable cerebral disorder such as a space-occupying lesion may lead to avoidable and irreversible brain damage. Table 4 outlines a list of cognitive and psychiatric disorders that may exhibit evidence of cognitive impairment. The diagnostic uncertainty pertaining to the assessment of chronic cognitive impairment is highlighted by the somewhat misleading term 'pseudodementia' used to denote a psychological aetiology.

The diagnostic challenges in this group of patients are exemplified by the complex differentiation between dementia and depression or 'depressive pseudodementia', where there are changes in behaviour, mood, intellectual functioning, and cognitive performance. Differentiation is complicated by the fact that depressed mood is a frequent prodrome (and possibly even a risk factor) for an emerging dementia such as Alzheimer's. Furthermore, depression is a common complication or consequence of Alzheimer's and other dementias. It is therefore essential in the clinical setting that the relative contributions of psychiatric and pathological factors in any given case are considered, and that assessment includes a thorough physical, neurological, and psychiatric examination.

Dementia

Dementia is a syndrome involving a pervasive impairment of higher cortical functions and resulting from widespread brain pathology. The aetiology and characteristic clinical features of dementia are described in detail in Chapter 24.13.8 and Chapter 30.2, summarized in Table 5, and the most important causes are shown in Table 6. In the investigation of dementia it is essential to identify or exclude reversible causes: this should therefore include a complete blood count, electrolyte and metabolic screen, thyroid screen, vitamin B_{12} and folate levels, syphilis serology, urinalysis, chest radiography, and electrocardiography, and head CT or MRI scan. These investigations are sufficient to diagnose most treatable dementias. Any uncertainty about the extent of cognitive impairment necessitates formal neuropsychological assessment using instruments such as the Weschler Adult Intelligence Scale (**WAIS**-R or WAIS-111), as well as standard memory and executive tests.

The presence of a family history of early-onset cognitive impairment may indicate the need for genetic screening for Alzheimer's and Huntington's disease after appropriate counselling. Other potential evaluations include MRI, electroencephalography, cerebrospinal fluid examination, and possibly cerebral blood flow and metabolism measures (single-photon emission computed tomography (**SPECT**) and positron-emission tomography (**PET**)). Brain biopsy can be of additional assistance in diagnosing the cause of the dementia when justified by the clinical setting. For example, any suggestion of Creutzfeldt–Jakob disease would make an electroencephalogram (**EEG**) essential, and focal neurological signs would indicate the need for neuroimaging to exclude a space-occupying lesion. A known history of HIV infection would warrant a lumbar puncture. Evidence of extrapyramidal disturbance should alert the clinician to the possibility of Wilson's disease, necessitating serum copper and caeruloplasmin level investigation and slit-lamp examination for Kayser–Fleischer rings. These issues are discussed more completely in Chapter 24.13.8.

New cases of psychotic disorder

Missing an underlying 'organic' diagnosis remains a continuing concern of clinicians responsible for the assessment and treatment of new cases of an apparent psychosis. Clinical experience and numerous case reports attest to the wide range of disorders that may emerge following the initial diagnosis of a 'functional' psychosis. One follow-up study of a sample of patients with first-episode schizophrenia found that 15 out of 268 cases studied had 'organic' disorders that appeared relevant to the mental state: 13 patients out of these 15 had salient features in the medical history or neurological signs that could have alerted the clinician to the underlying disorder, the two exceptions both having a diagnosis of neurosyphilis. HIV is increasingly prevalent in this population and one recent cohort identified a known diagnosis of HIV in 4 per cent, with many subjects not tested. An assessment of risk factors for HIV is therefore required in all new cases of psychotic disorder. Overall, the literature suggests that the risks of missing organic illness are low, provided that a thorough clinical assessment is performed.

Some debate remains over the degree of investigation appropriate for the onset of psychosis. Certainly patients with cognitive impairment, abnormal neurological signs, atypical illnesses not responding to treatment, or other indications from the history, warrant further investigation. Where appropriate, this should include neuroimaging, electroencephalography, syphilis serology, and other investigations indicated by the clinical picture. Increasingly, neuroimaging provides important information relevant to the management of a particular case, although in the absence of specific indications the identification of treatable neurological disease is low.

Focal cognitive disorders

A variety of neuropsychiatric syndromes may arise from regional cerebral impairments of diverse cause in the absence of generalized cognitive impairment.

Clinical features

Amnesic disorders

The essential clinical feature of an amnesic disorder is a profound impairment in new learning relative to any generalized cognitive impairment. The impairment in recent memory is usually associated with disorientation in time and the patient is unable to retain information for more than a few seconds. A common cause of the amnesic syndrome is the Wernicke–Korsakoff syndrome resulting from thiamine deficiency in association with chronic alcoholism or, occasionally, malnutrition or malabsorption. The Wernicke phase of this disorder is characterized by confusion, nystagmus, abducent and conjugate gaze palsies (ophthalmoplegia), and ataxia. These features are commonly accompanied by peripheral neuropathy. Prompt treatment with thiamine replacement is vital in order to avert a chronic and disabling amnesic disorder (the Korsakoff syndrome).

The neuropathology of amnesic disorders usually involves lesions within the limbic system, including the thalamus and posterior hypothalamus, medial temporal lobes, and mamillary bodies. The crucial pathology in Korsakoff's syndrome is thought to involve neuronal loss, gliosis, and microhaemorrhages that produce disruption of mammilothalamic circuits. Pathology elsewhere in the paraventricular and periaqueductal grey matter, the frontal lobes, and in white matter pathways traversing the diencephalons, are common accompanying features. Amnesia can also be seen in herpes simplex encephalitis, carbon monoxide poisoning, other causes of cerebral anoxia, thalamic infarction, subarachnoid haemorrhage, head injury, deep midline space-occupying lesions, or tuberculous meningitis.

Frontal lobe syndromes

Particular neuropsychiatric interest is attached to the consequences of damage to the anterior regions of the brain. These are frequently neurologically 'silent', but they can also result in remarkable alterations in behaviour and personality, with preservation of cognitive functions such as memory and intelligence. Thus, psychiatric manifestations may be the only signs of frontal brain disease, and psychiatric disturbance may be an impediment to medical management. Two clinical pictures that frequently coexist are recognized. The first is characterized by a loss of initiative, indifference, lack of motivation, with impoverished speech and communication to the extreme of mutism. The second is characterized by disinhibition, impulsivity (occasionally with aggression), lack of ability to sustain attention and concentration, and loss of sensitivity to social cues: in general, such patients are excessively talkative and they may confabulate spontaneously. Both such syndromes have been

subsumed under the term 'dysexecutive syndrome' (also known as 'strategy application disorder'), which attempts a unitary cognitive psychological perspective on the condition. Interesting parallels have been drawn between the clinical frontal lobe syndrome and the features of neurological conditions such as Parkinson's disease and psychiatric disorders such as the negative syndrome in schizophrenia: in both these examples it is thought that impaired dopaminergic neurotransmission in prefrontal brain regions gives rise to the particular symptomatology.

Temporal lobe syndromes

A variety of syndromes, depending upon the particular area affected, are recognized following temporal lobe damage. Personality disturbance may be seen, although usually with neurological impairments. A particular variant of this is the Kluver–Bucy syndrome following bilateral lesions to the medial and lateral temporal lobes: this results in irresistible impulses to touch objects and place them in the mouth, combined with a lack of initiative, placidity, and visual agnosia. Dominant lobe lesions may produce aphasia, 'surface' dyslexia, and/or dysgraphia, frequently accompanied by neurological impairments on the contralateral side. Non-dominant lesions may particularly affect facial, spatial, or autobiographical memory, or may appear to be cognitively 'silent'. There is a recognized association between temporal lobe lesions, particularly those giving rise to epileptic activity, and psychosis, which may bear striking similarities with that seen in schizophrenia or affective disorder. Last, severe bilateral medial temporal lobe damage usually gives rise to a profound amnesic syndrome, with an almost complete loss of the ability to learn new material (anterograde amnesia) and a variable degree of retrograde loss of memory, whilst pathology in the left inferior temporal gyrus can produce severe deficits in semantic memory.

Parietal lobe syndromes

Parietal lobe lesions are associated with two particular sets of higher cognitive impairments: first, loss of visuospatial abilities, resulting in apraxias and spatial disorientation including left–right disorientation; second, loss of higher sensory perception, resulting in astereognosis and body image disturbance, which may be of such severity that there is a denial of disability, anosognosia, and visuospatial neglect. There may also be involvement of the occipital lobe, which can produce additional homonymous field defects, occasionally associated with visual hallucinations.

Diencephalic syndromes

Lesions to the deep midline structures of the thalamus, hypothalamus, and brainstem are associated with an amnesic syndrome, particularly exemplified by the Wernicke–Korsakoff syndrome. Deep midline tumours can produce a similar picture. More posterior brainstem lesions are associated with hypersomnia and placidity, which may be insidious: this was memorably described in a patient with 'akinetic mutism' due to a juxtapituitary meningioma.

Investigation of focal cognitive disorders

Cognitive assessment

Neuropsychological tests require the patient's co-operation and may fail to discriminate reliably between a cognitive and psychiatric disorder. However, they may furnish important indications of localized cerebral dysfunction and assist the clinician in monitoring the progress of any cognitive impairment. A wide range of tests is available to evaluate the pattern of disability, and these may also contribute important information to assist rehabilitation.

Brain imaging

The most useful range of investigations in suspected neuropsychiatric disorders comprise brain imaging techniques, the technology and application of which has expanded greatly in recent years. The most widely available, computed tomography (CT), is able to visualize most cerebral lesions, but should nowadays be reserved for the investigation of acute progressive cerebral damage such as stroke or subarachnoid haemorrhage, where time is of the essence and management decisions must be made rapidly. In more insidious, less acute contexts, the brain imaging of choice is magnetic resonance imaging (MRI), which allows both a higher spatial resolution with fine anatomical detail and a choice of endogenous/exogenous tissue-contrast modalities, producing greater diagnostic yield. MRI is more sensitive than CT in identifying small vascular lesions or demyelination, but less sensitive in detecting calcified lesions. Sequential assessments of cerebral atrophy and quantitative approaches can facilitate the accurate assessment of disease progression.

The EEG is frequently employed in the investigation of neuropsychiatric disorder as it is both widely available and a sensitive, if relatively non-specific, indicator of cerebral dysfunction. Focal abnormalities are characteristic of epilepsy that may, in turn, reflect vascular change. Diffuse slowing (that is to say, a shift to lower frequency ranges) is a sensitive indicator of brain dysfunction arising from metabolic and degenerative processes that correlates with the degree of cognitive impairment, although with relatively little specificity. Characteristic EEG changes are associated with Huntington's disease (pronounced flattening of traces), sporadic Creutzfeldt–Jakob disease (repetitive and triphasic spike discharges), and in association with specific drugs. Medial temporal slowing can be suggestive of early Alzheimer dementia.

The imaging of cerebral metabolism, blood flow, and receptor density, using techniques such as SPECT and PET, provide an alternative perspective on cerebral dysfunction. These are proving increasingly valuable in neuropsychiatry, although more for research than clinical purposes.

The past decade has seen an enormous expansion in functional brain imaging that allows assessment of brain regional engagement during cognitive processing. Changes in regional cerebral blood flow using PET, or in regional haemoglobin oxygenation status using blood oxygenation level-dependent (**BOLD**) functional MRI (**fMRI**), can be measured during cognitive performance. Such dynamic assessments remain essentially research techniques, allowing measurement of regional brain function and functional connectivity between linked brain areas. However, they may well offer clinical value in diagnosis and disease stratification in the future.

Neuropsychiatric causes of psychiatric disorders

Neuropsychiatric causes of psychiatric disorder are shown in Table 7.

'Organic mood disorder'

A variety of medical conditions are associated with prominent affective disorder, including anxiety, elation, and depressive symptoms. In many of these there appears to be a direct relationship between the presence of brain disease and depression, and the latter does not just seem to reflect the disabling social consequences of chronic disease, although the 'psychological reaction' to the disablement may still be an important contributory factor. In this connection, the severity of the depression is poorly correlated with the 'objective' disability or the prognosis of the disorder, and some disorders, such as multiple sclerosis, can be associated with either euphoria or depression or mood swings between the two extremes.

'Organic personality disorder'

'Organic personality disorder' is an unhappy term employed in ICD-10 to denote acute or (more typically) insidious changes in personality, defined as a significant alteration in the habitual disposition and behaviour of a patient from their premorbid state. The syndrome is increasingly well recognized, although often in retrospect. Most prominently affected is the degree of emotional expression and levels of activity, in the absence of pronounced cognitive alterations, except where 'higher level' functions such as planning complex actions or anticipation of social and emotional consequences are affected. Causes comprise intracerebral insults and consequent damage, most commonly to the frontal, temporal, or subcortical regions, or a range of rare degenerative conditions (Table 8).

Specific conditions giving rise to acute, subacute, or insidious cognitive and behavioural change

This section attempts to address two aspects of psychiatric problems associated with medical conditions: (1) to prompt the recognition and exploration of psychiatric abnormality in 'high-risk' conditions where such associations are well recognized; and (2) to encourage appropriate medical examination and investigation in the presence of outwardly psychiatric abnormality.

Psychological or psychiatric disorder may become manifest as an adjustment reaction to medical disability, malaise, and handicap, and this can affect not only the

patient, but also family members, who often bear the practical burden of care. Psychological disorder may also result from a specific compromise of cerebral function, either directly or systemically mediated. For example, postoperative psychiatric disturbance is common and usually the result of infective, metabolic, or drug-induced delirium. However, the simple circumstances of operation may lead to the precipitation of disorientation in the presence of an insidious dementia, or a withdrawal syndrome in an alcohol-dependent individual. Furthermore, the emotional reaction in response to life-threatening and life-altering circumstances may be profound following major surgery. These factors interact, and the ultimate expression of mental disturbance depends upon a particular patient's premorbid disposition and social circumstances, as well as specific illness factors. Nevertheless, the recognition and specific ascertainment of the presence of mental disturbance in certain conditions has profound diagnostic and prognostic importance.

Neurological disorders

Cerebrovascular disorders

The psychiatric complications associated with stroke have illustrated the relevance of a neuropsychiatric perspective. Early studies recognized distinct emotional reactions associated with cerebral damage. These included the catastrophic reaction (often extreme or disproportionate emotional outburst to small demands); the indifference reaction (associated with fatuous mood, indifference to failures, and unilateral neglect and anosognosia); and pathological laughter/crying (also known rather pejoratively as emotional incontinence, where emotional displays occur seemingly spontaneously or to trivial provocation). More recent and carefully controlled studies have revealed that, contrary to expectation, the mood consequences of stroke are disproportionate to the objective disability and show a consistent relationship with the location of the lesion, with anterior lesions being strongly associated with depression.

Cerebral tumours

Cerebral tumours are frequently associated with psychiatric disability, ranging from 'understandable' reactions to the diagnosis to frank syndromes resulting from impaired brain function. Minor psychological disturbance including anxiety, depression, and occasionally hysterical symptoms may be seen before the medical diagnosis is made, and specific signs of cerebral pathology need to be excluded in this group of patients. The regional syndromes outlined above are notable in cases of primary or secondary cerebral tumours, in particular when the tumour is rapidly progressive or where multiple brain regions are involved with metastases. The most common adult-onset primary tumour is the ostensibly 'benign' meningioma, which is notoriously slow growing (estimates of growth indicate that tumours at diagnosis have often been present for some 10 to 15 years), and thus cerebral function is only slowly compromised. Coupled with their propensity for a frontal location, in which there may be few frank neurological signs, this can lead to tragic cases of progressive personality deterioration being overlooked. More dramatic impairments of cerebral function are particularly associated with rapidly progressive tumours in which raised intracranial pressure, irritative epileptic phenomena, and an overall distortion in brain structure may combine to produce delirium and dementia. There can be remote effects of malignant disease on cerebral function: hypercalcaemia may present with an acute confusional state or with other psychological/psychiatric manifestations, and some forms of lung carcinoma (in particular) secrete growth factor/endocrine hormones that result in neurodegenerative changes and a dementia-like picture. Last, both episodic and prolonged confusional states have been reported in malignant disease in the absence of a clear metabolic disturbance or focal brain involvement, for example in diffuse leptomeningeal disease.

Head injury

The most prominent group suffering head injury are young men who have sustained a motor vehicle injury. Recent neuropsychiatric interest in the condition stems from recognition that the problems suffered involve impairment in personality, affect, and social/occupational function more prominently than the objective dysfunction would suggest should be the case. The wide range of neuropsychiatric sequelae recognized after head injury are outlined in Table 9. The most disabling and distressing problems for both patients and carers are the emotional and behavioural effects and, in particular, the personality change.

Impairment of consciousness is characteristic after all but the most mild head injuries and features of delirium are often seen after severe injuries. The period of post-traumatic amnesia (**PTA**), representing the time that elapses between the moment of injury and the restoration of memory for everyday events, is an important predictor of outcome. This is correlated with personality change as well as intellectual impairment and neurological disorder. By contrast, memory loss for the events of the trauma itself appears to protect against the development of post-traumatic stress disorder.

Cognitive impairment following head injury is usually more apparent after a PTA exceeding 24 h, and testing reveals that performance and non-verbal intelligence are more vulnerable to the effects of trauma than verbal and vocabulary-based intelligence. Penetrating and localized injuries tend to result in more focal cognitive deficits dependent on the site of injury. Dysexecutive syndrome and short-term memory impairment are commonly seen.

Personality change is particularly common after severe head injury and frontal lobe damage and includes irritability, impatience, apathy, and lability of mood. There is an inability to learn from experience, with poor judgement and lack of initiative. Aggression and sexual disinhibition may necessitate high levels of subsequent care. Delusional disorders are frequently observed in the early stages of recovery and may reflect the persistence of disordered cognition; mood and anxiety disorders are also often seen. Premorbid alcohol misuse may have predisposed to the trauma and alcohol tolerance can decline markedly after severe injury. Problems with heavy drinking are not uncommon in this group, which may reflect poor insight and drinking in response to stressful circumstances. Caution is needed to exclude chronic subdural haematoma and post-traumatic epilepsy before ascribing emotional and behavioural change to a psychiatric diagnosis.

Complex rehabilitation strategies are often needed to manage this group of patients and novel pharmacological management are being assessed.

Epilepsy

In assessing epilepsy it is important to establish the extent of underlying cerebral damage giving rise to the epileptic discharge, as well as the nature and severity of any cognitive impairment. While compatible with normal intelligence, epilepsy is more common in patients with a learning disability, and is related to its severity, presumably as both epilepsy and a learning difficulty arise from underlying cerebral dysfunction. Thus the capacity of individuals to manage their epilepsy is highly variable and poses an important problem for management. Furthermore, there is a relationship between emotional state and definite epileptic seizures, indicating an interesting brain–mind relationship.

The neuropsychiatric consequences of epilepsy are best considered in terms of peri-ictal and interictal disorders, which are outlined in Table 10. Further information is available in Chapter 24.13.3.

Intracranial infections

Subacute encephalopathies

A well-recognized manifestation of acute cerebral infection is dramatic behavioural disturbance, which may involve violence and delirium—this forms a relatively common diagnostic problem, with particular value placed on a correct diagnosis. Of note are, first, that the groups relatively predisposed to the development of encephalitis are alcohol and drug misusers, where attitudes of medical staff may exert a pejorative effect; second, that the disordered behaviour may be sufficiently extreme to render medical examination and management difficult. While not rigorously studied, herpes simplex encephalitis is particularly implicated in such presentations, perhaps as the most common encephalitis and also because of tropism of the virus to frontal and temporal regions. Damage in these areas can lead to impairment of behaviour and language, and occasionally to acute psychotic features such as auditory hallucinations. Other forms of encephalitis can also produce striking behavioural change.

Neurosyphilis

In historical terms syphilitic infection is considered an archetype for neuropsychiatric disorder. It was formerly called the 'great mimic' in that a large variety of presentations were recognized, including frank neurological features, insidious cognitive deterioration, affective disturbance (particularly grandiose mania), and personality coarsening. Whilst its incidence is low, particular groups such as immigrants and those with coexistent HIV infection are at increased risk. More recently, partially treated neurosyphilis means that atypical presentations have become the norm. Although neurological features, such as the Argyll–Robertson pupil, may raise suspicions, these are rare; in practice, possible exposure needs to be elicited and routine laboratory testing carried out on a regular basis.

HIV and AIDS

Human immunodeficiency virus and the acquired immunodeficiency syndrome (**AIDS**) now form the most common infection 'cluster' associated with prominent psychiatric features. Primarily, the most common reactions are to the psychological impact of the diagnosis and its prognosis, which can result in profound depression or anxiety states and needs careful management. It is now recognized that HIV has a direct tropism for neuronal tissue and may result in a frank encephalopathy, producing progressive cognitive impairment in a so-called AIDS–dementia complex. The encephalopathy is also associated with features of mania. Gathering evidence suggests that this may be reversible using antiretroviral combination therapy.

New-variant Creutzfeldt–Jakob disease

New-variant Creutzfeldt–Jakob disease (**nvCJD**) commonly presents with neuropsychiatric problems, particularly with subtle memory disturbances (which may result from the disproportionate tropism of prions for the thalamus and diencephalon), intermittent delirium with violent outbursts, auditory hallucinations, and mutism. It is rapidly progressive, and while thalamic involvement can be recognized at an early stage on neuroimaging, the diagnosis can only be confirmed by brain biopsy with immunostaining for prion precursor protein. Less invasive tonsillar biopsy methods, which require specialist referral, have also been developed for diagnosis.

Neurodegenerative conditions

Multiple sclerosis and demyelinating disorders

A strong association with cognitive impairment (between 40 and 60 per cent) and mood disorder is recognized in multiple sclerosis. The cognitive deficits seem to subtend from disturbed 'frontal' or executive function, as does depressed mood, and both show little relationship with motor disability or with demonstrable lesion load.

Parkinson's disease and movement disorders

There are four particular aspects of neuropsychiatric disturbance relevant to the consideration of idiopathic and atypical parkinsonian syndromes, including multisystem atrophy:

1. A prodromal period without frank movement disorder is well recognized and may be typified by depression, personality change, and sensory changes, reflecting the onset of nigrostriatal degeneration. This must proceed to about 80 per cent loss before the development of motor signs relatively late in the disease process.
2. Dopaminergic replacement therapy, whether with cholinergic antagonists, dopamine agonists, or L-levodopa (**l-DOPA**), is associated with a substantial incidence of psychosis and delirium.
3. There is a strong association between Parkinson's disease and depression, with estimates of prevalence in community samples of about 7.5 per cent for major depressive disorder and 45 per cent for mild depressive symptoms. This association is widely divergent from objective disability, and probably reflects a prominent dopaminergic influence on mood and motivation.
4. A substantial proportion of sufferers develop 'diffuse' (extranigrostriatal) Lewy-body disease leading to focal and general cognitive deficits and dementia.

Particular mention should be made here of iatrogenic movement disorder attendant upon antipsychotic medication. All currently available antipsychotic agents have dopamine D2-antagonist activity and provoke parkinsonism, dyskinesias, and dystonia, although newer agents, in particular clozapine, have a lower propensity for this effect. Tardive dyskinesias and provoked parkinsonian syndromes do not necessarily remit on cessation of the antipsychotic agent.

Mental retardation

Although beyond the immediate scope of this chapter, the diverse conditions that give rise to learning disorders have been the subject of intense recent interest. In particular, their strong associations with attention-deficit hyperactivity disorder (**ADHD**), fetal alcohol syndrome, and autism have underscored a probable neuropsychiatric basis. From a neurodevelopmental perspective, attention deficit in children is strongly linked to conduct disorder and may have diverse consequences in adulthood, providing an archetype for the understanding of personality disorder, traditionally not viewed in neuropsychiatric terms. (See Chapter 24.21 for further discussion of some of these issues.)

Another relevant area of interest is the concept of behavioural phenotypes, where consistent behavioural abnormalities are associated with distinct genotypic abnormalities. For example, the frequently occurring chromosomal disorders of Down's syndrome, Turner's syndrome, and Klinefelter's syndrome have distinct and consistent neuropsychiatric aspects, albeit of variable degree. In Turner's syndrome, while overall IQ is variable, distinct difficulties in visuospatial function give rise to large verbal-performance IQ deficits.

Extracerebral disorders

Endocrine disorder

Endocrine disorder has an important influence on mental function and characteristic associations with neuropsychiatric disorder are found. The use of 'routine' blood tests of endocrine function has resulted in the earlier detection of problems, such that florid states are rarely seen. The focus of clinical interest in this area is upon 'preclinical' endocrine dysfunction, the complications arising from the disorder, and their treatment.

Diabetes mellitus

The neuropsychiatric aspects of diabetes mellitus, both insulin-dependent (**IDDM**) and non-insulin dependent (**NIDDM**) forms, may be considered in four areas:

1. The syndrome itself and the constraints of optimal management have a considerable impact on the lives of sufferers. The disorder is stigmatizing, especially in the young, and has significant associated morbidity and increased mortality, both factors that confer the risk of impaired psychological development, personality difficulties, and affective disorder. There is clear evidence of a relationship between 'stress', emotional disturbance, and impaired glycaemic control, although whether this is directly or indirectly (through neglect of diet and treatment) mediated is unclear.
2. Hypoglycaemic episodes are associated with frank behavioural disturbance and automatisms, as described below.
3. A series of studies have identified significant intellectual impairments in a subset of sufferers of diabetes, an attribute particularly associated with an earlier age of onset, and the frequency and severity of hypoglycaemic episodes. While this strongly suggests that episodic hypoglycaemia results in brain damage in this subgroup, the so-called 'hypoglycaemic encephalopathy', an alternative possibility is that these patients are less able to manage their diabetes. This is of particular importance for those suffering 'brittle' diabetes, whose dietary and insulin requirements prove unpredictable, and who manage only poor glycaemic control; also for those suffering hypoglycaemia-unawareness, where incipient hypoglycaemia fails to trigger counter-regulatory neuroendocrine responses, thereby resulting in unpredictable hypoglycaemic episodes.
4. Later complications of diabetes, particularly if the diabetes is suboptimally controlled, include cerebral atherosclerosis and an increased risk of stroke, focal neuropsychological syndromes, and multi-infarct dementia.

Thyroid disorder

It is rare for classical presentations of thyroid disease to be missed in clinical practice, but interesting to note that surveys reveal that more than 5 per cent of attendees at psychiatric consultations have an abnormality in thyroid function. While not all of this group would benefit from treatment for thyroid disease, it underscores the insidious nature of thyroid dysfunction, and the rule should be to exclude this whenever suspicions are aroused. In particular, hypothyroidism may commonly manifest as apathy, depression, memory impairment, and dementia, with prominent cognitive slowing. If recognized and treated early, the condition may be reversible, although in long-lasting cases of hypothyroidism it is rarely so. Neonatal hypothyroidism or 'cretinism' is now rare as a result of routine screening, but again this indicates the importance of adequate thyroid function for cerebral development. Occasionally, adult hypothyroidism may be associated with psychosis (most commonly a delirium with prominent agitation), famously termed 'myxoedema madness'.

Hyperthyroidism is almost ubiquitously associated with a subjective feeling of tension, irritability, and high arousal. Initially this may be confused with anxiety, but in more severe cases behaviour can be frankly disturbed, although the individual concerned generally retains insight. Hypomania has been reported, as has a paradoxical apathetic state—both are rare. Thyroid 'crises' with delirium are occasionally seen following radioiodine treatment.

It should be noted that both hypo- and hyperthyroidism are common consequences of lithium treatment of bipolar affective disorder. This introduces the possibility of confusing a thyroid disorder with the recurrence of affective disorder, although it should be effectively excluded by the routine monitoring of lithium treatment.

Cushing's syndrome

The excess endogenous corticosteroid production in Cushing's syndrome appears to give rise to apathy and depression, with irritability and occasionally frank behavioural disturbance. The depression is often marked and may progress to stupor. Interestingly, exogenous steroids can give rise not only to anxiety and irritability, but also to hypomania and euphoria, which are rarely seen in Cushing's syndrome itself.

Phaeochromocytoma

The episodic release of adrenergic hormones from phaeochromocytoma classically gives rise to dizziness, tremulousness, palpitations, and the subjective feeling of intense fear, leading to confusion with anxiety and panic attacks. The diagnostic feature is the elevation of blood pressure during an attack, and if such attacks are reported, routine testing for adrenergic metabolites should be conducted. Attacks are occasionally of sufficient intensity to lead to confusion and delirium.

Pituitary disorder

Panhypopituitarism with consequent adrenal failure produces a characteristic neuropsychiatric picture with apathy, fatigue, weight loss, inability to attend and concentrate, and memory impairment. The disorder may therefore be confused with chronic fatigue, depression, or even dementia. The weight loss has been confused with anorexia nervosa, although neither appetite disturbance nor distorted body image is usually found with panhypopituitarism. Acromegaly can also result in a rather characteristic apathy and lack of concern, with occasional depression and irritability, perhaps reflecting a degree of global hypopituitrism.

There are two other particular neuropsychiatric issues that arise in consideration of a pituitary disorder. First, cerebral irradiation of pituitary tumours may give rise to collateral damage to adjacent brain structures, and there are reports of memory impairment, perhaps reflecting diencephalic and/or hippocampal damage. Second, endocrine replacement therapy following pituitary ablation may be suboptimal; indeed, replacement of sex steroids is occasionally omitted because of their propensity to release sexual drive, which may be inappropriate.

Gonadal dysfunction

Considerable attention has been devoted to the influence of the menopause and the consequent fall in oestrogen and progesterone on the development of depressive disorders in women. However, given the association with changing social role, this is a complex area. The consensus is that a lack of these hormones does play a significant part in the development of minor depressive disorders, which are significantly ameliorated by hormone replacement therapy. As a corollary, recent studies of testosterone deficiency in men have identified an association with depression, anxiety, irritability, insomnia, weakness, fatigue, diminished libido, impotence, and poor memory. Given the significant confound with ageing, this remains a controversial area.

Further reading

Amiel SA (1997). Hypoglycaemia in diabetes mellitus—protecting the brain. *Diabetologia* **40**(Suppl. 2), S62–8.

Baddeley AD (1986). *Working memory*. Clarendon Press, Oxford.

Bain BK (1998). CT scans of first-break psychotic patients in good general health. *Psychiatric Services* **49**, 234–5.

Deary IJ, Frier BM (1996). Severe hypoglycaemia and cognitive impairment in diabetes. *British Medical Journal* **313**, 767–8.

D'Ercole A, *et al*. (1991). Diagnosis of physical illness in psychiatric patients using axis III and a standardised medical history. *Hospital and Community Psychiatry* **42**, 395–400.

Folstein MF, Folstein SE, McHugh PR (1975). 'Mini mental state'. A practical method for grading the cognitive state of patients for the clinician *Journal of Psychiatric Research* **12**, 189–98.

Frackowiak RSJ, *et al*. (1997). *Human brain function*. Academic Press, London.

Guinan EM, *et al*. (1998). Cognitive effects of pituitary tumors and their treatments: two case studies and our investigation of 90 patients. *Journal of Neurology, Neurosurgery, and Psychiatry* **65**, 870–6.

Harper C (1983). The incidence of Wernicke's encephalopathy in Australia—a neuropathological study of 131 cases. *Journal of Neurology, Neurosurgery and Psychiatry* **46**, 593–8.

Hebb DO (1945). Man's frontal lobe: a critical review. *Archives of Neurology and Psychiatry* **54**, 10–24.

Hodges JR (1994). *Cognitive assessment for clinicians*. Oxford Medical Publications, Oxford.

Jacobson R, Kopelman M (1998). *Organic psychiatric disorders*. In: Wilkinson G, Stein G, eds. *College seminars in adult psychiatry*. Gaskell, London.

Johnstone EC, Macmillan F, Crow TJ (1987). The occurrence of organic disease of possible or probable aetiological significance in a population of 268 cases of first episode schizophrenia. *Psychological Medicine* **17**, 371–9.

Kopelman MD (1994). Structured psychiatric interview. *British Journal of Hospital Medicine* **52**, 93–8 and see **52**, 277–81.

Kopelman MD (1995). The Korsakoff syndrome. *British Journal of Psychiatry* **166**, 154–73.

Lewis S, Higgins N (1996). *Brain imaging in psychiatry*. Blackwell Science, Oxford.

Lishman WA (1997). *Organic psychiatry, the psychological consequences of cerebral disorder*, 3rd edn. Blackwell Scientific Publications, Oxford.

Mendez MF, Cummings JL, Benson DF (1986). Depression in epilepsy. *Archives of Neurology* **43**, 766–70.

Raskind MA (1998). The clinical interface of depression and dementia. *Journal of Clinical Psychiatry* **59**(Suppl. 10), 9–12. [Review; 22 refs]

Ron MA, Feinstein A (1992). Multiple sclerosis and the mind. *Journal of Neurology, Neurosurgery, and Psychiatry* **55**, 1–3.

Rutter M., Taylor E, Hersov L (1994). *Child and adolescent psychiatry: modern approaches*. Blackwell Science, Oxford.

Sacks O (1995). *An anthropologist on Mars*. Picador, London.

Silver JM, Yudofsky SC, Hales RE (1994). Neuropsychiatry of traumatic brain injury. American Psychiatric Press, Washington DC.

Sternbach H (1998). Age-associated testosterone decline in men: clinical issues for psychiatry. *American Journal of Psychiatry* **155**, 1310–18.

Susser E *et al.* (1997). HIV infection among young adults with psychotic disorders. *American Journal of Psychiatry* **154**, 864–6.

Tandberg E, *et al*. (1996). The occurrence of depression in Parkinson's disease. A community-based study. *Archives of Neurology* **53**, 175–9.

Taylor D, Lewis S (1993). Delirium. *Journal of Neurology, Neurosurgery, and Psychiatry* **56**, 742–51.

Toone BJ, Garralda MF, Ron MA (1982). The psychoses of epilepsy and the functional psychoses: a clinical and phenomenological comparison. *British Journal of Psychiatry* **141**, 256–61.

World Health Organization (1992). The ICD-10 classification of mental and behavioural disorders. WHO, Geneva.

Zeidler M *et al*. (1997). New variant Creutzfeldt–Jakob disease: psychiatric features. *Lancet* **350**, 908–10.

26.4 Acute behavioural emergencies

Eleanor Feldman

This chapter covers the assessment and management of patients with acute behavioural disturbance in the general hospital.

Compromised cerebral function can lead to acute behavioural problems in any unit in a general hospital, but behavioural emergencies are most frequently encountered in accident and emergency departments and on wards to which deliberate self-harm patients have been admitted. The physician needs to be able to evaluate the causes of disturbance, understand how to calm the situation before resorting to sedation, know what emergency medication may safely be used if required, and have a confident understanding of his own, and other hospital staff's, legal rights and duties under local jurisdiction. Whilst all hospitals should have specialist help available from a psychiatrist, in many circumstances a physician will be the first doctor on the scene and may need to take immediate action to prevent harm.

Evaluating the causes of disturbance

A priority is to discover whether or not the patient is severely physically ill with compromised cerebral function and suffering from delirium. 'Delirium' is a term often used interchangeably with such phrases as 'toxic confusional state', 'acute confusional state', and 'acute organic brain syndrome'. It accounts for the majority of acute behavioural disturbances arising in patients in a general hospital who have no previous history of mental or behavioural disorder. Delirium in a young adult may signal severe life-threatening illness, but most patients will be elderly, many with a degree of dementia on to which the delirium is superimposed.

Delirium is a reversible organic mental syndrome with an acute or subacute onset, typically fluctuating in severity and often worse at night. Patients will have disturbed attention and concentration and no clear memory of events once they recover. They may be somnolent and have decreased psychomotor activity, or have the opposite with agitation and aggression. Mood changes occur, as do delusions (usually fleeting) and hallucinations, with the latter being in any sensory modality, commonly visual. Clinical features present in delirium are listed in Table 1.

Where mood disorder, delusions, and hallucinations are prominent, the patient's history and cognitive function are particularly helpful in distinguishing delirium from acute functional psychoses such as schizophrenia, mania, or psychotic depression. The most sensitive indicator of generalized cognitive dysfunction is disorientation in time, which may be subtle (see Table 1). The underlying cause of the delirium must be found and treated, but in the meantime any behavioural disturbance needs to be managed. Table 2 lists the causes of delirium.

It is worth noting that a patient who appears disorientated in person but shows no sign of other cognitive impairment is not delirious, but may be in a dissociative state or possibly be presenting with a factitious disorder.

How to calm the situation before resorting to sedation

Whatever the cause of a behavioural disturbance, there are general principles in the management of all such patients.

Staff behaviour

Disorientated and psychotic patients are often in a state of nightmarish terror, whilst patients disinhibited by drugs or alcohol are less in control of their aggressive tendencies. In all cases, staff need to remain calm and polite in their dealings with patients, as anxiety and hostility on the part of staff will only serve to escalate fear and aggression in the patient. Speech should be gentle, calm, and soft spoken, but also clear, confident, and honest. The patient should be treated with normal respect and staff should not forget to introduce themselves and explain what is happening at every point. The same few staff should have contact with a disorientated patient and, for inpatients, catering staff and cleaners should be kept away. Disorientated patients need to be reminded repeatedly where they are and what is happening. Non-verbal communication should mirror this calm and gentle approach. Touching the patient without permission or getting too close may be misinterpreted as an attack. An unpredictable patient should never be seen without support staff being present in the background and within earshot. Furthermore, the patient should not be backed into a corner of the room, and staff should remain close to the door. No attempts to control and restrain the patient should be made unless staff are trained in these techniques.

Facilities in an accident and emergency department

Accident and emergency (**A&E**) departments need an interview room designed for use with behaviourally disturbed patients. The room should be situated within sight and hearing of A&E staff, not isolated in an inaccessible part of the department, nor at the end of a corridor. It should be well lit, in a good state of decoration in quiet calming colours. No furniture or fittings should be usable as weapons. In the interests of safety the room should have more than one outwardly opening door and an observation window so that the occupants can be seen from outside. There should be an easily accessible 'panic button' with connection to the staff area nearby.

Facilities on an inpatient unit

Acutely disturbed inpatients are best managed in a well-lit, single-bedded room: delirious patients are more prone to visual misperceptions in the shadows of half light. The room should be sparsely furnished with no objects that can be used as weapons. The door should have an observational glass panel. If the room is not on the ground floor, the window in the room should be made safe, with reinforced glass and a means of preventing the window from being fully opened. The room should be fitted with an appropriate alarm system.

What emergency medication may be safely used if required?

It may be necessary to sedate a patient when all other efforts to calm them and make the situation safe have failed. The reality is that there is usually much less capacity to contain behavioural disturbance in a general hospital than would be the case on a psychiatric unit: there are usually no specialist psychiatric nurses available, ward layout is not designed for patients with disturbed behaviour, and other ill patients in the vicinity may be placed at risk. If non-drug calming measures fail, early intervention is desirable to bring disturbed behaviour under control as soon as possible. None the less, the decision to restrain and sedate a patient is not to be undertaken lightly. Whilst this experience may not be recalled by someone with delirium, it is very traumatic for a fully orientated and aware person, and this includes someone with psychosis. That person could develop post-traumatic stress disorder, and the experience may seriously compromise their future cooperation with the required medical and psychiatric treatment. The intervention must be carried out with kindness and as gently as possible. The general principles whereby medication can be used as safely as possible are summarized in Table 3.

Recommended drugs

Major tranquillizers

There are three major tranquillizers in common usage in emergencies: chlorpromazine, haloperidol, and droperidol. These are generally safe and effective for rapid tranquillization. All lower the seizure threshold and should be avoided in patients at risk of seizures, including those in alcohol withdrawal. They are also to be avoided in patients with pre-existing parkinsonism and they have proved dangerous in dementia with Lewy bodies. Hypotension is the most common of the potentially serious side-effects and is most frequent with chlorpromazine. Extrapyramidal side-effects are also relatively frequent, making the use of an antiparkinsonian drug such as procyclidine advisable prophylactically, especially in patients with organic brain syndromes. Acute dystonic reactions may be confused with the severe neck stiffness of meningitis or a spastic posture, thereby adding to diagnostic difficulty in organic brain syndromes of unknown cause. For a patient in spinal traction a dystonic reaction would be very dangerous. The most hazardous complication overall is cardiorespiratory arrest: the true incidence of this is unknown, but it is less of a risk with haloperidol than with chlorpromazine.

Haloperidol has been widely studied with regard to its rapidity of action: intramuscular injection brings about a quicker improvement than oral administration, with significant improvement within 30 min at minimum, although 1 to 2 h is more usual. The usual intramuscular dose is 5 to 10 mg, repeated every 60 min, to a maximum dose over 24 h of around 18 mg. Doses in the elderly should be much lower: Jacoby recommends a single small dose of no more than 2 mg oral haloperidol, with effect assessed after an hour, and in general no more than 6 to 9 mg given orally over 24 h. The patient should be assessed between doses, rather than given a regular regimen, since individual response is unpredictable and doses need careful titration to avoid oversedation and severe extrapyramidal effects.

Chlorpromazine is the least rapidly effective of the two commonly used major tranquillizers and carries the greatest risk of cardiovascular side-effects; it need not be used where haloperidol is available. If it is all that is available, then a dose of 50 to 100 mg chlorpromazine would be appropriate for a first dose to assess the patient's response. The oral route should always be offered first, but where this fails, the intramuscular route is generally preferred over the intravenous on grounds of safety and ease of access. A reasonable general assumption is about a 2:1 oral:parenteral equivalent dose.

In summary, haloperidol is the most rapidly effective major tranquillizer by the safe and convenient intramuscular route. Small doses of haloperidol should be preferred in the elderly.

Minor tranquillizers

The benzodiazepines are sedative drugs with low toxicity. The principal adverse effect is respiratory depression, and prolonged use results in tolerance and dependence. They raise the seizure threshold and may be used as anticonvulsants, so are helpful in cases where there is a risk of seizures and in other conditions where major tranquillizers are contraindicated. Flumazenil allows rapid reversal of respiratory depression. Its short half-life (1 h) means repeated administration may be necessary. There have been concerns about paradoxical disinhibition and release of aggression with benzodiazepines, but these have been overstated in the past and at less than 1 per cent are no greater than placebo.

In general, intramuscular lorazepam appears as effective as a sedative as intramuscular haloperidol, even in mania, and has fewer adverse effects. In a small study in patients with mania the peak reduction in agitation occurred 60 to 120 min after oral administration, but 45 to 75 min after intramuscular and 5 to 10 min after intravenous injection. It appeared more effective than haloperidol during the first 2 h if patients were already receiving antipsychotic drugs: 10 patients who received lorazepam on one occasion and haloperidol on another spent less time in seclusion after lorazepam medication. Doses reported in studies have ranged from 2 to 10 mg every 1 to 2 h. The maximum single dose reported is 40 mg and maximum daily doses from about 20 to 40 mg. No serious adverse effects have been reported with lorazepam use over 1 to 2 weeks. Ataxia has occurred above 10 mg/day, with nausea and confusion at the highest doses. When given by intramuscular injection lorazepam should be diluted with an equal volume of water or saline for injection.

Diazepam is poorly and erratically absorbed after intramuscular injection, making it less suitable for emergency sedation than lorazepam. It has a long half-life and accumulation is likely with repeated doses. Intravenous diazepam is effective in calming behavioural disturbance within 15 min, the diazemuls preparation causing less venous inflammation.

It is often useful to combine an antipsychotic with a benzodiazepine for the most effective safe sedation. The most studied combination has been parenteral haloperidol with lorazepam. It has been claimed that this reduces the total dose of antipsychotic required, but there are case reports where an intravenous combination of these drugs has caused cardiorespiratory arrest. In an open trial the combination of haloperidol and lorazepam given intramuscularly was effective more rapidly than either drug alone, occurring within 30 min in most patients, compared to nearing 60 min for most receiving a single drug.

The management of different syndromes

Psychosis

If a non-organic psychotic illness such as schizophrenia, mania, or psychotic depression is suspected, a psychiatrist should be contacted as soon as possible. Where disturbance is extreme and the patient represents a risk to themselves or others the recommended drug is haloperidol 5 to 10 mg orally or intramuscularly. Further doses can be given hourly according to response. Lorazepam 2 to 4 mg orally or intramuscularly can be added to treat patients who are extremely disturbed. The benzodiazepine antagonist, flumazenil, should be available in case of respiratory depression. Evidence from the notes of known psychiatric patients may suggest that higher doses may be required. The elderly should be treated with half the normal adult doses. Acute dystonic reactions to major tranquillizers respond to procyclidine 5 mg intramuscularly or orally.

Alcohol and drug states

Patients with alcohol withdrawal and disturbed behaviour should be treated acutely with diazepam 10 mg orally or lorazepam 2 mg intramuscularly. They should then be placed on an alcohol withdrawal regime including thiamine to prevent Wernicke's encephalopathy.

Patients suffering from acute drug or alcohol intoxication or drug withdrawal (but **not** alcohol withdrawal) should be treated with haloperidol 5 mg orally or intramuscularly, and continuing disturbance should be treated as for psychosis.

Care following tranquillization

Patients should not be left unattended in the hours following rapid tranquillization. Observations should be recorded every 15 min for 1 hour on a form detailing the following information:

1. Conscious level:
   a. awake and active,
   b. awake and calm,
   c. asleep but rousable,
   d. asleep and unrousable.
2. Respiratory rate, blood pressure, and oxygen saturation in patients in conscious levels 1(c) and 1(d)—arterial gases should be measured if oxygen saturation is less than 92 per cent.
3. Blood pressure should always be measured if antipsychotic drugs have been given.
4. Reassessment at 1 h to look for evidence of dystonia.

When the acute situation has been calmed, a decision should be made as to whether parenteral or oral medication should be used to keep things under control, also regarding the need for specialist advice. There should be further consideration of the overall treatment plan and levels of nursing and medical observation. There

should be a daily reassessment of mental state, and specialist advice should be sought if the patient remains disturbed after 3 days.

Drug interactions

Pharmaceutical formularies contain further information on drug interactions and these should be consulted for patients taking other drugs including alcohol, antiepileptic drugs, levodopa, and lithium.

Legal rights and duties under local jurisdiction

Unlike clinical matters, legal issues are limited by state and national boundaries. Most states and countries will have statute laws covering the treatment of mental disorder in situations of non-consent, but these laws may not allow treatment for coexisting medical disorders where the treatment of the latter is not a recognized treatment of the mental disorder. Clinicians require an understanding of local statute law and common law covering circumstances where patients have a diminished capacity to give meaningful consent to medical intervention in their best interests. This aspect of clinicians' training has often been neglected and the law in this area may be confusing for clinicians who must apply it in an emergency situation. Whilst psychiatrists may have a good understanding of statute law in relation to the treatment of mental disorder, they may not be so conversant with the legal issues surrounding non-consent for the treatment of physical illness. Hospital managers can assist their staff by drawing up guidance in conjunction with their legal advisors, the professions concerned, and any standing body or commission involved in the monitoring and regulation of statutory powers relevant to these circumstances. The author has discussed the legal situation in England and Wales in detail elsewhere (Feldman, in press).

Further reading

Anonymous (1991). Management of behavioural emergencies. *Drugs and Therapeutics Bulletin* **29**, 62–4.

Feldman EJ (2000). The use of the Mental Health Act and common law in mentally disordered general hospital patients. In: Ledingham JGG, Warrell D, eds. *Concise Oxford textbook of medicine*, pp.1443–6. Oxford University Press, Oxford.

Friedman T (2000). Medical management of acute behavioural disturbance in the general hospital. In: Peveler R, Feldman E, Friedman T, eds. *Liaison psychiatry: planning services for specialist settings*, pp.51–60. Gaskell, London.

Hodges JR (1994). *Cognitive assessment for clinicians*. Oxford Medical Publications, Oxford.

Jacoby R (1998). Drugs causing confusion and drugs to treat confusion. *Prescribers' Journal* **38**, 242–8.

Storer D (2000). Liaison psychiatry services in the accident and emergency department. In: Peveler R, Feldman E, Friedman T, eds. *Liaison psychiatry: planning services for specialist settings*, pp.14–26. Gaskell, London.

26.5.1 Grief, stress, and post-traumatic stress disorder

Jenny Yiend and Tim Dalgleish

Grief

Introduction

- Grief could be described as a natural response to an objectively significant loss, most commonly the death of a loved one, although whether the boundaries defining a 'significant loss' should be stretched further (and if so, how far) remains controversial. It involves primarily psychological reactions, although it may also lead to social and physical responses. 'Normal grief', which requires no clinical intervention, can be distinguished from 'pathological grief', also variously called 'atypical, traumatic, neurotic, morbid, complicated, or unresolved grief'. However, it should be stated clearly at the outset that the concept of an abnormal form of grief is not represented in current official diagnostic manuals, and as such is not an established clinical condition. Indeed, the lack of consensus over terminology in the literature illustrates the urgent need for quality research to establish clear and universally accepted diagnostic criteria. Despite this confusion, many professionals now recognize some form of abnormal grief response as an appropriate target for active intervention, the key features being: (1) an excessive intensity of the grief reaction (inappropriate within the culture); and (2) of an unusually prolonged duration.

Epidemiology

Considering the ubiquity of bereavement, the consequences of grief are of global importance. Incidence rates for pathological grief are not available; estimates of prevalence vary from 14 to 34 per cent, based primarily on samples drawn from the United States and Europe. In geographical areas prone to natural disasters and places engaged in active military conflict these rates will obviously rise significantly in line with death rates, and in such cases particular attention should be paid to the concurrent trauma that will have accompanied the loss (see under Stress, below).

Mortality

Much evidence shows that the bereaved in general (irrespective of whether a pathological response develops) are at greater risk of dying themselves than would be expected given mortality rates in the population at large. Mortality is elevated by a factor of two to three, applying not only to spousal loss, but also to parental, sibling, and child loss. These findings have proven robust across cultures and generations. The point of highest risk is the weeks and months immediately following the loss, although the data suggest it remains elevated for several years. This evidence should alert the clinician to pay particular attention to all forms of presenting grief, whether normal or pathological.

Within a bereaved population additional factors moderate this mortality risk, although it remains elevated for all subgroups. Thus, the younger bereaved are at a higher relative risk of death themselves than the older, as are widowers compared to widows (although remarriage selectively reduces the risk in widowers). The cause of bereavement is also important, mortality risk being particularly elevated for bereavements involving suicide, accidents, liver cirrhosis, and heart disease.

Hard evidence concerning the underlying reason for reduced longevity following bereavement is scarce, but factors both directly and indirectly related to the loss may be involved. For example, psychological consequences such as the loss of the will to live might directly lead to increased suicides and carelessness. Indirectly, the change in lifestyle necessitated by bereavement may lead to the adoption of health-impairing behaviours (neglect of diet, exercise, general well being) or may create psychological stress, which in turn could have serious negative consequences for health (see ' Stress' below).

Psychiatric comorbidity

Acute bereavement is associated with an increased risk for a range of psychiatric disorders, including major depression, panic disorder, generalized anxiety disorder (**GAD**), and post-traumatic stress disorder (**PTSD**). Any of these may occur comorbidly with pathological grief, when between 17 and 31 per cent will also meet criteria for major depression, 13 per cent panic disorder, 39 per cent GAD, and 9 per cent PTSD. However, these figures should be treated with caution since the criteria for 'pathological' and the timing and consistency of assessment can greatly affect the results. What is clear though, is that the rate of comorbidity following bereavement is significant, with estimates suggesting more than half of the bereaved suffer from two disorders. Hence, having diagnosed one disorder in the bereaved patient (be it psychiatric or pathological grief), the clinician should be particularly alert to the possible presence of additional, complicating disorders. It is unclear whether this comorbidity is best conceived as the presence of one disorder predisposing the patient to additional pathology, or simply the presence of two coexisting disorders whose symptoms may or may not aggravate each other. Whichever, it is essential that both domains of symptoms are separately monitored and, to the extent that it differs, treatment for both disorders is given.

Clinical features

Normal grief

The literature on normal grief reveals that we have yet to achieve a precise characterization of the process and a clear demarcation of its boundaries. Many theorists propose that there are distinct 'stages' of grief, and while opinions vary about the precise number and nature of stages, it is common to consider at least three. These are:

- an initial period of shock, including emotional numbing and disbelief. This stage may last from hours to weeks.
- a subsequent phase of acute mourning, involving an acknowledgement of the death together with intense emotional states that typically engulf the individual in periodic 'waves' of feeling. Somatic discomfort, social withdrawal, and preoccupation with thoughts of the deceased may accompany this. There may also be 'identification' with the deceased, in which the individual adopts characteristic behaviours, mannerisms or habits of the loved one, and may even experience physical symptoms associated with the cause of death ('grief facsimile symptoms'). This phase may last for several months.
- a period of restoration of normal function during which the characteristics of acute mourning are gradually replaced by feeling able to continue with life. A shift in focus occurs away from the deceased and towards the future. While memories and a sense of loss may remain, there is recognition of having grieved and a will

to move on.

More recent theories have placed less emphasis on chronological stages, preferring instead to consider particular domains or clusters of symptoms that may fluctuate in intensity throughout the period of grieving. Table 1 summarizes some of these symptoms of grief.

Pathological grief

As stated earlier, the concept of 'abnormal' grief superseding what might be construed as normal and therefore requiring medical intervention, is not currently an officially acknowledged pathology. However, many workers are calling for a set of universally accepted diagnostic criteria to be developed, and in the meantime Jacobs has proposed a preliminary set that should prove helpful to the practising clinician. These represent a consensus opinion drawn up at a recent conference of experts, and as such incorporate a variety of perspectives on grief and its manifestations. They are formulated in the American Psychiatric Association's *Diagnostic and statistical manual of mental disorders* (**DSM**) style and are reproduced in Table 2.

Several features are worth noting.

- First, diagnosis requires the actual death of a significant other, thereby excluding any other forms of loss (physical separation, loss of non-human objects: animals, body parts, material possessions). This remains controversial.
- Second, observable psychological distress in response to the death is essential, although it may be delayed. Thus a total non-response (the absence of any observable or reported signs of grief), which may be of concern to the practising clinician, would not warrant a positive diagnosis. In such cases the best approach may be close monitoring of the patient over time, together with probing for signs of intrusive thoughts or behaviours relating to the deceased, despite emotional numbing.
- Third, criterion B represents symptoms of particular severity. They fall into four broad categories: avoidance and numbing (1, 3, 4, 5); disorganized behaviour or experience (2, 6, 7, 9); identification symptoms (10); and anger (11). The recommendation is that at least four of these should be present for a positive diagnosis, in addition to the core response of distressing preoccupation listed under A.
- Fourth, criteria D and C are central to distinguishing normal from pathological grief. They embody the notion, described earlier and consistent throughout DSM-IV, that there must be significant impairment of functioning together with an abnormally long duration of symptoms. The latter is set, somewhat arbitrarily, at 2 months.

The core domains for positive diagnosis can therefore be summarized as severity, duration, and functioning.

Differential diagnosis

Bereavement and comorbid psychiatric disorder

DSM-IV criteria for the diagnosis of major depression include specific guidelines for the circumstances of bereavement. In effect, this acknowledges that depressive-like symptoms will be fairly ubiquitous following bereavement and criteria are therefore more stringent. Specifically, either a 2-month (rather than a 2-week) duration of symptoms is required, or alternatively the presence of particular symptoms such as marked functional impairment, psychotic symptoms, or suicidal ideation is necessary. Although no specific guidelines for other comorbid psychiatric disorders are given, the clinician would be well advised to apply similar principles of increased severity or extended duration before making a positive diagnosis.

Pathological grief and comorbid psychiatric disorder

Pathological grief, by the working definition given above, occurs exclusively in the circumstances of the death of a significant other. While this objective criterion helpfully restricts diagnosis, it remains necessary to distinguish between this and other possible psychiatric disorders that may follow bereavement.

- *Major depressive episode*—is distinguished by a pervasive and general depressed mood disturbance, in contrast to the episodic pangs of grief focused around the absence of the deceased. Other characteristic symptoms of pathological grief are absent (e.g. Criterion B: 1, 4, 5, 7, 8, 9, 10, 11)
- *Panic disorder and GAD*—are distinguished primarily by the absence of characteristic symptoms of pathological grief, and in the former by the presence of acute episodes of severe anxiety or panic attacks.
- *PTSD*—this is perhaps the most problematic differential diagnosis. Could pathological grief be construed as a specific example of PTSD? Only further research will resolve this question. For the present, we suggest that the following features be considered: pathological grief, in contrast to PTSD, does not require exposure to an objectively traumatic event (although this may occur in cases of violent death). In pathological grief, symptoms of avoidance and hyperarousal are less prominent than in PTSD. Symptoms of pathological grief are centred on the deceased person (pining, searching for them, sensitivity for signs of them in the environment), whereas those of PTSD centre around the traumatic event itself (re-experiencing the trauma, intrusive thoughts about the trauma, general hypervigilance). Finally, where both disorders are suspected, it is advisable to focus treatment initially on PTSD.

Treatment

Normal grief will resolve spontaneously over time. Treatment options for pathological grief fall into the categories of pharmacology, psychotherapies, cognitive/behavioural therapies, and self-help strategies. In common with other psychological disorders, maximum benefit may often be obtained by the prudent combination of drugs with psychological treatments. In practice, individual circumstances and the local availability of treatments will inevitably impose restrictions.

Pharmacology

The few studies available looking specifically at drug treatments following bereavement suggest that both tricyclic antidepressants and selective serotonin-reuptake inhibitors (**SSRIs**) may provide effective relief of symptoms. The tricyclics appear to be more confined in their effects, influencing primarily depressive symptoms, whereas the SSRIs may have a broader action, additionally counteracting symptoms reflecting trauma, such as avoidance and emotional numbing. SSRIs have the additional advantage of more tolerable side-effects, as well as being safer in overdose. Individual circumstances, side-effect profiles, and any known personal or family history of response to treatment can act as a guide in the selection of therapy.

Psychotherapies

Psychodynamic forms of psychotherapy tend to be favoured nowadays over those of a psychoanalytical persuasion. The former centre around the developing relationship between therapist and client, and focus on ongoing changes in the presenting psychological processes observed in the client. Current opinion suggests that this form of psychotherapy may yield more effective results within a shorter time frame than psychoanalytical techniques, which tend to focus more on a re-evaluation of personal history as the means to personal change.

Psychotherapies that have been used specifically to treat the bereaved include crisis intervention and brief dynamic psychotherapy. One study of crisis intervention psychotherapy, given immediately after the loss, lasted for several months and involved reviewing aspects of the lost relationship within the context of the psychodynamic relationship. A significant reduction in symptoms was noted. By contrast, in a study of a psychodynamic therapy starting several months after the loss, only a marginal symptom improvement was found.

Cognitive/behavioural therapies

Cognitive–behaviour therapy (**CBT**) is a popular form of treatment for many psychological disorders. It combines behavioural techniques, such as relaxation and exposure, with cognitive restructuring in which the patient is encouraged to identify and alter maladaptive styles of thinking that are thought to maintain ongoing psychological distress. While there is no data specifically considering the efficacy of this treatment for bereavement, it is likely to confer similar benefits to those noted elsewhere. In relation to separation anxiety disorder, which could be considered as a childhood analogue of pathological grief, a 60 per cent recovery to normal functioning has been reported, sustained over a follow-up year.

Exclusively behavioural techniques have also been used. These involve exposure to feared or avoided stimuli in order to produce habituation, and may also incorporate relaxation techniques to aid this process. Guided mourning and trauma desensitization are two such treatments. Both appear to selectively reduce somatic and avoidance symptoms, having less of an effect on depressive symptoms and preoccupations with the deceased.

Self-help

Self help groups should not be overlooked as a possible supplement to treatment, either to aid transition following successful treatment, or in a preventive capacity. Some evidence suggests that, with appropriately trained group leaders, benefits conferred may be equivalent to some of the more formal treatments discussed above.

Finally, the reader is referred to an excellent text by Jacobs, an expert in the area, which outlines one possible practical approach to treatment endorsed by the author (see p 81 therein), as well as a diagnosis/treatment algorithm (see p 76 therein).

Prognosis

Pathological grief responses may be chronic and unremitting without medical intervention, and a prolonged course of 2 years or more is likely where symptoms persist beyond the first year. In the case of normal grief, most of the acute symptoms of mourning may be expected to dissipate within several months to a year, but some level of emotional involvement may persist indefinitely. Those at higher risk of a pathological grief response may include the young, women, those who suffer multiple losses, and those who have suffered childhood loss. The risk is increased following sudden, unexpected, violent, or suicidal death. An ambivalent or insecure relationship to the deceased ('attachment disturbance') also increases risk, as do personality traits such as neuroticism, dependency, and schizoid personality. Finally, transient features displayed in an individual, such as the inability to accept an imminent death, or severe distress during a terminal illness, increase that individual's risk for pathological grief.

Prevention

Primary measures fall largely in the domain of social policy, such as gun control, safe driving practice, or healthy living styles. These can directly reduce deaths due to unnatural causes. However, the clinician may also play a role by moderating the impact of death, particularly when it is sudden or unexpected. This would include allowing ample time to be spent with a dying or indeed a deceased patient in a quiet and supportive atmosphere, which can enable associated others to more effectively assimilate their loss.

Secondary prevention might include screening bereaved populations at high risk. At the level of the individual this would involve ascertaining the risk profile of a recently bereaved patient and, where this is high, maintaining contact, monitoring progress, and providing early intervention where appropriate.

Tertiary prevention, to moderate the extent of disability, can be implemented by considering appropriate medium- to long-term treatments. Patients may well present late in the course of pathological grief, prompted only by severe functional impairment or social pressure. Although early intervention is preferable, an appropriate selection from the treatment options discussed above is still likely to confer some benefit.

Areas of controversy needing further research

Perhaps the major controversy of concern to clinicians is whether the concept of 'pathological grief' warrants a distinct diagnostic category, or whether it is best subsumed under existing pathologies such as PTSD or major depression. The high comorbidities and the question of the differential diagnosis support the latter view; factor analysis of symptom clusters, their differential response to drugs, and the distinct risk profiles support the former. However, the question of labelling becomes clinically unimportant, to the extent that the treatment for these pathologies overlaps.

Other controversies are:

- the model of distinct 'stages of grief', which some advocate more than others;
- the extent to which the absence of grief might be considered pathological;
- the duration of normal grief, which some argue is indefinite;
- the nature of the grief object, which some restrict to the death of an intimate, while others extend far more broadly to include non-death and non-human loss.

The following issues also warrant investigation:

- cultural differences in the expression and experience of grief;
- the factors responsible for the relationship between bereavement and increased health and mortality risks.

Stress and post-traumatic stress disorder

Introduction

Stress often refers to an external object, event, or situation that causes physical and psychological effects on an individual as a result of increased levels of arousal. These effects are usually experienced as unpleasant and undesirable, although there is a close correspondence with the excitement that occurs when a positive, desirable interpretation is adopted, for example during dangerous sports. The term 'stress' is perhaps more appropriately used to refer to the subjective experience of these effects, and the agent causing them is more accurately termed 'the stressor'.

Societal and lifestyle changes in developed nations, as well as media coverage, have given prominence to the role of stressors and their adverse effects, although in practice these are prevalent universally. While a moderate degree of stress can be helpful to enhance performance, chronic stress is indeed associated with negative outcomes. Stress can be a risk factor for various physical health problems, most notably coronary heart disease, infectious diseases, immune function, and cancer. In addition 'background stress' (the presence of low-level chronic stressors) is known to potentiate an individual's negative response to acute stressors, and thus can be considered a vulnerability factor for negative outcome. Within psychiatry the effect of stressors has been extensively studied. They are known to raise the probability of relapse and, more controversially, are believed to play a part in triggering the onset of some disorders. Examples include schizophrenia, where interventions developed to reduce the levels of interpersonal stress within families ('expressed emotion') have proved effective. Similarly, in major depression much research has been conducted on the role of 'life events' (for example, death of a spouse, loss of a job, going on holiday).

For the non-psychiatric patient the adverse effects of stressors can usually be addressed through lifestyle changes such as increased exercise, improved diet, relaxation techniques, reduction of working hours, and delegation of responsibilities. All these measures require an adjustment of personal priorities, which some may be unwilling to do. Where stress arises from unavoidable personal circumstances (for example, care-giving, financial or relationship problems), the role of the clinician includes referral to appropriate support services to enable the stressors to be addressed at their source.

An additional option, provided within the DSM-IV system, is a diagnosis of adjustment disorder. This may be appropriate where a discrete, identifiable stressor exists and causes either significant impairment in functioning, or distress beyond that which would normally be expected given the nature of the stress. However, this disorder (by definition) is time-limited by and closely coupled to the external stressor itself, although it may be classed as chronic where the stressor or its consequences persist indefinitely. Specifically, symptoms must commence within 3 months of the onset of the stressor and cease within 6 months of its termination. It is also of note that bereavement is specifically excluded as a qualifying stressor. Nevertheless, where these criteria are fulfilled and other psychiatric diagnoses have been excluded, the clinician may wish to offer appropriate psychological interventions as described below.

We will use the term 'trauma' where extreme stress occurs in response to an acute, intense episode brought about by a specific, objectively identifiable, external event. A trauma (defined under Clinical features, below) may be distinguished from a stressor primarily in terms of the objective intensity and severity of the experience or incident. It is now recognized that a proportion of people exposed to such a trauma go on to develop a clinical pathology, post-traumatic stress disorder (PTSD). PTSD, the subject of the rest of this section, was first introduced into the diagnostic nomenclature in 1980 with the publication of DSM-III, and it subsequently

appeared in the World Health Organization's *International classification of disease* (**ICD**) system in 1992.

Epidemiology

Traumatic events are common, estimates suggesting that most Americans will experience at least 1 trauma over a lifetime. The lifetime prevalence of PTSD in the general population is between 1 and 14 per cent according to DSM-IV, with a recent review suggesting this level is higher in women (10 to 12 per cent) than men (5 per cent). Estimates of lifetime prevalence among trauma victims vary widely according to the criteria and populations sampled, but somewhere between 3 and 58 per cent of people who experience a trauma will go on to develop PTSD at some time in their lives, although more recent reviews put the figure as high as 60 to 80 per cent. Clearly geographical factors will influence these figures, leading to significant increases in areas prone to natural disasters or human conflict. In common with most psychiatric disorders there appears to be high comorbidity in PTSD, with 80 per cent of sufferers meeting the criteria for at least one other psychiatric disorder.

Clinical features

The primary, essential feature for a positive diagnosis of PTSD is the prior experience of an objectively traumatic event. DSM-IV distinguishes two components: first, the nature of the event itself, which should involve an 'actual or perceived threat to life or physical integrity'—typical events including active combat, rape or other assault, natural disasters, and serious accidents. Witnessing such events is also included within the concept of 'experiencing'. Second, individuals should have an extreme emotional response to the event, which DSM-IV describes as intense fear, helplessness, or horror.

A pathological reaction to such a trauma is characterized by symptoms that fall into three clinically observed domains: re-experiencing, avoidance/numbing, and hyperarousal. Avoidance and numbing may be better considered separately, although DSM-IV does not do so. Typical examples of these symptoms are as follows.

- *Re-experiencing*—including nightmares, flashbacks, intrusive thoughts, and images relating to the trauma. Such symptoms have often been considered to be the hallmark of PTSD.
- *Avoidance*—typically anything that could remind the individual of or be associated with the trauma is avoided. This can include people, places, activities, and conversations.
- *Numbing*—emotional responsiveness is generally reduced. This may include an inability to experience certain feelings and feelings of detachment or other dissociative symptoms (e.g. depersonalization, dissociative amnesia, derealization).
- *Hyperarousal*—this includes insomnia, anger, irritability, hypervigilance, problems with concentration, exaggerated startle.

DSM-IV requires at least one symptom of re-experiencing, three of avoidance/numbing, and two of hyperarousal to be present for a positive diagnosis. It also currently specifies three subtypes of PTSD—acute, chronic (where symptoms have lasted under or over 3 months, respectively), and delayed onset (where 6 months or more has elapsed after the stressor before the emergence of symptoms).

Finally, DSM-IV introduced a new, related diagnostic category—acute stress disorder (ASD)—which essentially is an acute form of PTSD. The symptoms are identical, but the diagnosis can be made as early as 2 days' post-trauma, thereby encouraging earlier intervention. Persistence of symptoms beyond 1 month results in the diagnosis reverting to PTSD. For ASD three of five dissociative/numbing symptoms are required, reflecting the belief that these are predictive of longer term psychopathology. Although the diagnosis remains controversial, it does provide the clinician with a clear indication for early intervention in certain cases.

Differential diagnosis

Normal reactions to trauma

As with grief, the key features that distinguish PTSD from non-pathological reactions to trauma are intensity, duration, and functioning. The intensity of the pathological reaction is captured by the nature of the symptoms themselves, with the presence of numbing symptoms thought to be the most effective distinguisher of victims with PTSD from those without. In addition, symptoms must have been present for at least 1 month and must be causing clinically significant distress or impairment in functioning.

Other psychiatric conditions

Subsets of the features of PTSD often overlap with other psychiatric conditions, but distinguishing characteristics are usually present. First, in PTSD there is an instigating traumatic event, which is not required for any of the other anxiety disorders. Second, the symptoms of nightmares and flashbacks are specific for PTSD and do not characterize other anxiety disorders. Third, emotional numbing, which occurs in the place of the normally expected emotional reactions, is strongly and uniquely characteristic to PTSD.

PTSD, considered to be an anxiety disorder, shares several anxiety-related symptoms, particularly from the hyperarousal cluster. Hypervigilance, sleep disturbance, irritability, and concentration problems are all common to GAD. Similarly, fear and avoidance are common to the phobias. Intrusive thoughts may also occur in obsessive–compulsive disorder (**OCD**), major depression, and GAD. Conversely, PTSD sufferers may exhibit compulsive behaviours of the type associated with OCD, such as repetitive cleansing procedures, or continual checking of locks and security devices, perhaps following rape or other kinds of assault. Although rates of comorbidity are indeed high (see above), dual diagnoses should only be made where the full criteria for both disorders are met.

Treatment

Crisis intervention

This approach, also called 'psychological debriefing', aims to treat all survivors of a trauma in the hope of reducing subsequent pathology. It takes place in a single session, within days of the incident, most forms of treatment being given individually or in small groups. Typically, there are several structured phases including each individual sharing their own general perspective ('recreating the event'), their thoughts at the time, the worst aspect of the event, and their reactions to it. There is usually also a teaching element, covering common reactions to trauma and how to deal with them.

Without doubt participants subjectively feel this type of intervention to be helpful and valuable. However, the research findings on its efficacy are mixed: a recent review revealed that there have been few randomized controlled trials, but that these show little observable benefit, some even reporting a negative outcome. One current view holds that benefits are greater if the treatment is delayed for a week or so, until the initial shock subsides. Unless and until future empirical data supports their worth, clinicians should be cautious about the use of indiscriminatory immediate intervention. The treatment options discussed below are suitable for individuals who go on to develop PTSD following trauma.

Pharmacology

The main difficulty regarding drug treatment is the current dearth of clinical trials, hence what follows cannot be more than tentative advice, based primarily on clinical experience.

Antiadrenergic agents

Agents such as b-blockers are effective in the short term in reducing symptoms of hyperarousal and re-experiencing. Patients respond quickly, although tolerance is likely to develop. They are perhaps most appropriate for those whose individual prognosis is good, or where immediate symptom relief is required, before pursuing other treatment options.

Antidepressants

Most antidepressants provide at least some symptom relief, but the benefits are generally considered to be modest for classes such as tricyclics and monoamine oxidase inhibitors (**MAOIs**), and issues of side-effect profiles and overdose safety mean that SSRIs tend to be preferred. Recent data suggests that SSRIs may be

effective in reducing symptoms from all symptom domains, and therefore they are currently the preferred option for the long-term drug treatment of PTSD. Two additional points are worth noting. First, uncertainty persists about the speed of action of these drugs, with estimates for the onset of beneficial effects varying between 2 weeks and 1 month. Patients should therefore be prepared for some delay. Second, some SSRI side-effects, such as arousal and insomnia, although usually short-lived, will be particularly difficult for the patient with PTSD to tolerate.

Psychological treatments

Many different types of psychological therapy have been used to treat patients with PTSD, most appearing to impart some benefit in terms of symptom relief and improved psychological and social functioning, but longer term benefits remain unclear. Similarly, the potential for additional gains to be made by combining psychological treatments with each other or with drugs remains largely unexplored. Some of the commoner therapies are given below.

Cognitive–behavioural therapies

These are the psychological treatments of first choice for patients with PTSD because most are relatively brief and have a well-established efficacy, both from clinical experience and empirical research. Treatments are similar to those used for other anxiety disorders, such as specific phobia, but they focus specifically on trauma-related material. Therapies may differ in the particular components included, but generally they fall into one of two groups, exposure treatments or anxiety management.

Exposure treatments

Exposure treatments involve repeated exposure to trauma-related material on the basis that this will reduce undesirable responses, either through simple habituation or as a result of concurrent cognitive reprocessing. Treatments vary in the type of exposure used. Imaginal techniques involve the patient reliving (describing verbally, writing down, or role playing) the trauma within the treatment room. *In vivo* exposure involves confronting, in real-life but safe situations, places or objects that provide reminders of the trauma. Other variables include the length of exposure (brief or prolonged) and the level of arousal induced (high or low). Prolonged imaginal exposure is currently the favoured technique of many therapists for the treatment of PTSD because of its relative efficacy and time-efficiency.

Some forms of exposure treatment, for example systematic desensitization, adopt a hierarchical approach in which exposure is graded in difficulty, starting with least feared stimuli and progressing in tandem with patient improvement. Relaxation procedures may also be employed: these are known to be unnecessary for treatment efficacy, but may help to encourage patient participation in an initially unpleasant procedure.

Exposure techniques are time-efficient and easy for patients to learn. Good quality, consistent data supports their efficacy. However, as noted, compliance may be a genuine problem, particularly in those with prominent avoidance symptoms.

Anxiety management

Anxiety management training aims to teach patients to control and cope with heir symptoms, rather than focusing on elimination or cure. Stress inoculation training is one such technique that has been commonly used for PTSD. Treatment usually involves components of both education and skills training. The latter may include deep relaxation, quick relaxation, breathing control, thought stopping, and role play. Anxiety management programmes are more complex to administer and more intellectually demanding for patients than other treatment options. However, they are likely to be particularly appropriate for PTSD patients with symptoms of chronic, general arousal. In addition, they may be indicated at later stages, for example where maximal benefit has been achieved from other treatment options and the patient is left with residual symptoms.

Psychotherapies

Both psychodynamic and psychoanalytical techniques (see 'Grief' above) have been used to treat PTSD. Such therapies tend to vary enormously in nature, encompassing individual and group approaches and lasting anywhere from a few sessions to over a year. The available data reveals improvements following treatment, but methodological flaws preclude any clear conclusions. Where alternatives exist, it may well be advisable to pursue these options first.

Hypnotherapy

Hypnosis has reportedly been used with some success. However, there is a lack of sound published data to confirm this, although one controlled study suggests that hypnosis is effective and may be particularly suitable for reducing intrusive symptoms (re-experiencing cluster).

Eye-movement desensitization and reprocessing (EMDR)

This is a controversial treatment, largely because of the surprising claims made by its originator, the lack of rigorous scientific testing to confirm its supposed efficacy, and the lack of any obvious theoretical basis or justification for its beneficial effects. Treatment involves focusing on a disturbing trauma-related thought or image, while visually tracking a movement, for example the therapist's finger. The primary components are therefore production of saccadic eye movements and exposure. Until methodologically sound data is available, clinicians are advised to consider this technique with caution.

Prognosis

The evidence suggests that most treatments are effective in reducing symptoms and improving quality of life, but it seems that the magnitude of these benefits is limited. Many sufferers retain some symptoms despite having received optimal treatment. Although little data exists on the efficacy of combining treatment options (for example, drugs alongside psychological techniques), it may be appropriate to consider this in resistant cases. In addition, where residual symptoms persist, it may be appropriate to shift goals towards rehabilitation and the successful management of symptoms.

What factors influence the chance of recovery following exposure to trauma? Known risk factors for the development of PTSD include the following: the severity of (including proximity to) the trauma; bereavement as a result of the trauma; presence of a pre- or comorbid psychiatric disorder; certain personality traits, for example neuroticism; and the absence of adequate social and psychological support. All appear to be predictors of poor outcome. Recent reviews suggest that neither age nor ethnic group interacts with pathological response to trauma and that PTSD-symptom expression is similar across age groups and cultures.

Areas of controversy needing further research

Controversial issues largely overlap with those requiring further research in this relatively newly recognized area. Particular attention should be paid to:

- SSRIs as a drug treatment of first choice;
- combined treatment approaches;
- crisis intervention as an immediate, unselective, post-trauma intervention;
- the eye-movement desensitization and reprocessing technique;
- long-term outcomes;
- the new diagnostic category, acute stress disorder;
- postconcussional disorder—a category provided for further study (Appendix B, DSM-IV). Symptoms occur following a closed head injury with concussion and include cognitive (specifically, attention and memory), emotional (anxiety, depression, irritability), and physical (sleep problems, fatigue, headache) problems.

Further reading

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders*, 4th ed. APA, Washington DC.

Breslau N, *et al.* (1998). Trauma and post-traumatic stress disorder in the community. *Archives of General Psychiatry* **55**, 626–32.

Brom D, Kleber RJ, Defres PB (1989). Brief psychotherapy for posttraumatic stress disorders. *Journal of Consulting and Clinical Psychology* **57**, 607–12.

Bryant RA, Harvey AG (1997). Acute stress disorder: a critical review of diagnostic issues. *Clinical Psychology Review*, **17**, 757–73.

Davis LL, *et al.* (1997). Post-traumatic stress disorder and serotonin: new directions for research and treatment. *Journal of Psychiatry and Neuroscience* **22**, 318–26.

Foa EB, Rothbaum BO (1998). *Treating the trauma of rape: cognitive behavioral therapy for PTSD*. Guilford Press, New York.

Friedman MJ (1998). Current and future drug treatment for posttraumatic stress disorder. *Psychiatric Annals* **28**, 461–8.

Frueh BC, Brady KL, deArellano MA (1998). Racial differences in combat related PTSD: empirical findings and conceptual issues. *Clinical Psychology Review* **18**, 287–305.

Greenwood DC, *et al.* (1996). Coronary heart disease: a review of the role of psychosocial stress and social support. *Journal of Public Health Medicine*, **18**, 221–31.

Gump BB, Matthews KA (1999). Do background stressors influence reactivity to and recovery from acute stressors? *Journal of Applied Social Psychology* **29**, 469–94.

Irwin M, Pike J (1993). Bereavement, depressive symptoms and immune function. In: Stroebe MS, Stroebe W, Hansson RO, eds. *Handbook of bereavement: theory, research, and intervention*, pp 160–71. Cambridge University Press, Cambridge.

Jacobs S (1999). *Traumatic grief: diagnosis, treatment and prevention*. Brunner Mazel, Philadelphia.

Kim K, Jacobs S (1993). Neuroendocrine changes following bereavement. In: Stroebe MS, Stroebe W, Hansson RO, eds. *Handbook of bereavement: theory, research, and intervention*, pp 143–59. Cambridge University Press, Cambridge.

Kleber RJ, Brom D (1987). Psychotherapy and pathological grief: a controlled outcome study. Israeli *Journal of Psychiatry and Related Sciences* **24**, 99–109.

Marmar CR, *et al.* (1988). A controlled trial of brief psychotherapy and mutual help group treatment of conjugal bereavement. *American Journal of Psychiatry* **145**, 203–9.

Marshall RD, Spitzer R, Liebowitz MR (1999). Review and critique of the new DSM-IV diagnosis of acute stress disorder. *American Journal of Psychiatry* **156**, 1677–85.

Mawson D, *et al.* (1981). Guided mourning for morbid grief: A controlled study. *British Journal of Psychiatry* **138**, 185–93.

Prigerson HG, *et al.* (1995). The inventory of complicated grief: a scale to measure symptoms of maladaptive loss. *Psychiatry Research* **59**, 65–79.

Prigerson HG, *et al.* (1997). Traumatic grief as a risk factor for mental and physical morbidity. *American Journal of Psychiatry* **154**, 617–23.

Raphael B (1977). Preventive intervention with the recently bereaved. *Archives of General Psychiatry* **34**, 1450–4.

Shalev AY, Bonne O, Eth S (1996). Treatment of posttraumatic stress disorder: a review. *Psychosomatic Medicine* **58**, 165–82.

Shapiro F (1989). Eye movement desensitization: a new treatment for PTSD. *Journal of Behavior Therapy and Experimental Psychiatry* **3**, 211–17.

Shapiro F (1995). *Eye movement desensitization and reprocessing: basic principles, protocols and procedures.* Guilford Press, New York.

Simon RI (1999). Chronic posttraumatic stress disorder: a review and checklist of factors influencing prognosis. *Harvard Review of Psychiatry* **6**, 304–12.

Solomon SD (1997). Psychosocial treatment of posttraumatic stress disorder. *In Session-Psychotherapy in Practice* **3**, 27–41.

Solomon SD, Davidson JRT (1997). Trauma: prevalence, impairment, service use and cost. *Journal of Clinical Psychiatry* **58**(Suppl. 9), 5–11.

Stroebe MS, Stroebe W, Hansson RO, eds (1993). *Handbook of bereavement: theory, research, and intervention*. Cambridge University Press, Cambridge.

Weintraub D, Ruskin PE (1999). Posttraumatic stress disorder in the elderly: a review. *Harvard Review of Psychiatry* **7**, 144–52.

World Health Organization (1992). The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. WHO, Geneva.

26.5.2 The patient who has attempted suicide

Keith Hawton

Introduction

The term 'attempted suicide' is usually applied to all acts of deliberate self-harm, in other words deliberate self-poisoning or self-injury. In many ways this is a misleading term since the primary motivation for, or aim of, deliberate self-harm is often not death but some other purpose, such as communication of distress, blotting out an unbearable state of mind, or trying to change the behaviour of other people. Such 'non-suicidal' aims, however, often involve the use of the suicidal message to enhance the impact of such intentions. Because of this and the popularity of the term 'attempted suicide' among physicians, this terminology will be used throughout this chapter.

Attempted suicide is a major and increasing healthcare problem in most developed, and some developing, countries. In the United Kingdom, for example, there are approximately 170 000 general hospital presentations for self-poisoning or self-injury each year. Attempts occur in older children, but the behaviour becomes more common in adolescence, increasing rapidly in frequency in females from the age of 12 years, with peak rates in the late teens and early twenties. The increase with age occurs more slowly in males, rates peaking in the mid- to late twenties. In recent years there has been an increase in attempts by young males, which parallels the trend for completed suicide.

The vast majority of hospital-referred attempts involve self-poisoning. In the United Kingdom the substances most frequently involved are non-opiate analgesics, particularly paracetamol and paracetamol-containing compounds, and psychotropic agents, especially antidepressants and minor tranquillizers. Most self-injuries involve patients cutting themselves.

A wide variety of patient characteristics and problems can lead to attempted suicide. These include psychiatric disorders (especially depression, substance abuse, and anxiety disorders), personality difficulties, and poor coping resources. The more common life difficulties experienced by patients include interpersonal problems and broken relationships (especially in the young), employment difficulties, legal problems, and alcohol and drug misuse.

An important feature of attempted suicide is that it is often repeated, with at least 12 to 25 per cent of people repeating the act within a year. In the United Kingdom between 1 and 2 per cent die by suicide within a year and 3 to 5 per cent within 8 to 10 years. In settings where the patient population tends to be older the risk of suicide within a year of an attempt may be as high as 6 to 10 per cent.

This chapter focuses on the management of those who attempt suicide in the general hospital. It is imperative that patients should not only receive adequate physical care but that their psychiatric and psychosocial problems and needs are assessed.

Arrival of patients at the general hospital

In addition to the immediate assessment of the medical consequences of self-poisoning or self-injury, accident and emergency department staff should be capable of conducting a brief assessment of a patient's psychiatric status and risk. In particular, they need to determine whether a patient has a serious psychiatric disorder (for example, psychosis or severe depression) and/or is actively suicidal such that urgent attention by the psychiatric service is required. Dangerous tablets or other potential methods of self-harm should be removed.

Staff should be aware that a large number of patients leave hospital accident and emergency departments before a psychiatric assessment can be conducted. Such patients often have substance abuse disorders and a history of previous attempts, and may show behavioural disturbance in the department. Many have features associated with suicide risk, and tend to present to hospital with further repeat attempts more often than patients who are assessed in the accident and emergency department. These facts highlight the need for accident and emergency staff to have basic skills in assessment, and for them to be able to readily obtain urgent psychiatric assessment when they judge it to be necessary. Where a patient is thought to be at serious risk but wanting to leave hospital, medical staff can, in the United Kingdom at least, restrain the patient under common law until a psychiatric opinion can be obtained and, if necessary, a Mental Health Act order completed.

Medical care

Management of the medical complications of suicide attempts is dealt with in Chapter 26.6.1, but an obvious difficulty for physicians can arise with those patients who have attempted suicide and refuse potentially life-saving treatment for the physical consequences of their acts. The dilemma is whether to instigate such treatment against a patient's will. This problem most often presents in those who have poisoned themselves, such as with large overdoses of paracetamol, in which early treatment can prevent the development of potentially fatal liver damage.

In the United Kingdom the issue primarily comes down to one of mental capacity. To show that patients have the capacity to refuse treatment, they:

1. must be able to understand and retain information on the treatment proposed, its indications, and its main benefits, as well as possible risks and the consequences of non-treatment;
2. must be shown to believe that information; and
3. must be capable of weighing up the information in order to arrive at a conclusion.

If a clinician instigates treatment against a patient's wishes in spite of the patient appearing to have capacity, then the clinician is at risk of being accused of battery. Where the patient is judged as lacking capacity, essential treatment can either be instigated: (1) directly by a physician, or (2) after the patient has been placed on a Mental Health Order because of the degree of mental illness, in which case the treatment for the physical condition is given because the overdose is judged to be the result of mental illness.

In situations of dire emergency most clinicians would instigate essential treatment to save the patient's life and then try to sort out the legal issues afterwards. Such understandable action is unlikely to lead to successful litigation if the clinician acted in a way that he/she judged at the time to be in the patient's best interest.

Psychiatric assessment

Psychiatric assessment should not usually take place until a patient has recovered from any acute medical effects of an attempt. Clearly, more urgent assessment is indicated if the patient is severely disturbed or regarded as being at acute risk. In some centres, general medical and nursing staff may have to carry out these assessments, either because this is local policy or because of inadequacy of the local psychiatric service. General hospital staff should in any case be familiar with

how to conduct an assessment so that they can do this at times of emergency.

A semi-structured assessment procedure is recommended. The main factors that should be covered are listed in Table 1. A useful way of assessing the events and the patient's problems that preceded the act, the nature of the attempt, possible motivation, and suicidal intent, is to obtain a very detailed account of the few days leading up to the act. Whenever possible the patient's account should be supplemented by enquiry of other informants such as a partner, relatives, and friends. Information should also be sought from professionals and others involved in the patient's care, including the general practitioner.

Suicidal intent and other motives

Suicidal intent (that is to say, the extent to which the patient wished to die at the time of the attempt) can usefully be assessed by examining the circumstances of the act and the explanation given by the patient and by the relatives or friends. Circumstances suggesting high suicidal intent include:

- act carried out in isolation;
- act timed so that intervention unlikely;
- precautions taken to avoid discovery;
- preparations made in anticipation of death (e.g. making will, organizing insurance);
- preparations made for the act (e.g. purchasing means, saving up tablets);
- communicating intent to others beforehand;
- extensive premeditation;
- leaving a note;
- not alerting potential helpers after the act.

It is also important to take account of what the patient and others say about the purpose of the act. Approximately one-third of patients will say that they definitely wanted to die, although in some cases the circumstances of the act will suggest otherwise. There is a small but important group of patients who will claim they did not wish to die when the circumstances strongly suggest high suicidal intent—such patients may be at increased risk of making a repeat attempt, which has a high chance of being fatal. A useful questionnaire which can assist in the assessment of suicidal intent is the Beck Suicidal Intent Scale.

It is extremely important to recognize that the apparent physical danger of an overdose is a poor and potentially misleading measure of the extent to which a patient may have wanted to die. Many patients are ignorant of the relative dangers of substances taken in overdose, although increasing attention to suicidal behaviour by the media may be changing this. Thus a small overdose of a benzodiazepine hypnotic or even an antibiotic may represent a serious attempt at suicide for some patients, whereas a large overdose of a highly dangerous analgesic might be taken with low intent by others. People in the medical and allied professions represent an exception, and usually the danger of their acts is a good measure of intent. Very dangerous self-injuries are often associated with high suicidal intent, but this is not always so.

Assessment of the motives for deliberate self-poisoning and self-injury should be based on the precedents, circumstances of the act, the patient's account, that of other informants, and deduction by the clinician. Motivational reasons that frequently underly this behaviour include:

- to die;
- to escape from unbearable anguish;
- to get relief;
- to change the behaviour of others;
- to escape from a situation;
- to show desperation to others;
- to get back at other people/make them feel guilty;
- to get help.

Risk of a repeated attempt and of suicide

Estimation of the risk of repetition and of suicide following attempted suicide, both short-term and long-term, is a very important part of the assessment. Factors associated with an increased risk of a repeat include:

- previous attempt(s);
- personality disorder;
- alcohol or drug abuse;
- previous psychiatric treatment;
- unemployment;
- lower social class;
- criminal record;
- history of violence;
- age between 25 and 54 years;
- single, divorced, or separated.

Factors associated with an increased risk of suicide in this population include:

- older age;
- male gender;
- unemployed or retired;
- separated, divorced, or widowed;
- living alone;
- poor physical health;
- psychiatric disorder (particularly depression, alcoholism, schizophrenia, and 'sociopathic' personality disorder);
- high suicidal intent in current episode;
- violent method involved in current attempt (e.g. attempted hanging, jumping);
- leaving a suicide note;
- previous attempt(s).

It is essential, however, to recognize that such predictive measures are notoriously imprecise. For repetition, this is because the predictive factors are relatively crude and, whilst those patients who show the risk factors have a high risk of repeating, a substantial proportion of repeaters, possibly more than half, do not demonstrate many risk factors.

Coping resources and supports

Assessment of coping resources and supports should be based on past behaviour under stress and the patient's account of whom they can turn to for support. It is particularly important to assess whether the patient has specific difficulties in problem-solving as these can be an important target for psychosocial therapy. The best evidence for such difficulties will be a description of the methods used to solve problems in the past. It is always important to determine whether current problem-solving is impaired by depression or other psychiatric disorders.

Care after attempted suicide

Patients who have attempted suicide are frequently ambivalent about accepting help, or even frankly dismissive of it. However, this may be understandable in the context of acts that often represent attempts at interpersonal communication or have other functions unconnected with help-seeking. Furthermore, many patients

come from socioeconomic backgrounds in which help-seeking for emotional problems is rarely considered. Therefore clinicians may have to work hard in some cases to explain to patients how treatment might be of benefit. These factors also mean that a brief intervention, such as problem-solving, is likely to be more acceptable to a sizeable proportion of patients than more lengthy therapeutic approaches.

The assessment procedure can itself be highly therapeutic. Patients may be provided with their first opportunity to discuss their difficulties with a clinician. Joint interviews with family members can help highlight issues that need addressing and assist with communication problems.

Some patients thought to be at high risk of suicide refuse psychiatric treatment when this is judged to be essential. Management comes down to a judgement of whether the patient is suffering, or likely to be suffering, from a mental illness that necessitates hospital assessment and/or treatment. In most countries, if a patient thought to be at serious risk and/or mentally ill has presented to a general hospital following attempted suicide but is refusing to stay for a psychiatric assessment, accident and emergency department staff would be judged to be acting reasonably if they restrained the patient under 'common law' until a psychiatric opinion could be obtained.

Currently, there is inadequate evidence for the efficacy of treatments for patients who have attempted suicide, at least with regard to the prevention of repetition of attempts. This is partly to do with methodological flaws in the design of studies, and also because most studies have included too few patients. In one trial, repetition of attempts was reduced in multiple repeaters who received the depot neuroleptic flupentixol, compared with patients who received a placebo. Intensive and prolonged psychological therapy has been associated with promising results in female patients with a history of multiple acts of self-harm and borderline personality disorder. There are also promising results for brief problem-solving therapy.

Clinical services for patients who have attempted suicide

At one time the assessment of patients who attempted suicide was regarded as primarily the responsibility of psychiatrists. Increases in the clinical responsibilities of non-medical clinical staff and findings from research have resulted in a major change in the pattern of services in many places. In the United Kingdom it has been demonstrated that nurses, social workers, and other clinicians can assess these patients reliably, make effective aftercare arrangements, and provide effective therapy. This has resulted in official guidelines that reflect these findings.

It is imperative that staff of whatever discipline who are involved in this work have reasonable background experience and skills in the management of patients with emotional and psychiatric disorders, and that they be properly trained in the assessment and treatment of patients who have attempted suicide. They must also have support from senior psychiatrists, especially for patients with severe psychiatric disorders and where compulsory admission to hospital may be required. They must also be highly motivated and have good support systems in place, because working with such patients can be extremely demanding.

The functioning of a service for patients who have attempted suicide but do not require physical treatment in specialized settings (for example, in an intensive care unit) can be improved if they are admitted to one short-stay medical ward, rather than to a large number of wards. The attitudes of general medical and nursing staff to these patients can be negative, especially towards patients whose acts they perceive as having a low suicidal intent. Clinical experience shows that attitudes are far more favourable when admission to a single ward is possible. General medical and nursing staff in such wards acquire experience in managing these patients, and also develop closer working relationships with members of the service for attempted suicide.

The development of high-quality general hospital services for patients who attempt suicide should be a major element in any national or local suicide-prevention strategy.

Specific subgroups of patients

Alcohol and drug abusers

Many patients who attempt suicide have problems related to alcohol and drug abuse, and these factors, especially alcohol abuse, increase the risk of both repetition and eventual suicide. All attempters should be screened for substance abuse. Recognition of such problems in the general hospital may provide a special opportunity for treatment, which may be an important factor in preventing further suicidal behaviour as well as reducing physical and social harm.

Children and very young adolescents

Very young patients are usually admitted to a paediatric ward where this is available in the general hospital. It is advisable that all very young attempters be admitted to hospital rather than be dealt with in the accident and emergency department, since they require particularly careful and often prolonged assessment, including interviews with their families and the possible involvement of community statutory services (for example, social services).

The elderly

Attempted suicide in the elderly, while less common than in younger people, very often involves high suicidal intent. Routine admission to a medical bed is therefore also recommended for this group. Close links should be established with the local psychogeriatric service (if one exists) so that clinicians from the service can provide assessment and make arrangements for their aftercare.

Further reading

Bancroft J, *et al.* (1979). The reasons people give for taking overdoses: a further inquiry. *British Journal of Medical Psychology* **52**, 353–65.

Beck AT, Beck R, Kovacs M (1975). Classification of suicidal behaviors: I. Quantifying intent and medical lethality. *American Journal of Psychiatry* **132**, 285–7.

Crawford MJ, Wessely S (1998). Does initial management affect the rate of repetition of deliberate self harm? Cohort study. *British Medical Journal* **317**, 985.

Department of Health and Social Security (1984). *The management of deliberate self-harm*. HN **84**, 25. Department of Health and Social Security, London.

Eddleston K, Resvi Sheriff MH, Hawton K (1998). Deliberate self-harm in Sri Lanka—an overlooked tragedy in the developing world. *British Medical Journal* **317**, 133–5.

Hassan TB, *et al.* (1999). Managing patients with deliberate self harm who refuse treatment in the accident and emergency department. *British Medical Journal* **319**, 107–9.

Hawton K, Catalan J (1987). *Attempted suicide: a practical guide to its nature and management*, 2nd edn. Oxford University Press, Oxford.

Hawton K, van Heerinogen K (2000). *The international handbook of suicide and attempted suicide*. Wiley, Chichester.

Hawton K, *et al.* (1998). Deliberate self-harm: a systematic review of the efficacy of psychosocial and pharmacological treatments in preventing repetition. *British Medical Journal* **317,** 441–7.

Nordentoft M, *et al.* (1993). High mortality by natural and unnatural causes: a 10 year follow up study of patients admitted to a poisoning treatment centre after suicide attempts. *British Medical Journal* **306**, 1637–41.

Royal College of Psychiatrists (1994). *The general hospital management of adult deliberate self-harm*, Council Report CR32. Royal College of Psychiatrists, London.

Royal College of Psychiatrists (1998). *Managing deliberate self-harm in young people*, Council Report CR63. Royal College of Psychiatrists, London.

26.5.3 Medically unexplained symptoms in patients attending medical clinics

Christopher Bass and Michael Sharpe

Historical background

Throughout history, patients have presented with subjective somatic complaints that their doctors could not explain in terms of objectively identifiable disease. Medical advances have improved the precision with which disease can be identified, but has not solved the problem. Many complaints remain unexplained and continue to present a challenge to doctors.

The proposed explanations for medically unexplained symptoms have changed over the last 300 years; early ideas located their cause in a disturbance of a bodily organ, often the uterus. Attention then focused on the peripheral nervous system, later the central nervous system, and more recently the cause has been assumed to be in the patient's mental functioning. These changing aetiological theories have been reflected in the varied terms used for such complaints: the organ theory gave rise to hysterical and hypochondriacal; the nervous system theory to nervous, functional nervous illness, and neurasthenic; and the mental theory to psychological, psychogenic, and somatization.

Changes in theory led to changing approaches to management. Early treatments were focused on the organ believed to be giving rise to the symptoms. Consequently manipulations of, and even removal of, the female reproductive organs was practised. When the proposed explanation shifted to the peripheral nervous system the preferred treatments became tonics, electrical stimulation, and other means to regenerate nervous energy. As interest shifted to the central nervous system, hypnosis became a favoured treatment and was used by famous physicians such as Charcot. By the end of the nineteenth century many physicians came to consider hypnosis unnecessary and explanation and advice to be sufficient. It was only in the twentieth century that, with the rise in popularity of psychoanalysis, medically unexplained symptoms began to be seen as a 'mental' problem requiring psychiatric treatment. Much subsequent thinking has emphasized psychological and psychiatric theories and treatments.

At the beginning of the twenty-first century many doctors continue to regard medically unexplained symptoms as psychological in origin because there is no abnormality on standard investigation. However, it has become increasingly obvious that patients dislike the psychological approach, which they see as dismissive and stigmatizing. As a result of these conflicting views patients often seem to be left in a 'no man's land' between a psychiatric conceptualization, which they reject, and a biomedical approach that they see as rejecting them. Recent developments in neuroscience have suggested that many unexplained symptoms do have a basis in the functioning of the nervous system, as hypothesized more than 100 years ago. An approach that recognizes this, whilst also drawing on evidence-based psychological and psychiatric treatment, appears to be the most productive.

Introduction

Definition and terminology

Various terms have been used to describe symptoms that are unexplained by identifiable disease processes. These include:

- *Medically unexplained*—A simple operational term, but with the potential disadvantage of suggesting that psychophysiological explanations are not 'medical'.
- *Functional*—Originally meaning a disturbance of bodily function rather then structure, it is unfortunately used pejoratively to mean 'all in the mind'.
- *Somatization*—A widely used term implying a psychological problem expressed somatically. It should arguably be restricted to cases where the somatic symptoms are plausibly understood as an expression of identifiable emotional disorder.
- *Conversion*—Used specifically to refer to loss of function such as weakness of a limb. Implies (as does somatization) that the symptoms are due to a 'conversion' of psychological problems, usually without good evidence.
- *Somatoform*—A diagnostic category in the psychiatric classifications. Intended to be atheoretical but obviously linked to the idea of somatization.

In conclusion, there is no entirely satisfactory term: the best term scientifically is probably 'medically unexplained', or perhaps 'unexplained by identifiable disease', but many textbooks and computer databases use the term 'somatoform'.

Symptoms

Almost any symptoms can be medically unexplained. Common examples include:

- pain (including back pain, chest pain, abdominal pain, and headache);
- fatigue;
- dizziness;

- 'fits', funny turns, and feelings of weakness.

Syndromes

Unexplained symptoms have been grouped into various 'functional' syndromes. Each medical specialty has at least one (Table 1): for rheumatologists, prominent muscle pain and tenderness is fibromyalgia; for gastroenterologists, abdominal pain with altered bowel habit is irritable bowel syndrome; and for infectious-disease specialists, chronic fatigue and myalgia is a postviral or chronic fatigue syndrome.

It has been argued that these syndromes do not necessarily reflect separate conditions, but merely reflect the tendency of specialists to focus only on those symptoms most pertinent to their specialty. The research literature offers support for this hypothesis by revealing substantial overlap between syndromes in their constituent symptoms, proposed aetiological factors, and response to treatment.

The significance of medically unexplained symptoms and syndromes

Medically unexplained somatic complaints are a common and important but relatively neglected medical problem, constituting a major part of the work of most doctors. Whilst sometimes regarded as merely 'worried well', patients with medically unexplained complaints often suffer disability and distress at least as severe as that of those whose symptoms are explained by disease. Their doctors often find them difficult to help. They are also expensive to the healthcare system because they attend multiple specialist services and receive extensive, but unproductive, investigation and treatment. Many not surprisingly turn to unconventional treatments that are of unproven effectiveness. Some are financially exploited. This situation is clearly unsatisfactory.

Aetiology

The precise aetiology of many medically unexplained symptoms is unknown, but there is evidence that biological, psychological, and social factors all play a role. The degree to which each of these contributes probably varies from case to case. Rather than seeking a single factor, it is helpful to consider multiple factors and to distinguish between those that predispose to the development of medically unexplained symptoms, those that precipitate them, and those that act to perpetuate them. Table 2 provides a summary of possible aetiological factors, perpetuating factors being especially important since they are targets for treatment. For example, a person may be predisposed by virtue of genetics or childhood experience to develop irritable bowel syndrome. This may have been precipitated by a combination of infection and psychological stress. The factors that perpetuate it may include neurophysiological mechanisms, fear of gastrointestinal disease, social stress, chronic anxiety, and iatrogenic factors such as overinvestigation.

Epidemiology and classification

Epidemiology

Medically unexplained somatic symptoms are extremely common in the general population in all countries, some of whom will visit doctors, usually because of concern about the cause or because of severe discomfort and disability.

- *Primary care*—Medically unexplained symptoms are the principal reason for 25 to 50 per cent of all consultations in primary care. A minority are referred for a specialist opinion.
- *Hospital outpatient care*—At specialist outpatient clinics between one-quarter and one-third of new patients have symptoms that remain unexplained by disease. A small number of such patients are admitted to hospital.
- *Hospital inpatient care*—The proportion of medical inpatients with unexplained complaints is lower than amongst outpatients, but these patients can be particularly costly to the service. One Scandinavian study found that a relatively small number of patients with recurrent multiple medically unexplained symptoms (referred to as somatization disorder as described below) were consuming a significant proportion of the country's hospital inpatient budget.

Classification

There are, rather confusingly, parallel medical and psychiatric classification schemes for medically unexplained complaints. The former emphasizes the type of symptom and lists functional syndromes by organ system as in Table 1; the latter emphasizes the number of symptoms and associated psychological factors, with the main categories as listed in Table 3. The implication of these parallel classifications is that most patients will qualify for both a medical and a psychiatric diagnosis. Both may be useful in guiding management and prognostication, hence a combined medical/psychiatric diagnosis such 'irritable bowel syndrome/anxiety disorder' is probably more useful than either alone.

An alternative multidimensional classification system that combines both the medical and psychiatric approach has been suggested and may have additional clinical value. An example is shown in Table 4.

Pathogenesis and pathophysiology

Although the physiological mechanisms of symptom production are not fully understood in many cases, some physiological abnormalities and putative mechanisms have been identified, for example the effect of overbreathing in causing non-cardiac chest pain. These physiological mechanisms interact with psychological and social factors to perpetuate the illness. Hence, overbreathing gives rise to chest pain and paraesthesias, which is interpreted by the patient as a cardiac problem, leading to anxiety and the seeking of medical care. Medical investigation increases the anxiety, leading to further hyperventilation, and so on. Some suggested physiological mechanisms are listed in Table 5.

General clinical features

Pointers to a patient having medically unexplained complaints may be apparent before the initial consultation. The referral letter and medical notes may reveal frequent attendance at medical services, numerous negative (and often repeated) investigations, and a previous history of unsuccessful surgery.

At the consultation multiple symptoms are suggestive. However, the only way of confidently diagnosing complaints as medically unexplained is when the appropriate history, examination, and investigation reveal one or more somatic symptoms that remain unexplained by disease. It should be remembered that patients often have both symptoms that are explained by disease and others that are unexplained. An example is the frequent co-occurrence of epilepsy and non-epileptic attacks. Several general points are worth noting:

- Many (but not all) unexplained medical complaints are simply somatic symptoms of depression or anxiety.
- Most medically unexplained complaints reflect genuine suffering. The deliberate manufacturing of complaints with the intent to mislead is uncommon in ordinary medical practice. However, some degree of exaggeration and even frank deception in order to obtain financial gain is not uncommonly encountered in medicolegal practice.
- Although management may be based on general principles, psychiatric diagnosis may indicate additional specific and evidence-based treatment strategies.

Differential diagnosis

The main medical differential diagnosis is from symptoms due to disease. Difficulties are likely to involve unusual presentations of common diseases and rare diseases. Missed disease is always a concern, but once a patient has been carefully assessed the emergence of a 'missed' disease is the exception rather than the rule.

Management

The general principles of management are outlined in Table 6.

Patient assessment

When assessing the patient the main tasks are to:

- Understand the nature of the presenting symptoms. For example, what does the patient mean by their complaint of fatigue? Is it lack of energy (non-specific), sleepiness (suggesting a sleep problem), or lack of motivation (suggesting depression)?
- Find out what other symptoms the patient has. It is worth asking for an exhaustive list: the more symptoms, the more likely it is that they will be medically unexplained.
- Ask the patient what they think or fear is wrong with them. This can reveal the reason they are worried about the symptoms (for example, 'it could be cancer') and allow appropriately targeted education and reassurance to be given.
- Seek evidence of 'stress'. Life stresses may be a contributory factor. Furthermore, most patients find 'stress' to be a more acceptable explanation than psychiatric diagnoses such as depression or anxiety.
- Systematically seek evidence of depression and anxiety. This is often best done toward the end of the consultation so that the patient does not feel they are being dismissed as 'just psychiatric'. A useful approach is to empathize with the understandable distress resulting from the symptoms, thereby avoiding antagonizing the patient by giving the impression that the doctor believes that the cause of the symptoms is psychological.
- Physically examine the patient. This may reveal unsuspected signs of disease and also helps to convince the patient that they have been taken seriously and properly assessed.
- Perform appropriate investigations. It should be noted that misdiagnosis is relatively uncommon, and a balance needs to be struck between the risk of missing disease and the potential iatrogenic harm resulting from excessive investigation.

Giving reassurance explanation and advice

As well as being reassured that there is no evidence they have an unpleasant disease, patients benefit from being given a positive and credible explanation for their symptoms and practical advice on what to do next.

Reassurance

Giving appropriate reassurance is an important part of the medical consultation. This is most effective if based on the patient's actual concerns, so it is important to ask them what they are worried about before reassuring them. Many patients report that having a physical examination is particularly reassuring. A detailed explanation of what the tests that they have had do and don't show can also help. Clearly, it may be unwise to state categorically that the patient has no disease. However, it can be explained that it is not possible to do this, whilst emphasizing as unambiguously as possible that the probability they have the disease they fear is very low. Beware of the patient who repeatedly asks for reassurance about the same issue—they may have hypochondriasis (see below).

Explanation

Patients also need a positive explanation for their symptoms. It is nearly always unhelpful to explain that the symptoms are 'just psychological' or 'all in the mind'. Such statements are likely to reduce confidence in the doctor and may paradoxically increase the patient's concern about missed disease. It is also potentially harmful to suggest that the patient has a disease when they do not, or to collude with their idea that they do so. This may lead to inappropriate coping behaviour, for example obtaining a wheelchair rather than seeking rehabilitation. Rather, it is useful to describe a plausible physiological mechanism for the symptom that emphasizes the link with psychosocial factors and helps the patient to see their symptoms as reversible. For example, it can be explained that in irritable bowel syndrome psychological stress results in increased activation of the autonomic nervous system, leading to constriction of smooth muscle in the gut wall, which in turn causes pain. The symptoms may therefore be perpetuated by a vicious circle in which pain leads to anxiety, and anxiety leads to further pain. It can then be explained (perhaps using a diagram) that this mechanism is reversible by targeting these perpetuating factors. It is helpful to offer an optimistic prognosis, but an unrealistically precise prediction ('you will be better next week') is unwise as it is likely to lead to loss of faith in the doctor if not fulfilled.

Advice

A positive plan of action that specifies both what the patient can change and what the doctor can do is helpful. The patient can be advised how to overcome probable perpetuating factors, for example by resolving stress causing social problems or by practising relaxation. The doctor can offer to review progress, to prescribe (for instance, an 'antidepressant' drugs) and if appropriate to refer, for example, to physiotherapy or psychology. Action by the doctor gives the patient a sense that they are being taken seriously and not (as they may have experienced before) being dismissed. Writing to the patient as well as to the general practitioner to summarize the conclusions of the medical assessment and the proposed plan reinforces messages that may otherwise be easily forgotten.

'Antidepressant' drugs

Antidepressant drugs are most useful when the patient is depressed, but they can also be tried when he or she is not. A specific explanation of why they are being prescribed is needed if they are to be acceptable to the patient. Depending on the circumstances, one of the following two approaches is suggested. The first is to explain that the term 'antidepressant' is a misnomer: in fact the drugs are broad-spectrum 'nervous tonics' of proven value for sleep and pain as well as for depression. The second is to be explicit that they are being prescribed for depression, but emphasize that depression is understandable given the somatic symptoms the patient is suffering. Both these explanations minimize blame and stigma.

A systematic review of antidepressants for medically unexplained symptoms found them to be moderately effective overall. The odds ratio (**OR**) for improvement with antidepressant treatment compared with placebo was 3.4, with the size of effect similar across the different functional syndromes. However, there was a high dropout rate from treatment, emphasizing the need for careful explanation and follow-up to ensure adherence.

Psychological therapies

Explanation, reassurance, and advice are important psychological therapies. Where insufficient they can be reinforced by a formal psychological treatment. The most widely used are behavioural or cognitive–behavioural treatment (**CBT**), although other psychological treatments may have a role. CBT aims to help the patient to improve by examining their way of thinking about and coping with their symptoms. The treatment works by changing potentially illness-perpetuating beliefs and coping behaviours.

A systematic review of CBT for medically unexplained symptoms found that it was significantly superior to non-specific treatment in 70 per cent of trials. Individual behavioural components of CBT, such as graded exercise, have also been studied and are of value, but are probably less effective. Overall research supports the use of specialist CBT and other behavioural therapies in the management of patients with medically unexplained symptoms.

Referral for psychological or psychiatric management

The decision to refer will be based on the physician's assessment of the patient and an appraisal of the available services. Ideally, all medical clinics would have dedicated specialist psychiatric or psychology services: this is rarely the case in practice. It is wise to find out who is willing and interested in receiving referrals in the immediate locality. If the options are few, however, the healthcare team may wish to make the case for the provision of better services.

Reasons to refer include:

- very severe disability;
- suspected somatization disorder;
- specific service available, e.g. for chronic pain or chronic fatigue syndrome (**CFS**);

- patient remains distressed despite explanation and reassurance;
- suicide risk.

When explaining the referral to the patient it is wise to:

- Emphasize the reality of the patient's symptoms.
- Be positive about the service you are referring to.
- Do not prematurely imply you think the origin of their complaint is psychological.

A patient is more likely to attend for a referral if you have said: 'I see you have real and troublesome symptoms. I am pleased to tell you that they do not indicate a disease but am sorry to say that I do not have a simple cure I can prescribe. However, I can recommend and refer you to a specialist service for your problem', than if you have said: 'there is clearly nothing wrong with you; it must all be in your mind. There is nothing to do now but to refer you to the shrinks.'

What happens in practice?

In practice the ideal management described above is inconsistently applied and iatrogenic psychological and physical damage is probably common. The specific evidence-based treatments of antidepressants and CBT are rarely offered and frequently refused by patients. We could do better.

Prognosis

The prognosis for those patients whose symptoms are sufficiently severe and persistent for them to be referred to a specialist service is often poor. It is not uncommon for persistent symptoms and disability to persist for years, especially if untreated The prognosis is best for those patients who were well before the onset of the complaint and whose symptoms are expressions of uncomplicated depressive and anxiety disorders. It is worst for those patients with long-standing multiple symptoms.

Prevention

We do not know enough about the aetiology of medically unexplained complaints to implement primary prevention. Parenting behaviour and social factors such as the stigma associated with psychiatric illness and the nature of benefit and litigation systems appear to play a role.

Secondary prevention is important, as effective early management probably reduces the risk of chronicity, whereas poor explanation and overinvestigation probably perpetuates it.

Tertiary prevention is the effective management of chronic somatization and requires a proactive management plan as described above.

Quality of life and psychological aspects

Patients with medically unexplained symptoms may have considerable functional impairment and a markedly reduced quality of life. Two disorders in particular, fibromyalgia and chronic fatigue syndrome, are associated with marked functional impairment and can have important consequences on a person's ability to work. Some patients may become involved in medicolegal claims, especially if the symptoms were temporally related to an accident or injury.

Areas of uncertainty and controversy

Many aspects of medically unexplained somatic complaints are controversial. Perhaps the most controversial issue is whether they are best regarded as a psychiatric/psychological or as a medical problem. This issue has been particularly prominent in controversy over chronic fatigue syndrome/myalgic encephalomyelitis (**ME**).

Any clinician who sees patients with medically unexplained symptoms (and most do) is likely to agree that they are difficult to manage effectively. Furthermore, the core of this problem is frequently a clash between the physician and the patient in how the illness is viewed. A consideration of the changes in medical fashion for explaining such symptoms over the last few hundred years should encourage humility. Further study that integrates psychosocial and physiological perspectives is required, both to improve our understanding and to help us to explain these problems effectively to our patients.

Common syndromes

This section will cover specific aspects of the main psychiatric categories of medically unexplained symptoms listed in Table 3.

Predominant worry about disease—hypochondriasis

The central feature of hypochondriasis is a persistent preoccupation with the possibility that one has a serious and progressive physical disease. These fears have persisted despite the fact that repeated investigations and examinations have identified no adequate physical cause, and appropriate reassurance that there is no physical disease has been given.

Example

The patient is a young woman. Her urgent referral is faxed by her general practitioner (**GP**). She has attended the general practice clinic daily for the last 2 weeks. Her history is of headache. On enquiry her main concern is that she has developed a brain tumour. The background history reveals that her father died of a brain tumour 2 years ago; he was told it was only a 'tension headache' by the GP. Examination is normal but she is very anxious indeed and repeatedly asks for a brain scan and for your reassurance that she doesn't have a tumour.

Management and prognosis

Assessment should elicit the patient's specific fears. A useful question is: 'what is your worst fear?' The patient's catastrophic fear leads understandably to their behaviour and anxiety. For example, a patient with headache who thinks, 'I have a brain tumour' will be anxious and seek urgent medical attention. That is to say the somatic symptoms (headache) leads to thoughts (brain tumour) that in turn lead to anxiety and to certain behaviours (visiting doctors). It also causes muscular tension that leads to further pain and so on. Eliciting this causal sequence from the patient, and describing it to them, perhaps with a diagram, aids explanation. Reassurance that they do not have a tumour is appropriate. However, repeated reassurance can worsen concern about disease and should be avoided. Depressive disorder should be treated. CBT can be effective. Many acute cases resolve, but the condition can become chronic.

Somatic presentation of depression and anxiety

One of the commonest causes of medically unexplained somatic complaints is undiagnosed depression. Because depression is (erroneously) thought of as a purely 'mental illness', it is readily forgotten that it has somatic symptoms. These include:

- fatigue;
- increased pain complaints;
- loss of weight and appetite;
- loss of libido;
- in severe forms there may be negative ruminations on health that in rare cases can be delusional.

Another very common cause of unexplained symptoms is anxiety in a generalized form (generalized anxiety disorder) or episodic severe form (panic disorder). Somatic symptoms of anxiety include:

- fatigue;
- dizziness;
- paraesthesias;
- chest pain and palpitations;
- shortness of breath (especially 'getting enough breath in').

Example

A 50-year-old man presents to the Accident and Emergency department with chest pain and fatigue. Investigations are negative. Only after admission to hospital does a junior doctor examine the patient's mental state and find a persistent low mood, loss of interest, and negative thinking, with episodic anxiety associated with overbreathing and many somatic symptoms. The diagnosis is panic disorder and depressive disorder.

Management and prognosis

Treatment is by reassurance, explanation, and treating the depression and anxiety. This can be achieved by prescribing an antidepressant agent combined with explanation and active follow-up to ensure adherence. In some cases psychological treatment may also be required. The prognosis for recovery within 6 months is good, although there is a risk of relapse. Patients should continue to take the antidepressant drug for at least 6 months to prevent early relapse.

Simple somatoform disorders

This presentation refers to the patient with a single or small number of somatic complaints that do not appear to be simply expressions of depression or anxiety. This diagnosis includes undifferentiated somatoform disorder and somatoform pain disorder. The most common type of clinical problem in this group is the patient with chronic persistent pain in one site, for example chronic back pain or chronic headache. The physical symptoms may have commenced after a trivial injury or accident, but the subsequent disability is usually out of proportion to the organic findings. A medicolegal case may be pending, which often makes management more difficult.

Example

A middle-aged man presents with widespread pain. He believes that his symptoms are a result of occupational exposure to printing ink. He is medically retired. There is a history of depression some 12 months ago, but no evidence of current depressive or anxiety disorder. Examination is normal. The patient walks with a stick and seems concerned to demonstrate to you how ill he is.

Management and prognosis

Explanation and symptomatic treatment including cognitive–behavioural therapy is appropriate. Antidepressants may be tried on an empirical basis. The prognosis for chronic complaints is fairly poor, although improvement often occurs over many months. Physical rehabilitation may be useful.

Somatization disorder (Briquet's syndrome)—patients with chronic multiple complaints

'Somatization disorder', or Briquet's syndrome, is a term used to describe patients—mostly women—who have a lifelong history of multiple recurrent somatic complaints, which usually include conversion symptoms. If looked for, such patients are relatively easy to identify. Although the patient's current presenting complaint may be of only one or two symptoms, for example chest pain or shortness of breath, scrutiny of the past medical notes reveals numerous outpatient visits to different clinics with symptoms such as abdominal pain and bloating, diffuse muscular pains and tenderness, and chronic lassitude over previous years. A history of childhood abuse is common.

Example

The patient is a middle-aged woman, referred to as a 'heart sink' patient by the GP. She presents with dizziness, bloating of her stomach, and generalized weakness. She has three volumes of medical notes documenting presentations with a range of symptoms including pain in a number of sites, irritable bowel symptoms, menstrual problems, and transient loss of sight. She has had many investigations, a hysterectomy and three laparotomies, and she is taking many medications including oral opiates. However, review of her notes does not reveal any convincing evidence of any proven disease and examination reveals only a number of operation scars.

Management and prognosis

If possible, it is sensible to review the case notes before the patient arrives in the clinic. It is worth asking why the patient has been referred now: the GP may have become frustrated or angry, or helpless in the face of repeated complaints, or unable to cope with or contain the patient's distress. Management requires a reduction in the patient's and their general practitioner's expectation of medical 'cure', and a shift toward the development of coping strategies. Further investigation should be strictly limited to that which is clearly indicated, but may be difficult to avoid. Practical management strategies are listed in Table 7.

Depression is common in such patients, may give rise to some of the symptoms, and should be treated. Long-term follow-up by a hospital doctor, general practitioner, psychiatrist or a combination of these is desirable. A realistic aim is for limited improvement and prevention of iatrogenic harm. The condition is very likely to persist.

Conversion disorder

The presentation is loss of function of a body part, most often a limb, or abnormal body movements. This is not thought to be produced intentionally, as with factitious disorder and malingering, but rather 'subconsciously'. In reality this distinction is difficult. Persistence of the symptom even when the patient believes they are unobserved helps to differentiate intentional and non-intentional symptom production. Other oft-quoted signs such as 'belle indifference' are unreliable. In acute cases the symptoms may have arisen in the context of severe interpersonal stress or conflict. The patient may have experienced a family member with similar symptoms.

Example

A young woman is referred for unexplained weakness of her left side. She reports being unable to move her left arm and leg following a bang on the head. Enquiry reveals that she was sexually assaulted the week before her symptoms began. She has no explanation for the symptoms and seems unconcerned about them. On examination her left arm and leg appear weak, with a 'collapsing' pattern of weakness, but there are no abnormalities of reflexes. A magnetic resonance imaging (**MR**I) brain scan is normal.

Management and prognosis

For the acute case an early return to function should be encouraged. Physiotherapy may be useful. Depression should be sought and treated if present. The patient should also be offered an opportunity to talk about stressors. In chronic cases treatment is more difficult and may best be achieved via referral to a physical rehabilitation service. In chronic cases there may be long-term invalidism with dependence on state benefits.

Body dysmorphic disorder—requests for surgery to a body part in the absence of a conspicuous abnormality or deformity

Patients who are dissatisfied with some aspect of their bodily appearance or shape may find their way into the clinics of physicians and surgeons. Some of these patients are preoccupied with an imagined defect in appearance, such as a large nose. Others have more substantial concerns, such as disliking a limb sufficiently to

desire its amputation. Even in the presence of a physical abnormality, the person's concern appears to be markedly excessive.

Example

A 30-year-old woman complains about the shape of her nose and blames this for many failures in her life. She requests plastic surgery.

Management and prognosis

In the absence of a conspicuous physical abnormality or deformity, it is usually prudent to seek a psychiatric opinion. There is some evidence for the effectiveness of CBT. Surgery sometimes helps, but should only be carried out after careful consideration.

Other unusual presentations

Atypical eating disorders

Some patients who attend with symptoms of abdominal pain after food, vomiting, anorexia, or constipation may have a covert eating disorder. Judicious enquiry about weight loss, amenorrhoea, and laxative use may help to establish the correct diagnosis, which may have been present for many years. See Chapter 26.5.5 for further discussion.

Chronic constipation that has not responded to conventional treatment

Patients may report constipation for two or three decades that has proved unresponsive to appropriate treatment. There is often no identifiable organic cause, and some are referred to surgeons for colectomy. All such patients require a thorough psychiatric assessment before major decisions about treatment are implemented. Biofeedback has been shown to benefit some patients, but resources for this treatment are scarce.

Loin-pain haematuria syndrome

This is a rare syndrome that is poorly understood. Patients may abuse opiate analgesics, and a proportion is thought to simulate their pain. It is unwise to carry out renal autotransplantation, because the pain can recur on the opposite side of the body.

Factitious disorders

Patients with factitious disorder deliberately feign or simulate illness, which is in contrast to the disorders described above. The term 'factitious disorder' is preferable to the eponym Munchausen's syndrome, because this stereotype is often misleading—Munchausen's syndrome generally being applied to wandering, untreatable, male sociopaths. Over three-quarters of patients with factitious disorders are women, over half of whom work in medically related occupations. They often report a large number of childhood illnesses and operations. High rates of substance abuse, mood disorder, and personality disorder have been reported. Some patients exploit genuine disease (usually chronic) to create dramatic medical emergencies.

Example

A young man is repeatedly admitted with breakdown of an abdominal surgical wound. He is observed rubbing dirt into the wound. When confronted with this he discharges himself immediately.

Management and prognosis

A supportive confrontation is required. This means presenting the patient with the evidence that they have been manufacturing symptoms, together with the acknowledgement that they have an emotional problem and an offer of psychological help. Ideally, the physician and psychiatrist do this jointly. There is some evidence that psychological support following hospital discharge may be associated with an improved outcome. Patients with more stable social networks have a better prognosis than wanderers.

Malingering

Malingering is the deliberate simulation or exaggeration of physical or psychiatric symptoms for obvious and understandable gain, for example monetary compensation, disabled status, and avoidance of criminal prosecution or conscription. Doctors are often reluctant to diagnose malingering lest it adversely affect the individual's healthcare, occupation, or legal case. However, it does occur, especially in medicolegal contexts. Common examples include malingered cognitive deficit and postinjury back/neck pain in patients seeking compensation or disability payment. The initial physical injury may be established, but the length and severity of symptoms, disability, and distress are disproportionate. There is probably a continuum from minor embellishment of symptoms to frank malingering.

Example

A 30-year-old man is seen in order to produce a legal report for the purpose of claiming compensation. He reports severe back pain following a car accident 2 years ago and is in a wheelchair. There are no neurological signs and no muscle wasting. He is later seen lifting his wheelchair into the boot of his car.

Management and prognosis

Assessment of suspected malingering requires the methodical use of all sources of information. It may only be possible to make the diagnosis positively when covert surveillance reveals behaviour clearly at odds with the reported disability. If management is required it is by confrontation, although it is more common that the identification of malingering ends the behaviour (and the legal case).

Further reading

Barsky AJ, Borus JF (1999). Functional somatic symptoms. *Annals of Internal Medicine* **130**, 910–21.

Barsky AJ (1998). A comprehensive approach to the chronically somatizing patient. *Journal of Psychosomatic Research* **45,** 301–6.

Bass C, Gill D (2000). Factitious disorders and malingering. In: Gelder M, *et al*., eds. *New Oxford textbook of psychiatry* pp. 1126–32. Oxford University Press, Oxford.

Bowman ES (1998). Pseudoseizures. *Psychiatric Clinics of North America* **21,** 649–57.

Carson AJ, *et al.* (2000). Do medically unexplained symptoms matter? A prospective cohort study of 300 new referrals to neurology outpatient clinics. *Journal of Neurology, Neurosurgery and Psychiatry* **68,** 207–10.

Creed F, Mayou RA, Hopkins A (1992). *Medical symptoms not explained by organic disease*. The Royal College of Psychiatrists and the Royal College of Physicians of London, London.

Chambers J, Bass C, Mayou R (1999). Non-cardiac chest pain: assessment and management. *Heart* **82,** 656–7.

Drossman DA (1995). Diagnosing and treating patients with refractory functional gastrointestinal disorders. *Annals of Internal Medicine* **123,** 688–97.

Fink P, *et al.* (1999). Somatization in primary care. Prevalence, health care utilization, and general practitioner recognition. *Psychosomatics* **40,** 330–8.

Kroenke K, Swindle R (2000). Cognitive–behavioral therapy for somatization and symptom syndromes: a literature synthesis. *Psychotherapy and Psychosomatics* **69**, 205–15.

Mayou RA, Bass C, Sharpe M (1995). *Treatment of functional somatic symptoms*. Oxford University Press, Oxford.

O'Malley PG, *et al.* (1999). Antidepressant therapy for unexplained symptoms and CNS syndromes. *Journal of Family Practice* **48,** 980–90.

Sharpe M (1998). Doctor's diagnoses and patient's perceptions; lessons from the chronic fatigue syndrome. *General Hospital Psychiatry* **20,** 335–8.

Sharpe M, Bass C (1992). Pathophysiological mechanisms in somatization. *International Review of Psychiatry* **4,** 81–97.

Smith RC (1991). Somatization disorder: defining its role in clinical medicine. *Journal of General and Internal Medicine* **6,** 168–75.

Wessely S, Nimnuan C, Sharpe M (1999). Functional somatic syndromes: one or many?. *Lancet* **354,** 36–9.

26.5.4 Anxiety and depression

L. Chwastiak and W. Katon

Introduction

Depression and anxiety are more commonly seen in primary care than any other condition except hypertension. Over half of all patients with mental health disorders are cared for solely by a primary care provider, and many seek help from their physician because they attribute symptoms of depression and anxiety to a physical problem. Physicians fail to make an accurate diagnosis in at least 50 per cent of those with depressive or anxiety disorders. This failure occurs because of time limitations, lack of knowledge, focus on presenting physical symptoms, fear of opening 'Pandora's box', and the stigma associated with psychiatric illness. As a result, only about half of those patients with major depression receive the dose and duration of treatment with antidepressants that meets United States Agency for Health Care Policy and Research (**AHCPR**) guidelines. Fewer than half of patients with anxiety disorders in primary care are treated with specific medications or psychotherapy.

Early recognition of depressive and anxiety disorders is important because they can be treated effectively, often in primary care. The detection and treatment of anxiety and depression can prevent patient discomfort and unnecessary expensive diagnostic investigations, which are often ordered to investigate unexplained physical symptoms. More serious complications may also be prevented, such as the progression of psychiatric illness, loss of employment and impairment of social roles, decreased adherence to medical treatment, and suicide.

Epidemiology

Depression and anxiety are very common and frequently follow a chronic course. The National Comorbidity Survey (**NCS**) found, through a structured psychiatric interview, that almost 50 per cent of community respondents in the United States had at least one lifetime DSM-III-R (*Diagnostic and statistical manual of mental disorders, third edition, revised*) psychiatric disorder. Major depression was the most common disorder with a lifetime prevalence of 17.3 per cent and a 12-month prevalence rate of 10.3 per cent. The point prevalence of depression in Western industrialized nations is from 2.3 to 3.2 per cent for men and 4.5 to 9.3 per cent for women; lifetime risk is 7 to 12 per cent for men and 20 to 25 per cent women. The reasons for gender differences are poorly understood, but are thought to include endocrine, biological, and sociocultural factors: women have been found to have experience higher rates of childhood (especially sexual) abuse.

The prevalence rates of depression are higher for medical patients than for the general population: 6 to 10 per cent of primary care patients; and 10 to 14 per cent medical inpatients meet the criteria for major depression. In patients with at least two chronic physical illnesses, the 12-month prevalence rate of major depression was 12.5 per cent. Major depression is commonly associated with cardiovascular disease (prevalence rate of 25 per cent), cancer (20 to 45 per cent), cerebrovascular accidents (26 to 34 per cent), chronic pain (33 to 35 per cent), and Parkinson's disease (40 per cent). Depression and panic disorder are also common in 'medically unexplained' syndromes (Table 1, and see Chapter 26.5.3).

Some 16 per cent of people experience an anxiety disorder sometime during their lifetime. These disorders are commoner in women, with a 12-month prevalence of panic disorder in community samples of 3.2 per cent in women and 1.3 per cent in men, and a lifetime prevalence of 5 per cent and 2 per cent, respectively. Infrequent panic attacks occur in up to one-third of individuals. Only 26 per cent of patients with panic disorder present to a mental health setting: one-third present to an emergency room, and over a third to their primary physician.

About 80 per cent of psychiatric disorders are comorbid disorders: almost half the cases of depression and anxiety occur in the same patients at the same time. About 25 per cent of patients with major depression, dysthymia, and anxiety disorders also have a history of substance abuse.

Costs

Worldwide, depression is the leading cause of years lived with disability. In the United States in 1990, the estimated total annual cost of major depression was $44 billion, and for anxiety disorders it was $42.3 billion, of which $23 billion was attributed to non-psychiatric medical costs, reflecting the high degree of medically unexplained symptoms these patients experience. Patients with anxiety and depression make more emergency visits, primary care visits, and more telephone calls to their physicians, have more medical tests and evaluations, take more medications, and are more likely to be admitted to hospital for a medical disorder than patients who do not have these disorders. As many as 50 per cent of 'high utilizers' of medical care services have a current depressive or anxiety disorder. Even after controlling for age and pre-existing medical comorbidity, patients with depression receive two to four times as much non-psychiatric medical care as patients without depression, and those with panic disorder use three times as many services as other primary care patients.

In patients with chronic medical illness, depressive and anxiety disorders are associated with an amplification of physical symptoms, additional functional impairment, and a decreased ability to adhere to medication and important lifestyle changes (exercise and diet). Depression reduces the effectiveness of rehabilitation in older patients with stroke, Parkinson's disease, heart disease, fractures, and pulmonary disease. Effective treatment of major depression reduces physical symptoms and functional impairment in patients with chronic medical illness.

The indirect costs of depression and anxiety include mortality, absenteeism from work, and adverse effects on family roles. In the WHO study in primary care, depression was associated with 6.1 disability days per month—as much or more than eight chronic illnesses, including coronary artery disease and arthritis. Anxiety disorders usually strike people at the beginning of their working lives, and may last for many years. Some studies reported that over half of those with panic disorder were not working, and others were forced to take lower-paying or part-time jobs near their homes. The lost economic productivity is easier to quantify than the other indirect costs: lost earning time of family or friends bringing patients to treatment, decreased efficiency at work, the toll on families, the future costs to society of children reared by an unemployed or housebound parent. The role functioning of patients with panic disorder is substantially lower than that of patients with chronic medical illnesses, but higher than that of depressed patients.

Pathophysiology

Depressive and anxiety disorders are complex syndromes that are diagnosed based on clinical criteria. No clear anatomical, physiological, or biochemical explanation has been found. Pathophysiological hypotheses for these disorders involve endocrine characteristics, abnormalities in levels of particular neurotransmitters, and neuroanatomical changes.

Data support an underlying genetic component for major depression, panic disorder, generalized anxiety disorder, and obsessive–compulsive disorder. There are significantly higher rates of depression and panic disorder in first-degree relatives of patients with these disorders. Twin studies have shown a higher rate of

concordance for monozygotic compared to dizygotic twins in patients with panic disorder and major depression.

Mood disorders are associated with heterogeneous dysregulation of the biogenic amines, noradrenaline (norepinephrine) and serotonin being the two neurotransmitters most implicated. In animal models, long-term treatment with virtually all antidepressants is associated with a decrease in the sensitivity of postsynaptic b-adrenergic and type-2 serotonin receptors. Anxiety disorders are associated with abnormalities of the same neurotransmitters as well as in receptors of the neurotransmitter g-aminobutyric acid (GABA), an inhibitory neurotransmitter found mainly in the cerebral cortex.

A variety of neuroendocrine abnormalities have been reported in patients with mood disorders, but it is unclear whether these are the cause of the mood disorder, or reflect an underlying brain disorder. Among the more consistent observations in patients with major depression is dysfunction of the hypothalamic–pituitary–adrenal (**HPA**) axis, presenting as elevation of basal cortisol, dexamethasone-mediated negative feedback resistance, increased cerebrospinal fluid levels of corticotropin-releasing factor (**CRF**), and an ACTH response to challenge with exogenous CRF. These features appear to be markers of state rather than trait: they usually normalize after successful treatment.

There is evidence suggesting that panic disorder is associated with specific biological abnormalities in the central nervous system. Stimulation of the locus coeruleus in the pons increases anxiety, and selective serotonin-reuptake inhibitors (**SSRIs**), tricyclic antidepressants, monoamine oxidase inhibitors (**MAOIs**), and benzodiazepines all decrease the firing rates of neurones in this area.

Diagnosis and clinical manifestations

Patients with depression or anxiety initially present with physical complaints 50 to 70 per cent of the time. They often complain of vague symptoms or report multiple somatic symptoms in a variety of anatomical locations, or experience greater dysfunction than can be attributed to their known medical disorders. Patients with panic disorder often selectively focus on the somatic components of anxiety, such as chest pain or palpitations, attributing their increased anxiety and tension to the frightening nature of somatic symptoms.

When depression or anxiety are suspected, simple screening questions may be helpful: 'Are you feeling sad, blue, or depressed?' 'Have you lost interest and pleasure in most things you usually enjoy?' 'Do you have sudden episodes or attacks where your heart beats fast, your chest is tight, it feels hard to breathe and you feel shaky?' The physician should enquire about the patient's explanatory model for his or her symptoms: 'Why do you think you get these symptoms?'

History is the single best diagnostic tool. The physician should elicit the patient's concerns and fears, current life situation, family and other support systems, and concurrent medical problems. A family history of psychiatric problems (depression, anxiety, substance use disorders) should also be obtained. Screening questions about a childhood or adult history of physical or sexual abuse or domestic violence are important and can be woven into the usual questions about family medical history. Adverse childhood and adult traumatic experiences are associated with an increased risk of anxiety and affective disorders in adulthood. Over half of the depressed women in one primary care study reported experiencing physical abuse as adults.

The physician should explain carefully that anxiety and depression reflect a biological predisposition that is provoked during a period of life stress, but major depression is not the uniform outcome of any stressful event. Risk factors that predispose patients to anxiety or depression include a personal or family history of depression or substance abuse, serious medical illness, lack of social support, or a history of early childhood trauma or neglect. Risk factors that can precipitate an acute episode or perpetuate the disorder include poor physical health, divorce, poor interpersonal relationships, illness or death in a family member, low socioeconomic status, or a stressful work situation.

Major depressive episodes last at least 2 weeks and are characterized by at least five of nine criteria, including at least one of the two primary criteria of depressed mood and loss of interest or pleasure in nearly all activities (Table 2). There are two to three times as many people with 'minor' depressive symptoms that fall short of major depression criteria: these have a higher rate of spontaneous recovery.

Physical symptoms are predictive of a good response to treatment: patients with middle insomnia (awakening between 0200 and 0400 h) or a diurnal variation in mood are more likely to respond to antidepressant medications.

Several dimensions of depression severity should be assessed: the frequency and chronicity of depressive symptoms, the impact of depression on the patient's ability to function, the potential for suicide, and the presence of psychotic or manic symptoms. Dysthymic disorder is characterized by a chronically depressed mood that occurs most of the day, more days than not, for at least 2 years. Dysthymic disorder is associated with many of features of major depression, but differs in its onset, duration, persistence, and severity of symptoms. Dysthymia is associated with impaired functioning and may not remit spontaneously.

Patients with anxiety have cognitive, affective, and somatic symptoms. The key feature distinguishing panic disorder from other anxiety disorders is the episodic nature of the attacks. Panic attacks are characterized by the sudden onset of intense apprehension, fear or terror, and by the abrupt development of at least four of the symptoms listed in Table 3, reaching a peak within 10 min. Panic disorder is diagnosed when attacks are recurrent, produce persistent fear, or become significantly disruptive to the patient's life.

Laboratory studies

An adequate diagnostic work-up for anxiety or depression may include a complete blood screen, urinalysis, and routine laboratory tests of renal and liver function. Selected patients should receive an electrocardiogram or chest radiograph. Thyroid function studies are recommended in perimenopausal or postmenopausal women. Given the sizeable differential diagnosis for a patient presenting with anxiety or depression, the extent of work-up must be tailored for each case.

Suicide

The risk of suicide should be evaluated in all patients with depressive and anxiety disorders. The risk of suicide attempts and suicidal ideation more than doubles in depressed patients with comorbid anxiety or physical illness. Other risk factors for suicide include gender (elderly White males are at highest risk), alcoholism, severe medical illness, psychosis, and lack of social support. The topic of suicidal ideation can be approached gradually with a non-specific question such as, 'Do you ever feel so discouraged that life does not seem worth living?'

Asking about suicide will not increase a patient's risk. Enquiries about suicide can reassure the patient and enable the physician and patient together to make a plan to prevent suicide, including deciding together whether emergency psychiatric consultation or hospitalization is necessary. Physicians should consider using a 'no harm contract': meaning that patients are simply asked to contract with the physician in writing that they will contact the physician if they think that they are losing control of a suicidal impulse. Although data on the effectiveness of this technique are not available, it seems sensible and is standard clinical practice in many centres.

For further discussion of these issues see Chapter 26.5.2.

Treatment

Whilst depression and anxiety are usually recurrent or chronic disorders, their clinical course can be markedly improved with timely, evidence-based treatments. Many patients can be treated successfully by primary care or general physicians. A meta-analysis of 28 randomized controlled trials found an overall efficacy rate of 54 to 65 per cent for the treatment of depression in the primary care setting, which is comparable to the response of patients seen by psychiatrists.

Physicians need to educate patients about the nature of depression or anxiety and how symptoms can be managed. They should explore background problems, define treatment goals, and dispel negative perceptions (for example, that antidepressant therapy is addictive). Patients should be reassured that they are not 'crazy', nor are their symptoms a manifestation of their own failure or shortcomings. There are several important educational points to cover. Depression and anxiety are quite common and are associated with important physiological changes. With proper treatment, these disorders almost always improve or remit; but relapses and recurrences can occur, so follow-up is essential. Physicians should enquire about the patient's concerns regarding a diagnosis of depression or anxiety, and also their

worries about medical disorders. Raising the possibility of referral to a psychiatrist or other mental health professional early may make it easier to accept later.

Patients should be educated about the types of available treatments. Identifying the patient's desires and goals for treatment can provide a focus for the management of problems that can otherwise seem to be poorly defined and overwhelming. This can also increase patients' sense of participation in their care. Successful disease management programmes developed for asthma and diabetes that have emphasized educating patients to be partners in their medical care have resulted in significant improvements in adherence to treatment and in outcomes. Similar approaches appear to be successful in managing those with depression and panic disorder.

Pharmacotherapy

The decision to prescribe antidepressant therapy should be based on the number of symptoms, the level of dysfunction, and previous episodes of depression or anxiety. Before initiating antidepressant therapy, the physician should educate the patient regarding potential side-effects, the need to take medication regularly, and the usual time period and course to recovery.

The different classes of antidepressant drugs show virtually equivalent efficacy in the treatment of outpatients with major depressive disorder. The consensus on the pharmacotherapy of panic disorder also described equivalent efficacy of tricyclic antidepressants (**TCAs**), SSRIs, MAOIs, and high-potency benzodiazepines. For other forms of anxiety (social phobia and post-traumatic stress disorder), there is more evidence of the efficacy of SSRIs and MAOIs compared to TCAs.

The choice of medication should therefore be made on issues other than efficacy, such as side-effects, cost, adherence, and physician familiarity and comfort with prescribing particular agents. Primary care providers should become familiar with one or two medications with minimal side-effects from each of the major classes of antidepressants: TCAs, heterocyclics, and SSRIs. In each case the aim is to optimize treatment benefit and lower risk. Factors to be considered include the possibility of side-effects, history of response or a failure to respond, possible drug interactions, the presence of other psychiatric or medical conditions, familial response to a specific agent, and patient age.

The SSRIs have become the first-line treatment for major depression, dysthymia, and panic disorder, primarily because their improved side-effect profiles are associated with improved adherence to treatment. In the primary care setting, patients are significantly more likely to discontinue TCAs than SSRIs, those started on an SSRI being 7.5 times more likely to have a duration and an average dose of medication consistent with treatment guideline recommendations. The initial prescription of fluoxetine results in fewer side-effects, a lower rate of medication switching, and no difference in clinical outcomes, quality of life outcomes, or overall treatment costs when compared to TCAs. Although SSRIs are more expensive, total treatment costs per depressive episode are similar for patients treated with SSRI or TCA medication.

Randomized controlled trials have failed to show efficacy for antidepressant medication in patients with minor depression, largely because the placebo response rate was so high. Watchful waiting is appropriate for these patients, although the physician should recognize that they are at a higher risk of developing a major depressive episode.

Classes of medications

For a discussion of the pharmacology, side-effects, and interactions of tricyclic antidepressants, SSRIs, MAOIs, and other agents used in the treatment of anxiety and depression, see Section 26.6.1.

Selective serotonin-reuptake inhibitors (SSRIs)

The advent of the SSRIs (fluoxetine, paroxetine, sertraline, fluvoxamine, citalopram) and the newer atypical antidepressants (amfebutanone (buproprion), nefazodone, and venlafaxine) has significantly increased the number of patients receiving pharmacological treatment for depression and anxiety in a primary care setting. These new agents have fewer adverse side-effects and are safer for treating elderly patients and those with comorbid medical illnesses. SSRIs are also much safer than the tricyclic antidepressants if taken in overdose. They do not cause postural hypotension or cardiac conduction delay.

These agents have the advantage that the starting dose may also be an effective treating dose. This is most clearly the case for fluoxetine (20 mg), but may also be true for paroxetine (20 mg), sertraline (50 mg), and citalopram (20 mg). The frail elderly and those with liver disease require smaller starting doses (generally one-half of the recommended starting dose). Fluoxetine has the longest half-life (24–27 h) and a long-acting active metabolite (half-life of 7 days). This medication can be taken every other day, and the doses can eventually be given once or twice weekly to allow for smooth tapering when the drug is being withdrawn. The other SSRIs have half-lives of about 24 h and have no active metabolites with longer half-lives. This allows once-daily dosing and rapid washout. Amfebutanone (buproprion), venlafaxine, and nefazodone have shorter half-lives (14 h), so must be given at least twice a day; sustained release forms of amfebutanone and venlafaxine can be taken once daily.

Table 4 describes common side-effects of these medications: those such as anxiety, insomnia, nausea, headache, and agitation occur in fewer than 20 per cent of patients. When they do occur, they are usually mild and may respond to a reduction in the dose of medication. Sexual dysfunction (decreased libido and anorgasmia) may occur in up to one-third of patients treated with SSRIs; the addition of amfebutanone (buproprion) (75 to 150 mg a day in divided doses) or buspirone (20–40 mg) may alleviate these sexual side-effects. Strategies for the management of other common side-effects are listed in Table 5.

Patients with anxiety disorders are especially sensitive to the SSRI side-effects of jitteriness, restlessness, agitation, and insomnia. Low doses should be prescribed initially (5–10 mg paroxetine; 12.5–25 mg of sertraline; 25 mg fluvoxamine; 5 mg fluoxetine) for approximately 1 week, then gradually increasing to full therapeutic doses. To completely alleviate panic attacks in most cases requires: 20 to 50 mg paroxetine, citalopram, or fluoxetine; 50 to 200 mg of sertraline; and 100 to 300 mg of fluvoxamine. Similar schedules and dosages are used to treat patients with social phobia, post-traumatic stress disorder, and generalized anxiety disorder.

The SSRIs (but not amfebutanone (buproprion)) have all been shown to inhibit the cytochrome P-450 system in the liver, thus potentially leading to drug interactions (see Chapter 26.6.1). Amfebutanone is an effective antidepressant that may be especially useful because it does not cause the sexual side-effects common to the SSRIs, but it has been associated with a 1.5 per cent increase in prevalence of seizures compared to other antidepressants. The medication should not be given to individuals at risk for seizures (those with a history of seizures or head injury), and should never be administered in a single dose greater than 150 mg. Patients with a history of bulimia also have an increased risk of seizures on amfebutanone, and should not be prescribed this medication. The starting dose of 75 mg twice per day should be increased every week to achieve a therapeutic level of between 300 and 450 mg/day.

Tricyclic antidepressants

The heterocyclic medications include the tricyclic antidepressants and several other agents that are similar in structure, including maprotiline, amoxapine, and trazodone. These medications are similar in their side-effects and dosing strategies (see Chapter 26.6.1).

Low starting doses are required, with a gradual increase to a therapeutic level. For treating depression, generalized anxiety disorder, or panic disorder, physicians should begin at 10 to 25 of imipramine, gradually increasing the dose by 10 to 25 mg every 4 or 5 days. The ultimate dosage is variable, with some patient responding at a low dosage (50–100 mg), and others needing up to 300 mg.

Plasma levels can be measured, but these function primarily as crude indicators of whether or not the patient is taking the medication. Clinicians should treat the patient, and not the blood level, since many with high or low blood levels may do very well clinically. Nortriptyline is an exception, as it has a therapeutic window (50–150 ng/ml): levels below or above this are less likely to lead to remission of depression.

Other agents

For the small subgroup of patients with anxiety disorders who do not tolerate SSRIs or TCAs, high-potency benzodiazepines may be an effective second-line treatment. Alprazolam, lorazepam, and clonazepam have all been shown to be more effective than placebo for the treatment of panic disorder. Patients can be started at 0.25 mg clonazepam three times per day, with a gradual increase by 0.25 mg every 2 to 3 days until the attacks stop. Symptoms of generalized anxiety disorder can

be alleviated in most cases with a clonazepam dose between 0.25 to 0.5 mg twice daily and 1 mg two or three times daily, or lorazepam 0.5 to 1 mg three times daily. These agents are best used in conjunction with an SSRI or TCA at the beginning of treatment. After 6 to 8 weeks, when the antidepressant begins to have its optimal effects in treating anxiety symptoms, the benzodiazepine can usually be tapered with a 10 per cent dose reduction per week.

Buspirone is an azapirone with affinity for 5-HT1A and dopamine receptors that has been approved for the treatment of patients with generalized anxiety disorder. It is non-sedating, and there are no withdrawal symptoms with abrupt discontinuation. There are also no synergistic effects with alcohol. Buspirone is typically effective at doses of 30 to 60 mg, divided two or three times daily. Common side-effects include dizziness, nausea, headache, and nervousness. These can be reduced by using lower starting doses of 5 mg two or three times daily and advancing as tolerated.

Beta-adrenergic blockers may be useful for treating performance anxiety. Propranolol is used at doses of 10 to 80 mg per day (ideally taken 2 h before the anticipated exposure). Atenolol at 30 to 100 mg may be preferred since it has fewer CNS side-effects (exacerbating depression).

MAOIs are potentially the most effective class of medication for panic or certain types of depression. However, their regular use in primary care or by general physicians is precluded by the lack of familiarity with these agents, and the potential for hypertensive crisis that can ensue from not following a low-tyramine diet or taking an over-the-counter stimulant medication (like pseudoephedrine).

St John's Wort

Clinicians need to ask patients in a routine and non-judgemental manner about their use of alternative treatments, and should know enough about the more common ones to assess for deleterious effects or interactions. St John's wort has been widely used in Europe. A recent meta-analysis found it to be more effective than placebo and of similar effectiveness to low-dose TCAs in the short-term treatment of mild depression. However, a recent study comparing treatment with an SSRI versus St John's wort found that St John's wort was less effective in treating major depression. Gastrointestinal effects, including nausea, pain, loss of appetite, and diarrhoea occurred at a rate of 0.55 per cent in a German study of 3250 patients taking 300 mg three times a day. It may cause a sunburn-like reaction, mucosal inflammation, pruritis, and can lead to significant depression of the blood level of ciclosporin in organ transplant recipients, even leading to rejection.

Special issues in pregnancy

Mild postpartum 'blues' occur in 30 to 75 per cent of women immediately after delivery. Symptoms include labile mood, tearfulness, irritability, anxiety, and sleep and appetite disturbances lasting 4 to 10 days. If physical symptoms and depressed mood persist for 2 weeks, patients should be evaluated for postpartum major depression, which is relatively common, having a prevalence rate of approximately 10 per cent. Symptoms usually begin during the third trimester. A history of depression, limited social support, marital conflict, and ambivalence about the pregnancy increase the risk of depression during pregnancy and in the postpartum period.

Pharmacotherapy for depression during pregnancy requires an assessment of the risks and benefits of treatment for both mother and fetus. The risks of not treating depression may include suicide, poor maternal and fetal nutrition, an obstetric complication, and the continuation of depression into the postpartum period, with effects on mother and child bonding. Psychotherapy can be helpful in resolving interpersonal and psychosocial conflicts without exposing the mother or fetus to medications. Although psychotherapy is the first-line treatment for mild to moderate depression during pregnancy or after the birth of a child, antidepressant treatment may be warranted in severe major depression.

SSRIs are considered the first-line pharmacotherapy for depression during pregnancy. The Fluoxetine Pregnancy Database, based on 1103 prospectively reported pregnancies, reported rates of fetal malformation and spontaneous abortion similar to rates in pregnancies not exposed to fluoxetine. A prospective, controlled multicentre study to assess fetal safety and risks of the SSRIs (fluvoxamine, paroxetine, and sertraline) found that exposure to SSRIs was not associated with an increased risk for major malformations or higher rates of miscarriage, stillbirth, or prematurity. There do not appear to be adverse effects on global intelligence quotient, or language or behavioural development in preschool children exposed *in utero* to either tricyclic antidepressants or fluoxetine.

Data regarding the excretion of antidepressants in breast milk are limited. The American Academy of Pediatrics Committee on Drugs concluded: 'antidepressants are drugs whose effect on nursing infants is unknown but may be of concern'. Several studies have shown only small amounts of SSRIs in breast milk and infant serum samples. Children have been followed through their enrolment in kindergarten with no evidence of negative effects on global intelligence, language, or behaviour.

Monitoring

Regular visits are essential for patients with depression and anxiety. Brief visits every 2 weeks are usually indicated during the first 6 weeks of treatment to evaluate the dosage and side-effects of medications, and any changes in the patient's condition. With appropriate counselling beforehand and a regular discussion of side-effects, patients are more likely to adhere to a full medication trial, and this can also be encouraged by telephone monitoring. The physician should record the main symptoms that the patient presents at the beginning of treatment, and review these at each follow-up visit. After they have been on a therapeutic dose of medication for 4 weeks, the treatment response should be evaluated. Simply asking the patient to note the degree of progress on a scale of 1 to 5 can be helpful in assessing either an improvement or worsening of the target symptoms. Empirically validated, self-rating scales, such as the Patient Health Questionnaire, administered at initial diagnosis and at follow-up, can be useful in assessing treatment response. A 25 per cent or greater reduction in baseline symptoms constitutes a reasonable basis for extending the initial treatment. If there has been no response or only a partial response in symptoms, the dose of the medication should be increased to the upper therapeutic range.

Some two-thirds of patients with major depression respond to an antidepressant within 3 weeks after reaching a therapeutic plasma level. This success rate can be increased to 90 per cent by switching initial non-responders to another class of antidepressant, or by augmentation strategies using additional medication or psychotherapy. However, if a patient fails to improve adequately with first-line therapy, the diagnosis and the treatment plan must be reassessed. There may be unrecognized comorbid anxiety or substance abuse, and treatment failure is commonly due to inadequate dosing or lack of adherence to medication.

Once the patient is stabilized on a medication (usually 6–12 weeks), monthly visits are important for support. If chronic prophylactic treatment is necessary, visits every 3 months are usually appropriate. Consensus statements recommend a treatment period of between 6 and 9 months for a major depressive episode. Pharmacotherapy should be discontinued slowly over a period of 7 to 21 days. If tapered too quickly, almost all antidepressants (except fluoxetine) can produce withdrawal syndromes that include sleep disturbance, mood changes, anxiety, sensory disturbance, malaise, muscle aches, vertigo, sweating, fatigue, and gastrointestinal upset.

Patients who have remitted during the acute phase of pharmacotherapy for anxiety and depressive disorders remain at substantial risk of relapse during the subsequent 12 months: 37.1 per cent of depressed primary care patients experience further depressive symptoms during this time. There are three main risk factors are associated with relapse: (1) persistence of subthreshold depressive symptoms 7 months after the initiation of antidepressant therapy; (2) history of three or more previous episodes of major depression; and (3) chronic mood symptoms for more than 2 years. Patients with two of these risk factors are approximately three times more likely to relapse than those without. Between 50 and 60 per cent of patients who have had a single major depressive episode will have a second one; 70 per cent of those who have two episodes will have a third; and 90 per cent of those who have had three episodes will have a fourth. For patients with three or more depressive episodes or dysthymia and major depression, the AHCPR guidelines recommend treatment for 2 years or more.

The optimal duration of treatment for anxiety disorders has not been as well established by controlled studies. Treatment is recommended for 6 to 9 months after the first episode. Maintenance therapy should be considered for those with either a chronic history since adolescence, or three or more recurrences. Panic disorder is a recurrent or chronic disease in the majority of cases. In a review of 16 studies, most patients had improvement in symptoms with treatment, but few experienced complete resolution. As panic disorder progresses, attacks become more frequent, and are preceded by anticipatory anxiety. Patients may begin to associate environmental events with anxiety, leading to avoidance behaviour. The disorder may culminate in agoraphobia: being afraid to leave the house because of the association with panic attacks.

Electroconvulsive therapy (ECT)

ECT is still the most effective treatment available for the treatment of depression: it can be life-saving in some cases, and in the frail elderly it may be safer than antidepressants. Some reversible short-term memory loss is a common side-effect, but this reverts to normal in almost all cases. Patients with recurrent depression

who receive effective ECT treatment should be treated with prophylactic medication or maintenance ECT once the acute course of the treatment has finished.

Psychotherapy

Randomized controlled trials support the efficacy of psychosocial interventions provided to ambulatory medical patients with psychiatric disorders. Problem-solving skills and other behavioural techniques, most of which can be provided as simple self-help materials, are part of the psychosocial support that general physicians can provide in a disease management programme for depressive or anxiety disorders. As in all chronic illnesses, lifestyle changes should be reinforced: good sleep habits, adequate exercise, and minimization of caffeine and alcohol intake.

There are three short-term psychotherapies used to specifically target the symptoms of major depression: cognitive–behavioural therapy (**CBT**), interpersonal psychotherapy (**IPT**), and problem-solving therapy (**PST**). CBT is directed at the negative and distorted thinking patterns and subsequent maladaptive behaviours that often accompany depressive episodes. IPT helps the patient learn to manage the current interpersonal relationship difficulties that are sometimes related to the development and maintenance of depressive symptoms. PST helps activate patients to break down global problems to smaller units that they can begin to attempt to solve.

Patients with major depression treated with CBT and IPT experience as much relief from symptoms by 16 weeks and those with PST by 11 weeks as those taking medication alone. Given these findings, physicians need not rely as heavily on drug treatments as they typically have, and should consider psychosocial interventions if the patient prefers them and they are available.

In a meta-analysis of studies on panic disorder, psychological coping strategies involving relaxation training, cognitive restructuring, and exposure worked comparably with both antidepressant and benzodiazepine medications. The combined somatic exposure and cognitive therapy used by Barlow and Clark helps patients confront and alter maladaptive cognitions (for example, thoughts of a heart attack or stroke when experiencing a rapid heartbeat). A meta-analysis of the treatment of generalized anxiety disorder concluded that CBT is more efficacious than control treatments and at least equal in efficacy to anxiolytics.

Specialty referral

Referral to a psychiatrist should be considered when the physician is confused about the primary diagnosis, as in distinguishing an anxiety disorder from depression with anxiety or substance abuse disorder. Referrals should also be made when adequate treatment does not lead to an improvement in symptoms within 10 to 12 weeks**,** or several medication trials have failed. Patients with suicidal behaviour require specialty care (see Chapter 26.5.2).

Other disorders

Generalized anxiety disorder is characterized by constant, non-episodic anxiety that affects the patient for more than 6 months and interferes with normal function. In community samples, lifetime prevalence rates are 5.1 per cent. Medical problems such as hyperthyroidism should be ruled out as the cause of the symptoms of motor tension and autonomic hyperactivity. This disorder can be treated effectively with SSRIs, TCAs, buspirone, and benzodiazepines; the first three classes should be tried first, given their lower potential for habituation and dependence.

Post-traumatic stress disorder (**PTSD**) is a syndrome that can occur after a person experiences trauma outside the range of normal human experience (accidents, abuse, rape, and natural disasters). The lifetime prevalence in the general population is between 1 and 9 per cent. The most frequent traumas in civilian cases involve adult domestic violence and childhood abuse. Symptoms include flashbacks, nightmares, and/or severe distress (or numbness) to stimuli that concretely or symbolically resemble the event—or to many stimuli (generalization), hypervigilance, or other persisting signs and symptoms of autonomic arousal, and secondary depression. Cases of PTSD are occasionally seen in primary care or by general physicians, with these patients often having a combination of symptoms of PTSD, panic, major depression, and an increased risk of alcohol and drug abuse. The alcohol and drugs are often taken to try to blunt excessive anxiety symptoms. Treatment of PTSD usually includes medication and psychotherapy, recent trials showed SSRIs to be more effective than placebo.

Specific phobias are characterized by episodic anxiety in response to a specific precipitant: intense excessive fear makes patients avoid the situation. Examples of stimuli for simple phobias are airplanes, heights, and insects, although most patients with simple phobias do not seek care for the condition. Social phobia is the fear of humiliation or failure in public situations (such as public speaking, meeting strangers, and eating in restaurants), and many with this condition also have major depression (58.3 per cent of cases of social phobia), or other anxiety disorders in (panic disorder, 27.8 per cent; generalized anxiety disorder, 30.6 per cent). Since social phobia is often present from childhood or adolescence, patients often consider the marked social anxiety and avoidance as part of their personality. Recent studies have shown that SSRIs are more effective than placebo in the treatment of social phobia; effective cognitive–behavioural techniques have also been developed. Propranolol (10–80 mg) may be useful for treating non-generalized social phobia that occurs in one situation, such as public speaking.

Obsessive–compulsive disorder (**OCD**) is characterized by regular intrusive thoughts or obsessions about aggression, sex, religion, theft, or loss—or other covering mental rituals such as counting objects or letters. Patients may also have persistent rituals or compulsions that are so frequent or complex that they interfere with normal function. They experience these obsessions and compulsions as intrusive, silly, and upsetting. Most patients with OCD have experienced major depressive episodes. Pharmacological treatment is indicated: although clomipramine had been the first-line treatment for many years, recent randomized controlled trials have shown SSRIs to be as effective. Treatment with SSRIs should continue for at least 10 weeks before it is considered ineffective. Randomized trials have shown behavioural treatments to be effective; these focus on exposure to feared activities and prevention of compulsive responses.

Conclusions

The healthcare problems of depression and anxiety are as common, and as treatable, as asthma and hypertension. If a patient does not improve with initial management, he or she should be referred to a mental health specialist. Routine follow-up is essential: these disorders follow a relapsing and remitting course in 70 to 80 per cent of patients, and become chronic disorders in up to 20 per cent. Lifelong monitoring of symptoms with the patient and his or her family is required.

Further reading

Ballenger JC, *et al*. (1998). Consensus statement on panic disorder from the International Consensus Group on Depression and Anxiety. *Journal of Clinical Psychiatry* **59**(Suppl. 8), 47–54.

Ballenger JC, *et al*. (1999). Consensus statement on the primary care management of depression from the International Consensus Group on Depression and Anxiety. *Journal of Clinical Psychiatry* **60**(Suppl. 7), 54–61.

Brown C, Schulberg HC (1995). The efficacy of psychosocial treatments in primary care: a review of randomized clinical trials. *General Hospital Psychiatry* **17**, 414–24.

Clinical Practice Guideline Number 5 (1993). Treatment of major depression. *Depression in primary care*, Vol. 2, AHCPR publication 93–0551. US Dept Health Human Services, Agency for Health Care Policy and Research, Rockville MD.

Clum G, Surls R (1993). A meta-analysis of treatments for panic disorder. *Journal of Consulting and Clinical Psychology* **61**, 317–26.

Edlund MJ (1990). The economics of anxiety. *Psychiatry in Medicine* **8**(2), 15–26.

Katerndahl DA, Realini JP (1995). Where do panic attack sufferers seek care? *Journal of Family Practice* **40**, 237–43.

Katon W, *et al*. (1990). Distressed high utilizers of medical care: DSM-III diagnoses and treatment needs. *General Hospital Psychiatry* **12**, 355–62.

Katon W, *et al*. (1992). Adequacy and duration of antidepressant treatment in primary care. *Medical Care* **30**, 67–76.

Kessler RC, *et al*. (1994). Lifetime and 12-month prevalence of DSM-IIIR psychiatric disorders in the United States: results of the National Comorbidity Survey. *Archives of General Psychiatry* **51**, 8–19.

Kim HL, *et al*. St. John's wort for depression: a meta-analysis of well-defined clinical trials. *Journal of Nervous and Mental Disease* **187**(9), 532–8.

Kulin NA, *et al*. (1998). Pregnancy outcome following maternal use of new selective serotonin reuptake inhibitors: a prospective controlled multi-center study. *Journal of the American Medical*

Association **279**, 609–10.

Lin EH, *et al*. (1998). Relapse of depression in primary care: rate and clinical predictors. *Archives of Family Medicine* **7**, 443–9.

Mynors-Wallis LM, *et al*. (1995). Randomized controlled trial comparing problem solving treatment with amitriptyline and placebo for major depression in primary care. *British Medical Journal* **310**, 441–5.

Ormel J, *et al*. (1994). Common mental disorders and disability across cultures: results from the WHO Collaborative Study on Psychological Problems in General Health Care. *Journal of the American Medical Association* **272**, 1741–8.

Roy-Byrne P, *et al*. (1998). Pharmacotherapy of panic disorder: proposed guidelines for the family physician. *Journal of the American Board of Family Practice* **11**(4), 282–90.

Schulberg HC, Katon W (1998). Treating major depression in primary care practice: an update of Agency for Health Care Policy and Research Practice Guidelines. *Archives of General Psychiatry* **55**(12), 1121–7.

Simon GE, vonKorff M, Barlow W (1995). Health care costs of primary care patients with recognized depression. *Archives of General Psychiatry* **52**, 850–6.

Spitzer RL, *et al*. (1995). Health-related quality of life in primary care patients with mental disorders: results from the PRIME-MD 1000 study. *Journal of the American Medical Association* **274**, 1511–17.

Stein MB, *et al*. (1999). Social phobia in the primary care medical setting. *Journal of Family Practice* **48**(7), 514–19.

Wells KB, *et al*. (1989). The functioning and well-being of depressed patients: results from the Medical Outcomes Study. *Journal of the American Medical Association* **262**, 914–19.

26.5.5Eating disorders

Christopher G. Fairburn

Introduction

The term 'eating disorder' refers to a persistent and severe disturbance of eating habits which results in impaired physical health or psychosocial functioning. The disturbance should not be secondary to any general medical disorder or to any other psychiatric condition. Anorexia nervosa and bulimia nervosa are the best characterized of the eating disorders. Anorexia nervosa has been recognized for many years, with physicians in the 19th century providing particularly good accounts of the disorder. By contrast, bulimia nervosa was first described in 1979. Anorexia nervosa and bulimia nervosa are closely related, in that they have many features in common and some patients move from one disorder to the other. Other eating disorders are encountered, most of which appear to be variants of anorexia nervosa or bulimia nervosa. These disorders are classified as 'atypical eating disorders'.

Eating disorders should not be confused with obesity—a general medical condition in which there is excess body fat (see Chapter 10.5). Eating disorders may coexist with obesity, although in practice most people with an eating disorder have a normal or low body weight.

Anorexia nervosa and bulimia nervosa

Definition of anorexia nervosa

To make a diagnosis of anorexia nervosa, three features need to be present:

1. *A characteristic set of attitudes to shape and weight in which self-worth is judged largely, or even exclusively, in terms of shape and weight*—Whereas most people evaluate themselves on the basis of their perceived performance in a variety of domains (such as their relationships, work, sport, artistic ability, etc.), in anorexia nervosa shape and weight dominate self-evaluation. This overevaluation of shape and weight may be regarded as the core psychopathology of the disorder since most other features appear to be secondary to it.
2. *The active maintenance of an unduly low body weight*—This is the principal behavioural expression of the extreme concerns about shape and weight. For diagnostic purposes, an 'unduly low body weight' may be defined as a weight at least 15 per cent below that expected for the person's age, height, and sex, or as a body mass index below 17.5.
3. *Amenorrhoea* (in postmenarchal females who are not taking an oral contraceptive)—Although required for official diagnostic purposes, the symptom of amenorrhoea has little discriminatory value. A clinical diagnosis of anorexia nervosa may be made in its absence.

Definition of bulimia nervosa

Bulimia nervosa also has three necessary diagnostic features:

1. *The same core psychopathology as that seen in anorexia nervosa, with self-worth being judged in terms of shape and weight*.
2. *Repeated episodes of uncontrolled overeating*—These bulimic episodes (commonly referred to as 'binges') involve the consumption of unusually large amounts of food, given the circumstances, and a sense of loss of control at the time. It is this latter feature that distinguishes binge-eating from simple overeating.
3. *The regular practice of extreme weight-control behaviour*—People with bulimia nervosa diet intensely, they may overexercise, and many engage in self-induced vomiting and the misuse of laxatives or diuretics. This behaviour is an expression of their extreme concerns about shape and weight, although it is further encouraged by the episodes of loss of control over eating.

It should be noted that there is no weight criterion for bulimia nervosa. In practice, body weight is generally unremarkable. This is because the overeating and weight-control behaviour tend to cancel each other out. There are some patients with anorexia nervosa who have binges like those seen in patients with bulimia nervosa, and who could therefore be eligible for both diagnoses. In practice, both diagnoses are not given: instead, the convention is that the diagnosis of anorexia nervosa takes precedence over that of bulimia nervosa.

Epidemiology

Anorexia nervosa is largely confined to females aged between 10 and 30 years and to Western societies. Estimates of the incidence of the disorder range from 0.10 to 8.2 per l00 000 population per annum, the higher figure being likely to be the more accurate. Estimates of its prevalence amongst adolescent girls, the group most at risk, range from 0.2 per cent to 0.8 per cent. The disorder is uncommon among men with about 10 per cent of patients being male. The social class distribution seems to be uneven with an over-representation of cases from upper socioeconomic groups, but the extent to which this is a result of referral bias is not known. It has been suggested that the disorder has become more common over recent decades, but other explanations for the apparent increase cannot be ruled out. These include alterations in diagnostic practice, better detection, increased help-seeking, and changes in the demographic structure of the population. Irrespective of whether the incidence has increased, anorexia nervosa is a major cause for concern. The disorder has one of the highest mortality rates of any psychiatric illness and it can be extremely difficult to treat.

Bulimia nervosa affects a slightly older age group than anorexia nervosa, with most cases being in their twenties. It also appears to have a broader social class distribution. The disorder is considerably more common than anorexia nervosa, the prevalence rate among young women (15 to 40 years) being between 1 per cent and 2 per cent, most of whom are not in treatment. Whilst there are no satisfactory data on the incidence of bulimia nervosa, it seems that the disorder has become much more common since the early 1970s. From being viewed as an unusual variant of anorexia nervosa, bulimia nervosa is now the most common eating disorder seen in clinical practice. It rarely occurs among men and, like anorexia nervosa, it appears to be largely confined to Western societies.

General clinical features

Anorexia nervosa

In anorexia nervosa the overevaluation of shape and weight results in a pursuit of weight loss and thinness. To the extent that this is successful, the disorder is 'egosyntonic'; that is, it is not viewed by the person as a problem—indeed, it has been noted that patients view it more as an achievement than as an affliction. As a

consequence, there is little motivation to change or seek help.

The low weight is primarily the result of the strict and self-imposed restriction of food intake. Typically, the range of foods eaten is limited with those foods viewed as fattening being avoided. Except in long-standing cases, appetite persists and for this reason the term 'anorexia' is misleading. Restlessness and frequent intense exercising are common, and contribute to the low weight. Self-induced vomiting and the misuse of laxative and diuretics are practised by a subgroup, and some have occasional binges.

Driving the disturbed eating habits is the so-called 'body image disturbance'. This has several aspects. There is the core psychopathology, already mentioned, in which self-worth is judged largely in terms of shape and weight. In addition, there is sometimes a perceptual component involving overestimation of body size. This may be a consequence of the patients' frequent checking of their own body, and in some this becomes so distressing that it is abandoned, although the concerns about shape and weight persist. In almost every case there is preoccupation with thoughts about food, eating, shape, and weight, which may be expressed as an avid interest in cooking, nutrition, fitness, and health.

Depression, anxiety, irritability, lability of mood, and obsessional features are all common concomitants of anorexia nervosa. Typically, they get worse as weight is lost and improve with weight regain. Outside interests also decline and there may be marked social withdrawal. Suicidal thoughts may be present and the risk of suicide should always be kept in mind when assessing patients.

Bulimia nervosa

The clinical features of bulimia nervosa are similar to those of anorexia nervosa. There is the same over-evaluation of shape and weight, and this also leads to body checking and extreme methods of weight control. The main differences lie in the frequent bulimic episodes and the fact that body weight is generally unremarkable. The other important difference is that, as a consequence of the loss of control over eating, the disorder is 'egodystonic'; that is, it is viewed by the patient as a problem. This makes treatment much easier.

The binges involve the consumption of sizeable amounts on food (typically over 2000 kcal per episode) and they are a source of distress and shame. Typically, they are kept secret. In most cases, they are followed by self-induced vomiting or the taking of laxatives or diuretics, although there is a subgroup of patients who do not 'purge'. Between the binges, food intake is severely restricted. Depressive and anxiety symptoms are prominent in bulimia nervosa, more so than in anorexia nervosa, and some patients have problems with alcohol or drug misuse.

Physical and laboratory features

The physical abnormalities seen in anorexia nervosa have been the subject of much interest. They used to be viewed as evidence of a primary pituitary or hypothalamic disorder, but it is now thought that they are secondary to the disturbed eating habits and the patient's state of starvation. The physical abnormalities encountered in bulimia nervosa resemble those seen in anorexia nervosa except that they are less severe.

Symptoms and signs in anorexia nervosa

Many patients with anorexia nervosa have no physical complaints. However, systematic enquiry often reveals a heightened sensitivity to the cold and a variety of gastrointestinal symptoms such as constipation, fullness after eating, bloatedness, and vague abdominal pains. Other symptoms include restlessness, low sexual appetite, and poor sleep with early morning wakening. In females who are not taking an oral contraceptive, amenorrhoea is, by definition, present. Occasional patients complain of infertility.

On examination, the degree of emaciation may be striking. Growth may be stunted in those with a prepubertal onset and there may be failure of breast development. Unlike patients with hypopituitarism, axillary and pubic hair are preserved and there is no breast atrophy. A fine downy hair (lanugo) is commonly present on the back, arms, and side of the face. Typically, the skin is dry and the hands and feet are cold. There may be hypothermia. Blood pressure and pulse are low and some patients have dependent oedema.

Abnormalities on investigation in anorexia nervosa

Endocrine abnormalities

Many of the abnormalities encountered have been reproduced in studies of the physiological effects of dieting and starvation, and are reversed by the restoration of healthy eating habits and a normal weight. Luteinizing hormone-releasing hormone (**LHRH**) secretion is impaired and, as a result, levels of luteinizing hormone (**LH**), follicle-stimulating hormone (**FSH**), and oestradiol are low. There is an immature pattern of LH release. The LH response to LHRH is reduced, but the FSH response is normal or exaggerated.

Hypothalamic disturbance is also evident in a delayed thyroid-stimulating hormone (**TSH**) response to thyrotropin-releasing hormone (**TRH**). In addition, there is reduced peripheral conversion of thyroxine (T4) to triiodothyronine (T3), and an increased conversion of T4 to inactive reverse T3. These changes are seen in other chronic illnesses. T4 levels are in the low-normal range, whereas T3 levels are depressed. Clinical evidence of hypothyroidism includes sensitivity to cold, constipation, dry skin, and bradycardia.

Plasma cortisol levels are raised and the normal diurnal variation is lost. These changes are due in part to the increased half-life of cortisol seen in starvation, and in part to a relative increase in cortisol production. Growth-hormone levels are also increased, another secondary effect of starvation. Prolactin secretion is normal. Leptin levels are low.

Haematological changes

A normocytic normochromic anaemia is found in a minority of patients and is sometimes attributable to a low intake of iron or folate. Mild neutropenia is common. The erythrocyte sedimentation rate (**ESR**) is generally low.

Other metabolic abnormalities

Hypercholesterolaemia is frequently present. The mechanism is not understood. Increased serum beta-carotene may also be found and reflects increased dietary intake. Life-threatening hypoglycaemia very occasionally occurs, but may not present typically due to impaired sympathetic response. Dehydration is not uncommon, and electrolyte disturbance is found in those who vomit frequently or misuse large quantities of laxatives or diuretics.

Other abnormalities

Brain imaging studies have revealed enlargement of the cortical sulci and cisternes and dilatation of the ventricles. This appears to be reversible and has been termed 'pseudoatrophy'.

There is delayed gastric emptying and a prolonged gastrointestinal transit time. This may account for the common complaints of fullness after eating, bloatedness, and constipation. Acute gastric dilatation is a rare complication which can be provoked by episodes of extreme overeating or attempts at refeeding that are too vigorous.

In more long-standing cases bone mineral density is reduced, probably as a result of oestrogen deficiency and low weight. There is a heightened risk of fractures, particularly of the lumbar vertebrae. The bone loss appears to reverse with weight regain and the resumption of regular menstruation.

Symptoms and signs in bulimia nervosa

There are few physical complaints in bulimia nervosa. Those most commonly encountered are irregular or absent menstruation, weakness and lethargy, vague abdominal pains, and toothache. On examination, appearance is usually unremarkable. Salivary gland enlargement may be present: typically, this involves the parotids and gives the patient's face a slightly rounded appearance. Sometimes it is associated with a raised serum amylase level, the increase being in the salivary isoenzyme. The underlying pathophysiology is not understood. In those who vomit there may be calluses on the dorsum of the dominant hand (Russell's sign) due to the fingers being used to stimulate the gag reflex. Also, there may be significant erosion of the dental enamel particularly on the lingual surface of the upper front teeth. A minority of patients, particularly those who take large quantities of laxatives or diuretics, have intermittent peripheral or facial oedema.

Abnormalities on investigation in bulimia nervosa

Of most importance is the electrolyte disturbance which is encountered in about half of those who vomit or take laxatives or diuretics. Metabolic alkalosis, hypochloraemia, and hypokalaemia are the most common abnormalities and may account for the weakness and tiredness (and in rare instances hypokalaemic paralysis) experienced by some patients. The overall picture may resemble Bartter's syndrome. Severe electrolyte disturbance is occasionally encountered, particularly low potassium levels, but even when it is long-standing there may be surprisingly few accompanying symptoms. Despite concern about possible cardiac arrhythmias, nephrogenic diabetes insipidus, and the suggestion that chronic hypokalaemia may induce changes in the renal proximal tubular cells, aggressive treatment of this type of chronic electrolyte disturbance is rarely appropriate: instead, it should be monitored while treatment is focused on the eating disorder itself.

Endocrine abnormalities are also encountered in bulimia nervosa. They resemble those seen in anorexia nervosa, but are not as severe. They are thought to be secondary to the strict dieting and are probably reversible, given that the menstrual disturbance responds to the correction of the eating disorder.

Aetiology

It is generally accepted that anorexia nervosa and bulimia nervosa are the result of a complex interplay of physiological, psychological, and social processes. These have different influences at different stages in the development and subsequent course of the disorder. The understanding of the exact nature of these processes is limited, although there is increasing convergence between the findings of biological and psychosocial studies.

Development of anorexia nervosa and bulimia nervosa

Anorexia nervosa generally starts in mid-adolescence with a period of voluntary dietary restriction that proceeds to get out of control. Whereas everyday adolescent dieting is neither persistent nor extreme, in anorexia nervosa it becomes unremitting and intense. As a result body weight falls and a state of semi-starvation eventually develops. Concerns about shape and weight may pre-date the onset of the dieting or develop as weight is lost.

Bulimia nervosa starts in a similar way, although the age of onset is typically some years later. There is dietary restriction and it too leads to weight loss (sufficient to result in a period of anorexia nervosa in about 25 per cent of cases), but instead of dietary control being maintained, it becomes punctuated by episodes of binge-eating. This breakdown in control generally occurs within 2 years of onset. At first the episodes of overeating may be both modest in size and intermittent, but gradually they become larger and more frequent. As a result, the lost weight is regained and body weight returns to about the level at which the dieting first began. By this point the disorder tends to be self-perpetuating. At some stage in this sequence of events self-induced vomiting or laxative misuse may be adopted to compensate for the episodes of overeating. In practice they have the opposite effect, because these patients' belief in their effectiveness at preventing energy absorption undermines their attempts to control their eating.

Risk factors and processes

Epidemiological studies have implicated a variety of risk factors in the development of anorexia nervosa. These may be divided into four classes:

1. Being female, adolescent, and living in a Western society. These individuals are under social pressure to diet.
2. Being exposed to a microenvironment that further encourages dieting. This includes being brought up in a family in which there is intense interest in shape, weight, or eating, sometimes as a result of one or more family members having a frank eating disorder. Social or occupational pressures to diet are another example.
3. Being exposed to factors that increase the risk of psychiatric disturbance in general and depression in particular. These include a family history of psychiatric disorder, especially depression, and exposure to adverse childhood experiences such as parenting deficits and sexual and physical abuse.
4. The presence of the psychological traits of perfectionism, inflexibility, and low self-esteem.

A similar set of risk factors has been implicated in bulimia nervosa, although it seems that there is more exposure to social factors that encourage dieting. For example, there are strikingly raised rates of parental and childhood obesity (antedating the eating disorder), both of which are likely to sensitize individuals to their appearance and weight, and thereby make them prone to diet. The risk factors for psychiatric disturbance are also prominent, whereas the traits of perfectionism and inflexibility are less pronounced. One additional class of risk factor is parental substance abuse. How this might operate is not clear, although it does seem that those who develop bulimia nervosa are prone to mood fluctuations and that some may learn to modulate them by consuming large quantities of food, alcohol, or psychoactive drugs.

An important question is why people with bulimia nervosa have repeated episodes of loss of control over eating, whereas those with anorexia nervosa do not. Several factors appear to be relevant. First, the traits of perfectionism and inflexibility are less prominent, which may result in the person being less able to maintain strict self-control. Second, the mood fluctuations may interfere with dietary restraint. Third, the vulnerability to obesity, and perhaps therefore overeating, may also be relevant.

Genetic factors and neurobiological mechanisms

Family-genetic studies have demonstrated that eating disorders run in families, with familial aggregation being particularly evident in anorexia nervosa. The relative extent of genetic and environmental contributions is unclear: the findings of the few twin studies have been inconsistent and there have been no adoption studies.

The nature of any inherited vulnerability is not known. The liability appears not to be shared with that for other psychiatric disorders. However, there is some evidence of shared familial transmission between the various eating disorders, and between anorexia nervosa and 'obsessive–compulsive personality disorder', a personality construct that overlaps with certain of the traits mentioned earlier. Thus anorexia nervosa and obsessive–compulsive personality disorder may be common phenotypic expressions of a similar genotype. Clearly there may also be genetically determined abnormalities in the regulation of weight and eating habits.

Attempts to identify susceptibility genes have focused on those implicated in serotonin (5-hydroxytryptamine, **5-HT**) neurotransmission. This transmitter is of particular interest for a number of reasons: (1) it is known to have an important role in the control of both eating and mood; (2) there is some evidence of trait-related abnormalities in brain 5-HT function in both anorexia nervosa and bulimia nervosa; and (3) it has been found that dieting influences brain 5-HT function in women. There is some evidence, albeit inconsistent, that anorexia nervosa is associated with a polymorphism of the 5-HT $_{2A}$ receptor.

Maintaining factors and processes

Several processes maintain the dietary restriction that is central to anorexia nervosa. One is the potent and strongly reinforcing sense of self-control that these people get from restricting their eating. Another is the resulting weight loss, given the overevaluation of shape and weight. A third is the effect on others of refusing to eat, which can be of special significance when there are dysfunctional relationships. A fourth is secondary to certain aspects of the starvation state. For example, the social withdrawal and preoccupation with food and eating both narrow the focus of the person's interests, and the delayed gastric emptying produces feelings of fullness even after eating modest amounts of food.

In bulimia nervosa similar maintaining processes operate. For example, the periods of successful dietary control are strongly reinforcing, as is the initial weight loss. However, there are also important differences. First, since the disorder is usually kept secret, there may be no reinforcing effects on others. Second, since body weight is generally unremarkable, starvation-related mechanisms are less relevant. Third, the repeated binges, while being aversive and a stimulus to change,

strongly encourage further dieting as they undermine the sense of being in control and magnify fears of weight gain.

Treatment

Patients with eating disorders vary in the severity of their presenting symptoms and in their response to treatment. Some have a brief period of disturbance which spontaneously remits and does not recur; in others treatment is needed, but there is full and lasting recovery; while in others the disorder persists and may prove intractable.

The treatment of bulimia nervosa has been the subject of numerous randomized controlled trials. The most effective treatment is a specific form of cognitive–behaviour therapy. This is a psychological treatment which directly addresses both the disturbed eating habits and the abnormal attitudes to shape and weight. It involves about 20 sessions over 5 months and generally results in substantial improvement, with about half the patients making a complete and lasting recovery. Antidepressant drugs are the only pharmacological treatment to have shown promise. They result in a decline in the frequency of binge-eating and associated compensatory behaviour, and an improvement in mood, but their effect is not as great as that obtained with cognitive–behaviour therapy and, more importantly, it is often not maintained. Some patients respond to simple and brief forms of cognitive–behaviour therapy. These may be implemented in primary care. Indeed, a 'stepped care' management strategy has been advocated in which a simple treatment (such as cognitive–behavioural self-help) is used first, with more intensive and specialized treatments (such as full cognitive–behaviour therapy) being reserved for those who do not respond.

There has been relatively little research on the treatment of anorexia nervosa. This is for a number of reasons, including the low prevalence of the disorder and the fact that treatment may take a year or more. Another barrier to research and, indeed treatment, is the egosyntonic character of the disorder.

In principle, there are three aspects to the management of anorexia nervosa. The first is persuading patients that they need help, and maintaining their motivation thereafter, which is crucial given their reluctance to change. The second is weight restoration. The third is addressing the patient's overevaluation of shape and weight, eating habits, and general psychosocial functioning.

Weight restoration is needed to reverse the effects of starvation and of itself usually leads to substantial improvement in the patient's physical and psychological state. Weight restoration may be achieved on an outpatient, day-patient, or inpatient basis. Indications for hospital admission include risk of suicide, adverse home circumstances, and failure of outpatient treatment. Physical indications include a body mass index below 13.5, rapid weight loss, and the presence of medical complications such as marked oedema, severe electrolyte disturbance, hypoglycaemia, or significant intercurrent infection. Under such circumstances admission should be to a general medical ward or a psychiatric unit with good access to general medical help. Weight restoration may be achieved in either setting, but it is a great advantage if the staff are experienced in the management of patients with anorexia nervosa, and in a psychiatric unit it is generally easier to arrange the other aspects of treatment. Inpatient care should always be regarded as a preliminary to outpatient treatment.

There is no single way of addressing overevaluation of shape and weight, eating habits, and general psychosocial functioning. Leading approaches include various forms of family therapy, primarily for younger patients, and adaptations of cognitive–behaviour therapy. Training is needed to deliver these treatments, and they are best conducted on an outpatient basis.

Drug treatment has almost no role, although occasionally it is appropriate to use drugs to stimulate the resumption of regular menstruation, so long as body weight has reached a reasonable level and the patient is eating healthily.

Prognosis

Whilst at least half of those with anorexia nervosa recover in terms of their weight and menstrual function, sensitivity about shape and weight often persists and eating habits may remain disturbed. Up to one-quarter develop bulimia nervosa. The standardized mortality ratio is significantly raised, deaths being either a direct result of medical complications or due to suicide. The outcome in males appears to be similar to that in females. Few consistent predictors of outcome have been identified, exceptions being a long history and late onset, both of which are associated with a worse prognosis.

The outcome in bulimia nervosa is also varied, although it is substantially improved by cognitive–behaviour therapy. The mortality rate does not appear to be raised. No consistent predictors of outcome have been identified. Both anorexia nervosa and bulimia nervosa 'breed true', in that they do not seem to evolve into any other disorder.

Prevention

Programmes for the primary and secondary prevention of anorexia nervosa and bulimia nervosa have been developed, but none has been satisfactorily evaluated. The primary prevention programmes tend to focus on schoolgirls, the group most at risk. However, two specific difficulties have been encountered: first, there is a danger of magnifying concerns about shape and weight rather than reducing them; and second, there is potential for conflict between the content of these programmes and those directed at the prevention of obesity.

Pregnancy and childrearing

This topic has come to the fore with the emergence of bulimia nervosa, since pregnancy does not often occur in the course of anorexia nervosa. It is now clear that eating disorders can have untoward effects. They generally improve during pregnancy, but the amount of weight gained may be abnormally low or high. The effects on the fetus have yet to be established, although there have been reports of intrauterine growth retardation and low birth weight. Childrearing is impaired in some cases with adverse effects on the child's feeding and growth.

Atypical eating disorders

More than one-third of those who present for the treatment of an eating disorder do not meet the diagnostic criteria for anorexia nervosa or bulimia nervosa. Such patients have an 'atypical eating disorder', the equivalent North American term being an 'eating disorder not otherwise specified'. These eating disorders have received little attention. Many are similar in form to anorexia nervosa or bulimia nervosa and respond similarly to treatment.

Binge-eating disorder

The one atypical eating disorder to have been delineated is termed 'binge-eating disorder'. It is characterized by recurrent binge-eating in the absence of the extreme weight-control behaviour seen in anorexia nervosa and bulimia nervosa. The binge-eating occurs against the background of a general tendency to overeat. It seems to be more an habitual response to negative mood states than an intermittent breakdown of dietary restraint as occurs in bulimia nervosa. Not surprisingly, many of these patients are overweight or obese.

Little is known about the distribution of binge-eating disorder. It appears to affect an older age group than anorexia nervosa and bulimia nervosa, and cases among men are not uncommon. About 10 per cent of those attending obesity clinics have the disorder. The treatment that shows most promise is conventional behavioural weight control. As in bulimia nervosa, cognitive–behavioural self-help programmes help a subgroup of cases. The value of drug treatment is unclear.

Further reading

Andersen AE, Bowers W, Evans K (1997). Inpatient treatment of anorexia nervosa. In: Garner DM, Garfinkel PE, eds. *Handbook of treatment for eating disorders*, pp 327–53. Guilford Press, New York.

Fairburn CG, Marcus MD, Wilson GT (1993). Cognitive-behavioral therapy for binge eating and bulimia nervosa: a comprehensive treatment manual. In: Fairburn CG, Wilson GT, eds. *Binge eating: nature, assessment and treatment*, pp 361–404. Guilford Press, New York.

Fairburn CG, Brownell KD (2002). *Eating disorders and obesity: a comprehensive handbook*, 2nd edn. Guilford Press, New York.

Garner DM, Garfinkel PE (1997). *Handbook of treatment for eating disorders*. Guilford Press, New York.

Lilenfeld LR, Kaye WH (1998). Genetic studies of anorexia and bulimia nervosa. In: Hoek HW, Treasure JL, Katzman MA, eds. *Neurobiology in the treatment of eating disorders*, pp 169–94. Wiley, Chichester.

Mitchell JE, Pomeroy C, Adson DE (1997). Managing medical complications. In: Garner DM, Garfinkel PE, eds. *Handbook of treatment for eating disorders*, pp 383–93. Guilford Press, New York.

Russell GFM, Treasure J, Eisler I (1998). Mothers with anorexia nervosa who underfeed their children: their recognition and management. *Psychological Medicine* **28**, 93–108.

Wilson GT, Fairburn CG (2002). Treatments for eating disorders. In: Nathan PE, Gorman JM, eds. *A guide to treatments that work* 2nd edn. Oxford University Press, New York.

26.5.6 Schizophrenia, bipolar disorder, obsessive–compulsive disorder, and personality disorder

S. Lawrie

Recent research, using reliable diagnostic criteria based on clinical features since diagnostic laboratory tests are not available, has established that all of these conditions have both biomedical and psychosocial components. The *Diagnostic and statistical manual of mental disorders, 4th edition* (**DSM-IV**) has been most rigorously developed and is used here.

Schizophrenia

Schizophrenia and bipolar disorder are psychoses, that is to say they include phenomena that qualitatively differ from everyday experience. Schizophrenia is characterized by delusions, hallucinations, disorganized speech/behaviour, and negative symptoms.

Onset is in early adulthood (median age 25 years). The sex incidence is equal, but women tend to be affected later than men. The incidence is only 15 in 100 000 of the population per year, but the prevalence is about 5 in 1000 due to chronicity, and the lifetime risk is 1 per cent.

Aetiology

Genetic factors account for 80 per cent of the liability to schizophrenia, but no major genes have been identified. Having an affected relative increases the risk 5 to 50 times, depending on the relationship. Other risk factors include obstetric complications, developmental problems, and cannabis use, but these only double the risk. Stressful life events can be precipitants, but only in those otherwise predisposed.

There are subtle abnormalities of brain structure and function (particularly of the temporal and frontal lobes) in both chronic and first episode cases. Developmental changes in brain structure (for example, synaptic pruning) and function (for example, dopamine sensitivity) are thought to disrupt frontotemporal integration and bring on symptoms, but direct evidence is lacking.

Clinical features

Hallucinations and delusions are 'positive symptoms', that is, abnormal by their presence. Hallucinations are perceptions in the absence of stimuli. They are usually auditory 'voices' speaking the patients' thoughts or commenting on their actions. Hallucinations in other senses can occur but suggest a neurological disorder. Delusions are unshakeable false beliefs. Persecutory ('paranoid') delusions are common but occur in all psychoses. Delusions of passivity (actions or feelings 'made' by external forces) and other bizarre beliefs are more specific. The other positive symptom is 'thought disorder'—an illogical sequence of thoughts (as revealed in speech).

'Negative symptoms' are features that are abnormal by their absence. Common symptoms include a loss of emotion ('flat affect'), apathy, self-neglect, and social withdrawal. These may be prodromal, but are more common in chronic patients, and can be confused with depression or parkinsonism.

Differential diagnosis

Prodromal symptoms can be similar to depression. Delusions of passivity can be confused with obsessional ideas, but the latter are recognized as one's own. Drug intoxication can cause positive symptoms, but also disorientation. Neurological causes, for example temporal lobe epilepsy or brain tumours, are rare. The distinction of schizophrenia from bipolar disorder is based on whether psychotic or affective features predominate. Rarely, if both are present equally, a diagnosis of schizoaffective disorder is appropriate.

Management

Acute positive symptoms generally respond well to any antipsychotic drug (Table 1). These work by dopamine-receptor blockade. The main adverse effects are sedation, weight gain, and extrapyramidal syndromes (acute dystonia, akathisia, parkinsonism, tardive dyskinesia). These are best avoided by minimizing dosage, but dystonias and parkinsonism respond to anticholinergics. Medication should be continued for at least 2 years to reduce relapse rates.

Patients often refuse medication, due to adverse effects or lack of insight. Some are suitable for depot medication (intramuscular injections of esterified antipsychotics, see Table 1). The new 'atypical' antipsychotics have fewer adverse effects, but this is primarily because they are prescribed in relatively low doses. It is claimed that they are effective in those with negative and treatment-resistant positive symptoms, but clozapine is the only proven such treatment. Clozapine is the definitive atypical antipsychotic, with relatively high serotonin:dopamine-receptor blockade, but carries a considerable risk of neutropenia and agranulocytosis. These treatments are not contraindicated in pregnancy, as they confer only a small increased risk of teratogenicity and an untreated psychosis is more dangerous.

There are few effective non-drug treatments. Cognitive therapy may reduce symptoms and improve drug compliance. Illness education reduces relapse rates, as does teaching social skills, but these may primarily work by improving drug compliance.

Primary prevention is not a realistic prospect until better understanding of the pathogenesis of schizophrenia allows early detection. There is, however, some

evidence that earlier treatment with antipsychotics may be associated with a better prognosis.

Prognosis

The prognosis is generally poor. About 25 per cent of patients will only have one or two episodes, but most will suffer chronic symptoms, numerous relapses, unemployment, and social isolation. Most patients smoke heavily, and many abuse alcohol/drugs, resulting in a high premature mortality rate. Suicide is all too common, at 10 to 15 per cent over a lifetime.

Bipolar disorder (BPD)

The key features of bipolar disorder ('manic depression') are episodic increases or decreases in mood, thoughts, and activity, lasting at least 1 week. The prevalence, incidence, and lifetime risk of BPD are similar to schizophrenia (at 0.5 per cent, 0.01 per cent, and 1 per cent, respectively) and the sex incidence is also equal, but the mean age at onset is 21 years in both sexes in BPD.

Aetiology

The risk factors for BPD are similar to those for schizophrenia. Genetic influences are equally strong, but other associations are weaker in BPD. There may be specific abnormalities in monoamine metabolism and neuroendocrine function in BPD, for example a first 'high' may be precipitated by antidepressants, stimulants, steroids, or childbirth.

Clinical features

If 'hypomanic', such patients feel 'high', report rapid thoughts, have limitless energy, require little sleep, and are 'disinhibited', that is they are overfamiliar and take risks. They speak quickly ('pressure of speech') and jump between topics ('flight of ideas'), but with logical connections between thoughts. If psychotic ('manic'), their delusions and hallucinations are usually 'mood congruent', for example 'grandiose delusions' of special abilities.

If depressed, their mood is low, activities are not enjoyed ('anhedonia'), interests are diminished, energy is low, sleep is disturbed, appetite is reduced, and weight may fall. Patients typically think they are worthless, the future is hopeless, and they may be suicidal. Severe 'melancholic' depression is accompanied by early morning wakening, 'diurnal mood variation' (feeling worst in the morning), and 'psychomotor retardation' (head down, expressionless face, little spontaneous activity). Any psychotic symptoms are again mood-congruent, for example delusions of sin or guilt, 'voices' criticizing the patient.

Occasionally, one encounters 'mixed states', where patients have some features of (hypo)mania and depression simultaneously.

Differential diagnosis

Some one-third of patients will have several depressive episodes before their first 'high'. With no previous history, especially if old, (hypo)mania may rarely be attributable to thyroid disease or dementia. In established cases, the main differential is schizophrenia as individual symptoms can be similar (e.g. low mood and flat affect, thought disorder, and flight of ideas) but the key is that BPD is an episodic disturbance of mood.

Management

The treatment of depression is discussed elsewhere. 'High' patients may reject treatment, but most will later report feeling out of control and gratitude for being treated. (Hypo)mania generally responds well to antipsychotic drugs (see Table 1). Lithium is also effective, particularly with high serum levels (of about 1.0 mmol/l). Carbamazepine and sodium valproate are alternative 'mood stabilizers', with fewer adverse effects, but may be less effective. Valproate is the best treatment in 'rapid cycling disorder' (four or more illness episodes annually). Acute mixed episodes are probably best treated with mood stabilizers alone.

Prophylaxis is required if patients have two or more episodes in 5 years. Lithium (maintained at 0.5–1.0 mmol/l) is the treatment of choice as it reduces both (hypo)manic and depressive relapses, and may reduce suicide rates. Common adverse effects are dose-related and include diarrhoea, tremor, thirst, polyuria, and weight gain. Long-term effects can include hypothyroidism and renal impairment. Lithium is excreted in competition with sodium, so thiazide diuretics, non-steroidal agents, and dehydration can precipitate toxicity. Sudden vomiting, coarse tremor, sedation, or dysarthria require urgent medical treatment to avoid seizures, renal failure, and death.

Lithium is contraindicated in pregnancy and breast feeding, as it increases the rate of Fallot's tetralogy 10- to 20-fold and can cause neonatal toxicity. Carbamazepine and sodium valproate are associated with spina bifida. Best practice is therefore to use antipsychotics for symptom control until after delivery. Mood stabilizers should then be reinstituted, and bottle feeding commended, as relapse is very common postnatally.

Early diagnosis and self-medication can minimize relapses in established cases, but primary prevention is not presently possible.

Prognosis

Acute episodes generally respond to treatment, but 10 per cent of patients will be ill for a year and 50 per cent of (hypo)manic patients immediately become depressed. Recurrence is the norm, and becomes more frequent with age. Many patients have chronic symptoms between episodes, such that only 25 per cent of patients fully recover. Employment and social difficulties are common. Before treatment was possible, the annual mortality from cardiovascular collapse or suicide was 10 per cent. This is now the lifetime rate of suicide.

Obsessive compulsive disorder (OCD)

OCD is characterized by recurrent, unwanted thoughts that are recognized as one's own and/or repeated acts ('rituals') to relieve tension. Originally viewed as a neurosis, DSM-IV does not use this ambiguous term and classifies OCD as an anxiety state. OCD is common, with a prevalence of 1 per cent and lifetime risk of 2 to 3 per cent. The peak prevalence is in the fourth decade, with an earlier onset and slight excess in women.

Aetiology

OCD runs in families but twin studies are inconclusive. Obsessional personality is a risk factor, as are other personality disorders, childhood conduct disorder, and Tourette's syndrome. OCD can arise after lesions of the frontal lobes and basal ganglia; regions also implicated by studies of brain structure and function in patients. Follow-up functional imaging has found recovery is associated with normalized metabolism in these areas of the brain. Effective drugs all inhibit serotonin (5-hydroxytryptamine, **5-HT**) reuptake, and neurochemical evidence also suggests postsynaptic serotoninergic hypersensitivity and/or low synaptic 5-HT concentrations.

Clinical features

Patients are commonly ill for years before they come to medical attention. Obsessions are thoughts that are recurrent, resisted, and unwanted, but regarded as one's own. The thought may be a fear of contamination, excessive doubt, a somatic concern, a desire for precision, or an aggressive or sexual impulse. Usually, several obsessions coexist. Obsessional slowness and precision are more common in men.

Compulsions are obsessional acts, based on these thoughts and usually performed as a means to reduce anxiety, that are not in themselves pleasurable. Checking, cleaning, and counting are the most common acts. Most are performed in private as they are recognized as senseless.

Differential diagnosis

Rituals and superstitions are common in childhood, but are only pathological if they cause distress. Distinguishing OCD from depression/anxiety can be difficult, as secondary depression is frequent in OCD, and thoughts in depression and anxiety can have an obsessional quality. Obsessions can also be similar to phobias, but obsessives actually seek out anxiety-provoking stimuli whereas phobics avoid them.

Management

Both drugs and psychotherapy are effective in OCD. The tricyclic antidepressant clomipramine and the 'selective' serotonin-reuptake inhibitors (**SSRIs**) work, but high doses may be required. Clomipramine is probably most effective, particularly in treating patients with comorbid depression, but has unpleasant adverse effects (sedation, weight gain, anticholinergic). Maintaining drug treatment for 1 year reduces relapse rates.

The most effective psychotherapeutic techniques are behavioural. 'Exposure' to anxiety-provoking stimuli (with or without therapist 'modelling') and 'response prevention' (avoiding rituals by persuasion, monitoring, or adopting alternative behaviours) should also involve family members. This is more effective against compulsions than obsessions. Recent trials suggest that the combination of antidepressants and exposure may be best of all.

Prognosis

Response to treatment may take months and symptoms tend to recur if drugs are stopped. About 10 per cent of cases progressively deteriorate, particularly men with an early-onset OCD. Suicide was regarded as rare in OCD but recent studies challenge this view.

Personality disorder (PD)

PD is commonly seen as distinct from psychiatric 'illness' and untreatable. It is ironic that many PDs were originally described because they had links to particular psychiatric disorders, and that recent research has rediscovered these and found possible treatments.

PD is defined as culturally abnormal experience or behaviour, with onset in early adulthood, that is pervasive and inflexible, leading to distress or impairment. Only a brief description of subtypes (Table 2) and specific points is possible here.

Aetiology

Paranoid and borderline traits are clearly heritable, and schizotypy is genetically and biologically linked to schizophrenia. Paranoid and antisocial PDs are more common in men; borderline, histrionic, and dependent PDs in women. Childhood adversity is a general risk factor but specific effects are difficult to identify. Child sexual abuse, for example, may be linked to self-harm and substance abuse rather than any particular PD. Psychopaths have mild frontal lobe deficits—and similar head trauma related changes in personality have been described.

Clinical features

Many patients meet more than one set of PD criteria. Some borderline patients report auditory and visual hallucinations. Dysthymia (formerly depressive PD) and cyclothymia (also formerly a PD) are now seen as mild depression and bipolar disorder, respectively. Paranoid, schizotypal, obsessional, and avoidant PDs may be similarly reclassified in the future.

Differential diagnosis

The main differential is with the associated psychiatric disorder and other PDs. As a rule, the eccentric cluster can present similarly to schizophrenia, the emotional cluster-like bipolar disorder, and the anxious cluster with anxiety states. Borderline PD can cause diagnostic difficulties with both schizophrenia and BPD.

Management

Clinical management has conservative aims, but pressure is being put on psychiatrists to do more. Few treatments have been evaluated for PD *per se*. Cognitive therapy may generally reduce the frequency of deliberate self-harm. Dysthymia responds to antidepressants; paranoid, schizotypal, and borderline patients may benefit from antipsychotic drugs; and obsessional PD may respond to SSRIs.

Prognosis

Recurrent deliberate self-harm and eventual suicide is common. Patients with PDs who develop other psychiatric disorders also have a poor prognosis for that disorder.

Concluding remarks

Psychiatric disorders are common, involuntary, distressing, and disabling. Like most medical disorders, they are caused by biological, psychological, and social factors. They can be successfully treated, but are often chronic and associated with social rejection. Stigmatization remains a big problem, not least from doctors.

Further reading

Abramowitz JS (1997). Effectiveness of psychological and pharmacological treatments for obsessive–compulsive disorder: a quantitative review. *Journal of Consulting and Clinical Psychology* **65**, 44–52. [As recommended in the Cochrane Library]

Altshuler L, *et al*. (1996). Pharmacologic management of psychiatric illness during pregnancy: dilemmas and guidelines. *American Journal of Psychiatry* **153,** 592–606. [High quality systematic review]

American Psychiatric Association (1994). *The diagnostic and statistical manual of mental disorders*, 4th edition. APA, Washington DC. [Diagnostic criteria for all psychiatric disorders, with background information]

Cannon M, Jones P (1996). Epidemiology of schizophrenia. *Journal of Neurology, Neurosurgery and Psychiatry* **61,** 604–13. [Comprehensive review of the epidemiology of schizophrenia]

Daly I (1997). Mania. *Lancet* **349,** 1157–60. [Accessible review of bipolar disorder]

Frith C (1995). Schizophrenia: functional imaging and cognitive abnormalities. *Lancet* **346,** 615–20. [One of a generally excellent series of reviews about schizophrenia]

Haslam DRS, *et al*. (1997). The treatment of bipolar disorder: review of the literature, guidelines and options. *Canadian Journal of Psychiatry* **42,** Suppl. 2. [Best available synthesis of the literature on the treatment of bipolar disorder, although Cochrane reviews are in progress]

Hawton K, *et al*. (1999). Deliberate self-harm: the efficacy of psychosocial and pharmacological interventions. In: *The Cochrane Library*, Issue 3. Update software, Oxford. [Systematic review and meta-analysis of treatments for self-harm, including five personality disorder trials]

Johnson JG, *et al*. (1999). Childhood maltreatment increases risk for personality disorders during early adulthood. *Archives of General Psychiatry* **56,** 600–6. [Community-based longitudinal study of the antecedents of personality disorder]

Johnstone EC, Freeman CPL, Zealley AK, eds (1998). *Companion to psychiatric studies*, 6th edn. Churchill Livingstone, Edinburgh. [Detailed and well-referenced textbook of psychiatry]

McIntosh A, Lawrie SM (2001). Schizophrenia. In: *Clinical evidence* pp. 695–716. British Medical Journal, London. [Regularly updated summary of efficacious treatments for schizophrenia, including summaries of relevant Cochrane reviews]

26.6.1 Psychopharmacology in medical practice

P. J. Cowen

Introduction

Psychotropic drugs are widely used in medical practice so that most clinicians are likely to have under their care a number of patients receiving treatment with psychoactive medication (Table 1). Practitioners therefore need to have an understanding of the uses and side-effects of psychotropic drugs, particularly of the way in which such medication can interact with drugs used to treat other medical disorders.

The majority of psychotropic drugs are prescribed for the treatment of depressive and anxiety disorders. This reflects the frequency of these conditions in both primary care and general hospital settings; accordingly, drug treatment for anxiety and depression will often be instituted both by general practitioners and hospital clinicians. Similarly, while the principal use of antipsychotic drugs is in the treatment of schizophrenia, such agents are also frequently used in general hospitals in the management of organic psychoses. Finally, while treatment with mood-stabilizing drugs, such as lithium, will generally be initiated by psychiatrists, patients receiving long-term therapy may well require treatment for coexisting medical disorders, because of which a knowledge of the effects of lithium on different body systems and its liability to produce adverse drug interactions will be required.

Drug overdose

The effects of deliberate or accidental overdose of psychotropic drugs will also involve physicians (see Chapter 8.1 and Chapter 26.5.2). Related to this is the general point that when prescribing psychotropic drugs, particularly for depressed patients, the risk of overdose should always be considered. If such a risk is present the practitioner should: (1) ensure that medication is dispensed in small amounts; (2) consider asking a close relative to supervise the medication; (3) use a relatively non-toxic drug, if possible.

Pharmacokinetic factors

Most psychotropic drugs are highly lipophilic and well absorbed from the gastrointestinal tract. They are metabolized by the liver to water-soluble derivatives which are eliminated by the kidney, hence their half-life will be prolonged in patients with hepatic or renal impairment and in the elderly. Where psychotropic medication is added to another drug treatment the possibility of drug interaction must be considered. For example, selective serotonin-reuptake inhibitors are potent inhibitors of hepatic cytochrome P-450 enzymes and can thereby increase plasma levels of coadministered drugs such as warfarin.

Withdrawal of psychotropic medication

Psychotropic and many other classes of drugs produce neuroadaptive changes during their repeated administration. Readjustment has to occur when drug treatment is stopped, and this may appear clinically as a withdrawal or abstinence syndrome. Characteristic abstinence syndromes have been described for the antidepressants and anxiolytics, while the sudden discontinuation of lithium can provoke a 'rebound' mania, hence it is prudent to withdraw psychotropic drugs slowly whenever possible. It is clearly also important to be able to distinguish withdrawal syndromes from relapse of the disorder being treated.

Compliance and concordance with treatment

In psychotropic drug prescribing, compliance is an even greater problem than in general therapeutics. Psychoactive drugs frequently have unpleasant side-effects and, while side-effects are experienced early in treatment, several days may elapse before a therapeutic response is evident. In addition, patients may not see the need for treatment or believe that it can help them. Careful explanation accompanied by written instructions can help to ensure that necessary medication is taken.

It is increasingly recognized that the successful and safe use of medication requires a collaborative relationship between patient and doctor. The term 'concordance' may therefore be preferred to 'compliance', which carries the implicit assumption that the patient's job is to obey instructions. It is therefore important to acquire an understanding of the patient's attitude to his or her illness as well its treatment. For example, discussion that helps patients to weigh the advantages and disadvantages of drug treatment ('compliance therapy') has been shown to benefit those with schizophrenia.

Antidepressant drugs

All currently employed antidepressant drugs, through one mechanism or another, increase the activity of serotoninergic (5-hydroxytryptamine, **5-HT**) and/or noradrenergic neurones in the CNS. The pharmacological actions of both noradrenaline (norepinephrine) and 5-HT in the synapse are terminated by specific reuptake pumps that draw these neurotransmitters back into the presynaptic nerve ending. Most antidepressants potentiate the action of 5-HT and noradrenaline by blocking this reuptake process.

Tricyclic antidepressants

Pharmacology

Tricyclic antidepressants inhibit the neuronal uptake of noradrenaline (norepinephrine) and 5-HT. They have numerous other pharmacological properties, but these are thought to contribute to their adverse effect profile rather than their therapeutic activity. However, some of these adverse effects, for example, sedation, can prove beneficial in certain circumstances.

Principal drugs

These are amitriptyline, clomipramine, desmethylimipramine, dothiepin, doxepin, imipramine, lofepramine, and nortriptyline.

Indications and use

Tricyclic antidepressants are still the most widely prescribed drug treatment for the management of depressive illness, but their use, particularly in less severe depressive states, is waning in favour of newer compounds that are better tolerated (see below). However, none of the newer antidepressants is more efficacious than the tricyclics and their therapeutic activity in severely ill patients is not as well established. For this reason, unless there are specific contraindications, tricyclic antidepressants should be preferred in depressed inpatients or in those with marked melancholic features.

Depressed patients with prominent sleep disturbance and anxiety should be treated with a sedating antidepressant such as amitriptyline; for other patients, less sedating compounds such as lofepramine or nortriptyline can be used. To obtain tolerance to side-effects, it is usual to begin treatment at a low dose, for example 25 to 50 mg of amitriptyline at night, and to increase the amount over about 2 to 3 weeks to the usual therapeutic dose, which ranges between 75 and 200 mg daily for amitriptyline and imipramine. Tricyclic antidepressants have long half-lives and a single daily dose taken at night is usually appropriate. Patients should be warned about side-effects because this helps to ensure compliance in the early stages of treatment. They should also be advised that a clear therapeutic response may not appear for up to 2 to 4 weeks.

If treatment is successful, it is usual to continue the antidepressant for 4 to 6 months at the original dose if tolerance allows (so called 'continuation therapy'). This reduces the risk of early relapse by about half. Some patients with recurrent depressive illness require long-term prophylactic treatment with antidepressant drugs. This should be considered in those who have had more than two episodes of depression in the previous 5 years, particularly if the episodes have been severe in terms of symptomatology and impact on work and social functioning.

Side-effects

As well as inhibiting the uptake of noradrenaline (norepinephrine) and 5-HT, tricyclic antidepressants possess antagonist properties at a variety of neurotransmitter receptors, including muscarinic cholinergic receptors, a $_1$-adrenoceptors, and H_1-histamine receptors. These receptor-antagonist effects account for much of the adverse effect profile of these agents, particularly their anticholinergic properties (Table 2). Tricyclics also possess membrane-stabilizing effects that underlie their most serious side-effect of cardiotoxicity, which can be particularly problematic in tricyclic overdose, where ingestion of less than 1 g can sometimes prove fatal. Lofepramine**,** however, is relatively safe in overdose. Tricyclics should be used with caution in patients with cardiovascular disease. They also lower the seizure threshold and can thereby aggravate pre-existing epilepsy or sometimes cause seizures *de novo*.

Drug interactions

Tricyclic antidepressants antagonize the hypotensive effects of a $_2$-adrenoceptor agonists such as clonidine, but can be safely combined with thiazides and angiotensin-converting enzyme (**ACE**) inhibitors. The ability of tricyclics to block noradrenaline (norepinephrine) reuptake can lead to hypertension with systemically administered noradrenaline and adrenaline (epinephrine). Tricyclics should not be used in conjunction with antiarrhythmic drugs, particularly amiodarone. Plasma levels of tricyclics can be increased by numerous other drugs including cimetidine, sodium valproate, calcium-channel blockers, and selective serotonin-reuptake inhibitors (**SSRIs**).

Newer antidepressants

Principal drugs

These can be classified as follows:

- *selective 5-HT reuptake inhibitors (SSRIs)*: citalopram, fluoxetine, fluvoxamine, paroxetine, sertraline;
- *selective noradrenaline (norepinephrine)-reuptake inhibitors:* reboxetine;
- *selective noradrenaline- and serotonin-reuptake inhibitors:* venlafaxine;
- *monoamine-receptor antagonists*: mirtazapine, nefazodone, and trazodone.

Pharmacology

The actions of SSRIs are essentially confined to inhibition of 5-HT reuptake. Their use is associated with a sustained increase in brain 5-HT neurotransmission. By contrast, reboxetine inhibits only the reuptake of noradrenaline (norepinephrine). Venlafaxine is a potent blocker of 5-HT reuptake and at higher doses blocks the reuptake of noradrenaline as well. Nefazodone and trazodone have weak 5-HT-and noradrenaline-reuptake inhibiting properties. They also act as antagonists at 5-HT_2 receptors, and trazodone is a potent a $_1$-adrenoceptor antagonist. Mirtazapine is also a 5-HT $_2$-receptor antagonist, but in addition blocks inhibitory presynaptic a_2-adrenoceptors, resulting in an increased release of noradrenaline

All these compounds lack the cardiotoxicity and the anticholinergic effects of conventional tricyclic antidepressants. They are therefore safer in overdose and, in general, somewhat better tolerated. In the broad range of depressed subjects they have equal efficacy to tricyclics but, as noted above, it is not clear if this extends to the most severely ill patients.

Indications for use

The newer antidepressants should be used to treat patients in whom the use of tricyclic antidepressants is contraindicated because of their anticholinergic and cardiotoxic effects. In addition, some patients unable to tolerate a clinically effective dose of a tricyclic agent may find that one of the newer drugs causes fewer side-effects. The lack of sedation associated with SSRIs, reboxetine, venlafaxine, and nefazodone can be beneficial in outpatients striving to carry out their usual activities. Unlike tricyclic antidepressants, many of the newer drugs do not stimulate appetite and may therefore be appropriate in patients in whom weight gain would be undesirable. Finally, in patients where the risk of overdose cannot be minimized, the newer drugs may be preferred because of their lower acute toxicity.

Side-effects

The main adverse effects of the newer antidepressants are shown in Table 3. The major distinction between compounds is whether or not they are sedating. The sedating antidepressants have the advantage of improving sleep at an early stage but may impair cognitive function, while the reverse is true for SSRIs, venlafaxine, and reboxetine. Like tricyclic antidepressants, the newer compounds appear to lower seizure threshold to some extent, though this effect may be less with trazodone and SSRIs. SSRIs may increase the risk of upper gastrointestinal bleeding, particularly if given in conjunction with non-steroidal anti-inflammatory drugs (**NSAIDs**).

Drug interactions

SSRIs, with the exception of citalopram, slow the metabolism of numerous other drugs including warfarin, theophylline, anticonvulsants, antipsychotics, and tricyclic antidepressants. Dangerous interactions, characterized by 5-HT neurotoxicity, have been reported between SSRIs, venlafaxine, and monoamine oxidase inhibitors (**MAOIs**). This may be particularly problematic with fluoxetine, whose active metabolite norfluoxetine has a half-life of 7 to 10 days. At least 5 weeks should therefore elapse between stopping fluoxetine and prescribing a monoamine oxidase inhibitor. SSRIs may also produce 5-HT toxicity in combination with lithium. Trazodone and mirtazapine may increase the sedative effects of other centrally acting drugs. Nefazodone can raise plasma levels of terfenadine, causing a risk of cardiac arrhythmias. Nefazodone also elevates plasma levels of carbamazepine and digoxin. Reboxetine should not be given with other agents that might potentiate noradrenaline (norepinephrine) function (such as MAOIs) or increase blood pressure (such as ergot derivatives).

Monoamine oxidase inhibitors

Pharmacology

Monoamine oxidase inhibitors (MAOIs) block the enzyme monoamine oxidase, which deaminates the neurotransmitters, 5-HT, noradrenaline (norepinephrine), and dopamine. Monoamine oxidase exists in two forms, known as type A (which deaminates noradrenaline and 5-HT) and type B (which preferentially deaminates dopamine and tyramine). Conventional MAOIs irreversibly deactivate both type A and type B monoamine oxidase. This has two main consequences of importance for MAOI use: (1) there is a potential for serious food and drug interactions; and (2) the consequent drug and food restrictions need to be continued for 2 weeks after cessation of MAOI treatment so that new monoamine oxidase can be synthesized.

Recently a new MAOI, moclobemide**,** has been introduced. This differs from conventional MAOIs in that its inhibition of monoamine oxidase is reversible, and it selectively inhibits type A monoamine oxidase only. This leads to an increase in brain noradrenaline and 5-HT levels, but other amines such as tyramine are little affected. These factors make moclobemide much less likely than the older monoamine oxidase inhibitors to produce adverse food and drug interactions, giving it a significant safety advantage. However, while moclobemide has been shown to be effective in the treatment of moderately depressed outpatients, studies have not thus far demonstrated its efficacy in the patient groups for whom conventional MAOI treatment is currently reserved (see below).

Principal drugs

These are isocarboxazid, phenelzine, tranylcypromine, and moclobemide.

Indications and use

Conventional MAOIs are rarely used as a first choice of antidepressant, except where a patient is known to have responded to them in the past. They are usually reserved for subjects who have failed to respond to tricyclic antidepressants, newer antidepressants, or electroconvulsive therapy, where a very useful antidepressant effect can often be achieved.

Phenelzine and tranylcypromine are the two most commonly prescribed MAOIs. The usual therapeutic dose for phenelzine is between 30 and 90 mg daily. As with tricyclic antidepressants, patients should be informed about side-effects and advised that a therapeutic response from monoamine oxidase inhibitors may not be apparent for 3 to 4 weeks. Once a response is obtained, it is usually necessary to continue treatment for several months.

Side-effects

Monoamine oxidase inhibitors may cause the following side-effects:

- *central nervous system*: dizziness, muscular twitching, insomnia, confusion, mania;
- *cardiovascular:* tachycardia, postural hypotension, hypertension;
- *other:* dry mouth, blurred vision, impotence, peripheral oedema, hepatocellular damage, leucopenia.

Food and drug interactions

The major hazard of conventional MAOI treatment is through interaction with indirect sympathomimetics, that is, agents that release noradrenaline from nerve endings. The usual source of the interaction is tyramine in certain foodstuffs, especially cheese and meat extracts. Tyramine is usually metabolized by monoamine oxidase in the gut wall and liver, but in patients taking MAOIs large amounts may enter the systemic circulation, resulting in hypertension and even cerebrovascular accidents. Similar adverse effects have been reported when sympathomimetic drugs, such as amphetamines or ephedrine, are administered to patients taking monoamine oxidase inhibitors. Ephedrine or its derivatives are frequently present in cold cures: patients must therefore be warned against self-medication without seeking advice.

Hypertensive episodes resulting from the interaction of sympathomimetic drugs and monoamine oxidase inhibitors are best treated with an a_1-adrenoceptor antagonist. If one is unavailable, intramuscular chlorpromazine is an alternative.

MAOIs also produce important interactions with other commonly used drugs, including opiates, insulin, and oral hypoglycaemic agents. Except in special circumstances, combination with tricyclic antidepressants is best avoided. Combination with clomipramine, SSRIs, and venlafaxine can cause a 5-HT neurotoxicity syndrome and is contraindicated.

From the foregoing it will be apparent that conventional monoamine oxidase inhibitors should only be prescribed to patients capable of adhering to the necessary dietary restrictions. Written instructions listing prohibited foods should be provided. No additional medication should be given until the possibility of an adverse drug interaction has been excluded.

Moclobemide

Moclobemide is well tolerated, although insomnia and nausea may occur. Unlike conventional monoamine oxidase inhibitors, moclobemide does not cause significant interaction with tyramine, and adverse drug interactions also seem to be less likely. However, caution is recommended when prescribing with opiates, and combined use with SSRIs and sympathomimetic agents should be avoided. Because of the reversible nature of moclobemide's interaction with monoamine oxidase and its short half-life (about 3 h), normal monoamine oxidase activity is restored within a day of stopping treatment.

Mood-stabilizing drugs

Lithium

Pharmacology

Lithium salts have inhibitory effects on receptor-transduction systems, particularly second messengers such as cyclic-AMP and phosphoinositol. Lithium also produces marked increases in some aspects of brain 5-HT function.

Indications and use

The main uses of lithium are:

- prophylaxis of recurrent affective disorders, especially manic depressive illness;
- acute treatment of mania;
- augmentation of antidepressant medication in patients with resistant depression.

The excretion of lithium from the body is critically dependent on the kidney. Since there is little margin between the therapeutic plasma levels of lithium (0.5 to 0.8 mmol/l) and those causing toxicity (>1.2 mmol/l), the introduction of lithium therapy should be preceded by clinical and laboratory assessment of renal function. Renal function tests should include urinalysis and estimations of plasma creatinine and electrolyte levels, with measurement of creatinine clearance if there is any suggestion of impaired renal function.

Patients should initially be treated with 200 to 400 mg daily of lithium carbonate, usually as a single dose at night. Slow-release preparations of lithium are available, but their pharmacokinetics *in vivo* are very similar to those of the standard preparation. Dosage should be adjusted every 5 to 7 days on the basis of plasma lithium determinations obtained approximately 12 h after the last dose. For prophylaxis of recurrent mood disorders, plasma levels of 0.5 to 0.8 mmol/l are usually

satisfactory; but some patients—particularly those with an acute manic episode—may require higher levels (0.8 to 1.0 mmol/l). Most patients achieve adequate plasma levels with lithium carbonate dosages of between 600 and 1200 mg daily, and following this their lithium requirement is generally stable.

In the absence of clinical indications it is usually sufficient to check lithium levels every 2 to 3 months and repeat renal function tests every 6 months. Lithium can also cause hypothyroidism, so thyroid function tests should be performed prior to treatment and at 6-monthly intervals thereafter. If necessary, lithium can be combined with thyroxine replacement therapy. Sudden withdrawal of lithium in bipolar patients can cause an acute rebound mania and should be avoided if at all possible.

Side-effects

Many patients suffer from a fine tremor and nausea; diarrhoea may occur, especially at the start of treatment (Table 4). Some degree of thirst and polyuria is common, and a few patients develop nephrogenic diabetes insipidus, probably caused by lithium blocking the effect of ADH on the renal tubule. Most patients taking lithium have a demonstrable impairment of tubular concentrating ability, although this is rarely of clinical significance. Glomerular function is not usually affected by lithium, but glomerular damage and interstitial fibrosis have been reported following lithium toxicity. There are reports that long-term lithium treatment can occasionally cause long-term renal impairment. However, this risk is low provided the plasma concentrations of lithium are kept within the therapeutic range and episodes of toxicity are avoided.

Up to 80 per cent of the lithium filtered by the renal glomerulus is reabsorbed by the proximal tubule. Conditions such as diarrhoea and excessive sweating, which induce renal sodium retention, also result in increased lithium reabsorption by the kidney and elevated plasma lithium levels.

Drug interactions

Thiazide diuretics, through their effect on sodium excretion, increase lithium reabsorption and can produce lithium toxicity unless the dose of lithium is reduced and plasma concentrations carefully monitored. It is said that loop and potassium-sparing diuretics are less likely to alter lithium clearance, but it is prudent to monitor lithium levels carefully when using these drugs. Plasma lithium levels may also be increased by concomitant administration of NSAIDs, and a similar effect may be produced by metronidazole. Lithium levels may be increased by ACE inhibitors and lowered by theophylline. While the effects of lithium on cardiac conduction are usually considered benign, the effects of cardiac glycosides on conduction may be potentiated. Lithium can cause neurotoxicity (at normal plasma levels) with calcium-channel blockers and carbamazepine. Finally, lithium may increase the liability of antipsychotic drugs to cause extrapyramidal movement disorders.

Lithium toxicity

Acute lithium toxicity usually appears at plasma levels above 1.2 mmol/l. Early signs are coarse tremor, drowsiness, and dysarthria. Higher plasma concentrations (>2.0 mmol/l) can lead to seizures, coma, and death. Since lithium toxicity is potentially fatal, any suspicion of intoxication should lead to the immediate withdrawal of lithium treatment and close monitoring of serum lithium and plasma electrolyte and creatinine concentrations. Severely ill patients with high serum lithium levels may require dialysis.

Carbamazepine

Pharmacology

Like certain other anticonvulsant drugs, carbamazepine blocks neuronal sodium channels. The relationship of this effect to its therapeutic actions in affective disorder is uncertain. Similarly to lithium, carbamazepine facilitates some aspects of brain 5-HT neurotransmission.

Indications and use

Carbamazepine is effective in the acute treatment of mania and in the prophylaxis of bipolar affective disorder. It is used in patients who have difficulty tolerating or fail to respond to lithium therapy, when it may be given in combination with lithium.

The dose range of carbamazepine employed to treat patients with affective illness is similar to that used in the treatment of seizure disorders. Initial treatment should be with 100 mg of carbamazepine twice daily, with the dose increased according to tolerance over the next 2 to 4 weeks. The effective dose range in the treatment of bipolar disorder is generally between 600 and 1200 mg daily, although some patients require higher doses. Plasma level monitoring may be used to help avoid toxicity.

Side-effects

Dizziness, drowsiness, and nausea are common early in treatment, particularly with rapid dose titration, but tolerance to these effects usually develops. Persistent ataxia and diplopia may indicate plasma carbamazepine levels in the toxic range. A moderate degree of leucopenia is often seen during carbamazepine treatment and agranulocytosis can occasionally develop, such that it is prudent to monitor the white cell count as well as the carbamazepine level during treatment. Skin rashes are also quite common. Other rarer adverse effects include hyponatraemia and liver cell damage. Circulating thyroid hormone level may be lowered by carbamazepine treatment, but thyroid-stimulating hormone (**TSH**) levels generally remain in the normal range and clinical hypothyroidism is unusual. Carbamazepine can impair cardiac conduction and should be used with caution in patients with cardiovascular disease.

Drug interactions

Carbamazepine increases the metabolism of a number of other drugs, including tricyclic antidepressants, haloperidol, oral contraceptive agents, warfarin, and other anticonvulsants. A similar mechanism may underlie the decline in the plasma carbamazepine level sometimes seen during continued treatment. The carbamazepine level may be increased by erythromycin and by some calcium-channel blockers, such as diltiazem and verapamil. Reversible neurotoxicity has been reported when carbamazepine is combined with lithium.

Sodium valproate

Pharmacology

Valproate is a simple branch-chain fatty acid with a mode of action that is unclear, although there is some evidence that it can slow the breakdown of the inhibitory neurotransmitter g-aminobutyric acid (**GABA**). This action could account for its anticonvulsant properties, but whether it also underlies the psychotropic effects is unclear.

Indications and use

Like carbamazepine, sodium valproate was first introduced as an anticonvulsant. Recent studies have shown that it is clearly effective in the management of acute mania: the drug is licensed for this purpose in the United States, but not in the United Kingdom. In the United States, valproate is widely used in the longer term prophylaxis of bipolar disorder, but there is little evidence for this indication from randomized trials.

Valproate can be started at a dose of 400 to 600 mg daily, which may be increased once or twice weekly to between 1 and 2 g daily. Plasma levels of valproate do not correlate well with either its anticonvulsant or mood-stabilizing effects, but it has been suggested that efficacy in the treatment of mood disorders is usually apparent when plasma levels are above 50 µg/ml.

Side-effects

Common side-effects of valproate include gastrointestinal disturbances, tremor, sedation, weight gain, and transient hair loss. Serious side-effects are rare, but fatal hepatic toxicity has been reported, as has acute pancreatitis. Valproate may also cause thrombocytopenia and inhibit platelet aggregation, and increases in plasma

ammonia have been reported.

Drug interactions

Valproate potentiates the effects of central sedatives. It has been reported to increase the side-effects of other anticonvulsants (without necessarily improving anticonvulsant control). It may increase plasma levels of phenytoin and tricyclic antidepressants.

Antipsychotic drugs

Conventional (typical) and atypical agents

Pharmacology

Antipsychotic drugs, also known as major tranquillizers or neuroleptics, are a group of agents of varied structure that are used to treat schizophrenia and other psychoses. Conventional or typical antipsychotic agents have in common the ability to block dopamine receptors in the central nervous system, and it is likely that their antipsychotic effect is caused by blockade of dopamine D2 receptors in mesolimbic and mesocortical brain regions.

Atypical antipsychotic drugs have been developed more recently. These have a varied pharmacology, but a much lower likelihood than conventional agents of producing extrapyramidal side-effects at therapeutic doses. Some are highly selective dopamine D_2-receptor antagonists with selectivity for mesolimbic dopamine receptors, for example sulpiride and amisulpiride. Others (for example, risperidone, olanzapine, and quetiapine) have high affinities for the 5-HT_2 receptor that exceed their affinities for the D_2 receptor. Finally, clozapine is also a potent 5-HT_2-receptor antagonist but a weak D_2-receptor antagonist, which accounts for its particularly low risk of inducing extrapyramidal movement disorders.

Principal drugs

These are the:

- *conventional (typical) antipsychotic drugs*: chlorpromazine, haloperidol, flupentixol, fluphenazine, loxapine, pimozide, thioridazine, and trifluoperazine;
- *atypical antipsychotic drugs*: amisulpiride, olanzapine, quetiapine, risperidone, sulpiride.

Indications and use

Antipsychotic drugs are used mainly in the management of schizophrenia. They are also used to treat mania and are sometimes given to depressed patients who have psychotic symptoms or who are particularly agitated. Antipsychotic drugs are also used in the management of disturbed behaviour arising from other causes (for example, confusional states), but their use as non-specific tranquillizing agents should (if possible) be limited to short-term use because of potentially serious side-effects.

In the treatment of acute confusional states, haloperidol, in doses of 1.0 to 5.0 mg is often helpful. This can be administered either orally or parenterally, with the dose repeated after an hour if the patient remains disturbed. Cardiovascular and respiratory side-effects are unlikely with this drug, but acute dystonias can occur and should be treated appropriately (Table 5). Antipsychotic drugs such as risperidone are used for the treatment of confused elderly patients. It is worth noting that some groups of demented patients (particularly those with Lewy-body type dementia) may suffer severe extrapyramidal effects from comparatively low doses of antipsychotic drugs.

The treatment of patients with schizophrenia or mania with antipsychotic drugs requires careful monitoring and persistence because the full therapeutic response may be delayed for some weeks. Furthermore, the dose of antipsychotic drug required may vary considerably from patient to patient, and also within the same patient at different stages of the illness. Lower doses of conventional antipsychotic drugs are now employed for the treatment of these disorders, since positron-emission tomography (**PET**) imaging studies have revealed that an adequate blockade of dopamine D_2 receptors can be obtained with oral doses of 5 to 10 mg of haloperidol daily or 200 to 400 mg of chlorpromazine. Higher doses of these agents can produce sedation and behavioural calming, but at the expense of movement disorders and decreased compliance subsequently.

If a patient has responded to an antipsychotic drug it is usual to continue the medication for a number of months into remission. Frequently it is necessary to administer medication on a long-term basis to prevent relapse, in which case the use of long-acting intramuscular preparations will improve compliance. The decanoates of fluphenazine, flupentixol, and haloperidol are most commonly used.

Atypical antipsychotic drugs should be used when patients experience extrapyramidal movement disorders on low doses of typical agents. In addition, clozapine can be effective in up to 50 per cent of patients with schizophrenia whose symptoms have not responded to typical antipsychotic drugs. It is effective in the treatment of both positive and negative symptoms of schizophrenia; the latter often showing a poor response to typical agents. Whether the other atypical agents are effective in treatment-resistant patients is not yet fully clear. Clinical impression is that some patients can indeed show a useful response, but the frequency of a positive outcome may be less than with clozapine.

Side-effects

Movement disorders

Through their blockade of brain dopamine receptors, typical antipsychotic drugs produce a variety of extrapyramidal movement disorders that can mimic signs of basal ganglia disease (Table 5). Many patients exhibit symptoms of parkinsonism very similar to those of the idiopathic disorder, although tremor is less prominent. A side-effect that appears early in treatment is acute dystonia, which can present with abnormal postures or dramatic muscular spasms involving the face and limbs. Laryngeal spasm with respiratory distress can also occur. A history of recent antipsychotic drug use can help avoid misdiagnoses (it is not unusual, for example, for such reactions to be viewed as 'hysterical'). Another movement disorder that patients find very distressing is akathisia, which is a state of motor restlessness, often with agitation and dysphoria. Distinguishing this reaction from symptoms arising from the underlying psychiatric disorder may not be easy.

All these movement disorders may be treated by a reduction in dosage of the antipsychotic drug or by the introduction of anticholinergic medication such as benzatropine. However, anticholinergic drugs should not be prescribed routinely with antipsychotic medication because of the risk of misuse for their euphoriant effects.

Later in treatment, tardive dyskinesia may develop. This consists of involuntary repetitive movements, usually involving the tongue and lips, though other parts of the body may be involved. The condition may be associated with a supersensitivity of postsynaptic dopamine receptors in the basal ganglia. Unfortunately, this disorder cannot be treated easily, and anticholinergic medication may make it worse. If possible, the antipsychotic drug should be stopped, but this decision is often difficult because of the risk of relapse of the psychiatric disorder.

Atypical antipsychotic drugs are much less likely to cause movement disorders. Risperidone, however, is a potent D_2-receptor antagonist as well as a 5-HT_2-receptor antagonist and can produce some movement disorders at the upper end of its dose range (above 4 to 6 mg daily). Whether atypical antipsychotic drugs are generally less likely to cause tardive dyskinesia is not yet clear. The risk does appear to be less with clozapine, olanzapine, and risperidone.

Neuroleptic malignant syndrome

A rare but potentially very serious reaction to antipsychotic drugs is the neuroleptic malignant syndrome (Table 5). This is characterized by fever, rigidity, and altered consciousness, together with tachycardia and labile blood pressure. Laboratory investigations usually reveal a leucocytosis together with markedly raised levels of creatinine phosphokinase. Antipsychotic drug treatment should be withdrawn immediately if the neuroleptic malignant syndrome is suspected. Management in an

intensive care facility may be needed to deal with cardiovascular, respiratory, and renal complications. Treatment with a dopamine-receptor agonist, such as bromocriptine, and the antispasticity agent dantrolene may be beneficial.

Other side-effects of antipsychotic drugs

Antipsychotic drugs, especially chlorpromazine and thioridazine, can produce a variety of side-effects due to blockade of muscarinic receptors and a-adrenoceptors (Table 2). Other side-effects include:

- *endocrine*: elevated prolactin levels, amenorrhoea, and galactorrhoea;
- *skin*: rashes, pigmentation, and photosensitivity (especially phenothiazines);
- *other*: precipitation of seizures, hypothermia (especially chlorpromazine), cardiac arrhythmias (pimozide), weight gain, cholestatic hepatitis, leucopenia, and retinitis pigmentosa (thioridazine in doses >800 mg daily).

The most common side-effects of atypical antipsychotic drugs are shown in Table 6.

Drug interactions

Antipsychotic drugs potentiate the effects of other central sedatives. They may delay the hepatic metabolism of tricyclic antidepressants and antiepileptic drugs, leading to increased plasma levels of these agents. The hypotensive properties of chlorpromazine and thioridazine may enhance the effects of antihypertensive drugs. Antipsychotic drugs, particularly pimozide and thioridazine, can increase the Q–T interval and should not be given with other drugs likely to potentiate this effect, such as antiarrhythmics, astemizole, terfenadine, cisapride, and tricyclic antidepressants. There are also reports of an increased risk of cardiac arrhythmias when pimozide has been combined with clarithromycin and erthyromycin. Clozapine should not be given with any agent likely to potentiate its depressant effect on the white cell count such as carbamazepine, cotrimoxazole, and penicillamine. SSRIs slow the hepatic metabolism and increase blood levels of several antipsychotic drugs, including haloperidol, risperidone, and clozapine.

Antianxiety agents

Benzodiazepines

Pharmacology

Benzodiazepines enhance the action of the neurotransmitter g-aminobutyric acid (GABA) in the central nervous system by binding to a specific benzodiazepine receptor located in a complex with a GABA receptor and a chloride-ion channel. The pharmacological effects of benzodiazepines are attributed to the facilitation of GABA neurotransmission.

Principal drugs

These are alprazolam, chlordiazepoxide, diazepam, flurazepam, lorazepam, lormetazepam, nitrazepam, and temazepam.

Indications and use

The prescription of benzodiazepines is now decreasing following concern about their liability to produce dependence. Alternative therapies are available for most anxiety-related disorders, and it is recommended that the drug treatment of anxiety and insomnia should be limited to a few weeks' duration. The major indication for the use of benzodiazepines is to help patients in a crisis, when generalized anxiety and insomnia are causing functional impairment and reducing their ability to cope. Patients should be advised that the drug treatment will be of short duration to help them manage their immediate difficulties.

All benzodiazepines have hypnotic and anxiolytic properties. The major distinction of clinical importance is their length of action. Derivatives with short half-lives, such as temazepam, are suitable hypnotics because of their relative lack of a hangover effect. Other benzodiazepines, for example diazepam, have long half-lives and are metabolized to active compounds. These may be used for the treatment of anxiety, either in the form of regular dosing, or on the now preferred 'as required' basis with an agreed maximum daily dose.

Side-effects and interactions

Benzodiazepines have a low acute toxicity. Their adverse effects are extensions of their clinical effects and include the following: drowsiness, psychomotor impairment, dizziness, ataxia, and paradoxical aggression (rare). Benzodiazepines potentiate the effects of other centrally acting sedatives, particularly alcohol. The effects of benzodiazepines are potentiated by cimetidine.

Patients who have taken clinical doses of a benzodiazepine for more than a few months may show a withdrawal syndrome when the medication is stopped. In many respects this syndrome resembles an anxiety state, but perceptual disturbances and acute dysphoria may also occur. It is thus apparent that benzodiazepines can cause physical dependence, and, although the withdrawal syndrome is less severe than that seen following the cessation of barbiturates, patients frequently find it extremely difficult to stop their medication. A gradual reduction is usually best. Generally, withdrawal from a long-acting benzodiazepine is easier than from a short-acting preparation, and if patients taking short-acting benzodiazepines have difficulty withdrawing, then a switch to a long-acting preparation may be helpful.

Other drugs that increase brain GABA function

Zopiclone, zolpidem, and zaleplon

Zopiclone is a cyclopyrrolone derivative marketed for the treatment of insomnia. It binds to a site close to the benzodiazepine receptor and thereby facilitates brain GABA function. By contrast to the benzodiazepines, which reduce the amount of slow-wave (deep) sleep at night, zopiclone has little effect on sleep architecture and is relatively free from a daytime 'hangover' effect. It is claimed to be less likely than benzodiazepine hypnotics to produce tolerance and withdrawal effects, but cases have been reported. The most common adverse effects of zopiclone include bitter taste, nausea, dry mouth, irritability, and headache.

Zolpidem is an imidazopyridine derivative, which also binds to a site close to the benzodiazepine receptor. It has a very short duration of action and, like zopiclone, has little effect on sleep architecture or daytime performance. Its possible adverse effects include nausea, dizziness, headaches, and diarrhoea. Zoleplon is a pyrazolopyrimidine derivative with pharmacological properties similar to zopiclone and zolpidem. However, its half-life is only about an hour which means that it can be used to re-induce sleep after nocturnal waking.

Clomethiazole

Clomethiazole also binds at the GABA-receptor complex, but its clinical effects resemble those of barbiturates rather more than those of the benzodiazepines. It can cause serious respiratory depression in overdose, particularly in combination with alcohol. Because of its short half-life (4 to 5 h), clomethiazole is also used as a hypnotic in the elderly, where dependence has not been reported.

Drugs altering monoamine function

Buspirone

Buspirone is a 5-HT_{1A}-receptor agonist structurally unrelated to benzodiazepines. It is effective in the treatment of generalized anxiety disorder (less so in patients previously exposed to benzodiazepines) but has a slow onset of action (1 to 3 weeks). Unlike benzodiazepines, buspirone does not cause significant sedation or

cognitive impairment and appears unlikely to cause dependence. It does not have hypnotic properties. Side-effects include nervousness, dizziness, and headache.

Other drugs

Tricyclic antidepressants and SSRIs are effective in the management of patients with a range of anxiety disorders, including generalized anxiety. They are generally preferred to benzodiazepines for the treatment of agoraphobia with and without panic attacks. SSRIs are also effective in the treatment of obsessive–compulsive disorder and social phobia, but tricyclics (with the exception of clomipramine) are not.

Further reading

Cowen PJ (1997). Pharmacotherapy for anxiety disorders: drugs available. *Advances in Psychiatric Treatment* **3**, 66–71.

Ferrier IN, Tyrer SP, Bell AJ (1999). Lithium therapy. *Recent Topics from Advances in Psychiatric Treatment* **2**, 76–83.

Nutt D, Bell C (1997). Practical pharmacotherapy for anxiety. *Advances in Psychiatric Treatment* **3**, 79–85.

Porter R, Ferrier N, Ashton H (1999). Anticonvulsants as mood stabilisers. *Advances in Psychiatric Treatment* **5**, 96–103.

Richelson E (1999). Receptor pharmacology of neuroleptics: relation to clinical effects. *Journal of Clinical Psychiatry* **60**(Suppl. 10), 5–14.

Spigset O, Martensson B (1999). Drug treatment of depression. *British Medical Journal* **318**, 1188–91.

Stahl SM (2000). *Essential psychopharmacology*, 2nd edn. Cambridge University Press, Cambridge.

Stahl SM (1999). Selecting an atypical antipsychotic by combining clinical experience with guidelines from clinical trials. *Journal of Clinical Psychiatry* **60**(Suppl. 10), 31–41.

26.6.2 Psychological treatment in medical practice

Michael Sharpe and Simon Wessely

What is psychological treatment?

Psychological treatments in medicine come in two main forms. First are formal psychological interventions, which usually have particular labels and content, such as counselling or behaviour therapy, and which are rarely delivered by physicians. Second, and arguably more important, is the doctor–patient interaction itself. This can also be regarded as a psychological treatment; an intervention that has an important and unavoidable psychological impact, whether for good or ill.

Psychological treatment and the medical consultation

The psychotherapeutic importance of the medical encounter was perhaps best acknowledged and described at a time when physicians had less to offer in terms of biological therapies, and consequently placed more emphasis on what they could achieve by talking with the patient. Hence, one less-welcome consequence of the increasing specialization and dependence on technology in modern medicine has been a relative neglect of the psychological aspects of the encounter, most particularly in hospital medicine. Despite the power of modern drug and surgical therapies, there remains a great and possible increasing need for the psychotherapeutically helpful medical consultation. This is especially the case when there is diagnostic uncertainty and when the patient is distressed.

Psychotherapeutic consultations (doing good)

The key ingredients of the psychologically helpful consultation are those described by Jerome Frank as factors common to all psychological treatments, including:

- establishing a good, confiding, and collaborative relationship with the patient;
- giving an explanation to the person of what is wrong with them;
- offering a clearly described plan of action;
- giving a positive message.

Medical encounters are often suboptimal in these general, but important, non-specific factors. The nature of the encounter and its context may not provide an opportunity for the patient to confide in the doctor. The doctor may convey verbally or non-verbally that he does not wish to hear the patient's story or concerns. A positive explanation for the person's symptoms is often lacking and there may be no clear plan of action.

Simple steps can be taken to remedy these shortcomings. It is desirable that attention be paid to the physical arrangements for the consultation. The days when patients were first told that they have cancer on an open ward round may be gone, but many consultations offer scant privacy and opportunity for the patient to ask questions. Some consultations will take time. Whilst to some extent others such as nurses can supplement the doctor's role, it is important that time and privacy are recognized and insisted on as essential therapeutic tools. Taking an interest in the patient's symptoms (even those that are not of diagnostic value) and their fears about these (even if they appear illogical) is essential. Unless this is done the patient may feel they have not been listened to and ignore subsequent advice. It is hard to give effective reassurance if one does not know what the patient fears: time spent on such matters is not therefore a distraction from the diagnostic process, but critical to the overall aim of helping the patient to feel better.

A clear explanation and plan of action is important. All too often patients complain that doctors told them what they didn't have, but not what they did have. This usually happens when the patient's symptoms are not adequately explained by identifiable disease (see Chapter 26.5.3), but even when a clear explanation cannot be given, a positive plan of action usually can. There is evidence that a positive approach has a beneficial effect on outcome.

The provision of hope and an expectation of improvement has long been an ingredient of the doctor–patient relationship. This should not be false hope, for example if the patient has a terminal condition the message should not be a false message that the condition can be cured, but rather that their symptoms will be managed and the doctor will provide ongoing help for them.

Psychological iatrogenesis (doing harm)

Iatrogenesis is not only the result of prescribing the wrong drug or doing the wrong operation: what doctors say to patients can also have powerful negative effects. These include:

- *Dismissive messages*: for example, telling a patient with medically unexplained symptoms 'there is nothing wrong with you—it's “all in the mind”, you are imagining it'. This is not only likely to damage your relationship with the patient, but may also send him or her into the arms of less scrupulous practitioners.
- *Excessively optimistic predictions*: for example, telling a patient who has not yet been adequately assessed 'I'm sure it's nothing serious'. This may lead to the patient losing faith in the doctor who made the predictions (and to legal redress) if it turns out to be incorrect.
- *Excessively negative predictions*: for example, telling a person with possible multiple sclerosis 'its probably best if you come to terms with the idea of a wheelchair'. This is likely to lead to the patient becoming unnecessarily distressed.
- *Ill-considered or unhelpful explanation for the illness*: for example, telling the person who is depressed that it's 'probably a virus'. This can set them off on the wrong track for how they should cope and what treatment they should seek.
- *Poorly thought out or ill-informed advice*: for example, telling a patient with indigestion to avoid all foods that are associated with the symptoms. For some patients this can fuel excessive dietary restriction. An elegant demonstration of this type of problem was found in a study of schoolchildren whose parents were told (sometimes incorrectly) that their children had abnormal hearts and should avoid exertion: at follow-up the children with normal hearts were as disabled as those with heart disease.

Specific psychological treatments

Specific psychological treatments may be broadly divided into simple counselling interventions, more intense but short-term psychotherapeutic interventions such as cognitive–behavioural therapy, and longer-term treatments such as long-term psychotherapy. These have a potentially important role in medical practice in improving quality of life and adherence to recommended treatments, and there is some evidence that they may even improve survival. Certainly there is empirical support for better integration of psychological therapies into standard medical care.

Counselling

Simple counselling may have a role in helping people express distress and talk through acute problems, for instance adjustment to a worrying diagnosis such as that of cancer. Whereas the provision of reassurance and emotional support is clearly important, whether or not this qualifies as a specific 'treatment' or needs to be given by a mental health professional is doubtful. Rather it should be regarded as a generic skill to be possessed by all health workers.

Short-term specific psychological therapies

Short- and medium-term specific therapies such as cognitive–behavioural therapy and interpersonal therapy are of documented effectiveness. Both are as effective as antidepressant medication in the treatment of depression. Cognitive–behavioural therapy is also of value in the treatment of patients with anxiety and panic disorders and has a specific role in those with unexplained somatic symptoms (see Chapter 26.5.3). It has been shown in randomized trials to be of benefit in patients with chronic unexplained pain, chronic fatigue syndrome, hypochondriasis, irritable bowel syndrome, non-cardiac chest pain, and other poorly understood conditions.

Cognitive–behavioural therapy is based on a collaborative relationship. The cognitive aspects refer to the patient being helped to re-evaluate and optimize their understanding of their illness. The behavioural part involves some form of target setting, activity scheduling and trying out new ways of coping. Interpersonal therapy is similar, but focuses on the person's relationships and social roles, rather than on their thoughts and behaviours.

Long-term psychotherapy

For persons with problems not amenable to brief therapy, such as ongoing difficulty in adjusting to disease and/or personality problems, there is a case for longer term psychological therapy, but the availability of such therapy is very limited.

Making a referral for specialist psychological treatment

The first requirement for making a referral for specialist psychological treatment is to ensure that there is a service available and to find out what types of referrals are accepted. The second is to make sure that the patient understands why they have been referred and to establish that they are likely to attend.

Services

Psychological treatment services will ideally be located in organizational and geographical proximity to the medical consultation. In reality they often are not. It is therefore desirable for the physician to familiarize themselves with what is available, how long the patient will have to wait, and how they will be received before the need to make a referral arises.

Making the referral

It is often helpful to discuss the referral with the service to ensure it is appropriate. For example, the psychological problems may appear obvious to the physician but regarded as untreatable personality characteristics by the service being referred to. If medical investigation or treatment is ongoing it will help if uncertainties are made explicit in the referral letter and new findings communicated as they arise.

Explanation to the patient

The first and perhaps most important aspect of explaining such a referral to the patient is to make is clear to them that you are not implying that their illness is 'all in the mind', but rather that there is a psychological aspect that deserves attention. The second is to convey a positive attitude towards psychotherapy as being a sensible approach with a realistic chance of helping them. It helps if you have some knowledge of what they can expect. Finally, it can be important, if appropriate, to indicate that you will see the patient after the psychological treatment, meaning that you do not simply regard this as a way of disposing of them.

Summary

There has been an unfortunate separation of biomedical and psychosocial aspects of patient management in our medical system. This is maintained by a number of factors, including geographical separation of services and the short time available for medical consultations. None the less, all medical interactions have an inescapable psychological component and the potential to help or harm the patient. There are basic aspects of a consultation that will maximize the opportunity of doing good; there are other ways of handling consultations that may do harm and should be avoided.

Specific psychological treatments have an important role, especially for those who are distressed in relation to a medical condition, and for those who have unexplained physical symptoms. Whilst there is a general supportive role for counselling, most of the evidence of effectiveness is in support of short- and medium-termed structured psychological treatment such as cognitive–behavioural therapy. A small number of patients, such as those with long-standing problems adjusting to an illness or personality problems, might benefit from longer term psychotherapy, although the evidence for the effectiveness of this is limited.

Further reading

Andrews G (1996). Talk that works: the rise of cognitive behaviour therapy. *British Medical Journal* **313**, 1501–2.

Clark DM, Fairburn CG (1997). *Science and practice of cognitive behaviour therapy*. Oxford University Press, Oxford.

Frank JD (1967). *Persuasion and healing*. Johns Hopkins Press, Baltimore, MD.

Guthrie E (1996). Emotional disorder in chronic illness: psychotherapeutic interventions. *British Journal of Psychiatry* **168**, 265–73.

Kroenke K, Swindle R (2000). Cognitive–behavioral therapy for somatization and symptom syndromes: a literature synthesis. *Psychotherapy and Psychosomatics* **69**, 205–15.

Price JR (2000). Managing physical symptoms: the clinical assessment as treatment. *Journal of Psychosomatic Research* **48**, 1–10.

Spiegel D (1999). Healing words: emotional expression and disease outcome. *Journal of the American Medical Association* **281**, 1328–9.

Thomas KB (1987). General practice consultations: is there any point being positive? *British Medical Journal* **294**, 1200–2.

26.7.1 Alcohol and drug dependence

Mary E. McCaul and Gary S. Wand

Diagnosis

In the 1970s, Edwards and Gross first characterized the alcohol-dependence syndrome. Today, this syndrome is the basis of alcohol- and other drug-dependence criteria for both major diagnostic systems: the *World Health Organization international classification of diseases, tenth revision* (**ICD-10**) and the *American Psychiatric Association diagnostic and statistical manual of mental disorders, fourth edition* (**DSM-IV**). For a diagnosis of dependence, both systems require a clustering of at least three of the following symptoms during a 12-month period:

1. tolerance (i.e. increasing amounts of drug needed to achieve the desired effect);
2. characteristic physiological withdrawal;
3. difficulties in controlling onset, termination, or amount of substance use;
4. neglect of important social, occupational, or recreational activities because of substance use;
5. increased time required to obtain, use, or recover from substance use; and
6. continued use despite persistent negative physical or psychological consequences.

In DSM-IV, dependence is subtyped for the presence/absence of physiological withdrawal and tolerance, and recovery status is characterized for duration (early versus sustained remission) and symptom status (partial versus full remission). International studies have found remarkable consistency in the dependence syndrome across diverse geographical and cultural settings, suggesting that fundamental biological processes underpin the disorder.

By contrast, cross-cultural reliability has not been established for the less severe diagnoses of harmful use (ICD-10) and substance abuse (DSM-IV), defined by the negative consequences of alcohol and drug use. This may stem from the greater subjectivity of defining 'harm' within a particular social and cultural context, compared to the more physically based dependence symptoms of withdrawal and tolerance. Similarly, specific levels of hazardous alcohol consumption have not been defined: acceptable amounts vary across cultures, time, and individuals as a function of health, pregnancy, age, and gender.

It is sometimes difficult to disentangle the signs and symptoms of substance abuse and dependence from other common psychiatric and medical conditions. Given their prevalence, alcohol and other drug use should be explored carefully with every patient; there are few socioeconomic, racial/ethnic, or educational predictors of who may experience these problems. Patients should be queried for substance use versus abstention, recent, lifetime, and heaviest use patterns, and finally any problems associated with use. Particular attention should be given to those who report family members with alcohol or drug problems. Additionally, certain behaviours often suggest hazardous levels of substance use; these include cigarette smoking, missing work, neglecting family responsibilities, poor nutrition and hygiene, and high rates of injury and accidents.

This chapter focuses on the harmful use of alcohol, stimulants, and opioids because of their prevalence and severe medical and psychosocial consequences. Although of importance, marijuana, nicotine, sedative, hallucinogen, and inhalant use will not be covered (some of these are discussed in Chapter 26.7.3, and for review, see APA 1994).

Epidemiology of alcohol and drug use, abuse, and dependence

There is considerable cultural variation in alcohol- and drug-use patterns and definitions of harmful or pathological levels of consumption. In North America and Europe the use of alcohol is widespread: in a recent multinational study, approximately one-quarter of primary healthcare patients reported at least one alcohol-related problem in the last year. The North American and European lifetime prevalence of alcohol dependence is approximately 9 per cent (range: 5.5 per cent in The Netherlands to 14.3 per cent in the United States) and the prevalence of drug dependence is about 4 per cent (range: 0.7 per cent in Mexico to 7.5 per cent in the United States). Alcohol-related problems account for approximately 4 per cent of the total burden of disease and injury in the world.

In comparison to alcohol use, the prevalence of drug use in Europe and North America is considerably lower, with approximately one-third of persons reporting lifetime drug use. Globally, the use of substances other than alcohol is generally considered socially aberrant, and rates of use remain low. In the United States there was a large increase in the prevalence of cocaine use during the 1970s and 1980s, and currently about 10 per cent of people report lifetime cocaine use: 2 per cent report use in the past year and 1 per cent report current use. In the United States less than 1 per cent of the population reports heroin use in the past year. Use of most illicit drugs peaks in young adults and then declines with age.

Aetiology of alcohol and drug dependence

Many people have tried alcohol or drugs, but only a subset develop dependence, with an interplay of genetic, psychological, and environmental factors increasing dependence vulnerability. A family history of alcohol or drug dependence remains one of the strongest predictors of risk, with heritability estimates for alcohol dependence ranging from 45 to 65 per cent for men and women. Genetic factors associated with an increased risk for alcohol dependence may include inborn abnormalities in dopamine, opioid, and serotonin neurotransmitter systems.

Drugs of abuse influence several different brain neurotransmitter systems, many of these primary responses leading to secondary effects involving dopamine. For example, morphine and heroin first bind to opioid receptors, which then increase the activity of midbrain mesolimbic dopamine neurones that send projections to interconnected forebrain structures such as the prefrontal cortex and striatum. The nucleus accumbens, a region at the base of the striatum, is the key zone that mediates the rewarding effects of drugs such as amphetamine and cocaine, which act directly by increasing dopamine levels at this site, as does ethanol ingestion. Opioid antagonists block ethanol-induced release of nucleus accumbens dopamine, implicating opioidergic activity as an intermediary between ethanol exposure and dopamine release. Genetic, as well as psychological and environmental effects, on this key mesolimbic system may contribute to individual differences in dependence vulnerability.

Psychological factors also appear to increase a person's vulnerability for alcohol or drug dependence. A recent international study found strong associations between anxiety and mood disorders and substance-use disorders, despite large differences in their prevalence across study sites. Anxiety disorders were very likely to pre-date the onset of the substance-use disorders, suggesting aetiological significance. By contrast, mood disorders did not indicate as strong a temporal relationship. In a recent United States household survey, over three-quarters of alcohol-dependent persons were diagnosed with at least one additional psychiatric disorder, with anxiety disorders having the highest prevalence, 61 per cent of women and 36 per cent of men meeting lifetime diagnostic criteria. Some types of people are at increased risk for alcoholism as a result of antisocial tendencies and high excitement-seeking behaviours. Antisocial personality disorder increases the risk of alcohol dependence approximately fourfold for men and over fivefold for women. Cross-cultural consistency in patterns of comorbidity suggests that, while cultural factors may influence the availability and type of substance exposure, the associations between psychopathology as risk factors and the sequelae of substance disorders are probably independent of particular cultural norms and standards.

Finally, environmental and social processes—including marital discord, parental hostility, poor parental monitoring, and high parental tolerance of adolescent drinking—influence a person's vulnerability for alcohol and drug dependence.

Laboratory diagnosis of alcohol- and drug-use disorders

A diagnosis of alcohol or drug dependence is not based primarily on laboratory findings, but rather on the behavioural manifestations of uncontrolled substance use. None the less, laboratory assessments can be useful for confirming recent alcohol or drug use, supporting other evidence of regular use, and providing information about alcohol- and drug-related physical problems. For patients enrolled in treatment, periodic testing can be used to monitor their progress and encourage accurate verbal reports of recent use.

To confirm recent use, measurements are made of current alcohol or drug concentrations in blood, urine, or other body fluids. Blood alcohol levels are measured in milligrams of alcohol per decilitre of blood and can be readily obtained from breath or blood samples. Impairment from alcohol is common above 50 mg/dl and climbs steeply as the blood alcohol level reaches 100 mg/dl and higher: a very elevated value without significant impairment indicates high alcohol tolerance resulting from chronic, heavy drinking. Urine or blood toxicology screens are commonly effective in detecting most illicit drugs for up to 72 h following use.

Most markers (including liver enzymes—aspartate aminotransferase (**AST**), alanine aminotransferase (**ALT**), gamma-glutamyl transferase (**GGT**)—and estimation of macrocytosis, mean cell volume (**MCV**)) have not provided sufficient sensitivity and specificity for use in widespread screening for alcohol problems. Various non-alcohol-related diseases produce similar changes in these markers to those produced by excessive alcohol use. Recent findings indicate that an increase in carbohydrate-deficient transferrin (**CDT**) levels is highly specific for heavy alcohol use (>50 g ethanol/day for at least 1 week), although its usefulness in women or for screening heterogeneous populations of drinkers may be limited.

Alcohol pathology, treatment, and pregnancy complications

Alcohol-related pathology

Age-adjusted morbidity and mortality rates increase as a function of the amount and duration of alcohol consumption, and these rates are increased two- to threefold among chronic, heavy drinkers. Alcohol consumption affects virtually every major organ system. The primary causes of excess mortality include liver disease, severe respiratory infections, cancer of the upper respiratory and digestive systems, cardiovascular disease, suicides, and violence.

Gastrointestinal system

Heavy alcohol use is associated with acute abdominal pain, nausea, and vomiting. With regular, heavy alcohol use, a variety of medical complications involving the gastrointestinal tract and related organ systems can develop. Oesophageal disorders include oesophagitis, oesophageal varices, and oesophageal mucosal tears with bleeding. Common upper gastrointestinal symptoms are gastritis, duodenitis, and ulcer disease. Some of these effects result from direct mucosal irritation by alcohol.

Liver

Because the liver receives portal blood directly from the intestines and is the primary site of alcohol metabolism, liver damage is one of the most common health consequences of chronic, heavy drinking. Two main types of alcohol-related liver injury are inflammation (alcoholic hepatitis) and progressive scarring (fibrosis or cirrhosis), the mechanisms of the toxic effects being shown in Table 1. Despite multiple pathways for alcohol-induced hepatic injury, only some chronic, heavy drinkers experience serious liver damage. Vulnerability to liver injury may result from genetic variations in the enzymes that metabolize alcohol (ADH and **ALDH** (aldehyde dehydrogenase)) and in cytochrome P4502E1 activity. Women experience higher rates of hepatic injury at lower cumulative alcohol levels than men. This been attributed to a toxic interaction of female sex hormones and alcohol-metabolizing enzymes, also to lower levels of gastric ADH in women compared to men, resulting in reduced first-pass metabolism in the stomach and higher blood alcohol levels following equivalent alcohol doses. Hepatic injury is also facilitated by nutritional factors, including depletion of antioxidant vitamins and glutathione and a diet high in polyunsaturated fats or iron. Finally, other medical conditions including infection with hepatitis B and C viruses are known to increase the risk of liver damage in alcoholics. These issues are discussed in more detail in Section 14.20.

Cardiovascular

In healthy individuals, moderate alcohol use reduces mortality from atherosclerotic cardiovascular disease, due in part to alcohol's effects of decreasing low-density lipoprotein (**LDL**) and increasing high-density lipoprotein (HDL) cholesterol. By contrast, heavy drinking is associated with an increased risk for cardiac arrhythmias, cardiomyopathy, and sudden cardiac death. Heavy alcohol consumption is also associated with systolic and diastolic hypertension, significantly increasing the risk of stroke by 250 to 450 per cent. In those with established arrhythmias, hypertension, or hyperlipoproteinaemia, even moderate alcohol use may aggravate symptoms.

Pulmonary

At high doses, alcohol decreases respiratory rate, airflow, and oxygen transport, hence increasing pulmonary disease symptoms in affected patients. Alcohol also reduces key pulmonary defences against infection, including: mucociliary clearance; macrophage mobilization, killing, and clearance; and phospholipid metabolism. These actions directly contribute to the increased rates of pulmonary infections (e.g. pneumococcal and Gram-negative pneumonias) in chronic, heavy drinkers.

Neurological

Chronic, heavy alcohol consumption causes structural changes in the brain, particularly in the cerebellum, limbic system, diencephalon, and cerebral cortex. Enlargement of the ventricles and widening of the fissures and sulci over the cerebral hemispheres suggest cortical atrophy. Severely dependent patients may experience a significantly decreased blood flow in the frontal, cortical, and periventricular regions of the cerebral cortex.

A variety of cognitive deficits have been associated with regular, heavy alcohol use, including slowed information-processing, poor attention, difficulties with abstraction, solving problems, and learning new information, and reduced visuospatial abilities. Chronic irreversible damage includes ataxia and gait disturbances, polyneuropathy, dementia, and the Wernicke–Korsakoff syndrome. These are discussed further in Chapter 24.15.

Endocrine

Endocrine abnormalities result from the direct toxic effects of chronic, heavy alcohol use and indirect effects associated with alcohol-related liver disease and malnutrition. Chronic alcohol exposure is particularly damaging to the gonadal axis, resulting in impaired sex hormone production. Male alcoholics often develop gynaecomastia, impotence, and testicular atrophy. Alcohol-dependent women often develop menstrual abnormalities. Both sexes have a higher incidence of osteoporosis resulting in part from reduced sex hormone levels.

Chronic, heavy alcohol use is also associated with activation of the hypothalamic–pituitary–adrenal axis (**HPA**), especially during acute alcohol withdrawal. Some

alcoholics develop clinical and biochemical features of Cushing's syndrome or 'the psuedo-Cushing syndrome'. By contrast, HPA responsiveness is temporarily dampened following alcohol withdrawal. As alcohol-dependent individuals cycle through periods of intoxication and withdrawal, the HPA cycles through hyper- and hypoactivity. This alcohol-induced cyclical pattern of corticotropin-releasing factor (**CRF**) and cortisol secretion may induce various pathological states, for instance episodes of sustained hypercortisolism may exacerbate osteoporosis, diabetes mellitus, and hypertension, as well as impairing growth, reproductive ability, and immune function. Further, hypercortisolism accompanying alcohol withdrawal increases excitatory amino acid levels within the central nervous system, resulting in neurotoxicity and worsening withdrawal symptoms such as seizures.

Other alcohol-related endocrine abnormalities are shown in Table 2.

Cancers

Heavy alcohol consumption significantly increases the risk of oesophageal cancers through the local actions of alcohol-metabolizing enzymes on oesophageal cells, and by the increased production of cytochrome P4502E1 in the oesophageal mucosa. This risk is considerably increased by smoking, which has a strikingly high prevalence in heavy drinkers. Other cancers increased by chronic heavy alcohol use include breast, thyroid, skin, laryngeal, and nasopharyngeal. Compromised immune function associated with heavy drinking may contribute to these elevated cancer rates.

Injury

Accidental injuries are a major cause of increased morbidity and mortality among chronic, heavy drinkers. Alcohol use has been implicated in 15 to 63 per cent of fall fatalities, 33 to 61 per cent of burn fatalities, and 44 per cent of fatal traffic accidents. In a study of emergency-room patients admitted for blunt or penetrating trauma, almost half had a positive blood alcohol level (as do half of all those who die from unintentional injuries).

Treatment of alcohol disorders

Inpatient settings have traditionally been used for the early phases of treatment, particularly acute detoxification and short-term residential programmes. Outpatient settings have provided longer term, abstinence-maintenance treatment. With growing concern over the cost-effectiveness of services, outpatient utilization has increased across all treatment phases.

Alcohol-withdrawal management

Alcohol withdrawal is potentially life-threatening. However, fewer than 10 per cent of alcohol-dependent patients are at risk for severe withdrawal symptoms, which: appear within the first 24 h after drinking cessation; intensify; and then decrease over 2 to 3 days. Symptoms generally result from disinhibition of the g-aminobutyrate (**GABAergic**) system and the resulting overactivation of the autonomic nervous system. Primary symptoms include tremor, sweating, headache, restlessness, anxiety, nausea, vomiting, disorientation, hallucinations, and seizures. The level of alcohol consumption typically determines symptom severity, but comorbid medical conditions can exacerbate the problem and complicate its management.

Repeated, unmedicated withdrawal episodes may increase the risk of future alcohol withdrawal seizures (kindling), and hence withdrawal management using benzodiazepines is recommended for most patients. Symptom-driven protocols that base medication frequency and dosage on regular withdrawal severity assessments are increasingly preferred over traditional fixed-dose regimens. These decrease the total amount and duration of medication and withdrawal-symptom severity. For most patients, alcohol withdrawal can be successfully treated on an outpatient basis. (See Chapter 26.7.3 for further discussion.)

Pharmacotherapy for alcohol dependence

There is considerable interest in developing medications to use in conjunction with psychosocial therapies to improve treatment retention and reduce relapse. Until recently, the only pharmacotherapy for alcoholism was disulfiram (Antabuse), an alcohol-sensitizing medication that produces flushing, nausea, vomiting, increased blood pressure, and heart rate when combined with alcohol. However, there is limited empirical support for the effectiveness of disulfiram because of poor patient acceptance and compliance.

As described above, various neurotransmitter systems contribute to alcohol reward, tolerance, and dependence, so providing an empirical basis for the development of pharmacotherapies to treat alcohol craving or reduce alcohol reward or intoxication. If persistent and intrusive thoughts about drinking (in other words, craving) and within-treatment 'slips' could be effectively treated, this would improve patients' retention in psychosocial services and decrease their risk of relapse. In clinical trials, opioid receptor antagonists (such as naltrexone) have decreased craving and alcohol consumption, whilst improving objective markers of drinking such as liver function test results. Of particular interest, placebo-treated subjects who drank were far more likely to progress to persistent heavy drinking than naltrexone-treated subjects who drank. Acamprosate, an *N*-methyl-D-aspartate (**NMDA**)-receptor antagonist, also increases abstinence and improves treatment retention; it is widely marketed in Europe and is currently under a Food and Drug Administration (**FDA**) review for alcoholism treatment in the United States.

Alcohol-dependent patients with a concurrent psychiatric disorder are at an increased risk of substance-abuse treatment non-compliance, relapse, psychosocial and interpersonal problems, greater severity of psychiatric symptoms, and suicide attempt and completion. The psychiatric disorder often requires treatment concurrently with the substance-use disorder and may not improve simply as a result of reduced alcohol and drug use (for example, serotonergic medications may be effective in reducing drinking in alcohol-dependent patients with concurrent major depression, possibly secondary to improvements in mood and overall psychosocial functioning). A newer non-benzodiazepine anxiolytic, buspirone, has also been found to increase substance-abuse treatment participation and duration **,** and improve outcomes on measures of anxiety, depression, and global psychopathology compared to placebo treatment.

Psychosocial treatments

Most substance-abuse patients are treated in outpatient rather than inpatient facilities (that is, detoxification, residential rehabilitation, therapeutic communities). Outpatient programmes typically offer assessment and diagnosis, individual and group therapy, and referral for other needed services. Recently, outpatient services have increasingly introduced focused, empirically validated treatment interventions, such as cognitive–behavioural skills training programmes, relapse-prevention groups, and marital and family therapy. Family involvement in treatment and patient participation in community-based, self-help programmes such as Alcoholics Anonymous (**AA**) can improve long-term, post-treatment outcomes.

Alcohol-related pregnancy complications

Although many women reduce unhealthy behaviours, including alcohol and drug use and cigarette smoking during pregnancy, approximately one in five women in the United States reports drinking alcohol while pregnant. Among pregnant drinkers, 8 per cent drink on 6 or more days during the month, and 30 per cent consume three or more drinks per drinking day, hence alcohol use remains high during pregnancy despite well-documented and publicized maternal and fetal risks. Because no safe alcohol limits have been established during pregnancy, women are targeted for early intervention at lower alcohol-use levels during pregnancy.

At its most extreme, alcohol consumption during pregnancy can result in the fetal alcohol syndrome, characterized by facial dysmorphology, growth retardation, and CNS disorders, which is the most common preventable cause of mental retardation. Other more common fetal effects have been labelled 'alcohol-related birth defects', and include: reductions in weight, height, and head circumference; decreased cognitive abilities and school achievement; and, possibly, an increased risk of behavioural problems such as attention deficits and impulsiveness. The current global estimated incidence of fetal alcohol syndrome is approximately 1 in 1000 live births in the general obstetric population and 4.3 per 100 births among heavy drinkers.

Pathology, treatment, and pregnancy complications of stimulant drugs

Stimulant-related pathology

Stimulants including amphetamines and cocaine work primarily on three CNS neurotransmitters, norepinephrine (noradrenaline), serotonin, and dopamine. They acutely increase norepinephrine neuronal activity by increasing presynaptic synthesis and blocking reuptake from the synaptic cleft. Effects of which include

hypertension, tachycardia, dilated pupils, diaphoresis, vasoconstriction, and tremor. Norepinephrine levels in brain decrease in the longer term, leading to depression, confusion, restlessness, suicidal ideation, and irritability. Stimulants also acutely diminish CNS serotonin activity, contributing to insomnia. Effects on dopamine transmission, particularly in ventral tegmental/corticomesolimbic regions, are thought to mediate the high potential of stimulants for abuse. Acutely, dopamine transmission is increased, but chronic cocaine use depletes central dopamine levels, such that short-term euphoria, increased energy, and alertness is followed by depression and subsequent drug administration.

Drug-related pathology often results from the method of administration or lifestyle issues (for example, prostitution) rather than direct drug toxicity. Chronic intranasal administration increases rhinitis, maxillary sinusitis and necrosis, and perforation of the nasal septum. Smoked cocaine increases pulmonary complications including pneumonitis, obliterative bronchiolitis, asthma, and pulmonary haemorrhage. Finally, intravenous administration increases the risk of infectious diseases, including human immunodeficiency virus (**HIV**) infection, endocarditis, hepatitis, and sepsis, as well as abscesses and cellulitis at the injection site.

Cocaine use is associated with a number of other serious medical problems, particularly cardiovascular complications, including cardiac arrhythmias, myocardial infarction, myocarditis, cardiomyopathy, endocarditis, and aortic dissection. Myocarditis was present in approximately 20 per cent of regular cocaine users in one autopsy study.

A serious neurological complication of cocaine use and a leading cause of cocaine-associated deaths is seizures. The seizure threshold decreases with repeated cocaine use and associated kindling. Additionally, acute cocaine use increases the risk of stroke secondary to focal artery vasospasm, thrombosis, and elevated blood pressure. Chronic use can result in significant hypoperfusion in the frontal, periventricular, and temporoparietal areas, changes that have been linked to deficits in attention, concentration, learning, visual and verbal memory, and visuomotor integration.

Stimulant use can lead to significant psychiatric symptoms, both acutely and chronically. Acute symptoms include grandiosity, impulsiveness, and aggression. Common complications of chronic use include panic attacks, paranoia, depression, delirium, and hypersomnia.

Medical problems are more frequent among female than male drug abusers. Common complaints include infections, anaemia, sexually transmitted diseases (particularly gonorrhoea, trichomonas, and chlamydia), hepatitis, urinary tract infections, and gynaecological problems. Substance-abusing women are at an increased risk of developing a variety of reproductive dysfunctions compared with other women, including amenorrhoea, anovulation, luteal-phase dysfunction, ovarian atrophy, spontaneous abortion, and early menopause.

Treatment of stimulant-use disorders

Withdrawal management

Stimulant withdrawal is characterized by primarily psychological rather than physical symptoms. The acute 'crash' following discontinuation of intense cocaine use is notable for depression, agitation, cocaine craving, and hypersomnia. Prolonged withdrawal symptoms include anhedonia, anergia, and cocaine craving.

Pharmacotherapy

There has been little progress in identifying effective pharmacological treatments for stimulant dependence and, as a result, psychosocial treatments remain the mainstay of care. Because chronic cocaine use depletes brain dopamine levels, pharmacotherapeutic research has focused predominantly on dopamine agonists. Antidepressants have also been explored for the treatment of dysphoric symptoms accompanying cocaine cessation.

Psychosocial treatments

Retention for longer rather than shorter periods in outpatient treatment is associated with improved long-term, postdischarge outcomes across alcohol, cocaine, and opiate treatments. For cocaine-dependent patients receiving outpatient psychosocial treatments, retention in care for 90 days or longer is associated with a reduced risk of cocaine relapse. Abstinence-focused treatments appear most effective for promoting long-term recovery.

Stimulant-related pregnancy complications

Approximately 2 per cent of pregnant women in the United States report illicit drug use. Across a variety of drug classes, drug-abusing women experience a clinically significant increase in obstetric complications compared with non-drug-using women. Maternal cocaine abuse is associated with intrauterine growth retardation, decreased birth weight, length and head circumference, anaemia, increased risk of intrauterine or perinatal infections (hepatitis, sexually transmitted diseases (**STDs**) and HIV), and cardiac abnormalities. Maternal complications specifically associated with cocaine use include reduced maternal weight gain, precipitous delivery, placental abruption, preterm labour and delivery, and spontaneous abortion. Drug-using women generally participate in fewer prenatal services than non-drug-using women, compromising the delivery of adequate prenatal care for these more complicated pregnancies. Maternal drug use also adversely affects infant health, the most obvious neonatal complication associated with maternal stimulant abuse being drug withdrawal, including tremors, irritability, high-pitched and excessive crying, poor feeding, and abnormal sleep patterns. The overall morbidity for drug-exposed neonates is over twice that of non-drug-exposed neonates.

Pathology, treatment, and pregnancy complications of opioid drugs

Opioid-related pathology

As described above for stimulants, the pathology associated with regular heroin use often results from the method of drug administration or lifestyle issues. Approximately one-quarter of heroin-dependent individuals die from homicide, suicide, accidents, and infectious disease within 10 to 20 years of initiating regular use.

Infections including abscesses, cellulitis, endocarditis, and septicaemia are common among opioid users: many stem from intravenous drug administration, the use of unsterile injection equipment, and contaminated drugs.

Infection with the hepatitis C virus, associated with an increased mortality rate, is a growing concern among intravenous drug users. In the United States, as many as two-thirds of patients seeking drug treatment are positive for hepatitis C; hepatitis B is also fairly common. Both contribute to the liver dysfunction frequently observed in these patients.

More than 30 per cent of adult and 50 per cent of paediatric **AIDS** (acquired immunodeficiency syndrome) cases in the United States are directly or indirectly associated with intravenous drug use, and in both the United States and Europe over one-third of drug-treatment patients are infected with HIV. As a result, more-aggressive, harm-reduction strategies are recommended by some to reduce the spread of HIV, and needle-exchange programmes in which drug-dependent persons can exchange used for new injection equipment are increasing. Other sexually transmitted diseases including syphilis, gonorrhoea, trichomonas, and chlamydia also are common, especially among female drug users.

Tuberculosis (**TB**) emerged in the 1990s as a significant concern among intravenous drug users, with its rising prevalence attributed to increased susceptibility to TB infection among HIV-positive, immunocompromised drug users, and the emergence of antibiotic-resistant strains of the tubercle bacillus.

Treatment of opioid-use disorders

Withdrawal management

Opioids produce a very characteristic withdrawal syndrome, although the speed of onset and duration vary as a function of the half-life of the particular drug. Heroin or morphine withdrawal begins 8 to 12 h following the last dose and subsides over 5 to 7 days. Methadone withdrawal begins 12 to 16 h after the last dose, peaks in approximately 3 days, and can persist for 3 weeks or longer. Typical symptoms include: gastrointestinal distress such as cramping, diarrhoea, nausea, vomiting; arthralgias or myalgias; increased respiratory rate; lacrimation; yawning; rhinorrhoea; piloerection; anxiety, restlessness and irritability; tachycardia and hypertension.

Clonidine, marketed for the treatment of hypertension, is widely used for opioid-withdrawal management and is mainly effective for alleviating milder withdrawal symptoms and anxiety-related complaints. However, because of the high morbidity and mortality associated with heroin use, opioid-substitution therapy is the treatment of choice in the United States and Europe, and generally recommended for opioid-dependent people who have failed in prior drug-free treatments. Opioid-substitution therapy replaces the use of illicit, short-acting, injected, impure drugs with longer and orally active medications of known potency and purity. Outcomes include withdrawal symptom suppression, reduced drug craving, blockade of the euphoric effects of heroin or other opioid administration, and patient engagement in rehabilitative services. Individual and group counselling and urine testing for illicit drugs are part of routine care in most programmes. HIV/AIDS has heightened the importance of treatment approaches focused on reducing or eliminating injection drug use as opposed to achieving drug-free status. Methadone is commonly used for short- and long-term management of heroin and other opioid withdrawal. Buprenorphine, a mixed agonist and antagonist drug, may offer particular advantages for opioid-withdrawal management, although it does not have current United States approval. The details of pharmacotherapeutic management are discussed in Chapter 26.7.3.

Because of the breadth of psychosocial problems associated with chronic opioid use, individual and group counselling sessions are standard treatment components. Patients evidence their greatest improvements in comprehensive care programmes that integrate medical, psychiatric, family, legal, and social services.

Opioid-related pregnancy complications

Complications associated with regular heroin use include spontaneous abortion, amnionitis, chorioamnionitis, intrauterine growth retardation, placental insufficiency, postpartum haemorrhage, pre-eclampsia, premature labour, premature rupture of membranes, eclampsia, toxaemia, and placental abruption. Symptoms of neonatal drug withdrawal are very similar to those of adults, and include CNS hyperirritability, loose stools and other gastrointestinal dysfunction, nasal stuffiness, yawning, sneezing, increased lacrimation, and fever. As described above for stimulants, there is a high frequency of other fetal/neonatal complications, including suboptimal APGAR scores, low birth weight, prematurity, and an increased risk of intrauterine or perinatal infections.

Maternal and infant morbidity and mortality are significantly improved among women receiving opioid-replacement treatment, a major benefit for the fetus being the prevention of repeated, medically unsupervised withdrawal episodes. Maintenance has also been shown to improve the frequency of prenatal care, decrease illicit drug use, improve nutritional status, and decrease risky lifestyles. Infants of methadone-maintained mothers need to be closely monitored and, in most cases, treated for opiate withdrawal, typically using paregoric. Symptom onset is often delayed and duration may be unusually prolonged because of reduced metabolic clearance.

Prevention and early intervention

Given the high prevalence of alcohol and drug disorders, all physicians treat patients with these problems. It is estimated that as many as 20 per cent of primary care patients and 25 to 40 per cent of inpatients are alcohol-dependent or problem drinkers. Physicians should routinely enquire about drinking and other drug use (prescribed and illicit), and facilitate the referral and management of patients with these problems. Most alcohol-related health consequences are experienced not by those who are alcohol-dependent, but by the much larger group of hazardous drinkers whose problems are at an early or relatively mild stage. Brief advice on setting safe drinking limits can have a significant and sustained impact on drinking levels in this at-risk population. These interventions begin with a focused assessment of alcohol and drug use and related problems, followed by a brief, highly directive interaction in which the health professional provides personalized feedback based on the assessment findings (for instance, elevated liver function test results or other medical problems, absenteeism or lateness at work, marital distress). Finally, the provider offers specific drinking-reduction strategies, such as goal-setting for 'sensible' drinking and assigning relevant reading materials. Brief interventions reduce overall alcohol use, binge-drinking episodes, and hospital admissions and are easily conducted in various healthcare settings, including that of primary care. (For further discussion see Chapter 26.7.2.)

Areas of uncertainty/controversy

Controlled alcohol use versus abstinence

Controlled drinking as a therapeutic goal for alcoholism treatment remains controversial, particularly in the United States. Less than one-quarter of surveyed treatment programmes in the United States found controlled drinking acceptable, compared with approximately three-quarters of surveyed United Kingdom programmes. Even the United States and United Kingdom programmes that endorsed controlled drinking recommended it as a therapeutic goal for fewer than 25 per cent of patients based on their dependence severity, drinking history duration, psychological dependence, family history of alcoholism, prior treatment outcomes, and current health damage. Although controlled drinking may offer a harm-reduction option in a subset of treatment-resistant patients, abstinence generally produces the most positive long-term outcomes for patients in treatment.

Pharmacotherapy versus drug-free treatment

Pharmacotherapy for abstinence maintenance continues to be controversial. In part, this may stem from treatment providers' concerns about cross-addiction, but it also reflects the current paucity of medication options for alcoholism treatment. It is very likely that in the next decade new and more effective pharmacotherapies will be identified, and physicians will need to encourage the appropriate use of these medications by their patients since it is unlikely that the momentum for change will come from within specialty alcoholism-treatment facilities.

Smoking cessation as a concurrent treatment goal

While smoking rates have declined in the general population, estimates of smoking prevalence among alcohol- and drug-dependent persons have generally been stable and in the range of 75 per cent to 90 per cent. Indeed, smoking can now be considered as a risk marker of heavy alcohol and drug use, particularly among pregnant women. Historically, treatment professionals have discouraged a concurrent cessation of nicotine and other substance use. However, patients in treatment for alcoholism who decrease or eliminate cigarette use have comparable or slightly improved rates of alcohol abstinence compared with patients who continue smoking. Thus, smoking cessation concurrent with or in close proximity to substance-abuse treatment may decrease the risk of post-treatment relapse. This is a critical area for future research and intervention, given the excessive morbidity and mortality associated with nicotine dependence among alcohol and other drug abusers.

Further reading

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders, 4th edition*. American Psychiatric Association, Washington, DC.

Fleming MF, *et al*. (1997). Brief physician advice for problem alcohol drinkers. A randomized controlled trial in community-based primary care practices. *Journal of the American Medical Association* **277**, 1039–45.

Gianoulakis C (1996). Implications of endogenous opioids and dopamine in alcoholism: human and basic science studies. *Alcohol and Alcoholism Supplement* **1**, 33–42.

Lieber CS (2000). Alcoholic liver disease: new insights in pathogenesis lead to new treatment. *Journal of Hepatology* **32**(Suppl.), 113–28.

Ling W, Rawson RA, Compton MA (1994). Substitution pharmacotherapies for opioid addiction: from methadone to LAAM and buprenorphine. *Journal of Psychoactive Drugs* **26**, 119–28.

Mayo-Smith MF (1997). Pharmacological management of alcohol withdrawal. *Journal of the American Medical Association* **278**, 144–51.

Merikangas KR, *et al*. (1998). Comorbidity of substance use disorders with mood and anxiety disorders: results of the International Consortium in Psychiatric Epidemiology. *Addictive Behaviors* **23**, 893–907.

Reich T, *et al*. (1999). Genetic studies of alcoholism and substance dependence. *American Journal of Human Genetics* **65**, 599–605.

Swift RM (1999). Drug therapy for alcohol dependence. *New England Journal of Medicine* **340**, 1482–90.

Wagner CL, *et al*. (1998). The impact of prenatal drug exposure on the neonate. *Obstetric and Gynecology Clinics of North America* **25**, 169–94.

Wand GS, Froehlich JC (1991). Alterations in hypothalamo-hypophysial function by ethanol. In: Muller E, MacLeod R, eds. *Neuroendocrine perspectives*, Vol. 9, pp 45–126. Springer-Verlag, New

York.

26.7.2 Brief interventions against excessive alcohol consumption

Nick Heather and Eileen Kaner

Introduction

Harm caused by excessive drinking extends far beyond 'alcoholism' or severe alcohol dependence. Alcohol-related problems can be of many different kinds—medical, interpersonal, social, psychological, financial, vocational, legal—and can be associated with various patterns of drinking. For example, problems related to acute alcohol intoxication, such as accidents, violence, and public disorder offences, need not involve high levels of regular consumption or dependence. Even in the area of medical harm, research has shown that patients who develop chronic liver disease are usually only mildly dependent on alcohol, probably because those who escape florid symptoms of dependence are able to sustain a consistent pattern of heavy drinking over many years. When those at risk of harm are added to those who have already incurred it, the number of individuals whose lives may be adversely affected by their drinking becomes very large.

Among specialists in the treatment and prevention of alcohol problems, there is now a consensus that a prior focus on the alcoholism concept had distracted attention from the full range of alcohol problems that occur. This is not to imply, of course, that patients suffering from severe alcohol dependence should be ignored, only that the scope of treatment and preventive efforts should be broadened in an attempt to reduce the aggregate of alcohol-related harm in our society. Brief interventions in medical practice have a crucial role to play in this strategy.

Definitions

Excessive alcohol consumption

'Excessive alcohol consumption' is a term that includes both 'hazardous' and 'harmful' drinking. In the *International classification of diseases* (10th revision), the hazardous use of a psychoactive substance is defined as: 'An occasional, repeated or persistent pattern of use...which carries with it a high risk of causing future damage to the medical or mental health of the user but which has not yet resulted in significant medical or psychological ill effects'. Harmful use is defined as: 'A pattern of use which is already causing damage to health. The damage may be physical or mental'.

Hazardous and harmful alcohol consumption can also be defined by drinking limits recommended by medical authorities in various countries. In the United Kingdom, a joint working group of the Royal Colleges of Physicians, Psychiatrists, and General Practitioners in 1995 defined low risk (or 'sensible') consumption as up to 21 units/week for men and 14 units/week for women, increasing risk (here hazardous) as 22 to 50 units/week for men and 15 to 35 units/week for women, and high risk (here harmful) consumption as above 50 units/week for men and above 35 units/week for women (one unit = 8 g ethyl alcohol). Subsequently, a United Kingdom government report on sensible drinking effectively increased the drinking limits, by advising the public that men could drink up to 4 units/day and women up to 3 units/day. These revised recommendations have been severely criticized and medical practitioners are best advised to continue using the previous limits.

Brief interventions

As delivered by non-alcohol specialists, brief interventions refer to a collection of methods incorporated into routine practice aimed at helping patients who drink excessively to reduce their consumption to low-risk levels. These interventions are often called 'opportunistic' because they typically take advantage of opportunities that arise when people present to a medical facility for reasons unconnected with a possible alcohol problem. Brief interventions are normally restricted to individuals with only low levels of alcohol dependence or alcohol-related problems, those more seriously impaired usually being referred to specialist services. However, this is not always the case, and some doctors feel able to offer intensive treatment to more serious cases, often with advice or help from specialists in the form of shared-care arrangements. Although normally directed at a reduced drinking goal, there is no reason why brief interventions should not be targeted at total abstinence in appropriate circumstances or if the patient prefers it. However, for patients with relatively low levels of dependence and problems, insistence on abstinence is almost always a disincentive to a change in behaviour.

Prevalence

The latest United Kingdom General Household Survey in 1996 showed that 27.5 per cent of adult (over 16 years of age) males and 13 per cent of adult females reported drinking over the limits recommended by the Royal Colleges. Among young (16–24 years of age) men and women, the figures rose to 35 per cent and 21 per cent, respectively. There are reasons for believing that even these figures may be underestimates, but they nevertheless reveal the enormous extent of excessive drinking in the general population of the United Kingdom.

Patients with alcohol problems consult their general practitioners nearly twice as often as the average, their most common problems being gastrointestinal, psychiatric, and accidents. Some 40 harmful drinkers and a further 100 hazardous drinkers would be expected in every 2000 patients in primary care, but it is likely that the primary care physician will be unaware of the problem in more than half of these. In surgical and general medical wards, estimates range up to 30 per cent of all male admissions and 15 per cent of female admissions. Again, few of these patients will be identified as excessive drinkers. It is also well established that excessive drinkers are over-represented among patients of accident and emergency services.

Aims and targets of brief interventions

Opportunistic identification and brief intervention for excessive drinking are often justified as a means of early intercession and secondary prevention of alcohol problems. The attempt is made to help the patient reduce consumption or abstain before seriously adverse consequences arise, and before alcohol dependence and problems have reached levels that make intensive treatment difficult. However, as noted above, brief interventions can also be seen as making an important contribution to the public health approach in reducing alcohol-related harm at the population level. As public health measures, there is no incompatibility between the widespread implementation of brief interventions in medical practice and the adoption of fiscal, legislative, and other alcohol-control policies. These two strategies can be seen as mutually reinforcing and as acting synergistically to reduce and prevent alcohol-related harm.

Brief interventions in the sense defined here should be clearly distinguished from briefer forms of treatment given by specialist alcohol or addiction agencies. Among a number of differences, brief treatment by specialists typically takes longer to deliver than the kind of interventions we are considering here. Various types of brief intervention will be described below, but here it may be noted that they range in length from a few minutes of structured advice, up to perhaps five sessions of counselling spread over 6 months. Doctors, nurses, and other healthcare professionals (for example, specially employed counsellors) can be involved in the delivery of brief interventions.

Effectiveness of brief interventions

There is now abundant evidence that brief interventions delivered in medical settings are effective in leading to reduced alcohol consumption. Randomized controlled trials in general practice demonstrating the effectiveness of such interventions have been carried out in the United Kingdom, the United States, Canada, and Australia, and other trials are currently underway in several non-English-speaking countries. These studies indicate that intervention reduces drinking by an average of about 25 per cent. The best estimate of the proportion of patients who show a good outcome is 15 per cent among men and somewhat less among women, although it is possible than many women will respond simply by having their attention drawn to their drinking in an assessment (that is to say, without a specific intervention). Studies conducted in the 'real-world' conditions of general practice show somewhat less benefit than in those carried out under optimal research conditions, but nevertheless they support the effectiveness of brief interventions. Research in general hospital wards has been less extensive, and evidence for their effectiveness is less impressive, although still positive. Studies of the effectiveness of interventions in accident and emergency departments are just beginning.

It may be that some doctors will find the success rates described above unacceptably low. They may recall many patients who have been advised to cut down, but who have ignored this advice. However, given that up to 80 per cent of the population consult a healthcare professional at least once a year, brief interventions—if widely implemented—would undoubtedly represent a powerful means of reducing excessive drinking in the population at large. From the clinical perspective, patients who do not respond at first may do so on subsequent occasions, and even if advice seems to be ignored, it may well influence an evolving process of behaviour change. The disjunction in aims between the public health approach, which regards even very low success rates as beneficial, and the clinical perspective of most medical practitioners, in which the welfare of the individual patient is paramount, must be recognized, but the widespread and consistent implementation of brief interventions can serve both causes.

Practical approach to brief interventions

Identifying excessive drinkers

Screening is the process of identifying patients whose alcohol consumption places them at increased risk of psychological or physical complications and who might benefit from early detection and brief intervention. There are a number of laboratory indicators of excessive alcohol consumption, such as mean corpuscular volume (**MCV**), g-glutamyl transferase (**GGT**), and blood alcohol concentration (**BAC**). However, in medical practice, standardized questionnaires have been found to have a greater sensitivity and specificity than laboratory indicators; they are also far less intrusive and more acceptable to patients.

Although there are a number of standardized questionnaires, most were developed to detect a severe level of alcohol dependence (in other words, 'alcoholism'). The Alcohol Use Disorders Identification Test (**AUDIT**) is the only standardized instrument designed specifically to detect hazardous and harmful drinking in both primary and secondary healthcare settings. AUDIT is a 10-item questionnaire that includes items on drinking frequency and intensity (binge drinking), together with experience of alcohol-related problems and dependence (see Fig. 1). At a score of 8 out of a possible 40, the ability of AUDIT to detect genuine excessive drinkers (sensitivity) and to exclude false cases (specificity) is 92 per cent and 93 per cent, respectively.

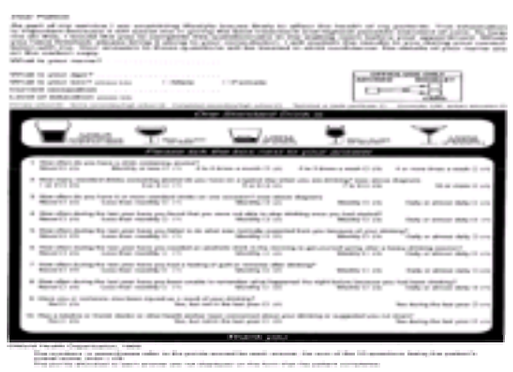

Fig. 1 The audit questionnaire.

Management of excessive drinkers

Patients who are drinking at hazardous levels (AUDIT 8–12 for men and 7–12 for women) should be offered brief intervention. Excessive drinkers who show prima facie evidence of significant alcohol dependence (AUDIT 13+ for women and 15+ for men) should normally receive a fuller assessment of their dependence and alcohol-related problems, including the impact of their drinking on their social functioning. Unless skills and resources exist to assess and treat these patients at the generalist level, they should be offered referral to a specialist alcohol or addiction treatment agency.

Types of brief intervention

The most basic form of brief intervention is simple advice to cut down, delivered in a persuasive but non-judgemental fashion and with clear guidance on consumption targets. This advice should, as far as possible, be personalized by taking into account the particular circumstances of the individual patient, their level of consumption in relation to population norms for their sex, and an appeal to any specific alcohol-related difficulties they may recognize as applying to them, including social and psychological as well as medical problems.

Advice can be supported by the offer of self-help material, and a follow-up appointment to check on their progress in cutting down will also be helpful. Successive feedback of GGT readings or other laboratory markers of alcohol consumption can be a powerful motivator. All this can be accomplished in a 5 to 10 min consultation, and specially developed brief intervention packages are available to assist this task. One such package is the Drink-Less Programme, which was developed for a WHO Collaborative Project on brief interventions in primary healthcare. It has been adapted for use in the United Kingdom, with separate versions for doctors and nurses, and is available at a small cost to cover production expenses from the Department of Primary Health Care, University of Newcastle upon Tyne.

Some doctors may prefer to spend more than 5 to 10 min on intervention, or may have the assistance of nursing or other colleagues with the time and expertise to devote to intervention. In either case, the opportunity arises to engage the patient more thoroughly in the change process. One possibility here is to use a condensed form of a type of treatment that is frequently offered in specialist alcohol treatment centres, known as cognitive–behavioural therapy. In this approach, the attempt is made to identify the types of stress and 'high-risk situations' that lead to excessive drinking in the individual case, and then train the patient in ways of coping with these stressors or situations without heavy drinking. Attention is also paid to modifying cognitive factors (for example, unhelpful beliefs about alcohol, subconscious positive expectations about its effects, lack of confidence in the ability to alter drinking behaviour) that may impede the patient's attempts to change. This form of intervention could be implemented in one session of, say, 40 min, but would probably be more effective if delivered over a series of three to five meetings over several months.

Condensed cognitive–behavioural therapy is most suitable for patients who clearly recognize that damage to their lives is being caused by alcohol and who are ready to try to change their drinking. With patients who refuse to recognize the possibility of harm and who are firmly set against change, probably nothing much can be done except to offer educational material and ask them to return if they change their minds. Between these two extremes, however, there is a large group of patients who will be fluctuating in their concern about their drinking, will be experiencing conflict about the advantages and disadvantages of heavy drinking, and will be ambivalent about attempting to cut down or abstain. For these patients, the approach that should be used is known as 'brief motivational interviewing'. This is based on the fact that ambivalence about changing behaviour is a common feature in healthcare consultations. Although some patients will respond immediately to the delivery of health advice, others will be less ready to change their drinking behaviour and direct persuasion may cause such patients to become defensive. The most effective way of helping ambivalent patients to change is to explore their conflict and encourage them to express their own reasons for concern and arguments for change. It must be emphasized that both condensed cognitive–behavioural therapy and brief motivational interviewing should not be attempted without proper training.

Implementing brief interventions

In a survey of general practitioners (**GPs**; primary care physicians) carried out in the English Midlands during 1995 to 1996, it was found that the levels of detection and intervention for excessive drinking were low. As in studies of junior hospital doctors, it appeared that the GPs did not routinely enquire about alcohol, and that any enquiries were mainly restricted to new patient registrations or those with obvious physical symptoms. Compared with earlier surveys, there was an increase in numbers of GPs who felt that working with alcohol issues was a legitimate part of medical practice, but fewer doctors saw themselves as being effective in this work. The main barriers to implementing brief interventions were stated as insufficient time and training and lack of help from government policy; the main incentives related to the availability of appropriate support services and the proven effectiveness of brief interventions.

As this chapter has shown, there is very good evidence for the effectiveness of brief interventions and it is clearly necessary to disseminate this information more widely. With regard to other barriers and incentives, strenuous efforts are now being made to persuade policy-makers and decision-makers at national, regional, and district levels to create the conditions that are needed to support the widespread implementation in routine medical practice of brief interventions against excessive alcohol consumption.

Further reading

Anderson P (1996). *Alcohol and primary health care*. WHO Regional Publications, European Series No. 64. World Health Organization, Copenhagen. [A comprehensive and authoritative guide to alcohol issues encountered in primary healthcare]

Faculty of Public Health Medicine/Royal College of Physicians (1991). *Alcohol and the public health: the prevention of harm related to the use of alcohol*. Macmillan, London. [An excellent coverage of the wider context of alcohol-related harm]

Heather N (1995). Brief intervention strategies. In: Hester RK, Miller WR, eds. *Handbook of alcoholism treatment approaches: effective alternatives*, 2nd edn, pp. 105–22. Allyn and Bacon, Needham Heights. [A review of brief intervention approaches and the evidence of their effectiveness]

Heather N, Robertson I (1998). *Problem drinking*, 3rd edn. Oxford University Press, Oxford. [An introduction to the broad changes in theory, research, and practice in the alcohol field over the last 20–30 years]

Israel Y, *et al*. (1996). Screening for problem drinking and counseling by the primary care physician–nurse team. *Alcoholism: Clinical and Experimental Research* **20**, 1443–50. [One of the best trials of brief interventions yet published, paying particular attention to issues of acceptability to physicians, nurses, and patients]

Kaner E, *et al*. (1999). Intervention for excessive alcohol consumption in primary health care: attitudes and practices of English general practitioners. *Alcohol and Alcoholism* **34**, 559–66.

Rollnick S, Kinnersley P, Stott N (1993). Methods of helping patients with behaviour change. *British Medical Journal* **307**, 188–90. [A key article on the motivational interviewing approach in medical settings]

Rollnick S, Mason P, Butler C (1999). *Health behaviour change: a guide for practitioners*. Churchill Livingstone, Edinburgh. [A recent and highly recommended guide to the negotiation of behaviour change, including drinking, in healthcare settings]

Royal Colleges of Physicians, Psychiatrists and General Practitioners (1995). *Alcohol and the heart in perspective: sensible limits reaffirmed*. Report of a joint working party. Royal College of Physicians, London.

Sanchez-Craig M (1990). A brief didactic treatment for alcohol and drug-related problems. *British Journal of Addiction* **85**, 169–77. [A guide to condensed cognitive–behavioural therapy for alcohol and other drug problems]

Saunders JB, *et al*. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on early detection of person with harmful alcohol consumption—II. *Addiction* **88**, 791–804.

Wallace PG, Cutler S, Haines A (1988). Randomised controlled trial of general practitioner intervention in patients with excessive alcohol consumption. *British Medical Journal* **297**, 663–8. [The first and still widely quoted trial showing a clear benefit of brief interventions in general practice]

26.7.3 Problems of alcohol and drug users in the hospital

Carol Ann Huff

Introduction

Alcohol and drug use are two of the most common problems facing physicians and nurses caring for patients in a general medical or surgical unit. It is estimated that up to 25 per cent of adult inpatients have a problem with alcohol or drug use. The implications of this are significant, and many studies have shown that those who abuse alcohol and/or drugs have higher rates of healthcare utilization, more complications, and longer hospital stays. Unfortunately, many of these problems go undetected or are not detected until the patient is in active withdrawal, which complicates their care. When an alcohol or drug use problem is not identified, both physicians and patients miss an important opportunity for counselling and effective early interventions.

There are many reasons why a diagnosis of alcohol or drug abuse may be missed. These include failure on the part of physicians to obtain a complete history of alcohol and drug use, including the quantity and pattern of use of both illicit and prescription medications as well as any personal or family history of substance abuse. Patients contribute to missed diagnoses through denial and minimization of their use of alcohol and drugs, in many cases because they fear social, occupational, legal, and insurance repercussions that may result from the identification of a drug or alcohol problem. These concerns are not unfounded and cannot be fully overcome, but they can be minimized by asking open-ended questions in a non-judgemental and supportive manner.

Central to the success of efforts to deal with those who abuse alcohol or drugs are educational programmes on the nature of addiction and its successful treatment. This in turn leads to the development of a supportive environment, which affords excellent patient care, minimizes ward disruptions, and encourages patients to seek assistance for alcohol or drug-related problems in conjunction with their other medical needs.

Epidemiology

Alcohol and drug use are common problems affecting people of all religions, socioeconomic classes, and geographical areas. Epidemiological studies estimate the lifetime prevalence of alcohol abuse at approximately 13 per cent and problem drug use at 6 per cent. Men are twice as likely as women to develop a drug or alcohol problem. Around 30 per cent of patients with mental illness have a concomitant substance abuse disorder, ranging from approximately 25 per cent in those with anxiety disorders to 50 per cent in those with schizophrenia. Similarly, half of the patients who abuse alcohol or drugs also have a mental illness. Polysubstance use is common and is seen in more than 75 per cent of patients seeking treatment for substance abuse.

It is estimated that more than 50 per cent of accidents and traumas are related to alcohol and or drug use. Many medical illnesses are directly or indirectly related to such use. These include gastrointestinal disorders (for example, bleeding, cirrhosis, and pancreatitis), infectious diseases (for example, pneumonia, human immunodeficiency virus (**HIV**) infection, and hepatitis) and cardiac problems (for example, ischaemia, infarction, and arrhythmias). Patients presenting with these and other medical problems are often the ones in whom a diagnosis of alcohol or drug dependence is not suspected and thus missed entirely, or only made when the patient is in withdrawal.

Clinical features and treatment of intoxications

Alcohol

The clinical picture of alcohol intoxication depends on the rate of ingestion, metabolism, and tolerance of an individual to alcohol's effects. Although legal intoxication is a blood alcohol level of 100 mg/dl, behavioural, psychomotor, and cognitive changes can be seen at levels as low as 20 to 30 mg/dl in those without tolerance. Euphoria occurs at 25 to 50 mg/dl, incoordination at 50 to 100 mg/dl, ataxia at 100 to 200 mg/dl, stupor at 200 to 400 mg/dl, and coma at 400 to 500 mg/dl. Some people become somnolent after modest alcohol ingestion, do not experience euphoria, and therefore rarely abuse alcohol.

The neurological signs of intoxication include slurred speech, impaired coordination, nystagmus, and gait disturbance. Signs of increased sympathetic activity including tachycardia, hypertension, mydriasis, and skin flushing often accompany these changes. Whilst patients may present with what appears to be mere alcohol intoxication, physicians must also be aware of the potential for superimposed problems requiring acute care including hypoglycaemia, subdural haematoma, systemic infections (including aspiration pneumonia), and other ingestions (including methanol, antifreeze, and sedatives).

Alcoholic coma is a medical emergency with a mortality rate approaching 5 per cent. Management requires prompt recognition, careful clinical examination, and an expeditious history, usually from a friend, of the amount and rate of alcohol ingested, as well as any other drugs or medications. If the patient is stuporous or has excessive secretions, he or she should be intubated immediately to provide airway protection and ventilatory support. Once the patient's cardiopulmonary status is stabilized, full investigation—including arterial blood gases, serum chemistries, toxicology screens, and imaging studies—can be undertaken to look for complicating factors when clinically appropriate. Patients may require close monitoring in an intensive care unit with respiratory, cardiovascular, and haemodynamic support and, in some cases, haemodialysis.

Opiates

An opiate overdose can be a life-threatening emergency and requires prompt recognition and institution of therapy. The characteristic signs include varying degrees of clouded consciousness (ranging from somnolence to obtundation), pinpoint pupils, and marked respiratory depression. Aside from protection of the airway, administration of oxygen, and provision of respiratory support as required, treatment involves the administration of an opiate antagonist, usually naloxone. If intravenous access is not available, then this should be administered intramuscularly. The starting dose is 0.2 to 0.4 mg and the onset of action is rapid, typically seen within 2 to 3 min of intravenous administration or 15 min after intramuscular administration. If no response is seen, the dose may be increased. Patients typically respond to doses of less than 2 mg of naloxone, and if the maximum dose of 10 mg is given and the patient has not responded it is unlikely that an opiate overdose is the cause of the problem. The half-life of naloxone is about 90 min, hence re-dosing may be needed, particularly if the overdose is due to a long-acting agent, such as methadone. Pulmonary oedema is often associated with the severe respiratory distress and may improve with restoration of the respiratory drive, but may also require positive-pressure ventilation to fully alleviate it.

Withdrawal syndromes

The clinical features and treatment of withdrawal syndromes are summarized in Table 1.

Alcohol

Clinical features of alcohol withdrawal

Chronic exposure to alcohol leads to upregulation of neural mechanisms within the central nervous system to counteract the depressant effects of alcohol. When the amount of alcohol ingested is diminished or abruptly stopped, these adaptive mechanisms are left unopposed and a hyperexcitable state ensues. This hyperexcitable state can lead to a range of symptoms, from tremulousness and disordered perceptions to seizures and frank delirium, also known as delirium tremens. This spectrum of findings is known collectively as the alcohol withdrawal syndrome.

The first signs of withdrawal begin 6 to 8 h after the alcohol-dependent person's last drink. Tremor is the earliest, most common, and most easily recognized sign. It is coarse, generalized, rapid, and intensified by motor activity and stress. It may be severe enough to interfere with basic motor activities. At this stage, patients are often irritable and complain of nausea and vomiting. In the absence of resumed alcohol consumption or treatment, many go on to develop signs of sympathetic overactivity including diaphoresis, tachycardia, mild hypertension, facial flushing, and increased body temperature. These symptoms peak in intensity between 48 and 72 h and gradually subside by day 4 to 5 unless alcohol consumption is resumed. Anxiety, insomnia, and mild autonomic dysfunction may persist for up to 6 months after alcohol cessation and can predispose some patients to early relapse.

Some 25 per cent of patients with tremor and autonomic instability develop perceptual disturbances. These include hyperacusis, vivid nightmares, and, in some cases, auditory hallucinations which peak in intensity 24 to 36 h after alcohol intake is stopped. The auditory hallucinations can last for several weeks and are termed 'alcoholic hallucinosis'. This condition can be distinguished from schizophrenia by its temporal association to alcohol cessation and its lack of recurrence unless drinking is resumed.

Generalized tonic–clonic seizures occur in up to one-third of patients with chronic alcohol dependence when their alcohol ingestion is stopped. The seizures are usually isolated and begin within the first 12 to 24 h of stopping, occasionally they occur in groups of three or four within a 6-hour period. Alcohol withdrawal accounts for 15 per cent of all seizures, and alcoholics who have seizures during one episode of withdrawal are likely to have them during subsequent episodes of withdrawal. The seizures are self-limited and do not require treatment beyond that given to the patient for alcohol withdrawal, unless another cause is found. Only rarely do alcohol withdrawal seizures progress to status epilepticus.

Delirium tremens is the most severe form of alcohol withdrawal and is seen in about 5 per cent of alcoholics. It is characterized by a state of agitated arousal, global confusion, and disorientation. Patients are delusional, have vivid hallucinations, and insomnia. They exhibit signs of sympathetic hyperactivity: fever, tachycardia, mydriasis, and diaphoresis. The onset is 2 to 4 days after alcohol cessation and may be the first sign of withdrawal in a previously unrecognized alcoholic. Patients are terrified, combative, and often destructive. An episode of delirium tremens may last from 24 to 72 h and often ends as abruptly as it starts. Relapses occur and may continue for days to weeks, often with intervening periods of lucidity. Aggressive treatment with benzodiazepines is essential to calm the patient, control his or her behaviour, and ensure that they do not harm themselves or others.

Treatment of alcohol withdrawal

There are three main aspects to the treatment of alcohol withdrawal. First, one must perform a thorough, yet expeditious, evaluation to look for coexisting medical illnesses. This can usually be accomplished through a careful history, physical examination, and selected laboratory studies. Second, one must ensure adequate nutrition. Patients who abuse alcohol often consume most of their daily calories in the form of alcoholic beverages, which contain carbohydrates but are devoid of minerals, protein, and vitamins. As such, most alcoholics are deficient in folic acid, thiamine, pyridoxine, and nicotinic acid and need to have these vitamins replaced. The most important is thiamine and the patient should be given 100 mg a day for 5 to 7 days. The first dose should be given intravenously or intramuscularly to ensure absorption, and later be changed to an oral preparation. In addition to thiamine, patients should be given a multivitamin and folic acid preparation daily for at least 1 week. Thiamine should be given prior to or concurrent with the administration of dextrose-containing fluids, as glucose increases the need for thiamine and may precipitate or worsen a Wernicke's encephalopathy if given to patients who are thiamine-deficient.

The third aspect of treatment is to replace the central nervous system depressant effect of alcohol with a pharmacological agent that can be tapered over 3 to 5 days. There are several drugs that can be used, but benzodiazepines have the highest margin of safety and are therefore preferred. Diazepam and chlordiazepoxide are the most commonly used, in part because of their longer half-lives. Lorazepam or oxazepam are preferred in patients with hepatic dysfunction, as they do not require hepatic metabolism.

Patients with only a slight tremor or minimal autonomic hyperactivity do not require pharmacotherapy. If additional symptoms are present or patients are in moderate to severe distress, a benzodiazepine should be given. The goal of treatment is to keep the patient sleepy, but easily rousable. Each patient's requirements will therefore differ depending on their tolerance, sex, age, and concomitant medical problems. As a general rule, the average male can be managed with diazepam 10 mg four times a day for the first 1 to 2 days, followed by 10 mg twice daily for 2 days, and 10 mg once a day for 2 days. If chlordiazepoxide is used, the average dose is 25 to 50 mg four times a day for 2 days, followed by 20 mg a day for 2 days, then 5 mg a day for 2 days. An individual patient's requirements may differ and can best be determined by frequently monitoring vital signs, symptoms, and mental status. In women, the starting doses are usually reduced by 20 per cent. The goal of early recognition and treatment is to prevent the development of delirium tremens.

Delirium tremens is a medical emergency and should raise suspicion as to the presence of an intercurrent illness, such as pancreatitis, pneumonia, hepatic failure, or a subdural haematoma. It requires an expeditious assessment, followed by the rapid administration of intravenous benzodiazepines. Between 5 and 10 mg of diazepam may be given every 5 to 15 min until the patient is calm. Maintenance therapy is needed every 1 to 4 h after this, and may exceed 200 mg a day for a period of 3 to 5 days. The goal of treatment is to calm the patient without oversedation until the symptoms have passed. Although this syndrome is being increasingly recognized and treated, it still has a reported mortality rate of 1 to 5 per cent, which is usually due to an intercurrent illness.

Carbamazepine is effective in treating mild to moderate alcohol withdrawal, but data are limited on its efficacy in preventing seizures and delirium tremens, and its use is limited by side-effects in up to 10 per cent of patients. Although commonly used in emergency settings, there is a paucity of data to support the use of phenytoin in treating alcohol or drug-withdrawal seizures and thus, if started, it should not be continued beyond 5 to 7 days unless an underlying seizure disorder is identified. Antipsychotic drugs have no role in the treatment of mild withdrawal. They are sometimes used as adjuncts in the treatment of delirium tremens, although caution must be exercised as they can lower the seizure threshold.

Opiates

Clinical features of opiate withdrawal

The onset and duration of withdrawal depends on which opiate the patient is dependent upon. Opiates with short half-lives have a more rapid onset and shorter duration of symptoms than opiates with longer half-lives. Withdrawal from morphine or heroin usually begins 6 to 8 h after the last dose in a tolerant person and lasts 5 to 7 days. Methadone has a half-life of 22 to 24 h and withdrawal begins more slowly and lasts much longer. **LAAM** (L-a-acetylmethadol) has the longest half-life and thus, the longest withdrawal syndrome in the absence of opiate replacement.

Although uncomfortable, opiate withdrawal is not life-threatening. The symptoms are the opposite of the acute effects of the drugs. They include fatigue, anxiety, irritability, and insomnia. Patients complain of abdominal cramping, nausea, vomiting, and diarrhoea. They frequently yawn, experience rhinorrhoea, excess lacrimation, and increased bronchial secretions. Sweating is common, may be profuse, and is often associated with piloerection. Patients complain of bone pain and myalgias and say that they feel as though they have influenza. These symptoms are accompanied by intense craving for opiates. The physical signs of opiate withdrawal include mydriasis and mild elevations in blood pressure, body temperature, and respiratory rate.

Treatment of opiate withdrawal

Many patients with opiate dependence will have attempted self-detoxification without significant success. As such, when these patients are admitted to the hospital, it is best to treat their withdrawal symptoms pharmacologically. This not only alleviates their discomfort, but also improves the ability of the entire healthcare team to treat their presenting problems and to assess their willingness to seek treatment for their substance abuse. Several approaches can be used, including the use of

opiate replacement therapy or through the use of non-opiate medications to treat the symptoms of withdrawal.

Opiate withdrawal is most effectively treated by the administration of opiates. This can be accomplished by giving a long-acting opiate such as methadone, or by using an opiate with agonist and antagonist properties, such as buprenorphine. When methadone is used, the principle is to give enough methadone on the first day to alleviate the patient's symptoms and then to decrease the dose by 10 to 20 per cent per day over the next 5 to 10 days. Estimating the amount of methadone that an individual patient needs is difficult, and it is generally best to give a test dose of 10 to 20 mg and monitor the patient's response over the next 1 to 2 h. This ensures that the patient is not overmedicated and also that the symptoms are improving. If not, an additional dose or doses may be given then or 12 h later if symptoms recur. Although there are no special licensing requirements for the use of methadone in hospital patients in the United States, its use in the outpatient treatment of opiate dependence is restricted to licensed facilities. As such, when therapy needs to be continued beyond the length of hospital stay, either for detoxification or maintenance, the patient must be referred to a licensed treatment facility. This makes its use in treating medically ill inpatients challenging, particularly as the length of hospital stays continues to decline.

An alternative approach is to use buprenorphine, an opioid with partial μ-agonist/antagonist properties. As such, it treats the symptoms of opiate withdrawal, and blocks the effects of additional opioid stimulation in a dose-dependent manner. It has poor bioavailability and therefore must be given sublingually or parenterally. Its advantages include less physical dependence, a lower risk of overdose, and a less intense withdrawal syndrome. Its half-life is approximately 3 h. The initial dose is usually 0.3 to 0.6 mg intramuscularly or 2 to 4 mg sublingually, repeated at 6 to 8 h intervals and tapered over 4 to 5 days.

The symptoms of opiate withdrawal can also be managed with non-opiate medications. The a2-agonists, such as clonidine, decrease the sympathetic nervous system overactivity in patients with opiate withdrawal. They decrease the patient's nausea, vomiting, abdominal cramping, and diarrhoea, but do little to alleviate myalgias, back pain, and craving for opiates. Most patients respond to oral doses of clonidine 0.1 to 0.3 mg given every 6 to 8 h for the first 24 to 48 h, followed by a taper over the next 3 to 5 days. Dose-limiting side-effects are orthostatic hypotension and somnolence. Clonidine neither shortens the duration of withdrawal nor prevents relapse.

Symptomatic treatment of abdominal cramps and diarrhoea can be achieved with Lomotil® (diphenoxylate and atropine) or loperamide and dicyclomine. Both diphenoxylate and loperamide are opioids with low addictive potential as they are poorly absorbed from the gastrointestinal tract. Myalgias and bone pain may be treated with non-steroidal anti-inflammatory drugs (**NSAIDs**) such as ibuprofen or naproxen, and rhinorrhoea can be managed with nasal decongestants such as pseudoephedrine. Insomnia should be managed expectantly: the routine use of sedatives is discouraged as they have no specific effect on opiate withdrawal and can also be misused.

Cocaine

Withdrawal from cocaine leads to intense neuropsychological symptoms, including dysphoria and intense cravings for more cocaine. Despite these symptoms, which are often severe, cocaine withdrawal is not life-threatening and has no associated physiological instability. Patients with a long history of cocaine use or recent bingeing may also complain of depression, anxiety, anhedonia, and profound fatigue. Nevertheless, patients can usually be managed with reassurance and supportive care, but may require psychiatric assistance if they are suicidal. On rare occasions, a brief course of benzodiazepine therapy (less than 24 h) may be needed if cravings and anxiety are of an intensity to jeopardize the team's ability to care for the patient.

Sedative/hypnotics

Clinical features of sedative/hypnotic withdrawal

The development of a sedative withdrawal syndrome varies in onset and severity based on the half-life of the drug involved, the degree of dependence, and the duration of daily use. Benzodiazepines are the most commonly prescribed sedatives, and although safer than barbiturate and non-barbiturate hypnotics, they still have the potential for abuse and the development of dependence. Benzodiazepine dependence can occur in as little as a month if higher than usual doses are taken on a daily basis, or after several months when standard doses are used. Withdrawal can therefore be seen in a wide range of patients when these drugs are stopped, including those who are unaware of their physical dependence.

The signs and symptoms of sedative withdrawal are variable and do not always follow a specific sequence in their development. Mild irritability, tremor, diaphoresis, and sleep disturbances are common. Physical findings can include orthostatic hypotension, tachycardia, fever, seizures, and delirium. With long-acting preparations, such as diazepam, the onset of symptoms may be delayed for 24 to 48 h after the drug is discontinued and does not peak in intensity until 5 to 7 days later. Symptoms may appear sooner when a shorter acting agent such as alprazolam is involved. A previously unappreciated sedative dependence may first be recognized 2 to 3 days into a patient's hospital stay, when some or all of the above symptoms and signs appear. Thus, it is important to consider benzodiazepine dependence in patients who are symptomatic, even if a history of benzodiazepine use has not been previously elicited.

Treatment of sedative/hypnotic withdrawal

Treatment of sedative/hypnotic withdrawal requires close medical attention, with monitoring for seizures. Detoxification is best accomplished with sedative substitution and gradual tapering in a controlled setting. As with opiate dependence, it is usually best to substitute an agent with a longer half-life, such as diazepam, clonazepam, or phenobarbital for the sedative of abuse. Each patient's level of dependence needs to be determined by giving a test dose and monitoring the patient's response, rather than relying on the patient's historical reports. Tolerance to the subjective effects of sedatives develops quickly; while tolerance to sedation remains low, unexpected central nervous system depression may occur if too large a dose is given. In the absence of sedation or intoxication, it is likely that the patient's tolerance is higher and indicates that a larger dose will be needed. Once the patient's requirements are determined, the dose is divided and can be gradually reduced over 14 to 21 days. This time course may require extension in patients with severe symptoms or a history of withdrawal seizures.

Pain management in patients with substance abuse disorders

Adequate pain management is an integral component of the successful care of all patients who require admission to hospital. This is particularly true for patients with a concomitant alcohol or drug use problem as they are more likely to sustain traumatic injuries and have a higher rate of medical and surgical illnesses. As pain is primarily subjective, physicians and nurses must rely on the patient's assessment of the adequacy of analgesia. Physical signs such as tachycardia, diaphoresis, and hypertension that are associated with acute pain are neither sensitive nor specific, and thus cannot be relied upon in assessing the adequacy of treatment. This is further complicated in patients with problems of dependence, as these findings may also be signs of withdrawal. The complex behavioural and psychological phenomena associated with addiction can alter the patient's perception of pain, and some studies suggest that patients with alcohol or drug-dependence may have a lower tolerance for pain.

Managing pain in this population can therefore be difficult even for experienced clinicians. Yet, by using a systematic approach, one can make a good assessment of the problem and provide the best chance for a successful outcome, diminishing the chances of developing an adversarial relationship between the patient and the medical staff.

The six steps to consider are listed in Table 2.

1. The first and most important step is to identify the source of the patient's pain and direct primary therapy towards it. Examples of this include antibiotics and debridement for an abscess and realignment of a broken bone.
2. The next step is to determine whether the patient has a history of active or remote substance abuse. This helps in determining the dose and frequency of analgesia, as those with a history of active opiate use or who are in methadone maintenance programmes are likely to be tolerant to opiates. Physicians should pay close attention to symptoms of anxiety and depression which can not only influence one's perception of pain, but may require concomitant intervention.
3. The third step involves selecting the appropriate opioid analgesic based on a clear understanding of the pharmacology of the agent selected and the person in whom it is to be used. Patients with tolerance will need higher doses at more frequent intervals to achieve the same analgesic effect as those who are not tolerant. Using opiates with both agonist and antagonist properties is generally not a good way to manage acute pain in dependent patients, as these drugs have lower analgesic potential due to difficulty in overcoming their antagonist activity. While patient-controlled analgesia is very useful in the non-addicted patient, its use in those with drug dependence is usually discouraged because of concerns about the 'high' that patients can receive from intravenous bolusing of opiates.

4. The fourth step involves the addition of non-opioid analgesics, including NSAIDs such as ibuprofen, or ketorolac if parenteral administration is needed. These medications can be effective, usually as adjuncts to opiates, in the treatment of acute pain. They are most helpful in cases where dose-limiting toxicities of opiates have been reached without optimal analgesia. Regional nerve blocks should also be considered in the appropriate clinical setting.
5. The fifth step is the recognition and prevention of aberrant behaviour. Patients with an active drug or alcohol problem can frequently have difficulty with limit-setting and, despite optimal medical care and pain management, may continue to seek drugs, either in the form of prescribed medications or through illicit use during their hospital stay. Patients may tamper with intravenous catheters and infusion devices, may attempt to crush oral medications for intravenous administration, or may surreptitiously use illicit drugs or alcohol. These behaviours are unacceptable and clear limits must be set. This is often best accomplished by discussing with the patient what is acceptable behaviour and which behaviours will not be tolerated. Utilization of behaviour contracts, limitation of visitors, and supervised medication administration have all been employed with some success in individual situations.
6. The last step includes early consultation with substance abuse and psychiatric services to assess the patient's readiness for treatment, to encourage change, and to allow for appropriate referrals to aftercare treatment. In using a multidisciplinary approach to the patient's care, the medical and substance use issues can be addressed simultaneously. This is beneficial both to patients and healthcare providers, and is essential when behavioural management becomes an issue.

Summary

Alcohol and drug dependence are common among patients admitted to a general medical or surgical unit. These problems are frequently missed or may not be detected until a patient exhibits physical or psychological signs of withdrawal. As awareness of the problem increases and physicians become more comfortable recognizing and treating them, these problems will differ little from other medical conditions that are commonly encountered in the hospital and clinic settings.

Further reading

Cheskin LJ, Fudala PJ, Johnson RE (1994). A controlled comparison of buprenorphine and clonidine for acute detoxification from opioids. *Drug and Alcohol Dependence* **36**, 115–21. [Demonstrated that buprenorphine and clonidine have similar efficacy in opiate withdrawal, but buprenorphine has fewer side-effects.]

Gossop M (1988). Clonidine and the treatment of the opiate withdrawal syndrome. *Drug and Alcohol Dependence* **21**, 253–9. [Nice review.]

Jaffe JH (1990). Drug addiction and abuse. In: Gilman AG, *et al*, eds. *Goodman and Gilman's the pharmacological basis of therapeutics*, 8th edn, pp 522–73. Pergamon Press, New York.

Lader M, Morton S (1991). Benzodiazepine problems. *British Journal of Addiction* **86**, 823–8.

Mayo-Smith MF for The American Society of Addiction Medicine Working Group on Pharmacological Management of Alcohol Withdrawal (1997). Pharmacological management of alcohol withdrawal. *Journal of the American Medical Association* **278**, 144–51. [A meta-analysis providing evidence-based guidelines.]

Moore RD, *et al.* (1989). Prevalence, detection and treatment of alcoholism in hospitalized patients. *Journal of the American Medical Association* **261**, 403–7. [Demonstrated underdiagnosis of alcohol abuse in hospital patients.]

Portenoy RK, Payne R (1997). Acute and chronic pain. In: Lowinson JH, *et al.*, eds. *Substance abuse: a comprehensive textbook*, pp 563–89. Williams and Wilkins, Baltimore, MD.

Regier DA, *et al.* (1990). Comorbidity of mental disorders with alcohol and other drug abuse: results from the epidemiologic catchment area (ECA) study. *Journal of the American Medical Association* **264**, 2511–18. [Population-based study estimating the lifetime prevalence of alcohol and drug disorders in patients with comorbid psychiatric illness.]

Samet JH, Stein MD, O'Connor PG (1997). Alcohol and other substance abuse. *Medical Clinics of North America* **81**, 831–1075. [Concise review of many aspects of alcohol and drug abuse.]

Smith DE, Wesson DR (1999). Benzodiazepines and other sedative-hypnotics. In: Galanter M, Kleber HD, eds. *Textbook of substance abuse treatment*, pp 239–50. American Psychiatric Press, Washington DC. [Comprehensive review.]

US Department of Health and Human Services (1999). *Results of the 1998 National Household Survey on Drug Abuse*. National Institute on Drug Abuse, Rockville, MD. [National survey of the prevalence of alcohol and drug use.]

27 Forensic medicine and the practising doctor

Anthony Busuttil

Introduction

The interface and borders between the law and the medical profession are becoming increasingly wider and more far-reaching, and a doctor has to keep in mind his or her legal responsibilities and duties from the very first day of their practice. Sound ethical principles should form the backbone of professional clinical practice, ensuring competence and integrity of the medical practitioner. The doctor–patient relationship is a partnership based on mutual trust, respect, and confidence with the fundamental rights of every patient being adhered to always and respected fully. These patient rights are succinctly enshrined in the *Declaration of Lisbon* agreed to in September 1981 by the 34th Assembly of the World Medical Association (Box 1).

Box 1 Declaration of Lisbon

- Recognising that there may be practical, ethical and legal difficulties, a physician should always act according to his/her conscience and always in the best interest of the patient. The following Declaration represents some of the principal rights that the medical profession seeks to provide to patients. Whenever legislation or governmental action denies these rights of the patient, physicians should seek by appropriate means to assure and restore them

a. The patient has the right to choose his physician freely.
b. The patient has the right to be cared for by a physician who is free to make clinical and ethical judgements without outside interference.
c. The patient has the right to accept or refuse treatment after receiving adequate information.
d. The patient has the right to expect that his physician will respect the confidential nature of his medical and personal details.
e. The patient has the right to die with dignity.
f. The patient has the right to receive or decline spiritual and moral comfort, including the help of a minister of an appropriate religion.

In addition to dealing with natural illnesses, the services of the medical practitioner, whether in primary care or in hospital practice, are often called upon when injuries and other forms of abuse have taken place, also when death has occurred, particularly if death was sudden and unexpected. As a direct consequence of this, the doctor may acquire information that suggests a suspicious and potentially criminal event. In such instances the doctor's duty of care and bond of confidentiality to the patient must be carefully balanced against his duties as a citizen of a country in which homicide cannot go undetected and crime cannot be condoned.

The courts

It is to be recalled at all times that the system in the courts in Britain, in contrast to other parts of Europe, is very firmly based on the so-called adversarial system, with the sole exception of the HM Coroners' Courts at which an inquisitorial system is in place. In the adversarial system within the criminal courts, the object of the forensic exercise is for those acting for the prosecution on behalf of the State (Her Majesty in England, Wales, and Northern Ireland and the Lord Advocate—or Procurator Fiscal—in Scotland) to prove that the charges laid against the person in the dock (the 'plaintiff' in England and the 'accused' in Scotland) can be proved 'beyond reasonable doubt'. The charges are drafted in terms of alleged contraventions of the Statutes of Criminal Law or of 'common law'. This is a universal system of unwritten law, applied commonly and universally throughout the land, which employs a set of principles, tenets, and maxims that can provide answers to legal problems and can be applied by the judiciary in deciding on cases. Over the years common law has been interpreted and re-moulded by the decisions of one generation of judges after another; these are written down and can be alluded to and quoted in decisions and judgments made by other judges. As a hierarchy of courts developed, thereby enabling an appeal against a decision reached in a lower court to a more senior court, decisions taken by higher courts became binding on lower courts—the concept of 'precedent'. The courts do this by adducing evidence taken under oath or affirmation from witnesses in open court. This enables in minor (summary) cases, the judge, or in serious cases (indictable or solemn cases), the members of the jury (15 in Scotland, 12 in the rest of the United Kingdom) to come to decision. If the verdict is one of 'guilty', the judge will sentence the accused according to tariffs—fines, custodial sentences, community work, etc.—laid down in Statute. The object of the defence is to attempt to cast doubt and demolish the evidence that is being given: they do so through their own witnesses, of which the accused may be one, and by cross-examination of the prosecution witnesses.

The medical practitioner in active employment is usually exempt from jury service but may have to give evidence to fact either as an 'ordinary' witness, like any other citizen, or as a 'professional' witness, by divulging to the court information which he or she has gathered in their professional capacity, for example about injuries, physical or mental illness of the persons in the dock, or other witnesses who have suffered injury. In matters of a scientific and medical nature, when the courts wish to be informed and have explained to them matters which do not fall within the ambit of common knowledge and common sense, persons of a sufficient experience and expertise in the particular subject who can suitably assist may be called as witnesses. By giving information about these matters to the court in this capacity the witness is referred to as an 'expert' (*ex*: from; *peritia*: specialized skill). They are not there to come to a decision on behalf of the court, but solely to assist to the best of their capacity with information, and if required, also to give opinions based on their professional experience and expertise. Details of their qualifications and specialization would be brought to the attention of the courts at the commencement of the expert evidence.

In HM Coroners' courts, whose sole limit is the investigation of death, the questions are asked of all the witnesses by the Coroner, who has to come to a decision at the end of the public 'inquest' as to the identity of the person who has died, when and where they died, and how they came about their death. There is no such office in Scotland, where death investigations are carried out in private by the Procurator Fiscal. There are no public inquests except for deaths in custody and in the course of work as a consequence of the employment; in all other cases the Lord Advocate decides whether 'in the public interest' a 'Fatal Accident Inquiry' is required.

In civil cases one party—persons, or company, or corporate body—sues another party, and the arbiter in almost all such cases is a judge or judges. Both sides can call witnesses, including expert witnesses, to give evidence, and all witnesses can be cross-examined. The only remedy in the civil courts is a pecuniary one, namely

the award of monetary damages. The case requires to be proven 'on the balance of probabilities'.

Duties at a death

Death can be said to have occurred in the body of a person when there is either:

- an irreversible cessation of all function of the brain; or
- an irreversible cessation of the circulation of blood.

In this country there is no statutory definition of death as exists elsewhere, for example in the United States or in New South Wales, Australia. There is also no statutory requirement that the diagnosis of the fact of death is invariably made by a doctor; it is however customary that all diagnoses of death are confirmed by a medical practitioner.

When a doctor is called to a person thought to be have died, six principal responsibilities have to be considered (Fig. 1):

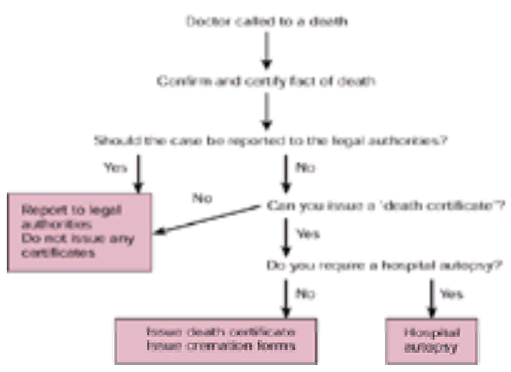

Fig. 1 Doctor called to a death.

1. To confirm the **fact of death** and **document** this at the time.
2. To exclude on medical grounds, where possible, any **suspicions of foul play or negligence** in relation to the death.
3. To identify whether there is a **requirement to report the death** to the appropriate authorities who have a statutory (Her Majesty's Coroners in England, Wales, and Northern Ireland) and/or 'common law' (Procurators Fiscal in Scotland) duty to investigate deaths.
4. To issue a **medical certificate of the causes of death** (commonly referred to as the **Death certificate**) when in a position to do so.
5. If **unable to issue** a Death certificate, to **refer the death to the legal authorities**.
6. If appropriate, and if the medical practitioner is suitably qualified, to issue **other certification** which may be required in relation to the death and the disposal of the decedent.

Confirming and documenting death

In the recently deceased it is essential that a formal clinical examination is carried out to ensure that the pumping action of the heart and breathing have both ceased. This should be done by auscultation over the chest for a timed period of about 2 min. It is also useful to feel the tension in the eyes, which decreases quite promptly after death (due to a lack of blood pressure) and to look into the pupils, which assume a mid-position: the corneal and light reflexes disappear. If the cornea is still transparent and moist (usually until about 10 min after death if the eyelids were closed) and if in any doubt, it may be useful to examine the eye with an ophthalmoscope: the blood in the retinal veins breaks up into segments within 10 s of clinical death in the majority of instances, a phenomenon variously referred to a 'railroading' or 'cattle-trucking'. In all deaths, there is also blanching of all retinal vessels if the eyeball is pressed with a finger.

In the presence of severe mutilation and burning, or if the body is showing obvious features of decomposition, these tests are obviously superfluous and to carry them out would not be sensible.

Great care must be taken in situations in which vital functions and general metabolism may have been decreased to such an extreme as to simulate death. For a doctor to be caught out in such situations is not only a major embarrassment but may lead to civil action for damages and disciplinary procedures for erroneous certification. Such 'apparent death' or 'suspended animation' can take place in instances where there has been:

- an overdose of CNS depressant drugs
- hypothermia from exposure or medical complaints
- (electrocution)
- (drowning)
- (psychiatric catatonic states).

When in any doubt, particularly when such conditions are present in combination, and particularly in the young, it is always wise to attempt full resuscitation (or 're-animation' as it is referred to in Europe) (see Chapter 16.3).

In all situations when death has been medically confirmed, this must be fully documented on the case notes with the entry timed, dated, and signed. This should be done at the time or as soon afterwards as is reasonable, and the entry should list the clinical tests carried out.

Brain death

The diagnosis of brain death (or 'brainstem death') in patients in an apnoeic coma should be done accurately and positively, even in the presence of a beating heart and machine-maintained respiration, by an appropriately qualified and suitably experienced—at least 5 years' registration—medical practitioner. Tests have to be carried by two doctors and on two occasions several hours apart. Details are discussed in Chapter 16.6.3. In brief:

- Any potentially reversible condition that has led to cerebral depression must be positively excluded. The coma must not be due to CNS depressant drugs, neuromuscular blocking agents (muscle relaxants), hypothermia, or metabolic abnormalities, particularly hypoglycaemia, renal or hepatic failure.

Once all these conditions have been excluded, the simple tests used to confirm that brain death has occurred are that:

- both pupils should be fixed, though not necessarily equal and rounded, and do not react to light;
- no response occurs to corneal stimulation with cotton wool;
- no response is found to the presence of the endotracheal tube or any cough response to suction applied to the tracheal lumen;
- no eye movements occur when 20 ml of ice-cold water are injected into the external auditory meati, having previously established clear access to the ear drums;
- no motor cranial nerve responses are elicited to painful stimuli, e.g. ear-lobe pinching, supraorbital pressure;
- there is no spontaneous breathing with the onset of hypercapnia by disconnecting the respiratory for a sufficiently lengthy period to allow the $PaCO_2$ to build up.

If brain death is thus diagnosed, the time when this has been definitively established can be taken as the 'time of death', and not the time that the respiratory support is withdrawn. Any subsequent harvesting of organs will therefore be carried out on a cadaver in whom artificial respiratory and other support has been retained.

The time of death

The actual time of death will always precede the time that 'life is pronounced extinct', and the duration of this period—often referred to as the 'postmortem interval'—is frequently unknown and unknowable. Forensic pathologists are often involved in making estimates of this period by utilizing such phenomena as cooling of the body after death, the onset and distribution of rigor mortis, etc. This is an area fraught with problems and inaccuracies, and one that the non-forensically qualified medical practitioner should best refrain from venturing into and giving any professional opinions on. In those instances where the estimation of the postmortem interval is of specific importance, this matter should be referred to those who are forensically qualified.

Excluding foul play

Doctors, similarly with all other professional persons, have an overriding duty to the 'society' in which they practise, that allows breaches of medical confidentiality when it comes to ensuring that crime is prevented, fully investigated, and detected. Thus, if a medical practitioner suspects or has good reason to believe that a particular death was due to a criminal or negligent act, it is his or her unalienable public duty to ensure that this incident is reported to the police, and that no certification is completed, no matter how obvious the cause of death may be. If the doctor has acted in 'good faith' in such instances, even if their suspicions are eventually found to be unfounded on due investigation, they do not lay themselves open to civil litigation by informing about the death.

No matter the age of the decedent, but particularly so in the elderly and in children, other considerations should not obscure or side-track the doctor from accepting the possibility that death was due to a criminal or negligent act. In this respect, it may be of some importance, as part of the diagnosis of the fact of death, to look carefully for petechial facial haemorrhages in all decedents, particularly around the eyes, behind the ears, on the labial mucosa, and specifically in the conjunctivae. In the absence of known coagulation problems, their presence on the face should always be taken seriously, especially in babies, and if found should always raise the possibility of death being due to a mechanical form of asphyxia, suggesting the possibility of another party's involvement in the death.

Reporting deaths to the legal authorities

Certain categories of deaths are always reportable to the legal authorities. These include all violent deaths, deaths from all types of accidents, including medical mishaps—no matter how long before the death this accident took place, deaths from suicide and suspected suicide, deaths in legal custody and in secure mental institutions, deaths in fires and explosions, from suspected poisoning, from industrial diseases, and from other diseases which by law are notifiable—these include tuberculosis and hepatitis (but not HIV-related deaths). Deaths in which a medical, surgical, or therapeutic mishap may have contributed to or caused the death of the patient always have to be reported.

If the cause of death is not known and yet there is no other reason why the death should be reported, that death is an uncertified death and thus has to be reported. In such instances this notification should precede any attempts to secure an autopsy.

Death certification

This is a privilege accorded to doctors as a consequence of fulfilling the criteria for their registration with the General Medical Council, and thus any misuse or abuse of this principle would render the doctor liable to a disciplinary procedure. The death certificate is an important statutory document and also a very important public health record: it should be filled in carefully and with all due consideration.

The certifier of the Causes of Death does so to 'the best of my knowledge and belief' and records both the immediate causes of death (Part I), these to be placed in a sequence with the initial line (a) being the condition which chronologically resulted from the condition in the second line (b) and so on. Other conditions that have contributed to the death or accelerated it should be listed in Part II of the certificate.

The mode of dying, for example cardiac arrest, cardiac failure, coma, are inappropriate terms to use in this context, except if they are qualified by the underlying causative pathological condition, such as ischaemic heart disease. Terms such as 'senility' and 'old age' should strictly refer to decedents above the age of 80 years, and then only when there was no further recent superimposed pathology. 'Natural causes' is not acceptable. However, it is accepted that in certain instances the determination of the cause of death would have to await laboratory studies, for example overdose deaths, and in such instances a death certificate indicating this may be acceptable.

Reporting to the legal authorities

The bereaved are understandably often in a very distressed state, and great care and sensitivity should be exercised in ensuring that their grief is not made more acute. Any delays in certification would compound such grief, yet the doctor should not feel pressurized, and every effort should be made to follow the rules and regulations strictly. Religious observances and rites, and other social considerations and conventions may also be brought to bear on the doctor. Although the family of the deceased should be heard out with deference and respect, the certifying doctor should not compromise his or her position in any way. If the doctor cannot certify the death or is bound to report it for any other reason, then the case should be referred further.

Consent to a hospital autopsy and to the retention of organs or tissues therefrom can only be sought if the cause of death is known and all other legal requirements have been abided by. This consent should be given in writing by the next-of-kin on the forms designated for this purpose.

Other certification

In the United Kingdom the disposal of the deceased's body by cremation is covered by Statute and Statutory Regulations. For human remains to be disposed of by cremation a series of forms have to be endorsed by medical practitioners. Form B can be signed by any registered doctor and gives details of the death and its causes; form C, a confirmatory certificate, can only be signed by a doctor who has been fully registered for 5 years. Both doctors need to have inspected the body after death, have conferred, and the second doctor needs to have spoken to some other person who had treated, nursed, or been otherwise directly involved in the patient's last illness. The duty of the 'Medical Referee to the Crematorium' is to scrutinize these certificates and the application for cremation (Form A), and if fully satisfied authorize the cremation to proceed. In cases reported to the legal authorities the Coroner or Procurator Fiscal signs the forms that should otherwise have been signed by the two medical practitioners. In all instances, permanent pacemakers and radioactive implants have to be removed prior to cremation.

If the body has to be transported abroad or to other parts of Great Britain, there may be a requirement of further certificates to enable this to take place. These include certificates from the legal authorities enabling the movement of the body outwith their jurisdiction and confirming they require no further access to it, a 'freedom from infection' certificate and often an 'embalming certificate'.

Particular causes of death

Deaths resulting from and in the course of medical care

No matter how vigilant and caring medical and surgical treatment may be, occasional deaths will occur in the course of treatment as a direct consequence of the treatment. This may due to an allergic or idiosyncratic response to medication, and much more rarely through error, such as giving an excessive dose of a drug, or from accidents (for example, intra-arterial versus intravenous injection) and mishaps (for example, internal bleeding after a liver biopsy). For this reason, deaths that occur during operation or in the early postoperative period, deaths during investigative procedures, in the course of the administration of a general or local anaesthetic, or in the progress of clinical trials, invariably become the subject of a legal investigation.

Such deaths in the course of medical care raise the spectre of litigation and claims of medical negligence. There should be absolute transparency in divulging all the facts about such deaths, which should always be investigated by a pathologist who is completely independent of the hospital or other establishment in which the death has occurred, and who has some previous experience in such investigations. The investigating pathologist requires access to all the medical notes of the deceased patient, full statements from the doctors and nursing staff involved, detailing their involvement (disconcertingly, often acquired by the police), a thorough examination of any equipment used, access to batches of drugs used and blood samples collected **premortem**. Such investigations will invariably require the

assistance and participation of a number of 'experts' in other fields.

Sudden infant deaths

The careful investigation of death in infancy and childhood has led to major successful prophylactic campaigns; perhaps if unexpected adult deaths were to be looked at as carefully, similar preventive measures could be implemented. These deaths have also led to the production of universally acceptable protocols for postmortem examination that involve photography, radiology, microbiology, virology, immunology, genetics, etc. If this approach were to be emulated in other death investigations, the end-product therefrom would be much enhanced.

As infectious diseases no longer take a major toll in infancy in most developed countries, and as serious congenital conditions are no longer as prevalent, the most important cause of the sudden death of infants after the first month of life and within the first year is the 'sudden infant death syndrome' (**SIDS**, sometimes referred to as 'cot death' or 'crib death'). The original definition, still applicable, is of 'sudden death of any young child that is unexpected by the history, and in which a thorough post mortem examination fails to demonstrate an adequate cause of death'. Most pathologists would also wish to have a thorough inspection of the scene of the death to exclude potentially noxious environmental agents, for instance carbon monoxide exposure.

The diagnosis of SIDS is therefore only made by carefully and meticulously excluding any other causes, and is a morbid anatomical diagnosis rather than a clinical one; an autopsy is therefore a *sine qua non* to reach this diagnosis. It is especially important to exclude congenital metabolic abnormalities such as medium-chain acyl coenzyme A dehydrogenase (**MCAD**) deficiencies by appropriate testing (plasma/blood spot acylcarnitine profiles in MCAD deficiencies) and thus alert the family to possible further recurrences. Trauma and poisoning by alcohol, or with other sedative or anxiolytic preparations, also have to be excluded specifically on appropriate autopsy samples; the 'Münchhausen syndrome by proxy', first described by Meadow in 1977, is another condition to be aware of.

The incidence of SIDS has decreased dramatically in many industrialized countries as a result of major public health educational campaigns advising parents about the risk factors and means of prevention. Overheating of the child is one such risk factor. Parents are instructed to prevent this by removing the child's headgear when indoors, and by ensuring that the sleeping child does not wear excessive clothing or has too many bedclothes, by preventing the ambient bedroom temperature from being too high, and by seeking medical advice when the child appears feverish. The dangers of cigarette smoking close to the baby have also been emphasized as an important risk factor for this and other childhood complaints. The 'back-to-sleep' campaign, that is ensuring that the child is placed in the prone sleeping position, is based on another epidemiologically established important risk factor. Paradoxically, immunization of the child for common childhood illnesses, once thought related to SIDS, has been shown to be protective and is thus further encouraged in this connection.

The actual pathogenetic cause of SIDS is still uncertain in the majority of instances: hypoxia, cardiac arrhythmias, hypoglycaemia, loss of vascular tone, reflex apnoea, and heat shock have all been proposed, as well as the 'superantigen' effects of bacterial toxins. Other important aspects of the epidemiology of this condition is the seasonal incidence, the familial recurrence, the increased incidence in boys, and the increased association with certain ethnic groups, such as Native Americans and Australian Aborigines, and its absence in others, for example immigrant families from the Indian Subcontinent. The pathognomonic feature of SIDS, which is yet to be explained, is the finding of diffuse, internal petechial haemorrhages overlying the thymus and beneath the pleura and the pericardium.

Sudden unexpected nocturnal deaths in adults

The syndrome of sudden unexpected nocturnal deaths (**SUND**) occurs in young adults and adolescents, mainly in immigrant workers from SE Asia employed in Singapore and Saudi Arabia, and in refugees from the Far East. These decedents are usually employed in manual jobs in the building or the gardening trades, and have been residing in their adoptive country for several months. They are almost exclusively males who smoke or who are passively exposed to smoke in their environment, and who have a recent history of a mild upper respiratory tract infection; petechiae may also be found internally. These deaths mostly occur during sleep and at night.

Various theories have been put forward, including vitamin B and other nutritional deficiencies, familial cardiac arrhythmias, brainstem epilepsy, *Pfeifferinella mallei* infection, stress and homesickness, bacterial toxin production from nasopharyngeal colonization by *Staphylococcus aureus,* but no specific and recurring cause has been identified.

It is also the case that a full autopsy, with comprehensive toxicological and histological investigations, in the occasional sporadic death of an adolescent native of this country fails to yield a cause of death. Such deaths have been labelled as 'deaths from SUND' in Great Britain. However, it must be kept strictly in mind that this diagnosis is also one of exclusion, and should be used sparingly and appositely.

Survivors of violence

Sexual assaults

Over the last few decades the police have appropriately received positive and favourable publicity regarding the manner in which they deal with the survivors of alleged sexual abuse, and investigate their formal complaints. This has enabled more of those who have been abused in this manner to come forward and report the abuse suffered. In spite of this, however, it is not infrequent that the first disclosure of such abuse, particularly abuse which had occurred some time previously—on occasions, several years earlier—is initially made to a doctor in the course of a confidential consultation, perhaps on a totally unrelated matter. The medical practitioner is thus placed in a situation in which they are party to highly confidential and sensitive information relating to a potentially very serious crime, whose investigation requires careful and specialized investigation. This requires the careful interviewing of the survivor of this crime, the description of general and genital injuries, the meticulous collection of trace evidence, and of course the presentation of all this expertly in the criminal courts.

In this situation—as in many other similarly problematic circumstances—the doctor has to determine for himself whether or not, in the eyes of the law, the person disclosing abuse has full competence to take decisions. In Scotland and the rest of the United Kingdom, the legal age above which legally valid consent to medical treatment can be given is 16 years. If the patient is competent in terms of age and of mental and physical faculties, then in all such instances the doctor must firmly put to them the option of immediately involving the police. The police are much more knowledgeable about the process of investigation of allegations of sexual abuse and better equipped for the purpose than any medical practitioner is likely to be. If after due consideration the patient does not wish to inform the police for any or no reason, then patient confidentiality must be maintained.

All efforts should be made to ensure that any problems related to possible sexual abuse are adequately dealt with, including any worries that the patient may have about the possibilities of an unwanted pregnancy, sexually transmitted diseases including HIV infection, and the physical and mental trauma sustained. If the police are brought into the picture, the patient may still require some further support.

In Scotland, below the age of 16 years, the consent by a minor is only competent if in the view of the particular medical practitioner that patient fully understands the implications of the treatment being offered, and this consent can be extended to consent to medical procedures. Elsewhere, medical decisions of all types are also governed by the child's understanding of the proposed line of action or procedure—referred to as ' *Gillick*-competence' in the United Kingdom. If the consent to involve the police is not forthcoming from a minor, then the doctor has to decide whether in the interest of the particular young patient, this decision should be overruled, and perhaps whether those with parental responsibility for the child are to become involved at this stage. This decision-taking tightrope has to be negotiated very carefully, with the best interests of the patient always paramount and with the medical practitioner acting 'in good faith'. Advice from more senior colleagues, from forensic medical practitioners, and from the medical defence unions is often invaluable in such instances.

In the case of young children who are 'not-legally competent', appropriate multidisciplinary guidelines have been put together by every health authority and health board. In these a close co-operation between the health services, the police, the education department, and the social work department forms the basis of the investigation and further management of these cases. These guidelines should be adhered to strictly.

If the police are not involved it may be very difficult to collect evidence that would stand up in a court of law. If the doctor is inexperienced in such examinations then their competence and expertise will be called into serious question by the courts in any eventual adversarial criminal court case; a gynaecological or surgical colleague may have to be involved. For instance, the taking of swabs for seminal fluid analysis may fall short in terms of the unbroken continuity of the chain of evidence and the exclusion of cross-contamination that the courts would always require. These difficulties should be brought specifically to the attention of any patient

who is reluctant to involve the police.

Another way to ensure that any physical evidence of injury to the genital area is recorded permanently at the time of the medical examination is to utilize videocolposcopy, as is now almost invariably performed in examinations of this type in prepubertal children to avoid problems with second examinations and nuances of varying interpretations.

It is a fact of life that a very significant number of prosecutions initiated in sexual assaults fail to produce a conviction. By the very nature of this crime, these incidents are usually perpetrated in private with no eye-witnesses to the event. It often boils down to the oral evidence given in court by the two parties, and to which of the two, tested by cross-examination, the members of the jury are prepared to accept.

The medical notes

On the principle that contemporaneous recording will always provide good evidence in court, and in many cases that will be the best evidence, it is essential that all members of staff keep regularly annotated and thorough medical notes that are adequate, comprehensive, and comprehensible. Conciseness is not an issue at all in these instances: the notes made may be telegrammatic, provided that they convey all that has transpired on that occasion in terms of how the patient was dealt with and managed.

It is important to record dates and timings, to ensure that each page bears the name of the patient (or a 'sticker' bearing their details).

In those patients who allege assault, or have been otherwise injured, the manner and method of presentation, as well as the triage procedures all have to be documented in full. It may be said that some of these issues are normally attended to by other clerical staff; however, it does no harm when urgency and the vagaries of practice has put the system out of kilter, and indeed sometimes saves the day, for the doctor to record such important details, or at least ensure that they have been properly recorded.

The narrative given of the presenting complaint should be carefully recorded at the time, and what is even more important, it is essential to document who gave the initial information that has found its way on to the notes. If it is the ambulance crew or the accompanying relative or police officer from which you obtained this information, indicate so, as this renders it second-hand or 'hearsay information' in a forensic context. If the patient has given you specific details about how their injuries were acquired, transcribe these into text. In cases that may end up in court it is important not to attempt to précis, filter, or alter the information as originally given, perhaps in an attempt to make it sound more plausible and coherent: it should be documented as it was imparted to the doctor at the time.

Always indicate the findings on clinical examination. State what you did and when, and the investigations that were carried out by you (for example, blood pressure, peritoneal lavage), or asked for, either at the accident and emergency department itself (such as breath testing for alcohol, urinalysis for blood), and/or elsewhere (such as blood gases, serum electrolytes). It is also important to ensure that when the reports of such tests are available that these are also quoted in the notes, even if this information has been given to you over the telephone.

If radiographs have been ordered, make sure that in addition to your own viewing thereof and a recording of the diagnosis made by you on your personal 'reading' of them at the time, that the films are also subsequently reported on by the radiology department. These reports will serve as confirmation of your diagnosis and should always find their way into the patient's notes. For instance, it is of little assistance if you believe that there was a fracture of the maxilla or of the nasal bones, both clinically and on radiography, but this has not been confirmed anywhere in the notes in an 'official' radiological report.

If there has been a referral to other units (for example, neurosurgery, maxillofacial surgery, or burns units), this must also be recorded, preferably with a copy of the letter of referral. Consultations with other colleagues, for instance the physicians on call, the otorhinolaryngological specialist registrar, should be fully documented for any future reference.

Describing wounds

It is absolutely essential that wounds in injured persons are carefully described in the medical notes. This holds true whether the presentation is in hospital or in primary care. Domestic violence is on the increase, and the setting up of primary care-manned units in the community is becoming more frequent. It is not a valid excuse to claim that because the doctor to whom the patient presented initially was not an accident and emergency doctor, there was no obligation on them to record appropriate details. (See Box 2.)

Box 2 Wounds

- **In a systematic way:**
- *For each wound:*
- Define the wound.
- Locate the wound in relation to fixed anatomical points.
- Measure the wound (with its edges in apposition).
- Describe its edges, its immediate surroundings, and its floor.
- Describe the wound in terms of its orientation or pattern.
- State whether recent or old.
- Discuss its severity, either individually or collectively.
- Record the baseline general physiological parameters of the patient, e.g. pulse rate, blood pressure, respiratory rate, Glasgow Coma Scale.
- Consider sketching the wound freehand or on a line diagram.
- Consider photographing the wound.

A wound in the medicolegal sense is any traumatically induced abnormality, and this ranges from erythema to abrasion to cuts through the skin or mucous membranes to any internal injury.

For each external 'wound', describe its shape (vertical, transverse, lozenge-shaped), and its exact location—the latter by reference to standard fixed anatomical sites (for example, suprasternal notch, the prominence of the seventh cervical vertebra), and not variable ones such as the nipple or the umbilicus.

The wound should be measured with some degree of accuracy, and if the wound happens to be oriented in any particular manner, this should also be recorded; similarly document any collar of abrasion and bruising around it, and any pattern in the wound itself.

In forensic practice, trivial wounds that may not require any active treatment may be as important as those that are more serious. Abrasions in the form of fingernail scratches may be as important as any full-thickness lacerations that may be coexistent on the same patient. Therefore try to refer to all wounds in your description.

Be careful to define each wound appropriately. By definition, a laceration indicates a wound with very irregular (and perhaps bruised) edges, and the presence of bridging of incompletely damaged tissue in its base. Furthermore, a laceration, again by definition, is also a wound caused by a blunt-force injury—that is to say a force that has stretched the skin excessively, and more usually over a bony point or surface, causing the elasticity of the epidermis and dermis or of a mucosal surface (such as the lip, the vagina, or anus) to be so exceeded that, as a consequence, there is splitting and tearing apart of the skin or mucosa at the site of application of the force.

A sharp and pointed object will produce an incision or an incised wound with clean-cut straight undamaged edges. However, it is often impossible to indicate what specific weapon did cause the injury; for example, a sharp shard of glass, a kitchen knife, and the sharp edge of a tin can all produce incised wounds which look identical, even to someone with plenty of experience in wound interpretation. In the case of lacerations and bruises the difficulty may be even more pronounced. All that one would be able to say with any degree of accuracy is that among other objects that could have produced the wound, its appearances are consistent with having been produced by the particular weapon that is being suggested.

If there is a penetrating stab wound, it is essential to record the length of the wound track after it has been probed or explored, also to what depth the wound has extended, and, if this has been identified, which direction the track leads away from the skin. If the wound has penetrated beyond the skin, it would be important to denote which layers have been breached.

Although it may prove impossible to record individually all the wounds sustained, the use of simple line-drawings with the inclusion thereon of brief comments may be very effective.

Also recall that fractures, dislocations, and internal injuries fall under the category of wounds in this context. Any foreign bodies which have been retrieved from wounds, no matter how banal they may look (for example, grit, glass), may have very important evidential value and should never be discarded.

Wounding in the legal context

The legal practitioner, when looking through descriptions of wounds, often has different priorities and different questions from medics in their mind. Occasionally, these may not be immediately apparent or deemed relevant by the medic, but these queries may be expressed in writing or in court. Samples of such questions include: 'How much force was required to produce the wound under review?'; 'Could the particular wound have been inflicted accidentally, or as part of a self-defence type of response by the patient to an assault on him?'; 'Was the wound inflicted by a right-handed or left-handed assailant?'; 'Were all the wounds inflicted in the course of one assault?'

It is the counsel of perfection only to answer such questions if one feels experienced and fully competent to so do. An off-the-cuff remark on such matters that cannot be substantiated on robust cross-examination may cost dearly in lost face, and perhaps even in reputation, in the witness box.

It is important to be very circumspect about the ageing of wounds; interindividual variation is such that one can only provide general answers to questions on this matter. If a wound is showing healing as demonstrated by scabbing, then it will be about 2 days old, and one which is scarred almost a week old, but statements must be as general as that. Ageing of bruises is particularly fraught, in that the colour change that can be seen as the haemoglobin that has extravasated into tissues is changed to bilirubin and biliverdin, depends on a number of variable local and systemic factors.

Of specific importance to the criminal justice system are also such matters as the severity of the wound in question, and whether a particular wound could be considered as being life-threatening. On the basis of the 'soil and seed' concept, severity should be assessed by the damage produced by the trauma, the amount of blood lost, the degree of surgical shock present; also on the amount of days lost from work, the age of the patient, associated medical conditions that decrease the rate of healing, the ease with which it could be dealt with medically, etc. Indeed, any answers to questions about severity should always be predicated by a series of reasons indicating why the opinion given is being proffered. In doing so it may be useful to distinguish between whether or not the particular wound is 'serious' and 'life-threatening', or whether wounds in that specific anatomical location (for example, neck, anterior chest wall) in general terms are serious and life-threatening. For example, a penetrating stab wound of the chest may not have actually produced a pneumothorax, but had it been slightly deeper or its track slightly more medial or more lateral it could have: thus within these caveats, the wound can be considered as serious and potentially life-threatening. The fact that a particular injury can be salvaged with relative ease in a hospital does not necessarily detract from its degree of severity. Any inevitable or avoidable delay in seeking or obtaining medical help, any intervening wound infection, etc., should also be listed to enable a more balanced assessment of wound severity.

The after-effects of the wounding are also of importance. Any scarring left behind by the wounding, which may be considered as cosmetically disfiguring, even if surgical in origin, can increase the 'legal' severity of the wound, and similarly any residual pain and stiffness of an injured joint.

Photographs of injury

Photographs of wounds may be extremely useful if in due course the case comes to a court hearing: pictures taken prior to stapling or suturing may clearly convey to a jury more poignantly the degree and variety of injury sustained. This may require close co-operation with the police, and above all the 'informed' consent of the patient, if at the time they are in a state in which they are legally capable (*capax*) to give this. If the patient is unconscious and the case is very likely to have been the result of criminal violence, there should be close co-operation with the police. If the management of the patient will not suffer adversely, any reasonable requests for photography made by the police should be considered—even if only a Polaroid—and if at all possible acquiesced to. In all such instances there must always be a careful and considered balance between one's professional obligation as a medical practitioner to provide optimal care and confidentiality for the patient, and one's duty as a citizen of a country in which violence cannot be condoned and for which its perpetrators are brought to justice in the course of a fair trial.

Photographs of wounds after they have been debrided and sutured may not be as useful as photographic documentation prior to such treatment, but they are better than nothing. Patterned injuries which may have to be matched to other weapons (for example, footwear imprints, imprints from blows) may need to be photographed in black and white, in colour, and under different light sources (such as ultraviolet light). If photography cannot be used for any reason, then consider sketching the wound freehand or use anatomical outline drawings to indicate the location and appearances: this will also economize on text.

Human bites

On occasions, human bites may be the presenting injuries. In these instances photography may be essential, and furthermore, valuable information can be gained from appropriately thorough and specialized examination. There may be enough material on the skin to secure a DNA profile of the perpetrator, and an odontological opinion may be able to produce a dental chart of the offending jaws for matching purposes. This cannot be done without the involvement of the police at an early stage, and, as always, the patient's own informed consent would be required.

Intoxication

Alcohol tends to feature prominently in persons who have been assaulted or accidentally injured, and it may be useful for the purposes of the courts to document the degree of intoxication observed in the victims or in the perpetrators. Although 'alcohol' can often be smelled on the breath, it must be remembered that it is not ethanol that is being picked up, but congeners such as esters and other organic compounds that have been consumed together with the alcohol. As with alcohol, these are excreted for a lengthy period after drinking has stopped.

Alcohol is absorbed from the upper gastrointestinal tract, mostly the duodenum, and disseminated uniformly and in an unimpeded fashion throughout all body compartments that contain water. Hence, the amount of water in the body, which relates to body weight and to gender, will influence the eventual concentration and therefore the effects of the alcohol consumed. Typically, the consumption of 1 pint (568 ml) of ordinary beer or a double public-house measure (about 55 ml) of spirits (40 per cent (v/v)—alcohol concentration) will result in a blood alcohol (ethanol) concentration of 30 mg of alcohol per 100 ml of blood, or 13 µg of alcohol per 100 ml of breath. The corresponding legally prescribed limits for driving in Britain are 80 mg for blood and 35 µg for breath.

One should only carry out a formal blood alcohol estimation if this has potential therapeutic indication, and then only with the knowledge and consent of the patient. Alcohol is a CNS depressant and the effects of alcohol intoxication can be elicited by tests of neuromuscular coordination and of higher central functions, including the Glasgow Coma Scale. The eyes will also show evidence of sustained lateral nystagmus and the pupils will be dilated.

In those who may be under the influence of 'controlled substances', with or without additional alcohol intoxication, a full neurological examination should be carried out and recorded. However, it is often unhelpful and profligate to carry out urinary or blood tests for the presence of drugs.

Access to information

The same rules of medical confidentiality apply wherever the patient is seen. Relay of information to others has to be carefully controlled. The police frequently seek information, either acutely and/or after the patient has been discharged. It is therefore important that some basic rules are laid down, enabling the police to know what information will and what will not be divulged to them, without the patient first having been approached and their formal consent obtained.

In terms of statute in the United Kingdom, the police have every right to ask for full details of those persons whom they believe have been driving a 'mechanically

propelled vehicle' which has been involved in a collision, and who have been admitted to hospital. Similarly, in relation to acts of terrorism, the police have a right to obtain information. In other instances it is a question of whether, in terms of their inquiries, information should be divulged to police officers who are seeking it. The admission of a person who is likely to die (or as they would put it, 'a condition that is likely to prove fatal'), whether this is the result of an accident or a criminal assault, is cause enough to bring information to the attention of the police.

If the patient or their legal representative requests a copy of the medical notes, this legitimate request cannot be refused unless it can be shown—if need be to the scrutiny of a Crown Court judge (or a Sheriff, in Scotland)—that disclosure of the notes may disclose the identify of a third party or be detrimental to the physical and/or mental heath of that particular patient. Barring this, such records should be handed over: staff in local hospital edical records' departments are trained to deal with such requests appropriately.

The forensic use of molecular biological techniques—DNA profiling evidence

DNA-based evidence has revolutionized forensic practice over the last few years. Based on the principles that all cells of an organism are derived from one fertilized ovum, and that mitotic division of cells is uniform and precise, all cells inside an organism that contain a nucleus retain an identical DNA content. Thus, DNA from hair, buccal cells, semen, and white blood cells derived from the same individual is identical, and it is possible to determine the exact origin of a particular cell or group of cells if their DNA can be matched to that of some other cells of the same origin. Furthermore, as half of a cell's DNA has originated from each of the parents, it is also possible to derive the genetic origin of a particular cell by comparison of its DNA content with that of its parents.

Variable number of tandem-repeat loci

Groups of DNA loci that are used extensively in forensic analysis are those counting variable numbers of tandem repeats (**VNTR**). These are not genes, since they do produce any known product, and those used in forensic analysis have no known biological role. They are thus less likely to be influenced by natural selection, which can lead to different frequencies in different populations. A typical VNTR region contains 500 to 10 000 base pairs, containing many tandemly repeated units, each 15 to 35 base pairs in length. The exact number of repeats, and hence the length of the VNTR, varies from one allele to another, and different alleles may be identified by their relative lengths. VNTRs have a very high mutation rate, leading to changes in their length, with an individual mutation usually resulting in a change in length by only one or a few repeating units. This leads to a very large number of alleles, often 100 or more, no one of which is common, although only 15 to 25 can be distinguished practically. This means that the number of possible pairs of alleles forming the genotype at a locus is considerable, and given that testing of several different such loci can be combined, the total number of genotypes becomes enormous. For example, for n alleles there are n homozygous genotypes and $n(n-1)/2$ heterozygous ones, in other words: if $n = 20$, there are a total of 210 genotypes, and if four loci are examined with 20 alleles each, then 210^4 or about 2 billion genotypes are possible (assuming that all four alleles are inherited independently).

The main uses of the DNA profiling method in forensic cases are:

1. The identification of crime suspects from trace evidence left behind at the scene, e.g. a specimen of blood from the deceased is found to match stains on the clothing, etc. of the accused person.
2. The elimination of crime suspects in crimes where there has been deposition of body fluids, e.g. the DNA profile of the seminal sample taken from the vagina of the victim does not match that obtained from the white cells of the peripheral blood of the alleged perpetrator of the rape.
3. Paternity testing, when it is necessary to establish which one of two or more males is the actual biological father of a particular child, e.g. in incest or rape cases. Fetal tissue is also suitable for such testing (and is often used).
4. Identification of an unknown person or of unknown mutilated human remains, by comparing material taken at postmortem examination with material, such as hair or blood, which is authenticated as belonging to a particular person during life. This has been extended to buried remains, e.g. the Romanovs, Mengele.
5. Mass disasters—this technique assists with the identification of individuals provided a premortem sample is available for comparison, and has the ability to match together different and separated parts of the same body.
6. Mass screening ('a genetic man-hunt') of a well-circumscribed population from which the murderer or rapist is known to have originated. This was the first use of DNA profiling in a criminal context in the United Kingdom.
7. DNA databases—allowing crimes committed by the same perpetrator in which body fluids have been deposited at the site of the crime to be associated, i.e. serial crimes and previous offenders to be linked with a specific 'new' crime, or resolution of historical unsolved cases.
8. Disputed maternity, e.g. in cases of infanticide and child destruction, when it requires to be proven that a particular child is indeed the offspring of a particular woman.
9. Settlement of immigration problems in relation to the admission into a country of blood relatives rather than 'friends'.

In the civil courts, DNA profiling has also been proving useful in such cases as:

1. divorce (associated with alleged adultery);
2. disputed paternity (and more rarely maternity): in settling estates after death;
3. in immigration disputes (when a country only allows entry of certain closely related relatives, born outside that country, of newly established residents);
4. in disputed pedigrees of animals and origins of biological material.

The forensic applications of DNA profiling are not universally available, and therefore these tests are yet to completely replace and supplant conventional blood grouping methods involving the ABO and isoenzyme (for example, **PGM** (phosphoglucomutase)), systems. In some countries such as the United Kingdom and Switzerland this has already taken place.

The important limitation of DNA profiling is the amount of DNA-containing material available for carrying out the appropriate testing. This is particularly the case when dealing with peripheral blood: only the leucocytes within it that can be of assistance, meaning that a substantial quantity of these cells must be available.

Polymerase chain reaction techniques

In 1987 the polymerase chain reaction (**PCR**) technique was introduced, enabling the rapid and specific, *in vitro*, enzymatic amplification of DNA fragments. This laboratory synthesis utilizes a heat-stable enzyme, DNA polymerase (formerly obtained from a thermophilic bacterium found in hot natural springs and called '*Thermus aquaticus*'—thus '*Taq* polymerase'). The other essential reagents are primers, which are small complementary fragments of single-stranded DNA, also produced artificially. With the correct reaction conditions, a pair of primers can be made to attach themselves to single complementary strands of DNA, and in the presence of an excess of bases in the solution, marshal the formation of replicas of the original DNA fragments. Repeated reaction cycles can be conducted until sufficient quantities of the product are formed to enable standard DNA profiling techniques to be carried out.

The PCR technique is so sensitive that even a very small amount of tissue, such as a single hair root, a single buccal cell, or a single spermatozoon, may be sufficient to produce a DNA profile that is adequate for matching purposes. It can also amplify denatured DNA, and even material from paraffin-embedded histology blocks and formaldehyde-fixed tissue is suitable for replication.

Very strict control of laboratory technique is required in conducting PCR procedures as any minute amount of DNA present in the sample will be replicated. Controls are of the essence in demonstrating the absence of contamination. Protective clothing is necessary when evidence is being collected at scenes of crime, as is careful attention to detail in terms of the collection of material for DNA analytical procedures.

HLA-DQa

A further development is related to one of the genes that controls transplant rejection, *DQA*, coding for a protein HLA-DQa that shows substantial variation in its base sequence from one individual to another. The most variable fragment of this gene may be readily amplified by PCR to distinguish eight different alleles, of which six are used in forensic practice, and thus 21 different combinations of two alleles—6 homozygous and 15 heterozygous. In practice, a reverse blot is used, with the nylon membrane containing preattached probes specific for the individual alleles, making a quick, reliable, and very useful preliminary test if one is required. On average the DQa profile of a person is identical with that of about 7 per cent of the population.

Mitochondrial DNA

Mitochondrial DNA contains a segment called the control region, which is highly variable and has the following additional properties:

1. This DNA can survive in extensively decomposed tissues.
2. It can be successfully amplified, even from the most unpromising tissues such as bones.
3. It is strictly maternally inherited.

Analysis of mitochondrial DNA has been used most frequently in looking at historical cases, and in instances where nuclear DNA is in short supply or where it has been denatured by, for example, contact with soil and the bacteria therein.

Other advances

Microsatellites or short tandem-repeat loci (STRL)

Microsatellites are much shorter than the minisatellites (VNTRs) and comprise 2- to 4-base pair repeats only. They are very common, are distributed widely throughout the genome, and can be amplified singly or together by PCR. Individual specificity can be achieved by typing several loci sequentially. The method can be used with severely decomposed and burnt bodies, using the DNA from a portion of spared voluntary muscle or red bone marrow.

Amplitype polymarker (PM) DNA

This method analyses several loci simultaneously: **LDLR** (low-density lipoprotein receptor), **GYPA** (glycophorin A, the MN blood groups), **HBGG** (haemoglobin gammaglobulin), D7S8 (an anonymous genetic marker on chromosome 7), and **GC** (group-specific component). Each has two to three alleles per locus.

Minisatellite repeat mapping or digital typing

This analyses for length variation and detects sequence differences within the base sequences repeated in VNRTs.

DNA identikit

In the future it may be possible to describe physical (for example, blue eyes, red hair) or other characteristics of a subject from their DNA sequence, thereby obtaining hints as to their actual identity.

Estimation of the population frequency of a DNA pattern

DNA 'exclusions' are easy to interpret: if technical artefacts can be excluded, a non-match is definitive proof that two samples have come from two different sources. However, 'DNA inclusions' cannot be interpreted without knowledge of how often a match might be expected to occur in the general population at random. In simple terms, if two DNA profiles match then there are two logical possibilities: first, that the DNA profile at the scene, on clothing, etc. is actually that of the suspect; and second, that it comes from someone else who has the same profile as the suspect. The commoner the particular DNA profile, the greater is the likelihood that it could come from someone other than the suspect; by contrast, if the suspect's profile happens to be a rare one, then the chances of this are much less likely. Thus the frequency of a particular profile in a particular population must be known to enable this comparison.

The statistical analysis of profiling data has caused problems to the courts: judiciary, jurors, and lawyers alike. Population frequencies of various profiles have been established, a standard method being to count occurrences in a random sample of the appropriate population and then to use classical statistical formulas to place upper and lower confidence limits on the estimate. Because estimates used in forensic science should avoid placing undue weight on incriminating evidence, an upper confidence limit of the frequency should be used in court, which is appropriate in the forensic context because any loss of power can be offset by studying additional loci.

The product rule should be used to estimate the frequency of a particular DNA profile frequency. If the race of the particular individual is known, the database for that particular race should be used; if not, calculations for all prevalent racial groups should be used. The probability that two randomly chosen individuals have a particular phenotype is the square of its frequency in the population. The probability that two randomly chosen persons have the same unspecified genotype is the sum of the squares of the frequencies of all the genotypes. If there are *n* loci, and the sum of the squares of the genotype frequencies at locus 1 is *p*1, then the exclusion power is (1 minus (*p*1, *p*2.....*pn*)).

Population frequencies that are quoted for DNA purposes are not based on actual counting but on theoretical models based on principles of population genetics. Each matching allele is assumed to provide statistically independent evidence, and the frequencies of the individual alleles are multiplied together to calculate a frequency of the complete DNA pattern. Although a databank might contain only 500 persons or less, multiplying the frequencies of enough separate events might result in an estimation frequency of their all occurring in the same person of $1{:}10^{9}$. However, the scientific validity of this multiplication rule depends on whether the events are actually statistically independent. In organizing databanks it is essential that ethnic groups are considered separately, and in the United Kingdom databanks for Caucasian, Afro–Caribbeans, Indians, Pakistanis, Chinese, etc. have yet to be established.

Further reading

Aitken CGG (1995). *Statistics and the evaluation of evidence for forensic scientists*. Wiley, Chichester.

Balarajan R, Reileigh VS, Botting B (1989). Sudden infant death syndrome and post neonatal mortality in immigrants in England and Wales. *British Medical Journal* **298**, 716–20.

Balding DJ, Donnelly P (1994). How convincing is DNA evidence? *Nature* **368**, 285–6.

Beckwith JB (1970). Discussion of terminology and definition of the Sudden Infant Death Syndrome. In: Bergman AB, Beckwith JB, Ray GC, eds. *Proceedings of the Second International Conference on the Causes of Sudden Death in Infancy*, pp 14–22. University of Washington Press, Seattle, WA.

Blackwell CC (1999). Sudden infant death syndrome. *FEMS Immunology and Medical Microbiology* **25**(1–2), Special issue.

Blackwell CC, *et al*. (1994). SUND among Thai immigrants in Singapore: the possible role of toxigenic bacteria. *International Journal of Legal Medicine* **106**, 205–8.

Budowle B, *et al.* (1995). Validation and population studies of the loci LDRL, GYPA, HBGG, D7S8 and Gc (PM loci), and the HLD-DQa using a multiplex amplification and typing procedure. *Journal of Forensic Science* **40**, 45–54.

Busuttil A (1993). Domestic violence. In: Mason JK, ed. *The pathology of trauma*, 2nd edn, Chapter 10, pp. 121–37. Hodder and Stoughton, London.

Comey CT, *et al.* (1993). PCR amplification and typing of HLA-DQ a gene in forensic samples. *Journal of Forensic Science* **38**, 239–49.

Evett IW, *et al*. (1996). Establishing the robustness of short-tandem-repeat statistics for forensic applications. *American Journal of Human Genetics* **58**, 398–407.

Hazelwood RR, Burgess AW, eds (1995). *Practical aspects of rape investigation*. CRC Press, Boca Raton, FL.

Jeffreys AJ, *et al*. (1991). Minisatellite repeat coding as a digital approach to DNA typing. *Nature* **354**, 204–9.

Karlsson T, Ormastad K, Rajs J (1988). Patterns in sharp force fatalities—a comprehensive forensic medical study. *Journal of Forensic Science* **33**, 448–61.

Kevorkian J (1961). The *fundus oculi* as a '*post mortem* clock'. *Journal of Forensic Science* **6**, 261–8.

Knight B, ed. (1995). *The estimation of the time since death in the early post-mortem period*. Edward Arnold, London.

Lander S, Budowle B (1994). DNA fingerprinting laid to rest. *Nature* **371**, 735–8.

Langlois NEI, Gresham GA (1991). The ageing of bruises: a review and study of colour changes with time. *Forensic Science International*, **50**, 227–38.

Millroy CM, Rutty GN (1997). If a wound is 'neatly incised' it is not a laceration? *British Medical Journal* **315**, 1312.

Morris JA, Haran D, Smith A (1987). Hypotheses: common bacterial toxins as a possible cause of the sudden infant death syndrome. *Medical Hypotheses* **22**, 211–22.

Ormstad K, *et al.* (1986). Patterns in sharp force fatalities—a comprehensive medical study. *Journal of Forensic Science* **31**, 529–42.

Pallis C (1983). *The ABC of brain stem death*. British Medical Journal, London.

Rao VG, Wetli CV (1988). The pathological significance of conjunctival petechiae. *American Journal of Forensic Medicine and Pathology* **9**, 32–4.

Raza NW, *et al.* (1999). Exposure to cigarette smoke, a major risk factor for SIDS: effects of cigarette smoke on inflammatory responses to viral infection and bacterial toxins. *FEMS Immunology and Medical Microbiology* **25**, 145–54.

Royal College of Physicians and the Royal College of Pathologists (1982). Medical aspects of death certification. A Joint Report of the Royal College of Physicians and the Royal College of Pathologists. *Journal of the Royal College of Physicians, London* **16**, 205–18.

Tomlin PJ (1967). 'Railroading' in retinal vessels. *British Medical Journal* **3**, 722–3.

Valdes-Dapena M (1992). A pathologist's perspective on the sudden infant death syndrome. *Pathology Annual* **27**, 133–64.

Webb E, *et al*. (1999). A comparison of fatal with non-fatal injuries in Edinburgh. *Forensic Science International* **99**, 179–87.

Weir BS (1993). DNA fingerprinting report. *Science*, **260**, 473.

Weir BS, Hill WG (1993). Population genetics of DNA profiles. *Journal of Forensic Science* **33**, 219–26.

Wilson MR, *et al.* (1993). Guidelines for the use of mitochondrial DNA sequencing in forensic science. *Crime Laboratory Digest* **20**, 69–77.

Winton R (1982). The Declaration of Lisbon. *Medical Journal of Australia* **1**, 101–4.

Wroblewski B, Ellis M (1970). Eye changes after death. *British Journal of Surgery* **56**, 69–72.

28Sports and exercise medicine

R. Wolman

Introduction

Traditionally, sports medicine has concentrated on injuries that occur during exercise, and therefore has come under the umbrella of orthopaedic surgery. However, with the pursuit of sporting excellence a range of different exercise-related medical disorders are now recognized. These are associated with intense levels of training and may have a detrimental effect on long-term health, as may drugs (and nutrients) that are frequently taken to enhance training and performance. Physicians are therefore increasingly being confronted with medical problems related to sport. This section will cover some of the medical disorders that occur with sport and physical training. It will also address the use of drugs and some of the common overuse injuries that occur with sport.

Developments in exercise physiology over the last 30 years have led to improved training regimes for athletes. A spin-off from this has been the recognition of the benefits of exercise in health promotion and disease management. Exercise prescription now forms an important part of some treatment programmes, with evidence for its use in the management of a range of disorders, including heart disease, diabetes, obesity, hypertension, osteoporosis, arthritis, back pain, chronic fatigue syndrome, and depression. These will be covered in other sections.

Female athlete triad

History

Amenorrhoea in athletes was first recognized in the late 1970s. Prior to this it was unusual for women to train sufficiently hard to develop this syndrome. Since 1980 there has been a growth in the popularity of aerobic sports and in the number of endurance competitions for women, whose first Olympic marathon was in 1984 and first 10 000 metres was in 1988. The other important factor has been the fashion for thinness, which really began in the 1970s. These two changes in female behaviour are the main factors responsible for the development of this syndrome.

In the early 1980s it was thought that training intensity was the main underlying aetiological factor, but studies in the late 1980s indicate that a high proportion of these athletes also have disordered eating habits. It was also thought that the bone density of amenorrhoeic athletes would be normal, as the high levels of exercise would compensate for the low levels of oestrogen. However, studies from 1984 onwards have shown that amenorrhoeic athletes have low bone density. This female athlete triad therefore consists of disordered eating, amenorrhoea, and osteoporosis.

Incidence and aetiology

The female athlete triad is associated with endurance sports. The incidence of amenorrhoea varies in different sports and is a reflection of the requirements for that particular activity, in terms of training intensity, calorie restriction, and age group (Fig. 1).

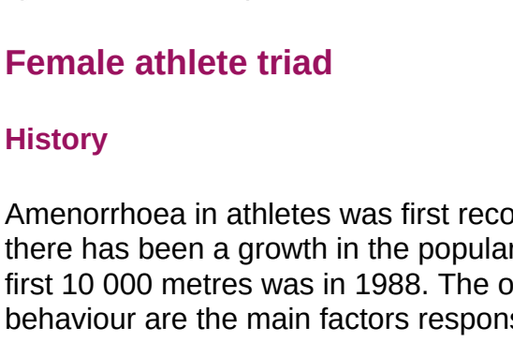

Fig. 1 The incidence of amenorrhoea amongst élite athletes in different sports. (Data taken from Wolman RL, Harries MG (1989). *Clinical Sports Medicine* **1**, 95–100, with permission.)

Training intensity can be difficult to quantify in certain sports, but in runners it is relatively easy as the number of miles run per week provides an accurate estimate. Some work has shown that the incidence of amenorrhoea increases as weekly training mileage increases, with an incidence of about 50 per cent in those running more than 80 miles per week.

Disordered eating is commonly seen in female athletes and may occur in over 60 per cent of competitors in sports such as gymnastics. In many cases this is the result of constant pressure from coaches, and sometimes parents, to maintain a prescribed body weight and appearance. In some of these cases the eventual outcome may

be an overt eating disorder and anorexia nervosa. The two most relevant nutritional deficiencies are of calories and calcium.

The importance of calorie restriction is seen in rowers, where the incidence of amenorrhoea is significantly higher amongst lightweights (who have to be below 59 kg in order to compete) than their heavyweight counterparts. Both groups have similar training regimes, but the lightweights frequently consume restricted diets in order to 'make the weight' for competition. Furthermore, nutritional studies on runners show that those with amenorrhoea have a lower daily calorie intake than their eumenorrhoeic counterparts.

Age is also important, with athletes in their late teens being more vulnerable to menstrual irregularity than those in their twenties. In activities such as gymnastics and ballet, where there are many teenage performers, there is a high incidence of amenorrhoea, both primary and secondary.

Pathophysiology

Endurance training is associated with menstrual dysfunction. At relatively low levels of training a shortened luteal phase may occur, which can be associated with reduced progesterone levels and anovulatory cycles. These abnormalities become more frequent as training intensity increases and eventually cycles may become irregular (oligomenorrhoea) or absent (amenorrhoea). In those sports where training starts before puberty there is often a delay in the menarche. Typically this is seen in gymnasts and ballerinas where, on average, the menarche is delayed by 1 year. However, primary amenorrhoea may result, and in some cases the athlete may be in their twenties before menstruation begins.

There are many similarities between the female athlete triad and anorexia nervosa, and many believe that these are part of the same spectrum of ill health. The psychological profiling of patients with both disorders is very similar. Furthermore, in both there is disordered eating, energy imbalance, and low body weight. The aetiology of the amenorrhoea is also similar, with slowing of the gonadotrophin-releasing hormone (**GnRH**) pulse generator in terms of both amplitude and frequency, leading to a reversible hypogonadotrophic hypogonadism with severe impairment of oestrogen production, which seems to be a reversion to the prepubertal pattern of gonadotrophin release.

Over the last 15 years or so, attention has been directed towards exploring the factors responsible for the suppression of the GnRH pulse generator. There are several hypotheses, including endorphin release, central 'stress', and energy deprivation, in each of which hormones are released that may influence hypothalamic function, including cortisol, insulin-like growth-factor binding protein-1 (**IGFBP-1**) and leptin (aside from endorphin). The inhibitory action of opioids on the GnRH pulse generator is now well established. However, although endorphin levels increase with acute aerobic training (which may account for the so-called 'runners high'), there is much less evidence that they remain elevated with regular exercise, hence endorphin release alone is unlikely to account for the gonadal suppression seen in athletes.

Serum cortisol levels are elevated in amenorrhoeic athletes compared to their eumenorrhoeic counterparts. This represents central 'stress' and occurs as a result of the central activation of corticotrophin-releasing hormone (**CRH**). CRH increases GnRH sensitivity to opioid inhibition, and therefore in combination with endorphin release provides a possible mechanism for amenorrhoea. The raised cortisol level may also adversely affect bone density (see below).

An alternative hypothesis is that amenorrhoea occurs as a result of energy deprivation (Fig. 2). Several independent studies have demonstrated that the energy (calorie) intake of amenorrhoeic athletes is significantly lower than their eumenorrhoeic counterparts, even when matched for exercise intensity. This produces a relative energy imbalance, that is to say a low-energy intake in the diet compared to a high-energy output in the form of exercise, which results in weight loss. Amenorrhoeic athletes, like those with anorexia nervosa, have a low body mass index (usually below 18) and low body fat (usually below 17 per cent). Furthermore, they have a lower resting metabolic rate and reduced levels of insulin, insulin-like growth factor-1 (**IGF-1**), and tri-iodothyronine. These changes are seen in situations of energy deprivation, but it is uncertain whether any of these three hormones act as a metabolic signal influencing the release of GnRH.

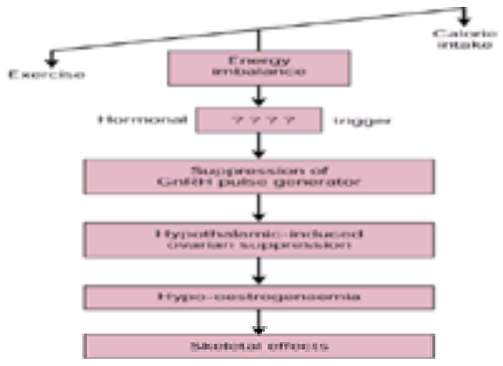

Fig. 2 Energy deprivation hypothesis for the female athlete triad.

Two other possible candidates are IGFBP-1 and leptin. The level of IGFBP-1 is directly suppressed by insulin, and in amenorrhoeic athletes this is elevated. Serum leptin levels are related to body fat and are reduced in anorexia nervosa and probably in the female athlete triad. Further work is needed to determine whether either of these hormones have a direct effect on the GnRH pulse generator.

Skeletal effects

Bone density

The bone density of amenorrhoeic athletes is reduced. This is seen most obviously in the spine where there is a high proportion of trabecular bone, but other sites such as the proximal femur and wrist may also be affected. The fall in bone density is predominantly due to low oestrogen levels, but other factors may be important since bone density does not always improve when oestrogen replacement is given (see below). Levels of IGF-1, which has important anabolic effects on the skeleton, are reduced, whilst those of cortisol, which has a catabolic action, are elevated in the female athlete triad: either of these hormonal changes could enhance the fall in bone density.

The type of exercise undertaken modifies the fall in bone density. For example, amenorrhoeic rowers have a higher spinal bone density than amenorrhoeic runners, presumably due to intense exercise involving the trunk. Amenorrhoeic gymnasts have higher spinal and femoral neck bone density than amenorrhoeic runners, probably due to weight training and jumping activity involved in gymnastic training.

With short episodes of amenorrhoea (6 to 12 months) the fall in bone density is reversible once normal menstruation is restored. With longer episodes, however, the changes may become irreversible and bone density remains persistently low. Occasionally the bone loss is severe, leading to bone densities similar to those seen in postmenopausal women, and in this subgroup there is a significant risk of osteoporotic fracture. More commonly, bone density reduction is less extreme and the risk is of premature osteoporosis (10 to 15 years early).

Other skeletal effects

Stress fractures occur more frequently in amenorrhoeic athletes, which may be related to low bone density. Stress injuries to bone commonly occur in athletes: these are usually repaired, preventing the development of a full stress fracture. The repair mechanism may be less effective in those with amenorrhoea who have low levels of oestrogen and IGF-1, both of which are important in maintaining skeletal integrity.

Athletes with delayed menarche and primary amenorrhoea may have delayed skeletal maturation, including delayed epiphyseal closure that may increase the risk of

epiphyseal injury.

Investigation and management

It is important to exclude other causes of amenorrhoea. This will include taking an accurate history to establish a relationship between training and menstrual abnormalities. Investigations should aim to exclude other causes of amenorrhoea (see Chapter 12.2 and Chapter 12.8.1) and the serum tri-iodothyronine level should be measured, which is likely to be low in the female athlete triad. A nutritional screen is helpful to assess calorie and calcium intake, and bone density should be measured (Table 1).

Once the diagnosis is established the most effective treatment is to re-establish natural menstruation (Table 2). This can be achieved with a combination of reducing training intensity (with the help of the coach) and increasing calorie intake (with the aid of a dietitian). It is very important to educate the athlete about both the short- and long-term risks of remaining amenorrhoeic, otherwise many athletes will not accept this type of intervention. Psychological intervention may be necessary in those athletes where an eating disorder is apparent (see Chapter 26.5.5).

In those athletes who remain amenorrhoeic despite attempts at adjusting their training and diet, oestrogen replacement (either the oral contraceptive or hormone-replacement therapy) should be given. Unfortunately some athletes may have difficulty tolerating this. Furthermore, anecdotal experience suggests that this is not always effective, probably for the reasons given above. Calcium supplements should be considered, especially in those with low intakes, and vitamin D supplements may also be helpful. Experience with bisphosphonates and raloxifene in this age group is too limited to offer clear advice.

Progress should be monitored with bone densitometry. In those who remain amenorrhoeic either the progressive fall in bone density or recurrent injuries will eventually force them to make lifestyle adjustments in terms of training and nutrition. By then it may be too late, hence emphasis needs to be placed on education and counselling at an early stage.

Overtraining syndrome

Introduction

This condition and its associated symptoms are well recognized in the athletic population, but the pathophysiology is poorly understood. It tends to occur in athletes doing high-intensity endurance training and is rarely seen in those who partake in strength and power sports. The most common presentation is with underperformance, for example a worsening of the times for the athletes' favoured events. Often the initial response is to assume that the training is inadequate and therefore to increase the intensity even more. This will perpetuate the problem, eventually forcing the athlete to seek medical advice.

Aetiology

Athletes must train hard to improve their performance. This can lead to transient fatigue and underperformance, hence intensive training should be followed by a period of relative rest to allow regeneration and recovery. This cyclical method of training is called 'periodization' and produces adaptation, allowing progressive increases in training intensity and performance. During the period of intense training, known as 'overreaching', it is common for the athlete to complain of muscle soreness, fatigue, and stress-related symptoms. As part of the stress response serum cortisol may rise, whilst testosterone may decrease, and creatine phosphokinase may be elevated as a reflection of transient muscle damage. These clinical and biochemical features should recover during the rest period. Overreaching is a necessary part of training if the athlete's performance is going to improve.

In some situations the athlete may not allow sufficient time for rest in between periods of heavy training. This leads to under-recovery and prevents full adaptation to increased training loads. This may occur when there is a rapid increase in training intensity, with very prolonged training, and at times of stressful competition. It may also occur when the athlete is exposed to other stresses, including intercurrent infection, travel across several time zones, or other unrelated psychoemotional stresses. In the adolescent athlete it may be seen during a rapid growth spurt. This combination of stresses may lead to underperformance, which will usually respond to a rest period of 2 weeks, but if the athlete fails to recognize the early signs the symptoms become more severe and may require a more prolonged rest period to achieve full recovery.

Pathophysiology

Over the last 15 years or so there have been many studies investigating the pathophysiology of the overtraining syndrome, and these have generated several hypotheses. Unfortunately studies are not always directly comparable because they use different definitions for the syndrome and may perform their evaluations at different stages in its evolution.

Neuroendocrinological features

There are effects on the hypothalamic–pituitary–adrenal axis. In some overtrained athletes there is a rise in salivary cortisol levels and a low testosterone:cortisol ratio. This reflects a rise in free-cortisol levels, which correlates with the depressed mood state seen in some athletes. There is a decreased pituitary ACTH response to insulin-induced hypoglycaemia, and reduced adrenal responsiveness to ACTH. The latter effect leads to a compensatory elevation of ACTH in the early stages of the overtraining syndrome, but in an advanced stage both ACTH release and the cortisol response may be reduced. There is also evidence of sympathetic involvement, with increased resting noradrenaline (norepinephrine) levels associated with decreased nocturnal excretion.

Central fatigue and the branched-chain amino acids

Prolonged endurance exercise leads to a depletion of glycogen stores in muscles and the liver, at which point muscle requires alternative sources of energy. Mobilization of fatty acids provide this, producing increased levels in plasma, where they are bound to albumin. The branched-chain amino acids (valine, leucine, and isoleucine) are another alternative fuel for muscle, with an increased rate of oxidation during exercise. In plasma, fatty acids compete with tryptophan for binding to albumin, hence an increase of bound fatty acids leads to an increase in plasma levels of free tryptophan. Tryptophan and the branched-chain amino acids pass the blood–brain barrier in competition for the same amino acid carrier, an increase in free plasma tryptophan and a decrease in branched-chain amino acids favouring the entry of free tryptophan into the brain. Tryptophan is converted to the neurotransmitter 5-hydroxytryptamine (5-HT, serotonin) in the brain, which plays a role in the induction of sleep and fatigue and has an inhibitory effect on the hypothalamus. This hypothesis might therefore explain some of the effects that are seen in the overtraining syndrome, but human overtraining studies, including interventions with glycogen and with branched-chain amino acids, have so far been inconclusive.

Glutamine and the immune system

The relationship between exercise and the immune system remains controversial. The leucocytosis of exercise is well recognized, and there is also evidence of impaired neutrophil function following intensive exercise, which may explain why upper respiratory tract infections are probably more common following a marathon. They also seem to be more frequent in the overtraining syndrome, although it is uncertain whether this is cause or effect.

Glutamine metabolism provides one possible explanation for the relationship between immune function and exercise. Glutamine is a free amino acid that is utilized at high rates by rapidly dividing cells such as leucocytes, and therefore considered important for normal immune function. Plasma levels fall with intensive exercise, possibly due to the increased glutamine requirement for gluconeogenesis and increased uptake by the liver, gut, and kidney. There is evidence that glutamine supplements may reduce the risk of infection in endurance athletes and also that plasma glutamine levels can act as a marker of the overtraining syndrome, but more research is needed in this area.

Clinical features

The overtraining syndrome is characterized by performance deterioration refractory to normal regeneration strategies. Variable combinations of symptoms are seen in association with this (Table 3). The athlete will commonly complain of fatigue and heaviness in the muscles, which they tend to ignore in the early stages. Sleep disturbance is common with difficulty getting off to sleep, early waking, and feeling unrefreshed on waking. They may complain of depression, anxiety, and irritability,

with loss of appetite and of libido. There may also be a history of upper respiratory tract infections.

The resting heart rate may be elevated, but this gradually returns to normal (usually very low in an élite endurance athlete) as recovery occurs. Physiological testing may show an increased heart rate response to exercise, a reduced maximum oxygen consumption ($\dot{V}O_2$**max**), and an impaired heart rate recovery following exercise. Maximum power output may also be reduced.

Blood tests show non-specific changes consistent with heavy training such as dilutional anaemia and raised muscle enzymes (in some cases the creatine phosphokinase level may be increased several thousand times). There may be evidence of a recent viral illness such as a positive Paul Bunnell test. It is also important to recognize that underperformance may be secondary to an underlying medical disorder such as anaemia, diabetes, or thyroid disease: these secondary causes are rare, but should always be considered.

Management

Prevention

As our understanding of this syndrome improves it may be possible to prevent it from occurring. This is not possible at present, because although it is related to prolonged intensive exercise, the ability to withstand the stress of training varies significantly between different athletes. Furthermore, there is currently no screening test that can reliably predict the onset of the syndrome. Resting heart rate can provide a guide: this is consistently low (within a couple of beats) in the healthy endurance athlete, but with overreaching and overtraining this may increase significantly (by 5–10 beats/min) and only returns to the previous low level when recovery occurs. Psychological assessment can also provide a guide, with the **POMS** (Profile of Mood State) questionnaire proving to be useful in this respect.

It is important to incorporate rest days into any training regime. Racehorses can develop a syndrome similar in many respects to the human overtraining syndrome, and in one study where they were given an alternate-day regime of hard training day/easy training day the frequency of the equine equivalent of the overtraining syndrome decreased and performance improved. There have now been several human studies where rest days have been incorporated into the training regime (for example, four heavy training days, two light training days, and one rest day per week): these seem to be associated with a reduction in the frequency of the syndrome.

It is also important to ensure that dietary intake is sufficient at times of intensive training, especially in terms of carbohydrate and hydration. It is surprising how many athletes are unaware of the nutritional components of diet and have an intake that is inadequate for the intensity of their training.

Treatment of the established syndrome

As many of the symptoms of this syndrome are non-specific, it is important to exclude other causes of underperformance and fatigue, including infection, thyroid disease, diabetes, and anaemia. In most cases tests for these disorders will be negative.

The treatment of the established syndrome requires rest to allow regeneration and recovery. Most athletes will only accept absolute rest for a few days, hence it is important to follow this up with a period of relative rest, that is allowing very low intensity training. The exercise can then be slowly progressed, but it may take up to 12 weeks to achieve full recovery. Many athletes are pleasantly surprised about how well they have maintained their performance despite 12 weeks of light training.

In addition to relative rest, it is important to adopt a holistic approach to treatment. This includes assessing other coexistent stresses and making use of relaxation techniques. Nutrition is important and further advice on this should be offered as appropriate. Athletes should monitor their progress by measuring their resting heart rate, which will decrease as recovery occurs.

Medical complications in sport

Although injuries are the dominant feature of sports medicine, there are several well-recognized medical disorders that occur as a result of sport and physical activity. Some are associated with acute bouts of exercise and others with more prolonged periods of training. The female athlete triad is discussed above, and sudden cardiac death and exercise-induced asthma are dealt with elsewhere.

Delayed-onset muscle soreness, rhabdomyolysis, and heat stroke

Strenuous exercise can produce transient damage to the muscle. Delayed-onset muscle soreness can occur several hours after a bout of unaccustomed, intensive eccentric exercise, reaching a peak between 1 and 3 days. It is associated with objective muscle weakness, which can last for up to 10 days. There are increased serum levels of creatine kinase and myoglobin, with muscle oedema and structural change revealed on magnetic resonance imaging (**MRI**) (*T*2-weighted sequences) and muscle biopsy, respectively. These findings are most obvious after 2 to 3 days. The structural damage fully repairs and the symptoms and laboratory abnormalities resolve by 10 days. This phenomenon is commonly seen following marathon running, when creatine kinase levels may rise to over 2000 IU/l (Fig. 3).

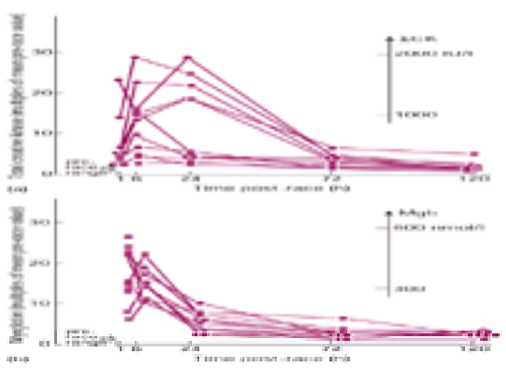

Fig. 3 Changes in (a) serum total creatine kinase (CK) and (b) plasma myoglobin in young men who had completed a marathon (running times: mean, 194 min; range, 163–280 min). (Taken from Young, *et al*. (1984). *European Journal of Clinical Investigation* **14**, 2, 58, with permission.)

There is less muscle damage when a bout of similar intensive eccentric exercise is repeated. This is known as the 'repeated bout effect' and is probably due to a series of adaptations taking place at neural, connective tissue, and cellular levels. These changes make the muscle more resistant to damage with further bouts of intensive exercise.

It is rare for large amounts of myoglobin to be released from muscle during exercise ('exertional rhabdomyolysis'), but when it does occur it can be life-threatening by causing acute renal failure, often with severe hyperkalaemia (see Chapter 20.4). This is associated with intensive eccentric exercise, dehydration, and hyperthermia and forms part of the syndrome of heat stroke.

There is high heat production with exercise: 75 per cent of the energy expended is converted to heat during running. Heat dissipation therefore becomes extremely important, with evaporation through sweating being the most important mechanism, and anything that impairs sweat evaporation will put the athlete at risk of hyperthermia and heat stroke. This can occur in extreme environmental conditions such as high temperature and high humidity, and is more likely when the athlete is dehydrated. A poorly prepared athlete lacking fitness, who is inadequately acclimatized to the heat, has had a recent illness (for example, a cold or gastroenteritis), or wears excessive clothing is more at risk. Nowadays heat stroke is more commonly seen in fun runs than in marathons, when there are often a large number of poorly prepared participants. Furthermore, it can occur when the environmental conditions are not particularly extreme, suggesting that heat production during exercise is the most important factor.

The pathophysiology, clinical features, and management of heat stroke is dealt with elsewhere (Chapter 8.5.1). As patients with heat stroke require immediate treatment, on-site resuscitation facilities should be available at competitions and fun runs. Management includes the use of intravenous therapy and increasing heat loss through evaporation (the patient should be put in the shade and then wetted with lukewarm water and fanned). It is possible to minimize the risk of developing exertional heat stroke by paying attention to certain recommendations (Table 4).

Exercise-induced gastrointestinal symptoms

Up to 50 per cent of long-distance runners will complain of gastrointestinal (**GI**) symptoms. These include reflux/heartburn, intestinal cramps, the urge to defecate, and diarrhoea, which may be bloody. The lower GI symptoms are probably due to reduced blood flow (splanchnic blood flow decreases by up to 80 per cent with intensive endurance activity) and possibly mechanical (jarring) stress on the gut (as symptoms are more common in runners than cyclists). The relative gut ischaemia may lead to the release of several GI hormones, including secretin, glucagon, and vasoactive intestinal polypeptide: the latter can remain elevated for up to 2 h after the termination of exercise. These hormones will increase secretion into the gut while also reducing absorption, effects that are enhanced by dehydration.

Exercise-induced GI symptoms can usually be controlled with appropriate advice and tend to decrease with adequate training. Adaptive changes occur with gradual increases in training volume and intensity, such that there may be an improvement in blood flow through the splanchnic circulation. It is also important for athletes to be adequately hydrated both before and during exercise: they should therefore take account of the ambient temperature and humidity and adapt their fluid intake accordingly. Hypertonic drinks should be avoided as they will tend to increase the risk of dehydration.

Exercise-related anaemia

Haemoglobin concentrations in highly trained endurance athletes are often at the lower end of the normal range or even just below it. In most cases this reflects a dilutional state where, although red cell mass increases with exercise, there is a proportionally greater increase in plasma volume. A true runners' anaemia may be caused by faecal blood loss (see above), also from intravascular haemolysis caused by high foot-impact forces ('march haemoglobinuria'). A similar traumatic haemoglobinuria also occurs in conga-drum players due to high impact on the hands.

Fitness to exercise

Sudden death in sport

Although it often attracts headlines in the press, sudden death in sport is very rare. The frequency of sudden death varies from about 0.5 per 100 000 to 6 per 100 000 per year, depending on the age group being assessed, the level of underlying fitness, and the intensity of activity. Fatal arrhythmia seems to be the most common mechanism of death (Table 5). Atherosclerotic coronary artery disease is the most common cause in those over 40 years of age: in younger athletes underlying causes include cardiomyopathy, valvular heart disease, and Marfan's syndrome.

Myocarditis (see Chapter 15.8.1) is another possible cause of sudden death in sport. This is usually viral in origin, in particular cocksackie B virus, but a series of sudden death cases amongst Swedish orienteers was found to be associated with *Chlamydia pneumoniae* myocarditis. Cardiac concussion is a rare cause of sudden death, thought to be due to a dysrhythmia resulting from a non-penetrating precordial blow from a projectile, such as a cricket ball or an ice hockey puck.

Screening

Although it is accepted that people with certain cardiac risk factors, such as those with hypertrophic cardiomyopathy or aortic stenosis, should avoid competitive sports, there is limited value in cardiac screening of athletes. This is because of the rarity of these abnormalities, the rarity of sudden death, and the cost of screening. It is estimated that 200 000 athletes would have to be screened to find one at-risk case. There is also limited predictive accuracy of some of the cardiac investigations. This is the case with electrocardiography (**ECG**), where it may be difficult to distinguish the physiological changes of the athletes' heart with the pathological changes seen with hypertrophic cardiomyopathy. Further cardiac investigation should probably be restricted to those with a history of cardiac and/or exercise-related symptoms, a relevant family history, or abnormalities found on cardiac auscultation.

Prevention of sudden death in sport

Athletes with confirmed myocarditis should be withdrawn from competitive sports for at least 6 months, while those with more general viral illness should abstain from sport until they have recovered.

Sudden death is most likely to occur in high-intensity competitive sports (for instance, squash), and in this situation the athlete should be offered a basic medical assessment (personal and family history and physical examination) prior to competition. The need for a medical assessment is less important in those participating in recreational sports. Although acute bouts of exercise increase the risk of cardiac death, this transient increase in risk is outweighed by the cardiac benefits of habitual exercise. The importance of graded increases in exercise intensity, allowing cardiac and musculoskeletal adaptations to occur, should be stressed.

Overuse injuries

Overuse injuries are the most common type of injury seen in a sports medicine clinic. It is usually relatively straightforward to make the diagnosis from the history and examination. However, it is equally important to determine the aetiological factors responsible for the injury to prevent a recurrence. These can be divided into training methods, equipment, and biomechanical factors (see Table 6).

Injuries can occur when the training is increased too quickly, when there is an inadequate warm-up or cool-down period, and when there is inadequate flexibility training to complement the overall programme. Injuries can also occur when the athlete suddenly changes from one surface or gradient to another. Equipment factors are also important, which is especially the case for footwear in weight-bearing sport, where it is vital that the shoes provide adequate support for the sport being undertaken and are not overly worn out. In racket sports the size, weight, and string tension of the racket are important factors that may influence the risk of injury.

It is important to consider biomechanical factors in athletes as the repetitive nature of their activities (for example, running action, tennis serve, or cricket bowling) may magnify any minor malalignment. It is therefore necessary to assess for various factors, including leg-length difference, wide pelvis, pelvic tilt, femoral neck anteversion, and tibia varum. The shape of the foot on standing should also be considered, as both overpronated and rigid, high-arched feet can cause problems. Variations in anatomy of the bones may also increase the risk of injury, such as the hooked acromium and rotator cuff impingement in throwers and the os trigonum and posterior ankle impingement in dancers and footballers.

The type of injury will depend on the age of the athlete as this determines the weakest point in the musculoskeletal chain. In the growing adolescent the point of attachment of the tendon to bone is vulnerable, and injuries such as Osgood–Schlatter's disease at the tibial tubercle and Severs disease at the calcaneus are commonly seen. In the young adult tendinitis and injuries at the musculotendinous junction are particularly common, while in the elderly athlete degenerative injuries of the tendon and joint are seen.

Stress (or fatigue) fractures occur as a result of bone overload. They occur when training is increased too quickly, as in rapid preparation for a marathon, or when there are biomechanical factors that increase the stress load on bone, for example rigid, high-arched feet. They also occur more commonly in amenorrhoeic athletes than in their eumenorrhoeic counterparts. Clinically the athlete usually has point tenderness. A bone scan or MRI will allow an accurate diagnosis to be made, although the plain radiograph may be negative.

Drugs and ergogenic aids in sport

Introduction

The use of drugs and nutritional supplements in competitive sports is widespread. The main reason for this is to enhance performance, but supplements are also

taken to improve general health and to increase resistance to infection. In theory, this would reduce the risk of developing coincidental medical problems and hence minimize interruptions to training. The scientific evidence for many of these substances is flimsy, but athletes are often prepared to try them on the basis of the anecdotal experience of fellow athletes, advice from their coach, or even from suppliers at the local gym. Furthermore, surveys suggest that athletes are willing to take substances for short-term performance enhancement, even if it puts their long-term health at risk. The banning of some substances has had an effect on some athletes, but others continue to take them and are prepared go to extremes to avoid detection.

With most performance-enhancing drugs there is a large interindividual response, but the cause for this is not fully understood. Studies on most drugs and nutritional supplements have therefore been unable to demonstrate consistent efficacy. There is a large potential placebo effect, which may obscure any pharmacological action. Large randomized controlled trials would be helpful, but these are virtually impossible to do because most athletes would be unwilling to accept a placebo. Furthermore, if there is any possibility of the drug enhancing performance it is likely already to have been banned by the sports governing body. Studies on the non-athletic population, who exercise at much lower levels, may not be representative of the effects seen in athletes. Moreover, athletes often take a variety of substances in extremely large doses, which makes it difficult to compare the results with the lower doses given in the general population in terms of efficacy and (especially) safety.

Anabolic agents

Anabolic–androgenic steroids

Testosterone, first synthesized in the 1930s, has both anabolic and androgenic effects. Since then several synthetic derivatives have been produced in an attempt to provide a steroid with predominant anabolic actions. The first recorded use of anabolic–androgenic steroids (**AAS**) was during the Second World War when they were given to German troops to increase their aggressiveness. In 1952 the Russian weightlifting team was suspected of taking them when they won three gold medals at the Helsinki Olympic Games. This increased the interest in these drugs, and by 1958 a United States pharmaceutical company had developed AAS. The dangers gradually became apparent, but it was not until 1975 that the International Olympic Committee banned their use.

AAS enhance muscle strength and power but have no effect on aerobic performance. They are also thought to have psychological effects and can promote aggressiveness, which in its most extreme form may lead to 'roid rage'. Athletes may also experience euphoria and reduced fatigue while taking AAS, a combination of effects that may allow them to train harder and for longer.

Although AAS will increase muscle strength in non-exercising people, the greatest increases in strength occur when AAS are taken in conjunction with resistance exercise (Fig. 4). The reason for this may be that physiological stress occurs with intense resistance exercise and leads to increased levels of glucocorticoids, thereby producing a catabolic state with a negative nitrogen balance that anabolic steroids can reverse.

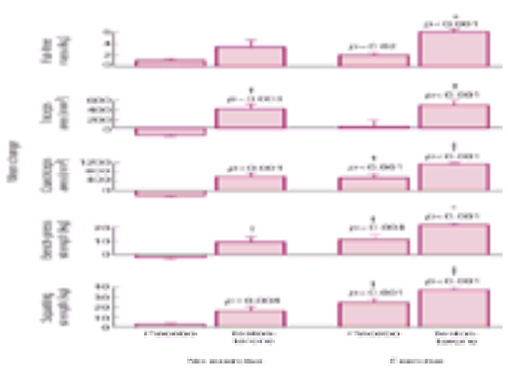

Fig. 4 A randomized trial of exercise and supraphysiological doses of testosterone in men over a 10-week period. Changes from baseline in mean (± SE) fat-free mass, triceps, and quadriceps cross-sectional areas, and muscle strength in the bench-press and squatting exercises. Note the greatest effect is obtained with a combination of exercise and testosterone.

Human growth hormone

The physiological effects of human growth hormone (**HGH**) include nitrogen retention, increased protein synthesis and tissue growth, and an increase in fat-free mass. Recombinant HGH was first produced in 1984 and (although expensive) its use by athletes has since increased.

The anabolic effects of HGH make it potentially valuable as an ergogenic aid for athletes participating in power sports. Furthermore, as it is not detected by most drug-test screenings performed by sports governing bodies, it is an attractive option when compared to anabolic steroids. However, so far there has been only limited research to determine its effectiveness in increasing muscle strength and power.

b_2-Agonists

This group of drugs was initially banned by the sporting bodies because of their stimulant action. However, in the last few years they have gained interest as muscle strengthening agents when taken orally. During the 1992 Barcelona Olympics, two American power athletes were found to have taken clenbuterol.

Animal studies show that clenbuterol can produce muscle hypertrophy. Studies of b_2-agonists in healthy men, in patients following lower limb surgery, and those with spinal cord injury, all showed gains in muscle strength following an exercise programme when compared to a placebo group. However, there is very little data on their effect in highly trained athletes. Although they act as bronchodilators, there is no evidence that they can improve aerobic fitness in non-asthmatics.

Creatine

Creatine was first recognized as an ergogenic aid in the early 1990s and was used by athletes in the 1992 Barcelona Olympics. It is an amino acid present in meat, normal daily intake being less than 1 g. The highest concentration of creatine is in skeletal muscle, in particular in type-II muscle fibres, where most is in the form of creatine phosphate, providing a source of energy and assisting in the restoration of ATP following exercise. Creatine phosphate may also assist in buffering when the pH falls due to lactic acid accumulation during exercise.

Supplements are usually taken in high dosage for a limited period (for example, 25 g for 6 days), following which muscle concentrations of free creatine and creatine phosphate remain elevated for several weeks. It is generally pointless to take high-dose supplementation beyond this period as the majority of muscle creatine uptake occurs during the first few days, but some athletes take a maintenance dose of 2 g/day.

Following this regime many athletes report that they are able to increase their training loads, and studies suggest that supplementation does improve high-intensity performance, especially when repeated exercise bouts are carried out (for example, repetitive sprinting and cycling). The ergogenic effects also extend to strength and power events such as weightlifting. No effect is seen in predominantly aerobic events.

Overall, there is only limited information on the effects of creatine supplementation, but these are consistent with the role of creatine phosphate in enhancing the restoration of ATP. The increase in muscle strength possibly occurs by allowing greater training intensity, leading to an enhanced training response. Within a few days of taking creatine there is a weight gain of between 2 and 5 kg: a large proportion of which can be accounted for by an increase in intracellular fluid volume (it is well known that urine volume decreases during the period of supplementation), but there may also be an increase in fat-free mass.

There are theoretical concerns regarding the side-effects of using large doses of creatine for prolonged periods. These include adverse effects on the kidney,

although, so far, there have been no confirmed reports of this. There are also anecdotal reports of increases in muscle cramps and concerns regarding the impact of weight gain, especially in athletes competing in weight-category sports (for example, boxing and wrestling). In these sports the athlete often has to lose weight rapidly to 'make the weight' for competition, and if they have gained extra weight with the use of creatine they may have to severely dehydrate to achieve such weight reduction.

As supplementation with creatine seems capable of enhancing performance there are ethical issues regarding its use. Some feel it should be banned, others argue against this on the basis that it is a component of a normal meat diet and is therefore no different to taking carbohydrate supplements.

Stimulants

Amphetamines, ephedrine, and cocaine

These stimulate the central and sympathetic nervous systems. As sympathomimetic agents they increase heart rate, blood pressure, metabolic rate, and plasma levels of free fatty acids. These actions could theoretically enhance aerobic performance, but not without risk. There were at least two amphetamine-related deaths in cyclists in the early years of their use (one in the Rome Olympic Games in 1960 and the other in the 1968 Tour de France) and there have also been several deaths in athletes associated with the use of both ephedrine and cocaine.

Research on the use of amphetamines suggests that they have little direct physiological effect during exercise but that they can mask pain and fatigue during activity. This may allow athletes to exercise closer to their limit, which may produce a positive effect, especially on endurance performance. However, this could also have a detrimental effect on health by inhibiting the athlete's awareness of early warning signs (for example, injury or dehydration). The same may apply to ephedrine and cocaine. Ephedrine is also used by athletes to reduce fat and increase fat-free mass, but there is little evidence for this effect. Cocaine can produce euphoria and also lead to addiction, which has caused great concern regarding its use amongst athletes.

Caffeine

Caffeine, which is chemically related to the theophyllines, is metabolized in the liver to produce dimethylxanthine. When used prior to prolonged exercise (in a dose of 3–6 mg/kg body weight, 1 h before exercise) it can delay fatigue and enhance endurance performance, although there is significant individual variation. Although caffeine is known to have effects on the central nervous system, adipose tissue, and skeletal muscle, the mechanism by which it reduces fatiguability is unclear. It has several unwanted effects, including insomnia, headache, and gastrointestinal irritation, but its diuretic effect is of particular concern, especially in athletes competing in hot climates.

The International Olympic Committee considers caffeine to be a performance-enhancing drug. However, it would be impossible to have an outright ban on a substance that is present in so many foods and drinks, hence an athlete found to have a urine concentration of more than 12 mg/l is deemed to be guilty of a doping offence. This has its limitations as some athletes can obtain a performance-enhancing effect at urine concentrations well below this level.

Other agents

b-Blockers

By reducing heart rate this group of drugs reduces aerobic capacity and therefore decreases endurance. However, they are effective in skill sports, with studies confirming that shooters improve their performance when taking b-blockers. These drugs probably exert their effect by reducing hand tremor and heart rate: top-class shooters tend to shoot between one heartbeat and the next, and therefore a reduction in heart rate is helpful. They have also been shown to be effective in treating stage fright in musicians.

Diuretics

Athletes use diuretics when they need to lose weight rapidly. This occurs in sports such as boxing, judo, and light-weight rowing when the individual needs to make a particular weight classification for competition. Diuretics are very effective in producing short-term weight loss, in the order of 4 per cent over 24 h. However, this can lead to dehydration and electrolyte disturbance, both of which can affect performance. Frequently the athlete will have up to 20 h between the weigh-in and competition: this gives sufficient time to replace the fluid loss, but is insufficient to re-establish normal electrolyte balance, which can lead to medical complications during competition, including renal failure, severe hyperkalaemia, and rhabdomyolysis.

Erythropoietin and blood doping

A modest increase in red cell mass of up to 5 per cent occurs with adaptation to endurance training. This can take several months. However, some athletes artificially increase their red cell mass either by infusing previously stored red cells or by the use of erythropoietin. Infusing red cells has probably been used since the early 1970s in sports such as distance running, cycling, and cross-country skiing. In 1984 the United States men's cycling team confessed to using it during the Olympics, and won gold medals. Erythropoietin has probably been used by athletes since the late 1980s and may well have been responsible for a number of deaths seen in cyclists in the last 12 years.

Although homologous transfusions are used by athletes, autologous transfusion is probably more common. During this process several units of blood are removed from the athlete and then stored for several weeks, while the blood count is naturally restored to normal. The red cells are then reinfused, thereby increasing the red cell count, which will be sustained for a few weeks. The alternative method of increasing red cell mass is to administer erythropoietin, when the red cell count rises gradually over several weeks and remains elevated as long as the treatment continues.

Physiological studies confirm improved exercise performance with the use of these techniques to increase the blood count. Maximal aerobic power, submaximal endurance, and race performance have all been shown to improve. Blood doping may also provide a thermoregulatory advantage for those exercising in the heat, and some benefit for those exercising at altitude.

Medical risks include those associated with transfusions. Even autologous infusions can cause problems through clerical errors and mishandling of the stored blood product. Risks also occur in association with the high haematocrit, with blood viscosity rising exponentially as the haematocrit increases above 30 per cent thus leading to an increased risk of thromboembolic events. There have now been numerous deaths related to the use of blood doping by athletes.

Although blood doping is banned by all the main sports governing bodies there is no reliable test for detecting either autologous red cell infusion or erythropoietin administration. It is therefore difficult to prove that an athlete has used this technique and to provide a consistent deterrent.

Bicarbonate

A metabolic acidosis occurs with anaerobic exercise and is thought to be responsible for the progressive fatigue that occurs. By inducing a metabolic alkalosis prior to exercise, it may be possible to delay the onset of fatigue and improve exercise performance. Several studies have been undertaken to assess the effect of pre-exercise bicarbonate ingestion on performance. There are conflicting results from these studies, some showing a benefit, others not. The reason for these differences is probably due to variations in the duration and intensity of exercise, the dosage of bicarbonate, and the length of time between taking bicarbonate and the onset of exercise. Bicarbonate is most likely to have an effect with exercise of only a few minutes duration, when given 2 to 3 h before exercise, and at a dose of about 0.3 g of sodium bicarbonate per kg body weight. Side-effects of this ingestion include vomiting and diarrhoea, which may limit the potential benefits. Bicarbonate is not banned by the sports governing bodies.

Further reading

Bhasin S, *et al*. (1996). The effects of supraphysiologic doses of testosterone on muscle size and strength in normal men. *New England Journal of Medicine* **335**, 1–7. [Comprehensive review]

Brouns F, Beckers E (1993). Is the gut an athletic organ? Digestion, absorption and exercise. *Sports Medicine* 15(4): 242–257. [Comprehensive review]

Budgett R (1998). Fatigue and underperformance in athletes: the overtraining syndrome. *British Journal of Sports Medicine* **32**, 107–10. [Comprehensive clinical review]

Budgett R, *et al*. (2000). Redefining the overtraining syndrome as the unexplained underperformance syndrome. *British Journal of Sports Medicine* **34**, 67–8. [Summary of current thinking]

Clarkson PM, Thompson HS (1997). Drugs and sport. Research findings and limitations. *Sports Medicine* **24**, 366–84. [Comprehensive review]

Foster C, Lehmann M (1997). Training/overtraining: the first Ulm symposium. *Medicine and Science in Sports and Exercise* **30**, 1137–78. [Comprehensive scientific review]

Futterman LG, Myerburg R (1998). Sudden death in athletes. *Sports Medicine* **26**, 335–50. [Comprehensive review]

Huston TP, *et al*. (1985). The athlete heart syndrome. *New England Journal of Medicine* **313**, 24–30. [Comprehensive review, especially of ECG changes in athletes]

Jenkins PJ, Grossman A (1993). The control of the gonadotrophin-releasing hormone pulse generator in relation to opioid and nutritional cues. *Human Reproduction* **8**, 154–61. [Comprehensive review]

Loucks AB, *et al*. (1992). The reproductive system and exercise in women. *Medicine and Science in Sports and Exercise* **24**, S288–S293. [Comprehensive review]

Maron BJ (1986). Structural features of the athletes' heart as defined by echocardiography. *Journal of the American College of Cardiology* **7**, 190–203. [Comprehensive review of echo changes in athletes]

Maughan RJ (1999). Nutritional ergogenic aids and exercise performance. *Nutrition Research Reviews* **12**, 255–80. [Comprehensive review]

McHugh MP, *et al*. (1999). Exercise-induced muscle damage and potential mechanisms for the repeated bout effect. *Sports Medicine* **27**, 157–70. [Comprehensive review]

Mittleman MA, *et al*. (1993). Triggering myocardial infarction by heavy physical exertion. *New England Journal of Medicine* **329**, 1677–83). [Important trial demonstrating the risk of myocardial infarction with unaccustomed physical activity]

Neely FG (1998). Biomechanical risk factors for exercise-related lower limb injuries. *Sports Medicine* **26**, 395–413. [Comprehensive review]

Otis CL, *et al*. (1997). American College of Sports Medicine position stand. The female athlete triad. *Medicine and Science in Sport and Exercise* **29**, i–ix. [Comprehensive review]

Renstrom PAFH (1999). An introduction to chronic overuse injuries. In: Harries MG, *et al*., eds. *Oxford textbook of sports medicine*, pp 633–48. Oxford University Press. [Comprehensive review]

Robinson TL, *et al*. (1995). Gymnasts exhibit higher bone mass than runners despite similar prevalence of amenorrhoea and oligomenorrhoea. *Journal of Bone and Mineral Research* **10**, 26–35. [Important trial showing difference between gymnasts and runners]

Sawka MN, *et al*. (1996). American College of Sports Medicine position stand. The use of blood doping as an ergogenic aid. *Medicine and Science in Sport and Exercise* **28**, i–viii. [Comprehensive review]

Walsh NP, *et al*. (1998). Glutamine, exercise and immune function. *Sports Medicine* **26**, 177–91. [Comprehensive review of the role of glutamine]

Wolman RL, *et al*. (1990). Menstrual status and exercise are important determinants of spinal trabecular bone density in female athletes. *British Medical Journal* **301**, 516–18. [Important trial showing difference between rowers and runners]

Zanker CL, Swaine IL (1998). The relationship between serum oestradiol concentration and energy balance in young women distance runners. *International Journal of Sports Medicine* **19**, 104–8. [Important trial demonstrating the effects of energy imbalance in athletes]

29 Adolescent medicine

R. Viner

Introduction

Young people between 10 and 20 years of age comprise 12 to 15 per cent of the population in most developed countries, and are increasingly recognized as a distinct patient group requiring a special approach. Few diseases are unique to adolescence, but adolescents have a distinct epidemiology of disease and present a constellation of symptoms and problems not found in children or adults. Special communication skills are needed to take an accurate history and elicit clinical signs in young people. The effective treatment of illness in adolescence requires adept management of compliance, consent, and confidentiality, and relationships between the young person and their family.

The most serious health problems affecting young people are primary care issues, including teenage pregnancy, drug misuse, mental health problems, and violence. However, young people with acute or chronic medical conditions present management challenges to physicians—challenges that are mounting with the increasing incidence of chronic illness in adolescence.

What is adolescence?

Strictly speaking, adolescence is the period between childhood and adulthood. Theoretical definitions abound; Freud saw adolescence as the period of recapitulation of the childhood Oedipal complex, while Erickson claimed that the struggle between Identity and Role Confusion typified adolescence. The World Health Organization defines adolescence as between 10 and 20 years of age. However, chronological definitions take little account of the timing of the developmental changes at the heart of the concept of adolescence.

Some have suggested that adolescence is merely a socially constructed rite of passage. These claims ignore the biological changes of puberty and psychological developments driven by increasing maturation of the central nervous system. The most useful definition is that adolescence is a period of biopsychosocial maturation leading to functionally independent adult life.

Adolescent development

All clinical interactions with adolescents must be seen against a developmental background. The timing and tempo of the events of biological, psychological, and social development each proceed independently, although they are deeply intertwined (Table 1).

Biological changes

The biological changes of adolescence are puberty, the pubertal growth spurt, and accompanying maturational changes in other organ systems. These include maturation of enzyme systems (such as cytochrome P-450), accretion of peak bone mass, and the development of sexually dimorphic adult patterns in blood lipids, haemoglobin, and red cell indices. It is important for physicians to be able to identify the pubertal stage in adolescents, particularly in chronic illness. The defining event of puberty in girls is the menarche. The biological changes of adolescence are universal to all races, as are psychological changes driven by increasing brain myelination. However, most psychological and social development is culture-specific, varying with the social and cultural norms of childhood and adult roles in society.

Psychological development

Before adolescence, children think concretely, understanding only the immediate and short-term consequences of actions or events. Ideas and concepts can be manipulated only by using concrete representations. From the age of 12 years onwards, thought patterns begin to become formal operational or abstract, with the ability to manipulate ideas rather than things, imagine the future, and conceive multiple outcomes of actions. These capacities are important for the development of a settled personal and sexual identity.

Social development

The essential social tasks of adolescence are developing a sense of personal identity, moving from dependence to relative independence, and developing mature relationships with peers. Rather than achieving 'independence' from the family; adolescence involves renegotiating family relationships from a position of increasing adult equality. These challenges will exist across all cultures; however, the point at which successful completion is expected varies greatly.

Why is medicine different when dealing with adolescents?

Clinical medicine with adolescents is different because adolescents have special patterns of disease and special needs in the management of illness. Few diseases, aside from disorders of puberty, are specific to adolescence. However, the causes of mortality and morbidity in adolescents are distinct from both children and adults. Environmental and social causes of mortality (for example, accidents and suicide) account for a larger proportion of total adolescent mortality than at any other age (Fig. 1). Young people are one of the only groups in which overall mortality rates have fallen little during the past 40 years.

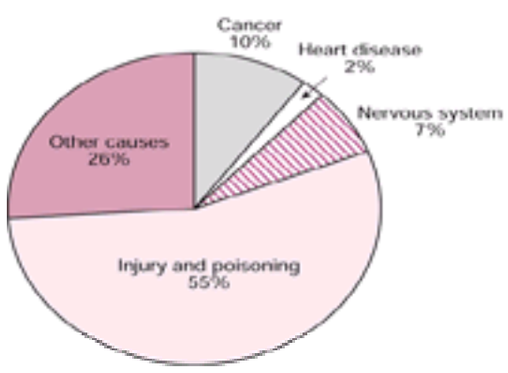

Fig. 1 Causes of death for 15- to 19-year-olds in the United Kingdom, 1997 (*Mortality statistics*, The Stationery Office. DH2, No. 23).

Common paediatric or adult diseases present with different patterns during adolescence. Type I diabetes has its peak age of incidence around 12 to 14 years. Puberty accelerates the progression of diabetic complications and the metabolic control of diabetes is poorer during adolescence than at any other age, due both to the growth-hormone excess of puberty and the psychosocial challenges of chronic illness management. Cancer during adolescence is remarkable for its combination of 'late' presentations of paediatric cancers, 'age-specific' cancers of adolescence (for example, bone tumours), and early-onset 'adult-type' carcinomas.

Medical challenges of adolescent development

The real challenges of medicine in adolescents come from the reciprocal impacts of adolescent development on disease management and quality of life. This is especially true in chronic conditions (Table 2) that may retard adolescent development, producing pubertal and growth delay, delayed social independence, poor body and sexual self-image, and educational and vocational failure. Physicians and paediatricians must monitor growth and pubertal development in chronic illness, until well into the early twenties.

Being chronically ill, having a visible disability, or being required to adhere to difficult treatment regimens is specially difficult during adolescence. Alienation from the peer group and absence from school cause social isolation, failure of socialization, and educational underachievement. The importance of helping young people with chronic illness or disability to develop independent adult living and vocational skills has been shown in longitudinal follow-up studies.

Conversely, developmental issues affect the management of illness. Poor adherence (compliance) and poor disease management are almost 'the rule' in adolescence. Immature abilities to imagine future consequences make the prevention of long-term complications of illness a poor motivator for compliance. Medical advice may be rejected as part of growing independence from parents, particularly in chronic paediatric illnesses where medical staff have become medical 'parents'. Adherence and disease control are also put at risk by the developmental need to explore possible modes of future behaviour, no matter how dangerous (derogatively referred to as 'adolescent risk-taking'). Risky behaviour such as smoking, alcohol, and recreational drug use are as common in adolescents with chronic illness or disability as in the general population.

The management of ill-health in adolescence

Most doctors have adolescent patients, but few are experienced or skilled in dealing with such patients. American studies suggest that only around one-third of physicians and paediatricians enjoy working with adolescents, and that another third have very little interest in caring for them. The effective clinical management of any disease in an adolescent requires a non-judgemental communication style, a knowledge of adolescent development, an awareness of consent and confidentiality issues, and an ethnographic approach that aims at understanding the health beliefs and situations in which the young person manages their disease.

Communication with young people

Neither the standard paediatric consultation (doctor communicates with parents) nor the standard adult consultation (doctor communications solely with patient) is appropriate for adolescents. It is best to see young people together with their parents, and also by themselves. While time consuming, this is essential for obtaining an accurate history, understanding their motivations and goals, and for getting accurate information about risky behaviour.

Frameworks have been developed for best practice with young people, the best known being **HEADSS** which reminds clinicians to cover the important domains of: home life; education; activities (friendships, social relationships, exercise); drugs; sexuality (intimate relationships, risky behaviour); and suicide (depression and self-harming). But having a framework is not enough; the key skills required for effective communication with young people are an understanding of adolescent development, empathy, respect and a non-judgemental attitude, understanding the link between physical and emotional well being, and provision of a physically and emotionally safe environment in which a clinical interaction can take place.

Table 3 summarizes these points.

Confidentiality and consent issues

Adolescents require from clinicians confidentiality, respect, and clinical excellence. Services that are not considered to be confidential are less likely to be used by young people. Confidentiality (including keeping confidentiality from parents) should be assured to young people, unless they are found to be at risk from suicide, sexual abuse, or they reveal plans to harm others.

Adolescents can fall into a no-man's land between parental rights over minors and adult rights. In most countries, including the United Kingdom, adolescents are now deemed to have adult rights to consent to treatment if they are legally competent, regardless of their parents' wishes. The legal criteria for competence differ between countries, but usually require the ability to give informed consent and understand the benefits and risks of treatment or non-treatment. In the United Kingdom, competence is presumed over the age of 18 years, while adolescents between 16 and 18 years can consent to treatment but cannot refuse life-saving treatment. Under 16 years of age, adolescents are presumed incompetent unless they demonstrate otherwise.

Adherence issues

There is little evidence that young people with chronic diseases are any less compliant to medical regimens than adults. Many struggle with the responsibility of organizing difficult regimens; others manipulate their regimen as part of their conflict with parents; and many adhere to a regimen of their own choosing—but one that may have little relationship to that prescribed. Practical measures to improve adherence to medical regimens are outlined in Table 4. The most important aspect is to 'decriminalize' non-adherence by recognizing that some non-adherence is universal, and to work with the adolescent to tailor the regimen to meet their health goals.

Transition

Adolescents with continuing health problems will require transfer from paediatric to adult services. Much more than just a clinic transfer, transition requires a change from the family-centred developmentally focused paediatric approach (which infantilizes the adolescent) to an adult medical culture that acknowledges patient autonomy and reproduction and employment issues but neglects growth, development, and family concerns.

Traditional methods of transfer of care by referral letter can lead to adolescents settling poorly into the new service or even dropping out of medical supervision altogether. This period is particularly dangerous for those diseases where adult services or skills are poorly developed, such as in 'paediatric' metabolic diseases or congenital heart disease. All paediatric specialist clinics should have transition policies and guidelines, especially where many adolescents are transferring. Preparation for transition should begin in early adolescence, and young people should only move to adult care when they have the necessary experience to survive independently in the adult service.

Drug, alcohol, sex, and health promotion

By 15 years of age, around 24 per cent of adolescents in Britain are regular smokers, 38 per cent will be regular alcohol consumers, and around 25 per cent are sexually active. Young people with chronic illness have similar rates of risk, although few physicians (especially paediatricians) address smoking, alcohol, or sex issues with young people with chronic conditions. Healthy behaviour begun in adolescence continues into adult life, and health promotion during early adolescence can discourage smoking, drug use, and unsafe sexual behaviour. Over 70 per cent of adolescents visit a doctor every year; each clinical interaction should provide an opportunity for health promotion.

Psychological issues of illness in adolescence

Epidemiological studies show that up to 20 per cent of adolescents may suffer mental health problems, most not being serious and frequently presenting with physical symptoms. Conversely, many young people with chronic medical conditions suffer adverse psychological sequelae, particularly depression and anxiety and adjustment disorders. During the developmental changes of adolescence, the psyche and soma are inextricably interrelated. Assessment and management of the reciprocal psychosocial impacts of adolescence and chronic illness (see Table 2) are a central part of medicine for adolescents. Severe or chronic illness in adolescence should be managed by multidisciplinary teams, including mental health, social, and youth workers and teachers as well as doctors and nurses.

Conclusions

Adolescent medicine demands that special attention be paid to communication, compliance, and risk behaviour, as the person develops. Greater skill in dealing with young people is required in all areas of the medical profession. These skills can be learned; evidence from randomized trials suggests that training in techniques for communicating with young people is effective and is valued by doctors and young people alike.

Further reading

Kramer T, Garralda ME (1998). Psychiatric disorders in adolescents in primary care. *British Journal of Psychiatry* **173**, 508–13.

British Medical Association (2001). *Consent, rights and choices in health care for children and young people.* BMJ, London.

Kyngas HA, Kroll T, Duffy ME (2000). Compliance in adolescents with chronic diseases: a review. *Journal of Adolescent Health* **26**, 379–88.

Goldenring JM, Cohen E (1988). Getting into adolescent heads. *Contemporary Pediatrics* **July**, 75–90.

Leffert N, Petersen AC (1995). Patterns of development during adolescence. In: Rutter M, Smith DJ, eds. *Psychosocial disorders in young people*, pp 67–103. Wiley, London.

MacKenzie RG (1990). Approach to the adolescent in the clinical setting. *Medical Clinics of North America* **74**, 1085–95.

Sanci LA, *et al.* (2000). Evaluation of the effectiveness of an educational intervention for general practitioners in adolescent health care: randomised controlled trial. *British Medical Journal* **320**, 224–30.

Viner RM (1999). Transition from paediatric to adult care. Bridging the gaps or passing the buck? *Archives of Disease in Childhood* **81**, 271–5.

Viner RM, *et al*. (2000). *Improving adherence to treatment in adolescents with chronic conditions: a practical evidence-based approach.* Society for Adolescent Medicine, San Diego, CA.

White PD (1999). Transition to adulthood. *Current Opinion in Rheumatology* **11**, 408–11.

Zirinsky L (1993). The psychological impact of illness in adolescence. In: Brook CDG, ed. *The practice of medicine in adolescence*, pp 25–34. Edward Arnold, London.

30.1 Medicine in old age

John Grimley Evans

Few, if any, diseases occur only in old age. The speciality of geriatric medicine is defined less in terms of the diseases it treats than in the range of responsibility it accepts. This responsibility embraces preventative care, health promotion, and diagnosis and treatment of acute illness followed by rehabilitation and resettlement of patients in the community. Some diseases are so much more common in later life that geriatricians will necessarily have more experience in managing them than will some other physicians. However, the great majority of what is to be found in a textbook of medicine will apply to older as well as to younger adults. Indeed it is an ethical duty of doctors to assume, in the absence of evidence to the contrary, that treatments that are effective for younger adults are at least as effective for those who are older.

Some aspects of medical practice need to take account of common age-associated changes in physiology. This chapter will briefly review some of these areas, but it begins with an outline of the background to medicine in later life provided by the universal processes of human ageing.

Ageing

Table 1 summarizes the sources of differences between young and old people. True ageing comprises those processes whereby differences arise because older people change from what they were when younger. However, not all differences between young and old people are due to ageing:

1. Selective survival leads to very old people showing genetic, sociobehavioural, and psychological differences from younger members of the same ethnic groups. Not surprisingly, differences include a lower prevalence of genes, social factors, and lifestyles associated with the risk of fatal diseases. Psychological variables with survival value include higher intelligence, better education, and a will for self-determination.
2. Cohort effects are prominent as causes of differences between young and old in changing societies. Apart from the effects of poverty and poor nutrition in early life, people born 70 years ago were raised and educated in a society very different from that of young people today. In longitudinal studies, where people are tested against their own former selves, declines in mental abilities appear less dramatic and later than in cross-sectional comparisons with younger individuals. Unfortunately most popular notions of the effects of ageing are based on uncritical acceptance of the findings of cross-sectional studies. Differences between young and old in psychological function partly reflect changes in educational emphasis, for example on computer skills rather than irregular Latin verbs. Also relevant, however, are changes in the cultural valuation of matters such as verbal abilities and good manners over the decades. Older people in England speak more slowly than the young not because they are slower witted but because 70 years ago it would have been considered ill-bred to talk as fast as is the custom nowadays. Another problem with cross-sectional studies is that they distort the pattern of ageing by blurring differences between individuals. Many individuals show preservation of mental function until the last year of life when abilities decline rapidly in what is sometimes termed the 'terminal drop'. With age the proportion of individuals who are in this phase will increase, so lowering the average performance of age groups. The true ageing pattern of a relatively constant level of performance followed by abrupt decline is therefore obscured by an appearance of continuous progressive decay.
3. Differential challenge. Since ageing is characterized by loss of adaptability it can only be accurately assessed by presenting individuals of different ages with similar challenges. In practice, society is organized so that older people may be faced with more severe challenges than are the young, and their poorer outcomes may then be attributed to ageing rather than inequity. This is particularly important in assessing the benefits potentially available to older people from medical interventions. There is abundant evidence from the United Kingdom and the United States that older people are on average provided with poorer quality health care than are younger adults. In the United States this is well documented for cancer treatment and in the United Kingdom for treatment of heart disease. In both countries the problem seems to arise from ageist prejudice among health workers at a local level in that its effects vary from district to district and hospital to hospital. Although less readily documented, there can be little doubt that ageism is as rife in primary as in secondary care. American studies have shown that primary care physicians spend less time with older patients than with younger ones even though the problems of the former are the more numerous, serious, and complex.

True ageing

Ageing, in the sense of senescence, is a progressive loss of adaptability of an individual organism as time passes. As we grow older the homoeostatic mechanisms on which survival depends become on average less sensitive, slower, less accurate, and less well sustained. Sooner or later we encounter a challenge from the internal or external environment to which we can no longer mount an effective response and we die. An increase in the risk of death with age is therefore the biological hallmark of senescence. In the human species death rates, which are high in the early years of life, fall to a nadir around the age of 12 to 13 at which point ageing first becomes manifest as rates turn upwards and continue through late adolescent perturbations due to violent and accidental deaths, to mount continuously and broadly exponentially throughout adult life (Fig. 1). The lowest point on the age-specific mortality curve which marks the onset of manifest senescence has been constant in England and Wales for over 100 years. It presumably therefore represents the point of maximum biological fitness. This is, in fact, to be expected since evolutionary pressure will lead to maximum fitness at the time of onset of reproductive capacity.

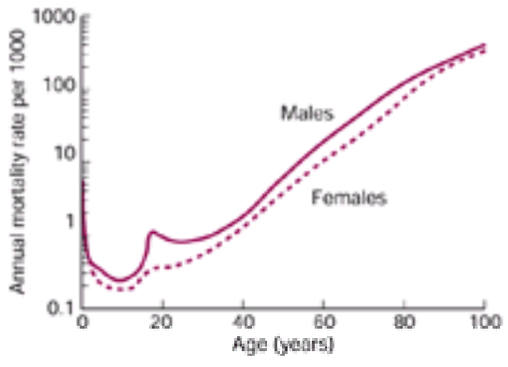

Fig. 1 Sex- and age-specific total mortality rates. England and Wales 1988.

As individuals, we age at different rates and with different patterns. Although average performance on measures of physical and psychological function may decline with age, interindividual variance increases and there will be many people in their eighties or beyond performing within a range normal for young adults. It is important to assess older people as individuals when making judgements about their function and capacity to benefit from medical interventions and not to assume that they are all average members of their age group.

Primary ageing

Primary ageing is the product of interactions between intrinsic, genetically determined, factors and extrinsic factors in lifestyle and environment. Some interactions are specific. A high dietary sodium intake will cause hypertension in genetically predisposed individuals, and excess dietary calories will lead to obesity and type 2 diabetes with its associated problems in people carrying the so-called 'thrifty genes' that allowed our intermittently starving ancestors to survive periods of famine by laying down body fat during the good times. Other interactions have more general effects: for example lack of exercise hastens the ageing of bone, muscle, and the cardiovascular system. Cigarette smoking impairs lung function, has anti-oestrogenic effects, brings forward the age of the menopause, and reduces bone density in addition to causing a variety of cancers.

Extrinsic factors in ageing are identified by epidemiological studies of people ageing under different conditions and by interventional studies of the effects of lifestyle or environmental modification. Epidemiological studies can be difficult to interpret because of clustering of lifestyle factors. Thus women who are health conscious enough to take hormone replacement therapy are more likely to be non-smokers, take regular exercise, have regular health checkups, and watch their weight and blood pressure. Statistical methods aimed, for example, at detecting a specific effect of hormone therapy on cardiovascular disease by 'adjusting' for the other factors are unreliable for ineluctable statistical reasons. Only a randomized placebo-controlled trial could settle this particular issue, but randomized trials of lifestyle modifications are rarely practicable.

The strongest evidence that extrinsic factors must be important comes from changes over time in the pattern of ageing. There have been dramatic improvements in the incidence of coronary heart disease and in some cancers over recent years. Conversely, there has in several countries been an increase in the age-specific incidence of fractures of the proximal femur. Overall, however, over the last 20 years older people in the United States have been living longer and also enjoying a falling prevalence of chronic disability. The presumption is that this is because of the adoption of healthier lifestyles including the use of preventative measures such as control of blood pressure, together with the rational and timely deployment of health interventions which reduce disability, such as hip replacement and coronary surgery. We have no idea whether similar changes are occurring in the United Kingdom or elsewhere, but the American evidence tells us that it could be made to happen.

Intrinsic ageing is due to cumulative damage to cells and their components which comes about because the body's systems of damage control are less than 100 per cent efficient. Damage control comprises processes of prevention, detection, repair, and replacement. Damage occurs from a variety of sources including heat, radiation, glycation (non-enzymatic crosslinking by sugar molecules of proteins, nuclear and mitochondrial DNA, and other biological polymers), and from free oxygen radicals generated in particular by mitochondrial metabolism.

Damage control is expensive in terms of energy. Although it might be biologically feasible to attain 100 per cent accuracy in damage control, organisms which achieved potential immortality in this way would still die from accident, predation, disease, famine, or warfare. They would therefore be at an evolutionary disadvantage in competition with organisms that devoted somewhat less of their resources to retarding ageing but were thereby able to maintain a higher average reproduction rate. Evolutionary pressure towards a longer lifespan will arise as environments become safer, when a slower reproduction rate can be more than compensated for by ensuring greater survival of offspring by strategies such as choice of breeding season or parental care. None the less, for any species in a specific ecological niche, investment of resources in damage control to retard ageing and prolong lifespan will always be at a level less than is necessary to abolish ageing. Thus, although maximum lifespan has increased enormously over the history of our species, we have not completely eliminated the accumulation of damage that manifests as ageing.

The fact that our lifespan has increased so rapidly, however, suggests that a fairly small number of genes may have an important effect. For the reasoning outlined above these genes are likely to be relevant to processes of damage control. Systematic comparisons of the genomes of centenarians with younger people are under way in the hope of identifying 'longevity assurance genes' whose mechanisms of action might be manipulable. Meanwhile, experimental attempts to slow intrinsic ageing are aimed at reducing the sources of damage, particularly from free oxygen radicals. Interventions under investigation include increasing the levels of free radical scavengers in cells, and reducing the production of radicals by limiting food intake and inessential mitochondrial metabolism.

Secondary ageing

Secondary ageing refers to adaptations to primary ageing changes. At the species level the female menopause is thought to be an adaptation to the age-associated increase in risk of maternal and infant death with age. In terms of getting genes into succeeding generations a woman in middle age will do better to give up increasingly dangerous and ineffective efforts to produce children containing 50 per cent of her genes and instead to devote her energies to the survival of her grandchildren each containing 25 per cent of her genes. At the individual level secondary ageing is most obvious in psychological adaptations to age-associated changes in memory and physical capabilities. Adaptations to changes in memory may range from minor obsessive traits to an old person's refusal to go shopping for fear they might get lost. The first is not obsessive–compulsive disorder and the second is not agoraphobia. Doctors need to be aware in a general way of the possibility of secondary ageing in order to resist unthinking attempts to 'normalize' some physiological parameter or aspect of behaviour where the deviation from the normal is in fact adaptive.

Sex differences in ageing

As Fig. 1 shows, death rates in females are lower than in males at all ages. In Westernized societies women outlive men by 5 to 6 years on average. Epidemiological evidence suggests that in the United Kingdom nearly four of these extra years are due to intrinsic differences between the sexes while the remining difference has developed during the twentieth century due to extrinsic effects. There is a paradox in that although women outlive men they are much more likely to become disabled and dependent in old age. As discussed below, the physical basis for this probably lies largely with sex differences in muscle bulk rather than particular disease entities. However, the impact of disability on an elderly woman is increased by the likelihood that she married a man older than herself and so has outlived her husband (and all too often his occupational pension) by more than the average sex difference in lifespan. It has also to be admitted that a woman is more likely to be capable of looking after a disabled husband than man is of caring for disabled wife. The sum of these biological, environmental, and social factors emerges in dependency rates in later life. Although one man in seven who reaches the age of 65 in the United States can expect to spend a year or more in a nursing home before death, the proportion for women is one in three.

Age-associated changes of medical significance

Anatomical and physiological changes associated with ageing are described in almost every body system. Most start to become apparent in early or middle adult life

but their magnitude and practical significance vary considerably.

The cardiovascular system

Interpretation of changes in cardiovascular function in old age is made difficult by the age-associated increase in prevalence of ischaemic heart disease. The principal anatomical alterations in the cardiovascular system include an increase in the amount of fibrous tissue in the skeleton of the heart and in the myocardium and valves and an accumulation of lipofuscin in the myocardial fibres. This last change has no evident functional significance. There is also an increase in amyloid material in the aged heart, but again this does not usually appear to have clinical significance. There is a decrease in the elasticity of the aorta and its main branches, accompanied by an increase in the diameter and length.

There is a minor decline in resting heart rate and maximum heart rate on exercise. These changes probably relate to a decrease in the number of pacemaker cells in the sinoatrial node and to alterations in their reactivity to sympathetic and parasympathetic stimuli. Cardiac output at rest does not fall with age, but maximum exercise-induced cardiac output falls to a variable extent. Mean arterial venous oxygen differences are unaltered. Since the older heart in exercise shows a smaller increase in cardiac rate than is seen in younger patients, the increased cardiac output is achieved by a relatively greater increase in stroke volume than in younger adults. The older heart is essentially more dependent than the young on the Frank–Starling mechanism for increasing output.

Owing to diminished compliance due to heart wall hypertrophy and crosslinking of proteins, the older heart takes longer to fill during diastole then does the young heart. The older patient is therefore relatively intolerant of tachycardia since at very fast rates stroke volume and cardiac output will decline. In addition, the low compliance of the ventricular wall leads to an increase in the relative importance of atrial output to diastolic filling.

Management of heart disease in older people

For various, though less than cogent, reasons older people tend to be omitted from randomized controlled trials, but it has to be assumed for ethical reasons that treatments shown to be effective for younger patients will be at least as effective for the old. In some instances benefits will be greater in later life because of the increase in background risk of morbidity or mortality. This is documented in the use of thrombolytic treatment for acute myocardial infarction and for the benefits of b-blockers after infarction. A recent study (The Heart Outcomes Prevention Evaluation Study) of the benefit of the angiotensin-converting enzyme inhibitor ramipril in secondary prevention of cardiovascular events in high-risk patients showed benefits at least as great in patients aged over 65 as in those younger.

While lack of evidence relating specifically to relevant age groups should not prevent older people from receiving treatment, care is needed with treatments for which an age-associated increase in the risk of undesirable side-effects may be expected. For example, anticoagulation in atrial fibrillation may need more careful and frequent supervision for an older person than for a younger one. The benefit of b-blockade in improving survival in chronic heart failure has so far only been demonstrated for patients aged up to 80, and since the likelihood of ill-effects of b-blockade increases with age, treatment of older patients should be introduced with low doses and increased slowly ('start low, go slow').

Blood pressure

Ordinary vascular pressures are unchanged (except for a slight rise on exercise) as is pulmonary blood volume. Mean systemic arterial pressure rises in most, but not all, populations, and presumably represents a response of susceptible individuals to environmental factors such as excessive salt intake and perhaps sociocultural stress. In populations in which blood pressure rises, systolic arterial pressure increases more than diastolic, probably as an expression of reduced elasticity of the large arteries. As ageing progresses, diastolic pressure rises to a peak and then falls. The combination of these processes leads to the increasing prevalence in old age of isolated systolic hypertension. This is a risk factor for cardiovascular disease and especially for stroke. Trials show that up to at least the age of 85 reduction of systolic pressure reduces the risk of stroke. Indeed this is another area in which, in terms of numbers needed to treat, reduction of blood pressure in older patients is more effective than at younger ages.

Trials of the treatment of high blood pressure in older people have used conventional drugs, and the assumption has been that it is the fall in blood pressure that matters and the drug used is less important. In terms of average effectiveness b-blockers tend to become less effective as hypotensive agents with age while calcium channel blockers become more effective. Concerns expressed over the long-term safety of calcium channel blockers are not convincing. Conventional treatment is of stepped-care type in which drugs are added in sequence if the response is inadequate. Recent studies in younger patients have emphasized individual variability in responsiveness to different classes of antihypertensive drug such as angiotensin-converting enzyme inhibitors, diuretics, b-blockers, and calcium channel blockers. The varying efficacy probably reflects differences in the underlying mechanisms of the hypertension in individual patients. Instead of launching immediately into stepped care, therefore, a formal trial of different classes of drugs should first be undertaken in the hope of finding a suitable monotherapy. These findings need to be verified in older people, but the idea is intuitively attractive.

An important consideration in managing hypertension at older ages is impairment of the responses of blood pressure to postural change. Baroreceptor responses are often blunted in later life, possibly as a result of sclerotic changes in the carotid sinus, and there may in addition be failure of central and perhaps peripheral vasoconstrictor responses to falling blood pressure. Older people are more susceptible to the risk of hypovolaemia. Postprandial hypotension is also common in older populations and may be a constraint on management of hypertension. When assessing older people taking medication that may affect blood pressure, lying and standing pressures should be used routinely. The role of ambulatory blood pressure monitoring in the management of treatment remains to be defined, but can be helpful in identifying episodes of unexpectedly low pressure that are missed in the clinic.

Atrial fibrillation

As noted earlier, loss of cardiac wall compliance with age increases the importance of the atrial phase of diastolic ventricular filling. Not surprisingly, therefore, the onset of atrial fibrillation, particularly if it is associated with a high ventricular rate, often has serious consequences for an older patient. Intermittent atrial fibrillation is also one of the recognized causes of recurrent syncope at later ages. Rate control is often a matter of clinical urgency for an older patient following the onset of atrial fibrillation, and electrical conversion may be required. Unless some definable precipitant such as myocardial infarction or thyrotoxicosis is present the likelihood of subsequent recurrence and permanence of atrial fibrillation rises with age. Longer-term management needs to take this into account.

Large trials have now established that in the absence of contraindications, patients aged 60 and over with atrial fibrillation should receive long-term anticoagulant therapy in order to reduce their three- to fivefold increased risk of stroke. Patients with associated valvular disease or a dilated left atrium are at even higher risk of stroke. Even in the absence of overt stroke, computed tomography scanning has shown that older patients in atrial fibrillation have a higher prevalence of 'silent' cerebral infarcts; anticoagulant therapy may therefore reduce the subsequent incidence of multi-infarct dementia.

Oral anticoagulants have more powerful pharmacodynamic effects in later life and the higher risk of complications calls for more intensive medical supervision of an old person taking anticoagulation than of a younger one. The target international normalized ratio of 2 to 2.5 is appropriate and it is prudent to reduce pre-existing high blood pressure. Where anticoagulants are thought to carry unacceptable risks, owing for example to poor compliance, intermittent high alcohol intake, or frequent falls, antiplatelet therapy should be considered. Published trials relate mostly to aspirin, and newer drugs such as clopidogrel require further evaluation in older patients.

In young patients, paroxysmal atrial fibrillation does not seem to be associated with an enhanced risk of stroke. The situation is less clear for patients aged over 60, some of whom will develop intermittent atrial fibrillation as they progress towards chronic established atrial fibrillation. There is epidemiological evidence that the risk of stroke may be greatest during this phase. Geriatricians will therefore normally anticoagulate an elderly person with intermittent fibrillation.

Respiratory system

The interpretation of changes in respiratory function with age is complicated by a high prevalence of cigarette smoking in many populations. Total lung volume does not alter but vital capacity falls and residual volume increases with age so that over the age of about 60 the critical closing volume exceeds the functional residual volume. Ventilatory capacity falls with age at a slower rate in non-smokers than in smokers. Ventilation-perfusion inequality increases slightly, probably mainly as a result of increasing inequality of ventilation. These changes in lung function reflect a decrease in elasticity of the lungs and of respiratory muscular strength but the overall consequences are minimal in terms of their effect on the blood gases. Closure of small airways during resting breathing can produce crepitations at the lung basis posteriorly in older patients. As an isolated finding these should not be overinterpreted as a sign of left ventricular failure. These same changes can also lead to

areas of atelectasis in ill older patients or following surgery. Physiotherapy for older patients after surgery should therefore concentrate on expanding the lung bases rather than clearing sputum.

The principles of management of respiratory disease do not vary with the age of the patient. Diagnostic probabilities may need to be adjusted for age and for cohort effects. Cough due to left ventricular failure or to oesophageal reflux disease may become more common, and older people are more likely to have experienced past industrial exposures, for example to asbestos and coal dust.

Renal function and fluid and electrolyte balance

Old people are particularly susceptible to disorders of fluid and electrolyte balance. This is due in part to age-associated changes in physiology and partly to a higher incidence of challenges to homoeostasis from disease and drugs.

Renal function

Age-related changes in renal function are well documented. The glomerular filtration rate falls as do tubular reabsorptive and secretory capacities. Typically, these changes exceed the decline in lean body mass so that serum urea and creatinine concentrations rise slightly. Several formulae have been devised to try to estimate the glomerular filtration rate from serum creatinine taking age, weight, and sex into account. Various forms of the Cockroft–Gault equation are available as a means of estimating creatinine clearance from serum creatinine and body weight. This formula is only approximate and can mislead in extreme old age and in the presence of obesity where calculation based on ideal body weight or lean mass body weight may be more accurate. Where accuracy is important, such formulae are no substitute for formal creatinine clearance based on a 24-h urine collection.

The response to an acid load is impaired and the maximum rate of secretion of hydrogen ions falls. Changes in blood pH in response to an acid load are therefore greater in magnitude and longer in duration than in younger people. Response to antidiuretic hormone is reduced and water conservation less efficient.

Thirst

A delayed response to fluid deprivation due to impaired thirst mechanisms in old age may compound the effect of delayed renal responses to changes in fluid status. The sensation of thirst is decreased in later life. In an experimental study of 24 h of fluid deprivation younger volunteers felt thirsty while older volunteers did not. When given access to water the older volunteers drank less than the young. Similar differences between young and old were found following infusions of hypertonic saline in which the younger volunteers were able to adjust serum osmolality by water ingestion more accurately than the old. It is not clear, however, to what extent the changed physiology is due to insensitivity of the osmoreceptors and baroreceptors rather than to a deficit in the opioid-mediated drinking drive.

The decline in thirst sensitivity with age is one reason why older people are at enhanced risk of dehydration. This effect may be exaggerated if an old person voluntarily restricts fluid intake in the hope of controlling urinary urgency or incontinence, or to avoid having to call for nursing assistance when in hospital.

Conversely some elderly people, often hypertensive women, show an enhanced thirst response to diuretics (particularly amiloride) with increased water intake and hyponatraemia.

Other age-associated changes

Average renin and aldosterone levels diminish with age with consequent impairment of sodium conservation. Renal concentrating ability is also reduced, which is partly due to a decrease in medullary hypertonicity and partly to impaired renal responsiveness to vasopressin which may result in excessive water losses. The ability of older patients to cope with volume expansion is also impaired and older individuals take longer to excrete a sodium load. There is also a possibility that vasopressin secretion may be impaired.

Secretion of atrial natriuretic peptide in response to hypervolaemia may be reduced in older people and the responsiveness of the kidney to atrial natriuretic peptide may also be reduced.

Implications for clinical care

Assessment of fluid status is an essential component of the evaluation of an elderly patient who is unwell. Sometimes a degree of fluid deprivation will have preceded the illness because some older patients deliberately minimize their fluid intake with the hope of reducing problems of urinary urgency and incontinence. During an illness the most sensitive assessment of fluid balance is by daily weighings, but every attempt should be made to maintain accurate fluid balance charts as well. Infections, commonly pneumonia or urinary tract infections, are frequent causes of dehydration. This is in part due to loss due to fever, which may be overlooked if body temperature is not measured with especial care and after a patient has recovered from travelling in a cold ambulance to hospital. Older people are also more susceptible than the young to the effects of high environmental temperature.

Many old people are on long-term diuretic treatment which can exacerbate the effects of illness as well as causing problems in their own right. In addition to hypovolaemia and postural hypotension, hyponatraemia and disorders of potassium balance can be caused by diuretic therapy. Diuretics may precipitate hyperuricaemia and gout in older patients and destabilize diabetes. The significance of the lipid-raising effect of diuretics on lipid levels in old age is less clear. As noted above, some older people, especially women with hypertension, seem particularly susceptible to diuretics and overcompensate for diuresis with excess water intake so leading to hyponatraemia.

As far as possible dehydration should be corrected by oral intake. Intravenous therapy provides for more rapid and accurate correction but care must be taken over the rapidity of correction of hypo- or hypernatraemia. Subcutaneous fluid replacement is now established as an option where intravenous access is difficult or fluid correction is less than urgent. The infusion needle can be placed subcutaneously in the abdominal wall, the axilla, or the subclavicular area, but for a confused patient between the shoulder blades can be a less troublesome site. Hyaluronidase, 1500 iu, can be given as a bolus through the cannula if the infusion runs too slowly but is expensive, does not improve comfort, and is not usually necessary with an older patient. The infusion site should be changed every 24 or 48 h. Normal saline is well-tolerated by this route, and 5 per cent dextrose can also be given safely at doses up to 1 ml/min (1.5 litres per day). If necessary, potassium chloride up to 40 mmol can be added to each litre of solution, but this may increase complication rates due to local inflammation and secondary infection of the infusion site. Colloid and hyperosmolar solutions should not be given subcutaneously.

Urinary incontinence

This socially disabling condition affects approximately 2 per cent of middle-aged men and 12 per cent of middle-aged women and increases with age to a prevalence of around 8 per cent in men and 16 per cent in women over the age of 75 years. A number of classifications of urinary incontinence have been proposed and Table 2 presents a common version. The patient with acutely developing incontinence due to illness or injury should be reassured and every effort should be made to ensure that it does not continue into a chronic form. As with all problems of older people, possible iatrogenic causes must be reviewed, including rapidly acting diuretics and sedative drugs. Excessive urine output due to hyperglycaemia or hypercalcaemia may present with urinary incontinence, but nocturnal urinary frequency and incontinence due to unrecognized heart failure is much more common. Faecal impaction is another common remediable cause of urinary incontinence in hospitalized older people.

Many patients with dementia become incontinent, but incontinence should never be attributed solely to dementia unless the latter is severe and until all treatable causes have been excluded. Functional incontinence due to the inability of an old person to reach a toilet in time is unnecessarily prevalent, particularly in hospitals where old people with mobility problems may have their beds too far from the toilet or where nurses cannot or do not answer bells promptly.

Overflow incontinence when the bladder is only able to overcome an outlet obstruction at high volumes is common in men with prostate difficulties. It is diagnosable by postvoiding ultrasound examination and often requires surgical intervention. Where that is not feasible, or if a patient declines surgery, medical approaches including a-adrenergic blockers (given with care on account of the risk of hypotension) may be helpful. Where there is prostate enlargement there is some evidence for benefit from finasteride which inhibits 5a-reductase, which metabolizes testosterone to the more potent dihydrotestosterone. This can lead over a period of some

months to a reduction in prostatic size and improvement in urinary flow rates and obstructive symptoms. There are some antiandrogenic side-effects including impotence and decreased libido, and the drug has to be handled with care by women of child-bearing potential, a fact which may be relevant to carers or nurses involved in dispensing the drug.

Urge incontinence is usually due to instability of the detrusor muscle of the bladder and may also arise through irritation of the bladder by infected urine when it responds to appropriate antibiotic therapy. After identifying the sensitivities of the infecting organism, antibiotic treatment is given and urine culture repeated. If the incontinence is not improved despite eradication of the infection antibiotics should not be continued. Studies have shown no benefit from treatment of asymptomatic urinary infections in older patients. The contractions of an unstable bladder may be inhibited by anticholinergic drugs but these must be used with care as they may induce glaucoma and constipation. Lipid-soluble anticholinergic drugs may also cross the blood–brain barrier and cause delirium. The drugs most commonly in use on account of claims of some degree of specificity for bladder receptors are oxybutynin and tolterodine.

The certain diagnosis of unstable bladder and associated urethral dysfunction requires invasive investigation including cystometry. In practice most cases of urinary incontinence in older patients can be dealt with empirically, and invasive investigations should be reserved for the more intractable cases. It is useful for medical and nursing teams providing care for older people in the community and residential settings or hospitals to agree on a general plan for approaching problems of incontinence. Figure 2 outlines an example of such an approach. In many instances urinary incontinence may be a problem only at night. A single dose of imipramine in the evening may be of benefit, partly through its anticholinergic effects. Setting an alarm to wake the patient to empty his or her bladder prophylactically may also be of benefit. For some patients a small dose of a rapidly acting diuretic in the afternoon followed by restriction of fluids until bedtime will enable the cardiovascular system to reabsorb fluid pooled during the day in the lower limbs without an inconvenient increase in urinary output. The use of dDAVP or other analogues of vasopressin at bedtime in order to produce an antidiuresis during the night has been reported but it carries a high risk of fluid overload and hyponatraemia in older patients.

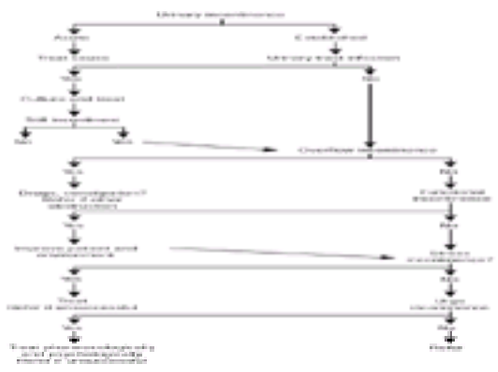

Fig. 2 Illustrative algorithm for management of urinary incontinence in elderly patients 'Refer' indicates referral to specialist urologist or gynaecologist.

Long-term catheterization may need to be considered in intractable situations. This is an invasive procedure with a risk of complications and should be undertaken only after the agreement of the patient. Some older patients will choose catheterization in preference to incontinence pads. Catheters should be inserted and changed under strict asepsis and with consideration of antibiotic cover for patients with cardiac valvular disease or joint prostheses. The urine of an older patient with an indwelling catheter will almost inevitably become infected. Although treatment of symptomatic infection must be considered, and the risk of ascending pyelonephritis and haemosepsis kept in mind, repeated courses of antibiotics can lead to colonization with highly resistant organisms. This is particularly undesirable in institutional care. Although a somewhat old-fashioned approach, manipulation of urinary pH by potassium citrate can help and urinary antiseptics such as hexamine hippurate can also be effective. The last should not be given with alkalinizing agents, and although theoretically less effective where an indwelling catheter prevents bladder filling it can be helpful. It should be avoided in renal failure.

Where possible, leg bags should replace the undignified bag-on-a-stand with which so many hospitals unnecessarily advertise their older patients' disability. Problems of backflow when the patient's bladder is lower than the bag when sitting usually indicate a need for a more suitable chair than for putting the bag on the floor.

For older men who are incontinent only at night a condom-type appliance with drainage bag may be helpful, although leakage and discomfort are common complications. No reliable analogous device has yet been produced for women.

A wide range of incontinence pads and holders is available. Specialist incontinence nurses are available in many hospitals and primary care districts who can advise on the most suitable treatment for individual patients. Charities concerned with disabled people are also a valuable source of expert advice.

The gastrointestinal system

Gastric atrophy becomes increasingly common as age advances but many older people retain full capacity for hydrogen ion secretion. Minor atrophic changes in the mucosa of the small bowel are described but there is no significant change in the absorption of nutrients that are absorbed by passive diffusion. Bacterial overgrowth syndrome may occur in an older patient with an anatomically normal bowel, and needs to be included in the differential diagnosis of diarrhoea, partially intermittent diarrhoea, and malabsorption. The increased risk with age is thought to reflect gastric hypoacidity and reduced motility of the small bowel. Reduced motility of the large bowel is also common and may be associated with diverticular disease or overuse of purgatives.

There is an age-associated reduction in hepatic mass and blood flow, with some consequent changes in first-pass metabolism of drugs and gastrointestinal hormones, but no significant changes in hepatic function as measured by conventional tests.

Gastrointestinal disease

Older people are liable to the full range of gastrointestinal disorders. Even coeliac disease may present for the first time in old age. Gastrointestinal haemorrhage is common and frequently associated with the ingestion of non-steroidal anti-inflammatory drugs. Management of an older patient is along conventional lines but age-associated loss of adaptability should be compensated for by an increased readiness to institute invasive monitoring and to intervene endoscopically or operate sooner rather than later in deteriorating situations. Uncertainty among junior hospital medical staff can be reduced by locally agreed guidelines for management of what is a common but dangerous emergency. Guidelines will only be effective if adhered to by surgeons as well as physicians.

Constipation is a common complaint among older people, but before initiating treatment it may be helpful to ensure that a patient's expectations of bowel function are reasonable. The implications of a recent change of bowel habit must also not be missed. Where chronic constipation is a problem, a daily intake of 10 to 20 g of dietary fibre should be aimed at, with if necessary a stool softening agent such as lactulose. Doses of stimulant laxatives such as senna may be necessary but should be avoided as much as possible.

Faecal incontinence

Recurrent faecal incontinence, with our without associated urinary incontinence, affects about 1 per cent of people aged over 65. Faecal and double incontinence are associated with a range of anorectal and neurological disorders including dementia. In later life, however, the commonest cause is rectal or sigmoid overloading due to constipation (faecal impaction). In middle-aged women a pelvic neuropathy, sometimes possibly a consequence of childbirth, is encountered as a risk of faecal incontinence and may be improved by surgical intervention. Any form of acute or chronic diarrhoea will also lead to incontinence in a proportion of cases. Faecal impaction most commonly occurs in the rectum and can be readily diagnosed by rectal examination. 'High' impaction can occur at the rectosigmoid junction and above and may need to be diagnosed by a plain abdominal radiograph. High impaction must always raise the possibility of obstruction by a carcinoma or inflammatory stricture due to diverticulitis. Treatment of impaction requires enemas and suppositories, rarely manual evacuation, and then a regular regime of a stool softening

medication such as lactulose.

In some cases where faecal incontinence is an intractable problem it can be ameliorated by giving constipating medicine such as loperamide with bowel lavage once a week. This is not an easy regime to establish but can be of help for an old person who wishes to remain in his or her home or to ease nursing problems in residential care.

The locomotor system

Muscle mass, strength, and power

Significant and progressive alterations in average body composition appear in the fifth decade of life. There is a decline in lean body mass with a corresponding fall in oxygen consumption largely attributable to a decline in muscle mass. This is also associated with a decline in muscle strength and power. (As in physics, strength is conceptualized as maximum force and power as the maximum rate of doing work.) The consequences of this trend are more prominent in women who, on average, start adult life with less muscle tissue than men. The age-associated decline is such that by the age of 80 the great majority of women in economically advanced societies are unable to rise from a chair without using their arms to help. This is probably the chief reason why disability levels are so much higher in old women than old men. Improvement in muscular power can be achieved even in the very old by appropriate exercise regimes, which should be considered as part of the rehabilitation programme of an older person who has had to be off his or her feet for more than a day or two.

Bone and joints

Fractures: osteoporosis

The major changes with age in the bony skeleton, which include a steep increase in the prevalence of osteoporosis after middle age, are discussed in Section 19. Loss of bone tissue with age occurs in all humans but appears to vary in severity with place, time, and race. The most important manifestation of the decline in bone mass is a reduction in the mechanical strength of bone and an increase in the tendency to fracture. Although many fractures increase in risk with age and the rising prevalence of osteoporosis, the three 'classical' osteoporotic fractures are those of the vertebrae, the distal forearm, and the proximal femur. Vertebral fractures start to appear at the time of menopause in women and increase in prevalence thereafter. The great majority of limb bone fractures in old age are caused by simple falls. The epidemiological pattern of fractures is partly determined by the causes of falls, the speed of protective responses, and the presence or absence of 'passive' protective factors such as floor coverings, clothing, muscle, or subcutaneous fat. Women are more likely than men to fall and there is an exponential increase in the risk of falling from about the age of 60. It is therefore not surprising that there is a similarly exponential increase in the risk of proximal femoral fracture with a much higher risk in women over the same age range. However, distal forearm fractures which show a steep increase around the age of menopause do not continue to increase in incidence through old age. This probably indicates that the older the person is the less likely he or she is in a fall to throw out an arm in time as a protective response.

Fractures: osteomalacia

Osteomalacia is now rare as a cause of falls and fractures in old age, but it may need to be considered in an older person who has been housebound, has a low dietary vitamin D intake, and takes drugs such as antiepileptics that induce hepatic enzymes that destroy vitamin D derivatives. Age-associated changes in the ability to synthesize vitamin D in the skin and to absorb calcium from the gut are coupled with a reduction in the renal 1a-hydroxylase activity that metabolizes vitamin D into its highly active 1,25-dihydroxy vitamin form. Relative vitamin D deficiency may play a part in the genesis of osteoporosis. There is a growing literature suggesting that minor degrees of vitamin D deficiency during the winter in temperate latitudes may lead to compensatory increases in parathyroid hormone secretion and negative bone balance. This can be prevented by a daily intake of 400 to 600 iu of vitamin D from October to April.

Arthritis

The reasons for the virtual universal occurrence of osteoarthritic changes in many joints in later life are uncertain but may include the mechanical effects of time-related wear and tear, age-associated changes in the metabolism of joint cartilage, and subchondral bone, or a disease process unrelated to age. The epidemiology of knee osteoarthritis suggests that wear and tear from occupation injury and obesity is important. Obesity seems both to predispose to arthritis of the knee and to be a factor in making the arthritis painful. Weight loss is therefore an important therapeutic approach. In contrast, osteoarthritis of the hip seems often to be a long-term consequence of minor (or major) forms of congenital dysplasia of the hip and the effects of occupation and obesity are much less clear. There is evidence of a genetic predisposition to the syndrome of generalized osteoarthritis.

Falls

Falls and the fear of falls are important causes of morbidity among older people. Community surveys indicate that a quarter of people aged 65 to 69 fall at least once in the course of a year and this annual prevalence doubles by the age of 80. Women are more liable to falls than men. This may partly reflect greater activity but is also related to lower muscular strength. Falls are a cause of direct and potentially fatal injuries such as fractures and head injury. Old people who are unable to get themselves up again may suffer the additional problems arising from a 'long lie'. These include hypothermia, pressure sores, and rhabdomyolysis. The last can in severe cases lead to acute renal failure and serum muscle enzymes should be checked in old people presenting to medical care after a fall with long lie. Significant haemorrhage following falls is an increasing problem as more older people are being prescribed anticoagulants for atrial fibrillation. A history of falls is a relevant issue in deciding whether an older person should be established on anticoagulants.

Falling, especially if associated with inability to rise again, is an extremely unpleasant experience for most older people and fear of further falls can lead to a form of 'postfall syndrome' in which the patient becomes morbidly afraid of falls, progressively more immobile (which may increase the risk falls as much as decreasing it), and socially isolated. Old people with this condition may seek premature institutional care or may be pressurized into care by worried family or neighbours. Falls by old people that come to medical attention need therefore to be taken seriously and possible preventive interventions sought. There have also been some studies of primary prevention at a community level but the cost-effectiveness of these approaches has not yet been established.

Causes of falls

Some general age-associated changes contribute to the rise in risk of falls. The syndrome of non-rotatory dizziness, which is common in later life and epidemiologically associated with an enhanced risk of falls, probably represents a temporary failure of the brain to achieve a coherent integration of positional data from eyes, inner ear, and proprioceptors. Proprioceptive information is reduced by increased variance in neural conduction time from peripheral tissues, and by damage to receptors in joint capsules by arthritis in the neck and elsewhere. The older patient therefore becomes increasingly dependent on vision for spatial orientation. Poor lighting levels or a visually confusing environment, as experienced on a moving escalator for example, can be particularly hazardous.

A large number of more specific risk factors for falls have been identified. These are typically classified as intrinsic to the patient and extrinsic in the environment, but as with all such dichotomies interactions are important. Interpretation in terms of causality can be uncertain because of confounding factors. In observational studies sedative and antidepressant drugs emerge as commonly associated with an increased risk of falls. Long-acting benzodiazepines and antidepressants are probably directly associated with an increased risk, but a link between diuretic therapy and falls found in some studies is probably more often mediated by the cardiovascular disease for which the drugs have been prescribed.

Some falls are no more than a misfortune that afflicts all of us by chance. An older person who falls only once or twice in a year will commonly be found to have no specific remediable cause. Old people who fall more than twice in a 12-month period should be investigated further. Extrinsic causes in the home should be identified, and the help of a skilled occupational therapist may be needed. Inadequate lighting, slippery floors, inadequate handholds, sloppy footwear, and loose rugs are common hazards. Where falls are not readily explicable in terms of an identifiable environmental hazard medical appraisal is needed. Table 3 lists some of the commoner medical causes of falls. Patients who find themselves on the floor with no memory of falling may have been unconscious at least momentarily and this suggests cardiac dysrhythmia, syncope, or epilepsy. Such a history may not be obtained consistently from patients subsequently shown to suffer from syncope, however, and this possibility has to be borne in mind for any older patient suffering repeated falls. Full investigation requires tilt-table and carotid sinus massage testing. The mechanism of syncope may be predominantly cardioinhibitory in which case a demand pacemaker can be of benefit. Syncope that is mediated by

systemic hypotension is less amenable to treatment.

Ambulatory electrocardiographic monitoring is a commonly requested investigation for older patients suffering falls or intermittent lapses of consciousness. Findings can be difficult to interpret in the absence of symptomatic events during recording because the prevalence of intermittent cardiographic abnormalities is high in later life. None the less, intermittent arrhythmias or significant pauses may be identified that justify a trial of therapy if compatible with the clinical history.

Where no remediable causes of falls can be established thought needs to be given to tertiary prevention of the consequences of further falls. A physiotherapist should train the older person in getting up after a fall or in moving across the floor to an alarm or telephone to summon help. An alarm system may need to be installed in the patient's home provided he or she can be trained to use it. Otherwise, some system of regular surveillance by statutory services, volunteers, or good neighbours may be more useful. The risks of further falls need to be discussed fully with the patient and with concerned relatives. The right of an old person to continue to live in his or her own home, even where that carries some risk, may need to be defined for relatives pressing for institutionalization to relieve their anxieties rather than the patient's.

The endocrine system

Historically, there have been many attempts to explain ageing as a consequence of sequential endocrine failure, in the hope that suitable replacement therapy might halt or reverse the process. Apart from the obvious changes in ovarian function with the menopause, there are minor declines in circulating thyroid hormone levels, a reduction in the rate of secretion of insulin in response to raised blood sugar levels, and a decline in tissue sensitivity to insulin. The release of antidiuretic hormone in response to osmotic loads increases, perhaps in association with reduced renal responsiveness and changes in the sensitivity of blood volume receptors. There is little evidence of any abnormality of parathyroid, adrenal, or pituitary function as a universal feature of old age. The response of the adrenals and of adrenocorticotrophin secretion to stress is essentially unaltered. Indeed, following injury ACTH secretion is more prolonged in older patients than younger but the significance of this is unclear.

There is a gradual age-associated decline in average testosterone levels in ageing men. There is no abrupt change corresponding to the female menopause, nor, on evolutionary grounds, would one expect there to be. Testosterone implants are available in the private medical sector in some countries but there is at present no scientific justification for the claimed benefits of 'normalizing' testosterone at young adult levels. Some older men have been found to have very low levels of growth hormone and in uncontrolled experiments show increases in muscle bulk and strength when replacement therapy is provided. Functional benefits have yet to be demonstrated, and side-effects from growth hormone can be severe.

Blood levels of dehydroepiandrosterone, a weak androgen, are high in fetal life, decline after birth, and then rise again from puberty into early adult life. There follows a steep decline with age and there have been reports of improvements in function and wellbeing in later life from supplements. Again, larger and better trials are needed.

Diabetes mellitus

The clinical presentation of diabetes in old age may differ from that of younger patients. Many patients are diagnosed as a result of routine testing during a medical or surgical illness and some because of the development of disorders associated with diabetes such as peripheral vascular disease or cataract. Relatively few present because of classical symptoms such as weight loss and polyuria but a small proportion present with life-threatening metabolic decompensation in a hyperosmolar state.

There is now sufficient evidence to justify trying to control hyperglycaemia. Dietary treatment and oral hypoglycaemic agents are firstline treatment but insulin should not be withheld if it is necessary for control of hyperglycaemia. Shorter-acting oral hyperglycaemic drugs such as gliclazide are preferred and longer-acting drugs such as chlorpropramide and glibenclamide should not be used for older patients.

Diabetes mellitus interacts with other risk factors for cardiovascular disease. Diabetic patients of any age should be persuaded to give up smoking. Control of blood pressure is very important, with evidence to support the use of angiotensin-converting enzyme inhibitors in preference to other classes of drugs. Hypercholesterolaemia should be reduced.

The specific complications of diabetes occur more frequently in older than in younger patients. The majority of patients blind from diabetes are aged over 60 and the prognosis of diabetic retinopathy in old age is less favourable than at younger ages. Photocoagulation remains effective, and the results of cataract extraction are excellent except when retinopathy is contributing to the visual impairment.

Foot care is an important aspect of the management of diabetes and it is important to bear in mind that peripheral neuropathy and vascular disease may render the elderly diabetic patient particularly prone to pressure sores of the heels if confined to bed for any length of time.

Hypothyroidism

Most cases of hypothyroidism in old age are of autoimmune origin, although previous thyroid surgery and radio-iodine therapy account for a proportion. Classical clinical signs such as cold intolerance, hair loss, and coarsening of the skin are less common as presentations in old age than an insidious decline in health and mobility with psychiatric manifestations, particularly depression. Hypothyroid coma, sometimes associated with severe headache and fits, and hypothermia are less common but important presentations. The most physical signs are a change in voice and delayed relaxation in tendon reflexes, often most easily recognized in the arm reflexes of an older patient.

The principle of 'start low, go slow' in prescribing for older people is important when starting thyroid replacement therapy, particularly if myocardial ischaemia is known to be present. A problem in management may be failure to comply with treatment and the responsible doctor must ensure that lifelong treatment is in fact lifelong. It may be necessary to enlist the aid of a relative, neighbour, or visiting nurse in ensuring compliance.

Hyperthyroidism

Hyperthyroidism is less common in old age than hypothyroidism but is even more likely to present in an 'atypical' manner. Weight loss, anorexia, gastrointestinal, and cardiovascular symptoms predominate. Cardiac failure, often associated with atrial fibrillation resistant to digitalis, is a common cardiac presentation. Eye signs and goitre are less common in elderly patients, and apathy or depression, rather than tremor and agitation, may be prominent.

Treatment is begun with carbimazole followed when thyroid function is returned to normal by radio-iodine. Thyroid replacement therapy is often required later either because of an ablative dose of radio-iodine or because of subsequent decline in thyroid function. Careful follow-up is therefore necessary.

The nervous system

Age changes in the nervous system are among the most important because of their significance in the psychology and psychiatry of ageing and in the production of disorders of movement. Numbers of neurones in some parts of the nervous system, for example the motor neurones of the spinal cord, the Purkinje cells of the cerebellum, the cells of the substantia nigra, and parts of the neocortex, fall with age. There are no changes in other parts, for example in several brainstem nuclei. Anatomical abnormalities of the neurones that remain are also described including the accumulation of lipofuscin and the loss of dendrites and dendritic spines in cortical neurones. The anatomical basis of alterations in cortical function may thus be due both to a reduction in the number of neurones and in the connections between them. There is a reduction in peripheral nerve conduction velocities, both motor and sensory, differing in different nerves and reflecting a fallout of fibres of all sizes. There is also an increase in the variance of nerve conduction velocities which probably contributes to the inaccurate transmission of information and may impair the cognitive function of the brain as a parallel computer. The increase in nerve conduction variance may explain the frequent bilateral loss of ankle tendon reflexes above the age of 60. Loss of vibration sense in the feet and ankles may be another manifestation of the same process. Although commonly of no clinical significance, such findings need to be interpreted with care as it is important not to overlook a peripheral neuropathy, due for example to vitamin B_{12} deficiency.

Temperature control

Central autonomic nuclei, for example the intermediolateral cells in the spinal cord and first- and second-order autonomic neurones outside the central nervous system, show a fall in cell numbers with age. Associated changes in autonomic function include alterations in the control of heart rate (see above) and abnormalities of temperature regulation. Older people are more susceptible to hypothermia and to heat stroke. The threshold for appreciation of skin temperature changes may increase by as much as tenfold over the age range. This contributes to older people's inability to recognize temperature change and to respond appropriately. There is reduced cutaneous vasoconstriction in response to cold and impaired vasodilatation and sweating in response to increase in body temperature. Reduction in sweating is due in part to a reduction in the number of sweat glands. Failure of shivering is common and makes hypothermia more likely, while coexisting undernutrition, even of brief duration, reduces hepatic thermogenesis.

Hypothermia

In many countries, especially the United Kingdom, there is an excess of deaths during winter. This is mostly due to influenza and to cardiovascular disease and the contribution from hypothermia is very small. None the less the diagnosis is important as it can easily be missed in its early stages and carries a high fatality. Most elderly patients with hypothermia present after a period of 2 or 3 days of cold weather, but it is entirely possible for hypothermia to develop in hospital in midsummer at United Kingdom latitudes especially if disease or drugs play a part. The fall in central temperature occurs over 24 to 48 h and its principal manifestations are progressive ataxia and slowing of cerebration continuing to stupor and finally coma. On clinical examination, the skin on unexposed surfaces such as the axilla, chest, and abdomen feels cold to the touch. There may be oedema of the face and eyelids resembling that of myxoedema but due to redistribution of fluid between the intra- and extracellular compartments and resolving with correction of hypothermia. The pulse is slow unless severe physical illness has raised it towards normal, and the blood pressure is difficult to record or is low. Tendon reflexes are normal unless there is associated hypothyroidism. The diagnosis is made by recording the rectal or other form of core temperature, and this should be done without delay on all occasions when an oral temperature is recorded at 35 °C or less. The electrocardiogram may show characteristic bradycardia and J waves (Fig. 3), but in severe hypothermia there is often atrial fibrillation with a slow ventricular response.

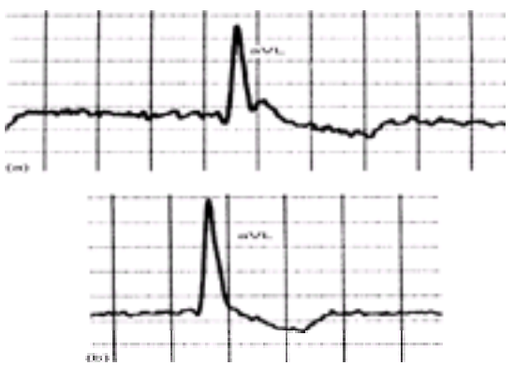

Fig. 3 An Osborne (J) wave seen in the electrocardiogram of a patient during hypothermia (a) and after recovery (b). The characteristic deflection lies between the QRS complex and the beginning of the ST segment. The pathophysiology of the J wave is uncertain.

Hypothermia in elderly patients carries a high fatality. Although patients with hypothermia due to drugs commonly recover, those with severe physical illness as a cause usually die. Management is rendered difficult by the absence of comparative trials. There is no rational basis for the use of steroids, and intravenous fluids are dangerous because they may produce pulmonary oedema. Intravenous glucose is not metabolized and insulin is ineffective in the hypothermic state. In some younger patients, particularly those suffering acute accidental hypothermia, there may be benefits from rapid warming, but in older patients slow rewarming is the usual practice. A rise in body temperature, optimally 0.5 °C/h, is obtained by exposing the patient to a relatively high ambient temperature of 30 °C. Improvement is shown by an increase in rectal temperature and pulse rate, maintained blood pressure, and improvement in level of consciousness.

An elderly patient who has recovered from hypothermia should be regarded as at risk for further episodes, and relatives and social agencies should be alerted. The long-term prognosis is better in patients who have survived an episode of hypothermia due to identifiable causes.

Hyperthermia

An increase in mortality on continuing care wards and among elderly people living at home has been recognized during heat waves. It is not known how many of the extra deaths are due specifically to hyperthermia, in the sense of rise in body temperature, rather than to dehydration or general stress. As with hypothermia part of the problem seems to be the failure of the older person to recognize that a problem is developing and for many a cool environment is not easily available. Adequate fluid intake should be recommended and tepid bathing may be the only available way of preventing a rise in body temperature.

Vision

Age-associated changes in the eye and ear are well documented. Decrease in elasticity of the lens begins early in life but only becomes symptomatic in the fifth decade when presbyopic hypermetropia results from failure of accommodation due to changes in the lens itself and in its capsule and suspensory ligaments. Cataracts increase in prevalence with age but vary in frequency between racial groups and with factors such as diabetes and family history. The media of the eye become less translucent with age and there is more scattering of light. Older people therefore have lower contrast sensitivity than the young and need more light and sharper contrast in reading material and in important environmental cues such as marker strips along the edges of steps. At a practical level, use of the inverse square law in placing lights nearer to what they are required to illuminate may be preferable to simply increasing their wattage. The eye media also become yellow with age so that the older eye is less sensitive to blue, a problem that can on occasion lead to mistakes in medication as the older patient fails to register which tablet is 'the blue one'. The public environment, including buildings such as hospitals, is often unnecessarily difficult for older people to use because of failure by architects to understand the visual problems of an ageing population.

Driving

Vision is an important determinant of fitness to drive a car. The loss of the right to drive can interfere seriously with an older person's quality of life as well as having a profound symbolic impact as a sign of disability and social marginalization. The common notion that older drivers are dangerous is exaggerated. In terms of preventing fatal road accidents the most effective intervention would be to refuse licences to males aged under the age of 25. Accident rates are very high for drivers, especially men, in early adult life where high-speed accidents with high injury and fatality rates are characteristic. Rates then fall into middle age but begin to rise again after the age of 70. At later ages, however, accidents tend to occur at low speeds with correspondingly low death and injury rates, and typically involve side collisions at road intersections. One reason for this may be a diminution in the size of the functional visual field with age. There is laboratory evidence that the functional visual field can be enlarged by training but the practical impact is not yet convincing. As with so many other aspects of modern life, road design could be improved to make travel safer and pleasanter for the increasing numbers of older drivers in society.

In addition to vision, a range of attributes including cognitive function and physical ability in manipulating controls, contribute to a person's ability to drive safely. In the absence of gross abnormalities, an older person's driving ability can only be assessed in a road test. Clinical examination and computer simulation are not adequate substitutes.

Specialized testing may also lead to useful advice about modifications to an old person's car that will enhance safety.

Hearing

A degree of high-tone deafness (presbyacusis) is probably universal in humans. It reflects several mechanisms and although it is partly an intrinsic true age change,

there is little doubt of the importance of prolonged exposure to industrial and other environmental noise. Loss of high-tone hearing with the associated difficulty in following one voice against a background of others is socially and occupationally disabling. Minor degrees of deafness also have a more subtle effect on cognitive performance. Normal language has a high degree of informational redundancy, so that in conversation the hearer has often already understood and is preparing an answer before the speaker has concluded. Hearers who have to concentrate on listening to complete sentences in order to be sure they have understood are at a disadvantage in terms of processing and reacting to what is being said. This is all too often interpreted as cognitive impairment, and in confrontational situations, such as legal proceedings, can be exploited by unscrupulous interlocutors.

Psychiatric disorders in later life

Physicians should be aware that there is always a psychological element to physical disease. At one extreme physical symptoms and even signs can be a manifestation of a somatization syndrome. At the other extreme there will inevitably be a psychological reaction to physical illness. This may include a panic reaction, but more commonly fear and anxiety manifest in different ways ranging from denial of symptoms to morbid preoccupation with them. It is part of a physician's duty to recognize the psychological dimension to a patient's illness and to respond to it appropriately. All this is true at any age but requires particular thought with older patients for whom the possibility of death is a constant presence and disability a constant dread. These problems may be compounded with a degree of cognitive impairment, background depression, and an increased susceptibility to delirium. For practical reasons the physician will need to deal directly with a broad range of such problems but should be ready to invoke the aid of psychogeriatrician colleagues in situations that are less than straightforward.

Delirium

Delirium, one form of acute confusional state, can affect an acutely ill patient of any age but becomes more common in later life. Any toxic febrile condition can precipitate delirium as can primary insult to the brain such as stroke, meningitis, encephalitis, or subarachnoid haemorrhage. The mechanisms of delirium remain obscure but there is suggestive evidence that in toxic states it may involve the leakage across the blood–brain barrier of neuroactive compounds, normally not present in the bloodstream or normally excluded from passage into the brain. In some instances the patient may be quiet, drowsy, and withdrawn, perhaps quietly muttering, but more commonly a delirious older patient is agitated, restless, noisy, paranoid, and sometimes aggressive. The diagnosis must be suspected in old person who shows an abrupt deterioration in cognitive function or an abrupt change in personality. In toxic conditions such as urinary infection, haemosepsis, or pneumonia the delirium may appear before any other signs of infection such as fever or leucocytosis.

The central element of management is to diagnose and treat the underlying cause and in some instances this may call for the 'blind' institution of antibiotic therapy while cultures of blood and urine are awaited. Although an agitated depression can be made more manageable by sedative drugs such as haloperidol these bring with them a risk of secondary complications and as far as possible doses should be kept to a minimum and the patient's disturbed behaviour handled by skilled nursing. This will be facilitated by an attempt to understand what the patient is experiencing as the cause for his or her behaviour. In a delirious state consciousness is clouded and experience may have a dreamlike quality with all that means in a sense of ill-understood dread and powerlessness to escape or defend oneself. Memory is impaired so that although careful explanation to the patient of who people are and what is happening is an important part of care it may need to be repeated at frequent intervals. Attention is disrupted so that the patient may not be listening when spoken to or may focus on some unimportant feature of the environment which may take on particular and often menacing significance. This is often coupled with misinterpretation of what is seen or heard so that a smoke alarm in the ceiling becomes a Martian death ray machine, or a pop song on a distant television set becomes the howling of a fellow prisoner in a torture chamber. In some instances, characteristically where alcohol or certain drugs have played a part in the delirium, visual hallucinations may occur which are often of a frightening or threatening kind.

Where possible a delirious patient should be nursed in a quiet room with good lighting and preferably in the company of a well-known friend or family member. Extraneous noise should be avoided as far as possible and constant reassurance and explanation provided. It is reassuring to delirious old people if doctors are dressed and behave like doctors, and nurses like nurses of more gracious times. Being addressed in old age by some overfamiliar ambiguous stranger using one's forename can be alarming. All forms of physical restraint are terrifying and should not be used.

Patients with pre-existing brain disease such as dementia are more prone than average to delirium but one of the differential diagnoses of delirium is of the acute panic and alarm of a demented person removed from their familiar environment. It is important to recognize this syndrome of 'decompensated dementia' in an older patient since if they are not returned as quickly as possible to their familiar environment their cognitive hold on reality may become permanently disrupted. The accident and emergency unit of a busy general hospital is a common setting for this problem.

Dementia

Dementia is distressingly common in later life with a prevalence of approximately 5 per cent over the age of 65. In its fully developed form it is conceptualized as an acquired global impairment of cognitive function. At earlier stages it is diagnosed on the basis of progressive impairment in two or more areas of cognition (memory, language, visuospatial and perceptual ability, thinking and problem-solving, personality) sufficient to interfere with work, social function, or relationships, in the absence of delirium or major 'non-organic' psychiatric disorders such as depression (section XXX). In the earlier stages, diagnosis may require formal neuropsychological testing. At present the diagnosis is essentially clinical, although functional neuroimaging is showing promise as a diagnostic aid.

The commonest cause of dementia in old age is Alzheimer's disease with cardiovascular causes second. Normal pressure hydrocephalus, classically associated with the triad of urinary incontinence, apraxia of gait, and mild cognitive impairment, is always sought for even though operative ventricular shunting often confers no benefit. Dementia with cortical Lewy bodies is being increasingly recognized. Although commonly described as a subacute disorder it is likely that more chronic forms will become increasingly recognized. Mood disorders, particularly depression, are common in the early stages and visual hallucinations are an early and persistent feature. Parkinsonian symptoms and signs are commonly present but are rarely severe. The importance of recognizing the disease lies in the particular sensitivity of sufferers to the ill-effects of phenothiazine drugs, which should be avoided.

Unfortunately there is little in the way of specific treatment for dementia. Patients with vascular dementia are commonly prescribed aspirin, although there is little evidence at present to support this practice. Patients in atrial fibrillation with stepwise progressive dementia suggestive of cerebral emboli will normally be offered anticoagulation, although careful supervision of dosage and monitoring of the international normalized ratio may be required. The memory defect of Alzheimer's disease is associated with a deficiency of acetylcholine in the brain and drugs which inhibit cholinesterase in the brain are now becoming available. The first of these, tacrine, had too severe a side-effect profile to be clinically useful but its successors, donepezil, galantamine, and rivastigmine, are proving more acceptable. Both drugs bring about a small improvement in cognitive function but, as to be expected, do not retard the continuous decline in function associated with the underlying dementing process. There is evidence that some patients with clinically diagnosed Alzheimer's disease do not respond to these drugs and it is important that if they are tried some formal measurement of cognitive function be applied at baseline and then a decision taken at a 6-week review as to whether the drug should be continued. If cognitive function has declined further the drug is not worthwhile. If there is been improvement or maintenance of function further review should take place at 12 weeks. Side-effects of these drugs are to be expected due to their cholinergic properties and include nausea, stomach cramps, and diarrhoea.

Depression

Depression is probably no more common in old age than at any other time of life but its effects can be more prominent and disabling. The present generation of older people are often unwilling to acknowledge that they may have a mental illness so that the clinical presentation and treatment may be complicated by denial and somatization. The classical features of depression may be present but more subtle manifestations such as a change of personality or behaviour, self-neglect, and asocial behaviour may be the presenting feature. Late-onset alcoholism or other forms of drug abuse, usually of sleeping tablets or pain killers, may also be symptoms. It is also important to recognize depression complicating physical disease; for example, rehabilitation of a stroke patient is often interrupted by the understandable onset of a depressive illness. As a physical sign avoidance of eye contact by an older patient can be very significant even if other aspects of demeanour are not typical of depression. Treatment is along conventional lines and old people respond as well as young to antidepressants. Doses should start low and be increased with care, particularly with drugs such as the tricyclics which because of their anticholinergic properties may induce delirium. The risk of suicide increases steeply with age and is particularly high in older men. Where a suicide is thought to be a high risk urgent psychogeriatric help should be sought and in such circumstances, or in the case of an old person whose health is compromised by withdrawal and refusal to eat, electroconvulsive therapy may be life saving.

Paraphrenia

Although for a long time thought to be a form of late-onset schizophrenia, this syndrome is now suspected more often to have its basis in organic brain damage. It is usually readily recognized as a primary psychiatric illness but may occasionally present to the physician, sometimes in a patient brought up to an accident and

emergency unit because of abnormal behaviour. Chronic undernutrition due to delusional ideas about food being poisoned or unsafe is one of the causes for the geriatric syndrome of 'failure to thrive'. In contrast with earlier-onset schizophrenia, personality is usually well preserved and although the patient may have alarming delusions they rarely become dangerously aggressive. The typical patient is an elderly solitary female, somewhat deaf with prominent semistructured auditory hallucinations. Ideas of being subjected to influence from outside, extraterrestrial aliens, or merely the television set are common and auditory noises attributed to the neighbours may lead to friction. The physician should not necessarily leap to the conclusion that a patient's description of strange noises is necessarily delusional. The old lady diagnosed as paraphrenic before the family of illegal immigrants living in her roof space was discovered is a possibly apocryphal but cautionary tale. Patients with paraphrenia require skilled psychogeriatric care, but can often be supported in their own homes if visiting psychiatric nurses or social workers can establish adequate rapport.

Unusual personalities

Old people display the same range of personalities as seen at any other age, from the delightfully eccentric to the perfectly odious. A syndrome of self-neglect, commonly, though unhelpfully, called the Diogenes syndrome, is well recognized. Characteristically this affects an old person living alone with a good work record and rather rigid personality, who for obscure reasons accumulates enormous piles of rubbish through a pathological inability to throw anything away. Such a patient may come to medical attention through the consequences of self-neglect, but sometimes as a victim of burglary and violence as a consequence of their being assumed by local villains to be the archetypal rich old miser. Complaints from neighbours about rat infestations, fire risks, or odours may also precipitate medical attention.

Other forms of personality disorder may resurface in old age, perhaps when protective spouses die or families and other carers reach the limit of their tolerance. Placement problems can arise since aggressive, abusive, or even merely 'difficult' behaviour can be hard to accommodate where some form of collective living is required.

Clinical pharmacology and the older patient

In epidemiological studies, the incidence of adverse effects of drugs increases with age. It has been suggested that at least 10 per cent of hospital admissions of older people in the United Kingdom are due in whole or part to adverse reactions to drugs. Higher incidences have been reported from general practice. Conventionally, adverse drug reactions are categorized into the idiosyncratic, usually due to host factors such as allergy or genetic susceptibility, and the dose-related, which are undesirably intense effects of the drug's pharmacological actions. In some instances adverse drug reactions are related to both dose and duration of exposure; the adverse effects of long-term high-dose steroids in older people being an example. Most adverse drug reactions affecting older people are dose-related or dose and duration-related. The chief reason for this is that more older people than younger take medications and are also more likely to be taking multiple medications with the risk of interactions. Although patients suffering from cognitive or visual impairment are at risk of making mistakes with their medications, there is no evidence that older people are any worse than younger ones at following treatment advice. Unnecessarily complex drug regimens may cause problems in compliance, and care in prescribing is an important aspect of good-quality medical care for older people. There are some age-associated changes in pharmacokinetics and, probably, pharmacodynamics that increase the risk of adverse drug reactions.

Bioavailability

In general there is no significant change with age in intestinal absorption of drugs that are absorbed, as most are, by passive diffusion. Absorption of substances such as iron, thiamine, calcium, and vitamin B_{12} which undergo active transport across the intestinal mucosa may be lower in older people. Loss of gastric acidity and bacterial overgrowth in the small intestine increase in prevalence with age and may affect drug absorption. The absorption of levodopa increases with age, probably because of reduced dopa decarboxylase activity in the gastric mucosa.

Hepatic metabolism

First-pass metabolism in the liver has an important effect on the bioavailability of drugs. The size of the liver declines with age, as does the density of its blood supply. For drugs such as propanolol and morphine that undergo significant first-pass metabolism, a higher proportion of drug absorbed from the intestine will reach the systemic circulation. Hepatic drug metabolism has been classified into phase 1 (oxidation–reduction) reactions mediated by the mixed-function oxidase system and phase 2 (conjugation) reactions. The clearance of some drugs (chlordiazepoxide and diazepam) metabolized by phase 1 reactions is retarded with age, but conjugation reactions seem unimpaired. Phase 2 reactions are, however, a potential site for important interactions between drugs that share common pathways of elimination. As with all age-associated phenomena, these are generalizations based on averages derived from groups of people of different ages. They reflect in part the increasing prevalence with age of 'frail' people with multiple disease and acquired impairments, and may not apply to individuals who are fit and healthy.

Body composition

Average body composition changes with age. Even though total body weight may not alter, muscle mass and total body water fall, while fat increases. These changes affect volumes of distribution of drugs and also have consequences for drug binding and retention. Water-soluble drugs such as digoxin, gentamicin, theophylline, and cimetidine have reduced volumes of distribution in older patients. Although, in acute single-dose studies, these drugs may produce higher serum levels in older than in younger patients, higher levels will lead to more rapid excretion. There is therefore no consistent effect in steady state conditions. Lipid-soluble drugs such as hypnotics and anaesthetics, diazepam and thiopental for example, have increased volumes of distribution contributing to prolonged serum half-lives in older people.

Renal function

Particular care is required in prescribing drugs that undergo renal elimination. As noted earlier, most individuals show an age-associated decline in glomerular filtration, which may be intensified by illness or medications, and which can retard the elimination of drugs, such as digoxin, that are largely excreted renally. Drugs eliminated by renal tubular secretion, such as penicillins and aminoglycosides, are also affected by the age-associated decline in renal mass and loss of nephrons.

Blood–brain barrier

The blood–brain barrier comprises the mechanical barrier provided by the endothelial cells of the cerebral vasculature with their characteristic tight (non-porous) junctions, and the metabolic barriers provided by the glial cells. It is not clear whether there is any general age-associated change in the efficiency of the blood–brain barrier. An increased susceptibility of older people to the adverse effects of benzodiazepines is well documented, but it is not known whether this is a pharmacodynamic effect at receptor level or where such drugs enter the brain in higher concentration in later life. Clinically, patients with cerebrovascular disease often seem more susceptible than average to drugs acting on the central nervous system, and this should be borne in mind when prescribing, but again it is not known whether this represents increased permeability of the blood–brain barrier or a pharmacodynamic effect.

Protein binding

There is a very small reduction in serum albumin concentration in healthy elderly people which has no clinical significance. The much greater reduction in albumin levels in many elderly people in hospital is due to the effects of disease and subnutrition and is a predictor of poor prognosis. Low serum albumin levels might in theory increase the proportion of unbound and metabolically active drugs that are normally highly protein bound. Interactions due to one drug displacing another from protein binding sites is also an enhanced theoretical possibility. Bound drugs that might be affected are particularly warfarin, tolbutamide, and phenytoin, and common displacing drugs are aspirin and sulphonamides. Other things being equal, any effect should be transient as the unbound drug is also more rapidly eliminated, but with older patients it is wise to be alert to any avoidable possibility of harm.

Pharmacodynamic effects

Age-associated increases in the effects (including adverse effects) of some drugs are not readily explicable in terms of gross pharmacokinetic changes, and pharmacodynamic effects at receptor level have been proposed. Such effects may contribute to the increased susceptibility of older people to benzodiazepines and warfarin, and to the age-associated reduction in sensitivity of b_1 receptors to adrenergic b-blockers. The susceptibility of older people to gastric complications of non-steroidal anti-inflammatory drugs may also have a pharmacodynamic element.

Preventing adverse drug effects

The possibility of an adverse drug reaction needs to be included in every differential diagnosis considered for an older patient. It is important to seek information about over-the-counter and self-prescribed preparations (perhaps 'borrowed' from a spouse or neighbour) as well as prescribed drugs. The use of over-the-counter drugs by older people varies between countries and social classes but is generally increasing. Table 4 lists some of the commoner adverse drug reactions seen in clinical practice. In addition to the effects listed in Table 4, older people often feel non-specifically unwell when taking drugs, especially antibiotics. Another frequent problem arises from oesophageal dysmotility, which is common in later life and can cause temporary delay in the clearance of tablets or capsules from the lower oesophagus into the stomach. This can cause local oesophagitis, which may be misinterpreted as ischaemic heart disease or an indicator of gastro-oesophageal reflux. Although well recognized as an adverse effect of alendronate, virtually any drug can cause the problem, and antibiotic capsules are among the most common. The remedy is for the patient to take tablets or capsules while standing and to wash them down well with a glass of water. Some older patients find that following tablets or capsules with a small piece of bread will stimulate enough oesophageal peristalsis to clear the drugs into the stomach.

Of the drugs causing adverse effects listed in Table 4, diuretics are probably the commonest offenders because they tend to be over-prescribed in general practice, usually for minor stasis oedema in older women. In terms of severity and permanence of damage, steroid-induced osteoporosis is one of the most serious complications of drug therapy for older people. If there is a possibility of steroid therapy becoming prolonged, as when prescribed for polymyalgia rheumatica or giant-cell arteritis, treatment to prevent osteoporosis should be initiated from the time of first prescription. For older patients this will normally consist of oral supplements of vitamin D and calcium with a bisphosphonate.

Table 5 outlines the principles of safer prescribing for older patients. Most of these are self-evident. In choosing a drug for an older patient a common problem, already alluded to, is that all too often there is no evidence on effectiveness specific to older people. Other things being equal, one is wise to choose a class of drug for which there is relevant evidence rather than extrapolate from data on younger patients. In general, with drugs with powerful and potentially dangerous effects, shorter-acting forms are preferable to longer, even though this may complicate dosage regimens. Thus, with oral hypoglycaemic drugs, tolbutamide or gliclazide should be prescribed in preference to chlorpropamide or glibenclamide. Benzodiazepines are best avoided entirely for older patients, but if essential for sleep disturbance only shorter-acting forms should be used and only for short periods. In a range of studies, longer-acting benzodiazepines have been consistently associated with falls. Although allegedly short acting, temazepam has longer effects on psychological function than its plasma half-life would suggest and is best regarded as a longer-acting drug.

Various prescribing practices are aimed at helping patients to avoid mistakes in medication. If possible drugs should be chosen to minimize the number of different times a day that they have to be taken. Combinations of four and three times a day regimes can be particularly troublesome if followed religiously. Patients should be helped by knowing what each of the medications they are taking is intended to do. Although there is little direct evidence in support, it is common geriatric practice to try to offer older patients drugs with distinctive shapes and colours rather than a collection of anonymous white tablets. Such a policy has to be consistent across repeat prescriptions and may call for specific rather than generic prescribing with attendant increases in costs. Patients discharged from specialist geriatric or psychogeriatric departments are often given cards with specimens of each prescribed tablet attached with transparent tape against a description of its purpose and dosage schedule. This can be helpful provided that general practitioners and pharmacists continue to provide the same brands of drug. A more reliable approach is to make use of one of the various forms of box or packet in which tablets can be sorted, by pharmacist or carer, into separate compartments labelled by day and time. These can be helpful to patients and also to carers who need to check whether drugs have been taken or not.

Arrangements should be made for regular and frequent review of medications given to older people, not least because dosage adjustments are often required with longer-term therapy. Even more important, no drug should be prescribed without thought to when it should be stopped, and unintended continuation of treatments is one source of unnecessary morbidity in old people. Pill counts are helpful in detecting inadvertent non-compliance with therapy. Surveillance of drug therapy for older patients is also important if adverse effects are to be detected promptly. It may be constructive to involve carers in the surveillance process. While it is clearly wise practice to warn patients and carers about possible adverse effects, comprehensive warnings of the type included in packet inserts can frighten people into not taking the drug at all.

General approaches to medical care for older people

Conditions such as stroke, cancer, heart disease, and dementia increase in incidence and prevalence with age, but old people do not suffer from any diseases that never afflict younger adults. It is in the treatment of the patient rather than of the disease that medical care for older people has to provide particular emphasis and sensitivity. Central is the loss of adaptability characteristic of age. Older patients may have little physiological reserve to cope with even minor shortcomings in care. Loss of adaptability may need to be compensated for in the design of medical services. More frequent checking of the international normalized ratio of older patients taking anticoagulants, and readier deployment of invasive monitoring for older patients after trauma or at risk of cardiovascular instability are examples.

Table 6 sets out some of the characteristics of illness in later life of which health workers must be aware and to which health services should be ready to respond. All of these characteristics are directly or indirectly aspects of the loss of adaptability that is the fundamental property of ageing. The first four underlie a need for rapid access to high-quality diagnostic and treatment facilities when an older person falls ill. Because of the frequent conjunction of multiple diseases with non-specific presentation, older people often need more investigations than do younger patients to establish an accurate diagnosis. An elderly patient with pneumonia may present with delirium or falls before any localizing signs appear in the chest. In infections, fever may appear late and may be missed if core temperature is not accurately measured or the patient is not given time to recover from a cold ambulance drive. Visceral pain from the peritoneum in acute appendicitis or intestinal perforation, or from the heart in myocardial infarction, may be reduced or absent in older patients. The slowness of the aged body in mounting its defences can lead to rapid deterioration unless an accurate diagnosis is made and correct treatment instituted urgently.

Reduced adaptability in old age also leads to a high incidence of secondary complications both of disease and treatment with consequent need for careful surveillance by medical and nursing staff. Good-quality medical and nursing care depends on scrupulous attention to detail, since even the smallest error of judgement or lack of observation may have serious consequences for the patient. A typical example lies with the development of pressure sores. A young person can usually lie immobile for 4 h without suffering serious pressure sores. An older person can develop sores in half that time. Pressure sores are most likely to develop if a patient is lying on a hard surface; particular hazards include time spent lying on a hospital trolley in a casualty department or on an operating table while junior orthopaedic surgeons spend an age doing their first hip replacement. High-quality care for an older population puts demands on the managers as well as the practitioners in health services.

Specialist geriatric care

The last two items in Table 6 underlie the need for specialist geriatric rehabilitation teams to be closely linked to the working of acute medical and surgical services. Most older patients, approximately 70 per cent of those referred to British hospitals as medical emergencies, have fairly straightforward illnesses such as pneumonia, myocardial infarction, or deep venous thrombosis. Treatment can be along normal lines, and patients discharged directly back home. The average length of stay will need to be longer than for younger patients because older people need longer to recover full function after a debilitating illness, and they are more likely to be living alone with no one to support them during convalescence. A proportion of older people admitted to medical wards, typically 10 to 15 per cent, have complex illnesses and functional problems that call for the multiprofessional approach of a specialist geriatric service if best outcomes are to be attained. Such patients need to be identified early on in their illness, and the ideal way of ensuring this is for physicians with special responsibility for older people to be part of the clinical team on acute medical wards. This approach ensures that the majority of older people, whose chief need is unimpeded access to the skills of other specialties, are not disadvantaged as they may be if admitted initially to a purely geriatrics service.

The approach to an elderly patient

Table 7 outlines the four stages that should structure the approach to elderly patients, especially those with complex problems.

Assessment

Assessment may require contributions from all members of the core geriatric team—doctor, nurse, occupational therapist, physiotherapist, and social worker. In addition to dealing with technical matters, this stage should also be seen by the team as an opportunity for 'getting to know' the patient as an individual and for building mutual trust and friendship. Functional assessment is best documented in terms of performance as measured using standard scales agreed both among the

geriatric team and with the relevant social services of community teams who will care for the patient after discharge. A wide range of scales is available, but one of the commonest in use for activities of daily living is the Barthel scale (Fig. 4). This is robust and reliable if the rules of administration are agreed and followed. It is useful as an indicator of needs for help but is primarily applicable to patients with moderate to severe disabilities. At higher levels of performance, scales for assessment of instrumental activities of daily living such as ability to use the telephone or travel on public transport may be more relevant.

Fig. 4 The Barthel activities of daily living scale with instructions for administration. (Formatted by courtesy of Dr S. J. Winner.)

Assessment of mental function usually calls for a global performance scale and may also require assessment of mood. The abbreviated mental test score (Table 8) is widely used in British hospitals and is easily applied but was developed for use in geriatric and psychogeriatric inpatient units; a patient scoring at the significantly low score of 7 out of 10 is quite severely impaired. Milder degrees of impairment are better detected by the Folstein mini-mental status examination. Neither the mental test score nor the mini-mental status examination is a diagnostic instrument, and reduced scores may reflect any cause of impaired cognitive function including dementia, depression, or delirium. If depression is suspected, a screening questionnaire such as the geriatric depression scale may be helpful but should not be regarded as an adequate substitute for a skilled and sensitive clinical assessment.

Understanding a patient's cultural and educational background is an important stage in ensuring good communication and information needs. Protection of patients' autonomy is imperative but forcing patients to make worrying decisions they do not understand is to mistake the form for the substance. The most important resource available to an older patient is often his or her family who must be appropriately involved in planning care. 'Appropriately' implies taking their needs and priorities into account but not letting them over-ride those of the patient.

Setting the objectives of care

The second stage of setting objectives of care is essentially a dialogue and negotiation between the doctor who knows what could effectively be done, and the patient who decides what should be done. This should be an essential stage of care for any patient but is often omitted with younger patients for whom it is usually assumed, not always appropriately, that prolongation of life is the only objective to be considered. Older people may have other priorities; they may value dignity and independence more than life. If alone in the world they may have the privilege of being able to please themselves without having to worry about the effects their decision may have on others. It is at this stage of care that issues of quality of life are most important. Standard questionnaires are of limited worth because they assume that everyone shares the same system of values. In identifying what a person enjoys in life, and the effects that different treatment options may have on a particular patient's quality of life, more individualized methods are required. Psychometrically based questionnaires responsive to the value systems of individual patients are under development, and interactive computer programs can help with specific issues such as choice of treatment of prostatic hypertrophy, but for older patients with complex problems it has yet to be shown that anything is more reliable than compassionate human interchange. Once objectives have been agreed they need to be specified clearly in the notes and, subject to the patient's agreement, explained to relevant family members and future carers.

The management plan

Formulating a management plan calls for appropriate and timely treatment for any relevant acute diseases together with rehabilitation of the patient for resettlement in the future abode of choice, usually his or her own home. In order to achieve this the management plan must provide for bridging any 'ecological gap' between what patients can do and what their homes are going to demand. The gap is closed by improving the patient through therapeutic interventions and reducing the demands of the environment by prosthetic interventions. Therapeutic interventions often involve the whole geriatrics team. A hemiplegic stroke patient, for example, might be helped by the doctor improving exercise tolerance through tighter management of heart failure, the physiotherapist improving walking, the occupational therapist teaching him or her how to dress, the nurse devising the best programme for the diuretic therapy and ensuring proper nutrition, and the social worker steadying morale and sense of security about future support in the community. At the same time that the therapeutic programme is under way plans should be going forward for any prosthetic changes to the patient's home that are likely to be needed. These may include aids and adaptations such as extra banisters on stairs, grab rails in the bathroom, a raised toilet seat, fitting an alarm, as well as providing for personal help for shopping, bathing, supervision of medication, or other needs. In difficult situations rehousing, or, in the worst case, the ultimate prosthetic environment of institutional care might have to be arranged. It is important that the therapeutic and prosthetic programmes are managed in step since it can be extremely demoralizing for a patient who is physically and psychologically ready to go home to have to wait because aids and services are not yet in place.

Home visits with the patient prior to discharge can be very helpful both in identifying what prosthetic input is necessary and also in reassuring an apprehensive patient and possibly anxious relatives that a return home is feasible. The team must pay careful attention to the needs of relatives and others involved in the patient's care. Those who will make a personal contribution may need to spend time in the rehabilitation ward under the tutelage of nurses or therapists learning any necessary skills in helping the patient both physically and psychologically. An important element in this may be to prevent carers from being too helpful, as that can lead later to the patient becoming unnecessarily dependent. A less welcome aspect of the rehabilitation team's dealing with relatives is in relation to objections, often from those who will be contributing least to a patient's later care, that attempts at return into the community are impractical and 'she would be much better off in a nursing home'. This situation is often more complex than it seems in that it may reflect family tensions arising from half a century of unspoken animosities. The problem may also arise because relatives think that responsibility for the success or failure of the discharge will fall upon them, or feel guilty that they will not be able to contribute, or fear criticism for not doing so. The first step in negotiations is therefore always to try to find out what social and psychological crosscurrents are running—an activity in which all members of the team need to acquire skill, although the social worker is the most specifically trained. Discussion with a primary care physician who has known the patient and family for years can often prove illuminating. Tact and avoidance of confrontation are called for, followed sometimes by delicate diplomacy in negotiating between differing factions in a patient's family. It may surprise some to learn that this aspect of the work of specialist geriatrics departments, far from being seen as an impediment to rational medicine, is actually a source of great professional satisfaction when done well.

A follow-up visit 10 days or so after discharge is also helpful, both to make sure that all is well and for the rehabilitation team to savour the satisfaction of seeing their patient happily back in his or her natural habitat or to learn from any insufficiencies.

Regular review of a patient's progress is usually formalized in a weekly multiprofessional meeting. The meeting is chaired by the clinically responsible doctor, but decisions are collective and must be formally recorded in the patient's notes. In addition to a general review of progress towards treatment goals, an important task for the meeting is to recognize when progress is not as had been expected. This may mean that the original plan was inappropriate, either in setting unrealistic goals or in specifying a less than optimum treatment programme. More often the problem arises because some new factor has intervened to interrupt the patient's progress. This may range from medical problems such as a depressive illness (a particularly common sequela in stroke), a drug side-effect, or urinary tract infection, to psychosocial factors such as discouragement from something a visitor has said, or some adverse change in the ward environment. The new factor should be dealt with or the management plan or treatment goals modified as appropriate. Relevant family members and carers, as well as the patient, should be informed if significant changes to aims and means of rehabilitation are made.

The future

Populations are ageing throughout the world and for more than a decade the majority of the world's population of people aged over 65 have been living in developing

rather than developed countries. The older patient is now the norm, and the practice of medicine and the design of services must continue to adapt. The demand for medical care will rise as the numbers of older people increase and new medical technologies become increasingly appropriate for frailer patients. These changes must be planned for, but in so far as they represent the fact that more people are living longer and more actively, they should be welcomed as one aspect of the progress of civilization. Emphasis in health care needs to shift from the prolongation of life at all costs to the prevention and cure of disability. This needs to apply at all ages. Age-based rationing of health care is not compatible with the democratic ideal, nor would it make economic sense as it would merely substitute the long-term costs of disablement for the short-term price of treatment. But in association with the development of rational heath care systems there needs to be a public health strategy for the lifelong primary prevention of disability. Because of the age-associated loss of adaptability, the later in life a potentially disabling disease strikes the more likely the patient is to die from it rather than linger on in disability. Postponement of disease is prevention of disability. Our collective aim is to spend a long time living and a short time dying.

Further reading

Browne JP *et al.* (1994) Individual quality of life in the healthy elderly. *Quality of Life Research* **3**, 235–44.

Dickerson JEC *et al.* (1999). Optimisation of antihypertensive treatment by crossover rotation of four major classes. *The Lancet* **353**, 2008–13.

Fiatarone MA *et al.* (1994). Exercise training and nutritional supplementation for physical frailty in very elderly people. *New England Journal of Medicine* **330**, 1769–75.

Grimley Evans J *et al.*, eds (2000). *The Oxford textbook of geriatric medicine*, 2nd edn. Oxford University Press, Oxford.

Holliday R (1995). *Understanding ageing* Cambridge University Press, Cambridge.

Manton KG, Gu X (2001). Changes in the prevalence of chronic disability in the United States black and nonblack population above age 65 from 1982 to 1999. *Proceedings of the National Academy of Sciences of the USA* **98**, 6354–9.

Phillips PA *et al.* (1984). Reduced thirst after water deprivation in healthy elderly men. *New England Journal of Medicine* **311**, 753–9.

Salim A *et al.* (1998). Subcutaneous hydration in the elderly. In: Armand MJ *et al*, eds. *Hydration and aging* pp. 201–8. Springer Publishing Company, New York.

The Heart Outcomes Prevention Evaluation Study Investigators (2000). Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *New England Journal of Medicine* **342**, 145–53.

30.2 Mental disorders of old age

Robin Jacoby

Mental disorders of old age are of the greatest importance because:

- demographic trends in **all** countries are leading to a marked increase in the number of elderly persons, especially the very old;
- this age group is particularly prone to debilitating mental illness;
- presentations of mental illness in the elderly may differ from those at younger ages and go unrecognized;
- the pattern of morbidity differs; the elderly suffering more from organic disorder—delirium and dementia;
- mental and physical illness often occur together, the one sometimes masking the other.

Elderly inpatients on non-psychiatric wards show high rates of mental disorder of various sorts. Community rates are also high. For instance, the prevalence of dementia is about 7 per cent over 65 years of age, but approaches 20 per cent in those over 80. For major depressive illness defined by ICD-10 or DSM-IV the prevalence is no more than 3 per cent, but for pervasive depressive symptoms, which most psychiatrists would consider in need of treatment, it is around 12 per cent. There are no valid data for the prevalence of very-late-onset schizophrenia-like psychosis. The closure of large mental institutions in most developed countries has resulted in a greatly increased number of elderly patients in the community suffering from chronic schizophrenia that began much earlier in their lives.

Assessment

Factors to be considered are:

- the setting in which the patient is examined;
- the patient's ability to provide information.

Assessment in the patient's own home gives invaluable clues about premorbid adjustment, activities of daily living, and even causes of an abnormal mental state—for example, a waste bin full of empty whisky bottles. Patients in general hospital wards should be assessed in a side-room if possible. Many elderly patients are unable to give a history. It is therefore essential to find reliable informants, even if this means disturbing neighbours or making long-distance telephone calls. Doctors visiting patients outside hospital should carry equipment for physical examination. A low-reading thermometer, sphygmomanometer, and sugar-detection urine sticks may also prove invaluable at home in detecting the cause of mental disturbance.

Examination of the mental state of an elderly person is essentially the same as for any adult, but with differences of emphasis. More attention is paid to the level of consciousness, particularly subtle fluctuations throughout 24 h. For patients with suspected dementia a full cognitive examination is undertaken, frequently in several brief sessions because patients become easily fatigued and unable to co-operate. Tired or inattentive patients may fail a test that they might otherwise have been able to perform satisfactorily. Systematic evaluation of all higher cerebral functions, such as memory, praxis, and language, is essential. On general hospital wards or home visits the routine administration of standardized questionnaires is of value as an alerting mechanism, but not a substitute for full evaluation. The Mini-Mental State Examination (**MMSE**) is the preferred questionnaire because it assesses a range of higher cerebral functions, not just memory and orientation.

Because the capacity to live independently is frequently compromised in elderly people and becomes the crucial determinant of social outcome, it is essential to assess the patient's ability to perform activities of daily living (**ADL**)—dressing, feeding, toilet care, and so on. Impairment in ADL may reveal dyspraxia or agnosia undisclosed by formal clinical tests. The informant's account is important, as patients' assessments of their own abilities may be misleading.

Delirium

The clinical features of delirium are described elsewhere. The elderly are particularly susceptible because:

- They more commonly suffer from the physical disorders that cause delirium: hypoxia, infections, toxicity (especially from drugs), and metabolic and CNS disorders.
- Many have a decreased cerebral reserve because of incipient or overt dementia.

An underlying physical cause for delirium must therefore be sought assiduously, as it is generally more important to treat this than the mental symptoms. Physical illnesses causing delirium can be relatively minor: for example, a urinary tract infection, especially in patients with dementia. Because the elderly generally suffer high physical morbidity they tend to receive more prescribed medication. Single or multiple drugs, as well as drug withdrawal effects, are potent causes of delirium. Management is essentially that of the underlying cause and psychotropic drugs should be avoided whenever possible. However, short-term treatment of mental symptoms and behavioural disturbance with a small dose of a neuroleptic drug is sometimes unavoidable to allow medical treatment to proceed unhindered.

Dementia

The importance of dementia lies not only in the suffering it causes to patients but also in the demands it makes upon family caregivers and the medical and social services. The clinical features of the dementia syndrome are described in Chapter 24.13.8. Global impairment is required for diagnosis, and memory loss alone, of which elderly people invariably complain, is **not** a sufficient criterion. Accurate diagnosis is required because:

- treatable or arrestable causes may be discovered;
- management can be planned and implemented;
- caregivers are helped by a clearer understanding of the process affecting the patient and its likely course;
- knowledge of the underlying conditions is accrued and advanced.
- treatment with a central cholinesterase inhibitor may be indicated.

Rational management requires an appreciation of the practical implications of cognitive impairment: for example, ensuring nutrition in a patient with dyspraxic inability to feed herself. However, the non-cognitive, behavioural, and psychiatric disturbances in dementia are the manifestations which cause most problems to caregivers and often require management by specialist services. Patients should be maintained at home for as long as possible, both for humane reasons and because they will have lost the capacity to adapt easily to a new environment. This is best achieved through a continuing partnership between clinicians, family caregivers, and local statutory and voluntary social services. Such interagency co-operation and discussion facilitates access to community services, which include those brought to the patient—meals, nursing, bathing, house cleaning—and day care away from home. Intercurrent illness should be treated along with other medical problems, such as incontinence, which is due more commonly to urinary tract infections and faecal impaction, respectively, than to the underlying disease processes of dementia. Attention must be given to principal caregivers (most frequently spouses or adult daughters, but sometimes unrelated neighbours), who show high levels of psychiatric morbidity themselves due to their burden of care.

Affective illness

Depression

Depressive illness in the elderly differs only in emphasis from that in earlier life, varying from mild dysthymia to major psychosis with high risk of suicide. Psychosocial factors are as important, but genetic loading is usually lower in patients with late-onset depression. Cerebral organic change is more common in late-onset cases.

Psychomotor retardation is frequent, as in younger patients, but anxiety and agitation are also characteristic. Delusions of guilt and unworthiness are typical, but hypochondriacal or nihilistic beliefs also often occur. Histrionic or other bizarre behaviour, which is out of premorbid character, is also seen. Some patients with severe depression perform badly on cognitive tests, so called 'depressive pseudodementia', which may lead to a wrong diagnosis of organic dementia,. Hypochondriacal delusions together with anorexic weight loss can be mistaken for physical illness, such as cancer. Some patients may even deny low mood ('masked depression'), but the presence of other typical affective manifestations and a favourable response to antidepressant treatment reveal the correct diagnosis. The clinician should not avoid treating depression in patients, such as those with severe physical illnesses, simply because they appear to have a valid reason to be unhappy.

The elderly are vulnerable to the unwanted effects of psychotropic drugs but respond well to standard antidepressants. Selective serotonin-reuptake inhibitors (**SSRIs**) are usually preferred as first-choice drugs because of their safety and fewer side-effects. However, tricyclics, such as amitriptyline and imipramine, are highly effective, especially in major depression, if care is taken to observe the general principle of small initial doses increasing slowly to lower therapeutic doses than are given to younger patients. The elderly also respond well to electroconvulsive therapy (**ECT**), regarded by many as the treatment of choice for deluded and/or suicidal patients. Extreme age, dementia, and physical infirmity are **not** contraindications to electroconvulsive therapy. After recovery from an episode of major depression, antidepressant treatment should be continued for at least 2 years, and probably indefinitely.

Mania/hypomania

Manic illness, which accounts for about 5 per cent of psychogeriatric admissions, is usually less florid than in younger patients. Mixed manic and depressive pictures have been described as typical, but are probably not more common in old age than before. Cerebral organic disease is found in a high proportion of cases. Diagnosis is sometimes difficult to make because patients may present with a rather non-specific psychosis, or one mimicking delirium. The clinician should therefore enquire about a previous history of affective illness. Secondary mania, defined as a first onset in close temporal relationship with a physical illness or drug treatment in patients with no previous history of affective disorder, is also seen. Patients respond well to treatment with neuroleptics and lithium carbonate in the acute phase. Lithium is also widely used to prevent relapse. Here, frequent monitoring of renal and thyroid function is essential, and the serum lithium level should be maintained at the lower end of the therapeutic range, that is to say around 0.6 mmol/l.

Paranoid disorders

Paranoid ideas, usually but not invariably of persecution, are common in dementia and affective disorder. In dementia they tend to be transient, variable, and unsystematized. In affective disorder they are usually mood-congruent, for example persecution deserved because of guilt. Persecutory ideas are also seen in paranoid personalities and acute paranoid reactions, but very-late-onset schizophrenia-like psychosis * is the main paranoid syndrome of old age, equivalent to, but distinct from, paranoid schizophrenia of earlier adult life. The following aetiological factors have been consistently reported:

- a high female to male ratio (up to 7:l);
- a family history of schizophrenia intermediate between the general population and younger schizophrenic patients;
- social isolation and poor premorbid interpersonal relationships;
- low marriage and fecundity rates compared with age-matched peers;
- long-standing deafness.

The clinical picture is sometimes indistinguishable from early-onset paranoid schizophrenia, but the patient usually presents with a few simple delusions and associated auditory hallucinations. For example, she may complain of hearing the neighbours plotting against her, or impugning her sexual virtue. Medication with atypical antipsychotic drugs, such as risperidone or olanzapine, to minimize the risk of extrapyramidal side-effects, is the treatment of choice. However, neuroleptics rarely extinguish delusional beliefs that become 'encapsulated', meaning set aside or compartmentalized and not permitted to intrude into so many aspects of mental and daily life, as happens during the acute phase of illness. Adherence to treatment is often poor but can be improved with the help of a community psychiatric nurse.

Neurotic, personality, and other disorders

These are too complex to discuss briefly and the reader is referred to specialist texts. However, some general principles can be stated. It is essential to differentiate lifelong neurosis from neurotic symptoms of first onset in old age because the latter, predominantly anxiety and phobia, are most likely to indicate an underlying depressive illness. Elderly patients with lifelong neurosis have frequently adapted to their disability in spite of continued suffering. Mild and moderate lifelong personality deviations are common, but severe disorders are occasionally encountered, such as the senile squalor (Diogenes) syndrome, which can require compulsory removal from home under mental or public health legislation. Illicit drug taking is rare among the present generation of elderly, but the abuse of prescribed drugs, notably benzodiazepines, is common. Alcohol abuse is increasingly recognized as a cause of mental and social disability in elderly people.

---

*Recent international agreement states that cases with onset between 40 and 59 years of age should be termed 'late-onset schizophrenia'; whereas cases from 60 years onwards should be called 'very-late-onset schizophrenia-like psychosis'.

Further reading

Folstein MF, Folstein SE, McHugh PR (1975). 'Mini Mental State'. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* **12**, 189–98.

Jacoby R, Oppenheimer C, eds (2002). *Psychiatry in the elderly*, 3rd edn. Oxford University Press, Oxford.

31 Palliative care

Robert Twycross and Mary Miller

In Western countries, about 25 per cent of the population die of cancer. Even more die of progressive non-malignant disease. In this chapter the focus is on symptom management in patients with far-advanced cancer.

Palliative care is far more than symptom relief. It addresses physical, psychological, social, and spiritual aspects of suffering, thereby helping patients to come to terms with their impending death as constructively as they can while living as actively and creatively as possible. Palliative care also provides a parallel support system to help families cope during the patient's illness and in bereavement; it is best provided by a multiprofessional team.

General principles of symptom management ('EEMMA')

The principles underlying management are:

- *evaluation*: diagnosis of each symptom before treatment;
- *explanation*: explanation to the patient before treatment;
- *management*: individualized treatment;
- *monitoring*: continuing review of the impact of treatment;
- *attention to detail*: no unwarranted assumptions.

Evaluation is based on probability and pattern recognition. For example, hiccup in advanced cancer is mostly associated with gastric stasis or distension, and the most common cause of pruritus is dry skin. Symptoms are not always caused by the disease itself but by treatment, debility, or a concurrent second disorder. Some symptoms are caused by multiple factors. **Explanation** by the doctor of the cause(s) of a symptom does much to reduce its psychological impact on the sufferer.

Management falls into three categories:

- correct the correctable
- non-drug measures
- drugs.

By adopting a multimodality approach, although the underlying disease cannot be cured, it is generally possible to obtain significant, and sometimes complete, relief.

Drugs for a persistent symptom should be prescribed regularly on a prophylactic basis. The use of drugs 'as needed' instead of regularly is the cause of much needless distress. Although many symptoms respond to a combination of non-drug and drug measures, often the main part of the management of symptoms such as anorexia, weakness, and fatigue is helping the patient (and family) accept the irreversible physical limitations of terminal disease.

Patients vary, and it is not always possible to predict the optimum dose of opioids, laxatives, and psychotropic drugs. Adverse effects may also jeopardize patient compliance. **Monitoring** is crucial, with adjustments made as necessary.

Attention to detail is important at every stage of symptom management. It is equally important in relation to the non-physical aspects of care—all symptoms are exacerbated by anxiety and fear.

Pain

At diagnosis, between 20 and 50 per cent of patients with cancer have pain. Prevalence varies according to the primary site of the cancer and metastatic spread. In advanced cancer, 75 per cent of patients have pain and two-thirds of these have multiple pains.

Evaluation

Each pain described by the patient should be recorded on a body chart with a comment about the probable cause (Fig. 1). In addition, pain may be classified as:

- nociceptive, i.e. pain caused by physical and/or chemical stimulation of free nerve endings;
- neuropathic, i.e. pain caused by compression or injury to the peripheral or central nervous systems.

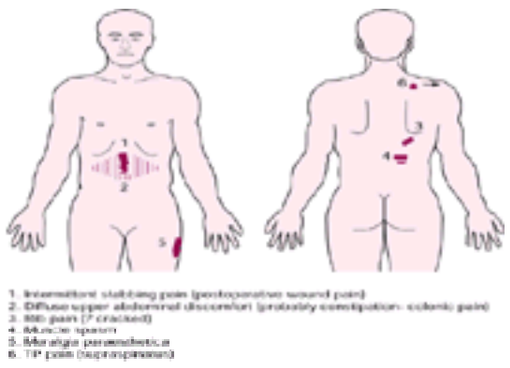

Fig. 1 Pain chart of a 63-year-old woman with cancer of the pancreas, 10 days postoperatively. TP, myofascial trigger point.

Nerve compression pain, like nociceptive pain, is aching in character. On the other hand, nerve injury pain tends to be superficial and burning, and associated with allodynia (light touch caused pain). There may also be spontaneous stabbing pain with or without an underlying aching component. Peripheral neuropathic pain is neurodermatomal in distribution, not local to the lesion, and there may be associated numbness.

A history of analgesic use is part of the evaluation. Not all pains respond equally to analgesics, and information about the benefits of specific medication (or lack of it) may help to identify the underlying mechanism. A detailed pain history coupled with physical examination is generally sufficient to guide management. Further radiological investigations are necessary in only a minority of patients.

Management

- Although analgesics are the mainstay of management, it is important not to overlook correctable causes of pain or ignore other treatment modalities (Box 1). Analgesics can be divided into three classes: non-opioid (antipyretic), opioid, and adjuvant. Their use is governed by the following maxims:
- *By the mouth*—the oral route is the preferred route, including for morphine and other strong opioids.
- *By the clock*—analgesics should be given regularly and prophylactically for persistent pain; 'as needed' medication is irrational and inhumane.
- *By the ladder*—use the 3-step analgesic ladder (Fig. 2).
- *Individualized treatment*—analgesic combinations and doses should be determined individually for each patient.
- *Use adjuvant medication*—meaning adjuvant analgesics, antidotes for adverse effects (e.g. laxatives, antiemetics), and psychotropic medication.

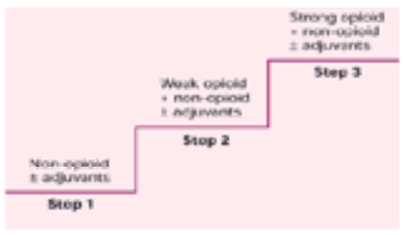

Fig. 2 The World Health Organization analgesic ladder for cancer pain management.

Box 1 Pain management in cancer

Modification of the pathological process	**Psychological**	
Radiation therapy	Relaxation	
Hormone therapy	Cognitive–behavioural therapy	
Chemotherapy	Psychodynamic therapy	
Surgery		
	Interruption of pain pathways	
Analgesics	Local anaesthesia	
Non-opioid (antipyretic)		lidocaine (lignocaine)
Opioid		bupivacaine
Adjuvant	Neurolysis	
	corticosteroids	chemical (e.g. alcohol, phenol)
	antidepressants	cold (cryotherapy)
	antiepileptics	heat (thermocoagulation)
	muscle relaxants	Neurosurgery
	antispasmodics	cervical cordotomy
Non-drug methods	**Modification of way of life and environment**	
Physical	Avoid pain-precipitating activities	
	massage	Immobilization of the painful part
	heat	cervical collar
	transcutaneous electrical nerve	surgical corset
	stimulation (TENS)	slings
		orthopaedic surgery

Walking aid
Wheelchair
Hoist

Non-opioids

The non-opioid (antipyretic) analgesics comprise paracetamol/acetaminophen and the non-steroidal anti-inflammatory drugs (**NSAIDs**). Paracetamol can be taken by two-thirds of patients who are hypersensitive to aspirin. NSAIDs and paracetamol can be used together with an additive effect. The main drawback with paracetamol is the frequency of administration, generally every 4 to 6 h.

NSAIDs inhibit cyclo-oxygenase (**COX**), an important enzyme in the arachidonic acid cascade which results in the production of tissue and inflammatory prostaglandins. COX exists in two forms; COX-1 is present in all normal tissues, whereas COX-2 is more limited in its distribution but massively induced by inflammation (Fig. 3). Recently introduced COX-1 sparing NSAIDs and specific COX-2 inhibitors cause less gastric toxicity than most dual COX inhibitors.

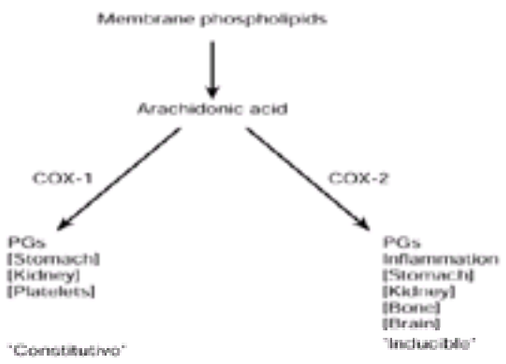

Fig. 3 The different effects of COX-1 and COX-2.

NSAIDs are of particular benefit for pains associated with inflammation, for example soft tissue infiltration and bone metastases. NSAIDs differ in their effect on platelet function. In patients undergoing chemotherapy or with thrombocytopenia from another cause, it is best to use an NSAID which has no effect on platelet function (Table 1).

Weak opioids

There is little to choose between codeine, dextropropoxyphene, and dihydrocodeine in terms of efficacy. Pentazocine is not recommended; it acts for only 2 to 3 h and often causes psychotomimetic effects (hallucinations, feelings of unreality, dysphoria). Generally, a weak opioid should be added to a non-opioid, not substituted for one. If a weak opioid is inadequate when given regularly in an optimal dose, change to morphine (or an alternative strong opioid); do not switch from a weak opioid to a weak opioid.

Tramadol forms a bridge between the classic weak and the classic strong opioids. By injection, it is one-tenth as potent as morphine. By mouth, it is about one-fifth as potent as morphine because of its high oral bioavailability. Tramadol has a dual mechanism of action, partly via opioid receptors and partly by blocking the presynaptic reuptake of 5-hydroxytryptamine (5-HT, serotonin) and noradrenaline (norepinephrine), similar to a tricyclic antidepressant. Tramadol causes less constipation than codeine and morphine.

Strong opioids

The use of strong opioids is dictated by therapeutic need, not by brevity of prognosis.

Morphine is the strong opioid of choice for treating cancer pain (Box 2), and is generally given with an NSAID (and/or paracetamol). Morphine is available as tablets (for example, 10 mg, 20 mg) or in a solution (for example, 2 mg and 20 mg in 1 ml). Modified-release preparations are mostly administered twice a day, some once a day.

Box 2 Starting patients on oral morphine

- Morphine is indicated in patients with pain that does not respond to the optimized combined use of a non-opioids and a weak opioid.
- The starting dose of morphine is calculated to give a greater analgesic effect than the medication already in use:
- If the patient was previously receiving a weak opioid, give 10 mg every 4 h or modified-release 20–30 mg every l2 h.
- If changing from another strong opioid, a much higher dose of morphine may be needed (Table 2).
- If the patient is frail and elderly, a lower dose (e.g. 5 mg every 4 h) helps to reduce initial drowsiness, confusion, and unsteadiness.
- Because of cumulation of an active metabolite, a lower and/or less frequent regular dose may be preferable in renal failure, e.g. 5–10 mg every 6 h.
- If the patient takes two or more 'as needed' doses in 24 h, the regular dose should be increased by 30 to 50 per cent every 2 to 3 days.
- Upward titration of the dose of morphine stops when the pain is relieved or intolerable undesirable effects supervene. In the latter case, it is generally necessary to consider alternative measures. The aim is to have the patient free of pain and mentally alert.
- *Modified-release morphine may not be absorbed satisfactorily in patients troubled by frequent vomiting or those with diarrhoea or an ileostomy. M/r morphine should be used with caution if there is evidence of renal failure.*
- **Scheme 1: ordinary (normal-release) morphine tablets or solution**
- morphine given 4-hourly regularly 'by the clock' with 'as needed' doses of equal amounts up to 1-hourly;
- after 1–2 days, adjust the dose upwards if the patient still has pain or is using two or more 'as needed' doses per day;
- continue with 4-hourly doses regularly with 'as needed' doses of equal amounts up to 1-hourly;
- increase the regular dose by 30 to 50 per cent every 2 to 3 days until there is adequate relief throughout each 4-h period;
- *a double dose at bedtime obviates the need to wake the patient for a 4-hourly dose during the night* .
- **Scheme 2: ordinary (normal-release) morphine and modified-release morphine**
- begin as for Scheme 1;
- when the 4-hourly dose is stable, replace with modified-release morphine every l2 h, or once daily if a 24 h preparation is prescribed
- each l2-hourly dose will be *three times* the previous 4-hourly dose; a once-daily dose will be *six times* the previous 4-hourly dose, rounded to a convenient number of tablets;
- continue to provide ordinary morphine solution or tablets for 'as needed' use. Give the equivalent of a 4-hourly dose, i.e. 1/6 of the total daily dose.
- **Scheme 3: modified-release morphine and ordinary (normal-release) morphine**
- starting dose generally modified-release morphine 20-30 mg 12-hourly or 40–60 mg once daily;
- use ordinary morphine tablets or solution for 'as needed' medication; give about 1/6 of the total daily dose;
- increase the dose of modified-release morphine by 30 to 50 per cent every 2–3 days until there is adequate relief around the clock.
- Supply an antiemetic in case the patient becomes nauseated, e.g. haloperidol 1.5 mg to be taken immediately and then regularly at bedtime.
- Prescribe laxatives, e.g. co-danthrusate or senna ± docusate; adjust the dose as necessary.
- Suppositories and enemas continue to be necessary in about one-third of patients.
- *Constipation may be more difficult to manage than the pain.*
- Warn all patients about the possibility of initial drowsiness.
- If swallowing is difficult or there is persistent vomiting, morphine may be given per rectum by
- suppository; the dose is the same as by mouth. Alternatively give half the oral dose by injection,
- or one-third as diamorphine, preferably by subcutaneous infusion.
- For outpatients, write out the drug regimen in detail with times, names of drugs and amount to
- be taken. Arrange for follow-up.

Morphine rarely causes respiratory depression in patients with pain because pain is a physiological antagonist to the central depressant effects of morphine. However, morphine is potentially dangerous in renal failure. Its main metabolites are morphine-3-glucuronide (**M3G**) and morphine-6-glucuronide (**M6G**). Both cumulate in renal failure but, whereas M3G is inactive, M6G is more potent than morphine. In this circumstance, there is danger of sedation and respiratory depression if the dose is not closely monitored and, if necessary, reduced both in quantity and frequency.

Psychological dependence (addiction) does not occur when morphine is used appropriately as an analgesic. Physical dependence—a phenomenon seen with many psychoactive drugs when taken for prolonged periods (months rather than weeks)—does not prevent a step-by-step reduction in the dose of morphine if the pain ameliorates, for example as a result of radiotherapy or a nerve block. Tolerance to morphine is not a practical problem.

Other modes of administration are necessary at times (Fig. 4). Transdermal fentanyl patches, available in several sizes (25, 50, 75, 100µg/h over 3 days), are an alternative option for patients on a stable dose of morphine, and for those with a tablet phobia or dysphagia. Patients who have not been taking morphine should always start on the lowest dose. Fentanyl is less constipating than morphine. However, acquisition costs are much higher.

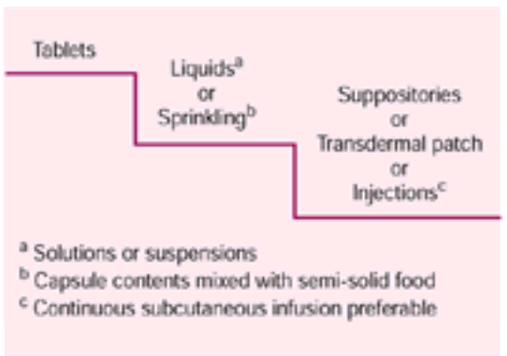

Fig. 4 Alternatives routes of administration.

Injections are indicated when there is:

- persistent nausea and vomiting;
- severe dysphagia (alternatively use transdermal fentanyl);
- extreme weakness;
- a diminished level of consciousness.

In the United Kingdom, diamorphine hydrochloride (diacetylmorphine) is used when injections are necessary. It is more soluble than morphine salts and large amounts can be given in small volumes. By injection, diamorphine is twice as potent as morphine. Because of rapid deacetylation, however, diamorphine by mouth is essentially a prodrug for morphine and is only marginally more potent than morphine.

There are several opioid receptor subtypes (µ, _, †). Opioids differ in their receptor-site affinity, intrinsic activity, and concomitant non-opioid effects. These differences can be capitalized on in patients who are intolerant of morphine (mainly a µ-receptor agonist) by converting, for example, to methadone (µ-receptor agonist with non-opioid properties) or oxycodone (µ-, _-receptor agonist).

When switching from another strong opioid to oral morphine (or vice versa), the initial dose depends on the relative potency of the two drugs (Table 2). The main reason for changing from morphine to another opioid is intolerance of morphine, for example marked dysphoria, sedation, and/or persistent hallucinations. Pethidine is not recommended for cancer pain—it acts for only 2 to 3 h.

Adjuvant analgesics

Adjuvant analgesics are miscellaneous drugs that relieve pain in specific circumstances. The main ones are corticosteroids, antidepressants, antiepileptics, muscle relaxants, and antispasmodics.

Corticosteroids

Corticosteroids are particularly useful for pain associated with nerve root or nerve trunk compression, and with spinal cord compression. Dexamethasone, 4 to 8 mg once a day, is prescribed for the former and 12 to 20 mg (sometimes more) for cord compression together with radiotherapy.

Antidepressants and antiepileptics

Nerve injury pains often do not respond well to non-opioids and opioids because of central (dorsal horn) sensitization. Additional measures are commonly necessary to obtain satisfactory relief. These aim to:

- dampen the hyperexcitability of damaged peripheral nerves;
- inhibit the glutamate excitatory system in the dorsal horn;
- enhance the g-aminobutyric acid (**GABA**) inhibitory system in the dorsal horn (Fig. 5).

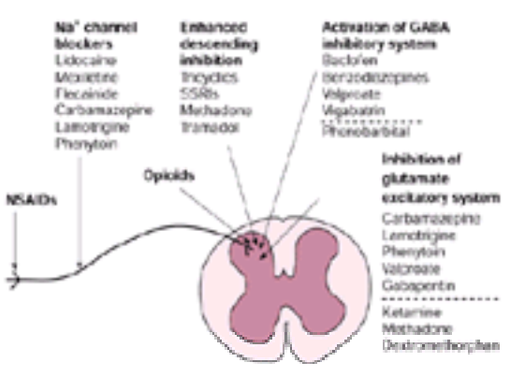

Fig. 5 Impact of analgesics and adjuvant analgesics on peripheral nerves and the dorsal horn of the spinal cord. Drugs below the dotted lines are channel blockers; the rest are receptor ligands.

For example, amitriptyline 25 to 75 mg at night (an antidepressant) and sodium valproate 400–1000 mg at night (an antiepileptic) can be used either singly or together. These should generally be given in addition to morphine and a non-opioid analgesic if the nerve injury pain is associated with an infiltrating cancer, but they may well be effective alone in 'pure' nerve injury pain (for example, chronic postsurgical incision pain, postherpetic neuralgia).

Gabapentin is the only antiepileptic drug that is specifically licensed in the United Kingdom for the relief of neuropathic pain. The effective dose varies between 100 and 1200 mg given three times a day. Like many drugs acting on the central nervous system, gabapentin can cause drowsiness; most patients with cancer cannot tolerate more 600 mg three times a day.

With nerve injury pain which does not respond to such measures, it may be necessary to supplement or replace the antidepressant and antiepileptic with a glutamate (**NMDA**, *N*-methyl-D-aspartate) receptor-channel blocker, for example methadone and ketamine. A few patients (<1 per cent) require spinal analgesia (for example, morphine plus bupivacaine, with or without clonidine) or a neurolytic procedure to obtain adequate relief. Some patients derive benefit from other non-drug measures, for instance transcutaneous electrical nerve stimulation (**TENS**).

Antispasmodics and muscle relaxants

Muscle spasm pain (cramp) secondary to underlying bone pain and/or skeletal deformity is common in cancer patients. Myofascial trigger-point pains also occur. For these the correct approach is not more analgesics but explanation, physical therapy (massage and local heat), diazepam, and relaxation therapy. Trigger points can be injected with a local anaesthetic and a corticosteroid, for example bupivacaine 0.5 per cent and depot methylprednisolone.

Antispasmodics such as hyoscine butylbromide and glycopyrronium (given by subcutaneous infusion) are necessary in some patients with inoperable endstage intestinal obstruction.

Nausea and vomiting

Nausea and vomiting occurs in about half of the patients with advanced cancer. Intestinal obstruction, gastric stasis, drugs, and biochemical abnormalities are responsible in about 80 per cent of cases. It is often the sequence of events (plus an appropriate level of suspicion) that points to the likely cause.

In **acute** bowel obstruction there is typically a single discrete lesion, whereas in **chronic** obstruction (persistent or remittent) there may well be several sites of partial obstruction in both small and large bowels. Retroperitoneal disease may cause visceral neuropathy and functional obstruction. In consequence, the quartet of symptoms and signs that point to a diagnosis of acute intestinal obstruction (abdominal distension, pain, vomiting, and constipation) is often not so obvious in patients with advanced cancer and chronic obstruction. For example, distension may be minimal because of multiple intra-abdominal malignant adhesions. Bowel sounds vary from absent to overactive with borborygmi; tinkling bowel sounds are unusual. Some patients have diarrhoea rather than constipation.

A degree of gastric stasis is present in many patients with advanced cancer. Diagnosis depends on a high level of clinical suspicion and pattern recognition (Box 3). There are several causes of gastric stasis and multiple factors may be responsible:

- dysmotility dyspepsia (often long-standing);
- drugs, e.g. opioids, antimuscarinics;
- cancer of the head of the pancreas (disrupts duodenal transit);
- retroperitoneal disease leading to neural dysfunction;
- spinal cord compression;
- paraneoplastic autonomic neuropathy;
- diabetic autonomic neuropathy.

Box 3 Clinical features of gastric stasis

- Some or all of the following symptoms may occur:

  - early satiety
  - postprandial fullness
  - epigastric bloating
  - epigastric discomfort
  - heartburn
  - belching
  - hiccup
  - nausea
  - retching
  - vomiting

- On examination there may be:

  - epigastric distension
  - a succussion splash

  The absence of these features does not rule out symptomatic gastric stasis.

  If associated with autonomic neuropathy, there is often evidence of other autonomic abnormalities, e.g. orthostatic hypotension without a compensatory tachycardia.

  •bowel sounds are generally normal but may be decreased if the stasis is opioid-induced

  The use of metoclopramide (a prokinetic drug) often leads to improvement and the resolution of the succussion splash.

Management

Correct the correctable

Medication may need to be modified, ascites drained, and hypercalcaemia corrected (unless coincidental in a moribund patient). Surgery for bowel obstruction should be considered if all the following criteria are all fulfilled:

- an easily reversible cause seems likely, e.g. postoperative adhesions or a single discrete neoplastic obstruction;
- the patient's general condition is good, i.e. does not have widely disseminated disease and has been independent and active; and
- the patient is willing to undergo surgery.

Surgical intervention for bowel obstruction is contraindicated in each of the following circumstances:

- previous laparotomy findings preclude the prospect of a successful intervention;
- diffuse intra-abdominal carcinomatosis as evidenced by diffuse palpable intra-abdominal tumours;
- massive ascites that reaccumulates rapidly after paracentesis.

Non-drug measures

Make sure that the patient is not assailed by food smells from the kitchen. Offer only small helpings of food. Possibly try an acupressure band on one or both wrists.

Drugs

The initial choice generally lies between four drugs: metoclopramide, haloperidol, hyoscine butylbromide, and cyclizine (Box 4). Second-line drugs may need to be added or substituted in some patients.

Box 4 Guidelines for the management of nausea and vomiting in palliative care

1. 1.After clinical evaluation, document the most likely cause(s) of the nausea and vomiting in the patient's case notes.
2. Ask the patient to record their symptoms and response to treatment.
3. Treat correctable causes/exacerbating factors, e.g. drugs, constipation, severe pain, infection, cough, hypercalcaemia. *Correction of hypercalcaemia may not be appropriate in a dying patient*.
4. Anxiety exacerbates nausea and vomiting from any cause and may need specific treatment, pharmacological and/or psychological.
5. Prescribe the most appropriate first-line antiemetic for the most likely main cause: immediately, regularly, and 'as needed'. Give subcutaneously if continuous nausea or frequent vomiting, preferably by subcutaneous infusion.
6. **First-line antiemetics:**
7. *Prokinetic antiemetic*
8. For gastritis, gastric stasis, functional bowel obstruction:
9. metoclopramide 10 mg immediately by mouth, and then four times a day; or 10 mg immediately subcutaneously, and 40–100 mg/24 h as a subcutaneous infusion, and 10 mg 'as needed' up to four times daily.
10. *Antiemetic acting principally in chemoreceptor trigger zone (area postrema)*
11. For most chemical causes of vomiting, e.g. morphine, hypercalcaemia, renal failure:
12. haloperidol 1.5–3 mg immediately by mouth, and at bedtime; or 2.5–5 mg immediately subcutaneously, and 2.5–10 mg/24 h by subcutaneous infusion and 2.5–5 mg 'as needed' up to four times daily.
13. Metoclopramide also acts here.
14. *Antispasmodic and antisecretory antiemetic*
15. If bowel colic and/or need to reduce gastrointestinal secretions:
16. hyoscine butylbromide 20 mg immediately subcutaneously, and 80–160 mg/24 h by subcutaneous infusion, and 20 mg 'as needed' up to every lR.
17. *Antiemetic acting principally in the vomiting centre*
18. For organic bowel obstruction, raised intracranial pressure, motion sickness:
19. cyclizine 50 mg immediately by mouth, and two or three times per day; or 50 mg immediately subcutaneously, and 100–150 mg/24 h by subcutaneous infusion, and 50 mg 'as needed' up to four times daily.
20. Review the dose of antiemetic every 24 h, taking note of 'as needed' use and the patient's diary.
21. If there is little or no benefit, despite optimizing the dose, have you got the cause right?
22. •if no, change to an alternative first-line antiemetic and optimize
    - if yes, provided the first-line antiemetic has been optimized *add* or *substitute* the second-line antiemetic
    - for patients with obstructive vomiting, follow the steps in Fig. 6.

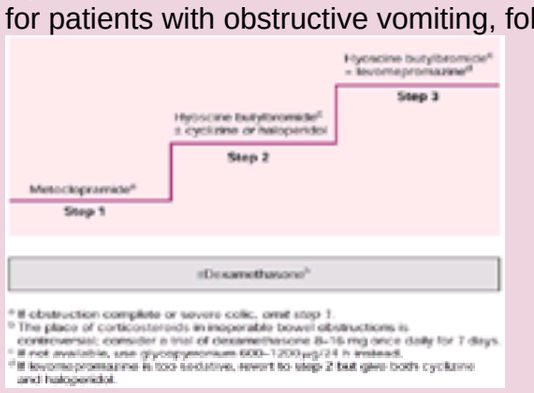

Fig. 6 Antiemetics for endstage bowel obstruction. These are normally given by subcutaneous injection/infusion. (a) If the obstruction complete or there is severe colic, start at step 2. (b) The role of corticosteroids in bowel obstruction is controversial; consider a trial of dexamethasone 8–16 mg once a day for 5 days. (c) If not available, use glycopyrronium 600–1200 µg/24 h instead. (d) If levomepromazine is too sedative, revert to step 2 and add haloperidol.

23. **Second-line drugs for nausea and vomiting**
24. *Broad-spectrum antiemetic*
25. If first-line antiemetics (in appropriate combination, dose, and route) are inadequate:
26. *levomepromazine (methotrimeprazine) 6–12.5 mg by mouth or subcutaneously immediately, at bedtime, and* 'as needed' up to four times daily.
27. *Corticosteroids*
28. Adjuvant antiemetic
29. *dexamethasone 8–16 mg immediately by mouth or subcutaneously, and once daily; consider reducing the dose after 7 days.*
30. *5-HT3-receptor antagonist*
31. Use when there is a massive release of 5-HT (serotonin) from enterochromaffin cells or platelets, e.g. chemotherapy, abdominal radiation, bowel obstruction, renal failure:
32. tropisetron 5 mg *immediately by mouth or subcutaneously, and once daily*.
33. Some patients with nausea and vomiting need more than one antiemetic.
34. *Do not prescribe a prokinetic and an antimuscarinic drug concurrently because the latter blocks the cholinergic final common pathway through which prokinetics act*.
35. Except in organic bowel obstruction consider converting to the oral route after 72 h of good control with injections.
36. Continue antiemetics indefinitely unless the cause is self-limiting.

A systematic approach for inoperable bowel obstruction is shown in Fig. 6. For those with severe colic or who are not passing flatus, prokinetic drugs are contraindicated and step 1 is omitted. Bulk-forming, osmotic, and stimulant laxatives should be stopped, but patients may well benefit from a faecal softener, for example docusate 100 to 200 mg once to twice a day. Octreotide, a somatostatin analogue, has intestinal antisecretory properties and is occasionally indicated. It is much more expensive than hyoscine butylbromide but has no antimuscarinic effects. A venting gastrostomy is rarely necessary for symptom relief in patients with advanced cancer and bowel obstruction.

Patients who are inoperable and are managed by drug therapy should be allowed to drink and eat small amounts of their favourite beverages and food. Some patients find that they can manage food best in the morning. Antimuscarinic drugs and diminished fluid intake often result in a dry mouth. This is generally relieved by conscientious mouth care. A few millitres of fluid every 30 min, possibly given as a small ice cube, is often helpful.

Constipation

Constipation (difficulty in defaecation) is common in advanced cancer. Diminished food and fibre intake, lack of exercise, and drugs are common causal factors.

Evaluation

Many patients who are constipated do not need a rectal examination (by definition, an assault which must be justifiable). On the other hand, a rectal examination is essential in patients with faecal leakage or diarrhoea to confirm or exclude faecal impaction.

In non-obese patients, firm faeces are often palpable in the left iliac fossa and left side of the abdomen. Faeces may also be palpable in the transverse colon, and occasionally in the ascending colon as well, together with caecal distension and tenderness. Bowel sounds are variable. Sometimes it is not clear whether the problem is severe constipation, chronic intestinal obstruction, or both. Although plain radiographs confirm the presence of retained faeces, they cannot reliably differentiate between the two conditions. Pattern recognition and probability may enable a presumptive diagnosis to be made, but sometimes only the passage of time and the response (or lack of response) to treatment confirm whether it is just constipation or obstruction.

Management

Laxatives are the mainstay of treatment for patients with a limited physical capacity and reduced intake of food and fibre. For patients not taking an opioid, start with senna or bisacodyl tablets, adding a faecal softener (for example, docusate) if necessary. Alternatively, a combination preparation (for example, codanthramer, codanthrusate) can be substituted (Box 5). Some patients do better with lactulose but may require 30 ml twice daily or more. Opioids cause constipation by decreasing propulsive activity and increasing non-propulsive activity in both the small and large bowel; peristalsis is impeded and absorption of fluid and electrolytes is facilitated. So-called stimulant laxatives act principally by reducing intestinal ring contractions, thereby facilitating propulsive activity.

Box 5 Management of opioid-induced constipation

- Ask about the patient's past (premorbid) and present bowel habit and use of laxatives; record date of last bowel action.
- Do a rectal examination if faecal impaction is suspected or if the patient reports diarrhoea or faecal incontinence (to exclude impaction with overflow).
- For inpatients, keep a daily record of bowel actions.
- Encourage fluids generally, and fruit juice and fruit specifically.
- When an opioid is first prescribed, prescribe co-danthrusate [a] (one capsule at bedtime) prophylactically; although occasionally appropriate to optimize a patient's existing bowel regimen, rather than change automatically to co-danthrusate.
- If already constipated, prescribe co-danthrusate (two capsules at bedtime).
- Adjust the dose every few days according to results, up to three capsules three times a day.
- If the patient prefers a liquid preparation, use co-danthrusate suspension; 5 ml is equivalent to one capsule.
- If more than 3 days since the last bowel action, 'uncork' with suppositories, e.g. bisacodyl 10 mg and glycerol 4 g.
- If suppositories are ineffective, administer a high phosphate enema; possibly repeat the next day.
- If the maximum dose of co-danthrusate is ineffective, switch to an osmotic laxative, e.g. lactulose 20–30 ml twice daily or macrogol 3350 1 to 3 sachets daily ± a reduced dose of co-danthrusate.
- If co-danthrusate causes abdominal cramps, divide the total daily dose into smaller more frequent doses, e.g. change from co-danthrusate two capsules twice daily to one capsule four times daily or change to lactulose or macrogol 3350.
- An osmotic laxative may be preferable to co-danthrusate in patients with a history of colic with other colonic stimulants.

About one-third of patients receiving morphine continue to need rectal measures (suppositories, enemas, digital evacuation) either regularly or intermittently. Danthron-containing laxatives are inadvisable in patients with faecal leakage or incontinence because danthron can cause a contact skin burn. Because of concern that it may be carcinogenic, danthron-containing laxatives are now only licensed for use in terminally ill patients.

Dyspnoea

Dyspnoea is an unpleasant subjective awareness of difficulty in breathing. Objective signs generally include tachypnoea (an increased rate of respiration) and sometimes hyperpnoea (an increased depth of respiration). Dyspnoea becomes more common as death approaches; overall it is experienced by about 70 per cent of patients with advanced cancer.

Evaluation

In most patients with advanced cancer, dyspnoea is caused by several factors, for example chronic obstructive pulmonary disease, progressive intrathoracic malignant disease, anaemia of chronic disease, weakness (Box 6). The history and examination are often sufficient to determine the main causes.

Box 6 Causes of breathlessness in advanced cancer

Caused by cancer	**Related to cancer and/or debility**
Pleural effusion(s)	Anaemia
Obstruction of main bronchus	Atelectasis
Replacement of lung by cancer	Pulmonary embolism
Lymphangitis carcinomatosa	Pneumonia
Mediastinal obstruction	Empyema
Pericardial effusion	Weakness
Massive ascites	
Abdominal distension	**Concurrent causes**
Cachexia–anorexia syndrome	Chronic obstructive pulmonary disease (COPD)
Caused by treatment	Asthma
Pneumonectomy	Heart failure
Radiation-induced fibrosis	Acidosis
Chemotherapy	
bleomycin	
doxorubicin	

Management

Non-drug measures

The key to successful management at the end of life is at an earlier stage, when dyspnoea on exertion first becomes a symptom. Acknowledgement of and discussion about the terror associated with acute episodes of breathlessness, for instance when climbing stairs, is very important. Referral to a physiotherapist for breathing advice and relaxation techniques is important.

Correct the correctable

Specific treatment should be given for specific causes, for example: bronchodilators for reversible airways obstruction; diuretics for cardiac failure; aspiration and pleuradesis for pleural effusion; and radiotherapy or stenting for superior vena caval obstruction.

Drugs

When the patient is close to death, breathlessness may be present at rest as well as on exertion, or become apparent on minimal activity. Often the patient is bedbound, more because of breathlessness than weakness. In this situation, an opioid (to slow the respiratory rate towards normal), an anxiolytic, and oxygen therapy may all be helpful (Box 7).

Box 7 Relief of breathlessness at rest in a dying patient

- **Bronchodilators**
- The use of bronchodilators should be reviewed and tested clinically for efficacy; it may be necessary to use a spacer to ensure adequate inhalation from an inhaler, or to convert to a nebulizer. Benefit is not always correlated with improvement in peak flow.
- **Opioids**
- Morphine reduces respiratory drive and can be used to ease the sensation of dyspnoea:
- if on morphine for pain, increase the dose by 50 per cent;
- if not on oral morphine, 5–6 mg every 4 h is a good starting dose.
- Nebulized morphine is not recommended; it is no better than saline.
- **Anxiolytics**
- Diazepam by mouth is a good choice:
- 5–10 mg immediately and bedtime;
- 2–5 mg in the very elderly;
- reduce dose after several days if drowsy.
- Use midazolam subcutaneously for patients who find tablets difficult to take:
- 2.5–5 mg immediately and 'as needed';
- 10–20 mg/24 h by subcutaneous infusion.
- **Oxygen**
- Oxygen 4 L/min, preferably via nasal prongs, should be tried to see if benefit gained either continuously or before and during activity (e.g. moving from bed to commode or chair). The benefit of oxygen is not dependent on correction of hypoxaemia; a trial of therapy is the only way to determine benefit, not improvement in blood gases.

Anorexia

Anorexia (diminished or absent appetite) is normal in advanced cancer.

Evaluation

Anorexia may be primary (cachexia–anorexia syndrome) or secondary. Secondary anorexia may be caused by:

- medication which causes dyspepsia or nausea, e.g. NSAIDs, opioids, antibiotics, selective serotonin-reuptake inhibitors (**SSRIs**);
- nausea;
- pain;
- altered taste;
- difficulty feeding because of weakness, dysphagia, or a sore mouth;
- fatigue;
- psychological distress, e.g. fear, anxiety, depression.

Patients are generally more tolerant of anorexia than their families, for whom it is a source of great concern. Many patients state that their carers try too hard to encourage eating, resulting in conflict. Eating is an important social interaction and anorexia may be interpreted by the family as giving in to the cancer. Explanation should be given about the cause(s) of anorexia and the limitations of treatment.

Management

Correct the correctable

Review the current medication for a possible cause; treat pain, nausea, and sore mouth. Obtain dietary advice to minimize the impact of an altered sense of taste.

Non-drug measures

'A little of what you fancy when you fancy' is good advice. Frequent snacks high in calories and low in bulk are preferable to traditional meals.

Drugs

If anorexia appears to be mainly due to early satiety, a prokinetic agent should be tried, for example metoclopramide 10 mg four times daily by mouth.

Prednisolone (15 to 20 mg once a day) or dexamethasone (2 to 4 mg once daily) help about 50 per cent of patients. To avoid cumulative adverse effects, a corticosteroid should be discontinued if there is no benefit after a week or, if effective, reduced to a maintenance dose. Patients and families should be forewarned that the benefit often lasts for only a few weeks.

Medroxyprogesterone acetate (400 mg once daily up to 500 mg twice daily) or megestrol acetate (160 mg once daily up to 800 mg once daily) often lead to weight gain after 3 to 4 weeks, particularly in patients with breast cancer, as a result of an increase in both fat and body water. The effect may last for months. Progestogens are much more expensive than corticosteroids.

Cachexia

Cachexia (marked weight loss and muscle wasting) occurs in up to 80 per cent of patients with advanced cancer. Muscle wasting results in weakness and fatigue. The incidence is highest in cancers of the gastrointestinal tract and lung.

Evaluation

Cachexia is caused by several interrelated metabolic disturbances (Box 8). Concurrent exacerbating factors may be reversible and should be considered.

Box 8 Causes of cachexia in advanced cancer

Paraneoplastic	**Concurrent**
Cytokines produced by host cells and tumour (e.g. TNF, IL-6, IL-3 [a])	Anorexia ® deficient food intake
	Vomiting

Abnormal host metabolism of protein, carbohydrate, and fat	Diarrhoea
	Malabsorption
Increased metabolic rate ® increased energy expenditure	Bowel obstruction
Nitrogen trap by the tumour	Debilitating effect of surgery, radiotherapy, chemotherapy
	Ulceration } excessive loss
	Haemorrhage of body protein

Management

Correct the correctable

As for anorexia (see above).

Non-drug measures

Avoid routine weighing. Explain that 'forced feeding' cannot correct primary cachexia. Some patients benefit psychologically from powdered or liquid nutritional supplements, and a few gain weight. Dietary advice is important if there are changes in taste sensation.

Drugs

As for anorexia. In most patients the effect is small or non-existent. Management is often best focused on acceptance and adaptation.

Dehydration

As the dying patient becomes weaker, intake of both food and fluid becomes less. Decreased fluid intake may be caused by dysphagia, nausea, anorexia, lack of energy and interest, and a reduced level of consciousness, either singly or in combination. There is a big difference between acute and chronic dehydration. Whereas the former is accompanied by intense thirst, this is not the case in dying patients when there is a slow progressive reduction in fluid intake. Conscientious mouth care (cleaning and moistening) is generally all that is called for. As a general rule, intravenous fluids should not be administered if the decreased intake is best interpreted as part of the process of dying. Guidelines from the British Medical Association support this approach.

Confusion

An acute confusional state (delirium) eventually occurs in most dying patients. This typically manifests as disorientation, bewilderment, and drowsiness, often compounded by a spectrum of cognitive disturbances such as poor concentration, impairment of short-term memory, misinterpretation, paranoid ideas, hallucinations, rambling incoherent speech, agitation, and noisy aggressive behaviour.

Evaluation

Multiple factors may contribute to confusion (Box 9). Biochemical investigations are generally contraindicated in patients close to death. It is important not to overlook urinary retention and faecal impaction.

Box 9 Common causes of acute confusion in the dying

Cancer	Drugs
paraneoplastic effect	sedative
cerebral involvement	psychostimulant
Infection	antiparkinsonian
Dehydration	Drug withdrawal
Change of environment	psychotropics
Unfamiliar excessive stimuli	alcohol
too hot	nicotine
too cold	Biochemical derangement
wet bed	hypercalcaemia
crumbs in bed	hyponatraemia
creases in sheets	hypoglycaemia
pain	hyperglycaemia
constipation	Cerebral anoxia
retention of urine	anaemia
pruritus	cardiac failure
Anxiety	Organ failure
Depression	hepatic
Fatigue	renal

Management

Correct the correctable

The patient's drug regimen should be reviewed and a reduction in psychoactive drugs considered, for example amitriptyline 75 mg once a day reduced to 25 mg once a day. Occasionally nicotine or alcohol withdrawal may be the main cause (see below). Pneumonia often precipitates confusion. Treating pneumonia with antibiotics is generally inappropriate in a debilitated, bedbound, dying patient.

Non-drug measures

The presence of a close relative or friend is generally helpful. Visual cues (day, date, soft light, and photographs) and auditory cues (favourite music) may help (Box 10).

Box 10 Treatment of acute confusion

- **Non-drug measures**
- Explanation to patient, family, nurses.
- Stress that patient is not going mad.
- Stress that almost always there are lucid intervals.
- Continue to treat patient as a sane, sensible person.
- Be aware that illusions, hallucinations, and nightmares may reflect unresolved fears and anxiety.
- **Drugs**
- Use drugs only if symptoms are marked, persistent, and cause distress to the patient and/or family. Review sooner rather than later if a sedative drug is prescribed in case it exacerbates symptoms.
- *Specific*
- Dose reduction of present psychotropic medication?
- If hypoxic or cyanosed, give oxygen.
- If cerebral tumour, give dexamethasone 8–16 mg once daily. *Correction of cerebral oedema may not be appropriate in a dying patient.*
- If nicotine withdrawal, give nicotine either as a nasal spray or transdermal patch.
- *General*
- Haloperidol 1.5–5 mg by mouth or subcutaneously, particularly if hallucinations and paranoid ideas.
- Diazepam 5–10 mg by mouth or midazolam 5–10 mg subcutaneously if still restless after two doses of haloperidol (or give concurrently if there is myoclonus):
- initial dose depends on previous medication, weight, age, and severity of symptoms;
- subsequent doses depend on initial response;
- daily- or twice-daily maintenance doses are generally adequate;
- sometimes more frequent administration is necessary.
- In extreme situations, it may be necessary to deeply sedate a dying agitated patient, e.g.:
- levomepromazine 25–50 mg immediately and 50–200 mg/24 h subcutaneously;
- phenobarbital 100–200 mg immediately and 800–1600 mg/24 h subcutaneously.

Drugs

If the patient is agitated, haloperidol should be given, for example 1.5 to 5 mg by mouth or 2.5 to 5 mg subcutaneously. Diazepam or midazolam should be added if the patient remains agitated after 5 to 10 mg of haloperidol (Box 10). If nicotine or alcohol withdrawal is suspected and the patient is unable to smoke or drink, specific treatment should be considered—for instance, transdermal nicotine patches, or a benzodiazepine with or without thiamine for delirium tremens.

Terminal anguish

Terminal anguish is a tormented state of mind that relates to long-standing unresolved emotional problems and/or interpersonal conflicts, or to long-hidden unhappy memories often with a guilty content. These problems have festered in the mind but have never been brought into the open. The possibility of such an outcome highlights the need to make every effort to deal with psychological 'skeletons in the cupboard' before the patient becomes too weak to be able to address them. A few patients, however, resist every attempt to explore what has been suppressed.

Terminal anguish is managed in the same way as an agitated confusional state, for instance by combining an antipsychotic and a benzodiazepine. Large doses may be required and sometimes the more sedative levomepromazine (methotrimeprazine) is preferable to haloperidol. A dose of 50 to 100 mg per 24 h is often very sedative, but this (or more) is what may be necessary to ensure calm. On rare occasions, it may be necessary to use subcutaneous phenobarbital (Box 10).

Death rattle

Evaluation

The inability to clear secretions from the oropharynx and trachea results in noisy (rattling) breathing as the secretions oscillate with respiration. When a death rattle develops the patient is almost always unconscious and untroubled by the secretions. It helps the family to appreciate this, but some families still find it distressing.

Management

Non-drug measures

Placing the unconscious patient in the recovery (semi-prone) position aids the drainage of secretions from the mouth. Suctioning is generally contraindicated; it can distress a patient who is otherwise settled.

Drugs

Because antimuscarinic drugs do not dry up secretions already present, it is important to act at the first sign of rattling. Hyoscine hydrobromide (0.4 to 0.6 mg immediately subcutaneously and 1.2 mg by subcutaneous infusion over 24 h) or hyoscine butylbromide (Buscopan; 20 mg subcutaneous at once and 20 to 40 mg by infusion over 24 h) are both widely used. Glycopyrronium may be used instead, for example, 0.2 mg immediately subcutaneously and 0.6 mg by infusion over 24 h). Such measures are successful in about 50 to 60 per cent of patients. Intravenous diuretics are of benefit if there is concomitant left ventricular failure, which is unusual.

Ethical considerations

The cardinal principles that underpin clinical practice, including palliative care, are:

- respect for patient autonomy (patient choice);
- beneficence (do good);
- non-maleficence (minimize harm);
- justice (fair use of available resources).

These four principles need to be applied against the background of respect for life and an acceptance of the ultimate inevitability of death. Thus, in practice, three dichotomies need to be held in balance:

- the potential benefits of treatment versus the potential risks and burdens;
- striving to preserve life but, when the burdens of life-sustaining treatments outweigh the potential benefits, then withdrawing or withholding such treatments and providing comfort in dying;

- individual needs versus the needs of society.

Principle of double effect

The principle of double effect states that:

> A single act having two possible foreseen effects, one good and one harmful, is not always morally prohibited if the harmful effect is not intended.

This is a universal principle without which the practice of medicine would be impossible. It follows inevitably from the fact that all treatment has an inherent risk. However, discussions of the principle of double effect are often limited to the use of morphine or similar drug to relieve pain in terminally ill patients. This gives the false impression that the use of morphine in this circumstance is a high-risk strategy. When correctly used morphine (and other strong opioids) are very safe drugs, safer than NSAIDs, which are widely prescribed with relative impunity. The use of both classes of analgesic is justified on the basis that the benefits of pain relief far outweigh the risk of serious adverse effects. Indeed, clinical experience suggests that those whose pain is relieved live longer than would have been the case if they had continued to be exhausted and demoralized by severe unremitting pain.

However, the intended aim of treatment must be the relief of suffering and not the patient's death. Although a greater risk is acceptable in more extreme circumstances, it remains axiomatic that effective measures that carry less risk to life should normally be used. Thus, in an extreme situation, although it may occasionally be necessary (and acceptable) to render a patient unconscious, it remains unacceptable (and unnecessary) to cause death deliberately. Deliberate hastening of death by intentionally giving an overdose of one or more drugs (euthanasia) is illegal in almost all countries. Palliative care and euthanasia are essentially mutually exclusive philosophies.

Appropriate treatment

A doctor is not obliged legally or ethically to preserve life 'at all costs'. Priorities change when a patient is clearly dying. There is no obligation to employ treatments if their use can best be described as prolonging the process of dying. A doctor has neither duty nor right to prescribe a lingering death. In palliative care, the primary aim of treatment is not to prolong life but to make the life that remains as comfortable and as meaningful as possible. Part of the art of medicine is to decide when to allow death to occur without further medical impediment.

However, it is not a question of to treat or not to treat but what is the most appropriate treatment given the patient's biological prospects and his personal and social circumstances? Appropriate treatment for an acutely ill patient may be inappropriate in the dying (Fig. 7). Nasogastric tubes, intravenous infusions, antibiotics, cardiac resuscitation, and artificial respiration are all primarily support measures for use in acute or acute-on-chronic illnesses to assist a patient through the initial crisis towards recovery of health. The use of these measures in patients who are irreversibly close to death is generally inappropriate (and therefore bad practice) because the burdens of such treatments exceed their potential benefits.

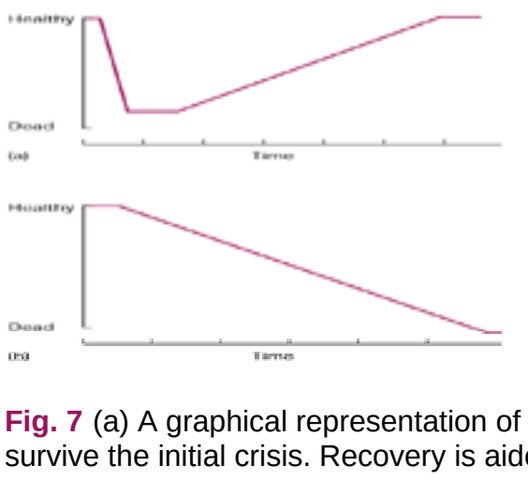

Fig. 7 (a) A graphical representation of acute illness. Biological prospects are generally good. Acute resuscitative measures are important and enable the patient to survive the initial crisis. Recovery is aided by the natural forces of healing; rehabilitation is completed by the patient on his own, without continued medical support. (b) A graphical representation of terminal illness. Biological prospects progressively worsen. Acute and terminal illnesses are therefore distinct pathophysiological entities. Therapeutic interventions that can best be described as prolonging the distress of dying are essentially futile and inappropriate.

Although the possibility of unexpected improvement or recovery should not be totally ignored, there are many occasions when it is appropriate to 'give death a chance'. Interest in hydration and nutrition often becomes minimal as death draws near. The patient's disinterest or positive disinclination is part of the process of letting go.

At the end

Although eventually you may feel powerless in the face of approaching death, patients are generally more realistic. They know you cannot perform a miracle and time is limited. Despite possibly having nothing new to offer, it is important for the doctor to:

- continue to visit;
- quietly indicate that, 'The important thing now is to keep you comfortable';
- simplify medication;
- arrange for medication to be given sublingually, rectally, or by continuous subcutaneous infusion when the patient cannot swallow;
- continue to inform the family of the changing situation;
- control agitation, even if it results in sedation;
- listen to the nurses.

Further reading

BMA Report (1999). *Withholding and withdrawing life-prolonging medical treatment.* BMA, London.

Doyle D, Hanks GWC, MacDonald N, eds (1997). *Oxford textbook of palliative medicine*, 2nd edn. Oxford Medical Publications, Oxford.

National Council for Hospice and Specialist Palliative Care Services (1997). Artificial hydration (AH) for people who are terminally ill. *European Journal of Palliative Care* **4,** 124.

Sindrup SH, Jensen TS (1999). Efficacy of pharmacological treatments of neuropathic pain: an update and effect related to mechanism of drug action. *Pain* **83,** 389–400.

Twycross R. (1999). *Introducing palliative care*, 3rd edn. Radcliffe Medical Press, Oxford.

Twycross R, Wilcock A, Charlesworth S, Dickman A (2002). *Palliative care formulary*, 2nd edn. Radcliffe Medical Press, Oxford.

32Reference intervals for biochemical data

P. A. H. Holloway and A. M. Giles

Classification of laboratory results

Intervals are presented in the following format;

- for individual tests see general text:
- Everyday tests and enzymes Table 1
- Blood gases Table 2
- Paediatric reference ranges Table 3
- Hormones Table 4
- Tumour markers Table 5
- Vitamins and related tests Table 6
- Lipids and lipoproteins Table 7
- Proteins and immunoproteins Table 8
- Trace elements and metals Table 9
- Urinary values Table 10
- Faecal values Table 11
- Cerebrospinal fluid Table 12
- Functional tests Table 13
- Therapeutic drugs Table 14
- Common drug toxicology Table 15

Introduction

The precise quantitation of a substance in easily accessible body fluids is an integral part of the clinical assessment of patients. The results are used in screening for disease as well as in diagnosis and for monitoring the response to therapy in established disease. Much diagnostic weight rests on single determinations and patterns of biochemical tests. To this end it is important to consider biological variations between healthy individuals, inherent variations in laboratory methods, and the errors of sampling and hospital practice which can influence every determination. The first (and the last) are the provinces of the physician ordering the test. The second is the concern of the laboratory which provides quality control and the reference intervals for the test.

An important growth area in diagnostic pathology is emerging in the field of 'point-of-care' testing (**POCT**), particularly in the critical care environment. Whilst technology has advanced to enable rapid analytical turnaround times in POCT—usually far less than 10 min from sample withdrawal to result—with consequent hastening to clinical decision-making, the laboratory responsibility for interpretation and overseeing test results is partially devolved to the nurse or doctor at the bedside. In this context there is greater need for an understanding of the appropriate reference intervals as well as of any limitations on the analytical precision when compared to central laboratory methods.

'Normal range' and 'abnormal' results

Clinical diagnostic decisions may depend equally upon finding a 'normal' or 'abnormal' result for any test requested. The physician should be clear as to the meaning of these terms. An important task of the clinical biochemist is therefore to provide relevant sets of reliable reference data. For any individual the ideal reference value for an analyte should be that obtained when that individual is healthy. However, in practice, laboratory test results are interpreted by comparison to traditional, but often inadequately, defined reference intervals (formerly termed 'normal' ranges). The wide belief that biological data assume a gaussian distribution is inappropriate. Most biological data are not symmetrically distributed and require statistical tools that assume other kinds of distribution or are independent of distribution form. Ideally, each laboratory should establish sets of reference intervals derived from a local reference population.

Reference interval

In the past many texts quoted a 'reference range' defined as the mean ± 2 standard deviations from the mean of results obtained from the reference population. This is, however, not now the commonest method applied to clinical biochemistry tests. Many of the 'intervals' quoted in clinical practice are in fact derived from a skew-distribution, which is calculated from the geometric mean to include 95 per cent of values obtained from what is considered to be a 'healthy' population (95 per cent confidence intervals). The merits of a diagnostic test are determined by the relationship between the data for healthy and unhealthy populations, and an example of a relatively poor test is given in Fig. 1. By whatever criteria it is obtained, the reference interval is compounded of both physiological variation and the irreducible error. More elaborate statistical handling of human biochemical data is available for individual tests, but is not generally required in making a diagnosis. By way of warning, however, it will readily be appreciated that if this criterion of health (that is, the biochemical results within 95 per cent limits for the given value), is applied to multiple tests in any individual, then with a battery of say 12 tests only 50 per cent of 'normal' individuals will be found to be 'healthy'. An important example of the limitations of strict referral to the 'reference interval' is highlighted by plasma total cholesterol measurements within certain populations, where established reference intervals may be considered to contain a considerable proportion of individuals with 'undesirable' or 'unhealthy' values. In such situations it may be appropriate to define an 'ideal' or 'optimum' range for that parameter. Other difficulties arise from situations where the reference interval has not been constructed from a population of matched ethnic mix.

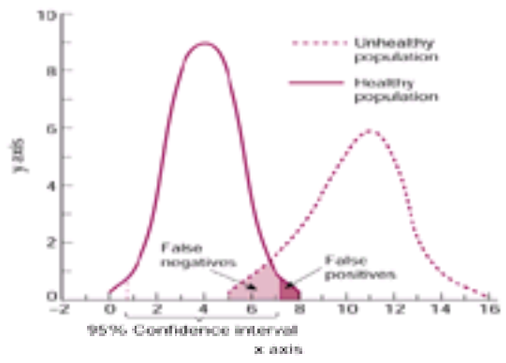

Fig. 1 Theoretical distributions of an analyte in healthy (symmetrical) and unhealthy (negatively skewed) populations.

In conclusion, the use of the following tables of reference intervals for biochemical tests must therefore include an appreciation of the limitations in reference intervals of analytical variation and differences between methods of analysis, as well as sex, age, posture, diet, and other biological variables, as well as sampling, that contribute to such an operation. The physician should constantly bear in mind the need to see individual results of laboratory tests in their clinical context, and not to

hesitate in questioning an unexpected or unlikely result.

The reference intervals given here are predominantly those established for laboratory investigations available at the John Radcliffe Hospital in Oxford, in the United Kingdom. Throughout the tables intervals are given in SI and conventional units, together wherever possible with the factor for converting from SI to conventional. Where appropriate, specific, specimen collection details are given, although it should be recognized that with varied methodologies for most assays it is not possible to make these prescriptive.

33 Emergency medicine

J. D. Firth, C. A. Eynon, D. A. Warrell, and T. M. Cox

1 Heart and circulation

1.1 Cardiac arrest

See Chapter 16.3 in main text

Clinical features	History
	1. Sudden collapse

	Examination 1. Patient unresponsive 2. Airway, breathing—no respiration or agonal breathing 3. Circulation—pulse not palpable
Immediate management	See Fig 1 and Fig 2

Fig. 1 European Resuscitation Council guidelines for adult basic life support.

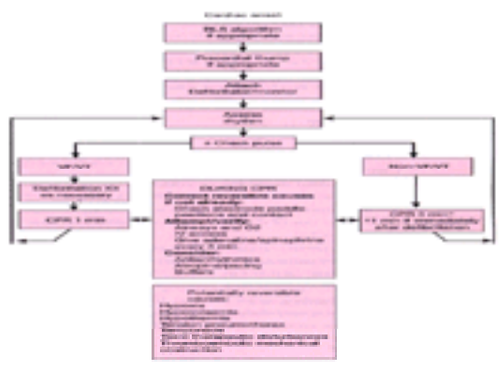

Fig. 2 European Resuscitation Council guidelines for advanced life support.

1.2 Cardiorespiratory collapse: the patient *in extremis*

See Chapter 16.1 in main text

Clinical features	**History**

A patient who is in extremis is unlikely to be able to give a lucid history and may die during (unwise) interrogation, but the following clues may be elicited and be very useful diagnostically:

1. Chest pain—suggests myocardial infarction or other cardiorespiratory catastrophe
2. Chest and back pain—dissection of thoracic aorta must be seriously considered
3. Abdominal pain—suggests ruptured abdominal aortic aneurysm or other intra-abdominal emergency
4. Recent surgery—pulmonary embolism likely
5. High fever/rigors—suggests infective cause
6. Recent travel to relevant area—malaria until proven otherwise

Examination
Airway and breathing 1. Is the airway patent? 2. Is the patient making a respiratory effort, and is the chest expanding with it? 3. Is the chest expanding symmetrically? Could there be a tension pneumothorax? (trachea deviated, mediastinum shifted, absent breath sounds on hyperinflated side of the chest, see Emergency Medicine section 2.3) 4. Widespread crackles in the chest—suggests pulmonary oedema in this context (see Emergency Medicine, section 1.8). 5. Does the patient look as though they could keep this breathing up for the next 10 min?—If not, the patient is very likely to need respiratory support. Call for assistance from the ICU immediately
Circulation 1. Do the peripheries feel cold or warm?—if warm, sepsis is likely 2. Pulse rate and rhythm—if rate <60/min or >120/min, consider whether arrhythmia is primary cause of hypotension 3. Blood pressure • Is there a postural drop if the patient is moved from lying to being propped up? If so, indicates intravascular volume depletion in this context. • Does BP fall substantially on inspiration? If so, indicates large intrathoracic pressure swings with breathing (likely in upper airway obstruction or asthma) or cardiac tamponade 4. What is the JVP? • If low, indicates intravascular volume depletion or dilated circulation • If high, suggests primary cardiorespiratory problem
General 1. Rash—purpura suggests meningococcal or other septicaemia 2. Temperature—high fever suggests infection 3. Loss of left radial pulse, or BP lower in left arm than right arm, indicates aortic dissection 4. Abdominal tenderness/peritonism—suggests ruptured abdominal aortic aneurysm or other intra-abdominal emergency See Table 1 for further information

Immediate management

Airway and breathing

1. Ensure airway is clear: consider oropharyngeal airway
2. Keep oxygen saturation >92 per cent (monitor using pulse oximetery), giving high flow oxygen (10 l/min) by face mask with reservoir bag if needed
3. If tension pneumothorax, decompress immediately (see Emergency Medicine, Section 10.4.3.1).
4. Give intravenous naloxone (0.8–2.0 mg repeated at intervals of 2–3 min to a maximum of 10 mg) if there is any suspicion that patient has received opioids
5. Consider elective intubation and ventilation

Circulation

Obtain IV access using a safe technique (see Emergency Medicine, section 10.1)

Also: begin resuscitation according to volume status as indicated in Table 2

1. Insert urinary catheter and monitor fluid input/output hourly in any patient with cardiorespiratory collapse.
2. Give broad spectrum antimicrobial cover to any patient with unexplained cardiorespiratory collapse, e.g. cefotaxime 1 g intravenously twice daily, as dictated by clinical suspicion of likely pathogen (see Emergency Medicine, section 7.7)

Key investigations See Table 1

Further management Determined by underlying condition.

1.3 Acute myocardial infarction (AMI)

See Chapter 15.4.2.3 and Chapter 15.4.2.4 in main text

Clinical features

History

1. Ischaemic chest pain
2. Cardiorespiratory collapse
3. May be non-specific or silent, especially in the elderly or in diabetics

Examination

May be normal, but look for

1. 'Pump failure'—cool peripheries, hypotension
2. Pulmonary oedema—see Emergency Medicine, section 1.8.
3. Cardiac—gallop rhythm, murmurs

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Otherwise

1. Give high flow oxygen by facemask.
2. Give aspirin 300 mg p.o. immediately, chewed or dispersed in water (if not given before admission to hospital)
3. Give adequate analgaesia, e.g. (i) diamorphine by slow intravenous injection at 1 mg/min, usual maximum initial dose is 5 mg, but may be repeated if necessary, or (ii) morphine by slow intravenous injection at 2 mg/min, usual maximum initial dose is 10 mg, but may be repeated if necessary. Both to be accompanied by appropriate antiemetic, e.g. metoclopramide 10 mg IV over 1–2 min, or cyclizine 50 mg IV over 1–2 min (caution in severe heart failure)
4. If appropriate, give thrombolysis as soon as possible (Table 3 and Table 4)
5. If appropriate, consider percutaneous intervention (Table 5)

Key investigations

To establish the diagnosis

1. ECG—looking for ST segment elevation and/or (presumed or proven) new bundle branch block
2. Cardiac biochemical markers (troponins, CK-MB)

Other important tests

1. Assess risk factors for ischaemic heart disease, e.g. cholesterol
2. As indicated by clinical examination, e.g. chest radiograph to look for pulmonary oedema; echo-cardiography to assess LV function or cause of pansystolic murmur (?mitral valve dysfunction, ?ventricular septal defect)

Further management

Consider

1. b-Blockade
   - Early—if no contraindication (e.g. hypotension, heart failure, heart block) give, e.g. atenolol 5 mg IV over 5 min, repeated after 10–15 min
   - Long term—if no contraindication continue oral b-blockade for at least 2–3 years.
2. Angiotensin converting enzyme inhibition
   - Early—start within 24 h in patients who are nor-motensive and continue for at least 5–6 weeks
   - Long term—recommended for any patient with left ventricular dysfunction
3. Long term aspirin (75–150 mg/day)
4. Long term statin (lipid lowering agent) will benefit most if not all patients after AMI

Note

1. Treat complications, e.g. venodilator or diuretic for pulmonary oedema. Severe heart failure/shock may require ventilation, inotropes +/- intra-aortic balloon pump
2. Diabetic patients will benefit from intensive insulin therapy during admission with AMI and afterwards
3. For all patients: give advice regarding lifestyle issues before and after discharge from hospital—smoking, diet, exercise—also regarding resumption of normal activities. Consider referral to cardiac rehabilitation services
4. Consider need for specialist cardiological opinion and/or investigation by cardiac stress test (e.g. treadmill exercise tolerance test) and/or coronary angiography

1.4 Unstable angina or non-ST segment elevation myocardial infarction (non-Q-wave myocardial infarction)

See Chapter 15.4.2.3 and Chapter 15.4.2.4 in main text

Clinical features

History

1. Ischaemic chest pain at rest or on minimal exertion
2. Chest tightness/breathlessness

Examination

Usually no specific signs, but may be

1. 'Pump failure'—cool peripheries, hypotension
2. Pulmonary oedema—breathing difficulty, pulmonary crackles (see Emergency Medicine, section 1.8)
3. Cardiac—gallop rhythm, murmurs

Immediate management

1. Give high flow oxygen by facemask.
2. Give aspirin 300 mg orally immediately, chewed or dispersed in water (if not given before admission to hospital)
3. Give thienopyridine, e.g. clopidogrel 300 mg orally (then 75 mg daily)
4. Give nitrate, e.g. (1) sublingual glyceryl trinitrate (GTN), 0.3–1 mg repeated as required; (2) buccal GTN, up to 5 mg, with tablet placed between upper lip and gum and left to dissolve; (3) intravenous infusion of isosorbide dinitrate at initial dose of 2 mg/h (increasing as necessary to maximum of 20 mg/h to relieve pain and as limited by hypotension)
5. If pain not relieved by nitrate give adequate analgaesia, e.g. (1) diamorphine by slow intravenous injection at 1 mg/min (usual maximum initial dose is 5 mg, but may be repeated if necessary), or (2) morphine by slow intravenous injection at 2 mg/min (usual maximum initial dose is 10 mg, but may be repeated if necessary). Both to be accompanied by appropriate antiemetic, e.g. metoclopramide 10 mg IV over 1–2 min, or cyclizine 50 mg IV over 1–2 min (caution in severe heart failure)
6. Give low molecular weight heparin, e.g. enoxaparin 1 mg/kg (100 units/kg) every 12 h, or dalteparin 120 units/kg every 12 h (maximum 10000 units twice daily), unless contraindicated
7. Give intravenous or oral b-blocker (see Emergency Medicine section 1.3) unless contraindicated
8. Consider heart-rate-lowering calcium antagonist (e.g. diltiazem or verapamil) if b-blocker is contra-indicated in patient without left ventricular dysfunction
9. Consider glycoprotein IIb/IIIa inhibitor in high risk groups, e.g. those with ST segment depression and/ or troponin positive, or those receiving urgent percutaneous intervention for unstable angina or non-ST segment elevation AMI. Agents tested in large scale randomized trials include abciximab, epti-fibatide, and tirofiban
10. Consider percutaneous intervention (Table 5)

Key investigations

To establish the diagnosis

1. ECG—looking for transient ST segment shift with pain; T wave changes are less specific and ECG may be normal
2. Cardiac biochemical markers (troponins, CK-MB)

Other important tests

As for acute myocardial infarction (see Emergency Medicine section 1.3)

Further management

1. Angiotensin converting enzyme inhibition— recommended for any patient with left ventricular dysfunction
2. Long-term aspirin (75–150 mg/day).
3. Long-term statin (lipid lowering agent) will benefit most patients with ischaemic heart disease

Consider:

4. Clopidogrel 75 mg/day

Note

1. For all patients: give advice regarding lifestyle issues before and after discharge from hospital—smoking, diet, exercise—also regarding resumption of normal activities. Consider referral to cardiac rehabilitation services
2. Consider need for specialist cardiological opinion and/or investigation by cardiac stress test (e.g. treadmill exercise tolerance test) and/or coronary angiography

1.5 Dissection of the thoracic aorta

See Chapter 15.14.1 in main text

Clinical features

History

1. Chest pain, particularly if of sudden onset, tearing in quality, and radiating to the back
2. Collapse

Examination

1. Patient will usually look very unwell: cool peripherally, hypotensive
2. Look for loss/reduction of one or more peripheral pulses: most likely is compromise of the left sub-clavian artery. Check left radial pulse in comparison with right; measure blood pressure in both arms; any deficit on the left strongly supports the diagnosis of aortic dissection. Examine also for reduction of carotid or femoral pulse(s)
3. Evidence of focal ischaemia, eg. focal neurological deficit ('stroke')
4. Could the patient have Marfan's syndrome? (risk factor)

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. The key to correct management is a high index of clinical suspicion that aortic dissection might be the diagnosis. Most patients with chest pain and circulatory collapse have acute myocardial infarction, the management for which (thrombolysis) could clearly be fatal in the patient with aortic dissection
2. Give high flow oxygen by facemask
3. Give adequate analgaesia, e.g. (i) diamorphine by slow intravenous injection at 1mg/min (usual maximum initial dose is 5mg, but may be repeated if necessary) or (ii) morphine by slow intravenous injection at 2 mg/min (usual maximum initial dose is 10 mg, but may be repeated if necessary). Both to be accompanied by appropriate antiemetic, e.g. metoclopramide 10 mg IV over 1–2 min, or cyclizine 50 mg IV over 1–2 min (caution in severe heart failure)

Key investigations

To establish the diagnosis

1. CT angiography of chest
2. Transoesophageal echocardiography

Other important tests

1. Chest radiograph—look for widened mediastinum
2. ECG—may have features of acute myocardial infarction (usually inferior) if dissection has compromised a coronary artery (usually right coronary artery)
3. Cardiac biochemical markers—to exclude acute myocardial infarction
4. Full blood count, clotting screen, electrolytes, renal and liver function tests—may give a lead to an underlying medical condition and will establish baseline
5. Group and save/crossmatch blood.

Further management

1. Reduce blood pressure using agents that will not cause tachycardia or increase the rate of cardiac ejection, e.g. titrate IV labetolol (initial dose 1 mg/min) or esmolol (50–200 µg/kg/min) to achieve SBP <110 mmHg. If blood pressure remains too high, add intravenous infusion of sodium nitroprus-side (0.5–8 µg/kg/min) after b-blockade established (pulse <60 /min)
2. Obtain opinion from cardiothoracic surgeon: immediate surgical repair will usually be the best management for patients with dissection of the ascending aorta (Stanford Type A) who are in reasonable condition

1.6 Bradycardia

See Chapter 15.2.3 and Chapter 15.6 in main text

Clinical features

History

1. Syncope or presyncope
2. Fatigue/breathing difficulty
3. Drugs (especially b-blockers)

Examination

The most important immediate issue is to decide whether or not the circulation is compromised: is the patient cool peripherally? What are the rate, rhythm, and blood pressure? Is there pulmonary oedema (see Emergency Medicine, section 1.8)?

If seen in the presence of bradycardia, note rate and

1. Abnormal rhythm, e.g. dropped beats in second degree AV block
2. Other cardiovascular abnormality, e.g. cannon waves in JVP in third degree (complete) AV block
3. Temperature (hypothermia—see Emergency Medicine, section 9.4)

Immediate management

Obtain ECG

If the patient is haemodynamically compromised

1. Give atropine, 0.3–1.0 mg IV, repeated as necessary
2. Consider isoprenaline, 0.5–10 µg/min by IV infusion
3. Consider temporary pacing (see Emergency Medicine, section 10.3)
4. Consider glucagon 50–150 µg/kg intravenously in 5 per cent glucose in cases of b-blocker overdose, with precautions to protect the airway in case of vomiting (NB unlicensed indication and dose)

Key investigations

To establish the diagnosis

12-lead ECG

Other important tests

1. Electrolytes (particularly potassium)
2. Cardiac biochemical markers (depending on context)
3. Chest radiograph—look at heart size and for evidence of pulmonary oedema
4. 24 h ECG monitor (if symptoms intermittent and 12-lead ECG not diagnostic)
5. Echocardiography (if clinical suspicion that heart is structurally abnormal)

Further management

Dependent on diagnosis. If not reversible likely to require permanent pacing

1.7 Tachycardia

See Chapter 15.2.3 and Chapter 15.6 in main text

Clinical features

History

1. Syncope or presyncope
2. Palpitations
3. Fatigue/breathing difficulty
4. Chest pain

Examination

The most important immediate issue is to decide whether or not the circulation is compromised: is the patient cool peripherally? What are the rate, rhythm and blood pressure? Is there pulmonary oedema (see Emergency Medicine, section 1.8)?

Physical examination is unlikely to aid diagnosis of the particular type of tachycardia, excepting for the presence of an irregularly irregular rhythm in atrial fibrillation. However, note the following:

1. Jugular venous pulse—absence of 'a' waves in AF; rapid flutter waves in atrial flutter; cannon waves in ventricular tachycardia
2. First heart sound—variable intensity in AF
3. A dilated heart increases the chance that tachycardia is ventricular in origin

Immediate management

Obtain ECG

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

FOR ARRHYTHMIAS THAT ARE POORLY TOLERATED, SYNCHRONISED DC SHOCK (UNDER GENERAL ANAESTHESIA OR DEEP SEDATION) USUALLY PROVIDES RAPID RELIEF.

Management otherwise depends upon clinical context and type of tachycardia

Key investigations

To establish the diagnosis

1. 12-lead ECG (see Table 6)

Other important tests

1. Electrolytes (particularly potassium)
2. Cardiac biochemical markers (depending on context)
3. Chest radiograph—look at heart size and for evidence of pulmonary oedema
4. 24 h ECG monitor (if symptoms intermittent and 12-lead ECG not diagnostic)
5. Echocardiography (if clinical suspicion that heart is structurally abnormal)
6. Thyroid function tests (in atrial fibrillation)

Further management

Uncertain of the diagnosis of a broad complex tachycardia? See Table 7

With severe haemodynamic compromise Atrial fibrillation/flutter

- DC cardioversion, or
- Amiodarone, 5 mg/kg over 20–120 min followed by 1200 mg/24hrs until sinus rhythm restored (into central venous catheter), or
- Sotalol, 1.5 mg/kg intravenously over 30 min.

Atrioventricular nodal re-entry (AVNRT) and atrioventricular reentry tachycardias (AVRT) (supraventricular tachycardias, SVTs)

- Adenosine, 3 mg intravenously given over 2 s, if necessary followed by 6 mg after 1–2 min, and then by 12 mg after a further 1–2 min (note— contraindicated in those with asthma, and patients taking dipyridamole are very sensitive, requiring reduced initial dose of 0.5–1 mg)
- Verapamil, 5–10 mg by slow intravenous injection over 2–3 min is an alternative in patients with asthma, but NOT in those who might have ventricular tachycardia, or in those who are receiving b-blockers

Ventricular tachycardia

- DC cardioversion (see Emergency Medicine, section 1.1)

Without severe haemodynamic compromise Atrial fibrillation/flutter

Duration <48 h or trans-oesophageal echocardiography shows no intracardiac thrombus

- Consider prompt chemical or synchronised DC cardioversion.
- Flecainide (Class 1C) 2 mg/kg intravenously over 30 min if there is no evidence of ischaemic heart disease or left ventricular dysfunction
- Amiodarone or sotalol (Class III) can be used to restore sinus rhythm and maintain it
- Digoxin is useful for rate control only but will not restore sinus rhythm. If digoxin is ineffective in controlling ventricular rate, and cardioversion is unsuccessful or inappropriate, consider adding verapamil or b-blocker

Duration >48 h or thrombus on trans-oesophageal echocardiography

- Anticoagulate for 4–6 weeks before synchronised DC cardioversion

Note

- Atrial fibrillation arising in the context of intercurrent illness is usually best managed by treatment of the underlying medical condition and with digoxin (plus or minus verapamil or b-blocker) to control ventricular rate. The patient is likely to return to sinus rhythm when the underlying condition has resolved

Atrioventricular nodal re-entry (AVNRT) and atrioventricular re-entry (AVRT) tachycardias (supraventricular tachycardias, SVTs)

- Vagal stimulation by respiratory manoeuvres (Valsalva), prompt squatting, or pressure over one carotid sinus (but not the latter in those with recent ischaemia, digitalis toxicity, or in the elderly)
- Adenosine if vagal stimulation fails
- Other options include verapamil, b-blocker, flecainide, sotalol, or amiodarone

Ventricular tachycardia

- Consider synchronized DC cardioversion.
- Lignocaine (lidocaine) 50–100 mg as intravenous bolus over a few min followed immediately by infusion of 4 mg/min for 30 min, 2 mg/min for 2 h and then 1 mg/min up to 24 hr.
- Other antiarrhythmics that can be used includeamiodarone, sotalol, procainamide and disopyramide—but seek expert help.

Torsade de pointes

This form of ventricular tachycardia requires particular treatment

- Give magnesium sulphate, 8 mmol of magnesium over 10–15 min, repeated once if necessary
- If torsade is associated with bradycardia and pauses, consider isoprenaline infusion or overdrive atrial/ventricular pacing to increase heart rate

1.8 Pulmonary oedema

See Chapter 15.15.2.2 in main text

Clinical features

History

1. Breathing difficulty
2. Orthopnoea, paroxysmal nocturnal dyspnoea
3. Palpitations
4. Chest pain
5. Ankle oedema
6. Of any cardiac disorder

Examination

1. How unwell is the patient? If very ill, see Emergency Medicine, section 1.2
2. Respiratory rate, cyanosis, peripheral circulation (cold, clammy), pulse rate and rhythm (?arrhythmia, see sections 1.6 and 1.7), blood pressure, JVP (likely to be elevated), apex beat (displaced in congestive cardiac failure), heart sounds (gallop rhythm, murmurs), crackles and/or wheezes in chest, peripheral oedema (suggests biventricular failure in this context)

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. Sit the patient up
2. Give high flow oxygen via reservoir bag to achieve PaO_2 >92%
3. Give frusemide 40–80 mg intravenously If not improving rapidly
4. Give either
   - Diamorphine by slow intravenous injection at 1 mg/min (usual maximum initial dose is 5mg, but may be repeated if necessary), or
   - Morphine by slow intravenous injection at 2 mg/min (usual maximum initial dose is 10 mg, but may be repeated if necessary)
   - Both to be accompanied by appropriate antiemetic, e.g. metoclopramide 10 mg IV over 1–2 min (not cyclizine in severe heart failure)
5. Unload with intravenous nitrate, e.g. isosorbide dinitrate 2–20 mg/h
6. Consider elective ventilation: non-invasive or after endotracheal intubation

Key investigations

To establish the diagnosis

Chest radiograph

Other important tests

1. ECG—look for arrhythmia or acute myocardial infarction
2. Cardiac biochemical markers

Further management

Depending on clinical context

1. Acute myocardial infarction—see Emergency Medicine, section 1.3
2. Arrhythmia—see Emergency Medicine, sections 1.6 and 1.7
3. Acute mechanical cause—e.g. aortic incompetence, mitral regurgitation, ventricular septal defect—may require surgical intervention

1.9 Deep venous thrombosis and pulmonary embolus

See Chapter 15.15.3.1 in main text

Clinical features

History

Deep venous thrombosis

1. Calf/leg pain
2. Calf/leg swelling
3. Features to suggest PE

Pulmonary embolus

1. Shortness of breath, developing over hours, days, or (sometimes) weeks
2. Pleuritic chest pain, haemoptysis (lung infarction, peripheral emboli)
3. Circulatory collapse (massive PE)
4. Features to suggest DVT

Deep venous thrombosis and pulmonary embolus Risk factors—immobilization, recent surgery, previous episodes, malignancy, travel, family history etc.

Examination

Deep venous thrombosis

1. Calf/leg swelling—measure circumference 10 cm below tibial tuberosity: difference between sides of >1cm likely to be significant
2. Calf tenderness; palpable cord; positive Homan's sign
3. Dilated superficial veins; leg feels warmer than the other
4. Check for signs of PE
5. Consider alternative diagnoses—especially Baker's cyst, cellulitis, haematoma in muscle

Pulmonary embolus

1. May be no abnormal signs
2. Tachypnoea (70% of cases), crackles (50%), tachycardia (30%), pleural rub (<10%)
3. Circulatory collapse with cool peripheries, hypotension, and cyanosis. Look particularly for signs of right heart strain: elevated JVP, parasternal heave, S3 over right ventricle, loud P2
4. Check for signs of DVT
5. Consider alternative diagnoses—especially pneumonia, musculoskeletal pain

Note

1. Low grade fever is common in both DVT and PE
2. In cases of DVT or PE—perform rectal/pelvic examination (before discharge from hospital)

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2
If index of clinical suspicion for PE is high, start IV standard (unfractionated) heparin pending the results of investigation

Key investigations **To establish the diagnosis**

Tests commonly used to demonstrate the presence of thrombus/embolus are as follows:

- DVT—venous ultrasonography, contrast venography
- PE—lung ventilation/perfusion (VQ) scan, CT pulmonary angiogram, or pulmonary angiogram

Many patients referred for medical opinion have a low probability of having DVT or PE and not all require imaging to exclude DVT or PE. Follow management algorithms as follows:

- DVT—see Table 8
- PE—see Table 9

Other important tests

Pulmonary embolus

1. ECG—commonest abnormality is sinus tachycardia and/or non-specific ST segment or T wave abnormalities. Look for signs of right heart strain, e.g. T wave inversion in V1/V2, S1Q3T3, axis shift
2. Chest radiograph—look for atelectasis or pulmonary parenchymal abnormality, also pleural effusion. May be normal
3. Arterial blood gases—look for hypoxia; but normoxia does not exclude PE

Deep venous thrombosis and pulmonary embolus

1. Full blood count, electrolytes, renal and liver function tests—may give a lead to an underlying medical condition and will establish baseline
2. At a later stage a thrombophilia screen may be appropriate, also investigations dictated by clinical findings or investigations detailed above

Further management

1. Anticoagulation with standard (unfractionated) heparin (Table 10) *or* low molecular weight heparin (e.g. Tinzaparin 175 units anti-Factor Xa IU/kg subcutaneous o.d.) *until* oral anticoagulation (usually with warfarin, Table 11) is established
2. In cases with circulatory collapse consider thrombolysis, e.g.
   - Streptokinase by intravenous infusion of 250000 units over 30 min, then 100000 units/h for 12–72 h, OR
   - Alteplase 10 mg by intravenous infusion over 1–2 min, followed by 90 mg over 2 h (maximum 1.5 mg/kg in patients of <65kg)

Notes

1. No monitoring of low molecular weight heparin treatment is required
2. Methods of reversing anticoagulation are shown in Table 12

1.10 Cardiac tamponade

See Chapter 15.9 in main text

Clinical features **History**

1. There are no specific features to indicate this condition
2. Can follow acute myocardial infarction, aortic dissection, cardiac trauma (including iatrogenic with cardiac catheterization)
3. There may be evidence of a condition that can cause pericardial effusion, e.g. tuberculosis, cancer, advanced renal failure

Examination

The key to making this rare but very important (because treatable) diagnosis is to consider it in any patient with unexplained cardiorespiratory collapse. Look for:

Signs of tamponade

1. Grossly elevated JVP that rises (if its top can be seen) on inspiration (Kussmaul's sign)
2. Pulsus paradoxus—meaning an exaggerated fall in systolic blood pressure on inspiration (normal <10 mmHg), but a rapid screening test for severe cases is to ask 'does the radial pulse disappear on inspiration'?

Evidence of a (large) pericardial effusion, although these will NOT be present unless there is a pre-existing effusion

3. Increased area of cardiac dullness
4. Quiet heart sounds

Immediate management

If the patient is *in extremis* proceed as in Emergency Medicine, section 1.2
As soon as the diagnosis of cardiac tamponade is established, perform or arrange for immediate/urgent pericardial aspiration (see Emergency Medicine, section 10.2.3)
Give colloid, e.g. gelofusin 500 ml by rapid intravenous infusion, to support blood pressure

Key investigations **To establish the diagnosis**

1. Echocardiography.
   - The most sensitive test for the presence of pericardial fluid.
   - Diastolic collapse of right ventricle or right atrium and a striking increase in the amplitude of septal motion with respiration indicate severe circulatory embarrassment
2. Cytology and culture of pericardial fluid

Other important tests

1. Chest radiograph—look for globular heart (almost invariably with clear lung fields)
2. ECG—look for low voltage QRS complexes and electrical alternans (in large pericardial effusion) and for evidence of acute myocardial infarction

Further management

As determined by underlying condition

1.11 Accelerated ('malignant') hypertension

See Chapter 15.16.3 in main text

Clinical features

History

1. Headache
2. Blurring of vision
3. Drowsiness
4. Epileptic fits

Examination

1. Blood pressure—will usually be grossly elevated with diastolic pressure >130 mmHg, but note that accelerated hypertension can occur at lower pressures than this and the diagnosis is established not by a particular elevation of blood pressure but by signs of fibrinoid necrosis
2. Ocular fundi
   - Grade III retinopathy: flame-shaped superficial haemorrhages, 'dot and blot' haemorrhages, cotton wool spots (retinal microinfarcts), hard exudates
   - Grade IV retinopathy: as Grade III + papilloedema (Note that there is no difference in management or prognosis of patients with Grade III or Grade IV disease)
3. Urine—stix testing shows proteinuria and haematuria, microscopy may show red blood cell casts

Also look for signs of

4. Pulmonary oedema—see Emergency Medicine, section 1.8
5. Aortic dissection—see Emergency Medicine, section 1.5
6. Scleroderma—scleroderma renal crisis

Immediate management

In an uncomplicated case*:*

1. Admit to hospital
2. Bed rest
3. No smoking (causes an acute rise in blood pressure)
4. Aim to lower diastolic pressure into range 100–105 mmHg over 2–3 days using:
   - Atenolol 25–50 mg orally, or
   - Nifedipine 10–20 mg of modified release preparation orally (tablets, not sublingual)
   - Further dosing determined by response
   - Maximum initial fall in blood pressure should not exceed 25% of presenting value

   In a complicated case (aortic dissection, epileptic fitting, acute pulmonary oedema, oral medication not possible) use intravenous infusion of:
   - Labetolol, initial bolus of 20 mg, then at 20 mg/h, increased as necessary every 30 min to maximum of 120 mg/h, or
   - Sodium nitroprusside at initial dose of 0.25–0.5 µg/kg/min, increasing up to 8 µg/kg/min

Key investigations

To establish the diagnosis

Accelerated hypertension is a clinical diagnosis

Other important tests

1. ECG—looking for evidence of left ventricular hypertrophy and acute myocardial ischaemia
2. Chest radiograph—looking for heart size, pulmonary oedema, and (if chest/back pain) for aortic dissection
3. Electrolytes and renal function—if serum creatinine >250 µmol/l renal function is likely to deteriorate further (at least in the short term)
4. 'Autoimmune/vasculitic' serology—ANCA, ANA etc.—for evidence of multisystem disorder that can present with accelerated phase hypertension and which (if present) will require specific treatment
5. CT angiography of chest if aortic dissection suspected

Further management

When acute emergency is controlled, all patients that have suffered from accelerated phase hypertension require thorough investigation for secondary causes of hypertension

1.12 Anaphylactic shock

See Chapter 16.4 in main text

Clinical features

History

1. Facial, tongue or throat swelling
2. Stridor or wheeze
3. Sudden collapse
4. Premonitory aura—apprehension, light-headedness, dizziness, tingling or itching of skin
5. Exposure to preciptant

Examination

1. Cyanosis
2. Hypotension
3. Facial, tongue, or throat swelling
4. Stridor or wheeze
5. Urticaria, angio-oedema, skin erythema, or extreme pallor

Immediate management

1. High flow oxygen (10 l/min) by face mask with reservoir bag to keep PaO_2 >92%.
2. Adrenaline (epinephrine)
   - Give 0.3–0.5 ml of 1:1000 (0.3–0.5 mg) intramuscularly, repeated every 5–10 min as needed
     If this is ineffective, or if the patient is about to die
   - Give 1–4 mg (1–4 ml) of 1:1000 adrenaline nebulized with oxygen, and
   - Make up 1:100000 preparation of adrenaline by drawing up 1ml of 1:1000 adrenaline (total of 1 mg) in 20 ml syringe, adding 9 ml of 0.9% saline to give total volume of 10 ml. Discard all but 2 ml (leaving 200 µg of adrenaline in the syringe), and then draw up further saline to a total volume of 20 ml, giving a final concentration of 10 µg/ml. Give 0.75–1.5 mg/kg of 1:100000 adrenaline IV at 10–20 µg/min (1–2 ml/min) initially, repeated as necessary
3. Colloid—give 10–20 ml/kg as rapid intravenous infusion if patient is hypotensive
   Second line therapy, after cardiorespiratory stability has been achieved:
4. Give H_1-blocker, eg. chlorpheniramine 10–20 mg IV, repeated up to 40 mg in 24 h (change to oral when patient tolerates)
5. Give H_2-blocker, e.g. ranitidine 50 mg IV three times daily (change to oral when patient tolerates)
6. Give hydrocortisone 5 mg/kg IV, then 2.5 mg/kg IV four times daily (change to oral prednisolone 40 mg daily when patient tolerates)
7. Give salbutamol 5 mg (repeated as necessary) via oxygen-driven nebulizer if bronchospasm is a persistent problem

Key investigations

To establish the diagnosis

1. Anaphylaxis is a clinical diagnosis
2. Mast cell tryptase

Other important tests

1. ECG, chest radiograph, electrolytes, renal function, arterial blood gases (depending on context)

Further management

1. Determination of allergen (if any)
2. Advice regarding avoidance
3. Medic Alert bracelet
4. Instruction regarding self-injection of adrenaline and supply of appropriate medication, e.g. EpiPen ™

2 Respiratory

2.1 Acute on chronic respiratory failure

See Chapter 17.6 and Chapter 17.7 in main text

Clinical features

History

1. Chronic respiratory condition—usually chronic obstructive pulmonary disease
2. Recent increase in breathlessness
3. Evidence of infection—fever, sweats, increased sputum production, increased sputum purulence
4. 'Cor pulmonale'—worsening ankle oedema

Examination

1. Cyanosis
2. Respiratory rate
3. Temperature
4. Evidence of carbon dioxide retention—drowsiness, asterixis, metabolic flap
5. Chest signs—of chronic respiratory condition, of infection, and exclude pneumothorax
6. Signs of cor pulmonale—elevated JVP, right ventricular heave, right ventricular gallop, loud P2, congested liver, ascites, peripheral oedema
7. Check PEFR if patient is able to use PEF recorder
8. Check pulse oximetry.

Is the patient getting exhausted? Remember that a 'normal' respiratory rate in the patient who looks very tired may mean that they are close to death.

Immediate management

The patient who is extremely ill

If the patient is *in extremis,* proceed as in Emergency Medicine, section 1.2, with the exception that a high concentration of inspired oxygen should NOT be given to patients who are KNOWN to have acute on chronic respiratory failure. If the patient is known to have chronic respiratory failure:

1. Give controlled oxygen (24–28% or 1–2 l/min by nasal prongs)
2. Initiate other aspects of management listed below
3. Check arterial blood gases, adjusting inspired oxygen concentration if allowed by clinical response, PaO_2, $PaCO_2$, and pH (pH, not hypoxia, is the most important factor related to survival in patients with acute on chronic respiratory failure)
4. Consider need for urgent intubation and ventilation if matters do not improve rapidly

Note

If it is UNCERTAIN whether or not a patient has acute on chronic respiratory failure, then high concentration oxygen should be given to all patients who are extremely ill. All such patients require continued close monitoring of their clinical state and arterial blood gases, allowing (amongst other things) detection of the few who will have acute on chronic respiratory failure and lose their respiratory drive in response to high concentration oxygen

The patient who is moderately unwell

1. Give 24% oxygen, increasing concentration to 28% (or 1 to 2 l/min by nasal prongs) depending on the results of subsequent blood gas analysis
2. Give nebulized b_2-agonist, e.g. salbutamol 2.5–5 mg, terbutaline 5–10 mg, using air as the driving gas, repeated as required
3. Give nebulized anticholinergic, e.g. ipatropium bromide 500 µg (can be combined with b_2-agonist), repeated as required
4. Give diuretic, e.g. frusemide 40–80 mg IV, if evidence of fluid overload
5. Give corticosteroid, e.g. hydrocortisone 100 mg IV twice daily or prednisolone 30 mg orally once daily
6. Give antibiotic that will cover likely respiratory pathogens if two of the following symptoms are present—increased breathlessness, increased sputum volume, or increased sputum purulence, e.g. amoxycillin 250 mg orally three times daily or (if allergic to penicillin) clarithromycin 250–500 mg orally twice daily (intravenously if oral administration not possible)
7. Consider aminophylline, loading dose (in patient not previously treated with theophylline) of 5 mg/kg given intravenously over 20 min, then an infusion of 0.5 mg/kg/h aiming for serum concentration in the range 10–20 mg/l
8. Consider need for non-invasive positive pressure ventilation if patient does not improve.

Note—use intravenous fluids to correct and prevent dehydration

Key investigations

To establish the diagnosis

1. Chest radiograph—looking for focal consolidation and to exclude pneumothorax
2. Sputum culture To determine severity and monitor response to treatment
3. Arterial blood gases
4. Serial measurements of peak flow

Other important tests

1. Full blood count
2. Electrolytes, renal and liver function
3. ECG

Further management

1. Optimization of treatment for chronic pulmonary condition, usually chronic obstructive pulmonary disease
2. Emphasize need to stop smoking

2.2 Tension pneumothorax

See Chapter 16.2 and Chapter 17.12 in main text

Clinical features	**History** 1. Collapse with extreme difficulty in breathing **Examination** 1. Patient looks as though they are about to die 2. Gasping respiratory effort 3. Cyanosis 4. Chest looks asymmetrical, being prominent on side of tension 5. Tracheal deviation, away from side of tension 6. Mediastinal shift, away from side of tension, most reliably detected by percussion of cardiac dullness 7. Chest is silent on side of tension, the only breath sounds being heard in the opposite axilla
Immediate management	Insert needle to decompress chest, see Emergency Medicine, section 10.4.3
Key investigations	**To establish the diagnosis** 1. Tension pneumothorax is a clinical diagnosis to be treated immediately without delay for investigation **Note** 1. The signs of tension pneumothorax are not subtle, but you will not make the diagnosis unless you consider it and seek the presence of the signs listed above 2. If a patient appears to be dying and you think that they might have a tension pneumothorax, then— after calling for help and initiating resuscitation (see Emergency Medicine, section 1.1)—there is nothing to be lost (and potentially much to be gained) from an attempt at chest decompression **Other important tests** Chest radiograph will confirm diagnosis of pneumothorax after decompression
Further management	Insertion of chest drain (see Emergency Medicine, section 10.4.3.3) after tension has been relieved

2.3 Upper airway obstruction

See Chapter 17.8.1 in main text

Clinical features	**History** 1. Extreme difficulty in breathing 2. Coughing /choking 3. Noisy breathing 4. Difficult/unable to speak 5. 'Something stuck' **Examination** 1. Extreme but ineffective respiratory effort 2. Cyanosis 3. Drooling (cannot swallow saliva) 4. Stridor
Immediate management	1. Heimlich manoeuvre if the patient has inhaled a foreign body • Patient sitting or standing—rescuer stands or kneels behind patient, encircling the patient's waist with their arms, placing one fist just above the navel (well below xiphoid process) and using their other hand to press the fist into the patient's abdomen with a quick upward thrust. Repeat as necessary • Patient lying—place patient on their back. Rescuer kneels astride patient and puts the palm of one hand between the navel and xiphisternum, places their other hand on top of this, and pushes upwards and inwards 2. If Heimlich manoeuvre is inappropriate or has failed • If there is time and you have the expertise—spray the pharynx with local anaesthetic (e.g. 5% cocaine and adrenaline) and examine the pharynx and upper airway by indirect laryngoscopy to establish the cause of obstruction and allow (if possible) its removal (with finger sweep under direct vision or long handled forceps) or passage of an endotracheal tube • If there is time and you are not experienced in upper airway management—call immediately for help from anaesthetic or ENT colleagues
Key investigations	**To establish the diagnosis** 1. Upper airway obstruction is a clinical diagnosis **Other important tests** As dictated by cause of obstruction
Further management	As dictated by cause of obstruction See Emergency Medicine, section 10.4.2.3

2.4 Asthma

See Chapter 17.4.4 in main text

Clinical features	**History** 1. Worsening asthma 2. Increasing difficulty breathing 3. Decrease in exercise tolerance 4. Increasing wheeze 5. Chest tightness 6. Cough 7. Difficulty in speaking 8. Fall in self-monitored peak flow 9. Failure to obtain improvement with use of regular b_2-agonist

Precipitating factor

1. Exposure to known precipitant, eg. exercise, cold air, dusty environment
2. Respiratory infection, e.g. upper respiratory tract infection

Examination

Moderate uncontrolled acute asthma

1. Breathlessness
2. Wheeze
3. Chest tightness
4. Peak flow 50–70% of predicted or personal best

Acute severe attack

1. Cannot complete sentences in one breath
2. Increased respiratory rate: >25 breaths/min
3. Use of accessory muscles of respiration
4. Tachycardia: >110/min
5. Peak flow <50% of predicted or personal best

Life-threatening asthma

1. Exhaustion or poor respiratory effort
2. Inability to speak.
3. Altered level of consciousness.
4. Cyanosis
5. Silent chest.
6. Hypotension or bradycardia
7. Peak flow <33% of predicted or personal best (or unrecordable)

Note

1. A 'normal' respiratory rate is consistent with the patient being near to death if they are exhausted
2. Always check carefully for signs of pneumothorax
3. Always check pulse oximetry
4. Asking the patient to count out loud as far as they can on a single breath provides a rapid, quantitative, and repeatable measure of respiratory function

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Moderate uncontrolled acute asthma

1. b_2-Agonist via spacer and mask or nebulizer (see below)
2. Oral prednisolone 30 mg once daily
3. Inhaled steroids—commence or increase dose

Acute severe attack

1. Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%
2. Salbutamol 5 mg or terbutaline 10 mg via oxygen-driven nebulizer, repeated up to every 15–30 min as needed, and then 4 hourly
3. Steroids—hydrocortisone 200 mg intravenously four times daily or prednisolone 30–60 mg orally once daily

Life-threatening attack or patient failing to improve

4. Add ipatropium 0.5 mg to nebulized b_2-agonist
5. Give aminophylline, loading dose (in patient not previously treated with theophylline) of 5 mg/kg given intravenously over 20 min, then an infusion of 0.5 mg/kg/h aiming for serum concentration in the range 10–20 mg/l. Omit loading dose if patient already taking oral theophylline

Consider intravenous salbutamol (3–20 µg/min) or terbutaline (1.5–5 µg/min) infusion

Note

1. IF THE PATIENT IS DETERIORATING, CALL FOR HELP FROM THE INTENSIVE CARE UNIT SOONER RATHER THAN LATER—ELECTIVE INTUBATION AND VENTILATION IS BETTER THAN THAT DONE AFTER CARDIORESPIRATORY ARREST (SEE CHAPTER 16.1)
2. Use intravenous fluids to correct and prevent dehydration

Key investigations **To establish the diagnosis**

Acute asthma is a clinical diagnosis

Other important tests

1. Chest radiograph—exclude pneumothorax
2. Arterial blood gases—markers for life-threatening asthma being:
   - Normal or high $PaCO_2$ (>5 kPa)
   - Low pH
   - Severe hypoxia (*P*aO2 <8 kPa) in spite of high flow oxygen treatment
3. Electrolytes, renal and liver function, full blood count

Further management

1. Optimization of long-term asthma management
2. Education regarding how to recognize severe attacks and how to respond when they develop

2.5 Pneumonia

See Chapter 17.5.2.1 in main text

Clinical features

History

1. Breathing difficulty
2. 'flu-like prodrome
3. High fever, sweats, rigors
4. Pleuritic chest pain
5. Sputum production (but note that this is not expected in atypical pneumonia)
6. Travel
7. Pet birds

Examination

1. Fever
2. Respiratory—cyanosis, respiratory rate, focal lung signs (consolidation, pleural rub, pleural effusion)
3. Circulation—peripheral perfusion (hot or cold), pulse, blood pressure
4. Look at the sputum (if any)
5. Always check pulse oximetry

British Thoracic Society definition of severe pneumonia states that one or more of the following must be present in a patient with clinical and/or radiological signs of pneumonia:

1. Respiratory rate >30/min
2. Systolic blood pressure <90 mmHg
3. PaO_2 <8 kPa (breathing room air)
4. Multilobar involvement on chest radiograph
5. Blood urea >7 mmol/l

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%
2. Appropriate antimicrobial agent

British Thoracic Society guidelines for treatment of community-acquired pneumonia

- Mild/moderate pneumonia
- Oral therapy with extended spectrum penicillin (eg. amoxicillin 250–500 mg three times daily) alone or with a macrolide (eg. clarithromycin 250–500 mg twice daily). Omit the penicillin in patients with penicillin allergy
- Severe pneumonia
- Intravenous therapy with a second- or third-generation cephalosporin (e.g. cefotaxime 1g twice daily) plus a macrolide (e.g. erythromycin 500 mg four times daily)
- Suspected legionnaire's disease
- High-dose intravenous erythromycin (1g four times daily) plus consider adding oral rifampicin (0.6–1.2 g daily in two to four divided doses)

In areas/countries where there is serious concern that *S. pneumoniae* may be resistant to penicillin and other agents

- Mild/moderate pneumonia
- Second- or third-generation cephalosporin (e.g. cefotaxime 1g intravenously twice daily) plus macrolide (e.g. erythromycin 500 mg orally or intravenously four times daily), OR fluoroquinolone (e.g. levofloxacin 500 mg orally or intravenously once or twice daily) alone
- Severe pneumonia
- Second/third-generation cephalosporin (e.g. cefotaxime 1 g intravenously twice daily) plus macrolide (e.g. erythromycin 500 mg intravenously four times daily), OR second-/third-generation cepha-losporin (e.g. cefotaxime 1g intravenously twice daily) plus fluoroquinolone (e.g. levofloxacin 500 mg intravenously twice daily)

Note

1. If staphylococcal pneumonia is suspected, add flucloxacillin 1g intravenously four times daily
2. See Chapter 17.5.2.2 and Chapter 17.5.2.3 for discussion of antimicrobial treatment of patients with hospital-acquired pneumonia or pneumonia in the immuno-compromised
3. Intravenous fluids to maintain adequate hydration

Key investigations

To establish the diagnosis

1. Chest radiograph—looking for focal consolidation (lobar pneumonia) or more widespread interstitial shadowing.
2. Blood culture.
3. Sputum culture.
4. Blood sample for serological testing.

To establish severity

Arterial blood gases—if patient is very ill or pulse oximetry shows PaO_2 <92%

Other important tests

Full blood count, electrolytes, renal and liver function

Further management

Follow up chest radiograph to ensure complete resolution

3 Gastrointestinal and hepatological

3.1 Upper gastrointestinal haemorrhage

See Chapter 14.3.2 in main text

Clinical features

History

1. Haematemesis or 'coffee-ground' vomiting
2. Melaena
3. Presyncope
4. Indigestion or reflux or medication for these symptoms
5. Retching before haematemesis (consider Mallory Weiss tear)
6. Previous upper gastrointestinal investigation or surgery
7. To suggest recent development of anaemia
8. Drugs that predispose to upper gastrointestinal haemorrhage—aspirin, non-steroidal anti-inflammatory agents, anticoagulants
9. Risk factors for, or presence of, chronic liver disease (consider varices)
10. Anorexia and weight loss (consider malignancy)

Examination

1. State of circulation—temperature of peripheries, pulse rate, blood pressure, jugular venous pressure
2. Mucous membranes—chronic anaemia
3. Evidence of chronic liver disease—jaundice and other manifestations (consider varices)
4. Evidence of portal hypertension—especially splenomegally (consider varices)
5. Lymphadenopathy—especially in left supraclavicular fossa (consider malignancy)
6. Abdomen—for epigastic mass (consider malignancy)
7. Rectal examination—for blood/melaena

Notes

1. The most reliable signs of intravascular volume depletion are postural hypotension (sitting versus lying) and a low jugular venous pressure
2. Clinical assessment of severity, see Table 13

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. Establish intravenous access with one or more large-bore peripheral venous cannulae (look in the antecubital fossae in the patient who is shut down). If you cannot do this, then insert femoral venous catheter (see Emergency Medicine, section 10.1.3). DO NOT ATTEMPT TO INSERT AN INTERNAL JUGULAR OR SUBCLAVIAN VENOUS CATHETER INTO A PATIENT WHO OBVIOUSLY HAS SEVERE INTRAVASCULAR VOLUME DEPLETION (see Chapter 16.1 for discussion).
2. If clinical evidence of intravascular volume depletion, give 1000 ml of intravenous fluid (colloid, e.g Gelofusin ™, or 0.9% saline) as fast as possible. Repeat clinical examination. If the patient still has intravascular volume depletion, give further 500 ml of fluid as fast as possible. Repeat cycle until arterial pressure and jugular venous pressure restored towards normal, then slow down rate of infusion. Use blood instead of colloid/saline as soon as it is available
3. Cross-match blood for transfusion
4. Consider need for urgent upper gastrointestinal endoscopy
5. If oesophageal varices—see Table 14

Also

1. Keep oxygen saturation >92% (monitor using pulse oximetery), giving high flow oxygen (10 l/min) by face mask with reservoir bag if needed
2. Insert urinary catheter and monitor fluid input/output hourly in any patient with substantial gastrointestinal haemorrhage—a satisfactory urine output is the best gauge of adequate resuscitation
3. Correct any coagulopathy—see Emergency Medicine, section 1.9, Table 3 (iatrogenic overanticoagulation) and section 9.1 (disseminated coagulation)
4. Nurse to avoid aspiration, and do not insert nasogastric tube, which makes this more likely

Key investigations

To establish the diagnosis (and also potentially therapeutic)

1. Upper gastrointestinal endoscopy
   - Within 24 h of admission in anyone with a substantial gastrointestinal bleed
   - Urgently if oesophageal varices are suspected or the patient is actively bleeding

See Table 13 for assessment of risk of rebleeding and mortality after endoscopy

Other important tests

1. Full blood count—but remember that the initial haemoglobin concentration is a poor estimate of the volume of acute blood loss
2. Electrolytes, renal and liver function tests
3. Coagulation screen
4. To pursue possibility and causes of chronic liver disease (if clinically indicated)

Further management

1. Inform surgical colleagues of all cases of substantial gastrointestinal haemorrhage immediately
2. Dependent on cause of haemorrhage, e.g.
   - Acid suppression for ulcer healing—high dose intravenous proton pump inhibitor, e.g. omeprazole 80 mg bolus followed by 8 mg/h
   - Eradication of *H. pylori*

3.2 Lower gastrointestinal haemorrhage

See Chapter 14.3.2 in main text

Clinical features

History

1. Haemorrhoids
2. Abdominal pain—if long-standing and intermittent may suggest diverticular disease, if severe may indicate mesenteric ischaemia
3. Previous lower gastrointestinal investigation or surgery
4. To suggest recent development of anaemia
5. Anorexia, weight loss, recent alteration in bowel habit (consider malignancy)
6. Drugs that predispose to gastrointestinal haemorrhage—aspirin, non-steroidal anti-inflammatory agents, anticoagulants
7. Risk factors for, or presence of, chronic liver disease (consider rectal varices)
8. Family history—colonic polyps/neoplasia, hereditary haemorrhagic telangiectasia

Examination

1. State of circulation—temperature of peripheries, pulse rate, blood pressure, jugular venous pressure
2. Mucous membranes—chronic anaemia
3. Jaundice—suggests malignancy or chronic liver disease
4. Lymphadenopathy—suggests malignancy
5. Abdomen—for localized tendeness, peritonism or palpable mass
6. Rectal examination—for piles and blood
7. Peripheral vasculature—generalized disease increases likelihood of mesenteric ischaemia
8. Telangiectasiae on skin or mucosae

Notes

The most reliable signs of intravascular volume depletion are postural hypotension (sitting versus lying) and a low jugular venous pressure

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. Establish intravenous access—as Emergency Medicine, section 3.1
2. If clinical evidence of intravascular volume depletion, resuscitate as described in Emergency Medicine, section 3.1
3. Cross-match blood for transfusion

Also

4. Keep oxygen saturation >92% (monitor using pulse oximetery), giving high flow oxygen (10 l/min) by face mask with reservoir bag if needed
5. Insert urinary catheter and monitor fluid input/output hourly in any patient with substantial gastrointestinal haemorrhage—a satisfactory urine output is the best guage of adequate resuscitation
6. Correct any coagulopathy—see Emergency Medicine, Section 1.9, Table 3 (iatrogenic overanticoagulation) and Section 9.1 (disseminated intravascular coagulation)

Key investigations

To establish the diagnosis

In all patients

1. Proctoscopy and rigid sigmoidoscopy

As required:

2. Colonoscopy
3. Mesenteric angiography

Other important tests

1. Full blood count
2. Electrolytes, renal and liver function tests, coagulation screen, inflammatory markers
3. To pursue possibility and causes of chronic liver disease (if clinically indicated)

Further management

1. Inform surgical colleagues of all cases of substantial gastrointestinal haemorrhage immediately
2. Dependent on cause of haemorrhage

3.3 Acute colitis

See Chapter 14.11 and Chapter 14.17 in main text

Clinical features

History

1. Bowel motions—frequency and type (blood, mucus, pus)
2. Abdominal pain
3. Rapidity of onset
4. Systemic features—fever, malaise, anorexia
5. Previous episodes/known colitic disease
6. Recent diet (contaminated or infected food)
7. Have close contacts also been ill?
8. Recent antibiotic treatment (consider *C. difficile*)
9. Use of non-steroidal anti-inflammatory agents
10. Associated vomiting
11. Travel
12. Risk factors for HIV (in some cases)

Examination

1. State of circulation—temperature of peripheries, pulse rate, blood pressure, jugular venous pressure
2. Signs of toxicity—fever
3. Mucous membranes—chronic anaemia, ulceration, Candida
4. Abdomen—for distension, localized tenderness, peritonism or palpable mass, or altered bowel sounds (absent, reduced)
5. Rectal and perineal examination—for fistulae and nature of stool (blood, pus)
6. Peripheral vasculature—generalized disease increases likelihood of ischaemic colitis
7. Peripheral oedema—suggests hypoproteinaemia and chronic disease in this context

Notes

The most reliable signs of intravascular volume depletion are postural hypotension (sitting versus lying) and a low jugular venous pressure

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. Fluid and potassium resuscitation as necessary—see Emergency Medicine, section 3.1 and section 5.5.
2. Most cases of acute colitis do not require antimicrobial therapy and settle with rehydration and time, the results of stool culture and rectal biopsy (which should be available in 24–48 h) being used to guide further treatment decisions. However, patients who are very ill with marked systemic symptoms and bloody diarrhoea (indicating probable colitis) should be given antimicrobial therapy pending culture results. Treat empirically with, e.g. ciprofloxacin (500–750 mg orally twice daily, or 200–400 mg intravenously twice daily) and metronidazole (400 mg orally three times daily or 500 mg intravenously three times daily).
3. Also, in cases of known colitis (and to be considered in those with new and undiagnosed presentations of colitis), give steroids to those who are very ill, e.g. hydrocortisone 100 mg intravenously every 6 h or prednisolone 60 mg orally once daily

Note the features of a severe acute attack of ulcerative colitis (Table 15)

Key investigations

To establish the diagnosis

In all patients

1. Abdominal radiograph—to assess extent of inflammation and to exclude toxic megacolon (required before proctoscopy/sigmoidoscopy), and erect chest radiograph—looking for air under diaphragm (perforation)
2. Flexible or rigid sigmoidoscopy and rectal biopsy
3. Stool—microscopy, culture and testing for *C. difficile* toxin
4. Blood cultures

Other important tests

1. Full blood count
2. Group and save or crossmatch blood
3. Electrolytes, renal and liver function tests, inflammatory markers, coagulation screen

Further management

1. Inform surgical colleagues of all cases of acute colitis, urgently if radiography shows perforation or toxic dilatation
2. Nurse in side room (if possible) until infective cause excluded
3. Further management dependent on cause of colitis
4. Note that suspected or proven food poisoning and typhoid are notifiable diseases in the UK

3.4 Acute hepatic failure

See Chapter 14.21.3 in main text

Clinical features

Definitions

1. Acute hepatic failure is hepatocellular jaundice, hypertransaminasaemia, and prolongation of the prothrombin time associated with an acute liver disease
2. Fulminant hepatic failure is acute liver failure with hepatic encephalopathy, most definitions specifying that this must occur within a particular time (variable) from the onset of clinical evidence of liver disease (usually jaundice)

History

1. Jaundice—not always present in fulminant hepatic failure
2. Confusion/drowsiness—note timing of onset of mental changes in relation to jaundice
3. Relevant to cause of acute liver failure, e.g. paracetamol overdose, full drug history (prescribed and non-prescribed), risk factors for viral hepatitis
4. Is there a background of chronic liver disease?—alcohol, risk factors for viral hepatitis
5. Autoimmune conditions (associated with autoimmune chronic active hepatitis)
6. Family history (Wilson's is a rare cause of fulminant hepatic failure)

Examination

1. State of circulation—vital signs are normal in the early stages. Tachycardia and hypotension occur later. Hypertension and bradycardia are very late and sinister signs of cerebral oedema
2. Jaundice
3. Liver—usually tender, but normal size or only slightly enlarged in acute hepatic failure. If hepatomegaly consider hepatic venous obstruction (Budd Chiari), malignant infiltration, chronic liver disease
4. Ascites—if substantial consider Budd Chiari
5. Encephalopathy
   - Grade 1—mild confusion, irritability, decreased attention
   - Grade 2—drowsiness, lethargy, inappropriate behaviour
   - Grade 3—somnolent but rousable, disorientated
   - Grade 4—coma
6. Signs of chronic liver disease

Notes

Focal neurological signs are not expected in acute hepatic failure. If present they suggest a focal cerebral lesion, most likely haemorrhage in this context

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Acute hepatic failure

1. Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92% (monitor with pulse oximetry)
2. Treat/prevent hypovolaemia—give 4.5% serum albumin intravenously to keep CVP at +10 cm of water
3. Treat/prevent hypoglycaemia—give 50% glucose intravenously (central line) at 5–10 ml/h (monitor BM stix regularly)
4. *N*-Acetyl cysteine by intravenous infusion
   - Paracetamol overdose 150 mg/kg in 200 ml 5% dextrose over 15 min, then 50 mg/kg in 500 ml 5% dextrose over 4 h, then 100 mg/kg in 1000 ml 5% dextrose over 16 h
   - Other diagnosis 150 mg/kg in 1000 ml 5% dextrose over 24 h
5. Give prophylactic broad spectrum antibiotic, eg. cefotaxime 1 g intravenously twice daily
6. Give prophylaxis against gastrointestinal stress ulceration, e.g. ranitidine (150 mg orally twice daily, 50 mg intravenously three times daily)

Hepatic encephalopathy To prevent or treat

1. Removal or correction of precipitating factors
   - Drugs—stop all if possible, particularly sedatives/hypnotics and diuretics
   - Fluid and electrolyte balance—maintain carefully. Avoid/treat dehydration, hypoglycaemia, hypokalaemia, hypophosphataemia
2. Minimize absorption of nitrogenous substances

   The following treatments may or may not be given

   - Give enemas ($MgSO_4$ or phosphate) to encourage bowel emptying
   - Give disaccharide laxative, e.g. lactulose 30–50 ml three times daily, dosage then adjusted to produce 2–3 soft stools daily
   - Give broad spectrum poorly-absorbed antibiotic, e.g. neomycin 1 g four times daily by mouth
3. If Grade 3 or 4 encephalopathy, also
   - Intubate and ventilate
   - Give parenteral feeding

Notes

1. Hyponatraemia is common and due to water excess rather than sodium deficiency. It should be treated with fluid restriction and not by infusion of saline
2. If there is a history of chronic high alcohol intake or malnourishment, give thiamine intravenously BEFORE giving glucose to avoid risk of precipitating Wernicke's encephalopathy, e.g. Pabrinex ™ intravenous high potency injection, 10 ml (2 ampoules) over 10 min (repeated three times daily)
3. Insert urinary catheter and monitor fluid input/output hourly in any patient with acute hepatic failure
4. Cerebral oedema
   - Avoidance—Avoid overfilling with intravenous fluids
   - Treatment—Nurse in quiet room with trunk and head elevated at 40°; consider transfer to facility where intracranial pressure can be monitored; consider mannitol 1 g/kg as intravenous bolus of 20% solution (if plasma osmolality <315 mosmol/ kg and the patient is not oliguric), repeated 4 hourly (0.5 g/kg) if previous infusion induced a diuresis

Key investigations

To establish the presence of acute liver failure

1. Liver blood tests—bilirubin, transaminases (ALT, AST, gGT)
2. Prothrombin time/coagulation screen

To establish the cause of liver disease

If no history of paracetamol overdose

1. Hepatitis B core IgM, hepatitis A IgM, liver autoantibodies, immunoglobulins
2. Abdominal ultrasound and Doppler of hepatic veins—looking for size/echogenicity of liver, splenomegaly, signs of Budd Chiari
3. If <40 years: serum copper and caeruloplasmin; ophthalmic examination for Kayser-Fleischer rings (Wilson's disease)

Note

1. Tap ascites if present—microscopy, culture, and sensitivity. Culture/swab blood, urine, nasal, high vagina
2. Do not correct coagulopathy unless the patient is bleeding: the prothrombin time is an important prognostic indicator
3. Where the prothrombin time (in s) is greater than the time after a paracetamol overdose (in h), there is a substantial risk of developing acute liver failure

Other important tests

1. Full blood count
2. Glucose, renal function tests, amylase
3. Arterial blood gases

Further management

1. Discuss all cases of acute hepatic failure with a specialist (transplant) centre (see Table 16): urgent orthotopic liver transplantation may be required and appropriate
2. Dependent on cause of hepatic failure

3.5 The acute abdomen

See Chapter 14.3.1 in main text

Clinical features

History

1. Abdominal pain—duration, constant or colicky, where is it worst (point with one finger), radiation
2. Gastrointestinal symptoms—anorexia, nausea, vomiting, diarrhoea, constipation (precisely when were the bowels last open), blood in vomit or stool
3. Urinary symptoms—frequency, pain on micturition, haematuria
4. Gynaecological symptoms—last menstrual period, vaginal discharge
5. To suggest sepsis—sweats, fevers, rigors
6. History of gastrointestinal problems—indigestion, peptic ulceration, gallstones, pancreatitis
7. History of atheromatous vascular disease—ischaemic heart disease, cerebrovascular disease, peripheral vascular disease (increase the likelihood of bowel ischaemia or of abdominal aortic aneurysm, also relevant to surgical risk)

Examination

1. State of circulation—temperature of peripheries, pulse rate, blood pressure, jugular venous pressure
2. Signs of toxicity—fever
3. Foetor
4. Abdomen
   - Inspection—distension, movement on respiration
   - Palpation—tenderness, guarding, rigidity, rebound tenderness, palpable mass
   - Auscultation—bowel sounds
   - Check all hernial orifices and abdominal aorta
5. Rectal examination—for tenderness and nature of stool, blood in stool
6. Vaginal examination—tenderness, pelvic mass
7. Test urine for blood

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. Establish intravenous access—as Emergency Medicine, section 3.1
2. If clinical evidence of intravascular volume depletion, resuscitate as described in Emergency Medicine, section 3.1
3. Immediate liaison with surgical colleagues
4. Provide effective analgaesia e.g.
   - Non-steroidal anti-inflammatory agent: e.g. diclofenac 75 mg intramuscularly, repeated after 30 min if necessary
   - Opioid: e.g. morphine 5 mg subcutaneously plus 5 mg intramuscularly, repeated if necessary and accompanied by appropriate antiemetic, e.g. metoclopramide 10 mg IV over 1–2 min, or cyclizine 50 mg IV over 1–2 min
5. Nasogastric tube
6. Urinary catheter

Key investigations

To establish the diagnosis

1. Abdominal radiograph—is there intestinal obstruction?
2. Erect chest radiograph—is there gas under the diaphragm indicating intestinal perforation?
3. Serum amylase—a substantial increase suggests pancreatitis
4. Abdominal ultrasound—?free fluid/swollen appendix/ovarian cyst
5. Abdominal CT scan

Note

Patients with generalized peritonitis require an urgent LAPAROTOMY provided that pancreatitis has been excluded. DO NOT DELAY. If the patient requires resuscitation, then make arrangements for theatre whilst initiating resuscitation and continue to resuscitate in the anaesthetic room. Do not wait 'until the patient is a bit better' before involving anaesthetic and surgical colleagues

Other important tests

1. Full blood count
2. Group and save or crossmatch blood
3. Electrolytes, renal and liver function tests
4. Coagulation screen

Further management

Dependent on the cause of the acute abdomen

Note

1. Adhesive small bowel obstruction may resolve with conservative management
2. Remember rare 'medical' causes of abdominal pain, e.g. pneumonia, shingles, drugs (digoxin), diabetes, sickle cell crisis, porphyria, familial mediterranean fever … remember also that these are rare: if in doubt, diagnose a common condition

4 Renal

4.1 Acute renal failure

See Chapter 20.4 in main text

Clinical features

History

1. There are no specific features to suggest acute renal failure: presentation is dominated by the precipitating condition
2. Previous renal or urinary tract disease
3. Drugs, prescribed and non-prescribed
4. Evidence of multisystem disease
5. ALWAYS seek the results of previous tests of renal function

Examination

1. State of circulation—temperature of peripheries, pulse rate, blood pressure, jugular venous pressure
2. Evidence of infection—fever, localizing signs
3. Breathing—evidence of pulmonary oedema or acidosis (Kussmaul)
4. Abdominal—is the bladder palpable? (obstruction)
5. Rectal—is there pelvic malignancy? (obstruction)
6. General—signs indicating multisystem disorder: rash, joints, eyes, nose. Are muscles swollen/tender? (rhabdomyolysis)

Note

The most reliable signs of intravascular volume depletion are postural hypotension (sitting versus lying) and a low jugular venous pressure

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Treatment of life-threatening complications

1. Hyperkalaemia—see Emergency Medicine, section 5.4
2. Pulmonary oedema—see Emergency Medicine, section 1.8
3. Severe acidosis, causing circulatory compromise
4. 'Gross uraemia', causing encephalopathy or bleeding Aside from immediate life-saving medical treatments, patients with these features will need urgent renal replacement therapy (preferably by haemodialysis or haemofiltration, as dictated by clinical context) unless their renal function can be restored rapidly

Optimization of intravascular volume—many patients presenting with acute renal failure will be volume deplete

1. Establish intravenous access—as Emergency Medicine, section 3.1
2. If clinical evidence of intravascular volume depletion, resuscitate as described in Emergency Medicine, section 3.1

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

Make diagnosis of cause of renal failure

1. Is it acute or chronic?—previous biochemical measurements; renal size on ultrasonography (small kidneys indicate chronic disease)
2. Is it due to urinary obstruction?—history of problems with urinary flow, urinary stones etc.; dilated pelvicalyceal system on ultrasonography (but beware of obstruction without dilatation)
3. Is it due to renal inflammation?—dipstick proteinuria and haematuria; urinary red cell casts
4. Is it due to prerenal failure/acute tubular necrosis?— clinical context; evidence of circulatory compromise/intrvascular volume depletion

Note

1. Stop all drugs that can be haemodynamically deleterious to renal function unless there is a very pressing indication for them, e.g. non-steroidal anti-inflammatory agents, angiotensin converting enzyme inhibitors, angiotensin II receptor blockers; also stop all nephrotoxic agents (e.g. aminoglycosides) and substitute non-toxic alternative
2. Insert urinary catheter and monitor fluid input/output hourly in any patient with acute renal failure— remove after 24 h if the patient is anuric/oliguric

Key investigations

To establish the diagnosis

Renal function tests—acute renal failure is usually diagnosed clinically on the basis of rapid rise in serum creatinine

Other important tests

1. ECG—looking for manifestations of hyperkalaemia
2. Electrolytes—especially potassium
3. Full blood count, coagulation screen, liver function tests
4. Creatine kinase (rhabdomyolysis)
5. Blood and other cultures—if clinically indicated
6. Autoimmune/vasculitic screen (anti-GBM, ANCA, ANA, immunoglobulins, cryoglobulins)—if clinically indicated
7. Ultrasonography of urinary tract—to determine renal size and look for evidence of obstruction
8. Chest radiograph—looking for pulmonary oedema or (less likely) evidence of lung haemorrhage in pulmonary-renal syndrome
9. Arterial blood gases—quantitate acidosis

Further management

Dependent on the cause of acute renal failure

Note

1. When intravascular volume has been restored to normal (JVP clearly visible/CVP in normal range; no postural drop in blood pressure), fluid input should then be given in equal volume to measured output of urine and other fluids, plus an allowance (500–1000 ml/day) for insensible losses. The prescription of fluid should be refined on the basis of (at least) twice daily clinical examination and daily measurement of the patient's weight
2. Precise diagnosis of the cause of acute renal failure due to renal inflammation (glomerulonephritis, tubulointerstitial nephritis, vasculitis) will probably require renal biopsy
3. If imaging suggests urinary obstruction, then this requires urgent relief, e.g. by urethral catheterization, suprapubic catheterization or percutaneous antegrade nephrostomy as appropriate

4.2 Rhabdomyolysis

See Chapter 20.5 in main text

Clinical features

Rhabdomyolysis is the breakdown of muscle fibres, when leakage of potentially toxic cellular contents into the circulation can lead to hypovolaemia, acidosis, hyperkalaemia, acute renal failure, and disseminated intravascular coagulation.

History

Muscular symptoms

1. Pain, tenderness—focal or generalized
2. May be none

Related to cause

1. Focal muscle damage
   - Obvious—e.g. crush injury, high-voltage electrical injury
   - Not so obvious—e.g. ischaemic injury following arterial embolus to limb; pressure damage following prolonged immobilization (commonly coma)
2. Generalized muscle damage
   - Excessive muscular activity

     Severe exercise—e.g. marathon running

     Epileptic fitting—prolonged (see Emergency Medicine, section 6.5)

     Status asthmaticus

     Severe dystonia

     Acute psychosis
   - Infections

     Septicaemia—see Emergency Medicine, section 7.7

     Viral myositis—e.g. influenza
   - Toxins

     Prescribed drugs—e.g. HMG CoA reductase inhibitors

     Substance abuse—e.g. alcohol, barbiturates, opioids, methanol, ethylene glycol (antifreeze), cocaine, amphetamine, ecstacy (MDMA), LSD (lysergic acid diethylamide)

     Other—e.g. snake bite, spider (black widow), bee sting (multiple), carbon monoxide, toluene, hemlock (quail that have eaten hemlock)
   - Heatstroke (see Emergency Medicine, section 9.3)
   - Malignant hyperpyrexia (see Emergency Medicine, section 9.3)
   - Neuroleptic malignant syndrome (see Emergency Medicine, section 9.3)
   - Myopathies

     Consider particularly if rhabdomyolysis occurs without clear precipitant

     Metabolic—ask for history of intermittent muscular fatigue/pain, e.g. McCardle's syndrome Inflammatory—e.g. polymyositis
   - Metabolic /endocrine

     Hypothyroidism

     Electrolyte disturbance—e.g. hypokalaemia Diabetic ketoacidosis

Examination

General

1. Vital signs—temperature, pulse rate, blood pressure, respiratory rate
2. Full physical examination

For cause of rhabdomyolysis

1. Muscles
   - Are any swollen or tender?
   - Is there a compartment syndrome?
2. Ischaemia
   - Are legs and arms all well perfused?
   - Can you feel all peripheral pulses?
3. Pressure damage
   - Look especially at the back of the head, spine, pelvis and heels—pressure sores indicate likelihood of pressure damage to muscles
4. Systemic condition
   - Rash—septicaemia (common), dermatomyositis (very rare)
   - Slow relaxing tendon jerks (hypothyroidism)

Immediate management

As for acute renal failure: see Emergency Medicine, section 4.1

To prevent rhabdomyolysis from leading to renal failure

1. Restore intravascular volume rapidly: see Emergency Medicine, section 3.1
2. Monitor
   - Urine output—urethral catheter
   - Urinary pH—dipstick
3. Fluid

   Encourage brisk diuresis (urine output >150 ml/h) of alkaline urine (myoglobin more soluble at elevated pH)—when intravascular volume has been restored give:

   - 0.9% sodium chloride/5% dextrose (alternating bags), or 0.45% sodium chloride and 2.5% dextrose (same bag), at 200 ml/h—adjust rate to achieve urine output of approx 200 ml/h
   - 1.25% sodium bicarbonate (=150 mmol/l each of sodium and bicarbonate) at 25 ml/h—adjust rate to achieve urinary pH>7
4. Mannitol / diuretic

   If urine output remains low give:

   - Mannitol—1 g/kg as 20% solution intravenously over 30–60 min and/or
   - Diuretic—e.g. frusemide 40 mg (push) –500 mg (over 2 h) intravenously

Note

1. If urine output remains low, then infusion of fluid as described here will inevitably lead to overload— FLUID INFUSION MUST BE REDUCED OR STOPPED BEFORE PULMONARY OEDEMA DEVELOPS. Then proceed as for acute renal failure (see Emergency Medicine, section 4.1)
2. There is no randomized controlled trial proof of the efficacy of a regimen comprising high volume fluid infusion/alkalinization/mannitol such as that described here, but use of this (or similar) treatment produces outcomes far superior to historical controls
3. Do not correct hypocalcaemia with calcium (risk of inducing / worsening metastatic calcification)

Key investigations

To establish the diagnosis

1. Urine
   - Dipstick test positive for blood, but microscopy shows no red blood cells
2. Blood
   - Creatine kinase—grossly elevated in severe cases (>10,000 IU/l)

Other important tests

1. ECG
   - Look for features of hyperkalaemia (see Emergency Medicine, section 5.4)
2. Blood
   - Electrolytes

     Hyperkalaemia—POTENTIALLY LIFE THREATENING. May develop rapidly. See Emergency Medicine, section 5.4

     Hypocalcaemia, hyperphosphataemia, hyperuricaemia
   - Renal function
   - Liver function tests—elevated transaminases from muscle (also LDH)
   - Coagulate on screen—risk of disseminated intravascular coagulation
3. As dictated by clinical suspicion, e.g.
   - Blood cultures
   - Thyroid function tests
   - Muscle biopsy

Further management

1. Dependent on the cause of rhabdomyolysis
2. Compartment syndrome—measure compartment pressure. Consider fasciotomy if elevated

5 Metabolic and endocrine

5.1 Hypoglycaemia

See Section 12.11 in main text

Clinical features

History

1. Coma
2. Epileptic fitting
3. Confusion and/or delirium
4. Focal neurological signs (including hemiplegia, uncommon)

The patient may not be able to give any useful history: obtain as much information as possible from others in attendance (relatives, friends, ambulance crew, bystanders etc.). Ask in particular regarding:

1. Diabetes mellitus
2. Patient self-medication, and access to insulin/oral hypoglycaemic agents
3. Previous episodes
4. Alcohol and food consumption
5. Other medical conditions

Examination

Immediate priorities

1. Airway, breathing, circulation
2. Glasgow Coma Score
3. Bedside stick test for blood glucose
4. Other features
5. Typically very pale and shut down peripherally with a cold sweat
6. Evidence that the patient is diabetic: search for Medic Alert bracelet/necklace, medication (insulin, oral hypoglycaemic agents), documentation (glucose monitoring, outpatient clinics), sites of insulin injection
7. Evidence of chronic liver disease or endocrine disorder

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Give glucose after establishing hypoglycaemia by bedside stick test, as follows:

1. Patient alert and co-operative: give glucose 10–20 g by mouth (2 teaspoons sugar, or 3 sugar lumps, or one 23 g oral ampoule of Hypostop™ gel)
2. If impaired consciousness and not protecting airway: give glucose 50% solution, 50 ml intravenously (note that the solution is viscous and irritant if extravasated, hence give through large bore needle/cannula into large vein)
3. If impaired consciousness, not protecting airway and intravenous access not possible: give glucagon 1 unit (= 1 mg) intramuscularly
4. Repeat blood sugar measurement 10 min later: repeat glucose if still hypoglycaemic

Note

Hypoglycaemic symptoms are unusual if the plasma glucose is >2.5 mmol/l, but the threshold varies from person to person; hence it is appropriate to administer one dose of glucose intravenously (50% solution, 50 ml) to any patient with impaired consciousness whose plasma glucose is <3.0 mmol/l

Key investigations

To establish the diagnosis

Blood glucose—take sample through cannula BEFORE giving intravenous glucose

Other important tests

Hypoglycaemia in a known diabetic is unlikely to require further investigation. However, if the situation is not clear-cut, then a serum sample should be taken BEFORE intravenous glucose (or intramuscular glucagon) is given for serum insulin and C-peptide levels—to determine whether hypoglycaemia is due to endogenous or exogenous insulin

The following investigations may also be appropriate

Electrolytes, renal and liver function tests

- Blood and other cultures—if clinical suspicion of sepsis

Tests for endocrine disease—adrenocortical insufficiency, hypothyroidism, hypopituitarism

- Salicylate level—if possibility of overdose
- Chest radiograph—?aspiration in any patient who has been unconscious

Further management

Dependent on the cause of hypoglycaemia

Note

1. Hypoglycaemia may recur—patients who have been given intravenous glucose and recovered from hypoglycaemia should be observed for at least 12 h, longer if they have taken long acting insulin/oral hypoglycaemic agents
2. Education—most cases of hypoglycaemia occur in known diabetics and can be avoided by the patient checking their blood glucose and responding appropriately in the event of warning signals

5.2 Diabetic ketoacidosis

See Section 12.11 in main text

Clinical features

History

1. Polyuria and polydipsia
2. Drowsiness
3. To suggest precipitating condition—often infection, but can be any acute illness
4. Monitoring and treatment of diabetes (in known diabetics)—in particular recent details of blood glucose measurements and medication with insulin or oral hypoglycaemic agents

Examination

1. State of circulation/dehydration—temperature of peripheries, skin turgor, pulse rate, blood pressure, tongue and mucous membranes, eyes, jugular venous pressure
2. Breathing—in particular for indication of acidosis (Kussmaul) and for smell of ketones
3. Glasgow Coma Score.
4. Evidence of infection—fever, localizing signs, including careful examination of the feet and skin for ulceration/sepsis

Note

1. The most reliable signs of intravascular volume depletion are postural hypotension (sitting versus lying) and a low jugular venous pressure
2. Examine carefully for evidence of complications of diabetes, but not in the immediate emergency setting

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. Restoration of intravascular volume/hydration— patients will typically have a total body fluid deficit of 3–6 litres
   - Establish intravenous access—as Emergency Medicine, section 3.1
   - Give 1 litre of colloid as fast as possible (if hypotensive and shut down peripherally, if not, then omit this and proceed directly to give saline)
   - 0.9% saline, 1 litre over 30 min, then
   - 0.9% saline, 1 litre over 1 h, with potassium (see below), then
   - 0.9% saline, 1 litre over 2 h, with potassium (see below), then
   - 0.9% saline, 1 litre every 4–6 h until rehydrated, with potassium (see below)
   - When blood glucose <15 mmol/l, switch from saline to 5% dextrose infusion until eating normally (do NOT simply allow the glucose concentration to keep falling into the normal range, reducing the insulin infusion rate to low levels according to the sliding scale. The patient continues to require insulin in dosage sufficient to allow them to metabolize ketones effectively)
2. Correction of electrolyte imbalance—all patients will have a very substantial deficit in body potassium, even though serum potassium concentration will usually be elevated at presentation. Replace potassium as follows, monitoring the serum concentration every few hours:

Serum potassium (mmol/l)	Potassium (mmol) added to each litre of fluid replacement
<3	40
<4	30
<5	20

3. Correction of hyperglycaemia

Give insulin (actrapid 50 units mixed in 50 ml of 0.9 per cent saline) intravenously according to a sliding scale as follows:

Blood glucose, measured hourly (mmol/l; reagent stick	Insulin rate (units/h)
<4	0.5
4–7	1
7.1–11	2
11.1–15	3
15.1–19	4
19.1–24	5
>24	6

4. Correction of acidosis

Acidosis will correct with restoration of circulating volume and administration of insulin; hence most cases do NOT require administration of bicarbonate. However, consider giving sodium bicarbonate (1.26% solution, 500 ml by intravenous infusion over 1 h) in cases where there is profound acidosis (e.g. arterial pH<7.0) that is thought to be causing circulatory compromise

Also

1. Empty the stomach with nasogastic tube—gastroparesis/acute gastric dilatation is a particular risk in diabetic ketoacidosis, with a high risk of vomiting and aspiration, which can be fatal
2. Give prophylaxis against venous thromboembolism (high risk) with low molecular weight heparin, e.g. enoxaparin 40 mg by subcutaneous injection once daily

---

And
Treat any precipitating condition vigorously. Note that surgical attention may be required, in particular when there is foot sepsis

---

Note

1. Hyperosmolar non-ketotic diabetic coma (HONK)
   - Typically occurs in elderly patients with non-insulin dependent diabetes mellitus (NIDDM)
   - Glucose usually >40 mmol/l
   - Not ketoacidotic (by definition)
   - Look for plasma osmolality >350 mosmol/kg, calculated as 2 × (Na+K) + Urea + Glucose (all measured in mmol/l)
   - Give colloid and 0.9% saline as for diabetic ketoacidosis, but switch to 0.45% saline when intravascular volume deficit is replaced (no postural hypotension, JVP clearly visible) if serum sodium remains >150 mmol/l
   - Insulin requirements are typically low: hence use a reduced dose of insulin on the sliding scale to avoid hypoglycaemia

---

Key investigations

To establish the diagnosis

1. Blood glucose
2. Reagent stick test of urine for ketones

Other important tests

1. Serum electrolytes
2. Arterial blood gases
3. Full blood count, renal and liver function tests
4. 'Infection screen'—chest radiograph, urine and blood culture, swab any potentially infected site
5. ECG (may have silent infarct)

Further management

Education—most cases of diabetic ketoacidosis occur in known diabetics and can be avoided. The key issue to emphasize is that illness increases insulin requirements, hence diabetics who are ill:

1. Still need to take insulin, even if they are not eating
2. Should check their blood glucose regularly (up to every 2 h or so)
3. Should give themselves frequent appropriate doses of short-acting insulin if their blood glucose starts to rise
4. Should have access to a phone number that they can call for advice if they run into problems

5.3 Metabolic acidosis

See Section 11.11 in main text

Clinical features

History

In the Emergency Medicine context presents non-specifically with

1. Altered conscious level
2. Circulatory collapse
3. Hyperventilation.
4. Key points to establish
5. In what circumstances was the patient found?
6. History of diabetes mellitus
7. History of chronic renal failure
8. Overdose—most commonly salicylates
9. Consumption of poison—e.g. ethylene glycol, methanol, antifreeze

---

Note

Medical conditions that can cause profound metabolic acidosis (with normal anion gap, see below) include:

1. Severe diarrhoeal illness
2. Renal tubular acidosis

Examination

1. State of circulation—temperature of peripheries, pulse rate, blood pressure, jugular venous pressure
2. Breathing—in particular for indication of acidosis (Kussmaul) and for smell of ketones
3. Glasgow Coma Score

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Restoration of intravascular volume

1. Establish intravenous access—as Emergency Medicine, section 3.1
2. If clinical evidence of intravascular volume depletion, resuscitate as described in Emergency Medicine, section 3.1

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

Should bicarbonate be given? This is a contentious issue: if acidosis is severe (pH<7.0) and there is circulatory compromise, give intravenous sodium bicarbonate (e.g. 1.26% solution, 500 ml by intravenous infusion over 1 h; or an equivalent amount of bicarbonate as a more concentrated solution if the patient is fluid overloaded), then assess clinical response and repeat estimation of arterial blood gases.

Note

Correction of metabolic acidosis requires careful attention to serum potassium concentration: profound hypokalaemia can occur if this is neglected

Key investigations

To establish the diagnosis

1. Arterial blood gases—show metabolic acidosis (by definition)
2. Plasma glucose and reagent stick test for urinary ketones—to exclude diabetic ketoacidosis (see Emergency Medicine, section 5.2)
3. Plasma salicylate concentration—to exclude overdose
4. Renal function tests—to exclude uraemic acidosis
5. Plasma potassium—acidosis may be associated with hypokalaemia or hyperkalaemia, but with profound deficit in total body potassium in both situations. Close monitoring required
6. Blood lactate concentration—many types of severe illness are associated with lactic acidosis, especially overwhelming sepsis
7. Plasma bicarbonate concentration
8. Calculate the anion gap: are there unusual anions in the blood? The blood 'anion gap', calculated as $(Na^{+} + K^{+}) – (Cl^{-} + HCO_3^{-})$, usually equals 10–18 mmol/ l. If there is acidosis with a high anion gap, then there must be an unmeasured substance in the blood, in which case discuss measurement of specific toxins with a clinical biochemist

Other important tests

1. Full blood count, electrolytes, liver function tests
2. Blood paracetamol level (rarely causes profound acidosis, but combined overdoses are common)
3. Chest radiograph—consider aspiration in any patient with a depressed conscious level
4. Abdominal radiograph—in cases of unexplained normal anion gap acidosis: renal tubular acidosis may be associated with nephrocalcinosis

Further management

Dependent on the cause of metabolic acidosis

5.4 Hyperkalaemia

See Chapter 20.2.2 and Chapter 20.4 in main text

Clinical features

History

1. Hyperkalaemia does not produce specific symptoms. Patients may sometimes develop 'odd feelings' in their muscles, but these are rarely dramatic
2. Cardiac arrest
3. Context—almost always occurs in the context of acute or chronic renal failure

Examination

1. Hyperkalaemia does not produce specific signs
2. Cardiac arrhythmia

Immediate management

If there are ECG changes that are more severe than tenting of the T waves:

1. Give 10 ml of 10% calcium gluconate by slow intravenous injection, repeated as necessary until ECG shows clear evidence of returning towards normal

   If ECG changes are not severe, or after giving calcium gluconate:
2. Give 10–20 units of soluble insulin in 50 ml of 50% dextrose intravenously over 20 min, and/or
3. Give nebulized b_2-agonist, e.g salbutamol 10 mg

   These treatments will lower serum potassium concentration by 1–2 mmol/l over 20–30 min and buy a few hours of time, but hyperkalaemia will recur unless the cause can be treated rapidly, hence consider:
4. Referral to nephrological services for renal replacement therapy

Note

5. Intravenous infusion of sodium bicarbonate 50–100 mmol (approx. 300–600 ml of 1.26% solution or approx. 50–100 ml of 8.4% solution) can usefully be employed to treat hyperkalaemia in the setting of severe acidosis. In other cases it has no advantage over insulin/dextrose or b_2-agonist and has the disadvantages of not only requiring a substantial sodium/fluid load (a problem in those who are already overloaded), but also that concentrated solutions are chemically irritant and hence must be administered through central venous lines

Key investigations

To establish the diagnosis

1. ECG—the following changes occur progressively as the plasma potassium concentration rises
   - Tenting of T waves
   - PR interval lengthens and P wave diminishes before disappearing
   - QRS complex widens
   - 'Sine wave' pattern
2. Serum potassium concentration >5.5 mmol/l

Other important tests

1. Renal function tests
2. To determine cause of acute renal failure—if clinical context is appropriate

Further management

1. Ion exchange resins, eg. calcium resonium™ 15 g in water three of four times daily by mouth (with concurrent prescription of a laxative), or 30 g in methylcellulose solution given as an enema, retained for 9 h and then removed by irrigation—these can be helpful in patients with persistent (but not life-threatening) hyperkalaemia who would not otherwise require renal replacement therapy. Note, however, that ion exchange resins take at least 4 h to have any effect and are NOT an emergency treatment for hyperkalaemia
2. Stop all drugs that might exacerbate hyperkalaemia unless there is a very pressing need for them and no alternative is available, e.g. potassium supplements, potassium-sparing diuretics, angiotensin converting enzyme inhibitors, angiotensin II receptor antagonists, trimethoprim, heparin
3. Dependent on the cause of hyperkalaemia

5.5 Hypokalaemia

See Chapter 20.2.2 in main text

Clinical features

History

1. In almost all cases of hypokalaemia there are no symptoms (or only non-specific symptoms) attributable to the low plasma potassium concentration
2. Cardiac arrhythmia (rare)
3. Muscular paralysis (very rare)
4. Relevant to cause of hypokalaemia

Examination

1. Hypokalaemia does not produce specific signs
2. Cardiac arrhythmia
3. Muscular paralysis (very rare)

Immediate management

Emergency treatment is rarely required.

If life-threatening cardiac arrhythmia or muscular paralysis

- Give 40 mmol of potassium intravenously via volumetric pump over 1 h, then repeat measurement of serum potassium concentration and adjust rate of potassium infusion as appropriate
- If thyrotoxic periodic paralysis
  - Give propanolol 3 mg/kg orally

Key investigations

To establish the diagnosis

1. Defined by serum potassium concentration <3.5 mmol/l, severe <3.0 mmol/l

Other important tests

1. ECG – looking for flattening of the T wave, depression of the ST segment, and the development of a prominent U wave, also for arrhythmia
2. To determine cause of hypokalaemia

Further management Dependent on the cause of hypokalaemia

5.6 Hyponatraemia

See Chapter 20.2.1 in main text

Clinical features

History

1. Does not produce specific symptoms
2. Altered consciousness, epileptic fitting
3. Relevant to cause of hyponatraemia

Examination

1. Glasgow Coma Score
2. Fluid status
   - Intravascular volume depletion—low JVP, postural hypotension
   - Clinically normal volume status
   - Volume expansion—peripheral oedema

Immediate management

Chronic asymptomatic hyponatraemia

Do NOT aim to correct rapidly:

1. If intravascular volume depletion—give 0.9% saline intravenously until intravascular volume restored, then restrict water intake
2. If euvolaemic or hypervolaemic—restrict fluid intake to 1000 ml/day. Provide swabs to moisten the mouth and give the fluid allowance as ice cubes in aliquots throughout the day

Acute symptomatic hyponatraemia

- Infuse saline intravenously, with the aim of:
  - Effecting initial correction of serum sodium concentration at a rate of 1 mmol/l/h
  - Reducing the rate of correction/stopping infusion of saline as soon as the patient's neurological condition begins to improve, or when serum sodium is elevated into the range 120–125mmol/l. DO NOT ATTEMPT RAPID CORRECTION OF SODIUM CONCENTRATION INTO THE NORMAL RANGE (probably increases risk of inducing central pontine myelinolysis)

Saline concentration	Rate of infusion (ml/h)
0.9%	3.3 × body weight (kg)
1.8%	1.7 × body weight (kg)
3.0%	1 × body weight (kg)

Monitor the serum sodium concentration regularly and adjust infusion as required

Note

If glucocorticoid deficiency is possible, then give steroid replacement immediately, e.g. hydrocortisone 100mg intravenously 6 hourly, until the diagnosis is excluded

Key investigations

To establish the diagnosis

1. Defined by serum sodium concentration <130 mmol/l

Other important tests

1. Plasma and urinary osmolality
2. Urinary sodium concentration

Further management

Dependent on the cause of hyponatraemia

5.7 Hypercalcaemia

See Chapter 12.6 in main text

Clinical features

History

1. Does not produce specific symptoms
2. Acute hypercalcaemia—general malaise, anorexia, thirst, polyuria, constipation. In severe cases vomiting, confusion, coma
3. Chronic hypercalcaemia—urinary stones, abdominal pain, mental disturbance
4. Relevant to cause of hypercalcaemia

Examination

1. Acute hypercalcaemia does not produce specific signs
2. Fluid status
   - Intravascular volume depletion—postural hypotension, low JVP
   - Dehydration—reduced skin turgor, dry mucous membranes
   - Evidence of malignancy

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2.

1. Restoration of intravascular volume (if necessary)—as described in Emergency Medicine, section 3.1
2. Saline diuresis—give 0.9% saline intravenously at a rate of 1 l/6 h until calcium restored towards normal, assuming adequate urinary output (monitor carefully, and examine the patient regularly for signs of fluid overload). Give loop diuretic, e.g. frusemide 40–80 mg orally or intravenously twice daily, if urine output slow to increase

When diuresis initiated:

3. Bisphosphonate, e.g. disodium pamidronate, 15–60 mg by intravenous infusion at a rate of 1 mg/min, repeated as necessary to give up to a total of 90 mg over 2–4 days

Also:

4. Glucocorticoids, e.g. prednisolone 40–60 mg daily, if hypercalcaemia is due to sarcoidosis, vitamin D toxicity, or haematological malignancy

Note

If glucocorticoid deficiency is possible, then give steroid replacement immediately, e.g. hydrocortisone 100mg intravenously 6 hourly, until the diagnosis is excluded

Key investigations

To establish the diagnosis

1. Defined by serum calcium concentration >2.6 mmol/l, acute symptomatic cases usually >3.0 mmol/l.

Other important tests

1. Full blood count, electrolytes, renal and liver function tests
2. Serum PTH, immunoglobulins. Protein electrophoresis of serum and urine
3. Chest radiograph
4. Directed by clinical suspicion of malignancy

Further management

Dependent on the cause of hypercalcaemia

5.8 Addisonian crisis

See Chapter 12.7.1 in main text

Clinical features

History

1. Cardiovascular collapse
2. Context of non-specific symptoms compatible with glucocorticoid deficiency: tiredness, weakness, dizziness, anorexia, weight loss, gastrointestinal disturbance. May have salt craving
3. Related to cause: personal or family history of autoimmune/endocrine disease, steroid usage (and cessation), tuberculosis, recent flank pain (?adrenal haemorrhage/infarction)
4. May occur in context of septicaemia

Examination

1. State of circulation—temperature of peripheries, pulse rate, blood pressure, jugular venous pressure
2. Hyperpigmentation—palmar creases, scars and buccal mucosae
3. Loss of axillary and pubic hair in women
4. Vitiligo

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2.

1. Restoration of intravascular volume—give 0.9% saline intravenously as described in Emergency Medicine, section 3.1
2. Steroid, e.g. hydrocortisone 100 mg intravenously (give immediately, then every 6 h)

Key investigations

To establish the diagnosis

1. Serum cortisol and ACTH—taken at the time of venous cannulation for resuscitation
2. Short synacthen test—performed later

Other important tests

1. Electrolytes, glucose, renal function tests, calcium, full blood count
2. Autoantibodies (adrenal, thyroid, intrinsic factor)
3. Thyroid function
4. Plasma renin activity—to assess mineralocorticoid status (high renin in primary adrenal insufficiency; not high in secondary adrenal insufficiency, where mineralocorticoid reserve is normal)
5. Chest radiograph—small heart, ?evidence of TB.
6. Adrenal CT scanning (where not available, abdominal radiograph—when adrenal calcification suggests TB)

Further management	1. Long-term steroid replacement therapy: usually hydrocortisone (30 mg/day in divided doses), also fludrocortisone (50–150 µg/day) if mineralocorticoid deficient 2. Education—patients need to know that they will require increased steroid dosage at times of intercurrent illness. All patients must carry a steroid card. Medic Alert bracelet

5.9 Thyrotoxic crisis

See Chapter 12.4 in main text

Clinical features	**History** 1. Usually known thyroid disease 2. Compatible with thyrotoxicosis: weight loss, palpitations, heat intolerance, sweating, diarrhoea, tremor, agitation/anxiety/irritability 3. Precipitant of thyrotoxic crisis: infection, trauma, withdrawal of antithyroid drug therapy, radio-iodine treatment, iodinated contrast dyes, thyroid surgery, childbirth 4. Personal or family history of autoimmune/endocrine disease **Examination** 1. Hyperpyrexia 2. Profuse sweating 3. Extreme restlessness, confusion, psychosis, eventually progressing to coma 4. Cardiac arrhythmia—particularly fast atrial fibrillation. Eventually cardiorespiratory collapse 5. Signs of thyroid disorder—goitre, eye signs of Graves' disease
Immediate management	THYROTOXIC CRISIS IS A POTENTIALLY FATAL DISORDER THAT REQUIRES IMMEDIATE TREATMENT ON THE BASIS OF CLINICAL SUSPICION If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2. Restoration of intravascular volume—as described in Emergency Medicine, section 3.1 Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92% Give 1. Antithyroid drug: propylthiouracil is better than carbimazole in thyrotoxic crisis • Propylthiouracil 600 mg orally or via NG tube given immediately, then 250 mg every 6 h (may also be given rectally if severe vomiting prevents oral/NG route), or • Carbimazole 20 mg orally or via NG tube given immediately, then 20 mg every 6 h 2. Iodide, starting 1–4 h after the antithyroid drug • Aqueous iodine oral solution, e.g. Lugol's (iodine 5%, potassium iodide 10% in purified water) 5 drops orally or via NG tube every 6 h 3. Hydrocortisone, 200 mg intravenously, then 100 mg every 8 h 4. Propanolol, 1 mg by intravenous injection over 1 min, repeated if necessary every 2 min to maximum of 5 mg, then 40–80 mg orally every 6 h 5. Active cooling—cooling blankets, antipyretics (use paracetamol, not aspirin, which displaces thyroid hormone from thyroid-binding globulin) Consider 6. Digoxin for atrial fibrillation—may need larger dose than usual 7. Diuretics for pulmonary oedema Also 8. Specific treatment of precipitating event (if possible)
Key investigations	**To establish the diagnosis** Thyroid function tests—these confirm the diagnosis of hyperthyroidism, but note that the diagnosis of thyrotoxic crisis is made on clinical grounds. The severity of disturbance of the thyroid function tests does not correlate with the clinical picture **Other important tests** 1. Full blood count, electrolytes, renal and liver function tests, calcium 2. Autoantibodies (adrenal, thyroid, intrinsic factor) 3. ECG—arrhythmia, especially atrial fibrillation 4. Chest radiograph—pulmonary oedema, infection
Further management	Dependent on the cause of thyrotoxicosis

5.10 Pituitary apoplexy

See Chapter 12.2 in main text

Clinical features	**History** Most commonly 1. Sudden onset retro-orbital headache 2. Visual field defect Sometimes 3. Nausea and vomiting 4. Meningism 5. Altered conscious level 6. Diplopia Also 7. Compatible with hypopituitarism or hyperprolactinaemia: lethargy, reduced libido, oligomenorrhoea/ amenorrhoea, impotence, galactorrhoea **Examination** 1. Glasgow Coma Score 2. Vision—acuity and fields 3. Eye movements—looking for ophthalmoplegia 4. Signs of underlying pituitary disease, e.g. acromegaly, are rarely present

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2.

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

On clinical suspicion of diagnosis

1. Take blood for serum cortisol assay to establish baseline retrospectively
2. Take blood for urgent prolactin assay
3. Assume anterior pituitary dysfunction and give
   - Corticosteroid, e.g. hydrocortisone 100 mg intravenously (immediately, then every 6 h)

Key investigations

To establish the diagnosis

MRI (or CT) scan of pituitary fossa—looking for haemorrhage into pituitary adenoma or other tumour

Other important tests

1. Electrolytes, glucose, renal function, calcium, full blood count, coagulation screen
2. Anterior pituitary function—baseline tests: cortisol, thyroid function tests, prolactin, LH, FSH, oestrogen/testosterone

Further management All cases require

1. Full endocrine evaluation
2. Management dependent on hormonal deficiencies and the cause of pituitary apoplexy

Prolactin <1500 mU/l

1. If vision is severly affected—urgent surgical decompression
2. If vision is not severely affected—consider surgical decompression within one week (improves visual and endocrine outcomes)

Prolactin >1500 mU/l (suggests prolactinoma)

1. A conservative (non-surgical) approach may be adopted if there is no progressive visual or neurological deficit and prolactin levels are very high, suggesting a prolactinoma
2. Start immediate treatment with dopamine agonist drug such as bromocriptine or cabergoline

5.11 Acute porphyria

See Chapter 11.5 in main text

Clinical features

History

Intermittent episodes of:

1. Acute abdominal pain, vomiting and constipation
2. Severe proximal limb and/or back pain
3. Seizures, coma
4. Psychiatric disturbance

Notes

5. Family history—nearly all the acute porphyrias are dominantly inherited, but many carriers are latent
6. Rash—in variegate and hereditary coproporphyria (which can cause acute neurovisceral attacks) but NOT in acute intermittent porphyria
7. Precipitant—alcohol, sex steroids, drugs (see Table 17), anaesthetic agents, starvation

Examination

1. Cardiovascular—looking for sinus tachycardia, hypertension
2. Abdominal—may have signs indistinguishable from those of the acute 'surgical' abdomen, but tenderness is usually lacking
3. Neurological—Glasgow Coma Score (if appropriate); look for sensorimotor neuropathy, respiratory muscle weakness

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2.

If coma, as described in Emergency Medicine, section 6.1 Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

1. Stop all known precipitant drugs, especially any that have recently been prescribed. Consult Table 17 before prescribing ANY agent
2. Give 5% dextrose intravenously, 1 litre every 8 h (except if hyponatraemic, when give reduced volume of more concentrated dextrose solution). Start high carbohydrate diet when patient able to eat
3. Give haem arginate, 3 mg/kg once daily for 4 days (maximum 250 mg daily) by intravenous infusion in 0.9% saline over at least 30 min

Notes

1. Supplies of haem arginate can be obtained from Orphan Europe Ltd., 32 Bell Street, Henley on-Thames, Oxon RG9 2BH. Tel 44(0)1491–414333. e-mail: info.uk@orphan-europe.com. Also from the on-call pharmacist at the University College of Wales, Cardiff ([029] 2074 7747); St James' University Hospital, Leeds ([0113] 243 3144 or [0113] 283 7010); St Thomas' Hospital, London ([020] 7928 9292)
2. Seizures pose difficulties since many anticonvulsants precipitate or worsen porphyric attacks: temazepam, lorazepam, and midazolam are probably safe
3. Distress may be helped by chlorpromazine. Morphine and pethidine (safe) may be required
4. Propanolol is useful for controlling hypertension and extreme tachycardia

Key investigations

To establish the diagnosis

1. Detection of porphyrin precursors in fresh urine (which may rarely become red/purple/brown on standing)

Other important tests

1. Electrolytes—may cause profound hyponatraemia. Monitor serum sodium daily in the acute phase
2. Full blood count, renal and liver function tests, calcium
3. ECG
   Further investigation to exclude serious abdominal or neurological disease will be determined by clinical presentation, especially if excretion of haem precursors is normal, e.g.
4. Amylase/chest and abdominal radiograph/CT abdomen/senior surgical opinion
5. CT brain/lumbar puncture

Note

In a patient with known porphyria, the absence of excess porphobilinogen (PBG) or d-aminolaevulinic acid (ALA) in the urine renders acute porphyria an unlikely cause of the current illness

Further management

1. Seek expert advice to establish diagnosis and investigate family
2. Medic Alert bracelet important as warning to health care personnel in the future

6 Neurological

6.1 Coma

See Chapter 24.13.1 in main text

Clinical features

History

Coma is defined as a Glasgow Coma Score (GCS) <8, hence the patient will not be able to give any useful history. Obtain as much information as possible from others in attendance (relatives, friends, ambulance crew, bystanders etc.). Ask in particular regarding:

1. The circumstances in which the patient was found
2. Alcohol consumption
3. Diabetes mellitus
4. Epilepsy
5. Drugs of abuse, in particular opioids
6. Head injury
7. Regular medications
8. Past medical history

Examination

Initial survey

1. Airway, breathing, circulation
2. Reagent stick test for blood glucose (?hypoglycaemia)
3. Check for small pupils and slow respiratory rate (?opioid overdose)
4. Check temperature (?hypothermia)
5. Look for Medic Alert bracelet or necklace
6. Check Glasgow Coma Score (see Table 18)

Further examination

1. State of circulation—temperature of peripheries, pulse rate, blood pressure, JVP
2. Respiratory—look for evidence of aspiration
3. Neurological
   - Focal/lateralizing signs—a structural lesion is likely if these are present
   - Meningism
   - Movements (can be subtle)—status epilepticus
4. Tongue biting or incontinence of urine—suggest (but do not prove) epilepsy
5. Back of head and neck—for bruising or bleeding to suggest head injury
6. Ears and nose—for bleeding or CSF leak to suggest basal skull fracture
7. Search pockets etc. for clues—e.g. anticonvulsant tablets

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Nurse in recovery position (when injury to neck excluded).

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%. Consider oropharyngeal airway Patients with a GCS <8 are likely to need endotracheal intubation to protect and maintain their airway if they do not respond to glucose or naloxone. This is obligatory if they need to be moved from an area where they can be given intensive nursing care to one where they cannot, e.g. to CT scanner

Establish intravenous access

1. If hypoglycaemic—give 50 ml of 50% glucose (dextrose monohydrate) intravenously. IF IN DOUBT, TREAT
2. If possibility of opioid overdose—give naloxone 0.8–2 mg intravenously, repeated at intervals of 2–3 min to a maximum of 10 mg. IF IN DOUBT, TREAT
3. If hypothermic—start rewarming

Key investigations

To establish the diagnosis

1. Glucose.
2. CT brain—if diagnosis not clinically apparent and patient not improving rapidly. Look for:
   - Extradural, subdural, subarachnoid, or intracerebral haemorrhage
   - Signs of raised intracranial pressure
   - Focal ischaemia (may not be visible on early scan)
3. Blood film for malaria—if relevant travel history

Other important tests

1. Electrolytes, renal and liver function tests, calcium, full blood count
2. ECG—note that 'ischaemic' changes can occur in subarachnoid haemorrhage
3. Chest radiograph—?aspiration pneumonia
4. Arterial blood gases—if diagnosis not clear, or if PaO_2 <92% on air
5. Sepsis screen (selected cases)
6. Lumbar puncture (selected cases)
7. EEG (selected cases, ?non-convulsive status)

Further management

Dependent on the cause of coma.

6.2 Acute confusional state

See Chapter 24.8 and Section 30 in main text

Clinical features

History

Is the patient confused?:

1. Establish that the patient is not dysphasic rather than confused
2. Abbreviated Mental Test (AMT) score—a score of 6 or less is likely to indicate impaired cognition
   - Age
   - Time (to nearest hour)
   - What year is it?
   - Name of institution
   - Recognition of two persons (can the patient identify your job and that of a nurse?)
   - Date of birth (day and month)
   - Year of First World War
   - Name of present monarch
   - Count backwards from 20 to 1
3. The patient who is confused cannot (by definition) give an accurate and reliable account of themselves

Obtain as much information as possible from others in attendance (relatives, friends, ambulance crew bystanders etc). Ask in particular regarding:

1. The situation in which the patient was found
2. Any recent change in health, in particular:
   - Symptoms to suggest infection
   - Medications—especially any recent change
3. Previous cognitive function
4. Alcohol consumption (consider intoxication, withdrawal, Wernicke's encephalopathy)
5. Drugs of abuse (if relevant)
6. Regular medications
7. Past medical history
8. Social circumstances

Examination

1. General appearance—well-presented clothing and cleanliness indicates an acute problem or an assiduous carer.
2. Nutritional state—reflects previous weeks/months
3. Hydration state—reflects previous 48 h
4. Full physical examination—look in particular for:
   - Temperature—pyrexia or hypothermia
   - Pulse rate, blood pressure, JVP—hypotension from any cause can lead to confusion
   - Evidence of sepsis—in particular chest, urine, cellulitis
   - Neurological—focal signs (indicating a focal neurological lesion, most commonly stroke), head injury, Wernicke's encephalopathy
   - Evidence of organ failure—cardiac, respiratory, hepatic, renal
   - Urinary retention or faecal impaction
   - Hip or pelvic fracture.

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2.

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

If hypoglycaemic—give 50 ml of 50% glucose (dextrose monohydrate) intravenously. IF IN DOUBT, TREAT

1. Fluids—encourage oral intake, but if intravenous fluids are required, then insert venous cannula into flat site and bandage carefully
2. Treat any obvious precipitating condition—if none apparent then consider initiating antibiotic treatment for, e.g. urinary infection, on a 'best guess' basis (e.g. ciprofloxacin 500 mg orally twice daily, but note local hospital policy)
3. Anticipate and avoid problems:
   - Do not exacerbate confusion—nurse in lit room (darkness makes confusion worse), expose to limited number of staff (many people 'popping in' increase confusion), enlist assistance from relatives/carers/friends (a sensible person that the patient knows can be enormously helpful)
   - Pressure areas—appropriate mattress
   - Urine—try to avoid catheterization if possible (will make any infection harder to clear), but need to strike a difficult balance with concern for skin/pressure areas
   - Bowels—suppositories, laxative, enema as required
   - Venous thromboembolism—low molecular weight heparin, e.g. enoxaparin 20 mg subcutaneously once daily
4. Sedation—try to avoid if possible, but if necessary use risperidone 0.5 mg orally twice daily (increased in steps of 0.5 mg twice daily to 1–2 mg twice daily) or haloperidol 0.5–2 mg orally/intramuscularly two to three times daily. (Dosage of both drugs appropriate for the elderly—higher doses likely to be required for younger patients)

Key investigations

To establish the diagnosis

These will be guided by any clinical leads, but as nonspecific presentation is common, the following are advisable in almost all patients:

1. Reagent stick test for blood glucose.
2. Full blood count, electrolytes, renal and liver function tests, calcium, phosphate, cardiac enzymes, glucose, thyroid function, inflammatory marker (CRP or ESR)
3. Oxygen saturation—check arterial blood gas if PaO_2 <92%. on air
4. Sepsis screen—urine dipstick test, urine and blood culture
5. Chest radiograph
6. ECG

Other important tests

Guided by clinical findings or results of screening investigations, e.g. new focal neurological signs—imaging of brain by CT scan or MRI

Further management

Dependent on the cause of confusion

6.3 Acute stroke

See Chapter 24.13.7 in main text

Clinical features

History

May be difficult to obtain, particularly if the patient has dysphasia. If this is the case, get as much information as possible from others in attendance (relatives, friends, ambulance crew, bystanders etc.)

1. Focal neurological deficit—usually of sudden onset
2. Previous episodes—stroke, transient ischaemic attack, amaurosis fugax
3. Risk factors
4. Other medical conditions
5. Medications
6. Normal level of functioning—do they need help with activities of daily living?
7. Social circumstances

Examination

1. Airway, breathing, circulation
2. Neurological
   - Glasgow Coma Score
   - Nature of focal deficit (see Table 19)
3. Cardiovascular—pulse rate and rhythm (?atrial fibrillation), blood pressure, carotid bruits, cardiac murmurs, absent peripheral pulses

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2.

1. Nurse in recovery position if impairment of consciousness
2. Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%. Consider oropharyngeal airway
3. Establish intravenous access
4. Reagent stick test for blood glucose: if hypoglycaemic—give 50 ml of 50% glucose (dextrose monohydrate) intravenously. IF IN DOUBT, TREAT

Notes

1. Urgent neurosurgical assessment is required for patients with large cerebellar infarcts or haemorrhages or hydrocephalus, and for some cases with cerebral haemorrhage
2. There is no proven benefit for drugs in the limitation of neural damage, including corticosteroids, nimodipine, plasma volume expanders, barbiturates. or glutamate receptor antagonists. Patients treated rapidly (?within 3 h of stroke) may benefit from thrombolysis, but this should only be given in centres that use the treatment routinely, and preferably in the context of controlled trials

Key investigations

To establish the diagnosis

CT or MRI brain—also to distinguish between infarction and haemorrhage

Other important tests

1. Full blood count, electrolytes, renal and liver function tests, calcium, inflammatory markers (CRP or ESR), coagulation screen
2. ECG—look for arrhythmia or signs of recent myocardial infarction
3. Chest radiograph—?aspiration pneumonia
4. Echocardiography; ultrasound/Doppler examination of carotid arteries—in selected cases

Further management

Short term

1. Nursing and physiotherapy—protect pressure areas, attention to bladder and bowels, prevent contractures, aid recovery of function, psychological support
2. Hydration/nutrition—If swallowing impaired, stop oral feeding and start intravenous fluids
3. Blood pressure—this is commonly elevated immediately after a stroke, cerebral autoregulation is impaired, and aggressive attempts to reduce it are likely to cause more harm than good. If BP >220/130 then many physicians would treat, e.g. using modified release nifedipine 10 mg orally, but some would only do so if there was evidence that the hypertension were causing acute organ damage
4. Venous thromboembolism—high risk: use compression stockings
5. Antiplatelet therapy—usually aspirin 300 mg once daily – should be started to prevent recurrence as soon as haemorrhage has been excluded
6. Blood glucose—use intravenous sliding scale of insulin (see Emergency Medicine, section 5.2) to obtain good control in diabetics

Medium/long term

1. Rehabilitation and social support as required
2. Control of vascular risk factors—hypertension, hyperlipidaemia, cessation of cigarette smoking
3. Consider imaging of the carotid arteries in all patients who have made a reasonable recovery from a carotid territory stroke: endarterectomy may be indicated

6.4 Subarachnoid haemorrhage

See Chapter 24.13.7 in main text

Clinical features

History

1. Presentation is very variable: typically severe headache ('worst ever') of sudden onset, but can vary from minor symptoms to collapse/coma or sudden death
2. Previous episodes; recent unusual headache
3. Risk factors—hypertension, cigarette smoking, alcohol (binge drinking), adult polycystic kidney disease, connective tissue disorders (some)

Examination

1. Airway, breathing, circulation
2. Glasgow Coma Score
3. Focal neurological signs—in particular:
   - Third nerve palsy—posterior communicating artery aneurysm
   - Sixth nerve palsy—posterior fossa aneurysm, but usually a false localizing sign
   - Bilateral leg weakness—anterior communicating artery aneurysm
   - Dysphasia/hemiparesis—middle cerebral artery aneurysm
4. Neck rigidity
5. Retinal haemorrhages
6. Cardiovascular—arrhythmia, hypertension

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2.

1. Nurse in recovery position if impairment of consciousness.
2. Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%. Consider oropharyngeal airway
3. Establish intravenous access and resuscitate if volume depleted or dehydrated
4. Bed rest for all patients
5. Nimodipine 60 mg orally every 4 h, started within 4 days of subarachnoid haemorrhage and continued for 21 days, should be given to all patients with subarachnoid haemorrhage who are not hypotensive (systolic BP <110 mmHg). This is to prevent ischaemic neurological deficit
6. Keep arterial pressure <160/100 mmHg (using conventional agents)

Key investigations **To establish the diagnosis**

1. CT brain, without contrast, taking thin cuts through the base
2. Lumbar puncture—perform not earlier than 12 h after the ictus if CT normal: look for xanthochromia after centrifugation of CSF

Other important tests

1. Electrolytes, renal and liver function tests, full blood count, coagulation screen
2. ECG—note that 'ischaemic' changes can occur in subarachnoid haemorrhage
3. Chest radiograph—?aspiration pneumonia

Further management

Depends on the patient's clinical condition: if

- GCS = 12 or more, or
- GCS <12 with space occupying intracranial haemorrhage or hydrocephalus

Then surgery should be considered in patients with proven intracranial aneurysms: hence discuss with neurosurgical colleagues with a view to arranging fourvessel angiography (CT angiograms may be done first, and sometimes instead)

6.5 Status epilepticus

See Chapter 24.13.3 in main text

Clinical features

Definition

Continuous seizures or serial (two or more) discrete seizures between which there is incomplete recovery of consciousness

History

1. Loss of consciousness, usually with obvious fitting The patient will not be able to give any useful history. Obtain as much information as possible from others in attendance (relatives, friends, ambulance crew, bystanders etc.). Ask in particular regarding:
2. The circumstances in which the patient was found
3. Past history of epilepsy
4. Alcohol consumption
5. Any possible drug abuse
6. Diabetes mellitus
7. Regular medications
8. Any other medical history

Examination

Initial survey

1. Airway, breathing, circulation
2. Signs of injury—especially of tongue, which can compromise breathing
3. Respiratory—?aspiration
4. Glasgow Coma Score
5. Medic Alert bracelet/necklace

Further examination

1. Vital signs—temperature, pulse rate, blood pressure
2. Neurological
   - Pupil size and reactions
   - Other brainstem signs
   - Symmetry of tone and reflexes in the limbs
   - Neck stiffness
   - Note that focal signs may indicate focal pathology, but can be seen as a post-ictal phenomenon (i.e. Todd's paresis)
3. Search pockets etc. for clues—e.g. anticonvulsant tablets

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2.

1. Place in recovery position (if possible)
2. Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%. Consider oropharyngeal airway, but do not try to insert one against resistance i.e. when the patient is actually fitting
3. Establish intravenous access.
4. Reagent stick test for blood glucose—if hypoglycaemic: give 50 ml of 50% glucose (dextrose monohydrate) intravenously. IF IN DOUBT (GLUCOSE <3 mmol/l), TREAT
5. Anticonvulsant—first line
   - Lorazepam 0.1 mg/kg intravenously at 2 mg/min. THE FIRST LINE TREATMENT OF CHOICE
   - Diazepam 10–20 mg intravenously at a rate of 5 mg/min. This may be repeated after 30–60 min if necessary, and can be followed by infusion (add 10–40 mg of diazepam to 100 ml of 5% dextrose to make a solution containing 0.1 0.4 mg/ml) at a rate of e.g. 5 mg/h, adjusted according to clinical response, but with maximum dose of 3 mg/kg body weight over 24 h
6. Anticonvulsant—second line. If seizure activity still continues, consider
   - Fosphenytoin, 15 mg phenytoin-equivalent (PE)/kg body weight (fosphenytoin 1.5 mg = phenytoin 1 mg) by intravenous infusion at a rate of 100–150 mg PE/min, followed by 4–5 mg PE/kg daily in 1–2 divided doses. Dose adjusted according to clinical response and trough plasma phenytoin levels. THE SECOND LINE TREATMENT OF CHOICE
   - Phenytoin, 15 mg/kg body weight by intravenous infusion at a rate not exceeding 50 mg/min, followed by 100 mg every 6–8 h. Dose adjusted according to clinical response and trough plasmaphenytoin levels
7. Anticonvulsant—Third line treatments. If seizure activity still continues, consider
   - Phenobarbital (phenobarbitone), 20 mg/kg by intravenous infusion at a rate of not more than 50–75 mg/min, maximum dose 1000 mg. Note that this treatment may lead to respiratory depression. THE THIRD LINE TREATMENT OF CHOICE
   - Paraldehyde, 5–10 ml by deep intramuscular injection (not more than 5 ml at any one site), or 10–20 ml administered by enema as a 10% solution in physiological saline or mixed with an equal volume of olive oil. Note that this treatment is only to be used if other treatments listed above are not available
8. Anaesthesia. If seizure activity still continues after first, second and third line treatments (or earlier if required to achieve adequate airway protection/ventilation):
   - Call anaesthetist and arrange ICU admission for anaesthesia with thiopental or propofol. Ventilate with EEG monitoring until clinical and EEG epileptic activity ceases

Key investigations

To establish the diagnosis

Status epilepticus is a clinical diagnosis, although EEG is used to diagnose the very rare condition of non-convulsive status in a patient with unexplained coma

Other important tests

1. A reagent stick test for blood glucose should be performed in all patients.
   The intensity of further investigation depends on the context: the patient that is known to have epilepsy who has frequent prolonged seizures does not require extensive investigation after each and every one. In other cases:
2. Glucose, electrolytes, renal and liver function tests, calcium, creatine kinase, anticonvulsant level (if appropriate)
   And consider:
3. Arterial blood gases.
4. Chest radiograph—?aspiration
5. ECG
6. Sepsis screen
7. Toxicology screen
8. CT or MRI brain
9. Lumbar puncture—only after imaging to exclude raised intracranial pressure or intracerebral mass

Further management

Dependent on the cause of status epilepticus

6.6 Spinal cord compression

See Chapters 24.13.16 and 24.13.17 in main text

Clinical features

History

Cord compression

1. Leg weakness—developing over hours or days
2. Sensory symptoms—in particular, is there a sensory level, which can be suspended?
3. Bladder disturbance

Cause of cord compression

1. Back pain
2. Intervertebral discs—any previous problem?
3. Malignancy—any known previous, or any features to suggest this diagnosis, e.g. anorexia, malaise, weight loss
4. Infection—sweats, fevers, rigors. Tuberculosis. Risk factors for osteomyelitis or abscess, e.g. previous septicaemia (particularly staphylococcal), intravenous drug abuse, haemodialysis

Examination

Cord compression

1. Motor—look for increased tone, weakness, and hyperreflexia below the site of the lesion. Do the plantars go up or down?
2. Sensory—is there a sensory level, which can be suspended? In particular, check for sensory loss in the saddle area, which would suggest a cauda equina lesion
3. Bladder—is this palpable?

Cause of cord compression

1. General examination for signs of malignancy—e.g. cachexia, clubbing, lymphadenopathy, pallor, jaundice, chest/abdominal examination, pelvic mass
2. Suggestion of infective cause—temperature

Immediate management

1. Nurse on pressure-relieving mattress
2. Relieve urinary retention with urethral catheter (if appropriate)
3. EMERGENCY IMAGING AND CONSULTATION WITH NEUROSURGICAL COLLEAGUES
4. Specific treatments depending on precise diagnosis
   - Disc protrusion—surgical decompression
   - Metastasis—high dose steroids (e.g. methylprednisolone) and radiotherapy
   - Abscess—surgical decompression/drainage; antimicrobials. For an immunocompetent patient with a pyogenic abscess give intravenous antimicrobials as follows: third-generation cephalosporin, e.g. cefotaxime 1–2 g 12 hourly PLUS flucloxacillin 1–2 g 6 hourly PLUS metronidazole 500mg 8 hourly. Modify regimen when microbiological results are available.
   - Spinal cord tumours (rare)—neurosurgical intervention may be appropriate

Note

Acute spinal cord injury—give methylprednisolone 30 mg/kg as intravenous bolus over 1 h, followed by 4.0 mg/kg/h for 23 h

Key investigations

To establish the diagnosis

MRI spine, performed as an emergency (if this is not available, discuss best available imaging modality with radiological colleagues, e.g. CT scan, myelography)

Other important tests

1. Full blood count, electrolytes, renal and liver function tests, calcium, inflammatory markers, coagulation screen, immunglobulins, and protein electrophoresis
2. Urinary Bence Jones protein
3. Chest radiograph
4. Other tests as dictated by clinical suspicion, e.g. blood cultures, lymph node biopsy

Further management

Dependent on the cause of spinal cord compression

6.7 Acute inflammatory polyneuritis (Guillain Barré)

See Chapter 24.19 in main text

Clinical features

History

1. Sensory symptoms—begin distally and ascend symmetrically
2. Motor—weakness, usually ascending (but can sometimes be proximal), symmetrical. Muscle pain is common (particularly lower back or interscapular)
3. Legs usually worst affected, but can sometimes be arms
4. Progression usually occurs over days (no longer than 4 weeks, by definition) but can sometimes be more rapid
5. Patients often have upper respiratory tract or diarrhoeal illness (especially *Campylobacter jejuni*) in the few weeks prior to onset

Examination

1. Motor—reduced tone; lower motor neurone weakness, distal > proximal; areflexia. May have facial involvement and ophthalmoplegia (Miller Fisher syndrome)
2. Sensory—glove and stocking sensory disturbance, often mild
3. Respiratory—RESPIRATORY FAILURE DUE TO MUSCLE WEAKNESS IS AN AVOIDABLE CAUSE OF DEATH: check forced vital capacity and monitor frequently
4. Autonomic—look for variable pulse rate, variable arterial pressure, intestinal ileus, urinary retention

Immediate management

1. Respiratory
   - CONSIDER ELECTIVE ASSISTED VENTILATION SOONER RATHER THAN LATER IF THE PATIENT IS TIRING
   - Note that tracheal suction can trigger hypotension or bradycardia in the presence of autonomic dysfunction
2. Cardiac
   - Monitor ECG
   - Arrhythmias can be fatal—treat as appropriate
   - Use antihypertensive drugs with extreme caution (if at all) in the face of autonomic dysfunction
3. Fluids
   - If gag reflex impaired—stop oral feeding and start intravenous fluids
   - Will need to consider PEG feeding as an early option
4. Nursing and physiotherapy
   - Keep chest clear
   - Protect pressure areas
   - Attention to bladder and bowels
   - Prevent contractures: move all joints through their full range of movement daily
   - Aid recovery of function
   - Psychological support: emphasize that most cases recover well
5. Pain—give non-steroidal anti-inflammatory agents as required. Consider amitriptyline, carbamazepine, gabapentin
6. Compression stockings and low molecular weight heparin (e.g. enoxaparin 40 mg subcutaneously once daily)—to reduce the risk of venous thromboembolism
7. Intravenous immunoglobulin, 0.4 g/kg body weight/ day, for 5 days—give to all patients, excepting those with very mild disease

Key investigations **To establish the diagnosis**

Acute inflammatory polyneuritis (Guillain Barré syndrome) is primarily a clinical diagnosis: investigation may confirm it, but initial management is dictated by clinical suspicion

1. Nerve conduction studies—the earliest abnormality is impersistence or absence of F waves. Peripheral demyelination starts proximally in the nerve roots, hence distal conduction velocities and motor latencies are often normal early in the illness, even when there is profound weakness
2. Lumbar puncture—look for elevated protein (but not cells)
3. Anti GQ1b antibodies—present in all cases that are associated with ophthalmoplegia

Other important tests

Relevant to cause

1. Stool culture and serology for *Campylobacter jejuni*
2. Serology for atypical pneumonias
3. CSF analysis for viral infection

   Need to exclude

   Acute intermittent porphyria—see Emergency Medicine, section 5.11

General

1. Full blood count, electrolytes, renal and liver function tests, plasma calcium, magnesium and phosphate concentrations
2. ECG
3. Chest radiograph

Further management

Dependent on the nature of any residual disability. Significant weakness remains in about 10% of cases, especially those with the axonal form

6.8 Myasthenia gravis

See Chapter 24.22.2 in main text

Clinical features

History

Myasthenic crisis

1. Breathing difficulty due to muscular weakness in a patient with known myasthenia.

Presentation of myasthenia

2. Droopy eyelid(s)/double vision
3. Difficulty chewing, swallowing, talking (nasal speech), holding the head up
4. Limb weakness
5. Symptoms less severe in the morning, getting worse as the day goes on
6. Exacerbation by intercurrent illness, pregnancy, menses, and by some drugs

Examination

Myasthenic crisis

1. Exhaustion
2. Ineffective respiratory effort
3. Inability to clear airway secretions
4. Cyanosis
5. Low vital capacity
   Also
6. Check for focal lung signs

Myasthenia

Muscular weakness that becomes worse with repetitive effort (fatiguability)

Immediate management

Respiratory failure caused by muscular weakness in a patient with myasthenia can be due to a myasthenic crisis (attributable to the disease itself) or rarely to an overdose of anticholinesterases (cholinergic crisis). These cannot reliably be distinguished on clinical grounds, hence safe management consists of:

1. Airway, breathing, circulation
2. Intubate and ventilate
3. Stop all anticholinesterases.
   IF there is specialist expertise, AND in conjunction with someone skilled in intubation, then edrophonium chloride, 2 mg by intravenous injection, can be used to discriminate between underdosage and overdosage of cholinergic drugs

Key investigations

To establish the diagnosis

Myasthenic crisis is a clinical diagnosis

Of myasthenia gravis

1. Edrophonium chloride (Tensilon) test: after pretreatment with atropine (0.6 mg intravenously), give edrophonium 2 mg intravenously and look for transient improvement in e.g. ptosis, diplopia, dysarthria. If no improvement after 1–2 min give edrophonium 8 mg intravenously and watch for effect. Note that this test has limited sensitivity and specificity
2. Serum acetylcholine receptor antibodies—highly specific, being present in 85% of patients with generalized myasthenia
3. Electromyography: look for increased jitter, also decremental response to repetitive nerve stimulation— this test has good sensitivity and specificity

Other important tests

In myasthenic crisis

1. Arterial blood gases
2. Chest radiograph
3. Electrolytes, renal and liver function tests, calcium, phosphate, full blood count
4. Sepsis screen (if appropriate)

Further management

In myasthenic crisis consider the following:

1. Plasma exchange—e.g. 50 ml/kg body weight/day for 4 or 5 days
2. Intravenous immunoglobulin—e.g. 0.4 g/kg body weight/day, for 5 days

Long-term treatment of myasthenia. Consider:

1. Immunosuppression—usually prednisolone (starting at a low dose of e.g. 10 mg on alternate days) and azathioprine (2.5 mg/kg body weight/day)
2. Anticholinesterase, e.g. pyridostigmine bromide 30–120 mg at suitable intervals throughout the day (total daily dose 0.3–1.2 g). Together with antimuscarinic agent if needed
3. Thymectomy

6.9 Acute Wernicke's encephalopathy

See Chapter 26.3 and Chapter 26.7.3 in main text

Clinical features

History

1. Alcoholism—usually, but also other states of nutritional deficiency and protracted vomiting
2. Difficulty standing/walking
3. Confusion

The patient will almost certainly not be able to give a reliable history: corroborate as much information as possible from other sources (relatives, friends, general practitioner etc.)

Examination

Related to Wernicke's encephalopathy, the classic triad of:

1. Ophthalmoplegia
   - Horizontal and vertical nystagmus
   - Weakness/paralysis of lateral rectus muscles
   - Weakness/paralysis of conjugate gaze
2. Ataxia—predominantly affecting stance and gait, often without clear-cut intention tremor
3. Confusion, confabulation

Related to clinical context:

1. Cardiovascular—look for evidence of intravascular volume depletion and/or dehydration
2. Consider other complications of alcoholism
   - Acute alcohol withdrawal
   - Acute liver failure
   - Chronic liver disease and its complications
3. Consider other causes of an acute confusional state—see Emergency Medicine, section 6.2
4. Nutritional status

Immediate management

Thiamine—give parenteral thiamine immediately, usually in combination with other vitamins B and C, e.g. Pabrinex ™I/V high potency, 2–3 pairs of ampoules intraven- ously over 10 min every 8 h (each pair of ampoules contains ascorbic acid 500 mg, anhydrous glucose 1 g, nicotinamide 160 mg, pyridoxine hydrochloride 50 mg, riboflavin 4 mg, and thiamine hydrochloride 250 mg in a total of 10 ml). Note—facilities for treating anaphylaxis should be available

Key investigations

To establish the diagnosis

1. Wernicke's encephalopathy is a clinical diagnosis
2. Red cell transketolase—a reduced level confirms thiamine deficiency

To exclude other conditions

CT scan brain—should be done in all cases because of the high incidence of structural lesions, e.g. subdural haematoma, in this group of patients

Other important tests

Depending of clinical context, consider as for acute confusional state—see Emergency Medicine, section 6.2

Further management

1. After 3–5 days, switch from intravenous to oral vitamin replacement, e.g. thiamine 50 mg once daily + vitamin B tablets, Compound, Strong, 1–2 tablets three times daily + vitamin C 100 mg once daily
2. If alcohol withdrawal—see Emergency Medicine, section 8.2
3. Other aspects: as for acute confusional state—see Emergency Medicine, section 6.2—except avoid antipsychotics which lower seizure threshold
4. Long term—measures to help alcoholism

7 Infectious disease

7.1 Malaria

See Chapter 7.13.2 in main text

Clinical features

Clinical features

Falciparum malaria is the life-threatening form and the immediate concern in patients presenting to medical services in malarious areas, or who have travelled to such areas

Transmitted to man by the bite of an infected *Anopheles* mosquito. The interval between bite and first symptom is usually 7–14 days. Most patients with imported falciparum malaria present within 3 months of return from the malarious area, but a few present up to 1 year or more later

History

1. Risk of exposure to malaria—anyone who has travelled to a malarious area and presents to medical attention with a febrile illness has malaria until proved otherwise
2. Symptoms of malaria
3. Early—malaise, headache, backache, myalgia, anorexia, low grade fever.
4. Later—dizziness, nausea, vomiting, abdominal discomfort, diarrhoea, rigors and drenching sweats.

   Symptoms of cerebral malaria
5. Gradual decline in conscious level over several hours
6. Generalised epileptic convulsion without post-ictal recovery of consciousness (present in 50 per cent of adult cases).

Note

'Classical' tertian (48 h) or subtertian (36 h) periodicity of fever spikes is rarely seen in falciparum malaria

Examination

1. Vital signs—temperature, pulse rate, blood pressure, respiratory rate. A high fever (rising to >39°C) is typical, which can be of any periodicity
2. General—anaemia, jaundice
3. Abdominal—ook for moderate tender enlargement of liver and/or spleen
4. Neurological—look for signs of cerebral malaria
   - Glasgow Coma Score—reduced (by definition in cerebral malaria)
   - Focal signs—note presence of dysconjugate gaze, brisk tendon reflexes, ankle clonus, extensor plantar responses and absent abdominal reflexes; and of decorticate or decerebrate posturing in severe cases
   - Fundi—retinal haemorrhages are common (exudates and papilloedema also occur)

Notes

1. The following are NOT found in malaria
   - Lymphadenopathy
   - Rash—excepting herpes simplex 'cold sores' in some cases
   - Focal signs
2. Signs of hypoglycaemia may be misinterpreted as merely being manifestations of cerebral malaria

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

If clinical evidence of intravascular volume depletion, establish intravenous access and resuscitate as described in Emergency Medicine, section 3.1

If hypoglycaemic—give 50 ml of 50% glucose (dextrose monohydrate) intravenously, followed by infusion of 10% glucose at sufficient rate to maintain blood glucose concentration >3 mmol/l. IF IN DOUBT, TREAT

Antimalarial drugs for falciparum malaria (adult dosages)

Assume chloroquine-resistance

Patients who can swallow and retain tablets

Use ONE of the following regimen

1. Mefloquine
   - By mouth: 15–25 mg/kg of mefloquine base (maximum 1500 mg), divided into two doses 6–8 h apart
2. Proguanil with atovaquone ('Malarone')
   - By mouth: 4 tablets (each containing 100 mg proguanil and 250 mg atovaquone) once daily for 3 days
3. Artemether with lumefantrine ('Riamet')
   - By mouth: 4 tablets (each containing 20 mg artemether and 120 mg lumefantrine) twice daily for 3 days
4. Quinine—*the treatment of choice in many countries*
   - By mouth: 600 mg of quinine salt every 8 h for 7 days
   - Afterwards: when the 7 day course is completed, give tetracycline 250 mg four times daily for 7 days or doxycyline 100 mg daily for 7 days
5. Fansidar™ (each tablet containing pyrimethamine 25 mg and sulfadoxine 500 mg)
   - 3 tablets as single dose.

Patients with severe malaria or who cannot swallow and retain tablets

Use ONE of the following regimen:

1. Quinine—*the treatment of choice in many countries*
   - By intravenous infusion: loading dose of 20 mg/kg dihydrochloride salt (maximum 1400 mg) diluted in 10 ml/kg isotonic fluid and given over 4 h, then after 8 h give maintenance dose of 10 mg/kg (maximum 700 mg) over 4 h, repeated following further 8 h gaps until patient can swallow tablets to complete 7 day course
   - By intravenous infusion in the intensive care unit: loading dose of 7 mg/kg dihydrochloride salt by infusion pump over 30 min, followed immediately by 10 mg/kg (maintenance dose) over 4 h, repeated after 8 h gaps as above.
   - By intramuscular injection (if intravenous infusion not possible): loading dose of 20 mg/kg of dihydrochloride salt diluted to 60 mg/ml, by deep intramuscular injection (half dose into each anterior thigh) with strict sterile precautions, then 10 mg/kg every 8–12 h until patient can take oral medication
   - Afterwards: when the 7 day course is completed, give tetracycline 250 mg four times daily for 7 days or doxycycline 100 mg daily for 7 days
2. Quinidine—replaces quinine for parenteral treatment of malaria in the USA
   - By intravenous infusion: loading dose of 15 mg/kg of quinidine base given over 4 h, then 7.5 mg/kg given over 4 h three times daily until the patient can swallow and take quinine. Followed by tetracycline or doxycycline as above
3. Artesunate
   - By intravenous 'push': loading dose of 2.4 mg/kg followed by 1.2 mg/kg at 12 and 24 h, then 1.2 mg/kg daily for 6 days
4. Artemether
   - By intramuscular injection: loading dose of3.2 mg/kg on the first day (in one or two doses), followed by 1.6 mg/kg/day for 6 days

Other measures

1. High fever—control by fanning, tepid sponging, cooling blankets, antipyretics (e.g. paracetamol 15 mg/kg in tablets, or powder washed down an NG tube, or as suppositories)
2. Anaemia—transfuse with whole blood or packed cells if haematocrit falls to <20% or if there is severe bleeding
3. Urine output—insert urinary catheter to monitorclosely
4. Cerebral malaria—appropriate nursing care for the unconscious patient. Control convulsions (see Emergency Medicine, section 6.5). Consider elective intubation and ventilation if airway in danger of compromise
5. Hyperparasitaemia—consider exchange transfusion or haemophoresis in non-immune patients who are severely ill, who have deteriorated on convention altreatment, and who have parasitaemia >10%

Key investigations

To establish the diagnosis

Depends on the detection of parasitaemia (stop antimalarial chemoprophylaxis)

1. Repeated examination of thick and thin blood films (8–12 hourly for 72 h) by an experienced microscopist
2. Antibody detection technique, e.g. dipstick antigen-capture assay

Note
If patient remains unwell and no other diagnosis can be made, then consider therapeutic trial even if early smears are negative

Other important tests

1. Reagent stick test for blood glucose— ?hypoglycaemia
2. Full blood count—anaemia with evidence of haemolysis is usual. Neutrophilia is common, but white cell count can be normal or low
3. Electrolytes, renal and liver function, glucose, coagulation screen—mild hyponatraemia is common
4. Arterial blood gases
5. Blood culture—to exclude secondary bacterial septicaemia in those with an obvious focus of such infection and in patients who are very unwell or have a raised blood white cell count
6. Depending on clinical context (mainly to exclude differential diagnoses)—CT brain, lumbar puncture, chest radiograph

Further management	Emphasize need for avoidance and prophylaxis with any future travel to malarious areas

7.2 Meningitis

See Chapter 7.11.3, Chapter 7.11.5, Chapter 7.11.12, and Chapter 24.14.1 in main text

Clinical features

Clinical features

Acute bacterial meningitis has a mortality of 70–100% if untreated and is the immediate concern in patients pre- senting to medical services

History

General symptoms

1. Early—malaise, headache, fever, vomiting, diarrhoea
2. Later—increasingly severe headache, photophobia, drowsiness
3. Very late—coma, convulsions

Localizing (if meningitis secondary to infection elsewhere)

4. Respiratory—pneumococcal disease (pneumonia)
5. Ear—*H. influenzae* (otitis media)

Also

6. Contact with a case of meningitis
7. Previous history of meningitis
8. History of immunodeficiency
9. Pregnancy—increased risk of Listeria
10. Travel history—particularly meningococcal disease

Examination

1. Vital signs—temperature, pulse rate, blood pressure, respiratory rate
2. General
   - Skin: petechiae/purpura—characteristic of meningococcal disease, but not specific
   - Conjunctivae/palate: petechiae—characteristic of meningococcal disease, but not specific
   - Posture—patients with severe meningism often lie with the back and neck in hyperextension
3. Neurological
   - Meningism—neck stiffness. Kernig's sign (with the leg flexed at the hip, an attempt by the clinician to passively extend the knee is resisted by hamstring spasm)
   - Ocular fundi—papilloedema indicates raised intracranial pressure, but absence of papilloedema does not exclude this
   - Cranial nerve lesions—most commonly VIth (false localizing sign)
4. Other—in secondary meningitis there may be signs of primary focus, e.g. pneumonia, otitis media, CSF shunts/reservoirs

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

If clinical evidence of intravascular volume depletion, establish intravenous access and resuscitate as described in Emergency Medicine, section 3.1

Antimicrobial chemotherapy (empirical treatment, adult dosages)

Spontaneous meningitis

Drug	Dose	Route	Frequency	Duration*
Cefotaxime	2g	IV	4 hourly	2 weeks
				Or
Ceftriaxone	2g	IV	12 hourly	2 weeks
If high prevalence of penicillin-resistent pneumococci, then add				
Vancomycin	1g	IV	12 hourly	2 weeks
If underlying immunosuppression, pregnancy or age >65 years, then add				
Ampicillin	2g	IV	4-6 hourly	3 weeks

Post-traumatic meningitis

Community acquired

Treat as for spontaneous meningitis

Nosocomial

Drug	Dose	Route	Frequency	Duration*
Probability of *Pseudomonas* spp. high				
Ceftazidine	2g	IV	8 hourly	3 weeks
				Or
Meropenem	2g	IV	8 hourly	3 weeks
Probability of *Pseudomonas* spp. low				
Cefotaxime	2g	IV	4 hourly	3 weeks
				Or
Ceftriaxone	2g	IV	12 hourly	3 weeks

Shunt-associated meningitis

Insidious onset

Drug	Dose	Route	Frequency	Duration*
Vancomycin	1g	IV	12 hourly	2 weeks
				PLUS
Vancomycin	5–10g	IT	48–72 hourly	2 weeks

Acute onset

Treat as for nosocomial post-traumatic meningitis

Notes

* Antimicrobial therapy can be refined as soon as organism is isolated, otherwise patients with suspected bacterial meningitis should receive treatment with the regimen indicated.

Doses of antimicrobials to be adjusted in renal failure (especially vancomycin)

IV, intravenous; IT, intrathecal.

Key investigations **To establish the diagnosis**

1. Epidemiological data (any current epidemics)
2. Lumbar puncture to obtain specimen of CSF— looking in bacterial meningitis for:

General appearance

- Cloudy or purulent, but can be clear

Microscopy

- White cell count—usually raised (although can rarely be normal, i.e. <6 lymphocytes/µl) with neutrophils accounting for >80% of cells, but can have a lymphocytic pleocytosis in early bacterial meningitis or with *L monocytogenes*
- Gram stain—shows organisms in 50–80% of cases

Biochemical testing

- Glucose—usually reduced (<40% that of a parallel serum sample)
- Protein—usually elevated (>0.45 g/l)

Microbiological culture

Bacterial antigen detection for common pathogens

PCR for meningococcal disease

Notes

1. Give antibiotics immediately—before referral to hospital, and if in hospital before lumbar puncture— in cases of suspected meningococcal meningitis/septicaemia
2. Lumbar puncture should not be performed if there are
   - Symptoms or signs to suggest raised intracranial pressure, namely
   - Drowsiness/coma
   - Focal neurological signs
   - Loss of retinal vein pulsation/papilloedema
   - Bradycardia/hypertension
   - Local skin sepsis at the sight of puncture
   - Clinical suspicion of spinal cord compression
   - Bleeding diathesis
3. If symptoms or signs suggest raised intracranial pressure—arrange for CT brain to exclude space occupying lesion or cerebral oedema

Other important tests

For specific diagnosis

1. Blood culture
2. Throat swab—for viral and bacteriological culture
3. Skin lesion—disrupt with needle and make contact slide for Gram stain.
4. Blood sample—in EDTA (as full blood count) for bacterial PCR

Other

1. Full blood count
2. Electrolytes, renal and liver function, glucose, clotting screen
3. Arterial blood gases (severe cases)
4. Chest radiography—?pneumonia (pneumococcal disease), ?aspiration (if impaired conscious level)
5. CT/MRI brain—may demonstrate skull fractures or parameningeal septic foci

Further management

1. Meningitis is a notifiable disease
2. If meningococcal meningitis
   - Household and other intimate contacts—give prophylaxis (e.g. rifampicin 600 mg orally twice daily for 2 days, or ciprofloxacin 750 mg orally as single dose) and immunize if serogroup C or A
   - Staff—prophylaxis is not required unless mouth to mouth resuscitation given

7.3 Encephalitis

See Chapter 7.10.2, Chapter 7.10.12, Chapter 7.10.6.1, and Chapter 24.14.2 in main text

Clinical features

Clinical features

Encephalitis is an acute inflammation of the brain and/or spinal cord (encephalomyelitis) presenting as alteration of consciousness, convulsions and/or focal neurological signs. It is usually caused by an acute viral infection of the central nervous system (typically Herpes simplex, Japanese encephalitis, or an arthropod-borne virus), or it complicates a systemic viral infection such as measles (post-infectious encephalomyelitis) or vaccination (post-vaccinal encephalomyelitis). Case fatality is extremely variable but may exceed 40% when there is no antiviral therapy (e.g. Japanese encephalitis), and there is a high incidence of permanent neurological sequelae

History

General symptoms (after incubation period of a few days to 2 weeks)

1. Early—fever, headache, neck stiffness, vomiting
2. Later—psychiatric symptoms, altered consciousness, convulsions

Localizing symptoms

3. Altered behaviour, hallucinations, temporal lobe seizures—Herpes simplex encephalitis
4. Rashes—preceding illness (e.g. measles, varicella, post-infectious encephalomyelitis); concurrent (e.g. West Nile virus encephalitis)

Also

5. Recent vaccination (vaccinia, nervous tissue rabies vaccine)
6. Current seasonal epidemic (arthropod-borne encephalitides)
7. Travel history—to endemic area (e.g. Central Europe/Scandinavia—tick-borne encephalitis)

Examination

1. Vital signs—temperature, pulse rate, blood pressure, respiratory rate, Glasgow Coma Scale
2. General
   - Skin: rashes—West Nile virus, enteroviruses etc.
   - Mucous membranes: cold sores (Herpes simplex encephalitis)
3. Neurological
   - Meningism
   - Ocular fundi—papilloedema indicates raised intracranial pressure, but absence of papilloedema does not exclude this
   - Cranial nerve lesions—most commonly VI (false localizing sign)
4. Other—in post-infectious encephalomyelitis there may be signs of the preceding illness, e.g measles, varicella, mumps etc.

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, Section 1.2

If convulsing, as described in Emergency Medicine, Section 6.5

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

1. Antiviral treatment
   - Aciclovir—where it is affordable, treatment with aciclovir should be started immediately in all undiagnosed cases in which viral encephalitis is included in the differential diagnosis. Specifically, aciclovir is recommended for Herpes simplex, Herpes simiae (B), Herpes zoster, and Epstein-Barr virus encephalitides: dose 10 mg/kg every 8 h by intravenous infusion (reduced in renal impairment)
   - Ribavirin (tribavirin)—for the rare encephalitides associated with RNA virus infections (e,g, Lassa fever, Argentine haemorrhagic fever, Hanta virus, Crimean-Congo haemorrhagic feverm and Rift Valley Fever) ribavirin (tribavirin) has been recommended: 2 g loading dose by intravenous infusion, then 1 g every 6 h for 4 days, then 0.5 g 8 hourly for 6 days
2. Other measures
   - Corticosteroids—sometimes used for post-vaccinal encephalomyelitis (controversial)
   - Reduction of severe intracranial hypertension— intravenous mannitol or mechanical hyperventilation

Key investigations

To establish the diagnosis

1. Epidemiological data (any current epidemics)
2. Lumbar puncture to obtain specimen of CSF— looking in viral encephalitis for:

   Microscopy

   - White cell count—usually raised (but normal in 10–15% of patients with Herpes simplex enceph-alitis at first examination), with lymphocytes and other mononuclear cells predominant except in early infections
   - Gram stain to exclude bacterial meningoencepha- litis

   Biochemical testing

   - Glucose—usually normal or increased, but low levels have been reported
   - Protein—usually elevated into range 0.5–1.5 g/l

   Virology

   - PCR
   - Specific viral IgM (microcapture technique)
   - Viral isolation—e.g. mumps, enteroviruses, lymphocytic choriomeningitis virus
3. Other samples

- Skin lesions—immunofluorescence (Herpes zoster) and electron microscopy (Herpesviruses)
- Nasopharyngeal aspirate—measles
- Stool—enteroviruses
- Serology (acute/convalescent titres)—mumps, Coxsackie viruses, arthropod-borne viruses

Other important tests

1. Full blood count
2. Arterial blood gases (severe cases)
3. CT/MRI—may demonstrate focal lesions (e.g. Herpes simplex encephalitis) or cerebral oedema

Note

The diagnosis of viral encephalitis should not be made too hastily as the differential diagnosis is broad and other treatable causes (e.g. cerebral malaria, bacterial or fungal meningoencephalitides) may be ignored

7.4 Tetanus

See Chapter 7.11.20 in main text

Clinical features

Tetanus, caused by toxins of *Clostridium tetani* in contaminated wounds, remains common in some developing countries but is preventable by vaccination. It is now rare in developed countries but, because it is decreasingly familiar, is less likely to be diagnosed. The case fatality ranges from 20–60%, although in expert hands this may be reduced to 6%, even in severe cases

History

1. Recent wound, especially penetrating, contaminated or with necrosis, is identified in 75–85% of cases

Also

2. Problems in head, neck, mouth—trismus due to a painful local condition is an important differential diagnosis
3. Drugs—a dystonic drug reaction is an important differential diagnosis

Symptoms of tetanus

After an incubation period of usually 6–10 days (less than 15 days in 90% of cases):

- Non-specific—malaise, fever, sweating, and headache
- Suggestive—muscle stiffness (especially of the jaws), spasms, and dysphagia.

Examination

Features of tetanus

1. Muscles
   - Trismus, risus sardonicus, neck retraction
   - Rigidity of erector spinae and abdominal muscles (board-like rigidity)
   - Opisthotonos
   - Tonic contractions/spasms of the stiff muscles
   - Spasms of respiratory muscles and larynx threaten to cause asphyxia
   - Local tetanus may involve only muscles in the region of the wound, e.g. cephalic tetanus
2. Autonomic nervous system
   - Fluctuating heart rate, blood pressure, and temperature with sweating and hypersalivation

   Clinical grading is of prognostic significance:

   - I (mild)—trismus and generalized stiffness without respiratory embarrassment or spasms
   - II (moderate)—marked rigidity, brief spasms, mild respiratory embarrassment, and dysphagia
   - III and IV—frequent prolonged spasms, respiratory embarrassment with apnoeic spells, severe dysphagia. and cardiovascular abnormalities

   Also

   - Features of alternative diagnosis, e.g. local cause of trismus

Note

1. Incubation period less than 4 days and period of onset (trismus to first spasm) less than 48 h are associated with high mortality
2. In developed countries, patients are often elderly (missed childhood vaccination)

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, Section 1.2

If convulsing, as described in Emergency Medicine, Section 6.5

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

1. If apnoeic/asphyxiating/hypoxaemic—emergency tracheostomy, assisted ventilation with oxygen. See Emergency Medicine, section 10.4.2
2. Tetanus immune globulin

In all cases—give before manipulating the wound:

- EITHER equine tetanus immune globulin 10000 units intravenously (beware of anaphylaxis, see Emergency Medicine, section 1.12)
- OR (preferably, if available) human tetanus immune globulin 5000 units intravenously

3. The wound

Antibiotics to sterilize

- EITHER—Metronidazole 500 mg, orally (if possible) or intravenously, three times a day for 10 days
- OR—Benzyl penicillin 2 megaunits intravenously four times a day for 8 days

Thorough surgical débridement of the wound after tetanus immune globulin has been given

4. Other measures

- Sedatives/muscle relaxants
- Diazepam 5–20 mg three times a day by mouth (mild cases) or by continuous intravenous infusion (moderate-severe cases)
- Tracheostomy (moderate and severe cases)
- Neuromuscular blockade (pancuronium/vecuronium) and mechanical ventilation (severe cases)

Further management (severe cases in the intensive care unit)

1. Ventilatory support
2. Control of autonomic nervous system disturbances
   - Hypertension—cautious use of low dose short-acting b-blockers
   - Brady/tachyarrhythmias—treat only if causing significant haemodynamic disturbance, see Emergency Medicine, Section 1.7
3. Prevention of deep vein thrombosis—low molecular weight heparin

Note

Avoid use of excessive doses of diazepam

Key investigations

THE DIAGNOSIS OF TETANUS IS ENTIRELY CLINICAL

1. Wound swab—but failure to culture *Clostridium tetani* from the wound does NOT exclude the diagnosis of tetanus
2. Lumbar puncture—the cerebrospinal fluid is normal

Note

The differential diagnosis includes the many local causes of trismus, dystonic reactions to drugs, tetany, strychnine poisoning, meningitis, and rabies (cephalic tetanus)

Further management

Infection does not confer immunity: give full course of active immunization (tetanus toxoid) after recovery

7.5 Rabies

See Chapter 7.11.9 in main text

Clinical features

Clinical features

Rabies is a zoonotic viral infection of the central nervous system, endemic in domestic dogs and cats, wild carnivores, bats etc., in most parts of the world. It is transmitted to humans by bites of rabid mammals, usually dogs. The case fatality of rabies encephalomyelitis is virtually 100%, but the disease is preventable by modern post-exposure treatment started soon after the bite

History

1. Animal contact
   - History of dog (or other mammal) bite (but may be distant or forgotten, especially with insectivorous bat bites in USA) or a lick by a mammal on broken skin
   - Travel history to rabies endemic area
   - Post-exposure treatment—see Emergency Medicine, section 7.6
2. Incubation period
   - Usually between 20 and 90 days (extreme range 4 days to 19 years)
3. Prodromal symptoms
   - Non-specific—fever, chills, malaise, weakness, tiredness, headache
   - Suggestive—itching, pain, or paraesthesiae at the site of the healed bite wound
4. A few days later
   - Furious rabies—difficulty swallowing (especially water), causing spasms of breathing and great anxiety (with or without pain in the throat)
   - Extreme susceptibility to draughts, causing similar spasms
   - Bizarre behaviour
   - Periods of extreme excitement, hallucinations, terror, aggression with lucid intervals
5. After several more days
   - Lapse into coma and convulsions
   - Sudden death during a hydrophobic spasm
   - Paralytic rabies—ascending weakness with sensory symptoms often starting in the bitten limb; sphincter problems; dysphagia, drooling, and respiratory weakness

Examination

1. Wound
   - Evidence of healed bite
2. Neurological
   - Clinical examination may be normal
   - Excitable behaviour interspersed with lucid intervals
   - Furious rabies—violent, jerky spasms of inspiratory muscles associated with evident terror provoked by attempts to drink or exposure to a draught of air
   - Paralytic rabies—ascending flaccid paralysis with fasciculations, sensory loss, sphincter dysfunction
   - Weakness of muscles of deglutition and respiration
   - Excitable behaviour interspersed with lucid intervals
3. Autonomic nervous system
   - Signs of overactivity—hypersalivation, sweating, labile pulse rate and blood pressure

Immediate management	Although life can be prolonged by invasive, intensive care (tracheostomy, paralysis, mechanical ventilation, cardiac monitoring etc.), the chances of a successful outcome are so low that there is a strong case for palliative care to relieve pain and anxiety
Key investigations	To establish the diagnosis during life 1. Skin punch biopsy (hairy area, e.g. nape of neck) • Detection of virus by direct fluorescent antibody in nerves surrounding hair follicles 2. Saliva • Virus may be isolated 3. Blood • Rabies-neutralizing antibody titre—elevated in unvaccinated patient (but may be negative for 7 days after clinical illness has begun) 4. CSF analysis • May be normal, but protein usually elevated, and may have elevated white blood cell count • Rapid PCR (experimental) • Virus may be isolated • Rabies-neutralizing antibody titre—elevated in unvaccinated patient (but may be negative for 7 days after clinical illness has begun) 5. Other important tests • Given the appalling outcome of rabies it is important to pursue the possible differential diagnosis of a rapidly progressing encephalitis (see Emergency Medicine, section 7.3) if there is ANY doubt about the diagnosis To establish the diagnosis in the biting animal (dog etc.)—brain smear with detection of virus by direct fluorescent antibody or viral isolation. Euthanizing the biting dog and examining its brain immediately is now recommended, rather than observing it for onset of rabid symptoms over a 10-day period
Further management	Attempt to identify/capture/examine (by veterinarian)/test the animal responsible for the bite

7.6 Animal bites/stings

See Chapter 7.11.9 and Chapter 8.2 in main text

Clinical features	A very wide range of animals may inflict bites and stings. Serious consequences may result from trauma, envenoming, allergy or infection

History

1. Timing
   - The event is usually painful and memorable and so precisely timed by the victim.
2. Immediate symptoms—distress associated with a terrifying event and attributable to trauma, envenoming or allergy
   - Trauma
     - Pain, bleeding, dysfunction (depending on site and severity of injury)
   - Envenoming
     - Snake bite

       Local—pain, swelling, persistent bleeding, bruising, blistering, painful enlargement of draining lymph nodes

       Systemic—syncope/collapse (may be early and transient), spontaneous systemic bleeding (gums, nose etc.), vomiting, progressive weakness starting with ptosis, blurred vision, inability to open mouth, swallow, speak etc., generalized muscle aches and tenderness, passage of black urine (rhabdomyolysis)
     - Scorpion sting

       Local—very severe pain, mild swelling
       Systemic—vomiting, sweating, faintness, difficulty with breathing, muscle spasms
     - Spider bites

       Local—pain, sweating and gooseflesh (neurotoxic) or progressive skin changes (red, white, and blue sign; necrotic)

       Systemic—vomiting, faintness, colic and muscle spasms
     - Jellyfish stings

       Local—severe pain, blistering, contact rash
       Systemic—collapse, vomiting
     - Fish stings

       Local—very severe pain
       Systemic—rarely collapse
   - Allergy
     - Hymenoptera stings (bees, wasps, hornets, yellowjackets, ants)

       Local—pain, swelling (may be negligible)
       Systemic—early syncope and collapse, raised, itchy rash, swelling of mouth, lips, tongue, and gums, chest tightness, wheezing, asthma attack, abdominal colic, vomiting, diarrhoea (all of these may develop within a few minutes of the sting)
3. Delayed symptoms—attributable to infection

   Earliest onset at about 12 h (*Pasteurella multocida*) Local—pain, swelling, redness, heat, purulent discharge

   Systemic—sometimes severe generalized symptoms (sepsis)

Examination

1. Vital signs

   Temperature, pulse rate, blood pressure, respiratory rate, Glasgow Coma Scale
2. Trauma

   Injuries to soft tissues, joints, tendons, bones (crush fractures), body cavities (e.g. haemothorax), evisceration, dead tissue, foreign material in the wound(broken teeth, claws, earth etc.). May be severe/life-threatening

   Note that effects of trauma may be associated withe nvenoming and/or allergy and/or infection (e.g. marine coral cuts, stingray, and sea urchin injuries)
3. Envenoming—see History
4. Allergy—features of anaphylaxis (Emergency Medicine, section 1.12)
5. Infection

   Local—pain, swelling, redness, heat, purulent discharge

   Systemic—sepsis syndrome (see Emergency Medicine, section 7.7)

Note

1. Human bites
   - May be of medicolegal significance: document carefully, also any other evidence of injury (sketch and photograph)
   - High risk of infection with group A b-haemolytic streptococci, *Staph. aureus* (40% of wounds), *Haemophilus, Klebsiella, Eikenella corrodens*, and anaerobes
   - May be self-inflicted—typically lips, buccal cavity, fingers, clenched-fist injuries of knuckles
2. Dog, cat/other mammal bites
   - Associated with high risk of infection with a wide range of pathogens, notably *Pasteurella multocida*, *Capnoctyophaga canimorsus*, *Staph. aureus, Clostridium tetani*, and other anaerobic bacteria, rabies virus etc.

Immediate management

First-aid

1. Trauma

   Control pain and bleeding, contain wound with bandaging, give plasma expander if available, transport to medical care
2. Envenoming

   Snake bite

   - Immobilize the patient, especially the bitten limb
   - Avoid harmful remedies
   - Transport the patient to medical care
   - Neurotoxic bites only—pressure immobilization and splinting with a long crepe bandage
   - Venom ophthalmia (spitting cobras and rinkhals)—irrigate affected eye with liberal quantities of bland fluid (e.g. water, milk) and apply 1% epinephrine drops for pain (if available)

   Note

   - Tourniquets, incisions, suction, electric shock, cryotherapy, snake stones etc. should NEVER be used
3. Other bites and stings
   - Fish stings—immerse stung part in uncomfortably-hot but not scalding water (maximum 45°C)
   - Scorpion stings and other painful bites and stings—local 1% lignocaine with digital block or strong systemic analgesia (if available)
   - Jellyfish stings

     Box jellyfish (North Australia, Indo-Pacific)— wash area with dilute acetic acid/vinegar

     Atlantic jellyfish—apply a slurry of baking powder
   - Bee stings—remove the sting as quickly as possible
   - Tick bites—apply surgical spirit to the animal and prise out the mouth parts with forceps

Hospital management

If anaphylaxis, as described in Emergency Medicine, section 1.12

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

If clinical evidence of intravascular volume depletion, establish intravenous access and resuscitate as described in Emergency Medicine, section 3.1

1. Trauma
   - Explore wound under anaesthesia, débriding and removing foreign material.
   - Treat specific injuries to vital structures
   - Delayed primary suture.
2. Envenoming

   Antivenom treatment

   In cases of envenoming by snakes, fish, scorpions, spiders, box jellyfish, and ticks, administer antivenom intravenously (provided that an appropriate specific antivenom is available) if any of the following are present:

   - Paralysis (ptosis etc)
   - Spontaneous systemic bleeding (gums, GI tract etc)
   - Incoagulable blood
   - Shock, ECG abnormalities
   - Black urine (myoglobinuria, haemoglobinuria)
   - Severe/rapidly progressive local envenoming

   Beware of antivenom reactions (anaphylactic or serum): treat/prevent as follows:

   - Treatment with adrenaline (1/1000, 0.3–0.5 ml intramuscularly = 0.3–0.5 mg dose, repeated as necessary) plus anti-H_1 (e.g. chlorpheniramine 10–20 mg IV) plus corticosteroid (e.g. hydrocortisone 5 mg/kg IV) at the first sign of a reaction is preferred to routine prophylaxis EXCEPT in atopic subjects with severe asthma and those who have suffered previous reactions to antivenom. See Emergency Medicine, section 1.12

Key investigations

Trauma

- Appropriate radiological imaging to define extent of the injury

Snake bites

- Simple 20 min whole blood clotting test or rapid coagulation screen
- Stick test urine for blood: positive may indicate red blood cells (?disseminated intravascular coagulation) or myoglobin (rhabdomyolysis)
- Australia only—rapid EIA venom detection kit, using swab from the bite wound
- Tensely-swollen limbs—measure intracompartmental pressure as guide to fasciotomy

Other important tests

- Depending on clinical context/severity—ECG, full blood count, electrolytes, renal and liver function tests, muscle enzymes (creatine kinase), arterial blood gases, chest radiograph

Further management

Trauma (bites by large animals)

1. Definitive wound closure with skin grafts etc.
2. Infection risk
   - Bacterial

     Prophylactic antibiotics for severe/multiple wounds or wounds of the fingers or in response to cultures:

     Amoxicillin/clavulanic acid—(expressed as) amoxicillin 250 mg three times daily by mouth (prophylaxis, mild case) to 1 g three times daily intravenously (treatment, severe case)
     OR

     Second/third generation cephalosporin, e.g.

     cefotaxime 1–2g 6 hourly intravenously
   - Tetanus

     Give tetanus toxoid or, if unimmunized, consider tetanus immunoglobulin
   - Rabies

     Consider possibility of rabies exposure and (if appropriate) give **rabies post-exposure treatment**

     - Thorough wound cleaning—scrub under running tap with soap and water; irrigate with plain water
     - Apply virucidal agent such as 40–50% alcohol or 1% iodine
     - Avoid suturing.
     - Vaccination:
       - Start active vaccination using tissue culture vaccine: dividing one dose (0.5–1 ml) between 8 sites intradermally produces the most rapid antibody response) PLUS
       - Start passive immunization: give equine rabies immunoglobulin, 40 units/kg body weight, OR—preferably if available—human rabies immunoglobulin, 20 units/kg body weight, infiltrate around the wound, with the residue given intramuscularly distant from the site of rabies vaccination

Envenoming

1. Nursing
   - Avoid elevation of the bitten limb
2. Surgery
   - Débridement of necrotic tissue with immediate split skin grafting
   - Avoid hasty and unjustified fasciotomy (especially if the blood is still incoagulable).
3. Myoglobinuric renal failure—try to prevent by correcting hypovolaemia and acidosis and encouraging diuresis (see Emergency Medicine, section 4.2)

7.7 Septic shock

See Chapter 4.4 and Chapter 7.5 in main text

Clinical features	Clinical features
	Septic shock is a condition associated with severe infection in which there is hypotension (systolic blood pressure <90 mmHg) unresponsive to fluids or requiring vasoactive drugs for its correction. The causative septicaemia may be with Gram-positive or Gram-negative bacteria, yeasts, viruses, or protozoa. Failure of one or more organ systems is common

History

1. Systemic features
   - Early—malaise, lethargy, nausea, vomiting, fever, sweating, shivering/rigors
   - Later—restless, anxious, confused, agitated.
   - May develop rapidly (minutes—hours), e.g. meningococcaemia, or gradually
2. Localized features
   - Related to causative infection—e.g pneumonia, urinary tract infection, infected intravascular catheter, meningitis, after large bowel surgery etc.
3. Risk factors
   - Complication of surgery, instrumentation, burns, or other trauma.
   - Complication of preceding illness, e.g 'flu predisposing to staphylococcal pneumonia
   - Travel history—could the patient have malaria? (see Emergency Medicine, section 7.1)

Examination

1. Vital signs
   - Temperature—fever or hypothermia
   - Tachycardia
   - Tachypnoea
   - Hypotension
   - Peripheries warm (vasodilated) or cold and cyanosed (vasoconstricted)
   - Glasgow Coma Score
2. Evidence of the causative infection
   - Complete physical examination to look for focus of infection. Do not forget to examine the back and perineum/rectum (localized abscess).
3. Evidence of organ failure
   - Respiratory—central cyanosis (check pulse oximetry), crackles. Risk of adult respiratory distress syndrome (ARDS)
   - Renal – low urine output. Risk of prerenal renal failure or acute tubular necrosis
   - Liver—jaundice
   - Neurological—confusion
   - Haematological—abnormal bleeding/gangrene of extremities

Notes

1. Look for evidence of predisposition to infection—elderly, immunosuppressed, asplenic, malignant disease, artificial heart valve, prosthetic material etc.
2. Streptococcal toxic shock syndrome—erythematous rash, local severe pain and swelling (necrotizing fasciitis/myositis)
3. Staphylococcal toxic shock syndrome—diarrhoea, myalgia, rash (desquamating), often associated with menstruation/tampon use

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%

If clinical evidence of intravascular volume depletion, establish intravenous access and resuscitate as described in Emergency Medicine, section 3.1

1. Fluid/circulatory
   - 500 ml crystalloid IV every 30 min until central venous pressure is 8–12 mmHg
   - If mean arterial pressure <65 mmHg—give vasopressors
   - If mean arterial pressure >90 mmHg—give vasodilators
   - Central venous oxygen saturation <70% (measured in blood from pulmonary artery or central venous catheter)—transfuse packed red cells to increase haematocrit to at least 30% and give dobutamine initial dose 2.5 µg/kg/min intravenously, increasing until central venous oxygen saturation is 70% or higher
2. Respiratory
   - Consider early intubation and mechanical ventilation
3. Antibiotics

Give broad-spectrum, empirical treatment

Community-acquired septicaemia:

- Aminoglycoside (e.g. gentamicin 5 mg/kg intravenously once daily, assuming normal renal function) + broad-spectrum penicillin (e.g. amoxicillin, 500 mg 8 hourly to 1 g 6 hourly, intravenously)
  OR
- Broad-spectrum cephalosporin (e.g. cefotaxime, 1 g 12 hourly to 2 g 6 hourly, intravenously)

Hospital-acquired septicaemia:

- Aminoglycoside (e.g. gentamicin 5 mg/kg intravenously once daily, assuming normal renal function) + broad-spectrum anti-Pseudomonal penicillin (e.g. Tazosin® 2.25–4.5 g [= Piperacillin 2–4 g + tazobactam 250–500 mg] 6 hourly intravenously)
  OR
- Ceftazidime, 1 g 12 hourly to 2 g 8 hourly, intravenously
  OR
- Meropenem 500 mg–1 g 12 hourly intravenously
  OR
- Imipenem with cilastatin, 500 mg–1 g (of imipenem) 6 hourly intravenously

Pseudomonas infection suspected:

- Aminoglycoside (e.g. gentamicin 5 mg/kg intravenously once daily, assuming normal renal function) + broad-spectrum anti-Pseudomonal penicillin (e.g. Tazosin® 2.25–4.5 g [= Piperacillin 2–4 g + tazobactam 250–500 mg] 6 hourly intravenously)

Gram positive infection suspected:

- Add flucloxacillin 1–2g 12 hourly intravenously
  OR
- Vancomycin 1g 12 hourly intravenously (assuming normal renal function)

Anaerobic infection suspected:

- Add metronidazole 500 mg 8 hourly intravenously

Meningococcaemia

- Benzylpenicillin 2.4 g 4 hourly intravenously
  OR
- Cefotaxime 2 g 6 hourly intravenously

Streptococcal toxic shock syndrome

- Benzylpenicillin 1.2–2.4 g 6 hourly intravenously, or Cefotaxime 2 g 6 hourly intravenously
  PLUS
- Clindamycin 600–1200 mg 6 hourly intravenously

Note

1. Aminoglycosides, vancomycin—dosage dependent on renal function; always monitor levels
2. Other aspects
   - For patients with shock, acidosis, oliguria, or hypoxaemia with evidence of end organ dysfunction—consider recombinant human activated protein C
   - Supportive treatment for specific organ failure—mechanical ventilation, renal replacement therapy (haemofiltration, haemodialysis)
   - Surgical—e.g. urgent fasciotomy and débridement for streptococcal necrotizing fasciitis/myositis
   - Strict normalization of blood glucose between 4.4 and 6.1 mmol/l using intravenous infusion of Actrapid insulin on sliding scale

---

Key investigations

To establish the source of infection

1. Blood culture
2. Other cultures as determined by clinical signs or imaging, e.g. needle aspiration of fluid collections

Other

3. Full blood count—leucocytosis or leucopenia
4. Electrolytes, renal and liver function tests, glucose, clotting screen, muscle enzymes (creatine kinase, ?rhabdomyolysis)
5. Arterial blood gases—pH, *p*O2, *p*CO2, base excess

8 Psychiatry

8.1 Acute alcohol withdrawal

See Chapter 26.3 and Chapter 26.7.3 in main text

Clinical features	**History** Related to alcohol withdrawal 1. Agitation and anxiety 2. Tremor 3. Sweating 4. Insomnia 5. Nausea and vomiting 6. Hallucinations—tactile, visual, auditory 7. Grand mal seizures. Also 8. Drinking history—how much alcohol does the patient usually drink? Have they recently been drinking particularly heavily, or have they stopped?
	Examination Related to alcohol withdrawal: 1. Agitation and anxiety 2. Confusion 3. Tremor 4. Sweating 5. Tachycardia and hypertension Related to clinical context: 6. Cardiovascular—look for evidence of intravascular volume depletion and/or dehydration 7. Consider other complications of alcoholism • Wernicke's encephalopathy: • Ophthalmoplegia • Horizontal and vertical nystagmus • Weakness/paralysis of lateral rectus muscles • Weakness/paralysis of conjugate gaze • Ataxia—predominantly affecting stance and gait, often without clear-cut intention tremor • Acute liver failure • Chronic liver disease and its complications 8. Consider other causes of an acute confusional state–see Emergency Medicine, section 6.2 9. Nutritional status
Immediate management	1. Sedation • Patient can take oral medication—reducing schedule of chlordiazepoxide, e.g. 30 mg four times daily (day 1); 20 mg three times daily and 30 mg at night (day 2); 10 mg three times daily and 20 mg at night (day 3); 5 mg three times daily and 10 mg at night (day 4); 5 mg in the morning and 10 mg at night (day 5); 5 mg at night (day 6), then stop • Patient cannot take oral medication—clomethiazole (chlormethiazole), 0.8% solution, initially 2.5–7.5 ml/min (20–60 mg/min) until light sleep is induced from which the patient can easily be roused, with the rate of infusion then reduced to the lowest possible to maintain this state. Note—careful monitoring for respiratory depression is required: resuscitation facilities must be available. Switch to oral sedation when possible 2. Thiamine—give parenteral thiamine immediately, usually in combination with other vitamins B and C as PabrinexÔ I/V high potency, 2–3 pairs of ampoules intravenously over 10 min every 8 h (each pair of ampoules contains ascorbic acid 500 mg, anhydrous glucose 1 g, nicotinamide 160 mg, pyridoxine hydrochloride 50 mg, riboflavin 4 mg and thiamine hydrochloride 250 mg in a total of 10 ml). Note—facilities for treating anaphylaxis should be available Then 3. Glucose—treat/prevent hypoglycaemia. • DO NOT GIVE GLUCOSE BEFORE THIAMINE—DANGER OF PRECIPITATING WERNICKE'S ENCEPHALOPATHY • If hypoglycaemic—give 50 ml of 50% glucose (dextrose monohydrate) intravenously. IF IN DOUBT, TREAT • If not hypoglycaemic—start 5% dextrose infusion at 50 ml/h to prevent hypoglycaemia (if hyponatraemic used reduced volume of more concentrated dextrose solution)
Key investigations	**To establish the diagnosis** Acute alcohol withdrawal is a clinical diagnosis **Other Important tests** Depending of clinical context, consider as for acute confusional state—see Emergency Medicine, section 6.2
Further management	1. After 2 days, switch from intravenous to oral vitamin replacement, e.g. thiamine 50 mg once daily + Vitamin B tablets, Compound, Strong, 1–2 tablets three times daily + Vitamin C 100 mg once daily 2. Other aspects: as for acute confusional state—see Emergency Medicine, section 6.2—except avoid antipsychotics which lower seizure threshold 3. Long term—measures to help alcoholism

8.2 Drug overdosage

See Chapters 8.1 in main text

Clinical features

History

The overdose

1. Nature, time and quantity of drug ingested
2. Circumstantial evidence
3. Concurrent alcohol consumption

Also

4. Assessment of intent
5. Past medical history, medications and allergies
6. Past psychiatric history

Note

Be cautious in accepting the patient's account at face value. Assume that overdoses of multiple drugs are likely

Examination

Initial survey

1. Airway, breathing, circulation
2. Reagent stick test for blood glucose (?hypoglycaemia)
3. Check for small pupils and slow respiratory rate (?opioid overdose)
4. Check temperature (?hypothermia)
5. Check Glasgow Coma Score (see Table 18)

Further examination

6. Look for features indicated in Table 20.

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Nurse in recovery position if Glasgow Coma Score impaired

Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%.

Consider oropharyngeal airway or cuffed endotracheal tube depending on level of consciousness

ECG monitor—but do not treat arrhythmias unless these are associated with profound hypotension

Establish intravenous access

1. If hypoglycaemic—give 50 ml of 50% glucose (dextrose monohydrate) intravenously. IF IN DOUBT, TREAT
2. If possibility of opioid overdose—give naloxone 0.8–2 mg intravenously, repeated at intervals of 2–3 min to a maximum of 10 mg. IF IN DOUBT, TREAT
3. If hypothermic—start rewarming

Prevention of drug absorption—see Table 21

Specific antidote (if available)—see Table 22

If in doubt—discuss management with a Poisons Centre: the following single number for the UK National Poisons Information Service directs the caller to the relevant local centre: 0870 600 6266

Key investigations

To establish the diagnosis

1. Serum drug levels, e.g. paracetamol, salicylates, iron, theophylline, lithium
2. Save serum sample for measurement of other toxins after discussion with clinical chemist, e.g. paraquat

Note

Record time of blood sampling accurately on specimen tube and in notes

Other Important tests

1. Electrolytes, glucose, renal, liver and bone function tests, full blood count, clotting screen
2. ECG

Consider

3. Arterial blood gases
4. Carboxyhaemoglobin level
5. Chest radiograph—look for evidence of aspiration or pulmonary oedema
6. Abdominal radiograph

See Table 23

Further management

1. Dependent on the nature of overdose taken
2. As dictated by psychiatric condition (if any)

9 Other conditions

9.1 Disseminated intravascular coagulation

See Chapter 22.5.5 and Chapter 22.5.6 in main text

Clinical features

Disseminated intravascular coagulation (DIC) is a systemic disorder in which haemorrhage (main problem in 90% of cases) and thrombosis can occur at the same time. It involves the generation of intravascular fibrin and the consumption of procoagulants and platelets. May be acute or chronic (only acute discussed here)

History
Presence of DIC

1. Bleeding
   - Skin—extensive superficial bruising; oozing from venepuncture/intramuscular injection sites, around indwelling catheters/tubes
   - Mucosa—mouth, nose, gastrointestinal tract, (lungs), (renal tract)
   - Internal—brain, other organs
2. Thrombosis
   - Microthrombotic lesions
   - Skin—often on fingers/toes
   - Internal organs—dysfunction of brain, kidneys, lungs

Related to cause of DIC

1. Sepsis—bacterial, viral, fungal, parasitic (malaria)
2. Major trauma—including burns, surgery
3. Toxins—e.g. venoms (see Emergency Medicine, section 7.6)
4. Obstetric—placental abruption, eclampsia, amniotic fluid embolism
5. Cancer—metastatic carcinoma of stomach, colon, pancreas, breast, lung; mucin-secreting adenocarcinomas; leukaemia (especially acute promyelocytic leukaemia)
6. Blood transfusion—incompatible, massive
7. Liver disease—acute hepatic failure
8. Others—heatstroke (see Emergency Medicine, section 9.3), prosthetic devices (e.g. shunts, ventricular assist devices)
9. Idiopathic—purpura fulminans

Examination

1. Vital signs
2. Evidence of bleeding or thrombosis
3. Related to possible cause (see above)

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. Underlying cause
   Treat aggressively
   Give broad-spectrum antimicrobials to cover sepsis if diagnosis not clear (see Emergency Medicine, section 7.7)
2. When diagnosis of DIC established by laboratory testing, give (as appropriate)
   - Fresh-frozen plasma—to keep prothrombin time and activated partial thromboplastin time below a value 1.5 times the upper limit of control values
   - Cryoprecipitate/fibrinogen concentrates—to keep fibrinogen levels >1g/l
   - Platelets—to keep platelets >50×109/l
   - Blood (packed red blood cells)—to keep haematocrit >0.30

Note

1. If the patient continues to bleed/clot 4–6 h after initiation of treatment of underlying cause and the supportive measures described above, then—ONLY WITH EXPERT HAEMATOLOGICAL ADVICE—consider:
   - Antithrombin III, 100 U/kg intravenously over 3 h (loading dose), then continuous infusion of 100 U/kg/24h. Used in moderate/severe DIC when levels of antithrombin III are very low
   - Heparin, 20000–30000 units/24h, by continuous intravenous infusion—to inhibit further thrombogenesis. Used when thrombosis is the main clinical problem
2. Possibility of adrenal infarction (Waterhouse Friederichson)—give steroid (e.g. hydrocortisone 50–100mg 6hourly intravenously) if circulatory compromise

Key investigations

To establish the diagnosis
There is no single diagnostic test for DIC: look for the following

1. Appropriate clinical context
2. Platelet count—decreased
3. Prothrombin time and activated partial thromboplastin time—both increased
4. Fibrinogen/fibrin degradation products (FDPs) and/ or D-dimer—both present/elevated

Other haematological features that may be present include

1. Antithrombin III level—reduced: useful test for diagnosis and therapeutic monitoring
2. Fibrinogen—reduced
3. Fibrinopeptide A—breakdown product of fibrinogen, elevated
4. Thrombin time—prolonged
5. Blood film—may show red cell fragmentation/microangiopathic haemolytic anaemia

Other Important tests
Dependent on clinical context

Further management

Dependent on clinical context

9.2 Sickle cell crises

See Chapter 22.4.7 in main text

Clinical features

History

There are several clinical conditions

1. Pain crisis—severe pain in limbs, hips, back, chest or abdomen
2. Chest/lung syndrome—breathlessness, pleuritic chest pain
3. Brain/neurological syndrome—epileptic fits, transient ischaemic attacks, strokes And less commonly in adults
4. Aplastic crisis—presents with breathlessness and fatigue. Usually seen in children. Associated with parvovirus infection
5. Sequestration crisis—presents with profound anaemia. Usually seen in babies and young children when the spleen and/or liver enlarge rapidly due to trapping of red blood cells. Hepatic sequestration can occur in adults
6. Priapism
   Also
7. Previous sickle cell crises.
8. Precipitating factors—extremes of heat and cold, infections/fever (often upper respiratory tract, 'flu), heavy exercise, emotional stress, any situation producing hypoxia
9. Family history—patterns of crises may follow through generations

Note

1. The patient or their relatives/friends generally know that they have sickle cell disease and are often knowledgeable about the condition
2. The pain is
   - Genuine
   - Excruciating
   - Varies in character and location

Examination

1. Airway, breathing, circulation
2. Glasgow Coma Scale
3. Vital signs—pulse rate, blood pressure, respiratory rate, temperature Note particularly
4. General examination—pallor
5. Chest—local tenderness and signs of infection.
6. Bones—tenderness
7. Liver or spleen—look for enlargement due to hepatosplenic sequestration (particularly in children, uncommon in adults)
8. Priapism.
9. Infection—fever may be the only sign Remember that there may be no localizing signs

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

1. Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%. Monitor with pulse oximeter.
2. Fluid—establish intravenous access (may be difficult, asking patient's advice on best site is often helpful) and give 1 litre of 0.9% saline rapidly, then repeat 4– 6 hourly for duration of crisis (assuming satisfactory urine output)
3. Analgaesia
   - Intravenous e.g. (1) diamorphine by slow intravenous injection at 1 mg/min, usual maximum initial dose is 5 mg, but may be repeated if necessary, or morphine by slow intravenous injection at 2 mg/min, usual maximum initial dose is 10 mg, but may be repeated if necessary. Both to be accompanied by appropriate antiemetic, e.g. metoclopramide 10 mg IV over 1–2 min, or cyclizine 50 mg IV over 1–2 min
   - Intramuscular/subcutaneous—if it is not possible to establish intravenous access, then give diamorphine (0.05 mg/kg) or morphine (0.1 mg/kg) subcutaneously, repeated after 1 h if necessary
4. Warmth—wrap in warm blankets if the patient feels cold
5. Antibiotics—e.g. amoxicillin 500 mg intravenously every 8 h + benzypenicillin 1.2 g intravenously every 6 h
6. Prophylaxis against venous thromboembolism—give low molecular weight heparin, e.g. enoxaparin 20 mg subcutaneously once daily

Key investigations

To establish the diagnosis

Sickle cell crisis is a clinical diagnosis

Other Important tests

1. Full blood count, reticulocytes, group and save
2. Electrolytes, renal and liver function, glucose
3. Cultures—blood, sputum, urine, throat swab
4. Arterial blood gases
5. Chest radiograph—may show widespread patchy infiltrate that is difficult to distinguish from infection
6. HbS level (percentage of total Hb)—if exchange transfusion considered
7. Abdominal radiograph, serum amylase (abdominal syndrome)
8. Serology to detect acute parvovirus B19 infection (aplastic crisis)

Notes

Other investigations as dictated by clinical context, e.g.

CT scan brain if focal neurological signs

Do not order plain radiographs of all sites of pain

Further management

Seek expert advice

Chest syndrome with hypoxia, neurological symptoms or priapism (also seek urological advice) are all indications for exchange transfusion to reduce the HbS to <30% of total Hb

Consider hydroxyurea to reduce frequency of crises

9.3 Heat stroke

See Chapter 8.5.1 and Chapter 26.6.1 in main text

Clinical features

Hyperthermia is a failure of thermal homeostasis that allows the core temperature to rise above 40°C. It can result from exposure to environmental heat with/without prolonged physical exercise (especially if heat dissipating mechanisms are impaired) and/or from increased metabolic heat production. Heat stroke is hyperthermia with severe central nervous system abnormalities such as delirium, convulsions, or coma. Its case fatality ranges from 17–70%

History

Predisposing factors

1. Exposure to high ambient temperature ± high humidity (e.g. in a heatwave)
2. Prolonged strenuous physical exercise at any ambient temperature, especially if insulation from clothing is excessive and the patient is unacclimatized
3. Drugs
   - Neuroleptic malignant syndrome—psychiatriac/ recreational dopaminergic drugs, e.g. phenothiazines, thioxanthene, butyrophenones, amphetamines
   - Malignant hyperpyrexia—occurs in people with a rare genetic predisposition (autosomal dominant) when exposed to various inhaled or local anaesthetic agents
4. Previous history of heat intolerance

Symptoms

Heat exhaustion

1. General—irritability, weakness, lethargy, fatigue, dizziness, headache, muscle cramps/myalgias
2. Gastrointestinal—nausea, vomiting, diarrhoea
3. Respiratory – hyperventilation/tachypnoea
   Heatstroke
   Any or all of the symptoms of heat exhaustion, plus
4. CNS dysfunction—impaired judgement, abnormal behaviour, disorientation, hallucinations, confusion, convulsions, loss of consciousness

---

Examination

1. Vital signs—core (rectal) temperature 40° or more (by definition), hypotension, tachycardia, tachypnoea
2. Related to temperature control mechanisms
   - Sweating—present (>50% cases)or absent (<50%). Hot dry skin is a late finding
   - Piloerection
3. General—weakness, dehydration, bleeding (disseminated intravascular coagulation)
4. Neurological—impaired consciousness (Glasgow Coma Scale), seizures, opisthotonus, decerebrate rigidity, cerebellar dysfunction, oculogyric crises, fixed dilated pupils
5. Signs of predisposing condition—e.g. obesity, skin disease, thyrotoxicosis

---

Note

Muscular rigidity is a feature of neuroleptic malignant syndrome and malignant hyperpyrexia but not of heatstroke

---

Immediate management

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2
Oxygen, high flow with reservoir bag if needed, to achieve PaO_2 >92%
If clinical evidence of intravascular volume depletion, establish intravenous access and resuscitate using isotonic saline as described in Emergency Medicine, section 3.1

1. Cooling
   Should be done rapidly
   Remove to shade or cooler place
   Remove clothes
   - External methods
     Tepid spongeing and fanning; ice packs to neck, axillae, groin; cover with wet sheet—as available
     Hypothermia bed/blanket
     Immersion in cold water with vigorous
     massage—effective at lowering body temperature but associated with more complications than evaporative cooling and not generally recommended
   - Internal methods
     Ice water gastric lavage, ice water rectal lavage, ice water intraperitoneal lavage, mechanical ventilation with cooled gases, cardiac bypass— anecdotal success reported for these treatments
2. Monitor urine output—consider urethral catheter
3. Correct fluid, electrolyte and acid-base abnormalities (see below)
4. Malignant hyperpyrexia and neuroleptic malignant syndrome
   - Stop drug/anaesthetic
   - Treat muscle spasms/rigidity—give dantrolene sodium by rapid intravenous injection, 1 mg/kg repeated as required to a cumulative maximum of 10 mg/kg

---

Notes

1. In the field—stop exertion at the first sign of heat exhaustion, remove excess clothing, move into shade, encourage oral fluids
2. Do not give salicylates—which do not reduce body temperature and may exacerbate coagulopathy
3. Do not give paracetamol (acetaminophen)—which does not reduce body temperature and may worsen hepatic damage

---

Key investigations

Heat stroke is a clinical diagnosis

Important tests

1. Electrolytes and renal function—risk of hypokalaemia, hyponatraemia, hypocalcaemia, hypomagnesaemia, hypophosphataemia, impaired renal function (hyperkalaemia if renal failure and rhabdomyolysis)
2. Glucose—may have hyperglycaemia; also risk of hypoglycaemia
3. Full blood count—high haematocrit indicates haemoconcentration
4. Liver blood tests—elevated transaminases almost always found in heatstroke and indicate hepatotoxicity and/or muscle damage
5. Muscle enzymes—elevated creatine kinase indicates muscle damage (rhabdomyolysis)
6. Albumin/serum proteins—elevated values indicate haemoconcentration
7. Coagulation tests—evidence of DIC (see Emergency Medicine, sectioin 9.1)
8. Amylase—?pancreatitis
9. ECG, troponin—evidence of myocardial damage
10. Arterial blood gases—?lactic acidosis

Other tests that may be indicated include

1. Chest radiograph—?aspiration
2. CT scan head—?alternative cause of CNS dysfunction

Further management

1. Monitor results of active cooling
   - Beware of seizures (see Emergency Medicine, section 6.5)
   - Beware of cardiac arrhythmias—only treat if causing significant haemodynamic compromise (see Emergency Medicine, section 1.7)
2. Supportive care as appropriate
   Respiratory failure—consider intubation and mechanical ventilation
   Circulatory failure—treat hypotension with volume repletion and, if necessary, vasopressor drugs
   Renal failure—in case of rhabdomyolysis, prevent renal damage by correcting acidosis and hypovolaemia and promoting diuresis (see Emergency Medicine, section 4.2)
   Sodium depletion (hyponatraemia, muscle cramps)—sodium repletion with isotonic saline

Note

After recovery—advice to prevent recurrence

9.4 Hypothermia

See Chapter 8.5.2 in main text

Clinical features

History

Two distinct contexts

1. Cold exposure, in patient of any age
2. Multifactorial cause, often in the elderly patient
   - Immobility/falls
   - Cognitive impairment
   - Alcohol
   - Vasodilating drugs
   - Autonomic dysfunction, eg. diabetes mellitus
   - Poor socio-economic conditions

Note

Elderly patients with hypothermia have often been found on the floor at home following a fall

Examination

Initial survey

1. Airway, breathing, circulation
2. Reagent stick test for blood glucose (?hypoglycaemia)
3. Check for small pupils and slow respiratory rate (?opioid overdose)
4. Temperature—using a low-range rectal thermometer: <35°C (by definition).
5. Check Glasgow Coma Score (see Table 18)

Futher examination

1. General appearance—cold, pale mottled skin (whereabouts on the body, if anywhere, does the skin feel warmer?). At 32–35°C, shivering; at <32°C, muscular rigidity
2. Cardiovascular—bradycardia and hypotension
3. Respiratory—look for evidence of aspiration, pneumonia or pulmonary oedema
4. Neurological
   - May range from mild inco-ordination to confusion, lethargy, and coma. Pupils may be dilated and nonreactive
   - Focal/lateralizing signs may indicate stroke that has precipitated hypothermia
5. Endocrine—could the patient be hypothyroid?

Immediate management

Depends on the clinical context

If cardiorespiratory collapse, as described in Emergency Medicine, section 1.2

Treat for hypoglycaemia and/or opioid overdose if appropriate

The patient with hypothermia of gradual onset (usually elderly, usually multifactorial cause)

1. Re-warm—slowly in warm room covered with a blanket or 'Bair hugger' (do not use foil blankets which retard re-warming)
2. Oxygen—give oxygen as necessary to keep PaO_2 >92%
3. Fluids—establish intravenous access. Note risk of pulmonary oedema, hence do not infuse fluid rapidly. If hypotensive, infuse 1 litre 0.9% saline over 2 h, checking for lung crackles and/or worsening gas exchange as this progresses. Repeat or slow rate of infusion as determined by clinical response
4. ECG monitoring—risk of ventricular tachycardia/ fibrillation
5. Antibiotics—as appropriate for pulmonary infection, which is common in this context, e.g. cefotaxime 1 g 12 hourly intravenously

The patient with cold exposure or severe hypothermia (<30°C) or with hypothermia of any cause complicated by life threatening arrhythmia
Re-warm—rapidly, using both:

1. Active external re-warming—apply heat to body surface, e.g. hot water bottles/warmed IV bags (not too hot, must be comfortably bearable against your own skin) in groins and axillae, warmed blankets, radiant heaters
2. Active core re-warming
   - Non-invasive—give warmed (42–46°C) humidified oxygen and warmed (43°C) intravenous fluids
   - Invasive—gastric, colonic, bladder, and peritoneal lavage with warmed (43°C) 0.9% saline solutions

Notes

1. Patients with hypothermia are best managed in an ICU setting: they may require treatment of arrhythmia and/or ventilatory support
2. Arrhythmias
   - Avoid use of catecholamines (arrhythmogenic)
   - Only treat if life-threatening (ventricular fibrillation or asystole)—For VF, attempt defibrillation up to three times, but not more until core temperature >30°C. The drug of choice is probably bretylium 5–10 mg/kg intravenously over 15–30 min, repeated after 1–2 h to total dose of 30 mg/kg. Magnesium sulphate (8 mmol of magnesium) given intravenously over 10–15 min (repeated once if necessary) has also been reported to be effective. Lignocaine (lidocaine) is not effective in hypothermic VF, and the International Liaison Committee on Resuscitation guidelines suggest avoiding this drug, also epinephrine and procainamide, because of the risk of accumulation to toxic levels
3. Diagnosis of death—this can be difficult. Patients with severe hypothermia can appear clinically dead. Resuscitative efforts must continue until the core temperature is 30–33°C, i.e. THE PATIENT IS NOT DEAD UNTIL THEY ARE WARM AND DEAD

Key investigations	**To establish the diagnosis** Hypothermia is defined as a core temperature <35°C

Other important tests

1. Full blood count, electrolytes, glucose, renal and liver function tests—note that severe hyperkalaemia is common in profound hypothermia. Serum potassium must be monitored closely during rewarming, even if initially normal
2. Calcium (low), amylase (high)—in pancreatitis, an important complication of hypothermia
3. Thyroid function tests—?hypothyroid
4. Arterial blood gases—to look for hypoxia and/or acidosis
5. ECG—look for sinus bradycardia, J wave ('junctional' wave—a broad slurred deflection that is superimposed on the distal limb of the QRS complex), prolonged QT interval. Note that the muscular tremor of shivering can lead to artefact on the ECG, which should not be confused with ventricular fibrillation
6. Chest radiograph—look for aspiration, pneumonia, pulmonary oedema

Further management	Dependent on the cause of hypothermia. Prevention of recurrence is likely to require socio-economic intervention in the elderly, e.g. provision of heating, increased supervision

10 Practical procedures

10.1 Central vein cannulation, arterial cannulation and invasive monitoring

10.1.1 Femoral vein cannulation

The optimum position is with the patient supine, but their head and torso can be propped up to an angle of 15 to 30° if this is more comfortable. The key landmark is the femoral pulse, which should be palpated one finger-breadth below the crease of the groin. The femoral vein lies one finger-breadth medial to the femoral artery (the mnemonic NAVY, Nerve Artery Vein Y-fronts, can be useful in remembering the anatomy). The needle should therefore enter the skin one finger-breadth medial to the femoral artery and one finger-breadth below the groin crease (Fig. 1). It should be advanced in the line of the leg, angled rostrally at about 60° to the skin, and with its bevel pointing forwards. When the vein is punctured the guidewire should pass directly up the femoral vein and into the inferior vena cava.

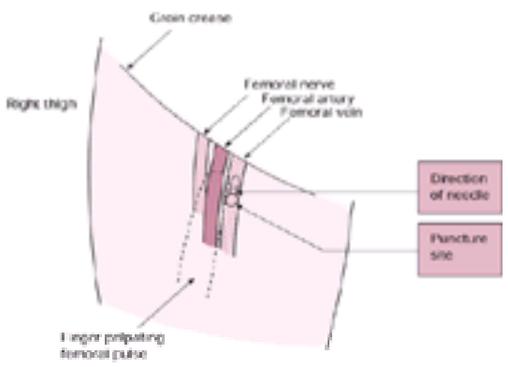

Fig. 1 The approach to the femoral vein.

10.1.2 Internal jugular vein cannulation

10.1.2.1—The low lateral approach

The patient is supine with the head turned away from the side of the puncture. A towel may be placed under both shoulders to extend the neck. After preparation of the skin and drapes, and insertion of local anaesthetic, the bed is tilted to a 25° head down position. The needle is inserted just lateral to the posterior border of the clavicular head of the sternocleidomastoid muscle, about one finger-breadth above the clavicle, with its bevel pointing caudally. It is then advanced parallel to the line of the clavicle and just behind the sternocleidomastoid muscle. The internal jugular vein, which lies superficially at this point, is cannulated close to its junction with the subclavian vein (Fig. 2(a)). As soon as the vein is entered the needle is angulated caudally to ease cannulation, the guidewire passing directly into the innominate vein. The risk of complications was lower with this technique than for any other method of central venous cannulation used in one series of over 5400 cases (see Chapter 16.1).

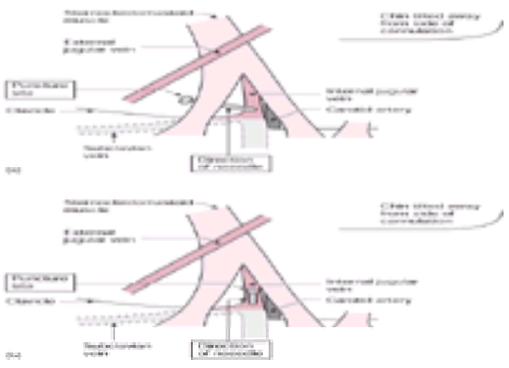

Fig. 2 (a)The low lateral approach to the internal jugular vein. (b)The axial approach to the internal jugular vein.

10.1.2.2—The axial approach

The patient is positioned as described for the low lateral approach to the internal jugular vein (Fig. 2(b)). The needle is inserted in the centre of the triangle defined by the sternal and clavicular heads of the sternocleidomastoid muscle and the clavicle itself. It should be angulated caudally, at about 60° to the skin, and in a line pointing towards the ipsilateral anterior superior iliac spine.

10.1.3 Subclavian vein cannulation

10.1.3.1—The infraclavicular approach (Fig. 3(a))

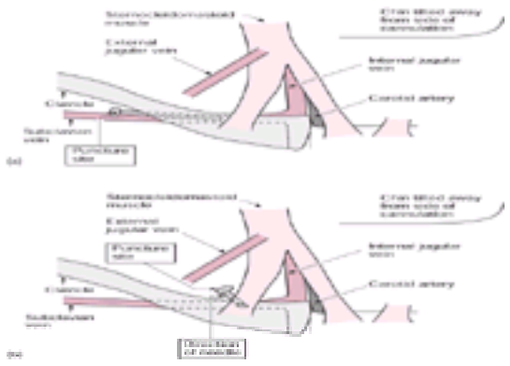

Fig. 3 (a)The infraclavicular approach to the subclavian vein. (b)The supraclavicular approach to the subclavian vein.

The patient is positioned as described for the low lateral approach to the internal jugular vein, excepting that instead of a towel being placed under both shoulders it should be positioned under the spine, allowing the shoulders to retract to reduce the risk of pneumothorax. The needle enters the skin below the mid-point of the lower border of the clavicle and is advanced under the clavicle towards the upper edge of the junction of the clavicle with the manubrium.

10.1.3.2—The supraclavicular approach (Fig. 3(b))

The patient is positioned as described for the infraclavicular approach to the subclavian vein. The needle is inserted into the angle between the superior border of the clavicle and the posterior border of the clavicular head of sternocleidomastoid and advanced caudally, medially and ventrally.

10.1.4 Pulmonary artery flotation catheter

Central venous cannulation should be performed as described above (section 10.1.2 and section 10.1.3) and a pulmonary artery (PA) catheter introducer inserted. Ensure that the balloon at the end of the PA catheter inflates completely and uniformly, and then slowly advance the catheter whilst watching the pressure trace on the monitor. The balloon should be inflated when the

catheter is advanced and deflated whenever the catheter is withdrawn. Pressure traces corresponding to the right atrium, the right ventricle and the pulmonary artery should be seen (Fig. 4). As a rough guide, the waveform should change for every 10 cm that the catheter is advanced. Inflation of the balloon when the catheter is in a medium sized pulmonary artery allows it to 'wedge' and occlude distal flow. To obtain valid readings, the catheter tip should reside in a region of the lung where pulmonary venous pressure exceeds alveolar pressure. The pressure recorded at the tip of the catheter (pulmonary capillary wedge pressure, PCWP) provides indirect measurement of the left atrial pressure, which reflects left ventricular end-diastolic pressure if the chamber is not diseased. Values for cardiac output and mixed-venous blood chemistries may also be directly measured. A number of variables such as systemic vascular resistance and left ventricular stroke work may be derived from values measured with a PA catheter.

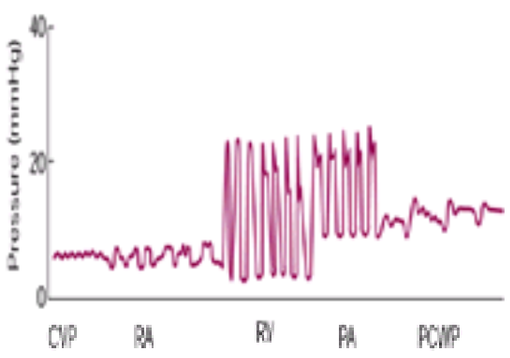

Fig. 4 Pressure tracings obtained on insertion of a pulmonary artery flotation catheter. CVP, central venous pressure, RA, right atrium, RV, right ventricle, PA, pulmonary artery, PCWP, pulmonary capillary wedge pressure.

10.1.5 Arterial puncture/cannulation

Before attempting to puncture or cannulate the radial artery, check the patency of the ulnar artery by applying pressure to the radial artery and asking the patient to clench their fist firmly. On relaxing the fist, the hand should pink up within 10 s (Allen test).

A 25G needle (orange) is perfectly adequate to obtain an arterial blood gas (ABG) sample from a radial artery. 18G (green) or 23G (blue) is needed for a femoral sample. Use either a preheparinized ABG syringe or draw up 1 ml of 1000 u/ml heparin into a syringe and then completely expel the heparin.

Lay the index and middle fingers of your non-dominant hand along the line of the artery as a guide (Fig. 5). For radial and brachial samples, hold the syringe at 45 to 60° to the skin and slowly advance in the line of the artery. For femoral samples, hold the syringe at 90° to the skin. A flash of blood into the syringe indicates successful puncture. Some syringes will fill to a predetermined volume, others require aspiration of 1 to 2 ml.

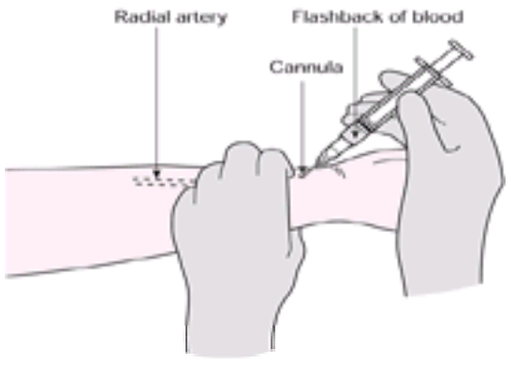

Fig. 5 Puncture of the radial artery.

After successful arterial puncture, press on the site for 3 min (5 if anti-coagulated) to prevent haematoma formation.

Arterial cannulation may be performed either with a cannula over a needle (similar to a venflon) or with a Seldinger technique. After preparation of the skin and insertion of local anaesthetic, the method of arterial puncture should be as described above, with the exception that for all arterial cannulations the needle should be inserted at 45° to the artery. Once arterial puncture has been confirmed, the cannula should be advanced over the needle, or the guidewire passed directly into the artery and the cannula then advanced over the guidewire.

10.2 Cardiac procedures

10.2.1 DC cardioversion

Synchronized cardioversion is the treatment of choice for symptomatic tachyarrhythmias. Conscious patients must be anaesthetized or sedated. Suitable monitoring and facilities for dealing with cardiac arrest should be available. Modern defibrillators incorporate a switch that allows the shock to be synchronised with the R wave of the ECG to reduce the risk of inducing ventricular fibrillation. Gel pads should be applied to the chest wall and cardioversion carried out in the same manner as for defibrillation. The energy required depends on the underlying rhythm. Synchronization means that there may be a delay between pressing the defibrillator buttons and the discharge of the shock when the next R wave occurs.

10.2.2 Cardiac pacing (temporary)

Indications for emergency/acute temporary cardiac pacing are shown in Table 24.

10.2.2.1 External (transcutaneous) pacing

10.2.2.1.1 Percussion pacing Percussion pacing can be used as a temporising measure in some patients with profound bradycardia causing clinical cardiac arrest. It is particularly useful for ventricular standstill where P waves are visible on the ECG. A series of gentle blows should be applied to the lower left sternal edge using the closed fist. Using trial and error, a site can sometimes be found which results in stimulation of the ventricular myocardium. If percussion does not produce a cardiac output, orthodox pacing or CPR should be instituted immediately.

10.2.2.1.2 Transcutaneous pacing Most modern transcutaneous pacing systems are integrated with an ECG monitor/defibrillator. Pacemaker electrodes should be placed in either an anterior-posterior position or in the conventional anterior-lateral configuration. The pacemaker should be set to demand pacing to prevent a stimulus from falling on the T wave following a sponteaneous heart beat, with the rate set at 60 to 90/min for adults. The pacemaker current should be set at the lowest setting and gradually increased to obtain capture of the myocardium and a palpable pulse. The current required to obtain capture is generally in the range 50 to 100 mA and will produce painful contraction of the patient's skeletal muscle. Conscious patients will require analgesia and/or sedation. If capture of the myocardium does not occur, alternative electrode placement should be tried.

Transcutaneous pacing is only a temporizing measure and arrangements should be made for urgent transvenous pacing.

10.2.2.2 Transvenous pacing (ventricular)

Temporary transvenous pacing can be achieved after cannulation of any central vein, but is most easily performed via the right internal jugular, right subclavian or right femoral vein, which can be cannulated as described in section 10.1.1, section 10.1.2 and section 10.1.3.

The conventional Seldinger technique of guidewire and dilators is used to allow placement of a sheath (preferably haemostatic) of sufficient size to accept passage of the pacing wire in the vein that has been cannulated. The pacing wire is passed down the sheath and advanced towards the heart, the aim being to maneouvre it under fluoroscopic guidance into a position where its tip is at the apex of the right ventricule, angulated slightly downwards. Key aspects of the technique are shown in Fig. 6. Common problems and their solutions are described in Table 25.

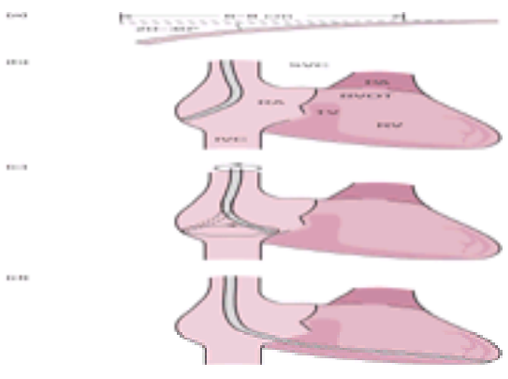

Fig. 6 Correct positioning of the electrode is helped if there is a 20 to 30° curve at the tip of the pacing wire. Mould the electrode to this shape using your fingers: it may need to be bent or straightened depending on its packaging. (a). Advance the wire until it lies vertically in the right atrium. It will usually assume a position where its tip points towards the free wall on the right side (b). Rotate the wire between your index finger and thumb until it points towards the patient's left (c). When it has done so, advance the wire steadily: it should pass through the tricuspid valve and along the floor of the right ventricle to the apex (d). SVC, superior vena cava; RA, right atrium; IVC, inferior vena cava; TV, tricuspid valve; RV, right ventricle; RVOT, right ventricular outflow tract.

After positioning the pacing wire, set the pacemaker to a rate of 70/min, or 10/min above the patient's ventricular rate, and to deliver a pulse of 3 V (or as directed by the manufacturer). A correctly positioned electrode should 'capture', such that each pacing spike is followed by a ventricular complex on the ECG. Establish the voltage threshold by gradually turning down the amplitude of the voltage delivered until capture is lost, which will usually be in the range 0.7 to1.0 V. To allow a safety margin, it is then appropriate in most circumstances to set the pacemaker to deliver a voltage of at least twice the threshold. Sensing can be checked only if there is spontaneous ventricular activity: this is best done by setting the pacemaker rate to between10 and 20/min below the spontaneous ventricular rate and looking on the ECG monitor and the pulse generator for evidence of pacing inhibition. Sensitivity is usually set to to its maximum. Common problems and their solutions are described in Table 26.

When the pacing wire is positioned appropriately and pacing is established, carefully remove the introducer sheath (in most cases), secure the wire with a strong suture (usually 2/0 silk), loop it once or twice on the skin, and then dress with a clear adhesive dressing.

Further reading

Fitzpatrick A, Sutton R (1992). A guide to temporary pacing. *British Medical Journal* **304**, 365–9.

Gammage MD (2000). Electrophysiology: temporary cardiac pacing. *Heart* **83**, 715–20.

10.2.3 Pericardiocentesis

Cardiac tamponade is the indication for pericardiocentesis as an emergency. Unless the patient is *in extremis* the procedure should, whenever possible, be performed with echocardiographic guidance by an operator experienced in the technique, as follows:

1. Two-dimensional echocardiography is used to assess the size, distribution and haemodynamic effect of the effusion.
2. The ideal entry site for pericardiocentesis is the point on the skin where the effusion is closest to the transducer and the fluid accumulation is maximal. The distance from the skin to the pericardial space is estimated, with the needle trajectory defined by the angulation of the hand-held transducer. A straight path that best avoids vital structures (also the internal mammary artery, which lies 3 to 5 cm lateral to the sternal margin) is chosen.
3. After preparation of the skin and insertion of local anaesthetic, a 16 to 18 gauge polytef-sheathed (or similar) needle attached to a saline-filled syringe is advanced in the predetermined trajectory, with continued gentle aspiration as it moves forward. On entering the pericardial fluid, the needle is advanced approximately 2 mm further, when the sheath is advanced over the needle and the steel core withdrawn.
4. The position of the sheath in the pericardial space can be confirmed by injecting 5 ml of agitated saline through it, whilst observing the pericardial space with 2D-echocardiography (optional).
5. Intrapericardial pressure can be directly measured with a manometer (optional); pericardial fluid can be sent for diagnostic tests (optional).
6. A guidewire is advanced through the polytef sheath, which is removed over the guidewire. A small stab incision of the skin is made at the entry site, following which dilators are used to allow the insertion of a larger sheath (6–8 F) through which a pigtail angiocatheter can be introduced. After the pigtail catheter has been inserted the introducer sheath is removed, leaving only the smooth-walled pigtail catheter in the pericardial space. (Note that this technique is preferred to that of introducing the pigtail catheter directly over the guidewire because the catheter tip can occasionally pull the guidewire out of the pericardial sac, particularly if this is sclerotic.)
7. Pericardial fluid is drained completely by syringe suction and the pericardial catheter is secured to the chest wall by suture and appropriate dressing.
8. If left on continuous drainage, pericardial catheters become plugged. It is therefore better to perform intermittent aspiration, every 4 to 6 h or as clinically indicated, leaving the catheter flushed with saline in between times. It can be removed when drainage has been reduced to less than 25 to 30 ml/day and follow-up echocardiography shows no significant residual effusion (sooner if the catheter is causing problems).

   If the patient is *in extremis* and/or echocardiography (with appropriate expertise) is not available, then a 'blind' subxiphoid approach is most often used:

1. Sit the patient up at an angle of 45°.
2. Insert the needle 3 cm below the xiphisternum at an angle of 30 to 45° to the skin and advance, applying gentle suction all the time (as above), in a line towards the patient's left shoulder.
3. If the needle touches the heart, it will usually provoke ectopic beats. Some authorities recommend that the aspiration needle is attached to the 'V' lead terminal of an ECG cable (using insulated wire with a clip on each end, or simply with sticky tape) to allow continuous monitoring. If the needle touches the heart, then the character of the ECG changes, most particularly with the appearance of gross ST-segment elevation if the needle touches the right or left ventricle.
4. When fluid is obtained, proceed as described above.

Further reading

Tsang TS, Freeman WK, Sinak LJ, Seward JB (1998). Echocardiographically guided pericardiocentesis: evolution and state-of-the-art technique. *Mayo Clinic Proceedings* **73**, 647–52.

10.3 Arterial blood gases (Table 27)

10.4 Airway and respiratory procedures

10.4.1 Mechanical support of ventilation

10.4.1.1 Continuous positive airways pressure

Continual positive airway pressure (CPAP) exerts a dilating force on the upper airway (hence its use in obstructive sleep apnoea), and also recruits collapsed alveoli and increases functional residual capacity. This improves lung compliance, reducing the work of breathing, which is a benefit in a range of clinical circumstances.

CPAP can be used for patients with acute or acute on chronic hypoxaemia who are not exhausted or in ventilatory failure (meaning elevated *p*CO2), e.g. acute pulmonary oedema, postoperative atalectasis, pneumonia. It is not appropriate and is contraindicated for patients who are too obtunded to cooperate, who are unable to protect their airway, who have haemodynamic instability or life-threatening arrhythmias, life-threatening hypoxaemia, or exhaustion.

CPAP is applied via a tight fitting face or nose mask, the usual range for pressure being 2.5 to 10 cmH2O. Once applied, patient comfort, respiratory rate and arterial blood gases should be monitored. Some patients are unable to tolerate the face mask: gastric distension, vomiting, aspiration, eye irritation, conjunctivitis, and facial-skin necrosis are other complications.

10.4.1.2 Non-invasive positive pressure ventilation

Masks that are used for CPAP can also be used to provide non-invasive positive pressure ventilation (NIPPV, often more simply referred to as NIV). The difference between the two treatments is that in CPAP a constant pressure is applied to the airway, but no airflow occurs in the absence of respiratory muscle activity. By contrast, in NIV a pulse of positive pressure is applied to assist respiration, the usual arrangement being that this is triggered by a sensor that detects a fall in pressure in the facial mask when the patient initiates a breath. If a positive pressure is also applied in the expiratory phase (EPAP) in addition to the pulse delivered to support inspiration, then then this is known as bilevel pressure support (BIPAP).

Contraindications for and complications of NIV are the same as those for CPAP.

10.4.1.3 Invasive ventilation

Invasive ventilation may be applied via a tracheal tube or tracheostomy. Ventilation can be adjusted by altering the minute volume (respiratory rate × tidal volume). Oxygenation is adjusted by altering inspired oxygen concentration and positive end-expiratory pressure (PEEP, which acts in a similar manner to CPAP by recruiting collapsed alveoli and reducing the work of breathing). Most ventilators for adults are volume generators that deliver a fixed tidal volume regardless of changes in lung mechanics. If the lungs become stiffer, then inflation pressure will increase to deliver the same tidal volume.

The change from inspiration to expiration is usually time cycled; that from expiration to inspiration is usually either time cycled or triggered by the patient if they are breathing spontaneously. The following values can be used as a guide when initially setting up a ventilator for an adult:

- Tidal volume should be 10–15 ml/kg
- Respiratory rate 10–12/min
- Ratio of inspiratory to expiratory time (I:E ratio) set at 1:2, but for patients with chronic obstructive pulmonary disease or asthma, a smaller I:E ratio is often used (e.g. 1:3) to prevent gas trapping and hyperinflation.
- Concentration of inspired oxygen depends on the clinical context: the patient with normal lungs who requires respiratory support because of respiratory muscle weakness does not need a high *F*iO2 (start with say 28 per cent), whereas the patient with severe problems with gas exchange, eg. bilateral pneumonia or acute respiratory distress syndrome, will require a high *F*iO2 (start with say 60–80 per cent).

Once ventilation is established, the various parameters should be adjusted (and others added, e.g. CPAP) according to the patient's clinical condition and the results of repeated measurement of arterial blood gases.

10.4.2 Management of the airway

10.4.2.1 Endotracheal intubation

Endotracheal intubation remains the gold standard for airway management as it provides a method of oxygenating and ventilating the patient, whilst securing the airway from vomitus and secretions.

Intubation should be preceded by ventilation with high concentration oxygen. The neck should be slightly flexed and the head extended (an assistant holding the neck in a neutral position if trauma to the cervical spine is suspected). The mouth should be inspected for loose teeth or dentures, which should be removed, as should any secretions or vomitus (by suction). A trained assistant should apply cricoid pressure to prevent passive regurgitation.

The laryngoscope should be introduced over the right side of the tongue, moving the tongue to the left. The tip of the blade should be positioned in the valecula (between the epiglottis and the base of the tongue) and lifted upwards and away from the operator to expose the vocal cords (Fig. 7(a)). The endotracheal tube should be introduced so that the cuff is positioned just beyond the cords (Fig. (b)).

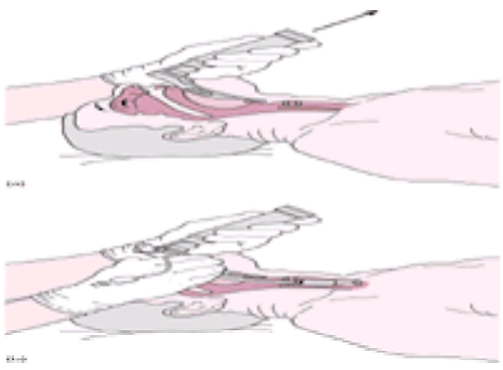

Fig. 7 (a) Position of the laryngoscope before insertion of the endotracheal tube. (b) Placement of the endotracheal tube.

After successful intubation, the patient should be ventilated with high concentration oxygen, the endotracheal tube secured and the tube cuff inflated. Positioning of the endotracheal tube should be confirmed by listening over the apices and the bases of the lungs, and over the stomach. If available, an end-tidal carbon dioxide monitor should be attached to the endotracheal tube.

10.4.2.2 Laryngeal mask airway

The laryngeal mask airway (LMA) is used widely in routine anaesthetic practice and is increasingly used for immediate airway management in cardiac arrest. Pulmonary aspiration associated with the use of a LMA is uncommon provided high inflation pressures are avoided.

The patient should be supine with the neck slightly flexed and the head extended (an assistant holding the neck in a neutral position if trauma to the cervical spine is suspected). The LMA should be held like a pen, and introduced into the mouth with the distal aperture facing towards the patient's feet. The tip should be applied to the palate and advanced until it reaches the posterior pharynx. The LMA is then pressed backwards and downwards until the resistance of the hypopharynx is felt (Fig. 8), when the cuff of the LMA should be inflated. If insertion is satisfactory, the end of the LMA will rise slightly. Positioning of the LMA should be confirmed by listening over the apices and the bases of the lungs, and over the stomach. If available, an end-tidal carbon dioxide monitor should be attached.

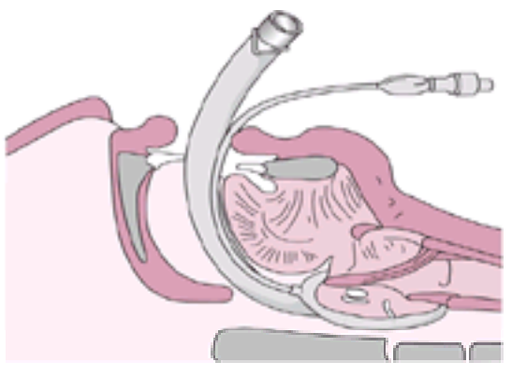

Fig. 8 Laryngeal mask airway inserted into the hypopharynx. The inflated cuff surrounds and isolates the entrance to the larynx.

10.4.2.3 Cricothyroidotomy

10.4.2.3.1 Needle cricothyroidotomy Insertion of a needle or a cannula (typically a large bore intravenous cannula) through the cricothyroid membrane is a useful emergency technique that allows short-term provision of oxygen until a definitive airway can be placed. The cannula should be connected to high flow oxygen with either a Y connector or a side hole cut into the tubing between the cannula and the oxygen supply (Fig. 9). Intermittent insufflation can be achieved by closing the Y connector or side hole with a thumb for one second and then releasing it for three seconds. Inadequate exhalation leads to accumulation of carbon dioxide, hence this technique of ventilation can only be used for 30 to 45 minutes.

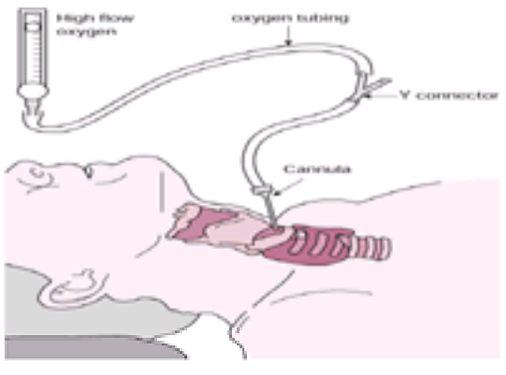

Fig. 9 Oxygenation via a cannula through the cricothyroid membrane.

10.4.2.3.2 Surgical cricothyroidotomy The skin over the cricothyroid membrane should be cleaned and local anaesthetic inserted (in patients who are conscious). A horizontal skin incision is made and extended through the cricothyroid membrane (Fig. 10). A curved haemostat (forceps) is then used to dilate the opening and a small, cuffed endotracheal tube or tracheostomy tube inserted. The position of the tube should be confirmed by auscultation of the lungs and over the stomach, and the tube then secured.

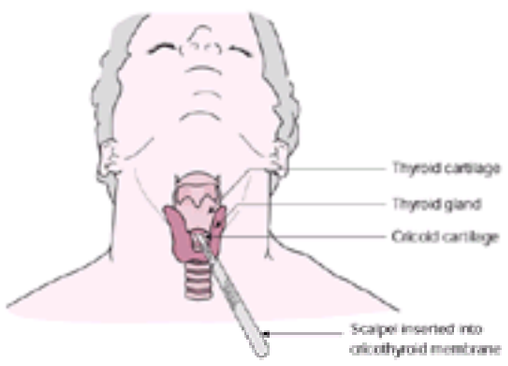

Fig. 10 Technique of surgical cricothyroidotomy.

10.4.3 Percutaneous procedures on the chest

10.4.3.1 Chest decompression

The rapidly deteriorating patient with clinical signs of a tension pneumothorax requires immediate needle decompression of the chest. The second intercostal space on the side of the tension pneumothorax should be identified, and an over the needle cannula or any hollow needle should be inserted in the midclavicular line, directing it just superior to the rib into the intercostal space. Listen for a sudden escape of air when the needle enters the pleural cavity. The cannula should be secured and arrangements made for an intercostal drain to be inserted as soon as the tension pneumothorax has been decompressed.

10.4.3.2 Chest aspiration

Chest aspiration may be considered for any symptomatic spontaneous pneumothorax, irrespective of its size. The advantages over intercostal tube drainage are that, if successful, needle aspiration means a shorter hospital stay, less pain, and less scarring.

The British Thoracic Society guidelines recommend an anterior approach using the second intercostal space in the midclavicular line. Other authors recommend a posterior approach, with the patient in a sitting position with the arms gripping the knees, and using the secone, third, or fourth intercostal space medial to the scapula.

The skin should be prepared and local anesthetic infiltrated down to the pleura. A 16 G cannula is inserted perpendicular to the skin and just over the superior border of the rib. The cannula is then connected to a three-way tap, the second port of which is connected to a 50 ml syringe, and the third to a length of tubing that runs to open under the surface a container of sterile water (Fig. 11). Aspiration should be continued until either resistance is felt or the patient coughs excessively.

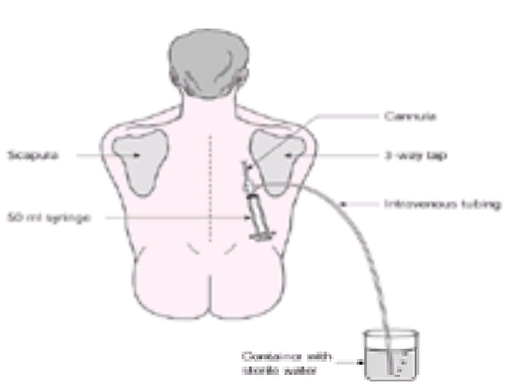

Fig. 11 Chest aspiration (posterior approach).

The success (or otherwise) of aspiration can be determined by repeat chest radiography. The procedure may be repeated if not successful or if the pneumothorax recurs. Success is less likely for older patients and in those with chronic lung disease or recurrent pneumothorax, also after aspiration of more than 2.5 litres.

10.4.3.3 Chest drain

Always confirm the correct side for chest tube insertion. The usual site is the fourth to sixth intercostal space anterior to the midaxillary line. Position the patient supine with the head of the bed slightly elevated and the patient's arm behind their head. Clean and drape the area for tube insertion.

Infiltrate local anaesthetic down to the parietal pleura (10–20 ml of 1 per cent lignocaine). Make a 2 to 3 cm transverse incision at the site and blunt dissect through the subcutaneous tissues, just over the superior surface of the rib. Puncture the parietal pleura with the end of the dissection forceps and insert a gloved finger into the incision to ensure that the pleural space has been entered.

Remove the trocar from the intercostal drain and slide the drain over your finger into the pleural cavity, when 'fogging' of the tube should be seen. Connect the end of the intercostal tube to an underwater-seal apparatus and confirm correct placement by ensuring that the fluid level is swinging with respiration.

Insert two 3/0 monfilament sutures at 90° to the line of the skin incision, one on either side of the chest drain, but do not tie them. They will be used to close the skin when the chest drain is removed (and are much better than a purse string suture, which produces an unsighly scar, for this purpose). Suture the tube in place with a separate 1/0 or 2/0 silk suture, tied around it as many times as its length allows. If the skin incision is gaping on either side of the drain, close this with one or more 3/0 sutures. Place a gauze dressing around the site and secure with strong tape, wrapping some of this around the tube to secure it firmly.

Obtain a chest radiograph to confirm satisfactory placement and effect of the chest drain.

10.5 Lumbar puncture

Ensure that there are no contraindications to lumbar puncture (LP), namely raised intracranial pressure, bleeding tendency, local sepsis, posterior fossa or spinal cord mass lesion.

The patient should be positioned on the bed (Fig. 12(a)) with their knees drawn up towards the chest to open the space between the spinous processes and with their

spine parallel to the bed. Prepare and drape the skin and locate the puncture site (L3/L4 or L4/L5).

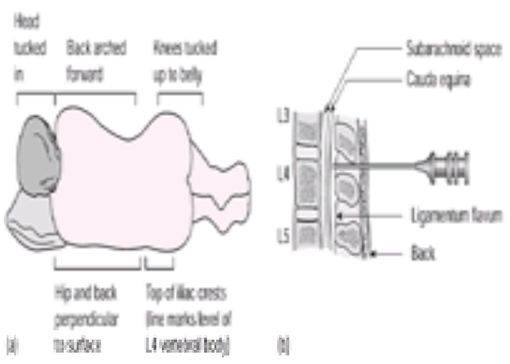

Fig. 12 (a) The patient should lie curled up to increase the spave between the vertebrae. (b) The needle should be slowly advanced until it penetrates the ligamentum liavum.

Anaesthetize the skin and subcutaneous tissues using 5 to 10 ml of 1 per cent lignocaine. Insert the LP needle at 90° to the skin. Advance slowly, aiming between two spinous processes (Fig. 12(b)). As the needle enters the dural space, there is a slight loss of resistance. Remove the stylet and ensure that CSF drips freely from the needle. If it does not, insert the stylet and advance the needle a few millimetres then check again.

Check the opening CSF pressure using a manometer (normally 6–15 cmH_2O) then collect CSF samples. The red cell count in consecutive samples can sometimes help to distinguish subarachnoid haemorrhage from a bloody tap, but this is not always reliable and the sample should also be examined for xanthochromia (oxyhaemoglobin and bilirubin) when subarachnoid haemorrhage is possible. Always send blood samples for glucose and protein estimation at the same time, the CSF glucose concentration normally being 60 to 80 per cent of the blood level.

The patient should be asked to remain lying flat for 2 to 4 h to reduce the severity of post-LP headache.